Level : 3

Task -1 : Restaurant Reviews

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

In [27]:
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
```

In [28]:
```
from nltk.sentiment import SentimentIntensityAnalyzer
from nltk.tokenize import word_tokenize
from nltk.corpus import stopwords
from collections import Counter
```

In [29]:
```
from nltk.sentiment import SentimentIntensityAnalyzer
from nltk.tokenize import word_tokenize
from nltk.corpus import stopwords
from collections import Counter
```

In [30]:
```
data_set = pd.read_csv(r"C:\Users\Shree\OneDrive\Desktop\FSDS_omkar sir\Datafiles\r
data_set
```

Out[30]:

	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	Ma
1	6304287	Izakaya Kikufuji	162	Makati City	Little Tokyo, 2277 Chino Roces Avenue, Legaspi...	Little Tokyo, Legaspi Village, Makati City	Le
2	6300002	Heat - Edsa Shangri-La	162	Mandaluyong City	Edsa Shangri-La, 1 Garden Way, Ortigas, Mandal...	Edsa Shangri-La, Ortigas, Mandaluyong City	Eds
3	6318506	Ooma	162	Mandaluyong City	Third Floor, Mega Fashion Hall, SM Megamall, O...	SM Megamall, Ortigas, Mandaluyong City	S
4	6314302	Sambo Kojin	162	Mandaluyong City	Third Floor, Mega Atrium, SM Megamall, Ortigas...	SM Megamall, Ortigas, Mandaluyong City	S
...	...	...	...	...	...	...	
9546	5915730	Naml\ Gurme	208	��stanbul	Kemanke�� Karamustafa Pa��a Mahallesi, Rlhtlm ...	Karak�_y	
9547	5908749	Ceviz A��acl	208	��stanbul	Ko��uyolu Mahallesi, Muhittin ��st�_nda�� Cadd...	Ko��uyolu	
9548	5915807	Huqqa	208	��stanbul	Kuru�_e��me Mahallesi, Muallim Naci Caddesi, N...	Kuru�_e��me	Kur
9549	5916112	A���k Kahve	208	��stanbul	Kuru�_e��me Mahallesi, Muallim Naci Caddesi, N...	Kuru�_e��me	Kur
9550	5927402	Walter's Coffee Roastery	208	��stanbul	Cafea��a Mahallesi, Bademaltl	Moda	

	Restaurant ID	Restaurant Name	Country Code	City	Address	Locality
					Sokak, No 21/B, ...	

9551 rows × 21 columns

In [39]:
```
cuisine_mode = data_set['Cuisines'].mode()[0]
print(cuisine_mode)

data_set['Cuisines'].fillna(cuisine_mode,inplace=True)

data_set.isnull().sum()
```

```
North Indian
```

Out[39]:
```
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
```

In [65]:
```
rating_texts=data_set['Rating text'].value_counts().reset_index()
rating_texts.columns = ["Rating-Type","Count"]
rating_texts
```

Out[65]:

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

In [66]:
```
sia=SentimentIntensityAnalyzer()
stop_words=set(stopwords.words('english'))
positive_review=[]
negative_review=[]
```

In [67]:
```
rating_texts=data_set['Rating text']
rating_texts
```

Out[67]:
```
0       Excellent
1       Excellent
2       Very Good
3       Excellent
4       Excellent
          ...
9546    Very Good
9547    Very Good
9548         Good
9549    Very Good
9550    Very Good
Name: Rating text, Length: 9551, dtype: object
```

In [68]:
```
# import nltk
# nltk.download('punkt')
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_word

    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 [69]:
```
# import nltk
# nltk.download('stopwords')
# sia=SentimentIntensityAnalyzer()
# stop_words=set(stopwords.words('english'))
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
```

Observation :

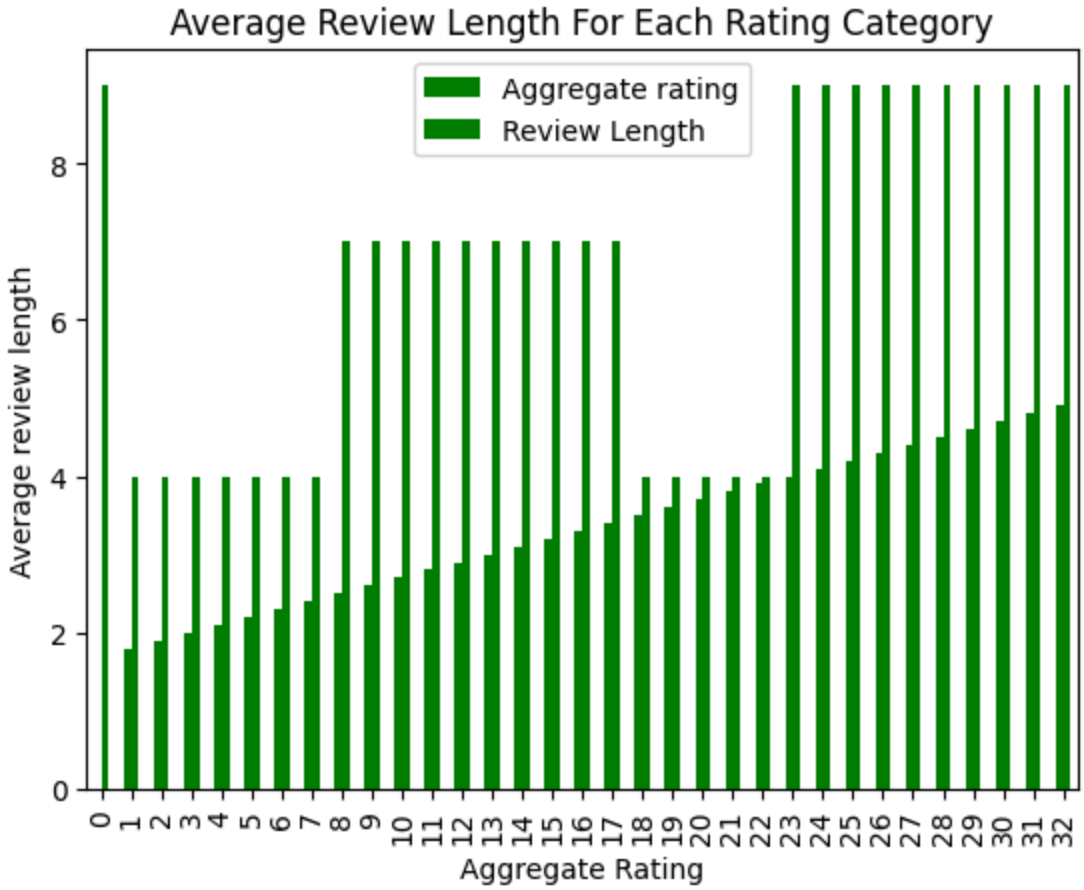
- positive Keywords - Good and Excellent
- Negative Keywords - Average , raated , poor

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

In [70]:
```
data_set['Review Length']=data_set['Rating text'].apply(lambda x: len(str(x)))
avg_rev_len=data_set.groupby('Aggregate rating')['Review Length'].mean()
avg_rev_df = pd.DataFrame(avg_rev_len).reset_index()
```

In [71]:
```
plt.figure(figsize=(10,10))
avg_rev_df.plot(kind='bar',color='green')
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

Task - 2 : Votes Analysis

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

In [72]:
```
cols = ['Votes','Restaurant Name']
df_votes_restaurants=data_set[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
7797      0         Anjali Resaturant
7798      0    Annapoorna Bhojanalya
7799      0      Anupam Eating Point
7800      0             Bikaner Rasoi
7804      0             Cake Knighter
...     ...                       ...
5806      0              Paras Corner
5807      0             Punjabi Tadka
5808      0  Sardar A Pure Meat Shop
5791      0    Brahm Point Fast Food
5793      0                 Curry Man

[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

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

In [73]:
```
cols = ['Votes','Aggregate rating']
df_corr_analysis = data_set[cols]
df_corr_analysis
```

Out[73]:

	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

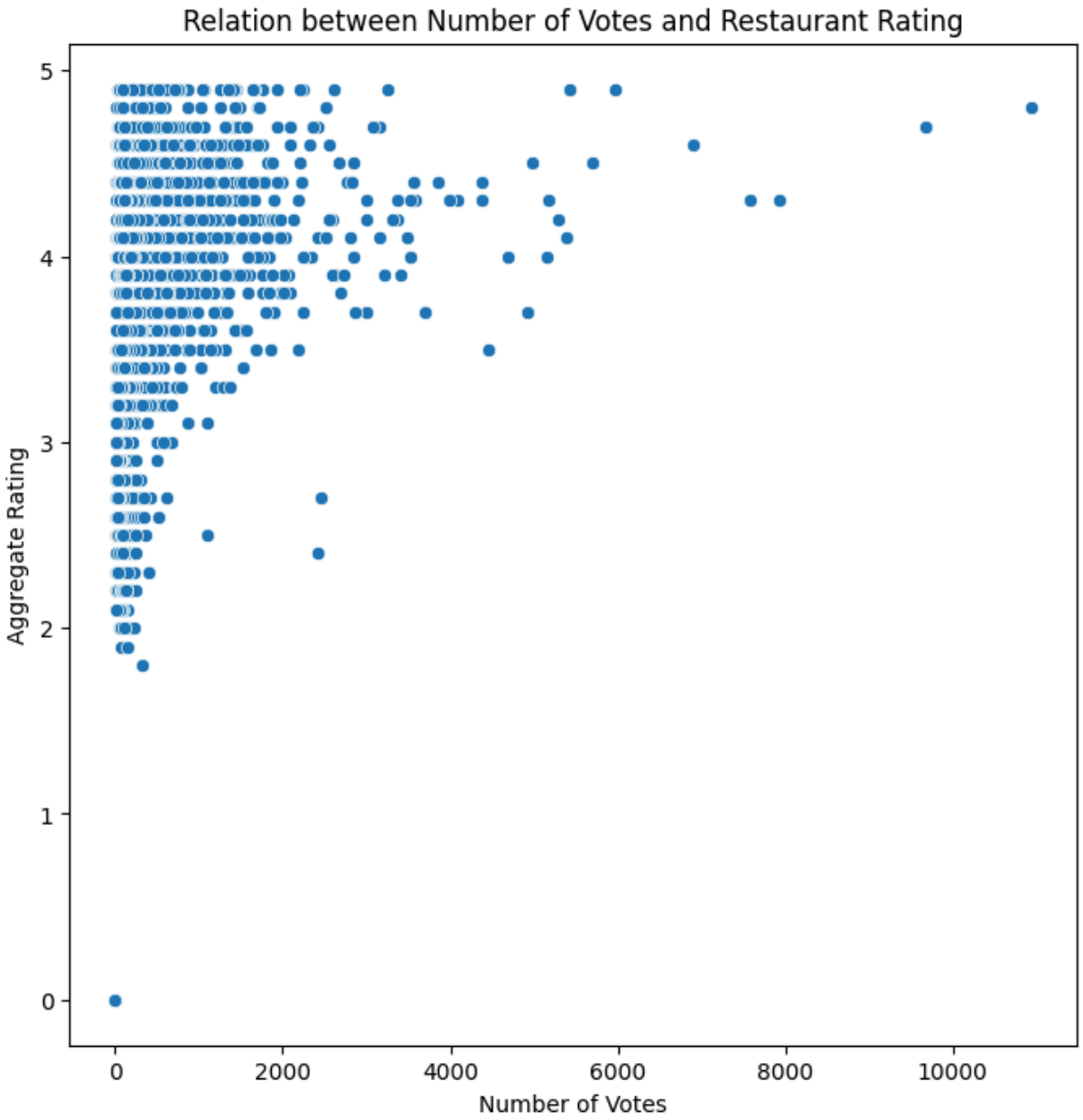
9551 rows × 2 columns

In [74]:
```
corr=df_corr_analysis.corr()
corr
```

Out[74]:

	Votes	Aggregate rating
Votes	1.000000	0.313691
Aggregate rating	0.313691	1.000000

In [75]:
```
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

Task : 3 = Price Range vs. Online Delivery and Table Bookin

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

In [76]:
```
data_set.head()
```

Out[76]:

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

5 rows × 22 columns

In [77]:
```
cols = ['Price range','Has Online delivery','Has Table booking']
df_analysis=data_set[cols].copy()
df_analysis['Has Online delivery']=df_analysis['Has Online delivery'].map({'Yes':Tr
df_analysis['Has Table booking']=df_analysis['Has Table booking'].map({'Yes':True,'
df_analysis
```

Out[77]:

	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 [78]:
```
summary_table=pd.pivot_table(df_analysis,index='Price range',values=['Has Online de
print('Summary Table:')
summary_table
```

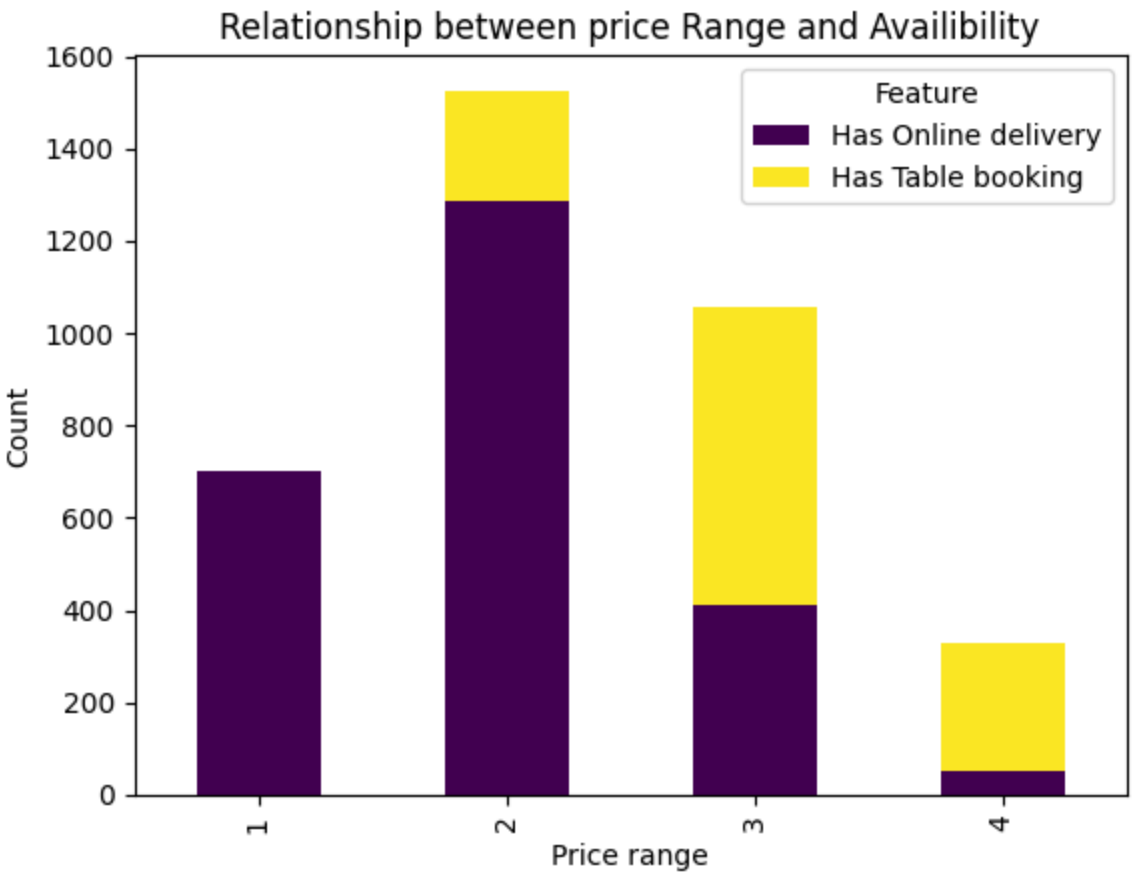
```
Summary Table:
C:\Users\Shree\AppData\Local\Temp\ipykernel_22976\948226973.py:1: FutureWarning: The
provided callable <built-in function sum> is currently using DataFrameGroupBy.sum. I
n a future version of pandas, the provided callable will be used directly. To keep c
urrent behavior pass the string "sum" instead.
  summary_table=pd.pivot_table(df_analysis,index='Price range',values=['Has Online d
elivery','Has Table booking'],aggfunc=sum)
```

Out[78]:

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

In [79]:
```
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>
```



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

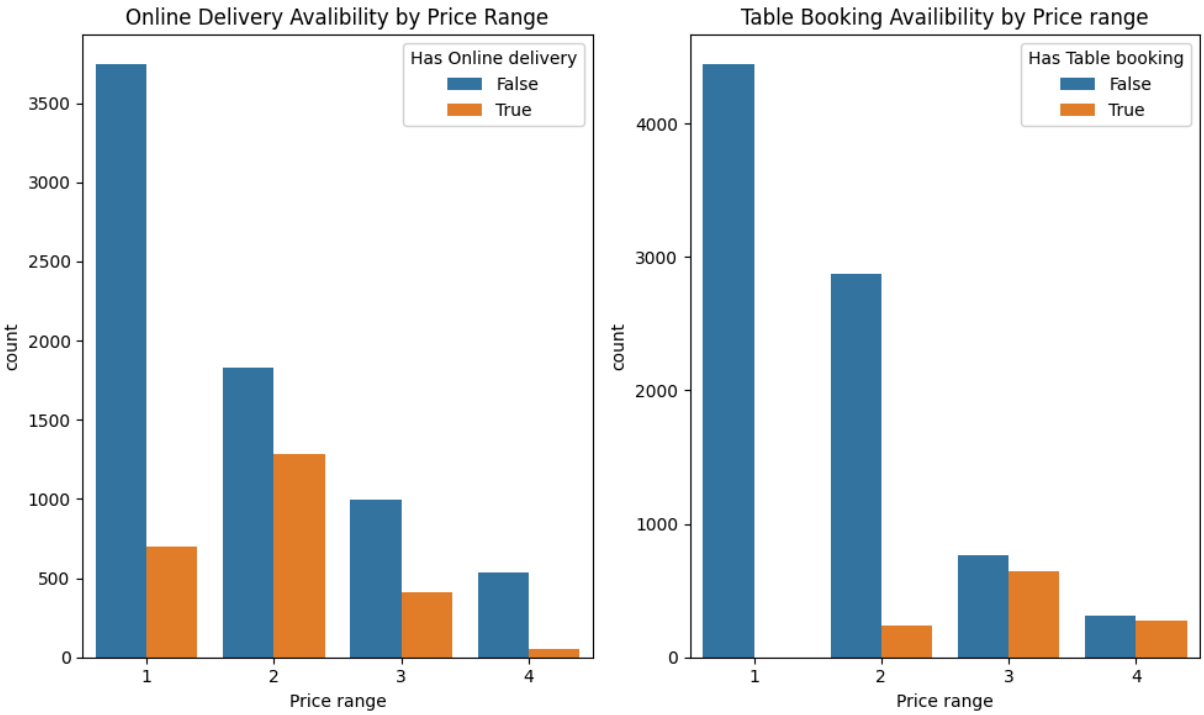
In [80]:
```
plt.figure(figsize=(10,6))

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