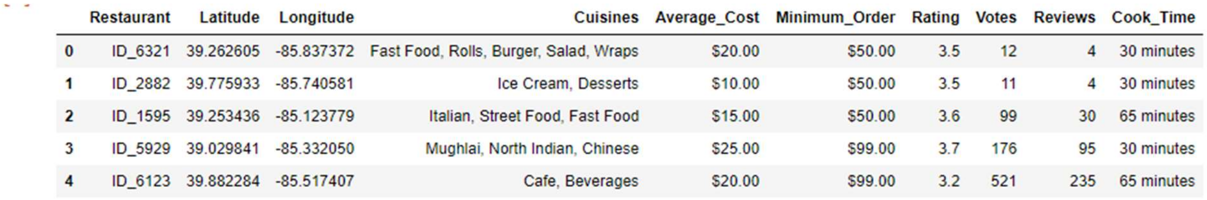
XTERN_2021: Data Science

Introduction: -

As a Data Science intern and part of a team, I was given a task to predict and analyze 4 emerging patterns of the FoodieX application. The FoodieX application was developed by other teams and I was responsible to do the analysis of the pattern which can lead and help to make better decisions for the company. The data was given in .CSV format and it has 2019 Rows. It has a column including Restaurant_ID, Latitude, Longitude, different cuisines, Average_Cost, Minimum_Order, Rating, Votes, Reviews, and Cook_Time of different Restaurants. I used python to draw my 4 findings. I have visualized them with some of the packages.

Data Cleaning: -

The original data looked like the below: -



	Restaurant	Latitude	Longitude	Cuisines	Average_Cost	Minimum_Order	Rating	Votes	Reviews	Cook_Time
0	ID_6321	39.262605	-85.837372	Fast Food, Rolls, Burger, Salad, Wraps	$20.00	$50.00	3.5	12	4	30 minutes
1	ID_2882	39.775933	-85.740581	Ice Cream, Desserts	$10.00	$50.00	3.5	11	4	30 minutes
2	ID_1595	39.253436	-85.123779	Italian, Street Food, Fast Food	$15.00	$50.00	3.6	99	30	65 minutes
3	ID_5929	39.029841	-85.332050	Mughlai, North Indian, Chinese	$25.00	$99.00	3.7	176	95	30 minutes
4	ID_6123	39.882284	-85.517407	Cafe, Beverages	$20.00	$99.00	3.2	521	235	65 minutes

First of all, I looked deep into the data frame to find various mistakes. In some of the columns I found missing values, some of the columns including "Rating", "Votes", and "Reviews" had values like "-" and "NEW". So I replaced the values "-" with "0" and other values were removed for data consistency.

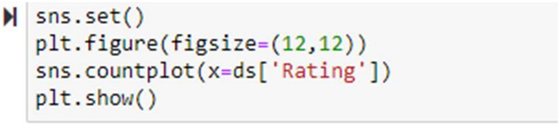
I also found the symbol like "$" in the "Average cost" and Minimum_Order" column. I have replaced them with float values. In the "Cook_Time", there was a "min" written after every value. I also have changed that value.

In the "Cuisines" section I changed the object as a list for descriptive analysis and better visualization of each cuisine.

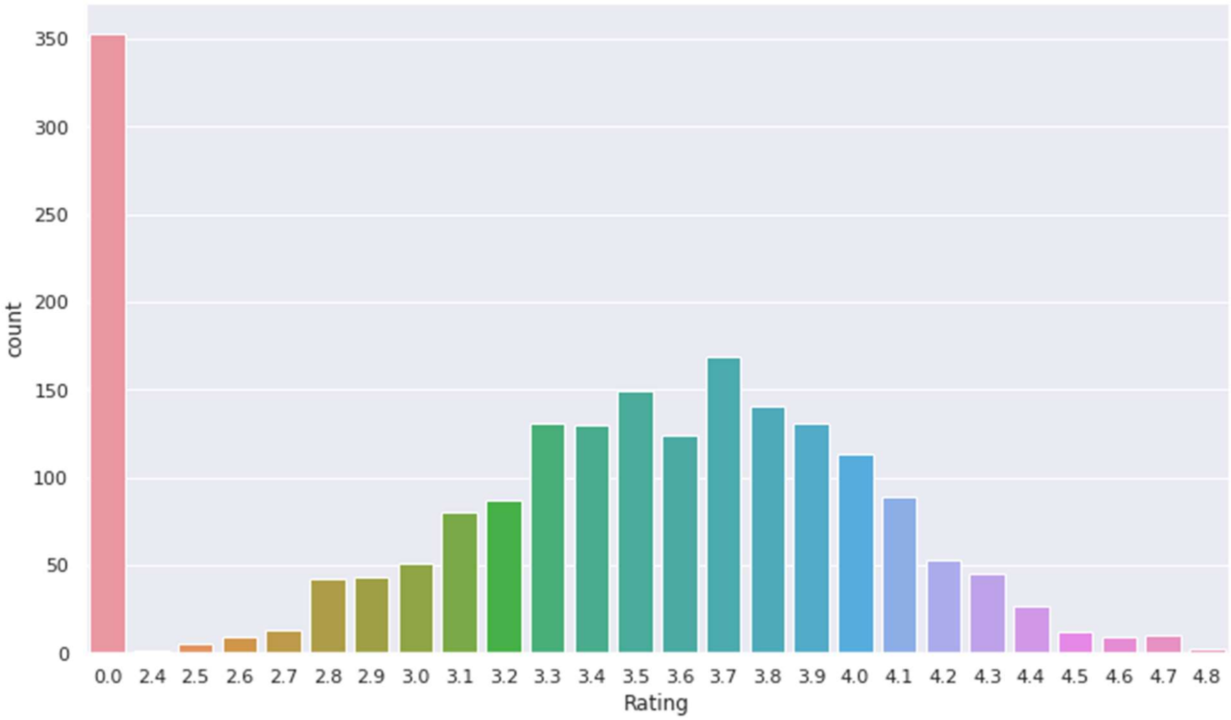
Finding: 1

How many users have given the Ratings?

First and foremost, I decided to describe and find the count plot using seaborn packages. I am specifically interested in Ratings. I want to know how many people had given the Ratings to the Restaurant. Ex. Sometimes only 1 person gave the ratings of 5 which results in the faulty results. I used a count plot to analyze the counts of observations in each categorical bin using bars. I found most of the restaurant has 3.5 to 3.8 ratings given by 150-180 users. This information helps us to think about which location we can improve our delivery boys to give fast and better delivery to the customer.



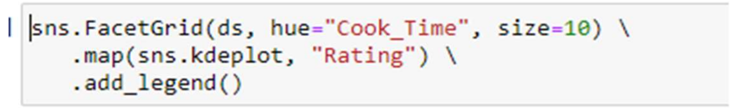
```
sns.set()
plt.figure(figsize=(12,12))
sns.countplot(x=ds['Rating'])
plt.show()
```



Finding: 2

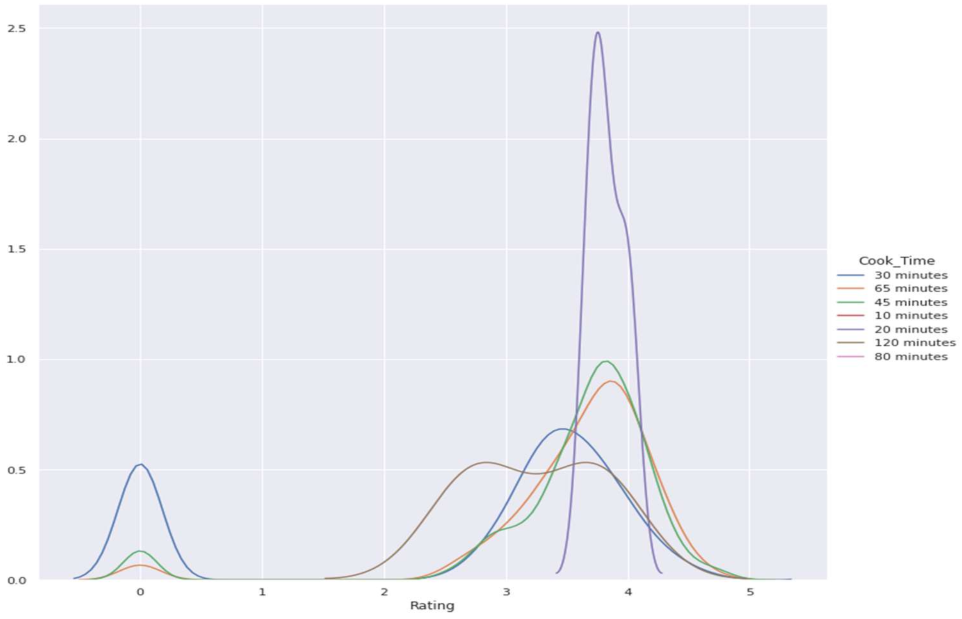
Is the Cooking time of Restaurant is related with Ratings of FoodieX?

Now, I decided to find the cooking time correlation with the Ratings given by the user. Is there any relation to giving more Ratings or fewer Ratings to the Restaurants in the terms of time? For that, I have used the seaborn package again and the KDE(Kernal Density Estimate) plot. So, we can visualize the distribution of observation in one pattern. It is analogous to the Histogram but gives an analysis in an interactive pattern.



```
sns.FacetGrid(ds, hue="Cook_Time", size=10) \
    .map(sns.kdeplot, "Rating") \
    .add_legend()
```

I was correct in the terms of what I think. People go give more Ratings to the Restaurants who deliver the food hot and in a shorter period of time. It will give the insight to think the chef and other work staffs of a restaurant are an important factor to increase ratings and our profits as well. We can improve the employee workflow to meet the timings. As we can see in the chart, people do give ratings between 3.5-4.0 to the restaurants who have 30 minutes of time. There is a conflict between restaurants that are serving within 10-20minutes. We can think from a customer point of view that, Customers do believe that only fast-food cuisines can be served within that short period of time. For other cuisines, it is not possible to make food and deliver the food in this lesser period.

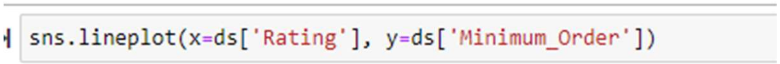


Finding:3

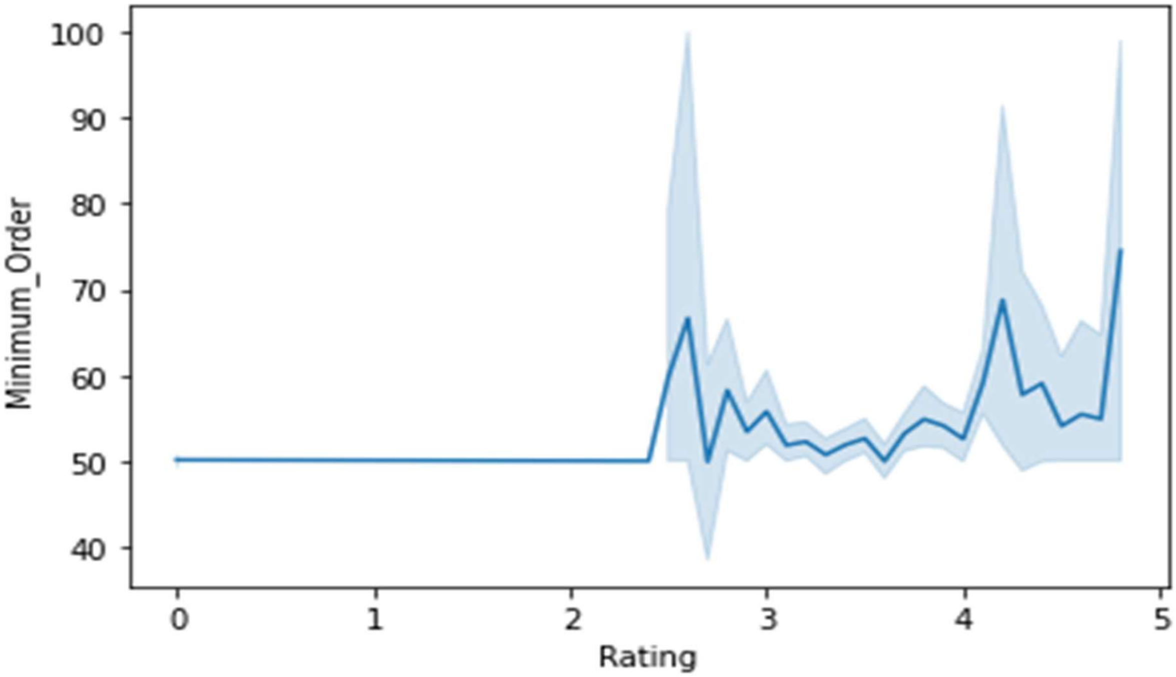
Comparing and Finding the patten of Ratings versus Minimum_Order

I want to dig more inside in between the data. So, I decided to find the pattern of Ratings and Minimum_Order. For that, I have used the line plot. The reason to choose the line plot is they are useful in that they show data variables and trends very clearly and can help to make predictions about the results of data not yet recorded.

I was impressed by the results of a comparison. My prediction was wrong this time. There is a significant variation in the ratings and Minimum_order. People do give higher ratings as well as the lower ratings who have the minimum order in between 90-100$ value. Additionally, the order value of fewer than 50 people does prefer to give only 0-2 ratings. It is giving insight that we don't need to worry and give stress to the minimum order value part.



```
sns.lineplot(x=ds['Rating'], y=ds['Minimum_Order'])
```

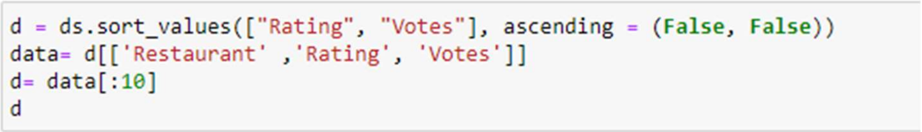


Finding 4:

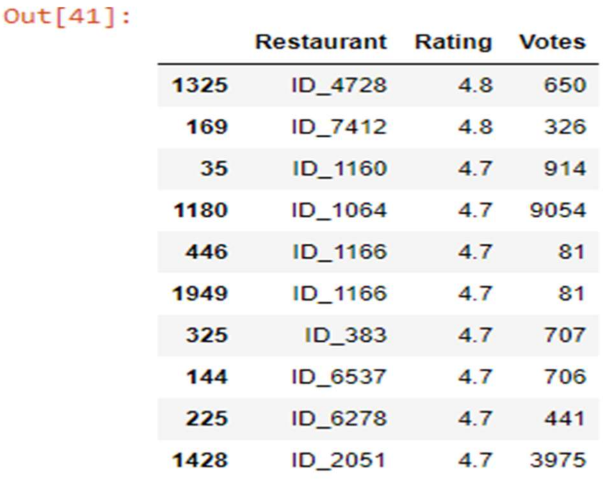
Finding the Most 10 Top Favorite Restaurants using Votes and Rating given by public.

I looked at the dataset. I thought that finding the top 10 restaurants would help us to develop our business strategy by introducing new things to the people on those locations to analyze whether it impacts the other location or not. For instance, by analyzing this pattern, we can give discounts and promotional offers to the customer who is using our application more frequently.

For that, I have first sorted values of the Rating and Votes column in a descending pattern. Then I consider the Restaurant ID and the other two variables. After that, I wrote the code to find only the first 10 rows. As you can see, below there are the results. We can use this restaurant Id's to start our new marketing strategies.



```
d = ds.sort_values(["Rating", "Votes"], ascending = (False, False))
data= d[['Restaurant' ,'Rating', 'Votes']]
d= data[:10]
d
```



Out[41]:

	Restaurant	Rating	Votes
1325	ID_4728	4.8	650
169	ID_7412	4.8	326
35	ID_1160	4.7	914
1180	ID_1064	4.7	9054
446	ID_1166	4.7	81
1949	ID_1166	4.7	81
325	ID_383	4.7	707
144	ID_6537	4.7	706
225	ID_6278	4.7	441
1428	ID_2051	4.7	3975

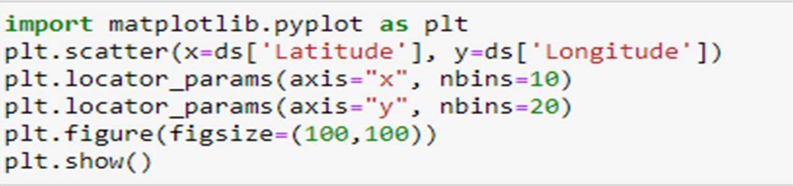
Finding:5

Analyzing the Restaurants' pick-up zones using Longitude and Latitude columns

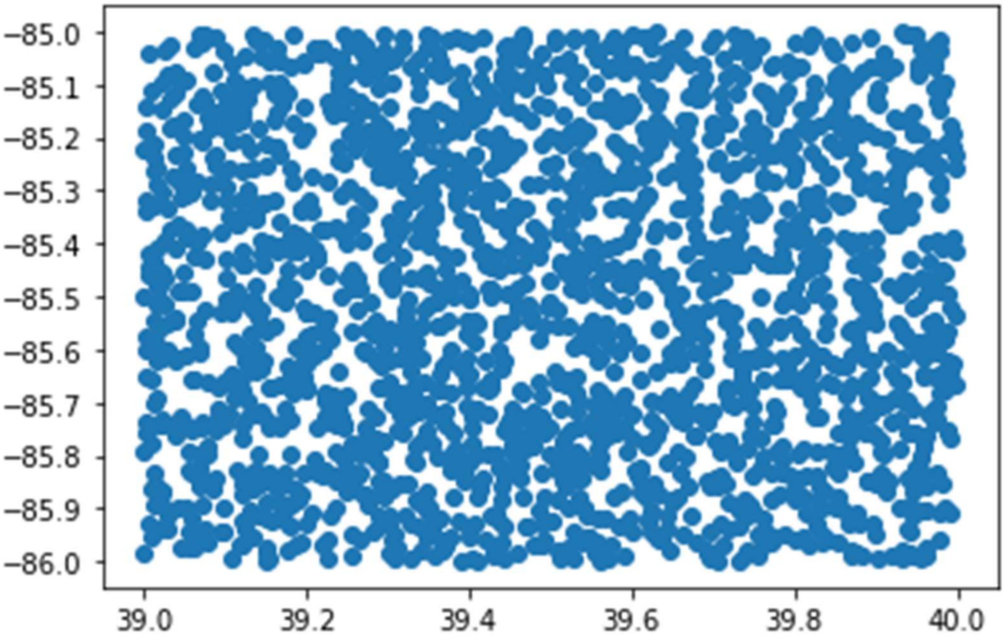
The dataset has the longitude and Latitude values of each restaurant and it will help us to find which are the most favorite pick-up zones of customers. We have used the matplot package here to plot those values as a scatter plot. The cluster algorithms like k-means to predict pickup zones more clearly.

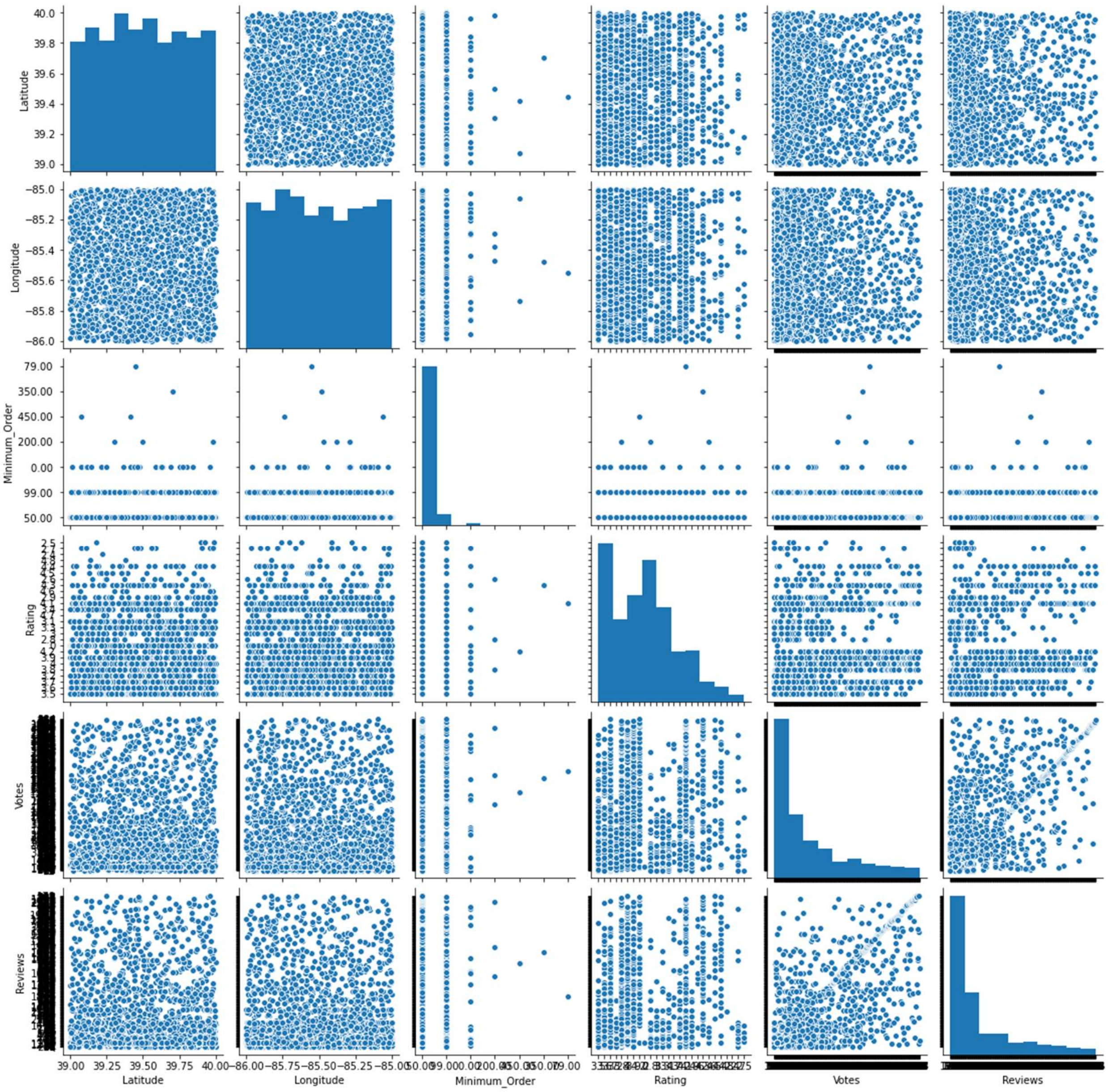
I first plotted everything on the scatter plot and tried to analyze the pattern. Then I have used k-means clustering algorithms to find the pickup zone of a restaurant. I have downloaded the .shape file of Indiana to visualize my pattern directly on the map.

I found the most pick up zones are in between 39.2-39.4 latitude and 85.4-85.6 longitude.



```
import matplotlib.pyplot as plt
plt.scatter(x=ds['Latitude'], y=ds['Longitude'])
plt.locator_params(axis="x", nbins=10)
plt.locator_params(axis="y", nbins=20)
plt.figure(figsize=(100,100))
plt.show()
```




Latitude
Longitude
Minimum_Order
Rating
Votes
Reviews
40.0
39.8
39.6
39.4
39.2
39.0
−85.0
−85.2
−85.4
−85.6
−85.8
−86.0
79.00
350.00
450.00
200.00
0.00
99.00
50.00
39.00
39.25
39.50
39.75
40.00
−86.00
−85.75
−85.50
−85.25
−85.00

Thank you.