U ovom radu korišćeni su podaci iz ankete Living Standards Measurement Study (LSMS), koju je u ime Svetske banke sproveo Republički zavod za statistiku u maju i junu 2007. godine na teritoriji Republike Srbije.
Ova anketa pruža detaljne informacije o životnim uslovima, ekonomskim aktivnostima i demografskim karakteristikama stanovnika Srbije, a njen cilj je da pomogne u boljem razumevanju faktora koji utiču na socijalni i ekonomski status građana.
Podaci iz LSMS-a omogućavaju analizu različitih aspekata života, a u ovom radu fokusiraću se na analizu mesečnih zarada, kako bih utvrdio koji faktori utiču na njihove varijacije u Srbiji i u kojoj meri svaki od njih doprinosi tim razlikama.
Prosečna mesečna zarada korišćena je kao pokazatelj ekonomske situacije, pružajući uvid u životne uslove prosečnog stanovnika zemlje. Takođe, ukupne godišnje zarade predstavljaju značajan pokazatelj ekonomske situacije, jer direktno utiču na bruto domaći proizvod zemlje.

Anketna pitanja za odabrane varijable:
plata - Neto prihod prethodnog meseca od glavnog posla
obrazovanje - U upitniku je ispitanicima data ISCED skala, varijabla prekodirana tako da predstavlja godine obrazovanja
obr3 - Tri obrazovne kategorije (osnovna, srednja, visoka škola)
starost - Godine ispitanika u trenutku anketiranja
satiRada - Koliko sati je ispitanik radio na glavnom poslu u toku prethodne nedelje
zene - dve kategorije (Zena, Muskarac)
urban - dve kategorije (Grad, Selo)
region - četiri kategorije (Beograd, Vojvodina, Zapadna Srbija i Šumadija, Južna i jugoistočna Srbija)

Pozivanjem df.head(10) dobijamo uvid u prvih 10 ispitanika.

```
1 df.head(10)
```

	zene	starost	satiRada	plata	region	urban	obr3	obrazovanje
0	Zena	47	42,00	28000	Sumadija i Zapadna Srbija	Grad	Srednja skola	12,00
1	Muskarac	58	42,00	9000	Sumadija i Zapadna Srbija	Grad	Osnovno skola ili manje	8,00
2	Muskarac	24	42,00	11000	Sumadija i Zapadna Srbija	Grad	Osnovno skola ili manje	8,00
3	Zena	40	48,00	23000	Sumadija i Zapadna Srbija	Grad	Srednja skola	12,00
4	Muskarac	46	48,00	31000	Sumadija i Zapadna Srbija	Grad	Visoko obrazovanje	16,00
5	Muskarac	49	42,00	11000	Sumadija i Zapadna Srbija	Grad	Visoko obrazovanje	16,00
6	Zena	40	42,00	11000	Sumadija i Zapadna Srbija	Grad	Srednja skola	12,00
7	Muskarac	46	42,00	14500	Sumadija i Zapadna Srbija	Grad	Srednja skola	12,00
8	Zena	43	42,00	19800	Sumadija i Zapadna Srbija	Grad	Visoko obrazovanje	14,00
9	Zena	32	4,00	12200	Sumadija i Zapadna Srbija	Grad	Srednja skola	12,00

Na prvi pogled izgleda da su varijable satiRada i obrazovanje celi brojevi, ali su učitani u float formatu.

Da bi dobili detaljniji uvid u bazu zovemo sledecu funkciju

```
1 df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 5141 entries, 0 to 5140
Data columns (total 8 columns):
 #   Column       Non-Null Count  Dtype
---  ------       --------------  -----
 0   zene         5141 non-null   category
 1   starost      5141 non-null   int8
 2   satiRada     5135 non-null   float64
 3   plata        5141 non-null   int32
 4   region       5141 non-null   category
 5   urban        5141 non-null   category
 6   obr3         5141 non-null   category
 7   obrazovanje  5141 non-null   float32
dtypes: category(4), float32(1), float64(1), int32(1), int8(1)
memory usage: 106.1 KB
```

Uočavam da varijabla satiRada (zajedno sa satnicom) ima 6 ispitanika koji nisu odgovorili, njih uklanjam iz baze.

```
1 df = df[df['satiRada'].isna() == 0]
```

```
1 print(df['satiRada'].unique(),'\n',df['obrazovanje'].unique())
```

```
[42,00 48,00 4,00 70,00 56,00 80,00 30,00 28,00 40,00 8,00 60,00 35,00
 0,00 5,00 52,00 50,00 75,00 20,00 45,00 84,00 99,00 18,00 12,00 16,00
 33,00 36,00 65,00 55,00 88,00 2,00 25,00 9,00 7,00 10,00 1,00 49,00 54,00
 15,00 24,00 53,00 17,00 90,00 66,00 39,00 72,00 100,00 85,00 47,00 74,00
 44,00 58,00 21,00 63,00 96,00 32,00 43,00 46,00 37,00 3,00 6,00 38,00
 22,00 26,00 91,00 14,00 62,00 105,00 64,00 98,00 126,00 68,00 78,00 77,00
 13,00 51,00 34,00 23,00 59,00 31,00 67,00 57,00]
 [12,00 8,00 16,00 14,00 4,00 11,00 0,00 18,00 10,00 20,00]
```

Vizualnim pregledom potvrđujem sumnju da su vrednosti varijabla obrazovanje i satiRada celi brojevi skladišteni kao brojevi sa pokretnim zarezom, te ih sledecom linijom pretvaram u cele brojeve.

```
1 df['obrazovanje'] = df['obrazovanje'].astype(int)
2 df['satiRada'] = df['satiRada'].astype(int)
```

Ova promena ima estetski efekat, jer rezultuje lakšim pregledom podataka, ali i poboljšava memorijsku efikasnost, jer celobrojni podaci zauzimaju manje memorije nego brojevi sa pokretnim zarezom. Iako na bazi ovih dimenzija promena neće doneti veliku uštedu u memoriji, ovakve optimizacije postaju važnije kada se radi sa većim bazama podataka.

```
1 df.describe()
```

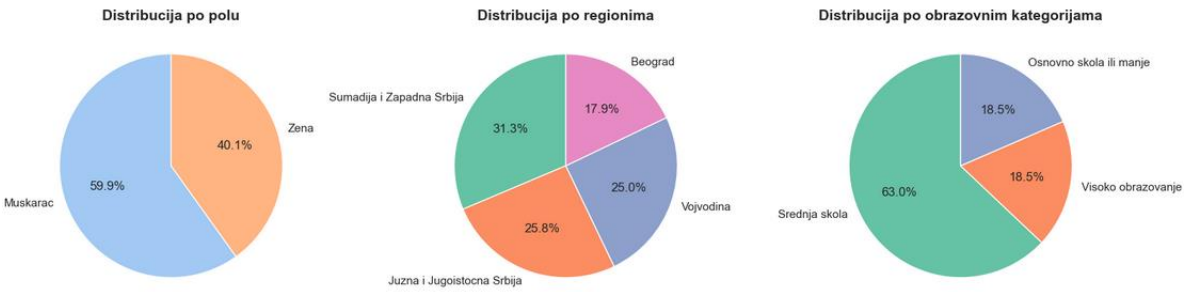
	starost	satiRada	plata	obrazovanje
count	5.135,00	5.135,00	5.135,00	5.135,00
mean	42,04	43,53	20.435,09	11,31
std	11,84	17,41	14.760,90	2,98
min	16,00	0,00	500,00	0,00
25%	33,00	40,00	12.000,00	11,00
50%	42,00	40,00	18.000,00	12,00
75%	51,00	50,00	25.000,00	12,00
max	84,00	126,00	230.000,00	20,00

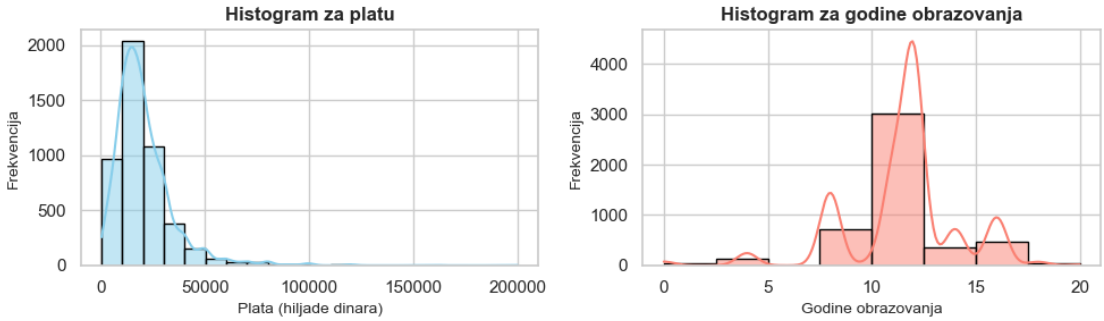
Uočavam problem u minimalnoj vrednosti varijable koja označava nedeljne sate rada i u maksimalnoj vrednosti varijable koja označava starost ispitanika. Kako je predmet ispitivanja isključivo radna populacija, iz baze izbacujem one koji imaju 0 radnih sati nedeljno i one koji su prešli starosnu granicu za penziju. U Srbiji 2007 godine starosna granica za muškarce bila je 65, a za žene 60 godina.

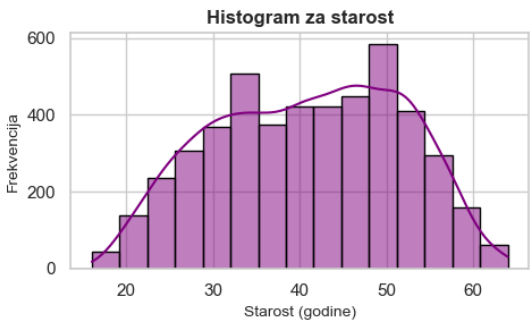
```
df = df[(df['satiRada'] > 0) & (((df['starost'] < 65) & (df['zene'] == 'Muskarac')) | ((df['starost']< 60) & (df['zene'] == 'Zena')))]
df.describe()
```

	starost	satiRada	plata	obrazovanje
count	4.776,00	4.776,00	4.776,00	4.776,00
mean	40,94	45,08	20.903,81	11,51
std	10,70	15,39	14.459,53	2,73
min	16,00	1,00	500,00	0,00
25%	32,00	40,00	12.000,00	11,00
50%	41,00	40,00	18.000,00	12,00
75%	50,00	50,00	25.000,00	12,00
max	64,00	126,00	200.000,00	20,00

Kako je baza sada kvalitativno očišćena, pre kvantitativnog čišćenja (izbacivanja outliera) korisno je vizualizovati podatke.







Kako je varijabla plata ciljana varijabla ovog rada bilo bi poželjno da njena raspodela ne odstupa značajno od normalne. Formalan Jarque - Berra (JB) test iz biblioteke scipy daje sledeću statistiku

```
stats.jarque_bera(df['plata'])
```

```
SignificanceResult(statistic=56177.39382059841, pvalue=0.0)
```

H0: Raspodela varijable plata se ne razlikuje od normalne
H1: Raspodela varijable plata se razlikuje od normalne

Statistika se poredi sa vrednošću hi kvadrat sa dva stepena slobode (5.99 za nivo značajnosti 5%). Kako je dobijena statistika veća od 5.99, odbacujem nultu hipotezu i zaključujem da se raspodela varijable plata značajno razlikuje od normalne raspodele.

Izbacivanje netipičnih vrednosti:
Možda izgleda intuitivno eliminisati netipične vrednosti za platu pomoću interkvartilne razlike, ali potrebno je malo dublje sagledati problem. Kako postoje podaci o broju nedeljnih radnih sati, varijabla plate se mora kontrolisati varijablom satiRada.
Na primer, ako ispitanik radi duplo više od proseka baze i ima duplo veću platu od prosečne, on ne predstavlja outlier jer ima prosečnu zaradu po radnom satu. S druge strane, ako ispitanik radi polovinu radnih sati od proseka baze i ima duplo veću platu od prosečne, on predstavlja outlier jer ima u proseku četiri puta veću zaradu po satu od nekoga ko radi prosečan broj sati.
Da bih preciznije identifikovao netipične vrednosti, formiram varijablu satnica kao odnos mesečne zarade i prosečnih mesečnih radnih sati. Kako je varijabla plata na mesečnom nivou, a satiRada na nedeljnom, sate množim sa 52 da bih dobio godišnji broj radnih sati, a potom delim sa 12 da bih dobio prosečan mesečni broj radnih sati.

```
df['satnica'] = df['plata'] / (df['satiRada'] * (52 / 12))
```

Na osnovu interkvartilne razlike izbacujem netipične vrednosti za nedeljne sate rada i preračunate satnice, eliminišući samo one ekstremne vrednosti koje ne prate uobičajene obrasce, dok se podaci koji odražavaju stvarne razlike u radnim satima i zaradama zadržavaju.

```
def izbaciOutliere(varijabla, df = df):
    Q1 = df[varijabla].quantile(0.25)
    Q3 = df[varijabla].quantile(0.75)
    IQR = Q3 - Q1

    donjaGranica = Q1 - 1.5*IQR
    gornjaGranica = Q3 + 1.5*IQR

    ukupnoIzbaceno = df[(df[varijabla] < donjaGranica) | (df[varijabla] > gornjaGranica)].shape[0]
    df = df[(df[varijabla] >= donjaGranica) & (df[varijabla] <= gornjaGranica)].reset_index(drop = True)
    print(f'Iz baze je izbacen svaki ispitanik kod koga je {varijabla} manja od {donjaGranica}, odnosno veca od {gornjaGranica}.')
    print(f'Iz baze je izbaceno  {ukupnoIzbaceno} outliera')
    print(f'Nakon izbacivanja outliera: {df.shape[0]}')
    return df
df = izbaciOutliere('satiRada', df = df)
df = izbaciOutliere('satnica',df = df)
```

```
Iz baze je izbacen svaki ispitanik kod koga je satiRada manja od 25.0, odnosno veca od 65.0.
Iz baze je izbaceno  766 outliera
Nakon izbacivanja outliera: 4010
Iz baze je izbacen svaki ispitanik kod koga je satnica manja od -43.269230769230816, odnosno veca od 256.7307692307693.
Iz baze je izbaceno  205 outliera
Nakon izbacivanja outliera: 3805
```

Nakon indirektnog čišćenja varijable plata baza izgleda ovako

```
df.describe()
```

	starost	satiRada	plata	obrazovanje	satnica
count	3.805,00	3.805,00	3.805,00	3.805,00	3.805,00
mean	40,57	44,05	19.246,63	11,53	102,84
std	10,61	7,38	9.384,32	2,50	50,75
min	16,00	25,00	1.000,00	0,00	3,85
25%	32,00	40,00	12.000,00	11,00	65,93
50%	41,00	40,00	18.000,00	12,00	92,31
75%	49,00	48,00	25.000,00	12,00	134,62
max	64,00	65,00	65.000,00	20,00	256,41

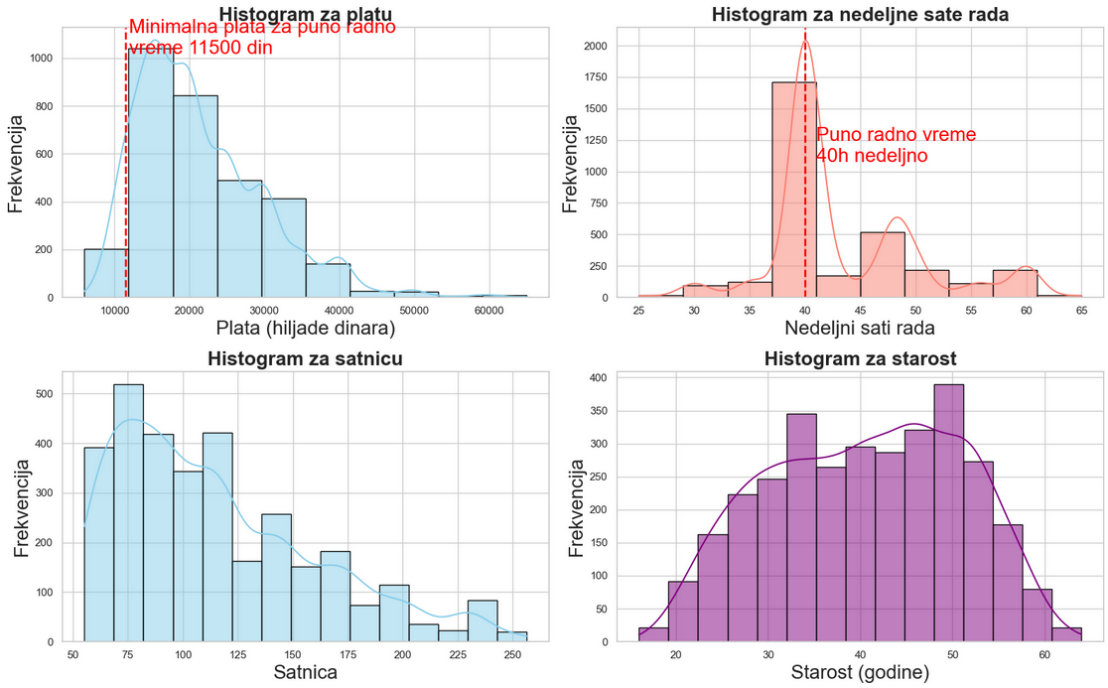
Iz izlaza vidim da je najmanja vrednost za satnicu 3.85 dinara. Kako je u vreme sprovođenja ankete minimalna propisana zarada po satu bila 55 dinara, izbacujem sve ispitanike kojima je preračunata satnica manja od minimalca. Takvih ispitanika ima 606.

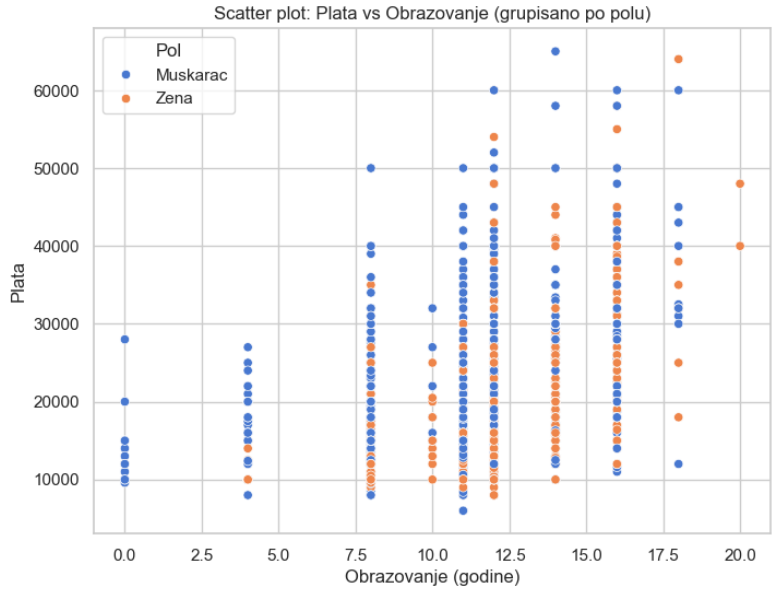
```
print((df['satnica'] < 55).sum())
df = df[df['satnica'] > 55]
df.describe()
```

```
606
```

	starost	satiRada	plata	obrazovanje	satnica
count	3.199,00	3.199,00	3.199,00	3.199,00	3.199,00
mean	40,52	43,48	21.420,56	11,80	115,09
std	10,37	6,89	8.570,89	2,38	45,74
min	16,00	25,00	6.000,00	0,00	55,29
25%	32,00	40,00	15.000,00	11,00	76,92
50%	41,00	40,00	20.000,00	12,00	103,85
75%	49,00	48,00	25.500,00	12,00	144,23
max	64,00	65,00	65.000,00	20,00	256,41

Početni broj ispitanika bio je 5141, a nakon čišćenja ostalo je 3199, što znači da je 38% podataka sadržavalo netipične vrednosti, bilo kvalitativne ili kvantitativne.





Na dijgramu raspršenosti može se primetiti da muškarci sa srednjom školom ili manje (do 12g obrazovanja) zarađuju više od žena.

Sledećom linijom koda dobijamo ispitanika sa najvećom zaradom

```
1 df.iloc[df['plata'].argmax()]
```

```
zene                     Muskarac
starost                        53
satiRada                       65
plata                       65000
region                    Beograd
urban                        Grad
obr3           Visoko obrazovanje
obrazovanje                    14
satnica                    230,77
Name: 440, dtype: object
```

Ponovljen test normalnosti daje statistiku JB = 982, što je i dalje veće od kritične vrednosti 5.99, ali se smanjila u odnosu na prethodnu vrednost (56177), što ukazuje na to da je postignut značajan napredak u eliminaciji netipičnih vrednosti.

```
1 stats.jarque_bera(df['plata'])
```

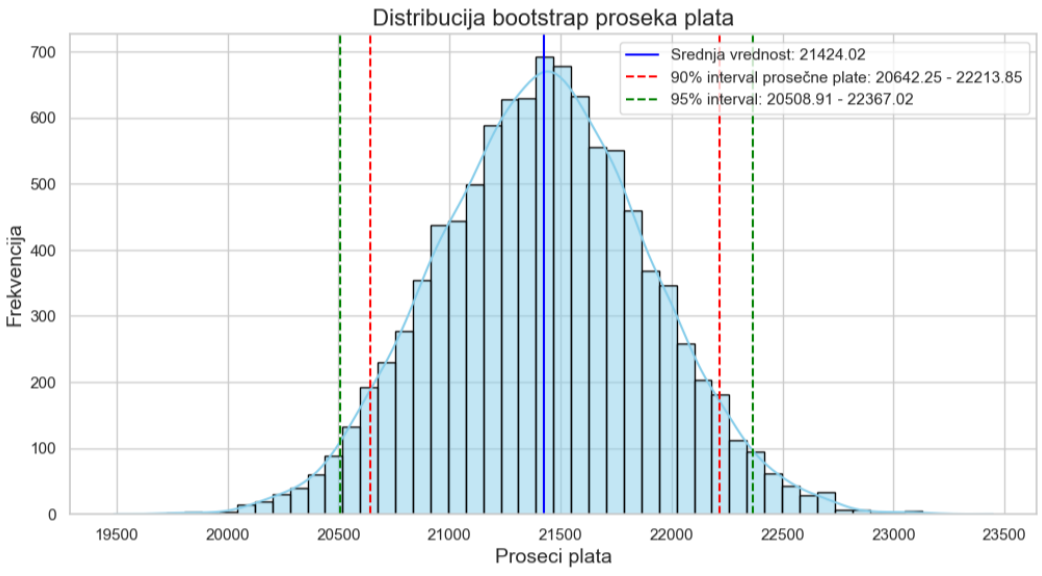
```
SignificanceResult(statistic=982.1632551748628, pvalue=5.320579964495498e-214)
```

Kako ću u ovom radu ocenjivati sredinu plate, dovoljno je da distribucija sredine plate iz slučajnih uzoraka ima normalnu raspodelu. Prema centralnoj graničnoj teoremi (CGT), u teoriji, ova pretpostavka bi trebala da važi za uzorake veće od 30, jer veliki uzorci obezbeđuju približavanje normalnoj distribuciji bez obzira na oblik osnovne distribucije. Međutim, kako bih dodatno osigurao validnost ove pretpostavke, odlučio sam da sprovedem dodatnu proveru normalnosti distribucije sredina plata koristeći bootstrap metodologiju.
Bootstrap je metod uzorkovanja sa vraćanjem (engl. resampling with replacement) koji omogućava generisanje velikog broja slučajnih uzoraka sa postojećih podataka. Iz baze se izvlači 10000 uzoraka obima 10% baze (n = 319) i iz svakog uzorka se čuva prosečna mesečna zarada i varijansa u zaradama (varijansa nije neophodna za CGT, ali će nešto kasnije biti potrebna). Definisana funkcija vraća dvojku (touple) listi izračunatih satistika. Na listi prosečnih plata urađen test normalnosti i dobijeno je JB = 2.55 što je manje od 5.99,
ne odbacujem nultu hipotezu i zaključujem da se raspodela sredina plata u uzorcima obima 319 ne razlikuje značajno od normalne raspodele.

```
1  def bootstrap(k, df = df , seed = 13,n = int(len(df) * 0.1)):
2      velicinaUzorka = len(df)
3
4      sredineUzoraka = []
5      varijansaUzorka = []
6      for i in range(k):
7          uzorak = df['plata'].sample(n = n , replace=True, random_state= seed + i)
8          sredineUzoraka.append(uzorak.mean())
9          varijansaUzorka.append(uzorak.var())
10     print(uzorak.shape)
11     return (pd.Series(sredineUzoraka), pd.Series(varijansaUzorka))
12 rezultatBs = bootstrap(10000)
13 print(f'Prosecna sredina od 10000 uzoraka je {form(rezultatBs[0].mean())}')
14 print(f'Jarque berra test normalnosti da je sredina plate normanlo raspodeljena {stats.jarque_bera(rezultatBs[0])}')
```

```
(319,)
Prosecna sredina od 10000 uzoraka je 21.423,96
Jarque berra test normalnosti da je sredina plate normanlo raspodeljena SignificanceResult(statistic=2.544900543653461, pvalue=0.28014435031229157)
```

Vizualizacija rezultata



Kako bi 90% odnosno 95% i 99% interval sadržao stvarnu prosečnu platu d se računa direktno iz objekta rezultatBs (iz realizovane raspodele uzimaju se kvantili od interesa)
d90 na histogramu predstavlja razdaljinu od plave do crvene linije, a d95 razdaljinu od plave do zelene linije.

```
d90 = (rezultatBs[0].quantile(0.95) - rezultatBs[0].quantile(0.05))/2
d95 = (rezultatBs[0].quantile(0.975) - rezultatBs[0].quantile(0.025))/2
d99 = (rezultatBs[0].quantile(0.995) - rezultatBs[0].quantile(0.005))/2
print('Da bi 90% interval ocene sredine plate sadrzao sredinu plate u bazi, odredjuje se d da bude', form(d90))
print('Da bi 95% interval ocene sredine plate sadrzao sredinu plate u bazi, odredjuje se d da bude', form(d95))
print('Da bi 99% interval ocene sredine plate sadrzao sredinu plate u bazi, odredjuje se d da bude', form(d99))
```

```
Da bi 90% interval ocene sredine plate sadrzao sredinu plate u bazi, odredjuje se d da bude 787,29
Da bi 95% interval ocene sredine plate sadrzao sredinu plate u bazi, odredjuje se d da bude 932,30
Da bi 99% interval ocene sredine plate sadrzao sredinu plate u bazi, odredjuje se d da bude 1.221,53
```

Za ocenu varijanse plate uzima se druga lista (u kojoj u sadržane varijanse plate) u objektu rezultatBs i uzima se njena srednja vrednost.

```
def obimUzorka(alfa, d):
    Z = stats.norm.ppf(1 - alfa/2)
    Sy2 = rezultatBs[1].mean()
    n0 = (np.square(Z) * Sy2) / np.square(d)
    n =int(1/ (1/n0 + 1/len(df)))
    return n
n90 = obimUzorka(10/100, d90)
n95 = obimUzorka(5/100, d95)
n99 = obimUzorka(1/100, d99)
print(f'Da bi sredina bila u 90% intervalu bira se uzorak od {n90}, za 95% interval {n95}, a za 99% interval {n99}')
```

```
Da bi sredina bila u 90% intervalu bira se uzorak od 291, za 95% interval 294, a za 99% interval 296
```

Dobija se da su veličine uzoraka potrebne za postizanje 90%, 95% I 99% intervala poverenja veoma slične. Ova sličnost može se objasniti činjenicom da je raspodela prosečnih plata u uzorcima bliska normalnoj raspodeli, što je i formalno testirano. Teorijski, za “savršeno” normalnu raspodelu za svaki nivo poverenja bi se dobio isti broj.
Razlika između širina intervala poverenja (d90,d95d i d99) srazmerno prati razliku između odgovarajućih Z vrednosti za 90% (Z=1.645), 95% (Z=1.960) i 99% (Z = 2.576) nivoa poverenja. S obzirom na to da je ova razlika mala, veličine uzoraka potrebne za postizanje ovih intervala su gotovo identične. Time se pokazuje da je uzorkovanje dovoljno stabilno za pouzdane ocene čak i pri različitim nivoima poverenja, uz male varijacije u veličini uzoraka.
Pošto su rezultati prilično slični i razlike u veličinama uzoraka su male, odlučio sam da uzmem uzorak od 296. Ovaj uzorak daje dovoljno precizne procene sredine plate za 95% interval poverenja, a takođe je dobar i za 90% i 99% intervale, što znači da mogu da budem siguran u rezultate za sve nivoe poverenja.

Kako sam bootstrap metodom proverio da li sredine uzoraka obima 320 imaju normalnu raspodelu, sada formalno testiram i za optimalnu veličinu uzorka. Kada se uzorak smanji, raspodela ne odstupa značajno od normalne.

```
rezultatBs = bootstrap(10000, n = n)
print(f'Prosecna sredina od 10000 uzoraka je {rezultatBs[0].mean()}')
print(f'Jarque berra test normalnosti da je sredina plate normanlo raspodeljena {(stats.jarque_bera(rezultatBs[0]))}')
```

```
(294,)
Prosecna sredina od 10000 uzoraka je 21421.20516360544
Jarque berra test normalnosti da je sredina plate normanlo raspodeljena SignificanceResult(statistic=1.0037000992280063, pvalue=0.6054095852412045)
```

Nakon što sam odredio optimalnu veličinu uzorka, proveravam koja numerička varijabla ima najvišu korelaciju sa ciljanom varijablom. Varijablu satnica ne uzimam u obzir jer je ona linearna kombinacija varijable plata.
Kako je najveća korelacija zavisne varijable sa varijablom obrazovanje, nju ću koristiti za količničko ocenjivanje, dok ću za regresiono ocenjivanje koristiti sve varijable.

```
print('Korelacija nedeljnih sati rada i plate\n',form(df.satiRada.corr(df.plata)),'\nKorelacija godina obrazovanja i plate\n', form(df.obrazovanje.corr(df.plat
```

```
Korelacija nedeljnih sati rada i plate
 0,19
Korelacija godina obrazovanja i plate
 0,41
Korelacija starosti i plate
 0,12
```

Pre nego što krenem sa uzorkovanjem, postaviću linearni regresioni model ću postaviti pomoću biblioteke statsmodels. Prvo je potrebno u bazu dodati konstantu pomocu statsmodels.add_const, nakon toga kategorijske varijable treba pretvoriti u veštačke pomoću pd.get_dummies i taj objekat spojiti sa numeričkim varijablama (uključujući konstantu).

```
df = sm.add_constant(df)
def vestackePromenjive(df):
    vestacke = pd.get_dummies(df[['region', 'zene','urban']], drop_first=True, dtype = float)
    x = pd.concat([df[['const','obrazovanje', 'satiRada','starost']], vestacke], axis=1)
    return x
X = vestackePromenjive(df)
X.head(10)
```

	const	obrazovanje	satiRada	starost	region_Vojvodina	region_Sumadija i Zapadna Srbija	region_Juzna i Jugoistocna Srbija	zene_Zena	urban_Grad
0	1,00	12	42	47	0,00	1,00	0,00	1,00	1,00
2	1,00	8	42	24	0,00	1,00	0,00	0,00	1,00
3	1,00	12	48	40	0,00	1,00	0,00	1,00	1,00
4	1,00	16	48	46	0,00	1,00	0,00	0,00	1,00
5	1,00	16	42	49	0,00	1,00	0,00	0,00	1,00
6	1,00	12	42	40	0,00	1,00	0,00	1,00	1,00
7	1,00	12	42	46	0,00	1,00	0,00	0,00	1,00
8	1,00	14	42	43	0,00	1,00	0,00	1,00	1,00
9	1,00	4	42	52	0,00	1,00	0,00	0,00	0,00
12	1,00	12	40	53	0,00	0,00	1,00	0,00	1,00

```
Y = df['plata']
model = sm.OLS(Y,X, hasconst=True).fit()
model.summary2()
```

Model:	OLS	Adj. R-squared:	0.264
Dependent Variable:	plata	AIC:	66049.4442
Date:	2025-01-10 22:10	BIC:	66104.0796
No. Observations:	3199	Log-Likelihood:	-33016.
Df Model:	8	F-statistic:	144.1
Df Residuals:	3190	Prob (F-statistic):	2.40e-207
R-squared:	0.265	Scale:	5.4093e+07

	Coef.	Std.Err.	t	P>|t|	[0.025	0.975]
const	-8032.7538	1278.5704	-6.2826	0.0000	-10539.6570	-5525.8506
obrazovanje	1440.6426	56.8535	25.3396	0.0000	1329.1696	1552.1156
satiRada	235.4926	19.1714	12.2835	0.0000	197.9031	273.0821
starost	115.9235	12.5749	9.2187	0.0000	91.2678	140.5792
region_Vojvodina	-2214.8099	398.0382	-5.5643	0.0000	-2995.2466	-1434.3733
region_Sumadija i Zapadna Srbija	-3109.4573	377.0626	-8.2465	0.0000	-3848.7670	-2370.1477
region_Juzna i Jugoistocna Srbija	-3406.7681	402.0411	-8.4737	0.0000	-4195.0532	-2618.4830
zene_Zena	-2172.9600	271.5610	-8.0017	0.0000	-2705.4119	-1640.5081
urban_Grad	1249.4440	278.7305	4.4826	0.0000	702.9348	1795.9531

Omnibus:	353.575	Durbin-Watson:	1.722
Prob(Omnibus):	0.000	Jarque-Bera (JB):	512.629
Skew:	0.834	Prob(JB):	0.000
Kurtosis:	4.032	Condition No.:	608

Iz izlaza se vidi da je ceo vektor objašnjavajućih promenjivih statistički značajan, a kofeicienti predstavljaju marginalnu promenu u plati.

Varijacije u vektoru X objašnjavaju 26,5% varijacija mesečne zarade. F statistika (144) je značajna, jer je veća od kritične vrednosti F(3190,8)=2.93 .
Reziduali nisu normalno raspoređeni (JB = 512 > 5.99), što govori da postoji problem u postavci modela. Pošto je plata pozitivno asimetrična, logaritamska transformacija bi mogla bolje opisati vezu između zavisne i nezavisnih promenjivih. Kako je cilj rada oceniti prosečnu platu, logaritamska transformacija nije optimalna, ali ću je ipak isprobati radi poređenja.

```
model = sm.OLS(np.log(Y),X, hasconst=True).fit()
model.summary2()
```

Model:	OLS	Adj. R-squared:	0.270
Dependent Variable:	plata	AIC:	1913.4886
Date:	2025-01-10 22:11	BIC:	1968.1239
No. Observations:	3199	Log-Likelihood:	-947.74
Df Model:	8	F-statistic:	148.5
Df Residuals:	3190	Prob (F-statistic):	6.46e-213
R-squared:	0.271	Scale:	0.10619

	Coef.	Std.Err.	t	P>|t|	[0.025	0.975]
const	8.5569	0.0566	151.0508	0.0000	8.4458	8.6680
obrazovanje	0.0651	0.0025	25.8346	0.0000	0.0601	0.0700
satiRada	0.0107	0.0008	12.5830	0.0000	0.0090	0.0124
starost	0.0053	0.0006	9.4485	0.0000	0.0042	0.0064
region_Vojvodina	-0.0976	0.0176	-5.5338	0.0000	-0.1322	-0.0630
region_Sumadija i Zapadna Srbija	-0.1274	0.0167	-7.6247	0.0000	-0.1601	-0.0946
region_Juzna i Jugoistocna Srbija	-0.1418	0.0178	-7.9589	0.0000	-0.1767	-0.1068
zene_Zena	-0.1012	0.0120	-8.4144	0.0000	-0.1248	-0.0777
urban_Grad	0.0584	0.0123	4.7301	0.0000	0.0342	0.0826

Omnibus:	32.317	Durbin-Watson:	1.718
Prob(Omnibus):	0.000	Jarque-Bera (JB):	22.054
Skew:	0.069	Prob(JB):	0.000
Kurtosis:	2.617	Condition No.:	608

Model gde je zavisna promenjiva ln(Plata) ima manju vrednost Bajesovog (BIC) i Akaikeovog (AIC) informativnog kriterijuma od lin lin modela. Log lin model ima reziduale koji su u većoj meri slični normalnoj raspodeli (JB = 22), iako još uvek nije moguće tvrditi da imaju potpuno normalnu raspodelu. Ove karakteristike sugerišu da log-lin model bolje odgovara podacima. Koeficient uz obrazovanje u ovom modelu iznosi 6.5, što znači da dodatna godina obrazovanja rezultuje većom zaradom za 6.5%.
U nekim poznatim studijama slični modeli pokazuju da koeficijent za obrazovanje varira između 10% i 14%, uz napomenu da obrazovanje često funkcioniše kao endogena promenjiva, što znači da može biti povezano sa neobserviranim faktorima koji takođe utiču na platu,što može dovesti do precenjivanja koeficijenta za obrazovanje u modelu.

```
Ytotal = df['plata'].sum()
Ybar = df['plata'].mean()
print(f'Total plate je {form(Ytotal)}, a prosek je {form(Ybar)}')
df = df.drop('satnica',axis =1 )
```

```
Total plate je 68.524.376,00, a prosek je 21.420,56
```

Prost slučajni uzorak:

```
1 psu(n)
```

```
Uzorak
       const  starost  satiRada     plata  obrazovanje
count 294,00   294,00    294,00    294,00       294,00
mean    1,00    41,96     42,83 20.730,72        11,50
std     0,00     9,99      6,09  8.459,21         2,55
min     1,00    17,00     25,00  8.000,00         0,00
25%     1,00    35,00     40,00 15.000,00        11,00
50%     1,00    43,00     40,00 20.000,00        12,00
75%     1,00    50,00     48,00 25.000,00        12,00
max     1,00    62,00     65,00 60.000,00        18,00

Df

         const  starost  satiRada     plata  obrazovanje
count 3.199,00 3.199,00  3.199,00  3.199,00     3.199,00
mean      1,00    40,52     43,48 21.420,56        11,80
std       0,00    10,37      6,89  8.570,89         2,38
min       1,00    16,00     25,00  6.000,00         0,00
25%       1,00    32,00     40,00 15.000,00        11,00
50%       1,00    41,00     40,00 20.000,00        12,00
75%       1,00    49,00     48,00 25.500,00        12,00
max       1,00    64,00     65,00 65.000,00        20,00
```

Kod za funckiju psu() će biti priložen u dodatku.
Slučajnim odabirom 294 ispitanika odabrali ljude koji su u proseku stariji, manje rade, manje su obrazovani. Prosečna plata u uzorku je manja oko 700 dinara.

Totali varijabli plata i obrazovanje u bazi i uzorku kada se koriguje ekspanzivnjim faktorom

```
Populacija:
Total mesecne zarade 68.524.376,00
Total godina obrazovanja 37.736,00
Uzorak (korigovan ekspanzivnim faktorom):
Total mesecne zarade 66.317.565,88
Total godina obrazovanja 36.788,50
```

KOLIČNIČKO OCENJIVANJE:

```
Količnik obeležja populacije: 1.815,889
Količnik obeležja uzorka: 1.802,671
Kolicnicka ocena sredine obeležja Y: 21.264,644
Stvarna sredina populacije 21.420,561
Kolicnicka ocena totala obeležja Y: 68.025.596,751
Stvarni total populacije 68.524.376,000
Pristrasnost varijable obrazovanje:
Za kolicnicku ocenu sredine -155,917
Za kolicnicku ocenu totala -498.779,249
Relativna pristrasnost : -0,728%
Standardna devijacija količničke ocena totala obeležja Y:
1.400.810,981
Standardna devijacija količničke ocena sredine obeležja Y:
437,890
Standardna devijacija količnika uzorka:
37,121
95% interval poverenja za sredinu mesecne zarade u populaciji (20406.395023372374, 22122.893348779566)
Raspon intervala za PSU je 1.716,498
```

95% interval poverenja sadrži stvarnu sredinu baze

REGRESIONO OCENJIVANJE:

```
 Varijabla urban_Grad, uklonjena (p vrednost 0,24)
 Varijabla zene_Zena, uklonjena (p vrednost 0,11)
Parametri modela:
                            OLS Regression Results
==============================================================================
Dep. Variable:                  plata   R-squared:                       0.274
Model:                            OLS   Adj. R-squared:                  0.259
Method:                 Least Squares   F-statistic:                     18.04
Date:                sub, 18 jan 2025   Prob (F-statistic):           9.37e-18
Time:                        16:44:51   Log-Likelihood:                -3028.3
No. Observations:                 294   AIC:                             6071.
Df Residuals:                     287   BIC:                             6096.
Df Model:                           6
Covariance Type:            nonrobust
=================================================================================================================
                                    coef    std err          t      P>|t|      [0.025      0.975]
-----------------------------------------------------------------------------------------------------------------
const                         -8263.8045   4168.203     -1.983      0.048   -1.65e+04     -59.680
obrazovanje                    1107.2416    170.707      6.486      0.000     771.246    1443.237
satiRada                        334.0071     70.085      4.766      0.000     196.061     471.953
starost                         125.3199     42.883      2.922      0.004      40.916     209.724
region_Vojvodina              -3566.8233   1315.704     -2.711      0.007   -6156.477    -977.170
region_Sumadija i Zapadna Srbija  -3469.7459   1234.119     -2.812      0.005   -5898.818   -1040.673
region_Juzna i Jugoistocna Srbija -5682.8526   1353.640     -4.198      0.000   -8347.174   -3018.532
==============================================================================
Omnibus:                       38.836   Durbin-Watson:                   1.872
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               54.146
Skew:                           0.862   Prob(JB):                     1.75e-12
Kurtosis:                       4.203   Cond. No.                         612.
------------------------------------------------------------------------------


Regresiona ocena sredine 21.073,673
 Standardna devijacije regresione ocene sredine
obrazovanje                       434.70
satiRada                          454.55
starost                           467.93
region_Vojvodina                  471.08
region_Sumadija i Zapadna Srbija  471.51
region_Juzna i Jugoistocna Srbija  465.06
dtype: float64
Pristrasnost regresionog ocenjivanja za ocenu sredine 342,955
Ako je poznat vektor objasnjavajucih promenjivih za celu populaciju druga ocena sredine je 21.073,673
A ako nije onda 20.730,718
90% interval poverenja za sredinu obelezja (20218.14812782354, 21929.197666517564)
Raspon intervala za PSU je 1.711,050
Tri ocene totala plate su: 67.414.679,598 67.414.679,598 66.317.565,881
A pravi total je 68.524.376,000
Srednja kvadratna greska kolicnickog ocenjivanja 216.058,076
Srednja kvadratna greska regresionog ocenjivanja 306.579,957
```

95% interval poverenja sadrži stvarnu sredinu baze

Kako je srednja kvadratna greška manja kod količničkog nego kod regresionog ocenjivanja za PSU(294), zaključujem da je količničko ocenjivanje preciznije.

Da bih proverio kako se ocene ponašaju sa povećanjem uzorka izvadio sam uzorak duplo veći od optimalnog. Postavljaću izlaze uporedno, karakteristike optimalnog uzorka će biti tamnije.

```
1 psu(2 * n)
```

Uzorak

	const	starost	satiRada	plata	obrazovanje
count	588,00	588,00	588,00	588,00	588,00
mean	1,00	41,19	42,79	21.136,96	11,75
std	0,00	10,07	6,34	8.260,45	2,42
min	1,00	17,00	25,00	8.000,00	0,00
25%	1,00	34,00	40,00	15.000,00	11,00
50%	1,00	42,00	40,00	20.000,00	12,00
75%	1,00	49,00	48,00	25.000,00	12,00
max	1,00	62,00	65,00	60.000,00	18,00

Df

	const	starost	satiRada	plata	obrazovanje
count	3.199,00	3.199,00	3.199,00	3.199,00	3.199,00
mean	1,00	40,52	43,48	21.420,56	11,80
std	0,00	10,37	6,89	8.570,89	2,38
min	1,00	16,00	25,00	6.000,00	0,00
25%	1,00	32,00	40,00	15.000,00	11,00
50%	1,00	41,00	40,00	20.000,00	12,00
75%	1,00	49,00	48,00	25.500,00	12,00
max	1,00	64,00	65,00	65.000,00	20,00

```
1 psu(n)
```

Uzorak

	const	starost	satiRada	plata	obrazovanje
count	294,00	294,00	294,00	294,00	294,00
mean	1,00	41,96	42,83	20.730,72	11,50
std	0,00	9,99	6,09	8.459,21	2,55
min	1,00	17,00	25,00	8.000,00	0,00
25%	1,00	35,00	40,00	15.000,00	11,00
50%	1,00	43,00	40,00	20.000,00	12,00
75%	1,00	50,00	48,00	25.000,00	12,00
max	1,00	62,00	65,00	60.000,00	18,00

Df

	const	starost	satiRada	plata	obrazovanje
count	3.199,00	3.199,00	3.199,00	3.199,00	3.199,00
mean	1,00	40,52	43,48	21.420,56	11,80
std	0,00	10,37	6,89	8.570,89	2,38
min	1,00	16,00	25,00	6.000,00	0,00
25%	1,00	32,00	40,00	15.000,00	11,00
50%	1,00	41,00	40,00	20.000,00	12,00
75%	1,00	49,00	48,00	25.500,00	12,00
max	1,00	64,00	65,00	65.000,00	20,00

Poređenjem srednjih vrednosti varijabli ova dva uzorka vidi se da veći uzorak ima sličnije karakteristike baze od manjeg uzorka za sve varijable osim za nedeljne sate rada gde je razlika između ova dva uzorka 0,04 radna sata nedeljno.

```
Populacija:
Total mesecne zarade 68.524.376,00
Total godina obrazovanja 37.736,00
Uzorak (korigovan ekspanzivnim faktorom):
Total mesecne zarade 67.617.126,99
Total godina obrazovanja 37.577,37
```

```
Populacija:
Total mesecne zarade 68.524.376,00
Total godina obrazovanja 37.736,00
Uzorak (korigovan ekspanzivnim faktorom):
Total mesecne zarade 66.317.565,88
Total godina obrazovanja 36.788,50
```

KOLIČNIČKO OCENJIVANJE:

```
Količnik obeležja populacije: 1.815,89
Količnik obeležja uzorka: 1.799,41
Kolicnicka ocena sredine obeležja Y: 21.226,19
Stvarna sredina populacije 21.420,56
Kolicnicka ocena totala obeležja Y: 67.902.569,25
Stvarni total populacije 68.524.376,00
Pristrasnost varijable obrazovanje:
Za kolicnicku ocenu sredine -194,38
Za kolicnicku ocenu totala -621.806,75
Relativna pristrasnost : -0,91%
Standardna devijacija količničničke ocena totala obeležja Y:
939.063,38
Standardna devijacija količničničke ocena sredine obeležja Y:
293,55
Standardna devijacija količnika uzorka:
24,89
95% interval poverenja za sredinu mesecne zarade u populaciji (20650.840524423442, 21801.531623688392)
Raspon intervala za PSU je 1.150,69
```

```
Količnik obeležja populacije: 1.815,889
Količnik obeležja uzorka: 1.802,671
Kolicnicka ocena sredine obeležja Y: 21.264,644
Stvarna sredina populacije 21.420,561
Kolicnicka ocena totala obeležja Y: 68.025.596,751
Stvarni total populacije 68.524.376,000
Pristrasnost varijable obrazovanje:
Za kolicnicku ocenu sredine -155,917
Za kolicnicku ocenu totala -498.779,249
Relativna pristrasnost : -0,728%
Standardna devijacija količničničke ocena totala obeležja Y:
1.400.810,981
Standardna devijacija količničničke ocena sredine obeležja Y:
437,890
Standardna devijacija količnika uzorka:
37,121
95% interval poverenja za sredinu mesecne zarade u populaciji (20406.395023372374, 22122.893348779566)
Raspon intervala za PSU je 1.716,498
```

Pristrasnost je veća nego u prvom uzorku, ali je standardna devijacija manja.
95% interval i dalje sadrži sredinu baze.

REGRESIONO OCENJIVANJE:

Parametri modela:

OLS Regression Results

Dep. Variable:	plata	R-squared:	0.242
Model:	OLS	Adj. R-squared:	0.232
Method:	Least Squares	F-statistic:	23.17
Date:	sub, 18 jan 2025	Prob (F-statistic):	7.34e-31
Time:	16:47:05	Log-Likelihood:	-6055.5
No. Observations:	588	AIC:	1.213e+04
Df Residuals:	579	BIC:	1.217e+04
Df Model:	8		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-7439.9401	3135.573	-2.373	0.018	-1.36e+04	-1281.456
obrazovanje	1376.5134	128.553	10.708	0.000	1124.026	1629.001
satiRada	218.5815	48.148	4.540	0.000	124.016	313.147
starost	114.8511	29.894	3.842	0.000	56.137	173.565
region_Vojvodina	-1393.3770	948.620	-1.469	0.142	-3256.534	469.779
region_Sumadija i Zapadna Srbija	-1759.8761	896.554	-1.963	0.050	-3520.771	1.019
region_Juzna i Jugoistocna Srbija	-2981.3055	968.014	-3.080	0.002	-4882.552	-1080.059
zene_Zena	-1861.0764	634.124	-2.935	0.003	-3106.541	-615.612
urban_Grad	1120.8712	639.607	1.752	0.080	-135.362	2377.104

Omnibus:	68.489	Durbin-Watson:	1.923
Prob(Omnibus):	0.000	Jarque-Bera (JB):	97.407
Skew:	0.825	Prob(JB):	7.05e-22
Kurtosis:	4.118	Cond. No.	652.

Varijabla urban_Grad, uklonjena (p vrednost 0,24)

Varijabla zene_Zena, uklonjena (p vrednost 0,11)

Parametri modela:

OLS Regression Results

Dep. Variable:	plata	R-squared:	0.274
Model:	OLS	Adj. R-squared:	0.259
Method:	Least Squares	F-statistic:	18.04
Date:	sub, 18 jan 2025	Prob (F-statistic):	9.37e-18
Time:	16:44:51	Log-Likelihood:	-3028.3
No. Observations:	294	AIC:	6071.
Df Residuals:	287	BIC:	6096.
Df Model:	6		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-8263.8045	4168.203	-1.983	0.048	-1.65e+04	-59.680
obrazovanje	1107.2416	170.707	6.486	0.000	771.246	1443.237
satiRada	334.0071	70.085	4.766	0.000	196.061	471.953
starost	125.3199	42.883	2.922	0.004	40.916	209.724
region_Vojvodina	-3566.8233	1315.704	-2.711	0.007	-6156.477	-977.170
region_Sumadija i Zapadna Srbija	-3469.7459	1234.119	-2.812	0.005	-5898.818	-1040.673
region_Juzna i Jugoistocna Srbija	-5682.8526	1353.640	-4.198	0.000	-8347.174	-3018.532

Omnibus:	38.836	Durbin-Watson:	1.872
Prob(Omnibus):	0.000	Jarque-Bera (JB):	54.146
Skew:	0.862	Prob(JB):	1.75e-12
Kurtosis:	4.203	Cond. No.	612.

U većem uzorku sve objašnjavajuće promenjive su statistički značajne.

```
Regresiona ocena sredine 21.267,60
 Standardna devijacije regresione ocene sredine
obrazovanje                          281,33
satiRada                             304,43
starost                              305,11
region_Vojvodina                     307,88
region_Sumadija i Zapadna Srbija     307,82
region_Juzna i Jugoistocna Srbija    306,15
zene_Zena                            307,54
urban_Grad                           305,57
dtype: float64
Pristrasnost regresionog ocenjivanja za ocenu sredine 130,65
Ako je poznat vektor objasnjavajucih promenjivih za celu populaciju druga ocena sredine je 21.267,60
A ako nije onda 21.136,96
90% interval poverenja za sredinu obelezja (20715.07213971545, 21820.13666132169)
Raspon intervala za PSU je 1.105,06
Tri ocene totala plate su: 68.035.066,48 68.035.066,48 67.617.126,99
A pravi total je 68.524.376,00
Srednja kvadratna greska kolicnickog ocenjivanja 123.952,82
Srednja kvadratna greska regresionog ocenjivanja 96.214,14
```

```
Regresiona ocena sredine 21.073,673
 Standardna devijacije regresione ocene sredine
obrazovanje                          434.70
satiRada                             454.55
starost                              467.93
region_Vojvodina                     471.08
region_Sumadija i Zapadna Srbija     471.51
region_Juzna i Jugoistocna Srbija    465.06
dtype: float64
Pristrasnost regresionog ocenjivanja za ocenu sredine 342,955
Ako je poznat vektor objasnjavajucih promenjivih za celu populaciju druga ocena sredine je 21.073,673
A ako nije onda 20.730,718
90% interval poverenja za sredinu obelezja (20218.14812782354, 21929.197666517564)
Raspon intervala za PSU je 1.711,050
Tri ocene totala plate su: 67.414.679,598 67.414.679,598 66.317.565,881
A pravi total je 68.524.376,000
Srednja kvadratna greska kolicnickog ocenjivanja 216.058,076
Srednja kvadratna greska regresionog ocenjivanja 306.579,957
```

Povećanjem uzorka smanjila se pristrasnost i standardna devijacija ocene.
95% interval sadrži sredinu baze.

Kako je srednja kvadratna greška manja kod regresionog ocenjivanja za PSU(588), zaključujem da je regresiono ocenjivanje preciznije.

Stratifikovani slučajni uzorak:

1) Regioni

Da bi uzorak bio stratifikovanpo regionima, baze je na slučajan način izabrano 96 ispitanika iz Šumadije, 72 iz Vojvodine, 68 iz Južne i jugoistočne Srbije i 58 iz Beograda. Na taj način u uzorku je dobijena identična raspodela po regionima kao i u bazi.

```
1 strat(n, 'region')
```

```
Strata
Sumadija i Zapadna Srbija     96
Vojvodina                     72
Juzna i Jugoistocna Srbija    68
Beograd                       58
Name: proportion, dtype: int32


Uzorak
Strata
Sumadija i Zapadna Srbija    0,33
Vojvodina                    0,24
Juzna i Jugoistocna Srbija   0,23
Beograd                      0,20
Name: proportion, dtype: float64

Df
Strata
Sumadija i Zapadna Srbija    0,33
Vojvodina                    0,24
Juzna i Jugoistocna Srbija   0,23
Beograd                      0,20
Name: proportion, dtype: float64
```

```
Uzorak
       const  starost  satiRada     plata  obrazovanje
count 294,00   294,00    294,00    294,00       294,00
mean    1,00    40,39     43,35 21.577,40        11,86
std     0,00    10,34      6,36  8.586,16         2,14
min     1,00    19,00     28,00  9.000,00         0,00
25%     1,00    32,00     40,00 15.000,00        11,00
50%     1,00    40,50     40,00 20.000,00        12,00
75%     1,00    49,00     48,00 25.000,00        12,00
max     1,00    64,00     65,00 60.000,00        18,00

Df
         const  starost  satiRada     plata  obrazovanje
count 3.199,00 3.199,00  3.199,00  3.199,00     3.199,00
mean      1,00    40,52     43,48 21.420,56        11,80
std       0,00    10,37      6,89  8.570,89         2,38
min       1,00    16,00     25,00  6.000,00         0,00
25%       1,00    32,00     40,00 15.000,00        11,00
50%       1,00    41,00     40,00 20.000,00        12,00
75%       1,00    49,00     48,00 25.500,00        12,00
max       1,00    64,00     65,00 65.000,00        20,00
```

```
1 psu(n)
```

```
Uzorak
       const  starost  satiRada     plata  obrazovanje
count 294,00   294,00    294,00    294,00       294,00
mean    1,00    41,96     42,83 20.730,72        11,50
std     0,00     9,99      6,09  8.459,21         2,55
min     1,00    17,00     25,00  8.000,00         0,00
25%     1,00    35,00     40,00 15.000,00        11,00
50%     1,00    43,00     40,00 20.000,00        12,00
75%     1,00    50,00     48,00 25.000,00        12,00
max     1,00    62,00     65,00 60.000,00        18,00

Df
         const  starost  satiRada     plata  obrazovanje
count 3.199,00 3.199,00  3.199,00  3.199,00     3.199,00
mean      1,00    40,52     43,48 21.420,56        11,80
std       0,00    10,37      6,89  8.570,89         2,38
min       1,00    16,00     25,00  6.000,00         0,00
25%       1,00    32,00     40,00 15.000,00        11,00
50%       1,00    41,00     40,00 20.000,00        12,00
75%       1,00    49,00     48,00 25.500,00        12,00
max       1,00    64,00     65,00 65.000,00        20,00
```

Proseci varijabla u stratifikovanom slučajnom uzorku su skoro identični onima u bazi, izlaz sa desne strane je poređenje sa PSU sa istim obimom.

```
Ocena prosečne plate stratifikovanim slucajnim uzorkom je : 21.571,69
Pristrasnost 151,13
Standardna devijacija ocene sredine stratifikovanim slucajnim uzorkom je 139,76?
95% interval ocene sredine stratifikovanim slucajnim uzorkom je (21296.62, 21846.76)
Ocena totala obeležja populacije (ukupna plata) stratifikovanim slucajnim uzorkom je : 69.007.835,08
Pristrasnost 483.459,08
Standardna devijacija ocene totala stratifikovanim slucajnim uzorkom je 447.104,67
95% interval ocene sredine stratifikovanim slucajnim uzorkom je (68127891.3, 69887778.86)
```

KOLIČNIČKO OCENJIVANJE:

```
KOLICNICKO OCENJIVANJE KOD STRATIFIKOVANOG UZORKA
KORELACIJA GODINA OBRAZOVANJA SA ZARADOM PO STRATUMIMA
Strata
Beograd                      0,48
Juzna i Jugoistocna Srbija   0,23
Sumadija i Zapadna Srbija    0,42
Vojvodina                    0,28
Name: obrazovanje, dtype: float64
Posebna količnička ocena prosecne mesecne zarade je:
 21.432,48
Pristrasnost 11,92
Standardna devijacija posebne količničke ocene sredine stratifikovanim slucajnim uzorkom je 427,75
95% interval ocene totala posebnom kolicnickom ocenom je (20591, 2714515)
Posebna kolicnicka ocena totala populacije 68.562.511,02
Pristrasnost 38.135,02
Standardna devijacija ocene totala stratifikovanim slucajnim uzorkom je 1.368.371,39

KOMBINOVANA KOLICNICKA OCENA
Kombinovana kolicnicka ocena sredine populacije je 21.450,62
Pristrasnost 30,06
Kombinovana kolicnicka ocena totala populacije je 68.620.535,99
Pristrasnost 96.159,99
```

1

```
Količnik obeležja populacije: 1.815,889
Količnik obeležja uzorka: 1.802,671
Kolicnicka ocena sredine obeležja Y: 21.264,644
Stvarna sredina populacije 21.420,561
Kolicnicka ocena totala obeležja Y: 68.025.596,751
Stvarni total populacije 68.524.376,000
Pristrasnost varijable obrazovanje:
Za kolicnicku ocenu sredine -155,917
Za kolicnicku ocenu totala -498.779,249
Relativna pristrasnost : -0,728%
Standardna devijacija količničke ocena totala obeležja Y:
1.400.810,981
Standardna devijacija količničke ocena sredine obeležja Y:
437,890
Standardna devijacija količnika uzorka:
37,121
95% interval poverenja za sredinu mesecne zarade u populaciji (20406.395023372374, 22122.893348779566)
Raspon intervala za PSU je 1.716,498
```

SSU ima manju pristrasnost i standardnu devijaciju ocene sredine.

REGRESIONO OCENJIVANJE:
Kako bih izbegao komplikacije ocenjivao sam prost linearni regresioni model u kojoj je objašnjavajuća promenljiva obrazovanje

```
                             const  obrazovanje
Beograd                   -1.366,79     2.168,19
Juzna i Jugoistocna Srbija  9.313,07       861,57
Sumadija i Zapadna Srbija   3.832,95     1.433,92
Vojvodina                   7.228,48     1.225,23
Ocene mesečne zarade po stratumima Strata
Beograd                      25.436,13
Juzna i Jugoistocna Srbija   19.420,64
Sumadija i Zapadna Srbija    20.556,48
Vojvodina                    21.415,97
dtype: float64
Pristrasnost po stratumima Strata
Beograd                       782,20
Juzna i Jugoistocna Srbija   -820,03
Sumadija i Zapadna Srbija      85,25
Vojvodina                     209,38
dtype: float64
Ocena sredine obeležja populacije (prosečna plata) stratifikovanim slucajnim uzorkom je : 21.462,51
Pristrasnost 41,95
Standardna devijacija ocene sredine stratifikovanim slucajnim uzorkom je 424,11
95% interval ocene sredine stratifikovanim slucajnim uzorkom je (20.627,83, 22.297,19)
Srednja kvadratna greska kolicnickog ocenjivanja 183.111,94
Srednja kvadratna greska regresionog ocenjivanja 181.625,31
```

```
Regresiona ocena sredine 21.073,673
 Standardna devijacije regresione ocene sredine
obrazovanje                        434.70
satiRada                           454.55
starost                            467.93
region_Vojvodina                   471.08
region_Sumadija i Zapadna Srbija   471.51
region_Juzna i Jugoistocna Srbija  465.06
dtype: float64
Pristrasnost regresionog ocenjivanja za ocenu sredine 342,955
Ako je poznat vektor objasnjavajucih promenjivih za celu populaciju druga ocena sredine je 21.073,673
A ako nije onda 20.730,718
90% interval poverenja za sredinu obelezja (20218.14812782354, 21929.197666517564)
Raspon intervala za PSU je 1.711,050
Tri ocene totala plate su: 67.414.679,598 67.414.679,598 66.317.565,881
A pravi total je 68.524.376,000
Srednja kvadratna greska kolicnickog ocenjivanja 216.058,076
Srednja kvadratna greska regresionog ocenjivanja 306.579,957
```

Za svaki region je ocenjena regresija: plata = β0 + β1 * obrazovanje. Pomoću dobijenog koeficienta β1 ocenio sam prosečnu platu za svaki stratum, plata u Beogradu je najprecenjenija, dok je u Južnoj i jugoistočnoj srbiji najpotcenjenija. Ocene za regione su ponderisane i dobijena je sveukupna ocena prosečne zarade koja je pristrasna na gore 42 dinara sa standardnom devijacijom 424. Pristrasnost regresionim ocenjivanjem je veća od pristrasnosti količničkim, dok je standardna devijacija ocene manja putem regresionog ocenjivanja.
Kako je srednja kvadratna greška manja kod regresionog ocenjivanja, zaključujem da je regresiono ocenjivanje u stratifikovanom slučajnom uzorku(294) za nijansu preciznije.

2) Pol

Da bi uzorak bio stratifikovanpo polu, baze je na slučajan način izabrano 173 muškaraca i 121 žena. Na taj način je u uzorku, kao i u bazi, 59% ispitanika je muškog pola. U ovom delu će tamniji izlazi biti iz uzorka sa regionalnom stratifikacijom.

```
1 strat(n,'zene')
```

```
Strata
Muskarac    173
Zena        121
Name: proportion, dtype: int32


Uzorak
Strata
Muskarac   0,59
Zena       0,41
Name: proportion, dtype: float64

Df
Strata
Muskarac   0,59
Zena       0,41
Name: proportion, dtype: float64
```

Uzorak

	const	starost	satiRada	plata	obrazovanje
count	294,00	294,00	294,00	294,00	294,00
mean	1,00	40,34	43,80	21.545,09	11,91
std	0,00	9,87	6,99	8.106,89	2,20
min	1,00	16,00	25,00	6.000,00	4,00
25%	1,00	32,00	40,00	15.000,00	11,00
50%	1,00	41,00	40,00	20.000,00	12,00
75%	1,00	47,75	48,00	25.450,00	12,00
max	1,00	63,00	65,00	45.000,00	18,00

Df

	const	starost	satiRada	plata	obrazovanje
count	3.199,00	3.199,00	3.199,00	3.199,00	3.199,00
mean	1,00	40,52	43,48	21.420,56	11,80
std	0,00	10,37	6,89	8.570,89	2,38
min	1,00	16,00	25,00	6.000,00	0,00
25%	1,00	32,00	40,00	15.000,00	11,00
50%	1,00	41,00	40,00	20.000,00	12,00
75%	1,00	49,00	48,00	25.500,00	12,00
max	1,00	64,00	65,00	65.000,00	20,00

Uzorak

	const	starost	satiRada	plata	obrazovanje
count	294,00	294,00	294,00	294,00	294,00
mean	1,00	40,39	43,35	21.577,40	11,86
std	0,00	10,34	6,36	8.586,16	2,14
min	1,00	19,00	28,00	9.000,00	0,00
25%	1,00	32,00	40,00	15.000,00	11,00
50%	1,00	40,50	40,00	20.000,00	12,00
75%	1,00	49,00	48,00	25.000,00	12,00
max	1,00	64,00	65,00	60.000,00	18,00

Df

	const	starost	satiRada	plata	obrazovanje
count	3.199,00	3.199,00	3.199,00	3.199,00	3.199,00
mean	1,00	40,52	43,48	21.420,56	11,80
std	0,00	10,37	6,89	8.570,89	2,38
min	1,00	16,00	25,00	6.000,00	0,00
25%	1,00	32,00	40,00	15.000,00	11,00
50%	1,00	41,00	40,00	20.000,00	12,00
75%	1,00	49,00	48,00	25.500,00	12,00
max	1,00	64,00	65,00	65.000,00	20,00

Prosečne vrednosti varijabli su gotovo identične, malu prednost ima regionalna stratifikacija (očekivano jer je podela po regionima detaljnija).

```
Ocena prosečne plate stratifikovanim slucajnim uzorkom je : 21.544,88
Pristrasnost 124,32
Standardna devijacija ocene sredine stratifikovanim slucajnim uzorkom je 136,15?
95% interval ocene sredine stratifikovanim slucajnim uzorkom je (21276.93, 21812.83)
Ocena totala obeležja populacije (ukupna plata) stratifikovanim slucajnim uzorkom je : 68.922.074,66
Pristrasnost 397.698,66
Standardna devijacija ocene totala stratifikovanim slucajnim uzorkom je 435.535,69
95% interval ocene sredine stratifikovanim slucajnim uzorkom je (68064899.72, 69779249.6)
Ocena prosečne plate stratifikovanim slucajnim uzorkom je : 21.571,69
Pristrasnost 151,13
Standardna devijacija ocene sredine stratifikovanim slucajnim uzorkom je 139,76?
95% interval ocene sredine stratifikovanim slucajnim uzorkom je (21296.62, 21846.76)
Ocena totala obeležja populacije (ukupna plata) stratifikovanim slucajnim uzorkom je : 69.007.835,08
Pristrasnost 483.459,08
Standardna devijacija ocene totala stratifikovanim slucajnim uzorkom je 447.104,67
95% interval ocene sredine stratifikovanim slucajnim uzorkom je (68127891.3, 69887778.86)
```

Pristrasnost i standardna devijacija su manje nego u prethodnom uzorku.

```
KOLICNICKO OCENJIVANJE KOD STRATIFIKOVANOG UZORKA
KORELACIJA GODINA OBRAZOVANJA SA ZARADOM PO STRATUMIMA
Strata
Muskarac   0,34
Zena       0,49
Name: obrazovanje, dtype: float64
Posebna količnička ocena prosecne mesecne zarade je:
 21.324,33
Pristrasnost -96,23
Standardna devijacija posebne količničke ocene sredine stratifikovanim slucajnim uzorkom je 415,61
95% interval ocene totala posebnom kolicnickom ocenom je (20506, 2637990)
Posebna kolicnicka ocena totala populacije 68.216.523,20
Pristrasnost -307.852,80
Standardna devijacija ocene totala stratifikovanim slucajnim uzorkom je 1.329.543,27

KOMBINOVANA KOLICNICKA OCENA
Kombinovana kolicnicka ocena sredine populacije je 21.329,98
Pristrasnost -90,59
Kombinovana kolicnicka ocena totala populacije je 68.234.594,23
Pristrasnost -289.781,77
```

```
KOLICNICKO OCENJIVANJE KOD STRATIFIKOVANOG UZORKA
KORELACIJA GODINA OBRAZOVANJA SA ZARADOM PO STRATUMIMA
Strata
Beograd                     0,48
Juzna i Jugoistocna Srbija  0,23
Sumadija i Zapadna Srbija   0,42
Vojvodina                   0,28
Name: obrazovanje, dtype: float64
Posebna količnička ocena prosecne mesecne zarade je:
 21.432,48
Pristrasnost 11,92
Standardna devijacija posebne količničke ocene sredine stratifikovanim slucajnim uzorkom je 427,75
95% interval ocene totala posebnom kolicnickom ocenom je (20591, 2714515)
Posebna kolicnicka ocena totala populacije 68.562.511,02
Pristrasnost 38.135,02
Standardna devijacija ocene totala stratifikovanim slucajnim uzorkom je 1.368.371,39

KOMBINOVANA KOLICNICKA OCENA
Kombinovana kolicnicka ocena sredine populacije je 21.450,62
Pristrasnost 30,06
Kombinovana kolicnicka ocena totala populacije je 68.620.535,99
Pristrasnost 96.159,99
```

Za količničku ocenu pristasnost je veća, a standardna devijacija manja.

```
REGRESIONO OCENJIVANJE
             const  obrazovanje
Muskarac 6.678,08     1.328,76
Zena       246,29     1.662,66
Ocene mesečne zarade po stratumima Strata
Muskarac   21.994,80
Zena       20.498,67
dtype: float64
Pristrasnost po stratumima Strata
Muskarac   -19,03
Zena       -74,11
dtype: float64
Ocena sredine obeležja populacije (prosečna plata) stratifikovanim slucajnim uzorkom je : 21.378,85
Pristrasnost -41,71
Standardna devijacija ocene sredine stratifikovanim slucajnim uzorkom je 411,97
95% interval ocene sredine stratifikovanim slucajnim uzorkom je (20.568,05, 22.189,66)
Srednja kvadratna greska kolicnickog ocenjivanja 181.994,46
Srednja kvadratna greska regresionog ocenjivanja 171.461,75
```

```
                             const  obrazovanje
Beograd                   -1.366,79     2.168,19
Juzna i Jugoistocna Srbija  9.313,07       861,57
Sumadija i Zapadna Srbija   3.832,95     1.433,92
Vojvodina                   7.228,48     1.225,23
Ocene mesečne zarade po stratumima Strata
Beograd                      25.436,13
Juzna i Jugoistocna Srbija   19.420,64
Sumadija i Zapadna Srbija    20.556,48
Vojvodina                    21.415,97
dtype: float64
Pristrasnost po stratumima Strata
Beograd                       782,20
Juzna i Jugoistocna Srbija   -820,03
Sumadija i Zapadna Srbija      85,25
Vojvodina                     209,38
dtype: float64
Ocena sredine obeležja populacije (prosečna plata) stratifikovanim slucajnim uzorkom je : 21.462,51
Pristrasnost 41,95
Standardna devijacija ocene sredine stratifikovanim slucajnim uzorkom je 424,11
95% interval ocene sredine stratifikovanim slucajnim uzorkom je (20.627,83, 22.297,19)
Srednja kvadratna greska kolicnickog ocenjivanja 183.111,94
Srednja kvadratna greska regresionog ocenjivanja 181.625,31
```

Linearni regresioni model ocenjuje prosečnu zaradu muškaraca sa 22000 dinara, a žena sa 20500 dinara.
U uzorku stratifikovanom prema polu SKG za regresiono ocenjivanje ima najmanju vrednost (čak za 5% manju).

3) Obrazovni nivo

Da bi uzorak bio stratifikovan po obrazovnom nivou, baze je na slučajan način izabrano 39 ispitanika sa osnovnim, 198 sa srednjim i 57 sa visokim obrazovanjem. Na taj način je u uzorku, kao i u bazi, 13%, 67% i 19% ispitanika imaju osnovno, srednje, odnosno visoko obrazovanje.

```
1 strat(n,'obr3')
```

```
Strata
Srednja skola              198
Visoko obrazovanje          57
Osnovno skola ili manje     39
Name: proportion, dtype: int32


Uzorak
Strata
Srednja skola             0,67
Visoko obrazovanje        0,19
Osnovno skola ili manje   0,13
Name: proportion, dtype: float64

Df
Strata
Srednja skola             0,67
Visoko obrazovanje        0,19
Osnovno skola ili manje   0,13
Name: proportion, dtype: float64
```

Uzorak

	const	starost	satiRada	plata	obrazovanje
count	294,00	294,00	294,00	294,00	294,00
mean	1,00	40,73	43,48	20.754,02	11,87
std	0,00	10,14	7,29	8.163,50	2,21
min	1,00	20,00	25,00	8.000,00	4,00
25%	1,00	32,00	40,00	15.000,00	11,00
50%	1,00	42,00	40,00	19.000,00	12,00
75%	1,00	48,75	48,00	25.000,00	12,00
max	1,00	63,00	65,00	50.000,00	16,00

Df

	const	starost	satiRada	plata	obrazovanje
count	3.199,00	3.199,00	3.199,00	3.199,00	3.199,00
mean	1,00	40,52	43,48	21.420,56	11,80
std	0,00	10,37	6,89	8.570,89	2,38
min	1,00	16,00	25,00	6.000,00	0,00
25%	1,00	32,00	40,00	15.000,00	11,00
50%	1,00	41,00	40,00	20.000,00	12,00
75%	1,00	49,00	48,00	25.500,00	12,00
max	1,00	64,00	65,00	65.000,00	20,00

Vektor objašnjavajućih promenljivih ovako stratifikovano uzoraka ima sličnije karakteristike bazi nego uzorak stratifikovan po polu i skoro iste karakteristike kada se uporedi sa regionalno stratifikovanim uzorkom. Dok zavisna varijabla odstupa mnogo više nego kod druga dva stratifikovana uzorka (skoro ista kao kod PSU). Čudno je da stratifikovani uzorak ima “veliko” odstupanje kao prost slučajni uzorak, ali za svaki slučaj želim da proverim da li postoji neka konceptualna greška zbog koje je plata u proseku manja kada se uzorak stratifikuje po obrazovnim nivoima.
U funkciji strat menjam argument seed kome je podrazumevana vrednost 1304 na 13040 kako bih dobio drugih 294 ispitanika

```
1 strat(n,'obr3')
```

```
Uzorak
       const  starost  satiRada     plata  obrazovanje
count 294,00   294,00    294,00    294,00       294,00
mean    1,00    40,82     43,22 21.854,13        11,73
std     0,00    10,07      6,97  8.522,41         2,52
min     1,00    16,00     25,00  8.000,00         0,00
25%     1,00    33,00     40,00 15.000,00        11,00
50%     1,00    42,00     40,00 20.000,00        12,00
75%     1,00    49,00     48,00 27.475,00        12,00
max     1,00    60,00     60,00 60.000,00        18,00

Df
         const  starost  satiRada     plata  obrazovanje
count 3.199,00 3.199,00  3.199,00  3.199,00     3.199,00
mean      1,00    40,52     43,48 21.420,56        11,80
std       0,00    10,37      6,89  8.570,89         2,38
min       1,00    16,00     25,00  6.000,00         0,00
25%       1,00    32,00     40,00 15.000,00        11,00
```

U ovom uzorku je prosečna zarada veća za 300 (u prošlom je bila manja za 700), tako da ne bih rekao da postoji konceptualni problem sa ovakvom stratifikacijom već sve varijacije pripisujem slučajnosti uzoraka i nastaviću da ispitujem karakteristike ovog uzorka.

```
Ocena prosečne plate stratifikovanim slucajnim uzorkom je : 21.846,22
Pristrasnost 425,66
Standardna devijacija ocene sredine stratifikovanim slucajnim uzorkom je 136,59?
95% interval ocene sredine stratifikovanim slucajnim uzorkom je (21577.4, 22115.04)
Ocena totala obeležja populacije (ukupna plata) stratifikovanim slucajnim uzorkom je : 69.886.066,48
Pristrasnost 1.361.690,48
Standardna devijacija ocene totala stratifikovanim slucajnim uzorkom je 436.950,18
95% interval ocene sredine stratifikovanim slucajnim uzorkom je (69026107.68, 70746025.27)


KOMBINOVANA KOLICNICKA OCENA
Kombinovana kolicnicka ocena sredine populacije je 21.984,63
Pristrasnost 564,07
Kombinovana kolicnicka ocena totala populacije je 70.328.822,07
Pristrasnost 1.804.446,07


REGRESIONO OCENJIVANJE
                            const  obrazovanje
Osnovno skola ili manje  11.876,54       959,22
Srednja skola            11.197,48       836,92
Visoko obrazovanje      -17.430,43     2.952,09
Ocene mesečne zarade po stratumima Strata
Osnovno skola ili manje   18.933,15
Srednja skola             21.003,33
Visoko obrazovanje        27.309,45
dtype: float64
Pristrasnost po stratumima Strata
Osnovno skola ili manje   2.144,68
Srednja skola               532,71
Visoko obrazovanje         -634,46
dtype: float64
Ocena sredine obeležja populacije (prosečna plata) stratifikovanim slucajnim uzorkom je : 21.943,96
Pristrasnost 523,40
Standardna devijacija ocene sredine stratifikovanim slucajnim uzorkom je 441,43
95% interval ocene sredine stratifikovanim slucajnim uzorkom je (21.075,18, 22.812,74)
Srednja kvadratna greska kolicnickog ocenjivanja 580.442,74
Srednja kvadratna greska regresionog ocenjivanja 468.806,38
```

Regresiona ocena je bolja od količničke ocene, ali su srednje kvadratne greške više nego duplo veće nego kod prethodna dva uzorka.
Kod regresionog ocenjivanja postoji konceptualni problem kada se stratifikacija obrazovnim kategorijama.

```
1 sorted(df['obrazovanje'].unique())
```

```
[0, 4, 8, 10, 11, 12, 14, 16, 18, 20]
```

Kako je obrazovanje kodirano po ISCED skali, za osnovno i srednje postoje tri , a za visoko obrazovanje četiri različite vrednosti. Znači kada se za svaki obrazovni nivo postavlja jednostavni linearni regresioni model u kome su godine obrazovanja objašnjavajuća promenljiva, gube se varijacije u objašnjavajućoj promenljivi, što nije negativno utiče na preciznost i varijansu dobijenih ocena. Zbog toga se ovakav vid stratifikacije ne preporučuje u praksi.