[Self-Project]

[ZOMATO DATA ANALYSIS]

Akash Vivek Chodankar | 190100009 | July 6, 2021

Categorical Variables

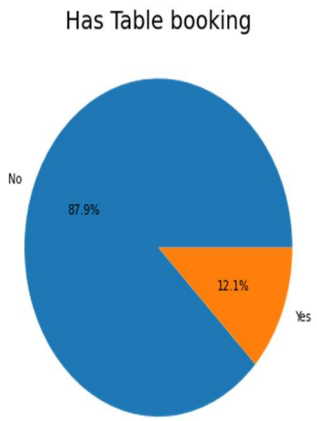
What do the binary variable have to say?

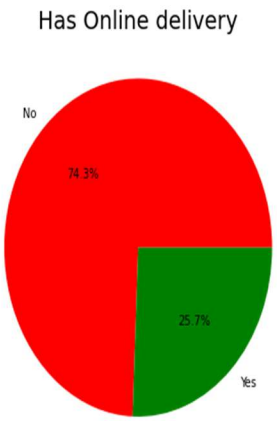
Each of the entries of 'Switch to Order Menu' is No. This indicates that no restaurant has this facility.

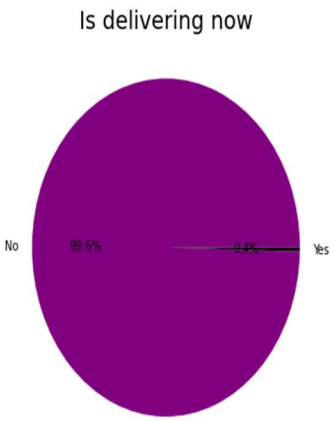
```
No     9551
Name: Switch to order menu
```

Below we can see that:

1) Nearly 12% of the restaurants have table booking.
2) Nearly 26% of them have the online delivery facility through Zomato.
3) Only 0.4% are delivering now.



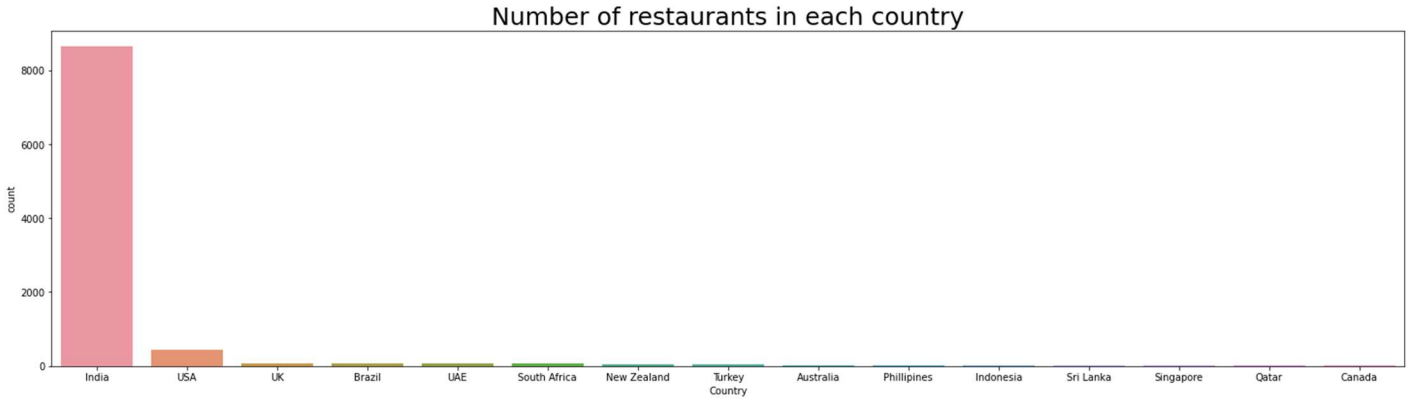


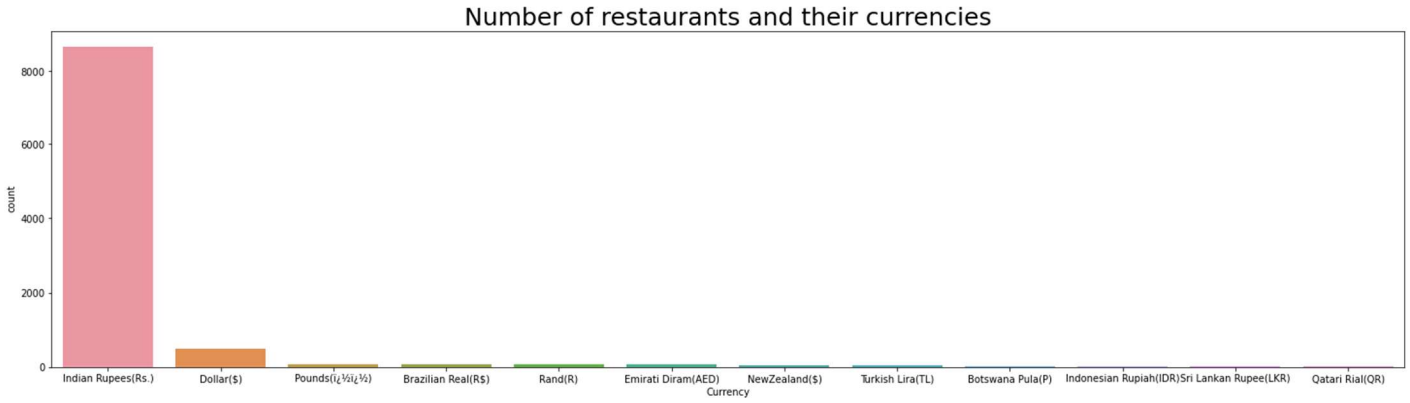


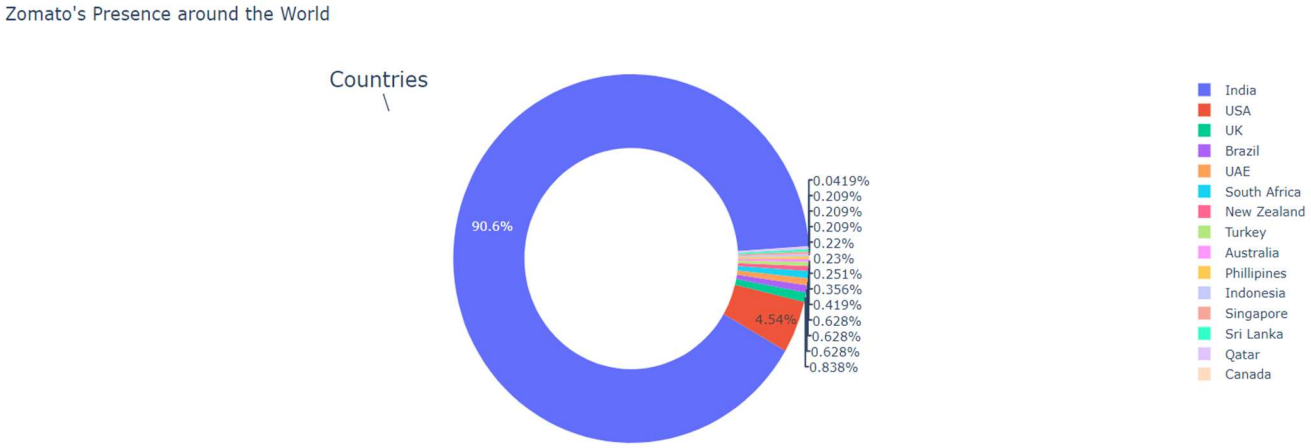
Countries and their Currencies

There are 15 countries in all. India has a share of about 90% and USA of about 5%. Qatar and Canada have the least share. The pie-chart gives a clear visualization of this.

As far has the currencies are concerned, there are only 12 of them indicating that some countries have the same currency.







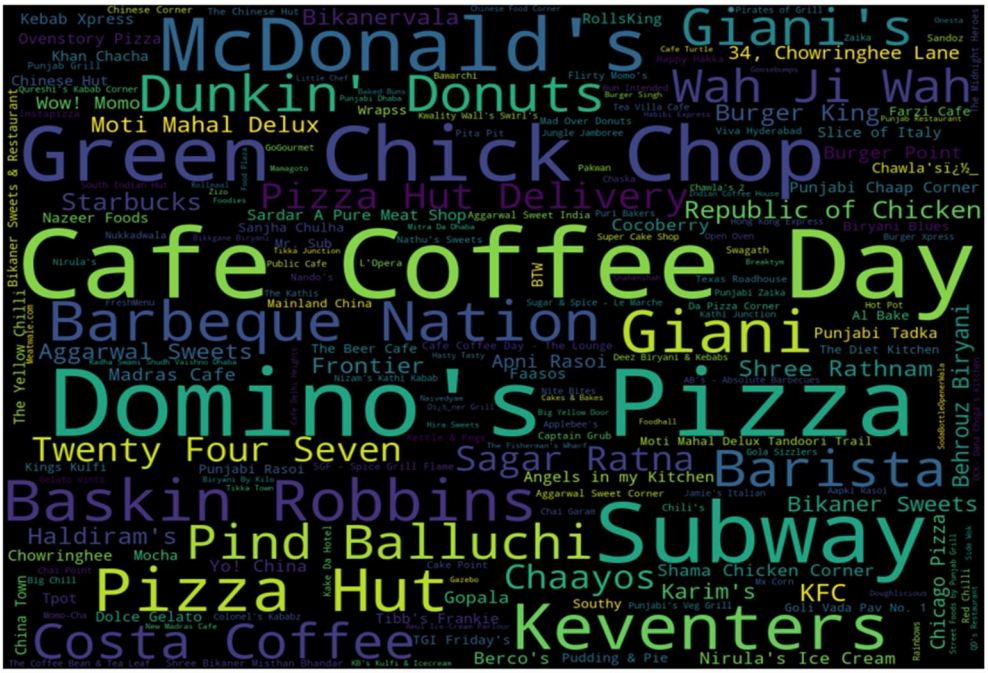
Which cities have the highest number of restaurants?

The below point plot shows us the trend very clearly. We can note that top 10 of the cities are from India (obviously because nearly 90% of the data is from India 😉!). In addition, the top 4 cities dominate the others.

Are there any food chains?

The below word cloud tells us that there actually are quite a few of those. Some of the major ones being, Café Coffee Day, Domino's Pizza, etc.

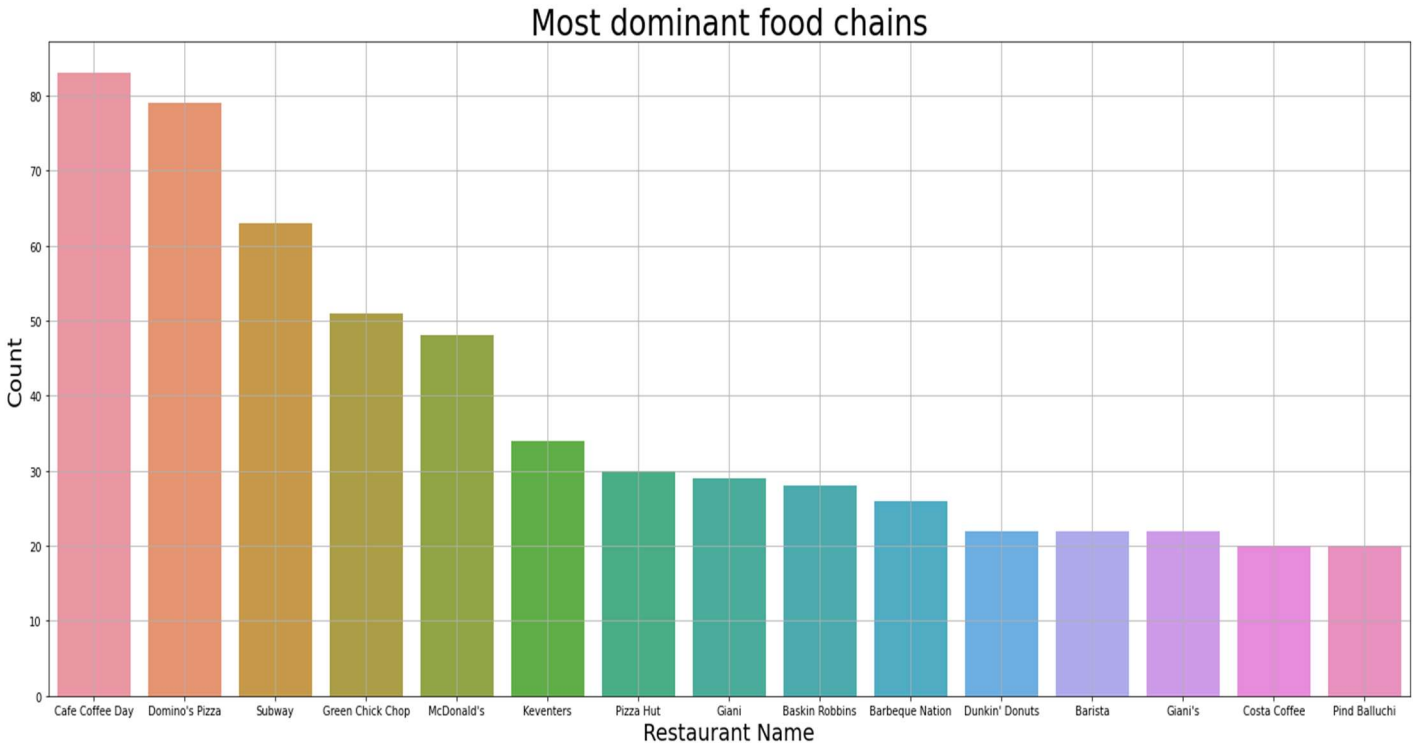
The table gives us the exact statistic of the 5 most dominant eateries on Zomato.



Restaurant	Count
Café Coffee Day	83
Domino's Pizza	79
Subway	63
Green Chick Chop	51
McDonald's	48

You really want to know which ones are the most dominant?

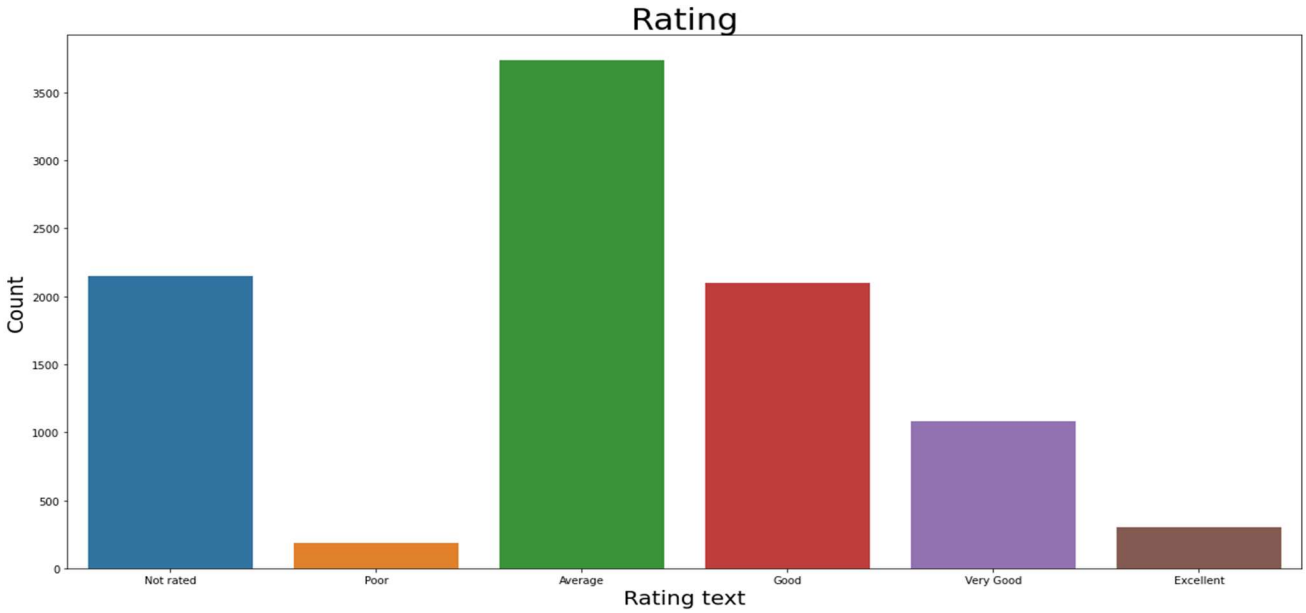
Below is a bar plot that shows you which ones are the most dominant and by what amount in a much elegant way.

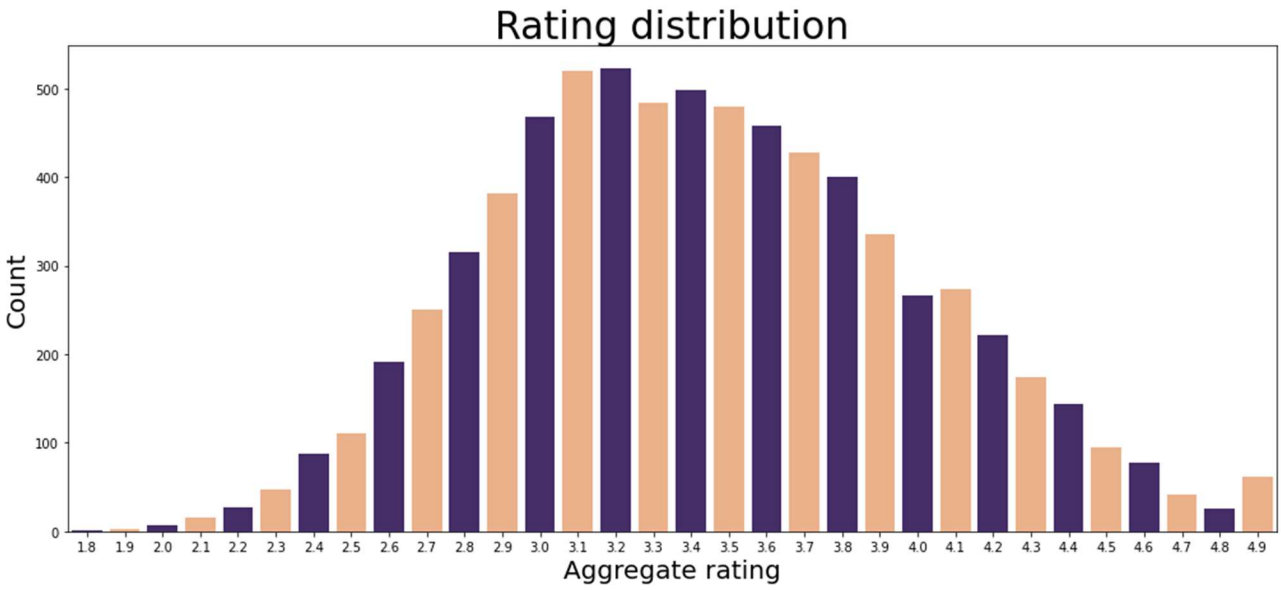


How does the overall rating distribution look like?

Below we can see:

1) A lot of restaurants are not rated.
2) The ones that are rated follow a near normal distribution.





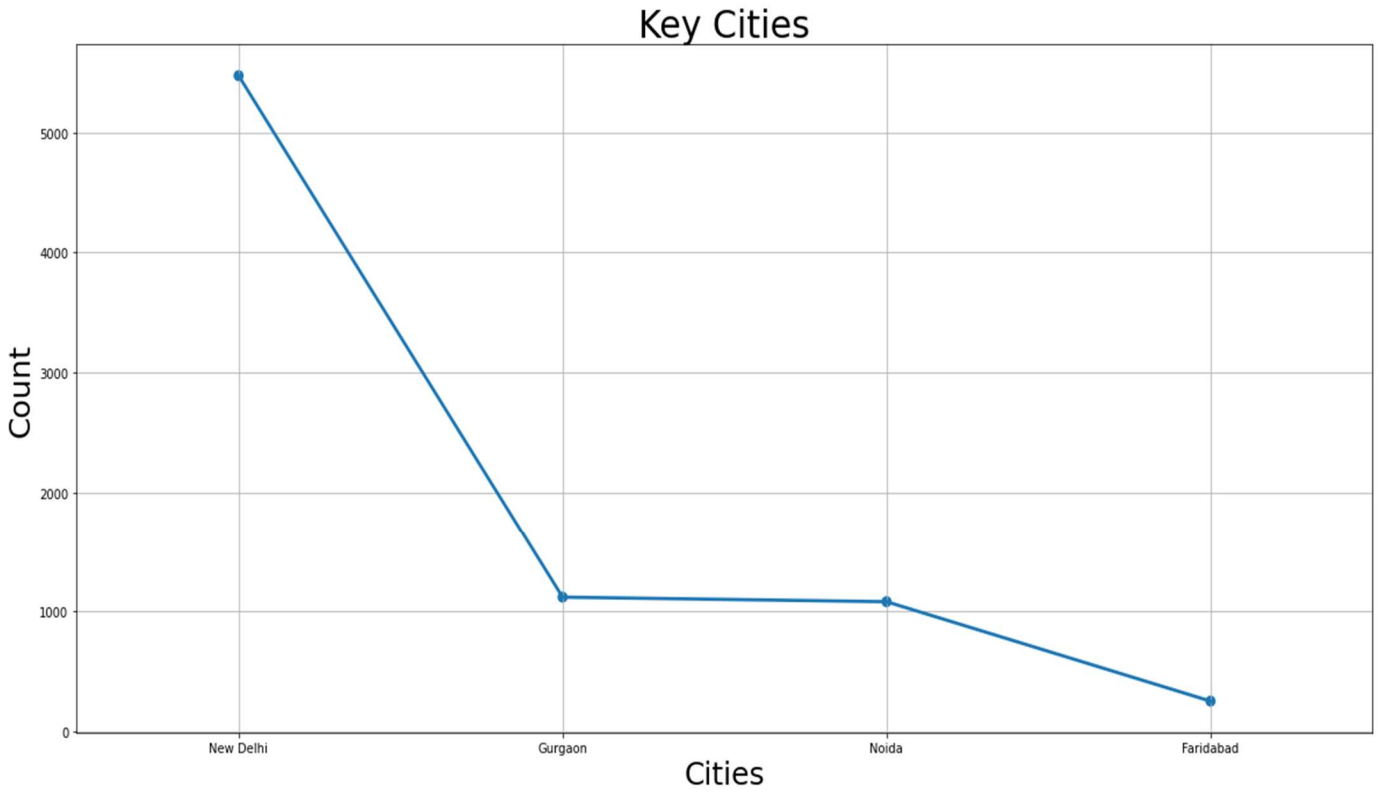
Analyzing Indian data

Why is this more practical?

Since a lot of data is from India and the most dominant locations are also located in India, it will be best to look at the Indian perspective of the data.

There are 8652 entries of Indian restaurants. Also, we do not have any null values.

After exploration we find that the data has 4 cities, New Delhi, Gurgaon, Noida and Faridabad which have the most restaurants.

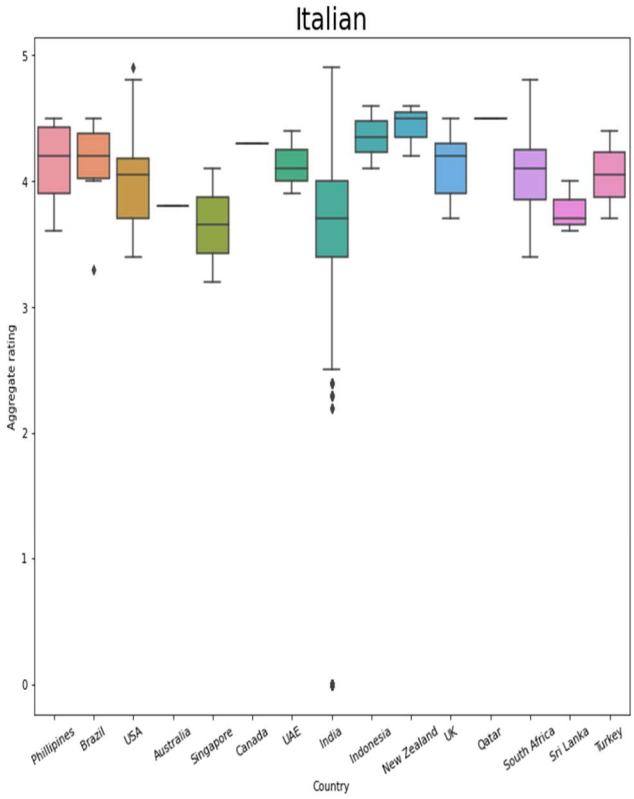


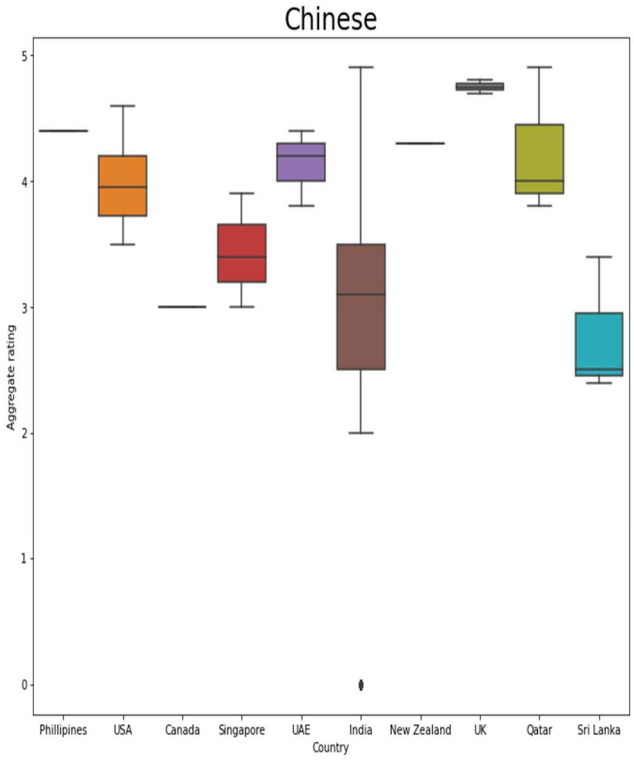
Question 1

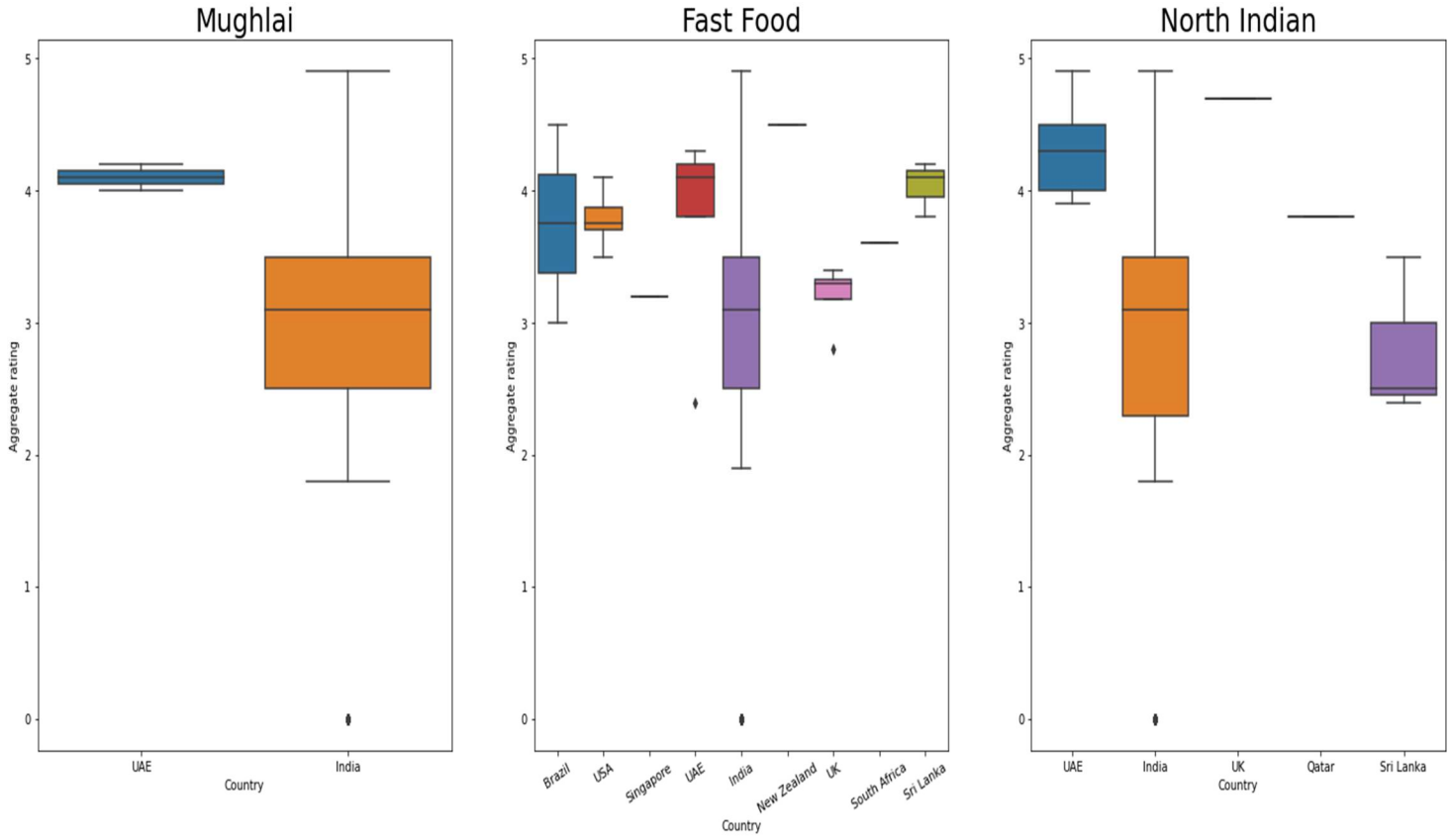
Are the prices of certain cuisine different in different countries?

Below we can see the top 5 cuisines followed by their price and rating variation in countries that they are offered. The price variation is of course a bad indicator because the cost is in different currencies.

North Indian	3960
Chinese	2735
Fast Food	1986
Mughlai	995
Italian	764



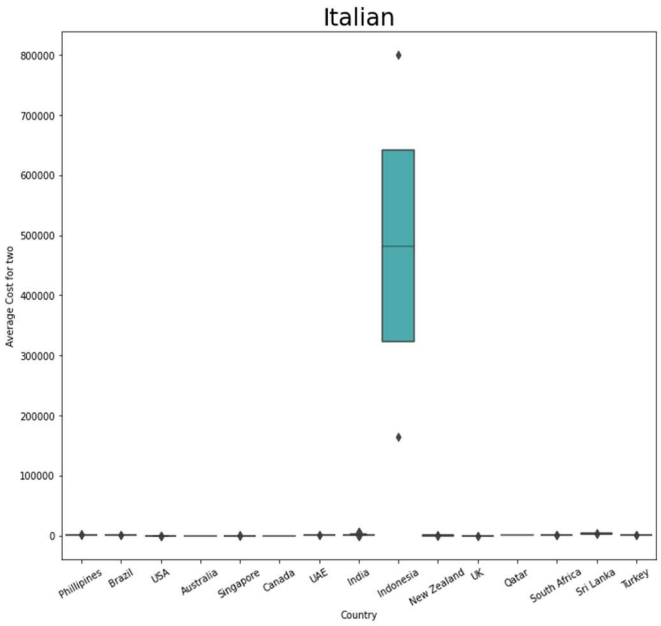


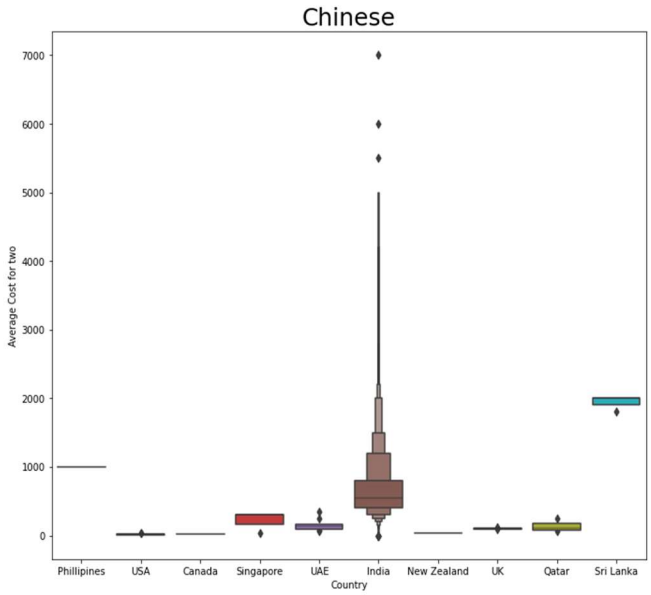


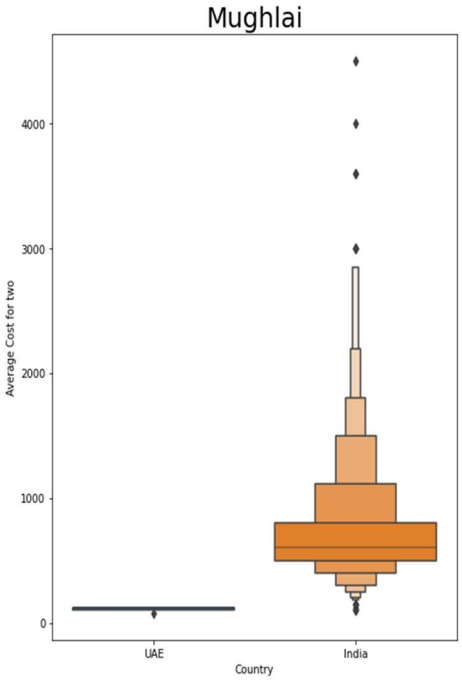
A possible reason for the rating variation is because of the quality of food offered as well as the preference of the people of that country.

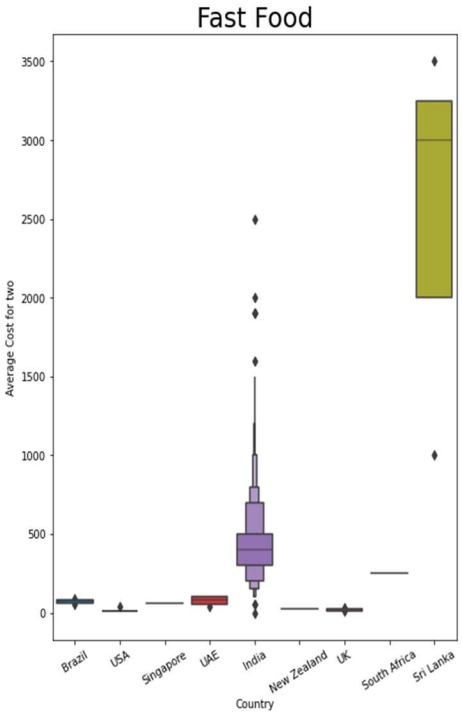
The price variation

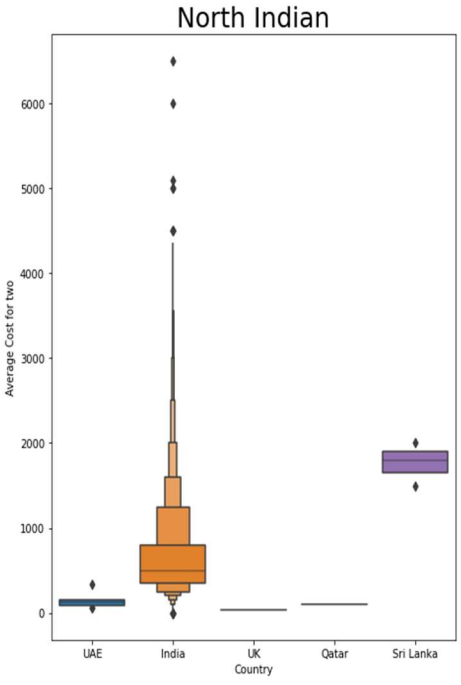
Below plots are not a clear indication of the variation as mentioned above.









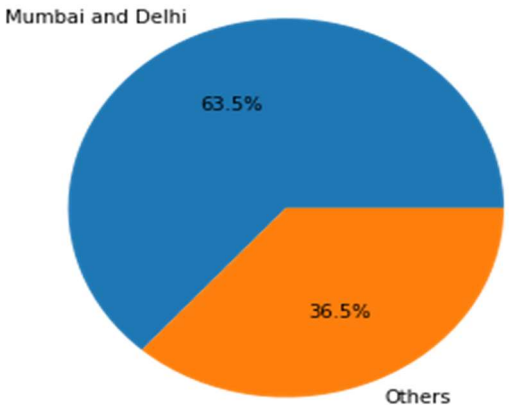


Question 2

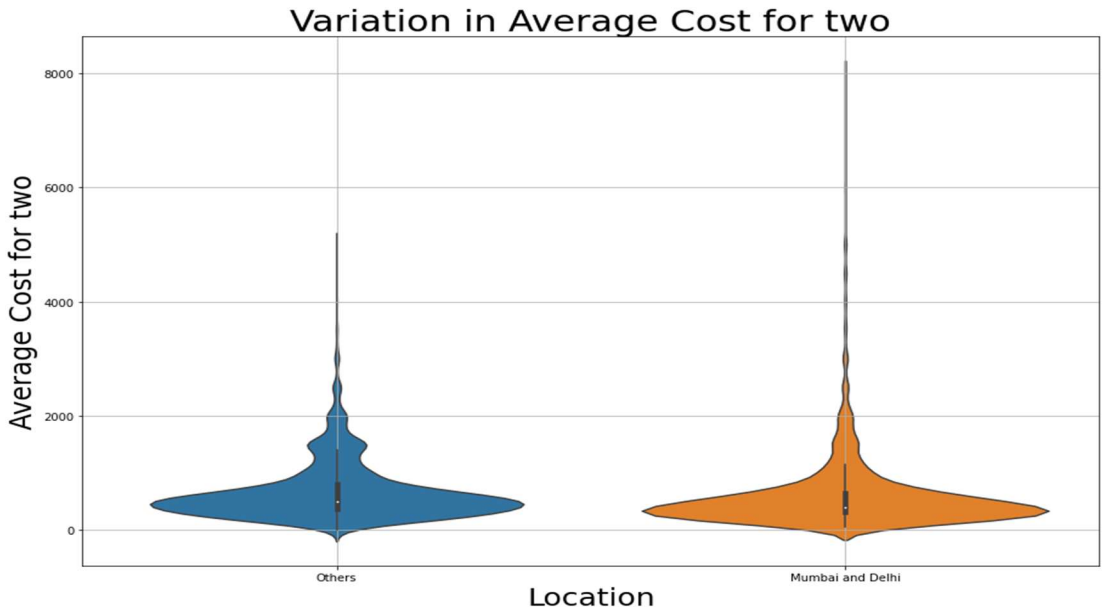
The Mumbai and Delhi tour

This dataset talks about only 20 restaurants from Mumbai, so the major contribution is from Delhi.

63% of the Indian restaurants are from Delhi and Mumbai.



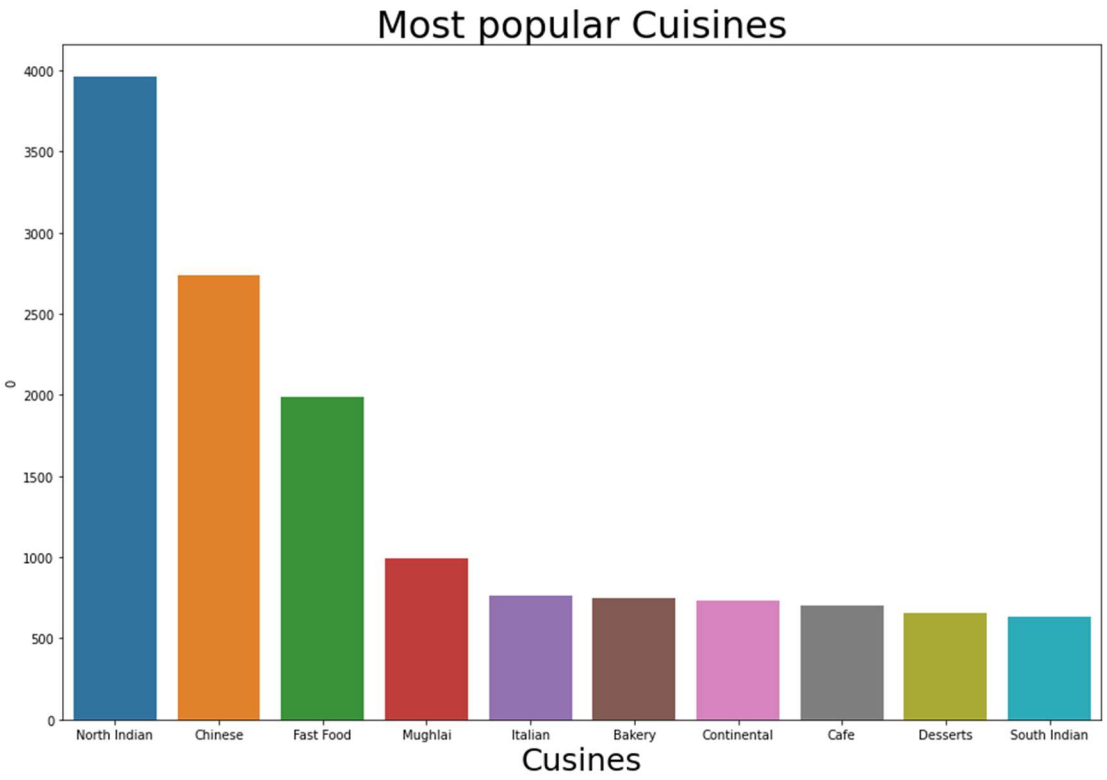
The violin plot gives us the distribution of the average cost for two people in these locations. The long tail for the second violin indicates that there is more variety in these two cities as compared to the others.



Question 3

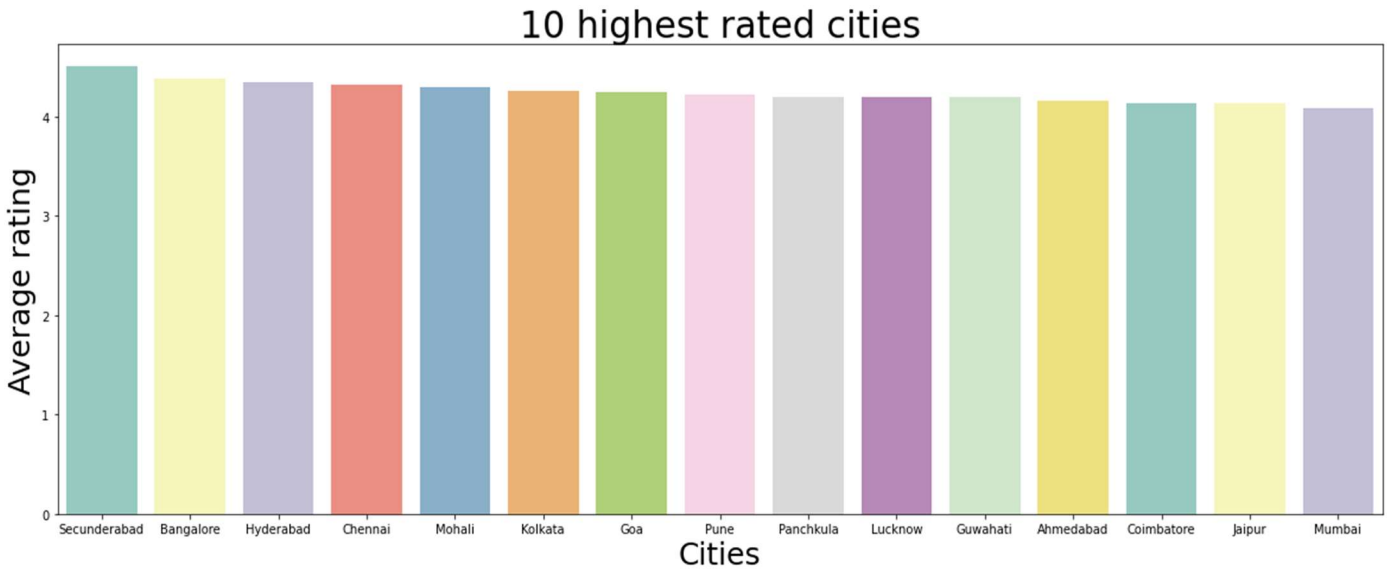
Which ones are the most popular cuisines?

North India, Chinese, Fast food, Mughlai and Italian are the 5 most popular ones. The plot below shows the other 5.



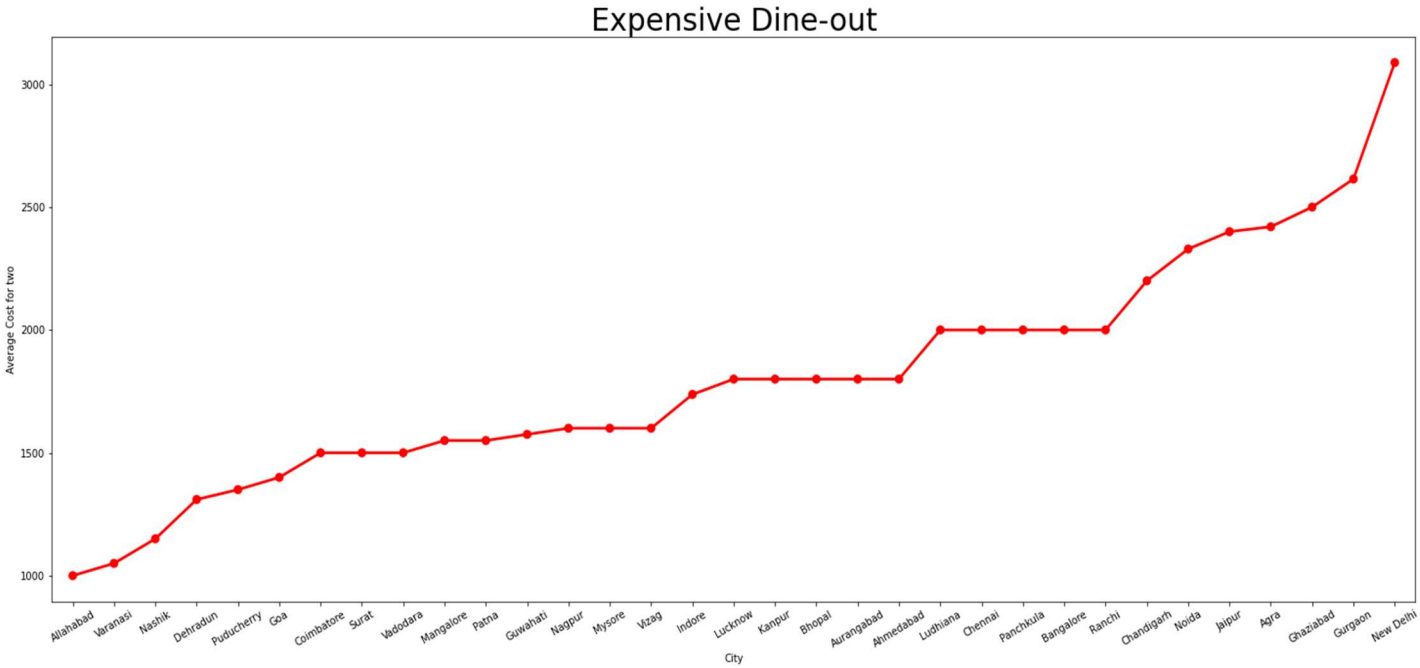
Which cities have the best food rating?

The arrangement is in descending order



Where are the most expensive dine-out services?

There are 33 cities with expensive dine-out services.



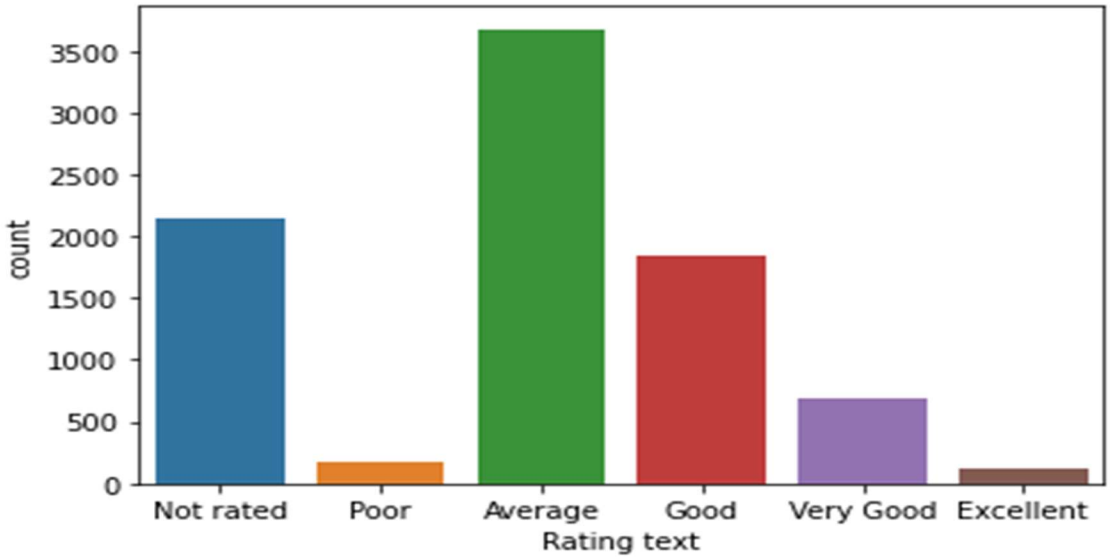
Which cuisines are served in the expensive dine-out restaurants?

Again, we have the word cloud to our rescue.

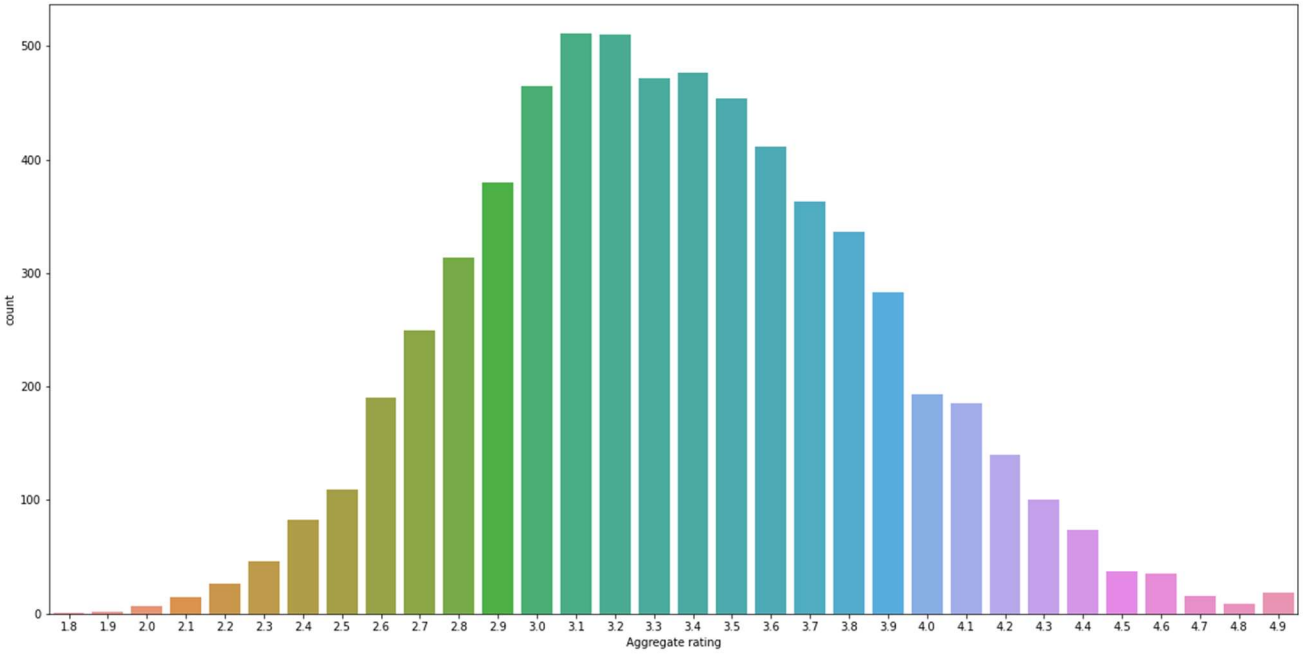
Question 4

Is it true that restaurants with higher votes have a higher rating?

Rating category visualization:

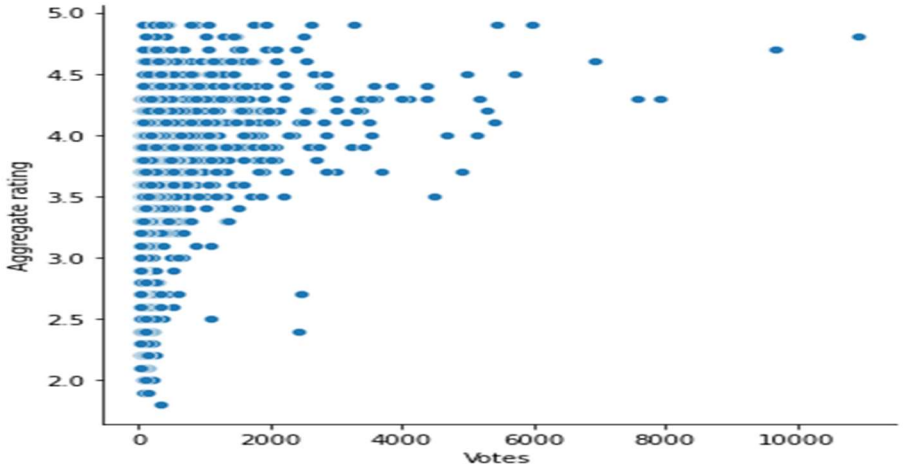


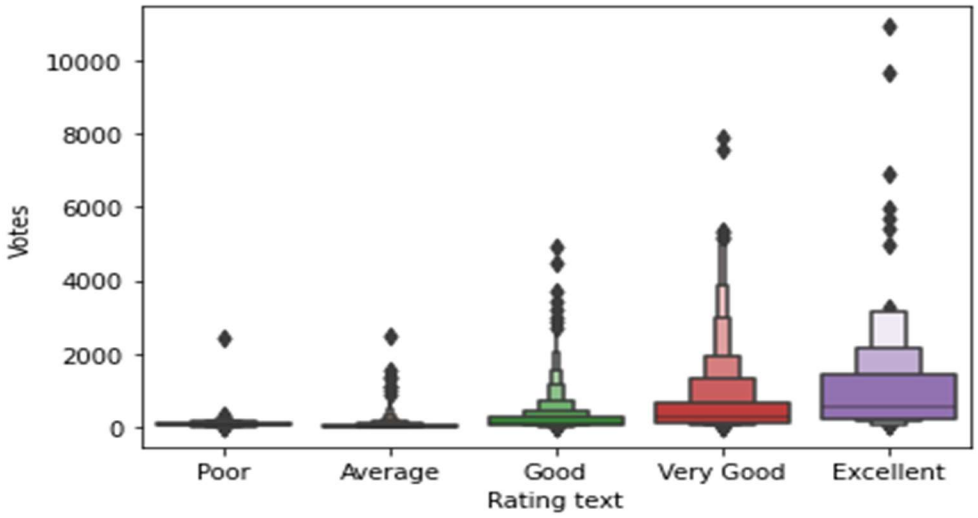
Distribution:



Votes vs Rating

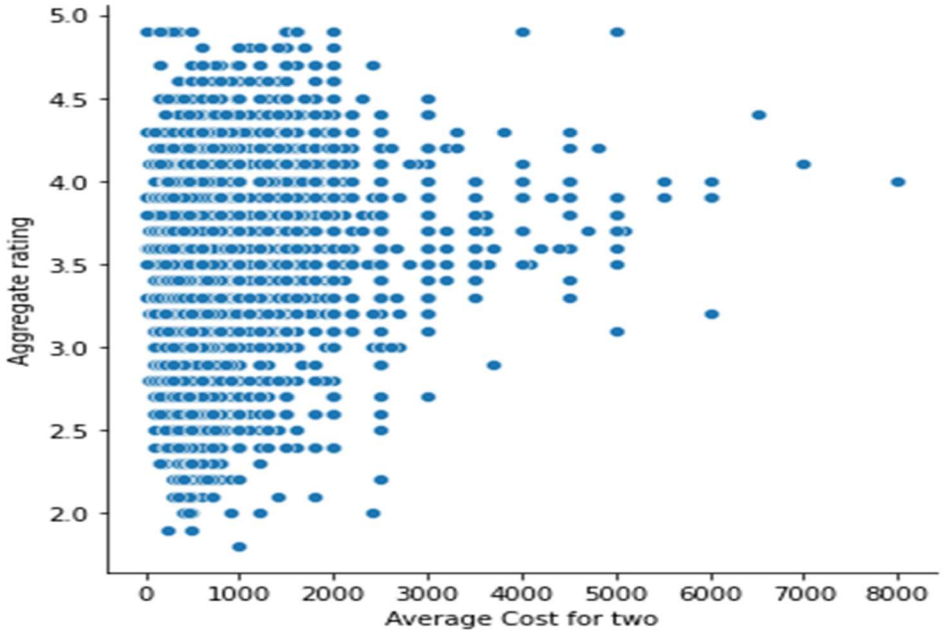
The higher rated restaurants do have higher votes, although there are quite a few restaurants which have low votes in spite of having a high rating. A possible reason could be that it is not tried by many. The correlation value for these two is 0.344, meaning they do have a direct proportionality.

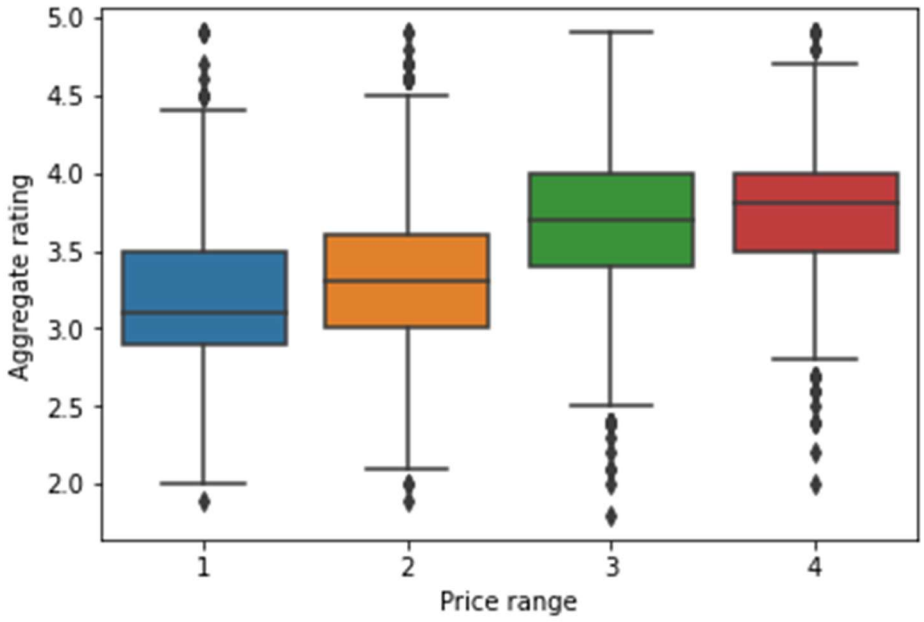




Price range vs rating

Expensive restaurants do show a higher rating. The correlation value for these two is 0.288, meaning they do have a direct proportionality.

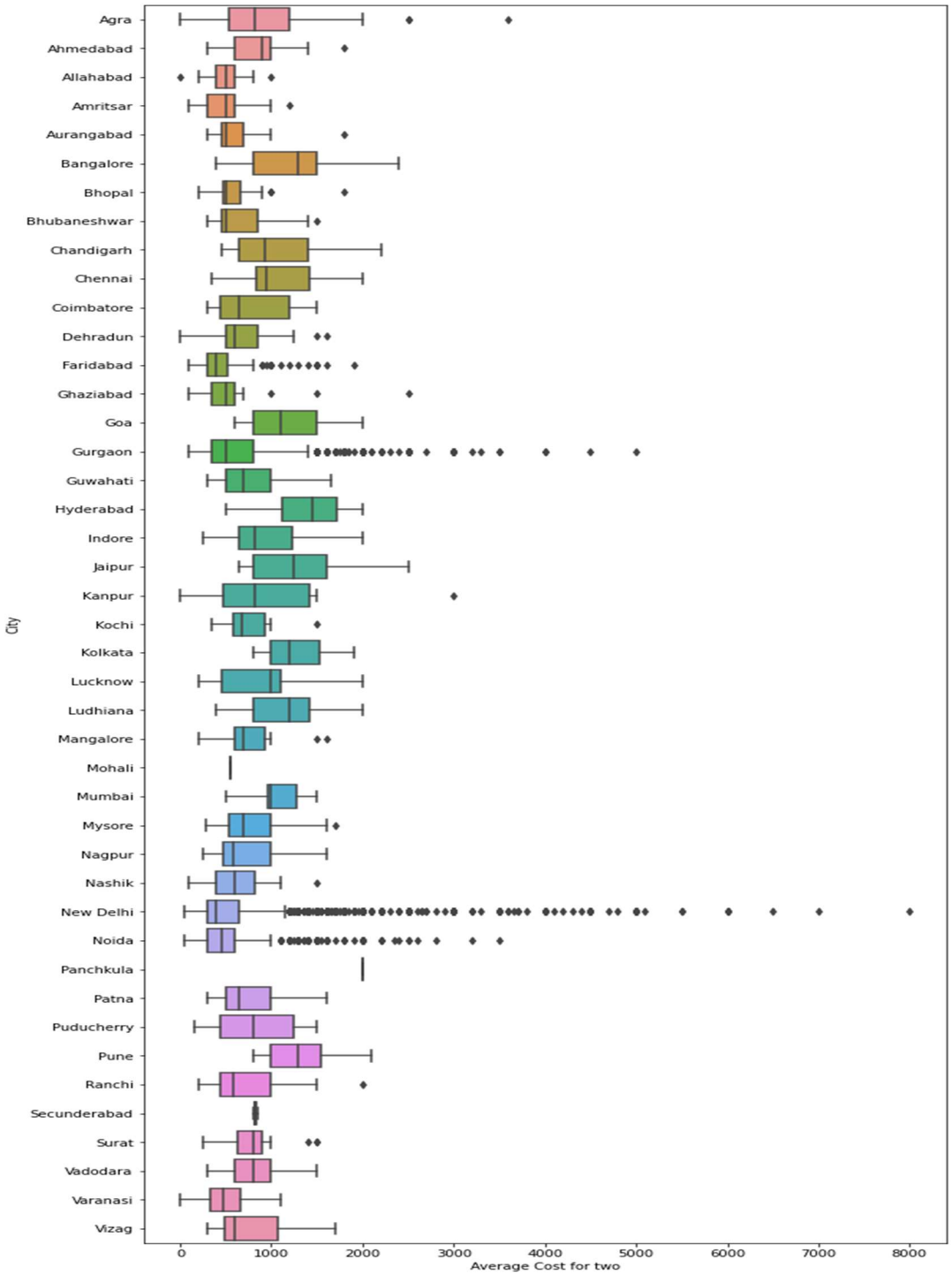




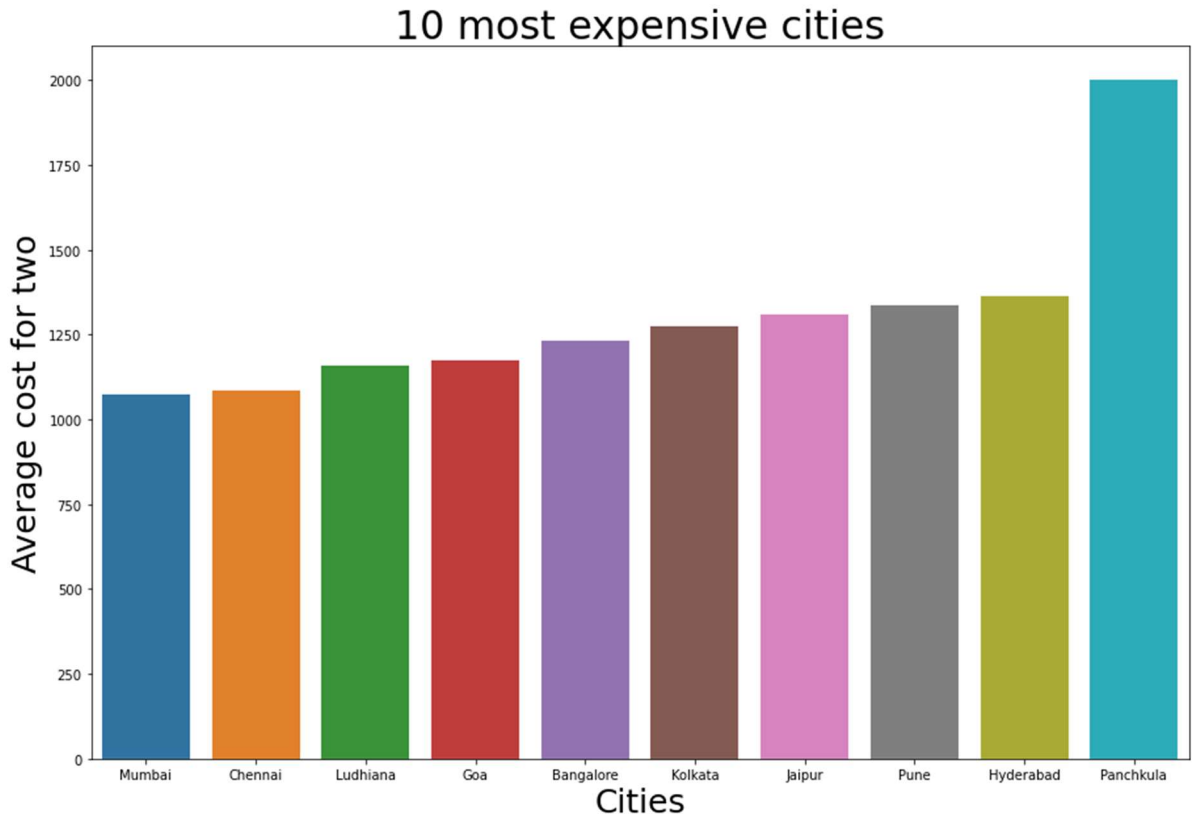
Question 5

What is the average cost for 2 in different cities?

The data has 43 different Indian cities. We can see the cost variation.

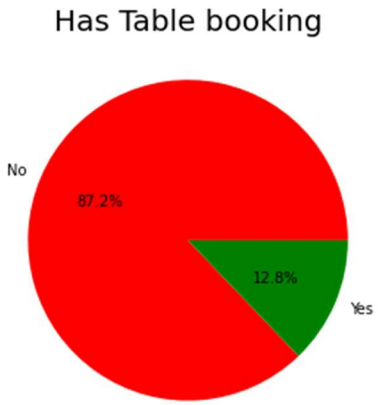


The least and the most expensive cities

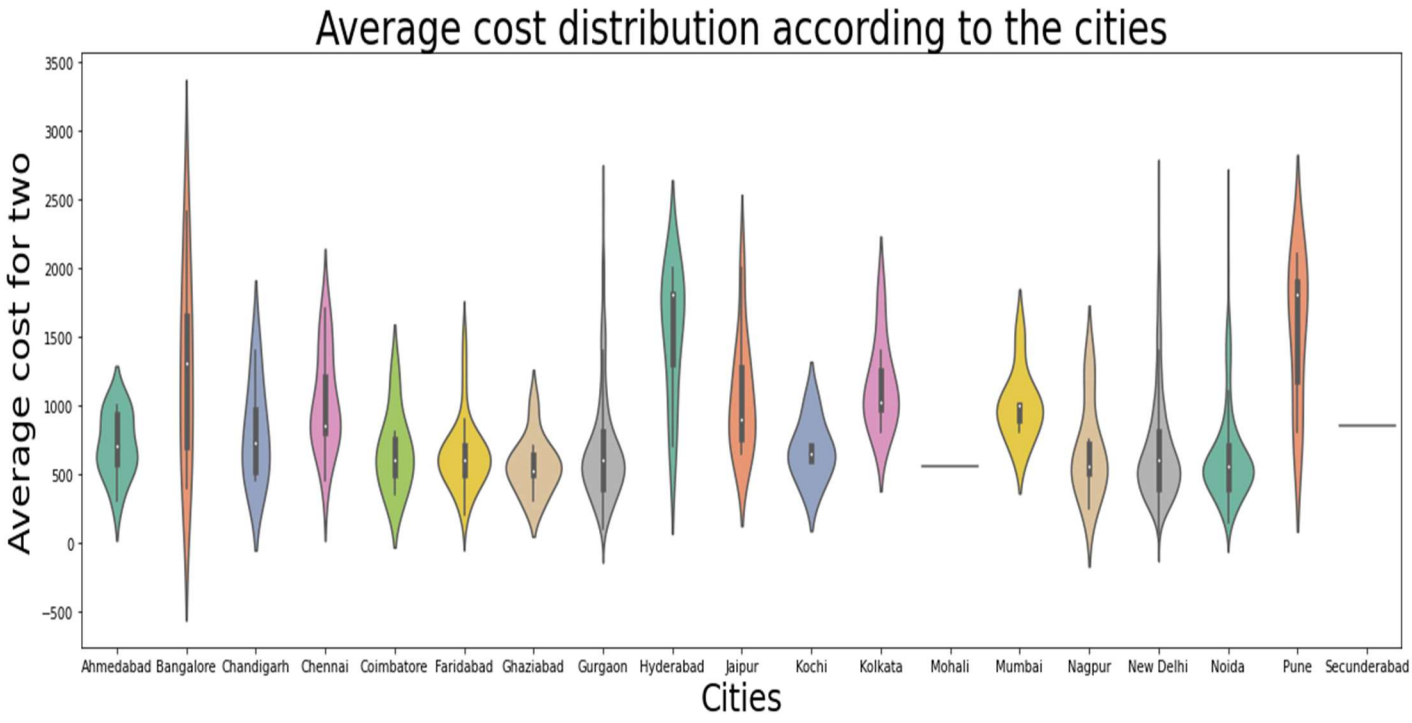


What is the average cost amongst restaurants that deliver via Zomato?

Only 12% delivery have table booking through Zomato.



For restaurants that deliver via Zomato, the average cost for 2 people is about 700 Rs. Below is the city wise distribution.



ML model

Predicting the rating category of an Indian restaurant

An attempt to predict the rating category (out of 5) of the restaurant was made using 2 model. First one was a logistic regression model and the second, a multilayer perceptron. Initially the latitude and the longitude variable were also used but the results improved after removing them. A possible reason could be that we do not represent the entire country well using this dataset. Most of the values are concentrated near New Delhi. We get an accuracy of nearly 65% in both cases.

Model 1, logistic regression:

	precision	recall	f1-score	support
0	0.00	0.00	0.00	148
1	0.70	0.94	0.80	2936
2	0.52	0.39	0.44	1484
3	0.49	0.16	0.24	546
4	0.56	0.05	0.09	97
accuracy			0.66	5211
macro avg	0.45	0.31	0.32	5211
weighted avg	0.61	0.66	0.61	5211

Model 2, Perceptron:

	precision	recall	f1-score	support
1	0.00	0.00	0.00	144
2	0.75	0.90	0.82	2939
3	0.52	0.48	0.50	1472
4	0.44	0.26	0.32	561
5	0.29	0.06	0.10	95
accuracy			0.67	5211
macro avg	0.40	0.34	0.35	5211
weighted avg	0.62	0.67	0.64	5211

For more details, do take a look at the code.

Thanks! 😊

Akash