SVEUČILIŠTE JURJA DOBRILE U PULI

FAKULTET INFORMATIKE U PULI

Skladišta i rudarenje podataka

**PRODAJA NEKRETNINA U MELBOURNE-U 2016 – 2018 GODINE**

SEMINARSKI RAD

Mateo Kocev, 0303104813, izvanredni student

Informatika

Pula, 22.5.2024.

### 2. Odabir dataseta / skupa podataka

Podaci su učitani iz CSV datoteke pronađene na poveznici: <https://www.kaggle.com/datasets/ronikmalhotra/melbourne-housing-dataset>

U skupu podataka pronalazimo razne informacije o prodaji nekretnina u periodu od veljače 2016. godine do ožujka 2018. godine koje na lokaciji imaju sagrađenu i kuću.

Skup podataka smo podijelili na dva dijela kako bi pokušali realizirati simulaciju realistične situacije gdje nam se podaci neće nalaziti u jednom izvoru, već je na nama da dostupne podatke obradimo i pripremimo za rad.

### 2.1. Opis problema i cilj analize

Tržište nekretnina u Melbourneu suočava se s nizom izazova koji utječu na trgovanje nekretninama. Cijene nekretnina mogu biti vrlo promjenjive te stvoriti nesigurnost u donošenju odluka kod kupaca, prodajnih agenata i investitora što dovodi do financijskog gubitka ili propuštenih prilika. Ovaj problem možemo riješiti analizom povijesnih podataka te identifikacijom trendova na tržištu pružajući uvid ne samo u opčenite trendove već gledajući performansu samih prodajnih agenata.

Problem nam također predstavlja veliku količinu faktora poput lokacije, veličine, dob nekretnine i dostupnih komoditeta koji mogu utjecati na cijene nekretnina. Teško je odrediti koji su najvažniji faktori koje moramo uzeti u obzir no kroz detaljnu analizu podataka ciljamo identificirati najveće utjecaje na kretanje tržišta kako bi mogli dovesti realne i konkurentne poslovne odluke.

Ciljamo optimizirati strategiju prodaje, kupnje i korištenja vremena kako bi minimizirali rizik te maksimizirali dobit financijskih sredstava donoseći pravilne poslovne odluke koristeći primjerene analitičke alate i modele. Također ciljamo održavati veću transparentnost i dostupnost informacija o često privatnim transakcijama kroz prikupljanje, saniranje i obrađivanje podataka kako bi naši prodajni agenti i investitori mogli dovesti preciznu poslovnu odluku.

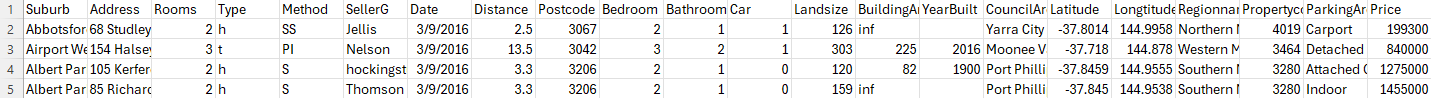
### 2.2. Analiza podataka

Skup podataka nam se sastoji o 22 atributa i ukupno 34,857 redaka gdje svaki individualni redak predstavlja zasebnu prodaju nekretnine. Skup podataka također sadrži prihvatljivu količinu kvalitativnih i kvantitativnih podataka među kojima nalazimo i veoma važnu vremensku dimenziju.

Atributi su sljedeći: Suburb, Address, Rooms, Type, Method, SellerG, Date, Distance, Postcode, Bedroom, Bathroom, Car, Landsize, BuildingArea, YearBuilt, CouncilArea, Latitude, Longitude, Regionname, Propertycount, ParkingArea, Price.

Po listi atributa primjećujemo nekoliko kategorija / vrsta atributa koje možemo podijeliti na sljedeći način:

* **Geografski atributi:** Suburb, Postcode, CouncilArea, Regionname, Latitude, Longitude predstavljaju neku lokaciju te opisuju razne aspekte lokacije u kojoj se nekretnina nalazi među kojima nalazimo regiju, kvart, poštanski broj i najvažnije koordinate koje nam daju preciznu informaciju o lokaciji.
* **Podaci o nekretnini:** Address, Landsize, buildingArea, YearBuilt, Propertycount, ParkingArea nam opisuju razne aspekte nekretnine koje bi potencijalni kupci uzimali u obzir za moguće modifikacije i buduće planove nadogradnje u reguliranom urbanom okolišu.
* **Podaci o kući:** Rooms, Type, Bedroom, Bathroom, Car nam opisuju detalje koji potencijalnim kupcima početno padaju na um te nam govore o dostupnim komoditetima kuće.
* **Podaci o prodavačima:** Skup podataka sadrži samo jedan atribut koji pripada prodavaču SellerG te predstavlja ime prodavača.
* **Podaci o prodaji:** Date, Method i Price opisuju informacije o prodaji nekretnine.



Slika 1: Prikaz atributa u CSV datoteci

Nakon analize svakog atributa možemo identificirati potrebne dimenzije za postizanje našeg cilja. Atributi koje ćemo koristiti primarno za analizu tržišta nekretnina glase:

* Latitude, Longitude, RegionName će nam dati prosječnu i preciznu informaciju o lokaciji te će nam pokazati moguće trendove vezane za lokaciju nekretnine.
* Karakteristike nekretnine i kuće poput Bedroom, Bathroom, Car i Landsize nam mogu ukazati trendove oko ponuđenih komoditeta koje potencijalni kupci traže te YearBuilt i Type nam ukazuje preference na vrstu kuće koja se cilja u prodajama
* Method i SellerG nam ukazuje trendove u načinu prodaje te nam pokazuje preferencije na tržištu.
* Date nam dodaje vremenski faktor gdje možemo ugledati promjenu trendova i performansi kroz vrijeme
* Price je najvažniji atribut u skupu s time da nam povezuje atribute i omogućuje spajanje atributa u svrhu usporedbe trendova

Iako neke atribute nećemo koristiti zbog nedostatka ili nezadovoljavajuće kvalitete podataka, skup nam pruža više nego dovoljno opširan skup podataka za izvođenje kvalitetne analize. Osim potencijalno malene količine iskoristivih redaka zbog nedostajećih podataka, naš skup zadovoljava ostale aspekte potrebne za izvršavanje projekta.

### 3. Izrada relacijskog modela i baze podataka

Kako bi stvorili bazu podataka i relacijski model za pripremu analize tržišta nekretnina, prvo ćemo primjeniti alate dostupne u programskom jeziku Python kako bi analizirali stanje integriteta podataka u našoj CSV datoteci. Sa Python programskim jezikom imamo dostupnu biblioteku *pandas*, koja nam omogućava rad na našim podacima putem manipulacije pandas specifičnog objekta *DataFrame* u koji učitavamo naše podatke te putem kojeg možemo jednostavno izvesti analizu, čišćenje i modifikaciju podataka u našoj CSV datoteci.

Nakon analize skupa koristeći Python alate, prelazimo na modeliranje i implementaciju baze podataka koristeći *sqlalchemy* i *MySQL Workbench*. sqlalchemy je biblioteka dostupna u Pythonu koja omogućava deklarativno mapiranje modela, tj. definiranje modela baze podataka koristeći Python klase, omogućava upravljanje sesijama za interakciju s bazom podataka gdje možemo implementirati promjene i upravljati relacijama između modela. U ovoj situaciji ćemo primjeniti sqlalchemy kako bi implementirali naš relacijski model u MySQL bazu podataka učitavajući podatke iz CSV-a direktno u MySQL bazu.

### 3.1. Analiza i čišćenje skupa putem implementacije u Pythonu

Implementacijom sljedeće Python skripte smo analizirali, očistili te raspodijelili skup podataka na dva dijela kako bi simulirali realnu situaciju gdje nam podaci dolaze sa više izvora.

Korišteni alati:

* Poveznica na službenu stranicu Python dokumentacije: [docs.python.org/3/](https://docs.python.org/3/)
* Poveznica na službenu stranicu pandas dokumentacije: [pandas.pydata.org/docs/](https://pandas.pydata.org/docs/)

1. import pandas as pd

2.

3. CSV\_FILE\_PATH = "Melbourne\_housing.csv"

4. df = pd.read\_csv(CSV\_FILE\_PATH, delimiter=',')

5.

6. print("Veličina skupa (broj redaka, broj stupaca):", df.shape)

7.

8. print("Nazivi stupaca:", df.columns.tolist())

9.

10. for column in df.columns:

11.     print(f"Jedinstvene vrijednosti u stupcu {column}: {df[column].nunique()}")

12.

13. print("Tipovi podataka po stupcu:\n", df.dtypes)

14.

15. print("Broj nedostajućih vrijednosti po stupcu:\n", df.isna().sum())

16.

17. print("CSV size before: ", df.shape)

18.

19. df = df.dropna()

20. print("CSV size after: ", df.shape)

21. print(df.head())

22.

23. df20 = df.sample(frac=0.2, random\_state=1)

24. df = df.drop(df20.index)

25. print("CSV size 80: ", df.shape)

26. print("CSV size 20: ", df20.shape)

27.

28. df.to\_csv("Melbourne\_housing\_PROCESSED.csv", index=False)

29. df20.to\_csv("Melbourne\_housing\_PROCESSED\_20.csv", index=False)

U skripti pozivamo biblioteku pandas te učitavamo CSV datoteku sa lokalnog foldera (skripta se mora nalaziti u istom folderu da bi korištena ruta radila) u pandas DataFrame objekt.

Putem *.shape* metode provjeravamo veličinu učitanog DataFrame-a



Slika 2: Ispis veličine DF-a u konzoli

Koristeći *.tolist()* metodu na kolonama DF-a smo ispisali sve učitane atribute



Slika 3: Ispis atributa u konzoli

Iterirajući kroz kolone DF-a ispisujemo količinu jedinstvenih vrijednosti u skupu podataka. Koristeći *.nunique()* dobijamo količinu jedinstvenih vrijednosti na osi.

A screenshot of a computer

Description automatically generated

Slika 4: Ispis količine jedinstvenih vrijednosti po atributu

Koristeći *.dtypes* metodu provjeravamo vrstu podataka koje nalazimo među atributima

A screenshot of a computer

Description automatically generated

Slika 5: Prikaz vrsta podataka u DF-u

Date je zapisan kao object umjesto podržane jedinice vremena u pandas DF-u no to se lagano ispravi prije ubacivanja u bazu podataka.

U zadnjem koraku analize provjeravamo količinu nedostajućih vrijednosti po stupcu kostiteći *.isna()* metodu te sumirajući vrijednosti po atributima.

A screenshot of a computer

Description automatically generated

Slika 6: Prikaz količine nedostajećih podataka po atributu

U sljedećem koraku provjeravamo početnu veličinu DF-a te brišemo sve retke sa null vrijednostima. Tijekom provjere preostalih podataka uočavamo da nam je ostalo oko 9000 preciznih podataka o prodajama. Iako smo izgubili veliki dio podataka, čisti i ispunjeni skup podataka će nam pružati preciznije rezultate tijekom analize.

Za kraj nasumično dijelimo skup podataka na 80% I 20% te ih spremamo u zasebne CSV datoteke koje se nalaze na lokalnoj ruti.