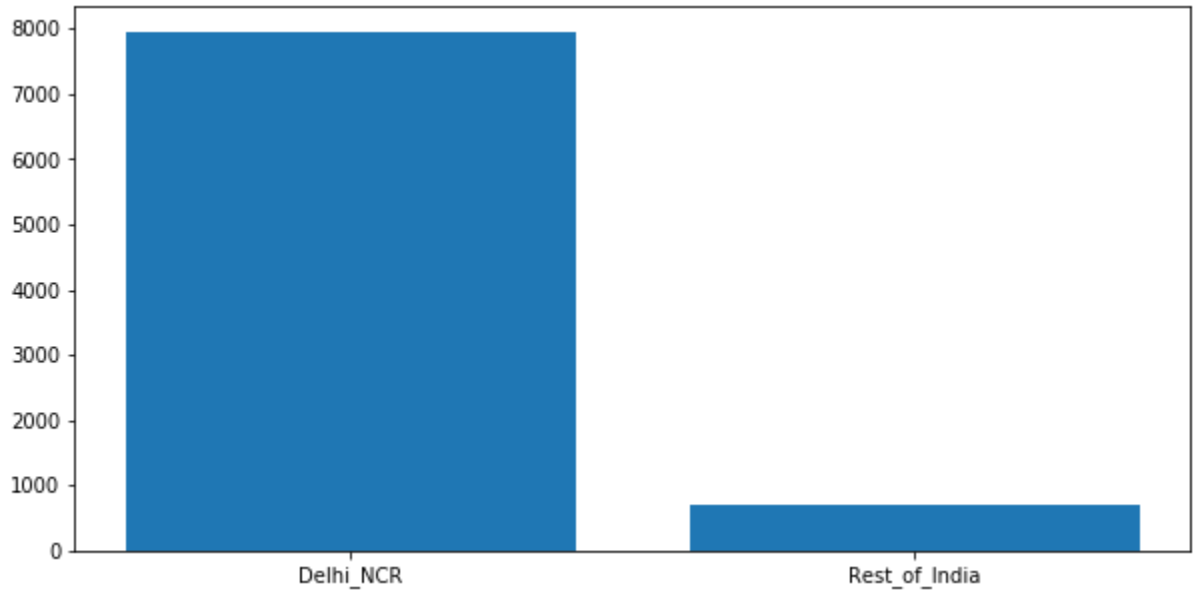
Zomato Project

The dataset is highly skewed toward the cities included in Delhi-NCR. So, we will summarise all the other cities in the Rest of India while those in New Delhi, Ghaziabad, Noida, Gurgaon, Faridabad to Delhi - NCR. Doing this would make our analysis turn toward Delhi-NCR v Rest of India.

Question 1.1 => Plot the bar graph of the number of restaurants present in Delhi NCR vs Rest of India?

Answer :=> Firstly I have read the Zomato CSV file through Pandas DataFrame. Then i did follow some steps that are given below:

- Filter the data which have country Code is 1.
- Declared the List "ncr" which contains the Delhi-NCR cities.
- Declared the Dict ("cities_restaurant") which contains the number of restaurants in Delhi_NCR and Rest_of_India.
- Define a function ("city") that counts the number of restaurants in delhi-ncr and rest of india. It puts the count values in the dictionary ("cities_restaurant").
- Finally we get all the data in the dictionary and draw a Bar Graph through Python Matplotlib.



Question 1.2 => Find the cuisines which are not present in restaurants of Delhi NCR but present in rest of India.Check using Zomato API whether this cuisines are actually not served in restaurants of Delhi-NCR or just it due to incomplete dataset ?

Answer => Firstly I have read the Zomato CSV file through Pandas DataFrame. Then i did follow some steps that are given below:

- Filter the data which have country Code is 1.
- Declared the List "ncr" which contains the Delhi-NCR cities.
- Combined Column is City_Cuisines from which I have figured out the data.
- I declared a function (city_cuisines) that fetch the cuisines and put into the SET (restaurant_cuisines & delhi_ncr_cuisines).
- I have got a new list of cuisines that are not available in delhi NCR restaurants.
- We also fetch the data form zomato API and match with the result what we are getting. If our data match with the actual result then it will return True otherwise False.

Output :

Cajun BBQ German Malwani

True

NOTE: Combined Column is City_Cuisines from which I have figured out the data. And this column has been used for further questions.

Question 1.3 => Find the top 10 cuisines served by the maximum number of restaurants in Delhi NCR and rest of India ?

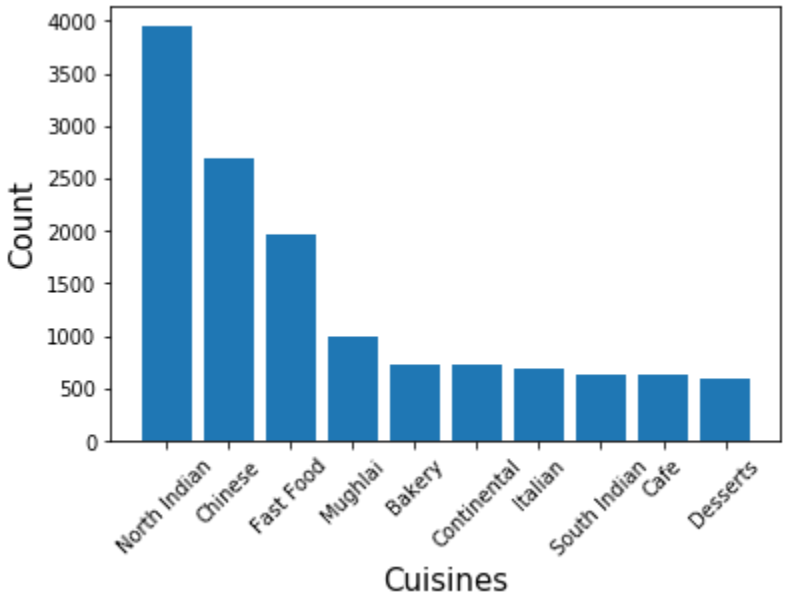
Answer => Firstly I have read the Zomato CSV file through Pandas DataFrame. Then i did follow some steps that are given below:

- Filter the data which have country Code is 1.
- I have made one dictionary named cuisines_with_count wherein cuisine name works as the key and value is the count. I have made a function inside which I am storing the count for each cuisine depending upon their occurrence.
- I made one list and the 0th column is for count and 1st column is for the cuisine name. I sorted the list based on the count and then I picked the top 10. After plotting on the graph we can see the list as below.

Output:

Top ten cuisines that are being served in Delhi-NCR and Rest of India are

1. North Indian
2. Chinese
3. Fast Food
4. Mughlai
5. Bakery
6. Continental
7. Italian
8. South Indian
9. Cafe
10. Desserts



Question 1.4=> Write a short detailed analysis of how cuisine served is different from Delhi NCR to Rest of India. Plot the suitable graph to explain your inference ?

Answer =>

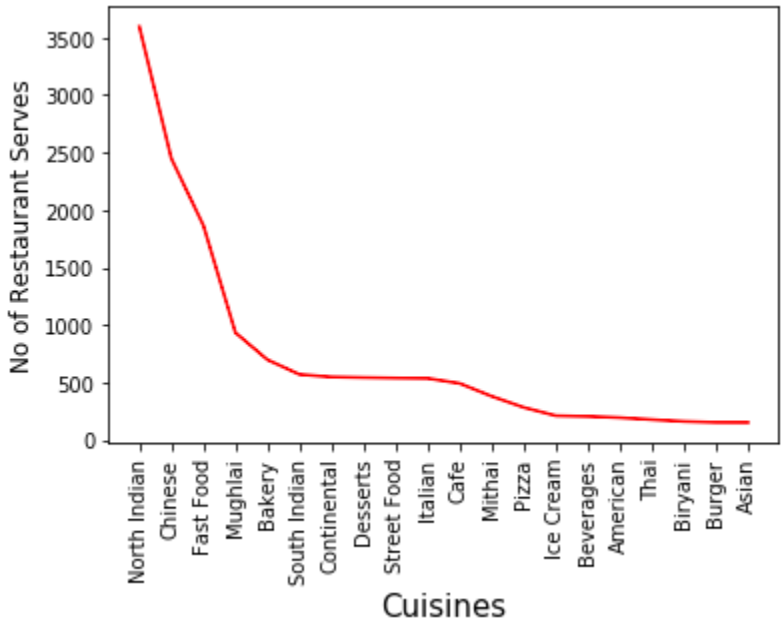
- From the graph we can see that North Indian, Chinese Cuisine are servered most in both places. But number of restaurants serves has a huge difference, North Indian cuisinse is served by more than 3500 restaurants in Delhi NCR.
- But this data changes drastically in Rest of India, and that is only 350 or so and same goes for Chinese cuisine.
- Third most servered cuisine in Delhi NCR is Fast Food wherein Rest India is Continental and there is some order difference in terms of restaurant serves. There are total 86 different cuisines that are offered in Delhi-NCR wherein Rest India offers 70 different cuisines.

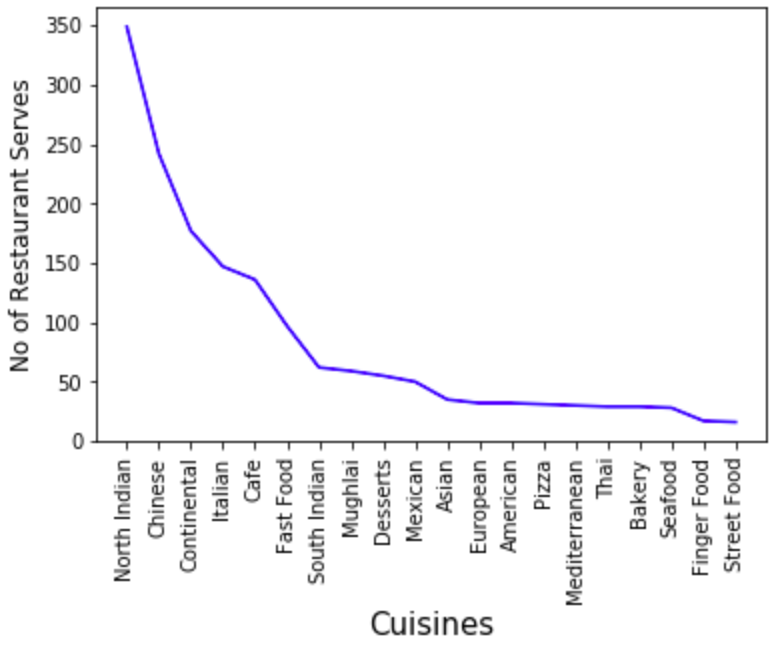
NOTE: Here I have made an column by combining two columns as mentioned earlier.

Output :

Delhi-NCR 86

Rest India 70





Question -2

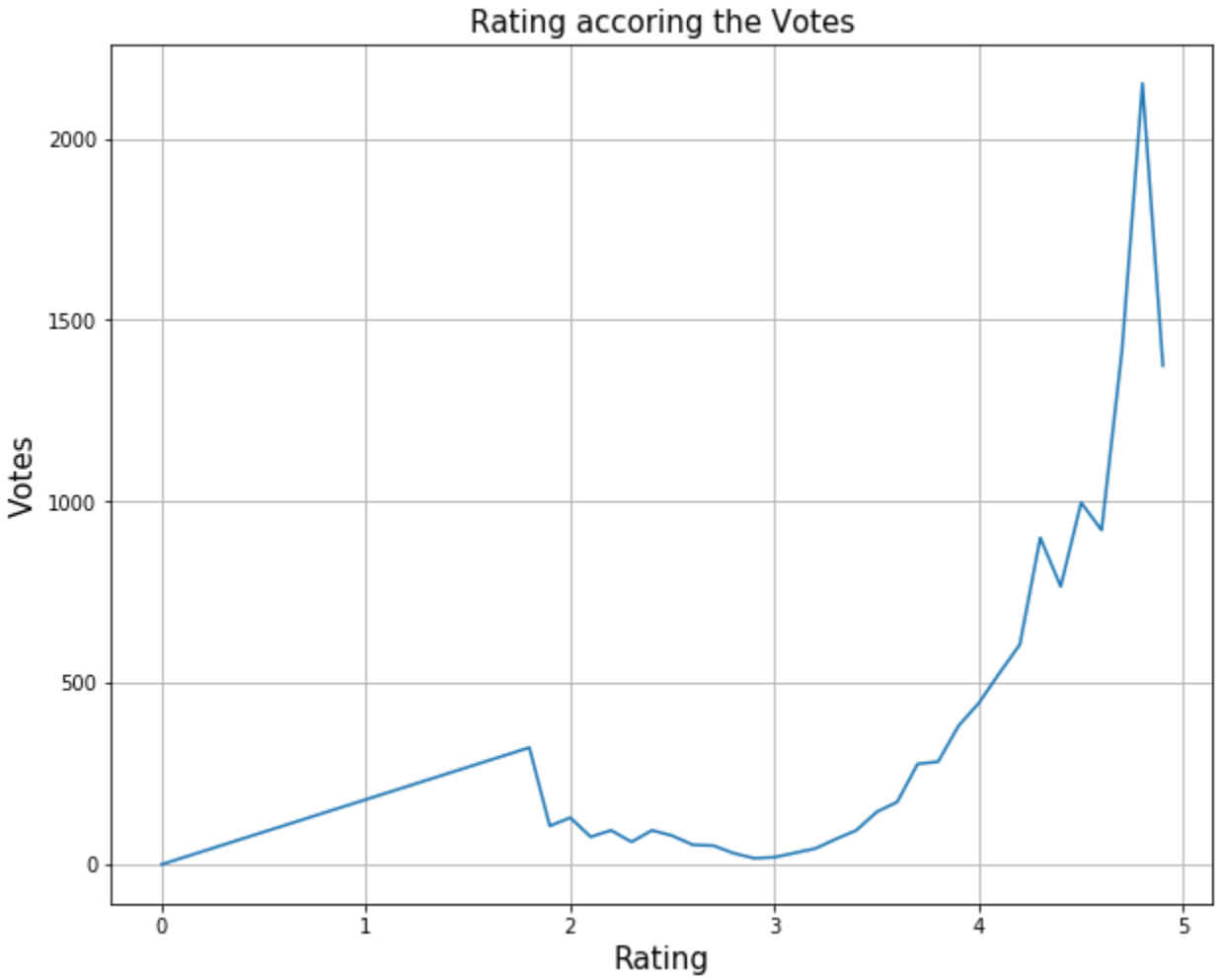
User Rating of a restaurant plays a crucial role in selecting a restaurant or ordering the food from the restaurant.

Question 2 => Write a short detailed analysis of how the rating is affected by Restaurant due following features: Plot a suitable graph to explain your inference.

Question 1 => Number of Votes given Restaurant.

Answer:

- I have got the idea of rating and how many votes they might have got, I did apply data compression technique as the number of votes can be so huge for a particular restaurant.
- To compress the data I have first added all the votes by the rating and maintained total restaurants which will help me to figure out the average voting for each restaurant.
- So first I have made one dictionary which will hold the rating as key and total number of votes & restaurant count as a value i.e a list 0th index is total number of votes and 1st restaurant count. Once I get these details I can make an average by doing total_votes//no_restaurants so for each restaurant we will get the compressed voting numbers.
- After plotting in the graph we can see that those restaurants get more votes. They usually get good ratings i.e 4-5 and voting range is 500 - 2000. Now if we see the low rating then we can find that there are less voting as compared to 4-5 rating range voting.



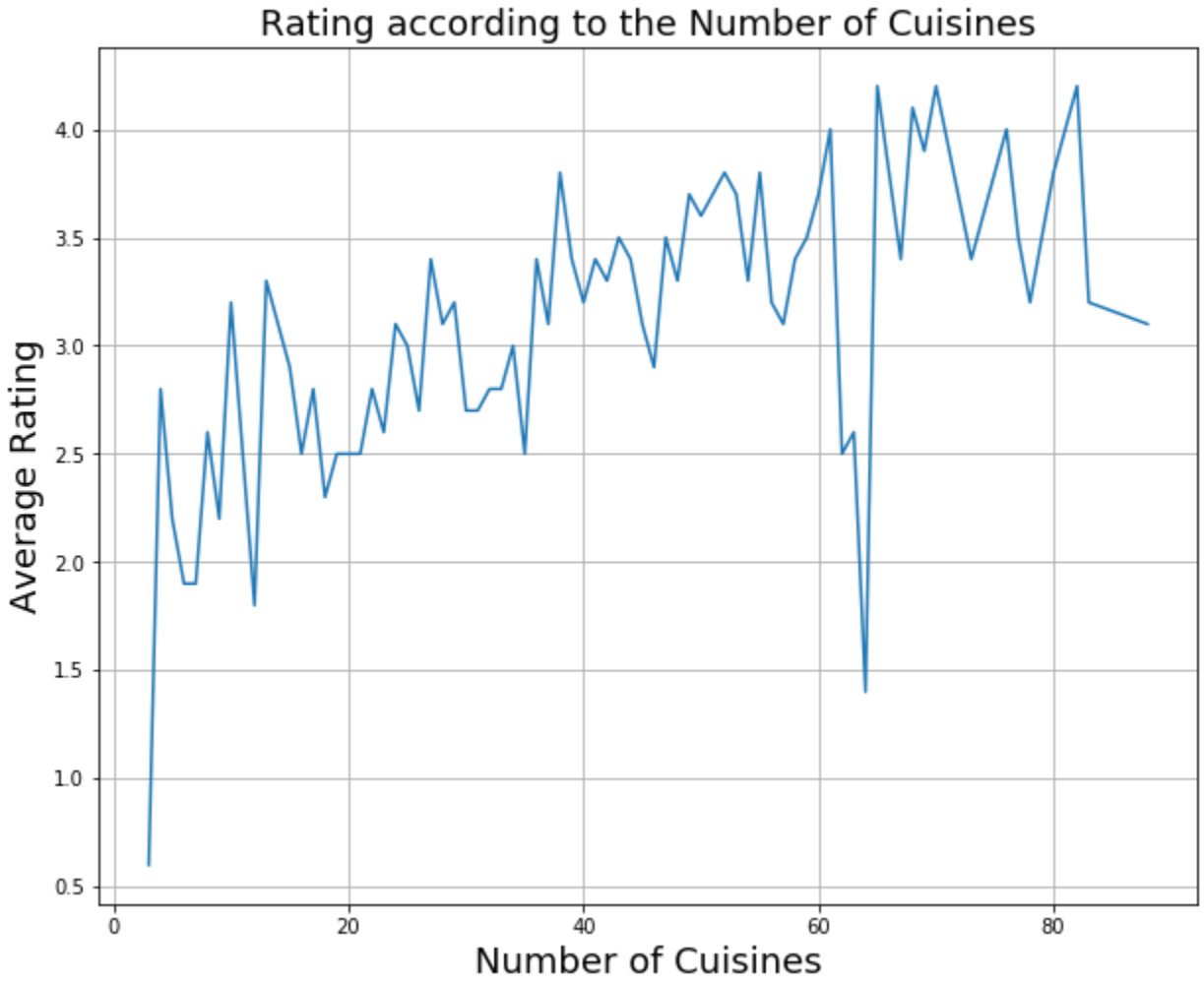
Question 2: => Restaurants serve more number of cuisines.

Answer:

NOTE: i am using data compression here. Data Compression is applied as numbers are huge and it might not fit in the graph properly.

- I have applied the same technique just in this case and for the below cases I haven't used any function to get the desired result. Here we need to check with the number of cuisines
- So I made a number of cuisines as keys and I have stored all the ratings given to a particular number of cuisines and side by side I have stored the number of restaurants.
- After doing this, I am finding the average rating for that particular number of cuisines right. And the rating will lie between the min rating and max rating obtained by the number of cuisines.

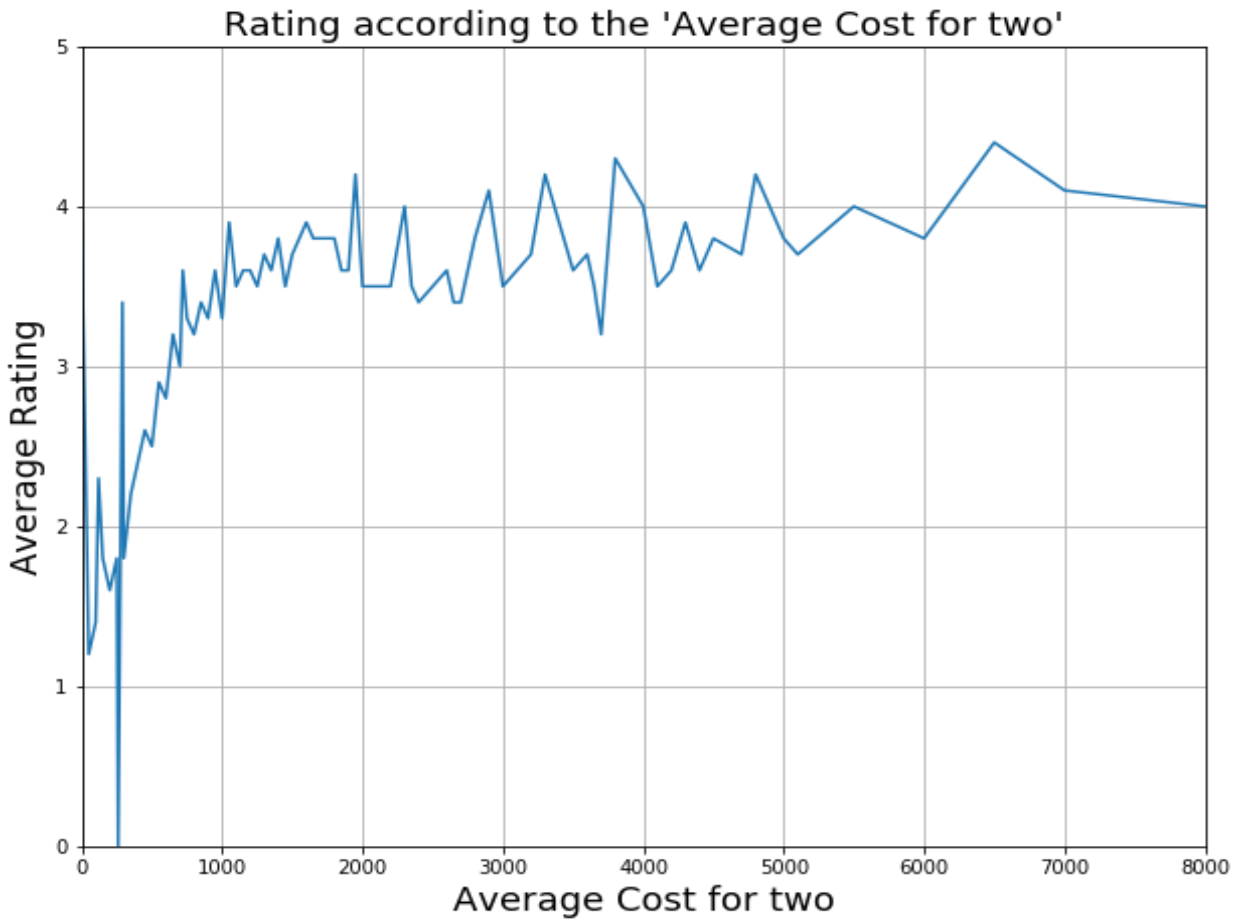
- To find out the average rating I have used total_rating/restaurant_count.
- After plotting the detail in the graph we can notice the higher number of cuisines offered more is the rating.



Question 3: => Average Cost of a Restaurant.

Answer:

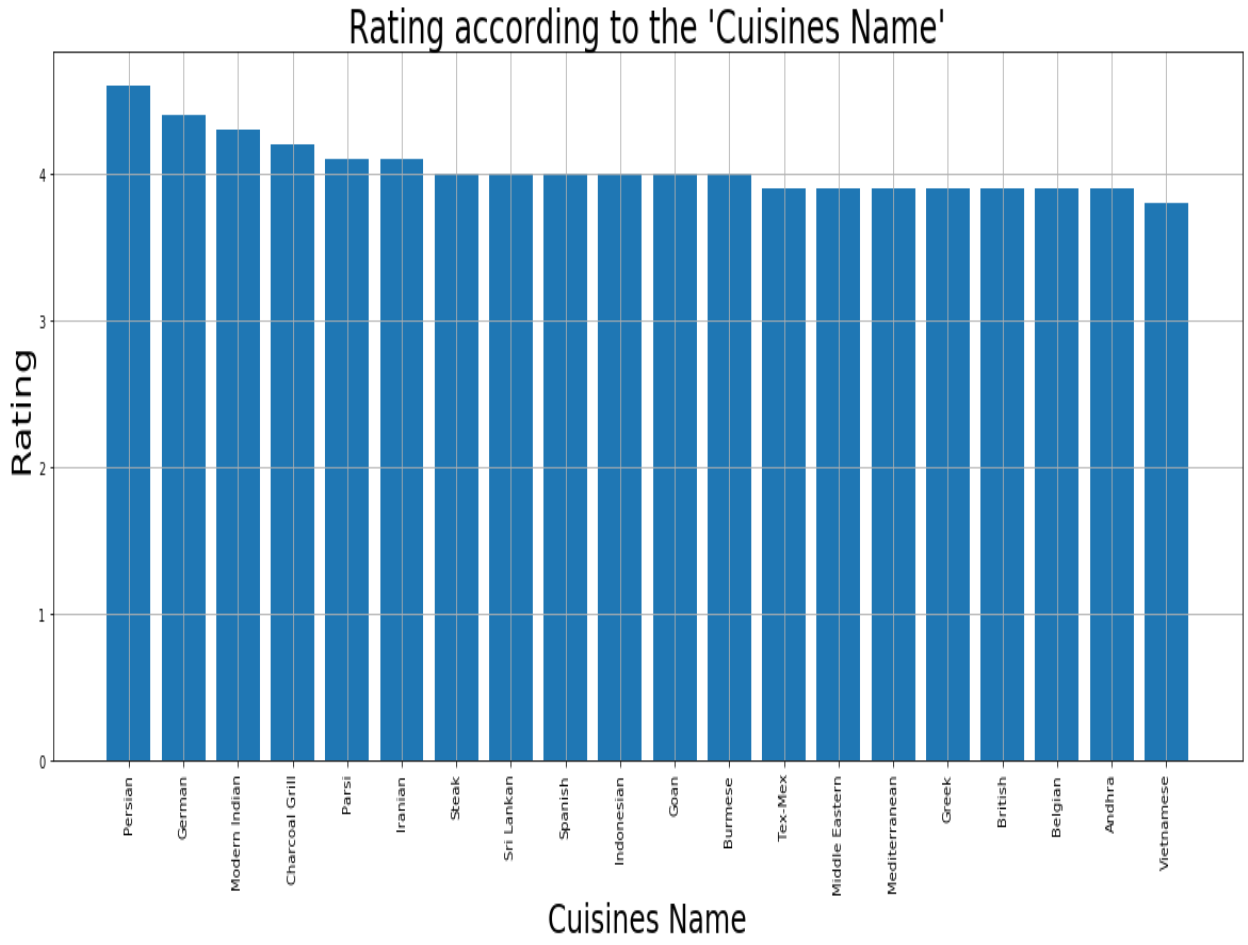
- Here I have used the cost of two as the key and store all the ratings for a particular cost and side by side I have stored the count of restaurants in order to get the average.
- Once all the data is stored in the dictionary I have found the average by total_ratings/restaurant_counts.
- After plotting the graph it is indicating that more cost of two more is the rating of the restaurant.



Question 4: => Restaurant serving some specific cuisines.

Answer:

- As earlier, here I have made one dictionary which is storing cuisines and ratings and the count of restaurants, wherein cuisine works as the key and total_ratings and number of restaurants.
- First I have summed up all the ratings and then I have divided by the total number of restaurants, which is helping me to get the average rating. Once I got the average rating for the cuisines I sorted them in descending order based on their ratings.
- After that I have plotted them in bar graph.



Question 2:=> Find the weighted restaurant rating of each locality and find out the top 10 localities with more weighted restaurant rating?

Weighted Restaurant Rating=Σ (number of votes * rating) / Σ (number of votes)

Answer:

- As we need to find the weighted ratings for each locality I have used one dictionary which will store the locality names and it will store all the values of all restaurants present in that locality.
- As formulation has been mentioned, I have applied that formula to get the weighted rating.
- Once that has been done, I have plotted the graph and from that graph

we can see these are the top 10 localites:

1 Hotel Clarks Amer, Malviya Nagar

2 Aminabad

3 Friends Colony

4 Powai

5 Kirlampudi Layout

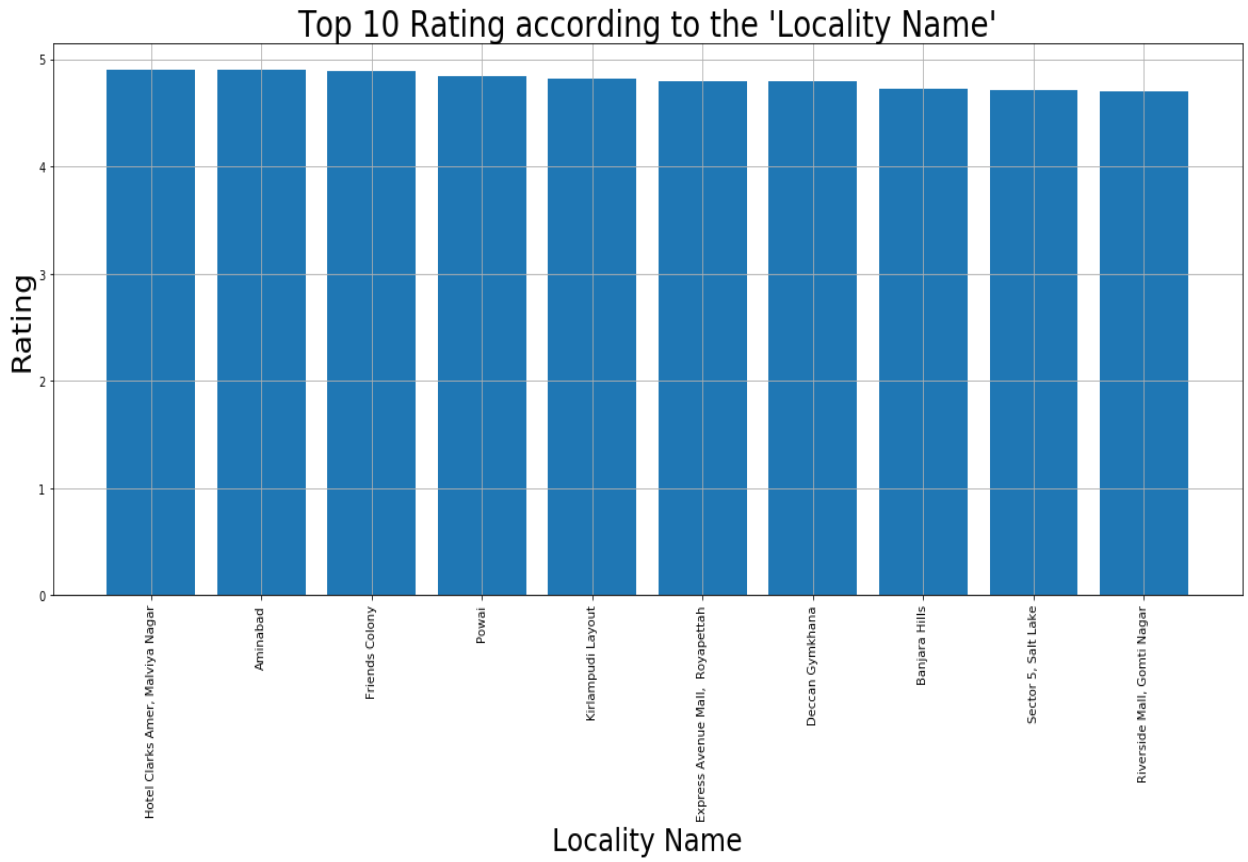
6 Express Avenue Mall, Royapettah

7 Deccan Gymkhana

8 Banjara Hills

9 Sector 5, Salt Lake

10 Riverside Mall, Gomti Nagar



Question 3

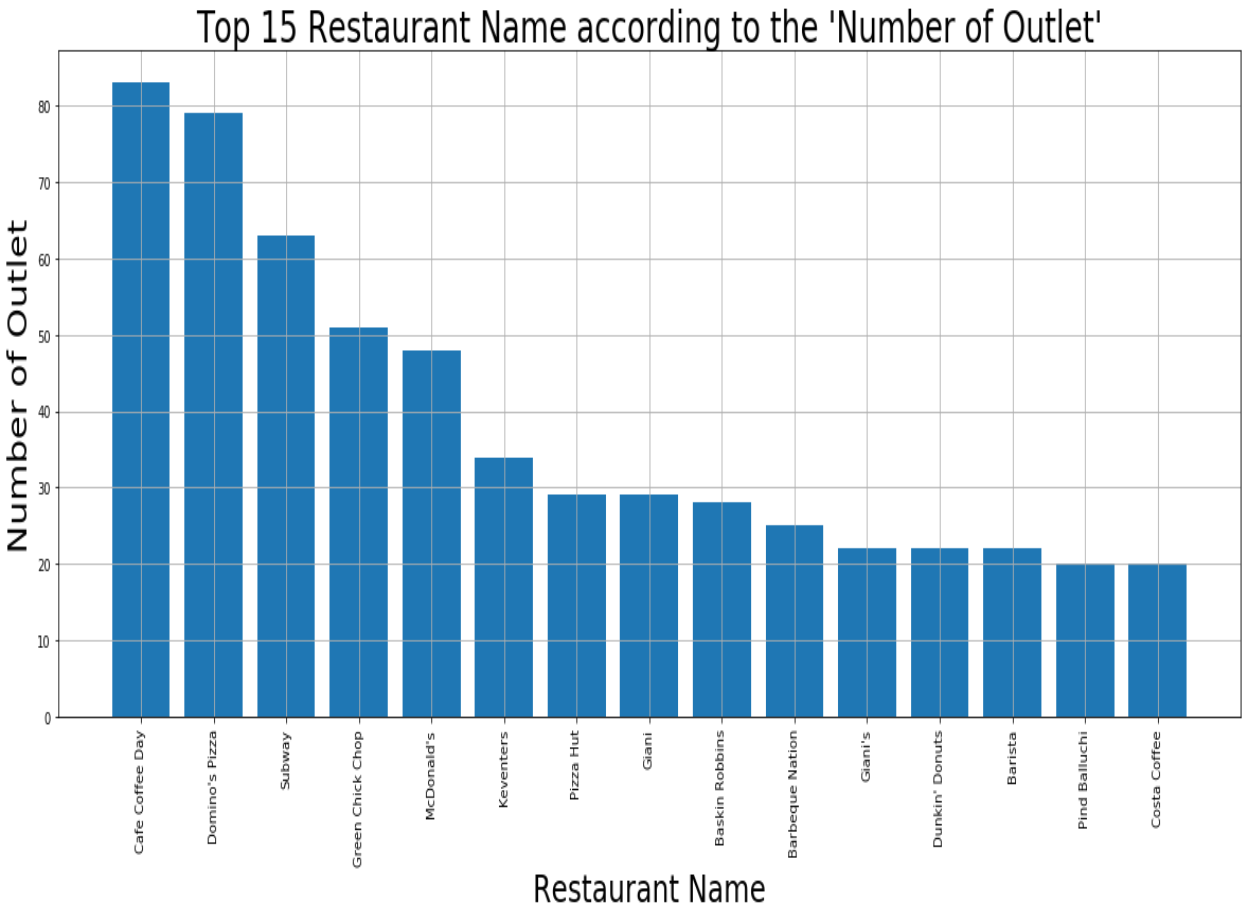
Qo 1 => Plot the bar graph top 15 restaurants that have a maximum number of outlets ?
Answer:

- Here I have considered only Indian Restaurants as it has been mentioned on the top of the Questions. Here I have filtered out only Indian Restaurants first by using boolean indexing.

- Once that's done I have made one dictionary in which I stored the outlet name and made it as key and frequency as the value. After iterating over all the names.

- I got my dictionary fully prepared for the data visualization but as we need to print in descending order so I take array “ans” that stored the data of dictionary. After that I sorted the list based on the frequency.

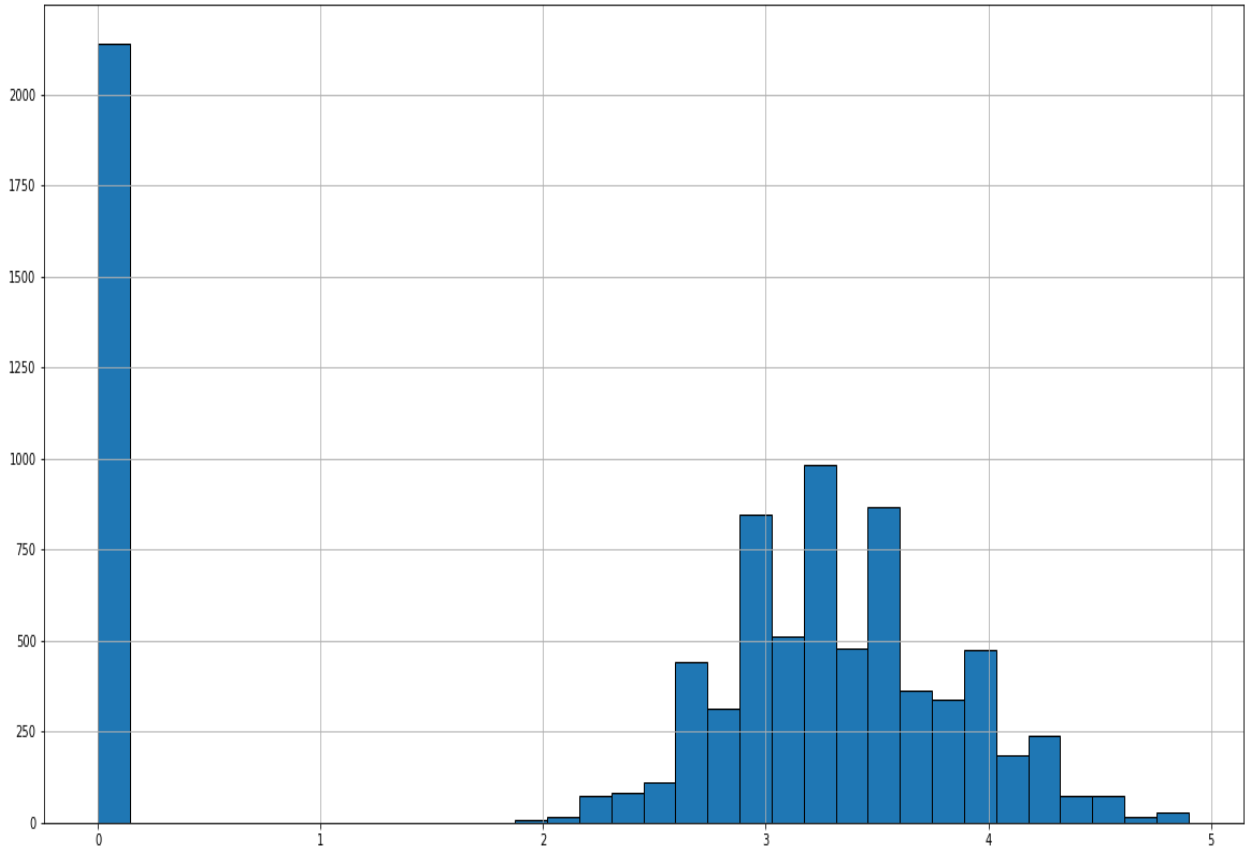
- Plot the answer.



Qo.2 => Plot the histogram of aggregate rating of restaurant(drop the unrated restaurant) ?

Answer:

- As mentioned in the question, first I need to drop all the data which have no Rating.
- After that I am picking values from only Aggregate rating column and storing it to ratings. Just to make sure there is no nan values I have checked and store the values which don't have value as nan.
- After that I plotted its graph and made the bin to auto so that it decide automatically and made xticks to 0-5 as our rating lies between 0-5. From the graph we can see that the number of zero rating is drastically high as compare to other ratings. If we exclude 0 rating then we can notice most of the restaurants got ratings in between 3-4.



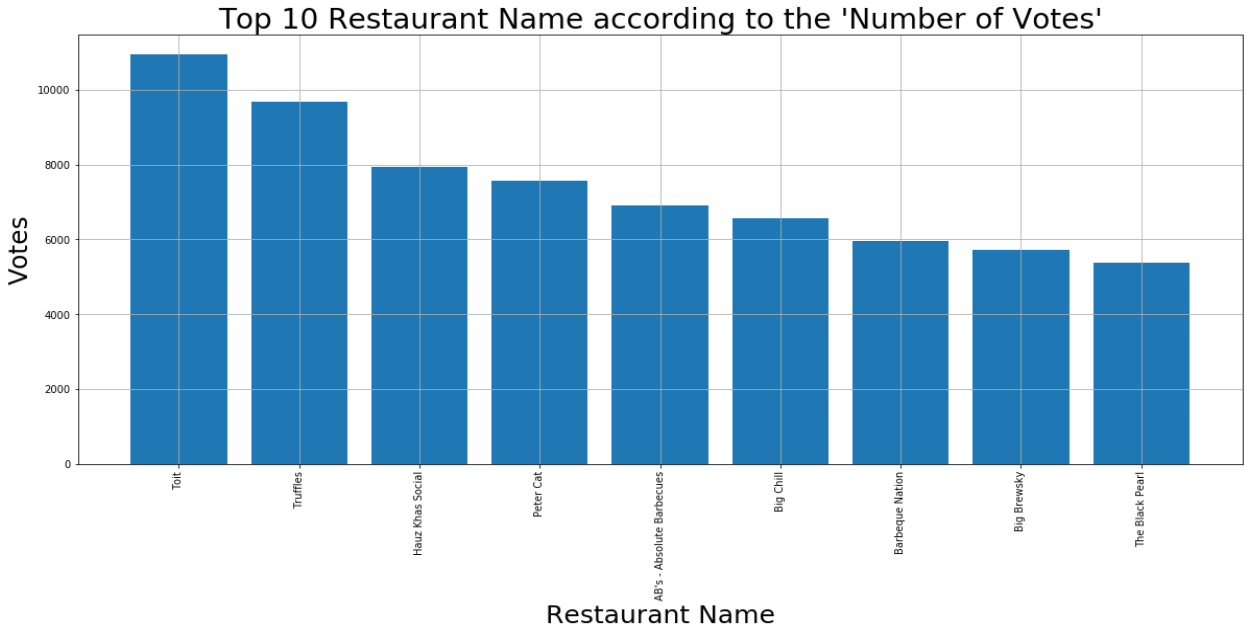
Qo 3 => Plot the bar graph top 10 restaurants in the data with the highest number of votes ?

Answer:

- To get the top 10 restaurants I have one dictionary to get the name of restaurant and number of votes but here one case that can happen is that with the same restaurant in a different location they get less votes and the other one gets an ample number of votes.
- I have made one column wherein I have combined name and locality & votes with ### & $$$. Reason for using two delimeter is that ### will be used to split into two parts i.e name_locality and votes; later $$$ has been used to split into two parts where one part is Name of the restaurant and the locality. I am storing the values after calling the function get_votes.
- Once I stored all the values I have made one list where 0th column contains votes and 1st contains name of the restaurant. Now we have the correct data now I simply sort them in descending order based on the votes and picked top 10 restaurants.

- Output :->

1. Toit
2. Truffles
3. Hauz Khas Social
4. Peter Cat
5. AB's - Absolute Barbecues
6. Big Chill
7. Barbeque Nation
8. Big Brewsky
9. AB's - Absolute Barbecues
10. The Black Pearl

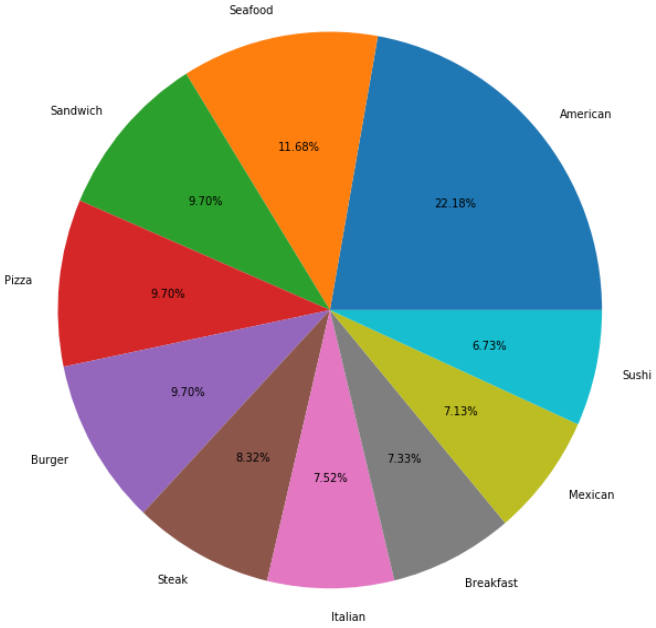


Qo 4 => Plot the pie graph of top 10 cuisines present in restaurants in the USA.

Answer:

- Here we need to filter out based on Country Code first and I have done and stored in zomato_usa_df. I have dropped all nan values present in the cuisines column. I have maintained on dictionary which will help me to keep track of cuisines count. One restaurants might be offering several cuisines. I have used one function inside which I am storing the count.
- Before that I am making sure that I am not skipping any cuisines. Cuisines column consists of all the cuisines names together so I have converted the type and then splited based on combination of space and comma; and updating the count which are in the cuisines list.
- After I got all the count stored in the dictionary then I have made one list of list which store values like 0th column will store the count and 1st column will store the cuisine name. After that I have sorted the list of list based on counts. Now I just took only top 10 cuisines names and its values. Ploting them in the pie chart we can notice

Rank	Cuisine	Percentage
1	American	22.18
2	Seafood	11.68
3	Sandwich	9.7
4	Pizza	9.7
5	Burger	9.7
6	Steak	8.32
7	Italian	7.52
8	Breakfast	7.33
9	Mexican	7.13
10	Sushi	6.73



Qo 5=> Plot the bubble graph of a number of Restaurants present in the city of India and keeping the weighted restaurant rating of the city in a bubble ?

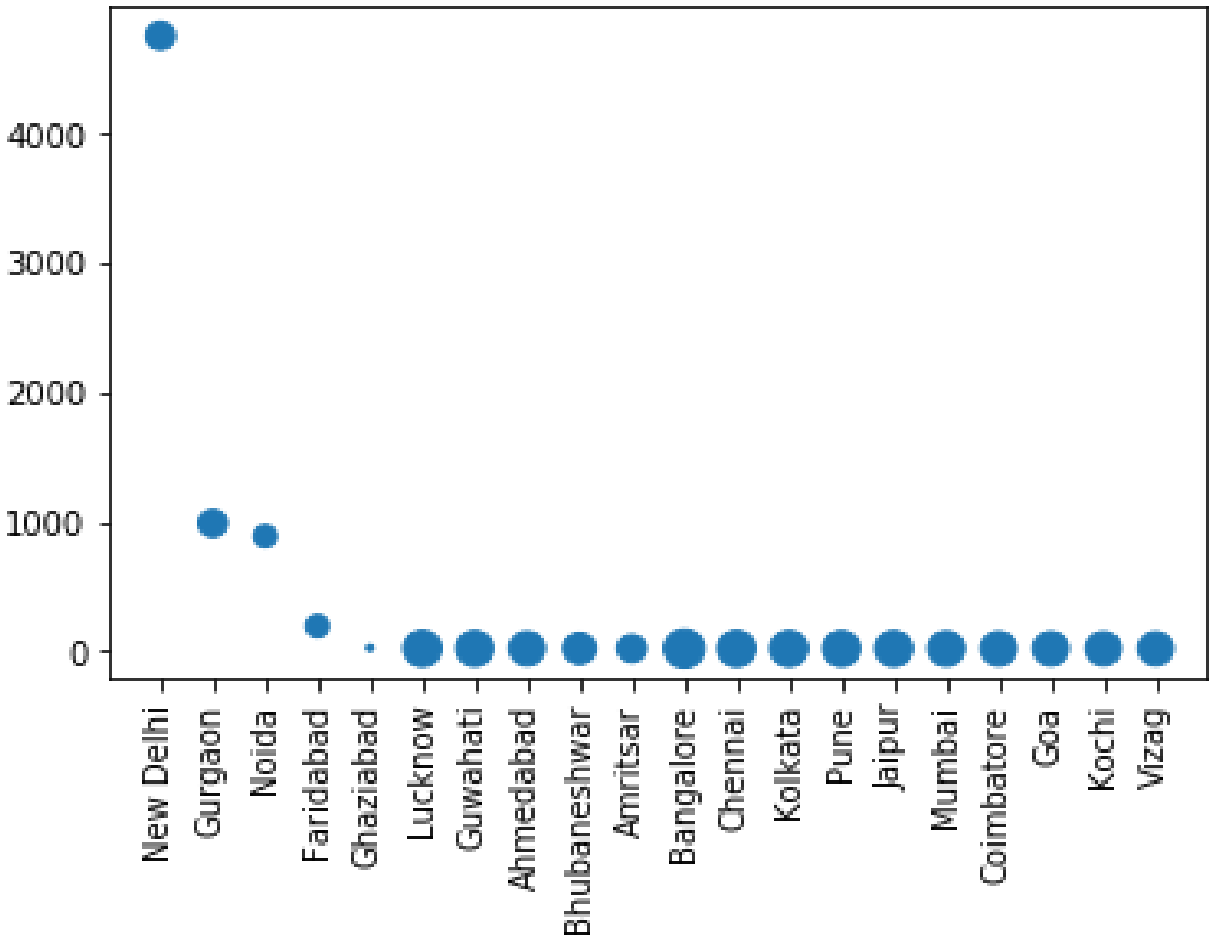
Answer:

- As we need to store the city weighted restaurant rating so I am making one city dictionary which will help me get the city names and the weighted ratings. I have used the same formula which have been mentioned in one of the previous question to find out the. But here I need to keep track of the count of restaurant present in the city so I am maintaining that by updating the count by one whenever I encounter the same city name. So basically here key is city and value is the list wherein 0th storing the summation of votes*ratings, 1st column storing the summation of votes and 2nd column storing the count of restaurants present in that city.
- Once I got the dictionary filled with values I have calculated the weighted rating and stored in a list named city_ratings with city name and total count. after that I have sorted based on the count of restaurants present in the city. So number of cities can be huge so I have considered top 20 cities in India. Delhi topped among the of cities. After that I have plotted them in the graph and provided the size of the bubble based on the weighted rating of the city.

NOTE: I have substracted 3 from each value if any value gets negative I have made it positive. As they are very close to each other so I have multiplied with 100 to get the proper visualization.

Output :

SL	City	Rating
1	New Delhi	3.77
2	Gurgaon	3.74
3	Noida	3.47
4	Faridabad	3.48
5	Ghaziabad	3.04
6	Lucknow	4.32
7	Guwahati	4.27
8	Ahmedabad	4.16
9	Bhubaneshwar	3.97
10	Amritsar	3.76
11	Bangalore	4.5
12	Chennai	4.32
13	Kolkata	4.3
14	Pune	4.28
15	Jaipur	4.28
16	Mumbai	4.22
17	Coimbatore	4.17
18	Goa	4.16
19	Kochi	4.14
20	Vizag	4.13


4000
3000
2000
1000
0
New Delhi
Gurgaon
Noida
Faridabad
Ghaziabad
Lucknow
Guwahati
Ahmedabad
Bhubaneshwar
Amritsar
Bangalore
Chennai
Kolkata
Pune
Jaipur
Mumbai
Coimbatore
Goa
Kochi
Vizag