**Exploring Yelp Reviews Data: Insights and Visualizations using OSEMN Process**

**Rahul Kumar Nalubandhu**

**Introduction** 

Conducting a review analysis gives businesses the opportunity to improve customer experience, identify service gaps and gain real time insights amongst many other benefits. This analysis dives into customer reviews entered in Yelp for businesses in different states throughout the United States. In this project I have 3 main questions.

2. Can you identify the top 5 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?
3. What is the relationship between the average stars of a restaurant and the number of reviews it has received?

* Do restaurants with more reviews have a higher average stars or is there no significant relationship?

**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.

**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.  
  
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 two questions, the exploration stage, the csv data is loaded into R-studio. To begin the exploration stage the review count field is plotted.

**Analysis and visualizations for 1st question:**

Question of interest?

**Analysis and visualizations for 2nd question:**

Question of interest?

Can you identify the top 5 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?

In this, we aimed to identify the top 5 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.

To achieve this, we followed these steps:

1. First, we created a dataset and uploaded it to cloud storage.
2. Next, we initiated a data flow in Dataprep by considering the Yelp academic business dataset. We then applied recipes to filter out null values and selected important categories such as business\_id, name, address, city, state, postal code, stars, review\_count, RestaurantsDelivery, RestaurantsTakeout, and BusinessAcceptsCards. We removed all null values and only considered businesses from the USA.
3. We proceeded to run the job and create a table in BigQuery.
4. We then modified the Python script to extract the relevant fields from the table to create a schema with the respective data types for each field.
5. We executed the job on the cluster and conducted an analysis to identify the top 5 cities with the highest average star ratings for restaurants that offer both delivery and takeout services and have a minimum of 50 reviews. We also analyzed the proportion of these highly rated restaurants that accept card payments compared to the overall number of restaurants in the same city.
6. To provide additional insight into the convenience and accessibility of these top restaurants in each city, we calculated the proportion of highly rated restaurants that accept card payments.
7. In the same python code, I also included the code to plot graphs.

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 5 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. Average Star Ratings of Top Cities:

This graph displays the average star ratings for the top 5 cities with the highest average ratings for restaurants that offer both delivery and takeout services and have a minimum of 50 reviews. By examining this graph, we can understand which cities have the highest-rated restaurants according to the specified criteria.

Chart, bar chart, histogram

Description automatically generated  
plot 1

1. Proportion of Card Accepting Restaurants in Top Cities:   
     
   This graph shows the proportion of highly rated restaurants that accept card payments in each of the top 5 cities identified in the first graph. This provides insight into the prevalence of card acceptance among the best-rated restaurants in those cities. It can help potential customers, business owners, and investors understand the convenience and accessibility of these top restaurants in terms of payment options.

Chart, application

Description automatically generated with medium confidence

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. However, the relationship is not strong, and other factors may contribute more significantly to a city's average star rating.

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. Further analysis could be conducted to identify other variables that may significantly influence a city's restaurant ratings.

**Analysis and visualizations for 3rd question:**

**Question of Interest ?**

Is there a significant difference in the average ratings of businesses that offer delivery services versus those that offer 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 two-sample t-test.

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 calculated the average rating for each group.  
  
so for this I performed the two-sample test as said earlier. In where we take the null hypothesis (H0) is that there is no significant difference in the average ratings between businesses offering delivery services and those offering takeout services.

The alternative hypothesis (H1) Is that there Is a significant difference between the two groups.

So why are we doing Two-Sample t-test?

The t-statistic measures how far the sample means are from each other, relative to the variation within each group.

A larger t-statistic indicates a larger difference between the group means. The p-value represents the probability of observing a t-statistic at least as extreme as the one calculated if there is no real difference between the group means.

A small p-value (typically less than 0.05) indicates that it is unlikely that the observed difference in means is due to random chance alone, and you can reject the null hypothesis (i.e., there is no difference between the two groups).

Hence, I updated the code to find t-static and p-value to say if we accept or reject the Hypothesis.

So, after doing some research I found out how to use python library stats to directly find t-static and p-value which is shown in the python code below.

After submitting the job in DataProc I got the p-value as 0.0 and t-statistic as -41.48343.

**Conclusions from result:**

Now from a negative t-statistic value it indicates that the mean of the delivery services group is lower than the mean of the takeout services group.

And a p-value of 0.0 indicates that the difference between the means of the two groups is statistically significant at the chosen significance level (typically 0.05).

Hence our results says that there is a significant difference in the average ratings between businesses that offer only delivery services and those that offer only takeout services, with businesses that offer only takeout services having a higher average rating.

**Visualization:**

Also, I generated an box plot for both the delivery services and Takeout services in the python using matplotlib library.

The reason for generating box plot for delivery and takeout services can help you visualize and compare the distribution of star ratings for both groups. It helps you identify any differences in central tendency, spread, and the presence of outliers between the two groups, which can be useful when interpreting the results of your statistical analysis.

Chart, box and whisker chart

Description automatically generated  
plot 3

From this boxplot we can conclude:

Delivery services: The line inside the box represents the median value of the distribution, which means that 50% of the businesses that offer only delivery services have a rating of 2.5 or lower, and the other 50% have a rating higher than 2.5.

Takeout services: The line is overlapping the top line of the box and is near to 4.0. This suggests that the median value for businesses offering only takeout services is around 4.0. This means that 50% of the businesses that offer only takeout services have a rating of 4.0 or lower, and the other 50% have a rating higher than 4.0.

The presence of an outlier in the takeout services box plot between 1.0 and 1.5 indicates that there is at least one business offering only takeout services with a rating that is significantly lower than the rest of the businesses in that group. To be precise the outlier is a little above 1.0, which means its rating is slightly higher than 1.0, but it is still much lower than the median rating of around 4.0 for takeout services.

In conclusion, the box plots show that businesses offering only takeout services generally have higher ratings than businesses offering only delivery services. The median rating for takeout services is around 4.0, while the median rating for delivery services is around 2.5. This further supports the earlier finding that there is a significant difference in the average ratings between the two groups, with takeout services having higher average ratings.

**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.

**Interpret Data**

For the 1st analysis, we can interpret that based ,,,,,,,

From the second analysis, we can interpret In this study, that we identified the top 5 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.

In the 3rd analysis, we can interpret as After conducting the t-test, the p-value is used to determine the significance of the results. If the p-value is less than the chosen significance level (e.g., 0.05), you can reject the null hypothesis and conclude that there is a significant difference in the average ratings between businesses offering delivery services and those offering takeout services. If the p-value is greater than the significance level, you cannot reject the null hypothesis, meaning there is no evidence to suggest a significant difference in average ratings between the two groups. Hence from our results we can say that we reject the null hypothesis and conclude there is a significant difference in the average ratings between businesses offering delivery services and those offering takeout services.

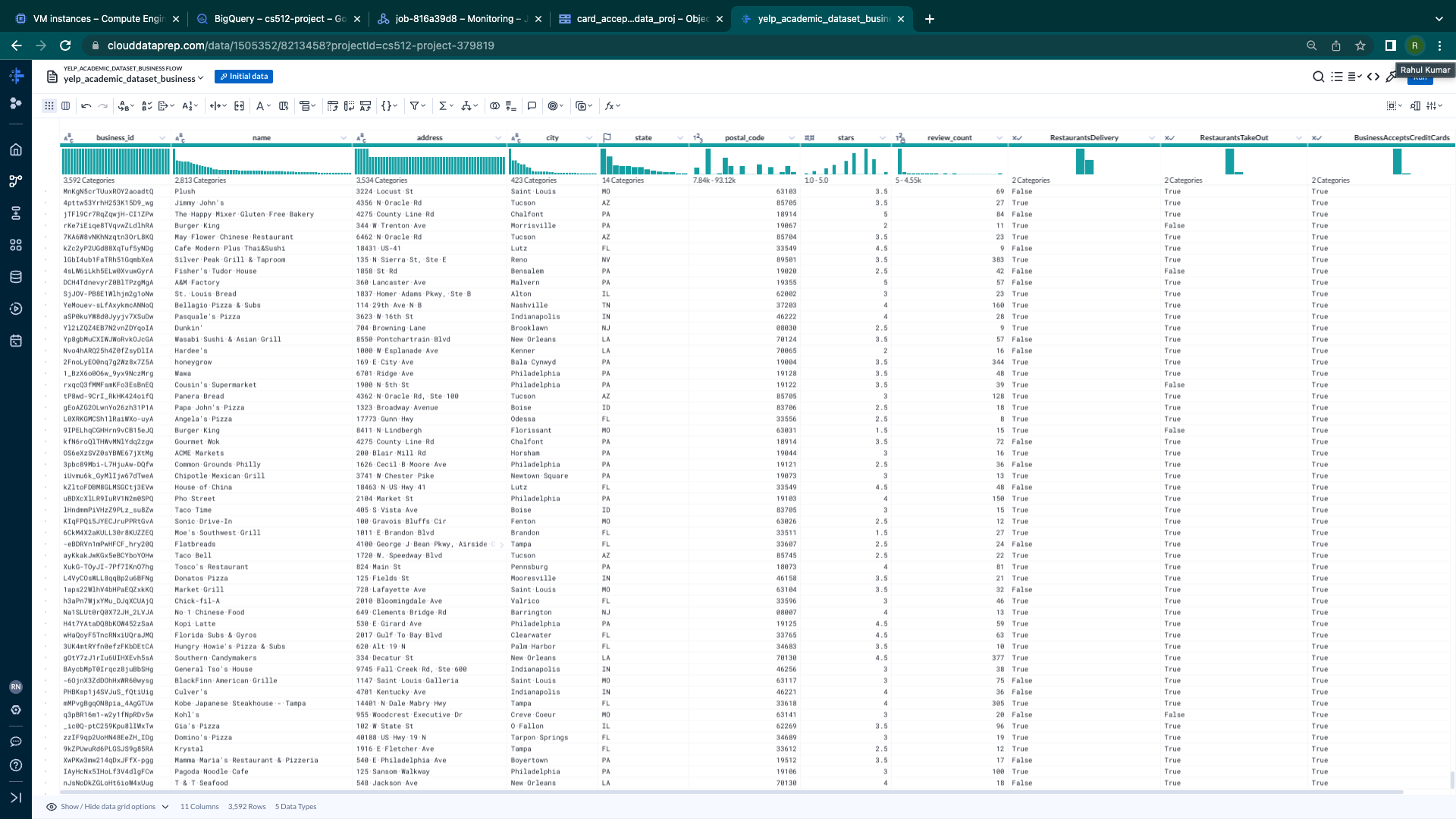
**Process how I loaded data into database:**

1. The first step was to extract the relevant data from the original Yelp dataset and convert it into two separate .csv files, containing the required fields.
2. The next step was to import the two .csv files into phpMyAdmin.
3. Within phpMyAdmin, a new table was created with headers for each respective column. During the import of the data, the header row was excluded, and appropriate data types were assigned to each column. The business\_id column was also set as the primary key.
4. The above step was repeated for the second .csv file containing business\_id and review count data.
5. A SQL query was then written to join the two tables using the primary key.
6. After running the SQL query, the resulting data was exported as a .csv file and loaded into R-studio for visualizations.

**Appendix**

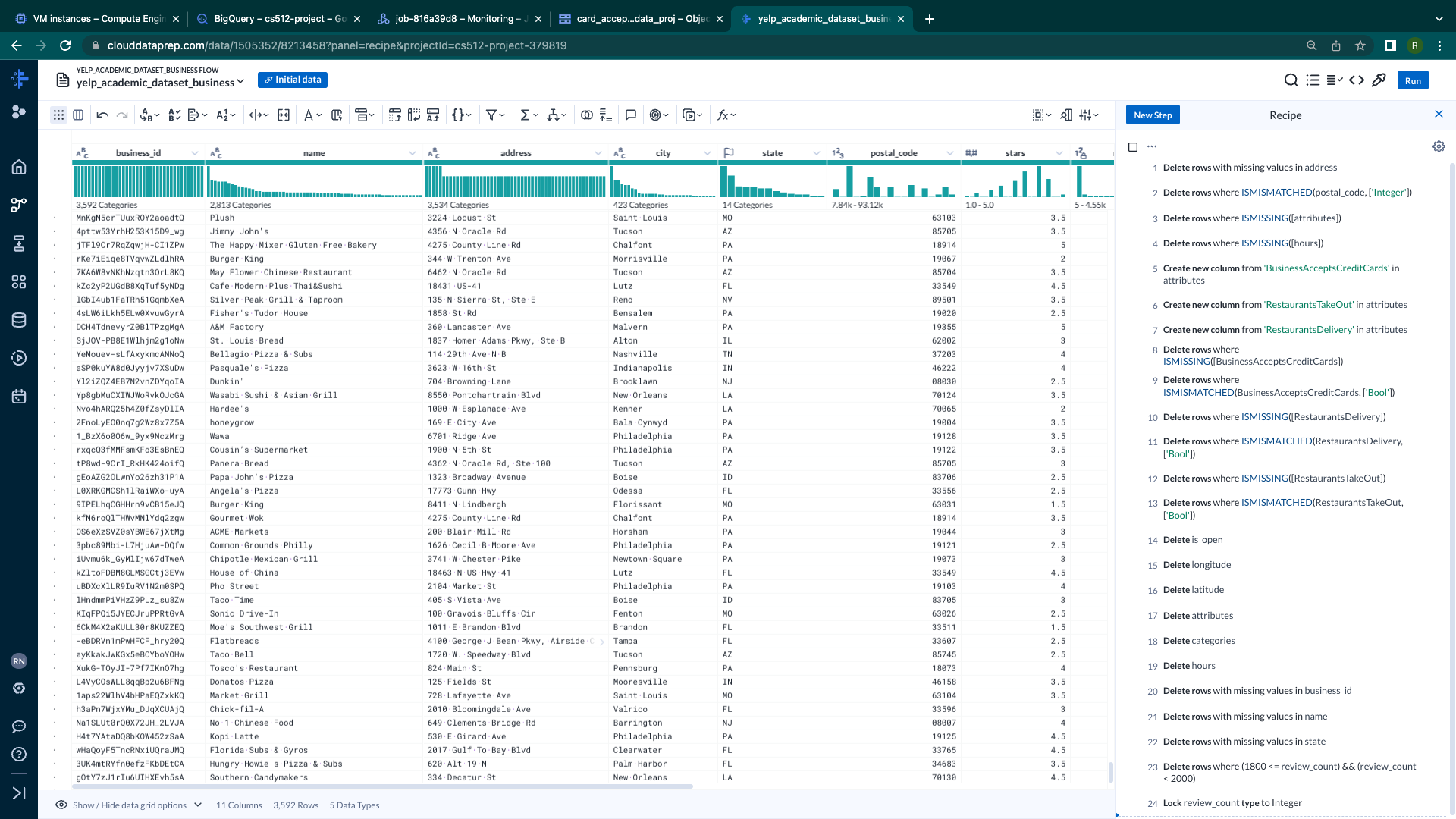
Initial Data:





Data Prep for 2nd question of intrest.

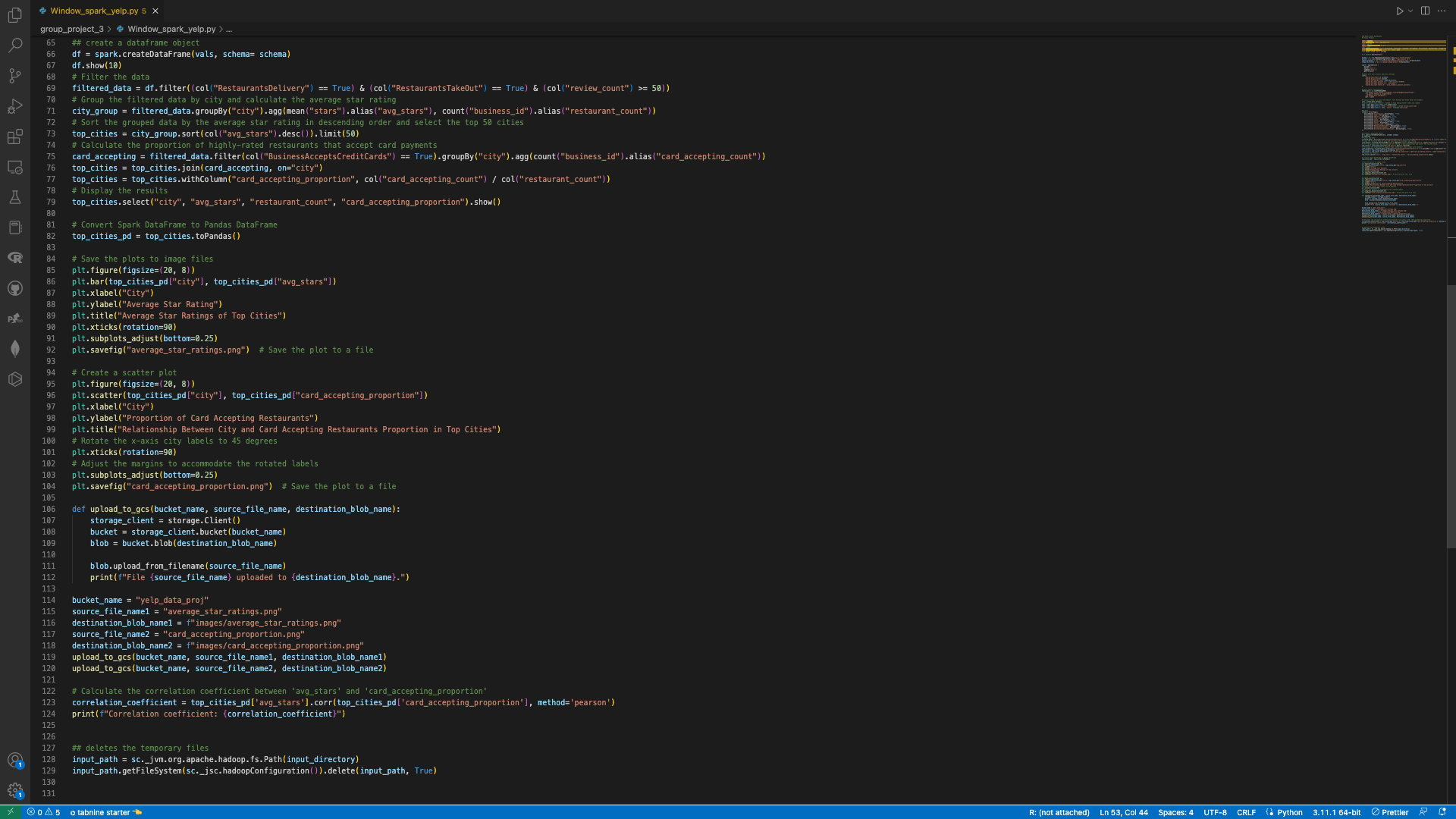
Recipes for 2nd question of intrest.



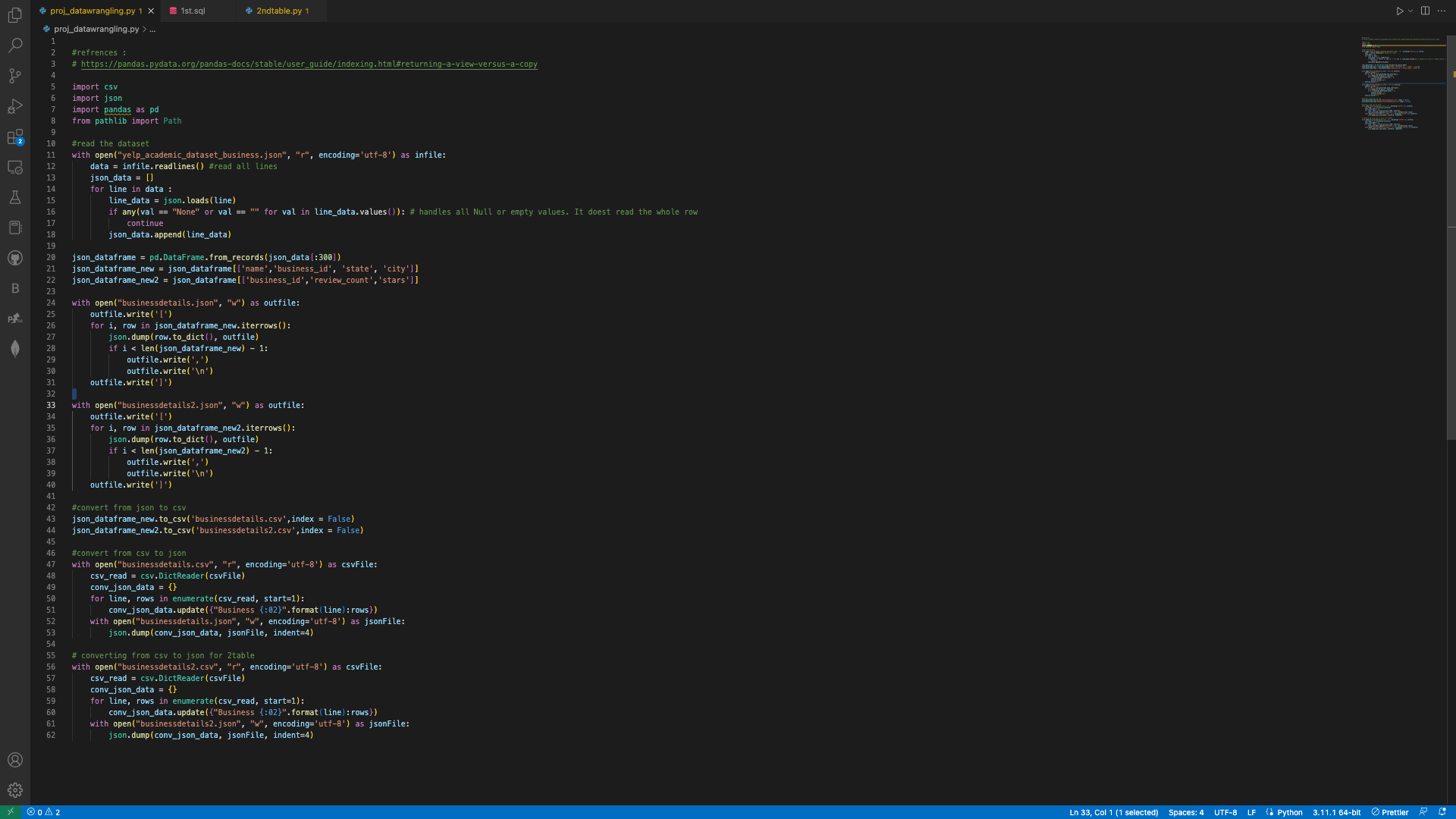
Python code for conversions:  for 2nd question.

A screenshot of a computer

Description automatically generated



 For 3rd question:



DataProc Samples : 

Graphical user interface, application, table

Description automatically generated

**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](#:~:text=Analyzing%20reviews%20left%20by%20your,what%20your%20customers%20truly%20want)  

*Convert CSV to JSON using python with new headers*. Stack Overflow. (1969, July 1). Retrieved January 27, 2023, from <https://stackoverflow.com/questions/73089933/convert-csv-to-json-using-python-with-new-headers>

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>

GeeksforGeeks. (2021, July 21). *Plot mean and standard deviation using GGPLOT2 in R*. GeeksforGeeks. Retrieved February 13, 2023, from <https://www.geeksforgeeks.org/plot-mean-and-standard-deviation-using-ggplot2-in-r/>