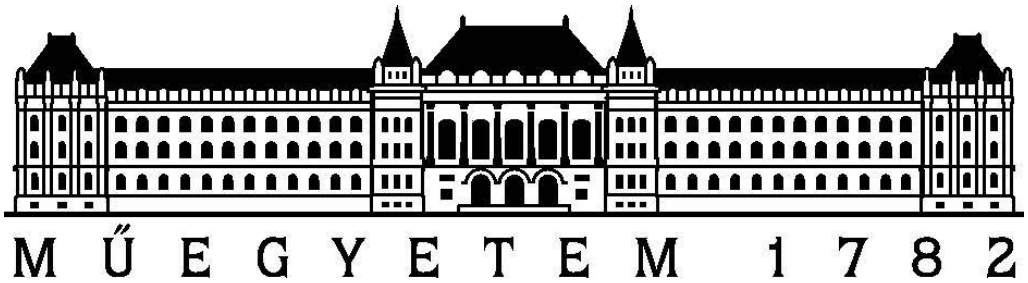


SAS Enterprise Guide

Készítette: Soltész Gábor

Tartalomjegyzék

Alapismeretek

Haladó ismeretek

Ábrajegyzék

Alapismeretek

1. Rövid ismertető

A SAS Enterprise Guide ötvözi a SAS szoftver világszinten elismert statisztikai és adatmanipulációs képességeit egy modern, Windows-os grafikus felhasználói felülettel. Segítségével a felhasználók könnyen és gyorsan végezhetnek statisztikai elemzéseket, adatösszesítéseket, leválogatásokat, az eredményt pedig látványos, jól áttekinthető formában publikálhatják.

Használata nem igényel programozói tapasztalatot. A szoftver lehetővé teszi, hogy bármilyen - a SAS rendszer által támogatott - adatot elérjünk, lefutassunk SAS procedúrákat, alkalmazásokat a SAS szerveren, az eredményeket pedig változatos jelentés, kimutatás, grafikon formátumban jelenítsük meg. Az eredmény automatikusan HTML-ben generálódik, weben publikálható, átemelhető MS Office alkalmazásokba, e-mailen terjeszthető, nyomtatható.

A SAS Enterprise Guide segítségével az elemzéseket megelőző adattisztítás és adatmanipuláció hatékonyan végezhető el. Az alkalmazás - amit a felhasználó többnyire egérkattintásokkal állít össze - elmenthető és egy későbbi időpontban ugyanaz a feladatsor végrehajtható. Az adatmanipulációs lehetőségek mellett egy nagyon komplex statisztikai eszköztár áll rendelkezésre. Az alapstatisztikák mellett klaszterelemzés, regresszió és számos egyéb elemző eszköz is elérhető a SAS Enterprise Guide-ban.

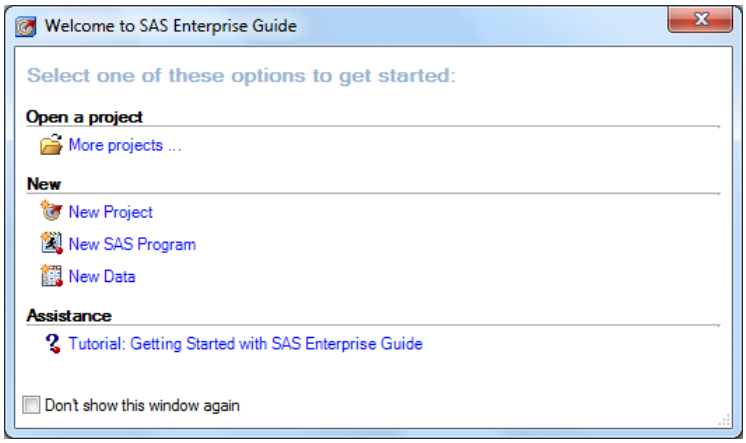
Funkcionalitás

- Adatelérés (SAS data set, PC fájl formátumok, MDDB-k)
- Adatkezelési feladatok (rangsorolás, standardizáció, formátum készítése, MDDB készítése és regisztrálása, véletlen mintavétel)
- Grafikus lekérdezés építõ (szűrés, rendezés, számított oszlopok képzése, táblák összekapcsolása)
- Leíró statisztikák (listázás, összesítő statisztikák, eloszlás elemzés, korreláció, egytényezős gyakoriság, összesítő táblák)
- Kereszttábla készítő
- ANOVA (t-teszt, egytényezős ANOVA, nem paraméteres egytényezős ANOVA, lineáris modellek)
- Regresszió (lineáris, nem-lineáris, logisztikus)
- Többváltozós elemzések (kanonikus korreláció, főkomponens-analízis, faktoranalízis, klaszteranalízis, diszkriminancia elemzés)
- Maradványérték analízis (élettáblák, arányos kockázatok)
- Alkalmasság vizsgálat (hisztogram, valószínűségi ábra, Q-Q diagram, P-P diagram, CDF diagram)
- Kontroll diagramok (átlag- és terjedelem-diagram, egyedi mérések ábrázolása, dobozábra, p diagram, np diagram, u diagram, c diagram)
- Pareto diagramok
- Idősorelemzés (idősoros adatok előkészítése, alap-előrejelzés, ARIMA modellezése és előrejelzése, regresszió autoregressziós hibával, paneladatok regresszióanalízise)
- Grafikus ábrázolás (oszlop-, kördiagram, terület-, vonal-, pont-, perec-, buborék-, felületdiagram, térkép, pénzügyi, vetület, radar)

2. A felhasználói felület bemutatása

A SAS Enterprise Guide röviden (SAS EG) indítását a START MENÜ / Minden program / SAS / Enterprise Guide 4.2 ikonnal tehetjük meg.

Indítsuk el a programot.

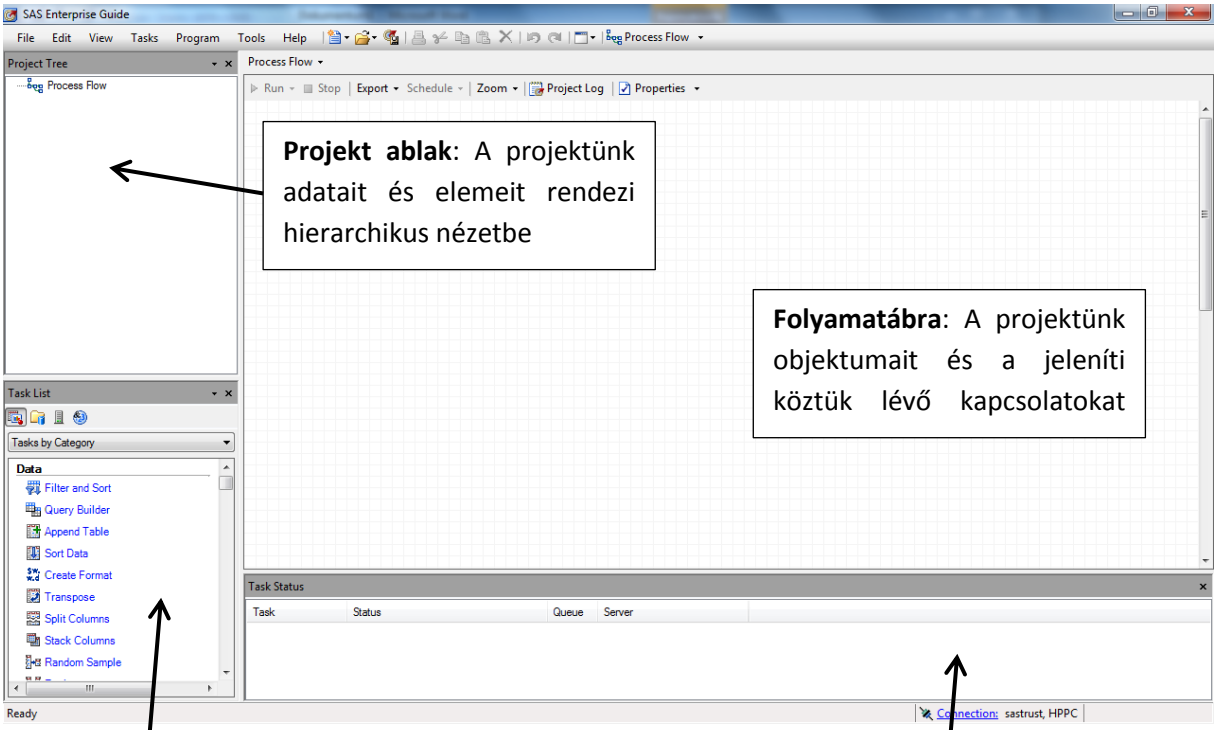


1. ábra Kezdő képernyő

Az 1. ábra mutatja, hogy első lépésben választhatunk, hogy egy már meglévő projektet vagy egy újat nyitunk meg. Továbbá lehetőségünk van új adattábla definiálására és SAS program írására is.

Válasszuk az Új projekt (New project) lehetőséget.

Ezek után megjelenik az alább látható kezelőfelület.



2. ábra Kezelőfelület

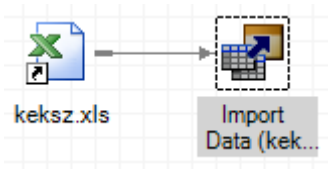
3. Adatok importálása és exportálása

A SAS EG lehetőséget biztosít, hogy más szoftver által készített adatállományokat olvassunk be és generáljunk belőle SAS kompatibilis adatállományt.

A következőnkben a gyakorlatokon már megismert kékszalag adatállományt (KEKSZ) fogjuk beimportálni. Ez az állomány jelenleg Excel típusú (.xls), azonban a további elemzések elvégzése végett szükséges ezeket SAS típusúba konvertálni. Ebben segít nekünk az importálás.

Válasszuk a File menü / Import data menüpontot.

Keressük meg a keksz adatállományunkat. Ha jól csináltuk, akkor a folyamatábra ablakban az alábbi folyamatot kapjuk.

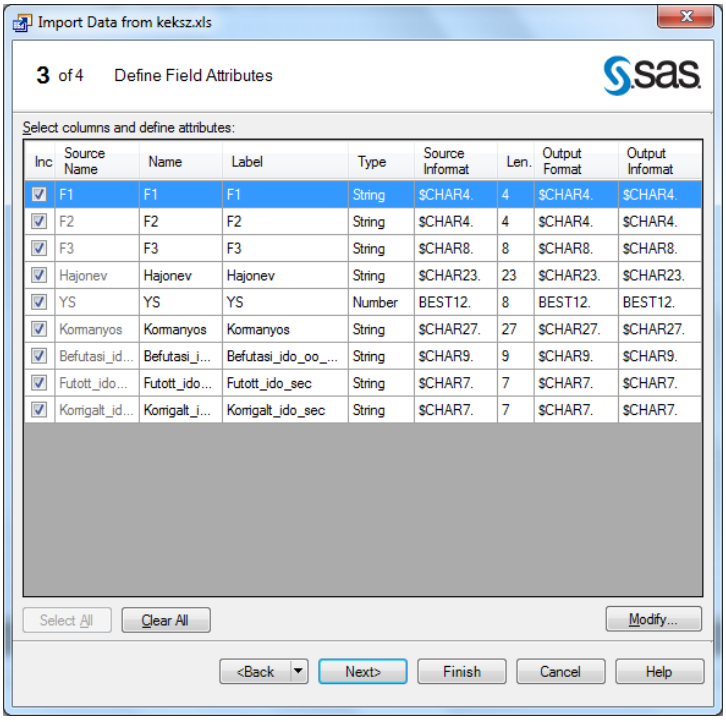


3. ábra Folyamatábra

A megjelenő ablakban klikkeljünk a Tovább gombra.

A következő ablakban kiválaszthatjuk, hogy melyik munkafüzetet (worksheet) kívánjuk beimportálni. Valamint itt van lehetőségünk megadni, hogy az adatállományunk első sorát másképp kezelje a SAS, ha az tartalmazza az oszlopok nevét. Ezek után válasszuk a Tovább gombot.

Ezt követően megjelennek az adatállomány oszlopai és azok jellemzői. Lehetőségünk nyílik, hogy kizárjuk bizonyos változókat illetve lehetőségünk van az egyes változók paraméterein változtatni.



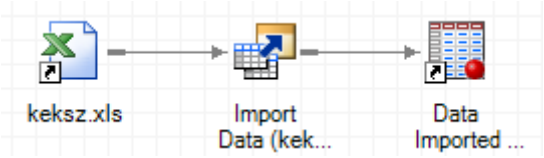
4. ábra Importálandó állomány szerkezete

Ne változtassunk a paramétereken, hanem így importáljuk be az adatállományt. (5. ábra)

	F1	F2	F3	Hajonev	YS	Kormanyos	Befutasi_ido_oo_pp_ss	Futott_ido_sec	Korrigalt_ido_s
1	1.	YS1	2002	Brokernet-Uniqua	63	Litkey Farkas	22.13.40	47 620	75 587
2	2.	YS1	92	Raffica	63	Király Zsolt	22.29.22	48 562	77 083
3	3.	YS1	91 GER	Telebox Due	63	Gerhard Müller	22.53.17	49 997	79 360
4	4.	YS1	115	Principessa	63	Rauschenberger...	22.53.36	50 016	79 390
5	5.	YS1	18	Liberty Sailing T...	67	Újhelyi Gáspár M...	00.36.27	56 187	83 861
6	6.	YS1	2	AC Sailing Team	63	Láng Róbert SVE	00.42.57	56 577	89 805
7	7.	YS1	1212	Mediacontact-Pi...	67	Soponyai Géza...	00.56.42	57 402	85 675
8	8.	YS1	1153	Sponzor Wanted	67	Vándor Róbert	00.58.44	57 524	85 857
9	9.	YS1	42	Raiffeisen	71	Pfeninberger An...	01.07.27	58 047	81 756
10	10.	70	6	Anna	78	Csoregh Zoltán	01.27.35	59 255	75 968
11	11.	70	1	Orpheus	78	Herkó Dezsö	01.28.14	59 294	76 018
12	12.	Ö	75/2	Sirocco	81	Cittel Lajos	01.43.32	60 212	74 336
13	13.	70	7	Irókéz	78	Litkey Bence KMP	01.49.22	60 562	77 644
14	14.	YS1	1005	Kacor	81	Gosztonyi András	01.57.03	61 023	75 337
15	15.	70	2	Capella	78	Pomucz Tamás	01.57.49	61 069	78 294
16	16.	99	134	BMW Sailing Te...	79	Vadnai Péter	02.05.24	61 524	77 878
17	17.	99	111	FMC Consulting	79	Paksi Álmos	02.27.02	62 822	79 522
18	18.	YS1	1	Barracuda	85	Bakos Tamás	02.34.09	63 249	74 411
19	19.	99	139	Racenace Grid	79	dr Szutrály Gera...	02.35.09	63 209	80 138

5. ábra Importált adatállomány

Az importálás folyamatábrája:



A folyamatábra jól szemlélteti, hogy Excel típusú adatállományból indultunk ki és végeredményben SAS adatállományt kaptunk.

Ha jobb egérgombbal klikkelünk az új SAS adatállományunkra (Data Imported…) és kiválasztjuk a Properties menüpontot, akkor láthatjuk, hogy az adatállomány a WORK könyvtárban található KEKSZ néven. Korábbi SAS programozási ismereteinkre támaszkodva tudhatjuk, hogy ez a könyvtár csak ideiglenes állományok tárolására alkalmas, hiszen ha a SAS-t, jelen esetben a SAS EG, bezárjuk, akkor ennek a könyvtárnak a tartalma elveszik.

Így ha a jövőben még használni akarjuk a KEKSZ adatállományt anélkül, hogy ismételten el kellene végezni az importálást, célszerű az adatállományt egy saját könyvtárba kiexportálni.

Exportáláshoz klikkeljünk az adatállományra, majd a menüből válasszuk az Exportálás menüpontot. (Export Imported Data from keksz.xls…)

A megjelenő ablakban válasszuk a Server-t, majd a SASMeta szervert. Itt lehetőségünk van a Libraries mappát választva egy korábban definiált SAS könyvtárba elmenteni az adatállományunkat, vagy a Files mappát választva a Description-ben megadott elérési útvonalon található könyvtárba exportálni a fájlt.

Exportáljuk most a SAS alapértelmezett (Files) könyvtárába.

Nevezzük el KEKSZ névvel az adatállományt, aminek a típusa legyen sas7bdat.

Végezetül ellenőrizzük, hogy a fájl megtalálható e az adott könyvtárban.

dmdata	2011.06.21. 23:34	SAS Catalog	33 KB
keksz	2011.07.05. 10:27	SAS Data Set	73 KB
parms	2011.06.21. 23:39	SAS Catalog	17 KB

4. Adatok beolvasása

Ez a pont alapvető fontosságú az adatelemzés szempontjából, hiszen az adatok általában rendelkezésre állnak, sőt gyakran már SAS formátumban, így annyi a dolgunk, hogy beolvassuk őket a további adatmanipulációs és elemzési feladatok elvégzéséhez.

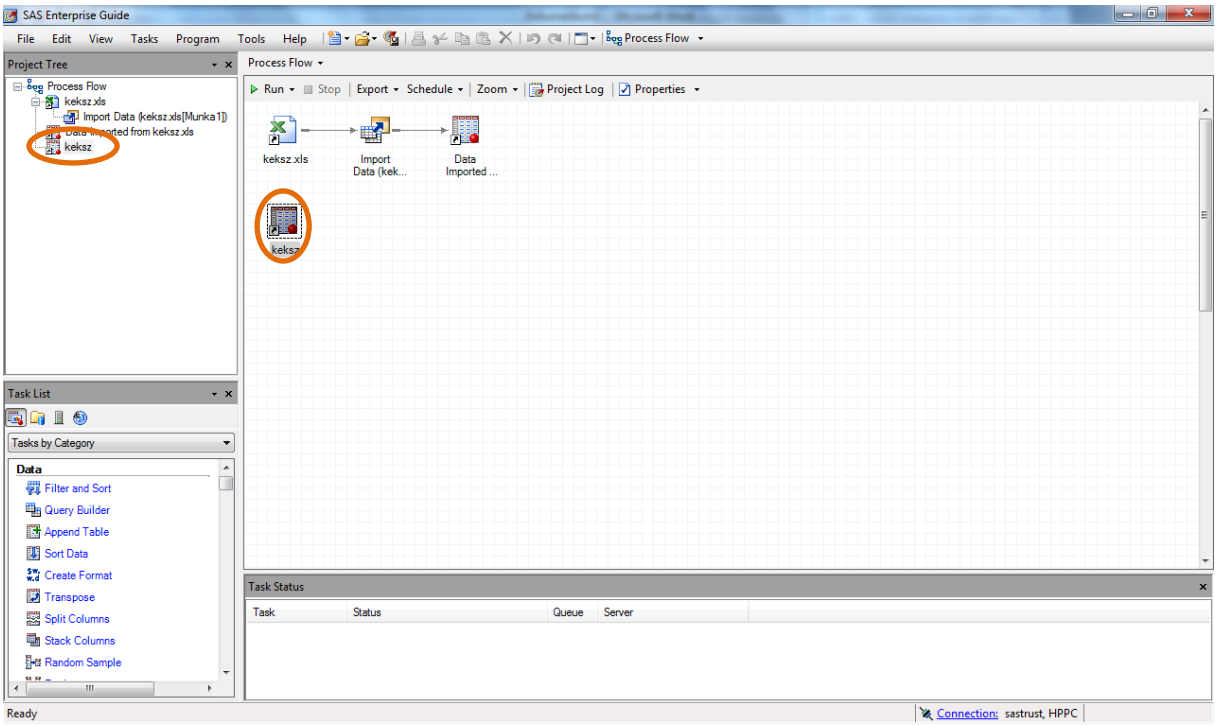
Adat beolvasásához klikkeljük jobb egérgombbal a Folyamatábra munkaterületre.

Válasszuk az Open menü Data menüpontját.

Adjuk meg az előző lépésben kiexportált KEKSZ adatállományt a beolvasandó állomány forrásaként.

Olvassuk be a fájlt.

A beolvasás eredményeként megjelenik az adatállomány tartalma. Valamint a Projekt Explorer ablakban és a Folyamatábra munkaterületet is megjelenik a beolvasott adatállomány.



6. ábra Beolvasás eredménye

4.1 Olvasási mód feloldása

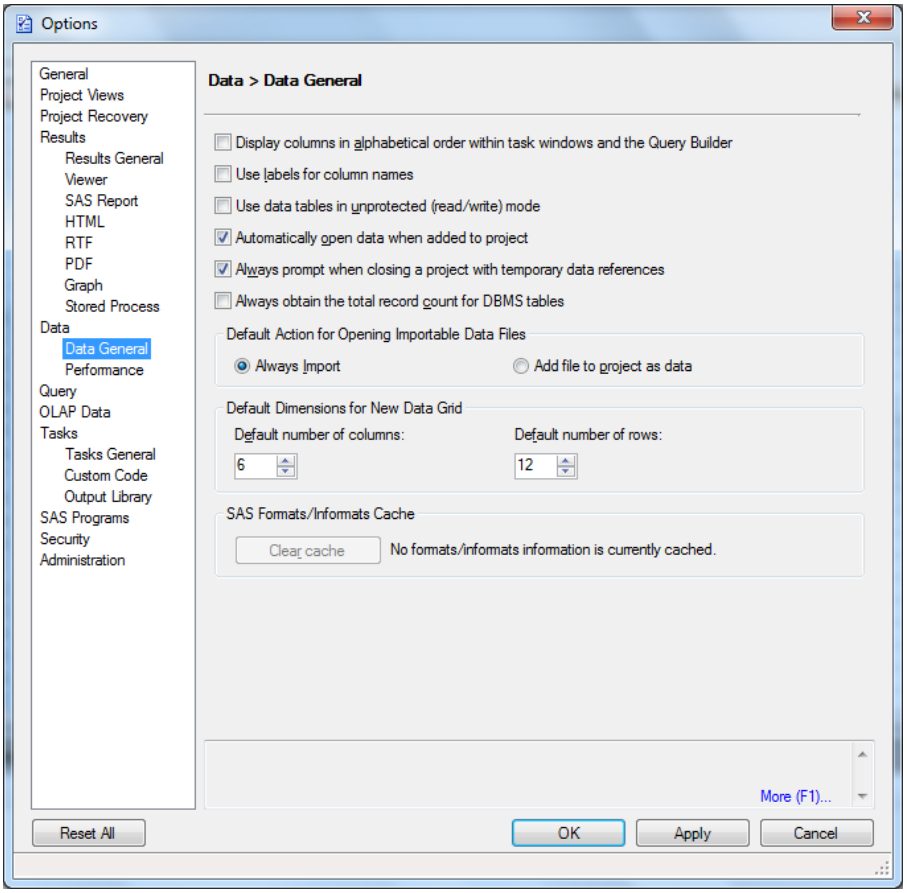
Fontos tulajdonsága a SAS EG-nek, hogy az adatok biztonsága érdekében azokat olvasási módban engedi megnyitni. Ez annyit jelent, hogy alapértelmezésben a felhasználó nem módosíthatja az adatállományok tartalmát.

Ha a korlátozást nem akarjuk globálisan feloldani, hanem csak az adott táblánk írásvédettségét akarjuk megszüntetni, akkor válasszuk az Edit menü Protect data menüpontját. Ha a menüpont előtt pipa található, akkor csak olvasni tudjuk a tartalmat. A jogosultságok bővítéséhez klikkeljünk a menüpontra. Ezek után a SAS figyelmeztet minket, hogy valóban fel akarjuk e oldani a korlátozást.

Ha olyan adatállományt nyitunk, meg amihez csak a felhasználók egy bizonyos körének van írási joga és mi nem tartozunk ezek közé, akkor ezzel a ponttal nem tudjuk feloldani a korlátozást. Ekkor

célszerű a tábláról egy másolatot készíteni a Query Builder segítségével, amiről a későbbiekben még részletesen olvashatunk.

A golbális korlátozás megszüntetéséhez válasszuk a Tools menü Options menüpontját. (7. ábra) A bal oldali menüben navigáljunk a Data Generál pontra. A korlátozás megszüntetéséhez pipáljuk ki a „Use data tables in unprotected (read/write) mode” utasítást.

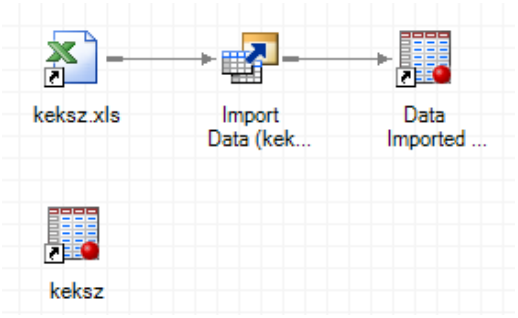


7. ábra Tools / Options menüpont

Megjegyzés: Ha egy adatállományon nincs jogosultságunk módosítani, akkor hiába állítjuk át globálisan, ez nem fogja befolyásolni a felhasználást, azaz a továbbiakban sem leszünk képesek módosítani a tartalmat.

5. Adatmódosítás

Jelen pillanatban a folyamatábra munkaterületén az alábbi objektumok és köztük lévő kapcsolatok láthatók.

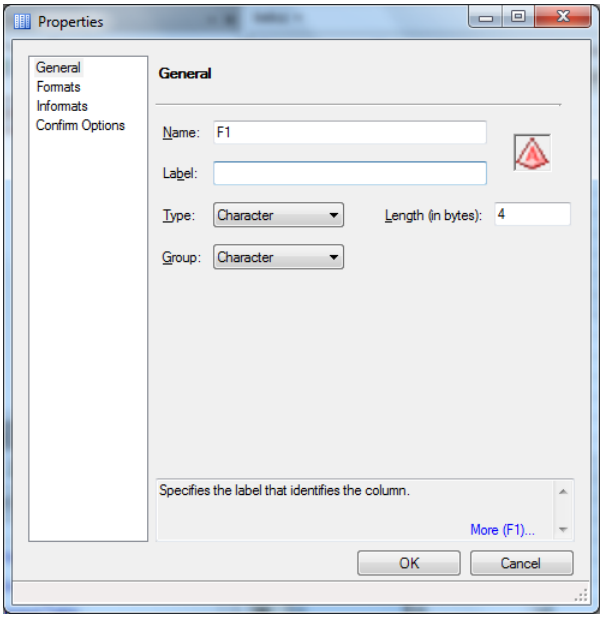


8. ábra Folyamatábrák

Nyissuk meg a KEKSZ adatállományt. Ha az adatállomány írásvédett, akkor szüntessük meg a korlátozást. Ezek után címkézzük fel az egyes változókat.

A megnyitott KEKSZ adatállományban válasszuk az F1 nevű oszlopot.

Klikkeljünk jobb egérgombbal az adott oszlop nevére és a menüből válasszuk a Properties menüpontot. (9. ábra)



9. ábra Tulajdonságok menüpont

A Label mezőbe adjuk meg a kívánt címke szövegét. Esetünkben legyen ID. Lehetőségünk van módosítani a mező típusán és méretén. Azonban arra figyelni, kell, hogy ha például a mező hosszát 2 hosszúra állítom, akkor azzal hibát generálok, hiszen a 110-119 közötti elemek azonosítója ebben az esetben 11 lesz, azaz a változó elveszti egyediségét.

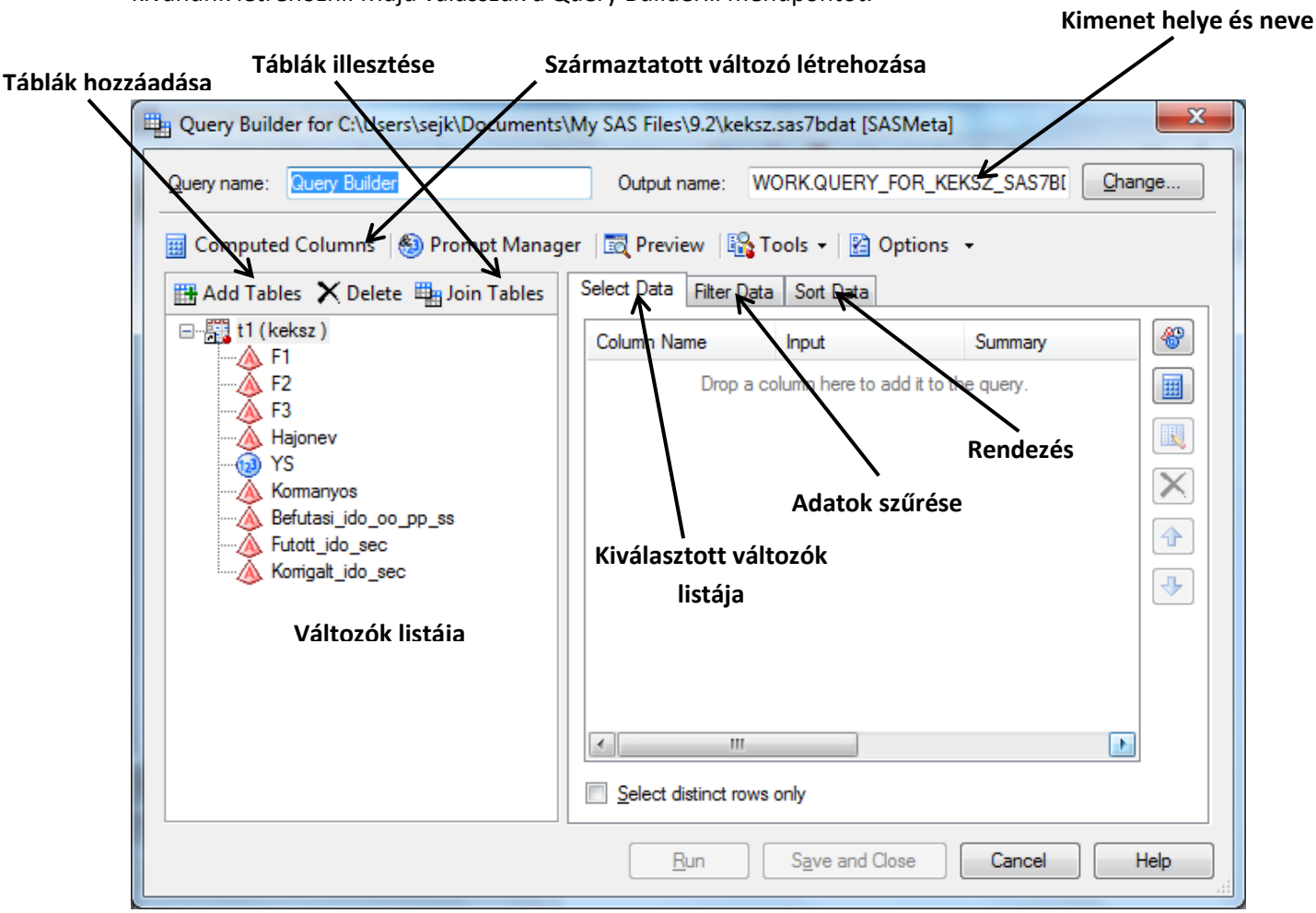
Ennél súlyosabb hiba, hogy ebben az esetben az adatállomány tartalma felülíródik és a korábbi tartalom így már nem állítható vissza. Esetünkben csupán az exportálást kell ismételten elvégezni a korábbi tartalom visszaállításához, azonban egy rendelkezésünkre bocsátott adatállomány esetén

nem biztos, hogy létezik az adott állományról másolat. Így azt javasoljuk, hogy az elemzést, mindig egy másolt adatállományon végezzük ezzel is kizárva az adatvesztés lehetőségét. A másolat készítéséről a következő pontban olvashatunk.

6. Másolat készítése egy adatállományról

A korábbiakban már többször felhívtuk a figyelmet, hogy ajánlatos másolatot készítenünk az elemzendő adatállományainkról. Ezt a Query Builder modul segítségével végezhetjük el a legkönnyebben. Mint ahogy a neve is mutatja ez a modul alkalmas lekérdezések készítéséhez. Amikor másolatot kívánunk készíteni, akkor az eredeti adatállomány összes változóját bevesszük a lekérdezésbe és nem alkalmazunk semmilyen szűrési megszorítást.

Másolat készítéséhez klikkeljünk jobb egérgombbal arra az adatállományra, amiről másolatot kívánunk létrehozni. Majd válasszuk a Query Builder… menüpontot.



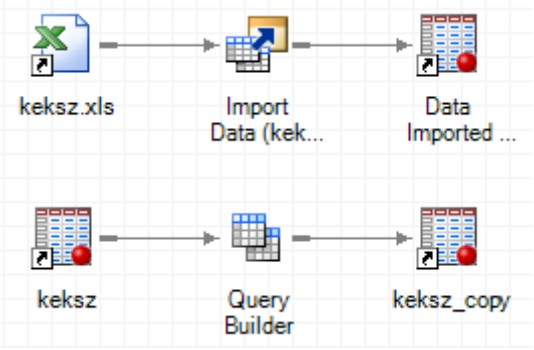
10. ábra Query Builder

Jelöljük ki a bal oldali ablakban található változókat és húzzuk át ezeket a jobb oldali ablakba (Selected Data). A keletkező állomány neve legyen KEKSZ_COPY, ha helye pedig a WORK könyvtár.

Output name: WORK.KEKSZ_COPY

Majd futtassuk a modult. (RUN)

A futás után a folyamatábránk munkaterülete az alábbiak szerint néz ki.



Ezzel biztosítottuk, hogy adatvesztés esetén is legfeljebb a KEKSZ_COPY adatállomány tartalma íródik felül. Így a visszaállításhoz elegendő a Query Builder csomópontot futtatnunk.

7. Új változó létrehozása egy adatállományban

Gyakran elengedhetetlen, hogy egy adatállományban ne hozzunk létre új változót. Így ebben a pontban bemutatjuk, hogyan lehet egy adatállományhoz új változót hozzáadni.

Nyissuk meg az előző pontban létrehozott KEKSZ_COPY adatállományunkat. Gondoskodjuk arról, hogy az adatállomány ne legyen írásvédett.

Korábbi tanulmányainkból tudhatjuk, hogy a változók típusa két nagy csoportba osztható. Vannak karakterek és numerikusváltozók. A SAS EG az adatállomány megnyitásakor kis piktogrammal mutatja az egyes változók típusát.

- Mennyiségi ismérveket (numerikus típusú változók) a SAS EG a következő ábrával szimbolizálja.  (A numerikusváltozók értékei egyébként jobbra rendezettek.)
- Míg karakter típusú változók esetén  ábra használatos.

Megvizsgálva az adatállományunkat megállapítható, hogy csak a YS (yardstick) szám az egyetlen numerikus típusú változó.

Tovább vizsgálódva észrevehetjük, hogy a Futott_ido_sec és a Korrigalt_ido_sec karakter típusú, ami a további elemzésekben problémát okozhat. Így első lépésben átkonvertáljuk ezeket a változókat.

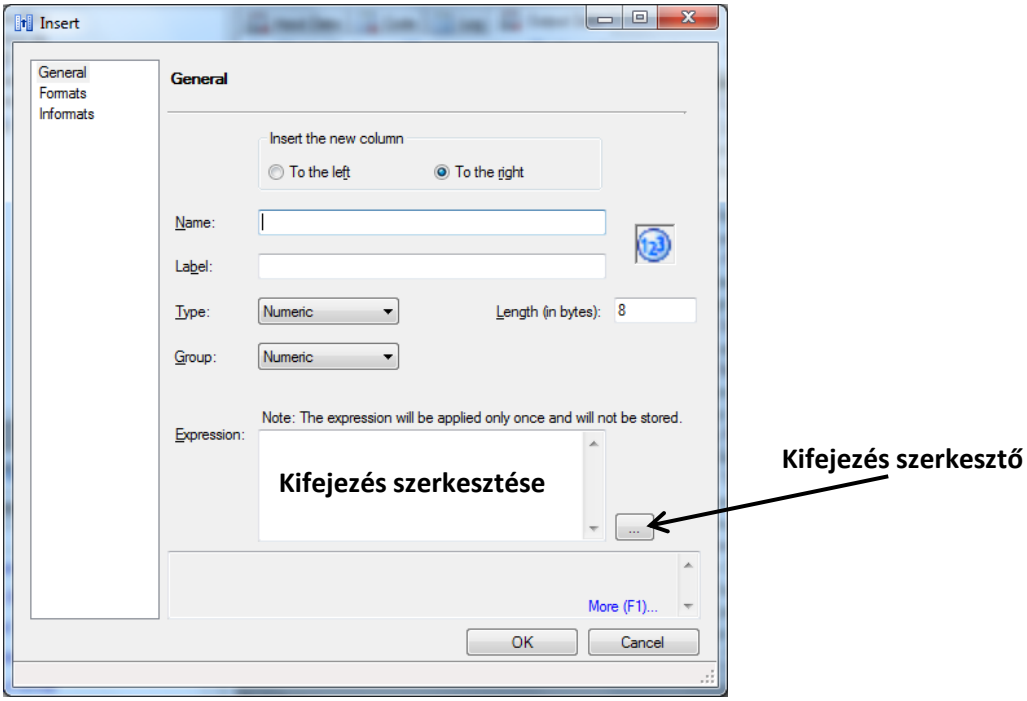
Futott_ido_sec	Korrigalt_ido_sec
47 620	75 587
48 562	77 083
49 997	79 360
50 016	79 390
56 187	83 861
56 577	89 805

11. ábra Karakter típusú változók

Nyissuk meg a KEKSZ_COPY adatállományt.

Jelöljük ki a Futott_ido_sec oszlopot az oszlopfejlécbe kattintással. (Jobb klikk)

A menüből válasszuk az Insert Column… menüpontot.



12. ábra Új változó létrehozása

Az új változó neve legyen: Futott_ido_sec_num, típusa numerikus és 8 hosszú. Válasszuk a kifejezés szerkesztőt.

A kifejezés szerkesztő bal oldalán található két mappa. Az első a SAS függvényeket tartalmazza. A második pedig az elérhető táblák változóit – esetünkben csak a KEKSZ_COPY változóit. A SAS függvények bemutatása nem célja ennek a dokumentumnak. Az alapvető függvények leírása megtalálható a SAS BASE programozás című dokumentumban. Valamint a SAS minden függvényhez ad ismertetőt a jobb oldali ablakban.

Térjünk vissza a konkrét feladatunkhoz. A 11. ábra jól látható, hogy a karakteres változó nem csak a karaktersorozat végén, hanem közben is tartalmaz szóközöket. Így olyan függvényt kell alkalmazni, ami egy adott karakter összes előfordulását eltávolítja egy adott sztringben.

A COMPRESS függvény alkalmas ennek megvalósítására.

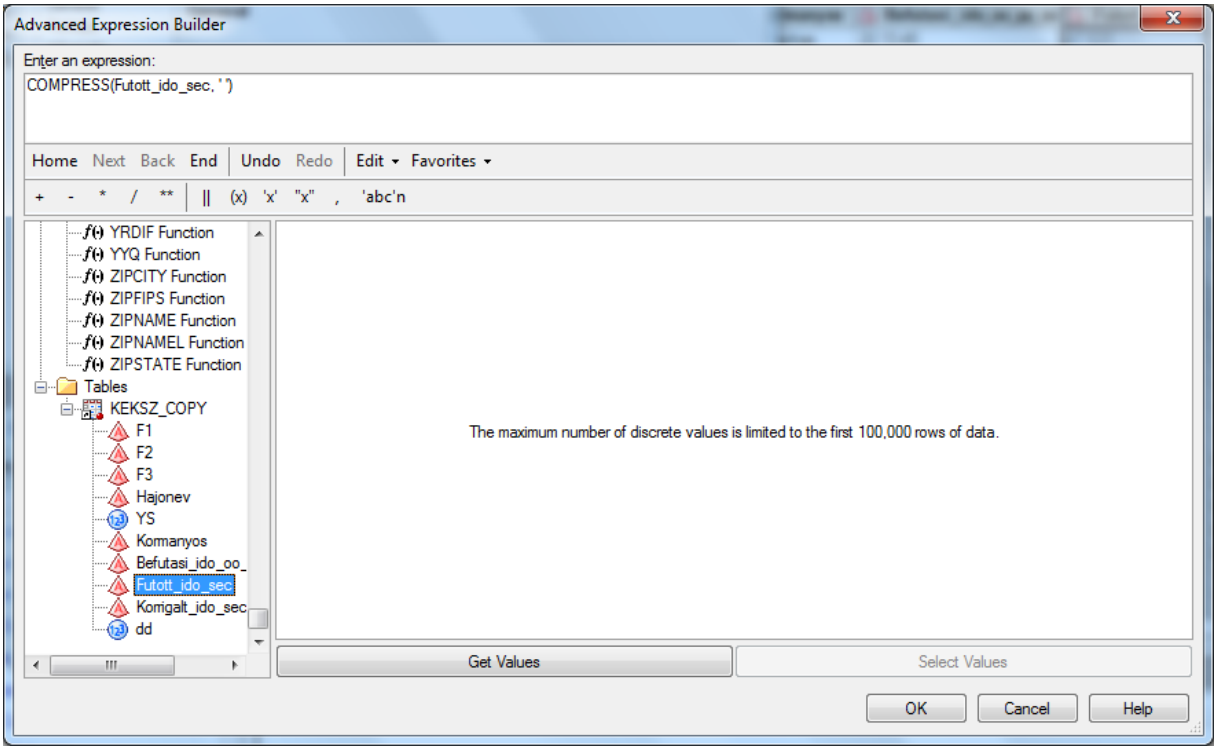
A függvény szintaxisa:

COMPRESS (<forrás változó>,<eltávolítani kívánt változó>)

A Function mappából válasszuk ki a COMPRESS függvényt, ekkor az ablak felső részén megjelenik a függvény és az esetleges paraméterei. Az első paraméterhez válasszuk a bal oldali ablakban a Tables mappa KEKSZ_COPY adatállományának Futott_ido_sec változóját. Ennek kiválasztása után a SAS az

első paramétert helyettesíti a kívánt változóval. A második változónak adjuk meg a szóköz karakter aposztrófok között a következő módon: ’ ’

A végeredményt a 13. ábra mutatja. Válasszuk az OK gombot.



13. ábra Kifejezés szerkesztése

A létrehozott változót a 14. ábra mutatja.

ss	Futott_ido_sec	Futott_ido_sec_num	Korrigalt_ido_sec
	47 620	47620	75 587
	48 562	48562	77 083
	49 997	49997	79 360
	50 016	50016	79 390
	56 187	56187	83 861
	56 577	56577	89 805
	57 402	57402	85 675
	57 524	57524	85 857
	58 047	58047	81 756

14. ábra Létrehozott változó

Alakítsuk át a Korrigált_ido_sec változót is numerikus típusú változóvá a fent ismertetett módon.

Az utolsó pontot feltétlenül végezzük el, mert a továbbiakban szükség lesz a numerikus típusú adatokra.

8. Változók és megfigyelések szűrése és rendezése

Az előző pontban létrehoztunk két numerikus típusú változót. Ha megnyitjuk az adatállományunkat, akkor láthatjuk, hogy a numerikus típusúakon kívül a karakteres változók is rendelkezésre állnak. Azonban ezek a karakteres változók feleslegesek, hiszen az elemzésekben ezeknek a numerikus változatát fogjuk felhasználni.

Így tehát a jobb áttekinthetőség és tárterület felhasználás optimálása végett célszerű ezeket a változókat eltávolítani. Ehhez a - Másolat készítése egy adatállományról - című pontban már bemutatott Query Buildert fogjuk alkalmazni.

Válasszuk ki a KEKSZ_COPY adatállományunkat, majd a felső menüsorból válasszuk a Task menüpontot. A legördülő listából a DATA menüpontján belül találjuk a Query Builder… csomópontot.

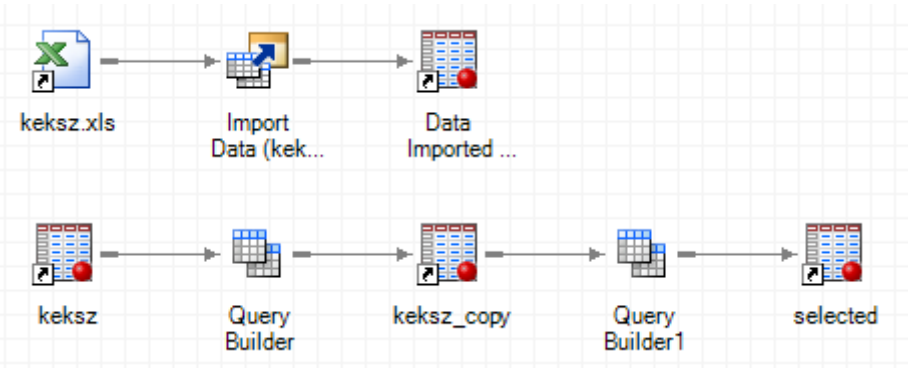
Ezzel a csomóponttal a változók számát tudjuk módosítani, azaz egy adatállomány oszlopainak a számát.

Nyissuk meg a Query Builder-t. (10. ábra) A jobb oldali ablakban válasszuk ki a kívánt változókat, esetünkben a Futott_ido_sec és a Korrigalt_ido_sec változókon kívül mindet.

Nevezzük el a kimenetet „Selected” néven. (10. ábra)

Ezek után futtassuk a csomópontot (RUN).

A folyamatábránk az alábbiak alapján néz ki.

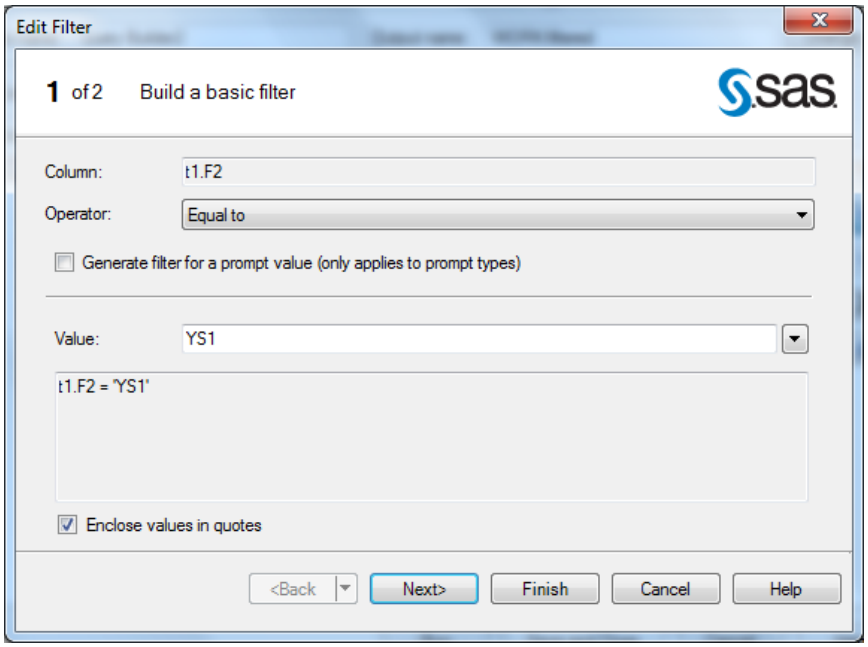


8.1 Megfigyelések szűrése

A Query Builder alkalmas arra is, hogy az adatállományainkat megszűrjük bizonyos feltételek alapján. Tételezzük, fel, hogy csak a YS1 hajóosztályba tartozó hajók neveire és futott idejükre vagyunk kíváncsiak. A hajóosztályokat az F2 változó tartalmazza.

A szűrés elvégzéséhez nyissuk meg a Query Builder csomópontot és válasszuk ki a Hajonev és a Futott_ido_sec_num változókat, hiszen a lekérdezésben csak ezekre a változókra leszünk kíváncsinak. Ha így futtatnánk a csomópontunkat, akkor nem a megfelelő eredményt kapnánk, hiszen csak az oszlopok számát szűrtük ezzel a megfigyelések száma változatlan.

A megfigyelések szűréséhez válasszuk a Filter Data fület (10. ábra) és adjuk meg a szűrési feltételt.

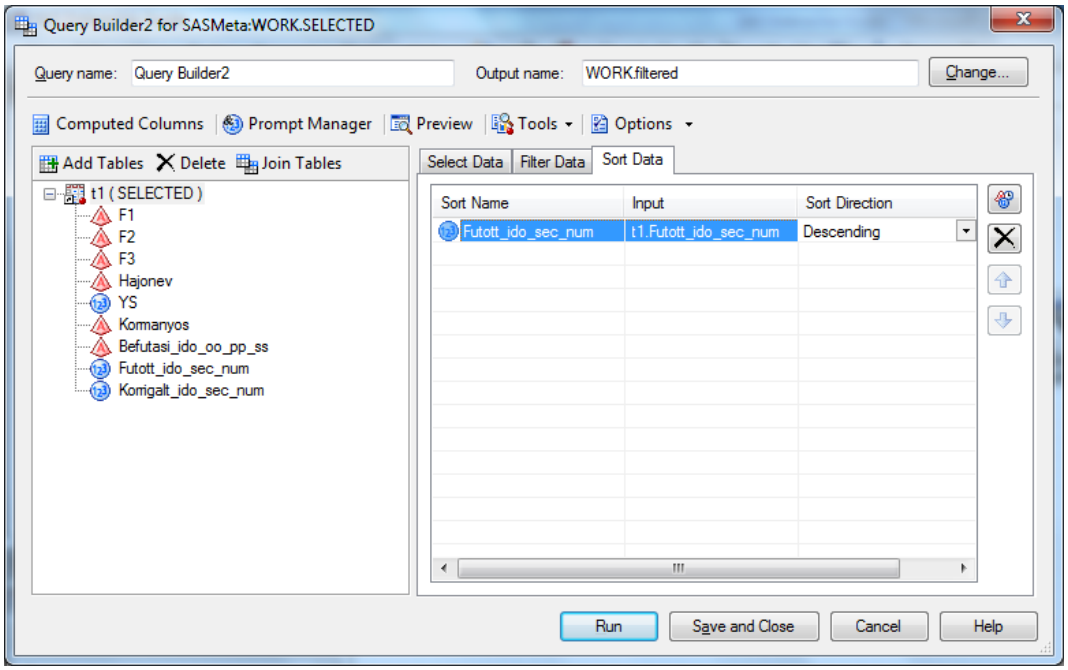


15. ábra Szűrési feltétel

Válasszuk ki az F2 változót és értéknek adjuk meg a YS1-t (15. ábra).

8.2 Megfigyelések rendezése

Ha az eredményeinket rendezni is szeretnénk, akkor válasszuk a Query Builder (10. ábra) Sort Data fülét. Nyilván csak olyan változóra van értelme szűrést megadnunk, amit a kimeneti adatállomány is tartalmaz, így válasszuk ki a Futott_ido_sec_num változót és állítsunk be csökkenő sorrendet. (16. ábra)



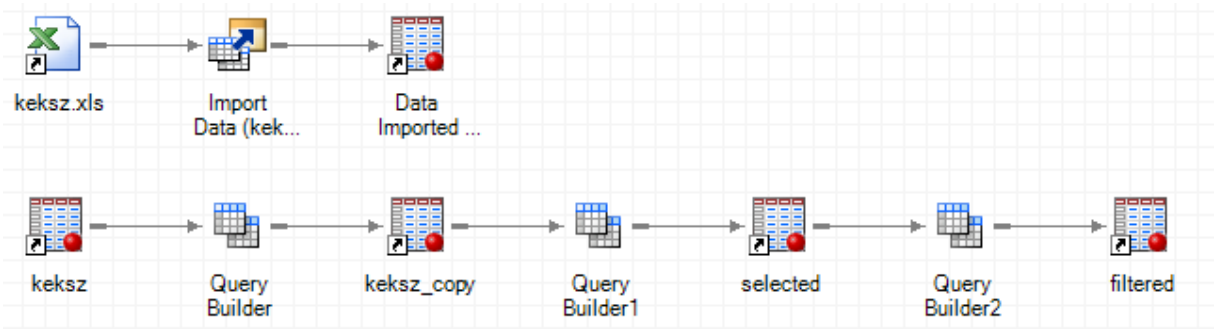
16. ábra Rendezési feltétel

Végezetül futtassuk a Query Builder csomópontunkat, miután a kimenetet elneveztük Filtered névvel. A kimenet eredménye alább látható.

	Hajonev	Futott_ido_sec_num
1	Sun Odyssey	126768
2	Seppuku	113997
3	A hajó	112959
4	Barbi	110191
5	Hudahupa	108987
6	Blue Star II.	106360
7	Royal Flash	105180
8	Hableány	103830
9	Virus	103239
10	Adagio	103226
11	Beerluck	101415
12	Lakinet.hu	99618
13	Mérges Rája	99372

17. ábra A szűrés és rendezés utáni eredmény

A folyamatábra pedig a következő szerint módosult.



18. ábra Szűrés és rendezés folyamatábrája

Haladó ismeretek

9. Több adatállomány összekapcsolása

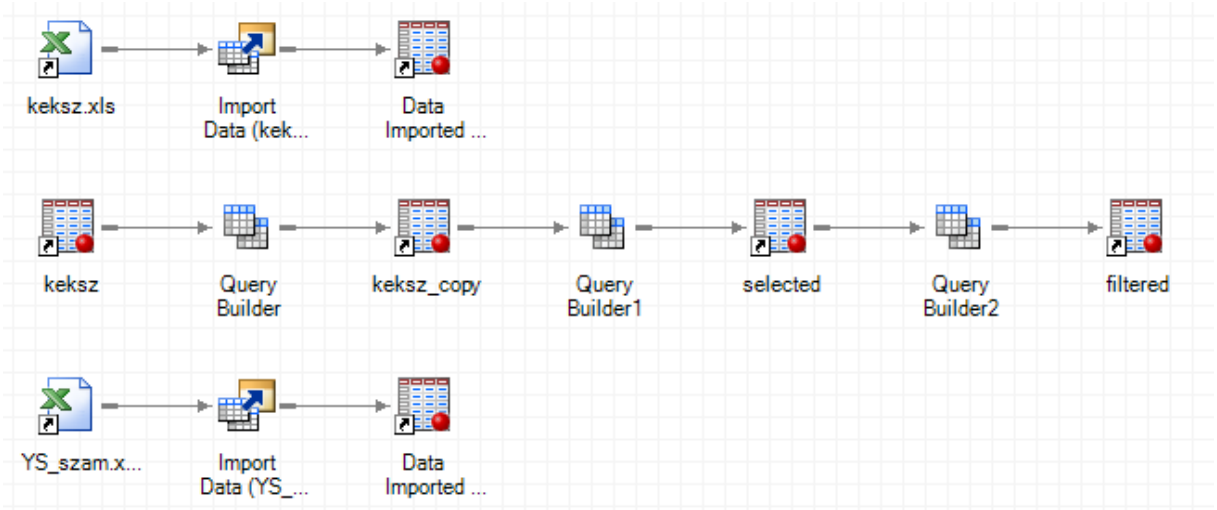
A SAS programozásnál már megismert YS adatállományt fogjuk összekapcsolni a KEKSZ adatállományunkkal. A YS állományt azonban az első lépésben be kell importálni, mivel Excel típusú fájl áll rendelkezésünkre.

Az importálás menetét nem részletezzük ismételten. Menete megegyezik a fent leírtakkal.

Az importálás után az alábbi adatállományt és folyamatábrát kaptuk.

	Osztaly	Rajtszam	Hajonev	Tipus	DSV_Ys	YS2010	YS2011	Megjegyzes	Grosz_Genua_Spi
1	22			22-es Schaerenk...	99	103	103		
2	30			30-as Schaerenk...	95	93	93		
3	33	1	BASA	Enter 36	94	94	94		37/38/130 7/8/Gen
4	33	3	PANDORA	Dehler 34 7/8	INFO	99	99		
5	33	5	CIMBORA	Enter 36	94	94	94		
6	33	14	BACCARA	Top 34 regatta	-	92	92		
7	33	32	FÜKE 32	Füke 32	-	99	99		
8	33	32	DEHLER-32	Dehler 32 JV, tav...			95	ideiglenes	
9	33	59	N.J.L	Elan 340 módosí...	(1,9 m ÖSV GPH...	84	84	karbon árboc+te...	
10	33	64	ENDORPHIN	Elan 333cs	(1,9 m ÖSV GPH...	96	96		
11	33	75	MADICKEN	Dehler DB1	97	89	89	átépített	30/32,8125 Top
12	33	80	FUJI	Elan 333cs+	95	94	94		

19. ábra YS adatállomány



20. ábra Folyamatábra az importálás után

Jelenleg rendelkezésünkre áll két adatállomány. A korábbiakban megismert KEKSZ adatállomány valamint ez az új, ami számos egyéb információt tartalmaz az egyes hajókról. A feladat, hogy összekapcsoljuk a két állományt.

Az összekapcsolás feltétele, hogy mind a két állomány tartalmazzon azonos típusú és azonos halmazból tartalmazó megfigyeléseket. A legfontosabb feltétel pedig az, hogy a kapcsolást biztosító változók értékei egyediek legyenek, hiszen esetünkben egy-egy irányú megfeleltetés áll fent. Létezik egy a többhöz (például egy személyhez tartozhat több autó és több telefonszám is) valamint több a többhöz kapcsolat is.

Első lépésben tehát tisztázni kell, hogy mi legyen a kulcsváltozó, azaz melyik változó szerint kapcsoljuk össze a két állományt. Ha azzal a feltételezéssel élünk, hogy mind a két állományban a hajónevek egyediek és egy hajónév csak egy hajót azonosít, akkor ez megfelelő kulcsváltozónak.

Vizsgáljuk meg közelebbről is ezt a változót. Látható, hogy a KEKSZ és a YS adatállományban a hajónevek nem azonos formátumúak. A YS állományban csupa nagybetűvel vannak írva a nevek. Így első lépésben ezeket kell átalakítani.

Most a változók átkódolásának egy egyszerűbb és gyorsabb változatát alkalmazzuk a már jól ismert Query Builder csomópont segítségével.

Jelöljük ki az importált (YS) állományt, majd kapcsoljuk hozzá egy Query Builder csomópontot.

Válasszuk a Származtatott változó létrehozása gombot. (10. ábra)

A megjelenő ablakban válasszuk az Új (New) gombot, hiszen egy új változót kívánunk létrehozni, majd az új ablakban az összetett kifejezés (Advanced exprission) lehetőséget és lépjük tovább.

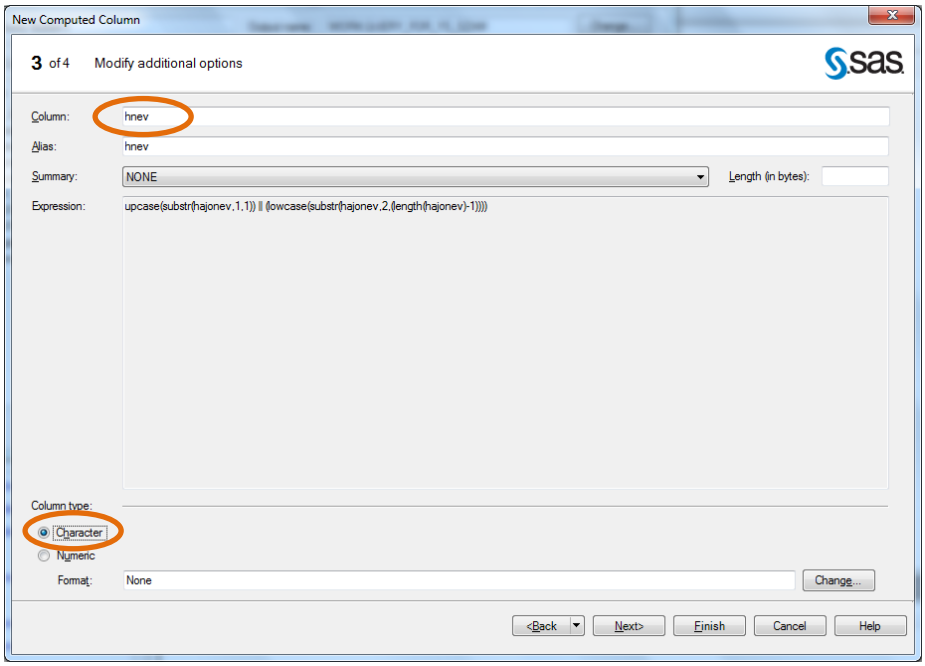
A következő ablakban írjuk be a kifejezésünket vagy építsük fel a segédfüggvények alapján. (13. ábra)

A kifejezés, amit alkalmazni fogunk alább látható:

```
upcase(substr(hajonev,1,1)) || (lowcase(substr(hajonev,2,(length(hajonev)-1))))
```

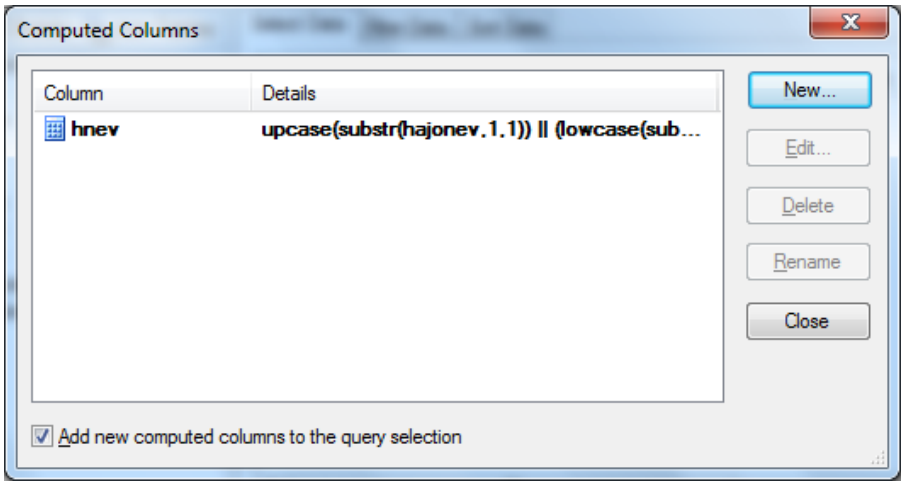
Magyarázat: Az a célunk, hogy csak az első karakter legyen nagy a többi kicsi, így a substr (részszting) függvény segítségével kiválasztjuk az első karaktert és azt az upcase függvénnyel nagybetűsre konvertáljuk, majd a || operátorral hozzáfűzzük a maradék részsztring kisbetűs (lowcase) változatát.

A kifejezés megadása után lépjünk tovább és állítsuk be a változó tulajdonságait az alábbiak szerint.



A változó neve legyen HNEV valamint biztosítsuk, hogy típusát tekintve karakter legyen.

Ezzel meg is szerkesztettük az új változónkat. Lépjünk tovább.



21. ábra Létrehozott új változó

A 21. ábra mutatja az eredményt. Ha alul kipipáljuk a felkínált lehetőséget, akkor a SAS automatikusan hozzáadja a származtatott változót a lekérdezni kívánt változók listájához.

A rajtszámra és a hajó típusára is szükségünk lesz ebből az adatállományból, így ezeket is válasszuk ki.

22. ábra Származtatott változó

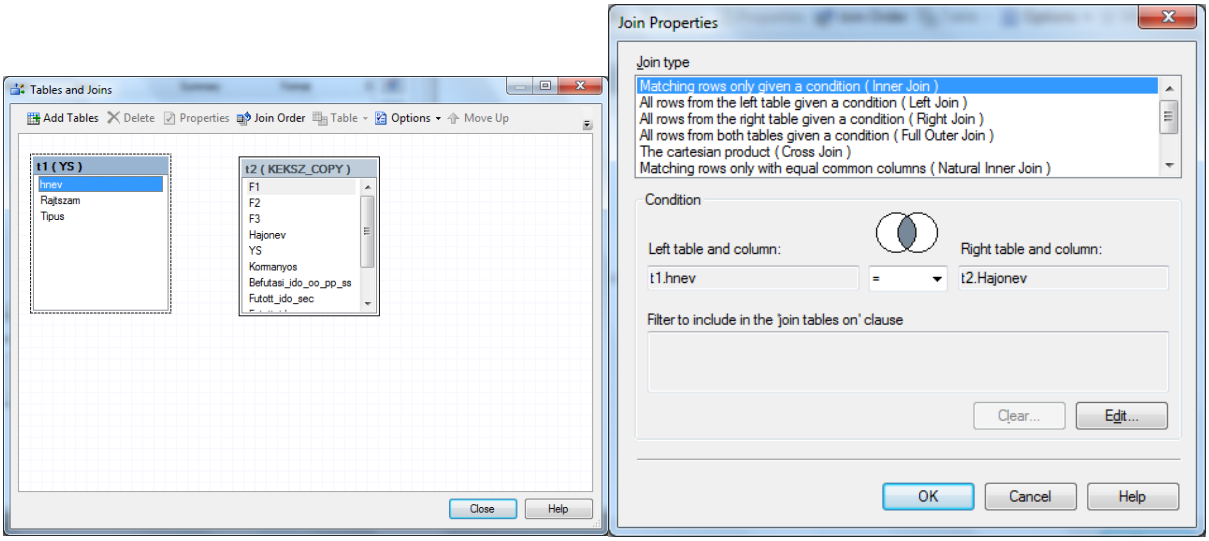
Futtassuk a csomópontot és nevezzük el YS névvel a kimenetet.

A következő lépés a kívánt táblák összekapcsolása. Ehhez vegyünk fel a YS tábla után egy Query Builder csomópontot és válasszuk a **Hiba! A hivatkozási forrás nem található.** látható Join Tables gombot.

A megjelenő ablakban láthatjuk a jelenlegi adatállományunkat. A felső menüsor első elemét választva (Add Tables) további táblákat adhatunk meg az összeillesztéshez.

Válasszuk ezt az opciót és a felkínált táblák közül adjuk hozzá a KEKSZ_COPY adatállományt.

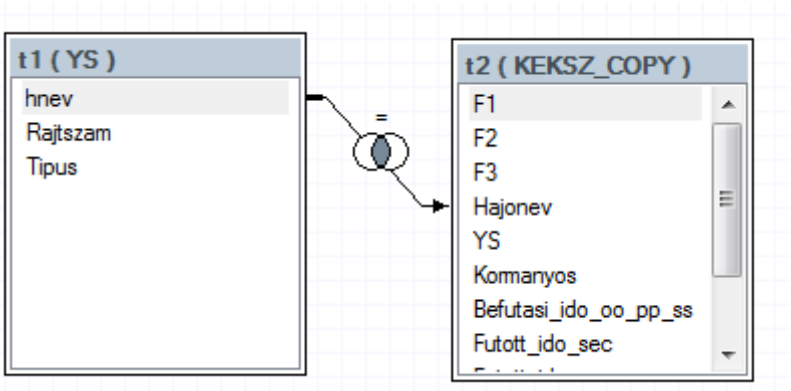
Ezek után a SAS figyelmeztet minket, hogy nem tudja összekapcsolni a két állományt. Ezt nekünk kell manuálisan megtennünk. A SAS azt az egyszerű utat követi, hogy megnézi vannak e azonos nevű és típusú változók. Mivel esetünkben ez nem teljesül, így nem tudta összekapcsolni az állományokat.



23. ábra Az összekapcsolás megadása

A 23. ábra jobb oldalán látható, hogy milyen kapcsolási lehetőségeket kínál fel a SAS számunkra. Az egyes kapcsolásokról részletesen a SAS BASE programozás könyvben olvashatunk.

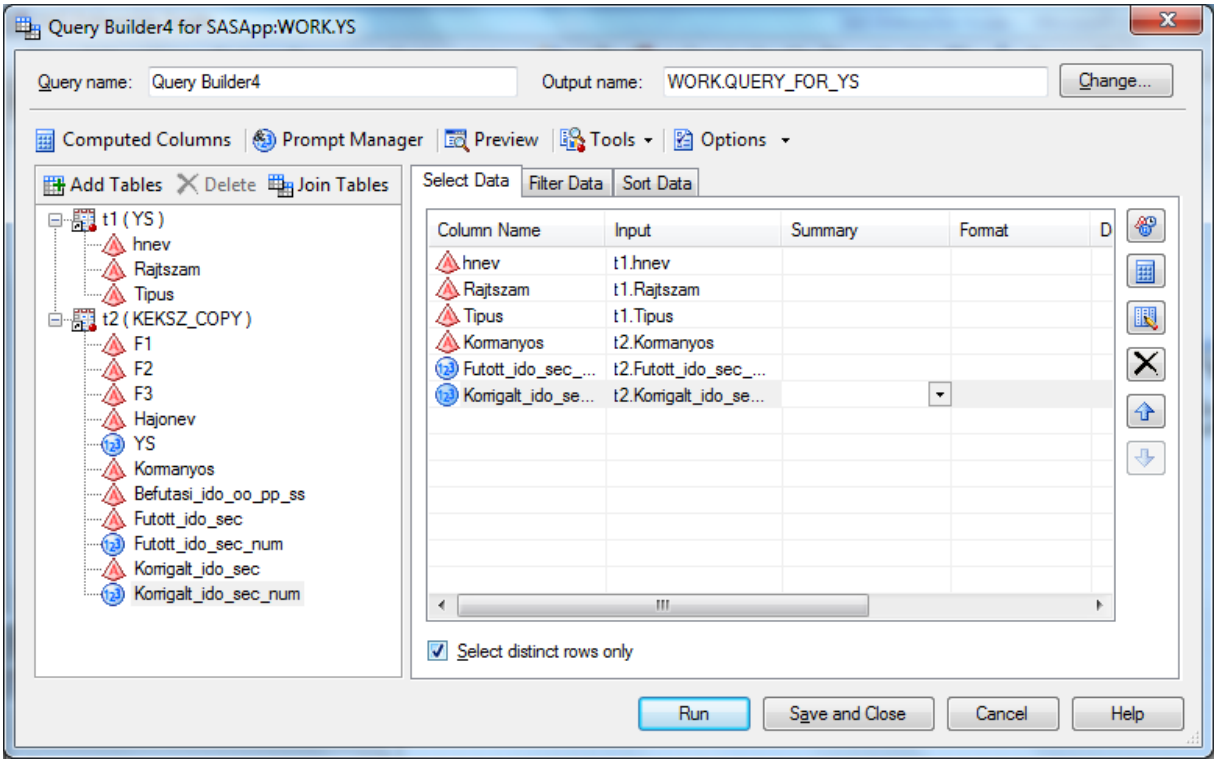
Jelen esetben válasszuk a belső illesztést (Inner Join). Ebben az esetben, a két adatállományból azok a megfigyelések kerülnek a kimenetre, amiknek a kulcs mind a kettőben megtalálható.



24. ábra Kapcsolás

Ezek után a Query Builder már a kapcsolt táblák összes változóját a rendelkezésünkre biztosítja, hogy összeállítsuk a kívánt kimenetet. Lehetőségünk van a korábban már ismertetett szűrésre és rendezésre.

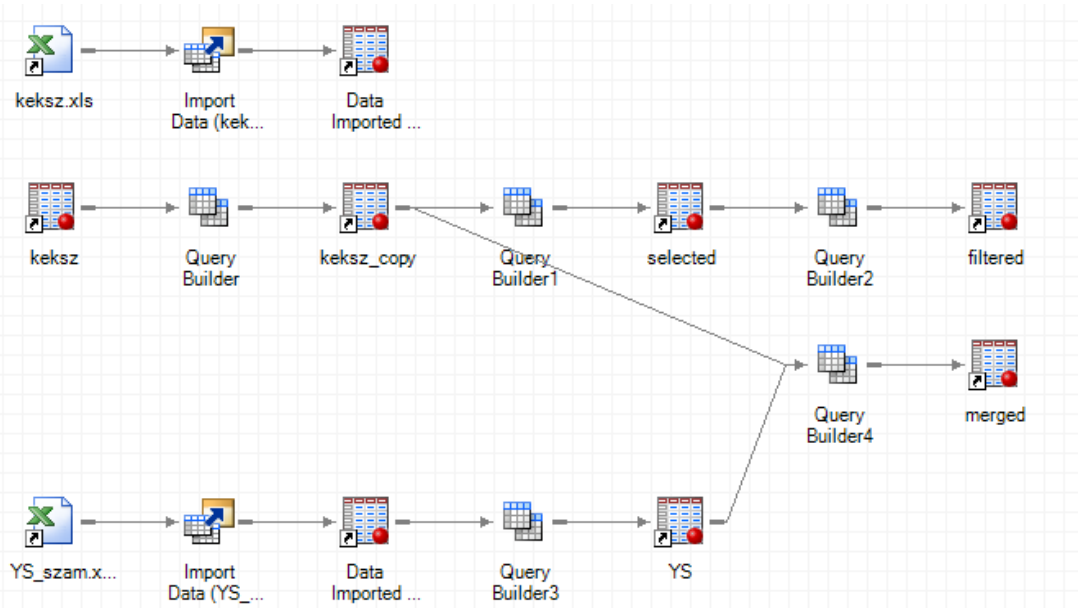
Válasszuk ki a 25. ábra látható változókat.



25. ábra Összekapcsolt táblák változóinak lekérdezése

Nevezzük el a kimeneti állományt merged névvel és futtassuk a csomópontot.

A folyamatábránk ekkor jól mutatja az összeillesztést.

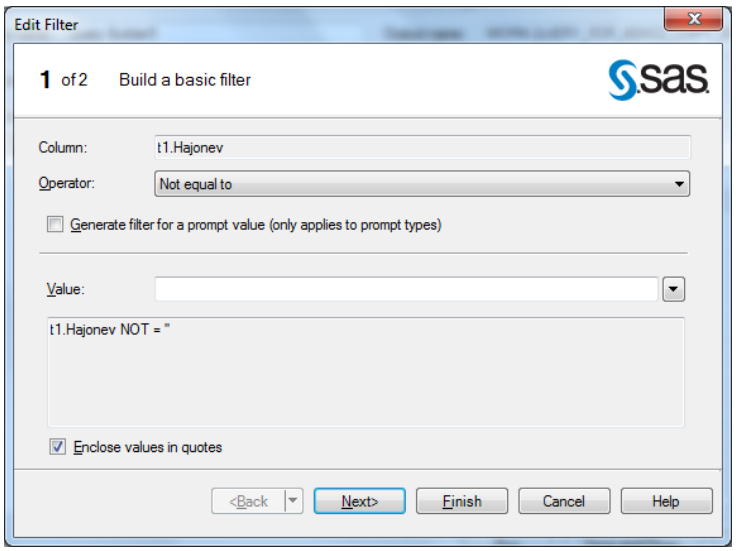


26. ábra Folyamatábra az összeillesztés után

Az eredményeket tanulmányozva észrevehetjük, hogy a feltevésünk, mely szerint a hajónév változó megfigyelései egyediek tévesnek bizonyult, hiszen számos hajónév több megfigyelés is található. Ráadásul súlyos hiba, ha a kulcsváltozó hiányzó értékeket is tartalmaz.

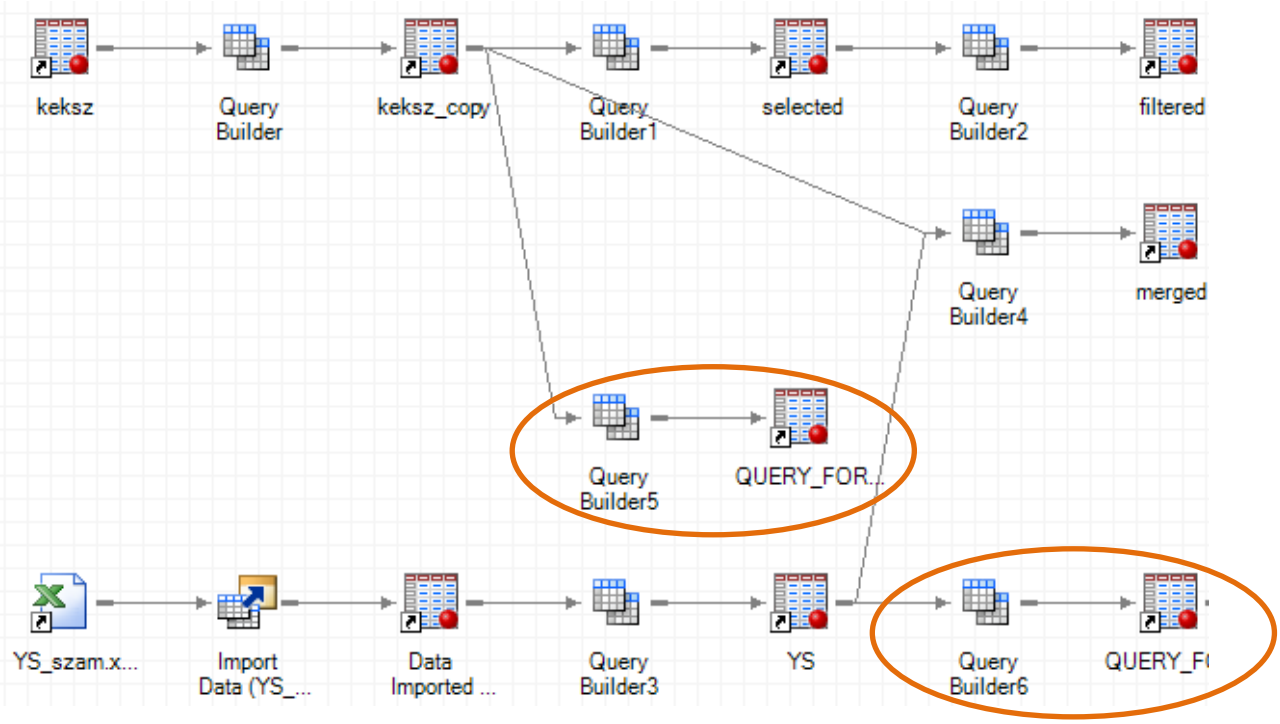
Így a továbbiakban korrigáljuk hibáinkat.

Alkalmazzuk a Query Builder csomópontot mind a két adatállományra (KEKSZ, YS) úgy, hogy a lekérdezett változók közé felvesszük az állományok összes változóit és a Filter Data fülön megadjuk, hogy a hajonev és a hnev változó se tartalmazzon hiányzó értéket, azaz például a hajonev változó esetén:



27. ábra Hiányzó (üres) értékek szűrése

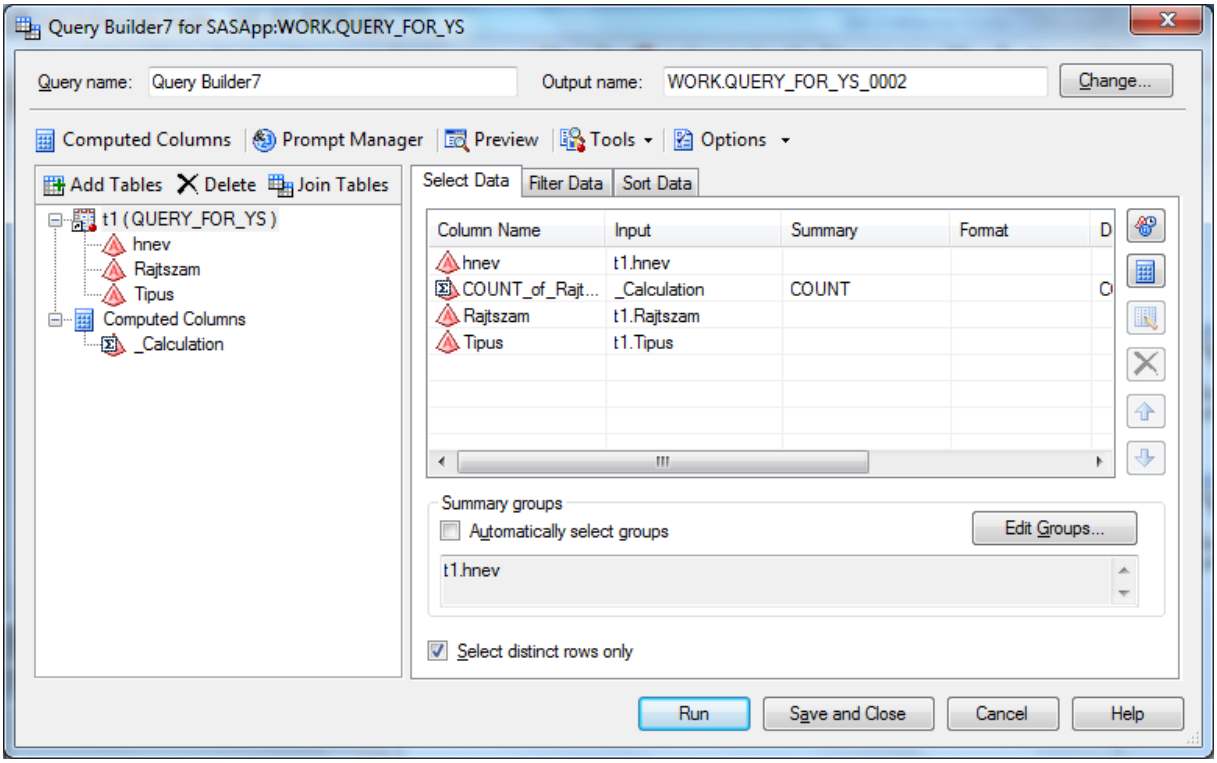
Futtatjuk a két Query Builder csomópontot, így megkapjuk az állományokat, amik már nem tartalmaznak üres hajóneveket.



Az ismétlődő megfigyelések eltávolítása összetettebb feladat.

Az elve a következő. Meghatározzuk, hogy az egyes rajtszámok hányszor szerepelnek a hajónév szerint csoportosítva. Ez nem azt jelenti, hogy a rajtszámokat csoportosítjuk, hanem itt a rajtszám csak, mint darabszám szerepel (mennyiségi ismérv). Létrehozhatnánk egy új változót, aminek az értéke konstans 1 lenne, és ezeket adhatnánk össze hajónév szerint csoportosítva. Ekkor azonos eredményre jutnánk.

Most is a Query Builder csomópontot alkalmazzuk, de most a rajtszám felvétele után a Summary oszlopban kiválasztjuk a COUNT opciót, azaz csak összeszámoljuk az esetek számát. Ezek után megadjuk a csoportosítandó változót úgy, hogy kivesszük a pipát az automatikus csoportosítás opció elől. Majd megadjuk a HNEV változót. Célszerű az összeszámlálás értéke szerint csökkenő sorrendbe rendezni a kimeneti állományt.



28. ábra Esetek összeszámlálása

Futtassuk a csomópontot.

	hnev	COUNT_of_Rajtszam	Rajtszam	Tipus
1	Márta	3	666	Birdie 24
2	Márta	3	980	Enter 29
3	Márta	3	AUT 430	H-boot+genua cs...
4	Skorpio	3	1088	Dehler Sprinta 70
5	Skorpio	3	1128	Larsen Maribo
6	Skorpio	3	233	Rebell MK 2
7	Aloha	2	16274	B 25 + spoiler
8	Aloha	2	441	Jeanneau 20
9	Anci	2	1062	Jeanneau S.O 3..
10	Anci	2	1209	Jeanneau S.O. 4...
11	Aquamarin	2	105	FPC 30 T
12	Aquamarin	2	488	H-boot+genua

29. ábra Eredmény (részlet)

Most már rendelkezésünkre állnak az egyes hajónevek gyakoriságai, így már meg tudjuk szűrni az adatállományt e változó alapján.

Alkalmazzuk a már jól ismert Query Builder vagy Filter and Sort csomópontot a korlátozás megvalósításához.

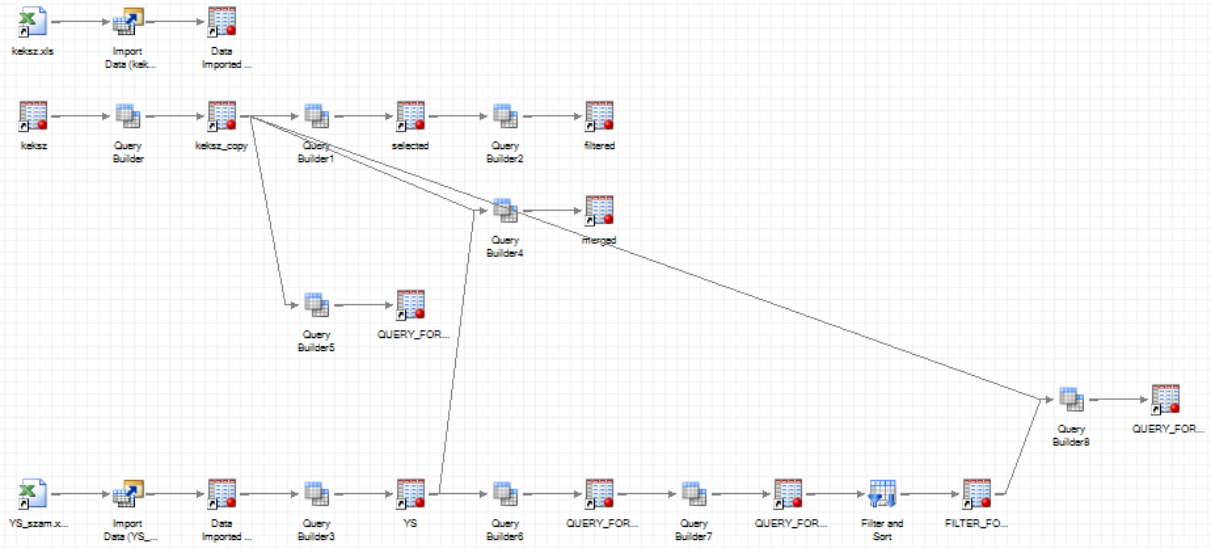
A kapott adatállományban csak olyan hajónevek lehetnek, melyek előfordulásának a száma egy. Mivel a KEKSZ adatállományban a kulcsváltozó egyedi volt, így már elvégezhető az összeillesztés.

Az összekapcsolás a korábbiak alapján történik a megfelelő adatállományok összekapcsolásával, így ennek részletezésétől eltekintünk.

A kapott eredmény egy részletét alább szemléltetjük.

	hnev	COUNT_of_Rajtszam	Rajtszam	Tipus	Kormanyos	Futott_ido_sec_num	Korrigalt_ido_sec_num
1	A hajó	1	914	Comet 38	Puska József	112959	122782
2	Accenture	1	1054	Dolphin 28cs+	Gál Attila	96900	89722
3	Adagio	1	1126	Dufour 365	Posta Péter	103226	108659
4	Adrienn	1	633	Dufour 36 Classic	dr.Juhász Béla	101382	100378
5	Albatrosz	1	467	Dehler Varianta...	Kétszeri Csaba	96390	93583
6	Allegro	1	597	Dehler 33	Bornemissza Lás...	91588	97434
7	Allure	1	1136	B 25 spoiler, mer...	Pandur László	135266	119704
8	Ami	1	198	B 25	Fürstall Ferenc	134594	116029
9	Amygdala	1	1149	Yolle 20	dr.Gellér László	103638	98703
10	Anna	1	1253	Saturn 25 (Satur...	Csoregh Zoltán	59255	75968
11	Aquanauta	1	906	Elan 333+	Takács Gábor	90165	95920
12	Aquarius	1	101	FPC 30 T	Molnár Imre	90219	87591
13	Ariadne	1	28	Egyedi	Jókuthy Miklós	113778	97246
14	Aries	1	692	Dehler Delenta 80	dr.Csiki Tamás	108581	106452
15	Asterix	1	983	B 24	dr.Szalai István	90583	100648

30. ábra Helyes összeillesztés eredménye



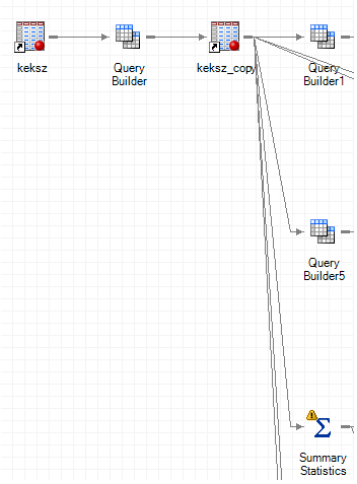
31. ábra A végső folyamatábra

10. Leíró statisztika készítése

10.1 Összegző statisztikák

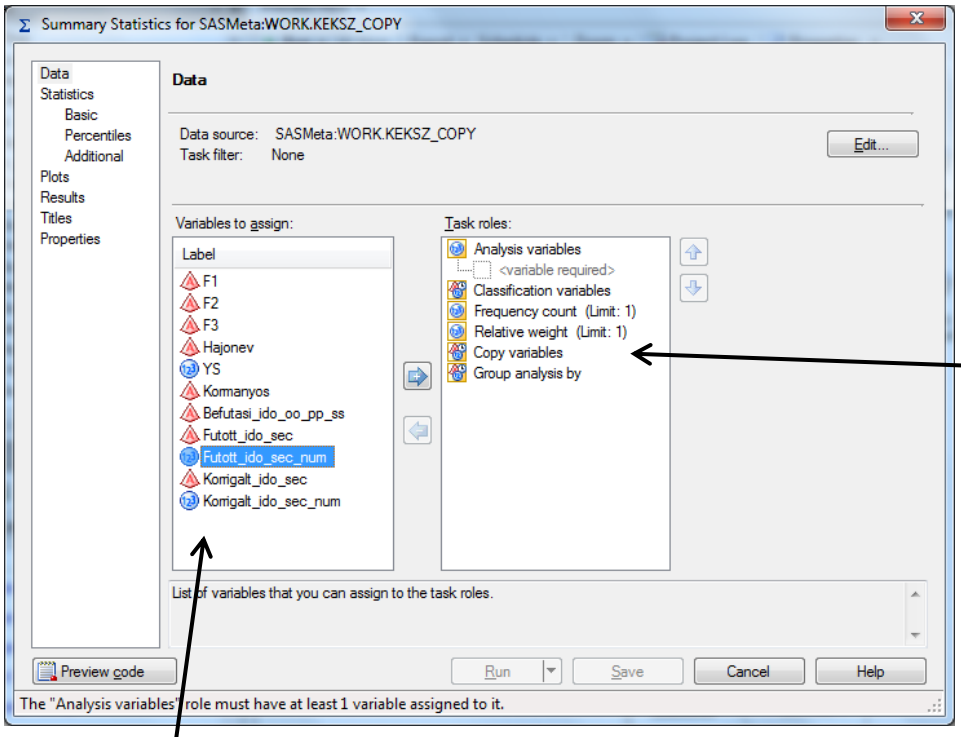
Leíró statisztikát a Task menü Describe menüpontja segítségével készíthetünk. Számos csomópont található ebben a részben melyek mindegyike az adatok megértésére és vizsgálatára szolgálnak.

Vizsgáljuk meg a KEKSZ_COPY adatállományunkat. A vizsgálathoz kössünk egy Summary Statistics csomópontot a KEKSZ_COPY csomóponthoz az alábbiak szerint.



32. ábra Summary Statistics csomópont alkalmazása

Fontos, hogy megértsük a csomópontok felparaméterezését, hiszen szinte az összes csomópont esetén hasonlóképpen kell eljárnunk.



33. ábra Felparaméterezés

Első lépésben meg kell adnunk, hogy melyik változót / változókat akarjuk elemezni (Analysis variables). Egyszerűen csak át kell húznunk a kívánt változókat a megfelelő szerepkörbe.

Megadhatunk csoportosító és osztályozó változót is.

Jelen esetben válasszuk ki a Futott_ido_sec_num és a Korrigalt_ido_sec_num változókat, mint elemzendő változók.

A 33. ábra bal oldalán lévő oszlopban válasszuk a Statistics elemet. Itt kiválaszthatjuk, hogy milyen statisztikai mérőszámokat kívánunk kiszámítatni. Az alapértelmezetteken kívül válasszuk a móduszt (mode) és a terjedelmet (range).

A Percentiles elemre klikkelve kiszámíthatjuk a kívánt percentiliseket. Kérjük le a mediánt és az alsó és felső kvantiliseket.

Az Additional elem alatt megadhatjuk a kívánt konfidencia szintet. Alapértelmezésben 95%.

A Plots elem alatt grafikonokat lehet lekérni. Válasszuk ki a hisztogramot.

Végezetül futtassuk a csomópontot.

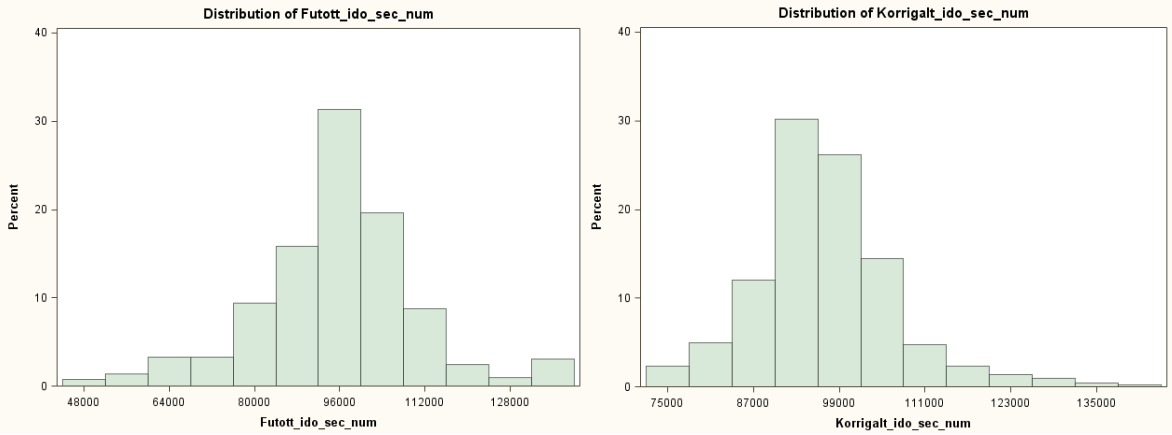
*Eredmény*ek:

Summary Statistics

Results

The MEANS Procedure

Variable	Mean	Std Dev	Minimum	Maximum	Mode	Range	N	Lower Quartile	Median	Upper Quartile
Futott_ido_sec_num	95954.00	15092.57	47620.00	138369.00	90977.00	90749.00	524	90149.50	96689.00	104862.00
Korrigalt_ido_sec_num	97027.41	9649.72	74336.00	140369.00	85042.00	66033.00	524	91216.00	96142.00	101792.50



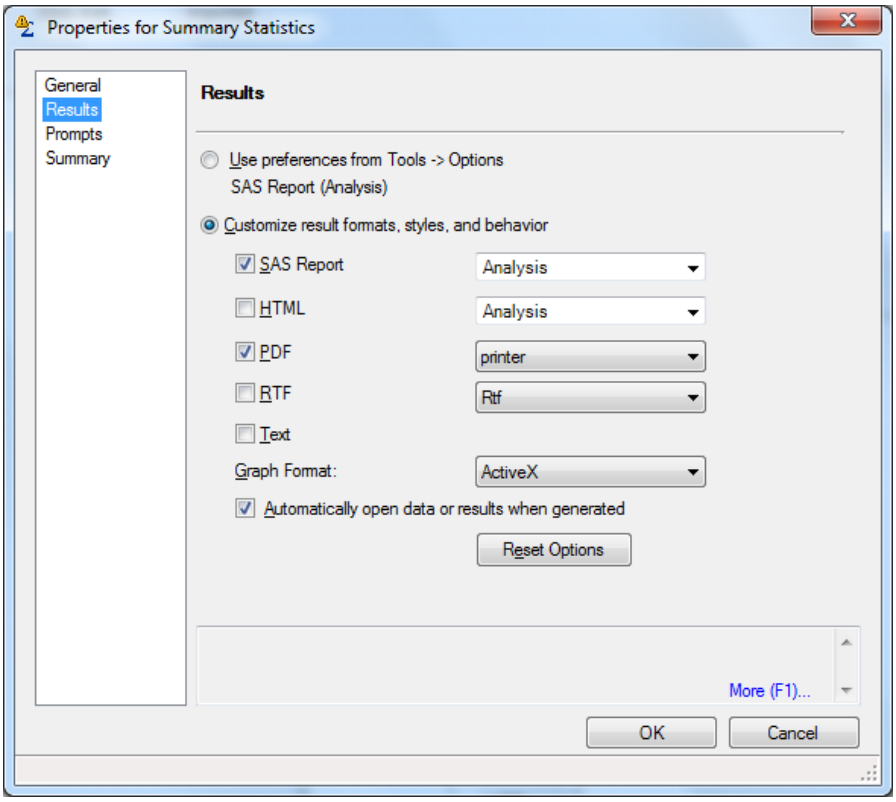
10.2 Kimenetek típusának megadása

Lehetőségünk van a kimenetek típusát megváltoztatni. Ehhez klikkeljünk jobb klikkek a Summary Statistics csomópontra. Válasszuk a Properties menüpontot.

A megjelenő ablak bal oldalán lévő listában válasszuk a Results pontot. (34. ábra)

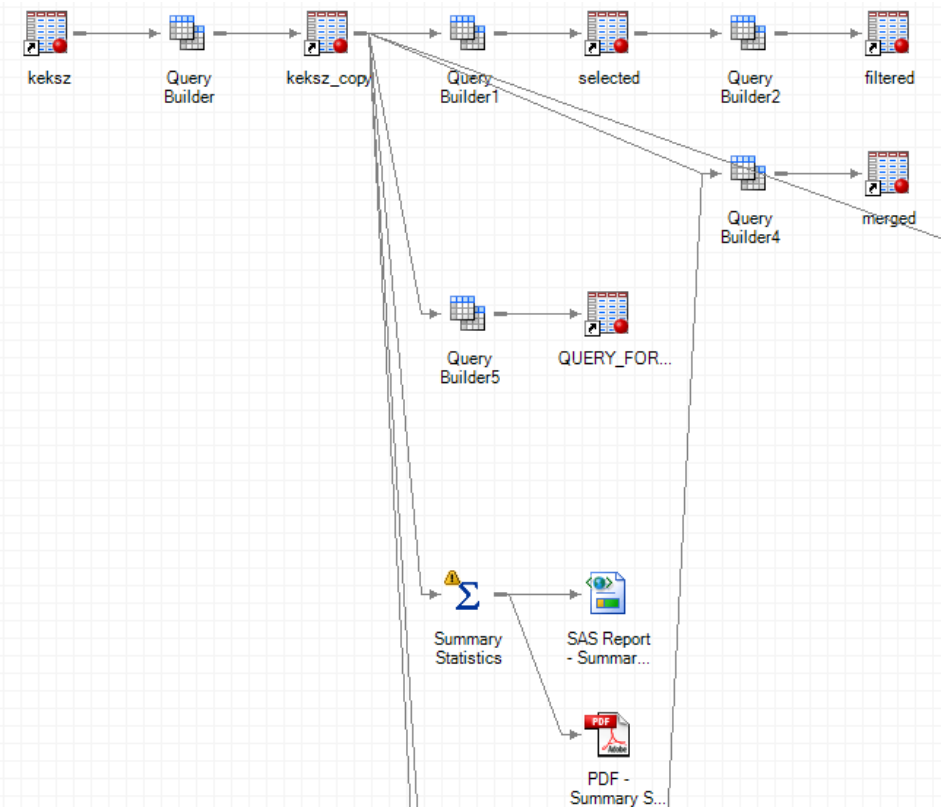
Válasszuk a Customize result formats, styles and behaviour lehetőséget.

Ezek után kiválaszthatjuk, hogy milyen típusokban jelenjenek meg az eredmények.



34. ábra Kimenetek típusának megadása

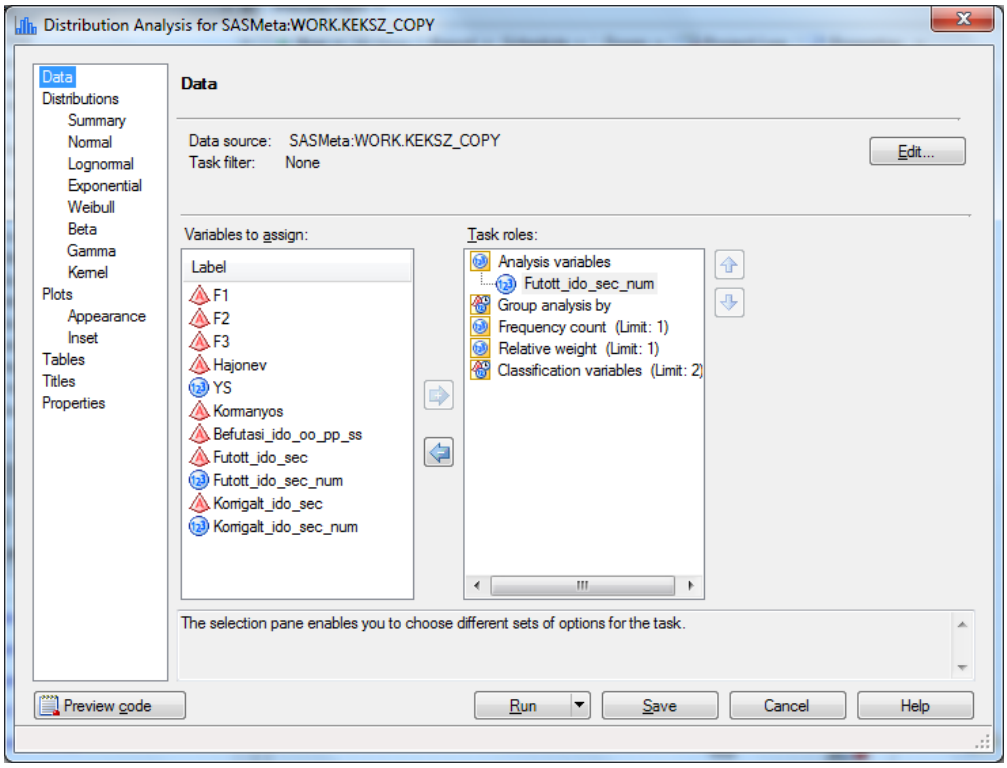
Globálisan a kimeneteket a Tools menü / Options / Results General menüpontban tudjuk megadni. Az itt kiválasztott kimenetek minden kimenetet generáló csomópontra érvényesek lesznek.



35. ábra PDF típusú kimenet generálása

10.3 Eloszlásvizsgálat

Eloszlásvizsgálatot a Task menü / Describe / Distribution Analysis csomópont segítségével tudunk elvégezni. Kössük ezt a csomópontot a KEKSZ_COPY csomópont után.



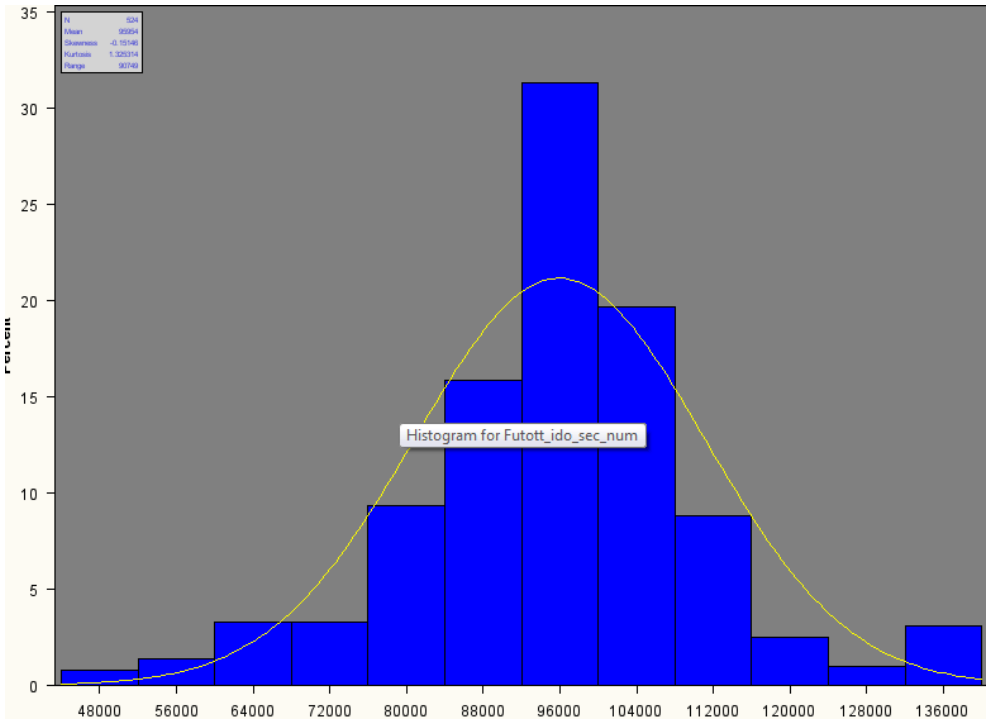
36. ábra Eloszlásvizsgálat

Válasszuk a Futott_ido_sec_num változót elemzendő változónak. A Summary fülön ki tudjuk választani, hogy milyen típusú eloszláshoz viszonyítsuk a kívánt változónk eloszlását. Válasszuk a Normált.

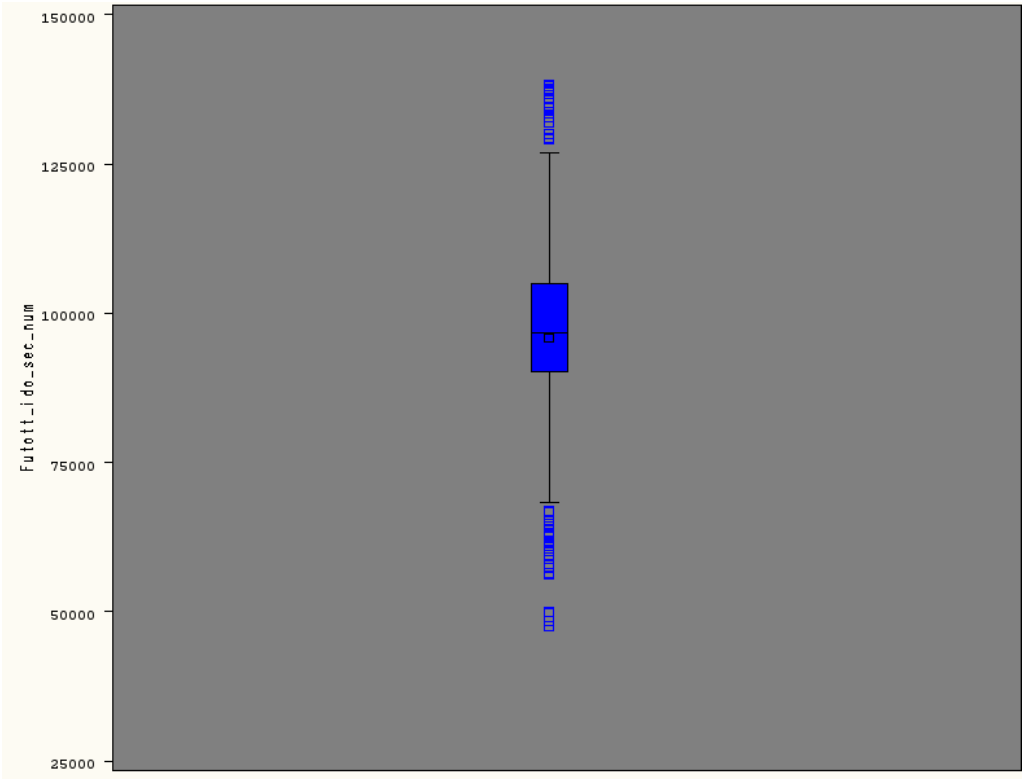
A Plots elem alatt kiválaszthatjuk a nekünk szükséges grafikon típusát. Legyen ez esetünkben hisztogram és a Boksz plot.

Az Inset elem alatt kiválaszthatjuk, hogy milyen mérőszámot kívánunk megjeleníteni a grafikonon. Válasszuk ki a csúcsosságot (Kurtosis) és a ferdeséget (Skewness).

Futtassuk a csomópontunkat.



37. ábra Eloszlásvizsgálat hisztogram alkalmazásával



38. ábra Box diagram

11. Korreláció vizsgálat

Korreláció vizsgálatot akkor végzünk, ha meg akarjuk határozni két mennyiségi ismérv közötti sztochasztikus kapcsolat (tendenciaszerű) erősségét, azaz azt akarjuk meghatározni, hogy egy jelenség alakulását egy másik jelenség milyen mértékben befolyásolja.

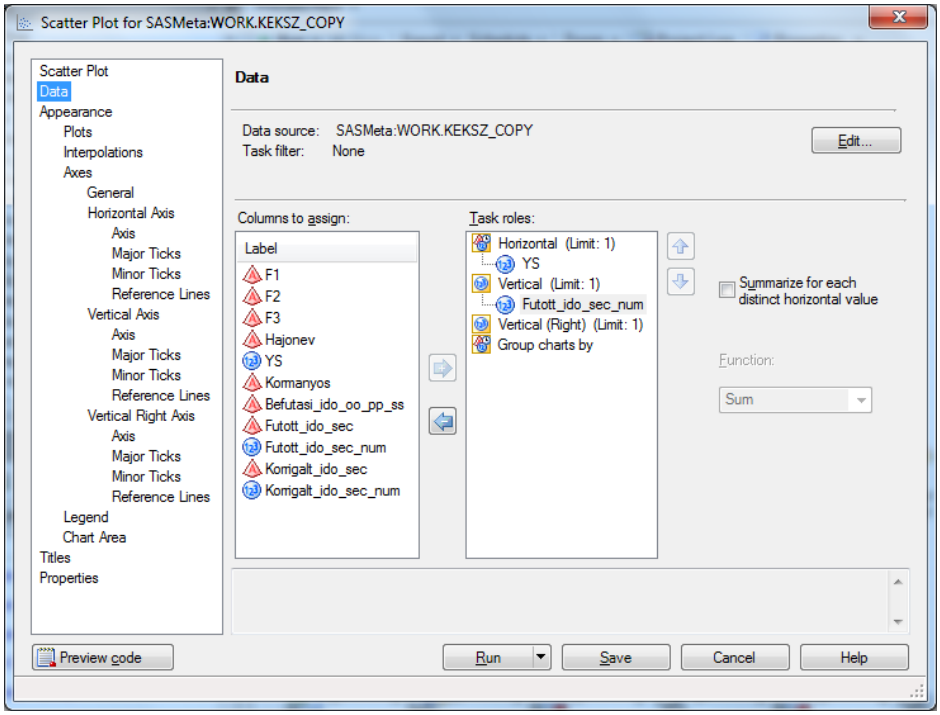
A korreláció vizsgálatot célszerű egy pontdiagrammal kezdeni.

11.1 Pontdiagram

Készítsünk pontdiagramot a YS-szám (yardstick szám) és a futott idő változókról.

Válasszuk ki a KEKSZ_COPY adatállományunkat és menjünk a Task menü / Graph / Scatter plot menüpontba. Ezzel kiválasztottuk a szükséges csomópontot.

2D-s ábrát fogunk készíteni az alábbiak szerint.



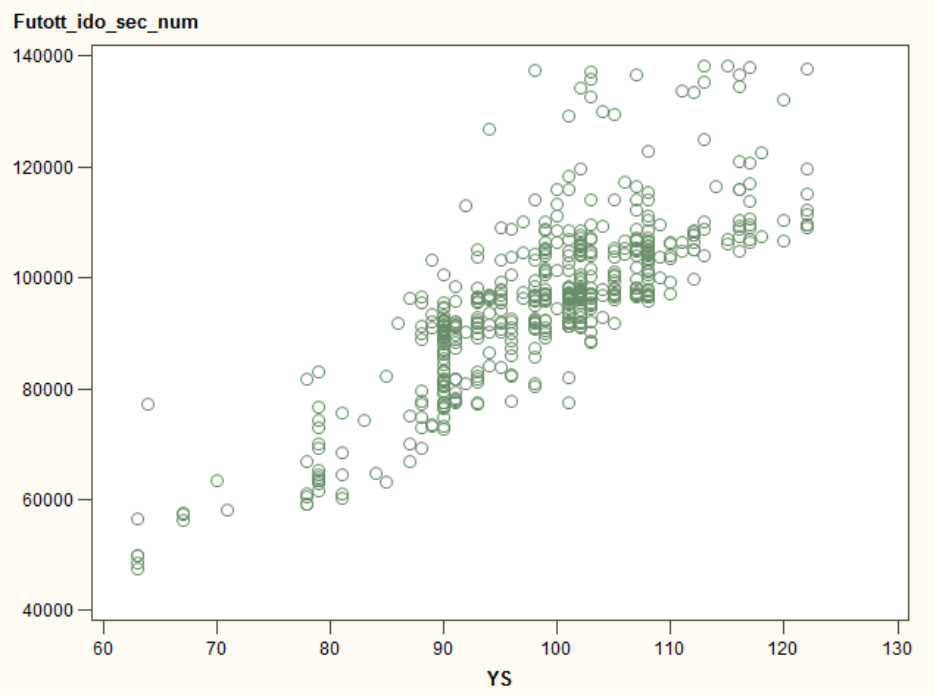
39. ábra Pontdiagram paraméterei

A vízszintes tengelynek állítsuk be a YS változót, míg a függőleges tengelyhez a Futott_ido_sec_num változót rendeljük.

Lehetőségünk van beállítani a tengelyeket és az osztáspontok számát, de ettől most tekintsünk el és futtassuk a csomópontunkat.

A kapott eredményeket a 40. ábra mutatja.

Az ábrát tanulmányozva megállapíthatjuk, hogy van értelme tovább vizsgálódni, hiszen az eredmény valamiféle kapcsolatot mutat.

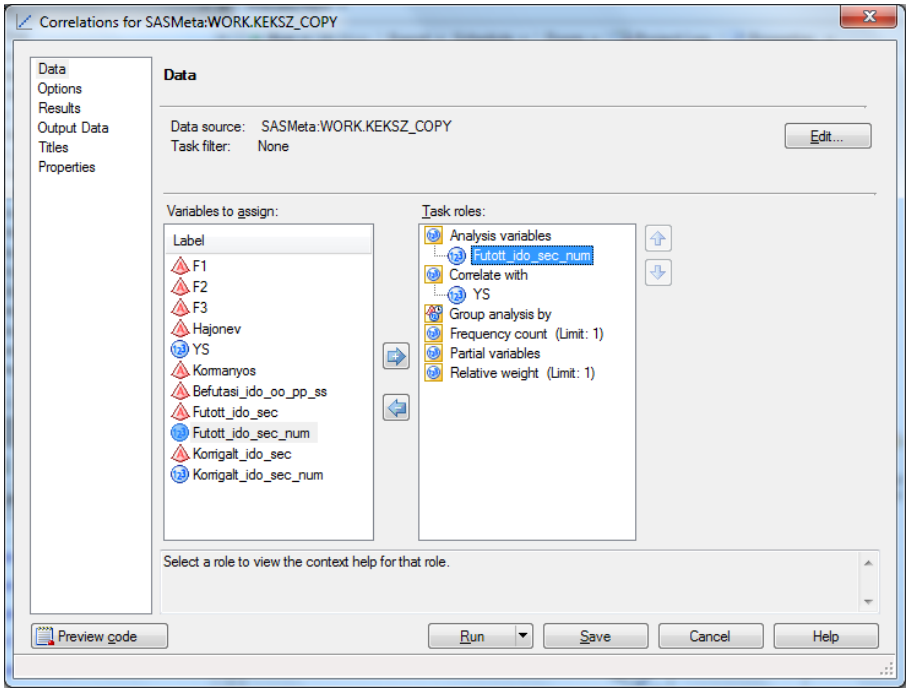


40. ábra Pontdiagram

11.2 Korreláció erősségének meghatározása

Ennek meghatározásához válasszuk a Task menü / Multivariate / Correlation menüpontot.

A megjelenő ablakban adjuk meg a függő (elemzendő változó) és a független (magyarázó) változót.



41. ábra Korreláció paraméterezése

Függő változó: Futott_ido_sec_num

Független: YS

1 With Variables:	YS
1 Variables:	Futott_ido_sec_num

Simple Statistics						
Variable	N	Mean	Std Dev	Sum	Minimum	Maximum
YS	575	99.17217	10.53664	57024	63.00000	129.00000
Futott_ido_sec_num	524	95954	15093	50279897	47620	138369

Pearson Correlation Coefficients Prob > |r| under H0: Rho=0 Number of Observations	
	Futott_ido_sec_num
YS	0.79782
	<.0001
	524

1. táblázat Korreláció eredménye

Az 1. táblázat jól látható, hogy a korreláció erőssége (R) 0,7978. Az alábbi állításokkal élve ez a kapcsolat erősnek mondható.

Kapcsolat értéke (R)	**Kapcsolat erőssége**
R = 0	Nincs kapcsolat
R < 0,5	Gyenge kapcsolat
R < 0,8	Közepesen erős kapcsolat
R > 0,8	Erős kapcsolat

2. táblázat Kapcsolat erőssége