ELECTIONEERING BUILD-UP AND ELECTION OUTCOMES

JESUBUKADE AJAKAYE

Abstract

Favourable electoral outcomes are only achievable with proper planning and formidable strategy development. Data science and analytics has gained popularity among electoral stakeholders and look to gain insights of their performance going into an elections. This study aimed to predict election outcomes through campaign finances. Preliminary findings revealed that higher electoral campaign finance lead to positive electoral outcomes and candidates with longer campaign days had a better chance of winning the election. The K-Nearest Neighbour algorithm was deployed to train the model which had a 95% model accuracy with AUC = 0.94. This study showed the ability of Data Science in real life applications which can aid preparations and lead to enhanced output or improve decision making.

Keywords: Campaign Finance, Data Science, Decision Making, Electoral Outcome, K-Nearest Neighbour

Table of Contents

List of Tables

List of Figures

Introduction

Elections have formed part of our democracy, almost anyone can show their intentions to vote or be voted for as long as the law allows. The electioneering process is about drawing up the best plan and developing the right strategies to triumph at the polls. Several factors influence the election outcome, as candidates canvass for votes during the campaign periods which require a lot of planning and take up bulk of finances which might come from personal funds, party funds or other contributions which includes lobbying stake holders in the country or region for financial backing. The application of data science and analytics has been a game changer and has been adopted several times to determine election outcomes. Unfortunately, only one person can win any contested seat therefore this study seeks to evaluate the possibility of a candidate winning using data science processes.

Motivation

Every electoral candidate registers their interest in the contest for a public position based on perceived support and capacity. They believe they have a chance of coming out on top. Previous elections worldwide have shown that paying attention to data has always played a major role in determining election outcomes. Several factors affect the outcome which include public perception, ideologies, and campaign strategies, among others. This study seeks to predict election outcomes through the build-up processes which include electioneering finance, campaign days and other socio-demographic factors.

Literature Review

Various research has looked at the US 2016 election at various angles and various factors as predicting the election outcomes [1], [2]. A recent study [3] carried out exploratory data analysis to show the relationship between election finance and election outcome using the data set. [4] stated that 1% of U.S. top earners dominated the Democratic and

the Republican parties which make the U.S. majorly a two-party system where total expenditure during the campaign has a linear relationship with votes scored during election. The study found that finances showed a relationship with getting high votes in elections and that money remains a powerful tool that influences election results. They further reported that Trump's election in the 2016 been studied was declared doom at some point during the campaign but a dramatic generation of funds switched the outcome of the elections for the senatorial and consequent impact on Trump's presidential campaign.

[5] opined that candidates benefit from incumbency in the U.S. especially in multimember district systems but decrease as the district magnitude grows. Also, other factors such as policies, service to constituency, developing reputation with "for show" projects and office perks can place an incumbent at an advantage going into elections. [6] also affirmed that there was a strong relationship between campaign financing and the results of congressional election but further observed that this relationship is dependent on the status of the candidate going into the elections. The more the challengers spent, the better their chances of winning in the elections while the more the incumbents spent, the worse the election outcome for them. Which could be that incumbents were spending bigger to cover their poor performances in office but were usually not saved by such jamboree.

These findings are also confirmed by [7] which studied how campaign financing affects winning chances in the US House of Representatives elections between 2000 and 2018 in all 50 states and Washington D.C. The study reported that campaign contributions and expenditures positively influence electoral outcomes while this is not very applicable to incumbents as electoral financing is less effective than for challengers. [8] used random forest model to predict the impact of campaign finance on congressional voting with more than 90% accuracy.

Technical Implementation

This section discusses the methodology/data science process, as shown in Figure 1, carried out during this research.

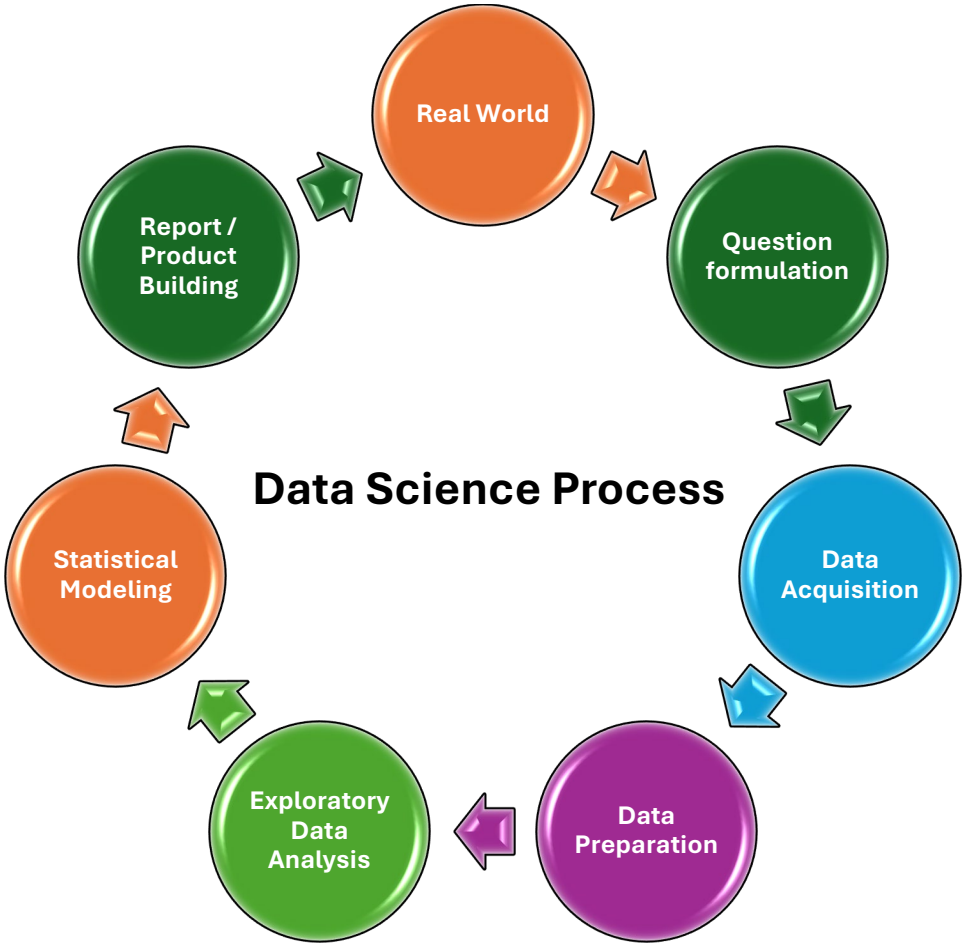

Figure 1: Data Science Process flow

Sources: Student resources from lecture notes

Data Information

The dataset was retrieved from Kaggle, and more information on the dataset was retrieved from the United States of America Federal Election Commission Official website (see Appendix A for links). The dataset is about the 2016 US General Election outcomes for the Presidential, Senatorial and House. The dataset has 51 columns as shown on Table 1.

Table 1: Data Description

S/N	Column Name	Description	Data type
1.	can_id	Identity number of the candidate in the US election	Nominal (AlphaNumeric)
2.	can_nam	Name of Candidate	Nominal
3.	can_off	Office the candidate is contesting for. P – Presidential, S – Senatorial, H – House of Representatives	Categorical
4.	can_off_sta	The State for the office the candidate is contesting for	Nominal
5.	can_off_dis	The district for the office the candidate is contesting for	Nominal
6.	can_par_aff	Candidate party affiliation	Categorical
7.	can_inc_cha_ope_sea	The status of the candidate in the election, OPEN seat, Incumbent or Challenger	Categorical
8.	can_str1	Candidate Address	String (Char)
9.	can_str2	Candidate Address	String (Char)
10.	can_cit	Candidate City	Nominal
11.	can_sta	Candidate State	Nominal
12.	can_zip	Address postal code	Numerical
13.	ind_ite_con	Individual itemised contributions	Numerical
14.	ind_uni_con	Individual unitemised contributions	Numerical
15.	ind_con	Total individual Contributions	Numerical
16.	par_com_con	Contributions from party committees	Numerical
17.	oth_com_con	Contributions from other committees	Numerical
18.	can_con	Candidate contributions	Numerical
19.	tot_con	Total contributions	Numerical
20.	tra_fro_oth_aut_com	Transfers from other authorised committees	Numerical
21.	can_loa	Candidate loans	Numerical
22.	oth_loa	Other loans	Numerical
23.	tot_loa	Total loans	Numerical
24.	off_to_ope_exp	Offsets to operating expenditures	Numerical
25.	off_to_fun	Offsets to fundraising expenditures	Numerical
26.	off_to_leg_acc	Offsets to legal and accounting expenses	Numerical
27.	oth_rec	Other receipts	Numerical
28.	tot_rec	Total receipts	Numerical

29.	ope_exp	Operating expenditures	Numerical
30.	exe_leg_acc_dis	Expenditures for legal/accounting purposes	Numerical
31.	fun_dis	Fundraising disbursements	Numerical
32.	tra_to_oth_aut_com	Transfers to other authorised committees	Numerical
33.	can_loa_rep	Candidate loan repayments	Numerical
34.	oth_loa_rep	Other loan repayments	Numerical
35.	tot_loa_rep	Total loan repayments	Numerical
36.	ind_ref	Individual refunds	Numerical
37.	par_com_ref	Party committee refunds	Numerical
38.	oth_com_ref	Other committee refunds	Numerical
39.	tot_con_ref	Total contributions refunds	Numerical
40.	oth_dis	Other disbursements	Numerical
41.	tot_dis	Total disbursements	Numerical
42.	cas_on_han_beg_of_per	Cash on hand at the beginning of the period	Numerical
43.	cas_on_han_clo_of_per	Cash on hand at the end of the period	Numerical
44.	net_con	Net contributions	Numerical
45.	net_ope_exp	Net operating expenditures	Numerical
46.	deb_owe_by_com	Debt owed by the committee	Numerical
47.	deb_owe_to_com	Debt owed to the committee	Numerical
48.	cov_sta_dat	Campaign start date	Datetime
49.	cov_end_dat	Campaign end date	DateTime
50.	winner	Outcome of the election for the candidate Y – Win, N - Lost	Categorical
51.	votes	Number of votes for the candidate	Numerical

Data Preparation and Cleaning

The data was prepared for analysis and modeling both on excel and using R programming language.

Excel – there activities were completed on excel to ensure that the data was ready to be processed on R programming language.

i. The value “N” was inputted in the winner column for candidates that lost in the election as these cells were empty before,

ii. The “$” sign and “,” were removed from the numerical columns as R was interpreted these fields as character and there was a slight difficulty casting them

to numerical values.

iii. Dataset was studied and candidates with campaign dates earlier than 6 months before the elections (or that lost at primaries) were dropped.

R Programming Language – after importing the CSV file, some activities were also completed as outline below:

i. Columns such as names, address and numerical columns marked as not required either because have been transformed to form new columns in the dataset or with very high percentage of missing values were dropped.

ii. Also, rows with missing values in the remaining columns were also removed.

iii. New calculated columns were created such as getting the actual campaign days for each candidate and the difference between net contribution and net expenditure of each candidate and previous columns not required again dropped.

iv. Columns with a long character of name renamed to a shorter on convenient name.

v. Casting of dataset was completed as categorical columns were factored.

Data Visualisation

Data Summary

Table 2: General statistics of dataset

	Net Contribution	Net Expenditure	Net Difference	Campaign Days
Min.	$10.00	$1.80	-$455,209,124.20	22 days
1st Qu.	$24,627.07	$31,436.57	-$5,524.28	292 days
Median	$ 215,364.44	$233,787.67	$2,899.68	542 days
Mean	$3,025,239.43	$3,217,854.87	-$192,615.44	470 days
3rd Qu.	$1,119,718.06	$837,293.34	$174,899.03	657 days
Max.	$2,526,103,377.00	$2,466,802,358.00	$59,301,019.00	841 days

Table 3: Spread of Candidates Matrix

winner	Candidate Office	H	P	S
Lost	Frequency	821	6	144
	Percentage	56.9%	0.4%	10.0%
	Row Percentage	84.6%	0.6%	14.8%
	Column Percentage	65.4%	85.7%	80.4%
Elected	Frequency	435	1	35
	Percentage	30.2%	0.1%	2.4%
	Row Percentage	92.4%	0.2%	7.4%
	Column Percentage	34.6%	14.3%	19.6%

Explorative Data Analysis (EDA)

This is an important aspect of data modelling which gives insight into the dataset under study. The researcher and readers are able to have firsthand understanding of the features and targets in the dataset through visualisation. A grouped bar chart was mostly used in this study to compare the net contribution, net expenditure and difference between both with various categories based on election outcome.

General Analysis

Election Outcome

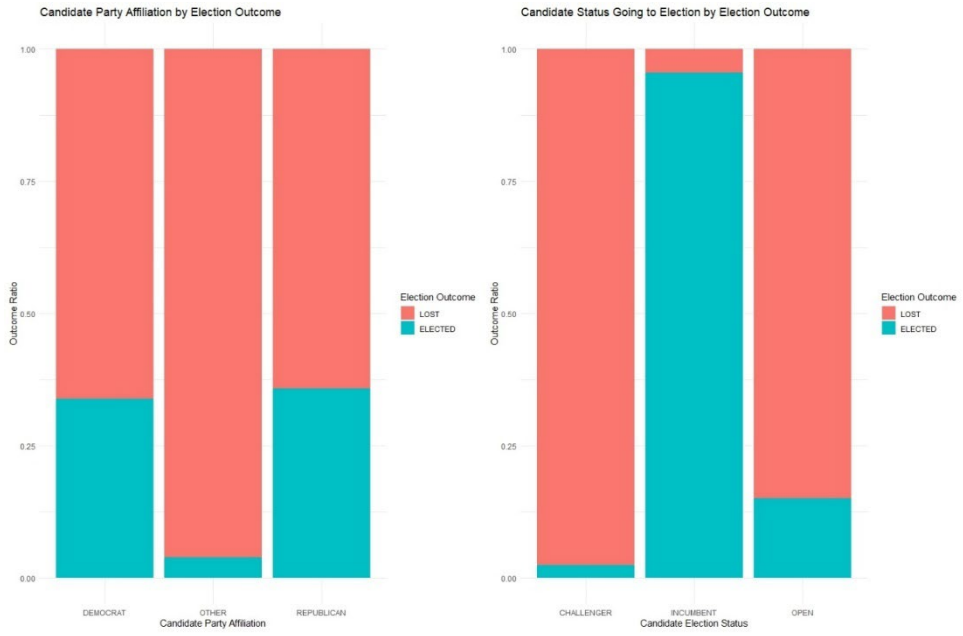

Figure 2: Candidate party affiliation and Election status by Election outcome

Findings from *Figure 2* revealed that Democrats and Republicans have about the same ratio of winning in an election and stand a better chance than candidate from other parties or individual candidates which showed that United States is majorly a two-party system. For Status going to election, Incumbent candidates stand a better chance of winning reelection than challengers.

Party Affiliation and Election Outcome

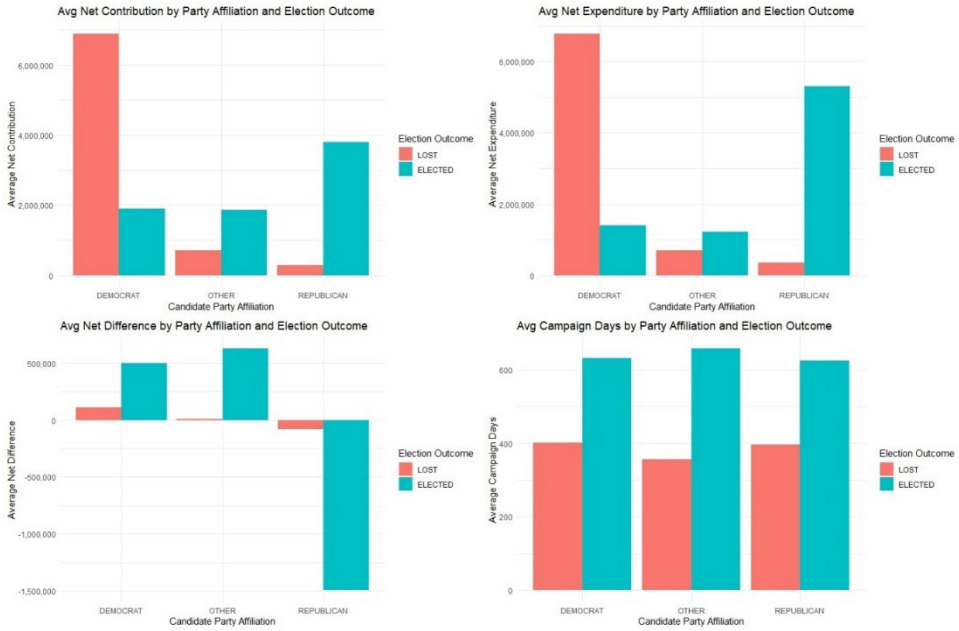

Figure 3: Average Net Contribution, Average Net Expenditure, Average Net Difference and Campaign days by Party Affiliation and Election Outcome

Findings revealed that Republican Candidates got more average contributions and spent more towards the elections and showed in their candidates getting elected more. Also, the left bottom showed that Republicans are more risk takers. They have a negative net average net difference which showed that most of their campaigns were ran on debt, but they won more sit in the elections when compared. The right bottom showed that there is no visible difference between the average campaign days spent by the 3 groups and election outcomes. Those that spent more days in the 3 groups won in their elections.

Election Status and Election Outcome

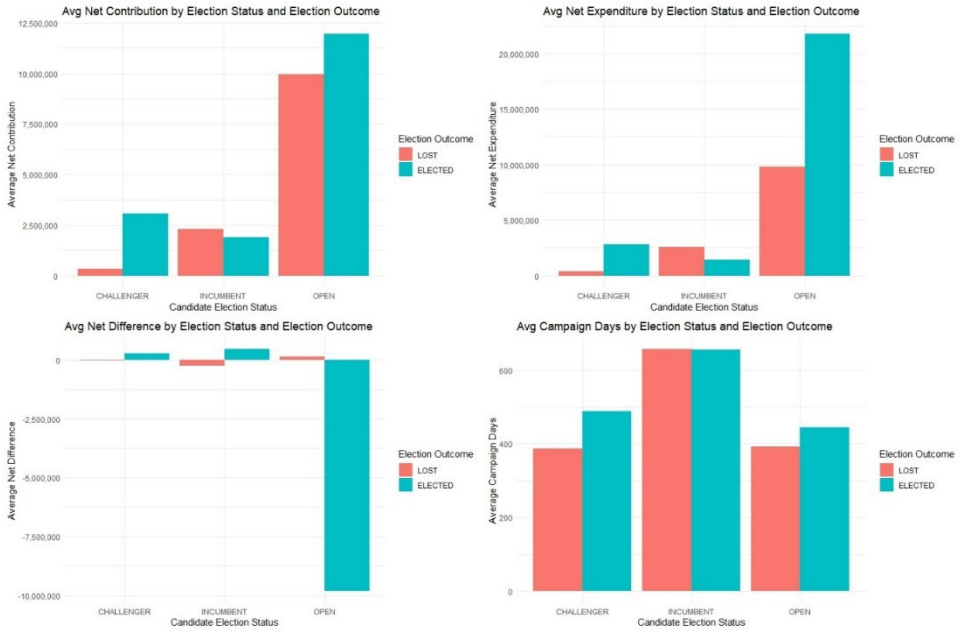

Figure 4: Average Net Contribution, Average Net Expenditure, Average Net Difference and Campaign days by Election Status and Election Outcome

Figure 4 revealed that the open seat candidate generated more contribution and spent more during the elections. This is more because of the presidential seat which was open, and the candidates were heavy spenders going into the elections. It is only in the incumbent category that the highest average contribution and expenditure has majority that lost the elections. The incumbent is able to gather more resources and have longer campaign days than the challenger and open candidates. There was visible difference between campaign days of the challenger and open candidates on election outcomes.

Presidential Election Analysis

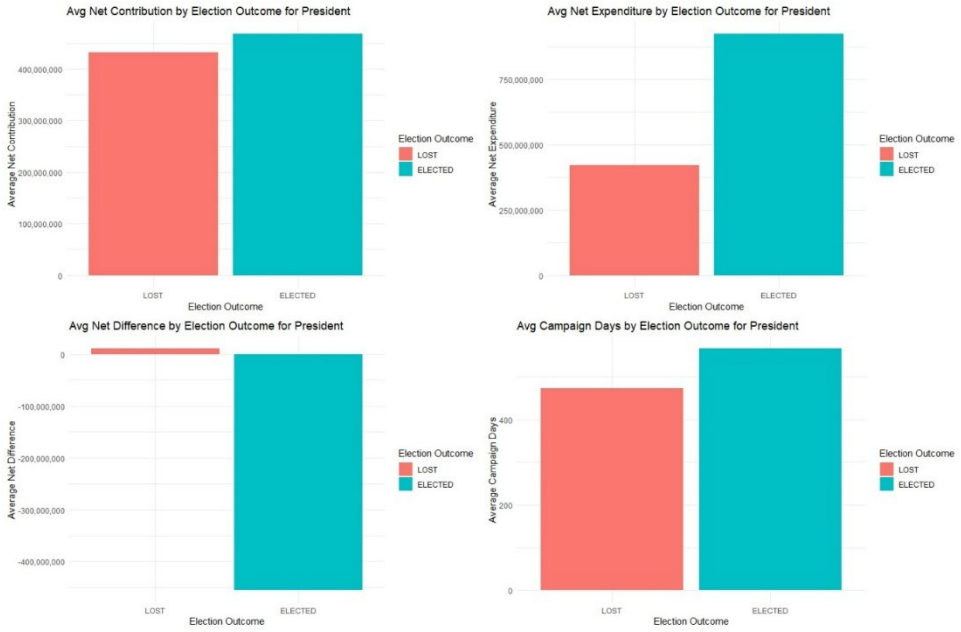

Figure 5: Average Net Contributions, Net Expenditure, Net Difference and Campaign days by Election Outcome

The elected candidate got more contribution and spent more on the average and recorded negative difference as debt going to the election. Also, the elected candidate recorded more campaign days on average when compared to unelected candidates.

Senatorial Election Analysis

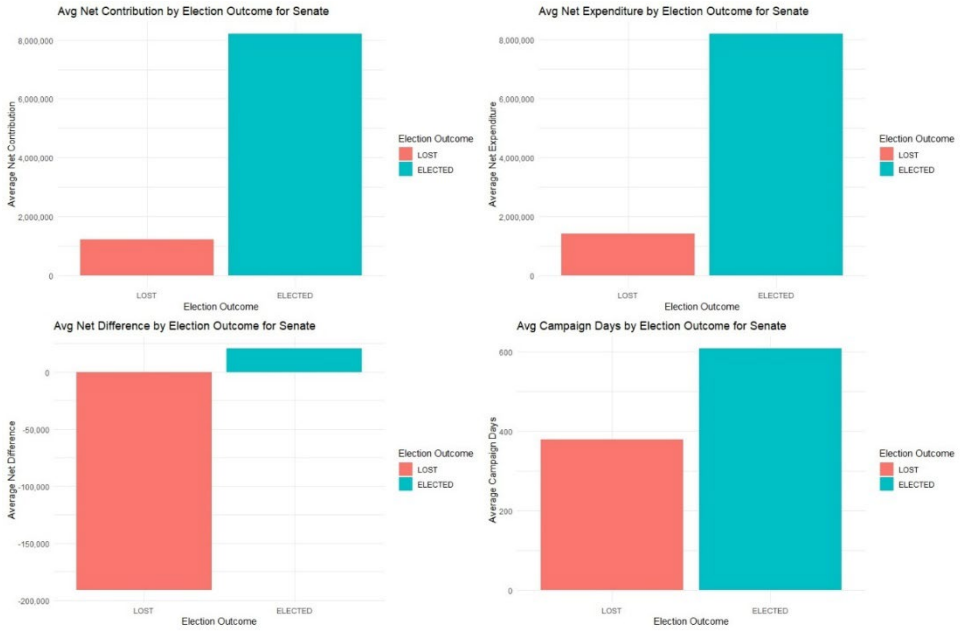

Figure 6: Average Net Contributions, Net Expenditure, Net Difference and Campaign days by Election Outcome

The elected candidates had more average net contributions, spent more average net expenditure and recorded longer campaign days than candidates that lost. Candidates that Lost had a negative net difference while Elected candidates had a positive net difference.

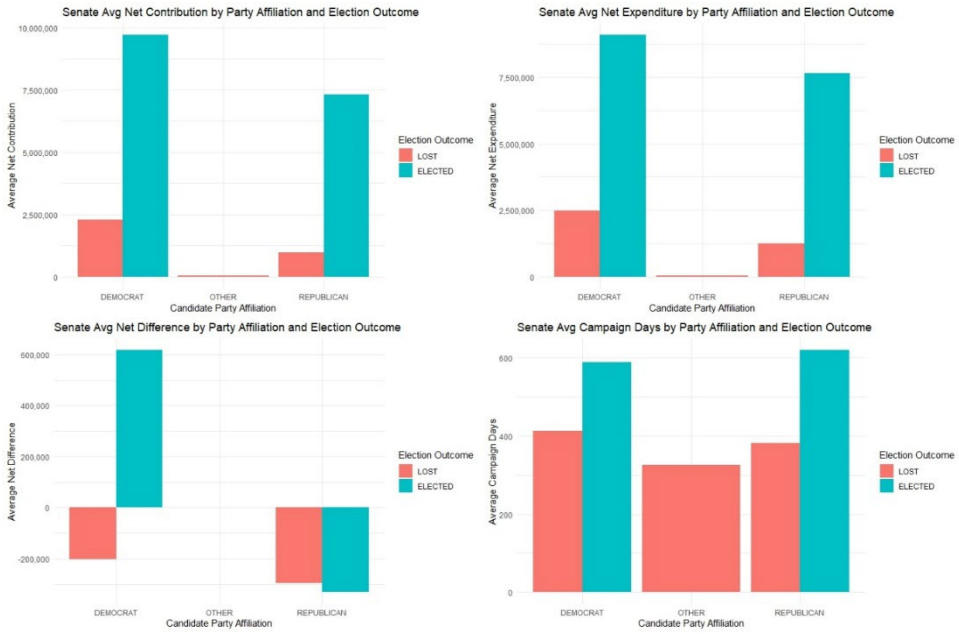

Figure 7: Average Net Contribution, Average Net Expenditure, Average Net Difference and Campaign days by Party Affiliation and Election Outcome

Democrats in the senatorial election had more net contribution and more net expenditure. This visual also revealed that higher average net contribution and net expenditure increase chances of getting elected. Also, longer campaign days increase chance of reelection.

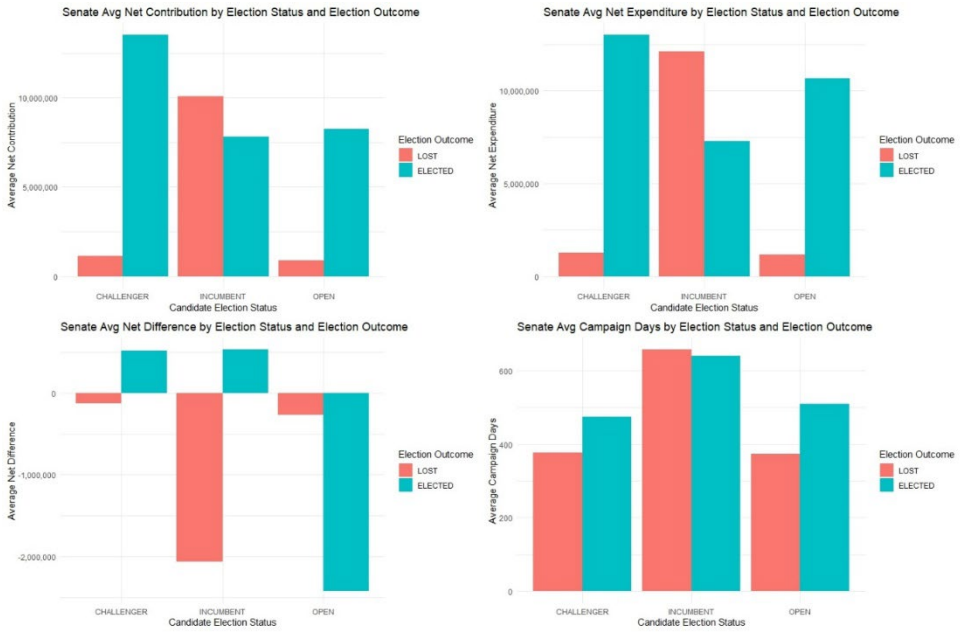

Figure 8: Average Net Contribution, Average Net Expenditure, Average Net Difference and Campaign days by Election Status and Election Outcome

Incumbents with more average net contribution and net expenditure lost reelection which may be due to other factors not considered under this study.

House of Representatives Election Analysis

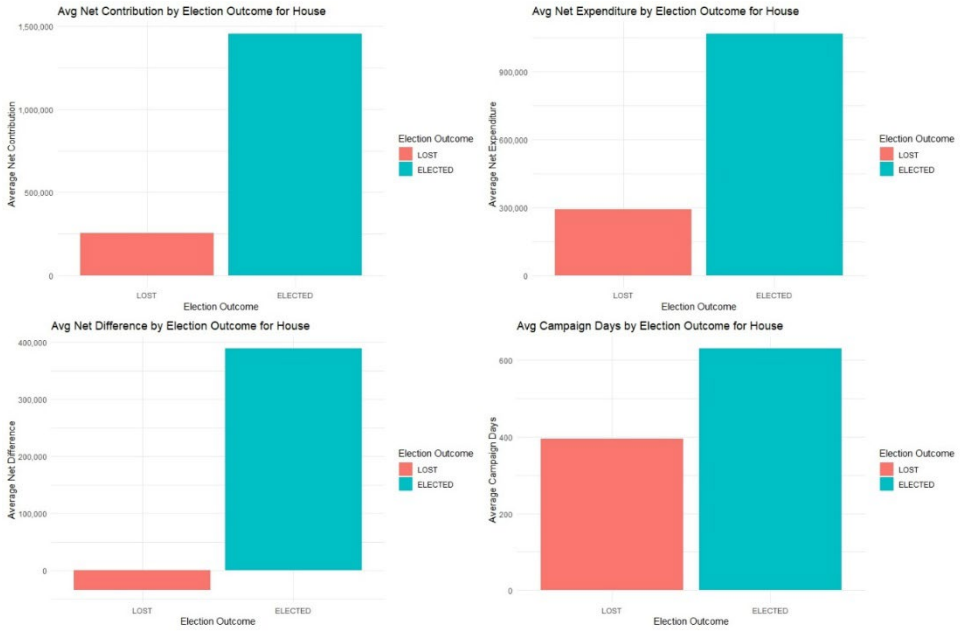

Figure 9: Average Net Contributions, Net Expenditure, Net Difference and Campaign days by Election Outcome

This finding also supports the findings in previous visualisations where getting more

contributions, spending more money and longer campaign days increases chances of getting elected in the US election.

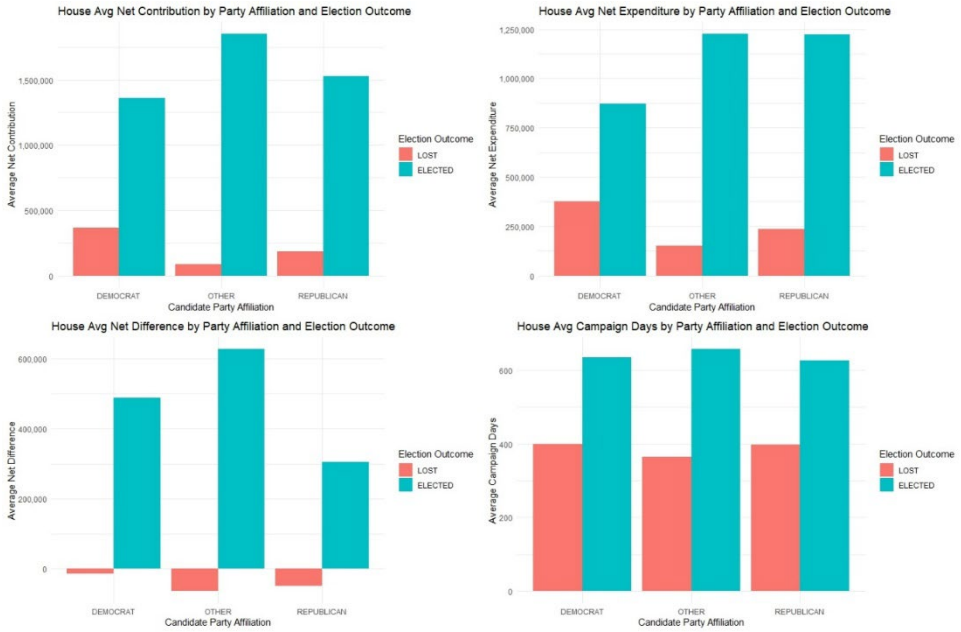

Figure 10: Average Net Contribution, Average Net Expenditure, Average Net Difference and Campaign days by Party Affiliation and Election Outcome

This finding revealed that negative net difference in the House of Representative election may lead to losing election. This is because candidates with negative net difference are mostly low spenders as shown in average net contribution and average net expenditure in *Figure 10*.

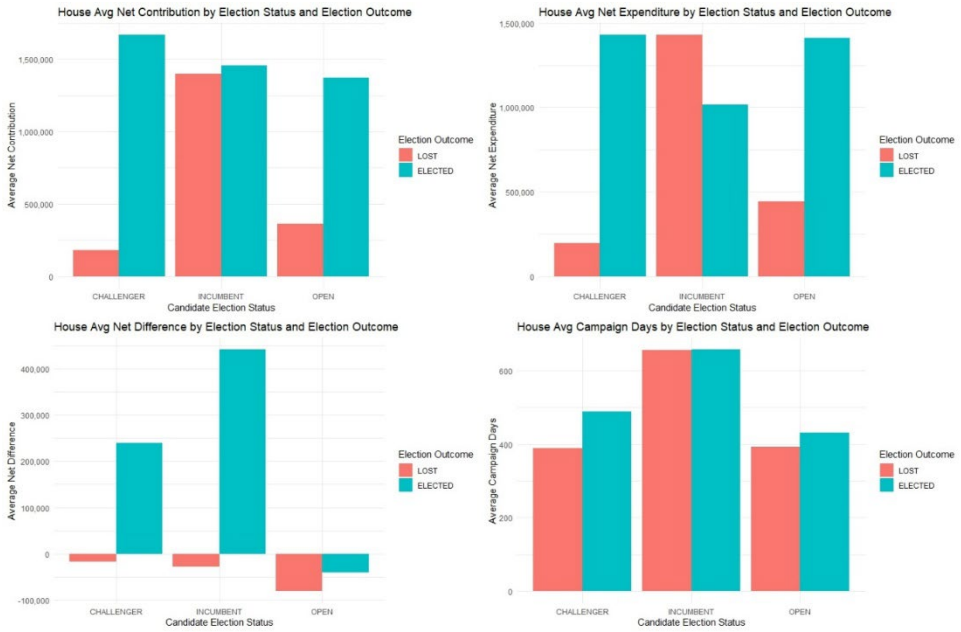

Figure 11: Average Net Contribution, Average Net Expenditure, Average Net Difference and Campaign days by Election Status and Election Outcome

Findings from this visual also revealed that funding and campaign days alone may not effectively predict reelection of incumbents as seen also in the Senatorial elections.

Data Preprocessing and Correlation Analysis

Outliers

Outliers were checked for in the dataset using Z-score. Numeric columns were scaled to ensure that accurate results were generated, and a Z-score function was applied. A Z-score below -3.29 or above 3.29 was used to determine outliers as explained in [9].

Table 4: List of Outliers in dataset

	can_of f	can_of f_sta	can_p ar_aff	can_st atus	net_con	net_ope_exp	win ner	net_diff	camp_ days
1	P	US	REP	OPEN	$468,441,873.4	$923,650,997.6	Y	-$455,209,124.2	566
1493	P	US	DEM	OPEN	$2,526,103,377.0	$2,466,802,358.0	N	$59,301,019.0	567

Table 4 reveals that the two key presidential candidates were outliers in the dataset who are also the biggest spenders. These candidates are important parts of the datasets therefore they are not removed from the dataset.

Population Proportion/Population Sample

The population proportion of the target was checked to determine if oversampling or under sampling will be carried out on the dataset.

Table 5: Spread of target outcomes in the dataset

winner	count	props
N	971	0.67
Y	471	0.33

The portion of about 67:33 is considered appropriate for modeling to reflect real life scenario as only one candidate can emerge for any position contested for by at least 2 candidates.

Correlation Analysis

As part of the preprocessing steps, categorical variables were encoded to numerical values. This is particularly for model build. The target variable was encoded for Y (elected class) to be 0 and N (not elected class) to be 1. All the features were the standardise using the scaling function in R.

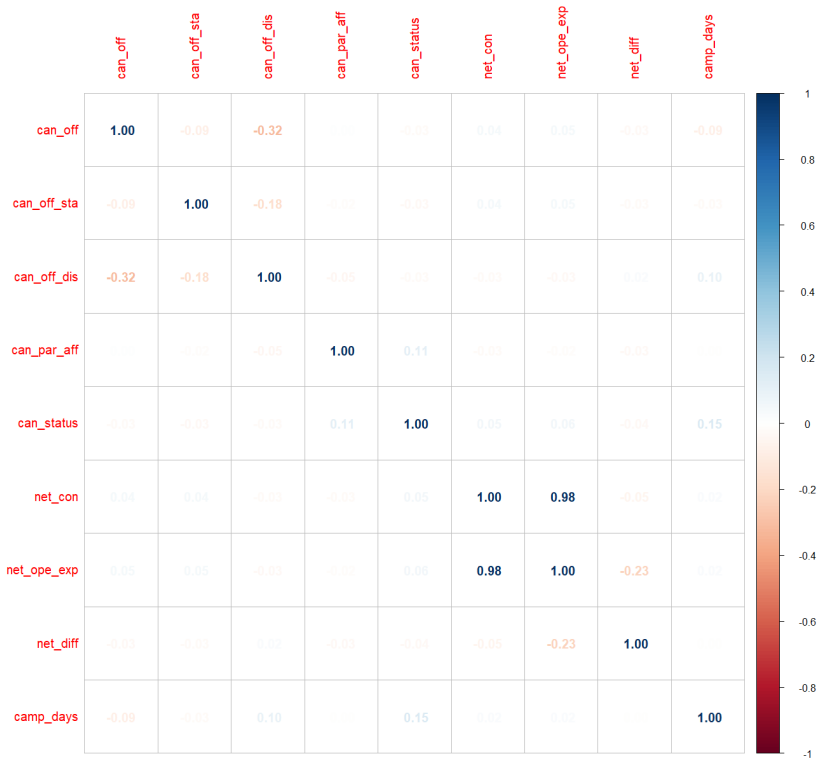

Figure 12: Correlation matrix

The correlation function was used to develop the correlation matrix in *Figure 12*. this showed there is a positive correlation between net contribution and net expenditure.

Table 6: Alias of correlated values

Model: winner ~ can_off + can_off_sta + can_off_dis + can_par_aff + can_status + net_con + net_ope_exp + net_diff + camp_days									
Complete:									
	(Intercept)	can_off	can_off_sta	can_off_dis	can_par_aff	can_status	net_con	net_ope_exp	camp_days
net_diff	0	0	0	0	0	0	1	-1	0

A test of multicollinearity was carried out using the Variance Inflation Factor (VIF) on the Generalised Linear Model (GLM) function. The first test for multicollinearity returned error "*there are aliased coefficients in the model*" which meant there were highly correlated values(columns) in the model. The alias function was used on the (GLM) function to further investigation of these correlated values and net difference was found to correlate with net contribution and net expenditure (see *Table 6*). Therefore, two (net contribution and net expenditure) of the three correlated values were dropped and the test of multicollinearity was carried out again.

Table 7: VIF values of Multicollinearity test

Column	can_off	can_off_sta	can_off_dis	can_par_aff	can_status	net_ope_exp	camp_days
Value	1.107151	1.076777	1.160339	1.014296	1.202369	1.016872	1.187294

Values from Table 7 showed that the remain columns have moderate multicollinearity (VIF < 5) therefore, these columns are used for training of the model.

Machine Learning Procedures

Supervised learning was the approach for the model training and a classification method is appropriate as the features are labelled which can help classify into the target classes. K-Nearest Neighbor was preferred as it was computationally inexpensive and easy to adopt. Checking for missing date returned "FALSE" and target variable was factored to ensure that the data was ready for the model building. Seed was set at 42 and the train-test split was carried out at an 80:20 ratio to ensure that enough data was available for

training and test.

An iterative process of training was carried out to determine the most effective number of k (neigbours) in the model and k was set at 3, 5, and 7. Training was performed and k = 5 was the best performing model therefore a training was performed to save the best performing model for future use.

Performance Evaluation

Evaluating the trained model helps to gauge the model's ability on dataset it has not been exposed to before. There are several ways of model performance evaluations, but this report used model accuracy as well as the ROC and AUC plots for performance evaluation.

Model Accuracy

The model has an accuracy of 95.14% which showed that the model performed well on the test data. The model also had a few False Positive (high specificity) which shows precision in the model performance.

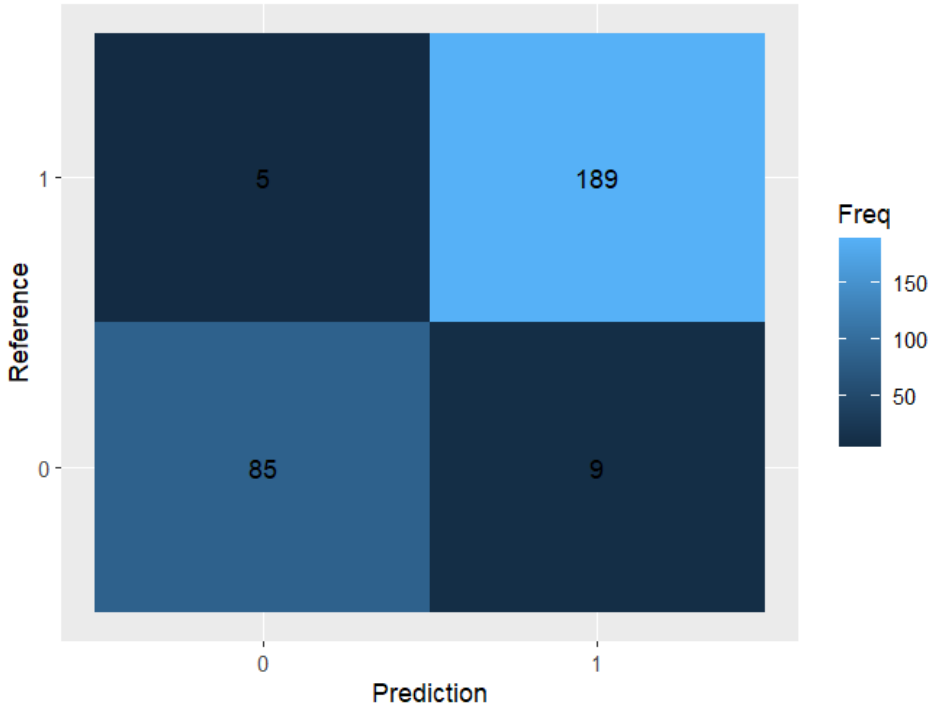

Figure 13: Confusion Matrix

Table 8: Confusion Statistics

Accuracy	0.9514
95% CI	(0.9198, 0.9732)
No Information Rate	0.6736
P-Value [Acc > NIR]	<2e-16
Kappa	0.8882
Mcnemar's Test P-Value	0.4227
Sensitivity	0.9043
Specificity	0.9742
Pos Pred Value	0.9444
Neg Pred Value	0.9545
Prevalence	0.3264
Detection Rate	0.2951
Detection Prevalence	0.3125
Balanced Accuracy	0.9392
'Positive' Class	0

ROC and AUC plots

To investigate further the model's performance, a Receiver Operator Characteristic (ROC) curve was plotted. The Area Under the Curve (AUC) = 0.94 affirming the high specificity and sensitivity of the model.

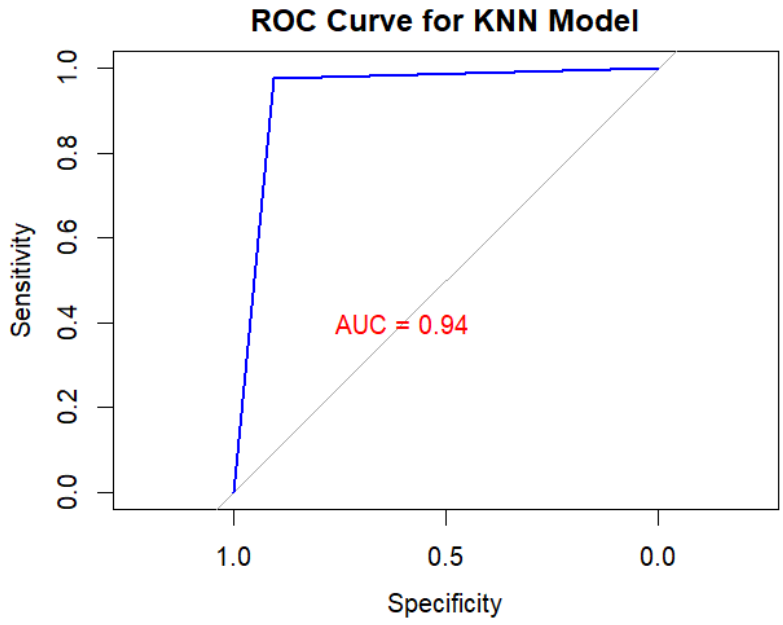

Figure 14: ROC Curve for the trained model

Findings and Technical Discussions

Findings from the study revealed that the U.S. is a 2-party system, and you have higher chances of winning in both which is confirmed by [4]. Major stakeholders are either in the Democratic party or the Republican party based on ideologies and this has been evident over centuries. Consequently, the analysis showed that the biggest spenders win elections (also, showed in the risk taking of republicans). [4-7] affirmed that campaign financing plays a major role in electoral outcomes as high financing increases the effect the campaign has in swaying voters and recording high votes at the polls.

Furthermore, the longer the campaign days the higher the chance of getting a positive electoral outcome. Infact, this can be said to be a snowballing effect from campaign financing. Candidates are able to plan more programs and have extended campaign days if required funds are available to support necessary action plans.

Incumbents showed more chances of winning reelection which is not largely dependent on funding and election campaign days this finding is corroborated by [5-7]. While it may be easier for incumbents to gain a favourable outcome in elections, studies have shown that this is dependent on performance appraisal by the voters. Finally, the model developed in this study is able to predict election outcomes.

Conclusion

Data science and analytics is an indispensable tool in our society today. The power data insights hold is unending and has a proper application that can predict outcomes with a high level of confidence. Participating in an election and achieving success is a factor of adequate preparation in various aspects which data science can be applied to all, not only in campaign financing. The overall effect of the campaign on the voters determines the electoral outcome at the end.

References

[1] Z. He, J. Camobreco and K, Perkins, "How he won: Using machine learning to understand Trump's 2016 victory," *Journal of Computational Social Science, Springer*, vol. 5, no. 1, pp 905-947, May 2022. doi: 10.1007/s42001-021-00147-3

[2] M. C. Karagosian, "From War Chests to Internet Fundraising: How Campaign Finances Influence Presidential Election Outcomes," *The Macksey Journal*, vol. 2, Article 130, 2021

[3] M. O'Brien (2022, May 2). "*Relationship between campaign finance and election results*". NYC Data Science Academy [Online]. Available: https://nycdatascience.com/blog/student-works/relationship-between-campaign-finance-and-election-results/ [Accessed: 14 November 2024].

[4] T. Ferguson, P. Jorgensen and J. Chen, "How money drives US congressional elections: Linear models of money and outcomes" *Structural Change and Economic Dynamics*, 61, pp. 527-545, Sept. 2019.

[5] K. Ariga, "When do political parties benefit from incumbents' personal votes? Comparative analysis across different electoral systems" *Electoral Studies*, 68, pp. 1-14, Aug. 2020. https://doi.org/10.1016/j.electstud.2020.102221

[6] G. C. Jacobson, "Measuring campaign spending effects in U.S. House Elections" in *Capturing Campaign Spending Effects*, H. Brady and R. Johnston, Eds. Ann Arbor: Michigan Press, 2006, pp. 199-220.

[7] T. Le, I. Onur, R. Sarwar and E. Yalcin, "Money in politics: How does it affect election outcomes" *SAGE Open*, vol. 14, no. 4, pp. 1-14, Oct.-Dec. 2024. DOI: 10.1177/21582440241279659

[8] M. Lalisse, "Measuring the Impact of Campaign Finance on Congressional Voting: A Machine Learning Approach" Working Paper 178, Institute for New Economic Thinking, Feb. 2020, https://doi.org/10.36687/inetwp178

[9] A. Soetewey, (2020, Aug 11). *Outlier detection in R*. Stats and R [Online]. Available: https://statsandr.com/blog/outliers-detection-in-r/ [Accessed: 14 November 2024]

Appendix

Appendix A: Related Links

Data Link https://www.kaggle.com/datasets/danerbland/electionfinance

Data Information Link https://www.fec.gov/data/browse-data/?tab=bulk-data