7PAM2000 APPLIED DATA SCIENCE 1

ASSIGNMENT 1: VISUALISATION

**Executive summary**

This research has been done using Python and Spyder considering the use of line, scatter, and histogram charts. The average number of seats gained by each party has been on the rise. It has been found that there is a strong correlation between victory probability and mean seats. The histogram shows a roughly balanced distribution of 225 seats. Timelines, connections between variables and distributions of data have been comprehended with the assistance of these illustrations. This type of study gives helpful information for making critical decisions regarding political forecasting.

**Table of Contents**

[Introduction 4](#_Toc150355554)

[Justification 4](#_Toc150355555)

[Implementation 4](#_Toc150355556)

[Data preparation 4](#_Toc150355557)

[Line Plot 8](#_Toc150355558)

[Scatter Plot 9](#_Toc150355559)

[Histograms 10](#_Toc150355560)

[Conclusion 11](#_Toc150355561)

[References 12](#_Toc150355562)

[Bibliography 12](#_Toc150355563)

[Appendix 13](#_Toc150355564)

[Appendix 1: Code 13](#_Toc150355565)

# Introduction

This research finds patterns in a dataset used for political forecasting, using line, scatter and histogram charts. Time series, relationships among variables and data distributions have been explored with the help of Python and Spyder. This has been helpful in demonstrating the complexity of the targeted dataset.

# Justification

This research work has been executed with the help of ***“Scatter, line and the histogram plot”***.

***Line plots***

Line graphs has been helped in identifying the patterns over a continuous variable. This chart is ideal for analysing the relationships between variables in a certain time frame. Connecting data points with lines aids in the identification of correlations and causal links, by highlighting trends, fluctuations, and patterns (Gehlen, 2020).

***Scatter plot***

The scatter plot has been known to be used to compare two continuous data. Dotted line graphs make it easy to see patterns, outliers, and trends in large amounts of data. This graph displays the strength and trend direction of associations between variables. The scatter plot's point dispersion can provide insight into data variability (Cox, 2023).

***Histogram***

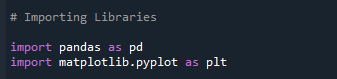
Histograms are used to display the distribution of a single variable. The histogram reveals the shape, central tendency, and dispersion of the data, by dividing the data into bins and quantifying the frequency or density of observations inside each bin. Modes, skewness, and normality can be identified with the assistance of histogram peaks, valleys, and symmetry (Wu *et al*. 2019).

Hence extensive data analysis is available via line, scatter, and histogram plots. Time trends can be seen in a line plot, relationships between variables can be shown in a scatter plot, and the distribution of a single variable can be seen in a histogram. Data patterns, correlations, and characteristics can be easily seen by analysts using these visualisations.

# Implementation

## Data preparation

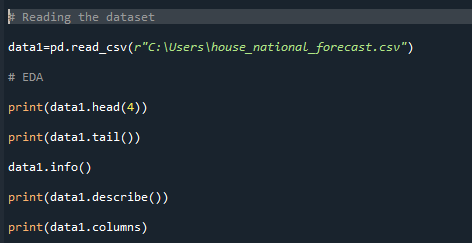
The selected visualization has been created with the help of Python programming language using the spyder. ***[Referes to appendix 1]***

****

#### Figure 1: Import Libraries

(Source: Spyder)

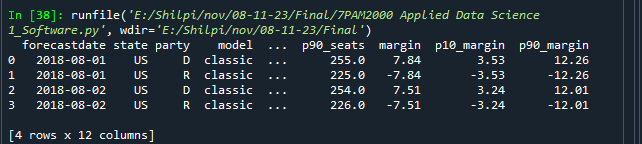
The above image shows the library files which need to be imported to perform this research. Data processing and visualisation libraries pandas and matplotlib.pyplot have been imported.

****

#### Figure 2: Reading the dataset and EDA

(Source: Spyder)

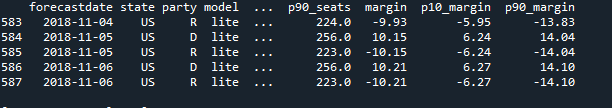
Using the 'pandas.read\_csv' command, the function retrieves a dataset from the specified position inside a file. The exploratory data analysis (EDA) phase begins with the presentation of the first four and last rows of the dataset by using the 'print(data1.head(4))' and 'print(data1.tail())' commands, respectively. The 'info()' method provides an overview of the structure of the dataset, while the 'description()' function provides statistical insights. 'print(data1.columns)' is the command that shows column names at the very end for further examination.



#### Figure 3: The first four rows of the selected dataset

(Source: Spyder)

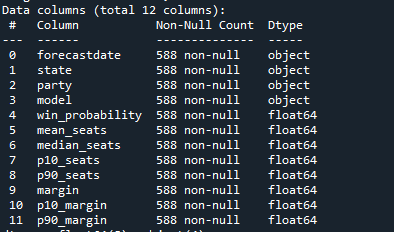
The above figure shows the first four rows that have been present on the selected dataset.



#### Figure 4: The last five rows of the selected dataset

(Source: Spyder)

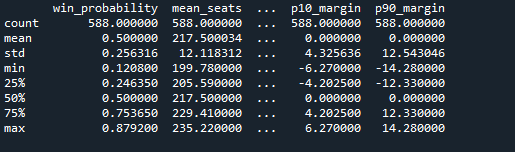
The last five rows of the chosen dataset have been contained in the above-shown figure.



#### Figure 5: Data information

(Source: Spyder)

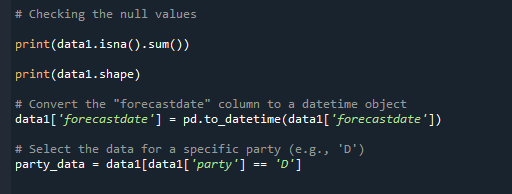
Based on the above figure it has been possible to identify the total number of non-null data columns and the corresponding data type continued the selected dataset.



#### Figure 6: Data description

(Source: Spyder)

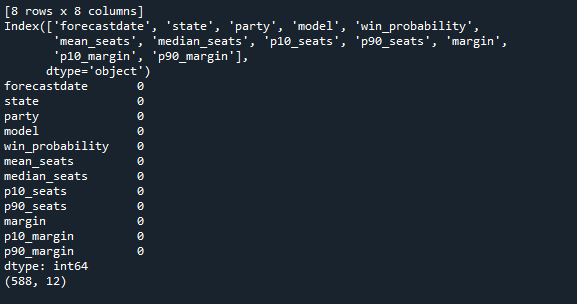
The overall summary of the dataset has been visualized in the above figure.

****

#### Figure 7: Checking Null Values

(Source: Spyder)

The dataset's shape is shown using the 'print (data1.shape)' and the code checks for null values using the ‘print(data1.isna().sum())’. The 'forecastdate' column is then transformed into a datetime object using the 'pd.to\_datetime()’ command. Then, the ‘party\_data’ variable is used to get data specific to a certain party ‘'D') for the sake of in-depth analysis and visualisation.

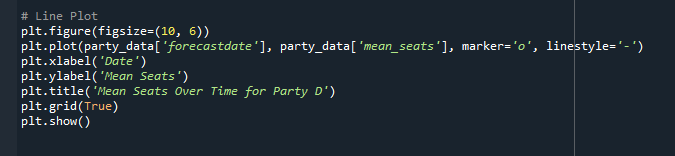


#### Figure 7: Null Values

(Source: Spyder)

Based on the above figure it can be said that the selected dataset does not contain any type of null value.

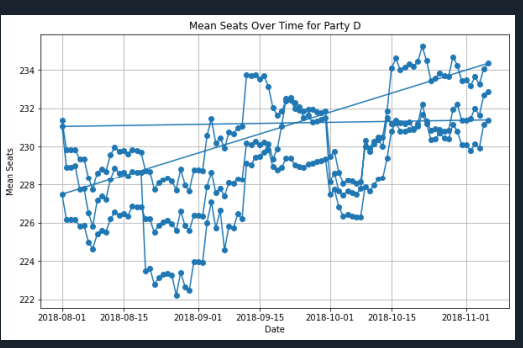
## Line Plot

****

#### Figure 8: Code of Line Plot based on Forecast Date and Mean Seats

(Source: Spyder)

The above figure describes the code for modelling a line plot based on two columns such as forecast date and mean seats of the chosen dataset. This line plot code is generated by using the Spyder programming IDE along with the plt of Python language. In order to get the line, plot the figure size is 10, 6 and the axis is labelled by the “Date” which carries “forecastdate” and the Y axis is labelled by “Mean seats” which carries “mean\_seats” values. The party\_data for this line plot is the “forecastdate” and “mean\_seats” and the title of this plot is mean seats over time for party D.

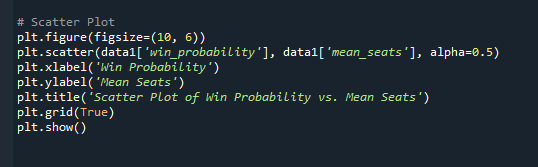
****

#### Figure 9: Line Plot of Mean Seats Over Time for Party D

(Source: Spyder)

The line plot of mean seats over time for Party D is explained by the above image. This plot is represented by the X and Y axis where the X-axis shows the Date and the Y-axis shows Mean Seats values based on the dataset. Based on the above line plot can say that the mean seat value for Party D is increased time by time which is described by the “trend line” of the graph.

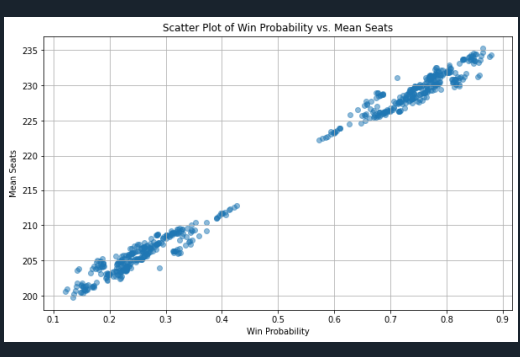
## Scatter Plot

****

#### Figure 10: Scatter Plot Code

(Source: Spyder)

The above image illustrates the code for creating a scatter plot based on the research dataset. In this plot figure size is 10, 6 and with the help of plt. scatter function, a scatter plot is created along with additional modifications in order to design d the plot. The data on the basis of the dataset such as data1 for this scatter are the “win\_probability” and “mean\_seats” along with the alpha size 0.5. The name of the X-axis is Win Probability and the Y-axis is Mean Seats. The title of this plot is a scatter plot of win probability vs. mean seats.

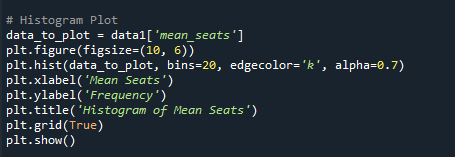
****

#### Figure 11: Scatter Plot of Win Probability vs. Mean Seats

(Source: Spyder)

The above figure shows the scatter plot of the win probability vs. mean seat column values. This scatter plot is described by the X and Y-axis where the X-axis represents the Win Probability whose maximum value is 0.9 and the Y-axis represents the Mean Seats whose maximum value is 235. Based on the data points of this plot which are coloured with a blue circle. can say that wind probability vs. meat sales has a strong relationship. These data points are maximum when the win probability value is o.2 and 0.8 and, in the middle, its value is 0.

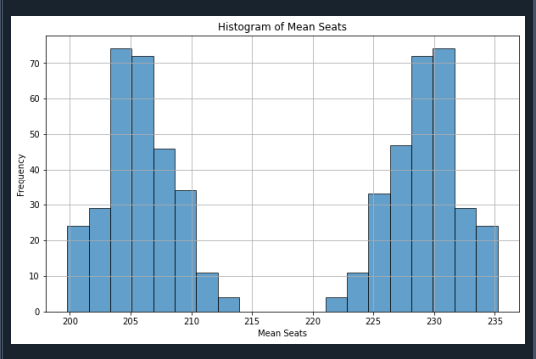
## Histograms

****

#### Figure 12: Code of Histogram Plot

(Source: Spyder)

The 'mean\_seats' column of the dataset is plotted as a Matplotlib histogram in the code. Histogram displaying mean seat data. It's a 10x6-inch figure. For clarity, the histogram has 20 bins, black borders ('k'), and 0.7 transparency. Histogram of Mean Seats, with Mean Seats on the x-axis and Frequency on the y-axis. Reading is facilitated by the grid. This visualisation depicts mean seat distribution, helping explain the dataset's central tendency and variability.

****

#### Figure13: Histogram Plot of Mean Seats

(Source: Spyder)

The above image shows the mean seats histogram plot which is represented by the X and Y axis. In this plot, the X-axis represents the Mean seats and the Y-axis represents the frequency. The histogram of expected seat values (‘mean\_seat’') has been found in the dataset. It has been found that the 225-seat threshold is the norm in political forecasting. There is a large amount of variation in the seat results between 220 and 235. Due to its symmetrical shape, the histogram is indicative of a normal distribution in which the projected frequencies are about equal on both sides of the central value.

There has been agreement or consistency in the predicted outcome if the number of seats is around 225, while uncertainty or disagreement can be reflected in the swings around that number. The collection seems to have a wide range of forecasts, which could improve political forecasting. The histogram is useful for spotting trends and outliers, by providing a fast and easy way to evaluate key characteristics of the dataset. This visualisation provides a visual summary of the expected mean seats and facilitates the exploration of political forecast factors for academics and analysts. The histogram makes the structure and critical aspects of the distribution clear, which aids political forecasters in understanding the data.

# Conclusion

Python and Spyder combine line, scatter, and histogram graphs to shed light on the nuances of a political forecasting dataset. The line plot indicates that the mean number of seats won by a party has increased over time. The scatter plot illustrates clustered patterns of winning probability and mean seat positions. The histogram shows a symmetrical distribution around 225 seats, indicating projected results unanimity. This intricate visual analysis sheds light on the dynamics of the dataset and enhances several aspects of political forecasting. The improved data knowledge that results from using these visualisations enables analysts to more easily detect patterns, outliers, and significant tendencies. Line, scatter, and histogram plots are three types of graphs that researchers and analysts can use to simplify political forecasting information and make more informed conclusions.

# References

Cox, N., 2023. AAPLOT: Stata module for scatter plot with linear and/or quadratic fit, automatically annotated.

Gehlen, M.H., 2020. The centenary of the Stern-Volmer equation of fluorescence quenching: From the single line plot to the SV quenching map. Journal of Photochemistry and Photobiology C: Photochemistry Reviews, 42, p.100338.

Wu, W., Federico, M.A.J. and Gan, S.K.E., 2019. The APD Skin Monitoring App for wound monitoring: image processing, area plot, and colour histogram. Scientific Phone Apps and Mobile Devices.

# Bibliography

McKinney, W., 2022. Python for data analysis. " O'Reilly Media, Inc.".

Mukhiya, S.K. and Ahmed, U., 2020. Hands-On Exploratory Data Analysis with Python: Perform EDA techniques to understand, summarize, and investigate your data. Packt Publishing Ltd.

Sahoo, K., Samal, A.K., Pramanik, J. and Pani, S.K., 2019. Exploratory data analysis using Python. International Journal of Innovative Technology and Exploring Engineering, 8(12), pp.4727-4735.

Stančin, I. and Jović, A., 2019, May. An overview and comparison of free Python libraries for data mining and big data analysis. In 2019 42nd International convention on information and communication technology, electronics and microelectronics (MIPRO) (pp. 977-982). IEEE.

Subasi, A., 2020. Practical machine learning for data analysis using python. Academic Press.

# Appendix

## Appendix 1: Code

|  |
| --- |
| ***“# Importing Libraries***  ***import pandas as pd***  ***import matplotlib.pyplot as plt***  ***# Reading the dataset***  ***data1=pd.read\_csv(r"C:\Users\house\_national\_forecast.csv")***  ***# EDA***  ***print(data1.head(4))***  ***print(data1.tail())***  ***data1.info()***  ***print(data1.describe())***  ***print(data1.columns)***  ***# Checking the null values***  ***print(data1.isna().sum())***  ***print(data1.shape)***  ***# Convert the "forecastdate" column to a datetime object***  ***data1['forecastdate'] = pd.to\_datetime(data1['forecastdate'])***  ***# Select the data for a specific party (e.g., 'D')***  ***party\_data = data1[data1['party'] == 'D']***  ***# Line Plot***  ***plt.figure(figsize=(10, 6))***  ***plt.plot(party\_data['forecastdate'], party\_data['mean\_seats'], marker='o', linestyle='-')***  ***plt.xlabel('Date')***  ***plt.ylabel('Mean Seats')***  ***plt.title('Mean Seats Over Time for Party D')***  ***plt.grid(True)***  ***plt.show()***  ***# Scatter Plot***  ***plt.figure(figsize=(10, 6))***  ***plt.scatter(data1['win\_probability'], data1['mean\_seats'], alpha=0.5)***  ***plt.xlabel('Win Probability')***  ***plt.ylabel('Mean Seats')***  ***plt.title('Scatter Plot of Win Probability vs. Mean Seats')***  ***plt.grid(True)***  ***plt.show()***  ***# Histogram Plot***  ***data\_to\_plot = data1['mean\_seats']***  ***plt.figure(figsize=(10, 6))***  ***plt.hist(data\_to\_plot, bins=20, edgecolor='k', alpha=0.7)***  ***plt.xlabel('Mean Seats')***  ***plt.ylabel('Frequency')***  ***plt.title('Histogram of Mean Seats')***  ***plt.grid(True)***  ***plt.show()***  ***# Pie Charts***  ***total\_mean\_seats = party\_data['mean\_seats'].sum()***  ***plt.figure(figsize=(8, 8))***  ***plt.pie(party\_data['mean\_seats'], labels=party\_data['forecastdate'], autopct='%1.1f%%')***  ***plt.title('Distribution of Mean Seats for Party D')***  ***plt.axis('equal')***  ***plt.show()”*** |