CSE 511 - DPS

Milestone 2:User Level Analysis

Abhi Dineshkumar Patel

Student ID: 1230038877

1. Analysis using spark.sql queries

1.1 Average Stars Given by Users

Inputs:

Dataset: Yelp review table.

Columns used: user_id, stars.

Query Logic:

Group reviews by user_id.
Compute the average stars given by each user.
Sort the results in descending order of average stars.
Limit to top 10 users.

Visualization:

A bar chart showing User ID on the x-axis and Average Stars on the y-axis.

Output:

Insights: The data indicates a significant clustering of user reviews achieving an average of 5 stars across all user IDs. This trend may suggest high levels of satisfaction, consistent service quality, or bias in the review collection process.
Decision-Making Value: The uniformity of 5-star reviews highlights an opportunity to further investigate review authenticity or customer feedback trends. If genuine, it could signal a strong value proposition or excellent customer experience, making this a favorable environment for retaining existing users or attracting new ones.

1.2 Top 10 Active Users by Review Count

Inputs:

Dataset: Yelp user table.

Columns used: name, review_count.

Query Logic:

Sort users by review_count in descending order.
Limit the results to the top 10 users.

Visualization:

A bar chart (log-scale y-axis) displaying User Name on the x-axis and Review Count on the y-axis.

Output:

Insights: Among the top 10 active users ranked by review count, the user "FOX" significantly outperforms others with a review count falling in the 10^4 to 10^5 range. All other users remain far below this threshold, highlighting FOX's unique activity level.
Decision-Making Value: Gym owners could focus on offering niche services to attract higher ratings and customer loyalty.

1.3 User Contributions Across Cities for Gyms

Inputs:

Dataset: Yelp gyms and reviews table.

Columns used: city, user_id.

Query Logic:

Join reviews with gym businesses.
Group by city and count distinct user_id.
Sort results by user count in descending order.

Visualization:

A bar chart with City on the x-axis and Number of Unique Users on the y-axis.

Output:

Insights: Tucson stands out with over 2,000 unique users contributing to gym reviews, indicating a highly active and engaged fitness community. In contrast, other cities like Oro Valley, Marana, and Green Valley each have fewer than 250 unique users, showing significantly lower participation.

Decision-Making Value: Tucson's high user engagement makes it a prime location for launching targeted marketing campaigns or community fitness events. The lower activity in Oro Valley, Marana, and Green Valley may indicate untapped potential for growth in user contributions or demand for fitness-related businesses.

1.4 Average Review Length Per User

Inputs:

Dataset: Yelp user and review table.

Columns used: user_id, name, text.

Query Logic:

Join user and review data.
Calculate average length of review text per user.
Sort in descending order by review length.
Limit to top 10 users.

Visualization:

A bar chart displaying User Name on the x-axis and Average Review Length on the y-axis.

Output:

Insights: Users such as Jerri, Penni, and others are consistently writing reviews with an average length of 5,000 characters. This uniformity across top users suggests a trend where detailed, elaborate reviews dominate the dataset, potentially driven by a specific review-writing behavior or platform incentives.

Decision-Making Value:Identifying these highly detailed reviewers offers an opportunity to collaborate with them for promotional content or community-building efforts. Additionally, their reviews could be used to gauge sentiment and provide deeper insights into customer experiences.

1.5 Yearly Trends of Gym Reviews

Inputs:

Dataset: Yelp review table.

Columns used: date.

Query Logic:

Extract review year from date.
Group reviews by year and count entries.
Sort by year in ascending order.

Visualization:

A line chart showing Year on the x-axis and Review Count on the y-axis.

Output:

Insights: The data reveals a steady growth in the number of reviews from 2005, reaching a peak of approximately 90,000 in 2019. However, there is a sharp decline in review activity post-2019, leveling off by 2022. This trend could reflect external factors such as changes in user behavior, platform policies, or broader economic and societal influences (e.g., the COVID-19 pandemic).
Decision-Making Value: The growth phase up to 2019 indicates increasing engagement and platform adoption, which can inform strategies for content optimization and user retention. The post-2019 decline suggests the need for interventions like re-engagement campaigns, exploring the causes of reduced activity, or adapting to external shifts (e.g., pivoting to new formats like video reviews or incentivizing reviews).

1.6 User Contribution by Gym ZIP Code

Inputs:

Dataset: Yelp gym and review tables.

Columns used: postal_code, user_id.

Query Logic:

Join gym and review tables.
Group by postal_code and count unique users.
Sort in descending order by user count.

Visualization:

A bar chart showing ZIP Code on the x-axis and User Count on the y-axis.

Output:

Insights: The analysis highlights that ZIP code 85719 has the highest user count, with over 500 unique users, showcasing it as a hub of engagement. The second-highest ZIP code, 85712, has around 300 users, while 85742 has the lowest user count at approximately 10. This variation in user distribution indicates significant disparities in user activity across ZIP codes.
Decision-Making Value: ZIP code 85719 presents a prime opportunity for targeted marketing efforts and service expansions due to its high user concentration. Conversely, ZIP code 85742 may require strategies to increase user engagement, such as localized promotions or community outreach to build a stronger presence in the area.

1.7 Reviews by User Experience Group

Inputs:

Dataset: Yelp user and review tables.

Columns used: yelping_since, user_id.

Query Logic:

Categorize users based on years since they joined Yelp.
Count reviews by each experience group.
Sort in descending order by review count.

Visualization:

A bar chart showing User Experience Group on the x-axis and Review Count on the y-axis.

Output:

Insights:The bar graph indicates that Experienced Users (with more than 6 reviews) dominate, showcasing their significant contribution to overall activity. In contrast, users in the 4-5 years and 2-3 years of experience groups have minimal engagement, with review counts between 0-1.
Decision-Making Value:The dominance of Experienced Users suggests they are the primary drivers of community engagement. Efforts could focus on sustaining their activity through loyalty rewards or exclusive features. On the other hand, the low activity levels in mid-experience groups indicate a potential drop in engagement over time. Retention strategies, such as re-engagement campaigns or gamified incentives, might boost participation in these segments.

1.8 Reviews by Membership Status

Inputs:

Dataset: Yelp user and review tables.

Columns used: elite, user_id.

Query Logic:

Categorize users into Elite, Non-Elite, and Regular groups.
Count reviews for each category.
Sort in descending order by review count.

Visualization:

A bar chart showing User Membership Status on the x-axis and Review Count on the y-axis.

Output:

Insights:The bar graph clearly shows that No-Elite users (5,264,589 users) contribute significantly with more than 5 reviews, suggesting high activity and ongoing engagement. On the other hand, Regular users (1,725,658 users) are engaged with 1-2 reviews on average, indicating more occasional participation.
Decision-Making Value:The higher review count for No-Elite users emphasizes their central role in the platform's activity, which may represent a broader user base that could be leveraged for community-building and content generation. The relatively lower review activity from Regular users suggests that while they represent a sizable portion of the user base, their engagement is still limited. Targeting Regular users with incentives or engaging features could help increase their participation and elevate the platform's content flow.

1.9 Reviews by Rating and Review Length

Inputs:

Dataset: Yelp review table.

Columns used: text, stars.

Query Logic:

Categorize reviews into Short, Medium, and Long.
Group by review length and star rating.
Count reviews in each group.

Visualization:

A bar chart showing Review Length and Rating on the x-axis and Review Count on the y-axis

Output:

Insights: The bar graph reveals that long reviews (high review length) with a 5-star rating have the highest review count of approximately 2.5 million, and they are associated with the largest number of users (2,385,634). As the review length decreases or the star rating drops, there is a noticeable decline in both the number of reviews and users. This suggests that longer, highly-rated reviews are more frequent and that a drop in either review length or rating leads to fewer reviews.
Decision-Making Value: This data shows a strong correlation between longer reviews and higher ratings, which could imply that users tend to provide more detailed feedback when they are highly satisfied. This insight can guide platforms in fostering more meaningful engagement by encouraging detailed, positive reviews. Additionally, understanding the drop-off in review length and star ratings could help in identifying areas for improvement in user satisfaction or content generation.

1.10 Reviews by User Elite Status and Review Length

Inputs:

Dataset: Yelp user and review tables.

Columns used: text, elite.

Query Logic:

Categorize reviews by length and user elite status.
Count reviews for each combination.

Visualization:

A bar chart with Review Length and User Status on the x-axis and Review Count on the y-axis.

Output:

Insights: The bar graph indicates that for Elite users, there is a clear trend where the review count decreases as review length decreases. Elite users tend to write longer reviews, and as the length of their reviews shortens, the number of reviews provided also drops. This suggests that Elite users, who are known for their higher engagement and expertise, tend to provide more detailed feedback, which may correlate with their higher status on the platform.
Decision-Making Value: This data highlights the quality of contributions from Elite users, with longer reviews being more frequent. The platform could leverage this insight to encourage in-depth reviews, particularly from users with higher engagement levels. It also reinforces the idea that Elite users provide more comprehensive and valuable insights, which can be a key factor for platforms looking to improve user engagement or content quality.