CSE 511 Milestone 2
Name – Shashank Navad
ASUID - 1229407428

Yelp Data Analysis: Nightlife in Arizona

Problem statement- This report analyzes user behavior and contributions related to nightlife businesses in Arizona using the Yelp dataset. The analysis focuses on user review patterns, elite status impact, and account longevity to provide insights into user engagement with nightlife establishments.

Methods- The analysis was conducted using Apache Spark for big data processing. Key steps included: filtering the dataset to focus on nightlife businesses in Arizona, analyzing user review counts and distribution, examining the impact of elite status on review behavior, and investigating user account age distribution.

Queries-

1. Distribution of users by review count
This query categorizes users based on their review count into four groups: 1-4 reviews, 5-19 reviews, 20-49 reviews, and 50+ reviews. It provides insights into user engagement levels.

2. Top 10 users with the most fans
This query identifies and ranks the most popular users based on their fan count, highlighting influential members of the Yelp community.

3. Average stars given by elite vs non-elite users
This analysis compares the rating behavior of elite and non-elite users, revealing potential differences in their reviewing patterns.

4. Distribution of users by account age
Users are grouped by account age into categories: <1 year, 1-2 years, 3-4 years, and 5+ years. This helps understand the user base's longevity and loyalty.

5. Sentiment analysis of reviews
This query analyzes sentiment and user behavior based on review count and elite status. It categorizes users into groups (Novice, Regular, Veteran, Elite) based on their review count, and distinguishes between Elite and Non-Elite users. For each category, it calculates average sentiment score, average star rating, number of users, and average review count.

6. User category preferences and sentiment
This query analyzes user preferences for different business categories based on their review activity and sentiment. It breaks down the business categories, groups the data by user and category, then calculates the number of reviews and average sentiment for each user-category combination.

7. User review quality and sentiment over time
This query analyzes user review quality and sentiment over time. It groups reviews by user and year, calculating average quality scores (based on useful, funny, and cool ratings), average sentiment scores, and review counts for each user annually.

8. User sentiment consistency across reviews per year
This query analyzes user sentiment consistency across reviews over multiple years. It calculates the average sentiment per user per year, then computes the overall average sentiment and standard deviation across years for each user. The results are filtered to include only users active for more than one year and ordered by the standard deviation of sentiment, showing users with the most variable sentiment first. This helps identify users whose review sentiments fluctuate significantly over time.

9. User engagement level and business rating correlation
This query analyzes the relationship between user engagement levels and business ratings. It groups users by their review count, categorizes them into five engagement levels, and then calculates average user ratings, average business ratings, and the correlation between these ratings for each engagement level.

10. User review length and sentiment analysis by elite status

This query analyzes review length and sentiment for elite and non-elite Yelp users. It calculates average review length and sentiment scores for each group, comparing these metrics to identify potential differences in review behavior based on user status.

Key findings-

1. User Engagement Distribution: There's a wide range of user engagement levels. The largest group (13,918 users) falls in the 5-19 reviews category, indicating a significant base of moderately active users. However, there's also a substantial number of highly active users, with 6,662 users having written 50+ reviews.

2. Elite vs Non-Elite User Behavior: Elite users, while comprising only 9.7% of the user base, tend to give higher ratings (average 3.99 stars) compared to non-elite users (3.68 stars). This could suggest different standards or experiences between these user groups.

3. User Longevity: Most users (35,909) have account ages of 5+ years, indicating a very stable and loyal user base. This longevity could impact the reliability and depth of reviews.

4. Influential Users: There's a small group of highly influential users. The top user, Katie, has 3,642 fans, significantly more than the next most popular user. This suggests the presence of "super users" who may have outsized influence on the platform.

5. There's a slight variation in sentiment scores for individual users across different categories suggesting that users may have preferences or varying experiences in different types of establishments.

6. Engagement Level and Rating Correlation: There's a strong positive correlation (0.256) between engagement level and star rating, suggesting that as the number of reviews increases, they also tend to be more positive.

7. Sentiment scores vary widely from highly negative -0.9972 to extremely positive 0.9998 showing diverse opinions among Elite reviewers.

Visualizations-
The analysis includes several key visualizations that offer insights into user behavior and engagement patterns on Yelp:

Distribution of Users by Review Count: A bar chart illustrating the number of users across different review count ranges (1-4, 5-19, 20-49, 50+).

Top 10 Users with Most Fans: A bar chart showcasing the top influencers in the Yelp community, displaying their fan counts.

Average Stars Given by Elite vs Non-Elite Users: This visualization compares the rating behaviors of elite and non-elite users.

User Account Age Distribution: A bar chart showing the distribution of user accounts by age (1-2 years, 3-4 years, 5+ years).

User Sentiment Analysis by Review Count and Elite Status: This visualization provides insights into how user sentiment varies with their review activity and elite status, highlighting trends in engagement and satisfaction.

Average Review Quality Score Over Time: A line graph analyzing how users' review quality changes over their account lifetime. Quality is determined by other users who found the review useful, funny or cool.

Correlation between engagement level and Rating: A scatter plot examining the relationship between engagement levels and the rating given.

Review Length Distribution by Elite Status: A boxplot depicting the relationship between average review length and elite status.

Conclusion-
This analysis provides a comprehensive view of user behavior on Yelp, focusing on review patterns, user engagement, and the impact of user characteristics on ratings. The findings can be valuable for understanding user dynamics and potentially improving business strategies.