CSE 511 - DPS

Milestone 1:Business Level Analysis

Abhi Dineshkumar Patel
Student ID: 1230038877

1. Analysis using spark.sql queries

1.1 Top 10 Cities by Number of Gyms

The data revealed Tucson as the leading city in Arizona, hosting the highest number of gyms. Other cities with a significant number of gyms include Oro Valley, Marana, and Green Valley. The results are visualized in the chart below:

Inputs:
>
> Dataset: gyms table filtered for businesses in Arizona (state = 'AZ').
>
> Columns used: city, state.

Query Logic:
> Group gyms by city.
> Count the total gyms in each city.
> Sort in descending order of gym count.
> Limit the results to the top 10 cities.

Visualization:
> Generate a bar chart showing cities on the x-axis and gym count on the y-axis.

Output:
>
> Insights: Tucson ranks highest with 61 gyms, followed by Oro Valley (21) and Marana (11).
>
> Decision-Making Value: Tucson has a high concentration of gyms, suggesting intense competition and demand in the area.

1.2 Top Gym Categories by Average Star Ratings

Gyms categorized under mostly all categories like "Massage Therapy" and "Fitness & Instruction" consistently scored the highest average ratings, with perfect scores of 5.0. This indicates that specialized services or classes attract higher customer satisfaction.

Inputs:
>
> Dataset: gyms table filtered for Arizona businesses.
>
> Columns used: categories, stars.

Query Logic:
> Group gyms by categories.
> Calculate the average star rating for each category.
> Sort by avg_rating in descending order.
> Limit results to the top 10 categories.

Visualization:
> Generate a bar chart showing gym categories (shortened for readability) on the x-axis and average ratings on the y-axis.

Output:
>
> Insights: Mostly all Categories like "Massage Therapy" and "Fitness & Instruction" received the highest average ratings of 5.0, indicating that customers value specialized services.
>
> Decision-Making Value: Gym owners could focus on offering niche services to attract higher ratings and customer loyalty.

1.3 Top-Rated Gyms with Over 100 Reviews

Among gyms with substantial customer engagement (100+ reviews), Chuze Fitness in Tucson emerged as the top-rated gym with an impressive 3.9 average rating from 148 reviews. LA Fitness and Banner Health also performed well in terms of both ratings and review counts.

Inputs:

Dataset: gyms and review tables joined on business_id.

Columns used: name, city, postal_code, stars, review_id.

Query Logic:

Perform an inner join between gyms and review.
Group by gym name, city, and postal code.
Calculate the average rating (avg_rating) and count total reviews (total_reviews).
Filter gyms with more than 100 reviews.
Sort by avg_rating in descending order and limit to the top 10 gyms.

Visualization:

Generate a bar chart showing gym names on the x-axis and average ratings on the y-axis.

Output:

Insights: Chuze Fitness in Tucson has the highest average rating (3.9) among gyms with over 100 reviews.
Decision-Making Value: High review counts with strong ratings suggest sustained customer satisfaction and trust.

1.4 Most Popular ZIP Codes for Gyms in Arizona

The ZIP code 85750 in Tucson recorded the highest average rating (4.37) and a significant total review count (581). Other well-performing ZIP codes include 85718 and 85745, indicating geographical clusters of high-performing gyms.

Inputs:

Dataset: gyms and review tables joined on business_id.

Columns used: postal_code, stars, review_id.

Query Logic:

Perform an inner join between gyms and review.
Group by postal_code.
Calculate the average rating and total review count for each ZIP code.
Sort by avg_rating and total_reviews in descending order.
Limit results to the top 10 ZIP codes.

Visualization:

Generate a bar chart showing ZIP codes on the x-axis and average ratings on the y-axis.

Output:

Insights: 85750 (Tucson) has the highest average rating (4.37) and review count (581).
Decision-Making Value: ZIP code clusters like 85750 indicate lucrative areas for gym businesses.

1.5 Correlation Between Number of Reviews and Ratings

The analysis of gyms with more than 20 reviews highlighted that businesses with higher review counts generally maintained solid average ratings. For instance, LA Fitness in Tucson had 410 reviews with a 3.27 average rating, indicating sustained customer interest and satisfaction.

Inputs:

Dataset: gyms and review tables joined on business_id.

Columns used: name, city, review_id, stars.

Query Logic:

Perform an inner join between gyms and review.
Group by gym name and city.
Calculate total reviews and average ratings for each gym.
Filter gyms with more than 20 reviews.
Sort by total reviews in descending order and limit to the top 10 gyms.

Visualization:

Generate a bar chart showing total reviews on the x-axis and average ratings on the y-axis.

Output:

Insights: Gyms with more reviews tend to maintain high ratings, indicating customer trust and satisfaction. For instance, LA Fitness in Tucson had 410 reviews with a 3.27 average rating.
Decision-Making Value: Fostering customer engagement (e.g., incentivizing reviews) can help build reputation and visibility.