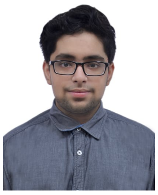
Exploratory Data Analysis on Election

PROJECT

by

Group 11

Anurag Choudhury
ID: 202318059
Course: Msc.Data Science



Aditya Tripathi
ID: 202318046
Course: Msc.Data Science

Karan Sharma
ID: 202318018
Course: Msc.Data Science

Course Code: IT 462
Semester: Winter 2024

Under the guidance of

Dr. Gopinath Panda

Dhirubhai Ambani Institute of Information and Communication Technology

May 2, 2024

Acknowledgment

I wanted to take a moment to express my sincere appreciation for the exceptional guidance and support you provided me during my project ”Election” Your mentorship has been invaluable and played a pivotal role in ensuring the success of this endeavor.

I consider myself very lucky to have had the opportunity to benefit from your expertise and mentorship. Your insightful advice, coupled with extensive knowledge in the field, has significantly influenced the quality and scope of the project. Your constructive feedbacks and suggestions not only helped me navigate challenges but deepened my understanding of the subject.

I would also like to show my gratitude to the entire team at DAIICT for fostering a culture of collaboration and innovation. The resources and facilities provided by the institution have been instrumental in facilitating comprehensive research and analysis, thereby enriching the outcome of the project.

Moreover, I am thankful to my peers and colleagues for their unwavering support and camaraderie throughout this journey. Their contributions have undeniably enhanced the development of the ”Election” project.

Entering into the ”Election” project has been an immensely fulfilling experience for me. I am confident that the knowledge and skills acquired during this endeavor will serve as a solid foundation for my future pursuits.

Once again, I want to express my deepest gratitude for your invaluable guidance and support. Your mentorship was indispensable, and I am genuinely appreciative of the opportunity to learn from you.

Sincerely,

Anurag Choudhury, 202318059
Aditya Tripathi, 202318046
Karan Sharma, 202318018

DECLARATION

We, [202318059, 202318046, 202318018] now declare that the EDA project work presented in this report is our original work and has not been submitted for any other academic degree. All the sources cited in this report have been appropriately referenced.

We acknowledge that the data utilized in this project has been sourced from https://www.eci.gov.in/ site. We affirm that we have complied with the terms and conditions specified on the website for accessing and using the dataset. We hereby confirm that the dataset employed in this project is accurate and authentic to the best of our knowledge.

We acknowledge that we have received no external help or assistance in conducting this project except for the guidance provided by our mentor, Prof. Gopinath Panda. We declare no conflict of interest in conducting this EDA project.

We have now signed the declaration statement and confirmed the submission of this report on 29 April 2024.

Anurag

Anurag Choudhury
ID: 202318059
Course: Msc.Data Science

Aditya

Aditya Tripathi
ID: 202318046
Course: Msc.Data Science

Karan

Karan Sharma
ID: 202318018
Course: Msc.Data Science

CERTIFICATE

This is to certify that Group 11 comprising Anurag Choudhury, Aditya Tripathi and Karan Sharma has completed an exploratory data analysis (EDA) project on the PROJECT, which was obtained from ECI site.

The EDA project presented by Group 11 is their original work. It was completed under the guidance of the course instructor, Prof. Gopinath Panda, who provided support and guidance throughout the project. The project is based on a thorough analysis of the PROJECT dataset, and the results presented in the report are based on the data obtained from the dataset.

This certificate is issued to recognize the successful completion of the EDA project on the PROJECT, which demonstrates the analytical skills and knowledge of the students of Group 11 in the field of data analysis.

Signed,
Dr. Gopinath Panda,
IT 462 Course Instructor
Dhirubhai Ambani Institute of Information and Communication Technology
Gandhinagar, Gujarat, INDIA.

May 2, 2024

Contents

List of Figures

List of Tables

ABSTRACT

This project conducts a comprehensive analysis of election data spanning the years 2008 to 2023, aiming to uncover trends, patterns, and insights crucial for understanding electoral dynamics over this period. Leveraging a diverse range of data sources, including voter turnout, candidate demographics, electoral outcomes, and campaign finance, the study employs advanced statistical techniques and data visualization methods to elucidate key phenomena and drivers shaping electoral processes.

The analysis encompasses a multi-dimensional examination of electoral landscapes, encompassing national, regional, and local elections across various jurisdictions. Through meticulous data preprocessing, cleansing, and normalization, the project ensures data integrity and consistency, laying a solid foundation for subsequent analyses.

Key focal points include the evolution of voter participation and engagement, shifts in voting patterns and geographical locations over the years, the emergence of polling stations and constituencies in an area, and the impact of gender on electoral outcomes. Additionally, the project delves into the pattern observing and gaining insights about election data.

Through rigorous statistical modeling , the project seeks to identify significant correlations, causal relationships, and predictive insights within the election data. Moreover, by employing machine learning algorithms and predictive analytics, the study endeavors to forecast future electoral trends and anticipate potential shifts in political dynamics.

The findings of this project hold implications for policymakers, political analysts, and electoral stakeholders, offering valuable insights into the complex interplay of factors shaping democratic processes. Ultimately, this research contributes to a deeper understanding of electoral phenomena, distribution and vote share among different genders in all over the country in different years in an ever-evolving political landscape.

Chapter 1. Introduction

1.1 Project idea

The objective of this project is to conduct an Exploratory Data Analysis (EDA) on state wise election from the year 2008 to 2023. Through this analysis, we aim to explore and understand the voting patterns and trends prevalent within the elections.

By employing exploratory data analysis techniques, we seek to uncover insights into the state elections in different years. This involves examining the distribution of vote shares, identifying different patterns related to voting i.e no. of seats in each constituencies or sharing of votes among different genders.

Understanding voting trends and patterns is crucial for comprehending the dynamics of elections. By analyzing historical data and current trends, we can gain insights into voter behavior, preferences, and the factors influencing electoral outcomes. This report aims to explore various trends and patterns observed in recent elections, drawing on data analysis and statistical methods.

1.2 Data Collection

Data Sources

To analyze the trends and patterns in voting and vote share in each state and in a year we obtain a dataset from Election Commission of India where in the dataset there are ideas about the voting shares in different places by different genders.

By leveraging these datasets, we can uncover a wealth of information that will enable us to understand and interpret the number of voters, number of constituencies and total seats based on geographical locations.

In the subsequent discussion, we'll explore the insights gleaned from these datasets, examining trends, patterns, and correlations to gain a comprehensive understanding of the election scenarios all over the country.

1.3 Dataset Description

Dataset: The data source for this project is a dataset from Election Commission of India. It consists of 100 rows and 33 columns

Features:

- **State / UT** Names of states and the union territories
- **Year of GE to SLA** Year in which state elections happened
- **Reserved SC Seats** Number of seats reserved for Schedule casts
- **Reserved ST Seats** Number of seats reserved for Schedule tribes.
- **Total No. of Seats** Total number of seats including all casts.
- **No. Of Male Contestants:** Male contestants who are going to participate in election.
- **No. Of Female Contestants** Female contestants who are going to participate in election.
- **No. Of Transgender Contestants** Transgender contestants who are going to participate in election.
- **Total number of contestants** Total number of contestants including all genders.
- **Elected Male Candidates** Male contestants who won
- **Elected Female Candidates** Female contestants who won
- **Elected Transgender Candidates** Transgender contestants who won
- **Total number of Candidates** Total number of winners
- **Elected Female percs** Percentage of elected female candidates
- **Forfeited Deposits Male** number of male candidates who forfeited there security deposit.
- **Forfeited Deposits Female** number of female candidates who forfeited there security deposit.
- **Forfeited Deposits Transgender** number of transgender candidates who forfeited there security deposit.
- **Total Deposits** Total forfeited deposits.
- **Average Contestants Per Constituency** Average number of Contestants Per Constituency in a state.
- **Male Electors** Number of Males who are eligible to vote
- **Female Electors** Number of females who are eligible to vote
- **Total Electors** Total number of people who are eligible to vote.

- **Male Electors Who Voted** Number of Males who voted in a city in a year
- **Female Electors Who Voted** Number of Females who voted in a city in a year
- **Total Votes Polled** Total number of votes counted
- **Male Polling Percentage Excluding Postal Ballots** Percentage of males who have given vote and postal votes are not counted
- **Female Polling Percentage Excluding Postal Ballots** Percentage of females who have given vote and postal votes are not counted
- **State Poll perc** Poll percentage in a state in a year
- **No of Polling Stations** Number of Polling stations in a city in a year
- **Average Number of Electors per PS** Number of people who are eligible to vote per Polling Stations
- **Postal Votes** Total number postal votes
- **Rejected Postal Votes** Postal votes which got rejected
- **NOTA** NOTA stands for "None of the Above." It's an option provided on ballot papers or electronic voting machines.

1.4 Packages required

Pandas (pd)

- **Why it's Required:** Pandas is essential for this project because it provides a robust framework for handling structured data, which is crucial for analyzing the monthly food price inflation dataset. Its functionalities streamline data preprocessing tasks such as cleaning, transforming, and aggregating, ensuring the dataset is ready for in-depth analysis.
- **Uses for the Report:** Pandas enables efficient loading of the election dataset into a DataFrame, providing powerful tools for indexing, slicing, and filtering data to extract relevant information for analysis. Pandas facilitates the exploration of key statistics and trends within the dataset through functions like describe(), mean(), median(), and value counts(), allowing for a comprehensive understanding of election patterns across different places at different times

Matplotlib (plt)

- **Why it's Required:** Matplotlib is indispensable for this project as it provides extensive functionality for creating various types of plots and visualizations, which are essential for exploring trends, patterns, and relationships within the monthly food price inflation dataset. Its versatility allows for the generation of static, interactive, and publication-quality visualizations, aiding in the interpretation and communication of findings effectively.

- **Uses for the Report:** Matplotlib enables the creation of diverse visualizations such as line plots, bar charts, histograms, and scatter plots to depict voting trends over time, compare voting percentage among genders across states, and analyze correlations with other election indicators.Matplotlib, along with tools like Basemap or Cartopy, can be used to visualize geographical variations in number of seats, candidates and polling percentage across Indian States

numpy:(np)

- **Why it's Required:** Numpy serves as a fundamental library for numerical computing in Python, offering efficient data structures and operations for handling arrays, matrices, and mathematical functions. Its seamless integration with Matplotlib and other scientific libraries makes it indispensable for data analysis tasks.
- **Uses for the Report:** Numpy plays a crucial role in data preprocessing, manipulation, and computation within the report. It enables efficient handling of large datasets, mathematical operations on arrays, and transformation of data for visualization and analysis. Numpy's array slicing, broadcasting, and vectorized operations streamline data processing workflows, enhancing the efficiency and accuracy of analytical tasks in the report.

scikit-learn:(sklearn)

- **Why it's Required:** Scikit-learn is a versatile machine learning library that provides a wide array of tools for data mining, analysis, and modeling. Its user-friendly interface and extensive range of algorithms make it invaluable for building predictive models, conducting feature selection, and evaluating model performance.
- **Uses for the Report:** Scikit-learn facilitates the implementation of machine learning algorithms for predictive modeling and pattern recognition tasks in the report. It enables the exploration of complex relationships within the election data, the development of predictive models for forecasting trends, and the evaluation of model accuracy using cross-validation techniques. Scikit-learn's comprehensive documentation and robust implementation of machine learning algorithms empower users to leverage advanced analytical capabilities in the report.

missingno:(msno)

- **Why it's Required:** Missingno is a Python library designed for visualizing and analyzing missing data within datasets. It provides intuitive visualizations that highlight missing values, enabling users to identify patterns of missingness and make informed decisions about data imputation and cleansing.
- **Uses for the Report:** Missingno enhances the data quality assessment process by visually identifying missing values and patterns within the election dataset. Its visualizations aid in understanding the extent and distribution of missing data, guiding data preprocessing steps such as

imputation or removal of incomplete records. By integrating missingno into the analysis workflow, the report ensures robust handling of missing data and maintains data integrity throughout the analytical process.

seaborn:(sns)

- **Why it's Required:** Seaborn is a powerful data visualization library built on top of Matplotlib, offering high-level functions for creating informative and visually appealing statistical graphics. It simplifies the creation of complex visualizations such as heatmaps, scatter plots, and distribution plots, making it ideal for exploratory data analysis and presentation.
- **Uses for the Report:** Seaborn enriches the election data analysis in the report by providing aesthetically pleasing and informative visualizations. Its intuitive syntax and built-in statistical functionalities enable users to generate insightful plots that showcase relationships, trends, and distributions within the data. By leveraging Seaborn's capabilities, the report enhances the interpretability and communicability of key findings, facilitating clearer insights into employment trends and patterns.

scipy:

- **Why it's Required:** Scipy is a comprehensive library for scientific computing in Python, offering a wide range of modules for numerical optimization, integration, interpolation, and statistical analysis. It complements Numpy by providing additional mathematical functions and algorithms for scientific computing tasks.
- **Uses for the Report:** Scipy augments the analytical capabilities of the report by providing advanced statistical functions and algorithms for data analysis. Its modules for hypothesis testing, regression analysis, and probability distributions enable rigorous statistical inference and hypothesis testing within the employment data analysis. By leveraging Scipy's rich functionality, the report ensures robust statistical analysis and sound interpretation of results, contributing to the credibility and reliability of the findings presented.

Chapter 2. Data Cleaning

After conducting a thorough analysis of our dataset, we identified missing values in four key columns:

1. No. Of Transgender Contestants
2. Transgender Candidates
3. Deposits Transgender
4. NOTA (None of the Above)

Rather than opting to remove the affected rows, we have chosen a more robust approach: imputing these missing values with the median. This decision was made to ensure that our dataset remains comprehensive and that we maintain the integrity of the data.
Since the distribution is not normal and skewed in the dataset, we have used median imputer
Imputing missing values with the mean is a prudent strategy, particularly for numerical columns like these. By doing so, we can effectively fill in the gaps in our dataset with values that are representative of the overall distribution, thereby minimizing the impact of missing data on our analyses.

This approach not only preserves the size and structure of our dataset but also ensures that any subsequent analyses or modeling efforts are based on a more complete and reliable set of data. By imputing missing values with the median, we aim to derive more accurate insights and make informed decisions based on the most representative data available.

2.1 Missing data analysis

We utilized a combination of bar plots and a matrix visualization to analyze the missing values within our dataset. The first three diagrams specifically highlight the presence of missing values across four key columns Additionally, the second diagram, a matrix representation, offers a comprehensive overview of missing values distribution across the dataset.

Upon careful examination, it became evident that the column labeled 'NOTA' exhibited the highest count of missing values compared to the other three columns.

To facilitate this analysis, we employed the 'missingno' package, which provided us with intuitive visualizations for assessing the extent and distribution of missing data within our dataset. By leveraging these visualizations, we were able to gain valuable insights into the patterns of missingness, enabling us to make informed decisions regarding our data imputation strategy.

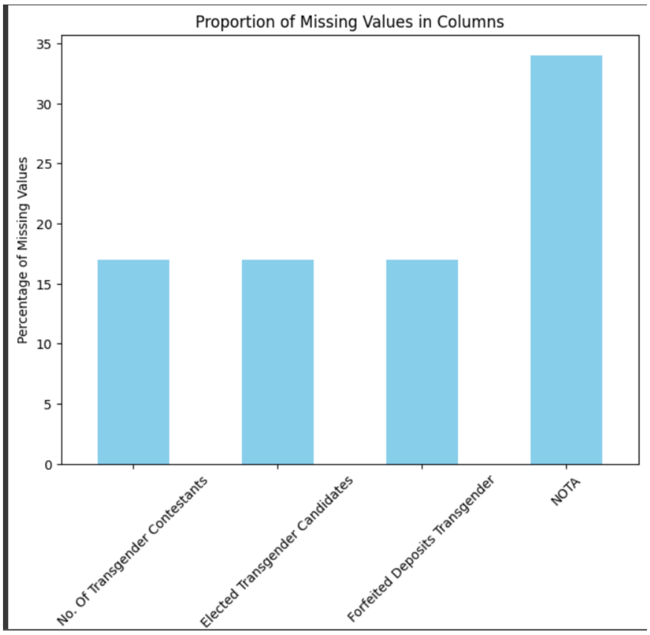


Figure 2.1: Proportion of missing values

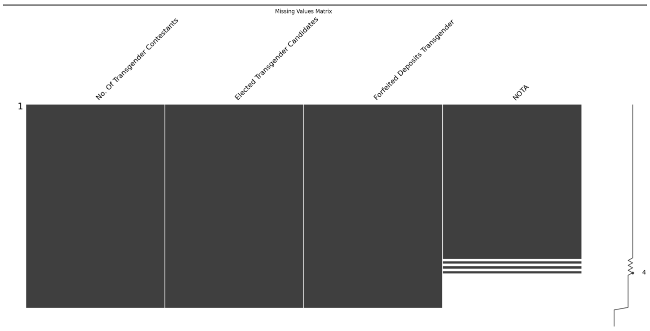


Figure 2.2: Missing Values Matrix

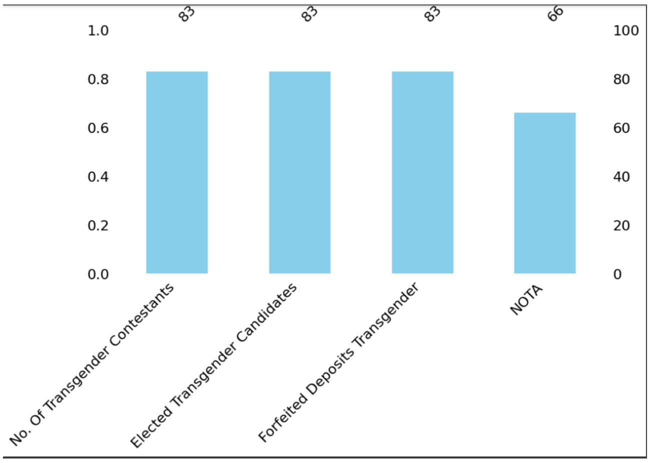


Figure 2.3: Missing Values Bar Plot

2.2 Imputation

In addressing the missing values within our dataset, we adopted the SimpleImputer module from scikit-learn. This module offers a straightforward method for handling missing data by replacing NaN (Not a Number) values with a specified statistic, such as the median.

Specifically, we applied the SimpleImputer to the four target columns:

1. No. Of Transgender Contestants
2. Elected Transgender Candidates
3. Forfeited Deposits Transgender
4. NOTA (None of the Above)

By utilizing the median as the imputation strategy, we aimed to minimize the influence of outliers and skewed distributions that may be present in the data.

Following the imputation process, the resulting imputed values are presented in the list of tables. These tables provide a clear overview of the imputed values for each column, allowing for easy reference and verification.

The choice of median as the imputation strategy ensures robustness against extreme values, and is reliable in a skewed distribution making it a reliable approach for handling missing data in numerical columns. This method enables us to maintain the integrity of the dataset while ensuring that the imputed values accurately reflect the central tendency of the respective columns.

Chapter 3. Visualization

We have done a brief analysis and studied about the different votting patterns and vote shares and analyzing other parameters affecting it by plotting different graphs

3.1 Univariate analysis

The bar graphs (3.1-3.4) in blue give the idea about the female vote share across different states in the years 2008, 2013, 2018, 2023. some union territories did not voted before and for the remaining states the pattern is almost identical.
The bar graph in green (3.5) shows the total number of seats in each year combing by different states. In 2010, 2015, 2020 the number of seats overall were less. The box plot (3.7) gives the number of elected male and elected females. From the plot we can say that there are less number of female candidates who got elected.
The line chart (3.6) gives the total votes polled in a year combining different states in a particular year. The pattern is similar to the number of seats in a year. As the number of seats were less the no. of polled votes were also less.

In fig 3.11, the blue represents the accepted postal votes and the red represents the rejected postal votes over the years in terms of 5 year difference.

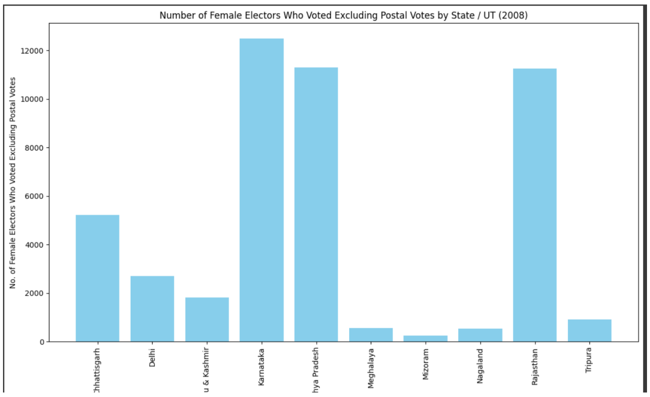


Figure 3.1: No. of Female Electors Who Voted in 2008

Table 3.1: Description of Columns

Column	Non-Null Count	Dtype
State / UT	100	object
Year of GE to SLA	100	int64
Reserved SC Seats	100	int64
Reserved ST Seats	100	int64
Total No. of Seats	100	int64
No. Of Male Contestants	100	int64
No. Of Female Contestants	100	int64
No. Of Transgender Contestants	83	float64
Total number of contestents	100	int64
Elected Male Candidates	100	int64
Elected Female Candidates	100	int64
Elected Transgender Candidates	83	float64
Total number of Candidates	100	int64
Elected Female perc	100	float64
Forfeited Deposits Male	100	int64
Forfeited Deposits Female	100	int64
Forfeited Deposits Transgender	83	float64
Total Deposits	100	int64
Average Contestants Per Constituency	100	object
Male Electors	100	float64
Female Electors	100	float64
Total Electors	100	float64
Male Electors Who Voted	100	float64
Female Electors Who Voted	100	float64
Total Votes Polled	100	float64
Male Polling Percentage Excluding Postal Ballots	100	float64
Female Polling Percentage Excluding Postal Ballots	100	float64
State Poll perc	100	float64
No of Polling Stations	100	int64
Average Number of Electors per PS	100	int64
Postal Votes	100	int64
Rejected Postal Votes	100	int64
NOTA	66	float64

Table 3.2: Description of Columns (with Missing Values)

Column	Non-Null Count	Dtype
State / UT	0	
Year of GE to SLA	0	
Reserved SC Seats	0	
Reserved ST Seats	0	
Total No. of Seats	0	
No. Of Male Contestants	0	
No. Of Female Contestants	0	
No. Of Transgender Contestants	17	float64
Total number of contestents	0	
Elected Male Candidates	0	
Elected Female Candidates	0	
Elected Transgender Candidates	17	float64
Total number of Candidates	0	
Elected Female perc	0	
Forfeited Deposits Male	0	
Forfeited Deposits Female	0	
Forfeited Deposits Transgender	17	float64
Total Deposits	0	
Average Contestants Per Constituency	0	
Male Electors	0	
Female Electors	0	
Total Electors	0	
Male Electors Who Voted	0	
Female Electors Who Voted	0	
Total Votes Polled	0	
Male Polling Percentage Excluding Postal Ballots	0	
Female Polling Percentage Excluding Postal Ballots	0	
State Poll perc	0	
No of Polling Stations	0	
Average Number of Electors per PS	0	
Postal Votes	0	
Rejected Postal Votes	0	
NOTA	34	float64

Table 3.3: Description of Columns (without Missing Values)

Column	Non-Null Count	Dtype
State / UT	0	
Year of GE to SLA	0	
Reserved SC Seats	0	
Reserved ST Seats	0	
Total No. of Seats	0	
No. Of Male Contestants	0	
No. Of Female Contestants	0	
No. Of Transgender Contestants	0	
Total number of contestents	0	
Elected Male Candidates	0	
Elected Female Candidates	0	
Elected Transgender Candidates	0	
Total number of Candidates	0	
Elected Female perc	0	
Forfeited Deposits Male	0	
Forfeited Deposits Female	0	
Forfeited Deposits Transgender	0	
Total Deposits	0	
Average Contestants Per Constituency	0	
Male Electors	0	
Female Electors	0	
Total Electors	0	
Male Electors Who Voted	0	
Female Electors Who Voted	0	
Total Votes Polled	0	
Male Polling Percentage Excluding Postal Ballots	0	
Female Polling Percentage Excluding Postal Ballots	0	
State Poll perc	0	
No of Polling Stations	0	
Average Number of Electors per PS	0	
Postal Votes	0	
Rejected Postal Votes	0	
NOTA	0	

Table 3.4: Statistical Summary for the ”No of Polling Stations” Column

Statistic	**Value**
Count	100.000000
Mean	31906.670000
Standard Deviation	34901.873059
Minimum	545.000000
25th Percentile	2924.500000
50th Percentile (Median)	21723.500000
75th Percentile	51870.500000
Maximum	174803.000000

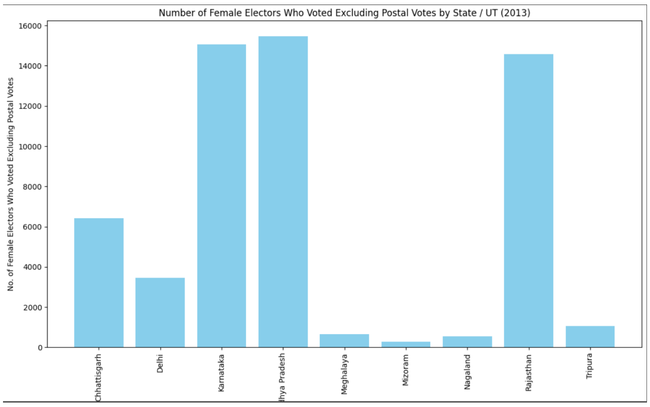


Figure 3.2: No. of Female Electors Who Voted in 2013

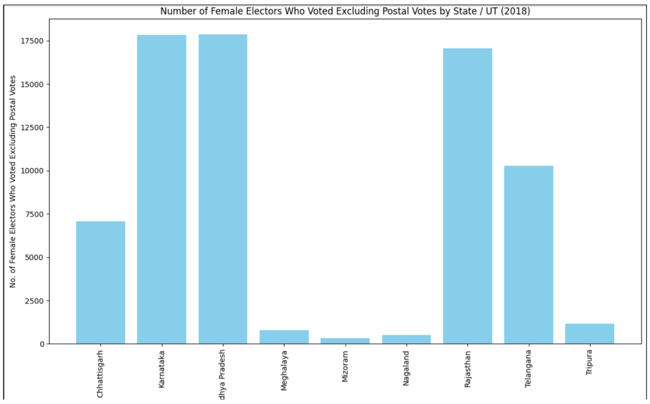


Figure 3.3: No. of Female Electors Who Voted in 2018

Table 3.5: Description of Columns after removing missing values

Column	Non-Null Count	Dtype
State / UT	100	object
Year of GE to SLA	100	int64
Reserved SC Seats	100	int64
Reserved ST Seats	100	int64
Total No. of Seats	100	int64
No. Of Male Contestants	100	int64
No. Of Female Contestants	100	int64
No. Of Transgender Contestants	100	float64
Total number of contestents	100	int64
Elected Male Candidates	100	int64
Elected Female Candidates	100	int64
Elected Transgender Candidates	100	float64
Total number of Candidates	100	int64
Elected Female perc	100	float64
Forfeited Deposits Male	100	int64
Forfeited Deposits Female	100	int64
Forfeited Deposits Transgender	100	float64
Total Deposits	100	int64
Average Contestants Per Constituency	100	object
Male Electors	100	float64
Female Electors	100	float64
Total Electors	100	float64
Male Electors Who Voted	100	float64
Female Electors Who Voted	100	float64
Total Votes Polled	100	float64
Male Polling Percentage Excluding Postal Ballots	100	float64
Female Polling Percentage Excluding Postal Ballots	100	float64
State Poll perc	100	float64
No of Polling Stations	100	int64
Average Number of Electors per PS	100	int64
Postal Votes	100	int64
Rejected Postal Votes	100	int64
NOTA	100	float64

Table 3.6: Transgender Data

State / UT	No. Of Transgender Contestants	Elected Transgender Candidates
Andhra Pradesh	3.0	0.0
Arunachal Pradesh	0.0	0.0
Assam	0.0	0.0
Bihar	1.0	0.0
Chhattisgarh	7.0	0.0
Delhi	1.0	0.0
Goa	0.0	0.0
Gujarat	0.0	0.0
Haryana	0.0	0.0
Himachal Pradesh	0.0	0.0
Jammu & Kashmir	0.0	0.0
Jharkhand	1.0	0.0
Karnataka	2.0	0.0
Kerala	1.0	0.0
Madhya Pradesh	6.0	0.0
Maharashtra	1.0	0.0
Manipur	0.0	0.0
Meghalaya	0.0	0.0
Mizoram	0.0	0.0
Nagaland	0.0	0.0
Odisha	2.0	0.0
Puducherry	1.0	0.0
Punjab	3.0	0.0
Rajasthan	0.0	0.0
Sikkim	0.0	0.0
Tamil Nadu	4.0	0.0
Telangana	2.0	0.0
Tripura	0.0	0.0
Uttar Pradesh	8.0	0.0
Uttarakhand	3.0	0.0
West Bengal	0.0	0.0

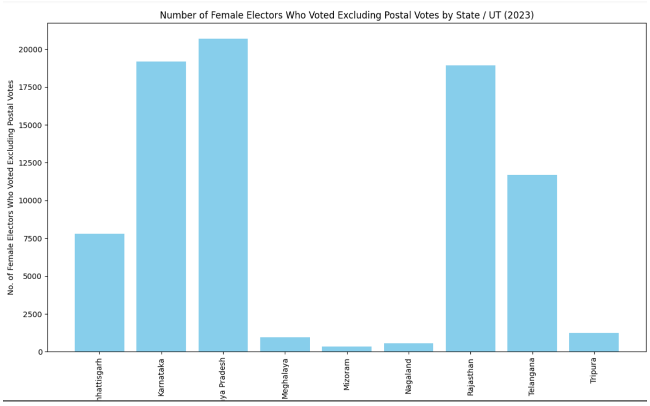


Figure 3.4: No. of Female Electors Who Voted in 2023

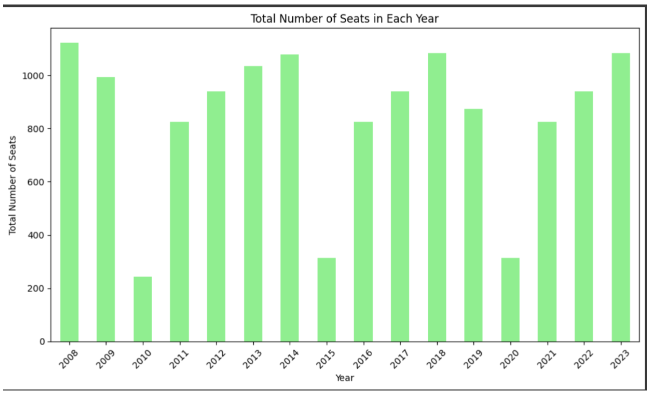


Figure 3.5: Total Number of Seats in Each Year

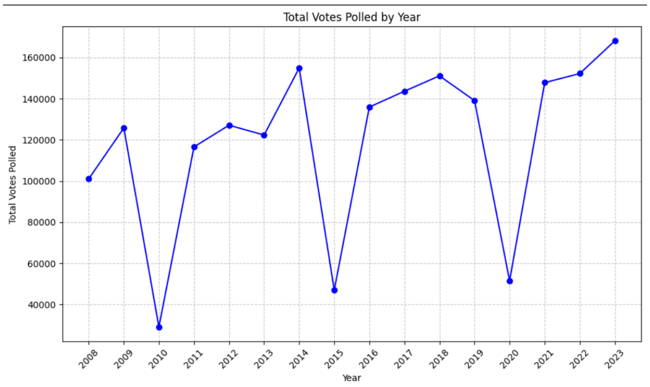


Figure 3.6: Total Votes Polled by Year

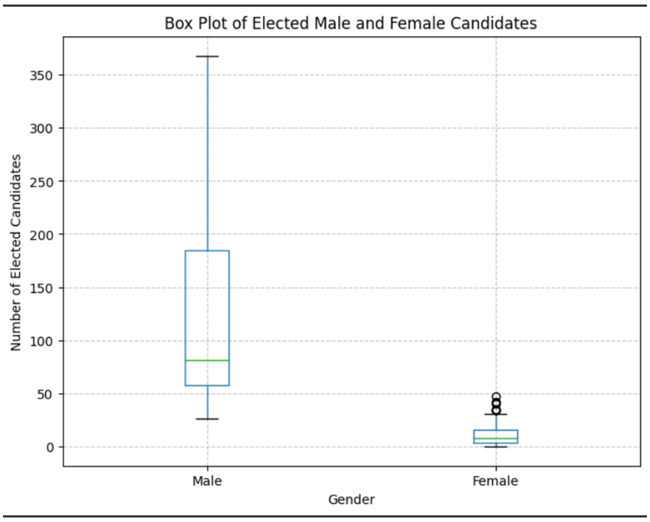


Figure 3.7: Box Plot of Elected Male and Female Candidates

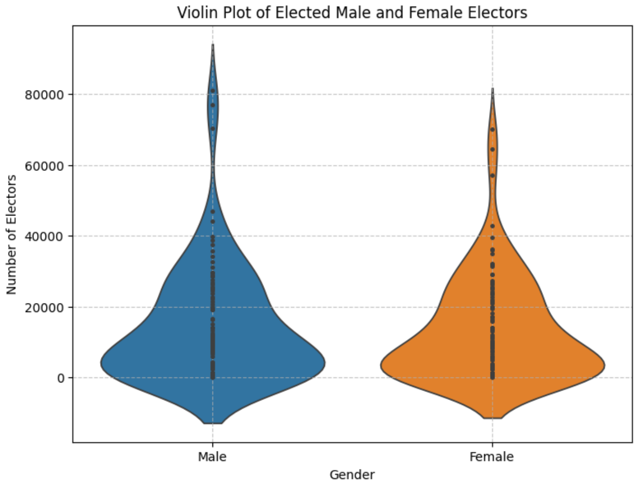


Figure 3.8: Violin Plot of Elected Male and Female Electors

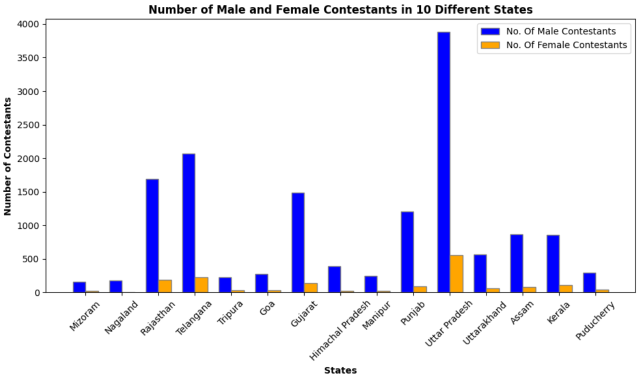


Figure 3.9: Number of Male and Female Contestants in 10 Different States

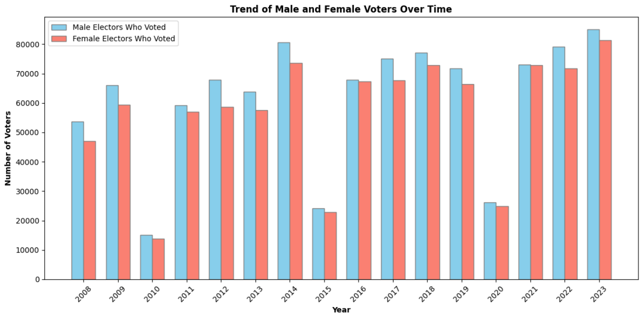


Figure 3.10: Trend of Male and Female Voters Over Time

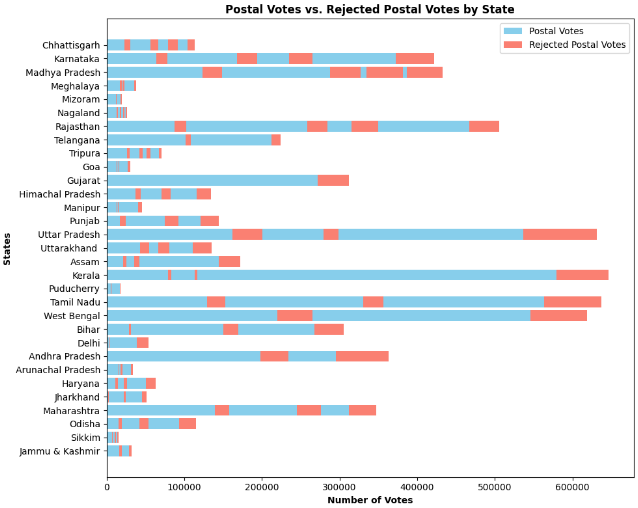


Figure 3.11: Postal Votes vs. Rejected Postal Votes by State

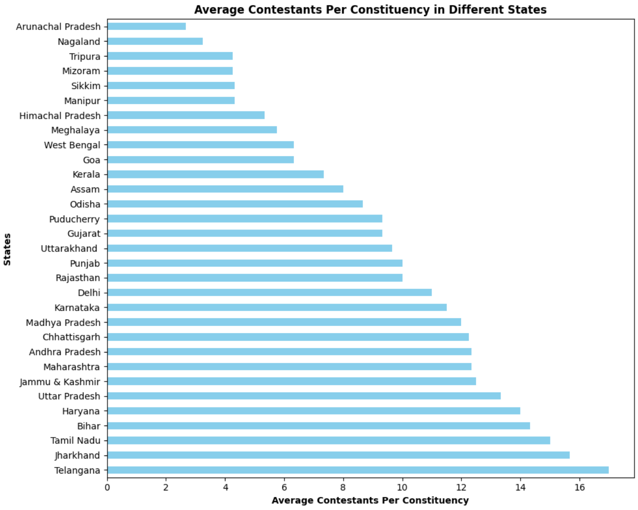


Figure 3.12: Average Contestants Per Constituency in Different States

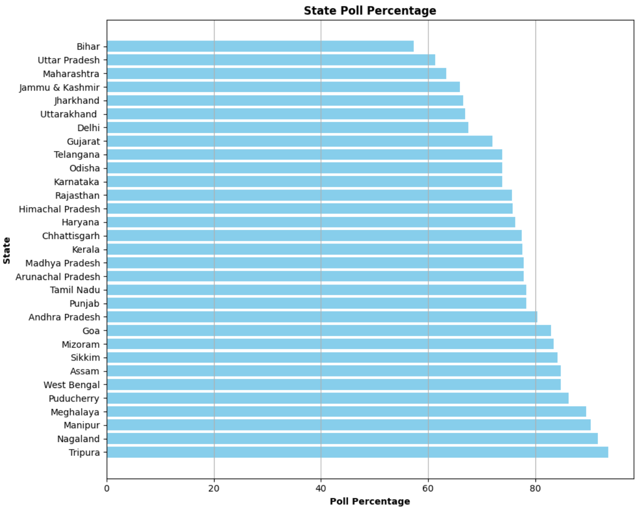


Figure 3.13: State Poll Percentage

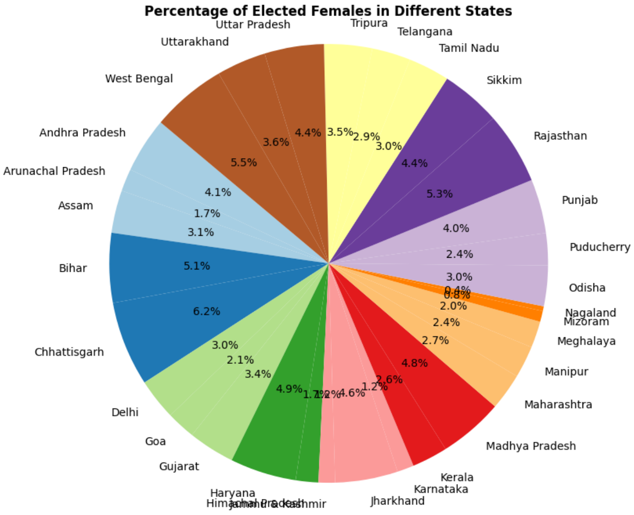


Figure 3.14: Percentage of Elected Females in Different States

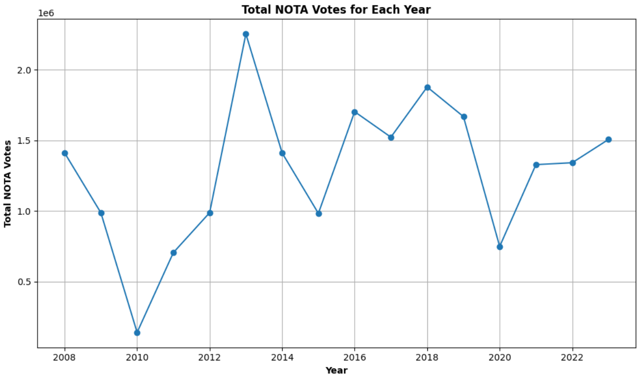


Figure 3.15: Total NOTA Votes for Each Year

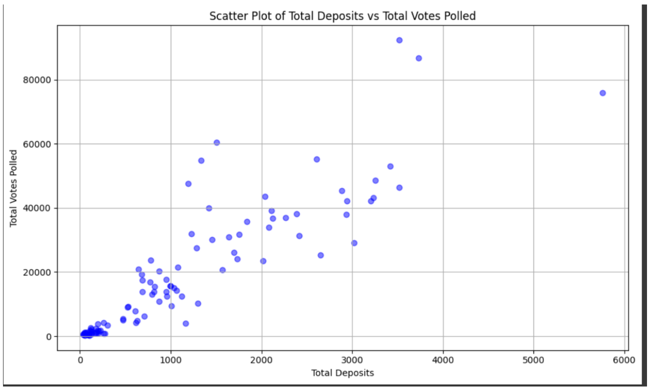


Figure 3.16: Scatter Plot of Total Deposits vs Total Votes Polled

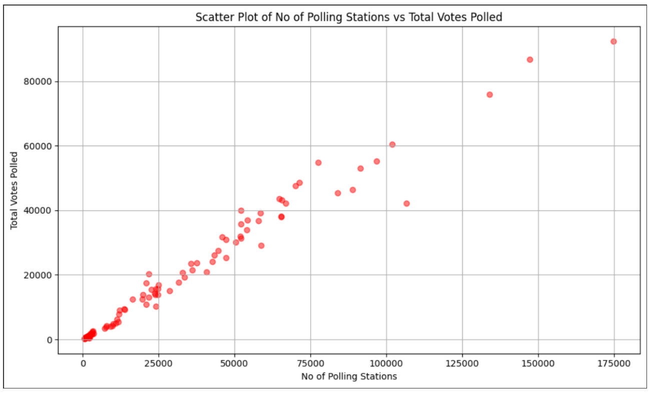


Figure 3.17: Scatter Plot of No of Polling Stations vs Total Votes Polled

3.2 Multivariate analysis

The blue scatter plot represents the relationship between total forfeited deposits and total number of votes polled in a year in different states. It is almost showing linear relationship between them.

The red scatter plot represents the relationship between number of polling stations and total number of votes polled in a year in different states. It is also showing linear relationship between them.

The heat map represents the relation between the average number of electors per polling station and total number of polling stations in a year in different states. The red areas represents strong correlation whereas the blue areas represent weak correlation.

The joint plots give a strong linear relationship between the number of total electors and the total votes polled. Also it gives linear relationship between male and female forfieted deposits.

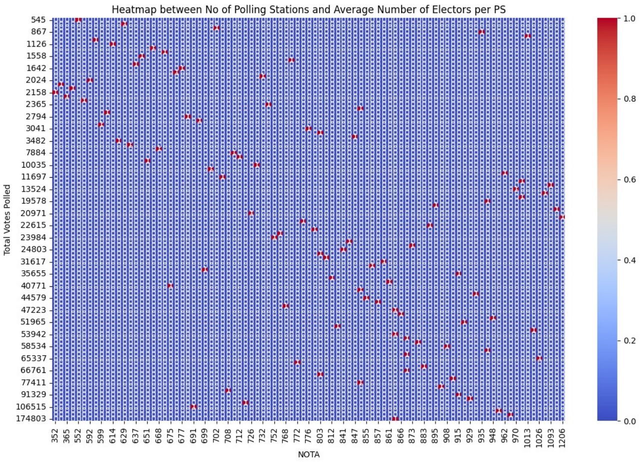


Figure 3.18: Heat map between No of Polling Stations and Average Number of Electors per PS

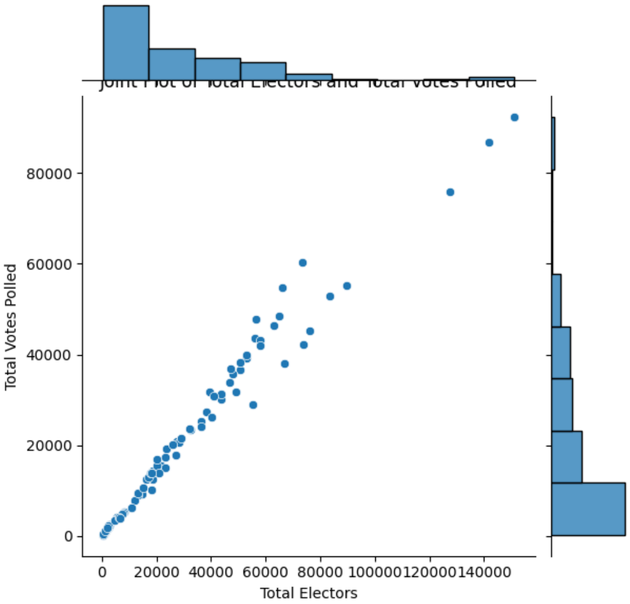


Figure 3.19: Joint Plot of Total Electors and Total Votes Polled

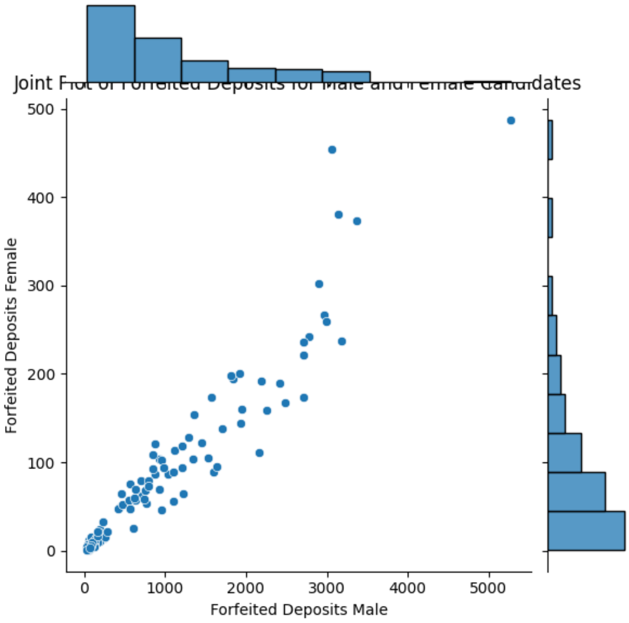


Figure 3.20: Joint Plot of Forfeited Deposits for Male and Female Candidate

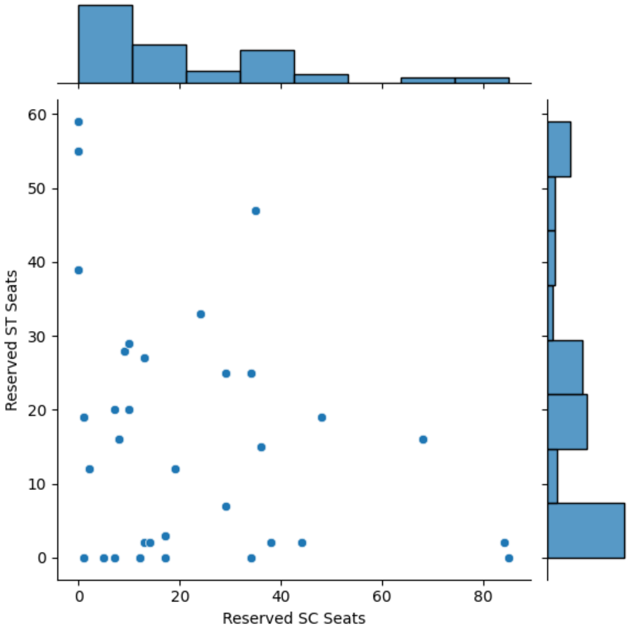


Figure 3.21: Joint plot between SC and ST seats

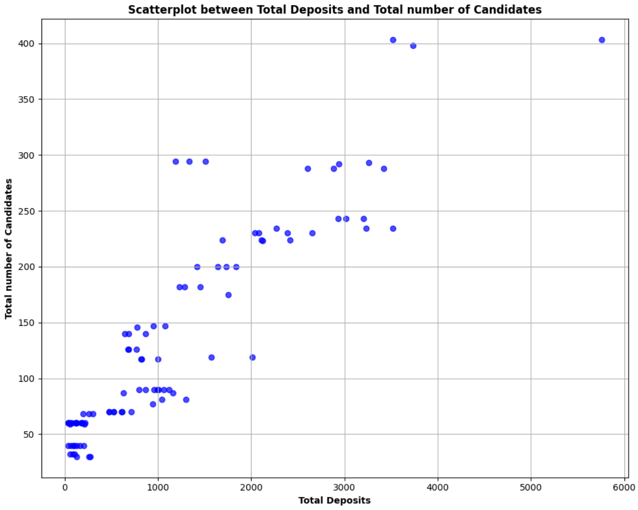


Figure 3.22: Scatterplot between Total Deposits and Total number of Candidates

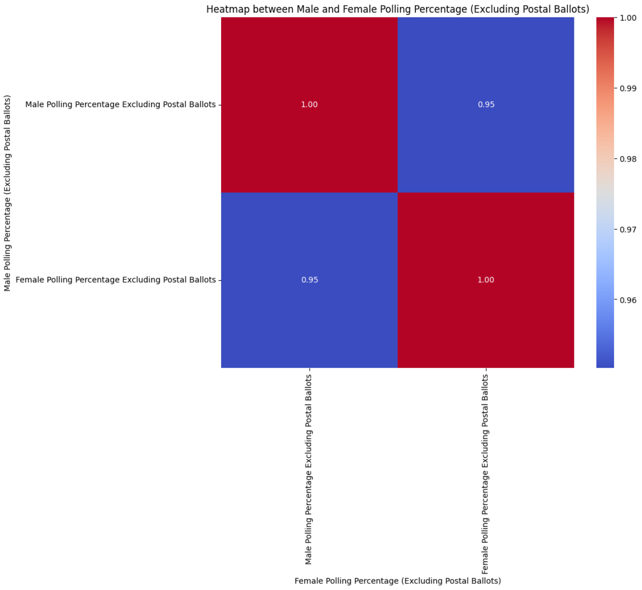


Figure 3.23: Heatmap between Male and Female Polling Percentage (Excluding Postal Ballots)

Chapter 4. Feature Engineering

Feature selection

1) Random Forest Algorithm

Intrinsic Feature Importance: After fitting the Random Forest model, intrinsic feature importances were obtained, which reflect how valuable each feature was in constructing the forest by measuring how much each feature decreases the weighted impurity in a tree. The top features were determined by sorting these importance values and selecting the most significant ones based on a predefined threshold or the top n features for further analysis or refinement of the model. **Algorithm:** Random Forest

Given:

- Training dataset $\{(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)\}$
- Number of trees T
- Number of features considered for splitting at each node m

For each tree $t = 1, 2, ..., T$:

1. We sample a bootstrap dataset D_t from the original dataset $\{(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)\}$ with replacement.
2. Grow a decision tree T_t using D_t with the following steps:
   (a) We select m features randomly from the total M features.
   (b) We choose the best feature among the selected features to split the node based on a splitting criterion (e.g., Gini impurity or information gain).
   (c) We repeat the splitting process recursively until a stopping criterion is met (e.g., maximum depth reached or minimum samples per leaf).
3. We store the decision tree T_t.

Prediction:

- For regression: Average the predictions of all trees.
- For classification: Take a majority vote among all trees.

2) Gradient Boosting Regressor

All selected features were used in the model training

$$
\begin{aligned}
F_0(x) &= \text{argmin}_\gamma \sum_{i=1}^{n} L(y_i, \gamma) \\
g_t(x_i) &= -\left[\frac{\partial L(y_i, F(x_i))}{\partial F(x_i)}\right]_{F(x)=F_{t-1}(x)} \\
F_t(x) &= F_{t-1}(x) + \nu_t \cdot h_t(x)
\end{aligned}
$$

Chapter 5. Model fitting

5.1 Regression

Gradient Boosting Regressor

1. Scaling Method Used:

None: Gradient Boosting constructs additive models by fitting successive trees to the residual errors made by previous trees, thus it inherently manages different feature scales.

2. Encoding Used:

OneHotEncoder: same method was used to transform categorical variables(name of the state) into a binary format that is into 1's and 0's format

3. Algorithm Used:

Gradient Boosting Regressor

Overview: Gradient Boosting is a machine learning technique for regression and classification problems, which produces a prediction model in the form of an ensemble of weak prediction models, typically decision trees.

Key Characteristics:

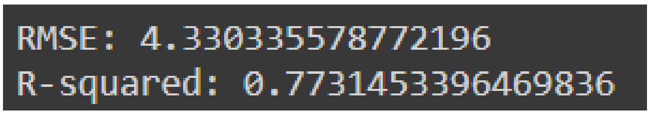
Sequential Learning: It builds trees one at a time, where each new tree helps to correct errors made by previously trained trees.
Loss Minimization: Uses a gradient descent algorithm to minimize the loss when adding new models.
Flexibility: Can optimize on different loss functions and provides several hyper parameter tuning options that make the function fit very flexible.

We have used 20 percent of the data for testing and rest 80 percent for training

5.2 ML algorithms

1) Random Forest Algorithm Intrinsic Feature Importance: After fitting the Random Forest model, intrinsic feature importances were obtained, which reflect how valuable each feature was in constructing the forest by measuring how much each feature decreases the weighted impurity in a tree. The top features were determined by sorting these importance values and selecting the most significant ones based on a predefined threshold or the top n features for further analysis or refinement of the model.

2) Gradient Boosting Regressor All selected features were used in the model training



```
RMSE: 4.330335578772196
R-squared: 0.7731453396469836
```

Figure 5.1: Root Mean Square Error

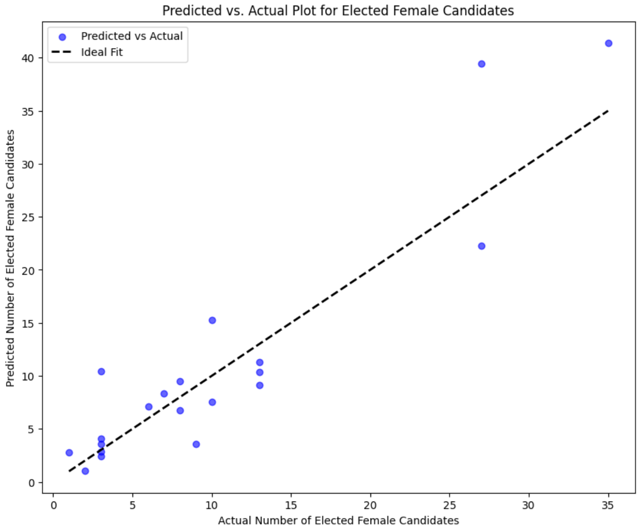


Figure 5.2: Gradient Boosting Regressor

3. Model Evaluation:

RMSE (Root Mean Squared Error): Measures the average magnitude of the errors between predicted and actual values, giving an absolute measure of fit. R-squared: Indicates the proportion of variance in the dependent variable that is predictable from the independent variables, offering insight into the goodness of fit.

Plot of Gradient Boosting Regression displays the importance of features for model building

Top-left quadrant (True Negative): The number 5 indicates that there are five instances where the model correctly predicted the negative class (e.g., 'Low NOTA'). Top-right quadrant (False Positive): The number 2 shows that the model incorrectly predicted the positive class (e.g., 'High NOTA') two times when it was actually the negative class. Bottom-left quadrant (False Negative): The number

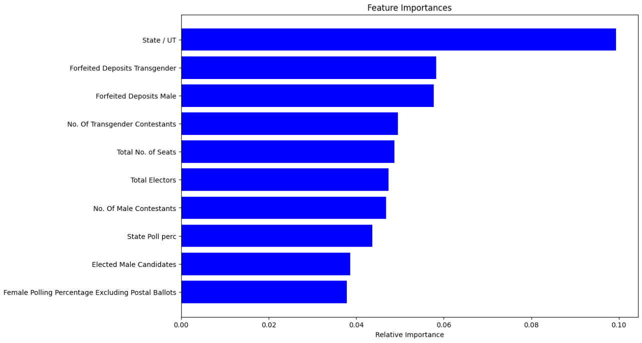


Figure 5.3: Feature Importance

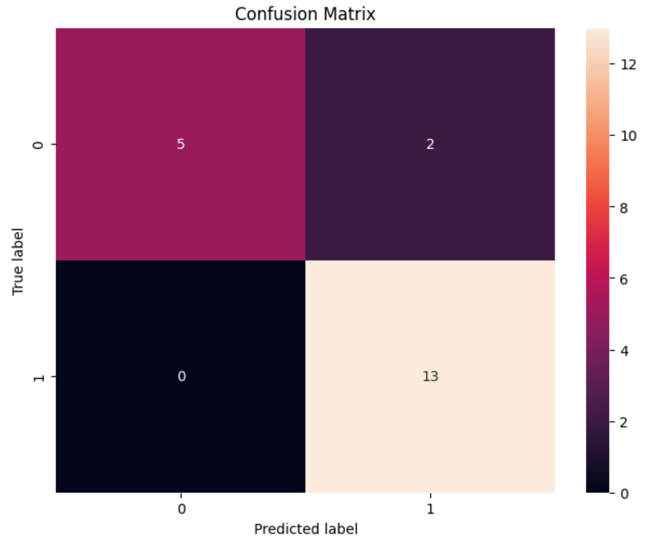
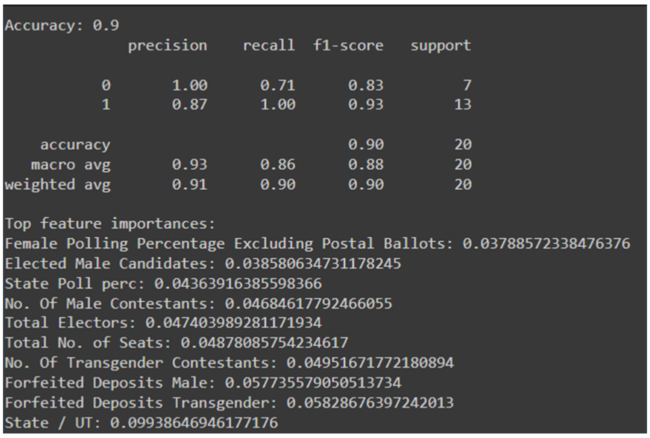


Figure 5.4: Correlation Matrix

Table 5.1: Classification Report

Class	Precision	Recall	F1-Score	Support
0	0.71	0.71	0.71	7
1	0.85	0.85	0.85	13
Accuracy			0.80	20
Macro Avg	0.78	0.78	0.78	20
Weighted Avg	0.80	0.80	0.80	20



```
Accuracy: 0.9
              precision    recall  f1-score   support

           0       1.00      0.71      0.83         7
           1       0.87      1.00      0.93        13

    accuracy                           0.90        20
   macro avg       0.93      0.86      0.88        20
weighted avg       0.91      0.90      0.90        20

Top feature importances:
Female Polling Percentage Excluding Postal Ballots: 0.03788572338476376
Elected Male Candidates: 0.038580634731178245
State Poll perc: 0.04363916385598366
No. Of Male Contestants: 0.04684617792466055
Total Electors: 0.047403989281171934
Total No. of Seats: 0.04878085754234617
No. Of Transgender Contestants: 0.04951671772180894
Forfeited Deposits Male: 0.057735579050513734
Forfeited Deposits Transgender: 0.05828676397242013
State / UT: 0.09938646946177176
```

Figure 5.5: Metrics for classification

0 indicates that there are no instances where the model predicted the negative class when it was actually the positive class. Bottom-right quadrant (True Positive): The number 13 signifies that the model correctly predicted the positive class 13 times.

Chapter 6. Conclusion & future scope

Random forest:

Voter Engagement and Gender Dynamics:

The feature 'Female Polling Percentage Excluding Postal Ballots' being a top influencer indicates that the engagement level of female voters is a significant predictor of NOTA votes. This may suggest that when female voter turnout is higher, there is a higher likelihood of NOTA votes being utilized, possibly reflecting a more critical stance among female voters towards the candidates offered. 'Elected Male Candidates' also appears as a key factor, which might imply that the number of male candidates elected (and, by extension, potentially the number of male candidates running) has a correlation with NOTA votes. This could be interpreted as a proxy for gender representation in politics; when voters perceive a gender imbalance, they might resort to NOTA as a form of protest.

'State Poll perc', reflecting overall voter turnout, suggests that general voter engagement in an election can be a bellwether for NOTA votes. This implies that higher voter turnout doesn't necessarily translate into fewer NOTA votes; rather, it might indicate that a more politically active electorate feels the need to express dissatisfaction through NOTA. The importance of 'State / UT' underscores regional differences in voter sentiment and behavior. Each state or territory's unique political climate, cultural factors, and issues at play can impact the inclination to vote NOTA

In summary, this analysis uncovers a multifaceted picture of voter behavior, where NOTA is used not only as a last resort but as a deliberate choice in response to the complexities of the electoral offerings. Understanding these patterns can provide valuable insights for political strategists, social scientists, and policymakers aiming to increase voter satisfaction and reduce the reliance on NOTA as an electoral statement.A, offering a tapestry of regional electoral narratives. Candidature and Electoral Outcomes.

In the analysis using the Random Forest Classifier to predict high NOTA voting instances, the model demonstrated a high degree of accuracy, with an overall correct classification rate of 90 percent. It performed exceptionally well in identifying instances of low NOTA votes, with perfect precision. However, it was slightly less precise but still very accurate in classifying high NOTA votes. The high recall for high NOTA votes indicates a strong sensitivity to detecting cases of voter dissatisfaction. The F1-scores were also high, particularly for high NOTA votes, suggesting a well-balanced model that is reliable for both classes.

Gradient Boosting :

Gradient Boosting Regressor tells a story about the factors contributing to the success of female candidates in elections. By focusing on the predictors of the number of elected female candidates, we gain insights into the elements that might enhance or impede their political victories.

With an RMSE value of 4.33, the model's predictions are close to the real outcomes, meaning it is effective at forecasting the number of female candidates elected within an acceptable error margin. An R-squared of 0.773 suggests a strong model that explains a substantial portion of the variation in the election results concerning female candidates.

6.1 Findings/observations

The graphs provide a comprehensive overview of female vote share across various states for the years. Notably, certain union territories did not participate in the voting process previously. However, for the remaining states, a consistent pattern emerges, indicating a nearly identical trend in female vote share across these years. Also there is a fluctuation in political representation over time.

Furthermore, in regards to the distribution of elected male and female candidates. Notably, it reveals a concerning trend of fewer female candidates being elected compared to their male counterparts. This observation underscores potential disparities in political representation and highlights areas for further investigation and action to promote gender equality in elected offices.

Also we have studied about the relationship between different variables like number of votes, forfeited deposits, polling stations and number of electors. From the plots we have found that there is a linear relationship in most of the cases with various degrees of correlation. It means that if there are more polling stations, then there will be more number of electors and voters and vice versa. So it signifies the importance of good infrastructure in different locations over time.

Analyzing Predictive Factors ,Total No. of Seats offers a foundational view of the election size. A greater number of seats may provide more opportunities for female candidates to win. No. Of Female Contestants is Directly linked to the outcome, more female contestants could logically lead to more victories, but this feature also suggests that simply having female candidates is not the sole determinant of their success. State / UT tells us Different states may have varying levels of gender sensitivity, cultural openness, and political maturity that can impact the election outcomes for female candidates. Year of GE to SLA: Watching the year-over-year changes could reveal evolving trends in gender inclusivity and the political empowerment of women. Total Electors, Female Electors, Male Electors: The size and composition of the electorate are crucial, as they provide a context for the voter base from which female candidates can draw support.

6.2 Challenges

Analyzing voting trends and patterns presents several challenges, which can complicate the interpretation of electoral dynamics and the development of effective strategies. Some of these challenges include:

1. **Data Availability and Quality:** Obtaining reliable and comprehensive data on voting behavior can be challenging, especially in regions with limited transparency or where electoral processes are not well-documented. Inaccurate or incomplete data can skew analysis results and hinder the identi-

fication of meaningful trends.

2. **Complexity of Factors:** Voting behavior is influenced by a multitude of interconnected factors, including demographics, Socioeconomic conditions, cultural values, infrastructure and historical context. Untangling these complex relationships requires sophisticated analytical techniques and interdisciplinary approaches, making it difficult to isolate the impact of individual variables.

3. **Temporal Dynamics:** Voting trends are subject to temporal dynamics, with preferences evolving over time in response to changing circumstances, events, and political climates. Short-term fluctuations can obscure long-term patterns, requiring careful consideration of temporal scales in analysis and interpretation.

4. **Sample Bias:** Surveys and polls used to gather data on voting behavior may suffer from sample bias, where certain demographic groups are over represented or underrepresented. Biased samples can skew results and lead to inaccurate conclusions about broader voter sentiments.

5. **Causality vs. Correlation:** Identifying causal relationships between variables in voting behavior is challenging, as correlations observed in data do not necessarily imply causation. Disentangling causality requires rigorous statistical methods and consideration of potential confounding factors.

6. **Regional Variations:** Electoral dynamics can vary significantly across regions, reflecting diverse political landscapes, cultural norms, and socio-economic conditions. Generalizing findings from one region to another can be problematic, requiring localized analysis and contextualization of results.

7. **Emerging Trends and Technologies:** Rapid advancements in technology and changes in communication channels are reshaping how voters engage with political information and interact with candidates. Understanding the implications of emerging trends such as social media influence, misinformation, and algorithmic bias poses ongoing challenges for researchers and analysts.

8. **Ethical Considerations:** The collection and analysis of voter data raise ethical concerns related to privacy, consent, biasness between people and potential misuse of information. Researchers must adhere to ethical guidelines and data protection regulations to ensure the responsible handling of sensitive information.

6.3 Future plan

The future of analyzing voting trends and patterns involves several key initiatives aimed at overcoming existing challenges and leveraging emerging opportunities:

1. **Advanced Data Analytics:** Continued advancements in data analytics techniques, including machine learning, natural language processing, and network analysis, will enable more sophisticated analysis of voting behavior. Predictive modeling and simulation tools can help anticipate electoral outcomes and assess the impact of policy changes and campaign strategies.

2. **Integration of Multiple Data Sources:** Integrating diverse data sources, including voter registration databases, social media interactions, demographic surveys, and election results, will provide

a more comprehensive understanding of voter behavior. Data fusion techniques and interoperable platforms will facilitate seamless integration and analysis of heterogeneous data sources.

3. **Real-time Monitoring and Feedback:** Developing real-time monitoring systems to track voter sentiment, campaign dynamics, and emerging issues will enable agile decision-making and response strategies. Interactive dashboards, sentiment analysis algorithms, and social network analysis tools can provide actionable insights to political stakeholders and electoral authorities.

4. **Ethical Data Governance:** Strengthening ethical principles and data governance frameworks to protect voter privacy, ensure data integrity, and mitigate potential biases and discriminatory practices. Transparency, accountability, and inclusively should guide the collection, analysis, and dissemination of voter data, fostering trust and confidence in electoral processes.

5. **Cross-disciplinary Collaboration:** Promoting interdisciplinary collaboration between political scientists, statisticians, data scientists, sociologists, and communication scholars will foster innovation and knowledge exchange in the study of voting behavior. Collaborative research projects, joint training programs, and interdisciplinary conferences can facilitate cross-pollination of ideas and methodologies.

6. **Public Engagement and Education:** Enhancing public awareness and understanding of voting trends and patterns through educational initiatives, public forums, and data literacy programs. Empowering citizens with the knowledge and skills to critically evaluate political information and engage in informed decision-making is essential for strengthening democratic participation and civic engagement.

7. **International Cooperation:** Promoting international cooperation and knowledge sharing to address global challenges in analyzing voting behavior, such as electoral fraud, disinformation campaigns, and foreign interference. Collaborative research networks, joint data-sharing agreements, and capacity-building initiatives can facilitate cross-border collaboration and mutual learning.

8. **Innovation in Electoral Technology:** Embracing technological innovations, such as blockchain voting systems, electronic voting platforms, and secure online voter registration tools, to enhance the efficiency, accessibility, and integrity of electoral processes. Investing in research and development of secure and user-friendly electoral technologies will help modernize voting systems and expand democratic participation.

9. **Gender Participation**: Equal participation is necessary for both the genders in order to select a proper candidate and to maintain political stability

10. **Proper Infrastructure**: Because of more number of polling stations, more electors will be able to give vote and the election process will be held smoothly resulting in proper election conduct.

Group Contribution

Member 1

Contributions of member 1 to the EDA final project: 1. Downloaded and finalised the dataset and made it usable for analysis 2. Done data preprocessing and analyzed the data and constructed plots 3. Report Making and finding insights

Member 2

Contributions of member 2 to the EDA final project: 1. Did feature engineering and made the data usable for applying ML algorithms 2. Applied Machine learning models and tested accuracy, recall, precession and f1 score. 3. Aided in Report making and finalizing it.

Member 3

Contributions of member 3 to the EDA final project: 1. Presented ideas about the model to use and report making 2. Provided resources regarding datasets

Short Bio

1. **Anurag Choudhury** is a passionate it enthusiat who loves doing work related to tech and data analysis. Done his bachelors degree in statistics from Gauhati University.

Throughout his career, Anurag has actively contributed to open-source projects and has a keen interest in exploring new technologies. He is proficient in programming languages such as Python and analytical skills like optimizing algorithms and a knowledge about machine learning

In addition to his technical skills, Anurag is an effective team player with excellent communication skills. He enjoys collaborating with colleagues and has a track record of delivering high-quality solutions within tight deadlines.

Outside of work, Anurag enjoys music, playing the guitar, and participating in music events.

2. **Aditya Tripathi** is a passionate it enthusiat who loves doing work related to tech and data analysis and a keen interest in mathematics. Done his bachelors degree in mathematics from Gautam Budh University.

Throughout his career, Aditya has actively contributed to open-source projects and has a keen interest in exploring new technologies. He is proficient in programming languages such as Python and analytical skills like optimizing algorithms and a knowledge about machine learning.

Aditya is an effective team leader with excellent leadership skills. He enjoys collaborating with colleagues, professors. Also he is a good solution provider.

Outside of work, Aditya enjoys debating and chess

3. **Karan Sharma** is a passionate it enthusiat who loves doing work related to tech and data analysis and a keen interest in computer science. Done his bachelors degree in Data Science from Gujarat University.

Throughout his career, Karan has actively contributed to open-source projects and has a keen interest in exploring new technologies. He is proficient in programming languages such as Python, C, and C++, and enjoys tackling challenging problems in software development and machine learning.

In addition to his technical skills, Karan is an effective guide with excellent communication skills. He enjoys collaborating with colleagues and has a track record of explaining insights to his colleagues and other people.

Outside of work, Karan enjoys badminton.

References

[1] Excel file link. *URL:* https://old.eci.gov.in/files/file/15617-year-wise-information-on-important-parameters-of-general-election-to-state-l

[2] Website url. *URL:* https://www.eci.gov.in/

[3] Other Website url. *URL:* https://data.gov.in/
URL: https://data.gov.in/resource/general-election-lok-sabha-parliamentary-constitue
URL: https://data.gov.in/resource/general-election-lok-sabha-parliamentary-constitue