Analysis of Arizona Businesses using Yelp Dataset

Varshini Rao Venkateshwara Rao
Robotics and Autonomous Systems (Artificial Intelligence)
Arizona State University
vvenka77@asu.edu

I. Introduction

The project analyzes user reviews of gyms in Arizona using the Yelp dataset. The gym industry plays a critical role in the health and wellness sector, making it a priority for businesses to understand both operational success and customer behavior. The analysis focuses on two milestones:
Business-Level Analysis: Investigating gym performance metrics, including ratings, reviews, and regional trends.
User-Level Analysis: Examining user behavior, contributions, and influence within the Yelp ecosystem.
The goal is to derive actionable insights that gyms can leverage to improve their services and customer engagement.

II. Description of the Solution

The solution involved conducting a comprehensive analysis of the Yelp dataset using distributed computing frameworks like Hadoop and Spark. The analysis was divided into two milestones—Business-Level Analysis (Milestone 1) and User-Level Analysis (Milestone 2)—to address different aspects of gym performance and user behavior. Below is a detailed breakdown:

A. Data Preparation

Dataset: The Yelp dataset included business, review, user, and check-in data for Arizona gyms. The dataset was originally in JSON format and contained over a million records.
Data Transformation: The JSON files were converted into Parquet format, a columnar storage file format, for faster processing and querying. Data filtering focused on gyms located in Arizona to narrow down the scope.
Environment Setup: A local Ubuntu 22.04 virtual machine was set up with pre-installed tools such as Hadoop, Apache Spark, PySpark, and Jupyter Notebook. Spark SQL was used to process and analyze the dataset.

B. Business-Level Analysis

Objective: Understand gym performance using business attributes and customer reviews.
Queries Developed:

- **Most Reviewed 5-Star Gym:** Extracted gyms with a 5-star rating and sorted them by review count to identify the most popular and customer-approved gym.
- **Top Zip Codes for Gym Activity:** Used review counts as a proxy for gym engagement and mapped the top-performing zip codes.
- **Seasonal Trends in Gym Reviews:** Grouped reviews by month to analyze customer activity patterns throughout the year.
- **Average Gym Ratings:** Calculated average ratings for each gym to highlight top performers and those requiring improvement.
- **Gyms with the Most Negative Reviews:** Focused on reviews with 2 stars or less to identify gyms with customer dissatisfaction.

Visualization: Created bar charts for queries such as review counts, seasonal trends, and top zip codes. A scatter plot was used to visualize the distribution of average ratings across gyms.

Insights: Identified gyms excelling in customer satisfaction and regions with the highest gym activity. Seasonal review trends revealed peaks and dips in customer engagement, guiding targeted promotions.

C. User-Level Analysis

Objective: Study user contributions and influence in the Yelp ecosystem.
Queries Developed:

- **Top Users by Review Count:** Ranked users by the number of reviews they left, identifying key influencers in the community.
- **Average Ratings by Users:** Calculated average ratings for individual users to assess general satisfaction trends and identify outliers.
- **Users with the Most "Useful" Votes:** Ranked users based on the number of "useful" votes their reviews received, highlighting contributors of valuable feedback.
- **User Account Age Distribution:** Analyzed the age of user accounts to evaluate retention and platform loyalty.
- **Most Active Contributors:** Combined review and tip counts to identify users with the highest overall contributions.
- **Gyms with the Most Fans:** Ranked gyms by their fan count to measure customer loyalty.
- **Retention of Returning Users:** Counted the number of users who repeatedly visited a gym, showcasing retention efforts.
- **Correlation Between Review Count and Ratings:** Explored the relationship between user review frequency and their rating patterns, uncovering biases.

- **Funniest Reviews:** Highlighted reviews with the most "funny" votes, showing the value of engaging content.
- **First-Time vs. Returning Users:** Compared counts of first-time users to returning users for each gym.

Visualization: Bar charts were used for metrics such as top users, useful votes, and retention rates. A scatter plot visualized the correlation between review count and average ratings. A grouped bar chart compared first-time and returning users.

Insights: Key influencers and loyal customers were identified, providing opportunities for targeted engagement. User behavior patterns revealed satisfaction levels, feedback quality, and gym retention strategies.

III. Results

A. Business-Level Analysis

Most Reviewed 5-Star Gym:

- **Finding:** "Let's Sweat" stood out as the most reviewed 5-star gym in Arizona, with over 50 reviews.
- **Use:** This gym sets a benchmark for customer satisfaction and engagement. Other gyms can study their operations, customer service practices, and marketing strategies to improve their own services.

Top Zip Codes for Gym Activity:

- **Finding:**The zip code 85719 had the highest engagement, with over 600 reviews, followed by 85712 and others.
- **Use:** These areas show high demand for fitness services. Gyms operating here should ensure adequate resources (staff, equipment, and space) to meet the demand. Businesses in less active areas can consider targeted marketing campaigns.

Seasonal Trends in Gym Reviews:

- **Finding:** January saw a peak in reviews, likely due to New Year fitness resolutions. October showed another increase as people prepared for the holiday season. Review activity dipped in February and March, reflecting waning enthusiasm after New Year resolutions.
- **Use:** Gyms can align their promotional efforts with these trends. For instance, offering New Year discounts or pre-holiday packages can boost memberships during peak seasons. During slower months, gyms can focus on customer retention strategies, like loyalty rewards or specialized classes.

Top-Rated and Low-Rated Gyms:

- **Finding:** Gyms like "Sierra Fitness" and "CrossFit Lanista" achieved perfect 5-star ratings, while others, such as "Bally Total Fitness," had ratings as low as 3.0.
- **Use:** High-rated gyms can leverage their ratings in marketing campaigns, while low-rated gyms need to address customer feedback to enhance their reputation.

Gyms with Most Negative Reviews:

- **Finding:** "Banner Health" and "LA Fitness" received the highest number of negative reviews, highlighting significant dissatisfaction.
- **Use:** These gyms can benefit from investigating and addressing specific complaints (e.g., customer service, cleanliness, or equipment maintenance) to rebuild trust.

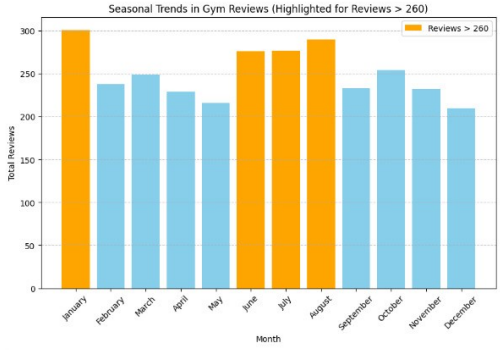

Fig. 1. An example of a bar chart showing seasonal trends in gym usage

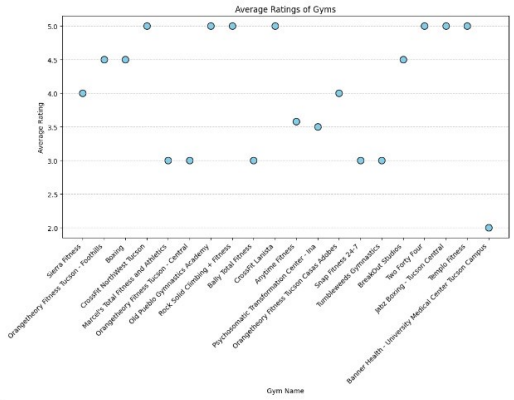

Fig. 2. An example of a scatter plot showing average ratings of gyms

B. User-Level Analysis

Top Users by Review Count:

- **Finding:** The top 10 reviewers were identified as key influencers who actively contribute to gym reviews.
- **Use:** Gyms can engage with these influencers by offering promotions or requesting feedback. Positive reviews from them can significantly boost a gym's visibility and reputation.

Average Ratings by Users:

- **Finding:** Most users provided moderately high ratings, indicating overall satisfaction. However, a few outliers consistently left lower ratings, possibly signaling unresolved issues.

- Use: Understanding user sentiment helps gyms improve services and target dissatisfied customers to rebuild relationships.

Users with the Most "Useful" Votes:

- **Finding:** Certain users received the highest "useful" votes, indicating their reviews and tips provided valuable feedback.
- **Use:** Gyms can analyze these reviews to implement constructive suggestions and improve customer experiences.

Account Age and Retention:

- **Finding:** Users with older accounts demonstrated loyalty, while newer accounts highlighted platform growth.
- **Use:** Loyal users can serve as brand ambassadors, while new users may benefit from introductory offers or onboarding programs.

Gym Retention and Fan Count:

- **Finding:** Gyms with high fan counts and retention rates, such as "Orangetheory Fitness," excelled in customer loyalty.
- **Use:** Such gyms set an example for building community and customer trust, which can inspire other businesses.

First-Time vs. Returning Users:

- **Finding:**Gyms that attracted more first-time users demonstrated effective marketing, while those with more returning users excelled in customer retention.
- **Use:** Striking a balance between acquiring new customers and retaining existing ones is essential for long-term growth.

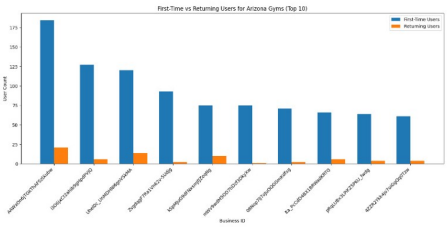

Fig. 3. An example of a grouped bar chart for comparison between first time and returning users

C. Applications of Findings

- **Business Strategy and Resource Allocation:** Gyms in high-demand areas can prioritize resource allocation, while those in low-demand areas can use targeted marketing to attract customers.
- **Marketing and Promotions:** Seasonal trends provide a roadmap for creating effective promotional campaigns during peak periods, such as New Year or October.
- **Customer Feedback Implementation:** Insights from top reviewers and "useful" votes can guide service improvements, enhancing customer satisfaction.
- **Retention Strategies:** High fan counts and returning user data underline the importance of loyalty programs and community-building efforts.
- **Data-Driven Decisions:** The correlation between reviews, ratings, and other factors enables gyms to adopt data-driven approaches for growth and operational efficiency.

IV. Contributions

A. Project Type

This was an individual project, meaning all aspects of the analysis, queries, and visualizations were conceptualized, implemented, and completed solely by me. The project was executed independently, ensuring that all insights and outputs reflected personal effort and problem-solving approaches.

B. Specific Contributions:

- **Data Analysis:** Formulated and implemented SQL queries to extract relevant insights from the dataset. Designed queries for both business-level analysis and user-level analysis, focusing on key factors like customer engagement, seasonal trends, user activity, and satisfaction metrics.
- **Visualization:** Created meaningful and intuitive visualizations such as bar charts and scatter plots to represent trends and findings effectively. Employed visual storytelling techniques to highlight key data points and actionable insights.
- **Insight Generation:** Interpreted data patterns to provide actionable recommendations for gyms, such as enhancing customer loyalty, targeting high-demand regions, and addressing service gaps.
- **Documentation:** Prepared detailed reports summarizing the findings, ensuring clarity and alignment with the project objectives.

C. Skills Acquired

- **Technical Skills:** Advanced proficiency in SQL for querying large datasets using concepts of Hadoop and Spark. Enhanced ability to create data visualizations using tools like Python (Matplotlib). Developed skills in data cleaning and preprocessing, ensuring the dataset was suitable for analysis.
- **Analytical Thinking:** Improved ability to interpret data trends and derive actionable insights from large, complex datasets.Gained a deeper understanding of customer engagement metrics, including the impact of reviews, ratings, and seasonal trends on business success.

References

[1] Yelp Open Dataset. "Dataset for academic research." https://www.yelp.com/dataset/documentation/main

[2] Apache Spark Documentation: https://spark.apache.org/docs/latest/api/python/index.html

[3] Ubuntu Virtual Machine Setup: https://askubuntu.com/questions/293348/how-do-i-set-up-an-ubuntu-virtual-machine-inside-ubuntu

[4] Apache Hadoop Documenation: https://hadoop.apache.org/docs/r1.2.1/hdfs_design.html