Election Systems in the United States and India

Zhenyu Wang

April, 2023

Abstract

This report is a summary of some key findings my JSC370 final project. This project presents some data analysis on the election system in the United States and India. It contains 3 sections, analysis of the United States, analysis of India, and compare and contrast between them.

The full analysis are delivered through interactive webpages (much more are here). For the detailed results, please refer to the webpages. And this report will attach few visualization since most of the visualizations are interactive, please refer to the websites for them too.

1 Introduction

Democracy is often touted as the best system of governance, as it provides citizens with a voice in decision-making and a mechanism for holding their leaders accountable. However, the pattern of democracy varies greatly across different countries and regions, with some nations thriving under a robust democratic system while others struggle to maintain even the most basic democratic institutions. With the help of data science, we will explore the patterns of democracy across the globe.

American democracy is often seen as a model for other countries around the world. The United States is one of the oldest and most stable democracies in the world, and its democratic institutions and traditions have been emulated by many other countries. India has the largest democratic exercises in the world, with over 900 million eligible voters across the country. They are are quite representative in the whole world. In this project, we will perform some data analysis on the election system in the United States and India, and compare and contrast between them.

Elections is the basement of any democratic systems around the world. They provide a mechanism for citizens to choose their representatives, participate in the political process, and hold their leaders accountable. Election is also a friendly way to use data science to analyze, as its data format is mostly numbers. Therefore, we will look into election datasets for both countries provided by Dataverse and Kaggle to discover the pattern of their democratic systems. Notice that this project is an exploratory project, we want to investigate the following general questions throughout the project:

- What are some key characteristics of the United States democratic system?
- What are some key characteristics of Indian democratic system?
- What are some difference between them?

2 Methods

2.1 Data Clean

For the United States dataset, we merge datasets that shared the same attributes for easier future analysis. And the resulting datasets are

- ‘usa_president_result‘: A dataset that contains president election result in both state and county level. Notice that, for data point in state level, we set ‘county_name‘ to be ‘NaN‘ for simplicty.
- ‘usa_state_level_result‘: A dataset that contains both senate and presidential election results, but only in state level.

There are many unique features in Indian election, this requires some special efforts in data clean. Some procedures are listed below.

- Splits of Parties

  For example, we have INC(I) and INC in the dataset, but refer to the same political party, the Indian National Congress. The difference is that INC(I) is used to differentiate between the Indian National Congress and its faction in the state of Tamil Nadu, which is known as the Indian National Congress (Indira). To ensure the consistency of the analysis, we treat every subdivision or splits of parties as their original parties. In R language, we do this through remove content inside the bracket by regex.

- Hung Parliament

  For example, there are 2 state level election in 2005 in Bihar. There was a fractured verdict in February 2005 Assembly Election, so another election was held in October 2005. This brought some challenges in the future analysis. Ideally, we want the election to be held periodically so that it is to obtain some patterns. And 2 data points in 2005 result in more weight on 2005's political situation if we want to apply any machine learning algorithm on it. Therefore, we remove the latter election such that political standing of the original situation is kept.

- Different Spellings of States

  There are some states have different 'name' but refer to the same place in the dataset. For example, we have Chattisgarh, Chhattisgarh, the latter one is the correct spelling. Different spellings may due to misspelling, official rename or abbreviation.

2.2 Exploratory Data Analysis

Multiple tools are used in the exploratory data analysis, we list most important ones.

2.2.1 Context Understand

Background research related to the topics is critical for this project. This is conducted through looking up from the literature and the internet. This information helps to apply suitable methods in the next step data analysis. For example, without understand the basic geographic knowledge about India, we are not able to perform spelling correct step in data clean.

2.2.2 Data Summarize

Calculate basic summary statistics for each variable to get a sense of the overall distribution and range of values. Look into range of each columns in the dataset or unique values in the dataset if applicable. This helps to identify obvious patterns within the data. Additionally, it helps to detect outliers, missing values, and errors that may need to be addressed before proceeding with further analysis. For example, by look into unique names of Indian party names, we can perform splits of parties in data clean stage.

2.2.3 Time Series Visualization

Time series visualization displays trends and patterns in data over time. It allows us to identify trends, seasonal patterns, cycles, and other patterns that may not be easily visible in tabular form. By visualizing the data over time, we can better understand the dynamics of the election results or other aspects of the data, and identify changes that may require further investigation. Multiple tools are used to perform time series visualization in this project, for example, line plots, box plots, animations.

2.2.4 Map Visualization

Map visualization is an essential tool for displaying geographic data and gaining insights into spatial relationships. It helps to reveal patterns and trends in data that may not be evident through other forms of data visualization. Maps are particularly useful for analyzing election data related because we clear see political boundaries through maps. By combining the map visualization and the time series visualization, we can understand how geographic patterns changed with data and vice versa.

2.2.5 Hypothesis Testing

Hypothesis testing tests whether an observed pattern or trend in the data is statistically significant or simply due to chance. In our analysis, we apply the Sieve-bootstrap Student's t-test for a linear trend.

2.3 Define Metric

Defining metrics is an essential method in data analysis. Metrics help to quantify and measure the data in a way that is relevant to the analysis goals, turn large scale dataset into interpret-able metric. Therefore, we can relate the concrete dataset with some vague conceptions. In this project, we defined **Number of Turning Points**. Given an region, this metric measures the times that this region flipped their election result, that is have a different winning party). Through this metric, we can connect it to political conceptions such as battle states (have high turning points) or safe states (have lower turning points).

2.4 Model Indian Party Coalition with Clustering

We want to model coalition among Indian political parties. This task does not come out of nowhere. Please refer to 3.2.1 and the website for the detailed motivation for this research task and method.

2.4.1 Feature Vector

We want to cluster political parties in India based on their voting results in same geographic regions. The intuition behind this is that for two parties A and B, in regions X and Y. If A and B got higher votes in X (comparing to other states they participated), but got lower votes in Y, A and B are more likely to have similar positions in the political spectrum.

We select the parties that participated elections in more than 20 states since other parties have too many missing values. We calculate each party's vote proportion (0 to 1) throughout all elections in each state (F(S, P) = total votes gained by that party P in state S / total votes in the state S), this is an element for the feature vector for that party. Each party has a feature vector with the length of number of states. For missing value in the vector, we replace it with the minimum in the data matrix since that party doesn't have any votes in that state, it is like a smoothing technique.

2.4.2 Hierarchical Clustering

Hierarchical Clustering is a widely used unsupervised learning method that groups similar objects or observations into clusters based on the similarity or distance between them. In this method, a dendrogram is created by recursively dividing or merging the clusters until all the objects belong to a single cluster. In our project, we apply hierarchical clustering with cosine distance metric on selected Indian political parties represented by feature vector.

3 Results

Here are some key findings in both countries in the exploratory data analysis.

3.1 Voting Population

3.1.1 India

From Figure 1 (website, India section), in 1977, 188 million voters voted for the national election in India. And this number grew with time significantly, in 2014, more than 500 million people voted. We can also tell this giant voting population from the size of assembly constituency. From Figure 4, In 2014, each state is about 1 million voters on average, which only determined one seat in the Lok Sabha.

3.1.2 The USA

We visualized population voting count in the United States in bar plots. Figure 1 is for the presidential election and senate election. We can tell that after 2008, democratic party always gain more population votes in presidential election. And we can also see a increasing trend of total voting population in the presidential election. In 2020 presidential election, the voting population hit 150 million.

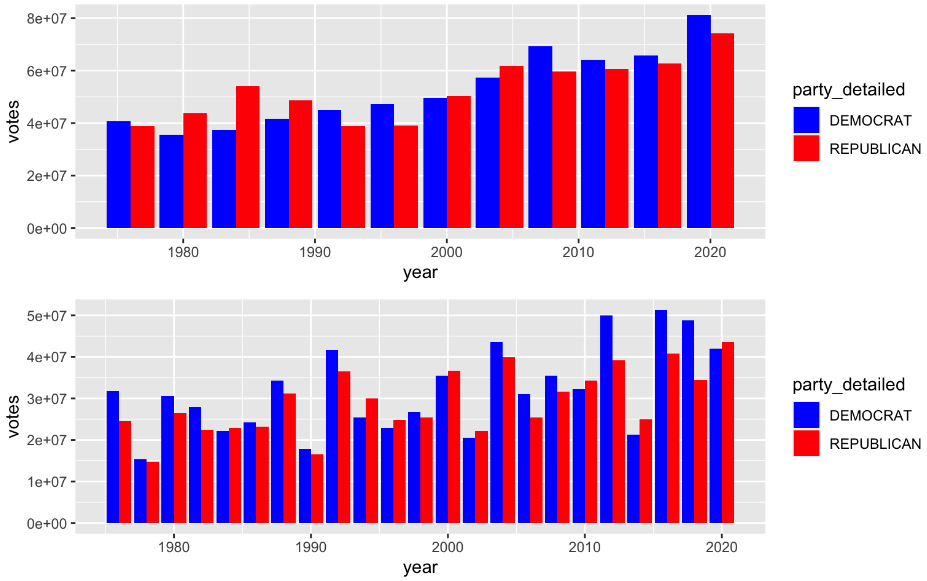

Figure 1: Voting count of presidential (above) / senate (below) elections (1976 to 2020)

3.2 Party System

3.2.1 India

From Figure 2 (website, India section), we can tell that the Indian National Congress (INC) being the dominant party until the 1990s. However, in recent years, the Bharatiya Janata Party (BJP) has emerged as a major contender. Notably, two parties often gained less than 50% of the popularity vote together. This implies India is not a two party system country.

Figure 3 (website, India section) is a cumulative line plot that displays the cumulative sum of voting count over different parties (sorted by number of votes by each party) in 1998. In this plot, the y-axis represents the cumulative sum of the proportion of votes, while the x-axis represents the rank of the party. The blue line is the cumulative sum, and the yellow line is the proportion contributed by a single party. We chose 1998 since INC and BNP both contributed to 25% of votings, which is somewhat representative of the whole situation.

We can see the pattern is that two parties eat up 50%. The third party only gained 5% of total votes, which implies the big difference between other parties and 2 major parties. From the third party to the 16th party, each of them gained 1% to 5% of the votings. We filtered out parties that gained less than 1% of the popularity votes for better visualization. From here, we can conclude that India is is a multi-party system with 2 major parties and a diverse political landscape consisting of other national and regional political parties (at least till 2014). Other medium size parties (with more than 1% popularity votes) though cannot make a huge difference to the national-wide situation, but can still be important in its region or some specific issues.

This gives us motivation to investigate in party coalition.

3.2.2 The USA

In the United States, there are a number of minor political parties that operate alongside the two major parties, the Democratic and Republican parties. These minor parties often represent specific ideologies or issues that are not fully addressed by the major parties. It is true argued that these parties generally have little chance of winning any elections. For example, in figure 2, we can tell it is hard for them to get even more than 1% population votes. Therefore, the Unites States is a typical two party system country.

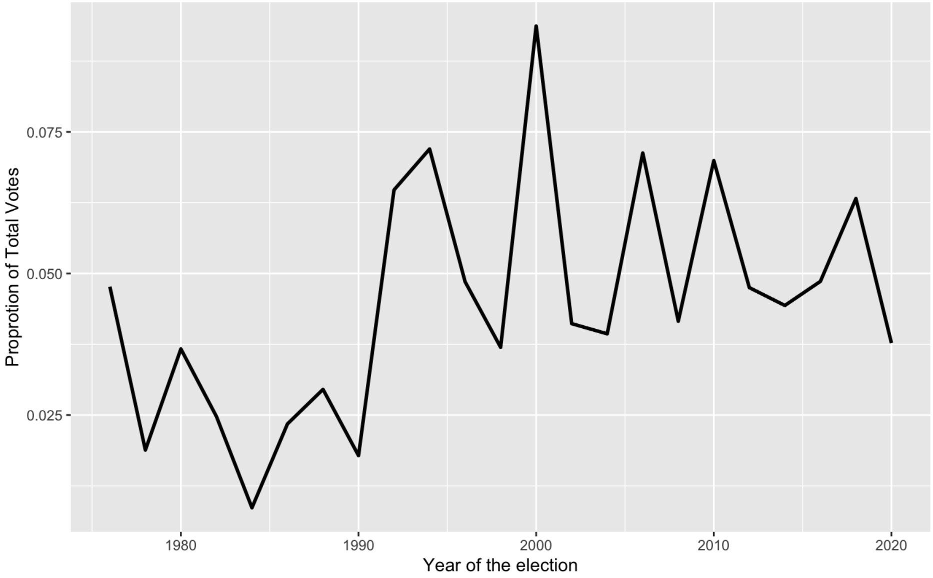

Figure 2: Proportion of Votes of Minor Parties in the Senator Election

3.3 Battle States, Safe States

Safe states are those where a particular party has consistently won elections over the years, and where the voters have a strong loyalty towards that party. On the other hand, battle states are those where there is no clear winner and where the political scenario is more dynamic, with voters being more likely to switch their allegiance between parties. We measure this using the number of turning points.

3.3.1 India

Figure 6 (website, Indian section) is a 2-dimenional histogram illustrated the distribution of number of turning points with two metrics.

Table 1 and 2 give us some candidates for battle and safe states. For example, Uttar Pradesh, the largest state as we mentioned before, is consistently changing their winning parties. In general, I would say the conception of ”battle state”, ”safe state” is not suitable for Indian situation. The electoral dynamics in India are constantly changing, the influence of national-level political parties in state elections is limited, there are many political coalition in the some states, our metrics of turning points may not even apply in this case (for example, 2nd and 3rd parties are coalition, together they beat the biggest party).

Table 1: Safe State Candidates

State	Number of Turning Points (Population)	Number of Turning Points (AC counts)
Gujarat	2	2
Kerala	0	2
Maharashtra	1	6
Meghalaya	1	2
West Bengal	6	1

Table 2: Battle State Candidates

State	Number of Turning Points (Population)	Number of Turning Points (AC counts)
Arunachal Pradesh	7	6
Haryana	3	8
Punjab	7	8
Rajasthan	6	8
Uttar Pradesh	7	5

3.3.2 The USA

Figure 3 illustrated distribution of number of turn-overs in each county in presidential election. The result is straightforward that most counties only voted for one party from 2000-2020 and had no turn-overs.

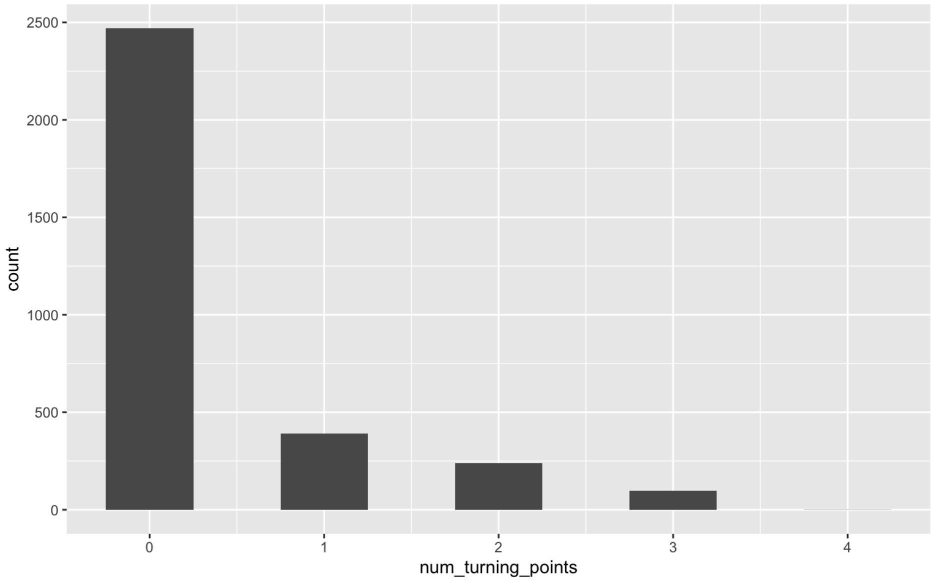

Figure 3: Distribution of number of turning points in each county in presidential election (2000-2020)

In figure 4, we illustrated distribution of number of turn-overs in each state in presidential election. We can say that most parties did not have any turn overs. Once a state has established a consistent voting pattern, it can be self-reinforcing, as voters tend to identify with a particular party and vote accordingly in future elections.

Among all these “safe states”, Iowa and Pinellas, Florida are two outliers in state level and county level.

3.4 Clustering Results

Some results we obtained are close to the real political spectrum (from manual data collection). Table 3 illustrated the results given group number equals 5.

From the table, we can see that group 1 included BJP and INC, although they are not in the political spectrum, they are too outstanding, and both of them could have larger amount of votes than other parties in most of states. Group 2 contains 2 left wing parties, group 3 contains 2 left parties. Group 4 and 5 had some wrong clustering (if the online source is correct). But in general, this result is good and can be interpreted with reasonable amount of information.

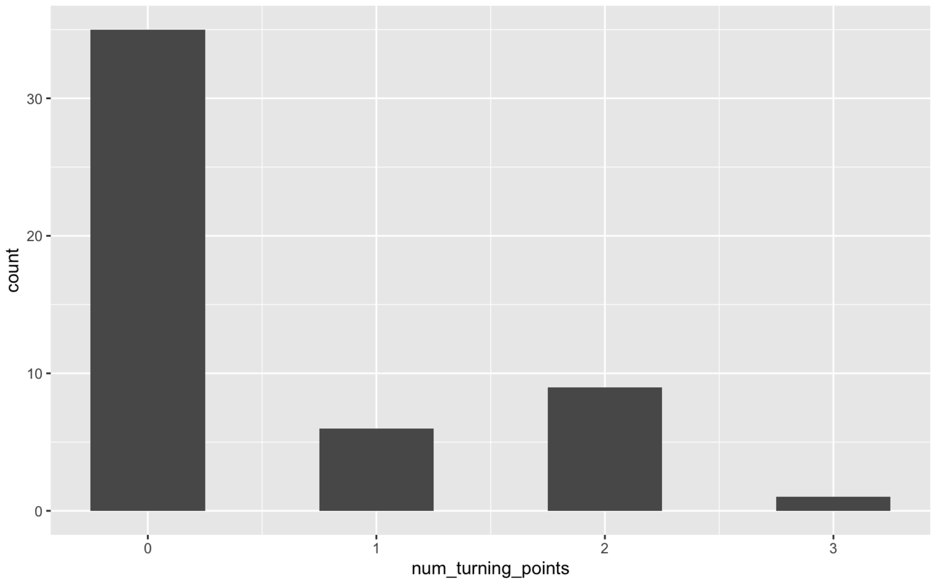

Figure 4: Distribution of number of turning points in each state in presidential election (2000-2020)

Table 3: Party Spectrum and Cluster Group

Party	**Spectrum (True)**	**Cluster Group (k=5)**
BJP	Right-wing	1
BSP	Centre-left	2
CPI	Far-left	3
CPM	Left-wing	3
INC	Centre-left	1
JD	Centre-right	4
LJP	Centre-right	4
NCP	Centre	5
RJD	Centre-left	4
SHS	Right-wing	5
SJP	Right-wing	4
SP	Left-wing	2

4 Conclusion and Summary

The United States of America and India are two of the largest democracies in the world, each with its own unique election system from. We compared and contrasted two systems based on data analysis from several aspects. The United States has a two-party system, while India has multiple parties. The United States has less voting population than India. The United States has more stable local voting pattern comparing to India. We also applied machine learning to strengthen the understanding of Indian party coalition.