PREDVIĐANJE CIJENA NEKRETNINA ALGORITMOM SLUČAJNIH ŠUMA

Hodžić, Alen

VISOKO UČILIŠTE ALGEBRA

DIPLOMSKI RAD

PREDVIĐANJE CIJENA NEKRETNINA ALGORITMOM SLUČAJNIH ŠUMA

Alen Hodžić

Zagreb, svibanj 2019.

Predgovor

Zahvaljujući studiju na Visokom učilištu Algebra imao sam tu čast ne samo upoznati ambiciozne kolege nego na predavanjima uvaženih predavača dobiti i poticaj koji me je pogurnuo dalje u mojoj karijeri. Njihova predanost i strpljenje je ono što me tjeralo da dam više od samoga sebe. Stoga se zahvaljujem profesorima Visokog učilišta Algebra posebno mentoru doc. dr. sc Leo Mršiću na pomoći pri izradi rada, te doc. dr. sc. Sandru Skansiju na njegovoj vjeri u mene i davanju korisnih savjeta tijekom studija. Svakako se moram zahvaliti i cjelokupnom osoblju Visokog učilišta Algebra na susretljivosti.

Rad je pisan kroz više etapa, ovisno o mom slobodnom vremenu, a jedan dio učenja i pripreme odradio sam i na Erasmus pripravničkom razdoblju u 2017. g. koje je realizirano kroz ured za međunarodnu suradnju Visokog učilišta Algebra.

Na kraju, ovaj rad posvećujem mojoj kćeri Laurori, mojoj snazi.

Prilikom uvezivanja rada, Umjesto ove stranice ne zaboravite umetnuti original potvrde o prihvaćanju teme diplomskog rada kojeg ste preuzeli u studentskoj referadi

Sažetak

Rad se bavi utjecajem raznolikog broja parametara na utjecaj cijene nekretnina na tržištu s primjerom na velikom tržištu Rusije. Podaci korišteni u radu dobiveni su od najveće banke u Rusiji i Istočnoj Europi, Sberbank, tijekom natjecanja na kaggle.com. Za analizu u ovom radu odabran je algoritam slučajnih šuma koji je, uspoređen sa drugim algoritmima, pokazao prednosti i mane u predviđanju tržišne cijene nekretnina. Osim samog algoritma slučajnih šuma napravljena je i analiza utjecaja stanja ruskog gospodarstva, u kontekstu razvoja tržišta nekretnina, u razdoblju od 2011. do 2016. godine. Samo gospodarstvo je utjecalo na kupovnu moć građana Rusije s obzirom na pad BDP-a tijekom ekonomskih sankcija za vrijeme sukoba u Ukrajini.

Ključne riječi: algoritam slučajnih šuma, predviđanje cijena nekretnina, detaljna analiza podataka, rusko tržište nekretnina.

Summary

Thesis main interest was to analyse influence between different parameters on real estate market prices prediction using data source provided by biggest Russian and East European bank, Sberbank, for competition on kaggle.com. Random Forest was chosen as an algorithm for prediction, and was compared with similar algorithms in order to compare and evaluate various approaches and results. Besides the algorithm itself, research include analysis of Russian economy form real estate market perspective, from 2011. to 2016. During that period Russia was under economic sanctions for it's Ukraine conflict which affected national GDP and consumer power, consequently influencing real estate market as well.

Key words: random forest, real estate prices prediction, exploratory data analysis, data science

Sadržaj

1. Uvod

Mnogi investitori u nekretnine ne prepoznaju važnost analize tržišta. Bilo da im nedostaju vještine i znanje da bi dovršili analizu tržišta ili jednostavno ne razumiju prednosti, analiza tržišta je podcijenjena imovina u ulaganjima u nekretnine. U stvarnosti, analiza tržišta najvažniji je element u procjeni ulaganja u nekretnine. Analiza tržišta čini osnovu svakog izračuna i odluke koja slijedi. Dakle, temeljito istraživanje i razumijevanje tržišta ključno je za dobro donošenje odluka.

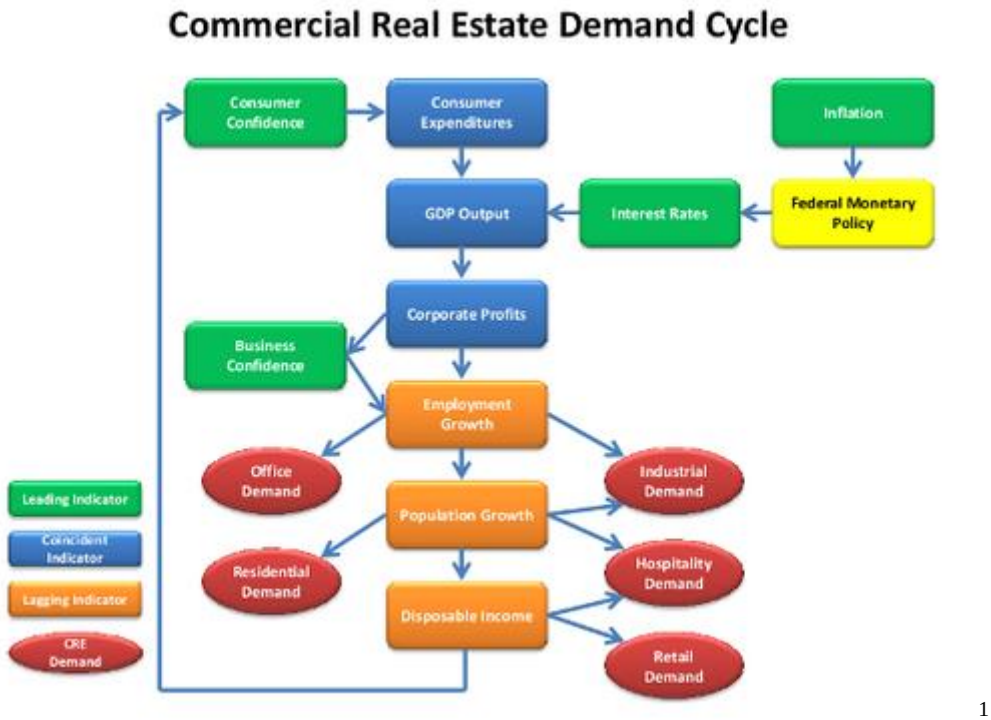
Analiza tržišta nekretnina sadrži nekoliko osnovnih dijelova, a svaki pruža ključne informacije potrebne za analizu vrednovanja i financijsku izvedivost bilo kojeg ulaganja u nekretnine. Prvi dio definira područje koje se razmatra. Definiranje područja više je nego samo pronalaženje granica partije, ali uključuje definiranje veličine ciljnog tržišta koje će najvjerojatnije generirati prihod. Precizno definiranje ciljnog tržišta i susjedstva omogućuje investitoru prepoznati konkurenciju i trenutnu ponudu kako bi zadovoljio trenutnu potražnju u tom području.

Drugi dio sadrži temeljitu analizu fizičkih i okolišnih čimbenika koji utječu na nekretnine. Fizički čimbenici uključuju stvari kao što su lokacija, prirodni resursi, topografija, uvjeti tla, klima, dostupnost vode i obrasci prijevoza. Na prvi pogled, neki od tih čimbenika možda se ne čine strašno važnima za analizu profitabilnosti ulaganja u nekretnine. U nekim slučajevima, međutim, lokalna zajednica pored oceana ili ugodne klime mogla bi biti sastavni dio gospodarstva, industrije i poželjnosti zajednice. Bilo bi nemoguće u potpunosti razumjeti dinamiku zajednice bez uvažavanja tih fizičkih karakteristika.

Osim ovih fizičkih čimbenika, analiza tržišta također može uključivati više informacija o značajkama susjedstva. To često uključuje detaljne informacije o pristupu susjedstva javnim dobrima i uslugama. Pristup i kvaliteta komunalnih usluga mogu biti izuzetno važni za razvoj komercijalnih nekretnina. Investitori u nekretnine moraju razmotriti ima li određena parcela dovoljan pristup komunalnim uslugama, kao i jesu li ta komunalna poduzeća sposobna ispuniti dodatne zahtjeve za servisiranjem novog razvoja. Ako ne, investitor mora uvjeriti lokalnu samoupravu da ulaže u poboljšanje komunalne usluge na tom području. Stoga se ne smije previdjeti dostupnost odgovarajućih komunalnih usluga i troškova izgradnje jer oni u konačnici mogu odrediti izvedivost bilo kojeg projekta nekretnina.

Nakon adresiranja fizičkih faktora lokacije, analiza tržišta procjenjuje ekonomske karakteristike i trendove na tom području. Svrha ove ekonomske analize je pružiti razumijevanje temeljne populacije, uvjeta poslovanja i buduće potražnje za određenom vrstom nekretnina. Trendovi u demografskim podacima pružaju uvid u buduće gospodarsko zdravlje regije. Na primjer, rastuća populacija općenito je dobar znak gospodarskog prosperiteta u regiji sve dok postoji sve veća mogućnost zapošljavanja za stanovnike. Raspodjela dobi, obrazovanje i dohodak također su važni pokazatelji regionalnih obrazaca rasta.

Šire ekonomske trendove u regiji, kao i na nacionalnoj razini, trebalo bi također razmotriti u analizi tržišta. Iako su sve nekretnine lokalne, veće, makroekonomske snage imaju odjeke na svim lokalnim tržištima. Kao rezultat toga, kamatne stope, sadašnje i predložene promjene poreznih politika, inflacije, rast BDP-a i stope nezaposlenosti treba analizirati u analizi tržišta. Svi ovi čimbenici igraju važnu ulogu u rastu ili smanjenju ekonomske baze koja okružuje predmetnu imovinu. CCIM institut ima lijep dijagram toka ciklusa potražnje koji sve to čini zajedno:



[1]

Grafikon 1-1 Ciklus potražnje za komercijalnim nekretninama

[1] https://www.propertymetrics.com/blog/2017/10/28/why-is-real-estate-market-analysis-so-important/

Istraživanje druge nove gradnje na tom području također bi trebalo biti dio analize tržišta. Potraga za građevinskim dozvolama može biti izvrstan pokazatelj prošlog razvoja, kao i nove ponude koja će biti na tržištu u budućnosti. Nova gradnja signal je da se četvrt smatra poželjnim, ali može biti i izvor konkurencije za stanare ili kupce. Ostala pitanja vezana za izgradnju su zoniranje i razvojni zahtjevi za novu gradnju. Analiza tržišta trebala bi istražiti prostorne i građevinske propise, kao i vremenski okvir, troškove i stav lokalnog odbora za planiranje. Samo ova tri čimbenika mogu odrediti je li ulaganje u nekretnine financijski izvedivo ili ne.

Sve u svemu, analiza tržišta trebala bi pružiti sveobuhvatnu sliku o predmetnoj imovini, lokaciji, susjedstvu i većim tržišnim ekonomskim pokretačima. Završni dokument trebao bi omogućiti čitatelju da razumije trenutne uvjete ponude i potražnje za ovom vrstom nekretnina, kao i sliku o tome kako se ti uvjeti mogu promijeniti u budućnosti. Također bi trebalo donijeti zaključke o promjenama u demografiji i propisima u četvrti te kako bi ti čimbenici, kao i ekonomija, mogli utjecati na predmetnu imovinu.

Za temu predviđanja cijena odlučio sam se osobnom željom i htjenjem ući dublje u tematiku predviđanja cijene nekretnina kako bi shvatio ne samo osnovne principe, nego i provjerio opće prihvatljive činjenice da su položaj nekretnine i sama površina najutjecajniji faktori pri određivanju cijene nekretnine. Vođen i osobnim iskustvom rada na porezu nekretnine gdje sam primijetio da standardne metode procjene nekretnine (prihodovne, poredbene, troškovne i hedoničke) imaju podosta nedostataka. Ti nedostaci se najviše očituju u krutosti same procjene, koja proizlazi iz:

- Nedovoljnog broja varijabli
- Nekorištenja modernih metoda strojnog učenja
- Nemogućnosti dinamičke prilagodbe utjecaja mikro i makroekonomskih elemenata

Kako sam u svom radu na porezu na nekretnine najviše koristio poredbenu metodu cijena nekretnina nije mi bilo jasno kako se mogu uspoređivati cijene nekretnina koje su kupljene prije par godina s cijenama nekretnina kupljenim ove godine. Iako tržište na kraju samo određuje cijenu nekretnina (u slučaju Hrvatske posrednici i prodavatelji nekretnina osobno), mislim da su cijene nerealne s obzirom na stanje gospodarstva. Pretpostavka je da su cijene

nekretnina nerealne zbog ulaska Hrvatske u Europsku uniju i otvaranja tržišta, intervencija države putem subvencija, iseljavanje stanovništva, u prvom koraku u Zagreb iz ostatka Hrvatske, a u drugom koraku prema inozemstvu i drugi nevidljivi čimbenici. Naravno za tako sveobuhvatnu analizu nemam dovoljno podataka zbog:

- Nepostojanja baze podataka na godišnjoj razini, kao i na dugoročnoj razini
- Nemogućnosti korištenja podataka kao posljedica Direktive o zaštiti osobnih podataka (GDPR)
- Nemogućnosti korištenja podataka zbog prirode posla kojim se bavim (Porezna uprava)
- Nezainteresiranosti cjelokupnog sustava i tržišta za uređenjem tržišta nekretnina
- Utjecaja mikro i makroekonomskih čimbenika zbog nedostatka vremena da se tome studiozno posvetim
- Kaskanja Hrvatske za ostatkom svijeta, naročito u pogledu primjene tehnologije
- Nesređenog katastra i uređenja vlasničkih prava
- Nepostojanja radova koji se bave problemom tržišta nekretnina u Hrvatskoj

U ovom trenutku za takvu analizu nema dovoljno elemenata da bih išao u dubinu kakvu bih želio. To nije projekt za jednog čovjeka.

Kada već pričam o trenutnom stanju u Hrvatsku bilo bi dobro pojasniti i utjecaj vanjskih čimbenika, najviše uzorkovanih globalnim kompanijama koje su ušle na hrvatsko tržište. Jedna od takvih je i Airbnb koja je neposredno? i digla cijene nekretnina najmom nekretnina u privatnom vlasništvu. Samim ulaskom na tržište promijenila je i samu strukturu tržišta gdje se odjedanput osjeća nedostatak nekretnina i za najam i na kupovinu. Time se pokazuje nespremnost hrvatske administracije na reakciju na konstantne promjene koje utječu na tržište nekretnina. Pod time mislim na:

- Poreznu politiku
- Radnu i demografsku politiku
- Gledanje kratkoročnih ciljeva umjesto dugoročnih
- Turističku i gospodarsku politiku

S jedne stane, uslijed otvaranja tržišta rasle su i cijene nekretnina i prihod od najma, konkurencija tržišta rada i plaće, a s druge strane dugoročno smo negativno utjecali na demografsko stanje države i ugrozili budućnosti države iseljavanjem stanovništva.

Mogli bismo pogledati slučaj Pariza, jednog od najvećih tržišta nekretnina u svijetu, gdje su pokušali intervenirati na cijene nekretnina donošenjem regulative koja bi spriječila nezakonito oglašavanje Airbnba. Uveden je zakon koji je limitirao broj dana najma na 120 dana. Pokušalo se spriječiti pretvaranje grada u prazne, turističke zone. Nešto slično događa se i u Dubrovniku, gdje domicilno stanovništvo iseljava iz jezgre da bi se napravilo mjesta

kapitalu. Zapravo glavni je problem Hrvatske nepostojanje dugoročne politike razvoja, pa ima prevelikih razlika među pojedinim regijama.

U ovom radu sam se okrenuo za mene jednostavnijem rješenju. Došao sam do gotovih podataka na kojima bih mogao testirati algoritam slučajnih šuma u odnosu na druge algoritme. Do podataka sam došao putem internetske stranice kaggle.com koja ima redovita natjecanja u strojnom učenju i primjeni algoritama na velike količine podataka. Podaci se sastoje od mikro i makroekonomskih varijabli te osobina nekretnina (površina, blizina škola, vrtića i drugih varijabli društvenog sadržaja) na ruskom tržištu. Ove podatke je Sberbank učinio dostupnima u skladu s pravilima natjecanja i bit će osnova mojih nastojanja u određivanju cijene nekretnina.

Osim dostupnih podataka provedena je i analiza utjecaja svjetskih kretanja na rusko tržište. Najvažniji utjecaji proizašli su iz sukoba u Ukrajini i nametnutih ekonomskih sankcija, koje su posredno utjecale i na rast cijena na ruskom tržištu nekretnina.

S te strane ne treba smetnuti s uma i ovisnost Europske unije o energentima koji dolaze iz Rusije. Europska unija ovisi o ruskom plinu i procjenjuje se da će ta ovisnost rasti. Zato uz neke realne pokazatelje treba paziti i na one političke, gdje Europska unija pokušava smanjiti ruski utjecaj, i trudi se ostati gospodarska i politička sila na globalnoj razini.

Nadalje, htio sam isprobati algoritam slučajnih šuma i provjeriti njegovu preciznost u predviđanju cijene nekretnina. Algoritmi koji će mi biti usporedba s algoritmom slučajnih šuma su višestruka regresijska analiza, algoritam pojačavanja gradijenta modela stabala i pojačavanje modela stabala XGBoost modelom. Očekujem da bi algoritam slučajnih šuma trebao slabije predviđati cijenu nekretnina. No važno je napomenuti i da se korištenjem jednostavnijeg algoritma cilja na bržu mogućnost predviđanja cijena nekretnina kao početnu točku predviđanja.

2. Metode procjene vrijednosti nekretnina

Da bi uopće mogao započeti procjenu cijena nekretnina prvo ću započeti sa danas najčešće korištenim metodama u procjeni cijena nekretnina. Nažalost u većini slučajeva analiza tržišta uopće nije prepoznata kao ključan element ulaganja u tržište nekretnina. Bilo to zbog nedostatka znanja ili nerazumijevanja važnosti analize tržišta, to je svakako korak koji se ne smije preskočiti. Analiza tržišta je temelj svake buduće odluke. U prvom koraku analize bi trebalo definirati veličinu tržišta na kojem želimo poslovati. Drugi korak analize se odnosi na okolišne uvjete u kojima se nekretnina nalazi. Pod okolišni uvjeti misli se na lokaciju, klimu, prometnu povezanost, dostupnost vode i kvaliteti zemljišta, pa onda i na dostupnost javno privatnih usluga (vrtić, škola, ugostiteljski objekti, sportske dvorane i sl.). Treći korak se odnosi na ekonomske karakteristike. Kakva je demografska struktura, da li populacija stari, kakav ću utjecat imati na buduće stanje tržišta, to su pitanja koja se moraju uzeti u obzir. Naravno osim lokalnih ekonomskih faktora moramo pogledati i širu sliku, utjecaj stanja u kojem je državna ekonomija, utjecaj regionalne i svjetske ekonomije, porezne politike, zaposlenosti, migracija stanovništva i ostalo.

Kako sam tijekom proteklih mjeseci i sam sudjelovao u radu u odjelu za nekretnine Porezne uprave mogu iz osobnog iskustva posvjedočiti o problemima koji se javljaju pri procjeni cijena nekretnina. Jedan od najvećih problema je to što ne postoji relevantna baza podataka koja bi olakšavala procjenu nekretnina u Hrvatskoj. Porezna uprava koristi isključivo poredbenu metodu procjene nekretnina, što u popriličnom broju procjena cijena nekretnina ne daje niti relevantne niti točne podatke na temelju kojih se vrši procjena cijene nekretnina.

Osobno smatram da je za procjenu cijene nekretnina potrebno više od podataka dobivenih putem poredbenih nekretnina temeljenih na prijašnjim procjenama Porezne uprave i podataka Geoportala Državne geodetske uprave.Potrebni su podaci koji daju širu sliku o kvaliteti nekretnine koja se procjenjuje. Pri tom mislim na podatke koji u potpunosti opisuju nekretninu, njen položaj, blizinu sadržaja koji oplemenjuje nekretninu, makroekonomska i mikroekonomska obilježja tržišta nekretnina i gospodarstva države u kojoj se procjena nekretnina vrši, etc. No krenimo redom, prvo ćemo nabrojati metode koje se koriste temeljem hrvatskih zakona i propisa.

2.1. Prihodovna metoda

Prihodovna metoda koristi se za procjenu vrijednosti onih nekretnina koje generiraju prihode. Pri tome je usredotočena na čiste prihode koji proizlaze iz korištenja nekretnine, a koji se definiraju kao razlika između tokova ukupnih prihoda i troškova gospodarenja. Logika koja stoji iza primjene prihodovne metode je ta da niti jedan potencijalni investitor neće za nekretninu platiti veći iznos od onog kojeg će kroz vrijeme moći vratiti kroz njen korištenje (Majčica, 2014). Prihodovna vrijednost nekretnine daje upravo tu informaciju, obzirom da predstavlja sadašnju vrijednost predviđenih novčanih tokova koji proizlaze iz korištenja nekretnine, diskontiranih na dan vrednovanja.

Ova se metoda najčešće koristi kako bi se odredila vrijednost poslovnih nekretnina, koje su Pravilnikom o metodama procjene vrijednosti nekretnina (NN 105/2015) definirane kao nekretnine koje se to su one nekretnine koje se prema ukupnom godišnjem prihodu koriste preko 80% za zakup pravnim osobama, obrtnicima i drugim poslovnim oblicima fizičkih osoba. Pravilnik također ističe kako očekivani prihodi, koji ulaze u izračun prihodovne vrijednosti, predstavljaju održive prihode koji se postižu na tržištu te kako se buduće korištenje nekretnine, koje će generirati prihode, ne smije temeljiti na špekulativnim očekivanjima nego isključivo na stvarnim okolnostima, prostornim planovima i slično. osim za poslovne prostore , Pravilnik propisuje i korištenje prihodovne metode u svrhu procjene vrijednosti određenih nekretnina koje imaju javnu namjenu, a ukoliko javna vlast razmatra zakup (za dječje domove, dječje vrtiće, domove za rehabilitaciju djece: poredbene privatne ustanove i slično) no tu je potrebno istaknuti kako se očekivanja o prihodima formiraju na temelju onih koji mogu proizaći iz poredbenih nekretnina (Slišković, 2016).

Tehnički se prihodovna vrijednost nekretnine može utvrditi primjenom dvaju metoda:

- metoda direktne kapitalizacije
- metoda sadašnje vrijednosti

Direktna kapitalizacija je znatno jednostavnija metoda, koja se masovno koristi za procjenu u praksi kao metoda za okvirnu procjenu vrijednosti objekata u kojima vlasnik prostora i korisnik nisu iste osobe (Majčica, 2014). Procjena se donosi na temelju posjedovanja podataka o očekivanim rentama *R*, odnosno godišnjem prihodu koji generira određena nekretnina i kamatne stope *i* koja predstavlja očekivani prinos. Prihodovna vrijednost je definirana kao omjer dvije navedene varijable, odnosno :

$$P=R/i \qquad (1)$$

Ovaj jednostavan odnos može se primijeniti i za procjenu očekivanog prinosa, odnosno zahtjevan stope kapitalizacije u situaciji kada su poznate informacije o prihodima cijeni. Stopa kapitalizacije je u tome slučaju jednaka omjeru godišnjeg prihoda i cijene nekretnine:

$$i=R/P \qquad (2)$$

Ovaj način je često korišten za procjenu stopa kapitalizacije unutar poredbene metode (Tica, 2014).

Nešto složenija metoda, koja obuhvaća veći broj koraka u odnosu na direktnu kapitalizaciju jest metoda sadašnje vrijednosti. Njome se predviđeni novčani tokovi pomoću kamatne stope diskontiraju na dan vrednovanja, odnosno svode na sadašnju vrijednost. Diskontiranje očekivanih prihoda ili drugih novčanih tokova je temeljni alat kojim se može procijenit vrijednost svega što u budućnosti donosi dohodak (Tica, 2014a).

Očigledno je kako se u primjeni ove metode treba voditi računa o tri faktora koji mogu utjecati na visinu procijenjene cijene. Jedan od njih je dužina razdoblja u kojem se očekuje da će nekretnina ostvarivati prihode. Drugi način je na koji se definira očekivani prihod. Za iznajmljene nekretnine, to je čisti prihod koji je razlika ukupnog prihoda i troškova gospodarenja (Majčica, 2014). Precizniji izračun podrazumijeva prvenstvenu procjenu bruto prihoda, iz kojeg se izračunava efektivni bruto prihod kada se uzme u obzir prosječna upražnjenost i troškovi rente. Tek potom se oduzimaju troškovi gospodarenja, a rezultat je neto operativni, odnosno čisti prihod (Tica, 2014). Konačno složen zadatak je i procjena stope kapitalizacije. U praksi postoji nekoliko načina, a jedan od njih je definicija dana jednadžbom 2. U nekim se zemljama se, pak, stope kapitalizacije ne procjenjuju, već se koriste određene standardizirane stope (Slišković, 2016).

U primjeni prihodovne metode posebice valja voditi računa o pretpostavkama koje su predviđene Pravilnikom. Naime, opće i pojednostavljena prihodovna metoda koje su njime propisane se temelje na pretpostavci o konstantnim prihodima i konstantnoj diskontnoj stopi. Ukoliko se očekuje promjena bilo koje od dvije navedene varijable, metoda više neće biti primjenjiva za procjenu tržišne, već investicijske vrijednosti, a u tim okolnostima se koristi metoda diskontiranog novčanog toka (DCF, engl. *Discounted Cash Flow*)u kojem se stope povrata korigiraju za njihov očekivani rast, ili se umjesto očekivanih čistih prihoda koriste alternativne mjere novčanih tokova. Ukoliko se pretpostavlja da oni nisu konstantni, koristiti će se njihovi ponderirani prosjeci (Tica, 2014). Obzirom da europski standardi vrednovanja

jasno razlikuju metode kojima se izračunava tržišna vrijednost od onih čiji je rezultat procjena investicijske vrijednosti, treba postojati oprez pri primjeni prihodovne metoda i njoj sličnih, investicijskih metoda procjene vrijednosti nekretnina (Majčica, 2014).

2.2. Poredbena metoda

Poredbena metoda počiva na ekonomskoj logici prema kojoj niti jedan investitor ili potencijalni kupac neće za nekretninu platiti više nego što su drugi subjekti platiti za usporedive nekretnine na tržištu. Ova metoda primjenu nalazi u procjenama vrijednosti izgrađenih i neizgrađenih zemljišta, poslovnih prostora, ali je posebno korištena na segmentu rezidencijalnog tržišta za procjenu vrijednosti stambenih jedinica. Obzorom da je za njezinu provedbu potrebna opsežna baza podataka u kojoj je sadržan velik broj transakcija, što omogućava pronalazak sličnih nekretnina koje se mogu međusobno usporediti, metoda je posebno pogodna za zemlje u kojima je tržište nekretnina dobro organizirano (Tica, 2014).

Dakle, kvalitetna baza podataka o prodanim nekretninama, koja sadrži informacije o kupoprodajnim cijenama, vremenu prodaje i što većem broju karakteristika je nužan uvjet za provedbu poredbene metode. Pravilnikom o metodama procjene vrijednosti nekretnina (NN 105/2015) je definirano izračunavanje poredbenih cijena temelju kupoprodajne cijene onih nekretnina koje sa procjenjivanom nekretninom pokazuju dovoljno podudarna obilježja. Drugim riječima, poredbena metoda omogućuje procjenu tržišne vrijednosti nekretnine na temelju kupoprodajnih cijena visoko usporedivih i nedavno prodanih nekretnina, uz pretpostavku da se opći vrijednosni odnosi na tržištu nisu promijenili (Majčica, 2014).

Za procjenu tržišne vrijednosti poredbenom metodom procjenitelj treba posjedovati informacije o nekoliko kupoprodajnih cijena poredbenih nekretnina. Ne postoji propisani broj poredbenih nekretnina na temelju kojega je moguće donijeti procjenu, no u praksi se smatra dovoljnim posjedovanje informacija o tri visoko usporedive nekretnine, dok se sedam do osam njih smatra optimalnim. Ipak, bez obzira na stupanj usporedivosti nekretnina koje služe kao temelj za procjenu poredbene vrijednosti, valja ipak imati na umu kako na tržištu gotovo ne postoje dvije identične nekretnine. Zbog karakteristike heterogenosti će se poredbene cijene najčešće morati korigirati za kvalitativna obilježja nekretnine (Tica 2014), odnosno nit će potrebno provesti interkvalitativno izjednačavanje.

Korekcija kupoprodajnih cijena koje postoje u bazi podataka se, s jedne strane provodi zbog razlika u karakteristikama usporedivih nekretnina i one koje se procjenjuje, a s druge strane zbog situacije u kojoj dolazi do promjene općih vrijednosti odnosa na tržištu (Majčica, 2014). Korekcijski faktori se najpreciznije mogu izračunati statističkim putem, no za njihovu procjenu potrebno je posjedovati bazu u kojoj se nalaze podaci o velikom broju stanova i njihovih karakteristika. U tom slučaju je cijenu nekretnina moguće modelirati kao funkciju njihovih karakteristika te se primjenom metode najmanjih kvadrata dolazi do cijene svake pojedine karakteristike, odnosno implicitnih cijena. Navedeni pristup naziva se hedoničkim pristupom određivanja cijena, a njegov naziv proizlazi iz pretpostavke da svaka pojedina karakteristika nekretnine donosi određeno zadovoljstvo korisniku i doprinosi njenoj ukupnoj cijeni (Slišković, 2016).

2.3. Troškovna metoda

Troškovna metoda počiva na logici prema kojoj nitko neće za nekretninu platiti više nego što iznosi trošak zemljišta i njezine izgradnje. Prema tome, potrebno je posebno procijeniti vrijednost zemljišta na kojem je nekretnina locirana i vrijednost izgradnje. Problem ovog pristupa leži u činjenici da troškovi i cijene obično divergiraju u veoma kratkom roku, jer sam proces izgradnje traje određeno vrijeme, a ni zemljište ponekad nije odmah raspoloživo za izgradnju (Majčica, 2014). Iz tog razloga je osobito prikladna za procjenu vrijednosti onih nekretnina koje nisu često predmet kupoprodaje i koje ne stvaraju prihode (škole, crkve i slično) (Tica, 2014). Ipak, Pravilnikom o metodama procjene vrijednosti nekretnina (NN 105/2015) je predviđena i uporaba troškovne metode kao potpore prihodovnoj metodi u određenim situacijama. Primjerice, moguće ju je koristiti za procjenu vrijednosti novih građevina kojima je svrha stvaranje prihoda, kako bi se utvrdila pokrivenost troškova gradnje budućim prihodima od najma ili zakupa nekretnine. Također se može primijeniti i kod starijih građevina koje zahtijevaju intenzivno održavanje i rekonstrukciju za procjenu vrijednosti naknadnog ulaganja u njih. Slijedom navedenog je moguće donijeti zaključak da je ova vrsta metode procjene vrijednosti posebno zanimljiva građevinskim poduzetnicima, a potpuno nebitna financijskim investitorima (Tica, 2014).

2.4. Hedonička metoda

Pristup modeliranju cijena koji se temelji na obuhvatu i ispitivanju statističke značajnosti što većeg broja internih karakteristika nekretnina, njihovih lokacijskih obilježja te ostalih karakterisktika tržišta u oblikovanju cijena se naziva hedoničkim (engl. *hedonic*) pristupom određivanja cijena nekretnina.

Prednosti hedoničkog pristupa su brojne. Osim što je proces modeliranja relativno jednostavan, ovim pristupom se ujedno obuhvaćaju i razlike u kvalitativnim karakteristikama, ali i promjene općih vrijednosnih odnosa kroz vrijeme (Majčica, 2014). Osnovna prednost hedoničkog modeliranja se zapravo sastoji u tome što se na temelju procijenjenih implicitnih cijena karakteristika može izračunati procijenjena vrijednost bilo koje nekretnine koja posjeduje te karakteristike. Ipak za razliku od onoga što se provodi u praksi (nekoliko usporedivih nekretnina), za preciznost procjene ovom metodom je potreban što veći broj nekretnina sa što većim brojem karakteristika. To predstavlja najveći nedostatak hedoničkog modeliranja, jer su takve baze podataka u pravilu rijetke (Slišković, 2016).

S obzirom da se bazira na velikim heterogenim uzorcima, hedonička metoda se ne koristi za pojedinačne procjene tržišne vrijednosti u praksi u Hrvatskoj. No to ne znači da ne postoji mogućnost njezine praktične primjene, Naime, implicitne cijene koje su rezultat dobro definiranog hedoničkog modela i kvalitetne baze podataka sa velike brojem promatranja mogu biti izvrstan temelj za masovnu procjenu cijene nekretnina (Slišković, 2016).

3. Primjena analitičkih metoda u procjeni nekretnina

Konvencionalne analitičke metode nisu se sposobne nositi sa današnjim izazovima primjene ogromne veličine podataka za dobivanje preciznih analiza u realnom vremenu. Korištenjem novih nekonvencionalnih metoda analize su dostupne u svakom momentu i svakom koraku. Štoviše uz standardne varijable procjene nekretnina mogu i obrađivati širok spektar drugih promjenjivih varijabli. Svjedočimo svakodnevnom promjenu naše okoline, od migracija stanovništva, otvaranja i zatvaranja restorana, promjene strukture stanovništva, promjene kupovne moći i ostalo. Danas u dinamičnom svijetu za dovoljno preciznu analizu moramo uzeti širok spektar varijabli i podataka da bi ostali u koraku sa konkurencijom. Današnja napredna analitika nam je već mobilno dostupna na našim pametnim telefonima. Jedan od takvih alata koji se koristi u radu je i algoritam slučajnih šuma, o čemu će biti riječi u nastavku.

3.1. Princip rada algoritma slučajnih šuma

Algoritam slučajnih šuma, kasnije RF, je općeniti naziv za skupinu metoda koje se koriste stablastim klasifikatorima {hx k (, , 1,...., Θ = k) } gdje je {Θk} skup jednoliko distribuiranih, međusobno potpuno neovisnih, vektora, a x ulazni vektorski uzorak. Prilikom treniranja, RF algoritam stvara velik broj stabala, od kojih se svako trenira na određenom broju uzoraka originalnog trening seta, i vrši pretragu samo po slučajno generiranom podskupu ulaznih varijabli kako bi odredio mjesto na kojem će se razgranati. Za klasifikaciju svako stablo unutar RF daje glas jednoj od klasa unutar skupa x. Izlaz klasifikatora ovisi o broju glasova stabala svakoj pojedinoj klasi (Breiman, 2001).

Trening set za svako stablo stvara se tako da se iz podataka za treniranje uzme određeni broj instanci ali na slučajan način. Iz tako stvorenog seta za treniranje stabla, jedna trećina instanci se odvaja. Ove instance se nazivaju oob instance (engl. *out of bag*) i koriste se za dobivanje nepristrane procjene greške klasifikacije. Također se koriste i za procjenu važnosti pojedinih varijabli ulaznih instanci. Kako se stablo stvara sve instance se puštaju duž stabla te se računaju njihove međusobne sličnosti. Ako se dvije instance nađu u istom konačnom čvoru njihova se međusobna sličnost povećava za 1. Kada se sve instance provuku kroz

stablo, sličnosti se normiraju tako da se podjele sa brojem stabala. Kod slučajne šume nema potrebe za unakrsnom validacijom ili korištenjem posebnog seta za testiranje kako bi se dobila nepristrana procjena greške uzoraka za testiranje. Svako stablo se stvara tako da se koristi podskup iz početnih podataka za učenje koji se naziva bootstrap podskup. Otprilike jedna trećina instanci se izostavlja iz bootstrap podskupa i ne koriste se pri izradi k-tog stabla. Sada svaki uzorak izostavljen pri stvaranju k-tog stabla, oob instance, treba pustiti niz k-to stablo da bi se dobila klasifikacija. Na ovaj način dobiva se klasifikacija testnog seta za svaku instancu u jednoj trećini svih stabala. Nakon završene obrade, neka je klasa j klasa koja je dobila najviše glasova svaki put kad je instanca n bila oob instanca. Omjer broja izlaza kada j nije bila jednaka pravoj klasi instance n s obzirom na sve instance naziva se procjena pogreške oob-a (Breiman, 2001).

U svakom stablu stvorenom u šumi zanemaruju se oob instance i zbrajaju se glasovi koji su ispravno doneseni s obzirom na klasu. Sada se na slučajan način permutiraju vrijednosti varijable m u oob instancama i te se instance puštaju niz stablo. Oduzima se broj glasova za ispravnu klasu oob instanci sa permutiranom m varijablom od broja glasova za ispravnu klasu ne upotrijebljenih oob instanci. Srednja vrijednost dobivene razlike u svim stablima unutar šume naziva se važnost varijable m. Ako je broj varijabli jako velik, znači svaka instanca se sastoji od većeg broja varijabli, moguće je obaviti klasifikaciju sa svim varijablama, pa opet ponoviti postupak samo sa najbitnijim varijablama(Breiman, 2001).

3.2. Konvergencija algoritma slučajnih šuma

Uz skup klasifikatora h1(x), h2(x),…, hK(x) te trening setom stvorenim od nasumce izabranih instanci iz distribucije slučajnih vektora Y, X, funkcija margine glasi:

$$mg(X,Y) = av_k\, I(h_k(X)=Y) - \max_{j \neq Y} av_k\, I(h_k(X)=j) \qquad (3)$$

gdje I() je indikatorska funkcija. Margina opisuje broj za koliko glasova prosječan broja glasova za X,Y za ispravnu klasu nadmašuje prosječan broj glasova za bilo koju drugu klasu. Što je veća margina to je klasifikator točniji. Greška prilikom generalizacije zapisuje se kao:

$$PE^* = P_{X,Y}(mg(X,Y) < 0) \quad (4)$$

Gdje ovi mali X i Y označavaju činjenicu da se vjerojatnost računa u X, Y prostoru. Kod slučajnih šuma h X hX k ( ) = (,Θk) . Kako se broj stabala povećava za gotovo sve sekvence Θ1,…. PE* konvergira prema:

$$P_{X,Y}(P_\Theta(h(X,\Theta)=Y) - \max_{j \neq Y} P_\Theta(h(X,\Theta)= j) < 0) \quad (5)$$

Ovaj rezultat objašnjava zašto slučajne šume ne generaliziraju što se više stabala dodaje šumi (Breiman, 2001).

3.3. Snaga i korelacija

Kod slučajnih šuma postoje dva parametra koji definiraju gornju granicu greške generalizacije. Oni ujedno služe kao mjera točnosti klasifikatora te njihove međusobne ovisnosti. Funkcija margine za slučajnu šumu glasi:

$$mr(X,Y) = P_\Theta(h(X,\Theta)=Y) - \max_{j \neq Y} P_\Theta(h(X,\Theta)= j) \quad (6)$$

a snaga skupa klasifikatora {h x(,Θ)} je:

$$s=E_{X,Y}\, mr(X,Y) \quad (7)$$

Uz pretpostavku da je $s \geq 0$, Čebiševljeva nejednakost glasi

$$PE^* \leq var(mr)/s^2 \quad (8)$$

Izraz koji još bolje opisuje devijaciju mr može se izvesti iz slijedećeg. Neka je

$$j(X,Y)=\arg\max_{j \neq Y} P_\Theta(h(X,\Theta)= j) \quad (9)$$

dalje možemo pisati

$$mr(X,Y) = P_\Theta(h(X,\Theta)=Y) - P_\Theta(h(X,\Theta)= \hat{}j(X,Y)) =$$

$$E_\Theta[I(h(X,\Theta)=Y) - I(h(X,\Theta)= \hat{}j(X,Y))] \quad (10)$$

Funkcija neobrađene margine glasi

$$rmg(\Theta,X,Y)= I(h(X,\Theta)=Y)-I(h(X,\Theta)= \hat{}j(X,Y)). \quad (11)$$

Potrebno je primijetiti da je mr(X,Y) očekivanje rmg(Θ,X,Y) s obzirom na Θ. Za svaku funkciju f identitet

$$[E_\Theta f(\Theta)]^2 = E_{\Theta,\Theta'} f(\Theta) f(\Theta') \quad (12)$$

Postoji ako Θ i Θ' su međusobno neovisne ali sa istom distribucijom. Dalje slijedi

$$mr(X,Y)^2 = E_{\Theta,\Theta'} rmg(\Theta,X,Y)rmg(\Theta',X,Y) \quad (13)$$

Koristeći gornji izraz možemo pisati

$$var(mr)=E_{\Theta,\Theta'} (covX,Y\ rmg(\Theta,X,Y)rmg(\Theta',X,Y))$$

$$= E_{\Theta,\Theta'} (\rho(\Theta,\Theta')sd(\Theta)sd(\Theta')) \quad (14)$$

Gdje ρ(Θ, Θ') predstavlja korelaciju između rmg(Θ, X, Y) i rmg(Θ', X, Y) uz fiksne Θ i Θ', a sd(Θ) standardna devijacija rmg(Θ, X, Y) uz fiksni Θ. Dalje je

$$var(mr) = \rho(E_\Theta sd(\Theta))^2$$

$$\leq \rho E_\Theta\ var(\Theta) \quad (15)$$

Gdje ρ predstavlja srednju vrijednost korelacije, i to

$$\rho= E_{\Theta,\Theta'} (\rho(\Theta,\Theta')sd(\Theta)sd(\Theta'))/E_{\Theta,\Theta'} (sd(\Theta)sd(\Theta')) \quad (16)$$

Gornja granica generalizacijske pogreške glasi:

$$PE^* \leq \rho (1-s^2)/s^2 \quad (17)$$

Ovaj izraz pokazuje kako dva osnovna elementa pogreške generalizacije slučajnih šuma su snaga pojedinog klasifikatora unutar šume i korelacija među njima. Izraz c/s2 predstavlja omjer korelacije i kvadrata snage. Da bi se shvatio način na koji slučajne šume funkcioniraju ovaj omjer biti će jako koristan, što je on manji to bolje. Omjer c/s2 za slučajnu šumu definiran je kao

$$c / s^2 = \rho / s^2 \quad (18)$$

U slučaju da ulazni podaci imaju samo dvije klase dolazimo pojednostavljenja. Funkcija margine glasila bi

$$mr(X,Y)=2P_{\Theta}(h(X,\Theta)=Y)-1 \quad (19)$$

Neobrađena margina izgleda 2 , IhX Y ((Θ) = −) 1 a korelacija ρ je između I ( ) hX Y ( ) ,Θ = i I (hX Y ( ) ,Θ = ′) . U slučaju da su vrijednosti Y +1 i -1 slijedi

$$\rho = E_{\Theta,\Theta'}\,[\rho(h(\cdot,\Theta),h(\cdot,\Theta'))] \quad (20)$$

3.4. Korištenje out-of-bag instanci za praćenje pogreške, snage i korelacije

U slučajnim šumama koristi se odvajanje (bagging) u tandemu sa odabiranjem slučajne varijable. Svaki novi trening set stabla zapravo je podskup iz ulaznog trening seta slučajne šume, koji se vraća natrag te se uzima slijedeći. Stablo se grana na osnovu navedenog podskupa koristeći nasumično odabiranje varijable.

Postoje dva razloga zašto se upotrebljava odvajanje (bagging). Prvi je taj da odvajanje povećava točnost kada se varijable uzimaju nasumce. Drugi razlog je taj da odvajanje omogućuje prikazivanje pogreške generalizacije (PE*) skupa stabala, te ocjene snage i korelacije. Ako imamo trening set T, konstruiramo klasifikatore h(x, TK) iz podskupa TK trening seta T te ih puštamo da rade i donose glasove kako bi smo dobili odvajajući (bagged) prediktor. Za svaki y, x u trening setu pamte se samo glasovi klasifikatora TK koji nisu vidjeli y, x. Ovo se naziva odvajajući oob (out-of-bag) klasifikator. Tada je oob ocjena generalizacijske greške zapravo greška oob klasifikatora na testnom setu. U svakom podskupu trening seta (bootstrap) oko jedne trećine instanci se izostavlja. Na osnovu toga oob ocijene se temelje na zbroju ocjena jedne trećine svih klasifikatora. Treba primijetiti da za razliku od unakrsne validacije kod koje ocjene u određenom postotku ovise jedna o drugoj, kod oob ocjena te međuovisnosti nema. Snaga i korelacija također mogu biti ocjenjene koristeći oob metode. Ovo daje interne ocjene koje mogu pomoći u razumijevanju klasifikacijske točnosti i kako ju poboljšati (Breiman, 2001).

3.5. Prednosti i nedostaci

Kao jedna od prednosti algoritma slučajnih šuma je ta što se može koristiti i za regresiju i za klasifikaciju, dajući joj širinu primjene. Osim toga sam algoritam je lagan za korištenje i daje relativno dobre rezultate predviđanja.

Jedan od problema koji se može pojaviti u strojnom učenju je prekomjerna specijalizacija modela (engl.*overfitting*) ali što se tiče algoritma slučajnih šuma taj problem se rješava dovoljnim brojem stabala. S druge strane sa velikim brojem stabala se algoritam slučajnih šuma usporava i postaje neupotrebljiv za obradu podataka u stvarnom vremenu. Uglavnom pomoću algoritma slučajnih šuma model se brzo istrenira, ali kako za predviđanje treba veći broj stabala, nije najbolji algoritam za aplikacije koje trebaju predviđanje u stvarnom vremenu.

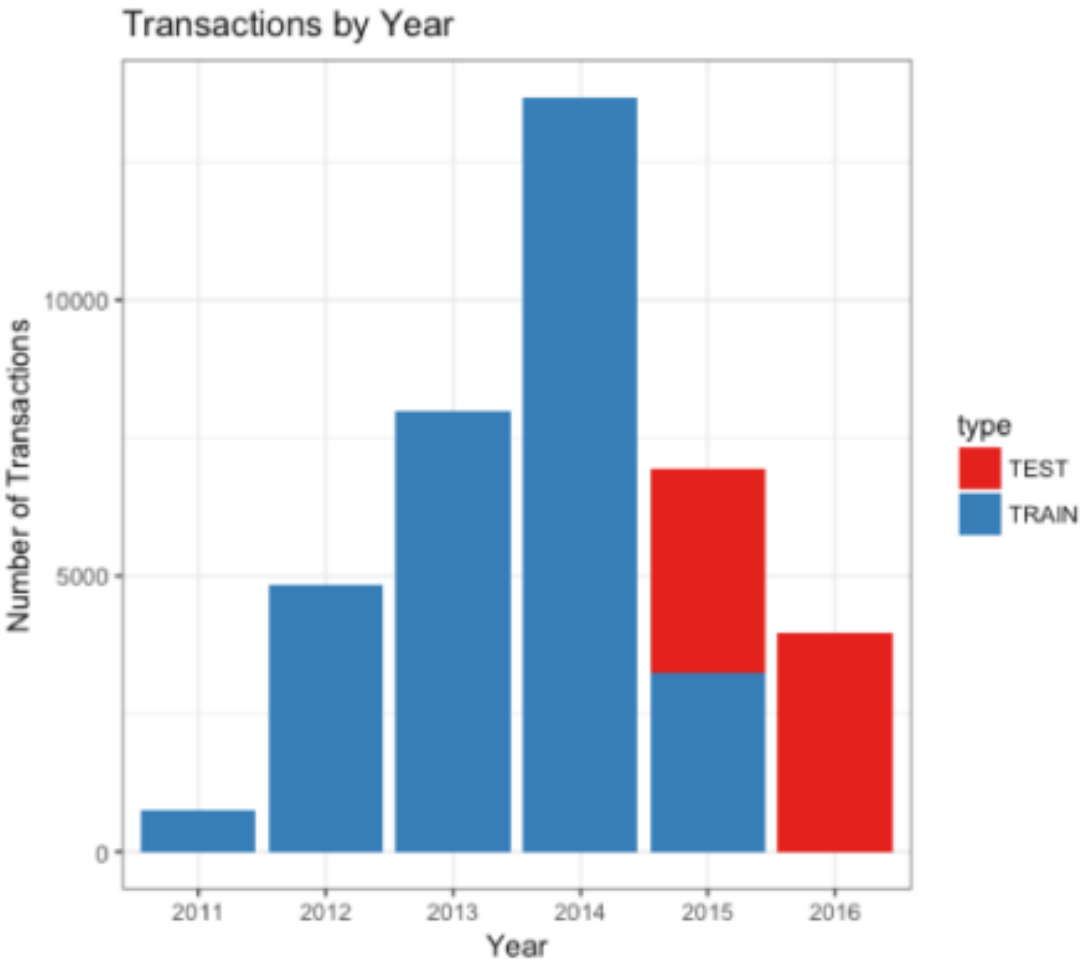
Moguće primjene algoritma su u bankarstvu (predviđanje kreditne sposobnosti klijenata, bankovnih prijevara), financije (predviđanje kretanja dionica), medicina (predviđanje bolesti temeljem povijesti bolesti) i e-trgovine (vjerojatnost prodaje proizvoda).

4. Analiza tržišta nekretnina i predviđanje cijena nekretnina korištenjem algoritma slučajnih šuma

Prije procjene cijene nekretnina prvo treba napraviti analizu podataka koji su dostupni da se vidi utjecaj koji imaju na procjenu cijene nekretnina. Za analizu podataka koristimo programski jezik Python verzija 3.7 koristeći Jupyter Notebook. Jupyter Notebook je web aplikacija otvorenog koda za dijeljenje programskog koda, vizualizacije i teksta koja se koristi za čišćenje podataka, statističku obradu podataka, vizualizaciju podataka, strojno učenje i sve druge stvari vezane uz moderan pristup izvlačenja informacija iz velike količine informacija.

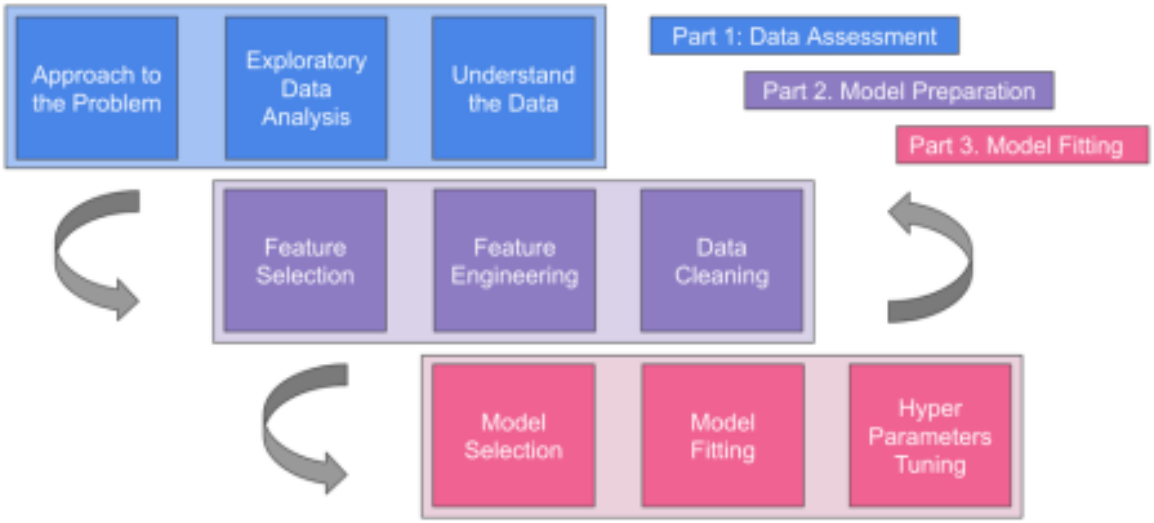
Osnova na kojoj se provodi analiza su podaci. Setovi podataka su preuzeti kao excel tablice, a sastoje se od podataka za treniranje, podataka za testiranje i povijesnih makroekonomskih podataka. Podaci za treniranje se sastoje od oko 21000 stvarnih transakcija u razdoblju od kolovoza 2011. do lipnja 2015. godine, uključujući i karakteristike nekretnina i okružja u kojem se nalaze. Podaci za treniranje imaju i istaknutu cijenu nekretnina kao ciljanu varijablu ovog predviđanja. Podaci za testiranje su realne transakcije za razdoblje srpanj 2015. do svibanj 2016. godine. Približna količina podataka za treniranje je oko 7000 transakcija i opisuju karakteristike nekretnina i okružja u kojem se nalaze. Naravno podaci za testiranje nemaju ciljanu varijablu odnosno cijenu nekretnina, jer će se cijena nekretnina odrediti algoritmom slučajnih šuma. Makroekonomski podaci se sastoje od približno 400 varijabli za predviđanje, kao što su demografski podaci, bruto domaći proizvod, inflacija, plaća, zaposlenost i druge značajke. Na grafikonu 4.1 je vidljivo raspored podataka za treniranje i podataka za testiranje. Najviše podataka za treniranje je iz 2014. godine., pa je pretpostavka da bi oni mogli činiti osnovicu za predviđanje cijene nekretnina podataka za testiranje. Osim gore navedenih preuzetih podataka imamo još i primjerak završne excel tablice u koju se ispisuje predviđena cijena nekretnina zajedno sa rednim brojem redaka podataka za testiranje.

Ispod prikaza strukture podataka nalazi se grafikon 4.2 koji opisuje tehničku skicu cjelokupnog postupka predviđanja cijene nekretnina. Postupak se odvija u više koraka, koje možemo ugrubo grupirati u tri dijela, procjena odnosno analiza podataka, pripremanje podataka za algoritam slučajnih šuma i podešavanje te samo predviđanje algoritmom slučajnih šuma.



Grafikon 4-1 Broj transakcija po godini uzimajući u obzir podatke za treniranje i testiranje

Prije nego što se bacimo na detaljnu analizu podataka upoznati ćemo ekonomsku situaciju ruske ekonomije. U tom dijelu ćemo se pozabaviti svim utjecajima na rusku ekonomiju, ponajviše vanjskim utjecajima uzrokovanih sukobom u Ukrajini.



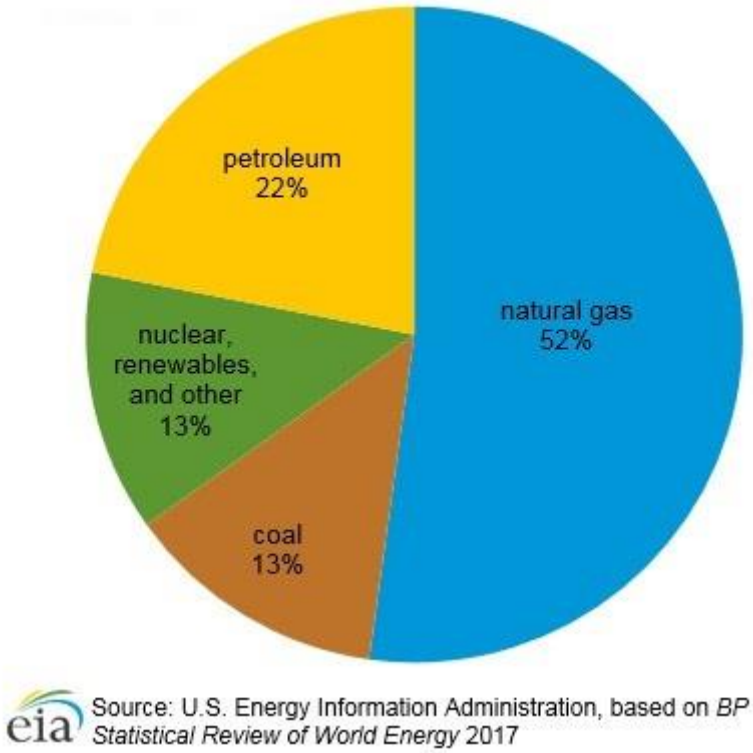
Grafikon 4-2 Tehnička skica postupka

Nakon toga uvodnog dijela ekonomske problematike prelazimo na detaljnu analizu podataka koje koristimo. Redom ćemo uspoređivati podatke koje utječu na cijene nekretnina da bi uvidjeli njihov utjecaj. Tu se najviše misli na podatke o samim nekretninama (površina, broj soba i sl.), pa onda na okružje u kojima se nekretnine nalaze (blizina škola, crkvi i ostalog sadržaja), demografske osobine (dob, spol i sl.) populacije u navedenim sredinama i na kraju analizu strukture podataka za treniranje i podataka na testiranje. Ukratko vidjet ćemo samu strukturu podataka i njihov međusobni odnos koji će nam pomoći u predviđanju cijene nekretnina. Kod za detaljnu analizu podataka se nalazi u prilogu naziva kod za analizu podataka.

U završnom dijelu koji se nalazi u prilogu, naziva kod za predviđanje cijene nekretnina, ćemo u tri koraka odraditi pripremu podataka i postaviti model za predviđanje cijene nekretnina. Priprema podataka se sastoji od čišćenja podataka, odabira najrelevantnijih podataka za predviđanje cijene nekretnina, postupak zamjene nedostajućih podataka odgovarajućim veličinama i sl.

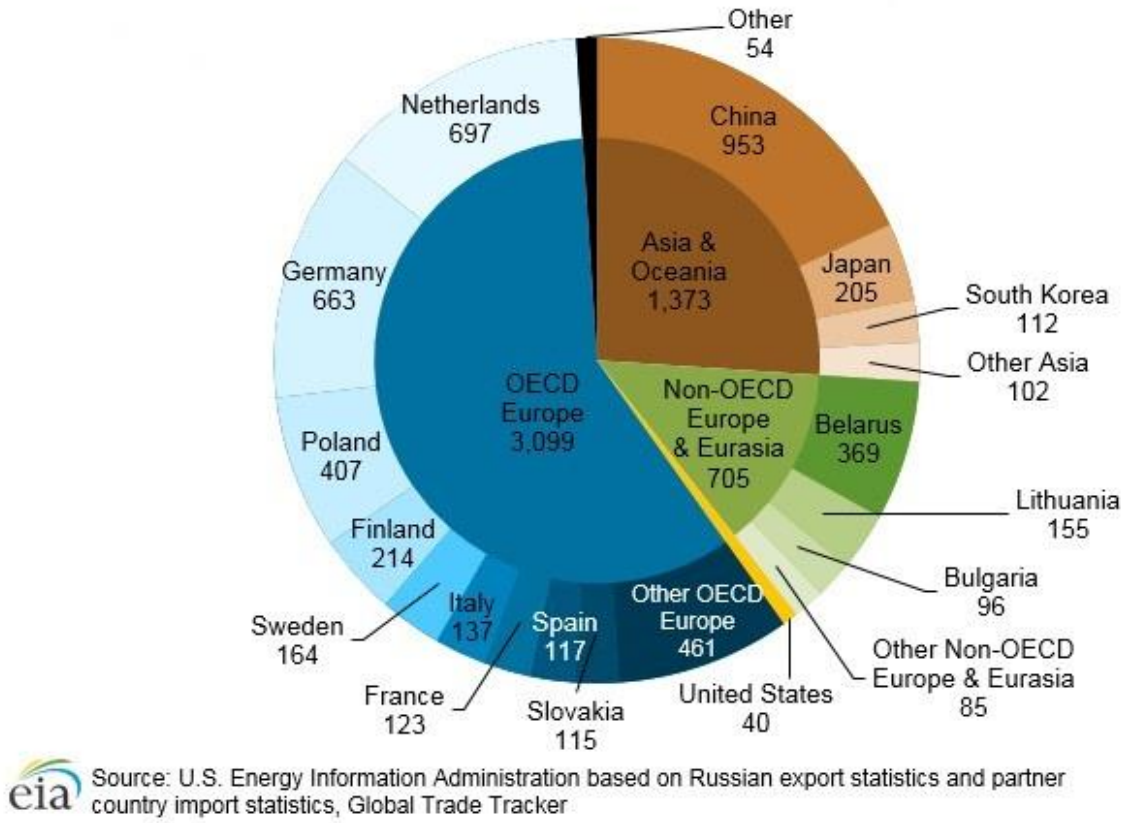
4.1. Nacionalna ekonomija na primjeru Rusije

Na kretanje cijena nekretnina utječe i ruska ekonomija. U ovom slučaju dva su najvažnija utjecaja za razdoblje od 2011. do 2016. godine bili utjecaj energetskog sektora i nametnutih sankcija.



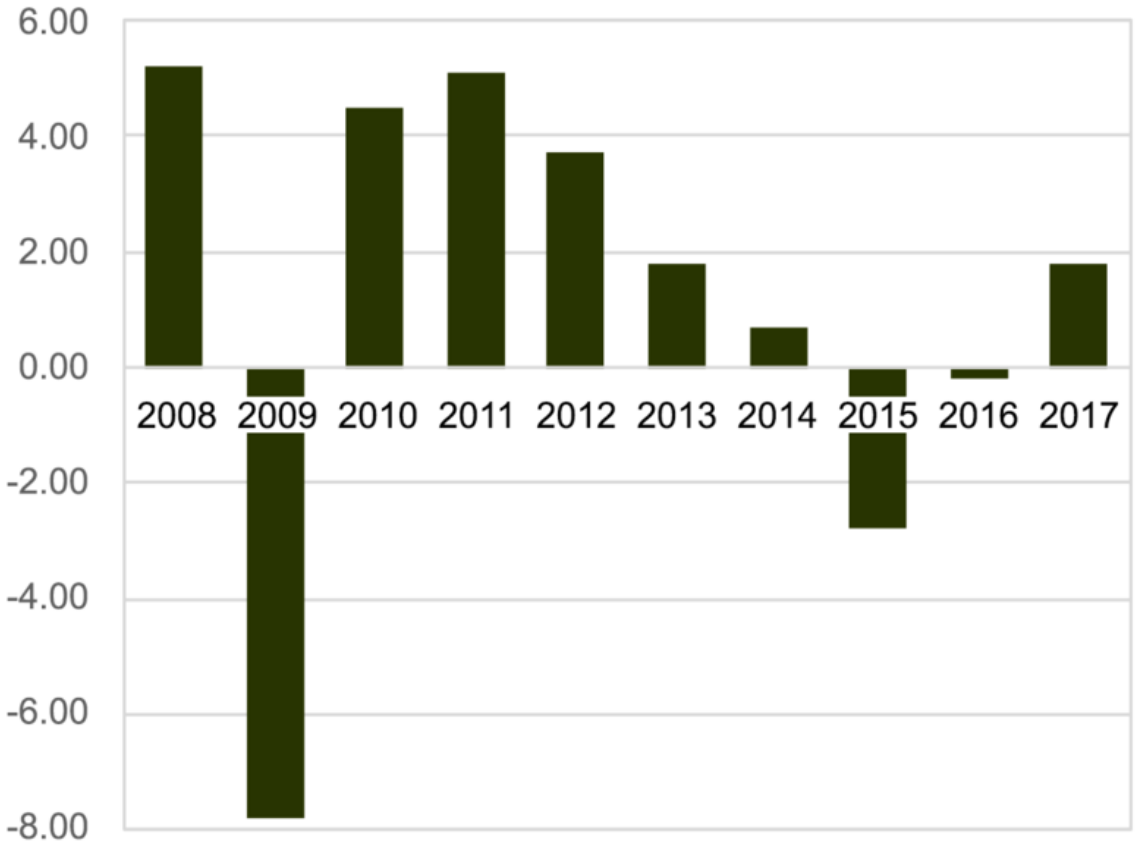
Grafikon 4-3 Ruska potrošnja primarnih energenata

Rusija i Europa su međusobno ovisne kada je u pitanju energija. Europa ovisi o ruskoj nafti i plinu. Više od trećine sirove nafte i 70% prirodnog plina 2016. je uvezeno iz Rusije. Naravno s druge strane i Rusija ovisi o prihodima koje ostvaruje sa članicama Europske Unije. Sankcije koje su joj 2014. nametnute, zbog vojnog sukoba sa Ukrajinom, od strane SAD-a i EU, učinile su rusku ekonomiju nestabilnom. Kada još uz to nadodamo pad cijene nafte od 50% u prvoj polovici 2014. godine pad BDP-a i porast inflacije je bio neminovan (EIA, 2019.).



Grafikon 4-4 Uvoznici ruske nafte 2016. godine

To je rezultiralo devalvacijom rublja u drugoj polovici 2014. godine, padom ruskog burzovnog indeksa za 30% u prosincu 2014. godine i smanjenjem robne razmjene (Wikipedia, 2017).



Grafikon 4-5 Stopa rasta BDP-a Rusije u postotku za razdoblje 2008.-2017.

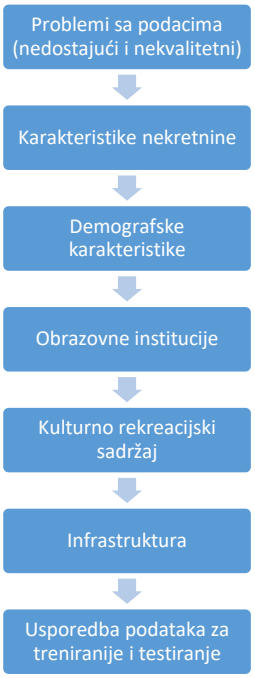
Nastavno na ekonomske sankcije 2014. godine i devalvaciju rublja u drugoj polovici 2014. godine dogodio se i pad stope rasta BDP-a u 2015. i 2016. godini. Na grafikonu se vidi oporavak u 2017.godini. S obzirom na dobivene podatke tek ćemo u nastavku vidjeti kako će kriza utjecati na predviđanje cijene nekretnina.

4.2. Detaljna analiza podataka

Nakon kratke uvodne analize stanja ruske ekonomije prijeći ćemo na detaljnu analizu podataka da bi uvidjeli utjecaj varijabli na predviđanje cijena nekretnina. Naravno na samom početku ćemo se pozabaviti općenitim karakteristikama podataka. U radu želimo vidjeti kakve su strukture podaci za treniranje, koja su obilježja i koje mogućnosti. Za početak treba

programski kod za inicijalizaciju podataka za treniranje. Sam kod za detaljnu analizu podataka se nalazi u prilogu.

Detaljnu analizu podataka ćemo podijeliti u više etapa, krenuvši prvo od problema koji se nalaze u podacima. U osnovi to su problemi sa podacima kojih nema i problemi sa podacima koji proizlazi iz krivih zapisa u bazama podataka. Nakon toga ćemo se redom posvetiti svakoj pojedinoj skupini karakteristika, od karakteristike nekretnina do utjecaja infrastrukture na cijenu nekretnina. Zadnja stavka bavi se odnosom podataka za treniranje i testiranje, koja nam je važna radi stavljanja tih podataka u perspektivu prilikom predviđanja cijene nekretnine.



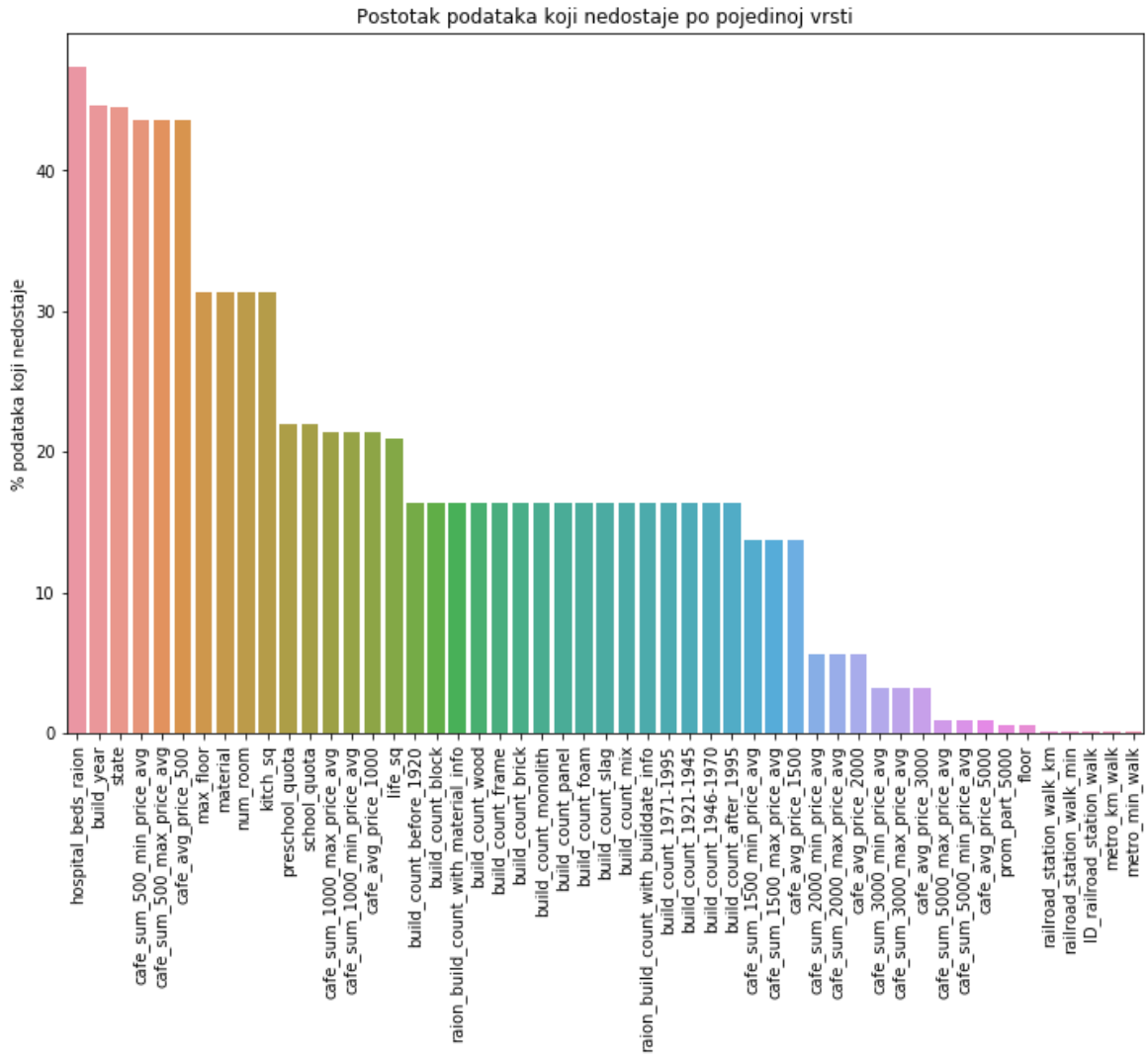
Grafikon 4-6 Skica postupka detaljne analize podataka

4.2.1. Podaci koje nedostaju

Podaci koji su na raspolaganju nisu cjelovit, zato vizualizacijom u podataka dolazimo do broja onih podataka koji nedostaju.

Kako vidimo od 291 stupca u njih 51 nedostaju podaci. Najviše ih fali u broju kreveta na 100 000 ljudi po području 47,4% Najmanje ih fali u vremenu potrebnom da najbliže stanice

podzemne željeznice 0,1 %. Samim time što je manji broj nedostajućih podataka to će ta vrsta podataka biti relevantnija odnosno procjena na temelju tih podataka će biti preciznija.



Grafikon 4-7 Postotak podataka koji nedostaje po pojedinoj vrsti podataka

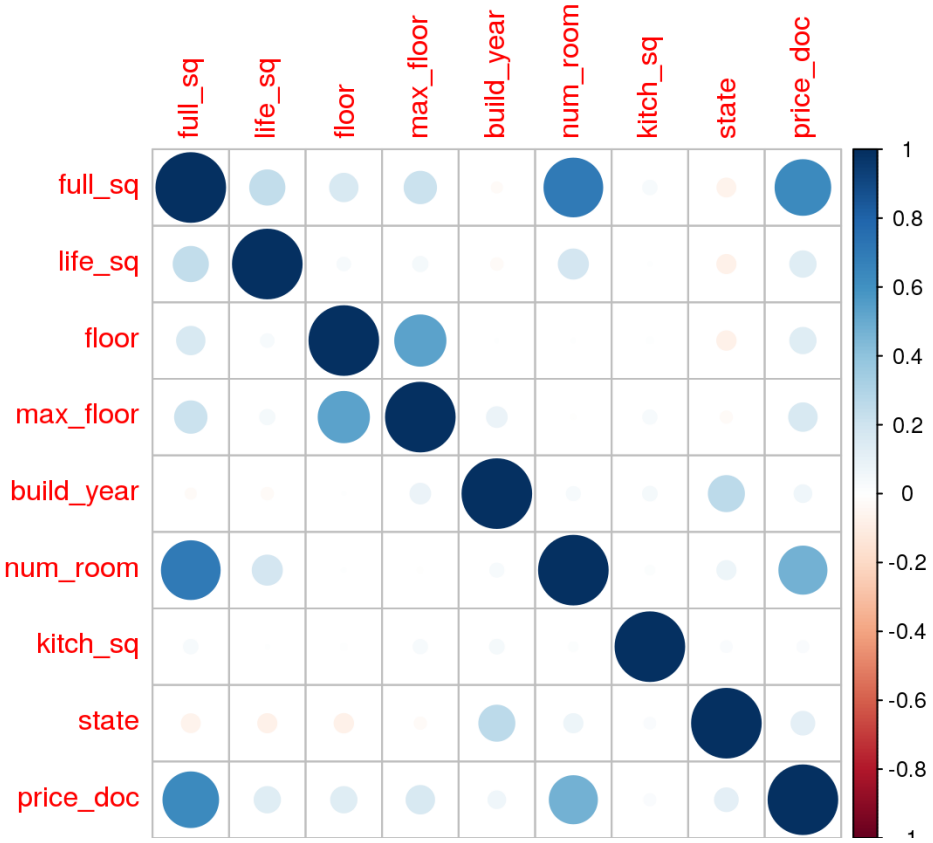
4.2.2. Ispravljanje problema sa podacima

Neke od podataka moramo ispraviti radi njihove neupotrebljivosti u samom procesu obrade podataka. Programskim kodom sam ispravio problem sa stanjem u kojem je pojedina nekretnina i godinom gradnje jer su se pojavile veličine koje nemaju smisla. Stanje nekretnina se određuje brojčanim veličinama od 1 do 4, dok se u podacima pojavio brojčani

podataka 33. U godini gradnje postoji brojčana veličina 20052009. Kako nismo sigurni da li ta veličina treba biti 2005 ili 2009 uzet ćemo srednju veličini 2007.

4.2.3. Karakteristike nekretnina

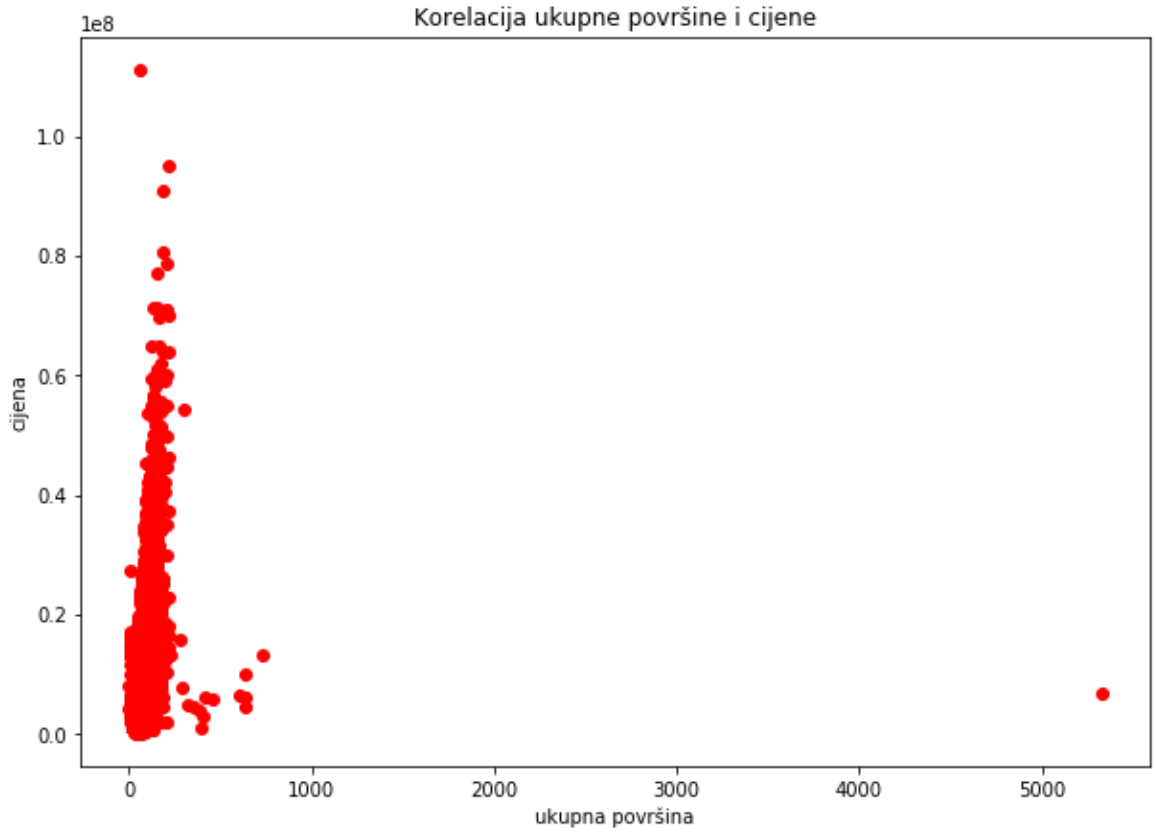
Sada ćemo pogledati korelaciju karakteristika nekretnina (ukupne površine nekretnine, neto korisne površine, godine gradnje, stanja u kojem je nekretnina i sl.) u odnosu na cijenu nekretnine.



Grafikon 4-8 Korelacija karakteristike nekretnina i cijene

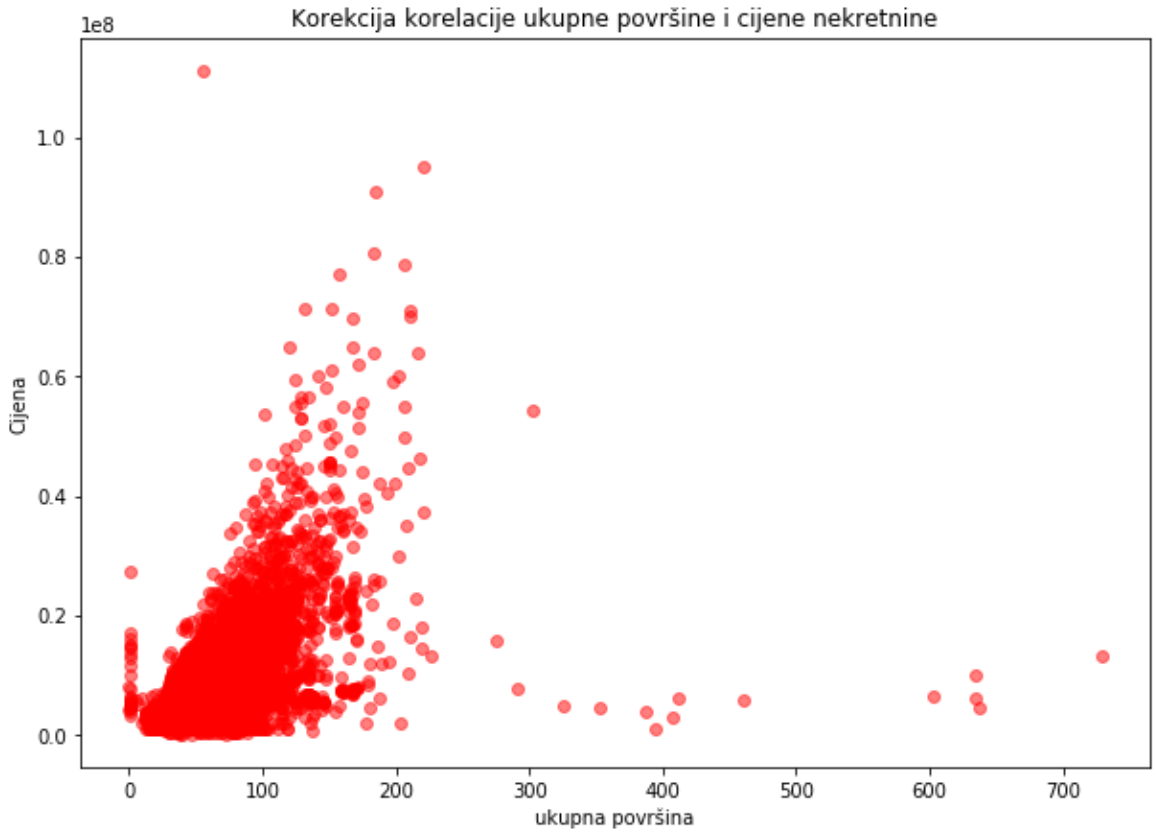
Iz grafikona proizlazi da su najveći pokazatelji koji utječu na cijenu nekretnine ukupna površina i broj soba u nekretnini. Zapanjujuće je što godina gradnje nije presudan pokazatelj u ukupnoj cijeni nekretnine. Moja pretpostavka je na to utječe postotak podataka koji

nedostaje za godinu gradnje kako je vidljivo iz Grafikona 4-1 gdje je postotak podataka koji nedostaje 45,4 %.



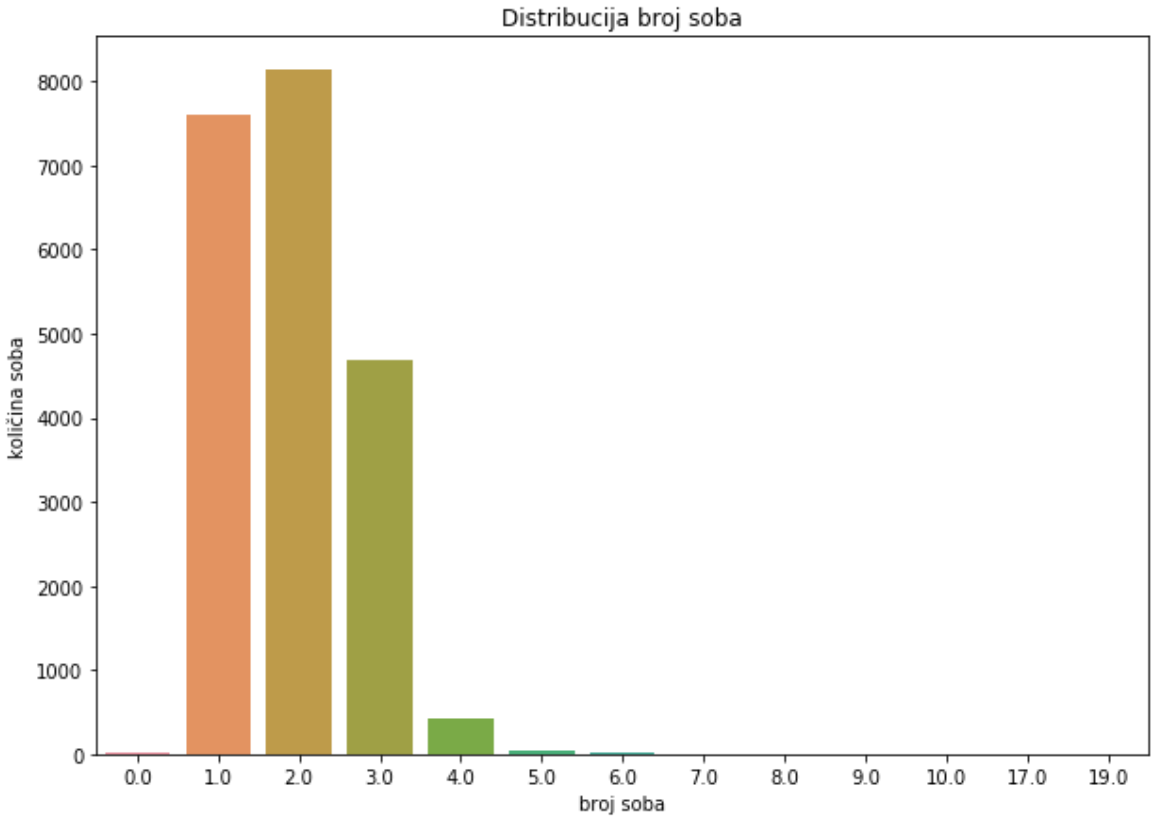
Grafikon 4-9 Korelacija ukupne površine i cijene

Grafikon 4-3 nam pokazuje jednu veličinu koja iskače iz inače pravilne distribucije korelacije ukupne površine i cijene, stoga će slijedeći korak biti eliminacija te veličine. Eliminacija te veličine se radi iz predostrožnosti i samoj činjenici da je to jedina takva veličina pa eliminacije iste ne bi trebala utjecati na detaljnu analizu podataka.



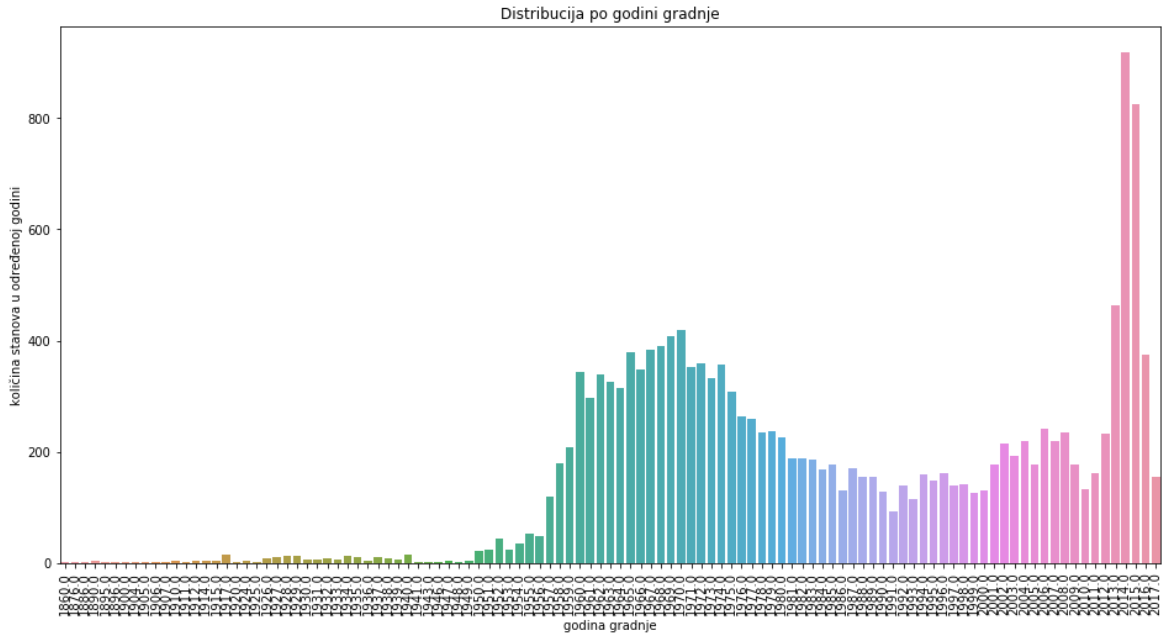
Grafikon 4-10 Korekcija korelacije ukupne površine i cijene nekretnine

Programskim kodom provjereno je koliko ima podataka za koje je neto površina veća od ukupne površine (neto površina je površina koje se odnosi na životni prostor odnosno od ukupne površine se oduzme površina balkona, lođa i sl.). Kako je taj broj 37 on je zanemariv u ukupnom skupu podataka.



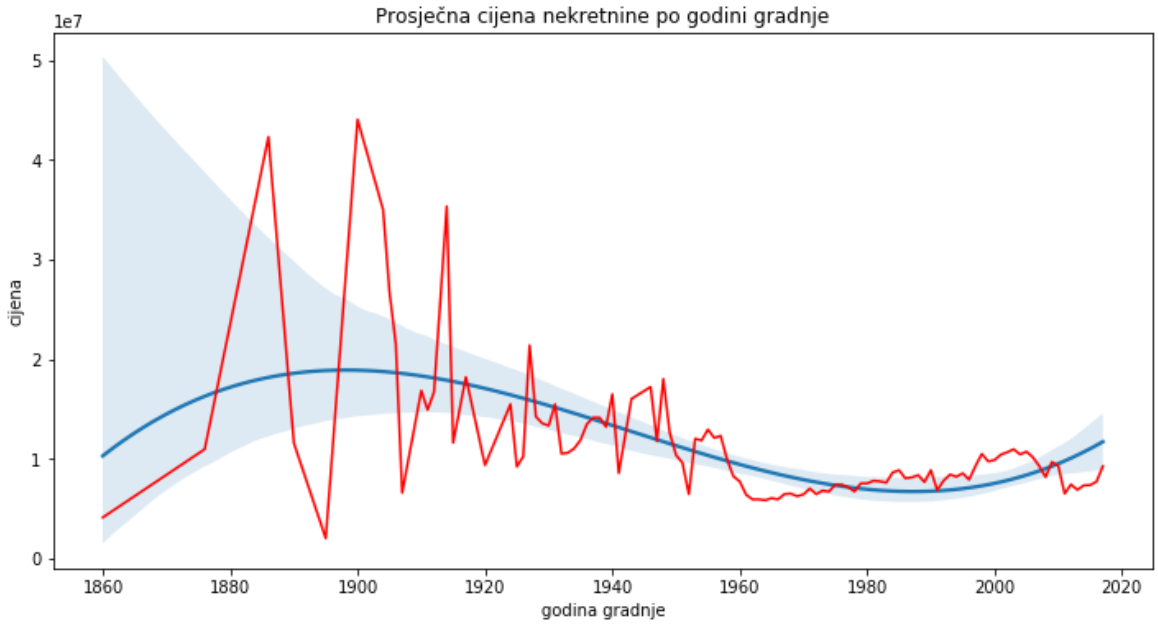
Grafikon 4-11 Distribucija broj soba

Kako je vidljivo većina stanova ima tri i manje soba, dok će se u idućem grafikonu vidjeti da su to većinom stanovi novogradnje. Naravno broj soba nam govori da su to većinom stanovi za obitelji srednjeg staleža. Kako je najveći broj stanova jednosoban i dvosoban, pa tek onda trosoban pretpostavka je da je kupovna moć ruskog kupca prosječna ili malo iznad prosjeka. U sljedećem grafikonu se vidi i utjecaj stanja ruske ekonomije na izgradnju nekretnina. U vrijeme komunizma bila je takoreći konstantna gradnja iz godine u godinu. Lagano sa približavanjem raspada Sovjetskog saveza je vidan i pad gradnje nekretnina. U novijem razdoblju rasta ruske ekonomije naglo je narasla i gradnja nekretnina.



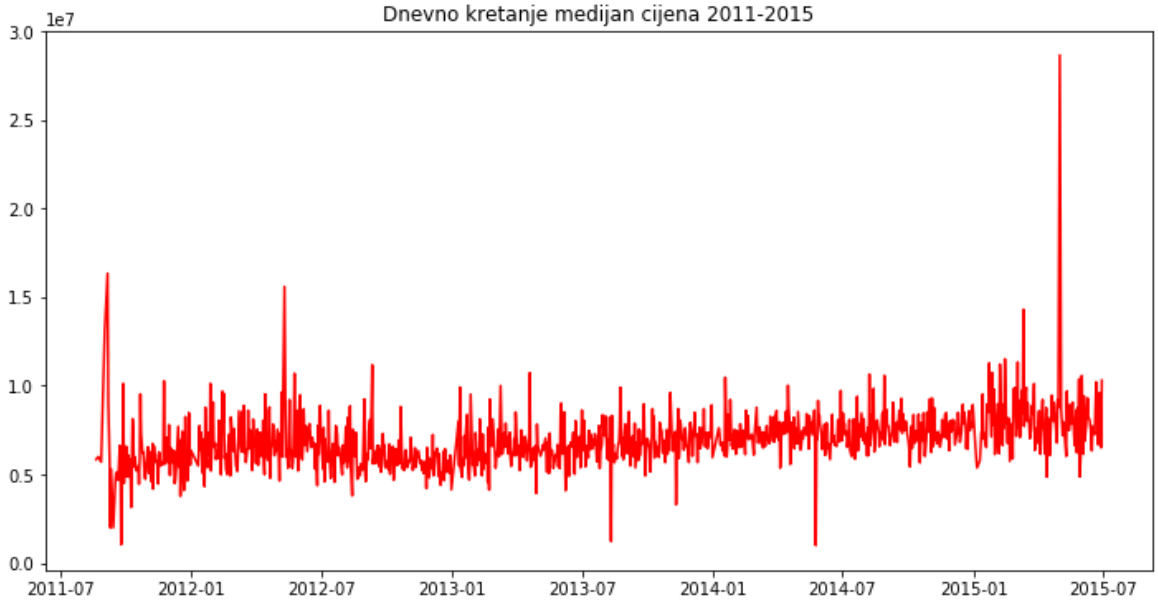
Grafikon 4-12 Distribucija po godini gradnje

Grafikon prosječne cijene nekretnine po godini gradnje treba uzeti sa rezervom zato jer do 1950. godine nema dovoljno pouzdanih podataka- Od otprilike 1960. godine se cijena nekretnina kreće u sličnim gabaritima sa blagim padovima i rastovima. Uspoređujući distribuciju po godini gradnje i prosječnu cijenu nekretnina po godini gradnje vidljivo je zašto oscilacije cijena nekretnina nisu u većem rasponu.

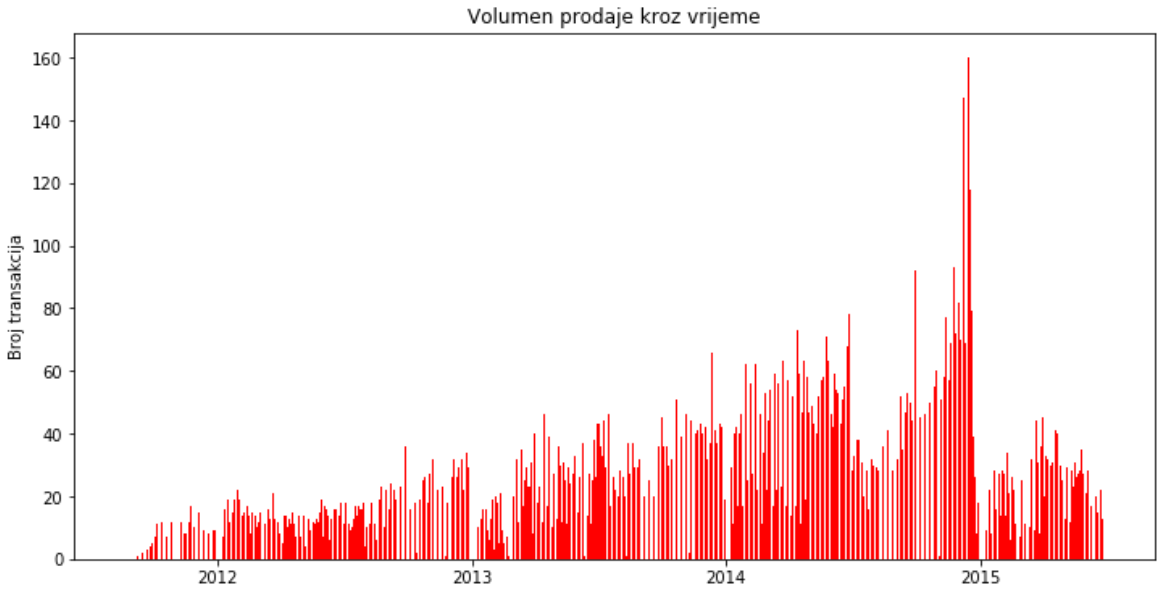


Grafikon 4-13 Prosječna cijena nekretnine po godini gradnje

Dnevni kretanje medijan cijena za razdoblje 2011. do 2015 godine nam potvrđuje malo osciliranje cijena nekretnina s iznimkama vršnih veličina. Vršne veličine pokazuju i direktan utjecaj sankcija 2014. godine na nagli pad cijena nekretnina kao i oporavak u 2015. godini.



Grafikon 4-14 Dnevno kretanje medijan cijena 2011-2015

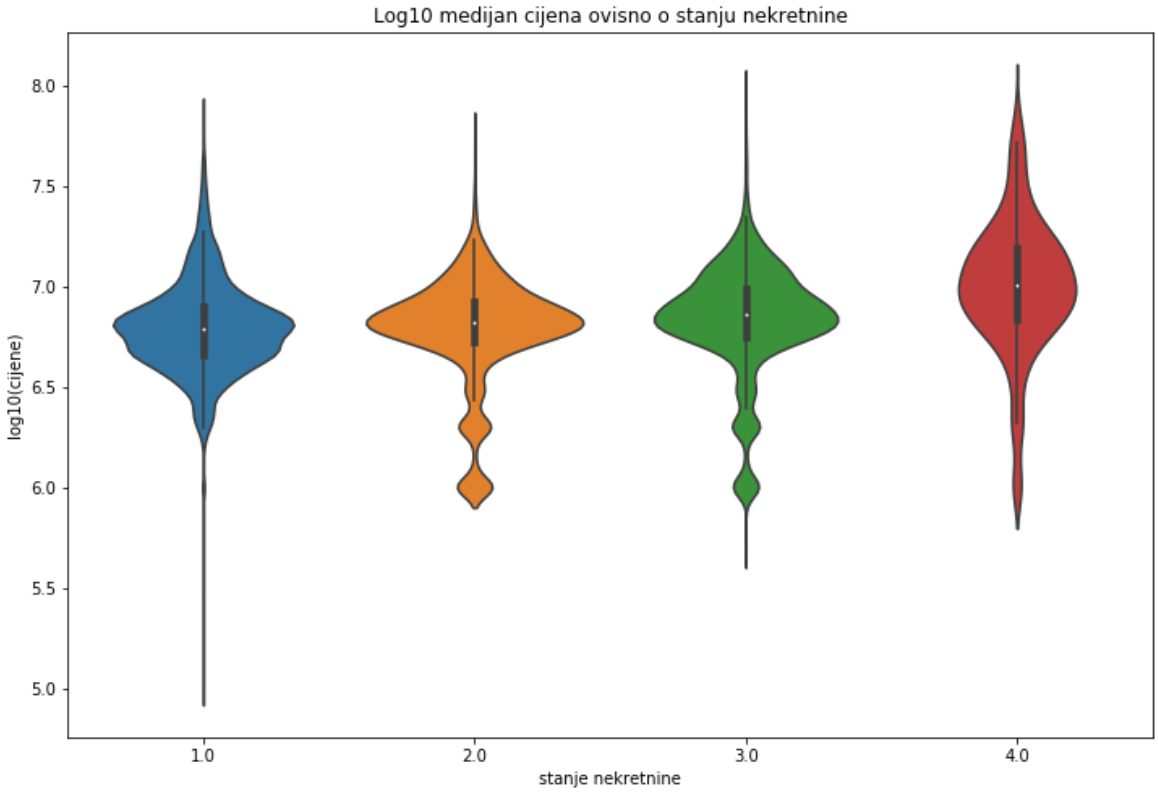


Grafikon 4-15 Volumen prodaje kroz vrijeme

Na grafikonu volumena prodaje se još ljepše vidi utjecaj rasta i pada BDP-a na cijene nekretnina. Osim utjecaja ekonomskih sankcija vidi se i zasićenje tržišta nekretnina ovisno o ponudi i potražnji. Na grafikonu cijena nekretnina po mjesecima vidljivo je da cijene nekretnina skaču na proljeće, dok u kasnu jesen naglo padaju. Pretpostavka je da ovisno o godišnjim odmorima, vremenskim prilikama i blagdanima ljudi mijenjaju svoje navike. Bez obzira na pretpostavku vidljiva su najpovoljnija razdoblja kupnje nekretnina, kao i aktivnosti rasta i pada tržišta nekretnina.

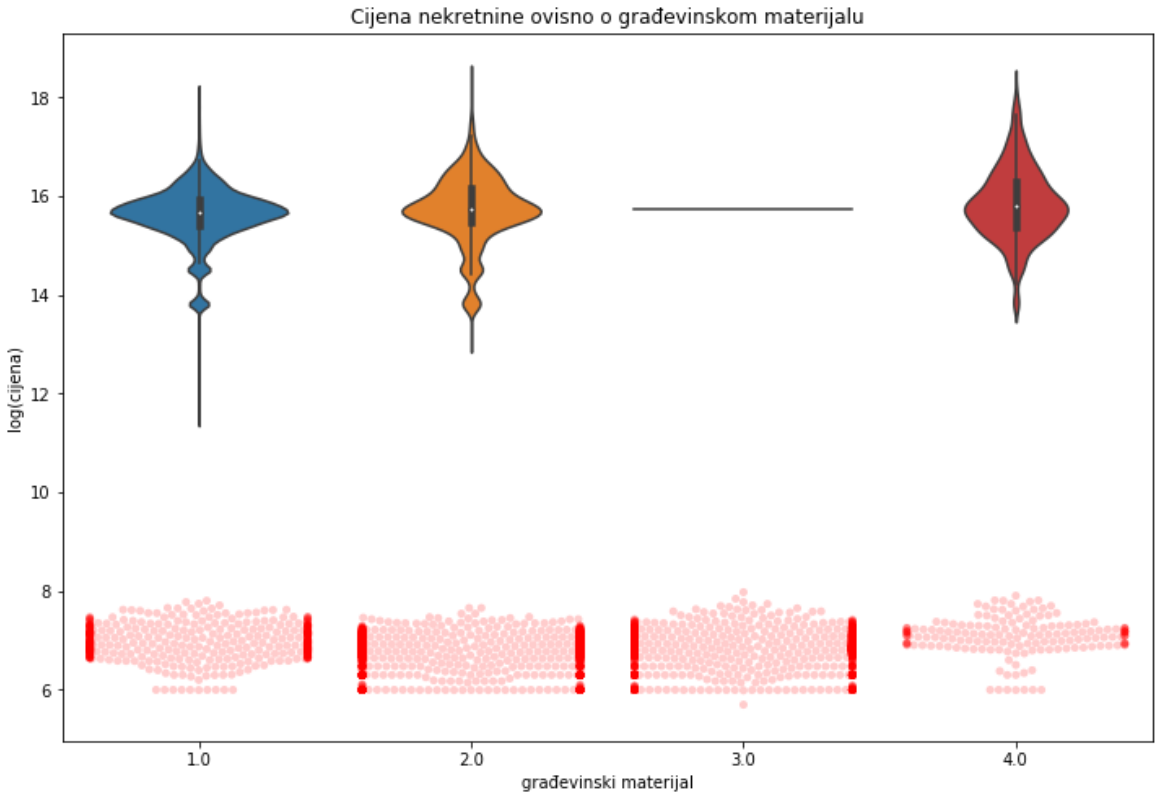
Grafikon 4-16 Cijena nekretnine po mjesecu u godini

Grafikon medijana cijena ovisno stanju nekretnine prikazane u logaritamskom mjerilu proizlazi da se nekretnine stanja 4 najskuplje mada ih ima najmanje u ukupnom volumenu. To je i shvatljivo zato jer i nekretnine u najboljem stanju i jesu najskuplje i ima ih manje. Tek nakon grafičkog prikaza stanja nekretnine nam je jasan raspon kriterija od 1 do 4 za ocjenu stanja nekretnine.



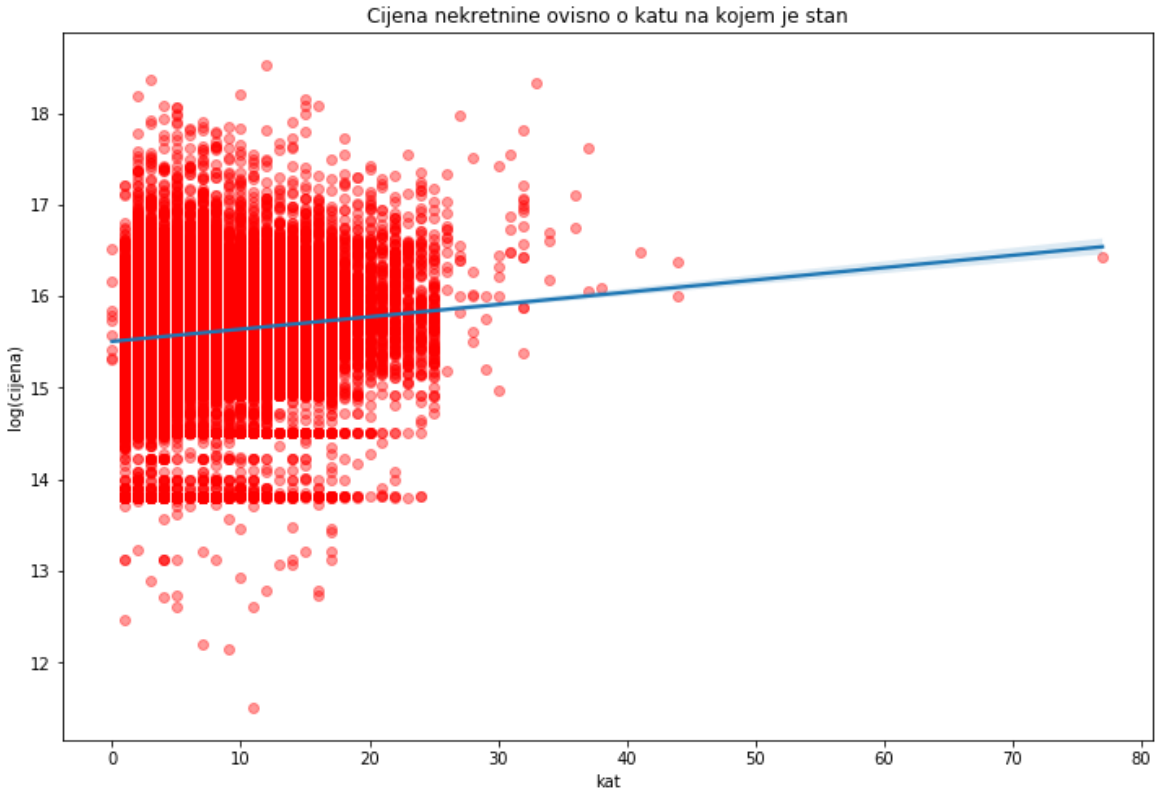
Grafikon 4-17 Log10 medijan cijena ovisno o stanju nekretnine

Vezano na grafikon ovisnosti cijena nekretnine o stanju nekretnine slijedi i grafikon ovisnosti cijene nekretnine o kvaliteti građevinskog materijala. Kao i kod stanja nekretnine raspon kvalitete građevinskog materijala je od 1 do 4. Ne samo da je raspon jednak nego je raspon kvalitete jednak. Oznaka sa brojem 4 predstavlja najkvalitetniji građevinski materijal. Najkvalitetniji građevinski materijal je po ovom grafikonu i najmanje zastupljen i naravno najskuplji.

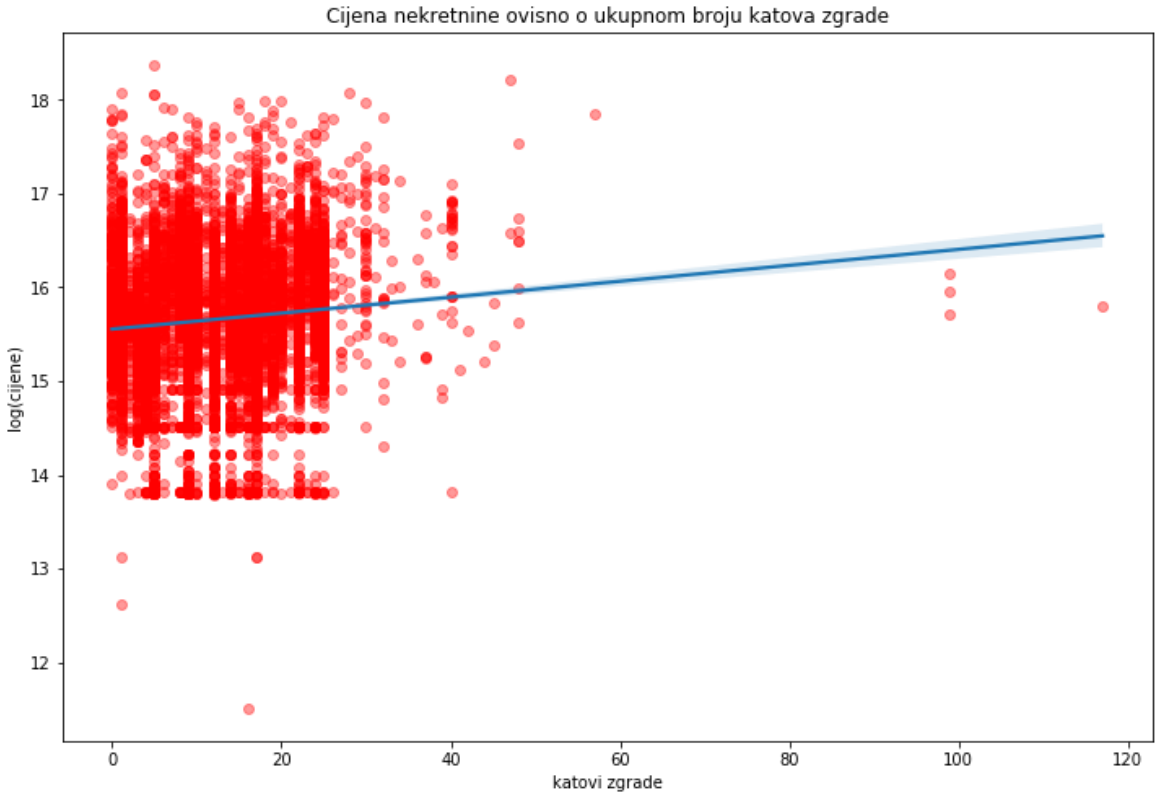


Grafikon 4-18 Cijena nekretnine ovisno o građevinskom materijalu

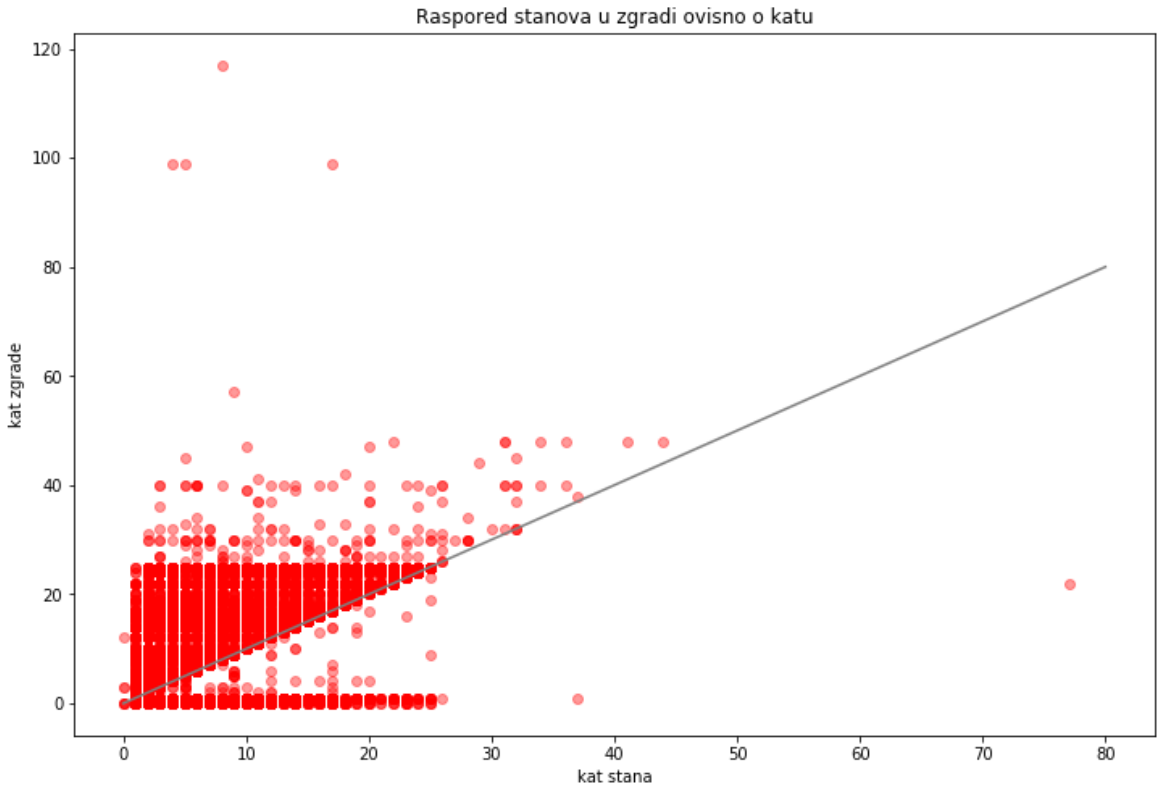
Sljedeća dva grafikona nam govore o utjecaju cijene nekretnine ovisno o katu na kojem je nekretnina i ovisno o broju katova koje nekretnina ima. Možemo povući paralelu s obzirom da broj katova utječe na lagano povišenje cijene nekretnine. Što je nekretnina na višem katu u zgradi sa više katova to je i cijena nekretnine veća. Razlika cijene nije ogromna ali vidljiv je rast cijene ovisno o katu nekretnine. I ovaj podatak ćemo uzeti sa rezervom, jer se očekuje da centar grada ima više nekretnina sa većim brojem katova., pa samim time cijene nekretnina rastu samo zbog pozicije nekretnine a ne broja katova.



Grafikon 4-19 Cijena nekretnine ovisno o katu na kojem je stan



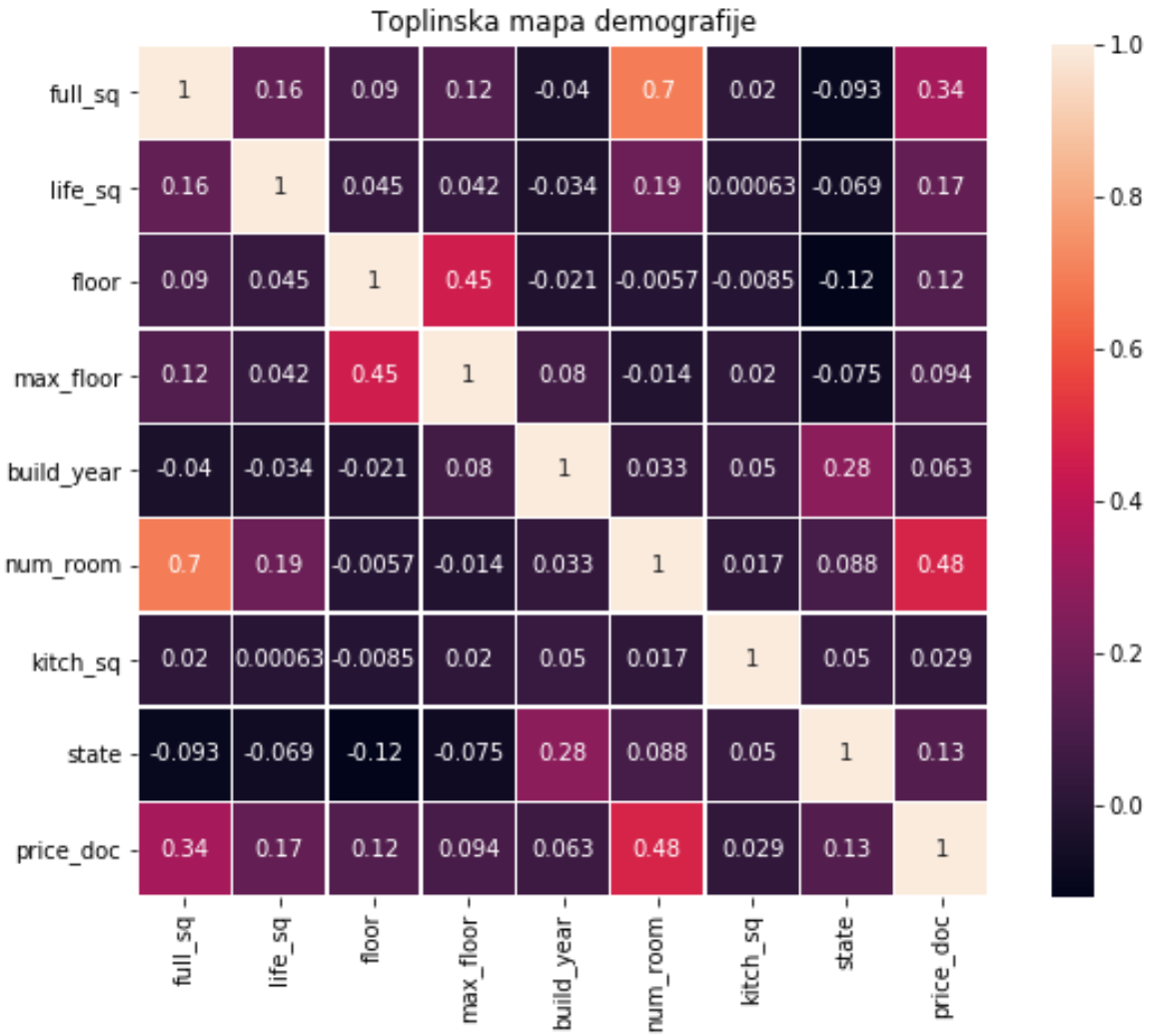
Grafikon 4-20 Cijena nekretnine ovisno o ukupnom broju katova zgrade



Grafikon 4-21 Raspored stanova u zgradi ovisno o katu

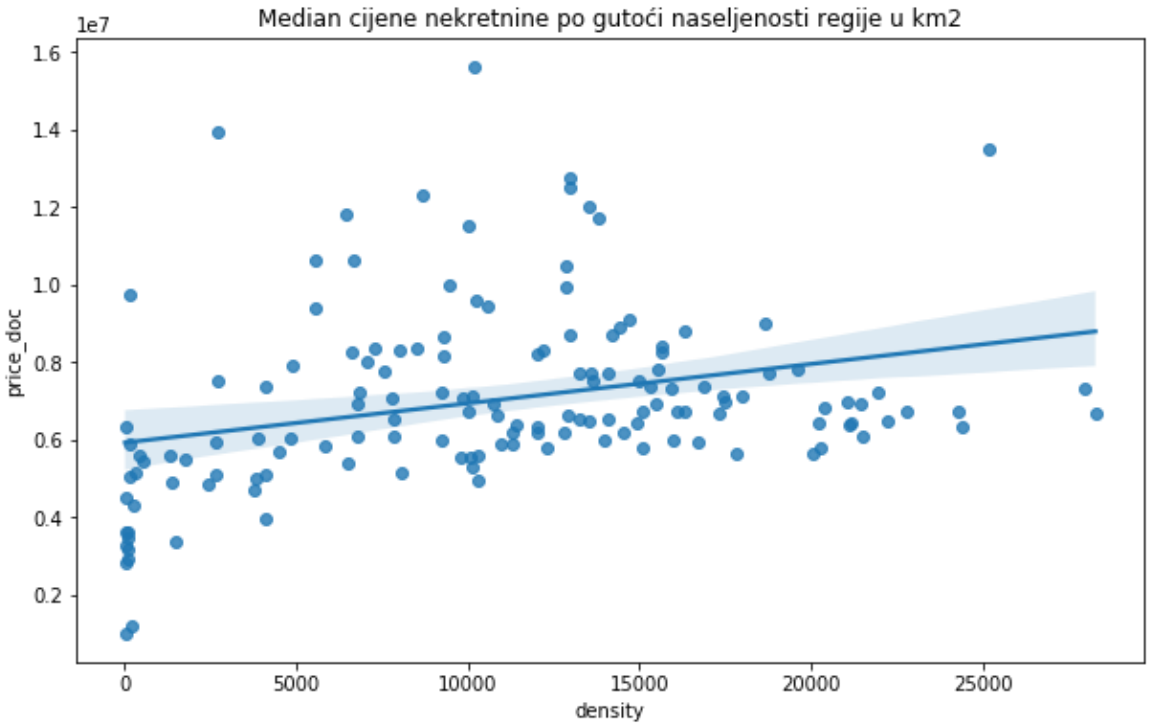
Grafikon rasporeda stanova u zgradi ovisno o katu nam pokazuje nelogičnost, jer proizlazi da ima nekretnina u kojima su stanovi na katu višem no što nekretnina ima katova. Kako vidimo iz programskog koda broj takvih nekretnina je 1493. Očito je da postoji greška prilikom zapisa podataka u bazu podataka.

4.2.4. Demografija



Grafikon 4-22 Toplinska mapa demografije

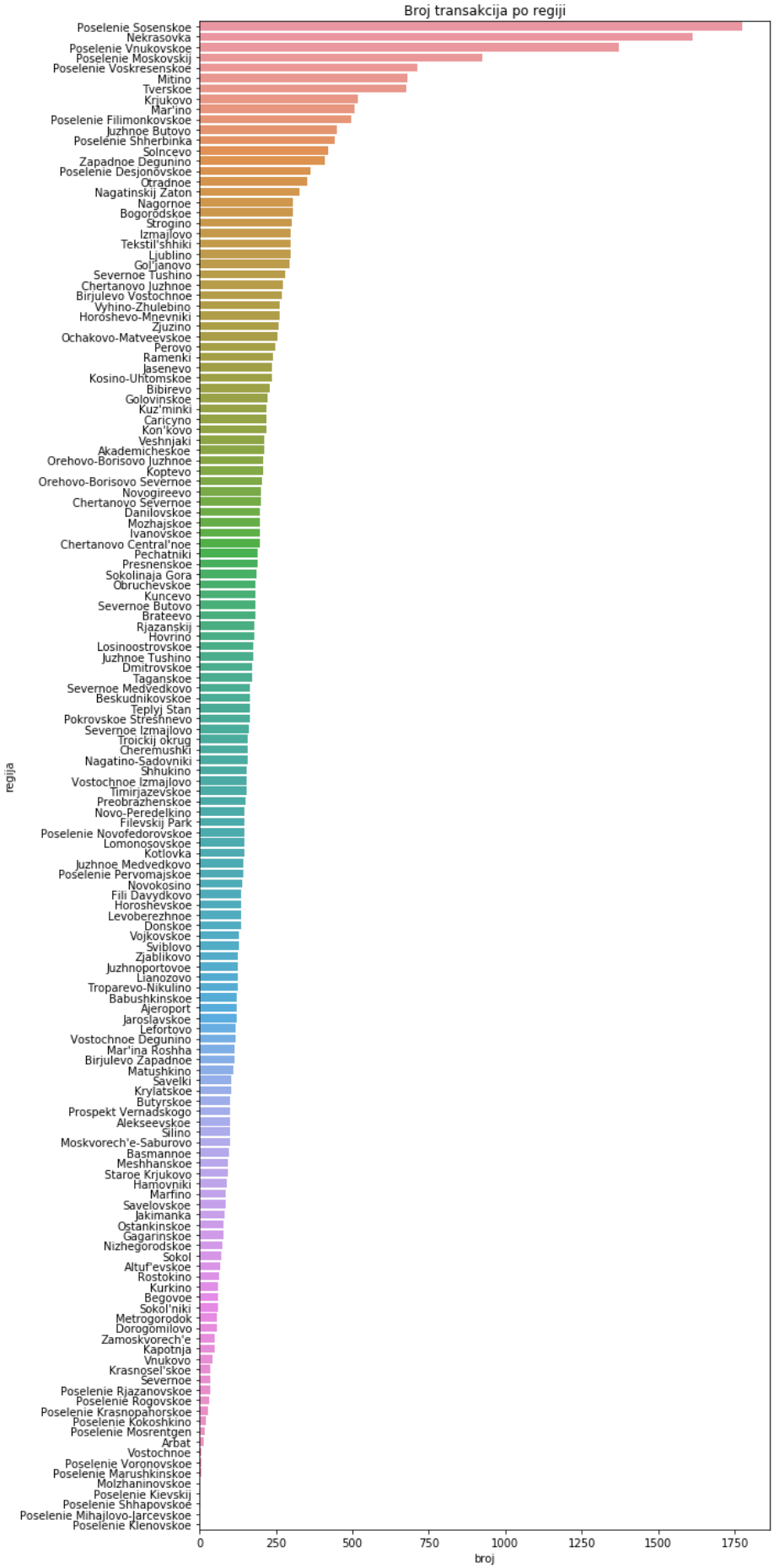
Iz toplinske mape demografije je vidljivo da je cijena nekretnina u korelaciji sa većinom veličina. Što je oznaka polja toplinske mape veća to je i veća korelacija. S obzirom na veličinu korelacija ta korelacija nije jaka. Najveća korelacija je 0.48, stoga ćemo provesti dodatnu analizu da uspostavimo pravu korelaciju između varijabli.



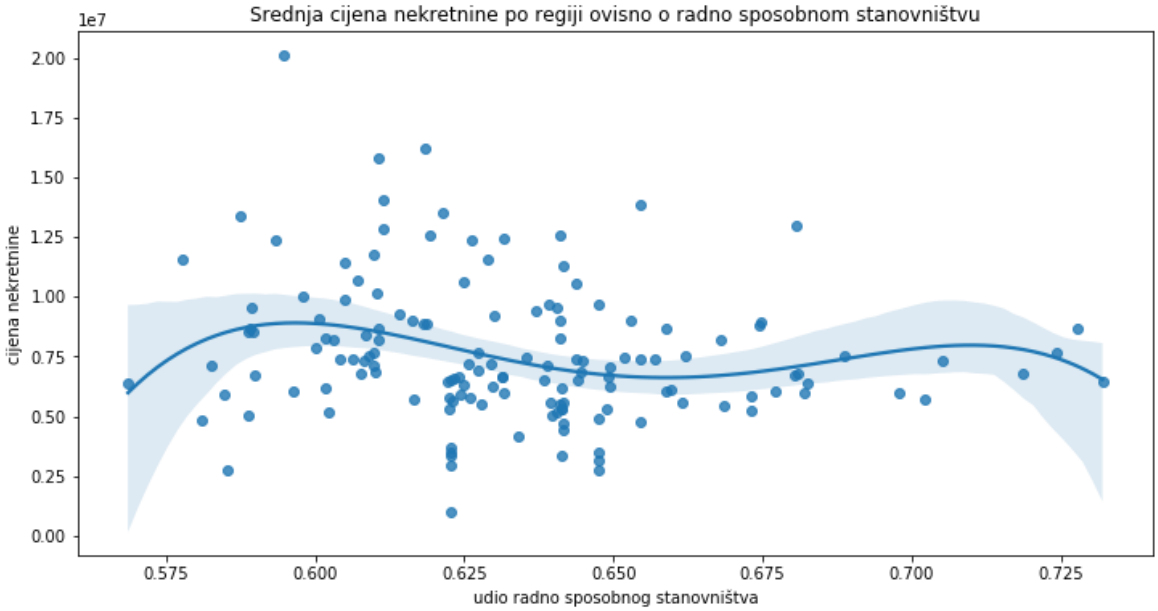
Grafikon 4-23 Medijan cijene nekretnine po gustoći naseljenosti regije u km2

Na gornjem grafikonu je vidljiv odnos gustoće naseljenosti u ovisnosti o vijene nekretnine. Površina je podijeljena sa 1000000 radi jednostavnijeg prikaza. Po nađenim podacima prosječna gustoća Moskve je 8537 po kvadratnom kilometru (Moscow Population 2019) što bi bilo približno u odnosu na prikazani grafikon. Iz podataka proizlazi da postoji regija sa gustoćom naseljenosti nula što je apsurdno. Najvažnije zaključak iz ovog grafikona je da s obzirom na gustoću naseljenosti raste i cijena nekretnine. Takav zaključak je realan u odnosu na postavke tržišne ekonomije.

Prikazati ćemo broj prodajnih transakcija nekretnina u odnosu na regiju. Poselenie Sosenskoe, Nekrasovka, Poselenie Vnukovskoe su imale najveći broj transakcija, štoviše prednjače u odnosu na druge regije.


Broj transakcija po regiji
Poselenie Sosenskoe
Nekrasovka
Poselenie Vnukovskoe
Poselenie Moskovskij
Poselenie Voskresenskoe
Mitino
Tverskoe
Krjukovo
Mar'ino
Poselenie Filimonkovskoe
Juzhnoe Butovo
Poselenie Shherbinka
Solncevo
Zapadnoe Degunino
Poselenie Desjonovskoe
Otradnoe
Nagatinskij Zaton
Nagornoe
Bogorodskoe
Strogino
Izmajlovo
Tekstil'shhiki
Ljublino
Gol'janovo
Severnoe Tushino
Chertanovo Juzhnoe
Birjulevo Vostochnoe
Vyhino-Zhulebino
Horoshevo-Mnevniki
Zjuzino
Ochakovo-Matveevskoe
Perovo
Ramenki
Jasenevo
Kosino-Uhtomskoe
Bibirevo
Golovinskoe
Kuz'minki
Caricyno
Kon'kovo
Veshnjaki
Akademicheskoe
Orehovo-Borisovo Juzhnoe
Koptevo
Orehovo-Borisovo Severnoe
Novogireevo
Chertanovo Severnoe
Danilovskoe
Mozhajskoe
Ivanovskoe
Chertanovo Central'noe
Pechatniki
Presnenskoe
Sokolinaja Gora
Obruchevskoe
Kuncevo
Severnoe Butovo
Brateevo
Rjazanskij
Hovrino
Losinoostrovskoe
Juzhnoe Tushino
Dmitrovskoe
Taganskoe
Severnoe Medvedkovo
Beskudnikovskoe
Teplyj Stan
Pokrovskoe Streshnevo
Severnoe Izmajlovo
Troickij okrug
Cheremushki
Nagatino-Sadovniki
Shhukino
Vostochnoe Izmajlovo
Timirjazevskoe
Preobrazhenskoe
Novo-Peredelkino
Filevskij Park
Poselenie Novofedorovskoe
Lomonosovskoe
Kotlovka
Juzhnoe Medvedkovo
Poselenie Pervomajskoe
Novokosino
Fili Davydkovo
Horoshevskoe
Levoberezhnoe
Donskoe
Vojkovskoe
Sviblovo
Ziablikovo
Juzhnoportovoe
Lianozovo
Troparevo-Nikulino
Babushkinskoe
Ajeroport
Jaroslavskoe
Lefortovo
Vostochnoe Degunino
Mar'ina Roshha
Birjulevo Zapadnoe
Matushkino
Savelki
Krylatskoe
Butyrskoe
Prospekt Vernadskogo
Alekseevskoe
Silino
Moskvorech'e-Saburovo
Basmannoe
Meshhanskoe
Staroe Krjukovo
Hamovniki
Marfino
Savelovskoe
Jakimanka
Ostankinskoe
Gagarinskoe
Nizhegorodskoe
Sokol
Altuf'evskoe
Rostokino
Kurkino
Begovoe
Sokol'niki
Metrogorodok
Dorogomilovo
Zamoskvorech'e
Kapotnja
Vnukovo
Krasnosel'skoe
Severnoe
Poselenie Rjazanovskoe
Poselenie Rogovskoe
Poselenie Krasnopahorskoe
Poselenie Kokoshkino
Poselenie Mosrentgen
Arbat
Vostochnoe
Poselenie Voronovskoe
Poselenie Marushkinskoe
Molzhaninovskoe
Poselenie Kievskij
Poselenie Shhapovskoe
Poselenie Mihajlovo-Jarcevskoe
Poselenie Klenovskoe
regija
0
250
500
750
1000
1250
1500
1750
broj

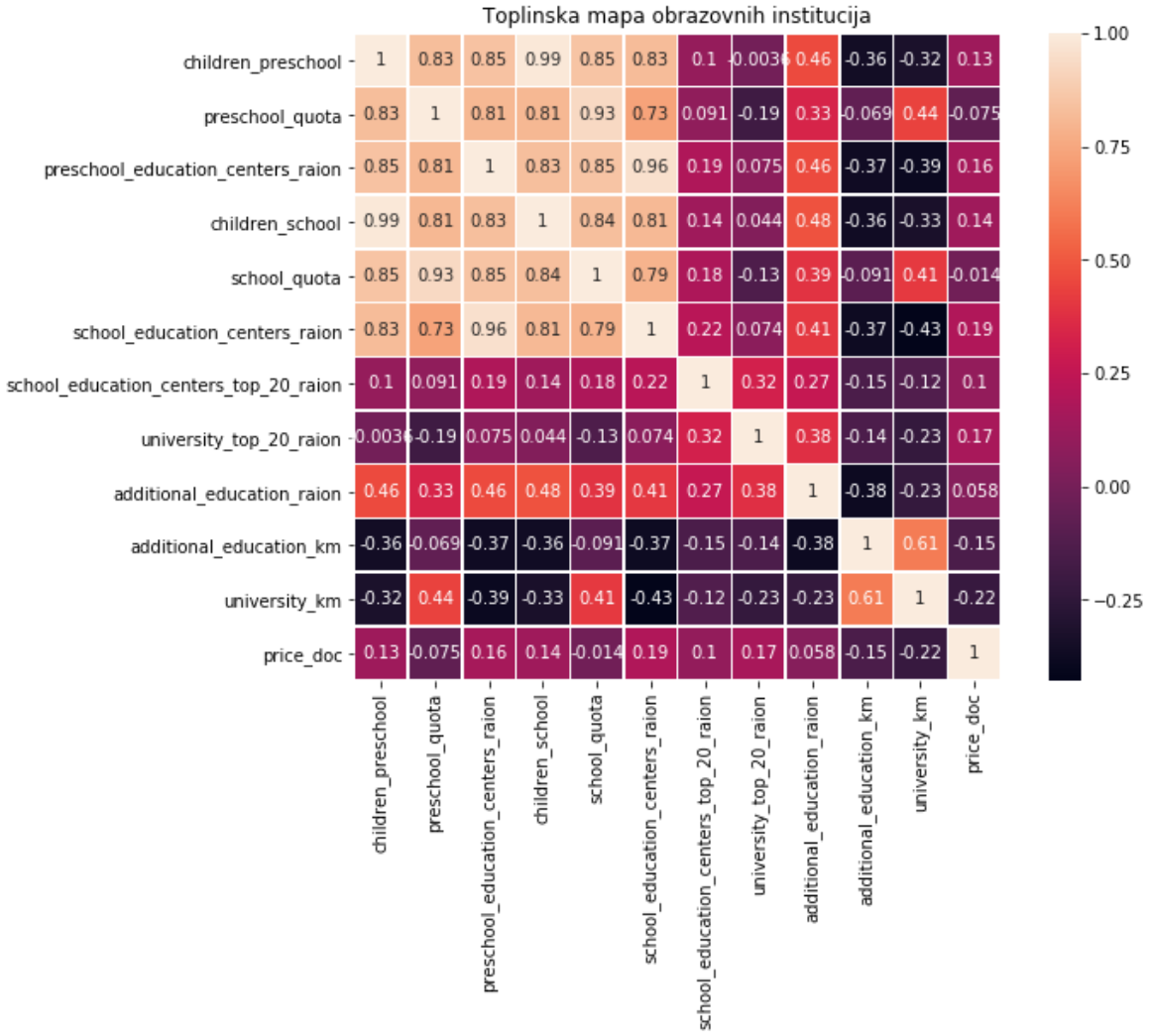
Grafikon 4-24 Broj transakcija po regiji



Grafikon 4-25 Srednja cijena nekretnine po regiji ovisno o radno sposobnom stanovništvu

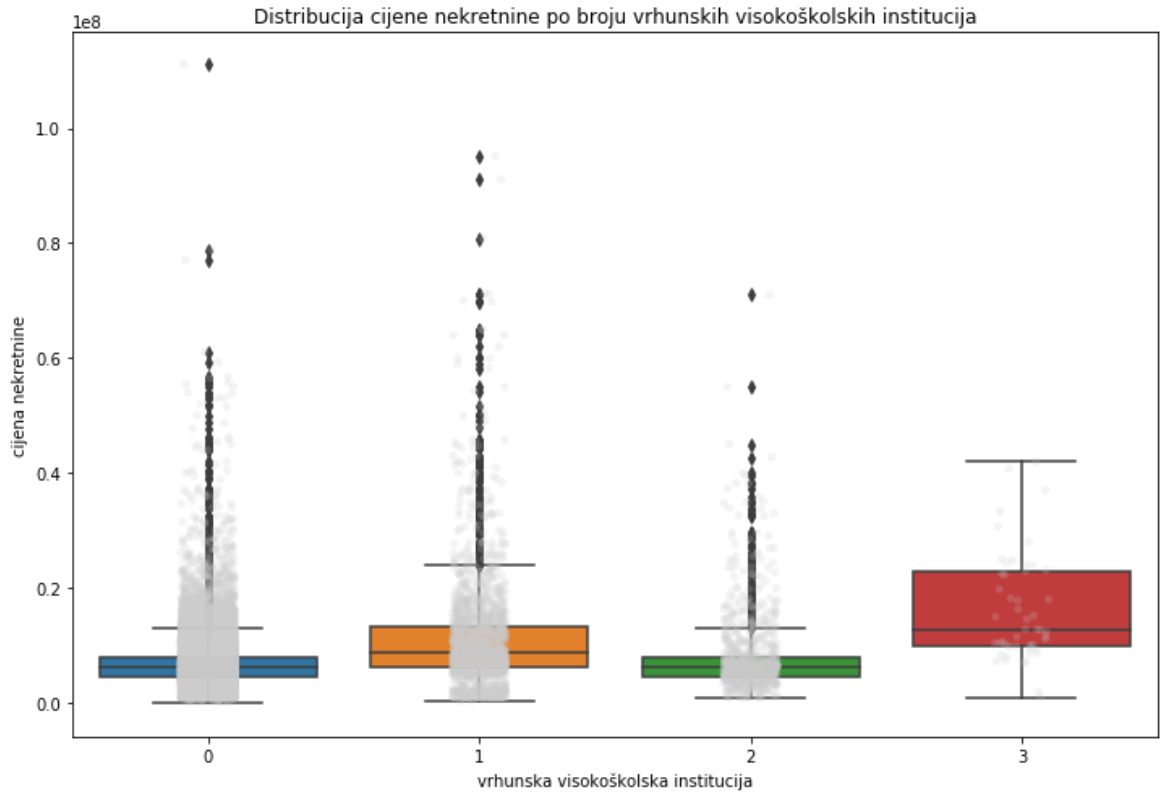
S obzirom da na grafikonu nema oscilacija, zaključujemo da je srednja cijena nekretnina neovisna o udjelu radno sposobnog stanovništva. Iz toga proizlazi da bez obzira na primanja pojedine regije one nisu u korelaciji sa cijenom nekretnine.

4.2.5. Obrazovne institucije



Grafikon 4-26 Toplinska mapa obrazovnih institucija

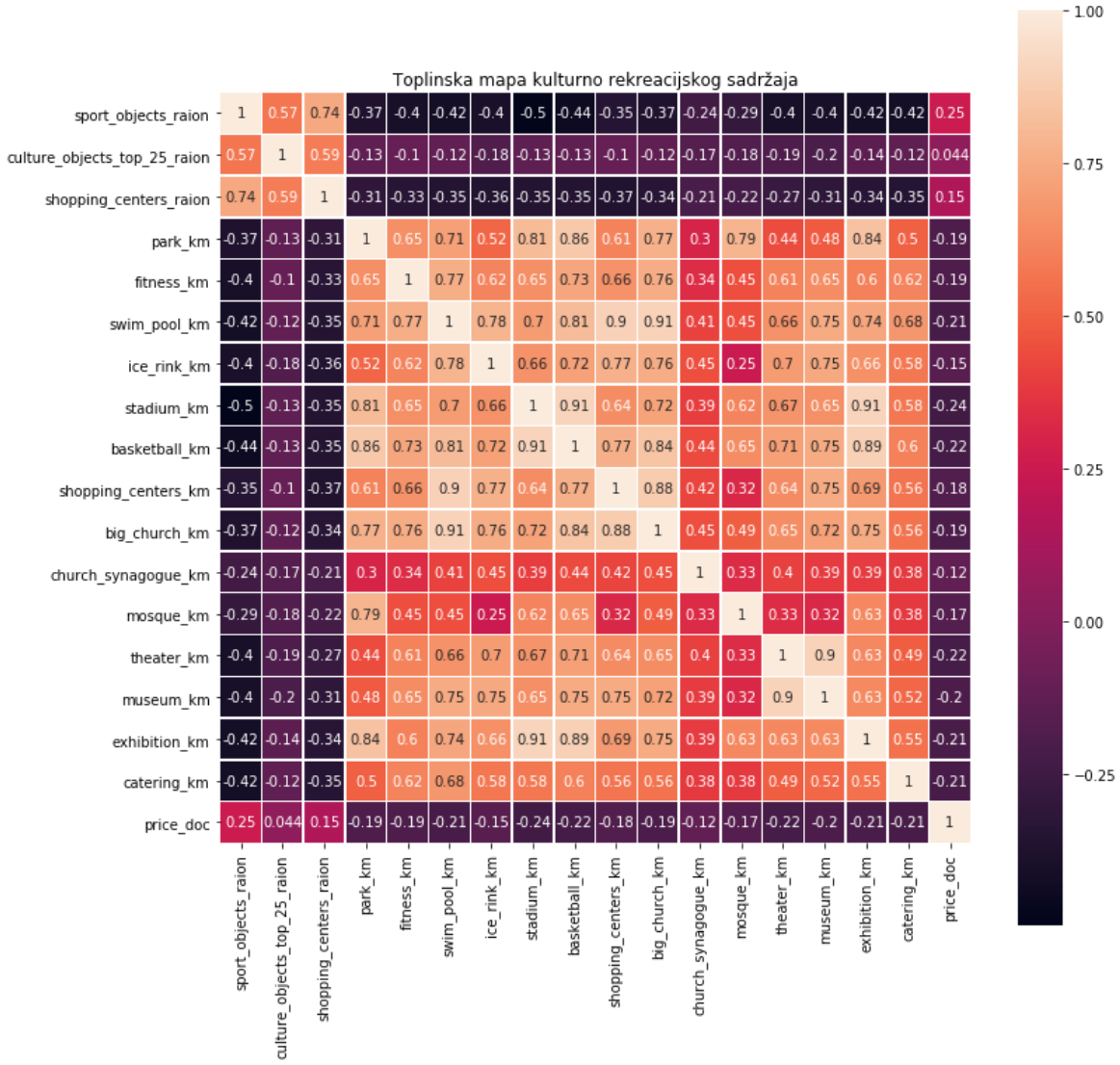
Slično kao i kod toplinske mape demografije niti ovdje se ne vidi velika korelacija između varijabli, osim školskih obrazovnih institucija međusobno. To nam je znak da pri modeliranju je dovoljno koristiti samo jednu od njih.



Grafikon 4-27 Distribucija cijene nekretnine po broju vrhunskih visokoškolskih institucija

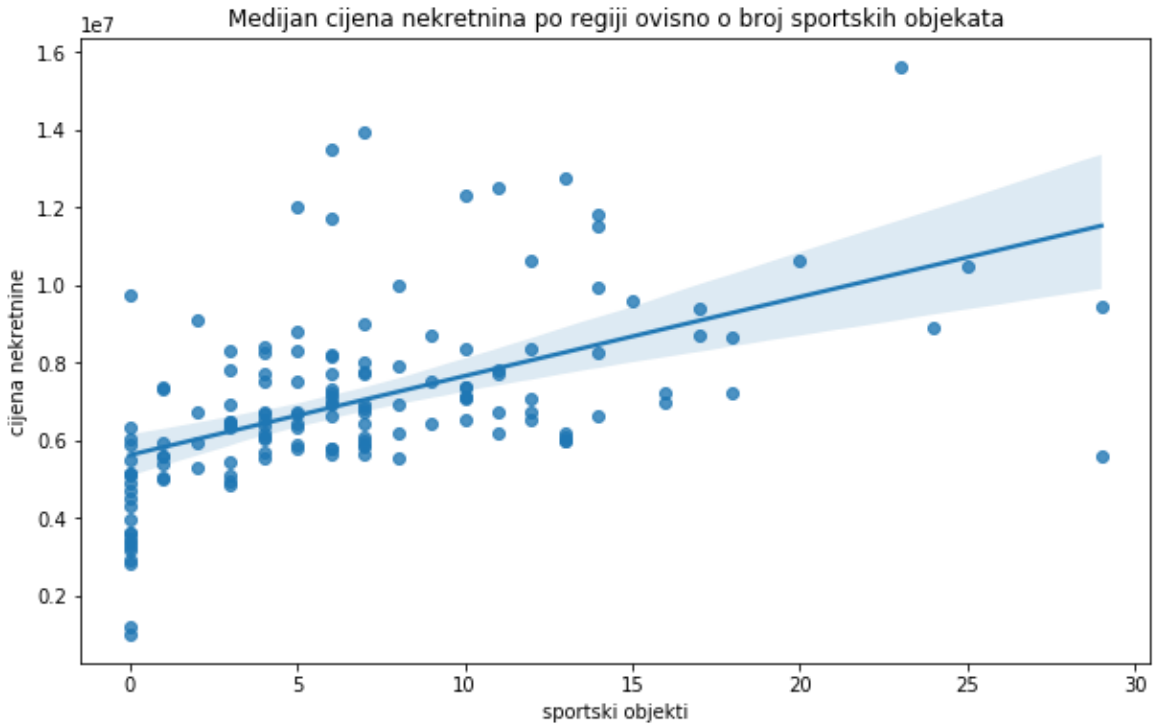
Kako je korelacija školskih obrazovnih institucija velika provjeriti ćemo koliko na cijenu nekretnina utječe blizina broja vrhunskih visokoškolskih institucija. Kada je broj vrhunskih visokoškolskih institucija 3 cijena je najveća. Programskim kodom je utvrđeno da broj regija koji imaju pristup tri vrhunske visokoškolske institucije je samo jedna i to Zamoskvorech'e. Kako je vidljivo i cijena nekretnine je najviša u toj regiji.

4.2.6. Kulturno rekreacijski sadržaj



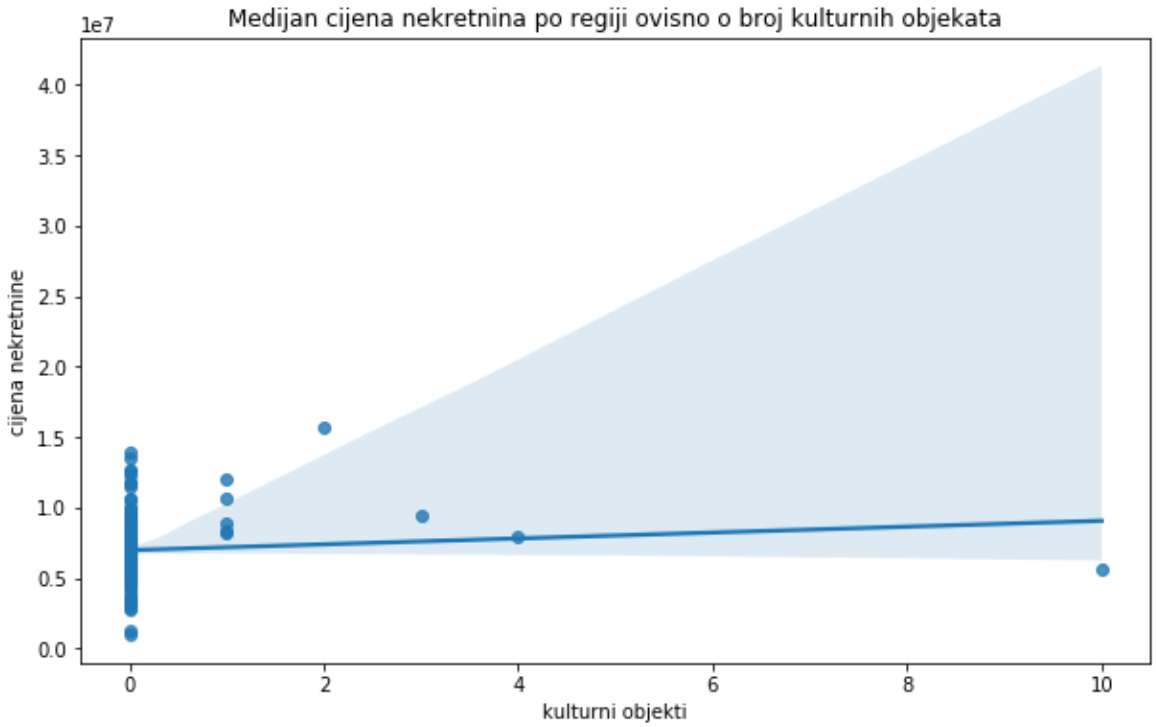
Grafikon 4-28 Toplinska mapa kulturno rekreacijskog sadržaja

Na toplinskoj mapi kulturno rekreacijskog sadržaja je vidljiva korelacija sportskog sadržaja. Zapanjuje negativna korelacija cijene nekretnine i većine varijabli. U nastavku ćemo pogledati odnos cijene nekretnine u odnosu na sportskih i kulturnih objekata te parkova..



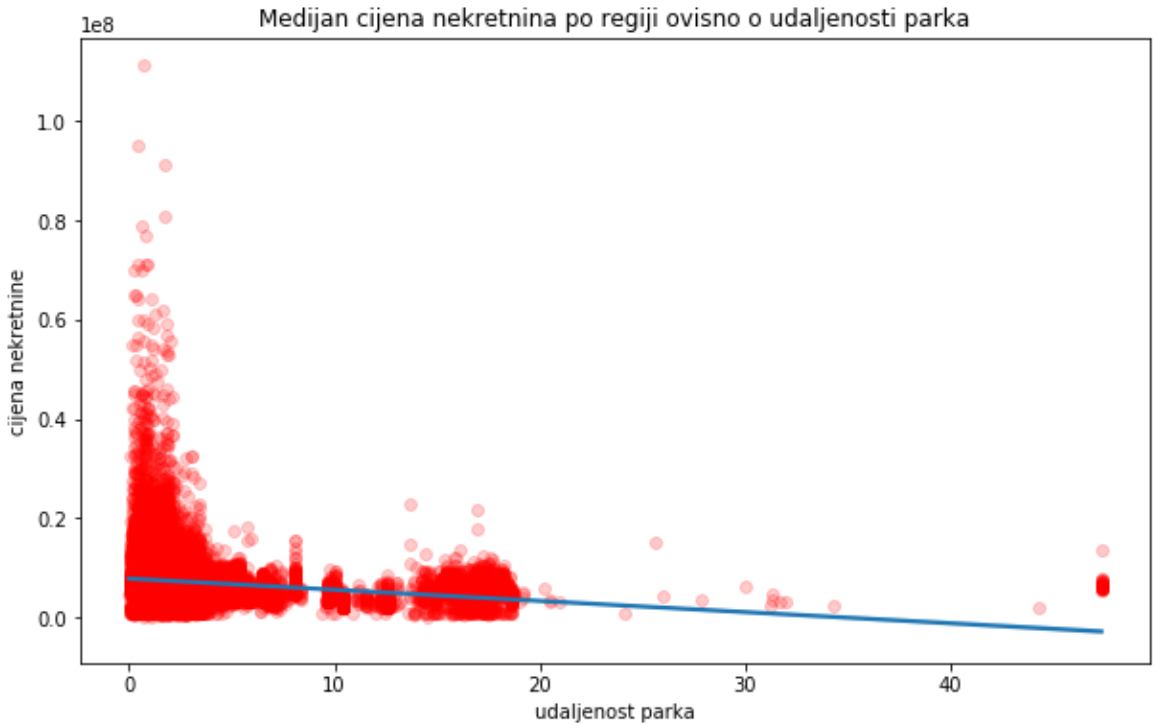
Grafikon 4-29 Medijan cijena nekretnina po regiji ovisno o broj sportskih objekata

Iz grafikona korelacije cijene nekretnina i broja sportskih objekata postoji vidljivo povećanje cijene nekretnina s obzirom na broj sportskih objekata. Utjecaj sportskih objekata je očito varijabla koja bi trebala biti dobar pokazatelj cijene nekretnina. Imat ćemo to na umu pri modeliranju algoritma za predviđanje cijena nekretnina.



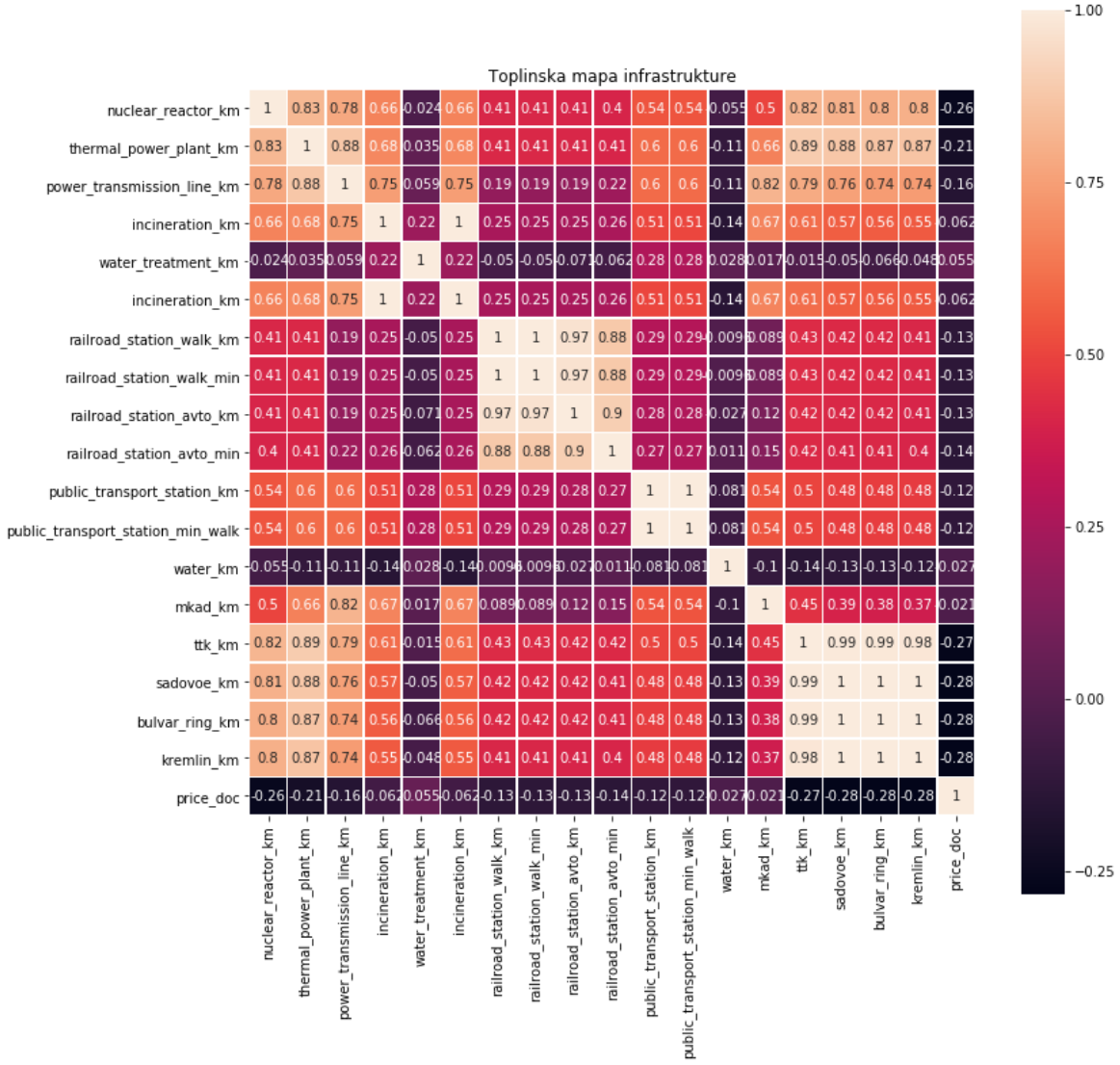
Grafikon 4-30 Medijan cijena nekretnina po regiji ovisno o broj kulturnih objekata

U odnosu na sportske objekte, utjecaj kulturnih objekata na cijenu nekretnina je nebitan zbog nepostojanja kulturnih objekata u blizini nekretnina. Kulturni objekti nam stoga neće biti potrebni u daljnjoj analizi.



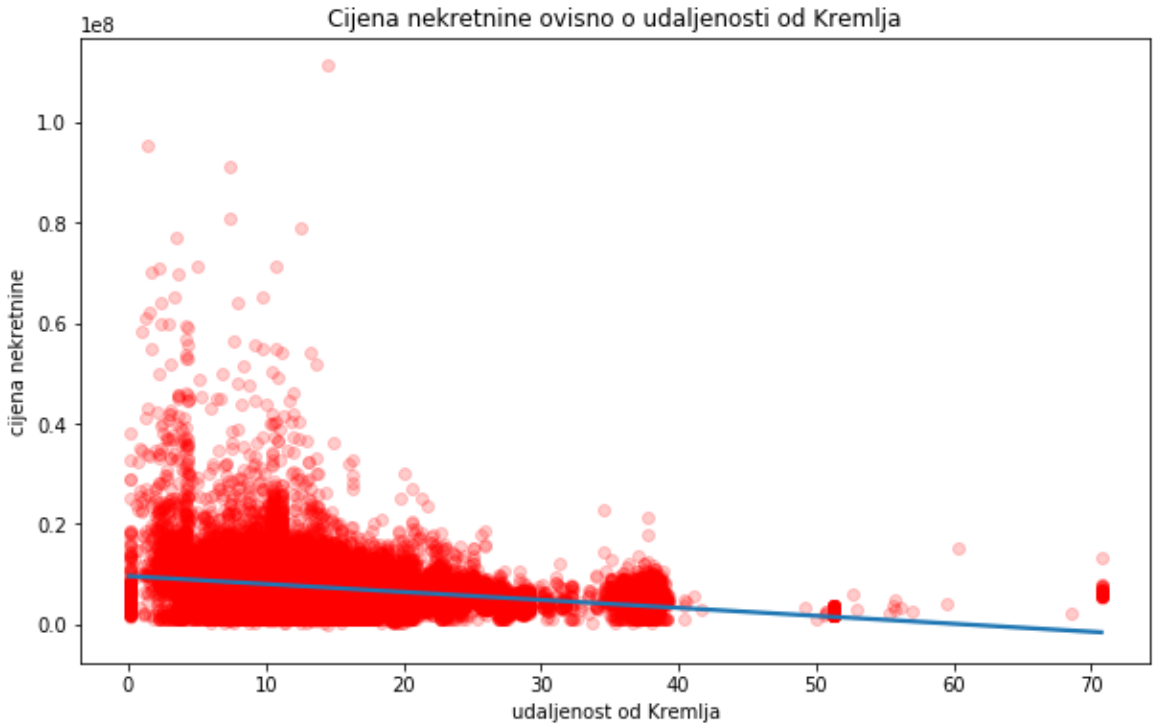
Grafikon 4-31 Medijan cijena nekretnina po regiji ovisno o udaljenosti parka

4.2.7. Infrastruktura



Kod 4-1 Toplinska mapa infrastrukture

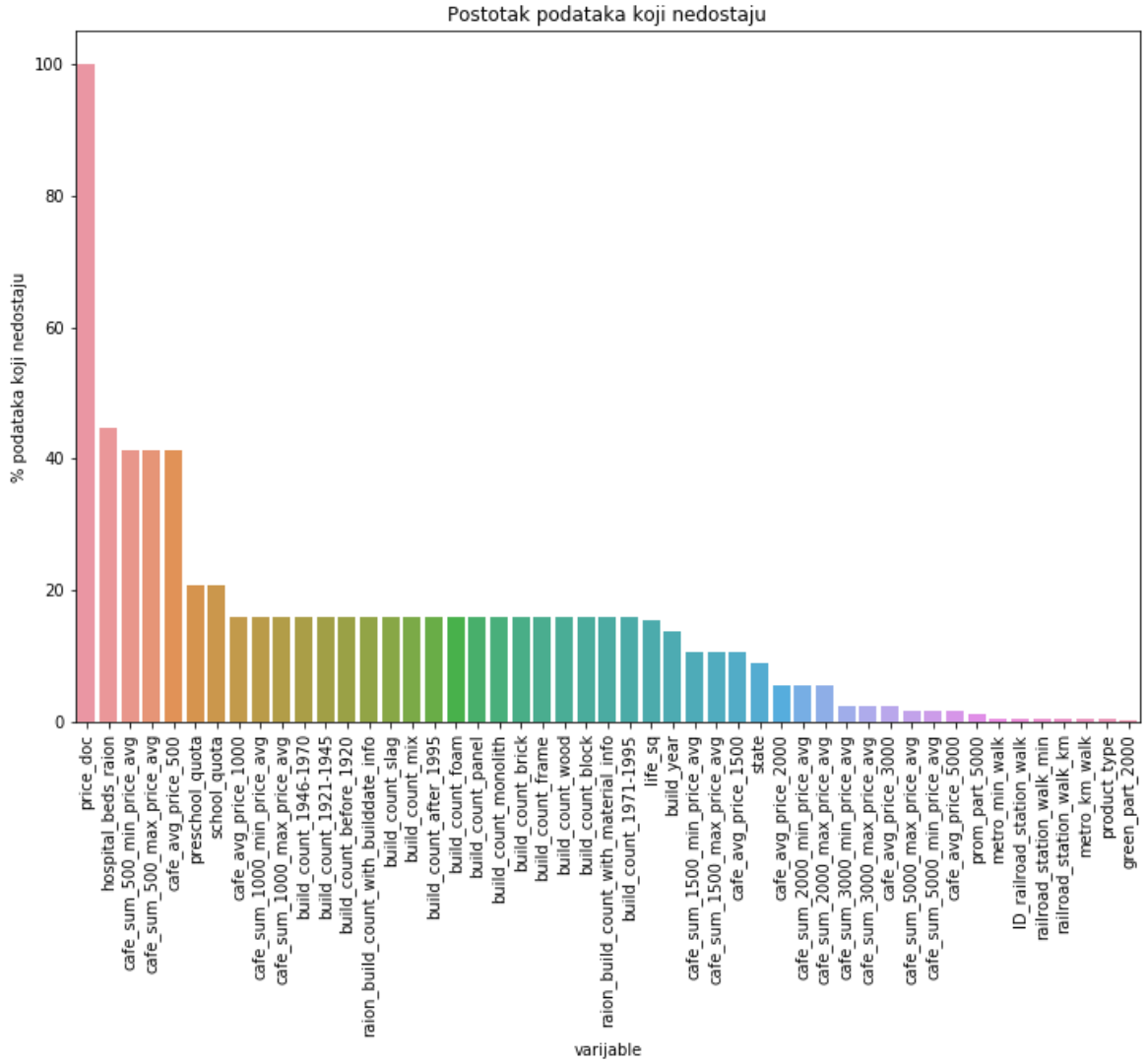
Toplinska mapa infrastrukture pokazuje najveću korelaciju između istog tipa infrastrukture, u ovom slučaju energetskih postrojenja, dok je negativna korelacija infrastrukture i cijene nekretnine. Pod pretpostavkom da će cijena nekretnine biti veća u blizini prijevoznog sredstva, ovom slučaju stanica javnog prijevoza, a manja s obzirom na udaljenost on energetskog postrojenja fokusirati ćemo se na druge varijable.



Grafikon 4-32 Cijena nekretnine ovisno o udaljenosti od Kremlja

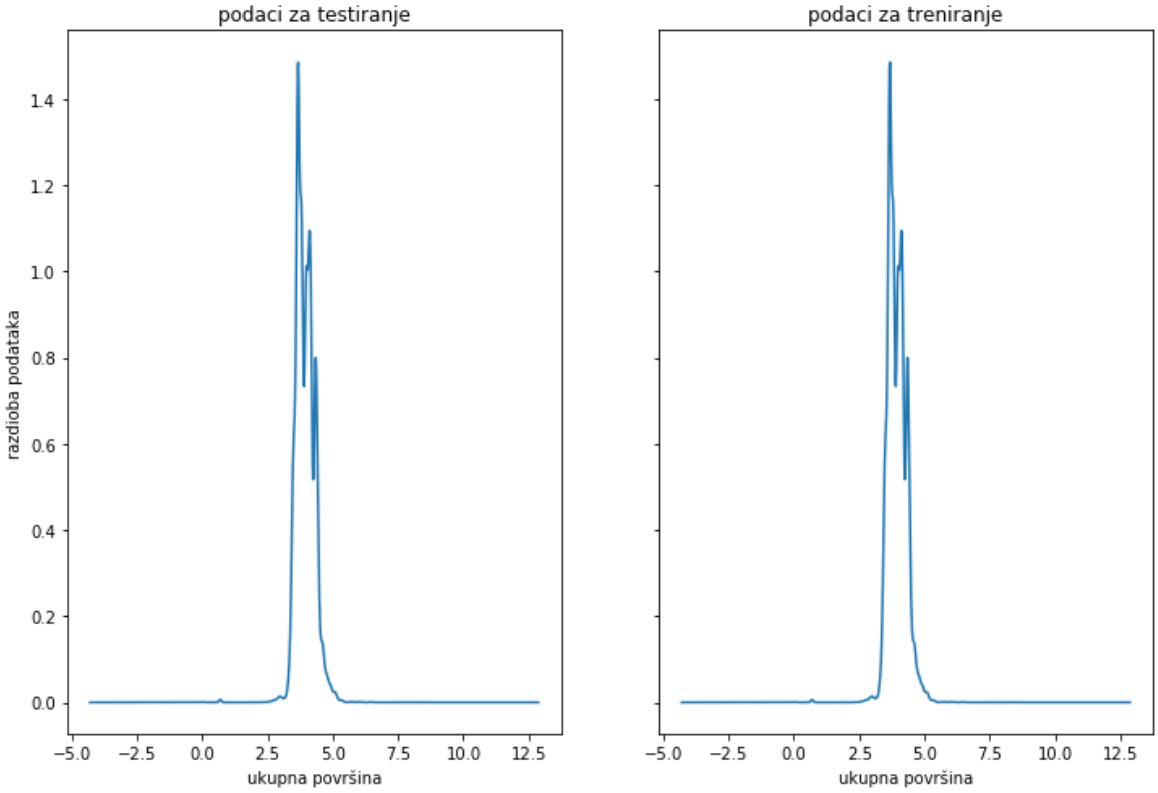
Grafikon ovisnosti cijene nekretnine od udaljenosti od Kremlja jednoznačno prikazuje veću cijenu što je Kremlj bliži. Takva korelacija je bila i očekivana, no osim toga i prikazuje da je veći broj nekretnina ipak bliži Kremlju nego dalji. Nekretnine koje bi valjalo uzeti u obzir su one do najveće udaljenosti od 40 kilometara. Izvan toga radijusa ne postoji dovoljan broj nekretnina da bi ušao u predviđanje cijena nekretnina.

4.2.8. Usporedba testnih podataka i podataka za treniranje

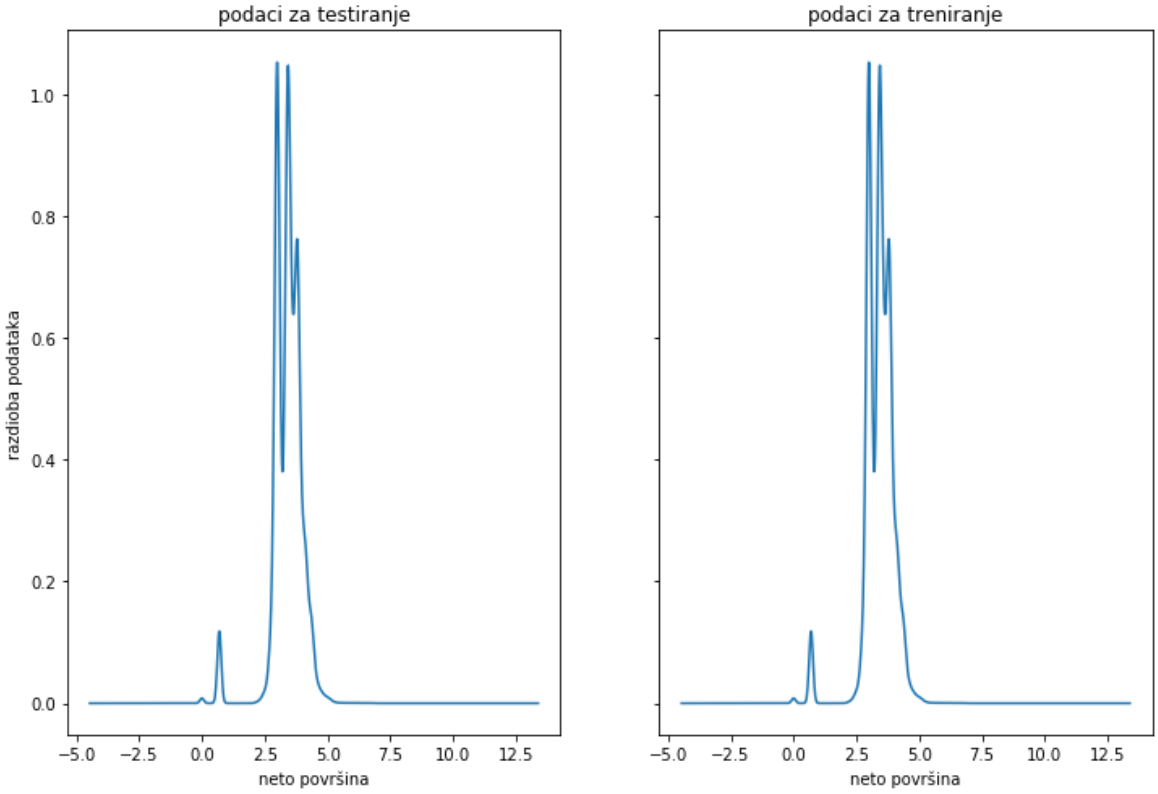


Grafikon 4-33 Postotak podataka koji nedostaju

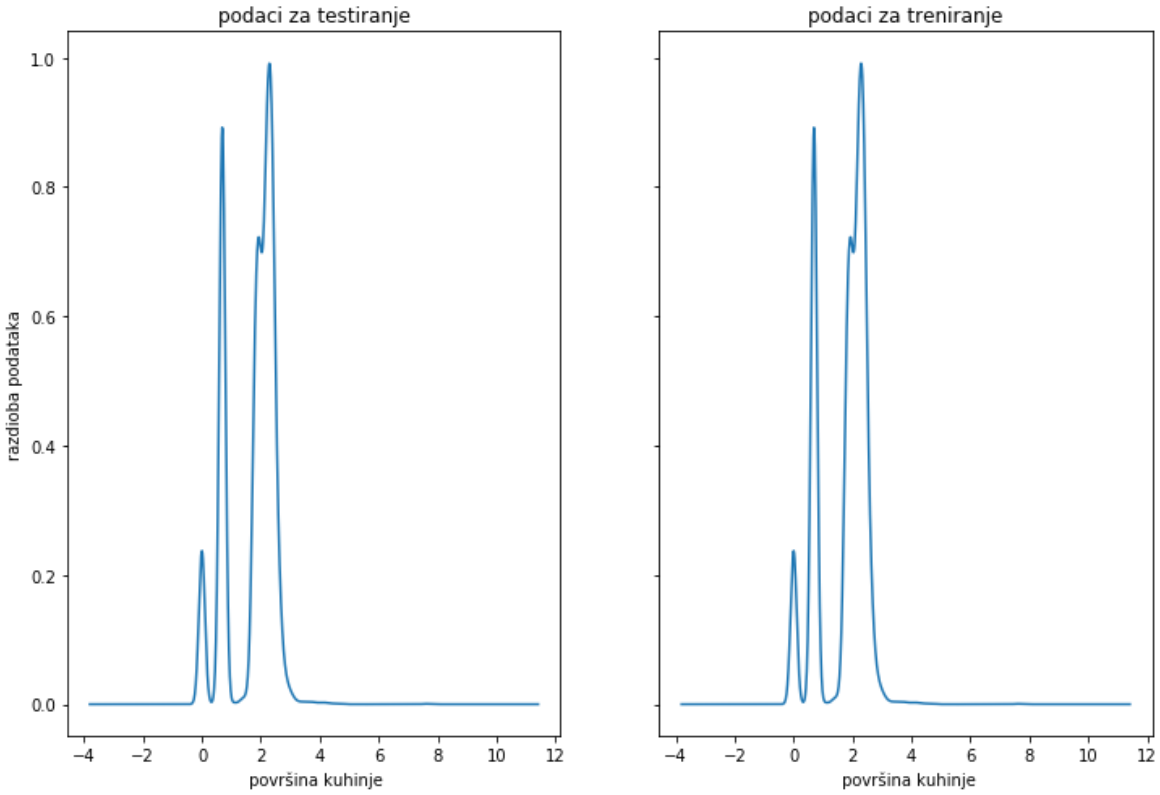
Za početak trebamo vidjeti strukturu nedostajućih podataka podataka za treniranje. Grafikon pokazuje da se to najviše odnosi na varijable vezane uz blizinu ugostiteljskih objekata. Do sada napravljena analiza nije pridavala veliku važnost tim varijablama pa u cjelokupnom predviđanju niti nemaju zamjetan utjecaj,



Grafikon 4-34 Usporedba po ukupnoj površini

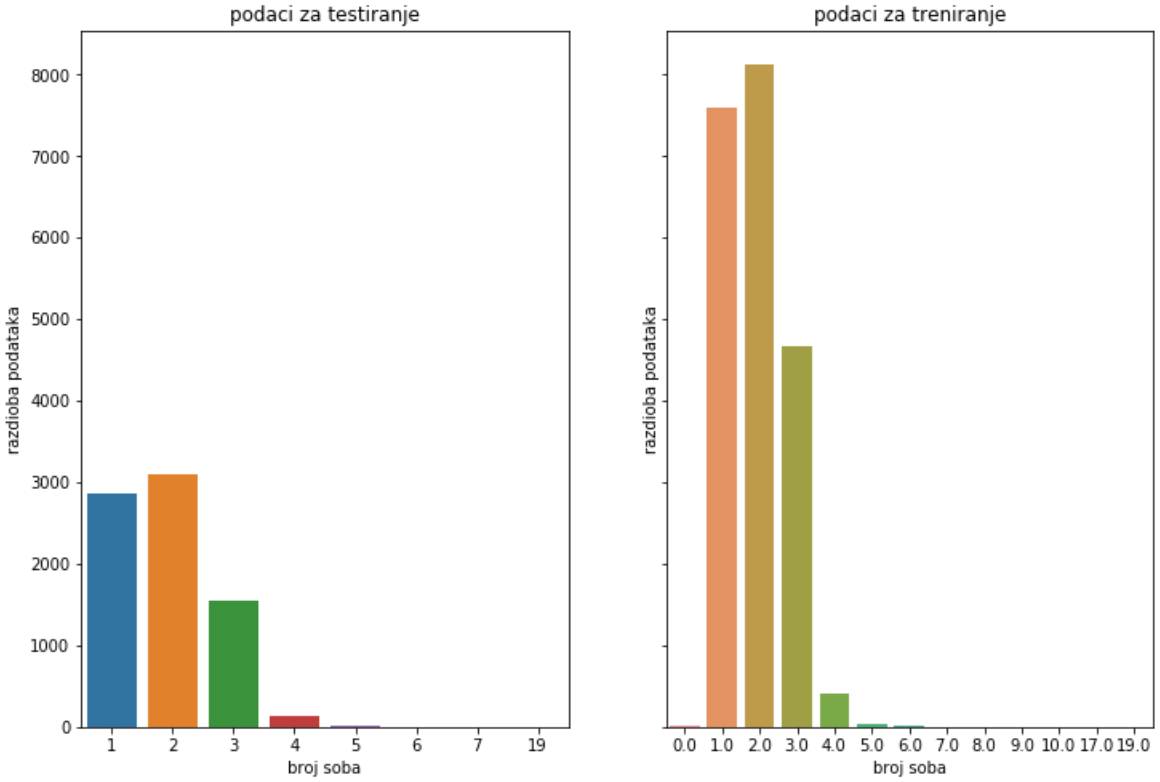


Grafikon 4-35 Usporedba po neto površini

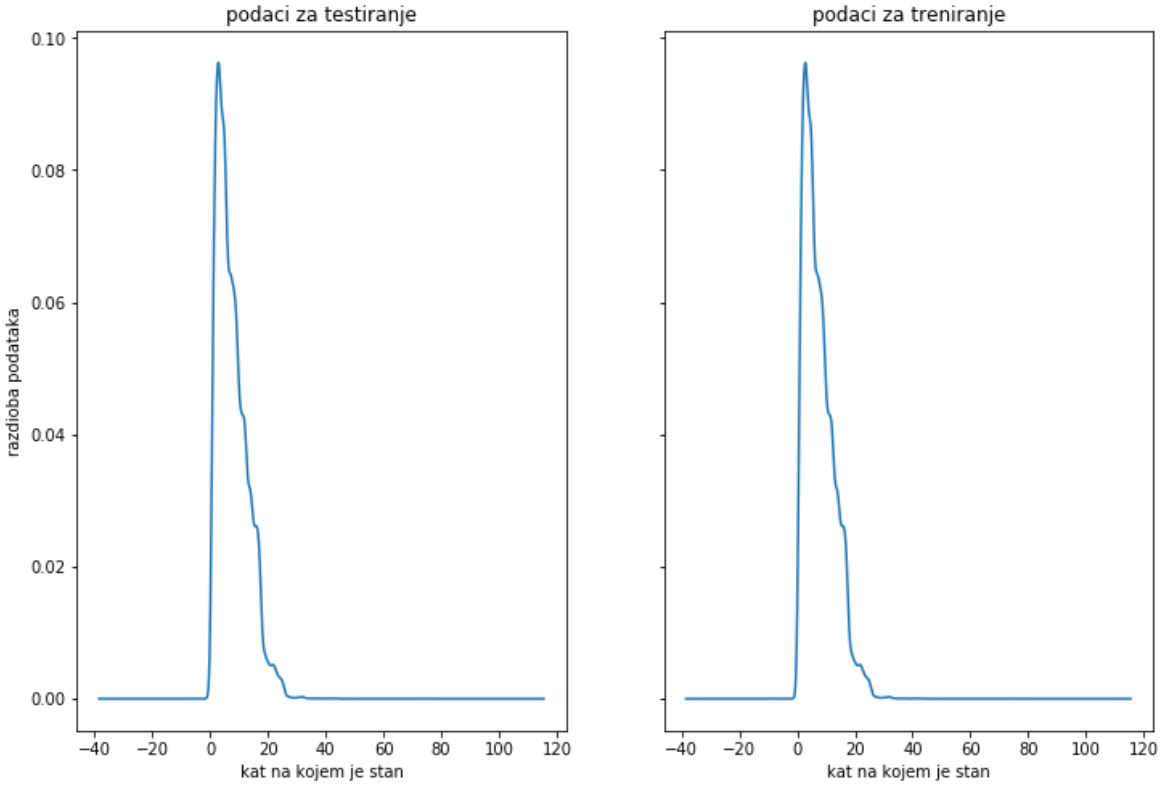


Grafikon 4-36 Usporedba po površini kuhinje

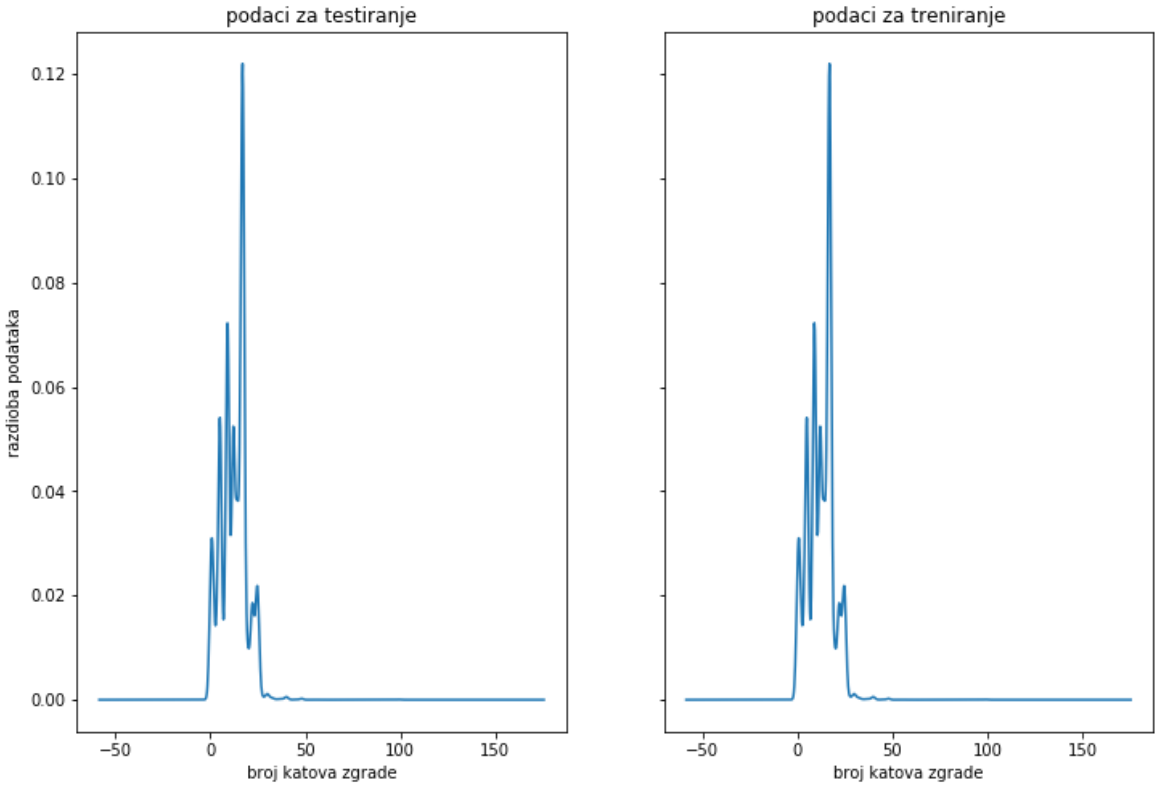
Usporedbom podataka za treniranje i podataka za testiranje vidljivo je da je njihova distribucija slična jedna drugoj za karakteristike vezane uz površinu nekretnina. Sličnost je vidljiva u distribuciji i ukupne i neto površine nekretnina te u površini kuhinje. U programskom kodu je pokazano da su to i tri najutjecajnije varijable. Ravnomjerna distribucija će nam pomoći pri predviđanju cijene nekretnina. U daljnjem koraku vidljiva je razlika u distribuciji broja soba u podacima treniranje i podacima za testiranje. Svaki od narednih grafikona radi usporedbu podataka za treniranje i podataka za treniranje dajući nam informaciju s kojim ćemo se problemima možda susretati u predviđanju cijene nekretnina. U biti nam ta usporedba stavlja u perspektivu struktura samih podataka.



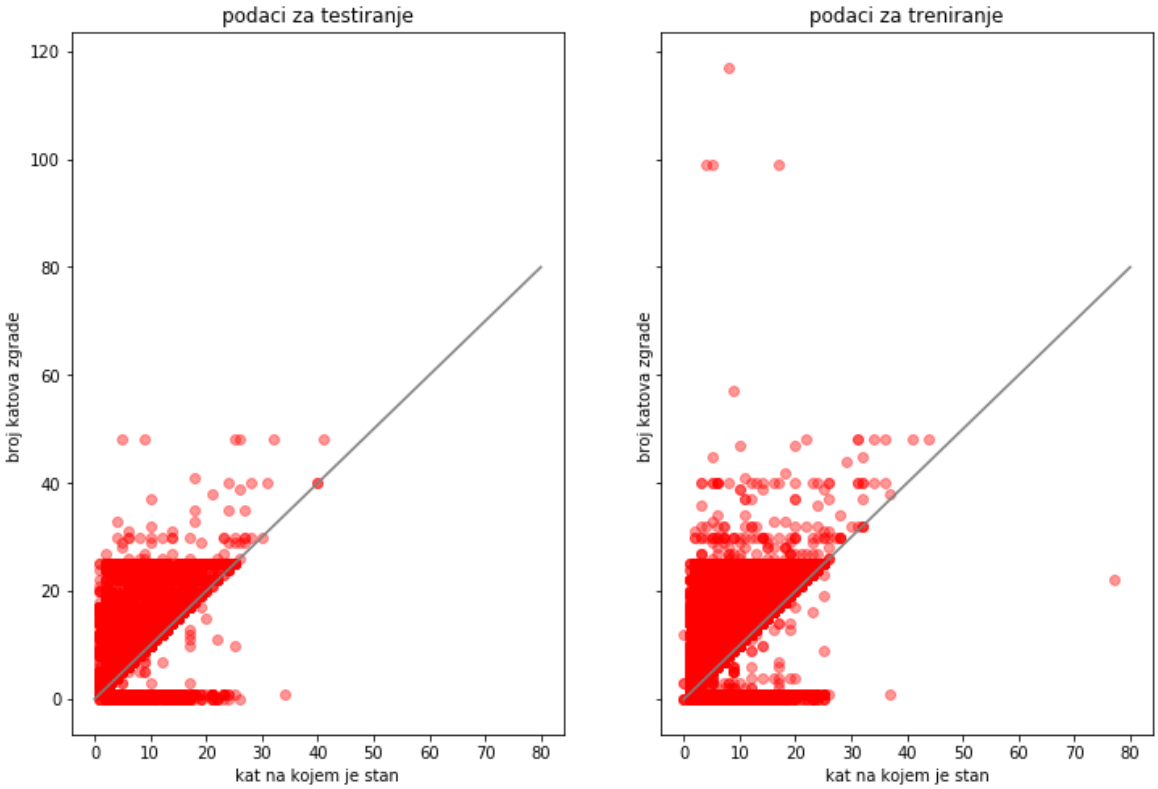
Grafikon 4-37 Usporedba po broju soba



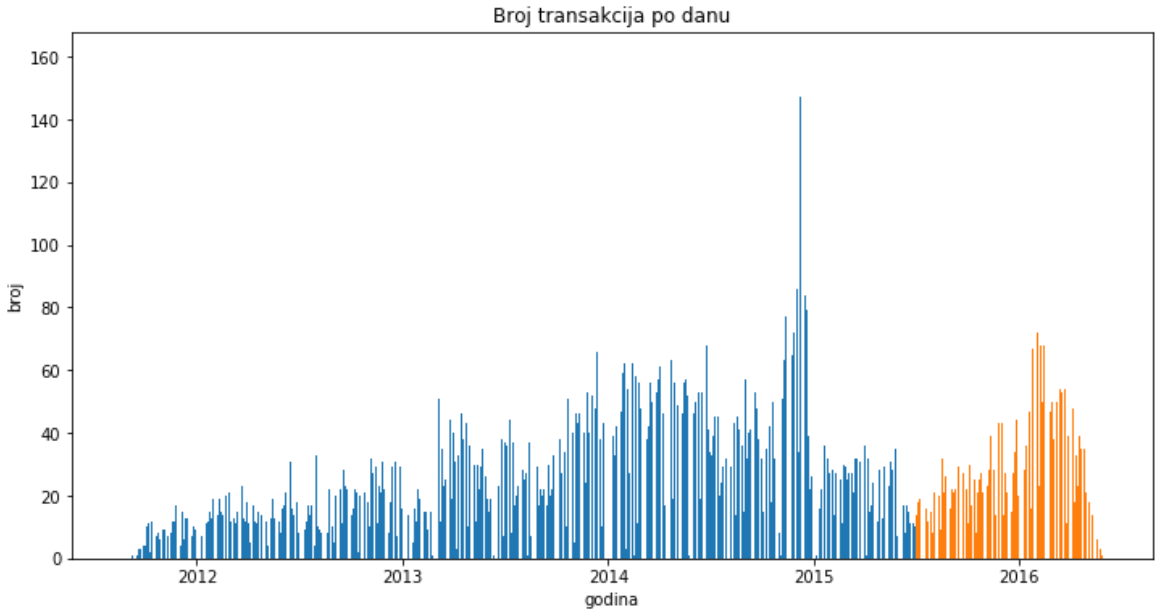
Grafikon 4-38 Usporedba ovisno na kojem se katu nalazi stan



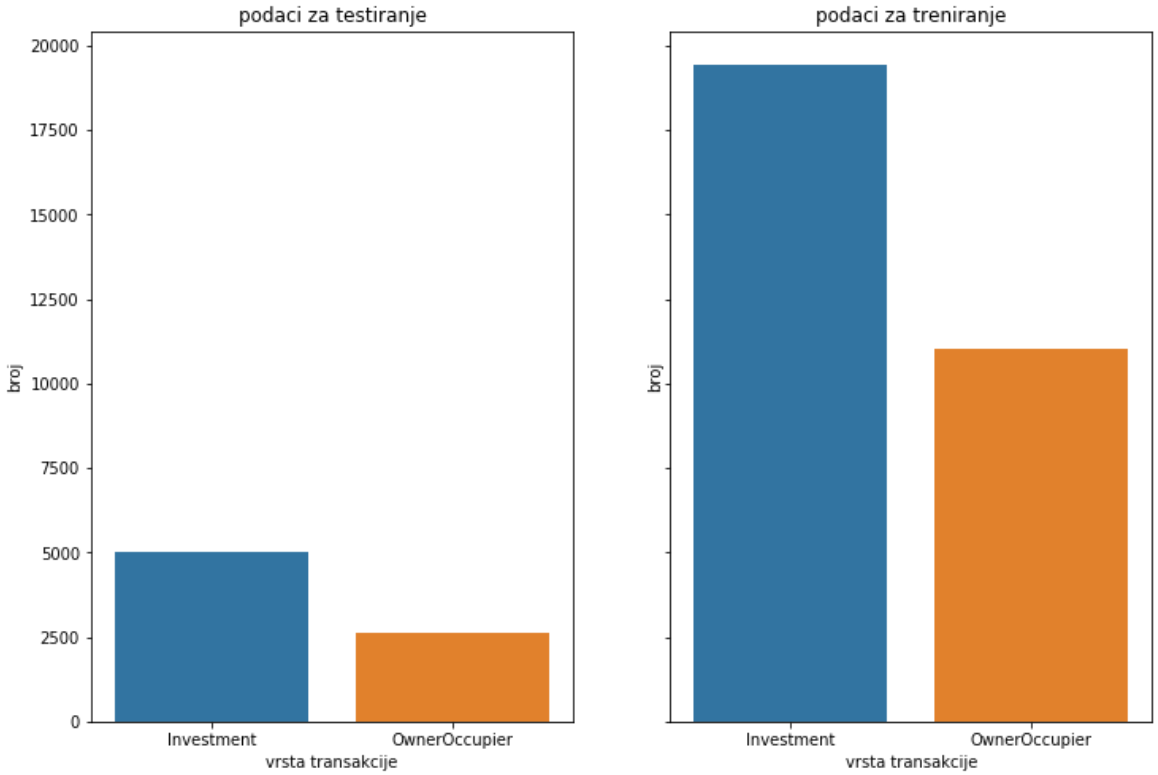
Grafikon 4-39 Usporedba ovisno o broju katova u zgradi



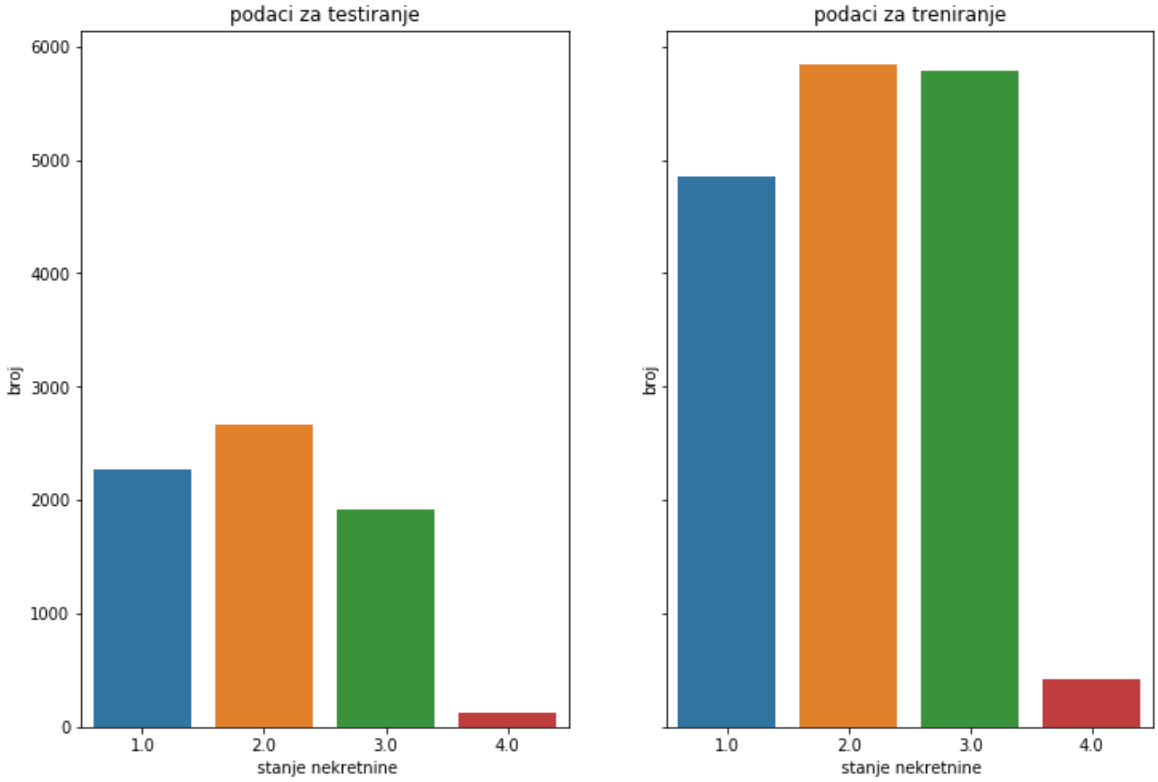
Grafikon 4-40 Usporedba ovisno o katu na kojem je stan i broju katova u zgradi



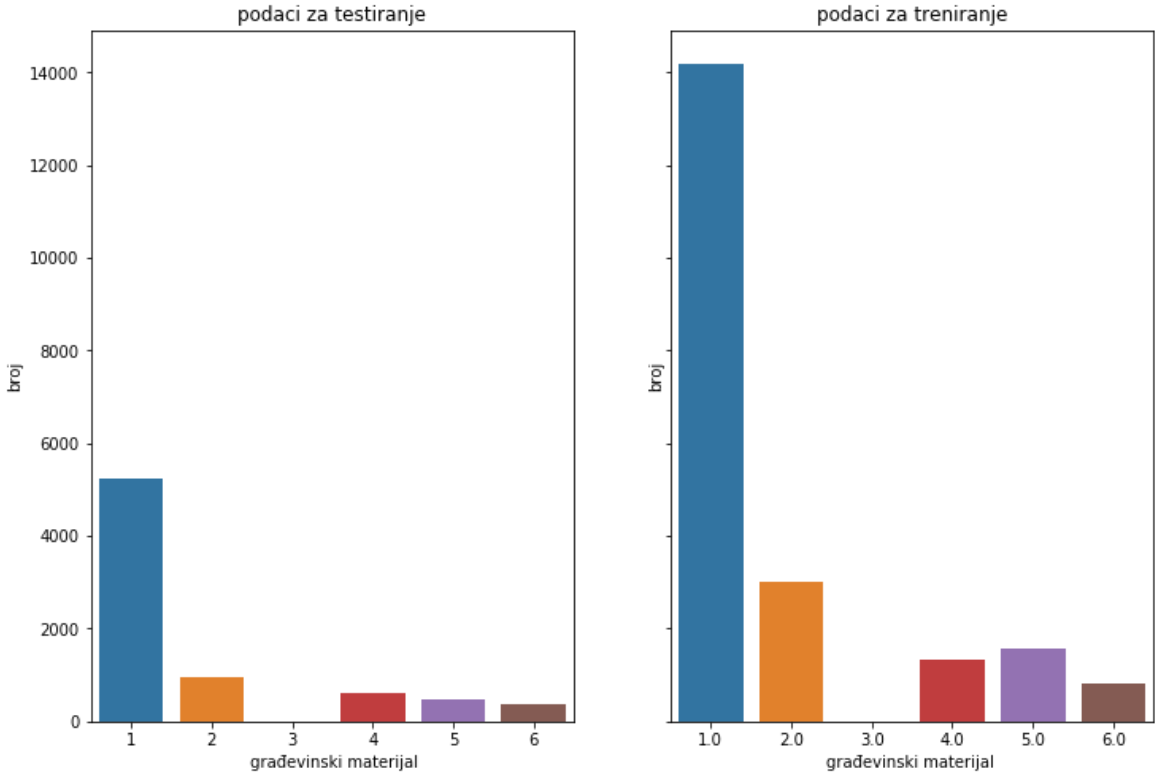
Grafikon 4-41 Broj transakcija po danu



Grafikon 4-42 Usporedba po vrsti transakcije



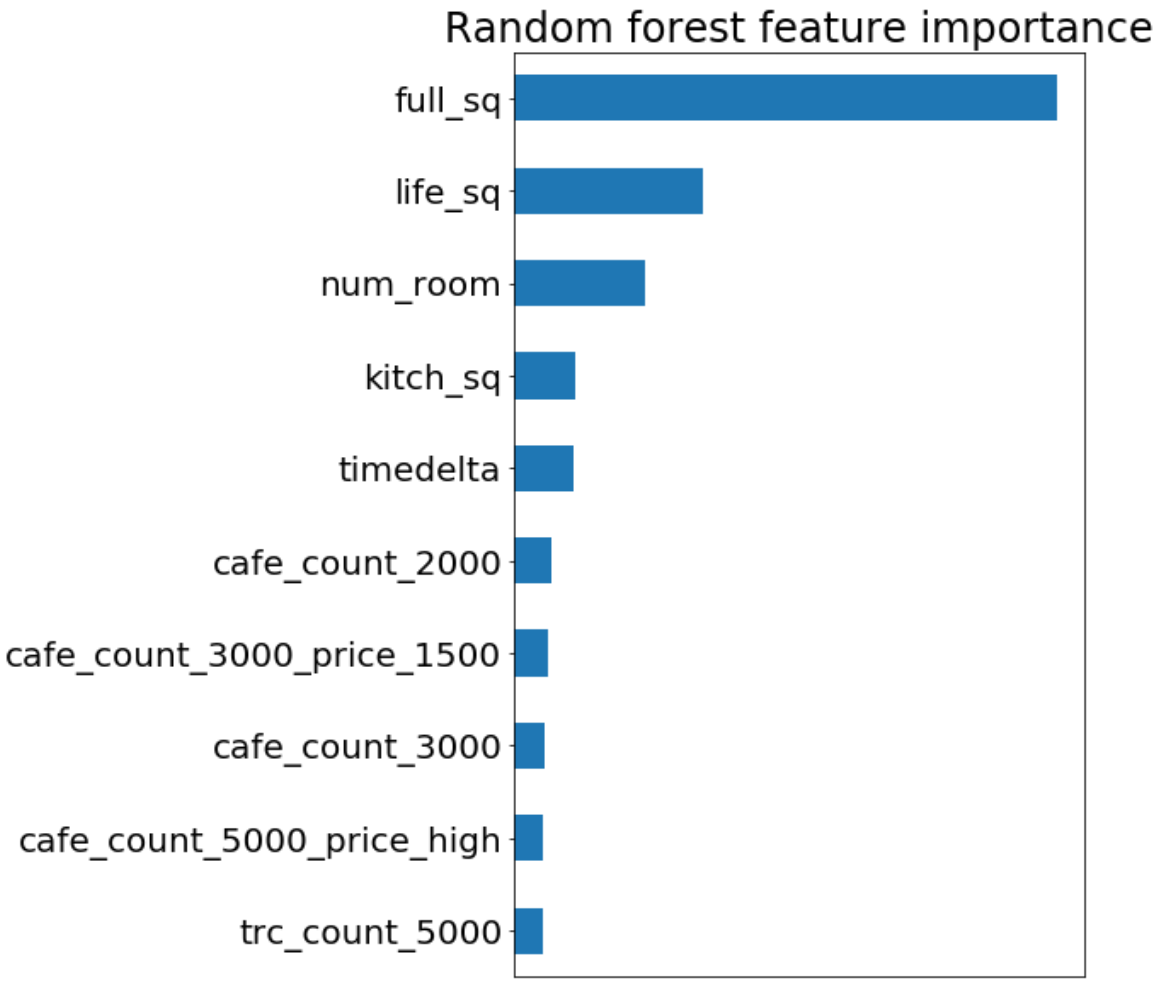
Grafikon 4-43 Usporedba ovisno o stanju nekretnine



Grafikon 4-44 Usporedba ovisno o građevinskom materijalu

4.3. Predviđanje cijena nekretnina

Prije nego što krenemo sa predviđanjem cijene nekretnina trebamo da utvrdimo koje su najvažnije varijable. Kako vidimo su varijable vezane uz karakteristiku nekretnine točnije površinu. Iznenađujuće je što su varijable vezane uz blizinu ugostiteljskih objekata u ovom slučaju kafića rangirane toliko visoko. Doduše njihov utjecaj je puno manji u odnosu na površinu nekretnine ali svejedno se nalaze u deset najvažnijih.



Grafikon 4-45 Deset najvažnijih značajki

Predviđanje cijene nekretnina ćemo odraditi u tri koraka. U prvom koraku treba učitati i pripremiti podatke, u drugom koraku treniramo algoritam slučajnih šuma, a u trećem radimo predviđanje cijene nekretnina. Algoritam slučajnih šuma je odabran kao jednostavan model pogodan za predviđanje cijene nekretnina. Uglavnom pomoću algoritma slučajnih šuma model se brzo istrenira, ali kako za predviđanje treba veći broj stabala, nije najbolji algoritam za aplikacije koje trebaju predviđanje u stvarnom vremenu. Broj stabala koji smo uzeli za predviđanje je 30 da bi dobili na brzini predviđanja algoritam slučajnih šuma.

Na kraju predviđanja ćemo napraviti usporedbu algoritma slučajnih šuma sa višestrukom regresijskom analizom, algoritmom pojačavanja gradijenta modela stabala i pojačavanjem modela stabala XGBoost modelom.

Za ocjenu modela ćemo koristiti korijen srednje kvadratne pogreške (engl. RMSE Root mean squared error). Korijen srednje kvadratne pogreške se definira kao korijen kvadratnih razlika veličine predviđanja i stvarne veličine ciljane varijable. U našem slučaju ciljana varijabla je cijena nekretnine.

4.3.1. Priprema podataka

Priprema podataka se sastoji od:

- Stvaranje vektora koji sadrži redne brojeve pojedinog predviđanja
- Stvaranja vektora koji sadrži ciljane varijable seta podataka za treniranje
- Stvaranje zajedničkog seta podataka za treniranje i testiranje
- Eliminacija rednog broja pojedinog predviđanja
- Pretvaranje datuma u broj
- Pretvaranje kategoričkih značajki u numeričke
- Zamjenom nedostajućih podataka prosječnom veličinom

Na početku trebamo Pythonove pakete koji će nam pomoći pri pripremi podataka, a to su redom numpy (linearna algebru i matematika), pandas (obrada podataka) i sklearn (algoritam slučajnih šuma).

Stvaranje vektora nam je potrebno da bi smo mogli vršiti matematičke operacije iz područja linearne algebre nad podacima. Kako završna excel tablica treba imati dva stupca, redni broj realne transakcije i ciljanu vrijednost odnosno predviđenu cijenu nekretnine, tako i te dvije veličine pretvaramo u vektore.

Isto tako da bi lakše i brže koristili podatke za treniranje i podatke za testiranje, stvaramo jedan zajednički set podataka. Nad tim zajedničkim setom podataka isto tako se provodi vektorizacija da bi mogli koristiti matematičke operacije iz područja linearne algebre. U daljnjem koraku, iz zajedničkog seta podataka, maknuti ćemo redne brojeve realnih transakcija. Taj korak radimo iz prevencije da bi spriječili da se na temelju toga podatka predviđa cijena nekretnine. Ako bi redni broj realne transakcije ostavili u zajedničkom setu došlo bi do mogućeg utjecaja na algoritam slučajnih šuma, čime bi algoritam učio na podatku na kojem nije ni trebao učiti.

Datum je važna dio informacije koji je nemoguće koristiti u formatu u kakvom dolazi. Datum nam stavlja svaku realnu transakciju u vremenski okvir. Format mm-d.d.-gg nam nije prikladan za obradu pa ga pretvaramo u jedan broj.

Sljedećim korakom ćemo isto tako pretvoriti kategoričke varijable u numeričke. Svaka kategorička varijabla je zapisana tekstualno, kao npr. ime regije i kao takva je neupotrebljiva u linearnoj matematici. Svaku takvu kategoričku varijablu ćemo isto tako pretvoriti u numeričku.

Kako je nemoguće napraviti predviđanje cijene nekretnine sa nepotpunim podacima popuniti ćemo nedostajuće podatke sa prosječnom veličinom. Postoje tri načina kako bi mogli zamijeniti nedostajuće podatke, osim već navedene prosječne veličine temeljene na postojećim podacima, imamo i najčešće korištenom veličinom i medijanom s obzirom na postojeće poznate podatke. Po strukturi podataka korištenje prosječne veličine je bila optimalno.

4.3.2. Treniranje algoritma slučajnih šuma

Za stvaranje modela koristimo iz sklearn-a (Python paket) RandomForestRegressor da bi definirali veličinu algoritma slučajnih šuma. Glavna značajka algoritma slučajnih šuma je naravno broj stabala koji ćemo koristiti. Pretpostavka je da veći broj stabala i bolje predviđa cijenu nekretnina, ali troši više resursa i duže vremenski traje, dok naravno obrnuto pretpostavka je da manji broj stabala brže i uz manje korištenje resursa lošije predviđa cijenu nekretnina.

Kako se vidi sam model je jednostavan za korištenje. Od ponuđenih parametara samog modela koristio sam:

- random_state - da bi prilikom poziva na programski kod se dobili isti rezultati kada se koristi ista veličina parametra (osim cijelog pozitivnog prirodnog broja parametar se može postaviti na nasumičan broj ili se može u potpunosti izostaviti)
- oob_score – da bi koristio oob metodu za predviđanje na novim nepoznatim podacima (može biti u dva stanja TRUE ili FALSE)
- max_features – da bi koristio maksimalan broj značajki prilikom grananja stabala (može biti prirodan i decimalan broj, niz ili se može izostaviti, u tom slučaju je jednak ukupnom broju značajki)
- min_samples_leaf – da bi odredio minimalan broj značajki nakon grananja stabla (može biti prirodan, decimalan broj ili se može izostaviti, u tom slučaju je 1)

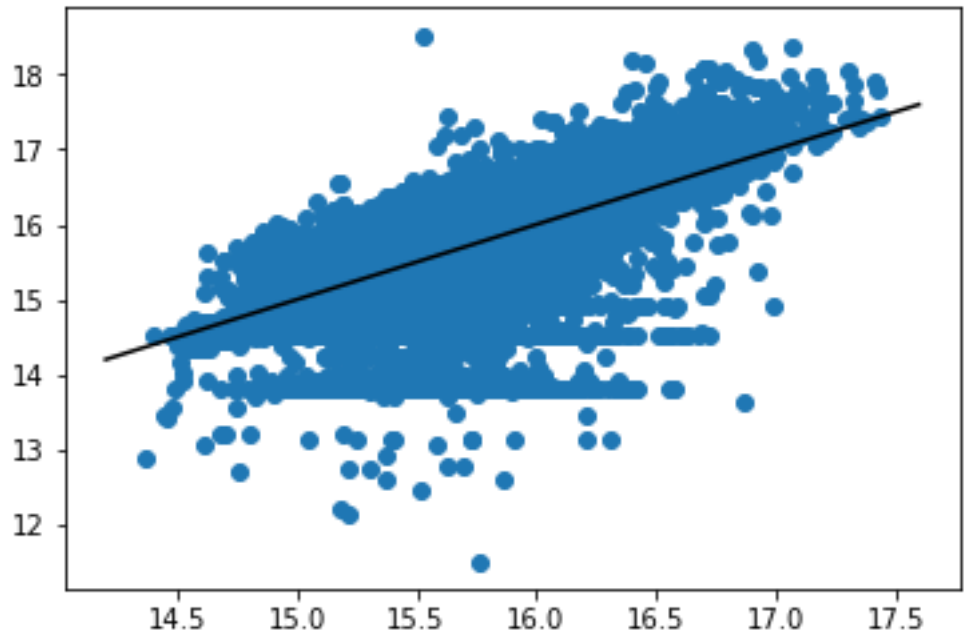
Model algoritma slučajnih šuma sam izvrtio u više navrata mijenjajući parametre pokušavajući dobiti čim bolji rezultat. Na kraju sam se odlučio za ovaj model

```
Model = RandomForestRegressor(n_estimators = 30,
                              random_state = 2017,
                              oob_score = True,
                              max_features = 20,
                              min_samples_leaf = 8)
```

Bez obzira na broj stabala rezultat se nije bitno mijenjao odnosno predviđanje većim brojem stabala nisam postigao bolje predviđanje cijene nekretnina. Stoga sam se odlučio za minimalan broj stabala bez da gubim na kvaliteti samog predviđanja cijene nekretnina, da bi uštedio na vremenu koje je potrebno za predviđanje cijene nekretnina.

4.3.3. Predviđanje i usporedba algoritma slučajnih šuma

Sama izlazna excel datoteka sa dobivenim cijenama nekretnina će biti dostupna u prilogu ovog rada. U ovom dijelu ću se pozabaviti odnosom stvarne greške i očekivane vrijednosti. Na sljedećem grafikonu je prikazana korelacija te dvije veličine. Iz priloženog se vidi da je raspodjela predviđene cijene i stvarne cijene relativno zadovoljavajuća. Predviđene cijene su zajedno grupirane i homogene su raspodjele. Zaključak je da je predviđanje dovoljno blizu stvarnoj veličini.



Grafikon 4-46 Grafikon odnosa stvarne greške i očekivane

Prije nego što usporedimo naš rezultat ćemo se ukratko upoznati sa drugim algoritmima. Višestruka regresijska analiza je u osnovi metoda linearne regresijske analize u kojoj se predviđa vrijednost jedne izlazne varijable na temelju dvije ili više ulaznih varijabli. Osim što vrijednost izlazne varijable ovisi o ulaznim (nezavisnim) varijablama višestruke regresijska analiza podrazumijeva i:

- Linearnu međuovisnost izlazne i ulaznih varijabli
- Ulazne varijable nisu međusobno jako korelirane
- Podaci na kojima se temelji analiza su nezavisno i nasumično odabrani
- Reziduali imaju normalnu razdiobu

U algoritmu pojačavanja gradijenta modela stabala za svako novo stablo se generira na temelju informacija iz prijašnjih stabala, time se stabla slabijeg učenja pojačavaju i daju stabla koja bolje uče.

Pojačavanje gradijenta stabala XGBoost modelom je jedan od češće korištenih modela zbog svoje brzine i točnosti predviđanja. XGBoost je napredniji model pojačavanja gradijenta.

Model Used	Best Public Score (Kaggle)	Features Used
Multiple Linear Regression	0.39561	kindergarten_km, school_km, metro_km_avto, public_healthcare_km, month, year, sub_area, full_sq, build_year, floor, max_floor, product_type
Gradient Boosting Trees Model	0.34711	full_sq, radiation_km, park_km, raion_popul, build_year, metro_km_avto, cpi, deposit_rate, public_healthcare_km, fitness_km, big_road1_km, shopping_centers_km, industrial_km, eurrub, metro_min_avto, max_floor, kindergarten_km, railroad_km, oil_urals, RTS_index
XGBoost	0.32695	All of 292 property-specific features plus additional engineered features

Grafikon 4-47 Rezultati dobiveni korištenjem drugih algoritama

Korijen srednje kvadratne pogreške za algoritam slučajnih šuma daje nam vrijednost od 0.47538. Vrijednosti bliže nuli imaju bolje predviđanje cijene nekretnina. Usporedbom sa rezultatima dobivenim dobivenih u grafikonu (A Data Scientist's Guide to Predicting Housing Prices in Russia, 2019.) algoritam slučajnih šuma sa rezultatom 0.47538 u odnosu na druge algoritme lošije predviđa. Algoritam slučajnih šuma kao algoritam je dosta grub i zahtjeva dodatne finese da bi se približio točnosti drugih algoritama. To je bila i očekivana pretpostavka na početku ovoga rada. Bitno je i zamijetiti na XGBoost model ima najbolju točnost bez obzira na to što koristi cijeli skup varijabli.

4.4. Smjernice za daljnja istraživanja

Zaključno ću navesti par smjernica koje su mi pomogle da oblikujem ovaj rad. Prva smjernica je svakako dobro poznavanje materije, bilo iskustvom ili proučavanjem dostupnog sadržaja. Važno dobrog poznavanja materije proizlazi iz nužnosti prepoznavanja relevantnih podataka i stavljanju u perspektivu istih. Druga smjernice tiče se postavljanja inicijalnih pretpostavki, koje ne smiju biti uvriježene i nepobitne nego podložne svakoj kritici. Pretpostavke su tu da se propitkuju i po potrebi mijenjaju. Treća smjernica odnosi se na pristup rješavanju pojedinog problema. Ne postoji unikatan pristup rješavanju problema, svaki problem zahtjeva drugačiji pristup. Stoga nikako se ne bi smjelo držati jednog recepta kojeg ćemo primjenjivati na sve. Treba nastojati naći adekvatniji pristup rješavanja problema ako to dostupni resursi dopuštaju (vrijeme, tehnologija, organizacija unutar tima, pritisak menadžmenta i slično.). Čak bih se usudio i reći čak i unatoč nedostatku resursa, inovacije

ne proizlaze iz hodanja lakšim putem. Četvrta smjernica bi bila da vlastite greške prilikom istraživanja pretvore u prilike za učenje a ne izvore frustracija. Greške su dio procesa, ne samo učenja nego i života. Peta smjernica se odnosi na algoritam slučajnih šuma koji se pokazao kao izvrstan inicijalni korak analize. Vrlo jednostavno testira važnost varijabli pa može služiti kao osnovna pretpostavka.

Zaključak

Krećući od analize podataka sreli smo se sa cijelim nizom problema i nedoumica. Sami preuzeti podaci su bili manjkavi (samo 20% podataka za treniranje je bilo u potpunosti cjelovito) sa puno nedostajućih podataka i vrijednostima koje nisu bile logične. Na takve manjkave podatke se nadovezala i situacija u kojoj se Rusija snašla zahvaljujući sukobu sa Ukrajinom i nametnutim ekonomskim sankcijama. Osobno sam očekivao da će cijene nekretnina u Rusiji imati puno veće i učestalije skokove i padove. Očito da rusko-europska međusobna ovisnost o uvozu-izvozu energenata bi više škodila Europi nego Rusiji u nekom dužem periodu. Potreba za energentima raste pa stoga će rasti i ekonomska snaga Rusije. Jedan od problema je i taj što alternativni izvori energije još uvijek ne mogu zamijeniti primarne.

Bez obzira na cijeli niz parametara ipak proizlazi da površina nekretnine još uvijek igra veliku ulogu u određivanju cijene nekretnine. Sami podaci u nekim trenutcima sakrivaju pravu vrijednost, poput rast cijene nekretnine u ovisnosti o visini kata i broju katova koji nekretnina ima. Znači osim sirovih podataka potrebno je poznavati i pozadinu stvari a ne na pretek zaključivati.

Druge karakteristike varijabli isto tako su nam jednoznačno potvrdile da blizina centru i blizina sportskim, kulturnim i obrazovnim institucijama znači i veću cijenu nekretnine. Osim toga na gušće naseljenom području cijenu nekretnina će tržište samo podignuti gonjeno osnovnim principima ponude i potražnje. Zanimljivo je bilo i vidjeti tako rečeno puls tržišta nekretnina ovisno o mjesecu u godini.

Usporedbom podataka za treniranje i podataka za testiranje vidjeli smo sličnosti u razdiobi podataka što nam je u jednu ruku pomoglo dok u drugu odmoglo. Jednom istrenirani podaci testirani u sličnom setu podataka dat će i bolje rezultate, što može biti poprilično varljivo. Kako je vidljivo u programskom kodu inicijalan veličina mjere za usporedbu, korijena srednje kvadratne pogreške, bila je 0,37598. Naravno kada smo testirali algoritam slučajnih šuma na novo i nepoznatom setu podataka ta veličina se popela na 0,47538.

Što se tiče algoritma slučajnih šuma sam po sebi pokazuje nedostatke u odnosu na novije algoritme. To se posebno odnosi na korištenje XGBoost algoritma koji finije ugađa parametre i samim time bolje predviđa. To ne znači da algoritam slučajnih šuma i dalje nema svoju vrijednost. Kada se kreće sa analizom i traži najoptimalnije model ne postoji unaprijed

shema koju treba slijediti. Barem ne što se tiče modela. Svakako treba provesti detaljnu analizu podataka da bi se znalo s čime se radi, koji su nedostaci, koje prednosti. Naravno treba imati i određeno znanje iz područja na koje se analiza primjenjuje ili surađivati sa stručnjakom iz toga područja. Kako za određenu vrstu problema se ne koristi uvijek isti princip, tako ne možemo očekivati da postoji magični štapić kojim ćemo riješiti problem uvijek istim načinom. Jedna od velikih stavki je očito i iskustvo koje dolazi i spremnost na konstantno učenje i unaprjeđivanje vlastitoga znanja.

Zaključno u radu su potvrđene obje pretpostavke. Površina je najvažnija varijabla u određivanju cijene nekretnina, dok položaj nekretnine svakako igra ulogu ali u manjem udjelu, te da algoritam slučajnih šuma predviđa slabije od algoritama slične svrhe.

Sve u svemu ovo je za mene bilo jedno zahvalno iskustvo jer je ovo prvi projekt koji sam sam odradio od početka do kraja. Naučio sam ne samo primijeniti naučeno već i istražiti i primijeniti ono što mi je do sada bilo nepoznato. To i je cilj obrazovanje, sposobnost snalaženja a ne učenje napamet.

Popis kratica

RF	*Random forest*	algoritam slučajnih šuma
oob	*Out Of the Bag*	metoda za mjerenje greške predviđanja
RMSE	Root mean squared error	korijen srednje kvadratne pogreške

Popis grafikona

Literatura

[1] MAJČICA, B. Procjena nekretnina-novi izazov za JLP(R)S, *Tim4pin Magazin* 12/2014, 92-104.

[2] TICA, J. Metode procjene vrijednosti nekretnina-tumačenje nove Uredbe i svjetska praksa. *Materijali sa specijalističkog seminara o metodama procjene vrijednosti nekretnina s naglaskom na prijedlog nove Uredbe i usporedbu sa svjetskom praksom.* 2014.

[3] TICA, J. Tumačenje Prihodovne Metode I Praktičan Primjer Izračuna Procjene. https://www.sudski-vjestaci.hr/_simpozij-2014-09_/Josip_Tica-Tumacenje_prihodovne_metode_i_praktican_primjer_izracuna_procjene.pdf 11.09.2018, rujan 2014a

[4] SLIŠKOVIĆ, T. Međuovisnost makroekonomske aktivnosti i tržišta nekretnina u Hrvatskoj. *Doktorski rad*. Sveučilište u Zagrebu, 2016.

[5] BREIMAN, L. Random Forests. *Machine Learning*, Volume 45, Issue 1, 5–32.

[6] STALLINGS, W. *Local Computer Networks*. London: John Wiley, 2006a.

[7] ATM FORUM, User-Network Interface (UNI) Specification, http://www.atmforum.com, travanj. 2010.

[8] BRADY, P.T. A statistical Analysis of On-off Patterns in 16 Conversation, *Bell System Technical Journal*, 47,1 (1998), 55-62.

[9] BRADY, N. A statistical Analysis of Use Case. *Proceedings of the 7th International Conference on Telecommunications ConTEL*, Zagreb, (2003), 45-52.

[10] LILYS, M. Final data structures. *Doktorski rad*. Sveučilište u Zagrebu, 2010.

[11] The US Energy Information Administration (EIA), https://www.eia.gov/beta/international/analysis.php?iso=RUS, ožujak 2019.

[12] Wikipedia: Russian financial crisis (2014-2017), https://en.wikipedia.org/wiki/Russian_financial_crisis_(2014%E2%80%932017), ožujak 2019.

[13] A Data Scientist's Guide to Predicting Housing Prices in Russia, https://www.r-bloggers.com/a-data-scientists-guide-to-predicting-housing-prices-in-russia/, ožujak 2019.

[14] Moscow Population 2019, http://worldpopulationreview.com/world-cities/moscow-population/, svibanj 2019.

„Pod punom odgovornošću pismeno potvrđujem da je ovo moj autorski rad čiji niti jedan dio nije nastao kopiranjem ili plagiranjem tuđeg sadržaja. Prilikom izrade rada koristio sam tuđe materijale navedene u popisu literature ali nisam kopirao niti jedan njihov dio, osim citata za koje sam naveo autora i izvor te ih jasno označio znakovima navodnika. U slučaju da se u bilo kojem trenutku dokaže suprotno, spreman sam snositi sve posljedice uključivo i poništenje javne isprave stečene dijelom i na temelju ovoga rada“.

U Zagrebu, datum.

Ime Prezime

Prilog

KOD ZA DETALJNU ANALIZU PODATAKA

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
color = sns.color_palette()
%matplotlib inline
pd.options.mode.chained_assignment = None  # default='warn'
pd.set_option('display.max_columns', 500)
train_df =
pd.read_csv("/Users/tnemet/Desktop/dataset/train.csv",
parse_dates=['timestamp'])
train_df['price_doc_log'] = np.log1p(train_df['price_doc'])
train_na = (train_df.isnull().sum() / len(train_df)) * 100
train_na = train_na.drop(train_na[train_na ==
0].index).sort_values(ascending=False)
f, ax = plt.subplots(figsize=(12, 8))
plt.xticks(rotation='90')
sns.barplot(x=train_na.index, y=train_na)
ax.set(title='Postotak podataka koji nedostaje po pojedinoj
vrsti', ylabel='% podataka koji nedostaje')
# state should be discrete valued between 1 and 4. There is a
33 in it that is cleary a data entry error
# Lets just replace it with the mode.
train_df.loc[train_df['state'] == 33, 'state'] =
train_df['state'].mode().iloc[0]
# build_year has an erronus value 20052009. Since its unclear
which it should be, let's replace with 2007
train_df.loc[train_df['build_year'] == 20052009,
'build_year'] = 2007
internal_chars = ['full_sq', 'life_sq', 'floor', 'max_floor',
'build_year', 'num_room', 'kitch_sq', 'state', 'price_doc']
corrmat = train_df[internal_chars].corr()
f, ax = plt.subplots(figsize=(10, 7))
plt.xticks(rotation='90')
sns.heatmap(corrmat, square=True, linewidths=.5, annot=True)
```

```
f, ax = plt.subplots(figsize=(10, 7))
plt.scatter(x=train_df['full_sq'], y=train_df['price_doc'],
c='r')
plt.title('Korelacija ukupne površine i cijene')
plt.xlabel('ukupna površina')
plt.ylabel('cijena')
f, ax = plt.subplots(figsize=(10, 7))
ind = train_df[train_df['full_sq'] > 2000].index
plt.scatter(x=train_df.drop(ind)['full_sq'],
y=train_df.drop(ind)['price_doc'], c='r', alpha=0.5)
ax.set(title='Korekcija korelacije ukupne površine i cijene
nekretnine', xlabel='ukupna površina', ylabel='Cijena')
(train_df['life_sq'] > train_df['full_sq']).sum()
f, ax = plt.subplots(figsize=(10, 7))
sns.countplot(x=train_df['num_room'])
ax.set(title='Distribucija broj soba', xlabel='broj soba',
ylabel='količina soba')
f, ax = plt.subplots(figsize=(16, 8))
plt.xticks(rotation='90')
ind = train_df[(train_df['build_year'] <= 1691) |
(train_df['build_year'] >= 2018)].index
by_df = train_df.drop(ind).sort_values(by=['build_year'])
sns.countplot(x=by_df['build_year'])
ax.set(title='Distribucija po godini gradnje', xlabel='godina
gradnje', ylabel='količina stanova u određenoj godini')
f, ax = plt.subplots(figsize=(12, 6))
by_price = by_df.groupby('build_year')[['build_year',
'price_doc']].mean()
sns.regplot(x="build_year", y="price_doc", data=by_price,
scatter=False, order=3, truncate=True)
plt.plot(by_price['build_year'], by_price['price_doc'],
color='r')
ax.set(title='Prosječna cijena nekretnine po godini gradnje',
ylabel='cijena',xlabel='godina gradnje')
f, ax = plt.subplots(figsize=(12, 6))
ts_df = train_df.groupby('timestamp')[['price_doc']].mean()
#sns.regplot(x="timestamp", y="price_doc", data=ts_df,
scatter=False, truncate=True)
plt.plot(ts_df.index, ts_df['price_doc'], color='r', )
ax.set(title='Dnevno kretanje medijan cijena 2011-2015')
import datetime
```

```
import matplotlib.dates as mdates
years = mdates.YearLocator()   # every year
yearsFmt = mdates.DateFormatter('%Y')
ts_vc = train_df['timestamp'].value_counts()
f, ax = plt.subplots(figsize=(12, 6))
plt.bar(ts_vc.index, ts_vc ,color='r')
ax.xaxis.set_major_locator(years)
ax.xaxis.set_major_formatter(yearsFmt)
ax.set(title='Volumen prodaje kroz vrijeme', ylabel='Broj
transakcija')
f, ax = plt.subplots(figsize=(12, 8))
ts_df =
train_df.groupby(by=[train_df.timestamp.dt.month])[['price_do
c']].median()
plt.plot(ts_df.index, ts_df, color='r')
ax.set(title='Cijena nekretnine po mjesecu po godini')
f, ax = plt.subplots(figsize=(12, 8))
ind = train_df[train_df['state'].isnull()].index
train_df['price_doc_log10'] = np.log10(train_df['price_doc'])
sns.violinplot(x="state", y="price_doc_log10",
data=train_df.drop(ind), inner="box")
# sns.swarmplot(x="state", y="price_doc_log10",
data=train_df.dropna(), color="w", alpha=.2);
ax.set(title='Log10 medijan cijena ovisno o stanju
nekretnine', xlabel='stanje nekretnine',
ylabel='log10(cijene)')
f, ax = plt.subplots(figsize=(12, 8))
ind = train_df[train_df['material'].isnull()].index
sns.violinplot(x="material", y="price_doc_log",
data=train_df.drop(ind), inner="box")
sns.swarmplot(x="state", y="price_doc_log10",
data=train_df.dropna(), color="r", alpha=.2);
ax.set(title='Cijena nekretnine ovisno o građevinskom
materijalu', xlabel='građevinski materijal',
ylabel='log(cijena)')

f, ax = plt.subplots(figsize=(12, 8))
plt.scatter(x=train_df['floor'], y=train_df['price_doc_log'],
c='r', alpha=0.4)
sns.regplot(x="floor", y="price_doc_log", data=train_df,
scatter=False, truncate=True)
```

```
ax.set(title='Cijena nekretnine ovisno o katu na kojem je
stan', xlabel='kat', ylabel='log(cijena)')
f, ax = plt.subplots(figsize=(12, 8))
plt.scatter(x=train_df['max_floor'],
y=train_df['price_doc_log'], c='r', alpha=0.4)
sns.regplot(x="max_floor", y="price_doc_log", data=train_df,
scatter=False, truncate=True)
ax.set(title='Cijena nekretnine ovisno o ukupnom broju katova
zgrade', xlabel='katovi zgrade', ylabel='log(cijene)')
f, ax = plt.subplots(figsize=(12, 8))
plt.scatter(x=train_df['floor'], y=train_df['max_floor'],
c='r', alpha=0.4)
plt.plot([0, 80], [0, 80], color='.5')
ax.set(title='Raspored stanova u zgradi ovisno o katu',
xlabel='kat stana', ylabel='kat zgrade')
train_df.loc[train_df['max_floor'] < train_df['floor'],
['id', 'floor','max_floor']].head(20)
demo_vars = ['area_m', 'raion_popul', 'full_all', 'male_f',
'female_f', 'young_all', 'young_female',
             'work_all', 'work_male', 'work_female',
'price_doc']
corrmat = train_df[demo_vars].corr()
f, ax = plt.subplots(figsize=(10, 7))
plt.xticks(rotation='90')
sns.heatmap(corrmat, square=True, linewidths=.5, annot=True)
ax.set(title='Toplinska mapa demografije')
train_df['area_km'] = train_df['area_m'] / 1000000
train_df['density'] = train_df['raion_popul'] /
train_df['area_km']
f, ax = plt.subplots(figsize=(10, 6))
sa_price = train_df.groupby('sub_area')[['density',
'price_doc']].median()
sns.regplot(x="density", y="price_doc", data=sa_price,
scatter=True, truncate=True)
ax.set(title='Median cijene nekretnine po gutoći naseljenosti
regije u km2')
f, ax = plt.subplots(figsize=(10, 20))
sa_vc = train_df['sub_area'].value_counts()
sa_vc = pd.DataFrame({'sub_area':sa_vc.index, 'count':
sa_vc.values})
```

```
ax = sns.barplot(x="count", y="sub_area", data=sa_vc,
orient="h")
ax.set(title='Broj transakcija po regiji', xlabel='broj',
ylabel='regija')
f.tight_layout()
train_df['work_share'] = train_df['work_all'] /
train_df['raion_popul']
f, ax = plt.subplots(figsize=(12, 6))
sa_price = train_df.groupby('sub_area')[['work_share',
'price_doc']].mean()
sns.regplot(x="work_share", y="price_doc", data=sa_price,
scatter=True, order=4, truncate=True)
ax.set(title='Srednja cijena nekretnine po regiji ovisno o
radno sposobnom stanovništvu',
       xlabel='udio radno sposobnog
stanovništva',ylabel='cijena nekretnine')
school_chars = ['children_preschool', 'preschool_quota',
'preschool_education_centers_raion', 'children_school',
                'school_quota',
'school_education_centers_raion',
'school_education_centers_top_20_raion',
                'university_top_20_raion',
'additional_education_raion', 'additional_education_km',
'university_km', 'price_doc']
corrmat = train_df[school_chars].corr()
f, ax = plt.subplots(figsize=(10, 7))
plt.xticks(rotation='90')
sns.heatmap(corrmat, square=True, linewidths=.5, annot=True)
ax.set(title='Toplinska mapa obrazovnih institucija')
school_chars = ['children_preschool', 'preschool_quota',
'preschool_education_centers_raion', 'children_school',
                'school_quota',
'school_education_centers_raion',
'school_education_centers_top_20_raion',
                'university_top_20_raion',
'additional_education_raion', 'additional_education_km',
'university_km', 'price_doc']
corrmat = train_df[school_chars].corr()
f, ax = plt.subplots(figsize=(10, 7))
plt.xticks(rotation='90')
sns.heatmap(corrmat, square=True, linewidths=.5, annot=True)
```

```
ax.set(title='Toplinska mapa obrazovnih institucija')
cult_chars = ['sport_objects_raion',
'culture_objects_top_25_raion', 'shopping_centers_raion',
'park_km', 'fitness_km',
               'swim_pool_km', 'ice_rink_km','stadium_km',
'basketball_km', 'shopping_centers_km', 'big_church_km',
               'church_synagogue_km', 'mosque_km',
'theater_km', 'museum_km', 'exhibition_km', 'catering_km',
'price_doc']
corrmat = train_df[cult_chars].corr()
f, ax = plt.subplots(figsize=(12, 7))
plt.xticks(rotation='90')
sns.heatmap(corrmat, square=True, linewidths=.5, annot=True)
ax.set(title='Toplinska mapa kulturno rekreacijskog
sadržaja')
f, ax = plt.subplots(figsize=(10, 6))
so_price =
train_df.groupby('sub_area')[['sport_objects_raion',
'price_doc']].median()
sns.regplot(x="sport_objects_raion", y="price_doc",
data=so_price, scatter=True, truncate=True)
ax.set(title='Medijan cijena nekretnina po regiji ovisno o
broj sportskih objekata',
       xlabel='sportski objekti', ylabel='cijena nekretnine')
f, ax = plt.subplots(figsize=(10, 6))
co_price =
train_df.groupby('sub_area')[['culture_objects_top_25_raion',
'price_doc']].median()
sns.regplot(x="culture_objects_top_25_raion", y="price_doc",
data=co_price, scatter=True, truncate=True)
ax.set(title='Medijan cijena nekretnina po regiji ovisno o
broj kulturnih objekata',
      xlabel='kulturni objekti', ylabel='cijena nekretnine')
train_df.groupby('culture_objects_top_25')['price_doc'].media
n()
f, ax = plt.subplots(figsize=(10, 6))
sns.regplot(x="park_km", y="price_doc", data=train_df,
scatter=True, truncate=True, scatter_kws={'color': 'r',
'alpha': .2})
ax.set(title='Medijan cijena nekretnina po regiji ovisno o
udaljenosti parka',
```

```
      xlabel='udaljenost parka', ylabel='cijena nekretnine')
inf_features = ['nuclear_reactor_km',
'thermal_power_plant_km', 'power_transmission_line_km',
'incineration_km',
                'water_treatment_km', 'incineration_km',
'railroad_station_walk_km', 'railroad_station_walk_min',
                'railroad_station_avto_km',
'railroad_station_avto_min', 'public_transport_station_km',
                'public_transport_station_min_walk',
'water_km', 'mkad_km', 'ttk_km',
'sadovoe_km','bulvar_ring_km',
                'kremlin_km', 'price_doc']
corrmat = train_df[inf_features].corr()
f, ax = plt.subplots(figsize=(12, 7))
plt.xticks(rotation='90')
sns.heatmap(corrmat, square=True, linewidths=.5, annot=True)
ax.set(title='Toplinska mapa infrastrukture')
f, ax = plt.subplots(figsize=(10, 6))
sns.regplot(x="kremlin_km", y="price_doc", data=train_df,
scatter=True, truncate=True, scatter_kws={'color': 'r',
'alpha': .2})
ax.set(title='Cijena nekretnine ovisno o udaljenosti od
Kremlja', xlabel='udaljenost od Kremlja', ylabel='cijena
nekretnine')
from sklearn.ensemble import RandomForestRegressor
from sklearn.preprocessing import LabelEncoder
X_train = train_df.drop(labels=['timestamp', 'id',
'incineration_raion'], axis=1).dropna()
y_train = X_train['price_doc']
X_train.drop('price_doc', axis=1, inplace=True)
for f in X_train.columns:
    if X_train[f].dtype == 'object':
        lbl = LabelEncoder()
        lbl.fit(X_train[f])
        X_train[f] = lbl.transform(X_train[f])
rf = RandomForestRegressor(random_state=0)
rf = rf.fit(X_train, y_train)
fi = list(zip(X_train.columns, rf.feature_importances_))
print('## rf variable importance')
d = [print('## %-40s%s' % (i)) for i in fi[:20]]
```

```
test_df =
pd.read_csv("/Users/tnemet/Desktop/dataset/test.csv",
parse_dates=['timestamp'])
test_na = (test_df.isnull().sum() / len(test_df)) * 100
test_na = test_na.drop(test_na[test_na ==
0].index).sort_values(ascending=False)
f, ax = plt.subplots(figsize=(12, 8))
plt.xticks(rotation='90')
sns.barplot(x=test_na.index, y=test_na)
ax.set(title='Postotak podataka koji nedostaju', ylabel='%
podataka koji nedostaju', xlabel='varijable')
all_data = pd.concat([train_df.drop('price_doc', axis=1),
test_df])
all_data['dataset'] = ''
l = len(train_df)
all_data.iloc[:l]['dataset'] = 'train'
all_data.iloc[l:]['dataset'] = 'test'
train_dataset = all_data['dataset'] == 'train'
f, ax = plt.subplots(nrows=1, ncols=2, figsize=(12, 8),
sharey=True)
all_data['full_sq_log'] = np.log1p(all_data['full_sq'])
all_data.drop(train_dataset)["full_sq_log"].plot.kde(ax=ax[0]
)
all_data.drop(~train_dataset)["full_sq_log"].plot.kde(ax=ax[1
])
ax[0].set(title='podaci za testiranje', xlabel='ukupna
površina',ylabel='razdioba podataka')
ax[1].set(title='podaci za treniranje', xlabel='ukupna
površina',ylabel='razdioba podataka')
f, ax = plt.subplots(nrows=1, ncols=2, figsize=(12, 8),
sharey=True)
all_data['life_sq_log'] = np.log1p(all_data['life_sq'])
all_data.drop(train_dataset)["life_sq_log"].plot.kde(ax=ax[0]
)
all_data.drop(~train_dataset)["life_sq_log"].plot.kde(ax=ax[1
])
ax[0].set(title='podaci za testiranje', xlabel='neto
površina',ylabel='razdioba podataka')
ax[1].set(title='podaci za treniranje', xlabel='neto
površina',ylabel='razdioba podataka')
```

```
f, ax = plt.subplots(nrows=1, ncols=2, figsize=(12, 8),
sharey=True)
all_data['kitch_sq_log'] = np.log1p(all_data['kitch_sq'])
all_data.drop(train_dataset)["kitch_sq_log"].plot.kde(ax=ax[0
])
all_data.drop(~train_dataset)["kitch_sq_log"].plot.kde(ax=ax[
1])
ax[0].set(title='podaci za testiranje', xlabel=' površina
kuhinje',ylabel='razdioba podataka')
ax[1].set(title='podaci za treniranje', xlabel=' površina
kuhinje',ylabel='razdioba podataka')
f, ax = plt.subplots(nrows=1, ncols=2, figsize=(12, 8),
sharey=True)
sns.countplot(x=test_df['num_room'], ax=ax[0])
sns.countplot(x=train_df['num_room'], ax=ax[1])
ax[0].set(title='podaci za testiranje', xlabel=' broj soba
',ylabel='razdioba podataka')
ax[1].set(title='podaci za treniranje', xlabel='broj
soba',ylabel='razdioba podataka')
f, ax = plt.subplots(nrows=1, ncols=2, figsize=(12, 8),
sharey=True)
all_data.drop(train_dataset)["floor"].plot.kde(ax=ax[0])
all_data.drop(~train_dataset)["floor"].plot.kde(ax=ax[1])
ax[0].set(title='podaci za testiranje', xlabel=' kat na kojem
je stan ',ylabel='razdioba podataka')
ax[1].set(title='podaci za treniranje', xlabel='kat na kojem
je stan',ylabel='razdioba podataka')
f, ax = plt.subplots(nrows=1, ncols=2, figsize=(12, 8),
sharey=True)
all_data.drop(train_dataset)["max_floor"].plot.kde(ax=ax[0])
all_data.drop(~train_dataset)["max_floor"].plot.kde(ax=ax[1])
ax[0].set(title='podaci za testiranje', xlabel=' broj katova
zgrade ',ylabel='razdioba podataka')
ax[1].set(title='podaci za treniranje', xlabel='broj katova
zgrade',ylabel='razdioba podataka')
f, ax = plt.subplots(nrows=1, ncols=2, figsize=(12, 8),
sharey=True)
ax[0].scatter(x=test_df['floor'], y=test_df['max_floor'],
c='r', alpha=0.4)
ax[0].plot([0, 80], [0, 80], color='.5')
```

```
ax[1].scatter(x=train_df['floor'], y=train_df['max_floor'],
c='r', alpha=0.4)
ax[1].plot([0, 80], [0, 80], color='.5')
ax[0].set(title='podaci za testiranje', xlabel=' kat na kojem
je stan ',ylabel='broj katova zgrade ')
ax[1].set(title='podaci za treniranje', xlabel='kat na kojem
je stan',ylabel='broj katova zgrade ')
years = mdates.YearLocator()   # every year
yearsFmt = mdates.DateFormatter('%Y')
ts_vc_train = train_df['timestamp'].value_counts()
ts_vc_test = test_df['timestamp'].value_counts()
f, ax = plt.subplots(figsize=(12, 6))
plt.bar(ts_vc_train.index, ts_vc_train)
plt.bar(ts_vc_test.index, ts_vc_test)
ax.xaxis.set_major_locator(years)
ax.xaxis.set_major_formatter(yearsFmt)
ax.set(title='Broj transakcija po danu', ylabel='broj',
xlabel='godina')
f, ax = plt.subplots(nrows=1, ncols=2, figsize=(12, 8),
sharey=True)
sns.countplot(x=test_df['product_type'], ax=ax[0])
sns.countplot(x=train_df['product_type'], ax=ax[1])
ax[0].set(title='podaci za testiranje', xlabel='vrsta
transakcije', ylabel='broj')
ax[1].set(title='podaci za treniranje', xlabel='vrsta
transakcije',ylabel='broj')
f, ax = plt.subplots(nrows=1, ncols=2, figsize=(12, 8),
sharey=True)
sns.countplot(x=test_df['state'], ax=ax[0])
sns.countplot(x=train_df['state'], ax=ax[1])
ax[0].set(title='podaci za testiranje', xlabel=' stanje
nekretnine', ylabel='broj')
ax[1].set(title='podaci za treniranje', xlabel=' stanje
nekretnine',ylabel='broj')
f, ax = plt.subplots(nrows=1, ncols=2, figsize=(12, 8),
sharey=True)
sns.countplot(x=test_df['material'], ax=ax[0])
sns.countplot(x=train_df['material'], ax=ax[1])
ax[0].set(title='podaci za testiranje', xlabel=' građevinski
materijal', ylabel='broj')
```

```
ax[1].set(title='podaci za treniranje', xlabel='građevinski
materijal ',ylabel='broj')
```

KOD ZA PREDVIĐANJE CIJENE NEKRETNINA

```
#korak 1: unos i priprema podataka
import numpy as np
import pandas as pd
from sklearn.tree import DecisionTreeRegressor
from sklearn.ensemble import RandomForestRegressor
from sklearn import  preprocessing
import datetime as dt
from sklearn.tree import export_graphviz
import matplotlib.pyplot as plt
%matplotlib inline

#unos podataka
train_df =
pd.read_csv("/Users/tnemet/Desktop/dataset/train.csv")
test_df =
pd.read_csv("/Users/tnemet/Desktop/dataset/test.csv")
macro_df =
pd.read_csv("/Users/tnemet/Desktop/dataset/macro.csv")
train_df.head()

#vektori
id_test = test_df.id
y_train = train_df["price_doc"]
x_train = train_df.drop(["id", "timestamp", "price_doc"],
axis=1)
x_test = test_df.drop(["id", "timestamp"], axis=1)
df_all = pd.concat([test_df, train_df])
df_all = df_all.drop("id", axis=1)

#vrijeme
fromDate = min(df_all['timestamp'])
df_all['timedelta'] = (df_all['timestamp'] -
fromDate).dt.days.astype(int)
print(df_all[['timestamp', 'timedelta']].head())
df_all.drop('timestamp', axis = 1, inplace = True)
```

```
#kategoričke varijable
for c in x_train.columns:
    if x_train[c].dtype == 'object':
        lbl = preprocessing.LabelEncoder()
        lbl.fit(list(x_train[c].values))
        x_train[c] = lbl.transform(list(x_train[c].values))

#nedostajući podaci
imputer = preprocessing.Imputer(missing_values='NaN',
strategy = 'mean', axis = 0)
x_train = imputer.fit_transform(x_train)
x_test = imputer.fit_transform(x_test)
# korak 2: treniranje algoritma slučajnih mreža
#stvaranje modela
Model = RandomForestRegressor(n_estimators = 30,
                              random_state = 2017,
                              oob_score = True,
                              max_features = 20,
                              min_samples_leaf = 8)

Model.fit(X = x_train, y =y_train )
# korak 3: predviđanje
ylog_pred = Model.predict(X = x_train)

# provjera greške treniranja
np.sqrt(np.mean((ylog_pred - y_train)**2))
# oko 0.37

#da bi vidjeli kako se model ponaša na podacima na novim
neviđenim podacima koristiti ćemo out of the bag instancu
np.sqrt(np.mean((Model.oob_prediction_ - y_train)**2))
# 0.47

#stvarna greška nasuprot očekivane greške
fig, ax = plt.subplots()
plt.scatter(Model.oob_prediction_, y_train)
x = np.linspace(*ax.get_xlim())
ax.plot(x, x, color = 'black')
plt.show()

#10 najvažnijih varijabli
```

```
df_ = pd.DataFrame(df_all.columns, columns = ['feature'])
df_['fscore'] = Model.feature_importances_[:,]
df_['fscore'] = df_['fscore'] / df_['fscore'].max()
df_.sort_values('fscore', ascending = False, inplace = True)
df_ = df_[0:10]
df_.sort_values('fscore', ascending = True, inplace = True)
df_.plot(kind='barh', x='feature', y='fscore', legend=False,
figsize=(6, 10))
plt.title('Random forest feature importance', fontsize = 24)
plt.xlabel('')
plt.ylabel('')
plt.xticks([], [])
plt.yticks(fontsize=20)
plt.show()

#izlazna excel datoteka sa predviđanjima
ylog_pred = Model.predict(x_test)
y_pred = np.exp(ylog_pred) - 1
output = pd.DataFrame({'id': id_test, 'price_doc':
y_predict})
output.to_csv('RandomForest.csv', index=False)
```