BA820 - Group 9 - Project Deliverable 2

Hospitality Intelligence: Extracting Insights from European Hotel Reviews

By Sricharan Mahavadi, Chen He, Eric Bai, Min Xu, Olimpia Borgohain

Link to Colab : [BA820-Project-Notebook]

Data Pre-Processing:

To address missing geographic data, we used Google Maps API to establish a geocoder object, which allows us to fetch geographical coordinates based on hotel addresses. We also capped ratings at 10 and filtered out outliers with review length under 20 words and more than 500 words and finished by merging dataframes and removing redundant data.

Exploratory Data Analysis:

To fully explore the dataset, we analyzed the dataset from the following four perspectives. Here are the key findings from our analysis:

- **Geographical Spread:** Cluster visualization is adopted for hotel geographical distribution with each marker showcasing hotel's name and review score (map 1.0).
- **Review Change over Time:** Hotel reviews reveal seasonal trends with a peak during summer and decline towards year-end, which align to peak travel season (plot 2.1).
- **Review Scores Variations:** Scores center around 8-9, suggesting a majority of positive reviews, while a longer left-skewed tail indicates various low reviews (plot 3.1 & 3.2).
- **Reviewers Nationalities:** The UK leads with 66.3% of reviewers, followed by the USA at 9.6% and Australia at 5.9% (plot 4.1). Some countries show higher satisfaction levels while others rate lower, possibly due to smaller sample sizes and less variance (plot 4.2)..

Analysis Plan: Our analysis will focus on 3 key parts:

- **Text preprocessing and sentiment Analysis:**
  - Extract useful features from reviews using NLP and other techniques
  - Improve the text-to-tags data pre-processing by adopting n-grams & word embedding techniques, preparing data for further analysis
  - Conduct sentiment analysis to improve understanding of customers' reviews
- **Customer segmentation:**
  - Apply K-Means, K-Means++ clustering on review patterns to group customers
  - Use hierarchical clustering to identify similar customer segments
  - Explore density-based clustering like DBSCAN to detect outliers
- **Recommendation system:**
  - Create a word cloud with a dropdown menu for customers to select their preferred hotel, showcasing popular tags associated with each choice.
  - Use market basket analysis to identify most frequently used tags for each hotel, apply apriori to develop a rule to set up recommendation mechanism
  - Enhance the recommendation process by enabling customers to select their nationality and desired hotel features.

Preliminary Results:

- **Market basket Analysis and Recommendation:**
  - Treated each dataset row as an individual transaction, combining 'Hotel_Name' with 'Tags' to represent items in a transaction.
  - Applied the Apriori rules, with hotel names as antecedents to highlight commonly associated tags in reviews.
  - Developed a function allowing customers to input a hotel name to see the most frequently mentioned tags.
- **Clustering:**
  - K-Means, K-Means++: Implemented these two clusterings on a downsampled data(~10k rows), employing the elbow plot method to determine the optimal value of K, which was found to be 4 and 5 respectively. (plot 5)
  - Hierarchical & DBSCAN: Implemented these two clusterings, with total 6 clusters, revealing distinct patterns, with one cluster containing the most values.
- **Dimensionality Reduction with PCA:**
  - Implemented PCA to address the high dimensionality of the dataset.
  - Utilized a scree plot to find out that, out of the initial 10 features, only 6 were needed to capture 85% of the dataset's variance. (plot 6)

Next Steps :

We noted that the 'Tags' column data missed critical aspects valued by customers. Thus, in the next step, we will:

- Conduct sentiment analysis on review columns to extract and identify features frequently mentioned in positive reviews, addressing the initial data limitation.
- Create a word cloud that enables customers to select a hotel, based on the outcomes of advanced word extraction and sentiment analysis.
- Introduce an interactive dropdown menu featuring the top 20 most frequently mentioned features from reviews, enabling user-preferred feature selection.

Team Collaboration : The following figure shows the division of work and responsibilities amongst the team members

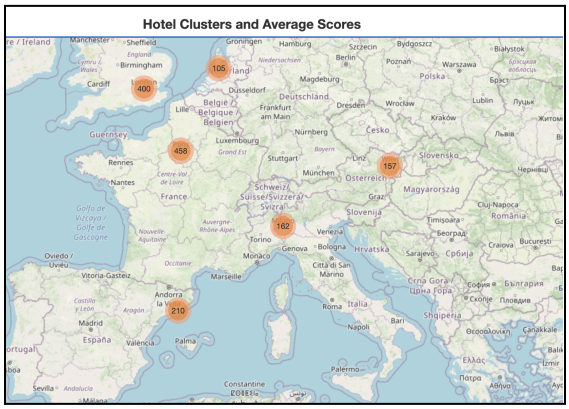
Team Member	Contributions
Chen	Data Cleaning (40%), Recommendation System 1 (50%), DBSCAN Clustering (30%)
Sricharan	Exploratory Data Analysis (30%), Recommendation System 1 (30%), Hierarchical Clustering (20%)
Eric	Initial Analysis (30%), Market Basket 2 (25%), PCA (25%), Recommendation System 1 (50%)
Min	Market Basket 1 (50%), Recommendation System 2 (20%), PCA (25%)
Olimpia	Market Basket 2 (25%), K-Means Clustering (50%), Dimensionality Reduction (50%)

A more detailed collaboration can be found in Fig C-1 and Collaboration Sheet

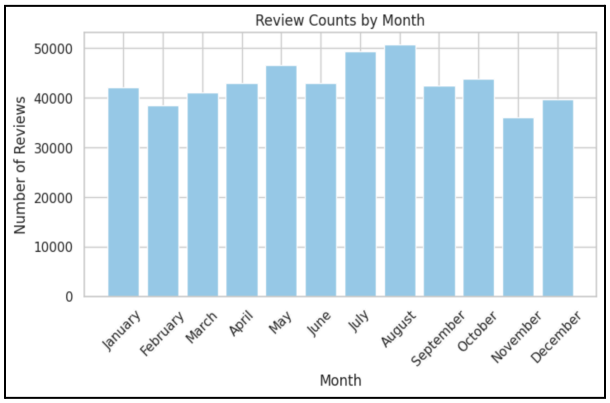
Kanban Board : A more detailed project tracking can be found here : BA820 Group 9 Kanban Board (Plot attached below)

Github Project link :A link to our project can be found here : Github Project

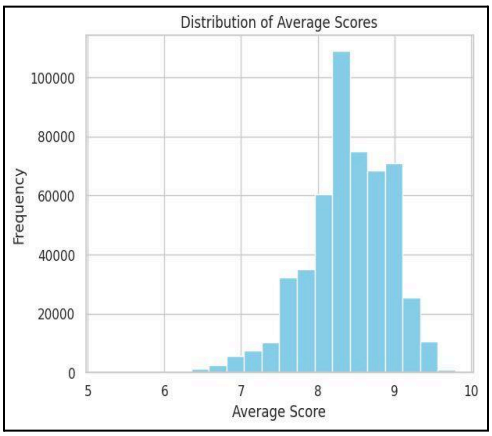
Appendix



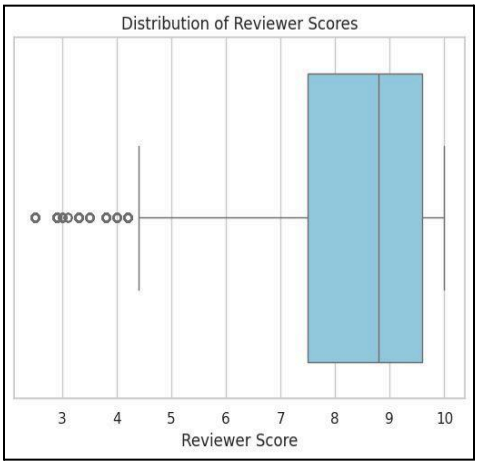
Map 1.0



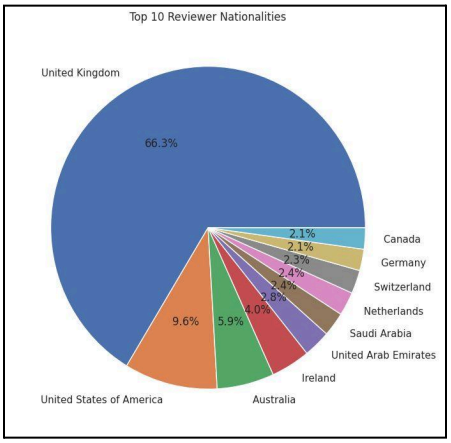
Plot 2.1



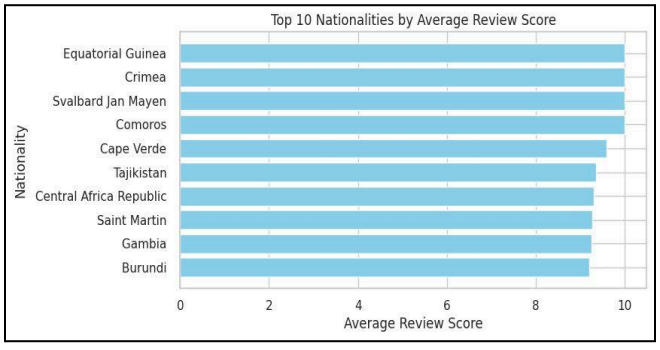
Plot 3.1



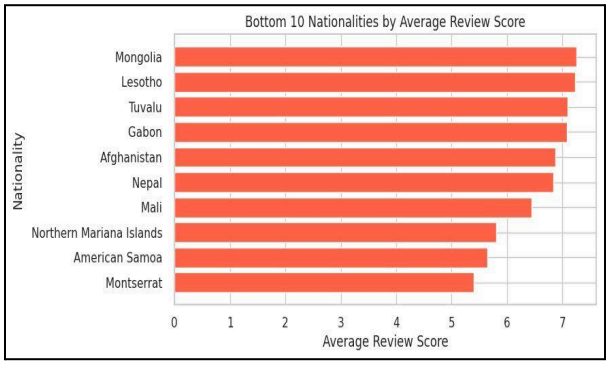
Plot 3.2



Plot 4.1

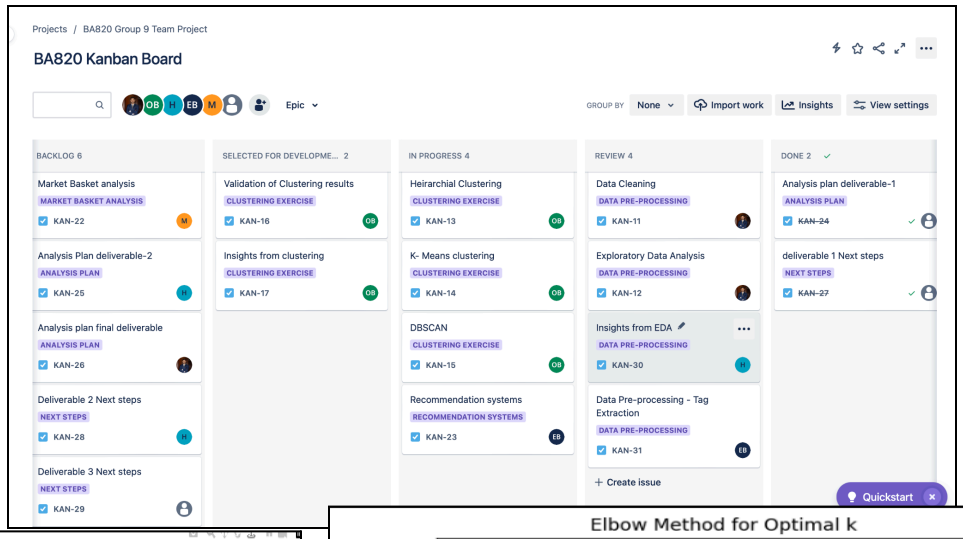


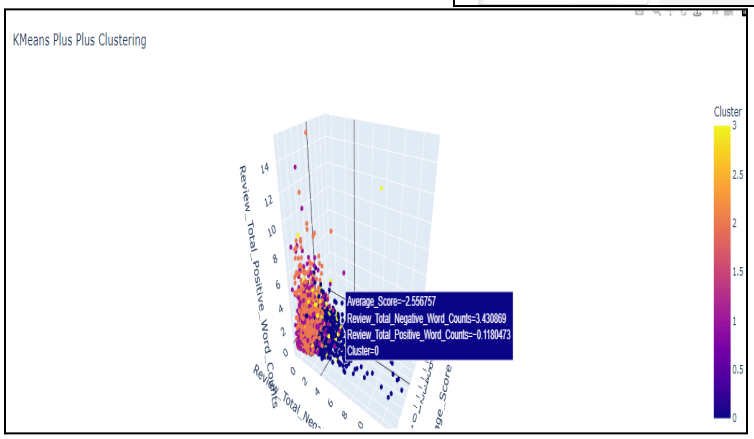
Plot 4.2



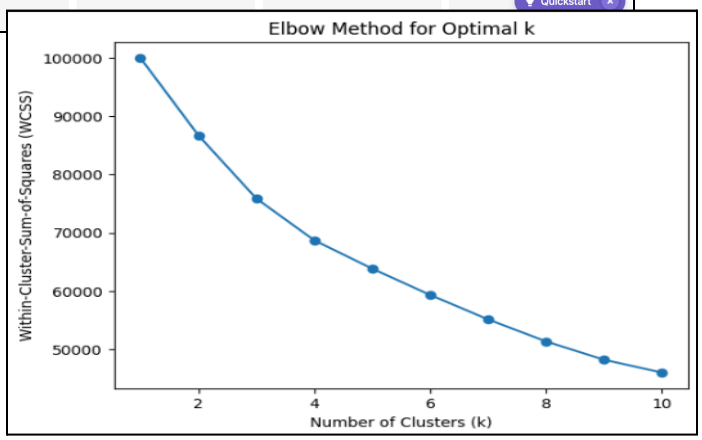
Plot 4.3

Kanban Board Interface:

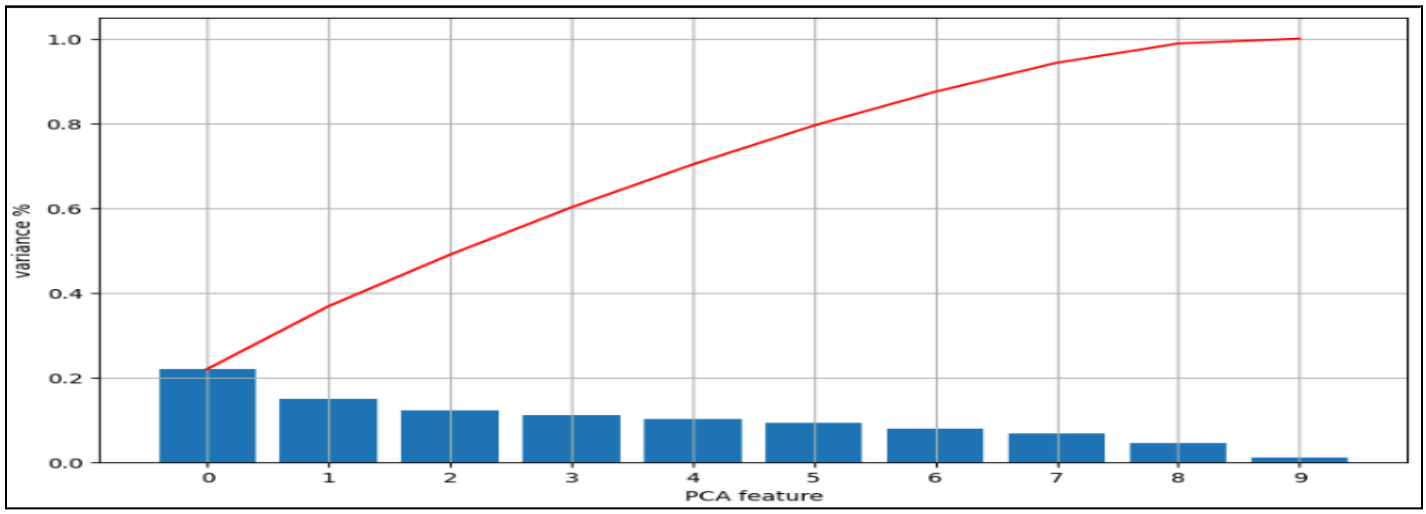




Plot 5.1



Plot 5.2



Plot 6

Team Contributions : GitHub