Syed Rahman
A20463481
Muzamil Faisal
A20547684

Analyzing Texas Congressional District 32: A Data-Driven Approach to Predicting Election Outcomes

Introduction

Congressional District 32 (CD-32) in Texas, which has a population of 755,403 according to the 2021 American Community Survey, is home to a vibrant and diversified population in Dallas County. A youthful, multicultural community is reflected in the district's demographic and economic landscape, and it has a big influence on the social and economic dynamics of the area. This report explores CD-32's demographic composition, economic trends, and voting patterns, intending to analyze the data to understand the district's distinctive political views. Through the lens of data analysis, this report documents findings, analysis techniques, and insights to provide a comprehensive understanding of Texas's Congressional District 32. This project intends to use previous election and demographic data to generate predictive insights for congressional representative elections in Texas's Congressional District 32. We intend to estimate election results by candidate, party winner, vote counts, and percentages, as well as project presidential election outcomes for this district. The algorithm will also estimate voter turnout, assisting in forecasting engagement trends and voter behaviors in future elections. By sifting data from previous elections and layering in demographic variables, we may find predictive factors and investigate how demographic and sociopolitical processes influence vote outcomes.

Data sources

Primary data sources for this project will include a mix of government databases, academic datasets, and local Texas-specific resources, providing detailed historical election results, demographic information, and insights on election dynamics. This research attempts to give a thorough analysis of CD-32's distinctive features by looking at both qualitative and quantitative components of the district. This research incorporates data from the U.S. Census Bureau's American Community Survey (ACS) 2021 1-Year Estimates to fully analyze CD-32. The strategy involves analyzing residential, economic, and demographic data to produce a comprehensive picture of the district. Demographic data such as age, income, education, ethnicity, and voter turnout will be sourced from ACS and Census data to create a socio-demographic profile of the district. Capitol.Texas.Gov serves as a central resource for datasets from past election cycles (e.g., *2018 Democratic Primary*, *2018 General*, *2020 Democratic Primary*, *2022 General*, etc.). Data from this source will be parsed to identify party-specific results for Democratic and Republican primaries and general elections.

Exploratory data analysis

Texas' Congressional District 32 has a diverse household and demographic structure, according to the 2021 American Community Survey. This process began with informal analysis, using summary statistics to gain a preliminary understanding of the district's age distribution, income brackets, and employment sectors. This district, with a population of 755,403, has a

young median age of 33.2 years, and children and young adults play an important role, with approximately 25% of the population under the age of 18. The district's gender ratio is balanced, with males accounting for 49% of the population, slightly less than females. It's essential to look at recent voting trends, demographics, and economic indicators in this area. CD32 is a highly diverse district, with Hispanic or Latino residents comprising 37.5% of the population, and significant Black (19.8%) and White (38.2%) populations as well as a developing Asian minority of 7.4%. The district's demographic and socioeconomic diversity can impact voter preferences and turnout. For instance, a large proportion of households in CD32 (42.8%) report using a non-English language at home, which is significantly higher than the national average of 21.7%, this can potentially influence campaign strategies and voter outreach (U.S. Census Bureau 2021 data).

Economically, the district contains 313,135 households, with married couples accounting for 35.2% and a sizable proportion of single-adult homes. Income levels show an 18.6% concentration between $50,000 and $74,999. The district's demographic shifts and economic factors will be critical. The median income in CD32 stands at $58,647, with nearly 15.8% of residents living below the poverty line, creating a mixed economic profile that could impact voting priorities around issues like healthcare, education, and economic support. Employment indicators point to a solid economy, with a 70.6% employment rate and a low unemployment rate of 5.4%, while high-skill industries employ 40.9% of the workforce in disciplines such as management, business, science, and arts. The sales and service industries also contribute to the district's broad labor market. This also influences preferences for policies supporting economic growth and job security as most people living here work a high skill job that pays more income particularly with 40.9% of residents employed in management, business, and science roles.

In past elections, CD32 has shifted from solid Republican representation to a more competitive split. In the 2020 House election, Democratic incumbent Colin Allred won with 51.9% of the vote against Republican Genevieve Collins, who received 47.8%. This indicates a tight race and suggests that the district leans slightly Democratic but remains highly competitive, influenced by both local issues and national party affiliations. Additionally, voter turnout in CD32, like much of Texas, often varies with presidential election cycles, increasing during these years compared to midterms. Texas as a state saw a turnout increase in 2020, with approximately 66% of eligible voters participating, reflecting a nationwide trend toward higher engagement (FEC 2020 data; Wikipedia election data for TX CD32). Politically, Donald Trump won the 2020 presidential election in Texas with 52.1% of the vote, while Joe Biden received 46.5%. Colin Allred currently represents Congressional District 32 in the United States House, while Senators Ted Cruz and John Cornyn represent Texas on the federal level.

This demographic, economic, and political profile, based on the US Census Bureau's thorough 2021 survey, presents a structured picture of District 32's distinct and diverse community. By integrating demographic data, voting history, and economic factors, we can build a model to anticipate election outcomes, voter turnout, and the percentage of votes by party for both congressional and presidential elections in CD32.

Data selection and data cleaning

The data preprocessing approach consisted of many needed stages that ensured the election-related datasets' quality and usefulness. Initially, CSV files from various election years and types—from 2018 to 2024 general elections, democratic and republican parties of the state of Texas—were thoroughly processed. The first step was to clean the data, which involved cleaning the County Voting District column to maintain just numeric values with a custom function. Following that, missing values in crucial columns, such as Voter Registration and Turnout, were handled correctly; missing voter registration data was filled with the median, while missing turnout data and vote counts were replaced with zeros, presuming no votes were cast. Rows with missing values in crucial identifiers such as County voting district and Voting District Key were removed to ensure data integrity.

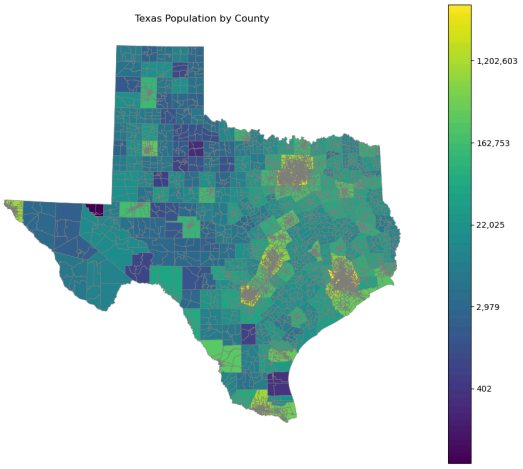
Following cleaning, data type consistency was guaranteed by converting important columns to numeric types as needed, particularly for voter registration and turnout figures, and standardizing identifier columns to string format for prospective merging procedures. More feature engineering was done to extract additional insights from the data, such as estimating the turnout rate as a fraction of registered voters and computing vote percentages for various candidates. This extensive preparation method not only cleaned and arranged the data, but also added important features to prepare it for analysis and modeling.

Observation of data

Once the district data is compiled, the code prepares the input features for prediction, utilizing the Turnout Rate and other relevant vote percentage columns. Missing values in the input data are handled using a mean imputation strategy to ensure completeness before making predictions. The regression model predicts a vote percentage of approximately 58.10% for Trump in the presidential election, while the classification model determines that Harris is likely to be the winner in district 32. Additionally, the predicted voter turnout rate for this district is approximately 28.03%, indicating a relatively low expected participation rate among eligible voters. In terms of congressional representation, the predicted vote percentage for the congressional candidate mirrors Trump's, at about 58.10%, with Harris again predicted to win. This comprehensive analysis provides valuable insights into the electoral dynamics within district 32. In terms of congressional representation, the estimated vote percentage for the congressional candidate is similar to Trump's, at around 58.10%, with Harris again expected to win. This extensive research sheds light on the election processes in district 32. Preliminary predictions suggest that, based on historical voting patterns, candidate preferences may lean similarly to previous election cycles. However, while the model estimates voter turnout and voting percentages, these predictions are based on historical data rather than actual future conditions. As a next step, we will further validate these findings against previous election results in District 32 to refine predictions of turnout and voting outcomes.

Spatial analysis

Currently, we have created a thorough map of Texas depicting population density throughout the state, which is an important initial step in our spatial research. The map uses a color gradient ranging from dark purple (low population density) to yellow (high population density). This color scheme efficiently depicts population discrepancies across Texas.



The eastern half of the state, particularly near large urban centers like Dallas, Houston, and San Antonio, has the highest population density, as evidenced by the use of yellow and green hues. This concentration of population is linked to economic activity and infrastructure development in major urban areas. Conversely, the western half of Texas, particularly in regions like the Panhandle and the Trans-Pecos, has much lower population density, as indicated by darker purple tones on the image. This distribution reflects the rural nature of these places, which have fewer residents and lower economic activity.

As we progress with our study, we intend to improve our analysis on Congressional District 32. This will entail mapping not only population density, but also socioeconomic and demographic differences within the district. To present a more complete picture of the area, we will look at crucial factors such as income levels, educational attainment, and ethnic diversity. The district's diverse demographics, with a strong Hispanic (37.5%) and African American (19.8%) population, will impact election outcomes and voter behavior. We also are currently finding more dataset to map the election result and turnout for our district along with Texas.

Plan for completion

To successfully complete our project on predicting the outcomes of congressional elections in Texas's Congressional District 32, we have outlined a comprehensive plan that addresses the necessary tasks, team responsibilities.

First and foremost, we must improve our data visualization abilities in order to effectively communicate voter turnout and other crucial metrics through graphical representations. This entails constructing plots and charts to better comprehend the district's voting habits. We have already collected and analyzed key datasets, such as historical voting statistics and demographic information, but we must improve our models to ensure they are efficient and reliable in predicting election results. We specifically want to know who won the congressional representative election, the number or percentage of votes cast, the outcomes of the presidential election, and voter turnout rates.

Given that our team consists of two members, we will split responsibilities to maximize our efficiency. One team member will focus on data analysis, including parsing the datasets to extract pertinent information about past elections, such as the number and percentage of votes received by each candidate in both congressional and presidential elections. This analysis will also extend to voter turnout rates among eligible voters in District 32. Meanwhile, the second member will concentrate on developing the visualization tools needed to present our findings graphically, allowing us to illustrate voter trends and demographic impacts effectively.

One of the most significant problems we encounter is the accuracy and completeness of our databases. If the data we use is out of date or improperly presented, our analysis may produce misleading results. For example, errors in voter registration data or underreporting of turnout numbers could have a substantial impact on our projections for election outcomes. Furthermore, Texas has undergone legal and political shifts in voter registration and suppression, complicating our capacity to gather reliable, comprehensive statistics. Targeting several data sources adds another level of complication. This can present technological issues, such as preserving data integrity during transformations and ensuring that all variables match correctly across datasets. Finally, when we develop predictive models, we must be aware of overfitting, which occurs when our models grow extremely complicated and specialized to our training data, thus reducing its generalizability to future elections. Balancing the complexity of our models with interpretability will be critical to ensure that our findings are actionable and meaningful.

To enhance and validate the model, we will iteratively test it against a range of past CD-32 election results. This procedure will involve testing forecast accuracy over numerous election cycles and fine-tuning model parameters to improve precision. In addition, we will include a broader range of features and real-time data updates, with a focus on comprehensive economic indicators such as household income as well as critical socio-demographic aspects. By incorporating these various variables, our model gains strength and adaptability, allowing it to account for the dynamic impact of economic conditions on voter behavior. These ongoing improvements attempt to create a more precise, actionable model for analyzing how economic developments and demographic shifts influence election dynamics in Texas Congressional District 32.

Results :

1. **Training Result :**

```
See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
  X.replace([float('inf'), -float('inf')], float('nan'), inplace=True)
Regression Model MSE: 3.077563365631778e-29
Classification Model Accuracy: 0.9959137628478315
```

2. **Validation Result**

```
See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
  X.replace([float('inf'), -float('inf')], np.nan, inplace=True)
Regression Cross-Validation MSE Scores: [2.43389288e-04 1.13511735e-28 1.04779254e-28 9.76089493e-29
 7.55417830e-01]
Mean MSE: 0.15113224380417992
C:\Users\muzam\AppData\Local\Programs\Python\Python312\Lib\site-packages\sklearn\model_selection\_split.py:737: UserWarning: The least populated class in
y has only 1 members, which is less than n_splits=5.
  warnings.warn(
Classification Cross-Validation Accuracy Scores: [0.98949611 0.99425921 0.99508636 0.99548748 0.99242899]
Mean Accuracy: 0.9933516320021702
```

3. **Prediction Result :**

```
Predicted vote percentage for TrumpR_24P_President: 0.5811688311688344
Predicted winner for President in district 850025: BidenD_24P_President_Percentage
Predicted voter turnout rate for district 850025: 28.03%
Predicted vote percentage for Congressional Representative: 0.5811688311688344
Predicted winner for Congressional Representative in district 850025: BidenD_24P_President_Percentage
```

References

U.S. Census Bureau. "American Community Survey 1-Year Estimates." *U.S. Census Bureau*, 2021, https://www.census.gov/programs-surveys/acs.

Texas Legislative Council. "Election Data." *Capitol.Texas.Gov*, n.d., https://www.capitol.texas.gov.

Federal Election Commission. "Federal Election Commission 2020 Election Data." *Federal Election Commission*, 2020, https://www.fec.gov.

Wikipedia contributors. "2020 United States House of Representatives Elections in Texas." *Wikipedia, The Free Encyclopedia*, 2020, https://en.wikipedia.org/wiki/2020_United_States_House_of_Representatives_elections_in_Texas.

OpenAI. *ChatGPT*. 2023, OpenAI, https://www.openai.com/chatgpt.