Mining Patterns within Non-Voters

IST 652-Scripting for Data Analysis
Group: 9
Team Members: Aadil Zikre, Lahari Chowtoori, Indraneel Timare

1. Introduction

The aim of this project was to uncover patterns and insights related to non-voter behavior in the United States. By analyzing a dataset of survey responses collected by Ipsos for FiveThirtyEight, we sought to better understand the reasons why many eligible voters choose not to participate in elections and how this impacts democratic representation.

Key areas of investigation in this study centered around three main questions related to non-voter attitudes and experiences:

Firstly, we examined how non-voters perceive the impact of elections on their personal lives. This included analyzing their opinions on whether they believe the actions and decisions of elected officials at various levels (local, state, federal) have a significant influence on their day-to-day experiences and well-being. Understanding these perceptions is crucial, as a belief that elections are irrelevant or disconnected from one's personal circumstances may contribute to a lack of motivation to participate in the democratic process.

Secondly, we explored non-voters' views on different systems of governance, with a particular focus on their opinions of democracy. This involved gauging their level of support for democratic principles and institutions, as well as their attitudes towards alternative forms of government such as rule by experts, a strong leader, or the military. Assessing these preferences helps shed light on whether disillusionment with democracy itself is a key factor driving non-participation, or if other issues are more salient.

Finally, we investigated the potential impact of personal tragedies on non-voters' political orientation and engagement. This included examining whether experiences such as job loss, health crises, or the death of a loved one may influence an individual's political attitudes or party affiliation. Analyzing these relationships is important for understanding how personal hardships might shape political behavior and whether they play a role in the decision to abstain from voting.

By delving into these three key areas, our aim was to develop a more nuanced and comprehensive understanding of the complex factors that contribute to non-voting behavior. Through analyzing non-voters' perceptions of electoral impact, their views on governance systems, and the influence of personal tragedies, we sought to identify patterns and insights that could inform efforts to boost civic engagement and strengthen democratic participation.

2. Data Source

The primary dataset used in this analysis came from a poll conducted by Ipsos on behalf of FiveThirtyEight, which was provided to us in the form of a structured CSV file. The dataset

included 8,327 survey responses from a sample representative of the U.S. population, with oversampling of young Black and Hispanic individuals.

To supplement the CSV file and provide context for the survey questions and responses, we were also given access to a detailed codebook in PDF format. This unstructured document described each survey question and the corresponding possible responses, allowing for a more comprehensive understanding and interpretation of the data.

Aristotle, a data vendor, matched 64% of the respondents to a voter file using names, zip codes and address segments. For our analysis, we focused on the 5,239 respondents who were matched to the voter file and eligible to vote in at least 4 elections, as well as an additional 597 respondents who self-identified as voting "rarely" or "never".

Data Preparation

The original dataset contained 5,836 entries across 119 features. To prepare the data for analysis, we first mapped the information in the CSV file to the corresponding questions and responses outlined in the codebook PDF. This process ensured that we had a clear understanding of the meaning and context behind each data point. During this stage, no data quality issues such as missing values were detected in the columns of interest.

One of the key challenges in this study was effectively preprocessing the unstructured codebook PDF to extract relevant information about the survey questions, subquestions, and possible answers. To tackle this task, we employed a combination of Large Language Models (LLMs) and Retrieval Augmented Generation (RAG) techniques.

LLMs, such as GPT-3, are powerful AI models that can understand and generate human-like text based on patterns learned from vast amounts of data. These models have shown remarkable capabilities in tasks such as question answering, summarization, and natural language understanding. In our preprocessing pipeline, we leveraged an LLM to analyze the structure and content of the codebook PDF and identify the relevant sections containing the survey questions and their corresponding answer options.

However, relying solely on LLMs for information extraction can sometimes lead to inconsistencies or errors, especially when dealing with complex document structures. To mitigate this issue, we employed Retrieval Augmented Generation (RAG), a technique that combines the strengths of LLMs with a retrieval component. RAG allows the model to access and incorporate relevant information from external sources during the generation process, improving the accuracy and contextual understanding of the extracted content.

In our case, we used RAG to enhance the LLM's ability to parse the codebook PDF by providing it with additional context and examples of correctly formatted survey questions and answers. This was achieved by creating a curated set of training examples that demonstrated how to properly extract and structure the relevant information from the PDF. By conditioning the LLM on these examples through RAG, we were able to guide the model towards more accurate and consistent parsing of the codebook content.

Once the LLM and RAG components identified and extracted the relevant questions, subquestions, and answer options from the PDF, we transformed this information into structured Python dictionaries. Each dictionary entry contained the question text as the key and a nested dictionary of subquestions and their corresponding answer options as the value.

This structured format allowed for easy integration with the main survey response data from the CSV file. Example output from the LLM is shown in Figure 1 below.

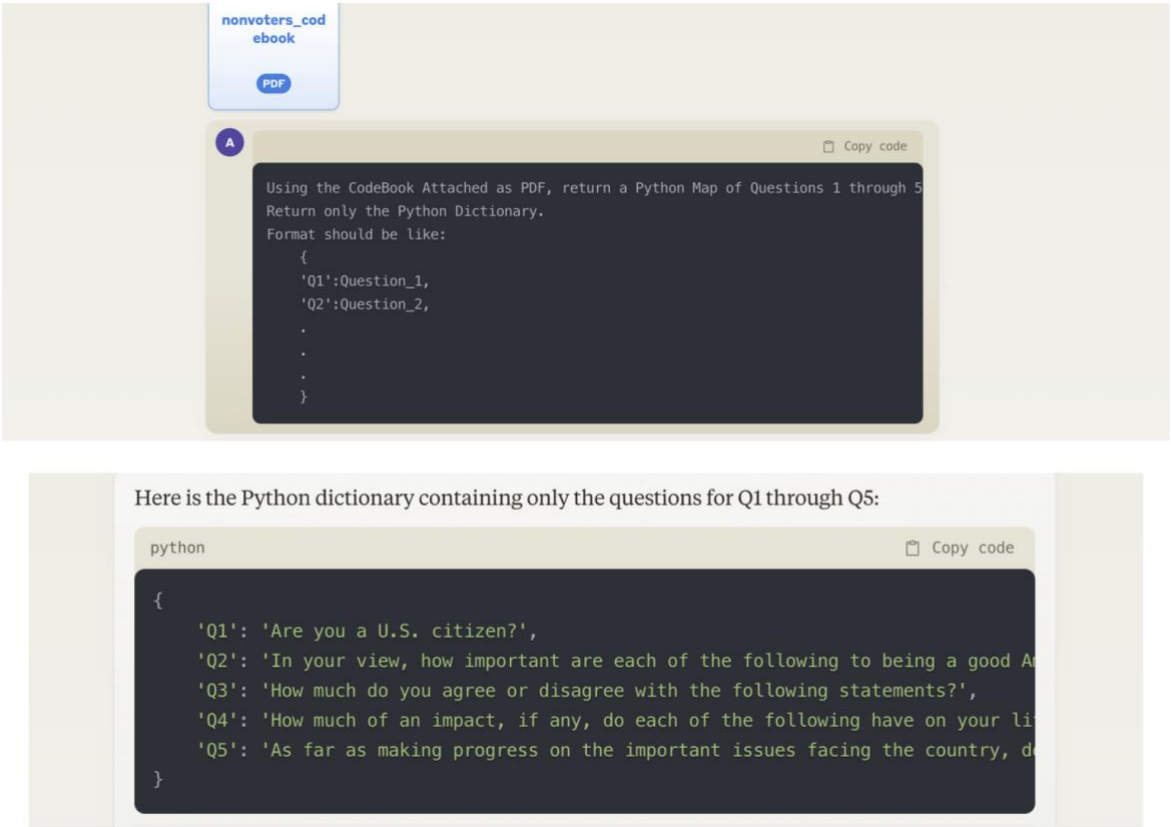

Figure 1 Retrieval Augmented Generation using an LLM, Claude Opus

To align our dataset with the key areas of research described in the introduction, we selected 25 essential columns that were most relevant to our investigation. These columns included information related to how non-voters perceive the impact of elections on their personal lives, their opinions on various systems of governance, and whether they have experienced personal tragedies that may influence their political orientation. Additional data engineering and feature generation steps were performed, such as categorizing respondents by age group to infer their generation, and economic status. This was done to enable more granular analysis and comparisons between most relevant subgroups.

3. Methodologies

To understand patterns and trends among non-voters, we employed two key analysis methods:

1. Distribution Analysis

   We visualized the frequency of various survey responses using value count bar plots. This allowed us to see the most common opinions and demographic characteristics.

2. Comparative Analysis

   To compare distributions of responses across different subgroups (e.g. by age or income level), we used grouped bar plots. This revealed interesting variations in non-voter attitudes.

To gain insights from the datasets, we formulated 3 hypotheses.

Hypothesis 1: *Non-voters do not believe that elections significantly impact their personal lives.*

If non-voters perceive that the outcomes of elections have little to no bearing on their day-to-day experiences, they may feel less motivated to participate in the voting process. This hypothesis suggests that a lack of perceived personal relevance could be a key factor contributing to non-voting behavior.

Hypothesis 2: *Non-voters have lost faith in democracy as an effective system of governance.*

Disillusionment with democratic institutions and processes may lead individuals to disengage from political participation, including voting. This hypothesis posits that if non-voters view democracy as broken, corrupt, or ineffective compared to other systems of governance, they may see little point in casting a ballot.

Hypothesis 3: *Influence of Personal Tragedies*

Experiencing personal tragedies significantly influences the political orientation of non-voters. Traumatic life events such as job loss, health crises, or the death of a loved one can have a profound impact on an individual's worldview and political attitudes. This hypothesis suggests that non-voters who have faced such tragedies may undergo shifts in their political beliefs or affiliations, potentially leading to a change in their voting behavior or a decision to abstain from voting altogether.

Based on these hypotheses, we will explore key areas related to non-voters' perceptions of electoral impact, their opinions on various systems of governance, and the potential influence of personal tragedies on their political orientation. By analyzing data relevant to these areas, we aim to test the validity of our hypotheses and gain insights into the complex factors that contribute to non-voting behavior. Through this investigation, we hope to identify potential strategies for increasing civic engagement and democratic participation among non-voters.

The results of our analyses were collated through a combination of visual representations and summary statistics. The value count bar plots and grouped bar plots provided clear visual depictions of the distribution and comparison of responses across different categories. These visualizations were supplemented by tables presenting the exact counts and percentages for each response option, allowing for a more precise understanding of the data.

From the original dataset containing 119 features, we carefully selected 25 essential columns that were most relevant to our key research areas and hypotheses. These fields were chosen to provide insights into non-voters' perceptions of electoral impact, their opinions on various systems of governance, and the potential influence of personal tragedies on their political orientation.

Fields related to Hypothesis 1 (Impact on Personal Life):

1. Q32_1: Impact of U.S. Congress on daily life
2. Q32_2: Impact of President on daily life
3. Q32_3: Impact of Supreme Court on daily life
4. Q32_4: Impact of State Government on daily life
5. Q32_5: Impact of Local Government on daily life
6. Q32_6: Impact of Law Enforcement on daily life
7. Q32_7: Impact of Media on daily life
8. Q32_8: Impact of Corporations/Wall Street on daily life

Fields related to Hypothesis 2 (Opinion of Governing Systems):

9. Q23_1: Opinion on the governing system of democracy
10. Q23_2: Opinion on the governing system of a strong leader
11. Q23_3: Opinion on the governing system of experts making decisions
12. Q23_4: Opinion on the governing system of the military ruling the country

Fields related to Hypothesis 3 (Influence of Personal Tragedies):

13. Q37_1: Experienced a natural disaster in the past year
14. Q37_2: Experienced a serious health issue in the past year
15. Q37_3: Experienced long-term unemployment in the past year
16. Q37_4: Experienced the death of a close friend or family member in the past year

Nine additional fields like age, and income were also included which provided important demographic and behavioral information that allowed for more granular analysis and comparisons between subgroups.

4. Program Overview

The Python scripts used for this analysis performed the following key functions in order:

1. Data retrieval and parsing of survey questions and responses from the codebook PDF
2. Null value checks to ensure data quality
3. Feature engineering to generate additional data fields like age group and economic status categories
4. Extending the functionality of the Pandas DataFrame structure used to manipulate the data
5. Generating visualizations of response distributions and comparisons between subgroups

After the preprocessing is complete, the script generates a file “filtered_nonvoters_data.csv” which is saved in the same directory as the script. As the visualizations created are more than 50 in the script, we do not export all the graphs as png. To view the plots, it is best to run the script in a Notebook Environment. The graphs are generated and shown in the interactive terminal in the notebook.

5. Results and Conclusions

Our analysis revealed several interesting insights about non-voters. Firstly, we found that newer generations increasingly feel elected officials have less impact on their lives compared to older generations (Figure 2). This belief may be contributing to declining voter turnout among young people. However, despite this trend, non-voters across all age groups still consider democracy to be the best system of governance (Figure 3). This suggests that lack of participation is not due to a disbelief in democracy itself, but rather other factors. Another key finding was that experiencing personal tragedies does not greatly influence the political orientation of non-voters (Figure 4). This indicates that individual hardships alone may not be a significant driver of political attitudes or voting behavior.

Perhaps most concerning was the discovery that an alarming percentage of nonvoters were found to have no political affiliation at all. This points to a general disinterest and disengagement from politics among a substantial portion of the electorate. Addressing this apathy and finding ways to make politics more relevant and accessible to all citizens should be a key priority for those working to strengthen democracy.

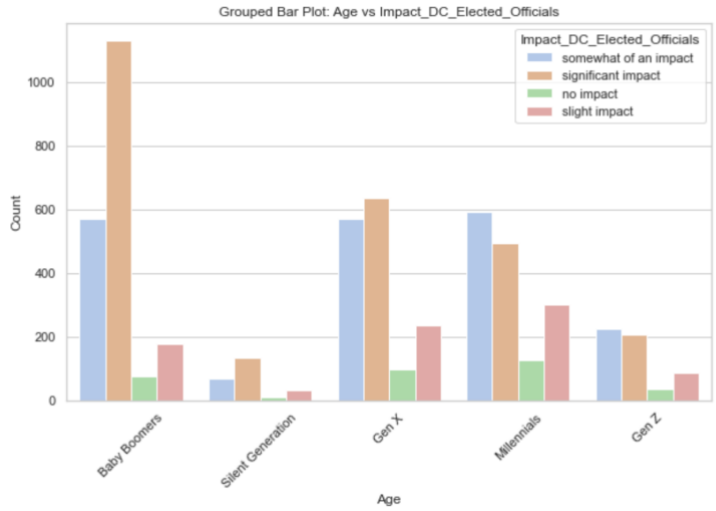

Figure 2 Plot that shows how older generations feel DC Elected Officials have significant impact on their lives whereas the newer generations feel the impact in only somewhat.

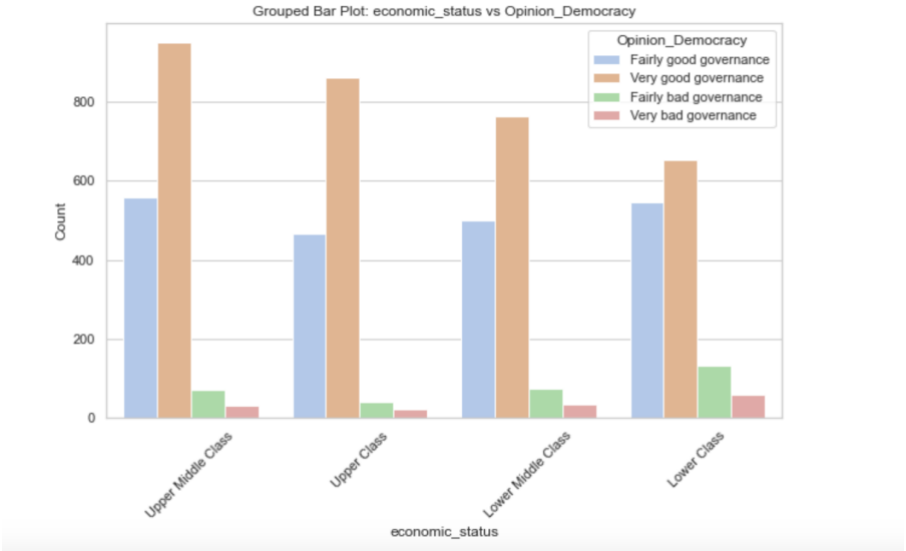

Figure 3 Plot that shows an overwhelming support for democracy across all economic classes

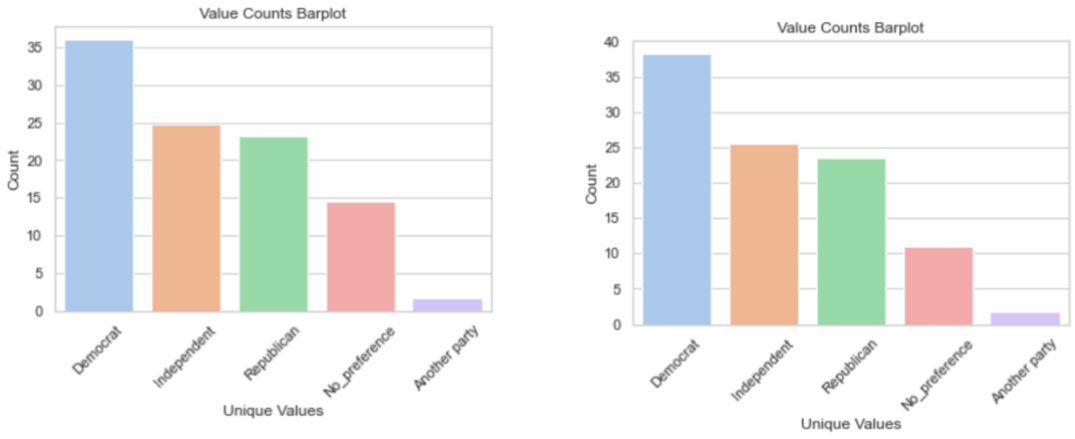

Figure 4 On Left is a distribution of Non Voters according to their political orientation, On Right, is the same distribution but for non-voters who had a family member affected by COVID

6. Limitations and Future Scope

While this study revealed some intriguing patterns, there were limitations that should be noted. One major constraint was that the dataset did not include a voter category that would allow direct comparison of the effect of various factors on likelihood of voting. Having data on the voting histories of survey respondents would enable more robust comparative analysis between voters and non-voters. Additionally, there was political and age bias in the sample, with older generations overrepresented. Future studies should aim for more balanced samples across age groups to ensure the findings are truly representative.

It's important to acknowledge that the scope of this project was limited to analyzing patterns and opinions among non-voters. To gain a more comprehensive understanding, further research is recommended to dive deeper into the root causes of increasing voter apathy, especially among younger citizens. This could involve incorporating additional data sources beyond surveys, such as social media sentiment analysis or qualitative interviews, to paint a vibrant picture.

Finally, as this study focused on the United States context, expanding the research to examine non-voter behavior in other countries and political systems could reveal interesting cross-cultural comparisons. Understanding how different social, economic, and political environments shape citizen engagement with democracy would be a valuable avenue for future scholarship.

Overall, this project highlights the importance of working to better understand non-voters in order to strengthen participatory democracy. By identifying key trends and potential factors influencing voter turnout, we can develop more targeted strategies to engage citizens and ensure more representative governance. Continued research in this area is crucial for the health and effectiveness of democratic systems worldwide.

All members of our group actively and equally participated in scripting, presentation building and report making. The success of this project was only made possible because of active participation and equal contribution from all team members.