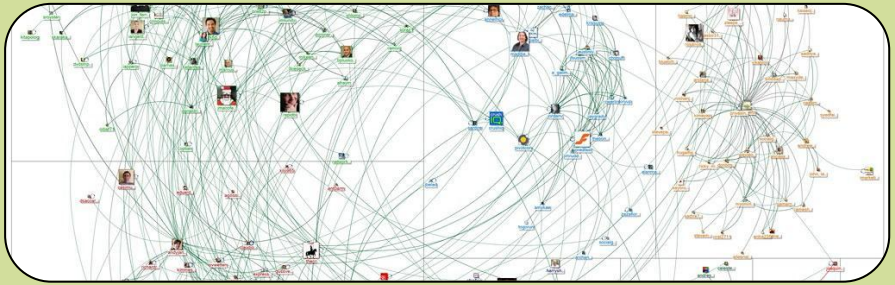
Data Science Project

Social Media Data Analysis



Oleksii Malovanyi | February 2025 | GitHub

INTRODUCTION

Social media is everywhere, and platforms like Twitter, Instagram, and Facebook generate huge amounts of data.

Challenge: Analyzing this data is complex but key to understanding user engagement, trends, and preferences.

This Project: uses simulated tweet data to explore user preferences based on likes.

TABLE OF CONTENTS

Project

Project Scope

Objective

Analyze simulated tweet data to understand user engagement (likes).

Key Steps

1. **Data Loading and Cleaning:** Handle missing values and outliers.
2. **Exploratory Data Analysis (EDA):** Visualize data distributions and relationships.
3. **Statistical Analysis:** Apply ANOVA and correlation analysis.
4. **Time Series Analysis:** Examine trends over time.
5. **Draw Conclusions:** Identify key insights and limitations.

Steps 1-3

Generation Loading and Inspection

Project

Data Generation

- Simulated a dataset of 1000 tweets.
- **Features:** tweet_id, category, likes, date, user_id.
- **Categories:** News, Sports, Entertainment, Tech, Food, Travel, Fashion.
- Intentionally included:
  - Skewed distribution of likes (exponential).
  - Outliers (a few tweets with very high likes).
  - Missing values (5% for 'category' and 'likes').

Data Loading and Inspection

- Data loaded from generated csv file using pandas.
- First 5 rows of data are printed for the user.

Data Cleaning

Step 4

Missing Values:

- **category**: Dropped rows with missing values (categorical feature).
- **likes**: Imputed missing values with the **median** (robust to outliers).

Date Conversion:

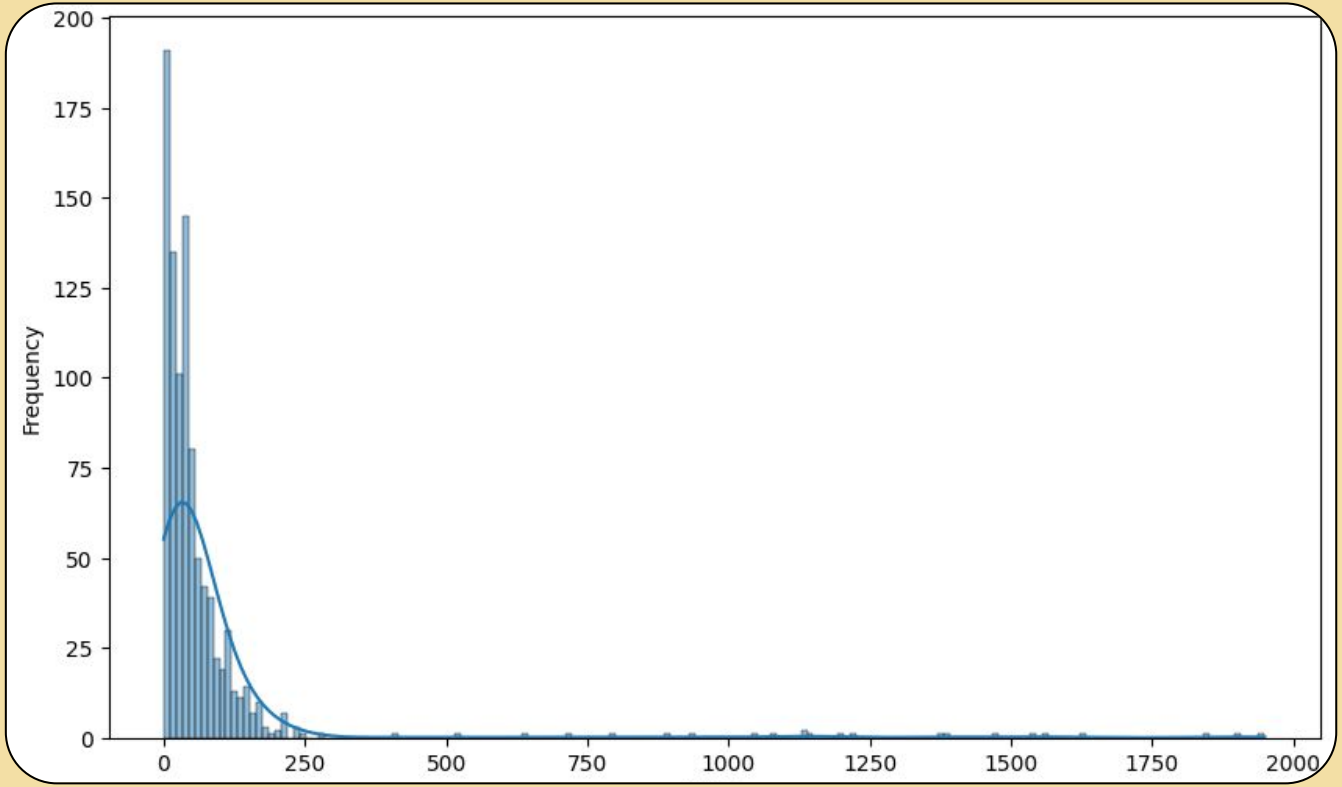
- Converted **date** column to datetime objects.

Step 5.1

Data Visualization

Distribution of Likes

Observation:
Highly skewed distribution. Most tweets have few likes, a few have many (outliers).



Step 5. 2

Data Visualization

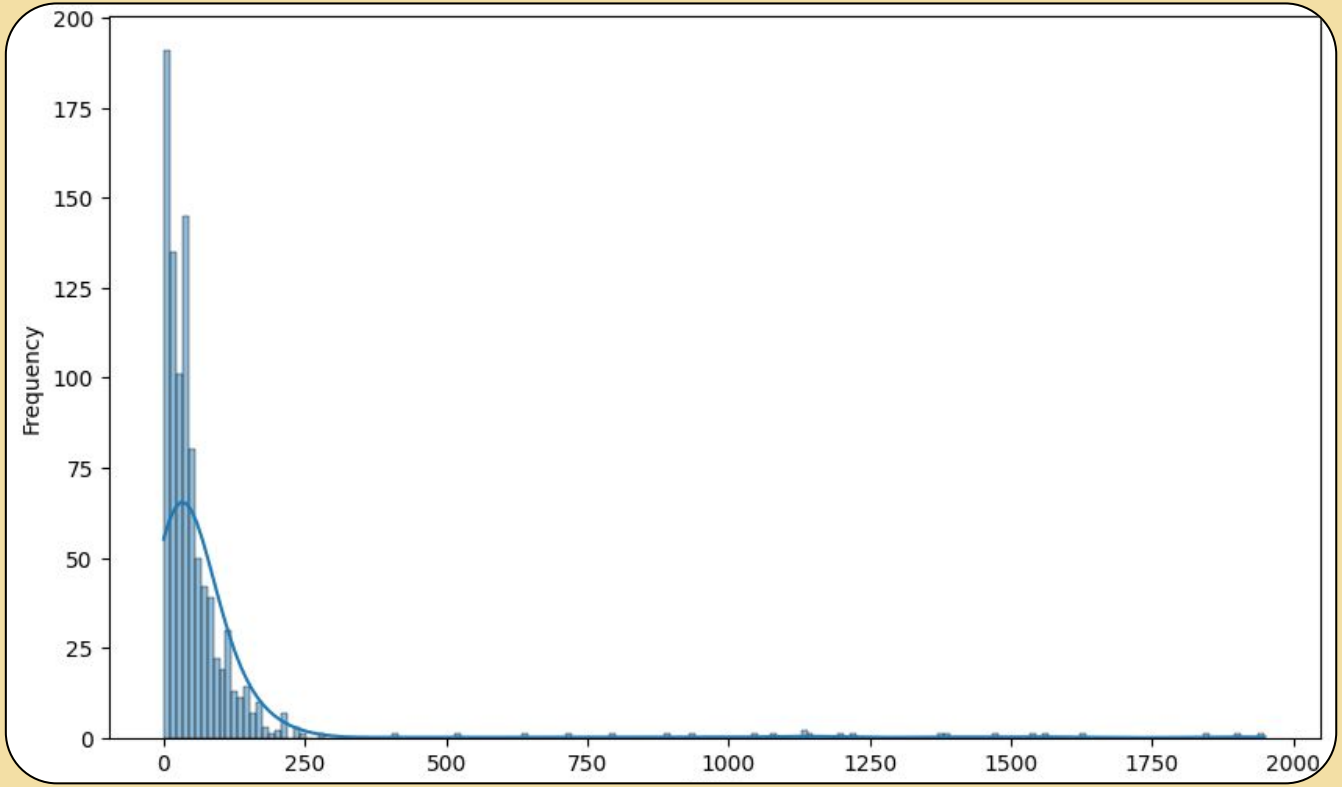
Likes by Category

Observation:

- Variation in distribution across categories.

- Some categories (e.g., Entertainment) have higher medians and more outliers.

- Other categories (e.g., Food) have tighter distributions.

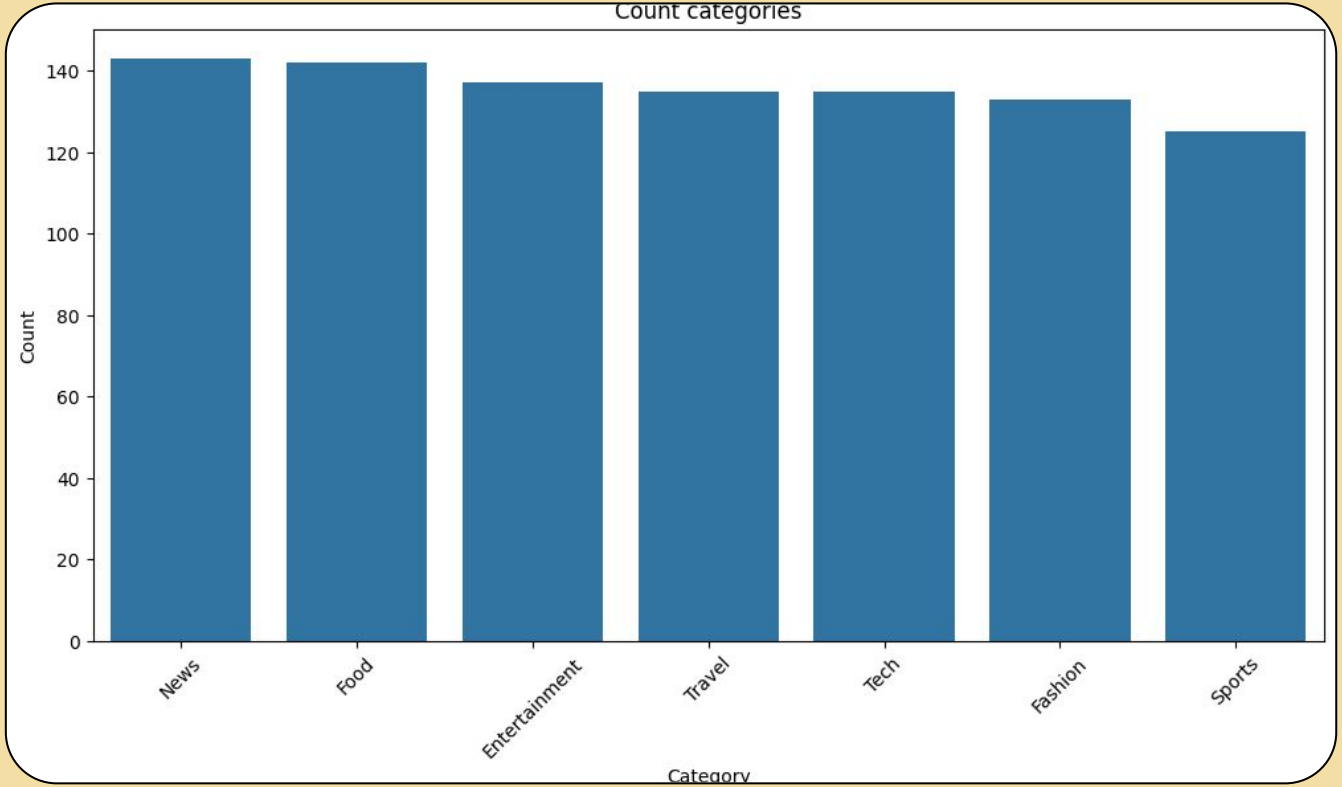


Step 5. 3

Data Visualization

Count categories

Observation: Categories have different count.

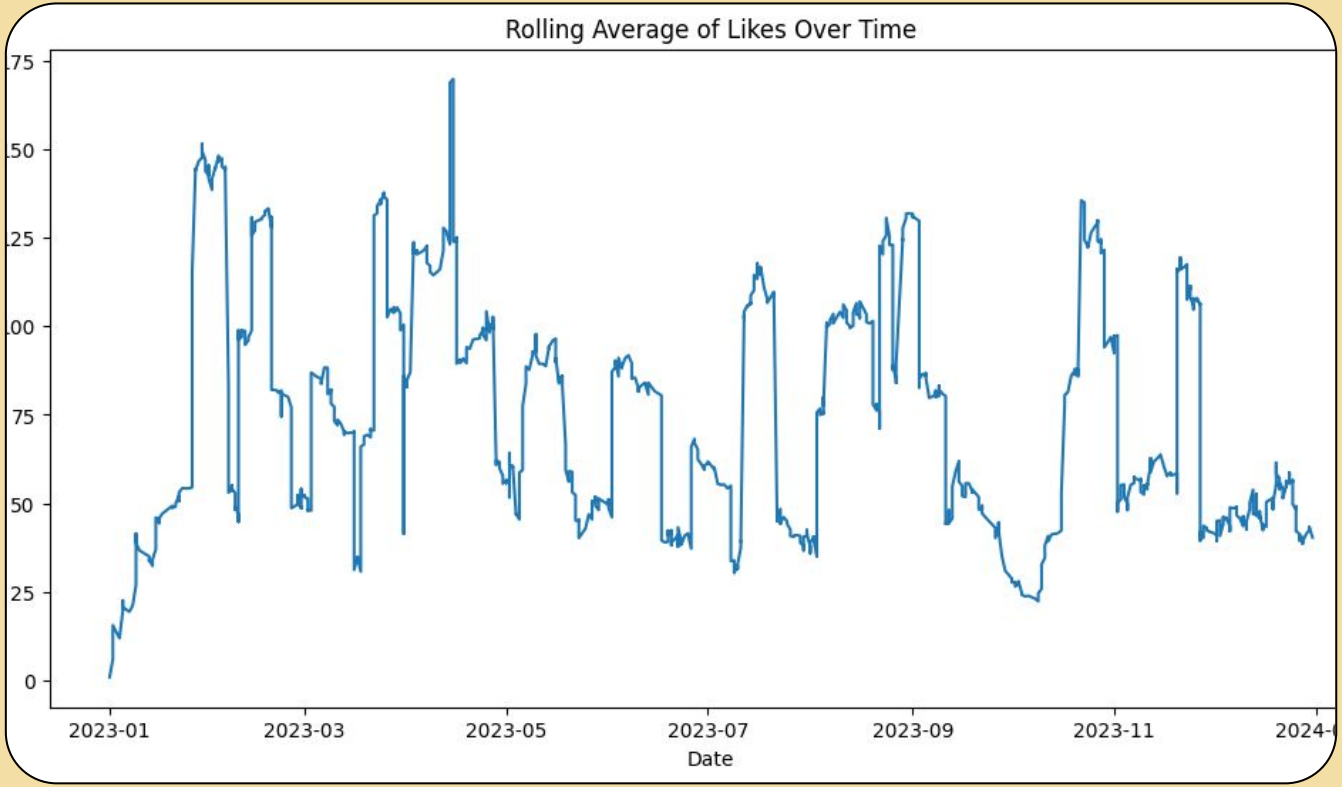


Step 5. 4

Data Visualization

Rolling Average of Likes Over Time

Observation: Fluctuations over time, showing periods of higher and lower average engagement. No strong overall trend.

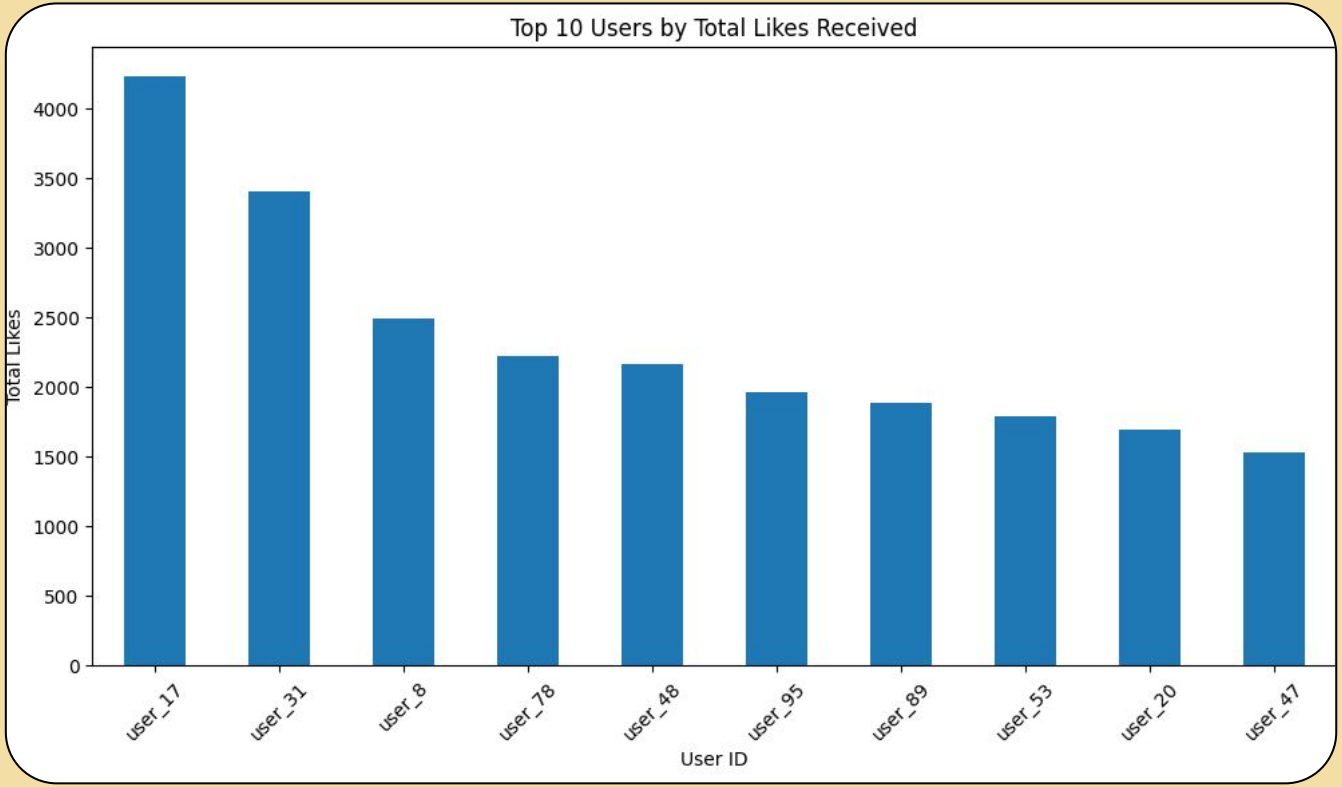


Step 5. 5

Data Visualization

Top 10 Users by Total Likes Received

Observation:
Top 10 user by Total Likes Received in descending order.



Step 6.1

Data Analysis

ANOVA Test

ANOVA Test

Purpose: Determine if there's a statistically significant difference in *mean likes* between categories.

Results:
- F-statistic: **0.90**
- P-value: **0.494**

Conclusion: There is NO statistically significant difference in mean likes between categories

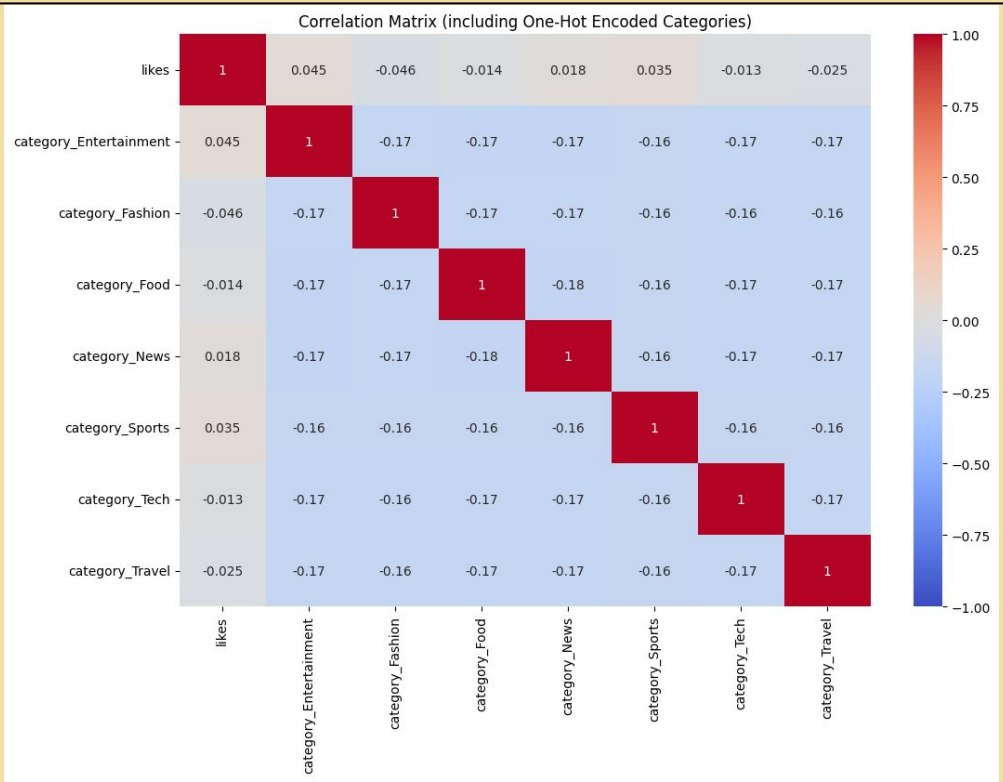
Step 6. 2

Data Analysis

Correlation Matrix

Purpose:
Examine linear relationships between 'likes' and one-hot encoded categories.

Observation:
Correlations are very close to zero, indicating a weak or no linear relationship. (Negative correlations between categories are expected due to one-hot encoding).

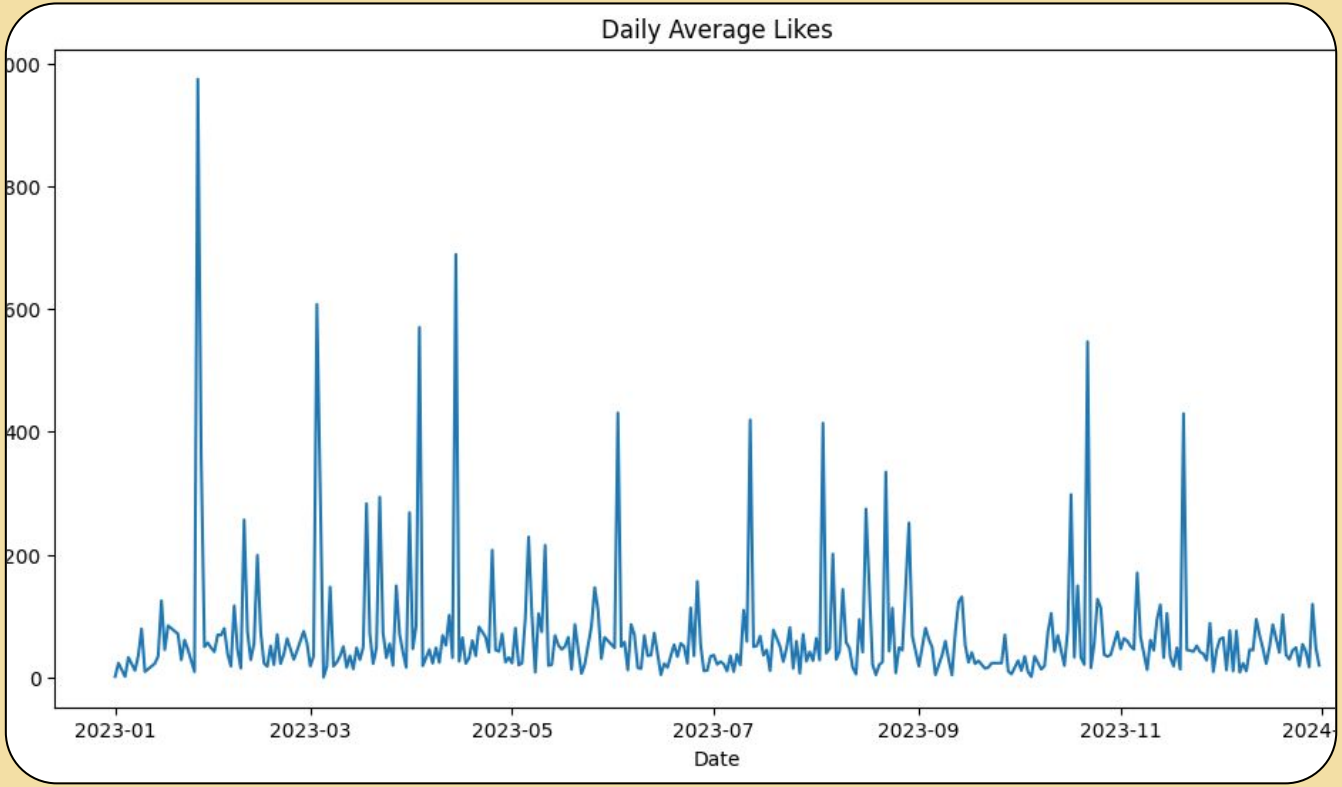


Step 6. 3

Data Analysis

Daily Average Likes

Observation: Considerable day-to-day variability. Suggests external factors (news, events) might influence engagement.



Insights

Summary of Findings

- **Likes Distributio**n: Highly skewed, most tweets have few likes, a few have many.

- **Likes by Category:** Variation in distribution, but ANOVA shows no statistically significant difference in *means*.

- **Correlation:** Very weak correlation between '*likes*' and '*category*'.

- **Time Series:** Fluctuations in rolling average and daily average likes.

Recommendations and Next Steps

- **Investigate Outliers:** Why do some tweets have very high likes?

- **Explore Temporal Patterns:**
  - Day of week/time of day effects?
  - Correlation with external events?
  - Deeper time series analysis.

- **Content Analysis:** Analyze tweet text (NLP) for keywords, topics, sentiment.

- **User Segmentation:** Identify different user groups.

- **Predictive Modeling:** Build a model to predict likes (requires more features).

Conclusion

1

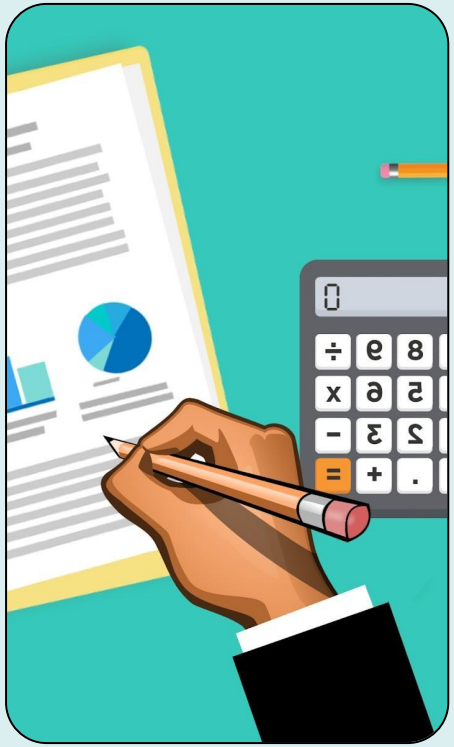
- This project analyzed simulated tweets to understand user engagement based on likes.
- Likes were unevenly distributed: most tweets had few likes, while a few had many

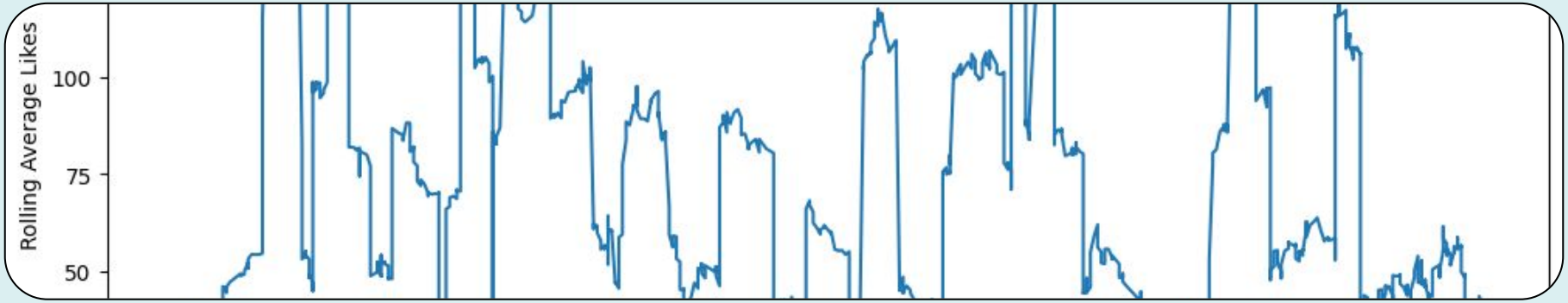
2

- No significant difference in mean likes between categories, and a weak correlation between likes and category.

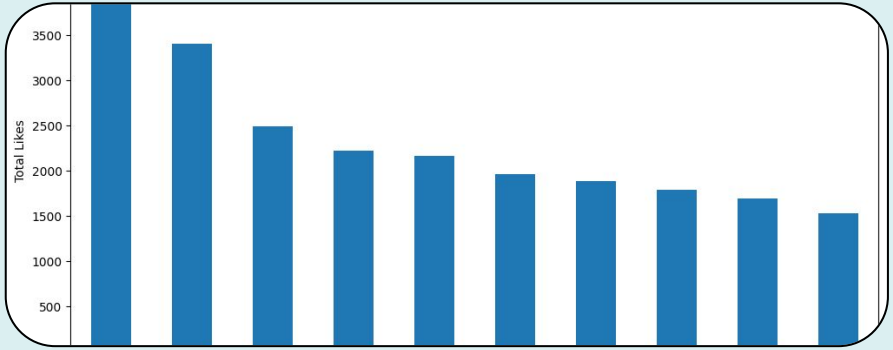
3

- Engagement changed over time, likely due to external factors.





THANK YOU



Oleksii Malovanyi
February 2025

Project on GitHub