Project Proposal

Yelp Dataset Analysis

Nathan Cronin

I. Preparation of Proposal

A.) *Client/Dataset*

The client I chose to work with was *Coffee King*, a new startup coffee company. *Coffee King* is aiming to appeal to a wide variety of clientele and seeking outside consulting to achieve business aims. My focus is to work with the *Yelp Review Dataset* to assist the new startup. This dataset is very expansive with just under 7 million reviews on 150 thousand distinct businesses ranging in over 25 states. This expanse of useful data will give us plenty of recommendations to offer the company.

B.) Importing and Cleaning the Data

The data was in csv file format, with five files corresponding to each table. Due to the size of the CSV files, they were uploaded into Google Cloud using partitions. After this, a dataset was created in Google Big Query in tabular format to begin analysis.

The files were mostly cleaned to begin with, however some data types needed to be changed for ease of analysis, and further exploration revealed some additional cleaning steps as well. These have been listed below:

- Data Types: The categories, attributes, and hours columns within the business table were changed to JSON format, as there were multiple values for each row.
- Unnecessary data: Only one business found to be in the UK, likely due to error so this was deleted from table. I also deleted longitude and latitude columns from the business table as not helpful for analysis.
- Null values: All nulls were checked for and replaced with more meaningful text using COALESCE function.
- Redundancies: There were six individual columns for compliments a user could recieve on their review. I concatenated these into an individual compliment column which I found to be more helpful.

C.) ***Initial Exploration of Data & Summary Statistics***

I used an information schema query to grab the columns and data types for tables: business and review.

Row	column_name	data_type
1	business_id	STRING
2	name	STRING
3	address	STRING
4	city	STRING
5	state	STRING
6	postal_code	STRING
7	stars	FLOAT64
8	review_count	INT64
9	is_open	BOOL
10	attributes	JSON
11	categories	JSON
12	hours	JSON

Row	column_name	data_type
1	text	STRING
2	cool	INT64
3	stars	FLOAT64
4	date	TIMESTAMP
5	funny	INT64
6	review_id	STRING
7	useful	INT64
8	business_id	STRING
9	user_id	STRING

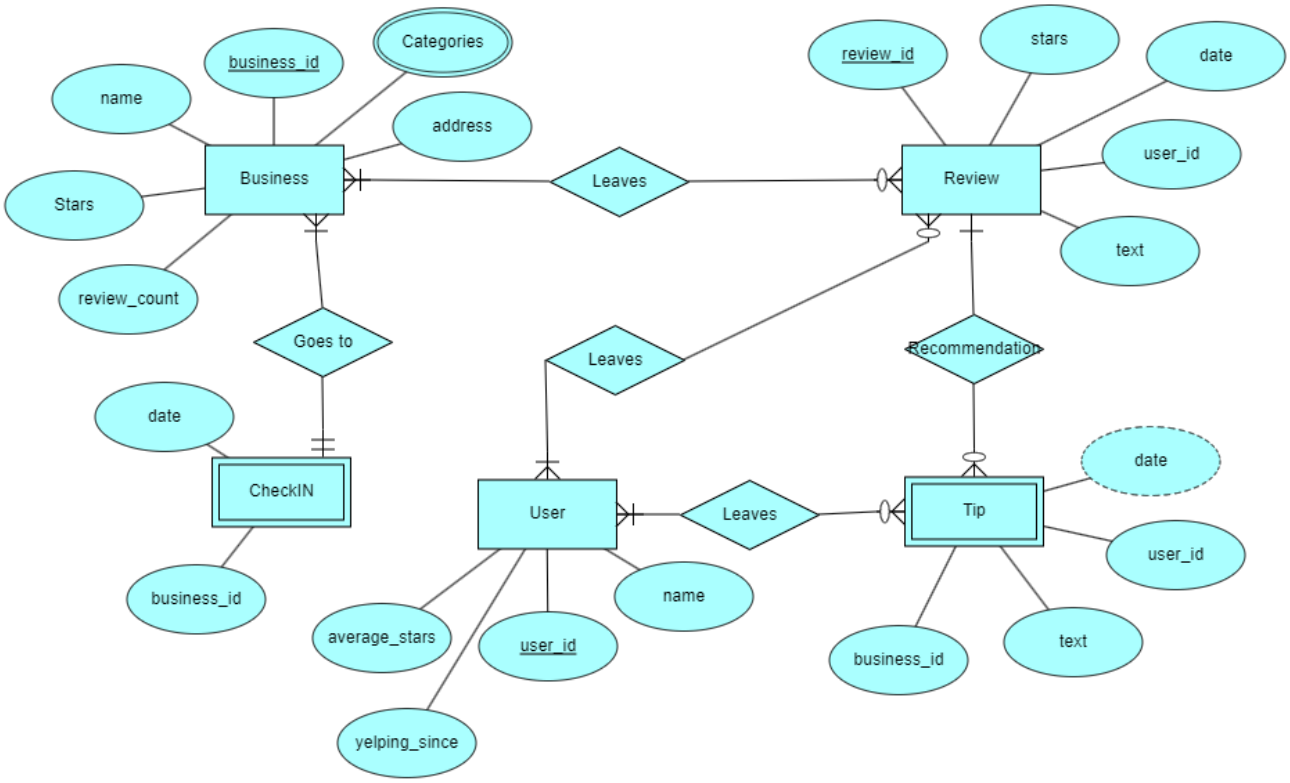
I also explored the reviews left by year to see the range of data available:

Row	min_year	max_year
1	2005	2022

Lastly, I wanted to take a look at the business count by state, as *Coffee king* noted that determining location was an influential factor for them. I limited results to the 10 most common. (AB is Alberta, Canada)

Row	state	business_count
1	PA	34039
2	FL	26330
3	TN	12056
4	IN	11247
5	MO	10913
6	LA	9924
7	AZ	9912
8	NJ	8536
9	NV	7715
10	AB	5573

D.) Entity Relationship Diagram



In summation, Users are able to leave reviews and/or tips on a business, while the Check In is used to monitor when customers arrive at the business.

Each table has their own unique primary key other than Checkin and tip. Tip is dependent on users and reviews while checkin is dependent on business_id as the foreign key. Of note, some columns were left off the ERD for page constraints.

II. Developing Proposal

A.) *Description*

This project aims to understand, analyze, and make inferences on the dataset available in order to create robust recommendations for *Coffee King.* All the data will be considered for sorting and analysis, however a focus will be placed on similar business types such as other coffee shops and restaurants. The audience for this project will be the decision-makers at *Coffee King* who will hopefully be receptive to the recommendations made to enhance the success of their budding business.

B.) Questions

Some questions that will be answered by the completion of this project are:

1. *What is the distribution of positive, neutral, and negative reviews for both businesses and users?*
2. *Does a higher review count for a business correlate with a higher or lower star rating?*
3. *What are some of the most influential factors of a user's experience that caused a review?*
4. *Does location of the business play any factor?*
5. *Is there any information detailed to see reasons as to why some businesses have closed?*

C.) Hypothesis

Along with these questions, I also came up with a few hypotheses regarding the data that I will be looking to prove or disprove in future analysis. My first hypothesis is that **businesses with a**

higher review count will have a higher star count on average. I based this on the fact that there is a higher count of reviews for businesses who are rated more highly. My second hypothesis is that **users with a very low review count (1-2) will generally have an extreme review.** I based this hypothesis on personal experience & behavioral analysis thinking that users are more prompted to review based on a very negative or positive experience. My last hypothesis is that **customer service is more influential in shaping the review than the quality of the product itself.**

D.) Approach

The approach to the analysis will be to group the “star rating” of businesses and users against other columns, such as State, Categories, review_count, and year. To assist with this, we will manipulate the star column to classify businesses as poor, average, and excellent. We will also do an in-depth analysis using descriptive and inferential statistics to get a sense of a baseline on how businesses perform. Lastly, a text sentiment analysis will be performed to determine what users value most when reviewing and visiting a business. The main focus will be to see what caused businesses to fail and others to flourish.