Yelp Analysis Final Report

Nathan Cronin

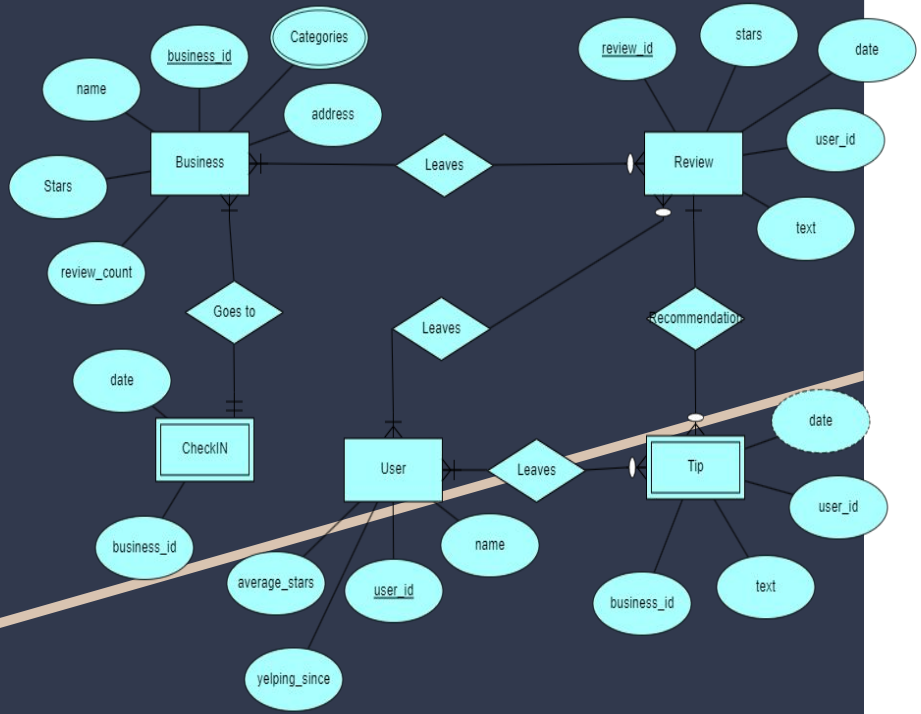
Client:

CoffeeKing Management

- A new startup coffee company providing a unique and novel experience to their customers

- Looking for insights & recommendations to improve their business
    - Location Selection
    - Hours of Operation
    - Clientele
    - Product Advice
    - General Recommendations

Yelp Dataset



- Free dataset taken from Yelp's website
- Consists of 5 unique tables
  - Business
  - Review
  - User
  - Check-In
  - Tip

- Data includes:
  - 7 million reviews
  - 150,000 unique businesses
  - 25 states
  - Reviews from 2005 - 2022

Questions & Hypotheses

Questions

- What is the distribution of positive, neutral, and negative reviews for both businesses and users?
- Does a higher review count correlate with a higher star rating?
- What are some of the most influential factors of a user's experience?
- Does location of the business play any factor?
- Is there any information detailed as to why some businesses closed and others are open?

Hypotheses

- On average, businesses with a higher review count will have a higher star count
- Users with a very low review count will generally have a more extreme review
- Customer service is more influential in shaping the review than the quality of the product itself

Approach

- Clean data to ensure accuracy and integrity
- Gather descriptive and inferential statistics using SQL queries for general insights
- Manipulate data to get data we need in right format using SQL & Excel
- Find any correlation that can explain a businesses star rating
- Write detailed queries to create a logistic regression model that will be able to forecast user sentiment
- Create visualizations to show location data and recommendations
- Be aware of limitations of dataset

Initial Exploration

Users

Min: 0 reviews

Avg: 23 Reviews

Max: 17,473 reviews

Row	quartiles	average_reviews	average_stars
1	1	1.0	3.27
2	2	3.0	3.64
3	3	10.0	3.78
4	4	79.0	3.83

quartiles

***Users with minimal reviews trended negative*

Businesses

Min: 5 reviews

Avg: 45 reviews

Max: 7568 reviews

Row	quartiles	average_reviews	average_stars
1	1	6.0	3.61
2	2	11.0	3.57
3	3	23.0	3.52
4	4	140.0	3.69

***Review count did not affect business stars*

Initial Exploration cont.

- The dataset was definitely left-skewed with significantly more positively rated businesses than negative

- Will be able to analyze positives but not negatives as well

All businesses

Row	stars	business_count
1	5.0	16307
2	4.5	27181
3	4.0	31124
4	3.5	26519
5	3.0	18453
6	2.5	14316
7	2.0	9527
8	1.5	4932
9	1.0	1986

Coffee Businesses

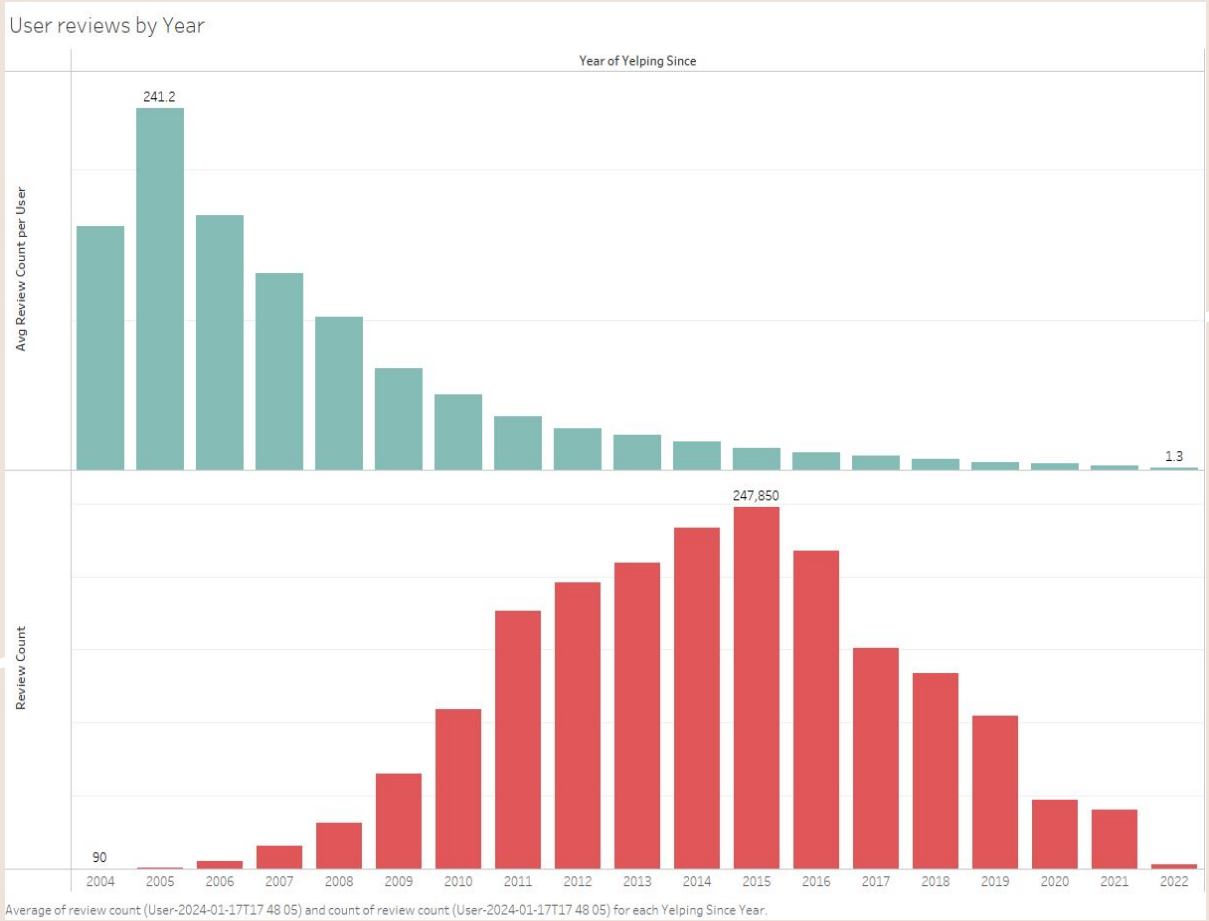
Row	stars	business_count
1	5.0	56
2	4.5	333
3	4.0	260

Avg # of reviews by user per year

- Right-skewed, negative relationship
- Individual users leaving less and less reviews every year
- Important to embrace wide range of clientele

Total # of reviews by year

- Normal distribution
- Reviews peaked in year 2015
- Shows necessity to make a solid first impression

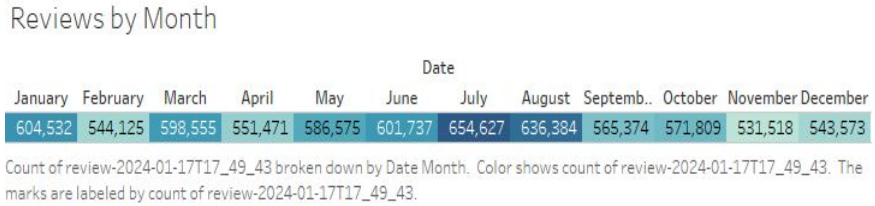


Hours of Operation Analysis

Hours Open	Business Count	Avg Star Rating
Sun-Sat: 7 AM to 2 PM	8	4.25
Sun-Sat: 8 AM to 3 PM	6	4.25
Sun-Sat: 7 AM to 5 PM	6	4.25
Sun-Sat: 7 AM to 4 PM	5	4.1
Sun-Sat: 8 AM to 2 PM	5	4.3

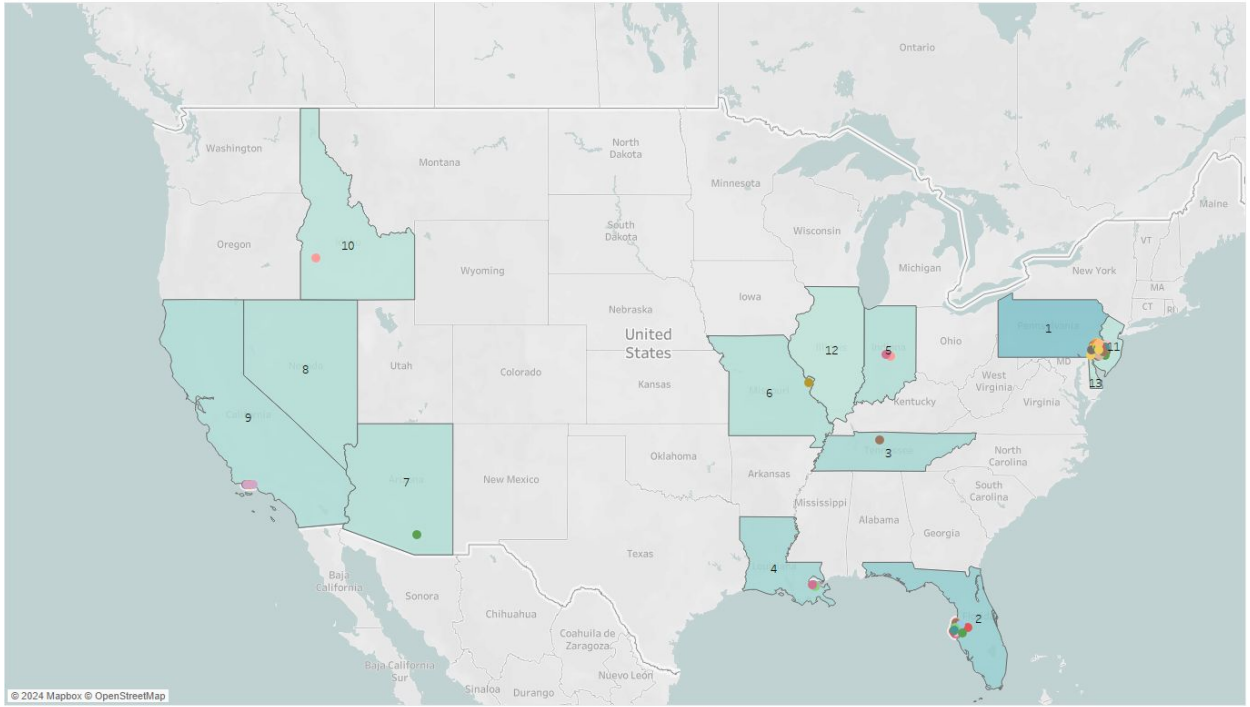
Most common hours of operation

- Expected some of these to be more common
- Many stores closed on low traffic days
- Reviews by month also shows more popular in the summer
- Recommendation: Be open later in the summer but stick to one of the first 2 times



Location Analysis

States Ranked by # of Coffee Shops



Map based on Longitude (generated) and Latitude (generated). For marks layer Tableau_coffee.State: Color shows sum of Review Count. The marks are labeled by Rank of Count of tableau_coffee. Details are shown for State. For marks layer Tableau_coffee.City: Color shows details about City. The view is filtered on City, which excludes Edmonton.

- Pennsylvania & Florida comprise a significant portion of the data
  - Specifically Philadelphia and Tampa Bay areas
- Data limited in a few select regions
- Theoretically if dataset contained entire population, recommendation would be LA due to limited stores

Modeling - Sentiment Analysis

- **Goal:** Create a model that will be able to effectively forecast reviewer's sentiments
  - Also, to sparse each review for individual words and analyze which words appear the most and seem to be the most impactful in shaping a review

- **Process:** To create the model the coffee business & review table had to be joined. Columns were also added to this new table splitting the data into test and training sets. Then each individual word was extracted in a new vocabulary table (ex bottom left)

- **Validation**: In order for the model to work, the data had to be binary - either considered a success or failure. **Success was measured as any review with a star rating greater than 3**

**Of note there are significantly more positive reviews, so false positives could be a potential bias*

Row	word	word_frequency	word_index
1	favorite	6725	107
2	sauce	2065	350
3	hours	1745	405
4	patio	1507	467
5	might	1340	519

Row	label	cnt
1	Positive	115344
2	Negative	12313

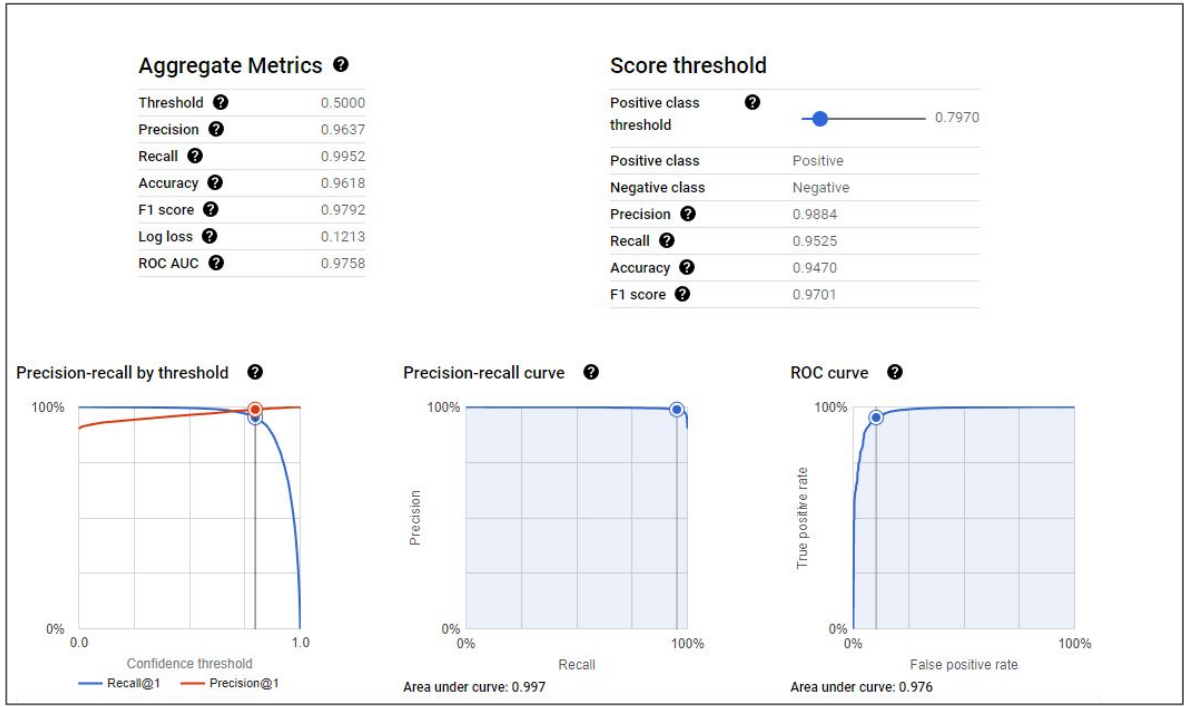
Sentiment Analysis Cont.

- We wanted to test if service/environment of a coffee shop seemed more important to customers than the actual quality of the food itself.
  - To analyze this, we took the vocabulary table we just created and queried it to see which category of word appeared more
    - We reject this hypothesis as clearly based on the numbers below
      - Still shows atmosphere & service is important in a location as they appear just 3x less than coffee itself

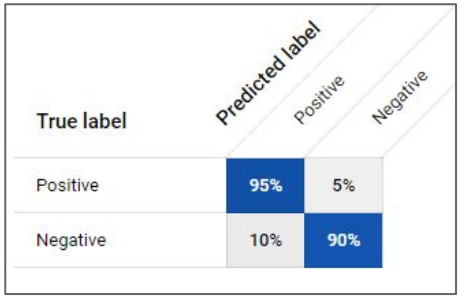
Row	word	word_frequency
1	food	27153
2	atmosphere	5751
3	service	14222
4	coffee	39033

Model Forecasting

Here you can see our Model Evaluation. The aggregate metrics are quite solid. A standard positive class threshold of 0.80 was chosen; meaning that anything above this threshold was classified as positive. This model should be pretty accurate for analysis



- Note the 10% false positive rate which we mentioned could occur earlier



Model Forecasting Cont.

- To input into the model predicted phrases and words for forecasting we attached the below query to the main:

```
WITH
  user_reviews AS (
      SELECT
        ROW_NUMBER()OVER() AS review_number,
        text,
        REGEXP_EXTRACT_ALL(LOWER(text), '[a-z]{2,}') AS words
      FROM (
            SELECT "Atmosphere"           AS text UNION ALL
            SELECT "Hot coffee"           AS text UNION ALL
            SELECT "Poor Service"         AS text UNION ALL
            SELECT "Hours"                AS text UNION ALL
            SELECT "Gross"                AS text UNION ALL
```

- The model then outputs it's prediction: positive or negative

Row	text	predicted_label
1	Atmosphere	Positive
2	Hot Coffee	Positive
3	Poor Service	Negative
4	Hours	Positive
5	Food	Positive
6	Gross	Negative

You can also mimic phrases and review segments

Row	text	predicted_label
1	The food was not good, but the coffee was great and the atmosphere was pretty good for getting work done	Positive

Recommendations & Actions

- **Have static hours for every day of the week from morning to afternoon**
- **Choose a location within the more populated states of the dataset**
  - **Preferably CA, as limited data and high population**
- **Place importance on appealing to a wide range of customers**
- **Recommend serving food with the coffee (most reviews in the dataset mentioned food)**
- **Place an importance on atmosphere and location as well**

Additional Hypotheses & Questions/Limitations

- Still some remaining questions on just how important or meaningful the reviews are to the success of the business.
- Unfortunately, unable to find any correlation. More data, such as sales data or other financials would have led to a much deeper root-cause analysis.
- Would of liked to see more complete data in some of the more populated states to get a better sense of population distribution
- Remaining questions on additional demographic data of users: gender, age, etc.
- Why do closed stores have similar star ratings to open stores?
- Best kind of atmosphere?