Szegedi Tudományegyetem
Juhász Gyula Pedagógusképző Kar

Csallner András Erik

Bevezetés az SPSS statisztikai programcsomag használatába

Jegyzet

SPORTINFORMATIKA SZAKIRÁNYÚ TOVÁBBKÉPZÉS

Szeged, 2015

Tartalomjegyzék

1. Bevezetés 4
2. Az SPSS felhasználói felülete 5
   2.1. A program felépítése és indítása 5
   2.2. Adat szerkesztő 6
   2.3. Adatbeviteli lehetőségek 10
      2.3.1. Az elsődleges adatbevitel 10
      2.3.2. A másodlagos adatbevitel 11
   2.4. Eredmények 12
   2.5. Menüelemek 12
   2.6. Feladatok 25
3. Leíró statisztika 26
   3.1. Alapfogalmak 26
      3.1.1. Helyzetmutatók 27
      3.1.2. Szóródásmutatók 29
      3.1.3. Alakmutatók 29
      3.1.4. Egyéb mutatószámok 30
   3.2. Példa a mutatószámok kiszámítására 31
   3.3. Feladatok 45
4. Faktoranalízis 46
   4.1. Alapfogalmak 46
      4.1.1. A faktoranalízis megvalósíthatóságának feltételei 47
      4.1.2. A faktorok számának meghatározása 48
      4.1.3. Faktorok rotálása 50
   4.2. Példa a faktoranalízisre 50
   4.3. Feladatok 64
5. Korreláció 65
   5.1. Alapfogalmak 65
   5.2. Példa a korreláció kiszámítására 68
   5.3. Feladatok 78
6. Regresszió 79
   6.1. Alapfogalmak 79
      6.1.1. Lineáris regresszió 79
      6.1.2. A legkisebb négyzetek módszere 80

6.1.3. Az illesztés és a becslés jósága 80
6.1.4. Hipotézisvizsgálat 81
6.1.5. Reziduálisok vizsgálata 82
6.2. Példa regressziószámításra 82
6.3. Feladatok 88
7. Kereszttábla elemzés 89
7.1. Alapfogalmak 89
7.1.1. A cellák tartalma 89
7.1.2. Kereszttábla statisztikák, a khi-négyzet próba 89
7.2. Példa kereszttáblák használatára 90
7.3. Feladatok 97
8. Klaszteranalízis 98
8.1. Alapfogalmak 98
8.1.1. A klaszterelemzés technikája 98
8.1.2. A klaszterelemzés korlátai 99
8.1.3. Vizsgálatok 99
8.1.4. Hierarchikus összevonó eljárások 100
8.1.5. Nem hierarchikus eljárások 101
8.2. Példa klaszteranalízisre 102
8.3. Feladatok 132
9. Irodalomjegyzék 133

1. Bevezetés

Amikor az olvasó kezébe veszi e tankönyv elektronikus vagy nyomtatott formáját, tudni kell, hogy szerkesztésnél a következő alapelveket tartottam szem előtt. Mindenekelőtt szerettem volna egy könnyen értelmezhető, minden SPSS vagy táblázat- és adatbázis kezelési előtanulmányokkal nem rendelkező olvasó számára is teljes értékűen használható könyvet készíteni az SPSS használatáról. A könyvben igyekeztem a már jól bevált, a témával foglakozó szakirodalom legjobb könyveinek szerkesztési alapelveit követni, és mind a témák sorrendiségében, mind a példákkal történő bőséges illusztrálásával szerettem volna segíteni a megértést. Mindemellett cél volt, hogy ne váljon túl dagályossá, amely már puszta méreteivel elriasztja a tanulni vágyó olvasót.

Reméljük, haszonnal forgatja ezt könyvet.

Egyben itt szeretném kifejezni köszönetemet Devosa Ivánnak és Maródi Ágnesnek, akik nélkül ez könyv nem készülhetett volna el.

2. Az SPSS felhasználói felülete

A Statistical Package for the Social Sciences, vagyis a társadalomtudományok számára kifejlesztett statisztikai programcsomag összetett adatbázisok gyors és hatékony feldolgozását teszi lehetővé. A rendszer megismerése során egy laikus is el tudja készíteni mindezt, ám akár saját parancssorozatot is tudunk a feladatokhoz rendelni. A fejezet a legfontosabbnak vélt menüpontokat mutatja be a legrészletesebben, mivel a program kiváló súgó menüvel (*HELP*) rendelkezik.

2.1. A program felépítése és indítása

A menürendszer nagyban hasonlít a Microsoft Office programcsomagnál megszokottakhoz: vannak olyan műveletek, melyek itt is ugyanúgy alkalmazhatóak – másolás, kivágás, beillesztés, törlés –, illetve találunk eltérőeket is – visszavonás csak az utolsóra terjed ki, a beillesztés (*PASTE*) pedig nem szúr be oszlopokat és sorokat, így adatvesztés lehetősége nagyobb figyelmetlen használat esetén.

Az indításkor megjelenő párbeszédablak azokat az első lépéseket tárja elénk, amik közül választhatunk:

Run the tutorial: olyan oktatóprogram, amely részletes leírást nyújt a használat során, így főleg a kezdők számára ajánlott

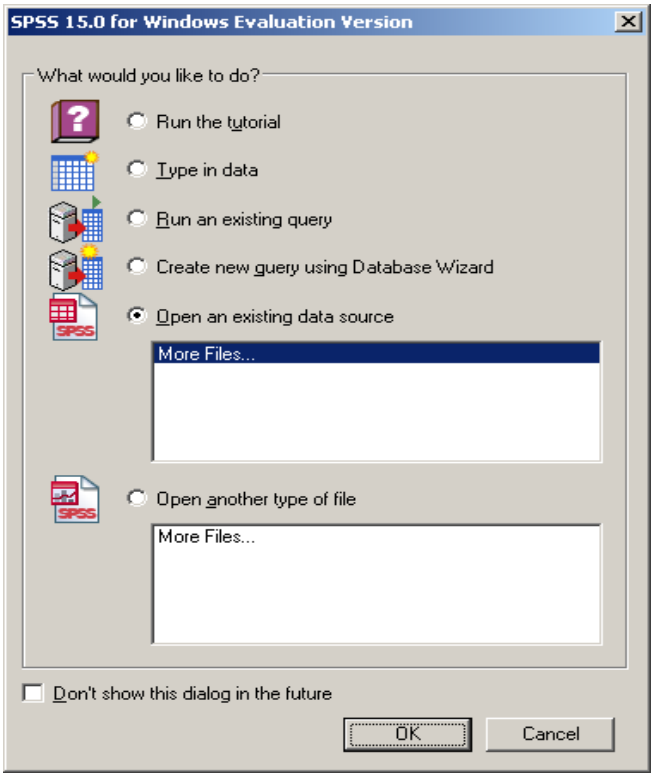
Type in data: adatok begépelését teszi lehetővé

Run an existing query: lefuttat egy már meglévő lekérdezést, illetve keresőkifejezést

Create new query using Database Wizard: adatok más adatbázisból történő bemásolására alkalmas

Open an existing dara source: egy már létező SPSS-adatállományt tölt be – az opció választása során meg kell ezt határozni

Open another type of file: más típusú fájt tölt be (nem SPSS) – szintén meg kell határozni még itt magát az adatbázis helyét.



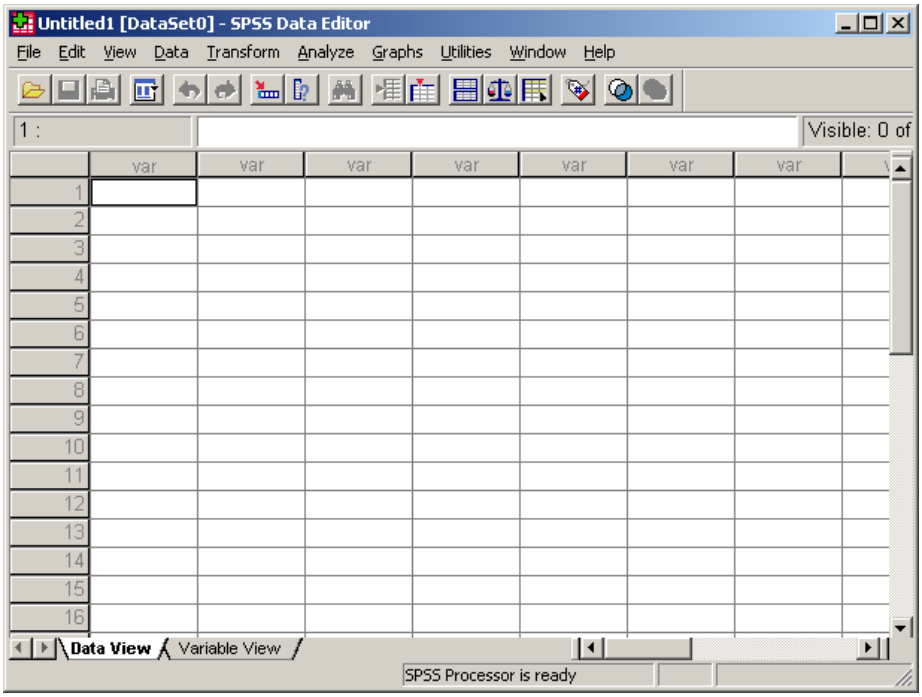
1. ábra

Új dokumentum készítését vagy a TYPE IN DATA lehetőség megjelölése, vagy a párbeszédablak elvetése, CANCEL, teszi lehetővé. Ezt követően egy Data Editor, vagyis adatszerkesztő ablak jelenik meg, ahol a változókat és az ezekhez tartozó adatokat rögzíteni lehet.

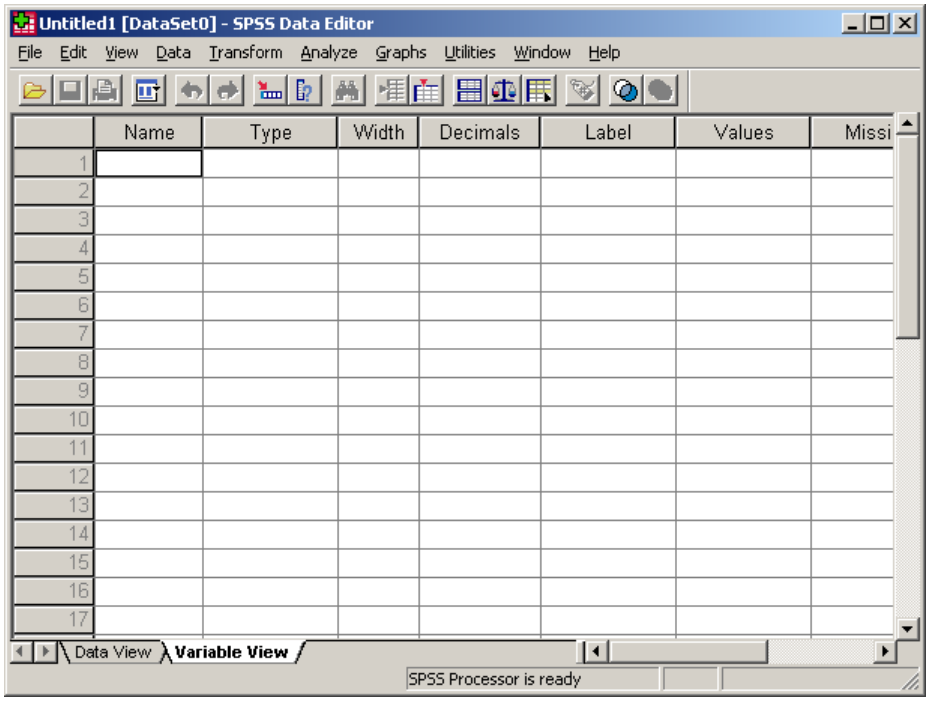
2.2. Adat szerkesztő

Mint már fentebb említettük, az SPSS menürendszere és a képernyő tartalmak elrendezése hasonlít az MS Office programcsomagból ismert programokéhoz, így ez a munkaablak a Microsoft Office Excellel mutat formai azonosságokat. Ennek a legszembetűnőbb jele a táblázatokra épülő szerkesztő panel. A többi elemzőprogramhoz hasonlóan itt is fontos, hogy melyik adat hol helyezkedik el. Az SPSS-ben a függőleges oszlopok alkotják a változókat, a vízszintes sorok pedig az ezekhez tartozó adatokat tartalmazzák, melyeket rekordoknak, eseteknek

„*case*” nevezünk. Az Excel alul található *munka1, munka2* elnevezésű füleinek itt a DATA VIEW [=adat nézet, vagyis maguk az adatok] (2.ábra) és a VARIABLE VIEW [=változó nézet, azaz az oszlopok nevei] (3.ábra) feleltethető meg. A lapok közötti váltást az egérrel a kívánt lap alsó fülére történő rákattintás vagy a Ctrl+t teszi lehetővé. Az előbbi lap az alapértelmezett, az utóbbi lap pedig kizárólag a változók szerkesztésére szolgál (új változók beállítása, paraméterek módosítása).



2. ábra



3. ábra

A **Variable View** ablakban a sorok tartalmazzák a változókat, az oszlopok pedig ezek tulajdonságait:

Name: a változó rövid nevét kell bevinni – azért célszerű rövidebb nevet megadni, hogy jobban átlátható legyen a táblázat. Amennyiben a mezőt üresen hagyjuk, az automatikusan generált elnevezés a VAR00001, ahol a sorszám értéke nő a sorokkal.

Type: a változó típusát és formáját kell meghatározni, de a műveletet soronként el kell végezni – a rendszer ad választási lehetőségeket:

Numeric (numerikus): legegyszerűbb formában jeleníti meg a számokat, ez a leggyakrabban használt forma például: 25704,20

Comma (vessző): jelölései: tizedesvessző (.) és ezres helyiérték (,). Például: 25,704.20

Dot (pont): jelölései: tizedesvessző (,) és ezres helyiérték (.). Például: 25.704,20

Scientific notation (tudományos alak): a számok normálalakban vannak: 1-10 közötti szám + a megfelelő hatvány szorzatai. Például: 587 = 5,87E2 = 5,87*100.

Date (dátum): év/hónap/nap sorrendjének beállítása.

Dollar (dollár): pénzben mért érték jelölésére alkalmas.

Custom currency (speciális pénzformátum): azok közül a pénzformátumok közül lehet választani, melyeket ezelőtt az OPTIONS menüben beállítottunk.

String (szöveges változó): szöveges adatok tárolása nyílt kérdés – egyéni válasz estén. Például: Miért…?

Width (szélesség): A Data View ablak rekordjai mennyi karaktert tartalmaznak cellánként.

Decimals (tizedes jegyek száma): mennyi karakter található a tizedesvessző után.

Label (címke): a változó jelentését, vagy magát a változót lehet magyarázni itt – a későbbi output táblák (VIEWER és a CHART EDITOR) is ezt jeleníti meg, valamint a Data View is, amennyiben a változó nevére irányítjuk az egeret.

Values (érték): a változó értékeinek definiálása itt lehetséges – pl. hőmérsékleti értékek. Ez a VALUE LABELS ablakban lehetséges az „Add" gombra kattintva, majd a folyamat befejeztével az „OK" gombot kell választani, és bezárni az ablakot.

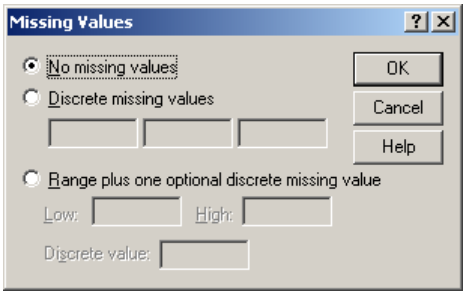
Missing (hiányzó érték): olyan értéket rendelünk hozzá, amely az adat hiányát, a nem kielégítő választ mutatja például: magyarországi átlaghőmérsékletek esetén 50°C-vagy ennél nagyobb adat. Ha ez a hozzárendelés nem történik meg, számos hibát eredményezhet. A hiányzó értéket általában 9-esekből álló olyan számmal jelöljük, ami érték nem fordul elő az adott változóban.

A MISSING VALUE ablakban van minderre lehetőség, ahol három változat közül választhatunk (4. ábra):

No missing values: ha nem adunk meg hiányzó értéket, azt a program egy .-tal jelöli.

Discrete missing values: egyedi kódot adhatunk meg a hiányzó értékekre (maximum 3 darabot).

Range plus one optional discrete missing value: számtartomány vagy egy tartományt és egy különálló értéket.



4. ábra

Columns (oszlopok): A Data View oszlopszélességének a mértéke, amely nem kisebb, mint maga a kitöltő szöveg hossza.

Align: A Data View cellatartalmainak igazítása: jobbra, balra, középre

Measure (mérési skála): a skálatípust kell megadni:

Scale: metrikus – intervallum- vagy arányskála.

Ordinal: sorrendi skála.

Nominal: névleges, nominális skála.

2.3. Adatbeviteli lehetőségek

Az adatok bevitele után regisztrálni kell azokat – erre két mód létezik: elsődleges (kódolási útmutató és a definiálandó változók ismeretével történő begépelés) és másodlagos (létező adatbázisok importálása).

2.3.1. Az elsődleges adatbevitel

A változók definiálása (INSERT VARIABLE) után a rekordok begépelése (INSERT CASES) történik, ez a sorok és oszlopok előtti jobb egérkattintással és az ikonokkal lehetséges. Ha a definiálást a Variable View ablakban tesszük, akkor egyedül a változók paramétereit kell beállítani – míg Data View esetében az új változó a legutolsó oszlopba kerül, és a paraméterek meghatározása egyedül dupla kattintással lehetséges.

A változókat célszerű előbb Variable View nézetben definiálni, és csak ezután átlépni Data View nézetbe.

A paraméterek meghatározása egyesével, vagyis cellánként történik: *name* és *label*: begépelés; *type, values, missing*: cella jobb oldalán lévő gombra kattintva a megjelenő panelt töltjük ki, a többi esetén a legördülő sáv lehetőségeit használjuk. A szöveges adatokat szöveges változókban tároljuk (*string*), a szám jellegű adatokat pedig numerikusban. Ha numerikus adatoknál nem szeretnénk tizedesjegyeket megjeleníteni, akkor a *decimals* értékét 0-ra kell megadni.
Ezt követően Data View nézetre váltunk, ahol a már meghatározott változókat töltjük ki az adatokkal.
Ha a View menüpontnál a Value Labels opciónál pipát látunk, akkor nem a változók nevei, hanem a megfelelő változó *label* mezőjének értéke látszik. Utóbbi pedig sokkal egyértelműbb lesz nem csak a táblázatban, de a kimeneti táblázatokban és grafikonokon is.

2.3.2. A másodlagos adatbevitel

Másodlagos adatbevitel esetén a kutató Excel fájlban megkapja az adatokat tartalmazó táblázatot. Ez esetben a FILE / OPEN DATABASE / NEW QUERY menüpontot választjuk, majd ezen belül is az „Excel fájl" lehetőségét, mert .xls kiterjesztésű a forrásként szereplő állomány.
A „Tovább" gomb használatával egy új ablak kerül elénk – ahol a „Browse" gomb megnyomásával – az importálni kívánt (jelen esetben .xls) fájl elérési útját kell megadni.
A forrásból a megfelelő adatokat tartalamzó munkalapot áthúzzuk a jobb oldali ablakba – mindezt úgy tehetjük meg, hogy az egérrel rákattintunk a mozgatni kívánt névre, majd áthúzzuk arra a helyre, ahova szeretnénk.
Az SPSS program felismeri a tartalmat, így a változókat és az eseteket is helyesen értékeli. A „Tovább" gomb kétszeri megnyomásával megkapjuk a végeredményt. A köztes állapotban az újrakódolásra, illetve a változók meghatározására van lehetőség.
Az így kapott adatbázis tartalmilag megegyezik az elsődleges adatbevitel során kapott eredménnyel, csupán a változók néhány paraméterében van különbség. Ami a .xls fájlban oszlopcím, az itt a név (*name*) lesz.
Létezik egy gyorsabb módszer is (OPEN FILE / DATA - .xls kiterjesztésű fájt kiválasztása) minderre, ám annak az a hátránya, hogy átalakítás előtt nem változtathatunk a változókon. Itt a program automatikusan beolvassa és a forrás fájl első sora alapján definiálja az SPSS változóit.

2.4. Eredmények

Ismertebb neve Output ablak (kimeneti, illetve statisztikai), amely a program lefuttatása után kapott eredményt szemlélteti táblázatok és diagramok formájában. Az elkészített statisztikák a program bal oldalán fa szerkezetben vannak, és piros nyíl jelzi, hogy melyik van kirészletezve a jobb ablakban. Az így kapott eredményeket lehet szerkeszteni, formázni dupla kattintás által. Ezt követően a jobb oldali egérgomb megnyomása során felkínált lehetőségekből kiválasztjuk a számunkra legmegfelelőbbet.

2.5. Menüelemek

File menü

A fájlkezelő műveleteket találjuk benne:

New: új adat- (*Data*) vagy output fájl (*Output*) létrehozása.

Open: már létező adatfájl vagy output állomány megnyitása.

Open Database: új vagy már létező SQL szervezésű adatfájl megnyitása.

Read Text Data: szövegformátumú állomány megnyitása.

Save: az adott fájl mentése a kijelölt helyre.

Save As: kijelöljük, hogy hova, milyen néven és fájltípusban legyen az adott fájl mentve – a „*Save*” első használat során automatikusan ideirányítja a felhasználót.

Save All Data: összes nyitott állomány mentése.

Mark File Read Only: az adott fájl megjelölése oly módon, hogy ezt követően csak olvasni lehessen, vagyis semmilyen javítás nem lesz engedélyezett rajta a későbbiek folyamán.

Rename Dataset: adatbázis el- vagy átnevezése – a fájlnév mellett az adatbázis is rendelkezhet saját névvel, amely akkor hasznos, ha több azonos nevű fájl létezik.

Display Data File Information: A .sav kiterjesztésű fájlokról ad információt egy külön ablakban. A WORKING FILE-ra kattintva a betöltött fájlról kapunk információkat (*variable*: változó, *position*: pozíciója, *label*: címke, *measurement level*: mérési szint, *column width*: oszlopszélesség, *alignment*: igazítás módja), míg

az EXTERNAL FILE során egy külső adatbázisról tudjuk meg ezeket az adatokat.

Cache Data: a *Cash Now* futtatása alatt az adatokon nem lehet változtatni, ám az adatok áttekintése gyorsabbá válik a folyamat után.

Print: nyomtatási beállítások megadása, amely az adott ablakra vonatkozik.

Print View: a várható eredmény megtekintésére szolgáló ún. nyomtatási kép.

Switch Server: szervergépre való csatlakozás.

Stop Processor: az SPSS számolási egységeinek leállítása, amely hibásan kiadott nagy számolási- és időigényű feladatoknál hasznos.

Recently Used Data: legutóbb használt .sav kiterjesztésű fájlok elérése.

Recently Used Files: legutóbb használt nem .sav kiterjesztésű fájlok elérése.

Exit: program bezárása.

Edit menü

Adatszerkesztéssel kapcsolatos programok, illetve utasítások tartoznak ide:

Undo: utoljára kiadott utasítás visszavonása.

Redo: az „Undo” során visszavontakat teszi érvényessé.

Cut: a kijelölt részlet kivágása.

Copy: a kijelölt részlet másolása – a *„Cut”* és a *„Copy”* során használtakat más alkalmazásba is be lehet illeszteni, mert azt a program a Windows vágóasztalára helyezi.

Paste: A *„Cut”* és a *„Copy”* során kijelöltek adott helyre másolása.

Paste Variables: előzőleg kiválasztottak bemásolása.

Clear: törlés – sorok és oszlopok törlésénél nem lesznek üres cellák.

Insert Variable: új változó, oszlop beillesztése a kijelölt helytől balra – ikonja oszlopok közötti kék ék.

Insert Cases: új eset, sor beillesztése a kijelölt hely fölé – ikonja sorok közti piros ék.

Find: változók keresésére lehet alkalmazni, esetekre nem – ikonja távcső.

Go to Case: megadott esethez viszi a kurzort – ikonja sor fölött álló piros nyíl.

Options: SPSS aktív ablakára vonatkozó beállítások.

Legfontosabb a GENERAL fül: vagy a változó nevét (*Display Name*), vagy azok jelentését (*Display Labels*) látjuk – később bevont változók esetén a jelentés szerepel a felsorolásban, a nevük zárójelben utánuk. Ha a változók nevét jelöljük be, akkor ez az adat áll rendelkezésre a későbbi statisztikai elemzések során.

Viewer: output ablakok beállítása (betűméret és stílus, szám).

Output Labels: az output ablakban megjelenő táblázatokban és grafikonokban a változó neve, jelentése vagy mindkét adat megjelenítésének a beállítása.

Charts: az output ablak grafikonbeállításai.

Interactive: állomány nyomtatási és mentési beállításai.

Pivot Tables: az output ablak táblázatainak formai beállításai.

Currency: pénznemek formai beállításai – tizedesek tagolása, / . ; toldalékokat tartalmaz.

Data: adatok beállításai: új numerikus adatok formátuma (DISPLAY FORMAT FOR NEW NUMERIC VARIABLES), vagy véletlenszám-generátor (RANDOM NUMBER GENERATOR).

View menü

Az aktív ablak megjelenítésére vonatkozó beállításait hajtjuk végre:

Status Bar: állapotsor beállítása, processzor ellenőrzése – ez az állapotsor aktív állapota során lehetséges.

Toolbars: eszköztár megjelenítése, illetve az itt megjelenő parancsok, ikonok beállítása.

Fonts: az éppen használt betű típusáért, stílusáért, méretéért felelős.

Grid Lines: ha aktív, akkor az ablak rácsozása látható, ha inaktív, akkor nem.

Value Labels (értékcímkék): amennyiben aktív, akkor a Variable View során meghatározott változójelentést használja a Data View ablak – ha inaktív, akkor a változó értékét.

Variables / Data: a két ablak között vált.

Data menü

Az adatkezelési lehetőségek:

Define variables properties: változók tulajdonságainak meghatározása/ megváltoztatása. Annyiban tér el a Variable View-tól, hogy itt az értékekhez tartozó esetek kilistázhatóak.

Copy data properties: adattulajdonságok másolása vagy egy külső forrásból ide, vagy pedig innen egy célfájlba.

Define Dates: a dátumformátumban lévő változók meghatározása – év, hónap, nap és másodperc pontosságú időpontok esetén.

Define Multiple Response Sets: többválaszos változó definiálása, amelyeket az ANALYZE / TABLES menüpontban a CUSTOM TABLES vagy a MULTIPLE RESPONSE SETS opció alatt lehet felhasználni táblázat részeiként. A másik lehetőség (ANALYSE – MULTIPLE RESPONSE – DEFINE SETS) csak akkor lesz aktív, ha a változó meghatározása már megtörtént. A menüpontok működése hasonló: azokat a változókat, amelyeket elemezni szeretnénk a „*Variables in set*" ablakba helyezzük, majd megadjuk, hogy egy értéket '*dichotomies*' (igen-nem válaszok esetén) vagy több értéket '*categories*' (több kategória esetén, felsoroláskor) számolunk össze – ha ezt az opciót választjuk, akkor minimum- és maximumérték megadása kötelező. Ezt követően nevet kell adni az új változónak, majd használni – ez az *„Add"* gomb megnyomása után lehetséges. Gyakran átkódolás szükséges ahhoz, hogy ezek csak a számunkra megfelelő értékeket tartalmazzák.

Identify Duplicate Cases: többször előforduló esetek azonosítása akár egy változóval, akár az összessel. A program meghatározza az ismétlődő (*Duplicate)* és az egyedülálló adatokat (*Unique/ Primary*). Az új változó *Primary Last* néven szerepel – hasonló elemekből álló kategóriában az utolsó elem az elsődleges szerepű.

Sort Cases: az esetek sorba rendezése az általunk megadott szempontok szerint.

Transpose: az adatbázis sorainak és oszlopainak felcserélése, amely során az eddigi funkciójuk is megváltozik.

Restructure: a „Transpose" menüpont kiegészítése – nemcsak a teljes adatbázist lehet felcserélni, hanem néhány általunk kijelöltet is.

Merge files: az esetek és a változók összefűzését teszi lehetővé egy vagy több állomány esetén.

Aggregate: adatok tömörítésére szolgál az általunk megadott összevonás által. Megkülönböztetünk csoportosító (*break variable*) és összevonni kívánt változót (*summaries of variables*). Az új változót vagy az eredeti adatbázisba helyezzük vissza, vagy pedig másik fájlban helyezzük el.

Orthogonal Design: a merőleges kivitelezést az összekapcsolt elemzések során használják, amelyet e könyv nem taglal.

Copy Dataset: az egész adatbázis másolása, amely során megbizonyosodik a program használója arról, hogy az elvégezni kívánt változtatások során az eredeti állományt nem írja felül.

Split File: a program csoportokra bontja az állományt, hogy ezeken hajtsa végre a statisztikai elemzést – aktiválása során a „*Split File On*" felirat jelenik meg, ikonja egy kettévágott adatbázis.

Select Cases: kizár az SPSS bizonyos elemeket az elemzésből az általunk megadott feltételeknek megfelelően – az eljárás során feleslegessé vált elemek fekete vonallal lesznek áthúzva, valamint a „*Filter On*" felirat látható.

Weight Cases: javítási lehetőség egyes elemek súlyozása során – túlprezentáltakat kisebb, az alulprezentáltakat nagyobb értékkel korrigálja a program.

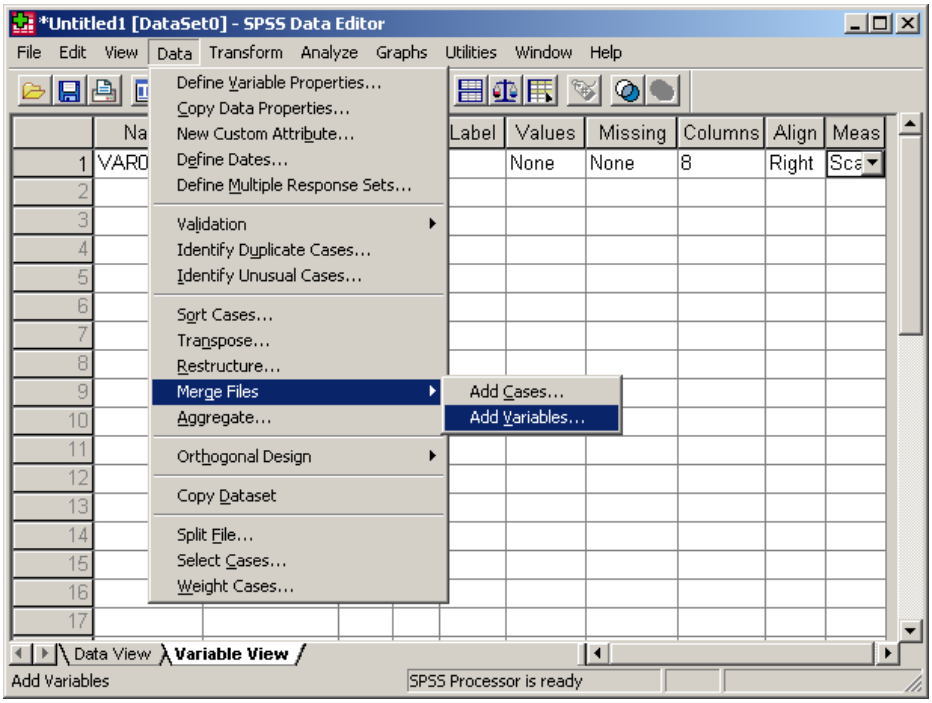
Data / Merge Files – Fájlok egyesítése

A régi állományhoz való hozzárendelés lépései:

1. lépés: a DATA menü MERGE FILES menüpontját kell használni:

Add Cases: a változók megegyeznek az eredeti adatbázissal, így csak az új eseteket illesztjük a régiekhez.

Add Variables: az esetek megegyeznek az eredetivel, így csak az új változókat illesztjük a régiekhez.



5. ábra

2. lépés: a hozzácsatolni kívánt fájl megadása külső adatfájlt használva (an external SPSS data file).

3. lépés: azok az adatok, amelyet a program párosítani tudott a „VARIABLES IN NEW ACTIVE DATASET" ablakban jelennek meg – „OK" gombra kattintva a kibővített adatbázis jelenik meg.

Data / Select Cases – Esetek kiválasztása

Ezt a menüpontot akkor használjuk, amikor az adatelemzésben nincs minden adatra szükség, mert csak az eseteket vizsgáljuk.

1. lépés: „DATA menü SELECT CASES" menüpont megnyitása – szűrőfeltételeket is meg kell adni az alábbi opciók segítségével:

All Cases: nincs szűrés, minden eset részt vesz az elemzésben.

If condition is satisfied: elemeket választunk ki relációs jelek, függvények, logikai feladatok során.

Random sample of cases: véletlenszerűen választja ki az eseteket vagy százalékos arány (*approximately*) vagy konkrét számmennyiség (*exactly*) megadásával.

Based on time or case range: sorrendiség vagy szűrési idő szerint választ.

Use filter variable: megadunk egy változót, amit a rendszer szűrőváltozóként használ.

Output: a szűrés eredményének sorsát adjuk meg.

Filter out unselected cases: a nem választott adatok az ablakban maradnak, ám figyelmen kívüliek – „Filter On" felirat.

Copy selected cases to a new dataset: új adatbázisba kerülnek a kiválasztott adatok.

Delete unselected cases: töröljük az adatbázisból a nem kiválasztott adatokat – használata nem ajánlott.

2. lépés: „IF" szűrőfeltétel alkalmazása során szűkítjük le az adatbázist, ezt követően a „CONTINUE" gombra kattintva lépünk tovább.

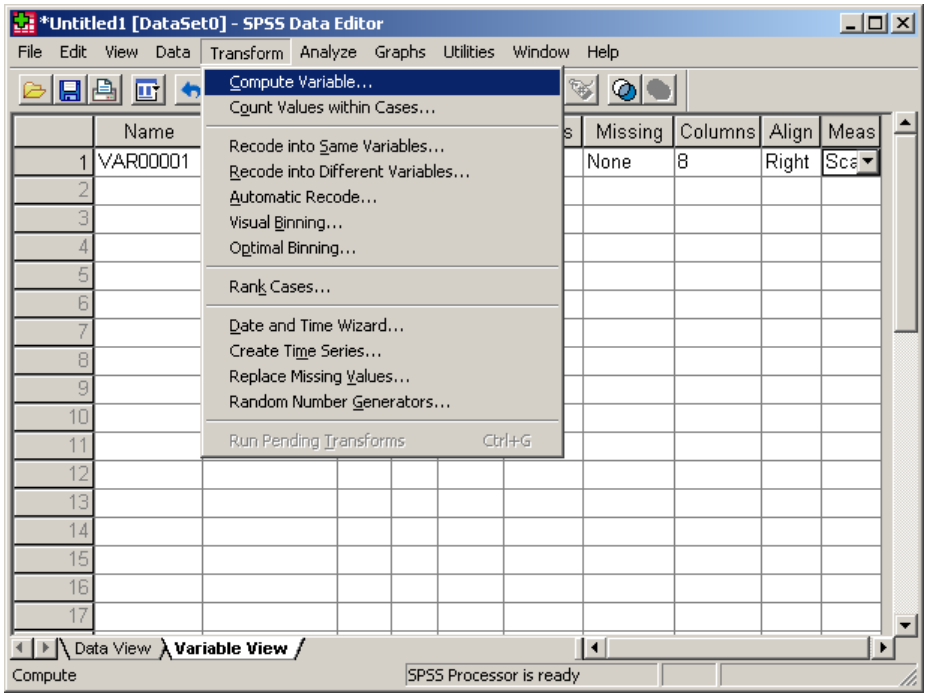
3. lépés: a „Filter out unselected cases" beállítás választása során a többi adat nem kerül törlésre, csupán figyelmen kívül hagyja a program az elemzés során – ezeknek az elemeknek a jelölése áthúzással történik, valamint egy új változó keletkezik, a „filter_$", amely értékei a 0-t (nincsenek benne az vizsgálatban, vagyis az áthúzott elemek) és az 1-et (benne vannak) vehetik fel. (*A szűrés visszaállítása nélkülözhetetlen a folyamat lefutása után – erre az „All cases" opció választásával van lehetőség.)*

Transform menü

Ez a menü is adatkezelési lehetőség – főleg akkor, ha régi változókból állítunk elő újat, vagy eseteket újrakódolunk.

Compute – új változó számítása

Ez a menüpont új adatbázis előállítását teszi lehetővé a régi felhasználásával – a kettő között vagy függvényszerű vagy logikai viszony van.



6. ábra

1. lépés: A TRANSFORM / COMPUTE választásával kapott ablakban a NUMERIC EXPRESSION panelbe azt a képletet kell bevinni, amely segítségével az új adatbázist szeretnénk létrehozni. A TARGET VARIABLE szövegdobozba pedig ennek a nevét írjuk be.

2. lépés: A TYPE AND LABEL opció használatával a változó tágabb értelemben vett jelentését adjuk meg a „*Label*" sorban, amely a „*Continue*" lenyomásával rögzül.

3. lépés: a szűrőfeltételek megadása az alapablakban lehetséges, ha az **„*If*"** gombra kattintunk.

4. lépés: A „*Continue*" majd az *„Ok"* egymást követő lenyomásával a végeredmény megjelenik.

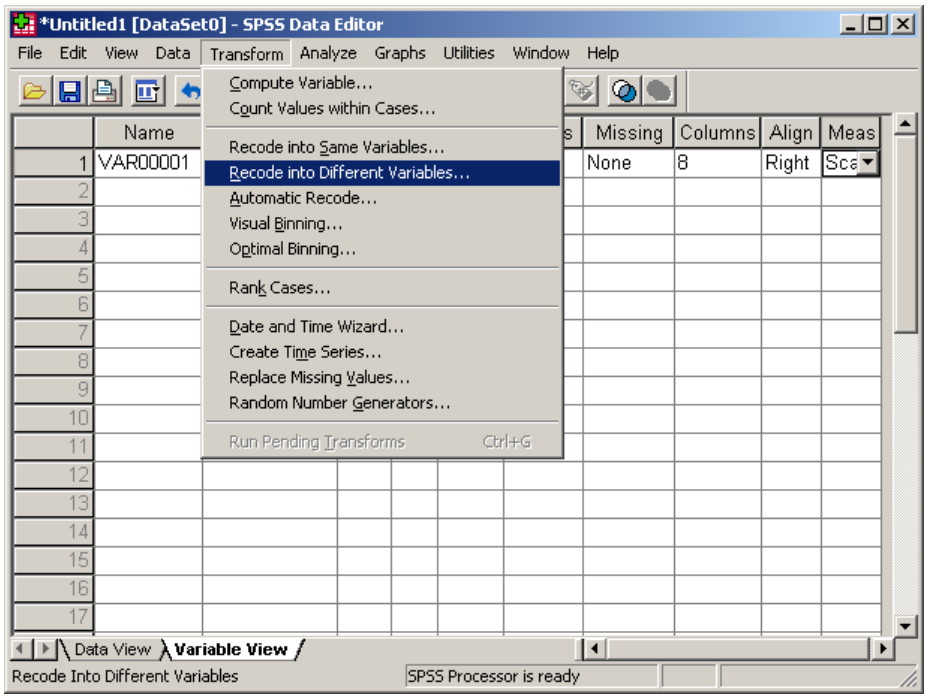
Recode – átkódolás

A már létező változók átkódolását, módosítását végezzük el ebben a menüben az alábbi lehetőségek szerint:

Into Same Variables (ugyanazokba a kódokba): ha az átkódolás után nincsen szükség az eredeti változóra, akkor ez felülírja a régit.

Into Different Variables (más változókba): megtartja a felülírandó adatot – új változó nevét és paraméterét meg kell adni – ez a módszer eredményesebb lehet a későbbi vizsgálatok során, hiszen nem történik adatvesztés.

1. lépés: TRANSFORM / RECODE / INTO DIFFERENT VARIABLES alkalmazása.



7. ábra

2. lépés: Pl. városokat jelképező kódok alapján történik a szelekció – ez azonban szöveges változót eredményez. Ezt a STRING VARIABLE → OUTPUT VARIABLE panel jelzi. Ez utóbbi nevét (*Name*) és címkéjét (*Label*) megadjuk – a változtatás mentése a „*Change*” gomb lenyomásával történik. Szűrőfeltételeket az „*If*”-fel adunk meg. A megfeleltetéseket (régi

változó – új változó) az „*Old and New Values*” gomb megnyomásával rögzítjük.

3.lépés: A „*Convert numeric strings to numbers*” opció bejelölése elengedhetetlen, hiszen szöveget alakítottunk adattá. Az „*Add*”-del rögzítjük a szelektálást. A program lehetőséget ad még a rendszer vagy a felhasználó hibájából adódó hiány jelölésére (*System- or user-missing*) és a nem szöveges változók esetén intervallumot adhatunk meg.

Az eredményt a „*Continue*” és az „*Ok”* egymást követő lenyomásával kapjuk meg. Az új változóhoz értéket kell még rendelni a *Variable View* lapon.

Az érték azonnali (Calculate Values Immediately) vagy használat előtti kiszámításának (Calculate Values Before Use) a beállítására is van lehetőség: EDIT / OPTIONS / DATA / TRANSFORMATION AND MERGE OPTIONS.

Count – az előfordulások megszámlálása

Olyan új változót hozunk létre, amely tartalmazza a régi változók együttes előfordulásait.

1.lépés: TRANSFORM / COUNT alkalmazása.

2.lépés: a panel kitöltése oly módon, hogy a „*Target Variable*” tartalmazza az új változó nevét, a „*Target Label”* a jelentését, a „*Numeric Variables*” pedig a csoportosítás alapjául szolgáló régi változókat tartalmazza. A „*Define Values*”-ra kattintva lépünk tovább.

3.lépés: „*Values to Count*” opcióra van szükség a feladat befejezéséhez: itt az előzetesen bevont változók értékeit kell megadni:

A, „*Value*” ablakba az értékeket bevisszük az „*Add*” segítségével

B, „*Range*” (terjedelem) lehetőségbe begépeljük, hogy mettől meddig foglaljuk bele a régi értékeket. Pl. *10 through 28*, azaz 10-től 28-ig.

C, „*Range / Lowest through value*”, amely során a felső korlátot adjuk meg.

A szűrőfeltételeket az „*If*”-fel lehet megadni.

Az új változó új oszlopként(!) jelenik meg.

Rank Cases – esetek rangsorolása

Az eseteket rangsorolja az megadott változók értékei alapján – ezt a „VARIABLES” ablakba visszük be, kivéve, ha csoportokon belül akarjuk a sorrendiséget kiszámoltatni (ez esetben a „BY” ablakot kell kitölteni). Be lehet állítani, hogy az 1. helyezéshez a legnagyobb (*Largest Value*) vagy a legkisebb érték (*Smallest Value*) kapcsolódjon: ASSIGN RANK 1 TO. A RANK TYPES segítségével speciális rangsorolási eljárásokat határozhatunk meg, míg a „TIES” annak megadására ad lehetőséget, hogy az azonos értékű változók milyen rangot kapjanak.

Automatic Record: Automatikus átkódolás

Azokat a változóértékeket kódolja át, amelyek nem alkalmasak a feldolgozásra, vagyis amelyeket statisztikai elemzéshez nem lehet felhasználni. Az átkódolandó változókat átnevezzük a „NEW NAME” segítségével – meghatározhatjuk azt is, hogy a folyamatot a rendszer a legnagyobb (RECORDE STARTING FROM HIGHEST VALUE) vagy a legkisebb elemmel kezdje.

Date / Time

Az idővel és a dátummal kapcsolatos változtatásokat és beállításokat hajthatunk itt végre.

Create Time Series: idősorok létrehozása

Jelen esetben az idősor olyan értékeket tartalmaz, amelyek időben egymás után következnek. Az ilyen jellegű változókból a „CREATE TIME SERIES” menüponttal más jellegű idősort lehet létrehozni. Az adatok olyan különböző időkben lezajlott megfigyeléseket tartalmaznak, amelyek között az eltelt idő, vagyis az intervallum egyenlő. A feldolgozandó adatok kiválasztását követően annak a függvénynek a típusát kell meghatározni („FUNCTION” panel”), amellyel az átalakítást hajtjuk végre – lehet szezonális ingadozás, simítás, mozgóátlagolás, stb. A folyamatot a „*Change*” feliratú gombra kattintva zárjuk.

Replace Missing Values: Hiányzó értékek pótlása

Akkor használjuk ezt az alkalmazást, ha nem használtunk hiányzóérték-kódot, és ha minden esetnél szükséges érvényes ismérvérték. A változók

megadása után a „METHOD” segítségével választjuk ki azokat a lehetséges értékeket, amelyekkel a hiányzó adatokat kívánjuk helyettesíteni – teljes átlag, szomszédos pontok átlaga, szomszédos pontok mediánja, lineáris interpoláció, az adott pontra vonatkozó lineáris trend.

Random Number Generators: Véletlenszám-generátor

Két fajtát tartalmaz a program: SPSS 12 COMPATIBLE-t és MERSENNE TWISTER-t. Míg az előbbi elavult, de a program 12-es verziójával kompatibilis, addig az utóbbi a modernebb és megbízhatóbb. Fix indulóértéket (Fixed Value) az ACTIVE GENERATOR INITIALIZATION menüpontban adhatunk meg.

Visual Bander: Változók kategorizálása

Numerikus változó elemzésénél szükség van az eredeti (folytonos) kategóriákba sorolt változatára is. Ide sorolható a jövedelem elemzése. Ehhez grafikai ábrát hoz létre a program, amelyben a felhasználó adja meg a kategóriák alsó és felső értékeit.

Az osztópontokat a MAKE CUTPOINTS panelben határozzuk meg:

A, *Equal Width Intervals*: azonos szélességű intervallumok létrehozása – megadjuk az első osztópontot, az osztópontok számát és szélességét.

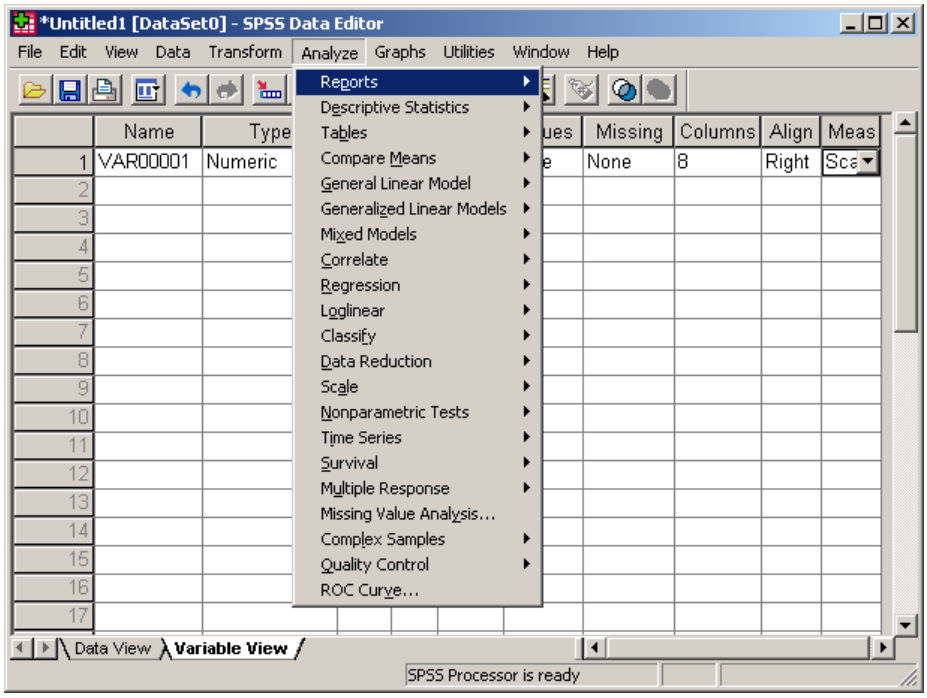
B, *Equal Percentiles*: ugyanannyi esetet tartalmazó intervallum, amelyben a szélesség nem mindig azonos: három kategória létrehozásához két osztópont szükséges (ezt a NUMBER OF CUTPOINTS pontban állítjuk be).

C, *Cutpoints at Mean and Selected Standard Deviations*: meghatározzuk, hogy az átlagon kívül mely szórásértékeknél legyenek osztópontok.

A kiinduló ábrához az „*Apply*” gomb megnyomásával jutunk vissza – ekkor már a két osztópont értékei megjelenítettek, ezeket a hisztogram kék vonallal jelöli. A MAKE LABELS paranccsal a program hozzárendeli az értékekhez (*Value*) a címkét (*Label*).

Analyze menü

A statisztikai számításokhoz szükséges eszközök többsége itt található, a fontosabb menüpontok tárgyalására a későbbi fejezetekben konkrét példákon keresztül kerül sor.



8. ábra

Graphs menü (ábrázolás)

Az itt található grafikonok, ábrák és diagramok a statisztikai elemzés eredményeit és adatait teszik szemléletesebbé, könnyen és gyorsan áttekinthetővé.

Interactive: az említett lehetőségek finombeállításai.

Ultilities menü

Variables: változók paramétereit egyesével megmutatja egy output ablakban.

OMS „***Output Managent System***" ***Control Panel*** (kimeneteli menedzsmentrendszer): a kiválasztott kategóriákat különféle kimeneti-formátumba írja – pl. sav, xml, html, text.

***OMS Identifiers*:** OMS parancsok írása

Data File Comments: adatfájl ellátása megjegyzésekkel

Define Sets: a változókat részhalmazra szűkíti az ide felvett változók megjelölésével és névvel történő ellátásával

Use Sets: az elemzés leszűkítése a változók egy adott részhalmazára

Menu Editor: menüsor szerkesztése, testre szabása

Window menü (ablakkezelés)

A felhasznált ablakok méreteinek beállítása.

Minimize All Window: összes ablakot lekicsinyíti, és a táblára helyezi

Split: képernyő felosztása oly módon, hogy a kisablakokat egymástól függetlenül lehessen mozgatni, ezáltal megtekinteni.

Lehetőség van még az ablakok közötti váltásra.

Help menü (segítség)

Részletes segítség kapható ezen keresztül a program használatáról angol nyelven.

2.6. Feladatok

1. Ismertesse az SPSS adatbeviteli lehetőségeit!
2. Ismertesse az SPSS menürendszerének felépítését!

3. Leíró statisztika

3.1. Alapfogalmak

Leíró statisztika (Descriptive Statistics): olyan eljárás, amelyben kijelentéseink pontosak, a populáció megegyezik a mintával.
Alkalmas:

- számszerű információk gyűjtésére, adatgyűjtésre;
- az információk rögzítésére és jellemzésére;
- grafikus ábrázolásra;
- csoportosításra, osztályozásra;
- egyszerűbb számtani műveletekre;
- az eredmények megjelenítésére.

Mért adatok: A jelenségeket, tulajdonságokat valamilyen mérőskálához hasonlítás alapján számértékkel jellemezzük.
Gyakoriság (abszolút gyakoriság): összeszámoljuk, hogy az egyes csoportokba hány adatot soroltunk.
Gyakorisági eloszlás: a csoportok és a hozzájuk tartozó gyakoriságok összessége. A statisztikai adatsokaságban előforduló lehetséges értékeket a gyakoriságukkal együtt gyakorisági eloszlásnak nevezzük.
Gyakorisági táblázat: A statisztikai adatsokaságban előforduló lehetséges értékeket a gyakoriságukkal együtt egy táblázatba rendezzük.
Egy statisztikai vizsgálat során egy kísérletet mindig többször végeznek el. Az egyes események bekövetkezési számát az esemény *gyakoriságának* nevezzük.
Relatív gyakoriság: Úgy kapjuk meg, hogy az abszolút gyakoriságot elosztjuk a kísérletek számával.
Kumulatív gyakoriság: A relatív gyakoriságok fokozatos összegzésével kapjuk meg.
Egyváltozós elemzések: Azt vizsgáljuk, hogy hogyan oszlanak meg az esetek egyetlen változó szerint, leírás céljából. *Változó* alatt itt a vizsgált jelenség valamely kiválasztott számszerű tulajdonságát értjük.
A statisztikai sokaság mérete általában nagy, ezért fontos, hogy néhány számmal jól tudjuk jellemezni az összegyűjtött adatokat. Ezeket a számokat *statisztikai mutató*knak nevezzük.

Az egyváltozós elemzéseknél leggyakrabban alkalmazott mutatókat négy csoportba sorolhatjuk:

Helyzetmutatók	Szóródási mutatók	Alakmutatók	Egyéb mutatók
Középértékek:	Terjedelem	Csúcsosság	Összeg
Átlag	Szórás	Ferdeség	Minimum
Módusz	Variancia (szórásnégyzet)		Maximum
Medián			Elemek száma
Kvantilisek			

1. táblázat

3.1.1. Helyzetmutatók

Középértékek: a minta eloszlásának alapvető tendenciáját mutatják.
Átlag *(Mean)*: számtani középérték. Az átlag a várható érték torzítatlan becslése.
Fajtái:

Számtani átlag: A számtani átlag az a szám, amellyel az átlagolandó értékeket helyettesítve azok összege nem változik. Kiszámításához összeadjuk az összes adatot, és elosztjuk annyival, ahány adat van.

Mértani átlag: A mértani átlag az a szám, amellyel az átlagolandó értékeket helyettesítve azok szorzata nem változik. Kiszámításához az átlagolandó értékek szorzatából az értékek számának megfelelő (n-dik) gyököt vonunk. Használata akkor célszerű, ha az átlagolandó értékek szorzata értelmezhető.

Harmonikus átlag: A harmonikus átlag az a szám, amellyel az átlagolandó értékeket helyettesítve azok reciprokjainak összege nem változik. Egy felhasználási módja lehet, amikor számtani átlagot kellene számolnunk, de a tényleges gyakoriságok nem ismertek, csak az értékösszegek vagy azok arányai.

Négyzetes átlag: A négyzetes átlag az a szám, amellyel az átlagolandó értékeket helyettesítve azok négyzetösszege nem változik. Úgy számítjuk ki, hogy az átlagolandó értékek négyzeteit összeadjuk, elosztjuk az elemek számával, majd az eredményből négyzetgyököt vonunk. Akkor használjuk, amikor az átlagolandó értékek között pozitív és negatív számok egyaránt vannak, de az előjelnek nincs jelentősége.

Módusz *(Mode)*: A leggyakoribb értéket jelenti a minta elemei között. Lehet több módusz is (például bimodális, trimodális). A módusz alkalmas a várható érték becslésére.

Medián *(Median)*: Az a közbülső érték a sorba rendezett értékek közül, amelyikhez képest a sorba rendezett értékek egyik fele nagyobb, a másik fele kisebb. A sorba rendezett értékek közül a középső, illetve ha két középső van, akkor ezek átlaga. A mediánra közelítő értéket kapunk interpolációval (a középső adatot tartalmazó intervallum alsó határához annyit kell hozzáadni, amennyi az intervallumhosszból arányosan jutna arra az adatra, amennyi az intervallum alsó határa és a középső adat között van). Szélsőséges értékek esetén használható.
Szimmetrikus eloszlás esetén a számtani átlag, a medián és a módusz értéke megegyezik.

Kvantilisek: Speciális helyzetmutatók, a medián általánosításai. Osztópontok segítségével a növekvő sorrendbe állított adataink egyenlő gyakoriságú osztályokra bonthatók.

Típusai:

- A medián 2 egyenlő részre osztja a nagyság szerint sorba rendezett sokaságot 1 osztópont segítségével.
- A tercilis 3 egyenlő részre osztja a nagyság szerint sorba rendezett sokaságot 2 osztópont segítségével.
- A kvartilis (quartilis) 4 egyenlő részre osztja a nagyság szerint sorba rendezett sokaságot 3 osztópont segítségével.
- A kvintilis (quintilis): 5 egyenlő részre osztja a nagyság szerint sorba rendezett sokaságot 4 osztópont segítségével.
- A decilis: 10 egyenlő részre osztja a nagyság szerint sorba rendezett sokaságot 9 osztópont segítségével.
- A percentilis: 100 egyenlő részre osztja a nagyság szerint sorba rendezett sokaságot 99 osztópont segítségével.

Ha a keresett kvantilis sorszáma törtszám, akkor értékét interpolációval kapjuk, ami annyit jelent, hogy a keresett adatot tartalmazó intervallum alsó határához annyit kell hozzáadni, amennyi az intervallumhosszból arányosan jutna arra az adatra, amennyi az intervallum alsó határa és a keresett adat között van.

3.1.2. Szóródásmutatók

Szóródásmutatók: azt mutatják meg, hogy az adatok az átlagtól kevésbé vagy jobban térnek el, azaz hogy az átlag körül mennyire szóródnak az adatok. Ily módon az átlag jóságáról is információval szolgálnak.

Terjedelem *(Range)*: a legnagyobb és legkisebb érték különbsége. Annak az intervallumnak a teljes hossza, amelyen belül a tényleges ismérvértékek mozognak.

Korrigált terjedelem: A terjedelem egy-egy adatra nagyon érzékeny, tehát nagyon nagy lehet ez az érték, ha van egy kiugró adat a többi között (nagyon nagy vagy nagyon kicsi), valójában pedig az adatok egy szám környékén tömörülhetnek. Ez kiküszöbölhető úgy, hogy a legnagyobb és legkisebb adatot (például a legjobban és legrosszabbul teljesítő tanulót) kihagyják az értékelésből. Amennyiben az alsó és felső 1-1%-ot hagyjuk el, a kapott eredményt alsó illetve felső centilisnek, míg az alsó és felső 10-10% elhagyása esetén alsó illetve felső decilisnek nevezzük. Tehát alsó decilis esetében a rangsorba rendezett adatok egytizede kisebb és kilenctizede nagyobb. Felső decilis esetén pedig fordítva. A minta nagysága határozza meg, hogy melyiknek van értelme.

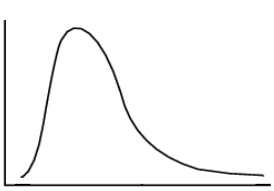
Átlagos eltérés: Azt mutatja meg, hogy az egyes ismérvértékek átlagosan mennyivel térnek el a számtani átlagtól. Hátránya, hogy az eltérések iránya, azaz előjele befolyásolja az értékét.

Szórás *(Standard Deviation):* Azt mutatja meg, hogy az adatok mennyire szóródnak az átlag körül, mennyire heterogén a minta. Valójában az egyes értékek átlagtól való eltérésének négyzetes átlaga. A szórás mindig nemnegatív szám (pozitív vagy nulla).

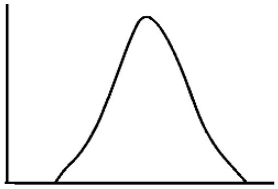
Variancia *(Variance)*: A szórás négyzete, szokás szórásnégyzetnek is nevezni. A négyzetfüggvény miatt hangsúlyosabban emeli ki az eltéréseket.

3.1.3. Alakmutatók

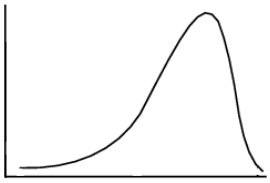
Ferdeség *(Skewness)*: Az eloszlás alakját vertikálisan leíró mutatószám. Az eloszlásnak az a tulajdonsága, hogy milyen irányban tér el a szimmetrikus eloszlástól. A szimmetrikus eloszlás ferdesége 0. Ha a gyakorisági eloszlásnak az oszlopos ábrázolása alapján (hisztogram) az eloszlás jobbra, azaz pozitív értékek irányába elnyúltabb, jobbra ferdének (skewed to right), ha balra, azaz a negatív értékek irányába torzított, akkor balra ferdének nevezzük (skewed to left).



9. ábra: balra ferde



10. ábra: szimmetrikus



11. ábra: jobbra ferde

Csúcsosság *(Kurtosis):* Az eloszlás alakját vertikálisan leíró mutatószám. Relatív fogalom, azt jelzi, hogy az eloszlás az azonos középértékű és szórású normális eloszláshoz viszonyítva az eloszlás csúcsos (jobban tömörül) vagy lapos (kevésbé tömörül). A pozitív értékek viszonylag csúcsos, míg a negatív értékek viszonylag lapos elosztást jeleznek.

3.1.4. Egyéb mutatószámok

Összeg *(Sum)*: A mintában lévő elemek összege.
Minimum *(Minimum)*: A mintában lévő elemek közül a legkisebb elem.
Maximum *(Maximum)*: A mintában lévő elemek közül a legnagyobb elem.
Esetszám *(Number of Cases)*: A megfigyelt esetek száma, a minta nagysága.

Változók / Mutatók	**Intervallum**	**Ordinális**	**Nominális**
Középértékek	Átlag, (Módusz, Medián)	Medián (Módusz)	Módusz
Szóródási mutatók	Szóródás, Variancia	Terjedelem	Gyakoriság, relatív gyakoriság
Alakmutatók	Ferdeség, Csúcsosság	-	-
Egyéb mutatók	Minimum, Maximum	Minimum, Maximum	-

2. táblázat

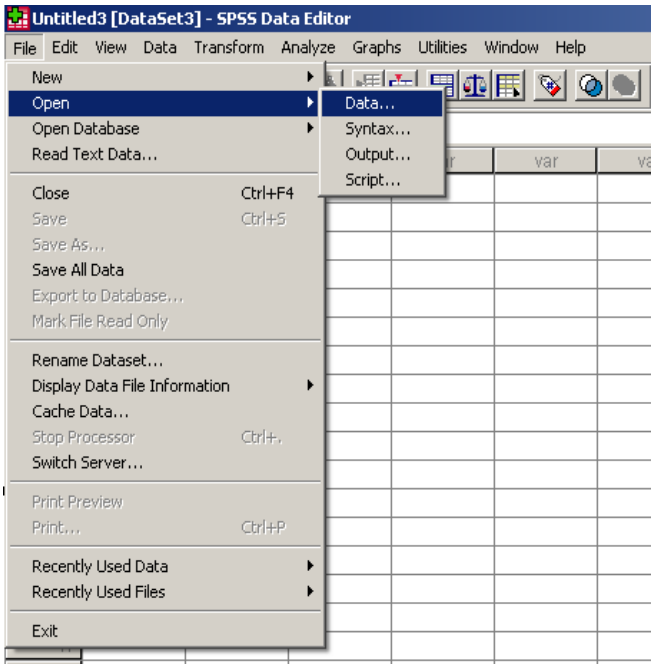
3.2. *Példa a mutatószámok kiszámítására*

A Floridai Egyetemen – 1989 ősze és 1991 tavasza között – a master képzésben részt vett hallgatók (forintra átszámított) kezdő fizetését szeretnénk megvizsgálni. A férfi és női hallgatók az egyetem 8 különböző főiskolai karán (1. agriculture – mezőgazdasági, 2. architecture – építés mérnöki, 3. building/construction – építészeti/épülettervezési, 4. business administration – üzleti tanulmányok, 5. forestry – erdészeti, 6. education – pedagógiai, 7. engineering – mérnöki, 8. fine arts – képzőművészeti) végezhettek.

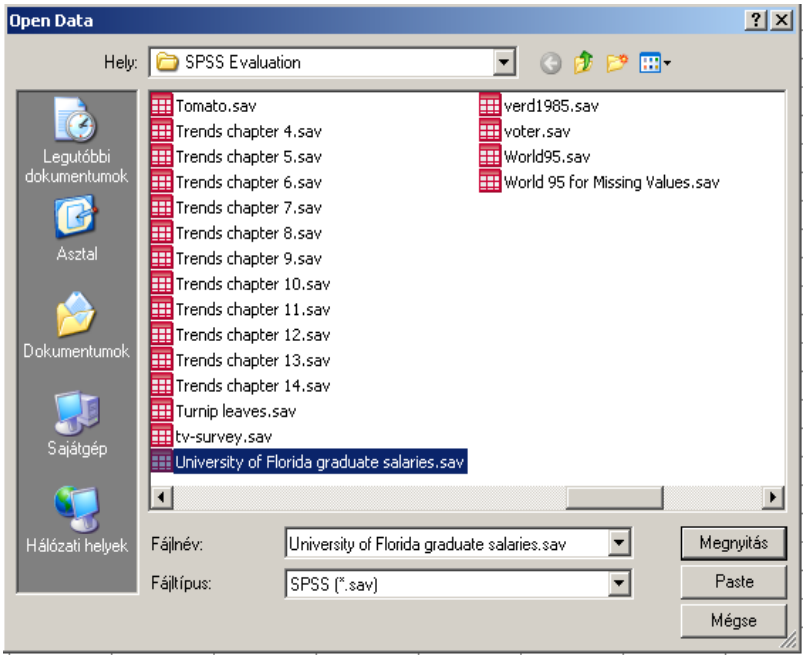
Első vizsgálatunk során kíváncsiak vagyunk arra, hogy mekkora a kezdő fizetések

- számtani,
- mértani, valamint
- harmonikus átlaga.

Az elemzés elvégzéséhez először is nyissuk meg az SPSS példái között található University of Florida graduate saleries.sav nevű fájlt.



12. ábra



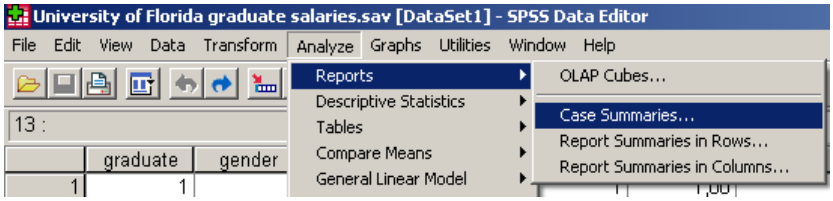
13. ábra

A File / Save As parancs segítségével mentsünk el a fájlt Floridai egyetemisták fizetése.sav néven.

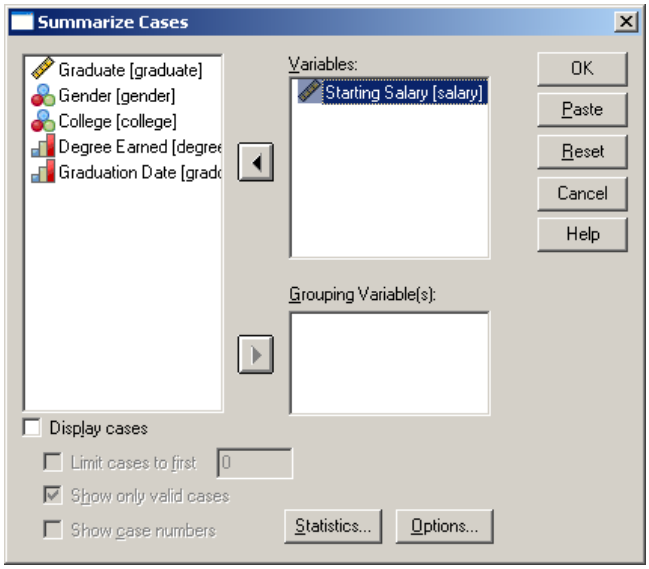
	graduate	gender	college	salary	degree	graddate
1	1	1	7	28900	1	1,00
2	2	1	7	28000	1	1,00
3	3	1	1	27500	1	1,00
4	4	1	7	30300	1	1,00
5	5	1	1	18000	1	1,00
6	6	0	7	31700	1	1,00
7	7	1	3	26000	1	1,00
8	8	1	7	25000	1	1,00
9	9	0	1	20000	1	1,00
10	10	1	1	18000	1	1,00

14. ábra: Floridai egyetemisták fizetése.sav

Az átlagok számításához válasszuk az Analyze / Reports / Case Summaries parancsot. Vigyük át a nyíl segítségével a Starting Salary (kezdő fizetés) változót a Variables alá, majd kattintsunk az OK gombra.

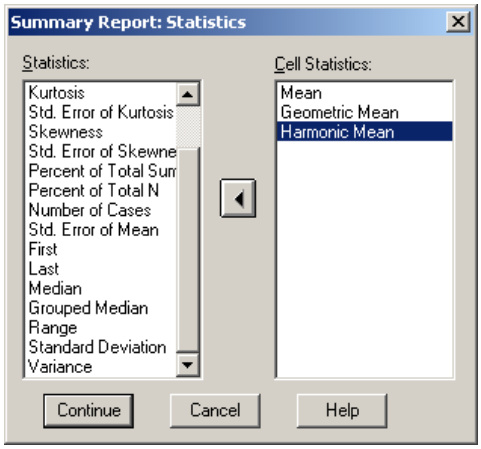


15. ábra



16. ábra

Ezután válasszuk ki a bal oldali listából az átlagot, geometriai átlagot és harmonikus átlagot és a Continue gombra kattintva megkapjuk az eredményeket.



17. ábra

Case Summaries

Starting Salary

Mean	Geometric Mean	Harmonic Mean
26064,20	25090,54	24005,04

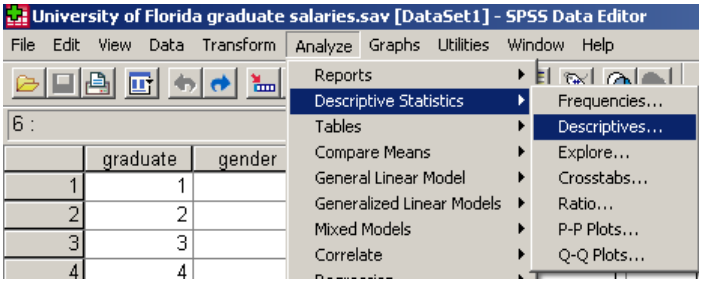
3. táblázat

A táblázatból leolvasható, hogy a kezdő fizetések számtani átlaga 26062,20 Ft, a geometriai átlaga 25090,54, a harmonikus átlaga 24005,04 Ft. Jól látható, hogy a három átlag nem egyforma.

A vizsgálatunk során szeretnénk megtudni, hogy:

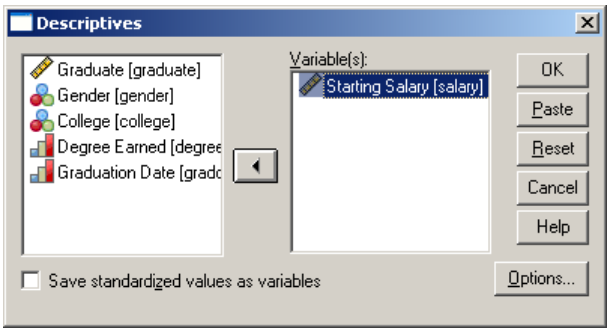
- Kinek a legnagyobb a kezdő fizetése?
- Kinek a legkisebb a kezdő fizetése?
- Mekkora az az összeg, amit legtöbben kapnak?
- Mekkora az az összeg, aminél ugyanannyian kapnak többet, mint ahányan kevesebbet?
- Mekkora a szórás, azaz mennyire tér el az egyes diplomások kezdő fizetése az átlagos kezdő fizetéstől?
- Mekkora az az összeg, amit összesen kapnak kezdő fizetésként?

A vizsgálat elvégzéséhez válasszuk ki az Analyze/Descriptive Statistics/Descriptives parancsot. Tehát a leíró statisztikák leíró menüpontját.



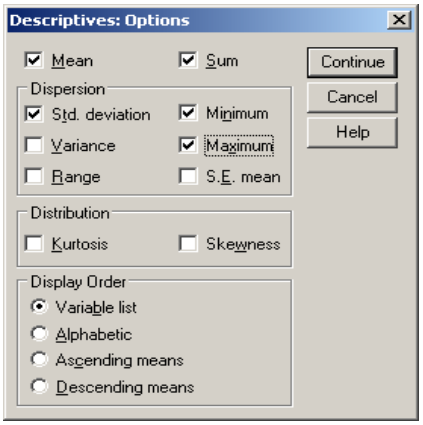
18. ábra

Az előugró Descriptives ablakban válasszuk ki, majd a nyíl segítségével vigyük át a Starting Salary (salary), azaz kezdő fizetés tételt a jobboldalra, a Variable(s) felirat alá. Majd kattintsunk az Options gombra. Ha véletlenül rossz tételt választottunk ki, akkor a nyíl segítségével vissza tudjuk vinni a baloldalra, majd a megfelelő elemet mozgassuk át.



19. ábra

Ekkor egy újabb ablak ugrik elő, a Descriptives: Options. Pipáljuk ki az egér segítségével a kiszámolandó értékeket: az átlagot (mean), az összegzést (sum), a legkisebb elemet (minimum), a legnagyobb elemet (maximum) és a szórást (Std. deviation), majd kattintsunk a Continue gombra. Display Order alatt állíthatjuk be azt, hogy a változók milyen sorrendben szerepeljenek, amennyiben több változónk van. (Variable list: az adatbázis sorrendjében, Alphabetic: ábécésorrendben, Ascending means: az átlagok szerint növekvő sorrendben, Descending means: az átlagok szerint csökkenő sorrendben.) Végül kattintsunk a Continue gombra a folytatáshoz, majd az Ok gombra.



20. ábra

Az Output ablakban megjelenik egy táblázat (4. táblázat), ahol láthatjuk a vizsgálatunk kérdéseinek válaszait:
Az átlagos kezdő fizetés: 26064,20 Ft.
A szórás: 6967,982 Ft.
A legkisebb kezdő fizetés: 7200 Ft.
A legnagyobb fizetés: 65500 Ft.
Összes kezdő fizetés: 28670625 Ft.
Valid után látható érték az érvényes esetek számát jelzi, vagyis azt, hogy hányan adták meg a kezdő fizetésük összegét.

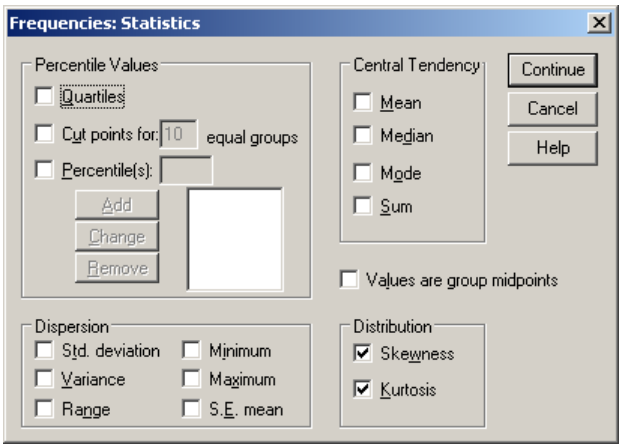
Descriptive Statistics

	N	Minimum	Maximum	Sum	Mean	Std. Deviation
Starting Salary	1100	7200	65500	28670625	26064,20	6967,982
Valid N (listwise)	1100					

4. táblázat

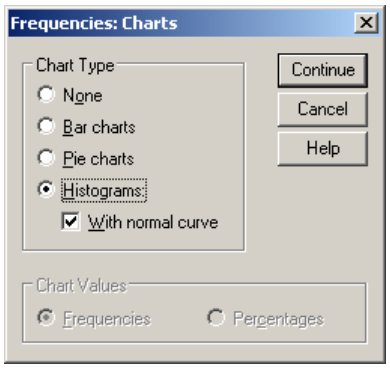
Amennyiben a táblázatokat szeretnénk átmásolni szövegszerkesztőbe, akkor kattintsunk a kívánt táblázatra, majd ez egér jobb gombjára, és válasszuk a Copy (másolás) parancsot, végül a szövegszerkesztőben a Szerkesztés/Beillesztés menüpontot. Így a táblázat könnyen formázható, az angol szavakat is átírhatjuk a magyar megfelelőikre.

Számoljuk ki a ferdeség és csúcsosság mutatóit, és ábrázoljuk hisztogram segítségével!
Az Analyze / Descriptive Statistics / Frequencies menüpontjában kattintsunk a Statistics gombra, és az előugró panelben a Discribution érték alatt található Skewness és Kurtosis értékeket pipáljuk ki, és nyomjuk meg a Continue gombot.

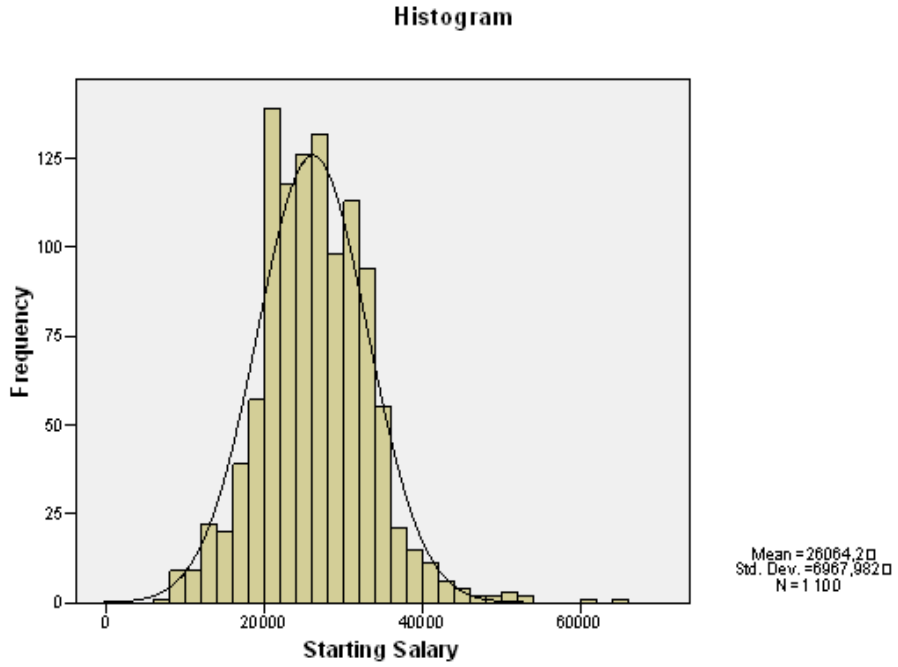


21. ábra

Ezután válasszuk a Chart gombot, majd jelöljük meg a Histograms és a With normal curve pontokat az alakzatok kirajzolásához.



22. ábra



23. ábra

Az ábrából kitűnik, hogy a kezdő fizetés alakzata szimmetrikus. Ezt jelzi az alábbi táblázat is.

Statistics

Starting Salary

N	Valid	1100
	Missing	0
Skewness		,488
Std. Error of Skewness		,074
Kurtosis		1,778
Std. Error of Kurtosis		,147

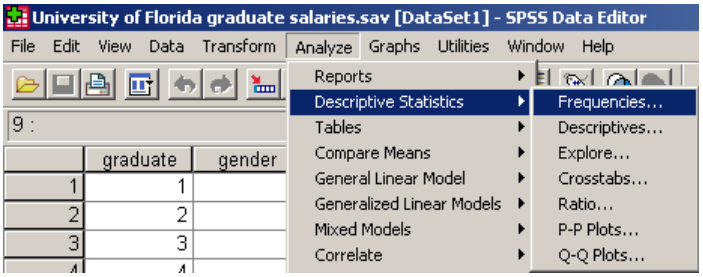
5. táblázat

A következő vizsgálat során az átlag, a módusz, a medián különbségére láthatunk példát, és ábrázoljuk őket.
A vizsgálatunk során az alábbi kérdésekre keressük a választ:

- Melyik főiskolai kart választották átlagosan?
- Melyik főiskolai karra jártak a legtöbben?
- Melyik az a főiskolai kar, amelyiket közepesen sokan választanak?

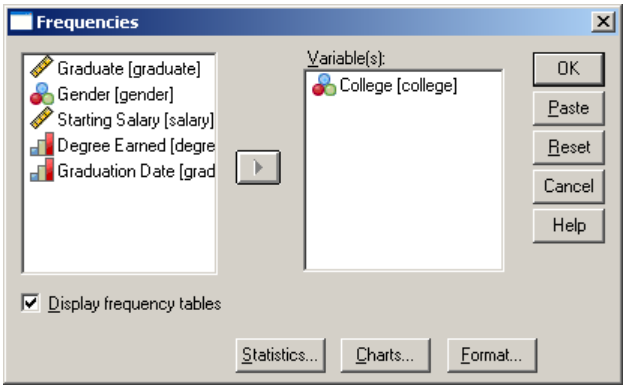
Emlékeztetőül, hogy milyen karok vannak a példában szereplő Floridai főiskolán: agriculture – mezőgazdasági, architecture – építőművészeti, building/construction – építészeti/épülettervezési, business administration – üzleti tanulmányok, forestry – erdészeti, education – pedagógiai, engineering – mérnöki, fine arts – képzőművészeti.

Az Analyze/Descriptive Statistics/Frequencies parancsot válasszuk a vizsgálat elvégzéséhez.



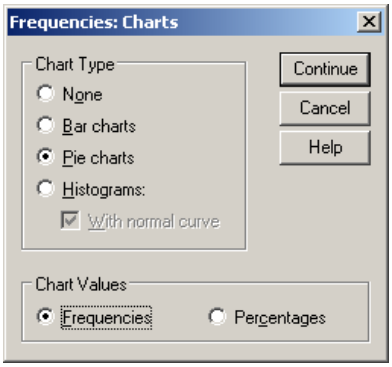
24. ábra

Vigyük át a Variable(s) alá a vizsgálni kívánt főiskolai kar (~College) változót, majd nyomjuk meg a Charts gombot.

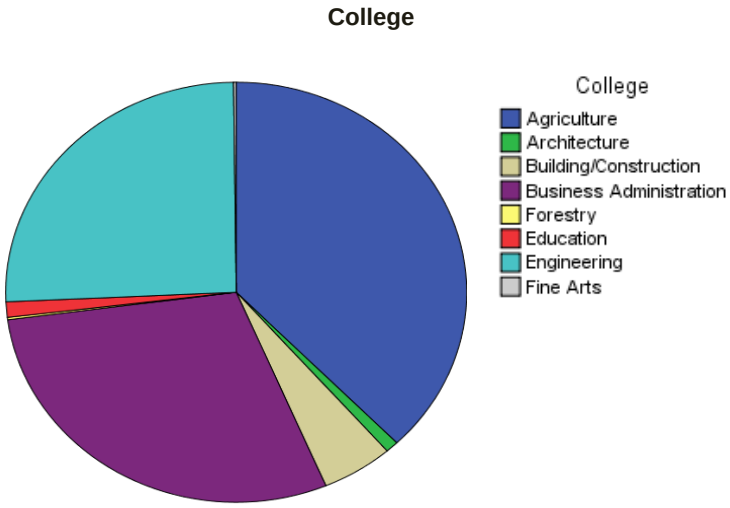


25. ábra

A kördiagramos ábrázoláshoz válasszuk a Chart Type alatti Pie charts-ot (26. ábra).

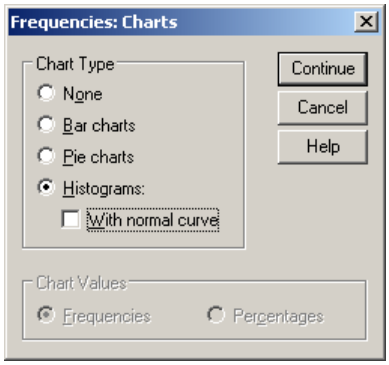


26. ábra

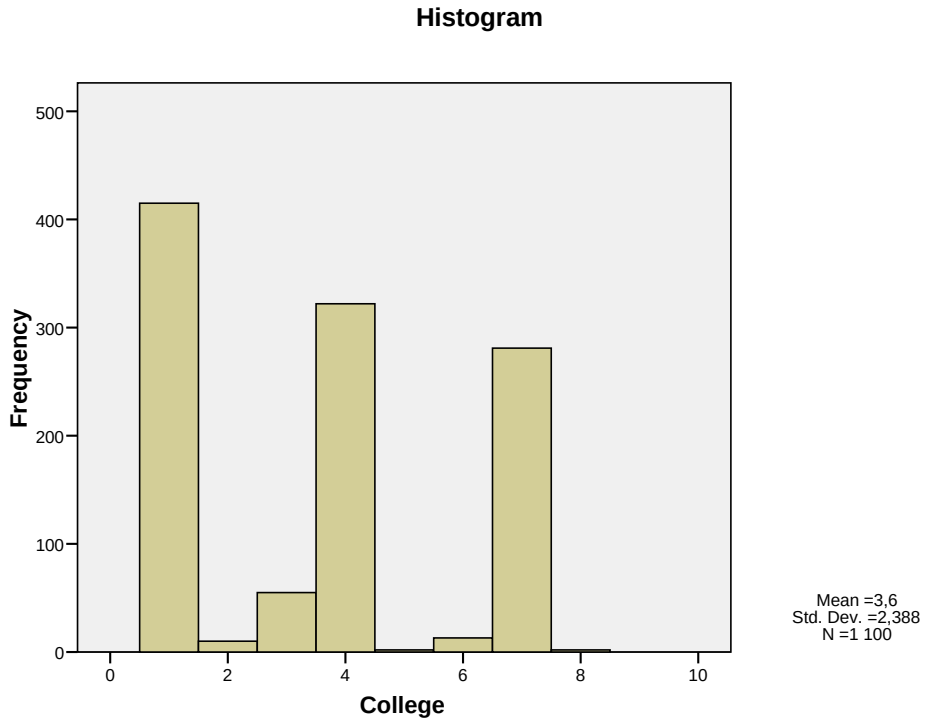


27. ábra

A Pie chart (tortadiagram) jól szemlélteti a 8 kar hallgatóinak a megoszlását (27.ábra).Mivel példánkban páros számú adat van, így a két középső értéket kell átlagolni. Ehhez a Pie Charts helyett a Histograms-ot kell választanunk (28. ábra).



28. ábra



29. ábra

A hisztogram (29. ábra) segítségével ábrázolt adatokról a legegyszerűbb leolvasni a medián értékét, hiszen csak meg kell keresni az oszlopok közül a középsőt, és az lesz a medián.

A diagramokkal együtt a gyakoriságokat tartalmazó táblázat is megjelenik az Output ablakban.

College

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Agriculture	415	37,7	37,7	37,7
	Architecture	10	,9	,9	38,6
	Building/Construction	55	5,0	5,0	43,6
	Business Administration	322	29,3	29,3	72,9
	Forestry	2	,2	,2	73,1
	Education	13	1,2	1,2	74,3
	Engineering	281	25,5	25,5	99,8
	Fine Arts	2	,2	,2	100,0
	Total	1100	100,0	100,0	

6. táblázat

A fenti táblázat (6.táblázat) a főiskolai karok különböző gyakorisági megoszlásait mutatja. Az abszolút gyakoriság (frequency) azt jelenti, hogy az adott kar hányszor szerepel a rangsorban. A legtöbb hallgató (415 fő) a mezőgazdasági főiskolai karra járt, majd ezt követi az üzleti tanulmányok kar (322 fő) és a mérnöki kar (281 fő). A vizsgálatban részt vett többi kar hallgatói már jóval kevesebben vannak. A Percent az adatok százalékos megoszlást jelenti.

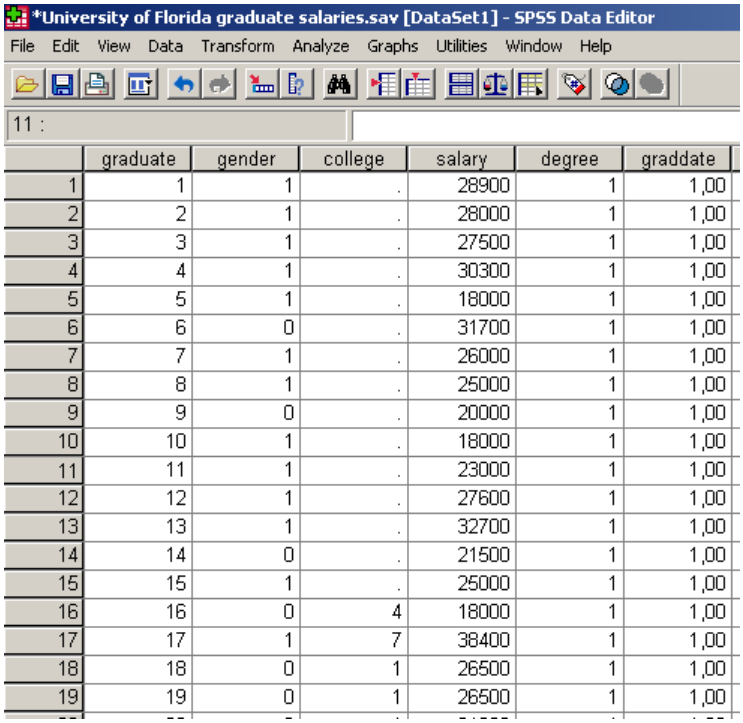
A relatív gyakoriság (Percent) az összelemszámhoz viszonyított gyakoriság (%-ban), azaz úgy kapjuk meg, hogy az abszolút gyakoriságot elosztjuk az elemszámmal és megszorozzuk százzal. Jelen esetben azt jelenti, hogy hány %-át teszik ki az egyes főiskolai karok hallgatói az összes kar hallgatóinak. Ez a szám például a mezőgazdasági főiskolai kar esetén 37,7 %.

A kumulatív relatív gyakoriság (Cumulative Percent) az adott sor és az azt megelőző sor – az első sor kivételével – relatív gyakoriságának összege százalékban kifejezve.

A Total, vagyis az összelemszám pedig a gyakoriságok összessége, azaz 1100 fő, ill. a relatív adatsoroknál nyilván 100% (kumulatív esetben nincs értelme).

Mivel a táblázatban a főiskolai karra vonatkozóan minden érték szerepel, így nem láthatunk különbséget az abszolút és a relatív gyakoriság között.

A két gyakoriság különbségének vizsgálatához töröljük ki a college oszlopból az első 15 értéket.



	graduate	gender	college	salary	degree	graddate
1	1	1	.	28900	1	1,00
2	2	1	.	28000	1	1,00
3	3	1	.	27500	1	1,00
4	4	1	.	30300	1	1,00
5	5	1	.	18000	1	1,00
6	6	0	.	31700	1	1,00
7	7	1	.	26000	1	1,00
8	8	1	.	25000	1	1,00
9	9	0	.	20000	1	1,00
10	10	1	.	18000	1	1,00
11	11	1	.	23000	1	1,00
12	12	1	.	27600	1	1,00
13	13	1	.	32700	1	1,00
14	14	0	.	21500	1	1,00
15	15	1	.	25000	1	1,00
16	16	0	4	18000	1	1,00
17	17	1	7	38400	1	1,00
18	18	0	1	26500	1	1,00
19	19	0	1	26500	1	1,00

30. ábra

Ezután ismét végezzük el a gyakorisági vizsgálatot (Analyze / Descriptive Statistics / Frequencies). Míg a mezőgazdasági főiskolai kar esetében az abszolút gyakoriság (Percent) 37,2%, addig a relatív gyakoriság (Valid Percent) 37,7%. A Valid érték jelzi, hogy hányan válaszolták meg a melyik kar hallgatója kérdést. A Missing érték, pedig azt jelzi, hogy van-e hiányzó érték, azaz létezik-e olyan személy, aki nem válaszolt a kérdésre (30.ábra). (Az előzőleg kitörölt 15 érték itt jelenik meg.)

A következő táblázatból (7. táblázat) kitűnik, hogy a relatív gyakoriság figyelmen kívül hagyja a hiányzó adatokat. Hiányzó értékek esetén tehát a relatív gyakoriság helyett az abszolút gyakoriságot használjuk.

College

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Agriculture	409	37,2	37,7	37,7
	Architecture	10	,9	,9	38,6
	Building/Construction	54	4,9	5,0	43,6
	Business Administration	320	29,1	29,5	73,1
	Forestry	2	,2	,2	73,3
	Education	13	1,2	1,2	74,5
	Engineering	275	25,0	25,3	99,8
	Fine Arts	2	,2	,2	100,0
	Total	1085	98,6	100,0	
Missing	System	15	1,4		
Total		1100	100,0		

7. táblázat

3.3. Feladatok

1. Sorolja fel a statisztikai helyzetmutatókat!
2. Sorolja fel a statisztikai szóródásmutatókat!
3. Nyissa meg a Cars.sav állományt és számítsa ki a helyzet- és szóródásmutatókat a horse (lóerő, teljesítmény) és a mpg (miles per gallon, fogyasztás) változókra!

4. Faktoranalízis

4.1. Alapfogalmak

Faktoranalízis: adattömörítésre és az adatstruktúra feltárására szolgál. A kiinduló változók számát úgynevezett faktorváltozókba vonja össze, amelyek közvetlenül nem figyelhetők meg.
A faktoranalízis struktúra-feltáró módszer, ami azt jelenti, hogy nincsenek előre meghatározott függő és független változók, hanem a változók közötti összefüggések feltárására törekszünk.
A faktoranalízis több, egymással korreláló változó összefüggését vizsgálja. Gyakran előfordul, hogy azok a változók, amelyeket mérni tudunk, nem a vizsgálni kívánt jelenséget legjobban jellemző változók. A módszer célja a közvetlenül nem megfigyelhető háttérváltozóknak, ún. faktoroknak a meghatározása a változók közti korrelációk alapján.
A faktoranalízis alkalmazásának akkor van létjogosultsága, ha az eredeti megfigyelési változók, vagy azok bizonyos csoportjai között erős összefüggés tapasztalható. Ezen felül az eredmények akkor lesznek gyakorlati szempontból jól értelmezhetők, ha a megfigyelési változók jól elkülöníthető csoportokba sorolhatók abból a szempontból, hogy az értékeket csoportonként közös háttérváltozók határozzák meg.
A faktoranalízis alkalmazása előtt meg kell vizsgálni, hogy az alábbi szükséges feltételek fennállnak-e:

- A faktoranalízisnek metrikus változókat kell feltételeznie, ugyanakkor a dummy változók (azaz 0 vagy 1 kimenettel rendelkező változók) használata is megengedett.
- A változók eloszlásával kapcsolatosan a normalitástól, homoszkedaszticitástól és a linearitástól való eltérés abból a szempontból fontos, hogy ezen feltételek megsértése csökkenti a változók közötti korrelációs együtthatók értékét.
- A változók közötti kapcsolat megléte, sőt a változók közötti multikollinearitás (ha nem tudjuk szétválasztani a független változók hatásait) kívánatos feltétel, ugyanis a változók közötti kapcsolat nélkül nem lehetne hasonló változókat találni és azokat egyetlen faktorba

tömöríteni. Elvárható, hogy minél több legyen a korrelált változó az adatbázisban és ezeknek a korrelációknak az értéke legyen 0,3-nál magasabb.

- Fontos a minta homogenitása, mert a faktoranalízis azt feltételezi, hogy a közös variancia az egész minta esetében fennáll.
- Minél nagyobb a mintanagyság, annál megbízhatóbb faktorokat eredményez az elemzés.

Meg kell keresni az eredeti változók azon csoportjait, amelyek egymással szorosabb korrelációban vannak, mint másokkal; ezeket a változókat tekintjük egy faktorhoz tartozónak. Ha sikerült ilyen csoportokat találnunk, a következő feladat a faktorok értelmezése, elnevezése. Legvégül a nagyszámú eredeti változót néhány faktorban összesíthetjük, és ezekkel, mint új változókkal dolgozhatunk tovább.

Az SPSS programban a faktoranalízis parancsot az ANALYZE/DATA REDUCTION/FACTOR menüpont alatt találhatjuk.

Az SPSS-ben több módszer is rendelkezésünkre áll annak kiderítésére, hogy adataink alkalmasak-e faktoranalízisre. Ezen módszerek közül néhány a faktoranalízis része, tehát az elemzés lefuttatása után derül ki, hogy az adatok/változók megfelelők-e valójában a faktoranalízisre.

4.1.1. A faktoranalízis megvalósíthatóságának feltételei

K**orrelációs mátrix**: az egyes változók közötti korrelációkat tükrözi, amelyek megléte alapvető feltétele a faktoranalízisnek, ugyanis nélküle nem lehetne a változókat faktorokba összevonni. Az erős korrelációk arra utalhatnak, hogy a változók alkalmasak a faktoranalízisre, hiszen az elemzésnek nem lenne sok alapja, ha a korrelációs mátrixban lévő értékek közel nullák lennének. Ugyanakkor a túlságosan magas korrelációk sem jók, ugyanis ez azt okozhatja, hogy a faktoranalízisnek nem lesz megoldása, ugyanis minden változó egy faktorba kerül. A „Descriptives" menüpontban állíthatjuk be a korrelációs mátrixot a korrelációs koefficiensek (Coefficients) és a szignifikanciaszint (Significance levels) bejelölésével.

Anti-image mátrix: az elemzés abból indul ki, hogy a változók szórásnégyzete felbontható magyarázott szórásnégyzetre (image) és nem magyarázott szórásnégyzetre (anti-image). A faktoranalízis során ezt a felbontást az anti-image kovariancia/korrelációs mátrixok mutatják. Az anti-image kovarianciamátrix átlón

kívüli elemei a variancia azon részét mutatják, amely független a többi változótól, ezért ezeknek az értékeknek lehetőség szerint alacsonynak kellene lenniük, míg az átlóban lévő elemek 1-hez közelítenek. Az anti-image korrelációs mátrixban elsődlegesen az átlóban lévő elemek fontosak, ugyanis ezek tartalmazzák az egyes változókra vonatkozó MSA-értékeket. Az MSA-értéke 0 és 1 között változhat, és azt mutatja meg, hogy az adott változó mennyire áll szoros kapcsolatban az összes többi változóval. Amennyiben egy változó MSA értéke 0,5 alatti, akkor ezt a változót valószínűleg ki kell zárni az elemzésből, míg ha 1 az értéke, akkor a változót a többi változó hiba nélkül becsli. Az anti-image mátrix parancs szintén a Descriptives menüponton belül található.

A **Bartlett-teszt** azt vizsgálja, hogy a változók az alapsokaságban korrelálatlanok-e vagyis hogy a korrelációs mátrixnak a főátlón kívüli elemei csak véletlenül térnek-e el a nullától. A faktoranalízis feltétele, hogy korreláljanak egymással a változók, lehetőleg minél erősebben.

A **Kaiser-Meyer-Olkin- (KMO) kritérium** az egyik legfontosabb mérőszám annak megítélésében, hogy a változók mennyire alkalmasak a faktoranalízisre. A KMO-érték az MSA értékek átlaga. Míg az MSA érték az egyes változókra vonatkozik, a KMO az összes változóra egyidejűleg. A KMO mutatószám jelentését a következőképpen ítélhetjük meg:

- KMO ≥ 0,9 kiváló
- KMO ≥ 0,8 nagyon jó
- KMO ≥ 0,7 megfelelő
- KMO ≥ 0,6 közepes
- KMO ≥ 0,5 gyenge
- KMO < 0,5 elfogadhatatlan.

4.1.2. A faktorok számának meghatározása

A faktorok számának meghatározására számos módszer áll rendelkezésre. Ilyen pl. az a priori kritérium, a Kaiser kritérium, a varianciahányad-módszer, a Scree-teszt (Könyökszabály).

A priori kritérium: a kutató a faktoranalízis megkezdése előtt dönt a faktorok számáról, ami maximum annyi lehet, amennyi kiinduló változó volt. Az SPSS-ben a faktorok számát a Factor Analysis menüpontban az Extraction parancs segítségével érhetjük el, ahol a „Number of factors”-t kell bejelölni és megadni a faktorok számát.

Kaiser kritérium: a sajátértéket használja, csak azokat a faktorokat vegyük figyelembe, amelyek sajátértéke legalább 1. A sajátérték a faktorok által az összes változó varianciájából magyarázott variancia. Ha egy faktor sajátértéke 1 alá csökken, akkor már kevesebb információt hordoz, mint egy változó, azaz azt a faktort már nem érdemes használni. A sajátérték az Extraction menüpontban az „Eigenvalues over" bejelölésével jelenik meg.

Varianciahányad-módszer: A faktorok számát meghatározhatjuk a variancia összesített (kumulált) százaléka alapján is, azaz annyi faktort hozunk létre, hogy elérjünk egy minimális összesített varianciaszintet, amelyre számos hüvelykujjszabály létezik. A természettudományokban az elfogadott varianciahányad minimálisan 95 százalék, míg a társadalomtudományi kutatásokban már 60 százalék is elfogadható. A varianciahányad-módszer a gyakorlati szignifikancián alapul, azaz ha magas varianciahányadot tudunk magyarázni, az azt jelenti, hogy az információ jelentős részét meg tudtuk tartani az elemzés során. A faktorok által magyarázott varianciát az SPSS alapesetben megadja.

A táblázatot három hármas egységre lehet osztani, az első a kezdeti értéket (Initial Eigenvalues), a második a faktoranalízis utáni értékeket (Extraction Sums of Squared Loadings), a harmadik pedig a rotáció utáni értékeket (Rotation Sums of Squared Loadings) tartalmazza. A „Total" oszlop a sajátértéket mutatja, a „% of Variance" az adott faktor által magyarázott varianciahányadot a teljes variancián belül, míg a „Cumulative %" oszlop az – adott faktoring - összesített varianciahányadot mutatja.

Az „Initial Eigenvalues" oszlopok a faktorok információtartalmát mutatják be standardizált formában, azaz itt annyi sort (komponenst) láthatunk, mint amennyi kiinduló változónk volt, és a sajátértékek összege megegyezik a komponensek számával. A táblázat „kezdeti" és a „faktoranalízis utáni" oszlopai majdnem teljesen megegyeznek egymással, ugyanakkor az utóbbi már csak az általunk kért, 1-nél nagyobb sajátértékű faktorokat tartalmazza. Az elemzés a faktorokat a magyarázott variancia nagyságának sorrendjében mutatja.

Scree-teszt (Könyökszabály): szintén segítséget nyújt a faktordimenziók számának meghatározásában. A Scree plot ábra valójában nem más, mint a sajátértékek ábrázolása a faktorok sorrendjében, ahol az *y* tengelyen mérjük a sajátértékeket, az *x* tengelyen pedig a faktorok számát. Habár egyedi variancia minden faktorban van, ugyanakkor ennek szintje az első faktornál nagyon

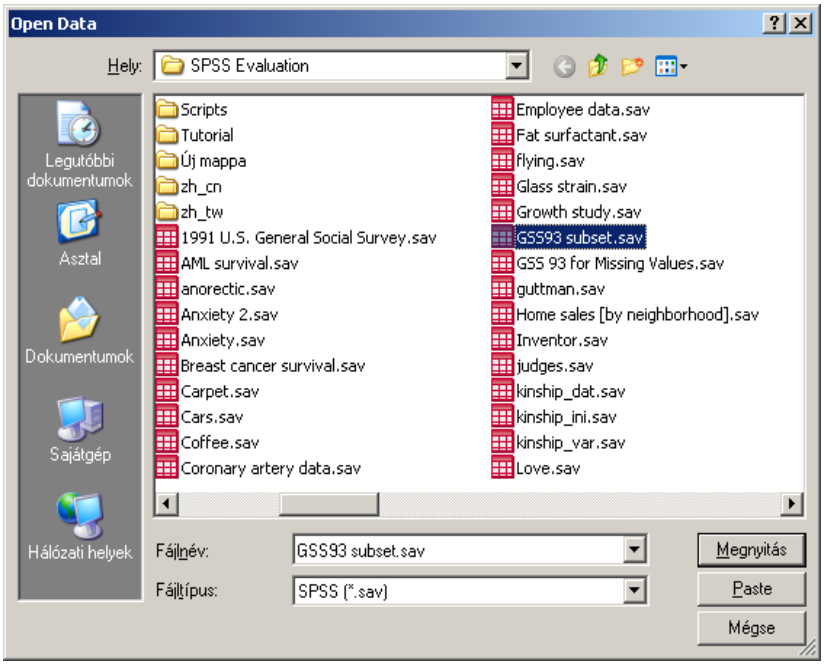
alacsony, és a közös variancia dominál, míg az utolsó faktornál ez fordított. A könyökszabály azt mondja ki, hogy a faktorok számát annyiban érdemes maximalizálni, ahol a görbe meredeksége hirtelen megváltozik és egyenesbe kezd átfordulni. A könyökszabály alapján tehát olyan faktorok is fontosak lehetnek, amelyek sajátértéke 1 alatt van.

4.1.3. Faktorok rotálása

A faktorkiválasztás (extrakció) során az elemzés elsődleges célja, hogy maximalizálja a főkomponensek varianciáját, amely eredményeként megkapjuk a rotálatlan faktorsúly-mátrixot. A faktorsúly az eredeti változó és az adott faktor közötti korrelációt mutatja, amelynek értéke a korrelációs együtthatókhoz hasonlóan -1 és 1 között változhat. A faktorkiválasztás során azonban előfordulhat, hogy olyan változók fognak korrelálni egy adott faktorral, amelyeknek semmi közük egymáshoz, ezáltal lehetetlenné téve az értelmezést. Ezen a problémán segít a forgatás, vagy más néven rotáció. A faktor-rotáció azt jelenti, hogy a faktorok tengelyeit elforgatjuk úgy, hogy egyszerűbb és értelmezhetőbb faktormegoldáshoz vezessen. A rotálási eljárást a faktoranalízis panelen belül a Rotation parancs alatt jelölhetjük meg. Itt kell kiválasztani a módszert és a Display keretben a „Rotated solution”-t.
A rotáció (forgatás) során nem változnak sem a kommunalitás, sem pedig az összes magyarázott variancia, csak a faktorok sajátértékei/magyarázott varianciái módosulnak. A rotáláson belül két típust különböztetünk meg: a derékszögű (ortogonális) (Varimax, Equimax, Quartimax) és a hegyesszögű (nem ortogonális) (Direct Oblimin, Promax) forgatási módszereket. A derékszögű esetében a tengelyek merőlegesen állnak egymásra, ezáltal a faktorok nem korrelálnak egymással, míg a hegyesszögű esetében ezek tetszőleges szöget zárnak be egymással, vagyis a faktorok korrelálni fognak egymással.

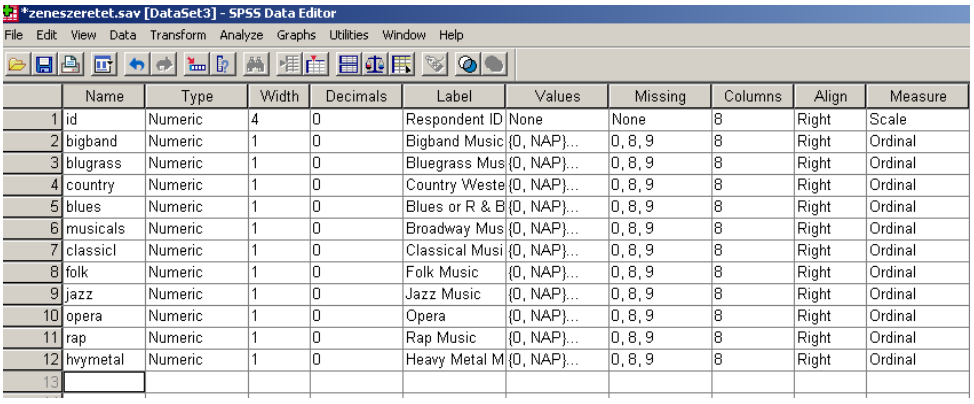
4.2. Példa a faktoranalízisre

Nyissuk meg az SPSS példaállományai közül a GSS93 subset.sav (31. ábra által jelzett) adatállományt.



31. ábra

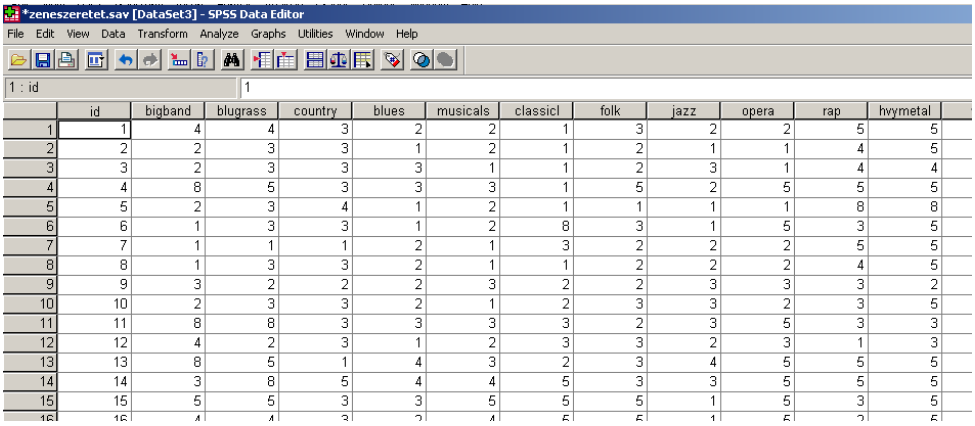
Megnyitás után váltsunk Variable View nézetre és az alábbi változók (id, bigband, bluegrass, country, blues, musicals, classical, folk, jazz, opera, rap, heavymetal) kivételével töröljük a többit, és mentsük el zeneszeretet.sav néven (32.ábra).



	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure
1	id	Numeric	4	0	Respondent ID	None	None	8	Right	Scale
2	bigband	Numeric	1	0	Bigband Music	{0, NAP}...	0, 8, 9	8	Right	Ordinal
3	blugrass	Numeric	1	0	Bluegrass Mus	{0, NAP}...	0, 8, 9	8	Right	Ordinal
4	country	Numeric	1	0	Country Weste	{0, NAP}...	0, 8, 9	8	Right	Ordinal
5	blues	Numeric	1	0	Blues or R & B	{0, NAP}...	0, 8, 9	8	Right	Ordinal
6	musicals	Numeric	1	0	Broadway Mus	{0, NAP}...	0, 8, 9	8	Right	Ordinal
7	classicl	Numeric	1	0	Classical Musi	{0, NAP}...	0, 8, 9	8	Right	Ordinal
8	folk	Numeric	1	0	Folk Music	{0, NAP}...	0, 8, 9	8	Right	Ordinal
9	jazz	Numeric	1	0	Jazz Music	{0, NAP}...	0, 8, 9	8	Right	Ordinal
10	opera	Numeric	1	0	Opera	{0, NAP}...	0, 8, 9	8	Right	Ordinal
11	rap	Numeric	1	0	Rap Music	{0, NAP}...	0, 8, 9	8	Right	Ordinal
12	hvymetal	Numeric	1	0	Heavy Metal M	{0, NAP}...	0, 8, 9	8	Right	Ordinal
13										

32. ábra

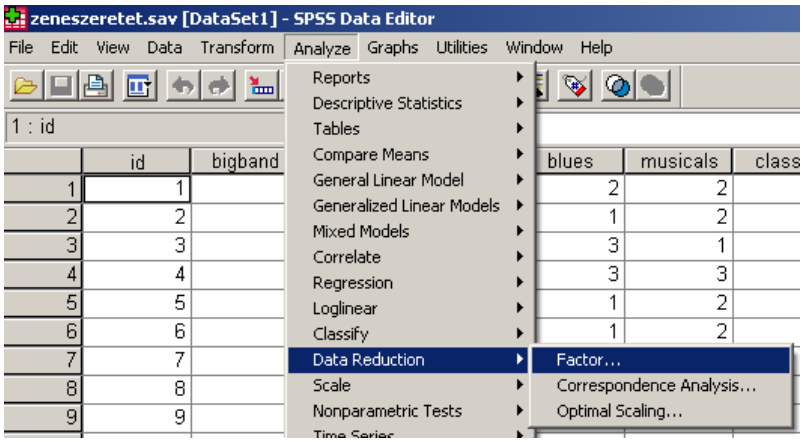
Az adatbázisban a különböző zenefajtákat megítélését láthatjuk (33.ábra) egy 5 fokozatú skálán (1=nagyon szeretem, 2=szeretem, 3=közömbös, 4=nem szeretem, 5=nagyon nem szeretem). A 0, 8, 9 értékek a hiányzó értékeket jelölik, ami azt jelenti, hogy nem válaszoltak a kérdésre, vagy nem volt megfelelő a válasz.



	id	bigband	blugrass	country	blues	musicals	classicl	folk	jazz	opera	rap	hvymetal
1	1	4	4	3	2	2	1	3	2	2	5	5
2	2	2	3	3	1	2	1	2	1	1	4	5
3	3	2	3	3	3	1	1	2	3	1	4	4
4	4	8	5	3	3	3	1	5	2	5	5	5
5	5	2	3	4	1	2	1	1	1	1	8	8
6	6	1	3	3	1	2	8	3	1	5	3	5
7	7	1	1	1	2	1	3	2	2	2	5	5
8	8	1	3	3	2	1	1	2	2	2	4	5
9	9	3	2	2	2	3	2	2	3	3	3	2
10	10	2	3	3	2	1	2	3	3	2	3	5
11	11	8	8	3	3	3	3	2	3	5	3	3
12	12	4	2	3	1	2	3	3	2	3	1	3
13	13	8	5	1	4	3	2	3	4	5	5	5
14	14	3	8	5	4	4	5	3	3	5	5	5
15	15	5	5	3	3	5	5	5	1	5	3	5

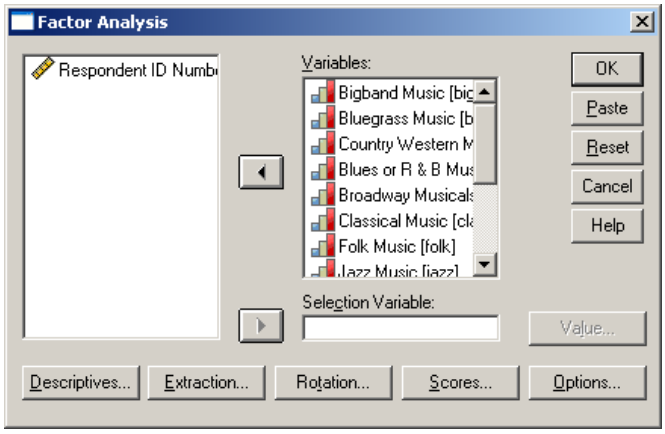
33. ábra

A faktoranalízishez válasszuk ki az Analyze / Data Reduction / Factor menüpontot (34.ábra).



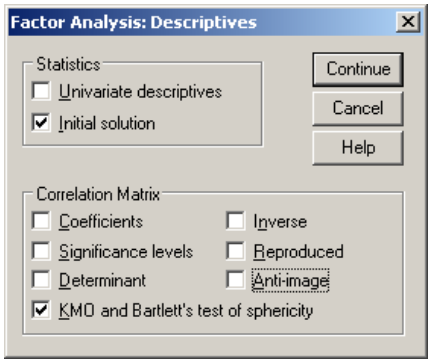
34. ábra

Vegyük át a nyíl segítségével a tizenegy zenetípust a Variables mező alá (35. ábra), majd a Descriptives gombra kattintsunk (35. ábra).



35. ábra

Vizsgáljuk meg, hogy az általunk kiválasztott változók alkalmasak-e a faktoranalízisre. Ehhez a KMO (Kaiser-Meyer-Olkin) értékét kell megvizsgálnunk. A Bartlett teszthez és a KMO megállapításához a Factor Analysis: Descriptives ablaknál a KMO and Bartlett's test of sphericity mezőt pipáljuk ki, majd kattintsunk a Continue gombra, hogy megtudhassuk a KMO nagyságát (36. ábra).



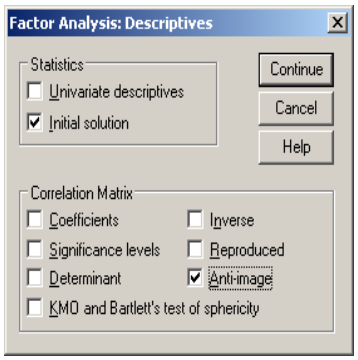
36. ábra

KMO and Bartlett's Test

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		,748
Bartlett's Test of Sphericity	Approx. Chi-Square	3048,818
	df	55
	Sig.	,000

8. táblázat

A KMO értékét mutatja a 8. táblázat. Jelen esetben: 0,748 a KMO, ami alapján a változók megfelelőnek tekinthetők (KMO>6), tehát alkalmasak a faktoranalízisre.
Miután eldöntöttük, hogy a változóink alkalmasak a faktoranalízisre, vizsgáljuk meg, hogy hány faktort kell képeznünk.
A Descriptives gombnál pipáljuk ki az Anti-image mezőt (37. ábra), majd a Continue gombot, és végül az OK-t, ami után megjelenik az Anti-image matrix (Anti-images Matrices), melyet a 9. táblázatban láthatunk.



37. ábra

Anti-image Matrices

		Bigband Music	Bluegrass Music	Country Western Music	Blues or R & B Music	Broadway Musicals	Classical Music	Folk Music	Jazz Music	Opera	Rap Music	Heavy Metal Music
Anti-image Covariance	Bigband Music	,608	-,082	-,051	-,057	-,217	-,030	-,005	-,064	-,069	,073	,025
	Bluegrass Music	-,082	,683	-,264	-,092	,035	,006	-,190	,012	-,002	,016	-,035
	Country Western Music	-,051	-,264	,765	-,010	,010	,108	-,077	,069	-,025	-,024	,062
	Blues or R & B Music	-,057	-,092	-,010	,640	-,001	-,001	-,014	-,313	-,015	-,061	-,027
	Broadway Musicals	-,217	,035	,010	-,001	,552	-,118	-,085	-,034	-,071	-,042	,061
	Classical Music	-,030	,006	,108	-,001	-,118	,500	-,143	-,049	-,230	,032	-,025
	Folk Music	-,005	-,190	-,077	-,014	-,085	-,143	,855	,032	-,033	,046	-,003
	Jazz Music	-,064	,012	,069	-,313	-,034	-,049	,032	,629	-,013	-,063	-,024
	Opera	-,069	-,002	-,025	-,015	-,071	-,230	-,033	-,013	,578	-,077	,019
	Rap Music	,073	,016	-,024	-,061	-,042	,032	,046	-,063	-,077	,829	-,281
	Heavy Metal Music	,025	-,035	,062	-,027	,061	-,025	-,003	-,024	,019	-,281	,854
Anti-image Correlation	Bigband Music	,825[a]	-,127	-,075	-,092	-,374	-,055	-,007	-,103	-,116	,103	,035
	Bluegrass Music	-,127	,672[a]	-,366	-,139	,057	,011	-,284	,019	-,003	,022	-,046
	Country Western Music	-,075	-,366	,569[a]	-,014	,015	,175	-,109	,100	-,037	-,030	,076
	Blues or R & B Music	-,092	-,139	-,014	,702[a]	-,001	,002	-,021	-,493	-,025	-,083	-,037
	Broadway Musicals	-,374	,057	,015	-,001	,816[a]	-,224	-,159	-,058	-,125	-,062	,088
	Classical Music	-,055	,011	,175	-,002	-,224	,770[a]	-,251	-,087	-,428	,050	-,038
	Folk Music	-,007	-,284	-,109	-,021	-,159	-,251	,808[a]	,049	-,054	,063	-,004
	Jazz Music	-,103	,019	,100	-,493	-,058	-,087	,049	,703[a]	-,022	-,087	-,033
	Opera	-,116	-,003	-,037	-,025	-,125	-,428	-,054	-,022	,814[a]	-,111	,028
	Rap Music	,103	,022	-,030	-,083	-,062	,050	,063	-,087	-,111	,550[a]	-,334
	Heavy Metal Music	,035	-,046	,076	-,037	,088	-,038	-,004	-,033	,028	-,334	,552[a]

a. Measures of Sampling Adequacy(MSA)

9. táblázat

Communalities

	Initial	Extraction
Bigband Music	1,000	,550
Bluegrass Music	1,000	,708
Country Western Music	1,000	,691
Blues or R & B Music	1,000	,771
Broadway Musicals	1,000	,629
Classical Music	1,000	,725
Folk Music	1,000	,581
Jazz Music	1,000	,769
Opera	1,000	,635
Rap Music	1,000	,650
Heavy Metal Music	1,000	,678

Extraction Method: Principal Component Analysis.

10. táblázat

A hosszú kommunalitási (Communalities) tábla (10. táblázat) a bemenő és kijövő kommunalitás értékeket mutatja a faktorokra, ami kezdetben ez az érték 1.
A táblázat alján a kiválasztott módszert láthatjuk, mely jelen esetben a főkomponens módszer (Principal Component Analysis). Ennek a módszernek az a lényege, hogy azokat a faktorokat választjuk ki, melyek a legtöbb varianciát magyarázzák meg.

Total Variance Explained

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	3,276	29,779	29,779	3,276	29,779	29,779
2	1,661	15,098	44,876	1,661	15,098	44,876
3	1,392	12,651	57,527	1,392	12,651	57,527
4	1,058	9,620	67,147	1,058	9,620	67,147
5	,728	6,619	73,766			
6	,658	5,978	79,744			
7	,566	5,150	84,893			
8	,496	4,510	89,404			
9	,421	3,823	93,227			
10	,397	3,608	96,835			
11	,348	3,165	100,000			

Extraction Method: Principal Component Analysis.

11. táblázat

A táblázat (11. táblázat) első oszlopa tartalmazza a kiinduló változóknak a számát, a második főoszlop mutatja a sajátértékeket és a varianciákat ennek a módszernek az alkalmazása után, a harmadik oszlopban a kiválasztott faktorokra jeleníti meg ugyanezeket.

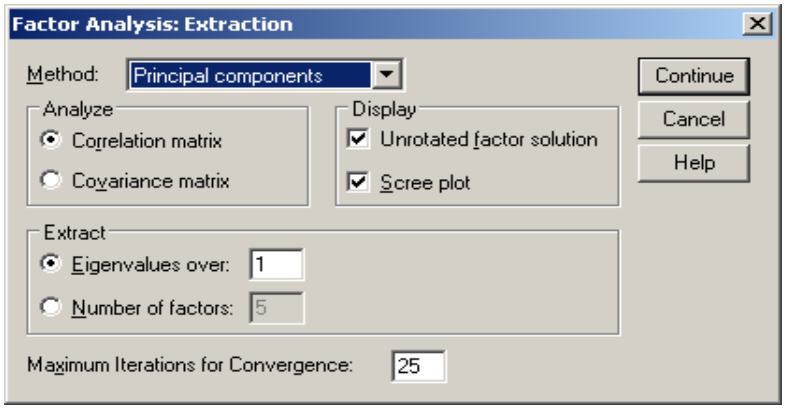
Láthatjuk, hogy 4 faktort különített el az Anti-image eljárás. Ezt a Total Variance Explained (11. táblázat jobb oldala) és a Component Matrix (12 t.áblázat) is mutatja. A négy faktor együtt a teljes variancia 67,147%-át magyarázza (ezt az utolsó oszlopnak az utolsó sorában láthatjuk), ami eléri a minimumként megfogalmazott 60%-ot.**Component Matrix(a)**

	Component			
	1	2	3	4
Bigband Music	,713	-,124	-,079	-,141
Bluegrass Music	,426	-,430	,584	,017
Country Western Music	,144	-,557	,600	,022
Blues or R & B Music	,531	,362	,333	-,497
Broadway Musicals	,743	-,033	-,266	,073
Classical Music	,741	,073	-,334	,243
Folk Music	,625	-,341	,091	,257
Jazz Music	,526	,491	,131	-,485
Opera	,712	,061	-,228	,267
Rap Music	,087	,592	,388	,376
Heavy Metal Music	-,018	,549	,404	,462

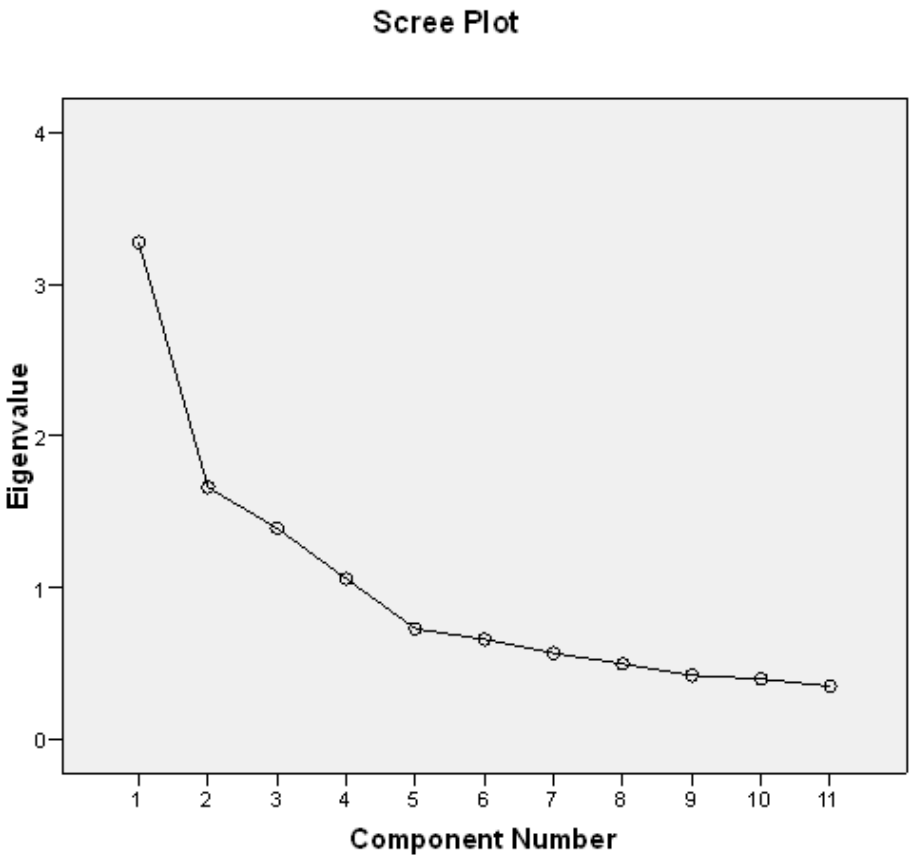
12 t.áblázat

A komponens mátrixból leolvashatjuk, hogy melyik változó melyik faktort jellemzi leginkább. A táblázatban szereplő értékek a faktorsúlyok.

Amennyiben nem megfelelő a négyes faktorszám, akkor a döntésben a Scree plot a segítségünkre lehet.Scree plot ábra az Extraction menüpontban a „Scree plot" bejelölésével kérhető (38.ábra).



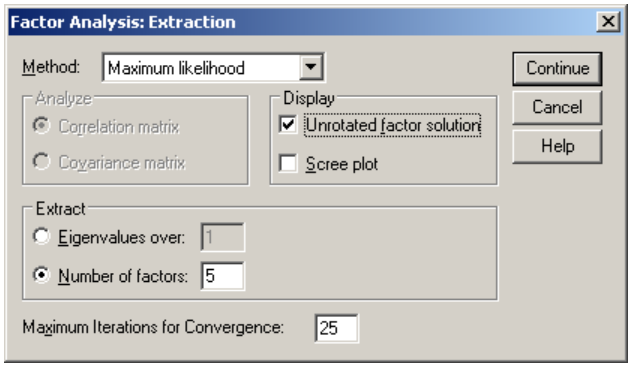
38. ábra



39. ábra

Az 39. ábra azt mutatja, hogy a 11 faktor meredeksége az elsőtől az utolsó felé haladva csökken. Az ábrán könyökpontokat (elbow-kritérium) kell keresni, olyan helyet, ahol törés van a görbén, mert azokon a helyeken romlik el a magyarázott varianciának a növekedése.

Az ábrán az 5 faktorszámnál találunk könyökpontot. Az 5 faktorszámnál lévő törés tehát megerősíti azt, hogy 4 faktoros megoldást kell választanunk.



40. ábra

Vizsgáljuk meg a Maximum likelihood eljárással, hogy mi történne, ha 5 faktort használnánk. Methodnál válasszuk a Maximum likelihood-ot, majd állítsuk a faktorok számát (Number of factors) 5-re (40. ábra).

Goodness-of-fit Test

Chi-Square	df	Sig.
16,466	10	,087

13. táblázat

A Maximum likelihood eljárás az 5 faktorra 0,087-es alacsony (<=0,1) szignifikanciaszintet adott eredményül (13. táblázat). Ez a 4 faktorral szemben még így is magasabb szignifikanciaszintet mutat. Tehát ez a módszer nem hozott megfelelő eredményt.

Rotated Component Matrix(a)

	Component				
	1	2	3	4	5
Broadway Musicals	,816	,139	,063	-,065	,006
Opera	,770	,070	-,026	,116	,174
Classical Music	,746	,115	-,187	,038	,384
Bigband Music	,678	,282	,261	-,162	-,054
Blues or R & B Music	,129	,863	,110	,084	,081
Jazz Music	,220	,843	-,097	,082	-,011
Country Western Music	2,99E-005	-,079	,885	-,029	-,016
Bluegrass Music	,073	,179	,697	,006	,466
Rap Music	,136	,095	,091	,823	-,333
Heavy Metal Music	-,154	,083	-,119	,796	,261
Folk Music	,425	,026	,261	-,036	,691

Extraction Method: Principal Component Analysis.

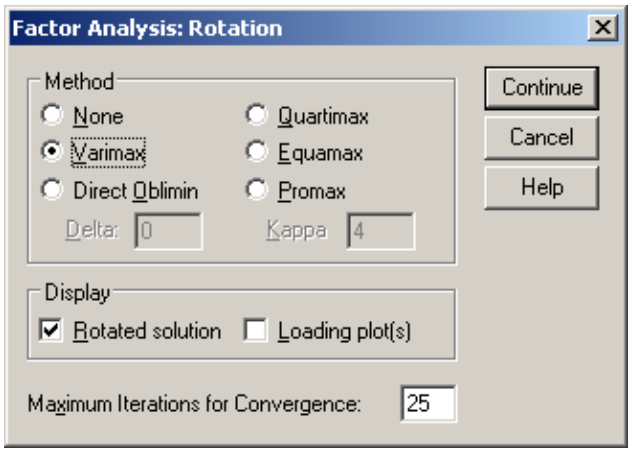
Rotation Method: Varimax with Kaiser Normalization.

Rotation converged in 8 iterations.

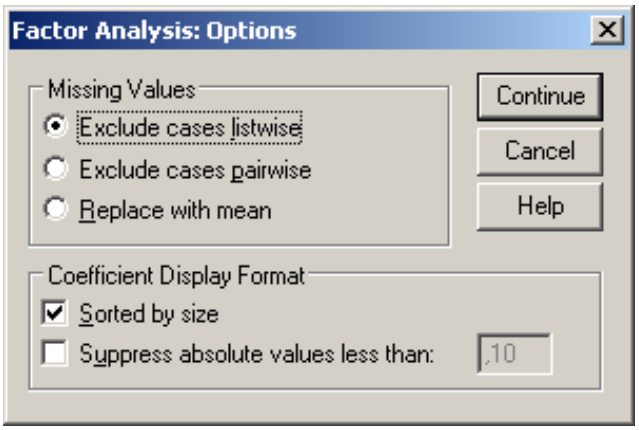
14. táblázat

Amennyiben az 5 faktoros megoldást választjuk, akkor a Folk zene egyedül külön faktorba kerülne (14. táblázat).

Maradjunk a 4 faktoros megoldásnál és rotációs eljárással alakítsuk át a mátrixunkat. Válasszuk a Varimax Methodot és a Continue gombot (41. ábra), majd az Options gombnál a Sorted by size mezőt pipáljuk ki a rendezés érdekében (42. ábra).



41. ábra



42. ábra

Rotated Component Matrix(a)

	Component			
	1	2	3	4
Classical Music	,841	,097	-,072	,046
Opera	,785	,090	,006	,103
Broadway Musicals	,764	,190	,033	-,091
Folk Music	,604	-,040	,463	-,012
Bigband Music	,597	,340	,206	-,189
Blues or R & B Music	,133	,850	,143	,105
Jazz Music	,204	,843	-,086	,099
Country Western Music	-,074	-,045	,825	-,058
Bluegrass Music	,164	,137	,813	,018
Heavy Metal Music	-,044	,018	-,012	,822
Rap Music	,020	,142	-,027	,793

Extraction Method: Principal Component Analysis.

Rotation Method: Varimax with Kaiser Normalization.

Rotation converged in 5 iterations.

15. táblázat

A 4 faktorra a Varimax rotálást alkalmazva a faktorok sokkal könnyebben értelmezhetőek (15. táblázat). Az oszlopokban az abszolút értékben 0,5-nél nagyobb számokat kell keresni (jelen esetben téglalappal jelöltük ezeket az értékeket). A kapott faktorokat a értelemszerű nevezzük el.
Az első faktorba tartozik: a Classical Music, Opera, Broadway Musicals, Folk Music és a Bigband Music.
A második faktort a Blues or R&B Music és a Jazz Music képzi.
A harmadik faktor a Country Western Music és a Bluegrass Music.
Az utolsó, negyedik faktor pedig a Heavy Metal Music és a Rap Music.

4.3. Feladatok

1. Ismertesse a faktoranalízis lényegét!
2. Ismertesse a faktoranalízis alkalmazhatóságának feltételeit!
3. Ismertesse a faktoranalízis menetét!

5. Korreláció

5.1. Alapfogalmak

A **korreláció**:

- A két változó közötti egyenes arányú, fordított arányú vagy hiányzó kapcsolat (pozitív, negatív vagy nem létező korreláció) lehet. Becslése az értékek ábrázolása alapján lehetséges.
- A korrelációs koefficiens legalacsonyabb értéke (abszolút értelemben): 0 (nincs lineáris korreláció), a legmagasabb +1,0 vagy -1,0 (tökéletes pozitív, ill. negatív lineáris korreláció)
- A korrelációs koefficiens értéke független a mértékegységektől.
- A kiugró értékek erősen befolyásolhatják a korrelációs együttható értékét. A kiugró érték lehet egy szabálytalan, torzult eloszlás eredménye, vagy lehet mérési hiba. A szóródási ábrán megjelenő, kiugró értékek esetén vizsgálatra van szükség. Célszerű ezeket kiküszöbölni. Ebben az esetben használható a Spearman féle rang-korreláció.
- Gyakran elkövetik azt a hibát, hogy a két változó közötti korrelációból ok-okozati összefüggésre következtetnek. Ha *x* és *y* között erős korreláció van, akkor ennek oka lehet:
    - az *y* változásai okozzák az *x* változásait
    - a *x* változásai okozzák az *y* változásait
    - egy harmadik faktor mind az *x*-et, mind az *y*-t egy irányba (vagy – negatív korreláció esetén – ellenkező irányba) befolyásolja.

A kapcsolat szorosságát, a függőség fokát mérnünk kell (16. táblázat). Ennek mérésére a ***korrelációs együttható*** a szokásos mérőszám, amelynek sok tulajdonsága hasonló a szóráséhoz. A korrelációs együttható egy statisztikai mutató, azaz egy minta korreláltsága leírására szolgál, miközben a populáció változói közötti kapcsolat erősségét a korrelációs együttható mint paraméter határozza meg.

r értéke	Kapcsolat iránya (előjele) és erőssége
$r=1$	Tökéletes pozitív kapcsolat (függvényszerű lineáris kapcsolat)
$0,7 \leq r < 1$	Erős pozitív kapcsolat
$0,2 \leq r < 0,7$	Közepes pozitív kapcsolat
$0 < r < 0,2$	Gyenge pozitív kapcsolat
$r=0$	Nincs lineáris kapcsolat
$-0,2 < r < 0$	Gyenge negatív kapcsolat
$-0,7 < r \leq -0,2$	Közepes negatív kapcsolat
$-1 < r \leq -0,7$	Erős negatív kapcsolat
$r=-1$	Tökéletes negatív kapcsolat (függvényszerű lineáris kapcsolat)

16. táblázat

Az összetartozó értékpárok halmazának mindegyik tagját egyenként átlagolhatjuk, és az egyes értékeknek a saját átlaguktól való eltérését vizsgálhatjuk. Az *x*, vagy az *y* szórásának számításakor ezen különbségek négyzeteit átlagoltuk (majd négyzetgyököt vontunk belőle), a korrelációs együttható számításakor az összetartozó különbségeket összeszorozzuk, és a szorzatok összegét (ezt más néven ***kovarianciának*** is nevezik) elosztjuk a négyzetes különbségek szorzatával. A korrelációs együttható két fontos tulajdonsága:

1. A korrelációs együttható értéke független változók esetében 0.
2. Lineáris függvénykapcsolatban lévő (nem sztochasztikus) változók esetében a korrelációs együttható abszolút értéke 1.

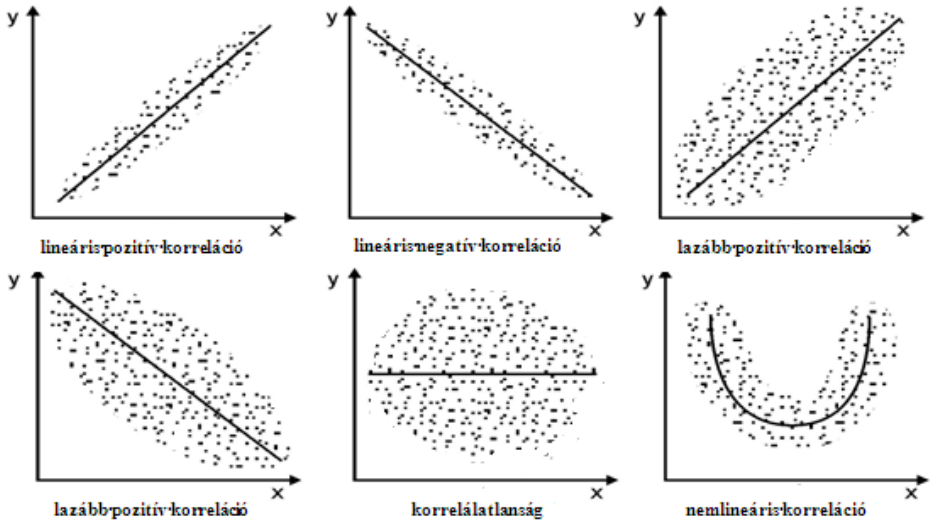
Minél szorosabb az összefüggés két változó között, annál jobban közelíti a korrelációs együttható abszolút értéke az 1-et; minél lazább összefüggés van két változó között, annál közelebb áll a korrelációs együttható értéke a 0-hoz (16. táblázat).

Fontos, hogy *a korrelációs együttható az egyszerű, közel lineáris sztochasztikus kapcsolat esetében használható statisztika*.

Ha két változó korrelációjának vizsgálata során az együttható értéke 0, akkor még nem biztos, hogy ezek függetlenek is! Ezért ilyenkor csak annyit mondhatunk: a két változó ***korrelálatlan***.

A két valószínűségi változó korrelációját egy elméleti korrelációs együttható írja le. Ennek értékét a gyakorlatban becsléssel közelítjük meg. A becsléshez a szokásos módszer szerint a populációból mintát veszünk, majd a minta korrelációs együtthatóját kiszámoljuk, és meghatározzuk a becslés hibáját. A

becslés hibájának ismeretében megmondhatjuk, hogy mekkora annak a valószínűsége, hogy a mintából számolt korrelációs együttható nem 0.
A korrelációs együttható előjele megmutatja, hogy az összefüggést jellemző egyenes emelkedő, vagy süllyedő. Görbére illeszkedő vagy annak mentén elhelyezkedő pontok ábrája jelzi, hogy a korreláció nem alkalmas az összefüggés jellemzésére, azaz nemlineáris a korreláció (43. ábra).



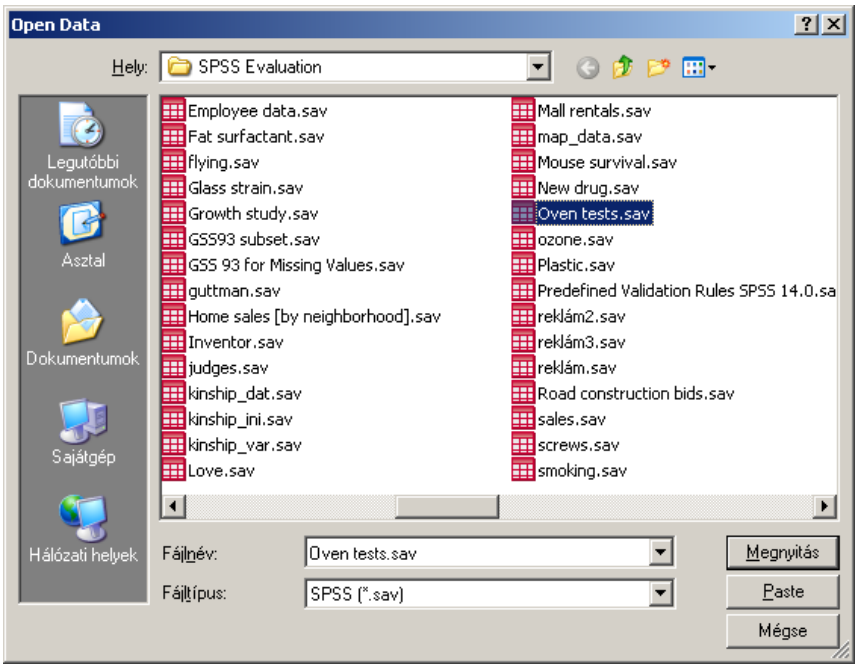
43. ábra

Gondolnunk kell arra is, hogy ha a korrelációs együttható szignifikáns, az még nem jelenti azt, hogy a változók között kapcsolat erős, vagy azt, hogy a kapcsolat jelentős lenne.
A korreláció mögött lehet ok-okozati viszony, de az is lehet, hogy a két korrelált változó nincs egymással ok-okozati kapcsolatban, hanem mind a kettő egy harmadik, közös októl függ. A korreláció magyarázata lehet a véletlen is, például, mind a két változó az idővel korrelált, és a közös tényezővel korrelált változók között gyakran van korreláció is. A tanulság, hogy az ok-okozati összefüggést logikai, vagy kísérleti úton kell bizonyítani.
Több változó esetében hasonló kapcsolat állapítható meg az ún. parciális korrelációs együttható segítségével. Ez azt mutatja meg, hogy mekkora lenne az *x*

és *y* közötti lineáris korreláció, ha egy vagy több másik változót állandó szinten tartanánk.

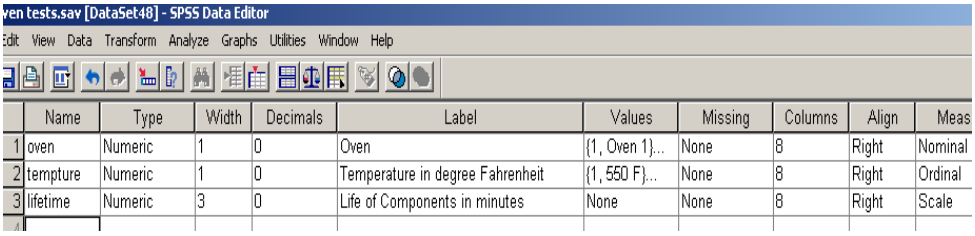
5.2. *Példa a korreláció kiszámítására*

Nézzünk egy példát a korrelációszámításra. Vásárolni szeretnénk egy sütőt, de nem tudjuk, hogy a három fajta sütő közül melyiket válasszuk. Kíváncsiak vagyunk arra, hogy van-e összefüggés a sütők fajtája és az élettartama között, valamint a sütő élettartama és a sütéskor használt hőmérséklet között.
A vizsgálathoz nyissuk meg (File/Open/Data) az Oven tests.sav nevű fájlt (44. ábra).



44. ábra

Töröljük ki Variable View nézetben az utolsó, számunkra nem lényeges sort (45. ábra), majd váltsunk Data View nézetbe (46. ábra).



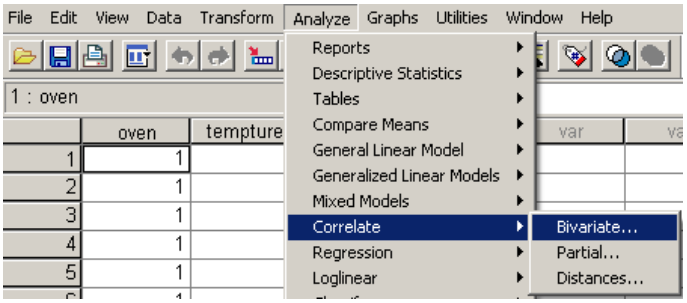
	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Meas
1	oven	Numeric	1	0	Oven	{1, Oven 1}...	None	8	Right	Nominal
2	tempture	Numeric	1	0	Temperature in degree Fahrenheit	{1, 550 F}...	None	8	Right	Ordinal
3	lifetime	Numeric	3	0	Life of Components in minutes	None	None	8	Right	Scale

45. ábra

	oven	tempture	lifetime
1	1	1	237
2	1	1	254
3	1	1	246
4	1	2	178
5	1	2	179
6	1	2	183
7	2	1	208
8	2	1	178
9	2	1	187
10	2	2	146
11	2	2	145
12	2	2	141
13	3	1	192
14	3	1	186
15	3	1	183
16	3	2	142
17	3	2	125
18	3	2	136
19			

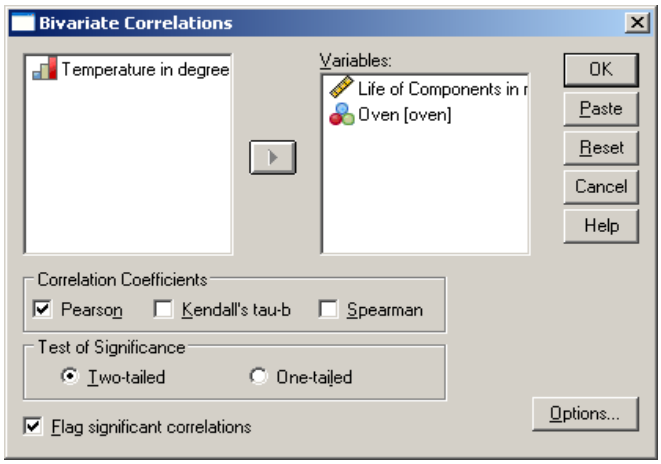
46. ábra

Végezzük el az összefüggésvizsgálatot a sütő fajtáit és az alkatrészek élettartamát figyelembe véve. Ehhez válasszuk az Analyze / Correlate / Bivariate parancsot (47. ábra).



47. ábra

A megjelenő Bivariate Correlation panelben (48. ábra) mozgassuk át a Variables alá az alkatrészek élettartamát (Life of components in minutes) és sütőket(Oven). A korrelációs kofficiensnél válasszuk a Pearson korrelációt, a kétoldali Two-tailed próbát, majd pipáljuk ki a Flag significant correlations-t, azaz jelezze csillaggal, ha szignifikáns a korreláció.



48. ábra

A megjelenő korrelációs táblázat (17. táblázat) viszonylag erős (-0,654) negatív korrelációt és szignifikanciát (csillagok jelzik) mutat a sütők és az alkatrészei élettartama között. A negatív előjel azt jelzi, hogy ez az összefüggés ellentétes. Minél kisebb a sütő fajtájának a száma, annál nagyobb a sütő alkatrészeinek élettartama. A Correlation is significant at 0.01 level azt jelenti, hogy a korreláció elfogadható legalább 1%-os szignifikanciaszint mellett.

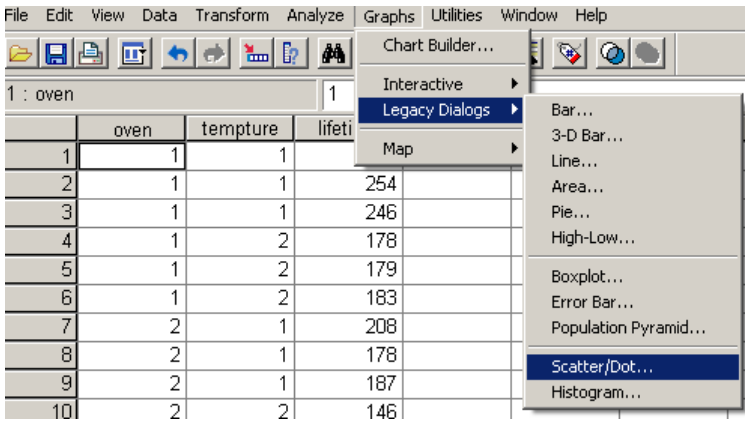
Correlations

		Life of Components in minutes	Oven
Life of Components in minutes	Pearson Correlation	1	-,654(**)
	Sig. (2-tailed)		,006
	N	16	16
Oven	Pearson Correlation	-,654(**)	1
	Sig. (2-tailed)	,006	
	N	16	16

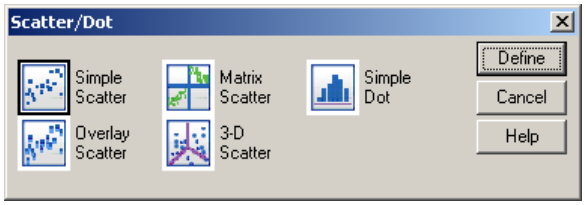
** Correlation is significant at the 0.01 level (2-tailed).

17. táblázat

Az adatokat jelenítsük meg pontfelhő diagram segítségével is, hogy szemléletesebbé tegyük a korrelációt. Válasszuk a Graphs/Legacy Dialogs/Scatter/Dot parancsát (49. ábra).



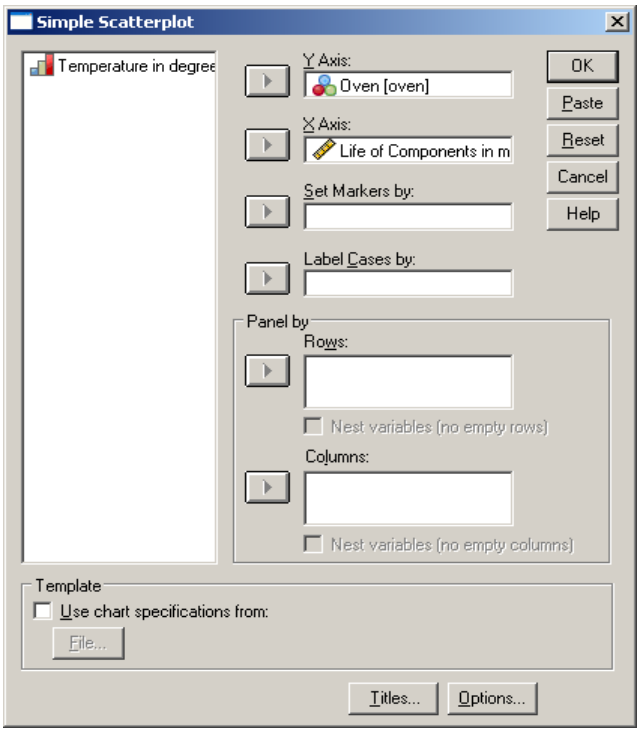
49. ábra



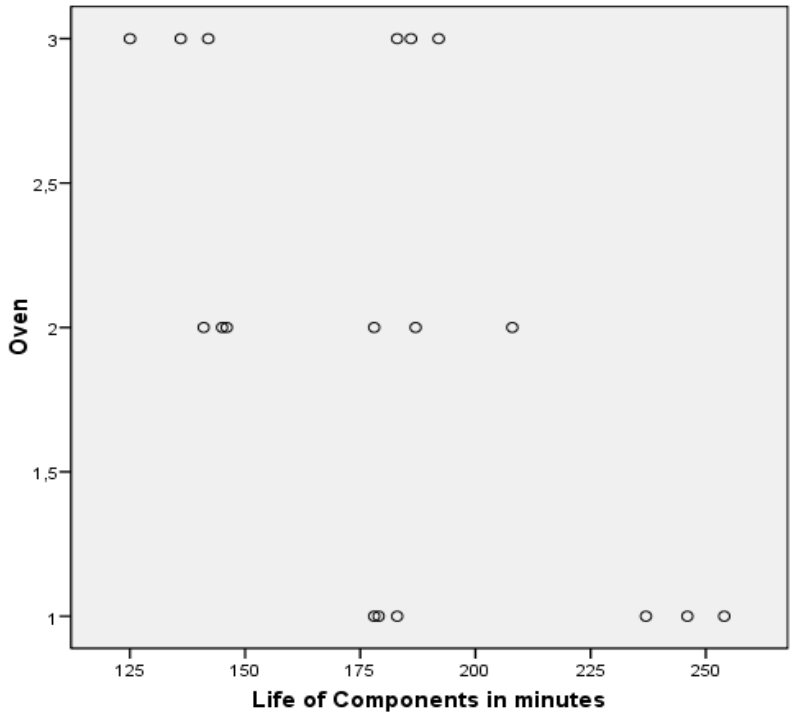
50. ábra

A Simple Scatter menüpontot válasszuk, majd nyomjuk meg a Define gombot (50. ábra).

A Simple Scatterplot ablakban (51. ábra) az X és Y (Axis) változók alá vigyük át a nyilak segítségével a sütőket (Oven) és az alkatrészek élettartamát (Life of Components in minutes), majd az OK gombra kattintva megjelenik ezek alapján a pontfelhő diagram.



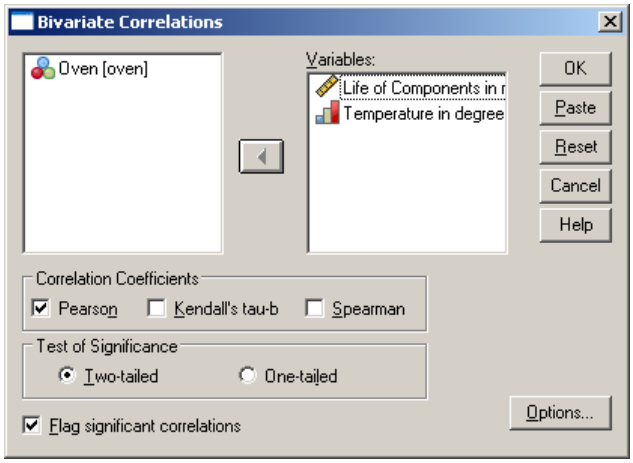
51. ábra



52. ábra

A megjelenő ábra a szórásdiagram (52. ábra) szemlélteti a korreláció elvégzésekor kapott eredményt**.** A szórásdiagram két, vagy több változó együttes elemzéséhez, a közöttük lévő összefüggések feltáráshoz nyújthat segítséget. A pontfelhő alakjából és elhelyezkedéséből következtethetünk az adott változók közötti összefüggésre.

Most vizsgáljuk meg, hogy a sütő hőfoka és az alkatrészek élettartama között van-e összefüggés. Az előbbiek során bemutatott menetet kövessük: Analyze/Correlate/Bivariate. Ebben az esetben a Life of Components in minutes és a Temperature in degree Fahrenheit változókat válasszuk ki (53. ábra).



53. ábra

A korrelációs táblázat (18. táblázat) azt mutatja, hogy a két érték között szignifikáns kapcsolat van. Ez a kapcsolat is negatív korrelációt mutat, tehát minél magasabb a sütő hőmérséklete, annál kisebb az alkatrészek élettartama.

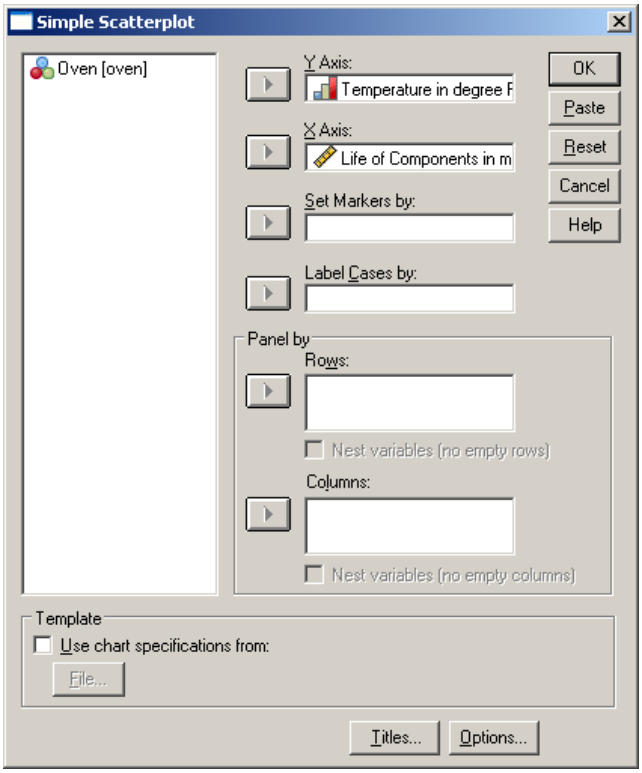
Correlations

		Life of Components in minutes	Temperature in degree Fahrenheit
Life of Components in minutes	Pearson Correlation	1	-,782(**)
	Sig. (2-tailed)		,000
	N	16	16
Temperature in degree Fahrenheit	Pearson Correlation	-,782(**)	1
	Sig. (2-tailed)	,000	
	N	16	16

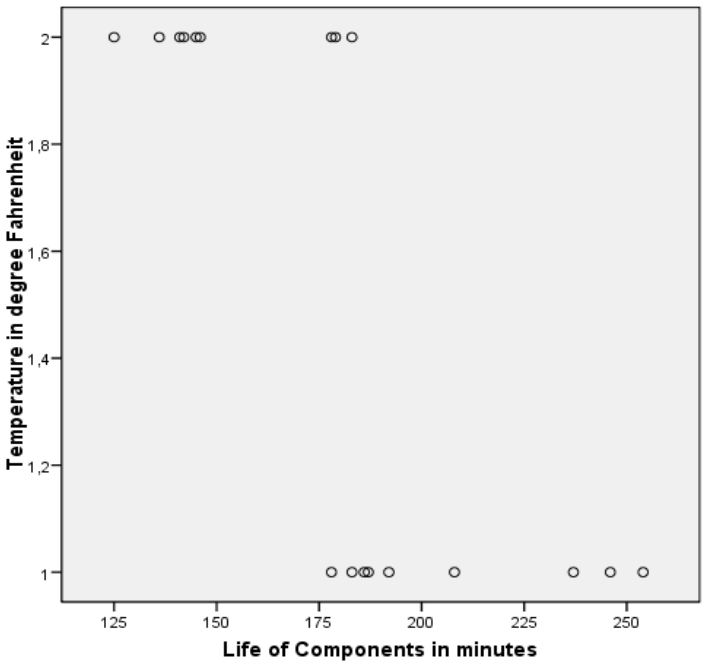
** Correlation is significant at the 0.01 level (2-tailed).

18. táblázat

Nézzük meg ezeknek a változóknak is a szórásdiagramját az előzőhöz hasonlóan. Graphs / Legacy Dialogs / Scatter / Dot után válasszuk a Simple Scatter ikont (54. ábra).


Simple Scatterplot
Oven [oven]
Y Axis:
Temperature in degree F
X Axis:
Life of Components in m
Set Markers by:
Label Cases by:
Panel by
Rows:
Nest variables (no empty rows)
Columns:
Nest variables (no empty columns)
OK
Paste
Reset
Cancel
Help
Template
Use chart specifications from:
File...
Titles...
Options...

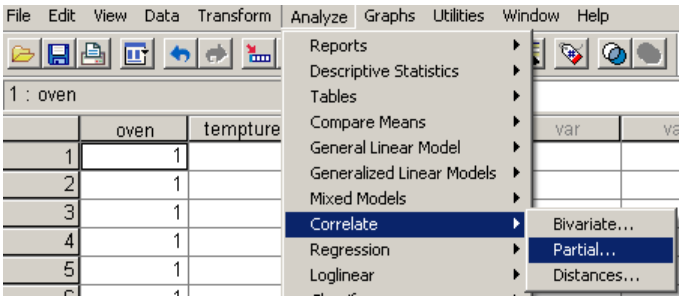
54. ábra



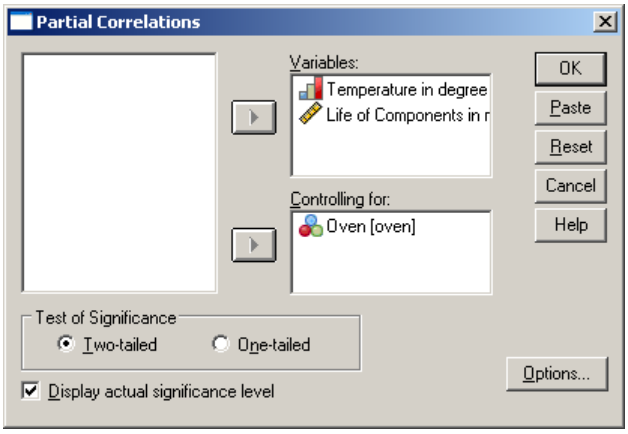
55. ábra

A szórásdiagram mutatja, hogy alacsonyabb hőmérsékleten az alkatrészeinek az élettartama magasabb (55. ábra).

A következőekben parciális korreláció segítségével vizsgáljuk meg a három változó közötti összefüggést. Ehhez válasszuk az Analyze/Correlate/Partial parancsot (56. ábra).

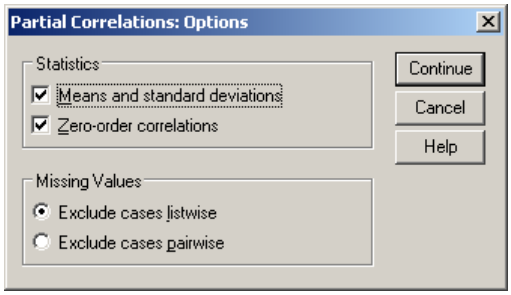


56. ábra



57. ábra

Válasszuk ki a Life of Components in minutes és a Temperature in degree Fahrenheit változókat, majd kattintsunk az Options gombra (57. ábra).



58. ábra

Az Options ablakban jelöljük be következőket: Means and standard deviations és Zero-order correlations (58. ábra).

A parciális korreláció segítségével megvizsgálhatjuk, hogy valóban szignifikáns-e a változók közötti összefüggés. Ennek segítségével megtudhatjuk, hogy a két változó közötti kapcsolat valódi összefüggés-e vagy egy harmadik változó hatásának tulajdonítható, ami mindkettővel összefüggést mutat. A parciális korreláció táblázatát is az előzőleg már említett feltételeknek megfelelően elemezzük (19. táblázat).

Correlations

Control Variables			Temperature in degree Fahrenheit	Life of Components in minutes	Oven
-none- [a]	Temperature in degree Fahrenheit	Correlation	1,000	-,748	,000
		Significance (2-tailed)	.	,000	1,000
		df	0	16	16
	Life of Components in minutes	Correlation	-,748	1,000	-,578
		Significance (2-tailed)	,000	.	,012
		df	16	0	16
	Oven	Correlation	,000	-,578	1,000
		Significance (2-tailed)	1,000	,012	.
		df	16	16	0
Oven	Temperature in degree Fahrenheit	Correlation	1,000	-,917	
		Significance (2-tailed)	.	,000	
		df	0	15	
	Life of Components in minutes	Correlation	-,917	1,000	
		Significance (2-tailed)	,000	.	
		df	15	0	

a. Cells contain zero-order (Pearson) correlations.

19. táblázat

5.3. Feladatok

1. Ismertesse a korreláció és a korrelálatlanság fogalmát! Térjen ki a korrelációs együttható jelentésére!
2. Milyen félrevezető tényezőkre kell figyelni a számított korreláció értelmezésénél?

6. Regresszió

6.1. Alapfogalmak

A *regresszió vizsgálat* célja két vagy több változó függvénykapcsolatának meghatározása, az összetartozó adatokból álló, tapasztalati adatsor analitikus közelítése előre megadott típusú matematikai összefüggéssel úgy, hogy a számított és a mért értékek eltérése minimális legyen. Az eltérések mértékét többféleképpen lehet megadni. Leggyakrabban a négyzetes hibák összegét szokták választani. A vizsgált jelenség természete szabja meg a közelítésre alkalmas függvény típusát. Eszerint megkülönböztetünk

- lineáris és
- nemlineáris

regressziót. A kapcsolt változók száma szerint ugyancsak eltérnek a modellek. Ilyen értelemben beszélünk két-, három- stb. változós regresszióról.

6.1.1. Lineáris regresszió

Az egyváltozós lineáris regresszió két – egy x független és egy y függő – folytonos változó összefüggésének jellemzése regressziós egyenessel.
A determinációs együttható (a korrelációs együttható négyzete), r^2 azt mutatja meg, hogy az x-től való függés mennyiben magyarázza meg az y variabilitását.
Ha r^2

- közelít a 0-hoz, akkor az x nem magyarázza az y-t, ha
- közelít 1-hez, akkor nagyon szoros az összefüggés.

Ha a két változó között van szignifikáns összefüggés, de az r^2 kicsi, az azt jelenti, hogy más tényezők is szerepet játszanak az y meghatározásában.
A legegyszerűbb regressziós kapcsolat két változó között a grafikusan egy egyenes vonallal jellemezhető lineáris függvénykapcsolat. Első kérdésünk az, hogy a két változó között van-e egy egyenessel leírható összefüggés? Ha igen, akkor megkeressük a legjobb ilyen egyenest. Az ennél bonyolultabb, nemlineáris függvénykapcsolatok, vagy a kettőnél több változó függvénykapcsolatának vizsgálata a statisztika haladó témái közé tartoznak.
A regressziós kapcsolatban mind a két változó függhet a véletlentől is, de az is előfordulhat, hogy csak az egyik esetében lényeges a véletlentől függő komponens. A továbbiakban mi a két esetet nem különböztetjük meg.

A regresszióban a két változó szerepe nem felcserélhető. A lineáris regresszió $y=ax+b$ képletében az egyik változó az x, a másik az y helyére kerül, és az x változó segítségével jósoljuk meg az y értékét. Itt elsősorban logikailag fontos, hogy a két változó szerepe nem felcserélhető (emlékezzünk arra, hogy a korreláció esetében a két változó közül egyik sem volt kitüntetett, azaz felcserélhetők voltak).
Gyakran az x változó esetében nem tételezzük fel, hogy a véletlen változás az x-t is közvetlenül érinti, hanem az x-t általunk választható rögzített és ismert értékként kezeljük, és a véletlentől való függés az y értékében jelenik meg. Az y tehát függ az x-től, de ezen kívül függhet a véletlen okozta ingadozástól is.
Hogyan határozzuk meg, hogy a pontok közé húzható rengeteg egyenes közül melyik az, amelyik az adatok összefüggését legjobban jellemzi? A grafikus ábrázolás pontdiagramja sejteti a lineáris összefüggést. Vonalzóval, „szemre" azonban általában lehetetlen megtalálni az egyenes és a pontok legjobb illeszkedését.

6.1.2. A legkisebb négyzetek módszere

A legjobb illeszkedést kiszámolhatjuk a legkisebb négyzetek módszerével. Nem hibázunk jelentősen, ha azt mondjuk, hogy a pontok és az egyenes távolságát minimalizálja a legkisebb négyzetek módszere. A valóságban a legkisebb négyzetek módszere azt az egyenest keresi meg, amelyre igaz az, hogy ha a pontoknak az egyenestől mért távolságait négyzetre emeljük, majd a kapott számokat összegezzük, akkor ez az összeg minimális lesz (nincs olyan másik egyenes, ami esetében kisebb ilyen összeget kapnánk). Ez legtöbbször nem azonos a távolságok összegével, sem annak négyzetével (mert általában nem mindegy, hogy előbb emelünk-e négyzetre és utána összegzünk, vagy pedig fordítva, előbb összegzünk és utána emelünk négyzetre), de igen hasonló tulajdonságú statisztika.

6.1.3. Az illesztés és a becslés jósága

Az angolszász szakirodalom a regresszió esetében használja még a determinációs koefficiens fogalmát is, amely az y értékek esetében a lineáris függvénynek tulajdonítható változásokat (szóródást) viszonyítja az összes szóródáshoz. Ha minden szóródást a lineáris komponens magyaráz, és nincs véletlennek tulajdonítható komponens, akkor ez a hányados 1. Ez a koefficiens könnyen

bizonyíthatóan azonos a korrelációs együttható négyzetével. A pontok szóródásának minél nagyobb részét tudjuk megmagyarázni a lineáris regresszióval, annál nagyobb ez az érték, annál közelebb áll 1-hez ez a hányados, és akkor annál nagyobb a korrelációs együttható abszolút értéke is. A regressziónak ez a tulajdonsága jól mutatja a korreláció és a lineáris regresszió fogalmainak rokonságát.

A korrelációhoz hasonlóan a két változó kapcsolata a regresszió esetében is többféle lehet. Ha a két változó között nincs kapcsolat, akkor a regressziós együttható értéke 0. Ha van kapcsolat, akkor a regressziós együttható értéke 0-tól eltérő.

A regressziós egyenes képletében mind a konstans tag, mind pedig az x változó együtthatója a véletlentől is függő mennyiség. Ismételt mintavétel esetében (a kísérlet ismétlésekor) várható, hogy egyik érték sem lesz pontosan ugyanaz, mint korábban volt, hanem szóródást fognak mutatni. Kivétel, hogy a regresszió esetében a független változó (x) esetében megengedhető, hogy az ne legyen valószínűségi változó, értékét a vizsgáló határozza meg, lehetőséget adva ezzel a jóslásra.

Fontos kérdés, hogy a regressziós együttható értéke eltér-e a 0-tól, másképpen fogalmazva van-e statisztikai értelemben vett összefüggés a két változó között, és milyen valószínűséggel helyes az ebben a kérdésben hozott döntésünk.

Ha a regressziós egyenest az egyik változó értékének ismeretében a másik becslésére kívánjuk használni, akkor tudnunk kell, hogy a becslés jósága függ a változók kapcsolatának erősségétől, azaz a korreláció szorosságától. Minél szorosabb a kapcsolat a két változó között, annál jobb az x alapján az y értékének a becslése.

6.1.4. Hipotézisvizsgálat

A hipotézisvizsgálathoz feltesszük, hogy a minta független, véletlenszerű mintavétellel vett elemekből áll, továbbá minden x értékre az y érték normális eloszlású valószínűségi változó.

A lineáris függés egyenletében mind a konstans tag, mind pedig a meredekség esetében a standard hibával képzett hányadosa a t-eloszlást követi, $n-2$ szabadságfokkal. Ennek alapján lehet véleményt kialakítani arról, hogy a számított értékeknek a nullától való eltérését vajon a véletlen okozta-e? A

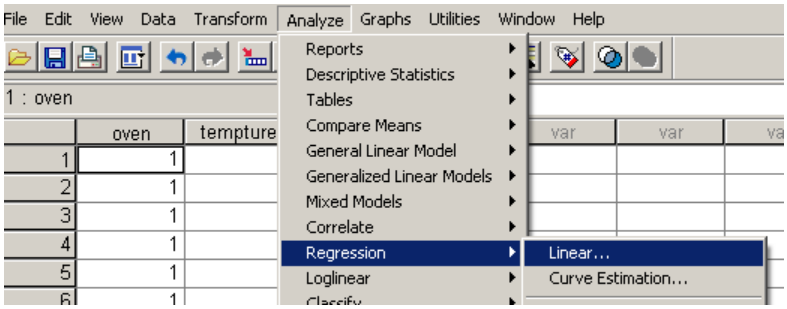
szignifikáns (0-tól eltérő) regressziós együttható (meredekség) azt jelzi, hogy a két változó kapcsolatát az adott valószínűség mellett nem a véletlen hozta létre.

6.1.5. Reziduálisok vizsgálata

Az egyes pontok és a regressziós egyenes közötti függőleges távolságokat reziduálisoknak is nevezik, és ezek képviselik az eljárásban elkülönített véletlentől függő komponenst. Ezek részletes vizsgálata fontos kiegészítése a változók kapcsolatának regresszióval történő vizsgálatának. Az SPSS tartalmaz eljárásokat a regresszió kiszámítása után a reziduális értékek táblázatokba foglalására, azok grafikus vizsgálatára. A reziduálisok ábrázolása jól mutathatja, ha a szóródás függ a független változó értékétől, ha az összefüggés eltér a lineáristól, ha az *x* tengely mentén egymás mellett lévő adatok nem függetlenek egymástól.
Minél kisebb az ábrán a vertikális szóródás, annál szorosabb a korreláció, és annál jobb az *y* érték becslése.
A görbe körüli szóródás adataiból az SPSS segítségével meghatározhatjuk a regressziós egyenes együtthatóinak standard hibáját. A standard hiba segítségével konfidencia intervallumok képezhetők, és az is vizsgálható, hogy független mintákból számított két regressziós egyenes paraméterei között van-e különbség.

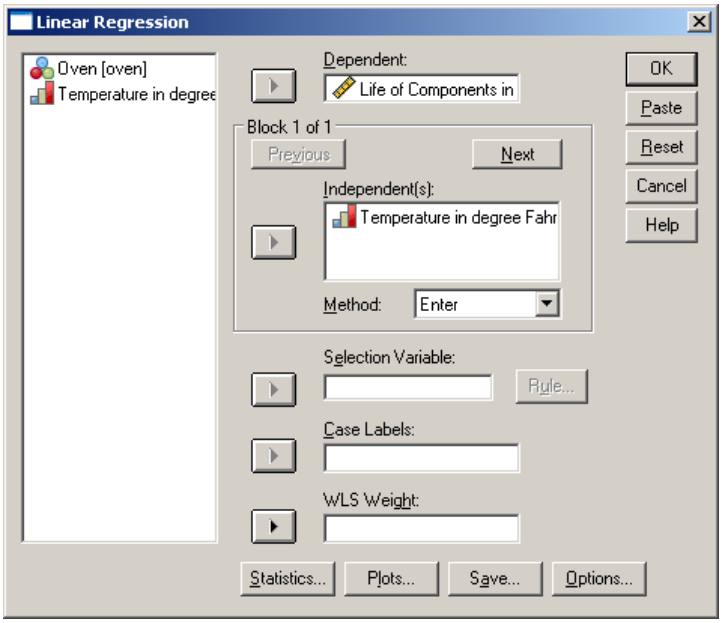
6.2. Példa regressziószámításra

A regresszióanalízis két változó közötti összefüggés leírását a korrelációs együtthatóhoz képest sokkal pontosabban határozza meg. Ennek szemléltetéséhez nyissuk meg az előző fejezetben megismert Oven.sav állományt, majd válasszuk az Analyze / Regression / Linear parancsot (59. ábra).



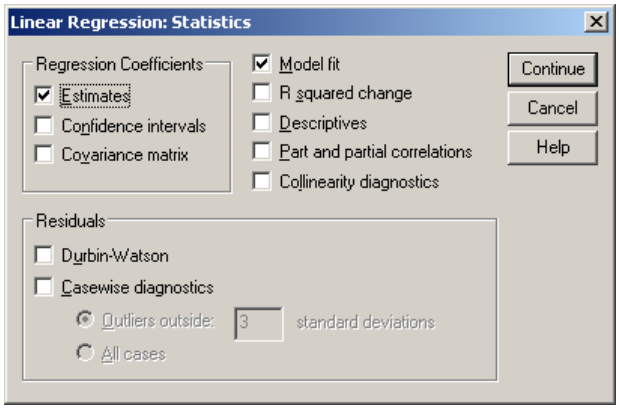
59. ábra

A megjelenő ablakban (60. ábra) a Dependent (függő változó) alá mozgassuk át a nyíl segítségével a Life of Components in minutes változót, míg az Independent(s) (független változó(k)) alá a Temperature in degree Fahrenheit változót, majd kattintsunk a Statistics gombra. A dependent a függő, míg az independent a független változót jelenti.



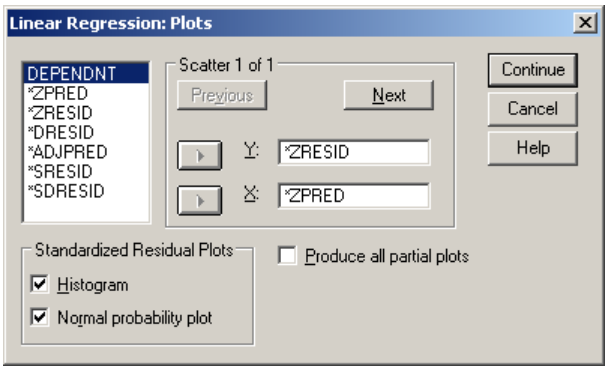
60. ábra

Az illeszkedésvizsgálathoz pipáljuk ki az Estimate és Model fit előtti négyzeteket, majd folytassuk a Continue, majd a Plots gombbal (61. ábra).



61. ábra

A homoszkedaszticitás (a hibatényező varianciája állandó) – mely a faktoranalízisnél is fontos – feltételének vizsgálatához a standardizált becsült értékre (ZPRED) és a standardizált reziduumokra (ZPRESID) lesz szükségünk. Ezért a Plots ablakban ezeket válasszuk ki (62. ábra).



62. ábra

A táblázatban (20. táblázat) az r értéke a korrelációs együttható értékét (0,748) mutatja, míg az R Square a determinációs együttható értékét (0,560), ez a teljes szórás százalékos magyarázatát (56 %) jelenti. Az Std. Error of the Estimate a becslés standard hibáját jelenti (25,937). Minél kisebb ennek az értéke, annál eredményesebb a vizsgálat.

Model Summary(b)

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,748(a)	,560	,532	25,931

a Predictors: (Constant), Temperature in degree Fahrenheit

b Dependent Variable: Life of Components in minutes

20. táblázat

Az ANOVA táblázat a regressziós egyenes által magyarázott (13667,556) és nem magyarázott (10758,444) szórásnégyzetet mutatja. Megtudhatjuk az F próba szignifikanciáját is, amelynek értéke kisebb, mint 0,05, tehát van kapcsolat (21. táblázat).

ANOVA(b)

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	13667,556	1	13667,556	20,326	,000(a)
	Residual	10758,444	16	672,403		
	Total	24426,000	17			

a Predictors: (Constant), Temperature in degree Fahrenheit

b Dependent Variable: Life of Components in minutes

21. táblázat

A t-próba szignifikancia szintje szintén kisebb, mint 0,05, így a hőmérsékletnek van befolyásoló ereje a sütő élettartamára. A Standardized Cofficients a regressziós egyenes meredekséget, míg az Unstandardized Cofficients adataiból a regressziós egyenes képletét lehet megtudni (22. táblázat).

Coefficients(a)

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta	B	Std. Error
1	(Constant)	263,000	19,328		13,607	,000
	Temperature in degree Fahrenheit	-55,111	12,224	-,748	-4,508	,000

a Dependent Variable: Life of Components in minutes

22. táblázat

A reziduálisokat az alábbi táblázat mutatja (23. táblázat).

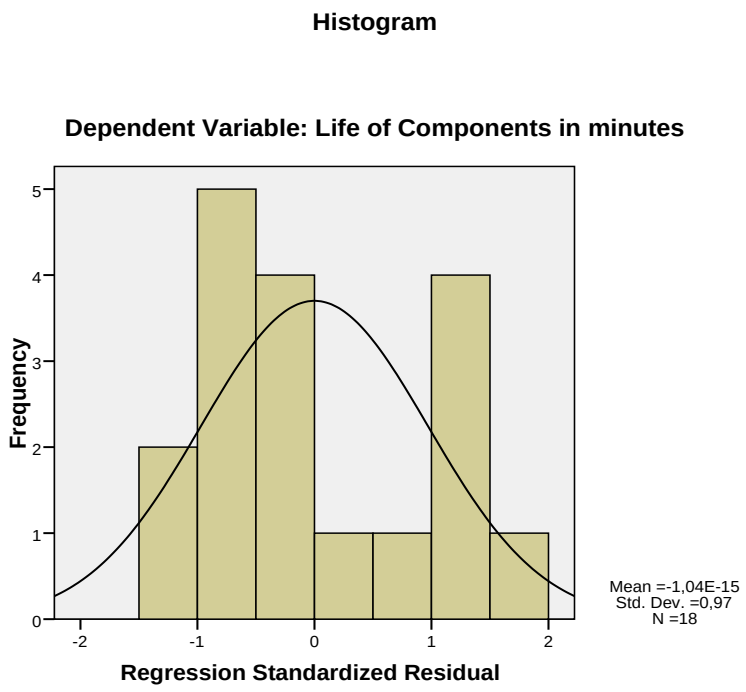
Residuals Statistics(a)

	Minimum	Maximum	Mean	Std. Deviation	N
Predicted Value	152,78	207,89	180,33	28,354	18
Residual	-29,889	46,111	,000	25,157	18
Std. Predicted Value	-,972	,972	,000	1,000	18
Std. Residual	-1,153	1,778	,000	,970	18

a Dependent Variable: Life of Components in minutes

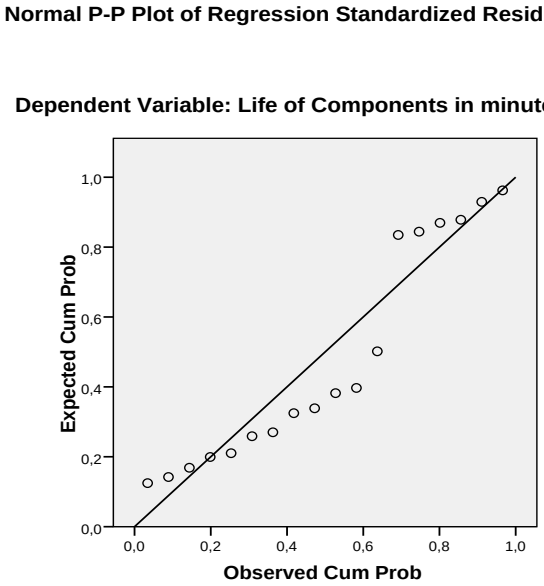
23. táblázat

A hisztogram segítségével azt a feltételt vizsgálhatjuk, hogy a rezidumok normálisan oszlanak-e el (63. ábra).



63. ábra

A 64. ábra a példa regressziós egyenesét mutatja meg, vagyis, hogy mennyire illeszkedik az egyenes a ponthalmazra.



64. ábra

6.3. Feladatok

1. Mi a regresszió lényege és célja, milyen típusai ismertek?
2. Definiálja a korrelációs és a determinációs együtthatók közötti kapcsolatot!
3. Mi az a legkisebb négyzetek módszere?
4. Mit értünk reziduális alatt?

7. Kereszttábla elemzés

7.1. Alapfogalmak

A ***kereszttábla*** változók közötti kapcsolat jellemzésére alkalmas adattábla. A mátrixban többnyire két nominális vagy ordinális változó értékeinek együttes eloszlása ábrázolható, azaz a változókhoz tartozó értékek kereszt-kombinációit jeleníti meg.
A kereszttáblák előnyei, hogy könnyen kiszámíthatók, az eredményei szemléletesek, a legalacsonyabb mérési szintű változók esetében is használhatók.
A kereszttáblákat két változó összefüggésének vizsgálatához használjuk. Ez a táblázat olyan cellákból áll, amelyek a két változó (oszlop- és sorváltozó) értékeinek minden kombinációja esetén kapott értékeket tartalmazza. Ezen cellák értékei szolgáltatnak információt a két változó közötti összefüggésről.

7.1.1. A cellák tartalma

A cellák elsődlegesen a két változó által meghatározott esetek számait, a gyakoriságot tartalmazzák (*Count*: ez a kulcsszó a táblázat bal felső sarkában látható). A második érték a sor százalék, amely a sor értékeinek a cellába eső hányadát mutatja *(Row Percentages*). A harmadik elem az oszlop százalék, amely az egész oszlop értékeinek a cellába eső hányadát mutatja (*Column Percentages*).
Az utolsó elem a táblázat százalék, amely a táblázat értékeinek a cellába eső hányadát mutatja (*Table Percentage*).
A táblázat alatt és tőle jobbra látható értékek a határértékek (*marginals*), amelyek az oszlop és sor változók százalék és számértékeit külön-külön tartalmazzák.

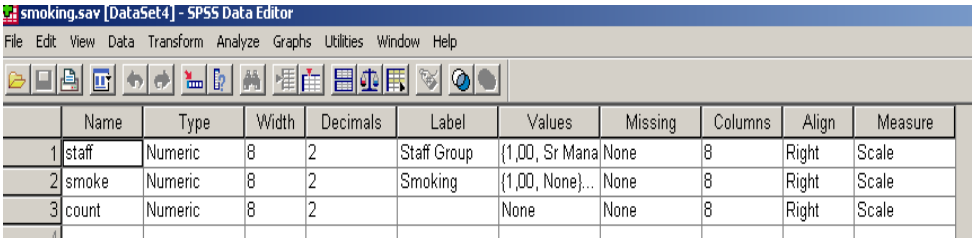
7.1.2. Kereszttábla statisztikák, a khi-négyzet próba

A kereszttáblában szereplő százalék- és számértékek nem elegendőek a két változó közötti kapcsolat jellemzésére. Egy lehetséges módszer erre a khi-négyzet próba.
Két változó akkor független, ha az egyes cellába eső esetek számát a peremeloszlások egyértelműen megadják
Egy, a statisztikában gyakran használt hipotézisvizsgálati módszer a Pearson-khi-négyzet próba. Ez a vizsgálat nagyon robusztus, azaz a számítás körülményei és

az adatok eloszlása nem nagyon befolyásolja a hipotézisvizsgálat megbízhatóságát.
A khi-négyzet próbával a nullhipotézist (H_0) ellenőrizhetjük, amely egy összefüggés-vizsgálati esetben a következő: A vizsgált változók között nincsen összefüggés. Amennyiben a khi-négyzet értékhez tartozó szignifikanciaszint 0.05-nél alacsonyabb, akkor elvetjük a nullhipotézist, ellenkező esetben megtartjuk. (A szignifikanciaérték a khi-négyzet eloszlás elméleti értékének az adatainkból kiszámított khi-négyzet értékkel való összehasonlításából származik.

7.2. Példa kereszttáblák használatára

Nyissuk meg a smoking.sav állományt az SPSS-ben (65. ábra). Az adatbázisban munkaköri megoszlásokat (staff) láthatunk és azt tudjuk meg, hogy ki milyen fokon dohányzik (smoke). A count egy előzetes számolást tartalmaz azonos értékek esetén (például 10 olyan titkárnő van, aki nem dohányzik) (66. ábra).



	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure
1	staff	Numeric	8	2	Staff Group	{1,00, Sr Mana	None	8	Right	Scale
2	smoke	Numeric	8	2	Smoking	{1,00, None}...	None	8	Right	Scale
3	count	Numeric	8	2		None	None	8	Right	Scale

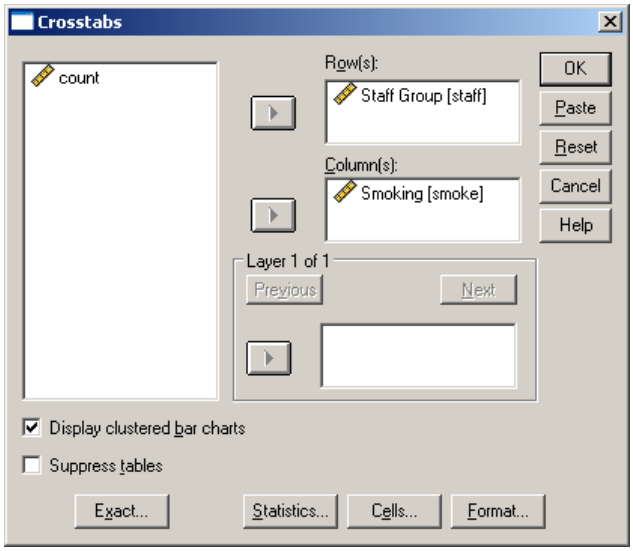
65. ábra

	staff	smoke	count
1	1,00	1,00	4,00
2	1,00	2,00	2,00
3	1,00	3,00	3,00
4	1,00	4,00	2,00
5	2,00	1,00	4,00
6	2,00	2,00	3,00
7	2,00	3,00	7,00
8	2,00	4,00	4,00
9	3,00	1,00	25,00
10	3,00	2,00	10,00
11	3,00	3,00	12,00
12	3,00	4,00	4,00
13	4,00	1,00	18,00
14	4,00	2,00	24,00
15	4,00	3,00	33,00
16	4,00	4,00	13,00
17	5,00	1,00	10,00
18	5,00	2,00	6,00
19	5,00	3,00	7,00

66. ábra

A *Crosstabs* (Kereszttáblák) az Analyze/Descriptive Statistics menüpont alatt érhető el.

Vigyünk egy vagy több változót a *Row(s)* (sorok) ill. *Column(s)* (oszlopok) dobozba. Jelen esetben a sor tartalmazza a Staff Group változót, míg az oszlop Smoking változót (67. ábra). A sorváltozók kategóriái adják meg a tábla sorait, az oszlopváltozó kategóriái pedig a tábla oszlopait. Minden egyes sor- és oszlopváltozó párhoz generálódik egy kereszttábla. A *Suppress Tables* opció kiválasztása után csak a statisztikákat fogja megjeleníteni a program, a táblákat nem.



67. ábra

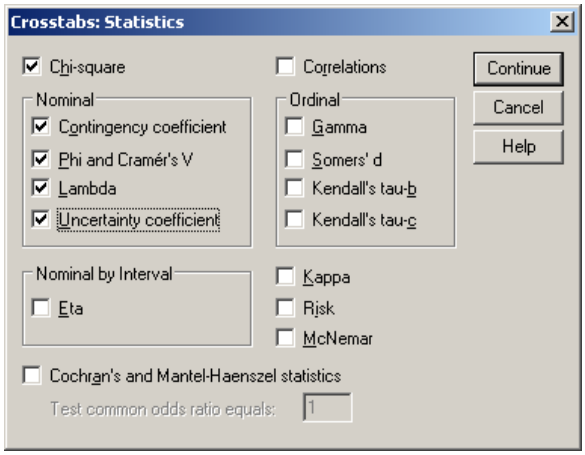
A *Statistics* gomb lenyomásával a sor- és oszlopváltozókra jellemző kétváltozós statisztikákat kérhetünk (68. ábra).

Válasszuk a Qhi-squere (khi-négyzet) vizsgálatot, majd nominális változók közül a Contingency cofficient, Phi and Cramér's V, Lambda, Uncertaninty cofficient vizsgálatokat.

A khi-négyzet statisztikát arra használjuk, hogy azt a hipotézist, miszerint a sor és oszlopváltozók függetlenek, ellenőrizhessük. Nem jól használható, ha bármelyik cellában a peremeloszlások alapján várható érték (expected value) kisebb 1-nél, vagy a cellák több mint 20%-ban ez az érték kisebb, mint 5. A Pearson khi-négyzet a legelterjedtebb forma, a likelihood-ratio khi-négyzet a maximum likelihood elméleten alapszik.

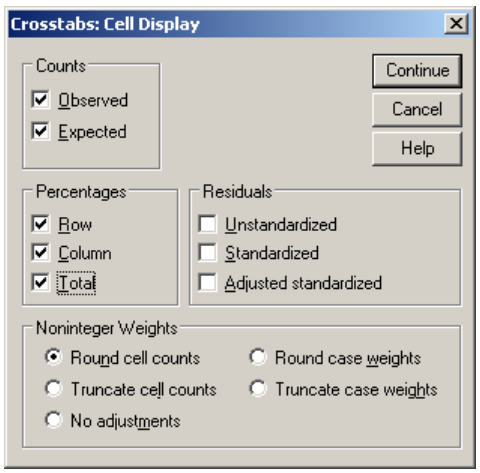
A *phi együttható* a khi-négyzetnek a mintanagysággal korrigált értéke. A *kontingencia együttható* a mintanagyságot használja a számításnál.

A *lambda* százalékos formában azt mutatja meg, hogy függő változót a független változó milyen mértékben képes előre jelezni. A Cramer V a legmegbízhatóbb mutató, aminek a számításához szükség van a mintanagyságra és a kevesebb lehetőséget felkínáló ismert kategóriák számára.



68. ábra

A *Cells* gomb lenyomása után a cellák tartalmát határozhatjuk meg. A *Format* gomb lenyomása után megjelenő párbeszédablakban a táblázat formátumát adhatjuk meg. Pipáljuk az Observed, Expected négyzeteket, majd a Row, Column, Total mezőket, hogy megjeleníthessük a sorokat, oszlopokat és ezek összesítését (69. ábra).



69. ábra

A kereszttábla megmutatja, hogy a minta összesen 486 főt tartalmaz, és nincs hiányzó érték (24. táblázat).

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
Staff Group * Smoking	486	100,0%	0	,0%	486	100,0%

24. táblázat

Az alábbi táblázat (25. táblázat) a Staff Group és a Smoking változók részletes megoszlását mutatja.

Staff Group * Smoking Crosstabulation

			Smoking						
			None	Light	Medium	Heavy	No Alcohol	Alcohol	Total
Staff Group	Sr Managers	Count	4	2	3	2	0	11	22
		Expected Count	4,7	3,3	3,7	1,5	1,0	7,7	22,0
		% within Staff Group	18,2%	9,1%	13,6%	9,1%	,0%	50,0%	100,0%
		% within Smoking	3,9%	2,7%	3,7%	5,9%	,0%	6,5%	4,5%
		% of Total	,8%	,4%	,6%	,4%	,0%	2,3%	4,5%
	Jr Managers	Count	4	3	7	4	1	17	36
		Expected Count	7,6	5,5	6,1	2,5	1,7	12,6	36,0
		% within Staff Group	11,1%	8,3%	19,4%	11,1%	2,8%	47,2%	100,0%
		% within Smoking	3,9%	4,1%	8,5%	11,8%	4,3%	10,0%	7,4%
		% of Total	,8%	,6%	1,4%	,8%	,2%	3,5%	7,4%
	Sr Employees	Count	25	10	12	4	5	46	102
		Expected Count	21,6	15,5	17,2	7,1	4,8	35,7	102,0
		% within Staff Group	24,5%	9,8%	11,8%	3,9%	4,9%	45,1%	100,0%
		% within Smoking	24,3%	13,5%	14,6%	11,8%	21,7%	27,1%	21,0%
		% of Total	5,1%	2,1%	2,5%	,8%	1,0%	9,5%	21,0%
	Jr Employees	Count	18	24	33	13	10	78	176
		Expected Count	37,3	26,8	29,7	12,3	8,3	61,6	176,0
		% within Staff Group	10,2%	13,6%	18,8%	7,4%	5,7%	44,3%	100,0%
		% within Smoking	17,5%	32,4%	40,2%	38,2%	43,5%	45,9%	36,2%
		% of Total	3,7%	4,9%	6,8%	2,7%	2,1%	16,0%	36,2%
	Secretaries	Count	10	6	7	2	7	18	50
		Expected Count	10,6	7,6	8,4	3,5	2,4	17,5	50,0
		% within Staff Group	20,0%	12,0%	14,0%	4,0%	14,0%	36,0%	100,0%
		% within Smoking	9,7%	8,1%	8,5%	5,9%	30,4%	10,6%	10,3%
		% of Total	2,1%	1,2%	1,4%	,4%	1,4%	3,7%	10,3%
	National Average	Count	42	29	20	9	0	0	100
		Expected Count	21,2	15,2	16,9	7,0	4,7	35,0	100,0
		% within Staff Group	42,0%	29,0%	20,0%	9,0%	,0%	,0%	100,0%
		% within Smoking	40,8%	39,2%	24,4%	26,5%	,0%	,0%	20,6%
		% of Total	8,6%	6,0%	4,1%	1,9%	,0%	,0%	20,6%
Total		Count	103	74	82	34	23	170	486
		Expected Count	103,0	74,0	82,0	34,0	23,0	170,0	486,0
		% within Staff Group	21,2%	15,2%	16,9%	7,0%	4,7%	35,0%	100,0%
		% within Smoking	100,0%	100,0%	100,0%	100,0%	100,0%	100,0%	100,0%
		% of Total	21,2%	15,2%	16,9%	7,0%	4,7%	35,0%	100,0%

25. táblázat

A Pearson-féle Khi-négyzet próba alapján megállapítható, hogy a két változó szignifikáns (26. táblázat). A Khi-négyzet értéke (χ^2) 117,025, míg a szabadságfok (df) 25.

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	117,025[a]	25	,000
Likelihood Ratio	150,226	25	,000
Linear-by-Linear Association	51,455	1	,000
N of Valid Cases	486		

a. 11 cells (30,6%) have expected count less than 5. The minimum expected count is 1,04.

26. táblázat

Az alábbi táblázatban (27. táblázat) látható, hogy a Lambda, a Goodman and Kruskal tau, Uncertainty Coffiecients szignifikancia szintje kisebb, mint 0,05. A Value értékeik a becslés hibavalószínűségének csökkenését jelzik, ha felszorozzuk őket százzal. A két változó (Staff Group és Smoking) Value értékei nem egyenlők, tehát a két változó nem azonos mértékben van hatással a másikra.

Directional Measures

			Value	Asymp. Std. Error[a]	Approx. T[b]	Approx. Sig.
Nominal by Nominal	Lambda	Symmetric	,113	,023	4,696	,000
		Staff Group Dependent	,094	,033	2,749	,006
		Smoking Dependent	,133	,019	6,780	,000
	Goodman and Kruskal tau	Staff Group Dependent	,065	,010		,000[c]
		Smoking Dependent	,071	,008		,000[c]
	Uncertainty Coefficient	Symmetric	,097	,010	9,548	,000[d]
		Staff Group Dependent	,097	,010	9,548	,000[d]
		Smoking Dependent	,096	,010	9,548	,000[d]

a. Not assuming the null hypothesis.

b. Using the asymptotic standard error assuming the null hypothesis.

c. Based on chi-square approximation

d. Likelihood ratio chi-square probability.

27. táblázat

A Phi, Cramer V, és a kontingencia együttható – azaz a szimmetrikus mutatók – mind szignifikánsak (Sig.<0,05). A kapcsolat erőssége a Phi alapján 0,491, Cramer V szerint 0,219, míg a kontingencia együttható szerint 0,441 (28. táblázat).

Symmetric Measures

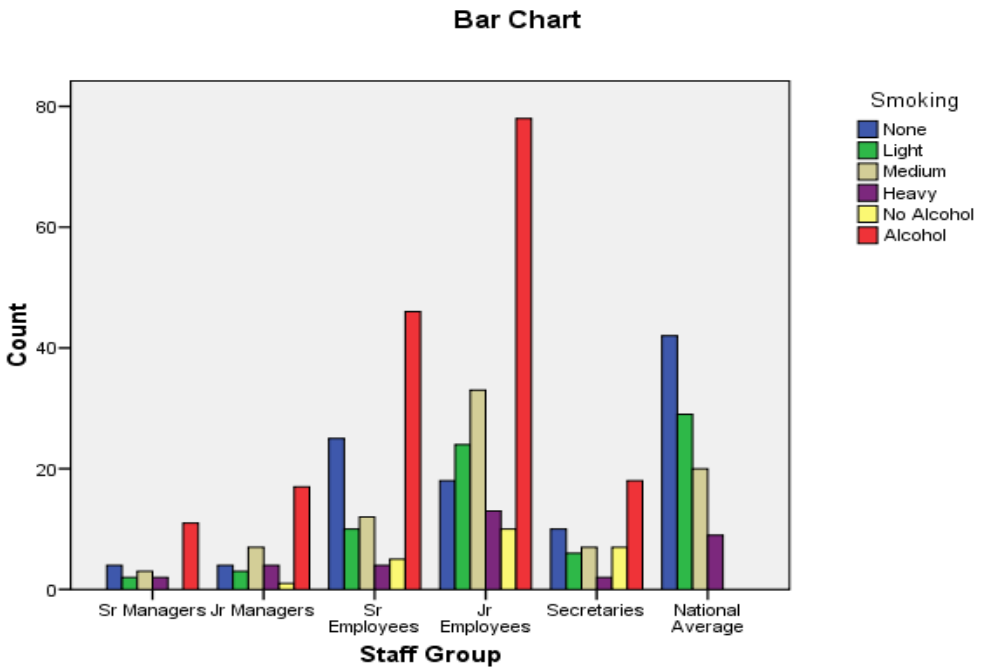
		Value	Approx. Sig.
Nominal by Nominal	Phi	,491	,000
	Cramer's V	,219	,000
	Contingency Coefficient	,441	,000
N of Valid Cases		486	

a. Not assuming the null hypothesis.

b. Using the asymptotic standard error assuming the null hypothesis.

28. táblázat

A kereszttábla eredményeit a Bar Chart jól ábrázolja (70. ábra).



70. ábra

7.3. Feladatok

1. Mi az a kereszttábla, milyen adatokból áll?
2. Milyen elemzések végezhetők el kereszttáblák segítségével?

3. Mire használjuk a kereszttábla statisztikáit?

8. Klaszteranalízis

8.1. Alapfogalmak

Elsősorban folytonos változók esetében alkalmazott statisztikai eljárás. Az eljárás a megadott változók segítségével csoportokat alakít ki. A csoportképzés távolságok mérésén alapul. Azokat tekintjük egy csoportban lévőknek, akik elkülönülten közel vannak egymáshoz. Az elemzés nehézsége leginkább abban áll, hogy a kialakult csoportoknak tudunk-e olyan nevet adni, ami jól leírja az adott csoportot a több csoporthoz képest. Tehát hasonló dolgok csoportosítását jelenti, gyakorlatilag az osztályozás szinonimájaként értelmezhetjük.
A **klaszteranalízis alapvető célja**, hogy a megfigyelési egységeket viszonylag homogén csoportokba rendezze, az elemzésbe bevont változók alapján.
A folyamat akkor sikeres, ha az egységek hasonlítanak csoporttársaikhoz, azonban eltérnek a más csoportba tartozó elemektől.

8.1.1. A klaszterelemzés technikája

Arra törekszünk, hogy a leginkább összetartozó elemek kerüljenek egy csoportba, a lehető legtöbb tulajdonság szerint.
Hierarchikus módszer: az adatok elemzése során hozzuk létre az osztályokat. Az összevonó eljárás során minden elem külön osztályba kerül a legközelebbiek összevonásával, míg a felosztó eljárás felülről lefelé, az egészet osztja külön osztályokba.
Az összevonó eljárások fajtái:

- A legközelebbi szomszéd módszere: a különálló elemeket egymástól való távolság szerint csoportosítjuk. Minél kisebb a távolság, annál jobb. Minden elem egymástól való távolságát kell számolni. A két legközelebbit kell összevonni. Addig kell folytatni, amíg van különálló. Ennek nyomonkövetése és ábrázolása dendrogrammal történik.
- A legtávolabbi szomszéd módszere.
- Centroid: az objektumok átlaga közötti távolságot jelenti.
- Csoportátlag: az össze lehetséges elemtávolság átlaga.
- Négyzetösszeg módszer.

Az értékek között korreláció szükséges, mert azok a korrelációs együtthatók szerepelhetnek benne, amelyek szignifikánsak, ezért szignifikancia vizsgálatot igényel.

8.1.2. A klaszterelemzés korlátai

- Nem vonhatók le következtetések a mintából az alapsokaságra, csak feltáró technikaként használható.
- Nincs egyetlen legjobb megoldás.
- Minden esetben létrehoz klasztereket.
- A megoldások a változóktól függnek.
- A kialakult csoportok függnek az egyedek adatbázisbeli sorrendjétől.

8.1.3. Vizsgálatok

Vizsgáljuk meg, hogy szükség van-e a skálák standardizálásra. Ez akkor fordulhat elő, ha nem egyforma skálákat használunk. A változókon végezzünk korrelációelemzést, hogy kiderítsük, elvégezhető-e az analízis. Ellenőrizzük a minta reprezentativitását. Meg kell vizsgálni, hogy vannak-e kiugró értékek, és amennyiben feltártuk azokat, akkor szüntessük meg. Vizsgálnunk kell a változó skálákat is. Ezeket a továbbiakban részletesebben kifejtjük.
Fontos eldöntenünk, hogy milyen hasonlósági- vagy távolságmértéket válasszunk. Bináris változó esetében mindkét típus fajtáiból választhatunk. Metrikus változó esetében távolságmértékeket alkalmazunk.
Válasszuk ki, hogy milyen a klasztermódszert szeretnénk használni: hierarchikus eljárást, nem hierarchikus eljárást, vagy a kettő kombinációját.
Gondoljuk át, hogy milyen szempontok alapján történik az elemzés, hány csoportot képezzünk, a csoportok számának változtatása hogyan hat az eredményekre.
Állapítsuk meg, hogy miben különböznek egymástól a klaszterek, értelmesen interpretálhatók-e az eredmények, szükség van-e új változók bevonására, és milyen nevet adjunk majd a kialakult klasztereknek.
Az elemzés érvényességének elemzése úgy történik, hogy különböző eljárásokat, vagy távolságmértékeket alkalmazunk és összehasonlítjuk az eredményeket. Az adatokat véletlenszerűen két részre osztjuk. A két almintán külön-külön elvégzett elemzések eredményeit összehasonlítjuk. Az elemzést többször lefuttatjuk az adatok sorrendjének megváltoztatásával.

Mint már említettük, a feltételek vizsgálata során fontos a kiugró értékek feltárása, mivel a klaszteranalízis rendkívül érzékeny az olyan egyedekre, amelyek jelentősen különböznek a többitől. Feltárásuk történhet egyszerű grafikus ábrázolással: pontdiagram, boxplot ábra vagy egyszerű láncmódszer segítségével. Ezek az elemek vagy ténylegesen „abnormális" megfigyelések, amelyek nem jellemzők az alapsokaságra, vagy a mintában szereplő egyedek alulreprezentálják az alapsokaságban lévő csoport nagyságát. Az első esetben tehát érdemes azokat kitörölni az adatbázisból.
Klaszterelemzés során fontos, hogy egyforma szintű metrikus skálákat használjunk. Ha a klaszteranalízis során különböző szintű metrikus skálákat alkalmazunk, teljesen torz összevonási sémát kaphatunk eredményül. A különböző skálák azonos szintre hozásához a standardizálást használjuk, amely során az átlagot kivonjuk az egyes értékekből és a különbséget elosztjuk a szórással. Így azonos szintű skálákat kapunk, lehetővé válik a különböző szintű skálán mért változók összehasonlítása. A standardizált skála szórása 1, az átlaga 0, a pozitív értékek átlag felettiek, a negatívak átlag alattiak
Szükséges a korrelációelemzés, mert a klaszterelemzés minden változót azonos súllyal kezel. Ha tehát két változó, vagy egy változócsoport tagjai egymással szoros korrelációs kapcsolatban vannak, akkor nagyobb szerepet kaphatnak az eredményekben. Ilyen esetben célszerű a változók valamilyen módon történő redukálása.
Bináris és metrikus változók esetén mind a távolságmértékeknél, mind a hasonlósági mértékeknél használatos az euklideszi távolság.

8.1.4. Hierarchikus összevonó eljárások

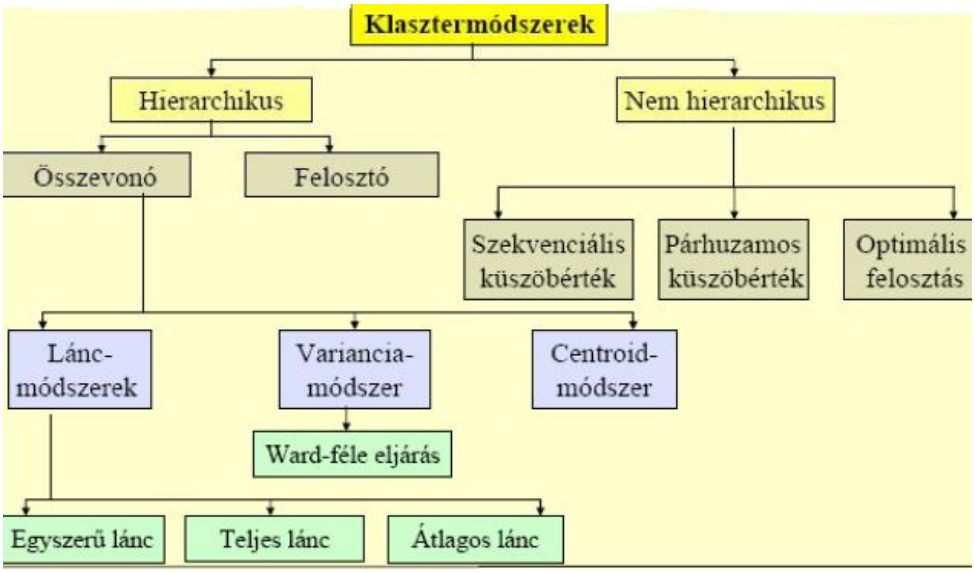
- Egyszerű láncmódszer (Single linkage): Azokat a megfigyelési egységeket vonja össze első lépésben, amelyek között legkisebb a távolság (legjobban hasonlítanak egymáshoz). Két klaszter közötti távolságot mindig a két legközelebbi pont távolsága határozza meg.
- Teljes láncmódszer (Complete linkage): Két klaszter közötti távolságot a két legtávolabbi pont határozza meg.
- Átlagos láncmódszer: Két klaszter távolságát az összes megfigyelési egység páronkénti távolságának átlaga definiálja. (általában előnyösebb, mint az előzőek).
- Hierarchikus összevonó eljárások.

- Ward-féle eljárás: Minden klaszterre kiszámolják az összes változó átlagát, majd minden megfigyelési egységre meghatározzák a négyzetes euklideszi távolságot. Minden lépésnél azt a két klasztert vonják össze, amelyeknél a klaszteren belüli szórásnégyzet növekedése a legkisebb.
- Centroidmódszer: Két klaszter közötti távolságnak az összes változó átlaga közötti távolsága. Ezeket minden lépés után újra számolják.

8.1.5. Nem hierarchikus eljárások

- Szekvenciális küszöbérték módszer: Kiválasztjuk a klaszter-középpontot, és minden egység, ami a középponttól egy előre meghatározott küszöbértéken belülre esik egy klaszterbe kerül. Ezután új középpontot választunk és csoportosítjuk a fennmaradó egységeket (egy egységet csak egy klaszter-középponttal lehet csoportosítani).
- Párhuzamos küszöbérték módszer: A klaszter-középpontokat itt egyidejűleg választjuk ki, a küszöbértéken belüli egységeket pedig a legközelebb eső középponthoz rendeljük.
- Optimális felosztás módszere: A megfigyelési egységeket a folyamat során újra hozzárendeljük más klaszterekhez is, hogy egy általános kritériumot optimalizálhassunk (pl.: adott számú klaszterre a klaszteren belüli távolságok átlagát).

Nagyobb esetszámnál (például 1500) a hierarchikus klaszterezés már körülményesebb, ezért célszerű például a K-közép (K-Means) módszert választani.
Előre meg kell határozni a létrehozandó klaszterek számát. Induláskor ismertnek tételezzük fel a klaszterközepeket, amelyeket mi is megadhatunk, de érdemes a programra bízni ezek kijelölését.



71. ábra

8.2. Példa klaszteranalízisre

Nyissuk meg a verd1985.sav állományt (72. ábra). A következő feladatban különböző életkori (age) kategóriákba tartozó és különböző családi állapotú (marital) egyedek adathalmazait szeretnénk csoportba rendezni matematikai (math) és nyelvi tesztjeiknek (language) megfelelően. Az adatbázisban egyéb változók is szerepelnek: pet (hány háziállatot tart), news (milyen újságot olvas), music (milyen zenét szeret), live (milyen típusú településen lakik), amelyeket most figyelmen kívül hagyhatunk (73. ábra).

verd1985.sav [DataSet4] - SPSS Data Editor

File Edit View Data Transform Analyze Graphs Utilities Window Help

1 : age | 1

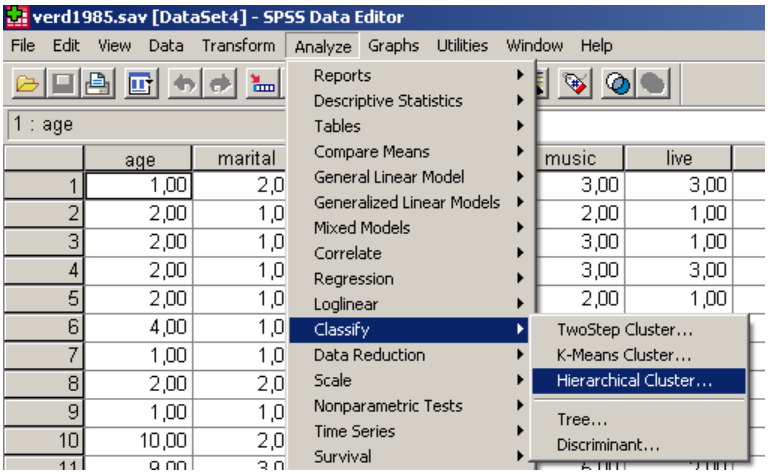
	age	marital	pet	news	music	live	math	language
1	1,00	2,00	2,00	3,00	3,00	3,00	3,00	4,00
2	2,00	1,00	1,00	3,00	2,00	1,00	3,00	3,00
3	2,00	1,00	2,00	3,00	3,00	1,00	1,00	2,00
4	2,00	1,00	5,00	3,00	3,00	3,00	3,00	3,00
5	2,00	1,00	3,00	2,00	2,00	1,00	1,00	1,00
6	4,00	1,00	1,00	3,00	3,00	1,00	3,00	3,00
7	1,00	1,00	1,00	4,00	2,00	1,00	2,00	2,00
8	2,00	2,00	2,00	2,00	5,00	3,00	2,00	2,00
9	1,00	1,00	1,00	2,00	3,00	2,00	2,00	2,00
10	10,00	2,00	1,00	2,00	3,00	1,00	1,00	1,00
11	9,00	3,00	2,00	2,00	5,00	2,00	1,00	4,00
12	8,00	3,00	1,00	4,00	1,00	1,00	2,00	3,00
13	10,00	2,00	1,00	2,00	5,00	2,00	1,00	2,00
14	5,00	2,00	2,00	3,00	1,00	1,00	1,00	2,00
15	8,00	2,00	1,00	4,00	1,00	3,00	3,00	4,00

72. ábra

	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure
1	age	Numeric	8	2	Age in years	{1,00, 20-25}...	None	8	Right	Ordinal
2	marital	Numeric	8	2	Marital status	{1,00, Single}...	None	8	Right	Nominal
3	pet	Numeric	8	2	Pets owned	{1,00, no}...	None	8	Right	Nominal
4	news	Numeric	8	2	Newspaper read most often	{1,00, None}...	None	8	Right	Nominal
5	music	Numeric	8	2	Music preferred	{1,00, Classical}...	None	8	Right	Nominal
6	live	Numeric	8	2	Neighborhood preference	{1,00, Town}...	None	8	Right	Nominal
7	math	Numeric	8	2	Math test score	{1,00, 0-5}...	None	8	Right	Ordinal
8	language	Numeric	8	2	Language test score	{1,00, 0-5}...	None	8	Right	Ordinal

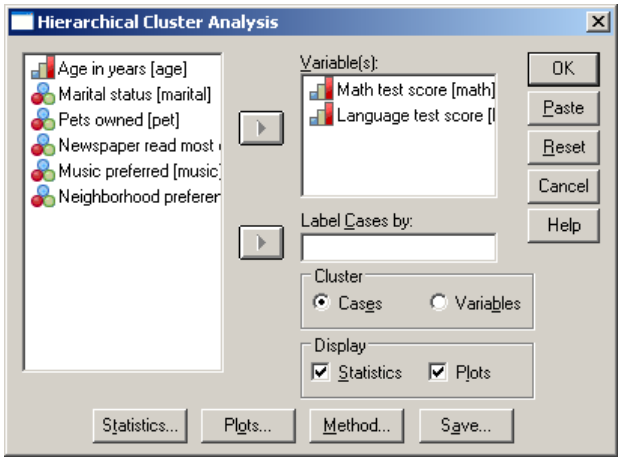
73. ábra

Mint láthatjuk, a math és language tesztpontszáma eltérő skálájú (a matematika hármas, míg a nyelv négyes skálás), ezért standardizálásra van szükségünk. Válasszuk az Analyze/Classify/Hierarchical Cluster parancsot (74. ábra).



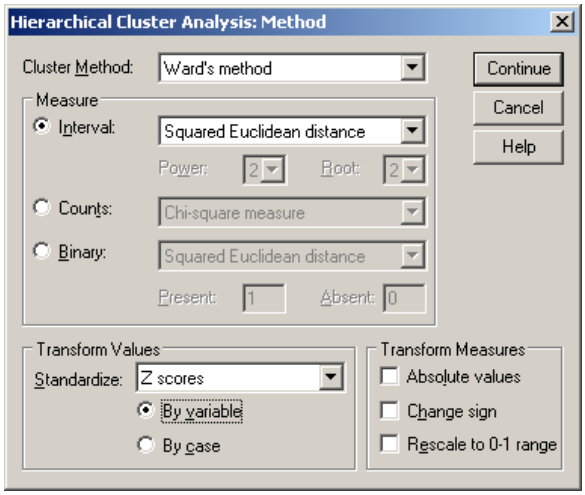
74. ábra

A Variable(s) alá mozgassuk át a vizsgálni kívánt Math test score és Languge test score változókat, majd kattintsunk a Method gombra (75. ábra).



75. ábra

A megjelenő ablakban válasszuk a Ward's methodot és Z score-nál a by variables mezőket (76. ábra).



76. ábra

Agglomeration Schedule

Stage	Cluster Combined		Coefficients	Stage Cluster First Appears		Next Stage
	Cluster 1	Cluster 2		Cluster 1	Cluster 2	
1	1	15	,000	0	0	11
2	13	14	,000	0	0	3
3	3	13	,000	0	2	10
4	5	10	,000	0	0	10
5	8	9	,000	0	0	6
6	7	8	,000	0	5	9
7	4	6	,000	0	0	8
8	2	4	,000	0	7	11
9	7	12	1,019	6	0	12
10	3	5	1,019	3	4	12
11	1	2	1,019	1	8	14
12	3	7	2,147	10	9	13
13	3	11	5,440	12	0	14
14	1	3	6,256	11	13	0

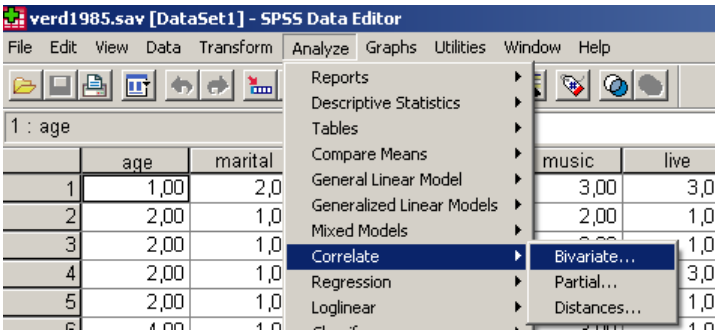
29. táblázat

Vertical Icicle

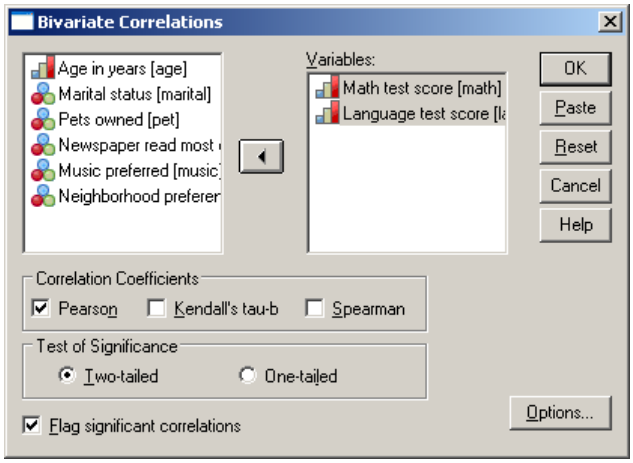
Number of clusters	Case																													
	11:Case 11		12:Case 12		9:Case 9		8:Case 8		7:Case 7		10:Case 10		5:Case 5		14:Case 14		13:Case 13		3:Case 3		6:Case 6		4:Case 4		2:Case 2		15:Case 15		1:Case 1	
1	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
2	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X		X	X	X	X	X	X	X	X	X	X
3	X		X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X		X	X	X	X	X	X	X	X	X	X
4	X		X	X	X	X	X	X	X		X	X	X	X	X	X	X	X	X		X	X	X	X	X	X	X	X	X	X
5	X		X	X	X	X	X	X	X		X	X	X	X	X	X	X	X	X		X	X	X	X	X		X	X	X	X
6	X		X	X	X	X	X	X	X		X	X	X		X	X	X	X	X		X	X	X	X	X		X	X	X	X
7	X		X		X	X	X	X	X		X	X	X		X	X	X	X	X		X	X	X	X	X		X	X	X	X
8	X		X		X	X	X	X	X		X	X	X		X	X	X	X	X		X	X	X		X		X	X	X	X
9	X		X		X	X	X	X	X		X	X	X		X	X	X	X	X		X		X		X		X	X	X	X
10	X		X		X	X	X		X		X	X	X		X	X	X	X	X		X		X		X		X	X	X	X
11	X		X		X		X		X		X	X	X		X	X	X	X	X		X		X		X		X	X	X	X
12	X		X		X		X		X		X		X		X	X	X	X	X		X		X		X		X	X	X	X
13	X		X		X		X		X		X		X		X	X	X		X		X		X		X		X	X	X	X
14	X		X		X		X		X		X		X		X		X		X		X		X		X		X	X	X	X

30. táblázat

A korrelációs vizsgálathoz válasszuk ki a már tanult Analyze/Correlate/Bivariate parancsot (77.ábra).



77. ábra



78. ábra

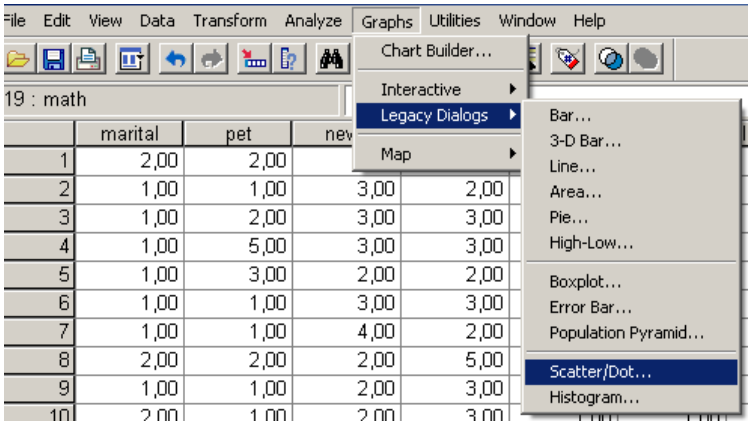
.Correlations

		Math test score	Language test score
Math test score	Pearson Correlation	1	,615(*)
	Sig. (2-tailed)		,015
	N	15	15
Language test score	Pearson Correlation	,615(*)	1
	Sig. (2-tailed)	,015	
	N	15	15

* Correlation is significant at the 0.05 level (2-tailed).

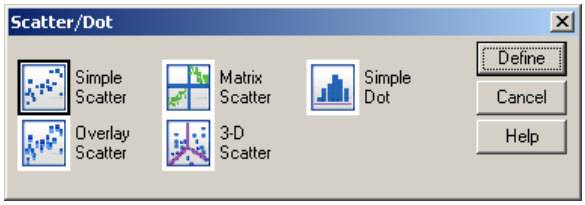
31. táblázat

A korrelációs analízisből látszik, hogy közepesen erős a korreláció a két változó között (31. táblázat). Így több érték fog egybeesni. Ennek ellenére most vizsgáljuk meg, hogy ha ez a feltétel nem teljesül, akkor mi történik.
Vizsgáljuk meg a továbbiakban pontfelhődiagram segítségével, hogy van-e kiugró érték az adatbázisban. Ehhez válasszuk a Graps/Legacy Dialogs/Scatter/Dot menüpontot (79. ábra).



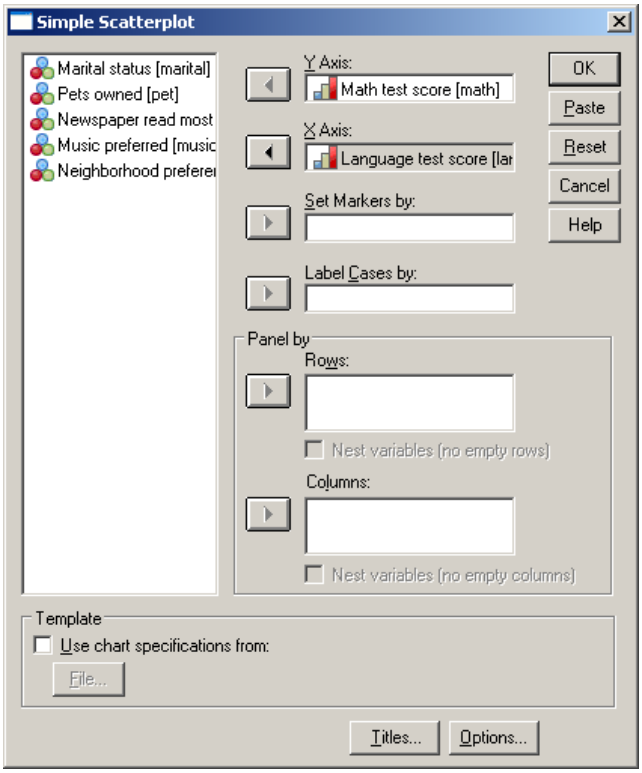
79. ábra

Az előugró panelben válasszuk ki a Simple Clustert az esetleges kiugró értékek szemléltetéséhez, majd nyomjuk meg a Define gombot (80. ábra).

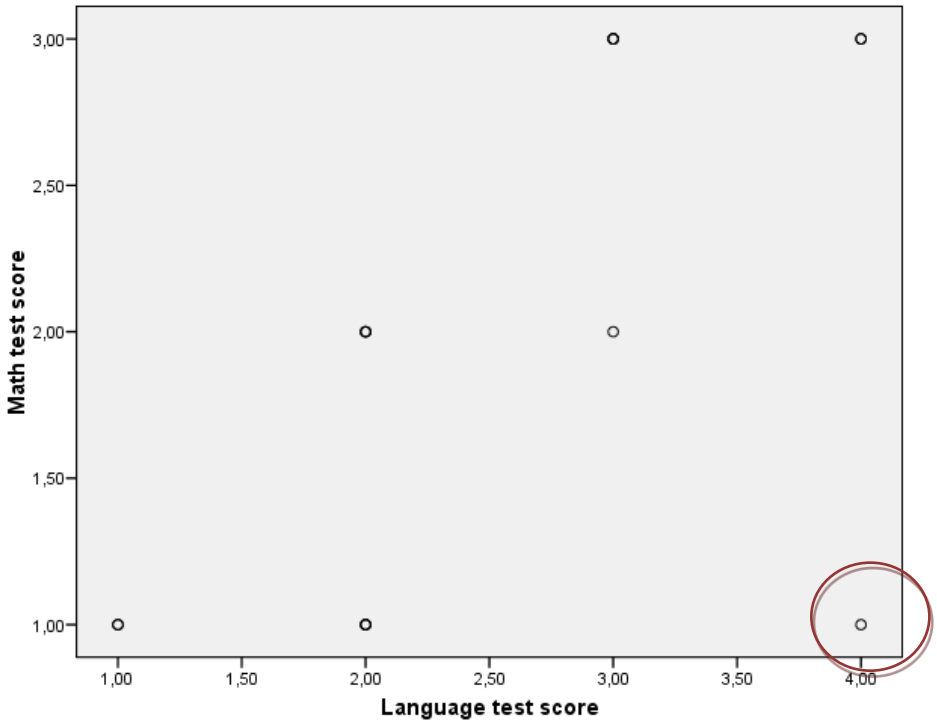


80. ábra

Ezután vigyük át a vizsgálandó változókat (Math test score and language test score) az Y Axis és X Axis alá. Amennyiben van egyedi azonosítóval rendelkező változónk, akkor a még jobb szemléltetés érdekében a Label Cases by (a pontok mellé írja az azonosítókat) vagy a Set Markers by (a pontokat színekkel látja el, majd az egyes színeket az azonosítóval párosítja) helyekre tehetjük (81. ábra).


Simple Scatterplot
Marital status [marital]
Pets owned [pet]
Newspaper read most
Music preferred [music
Neighborhood preferer
Y Axis:
Math test score [math]
X Axis:
Language test score [lar
Set Markers by:
Label Cases by:
OK
Paste
Reset
Cancel
Help
Panel by
Rows:
Nest variables (no empty rows)
Columns:
Nest variables (no empty columns)
Template
Use chart specifications from:
File...
Titles...
Options...

81. ábra



82. ábra

A pontfelhődiagram (82. ábra) azt mutatja, hogy van kiugró érték. Mivel viszonylag magas volt a korreláció és alacsonyak a skálák, így látható, hogy több érték is egybe esett.

Hogy szemléletesebbé tegyük a pontfelhődiagramot, hozzunk létre egyedi azonosítót (id) az egyes egyedeknek. Ennek érdekében váltsunk Variable View nézetre, majd írjuk be a név oszlopába az id változót, a tizedesvessző utáni értéket (decimals) csökkentsük 0-ra (83.ábra).

File Edit View Data Transform Analyze Graphs Utilities Window Help

	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure
1	marital	Numeric	8	2	Marital status	{1,00, Single}..	None	8	Right	Nominal
2	pet	Numeric	8	2	Pets owned	{1,00, no}...	None	8	Right	Nominal
3	news	Numeric	8	2	Newspaper rea	{1,00, None}...	None	8	Right	Nominal
4	music	Numeric	8	2	Music preferre	{1,00, Classica	None	8	Right	Nominal
5	live	Numeric	8	2	Neighborhood	{1,00, Town}...	None	8	Right	Nominal
6	math	Numeric	8	2	Math test scor	{1,00, 0-5}...	None	8	Right	Ordinal
7	language	Numeric	8	2	Language test	{1,00, 0-5}...	None	8	Right	Ordinal
8	id	Numeric	8	0		None	None	8	Right	Scale

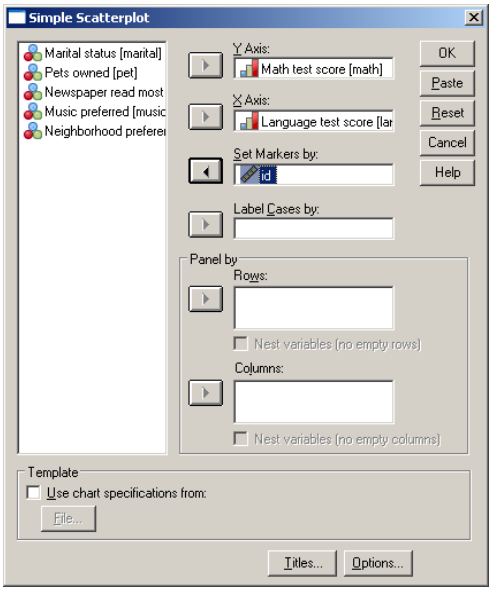
83. ábra

Ezután váltsunk vissza Data View nézetre, és gépeljük be az id változóhoz a sorok azonosítóit (84. ábra).

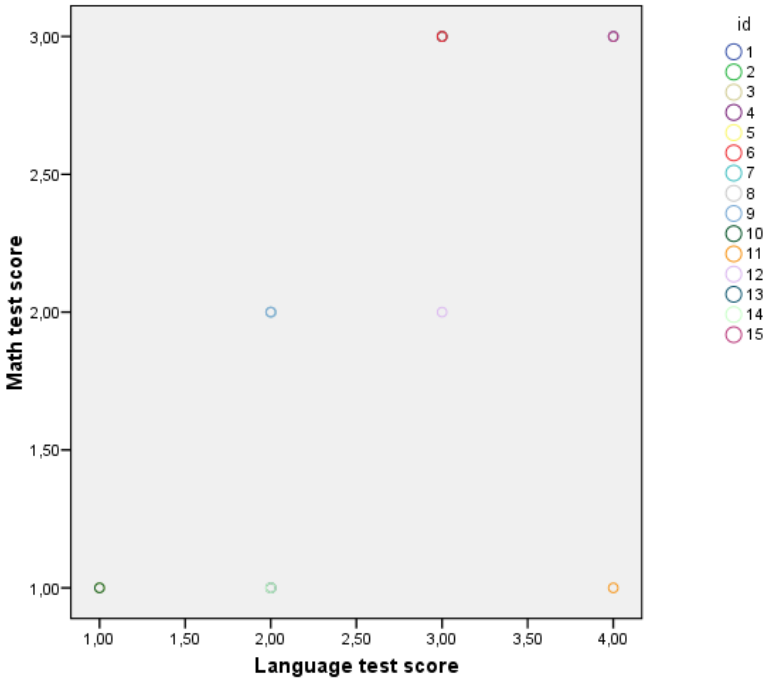
	marital	pet	news	music	live	math	language	id
1	2,00	2,00	3,00	3,00	3,00	3,00	4,00	1
2	1,00	1,00	3,00	2,00	1,00	3,00	3,00	2
3	1,00	2,00	3,00	3,00	1,00	1,00	2,00	3
4	1,00	5,00	3,00	3,00	3,00	3,00	3,00	4
5	1,00	3,00	2,00	2,00	1,00	1,00	1,00	5
6	1,00	1,00	3,00	3,00	1,00	3,00	3,00	6
7	1,00	1,00	4,00	2,00	1,00	2,00	2,00	7
8	2,00	2,00	2,00	5,00	3,00	2,00	2,00	8
9	1,00	1,00	2,00	3,00	2,00	2,00	2,00	9
10	2,00	1,00	2,00	3,00	1,00	1,00	1,00	10
11	3,00	2,00	2,00	5,00	2,00	1,00	4,00	11
12	3,00	1,00	4,00	1,00	1,00	2,00	3,00	12
13	2,00	1,00	2,00	5,00	2,00	1,00	2,00	13
14	2,00	2,00	3,00	1,00	1,00	1,00	2,00	14
15	2,00	1,00	4,00	1,00	3,00	3,00	4,00	15

84. ábra

Ismét menjünk a Graphs/Legacy Dialogs/Scatter/Dot menüponthoz, majd válasszuk a Simple Scatter-t és kattintsunk a Define gombra. A létrehozott azonosítónkat vigyük a Set Markers by mezőnévhez (85. ábra).



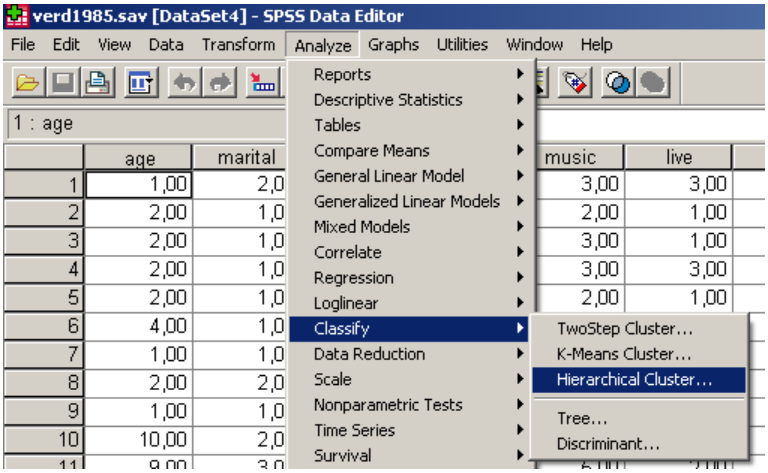
85. ábra



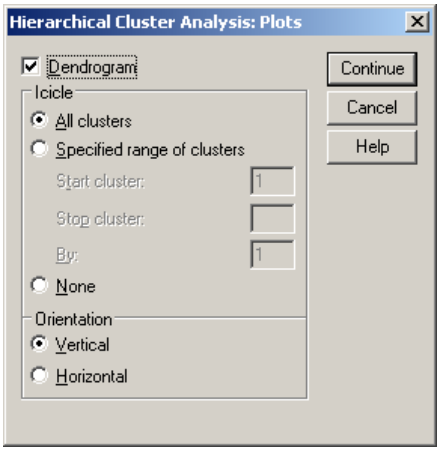
86. ábra

A szórásdiagramdiagram (86. ábra) jól szemlélteti, hogy nem mind a 15 elem esik más kategóriába, mivel egyes eredmények egybe esnek (ez a magasabb korreláció miatt lehetséges).

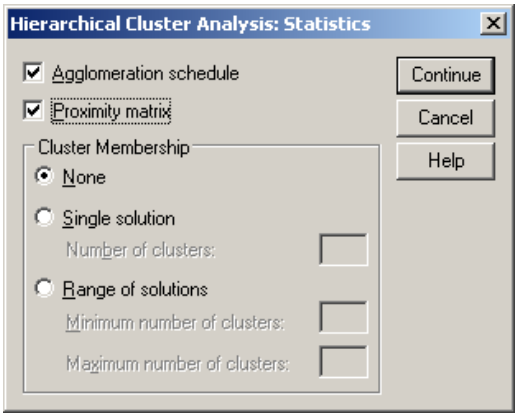
A kiugró érték megjelenítésének legalkalmasabb formája a dendrogram. Ehhez válasszuk az Analyze/Classify/Hierarchical Cluster paranancsnál (87. ábra) a Plots gombra (88. ábra) kattintva a dendrogramot és kattintsunk a Continue gombra.



87. ábra

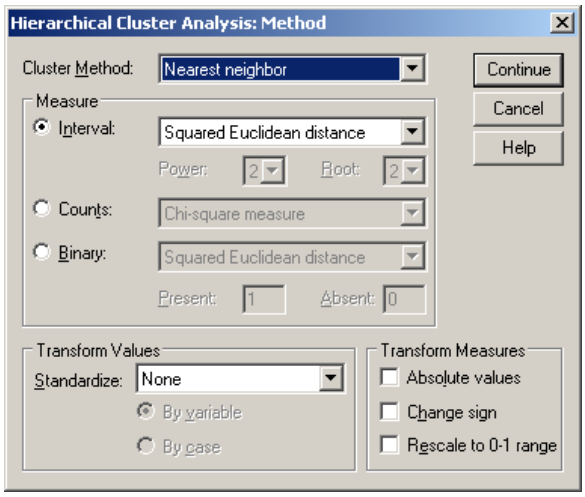


88. ábra



89. ábra

A Statistics gombra kattintva a proxy mátrixot és az Agglomeration schedule ábrát szeretnénk-e megjeleníteni (89. ábra), majd ismét a Continue gombra kattintsunk.



90. ábra

A Method gomra kattintva válasszuk Nearest neighbor (Legközelebbi szomszéd) módszert (90. ábra).

A klaszterek számának végső meghatározásában három szempontot vehetünk figyelembe. A hierarchikus klaszterelemzés során kapott összevonási táblázat (Agglomeration Schedule) (32. táblázat) Coefficients (koefficiens) oszlopában található érték ugrásszerű növekedése, másrészt a dendrogram, harmadrészt a lehetséges klaszterek szakmai értelmezhetősége.

Agglomeration Schedule

Stage	Cluster Combined		Coefficients	Stage Cluster First Appears		Next Stage
	Cluster 1	Cluster 2		Cluster 1	Cluster 2	
1	1	15	,000	0	0	13
2	13	14	,000	0	0	3
3	3	13	,000	0	2	10
4	5	10	,000	0	0	11
5	8	9	,000	0	0	6
6	7	8	,000	0	5	9
7	4	6	,000	0	0	8
8	2	4	,000	0	7	12
9	7	12	1,000	6	0	10
10	3	7	1,000	3	9	11
11	3	5	1,000	10	4	12
12	2	3	1,000	8	11	13
13	1	2	1,000	1	12	14
14	1	11	2,000	13	0	0

32. táblázat

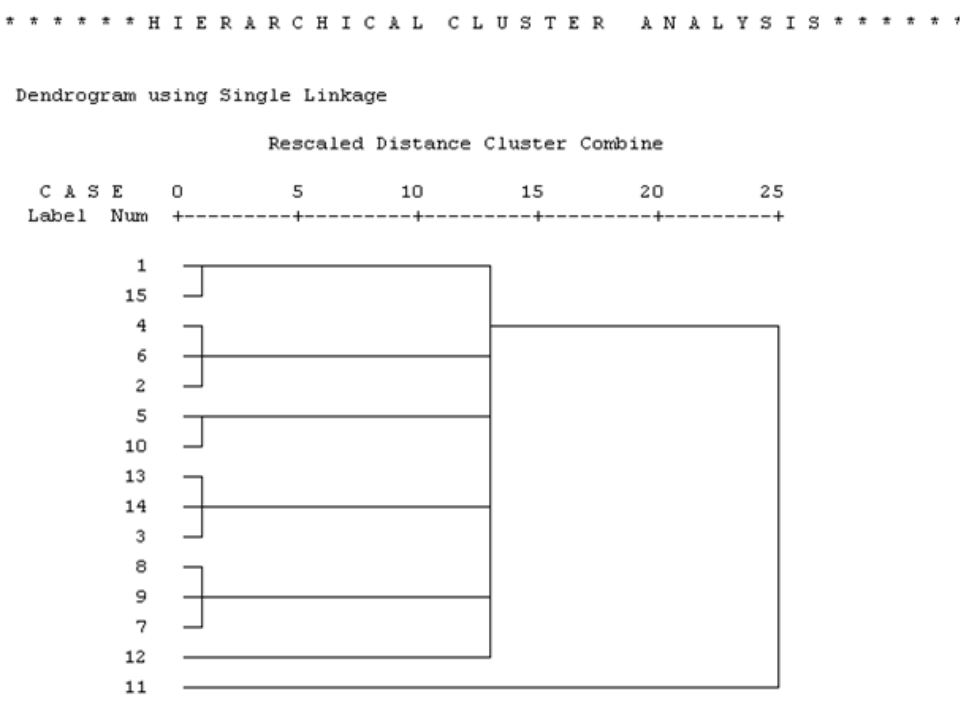
A dendrogrammal együtt kirajzolódik (33. táblázat) a jégcsap diagram (Icicle) különböző tájolással (Vertical/Horizontal), attól függően, hogy mit választottuk a Plots menüpontnál.

Vertical Icicle

Number of clusters	Case																													
	11		10		5		12		9		8		7		14		13		3		6		4		2		15		1	
1	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
2	X		X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
3	X		X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X		X	X	X	X
4	X		X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X		X	X	X	X	X		X	X	X	X
5	X		X	X	X		X	X	X	X	X	X	X	X	X	X	X	X	X		X	X	X	X	X		X	X	X	X
6	X		X	X	X		X	X	X	X	X	X	X		X	X	X	X	X		X	X	X	X	X		X	X	X	X
7	X		X	X	X		X		X	X	X	X	X		X	X	X	X	X		X	X	X	X	X		X	X	X	X
8	X		X	X	X		X		X	X	X	X	X		X	X	X	X	X		X	X	X		X		X	X	X	X
9	X		X	X	X		X		X	X	X	X	X		X	X	X	X	X		X		X		X		X	X	X	X
10	X		X	X	X		X		X	X	X		X		X	X	X	X	X		X		X		X		X	X	X	X
11	X		X	X	X		X		X		X		X		X	X	X	X	X		X		X		X		X	X	X	X
12	X		X		X		X		X		X		X		X	X	X	X	X		X		X		X		X	X	X	X
13	X		X		X		X		X		X		X		X	X	X		X		X		X		X		X	X	X	X
14	X		X		X		X		X		X		X		X		X		X		X		X		X		X	X	X	X

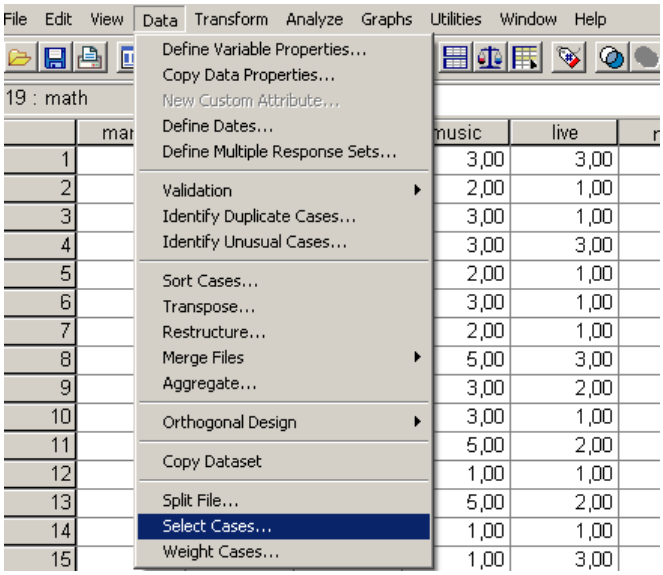
33. táblázat

A dendrogram segít eldönteni, hogy hány klasztert érdemes kialakítani. A dendrogramból (33. táblázat) jól látszik, hogy a 11-es a kiugró érték. El kell döntenünk, hogy ez a kiugró érték abnormális megfigyelés, vagy alulreprezentálja az alapsokaságban lévő csoport nagyságát.

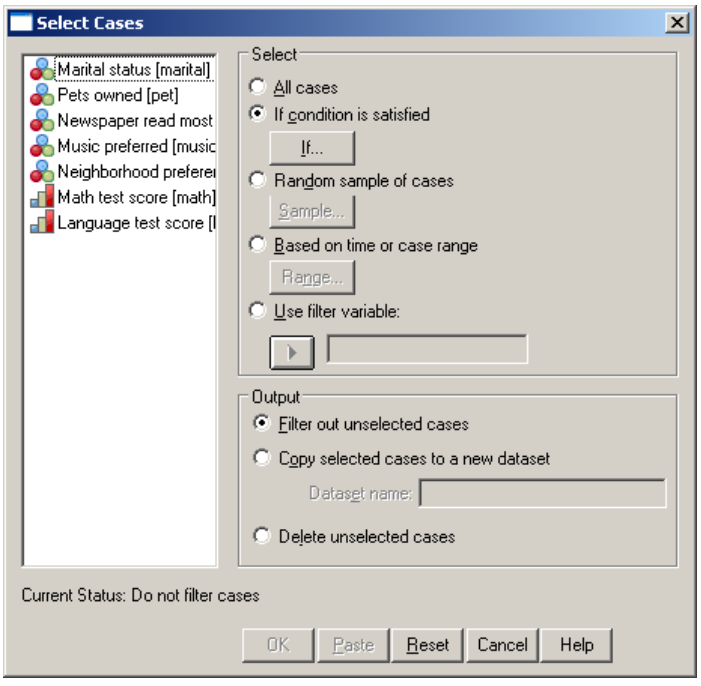


91. ábra

Mivel tudjuk, hogy melyik az az egyed (11-es számú), akit ki akarunk zárni, így nincs más dolgunk, mint kiszűrni. Ezért válasszuk a Data/Select Cases parancsát (92. ábra), majd If condition is satisfied alatt található If gombra kattintsunk (93. ábra).

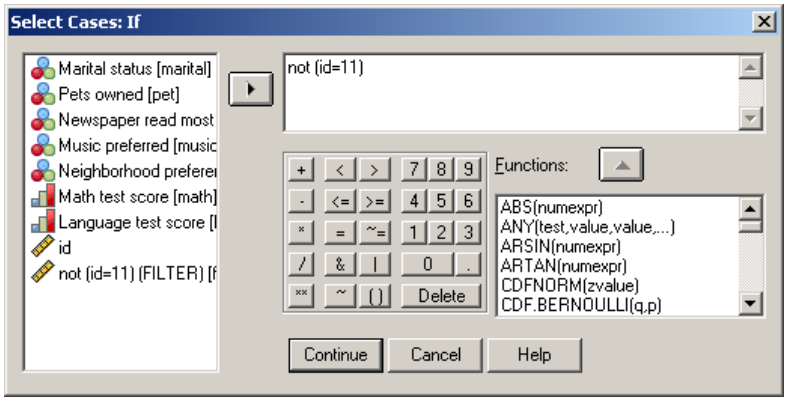


92. ábra



93. ábra

A szelektáláshoz egy tagadást kell alkalmaznunk, hiszen azt az egyedet nem szeretnénk, ha a vizsgálatainkban részt venne. Tehát a következő képletet alkalmazhatjuk: not (id=11). A jelen esetben a zárójel el is hagyható (94. ábra).



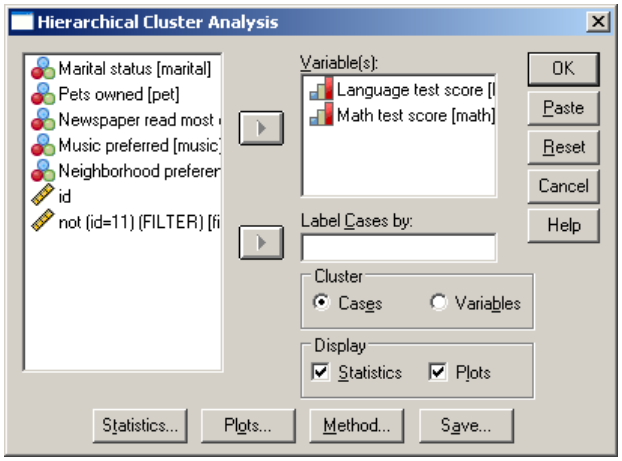
94. ábra

Az SPSS áthúzással jelzi, hogy melyik egyed nem fog szerepelni a vizsgálatban (95. ábra).

	marital	pet	news	music	live	math	language	id	filter_$
1	2,00	2,00	3,00	3,00	3,00	3,00	4,00	1	1
2	1,00	1,00	3,00	2,00	1,00	3,00	3,00	2	1
3	1,00	2,00	3,00	3,00	1,00	1,00	2,00	3	1
4	1,00	5,00	3,00	3,00	3,00	3,00	3,00	4	1
5	1,00	3,00	2,00	2,00	1,00	1,00	1,00	5	1
6	1,00	1,00	3,00	3,00	1,00	3,00	3,00	6	1
7	1,00	1,00	4,00	2,00	1,00	2,00	2,00	7	1
8	2,00	2,00	2,00	5,00	3,00	2,00	2,00	8	1
9	1,00	1,00	2,00	3,00	2,00	2,00	2,00	9	1
10	2,00	1,00	2,00	3,00	1,00	1,00	1,00	10	1
11	3,00	2,00	2,00	5,00	2,00	1,00	4,00	11	0
12	3,00	1,00	4,00	1,00	1,00	2,00	3,00	12	1

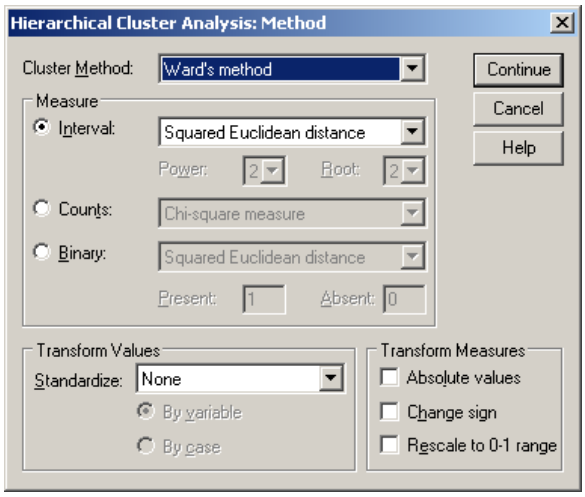
95. ábra

Ezt követően a Ward-eljárással haladunk tovább. Ez az eljárás akkor előnyös, ha a feltételeink teljesülnek, valamint a csoportok közel azonos szórásúak és minden csoport közel hasonló elemszámmal rendelkezik. Válasszuk az Analyze/Classify/Hierarchical Cluster parancsot (96. ábra).
Majd az előugró panelben válasszuk a Method gombot.


Hierarchical Cluster Analysis
Marital status [marital]
Pets owned [pet]
Newspaper read most
Music preferred [music]
Neighborhood preferer
id
not (id=11) (FILTER) [fi
Variable(s):
Language test score [l
Math test score [math]
Label Cases by:
Cluster
Cases
Variables
Display
Statistics
Plots
OK
Paste
Reset
Cancel
Help
Statistics...
Plots...
Method...
Save...

96. ábra

A Cluster Method lenyíló menüjéből válasszuk a Ward's methodot és a Transform Values lenyíló menüjéből a None-t (97. ábra).



97. ábra

A 34. táblázat egyrészt megmutatja az egyes elemek, klaszterek összevonási sorrendjét (Cluster Combine oszlopok), másrészt segít meghatározni, a megfelelő klaszterszámot. A legnagyobb szakadék megkeresése úgy történik, hogy meghatározzuk az egymást követő koefficiensek különbségét, és a szakadék előtti klasztermegoldást tekintjük a jó klasztermegoldásnak.

Ward Linkage

Agglomeration Schedule

Stage	Cluster Combined		Coefficients	Stage Cluster First Appears		Next Stage
	Cluster 1	Cluster 2		Cluster 1	Cluster 2	
1	1	15	,000	0	0	11
2	13	14	,000	0	0	3
3	3	13	,000	0	2	10
4	5	10	,000	0	0	10
5	8	9	,000	0	0	6
6	7	8	,000	0	5	12
7	4	6	,000	0	0	8
8	2	4	,000	0	7	9
9	2	12	,750	8	0	11
10	3	5	1,950	3	4	12
11	1	2	3,367	1	9	13
12	3	7	5,542	10	6	13
13	1	3	21,429	11	12	0

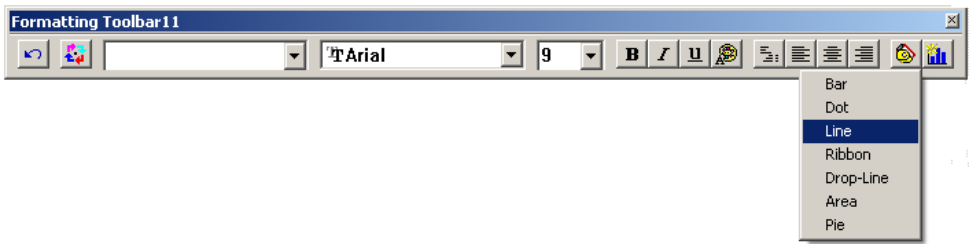
34. táblázat

Egy nagy ugrást (5,542-ről 21,429-re) láthatunk az utolsó két klaszter összevonása miatt. Ezt az ugrást megjeleníthetjük úgy, hogy a 4. táblázatra kétszer rákattintunk, majd kijelöljük egér segítségével az utolsó kofficienseket (coefficients) (35. táblázat) és a Formating Toolbarnál a Line diagramot választjuk ki (98. ábra).

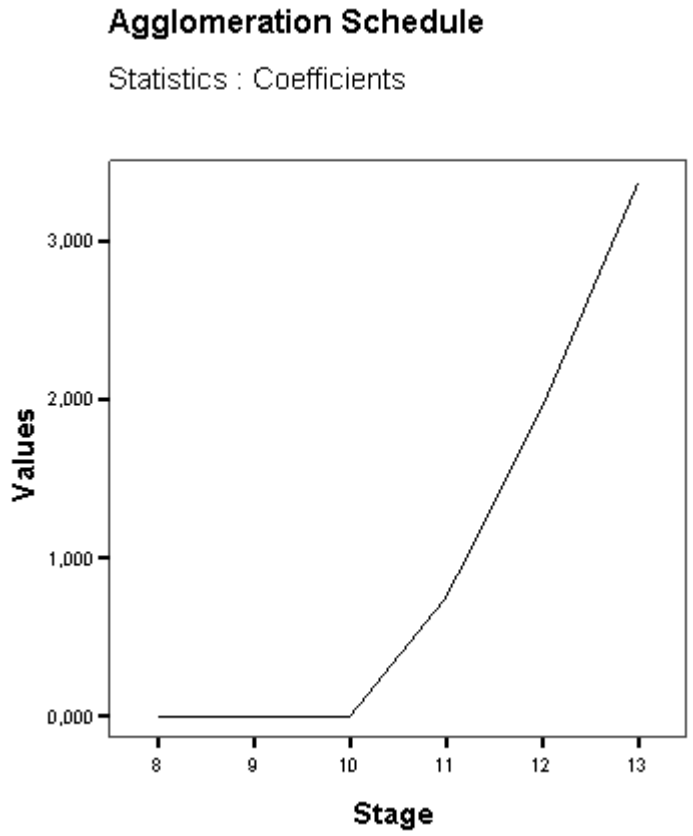
Agglomeration Schedule

Stage	Cluster Combined		Coefficients	Stage Cluster First Appears		Next Stage
	Cluster 1	Cluster 2		Cluster 1	Cluster 2	
1	1	15	,000	0	0	11
2	13	14	,000	0	0	3
3	3	13	,000	0	2	10
4	5	10	,000	0	0	10
5	8	9	,000	0	0	6
6	7	8	,000	0	5	12
7	4	6	,000	0	0	8
8	2	4	,000	0	7	9
9	2	12	,000	8	0	11
10	3	5	,000	3	4	12
11	1	2	,750	1	9	13
12	3	7	1,950	10	6	13
13	1	3	3,367	11	12	0

35. táblázat



98. ábra

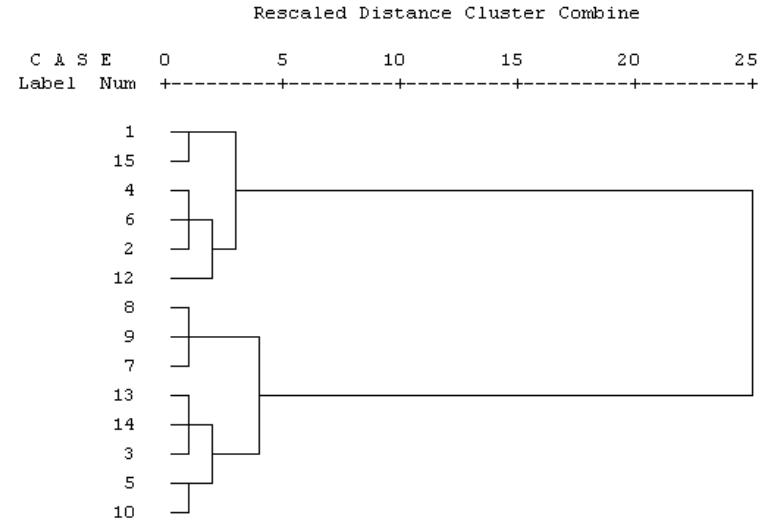

Agglomeration Schedule
Statistics : Coefficients
3,000
2,000
1,000
0,000
Values
8
9
10
11
12
13
Stage

99. ábra

Dendrogram

* * * * * * H I E R A R C H I C A L C L U S T E R A N A L Y S I S * * * * * *

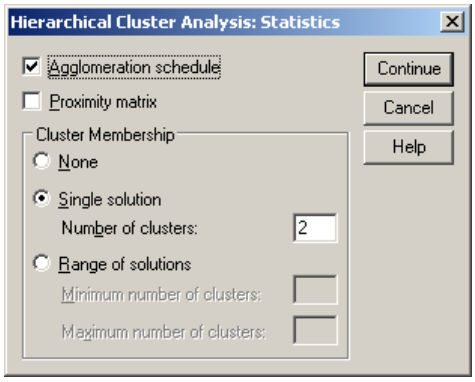
Dendrogram using Ward Method



100. ábra

A dendrogram (100. ábra) azt mutatja meg, hogy hány összevonás után hány klaszter maradt. A dendrogram alapján két klasztert célszerű létrehozni. Mentsük el a kétklaszteres javaslatot. Ehhez vissza kell térnünk az Analyze/Classify/Hierarchical Cluster parancsohoz, és ott válasszuk a Save gombot. A megjelenő ablakban a Single Solution (egyetlen megoldás) Number of clusters értékéhez írjunk kettőt (101. ábra). Amennyiben több klasztert sejtünk, akkor a Range of solutions menüpontot válasszuk, ahol a Minimum number of clusters (minimális klaszterszám) értékhez írjuk az általunk vélt legkisebb klaszterszámot, míg a Maximum number of clusters (maximális klaszterszám) értékhez a legnagyobb klaszterszámot.

A legnagyobb távolság a horizontális tengelyt tekintve 3 és 25 között fedezhető fel.



101. ábra

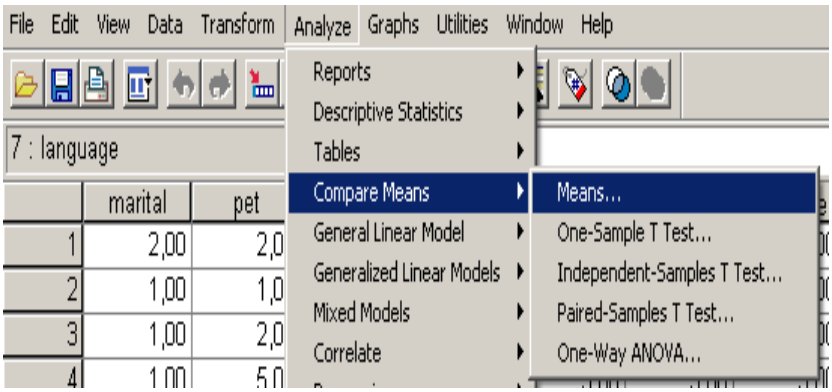
Az Output ablakban megjelenő alábbi ábra mutatja, hogy az egyes egyedek melyik klaszterbe esnek (36. táblázat).

Cluster Membership

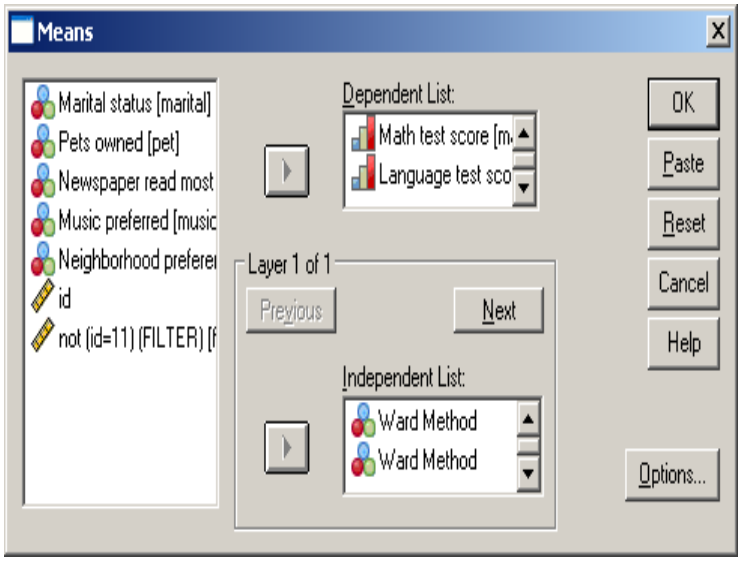
Case	2 Clusters
1	1
2	1
3	2
4	1
5	2
6	1
7	2
8	2
9	2
10	2
12	1
13	2
14	2
15	1

36. táblázat

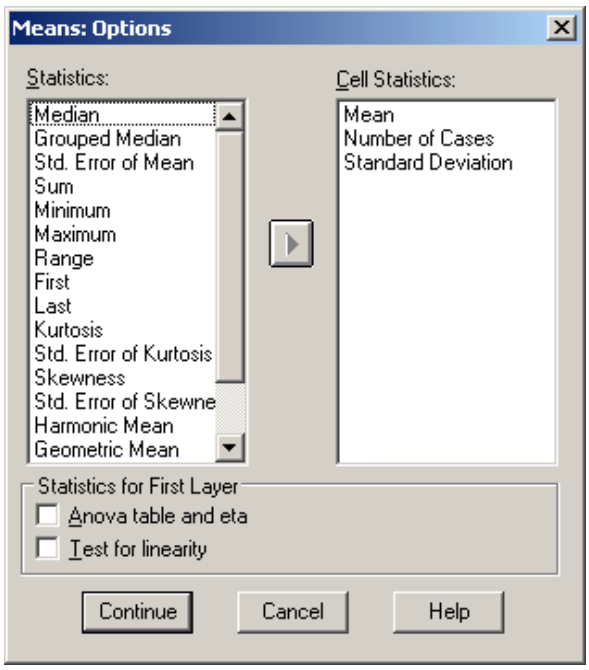
Az elemzést a klasztercentroidok (átlagok) alapján végezhetjük. Ehhez az átlag, elemszám és szórás értékeire lesz szükségünk. Az Analyze/Compare Means/Means parancsnál (102. ábra) a Dependent list-hez a Math and Language test score változókat, az Independent list-hez válasszuk a két Ward Methodot (103. ábra), majd az Options gombra kattintva keressük ki az átlag (mean), elemszám (number of cases), szórás (standard deviation) vizsgálatot (104. ábra).



102. ábra



103. ábra



104. ábra

A három klaszteres megoldás nem hozott megfelelő eredményt, mert a 3 klaszternél a szórás nagyon csekély (37. táblázat)

Math test score Language test score * Ward Method

Ward Method		Math test score	Language test score
1	Mean	2,8333	3,3333
	N	6	6
	Std. Deviation	,40825	,51640
2	Mean	1,0000	1,6000
	N	5	5
	Std. Deviation	,00000	,54772
3	Mean	2,0000	2,0000
	N	3	3
	Std. Deviation	,00000	,00000
Total	Mean	2,0000	2,4286
	N	14	14
	Std. Deviation	,87706	,93761

37. táblázat

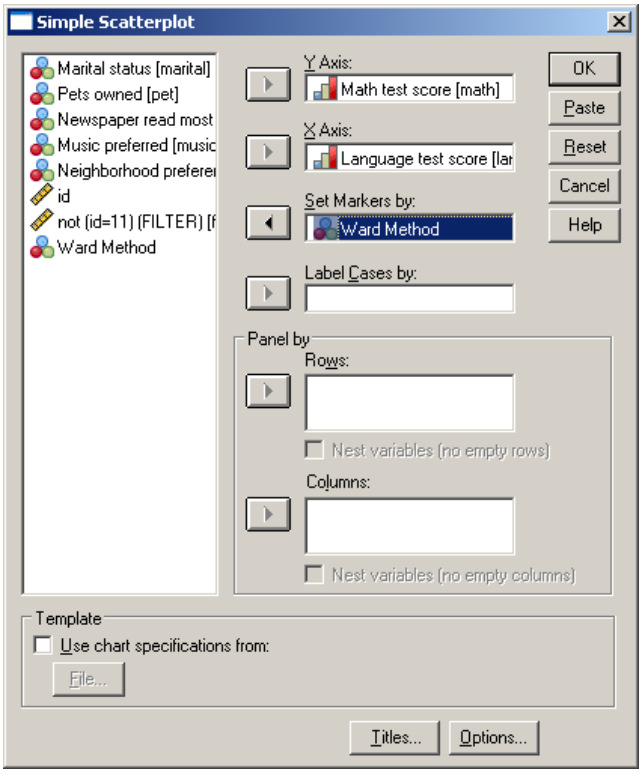
A két klaszteres megoldás jobb eredményeket hozott (38. táblázat).

Math test score Language test score * Ward Method

Ward Method		Math test score	Language test score
1	Mean	2,8333	3,3333
	N	6	6
	Std. Deviation	,40825	,51640
2	Mean	1,3750	1,7500
	N	8	8
	Std. Deviation	,51755	,46291
Total	Mean	2,0000	2,4286
	N	14	14
	Std. Deviation	,87706	,93761

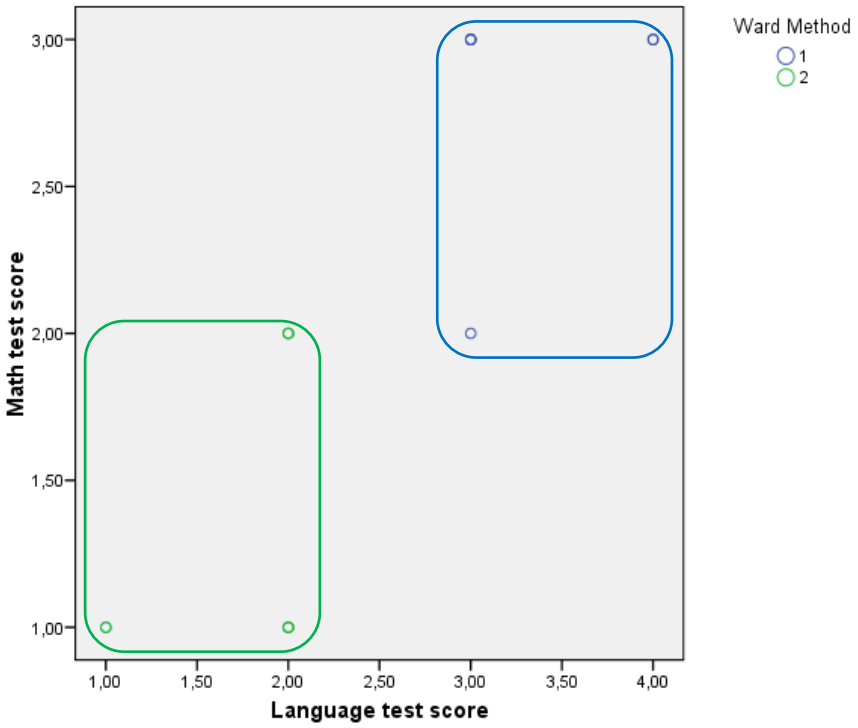
38. táblázat

A szórásdiagram segítségével érzékeltethetjük a két klasztert. Ehhez a Graphs/Legacy Dialogs/Scatter/Dot menüpontjában mozgassuk át a 2 klaszteres (Clu2_1) Ward Methodot (105. ábra).



105. ábra

Az ábrán kék és zöld alakzattal jelöltük a kialakult két klasztert (106. ábra). A két klasztert elnevezhetjük (például 1. klaszter: ügyes nyelv és matek tesztet írók, 2. klaszter: gyengébb nyelv és matek tesztet írók.)



106. ábra

8.3. Feladatok

1. Mi az klaszteranalízis lényege?
2. Milyen klaszterelemzési vizsgálati módszereket ismer?
3. Mi a hierarchikus és egyéb eljárások közti különbség?

9. Irodalomjegyzék

[1] Falus Iván, Ollé János: *Az empirikus kutatások gyakorlata. Adatfeldolgozás és statisztikai elemzés*, ISBN 978-963-19-6011-2, Nemzeti Tankönyvkiadó, Budapest, 2008.

[2] Sajtos László, Mitev Ariel: *SPSS kutatási és adatelemzési kézikönyv*, ISBN 978-963-9659-08-7, Alinea Kiadó, Budapest, 2007.

[3] *SPSS Base 15.0 User's Guide*, ISBN 978-0-13-613731-3, SPSS Inc. Chicago IL, 2006.

[4] Kassai Zsuzsanna: *Faktoranalízis SPSS alkalmazásával*. Szent István Egyetem, Gazdálkodás és Szervezéstudományi Doktori Iskola, Gödöllő, 2009. március 15. http://www2.szie.hu/tti/godolloi/kdi/aktual/tobbvalt_kassai_zsuzsanna.doc

[5] Mérési segédlet és útmutató az SPSS program használatához. BME Távközlési és Telematikai Tanszék, 2000. január. http://alpha.tmit.bme.hu/pub/meresek/3x/05/spss.rtf

[6] http://xenia.sote.hu/hu/biosci/docs/biometr/course/explore/statfv.html

[7] http://kompetenciameres.hu/OKM_szojegyzek.pdf