Regresszió

Pusztaházi Luca Sára
Ország Anna

Budapesti Műszaki és Gazdaságtudományi Egyetem
Matematika Intézet
Sztochasztika Tanszék

2017

Tartalomjegyzék

Elméleti háttér I.

- A regresszióanalízis során két vagy több véletlen változó között fennálló kapcsolatot modellezünk.
- Lehet lineáris és nemlineáris a regresszió a modell tulajdonságai alapján.
- A regresszió feladata az $Y = l(X_1, X_2, \ldots, X_n)$ függvény meghatározása.
- Y: függő (dependent) változó
- $X_1, X_2, \ldots, X_n$: magyarázó (predictor) változók

Elméleti háttér II.

- Cél: az Y valószínűségi változó közelítése az $l(X_1, X_2, \ldots, X_n)$ függvénnyel legkisebb négyzetes értelemben.
- Azaz az $\mathbb{E}(Y - l(X_1, X_2, \ldots, X_n))^2$ várhatóérték minimalizálása.
- Belátható, hogy ezt az $\mathbb{E}(Y|X_1, X_2, \ldots, X_n)$ feltételes várhatóérték elégíti ki.
- Ezt az együttes sűrűségfüggvényből $f(Y, X_1, X_2, \ldots, X_n)$ meg lehetne határozni, de általában ez nem adott, csak egy statisztikai mintánk van a függő és független változókra.

Elméleti háttér III.

- Lineáris regresszió

$$Y = a_1 X_1 + a_2 X_2 + \cdots + a_p X_p + b$$

- Multiplikatív modell

$$Y = b X_1^{a_1} X_2^{a_2} \ldots X_p^{a_p}$$

- Polinomiális regresszió

$$Y = a_1 X + a_2 X^2 + \cdots + a_p X^p + b$$

Módszertan I.

1. Adatok vizsgálata
    - A független változók normális eloszlásúak?
    - Vannak extrém értékű, kihagyható megfigyelések?
    - Pontdiagram alapján az $y - x$ párok lineáris kapcsolata fennáll?
    - A független változók közötti páronkénti korrelációk gyengék? Ha nem, akkor szakmai vagy statisztikai szempontok alapján válogatjuk ki a magyarázó változókat?

Módszertan II.

2. Az illesztés menete, a változók közötti szelekció végrehajtása
    - Melyek a statisztikai értelemben legerősebb magyarázó erővel bíró változók?
    - Létezik-e lineáris modell vagy minden becsült együttható nullának tekinthető?
    - Milyen tesztekkel és hogyan minősíthető a modell egésze?

Módszertan III.

3. A magyarázó változók közötti kapcsolatrendszer megfelelő?
    - Milyen mutatókra támaszkodhatunk annak mérésekor, hogy túlzott multikollinearitás fellépett-e?
    - Mely változók elhagyásával küszöbölhető ki a multikollinearitás?

Módszertan IV.

4. Modell diagnosztika, hibatagok viselkedése, kiugró pontok kezelése
    - Megfelelő magyarázó erejű modellt kaptunk?
    - A hibatagok normális eloszlásúak?
    - A hibatagok szórása azonos-e, nem lépett fel heteroszkedaszticitás?
    - Vannak nagyon erős hatást gyakorló megfigyelések a mintában? Ezek elhagyása indokolt?

Lineáris regresszió feladata

- Az $Y = a_1X_1 + a_2X_2 + \cdots + a_pX_p + b$ egyenletben az együtthatók és a konstans meghatározása
- Azaz:

$$\min_{a_1,\ldots,a_p,b} g(a_1,\ldots,a_p,b) = \min_{a_1,\ldots,a_p,b} \mathbb{E}(Y-(a_1X_1+\cdots+a_pX_p+b))^2$$

Lineáris regresszió elméleti megoldása I.

- b szerinti derivált:

$$\frac{\partial g}{\partial b} = -2\mathbb{E}(Y - (a_1 X_1 + \cdots + a_p X_p + b)) = 0$$

- Kifejezve b-t:

$$b = \mathbb{E}Y - a_1 \mathbb{E}X_1 - \cdots - a_p \mathbb{E}X_p$$

- a_i $(i = 1, \ldots, p)$ szerinti derivált:

$$\frac{\partial g}{\partial a_i} = -2\mathbb{E}\left((Y - (a_1 X_1 + \cdots + a_p X_p + b))\, X_i\right) = 0$$

Lineáris regresszió elméleti megoldása II.

- b-t behelyettesítve az a_i szerinti deriváltba:

$$
\begin{aligned}
0 &= -2\mathbb{E}\left(\left(Y - \left(a_1X_1 + \cdots + a_pX_p + b\right)\right)X_i\right) \\
&= -2\mathbb{E}\left(\left(\left(Y - \mathbb{E}Y\right) - \sum_{j=1}^{p} a_j(X_j - \mathbb{E}X_j)\right)X_i\right) \\
&= -2\left(\mathbb{E}\left(YX_i\right) - \mathbb{E}Y\mathbb{E}X_i\right) + \\
&\quad +2\sum_{j=1}^{p} a_j\left(\mathbb{E}\left(X_iX_j\right) - \mathbb{E}X_i\mathbb{E}X_j\right) \\
&= -2\,\mathrm{Cov}(Y, X_i) + 2\sum_{j=1}^{p} a_j\mathrm{Cov}(X_i, X_j),\ \ i = 1, \ldots, p.
\end{aligned}
$$

Lineáris regresszió elméleti megoldása III.

- Jelölés: $\mathbf{C} := \mathrm{Var}(\underline{X})$, $\underline{d} := \mathrm{Cov}(Y, \underline{X})$
- Ebből a következő egyenletrendszert kapjuk:

$$\mathbf{C}\underline{a} = \underline{d},$$

- melynek egyértelmű megoldása, ha $\mathbf{C}$ reguláris:

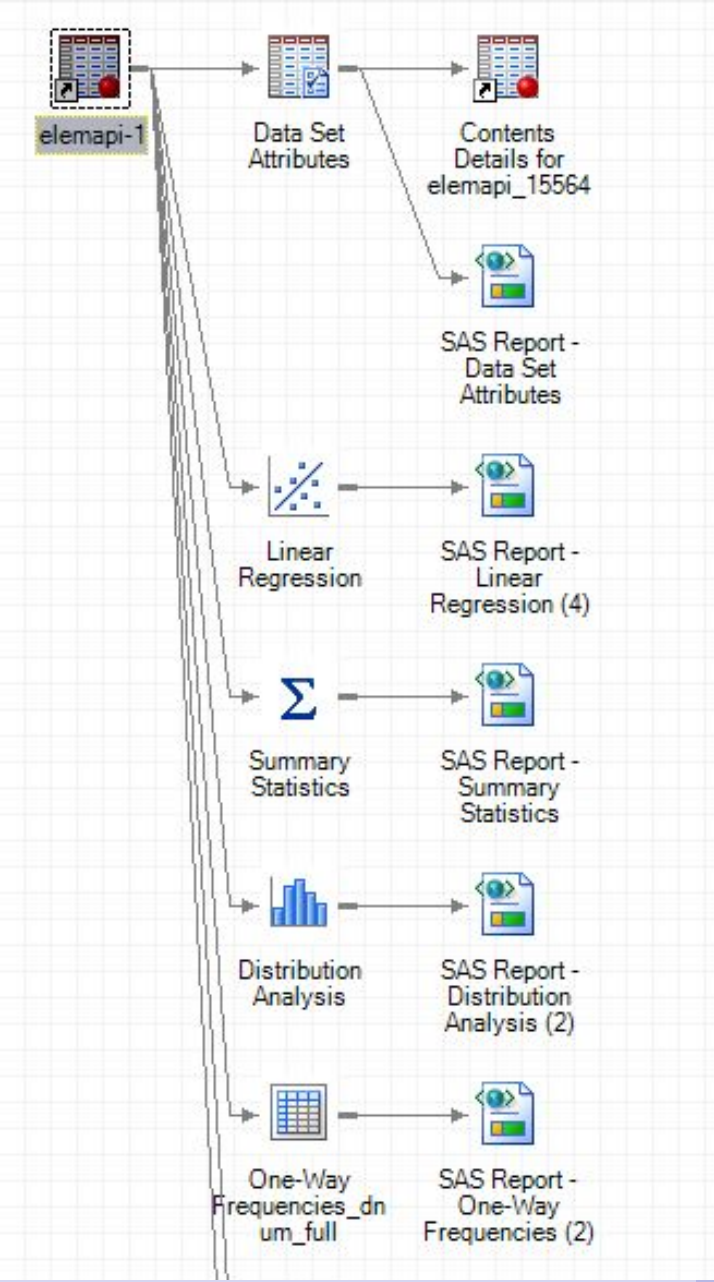
$$\underline{a} = \mathbf{C}^{-1}\underline{d}.$$

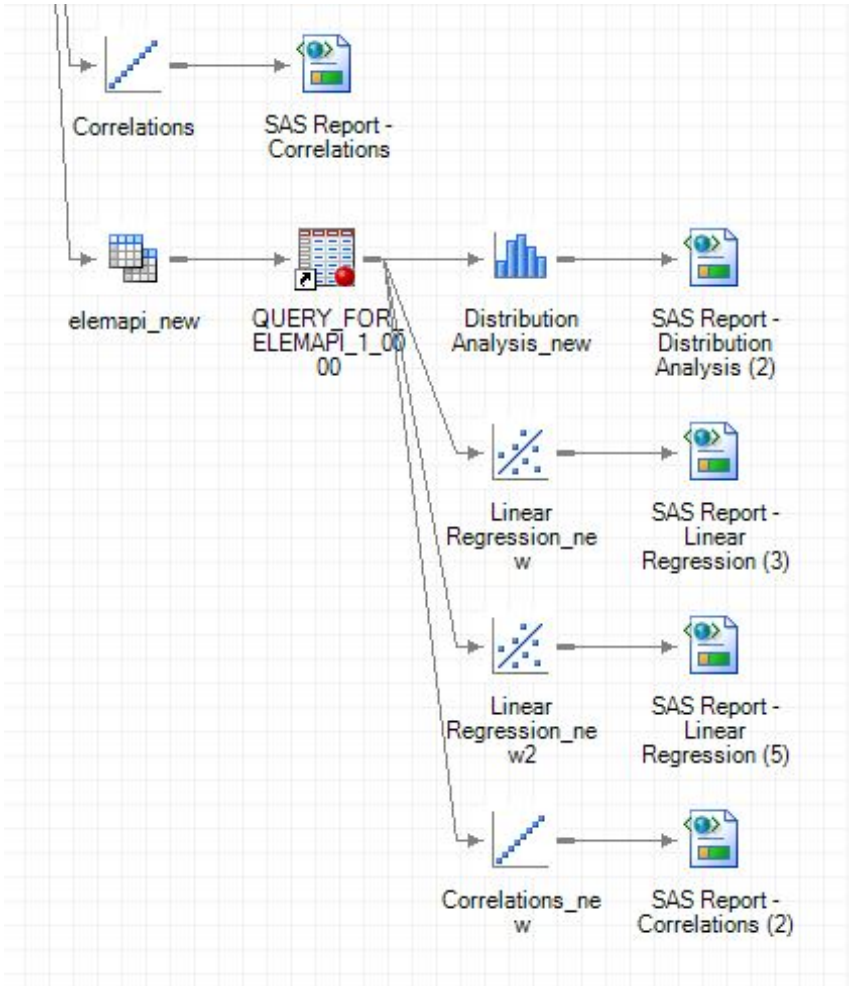
- A második derivált vizsgálatával belátható, hogy ez valóban minimum.

Adatállomány

- California Department of Education's Academic Performance Index (API) 1999-2000 Growth
- *api00*: API 2000
- *acs_k3*: átlagos osztálylétszám (k-3)
- *meals*: ingyenes vagy kedvezményes árú ebédprogramban résztvevő diákok százalékos aránya
- *full*: kitűnő ajánlással rendelkező tanárok százalékos aránya
- A 2000-es teljesítmény indexhez szeretnénk lineáris regressziót készíteni a másik három paraméter segítségével.
- *dnum*: körzetszám

Folyamatábra





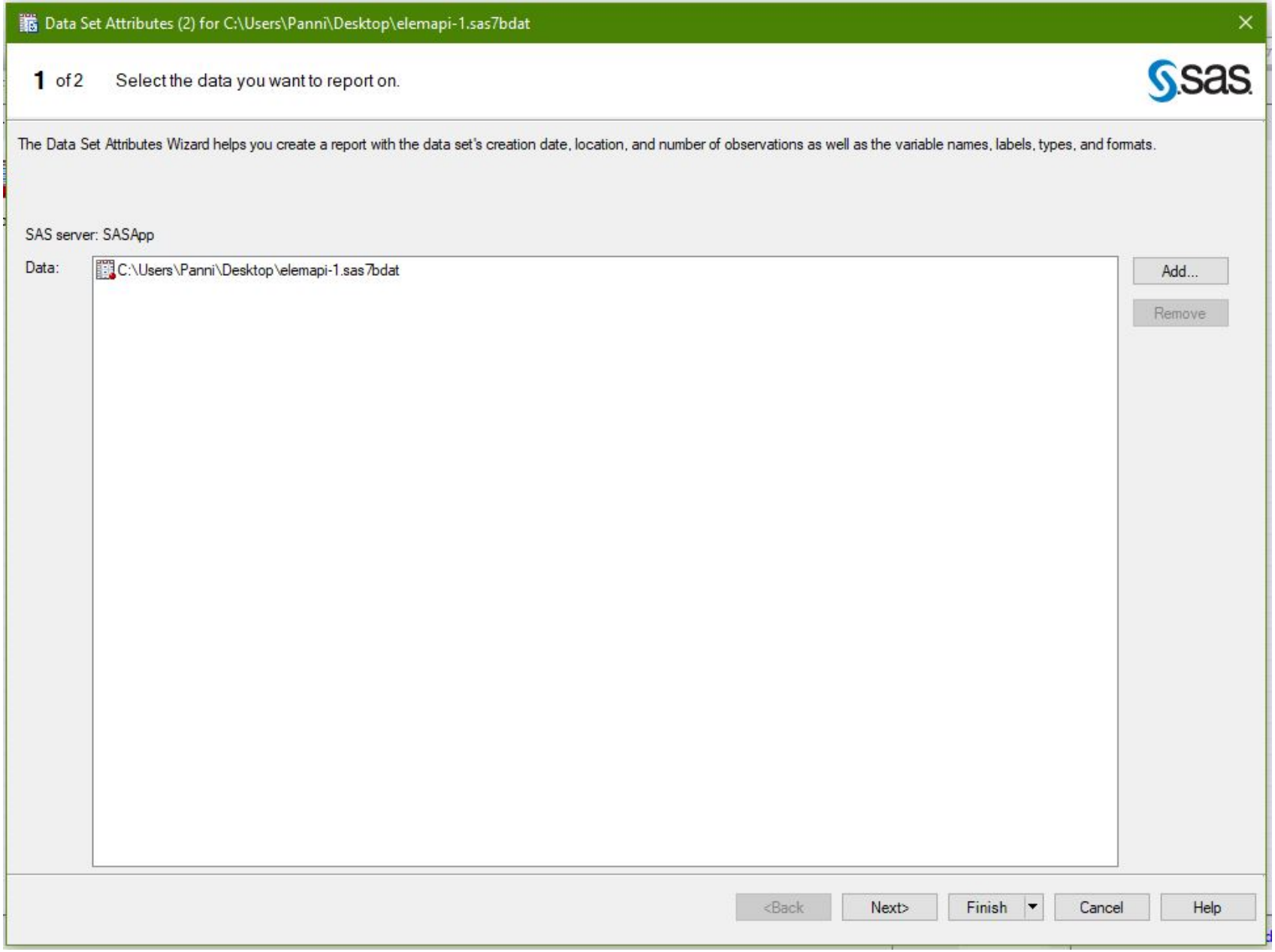
Útmutató a parancsokhoz

- Tasks/Data/Data Set Attributes
- Tasks/Regression/Linear Regression
- Tasks/Describe/Summary Statistics
- Tasks/Describe/Distribution Analysis
- Tasks/Describe/One-Way Frequencies
- Tasks/Multivariate/Correlation
- Tasks/Date/Query Builder

1. lépés – Ismerkedés az adatokkal

- Data Set Attributes: kapunk egy összefoglalót az adatokról, illetve megnézhetjük a változók típusát és leírását

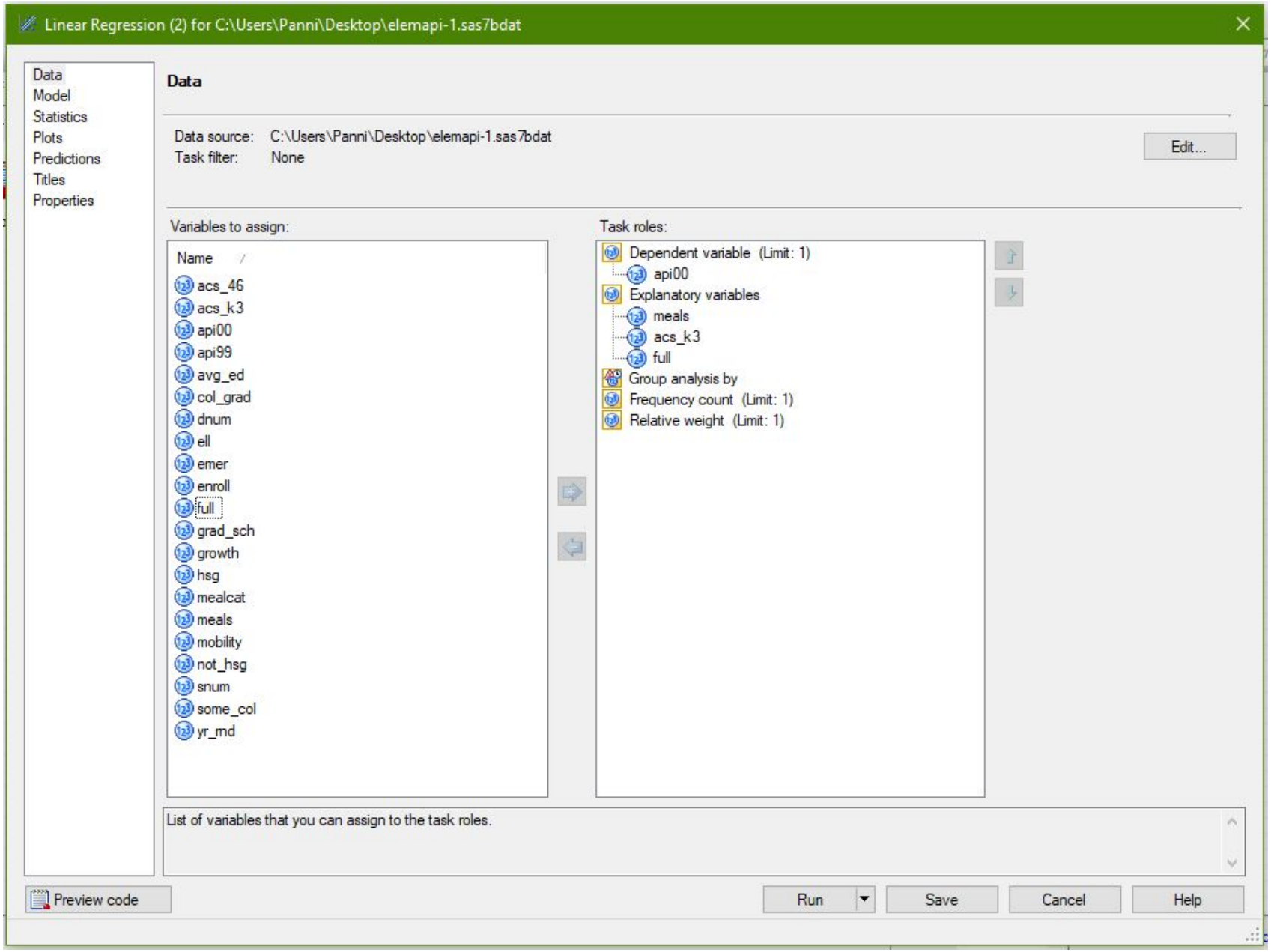
Data Set Attributes



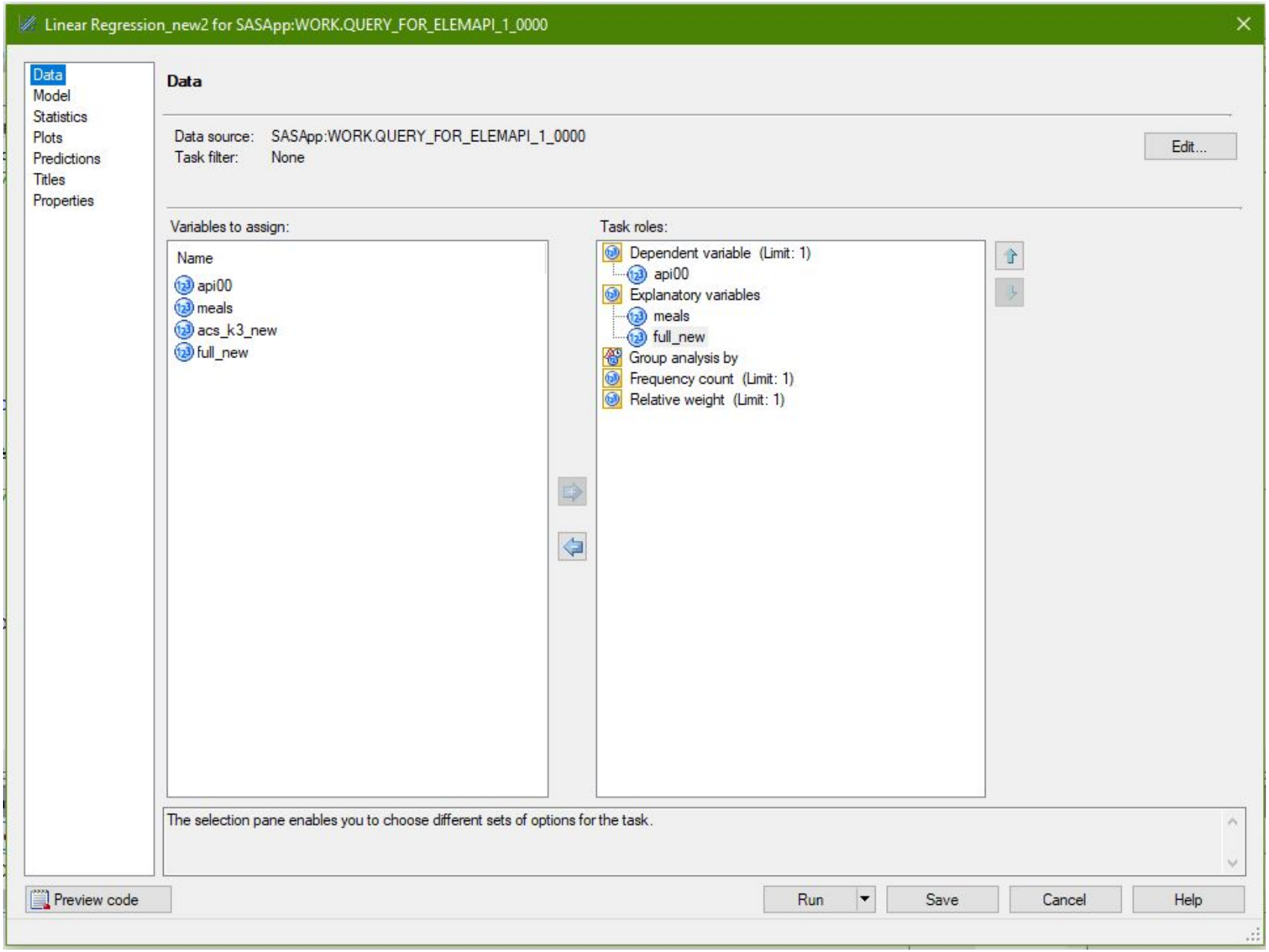
2. lépés – Lineáris regresszió a (talán rossz) adatokra

- H_0: a függő változó nem függ a magyarázó változóktól, azaz a regressziós egyenes meredeksége nulla.
- A futtatás után észrevehetjük, hogy az *acs_k3* és a *full* változók esetén nem utasíthatjuk el a null-hipotézist, azaz ezekkel nem lehet szignifikánsan közelíteni a teljesítmény indexet.
- A regressziós ábrán megfigyelhetjük, hogy érdekes a pontok eloszlása → nézzük meg alaposabban az adatokat!

Linear Regression



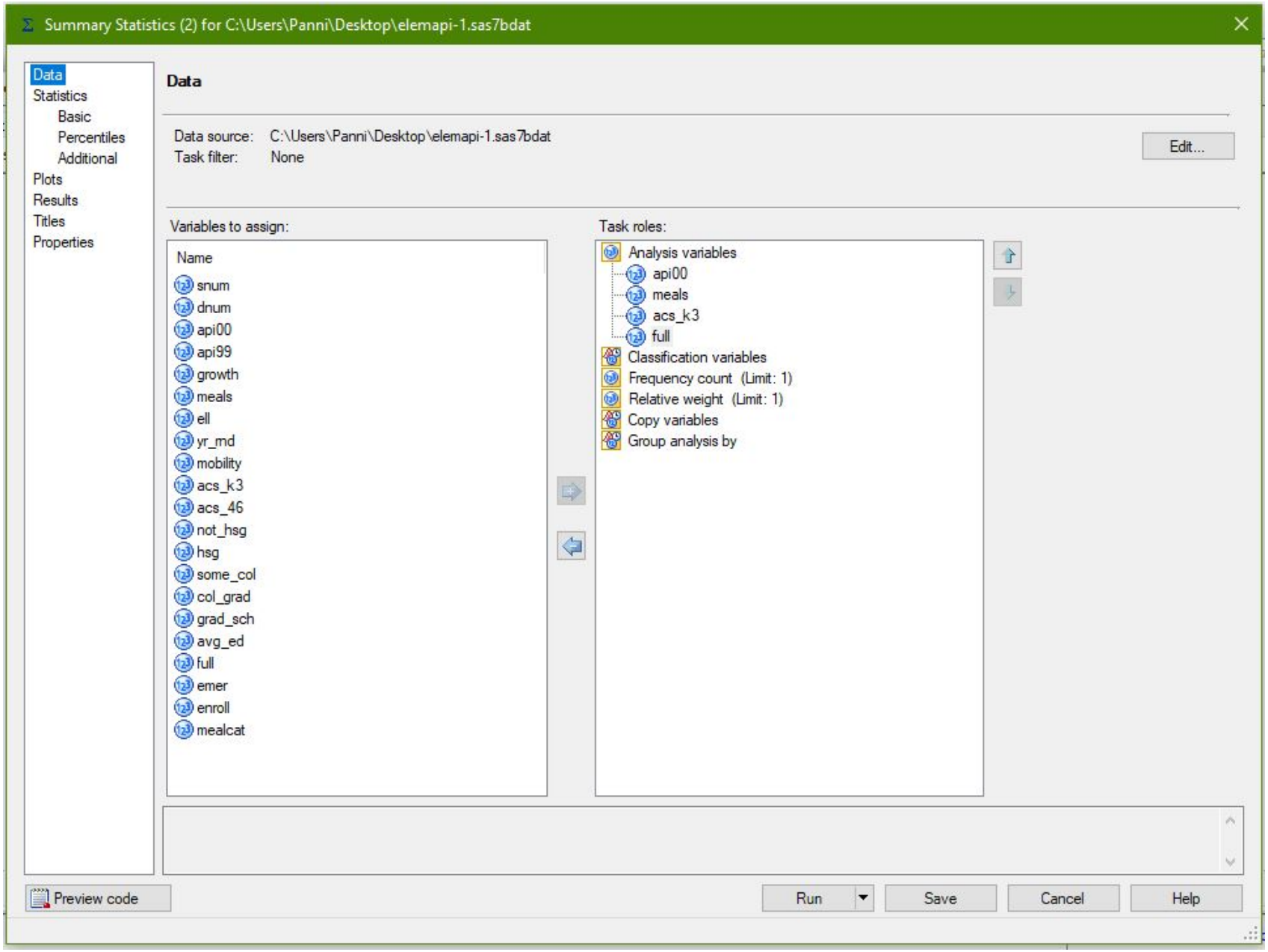
Linear Regression



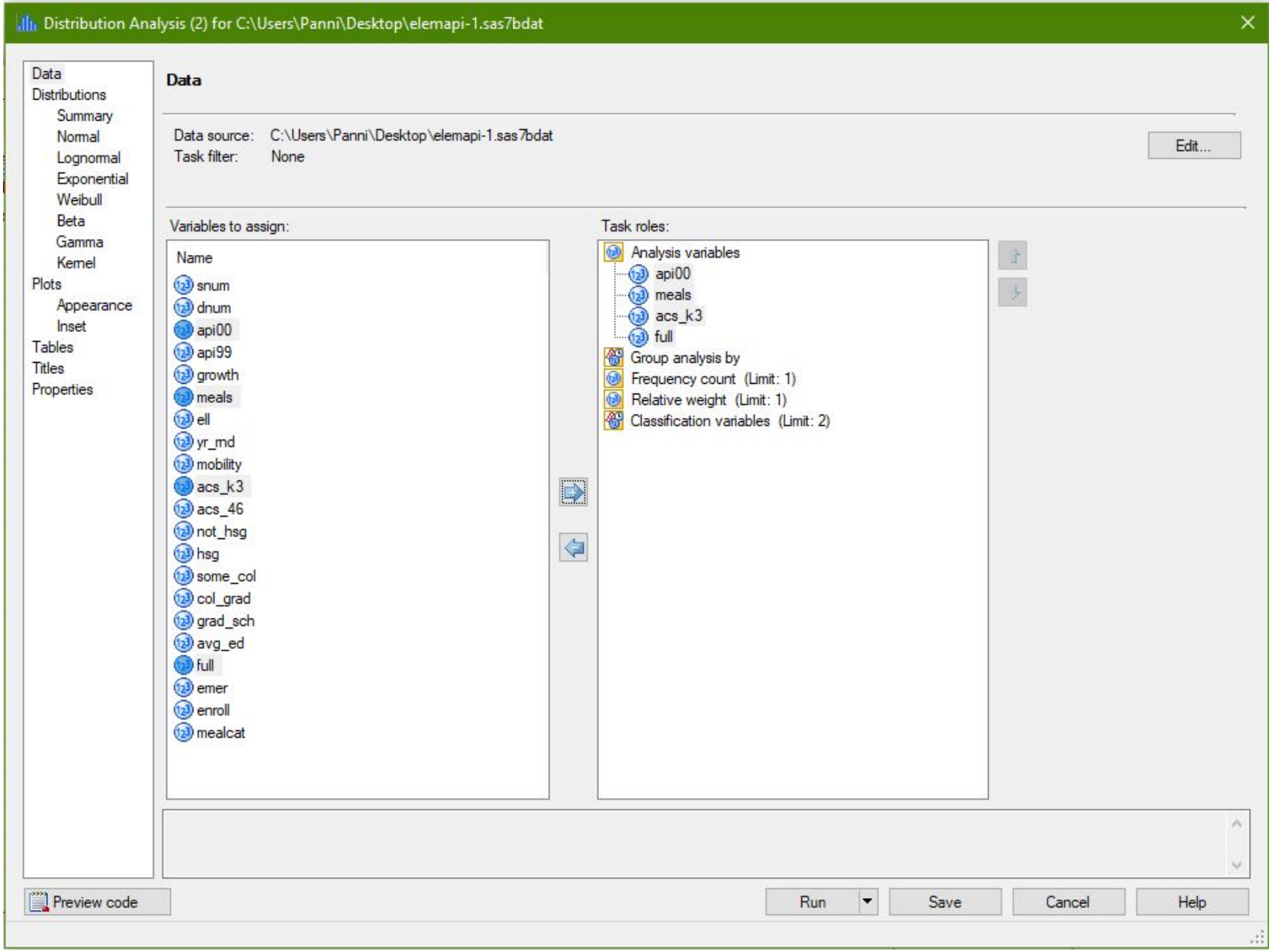
3. lépés – Mélyebb ismerkedés az adatokkal

- Summary Statistics: kapunk egy összefoglalást a felhasználandó változókról (min, max, átlag stb.)
- Figyeljük meg, hogy az *acs_k3* és a *full* változók esetén a minimum érték negatív, ill. nagyon kicsi – osztálylétszám és százalékos érték esetén ez nem tűnik jónak – lehet, hogy hibás az adat?
- Distribution Analysis: Figyeljük meg a fentebb kiemelt két változó hisztogramját!

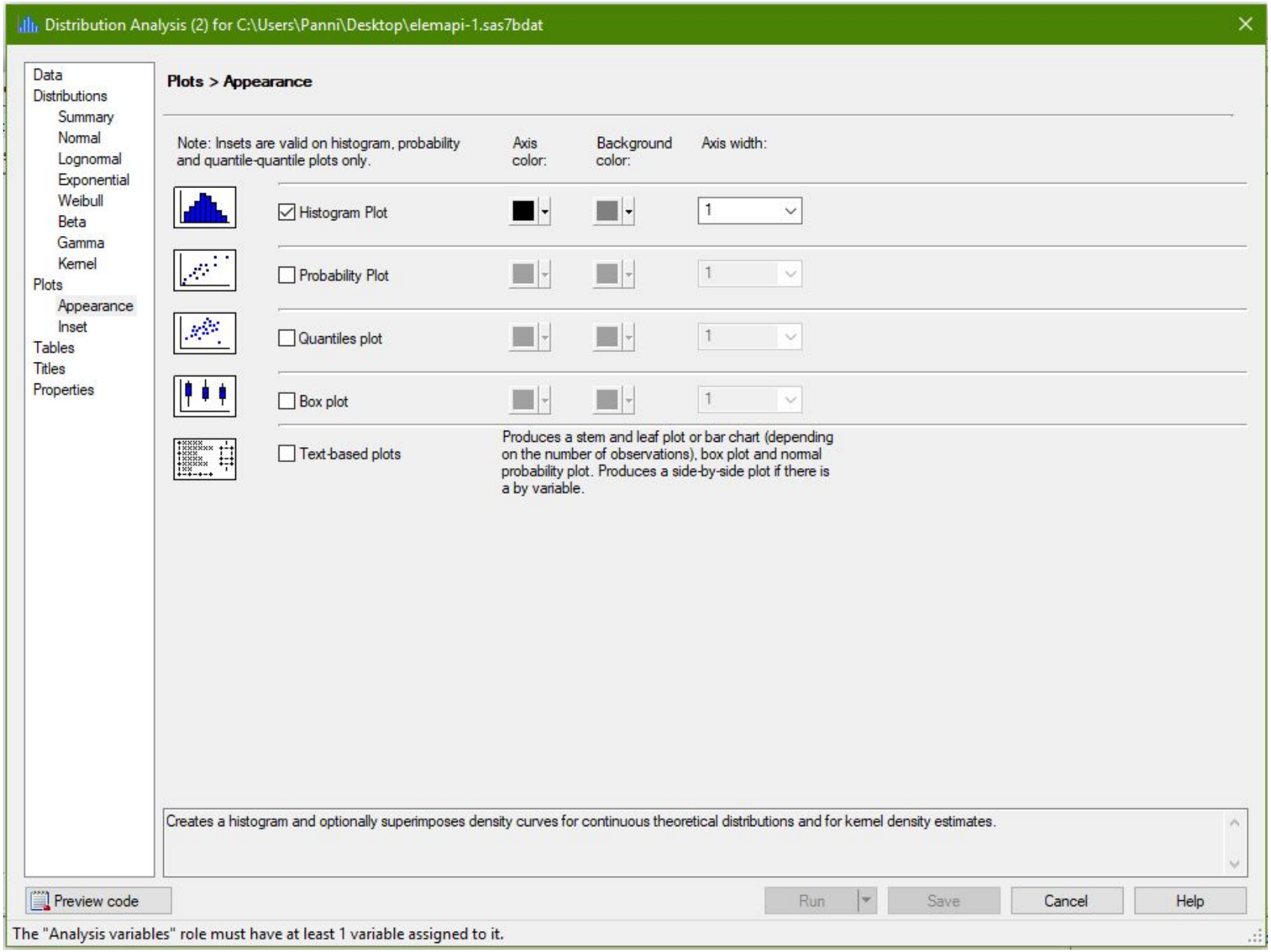
Summary Statistics



Distribution Analysis



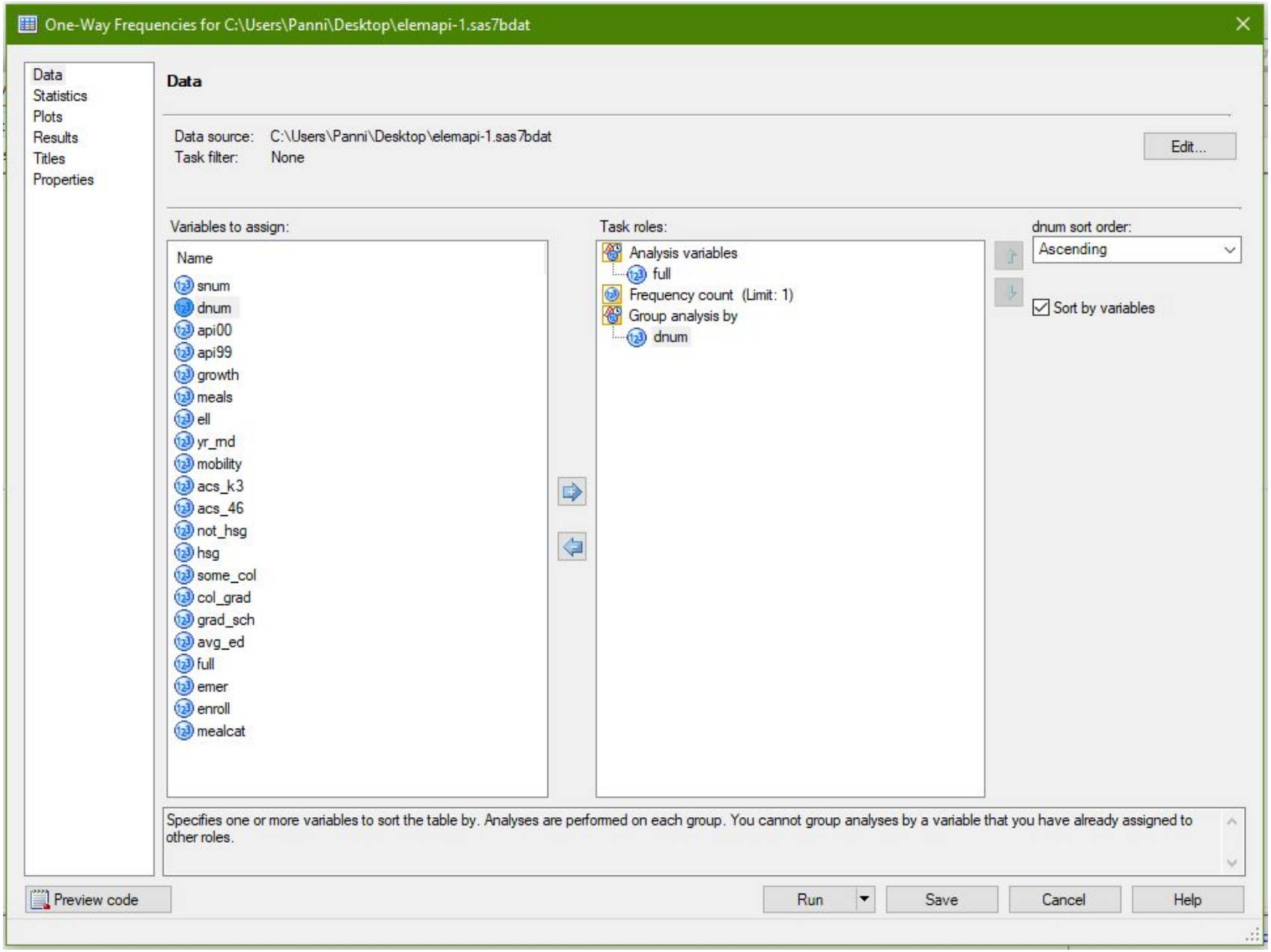
Distribution Analysis



3. lépés – Következtetés

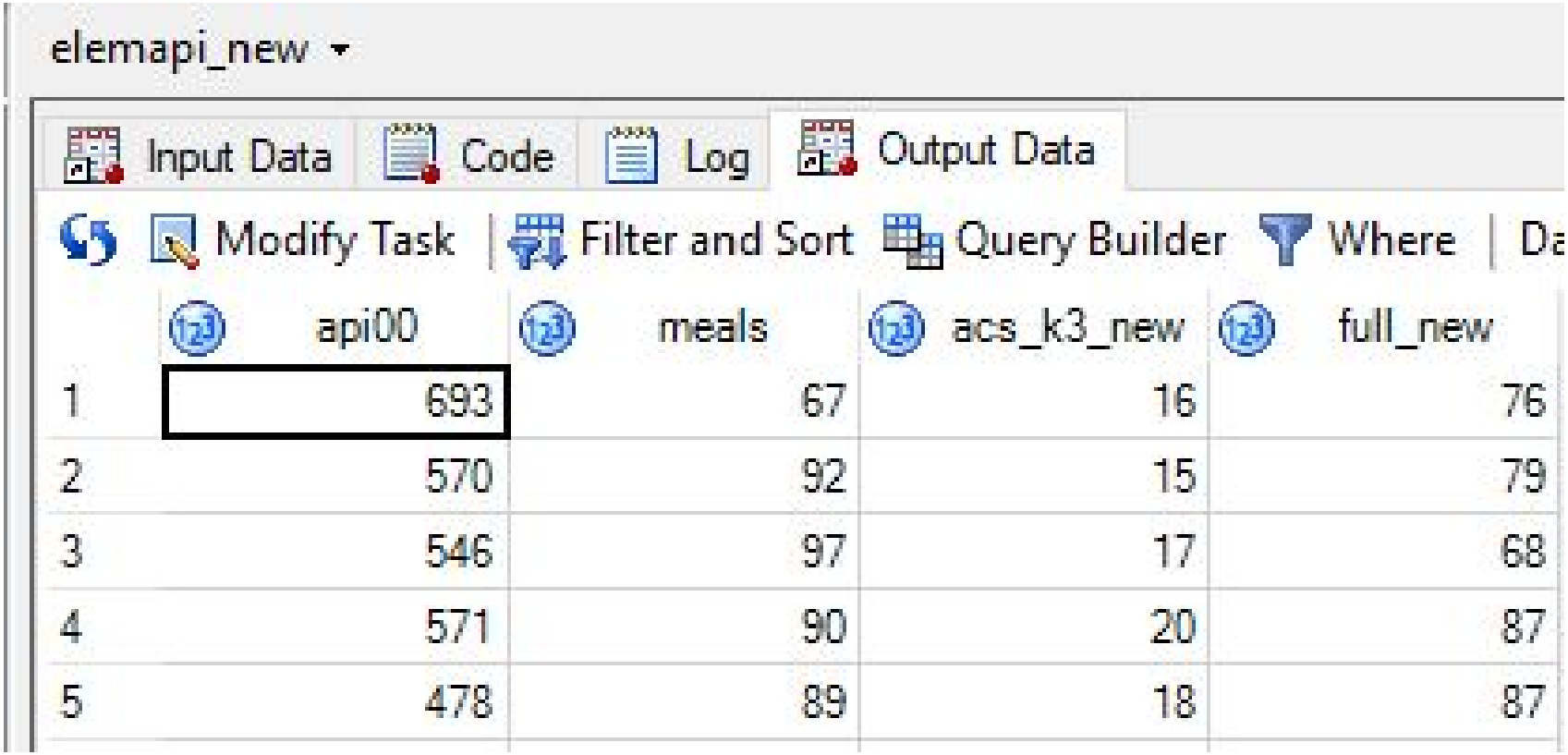
- Az *acs_k3* változó esetén azt gondoljuk, hogy elírás történt, így később az adatsor abszolútértékével fogunk tovább dolgozni.
- A *full* változó esetén készítünk körzetszám szerint egy gyakoriságtáblát, ahol azt fogjuk látni, hogy minden körzet 1 – 100-ig terjedő skálán adta meg a tanárok %-os értékelését, kivéve a 401-es körzet – így ebben az esetben értelmezési hibáról beszélhetünk.
- Megoldás: Az 1-nél kisebb adatokat megszorozzuk 100-zal.

One-Way Frequencies



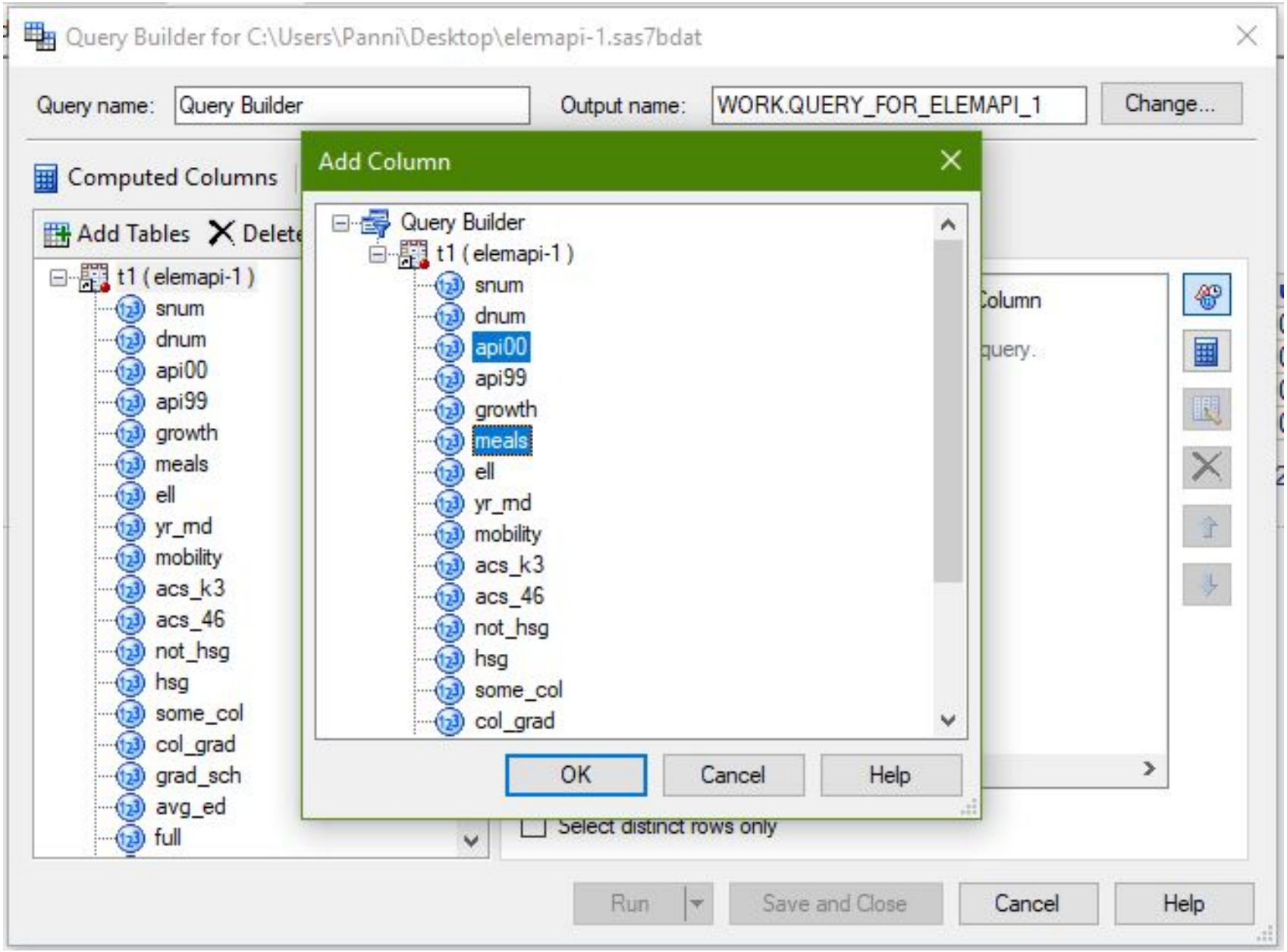
4. lépés – Adatsorok kijavítása

- A fentebb leírt módon létrehozunk egy új adattáblát a kijavított adatokkal.

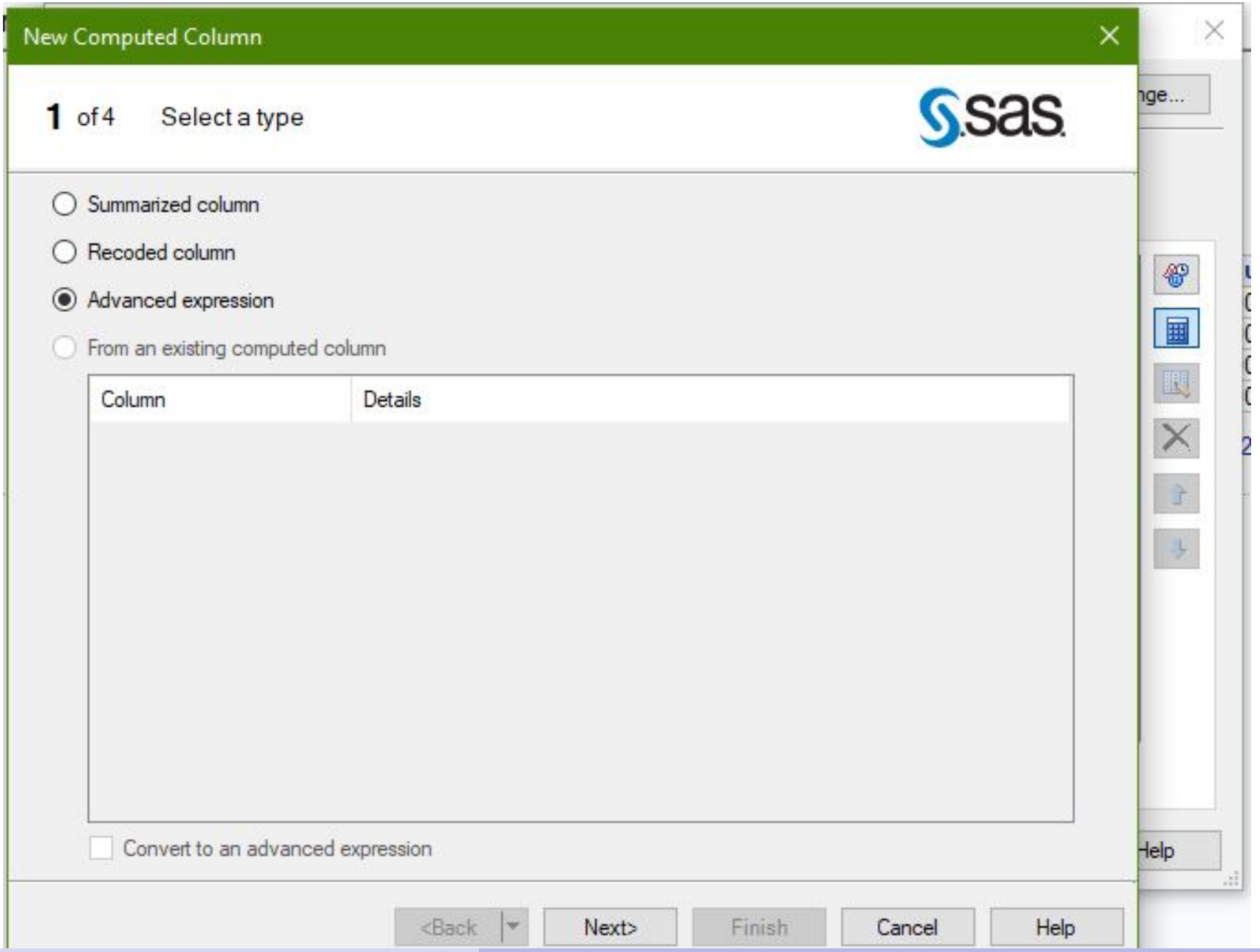


	api00	meals	acs_k3_new	full_new
1	693	67	16	76
2	570	92	15	79
3	546	97	17	68
4	571	90	20	87
5	478	89	18	87

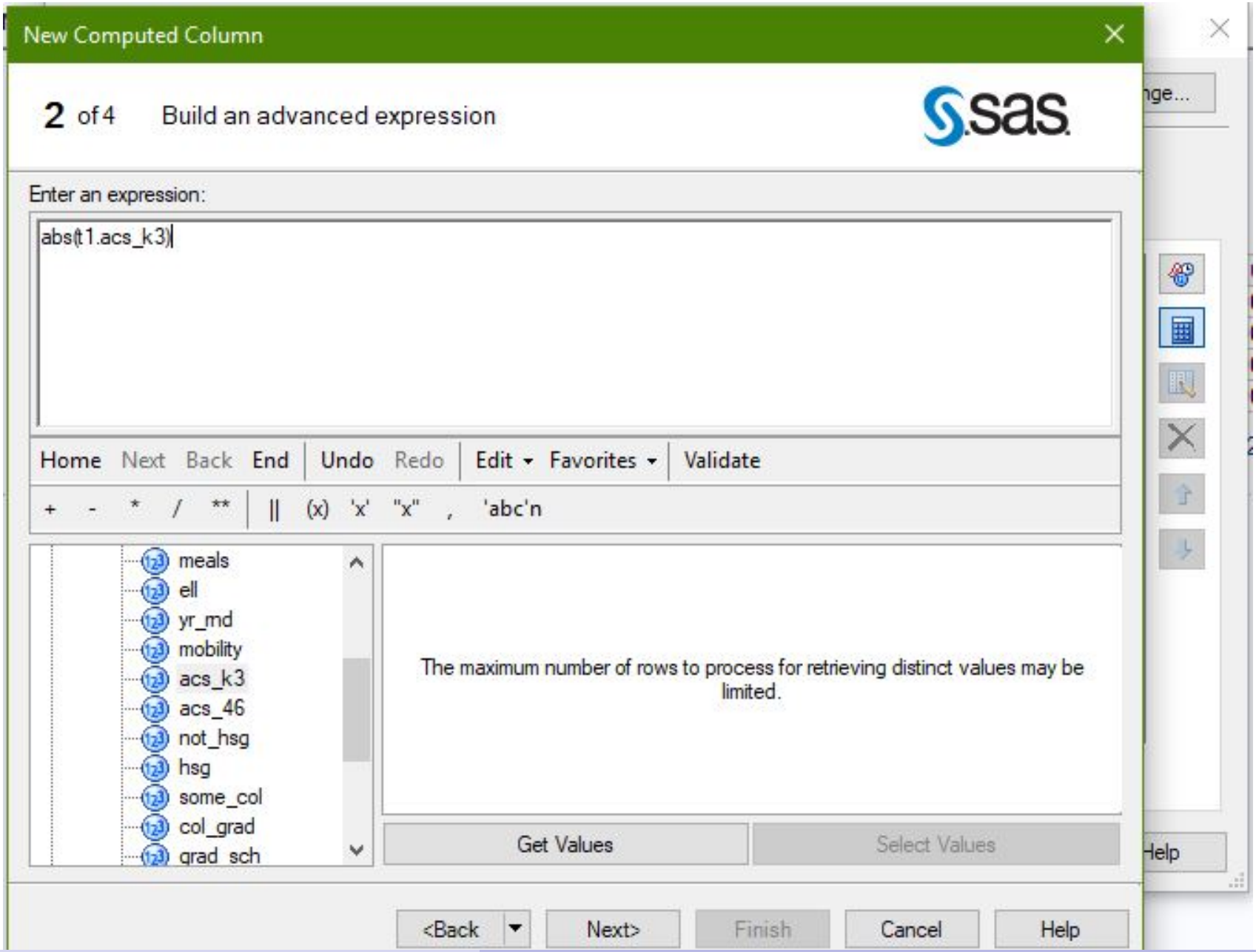
Query Builder



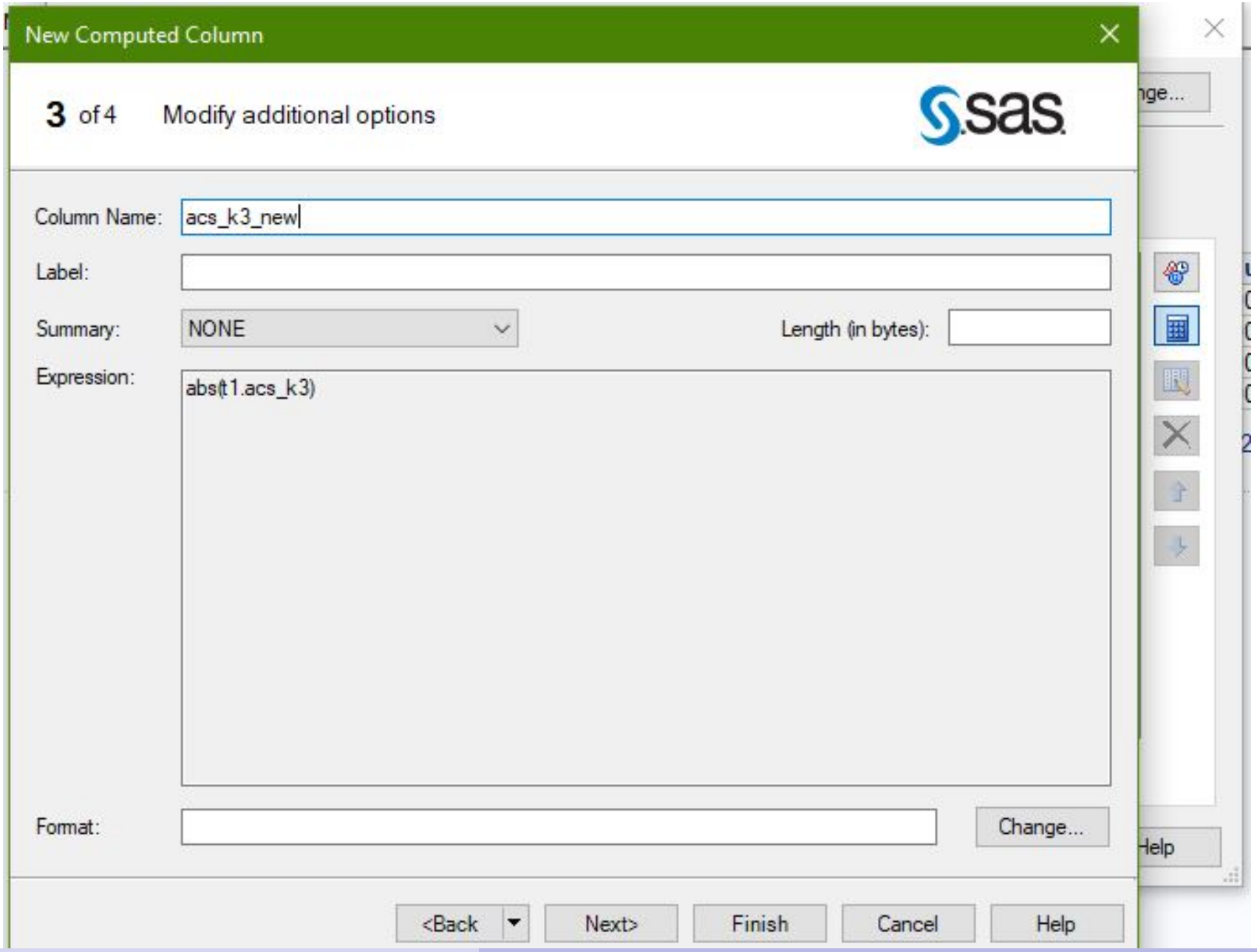
Query Builder



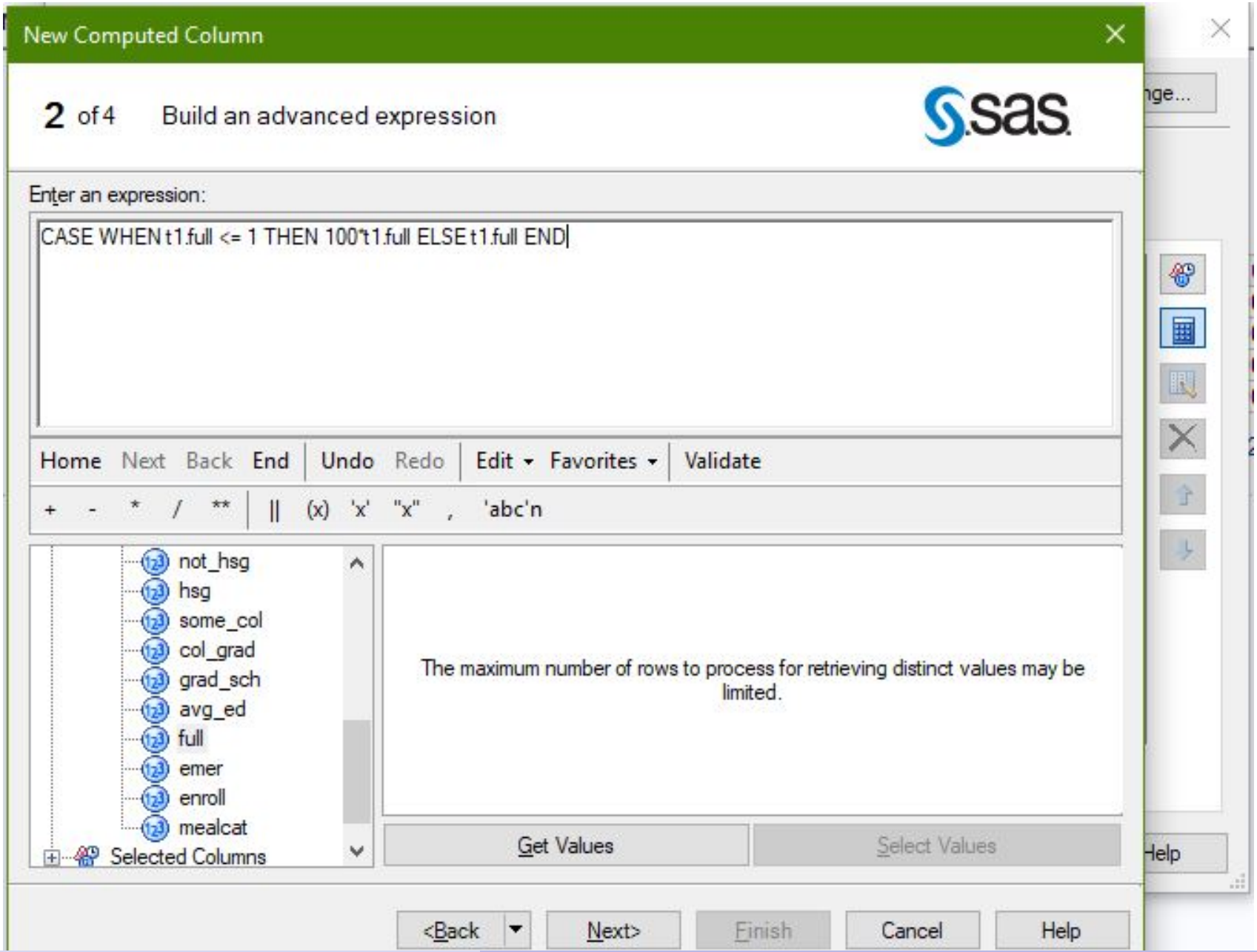
Query Builder



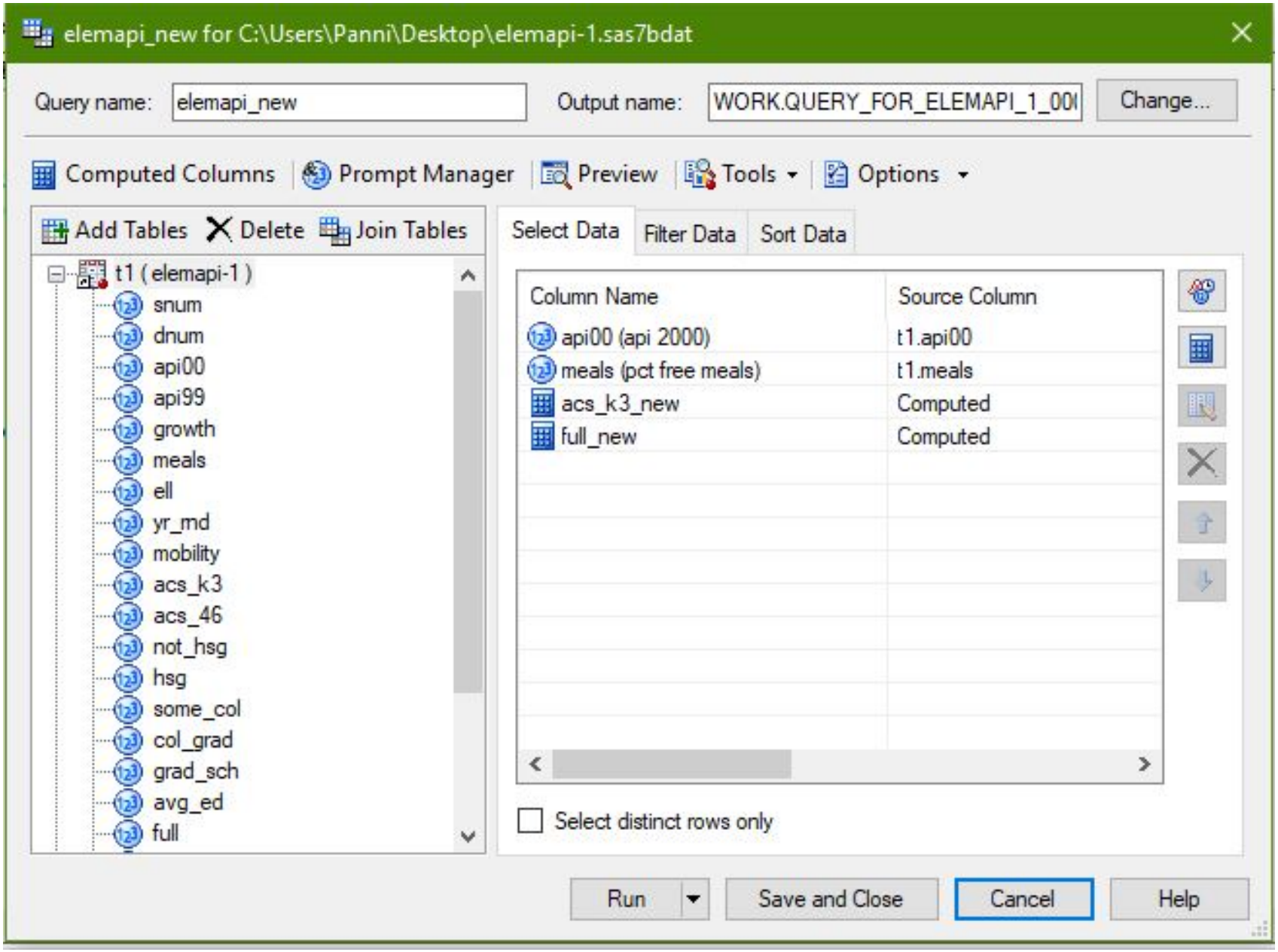
Query Builder



Query Builder



Query Builder



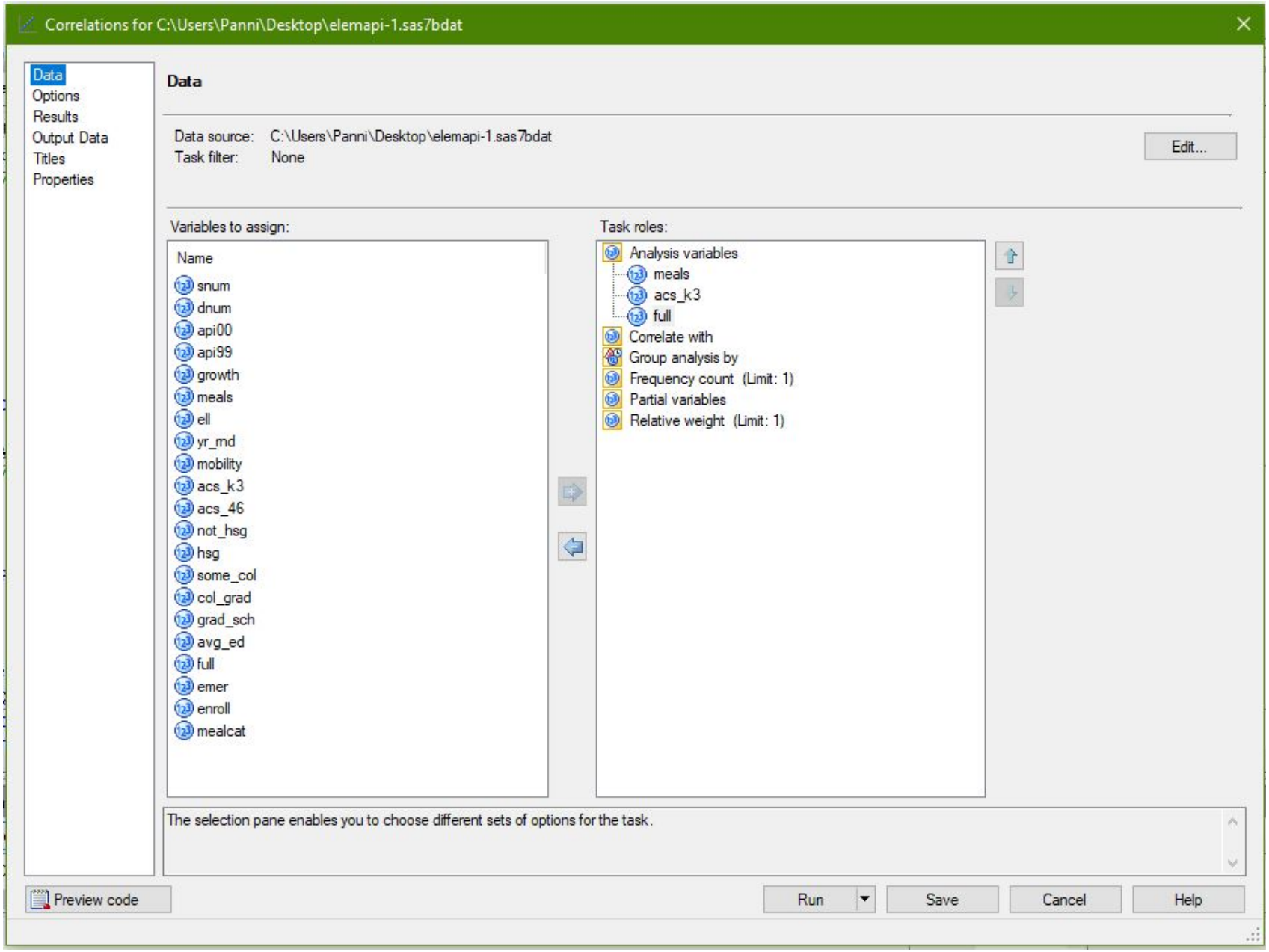
5. lépés – Az új adatok elemzése I.

- Az új adatokra ismét futtatunk egy Distribution Analysis-t, ahol láthatjuk, hogy reálisabb adatsort kaptunk.
- Futtatunk egy lineáris regressziót a jó adatokra, ahol láthatjuk, hogy nőtt az R^2, azaz a nagyobb a regresszió varianciájának a megmagyarázott része $\rightarrow$ jobb a regresszió!
- Viszont a hipotézisvizsgálatnál az osztálylétszám még mindig nem szignifikáns magyarázó változó $\rightarrow$ hagyjuk el a regresszióból!
- Ekkor a modell magyarázó ereje nem csökkent jelentősen, viszont kevesebb adatot (változót) használtunk $\rightarrow$ így ez az eddigi legjobb regressziónk!

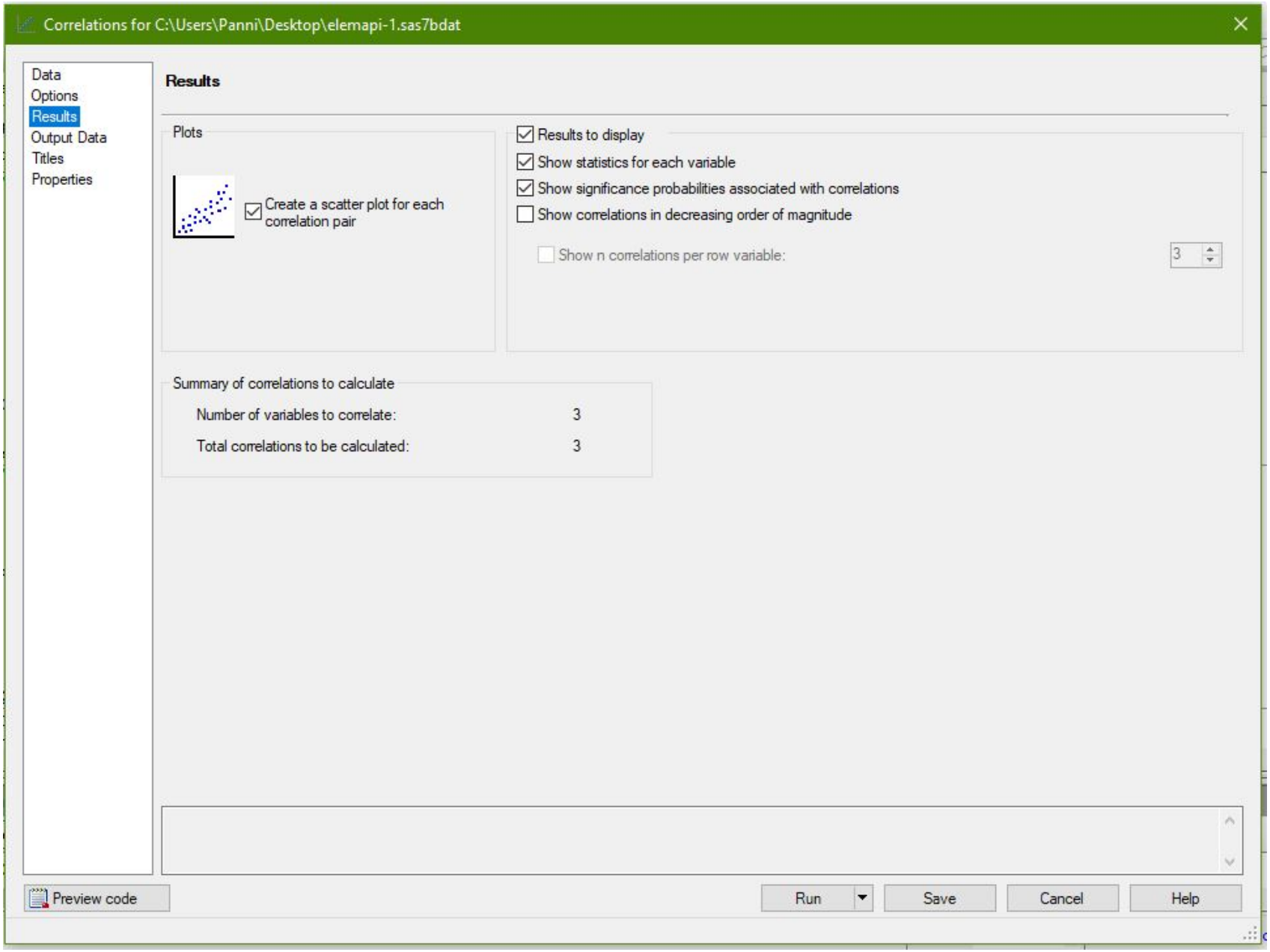
5. lépés – Az új adatok elemzése II.

- Az *acs_k3* elhagyhatóságából következik, hogy ezt a változót megmagyarázza a másik kettő.
- Most vizsgáljuk meg a korrelációt és vegyük észre, hogy ez valóban így van.

Correlation



Correlation



Vége :)

Köszönjük a figyelmet!