Mark Shinozaki
Cpts 451
Yelp: Classifying a ‘Popular’ and ‘Successful’ Business

1. Introduction

- In Milestone 2, we aim to classify sample data of Yelp businesses as “popular” and “successful” using various metrics. Many popular businesses attract a variety of customers due to substance of reviews or the quantity of those warm reviews, successful businesses have potentially lived a longer life in some communities therefore have more loyal customers and a larger pool of reviews and feedback. We propose specific metrics to identify successful businesses and popular businesses, in the process, describe our methodologies and data analysis in this report.

2. Description of Metrics

- In this section, we define two sets of metrics to classify businesses: “popular” metric and a “success” metric. These metrics utilize data from check-ins and reviews to provide insight into business performance.

3. Metrics for Popularity

- Metric 1: Average Check-ins per day
    - **Description**: This would measure the frequency of customer visits.
    - **Data Used**: Check-in data.
    - **Query**:

```
SELECT business_id, AVG(count) AS avg_checkins
FROM checkintable
GROUP BY business_id;
```

- Metric 2: Review Count
    - **Description**: Indicates the volume of customer feedback.
    - **Data Used**: Review Data
    - **Query**:

```
SELECT business_id, COUNT(*) AS review_count
FROM reviewtable
GROUP BY business_id;
```

4. Metrics for Success

- Metric 1: Business Longevity
    - **Description**: Measures the duration a business has been operational.
    - **Data Used**: First review date as a proxy for business start date.
    - **Query**:

```
SELECT business_id, MIN(date) AS start_date
FROM reviewtable
GROUP BY business_id;
```

Mark Shinozaki
Cpts 451
Yelp: Classifying a ‘Popular’ and ‘Successful’ Business

- Metric 2: Average Review Rating
    - **Description**: Reflects the quality of service/products.
    - **Data Used**: Review data.
    - **Query**:

```
SELECT business_id, ROUND(AVG(stars), 2) AS avg_rating
FROM reviewtable
GROUP BY business_id;
```

5. Data Pre-processing and Extraction

- To determine the most successful business using ‘business_id’, ‘zipcode’, and ‘numcheckins’ we can use a query that groups businesses by ‘zipcode’ and then selects the business with the highest ‘numcheckins’ within each ‘zipcode’ . This could help identify businesses with higher-than-average check-ins.
    - **Query**:

```
SELECT b.zipcode, b.business_id, b.name, b.numcheckins
FROM business b
JOIN (
  SELECT zipcode, MAX(numcheckins) AS max_checkins
  FROM business
  GROUP BY zipcode
) max_checkins_per_zip
ON b.zipcode = max_checkins_per_zip.zipcode
AND b.numcheckins = max_checkins_per_zip.max_checkins
ORDER BY b.zipcode;
```

- This query does the following:
    - Subquery(max_checkins_per_zip):
        - Groups the ‘business’ table by ‘zipcode’ and calculates the maximum number of check-ins (‘numcheckins’) for each ‘zipcode’.
    - Main Query:
        - Joins the ‘business’ table with the subquery on ‘zipcode’ and ‘numcheckins’
        - Selects the ‘zipcode’, ‘business_id’, ‘name’, and ‘numcheckins’ of the business that have the highest number of check-ins with each ‘zipcode’.