OSEMN Report
Final Project

Exploring Yelp Reviews Data: Insights and Visualizations using OSEMN Process

Rahul Kumar Nalubandhu

Introduction

Conducting a review analysis presents businesses with the opportunity to improve customer experience, identify service gaps, and gather real-time insights, along with numerous other advantages. This analysis dives deep into customer reviews posted on Yelp for businesses located in diverse states across the United States. By investigating three main questions in this project, we seek to gain a more comprehensive understanding of the restaurant landscape and uncover trends that impact both customers and businesses:

1) Can you identify the top 50 cities with the highest average star ratings for restaurants that offer both delivery and takeout services, and have a minimum of 50 reviews, while also analyzing the proportion of these highly rated restaurants that accept card payments compared to the overall number of restaurants in the same city?

2) What are the top three restaurants in Pennsylvania, Florida, Indiana, Tennessee, and Missouri? Rank the top 3 restaurants by review_count and only consider those with a star rating greater than or equal to 4. Are there any similarities in the business attributes for the top 3 restaurants in each state?

3) Is there a significant difference in the average ratings of businesses that offer delivery services only, takeout services only, both delivery and takeout services, and neither delivery nor takeout services?

Problem

The problem of having few or no customer evaluations might affect profits because customer reviews are an essential marketing tool for drawing in new clients and boosting revenue. Understanding which states and businesses experience the lowest levels of customer engagement can be learned by analyzing evaluations at the business, local, and state levels. With focused interventions, this analysis can increase customer interaction, better the customer experience, and eventually improve corporate performance.

Obtain Data

Yelp is a one-stop platform which enables customers to connect with businesses. More than 80 million people visit this platform in a month to find businesses and service providers. Customers are given the ability to leave reviews and request quotes from local businesses amongst many other things. In return, local business owners are given the ability to communicate with their customers and respond to reviews to build trust with their customers. The customer review data set is acquired directly through Yelp. The data set is 4.04GB (1 point) and split into multiple JSON files (2 points) which contain businesses, reviews, and user data. In addition, the data has punctuation (1 point) and has more than one type of related data (2 points). Based on the point system requirements provided, the yelp data is a 6-point data set.

OSEMN Report
Final Project

Scrub Data

The Yelp dataset was obtained from the Yelp website and then uploaded to a Google Cloud Storage bucket before analysis. After being imported from the cloud storage, the data completed the necessary transformations to remove any erroneous or pointless information. Only 3,592 rows of meaningful data were left in the dataset after filtering out the first 13,252 rows. The company name, ID, city, state, review count, stars, RestaurantsDelivery, RestaurantsTakeout, and BusinessAcceptsCards were among the fields that were taken into consideration. For the 2nd question states with the most data in the scrubbed dataset are Pennsylvania, Florida, Indiana, Tennessee, and Missouri.

The Python script was changed to extract the relevant fields and build a schema with the proper data types for each field after creating a table in BigQuery. Next, using this table, visualizations were created, and useful values were obtained. To protect the integrity of the analysis and ensure that the findings drawn were correct and reliable, all null values were also eliminated from the dataset.

Explore Data

For the 1st question, the exploration stage, the data is loaded into CloudStorage. To begin the exploration stage the data is then extracted using PySpark.

Analysis and visualizations for 1st question:

Question of interest?

Can you identify the top 50 cities with the highest average star ratings for restaurants that offer both delivery and takeout services, and have a minimum of 50 reviews, while also analyzing the proportion of these highly rated restaurants that accept card payments compared to the overall number of restaurants in the same city?

I wanted to find the top 50 cities with the highest average star ratings for eateries that provide both delivery and takeout and have at least 50 reviews. I also wanted to examine the percentage of these highly rated restaurants that accept card payments in relation to the total number of restaurants in the same city.

I took the following actions to achieve this:

1) I made a dataset first, then I uploaded it to cloud storage.
2) After that, I started a Dataprep data flow by considering the Yelp academic business dataset. Then, using recipes, significant categories such as business id, name, address, city, state, postal code, stars, review count, restaurantsDelivery, restaurantsTakeout, and businessAcceptsCards were chosen. I eliminated all null values and only considered US-based companies.
3) After that, the task was executed, and a table was created in BigQuery.
4) The appropriate fields were then extracted from the table using a modified Python script, which was used to produce a schema with the appropriate data types for each field.

OSEMN Report
Final Project

5) I completed the task on the cluster and carried out an analysis to determine the top 50 cities with the greatest average star ratings for eateries that provide delivery and takeout and have a minimum of 50 reviews. Also, the percentage of these highly rated restaurants that accept credit cards in relation to the total number of eateries in the same city was examined.
6) I computed the percentage of highly rated restaurants that take card payments to give further information about the accessibility and convenience of these top restaurants in each city.
7) I also included the code to plot graphs in the same Python code.

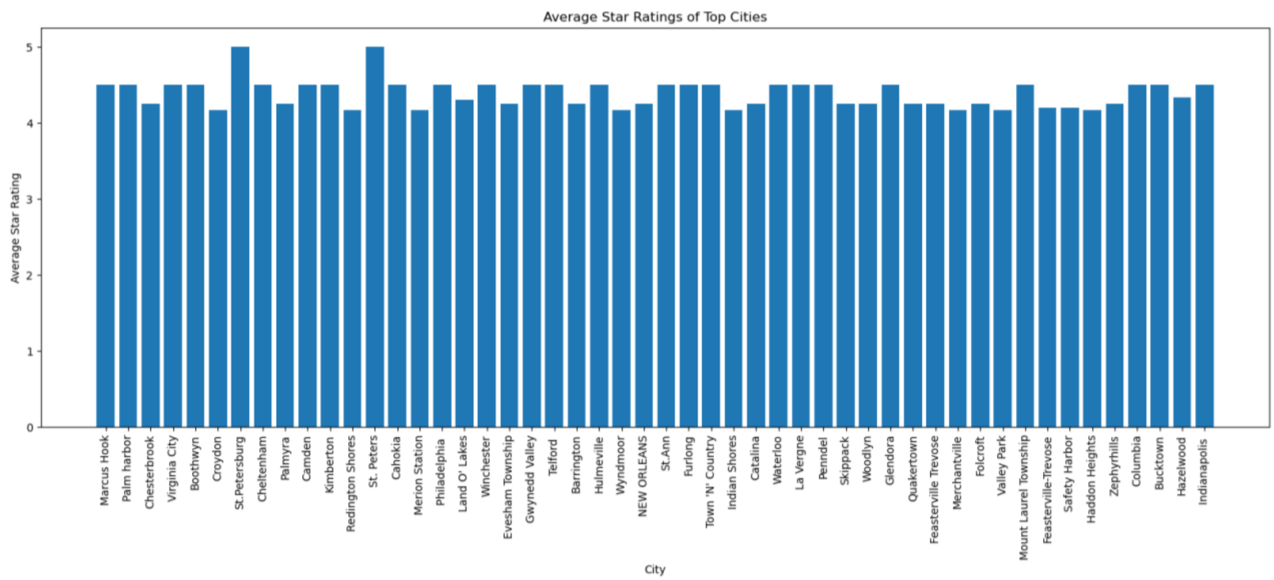
Why considers the proportion?

By examining the proportion of card-accepting restaurants, the analysis offers a better understanding of how payment-friendly the top restaurants are in the top 50 cities with the highest average star ratings. This information can be helpful for potential customers, business owners, and investors to determine the prevalence of card acceptance among the best-rated restaurants in a city.

So, calculating the proportion of highly rated restaurants that accept card payments adds another dimension to the analysis, providing a more comprehensive view of the restaurant landscape in the top cities.

Visualizations:

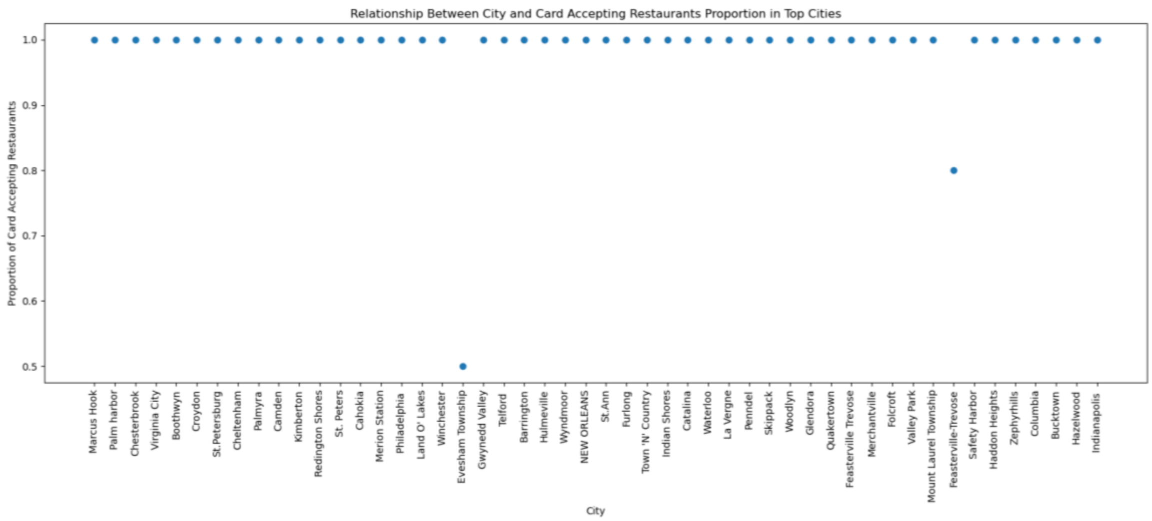
1) Top Cities' Average Star Ratings: This graph shows the average star ratings for the top 50 cities with the highest ratings for restaurants that provide both delivery and takeout and have at least 50 reviews. By looking at this graph, we may learn which cities have the best restaurants based on the given standards. As a result, St. petersbug and St. Peter's have the best reviews.



plot 1

2) Proportion of Card Accepting Restaurants in Top Cities:

   In each of the top 50 cities listed in the first graph, this graph displays the percentage of highly rated restaurants that accept credit cards. This reveals the frequency with which the top-rated restaurants in those cities accept cards. It can assist prospective customers, business owners, and investors in comprehending how simple and easy it is to make a payment at these top restaurants.



Plot 2

The proportion of card-accepting restaurants is generally high in these top cities, which suggests that most restaurants in these locations have adopted modern payment methods and provide convenience to their customers. The card-accepting proportion varies across cities, with some cities having all their restaurants accepting cards, while others have a slightly lower proportion.

The correlation coefficient between the average star ratings and the proportion of card-accepting restaurants is 0.15147935326504264. This value indicates a weak positive relationship between the two variables. It suggests that, in general, cities with higher average star ratings tend to have a slightly higher proportion of card-accepting restaurants.

```
Correlation coefficient: 0.15147935326504264
File average_star_ratings.png uploaded to images/average_star_ratings.png.
File card_accepting_proportion.png uploaded to images/card_accepting_proportion.png.
File ratings_distribution.png uploaded to images/ratings_distribution.png.
File ratings_distribution_both.png uploaded to images/ratings_distribution_both.png.
23/03/17 21:20:25 INFO org.sparkproject.jetty.server.AbstractConnector: Stopped Spark@3e83a
```

Fig1

In conclusion, the top cities with the highest average star ratings generally have a high proportion of card-accepting restaurants, providing convenience to customers. Although there is a weak positive correlation between the average star ratings and the proportion of card-accepting restaurants, other factors may have a more substantial impact on the average star ratings in these cities.

Analysis and visualizations for 2nd question:

Question of interest?

What are the top three restaurants in Pennsylvania, Florida, Indiana, Tennessee, and Missouri? Rank the top 3 restaurants by review_count and only consider those with a star rating greater than or equal to 4. Are there any similarities in the business hours that they operate or business attributes?

The states with the most data in the scrubbed dataset are Pennsylvania, Florida, Indiana, Tennessee, and Missouri. For this question, I was interested in seeing the top 3 restaurants in each state and how they compare to one another. The top 3 restaurants are identified by the following criteria: a rating greater than or equal to four, and review count which will determine the restaurant's rank within each state. The best way to accomplish this was by creating a SQL query in BigQuery. The SQL query starts off with a Common Table Expression (CTE) for the restaurant ranking.

Why we do Common Table Expression (CTE) ?

The powerful SQL construct known as the common table expression (CTE) aids in query simplification. CTEs function as virtual tables (with records and columns) that are created while a query is being run, used by the query, and then deleted once the query has finished running. below is the SQL Query used in BigQuery.

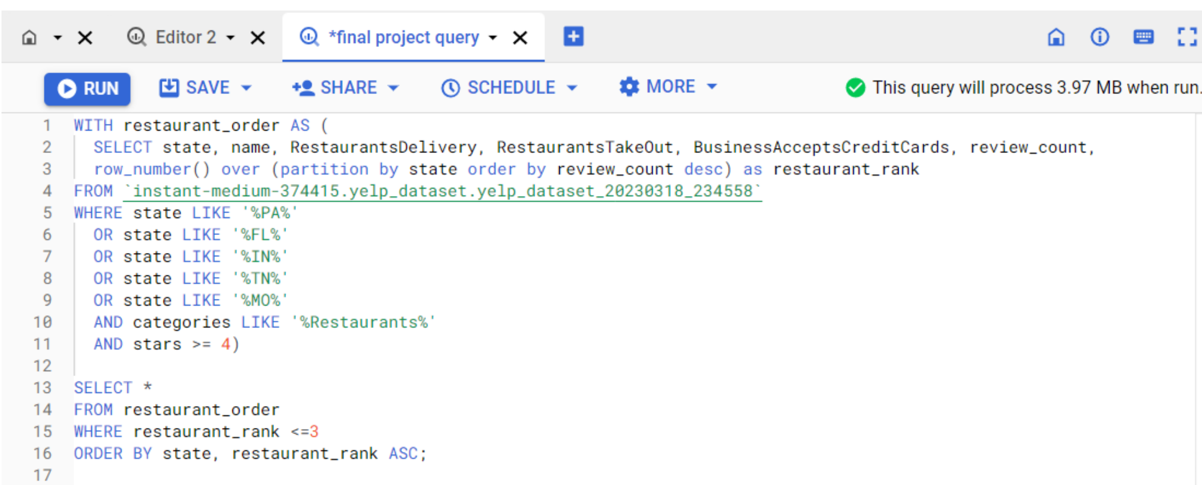


Fig2

To achieve this query, I followed these steps:

Step 1: Initialize a Common Table Expression (CTE)

- Set up a CTE to temporarily store data, including the new restaurant rank column.
- Choose the relevant columns from the original table.

Step 2: Generate restaurant ranks.

- Utilize the row_number() function along with the OVER clause to assign a unique rank to each restaurant in the dataset.

Step 3: Group data by state

- Incorporate "partition by state" in the query to ensure restaurant ranks are allocated within their respective states.

Step 4: Sort restaurants by review count

- Arrange the restaurants by their review count in descending order, which will rank them based on the number of reviews from highest to lowest.

Step 5: Apply filters for state and star rating.

- Include a WHERE clause in the CTE to focus on specific states and only consider restaurants with a star rating of four or higher.

Step 6: Retrieve top 3 restaurants in each state.

- Formulate a SQL query that selects all columns from the CTE table while limiting the output to the top 3 restaurants in each state based on their ranks.

Step 7: Organize the output by state and rank.

- Sort the results by state and restaurant rank in ascending order to facilitate easy interpretation and comparison.

Step 8: Export results to a CSV file

- Execute the query and save the outcome in a CSV file for subsequent analysis, reporting, or presentation.

The table below shows the top 3 restaurants in each state with their rank.

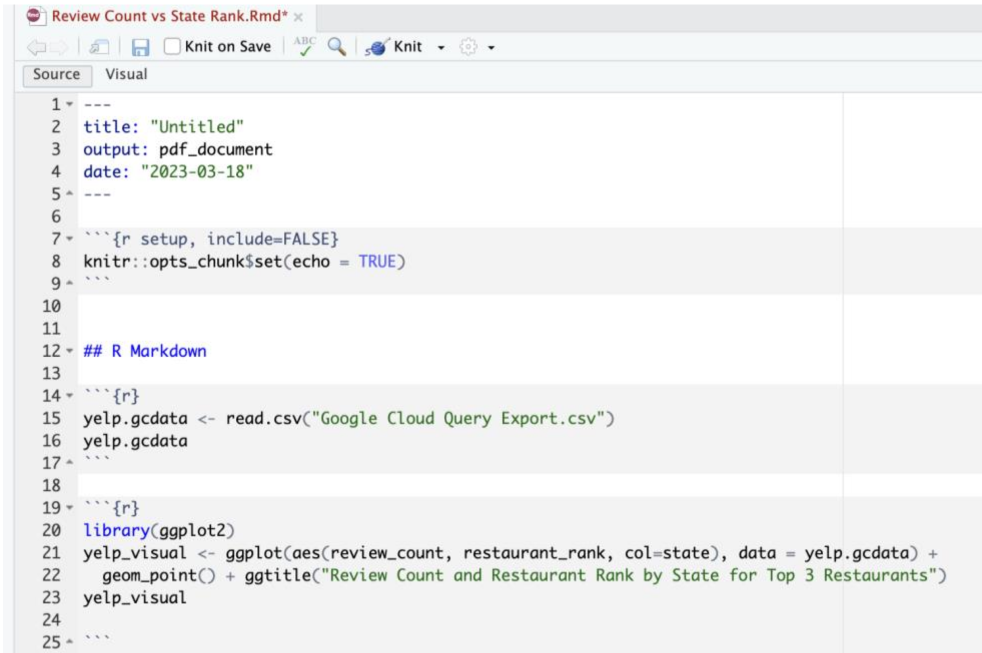
State	Name	Restaurants Delivery	Restaurants Take Out	Business Accepts Credit Cards	Review Count	Restaurant Rank
FL	Datz	TRUE	TRUE	TRUE	3260	1
FL	Bern's Steak House	FALSE	FALSE	TRUE	2924	2
FL	Oxford Exchange	TRUE	TRUE	TRUE	1868	3
IN	The Eagle	TRUE	TRUE	TRUE	2233	1
IN	St. Elmo Steak House	TRUE	TRUE	TRUE	2035	2
IN	Bakersfield	TRUE	TRUE	TRUE	1642	3
MO	Pappy's Smokehouse	TRUE	TRUE	TRUE	3999	1
MO	Broadway Oyster Bar	TRUE	TRUE	TRUE	2076	2
MO	Rooster - Downtown	TRUE	TRUE	TRUE	1984	3
PA	Reading Terminal Market	TRUE	TRUE	TRUE	5721	1
PA	Pat's King of Steaks	TRUE	TRUE	FALSE	4250	2
PA	Geno's Steaks	TRUE	TRUE	FALSE	3401	3
TN	Hattie Bâ€™s Hot Chicken - Nashville	TRUE	TRUE	TRUE	6093	1
TN	Biscuit Love: Gulch	TRUE	TRUE	TRUE	4207	2
TN	The Pharmacy	TRUE	TRUE	TRUE	3054	3

Table1

OSEMN Report
Final Project

Visualization:

To generate the plot, we exported the query output as .csv and then used r-studio to generate the plot.




```
---
title: "Untitled"
output: pdf_document
date: "2023-03-18"
---

```{r setup, include=FALSE}
knitr::opts_chunk$set(echo = TRUE)
```

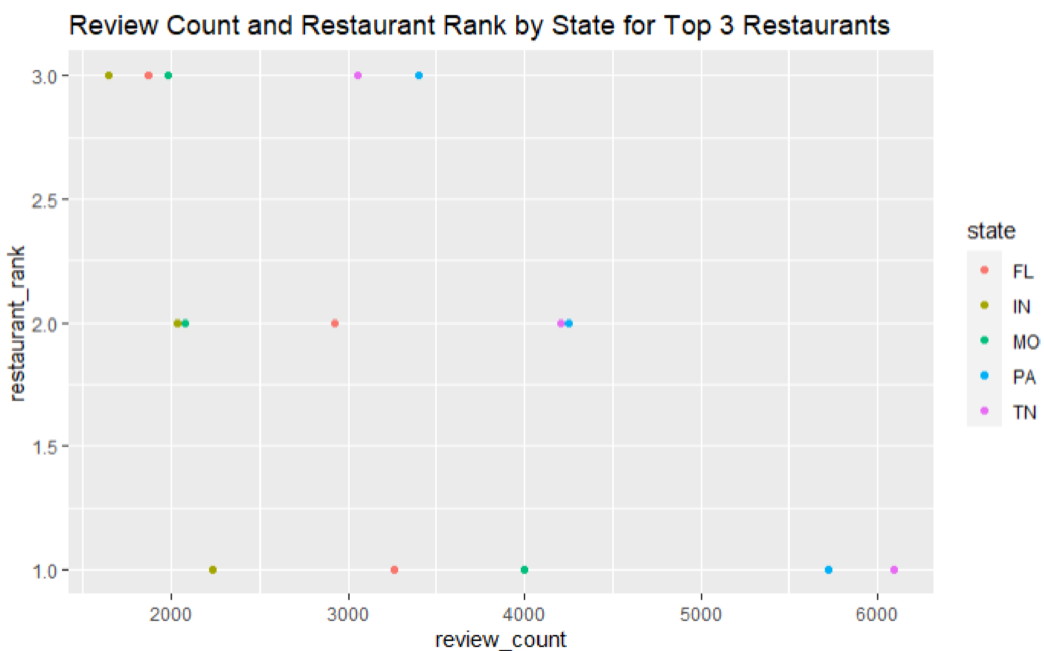

R Markdown

```{r}
yelp.gcdata <- read.csv("Google Cloud Query Export.csv")
yelp.gcdata
```

```{r}
library(ggplot2)
yelp_visual <- ggplot(aes(review_count, restaurant_rank, col=state), data = yelp.gcdata) +
  geom_point() + ggtitle("Review Count and Restaurant Rank by State for Top 3 Restaurants")
yelp_visual

```
```

The scatterplot displayed below has also been created to assess how the top ranked restaurants compare in each state based on review count and their restaurant rank.



plot 3

OSEMN Report
Final Project

Findings from the plot:

When compared to Florida, Indiana, and Missouri, Pennsylvania and Tennessee have the highest number of ratings for their top 3 restaurants, according to the scatterplot. When compared to the other states, Indiana has the fewest reviews for its top 3 restaurants. The query table reveals that the top-ranked restaurant in each state provides all of the characteristics we were interested in examining: delivery, takeout, and credit card acceptance. Finally, the top 3 restaurants in each state offer all three traits except for Florida and Pennsylvania, where only two of the three attributes are offered.

As a result, the examination of the top three restaurants in Pennsylvania, Florida, Indiana, Tennessee, and Missouri emphasizes the significance of customer interaction, as shown by the number of reviews, and the application of services that take into account client preferences. Except for a few establishments in Florida and Pennsylvania, the top-rated restaurants in each state typically provide a combination of delivery, takeout, and credit card acceptance. Businesses aiming to improve client happiness and their offers may find this information to be helpful.

Analysis and visualizations for 3rd question:

Question of Interest ?

Is there a significant difference in the average ratings of businesses that offer delivery services only, takeout services only, both delivery and takeout services, and neither delivery nor takeout services?

To answer this question, you would need to perform a hypothesis test to determine whether the difference in average ratings between the two groups is statistically significant. One way to do this is to use a Anova test. So here I found ANOVA test for 4 categories one is for average between delivery and takeout and next is for business that accepts both delivery and takeout and not accept any services.

For Data extraction and preprocessing we take the initial JSON data, and extract the following fields: stars (average rating), from attributes I took RestaurantsDelivery (whether the business offers delivery service) and RestaurantsTakeout (whether the business offers takeout service)

Now using this data I used PySpark, to filter and aggregate the data by businesses that offer delivery services and those that offer takeout services and business that offer both and those that offer none. Then calculated the average rating for each group.

Given the updated analysis where we now consider four categories of businesses, it is more appropriate to use ANOVA (Analysis of Variance) instead of the two-sample t-test. ANOVA allows us to simultaneously test the differences in average ratings among more than two groups, which is not possible with a two-sample t-test.

We perform the one-way ANOVA test with the following hypotheses:

- Null hypothesis (H0): There is no significant difference in the average ratings among the four categories of businesses.
- Alternative hypothesis (H1): There is a significant difference in the average ratings among at least two of the four categories.

Why are we using ANOVA ?

The two-sample t-test is only capable of comparing two groups at a time, which would require us to perform multiple t-tests to compare all four categories. However, performing multiple t-tests increases the risk of Type I errors, also known as false positives. This occurs when we mistakenly reject a true null hypothesis. To maintain a consistent level of statistical significance, we need a better approach.

One-way ANOVA comes to our rescue as it is specifically designed to handle the comparison of more than two groups. ANOVA allows us to test the null hypothesis that there is no significant difference among the group means simultaneously. This maintains the overall statistical significance level and reduces the risk of Type I errors.

Additionally, ANOVA is more efficient than conducting multiple t-tests because it only requires a single test to evaluate the differences among all groups. It achieves this by comparing the variance between the groups to the variance within the groups. By doing so, it generates an F-statistic, which measures the relative variation between the groups.

Conclusions from test:

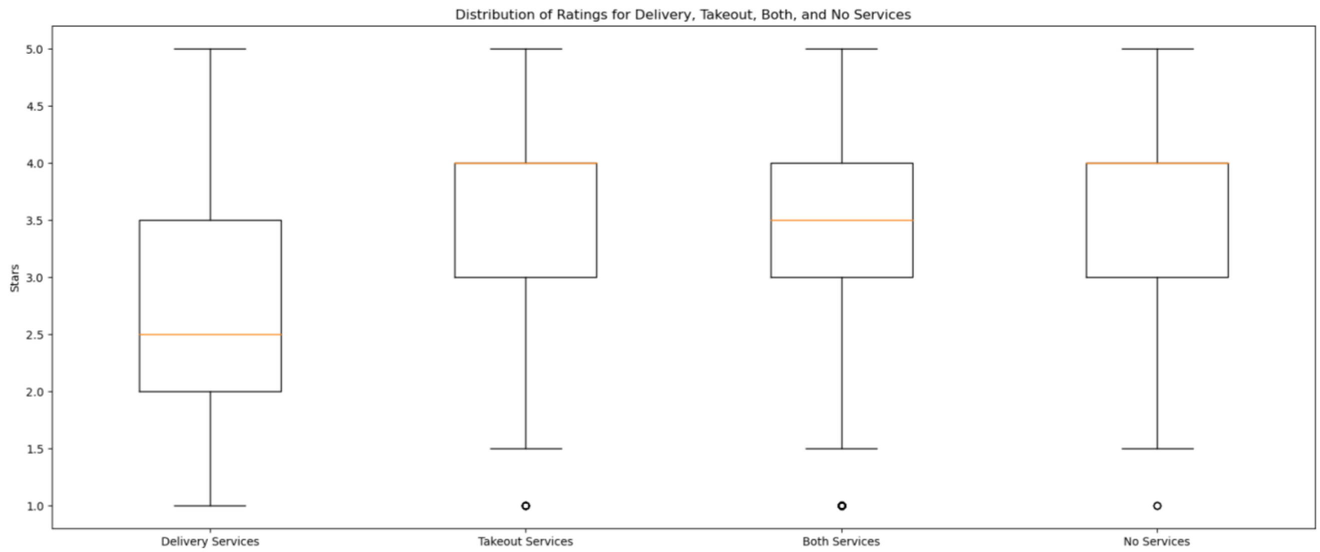
After conducting the ANOVA test, we obtained an F-statistic of 597.89 and a p-value of 0.0. The F-statistic value helps us understand the extent of the differences between the groups. A higher F-statistic value indicates a more significant difference between the group means.

Since the p-value is less than 0.05, we can reject the null hypothesis, concluding that there is a significant difference in the average ratings among at least two of the four categories of businesses (delivery services, takeout services, both delivery and takeout services, and no services).

Visualization:

Also, I generated box plots for delivery services only, takeout services only, both services, and neither service using the matplotlib library in Python.

The reason for generating box plots for these four groups is to help you visualize and compare the distribution of star ratings for each group. It helps to identify any differences in central tendency, spread, and the presence of outliers among the groups, which can be useful when interpreting the results of your statistical analysis.



plot 4

Summary of the boxplot:

Summary statistics for delivery_services:

summary	stars
count	1372
mean	2.777332361516035
stddev	1.0387197684817724
min	1.0
max	5.0

Summary statistics for takeout_services:

summary	stars
count	13259
mean	3.6687532996455237
stddev	0.7224428737562917
min	1.0
max	5.0

Summary statistics for both_services:

summary	stars
count	24457
mean	3.449728094206158
stddev	0.860909627361036
min	1.0
max	5.0

Summary statistics for no_services:

summary	stars
count	1702
mean	3.6818448883666277
stddev	0.759021387586643
min	1.0
max	5.0

Based on the box plot descriptions and summary statistics for delivery services, takeout services, both services, and no services, we can conclude the following:

Delivery services: Businesses offering only delivery services have a median rating of 2.5, with a lower quartile of 2.0 and an upper quartile of 3.5. This indicates that the middle 50% of ratings for delivery services lie between 2.0 and 3.5, suggesting that customer satisfaction is generally lower for businesses offering only delivery services.

Takeout services: Businesses offering only takeout services have a higher median rating of 4.0, with a lower quartile of 3.0 and an upper quartile of 4.0. This indicates that customer satisfaction levels are generally higher for businesses offering only takeout services compared to those offering only delivery services.

Both services: Businesses offering both delivery and takeout services have a median rating of 3.5, which is higher than delivery services but lower than takeout services. This suggests that offering both services may have a positive impact on customer satisfaction but not to the same extent as offering takeout services alone.

No services: Businesses offering neither delivery nor takeout services have a median rating of 4.0, which is on par with takeout services. This indicates that businesses in this category generally have good customer satisfaction levels, like those offering only takeout services.

In conclusion, the box plot analysis and summary statistics reveal that businesses offering only takeout services and businesses offering neither delivery nor takeout services generally have higher customer satisfaction levels, with median ratings of 4.0. Businesses offering both delivery and takeout services have a slightly lower median rating of 3.5, while businesses offering only delivery services have the lowest median rating of 2.5. This suggests that businesses may benefit more from focusing on takeout services or offering a combination of delivery and takeout services to improve customer satisfaction.

Model Data

After reviewing the dataset, a decision tree is the most suitable modeling technique for this analysis. A decision tree will contain the different factors that determine if businesses are likely to face less customer reviews due to the state they are in or even the demographics of their target customers. While this is deemed the best approach for this analysis, more data needs to be gathered and analyzed to properly complete this. Also calculated summary statistics for each group (delivery services, takeout services, both services, and no services) and compared their mean star ratings. This analysis provided insights into the average star ratings for different service categories.

Interpret Data

From the first analysis, we can interpret In this study, that we identified the top 50 cities with the highest average star ratings for restaurants offering both delivery and takeout services and having a minimum of 50 reviews. Our findings showed:

1) These top cities have a strong presence of high-quality restaurants, making them attractive for food enthusiasts.

2) The proportion of card-accepting restaurants is generally high in these cities, indicating widespread adoption of modern payment methods and convenience for customers. However, the proportion varies across cities.
3) The correlation coefficient between average star ratings and the proportion of card-accepting restaurants is 0.15147935326504264, indicating a weak positive relationship. Cities with higher average star ratings tend to have a slightly higher proportion of card-accepting restaurants, but other factors may contribute more significantly to a city's average star rating.

For the 2nd analysis, The analysis of the top 3 restaurants in Pennsylvania, Florida, Indiana, Tennessee, and Missouri reveals the following key insights:

1. Restaurants with a star rating of 4 or higher and a high review count indicate popularity and customer engagement.
2. Pennsylvania and Tennessee have the highest number of reviews for their top 3 restaurants, while Indiana has the lowest.
3. The #1 ranked restaurant in each state offers delivery, takeout, and accepts credit cards, adapting to customer preferences.
4. Florida and Pennsylvania have some top 3 restaurants that do not provide all three attributes, unlike Indiana, Tennessee, and Missouri.

In conclusion, customer engagement and catering to customer preferences are crucial for businesses to improve satisfaction. The top-ranked restaurants generally offer delivery, takeout, and credit card acceptance, with some exceptions in Florida and Pennsylvania.

In the 3rd analysis, we can interpret the following: After conducting the ANOVA test, the p-value is used to assess the significance of the results. If the p-value is less than the chosen significance level (e.g., 0.05), we can reject the null hypothesis and conclude that there is a significant difference in the average ratings among the four categories of businesses (delivery services, takeout services, both services, and no services). If the p-value is greater than the significance level, we cannot reject the null hypothesis, indicating there is no evidence to suggest a significant difference in average ratings between the groups. Based on our results, we can reject the null hypothesis and conclude that there is a significant difference in the average ratings among the four categories of businesses. Hence, the analysis suggests that businesses might benefit more from focusing on takeout services or offering a combination of delivery and takeout services to improve customer satisfaction. Businesses offering only delivery services should consider refining their services to enhance customer satisfaction levels.

OSEMN Report
Final Project

Process how I loaded data into database:

1. In the first stage, the Yelp dataset was downloaded and saved locally as a JSON file. After then, this file was put on a cloud storage system for later use.

2. After successfully uploading the file, a new dataset was created in BigQuery.

3. The JSON data file was loaded into Dataprep from cloud storage.

4. Within Dataprep, a new flow was developed for processing data.

5. Since the focus of this research was on firms in the United States, recipes were added in Dataprep to the data to eliminate any null values and exclude enterprises from Canada.

6. Once the data was filtered, the job was executed to create a new table in BigQuery, which was updated with each run.

7. BigQuery was utilized to answer the 2^{nd} question of the analysis.

8. For the first and third questions, PySpark was employed. A Python script was written to perform specific calculations and generate visualizations relevant to these questions.

Highlight how you used Spark in a parallelized computation:

In the python code for questions 1 and 3, Spark is used in a parallelized computation in several ways:

Loading and transforming data: Spark is used to load and transform data from Google BigQuery into a Spark Data Frame. The data is loaded in parallel across multiple worker nodes, which allows for faster processing of large datasets.

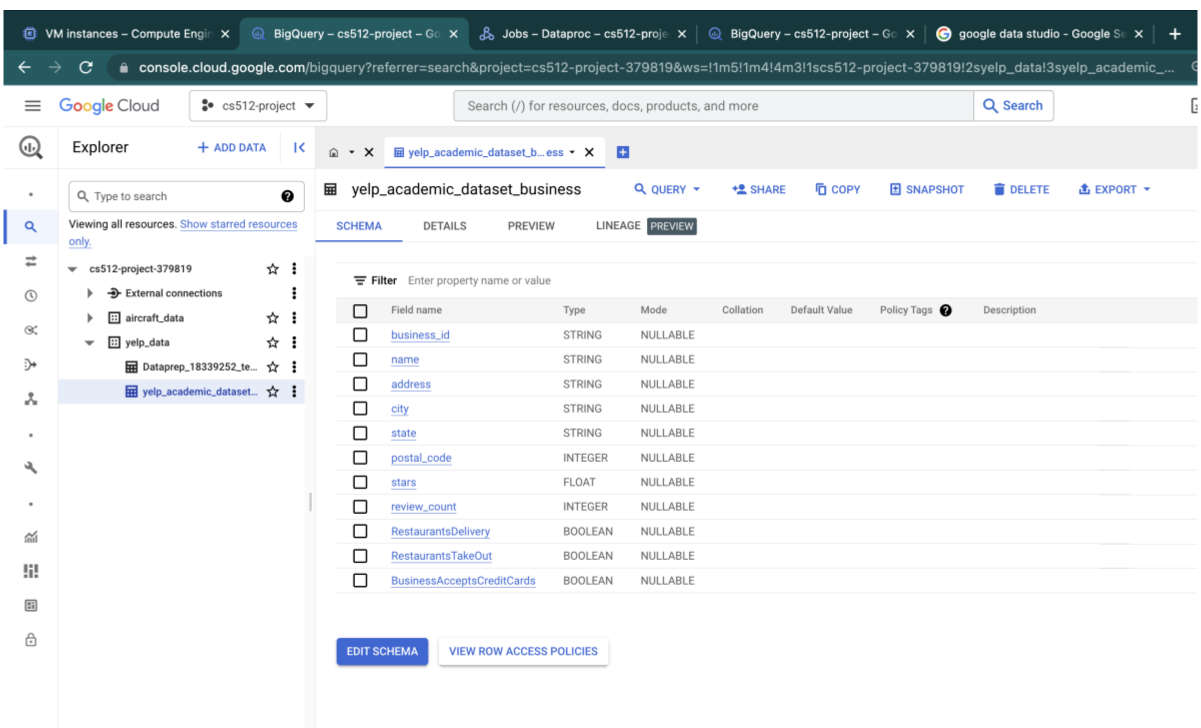
Filtering and grouping data: Spark is used to filter and group data in parallel across multiple worker nodes. For example, in question 1, the DataFrame is filtered to only include businesses that offer both delivery and takeout services and have at least 50 reviews. The filtered DataFrame is then grouped by city and the average star rating is calculated for each city. This process is done in parallel across multiple worker nodes, which allows for faster processing of large datasets.

Calculating statistics: Spark is used to calculate statistics such as mean, count, and correlation coefficient in parallel across multiple worker nodes. For example, in question 2, Spark's built-in mean() function is used to calculate the average star rating for each group (businesses that offer delivery services only, takeout services only, both services, and neither service). These calculations are done in parallel across multiple worker nodes, which allows for faster processing of large datasets.

OSEMN Report
Final Project

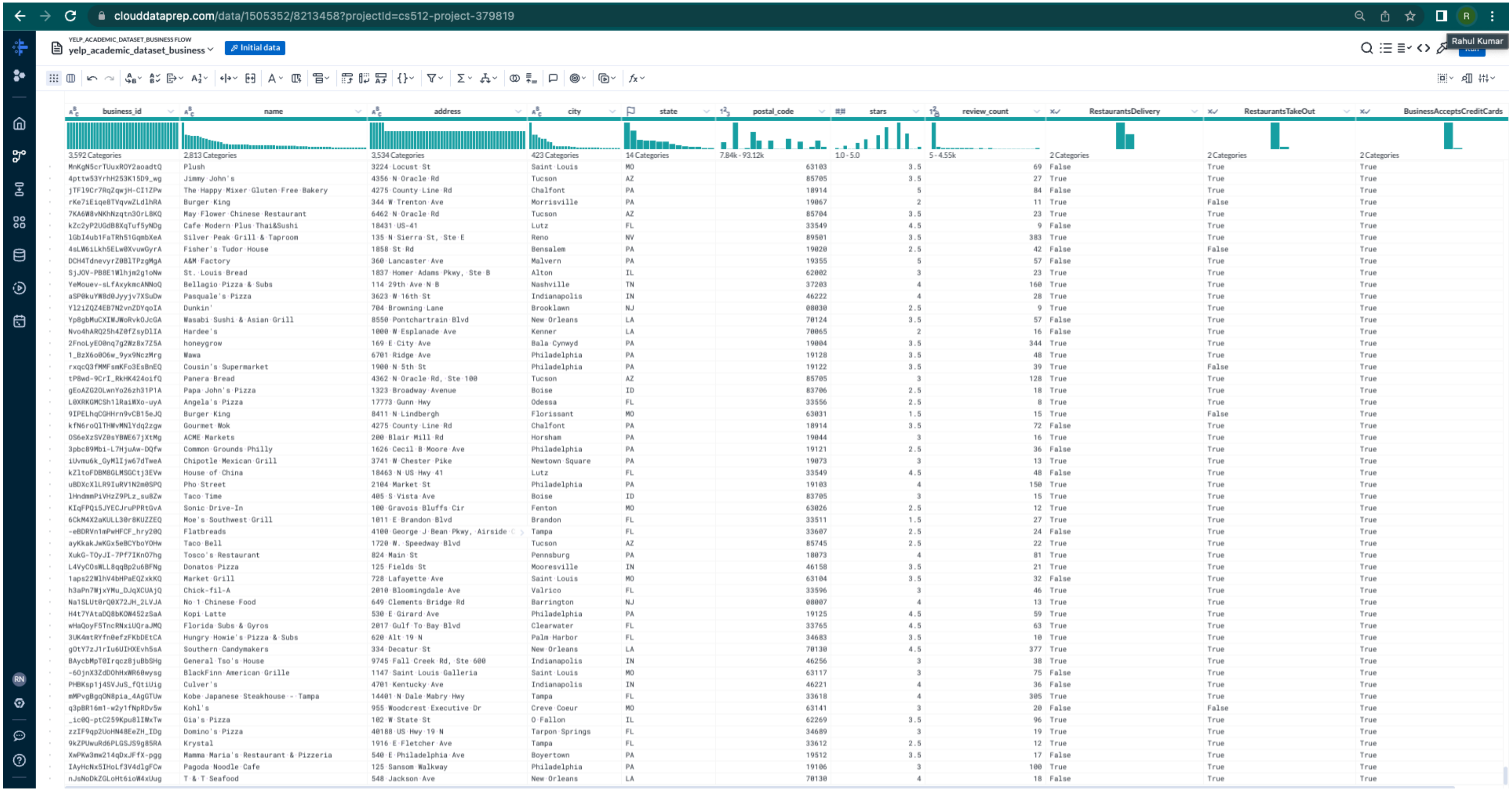
Appendix

Initial Data:



OSEMN Report
Final Project

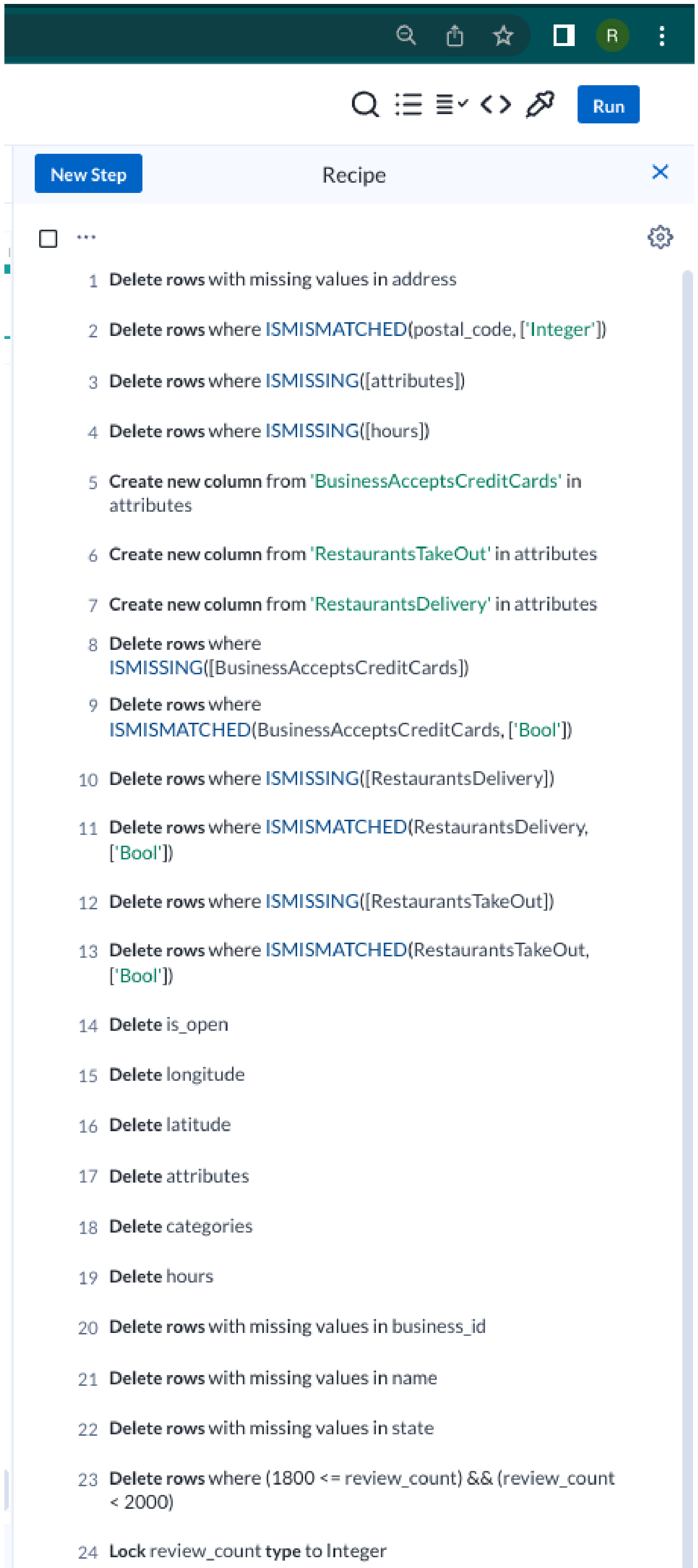
Data Prep for 1st question of interest.



business_id	name	address	city	state	postal_code	stars	review_count	RestaurantsDelivery	RestaurantsTakeOut	BusinessAcceptsCreditCards
3,592 Categories	2,813 Categories	3,534 Categories	423 Categories	14 Categories	7.84k - 93.12k	1.0 - 5.0	5 - 4.55k	2 Categories	2 Categories	2 Categories
MnKgN5crTUuxROY2aoadtQ	Plush	3224 Locust St	Saint Louis	MO	63103	3.5	69	False	True	True
4pttw53YrhH253K15D9_wg	Jimmy John's	4356 N Oracle Rd	Tucson	AZ	85705	3.5	27	True	True	True
jTF19Cr7RqZqwjH-CI1ZPw	The Happy Mixer Gluten Free Bakery	4275 County Line Rd	Chalfont	PA	18914	5	84	False	True	True
rKe7iEiqe8TVqvwZLdlhRA	Burger King	344 W Trenton Ave	Morrisville	PA	19067	2	11	True	False	True
7KA6W8vNKhNzqtn3OrL8KQ	May Flower Chinese Restaurant	6462 N Oracle Rd	Tucson	AZ	85704	3.5	23	True	True	True
kZc2yP2UGdB8XqTuf5yNDg	Cafe Modern Plus Thai&Sushi	18431 US-41	Lutz	FL	33549	4.5	9	False	True	True
lGbI4ub1FaTRh51GqmbXeA	Silver Peak Grill & Taproom	135 N Sierra St, Ste E	Reno	NV	89501	3.5	383	True	True	True
4sLW6iLkh5ELw0XvuwGyrA	Fisher's Tudor House	1858 St Rd	Bensalem	PA	19020	2.5	42	False	False	True
DCH4TdnevyrZ0BlTPzgMgA	A&M Factory	360 Lancaster Ave	Malvern	PA	19355	5	57	False	True	True
SjJOV-PB8E1Wlhjm2g1oNw	St. Louis Bread	1837 Homer Adams Pkwy, Ste B	Alton	IL	62002	3	23	True	True	True
YeMouev-sLfAxykmcANNoQ	Bellagio Pizza & Subs	114 29th Ave N B	Nashville	TN	37203	4	160	True	True	True
aSP0kuYW8d0Jyyjv7XSuDw	Pasquale's Pizza	3623 W 16th St	Indianapolis	IN	46222	4	28	True	True	True
Yl2iZQZ4EB7N2vnZDYqoIA	Dunkin'	704 Browning Lane	Brooklawn	NJ	08030	2.5	9	True	True	True
Yp8gbMuCXIWJWoRvkOJcGA	Wasabi Sushi & Asian Grill	8550 Pontchartrain Blvd	New Orleans	LA	70124	3.5	57	False	True	True
Nvo4hARQ25h4Z0fZsyDlIA	Hardee's	1000 W Esplanade Ave	Kenner	LA	70065	2	16	False	True	True
2FnoLyEO0nq7g2Wz8x7Z5A	honeygrow	169 E City Ave	Bala Cynwyd	PA	19004	3.5	344	True	True	True
1_BzX6o0O6w_9yx9NczMrg	Wawa	6701 Ridge Ave	Philadelphia	PA	19128	3.5	48	True	True	True
rxqcQ3fMMFsmKFo3EsBnEQ	Cousin's Supermarket	1900 N 5th St	Philadelphia	PA	19122	3.5	39	True	False	True
tP8wd-9CrI_RkHK424oifQ	Panera Bread	4362 N Oracle Rd, Ste 100	Tucson	AZ	85705	3	128	True	True	True
gEoAZG2OLwnYo26zh31P1A	Papa John's Pizza	1323 Broadway Avenue	Boise	ID	83706	2.5	18	True	True	True
L0XRKGMCSh1lRaiWXo-uyA	Angela's Pizza	17773 Gunn Hwy	Odessa	FL	33556	2.5	8	True	True	True
9IPELhqCGHHrn9vCB15eJQ	Burger King	8411 N Lindbergh	Florissant	MO	63031	1.5	15	True	False	True
kfN6roQlTHWvMNlYdq2zgw	Gourmet Wok	4275 County Line Rd	Chalfont	PA	18914	3.5	72	False	True	True
OS6eXzSVZ0sYBWE67jXtMg	ACME Markets	200 Blair Mill Rd	Horsham	PA	19044	3	16	True	True	True
3pbc89Mbi-L7HjuAw-DQfw	Common Grounds Philly	1626 Cecil B Moore Ave	Philadelphia	PA	19121	2.5	36	False	True	True
iUvmu6k_GyMlIjw67dTweA	Chipotle Mexican Grill	3741 W Chester Pike	Newtown Square	PA	19073	3	13	True	True	True
kZltoFDBM8GLMSGCtj3EVw	House of China	18463 N US Hwy 41	Lutz	FL	33549	4.5	48	False	True	True
uBDXcXlLR9IuRV1N2m0SPQ	Pho Street	2104 Market St	Philadelphia	PA	19103	4	150	True	True	True
lHndmmPiVHzZ9PLz_su8Zw	Taco Time	405 S Vista Ave	Boise	ID	83705	3	15	True	True	True
KIqFPQi5JYECJruPPRtGvA	Sonic Drive-In	100 Gravois Bluffs Cir	Fenton	MO	63026	2.5	12	True	True	True
6CkM4X2aKULL30r8KUZZEQ	Moe's Southwest Grill	1011 E Brandon Blvd	Brandon	FL	33511	1.5	27	True	True	True
-eBDRVn1mPwHFCF_hry20Q	Flatbreads	4100 George J Bean Pkwy, Airside C	Tampa	FL	33607	2.5	24	False	True	True
ayKkakJwKGx5eBCYboYOHw	Taco Bell	1720 W. Speedway Blvd	Tucson	AZ	85745	2.5	22	True	True	True
XukG-TOyJI-7Pf7IKnO7hg	Tosco's Restaurant	824 Main St	Pennsburg	PA	18073	4	81	True	True	True
L4VyCOsWLL8qqBp2u6BFNg	Donatos Pizza	125 Fields St	Mooresville	IN	46158	3.5	21	True	True	True
1aps22WlhV4bHPaEQZxkKQ	Market Grill	728 Lafayette Ave	Saint Louis	MO	63104	3.5	32	False	True	True
h3aPn7WjxYMu_DJqXCUAjQ	Chick-fil-A	2010 Bloomingdale Ave	Valrico	FL	33596	3	46	True	True	True
Na1SLUt0rQ0X72JH_2LVJA	No 1 Chinese Food	649 Clements Bridge Rd	Barrington	NJ	08007	4	13	True	True	True
H4t7YAtaDQ8bKOW452zSaA	Kopi Latte	530 E Girard Ave	Philadelphia	PA	19125	4.5	59	True	True	True
wHaQoyF5TncRNxiUQraJMQ	Florida Subs & Gyros	2017 Gulf To Bay Blvd	Clearwater	FL	33765	4.5	63	True	True	True
3UK4mtRYfn0efzFKbDEtCA	Hungry Howie's Pizza & Subs	620 Alt 19 N	Palm Harbor	FL	34683	3.5	10	True	True	True
gOtY7zJ1rIu6UIHXEvh5sA	Southern Candymakers	334 Decatur St	New Orleans	LA	70130	4.5	377	True	True	True
BAycbMpT0Irqcz8juBbSHg	General Tso's House	9745 Fall Creek Rd, Ste 600	Indianapolis	IN	46256	3	38	True	True	True
-6OjnX3ZdDOhHxWR60wysg	BlackFinn American Grille	1147 Saint Louis Galleria	Saint Louis	MO	63117	3	75	False	True	True
PHBKsp1j4SVJuS_fQtiUig	Culver's	4701 Kentucky Ave	Indianapolis	IN	46221	4	36	False	True	True
mMPvgBgqON8pia_4AgGTUw	Kobe Japanese Steakhouse - Tampa	14401 N Dale Mabry Hwy	Tampa	FL	33618	4	305	True	True	True
q3pBR16m1-w2y1fNpRDv5w	Kohl's	955 Woodcrest Executive Dr	Creve Coeur	MO	63141	3	20	False	False	True
_ic0Q-ptC259Kpu8lIWxTw	Gia's Pizza	102 W State St	O Fallon	IL	62269	3.5	96	True	True	True
zzIF9qp2UoHN48EeZH_IDg	Domino's Pizza	40188 US Hwy 19 N	Tarpon Springs	FL	34689	3	19	True	True	True
9kZPUwuRd6PLGSJS9g85RA	Krystal	1916 E Fletcher Ave	Tampa	FL	33612	2.5	12	True	True	True
XwPKw3mw214qDxJFfX-pgg	Mamma Maria's Restaurant & Pizzeria	540 E Philadelphia Ave	Boyertown	PA	19512	3.5	17	False	True	True
IAyHcNx5IHoLf3V4dlgFCw	Pagoda Noodle Cafe	125 Sansom Walkway	Philadelphia	PA	19106	3	100	True	True	True
nJsNoDkZGLoHt6ioW4xUug	T & T Seafood	548 Jackson Ave	New Orleans	LA	70130	4	18	False	True	True

OSEMN Report
Final Project

Recipes for 1st question of interest.



OSEMN Report
Final Project

Python code: for 1st and 3rd question.

```python
### Rahul Kumar Nalubandhu
## Final Project

import pyspark
from pyspark.sql import SparkSession
import pprint
import matplotlib.pyplot as plt
import json
from pyspark.sql.types import StructType, FloatType, LongType, StringType, StructField, BooleanType, IntegerType
from pyspark.sql.functions import col, count, mean
from google.cloud import storage

from scipy import stats


sc = pyspark.SparkContext()

bucket = sc._jsc.hadoopConfiguration().get('fs.gs.system.bucket')
project = sc._jsc.hadoopConfiguration().get('fs.gs.project.id')
input_directory = 'gs://{}/hadoop/tmp/bigquerry/pyspark_input'.format(bucket)
output_directory = 'gs://{}/pyspark_demo_output'.format(bucket)

spark = SparkSession \
  .builder \
  .master('yarn') \
  .appName('Yelp') \
  .getOrCreate()

#update with your project specific settings
conf={
    'mapred.bq.project.id':project,
    'mapred.bq.gcs.bucket':bucket,
    'mapred.bq.temp.gcs.path':input_directory,
    'mapred.bq.input.project.id': 'cs512-project-379819',
    'mapred.bq.input.dataset.id': 'yelp_data',
    'mapred.bq.input.table.id': 'yelp_academic_dataset_business',
}

## pull table from big query
table_data = sc.newAPIHadoopRDD(
    'com.google.cloud.hadoop.io.bigquery.JsonTextBigQueryInputFormat',
    'org.apache.hadoop.io.LongWritable',
    'com.google.gson.JsonObject',
    conf = conf)

## convert table to a json like object, turn PosTime and Fseen back into numbers
vals = table_data.values()
# pprint.pprint(vals.take(5))  #added to help debug whether table was loaded
vals = vals.map(lambda line: json.loads(line))
vals = vals.map(lambda x: {**x, 'review_count': int(x['review_count'])})
vals = vals.map(lambda x: {**x, 'stars': float(x['stars'])})

##schema
schema = StructType([
   StructField('business_id', StringType(), True),
   StructField('name', StringType(), True),
   StructField('address', StringType(), True),
   StructField('city', StringType(), True),
   StructField('state', StringType(), True),
   StructField('postal_code', StringType(), True),
   StructField('stars', FloatType(), True),
   StructField('review_count', LongType(), True),
   StructField('RestaurantsDelivery', BooleanType(), True),
   StructField('RestaurantsTakeOut', BooleanType(), True),
   StructField('BusinessAcceptsCreditCards', BooleanType(), True),
])

## create a dataframe object
```

OSEMN Report
Final Project

Window_spark_yelp.py 6

group_project_3 > Window_spark_yelp.py > ...

```python
## create a dataframe object
df = spark.createDataFrame(vals, schema= schema)
df.show(10)
# Filter the data // this is for 2nd question
filtered_data = df.filter((col("RestaurantsDelivery") == True) & (col("RestaurantsTakeOut") == True) & (col("review_count") >= 50))
# Group the filtered data by city and calculate the average star rating
city_group = filtered_data.groupBy("city").agg(mean("stars").alias("avg_stars"), count("business_id").alias("restaurant_count"))
# Sort the grouped data by the average star rating in descending order and select the top 50 cities
top_cities = city_group.sort(col("avg_stars").desc()).limit(50)
# Calculate the proportion of highly-rated restaurants that accept card payments
card_accepting = filtered_data.filter(col("BusinessAcceptsCreditCards") == True).groupBy("city").agg(count("business_id").alias("card_accepting_count"))
top_cities = top_cities.join(card_accepting, on="city")
top_cities = top_cities.withColumn("card_accepting_proportion", col("card_accepting_count") / col("restaurant_count"))
# Display the results
top_cities.select("city", "avg_stars", "restaurant_count", "card_accepting_proportion").show()

# this is for 3rd question of intrest

# filtering the data to create two separate DataFrames for businesses offering delivery services and those offering takeout services:

delivery_services = df.filter((col("RestaurantsDelivery") == True) & (col("RestaurantsTakeOut") == False))
takeout_services = df.filter((col("RestaurantsDelivery") == False) & (col("RestaurantsTakeOut") == True))


# Calculate the average star rating of the businesses offering delivery services
delivery_avg_rating = delivery_services.agg(mean("stars").alias("avg_delivery_stars")).collect()[0]["avg_delivery_stars"]
takeout_avg_rating = takeout_services.agg(mean("stars").alias("avg_takeout_stars")).collect()[0]["avg_takeout_stars"]

# Convert Spark DataFrame to Pandas DataFrame
top_cities_pd = top_cities.toPandas()

# Convert the 'stars' column of each group's DataFrame to Pandas Series to prepare for the t-test:
delivery_ratings_pd = delivery_services.select("stars").toPandas()["stars"]
takeout_ratings_pd = takeout_services.select("stars").toPandas()["stars"]

# Perform the independent two-sample t-test:
t_statistic, p_value = stats.ttest_ind(delivery_ratings_pd, takeout_ratings_pd)

# Display the results
print(f"T-statistic: {t_statistic}, P-value: {p_value}")

# Calculate the correlation coefficient between 'avg_stars' and 'card_accepting_proportion'
correlation_coefficient = top_cities_pd['avg_stars'].corr(top_cities_pd['card_accepting_proportion'], method='pearson')
print(f"Correlation coefficient: {correlation_coefficient}")

# Save the plots to image files
plt.figure(figsize=(20, 8))
plt.bar(top_cities_pd["city"], top_cities_pd["avg_stars"])
plt.xlabel("City")
plt.ylabel("Average Star Rating")
plt.title("Average Star Ratings of Top Cities")
plt.xticks(rotation=90)
plt.subplots_adjust(bottom=0.25)
plt.savefig("average_star_ratings.png")  # Save the plot to a file

# Create a scatter plot
plt.figure(figsize=(20, 8))
plt.scatter(top_cities_pd["city"], top_cities_pd["card_accepting_proportion"])
plt.xlabel("City")
plt.ylabel("Proportion of Card Accepting Restaurants")
plt.title("Relationship Between City and Card Accepting Restaurants Proportion in Top Cities")
# Rotate the x-axis city labels to 45 degrees
plt.xticks(rotation=90)
# Adjust the margins to accommodate the rotated labels
plt.subplots_adjust(bottom=0.25)
plt.savefig("card_accepting_proportion.png")  # Save the plot to a file
```

⊗ 0 ⚠ 6 tabnine starter

OSEMN Report
Final Project

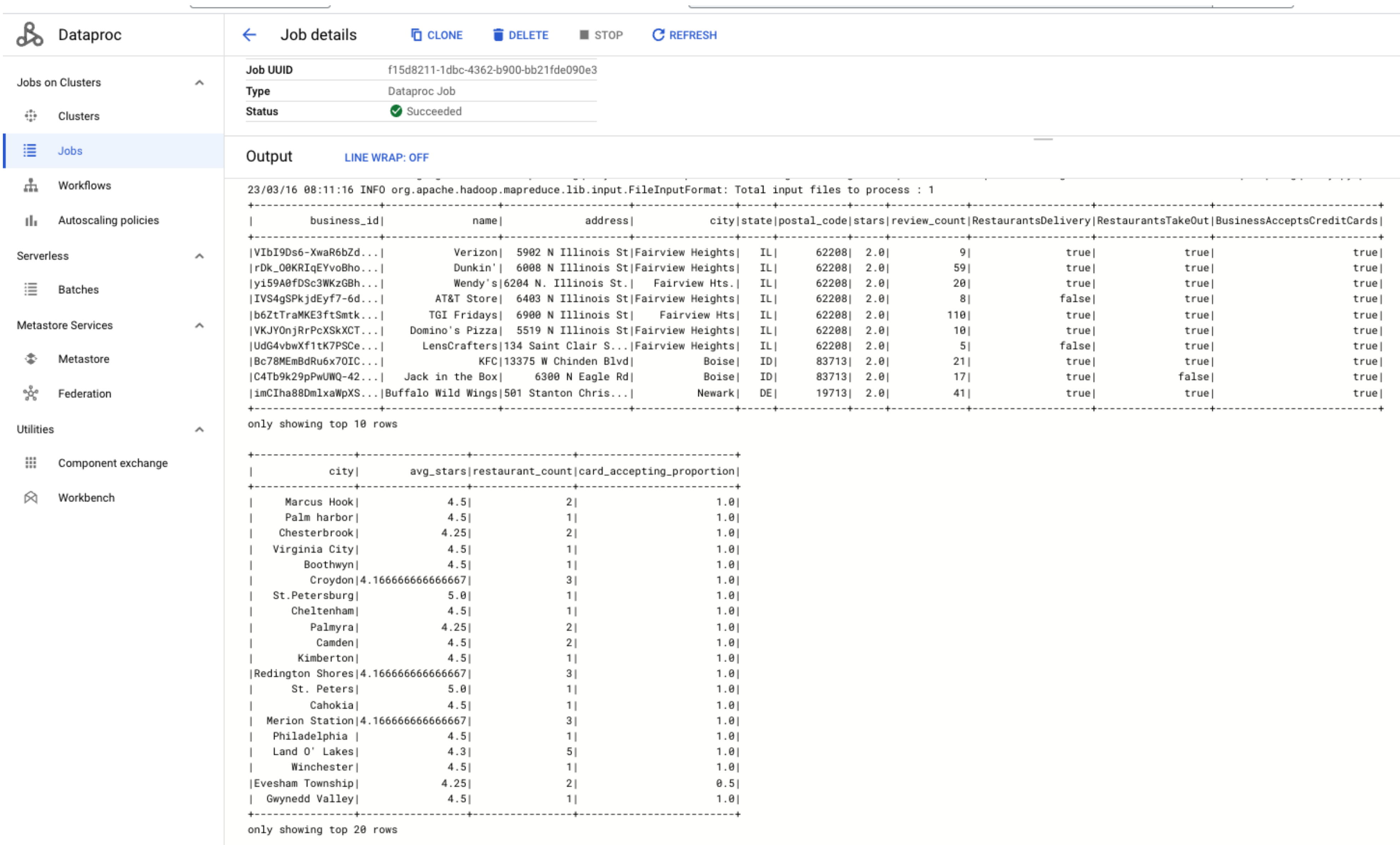
Window_spark_yelp.py 6 ×

group_project_3 > Window_spark_yelp.py > ...

```python
plt.figure(figsize=(20, 8))
plt.scatter(top_cities_pd["city"], top_cities_pd["card_accepting_proportion"])
plt.xlabel("City")
plt.ylabel("Proportion of Card Accepting Restaurants")
plt.title("Relationship Between City and Card Accepting Restaurants Proportion in Top Cities")
# Rotate the x-axis city labels to 45 degrees
plt.xticks(rotation=90)
# Adjust the margins to accommodate the rotated labels
plt.subplots_adjust(bottom=0.25)
plt.savefig("card_accepting_proportion.png")  # Save the plot to a file

plt.figure(figsize=(20, 8))
plt.boxplot([delivery_ratings_pd, takeout_ratings_pd], labels=["Delivery Services", "Takeout Services"])
plt.ylabel("Stars")
plt.title("Distribution of Ratings for Delivery and Takeout Services")
plt.savefig("ratings_distribution.png")


def upload_to_gcs(bucket_name, source_file_name, destination_blob_name):
    storage_client = storage.Client()
    bucket = storage_client.bucket(bucket_name)
    blob = bucket.blob(destination_blob_name)

    blob.upload_from_filename(source_file_name)
    print(f"File {source_file_name} uploaded to {destination_blob_name}.")

bucket_name = "yelp_data_proj"
source_file_name1 = "average_star_ratings.png"
destination_blob_name1 = f"images/average_star_ratings.png"
source_file_name2 = "card_accepting_proportion.png"
destination_blob_name2 = f"images/card_accepting_proportion.png"
upload_to_gcs(bucket_name, source_file_name1, destination_blob_name1)
upload_to_gcs(bucket_name, source_file_name2, destination_blob_name2)
source_file_name3 = "ratings_distribution.png"
destination_blob_name3 = f"images/ratings_distribution.png"
upload_to_gcs(bucket_name, source_file_name3, destination_blob_name3)



## deletes the temporary files
input_path = sc._jvm.org.apache.hadoop.fs.Path(input_directory)
input_path.getFileSystem(sc._jsc.hadoopConfiguration()).delete(input_path, True)
```

⊗ 0 ⚠ 6   tabnine starter

OSEMN Report
Final Project



23/03/16 08:11:16 INFO org.apache.hadoop.mapreduce.lib.input.FileInputFormat: Total input files to process : 1

business_id	name	address	city	state	postal_code	stars	review_count	RestaurantsDelivery	RestaurantsTakeOut	BusinessAcceptsCreditCards
VIbI9Ds6-XwaR6bZd...	Verizon	5902 N Illinois St	Fairview Heights	IL	62208	2.0	9	true	true	true
rDk_O0KRIqEYvoBho...	Dunkin'	6008 N Illinois St	Fairview Heights	IL	62208	2.0	59	true	true	true
yi59A0fDSc3WKzGBh...	Wendy's	6204 N. Illinois St.	Fairview Hts.	IL	62208	2.0	20	true	true	true
IVS4gSPkjdEyf7-6d...	AT&T Store	6403 N Illinois St	Fairview Heights	IL	62208	2.0	8	false	true	true
b6ZtTraMKE3ftSmtk...	TGI Fridays	6900 N Illinois St	Fairview Hts	IL	62208	2.0	110	true	true	true
VKJYOnjRrPcXSkXCT...	Domino's Pizza	5519 N Illinois St	Fairview Heights	IL	62208	2.0	10	true	true	true
UdG4vbwXf1tK7PSCe...	LensCrafters	134 Saint Clair S...	Fairview Heights	IL	62208	2.0	5	false	true	true
Bc78MEmBdRu6x7OIC...	KFC	13375 W Chinden Blvd	Boise	ID	83713	2.0	21	true	true	true
C4Tb9k29pPwUWQ-42...	Jack in the Box	6300 N Eagle Rd	Boise	ID	83713	2.0	17	true	false	true
imCIha88DmlxaWpXS...	Buffalo Wild Wings	501 Stanton Chris...	Newark	DE	19713	2.0	41	true	true	true

only showing top 10 rows

city	avg_stars	restaurant_count	card_accepting_proportion
Marcus Hook	4.5	2	1.0
Palm harbor	4.5	1	1.0
Chesterbrook	4.25	2	1.0
Virginia City	4.5	1	1.0
Boothwyn	4.5	1	1.0
Croydon	4.166666666666667	3	1.0
St.Petersburg	5.0	1	1.0
Cheltenham	4.5	1	1.0
Palmyra	4.25	2	1.0
Camden	4.5	2	1.0
Kimberton	4.5	1	1.0
Redington Shores	4.166666666666667	3	1.0
St. Peters	5.0	1	1.0
Cahokia	4.5	1	1.0
Merion Station	4.166666666666667	3	1.0
Philadelphia	4.5	1	1.0
Land O' Lakes	4.3	5	1.0
Winchester	4.5	1	1.0
Evesham Township	4.25	2	0.5
Gwynedd Valley	4.5	1	1.0

only showing top 20 rows

```
F-statistic for all 4 categories: 597.8902654613838, P-value: 0.0
Correlation coefficient: 0.15147935326504264
Summary statistics for delivery_services:
+-------+------------------+
|summary|             stars|
+-------+------------------+
|  count|              1372|
|   mean| 2.777332361516035|
| stddev|1.0387197684817724|
|    min|               1.0|
|    max|               5.0|
+-------+------------------+


Summary statistics for takeout_services:
+-------+------------------+
|summary|             stars|
+-------+------------------+
|  count|             13259|
|   mean|3.6687532996455237|
| stddev|0.7224428737562917|
|    min|               1.0|
|    max|               5.0|
+-------+------------------+


Summary statistics for both_services:
+-------+-----------------+
|summary|            stars|
+-------+-----------------+
|  count|            24457|
|   mean|3.449728094206158|
| stddev|0.860909627361036|
|    min|              1.0|
|    max|              5.0|
+-------+-----------------+


Summary statistics for no_services:
+-------+------------------+
|summary|             stars|
+-------+------------------+
|  count|              1702|
|   mean|3.6818448883666277|
| stddev| 0.759021387586643|
|    min|               1.0|
|    max|               5.0|
+-------+------------------+
```

OSEMN Report
Final Project

Collaboration:

I collaborated with my teammate, Sandra, on final project. For the report, I was responsible for questions 1 and 3, while Sandra worked on the 2nd question. However, I assisted Sandra with the BigQuery for her question. Later we exchanged questions, where I gave my 1st question to Sandra, and she gave me her question which is the 2nd in my report. And 3rd question is done on my own.

References

e-satisfaction. (n.d.). *7 reasons why customer reviews are important*. Retrieved January 27, 2023, from https://www.e-satisfaction.com/7-reasons-why-customer-reviews-are-important/#:~:text=Analyzing%20reviews%20left%20by%20your,what%20your%20customers%20truly%20want

Rana, D. S. (2022, September 12). *Learn Review Analysis: Why, how, data sources with Free Trial*. Learn Review Analysis | Why, How, Data Sources with Free Trial. Retrieved January 29, 2023, from https://www.repustate.com/blog/review-analysis/#:~:text=Review%20analysis%20allows%20you%20to,new%20sales%20opportunities%2C%20and%20more.

Indexing and selecting data - pandas 1.5.3 documentation. (n.d.). Retrieved January 29, 2023, from https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy.

Yelp for Business. (2023, January 23). *Plan, start, grow, and advertise your small business*. Yelp for Business. Retrieved January 29, 2023, from https://business.yelp.com/

PH717 module 9 - correlation and regression. The Correlation Coefficient (r). (n.d.). Retrieved February 12, 2023, from https://sphweb.bumc.bu.edu/otlt/MPH-Modules/PH717-QuantCore/PH717-Module9-Correlation-Regression/PH717-Module9-Correlation-Regression4.html

Pyspark.pandas.series.plot.box. pyspark.pandas.Series.plot.box - PySpark 3.3.2 documentation. (n.d.). Retrieved March 17, 2023, from https://sparkbyexamples.com/pandas/pandas-boxplot-from-dataframe/

Vijetha. (2022, September 28). *How to plot the boxplot from DataFrame?* Spark By{Examples}. Retrieved March 17, 2023, from https://sparkbyexamples.com/pandas/pandas-boxplot-from-dataframe/

Kapoor, V. (2022, September 9). *PySpark data analysis part-1*. Medium. Retrieved March 18, 2023, from https://medium.com/@vineet_kapoor_2019/pyspark-data-analysis-part-1-37ae66fd7909

Sirca, N. (2020, May 7). *ANOVA with python and SQL*. Medium. Retrieved March 18, 2023, from https://towardsdatascience.com/anova-with-python-and-sql-b37b68ebc2dc