For exploratory data analysis, we want to analyze the corpus with details by identifying the most frequent words in the corpus and this can be established by using WordCloud. The first step is to expand contractions. The shortened version of words “don’t” is changed to “do not”. The next step is by converting all documents into lowercase letters to standardize them. Any punctuation like commas, full stops, hyphens are removed. The next step is preparing data for Document Term Matrix where it provides the frequency of words in a collection of documents. In this section, the steps include removing stop words which are the most common words such as ‘I’, ‘this’, ‘is’.

1. Expand Contractions
2. Lowercase characters
3. Punctuation Removals
4. Stop Words Removal
5. Lemmatization

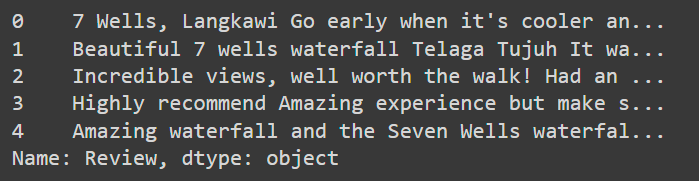


Figure 7: Before processing steps.

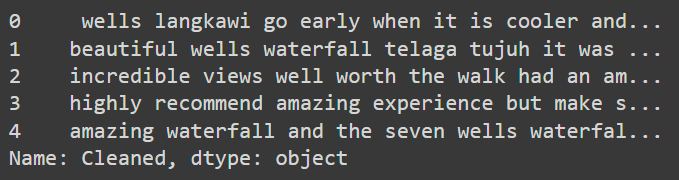


Figure 8: After processing steps.

The following step is lemmatization where words are converted into their base form. The SpaCy and wordcloud library is used for these steps. After that, a document-term matrix is created and wordcloud library is used to plot those words. Word clouds are the visualization of words where the bigger the size of the word, the more frequent it appears in the corpus.

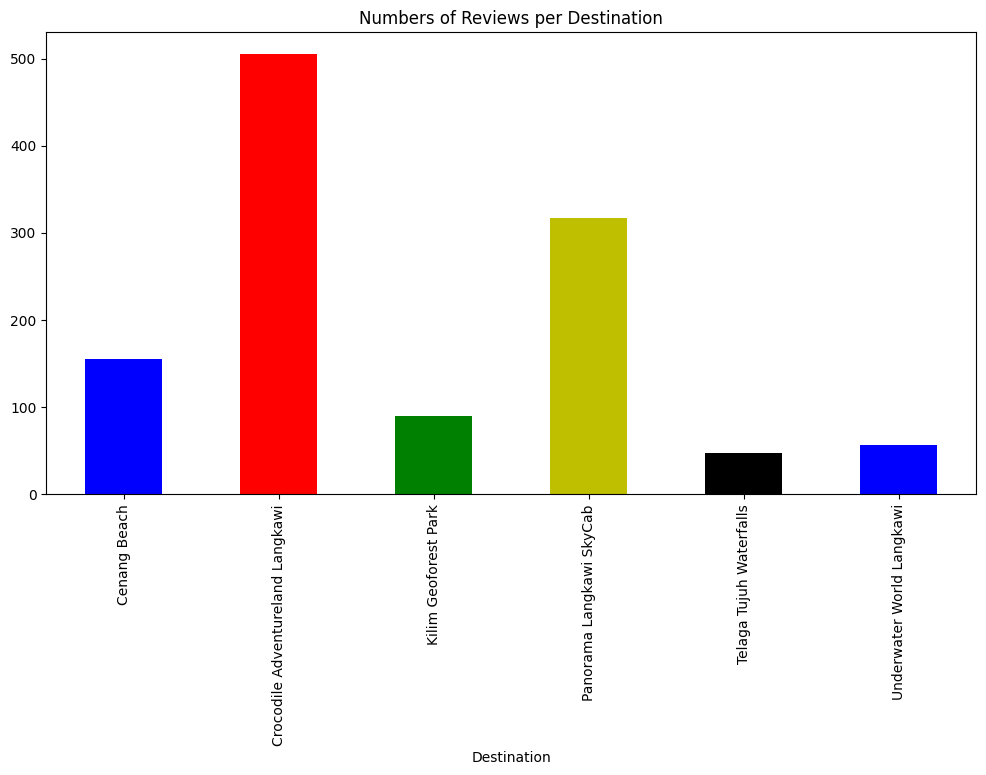


Figure 9: Number of Reviews per Destination.

Figure above is a bar chart detailing the numbers of reviews per destination in Langkawi Island. These places for visiting are Cenang Beach, Crocodile Adventureland Langkawi, Kilim Geoforest Park, Langkawi SkyCab, Telaga Tujuh Waterfalls, Underwater World Langkawi. Among the activities, Crocodile Adventureland Langkawi has received many reviews around 505 whereas Panorama Langkawi SkyCab has nearly 317 number of reviews. Other places have reviews comparatively lower, with Cenang Beach and Underwater World Langkawi receiving 155 and 57 reviews each, Kilim Geoforest Park contains 90 reviews, and Telaga Tujuh Waterfalls contains 47 reviews.

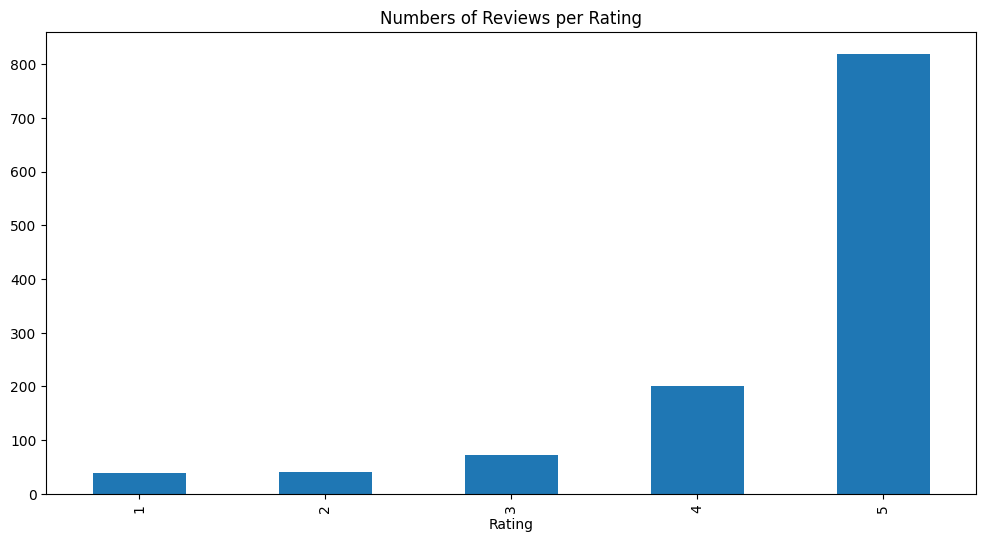


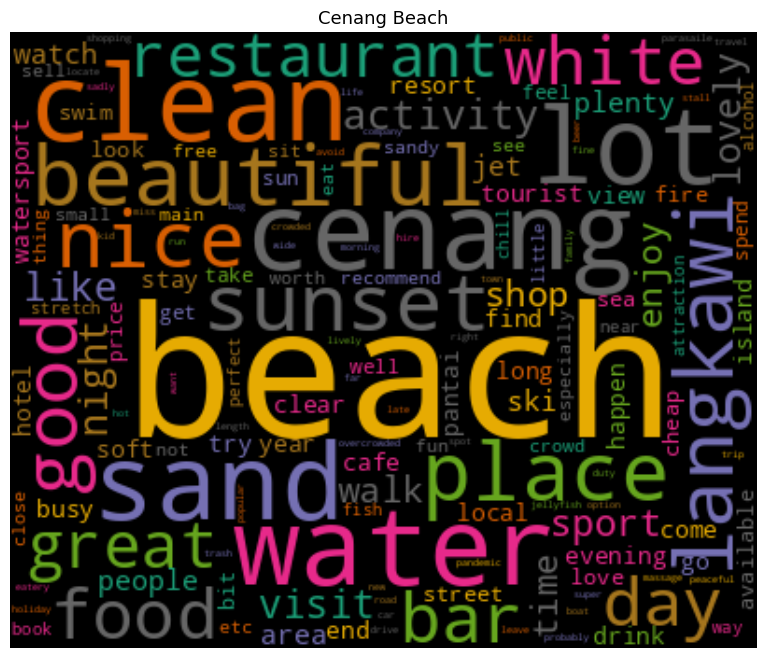
Figure 10: Number of Reviews per Rating.

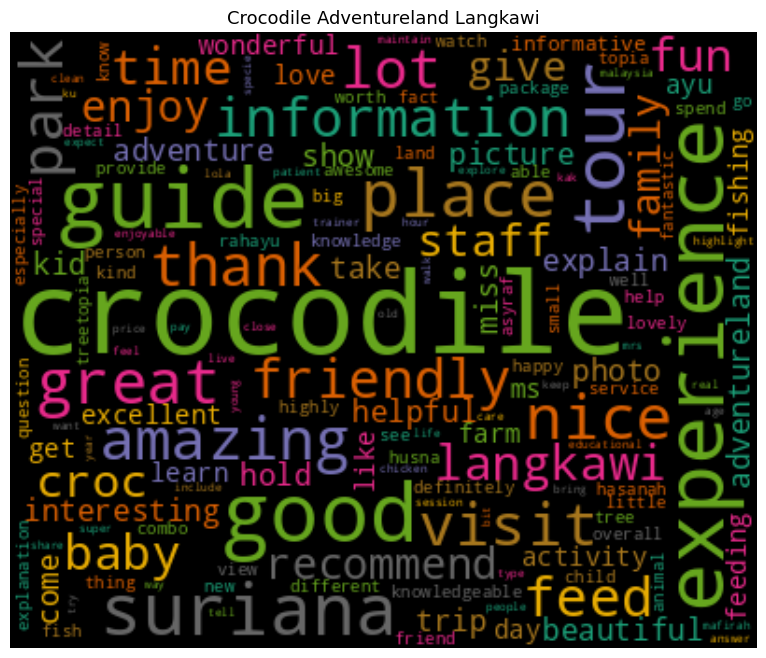
The second graph is a bar chart labeled “Numbers of Reviews per Rating,” which gives a breakdown of reviews based on the rating scale of 1 to 5 stars. The most numerous, about 819, of the reviews are positive, and rated 5-stars, which creates a positive impression. The number of reviews gradually reduces as the rating goes down, 4-star products having 201 reviews, 3-star products – around 72 reviews, while 1 and 2-star products received 41 and 38 reviews. That distribution provide the evidence that majority of visitors share positive impression of any place.

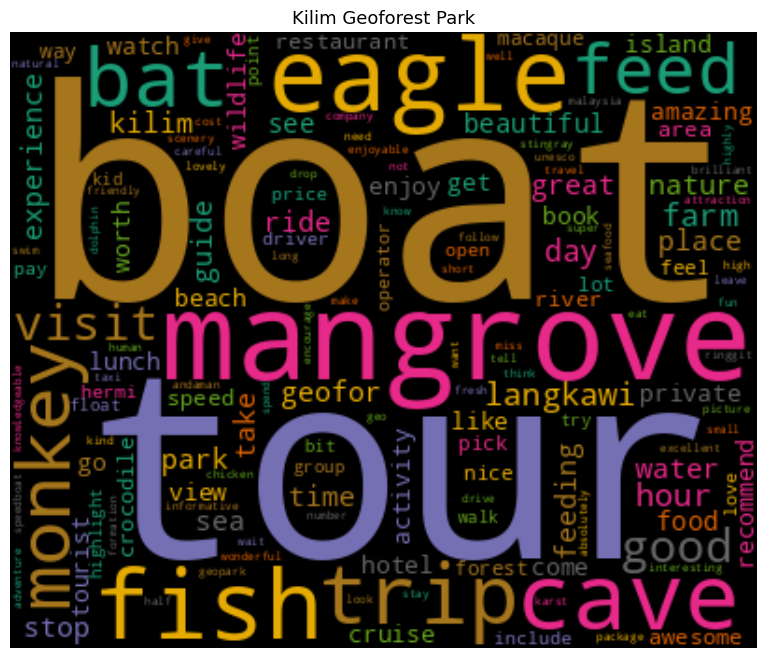


Figure 11: WordCloud of the corpus.

Figure above is a word cloud that shows what words are used most frequently in the reviews regarding Langkawi. Different sizes of words are used to allow a quantitative understanding of word frequency, with larger words pointing to higher frequency. Words like ‘Langkawi’, ‘visit’, ‘place’, ‘good’, ‘beach’, ‘crocodile’ and ‘time’ appear often in the cloud, indicating that many people in the review tend to use them and therefore probably highlight topics or aspects of the experience the writers like. Other distinguishable words include ‘nice’, ‘view’, ‘amazing’, ‘cable car’, ‘tour’, ‘guide’, ‘fun’ and ‘family’ suggesting that they represent important aspects of visitors’ experiences. The use of the word cloud also maps out positive descriptors such as ‘beautiful,’ ‘friendly,’ ‘great,’ and ‘enjoy,’ suggesting that the perception of the destinations in Langkawi may be largely positive. Furthermore, there are words and phrases like “feeding”, “riding”, “boat” and “park” that are characteristic of certain tourist attractions. In summary, the word cloud gives an idea about the general gist with regards to sentiments and topics touched in the reviews section.











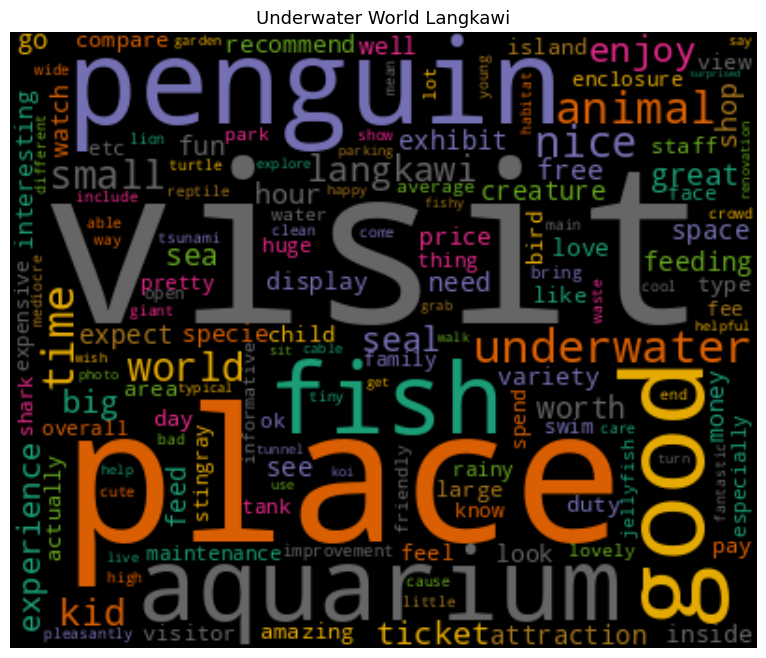


Figure 12: WordCloud for each destination.

Figure above presents six various types of the word clouds, which illustrate the most often used words in the reviews related to the given destinations in Langkawi. Some of the attractions include Cenang Beach, Panorama Langkawi SkyCab, Crocodile Adventureland Langkawi, Telaga Tujuh Waterfalls, Kilim Geoforest Park and the Underwater World Langkawi. In the case of Cenang Beach, the key words identified are “beach”, “sand”, “water”, and “sunset”, setting the context of a beautiful natural beach with activities to do. Analyzing the word frequency of the Panorama Langkawi SkyCab, words such as ‘view,’ ‘sky,’ ‘ticket,’ and ‘cable car,’ which are relevant to the location and the main attraction. In the case of Crocodile Adventureland Langkawi, the cloud consists of the words “crocodile”, “guide”, “tour”, and “show” suggesting that visitors to the park will get to see crocodiles and be guided on tours around them. The word cloud of Telaga Tujuh Waterfalls are “waterfall”, “walk”, “step”, “pool”, indicating that there is an emphasis on the natural environment and walking. The cloud words focusing on Kilim Geoforest Park are ‘boat’, ‘tour’, ‘mangrove’ and ‘fish’, reflecting key attractions including boat tours and mangrove fishing. Finally, Underwater World Langkawi’s clouds are “penguin”, “visit”, “fish” and “aquarium” and focus on many exhibits of marine life. In general, these word clouds present the most prevalent terms and experiences in relation to each place, thus offering insights into tourist experience.