INDEX - 20020139

Data Analytics Report - Insights from Uber Reviews

1. Introduction

This report presents a comprehensive analysis of the "Uber Reviews Without Reviewid" dataset, which comprises 12,000 reviews. The primary objective is to understand customer sentiment and service quality through text analytics, statistical examination, and clustering. The findings aim to inform Uber's strategic decisions regarding service improvements and customer engagement.

2. Data Overview and Preprocessing

2.1 Dataset Description

- **Source & Size**
  The dataset contains 12,000 records with 10 attributes, including:
  - **userName:** Reviewer identifier
  - **content:** Review text
  - **score:** Numeric rating (predominantly 5-star)
  - **thumbsUpCount:** Engagement metric
  - **at:** Timestamp of the review
  - **reviewCreatedVersion & appVersion:** Version details
  - **replyContent & repliedAt:** Response information (sparse)
- **Missing Values**
  - *userImage* is missing for all entries.
  - *replyContent* and *repliedAt* are nearly entirely missing (only 33 non-null values).
  - Version fields have partial missingness (~17–18% missing).
  - The key fields *content* and *score* are complete.

2.2 Data Cleaning & Feature Engineering

- **Cleaning -** Rows with missing *content* or *score* were removed.
- **Date Conversion -** The *at* column was converted to datetime format.
- **Derived Feature -** A new column, **Review_Length**, was created by computing the number of characters in each review. This helps in understanding how review verbosity might relate to ratings.

---

3. Exploratory Data Analysis (EDA)

3.1 Summary Statistics

- **Score Distribution -** The analysis revealed that the majority of reviews are 5-star ratings, with only a few instances of lower scores. This suggests an overall high level of customer satisfaction.
- **Review Length -** Although most reviews are brief, there is noticeable variability, indicating that while many customers leave short feedback, some provide more detailed commentary.

3.2 Visualizations

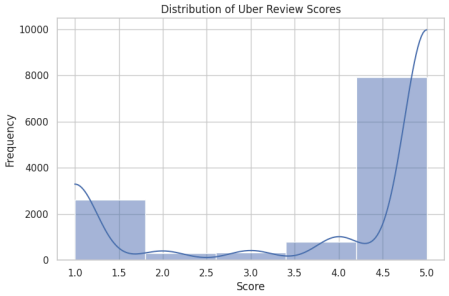


Figure 1. Histogram of Review Scores

This histogram demonstrates a strong skew towards 5-star ratings, confirming the overall positive sentiment.

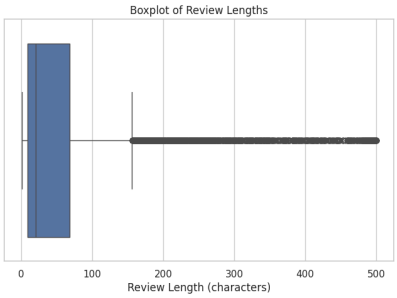


Figure 2. Boxplot of Review Lengths
The boxplot shows that while the median review length is short, outliers indicate a range of verbosity in customer feedback.

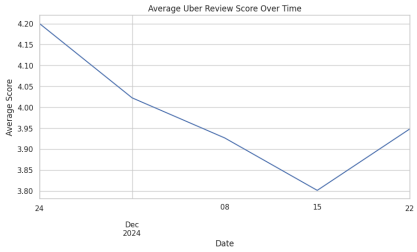


Figure 3. Time-Series Plot of Average Review Score

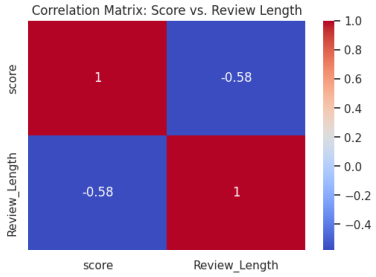
The weekly resampled time-series plot of the average review score remains consistently high, with minor fluctuations that may correspond to service updates or external events.

4. Statistical and Text Analysis

4.1 Correlation Analysis

A correlation matrix was computed to assess the relationship between the review score and review length.

Figure 4. Correlation Heatmap (Score vs. Review_Length)



The correlation is weak, suggesting that the length of a review does not significantly predict the given score.

4.2 Sentiment Analysis

Using TextBlob, each review was assigned a sentiment polarity score. The sentiment analysis indicates predominantly positive sentiment in the textual data.

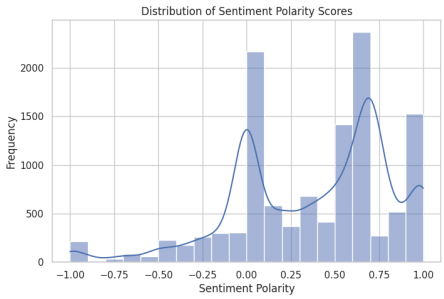


Figure 5. Sentiment Polarity Histogram
The histogram confirms that most reviews express positive sentiment, consistent with the high score distribution.

4.3 Clustering Analysis

The review text was vectorized using TF-IDF and then grouped using K-means clustering with *k = 5*. This segmentation revealed distinct thematic clusters that may relate to various aspects of service quality, such as:

- **Cluster 0:** Likely capturing general positive feedback.
- **Cluster 1:** Possibly reflecting comments on app performance.
- **Cluster 2:** Potentially associated with driver behavior.
- **Cluster 3 & 4:** Could represent niche topics like service reliability or specific operational feedback.

Figure 6. Bar Chart of Review Counts by Cluster