Milestone 1 Report: Business Level Analysis

Introduction

The objective of Milestone 2 is to analyze user behavior and contributions within the Yelp community, focusing on the chosen business category in Arizona (AZ). This analysis delves into user activity patterns, sentiment distribution of reviews, and variations in sentiment based on user characteristics. The aim is to uncover insights into user engagement, influence, and preferences within the specified business category. The analysis leverages Spark SQL for efficient querying, with results presented in a concise report highlighting the key findings and their implications. Category of business used - Automotive

Methods and Analysis

1. Distribution of Users Based on Review Count
Objective: Categorize users into groups based on their total review counts to understand participation levels.
Method: Grouped users into three categories: 1–10 reviews, 11–50 reviews, and >50 reviews. Calculated the proportion of users in each group and visualized using a pie chart.
Result: Nearly half of the users contributed between 1–10 reviews, while 31.7% wrote 11–50 reviews. A smaller group (21.6%) was highly active, contributing more than 50 reviews.
<Cell Number 57 in jupyter notebook>

2. Most Reviewed Businesses per User

Objective: Identify businesses in the automotive category that receive the most reviews from users.
Method: Filtered automotive businesses in Arizona, joined user and review data, and grouped by user and business to calculate review counts. Visualized using a bar graph.
Result: Discount Tire had the highest user engagement, followed by QuickTrip and Mister Car Wash.
<Cell Number 58 in jupyter notebook>

3. Sentiment Distribution of User Reviews
Objective: Analyze the distribution of positive, neutral, and negative reviews by users in the automotive.
Method: Classified reviews based on star ratings (positive: ≥4, neutral: 3, negative: ≤2). Grouped by user and visualized the sentiment distribution using a stacked bar chart.
Result: John had the highest number of reviews, with a significant proportion being positive. Other users, including Michael and David, also contributed substantially, with noticeable variations in sentiment distribution.
<Cell Number 59 in jupyter notebook>

4.Top 20 Users by Funny Votes
Objective: Identify the most engaging users based on the number of funny votes received for reviews.
Method: Grouped user reviews in the automotive category by user name and summed the funny votes. Sorted the results in descending order and visualized the top 20 users using a bar graph.
Result: John received the highest number of funny votes, significantly outperforming other users. Marshall and Jennifer also had notable contributions to humor-driven engagement.
<Cell Number 60 in jupyter notebook>

5. Top 20 Users by Useful Votes
Objective: Identify the users whose reviews were deemed the most useful by others in the automotive.
Method: Aggregate useful votes for each user by summing votes from all their reviews. Ranked users by their total useful votes and visualized the top 20 using a bar graph.

Result: John received the highest number of useful votes, followed by Michael and David, indicating their reviews were highly valued by the Yelp community.

<Cell Number 61 in jupyter notebook>

6. User Activity by Month

Objective: Examine the distribution of user review activity across months to identify trends.
Method: Extracted the month from review timestamps, grouped by month, and counted the number of reviews for each. Visualized the monthly review counts using a line chart.
Result: User activity peaked in January, followed by a steady decline and smaller peaks in March and August. Review activity dropped significantly in November, indicating possible seasonal patterns.

<Cell Number 62 in jupyter notebook>

7. Top 20 Users by Review Count

Objective: Identify the most active users in the automotive category based on the total number of reviews they have written.
Method: Grouped user review data by user name, calculated the total review count for each user, and ranked them in descending order. Visualized the top 20 users using a bar graph.
Result: John emerged as the most active reviewer, followed by Michael and David. These users contributed significantly to the review dataset in the automotive category.

<Cell Number 63 in jupyter notebook>

8. Top 20 Users by Compliments

Objective: Identify the users who received the most compliments for their reviews in the automotive category.
Method: Aggregated compliments (sum of hot, cool, funny, and other categories) for each user. Ranked users by their total compliments and visualized the top 20 using a bar graph.
Result: John received the highest number of compliments, followed by Marshall and Robert. These users demonstrated a strong influence and engagement within the Yelp community.

<Cell Number 64 in jupyter notebook>

9. Top 20 Users by Tip Count

Objective: Identify the users who provided the highest number of tips in the automotive category.
Method: Aggregated tip counts for each user in the automotive category and ranked users in descending order. Visualized the top 20 users using a bar graph.
Result: John was the most active user in giving tips, followed by Dr. Tim L and Yvonne. These users played a significant role in contributing additional insights to the Yelp community.

<Cell Number 65 in jupyter notebook>

10. Top 20 Users Reviewing Automotive Businesses

Objective: Identify the most active users contributing reviews for automotive businesses in Arizona.
Method: Filtered reviews for the automotive category in Arizona, grouped by user name, and aggregated review counts. Ranked users in descending order and visualized the top 20 using a bar graph.
Result: John was the top reviewer for automotive businesses, followed by Michael and David. These users demonstrated high engagement in the automotive category.

<Cell Number 66 in jupyter notebook>