Learn SQL Basics for Data Science Specialization Capstone Project (Yelp Academic Dataset)

01 SEP 2024

LOH YONG MING RAYMOND

Content

- Summary
- Research Questions
- Initial Hypotheses (3 W's)
- Analytical Approach
- Discuss The Challenges
- Entity Relationship Diagram (ERD)
- Initial Findings
- In-Depth Analysis
- Hypotheses Results

Summary

CoffeeKing is a newly established pioneering startup in the coffee industry, dedicated to providing a wide range of distinctive coffee experiences that appeal to a diverse customer base. To shape their business strategy, I will analyze the Yelp dataset, which offers a vast collection of user-generated reviews, focusing on selecting optimal locations and determining operating hours.

Research Questions (3 W's)

- Why Coffee business and consider opening a new coffee shop?
- Where should CoffeeKing establish its new location?
- What are the most favorable opening and closing times for CoffeeKing?

Initial Hypothesis

- Category Dominance: The coffee and restaurant category may have a higher count of occurrences compared to others.
- Urban Advantage: Urban areas may exhibit higher review scores compared to rural regions.
- Peak Operating Hours: Coffee shops typically experience peak business during the morning hours.

Analytical Approach

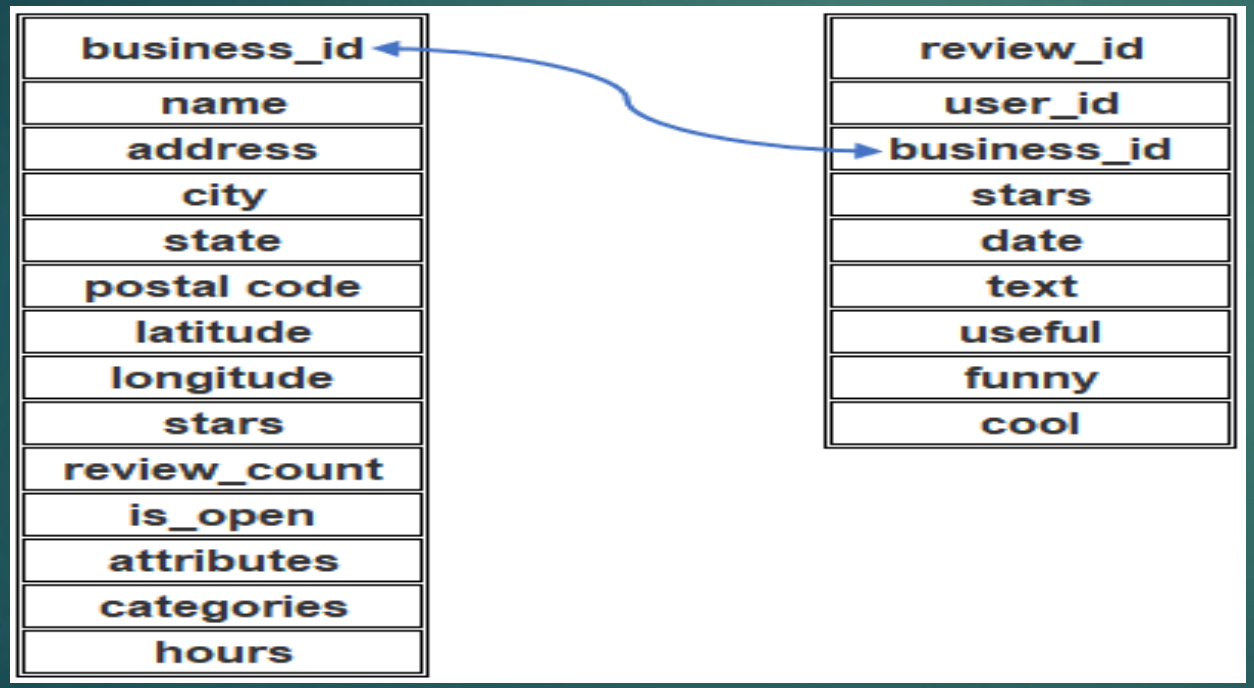
- Examine review counts, star ratings, location data, and review content to identify patterns and trends.
- Investigate potential relationships between operating hours and customer satisfaction as measured by star ratings.
- Analysis the word counts reviews to understand the context of different ratings.

Discuss The Challenges

- Due to the massive size of the JSON data, I'm facing frequent system hang and required to repeat the cycle of reimports of the python packages into Jupyter Notebook, and each loading process takes a long time.
- Even though this is a learn basics SQL data science specialization course, it requires Python knowledge, which I don't have. However, I was able to overcome this by doing my own research through self learning on system setup, Python coding, package installation & Import.

Entity Relationship Diagram

Below proposed ERD to show the relationships of data that are exploring, between business.json and review.json.



Initial Findings – Category Dominance

The query results indicate that the potential for a successful coffee shop business is high, based on the highest counts and frequent.

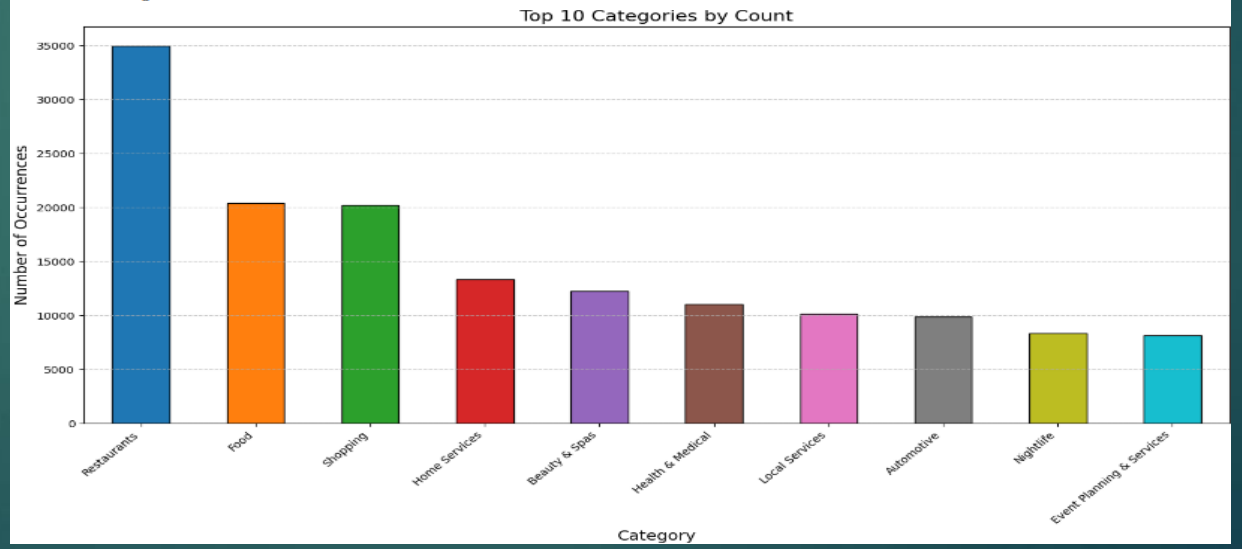
categories	
Coffee & Tea	4954
Coffee Roasteries	238
Coffeeshops	6
Coffee & Tea Supplies	5

Name: count, dtype: int64

	categories
3	Restaurants, Food, Bubble Tea, Coffee & Tea, B...
46	Arts & Entertainment, Music Venues, Internet S...
82	Restaurants, Automotive, Delis, Gas Stations, ...
85	Food, Restaurants, Salad, Coffee & Tea, Breakf...
89	Ice Cream & Frozen Yogurt, Coffee & Tea, Resta...

Top 10 Categories:

	Count	count
0	Restaurants	34987
1	Food	20419
2	Shopping	20186
3	Home Services	13322
4	Beauty & Spas	12263
5	Health & Medical	11046
6	Local Services	10138
7	Automotive	9879
8	Nightlife	8379
9	Event Planning & Services	8173



Initial Findings - Urban Advantage

The query results below indicate that urban cities have a higher number of reviews compared to rural cities.

	city	avg_rating	total_review
0	Philadelphia	4.050741	63006
1	New Orleans	4.075420	36091
2	Nashville	3.923305	24252
3	Tampa	3.910321	23584
4	Tucson	3.857586	18685
...	...	...	...
486	Chester	4.000000	5
487	Cumberland	4.000000	5
488	Eastampton Township	4.200000	5
489	Mehlville	4.400000	5
490	Kimmswick	5.000000	5

	business_id	name	address	city	state	postal_code	latitude	longitude	stars	review_count	is_open	attributes	categories
0	MTSW4McQd7CbVtyjqoe9mw	St Honore Pastries	935 Race St	Philadelphia	PA	19107	39.955505	-75.155564	4.0	80	1	{"RestaurantsDelivery": "False", "OutdoorSeati...	Restaurants, Food, Bubble Tea, Coffee & Tea, B...
1	MTSW4McQd7CbVtyjqoe9mw	St Honore Pastries	935 Race St	Philadelphia	PA	19107	39.955505	-75.155564	4.0	80	1	{"RestaurantsDelivery": "False", "OutdoorSeati...	Restaurants, Food, Bubble Tea, Coffee & Tea, B...
2	MTSW4McQd7CbVtyjqoe9mw	St Honore Pastries	935 Race St	Philadelphia	PA	19107	39.955505	-75.155564	4.0	80	1	{"RestaurantsDelivery": "False", "OutdoorSeati...	Restaurants, Food, Bubble Tea, Coffee & Tea, B...
3	MTSW4McQd7CbVtyjqoe9mw	St Honore Pastries	935 Race St	Philadelphia	PA	19107	39.955505	-75.155564	4.0	80	1	{"RestaurantsDelivery": "False", "OutdoorSeati...	Restaurants, Food, Bubble Tea, Coffee & Tea, B...
4	MTSW4McQd7CbVtyjqoe9mw	St Honore Pastries	935 Race St	Philadelphia	PA	19107	39.955505	-75.155564	4.0	80	1	{"RestaurantsDelivery": "False", "OutdoorSeati...	Restaurants, Food, Bubble Tea, Coffee & Tea, B...

Initial Findings - Peak Operating Hours

The query results below suggest that the most popular and in-demand operating hours are from 7 AM to 6 PM, with peak activity occurring from morning to afternoon.

	hours	total_review
0	{"Monday": "0:0-0:0", "Tuesday": "0:0-0:0", "Wednesday": "0:0-0:0", "Thursday": "0:0-0:0", "Friday": "0:0-0:0", "Saturday": "0:0-0:0", "Sunday": "0:0-0:0"}	18048
1	{"Monday": "8:0-18:0", "Tuesday": "8:0-18:0", "Wednesday": "8:0-18:0", "Thursday": "8:0-18:0", "Friday": "8:0-18:0", "Saturday": "8:0-18:0", "Sunday": "8:0-18:0"}	6647
2	{"Monday": "7:0-14:0", "Tuesday": "7:0-14:0", "Wednesday": "7:0-14:0", "Thursday": "7:0-14:0", "Friday": "7:0-14:0", "Saturday": "7:0-14:0", "Sunday": "7:0-14:0"}	4038
3	{"Monday": "7:0-15:0", "Tuesday": "7:0-15:0", "Wednesday": "7:0-15:0", "Thursday": "7:0-15:0", "Friday": "7:0-15:0", "Saturday": "7:0-15:0", "Sunday": "7:0-15:0"}	3676
4	{"Monday": "8:0-14:0", "Tuesday": "8:0-14:0", "Wednesday": "8:0-14:0", "Thursday": "8:0-14:0", "Friday": "8:0-14:0", "Saturday": "8:0-14:0", "Sunday": "8:0-14:0"}	3225

	hours	avg_rating	review_count
0	{"Monday": "8:0-18:0", "Tuesday": "8:0-18:0", "Wednesday": "8:0-18:0", "Thursday": "8:0-18:0", "Friday": "8:0-18:0", "Saturday": "8:0-18:0", "Sunday": "8:0-18:0"}	4.440349	33318333
1	{"Monday": "0:0-0:0", "Tuesday": "7:0-15:0", "Wednesday": "7:0-15:0", "Thursday": "7:0-15:0", "Friday": "7:0-17:0", "Saturday": "7:0-17:0", "Sunday": "7:0-15:0"}	4.000000	7316736
2	{"Monday": "7:0-21:0", "Tuesday": "7:0-21:0", "Wednesday": "7:0-21:0", "Thursday": "7:0-21:0", "Friday": "7:0-21:0", "Saturday": "7:0-21:0", "Sunday": "7:0-21:0"}	4.172897	4480866
3	{"Monday": "12:0-0:0", "Tuesday": "11:0-23:45", "Wednesday": "12:0-0:0", "Thursday": "12:0-0:0", "Friday": "12:0-0:0", "Saturday": "12:0-0:0", "Sunday": "12:0-0:0"}	4.000000	4330200
4	{"Monday": "8:0-14:0", "Tuesday": "8:0-14:0", "Wednesday": "8:0-14:0", "Thursday": "8:0-14:0", "Friday": "8:0-14:0", "Saturday": "8:0-14:0", "Sunday": "8:0-14:0"}	4.090698	4288555

In-Depth Analysis - Higher Rating By Word Frequencies

The word counts below represent reviews with a rating of 4 or higher. The larger the size of the words, the more frequently they appear throughout the entire review text.

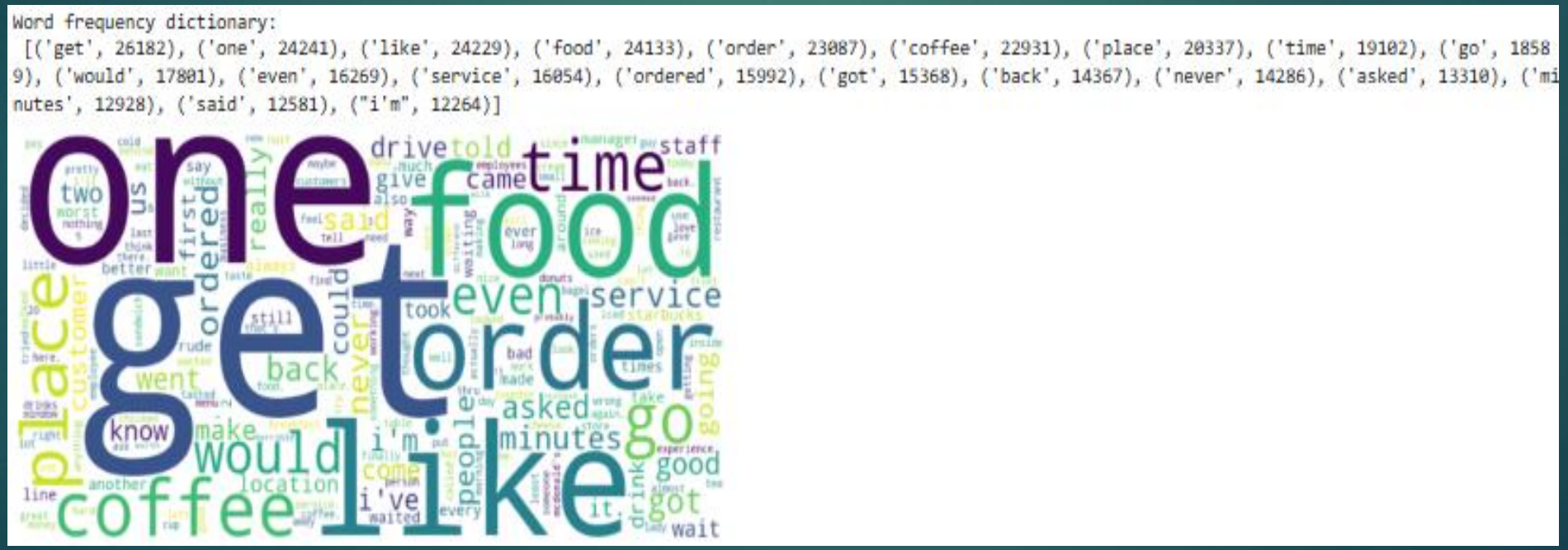
Word frequency dictionary:

[('coffee', 121954), ('place', 105005), ('great', 101577), ('good', 70606), ('food', 67076), ('like', 61529), ('get', 60292), ('one', 57203), ('love', 55491), ('also', 54445), ('really', 53008), ('go', 47277), ('always', 44603), ('best', 44343), ('time', 41320), ('little', 40848), ('definitely', 40544), ('would', 40319), ('back', 40206), ("i've", 39357)]

In-Depth Analysis - Lower Rating By Word Frequencies

The word counts below represent reviews with a rating of 3 or lower. The smaller the size of the words, the less frequently they appear in the overall review text.

Word frequency dictionary:
[('get', 26182), ('one', 24241), ('like', 24229), ('food', 24133), ('order', 23087), ('coffee', 22931), ('place', 20337), ('time', 19102), ('go', 1858
9), ('would', 17801), ('even', 16269), ('service', 16054), ('ordered', 15992), ('got', 15368), ('back', 14367), ('never', 14286), ('asked', 13310), ('mi
nutes', 12928), ('said', 12581), ("i'm", 12264)]



Hypotheses Results (3'Ws)

Why Coffee business and consider opening a new coffee shop?
The hypothesis is that growing consumer demand and favorable market trends make the coffee business a promising opportunity.

Where should CoffeeKing establish its new location?
The hypothesis is that a high-traffic, strategically located site with favorable demographics will boost the shop's success.

What are the most favorable opening and closing times for CoffeeKing?
The hypothesis is that peak hours, particularly in the morning and early afternoon, will optimize customer visits and revenue.

Hypotheses Results (Conclude)

Based on the query results from initial findings and in-depth analysis, the hypothesis is that the potential for establishing a successful coffee shop business is substantial. This conclusion is drawn from the observation that coffee shops with high review counts and frequent mentions in the data are associated with greater consumer interest and engagement. The high volume of positive feedback and frequent appearances in review texts suggest that these coffee shops are well-regarded and attract significant customer traffic, indicating a promising opportunity for new ventures in this market segment.

Thank you

Loh Yong Ming Raymond

- rlohym@gmail.com