Huy (Harry) Ky, Nhan Nguyen
CptS 451
Milestone 2 Businesses Classification Proposal
4/6/2025

Metrics for Classifying Yelp Businesses

Classifications:

- **Popular:**
  Businesses that attract a high volume of check-ins compared to their local peers, indicating strong customer engagement and foot traffic.
- **Expensive:**
  Businesses that are flagged for having premium pricing (typically with price values of 3 or 4), which often correlates with upscale products or services.
- **Successful:**
  Businesses that have demonstrated long-term viability and customer loyalty, as evidenced by consistent review activity, sustained operational years, and high ratings.

Popularity Metric

A business is considered popular if it has significantly more check-ins than others in the same category and zipcode, adjusted for population size. This uses check-in counts, category names, zipcodes, and population data from the zipcodeData table.

To compute the average check-ins per (zipcode, category) pair, we use the following SQL query:

```
CREATE TABLE avg_checkins_zipcat (
        zipcode VARCHAR(20),
        cname VARCHAR(100),
        avg_checkins NUMERIC
);

INSERT INTO avg_checkins_zipcat
SELECT b.zipcode, bc.cname, AVG(b.numCheckins)
FROM business b
JOIN business_category bc ON b.business_id = bc.business_id
GROUP BY b.zipcode, bc.cname;
```

Once the averages are calculated, the Popularity Score is determined using the formula:
Popularity Score = (Business Check-ins / Avg Check-ins for Zipcode+Category) * (1 + Zipcode Population / 100,000)

Success Metric

A business is considered successful if it has operated for many years, maintains consistent customer engagement, and receives high ratings. Success is based on review dates, number of reviews, and average rating.

We first calculate the years the business has been active and the average number of reviews it receives per year using the following SQL query:

```
CREATE TABLE business_lifetime (
	business_id VARCHAR(30),
	years_active INT,
	avg_reviews_per_year NUMERIC
);

INSERT INTO business_lifetime
SELECT business_id,
	EXTRACT(YEAR FROM MAX(date)) - EXTRACT(YEAR FROM MIN(date)) + 1 AS
years_active,
	COUNT(*) / (EXTRACT(YEAR FROM MAX(date)) - EXTRACT(YEAR FROM MIN(date))
+ 1) AS avg_reviews_per_year
FROM review
GROUP BY business_id;
```

The Success Score is then calculated using the formula:
Success Score = (Years Active) * (Avg Reviews per Year) * (Review Rating / 5.0)

Expensiveness Flag

The Expensiveness Flag identifies upscale businesses with premium prices. It's set when a business has a "price" attribute in the business_attribute table with a value of 3 or 4. The reason is the price ranges from 1 to 4, so 3 and 4 are on the expensive side.

The relevant SQL query is as follows:

```
SELECT business_id, value
FROM business_attribute
WHERE LOWER(aname) LIKE '%price%'
	AND value IN ('3', '4');
```