

# Naslov seminara

Ime i prezime

Fakultet / Predmet

29. svibnja 2025.

# Sadržaj

1	Uvod	2
2	Poslovna inteligencija	3
3	Skladišta podataka	4
4	Pronalazak i analiza dataseta	5
5	Inicijalna analiza podataka u Pythonu	5
6	Relacijski model podataka	9
7	Dimenzijski model podataka	14
8	Apache Spark, stvaranje i punjenje dimenzijskog modela	16
9	Interaktivni web-based dashboard i vizualizacija podataka	16

## 1. Uvod

Politički sustavi diljem svijeta generiraju ogromne količine podataka tijekom izbora – od registracije birača, izlaznosti, rezultata po strankama i kandidatima, do povijesnih trendova i promjena u izbornim jedinicama. Takvi podaci predstavljaju vrijedan resurs za analizu ponašanja birača, evaluaciju učinkovitosti izbornih procesa te donošenje strateških odluka političkih aktera i institucija koje nadziru izborni proces. Međutim, sama količina i kompleksnost podataka predstavljaju izazov u pogledu njihove obrade, skladištenja i analitičke interpretacije.

Cilj ovog seminara je prikazati cjelokupni proces izgradnje sustava za analizu izbornih rezultata temeljenog na skladištu podataka. U okviru rada razvijen je dimenzijski model koji omogućuje učinkovito pohranjivanje i dohvat podataka relevantnih za parlamentarne izbore, uključujući stranke, biračke jedinice, vremenske dimenzije, povijesne promjene i druge povezane entitete. Poseban naglasak stavljen je na korištenje naprednih koncepata poput sporo mijenjajućih dimenzija (SCD) i degradiranih dimenzija, kako bi se omogućila dugoročna analitika i točno praćenje promjena kroz vrijeme.

Nakon oblikovanja podatkovnog modela, pristupilo se implementaciji interaktivnog web-based dashboarda koji korisnicima omogućuje vizualno istraživanje podataka i donošenje zaključaka temeljenih na realnim brojkama i trendovima. Umjesto korištenja gotovih alata za dashboard analizu poput Power BI-ja ili Tableaui, odlučeno je da se razvije vlastito rješenje u tehnologijama otvorenog koda — Vue.js, Matplotlib i Metabase — kako bi se postigla maksimalna fleksibilnost i prilagodba specifičnim zahtjevima domene izbornih podataka. Gotova rješenja, iako moćna, često imaju ograničenja kada je riječ o povezivanju s kompleksnim podatkovnim modelima i implementaciji interaktivnosti prilagođene korisnicima.

Kroz ovaj seminar objašnjeni su teorijski temelji i praktična realizacija sustava, počevši od modeliranja i implementacije baze podataka, preko ETL procesa, pa sve do krajnje vizualizacije podataka putem modernog web sučelja. Time se ne prikazuje samo tehnička strana izgradnje skladišta podataka, već i važnost kvalitetnog dizajna, razumijevanja domene i potrebe za dostupnom i razumljivom prezentacijom informacija krajnjim korisnicima.

## 2. Poslovna inteligencija

Poslovna inteligencija (engl. *Business Intelligence*, BI) obuhvaća skup tehnologija, procesa i metodologija kojima se organizacijama omogućuje prikupljanje, integracija, analiza i prezentacija poslovnih podataka u svrhu donošenja informiranih odluka. Ključna svrha BI sustava je pretvaranje sirovih podataka u korisne informacije koje pomažu menadžmentu u prepoznavanju obrazaca, trendova i prilika na tržištu. U modernim poslovnim okruženjima, BI ne podrazumijeva samo izvještavanje i analizu povijesnih podataka, već uključuje i prediktivne modele, vizualizaciju podataka te integraciju s operativnim sustavima u stvarnom vremenu. Korištenjem naprednih analitičkih alata i vizualizacijskih tehnika, poslovna inteligencija postaje neizostavan dio strateškog planiranja i operativnog upravljanja u brojnim industrijama.

### 3. Skladišta podataka

Skladište podataka (engl. *Data Warehouse*) predstavlja centralizirani repozitorij strukturiranih podataka namijenjen analitičkoj obradi, donošenju odluka i podršci poslovnoj inteligenciji. Za razliku od operativnih baza podataka koje su optimizirane za česte transakcije i ažuriranja, skladišta podataka su dizajnirana za čitanje velikih količina povijesnih podataka, agregaciju, kompleksne upite i izvještavanje. Podaci u skladište dolaze iz različitih izvora putem procesa ekstrakcije, transformacije i učitavanja (ETL), pri čemu se čiste, usklađuju i konsolidiraju kako bi omogućili jedinstvenu i pouzdanu analizu. Arhitektura skladišta podataka obično se temelji na višeslojnom modelu, uključujući sloj izvora podataka, ETL proces, centralno skladište i sloj prezentacije kroz kojeg krajnji korisnici pristupaju podacima. Unutar samog skladišta podataka često se koristi dimenzijski model koji omogućuje učinkovito izvođenje analitičkih upita kroz tablice činjenica i dimenzija. Korištenjem skladišta podataka organizacije mogu prepoznati trendove, identificirati poslovne prilike te podržati strateško planiranje temeljem pouzdanih i sveobuhvatnih podataka. Suvremena skladišta podataka sve češće uključuju i elemente distribuiranih sustava, pohranu u oblaku te podršku za polustrukturirane podatke, čime proširuju mogućnosti analize u raznim kontekstima i skalabilnim okruženjima.

## 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 učitati sve podatke
    radi analize, ne samo prvih 2000 redova)
19 print(data.shape)
20
21 # Ispis imena stupaca i nedostaju ih vrijednosti
22 print(data.isna().sum())
23
24 # Ispis broja jedinstvenih vrijednosti po stupcima
25 print(data.nunique())
26
27 # Ispis tipova podataka po stupcima (analiza kvantitativnih 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.



(21643, 28)

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 entitete. 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

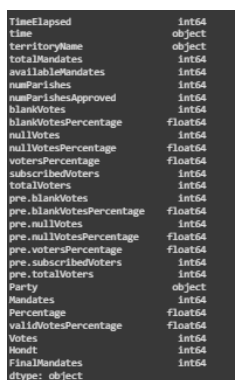


problem u manje kvalitetnim datasetima. Zadovoljavajući broj jedinstvenih vrijednosti omogućuje nam da uspješno razlikujemo, klasificiramo i grupiramo entitete po relevantnim kriterijima te izgradnju smislenih hijerarhija i veza unutar dimenzijskog modela.

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

Zatim je bilo potrebno stvoriti smisleni relacijski model za prikupljene podatke. Relacijski model podataka predstavlja način strukturiranja informacija pomoću međusobno povezanih tablica koje odražavaju stvarne entitete i odnose među njima. U kontekstu analize izbornih rezultata, to označava modeliranje tablica za stranke, teritorije, rezultate po biračkom mjestu, vremenske oznake i slično. Poanta relacijskog modela je omogućiti organiziranu, normaliziranu i nedvosmislenu pohranu podataka tako da se smanji redundancija i poveća dosljednost. Ključne prednosti relacijskog modela uključuju jasno definirane odnose putem primarnih i stranih ključeva, fleksibilnost u pisanju SQL upita, integritet podataka te mogućnost jednostavne nadogradnje sustava bez narušavanja postojećih struktura. Također, relacijski modeli su široko prihvaćeni

u poslovnom okruženju te podržani od strane gotovo svih standardnih baza podataka, što ih čini vrlo praktičnim za implementaciju.

S druge strane, relacijski model nije savršen. U analitičkim i BI sustavima koji se bave velikim količinama podataka i zahtijevaju brze agregacije, relacijski model može biti sporiji u usporedbi s denormaliziranim strukturama poput dimenzijskih modela. Također, složenost modela može otežati razumijevanje krajnjim korisnicima koji nisu tehnički potkovani i na barataju znanjem o ključevima, relacijama, kardinalnostima i sl. Unatoč tim izazovima, izgradnja dobrog relacijskog modela ostaje temelj za kvalitetnu obradu i pripremu podataka za napredne analitičke i poslovne potrebe.

Izrađena su 2 dijagrama: ER i EER. ER model koristi osnovne elemente poput entiteta, atributa i veza za opis strukture podataka, dok EER model dodaje dodatne mogućnosti poput nasljeđivanja između entiteta kako bi se preciznije prikazale složenije situacije iz stvarnog svijeta. Također, ER dijagram izrađen je "ručno" koristeći Lucidchart dok je EER dijagram automatski generiran obrnutim inženjerstvom u MySQL workbenchu.

Kao alat za izradu relacijskog (i kasnije dimenzijskog) modela izabran je MySQL. No, prije samog rada s bazom, bilo je potrebno konceptualno izmodelirati entitete iz izbornih podataka, pa je stoga najprije izrađen ER dijagram.

Za potrebe obrade i analize izbornih podataka, kreiran je sveobuhvatan relacijski model u programskom jeziku Python koristeći SQLAlchemy – objektno-relacijski mapper (ORM) koji omogućava definiranje strukture baze podataka koristeći klase i objekte. Kroz ovaj model uspostavljene su četiri glavne tablice koje reprezentiraju ključne entitete iz domene parlamentarnih izbora u Portugalu: države ("country"), izbori ("election"), političke stranke ("party") i izborni rezultati ("result"), uz dodatnu pomoćnu tablicu za povijesne podatke o izborima ("election\_history").

Prvo je definirana tablica `**Country**`, koja predstavlja teritorijalnu jedinicu – u ovom slučaju izborne jedinice koje su u modelu interpretirane kao države (ili "districts"). Svaka država ima svoj jedinstveni naziv i ID koji se koristi kao primarni ključ te se na njega referenciraju izbori koji se u toj državi održavaju.

Nakon toga, definirana je tablica `**Election**`, koja sadrži sve relevantne informacije o pojedinim izborima, uključujući godinu održavanja, broj dostupnih i ukupnih mandata, broj biračkih mjesta (parishes), broj odobrenih biračkih mjesta, broj i postotak praznih i nevažećih listića, postotak izlaznosti birača, broj upisanih birača te ukupan broj glasača. Tablica je povezana s tablicom "country" pomoću stranog ključa kako bi se znalo kojoj državi pojedini izbori pripadaju.

Tablica `**Party**` predstavlja političke stranke koje su sudjelovale na izborima. Svaka stranka ima jedinstveni naziv i identifikator.

Tablica `**Result**` povezuje izbore i političke stranke, te bilježi broj mandata koje je pojedina stranka osvojila, postotak glasova, postotak važećih glasova, ukupan broj osvojenih glasova i konačan broj osvojenih mandata. Time se omogućuje detaljna analiza rezultata izbora po strankama i povezivanje s konkretnim izbornim ciklusima i teritorijima.

Pomoćna tablica `**ElectionHistory**` dodaje dodatne informacije koje se odnose na povijesne podatke pojedinih izbora, poput broja praznih i nevažećih listića prije izbora,

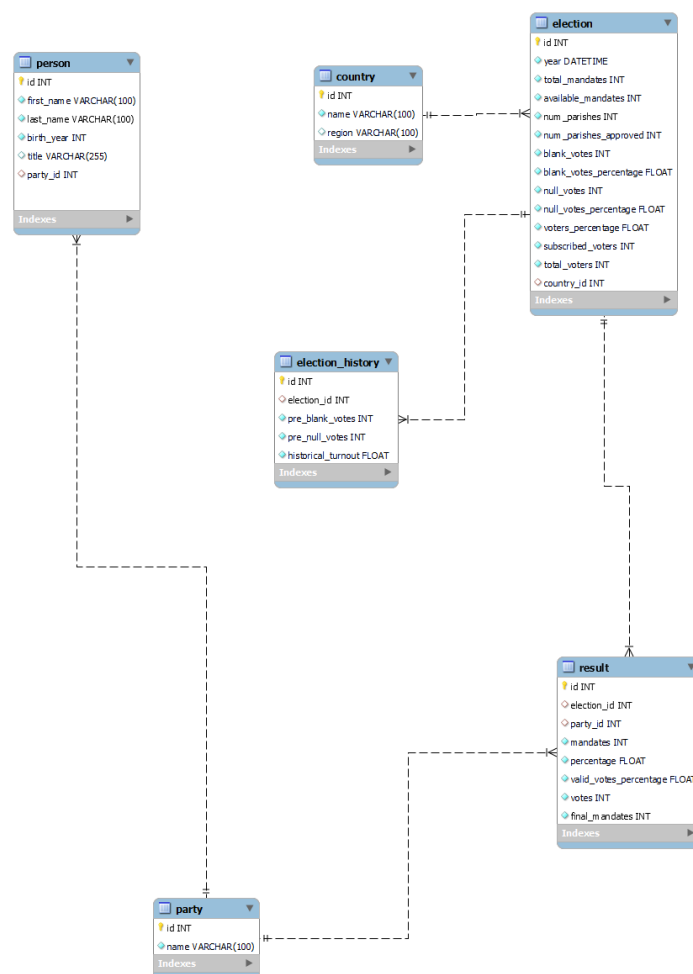
te povijesni postotak izlaznosti. Ova tablica je također povezana s tablicom "election" putem stranog ključa.

Podaci su u sustav učitani iz CSV datoteke koja sadrži prethodno obrađene izborne podatke. Svaki entitet u modelu najprije je inicijaliziran (npr. sve države i stranke su inicijalno stvorene i spremljene u bazu), a zatim su redovi CSV-a iterirani kako bi se za svaki unos stvorili novi izbori, rezultati i povijesni zapisi, pri čemu su uspostavljene sve potrebne veze između entiteta.

Ovakav relacijski model omogućuje:

preciznu i konzistentnu pohranu podataka – svaka informacija se nalazi na samo jednom mjestu (npr. ime stranke), normalizaciju podataka – smanjuje se redundancija, jednostavno postavljanje upita – zahvaljujući primarnim i stranim ključevima, mogućnost proširenja – ako bi se u budućnosti htjeli dodati novi entiteti poput kandidata, regija unutar država ili koalicija, model se može lako proširiti bez potrebe za rekonstrukcijom cijele baze, analitičku dubinu – moguće je pisati složene SQL upite za analizu povijesnih trendova, ponašanja birača ili uspjeha stranaka po teritorijima i godinama.

Ovakva struktura predstavlja dobar temelj za naprednu analitiku, vizualizaciju izbornih podataka, ali i buduću izgradnju aplikacije za prikaz rezultata ili predviđanje izbornih ishoda temeljem povijesnih uzoraka. Na ovaj način postignuta je maksimalna iskoristivost dostupnih podataka kroz sustavno i logično modeliranje relacija u domeni političkih izbora. Također, iako je ovaj relacijski model već sam po sebi poprilično zadovoljavajući, u idućoj iteraciji projekta (pretvorba u dimenzijski model) dodani su još neki atributi koji će predstavljati nove dimenzije za promatranje podataka



Slika 5: EER dijagram relacijskog modela

## Analiza EER Dijagrama Parlamentarnih Izborâ

U nastavku se nalazi detaljna analiza entitetsko-relacijskog (EER) modela koji prikazuje strukturu baze podataka namijenjene vođenju evidencije parlamentarnih izbora. Model uključuje entitete poput izbora, rezultata, političkih stranaka i osoba, te opisuje njihove međusobne veze i kardinalnosti.

### Entiteti i njihova svojstva

1. **person** Entitet **person** predstavlja osobu povezanu s nekom političkom strankom. Svojstva entiteta su:

- **id** – jedinstveni identifikator osobe (INT)
- **first\_name** – ime (VARCHAR)
- **last\_name** – prezime (VARCHAR)
- **birth\_year** – godina rođenja (INT)
- **title** – titula ili funkcija osobe (VARCHAR)

- `party_id` – strani ključ prema entitetu `party`
2. `party` Entitet `party` predstavlja političku stranku. Svojstva:
- `id` – jedinstveni identifikator stranke (INT)
  - `name` – naziv stranke (VARCHAR)
3. `country` Entitet `country` predstavlja izbornu jedinicu. Svojstva:
- `id` – identifikator izborne jedinice (INT)
  - `name` – ime izborne jedinice (VARCHAR)
  - `region` – regija kojoj jedinica pripada (VARCHAR)
4. `election` Entitet `election` sadrži podatke o određenom izbornom ciklusu. Svojstva uključuju:
- `id`, `year`, `total_mandates`, `available_mandates`, `num_parishes`, `num_parishes_approved`
  - `blank_votes`, `null_votes`, `blank_votes_percentage`, `null_votes_percentage`
  - `voters_percentage`, `subscribed_voters`, `total_voters`
  - `country_id` – strani ključ prema entitetu `country`
5. `result` Entitet `result` prikazuje rezultate izbora za pojedine stranke:
- `id`, `election_id`, `party_id` – strani ključevi
  - `mandates`, `final_mandates`, `percentage`, `valid_votes_percentage`, `votes`
6. `election_history` Entitet `election_history` sadrži povijesne informacije povezane s određenim izborima:
- `id`, `election_id` – strani ključ
  - `pre_blank_votes`, `pre_null_votes`, `historical_turnout`

### Veze i kardinalnosti

- **person – party:** Veza je **N:1**, jer više osoba može pripadati jednoj stranci.
- **election – country:** Veza je **N:1**, jer više izbora se može održati u jednoj izbornoj jedinici.
- **election\_history – election:** **1:1** veza, jer je povijest direktno vezana uz jedan izbor.
- **result – election:** **N:1**, jer se svaki rezultat odnosi na jedan izbor.
- **result – party:** **N:1**, jer se više rezultata može odnositi na istu stranku (kroz vrijeme).

EER model omogućava detaljno praćenje izbora, njihovih rezultata, pripadnosti osoba političkim strankama, te povijesnih podataka. Jasna hijerarhija veza i kardinalnosti omogućuje izvođenje složenih upita i potpunu analizu izbornog procesa. U daljnjem razvoju dimenzijskog modela neki su entiteti modificirani te su im dodani stupci koji omogućavaju razvoj hijerarhija i razvijena je logika za punjenje tih stupaca.

## 7. Dimenzijski model podataka

Dimenzijski model predstavlja jednu od temeljnih metoda modeliranja podataka u skladištima podataka, s glavnim ciljem optimizacije upita i omogućavanja brze i fleksibilne analize. Za razliku od relacijskog modela koji je usmjeren na normizaciju i konzistenciju podataka, dimenzijski model je denormaliziran kako bi omogućio lakše čitanje velikih količina podataka. Osnovni elementi dimenzijskog modela su tablice činjenica (*fact tables*) i tablice dimenzija (*dimension tables*).

Tablica činjenica sadrži mjerne podatke ili kvantitativne metrike koje korisnici žele analizirati. U kontekstu ovog modela, glavna tablica činjenica je `fact_election_result`, koja sadrži metrike kao što su broj mandata, broj glasova, postotak glasova, broj nevažećih i praznih listića, ukupno birača, izlaznost i druge agregatne vrijednosti povezane s izbornim rezultatima. Ključ svakog zapisa u tablici činjenica definiran je kombinacijom stranih ključeva koji vode prema odgovarajućim dimenzijama: izbori, stranke, vrijeme i povijest izbora.

Tablice dimenzija sadrže opisne podatke koji se koriste za filtriranje, grupiranje i označavanje činjenica. U promatranom modelu možemo uočiti više dimenzijskih tablica:

- `dim_election`: opisuje pojedine izbore, uključujući njihov identifikator te poveznicu na datum (`date_tk`) i zemlju (`country_id`).
- `dim_date`: omogućuje vremensku analizu podataka, uključujući detalje poput dana, mjeseca, godine i dana u tjednu.
- `dim_country`: predstavlja političku geografiju, uključujući regije i nazive zemalja (odnosno izbornih jedinica).
- `dim_party`: opisuje političke stranke koje sudjeluju u izborima, uključujući njihov naziv i političku orijentaciju.
- `dim_person`: dodatna dimenzija koja povezuje osobu sa strankom i omogućuje analiziranje političara, uključujući ime, prezime, godinu rođenja i titulu.
- `dim_election_history`: omogućuje povijesni kontekst pojedinih izbora, npr. koliko je bilo praznih i nevažećih listića u prethodnim verzijama, datume početka i završetka povijesnih razdoblja, kao i izlaznost.

Važan koncept u ovakvim modelima su sporo mijenjajuće dimenzije (*slowly changing dimensions*, SCD). Ove dimenzije se mijenjaju rijetko, ali je važno sačuvati staru vrijednost za potrebe analize povijesnih podataka. U ovom modelu `dim_election_history` predstavlja sporo mijenjajuću dimenziju, jer se rezultati i statistike prethodnih verzija

izbora zadržavaju u vremenskim intervalima `date_from` i `date_to`. Time se omogućuje točna vremenska analiza promjena u povijesti izbora.

Dodatno, `dim_person` također može predstavljati sporo mijenjajuću dimenziju ako se, primjerice, političar prebaci iz jedne stranke u drugu – za takav slučaj moguće je pratiti njegovu pripadnost kroz različite vremenske točke zadržavajući povijest.

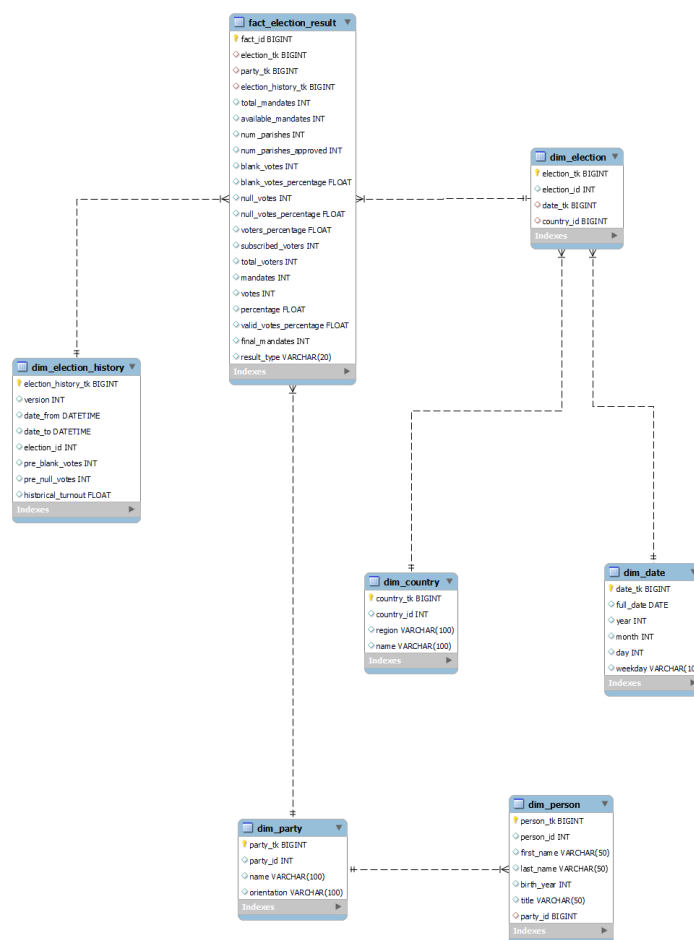
Također, treba napomenuti kako su neki od podataka izgenerirani te se nisu nalazili u originalnom datasetu. To su redom: degenerirana dimenzija `result_type` koja će biti detaljnije objašnjena u idućem paragrafu, `version`, `date_from`, `date_to` koji služe za praćenje promjena u glasovima na izborima kroz vrijeme (sa svakim novim dodavanjem obrađenih glasova, kreira se nova verzija `election_historyja`), `region` unutar `dim_country` koji služi za stvaranje dodatne hijerarhije unutar dimenzije i predstavlja geografsku regiju unutar Portugala. `Year`, `month`, `day` i `weekday` unutar `dim_date` koji omogućuju detaljniju vremensku analizu podataka te su kreirani ekstrakcijom dijelova `timestampa` iz dataseta. `Orientation` unutar `dim_party` koji predstavlja političku orijentaciju pojedine stranke (lijevo, desno ili centar) te `title` unutar dimenzije `person` koja označava titulu kandidata na izborima (doktor, magistar i sl.)

U modelu je prisutna i degradirana dimenzija (*degenerate dimension*), koja se odnosi na atribut `result_type` unutar tablice činjenica. Iako se radi o opisnom atributu (npr. *konačni*, *privremeni*), za njega nije definirana posebna tablica dimenzija. Umjesto toga, `result_type` je direktno pohranjen u tablici činjenica. Ovakav pristup je čest kada atribut nema dodatne opisne podatke i koristi se isključivo za filtriranje ili identifikaciju rezultata.

Što se tiče strukture modela, ovdje se ne koristi čista zvjezdasta shema (*star schema*), već tzv. snježna pahulja (*snowflake schema*). U snježnoj shemi dimenzije su dodatno normirane, čime se smanjuje redundantnost podataka, ali i povećava broj pridruživanja (*joinova*) prilikom izvršavanja upita. To se jasno vidi u relaciji između `dim_election` i `dim_country` te `dim_election` i `dim_date`, gdje su dimenzije izvedene putem dodatnih vanjskih ključeva, a ne direktno unutar tablice činjenica. Time je postignuta veća fleksibilnost i konzistencija, ali na račun nešto slabijih performansi u analitičkim sustavima.

Zaključno, ovaj model predstavlja dobro strukturiran dimenzijski model temeljen na "pahuljici", koji omogućuje složenu analizu izbornih rezultata kroz različite dimenzije, uključujući vrijeme, lokaciju, političke stranke, kandidate i povijesni kontekst. Korištenje sporo mijenjajućih dimenzija i višerazinskih veza omogućuje točnu i povijesno konzistentnu analitiku u političkom okruženju, dok upotreba degradirane dimenzije `result_type` dodatno pojednostavljuje klasifikaciju rezultata bez potrebe za dodatnim relacijama.





Slika 6: Snowflake shema dimezijskog modela

## 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

Za potrebe vizualizacije podataka iz dimenzijskog modela te omogućavanja korisnicima intuitivne i interaktivne analize izbornih rezultata, izrađen je prilagođeni web-based

dashboard. Umjesto korištenja gotovih alata kao što su Tableau ili Power BI, koji su često ograničeni u fleksibilnosti i prilagodbi specifičnim potrebama domene, odlučili smo se za izradu vlastitog rješenja. Bilo bi šteta osloniti se na generičke alate kad je moguće dizajnirati vizualizacijsko sučelje koje precizno odgovara strukturi i logici našeg dimenzijskog modela.

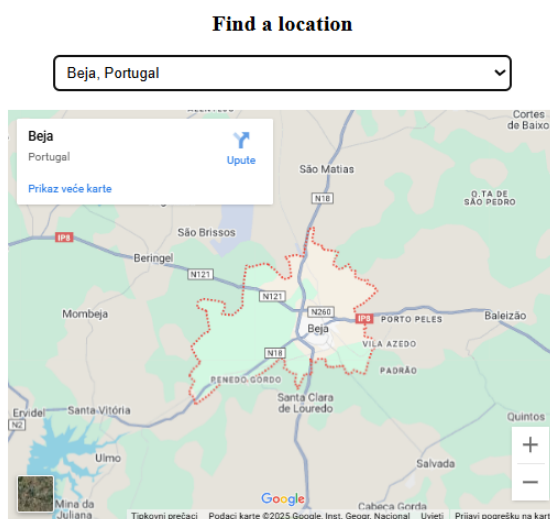
Za implementaciju su korištene sljedeće tehnologije:

- **Vue.js** – kao frontend JavaScript framework za izgradnju responzivnog i interaktivnog korisničkog sučelja.
- **Matplotlib** – za generiranje statičkih i dinamičkih grafova na strani servera, posebno prikladno za analize koje zahtijevaju preciznu kontrolu nad prikazom.
- **Metabase** – za jednostavno i brzo generiranje uvida i osnovnih vizualizacija direktno nad bazom, korišten pretežito tijekom razvoja i verifikacije podataka.

Ovaj pristup omogućio je kreiranje fleksibilnog dashboarda koji se može lako prilagoditi promjenama u modelu podataka, nadograditi dodatnim funkcionalnostima te povezati s analitičkim funkcijama specifičnim za političku domenu i praćenje izbornih trendova.

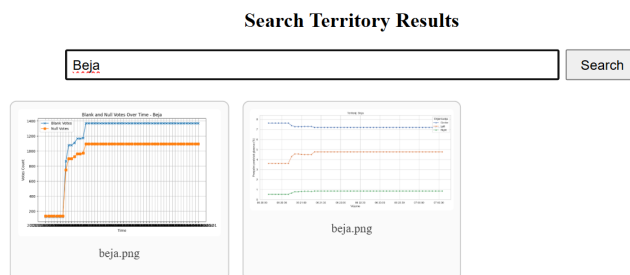
U nastavku će najprije biti prikazane funkcionalnosti dashboarda, a potom svi grafikoni i njihova interpretacija u kontekstu izbora.

- Implementiran je dinamički pronalazak pojedinih izbornih jedinica(`dim_country`) na interaktivnoj karti Portugala.



Slika 7: Interaktivna karta Portugala

- Omogućeno je pretraživanje rezultata po teritoriju, specifično prevladavajućih političkih orijentacija te pregled broja praznih i nevažećih glasova. Ta funkcionalnost odgovara OLAP Slice operaciji jer omogućava filtriranje po 1 dimenziji (teritorij)



Slika 8: Rezultati po teritoriju

- Moguće je pregledati sve kandidate, stranke kojima pripadaju i njihove autentične grbove te ih filtrirati po određenim kriterijima poput pripadnosti stranci ili godini rođenja. Ta funkcionalnost odgovara OLAP Dice operaciji jer omogućava filtriranje po više dimenzija istovremeno.

**Politicians and Their Parties**

Party:






















Birth Year:

Condition:

All

Enter year

Exact

 <b>Dr. João Silva</b> Born: 1985 Party: PSD	 <b>Eng. Carlos Pereira</b> Born: 1970 Party: PPD-PSD	 <b>Rafael Costa</b> Born: 1982 Party: B.E.
 <b>Prof. Miguel Fernandes</b> Born: 1978 Party: CDS-PP	 <b>Dra. Ana Santos</b> Born: 1975 Party: PAN	 <b>Bruno Almeida</b> Born: 1989 Party: PCTP/MRPP
 <b>Ricardo Neves</b> Born: 1967 Party: A	 <b>Eng. Fernanda Martins</b> Born: 1973 Party: L	 <b>Gustavo Lima</b> Born: 1988 Party: JFP
 <b>Dra. Patrícia Ramos</b> Born: 1990 Party: PDR	 <b>Tiago Moreira</b> Born: 1969 Party: PNR	 <b>Vasco Mendes</b> Born: 1972 Party: TUR
 <b>Dr. Diogo Faria</b> Born: 1960 Party: PPM	 <b>Helena Cruz</b> Born: 1985 Party: MPT	 <b>Prof. Sofia Teixeira</b> Born: 1978 Party: MAS
 <b>Eng. André Rodrigues</b> Born: 1955 Party: PCP-PEV	 <b>Catarina Barros</b> Born: 1992 Party: R.L.R.	 <b>José Nogueira</b> Born: 1983 Party: CH
 <b>Dra. Mariana Henriques</b> Born: 1976 Party: Bera	 <b>Filipe Coelho</b> Born: 1991 Party: NDG	 <b>Dr. Eduardo Gomes</b> Born: 1983 Party: FDP

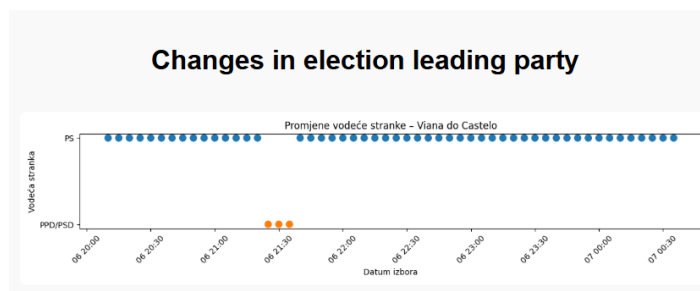
Slika 9: Pregled stranaka

- Zatim, kako je u skladištima podataka važno pratiti promjene kroz različite verzije podataka, sustav omogućuje pregled povijesnih rezultata izbora za svaki teritorij, pri čemu teritorij odgovara jednoj izbornoj jedinici. Iako podaci u odabranom datasetu ne prate promjene iz godine u godinu već samo kroz 1 izbore, omogućena je usporedba rezultata kroz različite izborne cikluse, što je dovoljno za analizu promjena u izbornim rezultatima tijekom vremena.

Za svaki teritorij moguće je vidjeti kako se mijenjao broj glasova po strankama kroz izbore, kako se ti brojevi izražavaju u postocima, koja je stranka bila vodeća u kojem ciklusu, te kako je izgledala konačna raspodjela mandata unutar teritorija. Ovakva analiza omogućuje korisnicima uvid u trendove i pomake u biračkom tijelu na razini izbornih jedinica, što može biti korisno za stranke, analitičare i istraživače.

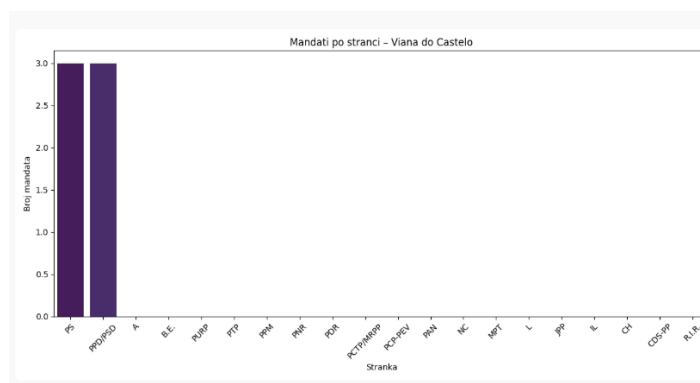
Konkretno, ti su podaci obrađeni na 4 načina: promjenama u vodećoj poziciji kroz vrijeme, broju mandata po stranci, postotku glasova kroz vrijeme i broju

glasova kroz vrijeme. Ti su podaci dostupni za svaku jedinicu, a primjera radi, odabran je teritorij Viana do Castelo



Slika 10: Pregled vodećih stranaka

Na grafikonu je vidljivo kako se mijenjala vodeća stranka kroz vrijeme te se može iščitati kako je gotovo cijelo vrijeme, osim kratkog perioda, u vodstvu bila stranka PS. To odgovara OLAP drill down operaciji jer se ne promatraju samo generalni podaci za izbornu jedinicu nego se točno promatraju za svaku stranku.



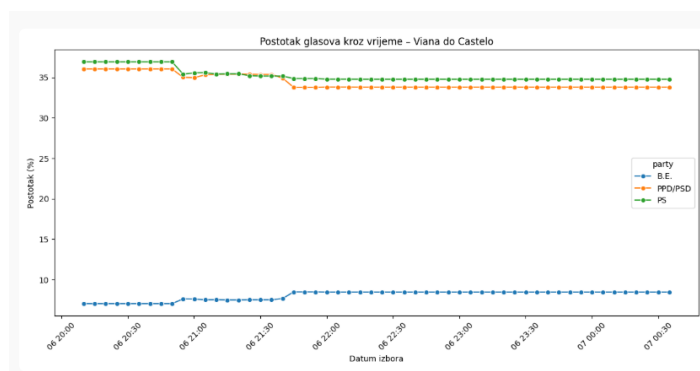
Slika 11: Broj mandata po stranci

Stupčasti grafikon prikazuje finalnu distribuciju mandata po strankama na odabranom teritoriju. Vidljivo je kako je na samom kraju došlo do ravnomjerne podjele od po 3 mandata za stranku PS i 3 za PPD/PSD. Također, treba napomenuti da broj mandata koje pojedini teritorij dodjeljuje nije fiksna, već ovisi o broju stanovnika i njihovom udjelu u ukupnom stanovništvu — što utječe na proporcionalnu dodjelu mandata.

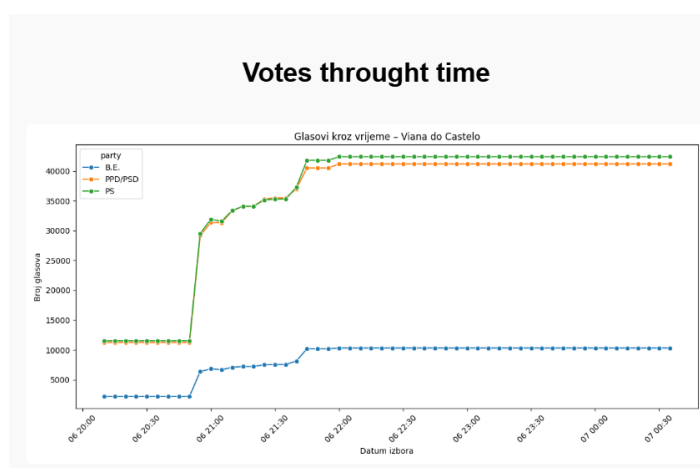
Ova analiza ilustrira primjenu dviju OLAP operacija: slice i roll-up. Prvo, korisnik odabire jedan specifičan teritorij — čime se primjenjuje slice operacija, odnosno filtriranje podataka prema jednoj vrijednosti dimenzije "teritorij". Nakon toga, koristi se roll-up operacija kako bi se mandatni rezultati agregirali po strankama, čime se prelazi s detaljnijih podataka (npr. pojedinačnih glasova ili izbornih jedinica) na višu razinu apstrakcije — ukupan broj osvojenih mandata po političkoj opciji.

Kombinacija ovih operacija omogućuje korisniku da istovremeno fokusira analizu

na specifičan teritorijalni kontekst, dok istovremeno dobiva sažetu, agregiranu sliku političke raspodjele snaga u tom području.



Slika 12: Promjena postotka glasova kroz vrijeme



Slika 13: Promjena broja glasova kroz vrijeme

Nadalje, iduća dva grafikona u kontekstu OLAP operacija predstavljaju analitički vrlo sličan pogled na podatke, budući da oba prikazuju promjenu kvantitativne vrijednosti kroz vrijeme za određenu stranku — jedan kroz apsolutni broj glasova, a drugi kroz postotak osvojenih glasova.

Ova analiza podrazumijeva primjenu kombinacije slice i drill-down operacija. Slice se koristi za filtriranje podataka prema određenoj stranci — fokusira se samo na jednu političku opciju iz cjelokupnog skupa podataka. Nakon toga, pomoću drill-down operacije, analitički pogled ide dublje u vremensku dimenziju, omogućujući korisniku da promatra kako se ta vrijednost mijenjala kroz pojedine izborne cikluse.

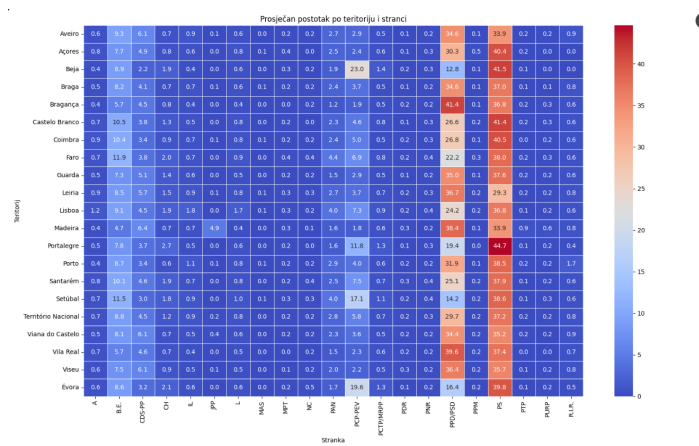
Uspoređivanjem oba grafikona paralelno, može se jasno uočiti način na koji se apsolutna brojka glasova preslikava u postotni udio, što dodatno doprinosi razumijevanju relativne snage stranke u kontekstu ukupnog broja birača. Time se ne samo promatra trend podrške određenoj stranci, već se omogućuje i dublja interpretacija rezultata, jer se broj glasova stavlja u odnos prema ukupnom

biračkom tijelu — što može otkriti promjene koje na prvi pogled nisu očite samo iz apsolutnih vrijednosti.

Sljedeća primijenjena OLAP operacija jest pivot. Iako u ovom kontekstu nije naročito korisna u usporedbi s prikazima temeljenima na operacijama poput slice, dice ili drill-down, ipak može pružiti određeni uvid u podatke. Pivot, za razliku od prethodno navedenih operacija koje otkrivaju nove obrasce ili omogućuju filtraciju i agregaciju, prvenstveno služi za reorganizaciju prikaza podataka - tj. za zamjenu redaka i stupaca radi preglednosti ili naglašavanja određenih odnosa među dimenzijama.

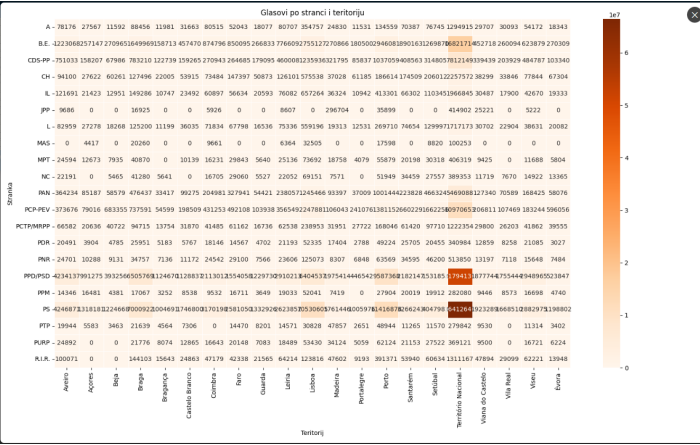
U konkretnom slučaju, pivot operacija omogućuje da se, primjerice, dimenzija "stranka" prikaže kao stupci umjesto redaka, dok se izborne jedinice prikazu po redovima, ili obrnuto. Takva promjena može pomoći u vizualnom uspoređivanju više entiteta istovremeno, ali sama po sebi ne dodaje novu informaciju, već samo prezentira postojeće podatke iz druge perspektive. Upravo zato, iako tehnički spada u skup osnovnih OLAP operacija, pivot u ovom slučaju ne doprinosi interpretaciji podataka u istoj mjeri kao operacije koje filtriraju, agregiraju ili istražuju vremenske ili prostorne obrasce.

Svejedno, njezina vrijednost leži u mogućnosti da korisnicima s različitim preferencijama u načinu čitanja tabličnih podataka ponudi alternativni prikaz koji im može olakšati preglednost i usporedbu — ali ne i otkriti nešto što već nije bilo prisutno.



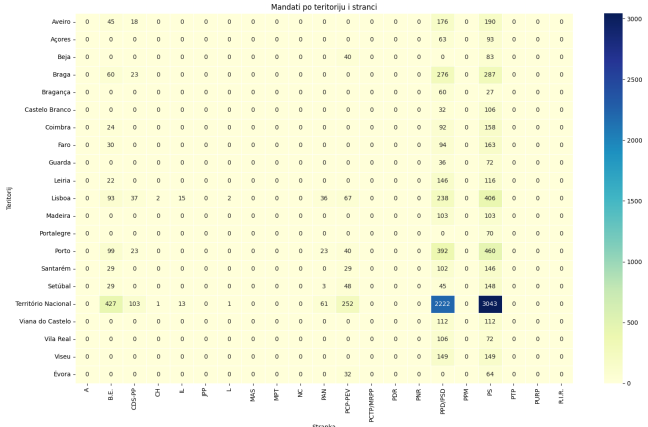
Slika 14: Prosječan postotak po teritoriju i stranci

Prvi pivot grafikon omogućuje uvid u prosječan postotak koji je svaka stranka osvojila u svakom teritoriju kad se u obzir uzme cijeli vremenski period izbora. Prema boji je vidljivo kako su pojedine stranke "dominantne" te osvajaju značajno više mandata kroz sve teritorije u odnosu na konkurentske strane. (konkretno, stranke PS i PPD/PSD)



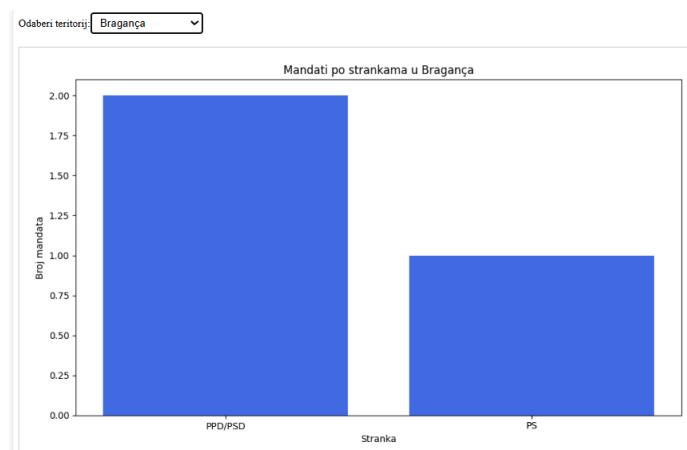
Slika 15: Broj glasova po stranci i teritoriju

Drugi pivot grafikon omogućuje prikazuje broj glasova svake stranke na svakom teritoriju što daje uvid u neke ključne detalje. Konkretno, omogućuje da se uvidi koja je stranka dominantna na kojem pojedinom teritoriju te omogućuje da se uvidi koji su teritoriji ključni u samim izborima (zbog svoje veličine i broja glasova/mandata koje oni daju).



Slika 16: Broj mandata po stranci i teritoriju

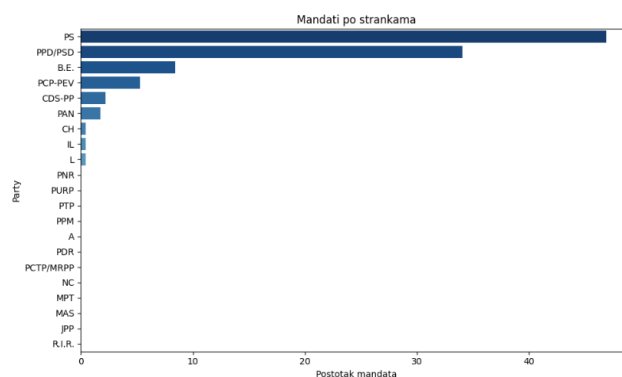
Za treći pivot koji prikazuje broj mandata po stranci i teritoriju vrijedi isto što i za drugi koji prikazuje broj glasova, no posebno su zanimljivi kada se promatraju u kombinaciji jer tada omogućuju uvid u čisto matematičko preslikavanje glasova u mandate (prije primjene D'Hondtove metode koja često nije "poštena" jer favorizira velike stranke, dok malim "lokalnim" strankama onemogućava značajnije političke uspjehe. Jer čak i ako osvoje dovoljan broj glasova, to često nije dovoljno za dobivanje mandata zato što D'Hondtova metoda ne vrednuje sve glasove u svim teritorijima jednako).



Slika 17: Slice konačnih rezultata po teritoriju

Zatim, omogućen je prikaz konačnih rezultata po teritoriju što je klasičan primjer slice operacije zato što ona uzima finalne rezultate i "filtrira" ih za 1 teritorij. Na sličan se način može prikazati dice kada bi se promatralo rezultate samo za određene stranke samo na određenim teritorijima. A ukoliko bi se uzelo ukupne rezultate svih stranaka i promatralo ih se (agregiralo) na razini cijele države, bio bi to roll-up.

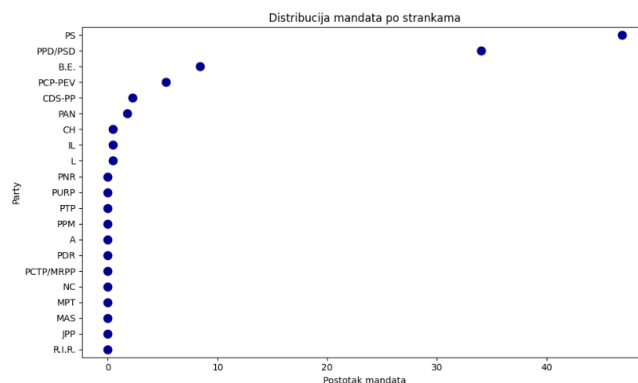
Naravno, omogućen je i prikaz konačnih rezultata na razini cijele države, jer filtriranje i OLAP operacije ne bi imale smisla kada ne bi bilo moguće vidjeti "sveukupne" podatke. Prikaz postotka ukupnih osvojenih mandata prikazan je na 3 načina te omogućava promatranje istih podataka iz 3 perspektive.



Slika 18: Stupčasti grafikon postotka mandata po strankama

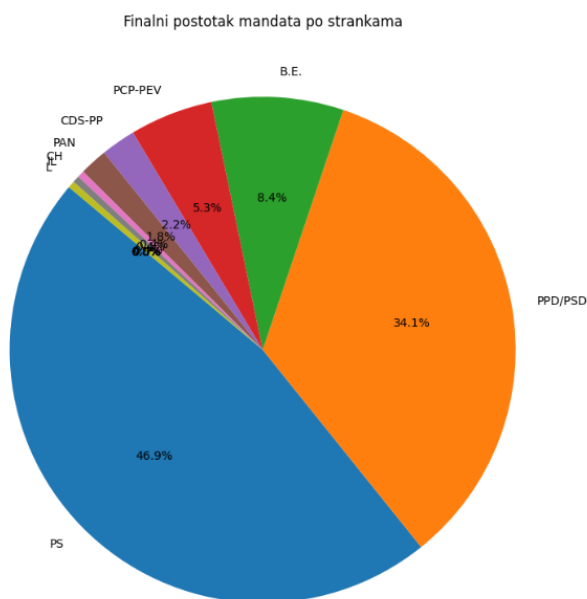
Prvi grafikon omogućuje uvid u postotak osvojenih mandata po strankama. Dodatno, daje i uvid u to koje su sve stranke uopće ušle u parlament. Taj bi se grafikon mogao zarotirati i na način da se stranke nalaze na x-osi no, ponovno, ta operacija ne bi dala neke značajne nove uvide u same rezultate.





Slika 19: Distribucija mandata po strankama u postotcima

Drugi grafikon omogućuje uvid u distribuciju mandata po strankama što je jasan indikator toga da nisu sve stranke jednako "popularne" te da PS i PPD/PSD osvajaju znatno više mandata od konkurencije.



Slika 20: Pie chart konačnih rezultata

Pie chart konačnih rezultata daje točan i direktan uvid u odnose među rezultatima stranaka i pobjednike izbora. Vidljivo je kako je pobjednička stranka osvojila više nego sve ostale stranke zajedno (s izuzetkom drugoplasirane stranke).

OVO JE NOTE KOJI ĆU OBRISAT; OSTALO MI JE JOŠ 5 BOTUNI ZA STISNAT I PRIKAZAT GRAFIKONE I OPISAT; ONDA JE VISUAL PART GOTOV