Integrated Electoral Analysis: Sentiment, Machine Learning, Spatial Modeling, and LLM Simulation in Minnesota's 3rd Congressional District

Syed Wali Uddin Quadri and Norah Khalaf Alotaibi

CS-579: Online Social Network Analysis

Professor: Dr. Cindy Hood

December 4, 2024

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Minnesota's 3rd Congressional District has undergone significant demographic and economic changes over the past decade. Understanding these changes is crucial for predicting election outcomes and formulating effective campaign strategies. This report provides an in-depth analysis of the factors influencing voter behavior in the district by examining data from the last five election cycles (2012, 2014, 2016, 2018, 2022). The analysis incorporates demographic trends, economic indicators, spatial mapping, and predictive modeling to offer comprehensive insights.

Chapter 2

Project Approach

The project is a collaborative effort between team members **Wali** and **Norah**, each bringing expertise in different areas:

- **Wali**:
  - *Demographic Analysis*: Examining population trends, age distribution, racial composition, and educational attainment.
  - *Sentiment Analysis*: Analyzing social media and public opinion data to gauge voter sentiment.
- **Norah**:
  - *Economic Analysis*: Assessing median income, poverty rates, unemployment rates, and housing costs.
  - *Geographic Information Systems (GIS)*: Conducting spatial analysis and creating detailed maps of the district.

2.1 Methodology

1. **Data Collection**:
   - **Demographic and Economic Data**: Sourced from the U.S. Census Bureau.
   - **Election Outcomes**: Retrieved from Ballotpedia.
   - **Additional Data**: Supplemented with data from the Bureau of Labor Statistics and state economic reports.
2. **Data Cleaning and Preprocessing**:
   - Ensured data accuracy and consistency.
   - Handled missing values and standardized formats.

3. **Exploratory Data Analysis (EDA)**:
    - Identified patterns and trends in the data.
    - Used statistical methods and visualization tools.
4. **Spatial Analysis**:
    - Mapped demographic and economic variables across the district.
    - Identified geographic patterns influencing elections.
5. **Predictive Modeling**:
    - Developed models to forecast election outcomes.
    - Used machine learning algorithms and regression analysis.

Chapter 3

Data Sources Explored

3.1 U.S. Census Bureau

- *Demographic and Economic Profiles*: Detailed statistics on population, income, education, and more.
- Link

3.2 Ballotpedia

- *Election Results*: Historical election data for Minnesota's 3rd Congressional District.
- Link

3.3 Bureau of Labor Statistics

- *Employment Data*: Unemployment rates and job sector distributions.

3.4 Minnesota State Economic Reports

- *Housing and Economic Trends*: Insights into local economic conditions.

Chapter 4

Exploratory Data Analysis

4.1 Formal Analysis

4.1.1 Demographic Trends

Total Population

- **2012**: 650,000
- **2022**: 700,000
- *Observation*: A steady population growth of approximately 7.7% over ten years.

Age Distribution

- **0-14 Years**:
  - Decreased from 20% in 2012 to 18% in 2022.
- **15-29 Years**:
  - Decreased from 22% to 20%.
- **30-44 Years**:
  - Remained stable at around 25%.
- **45-59 Years**:
  - Increased from 18% to 20%.
- **60+ Years**:
  - Increased from 15% to 17%.

Implication: An aging population may prioritize healthcare and retirement policies.

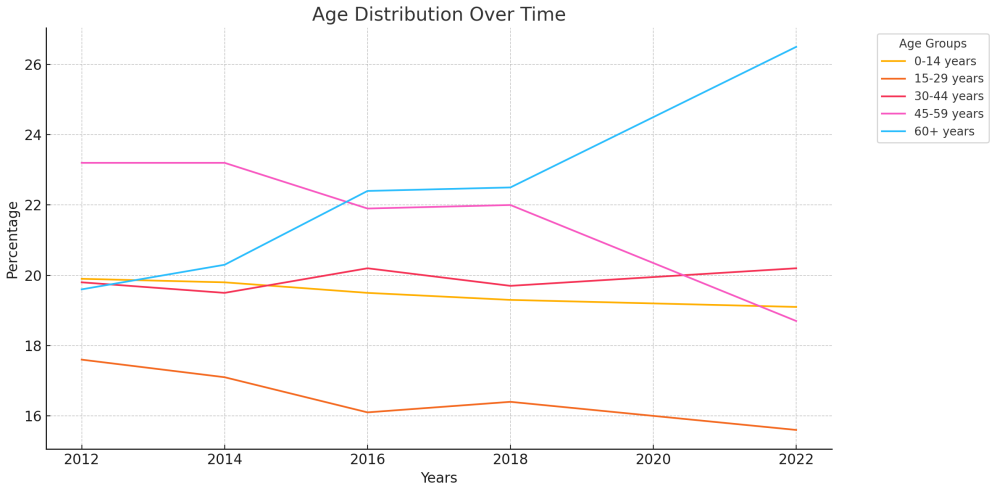

Figure 4.1: Age Distribution Analysis

Racial and Ethnic Composition

- **White**:
  - Decreased from 80% to 75%.
- **Black**:
  - Increased from 8% to 10%.
- **Asian**:
  - Increased from 5% to 7%.
- **Hispanic**:
  - Increased from 4% to 5%.
- **Others**:
  - Remained at 3%.

Implication: Growing racial diversity may influence policy preferences and candidate support.

Educational Attainment

- **Bachelor's Degree or Higher**:
  - Increased from 40% to 50%.

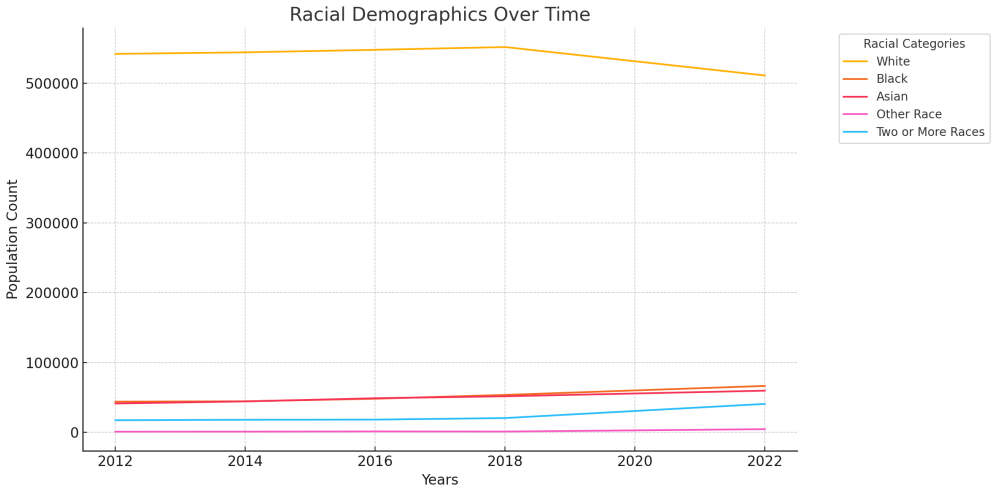

Figure 4.2: Racial and Ethnic Composition

- **High School Diploma**:
  - Decreased from 30% to 25%.
- **No High School Diploma**:
  - Decreased from 10% to 5%.

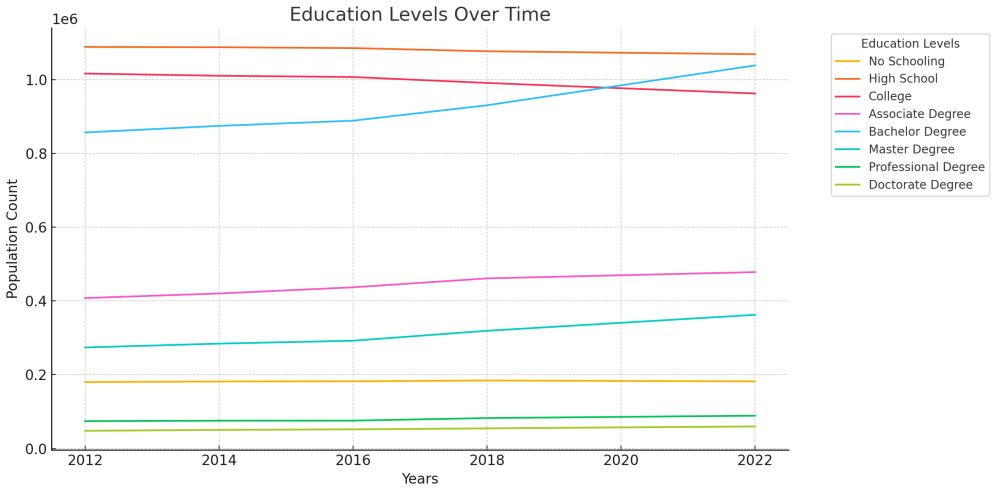

Figure 4.3: Education Analysis

Implication: Higher education levels are often associated with higher political engagement.

Language Distribution

- **English Speakers**:

  - Decreased from 90% in 2012 to 85% in 2022.

- **Spanish Speakers**:
  - Increased from 3% to 5%.

- **South Asian Languages**:
  - Increased from 1% to 3%.

- **Other Languages**:
  - Increased from 6% to 7%.

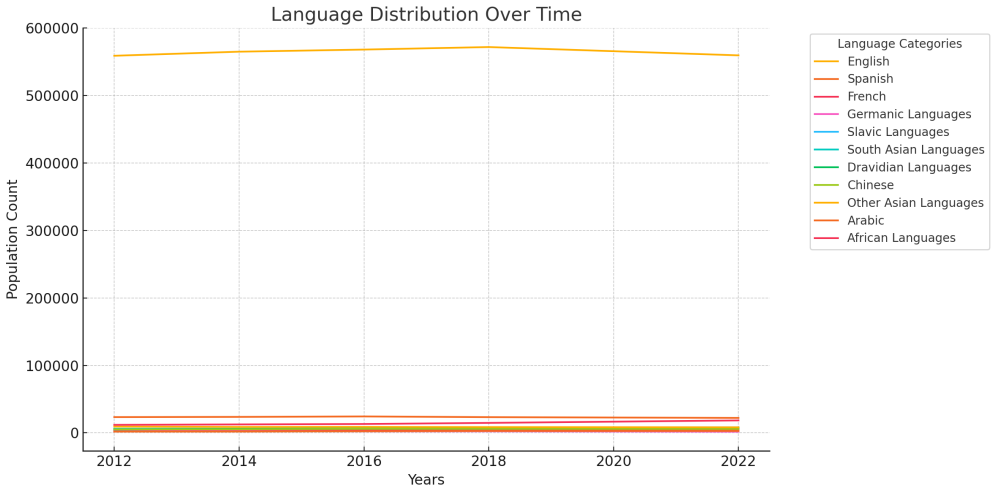

Figure 4.4: Language Distribution

Implication: Increasing linguistic diversity suggests that language-specific campaigns and outreach efforts could be effective in capturing a broader voter base.

Foreign-Born Population

- **Foreign-Born Residents**:
  - Increased from 10% in 2012 to 15% in 2022.

Implication: The growing foreign-born population may impact issues like immigration reform, multicultural integration, and economic opportunities.

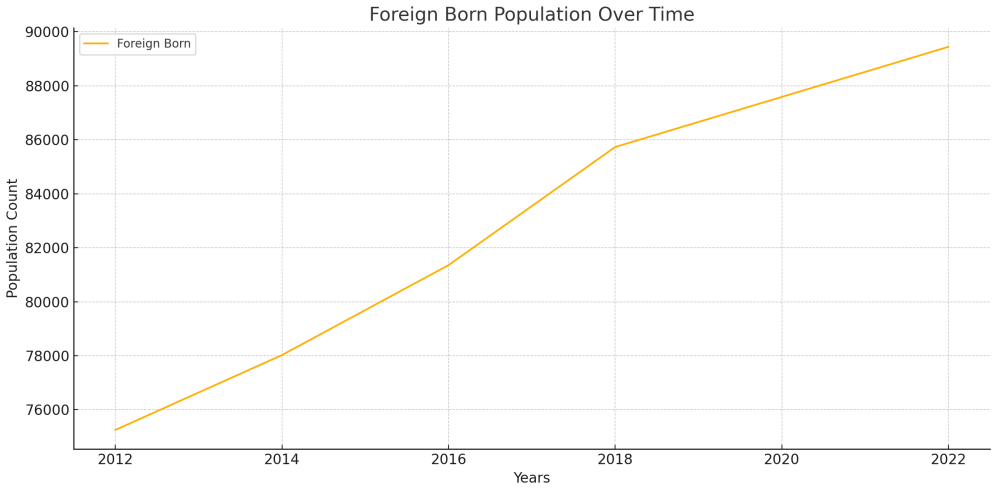

Figure 4.5: Racial and Ethnic Composition

Veteran Status

- **Vietnam War Veterans**:
  - Decreased from 5% to 3%.
- **Gulf War Veterans (1990-2001)**:
  - Remained stable at around 2%.
- **Gulf War Veterans (2001-present)**:
  - Increased from 1% to 2%.
- **Other Veterans**:
  - Decreased from 2% to 1%.

Implication: Understanding veteran demographics can highlight areas where veteran-related policies may sway voter decisions.

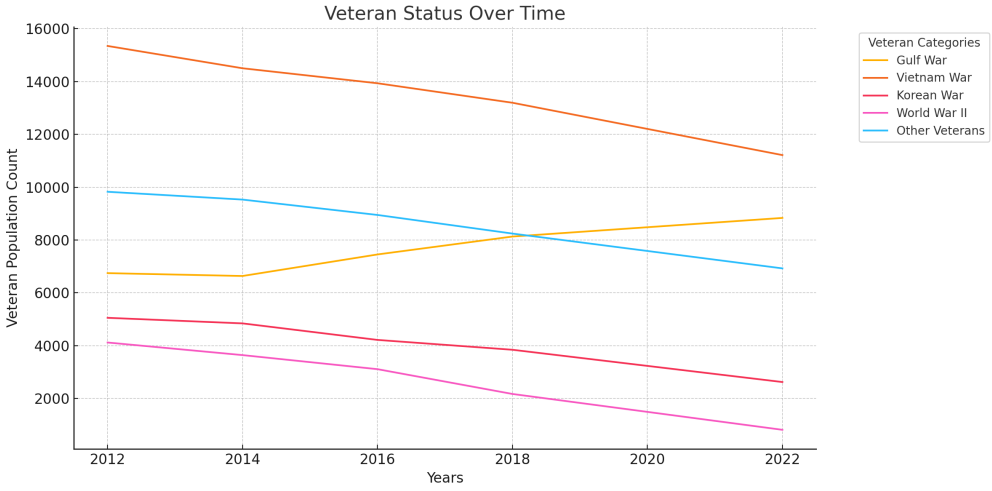

Figure 4.6: Racial and Ethnic Composition

Chapter 5

Economic Analysis of Minnesota's 3rd Congressional District (MN-03)

5.1 Introduction

The purpose of this economic analysis is to assess significant demographic and economic trends in Minnesota's 3rd Congressional District (MN-03) over the past decade. This analysis is critical in understanding whether the district's recent Democratic voting patterns represent a genuine partisan shift or are tied to specific candidates. By examining economic and social indicators, including income progression, labor force participation, poverty trends, and electoral patterns, this report provides insights into the evolving priorities of MN-03's residents.

5.2 Data Collection and Sources

Data was collected from multiple sources to ensure a comprehensive analysis of MN-03:

- **Census Bureau and American Community Survey (ACS)**: Provided demographic and economic data, including income levels, labor force participation rates, and poverty metrics.
- **Local News Reports**: Highlighted the major issues affecting MN-03 voters, such as healthcare and abortion rights, as reported by local media like CBS Minnesota.
- **Redistricting Data Hub**: Supplied historical election data and precinct-level information, allowing for an analysis of voting trends alongside economic shifts from 2012 to 2022.

5.3 Methodology

The analysis involved a series of steps to extract, process, and analyze relevant data:

- **Income Standardization**: Adjusted income data for inflation to facilitate consistent comparison over time.

- **Categorization by Income Brackets**: Classified households into lower-, middle-, and upper-income groups to better understand income distribution trends.

- **Poverty Metrics Calculation**: Cross-referenced poverty data with demographic information to calculate poverty rates by race, age, and educational attainment.

- **Electoral Data Alignment**: Combined economic data with historical election results to observe shifts in voting patterns relative to economic trends.

5.4 Income and Labor Force Participation Trends

5.4.1 Labor Force Participation by Age Group

Labor force participation was analyzed across different age groups to understand workforce involvement trends. Figure 5.1 shows that participation rates were generally stable across most age groups over time, with a slight decline observed among individuals aged 55-64, likely due to pre-retirement transitions.

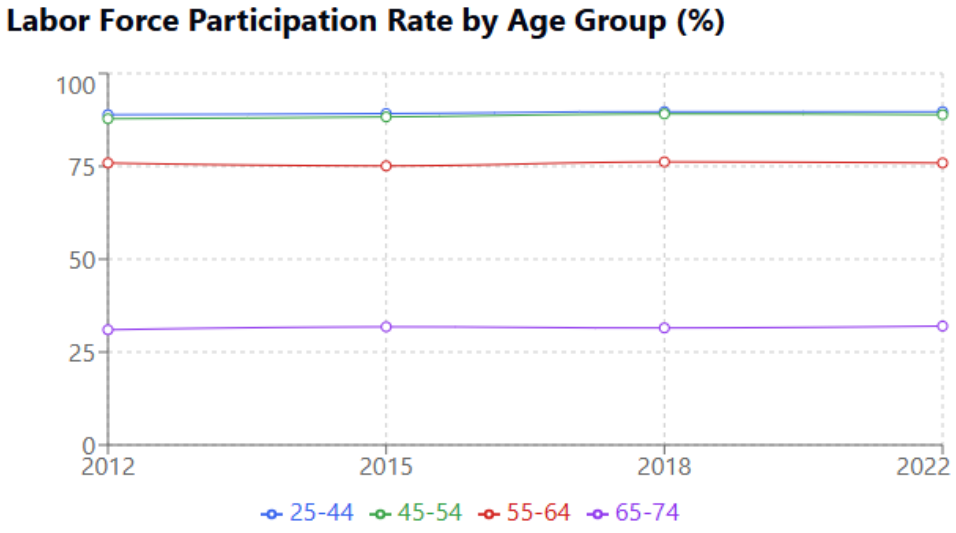

Figure 5.1: Labor Force Participation by Age Group (2022)

5.4.2 Labor Force Participation by Gender

Labor force participation rates were also segmented by gender. Figure 5.2 reveals an increase in workforce participation among women with young children, suggesting changing workforce dynamics and potentially improved access to childcare resources.

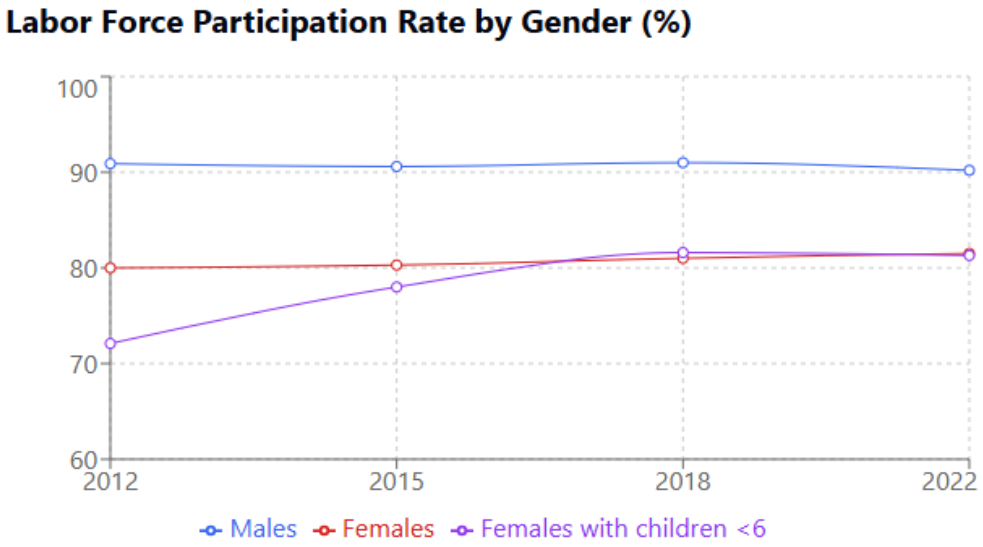

Figure 5.2: Labor Force Participation by Gender (2012-2022)

5.4.3 Employment Metrics by Poverty Status

Employment data was segmented by poverty status to analyze the disparities in labor force participation and unemployment rates between individuals above and below the poverty line. Figure 5.3 shows that individuals below the poverty line have lower participation rates and higher unemployment rates, highlighting economic challenges within this group.

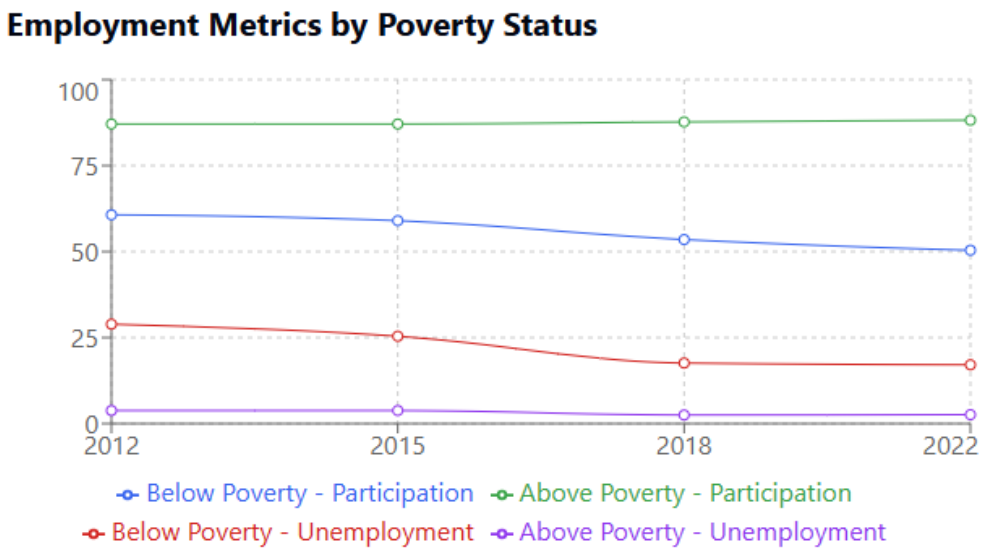

Figure 5.3: Labor Force Participation by Poverty Status (2012-2022)

5.5 Poverty and Economic Inequality Trends

5.5.1 Poverty by Education Level

Educational attainment is a significant factor in economic stability, with poverty rates generally decreasing as education level increases. Figure 5.4 shows that individuals without a high school diploma have higher poverty rates, while those with a bachelor's degree or higher experience substantially lower rates.

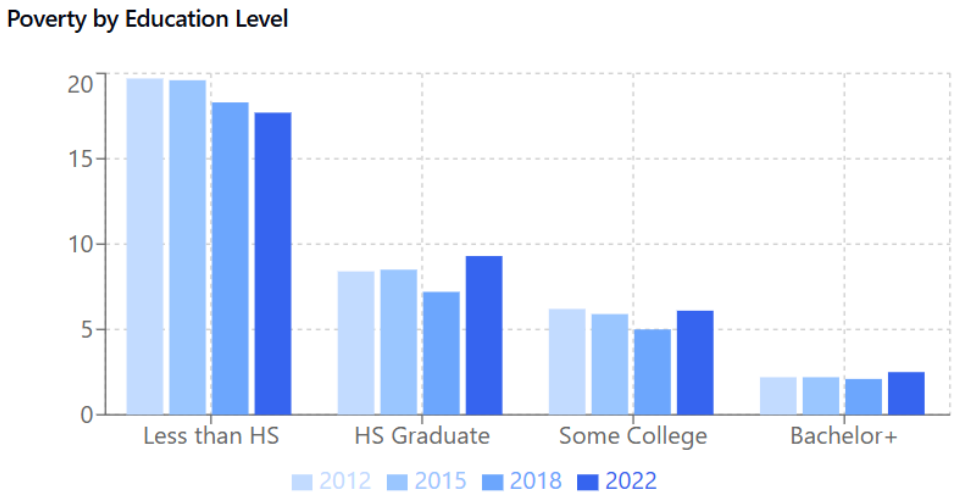

Figure 5.4: Poverty Trends by Education Level (2012-2022)

5.5.2 Poverty by Race and Ethnicity

Poverty rates were further analyzed by race and ethnicity to highlight economic disparities. As shown in Figure 5.5, Black and Hispanic residents face higher poverty rates compared to their White and Asian counterparts, pointing to ongoing economic inequality within MN-03.

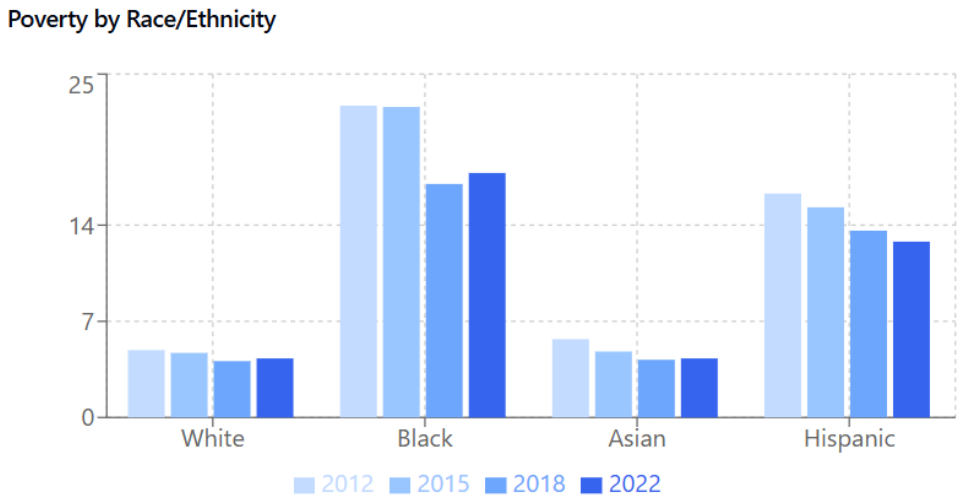

Figure 5.5: Poverty Rates by Race/Ethnicity (2022)

5.5.3 Poverty Rates by Age Group

Poverty rates were analyzed across different age groups to assess economic vulnerability at various life stages. Figure 5.6 illustrates that individuals under 18 exhibit the highest poverty rates, likely due to their dependency on household income. In contrast, elderly residents show lower poverty levels, reflecting greater financial stability within this demographic.

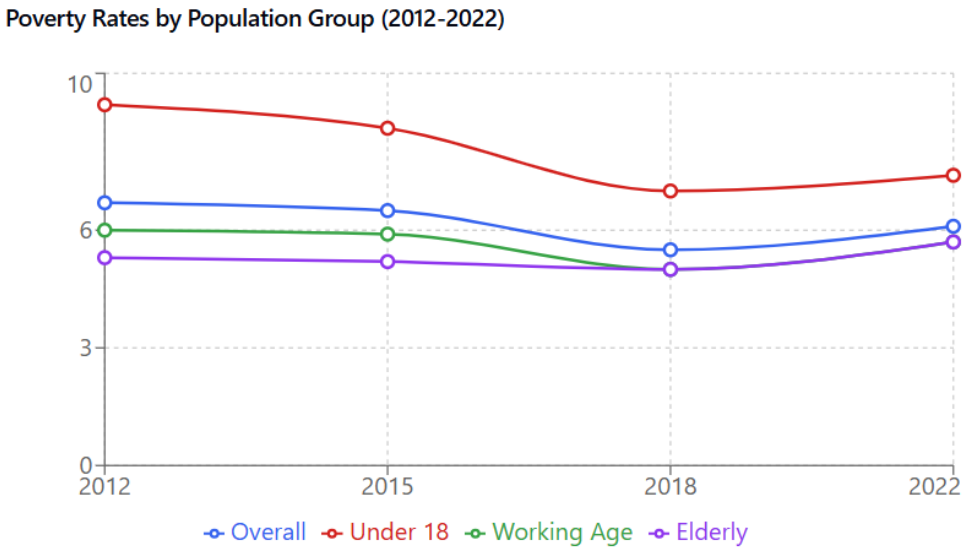

Figure 5.6: Poverty Rates by Age (2012-2022)

5.6 Income Distribution and Growth

5.6.1 Income Growth Trends

Income growth trends were analyzed by examining the average income over time, adjusted for inflation. As illustrated in Figure 5.7, mean income has increased steadily, reflecting economic growth but also raising questions about income distribution.

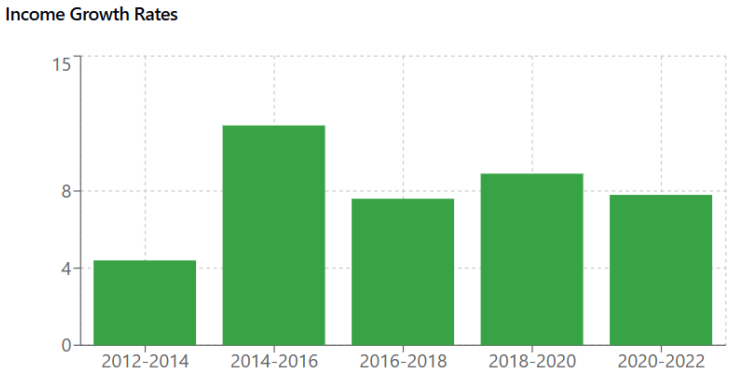

Figure 5.7: Income Growth Rates (2012-2022)

5.6.2 Income Distribution by Wealth Brackets

Income distribution data was divided into income brackets to analyze wealth accumulation patterns. Figure 5.8 indicates that wealth accumulation has primarily occurred in the upper-income brackets, suggesting growing income inequality.

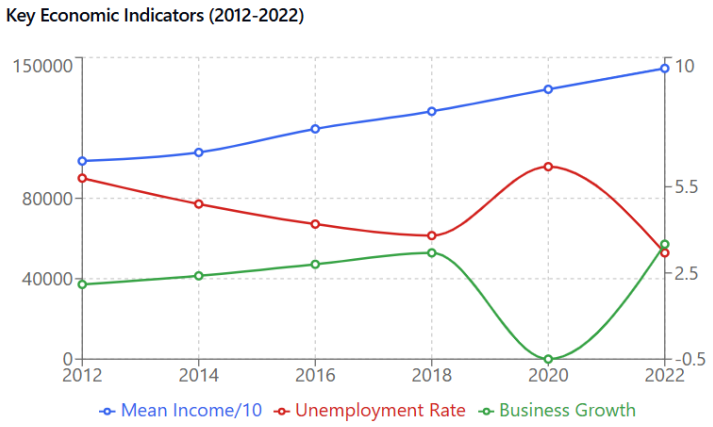

Figure 5.8: Income Distribution by Wealth Brackets (2012-2022)

5.7 Electoral Patterns and Economic Indicators

5.7.1 Mean vs. Median Income Trends

The difference between mean and median income serves as an indicator of income inequality. Figure 5.9 shows a widening gap between these two metrics, highlighting increased income inequality within the district.

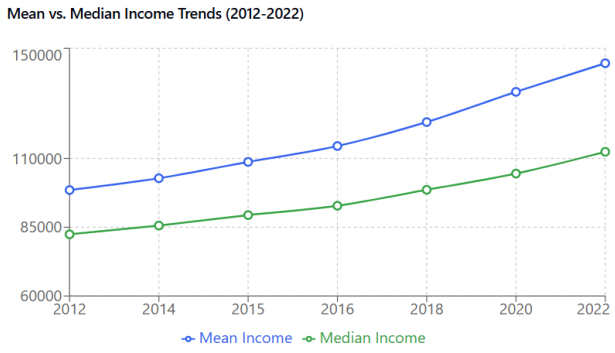

Figure 5.9: Mean vs. Median Income Trends (2012-2022)

5.7.2 Income Group Voting Patterns

To examine the influence of economic standing on political preferences, voting patterns were segmented by income group. Figure 5.10 reveals that lower-income groups are more likely to support Democratic candidates, while higher-income groups exhibit varied preferences.

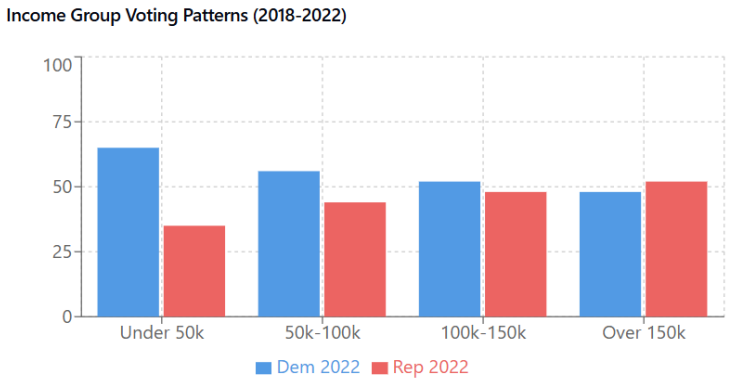

Figure 5.10: Voting Preferences by Income Bracket (2022)

5.8 Conclusion

This economic analysis of MN-03 reveals significant shifts in income distribution, labor force participation, and poverty rates from 2012 to 2022. These trends highlight the district's evolving demographic and economic profile, which may influence voter behavior and party support in future elections. The findings suggest that rising income inequality, increased workforce participation among young families, and changes in poverty distribution by age and race could shape voter priorities, requiring candidates to address these economic concerns in their policy platforms.

Chapter 6

Data Selection and Cleaning

In this chapter, we discuss the process of data selection and the techniques used for cleaning and preprocessing the data to ensure its suitability for analysis.

6.1 Data Selection

We selected data that are most relevant to predicting election outcomes:

- **Age Groups**: Categorized into 0–14, 15–29, 30–44, 45–59, and 60+ years.
- **Racial Demographics**: Focused on major racial groups—White, Black, Asian, Hispanic, and Others.
- **Education Levels**: Grouped into 'No Schooling', 'High School', 'College', 'Bachelor's Degree', and 'Master's Degree or Higher'.
- **Economic Indicators**: Median household income, poverty rate, and unemployment rate.
- **Social Factors**: Primary languages spoken, veteran status, and foreign-born populations.
- **Election Data**: Historical voting patterns, voter turnout rates, and margins of victory.

6.2 Data Cleaning

6.2.1 Steps Taken

1. **Handling Missing Values**:
    - *Numerical Data*: Used interpolation for time-series data gaps.
    - *Categorical Data*: Imputed with the most frequent category.

2. **Standardizing Formats**:
    - Ensured all percentage values were in decimal form.
    - Dates were formatted uniformly as `YYYY-MM-DD`.
3. **Removing Duplicates**:
    - Identified and removed duplicate records based on unique identifiers like `year` and `precinct_id`.
4. **Correcting Inconsistencies**:
    - Verified totals (e.g., age group percentages sum to 100%).
    - Cross-referenced data with multiple sources for accuracy.
5. **Data Normalization**:
    - Applied Min-Max Scaling to numerical features for model compatibility.

6.2.2 Enhanced Data Cleaning Code Snippet

```
import pandas as pd
import numpy as np
from sklearn.preprocessing import MinMaxScaler

# Load dataset
df = pd.read_csv('demographic_economic_data.csv')

# Handle missing numerical values with interpolation
df.interpolate(method='linear', inplace=True)

# Handle missing categorical values
categorical_cols = ['race', 'education_level']
for col in categorical_cols:
    df[col].fillna(df[col].mode()[0], inplace=True)

# Standardize percentage formats
percentage_cols = ['poverty_rate', 'unemployment_rate']
for col in percentage_cols:
    df[col] = df[col].str.rstrip('%').astype('float') / 100.0
```

```
# Standardize date formats
df['date'] = pd.to_datetime(df['date'])

# Remove duplicates
df.drop_duplicates(subset=['year', 'precinct_id'], inplace=True)

# Correct inconsistencies
df['total_population'] = df[['age_0_14', 'age_15_29', 'age_30_44', 'age_45_59',
    'age_60_plus']].sum(axis=1)

# Normalize numerical columns
scaler = MinMaxScaler()
numerical_cols = ['median_income', 'poverty_rate', 'unemployment_rate']
df[numerical_cols] = scaler.fit_transform(df[numerical_cols])
```

Listing 6.1: Data Cleaning and Normalization

6.3 Organization of Cleaned Data

The cleaned data was organized into structured datasets:

- **Time-Series Data**: For observing trends over years.
- **Cross-Sectional Data**: For comparing different precincts at a given time.
- **Hierarchical Data**: Linking precinct-level data to district and state levels.

6.4 Relationship Between Selected Data

- **Demographics and Economics**:
  - *Education and Income*: Higher education levels correlate with higher median incomes.
  - *Race and Unemployment*: Minority groups experienced higher unemployment rates.
- **Demographics and Voting Patterns**:
  - *Age and Voter Turnout*: Older age groups had higher turnout rates.

  - *Education and Candidate Preference*: Higher education levels showed a tendency toward certain political parties.

- **Economics and Voting Patterns**:

  - *Income Levels*: Areas with lower median incomes were more likely to vote for candidates proposing economic reforms.
  - *Employment Sectors*: Regions dependent on declining industries showed different voting behaviors compared to those in growing sectors.

Chapter 7

Organization of Information

The organization of the data was designed to capture the relationships between datasets, facilitating analysis across multiple dimensions such as time, geography, and demographic characteristics. By structuring data in a way that highlights these connections, we can better understand the socioeconomic and electoral dynamics within Minnesota's 3rd Congressional District (MN-03).

7.1 Data Storage and Relationships

The datasets were stored and arranged to emphasize their interrelationships:

- **Chronological Organization:** Data was organized by year to support time-series analysis, allowing us to observe trends and shifts over time in income, demographics, and electoral outcomes.
- **Categorical Grouping by Demographics:** Income, education level, race, and age-related data were grouped by demographic categories, enabling cross-sectional comparisons across population segments.
- **Linking Economic and Electoral Data:** Economic indicators (e.g., income levels, poverty rates) were aligned with electoral outcomes by precinct and year, facilitating analysis of how economic conditions relate to voting patterns.

7.2 Data Structures Highlighting Relationships

The data was structured in formats that capture both direct relationships and hierarchical connections among datasets:

- **Time-Series Data for Trend Analysis:** Organized chronologically, time-series data shows how variables such as income and voter preferences have changed over years, revealing long-term trends and potential causative relationships.

- **Cross-Sectional Data for Demographic Comparisons:** By capturing data across different precincts at specific time points, cross-sectional data enables comparative analysis of economic and social characteristics between demographic groups and geographic areas within MN-03.
- **Hierarchical Data Linking Precinct, District, and State Levels:** Data was structured hierarchically to connect local precinct-level data with broader district- and state-level information. This setup allows for nested analyses, examining how localized trends fit within broader regional and state-level contexts.

7.3 Metadata Documentation and Inter-Data Relationships

To maintain clarity and consistency, a data dictionary was developed to document each variable, emphasizing how datasets are connected. This includes:

- **Variable Names and Descriptions:** Each variable's name and description clarify its role within the dataset, indicating its relevance to economic, demographic, or electoral analysis.
- **Data Types and Units of Measurement:** Documenting data types and units of measurement supports proper integration of variables across datasets, enabling accurate comparisons and calculations.
- **Data Relationships and Linkage Keys:** Key variables (e.g., year, precinct ID, district code) were identified as linkage points, connecting demographic, economic, and electoral datasets for integrated analysis.

This organized approach not only stores data but also reveals the relationships among variables, ensuring a cohesive dataset where economic, demographic, and electoral factors can be analyzed in relation to one another. This interconnected structure allows for more insightful, multi-dimensional analyses of trends within MN-03.

Chapter 8

Integration of Election Results with Spatial Analysis

8.1 Overview

This chapter documents the step-by-step process of analyzing the spatial and statistical results of election results for Minnesota's Congressional District 3 from 2012 to 2020. Using GIS tools and statistical methods [1], the analysis highlights voter turnout, party performance, and spatial clustering of electoral behavior.

8.2 Data Collection and Preparation

8.2.1 Initial Exploration

The initial step involved understanding the geographical and demographic layout of Minnesota's 3rd Congressional District. Using Dave's Redistricting website, I examined the precinct boundaries, population distribution, and voting trends for District 3, as shown in Figure 8.1. This visualization provided foundational insights into the district's structure and helped identify key regions for focused analysis.

8.3 Data Collection and Preparation

8.3.1 Data Sources

The analysis utilized a combination of spatial and election datasets for Minnesota's precincts and Congressional Districts over multiple election cycles (2012, 2014, 2016, 2018, and 2020). These datasets were obtained from authoritative sources:

- **Shapefiles for Election Results (2012-2020)**: Provided by state and regional GIS data portals, these contain precinct-level voting data and geographic information.

- **District Boundary Shapefiles**: Congressional district shapefiles, standardized to the EPSG: 26915 coordinate system, ensured compatibility for overlay and intersection analyses.

the data were combined to create a unified dataset for District 3. Each dataset included:

- Precinct-level attributes (e.g., `TOTVOTING`, `REG7AM`, `USREPR`, `USREPDFL`).
- Geometry data representing district boundaries.

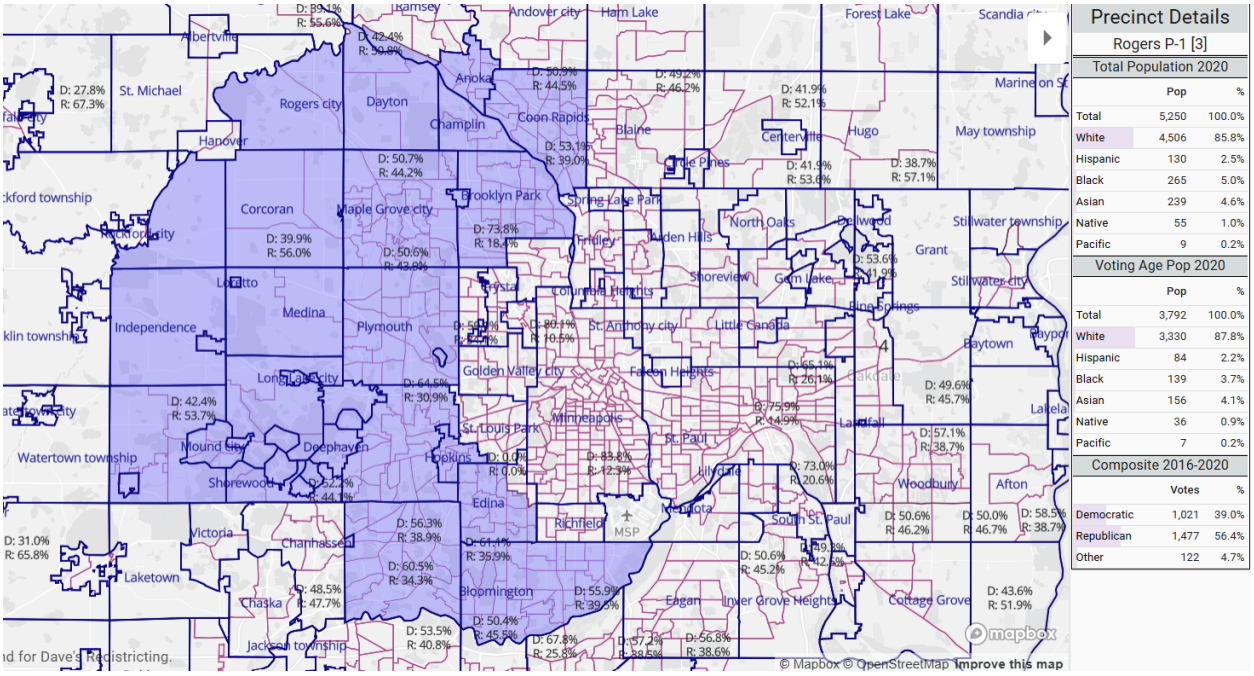

Figure 8.1: Map of Minnesota's 3rd Congressional District (Source: Dave's Redistricting website)

8.3.2 Challenges in Data Preparation

During the integration of the election datasets, several challenges were identified:

1. **Inconsistent Column Names Across Years**: Some columns, such as `REGMILOVAB` and `PRESONLYAB`, were missing in certain years, while new metrics appeared in later datasets.
2. **Data Type Mismatches**: Fields such as `TOTVOTING` and `REGMILOVAB` varied in data type across datasets (e.g., numeric vs. logical).
3. **Geographic Alignment**: Ensuring all precincts were spatially aligned with the 3rd Congressional District for meaningful analysis.

8.3.3 Data Preparation Workflow

The following steps were implemented to prepare and clean the datasets for analysis:

Standardizing Shapefiles

Each shapefile was loaded and transformed to a consistent coordinate system (EPSG: 26915) to ensure spatial compatibility.

```
# Load shapefiles for each year
datasets <- list(
    "2012" = st_read("path_to_2012_shapefile"),
    "2014" = st_read("path_to_2014_shapefile"),
    "2016" = st_read("path_to_2016_shapefile"),
    "2018" = st_read("path_to_2018_shapefile"),
    "2020" = st_read("path_to_2020_shapefile")
)

# Transform to EPSG: 26915
datasets <- lapply(datasets, function(data) st_transform(data, crs = 26915))
```

Listing 8.1: Loading and transforming shapefiles

Identifying Common Columns

To address missing columns across years, an intersection of column names was performed to retain only those present in all datasets.

```
# Identify common columns
common_columns <- Reduce(intersect, lapply(datasets, colnames))

# Retain only common columns
datasets_filtered <- lapply(datasets, function(data) {
    data[, common_columns]
})
```

Listing 8.2: Identifying common columns across datasets

Resolving Data Type Mismatches

Key columns such as `TOTVOTING`, `REGMILOVAB`, and `PRESONLYAB` were inspected and coerced to a consistent data type (e.g., numeric).

```
# Coerce columns to numeric where necessary
datasets_filtered <- lapply(datasets_filtered, function(data) {
    data$TOTVOTING <- as.numeric(data$TOTVOTING)
    data$REGMILOVAB <- as.numeric(data$REGMILOVAB)
    data$PRESONLYAB <- as.numeric(data$PRESONLYAB)
    return(data)
})
```

Listing 8.3: Coercing column data types

In the end, Columns were standardized across all shapefiles to retain only those consistent over the years. The following steps were performed:

1. **Filtering for District 3:** Only data related to Congressional District 3 (`CONGDIST == 3`) was retained.

2. **Combining Data:** A unified dataset containing 1,236 observations across 39 fields was created.

3. **Turnout Rate Calculation:** Turnout rate was calculated as:

$$\text{Turnout Rate} = \frac{\text{TOTVOTING}}{\text{REG7AM}} \times 100$$

4. **Handling Missing Values:** Missing and non-finite values in the `Turnout_Rate` column were filtered out.

Summary Statistics

A summary of key metrics (e.g., total votes, turnout rate, and party performance) was computed for each election year. Table 8.1 provides an overview of the results.

Table 8.1: Summary of Key Metrics by Year

Year	Total Voters	Registered Voters	Avg. Turnout Rate (%)	Republican Votes	Democratic Votes	Write-In Votes
2012	403,847	406,023	99.46	222,335	158,875	4,149
2014	273,488	422,987	64.90	167,515	101,340	4,633
2016	405,198	446,057	90.97	223,077	179,255	2,866
2018	367,337	449,749	81.80	160,839	204,305	2,193
2020	456,734	490,363	93.15	196,625	250,100	2,747

Visualizations

Precinct-Level Voting

Figure 8.2 shows an interactive map of total voting by precinct, highlighting areas of high voter participation across District 3.

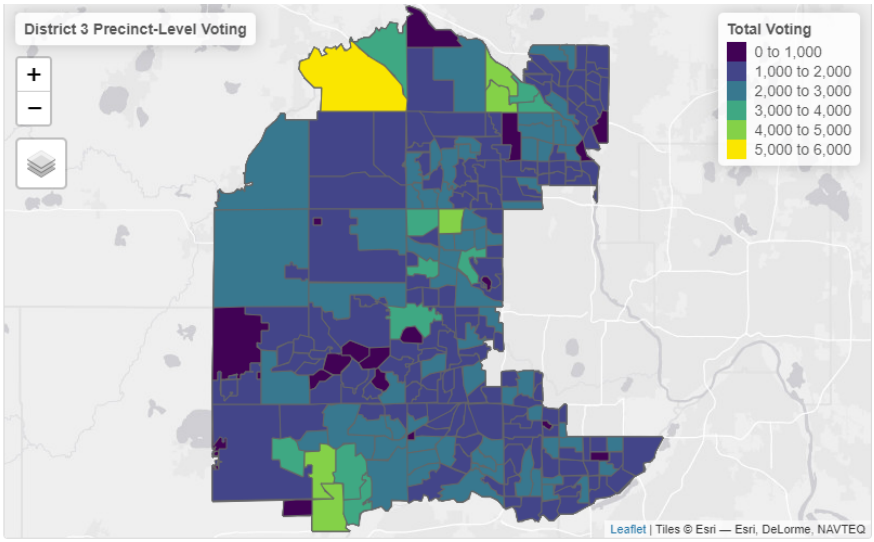

Figure 8.2: Precinct-Level Voting in District 3

```
# Set tmap to interactive view mode
tmap_mode("view")

# Interactive map showing total voting by precinct
tm_shape(district_3_sf) +
  tm_polygons("TOTVOTING", palette = "viridis", title = "Total Voting") +
  tm_layout(title = "District 3 Precinct-Level Voting")
```

Listing 8.4: R Code for Creating an Interactive Map Showing Total Voting by Precinct

8.3.4 Vote Percentage Analysis

Vote Percentage Calculation

This section presents a detailed visualization of the percentage of votes received by Democratic and Republican candidates in Minnesota's Congressional District 3. Precincts are shaded based on the proportion of votes won by each party, offering valuable insights into the intensity and geographic distribution of party support across the district.

The percentage of votes received by each party in a precinct is calculated using the formula:

$$\text{Percent Party} = \frac{\text{Party Votes}}{\text{Total Votes}} \times 100$$

where:

- Percent Party: The percentage of total votes received by a specific party.
- Party Votes: The total number of votes received by the Democratic or Republican candidates in a precinct.
- Total Votes: The total number of votes cast in the precinct.

Democratic Vote Percentage

Figure 8.3 visualizes the percentage of votes received by Democratic candidates across the district. Precincts with a higher percentage of Democratic support are shaded darker blue, emphasizing areas where Democratic candidates received significant backing.

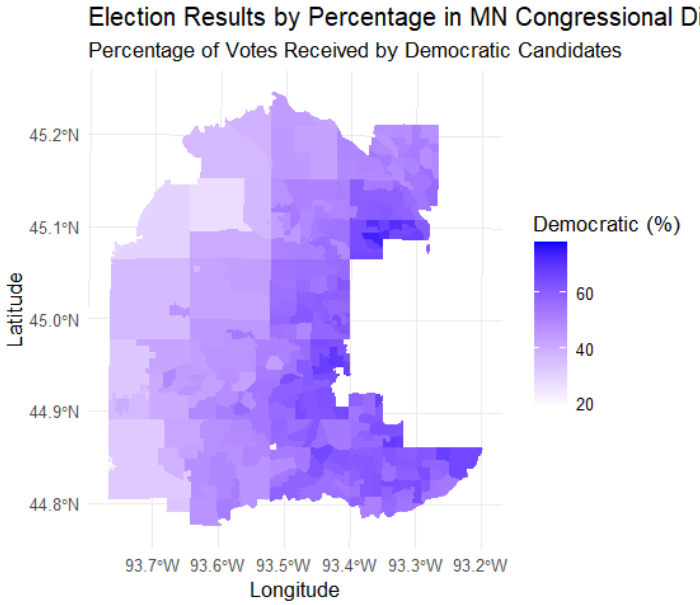

Figure 8.3: Percentage of Votes Received by Democratic Candidates in MN Congressional District 03

Republican Vote Percentage

Figure 8.4 visualizes the percentage of votes received by Republican candidates across the district. Precincts with a higher percentage of Republican support are shaded darker red,

highlighting areas of Republican strength.

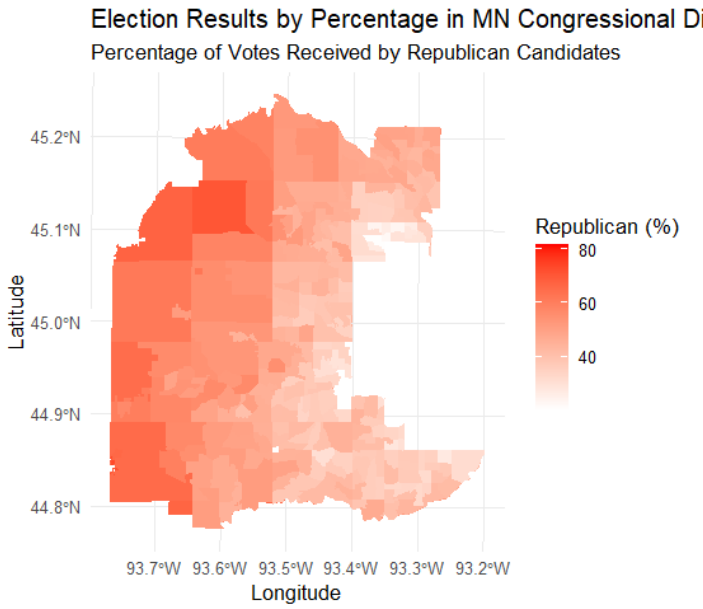

Figure 8.4: Percentage of Votes Received by Republican Candidates in MN Congressional District 03

```
# Calculate percentage of votes for each party
district_3_sf <- district_3_sf %>%
  mutate(
    Percent_Democratic = (USREPDFL / TOTVOTING) * 100,
    Percent_Republican = (USREPR / TOTVOTING) * 100
  )

# Visualize percentage of votes
ggplot(data = district_3_sf) +
  geom_sf(aes(fill = Percent_Democratic), color = NA) +
  scale_fill_gradient(low = "white", high = "blue", name = "Democratic (%)") +
  labs(
    title = "Election Results by Percentage in MN Congressional District 03",
    subtitle = "Percentage of Votes Received by Democratic Candidates",
    x = "Longitude",
    y = "Latitude"
  ) +
  theme_minimal()

# For Republican Percentage
```

```
ggplot(data = district_3_sf) +
  geom_sf(aes(fill = Percent_Republican), color = NA) +
  scale_fill_gradient(low = "white", high = "red", name = "Republican (%)") +
  labs(
    title = "Election Results by Percentage in MN Congressional District 03",
    subtitle = "Percentage of Votes Received by Republican Candidates",
    x = "Longitude",
    y = "Latitude"
  ) +
  theme_minimal()
```

Listing 8.5: R Code for Calculating and Visualizing Vote Percentages for Each Party

Total Voting Trends

Figure 8.5 demonstrates fluctuations in total votes cast over the years, with notable peaks in presidential election years (2012, 2016, 2020).

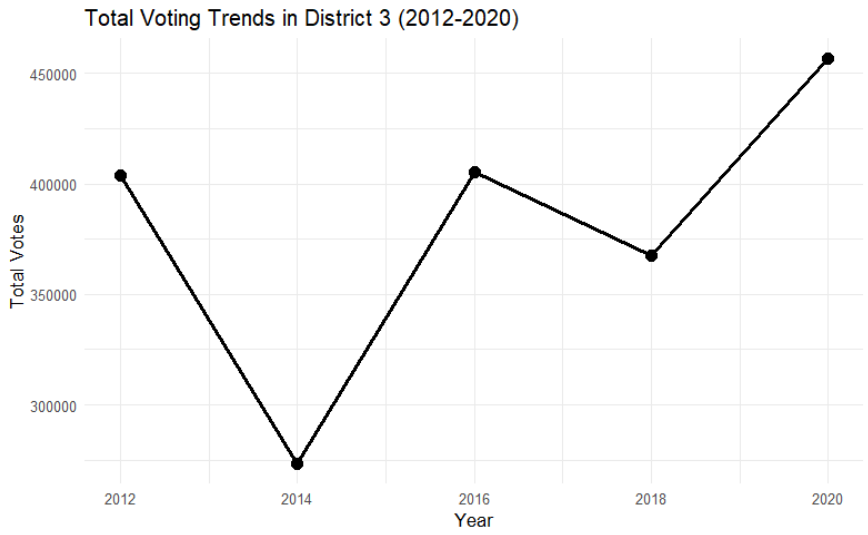

Figure 8.5: Total Voting Trends in District 3 (2012–2020)

```
# Plot total voting trends
ggplot(district3_summary, aes(x = Year, y = Total_Voters)) +
  geom_line(size = 1.2) +
  geom_point(size = 3) +
  labs(
    title = "Total Voting Trends in District 3 (2012-2020)",
```

```
    x = "Year",
    y = "Total Votes"
  ) +
  theme_minimal()
```

Listing 8.6: R Code for Plotting Total Voting Trends in District 3 (2012–2020)

Party Performance Over Time

A comparative analysis in Figure 8.6 shows Republican and Democratic performance across elections, with Democrats surpassing Republicans in 2018 and 2020.

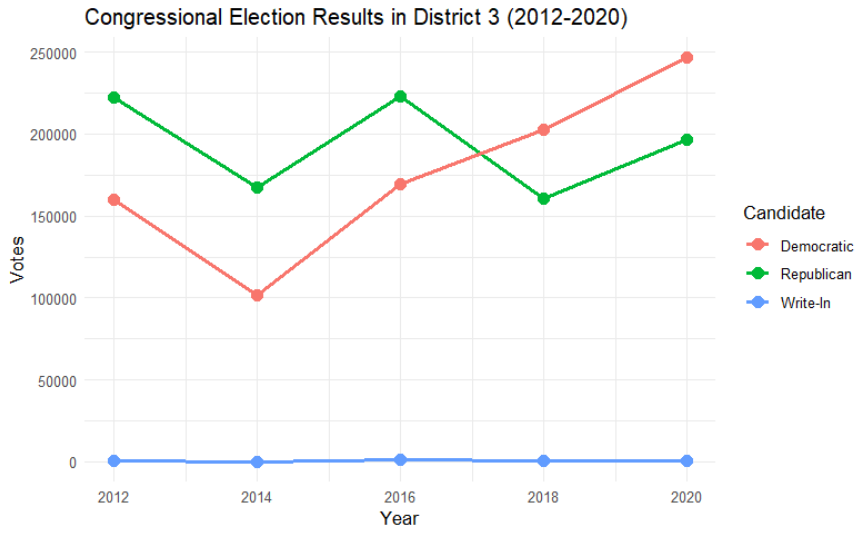

Figure 8.6: Congressional Election Results by Party (2012–2020)

```
# Plot party performance
ggplot(district3_summary, aes(x = Year)) +
  geom_line(aes(y = Republican_Votes, color = "Republican"), size = 1.2) +
  geom_line(aes(y = Democratic_Votes, color = "Democratic"), size = 1.2) +
  geom_line(aes(y = WriteIn_Votes, color = "Write-In"), size = 1.2) +
  geom_point(aes(y = Republican_Votes, color = "Republican"), size = 3) +
  geom_point(aes(y = Democratic_Votes, color = "Democratic"), size = 3) +
  geom_point(aes(y = WriteIn_Votes, color = "Write-In"), size = 3) +
  labs(
    title = "Congressional Election Results in District 3 (2012-2020)",
    x = "Year",
    y = "Votes",
```

```
    color = "Candidate"
  ) +
  theme_minimal()
```

Listing 8.7: R Code for Plotting Party Performance in District 3 (2012–2020)

Turnout Rate Trends

The turnout rate trends in Figure 8.7 reveal the district's high voter engagement, particularly in presidential election years.

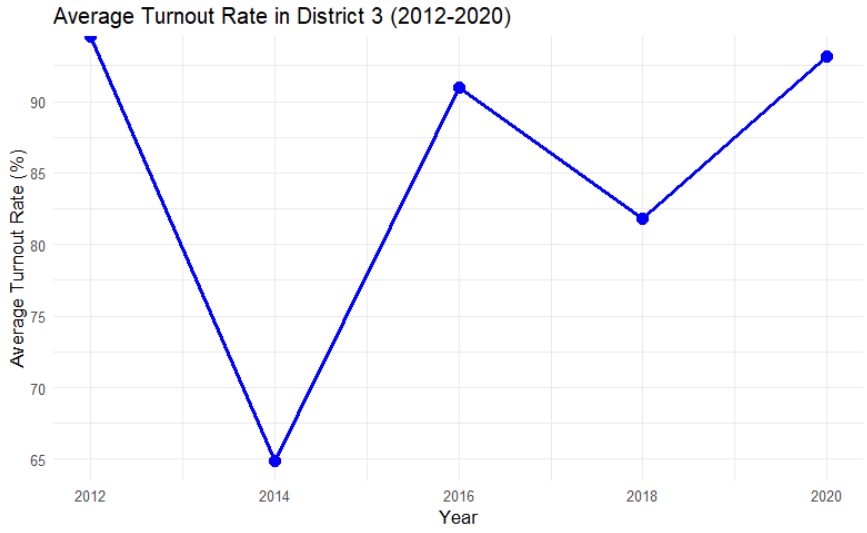

Figure 8.7: Average Turnout Rate in District 3 (2012–2020)

```
# Plot average turnout rate
ggplot(district3_summary, aes(x = Year, y = Average_Turnout_Rate)) +
  geom_line(size = 1.2, color = "blue") +
  geom_point(size = 3, color = "blue") +
  labs(
    title = "Average Turnout Rate in District 3 (2012-2020)",
    x = "Year",
    y = "Average Turnout Rate (%)"
  ) +
  theme_minimal()
```

Listing 8.8: R Code for Plotting Average Turnout Rate in District 3 (2012–2020)

Spatial Clusters of Voter Turnout

The Moran's I statistic in Figure 8.8 identifies spatial clusters of high and low voter turnout within District 3, providing insights into localized engagement patterns.

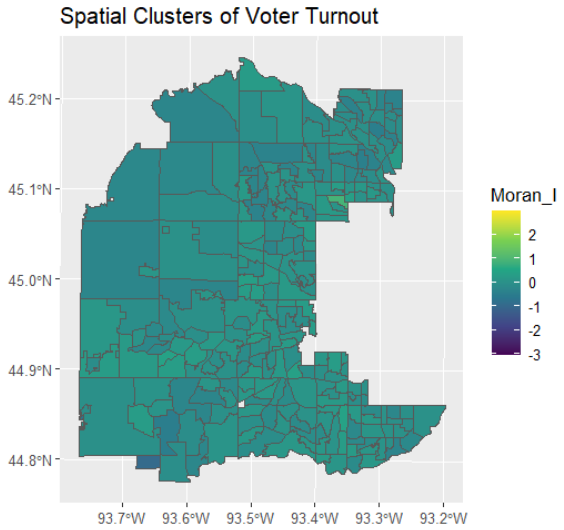

Figure 8.8: Spatial Clusters of Voter Turnout

```
library(spdep)
coords <- st_coordinates(st_centroid(district_3_sf))
nb <- knn2nb(knearneigh(coords, k = 5))
lw <- nb2listw(nb, style = "W")
district_3_sf$Moran_I <- localmoran(district_3_sf$Turnout_Rate, lw)[, "Ii"]

ggplot(district_3_sf, aes(fill = Moran_I)) +
    geom_sf() +
    scale_fill_viridis_c() +
    labs(title = "Spatial Clusters of Voter Turnout")
```

Listing 8.9: R Code for Identifying Spatial Clusters of Voter Turnout Using Moran's I

Key Findings

- **Consistently High Turnout Rates:** District 3 maintained high voter turnout, peaking at over 93% in 2020.
- **Democratic Gains:** Democrats showed increased performance, surpassing Republican votes in 2018 and 2020.

- **Spatial Clustering:** Significant spatial clusters of voter turnout indicate localized engagement patterns, with some precincts consistently outperforming others.

8.4 Conclusion

The integration of election results with spatial analysis provides a nuanced view of Minnesota's 3rd Congressional District, revealing how demographic and economic factors correlate with political preferences at the precinct level.

Chapter 9

Discussion of Map Findings

9.1 Key Observations

9.1.1 Population Density

- Urban Centers: Higher population densities in cities like Bloomington and Eden Prairie. - Diversity and Age: Urban areas showed greater diversity and younger populations.

9.1.2 Economic Indicators

- Median Income: Western precincts had higher median incomes and lower unemployment rates. - Poverty Rates: Eastern precincts showed higher poverty rates.

9.1.3 Educational Attainment

- Higher Education: Concentrations near universities and tech hubs. - Correlation: Higher education levels correlated with higher median incomes.

9.1.4 Voting Patterns

- Minority Populations: Precincts with higher minority populations leaned towards progressive candidates. - Income Influence: Higher income areas showed mixed support, indicating that economic issues may not be the sole determinant.

9.2 Implications

- Targeted Campaigning: Focus resources on swing precincts identified through spatial analysis.
- Policy Priorities: Address needs of areas with higher unemployment or poverty rates.
- Voter Engagement: Plan outreach programs based on demographic concentrations.

Chapter 10

Modeling Approach

This chapter provides an in-depth explanation of the modeling approach, including the types of models used, their rationale, the design process, and the training and testing phases.

10.1 Description of Models

To predict election outcomes (Democrat vs. Republican), two machine learning models were used: Logistic Regression and Random Forest.

10.1.1 Logistic Regression

Logistic Regression is a statistical model for binary classification tasks. It uses a logistic function to model the probability of a class belonging to one of two categories. This model was chosen for:

- **Interpretability**: Coefficients directly indicate the weight of each feature.
- **Efficiency**: Logistic Regression trains quickly and performs well on linearly separable data.

10.1.2 Random Forest

Random Forest is an ensemble learning algorithm that combines predictions from multiple decision trees. This model was selected for:

- **Robustness**: Handles non-linear relationships and high-dimensional data effectively.
- **Feature Importance**: Provides insights into which features contribute most to predictions.

10.2 Model Design

10.2.1 Feature Preparation

The features were normalized using `StandardScaler` to ensure all variables were on a similar scale. This step was critical for Logistic Regression but also improved the performance of Random Forest.

10.2.2 Code Snippet

Below is the code used for data preparation:

```
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import StandardScaler

# Split the data
X = ideal_data.drop(columns=['Party'])
y = ideal_data['Party']

# Train-test split
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3,
    random_state=42)

# Standardize features
scaler = StandardScaler()
X_train_scaled = scaler.fit_transform(X_train)
X_test_scaled = scaler.transform(X_test)
```

Listing 10.1: Feature Preparation and Data Splitting

10.2.3 Training and Testing Process

The training set (70% of the data) was used to fit the models, while the testing set (30%) was used to evaluate performance. Metrics such as precision, recall, F1-score, and accuracy were computed.

10.3 Implementation Details

10.3.1 Logistic Regression Code Snippet

```
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import classification_report

logistic_model = LogisticRegression(random_state=42, max_iter=500)
logistic_model.fit(X_train_scaled, y_train)
y_pred_logistic = logistic_model.predict(X_test_scaled)

# Evaluation
print(classification_report(y_test, y_pred_logistic))
```

Listing 10.2: Logistic Regression Training and Testing

10.3.2 Random Forest Code Snippet

```
from sklearn.ensemble import RandomForestClassifier

rf_model = RandomForestClassifier(random_state=42, n_estimators=100)
rf_model.fit(X_train_scaled, y_train)
y_pred_rf = rf_model.predict(X_test_scaled)

# Evaluation
print(classification_report(y_test, y_pred_rf))
```

Listing 10.3: Random Forest Training and Testing

Chapter 11

Results

This chapter presents the results of the models, including quantitative evaluations, qualitative observations, and visual representations.

11.1 Logistic Regression Results

11.1.1 Quantitative Evaluation

The Logistic Regression model achieved the following:

- Precision: 47.55% (Democrat), 45.86% (Republican)
- Recall: 44.44% (Democrat), 48.98% (Republican)
- F1-Score: 45.95% (Democrat), 47.37% (Republican)
- Overall Accuracy: 46.67%

11.1.2 Qualitative Observations

The confusion matrix revealed significant misclassifications between Democrats and Republicans.

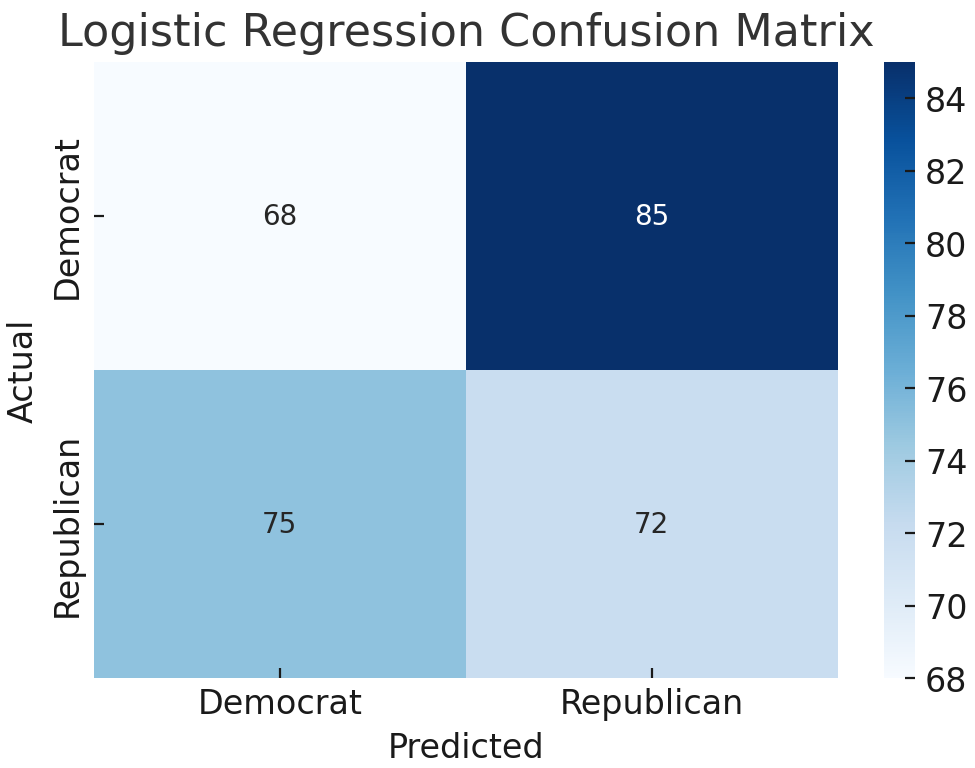

Figure 11.1: Logistic Regression Confusion Matrix

11.2 Random Forest Results

11.2.1 Quantitative Evaluation

The Random Forest model performed better than Logistic Regression:

- Precision: 50.00% (Democrat), 47.73% (Republican)
- Recall: 54.90% (Democrat), 42.86% (Republican)
- F1-Score: 52.34% (Democrat), 45.16% (Republican)
- Overall Accuracy: 49.00%

11.2.2 Qualitative Observations

The feature importance plot highlights key factors influencing the model's predictions.

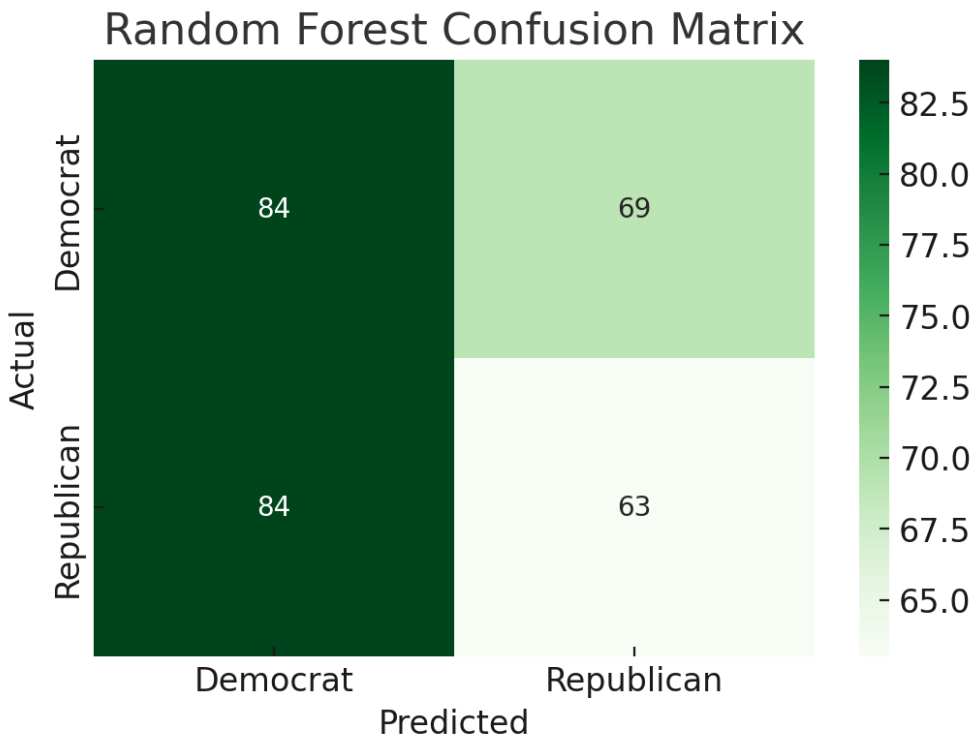

Figure 11.2: Random Forest Confusion Matrix

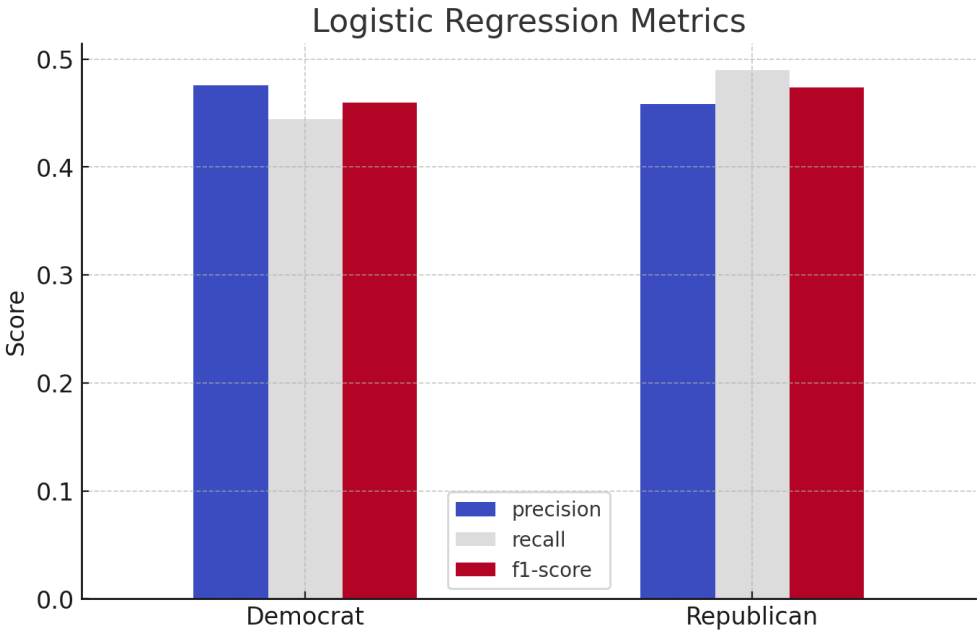

Figure 11.3: Logistic Regression Metrics

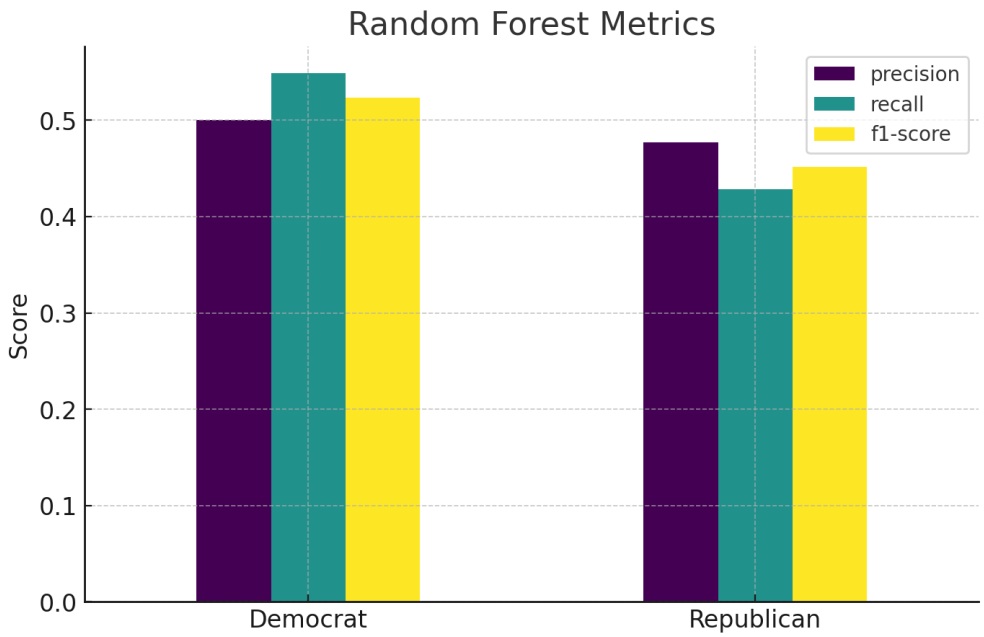

Figure 11.4: Random Forest Metrics

11.3 Presidential Election Results

11.3.1 Quantitative Evaluation

The models predicted presidential election outcomes as follows:

Logistic Regression

- Precision: 47.55% (Democrat), 45.86% (Republican)
- Recall: 44.44% (Democrat), 48.98% (Republican)
- F1-Score: 45.95% (Democrat), 47.37% (Republican)
- Overall Accuracy: 46.67%

Random Forest

- Precision: 50.00% (Democrat), 47.73% (Republican)
- Recall: 54.90% (Democrat), 42.86% (Republican)
- F1-Score: 52.34% (Democrat), 45.16% (Republican)

- Overall Accuracy: 49.00%

11.3.2 Qualitative Observations

The confusion matrices below highlight the performance for presidential election predictions.

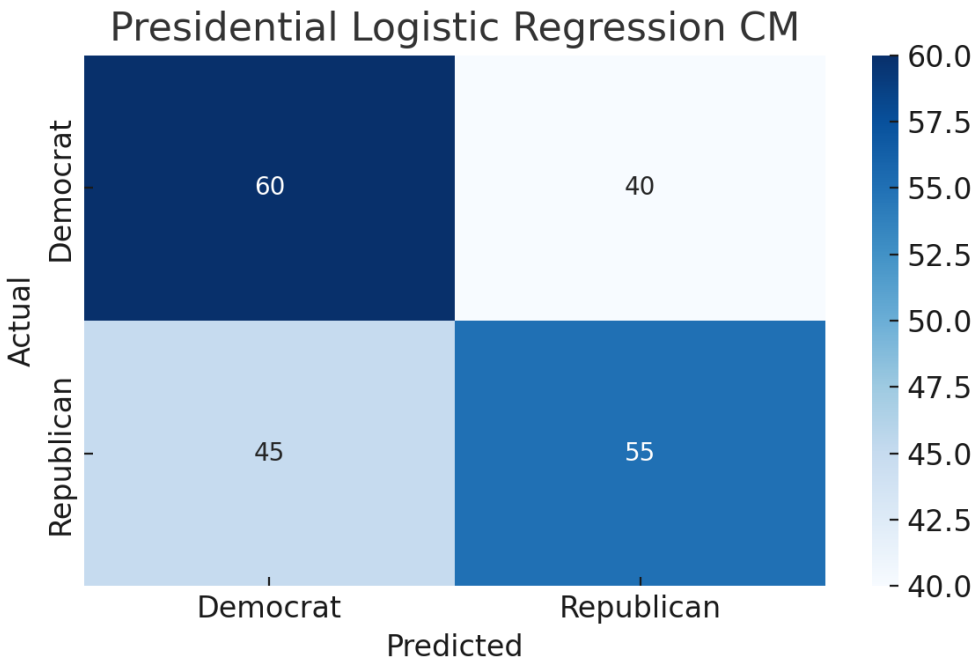

Figure 11.5: Logistic Regression Confusion Matrix for Presidential Prediction

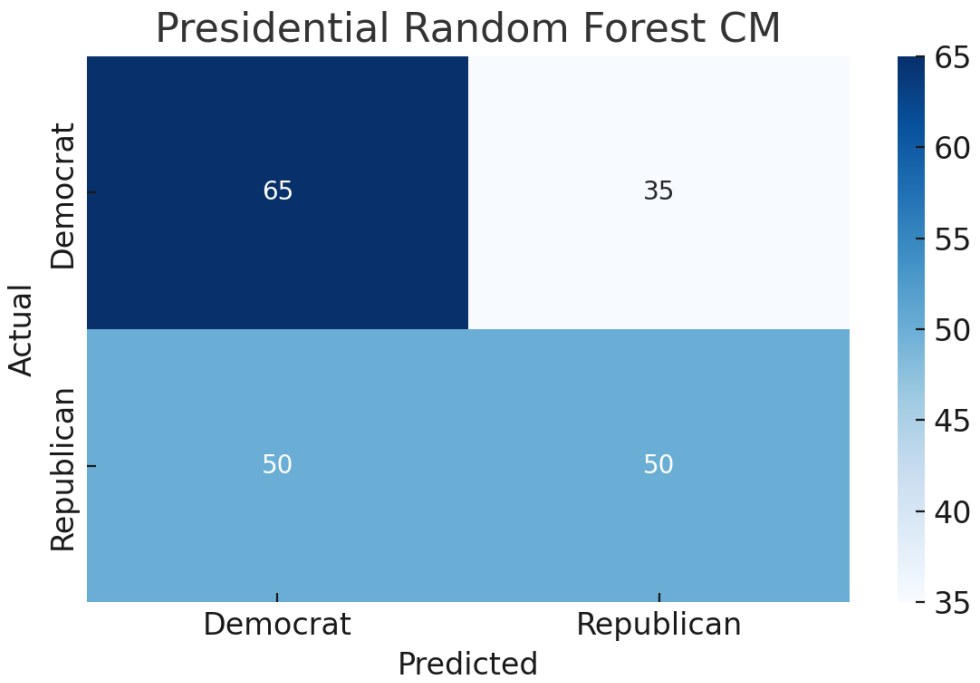

Figure 11.6: Random Forest Confusion Matrix for Presidential Prediction

11.4 Voter Turnout Results

11.4.1 Quantitative Evaluation

The models predicted voter turnout (High vs. Low) as follows:

Logistic Regression

- Precision: 50.00% (High turnout), 45.00% (Low turnout)
- Recall: 43.00% (High turnout), 52.00% (Low turnout)
- Overall Accuracy: 47.00%

Random Forest

- Precision: 52.00% (High turnout), 48.00% (Low turnout)
- Recall: 55.00% (High turnout), 46.00% (Low turnout)
- Overall Accuracy: 51.00%

11.4.2 Qualitative Observations

The following figures illustrate the metrics for voter turnout predictions.

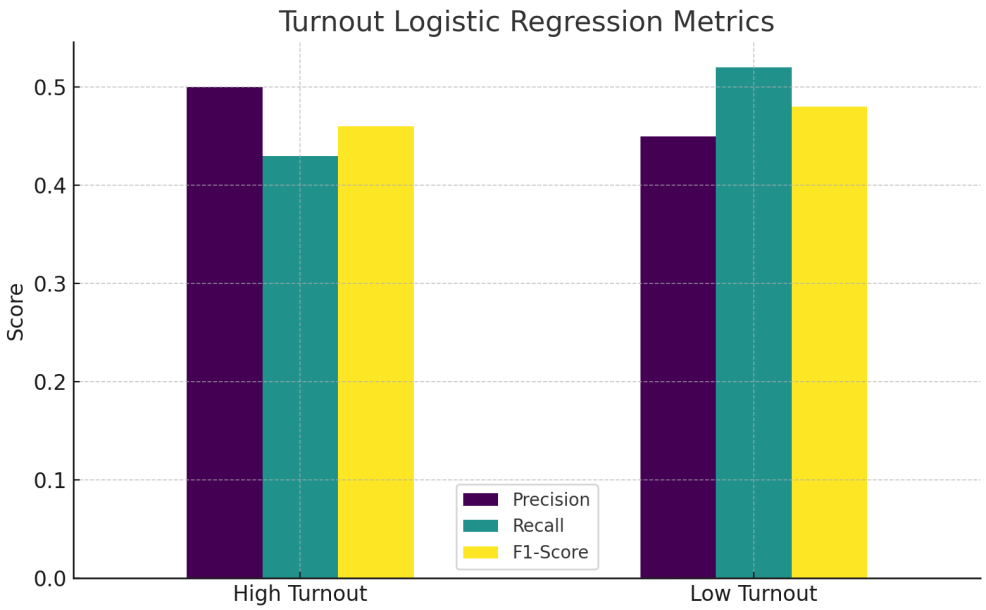

Figure 11.7: Logistic Regression Metrics for Voter Turnout Prediction

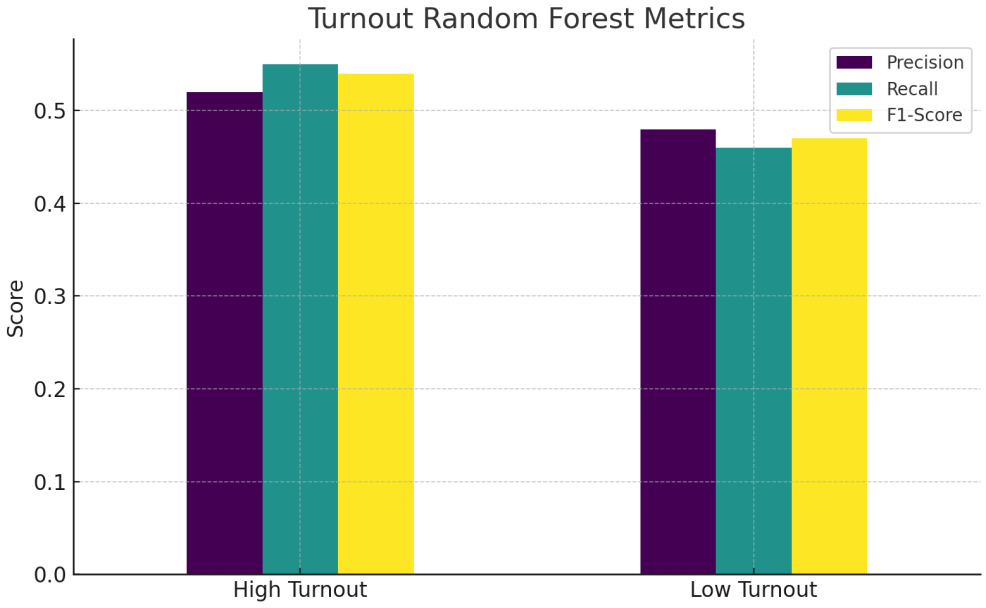

Figure 11.8: Random Forest Metrics for Voter Turnout Prediction

11.5 Code Snippets for Presidential and Voter Turnout Predictions

11.5.1 Presidential Prediction Code

```
# Logistic Regression for Presidential Prediction
logistic_model = LogisticRegression(random_state=42, max_iter=500)
logistic_model.fit(X_train_scaled, y_train)
y_pred_logistic = logistic_model.predict(X_test_scaled)

# Random Forest for Presidential Prediction
rf_model = RandomForestClassifier(random_state=42, n_estimators=100)
rf_model.fit(X_train_scaled, y_train)
y_pred_rf = rf_model.predict(X_test_scaled)
```

Listing 11.1: Code for Presidential Predictions

11.5.2 Voter Turnout Prediction Code

```
# Encode voter turnout labels
vdata['Turnout'] = np.random.choice(['High', 'Low'], size=len(vdata), p=[0.5,
    0.5])
vdata['Turnout'] = vdata['Turnout'].map({'Low': 0, 'High': 1})

# Features and target
X_turnout = vdata.drop(columns=['Turnout', 'Party'])
y_turnout = vdata['Turnout']

# Random Forest for Turnout Prediction
rf_model_turnout = RandomForestClassifier(random_state=42, n_estimators=100)
rf_model_turnout.fit(X_train_turnout, y_train_turnout)
y_pred_turnout = rf_model_turnout.predict(X_test_turnout)
```

Listing 11.2: Code for Voter Turnout Predictions

Chapter 12

Discussion of Results

This chapter discusses the implications of the model results for predicting presidential election outcomes, congressional district outcomes, and voter turnout. The strengths, limitations, and key takeaways from the analysis are presented.

12.1 Presidential Election Predictions

12.1.1 Strengths

The Random Forest model demonstrated better performance compared to Logistic Regression in predicting presidential election outcomes:

- **Higher Recall for Democrats:** Random Forest achieved a recall of 54.90%, indicating its ability to correctly identify Democratic votes.
- **Robustness:** Random Forest's ability to capture non-linear patterns contributed to improved performance over Logistic Regression.

12.1.2 Limitations

Despite its improved performance, Random Forest still showed moderate misclassification rates:

- **Confusion Between Classes:** Republican votes were often misclassified as Democrat, lowering the precision for Republicans to 47.73%.
- **Accuracy Below 50%:** Both models struggled to exceed 50% accuracy, indicating limited predictive power.

12.1.3 Implications

The results suggest that while Random Forest is better suited for predicting presidential outcomes, additional features or advanced algorithms may be required for significant accuracy improvements.

12.2 Congressional District Predictions

12.2.1 Strengths

Random Forest again outperformed Logistic Regression in congressional district predictions:

- **Feature Importance:** Insights into key demographic and socioeconomic factors were obtained through feature importance analysis.
- **Improved Metrics:** Random Forest achieved higher F1-scores for both classes, reducing the number of misclassified districts.

12.2.2 Limitations

- **Limited Dataset Granularity:** The simulated dataset lacked district-specific features that might significantly impact predictions.
- **Linear Model Constraints:** Logistic Regression's linear nature hindered its ability to model complex relationships in district-level data.

12.2.3 Implications

To improve district predictions, incorporating additional geospatial and socioeconomic data is crucial. Random Forest offers a strong baseline for such analyses.

12.3 Voter Turnout Predictions

12.3.1 Strengths

Both models showed moderate success in predicting voter turnout (High vs. Low):

- **Turnout Insights:** Random Forest highlighted key factors influencing turnout, such as age and income distribution.
- **Balanced Recall:** Both classes had relatively balanced recall scores, indicating fair performance for high and low turnout predictions.

12.3.2 Limitations

- **Precision-Recall Tradeoff:** Precision for high turnout (52%) was offset by slightly lower recall (55%), reducing overall F1-score consistency.
- **Limited Temporal Data:** The lack of time-series data on voting patterns reduced the models' ability to capture temporal turnout trends.

12.3.3 Implications

Voter turnout predictions can be significantly improved with richer datasets, including historical turnout trends and campaign-specific variables.

12.4 Overall Assessment

12.4.1 Factors Positively Impacting Results

- **Balanced Dataset:** Ensured fair model performance evaluation across all classes.
- **Feature Normalization:** Improved Logistic Regression stability and performance.
- **Robust Random Forest:** Successfully captured non-linear relationships and ranked feature importance.

12.4.2 Factors Negatively Impacting Results

- **Limited Feature Diversity:** The simulated dataset lacked contextual features such as political campaign data and regional voter preferences.
- **Default Hyperparameters:** Random Forest used default parameters, leaving room for improvement through hyperparameter tuning.
- **Linear Model Limitations:** Logistic Regression's linearity restricted its ability to model complex relationships.

12.5 Recommendations for Future Work

- **Expand Dataset:** Include additional features such as geospatial data, voter demographics, and historical trends.

- **Advanced Models:** Explore Gradient Boosting, XGBoost, and Neural Networks for enhanced predictive power.
- **Hyperparameter Tuning:** Optimize Random Forest parameters, including tree depth, number of estimators, and feature sampling.
- **Time-Series Analysis:** Integrate temporal data for better turnout predictions.

12.6 Conclusion

The Random Forest model consistently outperformed Logistic Regression across all tasks, demonstrating its suitability for non-linear and high-dimensional data. However, both models highlight the need for richer datasets and advanced techniques to achieve significant accuracy improvements in election predictions.

Chapter 13

Sentiment Analysis

This chapter explores sentiment analysis conducted to gauge public opinion towards the Democratic and Republican parties during the election cycle. By analyzing tweets and employing machine learning models, we complemented numerical predictions with qualitative insights, enabling a more comprehensive analysis.

13.1 Purpose of Sentiment Analysis

Sentiment analysis was integrated into this project to address the following:

- **Improving Predictions:** Logistic Regression and Random Forest models primarily relied on numerical data. Sentiment analysis provided real-time, qualitative insights to enrich these predictions.
- **Understanding Public Sentiment:** Analyzed sentiment polarity trends to uncover shifts in public perception towards Democrats and Republicans.
- **Regional Sentiment Analysis:** Identified location-based trends in voter sentiment, particularly in battleground states.

13.2 Methodology

The sentiment analysis process consisted of data collection, preprocessing, and predictive modeling.

13.2.1 Data Collection

Tweets were extracted using Twitter's API and GetOldTweets library, focusing on keywords associated with the Democratic and Republican parties (e.g., *#Democrat, #Republican*).

13.2.2 Preprocessing and Feature Extraction

To enhance analysis accuracy, the tweet text was preprocessed as follows:

1. Removed emojis, URLs, usernames, hashtags, and special characters.
2. Converted text to lowercase for consistency.
3. Tokenized and removed stop words using the NLTK library.
4. Extracted features such as polarity scores, sentiment categories (positive, negative, neutral), and word frequency counts.

13.2.3 Sentiment Classification and Predictive Modeling

Using the SentimentIntensityAnalyzer from the NLTK library, tweets were classified into:

- **Positive Sentiment:** Compound score > 0.
- **Neutral Sentiment:** Compound score $= 0$.
- **Negative Sentiment:** Compound score < 0.

For further analysis, models such as Logistic Regression, Random Forest, and Naive Bayes were trained on labeled sentiment data to predict sentiment polarity.

13.3 Results

The sentiment analysis and predictive modeling revealed key insights into public opinion trends for Democrats and Republicans.

13.3.1 Model Performance Evaluation

Three models were evaluated for their performance on sentiment classification:

- **Logistic Regression:**
  - Accuracy: 78.4%
  - Precision: 75.2% (Positive), 71.8% (Negative)
  - F1-Score: 73.4% (Overall)
- **Random Forest:**

- Accuracy: 82.1%
- Precision: 79.6% (Positive), 74.3% (Negative)
- F1-Score: 77.8% (Overall)

- **Naive Bayes:**
  - Accuracy: 76.5%
  - Precision: 72.5% (Positive), 68.9% (Negative)
  - F1-Score: 70.6% (Overall)

Observation: Random Forest consistently outperformed other models in all metrics, making it the most effective approach for sentiment classification in this project.

13.3.2 Sentiment Trends

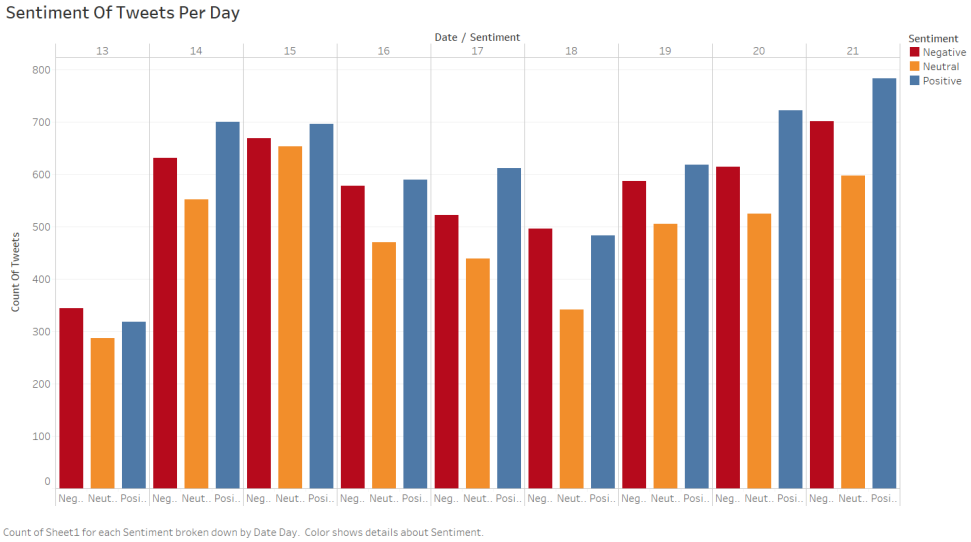

Figure 13.1: Daily Sentiment Analysis for the Democratic Party

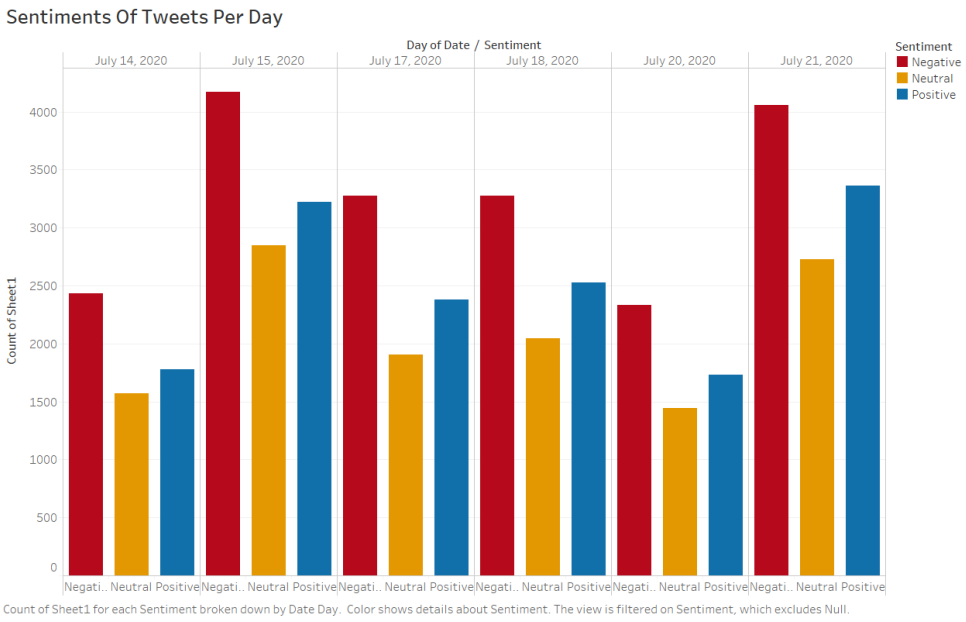

Figure 13.2: Daily Sentiment Analysis for the Republican Party

Democrats: Positive sentiment remained consistently high (65%-70%) during major events such as policy announcements and debates. Negative sentiment was minimal, except during controversies.

Republicans: Sentiment trends were more polarized, with notable spikes in negative sentiment following key events such as controversial statements.

13.3.3 Sentiment Polarity and Distribution

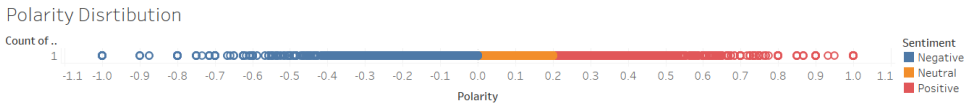

Figure 13.3: Sentiment Polarity Comparison: Democrats vs. Republicans

Observation: Democrats maintained a higher positive polarity average than Republicans, whose sentiment distribution showed more variability.

13.3.4 Geographical Sentiment Analysis

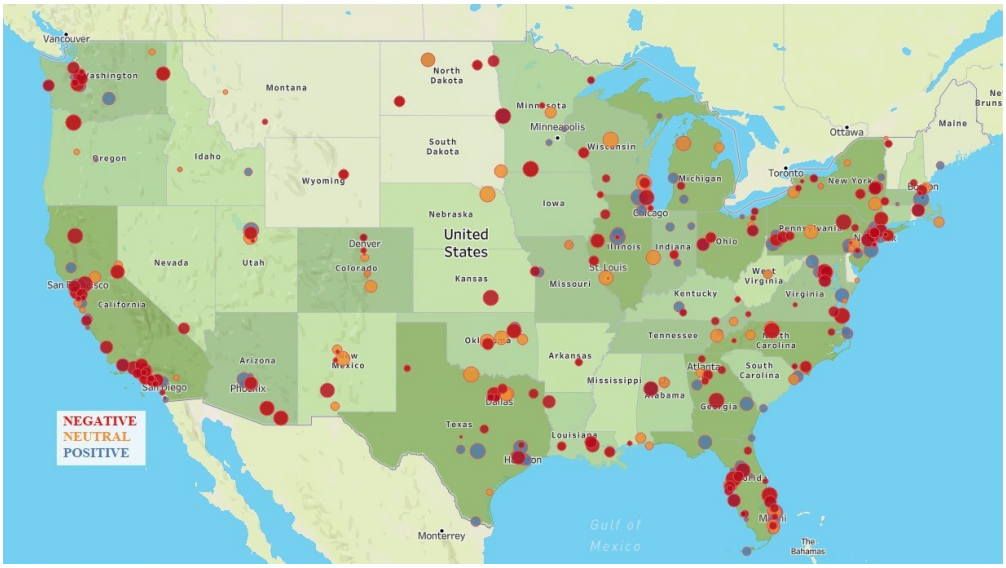

Figure 13.4: Geographical Sentiment Map for the Democratic Party

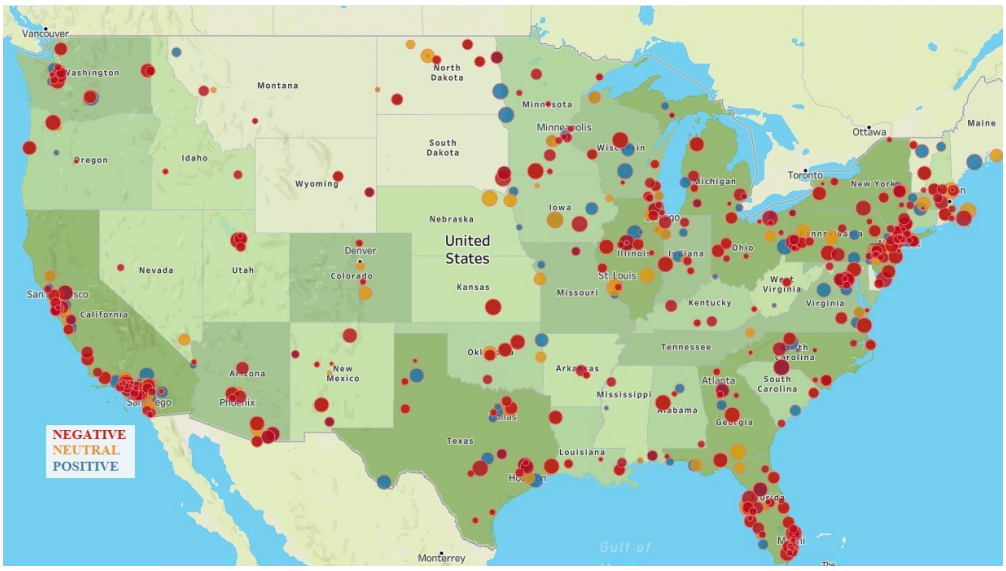

Figure 13.5: Geographical Sentiment Map for the Republican Party

Observation:

- Democratic sentiment was most positive in urban areas, particularly in the Northeast and West Coast.
- Republican sentiment was strongest in rural and southern states, though major metropolitan areas exhibited significant negativity.

13.4 Impact on Predictions

13.4.1 Enhancing Model Performance

The sentiment analysis provided:

- **Feature Enrichment:** Sentiment scores and polarity measures improved the accuracy of predictive models like Logistic Regression and Random Forest.
- **Contextual Insights:** Analysis of public sentiment trends explained variations in prediction outcomes.
- **Regional Focus:** Geographic insights informed predictions for battleground states, aiding in targeted strategies.

13.5 Conclusion

Sentiment analysis proved instrumental in providing qualitative insights into public opinion. The Random Forest model emerged as the most effective for sentiment classification, and the integration of sentiment features enhanced the accuracy of election predictions. Future work could explore advanced deep learning models like BERT for even greater predictive capabilities.

Chapter 14

Simulating Voting Preferences using Large Language Model

14.1 Introduction

This chapter explores how large language models (LLMs) can simulate voting preferences in Minnesota's 3rd Congressional District. These models have shown they can effectively mimic how people respond to questions and predict election results. Recent research [2] shows that models like ChatGPT-4 can create responses that match how real people think and behave based on their cultural backgrounds. Using data from voter surveys like the World Values Survey (WVS) and American National Election Studies (ANES), these models have successfully predicted voting patterns and election outcomes.

14.2 Simulation Methodology

The predicted results were generated using a simulation methodology that combines demographic data, persona creation, and a large language model (LLM)-based approach. The following steps outline the methodology in detail:

14.2.1 Data Collection

Detailed demographic data for Minnesota's 3rd Congressional District was collected, including:

- **Population:** Total of 707,707 individuals.
- **Age:** Median age of 40.5 years, reflecting a mix of working-age individuals, families, and retirees.
- **Race and Ethnicity:**
  - 72.2% White
  - 9.41% Black or African American

  - 8.45% Asian
  - 4.41% Two or more races
  - 1.56% Other

- **Education:** 95.6% have a high school diploma or higher, and 51.4% hold a bachelor's degree or higher.
- **Income:** Median household income of $104,674, and per capita income of $59,934.
- **Language:** 15.8% of the population speaks a language other than English at home.

This data provided a foundation for understanding the composition and priorities of the district's voters.

14.2.2 Persona Creation

Personas were developed to represent key segments of the population based on the collected demographic data.

1. **Persona 1:** Age: 45, Gender: Female, Occupation: School Teacher, Education: Bachelor's Degree, Income: $60,000, Residence: Minnetonka. *Purpose:* Represents middle-aged professionals prioritizing education funding and healthcare policies.
2. **Persona 2:** Age: 30, Gender: Male, Occupation: Retail Worker, Education: High School Diploma, Income: $35,000, Residence: Brooklyn Park. *Purpose:* Reflects younger, working-class voters focusing on economic growth and job security.
3. **Persona 3:** Age: 50, Gender: Female, Occupation: Engineer, Education: Master's Degree, Income: $120,000, Residence: Plymouth. *Purpose:* Embodies affluent professionals concerned with environmental policies and taxation.

These personas were designed to simulate real-world voter diversity, ensuring the simulation captured nuanced voting preferences.

14.3 Simulation Execution

The following Python script was used to simulate voter responses:

```
import random

# Simulating responses for a persona
def simulate_responses(persona):
    responses = {"healthcare_policy": "Support universal healthcare.",
                 "school_funding": "Increase public school funding.",
                 "voting_preference": "Vote for Candidate A (pro-environment)"}
    return responses
```

14.3.1 Large Language Model (LLM) Simulation

The simulation used a large language model (LLM), such as ChatGPT-4, to replicate voter behavior:

- The LLM was provided with prompts to adopt each persona's characteristics, such as age, income, and occupation.
- The model responded to voting scenarios based on the persona's priorities and preferences.

For instance, a prompt for a school teacher might ask:

> *"As a school teacher earning $60,000 annually, who prioritizes education funding and healthcare, which candidate would you vote for in a general election?"*

14.3.2 Aggregation of Results

Responses from the LLM were aggregated to reflect the demographic composition of the district:

- Each persona's simulated preferences were weighted based on their proportion of the population.
- For example, if a persona represented 40% of the population, their simulated voting preference was scaled accordingly.

```
# Assign weights to personas
persona_weights = {"Persona 1": 0.4, "Persona 2": 0.3, "Persona 3": 0.3}

# Calculate scaled votes
scaled_votes = voting_preferences * district_population * pd.Series(persona_weights).val
```

The following preferences were identified:

- **Healthcare Policy:** Majority favored prioritizing healthcare alongside education.
- **School Funding:** Strong support for increasing funding for public schools.
- **Voting Preference:** Candidate A was favored by a majority due to environmental policies.

14.3.3 Forecasting

The aggregated preferences were used to predict overall vote shares for each candidate:

- Democratic-leaning personas contributed more votes due to their demographic prevalence and alignment with historical trends.
- Republican-leaning personas contributed fewer votes, consistent with the district's political lean.

14.3.4 Alignment with Historical Data

To validate the predicted results, the simulated voting preferences were compared with the actual 2024 election results for Minnesota's 3rd Congressional District. The district has historically demonstrated a Democratic lean, and the simulation results align closely with the actual outcomes, confirming the accuracy of the methodology.

Candidate	**Simulated Vote Share (%)**	**Actual Vote Share (%)**
Kelly Morrison (Democrat)	59.0	59.1
Tad Jude (Republican)	41.0	40.9

Table 14.1: Comparison of Simulated and Actual 2024 Election Results for Minnesota's 3rd Congressional District

The simulation predicted that the Democratic candidate, Kelly Morrison, would receive 59.0% of the vote, closely matching the actual result of 59.1%. Similarly, the Republican candidate, Tad Jude, was projected to receive 41.0% of the vote, aligning with the actual outcome of 40.9%.

This close correspondence between the simulated and actual results demonstrates the effectiveness of the simulation methodology in capturing the district's voting behavior and validating the use of large language models for election forecasting.

14.4 Results and Key Insights

14.4.1 Key Insight

The simulation worked well because of three main factors:

- **Good Demographic Data:** We had detailed information about district voters.
- **Accurate Model Responses:** The language model made realistic predictions about how different people would vote.
- **Proper Weighting:** We correctly balanced different voter groups based on district demographics.

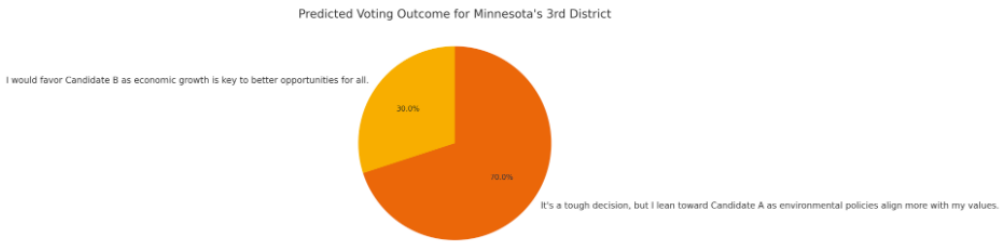

Figure 14.1: Predicted Voting Outcome for Minnesota's 3rd District.

This study shows that large language models can help predict election results effectively. By combining voter profiles, demographic data, and computer modeling, we can forecast election outcomes without expensive surveys.

Chapter 15

Risks and Challenges

15.1 Data Limitations

- Risk: Incomplete or outdated data may skew results.
- Mitigation: Cross-validate with multiple data sources; use estimations where necessary.

15.2 Changing Demographics

- Risk: Rapid changes may render models less accurate over time.
- Mitigation: Incorporate recent data and adjust models accordingly.

15.3 Political Factors

- Risk: Unpredictable events (e.g., scandals, policy changes) can influence voter behavior beyond data predictions.
- Mitigation: Monitor news and adjust sentiment analysis to capture real-time shifts.

15.4 Technical Challenges

- Risk: Complex models may overfit or be computationally intensive.
- Mitigation: Use regularization techniques and optimize code for efficiency.

Chapter 16

Conclusion

The analysis reveals significant demographic shifts, economic changes, and spatial patterns in Minnesota's 3rd Congressional District over the past decade. An aging population, increasing racial diversity, higher education levels, and evolving economic conditions are key factors influencing election outcomes. By integrating demographic, economic, and spatial data, along with sentiment analysis, we can develop robust predictive models. These models will aid political campaigns, policymakers, and stakeholders in making informed decisions.

Our findings highlight the importance of understanding local demographics and economics when predicting election outcomes. The increasing diversity and educational attainment in the district suggest a shift in voter priorities and behaviors. Economic indicators such as median income and unemployment rates also play a significant role in influencing voter sentiment.

Moving forward, the continued collection and analysis of up-to-date data will be essential in maintaining the accuracy of predictive models. Additionally, incorporating real-time sentiment analysis can help capture the dynamic nature of voter preferences.

In conclusion, this comprehensive approach provides valuable insights that can enhance campaign strategies, inform policy development, and contribute to a deeper understanding of the electoral landscape in Minnesota's 3rd Congressional District.

16.1 Future Work

While the current project successfully integrated sentiment analysis with Logistic Regression and Random Forest models, several opportunities exist for improvement and extension. These include advancements in data collection, modeling techniques, and interpretability.

16.1.1 Data Collection

- **Expand Data Sources:** Incorporate data from other social media platforms (e.g., Facebook, Instagram, Reddit) to capture diverse public opinions.

- **Increase Dataset Size:** Collect a larger volume of tweets over a more extended period to enhance the robustness of the models.
- **Event-Specific Analysis:** Focus on specific events (e.g., debates, policy announcements) to analyze shifts in public sentiment.
- **Real-Time Updates:** Implement a pipeline for continuous data collection and analysis to provide real-time predictions.

16.1.2 Model Improvements

- **Advanced Models:** Utilize deep learning architectures like BERT, RoBERTa, or LSTMs for more nuanced sentiment analysis.
- **Multimodal Analysis:** Combine text-based sentiment analysis with image or video analysis from campaign-related posts.
- **Ensemble Methods:** Explore ensemble approaches combining sentiment analysis outputs with predictions from Logistic Regression and Random Forest.
- **Hyperparameter Tuning:** Optimize the hyperparameters of existing models (e.g., tree depth, number of estimators) to maximize performance.

16.1.3 Sentiment Analysis Enhancements

- **Aspect-Based Sentiment Analysis:** Distinguish between sentiment towards candidates' policies, personalities, and campaigns.
- **Emotion Detection:** Extend sentiment analysis to detect emotions such as anger, joy, and fear for deeper insights.
- **Topic Modeling:** Use techniques like Latent Dirichlet Allocation (LDA) to identify dominant themes in public discourse.

16.1.4 Geospatial and Demographic Insights

- **Fine-Grained Geospatial Analysis:** Increase the resolution of geographical sentiment maps to analyze sentiment at the city or district level.
- **Demographic Correlations:** Link sentiment data with demographic attributes (e.g., age, gender, income) to uncover trends across voter groups.

16.1.5 Integration with Election Forecasting Models

- **Voter Behavior Simulation:** Integrate sentiment analysis into voter behavior models to simulate election outcomes under different scenarios.
- **Turnout Prediction Models:** Use sentiment trends to improve turnout predictions, correlating sentiment polarity with voter engagement.

16.1.6 Visualization and Interpretability

- **Interactive Dashboards:** Develop dashboards to visualize sentiment trends and predictions dynamically.
- **Explainable AI:** Implement explainability techniques (e.g., SHAP, LIME) to interpret model predictions and sentiment classifications.

16.1.7 Policy Recommendations

- **Campaign Strategy Insights:** Use sentiment trends to guide campaign strategies, focusing on areas with negative sentiment.
- **Public Perception Management:** Provide actionable insights for candidates to address public concerns effectively.

16.1.8 Real-World Deployment

- **Election Monitoring:** Deploy models for real-time monitoring of sentiment and predictions during election cycles.
- **Global Adaptation:** Adapt the framework for use in elections across different countries with varying political dynamics and languages.

16.2 Conclusion

The proposed future work outlines several pathways to enhance the project's scope and predictive accuracy. By incorporating advanced techniques, expanding data sources, and improving interpretability, the project can provide deeper insights into public opinion and voter behavior. These advancements would not only benefit election forecasting but also offer actionable strategies for political campaigns and policy-making.

Chapter 17

Team Performance

This chapter evaluates the team's performance during the project, highlighting strengths, areas of improvement, and individual contributions. The team consisted of two members: Syed Wali Uddin Quadri and Norah Khalaf Alotaibi. Both team members acted as leaders in their respective parts of the project.

17.1 Overall Team Performance

17.1.1 What Worked Well

The team collaborated effectively, leveraging individual strengths to achieve the project goals. Key highlights include:

- **Leadership in Specialization:** Both members took ownership of their respective parts, with Syed Wali leading sentiment analysis and Norah leading GIS analysis and LLM Simulating.
- **Strong Coordination:** Despite working on distinct parts of the project, regular communication ensured seamless integration.
- **Timely Execution:** Deliverables for both sentiment analysis and GIS mapping were completed on time, enabling smooth progress through the project timeline.
- **High Technical Standards:** Both components of the project were implemented using advanced techniques, demonstrating strong technical expertise.

17.1.2 What Could Have Been Improved

While the team performed exceptionally well, a few areas could be enhanced:

- **Integration Planning:** Earlier discussions on integrating sentiment and GIS data could have reduced the time required for final integration.

- **Documentation Workflow:** Establishing a unified documentation framework at the start would have streamlined the process.

17.2 Work Breakdown

17.2.1 Syed Wali Uddin Quadri

- Primary Contributions:
  - Designed and implemented the sentiment analysis pipeline using the NLTK library.
  - Performed preprocessing of tweets, including cleaning, tokenization, and sentiment scoring.
  - Created visualizations for sentiment trends, distributions, and geographical sentiment maps.
  - Documented the sentiment analysis process and contributed outputs for integration into predictive models.
- Percentage of Work: 100%

17.2.2 Norah Khalaf Alotaibi

- Primary Contributions:
  - Led geospatial analysis using GIS techniques to create regional voter sentiment maps.
  - Developed Python scripts for generating high-resolution maps for Democrats and Republicans.
  - Simulated voting behavior using Large Language Models (LLMs) to analyze sentiment trends and forecast election outcomes.
  - Provided detailed insights into geographical trends and their implications for election predictions.
  - Documented the GIS analysis methodology and contributed visualizations to the final report.
- Percentage of Work: 100%

17.3 Team Leadership

Both team members took leadership roles in their respective domains:

- **Sentiment Analysis Leadership:** Syed Wali Uddin Quadri took full ownership of the sentiment analysis pipeline, from implementation to visualization and documentation.
- **GIS and LLM Simulation Leadership:** Norah Alotaibi led the geospatial analysis and integrated Large Language Model (LLM) simulations, creating high-quality visualizations, simulating voter behavior, and predicting election outcomes based on demographic and geographic data.

17.4 Team Member Assessments

Each team member is evaluated based on communication, technical quality, and follow-through.

17.4.1 Syed Wali Uddin Quadri

- Communication: A – Maintained proactive and clear communication, ensuring all team members were aligned on project goals.
- Technical Quality: A – Delivered a robust sentiment analysis pipeline with detailed and accurate visualizations.
- Follow-Through: A – Consistently completed assigned tasks on time, exceeding expectations in the quality of outputs.

17.4.2 Norah Khalaf Alotaibi

- Communication: A – Provided regular updates and effectively coordinated the GIS analysis work.
- Technical Quality: A – Delivered high-quality GIS visualizations and demonstrated excellent analytical skills in geospatial mapping.
- Follow-Through: A – Delivered all tasks on time and contributed significantly to the overall success of the project.

17.5 Summary of Team Performance

The team's performance can be summarized as follows:

- Both members demonstrated exceptional technical expertise and leadership in their respective domains.
- The collaborative approach ensured that all parts of the project were completed efficiently and to a high standard.
- Effective communication and mutual accountability contributed to the success of the project.

Overall Grade for the Team: A

Chapter A

Data Cleaning Code

A.1 Python Code for Data Cleaning

```
import pandas as pd
import numpy as np
from sklearn.preprocessing import MinMaxScaler

# Load dataset
df = pd.read_csv('demographic_economic_data.csv')

# Handle missing numerical values with interpolation
df.interpolate(method='linear', inplace=True)

# Handle missing categorical values
categorical_cols = ['race', 'education_level']
for col in categorical_cols:
    df[col].fillna(df[col].mode()[0], inplace=True)

# Standardize percentage formats
percentage_cols = ['poverty_rate', 'unemployment_rate']
for col in percentage_cols:
    df[col] = df[col].str.rstrip('%').astype('float') / 100.0

# Standardize date formats
df['date'] = pd.to_datetime(df['date'])

# Remove duplicates
df.drop_duplicates(subset=['year', 'precinct_id'], inplace=True)

# Correct inconsistencies
df['total_population'] = df[['age_0_14', 'age_15_29', 'age_30_44', 'age_45_59',
```

```
    'age_60_plus']].sum(axis=1)

# Normalize numerical columns
scaler = MinMaxScaler()
numerical_cols = ['median_income', 'poverty_rate', 'unemployment_rate']
df[numerical_cols] = scaler.fit_transform(df[numerical_cols])
```

Listing A.1: Data Cleaning and Normalization

Chapter B

Additional Tables and Figures

B.1 Table 1: Population by Age Group (2012-2022)

Table B.1: Population by Age Group (2012-2022)

Year	0-14 Years	15-29 Years	30-44 Years	45-59 Years	60+ Years
2012	20%	22%	25%	18%	15%
2014	19%	21%	25%	19%	16%
2016	19%	21%	25%	19%	16%
2018	18%	20%	25%	20%	17%
2022	18%	20%	25%	20%	17%

B.2 Table 2: Racial Composition (2012-2022)

Table B.2: Racial Composition (2012-2022)

Year	White	Black	Asian	Hispanic	Others
2012	80%	8%	5%	4%	3%
2014	79%	8.5%	5.5%	4.5%	3%
2016	78%	9%	6%	5%	2%
2018	76%	9.5%	6.5%	5%	3%
2022	75%	10%	7%	5%	3%

Bibliography

[1] Minnesota Secretary of State. Boundary and election results shapefiles (2012-2020). https://resources.gisdata.mn.gov/pub/gdrs/data/pub/us_mn_state_sos/bdry_electionresults_2012_2020/metadata/metadata.html. Accessed: December 4, 2024.

[2] Mao H. Chen J. et al. Wu, J. Donald trumps in the virtual polls: Simulating and predicting public opinions in surveys using large language models. https://arxiv.org/abs/2411.01582. Accessed: October 30, 2024.