UIT2502 - Data Analytics and Visualization Laboratory

App Store : Sentiment Analysis for Strategic Insights

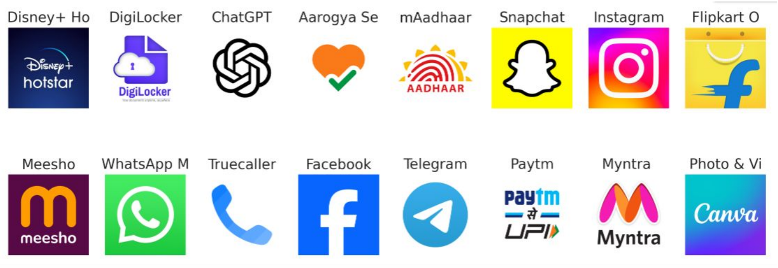
Team Members
B. Sasmitha (3122215002097)
S. Selcia (3122215002098)
Shashwat Shivam (3122215002099)

PROBLEM STATEMENT

- In the dynamic landscape of **mobile applications**, businesses seek to harness the power of **data science to understand user sentiments expressed through app** reviews on platforms like the Google Play Store.
- The objective is to **extract valuable insights** regarding user opinions on various apps, spanning genres from entertainment to utilities.
- This data process aims to enable businesses to adapt their app development and marketing strategies based on public sentiments, providing a comprehensive view of user satisfaction and concerns in the competitive app market.

DATA COLLECTION METHOD

1. Google Play Scraper Library: The `google-play-scraper` library was used for extracting information about mobile applications.A predefined list of Android app package names was specified and the list likely includes a variety of apps from different genres, industries, or developers.



2. Extracting App Details: For each app in the list, the script utilized the `app` function from the `google-play-scraper` library to extract general information about the app.

```
app_infos_df = pd.DataFrame(app_infos)
app_infos_df.head(2)
```

	title	description	descriptionHTML	summary	installs	minInstalls	realInstalls	score	ratings	reviews	...	videoImage	contentRating	contentR
0	Disney+ Hotstar	Disney+ Hotstar is your go-to video streaming ...	Disney+ Hotstar is your go-to video streaming ...	Watch the latest Live Sports, TV and Movies in...	500,000,000+	500000000	746170317	3.968907	11500732	3087364	...	None	Rated for 12+	
1	DigiLocker	DigiLocker is a key initiative under Digital I...	DigiLocker is a key initiative under Digital I...	DigiLocker - a simple and secure document wa...	50,000,000+	50000000	80111521	4.159125	438303	130649	...	None	Rated for 3+	

2 rows × 43 columns

DATA COLLECTION METHOD

3. User Reviews Extraction: The **`google-play-scraper`** library was employed to extract user reviews, considering parameters like language, country, sorting order, and filtering by score. Both the most relevant and newest reviews were collected, providing a diverse set of user opinions.

```
[66] app_reviews_df = pd.DataFrame(app_reviews)
     app_reviews_df.shape

     (19200, 13)
```

```
app_reviews_df.head(2)
```

	reviewId	userName	userImage	content	score	thumbsUpCount	reviewCreatedVersion	at	replyContent	repliedAt	appVersion	so
0	aacc23bf-d882-4893-9ccd-f0a7c7a970d6	Rahul Yadav	https://play-lh.googleusercontent.com/a-/ALV-U...	This app is so frustrating, especially for use...	1	22274	23.09.11.19	2023-10-02 19:30:38	None	NaT	23.09.11.19	most_
1	e9b116bc-42e0-40d3-b8e7-19ccc98e83a4	Snifa D'souza	https://play-lh.googleusercontent.com/a-/ALV-U...	Worst app ever. With a paid subscription, I'm ...	1	11910	23.09.05.4	2023-09-22 13:07:11	Hi! We recommend making sure that your app is ...	2023-09-23 04:27:38	23.09.05.4	most_

```
[68] app_reviews_df.to_csv('reviews.csv')
```

4. Data Storage: The collected information was stored in data frames for further analysis. Output files, such as CSV files (e.g., 'apps.csv' and 'reviews.csv'), were generated to persist the collected data.

STRENGTHS AND LIMITATIONS

STRENGTHS:

1. **Automation:** Efficient Python script using google-play-scraper for automated data collection.
2. **Rich Data:** Captures diverse app details and user reviews, yielding a comprehensive dataset.
3. **Customization:** Flexible parameters (language, country, sorting) enable tailored data collection.

LIMITATIONS:

1. **Platform Dependency:** It relies on Google Play Store structure, which is vulnerable to changes.
2. **Rate Limiting:** Potential slowdown due to Google Play Store's request limits.
3. **Legal/Ethical Considerations:** Adherence to platform terms crucial for ethical and legal compliance.

ACCURACY, RELIABILITY, REPRESENTATIVENESS

ACCURACY:

- The script relies on the robust google-play-scraper library, enhancing data accuracy.
- Occasional errors are addressed through effective error-handling mechanisms, minimizing inaccuracies.

RELIABILITY:

- Error management, implemented via try-except blocks, ensures reliability during data collection.

REPRESENTATIVENESS:

- The dataset strives for representation across diverse app genres and encompasses both positive and negative reviews.

STEPS TAKEN IN DATA PREPROCESSING

1. Text Preprocessing:

- ➢ **Tokenization:** CountVectorizer` to convert the text into a numerical representation.
- ➢ **Stop Word Removal:** Common English stop words were removed during tokenization to focus on meaningful content.

```
Transformed Text (CountVectorizer):
  (0, 5174)	1
  (0, 2981)	2
  (0, 7109)	1
  (0, 2514)	1
  (0, 2276)	1
  (0, 7101)	2
  (0, 3522)	1
  (0, 7156)	1
```

2. Numerical Preprocessing:

- ➢ **Imputation:** Missing values in numerical columns were filled with the mean value to handle potential data gaps.
- ➢ **Scaling:** Numerical features were standardized to bring them to a common scale, preventing features with larger magnitudes from dominating.

```
Transformed Numerical Features:
[[-0.68496017 -0.04865135]
 [ 0.12588077 -0.12386665]
 [ 0.12588077 -0.12386665]
 ...
 [ 0.12588077 -0.1231145 ]
 [ 0.12588077 -0.12386665]
 [ 0.93672172 -0.12386665]]
```

STEPS TAKEN IN DATA PREPROCESSING

3. Categorical Preprocessing:

- **One-Hot Encoding:** to convert categorical data into a numerical format
- The `handle_unknown='ignore'` parameter was used in one-hot encoding to handle new categories that may appear in the test set but not in the training set.

```
Transformed Categorical Features (One-Hot Encoding):
  (0, 4)	1.0
  (1, 0)	1.0
  (2, 4)	1.0
  (3, 2)	1.0
  (4, 6)	1.0
  (5, 3)	1.0
  (6, 3)	1.0
  (7, 3)	1.0
  (8, 7)	1.0
  (9, 7)	1.0
  (10, 1)	1.0
```

4. Column Transformation:

- Combines the preprocessing steps for text, numerical, and categorical data. Columns not specified in the preprocessing steps were passed through without transformation (`remainder='passthrough'`).

```
Transformed Data (Combined Text, Numerical, and Categorical Features):
  (0, 50)	1.0
  (0, 396)	1.0
  (0, 616)	3.0
  (0, 1094)	1.0
  (0, 1211)	2.0
  (0, 1338)	1.0
  (0, 1547)	1.0
  (0, 1565)	1.0
  (0, 1753)	2.0
  (0, 1893)	1.0
  (0, 2276)	1.0
  (0, 2423)	1.0
  (0, 2514)	1.0
  (0, 2675)	1.0
  (0, 2981)	2.0
  (0, 3021)	1.0
  (0, 3495)	1.0
  (0, 3522)	1.0
```

HANDLING INCONSISTENCIES

1. Handling Missing Values:

- **Imputation:** For numerical features (score and thumbsUpCount), the missing values were filled with the mean. This helps address missing values in numerical columns.

2. Handling Outliers:

- **Scaling:** Numerical features (score and thumbsUpCount) are scaled using StandardScaler, reducing the impact of outliers normalizing the range of the features.

3. Handling Inconsistencies:

- **One-Hot Encoding:** For the categorical feature appId, one-hot encoding is applied which ensures that categorical variables are represented in a consistent format suitable for machine learning models.

4. Handling Text Data:

- **Text Vectorization:** The CountVectorizer is used to convert text data in the content column into numerical features.

SERIES OF TRANSFORMATIONS

1. **Text Preprocessing (CountVectorizer):**
   - Applied CountVectorizer to 'content' column, Converted text reviews into a matrix of word counts, Removed stop words.
2. **Numerical Preprocessing (Imputation and Scaling):**
   - Imputed missing values in 'score' and 'thumbsUpCount' with mean. Standardized numerical features using StandardScaler.
3. **Categorical Preprocessing (One-Hot Encoding):**
   - One-hot encoded the 'appId' column. Converted categorical variable into binary vectors.
4. **Column Transformation (Combining Features):**
   - Used ColumnTransformer to combine transformed text, numerical, and categorical features.

```
21s
K-Neighbors Classifier Accuracy: 0.73

Random Forest Classifier Accuracy: 0.91

SVC Accuracy: 0.97

Logistic Regression Accuracy: 1.00
```

DESCRIPTIVE ANALYSIS

```
Mean Value of Score: 3.0
Median Value of Score: 3.0
```

```
Mean Thumbs Up Count: 355.8346875
Median Thumbs Up Count: 0.0
```

Scores:

- The median and mean scores are **equal**, indicating a **relatively balanced distribution.**

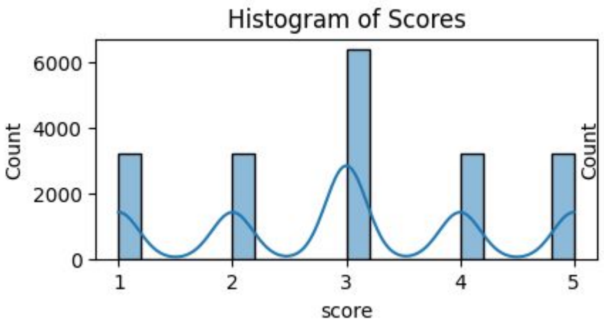
Thumbs-up Count:

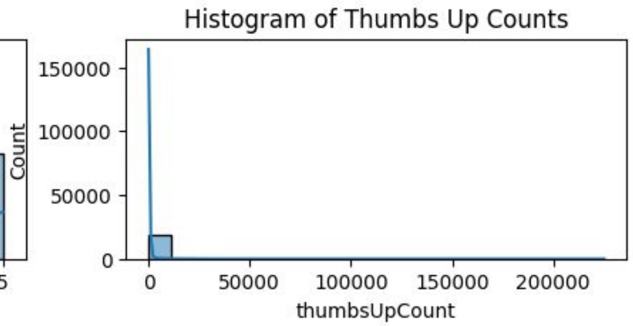
- For the thumbs-up count, the median of 0 suggests that there are likely **many reviews with no thumbs-up,**
- The mean is higher, indicating the **presence of reviews with a significant number of thumbs-up.**

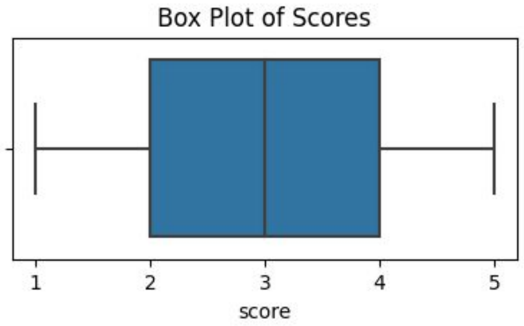
WORD CLOUD - Frequency of Words

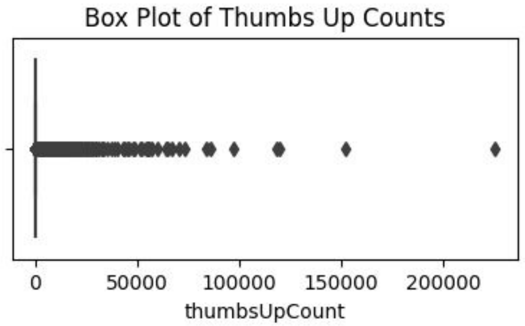
1. Words like "good", "great" clearly indicates positive sentiments expressed by users, which can be generated using the **WordCloud library in Python**.
2. Terms such as "problem," "issue," or "bug" might indicate areas of dissatisfaction.
3. Specific features like "app," "feature," and "option" stand out, providing insights into aspects frequently mentioned by users.

DATA VISUALIZATION









Scores:

- The histogram shows that a substantial number of reviews have scores around 3.0, indicating a moderate satisfaction level.
- The box plot of scores indicates a relatively balanced distribution, with median around 3.0. There are no apparent outliers.

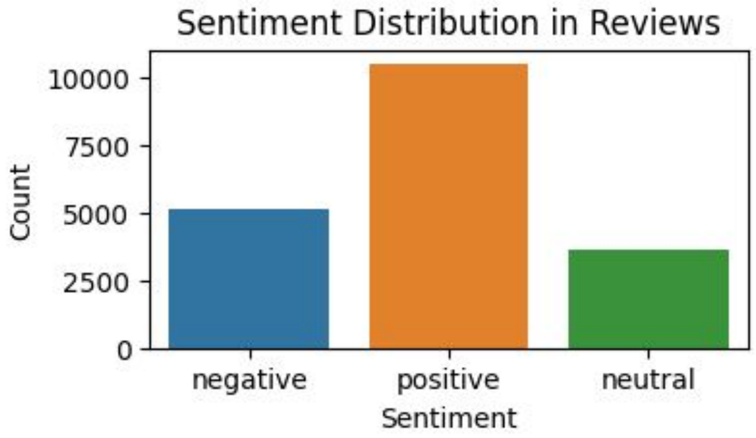
Thumbs-up Count:

- The histogram of thumbs-up count is right skewed, suggesting that most reviews have lower number of thumbs-up, with few having significantly higher counts.
- The box plots of thumbs-up count shows the presence of outliers, with few reviews having a considerably higher number of thumbs-up.

OVERALL SENTIMENT ANALYSIS

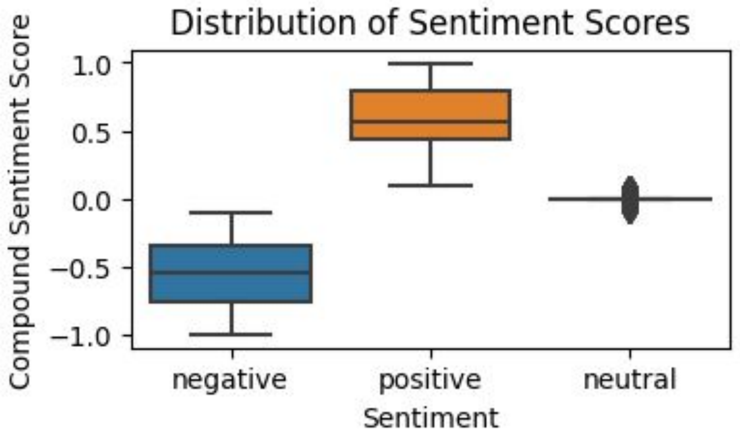
Sentiment Distribution Plot:

- The majority of reviews tend to have a positive sentiment.
- Negative sentiments are present but in a smaller proportion.
- The distribution of sentiments indicates a generally **favorable perception of the apps**, as most reviews fall into the positive category.

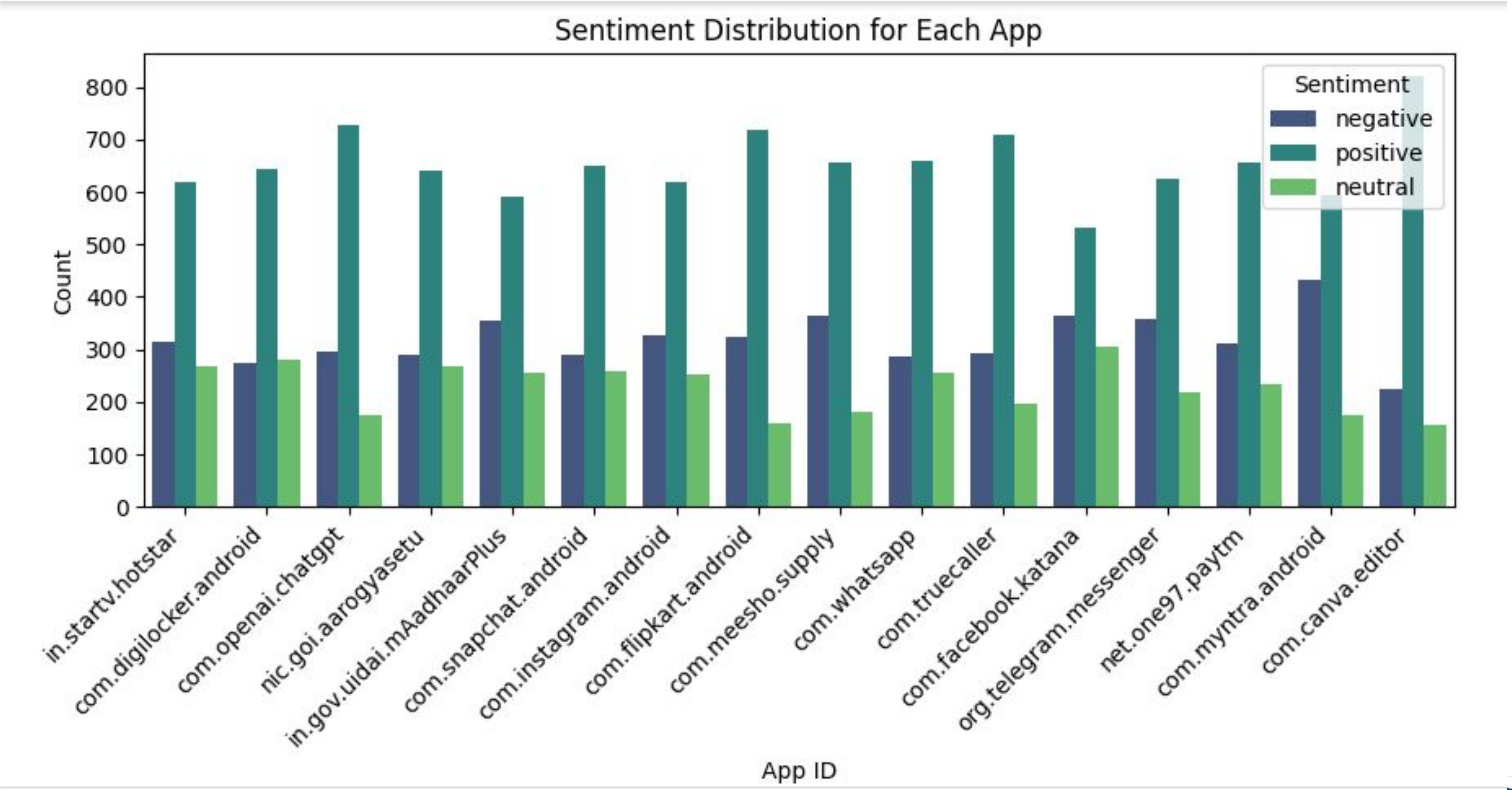


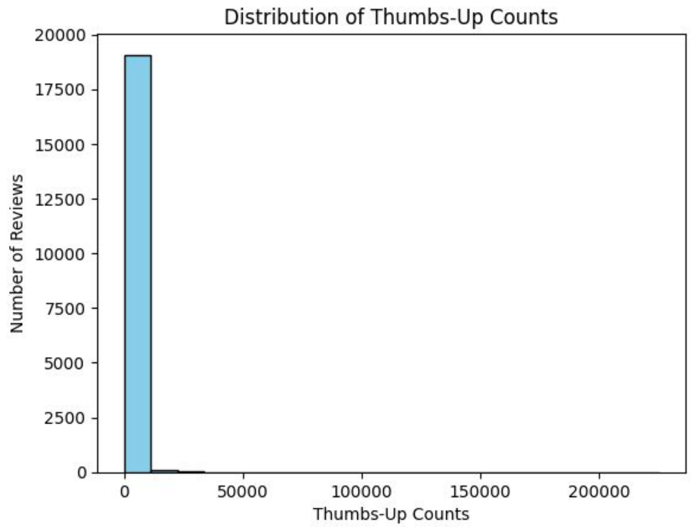
Box Plot of Sentiment Scores:

- Positive sentiments - higher median compound score, indicating a consistently positive tone in these reviews.
- Negative sentiments - wider spread, suggesting variability in the intensity of negative opinions.
- Neutral sentiments have a relatively lower median score, reflecting a more neutral tone in these reviews.

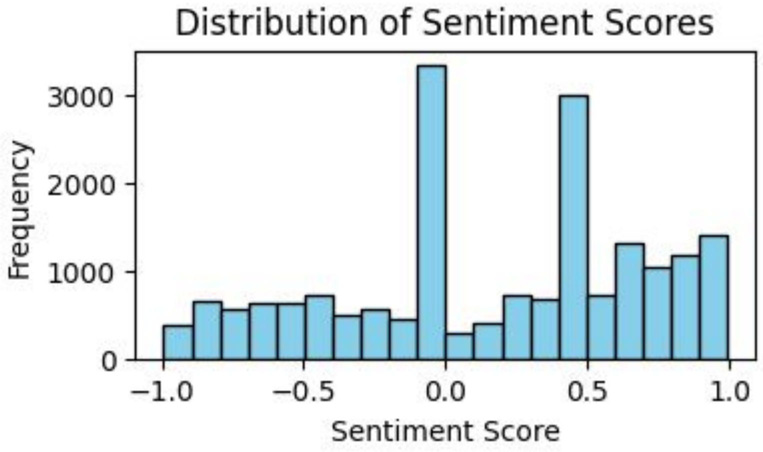


SENTIMENT DISTRIBUTION FOR EACH APP

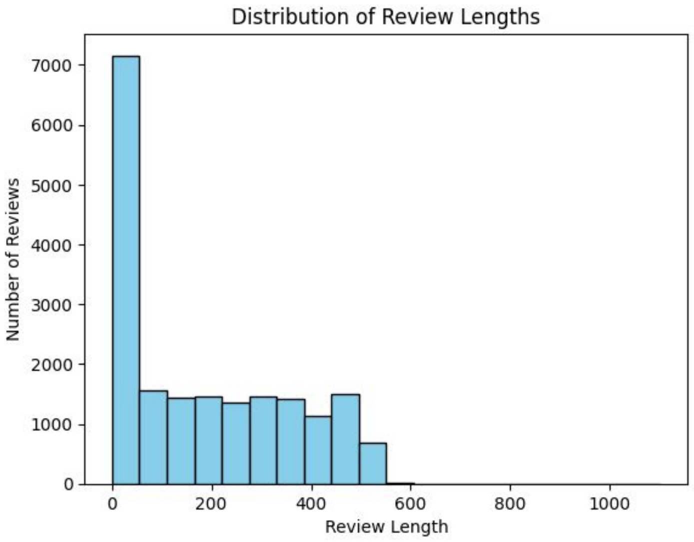




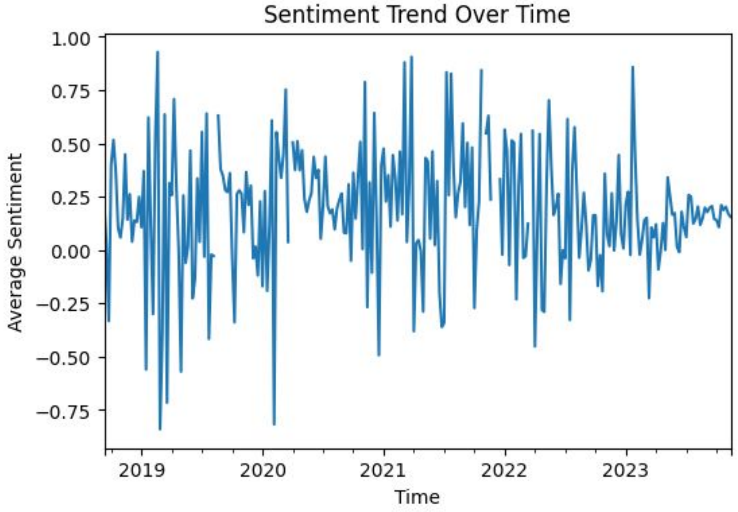
Long tail distribution

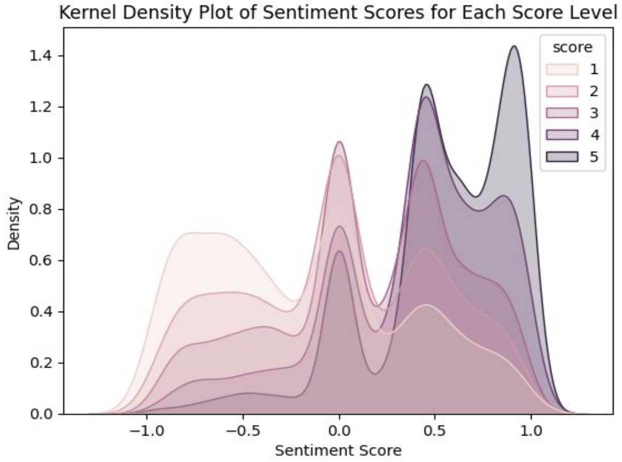


Bimodal distribution

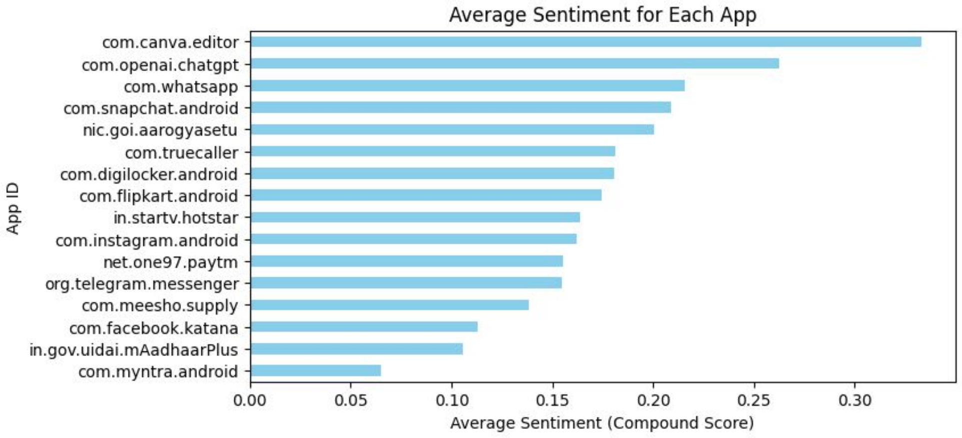


Positively skewed distribution

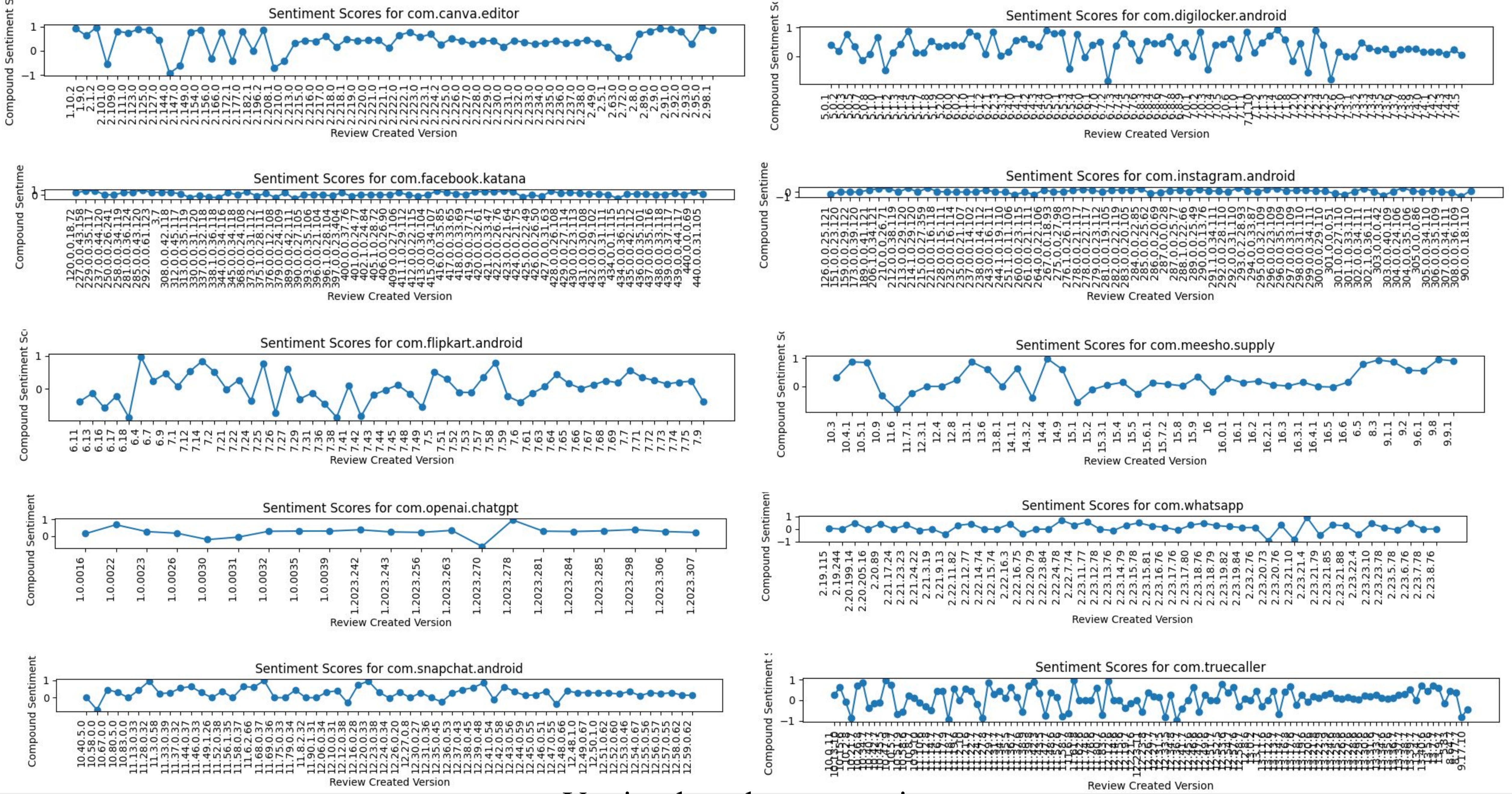




Score Segmentation



App Segmentation



Version-based segmentation

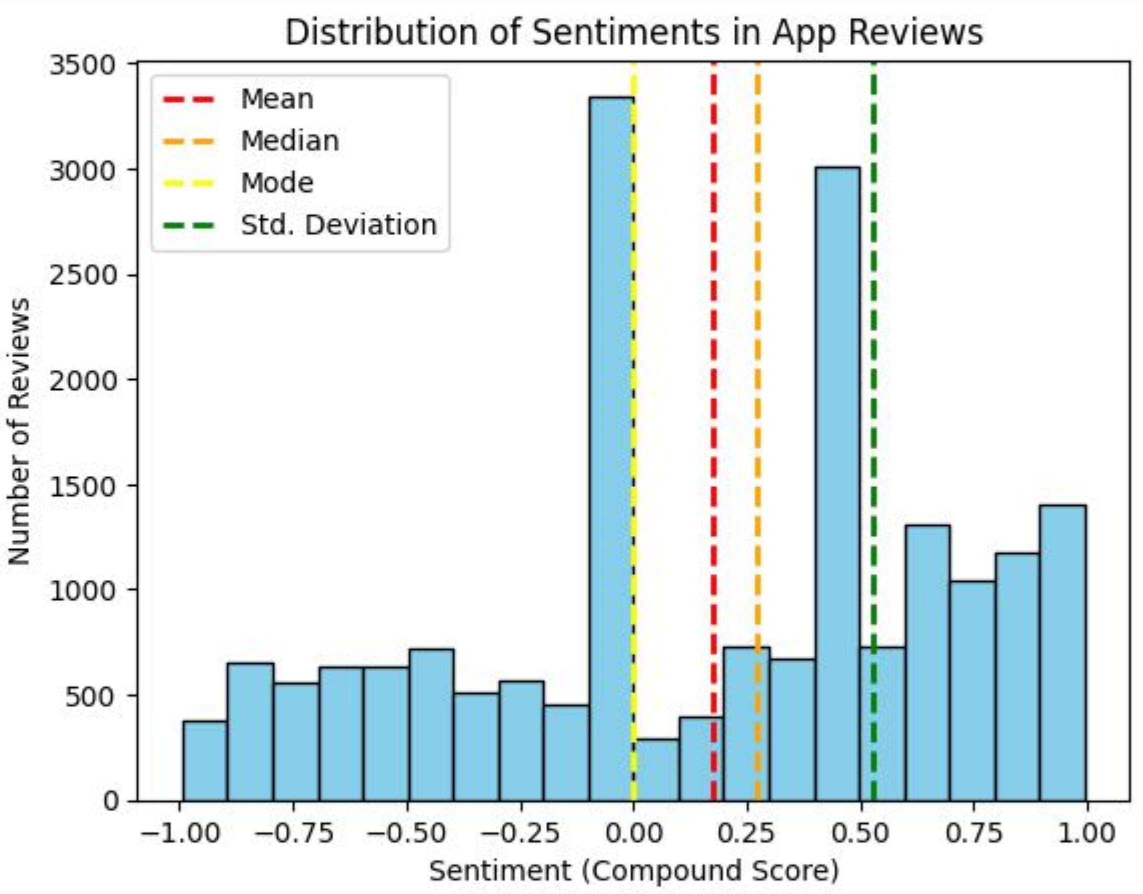
DESCRIPTIVE STATISTICS

1. Mean Sentiment: 0.176, which is slightly positive. The reviews are leaning towards positive sentiments.

2. Median Sentiment: 0.2732, which is also positive. The median is less affected by extreme values and gives a sense of the central tendency of sentiments.

3. Mode Sentiment: 0.0, indicating a significant number of reviews express neutral sentiments.

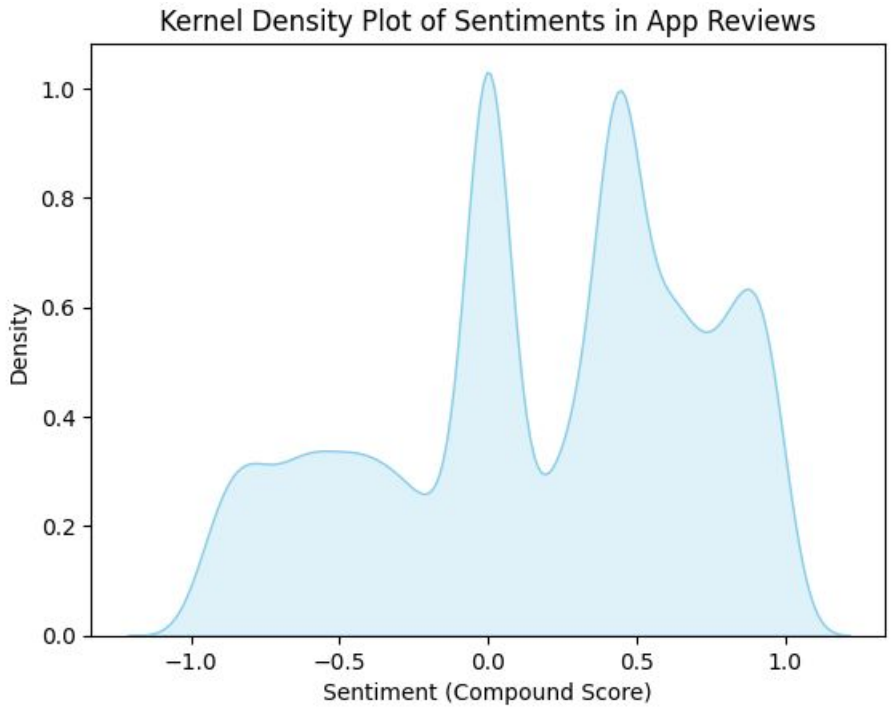
4. Standard Deviation Sentiment: 0.530, indicating a relatively high variability in sentiment scores. This suggests that there is a wide range of sentiments in the dataset, with some reviews being very positive and others very negative



CENTRAL TENDENCY & VARIABILITY

Central Tendency (Mean, Median, Mode):

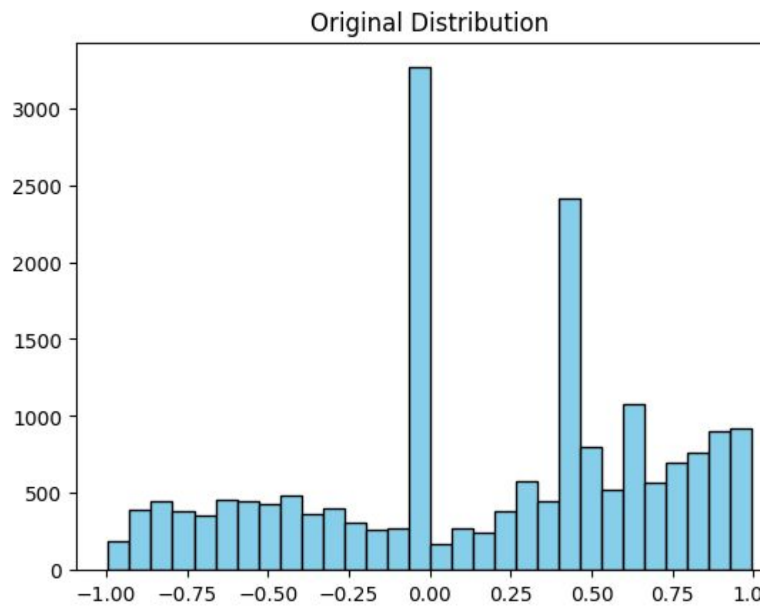
1. Mean - overall sentiment orientation of users is positive implying user satisfaction.
2. Median - slightly higher than the mean indicates distribution might be right skewed.
3. Mode - significant presence of neutral reviews.

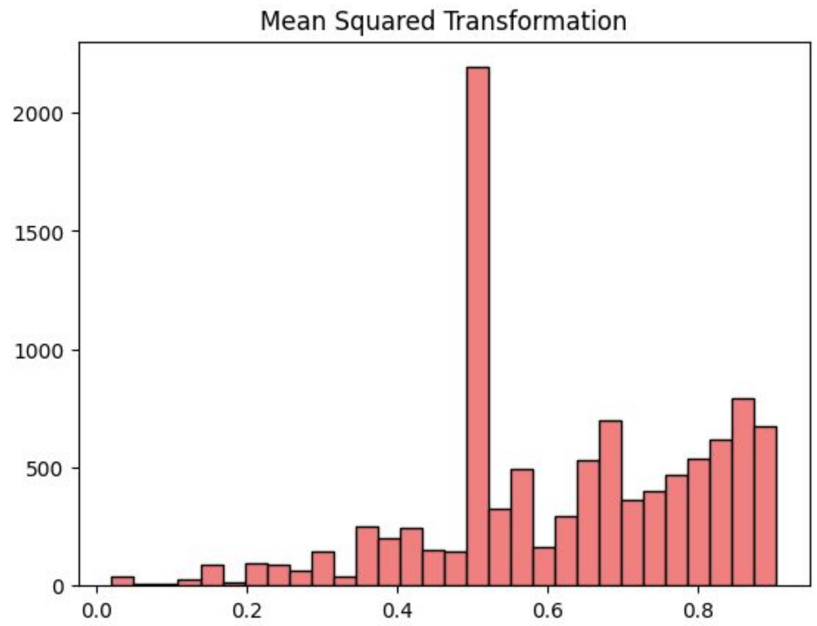


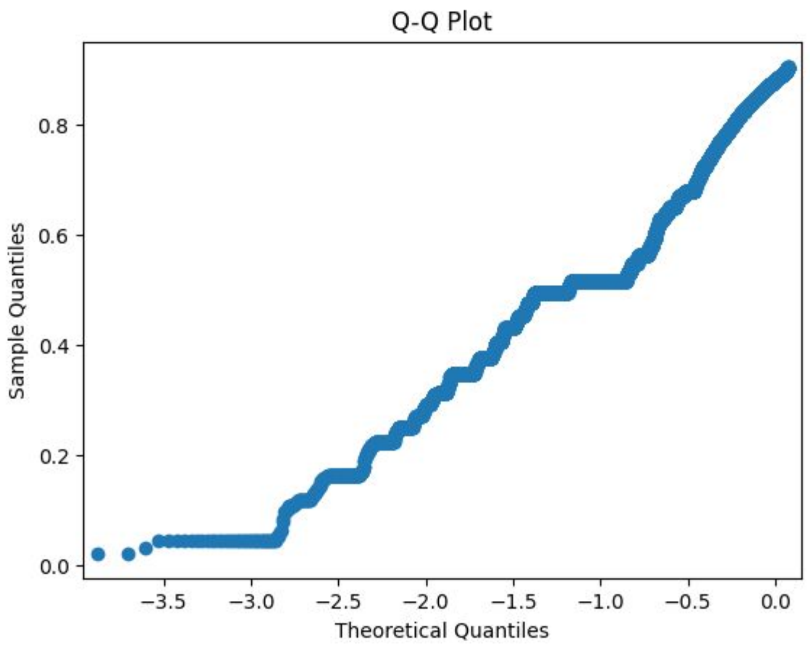
Variability (Standard Deviation):

1. Standard deviation - high standard deviation implies a wide variety of sentiment scores.

NORMAL CURVES







1. Histograms Plots: Checked the distribution visually through histograms and kernel density plots.

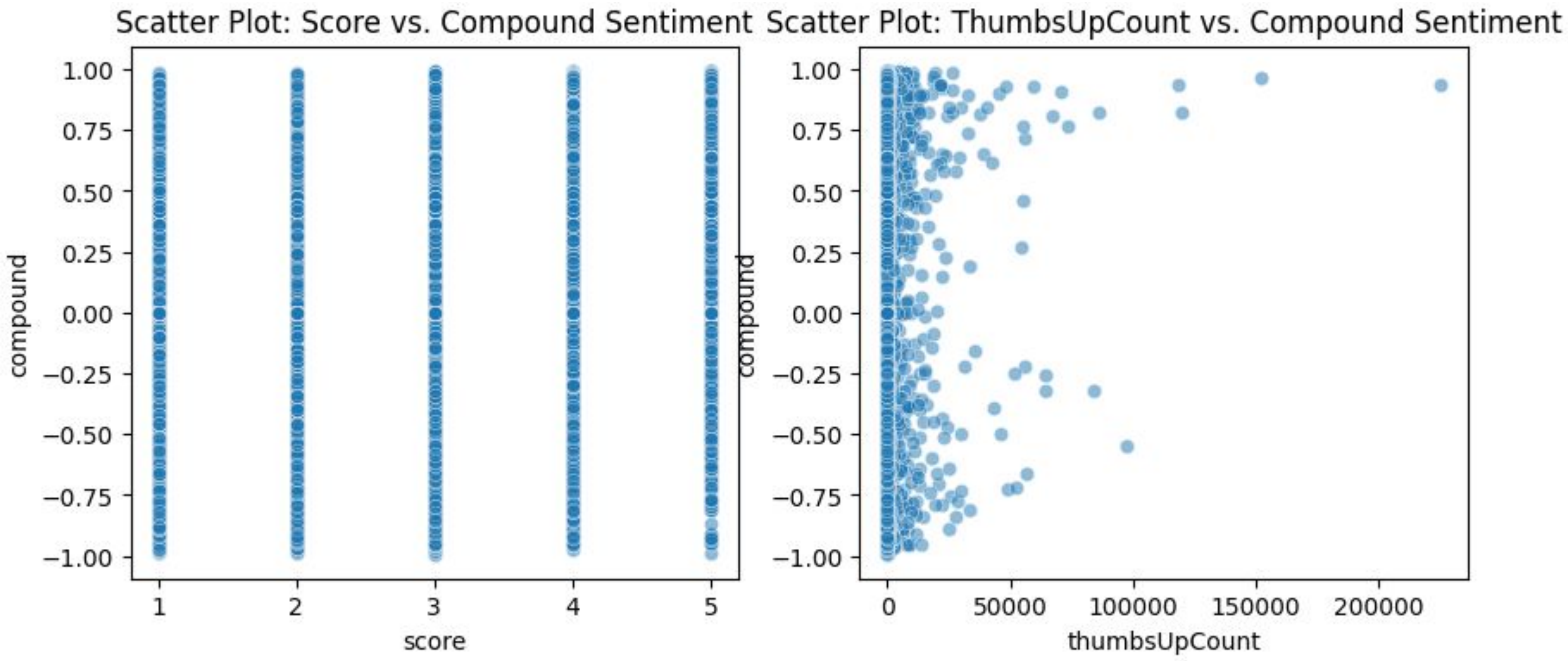
2. Q-Q Plots (Quantile-Quantile Plots): Examined if the data quantiles match those of a normal distribution.

3. Statistical Tests: Employed the Shapiro-Wilk Test to formally assess normality.

```
Shapiro Wilk Result: Data appears to be normally distributed.
```

CORRELATION COEFFICIENT

IV: score, thumbsupcount DV: compound




```
Correlation coefficient between "score" and "compound": 0.44
Correlation coefficient between "thumbsUpCount" and "compound": 0.01
```

CORRELATION COEFFICIENT

1. Correlation coefficient between "score" and "compound": 0.44 - positive correlation coefficient of 0.44 indicates a moderate positive linear relationship, if user's rating increases, there is a tendency for the sentiment compound score to also increase.

2. Correlation coefficient between "thumbsUpCount" and "compound": 0.01 - very small positive correlation coefficient of 0.01 indicates an extremely weak positive linear relationship almost no linear association between the number of thumbs-up counts and the sentiment compound score.

Interpretation:

- For "score," users who give higher ratings tend to have slightly more positive sentiment in their reviews.
- For "thumbsUpCount," the number of thumbs-up counts is not a strong predictor of the sentiment expressed in the reviews.

OLS REGRESSION MODEL

- **Extract the relevant columns** for the regression model, such as 'score,' 'thumbsUpCount,' and 'compound.'
- Use a **regression model**, such as Ordinary Least Squares (OLS), to **fit the chosen independent variables** ('score' and 'thumbsUpCount') to the **dependent variable** ('compound').
- Utilize libraries like Statsmodels or Scikit-learn to perform the **regression analysis** and obtain coefficients.

```
                            OLS Regression Results
==============================================================================
Dep. Variable:               compound   R-squared:                       0.195
Model:                            OLS   Adj. R-squared:                  0.195
Method:                 Least Squares   F-statistic:                     2327.
Date:                Wed, 15 Nov 2023   Prob (F-statistic):               0.00
Time:                        19:28:21   Log-Likelihood:                -12955.
No. Observations:               19200   AIC:                         2.592e+04
Df Residuals:                   19197   BIC:                         2.594e+04
Df Model:                           2
Covariance Type:            nonrobust
=================================================================================
                    coef    std err          t      P>|t|      [0.025      0.975]
---------------------------------------------------------------------------------
const            -0.3688      0.009    -42.439      0.000      -0.386      -0.352
score             0.1812      0.003     68.190      0.000       0.176       0.186
thumbsUpCount  3.561e-06   9.62e-07      3.701      0.000    1.68e-06    5.45e-06
==============================================================================
Omnibus:                      584.783   Durbin-Watson:                   1.877
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              448.780
Skew:                          -0.287   Prob(JB):                     3.54e-98
Kurtosis:                       2.518   Cond. No.                     9.43e+03
==============================================================================
```

COEFFICIENTS OF REGRESSION MODEL

- **Intercept (const): -0.3688**
  - Interpretation: When both 'score' and 'thumbsUpCount' are zero, the predicted 'compound' sentiment is approximately -0.369.
- **Coefficient for 'score': 0.1812**
  - Interpretation: Holding 'thumbsUpCount' constant, a one-unit increase in 'score' is associated with a 0.1812 increase in the predicted 'compound' sentiment.
- **Coefficient for 'thumbsUpCount': 3.561e-06**
  - Interpretation: Holding 'score' constant, a one-unit increase in 'thumbsUpCount' is associated with a very small (3.561e-06) increase in the predicted 'compound' sentiment.

	coef
const	-0.3688
score	0.1812
thumbsUpCount	3.561e-06

ADDRESSING LIMITATIONS AND POTENTIAL IMPROVEMENTS

Limitations:

- Assumes a linear relationship, which may not capture nonlinear patterns.
- Relies on certain assumptions like normality of errors.
- Sensitivity to outliers.

Potential Improvements:

- Explore non-linear relationships using polynomial terms or other regression techniques.
- Validate assumptions and consider transformations if needed.
- Address outliers through robust regression or outlier removal.

Z - TEST

Two independent groups in the sample that can be tested using a z-test could be:

Users who gave a high 'score' (e.g., rating above a certain threshold) versus users who gave a low 'score' (below the threshold).

1. Formulate Hypotheses:

- Null Hypothesis (H0): The mean 'compound' sentiment score for users with high 'score' is equal to the mean for users with low 'score'.
- Alternative Hypothesis (H1): The mean 'compound' sentiment score for users with high 'score' is not equal to the mean for users with low 'score'.

2. Collect Data:

- Obtain independent samples from the two groups based on a 'score' threshold.

```
Sample mean 1: 0.483752
Sample mean 2: 0.05517849999999999
Sample SD 1: 0.42545882781631666
Sample SD 2: 0.5072004347323783
Sample size 1: 200
Sample size 2: 200
```

3. Calculate Sample Means and Standard Deviations:

- Compute the sample mean and standard deviation for 'compound' in each group.

Z - TEST

4. Specify the Significance Level (α):

- Choose a significance level (e.g., 0.05).

```
Reject the null hypothesis: There is a significant difference between the means.
Z-statistic: 9.1553
```

5. Calculate the Z-Statistic:

- Use the formula for the z-statistic to compare the means of the two groups.

6. Determine Critical Region:

- Find the critical z-values corresponding to the chosen significance level for a two-tailed test.

7. Make a Decision:

- If the calculated z-statistic falls into the critical region, reject the null hypothesis. Otherwise, fail to reject the null hypothesis.

8. Draw Conclusions:

- Conclude whether there is enough evidence to suggest a significant difference between the means of the two groups.

ASSUMPTIONS & LIMITATIONS : Z - TEST

Assumptions:

1. Data should be normally distributed within each group.
2. Requires knowledge of the population standard deviation.
3. Data should be collected through random sampling.

Potential Limitations:

1. More reliable with larger sample sizes; smaller samples may favor the t-test.
2. Assumes knowledge of population standard deviation; t-test used when unknown.
3. Susceptible to outliers, impacting accuracy; consider transformations or non-parametric tests.
4. Deviation from normality affects reliability, especially with smaller samples; consider non-parametric tests.
5. Best suited for continuous data; categorical data may require alternative tests.
6. Requires parameters like population standard deviation, which may not be available.
7. Assumes a fixed Type I error rate; caution needed with multiple testing.

t - TEST

1. **Formulate Hypotheses:**
   - Null Hypothesis (H0): The mean 'compound' sentiment score for users with high 'score' (above a certain threshold) is equal to the mean for users with low 'score'.
   - Alternative Hypothesis (H1): The mean 'compound' sentiment score for users with high 'score' is not equal to the mean for users with low 'score'.
2. **Collect Data:**
   - Obtain dependent samples from the two groups based on a 'score' threshold.
3. **Calculate Sample Means and Standard Deviations:**
   - Compute the sample mean and standard deviation for 'compound' in each group.
4. **Specify the Significance Level (α):**

- Choose a significance level (e.g., 0.05).

5. **Calculate the t-Statistic:**

- Use the formula for the t-statistic to compare the means of the two groups.

t - TEST

6. Determine Critical Region:

- Find the critical t-values corresponding to the chosen significance level for a two-tailed test.

7. Make a Decision:

- If the calculated t-statistic falls into the critical region, reject the null hypothesis. Otherwise, fail to reject the null hypothesis.

8. Draw Conclusions:

- Conclude whether there is enough evidence to suggest a significant difference between the means of the two groups.

```
T-Statistic: 50.0789
P-Value: 0.0000
Reject the null hypothesis: There is a significant difference in mean 'compound' scores.
```

ASSUMPTIONS & LIMITATIONS : t - TEST

Assumptions of the t-Test:

1. Data in each group should ideally follow a normal distribution, but t-tests are robust with larger sample sizes.
2. Variances of the compared groups should be roughly equal, especially in the independent samples t-test.
3. Data should be measured on an interval or ratio scale.
4. Observations should be randomly and independently selected.

Potential Limitations:

1. T-tests can be sensitive to extreme values, particularly with small sample sizes.
2. Assumes independence between observations in different groups.
3. Performance may be limited with very small sample sizes, especially for non-normally distributed data.
4. P-values require careful interpretation and don't indicate effect size or practical significance.
5. Violating the equal variances assumption may impact result accuracy in independent samples t-test.

ANOVA

1. Formulate Hypotheses:

- Null Hypothesis (H0): The mean 'compound' values are equal across all groups.
- Alternative Hypothesis (H1): At least one group has a different mean 'compound' value.

2. Organize Data:

- Organize your data into groups. Each group should represent a distinct category or level.

3. Check Assumptions:

- Check the assumption of normality within each group. You can use normality tests or visual inspections like histograms.
- Check the assumption of homogeneity of variances, meaning that the variance within each group should be roughly equal.

4. Perform ANOVA:

- Use an ANOVA test to evaluate whether there are statistically significant differences in the means of the 'compound' values among the groups.

ANOVA

5. Interpret Results:

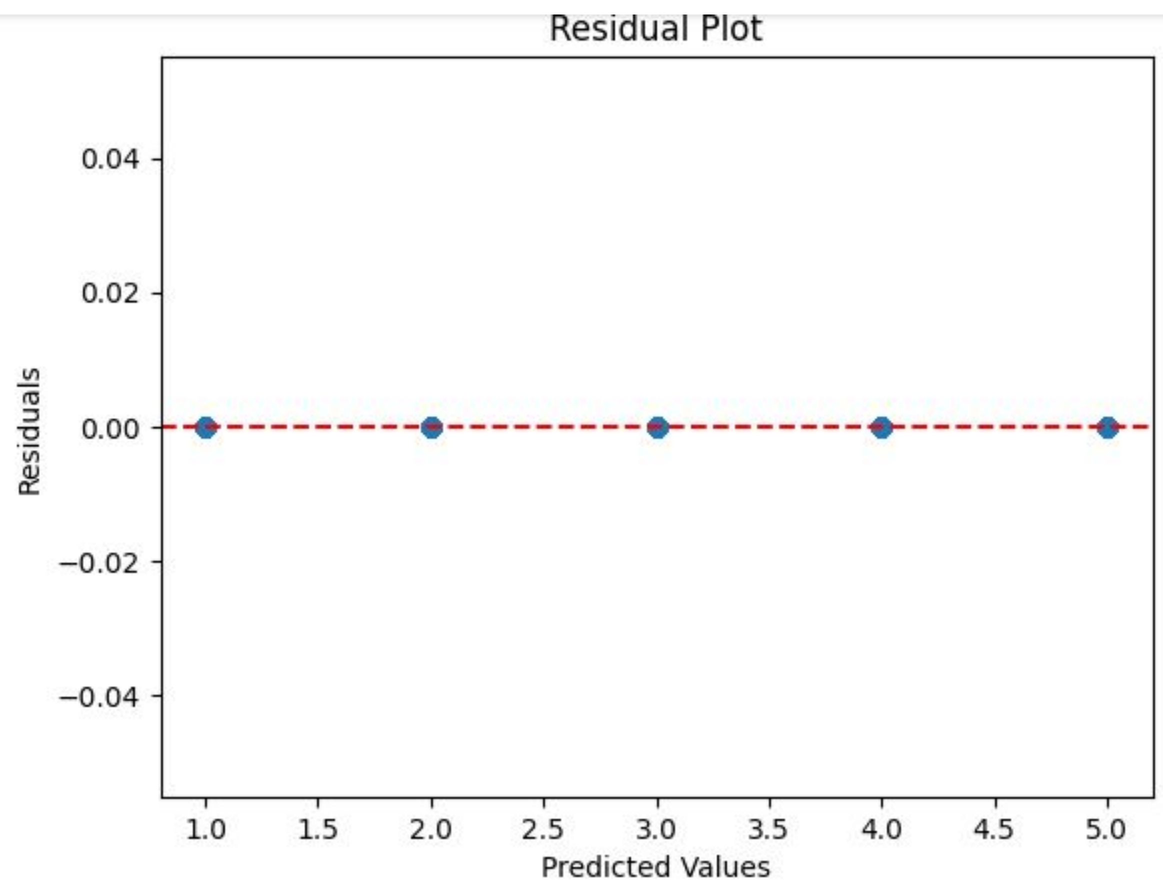
- Examine the F-statistic and associated p-value from the ANOVA output.
- If the p-value is below your chosen significance level (commonly 0.05), you reject the null hypothesis and conclude that there is a significant difference in at least one group mean.

6. Post-hoc Tests (if needed):

- If you have more than two groups and the ANOVA indicates significance, consider post-hoc tests (e.g., Tukey's HSD) to identify which specific groups differ from each other.

```
F-Statistic: 1918.3562
P-Value: 0.0000
Reject the null hypothesis: There is a significant difference between group means.
```

BUILDING AND VALIDATING LINEAR MODELS

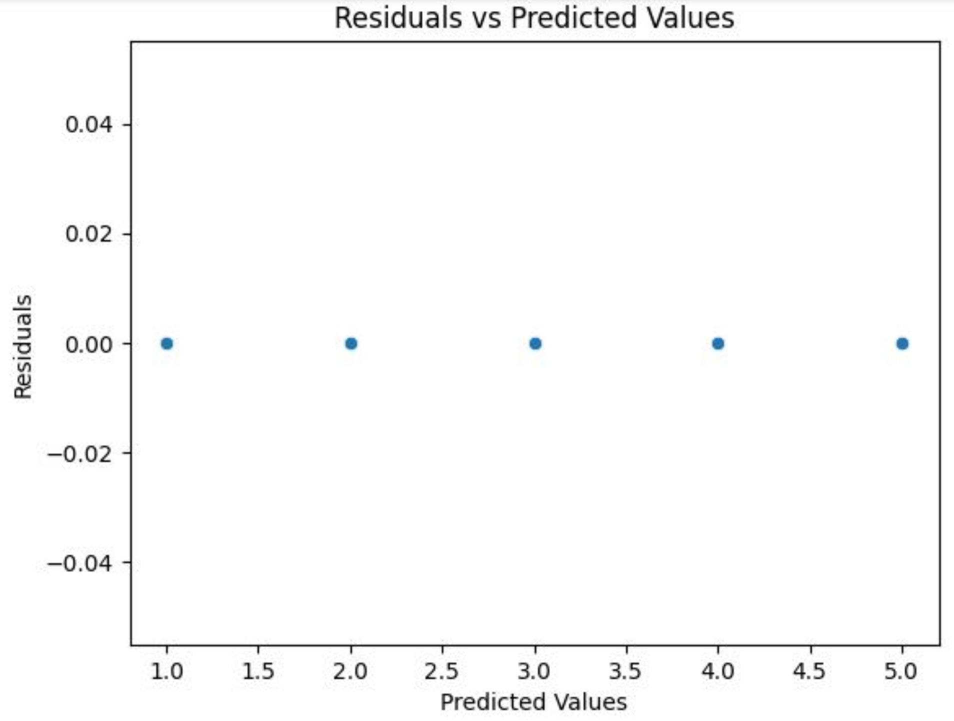



```
K-Neighbors Classifier Accuracy: 0.73

Random Forest Classifier Accuracy: 0.91

SVC Accuracy: 0.98

Logistic Regression Accuracy: 1.00
```



BUILDING AND VALIDATING LINEAR MODELS

Strategies to validate the performance :

1. Train-Test Split: Divide data into training and test sets to check model performance on unseen data.
2. Cross-Validation: Use k-fold validation to assess model robustness across different data subsets.
3. Residual Analysis: Check for patterns in residuals to verify model assumptions.
4. Diagnostic Plots: Create plots like Q-Q plots and scatterplots to evaluate model fit.
5. Performance Metrics: Employ metrics (e.g., R-squared, RMSE) to quantify model accuracy

Best Practises:

1. Evaluate Metrics: Use R-squared, RMSE, or MAE to assess performance.
2. Test Data Validation: Confirm performance on a separate unseen dataset.
3. Iterate and Refine: Improve the model based on validation results.
4. Interpret and Explain: Ensure model interpretability for stakeholders.
5. Document and Communicate: Record the process and findings clearly for effective communication.