Destination Analysis and Targeted Marketing for Egyptian Tourism

Ramy Anwar El-Adl   ramy.a.eladl@gmail.com

Mohamed Khaled   m.salem250@gmail.com

Amr Essam   amrmoustafa5@gmail.com

Omnia Gad   omnia.gad92@gmail.com

Nesma Silman   Nesma.silman@gmail.com

1. Introduction

Tourism is an essential aspect of Egypt's economy, and understanding the preferences and interests of tourists can help promote specific locations as an appealing tourist destination effectively. To achieve this goal, we will conduct comprehensive research and gather data from various sources. We will then use machine learning algorithms to analyze this data and identify patterns and trends that can help us understand tourists' preferences better.

By leveraging the power of machine learning, we will gain valuable insights into the key factors that motivate tourists to choose specific destinations. This information will enable us to develop effective marketing campaigns that target specific tourist segments and promote Egypt's unique attractions. We are excited to embark on this journey and look forward to working on this project. We are confident that our efforts will help position Egypt as a top tourist destination and boost the country's tourism industry.

2. Project Objectives

a. Destination Analysis: Conduct an in-depth analysis of popular tourist destinations in selected countries to identify key factors that attract visitors. This analysis will include factors such as cultural heritage, natural landscapes, historical significance, recreational activities, and overall experience.

b. Tourist Preferences: Collect data through online platforms to understand tourists' preferences and their experiences in various destinations. This will help identify the aspects of these destinations that resonate most with tourists.

c. Comparative Analysis: Compare the findings from different tourist destinations to identify common themes and patterns. Determine the unique selling points and competitive advantages of Egypt as a tourist destination in comparison to other countries.

3. Expected Deliverables

     I.   Comprehensive analysis of popular tourist destinations in selected countries.
    II.   Detailed report on tourist preferences and what they find appealing in these destinations.
   III.   Comparative analysis highlighting Egypt's unique selling points and competitive advantages.

4. Dataset and Features

We started the project by preparing a list of familiar tourist places in Egypt and Turkey, and then extended the data to add places from Thailand. This list was obtained from Chatgpt. Our initial data included each place and the category to which this place could be assigned. Below is a description of the features we extracted,

- **Location**: The city in which each place is located.
- **Country**: The country in which each place is located.
- **Historical Significance**: Relates to the importance of the location in terms of its historical events, heritage, and cultural value. It allows travelers to explore the rich history and traditions of the place.
- **Natural Beauty**: Encompasses the scenic landscapes, biodiversity, and natural wonders that the location offers. It appeals to travelers seeking aesthetically pleasing environments and outdoor experiences
- **Adventure Activities**: Includes various thrilling and exciting experiences such as hiking, rock climbing, ziplining, or any other adrenaline-pumping activities that cater to adventurous travelers.
- **Accessibility**: Refers to how easily a location can be reached or explored by different means of transportation. Good accessibility allows for smooth travel and exploration.
- **Shopping**: Describes the availability of shopping opportunities in the location, including markets, malls, and unique local products. It attracts travelers interested in purchasing souvenirs or experiencing the local retail scene.
- **Nightlife**: Represents the entertainment and social activities that take place during the evening and nighttime. It appeals to travelers looking for vibrant nightlife experiences, such as clubs, bars, or cultural events.
- **Water Sports:** Encompasses a variety of activities that take place in water bodies, like swimming, snorkeling, kayaking, or surfing. It attracts travelers who enjoy aquatic adventures.
- **Wildlife Viewing:** Focuses on the opportunities to observe and appreciate the local fauna and wildlife in their natural habitats. It appeals to nature enthusiasts and wildlife photographers.
- **Scuba Diving**: Specifically refers to the underwater diving activity that involves using self-contained underwater breathing apparatus (scuba). It allows travelers to explore the marine world and its diverse ecosystems.
- **Type**: This may refer to the categorization of the destination, such as whether it is a beach resort, a historic city, an eco-tourism destination, or an adventure sports hub. It helps travelers choose a location based on their preferred type of travel experience.

These columns contain categorical values that express the relevance of each feature to each place.

Next, we moved to the second step in our data collection. Our main source of data here was https://tripadvisor.com/. Which is considered the world's largest travel site. The site contains data about tourist destinations around the world and their reviews.

Using web scraping, we collected all the URLs of our places from Trip Advisor website, and we used these URLs to extract the reviews written about these places and their details (review date, title, star...etc.) as well as reviewers' data. We restricted the number of collected reviews to a maximum of 2000 reviews per place to optimize the time consumption.

Pre-processing steps were taken on the collected data before starting our analyses. We noticed that some URLs collected from scraping do not match our place of interest, so we used Levenstein distance technique to validate the data and use accurate URLs.

We performed sentiment analysis on reviews to identify their polarity using Vader. We also calculated the % of positive reviews per place so that we could filter out the places with a given % of positive reviews.

We noticed that the values in "Country" column are inconsistent. For example, some reviewers wrote their home country as "UK", some used "United Kingdom" and some used "London". So, we unified all the different names that belong to the same country by using geopy.geocoders Nominatim.

We wanted to understand how the trend of visits change per season. However, the site doesn't contain information about the date of visits. So, we assumed that the reviews were written nearly in the same season in which the visit took place, and we mapped the reviews date to each of the four seasons and used the data in our analysis.

We took steps to perform topic modeling on our data using Gensim library. However, the outcome wasn't informative enough. So, we skipped this step.

Finally, we used the above data to perform clustering on our places using K-means technique.
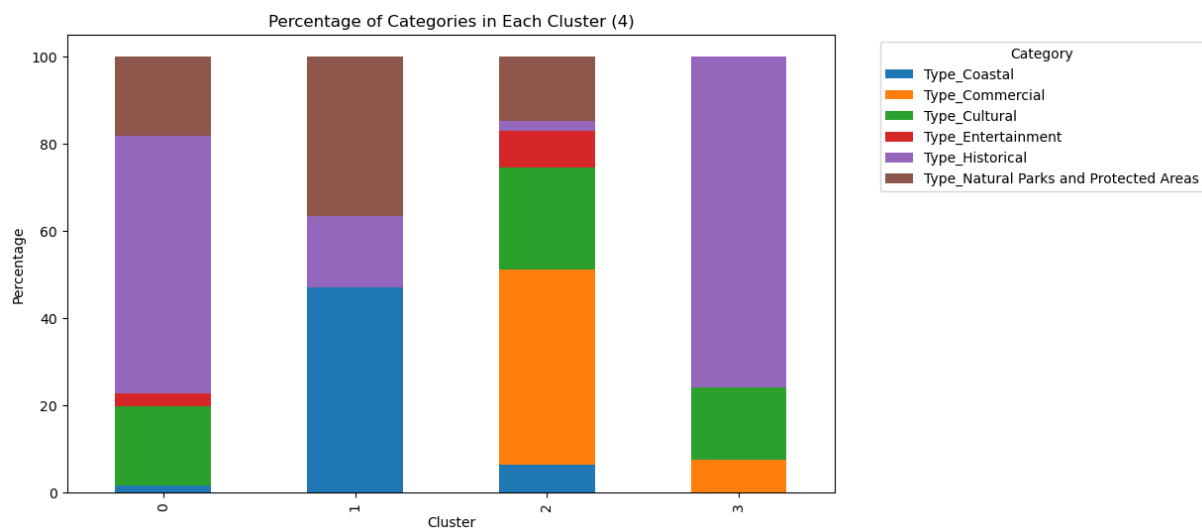

5. Methodology

We set out with the objective of grouping the list of locations by utilizing the categories gathered during the initial phase. Our goal was to generate distinct clusters, each consisting of comparable places. We experimented with creating 7, 6, 5, and 4 clusters. And we observed the Elbow Method and silhouette below,
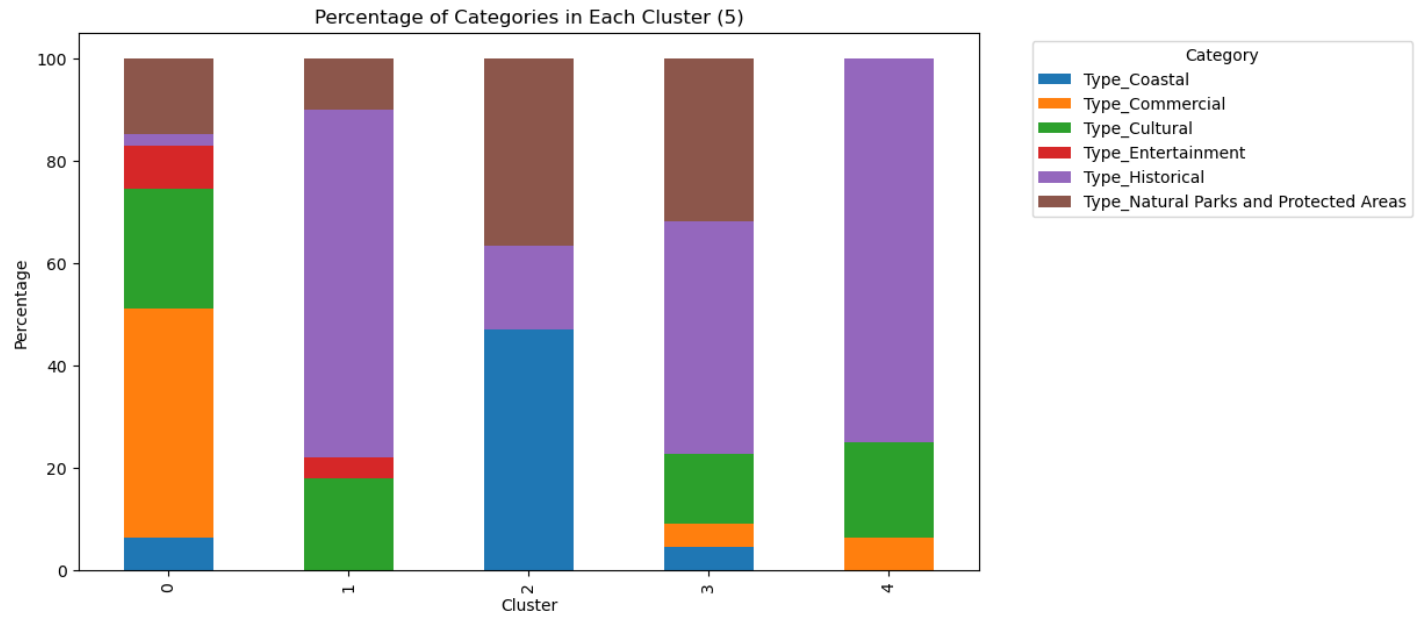
Elbow Method
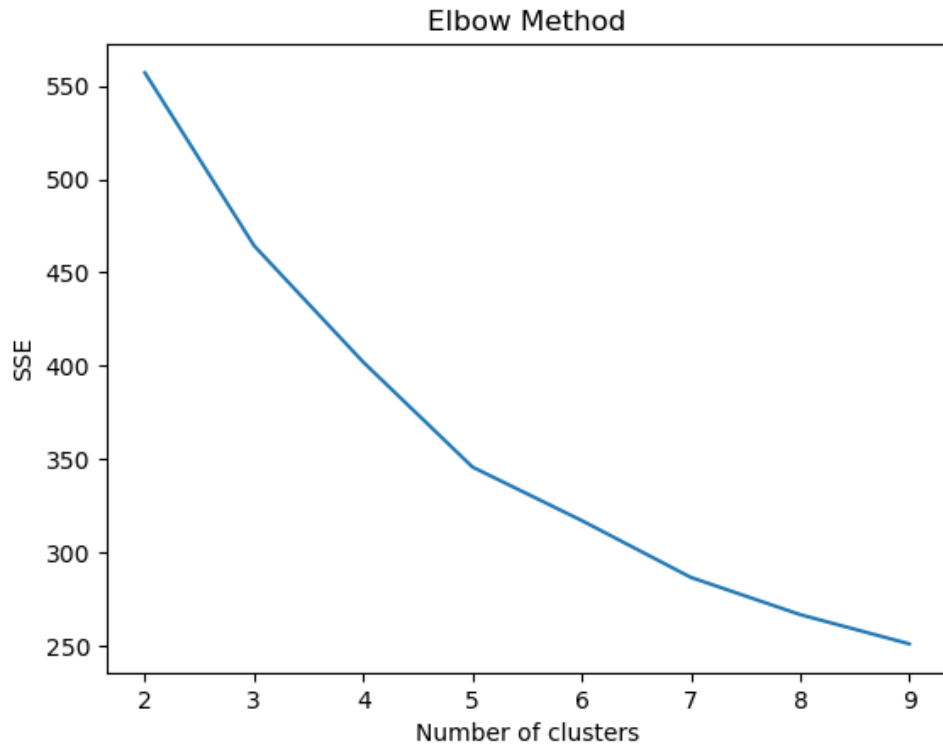
- Silhouette score for k=4: 0.27775321651688833
- Silhouette score for k=5: 0.298341387444546
- Silhouette score for k=6: 0.3213010496381913
- Silhouette score for k=7: 0.3218767204461577

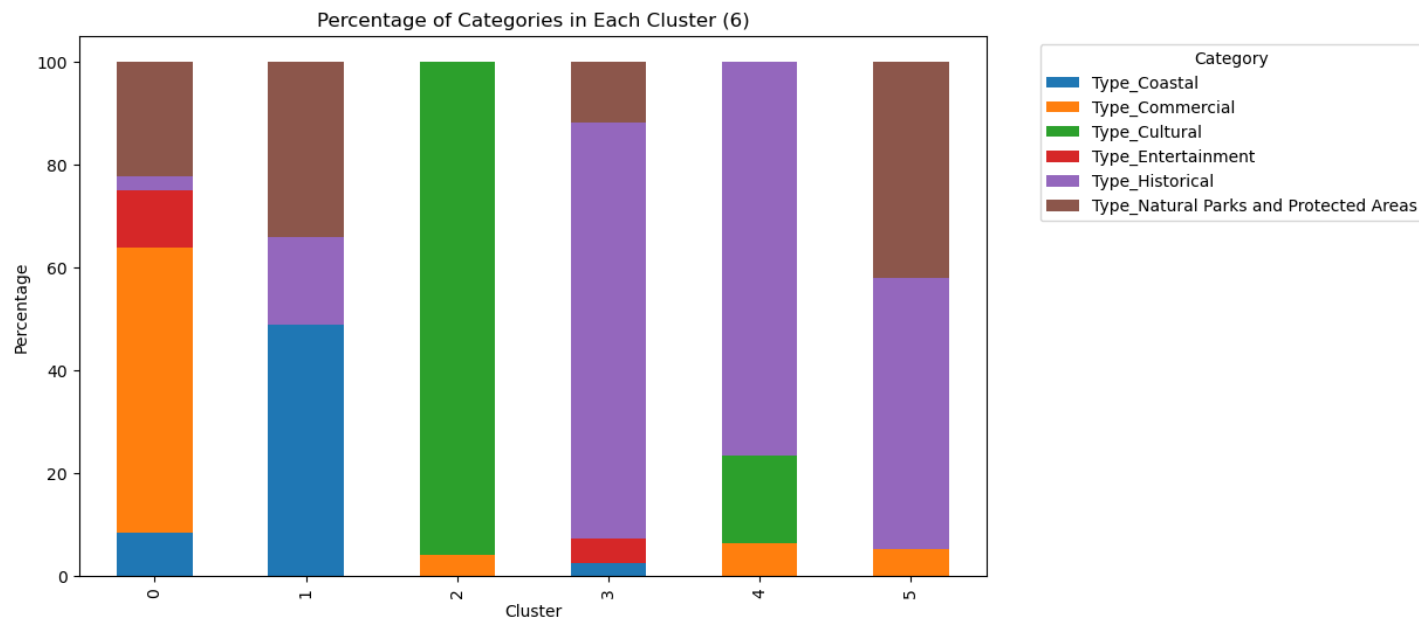**Visualizing the relation between the categorizes and the number of clusters:**
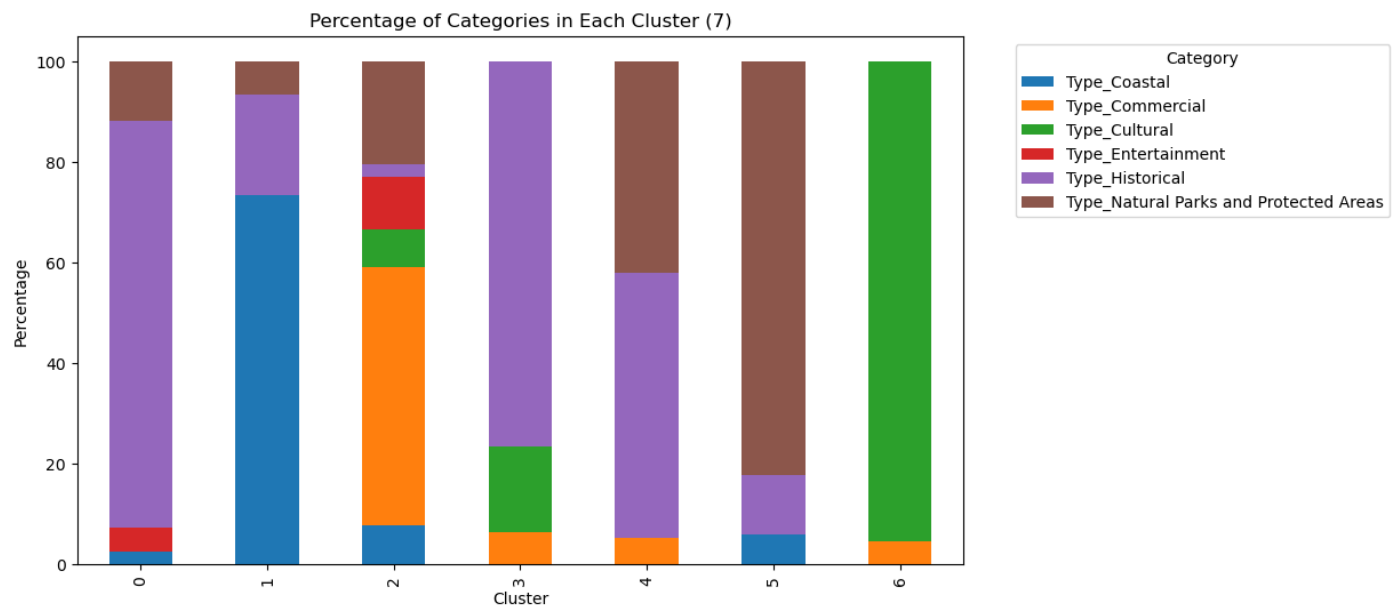
➢ **4 Clusters:**

Percentage of Categories in Each Cluster (4)

➤ **5 Clusters:**

Percentage of Categories in Each Cluster (5)

➤ **6 Clusters:**

Percentage of Categories in Each Cluster (6)

➢ **7 Clusters:**

Percentage of Categories in Each Cluster (7)

6. Conclusion

Following the above comprehensive analysis, we notice that categories are more dispersed for clusters 4, 5 & 7 than for cluster 6. Moreover, considering the Elbow method and silhouette outcome, we concluded that the optimal number of clusters for our study is 6-clusters.