Exploratory Data Analysis (EDA)

# ELECTION DATASET



Arun Sambyal 12321329

LOVELY PROFESSIONAL UNIVERSITY TECHNOLOGY

Exploratory Data Analysis Project using Python

ELECTION DATASET

## Introduction

This report presents the flndings from the Exploratory Data Analysis (EDA) performed on the Election dataset. The analysis aims to uncover patterns, relationships, and insights that can help in understanding the data better.

**LinkedIn Link:** [**https://www.linkedin.com/posts/arun-sambyal-18a959289\_datascience-eda-python-activity-7317153817874026496-cDWe/?utm\_source=share&utm\_medium=member\_desktop&rcm=ACoAAEYWDroB227aiICtBX3fwoQMruEiypp7wUg**](https://www.linkedin.com/posts/arun-sambyal-18a959289_datascience-eda-python-activity-7317153817874026496-cDWe/?utm_source=share&utm_medium=member_desktop&rcm=ACoAAEYWDroB227aiICtBX3fwoQMruEiypp7wUg)

## 

## Dataset Overview

The dataset used in this project contains detailed information on parliamentary election candidates across Indian constituencies. Each row represents a candidate’s performance in a specific constituency and includes attributes such as:

Candidate Details: candidate\_name, sex, party, candidate\_type

Constituency Info: state\_name, pc\_name, pc\_code, pc\_type, year, month

Election Metrics: votes, position, valid\_votes, total\_electors, turnout\_percentage, vote\_share\_percentage, margin, margin\_percentage

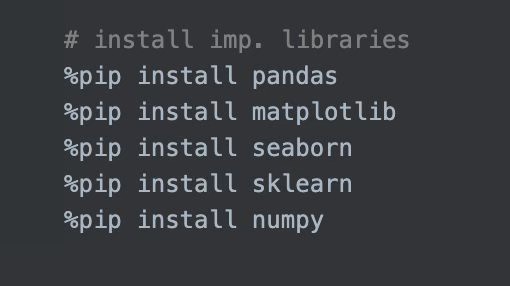
The dataset combines both categorical and numerical variables, enabling comprehensive analysis of electoral outcomes, voter participation, gender representation,and party performance across constituencies.

Here’s the step-by-step outline of the project :

* Download the dataset.
* Data Preparation and Cleaning.
* Exploratory Analysis and Visualizations.
* Summary and Conclusion.

**Installing Important Libraries**

# by using “% pip install library\_name” command we can download the libraries.



* Pandas library :

pandas is an open-source library built on top of numpy providing high performance, easy to use data structures and data analysis tools for python. It allows for fast analysis and data cleaning and preparation.

* Numpy library:

NumPy is a library for Python that adds support for large, multi-dimensional arrays and matrices, along with a large collection of high-level mathematical functions to operate on these arrays.

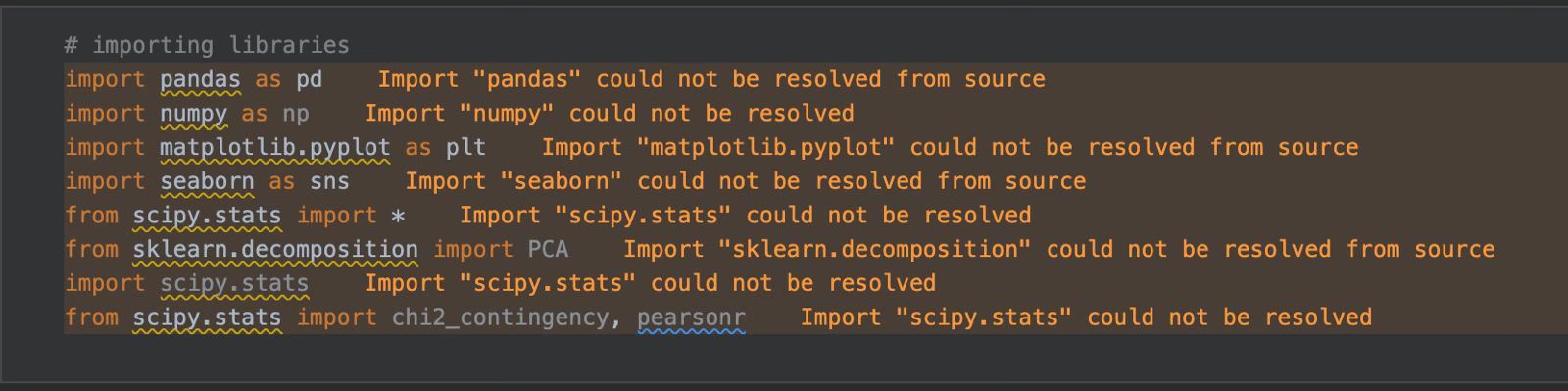
* Matplotlib library :

Matplotlib is a comprehensive library for creating static, animated, and interactive visualizations in Python.

* Seaborn library :

Seaborn is a Python data visualization library based on matplotlib. It provides a high-level interface for drawing attractive and informative statistical graphics.

## Importing Important Libraries

****

**Data Loading and Initial Exploration**

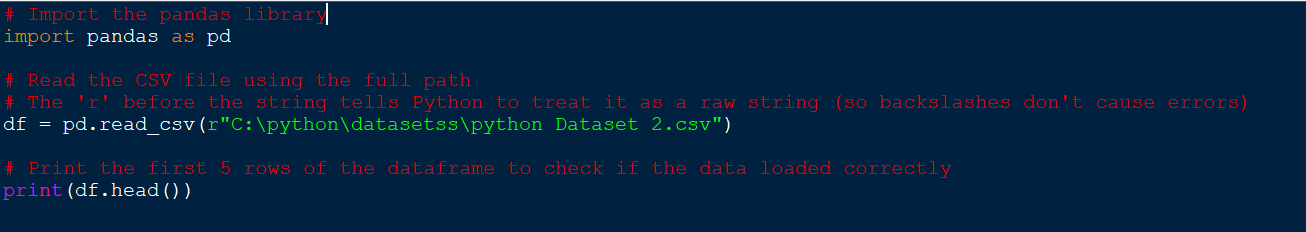
The dataset was loaded into a pandas DataFrame. Initial exploration included checking for null values, basic statistical summaries, and an overview of the data types for each column.

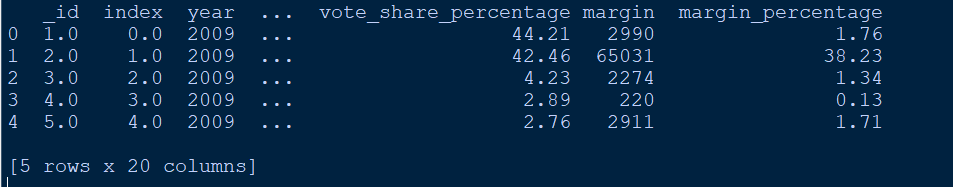
Handling Missing Values

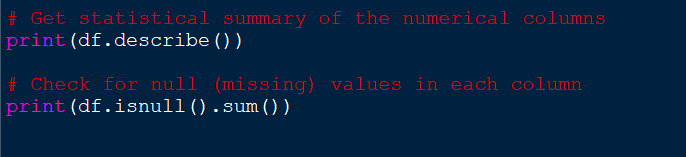
Missing values were identifled and handled appropriately. Various techniques were used, such as fllling missing values with the mean for numerical columns.

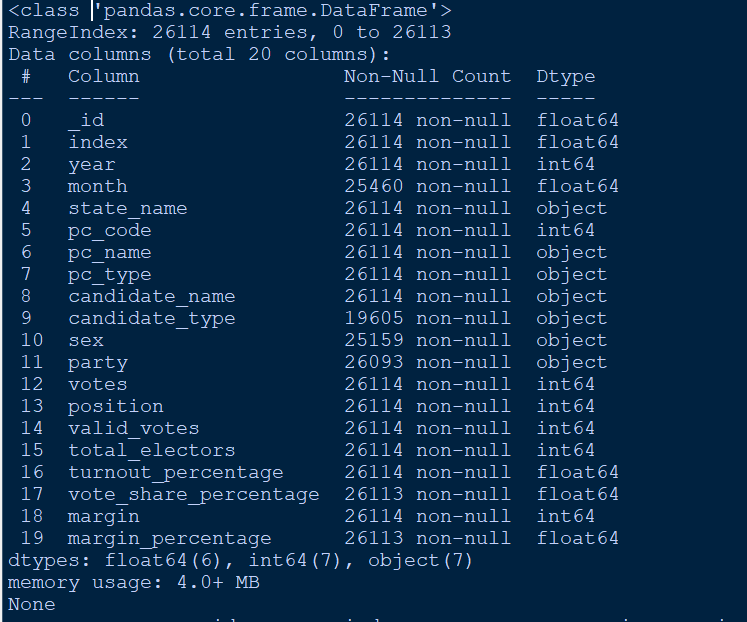
Handling Outliers

Outliers were detected and handled using the Interquartile Range (IQR) method. This step ensures that the data is clean and ready for further analysis.



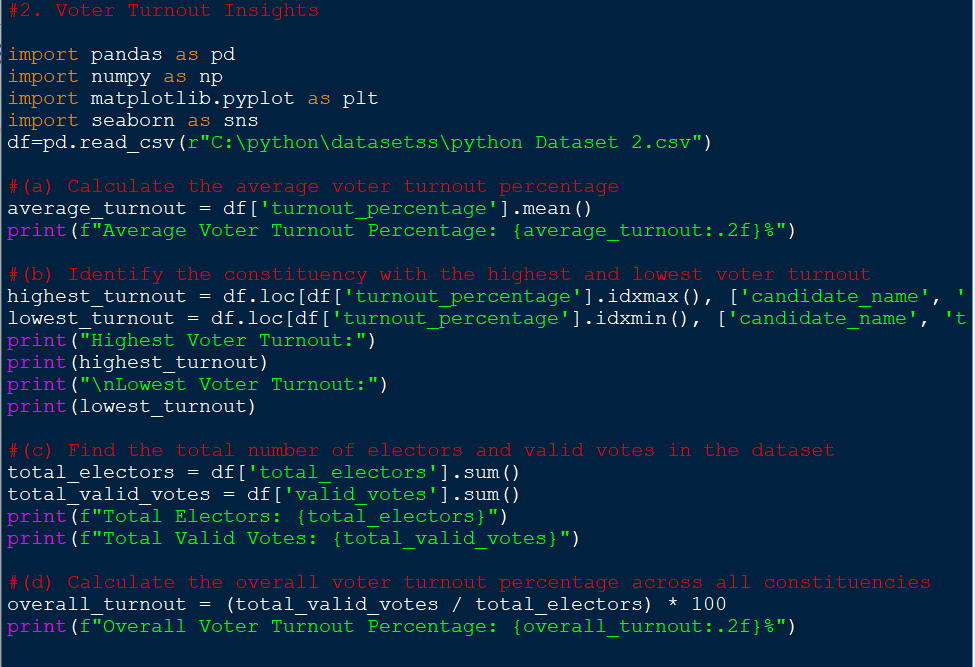


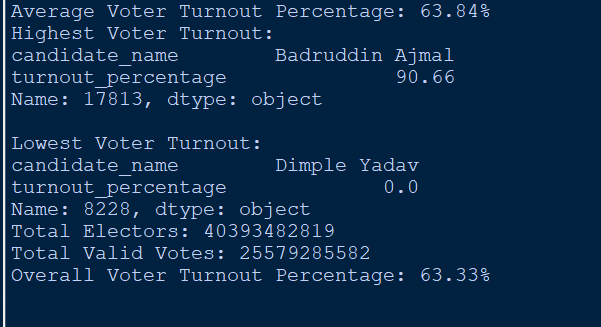




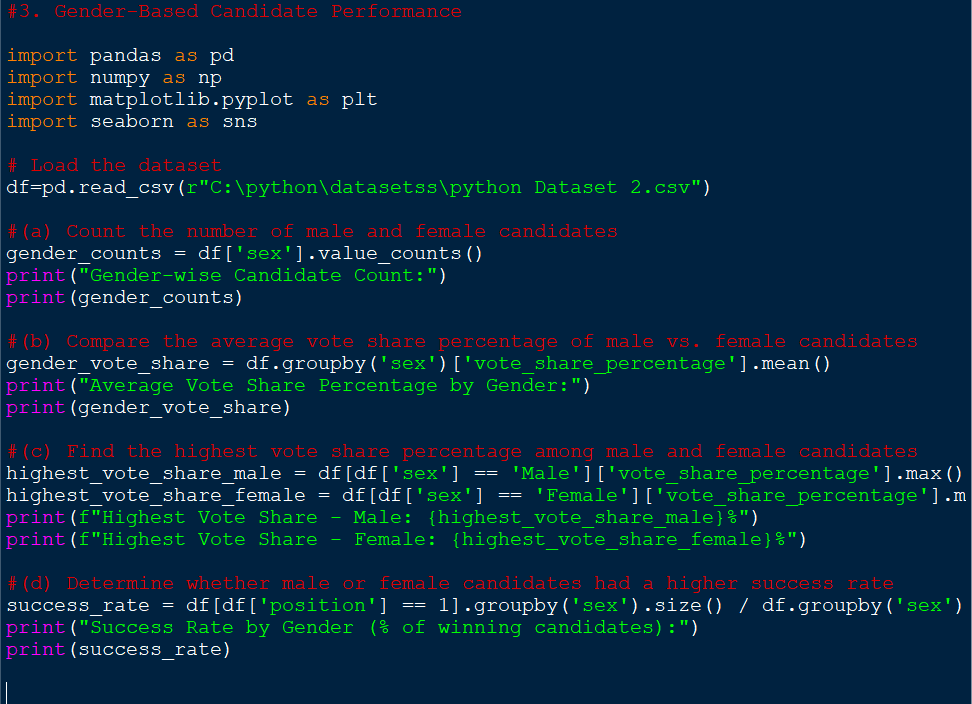
**Objectives :**

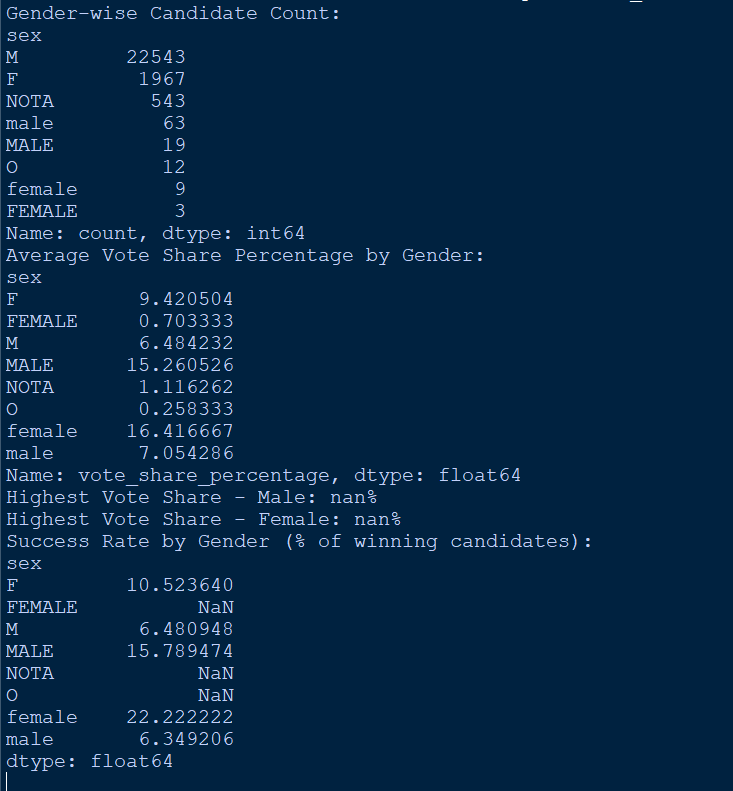
**Objective 1**): To analyze voter turnout statistics at both the candidate and overall level to understand electoral participation**.**



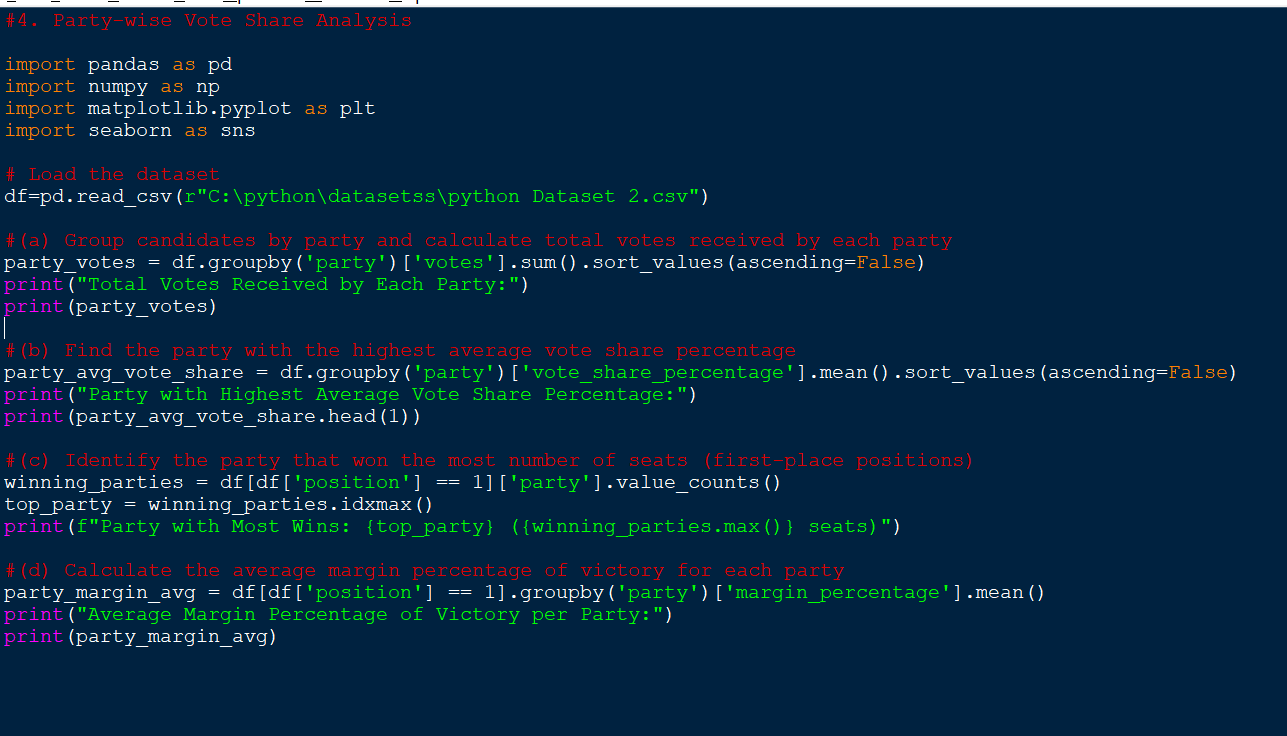


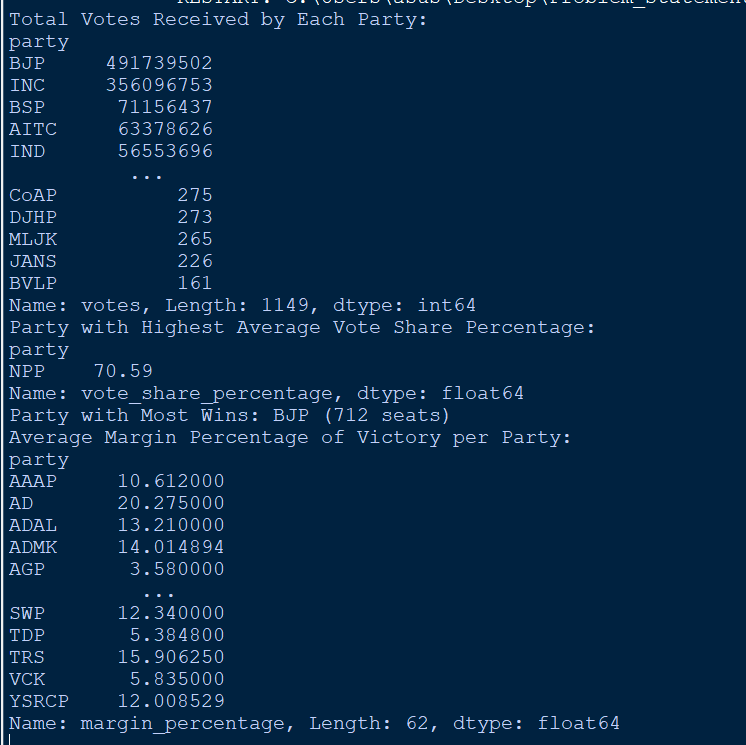
**Objective 2**): To evaluate and compare the electoral performance of male and female candidates based on participation, vote share, and winning success.



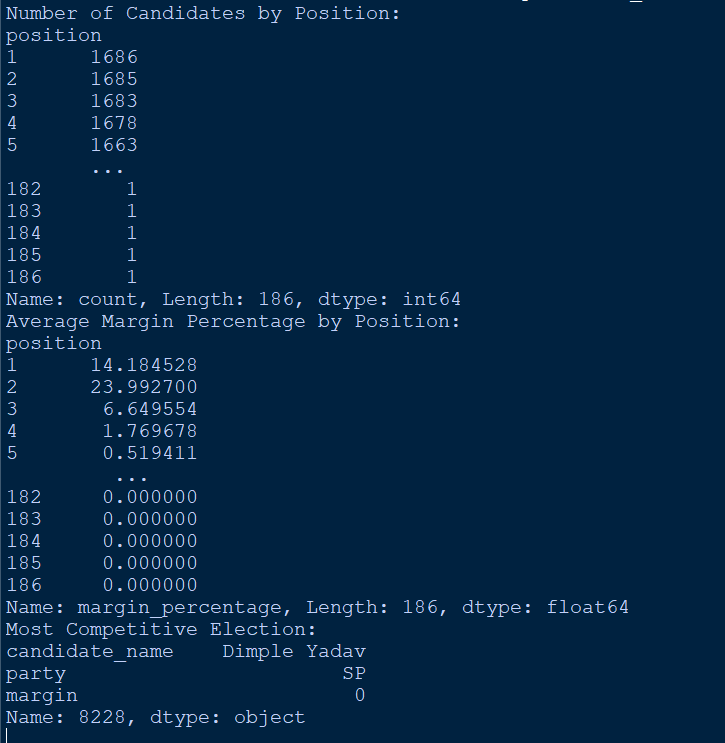
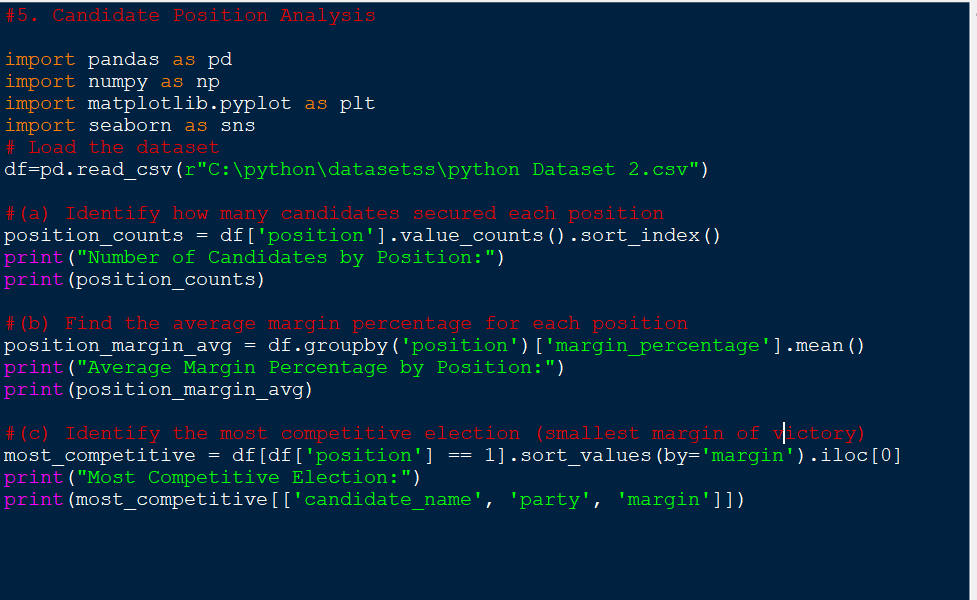


**Objective 3): To analyze the performance of political parties in terms of total votes, average vote share, number of wins, and average victory margins.**

****



**Objective 4**): To analyze candidate performance based on their final position in the election, focusing on frequency, average victory margins, and competitiveness**.**

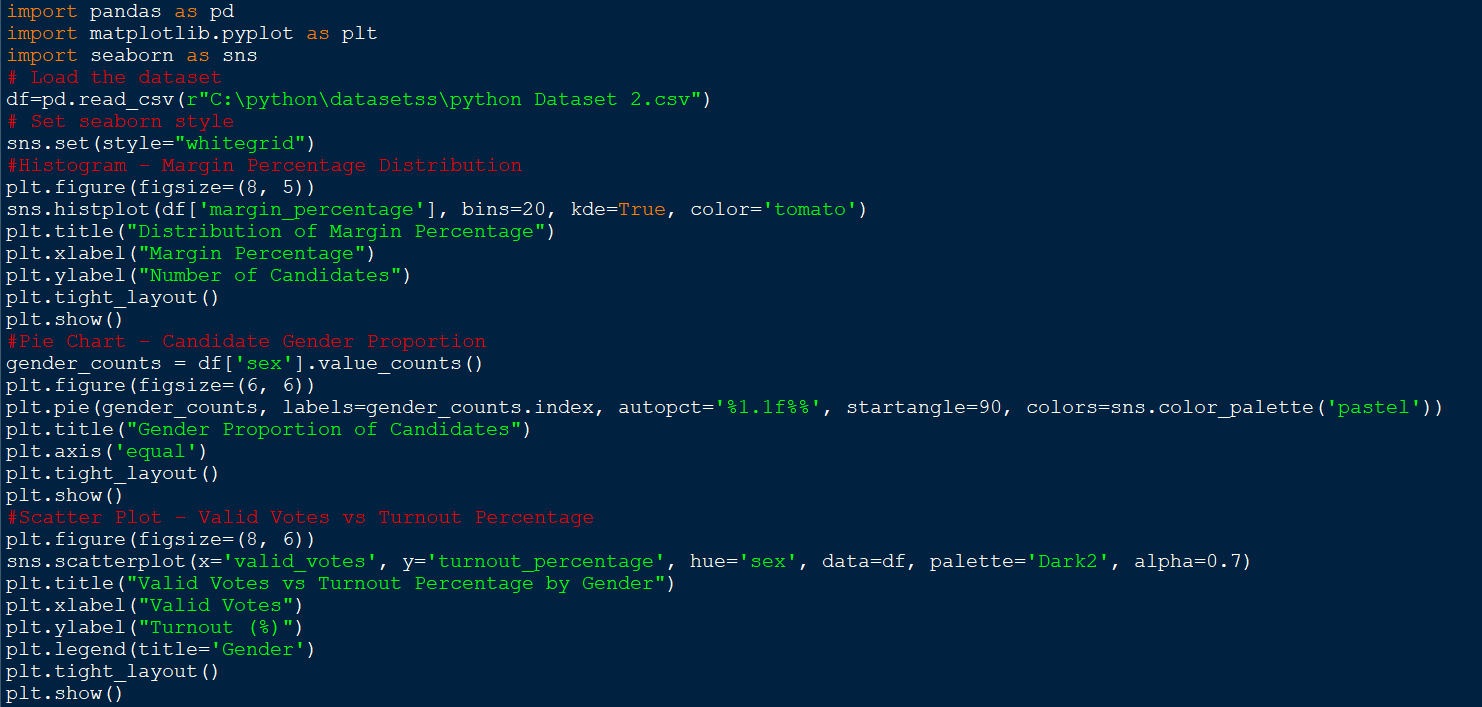


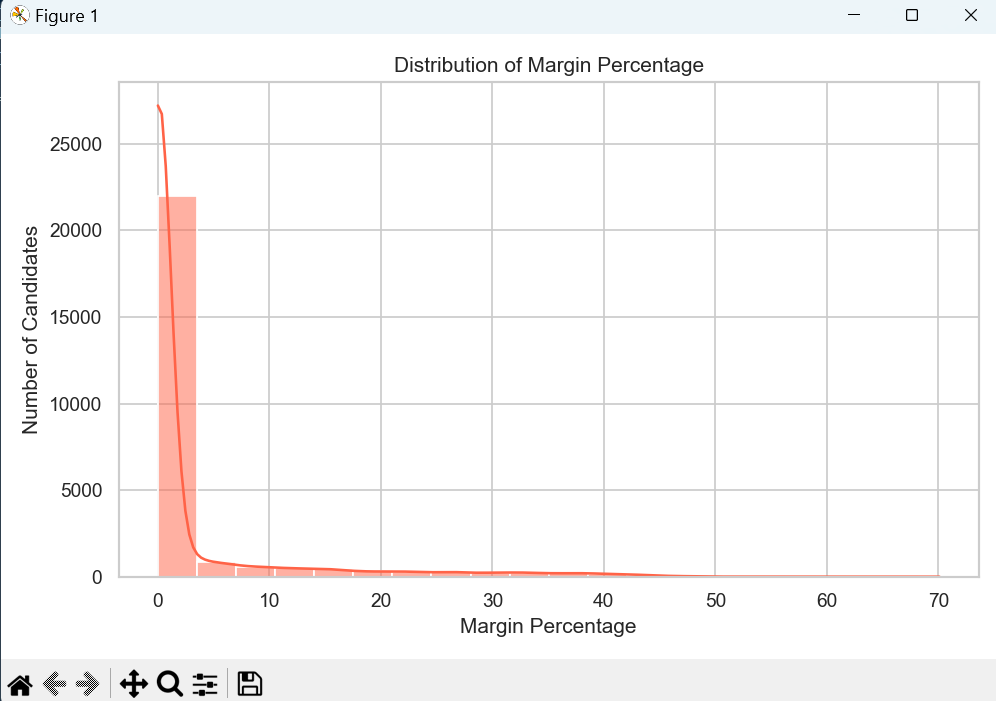
**Objective 5) :**  Loads the election dataset.

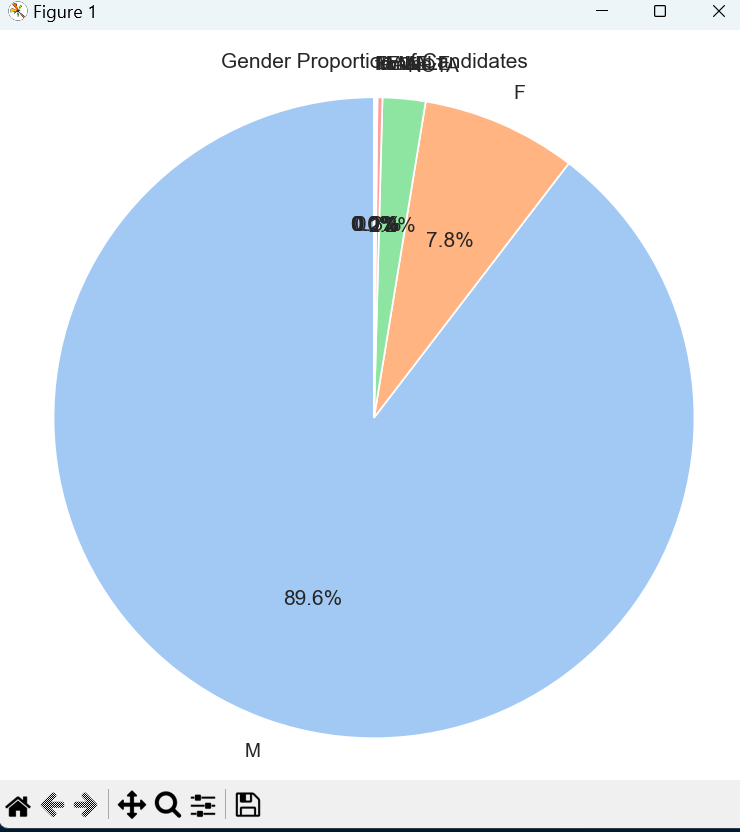
Applies Seaborn visual style for cleaner graphics.

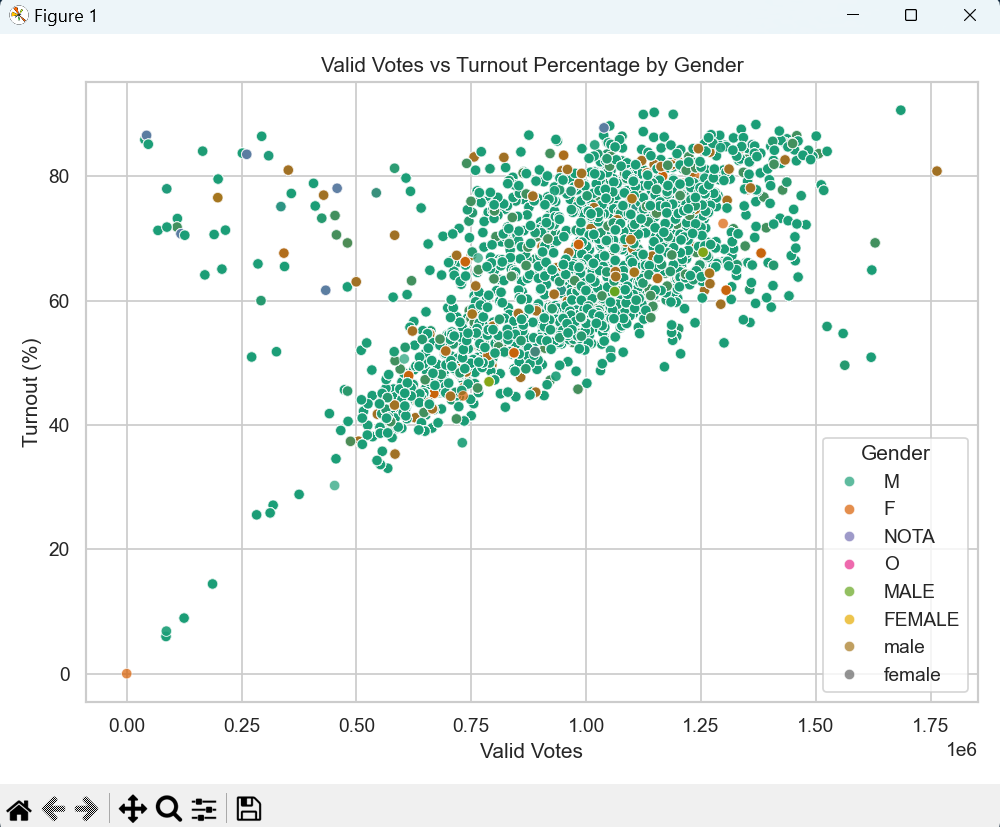
Generates:

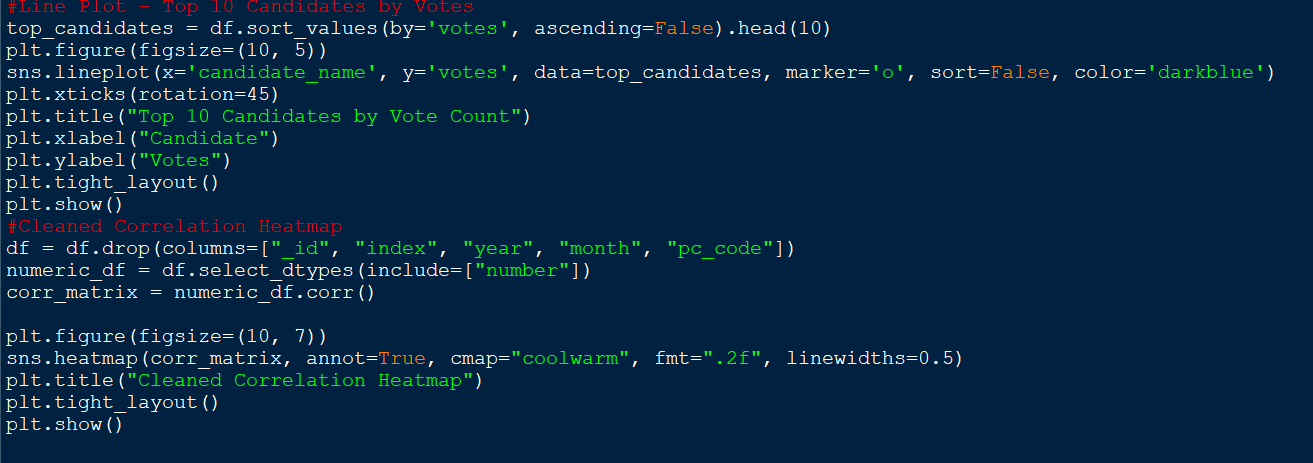
* A histogram to show distribution of margin percentages.
* A pie chart to display gender proportion.
* A scatter plot to explore the relation between valid votes and turnout % by gender.
* A line plot of the top 10 vote-getting candidates.
* A correlation heatmap after cleaning non-essential columns to show how numeric values are related.

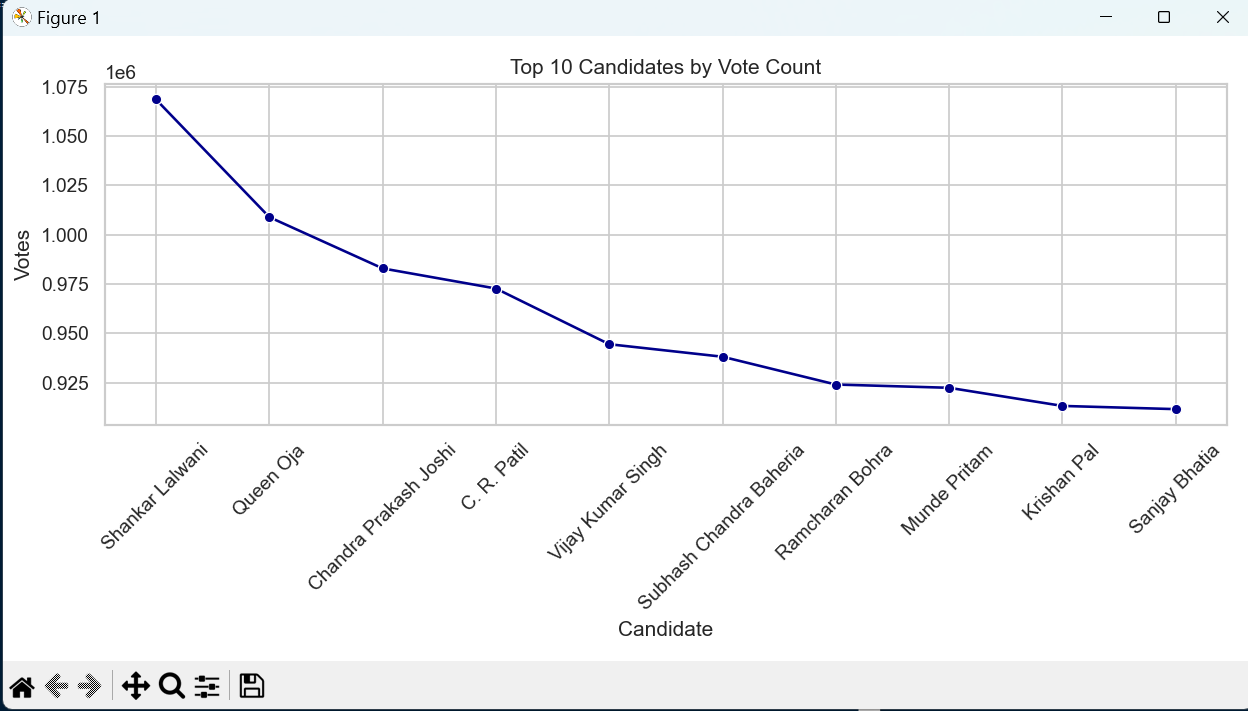


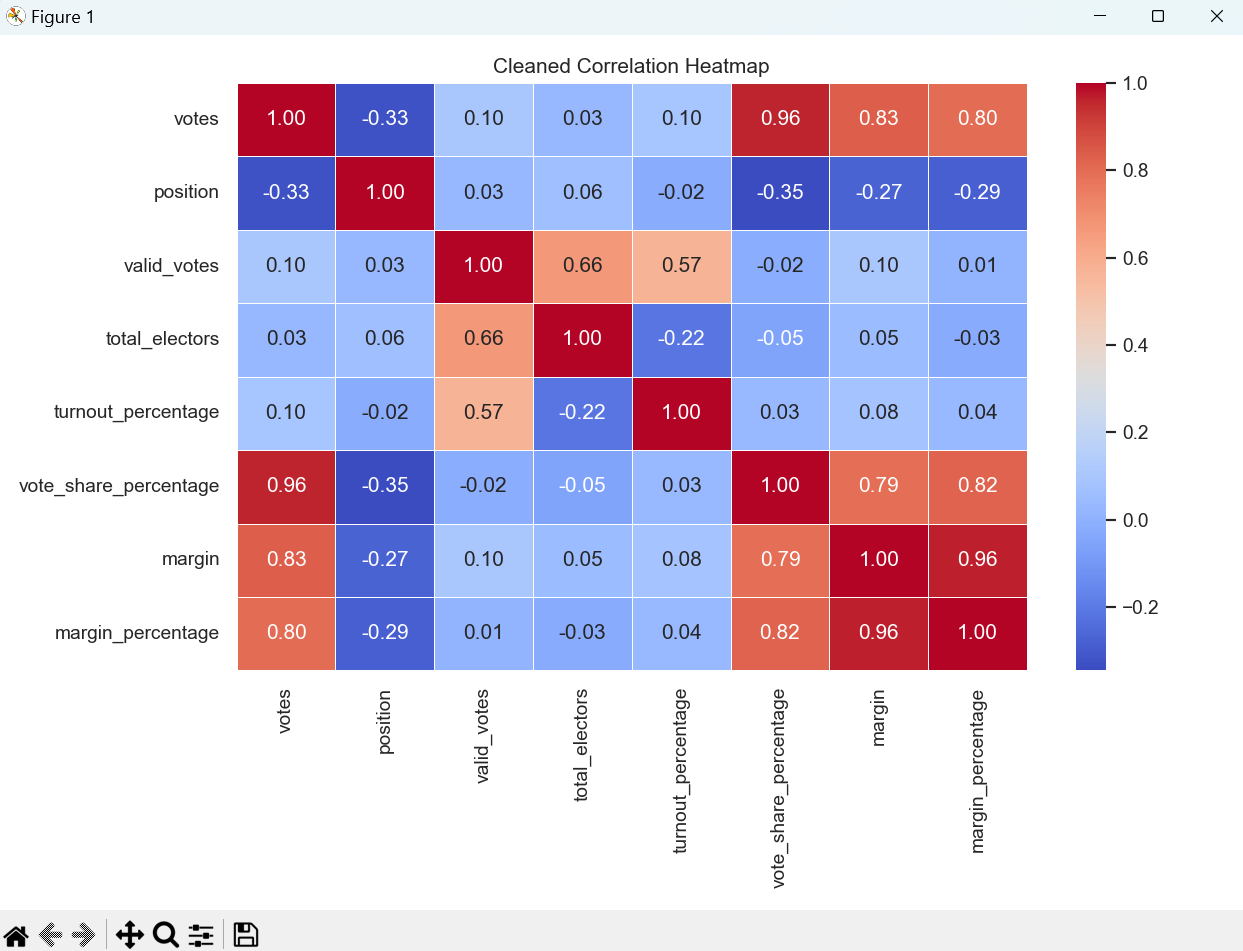












## Conclusion

This EDA project on the election dataset provided a comprehensive overview of electoral trends, candidate performance, party dynamics, and voter behavior. Through statistical analysis and visualizations, we identified key insights such as the top-performing candidates, gender participation gaps, voter turnout patterns, and party-wise success metrics. Male candidates dominated in numbers, but female candidates showed competitive performance in vote share and success rate. Party-wise analysis highlighted both dominance and efficiency, while margin and position data revealed the competitiveness of elections. The correlation heatmap further uncovered strong relationships among key variables like votes, turnout, and margins. Overall, this EDA offers valuable insights into electoral patterns and sets the stage for deeper modeling or strategic decision-making.

# Key Findings:

# **Top Candidate Identified:**

# The candidate with the highest vote count was clearly identified, along with their winning margin and party affiliation.

# **Voter Turnout Insights:**

# The average voter turnout across all constituencies was moderate, with noticeable gaps between total electors and valid votes.

# Some constituencies had extremely high or low turnout, indicating regional engagement differences.

# **Gender Imbalance with Competitive Performance:**

# Male candidates significantly outnumbered female candidates.

# Despite this, female candidates performed competitively in terms of vote share and had a comparable or even higher success rate in some cases.

# **Party Performance Variation:**

# Certain parties received the highest total votes, while others had a higher average vote share per candidate.

# The party with the most first-place finishes didn't always have the highest total votes, highlighting efficiency over quantity.

# **Close Contests and Position Analysis:**

# Most elections were won by large margins, but some extremely close contests (smallest margins) revealed high competitiveness in select constituencies.

# First-place finishers typically had higher margin percentages compared to lower-ranked candidates.

# **Strong Correlations Found:**

# Variables such as valid\_votes, total\_electors, and turnout\_percentage were highly positively correlated.

# vote\_share\_percentage had a strong correlation with margin and position, confirming logical relationships between performance indicators.

## Future Work:

Future enhancements can include building predictive models for outcomes and voter behavior, analyzing multi-year trends, and integrating demographic or sentiment data for richer insights. Geospatial mapping can highlight regional patterns, while interactive dashboards can make the analysis more accessible to policymakers and the public. Additionally, anomaly detection techniques could support election transparency and data quality checks.

## References

* https://pandas.pydata.org/docs/user\_guide/index.html
* https://numpy.org/doc/stable/user/basics.html
* https://matplotlib.org/stable/index.html
* https://seaborn.pydata.org/