Analysis of the Spoilt Votes in Kenya's 2013 Presidential Elections

By

Herman Githinji Mwangi

www.github.com/jojherman

Junior Researcher
Berlin, Germany

October, 2023

Abstract

This study explores factors influencing spoilt votes in Kenyan constituencies during the 2013 presidential elections. The total number of spoilt votes were **108,700**. It addresses a knowledge gap of how registered voters (**14,292,665**), cast votes (**12,277,194**), voting patterns (**'Landslide' or 'Close Race'**) influence the number of spoilt votes as well as the implications of the spoilt votes on constituency development fund (CDF) allocation (**KES: 138 Billion**). This data represents **289** out of the total 291 constituencies. The primary goal is to unveil these factors' impact on spoilt vote numbers, the effect of spoilt votes on CDF allocation, with broader implications for election integrity and policy in Kenya. The study employs quantitative analysis, focusing on linear regression, to provide empirical evidence. In conclusion, it contributes vital insights to election dynamics and the Kenyan electoral landscape.

Introduction

This study aims to contribute to a better understanding of the Kenyan electoral landscape and provide data-driven recommendations for policymakers to bolster the integrity of future elections in the country. These factors play a crucial role in shaping the electoral landscape, as they can impact the overall integrity of the election process. Additionally, the study explains the reciprocal relationship between spoilt votes and the allocation of Constituency Development Funds (CDF), exploring how this allocation is affected by the number of spoilt votes. The study's findings have broader implications for election integrity and policy in Kenya. This study adopts a quantitative approach, employing the tools of statistical analysis, particularly focusing on linear regression. By utilizing quantitative methods, the research aims to provide empirical evidence that can help policymakers, election authorities, and stakeholders make informed decisions and implement measures to enhance the fairness and credibility of future elections in Kenya.

Background to the Study

In the 2013 Kenyan presidential election, Uhuru Kenyatta emerged as the winner with **50.07 percent** of the vote, securing a first-round victory. Notably, the election witnessed an exceptionally high voter turnout, with **85.86 percent** of registered voters actively participating. Out of a total of 12,330,028 votes cast, 108,975 were deemed invalid, resulting in a spoilt vote percentage of roughly **0.88 percent**. These statistics reveal only **33.34 percent** of the total population registered as voters, underscoring the significance and democratic vitality of this pivotal election.

Key election information

Variable	Approximate Value
Cast Votes	12,330,028
Valid Votes	12,221,053
Invalid Votes	108,975
Registered Voters	14,352,533
Population (July 2012 est.)	43,013,341
Voter Turnout	85.86%
Spoilt Vote Percentage	0.88%
Registered Voter Percentage	33.34%
CLOSE RACE	55 constituencies
LANDSLIDE	234 constituencies
CDF Allocation (2013-2017)	KES 138 Billion

Table: Key Data Points

Study Objectives

- Summarize variables, including Spoilt votes, Registered voters, Cast votes, and CDF Allocation
- Analyze the correlations between these variables to understand their relationships
- Identify disparities in voter registration and participation and provide potential recommendations for democratic representation and the electoral process.

Summary of data

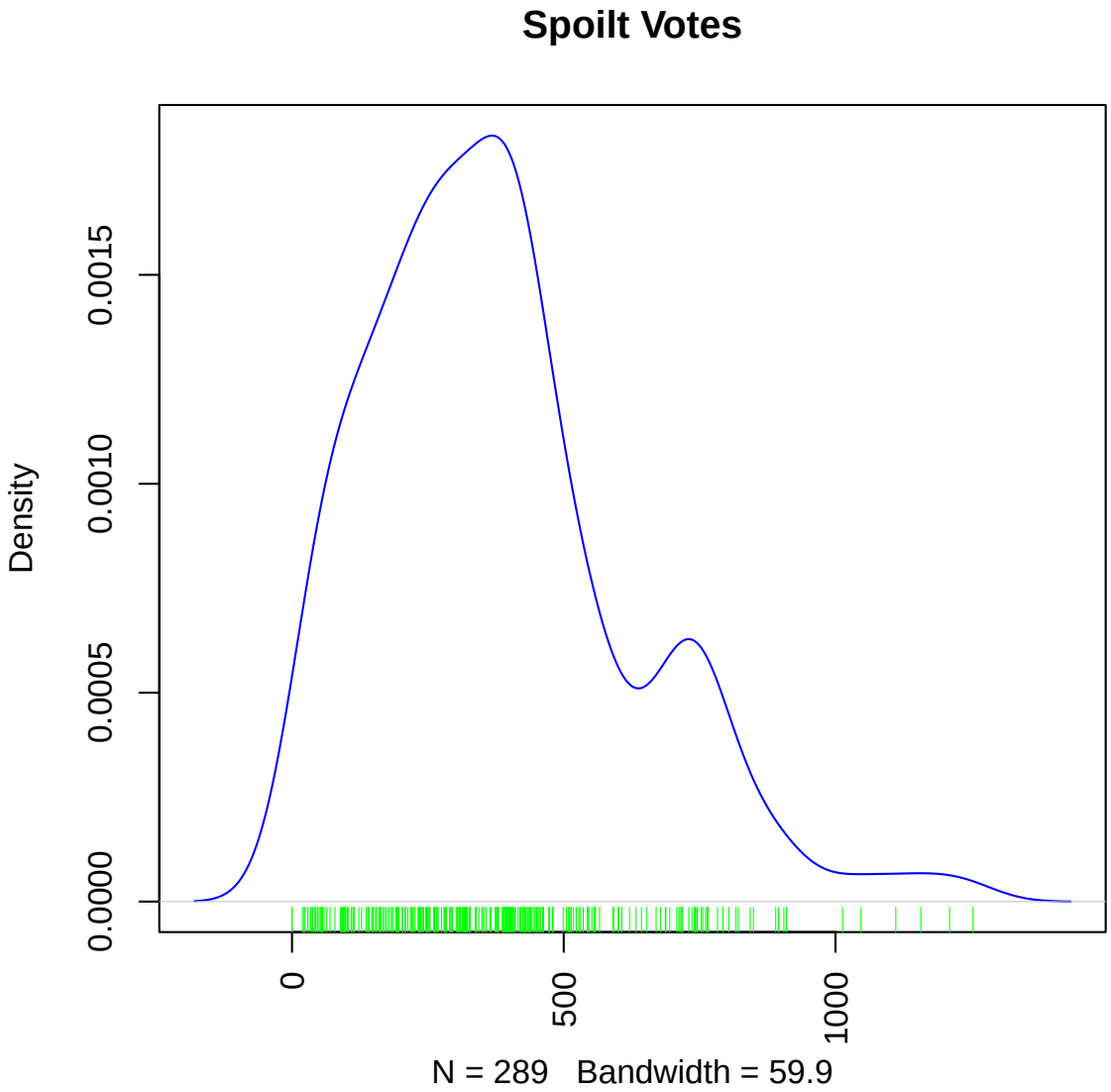
Summary of Spoilt, Registered, Cast, and Allocation respectively.

Variable	Min	Median	Mean	Max
Spoilt	0	344	3.761246e+02	1253
Registered	10574	44489	4.945559e+04	133279
Cast	8718	37917	4.248164e+04	110828
Allocation	415636890	474971364	4.801269e+08	682170481

Density Plot

Density plots are useful for visualizing the distribution and shape of a continuous variable, showing where the data is concentrated and how it is spread across different values.The probability density function of a continuous random variable describes how the probability of observing a specific value is distributed across the variable. The probability of obtaining the constituency with the highest frequency number of spoilt votes was approximately **0.0020** as shown

Density Plot for Spoilt Votes



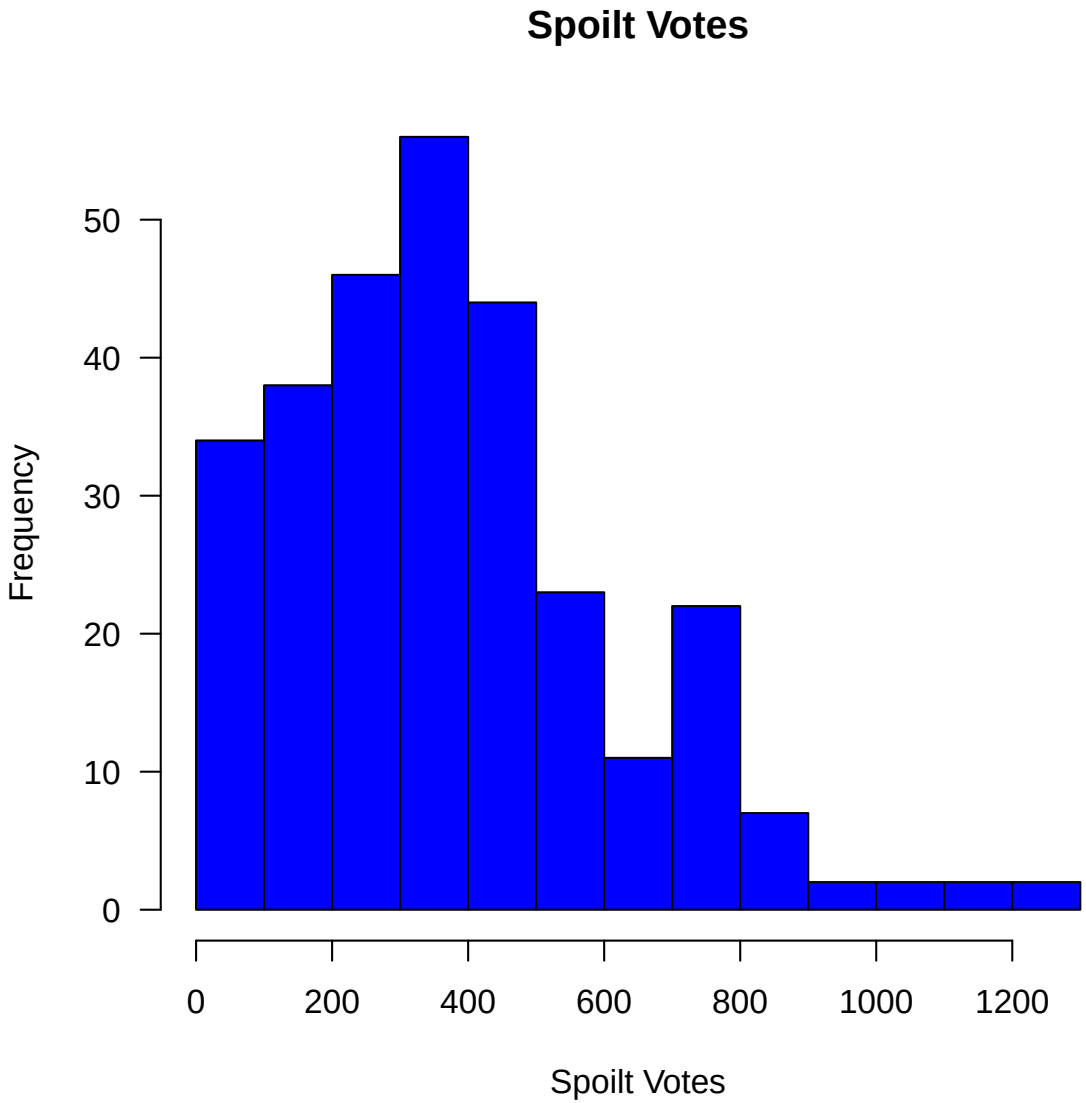
Histograms

Histograms are a graphical representation commonly used to display the distribution of a data variable and its corresponding frequency or count of observations within different intervals. Histograms are valuable for exploring and summarizing data, identifying central tendencies (such as modes and medians), detecting outliers, and assessing the spread or variation of the data as shown

Histogram of spoilt votes

The histogram in the next slide represents the distribution of the **Spoilt Votes** in the dataset. The x-axis ranges from 0 to 1200, divided into intervals of 100, and represents the Spoilt votes. The y-axis ranges from 0 to 60, divided into intervals of 10, and represents the frequency of these values.

Frequency Distribution of Spoilt Votes



Histogram of spoilt votes

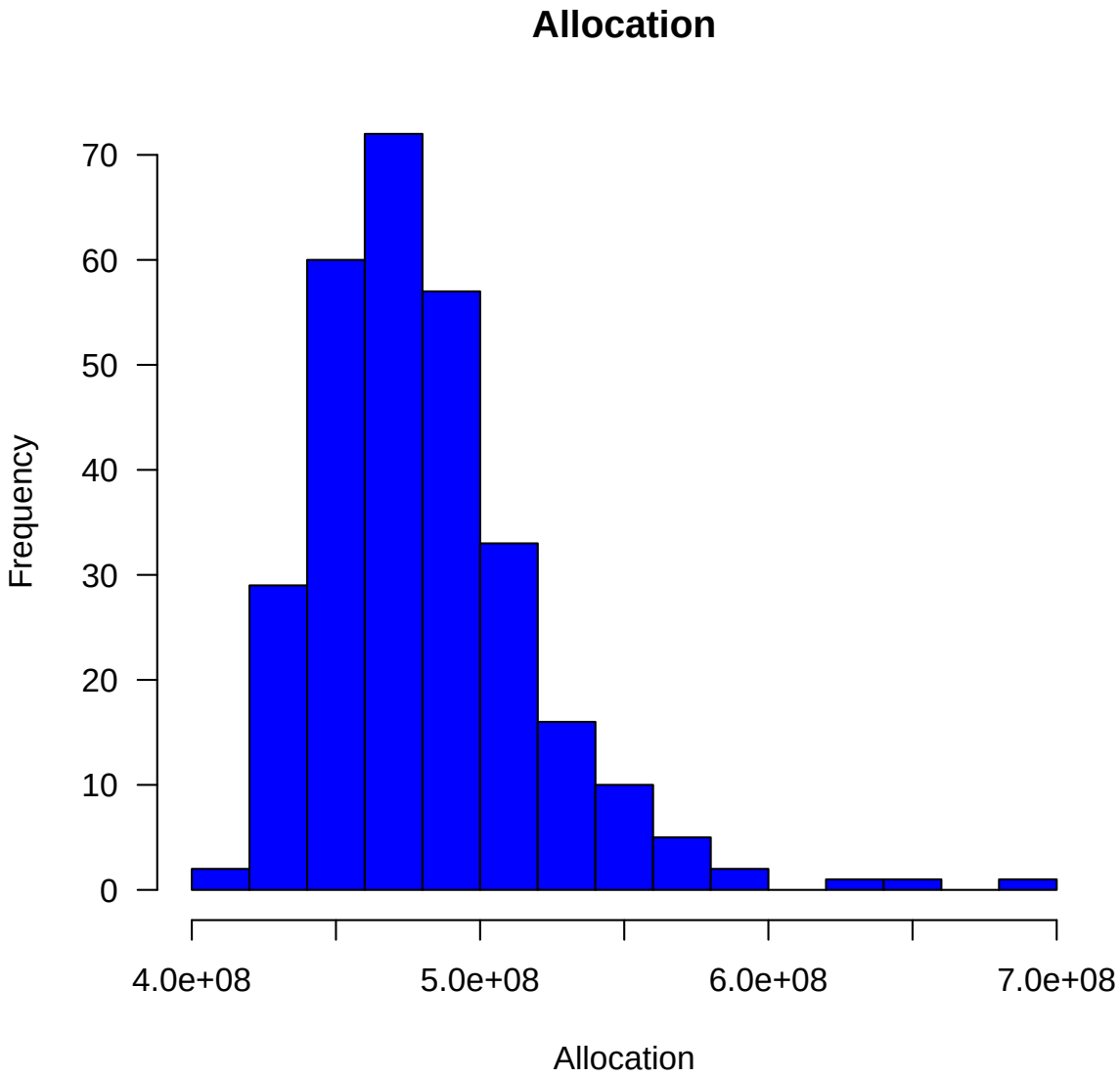
- From the histogram, we can observe that:
- The majority of the Spoilt Votes fall within the range of **300 to 400**.
- The percentage of Spoilt Votes that fall within the 300-400 range is approximately **19.378 percent of the total Spoilt votes**.
- The frequency count for Spoilt Votes falling within the range of 300 to 400 is **56**. This means there are 56 instances in the dataset where the Spoilt votes are between 300 and 400.
- The skewness of the Spoilt Votes is approximately **0.90**.

This value indicates a moderate positive skew, which means the distribution of Spoilt votes is **skewed to the right**.

Histogram of Allocated CDF funds

The histogram in the next slide represents the distribution of the Allocation of CDF funds in the dataset. The x-axis ranges from 400 million to 700 million, divided into intervals of 21.5 million, and represents the Allocation values. The y-axis ranges from 0 to the maximum frequency count 85 distributed in intervals of 10

Frequency Distribution of Allocation



Histogram of Allocation Values

From the histogram, we can observe that:

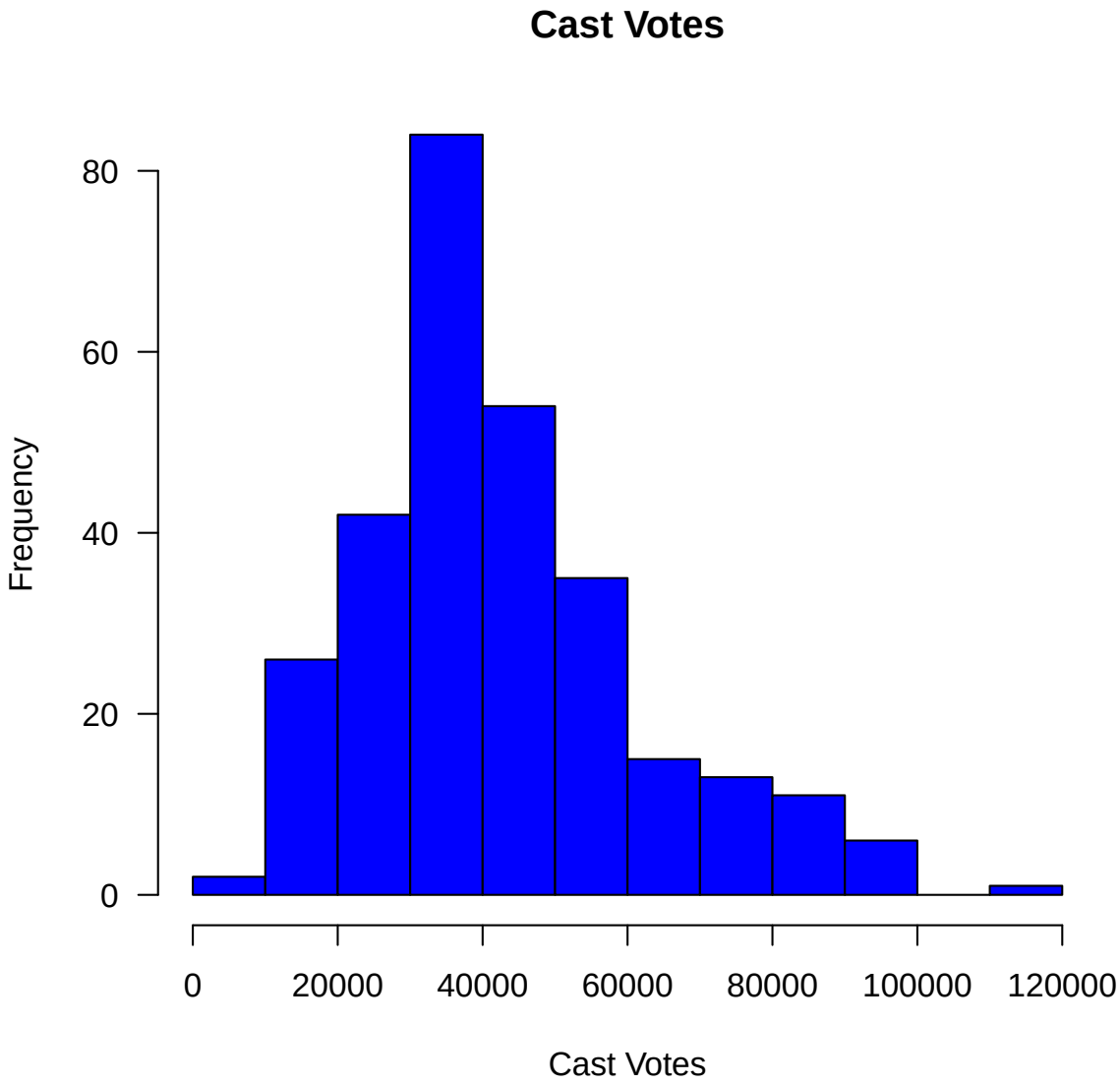
- The majority of the Allocated finds fall within the range of **464.5 million to 486 million**.
- The percentage of Allocation values that fall within the 464.5 million to 486 million range is approximately **29.41 percent** of the total "Allocation" values.
- The frequency count for Allocation values falling within the range of 464.5 million to 486 million is **85**. This means there are 85 instances in the dataset where the Allocation values are between 464.5 million and 486 million.
- The skewness of the Allocation values is approximately **1.45**.

This value indicates a moderate positive skew, which means the distribution of Allocation values is **skewed to the right**.

Histogram of Cast Votes

The histogram in the next slide represents the distribution of the Cast votes in the dataset. The x-axis ranges from 0 to 120,000, divided into intervals of 10,000, and represents the Cast votes. The y-axis ranges from 0 to the maximum frequency count 84

Frequency Distribution of Cast Votes



Histogram of Cast Values

From the histogram, we can observe that:

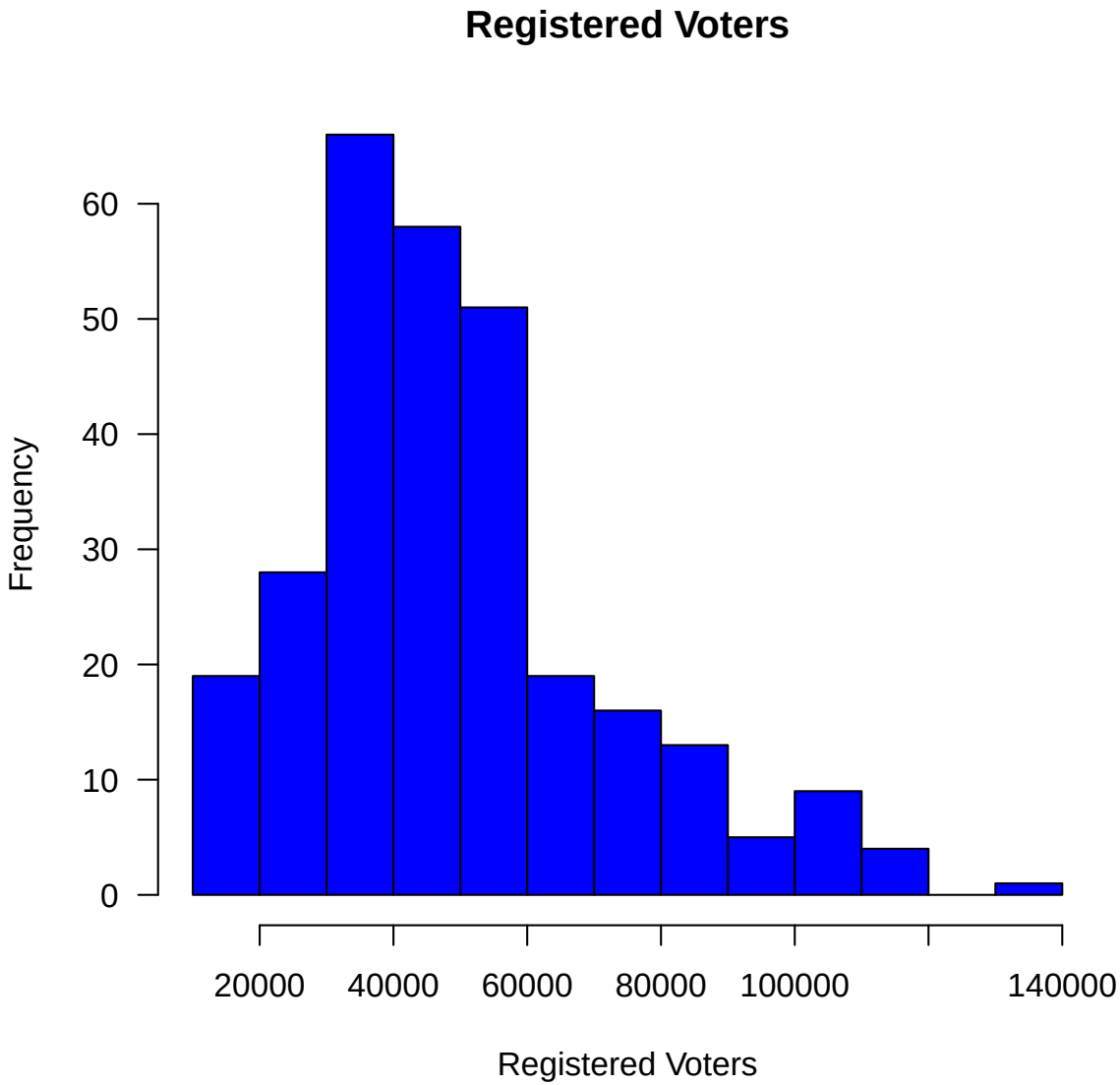
- The majority of the Cast votes fall within the range of **30,000 to 40,000**.
- The percentage of Cast votes that fall within the 30,000 to 40,000 range is approximately **29.07 percent** of the total Cast votes.
- The frequency count for Cast votes falling within the range of 30,000 to 40,000 is **84**. This means there are 84 instances in the dataset where the Cast votes are between 30,000 and 40,000.
- The skewness of the Cast votes is approximately **0.84**.

This value indicates a moderate positive skew, which means the distribution of Cast votes is **skewed to the right**.

Histogram of Registered Voters

The histogram in the next slide represents the distribution of the Registered voters in the dataset. The x-axis ranges from 0 to 140,000, divided into intervals of 10,000, and represents the Registered voters. The y-axis ranges from 0 to the maximum frequency count 66

Frequency Distribution of Registered voters



Histogram of Registered Voters

From the histogram,

- The majority of the Registered voters fall within the range of **30,000 to 40,000**.
- The percentage of Registered voters that fall within the 30,000 to 40,000 range is approximately **22.84 percent** of the total Registered voters.
- The frequency count for Registered voters falling within the range of 30,000 to 40,000 is **66**. This means there are 66 instances in the dataset where the Registered voters are between 30,000 and 40,000.
- The skewness of the Registered voters is approximately **0.96**.

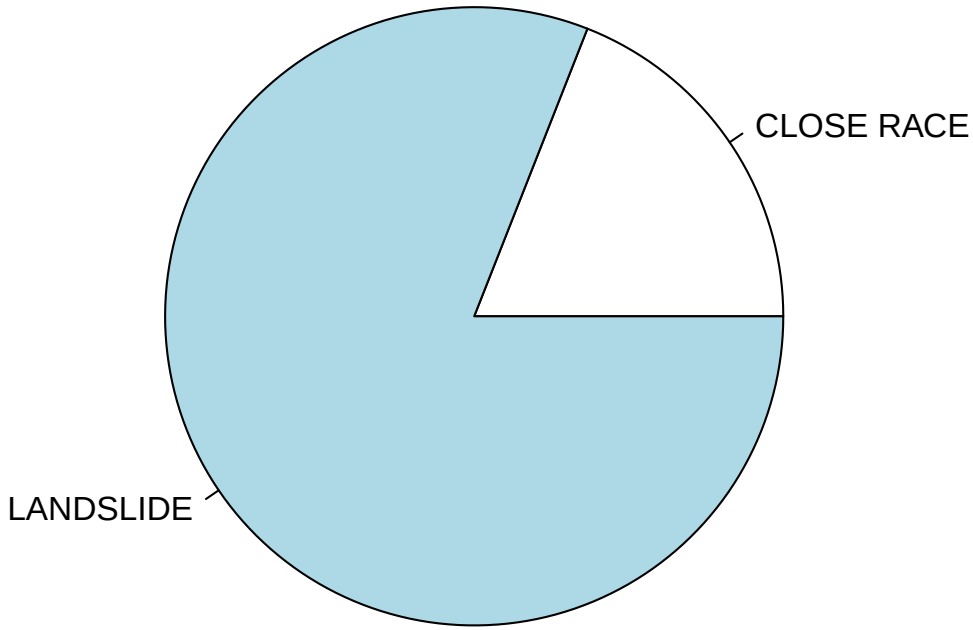
This value indicates a moderate positive skew, which means the distribution of Registered voters is **skewed to the right**.

Distribution of Voting Pattern

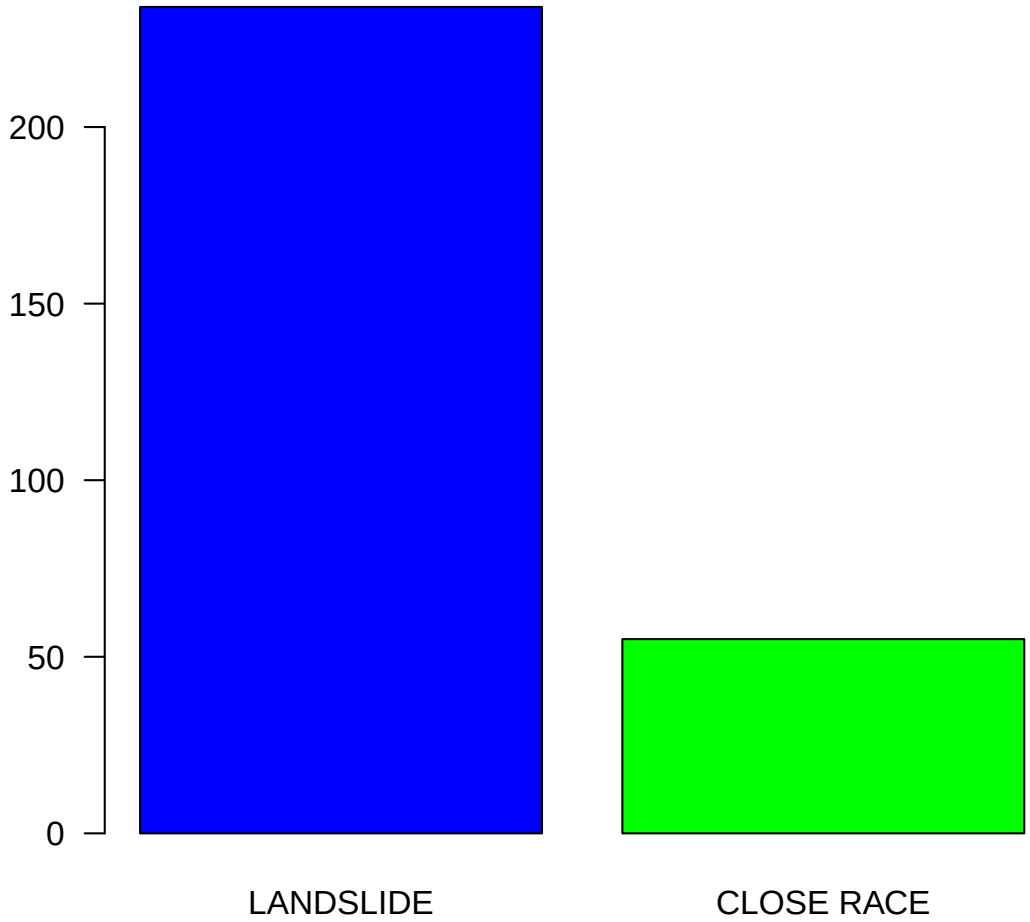
The data shows two distinct voting patterns:

- CLOSE RACE: Elections were fiercely competitive, with a narrow margin of victory. Outcomes were uncertain.
- LANDSLIDE: one choice dominated, securing a decisive and one-sided victory.

Pie Chart of the voting pattern



Bar graph of voting pattern



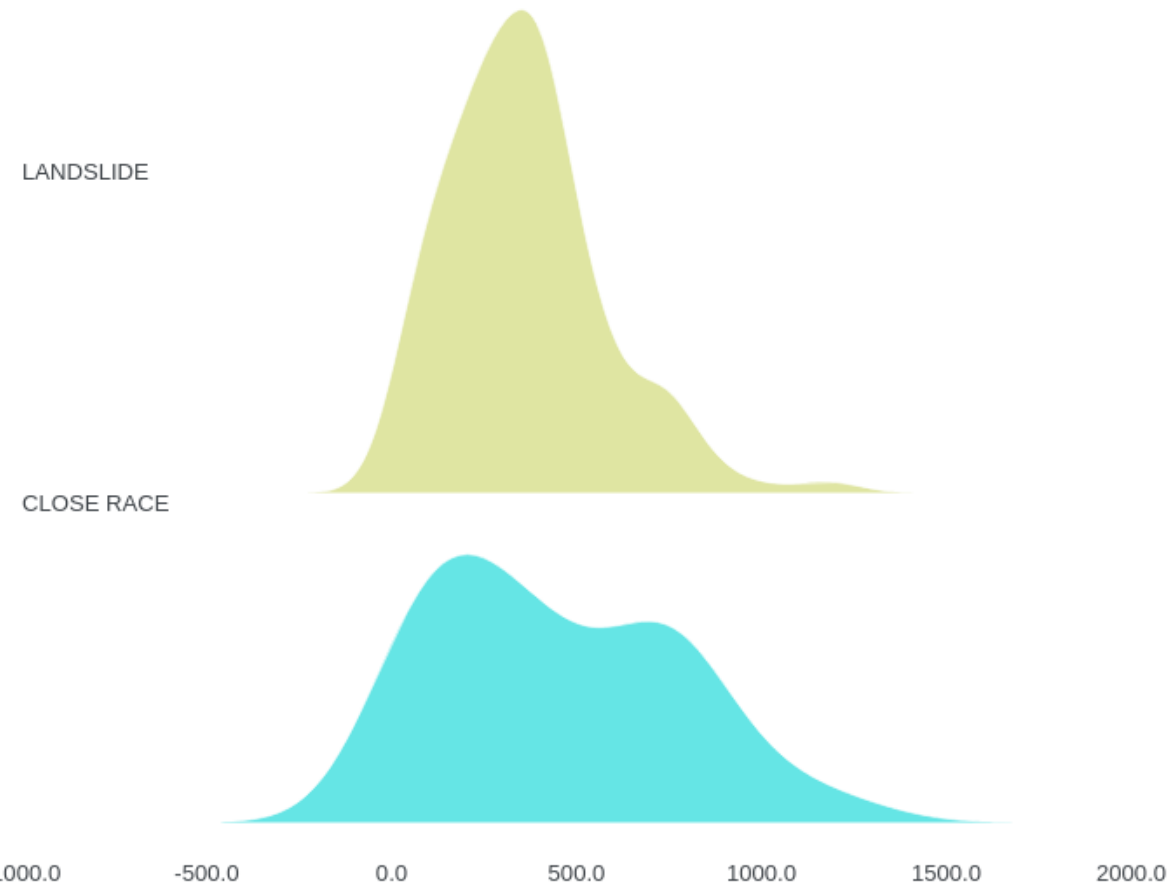
Pattern Counts

Voting pattern counts for "CLOSE RACE" and "LANDSLIDE"

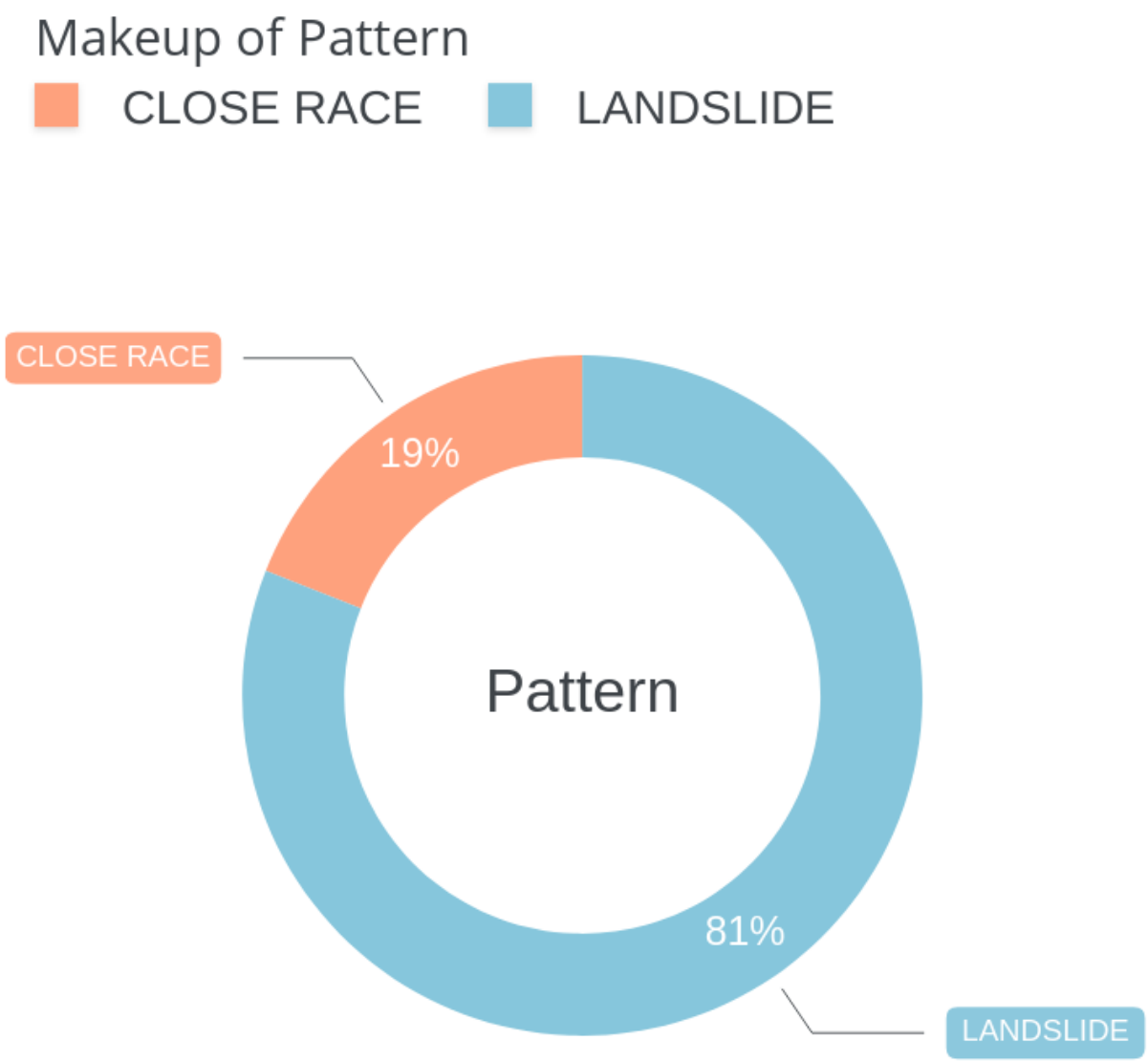
Pattern	Count
CLOSE RACE	55
LANDSLIDE	234

Pattern impacts Spoilt distribution

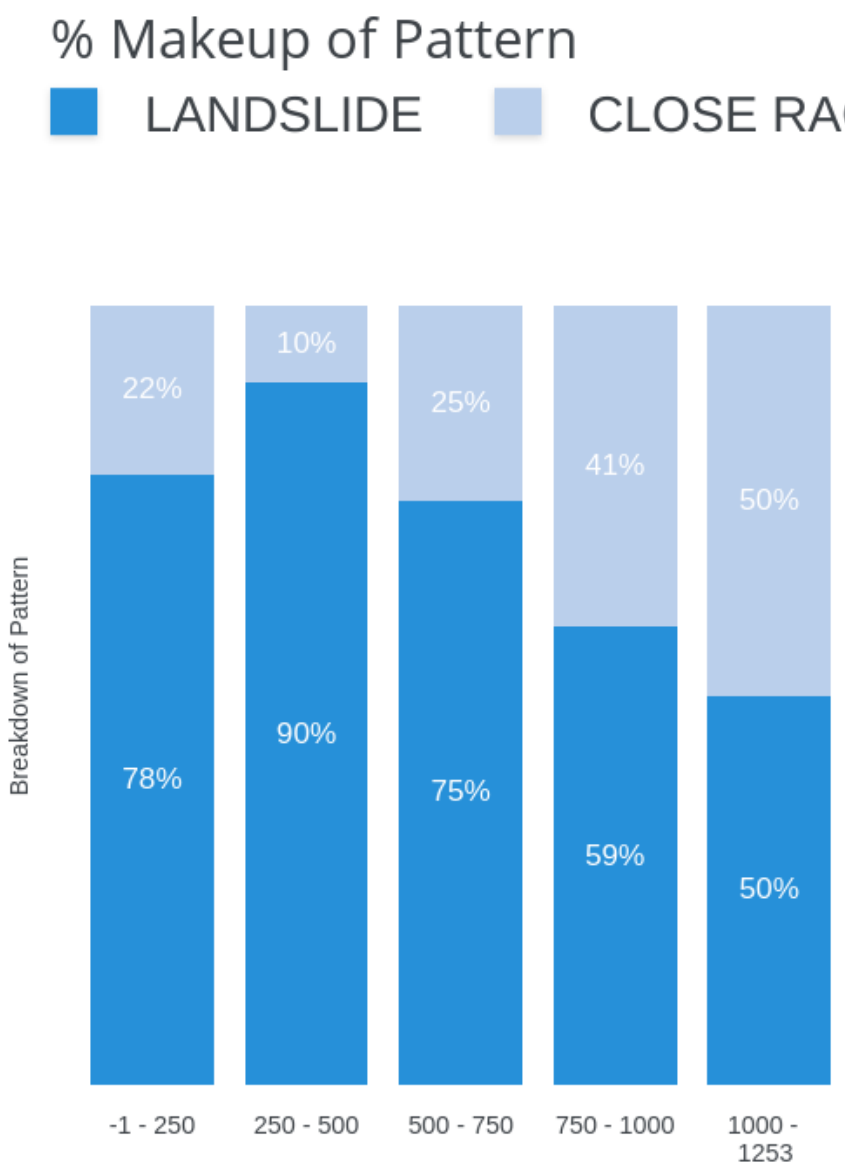
Distribution of Spoilt



81% of Patterns: LANDSLIDE

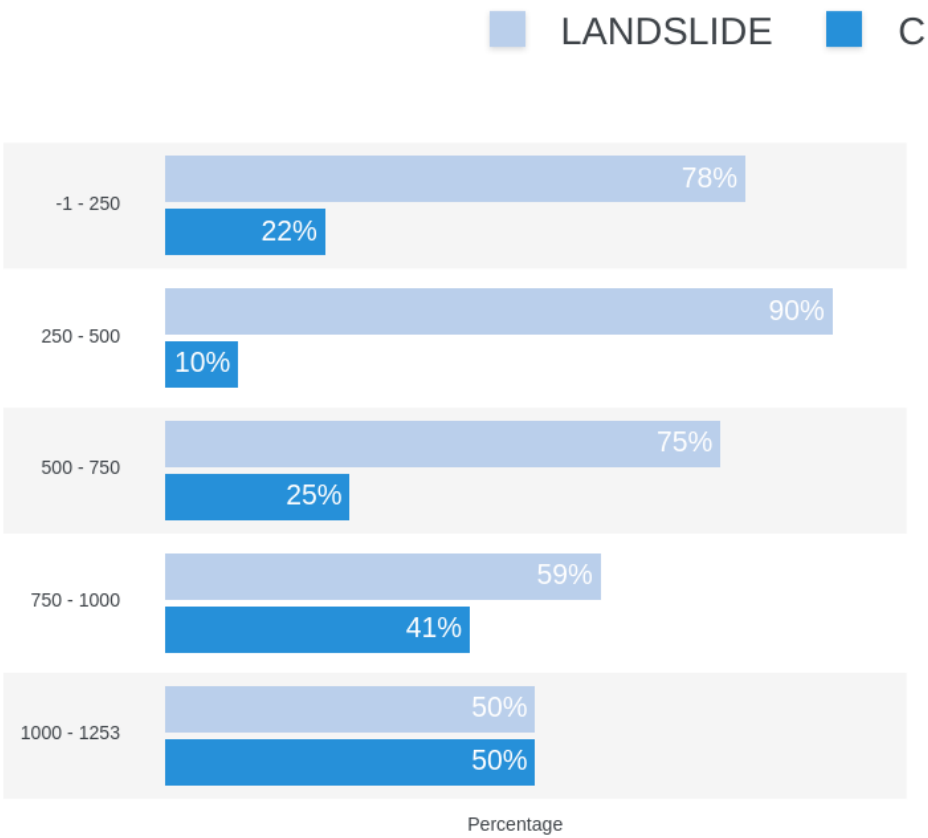


Pattern distribution varies by Spoilt



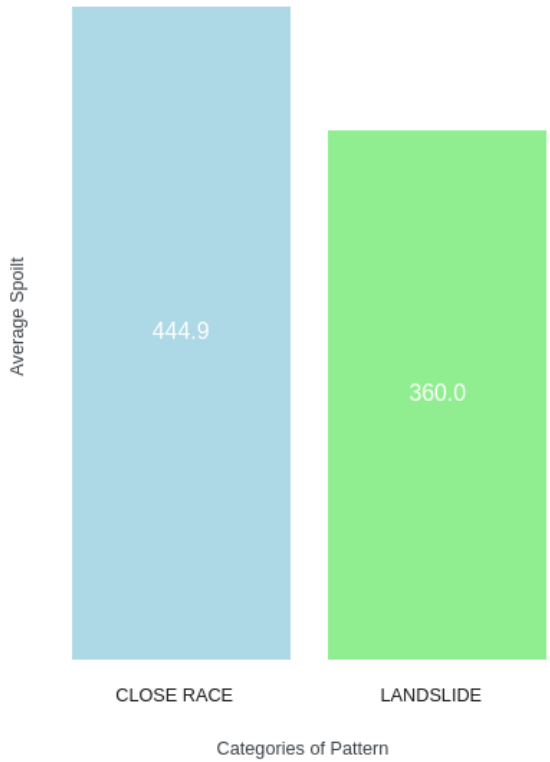
Spoilt: LANDSLIDE spikes in 250.0 - 500.0 segment

Distribution of Pattern across groups of Spoilt in percent



Pattern categories impact Spoilt significantly

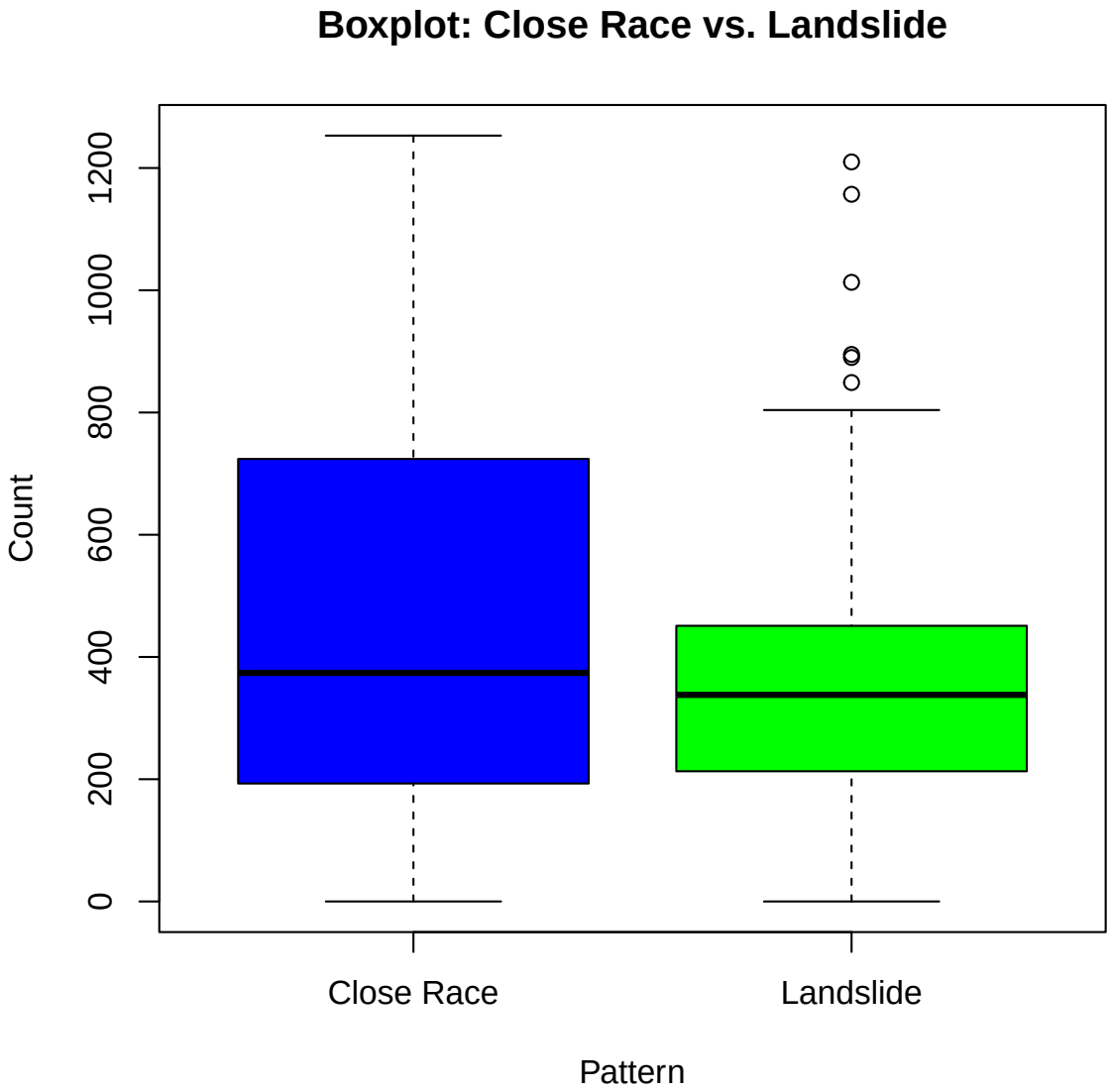
Average Spoilt



Voting Pattern Distribution

The box plot in the next slide shows the distribution of the number of spoiled votes in each category of election result pattern. The line in the middle of each box represents the median of the data, while the box itself represents the interquartile range (from the first quartile to the third quartile). The whiskers represent the range of the data within 1.5 times the interquartile range, and any points outside of this range are considered outliers and are represented as individual points. From this plot, we can see that the distribution of spoiled votes varies between different election result patterns.

Close Vs Landslide box Plot

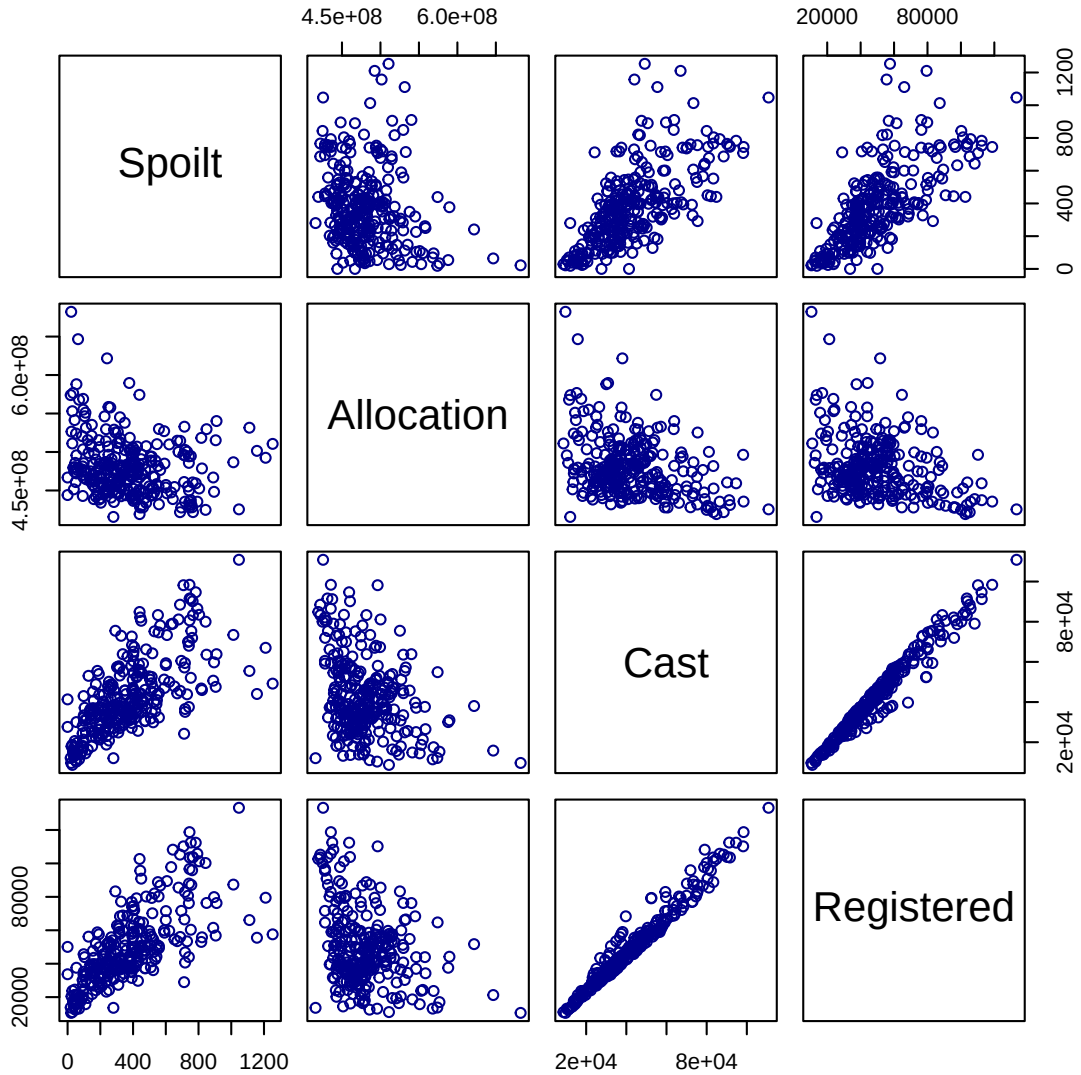


Plot interpretation

The box plot provides a visual representation of the relationship between the number of spoiled votes and the pattern of the election result. From the plot, we can observe the following:

- **LANDSLIDE pattern:**, the number of spoiled votes varies widely, as indicated by the length of the box and the whiskers. The median number of spoiled votes (represented by the line in the middle of the box) is relatively high, suggesting that in constituencies where the election result was a landslide, there tends to be a higher number of spoiled votes.
- **CLOSE RACE pattern:**, the number of spoiled votes also varies, but the range is narrower than for the "LANDSLIDE" pattern. The median number of spoiled votes is lower, suggesting that in constituencies where the election result was a close race, there tends to be a lower number of spoiled votes.

Correlation Plots



Correlation

As shown in the scatter plots, some variables are related directionally either positively or negatively. The correlation coefficients indicate the strength and direction of these relationships. Positive values suggest a direct (positive) relationship, where both variables tend to move in the same direction. Negative values indicate an inverse (negative) relationship, where the variables tend to move in opposite directions. The closer the correlation coefficient is to -1 or 1, the stronger the relationship.

Correlation Matrix

	Spoilt	Registered	Cast	Allocation
Spoilt	1.0000	0.7168	0.6795	-0.2077
Registered	0.7168	1.0000	0.9808	-0.3251
Cast	0.6795	0.9808	1.0000	-0.3431
Allocation	-0.2077	-0.3251	-0.3431	1.0000

Correlation Matrix - Part 1

- **Spoilt vs. Registered**: There is a moderately positive relationship (correlation coefficient of **0.7168**) between Spoilt and Registered. As the number of Registered voters increases, the number of Spoilt votes tends to increase, indicating that higher voter registration may lead to more Spoilt votes.
- **Spoilt vs. Cast**: Similar to the relationship with Registered, there is also a moderately positive relationship (correlation coefficient of **0.6795**) between Spoilt and Cast. This suggests that as the number of votes Cast increases, the number of Spoilt votes tends to increase as well.

Correlation Matrix - Part 2

- **Spoilt vs. Allocation**: There is a weak negative relationship (correlation coefficient of -**0.2077**) between Spoilt and Allocation. As Allocation increases, the number of Spoilt votes tends to decrease slightly, indicating a slight inverse relationship.
- **Registered vs. Cast**: There is a very strong positive relationship (correlation coefficient of **0.9808**) between Registered and Cast. This means that as the number of Registered voters increases, the number of votes Cast also significantly increases.

Correlation Matrix - Part 3

- **Registered vs. Allocation**: There is a moderately negative relationship (correlation coefficient of -**0.3251**) between Registered and Allocation. As Allocation increases, the number of Registered voters tends to decrease.
- **Cast vs. Allocation**: There is also a moderately negative relationship (correlation coefficient of -**0.3431**) between Cast and Allocation. As Allocation increases, the number of votes Cast tends to decrease.

Linear Regression Model

The `lm` function provides a comprehensive analysis of the relationship between the response variable, 'Spoilt votes' (Y), and the predictor variables: 'Registered Voters' (X1), 'Cast Votes' (X2), 'Voting Patterns' (X3), and 'CDF Allocation' (X4).

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \epsilon$$

Where:

Linear Regression Model

- Y: Response variable (Spoilt votes)
- X1: Number of registered voters
- X2: Number of cast votes
- X3: Voting patterns per constituency (Close race or landslide)
- X4: Allocation of CDF Funds per constituency
- $\beta_0, \beta_1, \beta_2, \beta_3, \beta_4$: Represent coefficients of respective explanatory variables on the response variable, and ϵ represents the error term

Data Variables

Variable	**Description**
Y	Represents Spoilt votes (invalid)
$X1$	Represents Registered Voters
$X2$	Represents Cast votes
$X3$	Represents Voting Patterns per constituency (Close race or landslide)
$X4$	Represents Allocation of CDF Funds per constituency

Hypothesis Testing

The hypothesis testing will help us determine whether any of the predictor variables have a statistically significant impact on the number of Spoilt votes in the linear regression model.

Hypothesis Testing

Null Hypothesis (H0). There is no significant relationship between the number of Spoilt votes and the predictor variables, including the number of Registered Voters, Cast Votes, Voting Patterns, and CDF Allocation. In mathematical terms, this can be expressed as H0:
$\beta_0 = \beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$.
Alternate Hypothesis (H1). There is a significant relationship between the number of Spoilt votes and at least one of the predictor variables, including the number of Registered Voters, Cast Votes, Voting Patterns, and CDF Allocation. In mathematical terms, this can be expressed as H1: At least one of $\beta_0, \beta_1, \beta_2, \beta_3, \beta_4$ is not 0.

	Dependent variable:
	Y
X1	0.013*** (0.002)
X2	-0.007** (0.003)
X3LANDSLIDE	-64.158** (24.834)
X4	0.00000 (0.00000)
Constant	13.808 (140.393)
Observations	289
R2	0.539
Adjusted R2	0.533
Residual Std. Error	163.304 (df = 284)
F Statistic	83.178*** (df = 4; 284)
Note:	*p<0.1; **p<0.05; ***p<0.01

Confidence Level Percentage

Confidence intervals for the linear regression model, Lower CI:2.5 vs Upper CI 97.5 percent

Variable	Lower CI	Upper CI
(Intercept)	-2.625348e+02	2.901504e+02
X1	8.645493e-03	1.738032e-02
X2	-1.168361e-02	-1.323314e-03
X3LANDSLIDE	-1.130398e+02	-1.527593e+01
X4	-4.348301e-07	6.305344e-07

Relationship between Variables

- **Number of Registered Voters (X1)**: An increase in the number of registered voters (with each additional unit) leads to an increase in Spoilt votes. For every one-unit increase in the number of registered voters, Spoilt votes are expected to increase by approximately 0.0130 units, which is a positive effect. The p-value for the effect of the number of registered voters on Spoilt votes is less than 0.05, indicating high statistical significance.
- **Number of Cast Votes (X2)**: An increase in the number of cast votes (with each additional unit) leads to a decrease in Spoilt votes. For every one-unit increase in the number of cast votes, Spoilt votes are expected to decrease by approximately 0.0065 units, indicating a negative effect. The p-value for the effect of the number of cast votes on Spoilt votes is less than 0.05, indicating statistical significance.

Relationship between Variables

- **Landslide Pattern (X3)**: The coefficient for Voting Patterns landslide is -**64.1579**, which means the presence of a landslide voting pattern is associated with a decrease in Spoilt votes by 64.1579 units. The p-value for Voting Patterns landslide is 0.010, which is less than 0.05. This suggests that Voting Patterns landslide is a statistically significant predictor of Spoilt votes.
- **Allocation (X4)**: The coefficient for CDF Allocation is **9.785e-08**, which means for every one-unit increase in CDF Allocation, the Spoilt votes are expected to increase by 9.785e-08 units. However, the p-value for CDF Allocation is 0.718, which is greater than 0.05. This suggests that CDF Allocation is not a statistically significant predictor of 'Spoilt votes'.

Key Insights

In summary, the analysis of factors influencing spoilt votes during the 2013 Kenyan presidential elections reveals the following key insights:

- **Number of Registered Voters (X1):** A higher number of registered voters is associated with an increase in spoilt votes, and this relationship is statistically significant.
- **Number of Cast Votes (X2):** An increase in the number of cast votes leads to a decrease in spoilt votes, and this effect is also statistically significant.
- **Landslide Pattern (X3):** The presence of a "Landslide" pattern is associated with fewer spoilt votes, with statistical significance.
- **Allocation (X4):** Spoilt votes do not significantly affect Constituency development funds allocation.

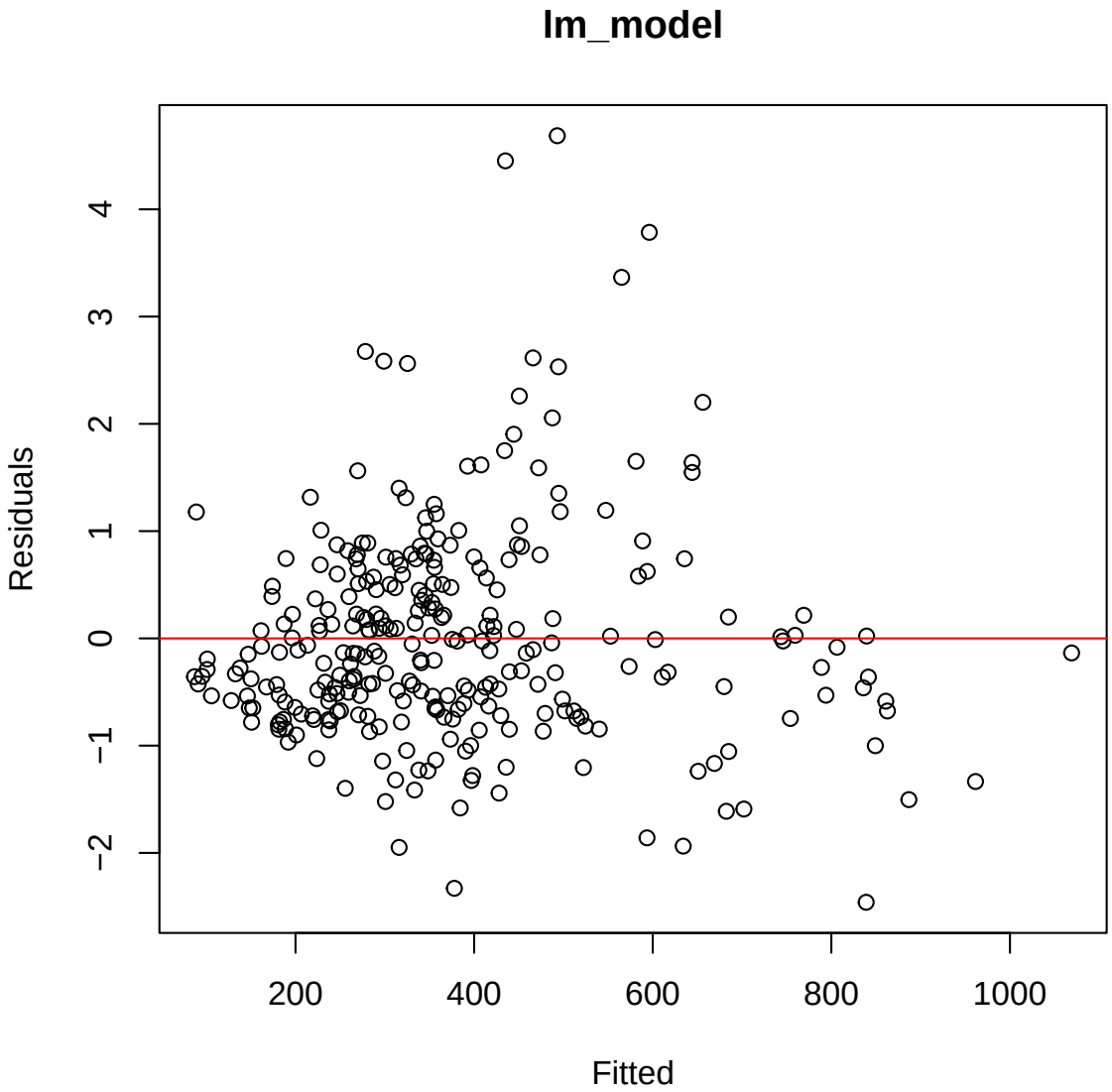
Assumptions and Credibility

It is important to note that this investigation is based on the underlying assumptions of the OLS linear estimator. The findings unveil instances where these assumptions are not valid, which affects the credibility of results. More comprehensive data could help improve the reliability of the analysis as shown below:

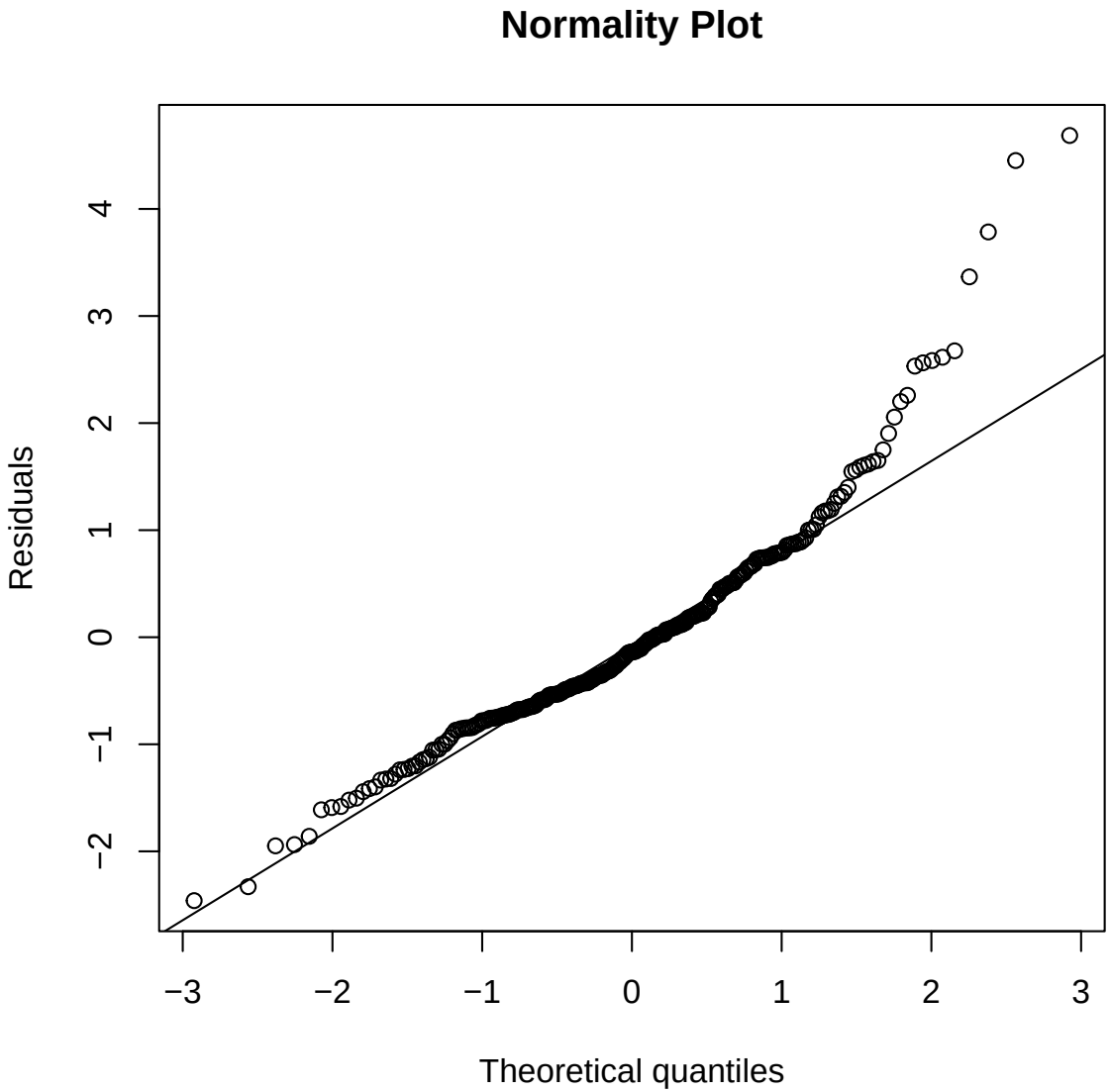
Assumptions of Linear Regression Analysis

- The mean of the errors is approximately **2.6673e-08**, which is very close to zero, suggesting that the assumption of zero mean errors is met.
- The Durbin-Watson statistic is **1.891**, close to 2, indicating no significant autocorrelation in the residuals, and the assumption of independent errors is met.
- The residuals versus predicted values plot and tests are shown below, to show whether the assumption of homoscedasticity is met.
- The Q-Q plot of residuals and tests is shown below to try and indicate a normal distribution.

Homoscedasticity



QQ Normality Plot



Homoscedasticity Test

- We start by testing the assumption of homoscedasticity (equal variances of residuals).
- The Breusch-Pagan test is commonly used for this purpose.
- The Breusch-Pagan test p-value:**0.0001386791**
- If p-value is below a significance level (e.g., 0.05), the assumption of homoscedasticity is violated.

Normality Test

- Next, we test the normality of residuals.
- The Shapiro-Wilk test can be used for this purpose.
- The Shapiro-Wilk test p-value: **0.0000000001676222**
- If p-value is below a significance level (e.g., 0.05), the assumption of normality is violated.

Model Tests and Statistics

- **Omnibus Test:** 77.318
  - **Interpretation:** The Omnibus test assesses the overall goodness of fit of the model. In this case, a value of 77.318 suggests that the model fits the data well.
- **Jarque-Bera Test:** 205.740
  - **Skew:** 1.225
  - **Kurtosis:** 6.329
  - **Prob(JB):** 2.11e-45
  - **Interpretation:** The Jarque-Bera test checks for normality of residuals. A high test statistic (205.740) and a very low p-value (2.11e-45) indicate that the residuals may not be normally distributed. Additionally, skew (1.225) and kurtosis (6.329) values can suggest non-normality and potential outliers.

Model Tests and Statistics

- **Condition Number:** 7.04e+09
  - **Interpretation:** The condition number assesses multicollinearity in the model. A very high condition number (7.04e+09) may indicate a strong multicollinearity issue among predictor variables.

Assumptions and Their Confirmations (Part 1)

- **Zero Mean Errors:**
  - **Assumption:** The errors have a mean value of zero.
  - **Confirmation:** The mean of the errors is very close to zero suggesting that this assumption is met..
- **Independent Errors (Durbin-Watson Test):**
  - **Assumption:** The errors are independent with no autocorrelation.
  - **Confirmation:** The Durbin-Watson statistic is close to 2, indicating no significant autocorrelation.
- **Homoscedasticity (Constant Spread in Residuals):**
  - **Assumption:** The spread of errors is consistent across different levels of predictor variables.
  - **Confirmation:** This assumption is violated as there's a non-constant spread in residuals as shown in the test and the plot.

Assumptions and Their Confirmations (Part 2)

- **Normality of Residuals (Jarque-Bera Test):**
  - **Assumption:** The errors follow a normal distribution.
  - **Confirmation:** The Jarque-Bera test results indicate non-normality of residuals as shown by the test and plot.
- **Multicollinearity (Condition Number):**
  - **Assumption:** Predictor variables are not strongly correlated.
  - **Confirmation:** A high condition number suggests strong multicollinearity

References

1. R Core Team (2019). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.
2. Wickham et al., (2019). Welcome to the tidyverse. Journal of Open Source Software, 4(43), 1686, https://doi.org/10.21105/joss.01686
3. Meyer, M. (2002). Uncounted Votes: Does Voting Equipment Matter? Chance Methods, 15(4), 33-38.
4. National Government Constituencies Development Fund (NGCDF). (2013-2017). Title of the Webpage. https://ngcdf.go.ke/allocations/. Retrieved October 23, 2023.
5. The Electoral Knowledge Network. (2013). Title of the Webpage. https://aceproject.org/ero-en/regions/africa/KE/kenya-constituency-summary-of-voter-turnout-the. Retrieved October 2023.