Sejal Sanas

Create a restaurant recommendation system based on user preferences.

In [1]:
```
# Import Libraries
import numpy as np
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
from sklearn.metrics import jaccard_score
from scipy.spatial.distance import pdist, squareform
pd.reset_option('display.max_rows')
```

In [2]:
```
import warnings
warnings.filterwarnings("ignore")
```

In [3]:
```
# Creating a Dataframe
df = pd.read_csv(r'C:/Users/Shejal Sanas/Downloads/Dataset.csv')
df.head()
```

Out[3]:

	Restaurant ID	Restaurant Name	Country Code	City	Address	Locality	Locality Verbose	Longitude	Latitude	Cuisines	...
0	6317637	Le Petit Souffle	162	Makati City	Third Floor, Century City Mall, Kalayaan Avenu...	Century City Mall, Poblacion, Makati City	Century City Mall, Poblacion, Makati City, Mak...	121.027535	14.565443	French, Japanese, Desserts	...
1	6304287	Izakaya Kikufuji	162	Makati City	Little Tokyo, 2277 Chino Roces Avenue, Legaspi...	Little Tokyo, Legaspi Village, Makati City	Little Tokyo, Legaspi Village, Makati City, Ma...	121.014101	14.553708	Japanese	...
2	6300002	Heat - Edsa Shangri-La	162	Mandaluyong City	Edsa Shangri-La, 1 Garden Way, Ortigas, Mandal...	Edsa Shangri-La, Ortigas, Mandaluyong City	Edsa Shangri-La, Ortigas, Mandaluyong City, Ma...	121.056831	14.581404	Seafood, Asian, Filipino, Indian	...
3	6318506	Ooma	162	Mandaluyong City	Third Floor, Mega Fashion Hall, SM Megamall, O...	SM Megamall, Ortigas, Mandaluyong City	SM Megamall, Ortigas, Mandaluyong City, Mandal...	121.056475	14.585318	Japanese, Sushi	...
4	6314302	Sambo Kojin	162	Mandaluyong City	Third Floor, Mega Atrium, SM Megamall, Ortigas...	SM Megamall, Ortigas, Mandaluyong City	SM Megamall, Ortigas, Mandaluyong City, Mandal...	121.057508	14.584450	Japanese, Korean	...

5 rows × 21 columns

In [4]:
```
df.columns
```

Out[4]:
```
Index(['Restaurant ID', 'Restaurant Name', 'Country Code', 'City', 'Address',
       'Locality', 'Locality Verbose', 'Longitude', 'Latitude', 'Cuisines',
       'Average Cost for two', 'Currency', 'Has Table booking',
       'Has Online delivery', 'Is delivering now', 'Switch to order menu',
       'Price range', 'Aggregate rating', 'Rating color', 'Rating text',
       'Votes'],
      dtype='object')
```

In [5]:
```
dfRS = df[['Restaurant ID','Restaurant Name','Cuisines','Aggregate rating','Votes']]
dfRS
```

Out[5]:

	Restaurant ID	Restaurant Name	Cuisines	Aggregate rating	Votes
0	6317637	Le Petit Souffle	French, Japanese, Desserts	4.8	314
1	6304287	Izakaya Kikufuji	Japanese	4.5	591
2	6300002	Heat - Edsa Shangri-La	Seafood, Asian, Filipino, Indian	4.4	270
3	6318506	Ooma	Japanese, Sushi	4.9	365
4	6314302	Sambo Kojin	Japanese, Korean	4.8	229
...	...	...	...	...	...
9546	5915730	Naml\ Gurme	Turkish	4.1	788
9547	5908749	Ceviz A��ac\	World Cuisine, Patisserie, Cafe	4.2	1034
9548	5915807	Huqqa	Italian, World Cuisine	3.7	661
9549	5916112	A���k Kahve	Restaurant Cafe	4.0	901
9550	5927402	Walter's Coffee Roastery	Cafe	4.0	591

9551 rows × 5 columns

Data Cleaning

In [6]:
```
# Gathering information of every columns

# Columns Description
def dataDesc():
    listItem = []
    for col in dfRS.columns :
        listItem.append(
            [col,
            dfRS[col].dtype,
            dfRS[col].isna().sum(),
            round(dfRS[col].isna().sum()/len(dfRS)*100,2),
            dfRS[col].nunique(),
            list(dfRS[col].drop_duplicates().sample(2).values)]
        )
    descData = pd.DataFrame(data = listItem,
                            columns = ['Column','Data Type', 'Missing Value',
                                       'Pct Missing Value', 'Num Unique', 'Unique Sample'])
    return descData

dataDesc()
```

Out[6]:

	Column	Data Type	Missing Value	Pct Missing Value	Num Unique	Unique Sample
0	Restaurant ID	int64	0	0.00	9551	[18241883, 18420675]
1	Restaurant Name	object	0	0.00	7446	[Shaan-E-Tandoorz, Tawa King]
2	Cuisines	object	9	0.09	1825	[Italian, Continental, North Indian, Italian, ...
3	Aggregate rating	float64	0	0.00	33	[4.4, 4.1]
4	Votes	int64	0	0.00	1012	[278, 184]

In [7]:
```
dfRS = dfRS.dropna()
```

Loading [MathJax]/extensions/Safe.js

In [8]:
```
dfRS
```

Out[8]:

	Restaurant ID	Restaurant Name	Cuisines	Aggregate rating	Votes
0	6317637	Le Petit Souffle	French, Japanese, Desserts	4.8	314
1	6304287	Izakaya Kikufuji	Japanese	4.5	591
2	6300002	Heat - Edsa Shangri-La	Seafood, Asian, Filipino, Indian	4.4	270
3	6318506	Ooma	Japanese, Sushi	4.9	365
4	6314302	Sambo Kojin	Japanese, Korean	4.8	229
...	...	...	...	...	...
9546	5915730	Naml\ Gurme	Turkish	4.1	788
9547	5908749	Ceviz A��ac\	World Cuisine, Patisserie, Cafe	4.2	1034
9548	5915807	Huqqa	Italian, World Cuisine	3.7	661
9549	5916112	A���k Kahve	Restaurant Cafe	4.0	901
9550	5927402	Walter's Coffee Roastery	Cafe	4.0	591

9542 rows × 5 columns

In [9]:
```
# Renaming the Columns
dfRS = dfRS.rename(columns={'Restaurant ID': 'restaurant_id'})
dfRS = dfRS.rename(columns={'Restaurant Name': 'restaurant_name'})
dfRS = dfRS.rename(columns={'Cuisines': 'cuisines'})
dfRS = dfRS.rename(columns={'Aggregate rating': 'aggregate_rating'})
dfRS = dfRS.rename(columns={'Votes': 'votes'})
```

In [10]:
```
dfRS
```

Out[10]:

	restaurant_id	restaurant_name	cuisines	aggregate_rating	votes
0	6317637	Le Petit Souffle	French, Japanese, Desserts	4.8	314
1	6304287	Izakaya Kikufuji	Japanese	4.5	591
2	6300002	Heat - Edsa Shangri-La	Seafood, Asian, Filipino, Indian	4.4	270
3	6318506	Ooma	Japanese, Sushi	4.9	365
4	6314302	Sambo Kojin	Japanese, Korean	4.8	229
...	...	...	...	...	...
9546	5915730	Naml\ Gurme	Turkish	4.1	788
9547	5908749	Ceviz A��ac\	World Cuisine, Patisserie, Cafe	4.2	1034
9548	5915807	Huqqa	Italian, World Cuisine	3.7	661
9549	5916112	A���k Kahve	Restaurant Cafe	4.0	901
9550	5927402	Walter's Coffee Roastery	Cafe	4.0	591

9542 rows × 5 columns

In [11]:
```
# Check for Duplicates
dfRS.duplicated().sum()
```

Out[11]:
```
0
```

In [12]:
```
dfRS['restaurant_name'].duplicated().sum()
```

Out[12]:
```
2105
```

In [13]:
```
dfRS['restaurant_name'].value_counts()
```

Out[13]:
```
Cafe Coffee Day            83
Domino's Pizza             79
Subway                     63
Green Chick Chop           51
McDonald's                 48
                           ..
The Town House Cafe         1
The G.T. Road               1
The Darzi Bar & Kitchen     1
Smoke On Water              1
Walter's Coffee Roastery    1
Name: restaurant_name, Length: 7437, dtype: int64
```

In [14]:
```
dfRS = dfRS.sort_values(by=['restaurant_name','aggregate_rating'],ascending=False)
```

In [15]:
```
dfRS[dfRS['restaurant_name']=="Domino's Pizza"].head()
```

Out[15]:

	restaurant_id	restaurant_name	cuisines	aggregate_rating	votes
3031	143	Domino's Pizza	Pizza, Fast Food	3.7	336
1844	5065	Domino's Pizza	Pizza, Fast Food	3.6	146
2448	15078	Domino's Pizza	Pizza, Fast Food	3.6	86
7618	18263236	Domino's Pizza	Pizza, Fast Food	3.6	24
8437	384	Domino's Pizza	Pizza, Fast Food	3.6	547

In [16]:
```
# Dropping duplicaes only keeping first Value.
dfRS = dfRS.drop_duplicates('restaurant_name',keep='first')
dfRS
```

Out[16]:

	restaurant_id	restaurant_name	cuisines	aggregate_rating	votes
9523	6000871	ukura a Sofras	Kebab, Izgara	4.4	296
3120	18222559	{Niche} - Cafe & Bar	North Indian, Chinese, Italian, Continental	4.1	492
9334	7100938	wagamama	Japanese, Asian	3.7	131
9454	6401789	tashas	Cafe, Mediterranean	4.1	374
4659	18361747	t Lounge by Dilmah	Cafe, Tea, Desserts	3.6	34
...	...	...	...	...	...
8692	18317511	#Urban Caf	North Indian, Chinese, Italian	3.3	49
6998	18336489	#OFF Campus	Cafe, Continental, Italian, Fast Food	3.7	216
2613	18311951	#InstaFreeze	Ice Cream	0.0	2
9148	18378803	#Dilliwaala6	North Indian	3.7	124
2459	3100446	#45	Cafe	3.6	209

7437 rows × 5 columns

In [17]:
```
dfRS['restaurant_name'].value_counts()
```

Out[17]:
```
ukura a Sofras          1
French Toast            1
Fourteen Eleven Tea Cafe    1
Fozzie's Pizzaiolo      1
Frasers                 1
                       ..
Pizza Street            1
Pizza Treat             1
Pizza Yum               1
Pizza  Bessa            1
#45                     1
Name: restaurant_name, Length: 7437, dtype: int64
```

In [18]:
```
dfRS = dfRS[dfRS['aggregate_rating']>=4.0]
dfRS
```

Out[18]:

	restaurant_id	restaurant_name	cuisines	aggregate_rating	votes
9523	6000871	ukura a Sofras	Kebab, Izgara	4.4	296
3120	18222559	{Niche} - Cafe & Bar	North Indian, Chinese, Italian, Continental	4.1	492
9454	6401789	tashas	Cafe, Mediterranean	4.1	374
9385	6113857	sketch Gallery	British, Contemporary	4.5	148
1837	18418247	feel ALIVE	North Indian, American, Asian, Biryani	4.7	69
...	...	...	...	...	...
1468	18408054	19 Flavours Biryani	Mughlai, Hyderabadi	4.1	84
2484	18233317	145 Kala Ghoda	Fast Food, Beverages, Desserts	4.2	1606
2292	2100784	11th Avenue Cafe Bistro	Cafe, American, Italian, Continental	4.1	377
751	2600031	10 Downing Street	North Indian, Chinese	4.0	257
351	17057397	'Ohana	Hawaiian	4.5	1151

1236 rows × 5 columns

Loading [MathJax]/extensions/Safe.js

In [19]:
```
# Split Cuisines into list
dfRS['cuisines'] = dfRS['cuisines'].str.split(', ')
dfRS
```

Out[19]:

	restaurant_id	restaurant_name	cuisines	aggregate_rating	votes
9523	6000871	ukura a Sofras	[Kebab, Izgara]	4.4	296
3120	18222559	{Niche} - Cafe & Bar	[North Indian, Chinese, Italian, Continental]	4.1	492
9454	6401789	tashas	[Cafe, Mediterranean]	4.1	374
9385	6113857	sketch Gallery	[British, Contemporary]	4.5	148
1837	18418247	feel ALIVE	[North Indian, American, Asian, Biryani]	4.7	69
...	...	...	...	...	...
1468	18408054	19 Flavours Biryani	[Mughlai, Hyderabadi]	4.1	84
2484	18233317	145 Kala Ghoda	[Fast Food, Beverages, Desserts]	4.2	1606
2292	2100784	11th Avenue Cafe Bistro	[Cafe, American, Italian, Continental]	4.1	377
751	2600031	10 Downing Street	[North Indian, Chinese]	4.0	257
351	17057397	'Ohana	[Hawaiian]	4.5	1151

1236 rows × 5 columns

In [20]:
```
# Exploding 'cuisines'
dfRS = dfRS.explode('cuisines')
dfRS
```

Out[20]:

	restaurant_id	restaurant_name	cuisines	aggregate_rating	votes
9523	6000871	ukura a Sofras	Kebab	4.4	296
9523	6000871	ukura a Sofras	Izgara	4.4	296
3120	18222559	{Niche} - Cafe & Bar	North Indian	4.1	492
3120	18222559	{Niche} - Cafe & Bar	Chinese	4.1	492
3120	18222559	{Niche} - Cafe & Bar	Italian	4.1	492
...	...	...	...	...	...
2292	2100784	11th Avenue Cafe Bistro	Italian	4.1	377
2292	2100784	11th Avenue Cafe Bistro	Continental	4.1	377
751	2600031	10 Downing Street	North Indian	4.0	257
751	2600031	10 Downing Street	Chinese	4.0	257
351	17057397	'Ohana	Hawaiian	4.5	1151

2971 rows × 5 columns

In [21]:
```
dfRS['cuisines'].value_counts()
```

Out[21]:
```
North Indian     270
Italian          237
Chinese          200
Continental      199
Cafe             177
                ...
Pub Food           1
Durban             1
Irish              1
Persian            1
Sunda              1
Name: cuisines, Length: 128, dtype: int64
```

In [22]:
```
# Cross Tabulate Restaurant Name and Cuisines
xTabRestoCuisines = pd.crosstab(dfRS['restaurant_name'],
                                dfRS['cuisines'])
```

In [23]:
```
xTabRestoCuisines
```

Out[23]:

cuisines	Afghani	African	American	Andhra	Arabian	Argentine	Asian	Asian Fusion	Australian	Awadhi	...	Teriyaki	Tex Me
restaurant_name													
'Ohana	0	0	0	0	0	0	0	0	0	0	...	0	
10 Downing Street	0	0	0	0	0	0	0	0	0	0	...	0	
11th Avenue Cafe Bistro	0	0	1	0	0	0	0	0	0	0	...	0	
145 Kala Ghoda	0	0	0	0	0	0	0	0	0	0	...	0	
19 Flavours Biryani	0	0	0	0	0	0	0	0	0	0	...	0	
...	...	...	...	...	...	...	...	...	...	...	...	...	.
feel ALIVE	0	0	1	0	0	0	1	0	0	0	...	0	
sketch Gallery	0	0	0	0	0	0	0	0	0	0	...	0	
tashas	0	0	0	0	0	0	0	0	0	0	...	0	
{Niche} - Cafe & Bar	0	0	0	0	0	0	0	0	0	0	...	0	
ukura a Sofras	0	0	0	0	0	0	0	0	0	0	...	0	

1236 rows × 128 columns

In [24]:
```
# Checking on restaurant name value
xTabRestoCuisines.loc['feel ALIVE'].values
```

Out[24]:
```
array([0, 0, 1, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0,
       0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
       0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
       0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
       0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
       0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0], dtype=int64)
```

In [25]:
```
# Resto Names Sample
dfRS['restaurant_name'].sample(20, random_state=101)
```

Out[25]:
```
439                      Mrs. Wilkes' Dining Room
9545                                     Baltazar
6921                                  Rose Cafe
108                          Big City Bread Cafe
2311                               Olive Bistro
120                          Transmetropolitan
4387                         Maxims Pastry Shop
9222                                      Meraki
9359                           Mimi's Bakehouse
2438                           Cappuccino Blast
8049                              Oh So Stoned!
9544                        Karak_y G_ll_o  lu
579                                  Via Delhi
376                       Tu-Do Vietnamese Restaurant
4088    Tian - Asian Cuisine Studio - ITC Maurya
153                            Boise Fry Company
172                           Ting's Red Lantern
3107                                Odeon Social
9513                                  The Sizzle
839                             Sree Annapoorna
Name: restaurant_name, dtype: object
```

In [26]:
```
# Measure Similarity
print(jaccard_score(xTabRestoCuisines.loc["Olive Bistro"].values,
                    xTabRestoCuisines.loc["Rose Cafe"].values))
```

```
0.3333333333333333
```

Loading [MathJax]/extensions/Safe.js

In [27]:
```
# Create Similarity Value DF
jaccardDist = pdist(xTabRestoCuisines.values, metric='jaccard')
jaccardMatrix = squareform(jaccardDist)
jaccardSim = 1 - jaccardMatrix
dfJaccard = pd.DataFrame(
    jaccardSim,
    index=xTabRestoCuisines.index,
    columns=xTabRestoCuisines.index)

dfJaccard
```

Out[27]:

restaurant_name	'Ohana	10 Downing Street	11th Avenue Cafe Bistro	145 Kala Ghoda	19 Flavours Biryani	1918 Bistro & Grill	2 Dog	22nd Parallel	3 Wise Monkeys	38 Barracks	...	Zoeys Pizzeria	Zo
restaurant_name													
'Ohana	1.0	0.0	0.000000	0.0	0.0	0.0	0.000000	0.0	0.0	0.000000	...	0.0	
10 Downing Street	0.0	1.0	0.000000	0.0	0.0	0.0	0.000000	0.0	0.0	0.200000	...	0.0	
11th Avenue Cafe Bistro	0.0	0.0	1.000000	0.0	0.0	0.0	0.166667	0.0	0.0	0.333333	...	0.0	
145 Kala Ghoda	0.0	0.0	0.000000	1.0	0.0	0.0	0.000000	0.0	0.0	0.000000	...	0.0	
19 Flavours Biryani	0.0	0.0	0.000000	0.0	1.0	0.0	0.000000	0.0	0.0	0.000000	...	0.0	
...	...	...	...	...	...	...	...	...	...	...	...	...	...
feel ALIVE	0.0	0.2	0.142857	0.0	0.0	0.0	0.166667	0.0	0.0	0.600000	...	0.0	
sketch Gallery	0.0	0.0	0.000000	0.0	0.0	0.0	0.000000	0.0	0.0	0.000000	...	0.0	
tashas	0.0	0.0	0.200000	0.0	0.0	0.0	0.000000	0.0	0.0	0.000000	...	0.0	
{Niche} - Cafe & Bar	0.0	0.5	0.333333	0.0	0.0	0.0	0.000000	0.0	0.0	0.333333	...	0.0	
ukura a Sofras	0.0	0.0	0.000000	0.0	0.0	0.0	0.000000	0.0	0.0	0.000000	...	0.0	

1236 rows × 1236 columns

In [28]:
```
# Resto Names Sample
dfRS['restaurant_name'].sample(20)
```

Out[28]:
```
3701                    Pho King Awesome
9113                       Hops n Grains
4176                          Cool Point
2435    L 14 - Renaissance Lucknow Hotel
7481                  Zerruco - The Ashok
230         Red Ginger Sushi, Grill & Bar
752                           Chi Kitchen
2389                              Cafe 17
48                    Braseiro da G  vea
5896                           Bistro 57
416              Giovanni's Shrimp Truck
3098                              Caffe 9
52                         Fil  de Ouro
2404               Gabbar's Bar & Kitchen
3658                         Owl is Well
4319                        Cafe Southall
7481                  Zerruco - The Ashok
660                        Nini's Kitchen
6997                  Echoes Satyaniketan
10                       Silantro Fil-Mex
Name: restaurant_name, dtype: object
```

Final Recommendation System

In [29]:
```
# Input Initial Restaurant Name
resto = 'Ooma'

sim = dfJaccard.loc[resto].sort_values(ascending=False)

sim = pd.DataFrame({'restaurant_name': sim.index, 'simScore': sim.values})
sim = sim[(sim['restaurant_name']!= resto) & (sim['simScore']>=0.7)].head(5)

# Merge The Rating
RestoRec = pd.merge(sim,dfRS[['restaurant_name','aggregate_rating']],how='inner',on='restaurant_name')
FinalRestoRec = RestoRec.sort_values('aggregate_rating',ascending=False).drop_duplicates('restaurant_n
```

In [30]:
```
FinalRestoRec
```

Out[30]:

	restaurant_name	simScore	aggregate_rating
0	Sushi Masa	1.0	4.9
2	Nobu	1.0	4.4
4	Ichiban	1.0	4.3
8	Osaka	1.0	4.2
6	Guppy	1.0	4.1

Conclusion: The above Data will show up to top 5 recommended restaurants with the best rating, the rating is also curated only 4 and above, so the reccomendation system porvide good rating objectively.