OSEMN Report
Midterm Project

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.

1) What is the distribution of the number of reviews for businesses in the dataset, and what can we conclude about the central tendency and spread of the data based on the mean and standard deviation?

2) How does the average and total log-scaled review count of businesses vary by state in the Yelp data?

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 issue with having little to no customer reviews is that this can negatively impact sales. Customer reviews are an important channel to attract customers and increase sales. A benefit in analyzing reviews at the business, city and state level will provide insight into which states and/or businesses have the least customer engagement and allow for proper intervention.

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

Prior to conducting the analysis for the yelp data set, the data is extracted and consolidated into a CSV and JSON files for the following fields: business name, id, city, state, and review count, stars.

OSEMN Report
Midterm Project

Visual studio is the primary application utilized to read and extract data in the python programming language for the first two question of interests and for the 3rd question I divided the data into 2 different tables as .csv files and imported them into phpMyAdmin and wrote a SQL Query to join them and generate data to answer the 3rd question. In addition, all null values are deleted from the data set. Null values are deleted to not compromise the integrity of the analysis.

Explore Data

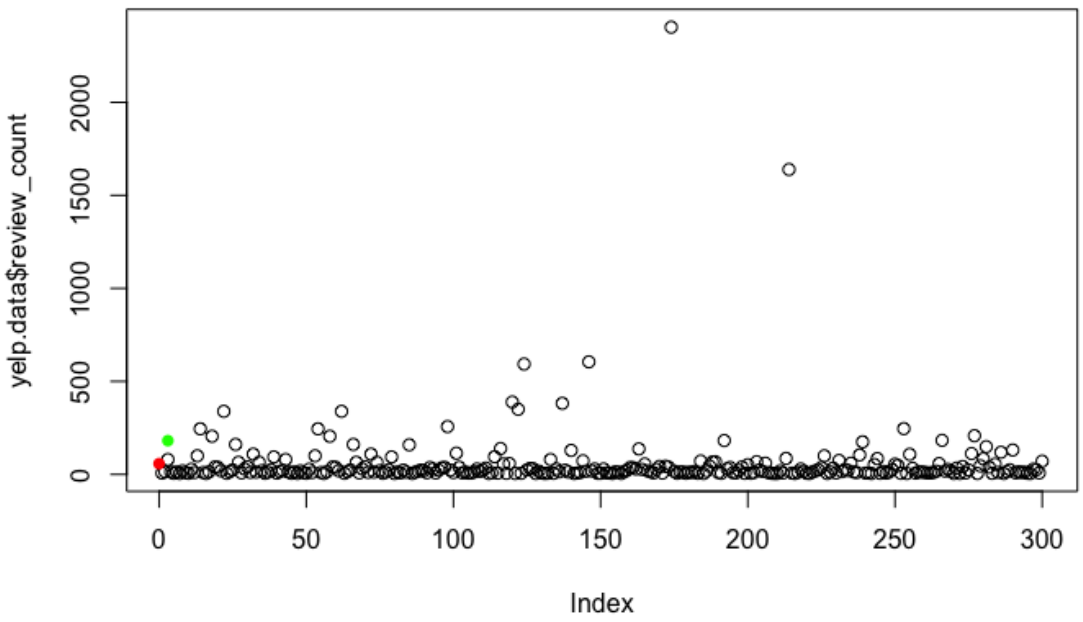
For the 1st two questions, the exploration stage, the csv data is loaded into R-studio. To begin the exploration stage the review count field is plotted.

Analysis and visualizations for 1st question:

Question of interest?

What is the distribution of the number of reviews for businesses in the dataset, and what can we conclude about the central tendency and spread of the data based on the mean and standard deviation?

The generated plot displays the distribution of the number of reviews for the dataset's businesses. Where x-axis is no of business and y axis is review_count of those business. Most of the businesses have fewer than 200 reviews, with a few outliers having more. The median number of reviews, which is around 57.67333, is represented by the red point on the plot. The standard deviation, or green point, is around 181.3557. We may infer from this plot that the majority of the dataset's firms have a relatively low number of reviews, while a few outliers have a high number. We can get a sense of the data's central tendency and distribution from the mean and standard deviation.



plot 1

OSEMN Report
Midterm Project

What is the use of finding mean and standard deviation ?

To visualize the central tendency and distribution of the data, the mean and standard deviation of the review count are computed and plotted. While the standard deviation provides an idea of how much variety there is in the review counts, the mean provides an estimate of the "average" number of reviews for businesses in the dataset. When taken as a whole, these statistics can shed light on the distribution of review counts and the degree of business fluctuation.

Analysis and visualizations for 2nd question:

Question of interest?

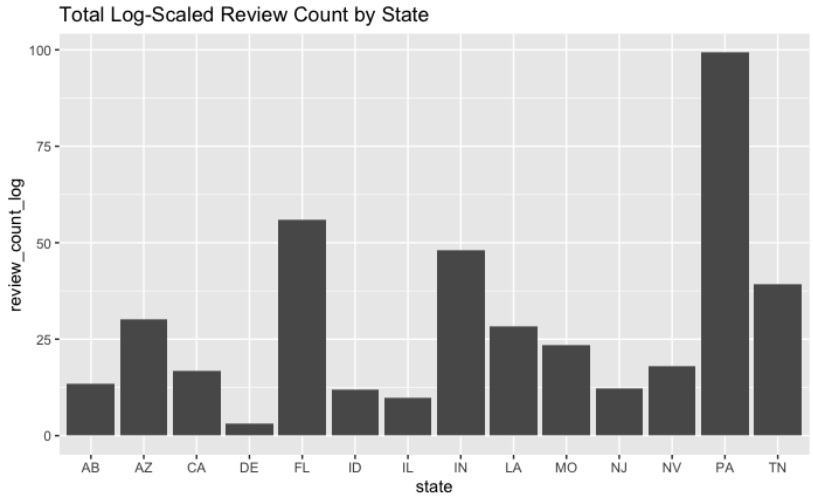
How does the average and total log-scaled review count of businesses vary by state in the Yelp data?

In this two bar charts, one is representing the total log-scaled review count of businesses by state and another representing the average log-scaled review count of businesses by state. The x-axis in both charts shows the states and the y-axis shows the log-scaled review count. By observing these plots, we can conclude that the log-scaled review count of businesses varies by state. Some states have higher total and average log-scaled review count compared to others, indicating a higher concentration of businesses with more reviews in those states.

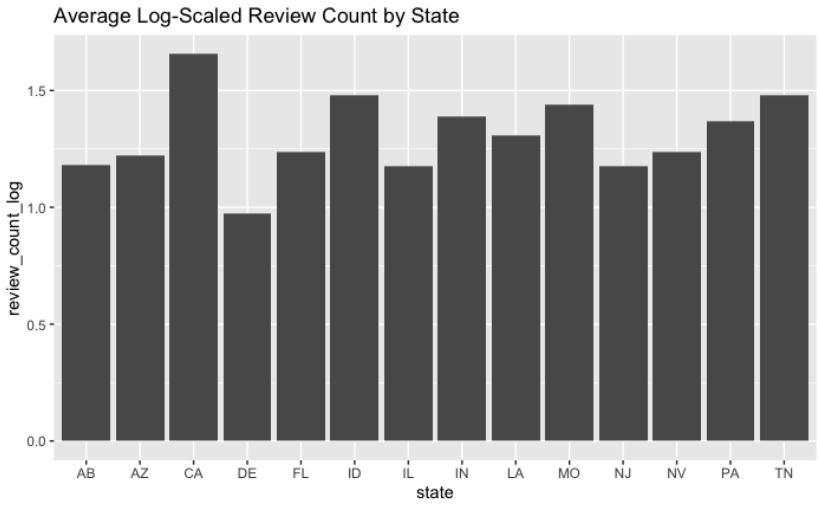
Why are we doing log-scaling here?

By using the log scaling on the review count data, it can provide a more informative representation of the distribution of review count values, especially if the original data has a skewed distribution.

The log scaling helps to compress the differences in the review count between states with a large difference in their review count, making it easier to see differences in the total review count among states with smaller review counts.



Graph 1



Graph 2

Analysis and visualizations for 3rd question:

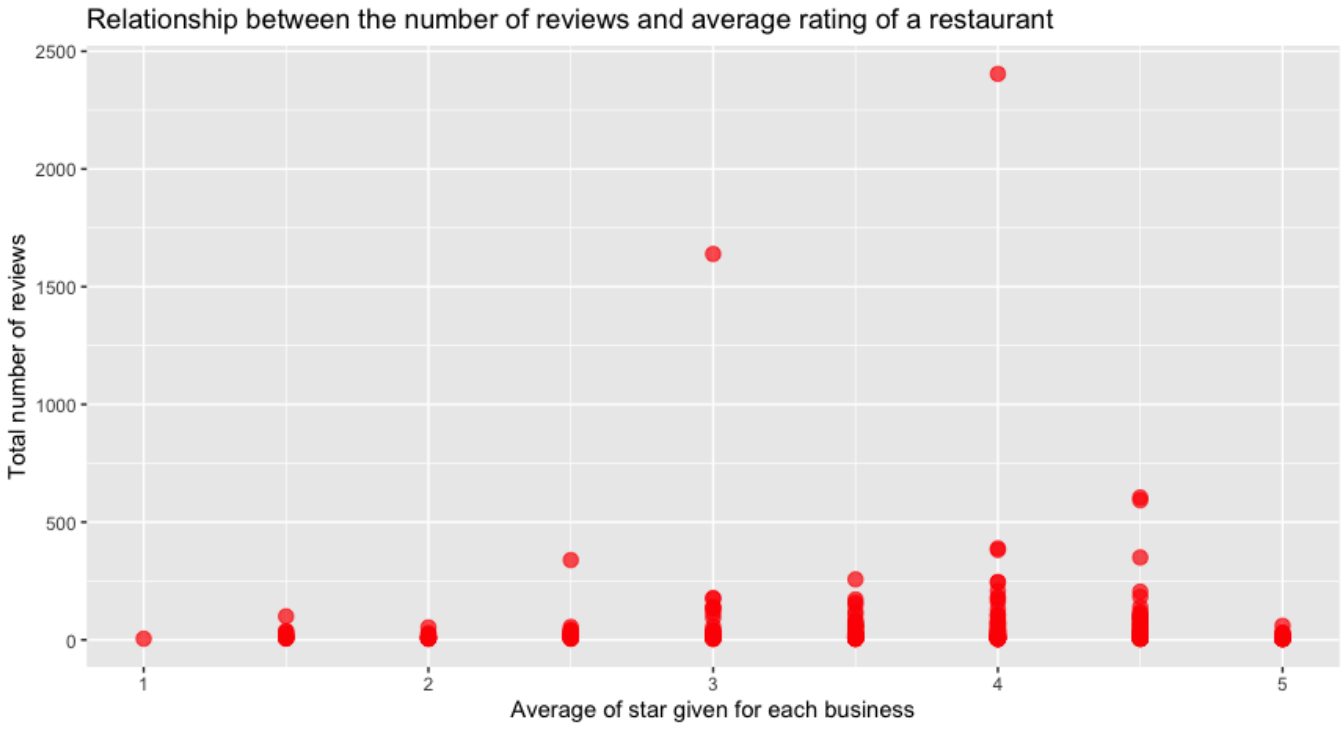
What is the relationship between the average stars of a restaurant and the number of reviews it has received?

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

For this question, you will need two data tables :

Restaurant information table - This table should contain information such as the restaurant name, id, state, city. etc.
Review information table - This table should contain information such as the stars, number of reviews, business_id for each restaurant. we will then need to join the two tables using relational database and the let's make the business_id as it is the primary key, hence you can access both the number of reviews and the average stars for each restaurant.

From there, you can create a scatter plot (plot2) showing the relationship between the number of reviews and average stars for each restaurant.



Plot 2

Hence From the generated plot, you can draw the following conclusions:

1. There doesn't seem to be a clear linear relationship between the number of reviews and the average stars of a restaurant. Some restaurants have many reviews but a relatively low stars, while others have fewer reviews but a higher average stars.

2. There is a large spread of data points, which means that there is a lot of variability in the relationship between the number of reviews and the average stars of restaurants.

3. There is a cluster of points near the 4.0 – 4.5-star mark with a high number of reviews. This suggests that there may be a relationship between high numbers of reviews and a high average stars.

Overall, the graph suggests that while there may be some correlation between the number of reviews and the average stars of a restaurant, it is not a strong or straight forward relationship.

Therefor we cannot still completely answer the question without finding the correlation hence the correlation coefficient can be found in r using the formulae :

correlation ← cor(data$ReviewCount, data$AverageStars)

correlation = 0.03157684

But why are we finding the correlation here?

Because the correlation coefficient measures the strength and direction of a linear relationship between two variables. In the context of the relationship between the number of reviews and the average stars of a restaurant, the correlation coefficient can help you determine if there is a strong positive, strong negative, or weak relationship between these two variables.

1. If the correlation coefficient is positive and close to 1, this suggests that there is a strong positive relationship between the number of reviews and the average rating of a restaurant, meaning that as the number of reviews increases, the average rating tends to increase as well.

2. On the other hand, if the correlation coefficient is negative and close to -1, this suggests that there is a strong negative relationship between the number of reviews and the average rating, meaning that as the number of reviews increases, the average rating tends to decrease.

3. If the correlation coefficient is close to 0, this suggests that there is no clear relationship between the number of reviews and the average rating.

Knowing the correlation coefficient can help you make informed conclusions about the relationship between these two variables and can inform future decisions, such as how to allocate resources to improve a restaurant's rating.

so, as we got the correlation coefficient as 0.03157684 which is considered a very weak or close to no linear relationship between two variables. This means that there is not a strong relationship between the number of reviews and the average stars of a restaurant, and that changes in one variable do not consistently predict changes in the other variable.

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 on the generated plot of the number of reviews for businesses in the Yelp dataset, we can observe that most of the businesses have a relatively low number of reviews, with only a few outliers having significantly higher numbers of reviews. The median number of reviews is approximately 57.67, indicating that half of the businesses have fewer reviews than this value, while the other half has more. The standard deviation, which is around 181.36, implies that there is considerable variation in the review counts of businesses in

the dataset. By using the mean and standard deviation, we can visualize the central tendency and distribution of the data, which can help us understand the degree of business fluctuation in the Yelp dataset.

From the second analysis, we can interpret that the log-scaled review count of businesses varies by state and some states have higher total and average log-scaled review counts than others. The use of log scaling on the review count data has allowed for a better representation of the distribution of review count values, which helps in identifying states with high and low review counts. The bar charts have provided a visual representation of the data that is easy to understand and interpret, making it easier to identify patterns in the data. Overall, this analysis has helped us to gain insights into how the review count of businesses varies by state in the Yelp data.

In the 3rd analysis, we can interpret as the correlation coefficient is near to zero which means that it indicates a very weak positive relationship between the two variables, meaning that as the number of reviews a restaurant has received increases, its average stars also slightly increase. However, this relationship is not strong enough to conclude that restaurants with more reviews have a higher average stars. Therefore, it can be concluded that there is no significant relationship between the average stars of a restaurant and the number of reviews it has received.

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:

Python code for conversions: for 1st and 2nd question.

```python
import csv
import json
import pandas as pd
from pathlib import Path
                    ----

#read the dataset
with open("yelp_academic_dataset_business.json", "r", encoding='utf-8') as infile:
    data = infile.readlines() #read all lines
    json_data = []
    for line in data :
        line_data = json.loads(line)
        # handles all Null or empty values. It doest read the whole row
        if any(val == "None" or val == "" for val in line_data.values()):
            continue
        json_data.append(line_data)

json_dataframe = pd.DataFrame.from_records(json_data[:300])
json_dataframe_new = json_dataframe[['name', 'state', 'city', 'review_count']]

with open("format_json.json", "w") as outfile:
    outfile.write('[')
    for i, row in json_dataframe_new.iterrows():
        json.dump(row.to_dict(), outfile)
        if i < len(json_dataframe_new) - 1:
            outfile.write(',')
            outfile.write('\n')
    outfile.write(']')

#convert from json to csv
json_dataframe_new.to_csv('format.csv',index = False)

#convert from csv to json
with open("format.csv", "r", encoding='utf-8') as csvFile:
    csv_read = csv.DictReader(csvFile)
    conv_json_data = {}
    for line, rows in enumerate(csv_read, start=1):
        conv_json_data.update({"Business {:02}".format(line):rows})
    with open("format.json", "w", encoding='utf-8') as jsonFile:
        json.dump(conv_json_data, jsonFile, indent=4)

```

OSEMN Report
Midterm Project

For 3rd question:

proj_datawrangling.py 1 | 1st.sql | 2ndtable.py 1

proj_datawrangling.py > ...

```python

#refrences :
# https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy

import csv
import json
import pandas as pd
from pathlib import Path

#read the dataset
with open("yelp_academic_dataset_business.json", "r", encoding='utf-8') as infile:
    data = infile.readlines() #read all lines
    json_data = []
    for line in data :
        line_data = json.loads(line)
        if any(val == "None" or val == "" for val in line_data.values()): # handles all Null or empty values. It doest read the whole row
            continue
        json_data.append(line_data)

json_dataframe = pd.DataFrame.from_records(json_data[:300])
json_dataframe_new = json_dataframe[['name','business_id', 'state', 'city']]
json_dataframe_new2 = json_dataframe[['business_id','review_count','stars']]

with open("businessdetails.json", "w") as outfile:
    outfile.write('[')
    for i, row in json_dataframe_new.iterrows():
        json.dump(row.to_dict(), outfile)
        if i < len(json_dataframe_new) - 1:
            outfile.write(',')
            outfile.write('\n')
    outfile.write(']')

with open("businessdetails2.json", "w") as outfile:
    outfile.write('[')
    for i, row in json_dataframe_new2.iterrows():
        json.dump(row.to_dict(), outfile)
        if i < len(json_dataframe_new2) - 1:
            outfile.write(',')
            outfile.write('\n')
    outfile.write(']')

#convert from json to csv
json_dataframe_new.to_csv('businessdetails.csv',index = False)
json_dataframe_new2.to_csv('businessdetails2.csv',index = False)

#convert from csv to json
with open("businessdetails.csv", "r", encoding='utf-8') as csvFile:
    csv_read = csv.DictReader(csvFile)
    conv_json_data = {}
    for line, rows in enumerate(csv_read, start=1):
        conv_json_data.update({"Business {:02}".format(line):rows})
    with open("businessdetails.json", "w", encoding='utf-8') as jsonFile:
        json.dump(conv_json_data, jsonFile, indent=4)

# converting from csv to json for 2table
with open("businessdetails2.csv", "r", encoding='utf-8') as csvFile:
    csv_read = csv.DictReader(csvFile)
    conv_json_data = {}
    for line, rows in enumerate(csv_read, start=1):
        conv_json_data.update({"Business {:02}".format(line):rows})
    with open("businessdetails2.json", "w", encoding='utf-8') as jsonFile:
        json.dump(conv_json_data, jsonFile, indent=4)
```

OSEMN Report
Midterm Project

CSV Samples:

This is for 1st question:

	A	B	C	D	E
1	name	state	city	review_count	
2	Abby Rappoport, LAC, CMQ	CA	Santa Barbara	7	
3	The UPS Store	MO	Affton	15	
4	Target	AZ	Tucson	22	
5	St Honore Pastries	PA	Philadelphia	80	
6	Perkiomen Valley Brewery	PA	Green Lane	13	
7	Sonic Drive-In	TN	Ashland City	6	
8	Famous Footwear	MO	Brentwood	13	
9	Temple Beth-El	FL	St. Petersburg	5	
10	Tsevi's Pub And Grill	MO	Affton	19	
11	Sonic Drive-In	TN	Nashville	10	
12	Marshalls	FL	Land O' Lakes	6	
13	Denny's	IN	Indianapolis	28	
14	Adams Dental	FL	Clearwater	10	
15	Zio's Italian Market	FL	Largo	100	
16	Tuna Bar	PA	Philadelphia	245	
17	Arizona Truck Outfitters	AZ	Tucson	10	
18	Herb Import Co	LA	New Orleans	5	
19	Nifty Car Rental	LA	Kenner	14	
20	BAP	PA	Philadelphia	205	
21	Roast Coffeehouse and Wine Bar	AB	Edmonton	40	
22	Barnes & Noble Booksellers	IN	Indianapolis	38	
23	Hibachi Express	IN	Indianapolis	20	
24	Romano's Macaroni Grill	NV	Reno	339	
25	Super Dog	TN	Nashville	6	
26	Indian Walk Veterinary Center	PA	Newtown	15	
27	H&M	CA	Santa Barbara	24	
28	The Green Pheasant	TN	Nashville	161	

format

OSEMN Report
Midterm Project

This is for the 2nd question

	name	state	city	review_count	review_count_log
1	Abby Rappoport, LAC, CMQ	CA	Santa Barbara	7	0.8450980
2	The UPS Store	MO	Affton	15	1.1760913
3	St Honore Pastries	PA	Philadelphia	80	1.9030900
4	Perkiomen Valley Brewery	PA	Green Lane	13	1.1139434
5	Sonic Drive-In	TN	Ashland City	6	0.7781513
6	Famous Footwear	MO	Brentwood	13	1.1139434
7	Temple Beth-El	FL	St. Petersburg	5	0.6989700
8	Tsevi's Pub And Grill	MO	Affton	19	1.2787536
9	Sonic Drive-In	TN	Nashville	10	1.0000000
10	Marshalls	FL	Land O' Lakes	6	0.7781513
11	Denny's	IN	Indianapolis	28	1.4471580
12	Adams Dental	FL	Clearwater	10	1.0000000
13	Zio's Italian Market	FL	Largo	100	2.0000000
14	Tuna Bar	PA	Philadelphia	245	2.3891661
15	Arizona Truck Outfitters	AZ	Tucson	10	1.0000000
16	Herb Import Co	LA	New Orleans	5	0.6989700
17	Nifty Car Rental	LA	Kenner	14	1.1461280
18	BAP	PA	Philadelphia	205	2.3117539
19	Roast Coffeehouse and Wine Bar	AB	Edmonton	40	1.6020600
20	Barnes & Noble Booksellers	IN	Indianapolis	38	1.5797836
21	Hibachi Express	IN	Indianapolis	20	1.3010300
22	Romano's Macaroni Grill	NV	Reno	339	2.5301997
23	Super Dog	TN	Nashville	6	0.7781513
24	Indian Walk Veterinary Center	PA	Newtown	15	1.1760913
25	H&M	CA	Santa Barbara	24	1.3802112
26	The Green Pheasant	TN	Nashville	161	2.2068259
27	Bar One	PA	Philadelphia	65	1.8129134
28	Domino's Pizza	TN	White House	8	0.9030900
29	Altitude Trampoline Park - Boise	ID	Boise	30	1.4771213
30	DeSandro on Main	PA	Philadelphia	41	1.6127839

OSEMN Report
Midterm Project

This is for 3rd question

A	B	C	D	E	F
business_id	review_count	stars			
Pns2l4eNsfO8kk83dixA6A	7	5			
mpf3x-BjTdTEA3yCZrAYPw	15	3			
tUFrWirKiKi_TAnsVWINQQ	22	3.5			
MTSW4McQd7CbVtyjqoe9mw	80	4			
mWMc6_wTdE0EUBKIGXDVfA	13	4.5			
CF33F8-E6oudUQ46HnavjQ	6	2			
n_0UpQx1hsNbnPUSlodU8w	13	2.5			
qkRM_2X51Yqxk3btlwAQlg	5	3.5			
k0hlBqXX-Bt0vf1op7Jr1w	19	3			
bBDDEgkFA1Otx9Lfe7BZUQ	10	1.5			
UJsufbvfyfONHeWdvAHKjA	6	3.5			
il_Ro8jwPlHresjw9EGmBg	28	2.5			
jaxMSoInw8Poo3XeMJt8lQ	10	5			
0bPLkL0QhhPO5kt1_EXmNQ	100	4.5			
MUTTqe8uqyMdBl186RmNeA	245	4			
rBmpy_Y1UbBx8ggHlyb7hA	10	4.5			
M0XSSHqrASOnhgbWDJIpQA	5	4			
8wGISYjYkE2tSqn3cDMu8A	14	3.5			
ROeacJQwBeh05Rqg7F6TCg	205	4.5			
WKMJwqnfZKsAae75RMP6jA	40	4			
qhDdDeI3K4jy2KyzwFN53w	38	4			
kfNv-JZpuN6TVNSO6hHdkw	20	4			
9OG5YkX1g2GReZM0AskizA	339	2.5			
4iRzR7OaS-QaSXuvYxEGKA	6	4			
PSo_C1Sfa13JHjzVNW6ziQ	15	5			
noByYNtDLQAra9ccqxdfDw	24	3			
tMkwHmWFUEXrC9ZduonpTg	161	4			
QdN72BWoyFypdGJhhI5r7g	65	4			
sqSqqLy0sN8n2lZrAbzidQ	8	3.5			
fvWn8oXXwbj2l79cochZyw	30	5			
Mjboz24M9NlBeiOJKLEd_Q	41	3			
8sshLb4UU7emeUDvtJWnpA	11	4.5			
kV_Q1oqis8Qli8dUoGpTyQ	109	3.5			
w_AMNol1iG9eay7ncmc67w	12	3			
aPNXGTDkf-4bjhyMBQxqpQ	65	3.5			

OSEMN Report
Midterm Project

JSON Samples :


```
{
    "Business 01": {
        "name": "Abby Rappoport, LAC, CMQ",
        "business_id": "Pns2l4eNsfO8kk83dixA6A",
        "state": "CA",
        "city": "Santa Barbara"
    },
    "Business 02": {
        "name": "The UPS Store",
        "business_id": "mpf3x-BjTdTEA3yCZrAYPw",
        "state": "MO",
        "city": "Affton"
    },
    "Business 03": {
        "name": "Target",
        "business_id": "tUFrWirKiKi_TAnsVWINQQ",
        "state": "AZ",
        "city": "Tucson"
    },
    "Business 04": {
        "name": "St Honore Pastries",
        "business_id": "MTSW4McQd7CbVtyjqoe9mw",
        "state": "PA",
        "city": "Philadelphia"
    },
    "Business 05": {
        "name": "Perkiomen Valley Brewery",
        "business_id": "mWMc6_wTdE0EUBKIGXDVfA",
        "state": "PA",
        "city": "Green Lane"
    },
    "Business 06": {
        "name": "Sonic Drive-In",
        "business_id": "CF33F8-E6oudUQ46HnavjQ",
        "state": "TN",
        "city": "Ashland City"
    },
    "Business 07": {
        "name": "Famous Footwear",
        "business_id": "n_0UpQx1hsNbnPUSlodU8w",
        "state": "MO",
        "city": "Brentwood"
    },
    "Business 08": {
        "name": "Temple Beth-El",
        "business_id": "qkRM_2X51Yqxk3btlwAQIg",
        "state": "FL",
        "city": "St. Petersburg"
    },
    "Business 09": {
        "name": "Tsevi's Pub And Grill",
        "business_id": "k0hlBqXX-Bt0vf1op7Jr1w",
        "state": "MO",
        "city": "Affton"
    },
    "Business 10": {
        "name": "Sonic Drive-In",
        "business_id": "bBDDEgkFA1Otx9Lfe7BZUQ",
        "state": "TN",
        "city": "Nashville"
    },
    "Business 11": {
        "name": "Marshalls",
        "business_id": "UJsufbvfyfONHeWdvAHKjA",
        "state": "FL",
        "city": "Land O' Lakes"
    },
    "Business 12": {
```

proj_datawrangling.py 1 | businessdetails2.json | 1st.sql | 2ndtable.py 1

businessdetails2.json > { } Business 04


```
{
    "Business 01": {
        "business_id": "Pns2l4eNsfO8kk83dixA6A",
        "review_count": "7",
        "stars": "5.0"
    },
    "Business 02": {
        "business_id": "mpf3x-BjTdTEA3yCZrAYPw",
        "review_count": "15",
        "stars": "3.0"
    },
    "Business 03": {
        "business_id": "tUFrWirKiKi_TAnsVWINQQ",
        "review_count": "22",
        "stars": "3.5"
    },
    "Business 04": {
        "business_id": "MTSW4McQd7CbVtyjqoe9mw",
        "review_count": "80",
        "stars": "4.0"
    },
    "Business 05": {
        "business_id": "mWMc6_wTdE0EUBKIGXDVfA",
        "review_count": "13",
        "stars": "4.5"
    },
    "Business 06": {
        "business_id": "CF33F8-E6oudUQ46HnavjQ",
        "review_count": "6",
        "stars": "2.0"
    },
    "Business 07": {
        "business_id": "n_0UpQx1hsNbnPUSlodU8w",
        "review_count": "13",
        "stars": "2.5"
    },
    "Business 08": {
        "business_id": "qkRM_2X51Yqxk3btlwAQIg",
        "review_count": "5",
        "stars": "3.5"
    },
    "Business 09": {
        "business_id": "k0hlBqXX-Bt0vf1op7Jr1w",
        "review_count": "19",
        "stars": "3.0"
    },
    "Business 10": {
        "business_id": "bBDDEgkFA1Otx9Lfe7BZUQ",
        "review_count": "10",
        "stars": "1.5"
    },
    "Business 11": {
        "business_id": "UJsufbvfyfONHeWdvAHKjA",
        "review_count": "6",
        "stars": "3.5"
    },
    "Business 12": {
        "business_id": "il_Ro8jwPlHresjw9EGmBg",
        "review_count": "28",
        "stars": "2.5"
    },
    "Business 13": {
        "business_id": "jaxMSoInw8Poo3XeMJt8lQ",
        "review_count": "10",
        "stars": "5.0"
    },
    "Business 14": {
        "business_id": "0bPLkL0OhhPO5kt1_EXmNQ",
```

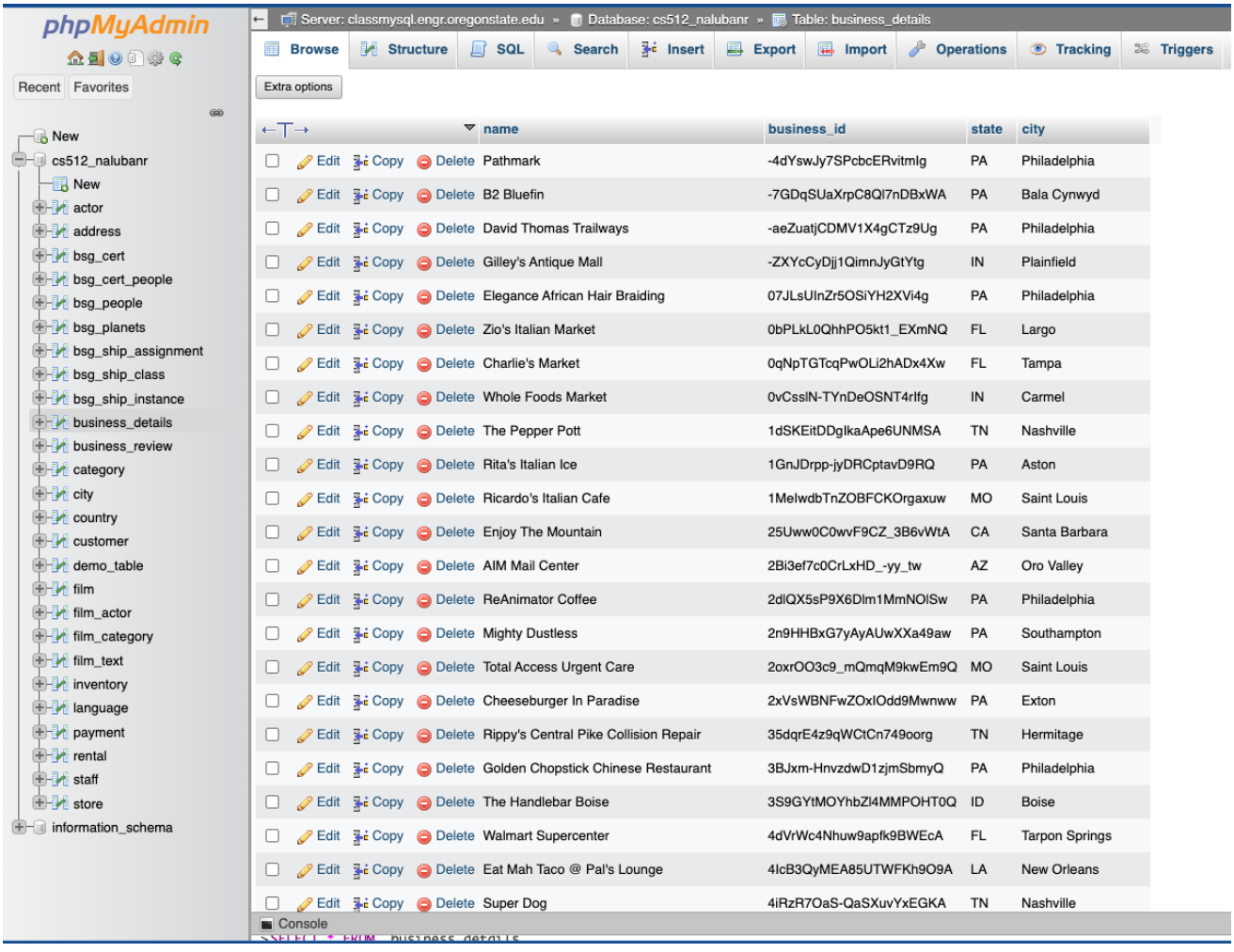
OSEMN Report
Midterm Project

SQL Query:

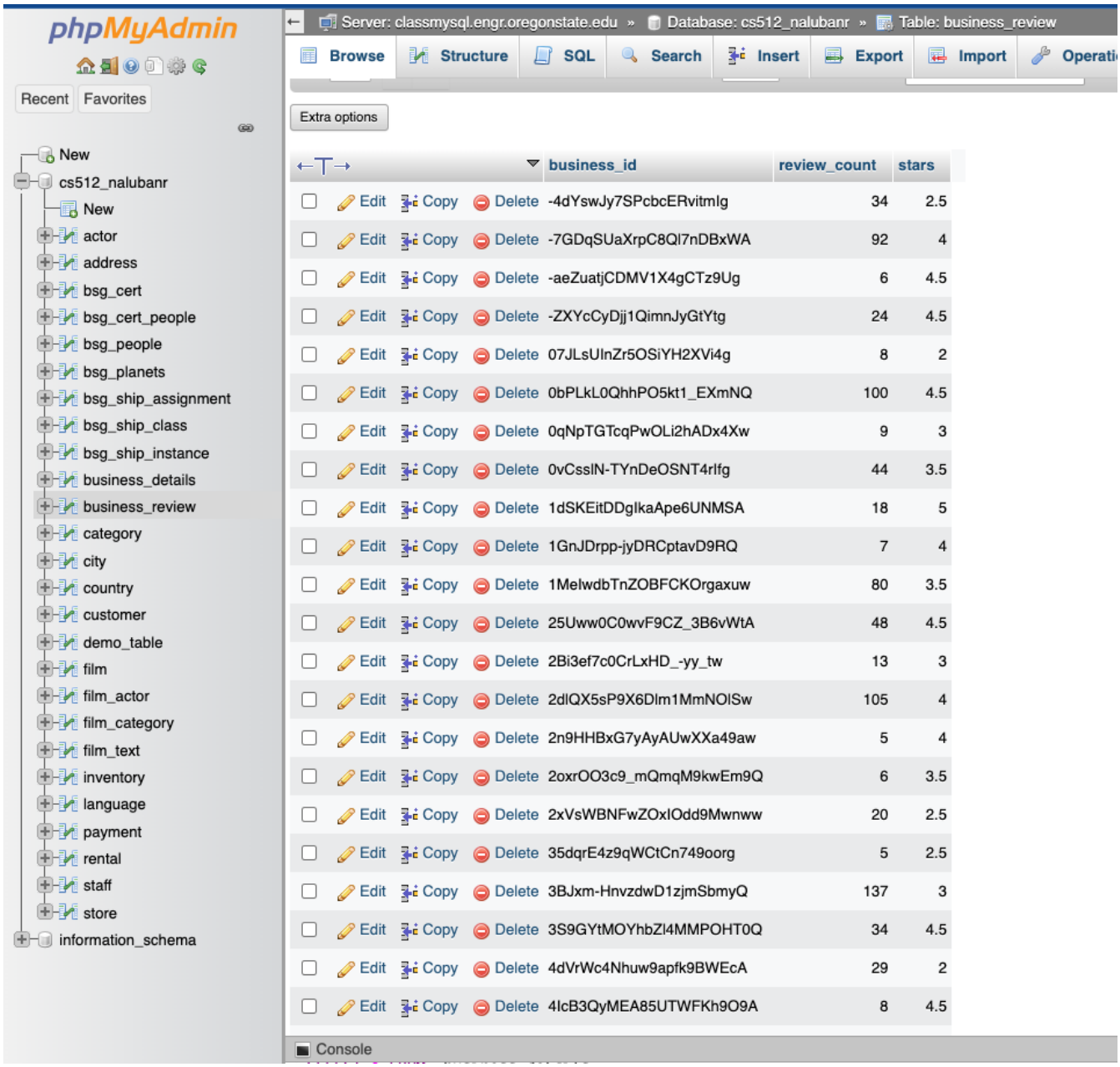
1st.sql

```sql
-- What is the relationship between the average rating of a restaurant and the number of reviews it has received?

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

SELECT D.business_id, D.name, D.state, D.city, SUM(R.review_count) AS ReviewCount, AVG(R.stars) AS AverageStars
FROM business_details AS D
JOIN business_review AS R
ON D.business_id = R.business_id
GROUP BY D.business_id, D.name, D.state, D.city;
```

phpMyAdmin: below are the 2 tables loaded into phpMyAdmin



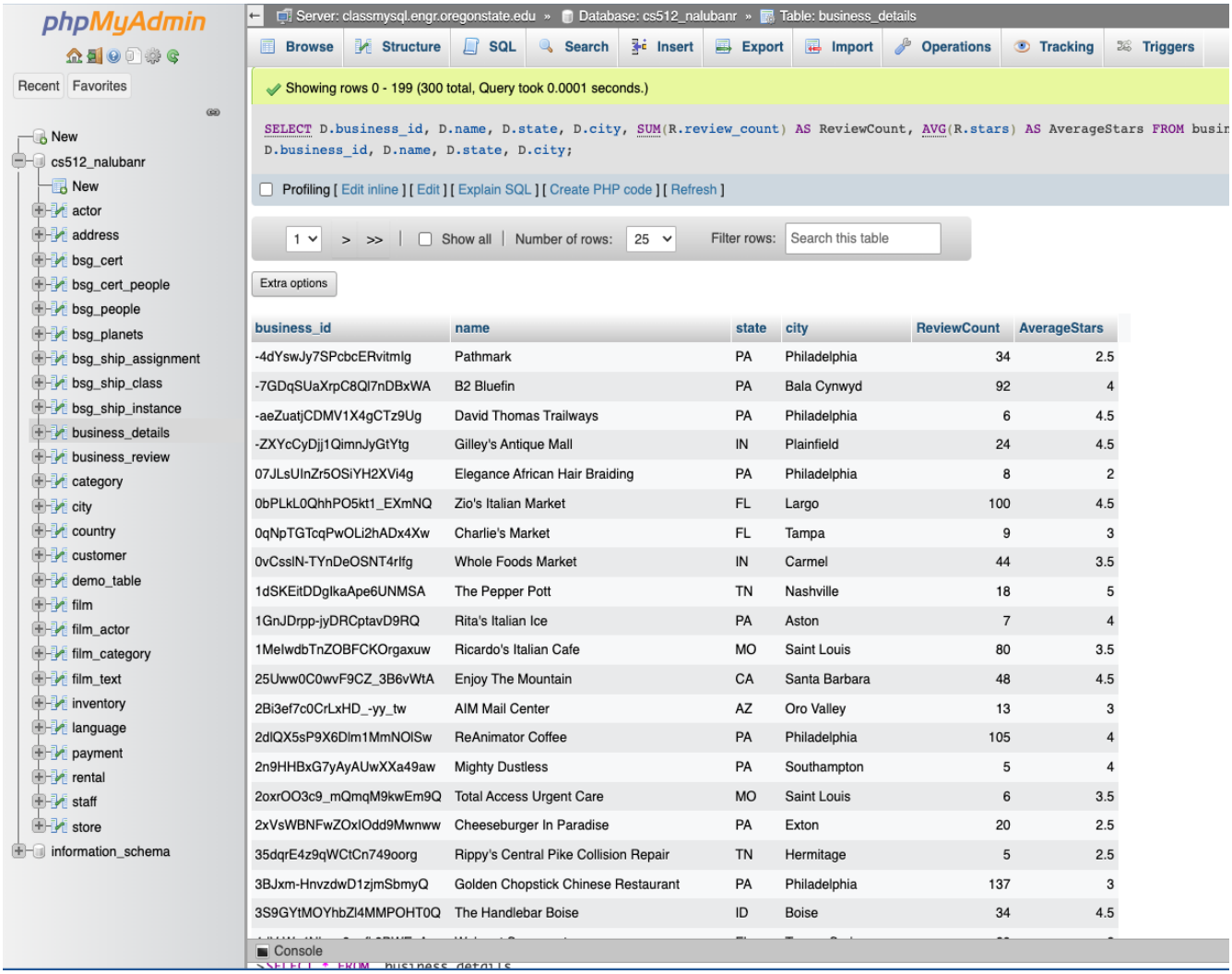
OSEMN Report
Midterm Project



OSEMN Report
Midterm Project

SQL query output:

This is the result_report generated by the query.



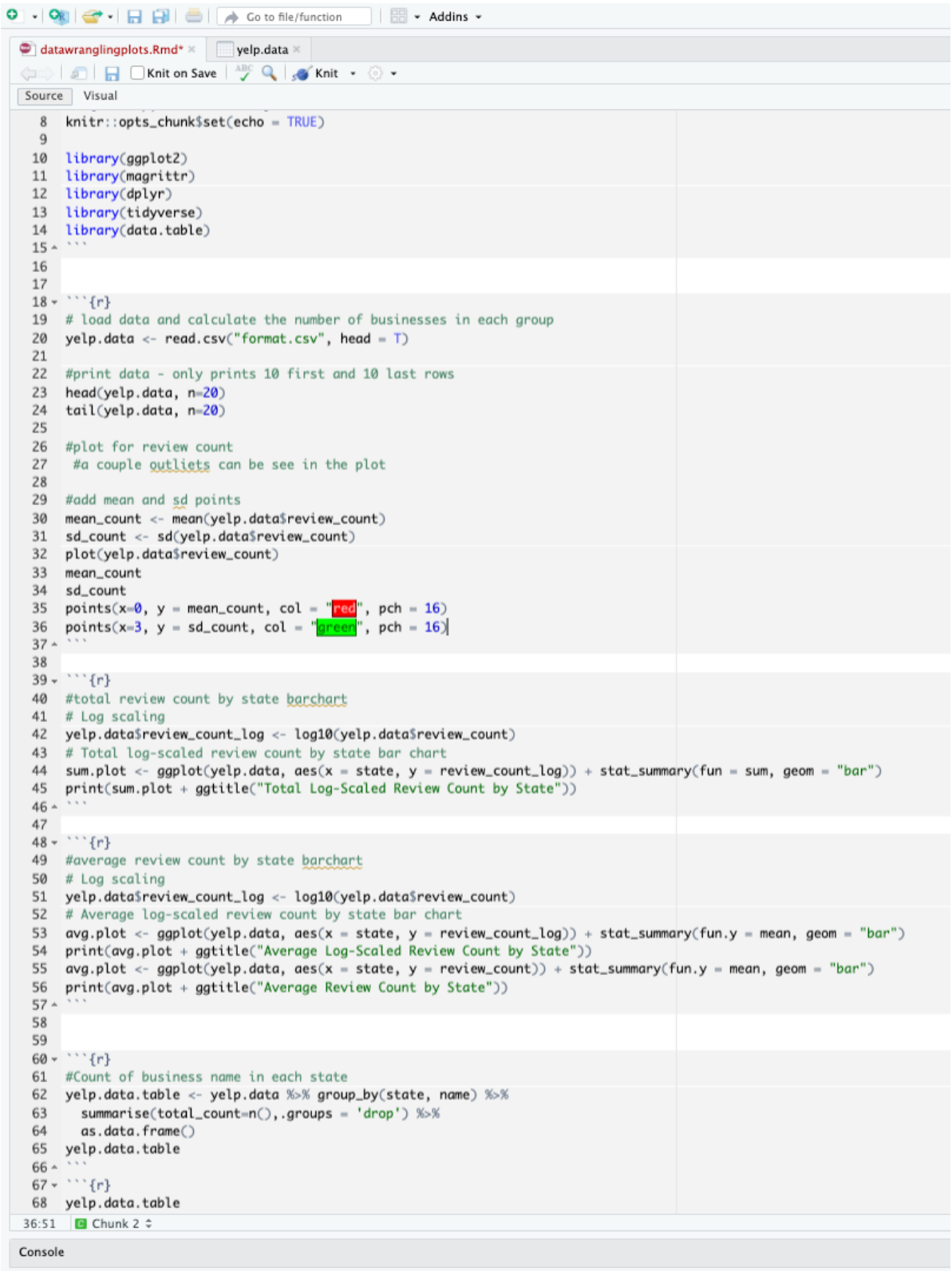
OSEMN Report
Midterm Project

This is the result_report exported as .csv.

business_id	name	state	city	ReviewCount	AverageStars
-4dYswJy7SPcbcERvitmIg	Pathmark	PA	Philadelphia	34	2.5
-7GDqSUaXrpC8Ql7nDBxWA	B2 Bluefin	PA	Bala Cynwyd	92	4
-aeZuatjCDMV1X4gCTz9Ug	David Thomas Trailways	PA	Philadelphia	6	4.5
-ZXYcCyDjj1QimnJyGtYtg	Gilley's Antique Mall	IN	Plainfield	24	4.5
07JLsUInZr5OSiYH2XVi4g	Elegance African Hair Braiding	PA	Philadelphia	8	2
0bPLkL0QhhPO5kt1_EXmNQ	Zio's Italian Market	FL	Largo	100	4.5
0qNpTGTcqPwOLi2hADx4Xw	Charlie's Market	FL	Tampa	9	3
0vCssIN-TYnDeOSNT4rIfg	Whole Foods Market	IN	Carmel	44	3.5
1dSKEitDDgIkaApe6UNMSA	The Pepper Pott	TN	Nashville	18	5
1GnJDrpp-jyDRCptavD9RQ	Rita's Italian Ice	PA	Aston	7	4
1MeIwdbTnZOBFCKOrgaxuw	Ricardo's Italian Cafe	MO	Saint Louis	80	3.5
25Uww0C0wvF9CZ_3B6vWtA	Enjoy The Mountain	CA	Santa Barbara	48	4.5
2Bi3ef7c0CrLxHD_-yy_tw	AIM Mail Center	AZ	Oro Valley	13	3
2dIQX5sP9X6Dlm1MmNOlSw	ReAnimator Coffee	PA	Philadelphia	105	4
2n9HHBxG7yAyAUwXXa49aw	Mighty Dustless	PA	Southampton	5	4
2oxrOO3c9_mQmqM9kwEm9Q	Total Access Urgent Care	MO	Saint Louis	6	3.5
2xVsWBNFwZOxlOdd9Mwnww	Cheeseburger In Paradise	PA	Exton	20	2.5
35dqrE4z9qWCtCn749oorg	Rippy's Central Pike Collision Repair	TN	Hermitage	5	2.5
3BJxm-HnvzdwD1zjmSbmyQ	Golden Chopstick Chinese Restaurant	PA	Philadelphia	137	3
3S9GYtMOYhbZl4MMPOHT0Q	The Handlebar Boise	ID	Boise	34	4.5
4dVrWc4Nhuw9apfk9BWEcA	Walmart Supercenter	FL	Tarpon Springs	29	2
4lcB3QyMEA85UTWFKh9O9A	Eat Mah Taco @ Pal's Lounge	LA	New Orleans	8	4.5
4iRzR7OaS-QaSXuvYxEGKA	Super Dog	TN	Nashville	6	4
59jkCUm4yHciKL1KW_1HJA	Roma Pizza & Italian Eatery	PA	Warrington	25	2.5
5BmQX4UVJY19mMtafMg7JA	Breadland Organic Whole Grain Bakery	AB	Edmonton	23	4
5BYQP7i7ckUlgSzX4Yhnmw	A-1 Expert Cleaning Services	PA	Levittown	5	5
6BF3ealzHtoNHSxKTFnTyg	Mister B's Cigar Box	IN	Avon	13	3.5
6bFx8j2KqPfmT0XEUS1RMg	Center City Emergency Dentist	PA	Philadelphia	120	3.5
6czqMScx8MRDirv7fxX7pw	Frankie's Fruit Bar	NV	Reno	34	4.5
6l_DA3uqOox50rBvjrQ_-g	Aloft Philadelphia Airport	PA	Philadelphia	175	3
7clCBqNbd-x2Wi96lZ6Miw	Bier Brewery and Tap Room	IN	Indianapolis	139	4.5

OSEMN Report
Midterm Project

R-Studio code : visualizations for 1st two questions.



```
knitr::opts_chunk$set(echo = TRUE)

library(ggplot2)
library(magrittr)
library(dplyr)
library(tidyverse)
library(data.table)
```

```{r}
# load data and calculate the number of businesses in each group
yelp.data <- read.csv("format.csv", head = T)

#print data - only prints 10 first and 10 last rows
head(yelp.data, n=20)
tail(yelp.data, n=20)

#plot for review count
 #a couple outliets can be see in the plot

#add mean and sd points
mean_count <- mean(yelp.data$review_count)
sd_count <- sd(yelp.data$review_count)
plot(yelp.data$review_count)
mean_count
sd_count
points(x=0, y = mean_count, col = "red", pch = 16)
points(x=3, y = sd_count, col = "green", pch = 16)
```

```{r}
#total review count by state barchart
# Log scaling
yelp.data$review_count_log <- log10(yelp.data$review_count)
# Total log-scaled review count by state bar chart
sum.plot <- ggplot(yelp.data, aes(x = state, y = review_count_log)) + stat_summary(fun = sum, geom = "bar")
print(sum.plot + ggtitle("Total Log-Scaled Review Count by State"))
```

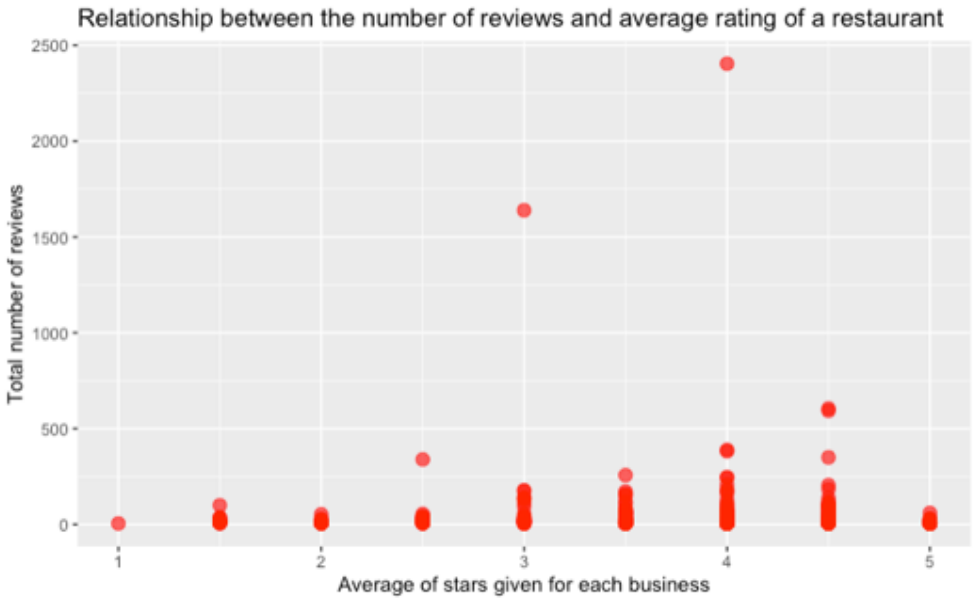
```{r}
#average review count by state barchart
# Log scaling
yelp.data$review_count_log <- log10(yelp.data$review_count)
# Average log-scaled review count by state bar chart
avg.plot <- ggplot(yelp.data, aes(x = state, y = review_count_log)) + stat_summary(fun.y = mean, geom = "bar")
print(avg.plot + ggtitle("Average Log-Scaled Review Count by State"))
avg.plot <- ggplot(yelp.data, aes(x = state, y = review_count)) + stat_summary(fun.y = mean, geom = "bar")
print(avg.plot + ggtitle("Average Review Count by State"))
```

```{r}
#Count of business name in each state
yelp.data.table <- yelp.data %>% group_by(state, name) %>%
  summarise(total_count=n(),.groups = 'drop') %>%
  as.data.frame()
yelp.data.table
```
```{r}
yelp.data.table
```

OSEMN Report
Midterm Project

This is the r-studio code for 3rd question analysis.

```
Source  Visual
4  library(ggplot2)
5  ```
6
7  ```{r}
8  # Load the data into R
9  data <- read.csv("~/Downloads/business_details (1).csv", head = T)
10 ```
11
12 ```{r}
13 # Calculate the correlation coefficient
14 correlation <- cor(data$ReviewCount, data$AverageStars)
15 cat(correlation)
16 ```

0.03157684

17
18
19 ```{r}
20 # AverageStars= It is the result of taking the average of the stars field for each business_id
21
22 ggplot(data, aes(x = AverageStars, y = ReviewCount )) +
23   geom_point(color = "red", alpha = 0.7,size = 3, shape = "circle") +
24   labs(x = "Average of stars given for each business", y = "Total number of reviews") +
25   ggtitle("Relationship between the number of reviews and average rating of a restaurant")
26 ```
```



This report has some data mixed from group project and the 3rd question is the different one and the visualizations developed for 3rd question uses 2 .csv files loaded into phpMyAdmin, and a query written in SQL is executed to relate the 2 tables and the data generated from the query is exported as .csv and loaded into r-studio for generating graph.

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

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/