# Univerzitet u Beogradu

# Elektrotehnički fakultet



## “Real Estate Price Predictor”

## Projekat iz predmeta Pronalaženje skrivenog znanja

|  |
| --- |
| Student: |
| Predrag Mitrović 2019/3173 |

Beograd, Jul 2020.

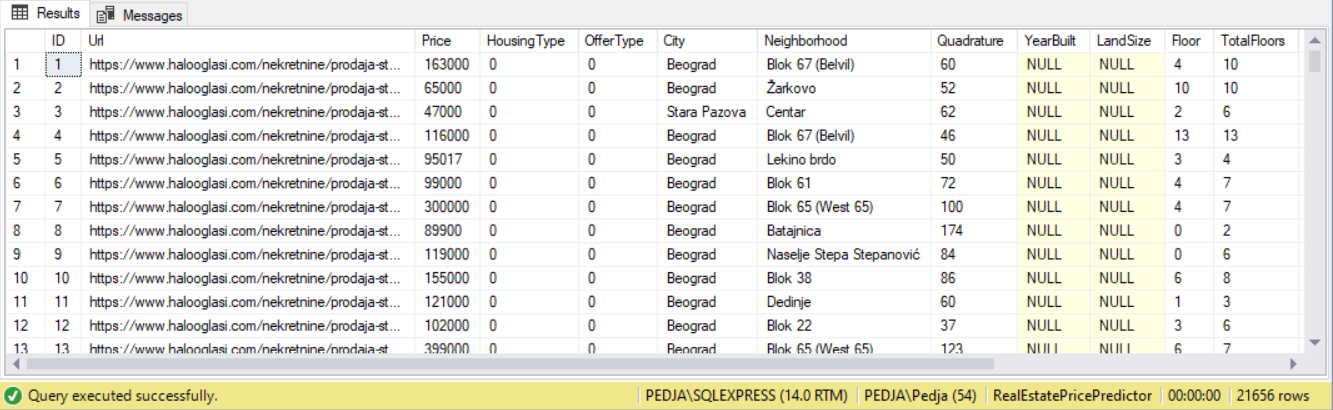
# Prikupljanje podataka

Kao izvor podataka korišćen je sajt <https://www.halooglasi.com/nekretnine>. Web crawler i web scraper su implementirani u programskom jeziku C#. Za skladištenje je korišćena Microsoft SQL Server baza podataka. Prilikom slanja HTTP zahteva ka sajtu je korišćena metoda rotirajućih proxy-ja. Proxy-ji su dohvaćeni sa sajta <https://gimmeproxy.com/api/getProxy> nisu uvek bili potpuno stabilni, dešavalo se da ne vraćaju odgovor, što je i očekivano s obizom na to da su besplatni. Prikupljanje podataka je podeljeno u dve faze. Prva faza je bila dohvatanje linkova ka oglasima i upisivanje u fajl, a druga faza je bila čitanje iz pomenutog fajla, parsiranje stranica sa zadatih linkova i upis u bazu podataka.

Ono što je bitno napomenuti jeste da dve informacije o nekretninama koje su tražene u tekstu zadatka ne postoje na sajtu, a to su godina izgradnje i broj kupatila.

Rezultat prve faze je 21656 redova u bazi podataka.

SELECT \* FROM RealEstates;

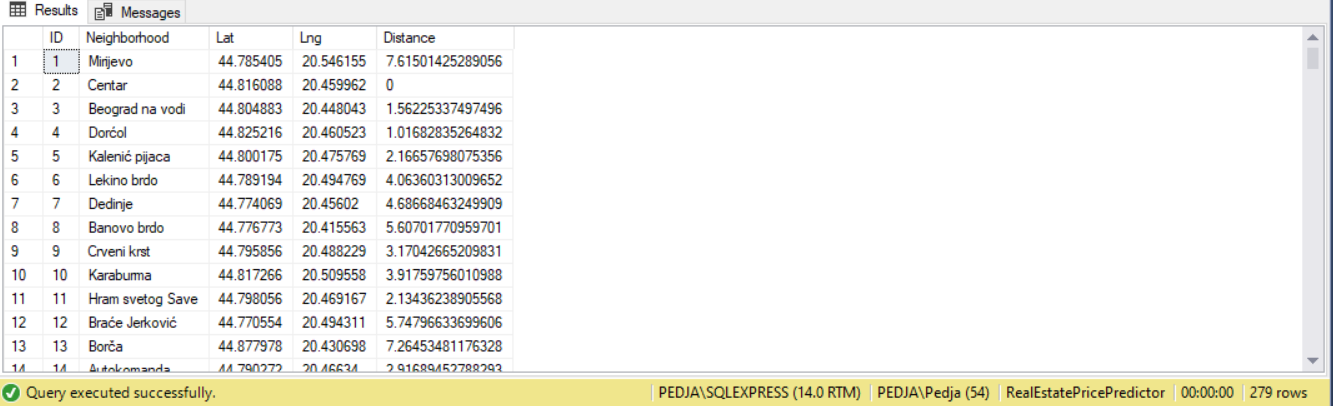


Tokom implementacije su korišćeni sledeći NuGet paketi:

* Entity Framework – Objektno relacioni maper, pri čemu je korišćena Code First tehnika
* HtmlAgilityPack – Parser koji omogućava jednostavno parsiranje HTML DOM-a pomoću XPATH-a
* Newtonsoft.Json – Paket za rad sa JSON objektima
* Selenium.WebDriver – Paket koji je korišćen kako bi se izvršio Javascript kod i pravilno interpretirao dinamički sadržaj dohvaćenih stranica pre nego što se krene sa parsiranjem

Osim tabele RealEstates, kreirana je i tabela NeighborhoodDistances koja sadrži podatke o tome koliko su beogradska naselja koja se pominju u oglasima daleko od centra grada. Tabela će biti korišćena u četvrtoj i petoj fazi projekta. Za podatke o geografskoj širini i dužini je korišćen Google-ov API, a distanca je izračunata pomoću klase GeoCoordinate iz System.Device.Location namespace-a. Distance su date u kilometrima.

SELECT \* FROM NeighborhoodDistances;



# Analiza podataka

Podaci su analizirani izvršavanjem upita nad bazom i dati su odgovori na pitanja koja su postavljena u tekstu zadatka:

1. Izlistati koliko je nekretnina za prodaju, a koliko za iznajmljivanje

SELECT COUNT(\*) as SellingCount

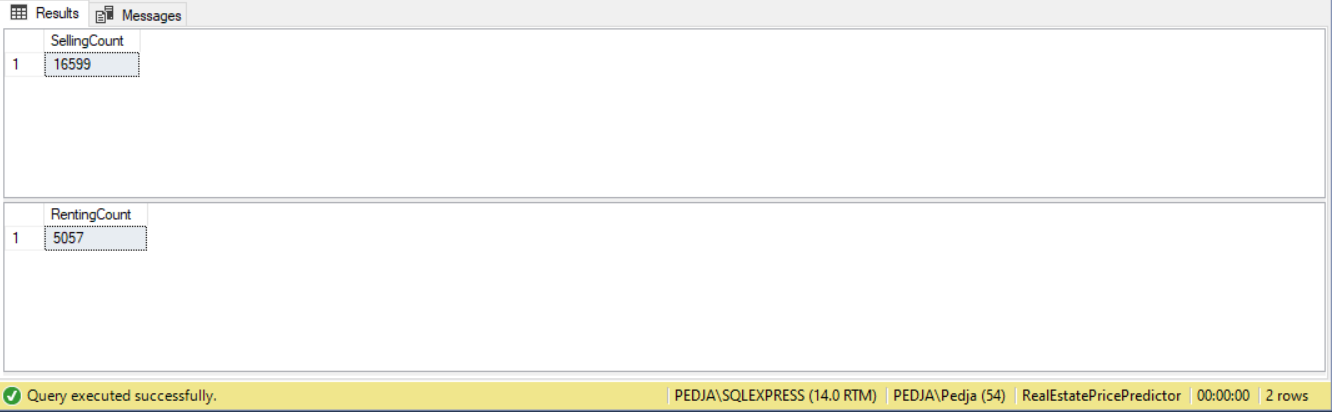
FROM RealEstates

WHERE OfferType = 0;

SELECT COUNT(\*) as RentingCount

FROM RealEstates

WHERE OfferType = 1;



1. Izlistati koliko nekretnina se prodaje u svakom od gradova (izlistati sve gradove, obuhvatiti i kuće i stanove

SELECT City, COUNT(\*) as SellingCount

FROM RealEstates

WHERE OfferType = 0

GROUP BY City

ORDER BY SellingCount DESC;



1. Izlistati koliko je uknjiženih, a koliko neuknjiženih kuća, a koliko stanova

SELECT COUNT(\*) as SellingNotRegistered

FROM RealEstates

WHERE HousingType = 0 AND Registered = 0;

SELECT COUNT(\*) as SellingRegistered

FROM RealEstates

WHERE HousingType = 0 AND Registered = 1;

SELECT COUNT(\*) as RentingNotRegistered

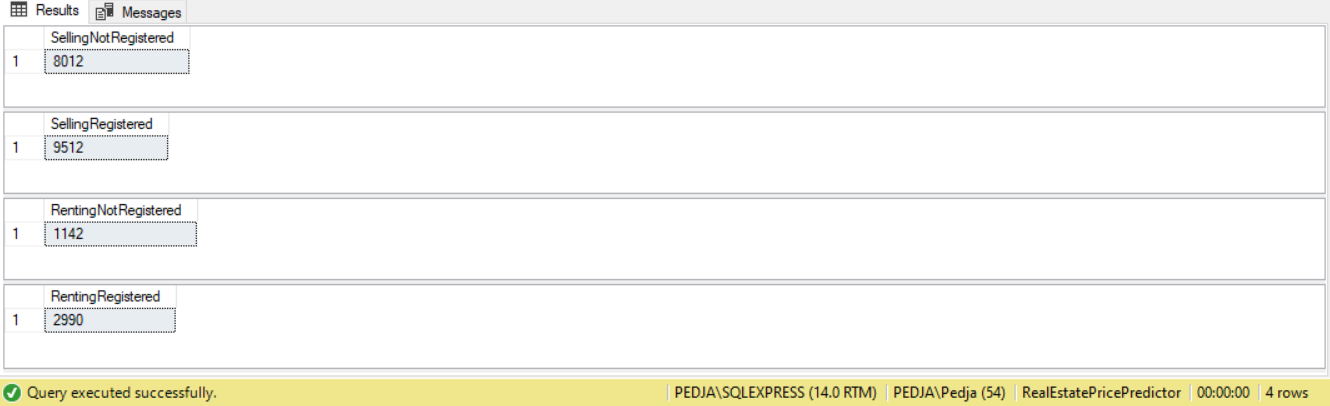
FROM RealEstates

WHERE HousingType = 1 AND Registered = 0;

SELECT COUNT(\*) as RentingRegistered

FROM RealEstates

WHERE HousingType = 1 AND Registered = 1;



1. Prikazati rang listu prvih 20 najskupljih kuća koje se prodaju, i 20 najskupljih stanova koji se prodaju u Srbiji

SELECT TOP(20) \*

FROM RealEstates

WHERE HousingType = 0

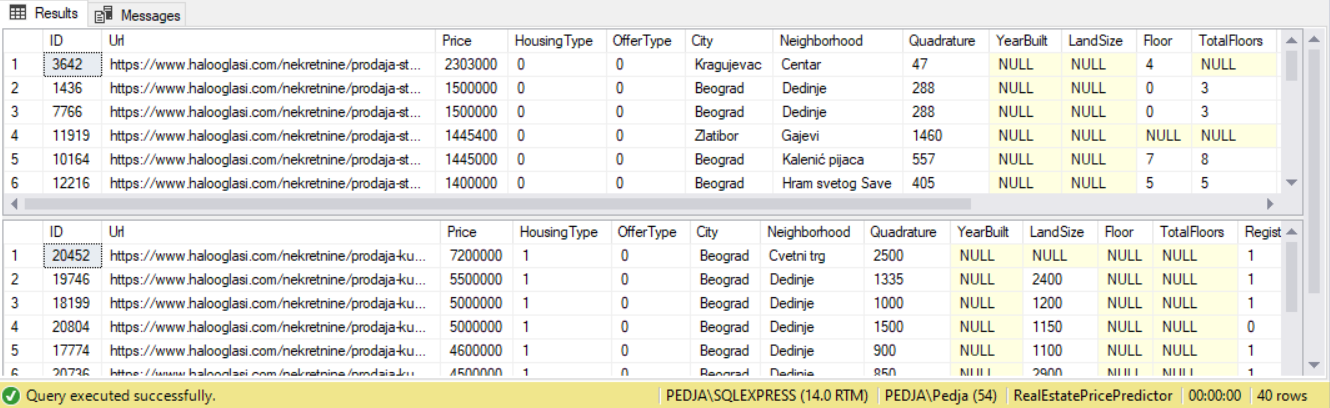
ORDER BY Price DESC; --stanovi

SELECT TOP(20) \*

FROM RealEstates

WHERE HousingType = 1

ORDER BY Price DESC; --kuce



1. Prikazati rang listu prvih 100 najskupljih kuća i 100 najskupljih stanova za iznajmljivanje (rentiranje)

SELECT TOP(100) \*

FROM RealEstates

WHERE HousingType = 0 AND OfferType = 1

ORDER BY Price DESC; --stanovi

SELECT TOP(100) \*

FROM RealEstates

WHERE HousingType = 1 AND OfferType = 1

ORDER BY Price DESC; --kuce



1. Prikazati rang listu svih nekretnina izgrađenih u 2019. godini, i izlistati ih opadajuće prema ceni prodaje, odnosno ceni iznajmljivanja

SELECT \*

FROM RealEstates

WHERE YearBuilt = 2019 AND OfferType = 0

ORDER BY Price DESC; --stanovi

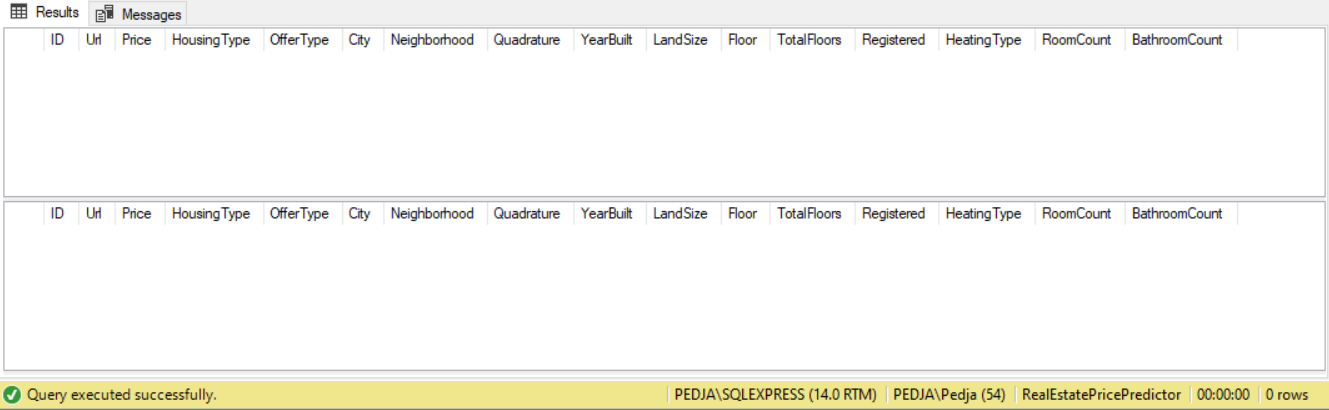
SELECT \*

FROM RealEstates

WHERE YearBuilt = 2019 AND OfferType = 1

ORDER BY Price DESC; --kuce

--Ne postoji podatak o godini izgradnje na website-u



1. Prikazati nekretnine koje imaju
   * najveći broj soba unutar nekretnine
   * najveći broj kupatila unutar nekretnine
   * najveću površinu zemljišta (samo za kuće)

SELECT \*

FROM RealEstates

WHERE RoomCount = (SELECT MAX(RoomCount) FROM RealEstates);

SELECT \*

FROM RealEstates

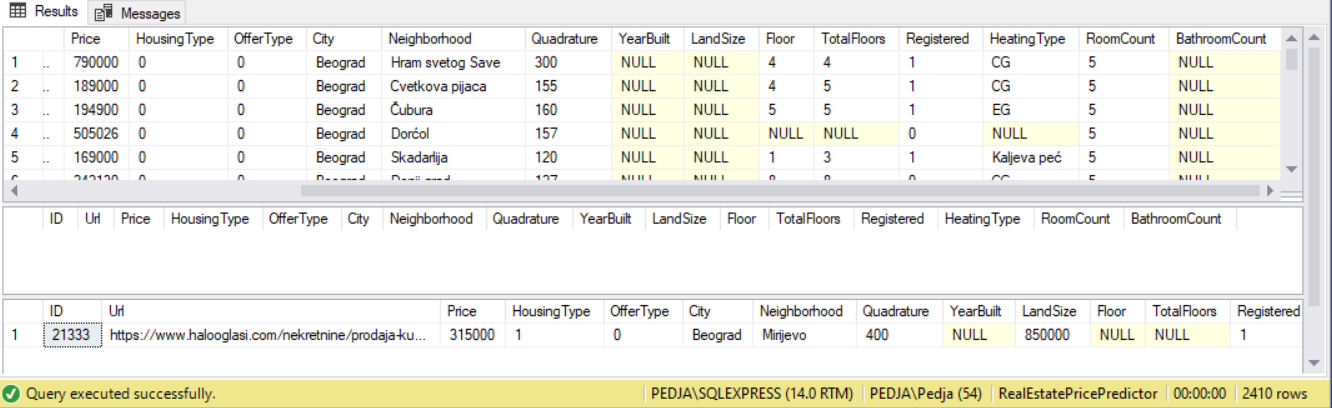
WHERE BathroomCount = (SELECT MAX(BathroomCount) FROM RealEstates);

--Ne postoji podatak o broju kupatila na website-u

SELECT \*

FROM RealEstates

WHERE HousingType = 1 AND LandSize = (SELECT MAX(LandSize) FROM RealEstates);



# Vizuelizacija podataka

Pri vizuelizaciji podataka su korišćeni upiti napisani u jeziku LINQ, ali je zbog jednostavnosti prikazan odgovarajući SQL upit koji bi dao isti rezultat.

Dati su odgovori na pitanja koja su postavljena u tekstu zadatka:

1. 8 najzastupljenijih delova Beograda koji imaju najveći broj nekretnina u ponudi (i u sekciji za prodaju, i u sekciji za iznajmljivanje, zbirno)

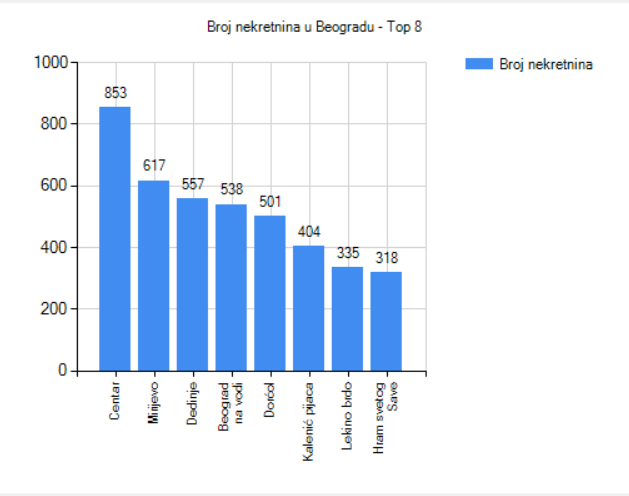
SELECT TOP(8) Neighborhood, COUNT(\*) as TotalCount

FROM RealEstates

WHERE City = 'Beograd'

GROUP BY Neighborhood

ORDER BY TotalCount DESC;



1. Broj stanova za prodaju prema kvadraturi, u celoj Srbiji (do 35 kvadrata, 36-50, 51-65, 66-80, 81-95, 96-110, 111 kvadrata i više)

SELECT '-35' as Quadrature, COUNT(\*) as TotalCount FROM RealEstates

WHERE HousingType = 0 AND OfferType = 0 AND Quadrature < 36

UNION

SELECT '36-50', COUNT(\*) FROM RealEstates

WHERE HousingType = 0 AND OfferType = 0 AND Quadrature BETWEEN 36 AND 50

UNION

SELECT '51-65', COUNT(\*) FROM RealEstates

WHERE HousingType = 0 AND OfferType = 0 AND Quadrature BETWEEN 51 AND 65

UNION

SELECT '66-80', COUNT(\*) FROM RealEstates

WHERE HousingType = 0 AND OfferType = 0 AND Quadrature BETWEEN 66 AND 80

UNION

SELECT '81-95', COUNT(\*) FROM RealEstates

WHERE HousingType = 0 AND OfferType = 0 AND Quadrature BETWEEN 81 AND 95

UNION

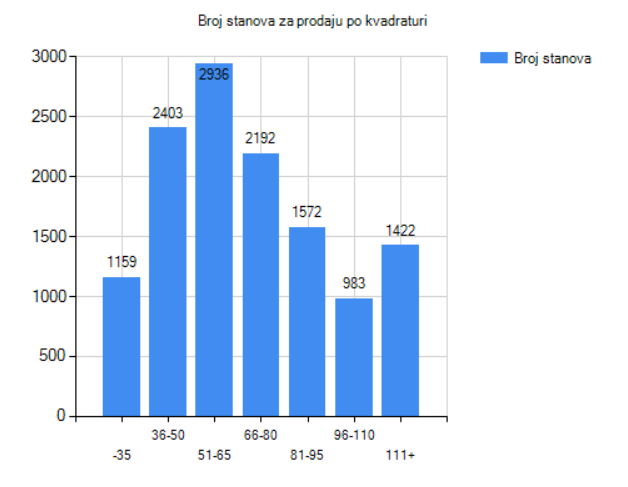
SELECT '96-110', COUNT(\*) FROM RealEstates

WHERE HousingType = 0 AND OfferType = 0 AND Quadrature BETWEEN 96 AND 110

UNION

SELECT '111+', COUNT(\*) FROM RealEstates

WHERE HousingType = 0 AND OfferType = 0 AND Quadrature > 110;



1. Broj izgrađenih nekretnina po dekadama (1950-1959, 1960-1969, 1970-1979, 1980-1989, 1990-1999, 2000-2009, 2010-2019), a obuhvatiti i sekcije za prodaju i za iznajmljivanje

SELECT '1950-1959' as Decade, COUNT(\*) as TotalCount FROM RealEstates

WHERE YearBuilt BETWEEN 1950 AND 1959

UNION

SELECT '1960-1969', COUNT(\*) FROM RealEstates

WHERE YearBuilt BETWEEN 1960 AND 1969

UNION

SELECT '1970-1979', COUNT(\*) FROM RealEstates

WHERE YearBuilt BETWEEN 1970 AND 1979

UNION

SELECT '1980-1989', COUNT(\*) FROM RealEstates

WHERE YearBuilt BETWEEN 1980 AND 1989

UNION

SELECT '1990-1999', COUNT(\*) FROM RealEstates

WHERE YearBuilt BETWEEN 1990 AND 1999

UNION

SELECT '2000-2009', COUNT(\*) FROM RealEstates

WHERE YearBuilt BETWEEN 2000 AND 2009

UNION

SELECT '2010-2019', COUNT(\*) FROM RealEstates

WHERE YearBuilt BETWEEN 2010 AND 2019

--Ne postoji podatak o godini izgradnje na website-u

1. Broj (i procentualni odnos) nekretnina koje se prodaju i nekretnina koje se iznajmljuju, za prvih 5 gradova sa najvećim brojem nekretnina (za svaki grad posebno prikazati grafikon BROJ\_ZA\_PRODAJU : BROJ\_ZA\_IZNAJMLJIVANJE)

SELECT City, OfferType, COUNT(\*) as TotalCount FROM RealEstates

WHERE City IN (

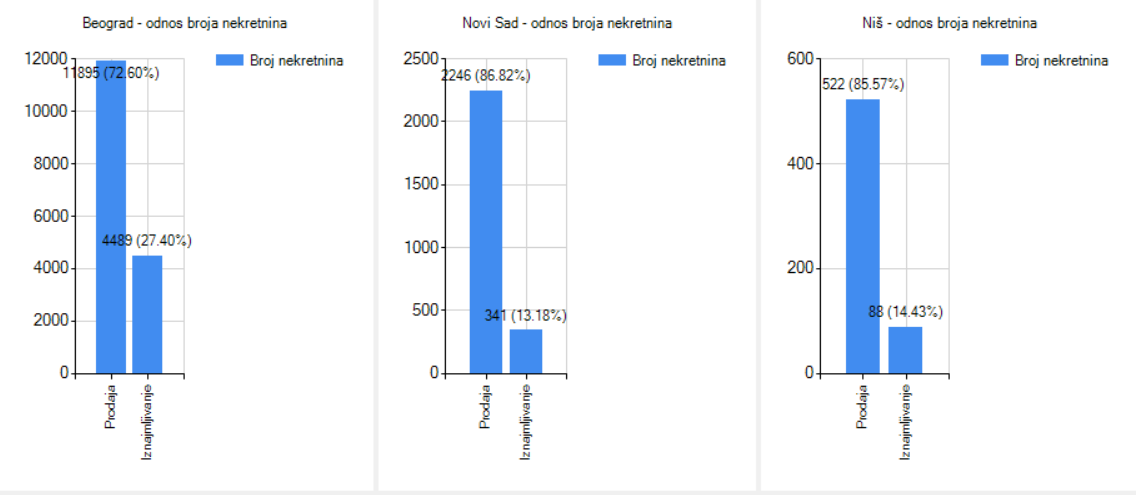
SELECT TOP(5) City FROM RealEstates

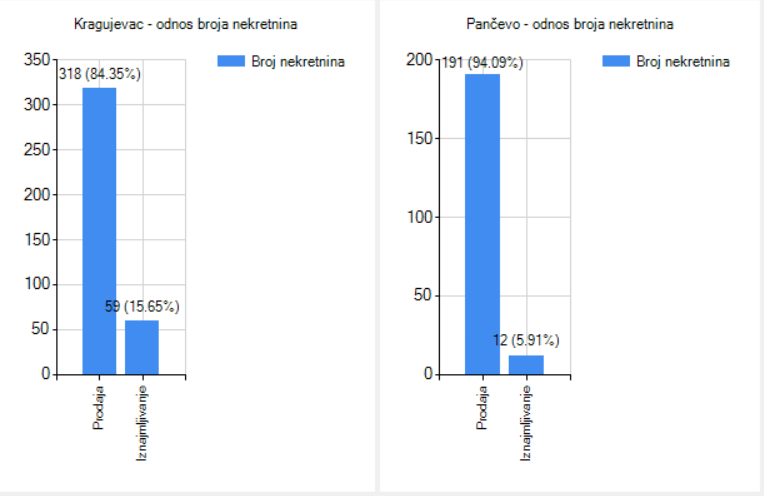
GROUP BY City

ORDER BY COUNT(\*) DESC

)

GROUP BY City, OfferType;





1. Broj (i procentualni odnos) svih nekretnina za prodaju, koje po ceni pripadaju jednom od sledećih opsega:

* manje od 49 999 €
* između 50 000 i 99 999 €
* između 100 000 i 149 999€
* između 150 000 €i 199 999 €
* 200 000 € ili više

SELECT '-49999' as Price, COUNT(\*) as TotalCount

FROM RealEstates WHERE OfferType = 0 AND Price < 50000

UNION

SELECT '50000-99999', COUNT(\*)

FROM RealEstates WHERE OfferType = 0 AND Price BETWEEN 50000 AND 99999

UNION

SELECT '100000-149999', COUNT(\*)

FROM RealEstates WHERE OfferType = 0 AND Price BETWEEN 100000 AND 149999

UNION

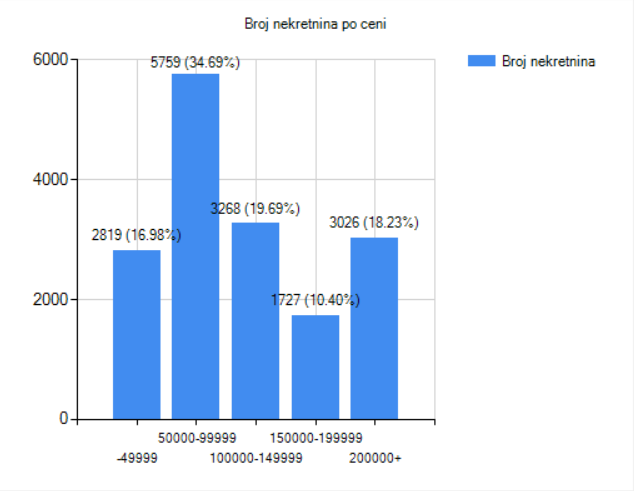
SELECT '150000-199999', COUNT(\*)

FROM RealEstates WHERE OfferType = 0 AND Price BETWEEN 150000 AND 199999

UNION

SELECT '200000+', COUNT(\*)

FROM RealEstates WHERE OfferType = 0 AND Price > 199999;

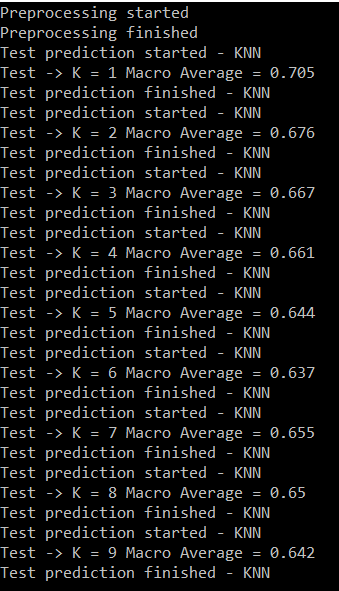


# Implementacija regresije

Prvi korak je bio izdvajanje stanova koji su na prodaju u Beogradu iz tabele RealEstate i njihovo slučajno mešanje. Zatim je 20% stanova odvojeno za skup za testiranje. Nakon toga su stanovima pridružene vrednosti iz tabele NeighborhoodDistances koje označavaju koliko je stan udaljen od centra grada. Na početku su kao feature-i odabrani distanca od centra grada, kvadratura, sprat, broj soba i uknjiženost. Kasnije je utvrđeno da uknjiženost ima mali uticaj na cenu, pa je taj feature uklonjen. Feature-i su normalizovani, odnosno sve vrednosti su skalirane da budu u opsegu između 0 i 1. Pre početka treninga iz preprocesiranog trening skupa je odvojeno 20% za validacioni skup, a sa ostalim podacima je započet trening. Trening traje sve do trenutka kada RMSE za validacioni skup ne počne da raste. Za learning rate su testirani opsezi između 10-6 do 103 i najbolji rezultat je postignut kada je learning rate između 0 i 1. Na kraju je odabrano da learning rate bude jednak 0,5. Kao metrika je korišćen RMSE i najbolji postignut rezultat na test skupu je ~66000. Obično se RMSE kreće u opsegu od 70000-90000.

# Implementacija klasifikacije

Preprocesiranje je identično kao u četvrtom zadatku, osim što u ovom slučaju ne postoji validacioni skup. Implementirane su tri funkcije za računanje distance: Euklidska, Menhetn i Čebišev. Najbolje se pokazala Menhetn distanca koja ima veoma slične rezultate kao Euklidska, dok se Čebišev distanca osetno lošije pokazala. Kao metrika je korišćen MacroAverage i najbolji postignut rezultat na test skupu je 0,74 za K = 1. Testirano je kako se ponaša klasifikator sa različitim vrednostima K. Pokazalo se da klasifikator sa manjim vrednostima K bolje predviđa klase. Na slici ispod se može videti primer gde je K između 1 i 10.



Sa povećanjem K do preporučene vrednosti (koren broja instanci iz trening skupa) K opada i zadržava se u opsegu 0,55 – 0,6. Za sada je ostavljena preporučena vrednost, a korisnik može preko standardnog ulaza promeniti vrednost K.

# Zaključak

Na slici ispod se može videti ispis nakon pokretanja programa koji komunicira sa korisnikom preko standardnog ulaza. RMSE na trening skupu je 85178, na validacionom skupu 76298, a na test skupu 88038. Za K = 89 MacroAverage vrednost je 0,557. Težina uz distancu je negativna i veće su težine dodeljene uz kvadraturu i broj soba (u odnosu na sprat) što su dva feature-a koja su usko korelisana.

