Project Zomato API-2

Q1.1

The dataset is highly skewed toward the cities included in Delhi-NCR. So, we will summarise all the other cities in 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.

Explanation:

Read the data from Zomato.CSV file. Filter the country by applying country code for India such that the data of India will be available. Create a dataframe which contains cities of Delhi NCR region and count the number of restaurants. Similarly create a dataframe for the Cities of rest of India and Count the number of restaurants.

Use a bar graph for plotting locations on X-axis and No of restaurants on Y-axis. It will clearly show the difference between the no of restaurants in these two regions.

Code:
```
import pandas as pd
import matplotlib.pyplot as plt

path="C:\\Users\\RAHUL\\Documents\\Project _Zomato API-2\\zomato.csv"

data=pd.read_csv(path,encoding='latin')
df=data.copy()
df=df[(df["Country Code"]==1)]

NCR_cities=['New Delhi','Ghaziabad','Noida','Gurgaon','Faridabad']

a=df[df.City.isin(NCR_cities)]
a=a['City'].value_counts().sum()

b=df[~df.City.isin(NCR_cities)]
b=b['City'].value_counts().sum()

plt.bar(['Delhi_NCR','Rest Of India'],[a,b], color='y', edgecolor='black')
plt.xlabel('Cities')
plt.ylabel('No. of restaurants')
plt.title("Number of Restaurants in Delhi NCR and Rest of India", color='green')
plt.show()
```
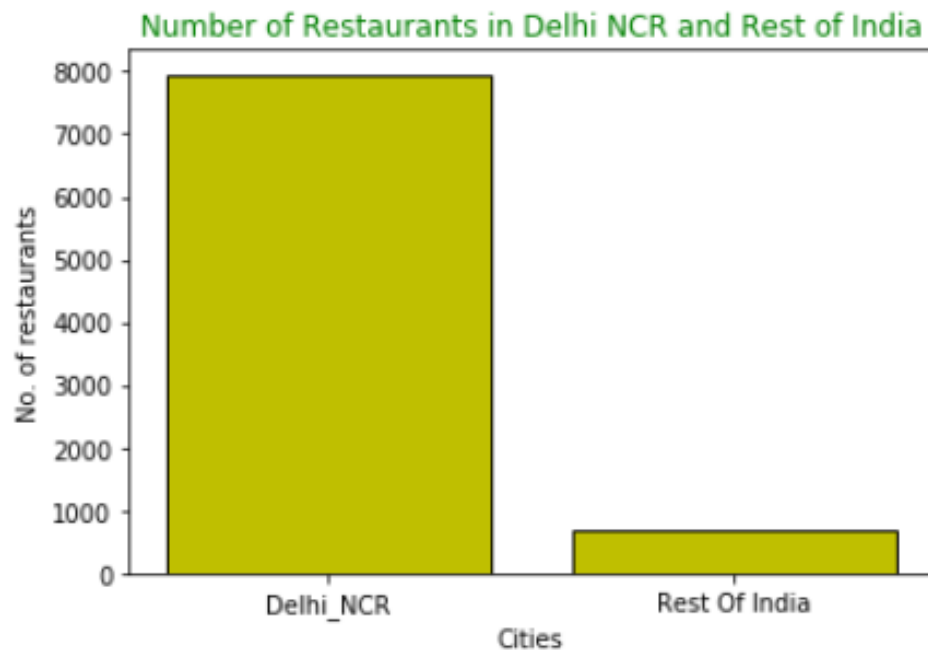
Number of Restaurants in Delhi NCR and Rest of India

Q1.2.

Find the cuisines which are not present in restaurant of Delhi NCR but present in rest of India. Check using Zomato API whether these cuisines are actually not served in restaurants of Delhi-NCR or just it due to incomplete dataset.

Explanation:

Read the data from Zomato.CSV file. Filter the country by applying country code for India such that the data of India will be available.

Created 2 sets one for cuisines of Delhi NCR and another one for Rest of India . After that traversed through the cuisines column and added the cuisines to the sets according to their locations. After that found out the elements that are present in rest of India cuisine set but not in Delhi NCR set.

According to the csv file that is provided to us, there are 4 cuisines that are present in rest of India but not in Delhi NCR , those are Malwani, BBQ, Cajun, German but according to zomato API BBQ and German are served in the restaurants of Delhi-NCR.

Output is incorrect due to incomplete information.

Code:

```
import pandas as pd

import matplotlib.pyplot as plt

path="C:\\Users\\RAHUL\\Documents\\Project _Zomato API-2\\zomato.csv"

data=pd.read_csv(path,encoding='latin')

df=data.copy()

df=df[(df["Country Code"]==1)]

NCR_cities=['New Delhi','Ghaziabad','Noida','Gurgaon','Faridabad']

servedInNCR=set()

serevedInROI=set()

for i in df.index:

    if df["City"][i] in NCR_cities:

        c=df["Cuisines"][i].strip().split(',')

        for j in c:

            servedInNCR.add(j.strip())

    else:

        c=df["Cuisines"][i].strip().split(',')

        for j in c:

            serevedInROI.add(j.strip())


for i in (serevedInROI- servedInNCR):

    print(i)
```

Output:

```
Malwani
Cajun
German
BBQ
```

Q1.3.

Find the top 10 cuisines served by maximum number of restaurants in Delhi NCR and rest of India.

Explanation:

After reading the data from zomato.csv file, filtered out those rows having Country code one. Created 2 dictionaries one for cuisines of Delhi NCR and another one for Rest of India, where keys are names of cuisine and values are number of restaurants in which it are served. After that sorted the cuisines according to the number of restaurants in which it is served for both the locations and at last printed the top 10 cuisines of Delhi NCR and Rest of India.

Code:

```
import pandas as pd

import matplotlib.pyplot as plt

import numpy as np

path="C:\\Users\\RAHUL\\Documents\\Project _Zomato API-2\\zomato.csv"

data=pd.read_csv(path,encoding='latin')

df=data.copy()

df=df[(df["Country Code"]==1)]

NCR_cities=['New Delhi','Ghaziabad','Noida','Gurgaon','Faridabad']

cusinies_NCR={}

b=df[df.City.isin(NCR_cities)]

for i in b.index:

    for j in b['Cuisines'][i].split(','):

        j=j.strip()

        if j in cusinies_NCR:

            cusinies_NCR[j]+=1

        else:

            cusinies_NCR[j]=1
```

```
d1=sorted(cusinies_NCR, key=cusinies_NCR.get, reverse=True)[:10]

print("Top 10 cuisines which are served in Delhi NCR region are:-")

for i in d1:

    print(i)

print()

a=df[~df.City.isin(NCR_cities)]

cus={}

for i in a.index:

    for j in a['Cuisines'][i].split(','):

        j=j.strip()

        if j in cus:

            cus[j]+=1

        else:

            cus[j]=1


d2=sorted(cus, key=cus.get, reverse=True)[:10]

print("Top 10 cuisines which are served in Rest of India are:-")

for i in d2:

    print(i)
```

Output:

Top 10 cuisines which are served in Delhi NCR region are:-
North Indian
Chinese
Fast Food
Mughlai
Bakery
South Indian
Continental
Desserts
Street Food
Italian

Top 10 cuisines which are served in Rest of India are:-
North Indian
Chinese
Continental
Italian
Cafe
Fast Food
South Indian
Mughlai
Desserts
Mexican

Q1.4.

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.

Explanation:

By plotting pie charts for Top 10 cuisines served in Delhi NCR region and rest of the India, we can conclude that:

1. In Delhi NCR and rest of India "North India" and "Chinese" cuisines are mostly served.
2. In Delhi NCR "fastfood" and 'Mughlai" is famous whereas in rest of India "Continental" and 'Italian' is famous.
3. South Indian and Desserts are almost equally famous in both the areas.
4. In Delhi NCR "Bakery" cuisines are more famous as compared to Cafe cuisines which are famous in rest of India.

Code:

```
import pandas as pd

import matplotlib.pyplot as plt

import numpy as np

path="C:\\Users\\RAHUL\\Documents\\Project _Zomato API-2\\zomato.csv"

data=pd.read_csv(path,encoding='latin')

df=data.copy()

df=df[(df["Country Code"]==1)]

NCR_cities=['New Delhi','Ghaziabad','Noida','Gurgaon','Faridabad']

cusinies_NCR={}
```

```python
b=df[df.City.isin(NCR_cities)]

for i in b.index:

    for j in b['Cuisines'][i].split(','):

        j=j.strip()

        if j in cusinies_NCR:

            cusinies_NCR[j]+=1

        else:

            cusinies_NCR[j]=1

d1=sorted(cusinies_NCR.items(), key=lambda x: x[1], reverse=True)

x1=[]

cities=[]

for i in range(10):

    x1.append(d1[i][1])

    cities.append(d1[i][0])

explode=[0,0,0,0.2,0.2,0.2,0.2,0.2,0,0]

plt.title("Top 10 cuisines served in Delhi_NCR", color='green', y=1.27)

plt.pie(x1, labels=cities, autopct="%.2f%%" , explode=explode, radius=1.55)

plt.show()

a=df[~df.City.isin(NCR_cities)]

cus={}

for i in a.index:

    for j in a['Cuisines'][i].split(','):

        j=j.strip()
```

```
        if j in cus:

            cus[j]+=1

        else:

            cus[j]=1

d2 =sorted(cus.items(), key=lambda x: x[1], reverse=True)

x2=[]

cities=[]

for i in range(10):

    x2.append(d2[i][1])

    cities.append(d2[i][0])

explode=[0,0,0,0.2,0.2,0.2,0.2,0.2,0,0]

plt.title("Top 10 cuisines served in Rest Of India", color='green', y=1.27)

plt.pie(x1, labels=cities, autopct="%.2f%%" , explode=explode, radius=1.55,)

plt.show()
```
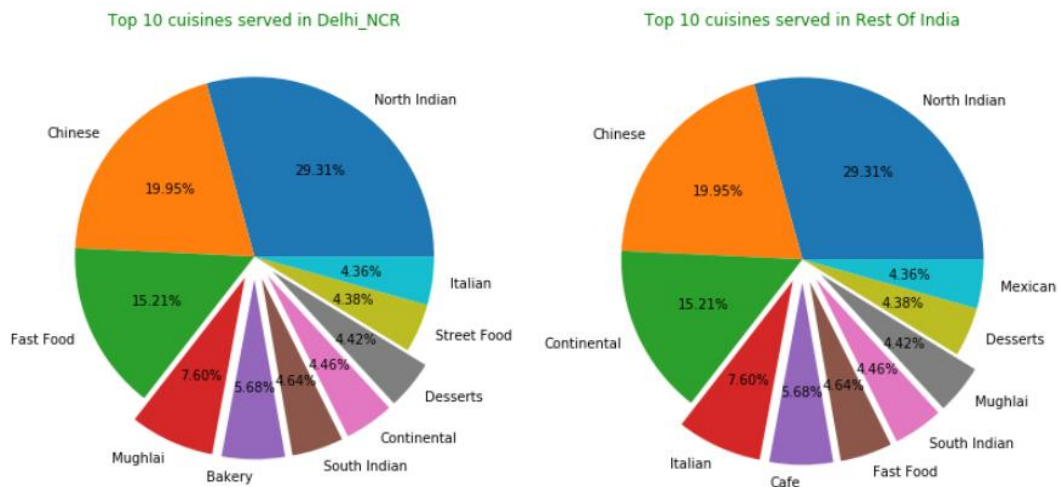
Visualization:

Q2.

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

Write a short detail analysis of how the rating is affected by restaurant due following features: Plot a suitable graph to explain your inference.

1. **Number of Votes given Restaurant**

Explanation:

From the graph(Number of Votes vs Aggregate Rating), we can say that restaurants that received votes less than around 400 rated between 2 and 5 , restaurants that received votes between 400 and 1000 rated between 3 and 5, restaurants that received votes between 1000 and 2000 rated between 3.5 and 5 and restaurants that received votes above 2000 generally rated above 3.4.

Code:

```
import pandas as pd

import matplotlib.pyplot as plt

import numpy as np

path="C:\\Users\\RAHUL\\Documents\\Project _Zomato API-2\\zomato.csv"

data=pd.read_csv(path,encoding='latin')

df=data.copy()

# df=df[(df["Country Code"]==1)]

plt.scatter(df["Votes"],df["Aggregate rating"], color='b', marker='.')

plt.xlabel('Number of votes')

plt.ylabel('Aggregate Rating')

plt.title('Number of Votes vs Aggregate rating\n', color='r')

plt.show()
```
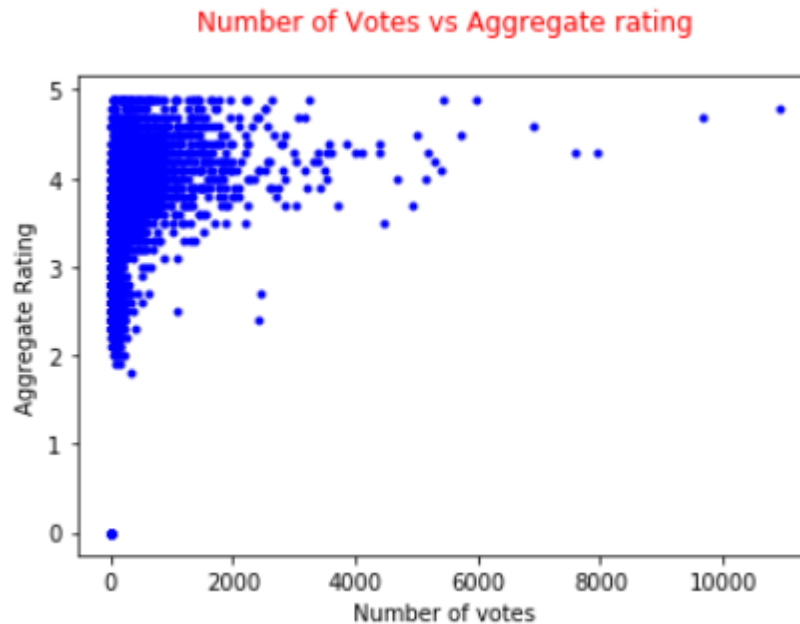
Visualization:

Number of Votes vs Aggregate rating

2. Restaurant serving more number of cuisines.

Explanation:

From the graph (Number of Cuisines Vs Aggregate Rating), we can say that restaurants that serve 1 and 3 cuisines usually rated between 2 and 5, restaurants that serve 2 and 4 cuisines generally rated between 1.7 and 5, restaurants that serve 5 cuisines usually rated between 2.2 and 4.8, there are some restaurants who serve 6 and 7 cuisines and are rated between 2 and 3, rest all the restaurants who serve 6 to 7 cuisines are generally rated between 3 and 4.8,restaurants that serve 8 cuisines are generally rated between 3 and 4.5.

Code:

```
import pandas as pd

import matplotlib.pyplot as plt

import numpy as np

path="C:\\Users\\RAHUL\\Documents\\Project _Zomato API-2\\zomato.csv"

data=pd.read_csv(path,encoding='latin')

df=data.copy()
```
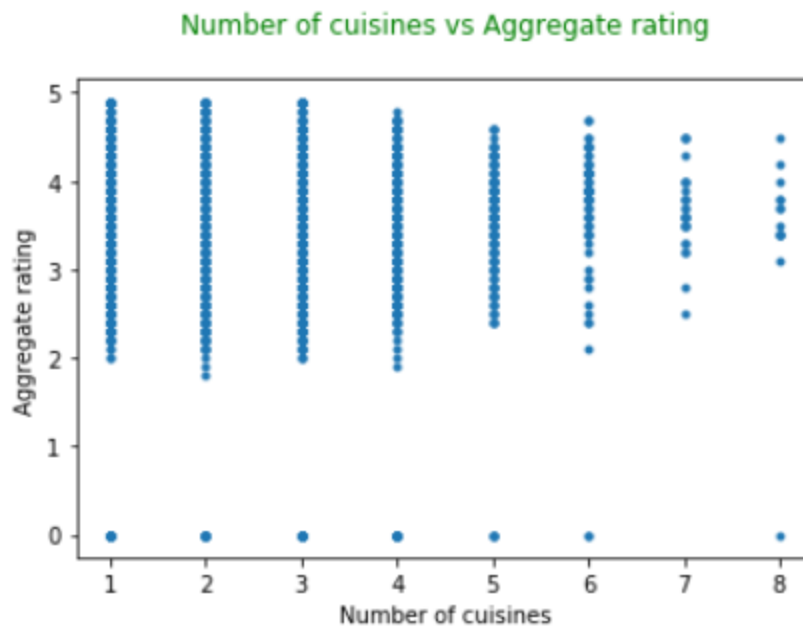
```
# df=df[(df["Country Code"]==1)]

df.dropna(subset=['Cuisines'], inplace=True)

num_of_cuisines=[]

for i in df['Cuisines']:

    num_of_cuisines.append(len(set(i.strip().split(','))))

plt.scatter(num_of_cuisines, df['Aggregate rating'],marker='.')

plt.xlabel('Number of cuisines')

plt.ylabel('Aggregate rating')

plt.title('Number of cuisines vs Aggregate rating\n', color='green')

plt.show()
```

Visualization:

3. **Average Cost of Restaurant**

Explanation:

From the graph(Average Cost vs Aggregate Rating), we can say that restaurants whose average cost for two less than 25000 are generally rated between 1.7 and 5 and all those restaurants whose average cost is above 25000 are usually rated above 3.

Code:

```
import pandas as pd

import matplotlib.pyplot as plt

path="C:\\Users\\RAHUL\\Documents\\Project _Zomato API-2\\zomato.csv"

data=pd.read_csv(path,encoding='latin')

df=data.copy()

# df=df[(df["Country Code"]==1)]

plt.scatter(df['Average Cost for two'], df['Aggregate rating'], marker='.')

plt.xticks(rotation=90)

plt.title("Average Cost vs Aggregate Rating\n", color='Green')

plt.xlabel('Average Cost')

plt.ylabel('Aggregate Rating')

plt.show()
```
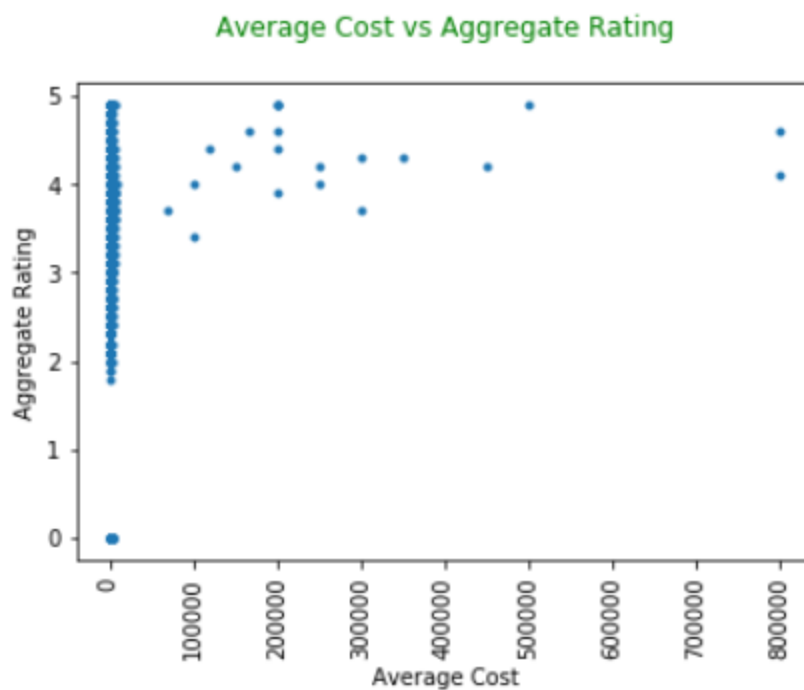
Visualization:

4. Restaurant serving some specific cuisines

Explanation:

From the graph (Number of Cuisines Vs Aggregate Rating), we can say that restaurants serving some specific number of cuisines are generally rated between 2 to 5. Restaurants serving more than 4 cuisines are rated between 2 and 5.

Code:

```python
import pandas as pd

import matplotlib.pyplot as plt

path="C:\\Users\\RAHUL\\Documents\\Project _Zomato API-2\\zomato.csv"

data=pd.read_csv(path,encoding='latin')

df=data.copy()

# df=df[(df["Country Code"]==1)]

df.dropna(subset=['Cuisines'], inplace=True)

num_of_cuisines=[]

for i in df['Cuisines']:

    num_of_cuisines.append(len(i.strip().split(',')))

plt.scatter(num_of_cuisines, df['Aggregate rating'], marker='.')

plt.xlabel('Number of cuisines')

plt.ylabel('Aggregate rating')

plt.title('Number of cuisines vs Aggregate rating\n', color='green')

plt.show()
```

Visualization:

Q3.

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).

Explanation:

After reading the data from zomato.csv file. Took 2 dictionaries, both of them having locality name as the key and one having value, summation of (votes*rating) and other one having value of the summation of number of votes of all the restaurants of that locality.
Then, took one more dictionary, for calculating and saving weighted restaurant rating for each locality.
At last, sorted the localities according to their weighted restaurant rating and filtered out the top 10 among them.

Code:

```
import pandas as pd

import matplotlib.pyplot as plt

path="C:\\Users\\RAHUL\\Documents\\Project _Zomato API-2\\zomato.csv"

data=pd.read_csv(path,encoding='latin')

df=data.copy()
```

```python
# df=df[(df["Country Code"]==1)]

votes_x_rating={}

v={}

for i in df.index:

    loc= df['Locality'][i].strip()

    votes=df["Votes"][i]

    rating=df["Aggregate rating"][i]

    votes_x_rating[loc]=votes_x_rating.get(loc,0)+(votes*rating)

    v[loc]=v.get(loc,0)+votes

weighted_rest_rating={}

for i in votes_x_rating:

    if v[i]==0:

        weighted_rest_rating[i]=0

    else:

        weighted_rest_rating[i]= votes_x_rating[i]/v[i]

d1= sorted(weighted_rest_rating, key=weighted_rest_rating.get, reverse= True)[:10]

for i in d1:

    print(i)
```

Output:

Sofitel Philippine Plaza Manila, Pasay City
Setor De Clubes Esportivos Sul
Gíçvea
The Milk District
Windermere
Paia
Kenwood
New Tampa
Venetian Village, Al Maqtaa
Deira City Centre Area

Q4. Visualization

1. **Plot the bar graph top 15 restaurants have a maximum number of outlets.**

Explanation:

After reading the data, created a dictionary where key is restaurant's name and value is the number of its outlets.
Then, created a list containing top 15 restaurants having maximum number of branches.
Then, Create one more list containing number of outlets of those restaurants.
At last created a bar graph between restaurants and number of outlets to indicate which the number of outlets each restaurants have. Also from the fraph we can see that which restaurant has most and least number of the outlets.

Code:

```
import pandas as pd
import matplotlib.pyplot as plt

path="C:\\Users\\RAHUL\\Documents\\Project _Zomato API-2\\zomato.csv"

data=pd.read_csv(path,encoding='latin')
df=data.copy()
# df=df[(df["Country Code"]==1)]

res_name={}
for i in df.index:
    res=df['Restaurant Name'][i]
    res_name[res]=res_name.get(res,0)+1

restaurants=sorted(res_name, key=res_name.get, reverse=True)[:15]
outlet_count=[]

for i in restaurants:
    outlet_count.append(res_name[i])

plt.bar(restaurants,outlet_count,color='y', edgecolor='r')
plt.xticks(rotation=90)
plt.xlabel('Retaurants')
plt.xlabel('Number of outlets')
plt.title("Restaurants vs Number of outlets", color='green')
plt.show()
```
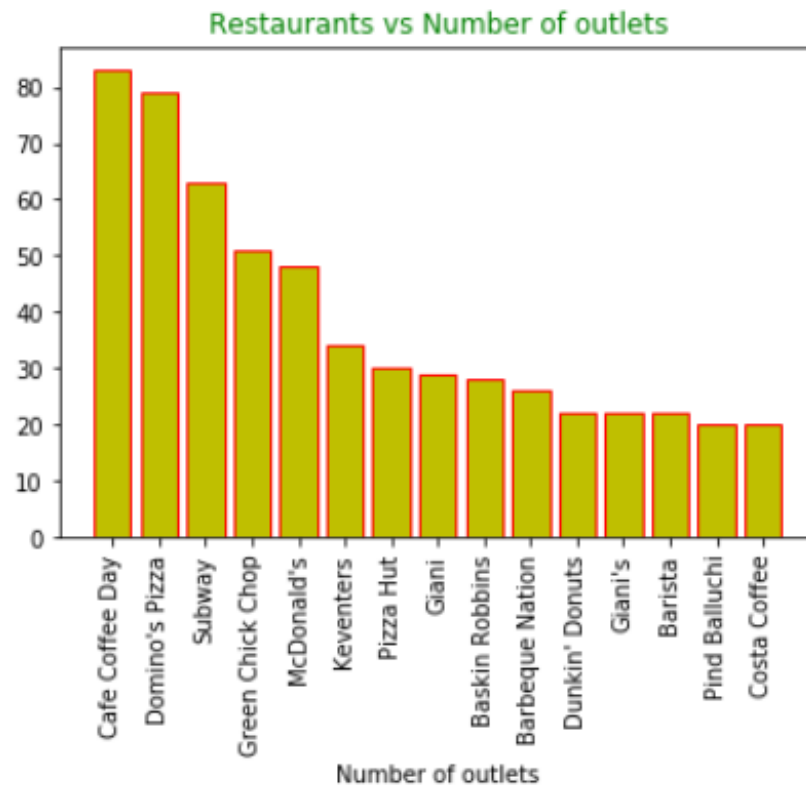
Visualization:

Restaurants vs Number of outlets

2. Plot the histogram of aggregate rating of restaurant (drop the unrated restaurant).

Explanation:

After reading the data from zomato.csv file, removed all the unrated restaurants. Unrated restaurants are those which have rating of 0. Then, created a list and appended aggregate rating of all the restaurants. At last created a histogram using the data of that list, to get the number of restaurants having particular rating and most restaurants having rating of 3.3 and then 3.7

Code:

import pandas as pd

import matplotlib.pyplot as plt

import numpy as np

path="C:\\Users\\RAHUL\\Documents\\Project _Zomato API-2\\zomato.csv"

data=pd.read_csv(path,encoding='latin')

```
df=data.copy()

# df=df[(df["Country Code"]==1)]

df=df[df['Aggregate rating']!=0.0]

agg_rating=[]

for i in df.index:

    agg_rating.append(df['Aggregate rating'][i])

p= np.arange(0,5.1,0.1)

plt.hist(agg_rating, bins=p, color='pink', edgecolor='black')

plt.xlabel('Rating')

plt.ylabel('Number of Restaurants')

plt.title('Rating vs Number of Restaurants', color='green')

plt.show()
```

Visualization:
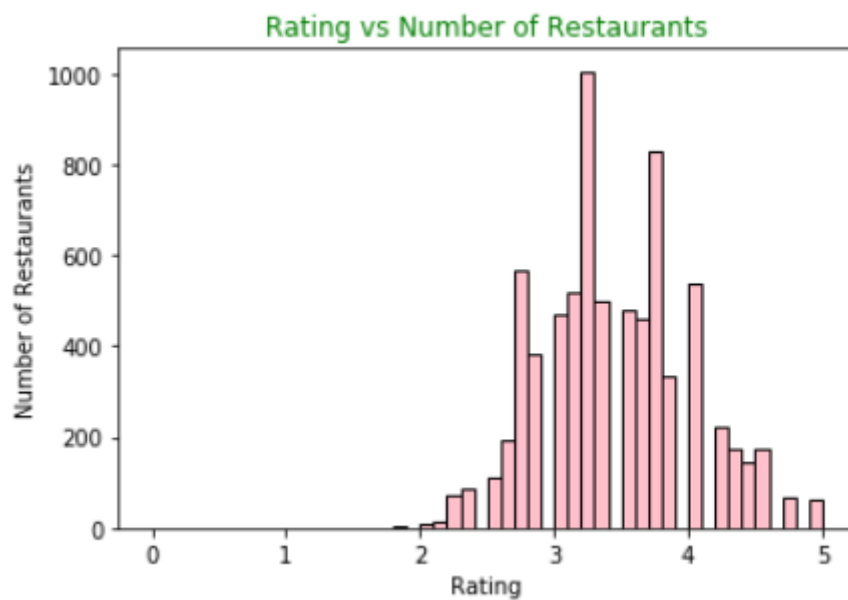
3. Plot the bar graph top 10 restaurants in the data with the highest number of votes.

Explanation:

After reading the data from zomato.csv file, created a dictionary having keys as the name of the restaurant and value as the number of votes they received. Created a new list having names of all the restaurants sorted in decreasing order according to their value, filtered out top 10 from that list and created 2 lists one having the names of the restaurants and other one having respected votes they had received. At last created a bar graph using those two list.

Code:

```
import pandas as pd

import matplotlib.pyplot as plt

import numpy as np

path="C:\\Users\\RAHUL\\Documents\\Project _Zomato API-2\\zomato.csv"

data=pd.read_csv(path,encoding='latin')

df=data.copy()

# df=df[(df["Country Code"]==1)]

df=df[df['Votes']!=0]

res={}

for i in df.index:

    votes=df['Votes'][i]

    res_name=df['Restaurant Name'][i].strip()

    res[res_name]=res.get(res_name,0)+votes

    sorted_res= sorted(res, key=res.get, reverse=True)[:10]

x,y=[],[]

for i in sorted_res:

    x.append(i)

    y.append(res[i])

    plt.bar(x,y, color='pink', edgecolor='black')

plt.xlabel('Restaurant')

plt.ylabel('Number of Votes')
```
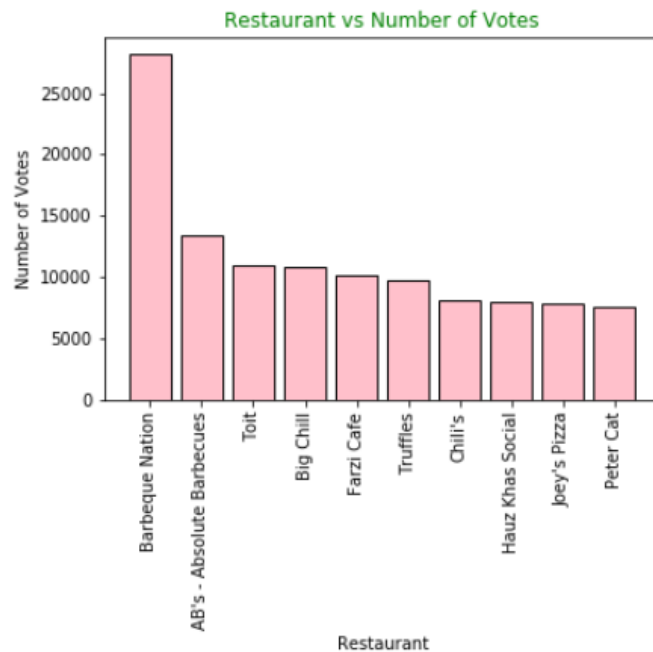
plt.xticks(rotation=90)

plt.title('Restaurant vs Number of Votes', color='green')

plt.show()

Visualization:

4. **Plot the pie graph of top 10 cuisines present in restaurants in the USA.**
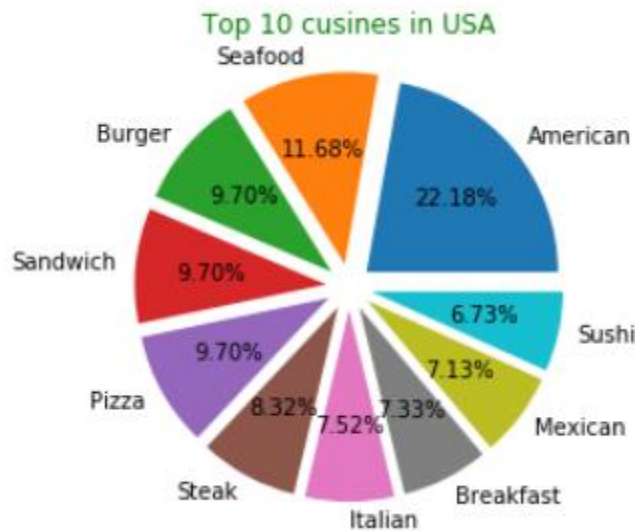
Explanation:

After reading the data from zomato.csv file, filtered out all the rows having country code 216. Dropped all the nan values from cuisines. Created a dictionary to keep a record of most served cuisine in maximum number of restaurants in the USA. Created a list of top 10 cuisines served using the values of the dictionary and another list containing the number of restaurants in which it is served. At last created a pie graph using the details of both the lists which clearly indicated the % distribution of the cuisines served in the restaurants of USA.

American and seafood are top cuisines which are present in most restaurants of USA.

Code:

```python
import pandas as pd

import matplotlib.pyplot as plt

import numpy as np

path="C:\\Users\\RAHUL\\Documents\\Project _Zomato API-2\\zomato.csv"

data=pd.read_csv(path,encoding='latin')

df=data.copy()

df=df[(df["Country Code"]==216)]

df.dropna(subset=['Cuisines'], inplace=True)

cus={}

for i in df.index:

    cuisine=df['Cuisines'][i].strip().split(',')

    for j in cuisine:

        j=j.strip()

        cus[j]=cus.get(j,0)+1

cus_name= sorted(cus, key=cus.get, reverse=True)[:10]

cus_count=[]

for i in cus_name:

    cus_count.append(cus[i])

exp=[0.12,0.12,0.12,0.12,0.12,0.12,0.12,0.12,0.12,0.12]

plt.pie(cus_count, autopct='%.2f%%', labels=cus_name, explode=exp)

plt.title('Top 10 cusines in USA', color='green')

plt.show()
```

Visualization:

Top 10 cusines in USA

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.**

Explanation:

After reading the data from zomato.csv file, filtered out all the rows having country code 1. Created 3 dictionaries all having key the city name and first one having value the multiplication of votes and rating and second one having value the number of votes and third one having value the number of restaurants in that city. After that created a list weighted_rest_rating having restaurant name as the keys and there weighted rating as their values. After that created three lists x to save the name of the cities, y to save the no of restaurant in a city and third one col to save the weighted restaurant rating. At last created a scatter graph between city and the number of restaurants and weighted restaurant rating of each city in the bubble.

Code:

import pandas as pd

import matplotlib.pyplot as plt

import numpy as np

path="C:\\Users\\RAHUL\\Documents\\Project _Zomato API-2\\zomato.csv"

data=pd.read_csv(path,encoding='latin')

```python
df=data.copy()
df=df[(df["Country Code"]==1)]
votes_x_rating={}
v={}
res={}
for i in df.index:
    city= df['City'][i].strip()
    votes=df["Votes"][i]
    rating=df["Aggregate rating"][i]
    votes_x_rating[city]=votes_x_rating.get(city,0)+(votes*rating)
    v[city]=v.get(city,0)+votes
    res[city]=res.get(city,0)+1
weighted_rest_rating={}
for i in votes_x_rating:
    if v[i]==0:
        weighted_rest_rating[i]=0
    else:
        weighted_rest_rating[i]= "%.2f" % (votes_x_rating[i]/v[i])
x,y=[],[]
col=[]
for i in weighted_rest_rating:
    x.append(i)
    y.append(res[i])
    col.append(float(weighted_rest_rating[i]))
print(weighted_rest_rating)
plt.scatter(x,y, alpha=0.85, marker='o', c=col)
plt.xlabel('Cities')
plt.ylabel('Number of Restaurants')
plt.title('Number of Restaurants vs Cities', color='green')
```

plt.xticks(rotation=90)

plt.show()

Visualization:

Number of Restaurants vs Cities