

PARLAMENTARNI IZBORI U PORTUGALU 2019.

Fakultet Informatike u Puli

Kolegij: Skladišta i rudarenje podataka

Autor: Antonio Labinjan

Mentori: izv.prof.dr.sc. Goran Oreški & Marijela
Miličević, mag. educ. inf.

- ▶ Politički izbori generiraju velike količine kompleksnih podataka
- ▶ Analiza tih podataka je ključna za razumijevanje ponašanja birača, uočavanje trendova i donošenje odluka
- ▶ CILJ: izgradnja sustava za analizu izbornih rezultata

UVOD

- ▶ Spomenuti sustav sastoji se od:
 - Dimenzijskog modela izbornih podataka
 - Skladišta podataka temeljnog na tom modelu
 - Interaktivnog dashboarda za analizu rezultata
 - NAGLASAK: fleksibilnost, dubinska analiza i vizualno istraživanje podataka

SUSTAV

- ▶ Biti će prikazan cijeli proces izrade skladišta i interaktivnog dashboarda kroz nekoliko koraka: od pronalaska samog dataseta, preko kreiranja relacijskog modela, kreiranja dimenzijskog modela, punjenja skladišta kroz ETL proces do vizualizacije podataka i kreiranja finalnog proizvoda

PROCES

- ▶ Ključni pojmovi koje treba spomenuti prije izlaganja daljnjih rezultata su svakako:
 - 1) POSLOVNA INTELIGENCIJA (BUSINESS INTELLIGENCE, BI)
 - 2) SKLADIŠTA PODATAKA (DATA WAREHOUSES, DW)

KLJUČNI POJMOVI

- ▶ Skup tehnologija, procesa i metoda za prikupljanje i analizu poslovnih podataka
- ▶ Cilj: donošenje informiranih poslovnih odluka
- ▶ Pretvara sirove podatke u korisne informacije
- ▶ Pomaže menadžmentu u prepoznavanju obrazaca, trendova i prilika
- ▶ Uključuje izvještavanje, analizu, prediktivne modele, vizualizaciju i integraciju u stvarnom vremenu
- ▶ Ključan alat za strateško planiranje i operativno upravljanje u suvremenom poslovanju

BUSINESS INTELLIGENCE (BI)

- ▶ Centralizirani repozitorij strukturiranih podataka za analitičku obradu
- ▶ Namijenjeno podršci poslovnoj inteligenciji i donošenju odluka
- ▶ Optimizirano za čitanje velikih količina povijesnih podataka, agregaciju i kompleksne upite

DATA WAREHOUSES

- ▶ Podaci se unose putem ETL procesa (Extract, Transform, Load)
- ▶ Osigurava čišćenje, usklađivanje i konsolidaciju podataka iz različitih izvora
- ▶ Arhitektura uključuje:
 - ▶ sloj izvora podataka
 - ▶ ETL sloj
 - ▶ centralno skladište
 - ▶ sloj prezentacije
- ▶ Često koristi dimenzijski model (fact i dimenzijske tablice)

DATA WAREHOUSES (2)

- ▶ Skladišta podataka omogućuju:
 - Prepoznavanje trendova
 - Praćenje povijesnih podataka
 - Ujednačenu uporbu više izvora podataka
 - Polustrukturiranje podatke
 - Izradu distribuiranih sustava

KLJUČNE ZNAČAJKE

- ▶ Pronalazak i analiza dataseta
- ▶ Izrada relacijskog modela podataka
- ▶ Izrada dimenzijskog modela podataka
- ▶ ETL proces koristeći Apache Spark

- ▶ Bonus (nije nužno dio same izrade skladišta): Kreiranje dashboarda za vizualizaciju podataka

KORACI IZRADE SKLADIŠTA

- ▶ PRONALAZAK I ANALIZA DATASETA U SKLADU SA ZADANIM UVJETIMA
- ▶ Veličina? Bar 15 000 x 20
- ▶ Dovoljno različiti podaci?
- ▶ Vremenska dimenzija?
- ▶ Kvantitativni i kvalitativni podaci?
- ▶ Null vrijednosti?

CHECKPOINT #1

- ▶ Koristi se dataset: "Election data 2019."
(<https://archive.ics.uci.edu/dataset/513/> [4])
- ▶ Bavi se analizom parlamentarnih izbora u Portugalu 2019. godine

DATASET

- ▶ Dataset sadrži 21.643 retka i 28 stupaca, čime zadovoljava uvjet minimalne veličine (15.000 redaka, 10 stupaca)
- ▶ Premali dataset bi otežao analizu jer iz njega ne bi bilo moguće uočiti pravilnosti, korelacije i uzorke

(21643, 28)

VELIČINA DATASETA

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	

- ▶ Null vrijednosti negativno utječu na kvalitetu analize, povećavaju kompleksnost obrade podataka i mogu dovesti do pogrešnih zaključaka.
- ▶ Dataset je u potpunosti "čist" - nema niti jednu nedostajuću vrijednost, što omogućuje pouzdaniju i precizniju analizu bez dodatne obrade

NULL VRIJEDNOSTI

- ▶ Analiza jedinstvenih vrijednosti otkriva raznolikost podataka, što je ključno za kvalitetnu i dubinsku analizu.
- ▶ Stupci s mnogo unique vrijednosti omogućuju detaljnu analizu po dimenzijama poput teritorija, stranaka ili vremenskih oznaka.
- ▶ Identifikacija nekonzistentnih kodiranja (npr. "HR" i "Hrvatska") pomaže u osiguravanju točnosti i ujednačenosti podataka.

UNIQUE VRIJEDNOSTI

TimeElapsed	54
time	54
territoryName	21
totalMandates	62
availableMandates	69
numParishes	20
numParishesApproved	219
blankVotes	329
blankVotesPercentage	146
nullVotes	331
nullVotesPercentage	107
votersPercentage	282
subscribedVoters	335
totalVoters	336
pre.blankVotes	323
pre.blankVotesPercentage	130
pre.nullVotes	329
pre.nullVotesPercentage	90
pre.votersPercentage	278
pre.subscribedVoters	331
pre.totalVoters	331
Party	21
Mandates	67
Percentage	1363
validVotesPercentage	1387
Votes	4029
Hondt	41
FinalMandates	17
dtype: int64	

- ▶ **Provjereni su tipovi podataka** kako bi se utvrdila prisutnost **kvantitativnih (numeričkih)** i **kvalitativnih (tekstualnih)** vrijednosti.
- ▶ **Kvantitativni podaci** (npr. broj glasova) omogućuju **statističke analize**, dok se **kvalitativni podaci** (npr. nazivi stranaka, teritorija) koriste za **grupiranje i filtriranje**.
- ▶ **Većina stupaca sadrži numeričke vrijednosti**, uz nekoliko object stupaca koji predstavljaju **stringove ili vremenske oznake**.
- ▶ Ispravna identifikacija tipova podataka je ključna za **točnu analizu, vizualizaciju i izbjegavanje grešaka** pri obradi.

```
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
```

TIPOVI PODATAKA

- ▶ Na samom kraju, bilo je potrebno ispisati koja su imena stupaca i što oni točno predstavljaju
- ▶ Radi preglednosti, taj će dio biti prikazan u odvojenoj sekciji prezentacije

IMENA STUPACA

TimeElapsed – proteklo vrijeme od početka brojanja glasova

Time - točan trenutak kad su podaci zabilježeni (timestamp)

TerritoryName – naziv izborne jedinice (savezna država Portugala)

TotalMandates – ukupan broj mandata koji se dodjeljuju u teritoriju

AvailableMandates – broj još neraspodjeljenih mandata

KLJUČNI ATRIBUTI DATASETA (1/3)

NumParishes – ukupan
broj biračkih mjesta

NumParishesApproved
– broj obrađenih
biračkih mjesta

BlankVotes – broj
praznih (nepopunjenih)
listića

BlankVotesPercentage
– postotak praznih
(nepopunjenih) listića

NullVotes – broj
nevažećih (nepravilno
ispunjenih) listića

NullVotesPercentage –
postotak nevažećih
listića

VotersPercentage –
postotak birača koji su
glasali

KLJUČNI ATRIBUTI DATASETA (2/3)

- ▶ 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 - | | -
- ▶ Pre.nullVotes - broj nevažećih listića u prethodnom izvještaju
- ▶ Pre.nullVotesPercentage - postotak nevažećih listića - | | -

KLJUČNI ATRIBUTI DATASETA (3/4)

- ▶ 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 stranke na koju se podaci odnose

KLJUČNI ATRIBUTI DATASETA (4/4)

- ▶ U dimenzijski model su kasnije dodani novi atributi koji nisu postojali u originalnom CSV-u, kako bi se omogućila dublja analiza rezultata izbora.
- ▶ Nakon potvrde kvalitete i strukture podataka, projekt je nastavljen izradom relacijskog modela.

ZAKLJUČAK

- ▶ Kreiranje relacijskog modela i EER dijagrama

CHECKPOINT #2

- ▶ Relacijski model strukturira podatke u povezane tablice koje odražavaju stvarne entitete i njihove odnose (npr. stranke, teritoriji, rezultati, vremenske oznake).
- ▶ Cilj je organizirana, normalizirana i dosljedna pohrana podataka uz smanjenu redundanciju i jasno definirane primarne i strane ključeve.
- ▶ Prednosti uključuju fleksibilnost SQL upita, integritet podataka i mogućnost nadogradnje bez narušavanja strukture.
- ▶ Nedostatci su sporije agregacije u analitičkim sustavima i veća složenost za korisnike bez tehničkog znanja.
- ▶ Izrađena su dva dijagrama:
 - ▶ – ER dijagram u Lucidchartu (ručno) za osnovno konceptualno modeliranje
 - ▶ – EER dijagram u MySQL Workbenchu (automatski), s naprednim elementima poput nasljeđivanja
- ▶ Kao alat za implementaciju odabran je MySQL, uz prethodnu izradu ER dijagrama kao temelj za daljnji razvoj baze.

RELACIJSKI MODEL

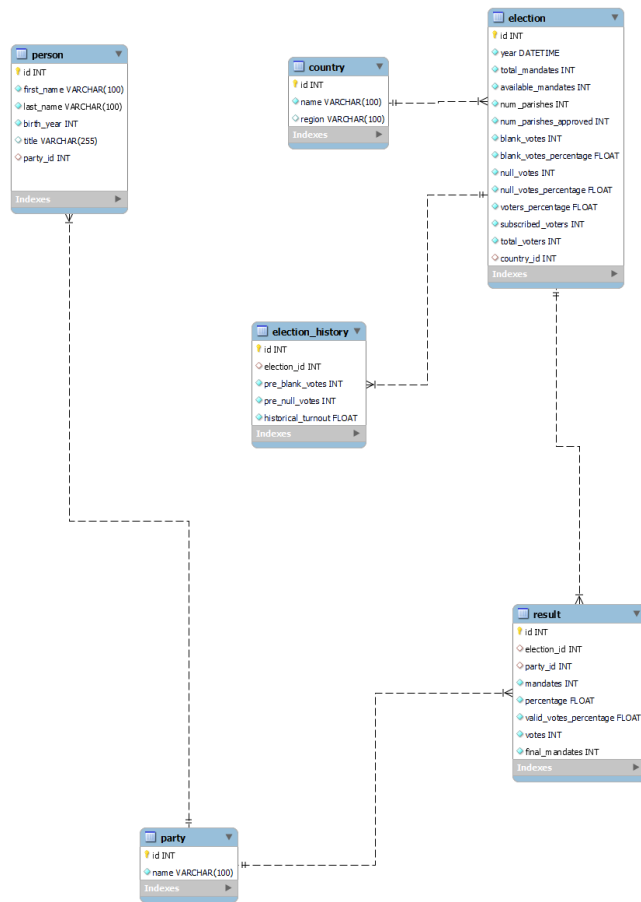
EER dijagram prikazuje relacijski model za praćenje parlamentarnih izbora, s jasno definiranim entitetima i njihovim vezama.

- ▶ Glavni entiteti su:
- ▶ – person (osobe povezane sa strankama)
- ▶ – party (političke stranke)
- ▶ – country (izborne jedinice)
- ▶ – election (pojedini izborni ciklusi)
- ▶ – result (rezultati izbora po strankama)
- ▶ – election_history (povijesni podaci vezani uz izbore)

SVOJSTVA RELACIJSKOG MODELA

- ▶ Svaki entitet ima precizno definirane attribute poput identifikatora, imena, postotaka, broja glasova i stranih ključeva koji osiguravaju povezivost tablica.
- ▶ Veze i kardinalnosti uključuju odnose poput N:1 između `person` i `party`, N:1 između `election` i `country`, te 1:1 između `election_history` i `election`.
- ▶ Model omogućuje detaljno praćenje izbornih rezultata, promjena kroz vrijeme i povezanost kandidata sa strankama i teritorijima.
- ▶ U kasnijim fazama razvoja, neki entiteti su prošireni dodatnim stupcima radi boljeg analitičkog potencijala i pripreme za dimenzijski model.

KARDINALNOSTI



EER DIAGRAM

- ▶ Izrada dimenzijskog modela (Snowflake schema)

CHECKPONT #3

Dimenzijski model je ključan za skladišta podataka i analitičke sustave

- ▶ Cilj: optimizacija upita i omogućavanje brze, fleksibilne analize
- ▶ Za razliku od relacijskog modela, koristi denormalizaciju radi boljih performansi
- ▶ Glavni elementi:
 - ▶ – Fact tablice (sadrže mjerljive podatke)
 - ▶ – Dimension tablice (sadrže kontekst i opisne podatke)
- ▶ Pogodan za rad s velikim količinama podataka i BI alatima

DIMENZIJSKI MODEL PODATAKA

- ▶ Sadrži mjerljive podatke/mjere povezane s izbornim rezultatima
- ▶ Kvantitativne vrijednosti
- ▶ Strani ključevi prema dimenzijskim tablicama

TABLICA ČINJENICA -
FACT_ELECTION_RESULT

dim_election

- ▶ Opisuje izbore; sadrži identifikator, datum (date_tk) i zemlju (country_id)

dim_date

- ▶ Omogućuje vremensku analizu: dan, mjesec, godina, dan u tjednu

dim_country

- ▶ Geografski podaci: regije i nazivi zemalja (odnosno izbornih jedinica)

dim_party

- ▶ Podaci o političkim strankama: naziv i politička orijentacija

dim_person

- ▶ Opis političara: ime, prezime, godina rođenja, titula; povezuje osobu sa strankom

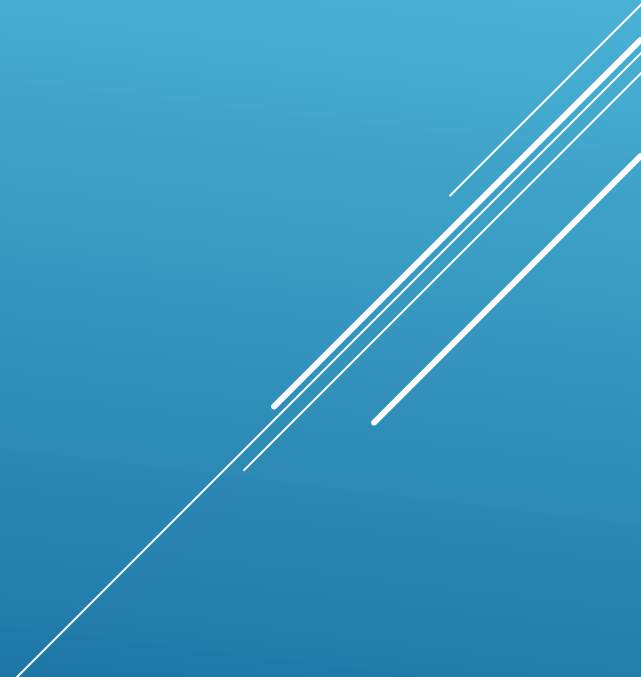
dim_election_history

- ▶ Povijesni kontekst izbora: broj praznih i nevažećih listića, razdoblja, izlaznost

DIMENZIJSKE TABLICE

- ▶ SPORO MIJENJAJUĆE DIMENZIJE (SCD)
- ▶ DEGENERIRANA DIMENZIJA
- ▶ IZGENERIRANI ATRIBUTI I DODATNI PODACI

DODATNE MOGUĆNOSTI DIMENZIJSKOG MODELA



- ▶ Dimenzije koje se rijetko mijenjaju, ali je važno zadržati povijest njihovih promjena
- ▶ Omogućuju analizu stanja kroz vrijeme
- ▶ Primjeri:
 - Dim_election_history -> sadrži vremenske intervale date_from i date_to za praćenje povijesti rezultata izbora
 - Dim_person -> omogućuje praćenje promjena pripadnosti političara različitim strankama kroz vrijeme

SPORO MIJENJAJUĆE DIMENZIJE

- ▶ Nema posebnu dimenzijsku tablicu
- ▶ Pohranjuje se izravno u fact tablici
- ▶ Koristi se isključivo za filtriranje i identifikaciju rezultata
- ▶ Result_type => označava tip rezultata
 - Može biti privremeni ili konačni
 - Svi su rezultati privremeni osim finalnog nakon kojeg nije bilo novih unosa

DEGENERIRANA DIMENZIJA

Neki od podataka nisu postojali u originalnom datasetu, već su ručno/AI generirani:

- ▶ `version`, `date_from`, `date_to`: praćenje verzija izbornih rezultata i intervala kada su one vrijedile
- ▶ `region` u `dim_country`: dodatna geografska hijerarhija
- ▶ `year`, `month`, `day`, `weekday` u `dim_date`: extractani su iz timestampova; omogućuju preciznu vremensku analizu
- ▶ `orientation` u `dim_party`: politička orijentacija (lijevo, desno, centar)
- ▶ `title` u `dim_person`: akademska titula kandidata (npr. doktor, magistar)

IZGENERIRANI ATRIBUTI I DODATNI PODACI

- ▶ Provedba ETL procesa koristeći Apache Spark

CHECKPOINT #4

- ▶ ETL (Extract, Transform, Load) je ključni proces u skladištenju podataka
- ▶ Omogućuje integraciju podataka iz različitih izvora u jedinstveno i strukturirano okruženje

ETL PROCES

- ▶ **Faza Extract (Dohvat podataka)**
- ▶ Izvori: relacijske baze, nestrukturirane datoteke, web servisi, aplikacijski sustavi
- ▶ Cilj: dohvat podataka iz heterogenih sustava

EXTRACT

- ▶ **Faza Transform (Transformacija podataka)**
- ▶ Operacije: čišćenje, normalizacija, obogaćivanje, konsolidacija, mapiranje
- ▶ Cilj: semantičko ujednačavanje i priprema za analizu

TRANSFORM

- ▶ **Faza Load (Učitavanje podataka)**

- ▶ Podaci se učitavaju u skladište podataka (Data Warehouse)
- ▶ Poštuju se pravila integriteta i optimiziraju performanse
- ▶ Osigurava se podatkovna konzistentnost

LOAD

- ▶ Povećava pouzdanost i točnost analitičkih sustava
- ▶ Smanjuje redundantnost i neusklađenost podataka
- ▶ Postavlja temelj za OLAP, izvještavanje i poslovno odlučivanje

VAŽNOST ETL PROCESA

Apache Spark

- ▶ Open-source distribuirani framework za obradu podataka
- ▶ Visoka skalabilnost i brzina zahvaljujući obradi u memoriji
- ▶ Nije klasičan ETL alat, ali izuzetno prikladan za takve zadatke

Primjena u projektu:

- ▶ Dohvat podataka iz više izvora (generiranih pri splittanju dataseta)
- ▶ Transformacija: filtriranje, čišćenje, strukturiranje
- ▶ Učitavanje u analitički model

Prednosti:

- ▶ Podrška za paralelizaciju i distribuciju
- ▶ Učinkovit rad s velikim količinama podataka

ODABRANI ALAT

- ▶ Napomena: ovo je subjektivno mišljenje autora prezentacije
- ▶ Kao alternativa ponuđen je Pentaho, no Spark ima određene prednosti:
 - Omogućen je rad isključivo kroz kod (bez GUI)
 - Veća fleksibilnost u implementaciji
 - Izbjegavanje rada s brojnim tipkama/tabovima/drag&dropovima
 - Lakša uporaba LLM-ova za debugging

PREDNOSTI U ODNOSU NA PENTAHO

- ▶ Dohvat sirovih podataka iz više izvora korištenjem opisanog alata
- ▶ Korištenje Spark dataframeova
- ▶ 2 modula:
 - Učitavanje CSV datoteka
 - Učitavanje datoteke s uključenim zaglavljima
 - Automatska detekcija tipova podataka
 - Povezivanje s MySql bazom podataka
 - Korištenje JDBC connectora (JAVA!)
 - Dohvat podataka iz pojedinačnih tablica opisanih u prethodnoj fazi projekta

EXTRACT NA DATASETU

```
# extract/extract_mysql.py
from spark_session import get_spark_session

def extract_table(table_name):
    spark = get_spark_session("ETL_App")

    jdbc_url = "jdbc:mysql://127.0.0.1:3306/elections_brazil?useSSL=
false"
    connection_properties = {
        "user": "root",
        "password": "root",
        "driver": "com.mysql.cj.jdbc.Driver"
    }

    df = spark.read.jdbc(url=jdbc_url, table=table_name, properties=
connection_properties)
    return df

def extract_all_tables():
    return {
        "country": extract_table("country"),
        "election": extract_table("election"),
        "election_history": extract_table("election_history"),
        "party": extract_table("party"),
        "person": extract_table("person"),
        "result": extract_table("result"),
    }
```

KOD

```
# extract/extract_csv.py
from spark_session import get_spark_session

def extract_from_csv(file_path):
    spark = get_spark_session("ETL Extract - CSV")
    df = spark.read.option("header", True).option("inferSchema", True)
        .csv(file_path)
    return df
```

- ▶ **Modularna organizacija** transformacijskog procesa:
 - ▶ Svaka dimenzijska tablica ima svoju zasebnu transformacijsku funkciju
 - ▶ Centralna skripta `pipeline.py` orkestrira izvođenje svih transformacija
- ▶ Obuhvaćene tablice:
 - ▶ Dimenzije: `dim_country`, `dim_date`, `dim_party`, `dim_election`, `dim_election_history`, `dim_person`
 - ▶ Činjenice: `fact_election_data`
- ▶ Transformacije kombiniraju podatke iz **MySQL baze** i dodatnih **CSV datoteka**
- ▶ Operacije uključuju:
 - ▶ Filtriranje, čišćenje, povezivanje (join) podataka
 - ▶ Konstrukciju konačnog **analitičkog modela** za OLAP obradu

TRANSFORM NA DATASETU (INTRO 1/3)

- ▶ Najsloženiji dio ETL procesa zbog:
 - ▶ Višestrukih međusobno povezanih skripti
 - ▶ Velikog broja potrebnih **joinova**
 - ▶ Potencijalnih grešaka koje mogu zaustaviti cijeli tijek transformacije

TRANSFORM NA DATASETU (HELL 2/3)

- ▶ Svaka dimenzijska tablica ima **zasebnu transformacijsku funkciju**
- ▶ Sve transformacije orkestrira **pipeline skripta** (`pipeline.py`)
- ▶ Prednosti modularnog pristupa:
 - ▶ **Preglednost koda**
 - ▶ **Višekratna iskoristivost**
 - ▶ **Lakše praćenje i izolacija pogrešaka**

TRANSFORM NA DATASETU (PIPELINE 3/3)

```
def run_transformations(raw_data):
    first_row = next(iter(raw_data.items()))
    print(first_row)

    print("\n Starting all transformations...\n")
    print(" [1] Transforming Country dimension...\n")
    try:
        country_dim = transform_country_dim(
            raw_data["country"],
            csv_country_df=raw_data.get("ElectionData")
        )
        print("[1] Country dimension complete\n")
    except Exception as e:
        print(f" [1] Country dimension failed: {e}")
        raise

    print(" [4] Transforming Election dimension...\n")
    try:
        election_dim = transform_election_dim(
            raw_data["election"],
            csv_election_df=raw_data.get("ElectionData")
        )
        print("[4] Election dimension complete\n")
    except Exception as e:
        print(f" [4] Election dimension failed: {e}")
        raise

    print(" [5] Transforming Election History dimension...\n")
    try:
        election_history_dim = transform_election_history_dim(
            raw_data["election_history"],
            csv_election_history_df=raw_data.get("ElectionData")
        )
        print("[5] Election History dimension complete\n")
    except Exception as e:
        print(f" [5] Election History dimension failed: {e}")
        raise

    print(" [6] Transforming Person dimension...\n")
    try:
        person_dim = transform_person_dim(
            raw_data["person"],
            csv_person_df=raw_data.get("ElectionData")
        )
        print("[6] Person dimension complete\n")
    except Exception as e:
        print(f" [6] Person dimension failed: {e}")
        raise

    print(" [3] Transforming Party dimension...\n")
    try:
        party_dim = transform_party_dim(
            raw_data["party"],
            csv_party_df=raw_data.get("ElectionData")
        )
```



```
        print(" [3] Party dimension complete\n")
    except Exception as e:
        print(f" [3] Party dimension failed: {e}")
        raise

    print(" [2] Transforming Date dimension...\n")
    try:
        date_dim = transform_date_dim(
            raw_data["election"],
            csv_date_df=raw_data.get("ElectionData")
        )
        print("[2] Date dimension complete\n")
    except Exception as e:
        print(f" [2] Date dimension failed: {e}")
        raise

    print(" [7] Transforming Election Data fact table...\n")
    try:
        election_data_fact = transform_elections_fact(
            country_dim,
            date_dim,
            party_dim,
            election_dim,
            election_history_dim,
            person_dim
        )
        print("[7] Election Data fact table complete\n")
    except Exception as e:
        print(f" [7] Election Data fact table failed: {e}")
        raise

    print(" All transformations completed successfully!")

    return {
        "dim_country": country_dim,
        "dim_date": date_dim,
        "dim_party": party_dim,
        "dim_election": election_dim,
        "dim_election_history": election_history_dim,
        "dim_person": person_dim,
        "fact_election_data": election_data_fact
    }
```



```
from transform.dimensions.country_dim import transform_country_dim
from transform.dimensions.date_dim import transform_date_dim
from transform.dimensions.party_dim import transform_party_dim
from transform.dimensions.election_dim import transform_election_dim
from transform.dimensions.election_history_dim import
    transform_election_history_dim
from transform.dimensions.person_dim import transform_person_dim
from transform.facts.elections_fact import transform_elections_fact

print(" Pokrećeno pipeline.py")
```

IZGLED PIPELINEA

- ▶ Koristi se funkcija `transform_elections_fact`
- ▶ Podaci dolaze iz CSV datoteke sa svim sirovim atributima o rezultatima izbora po teritorijima
- ▶ Izvršava se čišćenje podataka (npr. standardizacija imena teritorija)
- ▶ Slijedi niz JOIN operacija s dimenzijskim tablicama:
 - ▶ `dim_country`, `dim_election`, `dim_election_history`, `dim_date`
- ▶ Korišteni su LEFT JOIN-ovi za očuvanje svih zapisa iz izvornog skupa podataka, čak i ako nema pripadajućih vrijednosti u dimenzijama
- ▶ Izdvaja se skup relevantnih atributa i oblikuje se konačna fact tablica
- ▶ Generira se jedinstveni identifikator `fact_id` pomoću funkcije `row_number()`
- ▶ Fact tablica se transformira zadnja zbog očuvanja referencijalnog integriteta
- ▶ Joinovi su izazovni zbog usklađivanja naziva stupaca i tablica

TRANSFORMACIJA FACT TABLICE

Funkcija **transform_person_dim** gradi tablicu kandidata (**dim_person**)

- ▶ Normalizira ulazni DataFrame iz MySQL baze:
 - ▶ Uklanja prazna polja i duplikate
 - ▶ Čisti podatke (npr. `trim` na tekstualnim kolonama)
- ▶ Dodatna CSV datoteka s podacima, spaja se s MySQL izvorom koristeći **unionByName**
- ▶ Izvodi se **JOIN** s tablicom **dim_party** radi zamjene izvornog party ID-a s tehničkim ključem (**partytk**) => osigurava da se za svaku osobu evidentira kojoj stranci pripada
- ▶ Potrebna provjera spojeva preko primarnih i tehničkih ključeva radi ispravnosti podataka
- ▶ Na kraju se generira jedinstveni **surrogate ključ** **persontk** koristeći funkciju `row_number()`
- ▶ Time se osigurava jednoznačno indeksiranje tablice

TRANSFORMACIJA DIMENZIJE PERSON

- ▶ Funkcija **transform_party_dim** obrađuje podatke o političkim strankama
- ▶ Ulazni podaci dolaze iz MySQL baze, s dopunama iz CSV izvora
- ▶ U prvoj fazi:
 - ▶ Čišćenje podataka
 - ▶ Povezivanje s unaprijed definiranim rječnikom političkih orijentacija po ID-u stranke
- ▶ Obrada CSV izvora:
 - ▶ Dodatno čišćenje zapisa
 - ▶ Generiranje vlastitog ID-a pomoću `row_number()`
 - ▶ Dodjeljivanje orijentacije "unknown" za nepoznate vrijednosti
 - ▶ Spajanje s MySQL podacima
- ▶ Završna faza:
 - ▶ Generiranje jedinstvenog tehničkog ključa partytk za identifikaciju u skladištu podataka
- ▶ Dimenzija omogućuje analizu izbornih rezultata kroz prizmu političke orijentacije
- ▶ Orijentacije se pridružuju svakoj stranci korištenjem dictionaryja `party_orientations`

TRANSFORMACIJA DIMENZIJE PARTY

- ▶ Dimenzija datuma omogućuje analizu vremenskih aspekata podataka (trendovi po godinama, mjesecima, danima u tjednu)
- ▶ Funkcija **transform_date_dim** koristi vremenske podatke iz baze i CSV datoteka
- ▶ Podaci se obrađuju pomoću pomoćne funkcije **normalize_time_df** koja:
 - ▶ Automatski identificira kolone s datumima (npr. `time`, `year`)
 - ▶ Standardizira formate datuma
 - ▶ Pretvara vrijednosti u **TimestampType**
 - ▶ Uklanja nedosljednosti poput duplikata i NULL vrijednosti
- ▶ Izvorni podaci se spajaju u koherentnu tablicu
- ▶ Dodaju se dodatni atributi dimenzije: godina, mjesec, dan, dan u tjednu nastali ekstrakcijom `timestampa`
- ▶ Svakom zapisu se dodjeljuje jedinstveni tehnički ključ `date_tk`
- ▶ Omogućuje jednostavno vremensko filtriranje i grupiranje u OLAP sustavu

TRANSFORMACIJA DIMENZIJA DATE

- ▶ Analiza rezultata na teritorijalnoj bazi
 - Izborne jedinice (savezne države Portugala)
 - Čišćenje i formatiranje naziva
 - Uklanjanje tipičnih portugalskih slova koja bi stvarala probleme u Python encodingu
 - Uklanjanje duplikata
 - Nasumično punjenje dodatnog hijerarhijskog atributa region
 - Dodavanje tehničkog ključa

TRANSFORMACIJA DIMENZIJE COUNTRY

Transformacija i integracija podataka o izborima iz MySQL baze i opcionalnog CSV izvora

- ▶ Iz MySQL baze se uzimaju stupci: id, year i countryid
- ▶ Konverzija id i countryid u numeričke tipove
- ▶ Pretvorba godine u string pa u format datuma yyyy-MM-dd
- ▶ Uklanjanje redaka s nedostajućim vrijednostima i duplikata
- ▶ Ako postoji CSV, primjenjuje se isti postupak čišćenja i tipizacije
- ▶ Spajanje podataka iz oba izvora u jedinstveni DataFrame
- ▶ Join s tablicama dim_date i dim_country radi zamjene stvarnih vrijednosti surrogate ključevima
- ▶ Dodjeljuje se surrogate ključ electiontk korištenjem rownumber funkcije
- ▶ Podaci se sortiraju, čiste i pripremaju za učitavanje u skladište
- ▶ Osigurava se dosljednost, bez dupliciranja i s pravilnim vezama na ostale dimenzije

TRANSFORMACIJA DIMENZIJE ELECTION

- ▶ Korišten je drukčiji pristup jer je bilo potrebno dodavati mnogo novih atributa kojih nije bilo u originalnom datasetu, zbog potrebe praćenja podataka kroz vrijeme
- ▶ Python skripta puni dimenzijsku tablicu `dimelectionhistory` iz CSV datoteke koristeći Pandas i MySQL
- ▶ Koraci u skripti: učitavanje CSV-a, konverzija vremena, izračun povijesne izlaznosti birača (`historicalturnout`), umetanje ili ažuriranje podataka
- ▶ Provjera postoji li već zapis za teritorij s `iscurrent = TRUE`
- ▶ Ako postoji, zapis se ažurira: postavlja mu se `dateto` na trenutni datum i `iscurrent = FALSE`
- ▶ Unosi se nova verzija zapisa s povećanim brojem verzije i `iscurrent = TRUE`
- ▶ Dimenzija je složena zbog višestrukih referenciranja i povezanosti s drugim tablicama
- ▶ Koristi se koncept Slowly Changing Dimension Type 2 (SCD2) za praćenje povijesti promjena
- ▶ Svaka promjena teritorija bilježi se s vremenskim oznakama `datefrom` i `dateto`
- ▶ Omogućeno je praćenje kako su se atributi teritorija mijenjali kroz vrijeme
- ▶ Ključno za analitičke potrebe skladišta podataka

TRANSFORMACIJA DIMENZIJE ELECTION_HISTORY

- ▶ Funkcija `writesparkdftomysql` omogućuje učitavanje transformiranih podataka iz Spark DataFrame-a u MySQL bazu koristeći `write.jdbc()`
- ▶ Omogućuje fleksibilno i višekratno učitavanje podataka u fazi Load ETL procesa bez ručne intervencije, uz podršku za različite načine učitavanja (npr. `append`)

LOAD NA DATASETU

JESU LI PODACI DOISTA LOADANI?



country_tk	country_id	name	region
1	4	Braga	Center
2	11	Leiria	Center
3	12	Lisboa	Center
4	6	Castelo Branco	East
5	9	Faro	East
6	13	Madeira	East
7	10	Guarda	North

45	2019-10-06	2019	10	6	Sunday	2019-10-07 05:00:00
46	2019-10-06	2019	10	6	Sunday	2019-10-07 05:15:00
47	2019-10-07	2019	10	7	Monday	2019-10-07 05:30:00
48	2019-10-07	2019	10	7	Monday	2019-10-07 05:45:00

	party_tk	party_id	name	orientation
►	215	1	PS	Center
	216	2	PPD/PSD	Left
	217	3	B.E.	Center
	218	4	CDS-PP	Center
	219	5	PCP-PEV	Center
	220	6	PAN	Center
	221	7	CH	Right

	person_tk	person_id	first_name	last_name	birth_year	title	party_id	gender
►	1	1	Manuel	Martins	1962	Prof.	215	M
	2	2	Maria	Ferreira	1952	Prof. ^a	216	F
	3	3	Miguel	Lopes	1991	Dr.	217	M
	4	4	Tiago	Ferreira	1950	Dr.	218	M
	5	5	Joana	Silva	1995	Sra.	219	F

election_tk	election_id	date_tk	country_id
265	265	13	250
266	266	13	251
267	267	13	252
268	268	13	253
269	269	13	254

JESU LI PODACI DOISTA LOADANI?

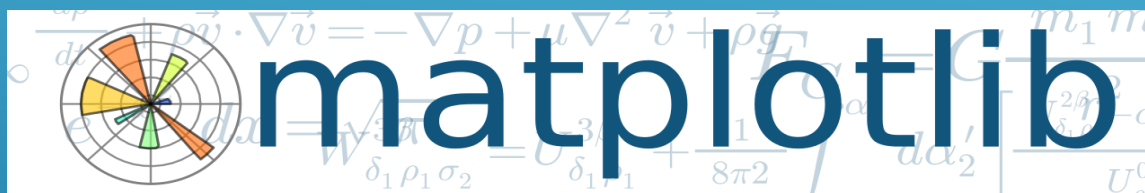
- ▶ Vizualizacija i analiza podataka iz skladišta

CHECKPOINT #5

- ▶ Izrađen web-based dashboard za interaktivnu analizu izbornih rezultata
- ▶ Odluka da se ne koriste alati poput Tableaui Power BI zbog ograničene fleksibilnosti
- ▶ Vlastito rješenje omogućuje potpunu prilagodbu strukturi dimenzijskog modela
- ▶ Precizno vizualizacijsko sučelje dizajnirano prema specifičnostima domene

CUSTOM DASHBOARD ZA VIZUALIZACIJU PODATAKA

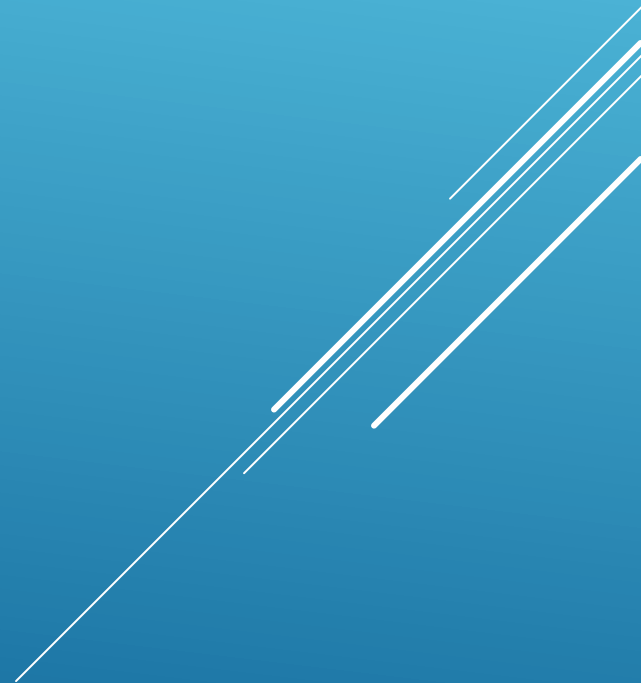
- ▶ Vue.js za responzivni i interaktivni frontend
- ▶ Matplotlib za generiranje statičkih i dinamičkih grafova na serveru
- ▶ Metabase za brzu izradu osnovnih vizualizacija tijekom razvoja i verifikacije



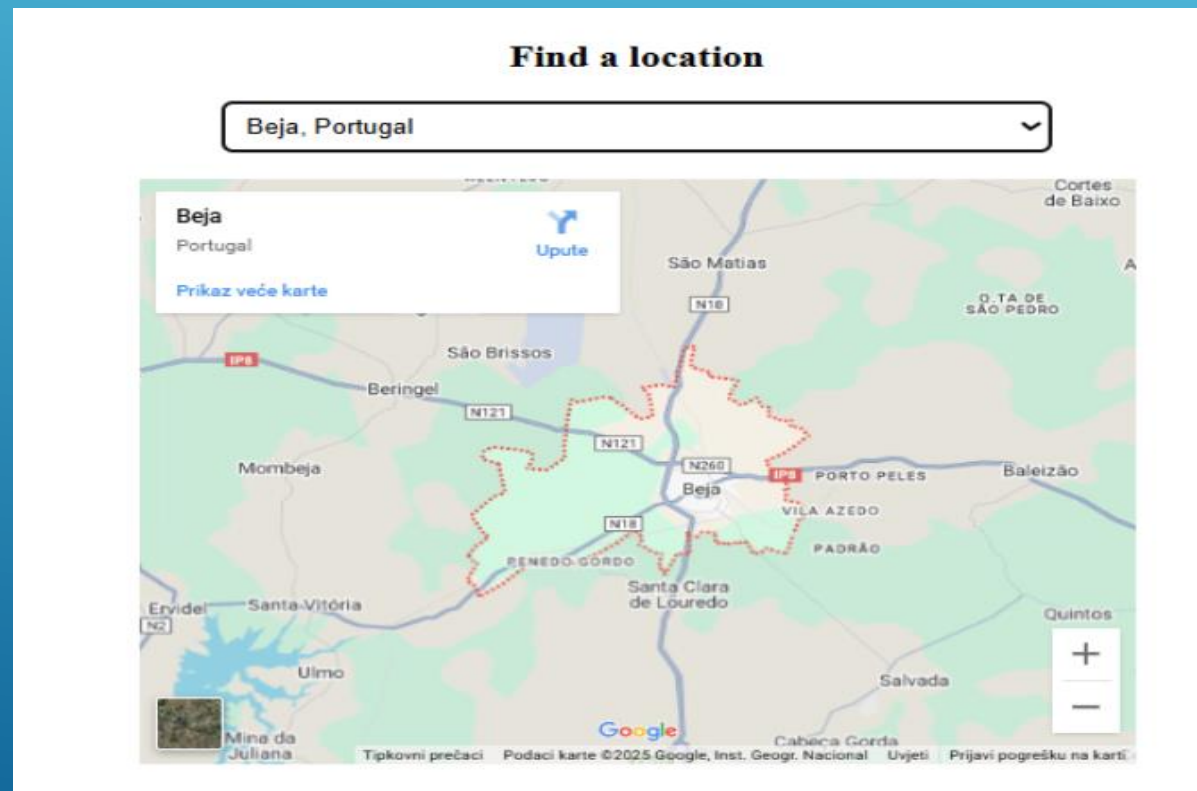
TEHNOLOGIJE



FUNKCIONALNOSTI DASHBOARDA



- Dinamički pronalazak izbornih jedinica (dim_country) na interaktivnoj karti

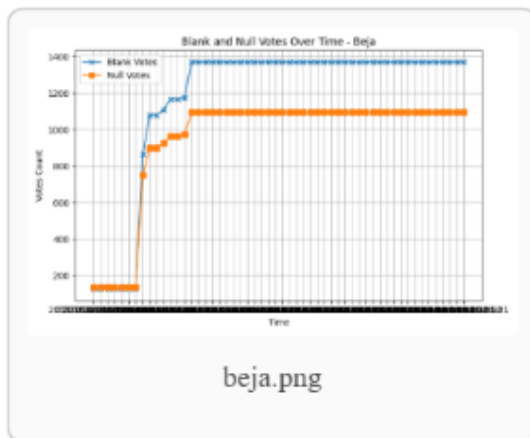


- ▶ Pretraživanje rezultata po teritoriju (prevladavajuće političke orijentacije & prazni i nevažeći glasovi)
- ▶ Ekvivalentno OLAP slice operaciji (filtriranje po 1 dimenziji; teritoriju)

Search Territory Results

Beja











Search



- ▶ Pregled kandidata, stranaka i grbova
- ▶ Filtriranje na temelju više kriterija (npr. stranka, godina rođenja)
- ▶ Ekvivalentno OLAP Dice operaciji (filtriranje po više dimenzija istovremeno)

Politicians and Their Parties

Party: Birth Year: Condition:

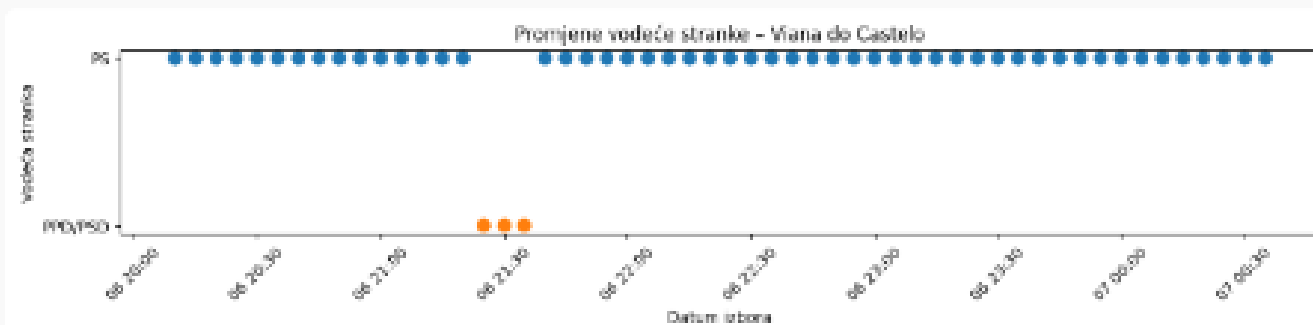
 Dr. João Silva Born: 1963 Party: PS	 Eng. Carlos Pereira Born: 1970 Party: PSD	 Rafael Costa Born: 1982 Party: S.E.
 Prof. Miguel Fernandes Born: 1958 Party: CDS-PP	 Dra. Ana Santos Born: 1975 Party: PAN	 Bruno Almeida Born: 1980 Party: PCIP-MDPP
 Ricardo Neves Born: 1967 Party: A	 Eng. Fernanda Martins Born: 1973 Party: I	 Gustavo Lima Born: 1985 Party: JPP
 Dra. Patrícia Ramos Born: 1990 Party: PDR	 Tiago Moreira Born: 1969 Party: PSN	 Vasco Mendes Born: 1972 Party: PLR
 Dr. Diogo Faria Born: 1989 Party: PPM	 Holena Cruz Born: 1981 Party: MP	 Prof. Sofia Teixeira Born: 1978 Party: MAS
 Eng. André Rodrigues Born: 1973 Party: PCP-PEV	 Caterina Barros Born: 1987 Party: S.I.E.	 José Nogueira Born: 1983 Party: CH
 Dra. Mariana Henriques Born: 1978 Party: IL	 Filipe Coelho Born: 1991 Party: NC	 Dr. Eduardo Gomes Born: 1963 Party: FTF

- ▶ Vrlo važan aspekt skladišta podataka
- ▶ Na raspolaganju je jako mali vremenski period za podatke (2 dana)
- ▶ Ipak dovoljno za vremensku analizu

PRAĆENJE PROMJENA KROZ
VRIJEME?

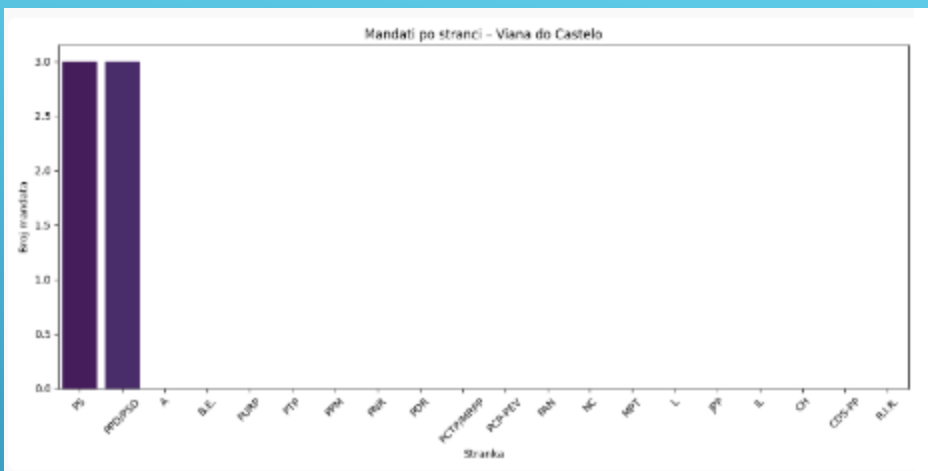
- ▶ Prikaz promjena broja glasova po strankama kroz izborne cikluse
- ▶ Izračun postotnog udjela glasova po strankama
- ▶ Identifikacija vodeće stranke u svakom izbornom ciklusu
- ▶ Vizualizacija konačne raspodjele mandata unutar teritorija
- ▶ Analiza omogućuje uvid u trendove i promjene u biračkom tijelu
- ▶ Podaci korisni za političke stranke, analitičare i istraživače
- ▶ Četiri vrste analize po teritoriju:
 - ▶ Promjene u vodećoj poziciji kroz vrijeme
 - ▶ Broj mandata po stranci
 - ▶ Postotak glasova kroz vrijeme
 - ▶ Broj glasova kroz vrijeme
- ▶ Primjer analize prikazan za teritorij Viana do Castelo

Changes in election leading party



Ekvivalentno drill-downu

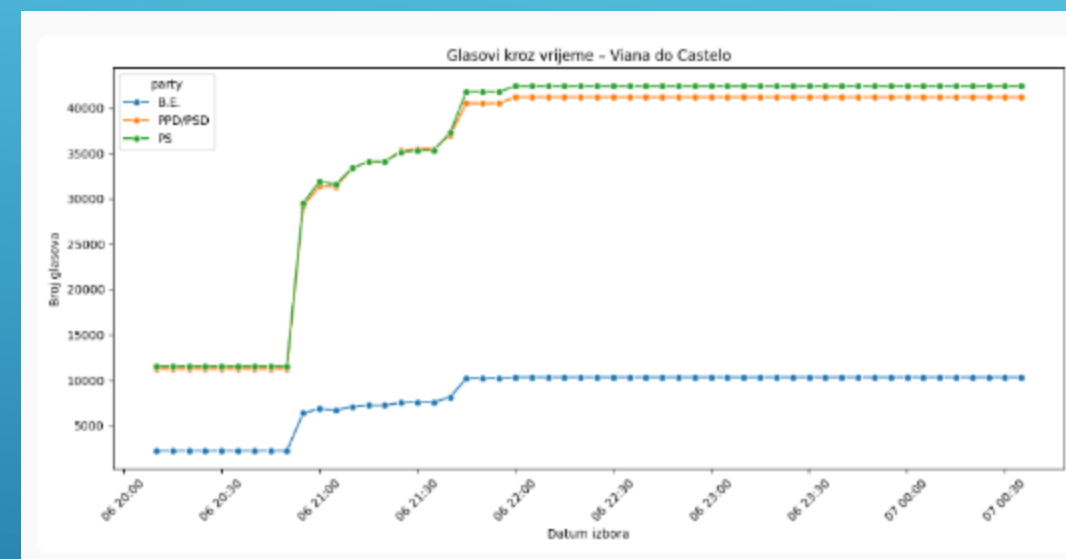
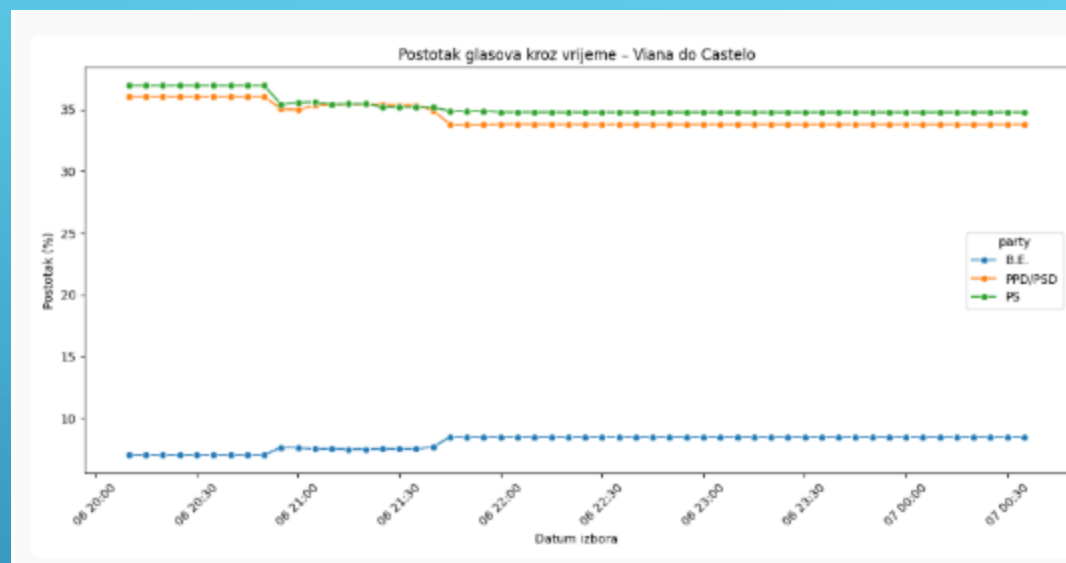
PREGLED VODEĆIH STRANAKA NA
ODREĐENOM TERITORIJU



Filtriranje po teritoriju -> SLICE
Agregiranje mandata po strankama -> ROLL-UP

Napomena: broj mandata po teritoriju nije fiksna te ovisi o broju stanovnika samog teritorija

PREGLAD BROJA MANDATA PO STRANCI NA ODREĐENOM TERITORIJU



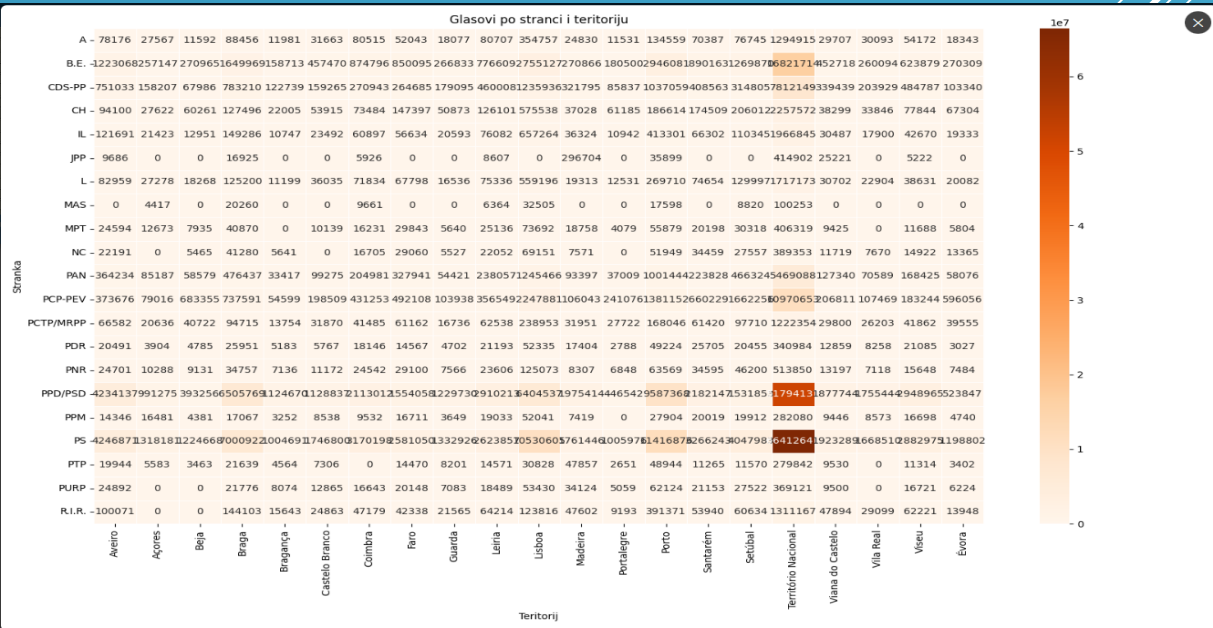
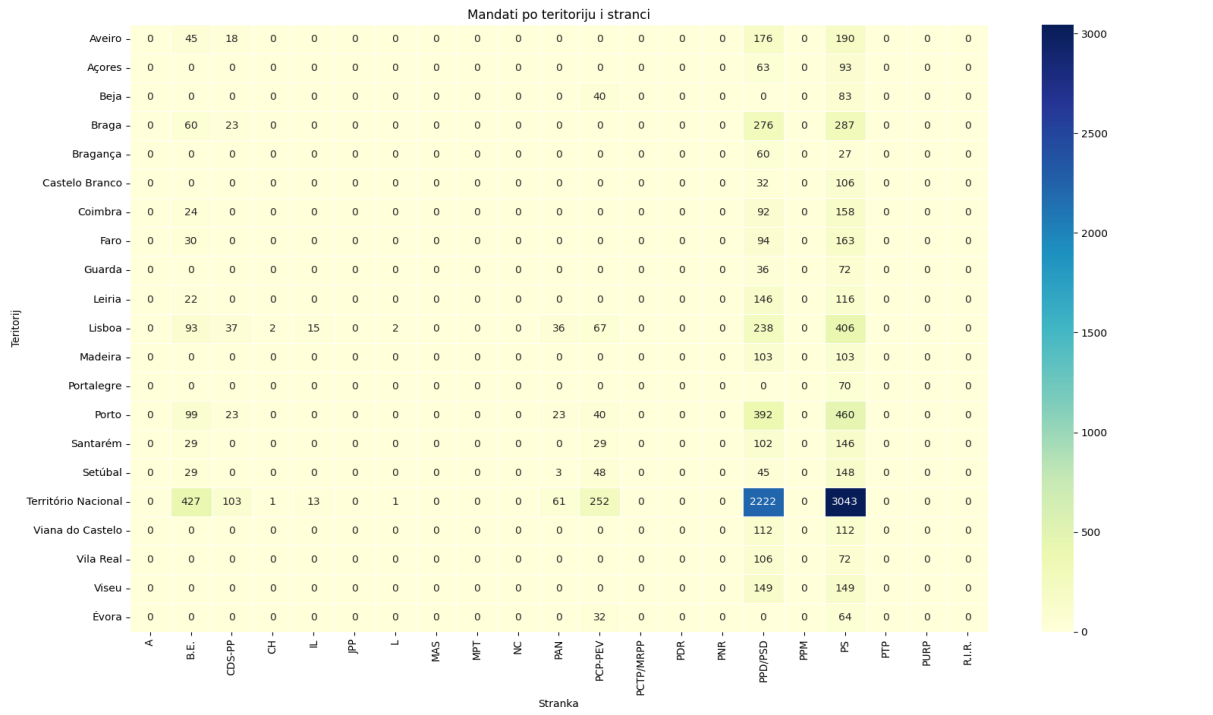
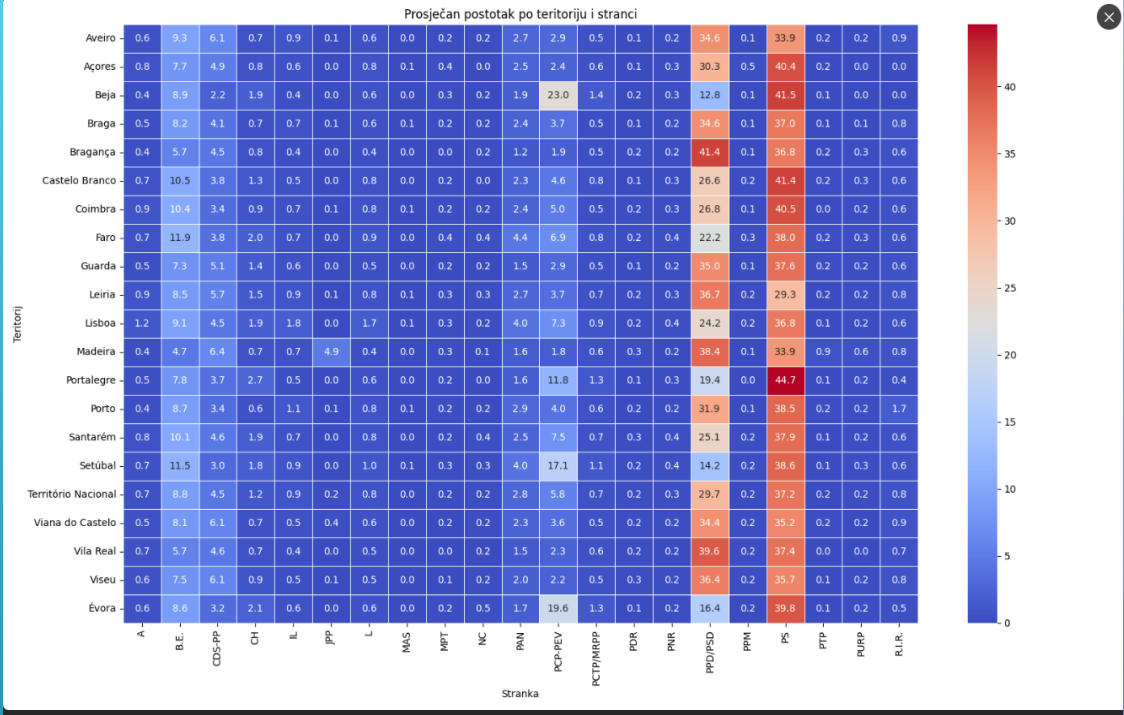
POSTOTAK GLASOVA & BROJ GLASOVA KROZ VRIJEME

- ▶ Dva prethodna grafikona prikazuju promjenu podrške stranci kroz vrijeme:
 - ▶ Apsolutni broj glasova
 - ▶ Postotak osvojenih glasova
- ▶ Primjena **slice** operacije za filtriranje podataka po jednoj stranci
- ▶ Primjena **drill-down** operacije za analizu po izbornim ciklusima
- ▶ Usporedba apsolutnih i relativnih vrijednosti za bolju interpretaciju
- ▶ Prikaz trenda podrške stranci kroz vrijeme
- ▶ Omogućuje analizu relativne snage u odnosu na sve birače
- ▶ Otkivanje suptilnih promjena koje nisu vidljive iz sirovih brojki
- ▶ Doprinos dubljem razumijevanju političke dinamike po teritorijima

OBJAŠNJENJE

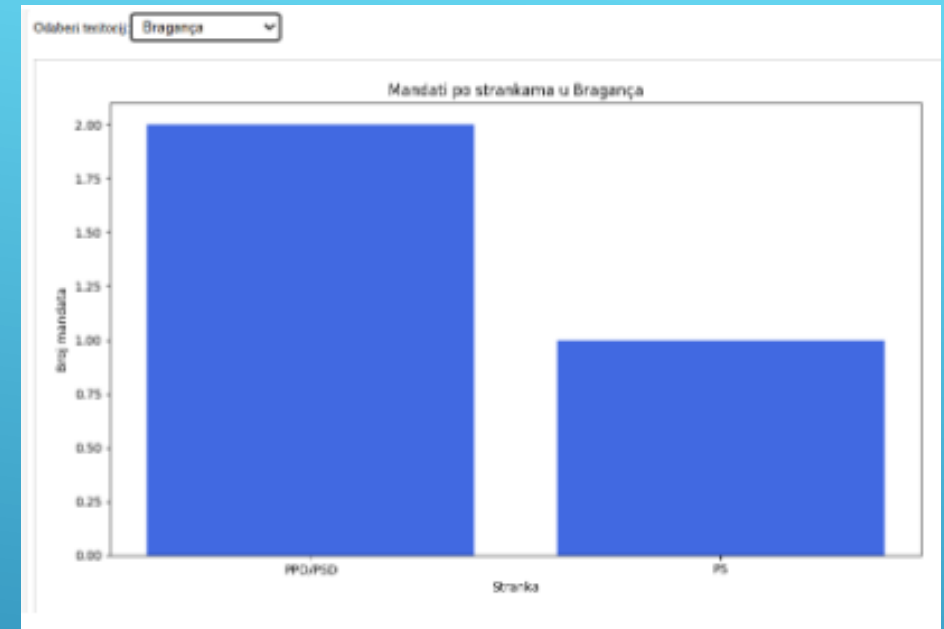
- ▶ **Pivot reorganizira prikaz podataka** - zamjena redaka i stupaca radi bolje preglednosti.
- ▶ **Ne otkriva nove informacije**, već daje alternativni pogled na iste podatke.
- ▶ Korisno za **vizualnu usporedbu više entiteta** (npr. stranke kao stupci, jedinice kao redovi).

PIVOT



PRIMJERI

- ▶ Klasični slice
- ▶ Dice ako bi se dodatno još filtriralo po stranci
- ▶ Roll-up ako bi se promatranje prebacilo na nivo cijele države

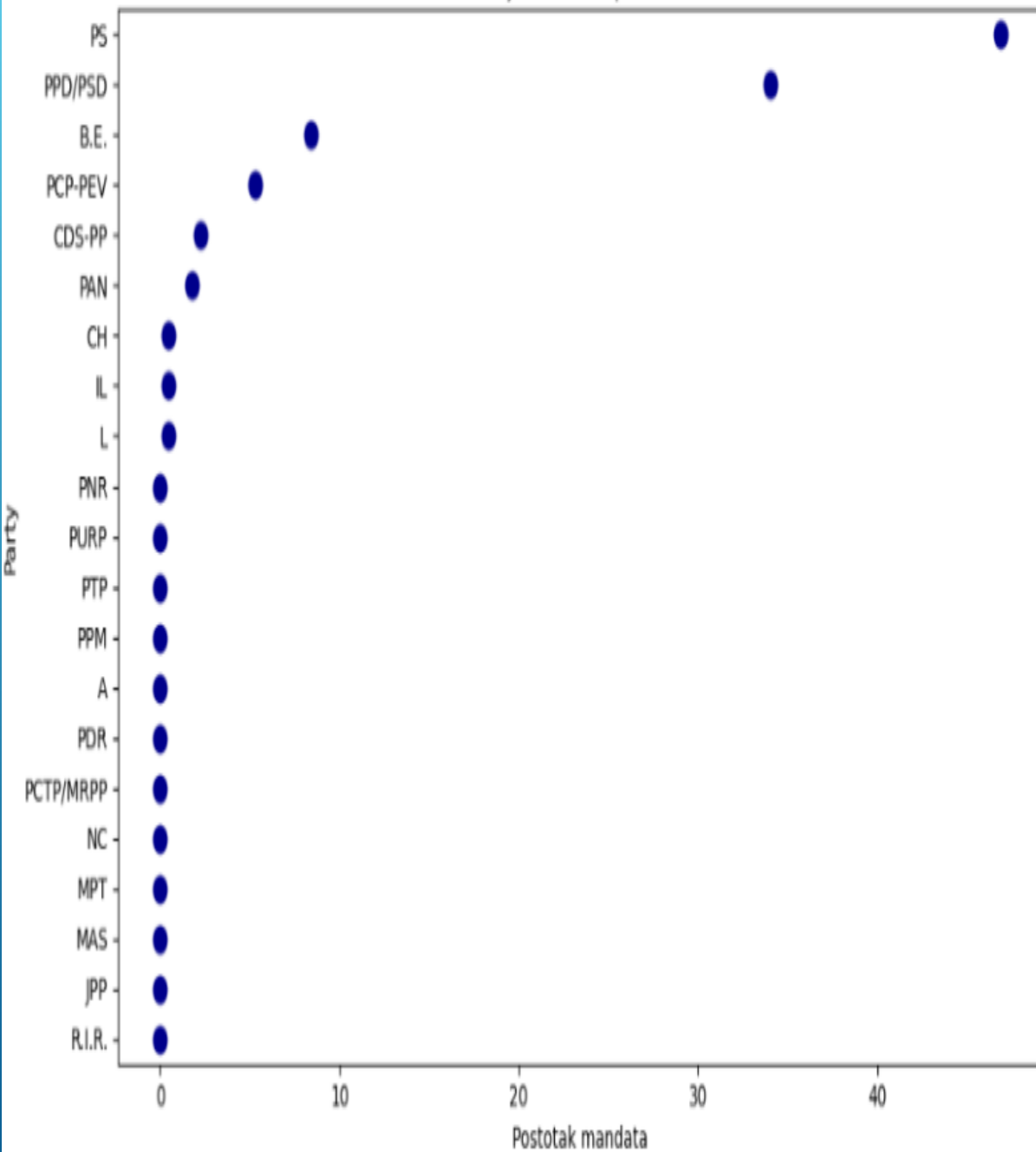


KONAČNI REZULTATI PO TERITORIJU

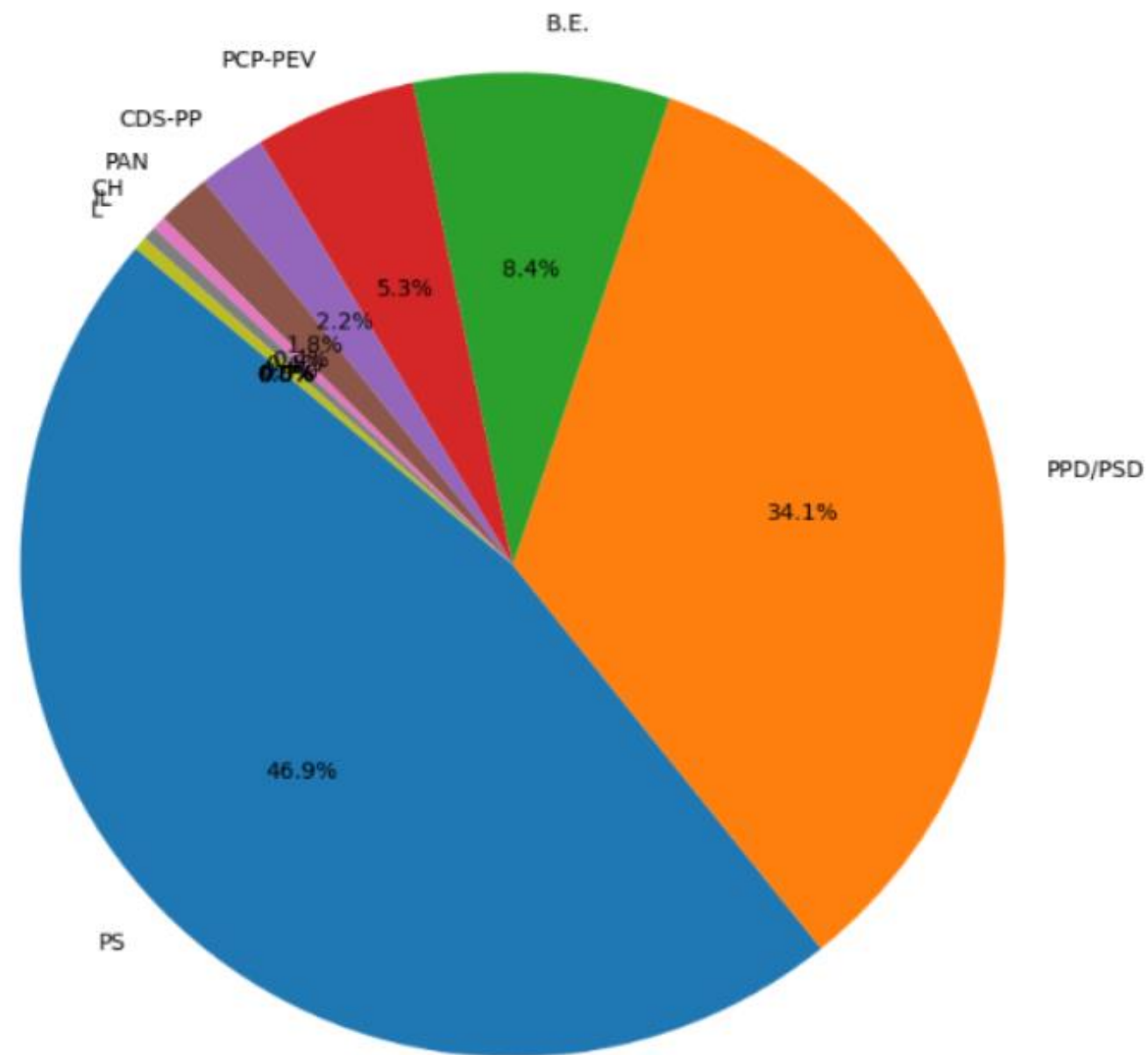
- ▶ 3 perspektive za promatranje konačnih rezultata
 - Stupčasti grafikon postotka mandata po strankama
 - Distribucija mandata u postocima
 - Pie chart konačnih rezultata

KONAČNI REZULTATI

Distribucija mandata po strankama

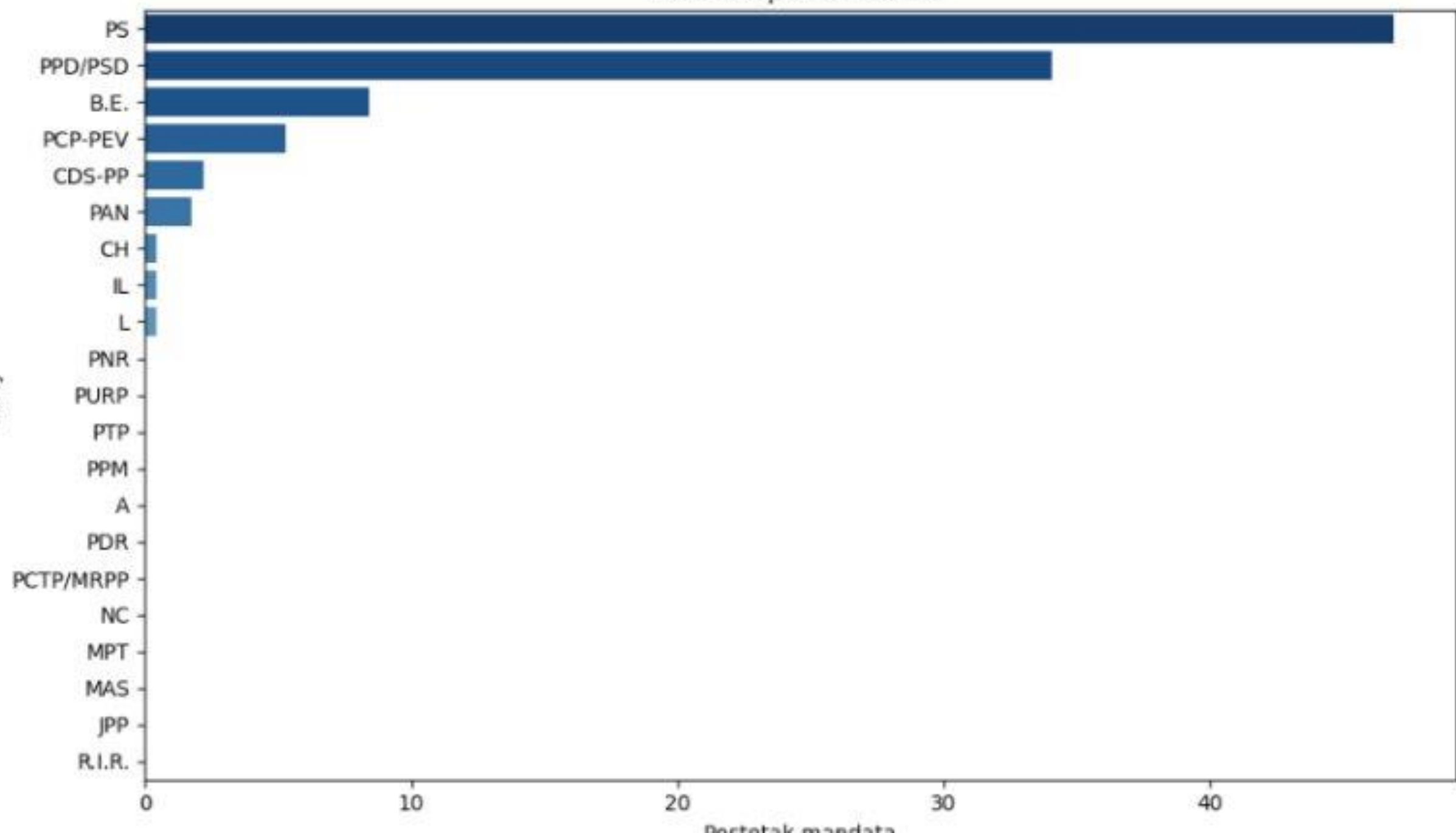


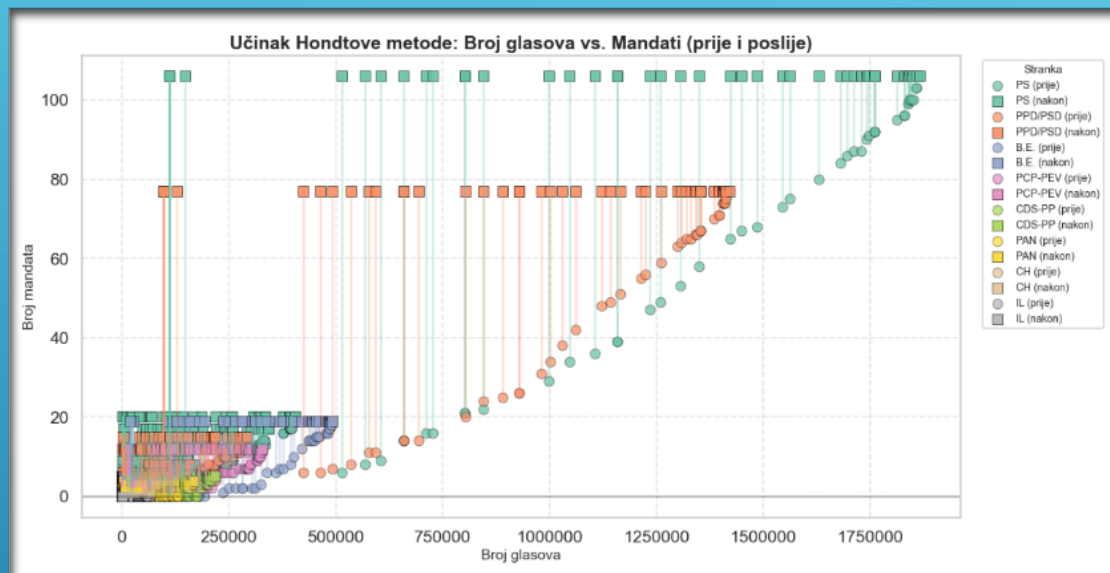
Finalni postotak mandata po strankama



Mandati po strankama

Party

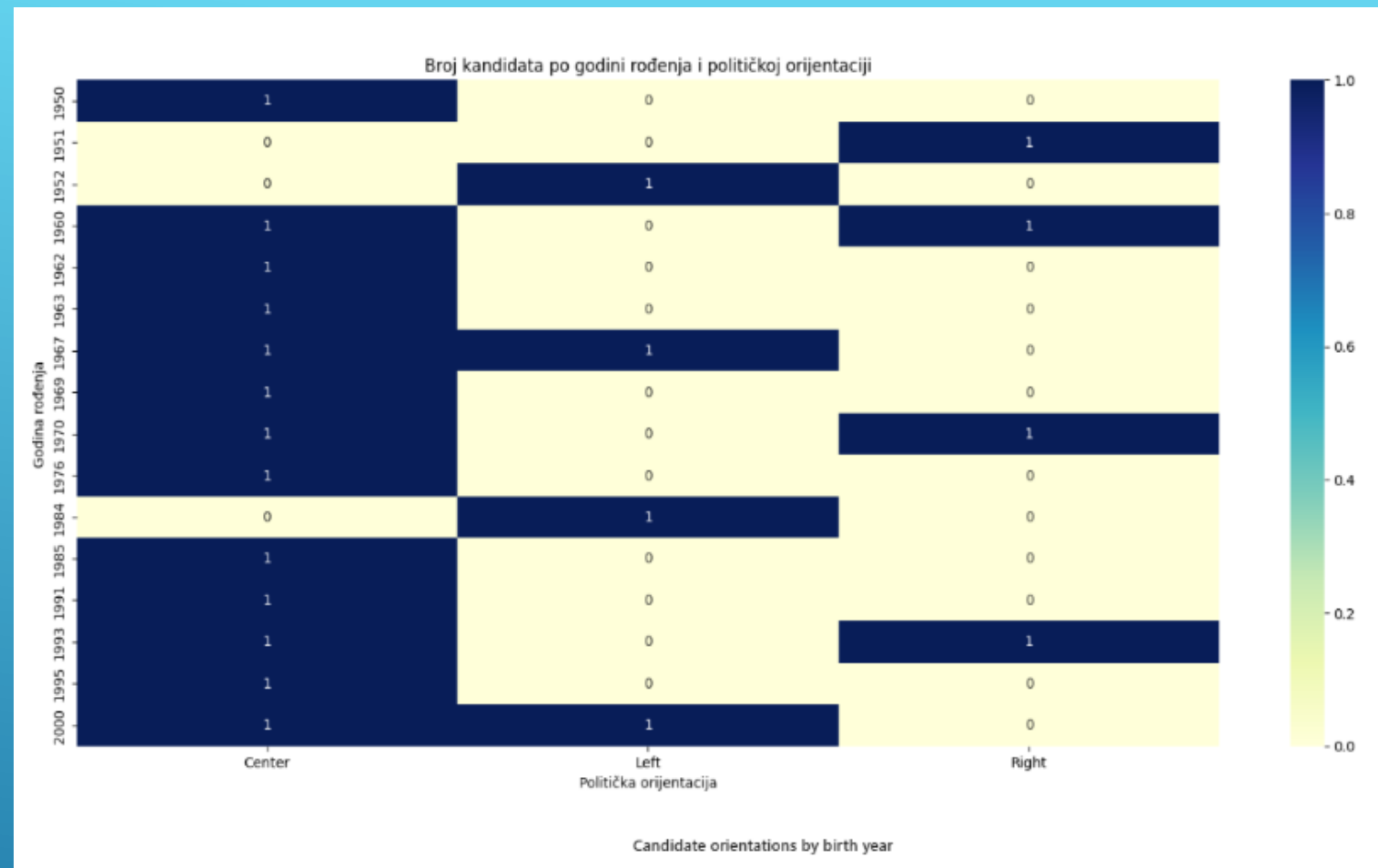




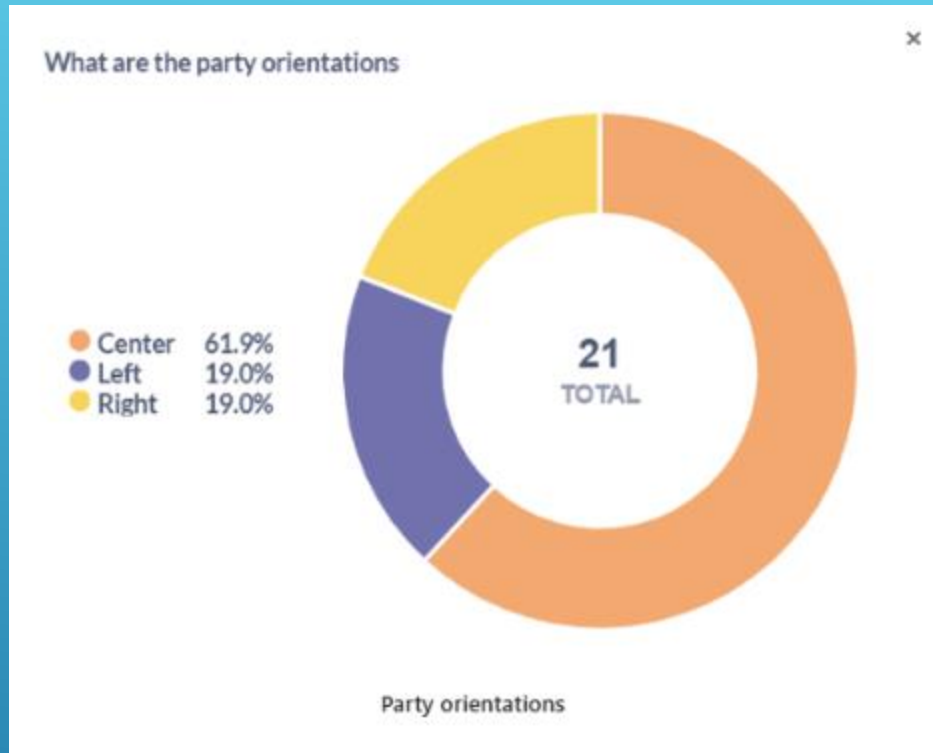
- ▶ Preračunavanje glasova u mandate
- ▶ Često nepošteno

D'HONDT

- ▶ 2 dimenzije => dice
- ▶ Heatmap grafikon



BROJ KANDIDATA PO GODINI ROĐENJA & POLITIČKOJ ORIJENTACIJI



Roll-up -> 21 stranka se svodi na 3 grupe

POSTOCI ORIJENTACIJA STRANKA

- ▶ Intuitivno: dice jer se filtrira po 2 dimenzije istovremeno
- ▶ Zapravo: slice jer u Portugalu muškarci i žene imaju različite službene titule

KANDIDATI PO SPOLU I TITULI

title	number	gender
Prof.	1	M
Prof. ^a	3	F
Dr.	6	M
Sra.	2	F
Dra.	1	F
Eng.	4	M
Eng. ^a	2	F
Sr.	2	M

- ▶ Roll-up -> promatramo srodne titule za muškarce i žene (npr. Dr i DrA)
- ▶ Agregacija

Title	Number
Prof.	4
Dr.	7
Sr.	4
Eng.	6

KANDIDATI PO TITULI NEOVISNO O
SPOLU

- ▶ Izgrađen je cjelovit sustav za analizu parlamentarnih izbora u Portugalu 2019.
- ▶ Proces je uključivao:
 - ▶ analizu i pripremu dataseta
 - ▶ izradu relacijskog i dimenzijskog modela
 - ▶ provedbu ETL procesa koristeći Apache Spark
 - ▶ razvoj vizualnog dashboarda za interaktivnu analizu
- ▶ Korištenjem vlastitog rješenja (umjesto gotovih alata) osigurana je fleksibilnost i dubinska kontrola nad prikazom i analizom podataka
- ▶ Omogućena je vremenska i prostorna analiza rezultata, kao i pregled strukture biračkog tijela po strankama i teritorijima
- ▶ Sustav je spreman za proširenje na druge izbore i domene uz minimalne prilagodbe

ZAKLJUČAK

- ▶ Rick Sherman: *Business Intelligence Guidebook: From Data Integration to Analytics*, Morgan Kaufmann, 2015
- ▶ Ralph Kimball & Margy Ross: *The Data Warehouse Toolkit*, John Wiley & Sons, 2013
- ▶ Kimball, Ross, Thornthwaite, Mundy, Becker: *The Data Warehouse Lifecycle Toolkit*, 2. izd., John Wiley & Sons, 2008
- ▶ UC Irvine Machine Learning Repository: *Portuguese Parliamentary Elections Dataset*
- ▶ (<https://archive.ics.uci.edu/dataset/513/>), pristupljeno 31. 5. 2025.

LITERATURA