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)

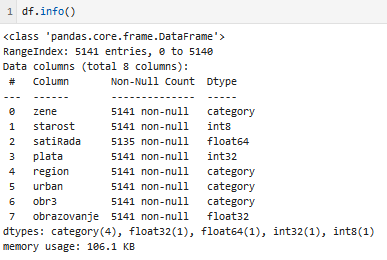
region - četiri kategorije (Beograd, Vojvodina, Zapadna Srbija i Šumadija, Južna i jugoistočna Srbija)

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



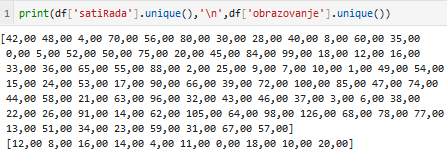
Na prvi pogled izgleda da su varijable satiRada i obrazovanje celi brojevi, ali su učitani u float formatu.

Da bi dobili detaljniji uvid u bazu zovemo sledecu funkciju



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

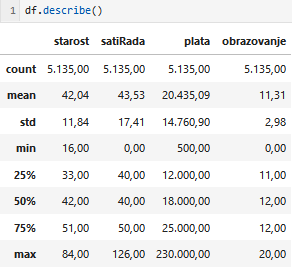




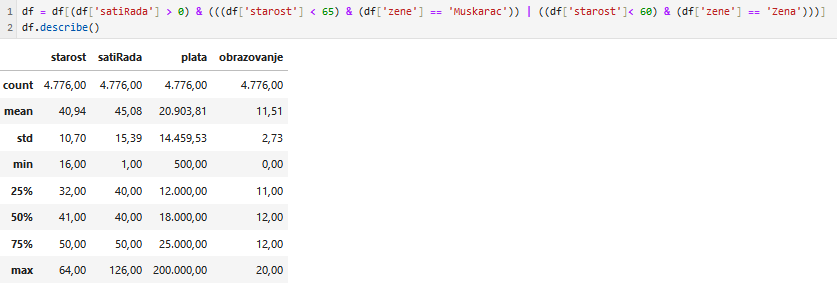
Vizualnim pregledom potvrđujem sumnju da su vrednosti varijabla obrazovanje i satiRada celi brojevi skladišteni kao brojevi sa pokretnim zarezom, te ih sledecom linijom pretvaram u cele brojeve.



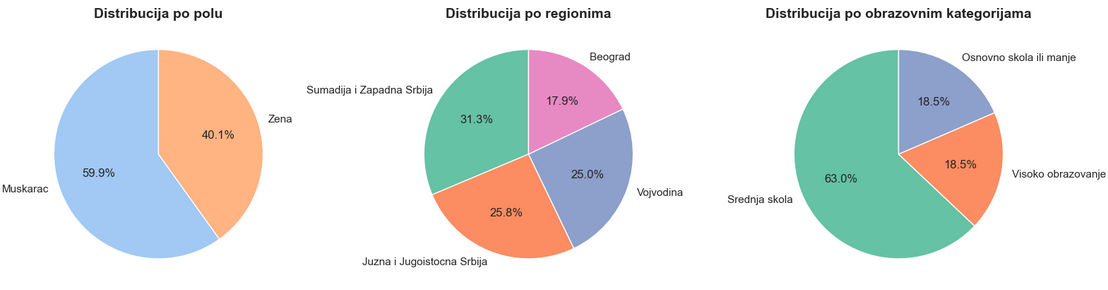
Ova promena ima estetski efekat, jer rezultuje lakšim pregledom podataka, ali i poboljšava memorijsku efikasnost, jer celobrojni podaci zauzimaju manje memorije nego brojevi sa pokretnim zarezom. Iako na bazi ovih dimenzija promena neće doneti veliku uštedu u memoriji, ovakve optimizacije postaju važnije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

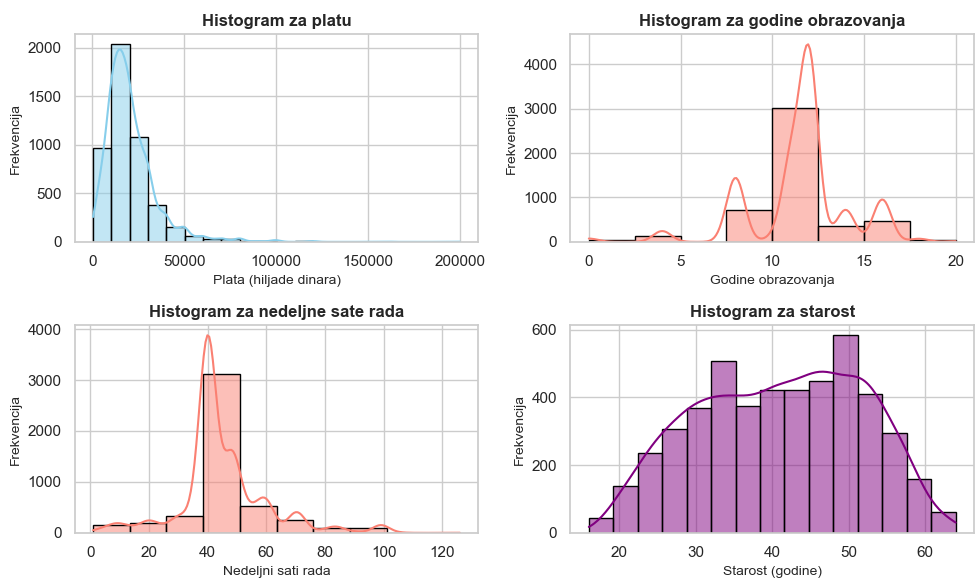


Uočavam problem u minimalnoj vrednosti varijable koja označava nedeljne sate rada i u maksimalnoj vrednosti varijable koja označava starost ispitanika. Kako je predmet ispitivanja isključivo radna populacija, iz baze izbacujem one koji imaju 0 radnih sati nedeljno i one koji su prešli starosnu granicu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



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



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:

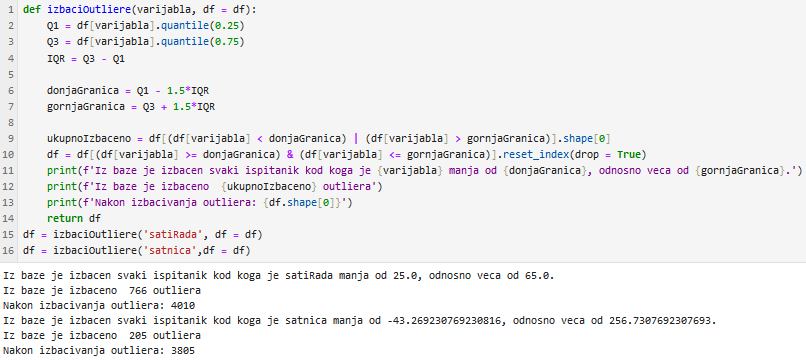
Možda izgleda intuitivno eliminisati netipične vrednosti za platu pomoću interkvartilne razlike, ali potrebno je malo dublje sagledati problem. Kako postoje podaci o broju nedeljnih radnih sati, varijabla plate se mora kontrolisati varijablom satiRada.

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.

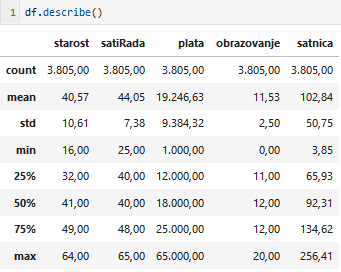
Da bih preciznije identifikovao netipične vrednosti, formiram varijablu satnica kao odnos mesečne zarade i prosečnih mesečnih radnih sati. Kako je varijabla plata na mesečnom nivou, a satiRada na nedeljnom, sate množim sa 52 da bih dobio godišnji broj radnih sati, a potom delim sa 12 da bih dobio prosečan mesečni broj radnih sati.



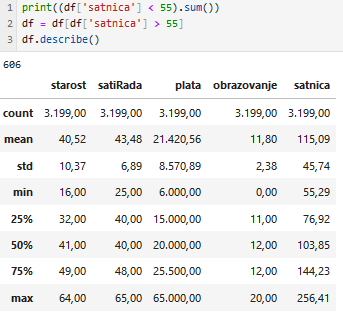
Na osnovu interkvartilne razlike izbacujem netipične vrednosti za nedeljne sate rada i preračunate satnice, eliminišući samo one ekstremne vrednosti koje ne prate uobičajene obrasce, dok se podaci koji odražavaju stvarne razlike u radnim satima i zaradama zadržavaju.



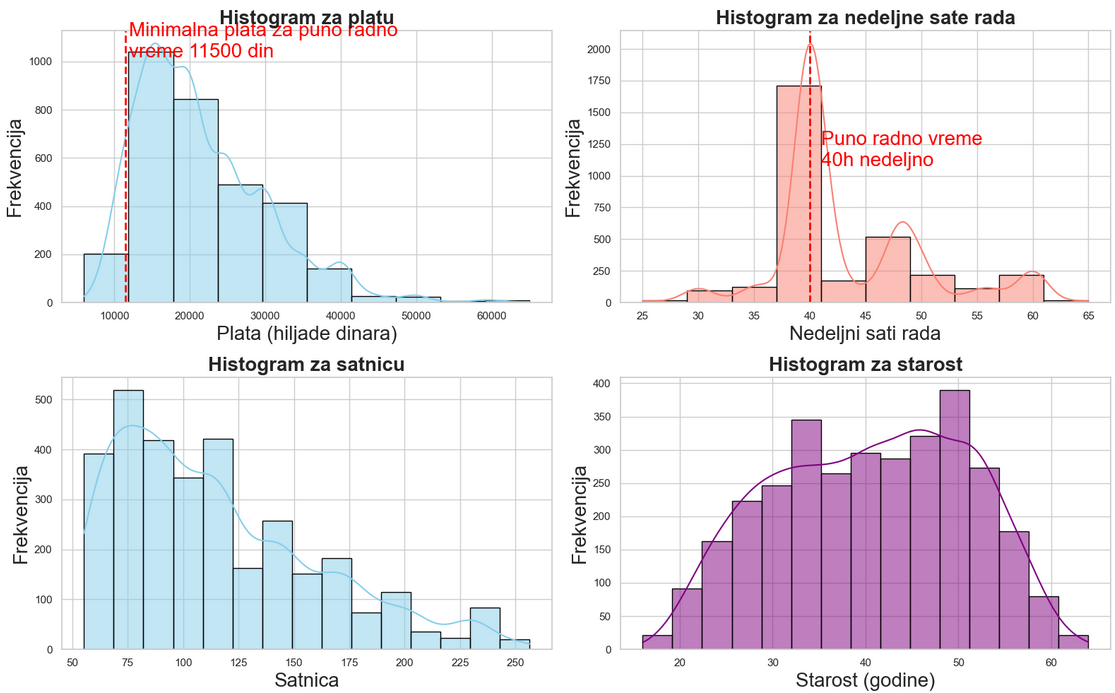
Nakon indirektnog čišćenja varijable plata baza izgleda ovako

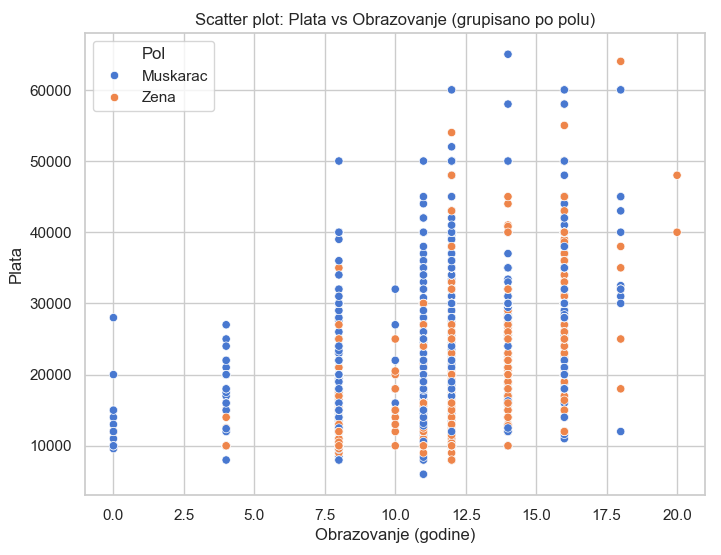


Iz izlaza vidim da je najmanja vrednost za satnicu 3.85 dinara. Kako je u vreme sprovođenja ankete minimalna propisana zarada po satu bila 55 dinara, izbacujem sve ispitanike kojima je preračunata satnica manja od minimalca. Takvih ispitanika ima 606.



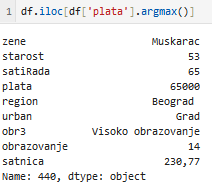
Početni broj ispitanika bio je 5141, a nakon čišćenja ostalo je 3199, što znači da je 38% podataka sadržavalo netipične vrednosti, bilo kvalitativne ili kvantitativne.





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



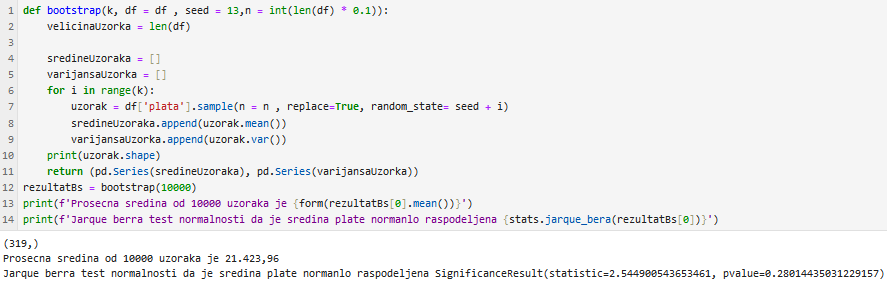
Ponovljen test normalnosti daje statistiku JB = 982, što je i dalje veće od kritične vrednosti 5.99, ali se smanjila u odnosu na prethodnu vrednost (56177), što ukazuje na to da je postignut značajan napredak u eliminaciji netipičnih vrednosti.



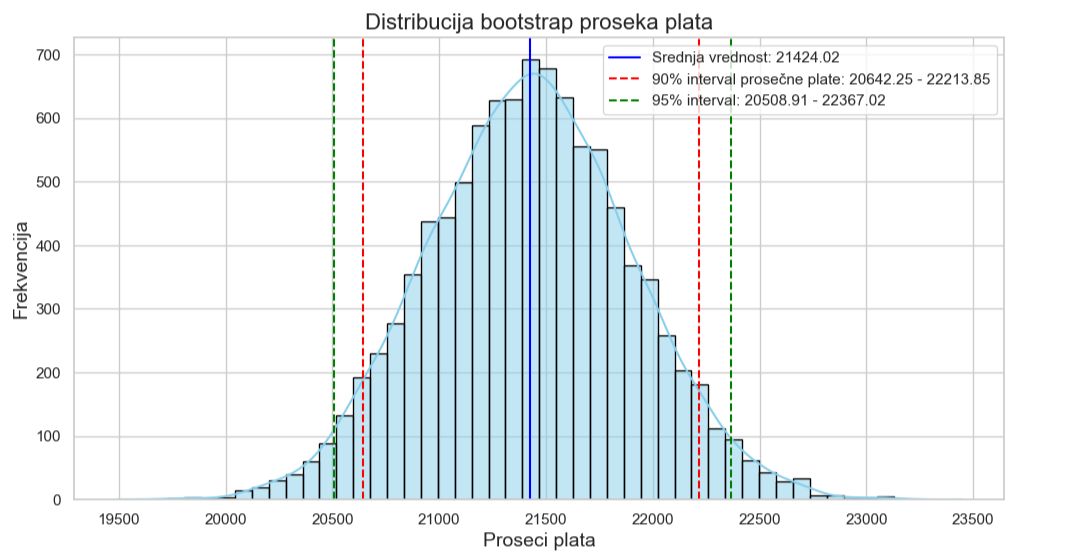
Kako ću u ovom radu ocenjivati sredinu plate, dovoljno je da distribucija sredine plate iz slučajnih uzoraka ima normalnu raspodelu. Prema centralnoj graničnoj teoremi (CGT), u teoriji, ova pretpostavka bi trebala da važi za uzorake veće od 30, jer veliki uzorci obezbeđuju približavanje normalnoj distribuciji 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.

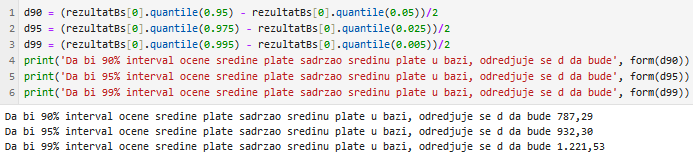


Vizualizacija rezultata

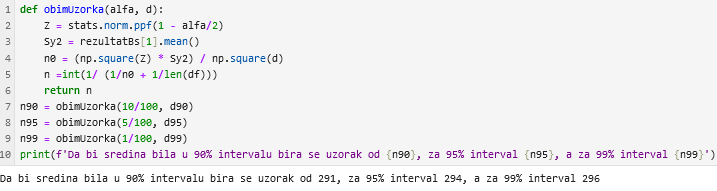


Kako bi 90% odnosno 95% i 99% interval sadržao stvarnu prosečnu platu d se računa direktno iz objekta rezultatBs (iz realizovane raspodele uzimaju se kvantili od interesa)

d90 na histogramu predstavlja razdaljinu od plave do crvene linije, a d95 razdaljinu od plave do zelene linije.



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

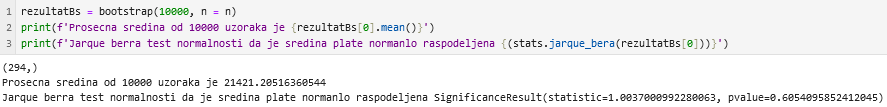


Dobija se da su veličine uzoraka potrebne za postizanje 90%, 95% I 99% intervala poverenja veoma slične. Ova sličnost može se objasniti činjenicom da je raspodela prosečnih plata u uzorcima bliska normalnoj raspodeli, što je i formalno testirano. Teorijski, za “savršeno” normalnu raspodelu za svaki nivo 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.

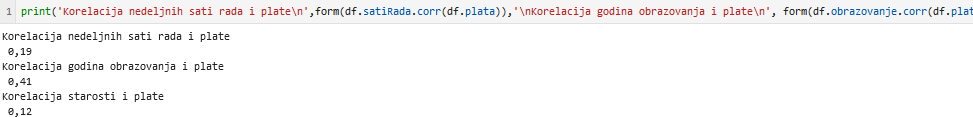
Pošto su rezultati prilično slični i razlike u veličinama uzoraka su male, odlučio sam da uzmem uzorak od 296. Ovaj uzorak daje dovoljno precizne procene sredine plate za 95% interval poverenja, a takođe je dobar i za 90% i 99% intervale, što znači da mogu da budem siguran u rezultate za sve nivoe poverenja.

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.

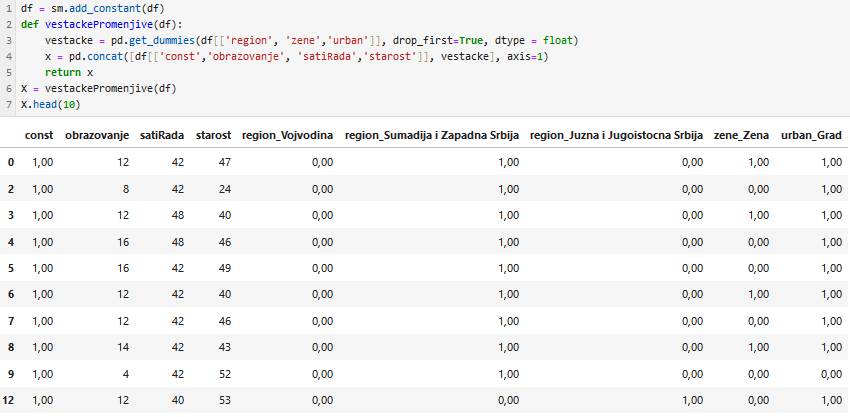


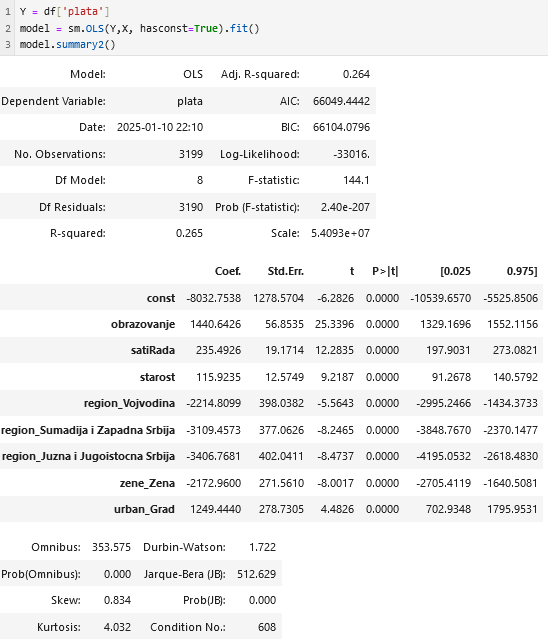
Nakon što sam odredio optimalnu veličinu uzorka, proveravam koja numerička varijabla ima najvišu korelaciju sa ciljanom varijablom. Varijablu satnica ne uzimam u obzir jer je ona linearna kombinacija varijable plata.

Kako je najveća korelacija zavisne varijable sa varijablom obrazovanje, nju ću koristiti za količničko ocenjivanje, dok ću za regresiono ocenjivanje koristiti sve varijable.



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).





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

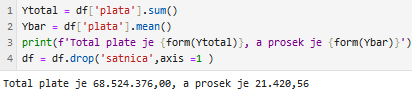
Varijacije u vektoru X objašnjavaju 26,5% varijacija mesečne zarade. F statistika (144) je značajna, jer je veća od kritične vrednosti F(3190,8)=2.93 .

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.

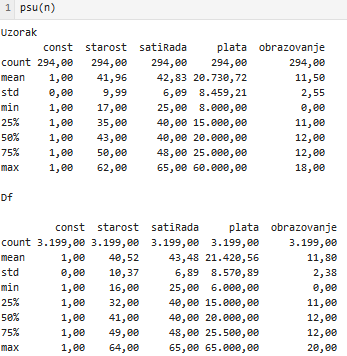


Model gde je zavisna promenjiva ln(Plata) ima manju vrednost Bajesovog (BIC) i Akaikeovog (AIC) informativnog kriterijuma od lin lin modela. Log lin model ima reziduale koji su u većoj meri slični normalnoj raspodeli (JB = 22), iako još uvek nije moguće tvrditi da imaju potpuno normalnu raspodelu. Ove 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.



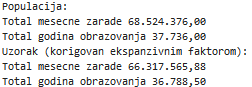
Prost slučajni uzorak:



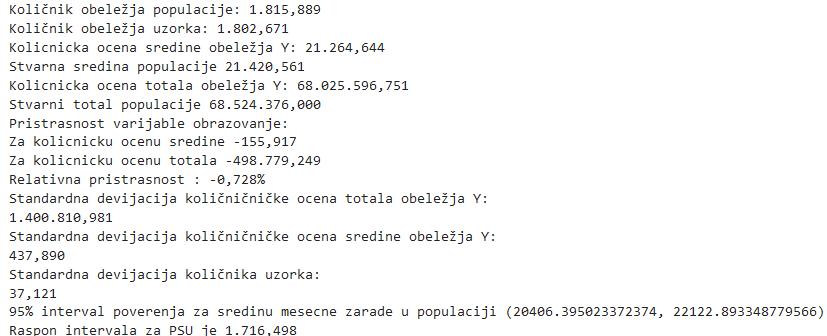
Kod za funckiju psu() će biti priložen u dodatku.

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

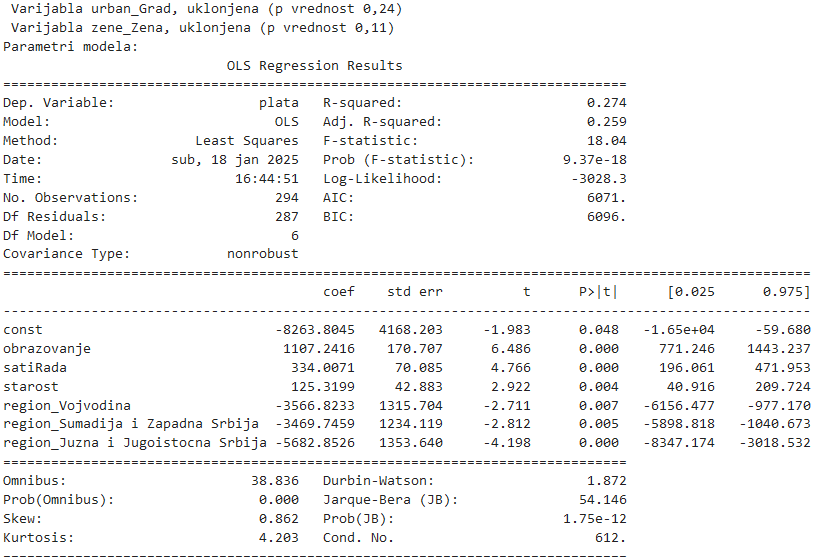


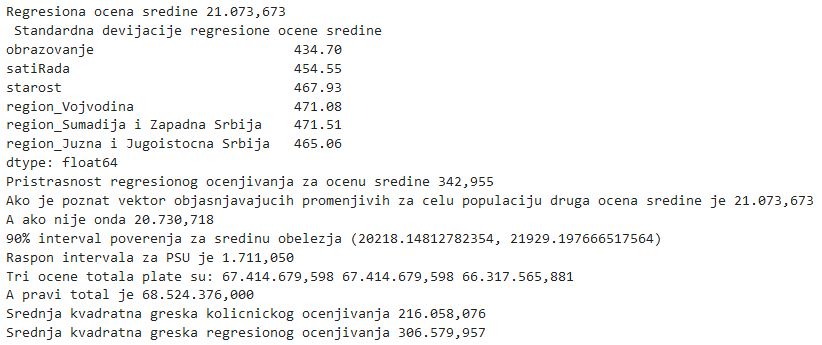
KOLIČNIČKO OCENJIVANJE:



**95% interval poverenja sadrži stvarnu sredinu baze**

REGRESIONO OCENJIVANJE:

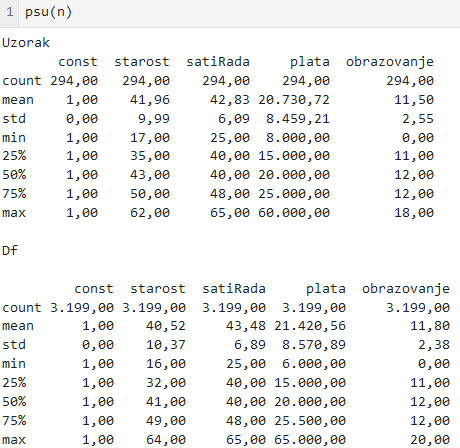
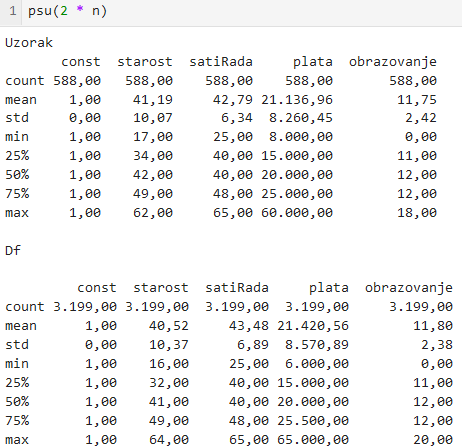




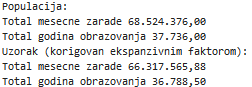
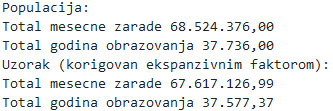
**95% interval poverenja sadrži stvarnu sredinu baze**

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

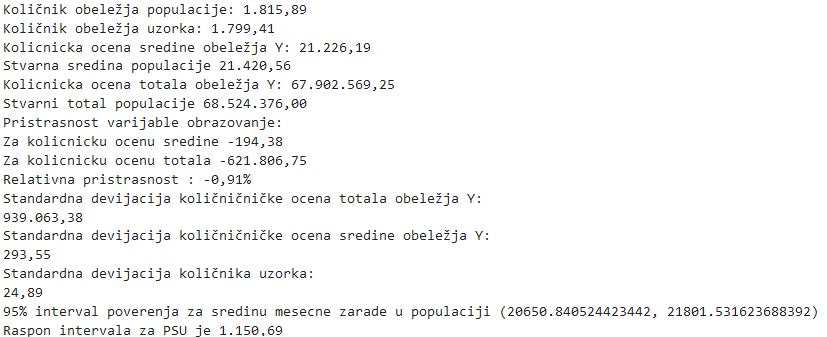
Da bih proverio kako se ocene ponašaju sa povećanjem uzorka izvadio sam uzorak duplo veći od optimalnog. Postavljaću izlaze uporedno, karakteristike optimalnog uzorka će biti tamnije.

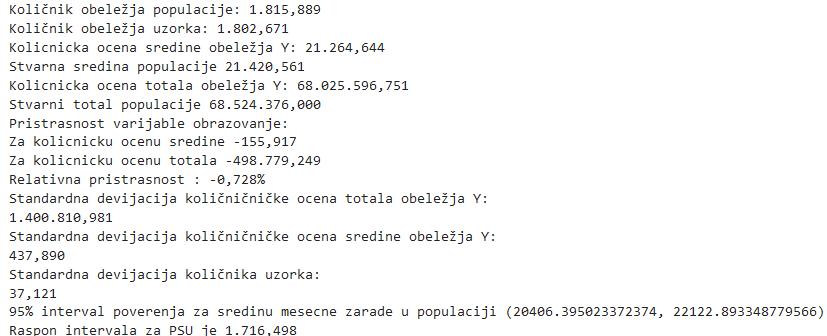


Poređenjem srednjih vrednosti varijabli ova dva uzorka vidi se da veći uzorak ima sličnije karakteristike baze od manjeg uzorka za sve varijable osim za nedeljne sate rada gde je razlika između ova dva uzorka 0,04 radna sata nedeljno.



KOLIČNIČKO OCENJIVANJE:

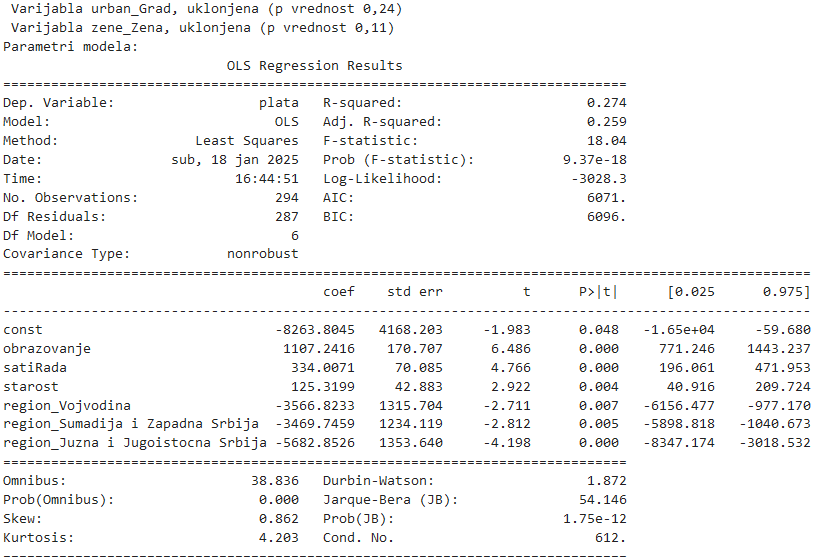
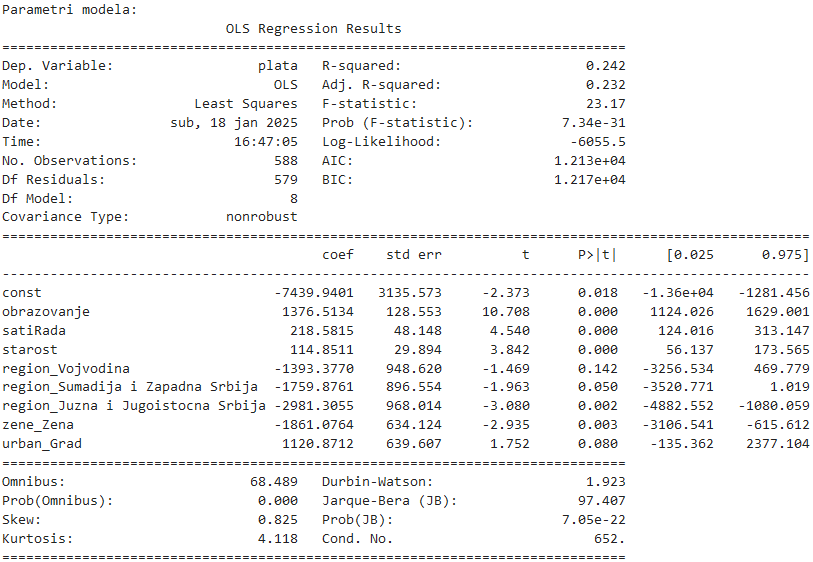




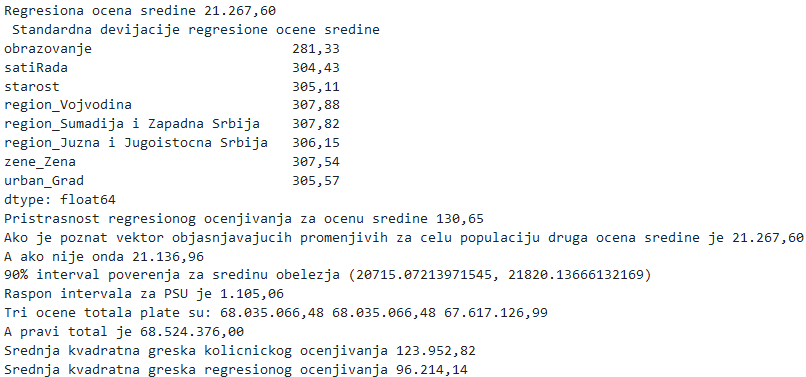
Pristrasnost je veća nego u prvom uzorku, ali je standardna devijacija manja.

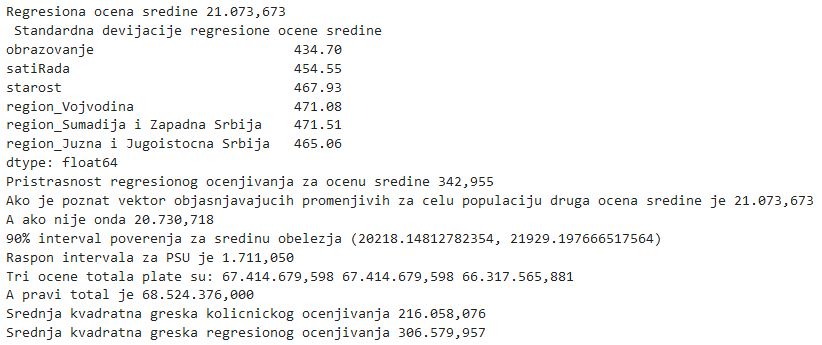
**95% interval i dalje sadrži sredinu baze.**

REGRESIONO OCENJIVANJE:



U većem uzorku sve objašnjavajuće promenjive su statistički značajne.





Povećanjem uzorka smanjila se pristrasnost i standardna devijacija ocene.

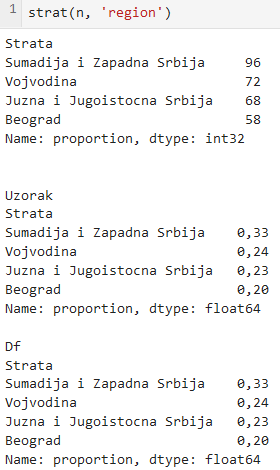
**95% interval sadrži sredinu baze.**

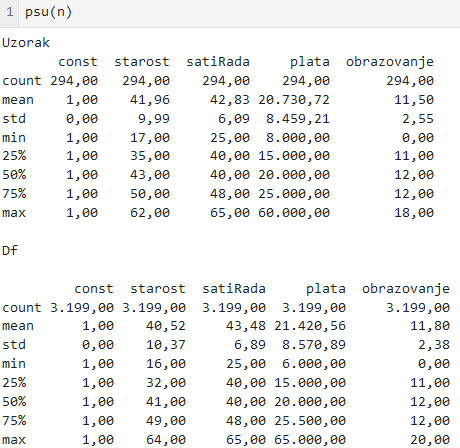
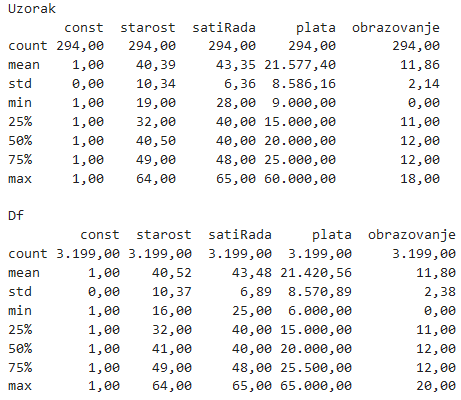
Kako je srednja kvadratna greška manja kod regresionog ocenjivanja za PSU(588), zaključujem da je regresiono ocenjivanje preciznije.

Stratifikovani slučajni uzorak:

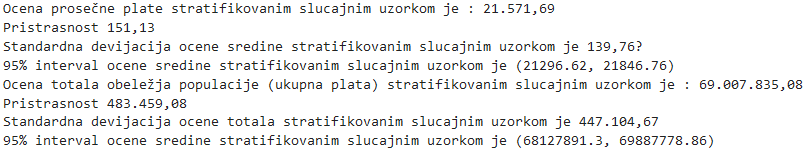
1. Regioni

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.

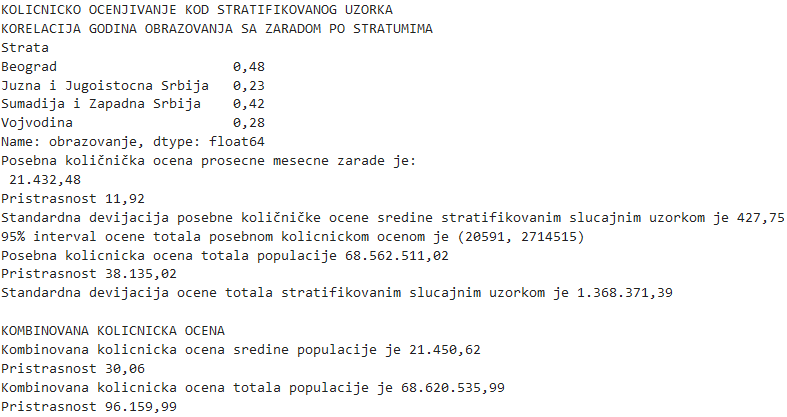
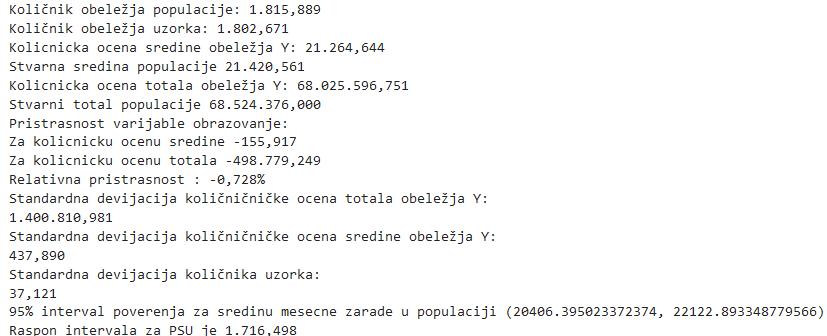




Proseci varijabla u stratifikovanom slučajnom uzorku su skoro identični onima u bazi, izlaz sa desne strane je poređenje sa PSU sa istim obimom.



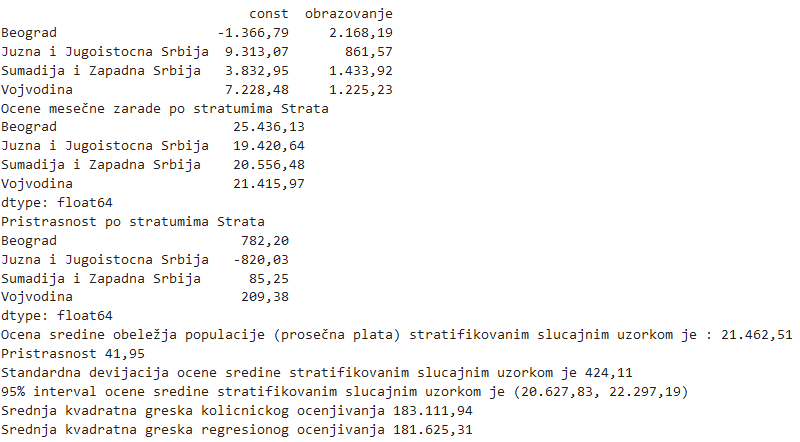
KOLIČNIČKO OCENJIVANJE:

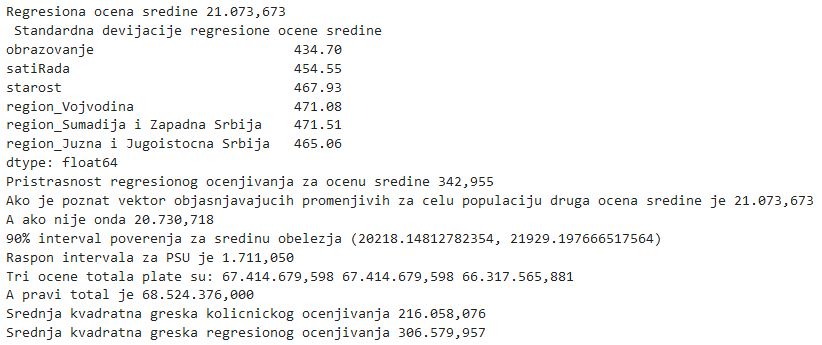
1

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



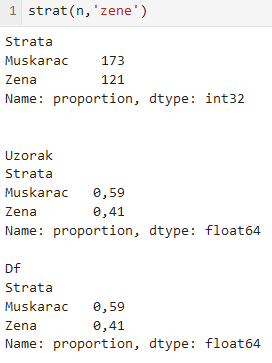


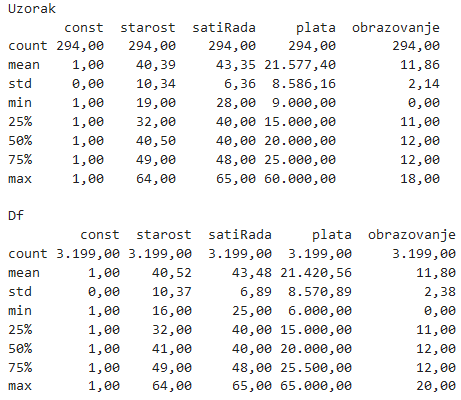
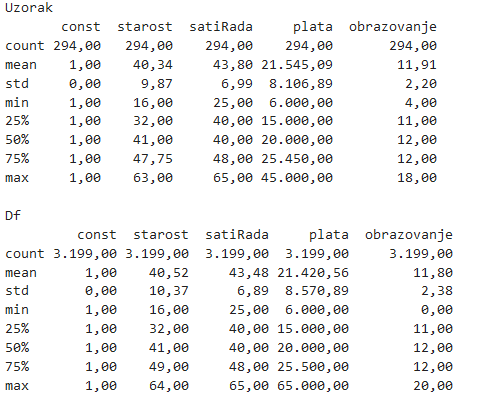
Za svaki region je ocenjena regresija: plata = β0 + β1 \* obrazovanje. Pomoću dobijenog koeficienta β1 ocenio sam prosečnu platu za svaki stratum, plata u Beogradu je najprecenjenija, dok je u Južnoj i jugoistočnoj srbiji najpotcenjenija. Ocene za regione su ponderisane i dobijena je sveukupna ocena 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.

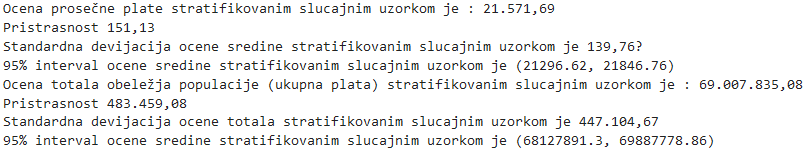
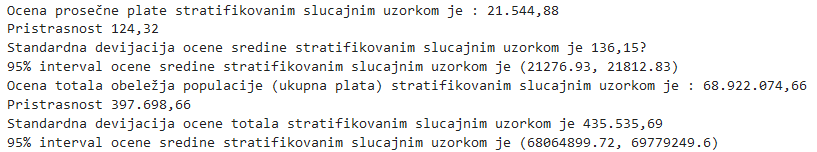
1. Pol

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.

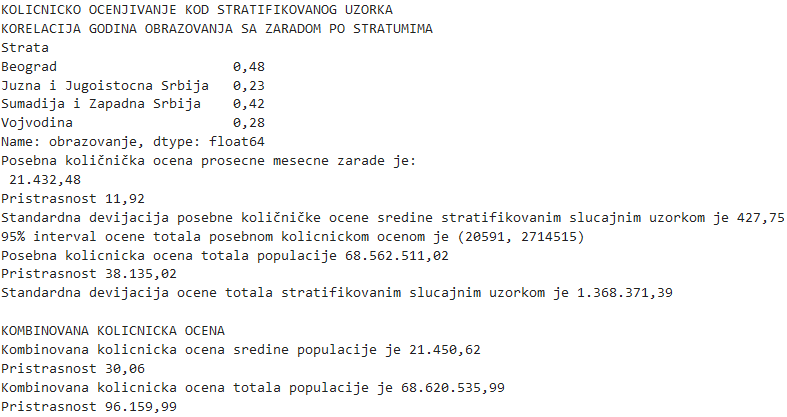
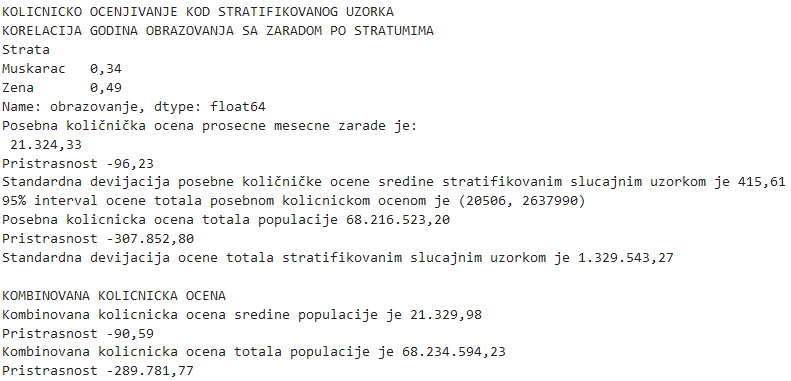




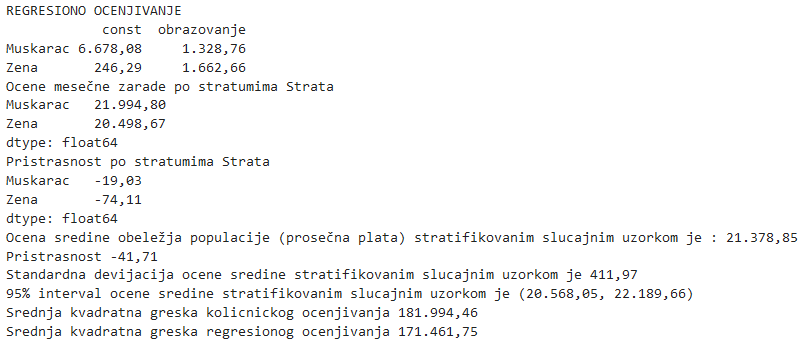
Prosečne vrednosti varijabli su gotovo identične, malu prednost ima regionalna stratifikacija (očekivano jer je podela po regionima detaljnija).

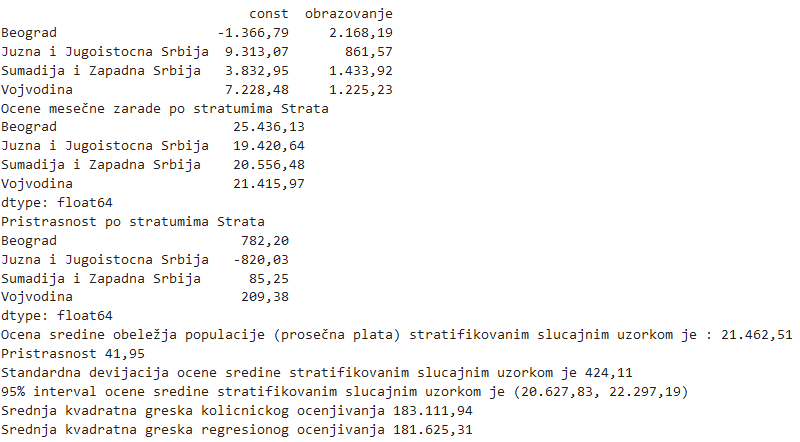


Pristrasnost i standardna devijacija su manje nego u prethodnom uzorku.



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



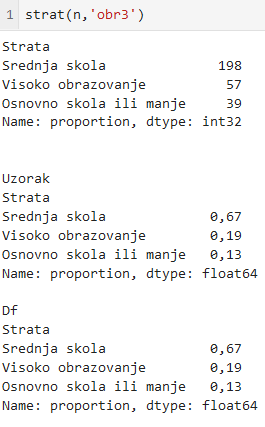


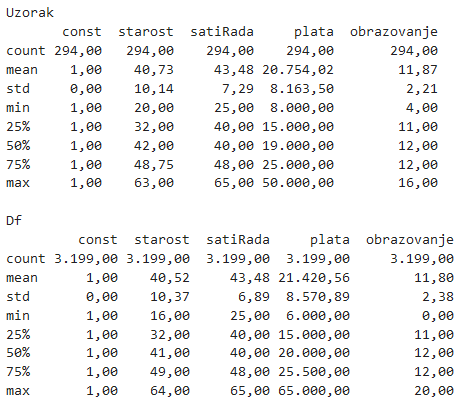
Linearni regresioni model ocenjuje prosečnu zaradu muškaraca sa 22000 dinara, a žena sa 20500 dinara.

U uzorku stratifikovanom prema polu SKG za regresiono ocenjivanje ima najmanju vrednost (čak za 5% manju).

1. Obrazovni nivo

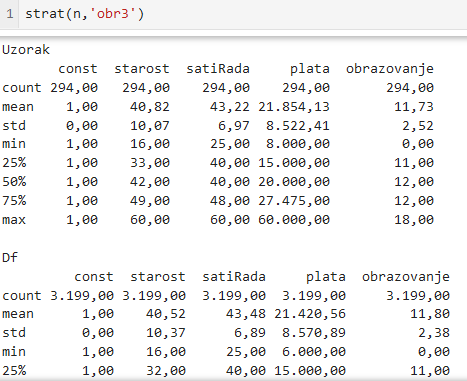
Da bi uzorak bio stratifikovan po obrazovnom nivou, baze je na slučajan način izabrano 39 ispitanika sa osnovnim, 198 sa srednjim i 57 sa visokim obrazovanjem. Na taj način je u uzorku, kao i u bazi, 13%, 67% i 19% ispitanika imaju osnovno, srednje, odnosno visoko obrazovanje.



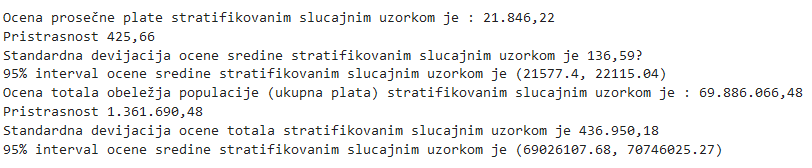


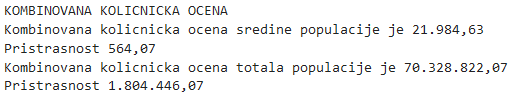
Vektor objašnjavajućih promenljivih ovako stratifikovano uzoraka ima sličnije karakteristike bazi nego uzorak stratifikovan po polu i skoro iste karakteristike kada se uporedi sa regionalno stratifikovanim uzorkom. Dok zavisna varijabla odstupa mnogo više nego kod druga dva stratifikovana uzorka (skoro 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.

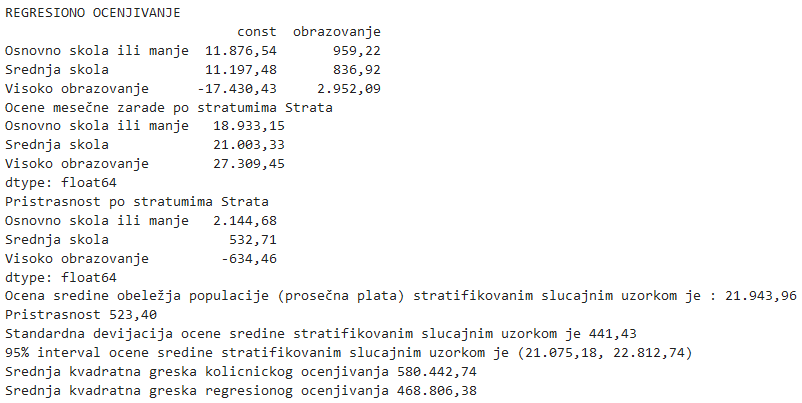
U funkciji strat menjam argument seed kome je podrazumevana vrednost 1304 na 13040 kako bih dobio drugih 294 ispitanika



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.

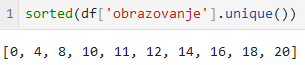






Regresiona ocena je bolja od količničke ocene, ali su srednje kvadratne greške više nego duplo veće nego kod prethodna dva uzorka.

Kod regresionog ocenjivanja postoji konceptualni problem kada se stratifikacija obrazovnim kategorijama.



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.