Task: Restaurant Reviews

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

In [45]:
```
import pandas as pd
from sklearn.feature_extraction.text import CountVectorizer
```

In [46]:
```
dt = pd.read_csv(r"C:\Users\HP\OneDrive\Documents\Cognifyz Internship Program\Dataset.cs
dt
```

Out[46]:

	Restaurant ID	Restaurant Name	Country Code	City	Address	Locality	Locality Verbose	Longi
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...	121.02
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...	121.01
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...	121.05
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...	121.05
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...	121.05
...	...	...	...	...	...	...	...	
9546	5915730	Naml\ Gurme	208	��stanbul	Kemanke�� Karamustafa Pa��a Mahallesi, R\ht\m ...	Karak�_y	Karak�_y, ��stanbul	28.97
9547	5908749	Ceviz A��ac\	208	��stanbul	Ko��uyolu Mahallesi, Muhittin ��st�_nda�� Cadd...	Ko��uyolu	Ko��uyolu, ��stanbul	29.04
9548	5915807	Huqqa	208	��stanbul	Kuru�_e��me Mahallesi, Muallim Naci Caddesi, N...	Kuru�_e��me	Kuru�_e��me, ��stanbul	29.03
9549	5916112	A���k Kahve	208	��stanbul	Kuru�_e��me Mahallesi, Muallim Naci Caddesi, N...	Kuru�_e��me	Kuru�_e��me, ��stanbul	29.03
9550	5927402	Walter's Coffee Roastery	208	��stanbul	Cafea��a Mahallesi, Bademalt\ Sokak, No 21/B, ...	Moda	Moda, ��stanbul	29.02

9551 rows × 21 columns

In [47]:
```
review_text = dt['Rating text']
review_text
```

Out[47]:
```
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 [48]:
```
# Create vectorizer
vectorizer = CountVectorizer()
```

In [49]:
```
vectorizer
```

Out[49]:
```
CountVectorizer()
```

In [50]:
```
# Generate word counts
word_counts = vectorizer.fit_transform(review_text)
```

In [51]:
```
word_counts
```

Out[51]:
```
<9551x7 sparse matrix of type '<class 'numpy.int64'>'
        with 12778 stored elements in Compressed Sparse Row format>
```

In [52]:
```
# Get vocabulary of words
words = vectorizer.get_feature_names()
```

```
C:\Users\HP\anaconda3\lib\site-packages\sklearn\utils\deprecation.py:87: FutureWarning:
Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and wil
l be removed in 1.2. Please use get_feature_names_out instead.
  warnings.warn(msg, category=FutureWarning)
```

In [53]:
```
words
```

Out[53]:
```
['average', 'excellent', 'good', 'not', 'poor', 'rated', 'very']
```

In [54]:
```
# Sum up counts per word
word_counts = word_counts.sum(axis=0).tolist()[0]

# Convert to dense array
#word_counts = word_counts.toarray()
```

In [55]:
```
word_counts
```

Out[55]:
```
[3737, 301, 3179, 2148, 186, 2148, 1079]
```

In [76]:
```
# pos_words = []
# for key, count in sorted(word_counts.items(), key=lambda x: x[1], reverse=True)[:5]:
#  pos_words.append(key)
```

In [82]:
```
# Get top positive words
# pos_words = [words[i] for i in sorted(word_counts.items(), key=lambda x: x[1], reverse
import numpy as np
```

In [85]:
```
array = np.array(word_counts)
```

Out[86]:
```
array([3737,  301, 3179, 2148,  186, 2148, 1079])
```

In [88]:
```
# Get top positive words
pos_words = [words[i] for i in array.argsort()[-5:][::-1]]
```

In [89]:
```
# Get top negative words
neg_words = [words[i] for i in array.argsort()[:5]]
```

In [90]:
```
# Print results
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

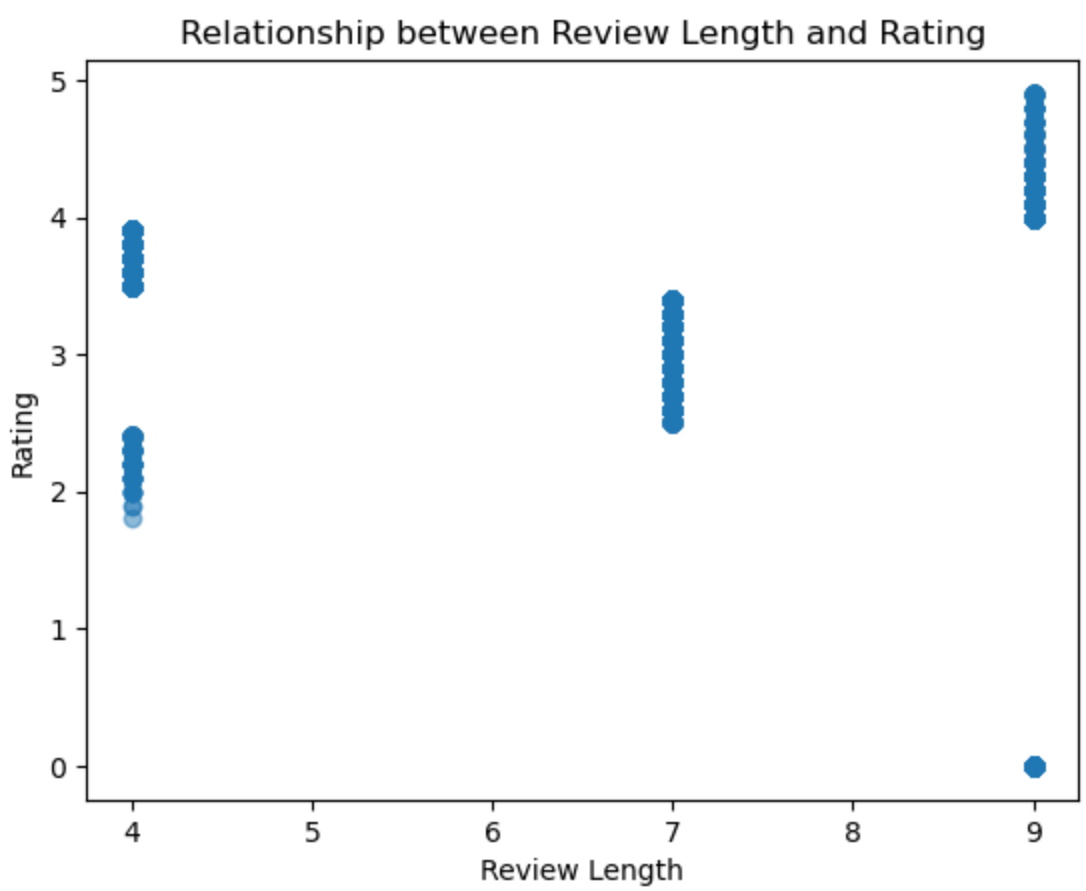
In [94]:
```
# Calculate the average review length
average_review_length = dt['Rating text'].str.len().mean()

# Print the average review length
print(average_review_length)
```

```
7.020730813527379
```

In [98]:
```
# Create a scatter plot of the review length and rating
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 [101…

```
# Calculate the correlation coefficient between review length and rating
correlation_coefficient = dt['Rating text'].str.len().corr(dt['Aggregate rating'])

# Print the correlation coefficient
print(correlation_coefficient)
```

```
-0.47888483813493266
```

In [ ]: