

# Naslov seminara

Ime i prezime

Fakultet / Predmet

22. svibnja 2025.

# Sadržaj

1	Uvod	2
2	Poslovna inteligencija	2
3	Skladišta podataka	2
4	Pronalazak i analiza dataseta	2
5	Inicijalna analiza podataka u Pythonu	3
6	Relacijski model podataka	7
7	Dimenzijski model podataka	8
8	Apache Spark, stvaranje i punjenje dimenzijskog modela	8
9	Interaktivni web-based dashboard i vizualizacija podataka	8

## 1. Uvod

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetur id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

## 2. Poslovna inteligencija

Nam dui ligula, fringilla a, euismod sodales, sollicitudin vel, wisi. Morbi auctor lorem non justo. Nam lacus libero, pretium at, lobortis vitae, ultricies et, tellus. Donec aliquet, tortor sed accumsan bibendum, erat ligula aliquet magna, vitae ornare odio metus a mi. Morbi ac orci et nisl hendrerit mollis. Suspendisse ut massa. Cras nec ante. Pellentesque a nulla. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Aliquam tincidunt urna. Nulla ullamcorper vestibulum turpis. Pellentesque cursus luctus mauris.

## 3. Skladišta podataka

Nulla malesuada porttitor diam. Donec felis erat, congue non, volutpat at, tincidunt tristique, libero. Vivamus viverra fermentum felis. Donec nonummy pellentesque ante. Phasellus adipiscing semper elit. Proin fermentum massa ac quam. Sed diam turpis, molestie vitae, placerat a, molestie nec, leo. Maecenas lacinia. Nam ipsum ligula, eleifend at, accumsan nec, suscipit a, ipsum. Morbi blandit ligula feugiat magna. Nunc eleifend consequat lorem. Sed lacinia nulla vitae enim. Pellentesque tincidunt purus vel magna. Integer non enim. Praesent euismod nunc eu purus. Donec bibendum quam in tellus. Nullam cursus pulvinar lectus. Donec et mi. Nam vulputate metus eu enim. Vestibulum pellentesque felis eu massa.

## 4. Pronalazak i analiza dataseta

U procesu pronalaska dataseta isprobano je nekoliko standardnih izvora poput Kagglea, no većina tema nije bila dovoljno interesantna ili pak nije bilo dovoljno redaka u datasetu što bi onemogućilo ozbiljnije daljnje obrade i ne bi pružilo adekvatne analitičke uvide. Naposljetku, odabran je dataset o portugalskim parlamentarnim izborima naen na UC Irvine Machine Learning Repositoryju: <https://archive.ics.uci.edu/dataset/513/>

Ovaj dataset zadovoljavao je više ključnih uvjeta: dovoljan broj redaka i atributa, dostupnost podataka za različite teritorijalne jedinice, mogućnost vremenske analize

te potencijal za dimenzijsko modeliranje po strankama, teritorijima i vremenu. Osim toga, tema izbora se pokazala zanimljivom za dublju obradu, vizualizaciju i izradu interaktivnog dashboarda.

Nakon odabira i preuzimanja dataseta, pristupilo se inicijalnoj obradi i upoznavanju s podacima koristeći Python u Google Colab notebook okruženju. U početku je skripta ispisala poruku "Hello World" radi provjere ispravnosti Python okruženja. Sljedeći korak bio je učitavanje biblioteke Pandas (Python and Data Science), koja se koristi za obradu i analizu strukturiranih podataka u obliku dataframeova.

Dataset je učitao iz CSV datoteke pomoću `pd.read_csv()`, pri čemu je prvotno analizirano samo prvih 2000 redaka kako bi se omogućio brzi pregled i izbjeglo opterećenje memorije. Ispisani su prvi redovi dataseta kako bi se dobio uvid u strukturu tablice i moguće varijable.

Daljnjom analizom ispisane su dimenzije cjelokupnog skupa podataka (`data.shape`), broj null vrijednosti po stupcima (`data.isna().sum()`), broj jedinstvenih vrijednosti po atributima (`data.nunique()`), kao i tipovi podataka po svakom stupcu (`data.dtypes`). Na kraju, korištenjem `value_counts()` za svaki stupac, dobiven je detaljan prikaz distribucije vrijednosti, što je pomoglo u identifikaciji potencijalnih dimenzija i metrika za kasnije faze modeliranja.

Za ostvaranje opisanih rezultata korišten je sljedeći kod:

## 5. Inicijalna analiza podataka u Pythonu

```
1 # Provjera da li je python instaliran na sustavu
2 print("Hello World")
3
4 # U itavanje potrebnih biblioteka
5 import pandas as pd
6
7 # U itavanje podataka iz CSV datoteke
8 PATH = "/content/ElectionData.csv"
9
10 data = pd.read_csv(PATH, delimiter=',')
11
12 # U itavanje prvih 2000 redova
13 data_first_2000 = pd.read_csv(PATH, delimiter=',', nrows=2000)
14
15 # Ispis prvih 5 redova
16 print(data_first_2000.head())
17
18 # Ispis dimenzija skupa podataka (potrebno je u itati sve podatke
19   radi analize, ne samo prvih 2000 redova)
20 print(data.shape)
21
22 # Ispis imena stupaca i nedostaju ih vrijednosti
23 print(data.isna().sum())
24
25 # Ispis broja jedinstvenih vrijednosti po stupcima
26 print(data.nunique())
```

```
27 # Ispis tipova podataka po stupcima (analiza kvanitativnih i
    kvalitativnih varijabli)
28 print(data.dtypes)
29
30 # Ispis broja jedinstvenih vrijednosti po stupcima
31 for column in data:
32     print(data[column].value_counts())
33     input("...")
34
35 # Ispis imena stupaca
36 print(data.columns.values)
```

Listing 1: Učitavanje i inicijalna analiza izbornog dataseta u Pythonu

Za dodatno utvrđivanje kvalitete dataseta bilo je potrebno ispuniti određene uvjete.



Slika 1: Dimenzije dataseta

Vidljivo je kako dataset ima 21643 retka i čak 28 stupaca što zadovoljava inicijalne uvjete veličine dataseta od bar 15000 redaka i 10 stupaca. Bilo je potrebno postaviti donju granicu veličine iz razloga što premali dataset ne bi pružao dobre mogućnosti analize jer se iz njega ne bi mogle izvući određene pravilnosti, korelacije i sl.

Nakon što je utvrđeno da je dataset dovoljno velik, ispisani su svi stupci i broj nedostajućih vrijednosti u njima. Poželjno je da dataset ima čim je manje mogući broj takvih vrijednosti zato što one smanjuju kvalitetu analize i ne pružaju potpuni uvid u promatrane entitetime. Nedostajuće vrijednosti mogu dovesti do iskrivljenih rezultata u statističkim analizama, jer se algoritmi mogu oslanjati na nepotpune ili neprecizne podatke. Osim toga, prisutnost null vrijednosti često zahtijeva dodatne korake obrade podataka poput popunjavanja nedostajućih vrijednosti, konverzije null vrijednosti u druge tipove (npr. pretvorba NULL u broj 0 ili u prazan string (" ")) ili filtriranja redaka, što može dodatno povećati kompleksnost analize i smanjiti interpretabilnost rezultata. U kontekstu poslovne inteligencije, takvi podaci mogu dovesti do donošenja pogrešnih poslovnih odluka ako se na njih ne obrati pažnja. Zbog svega navedenog, kvaliteta dataseta i njegova čistoća ključni su preduvjeti za izvođenje pouzdane i korisne analize. U ovom slučaju, dataset nije imao niti jednu nedostajuću vrijednost što je indikator njegove kvalitete i potpunosti i dodatna potvrda kako može biti korišten za implementaciju skladišta podataka.

```

TimeElapsed      0
time             0
territoryName     0
totalMandates     0
availableMandates 0
numParishes       0
numParishesApproved 0
blankVotes        0
blankVotesPercentage 0
nullVotes         0
nullVotesPercentage 0
votersPercentage  0
subscribedVoters  0
totalVoters       0
pre_blankVotes    0
pre_blankVotesPercentage 0
pre_nullVotes     0
pre_nullVotesPercentage 0
pre_votersPercentage 0
pre_subscribedVoters 0
pre_totalVoters   0
Party            0
Mandates         0
Percentage       0
validVotesPercentage 0
Votes           0
Hondt            0
FinalMandates    0
dtype: int64

```

Slika 2: Provjera broja nedostajućih vrijednosti

Nadalje, provjeren je i broj unique (jedinstvenih) vrijednosti u stupcima kako bi se utvrdila raznolikost podataka što je vrlo bitno u analizi zato jer raznolikost podataka izravno utječe na kvalitetu i dubinu analize koju možemo provesti. Stupci s velikim brojem jedinstvenih vrijednosti često sadrže ključne informacije koje omogućuju dublje razumijevanje podataka koji se analiziraju. Na primjer, atribut s mnogo različitih vrijednosti može predstavljati identifikatore poput naziva teritorija, političkih stranaka ili vremenskih oznaka, što nam omogućuje detaljnu analizu po tim dimenzijama.

S druge strane, stupci s vrlo malo jedinstvenih vrijednosti često predstavljaju kategorijske varijable koje se mogu koristiti za grupiranje podataka, što je korisno pri stvaranju dimenzijskih modela ili pri agregiranju podataka za vizualizaciju. U poslovnoj inteligenciji, razumijevanje strukture i raznolikosti podataka ključno je za izgradnju korisnih izvještaja, preciznih modela i donošenje odluka temeljenih na podacima. Analiza jedinstvenih vrijednosti pomaže i pri otkrivanju potencijalnih problema poput kodiranih vrijednosti koje predstavljaju iste entitete (npr. "HR" i "Hrvatska"), što može dovesti do neujednačenosti i netočnosti u interpretaciji zato što se ponekad isti real-life entitet može u datasetu evidentirati na više različitih načina.

```

TimeElapsed      54
time             54
territoryName     21
totalMandates     62
availableMandates 69
numParishes       98
numParishesApproved 219
blankVotes        329
blankVotesPercentage 146
nullVotes         331
nullVotesPercentage 187
votersPercentage  282
subscribedVoters  335
totalVoters       336
pre_blankVotes    323
pre_blankVotesPercentage 138
pre_nullVotes     329
pre_nullVotesPercentage 98
pre_votersPercentage 278
pre_subscribedVoters 331
pre_totalVoters   331
Party            21
Mandates         67
Percentage       1363
validVotesPercentage 1387
Votes           4829
Hondt            41
FinalMandates    17
dtype: int64

```

Slika 3: Provjera broja jedinstvenih vrijednosti

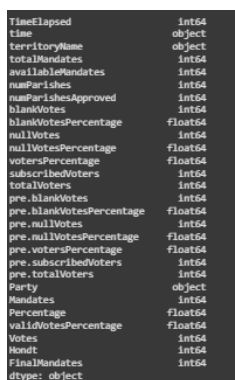
Vidljivo je kako svaki stupac ima zadovoljavajuć broj jedinstvenih vrijednosti što upućuje na to da dataset pruža dovoljno raznolikih informacija za provođenje kvalitetne analize. To znači da podaci nisu redundantni niti previše uniformni, što je čest

problem u manje kvalitetnim datasetima. Zadovoljavajući broj jedinstvenih vrijednosti omogućuje nam da uspješno razlikujemo, klasificiramo i grupiramo entitete po relevantnim kriterijima te izgradnju smislenih hijerarhija i veza unutar dimenzijskog modela.

Na primjer, ako stupac koji označava političke stranke ima dovoljno različitih vrijednosti, možemo analizirati ponašanje birača prema različitim strankama, promatrati distribuciju glasova kroz teritorije i vrijeme, te povezati rezultate s demografskim ili geografskim faktorima. Slično tome, stupac s nazivima teritorija omogućava prostornu analizu rezultata izbora, što je korisno za lokalizirano donošenje zaključaka i izradu ciljanih vizualizacija. Svaki će stupac biti detaljnije objašnjen u daljnjoj analizi.

U idućem koraku bilo je potrebno provjeriti tipove podataka u svakom stupcu kako bi se utvrdilo sadrži li dataset i kvalitativne i kvantitativne podatke. Kvantitativni podaci odnose se na brojeve dok kvalitativni podaci mogu biti u raznim formatima, no ovdje se specifično radi o stringovima. Kvantitativni podaci, poput broja glasova, omogućuju izvođenje statističkih analiza, izračun prosjeka, suma, postotaka te primjenu naprednijih metoda poput regresije ili klasifikacije. S druge strane, kvalitativni podaci – poput imena političkih stranaka ili teritorija – služe za kategorizaciju, grupiranje i filtriranje, što je osnova za stvaranje dimenzija u dimenzijskom modelu.

Identifikacija tipova podataka pomaže i u pripremi za obradu i vizualizaciju. Na primjer, numeričke vrijednosti mogu se prikazati u obliku grafikona, dok se kategorijske vrijednosti češće koriste za filtriranje rezultata u dashboardima. Također, razumijevanje strukture podataka omogućuje rano otkrivanje potencijalnih pogrešaka – primjerice, ako bi se brojevi podaci greškom učitali kao tekstualni, bilo bi onemogućeno izvođenje matematičkih operacija nad njima.



TimeElapsed	int64
time	object
territoryName	object
totalMandates	int64
availableMandates	int64
numParishes	int64
numParishesApproved	int64
blankVotes	int64
blankVotesPercentage	float64
nullVotes	int64
nullVotesPercentage	float64
votersPercentage	float64
subscribedVoters	int64
totalVoters	int64
pre-blankVotes	int64
pre-blankVotesPercentage	float64
pre-nullVotes	int64
pre-nullVotesPercentage	float64
pre-votersPercentage	float64
pre-subscribedVoters	int64
pre-totalVoters	int64
Party	object
Mandates	int64
Percentage	float64
validVotesPercentage	float64
Votes	int64
Hondt	int64
FinalMandates	int64
dtype:	object

Slika 4: Provjera tipova podataka

Iz analize je vidljivo kako prevladavaju numeričke vrijednosti (cjelobrojne i decimalne), ali i da postoje pojedini "object" stupci. U ovom slučaju object predstavlja stringove i timestampove.

Na samom kraju ispisana su sva imena stupaca i primjer prvih nekoliko redaka data-seta. Također, za shvaćanje konteksta samih podataka, potrebno je razjasniti što svaki pojedini stupac predstavlja, pa tako, redom:

Stupac	Opis
TimeElapsed	Vrijeme proteklo od početka brojanja glasova.
time	Točno vrijeme kada su podaci zabilježeni (timestamp).
territoryName	Naziv izborne jedinice (u konkretnom slučaju, radi se o saveznom državama unutar Portugala).
totalMandates	Ukupan broj mandata koji se dodjeljuju u teritoriju.
availableMandates	Broj još neraspodijeljenih mandata.
numParishes	Ukupan broj biračkih mjesta.
numParishesApproved	Broj biračkih mjesta čiji su rezultati obrađeni.
blankVotes	Broj praznih (nepopunjenih) listića.
blankVotesPercentage	Postotak praznih listića.
nullVotes	Broj nevažećih (nepravilno ispunjenih) listića.
nullVotesPercentage	Postotak nevažećih listića.
votersPercentage	Postotak birača koji su glasali.
subscribedVoters	Ukupan broj registriranih birača.
totalVoters	Ukupan broj birača koji su izašli na izbore.
pre.blankVotes	Broj praznih listića u prethodnom izvještaju.
pre.blankVotesPercentage	Postotak praznih listića iz prethodnog izvještaja.
pre.nullVotes	Broj nevažećih listića iz prethodnog izvještaja.
pre.nullVotesPercentage	Postotak nevažećih listića iz prethodnog izvještaja.
pre.votersPercentage	Odaziv birača u prethodnom izvještaju.
pre.subscribedVoters	Broj registriranih birača u prethodnom izvještaju.
pre.totalVoters	Broj birača koji su glasali u prethodnom izvještaju.
Party	Naziv političke stranke na koju se podaci odnose.

Tablica 1: Opis svakog stupca u skupu podataka

Isto tako, važno je istaknuti da su finalnom dimenzijskom modelu dodani neki novi atributi kojih nije bilo u ovom csv-u kako bi se omogućila bolja analiza s dubljim uvidom u rezultate izbora. Po završetku analiziranja dataseta i dodatne potvrde da je zaista dobar, nastavljena je daljna izrada projekta. Idući korak bio je izrada relacijskog modela.

## 6. Relacijski model podataka

Fusce mauris. Vestibulum luctus nibh at lectus. Sed bibendum, nulla a faucibus semper, leo velit ultricies tellus, ac venenatis arcu wisi vel nisl. Vestibulum diam. Aliquam pellentesque, augue quis sagittis posuere, turpis lacus congue quam, in hendrerit risus eros eget felis. Maecenas eget erat in sapien mattis porttitor. Vestibulum porttitor. Nulla facilisi. Sed a turpis eu lacus commodo facilisis. Morbi fringilla, wisi in dignissim interdum, justo lectus sagittis dui, et vehicula libero dui cursus dui. Mauris tempor ligula sed lacus. Duis cursus enim ut augue. Cras ac magna. Cras nulla. Nulla egestas. Curabitur a leo. Quisque egestas wisi eget nunc. Nam feugiat lacus vel est. Curabitur consetetuer.



## 7. Dimenzijski model podataka

Suspendisse vel felis. Ut lorem lorem, interdum eu, tincidunt sit amet, laoreet vitae, arcu. Aenean faucibus pede eu ante. Praesent enim elit, rutrum at, molestie non, nonummy vel, nisl. Ut lectus eros, malesuada sit amet, fermentum eu, sodales cursus, magna. Donec eu purus. Quisque vehicula, urna sed ultricies auctor, pede lorem egestas dui, et convallis elit erat sed nulla. Donec luctus. Curabitur et nunc. Aliquam dolor odio, commodo pretium, ultricies non, pharetra in, velit. Integer arcu est, nonummy in, fermentum faucibus, egestas vel, odio.

## 8. Apache Spark, stvaranje i punjenje dimenzijskog modela

Sed commodo posuere pede. Mauris ut est. Ut quis purus. Sed ac odio. Sed vehicula hendrerit sem. Duis non odio. Morbi ut dui. Sed accumsan risus eget odio. In hac habitasse platea dictumst. Pellentesque non elit. Fusce sed justo eu urna porta tincidunt. Mauris felis odio, sollicitudin sed, volutpat a, ornare ac, erat. Morbi quis dolor. Donec pellentesque, erat ac sagittis semper, nunc dui lobortis purus, quis congue purus metus ultricies tellus. Proin et quam. Class aptent taciti sociosqu ad litora torquent per conubia nostra, per inceptos hymenaeos. Praesent sapien turpis, fermentum vel, eleifend faucibus, vehicula eu, lacus.

## 9. Interaktivni web-based dashboard i vizualizacija podataka

obavezno reć...bila bi šteta koristiti postojeće dashboard alate koji su prilično ograničeni i zato je izrađen custom specifično za ovu namjenu Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Donec odio elit, dictum in, hendrerit sit amet, egestas sed, leo. Praesent feugiat sapien aliquet odio. Integer vitae justo. Aliquam vestibulum fringilla lorem. Sed neque lectus, consectetur at, consectetur sed, eleifend ac, lectus. Nulla facilisi. Pellentesque eget lectus. Proin eu metus. Sed porttitor. In hac habitasse platea dictumst. Suspendisse eu lectus. Ut mi mi, lacinia sit amet, placerat et, mollis vitae, dui. Sed ante tellus, tristique ut, iaculis eu, malesuada ac, dui. Mauris nibh leo, facilisis non, adipiscing quis, ultrices a, dui.