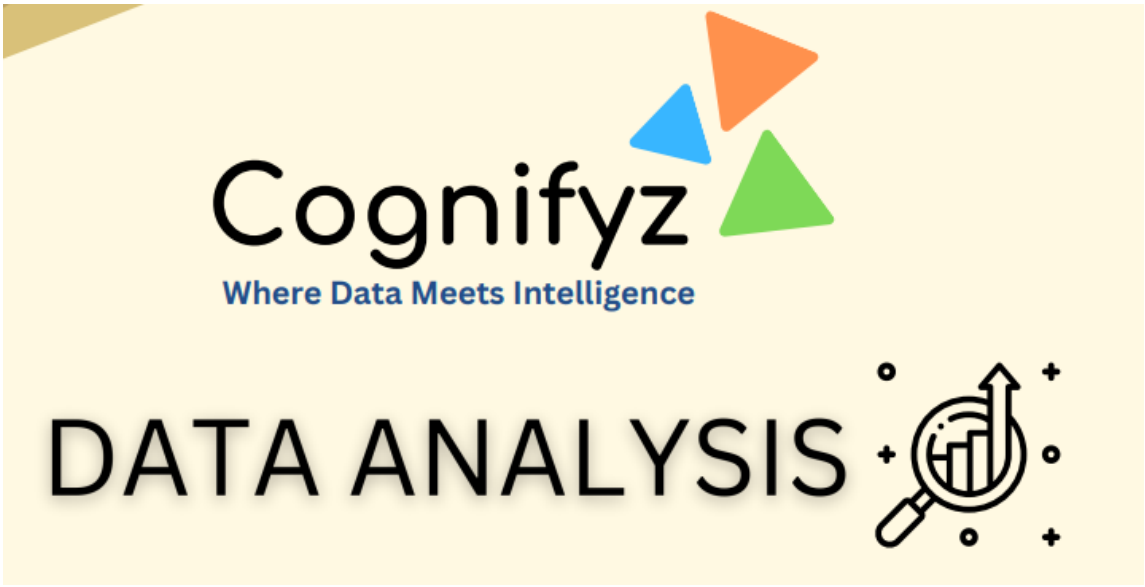


Level 3 - Objective

- Expertise in Python programming and Data Manipulation
- Extract valuable insights from large datasets and drive informed decision-making.
- Data cleaning and preprocessing data, performing statistical analysis, or creating data visualizations,
- Proficiency in Python will play a crucial role in delivering meaningful results.

1. Load Python Modules

In [14]:
```
# import python modules
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
```

In [15]:
```
# nltk modules
from nltk.sentiment import SentimentIntensityAnalyzer
from nltk.tokenize import word_tokenize
from nltk.corpus import stopwords
from collections import Counter
```

In [16]:
```
import nltk
nltk.download('vader_lexicon')
nltk.download('stopwords')
nltk.download('punkt')

```

```
[nltk_data] Downloading package vader_lexicon to
[nltk_data]     C:\Users\91956\AppData\Roaming\nltk_data...
[nltk_data]   Package vader_lexicon is already up-to-date!
[nltk_data] Downloading package stopwords to
[nltk_data]     C:\Users\91956\AppData\Roaming\nltk_data...
[nltk_data]   Package stopwords is already up-to-date!
[nltk_data] Downloading package punkt to
[nltk_data]     C:\Users\91956\AppData\Roaming\nltk_data...
[nltk_data]   Package punkt is already up-to-date!
```

Out[16]: True

2. Read the Dataset from CSV file - Using Pandas

In [17]:
```
# Read the csv file using pandas read_csv
restaurant_df=pd.read_csv("Dataset.csv")
restaurant_df
```

Out[17]:

	Restaurant ID	Restaurant Name	Country Code	City	Address	Locality	
0	6317637	Le Petit Souffle	162	Makati City	Third Floor, Century City Mall, Kalayaan Avenu...	Century City Mall, Poblacion, Makati City	C Mall,
1	6304287	Izakaya Kikufuji	162	Makati City	Little Tokyo, 2277 Chino Roces Avenue, Legaspi...	Little Tokyo, Legaspi Village, Makati City	L Lega Makat
2	6300002	Heat - Edsa Shangri-La	162	Mandaluyong City	Edsa Shangri-La, 1 Garden Way, Ortigas, Mandal...	Edsa Shangri-La, Ortigas, Mandaluyong City	Edsa S Ma
3	6318506	Ooma	162	Mandaluyong City	Third Floor, Mega Fashion Hall, SM Megamall, O...	SM Megamall, Ortigas, Mandaluyong City	SM Ma Cit
4	6314302	Sambo Kojin	162	Mandaluyong City	Third Floor, Mega Atrium, SM Megamall, Ortigas...	SM Megamall, Ortigas, Mandaluyong City	SM Ma Cit
...	...	...	...	...	...	...	
9546	5915730	Naml\ Gurme	208	��stanbul	Kemanke�� Karamustafa Pa��a Mahallesi, R\ht\m ...	Karak�_y	
9547	5908749	Ceviz A��ac\	208	��stanbul	Ko��uyolu Mahallesi, Muhittin ��st�_nda�� Cadd...	Ko��uyolu	K
9548	5915807	Huqqa	208	��stanbul	Kuru�_e��me Mahallesi, Muallim Naci Caddesi, N...	Kuru�_e��me	Kuru
9549	5916112	A���k Kahve	208	��stanbul	Kuru�_e��me Mahallesi, Muallim Naci Caddesi, N...	Kuru�_e��me	Kuru
9550	5927402	Walter's Coffee Roastery	208	��stanbul	Cafea��a Mahallesi, Bademalt\ Sokak, No 21/B, ...	Moda	

9551 rows × 21 columns

3. Basic Inspection on given dataset

In [18]:

```
def basic_inspection_dataset(table):

    print("top 5 rows - using head")
    print(table.head())
    print()

    print("bottom 5 rows using tail")
    print(table.tail())
    print()

    print("numbers of samples and columns")
    print(table.shape)
    print()

    print("numbers of samples ")
    print(len(table))
    print()

    print("numbers of entries in the data frame")
    print(table.size)
    print()

    print("Columns Names")
    print(table.columns)
    print()

    print("Columns dtypes")
    print(table.dtypes)
    print()

    print("Dataframe info")
    print(table.info())
    print()

    print()
    print("check the missing value in each column")
    print(table.isnull().sum())

    print()
    print("check the missing value in each column")
    print(table.isna().sum())

basic_inspection_dataset(restaurant_df)
```

```
top 5 rows - using head
   Restaurant ID         Restaurant Name  Country Code              City
\
0        6317637        Le Petit Souffle           162       Makati City
1        6304287        Izakaya Kikufuji           162       Makati City
2        6300002  Heat - Edsa Shangri-La           162  Mandaluyong City
3        6318506                    Ooma           162  Mandaluyong City
4        6314302             Sambo Kojin           162  Mandaluyong City

                                          Address  \
0  Third Floor, Century City Mall, Kalayaan Avenu...
1  Little Tokyo, 2277 Chino Roces Avenue, Legaspi...
2  Edsa Shangri-La, 1 Garden Way, Ortigas, Mandal...
3  Third Floor, Mega Fashion Hall, SM Megamall, O...
4  Third Floor, Mega Atrium, SM Megamall, Ortigas...

                                    Locality  \
0   Century City Mall, Poblacion, Makati City
1  Little Tokyo, Legaspi Village, Makati City
2  Edsa Shangri-La, Ortigas, Mandaluyong City
3      SM Megamall, Ortigas, Mandaluyong City
4      SM Megamall, Ortigas, Mandaluyong City

                                  Locality Verbose   Longitude   Latitud
e  \
0  Century City Mall, Poblacion, Makati City, Mak...  121.027535  14.56544
3
1  Little Tokyo, Legaspi Village, Makati City, Ma...  121.014101  14.55370
8
2  Edsa Shangri-La, Ortigas, Mandaluyong City, Ma...  121.056831  14.58140
4
3  SM Megamall, Ortigas, Mandaluyong City, Mandal...  121.056475  14.58531
8
4  SM Megamall, Ortigas, Mandaluyong City, Mandal...  121.057508  14.58445
0

                           Cuisines  ...          Currency Has Table booki
ng  \
0       French, Japanese, Desserts  ...  Botswana Pula(P)              Y
es
1                          Japanese  ...  Botswana Pula(P)              Y
es
2  Seafood, Asian, Filipino, Indian  ...  Botswana Pula(P)              Y
es
3                   Japanese, Sushi  ...  Botswana Pula(P)
No
4                  Japanese, Korean  ...  Botswana Pula(P)              Y
es

  Has Online delivery Is delivering now Switch to order menu Price range
\
0                  No                No                   No           3
1                  No                No                   No           3
2                  No                No                   No           4
3                  No                No                   No           4
4                  No                No                   No           4

   Aggregate rating  Rating color Rating text Votes
0               4.8    Dark Green   Excellent   314
1               4.5    Dark Green   Excellent   591
2               4.4         Green   Very Good   270
```

```
3              4.9    Dark Green   Excellent   365
4              4.8    Dark Green   Excellent   229

[5 rows x 21 columns]

bottom 5 rows using tail
      Restaurant ID           Restaurant Name  Country Code       City  \
9546        5915730              Naml\ Gurme           208  ��stanbul
9547        5908749             Ceviz A��ac\           208  ��stanbul
9548        5915807                   Huqqa           208  ��stanbul
9549        5916112              A���k Kahve           208  ��stanbul
9550        5927402  Walter's Coffee Roastery           208  ��stanbul

                                                 Address     Locality  \
9546  Kemanke�� Karamustafa Pa��a Mahallesi, R\ht\m ...     Karak�_y
9547  Ko��uyolu Mahallesi, Muhittin ��st�_nda�� Cadd...    Ko��uyol
u
9548  Kuru�_e��me Mahallesi, Muallim Naci Caddesi, N...  Kuru�_e��me
9549  Kuru�_e��me Mahallesi, Muallim Naci Caddesi, N...  Kuru�_e��me
9550  Cafea��a Mahallesi, Bademalt\ Sokak, No 21/B, ...         Moda

           Locality Verbose  Longitude   Latitude  \
9546     Karak�_y, ��stanbul  28.977392  41.022793
9547    Ko��uyolu, ��stanbul  29.041297  41.009847
9548  Kuru�_e��me, ��stanbul  29.034640  41.055817
9549  Kuru�_e��me, ��stanbul  29.036019  41.057979
9550         Moda, ��stanbul  29.026016  40.984776

                           Cuisines  ...          Currency  \
9546                         Turkish  ...  Turkish Lira(TL)
9547  World Cuisine, Patisserie, Cafe  ...  Turkish Lira(TL)
9548          Italian, World Cuisine  ...  Turkish Lira(TL)
9549                 Restaurant Cafe  ...  Turkish Lira(TL)
9550                            Cafe  ...  Turkish Lira(TL)

     Has Table booking Has Online delivery Is delivering now  \
9546               No                  No                No
9547               No                  No                No
9548               No                  No                No
9549               No                  No                No
9550               No                  No                No

     Switch to order menu Price range  Aggregate rating  Rating color  \
9546                   No           3               4.1         Green
9547                   No           3               4.2         Green
9548                   No           4               3.7        Yellow
9549                   No           4               4.0         Green
9550                   No           2               4.0         Green

     Rating text Votes
9546   Very Good   788
9547   Very Good  1034
9548        Good   661
9549   Very Good   901
9550   Very Good   591

[5 rows x 21 columns]

numbers of samples and columns
(9551, 21)
```

```
numbers of samples
9551

numbers of entries in the data frame
200571

Columns Names
Index(['Restaurant ID', 'Restaurant Name', 'Country Code', 'City', 'Addres
s',
       'Locality', 'Locality Verbose', 'Longitude', 'Latitude', 'Cuisine
s',
       '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')

Columns dtypes
Restaurant ID             int64
Restaurant Name          object
Country Code              int64
City                     object
Address                  object
Locality                 object
Locality Verbose         object
Longitude               float64
Latitude                float64
Cuisines                 object
Average Cost for two      int64
Currency                 object
Has Table booking        object
Has Online delivery      object
Is delivering now        object
Switch to order menu     object
Price range               int64
Aggregate rating        float64
Rating color             object
Rating text              object
Votes                     int64
dtype: object

Dataframe info
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 9551 entries, 0 to 9550
Data columns (total 21 columns):
 #   Column                Non-Null Count  Dtype
---  ------                --------------  -----
 0   Restaurant ID         9551 non-null   int64
 1   Restaurant Name       9551 non-null   object
 2   Country Code          9551 non-null   int64
 3   City                  9551 non-null   object
 4   Address               9551 non-null   object
 5   Locality              9551 non-null   object
 6   Locality Verbose      9551 non-null   object
 7   Longitude             9551 non-null   float64
 8   Latitude              9551 non-null   float64
 9   Cuisines              9542 non-null   object
 10  Average Cost for two  9551 non-null   int64
 11  Currency              9551 non-null   object
 12  Has Table booking     9551 non-null   object
 13  Has Online delivery   9551 non-null   object
```

```
 14  Is delivering now     9551 non-null   object
 15  Switch to order menu  9551 non-null   object
 16  Price range           9551 non-null   int64
 17  Aggregate rating      9551 non-null   float64
 18  Rating color          9551 non-null   object
 19  Rating text           9551 non-null   object
 20  Votes                 9551 non-null   int64
dtypes: float64(3), int64(5), object(13)
memory usage: 1.5+ MB
None


check the missing value in each column
Restaurant ID           0
Restaurant Name         0
Country Code            0
City                    0
Address                 0
Locality                0
Locality Verbose        0
Longitude               0
Latitude                0
Cuisines                9
Average Cost for two    0
Currency                0
Has Table booking       0
Has Online delivery     0
Is delivering now       0
Switch to order menu    0
Price range             0
Aggregate rating        0
Rating color            0
Rating text             0
Votes                   0
dtype: int64

check the missing value in each column
Restaurant ID           0
Restaurant Name         0
Country Code            0
City                    0
Address                 0
Locality                0
Locality Verbose        0
Longitude               0
Latitude                0
Cuisines                9
Average Cost for two    0
Currency                0
Has Table booking       0
Has Online delivery     0
Is delivering now       0
Switch to order menu    0
Price range             0
Aggregate rating        0
Rating color            0
Rating text             0
Votes                   0
dtype: int64
```

4. Handling Missing Values

In [19]:
```
#For a categorical variable, determine the most frequent value, known as the mode.
cuisine_mode = restaurant_df['Cuisines'].mode()[0]
print(cuisine_mode)

# fill the missing value with mode
restaurant_df['Cuisines'].fillna(cuisine_mode,inplace=True)

# check for missing values - for confirmation
restaurant_df.isnull().sum()
```

```
North Indian
```

Out[19]:
```
Restaurant ID           0
Restaurant Name         0
Country Code            0
City                    0
Address                 0
Locality                0
Locality Verbose        0
Longitude               0
Latitude                0
Cuisines                0
Average Cost for two    0
Currency                0
Has Table booking       0
Has Online delivery     0
Is delivering now       0
Switch to order menu    0
Price range             0
Aggregate rating        0
Rating color            0
Rating text             0
Votes                   0
dtype: int64
```

Level 3, Task 1:Task: Restaurant Reviews

3.1.1 Analyze the text reviews to identify the most common positive and negative keywords.

In [20]:
```
rating_texts=restaurant_df['Rating
text'].value_counts().reset_index()
rating_texts.columns = ["Rating-Type","Count"]
rating_texts
```

Out[20]:

	Rating-Type	Count
0	Average	3737
1	Not rated	2148
2	Good	2100
3	Very Good	1079
4	Excellent	301
5	Poor	186

In [21]:
```
sia=SentimentIntensityAnalyzer()
stop_words=set(stopwords.words('english'))
positive_review=[]
negative_review=[]
```

In [22]:
```
rating_texts=restaurant_df['Rating text']
```

In [23]:
```
for rating_text in rating_texts:
    tokens= word_tokenize(rating_text.lower())
    tokens=[token for token in tokens if token.isalpha() and token
not in stop_words]

    sentiment_score=sia.polarity_scores(rating_text)['compound']

    if sentiment_score>=0.05:
        positive_review.extend(tokens)
    elif sentiment_score<0.05:
        negative_review.extend(tokens)
```

In [24]:
```
positive_counts=Counter(positive_review)
negative_counts=Counter(negative_review)

num_top_keywords = 10
print('Top positive Review Keywords:')
for keyword, count in positive_counts.most_common(num_top_keywords):
    print(f"{keyword}:{count} times")

print()
print('Top Negative Review Keywords:')
for keyword, count in negative_counts.most_common(num_top_keywords):
    print(f"{keyword}:{count} times")
```

```
Top positive Review Keywords:
good:3179 times
excellent:301 times

Top Negative Review Keywords:
average:3737 times
rated:2148 times
poor:186 times
```

Observations

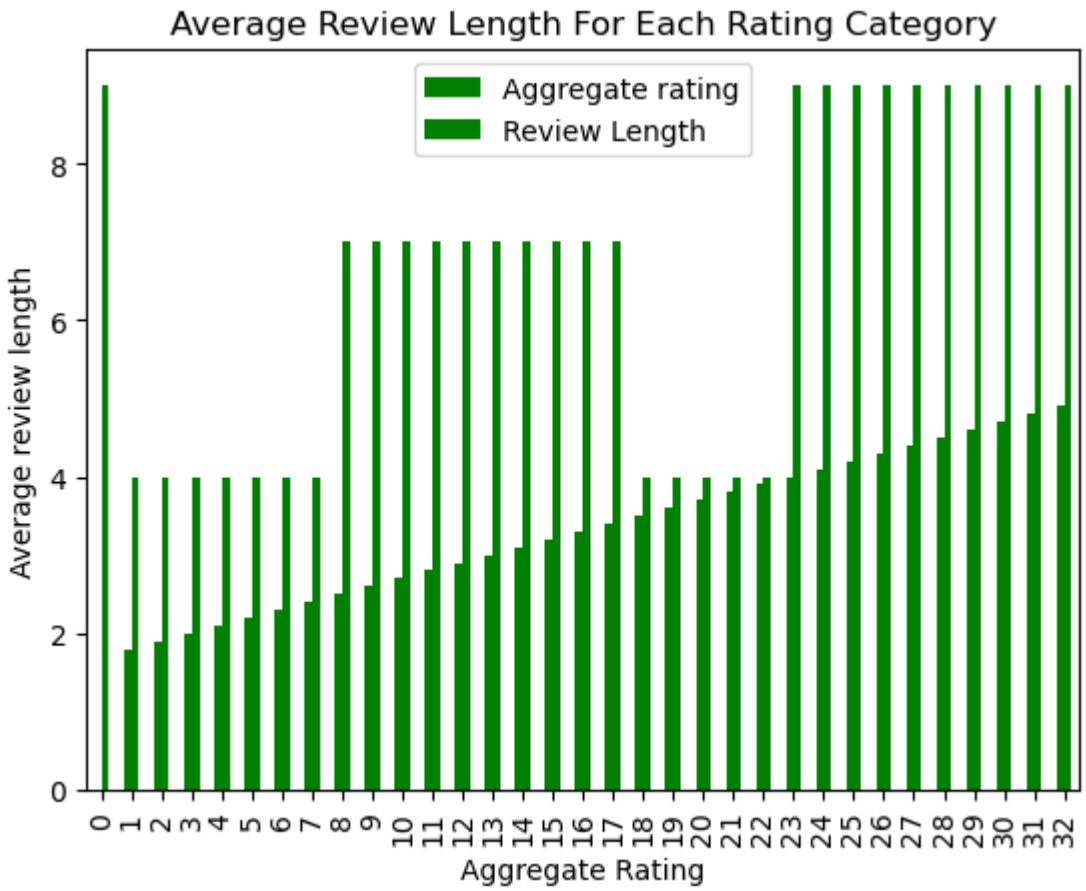
- Positive Keywords - good and excellent
- Negative Keywords - average, rated , poor

3.1.2 Calculate the average length of reviews and explore if there is a relationship between review length and rating.

In [25]:
```
restaurant_df['Review Length']=restaurant_df['Rating
text'].apply(lambda x: len(str(x)))
avg_rev_len=restaurant_df.groupby('Aggregate rating')['Review
Length'].mean()
avg_rev_df = pd.DataFrame(avg_rev_len).reset_index()
```

In [26]:
```
plt.figure(figsize=(10,10))
avg_rev_df.plot(kind='bar',color='green')
#plt.bar(x=avg_rev_df["Aggregate rating"],height=avg_rev_df['Review Length'])
plt.title('Average Review Length For Each Rating Category')
plt.xlabel('Aggregate Rating')
plt.ylabel('Average review length')
plt.show()
```

```
<Figure size 1000x1000 with 0 Axes>
```



Observations

- Relation between Agg Rating vs Avg Review Text length
    1. Agg Rating 1.8 to 2.4 - Avg Review text length - 4
    2. Avg Rating 2.5 to 3.4 - Avg Review text length - 7
    3. Avg Rating 3.5 to 3.9 - Avg Review text length - 4
    4. Avg Rating 4.0 to 4.9 - Avg Review text length - 9

Level 3 , Task 2 : Votes Analysis

3.2.1 Identify the restaurants with the highest and lowest number of votes.

In [27]:
```
cols = ['Votes','Restaurant Name']
df_votes_restaurants=restaurant_df[cols]
print()
print('Restaurant with highest Votes:')
print(df_votes_restaurants.sort_values(by="Votes").tail(1))

print()
print('Restaurant with lowest Votes:')
print(df_votes_restaurants.sort_values(by="Votes").head(90))
```

```
Restaurant with highest Votes:
     Votes Restaurant Name
728  10934            Toit

Restaurant with lowest Votes:
      Votes             Restaurant Name
5799      0         Khalsa Eating Point
7411      0   Radha Swami Chaat Bhandar
7414      0  Ram Ram Ji Kachori Bhandar
7415      0          Rana's Food Corner
7416      0         Sanjay Chicken Shop
...     ...                         ...
1185      0                Solty Hotel
1183      0                OMG Tiffinz
1181      0      Narayan Fast Food Home
1178      0        Gopi Sweets & Caters
3621      0             Baweja's Haandi

[90 rows x 2 columns]
```

Observations

- Restaurant with highest Votes
    1. Toit with 10934 Votes
- Restaurant with lowest Votes
    1. Many Restaurants have 0 Votes

3.2.2 Analyze if there is a correlation between the number of votes and the rating of a restaurant.

In [28]:
```
cols = ['Votes','Aggregate rating']
df_corr_analysis = restaurant_df[cols]
df_corr_analysis
```

Out[28]:

	Votes	Aggregate rating
0	314	4.8
1	591	4.5
2	270	4.4
3	365	4.9
4	229	4.8
...	...	...
9546	788	4.1
9547	1034	4.2
9548	661	3.7
9549	901	4.0
9550	591	4.0

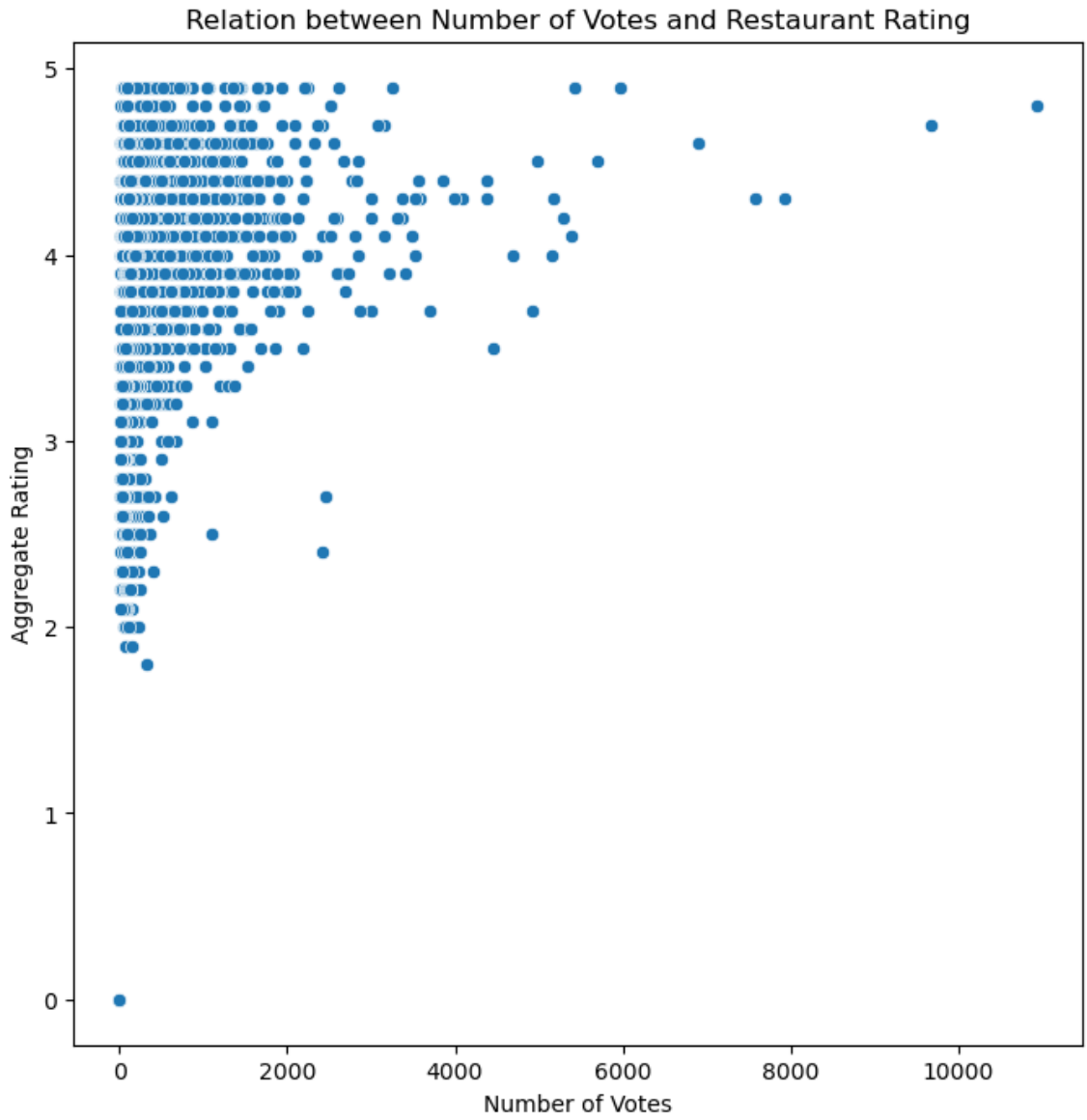
9551 rows × 2 columns

In [29]:
```
corr=df_corr_analysis.corr()
corr
```

Out[29]:

	Votes	Aggregate rating
Votes	1.000000	0.313691
Aggregate rating	0.313691	1.000000

In [30]:
```
plt.figure(figsize=(8,8))
sns.scatterplot(x='Votes',y='Aggregate rating',data=df_corr_analysis)
plt.title('Relation between Number of Votes and Restaurant Rating')
plt.xlabel("Number of Votes")
plt.ylabel('Aggregate Rating')
plt.show()
```



Observations

- Correlation between the number of votes and the rating of a restaurant is 0.31

Level 3 , Task 3 : Price Range vs. Online Delivery and Table Booking

3.3.1 Analyze if there is a relationship between the price range and the availability of online delivery and table booking

In [31]:
```
1 restaurant_df.head()
```

Out[31]:

	Restaurant ID	Restaurant Name	Country Code	City	Address	Locality	Locality Verbose	Lon
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.(
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.(
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.(
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.(
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.(

5 rows × 22 columns

In [32]:
```
cols = ['Price range','Has Online delivery','Has Table booking']
df_analysis=restaurant_df[cols].copy()
df_analysis['Has Online delivery']=df_analysis['Has Online delivery'].map({'Yes':True,'No':False})
df_analysis['Has Table booking']=df_analysis['Has Table booking'].map({'Yes':True,'No':False})
df_analysis
```

Out[32]:

	Price range	Has Online delivery	Has Table booking
0	3	False	True
1	3	False	True
2	4	False	True
3	4	False	False
4	4	False	True
...	...	...	...
9546	3	False	False
9547	3	False	False
9548	4	False	False
9549	4	False	False
9550	2	False	False

9551 rows × 3 columns

In [33]:
```
summary_table=pd.pivot_table(df_analysis,index='Price range',values=['Has Online delivery','Has Table booking'],aggfunc=sum)
print('Summary Table:')
summary_table
```

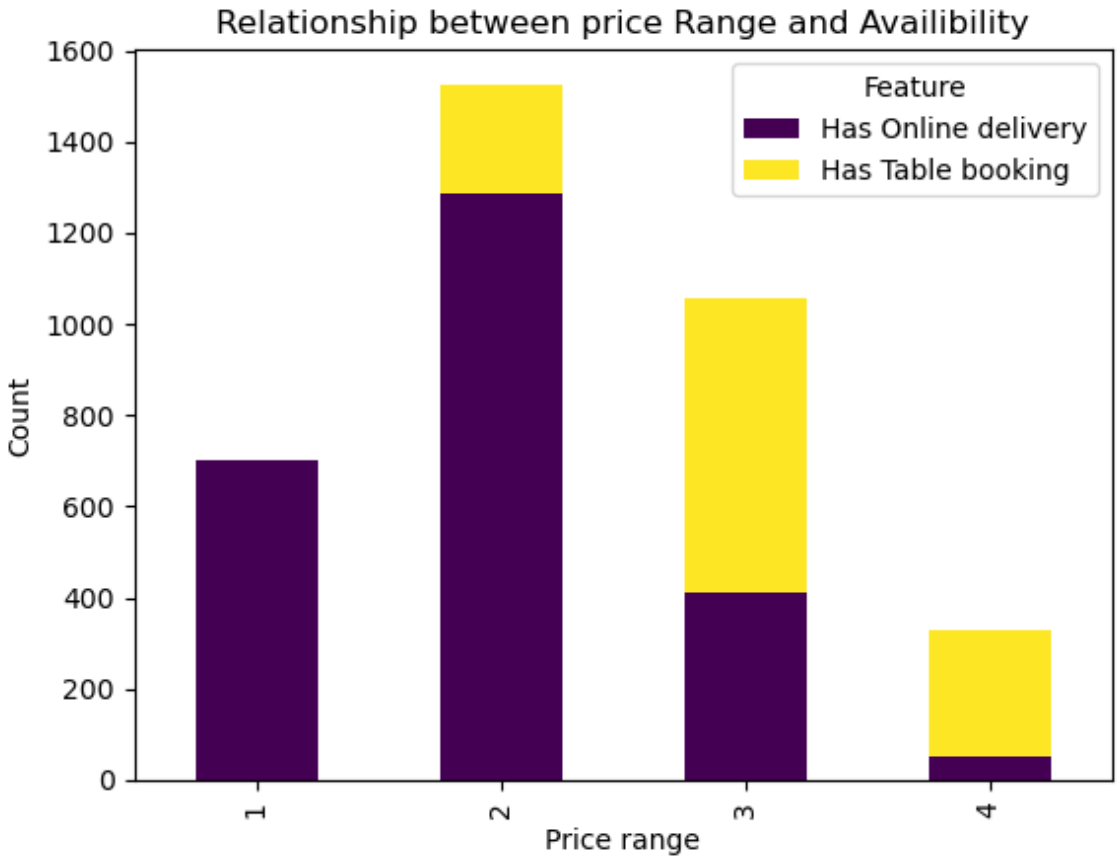
```
Summary Table:
```

Out[33]:

Price range	Has Online delivery	Has Table booking
1	701	1
2	1286	239
3	411	644
4	53	274

In [34]:
```
plt.figure(figsize=(10,8))
summary_table.plot(kind='bar',stacked=True,colormap='viridis')
plt.title('Relationship between price Range and Availibility')
plt.xlabel('Price range')
plt.ylabel('Count')
plt.legend(title='Feature',loc='upper right')
plt.show()
```

```
<Figure size 1000x800 with 0 Axes>
```



3.3.2 Determine if higher-priced restaurants are more likely to offer these services

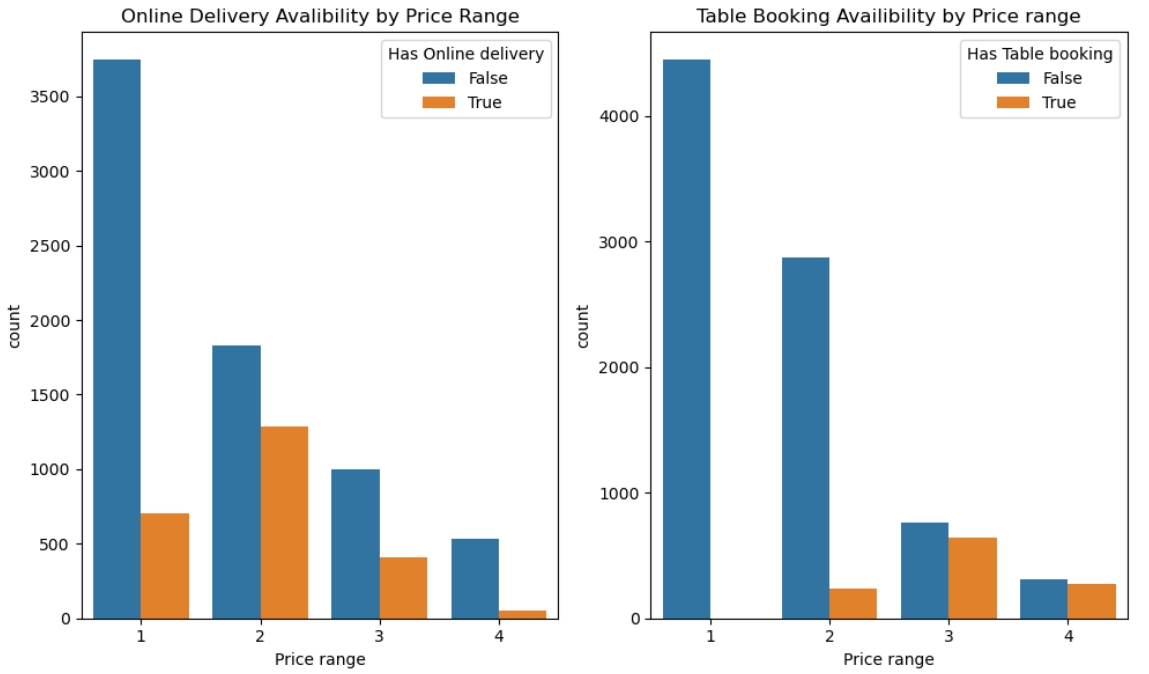
In [35]:
```
plt.figure(figsize=(10,6))

plt.subplot(1,2,1)

sns.countplot(x='Price range' , hue='Has Online delivery' , data=df_analysis)
plt.title('Online Delivery Avalibility by Price Range')

plt.subplot(1,2,2)
sns.countplot(x='Price range', hue='Has Table booking', data=df_analysis)
plt.title('Table Booking Availibility by Price range')

plt.tight_layout()
plt.show()
```



Observations

- The statement "higher-priced restaurants are more likely to offer these services" is not valid