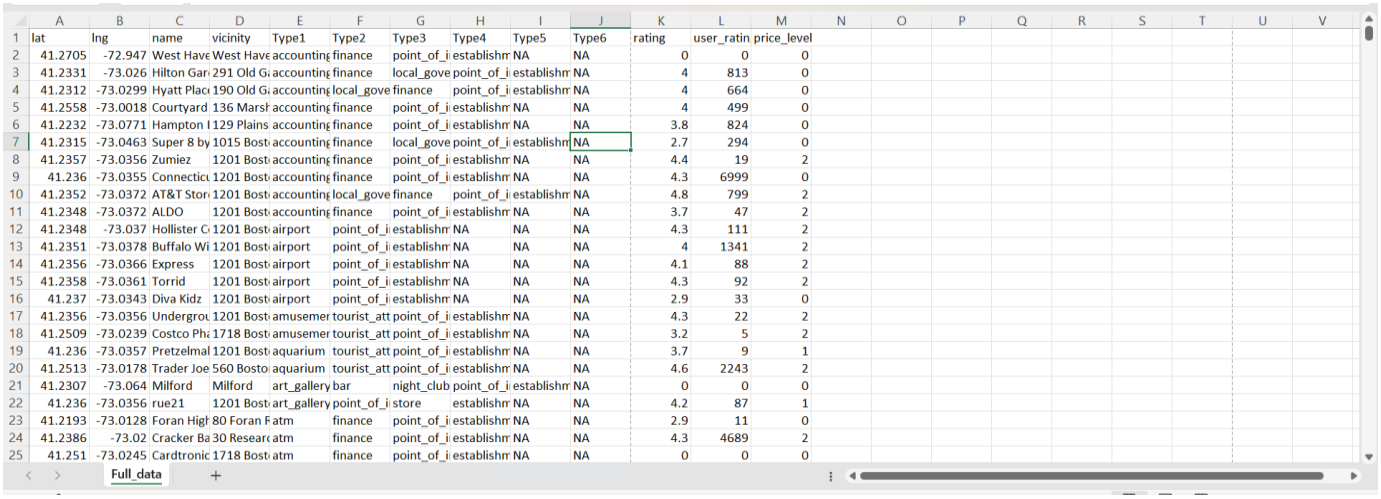
Google Maps
Team name: AMIGOS
Team Members:
Member 1: Jayanth Reddy Manda (00764288) jmand5@unh.newhaven.edu
Member 2: Snigdha Reddy Yeruva (00762316) syeru2@unh.newhaven.edu
Member 3: Jeevan Kumar Konduru (00806735) Jkond6@unh.newhaven.edu

Dataset:



	A	B	C	D	E	F	G	H	I	J	K	L	M
1	lat	lng	name	vicinity	Type1	Type2	Type3	Type4	Type5	Type6	rating	user_ratin	price_level
2	41.2705	-72.947	West Have	West Have	accounting	finance	point_of_i	establishm	NA	NA	0	0	0
3	41.2331	-73.026	Hilton Gar	291 Old G	accounting	finance	local_gove	point_of_i	establishm	NA	4	813	0
4	41.2312	-73.0299	Hyatt Plac	190 Old G	accounting	local_gove	finance	point_of_i	establishm	NA	4	664	0
5	41.2558	-73.0018	Courtyard	136 Marsh	accounting	finance	point_of_i	establishm	NA	NA	4	499	0
6	41.2232	-73.0771	Hampton I	129 Plains	accounting	finance	point_of_i	establishm	NA	NA	3.8	824	0
7	41.2315	-73.0463	Super 8 by	1015 Bost	accounting	finance	local_gove	point_of_i	establishm	NA	2.7	294	0
8	41.2357	-73.0356	Zumiez	1201 Bost	accounting	finance	point_of_i	establishm	NA	NA	4.4	19	2
9	41.236	-73.0355	Connectic	1201 Bost	accounting	finance	point_of_i	establishm	NA	NA	4.3	6999	0
10	41.2352	-73.0372	AT&T Stor	1201 Bost	accounting	local_gove	finance	point_of_i	establishm	NA	4.8	799	2
11	41.2348	-73.0372	ALDO	1201 Bost	accounting	finance	point_of_i	establishm	NA	NA	3.7	47	2
12	41.2348	-73.037	Hollister C	1201 Bost	airport	point_of_i	establishm	NA	NA	NA	4.3	111	2
13	41.2351	-73.0378	Buffalo Wi	1201 Bost	airport	point_of_i	establishm	NA	NA	NA	4	1341	2
14	41.2356	-73.0366	Express	1201 Bost	airport	point_of_i	establishm	NA	NA	NA	4.1	88	2
15	41.2358	-73.0361	Torrid	1201 Bost	airport	point_of_i	establishm	NA	NA	NA	4.3	92	2
16	41.237	-73.0343	Diva Kidz	1201 Bost	airport	point_of_i	establishm	NA	NA	NA	2.9	33	0
17	41.2356	-73.0356	Undergrou	1201 Bost	amusemer	tourist_att	point_of_i	establishm	NA	NA	4.3	22	2
18	41.2509	-73.0239	Costco Pha	1718 Bost	amusemer	tourist_att	point_of_i	establishm	NA	NA	3.2	5	2
19	41.236	-73.0357	Pretzelmal	1201 Bost	aquarium	tourist_att	point_of_i	establishm	NA	NA	3.7	9	1
20	41.2513	-73.0178	Trader Joe	560 Bosto	aquarium	tourist_att	point_of_i	establishm	NA	NA	4.6	2243	2
21	41.2307	-73.064	Milford	Milford	art_gallery	bar	night_club	point_of_i	establishm	NA	0	0	0
22	41.236	-73.0356	rue21	1201 Bost	art_gallery	point_of_i	store	establishm	NA	NA	4.2	87	1
23	41.2193	-73.0128	Foran High	80 Foran F	atm	finance	point_of_i	establishm	NA	NA	2.9	11	0
24	41.2386	-73.02	Cracker Ba	30 Researc	atm	finance	point_of_i	establishm	NA	NA	4.3	4689	2
25	41.251	-73.0245	Cardtronic	1718 Bost	atm	finance	point_of_i	establishm	NA	NA	0	0	0

Full_data

As shown in the above image we deal with multiple attributes such as:
Latitude & Longitude: Coordinates of a business so that we able to predict if user is available in available range of coordinates.
Name: Name of different businesses at coordinates.
Vicinity: Human readable location of a business.
Type: Indicates different attributes of a store such as availability of food, coffee, restaurant….
Rating: Indicates the rating of a business by the customer, it's one of the key parameter for the business model, it's one of the key parameter whether a new customer is willing to visit or not.
User rating: Number of ratings users made on business. so that if the business has more ratings with god rating it indicates good business.
Research question: Recommendation to the customer in a path from source to destination (Ex: let us Assume google maps knows user is from India, assume user is travelling from "Newhaven railway station " to "Hartford Railway Station" via car. If the user needs coffee, we recommend best coffee available stores near user based on business data available in google maps such as rating, price levels, user rating, etc....)
Univariate, Bivariate, and Multivariate Data Analysis

In a nutshell, the process of cleaning, transforming, visualizing, and analyzing the data to gain valuable insights to make more effective business decisions is known as **Data Analysis**.
Here, we will try to investigate data analysis techniques and see which techniques can be used with what kind of variables. Specifically, we will understand:

1. **Univariate Analysis**

a) **Bar chart**

b) **Count**

c) **Pie chart**

d) **Min, Max, Median, Mode, Mean**

e) **Histogram**

f) **Box Plot**

2. **Bivariate Analysis**

a) **Scatter Plot Analysis**

b) **Corelation Analysis**

To understand the definitions and the steps involved in data analysis we will import a dataset on which we will be implementing the data analysis operations on.

IMPORTING LIBRARIES

```
import numpy as np
import matplotlib.pyplot as plt
%matplotlib inline
import seaborn as sns
import math
```

Importing the Dataset

Here, we will be using the "Full_data.csv"

```
df=pd.read_csv("Full_data.csv")
```

In [3]: df

Out[3]:

	lat	lng	name	vicinity	Type1	Type2	Type3	Type4	Type5	Type6	rating	user_ratings_total	price_level
0	41.270548	-72.946971	West Haven	West Haven	accounting	finance	point_of_interest	establishment	NaN	NaN	0.0	0	0
1	41.233086	-73.026043	Hilton Garden Inn Milford	291 Old Gate Lane, Milford	accounting	finance	local_government_office	point_of_interest	establishment	NaN	4.0	813	0
2	41.231155	-73.029880	Hyatt Place Milford / New Haven	190 Old Gate Lane, Milford	accounting	local_government_office	finance	point_of_interest	establishment	NaN	4.0	664	0
3	41.255831	-73.001768	Courtyard by Marriott New Haven Orange/Milford	136 Marsh Hill Road, Orange	accounting	finance	point_of_interest	establishment	NaN	NaN	4.0	499	0
4	41.223208	-73.077087	Hampton Inn Milford	129 Plains Road, Milford	accounting	finance	point_of_interest	establishment	NaN	NaN	3.8	824	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...
2275	41.640696	-72.872586	T.J. Maxx	875 Queen Street, Southington	university	point_of_interest	establishment	NaN	NaN	NaN	4.2	482	1
2276	41.665909	-72.922815	ALDI	110 Middle Street, Bristol	university	point_of_interest	establishment	NaN	NaN	NaN	4.5	1050	1
2277	41.677566	-72.914300	Hartford County Tattoo	253 West Washington Street, Bristol	university	point_of_interest	establishment	NaN	NaN	NaN	4.8	336	0
2278	41.622782	-72.873652	VIP Very Intimate Pleasures	228 Queen Street, Southington	veterinary_care	point_of_interest	establishment	NaN	NaN	NaN	3.9	145	0
2279	41.611250	-72.901275	Target Mobile	600 Executive Boulevard South, Southington	veterinary_care	point_of_interest	establishment	NaN	NaN	NaN	4.5	6	2

2280 rows × 13 columns

Here, In the given dataset we have 13 columns and 2280 rows .
The different types of columns given here are latitude , longitude , Name of the place what we wants to search , address of the place , It is mentioned Up to 6 Different types we will get in that place , The total rating of the place – given out of 5 scale point , Number of users given rating on the place and the level of price .
Now lets get a summary of data using info method of the data frame

```
In [8]: print(df.info())
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2280 entries, 0 to 2279
Data columns (total 13 columns):
 #   Column              Non-Null Count  Dtype
---  ------              --------------  -----
 0   lat                 2280 non-null   float64
 1   lng                 2280 non-null   float64
 2   name                2280 non-null   object
 3   vicinity            2280 non-null   object
 4   Type1               2280 non-null   object
 5   Type2               2279 non-null   object
 6   Type3               1998 non-null   object
 7   Type4               1392 non-null   object
 8   Type5               401 non-null    object
 9   Type6               163 non-null    object
 10  rating              2280 non-null   float64
 11  user_ratings_total  2280 non-null   int64
 12  price_level         2280 non-null   int64
dtypes: float64(3), int64(2), object(8)
memory usage: 231.7+ KB
None
```

We can see that the data frame has 2280 entries and 13 columns, Non-Null count for the given entries and the Datatype of the entries.
The Datatype entries are In float64, int64 and object.

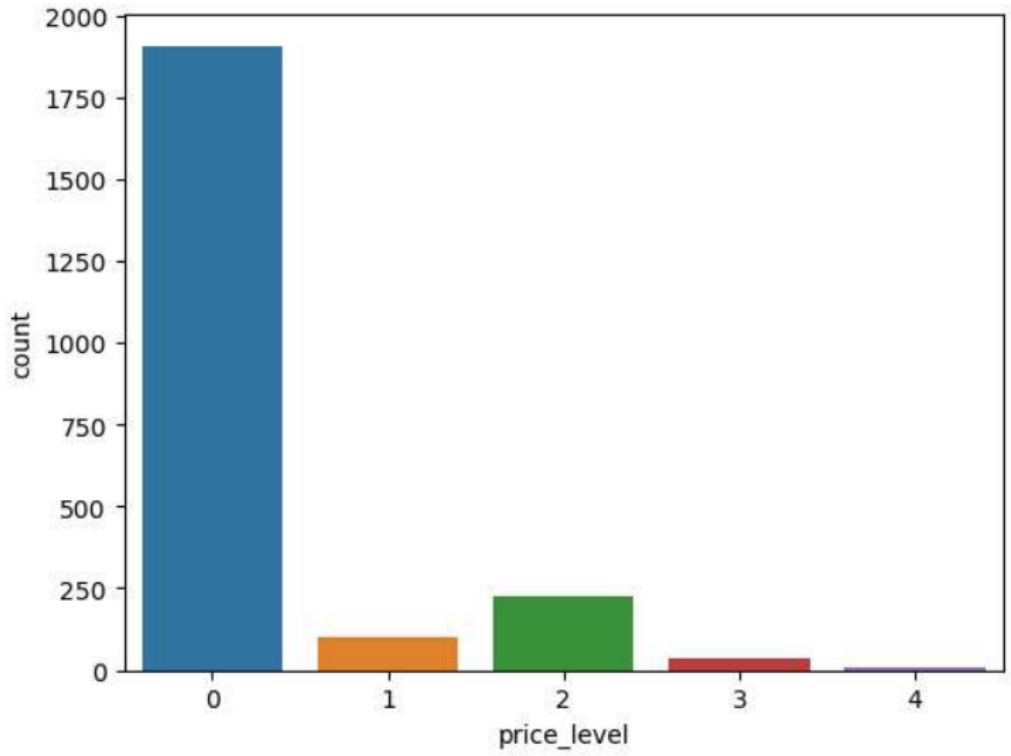
Univariate analysis: The most basic form of the data analysis technique. When we want to understand the data contained by only one variable and don't want to deal with the causes or effect relationships then a Univariate analysis technique is used.

Univariate Analysis of Categorical Variables

Bar Chart:

```
sns.countplot(df.price_level)
```

```
Out[20]: <AxesSubplot:xlabel='price_level', ylabel='count'>
```



Here, in this Bar Graph X-Axis refers to Price Level and Y-Axis refers to Count with 0,1,2,3,4 price levels .
The data set describes maximum amount of stores belongs to price 0 level.

Count:

```
import pandas as pd
df=pd.read_csv("Full_data.csv")
test= df.groupby(['price_level','name'])
test.size()
```

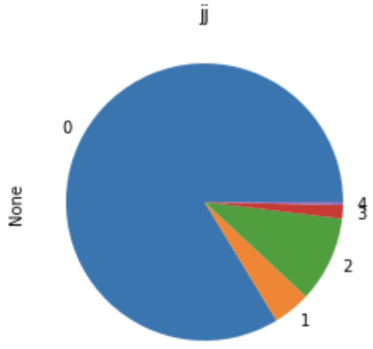
```
price_level   name
0             100 Great Meadow Rd Associates     1
              200 Fountain                       1
              2400 Degrees F                     2
              360 State                          8
              3PL Worldwide,Äé                   2
                                                ..
3             Whole Foods Market                 1
              ZINC New Haven                     1
4             Lux Bond & Green                   1
              The Capital Grille                 3
              Union League Cafe                  4
Length: 1159, dtype: int64
```

Categorical analysis describes the uniqueness of the character such as its frequency in the dataset.as the above dataset describes frequency of available same store names with particular price_level.(Ex: Pricel_level 0 consists of 8 “360 state” stores in all the locations of the dataset).

Pie Chart:

```
import matplotlib.pyplot as plt
import pandas as pd
df =  pd.read_csv('Full_data.csv')
a = df.groupby('price_level').size()
a.plot.pie(figsize=(4,4))
```

```
<Axes: title={'center': 'jj'}, ylabel='None'>
```



The above pie chart describes distribution of price levels across the data. Price level 0 has maximum amount of data.

Univariate Analysis of continuous Variables :
We will do the **univariate analysis of continuous variables**. We will first use the describe function to get the descriptive statistics of continuous variables.
Numerical Analysis:
Min, Max, Median, Mode, Mean:

```
df[['rating','user_ratings_total','price_level']].describe()
```

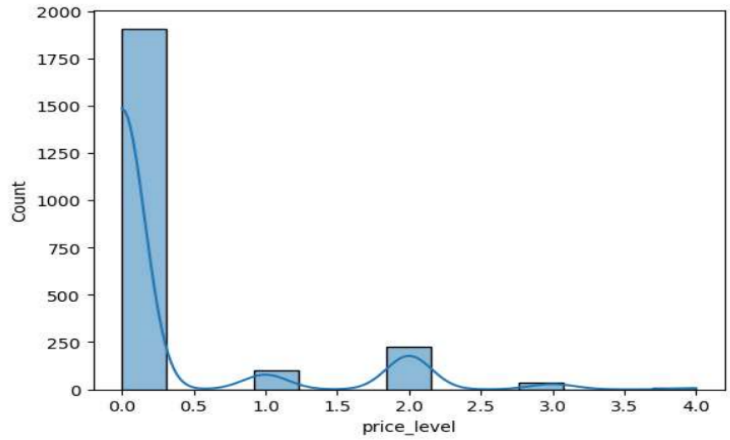
	rating	user_ratings_total	price_level
count	2280.000000	2280.000000	2280.000000
mean	2.746272	344.301316	0.305263
std	2.029706	918.820560	0.741054
min	0.000000	0.000000	0.000000
25%	0.000000	0.000000	0.000000
50%	3.800000	13.000000	0.000000
75%	4.500000	242.000000	0.000000
max	5.000000	13335.000000	4.000000

By using the describe function on the selected columns, we get the count, mean, std, min, max, 25th percentile, 50th percentile, and 75% percentile values of the columns for ratings , user ratings total and price level .

Histogram:
Now we will plot histograms for continuous columns to see the frequency distribution of values of columns.

```
sns.histplot(df.price_level,kde=True)
```

```
<AxesSubplot:xlabel='price_level', ylabel='Count'>
```

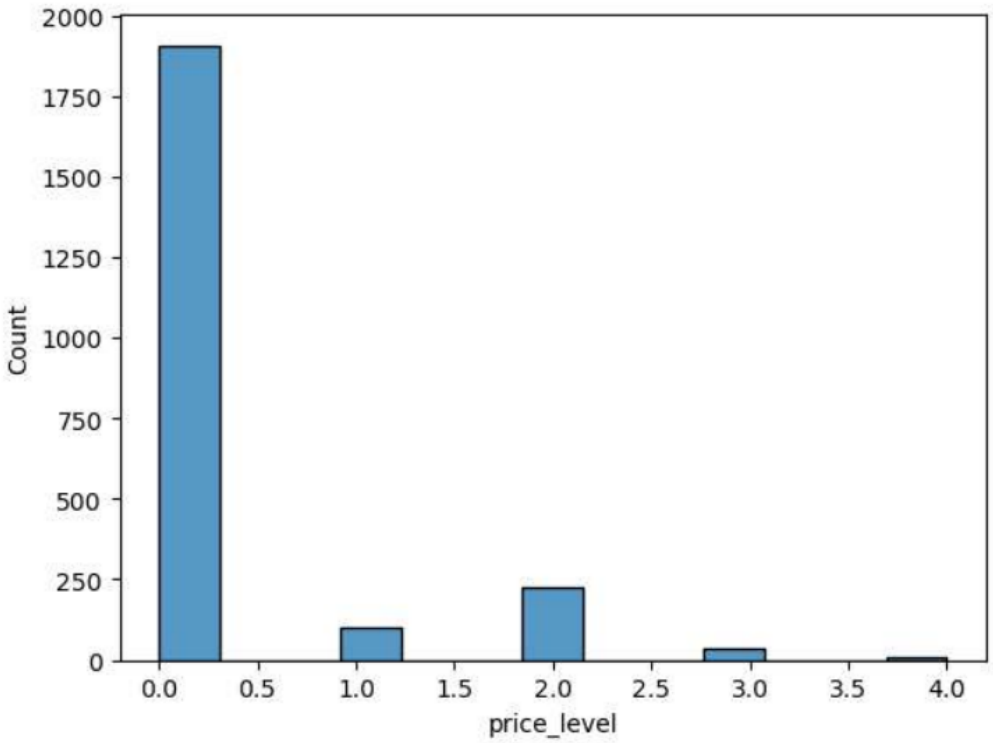


using the below line of cod

Here, In the above graph it is given the count & price level on X-Axis and Y-Axis for kde (Kernel density estimation) is true. The price level is 0.0 for more than 1750 entries, It is 1.0 level for about 100 entries, 2.0 for about 250 entries and so on….

```
In [19]: sns.histplot(df.price_level,kde=False)
```

```
Out[19]: <AxesSubplot:xlabel='price_level', ylabel='Count'>
```



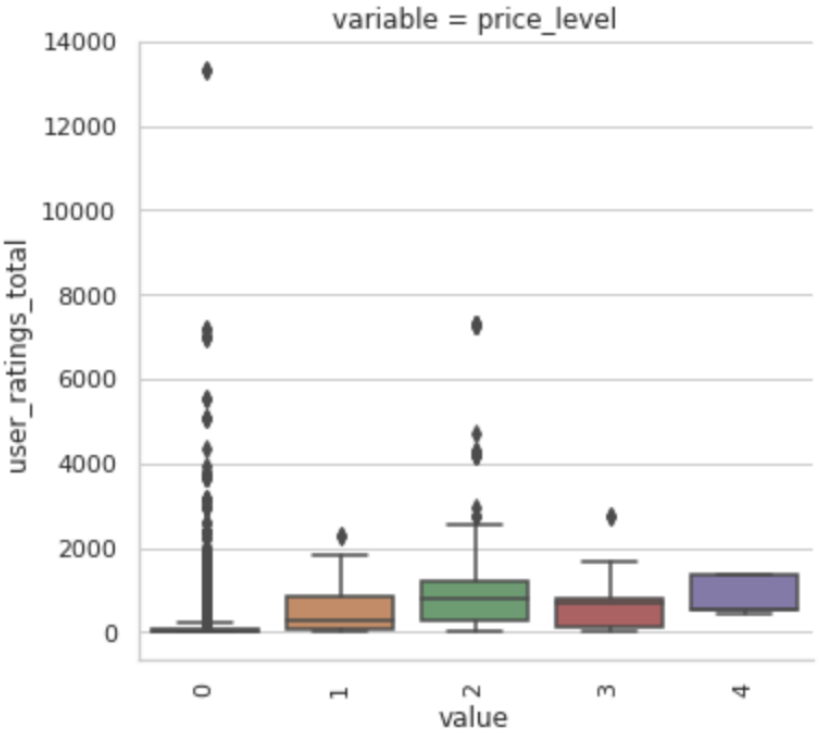
Here, In the above given graph it is given the count & price level on X-Axis and Y-Axis for kde (Kernel density estimation) is false.

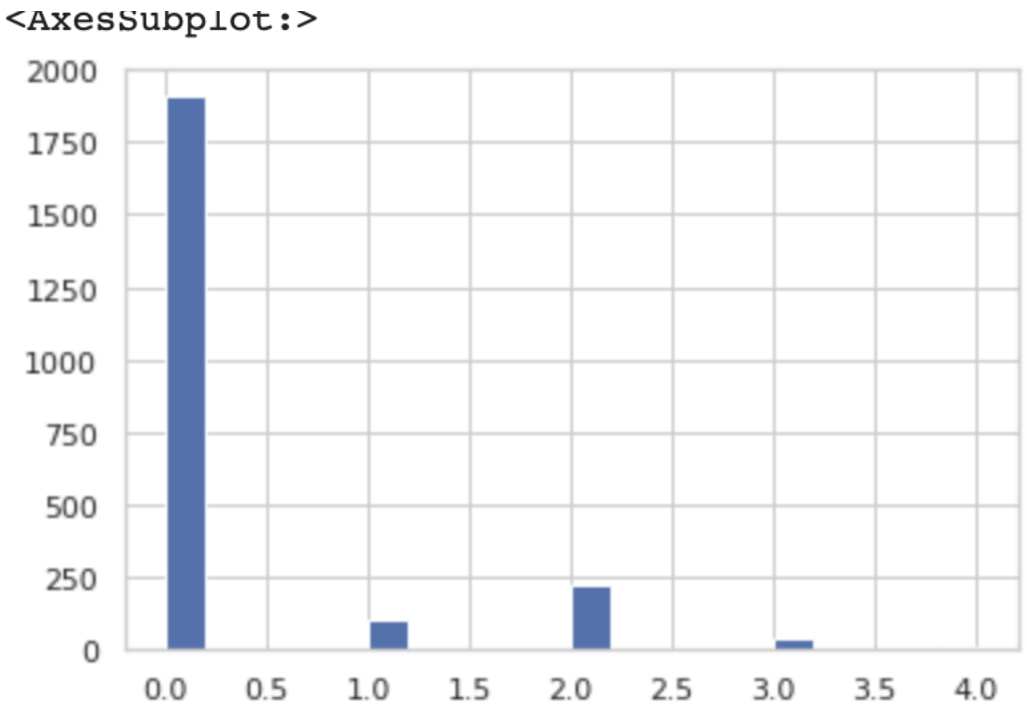
Boxplot:
A Boxplot is a graph that indicates how the values in the dataset are spread out. Boxplots are used to visualize the distribution of the data based on following parameters:

1. minimum
2. first quartile (Q1)
3. median
4. third quartile (Q3)
5. maximum

Advantages of Bar plots

1. Used to find out skewness of variables.
2. Used to find out outliers in a variable.
3. Used to find out if the data is symmetrical or not? How tightly the data is grouped?





Most of the skewness is aligned to the left side as most of the data consists of price level 0.

Bivariate Analysis

Bivariate analysis is slightly more analytical than Univariate analysis. When the data set contains two variables and researchers aim to undertake comparisons between the two data set then Bivariate analysis is the right type of analysis technique.

Bivariate Analysis of Continuous Variables:

Numerical & Numerical:

The first step in performing bivariate analysis between continuous variables would be to calculate correlations between them. Use ***corr*** function to construct the correlation matrix.

```
In [21]: df[['rating','user_ratings_total','price_level']].corr()
```

Out[21]:

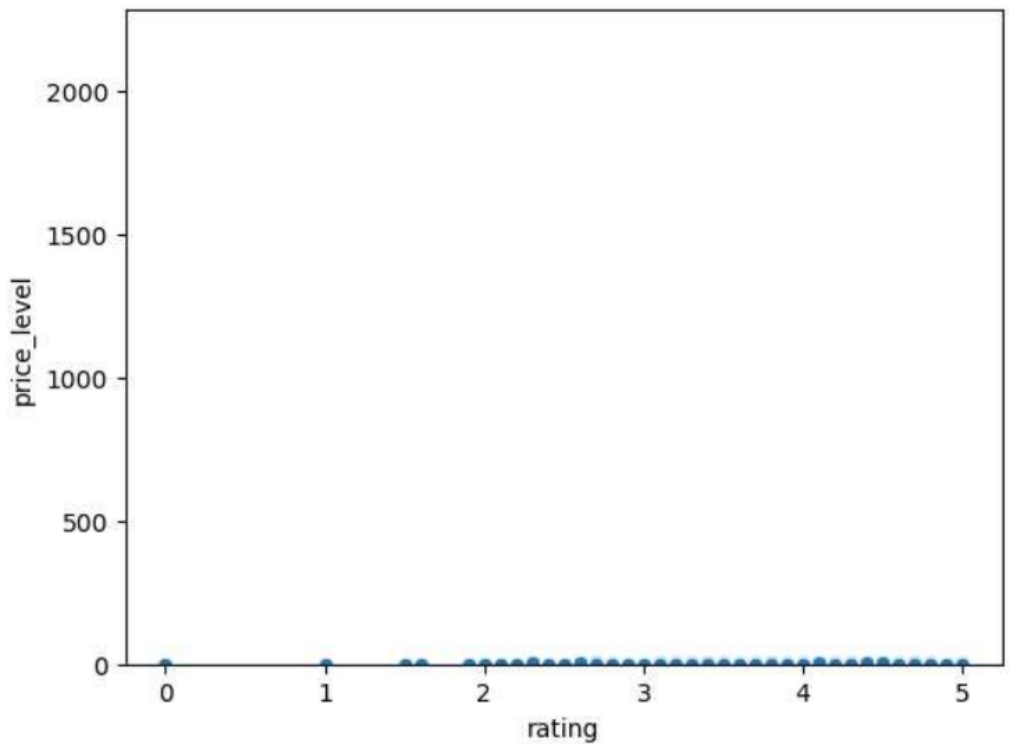
	rating	user_ratings_total	price_level
rating	1.000000	0.290899	0.295049
user_ratings_total	0.290899	1.000000	0.283290
price_level	0.295049	0.283290	1.000000

Though in this dataset, we don't see any strong correlation between any two continuous variables, in some datasets, continuous variables could be strongly correlated and the values of one might depend on others.

We can also draw line plots and scatterplots to see a relation between the two continuous variables.

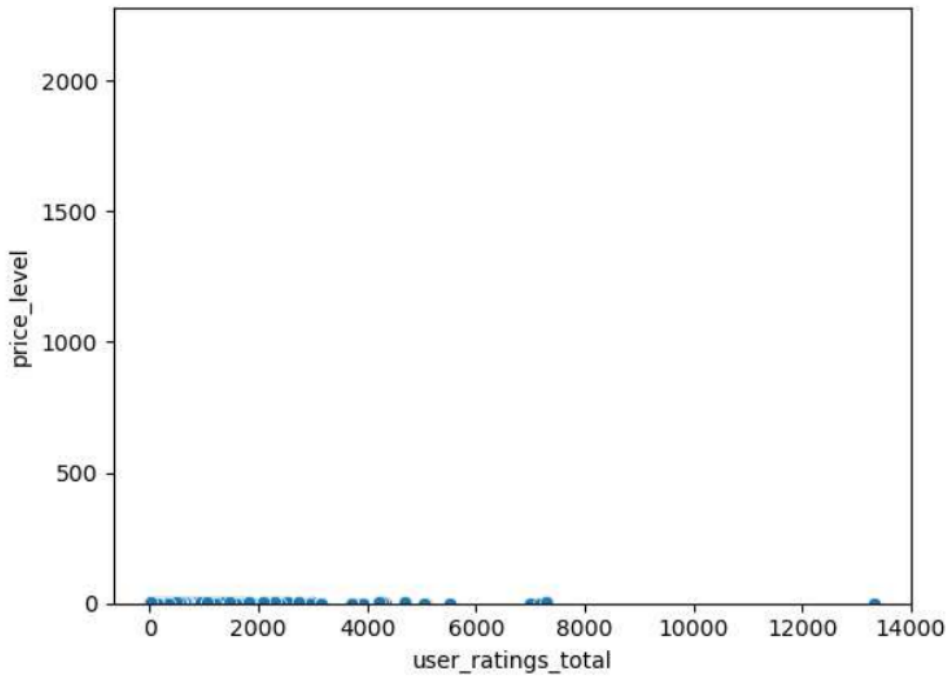
```
sns.scatterplot(df.rating,df.price_level)
plt.ylim(0,2280)
```

Out[22]: (0.0, 2280.0)



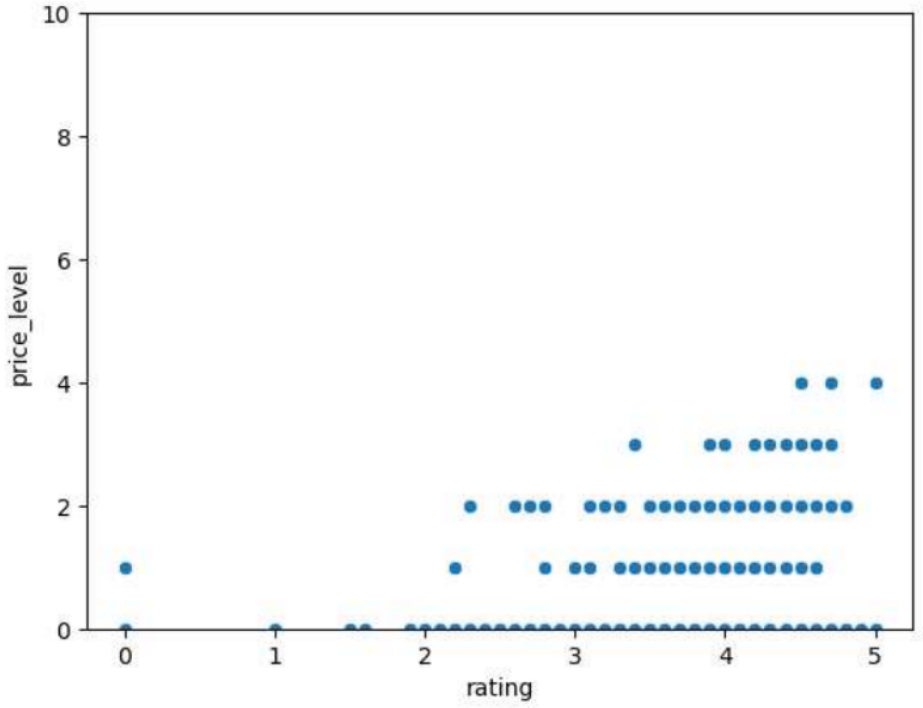
```
sns.scatterplot(df.user_ratings_total,df.price_level)
plt.ylim(0,2280)
```

Out[23]: (0.0, 2280.0)




```
sns.scatterplot(df.rating,df.price_level)
plt.ylim(0,10)
```

```
Out[24]: (0.0, 10.0)
```

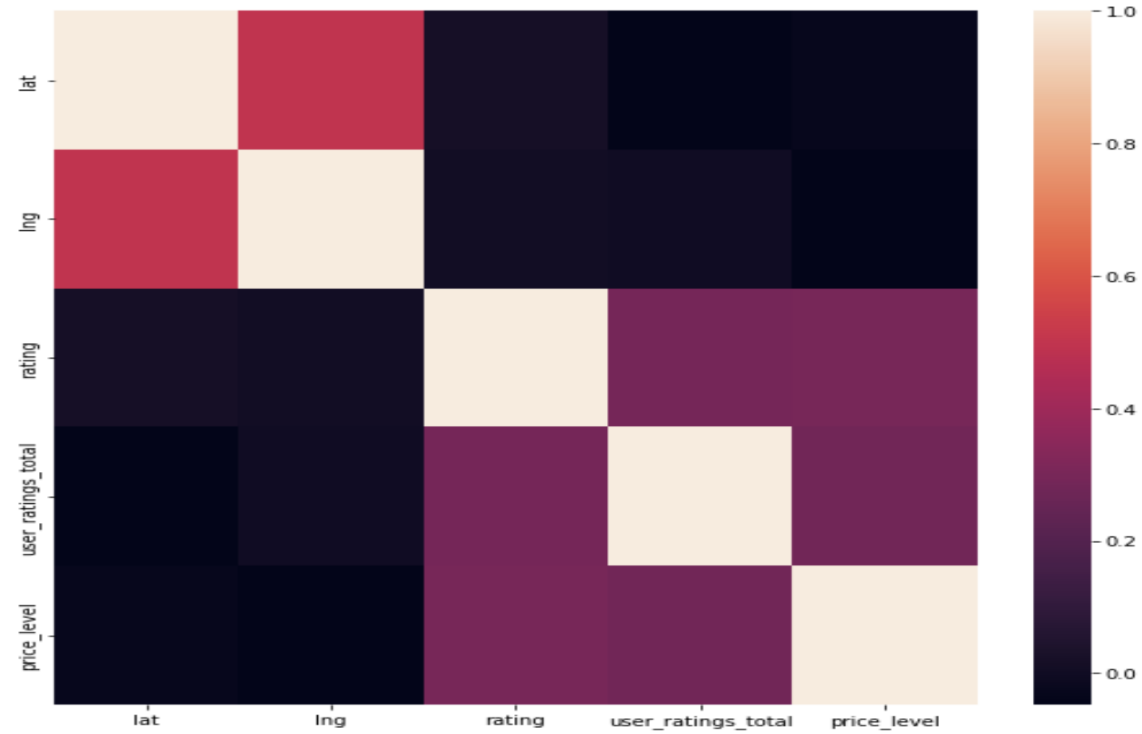


Here, the above scatter plots describes the rating to the price level.it each price level with different range of ratings.as we can find outlier such as price level 4 consists of high user ratings which can be considered as outlier.

Corelation Analysis :

```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
dataset = pd.read_csv('Full_data.csv')
numeric_data = dataset.select_dtypes(include=[np.number])
categorical_data = dataset.select_dtypes(exclude=[np.number])
corr = numeric_data.corr()
plt.figure(figsize=(10, 10))
sns.heatmap(corr)
print (corr['price_level'].sort_values(ascending=False)[:10]) #top 10 correlations
#print (corr['price_level'].sort_values(ascending=False)[-5:]) #least 5 correlations
```

```
price_level            1.000000
rating                 0.295049
user_ratings_total     0.283290
lat                   -0.029785
lng                   -0.046759
Name: price_level, dtype: float64
```



A correlation heatmap is a graphical representation of a correlation matrix representing the correlation between different variables. The value of correlation can take any value from 0 to Correlation between two random variables or bivariate data does not necessarily imply a causal relationship.

Conclusion :

At long last, we can say that exploratory information investigation is a demonstrated procedure that can help to handle on complex datasets. By utilizing representations and different strategies, we can uncover examples and connections that you probably won't have seen as in any case. We are able to find most of the stores present in the google maps have price level 0 with multiple user ratings, price level 4 ca be considered as exceptional with least rating and high price level which may not be recommended to the customer. We can find the count of each store with particular price level which can be useful that most of the same named stores have same price level.