USER ANALYSIS ON YELP DATASET

CSE 511- MILESTONE PROJECT

MANSI NANDKAR (1233870562)

INTRODUCTION:

Yelp is one of the most popular online review platforms where users share star ratings and detailed reviews, offering valuable insights into consumer opinions and preferences. It provides a centralized platform for users to express their experiences with businesses of all types and sizes. This analysis focuses on understanding user behavior, patterns, and sentiments by leveraging tools. The primary objective is to uncover key trends in user interactions, identify common behavioral patterns, and analyze sentiment within reviews. These insights aim to provide a deeper understanding of user preferences and expectations, enabling businesses to make data-driven decisions to enhance customer engagement, satisfaction, and loyalty.

IMPLEMENTATION:

1. Data Loading and Filtering: The Yelp dataset was loaded in JSON files using PySpark and then I filtered the business dataset to include only businesses in Arizona (AZ). Further, I filtered to focus specifically on salons by searching for "Salon" in the categories which had 554 entries. I created temporary views of the filtered datasets to enable SQL queries for data transformation.

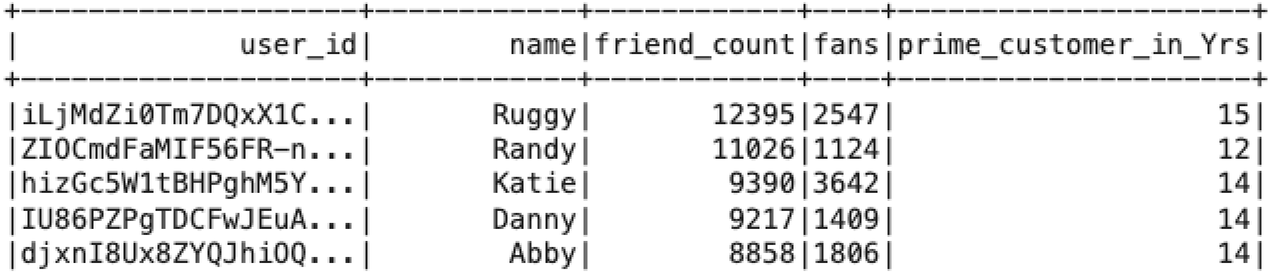
2. User Analysis: The key tasks involved analyzing the distribution of user ratings, identifying patterns in user preferences, and investigating attributes of highly active or influential users. This included understanding user engagement metrics, such as the number of reviews and average ratings, and visualizing trends to support the analysis. The tasks aimed to uncover the behavior and preferences that contribute to user satisfaction and loyalty.

3. Data Visualization: Various plotting techniques were employed to present the results of the analysis visually. Bar charts were used to depict rating distributions and geographical comparisons, making complex trends easier to interpret and communicate effectively.

4. Insight Generation: Insights derived from SQL queries and visualizations revealed critical trends in user behavior and sentiment on the Yelp platform. Factors such as positive sentiment in reviews, frequent engagement, and detailed feedback emerged as key contributors to user influence and satisfaction. These findings provide actionable recommendations for enhancing user engagement, improving platform features, and tailoring experiences to better meet user expectations.

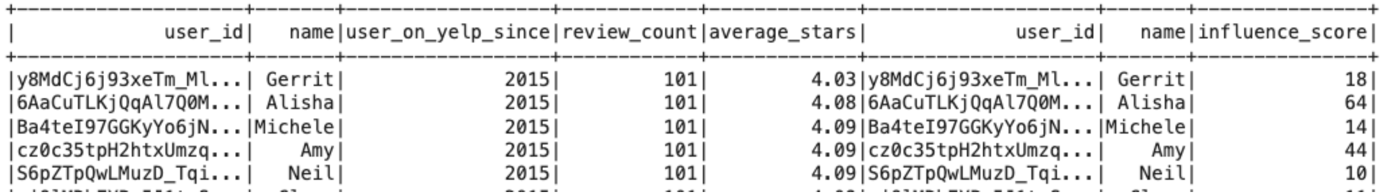
FINDINGS & ANALYSIS:

1. Ruggy has been a prime customer for 15 years with a friend count of 12395 and 2547 fans followed by Randy of 11026 friends and 1124 fans who has been prime customers for 12 years.



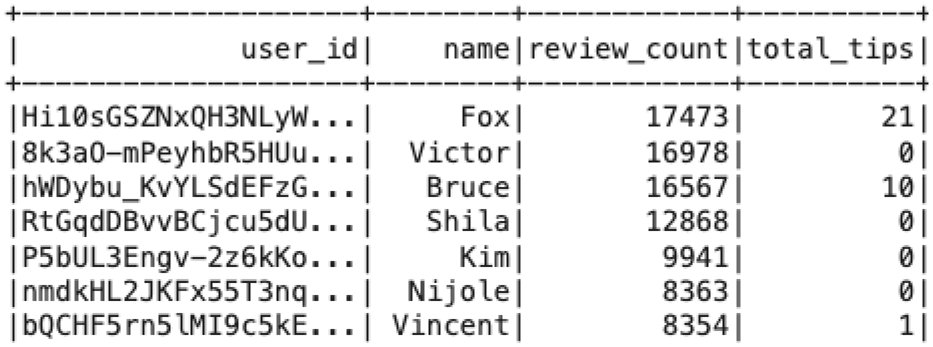
```
+--------------------+------------+------------+----+---------------------+
|             user_id|        name|friend_count|fans|prime_customer_in_Yrs|
+--------------------+------------+------------+----+---------------------+
|iLjMdZi0Tm7DQxX1C...|       Ruggy|       12395|2547|                   15|
|ZIOCmdFaMIF56FR-n...|       Randy|       11026|1124|                   12|
|hizGc5W1tBHPghM5Y...|       Katie|        9390|3642|                   14|
|IU86PZPgTDCFwJEuA...|       Danny|        9217|1409|                   14|
|djxnI8Ux8ZYQJhiOQ...|        Abby|        8858|1806|                   14|
```

2. The majority of the users who joined yelp from 2015 has a review count of 101 and did a user activity analysis by influence count which depends on the user getting cool, funny, writer and fans under compliment criteria.



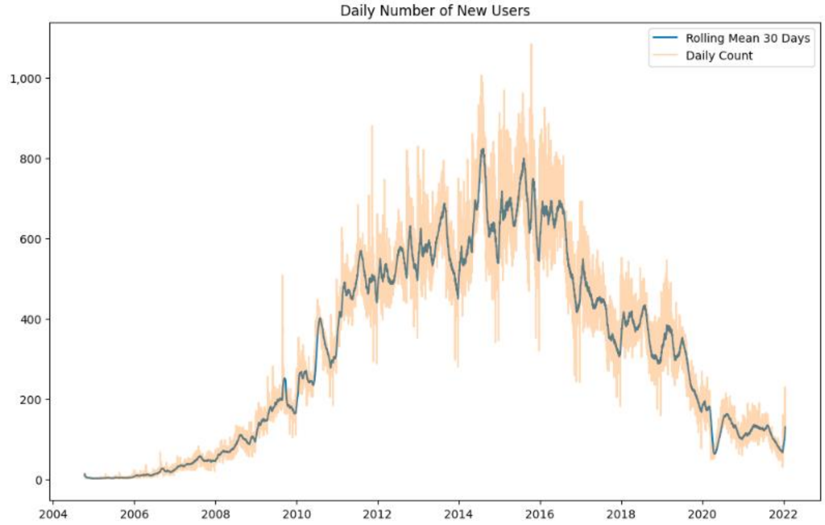
```
+--------------------+-------+------------------+------------+-------------+--------------------+-------+---------------+
|             user_id|   name|user_on_yelp_since|review_count|average_stars|             user_id|   name|influence_score|
+--------------------+-------+------------------+------------+-------------+--------------------+-------+---------------+
|y8MdCj6j93xeTm_Ml...| Gerrit|              2015|         101|         4.03|y8MdCj6j93xeTm_Ml...| Gerrit|             18|
|6AaCuTLKjQqAl7Q0M...| Alisha|              2015|         101|         4.08|6AaCuTLKjQqAl7Q0M...| Alisha|             64|
|Ba4teI97GGKyYo6jN...|Michele|              2015|         101|         4.09|Ba4teI97GGKyYo6jN...|Michele|             14|
|cz0c35tpH2htxUmzq...|    Amy|              2015|         101|         4.09|cz0c35tpH2htxUmzq...|    Amy|             44|
|S6pZTpQwLMuzD_Tqi...|   Neil|              2015|         101|         4.09|S6pZTpQwLMuzD_Tqi...|   Neil|             10|
```

3. Fox being the topmost elite user on Yelp who gave a total review of 17473 and a total tip of 21 followed by Victor for giving a total review of 16978. This helps us to analyze the behavior of elite users on user platform and their preferred activity.

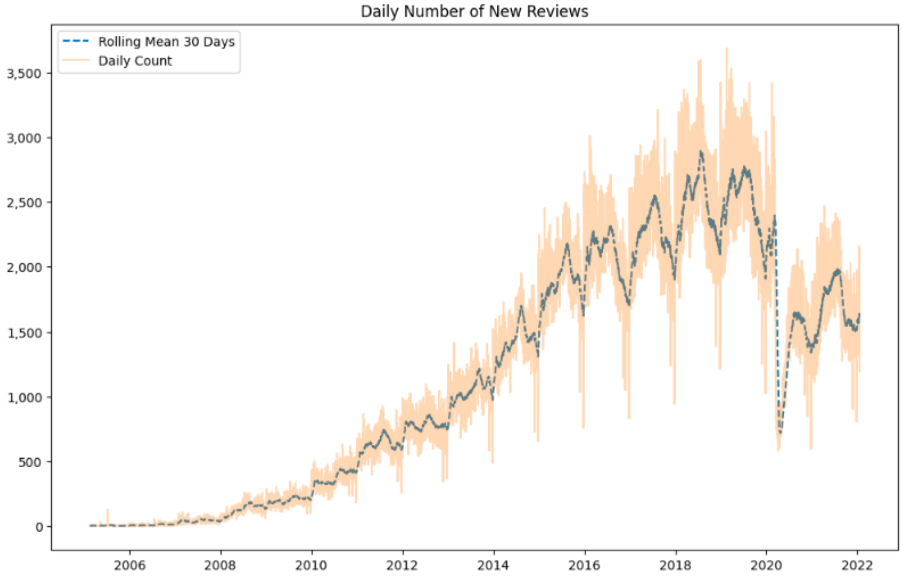


```
+--------------------+--------+------------+----------+
|             user_id|    name|review_count|total_tips|
+--------------------+--------+------------+----------+
|Hi10sGSZNxQH3NLyW...|     Fox|       17473|        21|
|8k3aO-mPeyhbR5HUu...|  Victor|       16978|         0|
|hWDybu_KvYLSdEFzG...|   Bruce|       16567|        10|
|RtGqdDBvvBCjcu5dU...|   Shila|       12868|         0|
|P5bUL3Engv-2z6kKo...|     Kim|        9941|         0|
|nmdkHL2JKFx55T3nq...|  Nijole|        8363|         0|
|bQCHF5rn5lMI9c5kE...| Vincent|        8354|         1|
```

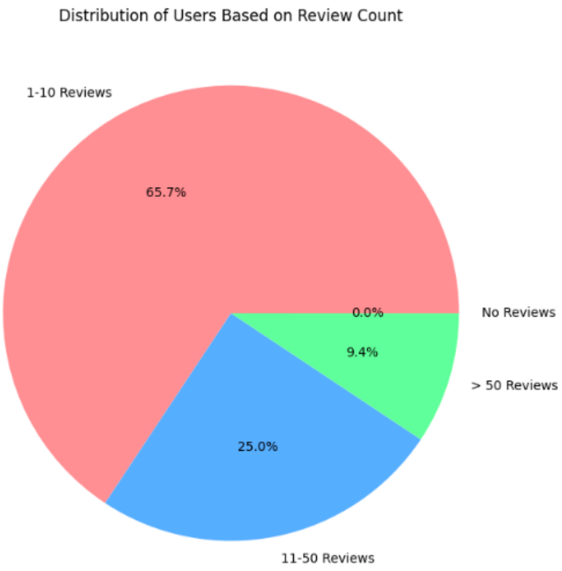
4. The daily number of new users experienced steady growth until around 2014, followed by a sharp decline, and has remained relatively stable at lower levels since then. The rolling 30-day mean captures long-term trends while smoothing short-term variability, suggesting potential seasonality or external factors influencing user acquisition over time.



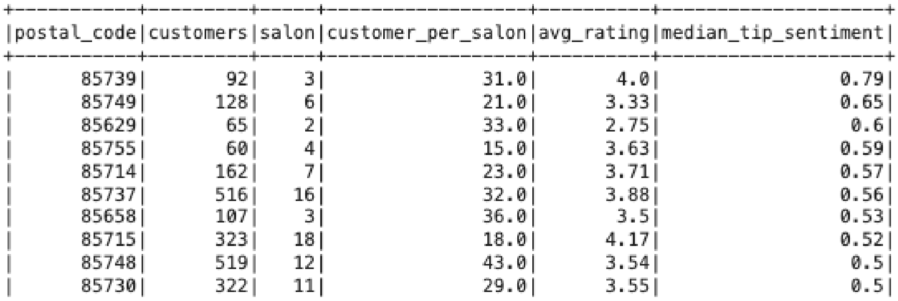
5. The daily number of new reviews shows steady growth from 2006, peaking between 2018 and 2020, before experiencing a notable decline and subsequent stabilization. The rolling 30-day mean highlights cyclical patterns, likely indicating seasonal trends or external factors affecting user engagement. The sharp drop around 2020 could correspond to global disruptions.



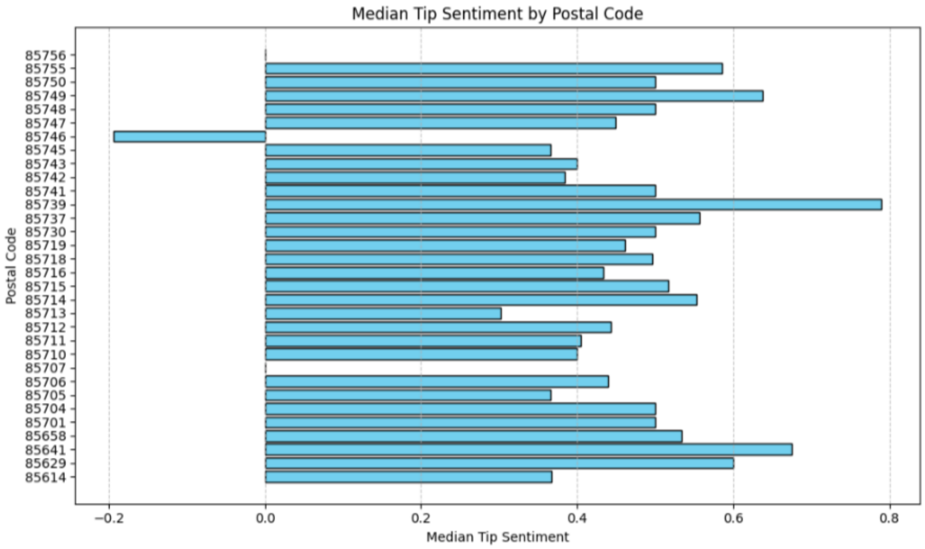
6. The analysis shows that 65.7% of users have 1–10 reviews, 25.0% have 11–50 reviews, and 9.4% have over 50 reviews, with no users having zero reviews. This indicates most users are occasional contributors presence on Yelp, while a small group of highly active users drives a significant portion of reviews.



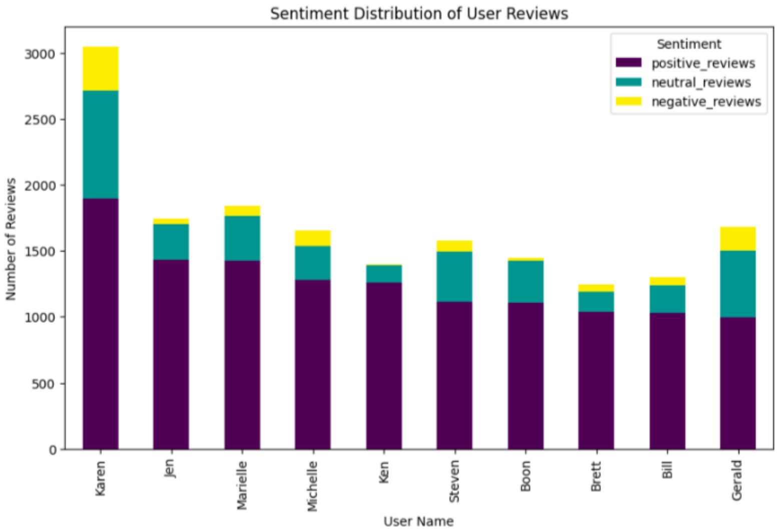
7. After analysing user behavior as per postal code area and found average rating and median of sentiment analysis of all the salons in that area showed 85739 with the highest median of sentiment of tips and 85746 with the lowest.



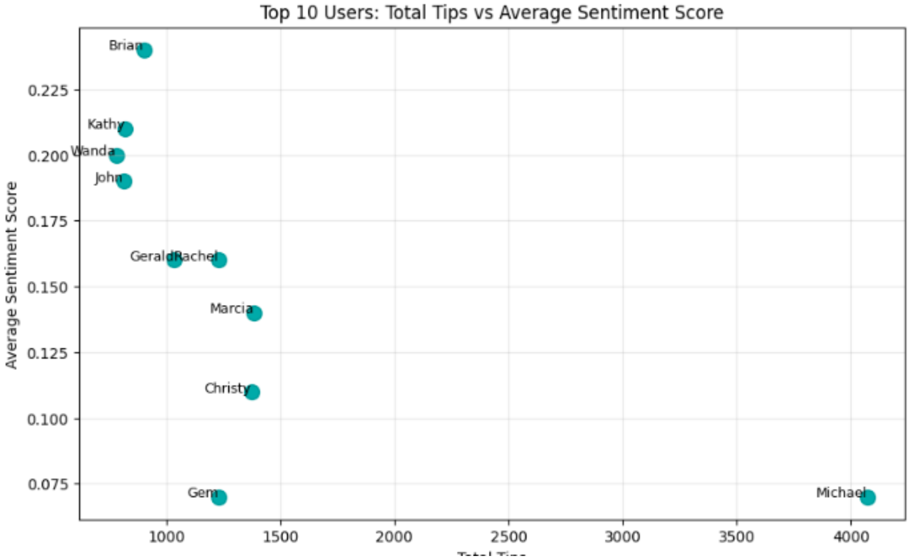
```
+-----------+---------+-----+------------------+----------+--------------------+
|postal_code|customers|salon|customer_per_salon|avg_rating|median_tip_sentiment|
+-----------+---------+-----+------------------+----------+--------------------+
|      85739|       92|    3|              31.0|       4.0|                0.79|
|      85749|      128|    6|              21.0|      3.33|                0.65|
|      85629|       65|    2|              33.0|      2.75|                 0.6|
|      85755|       60|    4|              15.0|      3.63|                0.59|
|      85714|      162|    7|              23.0|      3.71|                0.57|
|      85737|      516|   16|              32.0|      3.88|                0.56|
|      85658|      107|    3|              36.0|       3.5|                0.53|
|      85715|      323|   18|              18.0|      4.17|                0.52|
|      85748|      519|   12|              43.0|      3.54|                 0.5|
|      85730|      322|   11|              29.0|      3.55|                 0.5|
```



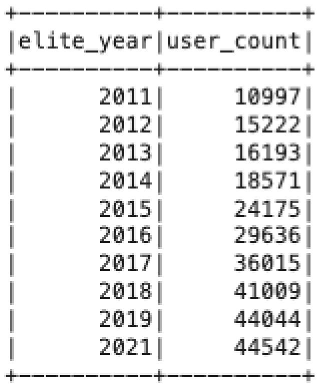
8. The bar chart illustrates the sentiment distribution of user reviews across different individuals. Karen leads with the highest total reviews, predominantly positive, followed by Jen and Marielle with significant positive and neutral contributions. Other users like Michelle, Ken, and Steven have balanced distributions but fewer total reviews. Negative reviews are consistently the smallest segment for all users, with Gerald and Karen contributing slightly more in this category.



9. The scatter plot shows the relationship between total tips and average sentiment scores for the top 10 users. Michael has the highest total tips but the lowest average sentiment score, indicating high activity but relatively less positive sentiment. Conversely, Brian, Kathy, and Wanda have the highest average sentiment scores with fewer total tips, suggesting they provide fewer but more positively received contributions.



10. The number of elite users steadily increased from 2011 to 2021, indicating consistent growth in user engagement and contributions over the years.


```
+----------+----------+
|elite_year|user_count|
+----------+----------+
|      2011|     10997|
|      2012|     15222|
|      2013|     16193|
|      2014|     18571|
|      2015|     24175|
|      2016|     29636|
|      2017|     36015|
|      2018|     41009|
|      2019|     44044|
|      2021|     44542|
+----------+----------+
```

CONCLUSION & RECCOMENDATIONS:

The analysis highlights active user engagement on the platform through reviews, tips, and check-ins, underscoring a dynamic and interactive community. Reviews capture a mix of positive and critical feedback, offering valuable insights into diverse user experiences. Popular categories like restaurants and cafes see higher engagement, while urban areas such as Tucson attract a significant concentration of activity. Sentiment analysis reveals both satisfaction and areas for improvement, emphasizing opportunities for businesses to better align with user expectations.

To leverage these insights, I recommend focusing on enhancing customer service and systematically addressing recurring complaints highlighted in reviews. Encouraging satisfied customers to leave positive reviews can boost credibility and attract new clientele. Businesses should utilize targeted marketing strategies tailored to popular categories and consider expanding operations in high-engagement urban areas. Regular analysis of user data and trends will enable informed decisions for continuous improvement and growth.