Election Data Analysis using Tableau

Ajinkya Thakare, Akshay Jaiswal, Akshay Anil Pagar, Janhavi Dahihande, Koushik Kumar Kamala, Sai Harshith Reddy, Daniel Sampreeth Eadara, Madhusudhan Shagam
Computer Engineering Department
San José State University (SJSU)
San José, CA, USA
Email: {ajinkya.thakare, akshay.jaiswal, akshayanil.pagar, janhavi.dahihande, koushikkumar.kamala, saiharshithreddy.gaddam, danielsampreethreddy.eadara, madhusudhan.shagam, }@sjsu.edu

***Abstract*—Elections are imperative to democracy and let citizens make a contribution in the governance while voicing out their opinions. However, low voter turnouts impair these objectives of elections while also causing wastage of the government resources. It is important to determine the root cause of low voter turnouts and use the same to suggest plausible solutions. Business Intelligence and Analytical tools are available which help perform such studies and determine the key indicators for low voter turnouts. At the same time, visualization tools can be used to depict the data underhand in order to better understand the observations and derive important patterns in the data. Under this project, we aim to utilize such intelligent tools to analyze and visualize the available election datasets and find patterns which can aid in understanding the low voter turnout reasons and help in suggesting possible solutions to the same.**

***Index Terms*—US-elections, analysis, tableau, voter turnout, R, pyspark, map-reduce**

I. Introduction

Elections are important to the quality of a country's governance and can either drastically advance or set back the long-term democratic development of a nation, as well as the goals of the government, regional and global foreign policies. Elections also decide what is essential for the citizens and give them an opportunity to have their say and, through expressing partisanship, satisfy their need to feel a sense of belonging. Voting in elections is a fundamental right and responsibility of all citizens in a democratic nation.

One of the biggest problems faced by the US elections is the low voter turnout. Voter turnout in the US has been around only 40% in midterm elections and 60% in presidential elections. Such low turnouts can lead to unequal representation among various parts of the population and their voices, thus turning out to be one of the major threats for the elections and in turn, the democracy itself. The government spends a huge amount of money on various facilities provided to the public during elections. Having a good estimate on the voter turnout will help estimate these expenses. Solving this problem could drastically affect the tides of election giving power to underrepresented social groups and saving money for the government.

II. Problem Statement

Low voter turnout impairs the objectives of any election - to let the citizens make a fundamental contribution in the democratic governance of the nation and voice out their opinions for the nation's betterment. It is, however, important to understand the reasons for this low turnout and provide any plausible measures that can be taken to fix the issues.

We aim to utilize the business intelligence and analytical tools to understand these reasons that lead to low voter turnout in the elections. Images, figures, and charts are proven to be effective tools to explain and demonstrate important pieces of information. We thus aim to use data visualization software and tools that can aid to visualize and understand the data available. This will help us better understand the various trends derived in the voter turnouts and point us to any hidden patterns in the data that can be used to help understand the problems better. Using these patterns, we can also think of any efficient solutions that can help mitigate the problems. Luckily, election data is readily available and we can use it to perform our analysis and find solutions to the problems underhand.

III. Dataset

The dataset we are using for this project is acquired from MIT education and science lab. The MIT Election Data and Science Lab is dedicated to the nonpartisan application of scientific principles to election research and administration. The dataset contains non-normalized indicator scores for the Elections Performance Index (EPI). [1] The EPI discusses and explores various election policies and performance on U.S. Presidential and midterm elections of all states.[2] The dataset contains data from 2008 to 2016 in a 2 years interval period.

The dataset is built to help administrators or any person who wants to analyze elections and factors affecting elections. The dataset is intended to recognize potential problems and estimate impact of any changes in elections.

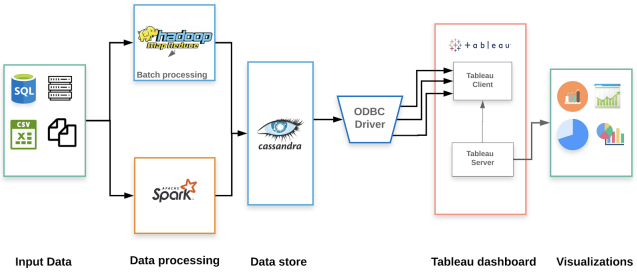
Some of the Election Performance Indicators (EPIs) are:

- state_abbv: State Abbreviation
- reg_rej: Registration rejection rate
- reg_rej: Registration rejection rate
- nonvoter_reg_pct: Percent non-voters because registration problem
- nonvoter_illness_pct: Percent non-voters because illness/disability
- wait: Wait time from SPAE

Here is a snippet from dataset:

```
state_abbv: "AL"
state_fips: "Alabama"
year: "2014"
website_pollingplace: "1"
website_reg_status: "1"
website_precinct_ballot: "0"
website_absentee_status: "1"
website_provisional_status: "1"
reg_rej: "0.017743733"
prov_partic: "0.001937722"
prov_rej_all: "0.000613034"
abs_rej_all_ballots: ""
abs_nonret: ""
uocava_rej: ""
uocava_nonret: "0.81782645"
eavs_completeness: "0.62521034"
post_election_audit: "0"
nonvoter_illness_pct: "0.16604702"
nonvoter_reg_pct: "0.06461484"
online_reg: "0"
wait: "3.71631"
residual: ""
pct_reg_of_vep_vrs: "0.80184573"
vep_turnout: "0.33212399"
```


IV. Project Architecture



Above diagram depicts the architecture for this project. The main components of the system are:

A. Input Data

Input data includes simple files provided by the MIT Elections lab and also Kaggle 2016 US election dataset which consists of democratic data on counties from the US census. This data is collected over a period of 18 years covering yearly elections.

B. Data Processing

Anticipating queries is an important part of Cassandra's storage model. We are using PySpark to generate and process these queries from the flat-files and CSVs available as datasets. Spark performs various operations using Hadoop's map-reduce to generate the final data to be stored in the Cassandra database. PySpark provides simple and comprehensive APIs with the advantage of easy code readability and maintenance.

C. Data Store

We have used Cassandra as our data store for this project. Given that the datasets used are flat-files or CSVs, a NoSQL database like Cassandra can rightly serve our purpose since there are little to no relationships between entities involved. Moreover, we can utilize the distributed nature of Cassandra to partition and replicate the data to support the scalability and availability of the data. We have thus used a 3 node Cassandra cluster, with the data replicated with a factor of 2 and partitioned horizontally using the state as a partition key.

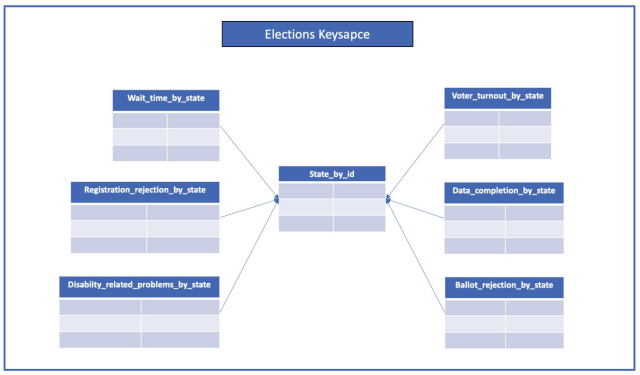
D. Tableau Dashboard

Tableau Desktop is a data visualization software that lets you see and understand the data in minutes. We can quickly build powerful calculations, drag and drop reference lines and forecast, and review statistical summaries. Tableau uses ODBC drivers to access the Cassandra database and can further be used to provide a number of visualizations.

E. Visualizations

Visualizations should tell a story using graphs and that's what we plan to do with various graphs like voter turnout trends, ballot rejection, and many more.

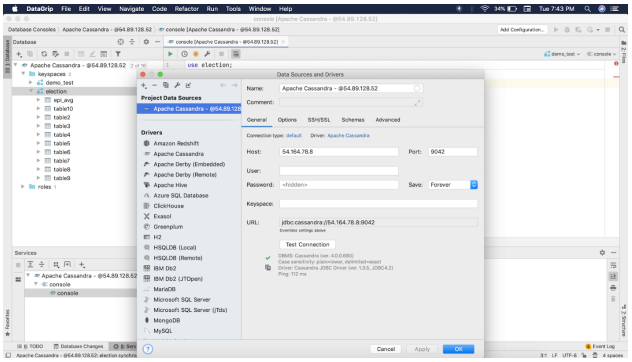
V. Database Design



Cassandra's keyspace is a namespace which focuses on the data replication on the nodes. Every node in the cluster contains a keyspace. For the Election data, we have created the Elections Keyspace. Each keyspace has different tables satisfying a particular query.

Since Cassandra follows a query-focused approach, we anticipated the queries needed for every visualization graph. To implement every query, we have different column families as will be discussed further. The MIT Election database provides us with various Key Performance Indicators. Every indicator will have a different column-family as shown in the Database design. For example, wait time with respect to state, voter turnout by state, registration rejection by state, data completion by state

- Testing Database connection using DataGrip:
  After the database setup, the next step is testing the database connection. For that, we have used a JetBrains IDE called DataGrip. In this, we can input our Hostname and port to test the connection. Also, after the connection is successful, we can see the various tables we have used in our visualizations. These tables will be used when we connect to Cassandra using Tableau.



VI. Data Analysis Approach

A. Relationship between disability vs turnout group by state

Disability in this context measures the number of people in a state who are unable to vote due to physical constraints or might face complexities while voting. The disability of a state is the total percentage of people who are disabled among the total number of effective voters from a state. The turnout of a state is the number of total people who turned up for the ballot among the total number of effective voters in that state. This analysis shows the turnover of the total disabled people among all the people who turned up to vote.

B. Online registration available vs Voter registration rate

Online registration availability of a state is an indication whether a particular state offers the ability to vote online or offer only updates or its only offline. The voter registration rate is the number of eligible people per state who actually report themselves to actively participate in the voting process. This analysis portrays the number of registrations made depending on the availability of online registration. Since online registration makes the process easier, the expected turnover is high in online registration enabled states.

C. Voting wait time vs voter registration rate

Voting wait time is the measure of average waiting time a voter has to endure in line to cast his vote or deposit a ballot already mailed to them. Wait time plays an important role in voting as less waiting time portrays the ease of voting. The analysis to be made is how the average voting wait time impacts the voter registration rate and in turn, how the increase in voter registration rate implies an upward trend in voter wait time.

D. Voter registration rate vs turnout

This analysis exhibits the relationship between the number of people who registered to vote and the number of people who actually turned up to vote. This analysis helps in studying factors such as the impact of campaigning, voter interest, demographics, and even the effect of weather.

VII. Visualization Design

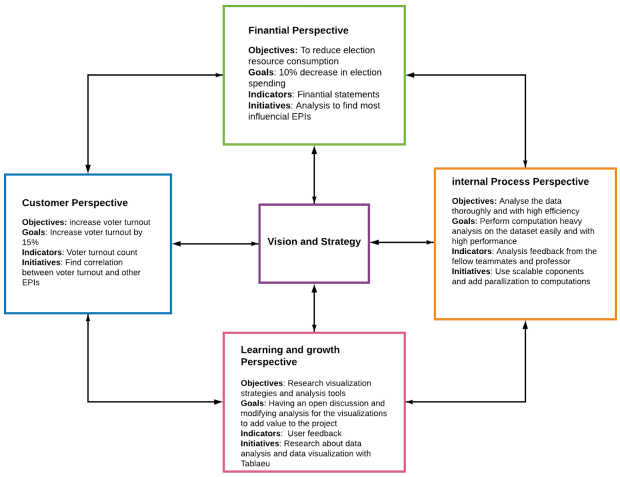
Since R is a powerful open-source library available for visualisations, and Tableau is a user-friendly software, we try to implement both. To achieve this, we used the tab admin utility, that connects to an instance of Rserve, and thereby makes the Rserve package available to the Tableau. This enables a great visualisations with deep statistical analysis, added with Tableau features like drag-and-drop.

We used tableau as the data visualization software in our aim to visualize and understand the data available. This helped us in understanding various trends derived in the voter turnouts and pointed us to any hidden patterns in the data that can be used to help understand the problems better.

The following graph is a combination of various graphs including the state-wise indicator average.

- The state-wise indicator average measures the performance index of the states in comparison to each other. Higher the index, better the performance.
- The registration vs turnout graph represents the state of Louisiana. It shows the registration, online registration rate, and the turnout rate per year from 2007 to 2017.
- The state turnout vs non-voter registration graph shows the inverse rate of the vep value between the turnout and the non-voter registration.
- The state turnout vs non-voters illness graph similarly shows that the turnout is inversely proportional to the Illness rate for each state.
- The state turnout vs the nonvoters - both graphs include the non-voters' illness and their registration graph plotted against the total turnout vep value. This too followed an inverse trend between the attributes.

VIII. Business Performance Metrics



The Business Performance Metrics are divided into 4 perspectives:

A. Financial Perspective

The objective of the Financial perspective is to reduce election resource consumption. The goal is to decrease the election cost by 10%. Financial statements work as the indicators for this perspective. One of the initiatives to be taken up is to analyze and find the most influential EPIs.

B. Internal Process Perspective

This perspective analyzes the data thoroughly with high efficiency. This perspective helps in performing heavy computation analysis on the dataset with ease. The analysis feedback from teammates and our Professor are used as the indicators for this perspective. We use scalable components and add parallelization to all the computations.

C. Learning and Growth Perspective

The objective of this perspective is to research the visualization strategies and analysis tools. The goal is to have an open discussion and modify the analysis for the visualizations to add value to the project The main indicator for this perspective is user feedback. The various initiatives under this perspective include research about data analysis and data visualization with Tableau.

D. Customer Perspective

The customer perspective objective is to increase the voter turnout and its goal is to increase voter turnout by 15%. The voter turnout count acts as the indicator for this perspective. The various initiatives taken under this perspective are finding correlation between voter turnout and other EPIs.

IX. Challenges

A. Finding key performance indicators

- There are a lot of factors that affect the election process. It is easy to get lost in the data and not finding anything helpful.
- We performed various manipulations and plotted multiple plots in order to find some pattern in the data and various indicators.
- We learned many new data analysis techniques for identifying these key performance indicators.

B. Finding correlation between indicators

- Each indicator represents one aspect of the election process and trying to find the correlation between these aspects needs a deep understanding of these indicators.
- Once we were familiar with what these indicators represent, we had to figure out the relationships between each of these indicators which was a daunting task with more than 15 indicators at play.
- We used python's pyplot library to plot these indicators and looking for some correlation between them.
- We also calculated correlation between each of these indicators using libraries like PySpark and map-reduce jobs in Hadoop.

C. Performing computation heavy tasks on the dataset

- Computation heavy tasks like finding correlation between key performance indicators take a lot of time and computation power.
- We used PySpark to use RDDs to store the data and Hadoop to run the Map-Reduce jobs for manipulation.

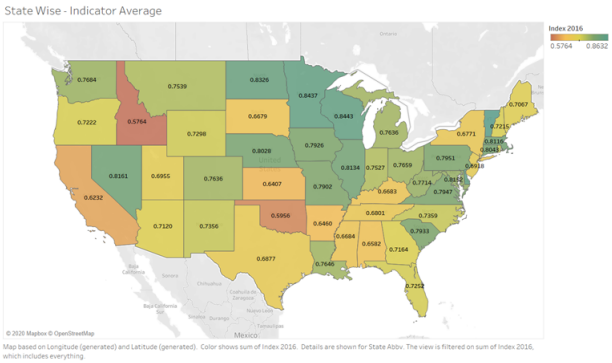
D. Using Cassandra with Apache PySpark and Hadoop

- Learning Cassandra, PySpark and Hadoop was a new experience for us.
- Using Cassandra along with PySpark and Hadoop proved to be a challenge since there are a lot of moving pieces and using them efficiently becomes a challenge.

X. Results

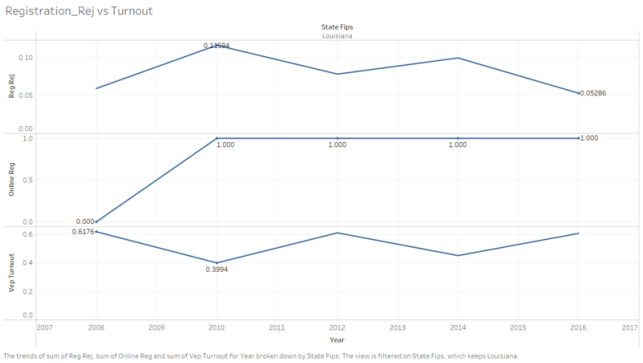
A. State wise Indicator Average Performance

Since the election administration is complex and involves so many activities. In order to find how particular states perform, both in isolation and in comparison with one another. So we came up with this visualization that used the EPI to combine the information from the possible indicators to produce a summary measure of the performance of the elections. So this visualization explains how each state performs.



B. Registration Rejection vs Turnout

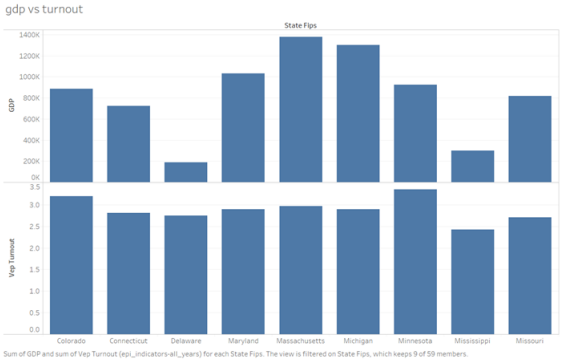
The increase in online registration has a decrease in the registration rejection rate which in turn is inversely proportional to the vep turnout rate. But there have been registration problems since the inclusion of the online registration problems and those people are rejected. If the problems can be rectified while registering, higher turnout can be expected.



C. Average GDP vs Turnout

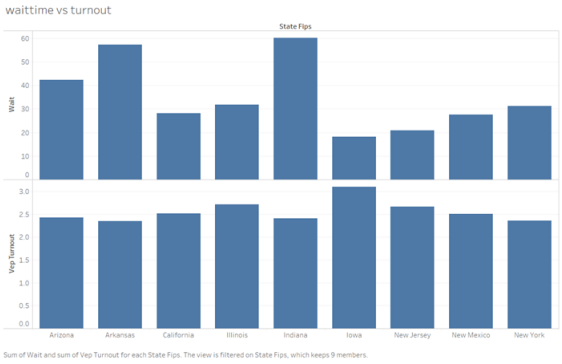
The average gdp has an effect on the turnout. It can be seen that the increase in the average gdp of an area is directly proportional to the vep turnout rate.[3] From the graph, we can infer that states with high economies or people who are rich tend to vote more. From this view, vote turnout can be

increased by concentrating on states with low economy and exploring more surveys in those areas.



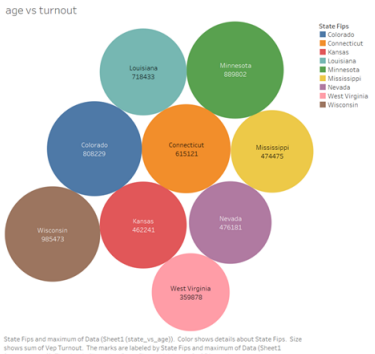
D. Wait time vs Turnout

The wait time impacts the turnout greatly. The delay in casting the vote despairs the voters and they restrain from casting their votes. The states with higher wait times experienced a serious decrease in the turnout while the states with quicker process has seen an upward trend.



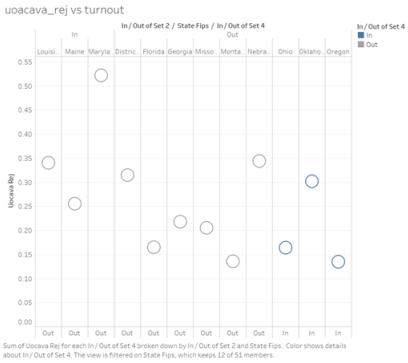
E. Age vs Turnout

There is a requirement for spreading awareness among the younger generation to cast their vote. The states with higher average populations experienced a higher turnout rate. Older people tend to vote more while the younger generation need to be motivated.



F. UOACAVA Rejection vs Turn out

The parameter uoacava_rej refers to the mail ballot rejected. These came in place when people are unable to exercise their votes because of unavailability on the dedicated election days.[4] This can also happen due to multiple reasons. Once these mail ballots are exercised, they may be rejected for various reasons. Like since the ballot requires the signature on it, some of the ballots were not at all signed, while even though signed but an improper one. This results in a decrease in the turnout.

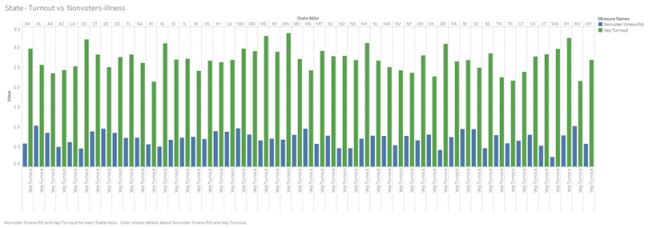


G. Website Precinct Ballot Vs Prov_rej

Generally, Website Precinct Ballot is indicator, where the states provides an option to locate their polling place, so that people can lookup and do cast their vote. Provisional rejections are the holding votes which are mostly because of wrong precinct ballots. So, we can infer from the graph, providing the ballot information reduced provisional rejections which increased turnout. So, providing ballot locations increases turnout.

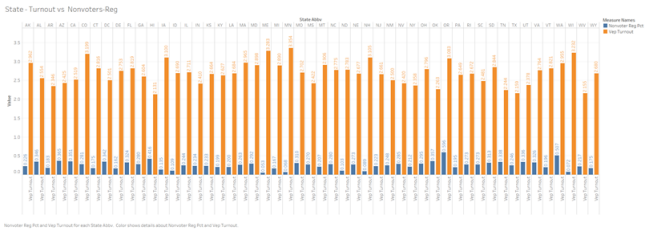
H. Non-Voter Illness Problems vs Turnout

For years, it has been of great concern for the physically disabled to cast their votes. This impacted the rate of turnout as those with disabilities are lower than those without disabilities. We can understand this from the below graph. It also explains that a person with disabilities or illness are more prone to not casting the vote than a normal person.



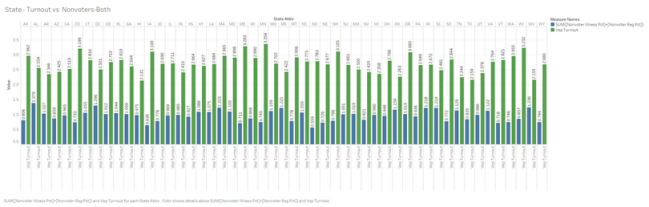
I. Non Voter Registration Problems vs Turnout

In this experiment, we focused on the registration impacts over turnout. In most of the states, a voter needs to be registered with the government to exercise his/her vote. So, we experimented between voters who are registered to those who are not registered. It has been clearly depicted from the visualization that the states without any pre-requisite of registration have high percent of exercising election. Like, if we see, North Dakota has removed the user registration requirement and thus the turnout for these states has increased greatly.



J. Non Voter Problems vs Turnout

The non voter has a variety of reasons for not casting their votes. These problems like illness, registration to cast the vote and more, have a serious effect on the turnout rate. As it can be observed in the graph, the rate of the non voter registration is inversely proportional to the turnout. Each state except North Dakota needs it's voters to register before casting their vote. But the increase in the non registration has affected the voting turnout.



XI. Conclusions

With this project, we were able to narrow down the factors which are influential towards voter turnout in US elections. We found that Wait time affects the voter turnout negatively. People don't like to wait too long to cast a vote. Voter registration rejections due to mistakes in the forms submitted occupy a significant chunk of the reasoning behind low voter turnout. We can help voters make fewer mistakes by educating them on the process. GDP of an area also affects the voter turnout such that states with higher overall GDP tend to vote in more consistently higher numbers. This speaks to the fact that lower-income areas are being neglected. Voters with ages 65 and above hold the majority in the voter population. This means voters aged below 65 are failing to vote and one of the reasons being the elections being held on Weekdays where the working population can't vote due to being on the job. Using this information, we can confidently say that the consumer will be able to take meaning action towards solving this problem.

References

[1] E. P. Index, "See http://epi. yale. edu/epi2012/rankings," *Note several international comparisons at this site*, 2012.

[2] U. E. A. Commission *et al.*, "2010 election administration and voting survey," Tech. Rep., 2011.

[3] B. C. Burden and C. Stewart III, *The Measure of American Elections*. Cambridge University Press, 2014.

[4] L. Schur, M. Adya, and M. Ameri, "Accessible democracy: reducing voting obstacles for people with disabilities," *Election Law Journal*, vol. 14, no. 1, pp. 60–65, 2015.