

UNIVERZITET U BEOGRADU

MATEMATIČKI RAD

SEMINARSKI RAD IZ STATISTIČKOG SOFTVERA 3 NA TEMU:

TEŽINA I VISINA ISPITANIKA

Asistent:

Marija Radičević

Student:

Jovana Protić 162/2012

Beograd, jun 2016. godine

Sadržaj:

1. Uvod	3
2. Učitavanje baze u SPSS i prilagođavanje radu	3
3. Izračunati BMI za izmerene vrednosti ispitanika.....	8
4. Izračunati prosečnu vrednost težine i visine za oba pola:	9
a. na osnovu izmerenih vrednosti	
b. na osnovu prijavljenih vrednosti	
5. Na osnovu BMI vrednosti izdvojiti muškarce sa idealnom masom	11
6. Da li je srednja vrednost visine naših ispitanika jednaka u slučaju muškaraca i žena?.....	13
7. Testirati vezu između izračunate i prijavljene težine.....	17
8. Napraviti najbolji linearni model za određivanje BMI-a u odnosu na ostale promenljive.....	19
9. Literatura.....	23

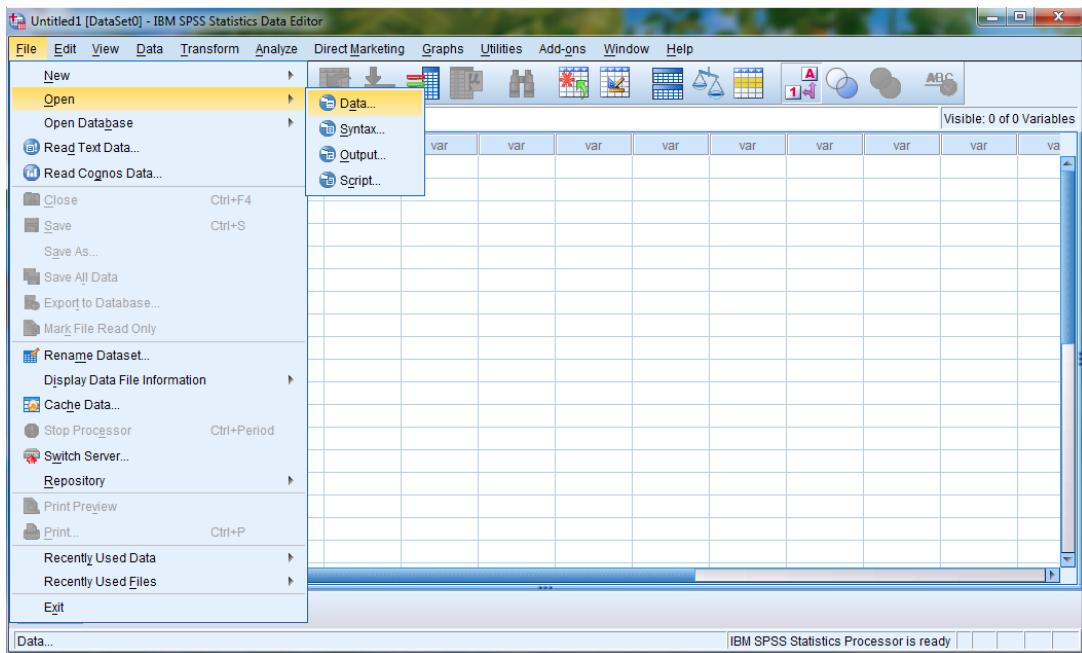
1. UVOD

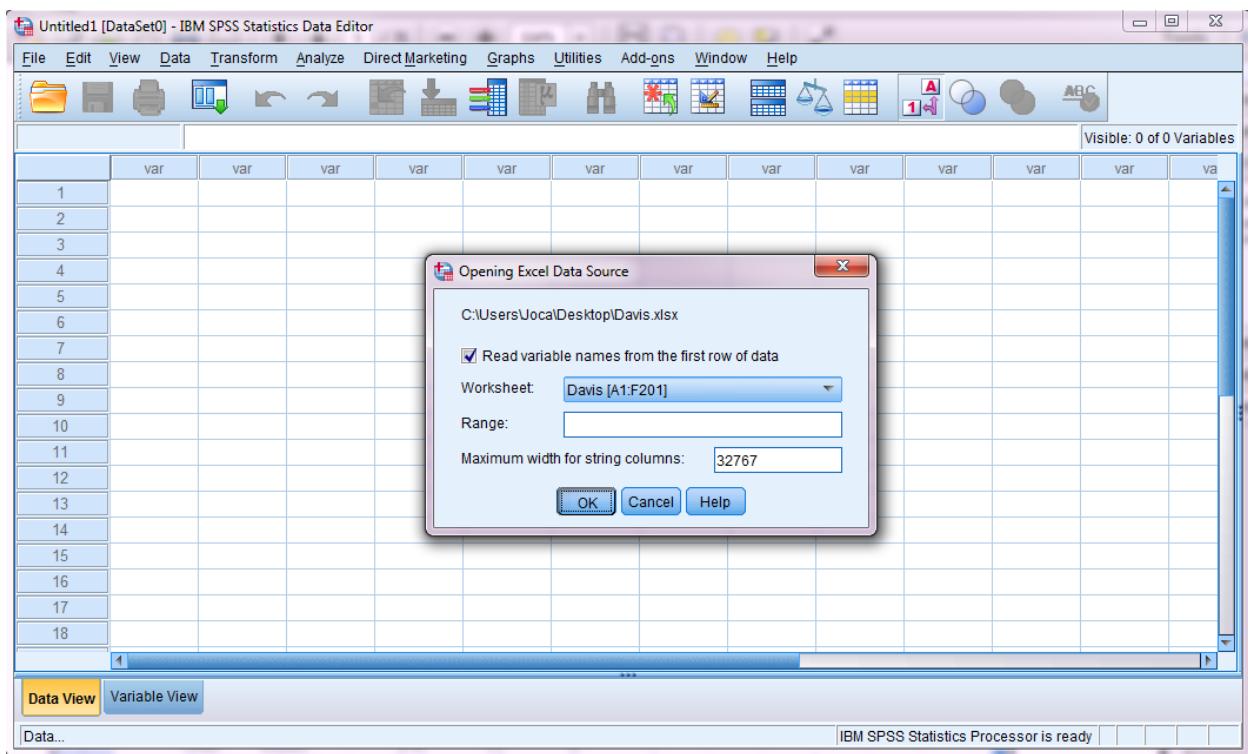
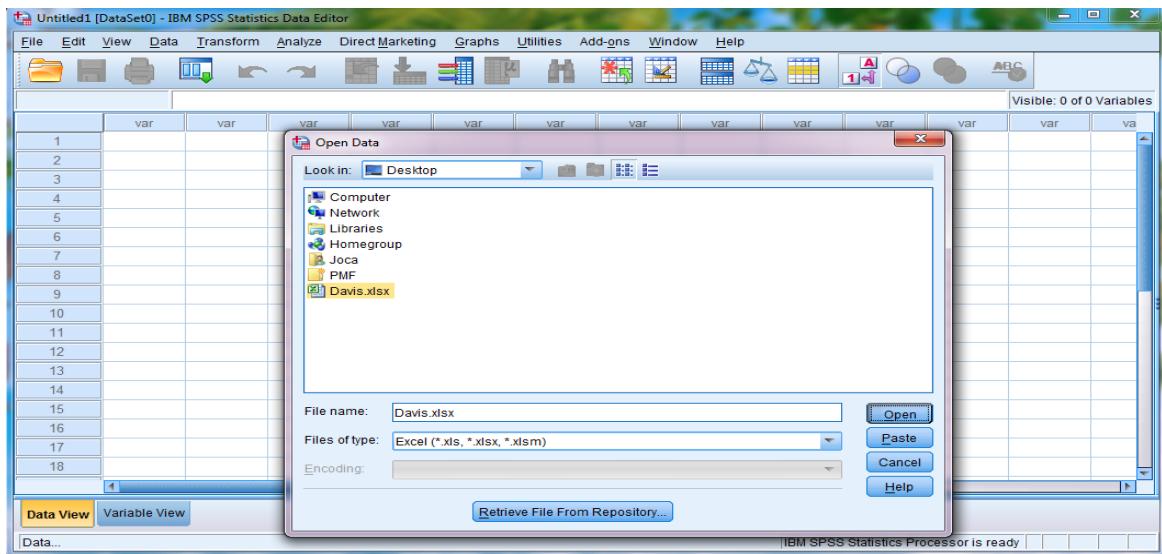
Radićemo sa bazom *Davis* koja u sebi sadrži podatke o tačnoj težini i visini naših ispitanika, kao i o vrednostima koje su ispitanici prijavili za svoju težinu i visinu. Baza se sastoji od 200 opservacija i 5 promenljivih. Promenljive su:

- **Sex-** pol naših ispitanika, F označava da je u pitanju žena, M da je u pitanju muškarac
- **Weight-** izmerena težina u kilogramima
- **Height-** izmerena visina u centimetrima
- **Repwt-** prijavljena težina u kilogramima
- **Reph-** prijavljena visina u centimetrima

2. UČITAVANJE BAZE U SPSS I PRILAGOĐANJE RADU

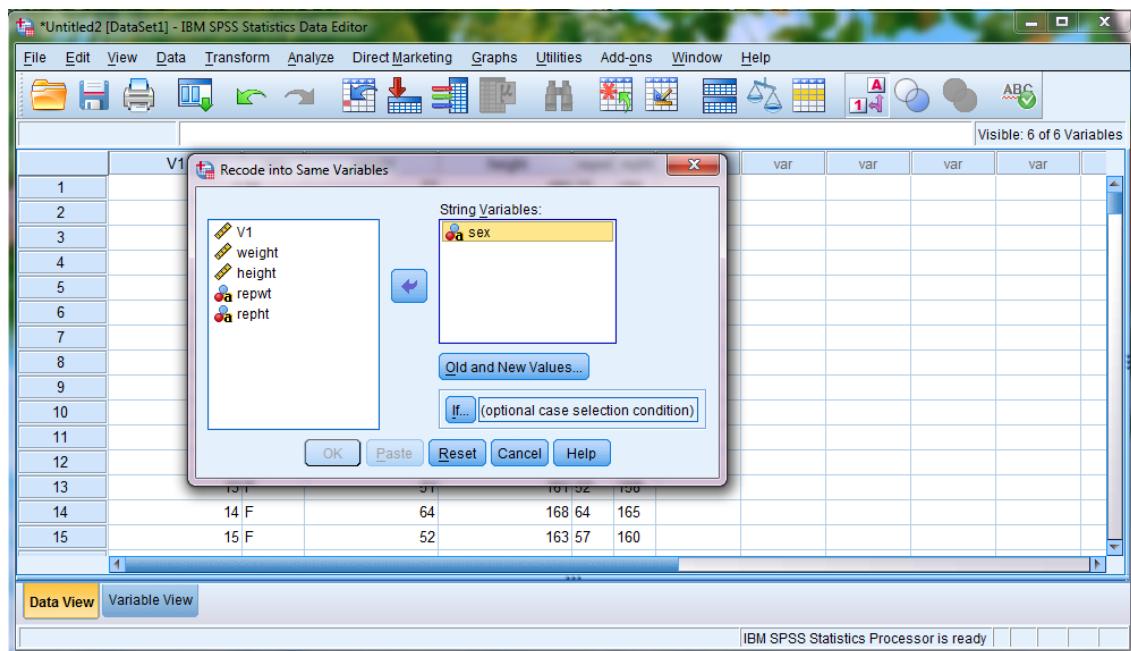
File → Open → Data...



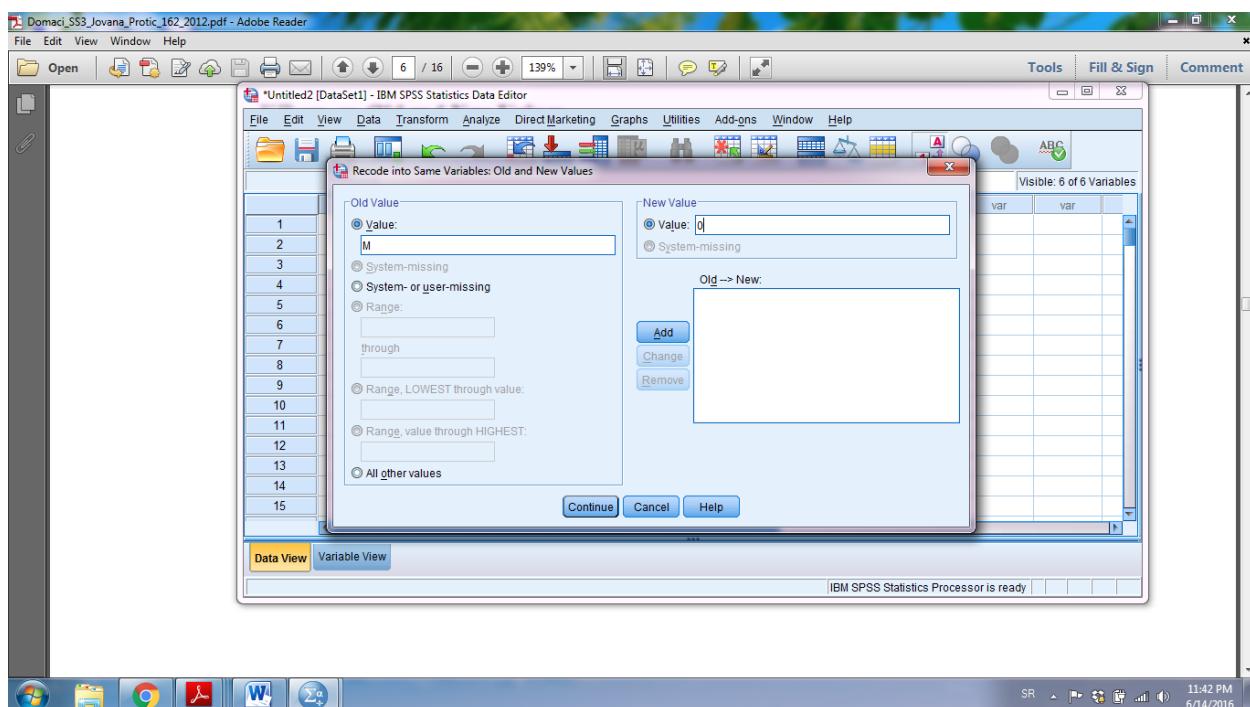


I ovim smo završili učitavanje naše baze. Sada, kada smo učitali našu bazu, vidimo da je promenljiva *sex* tipa String, a kako u SPSS-u možemo da radimo samo sa numeričkim promenljivim prilikom statističkih analiza, izvršićemo korekciju naše baze na sledeći način:

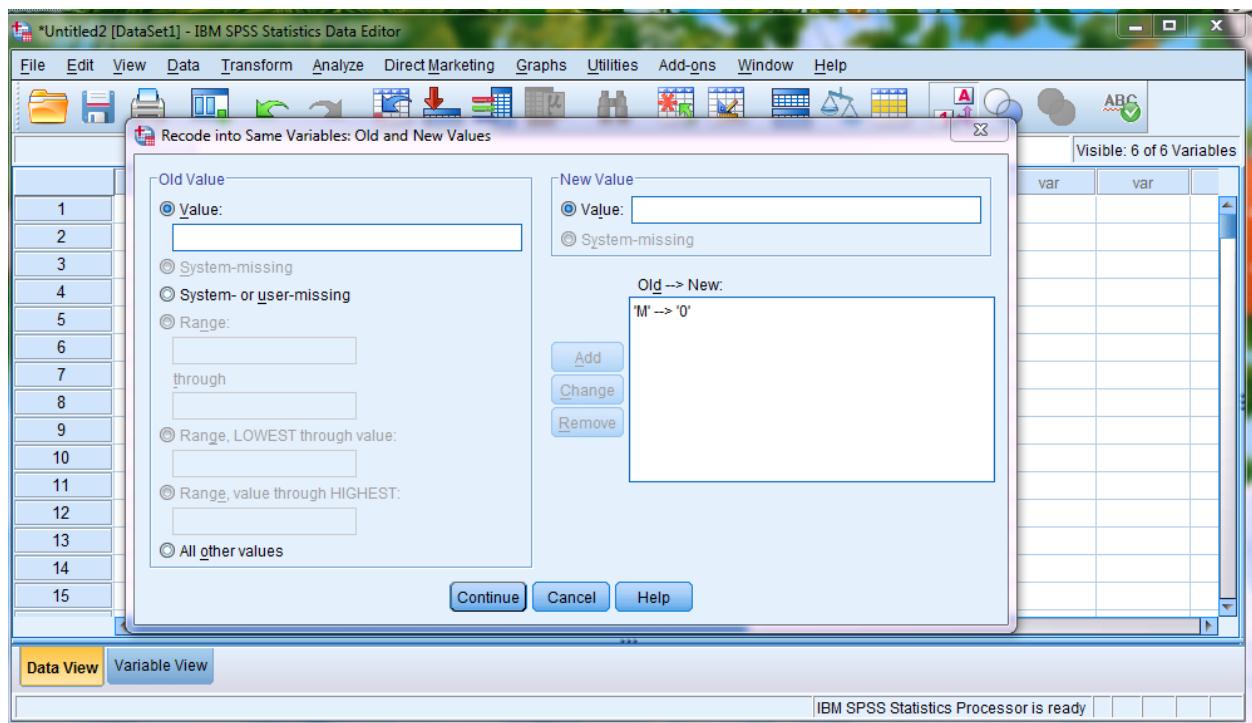
Transform → Recode into same variables...



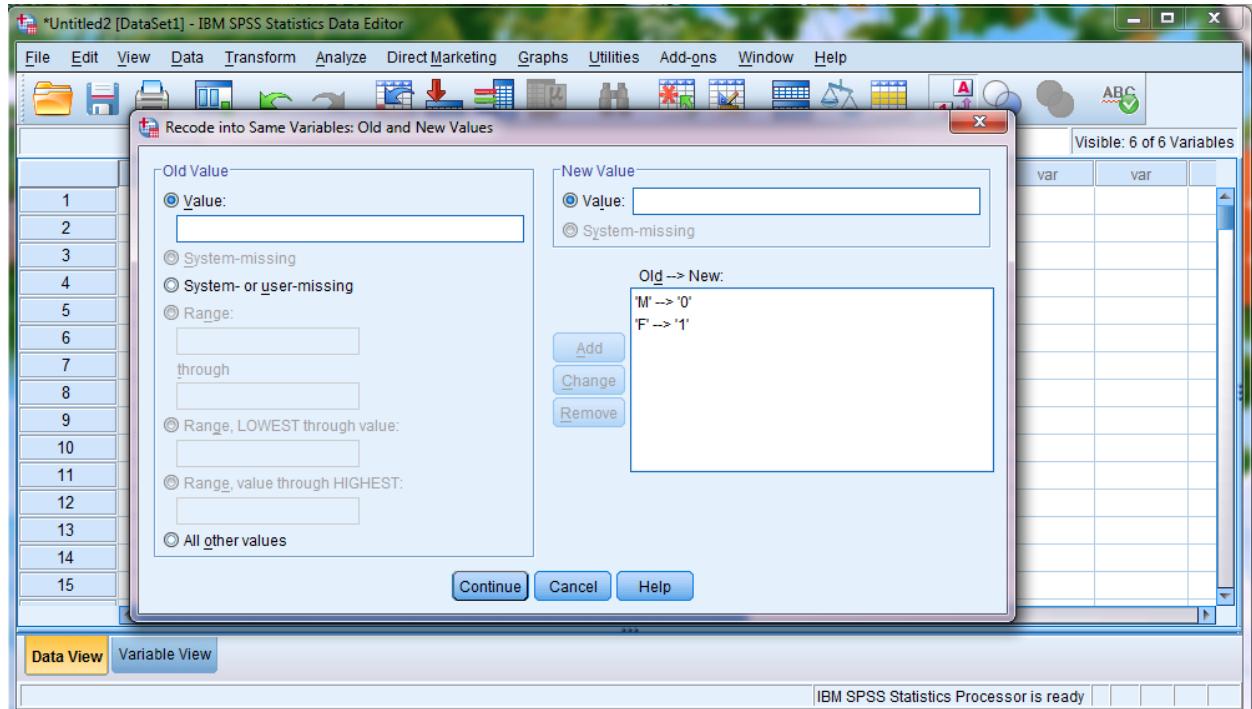
Sada klikom na *Old and New Values...*



Klikom na *Add* dobijamo:



Slično odradimo i za ženski pol pri čemu dobijamo:



Sada možemo promeniti i tip promenljive:

*Untitled2 [DataSet1] - IBM SPSS Statistics Data Editor

	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure
1	V1	Numeric	12	0		None	None	12	Right	Scale
2	sex	String	1	0		None	None	5	Left	Nominal
3	weight	Numeric	12	0		None	None	12	Right	Scale
4	height	Numeric	12	0		None	None	12	Right	Scale
5	repwt	String	3	0		None	None	3	Left	Nominal
6	rept	String	3	0		None	None	3	Left	Nominal
7										
8										
9										
10										
11										
12										
13										
14										
15										
16										
17										

Data View Variable View

IBM SPSS Statistics Processor is ready

Klikom na "... " pojavljuje nam se sledeći prozor:

*Untitled2 [DataSet1] - IBM SPSS Statistics Data Editor

	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure
1	V1	Numeric							Right	Scale
2	sex	String							Left	Nominal
3	weight	Numeric							Right	Scale
4	height	Numeric							Right	Scale
5	repwt	String							Left	Nominal
6	rept	String							Left	Nominal
7										
8										
9										
10										
11										
12										
13										
14										
15										
16										
17										

Variable Type

OK Cancel Help

The Numeric type honors the digit grouping setting, while the Restricted Numeric never uses digit grouping.

Data View Variable View

IBM SPSS Statistics Processor is ready

Čekiranjem *Numeric* naša promenljiva postaje numeričkog tipa.

Kako su nam i promenljive *repwt* i *rept* takođe tipa String i to ćemo promeniti sličnim postupkom kao malopre. Sada naša baza izgleda na sledeći način:

The screenshot shows the IBM SPSS Statistics Data Editor window. The menu bar includes File, Edit, View, Data, Transform, Analyze, Direct Marketing, Graphs, Utilities, Add-ons, Window, and Help. The toolbar contains various icons for file operations like Open, Save, Print, and Data manipulation. The main data grid displays the following data:

	V1	sex	weight	height	repwt	reph	var	var	var	var	var
1		1	0	77	182	77	180				
2		2	1	58	161	51	159				
3		3	1	53	161	54	158				
4		4	0	68	177	70	175				
5		5	1	59	157	59	155				
6		6	0	76	170	76	165				
7		7	0	76	167	77	165				
8		8	0	69	186	73	180				
9		9	0	71	178	71	175				
10		10	0	65	171	64	170				
11		11	0	70	175	75	174				
12		12	1	166	57	56	163				
13		13	1	51	161	52	158				
14		14	1	64	168	64	165				
15		15	1	52	163	57	160				

The status bar at the bottom of the SPSS window indicates "IBM SPSS Statistics Processor is ready". The taskbar at the bottom of the screen shows the Windows Start button, Task View, Google Chrome, File Explorer, and other icons.

3. BMI ZA IZMERENE VREDNOSTI ISPITANIKA

Formiraćemo promenljivu BMI koja nam prikazuje indeks telesne mase na osnovu izmerenih vrednosti težine i visine:

Transform → Compute variable...

The screenshot shows the 'Compute Variable' dialog box. The 'Target Variable:' field is set to 'BMI'. The 'Numeric Expression:' field contains the formula $\text{weight} / ((\text{height} * 0.01) * (\text{height} * 0.01))$. The 'Function group:' dropdown is set to 'All'. The 'Functions and Special Variables:' panel is empty. At the bottom, there is an 'optional case selection condition' field and buttons for OK, Paste, Reset, Cancel, and Help.

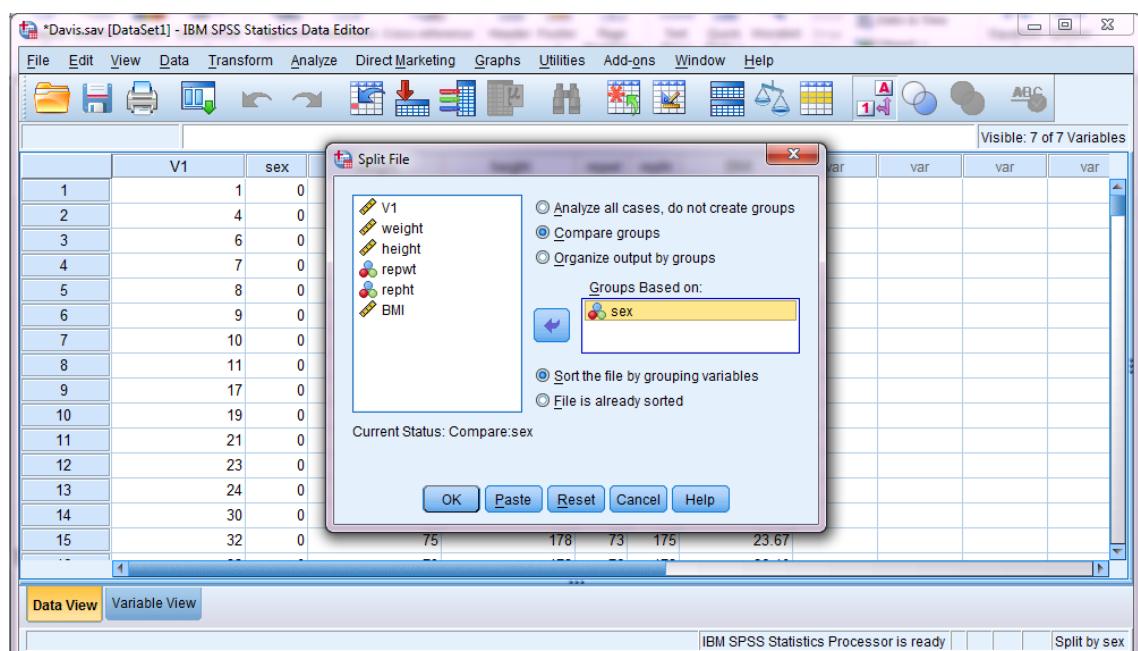
Sada naša baza izgleda:

	V1	sex	weight	height	repwt	rept	BMI	var	var	var	var
1		1	0	77	182	77	180	23.25			
2		2	1	58	161	51	159	22.38			
3		3	1	53	161	54	158	20.45			
4		4	0	68	177	70	175	21.71			
5		5	1	59	157	59	155	23.94			
6		6	0	76	170	76	165	26.30			
7		7	0	76	167	77	165	27.25			
8		8	0	69	186	73	180	19.94			
9		9	0	71	178	71	175	22.41			
10		10	0	65	171	64	170	22.23			
11		11	0	70	175	75	174	22.86			
12		12	1	166	57	56	163	510.93			
13		13	1	51	161	52	158	19.68			
14		14	1	64	168	64	165	22.68			
15		15	1	52	163	57	160	19.57			

4. IZRAČUNATI PROSEČNU VREDNOST VISINE I TEŽINE ZA OBA POLA

Prvo ćemo da splitujemo podatke:

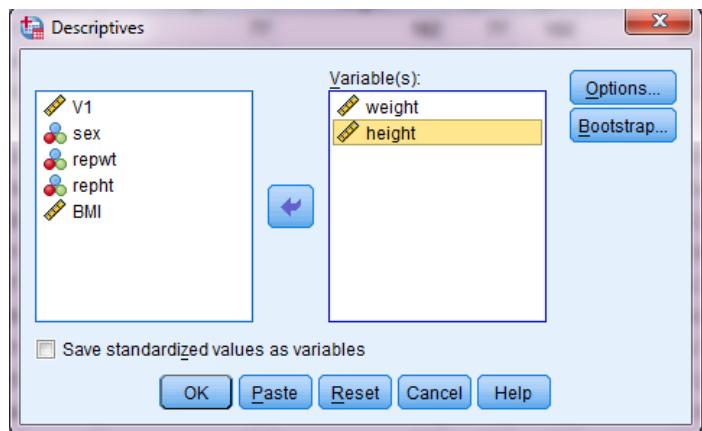
Data → Split File...



Splitovali smo u odnosu na pol, pa sada možemo nastaviti dalje sa radom.

a) U ODNOSU NA IZMERENE PODATKE

Analyze → Descriptive Statistics → Descriptives



Pri čemu dobijamo sledeće podatke:

Descriptive Statistics

sex		N	Minimum	Maximum	Mean	Std. Deviation
0	weight	88	54	119	75.90	11.890
	height	88	163	197	178.01	6.441
	Valid N (listwise)	88				
1	weight	112	39	166	57.87	12.383
	height	112	57	178	163.74	11.644
	Valid N (listwise)	112				

Vidimo da broj muškaraca među našim ispitanicima je 88, dok imamo 112 žena. Takođe vidimo da je minimalna težina muškaraca 54kg, dok najteži muškarac ima 119kg. Kod žena vidimo da najlakša žena ima 39kg, dok najteža ima 166kg. Što se visine tiče, kod muškaraca najniži ima 163cm, najviši 197cm. Dok najniža žena ima 57cm, a najviša 178cm. Kod muškaraca, prosečna težina je 75,9kg, a visina je 178,01cm. Dok su kod žena te vrednosti 57,87kg i 163,74cm.

b) U ODNOSU NA PRIJAVLJENE VREDNOSTI

Sličnim postupkom kao pod a) dobijamo sledeće vrednosti:

Descriptive Statistics

sex	N	Minimum	Maximum	Mean	Std. Deviation
0 repwt	82	56	124	76.56	12.287
repht	82	161	200	176.26	6.978
Valid N (listwise)	82				
1 repwt	101	41	77	56.74	6.737
repht	101	148	176	162.20	5.852
Valid N (listwise)	99				

Kako je u ovom slučaju broj muškaraca manji od ukupnog broja muškaraca, možemo zaključiti da nisu svi muškarci prijavili vrednosti za težinu i visinu. Isto se može zaključiti i u slučaju ženskog pola.

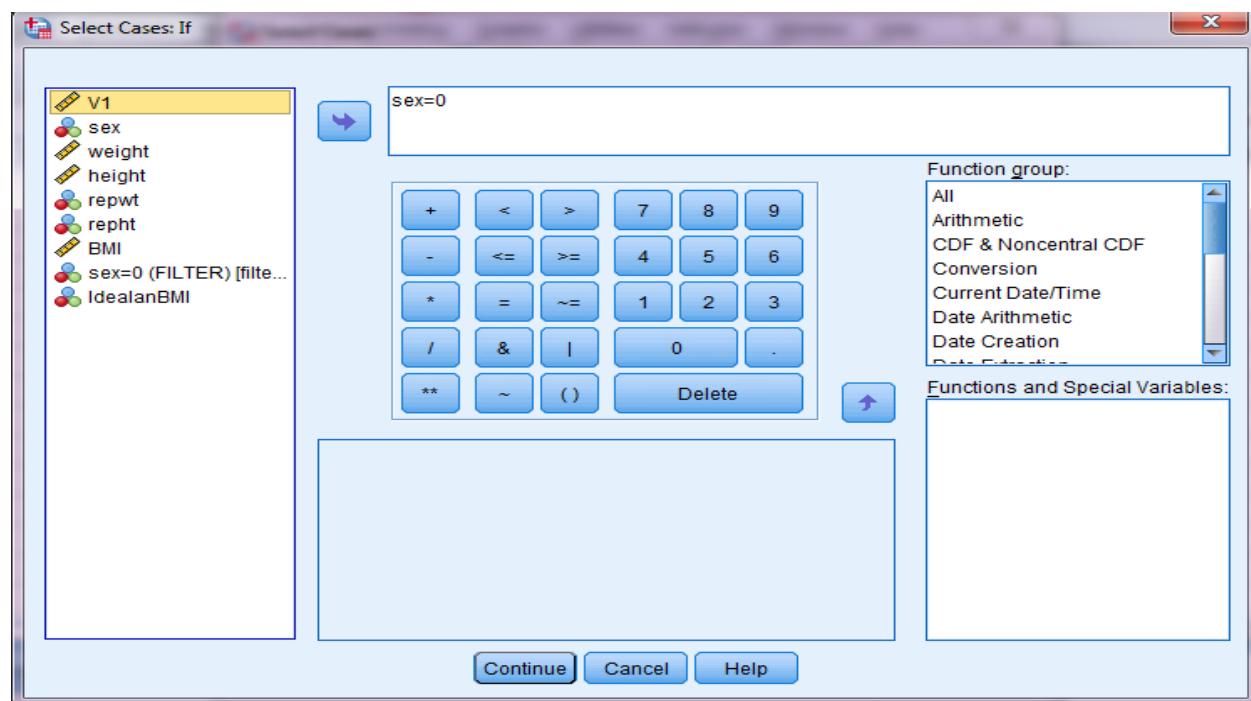
5. NA OSNOVU BMI IZDVOJITI MUŠKARCE SA IDEALNOM MASOM

Na osnovu tablice za BMI vrednosti, vidimo da idealnu masu imaju oni ispitanici čija je BMI vrednost između 18,5 i 24,9. Sada ćemo izdvojiti taj niz naših ispitanika.

Prvo treba da selektujemo samo muškarce, a to radimo na sledeći način:

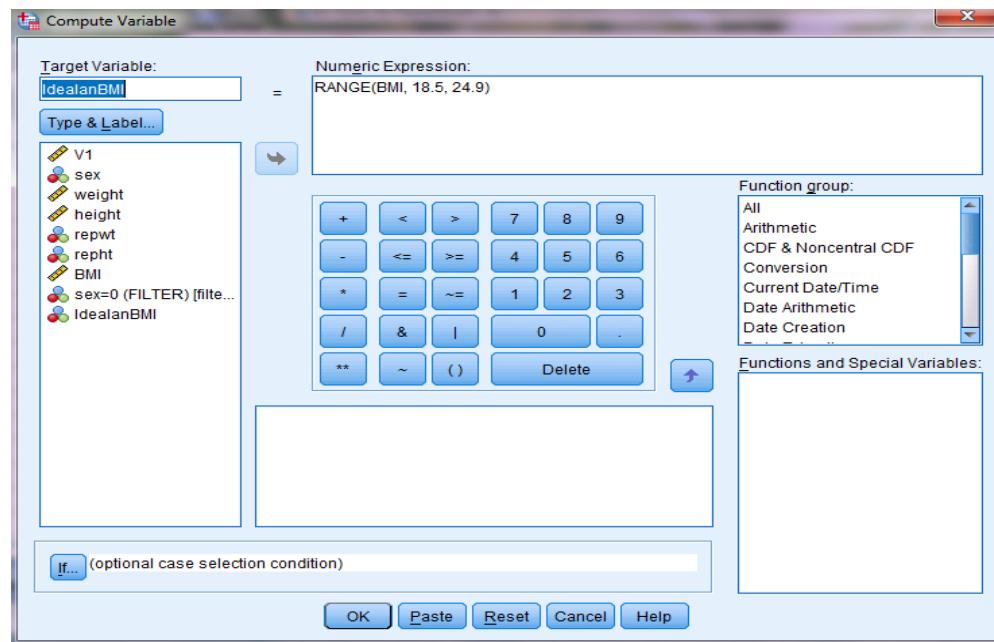
Data → Select Cases

Čekiramo funkciju *if*:



Sada možemo dalje nastaviti sa radom.

Transforme → Compute Variables...



Analyze → Descriptive Statistics → Frequencies

The screenshot shows the SPSS Data Editor with the 'Davis.sav' dataset open. The 'Data View' tab is selected. A 'Frequencies' dialog box is overlaid on the data view. It lists 'Variable(s):' as 'IdealanBMI' and has a checked 'Display frequency tables' option. The data table shows variables V1, sex, weight, height, repwt, rept, BMI, filter_\$, IdealanBMI, and var. The 'IdealanBMI' column contains values like 1.00, .00, 1.00, etc. The 'filter_\$' column contains 'Selected' for most rows. The 'Data View' tab is highlighted at the bottom.

Dobijamo:

Statistics

IdealanBMI

N	Valid	88
	Missing	0

IdealanBMI

	Frequency	Percent	Valid Percent	Cumulative Percent
Valid	.00	36	40.9	40.9
	1.00	52	59.1	59.1
Total	88	100.0	100.0	100.0

Vidimo da 59.1% muškaraca ima idealan BMI. Slično možemo uraditi ako želimo da ispitamo za žene, samo nam se uslov menja, tj stavljamo sex=1. Takođe, ako isključimo uslov izdvojiće se i muškarci i žene.

6. DA LI JE SREDNJA VREDNOST VISINE NAŠIH ISPITANIKA ISTA ZA MUŠKARCE I ŽENE?

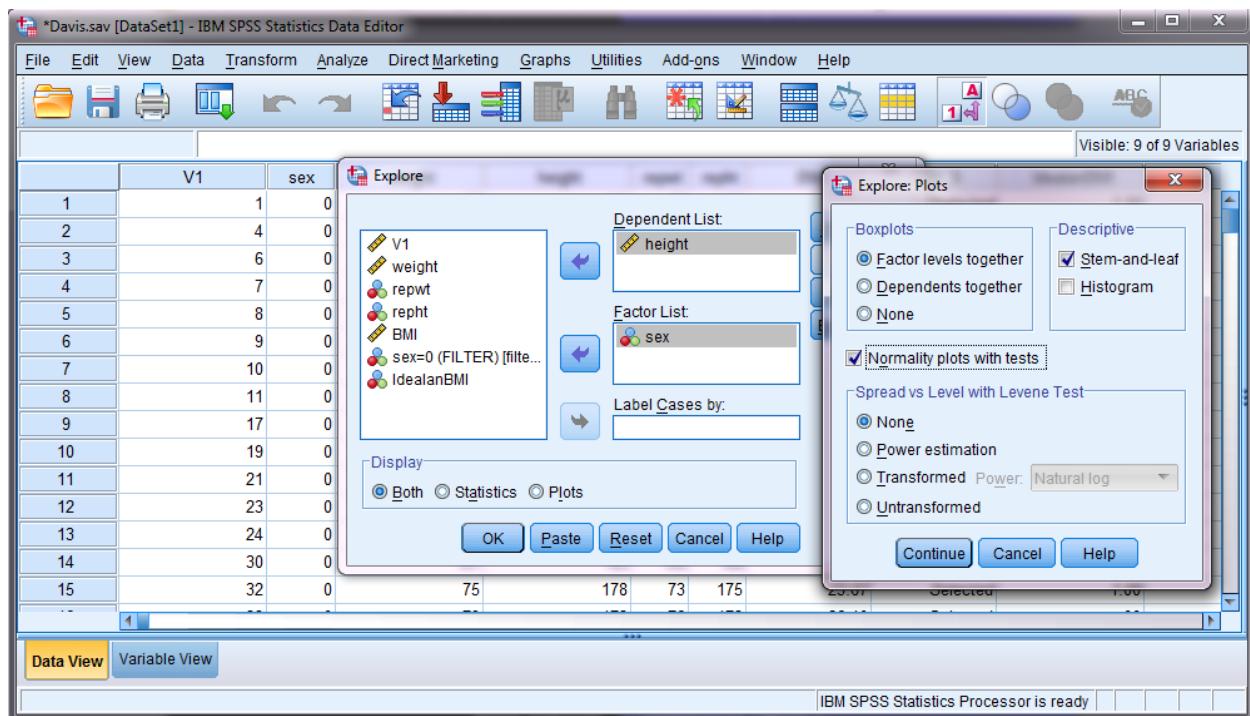
Pokušaćemo da iskoristimo Independent-samples T-test. Uslovi koji treba da budu ispunjeni:

- Neprekidnost zavisne promenljive
- Uzorci su iz populacije sa normalnom raspodelom
- Opervacije su međusobno nezavisne
- Nema značajnih autlajera
- Nepoznata je varijansa osnovnog skupa

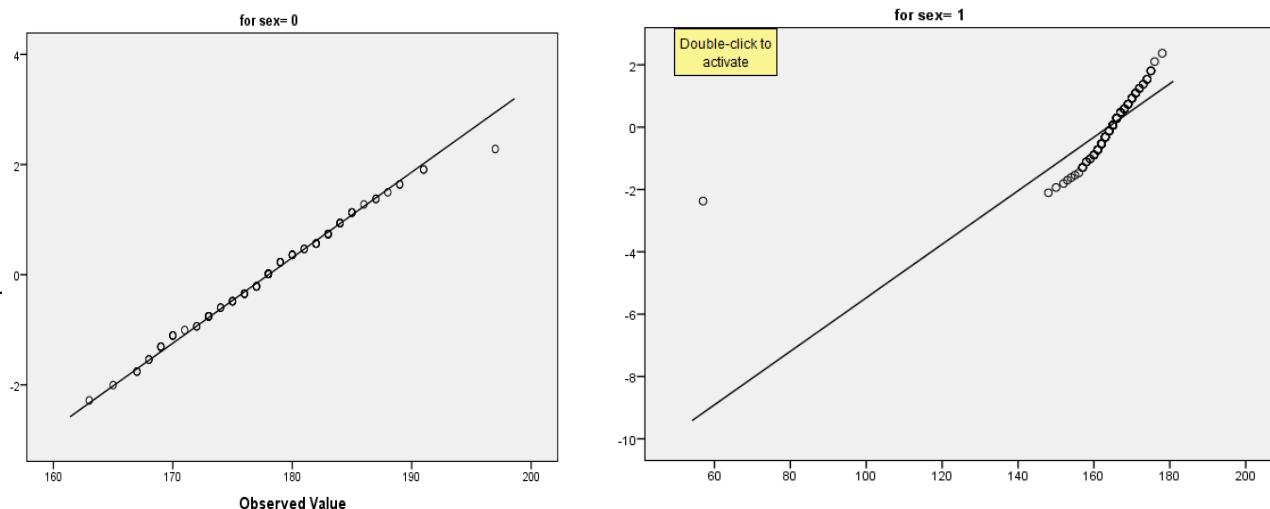
U našem slučaju, zavisna promenljiva je *Height* dok je nezavisna *Sex*. Sada ćemo ispitivati određene uslove:

Prvo ćemo proveriti da li je naša promenljiva *Height* normalno raspodeljena?

Analyze → *Descriptive Statistics* → *Explore...*



Pri čemu dobijamo sledeće grafike:



U slučaju ženskog pola i nismo baš sigurno za tačnost normalne raspodele, zato ćemo proveriti i pomoći testova:

Tests of Normality

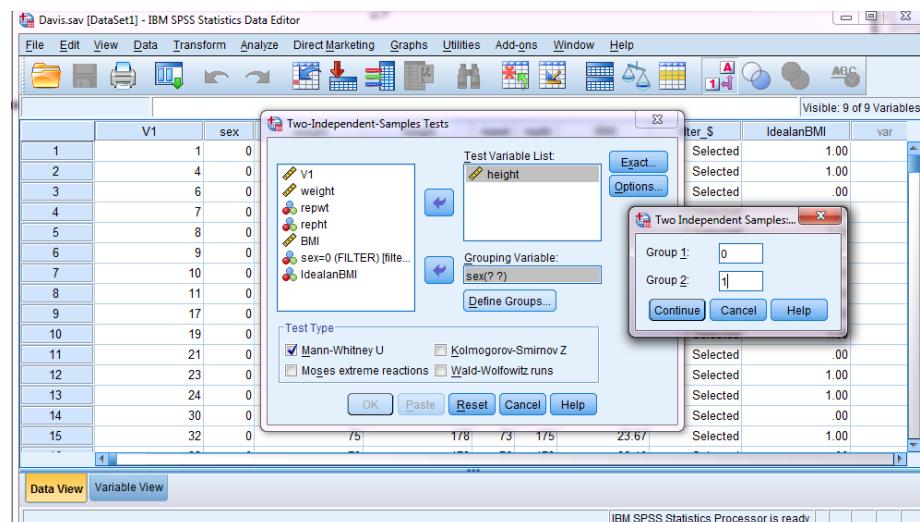
sex	Kolmogorov-Smirnov ^a			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
height 0	.069	88	.200*	.992	88	.872
height 1	.213	112	.000	.482	112	.000

*. This is a lower bound of the true significance.

a. Lilliefors Significance Correction

Kao i što smo mogli videti sa grafika, i na osnovu ova dva testa možemo videti da kod muškaraca važi normalna raspodela, dok kod žena to nije slučaj. Dakle, ne možemo koristiti navedeni test. Najpogodniji test u slučaju kada nam ne važi normalna raspodela je Mann-Whitney, koji poredi dve promenljive na osnovu medijana.

Analyze → Nonparametric tests → Legacy dialogs → 2 independent samples



Pri čemu dobijamo:

Mann-Whitney Test

Ranks

	sex	N	Mean Rank	Sum of Ranks
height	0	88	150.09	13207.50
	1	112	61.54	6892.50
	Total	200		

Test Statistics^a

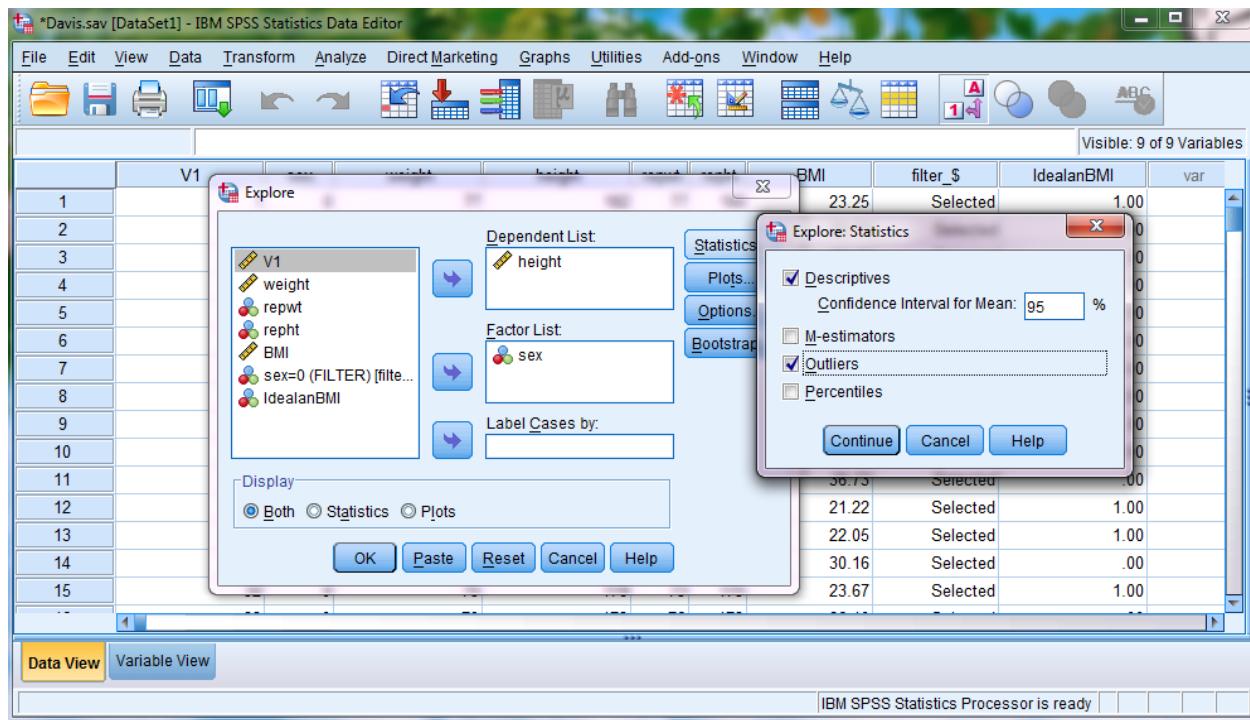
	height
Mann-Whitney U	564.500
Wilcoxon W	6892.500
Z	-10.747
Asymp. Sig. (2-tailed)	.000

a. Grouping Variable: sex

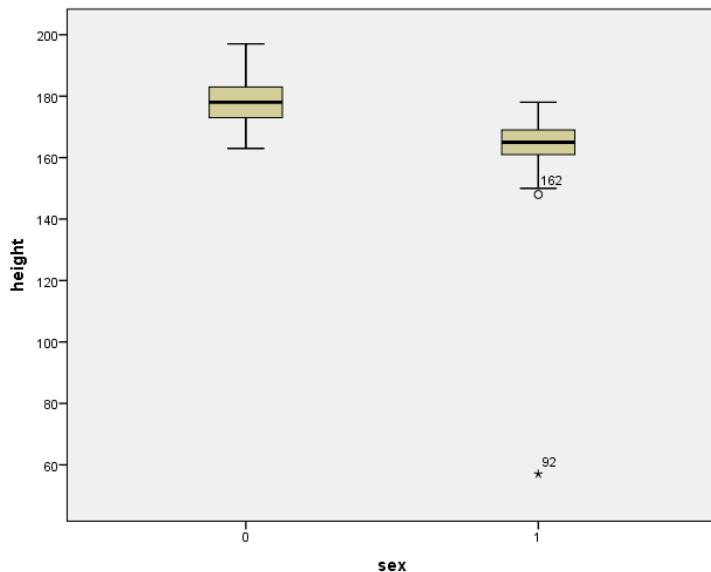
Vidimo da nam je vrednost p-statistike manja od praga unačajnosti 0.05, što znači da odbacujemo nultu hipotezu, tj. visina ne zavisi od pola naših ispitanika.

Sada ćemo da vidimo da li imamo značajnih autlajera:

Analyze → Descriptive statistics → Explore...



Dobijamo sledeći grafik:



Javljuju se samo dva autlajera kod žena prilikom 92 i 162 opservacije, dok kod muškaraca nema autlajera.

7. TESTIRATI VEZU IZMEĐU IZRAČUNATE I PRIJAVLJENE TEŽINE

Kada želimo da ispitujemo vezu između promenljivih to možemo uraditi na tri načina:

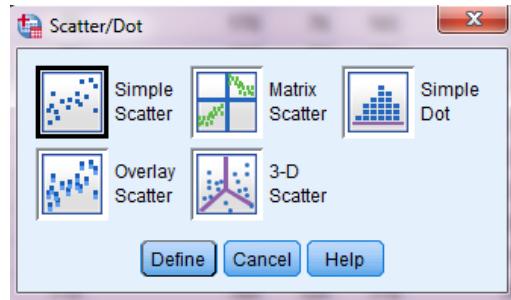
- Procedure za korelacionu analizu
- Procedure za regresionu analizu
- Hi-kvadrat test

Kako se Hi-kvadrat test koristi samo u slučaju kategorijskih promenljivih, a naše promenljive to nisu, ne možemo ga koristiti, pa ćemo primeniti korelacionu analizu.

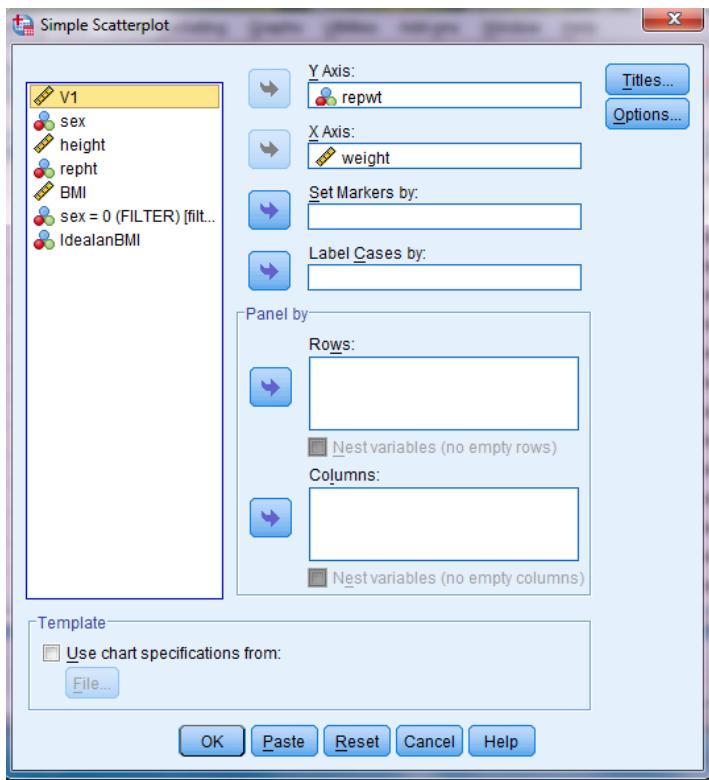
Ovom procedurom mi izražavamo stepen veze između promenljivih, pri čemu se mera korelacije izražava koeficijentom korelacije koji uzima vrednosti u intervalu $[-1, 1]$.

Za početak ćemo nacrtati dijagram raspršenosti:

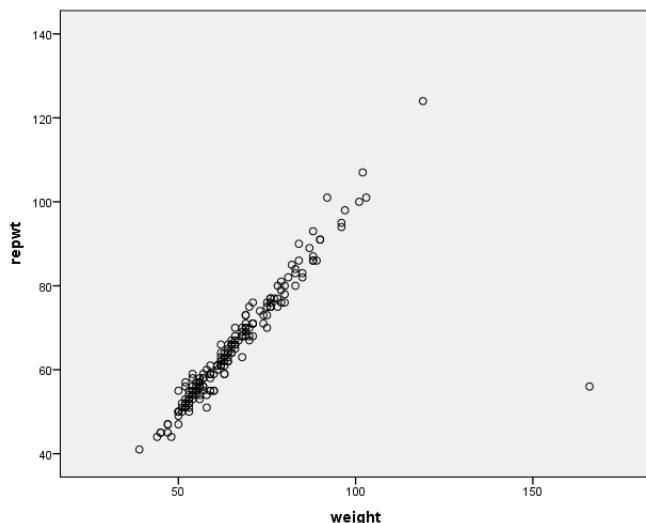
Graphs → Legacy Dialogs → Scatter/Dot



Označićemo *Simple Scatter* za ispitivanje proste korelacije, pa klikom na *Define* otvara nam se sledeći prozor:



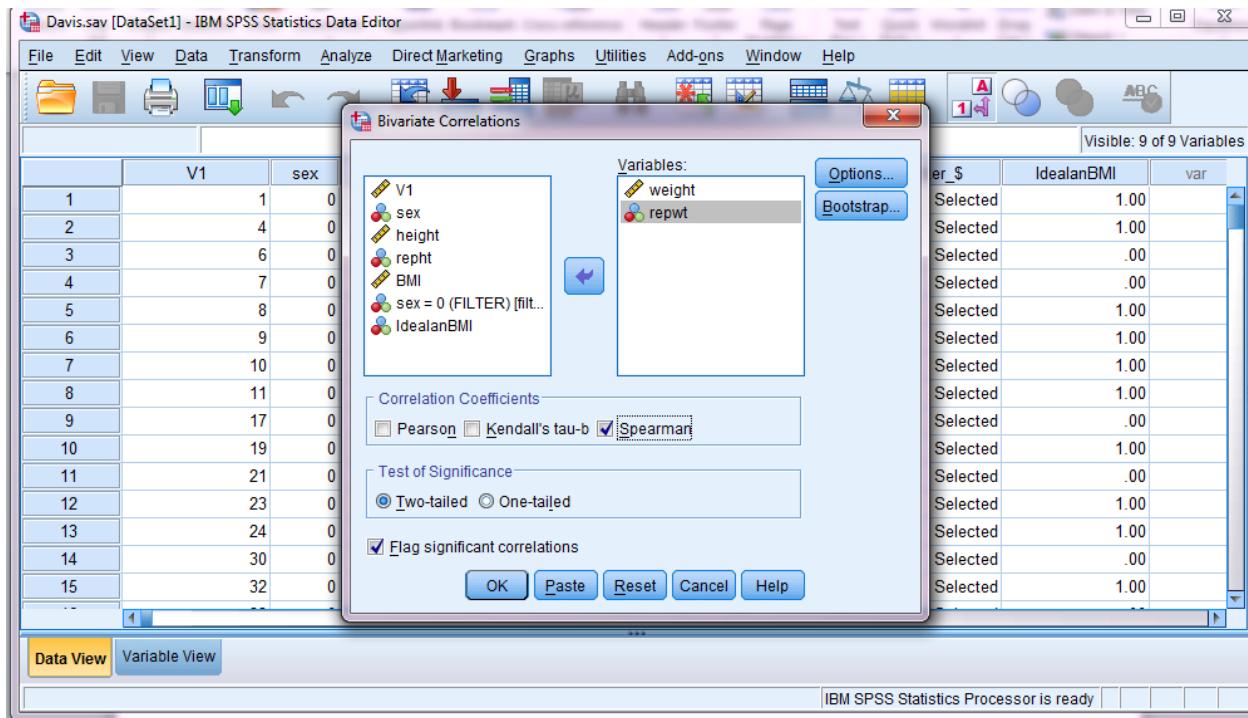
Kada smo označili promenljive čiju zavisnost ispitujemo, dobijamo sledeći grafik:



Vidimo da su tačke poprilično zbijene u jednu grupu, pa možemo zaključiti da postoji linearna veza između naših promenljivih.

Sada ćemo izračunati koeficijent korelacije. Kako naše promenljive nisu normalno raspoređene, koristićemo Spirmanov koeficijent korelacije.

Analyze → Correlate → Bivariate



Pokretanjem programa dobijamo sledeće vrednosti:

Correlations		
Spearman's rho	weight	Correlation Coefficient
		.962**
		.000
	N	200
		183
repwt	Correlation Coefficient	.962**
		.000
	N	183
		183

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

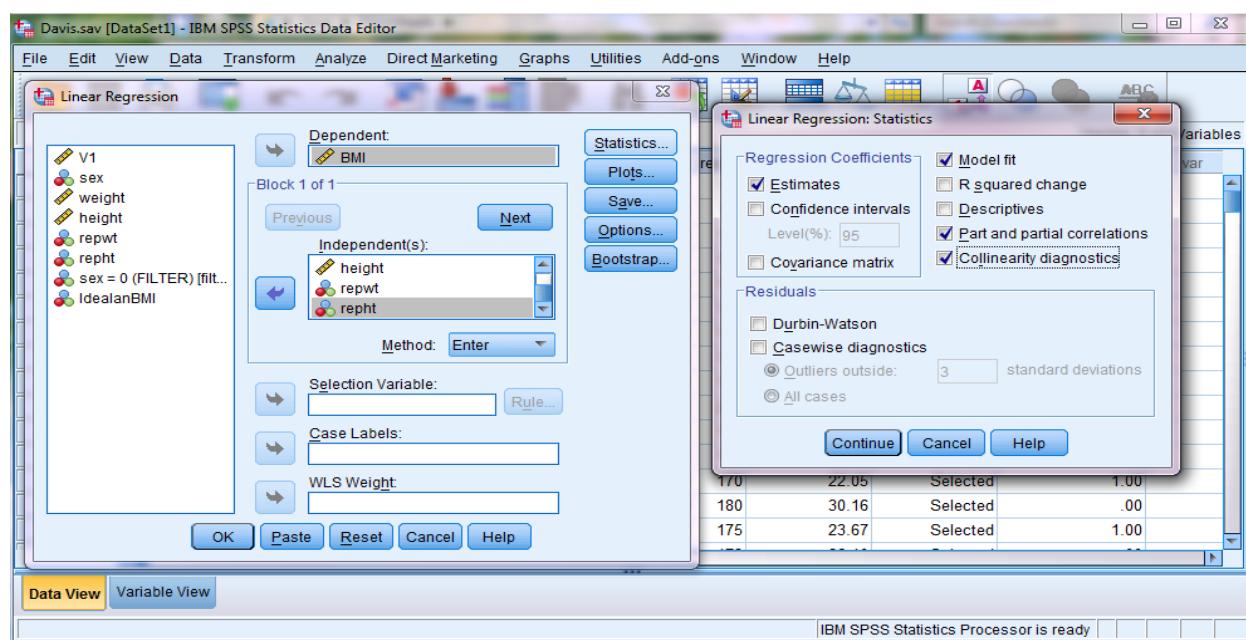
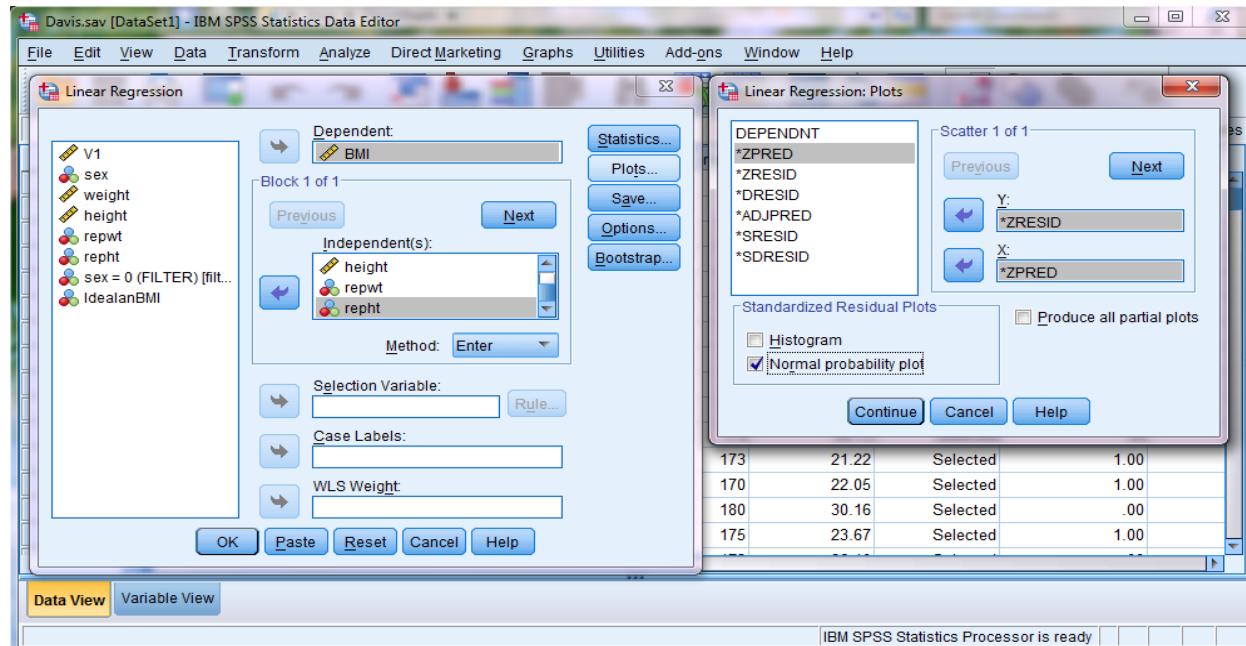
Vidimo da je koeficijent korelacijske 0.962 što nam još jednom potvrđuje vezu između ove dve promenljive.

8. NAPRAVITI NAJBOLJI LINEARNI MODEL ZA IZRAČUNAVANJE TEŽINE U ODносУ НА ОСТАЛЕ ПРОМЕНЛJИВЕ

Kako će na našu zavisnu promenljivu, tj. BMI uticati više nezavisnih promenljivih, pravimo višestruki linearni model. Pretpostavke za primenu višestrukog linearnog modela su sledeće:

- Opservacije su nezavisne
- Greška ima normalnu raspodelu sa očekivanjem 0 i konstantnom disperzijom
- Greške su međusobno nekorelirisane
- Nezavisne promenljive ne smeju biti savršeno korelisane
- Broj podataka u uzorku je značajno veći od broja parametara koji se ocenjuju
- Mora postojati linearna zavisnost između zavisne i bilo koje nezavisne promenljive, ili grupe ostalih

Analyze → Regression → Linear...



Pri čemu dobijamo:

Model Summary^b

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.988 ^a	.976	.976	5.68080

a. Predictors: (Constant), repht, weight, sex, height, repwt

b. Dependent Variable: BMI

ANOVA^a

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	233375.904	5	46675.181	1446.331	.000 ^b
	Residual	5647.502	175	32.271		
	Total	239023.406	180			

a. Dependent Variable: BMI

b. Predictors: (Constant), repht, weight, sex, height, repwt

Iz prve tabele možemo videti da nezavisne promenljive objašnjavaju čak 97,6% disperzije zavisne promenljive, što znači da je veza jaka. Iz druge tabele vidimo da je procenat objašnjenja disperzije značajan.

Coefficients^a

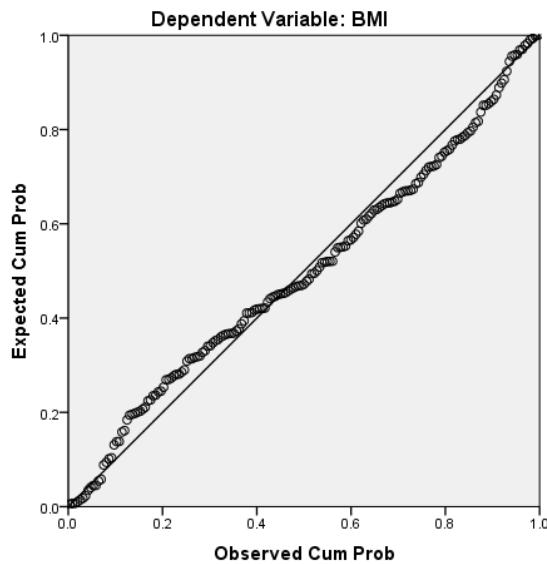
Model	Unstandardized Coefficients			t	Sig.	Correlations			Collinearity Statistics	
	B	Std. Error	Beta			Zero-order	Partial	Part	Tolerance	VIF
1	(Constant)	91.871	12.642	7.267	.000					
	sex	-1.365	1.348	-.019	.313	.029	-.076	-.012	.396	2.527
	weight	1.929	.148	.812	13.010	.000	.548	.701	.151	.035
	height	-2.454	.151	-.829	-16.213	.000	-.662	-.775	-.188	.052
	repwt	-1.431	.160	-.543	-8.937	.000	.018	-.560	-.104	.037
	reph	1.883	.151	.485	12.463	.000	-.015	.686	.145	.089
										11.235

a. Dependent Variable: BMI

Sada ćemo posmatrati p-vrednosti promenljivih. Značajne za naš model su samo one promenljive čija je p-vrednost manja od praga značajnosti, što znači da u našem slučaju samo promenljiva *sex* nije značajna.

Treba da proverimo i normalnost raspodele reziduala:

Normal P-P Plot of Regression Standardized Residual



Nemamo neka veća odstupanja, pa možemo pretpostaviti da važi Normalna raspodela.

Naš model izgleda:

$$Y = 91.871 + 1.929X_2 - 2.454X_3 - 1.431X_4 + 1.883X_5$$

Pored *Enter* opcije, tj standardne linearne regresije imamo još mogućnosti za pravljenje modela. Na primer, *Remove*, *Forward*, *Backward*, *Stepwise*. Možemo da pokušamo i sa nekom od njih, možda dobijemo još bolji model. Međutim, pozivanjem i *Forward* i *Stepwise* dobijamo:

Variables Entered/Removed ^a			
Model	Variables Entered	Variables Removed	Method
1	height		Stepwise (Criteria: Probability-of-F-to-enter <= .050, Probability-of-F-to-remove >= .100).
2	repht		Stepwise (Criteria: Probability-of-F-to-enter <= .050, Probability-of-F-to-remove >= .100).
3	weight		Stepwise (Criteria: Probability-of-F-to-enter <= .050, Probability-of-F-to-remove >= .100).
4	repwt		Stepwise (Criteria: Probability-of-F-to-enter <= .050, Probability-of-F-to-remove >= .100).

a. Dependent Variable: BMI

Pri čemu dobijamo isti rezultat kao i sa *Enter*.

LITERATURA

- <http://www.math.rs/p/marija-radicevic/kurs/324/%D0%A1%D1%82%D0%B0%D1%82%D0%B8%D1%81%D1%82%D0%B8%D1%87%D0%BA%D0%B8-%D1%81%D0%BE%D1%84%D1%82%D0%B2%D0%BD%D1%80-3/>
- <http://stat.uns.ac.rs/LLLprogramme/NP/TeachingMaterial/Uputstvo.pdf>
- <https://vincentarelbundock.github.io/Rdatasets/datasets.html>