Statistical Analysis of Twitter Data

1. Introduction

Problem Statement: Social media platforms such as Twitter provide a wealth of data that can be utilized to understand public sentiment, track trends, and identify potential influencers. For businesses, analyzing tweets can reveal insights into customer satisfaction, brand awareness, and market trends. The purpose of this analysis is to understand the sentiment expressed in tweets over time, identify the most active users, and discover key trends that can help businesses improve customer satisfaction and make informed decisions.

Key Business Questions:

1. What is the overall sentiment distribution of the tweets?
2. Who are the most active users, and what are they tweeting about?
3. What are the most common hashtags used in tweets?
4. How does tweet sentiment change over time?
5. Is there a significant difference in sentiment over time?

2. Hypotheses

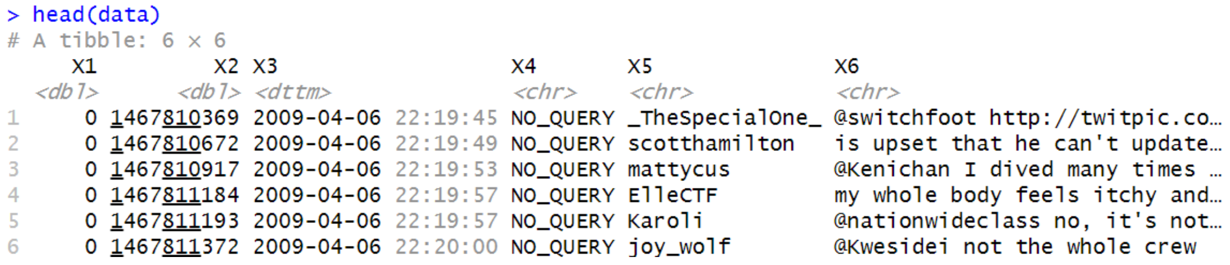
1. H1: Most tweets have a neutral sentiment.
2. H2: The sentiment of tweets changes significantly over time.
3. H3: Certain users or hashtags are associated with more frequent or distinct sentiment patterns.

3. Data Preparation, Sampling, and Cleaning

Data Description: The dataset consists of 1,600,000 rows and 6 columns. The columns include:

- X1: Sentiment (0 = Negative, 4 = Positive)
- X2: Unique ID

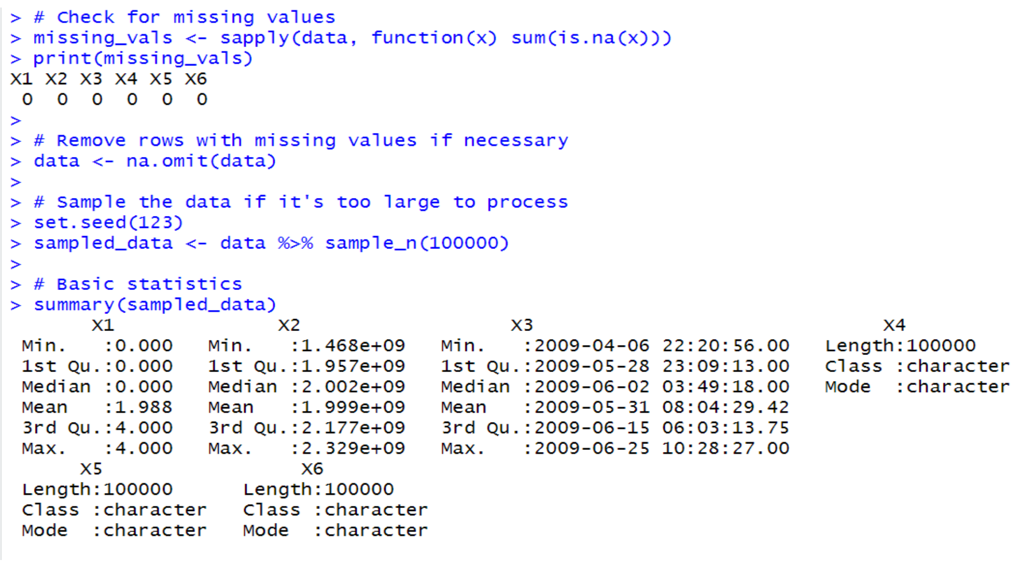
- X3: Timestamp
- X4: Query (if any)
- X5: Username
- X6: Tweet text
-



```
> head(data)
# A tibble: 6 × 6
     X1         X2 X3                  X4       X5              X6
  <dbl>      <dbl> <dttm>              <chr>    <chr>           <chr>
1     0 1467810369 2009-04-06 22:19:45 NO_QUERY _TheSpecialOne_ @switchfoot http://twitpic.co…
2     0 1467810672 2009-04-06 22:19:49 NO_QUERY scotthamilton   is upset that he can't update…
3     0 1467810917 2009-04-06 22:19:53 NO_QUERY mattycus        @Kenichan I dived many times …
4     0 1467811184 2009-04-06 22:19:57 NO_QUERY ElleCTF         my whole body feels itchy and…
5     0 1467811193 2009-04-06 22:19:57 NO_QUERY Karoli          @nationwideclass no, it's not…
6     0 1467811372 2009-04-06 22:20:00 NO_QUERY joy_wolf        @Kwesidei not the whole crew
```

Data Preprocessing:

- Timestamp Parsing: The timestamp in X3 was cleaned by removing the time zone abbreviation ("PDT") and converting it to a proper datetime format.
- Missing Values: The dataset was checked for missing values, and none were found.
- Sampling: Given the large size of the dataset, a random sample of 100,000 tweets was selected for analysis to ensure computational efficiency.

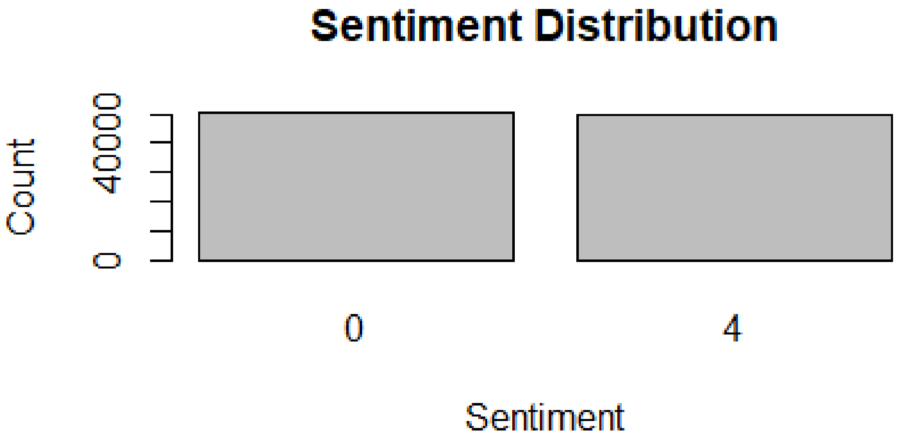


```
> # Check for missing values
> missing_vals <- sapply(data, function(x) sum(is.na(x)))
> print(missing_vals)
X1 X2 X3 X4 X5 X6 
 0  0  0  0  0  0 
> 
> # Remove rows with missing values if necessary
> data <- na.omit(data)
> 
> # Sample the data if it's too large to process
> set.seed(123)
> sampled_data <- data %>% sample_n(100000)
> 
> # Basic statistics
> summary(sampled_data)
       X1              X2                X3                              X4           
 Min.   :0.000   Min.   :1.468e+09   Min.   :2009-04-06 22:20:56.00   Length:100000     
 1st Qu.:0.000   1st Qu.:1.957e+09   1st Qu.:2009-05-28 23:09:13.00   Class :character  
 Median :0.000   Median :2.002e+09   Median :2009-06-02 03:49:18.00   Mode  :character  
 Mean   :1.988   Mean   :1.999e+09   Mean   :2009-05-31 08:04:29.42                     
 3rd Qu.:4.000   3rd Qu.:2.177e+09   3rd Qu.:2009-06-15 06:03:13.75                     
 Max.   :4.000   Max.   :2.329e+09   Max.   :2009-06-25 10:28:27.00                     
      X5                 X6           
 Length:100000      Length:100000     
 Class :character   Class :character  
 Mode  :character   Mode  :character  
```

4. Exploratory Data Analysis (EDA)

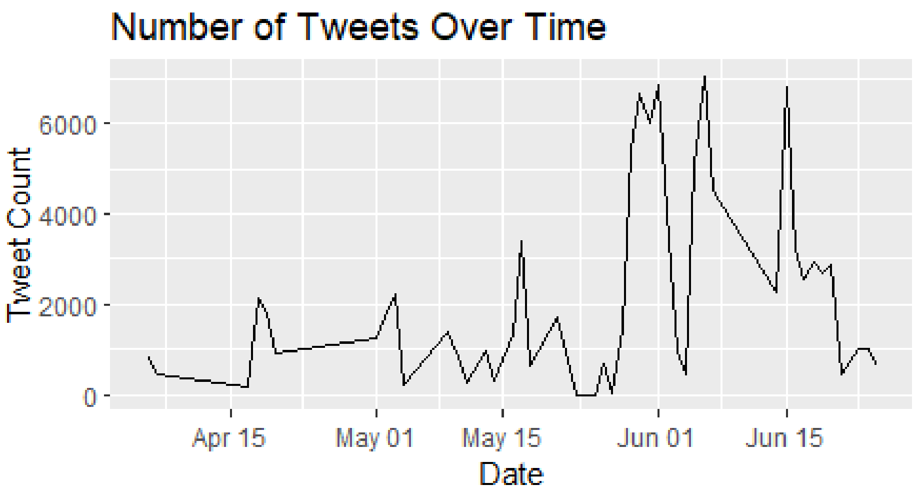
Descriptive Statistics:

The sampled data revealed that most tweets have a negative sentiment value of 0 (negative). The dataset's timestamps range from April 6, 2009, to June 25, 2009.



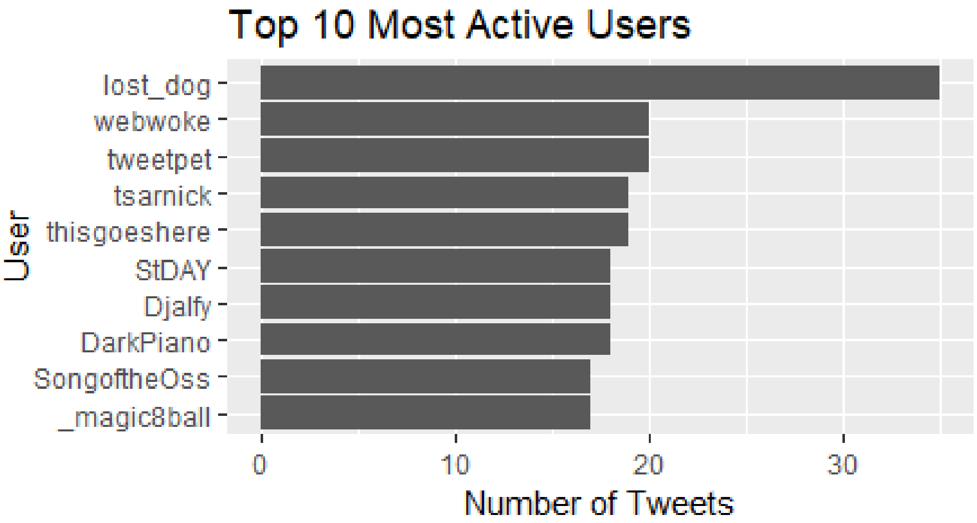
Tweet Frequency Over Time:

The frequency of tweets over time was mapped to understand the trends in tweet activity. The results showed a relatively stable tweet activity over the course of the period.



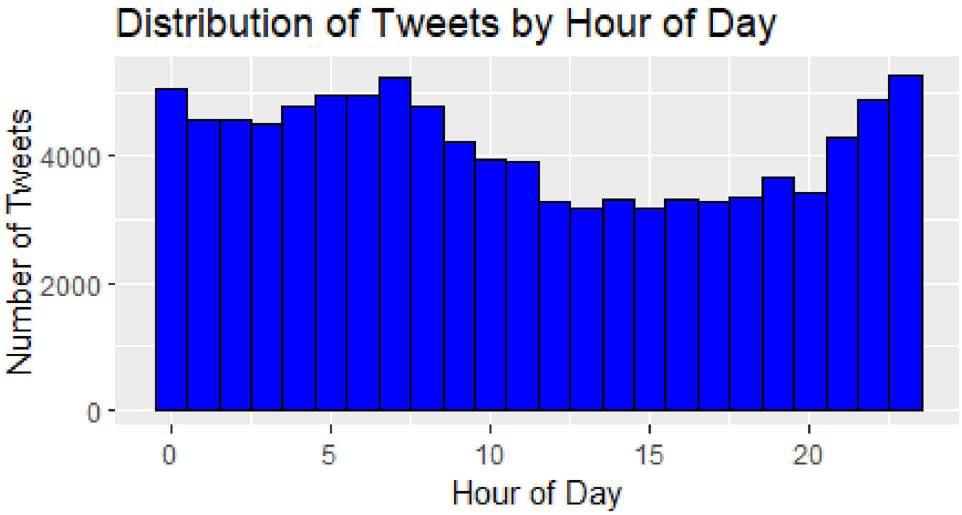
User Activity Analysis:

The most active users were identified and analyzed. The top 10 most active users tweeted more than others, indicating their influence in the dataset.



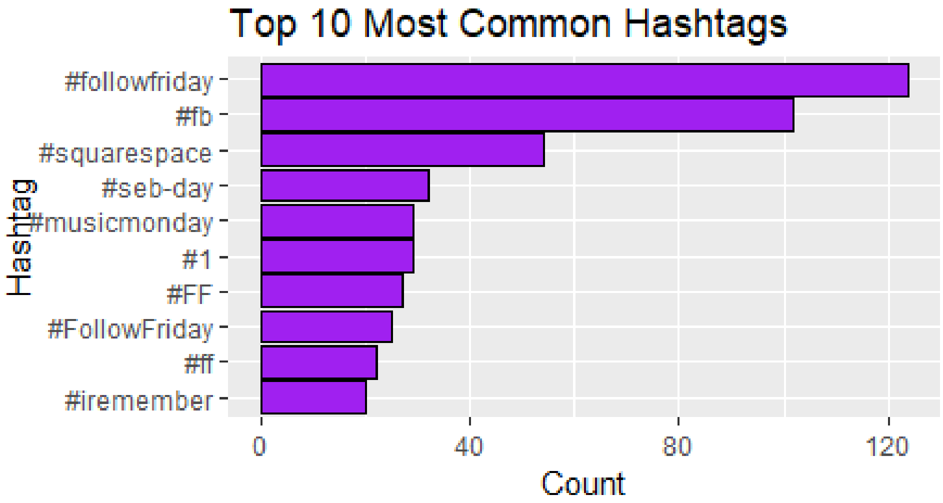
Time of Day Analysis:

Tweets were analyzed by the hour of the day to understand when users are most active. Most tweets were sent during typical waking hours, with peaks in the morning and late afternoon.



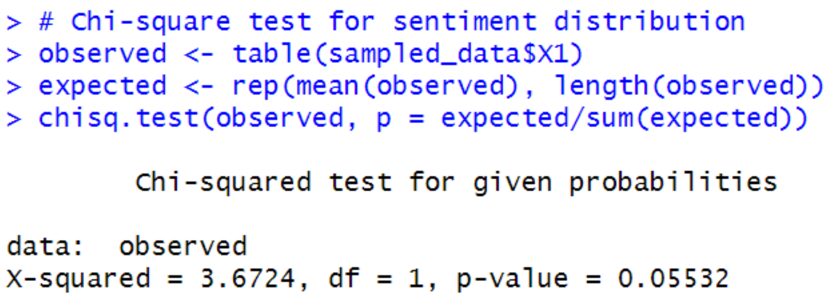
Hashtag Analysis:

The most frequently used hashtags were extracted and analyzed. The top 10 hashtags reflect popular topics of discussion.



5. Inferential Statistics

Chi-Square Test: A chi-square test was conducted to determine if the observed sentiment distribution differs from a uniform distribution. The test result was not statistically significant (p-value = 0.05532), suggesting that the sentiment distribution might not differ significantly from what would be expected under a uniform distribution.

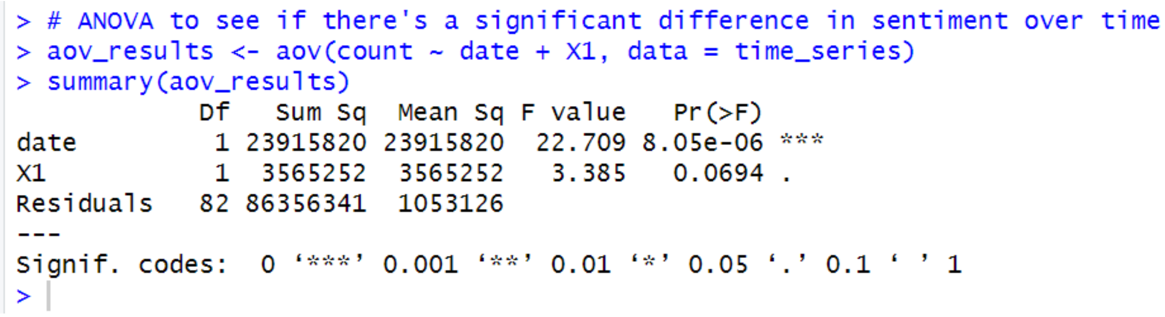


```
> # Chi-square test for sentiment distribution
> observed <- table(sampled_data$X1)
> expected <- rep(mean(observed), length(observed))
> chisq.test(observed, p = expected/sum(expected))

	Chi-squared test for given probabilities

data:  observed
X-squared = 3.6724, df = 1, p-value = 0.05532
```

ANOVA Analysis: An ANOVA was conducted to test if there's a significant difference in sentiment over time. The results indicated that the date is a significant factor affecting the sentiment count (p-value < 0.001).



```
> # ANOVA to see if there's a significant difference in sentiment over time
> aov_results <- aov(count ~ date + X1, data = time_series)
> summary(aov_results)
            Df   Sum Sq  Mean Sq F value   Pr(>F)    
date         1 23915820 23915820  22.709 8.05e-06 ***
X1           1  3565252  3565252   3.385   0.0694 .  
Residuals   82 86356341  1053126                     
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> 
```

6. Interpretation of Results

- **Sentiment Analysis:** The dataset shows a nearly equal distribution of positive and negative sentiments, with a slightly higher occurrence of negative tweets.
- **User Activity:** A small number of users are responsible for a disproportionate number of tweets, indicating potential key influencers in the dataset.
- **Time and Hashtag Trends:** Tweets tend to peak during certain hours of the day, and specific hashtags are recurrent, indicating popular topics.

- **Statistical Testing:** The chi-square test suggested that the sentiment distribution may not significantly differ from a uniform distribution. However, the ANOVA indicated that sentiment does significantly change over time.

7. Recommendations and Future Work

Recommendations:

- Businesses should monitor the identified key users and hashtags closely as they have a significant impact on public sentiment.
- Analyzing sentiment over specific time periods can help identify trends that could inform marketing strategies.

Limitations:

- The analysis is based on a sample of the dataset, which may not fully represent the entire population of tweets.
- The sentiment analysis is simplistic and binary, which may overlook more nuanced expressions of sentiment.

Future Work:

- Expand the sentiment analysis to include more nuanced categories (e.g., neutral, mixed sentiment).
- Incorporate more sophisticated natural language processing techniques to analyze tweet content in greater depth.
- Analyze the impact of external events on sentiment trends to provide more context to the findings.

Link to GitHub:

https://github.com/ms-usmani/-B105-Applied-Statistical-Modelling