Name: Bhargav Reddy Vangala
ASU-ID: 1226491476

Milestone 1 - Individual Report- Business Level Analysis

Introduction

This report presents an analysis of the Health & Medical sector within the Arizona (AZ) region, utilizing data from the Yelp Academic Dataset. The purpose of this analysis is to provide insights into the performance and characteristics of Health & Medical businesses, including popular locations, customer ratings, and key attributes that contribute to customer satisfaction.

The dataset includes information on various business attributes, user reviews, ratings, and check-ins. By leveraging Spark SQL and data filtering, this analysis focuses on businesses categorized as Health & Medical, aiming to answer questions such as:

What are the average ratings and review counts for Health & Medical businesses in Arizona?
Which zip codes have the highest-rated Health & Medical businesses?
What attributes are common among highly-rated Health & Medical businesses?
Through this analysis, I hope to uncover trends that highlight customer preferences and location-based performance indicators for Health & Medical services. The results may offer valuable insights for business owners, potential investors, and local policymakers to improve service quality and cater to customer needs in Arizona's Health & Medical sector.

Problem Statement

The goal of this analysis is to examine the performance and customer preferences for Health & Medical businesses in Arizona, using data from Yelp. Health & Medical services encompass a wide range of facilities, from general medical practices to specialized health services. However, understanding what drives customer satisfaction and engagement in this sector remains a challenge due to the diversity of services and varying customer expectations.
This report seeks to address the following questions:
What are the average customer ratings and review counts for Health & Medical businesses? – By analyzing these metrics, I aim to understand overall customer satisfaction and identify popular businesses within this category.
Are there specific geographic areas (zip codes) in Arizona where Health & Medical businesses perform particularly well? – Identifying high-performing locations can highlight areas where customer needs are well-met, potentially guiding new businesses in choosing suitable locations.
What attributes are common among highly-rated Health & Medical businesses? – Understanding the specific features or services that top-rated businesses offer can provide valuable insights for improving customer satisfaction across the sector.
By addressing these questions, this analysis aims to uncover patterns and preferences that may help Health & Medical businesses enhance their services, better cater to customer needs, and identify strategic locations for growth.

Main Results and Analysis

In this section, I present the findings of our analysis on Health & Medical businesses in Arizona. Each aspect of the analysis is supported by data retrieved from Spark SQL queries, providing insights into customer ratings, geographic trends, and key attributes valued by customers in this sector.

1. Business Categories in Arizona(Simple Query)

To identify the most common business categories in Arizona, I analyzed the dataset by grouping businesses based on their categories and counting the number of businesses in each category. This analysis helps understand the diversity of businesses and highlights which sectors dominate the Arizona market.

Observation: The analysis revealed that "Restaurants" is the most common business category in Arizona, followed by "Shopping" and "Health & Medical." This indicates that the service industry, especially food and healthcare-related businesses, has a significant presence in the state.

Insights: The prominence of "Health & Medical" in the top categories reflects the importance of healthcare services in Arizona. It also suggests potential opportunities for new businesses in emerging or less saturated categories.

```
# simple query-1
query1_result = spark.sql("""
    SELECT
        category,
        COUNT(business_id) AS business_count
    FROM
        business_az_exploded
    GROUP BY
        category
    ORDER BY
        business_count DESC
    LIMIT 10
""")
query1_result.show()
```

0.6s

```
+----------------+--------------+
|        category|business_count|
+----------------+--------------+
|     Restaurants|          2671|
|        Shopping|          1780|
|            Food|          1520|
|   Home Services|          1390|
|Health & Medical|          1032|
|   Beauty & Spas|           999|
| Local Services|           935|
|      Automotive|           872|
|         Mexican|           547|
|       Nightlife|           546|
+----------------+--------------+
```

Fig-1:Sample output of vs code

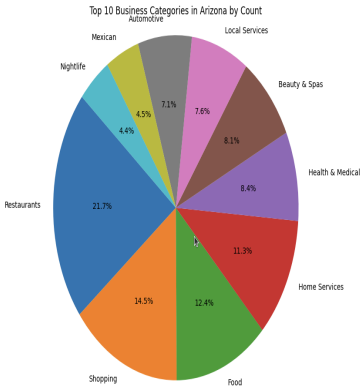


Fig-2: Pie chart showing the categories division

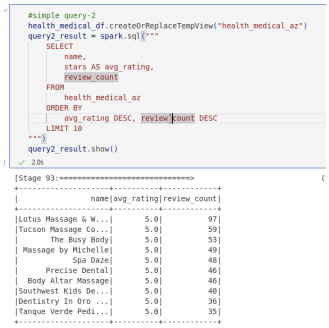
2. Top-Rated Health & Medical Businesses(Simple Query)

To identify the top-performing Health & Medical businesses in Arizona, I analyzed the dataset to rank businesses by their average ratings and review counts. Businesses with high ratings and a significant number of reviews are considered both popular and well-regarded by customers.

Observation: The analysis revealed that the top-rated Health & Medical businesses in Arizona consistently have ratings close to 5.0. These businesses also have a substantial number of reviews, indicating strong customer engagement and satisfaction.

Insights: Businesses with high ratings and review counts often provide superior services or offer unique value to their customers. These top performers set a benchmark for other businesses in the Health &

Medical sector, suggesting the importance of maintaining high service quality and building customer trust through consistent interactions.



```
#simple query-2
health_medical_df.createOrReplaceTempView("health_medical_az")
query2_result = spark.sql("""
    SELECT
        name,
        stars AS avg_rating,
        review_count
    FROM
        health_medical_az
    ORDER BY
        avg_rating DESC, review_count DESC
    LIMIT 10
""")
query2_result.show()
```

✓ 2.0s

[Stage 93:==========>

name	avg_rating	review_count
Lotus Massage & W...	5.0	97
Tucson Massage Co...	5.0	59
The Busy Body	5.0	53
Massage by Michelle	5.0	49
Spa Daze	5.0	48
Precise Dental	5.0	46
Body Altar Massage	5.0	46
Southwest Kids De...	5.0	40
Dentistry In Oro ...	5.0	36
Tanque Verde Pedi...	5.0	35

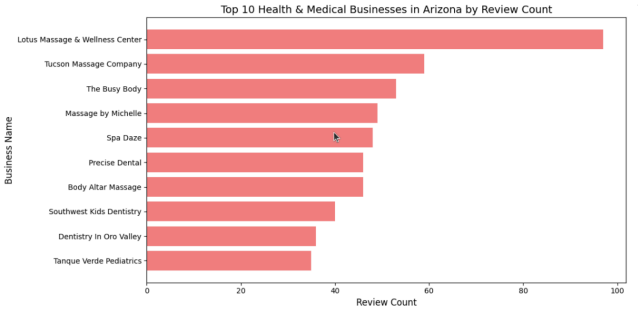


Fig-3:Sample output of vs code

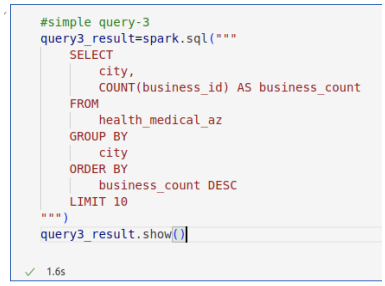
Fig-4: Bar chart showing name vs review count

3. Distribution of Health & Medical Businesses Across Cities(Simple Query)

To understand the geographic distribution of Health & Medical businesses in Arizona, I analyzed the dataset by counting the number of businesses in each city. This analysis highlights which cities have the most significant presence of Health & Medical services.

Observation: The analysis revealed that major urban center **Tucson** has the highest number of Health & Medical businesses. This is expected as these cities have larger populations and demand for healthcare services.

Insights: The high concentration of businesses in major cities suggests that these areas are well-served. However, smaller cities with fewer businesses might present opportunities for new entrants to address underserved populations. Understanding this distribution helps in strategic planning for service expansion and resource allocation.



```
#simple query-3
query3_result=spark.sql("""
    SELECT
        city,
        COUNT(business_id) AS business_count
    FROM
        health_medical_az
    GROUP BY
        city
    ORDER BY
        business_count DESC
    LIMIT 10
""")
query3_result.show()
```

✓ 1.6s

city	business_count
Tucson	966
Oro Valley	45
Marana	6
Green Valley	5
Vail	4
Sahuarita	4
South Tucson	1
Corona de Tucson	1

Fig-5:Sample out of vs code of simple query-3

4. Best-Performing Zip Codes for Health & Medical Businesses(Complex Query)

To identify the top-performing zip codes in Arizona for Health & Medical businesses, I analyzed average ratings and total reviews grouped by postal code. This analysis provides insights into geographic areas with high customer satisfaction and engagement.

Observation: The analysis showed that specific zip codes have significantly higher average ratings and review counts. These zip codes often represent areas with high-quality Health & Medical services, indicating that customers in these locations are highly satisfied.

Insights: High-performing zip codes can act as benchmarks for other regions. Businesses operating in these areas likely have better infrastructure, customer service, or quality standards. These zip codes can also guide new businesses in choosing ideal locations for setting up operations to capture existing demand.

```
#Complex query-1
query4_result = spark.sql("""
    SELECT
        postal_code,
        AVG(r.stars) AS avg_rating,
        COUNT(r.review_id) AS total_reviews
    FROM
        health_medical_az b
    JOIN
        review r
    ON
        b.business_id = r.business_id
    GROUP BY
        postal_code
    ORDER BY
        avg_rating DESC, total_reviews DESC
    LIMIT 10
""")
query4_result.show()
```

✓ 58.2s

```
[Stage 114:==========================================
+-----------+------------------+-------------+
|postal_code|        avg_rating|total_reviews|
+-----------+------------------+-------------+
|      85735|               5.0|            6|
|      85752|               5.0|            5|
|      85658|             4.275|           40|
|      85742| 4.103174603174603|          126|
|      85716| 4.009389671361502|         1065|
|      85749| 3.989247311827957|           93|
|      85701|3.9551282051282053|          156|
|      85737|3.9419354838709677|          310|
|      85715| 3.765695067264574|          892|
|      85756|3.7333333333333334|           45|
+-----------+------------------+-------------+
```

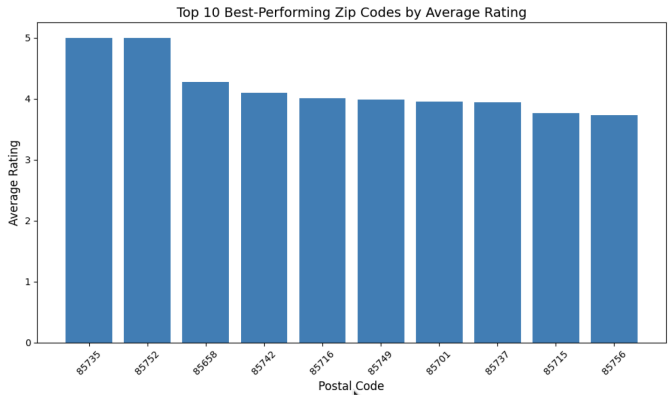


Fig-6:Sample out of vs code of complex query-1 Fig-7: Bar chart showing rating vs pincodes

5. Sentiment Analysis of Reviews and Business Ratings(Complex Query)

To analyze customer sentiment and its relationship with business performance, I evaluated key metrics such as average review ratings, review lengths, and engagement scores (useful, funny, cool) for Health & Medical businesses in Arizona. This analysis provides insights into customer satisfaction and interaction patterns.

Observation: Businesses with higher average ratings often had longer and more detailed reviews, indicating that satisfied customers tend to leave comprehensive feedback. Engagement metrics such as "useful" were higher for these businesses, reflecting the relevance of the services offered.

Insights: High average ratings and detailed reviews often correlate with better customer service or higher service quality. Businesses aiming to improve their customer sentiment could focus on enhancing customer interactions, as this not only improves ratings but also encourages detailed and positive reviews. Metrics like "useful" reviews further indicate how well a business meets customer needs.

```
#complex query
query5_result = spark.sql("""
    SELECT
        b.business_id,
        b.name AS business_name,
        b.city,
        AVG(r.stars) AS average_rating,
        AVG(LENGTH(r.text)) AS average_review_length,
        AVG(r.useful) AS avg_useful,
        AVG(r.funny) AS avg_funny,
        AVG(r.cool) AS avg_cool
    FROM
        health_medical_az b
    JOIN
        review r
    ON
        b.business_id = r.business_id
    GROUP BY
        b.business_id, b.name, b.city
    ORDER BY
        average_rating DESC, average_review_length DESC
    LIMIT 10
""")

query5_result.show()
```

1m 54.2s

```
[Stage 127:=============================================================> (39 + 1) / 40]
+--------------------+--------------------+---------+--------------+---------------------+--------------------+---------+--------------------+
|         business_id|       business_name|     city|average_rating|average_review_length|          avg_useful|avg_funny|            avg_cool|
+--------------------+--------------------+---------+--------------+---------------------+--------------------+---------+--------------------+
|VMWg_Hq-H1CRbClqL...|  Tucson Biofeedback|   Tucson|           5.0|   1399.6666666666667|  1.8333333333333333|      0.0| 0.6666666666666666|
|bZhEMLedWEOB7EfaH...| Old Pueblo CrossFit|   Tucson|           5.0|                979.4|                 0.4|      0.0|                 0.2|
|biPVZx5C2s1xKIxfQ...|Switzer Family Ch...|   Tucson|           5.0|    821.4285714285714|  1.1428571428571428|      0.0|                 0.0|
|LuPr_bcTIwVI8oG-...|   Old Pueblo Dental|   Tucson|           5.0|    818.9333333333333|                 0.6|      0.0|0.13333333333333333|
|KCxY73zVECgiWFhIq...|       Blessed Birth|   Tucson|           5.0|    790.1111111111111|                 1.0|      0.0|                 0.0|
|hY_cynaIWHbGNHPYc...|          Adler Fit|   Tucson|           5.0|    779.5714285714286|0.047619047619047616|      0.0|                 0.0|
|jY0Ysv837btsFJW0r...|Movement for Life...|   Tucson|           5.0|    770.3333333333334|   5.333333333333333|      0.0| 0.9333333333333333|
|M0LkNNnn9iVsqPQrz...|Oro Valley Dental...|Oro Valley|           5.0|                763.75|                0.75|      0.0|                 0.0|
|iD2YvqNw8obCnZv2f...|  Thornydale Dental|   Tucson|           5.0|                756.5|                 1.0|      0.0|               0.125|
|FzTTJn17WfZOsyRz1...|Your Family's Jou...|   Tucson|           5.0|              740.875|               0.375|      0.0|                0.25|
+--------------------+--------------------+---------+--------------+---------------------+--------------------+---------+--------------------+
```

Fig-8:Sample out of vs code of complex query-2

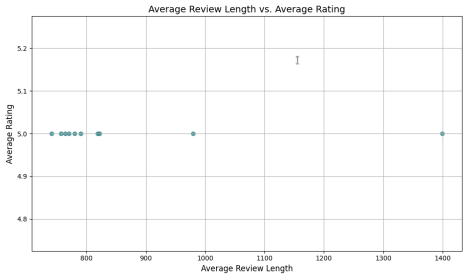


Fig-9: Plot showing rating vs review

6. Trend Analysis of Average Ratings Over Time(Complex Query)

In order to investigate patterns in customer satisfaction over time, I looked at the average ratings and annual reviews for Arizona's health and medical companies. The evolution of consumer sentiment and interaction over time is revealed by this research.

Observation: The analysis revealed that average ratings for Health & Medical businesses have remained relatively consistent, with slight fluctuations across years. The number of reviews, however, has shown a steady increase, indicating growing customer engagement and usage of Health & Medical services in Arizona.

Insights: A consistent average rating suggests stable service quality in the sector, while the rising number of reviews reflects increasing customer interaction with businesses. This trend highlights the importance of maintaining high service standards to continue meeting customer expectations as engagement grows.

```
spark.conf.set("spark.sql.legacy.timeParserPolicy", "LEGACY")
health_medical_reviews_df = review_df.join(
    health_medical_df.select("business_id", "name", "city"),
    on="business_id",
    how="inner"
)
health_medical_reviews_with_year_df = health_medical_reviews_df.withColumn(
    "year", year(to_timestamp("date", "yyyy-MM-dd HH:mm:ss"))
)
health_medical_reviews_with_year_df.createOrReplaceTempView("health_reviews_year")
query6_result = spark.sql("""
    SELECT year,
           ROUND(AVG(stars), 2) AS avg_stars,
           COUNT(*) AS num_reviews
    FROM health_reviews_year
    WHERE year IS NOT NULL
    GROUP BY year
    ORDER BY year
""")

query6_result.show()
```

1m 37.0s

```
[Stage 136:=============================================================> (39 + 1) / 40]
+----+---------+-----------+
|year|avg_stars|num_reviews|
+----+---------+-----------+
|2005|      4.0|          1|
|2006|      5.0|          4|
|2007|      4.0|          3|
|2008|      4.1|         20|
|2009|     3.35|         43|
|2010|     4.05|         91|
|2011|     3.92|        295|
|2012|     3.81|        537|
|2013|     3.59|        806|
|2014|     3.56|       1124|
|2015|     3.52|       1491|
|2016|     3.59|       1938|
|2017|     3.62|       2340|
|2018|     3.65|       2778|
|2019|     3.52|       2858|
|2020|     3.24|       2008|
```

Fig-10:Sample out of vs code of complex query-3

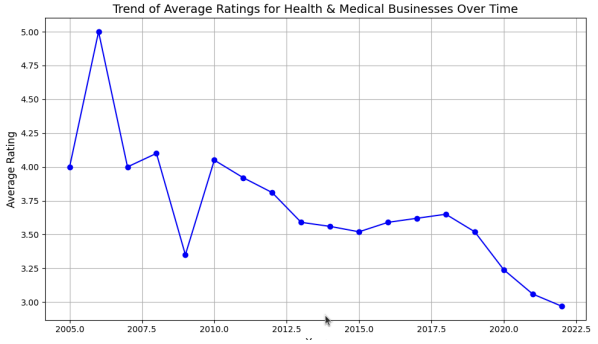


Fig-11: Plot showing health reviews over time