

UNIVERZITET U BEOGRADU
MATEMATIČKI FAKULTET



Seminarski rad iz Statističkog softvera 3

Asistent:

Marija Radičević

Student:

Jelena Cicvarić 158/2012

Beograd, septembar 2016. godine

Sadržaj:

1. Opis baze i cilj istraživanja.....	3
2. Učitavanje baze u SPSS.....	4
3. Analiza podataka.....	7
4. Zavisnost visine plate od ostalih promenljivih.....	13
5. Ispitujemo da li su srednje vrednosti visine plate iste i za muškarce i za žene.....	20
6. Linearni model.....	27
7. Literatura.....	34

1. Opis baze i cilj istraživanja

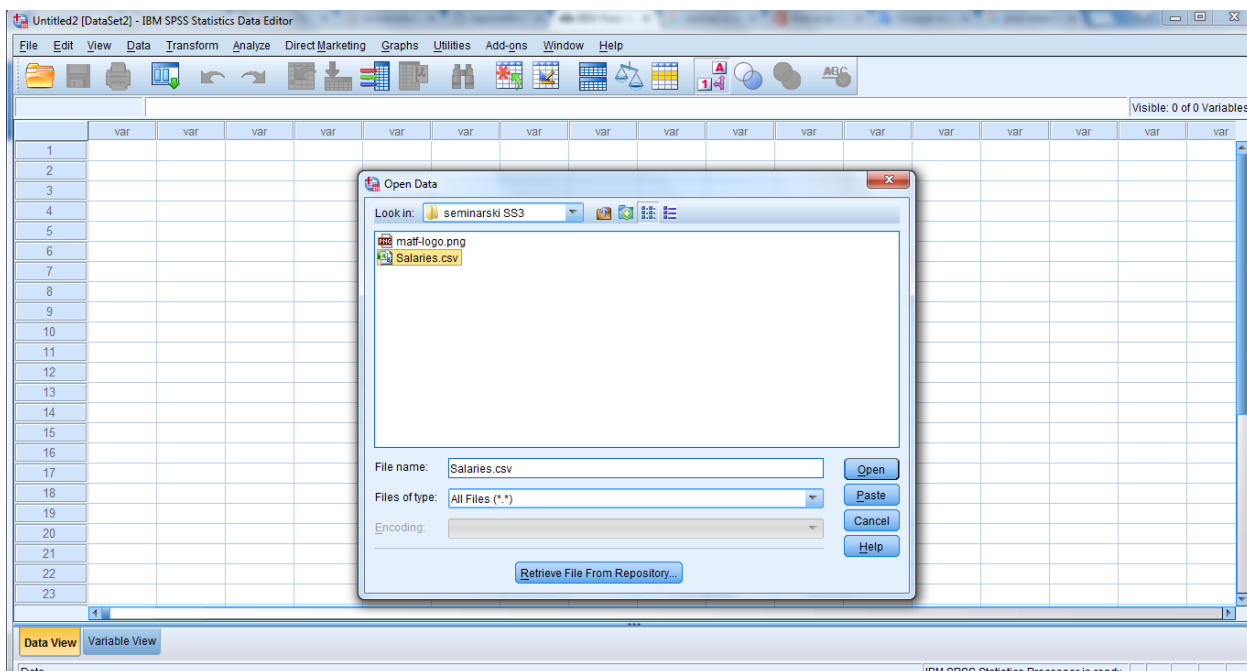
Koristimo bazu ***Salaries for Professors***, baza se sastoji od 397 opservacija i 6 promenljivih. Promenljive su:

- *rank* - kategorička promenljiva koja uzima vrednosti:
 - 1 = docent
 - 2 = vandredni professor
 - 3 = professor
- *discipline* - kategorička promenljiva koja uzima vrednosti:
 - 1 = teoretska odeljenja
 - 2 = primenjena odeljenja
- *yrs.since.phd* - numerička promenljiva koja predstavlja godine od doktorata
- *yrs.service* - numerička promenljiva koja predstavlja godine radnog staža
- *sex* - kategorička promenljiva, pol ispitanika, uzima vrednosti:
 - 0 = muški
 - 1 = ženski
- *salary* – numerička promenljiva koja predstavlja devetomesečnu platu (u dolarima)

Cilj istraživanja: želimo da proverimo koji faktori utiču na visinu plate, a medju najosnovnijim proveravamo da li visina plate zavisi od pola ispitanika.

2. Učitavanje baze u SPSS

File → Open → Data (Files of type promenimo na All files), čekiramo našu bazu
I kliknemo Open

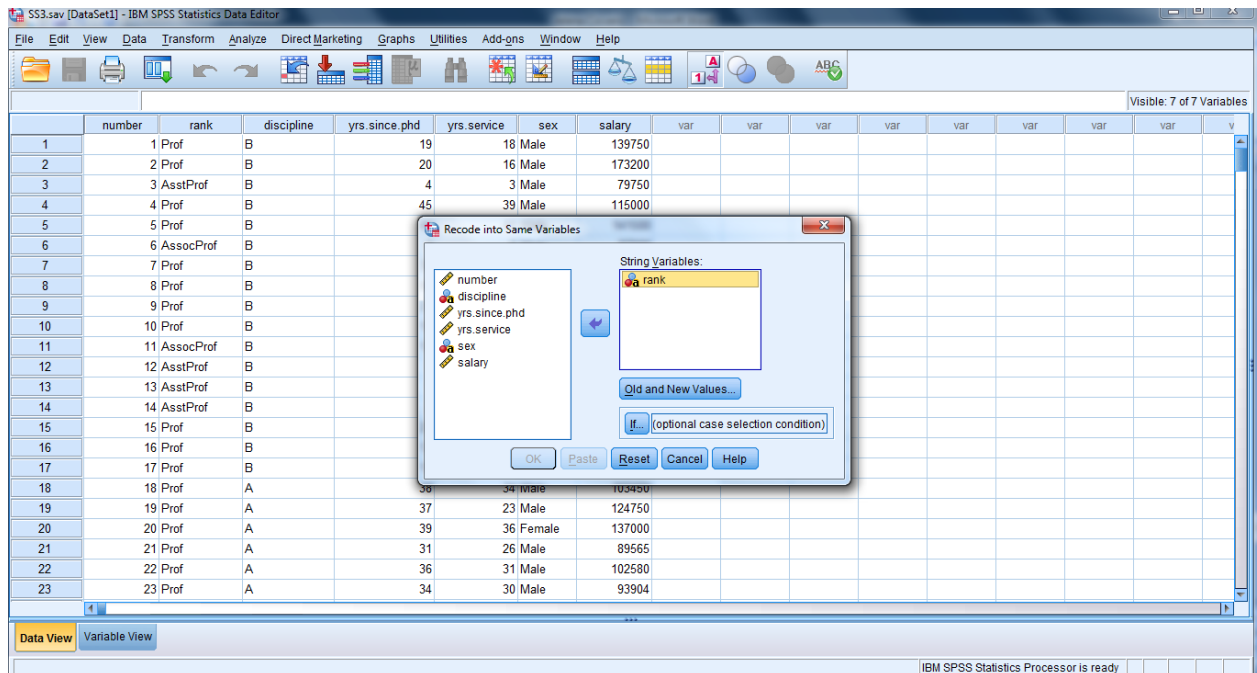


Visible: 7 of 7 Variables

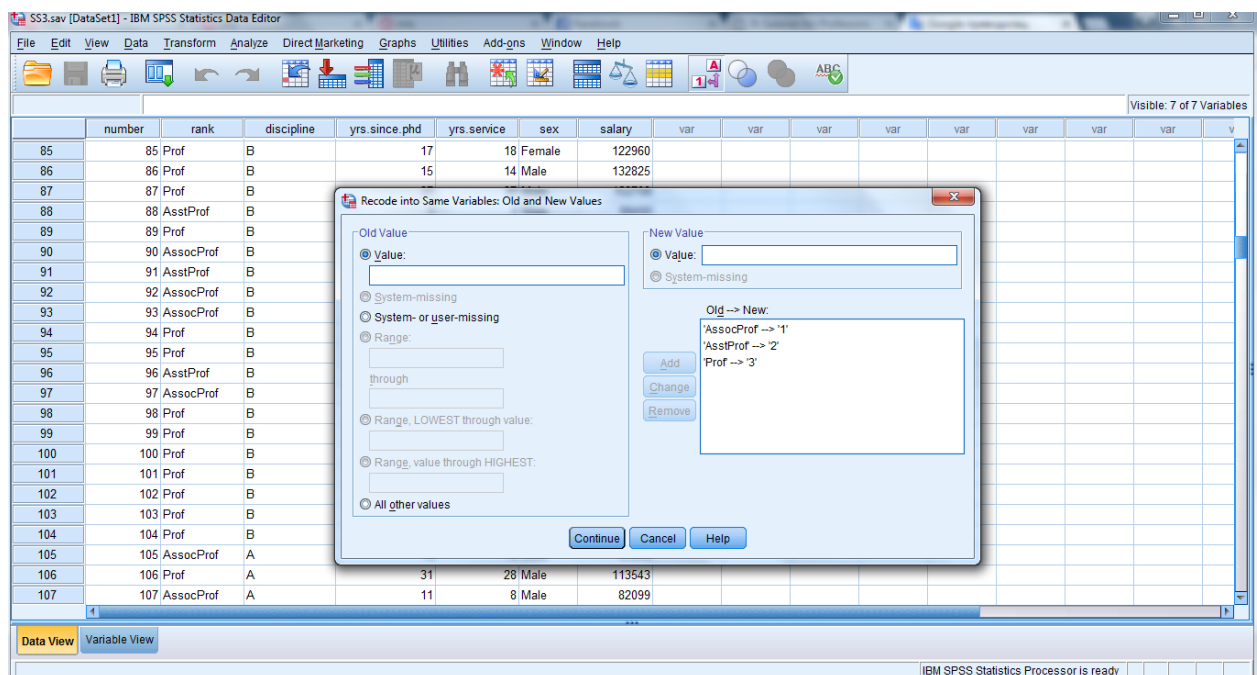
	number	rank	discipline	yrs. since phd	yrs. service	sex	salary
1	1	Prof	B	19	18	Male	139750
2	2	Prof	B	20	16	Male	173200
3	3	AsstProf	B	4	3	Male	79750
4	4	Prof	B	45	39	Male	115000
5	5	Prof	B	40	41	Male	141500
6	6	AssocProf	B	6	6	Male	97000
7	7	Prof	B	30	23	Male	175000
8	8	Prof	B	45	45	Male	147765
9	9	Prof	B	21	20	Male	119250
10	10	Prof	B	18	18	Female	129000
11	11	AssocProf	B	12	8	Male	119800
12	12	AsstProf	B	7	2	Male	79800
13	13	AsstProf	B	1	1	Male	77700
14	14	AsstProf	B	2	0	Male	78000
15	15	Prof	B	20	18	Male	104800
16	16	Prof	B	12	3	Male	117150
17	17	Prof	B	19	20	Male	101000
18	18	Prof	A	38	34	Male	103450
19	19	Prof	A	37	23	Male	124750
20	20	Prof	A	39	36	Female	137000
21	21	Prof	A	31	26	Male	89565
22	22	Prof	A	36	31	Male	102580
23	23	Prof	A	34	30	Male	93904

Primećujemo da imamo promenljive koje su tipa *String*, da bismo odradili potrebne statističke analize ove promenljive ćemo promeniti u numeričke. To radimo na sledeći način:

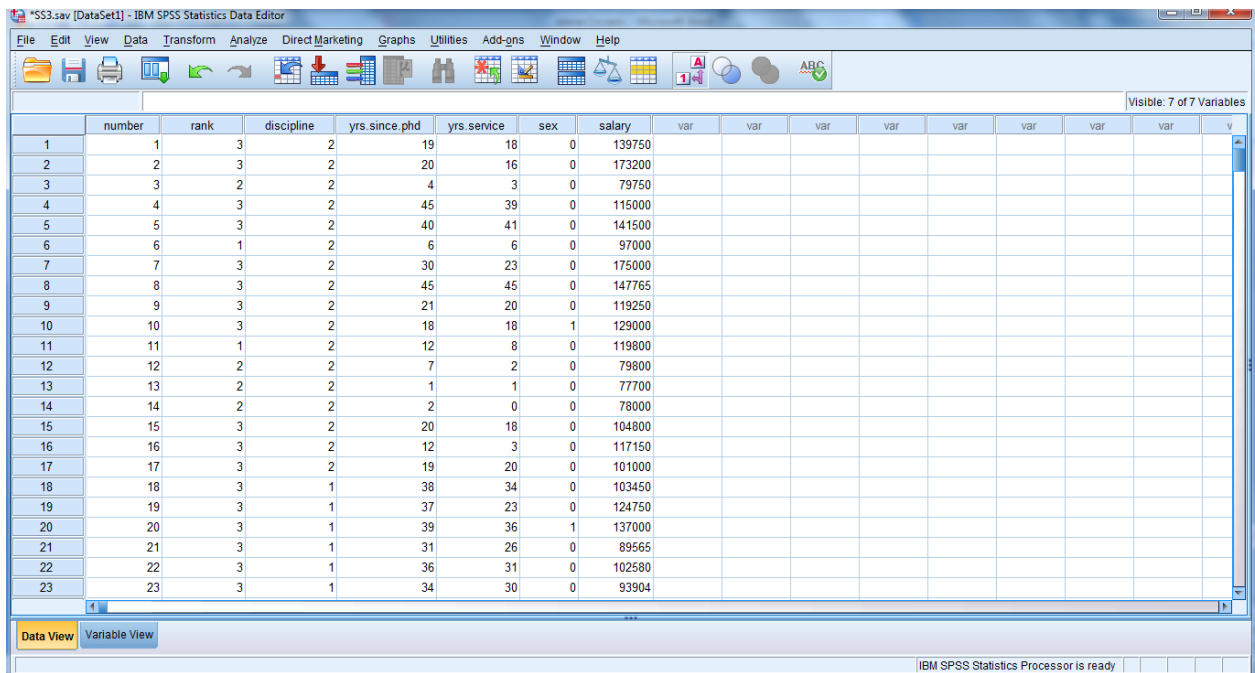
Transform → Recode Into Same Variables



Sada klikom na *Old and New Variables* prekodiracemo naše vrednosti:



Potpuno isto radimo i za ostale promenljive koje su tipa *String*. Naša baza izgleda ovako:



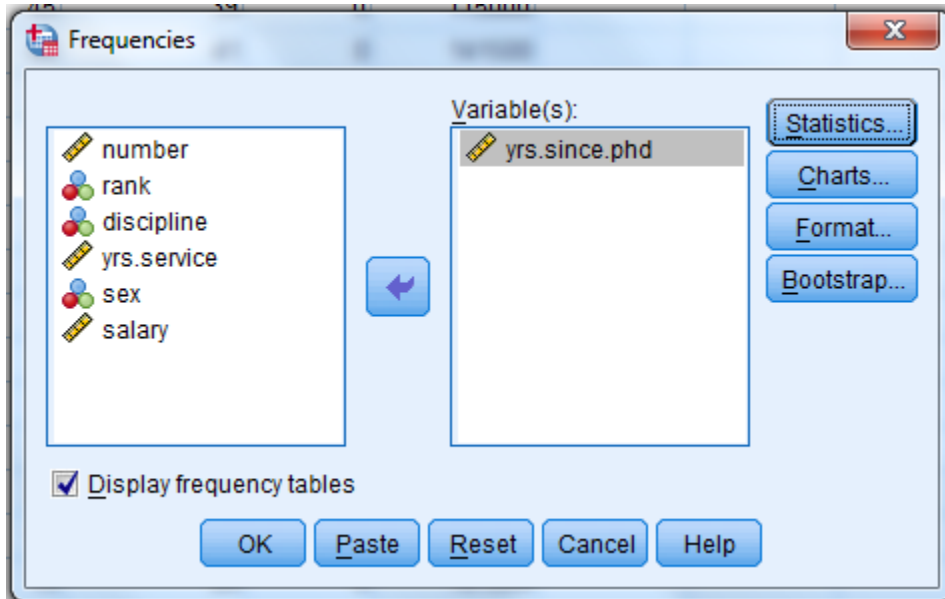
	number	rank	discipline	yrs. since.phd	yrs. service	sex	salary	var	var	var	var	var	var	var	var	v
1	1	3	2	19	18	0	139750									
2	2	3	2	20	16	0	173200									
3	3	2	2	4	3	0	79750									
4	4	3	2	45	39	0	115000									
5	5	3	2	40	41	0	141500									
6	6	1	2	6	6	0	97000									
7	7	3	2	30	23	0	175000									
8	8	3	2	45	45	0	147765									
9	9	3	2	21	20	0	119250									
10	10	3	2	18	18	1	129000									
11	11	1	2	12	8	0	119800									
12	12	2	2	7	2	0	79800									
13	13	2	2	1	1	0	77700									
14	14	2	2	2	0	0	78000									
15	15	3	2	20	18	0	104800									
16	16	3	2	12	3	0	117150									
17	17	3	2	19	20	0	101000									
18	18	3	1	38	34	0	103450									
19	19	3	1	37	23	0	124750									
20	20	3	1	39	36	1	137000									
21	21	3	1	31	26	0	89565									
22	22	3	1	36	31	0	102580									
23	23	3	1	34	30	0	93904									

Sada možemo da počnemo sa analizom podataka.

3. Analiza podataka

Za početak ćemo izračunati srednju vrednost, maksimum, minimum, standardnu devijaciju za svaku promenljivu. To radimo na sledeći način:

Analyze → Descriptive Statistics → Frequencies

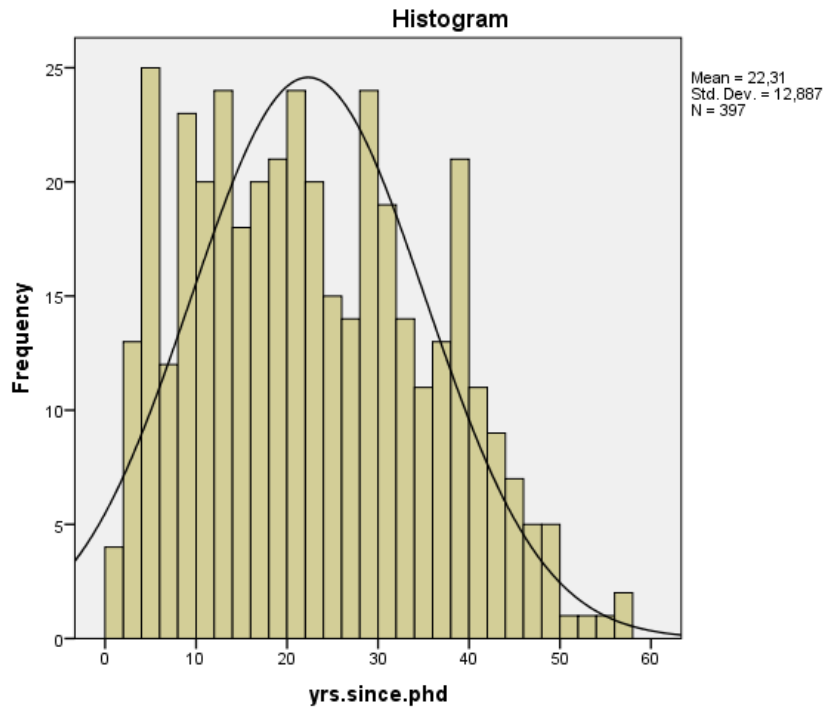


Statistics

yrs.since.phd

N	Valid	397
	Missing	0
Mean		22,31
Std. Error of Mean		,647
Median		21,00
Mode		4
Std. Deviation		12,887
Variance		166,075
Range		55
Minimum		1
Maximum		56
Sum		8859
Percentiles	25	12,00
	50	21,00
	75	32,00

Odavde vidimo da je prosečan broj godina otkako su završili doktorske studije 22,31. Takodje možemo napraviti histogram pritiskom na dugme Charts čekiramo Histogram I možemo čekirati da nam pokaže krivu normalne raspodele preko histograma.



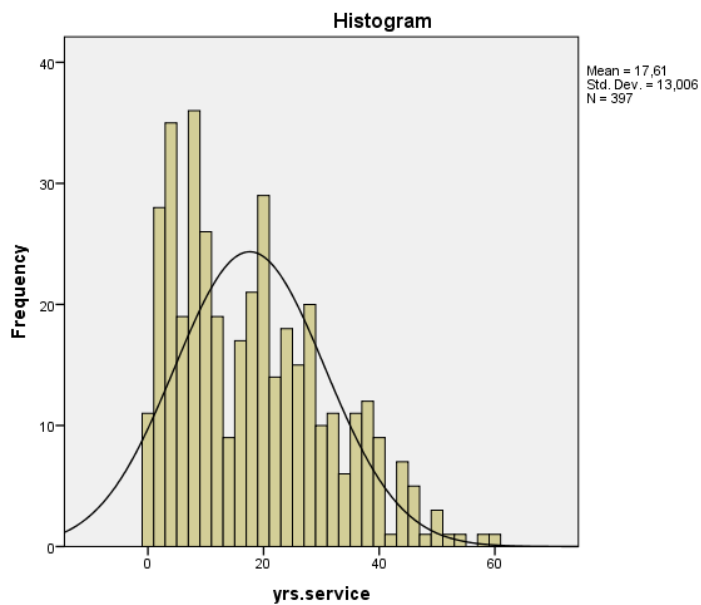
Isto radimo I sa ostalim promenljivim. Sada ispitujemo statistike za godine radnog staža:

Statistics

yrs.service

N	Valid	397
	Missing	0
Mean		17,61
Std. Error of Mean		,653
Median		16,00
Mode		3
Std. Deviation		13,006
Variance		169,157
Range		60
Minimum		0
Maximum		60
Percentiles	25	7,00
	50	16,00
	75	27,00

Vidimo da je prosečan broj godina radnog staža 17,61, da postoje neke osobe koje imaju 0 godina radnog staža tj. tek su počeli da rade I maksimum radnog staža je 60 godina.

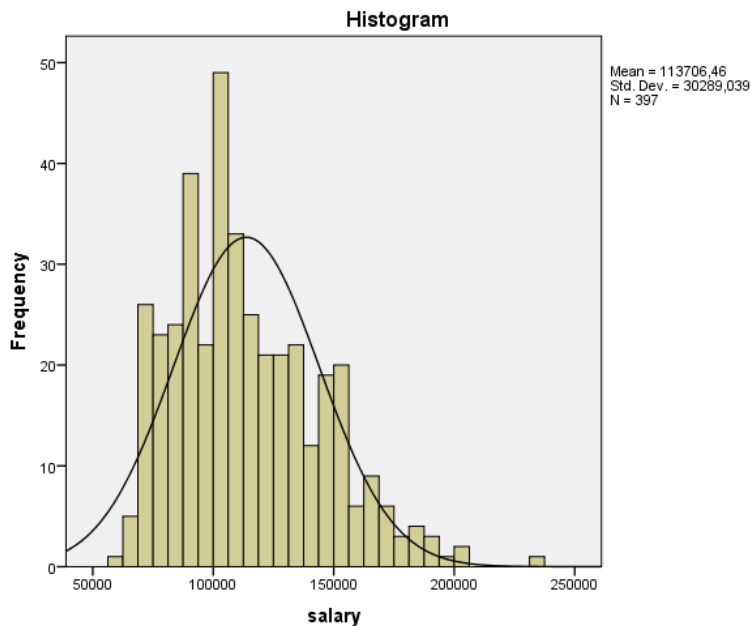


Proveravamo statistike za devetomesečnu platu:

Statistics

salary		
N	Valid	397
	Missing	0
Mean		113706,46
Std. Error of Mean		1520,163
Median		107300,00
Mode		92000
Std. Deviation		30289,039
Variance		917425865,1
Range		173745
Minimum		57800
Maximum		231545
Percentiles	25	91000,00
	50	107300,00
	75	134367,50

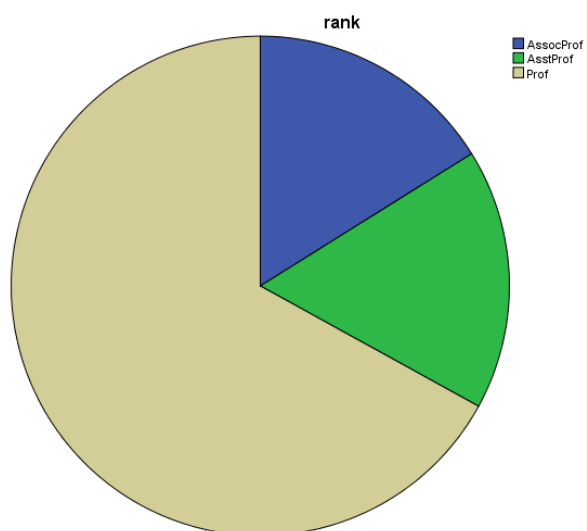
Prosečna plata iznosi 113.706,46 dolara, minimalna je 57.800, a maksimalna 231.545 dolara.



Sada ćemo proveravati za kategoričke promenljive, pa ćemo umesto histograma koristiti *Pie Charts*. Prvo proveravamo promenljivu rank:

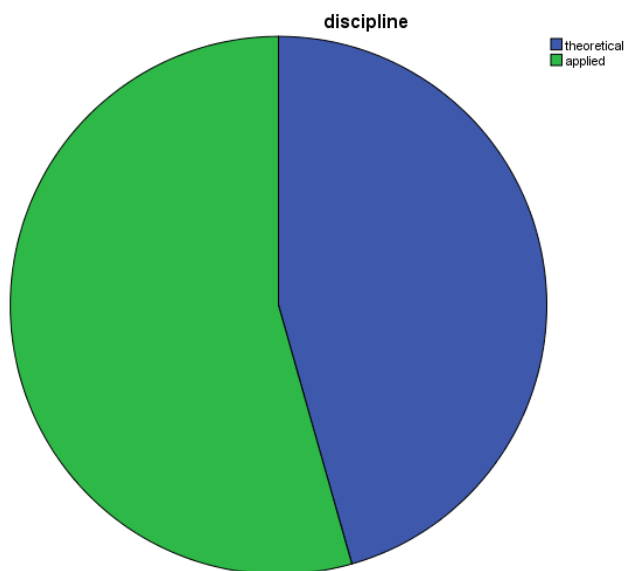
		rank			
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	AssocProf	64	16,1	16,1	16,1
	AsstProf	67	16,9	16,9	33,0
	Prof	266	67,0	67,0	100,0
	Total	397	100,0	100,0	

Vidimo da najviše ima profesora dok vandrednih profesora I docenata ima isto i dosta manje, a to mozemo videti I sa grafika.



Sada ćemo proveriti kako su rasporedjeni na teorijskoj I primenjenoj katedri.

		discipline			
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	theoretical	181	45,6	45,6	45,6
	applied	216	54,4	54,4	100,0
	Total	397	100,0	100,0	

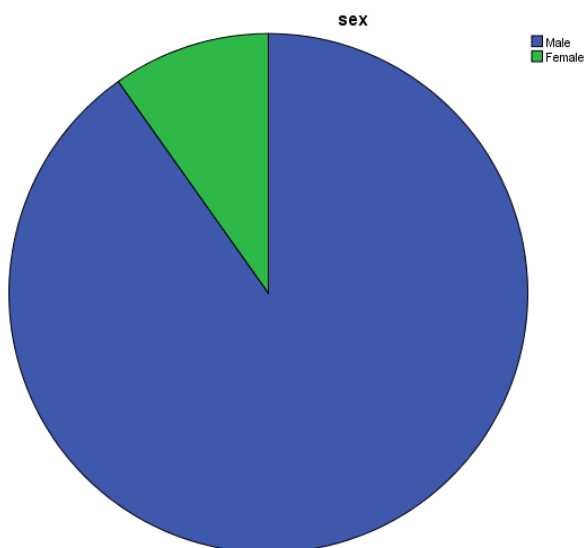


Vidimo da se tu brojke vrlo malo razlikuju, na primenjenoj katedri ima 35 više zaposlenih nego na teorijskoj.

Proveravamo još odnos žena I muškaraca:

sex

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Male	358	90,2	90,2	90,2
	Female	39	9,8	9,8	100,0
	Total	397	100,0	100,0	

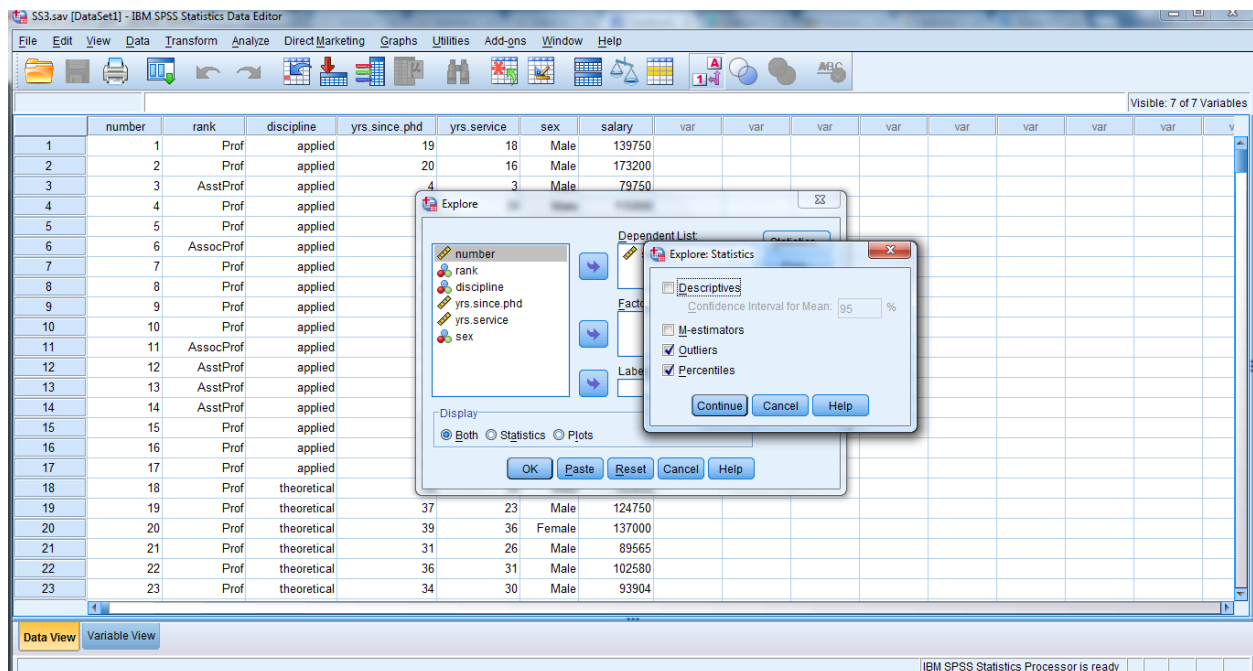


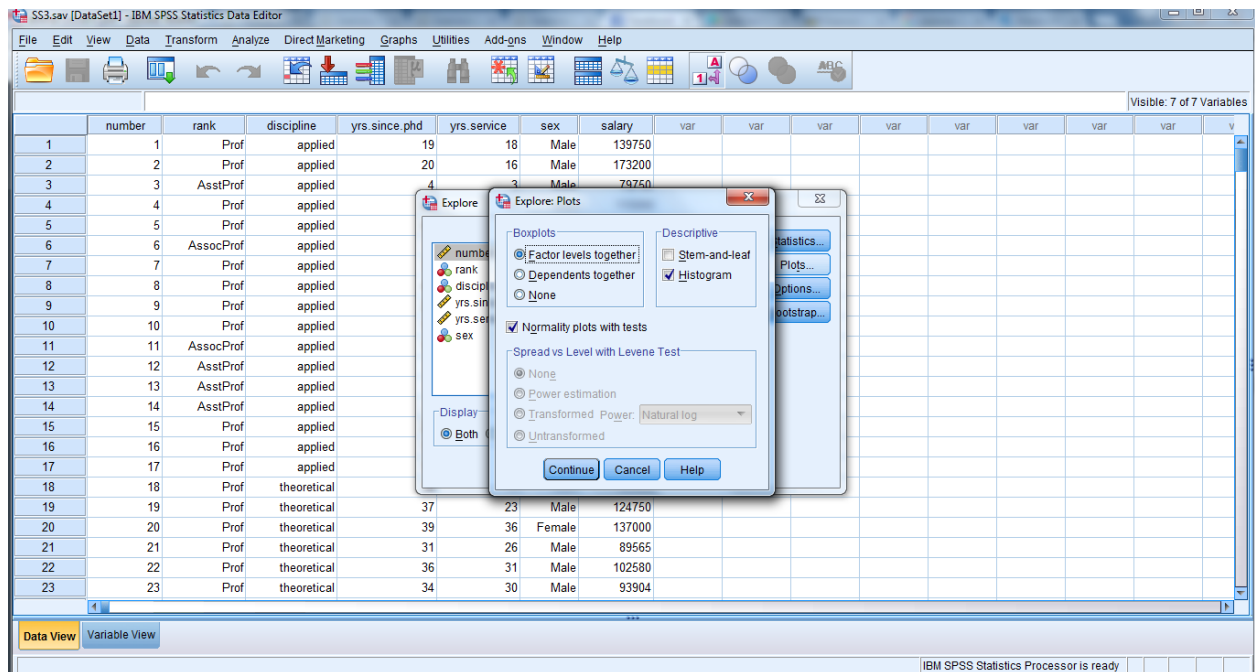
Vidimo da je broj muškaraca znatno veći od broja žena, 358 muškaraca i samo 39 žena.

4. Zavisnost visine plate od ostalih promenljivih

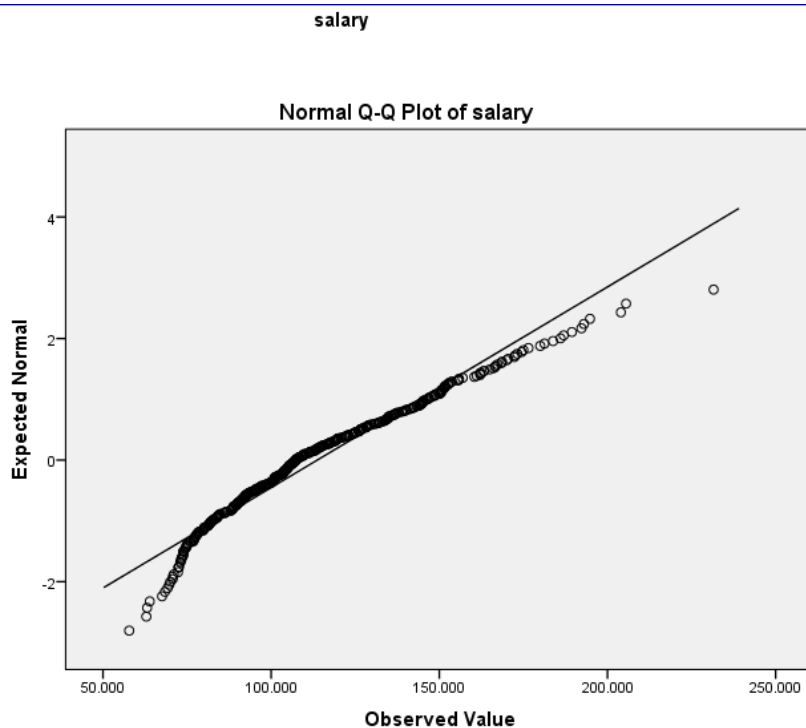
Da bismo odredili koji testove ćemo primenjivati prvo ćemo ispitati normalnost promenljive *salary*, kao i da li postoji autlejera.

Analyze → *Descriptive Statistics* → *Explore*





Dobijamo sledeće rezultate:

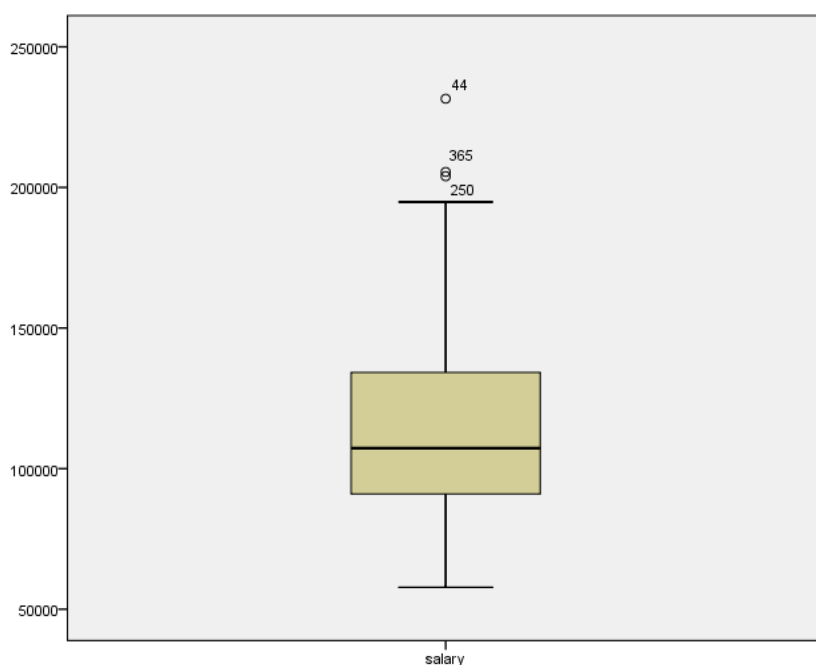


Tests of Normality

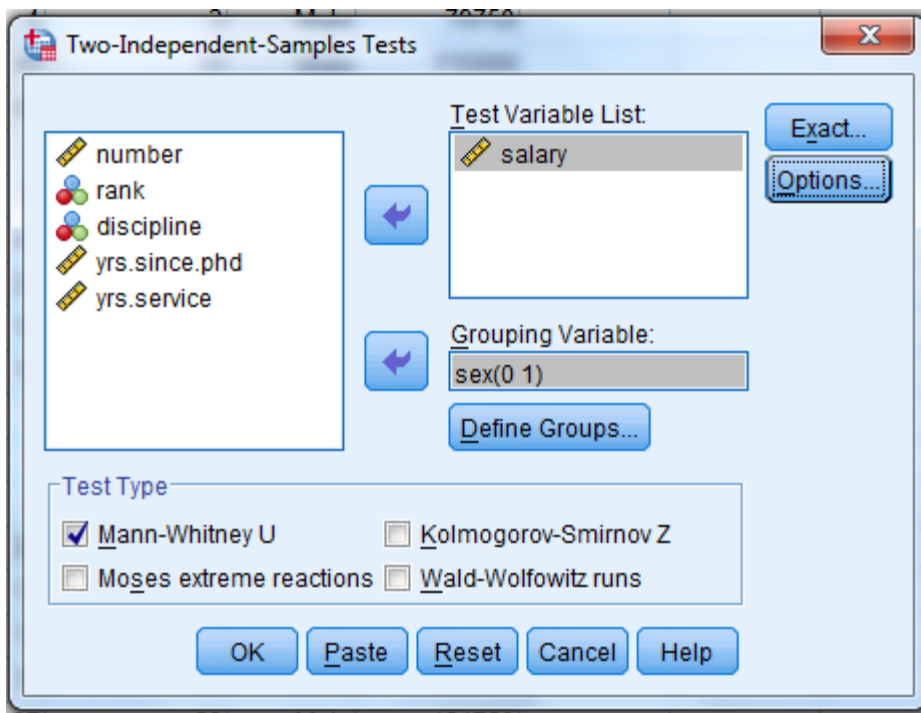
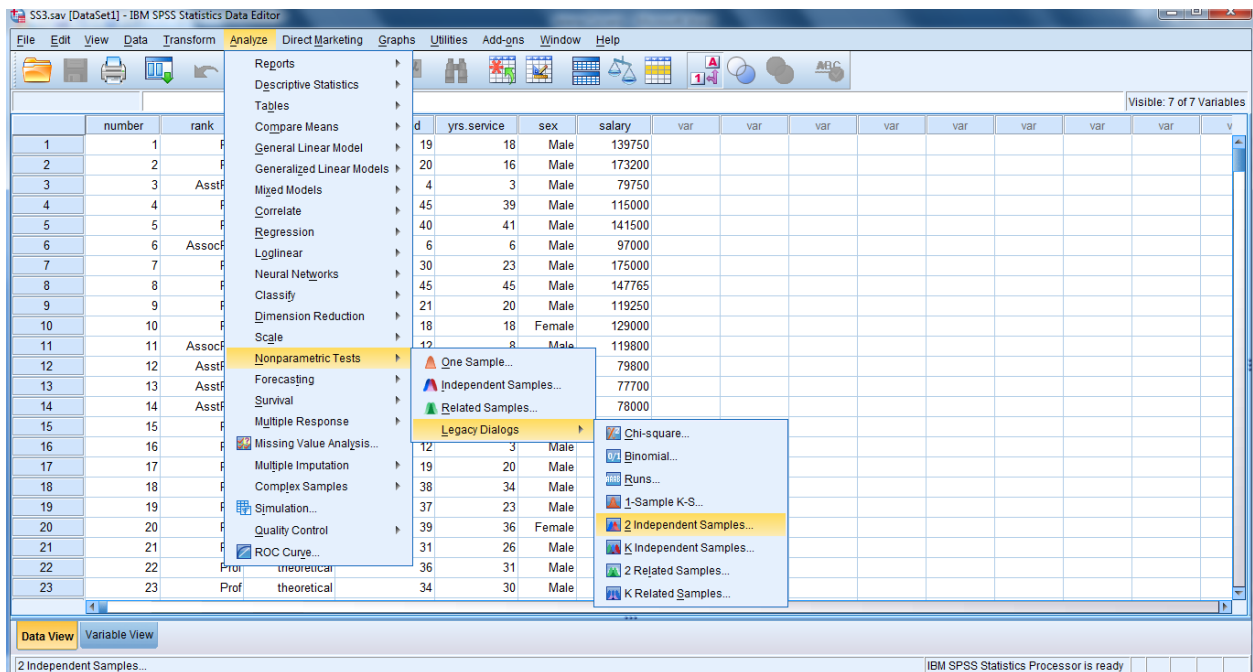
	Kolmogorov-Smirnov ^a			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
salary	,091	397	,000	,960	397	,000

a. Lilliefors Significance Correction

Sa grafika vidimo da promenljiva salary nema normalnu raspodelu, a to isto mozemo primetiti i iz testova normalnosti i kod Kolmogorov-Smirnov i Shapiro-Wilk testa p-vrednost testa je 0,00 a to je manje od 0,05 pa odbacujemo nultu hipotezu, tj. promenljiva salary nema normalnu raspodelu.



Takodje primećujemo da postoje autlejeri. Sada možemo zaključiti da ne možemo primenjivati testove koji zahtevaju normalnost i nepostojanje autlejera. Moramo koristiti neki od neparametarskih testova. Upotrebićemo Mann-Whitney test, on se nalazi u sledećim karticama.



Kada kliknemo Ok, dobijamo sledeće rezultate testa.

Mann-Whitney Test

Ranks

	sex	N	Mean Rank	Sum of Ranks
salary	Male	358	204,02	73040,50
	Female	39	152,88	5962,50
	Total	397		

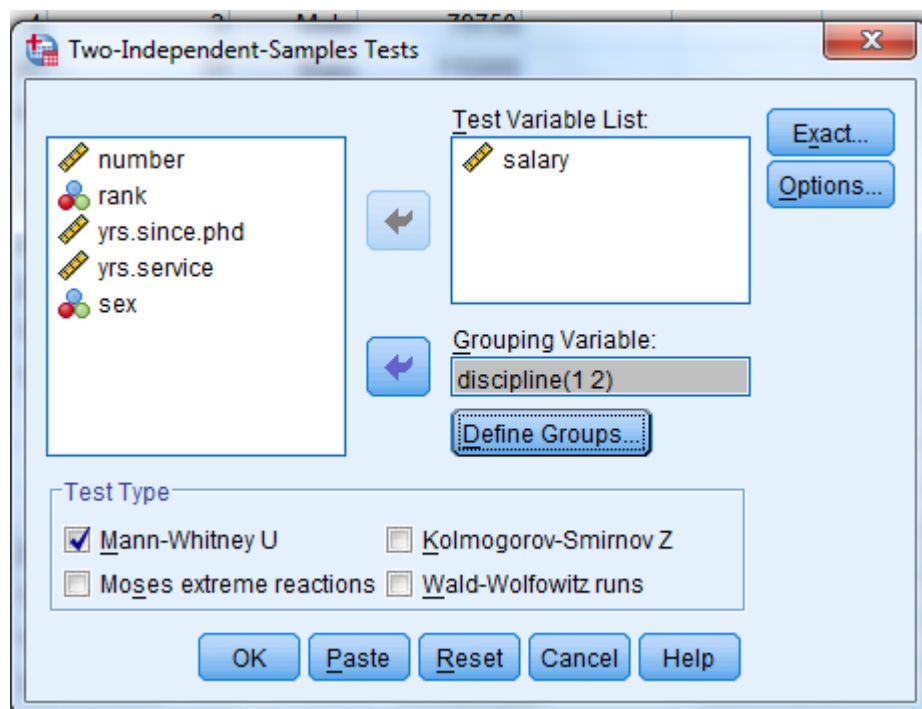
Test Statistics^a

	salary
Mann-Whitney U	5182,500
Wilcoxon W	5962,500
Z	-2,643
Asymp. Sig. (2-tailed)	,008

a. Grouping Variable: sex

Vidimo da je p-vrednost testa 0,008, što je manje od 0,05 pa odbacujemo nultu hipotezu, tj. visina plate ne zavisi od pola zaposlenih.

Sada ćemo proveriti da li visina plate zavisi od promenljive *discipline*, koja označava na kojoj katedri su zaposleni. Isto ćemo proveravati sa Mann-Whitney testom, jer koristimo promenljivu *salary* koja nema normalnu raspodelu i ima autlejera. Radimo na analogan način.



Mann-Whitney Test

Ranks

discipline		N	Mean Rank	Sum of Ranks
salary	theoretical	181	177,80	32181,00
	applied	216	216,77	46822,00
Total		397		

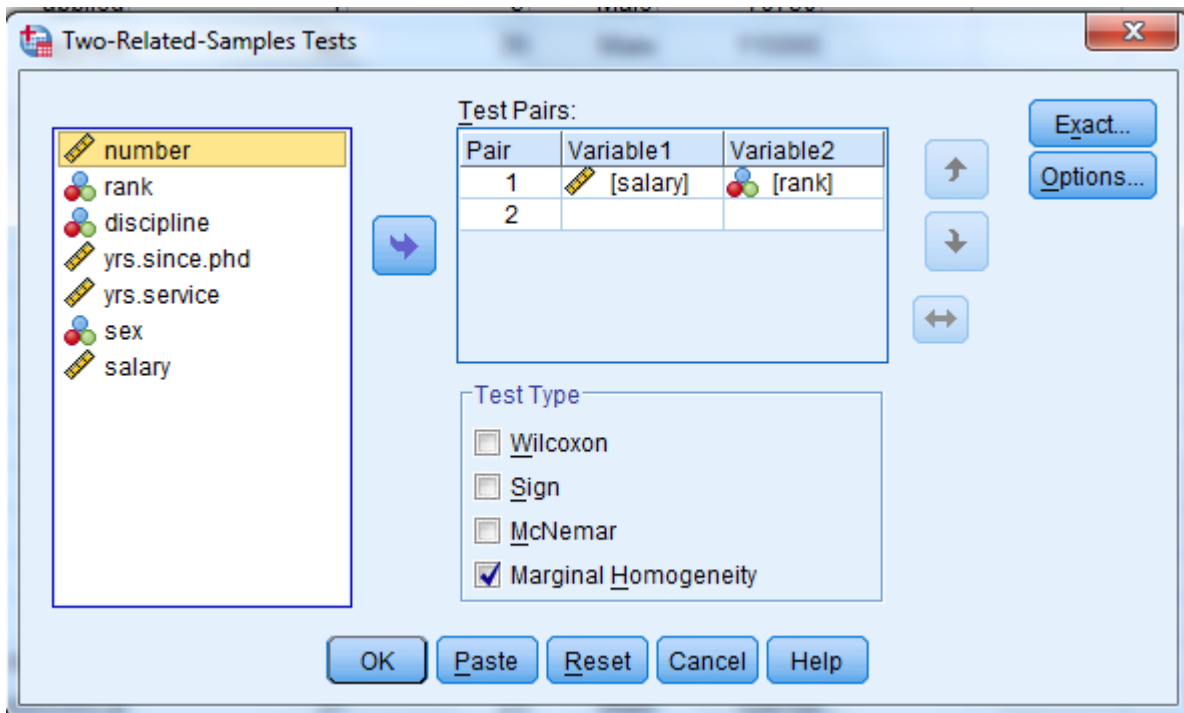
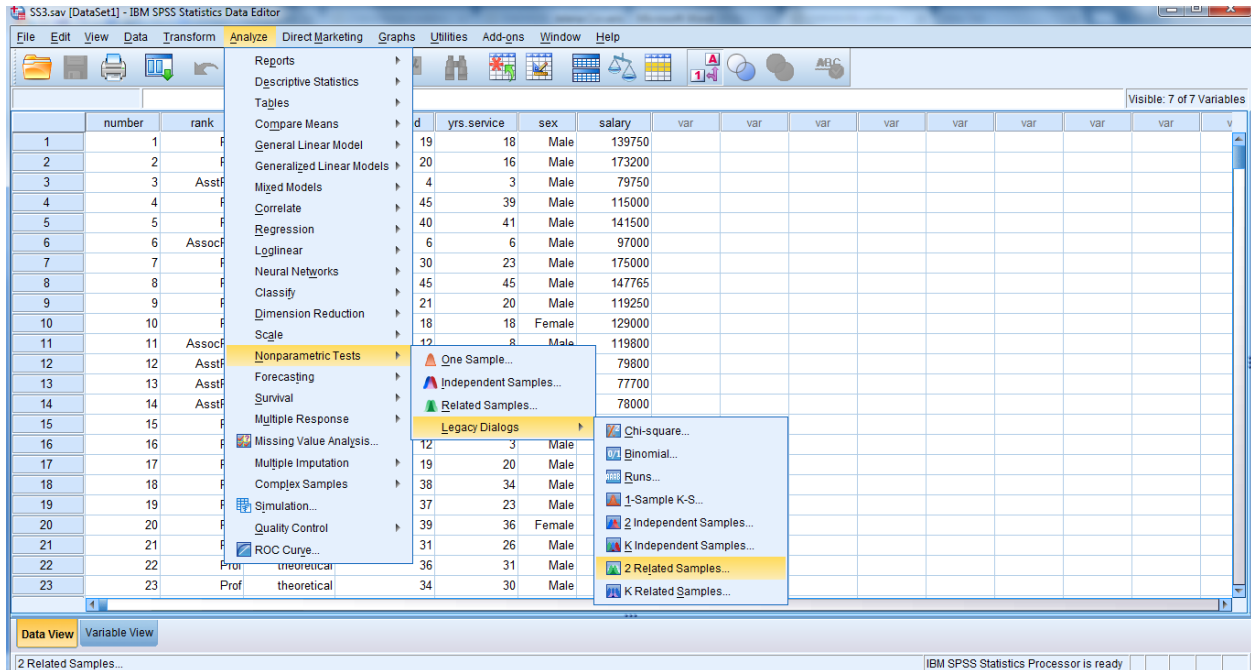
Test Statistics^a

	salary
Mann-Whitney U	15710,000
Wilcoxon W	32181,000
Z	-3,370
Asymp. Sig. (2-tailed)	,001

a. Grouping Variable:
discipline

Dobijamo jos manju p-vrednost (0,001) pa možemo zaključiti da visina plate ne zavisi ni od katedre na kojoj su zaposleni.

Proverićemo još da li visina plate zavisi od pozicije na kojoj su zaposleni. Za to ćemo koristiti test marginalne homogenosti, jer nam je promenljiva *rank* kategorijska sa 3 kategorije (docent, vandredni profesor i profesor).



Marginal Homogeneity Test

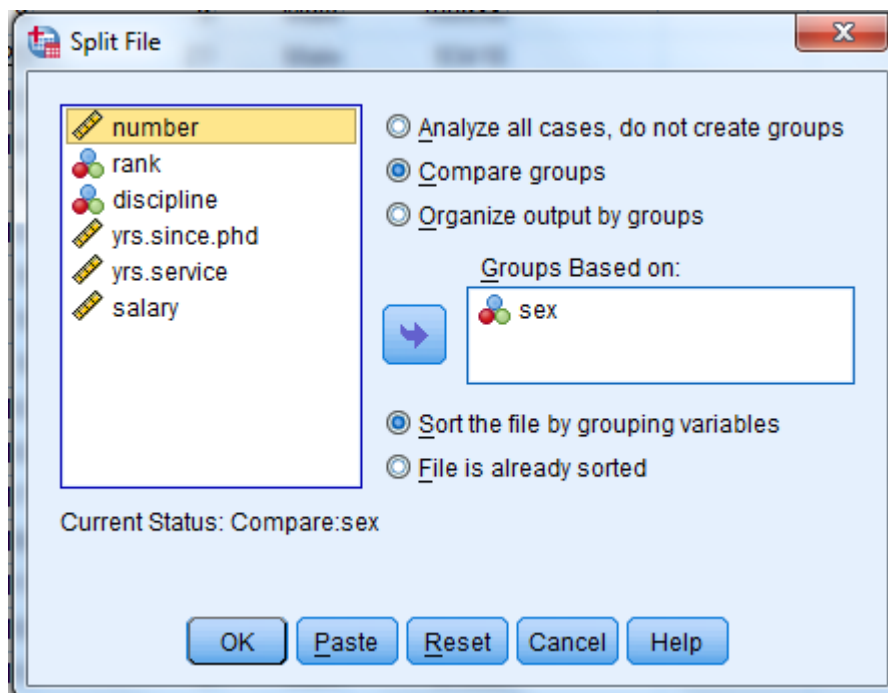
	salary & rank
Distinct Values	374
Off-Diagonal Cases	397
Observed MH Statistic	45141464,00
Mean MH Statistic	22571230,00
Std. Deviation of MH Statistic	1172171,132
Std. MH Statistic	19,255
Asymp. Sig. (2-tailed)	,000

Kao što vidimo p-vrednost testa je 0,000, pa odbacujemo nultu hipotezu, tj. visina plate ne zavisi od pozicije zaposlenih.

5. Ispitujemo da li su srednje vrednosti visine plate iste i za muškarce i za žene

Izračunaćemo koliko iznosi prosečna plata muškaraca, a kolika je ona kada su u pitanju žene. Prvo ćemo bazu podeliti u odnosu na pol.

Data → Split File



Sada kada smo podelili bazu možemo izračunati statistike.

Analyze → Descriptive Statistics → Frequencies

U prozor *Variables* prebacićemo promenljivu *salary*.

Statistics

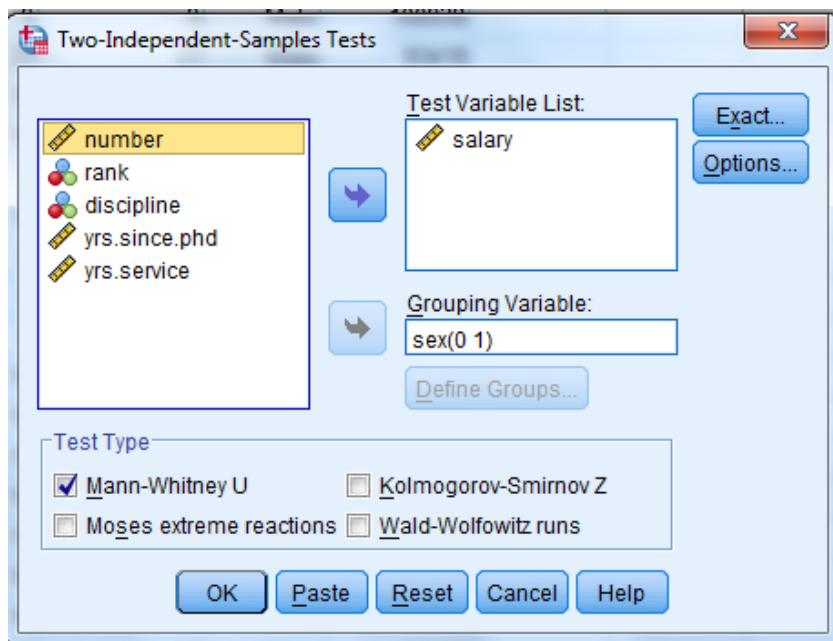
salary

Male	N	Valid	358
		Missing	0
	Mean		115090,42
	Median		108043,00
	Mode		74000 ^a
	Std. Deviation		30436,927
	Minimum		57800
	Maximum		231545
	Sum		41202370
Female	N	Valid	39
		Missing	0
	Mean		101002,41
	Median		103750,00
	Mode		72500 ^a
	Std. Deviation		25952,127
	Minimum		62884
	Maximum		161101
	Sum		3939094

a. Multiple modes exist. The smallest value is shown

Vidimo da prosečna plata kod muškaraca iznosi 115.090,42, minimalna je 57.800 a maksimalna 231.545. Dok je kod žena prosečna 101.002,41, minimalna 62.884 a maksimalna 161.101 . Primećujemo da su im prosečne plate slične, da je minimalna plata malo veća kada su u pitanju žene, dok je maksimalna plata kod muškaraca daleko veća nego kod žena.

Sada ćemo bazu podeliti po pozicijama na kojima su zaposleni, tj po promenljivoj *rank*. Bazu delimo kao u prethodnom slučaju, samo sada prebacujemo promenljivu *rank*. Želimo da ispitamo zavisnost visine plate u odnosu na pol na ovako podeljenoj bazi. Opet ćemo koristiti Mann-Whitney test iz istih razloga kao i ranije.



Mann-Whitney Test

Ranks

rank	sex	N	Mean Rank	Sum of Ranks
AssocProf	salary Male	54	33,50	1809,00
	Female	10	27,10	271,00
	Total	64		
AsstProf	salary Male	56	35,29	1976,00
	Female	11	27,45	302,00
	Total	67		
Prof	salary Male	248	134,15	33268,50
	Female	18	124,58	2242,50
	Total	266		

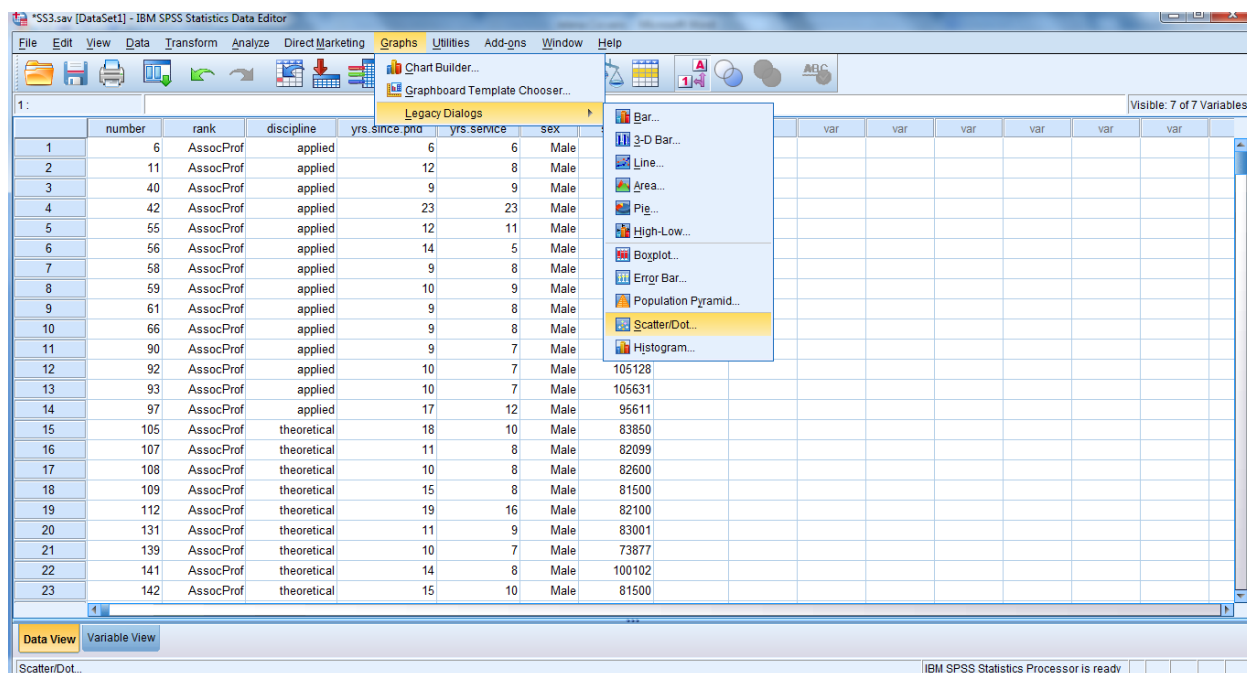
Test Statistics^a

rank		salary
AssocProf	Mann-Whitney U	216,000
	Wilcoxon W	271,000
	Z	-,998
	Asymp. Sig. (2-tailed)	,318
AsstProf	Mann-Whitney U	236,000
	Wilcoxon W	302,000
	Z	-1,219
	Asymp. Sig. (2-tailed)	,223
Prof	Mann-Whitney U	2071,500
	Wilcoxon W	2242,500
	Z	-,509
	Asymp. Sig. (2-tailed)	,611

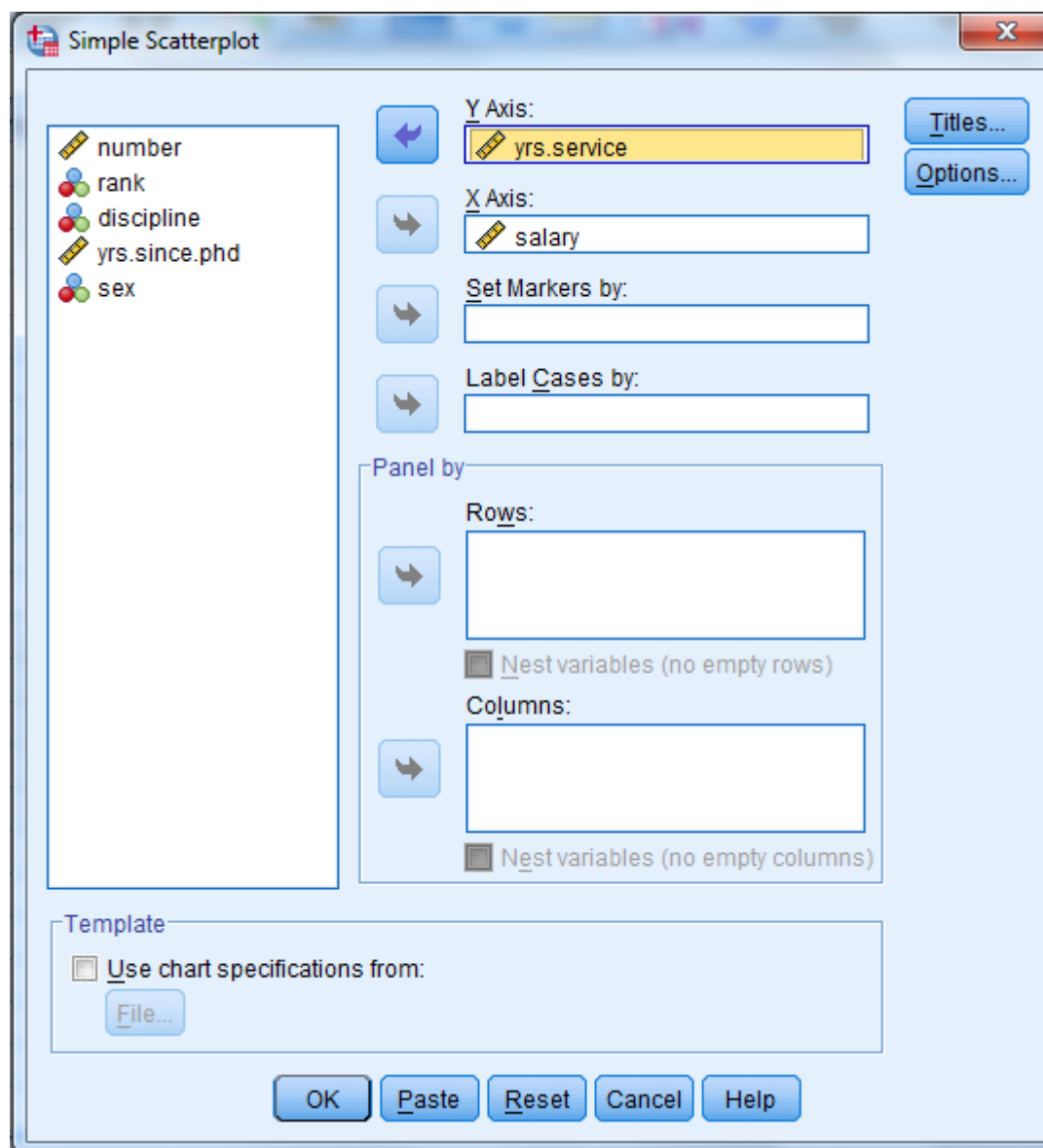
a. Grouping Variable: sex

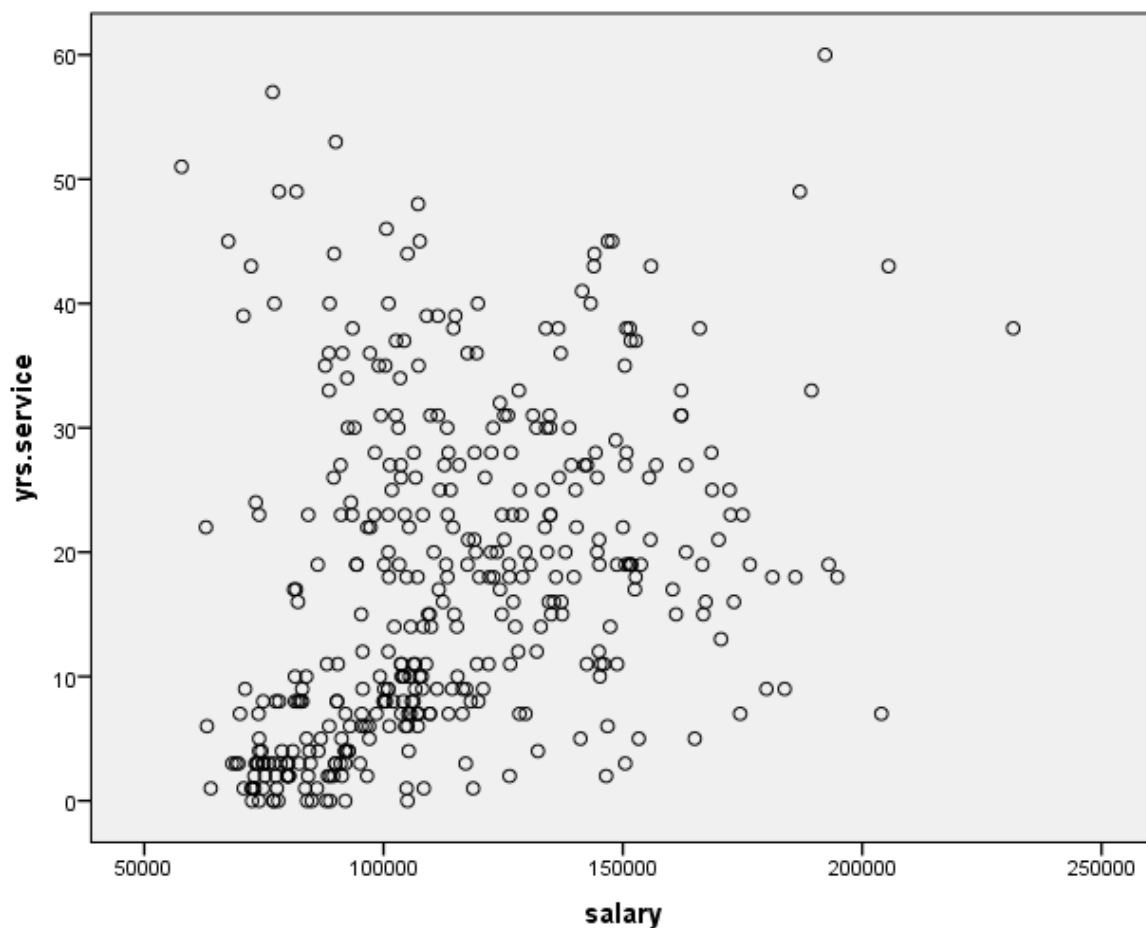
Možemo primetiti da su u sva tri slučaja p-vrednosti veće od 0,05, što znači da prihvatamo nultu hipotezu, odnosno visina plate zavisi od pola zaposlenih.

Želimo da proverimo vezu izmedju visine plate I godina radnog staža, kako su u pitanju numeričke promenljive ne možemo koristiti Hi-kvadrat test, pa ćemo primeniti korelacionu analizu. Prvo ćemo nacrtati dijagram raspršenosti.



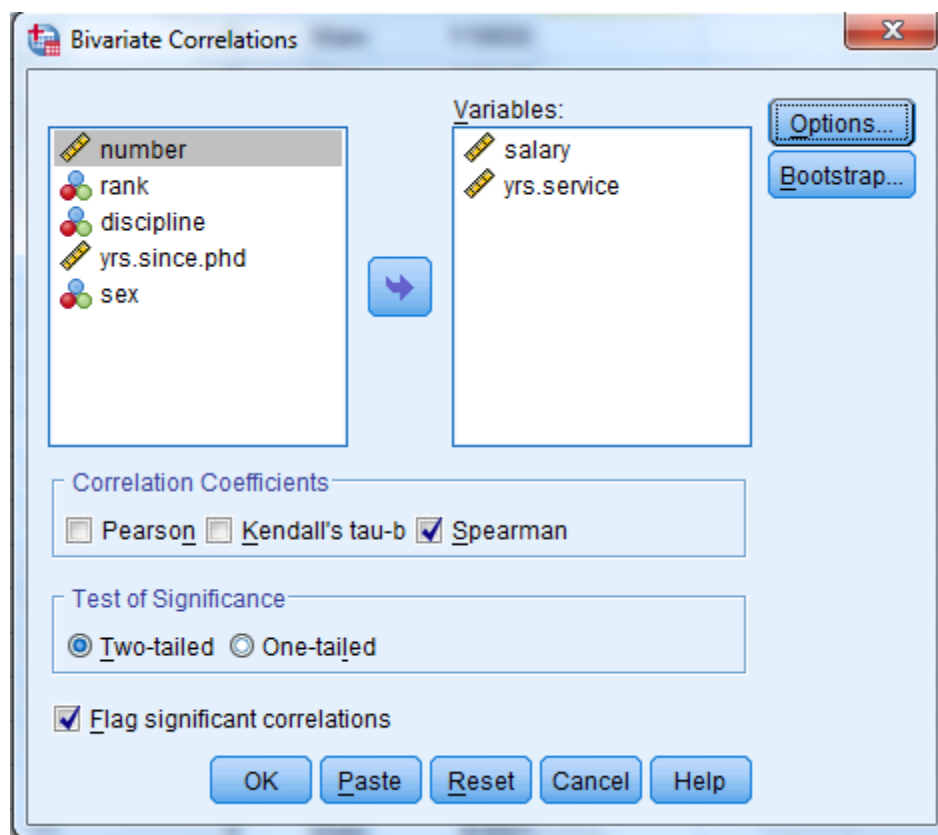
Kada kliknemo na Scatter/Dot za ispitivanje proste korelacije izaberemo Simple Scatter I kliknemo define, tada dobijamo sledeći prozor.





Odavde vidimo da postoji veza imedju ovih promenljivih. Sada ćemo izračunati koeficijente korelacije. Kako promenljiva salary nema normalnu raspodelu, ne moramo ni da proveravamo za drugu promenljivu, ne možemo koristiti Pirsonov koeficijent korelacije, pa ćemo koristiti Spirmanov koeficijent.

Analyze → Correlate → Bivariate



Correlations

			salary	yrs.service
Spearman's rho	salary	Correlation Coefficient	1,000	,425**
		Sig. (2-tailed)	.	,000
		N	397	397
	yrs.service	Correlation Coefficient	,425**	1,000
		Sig. (2-tailed)	,000	.
		N	397	397

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

Vidimo da je stepen korelacije 0,425. Možemo proveriti i Kendalov tau-b koeficijent, koji se koristi u slučaju kada postoje vrednosti koje se ponavljaju, trebalo bi da bude manji od Spearmanovog, jer u godinama radnog staža sigurno postoje vrednosti koje se ponavljaju. Sada ćemo osim Spearmanovog čekirati i Kendalov tau-b koeficijent.

Correlations

			salary	yrs.service
Kendall's tau_b	salary	Correlation Coefficient	1,000	,305**
		Sig. (2-tailed)	.	,000
		N	397	397
	yrs.service	Correlation Coefficient	,305**	1,000
		Sig. (2-tailed)	,000	.
		N	397	397
Spearman's rho	salary	Correlation Coefficient	1,000	,425**
		Sig. (2-tailed)	.	,000
		N	397	397
	yrs.service	Correlation Coefficient	,425**	1,000
		Sig. (2-tailed)	,000	.
		N	397	397

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

I vidimo da je Kendalov tau-b koeficijent korelacije 0,305 što je manje od Spirmanovog koji je 0,425. Takodje primećujemo da je p-vrednost testa 0,000 u oba slučaja, pa odbacujemo nultu hipotezu, odnosno možemo zaključiti da postoji linearna pozitivna veza izmedju visine plate I godina radnog staža.

6. Linearni model

Napravićemo model, koristićemo višestruku linearnu regresiju jer na visinu plate utiče više promenljivih.

Analyze → Regression → Linear

Linear Regression

Dependent: salary

Block 1 of 1

Previous Next

Independent(s):

- yrs.since.phd
- yrs.service
- sex

Method: Enter

Selection Variable:

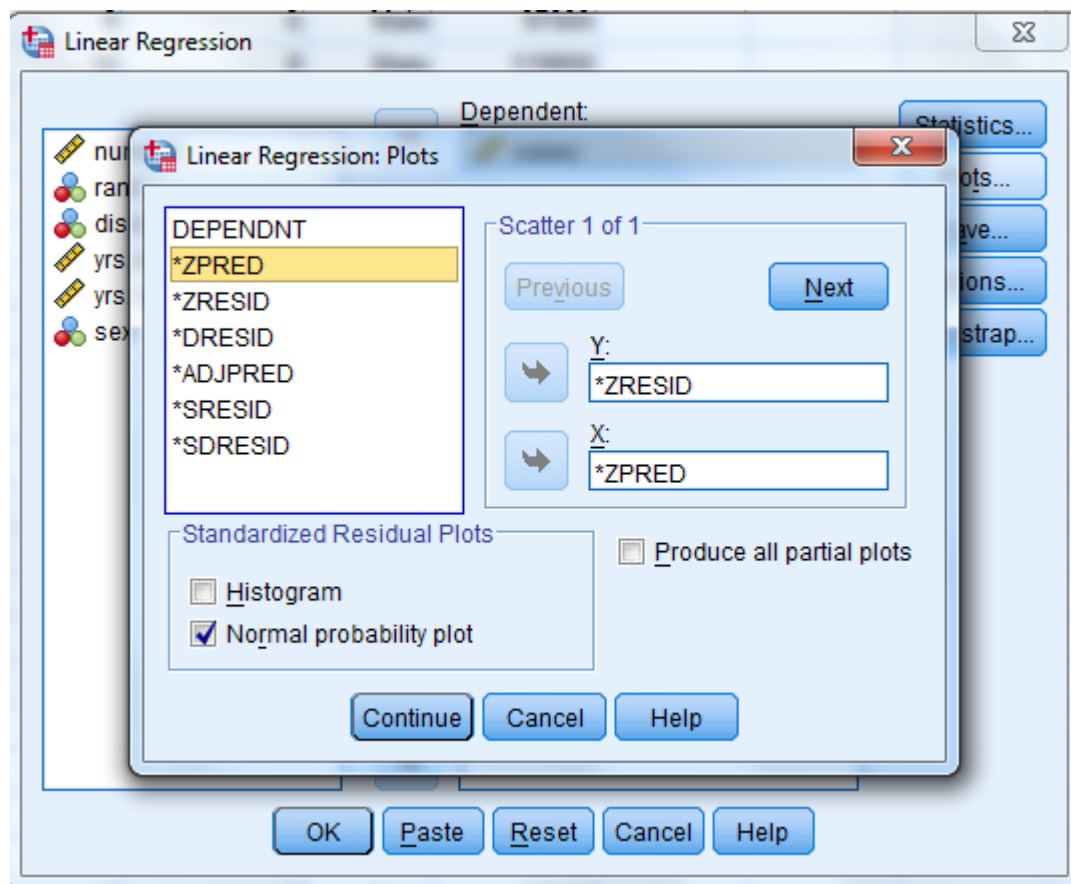
Case Labels:

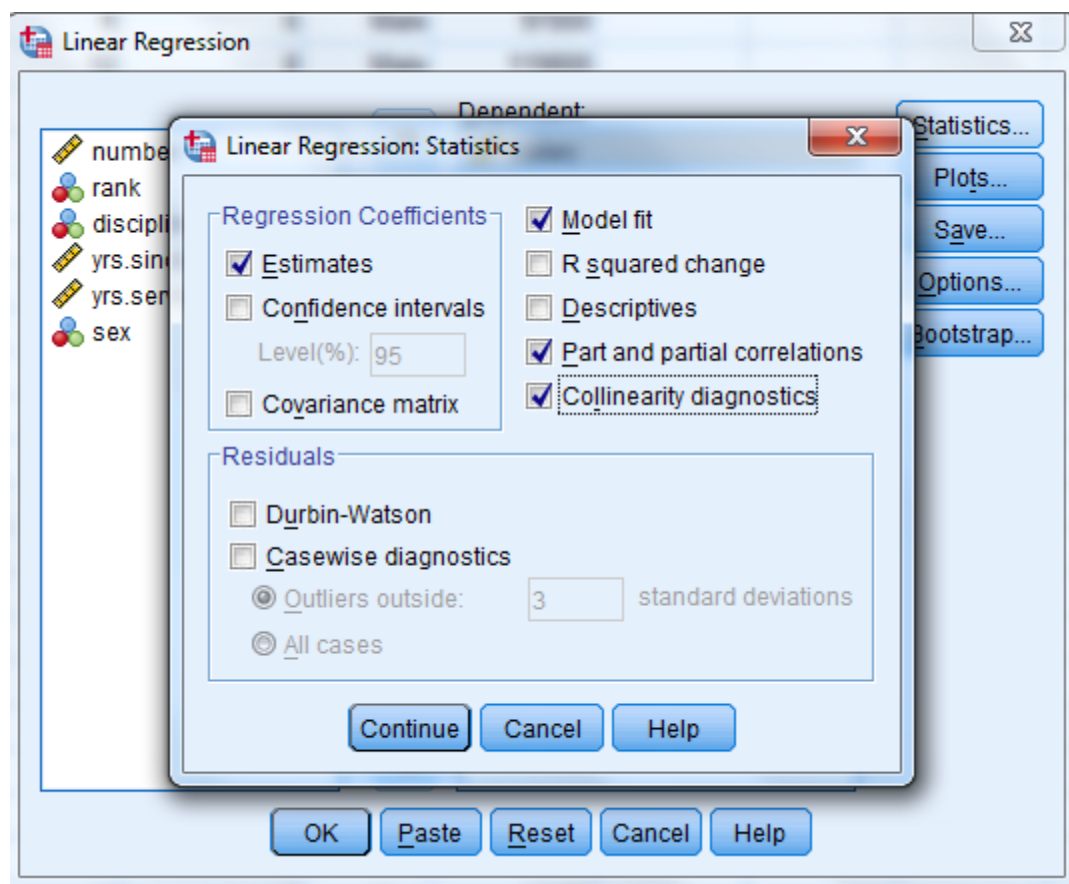
WLS Weight:

OK Paste Reset Cancel Help

Statistics... Plots... Save... Options... Bootstrap...

number rank discipline yrs.since.phd yrs.service sex





Kada kliknemo Ok, dobijamo sledeće rezultate:

Variables Entered/Removed^a

Model	Variables Entered	Variables Removed	Method
1	sex, discipline, rank, yrs. service, yrs. since.phd ^b	.	Enter

a. Dependent Variable: salary

b. All requested variables entered.

Model Summary^b

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,611 ^a	,373	,365	24131,122

a. Predictors: (Constant), sex, discipline, rank, yrs.service, yrs.since.phd

b. Dependent Variable: salary

ANOVA^a

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	1,356E+11	5	27123404633	46,579	,000 ^b
	Residual	2,277E+11	391	582311047,0		
	Total	3,633E+11	396			

a. Dependent Variable: salary

b. Predictors: (Constant), sex, discipline, rank, yrs.service, yrs.since.phd

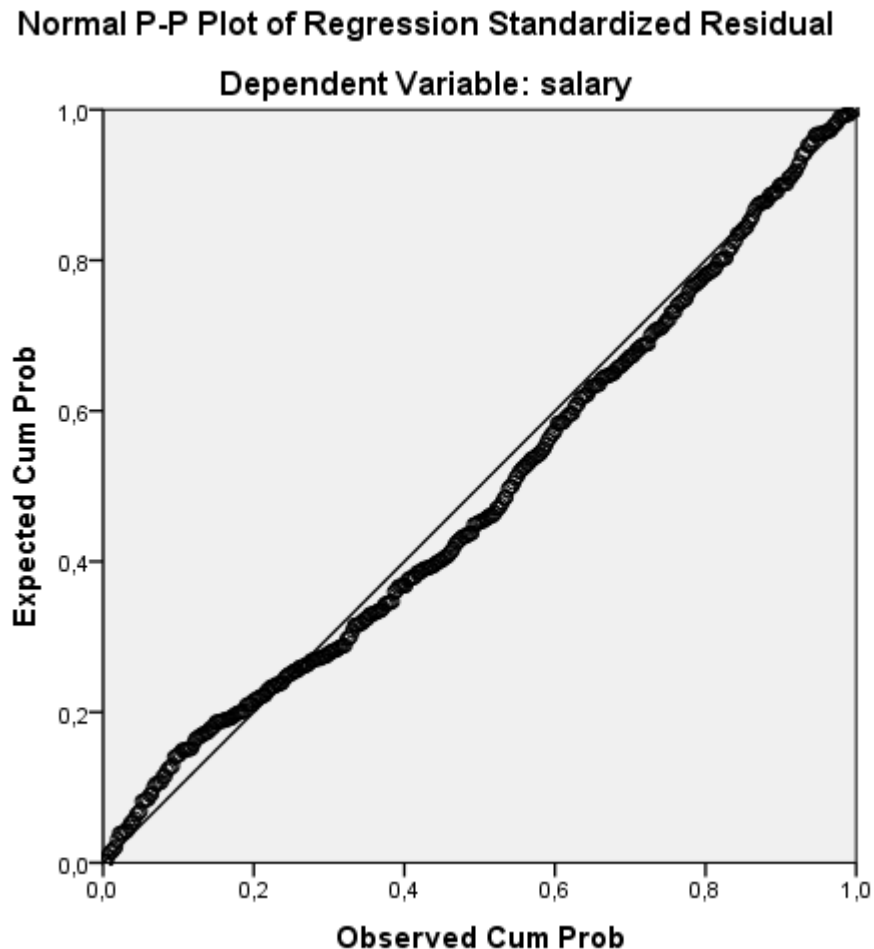
Iz druge tabele vidimo da nezavisne promenljive objašnjavaju 37,3% disperzije zavisne promenljive *salary*, iz toga mozemo zaključiti da je veza jaka.

Coefficients^a

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Correlations			Collinearity Statistics	
		B	Std. Error	Beta			Zero-order	Partial	Part	Tolerance	VIF
1	(Constant)	35505,288	6073,871		5,846	,000					
	rank	15691,030	1894,623	,392	8,282	,000	,522	,386	,332	,714	1,401
	discipline	15508,158	2503,739	,255	6,194	,000	,156	,299	,248	,943	1,060
	yrs.since.phd	1161,293	242,594	,494	4,787	,000	,419	,235	,192	,150	6,647
	yrs.service	-596,667	226,413	-,256	-2,635	,009	,335	-,132	-,106	,170	5,897
	sex	-5238,623	4130,810	-,052	-1,268	,205	-,139	-,064	-,051	,970	1,031

a. Dependent Variable: salary

Posmatramo koeficijente promenljivih, za naš model značajne promenljive su one čija je p-vrednost manja od 0,05 pa možemo zaključiti da promenljive *sex* i *yrs.service* nisu značajne za naš model. Proverićemo i normalnost reziduala:



Sa grafika možemo videti da nema značajnih odstupanja, pa možemo zaključiti da reziduali imaju normalnu raspodelu. Sada možemo napraviti model koji izgleda ovako:

$$Y = 35505,288 + 15691,03X_1 + 15508,158X_2 + 1161,293X_3$$

Nismo uključili pol i radni staž jer nam oni nisu značajni za model. Možemo koristiti i neki drugi metod osim *Enter*, ako promenimo metod u npr *Stepwise* dobijamo:

Variables Entered/Removed^a

Model	Variables Entered	Variables Removed	Method
1	rank		Stepwise (Criteria: Probability-of- F-to-enter <= , 050, Probability-of- F-to-remove >= ,100).
2	discipline		Stepwise (Criteria: Probability-of- F-to-enter <= , 050, Probability-of- F-to-remove >= ,100).
3	yrs.since.phd		Stepwise (Criteria: Probability-of- F-to-enter <= , 050, Probability-of- F-to-remove >= ,100).
4	yrs.service		Stepwise (Criteria: Probability-of- F-to-enter <= , 050, Probability-of- F-to-remove >= ,100).

a. Dependent Variable: salary

Dobijamo iste rezultate, mada nam ovaj metod izbacuje samo pol kao promenljivu koja nije značajna, ali je model bolji bez pola I radnog staža.

Literatura

- <https://vincentarelbundock.github.io/Rdatasets/datasets.html>
- <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%B5%D1%80-3/>