Task: Restaurant Reviews

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

In [1]:

```
import pandas as pd
from sklearn.feature_extraction.text import CountVectorizer
```

In [6]:

```
dt = pd.read_csv("C:\\Users\\MAHESH\Desktop\cognifyz\Dataset.csv")
dt
```

Out[6]:

	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
...	...	...	...	...	...	...	...	...	...	...
9546	5915730	Naml\ Gurme	208	??stanbul	Kemanke?? Karamustafa Pa??a Mahallesi, R\ht\m ...	Karak?_y	Karak?_y, ??stanbul	28.977392	41.022793	Turkish
9547	5908749	Ceviz A??ac\	208	??stanbul	Ko??uyolu Mahallesi, Muhittin ??st?_nda?? Cadd...	Ko??uyolu	Ko??uyolu, ??stanbul	29.041297	41.009847	World Cuisine Patisserie Cafe
9548	5915807	Huqqa	208	??stanbul	Kuru?_e??me Mahallesi, Muallim Naci Caddesi, N...	Kuru?_e??me	Kuru?_e??me, ??stanbul	29.034640	41.055817	Italian World Cuisine
9549	5916112	A???k Kahve	208	??stanbul	Kuru?_e??me Mahallesi, Muallim Naci Caddesi, N...	Kuru?_e??me	Kuru?_e??me, ??stanbul	29.036019	41.057979	Restaurant Cafe
9550	5927402	Walter's Coffee Roastery	208	??stanbul	Cafea??a Mahallesi, Bademalt\ Sokak, No 21/B, ...	Moda	Moda, ??stanbul	29.026016	40.984776	Cafe

9551 rows × 21 columns

In [7]:

```
review_text = dt['Rating text']
review_text
```

Out[7]:

```
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 [8]:

```
vectorizer = CountVectorizer()
```

In [10]:

```
vectorizer
```

Out[10]:

▾ CountVectorizer

CountVectorizer()

In [11]:

```
word_counts = vectorizer.fit_transform(review_text)
```

In [16]:

```
word_counts
```

Out[16]:

```
<9551x7 sparse matrix of type '<class 'numpy.int64'>'
 with 12778 stored elements in Compressed Sparse Row format>
```

In [32]:

```
words = vectorizer.get_feature_names_out()
```

In [34]:

```
words
```

Out[34]:

```
array(['average', 'excellent', 'good', 'not', 'poor', 'rated', 'very'],
      dtype=object)
```

In [19]:

```
word_counts = word_counts.sum(axis=0).tolist()[0]
```

In [20]:

```
word_counts
```

Out[20]:

```
[3737, 301, 3179, 2148, 186, 2148, 1079]
```

In [24]:

```
import numpy as np
```

In [26]:

```
array = np.array(word_counts)
```

In [27]:

```
array
```

Out[27]:

```
array([3737,  301, 3179, 2148,  186, 2148, 1079])
```

In [33]:

```
pos_words = [words[i] for i in array.argsort()[-5:][::-1]]
```

In [35]:

```
neg_words = [words[i] for i in array.argsort()[:5]]
```

In [36]:

```
print("Positive words:", pos_words)
print("Negative words:", neg_words)
```

```
Positive words: ['average', 'good', 'rated', 'not', 'very']
Negative words: ['poor', 'excellent', 'very', 'not', 'rated']
```

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

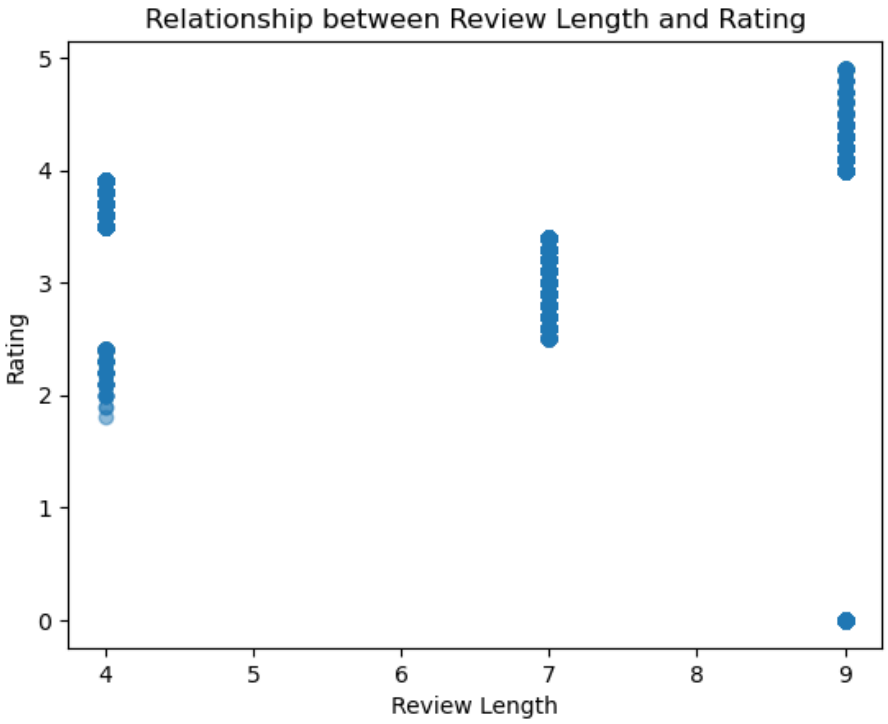
In [37]:

```
average_review_length = dt['Rating text'].str.len().mean()
print(average_review_length)
```

```
7.020730813527379
```

In [38]:

```
import matplotlib.pyplot as plt
plt.scatter(dt['Rating text'].str.len(), dt['Aggregate rating'], alpha=0.5)
plt.xlabel('Review Length')
plt.ylabel('Rating')
plt.title('Relationship between Review Length and Rating')
plt.show()
```



In [39]:

```
correlation_coefficient = dt['Rating text'].str.len().corr(dt['Aggregate rating'])

print(correlation_coefficient)
```

```
-0.47888483813492855
```

In [ ]: