BA820 - Group 9 - Project Deliverable 3

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

Project Motivation :

Hotels must exceed diverse guest expectations to thrive today. By unlocking insights from customer segments and hotel reviews at scale, hotels can understand what customers truly value, guiding improvements for more personalized, cozy stays.

Dataset and Source :

- The dataset is from Kaggle, with the original data web scraped from **Booking.com**. It's a comprehensive collection of **~515K** records from **1450+** luxury hotels across Europe, featuring customer reviews, ratings and geographical data. The source can be found **here**;
- The dataset contains 17 features, including text-based customer reviews, numerical data such as average score, latitude, longitude, as well as categorical data such as hotel name, reviewer nationality, and review tags.

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 lengths 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 the hotel's name and review score (map 1.0).
- **Review Change over Time:** Hotel reviews reveal seasonal trends with a peak during summer and a decline towards year-end, which aligns with 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 (plots 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 the most frequently used tags for each hotel, apply apriori to develop a rule to set up a recommendation mechanism
  - Enhance the recommendation process by enabling customers to select their nationality and desired hotel features.

Final 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 consequents to highlight commonly associated tags in positive reviews.
  - Created our own tags by extracting important hotel features from reviews.
  - Developed an interactive dropdown allowing customers to choose features they care about and return the hotel recommendation.
- **Clustering:**
  - Preprocessing
    - Used a subset of 10k rows and took the numerical columns
    - Standardized the numerical columns
  - Algorithms
    - Implemented KMeans, KMeans++, hierarchical, and DBSCAN clustering algorithms.
  - Analysis
    - Points having higher values of `Review_Total_Negative_Word_Counts` fall on the same cluster, suggesting that there is a distinct pattern associated with negative sentiments expressed in customer reviews.
    - Likewise, data points characterized by higher values of Review_Total_Positive_Word_Counts are divided into two clusters, determined by their respective average score values.
    - In the case of the DBSCAN algorithm, as the majority of points exhibit dense clustering, the choice of the `min_samples` parameter does not significantly impact the outcome. Regardless of the specific value selected for min_samples, the algorithm consistently results in a single cluster containing the majority of data points.

- **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)

- **Hotel Sentiment Analysis: Insights from Reviews**
  - Preprocessing
    - Cleaned and standardized a sample of 10,000 hotel reviews
    - Tokenized reviews after removing stopwords and lemmatization
  - Text Representation
    - Converted reviews into 300-dimensional Word2Vec embeddings
    - Created separate positive and negative sentiment centroids
  - Modeling
    - Calculated cosine similarity scores between review embeddings and sentiment centroids
    - Trained logistic regression model on 16,000 reviews (80% data)
    - Achieved 0.93 F1-score in classifying positive/negative sentiments on 4,000 test reviews
  - Analysis
    - Mean sentiment score: 0.089 (slightly positive)
    - 25th percentile: -0.051, 75th percentile: 0.245
    - Real time Sentiment Analysis of Customer Reviews
  - Recommendation System
    - Suggested top-rated hotels with positive sentiment for selected city
    - Minimum rating threshold adjustable from 0-10 scale

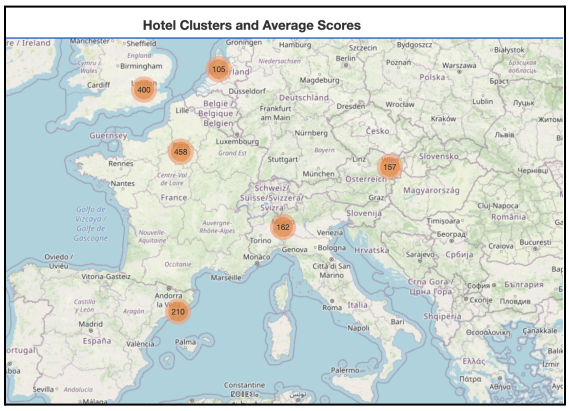
Team Collaboration :
The following figure shows the division of work and responsibilities among the team members

Team	Contributions
Chen He	Data Cleaning, Recommendation System, DBSCAN Clustering
Sricharan Mahavadi	EDA, Recommendation System, Hierarchical Clustering
Eric Bai	Initial Analysis, Market Basket Analysis, PCA, Recommendation System,
Min Xu	Market Basket, Recommendation System, PCA
Olimpia Borgohain	Market Basket, KMeans, Dimensionality Reduction,

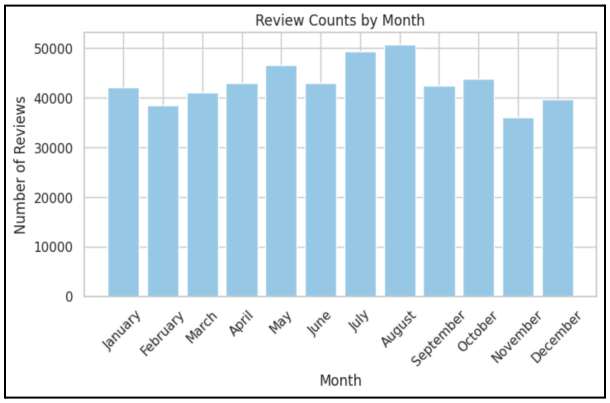
A more detailed collaboration can be found in Fig C-1 and Collaboration Sheet

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

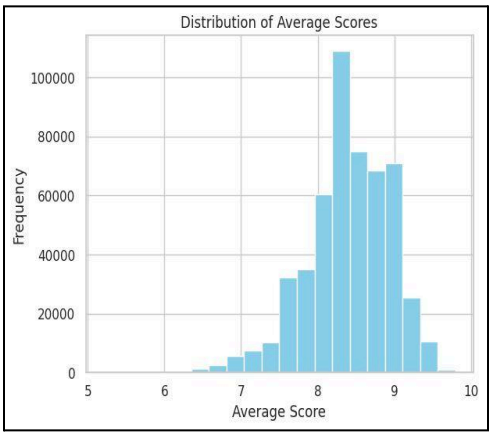
Appendix



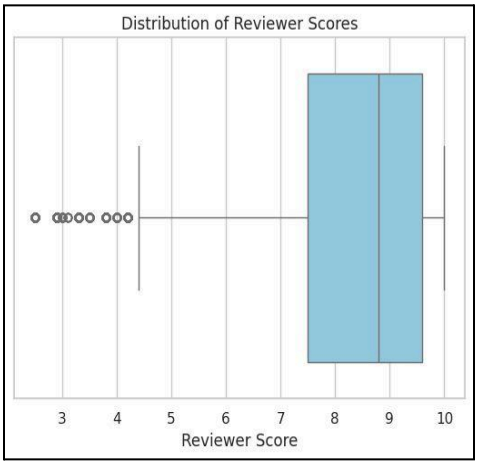
Map 1.0



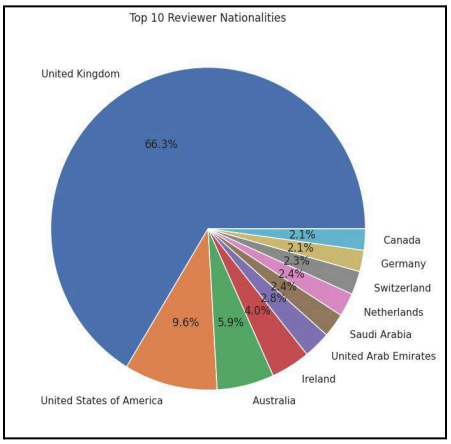
Plot 2.1



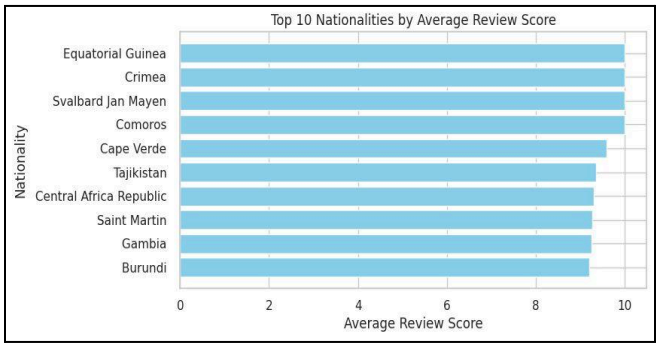
Plot 3.1



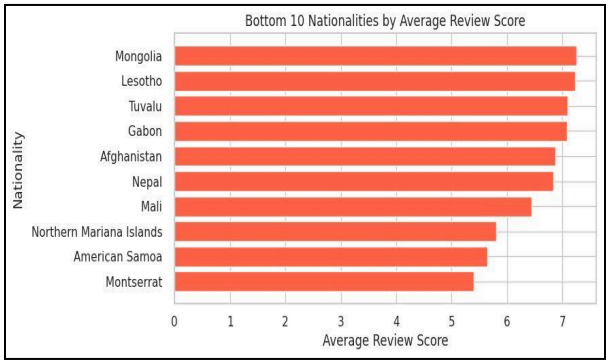
Plot 3.2



Plot 4.1

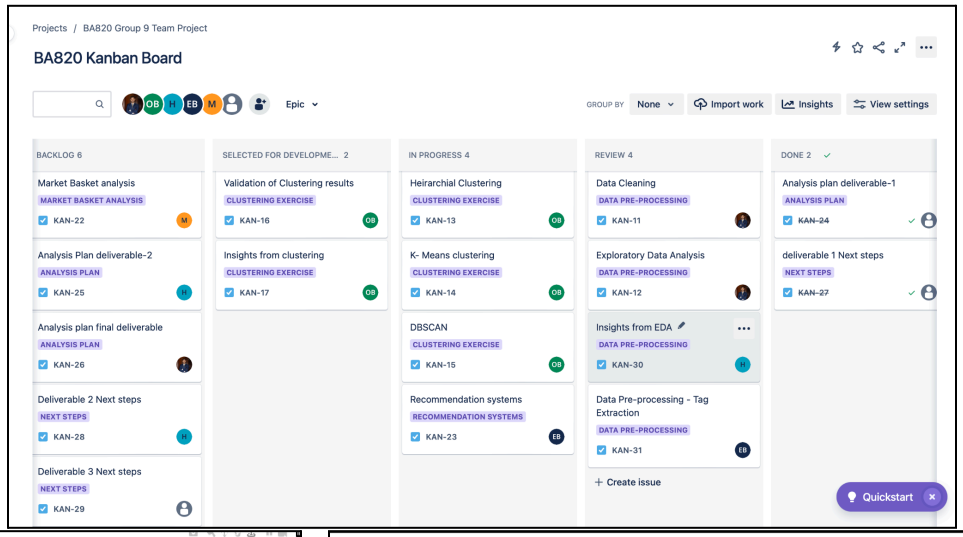


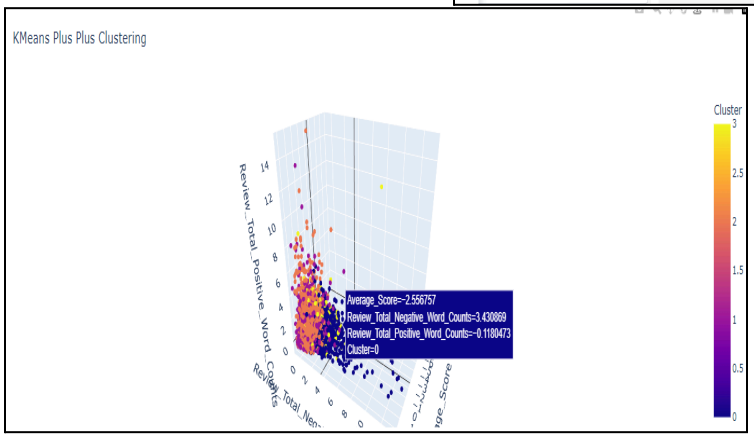
Plot 4.2



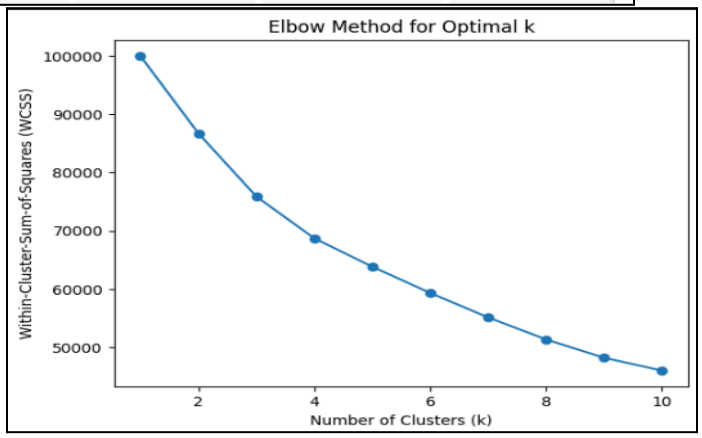
Plot 4.3

Kanban Board Interface:

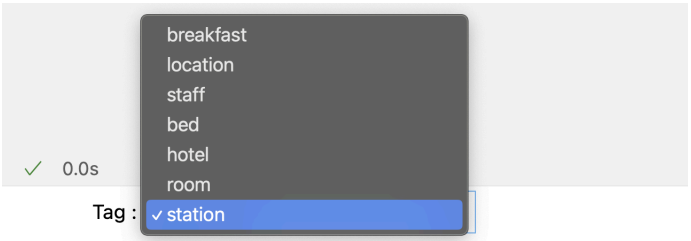




Plot 5.1



Plot 5.2




```
Recommended hotel: the grosvenor
Recommended hotel: nh collection amsterdam barbizon palace
```

New Review analysis:

```
print("The review is not positive.")

/usr/local/lib/python3.10/dist-packages/ipykernel/ipkernel.py:283: DeprecationWarning: `should_run_async` will not call `tra
  and should_run_async(code)
[nltk_data] Downloading package punkt to /root/nltk_data...
[nltk_data]   Package punkt is already up-to-date!
Enter your review: The Hotel is awesome
```

```
print("The review is not positive.")

/usr/local/lib/python3.10/dist-packages/ipykernel/ipkernel.py:283: DeprecationWarning: `should_run_async` will not call `tra
  and should_run_async(code)
[nltk_data] Downloading package punkt to /root/nltk_data...
[nltk_data]   Package punkt is already up-to-date!
Enter your review: The Hotel is awesome
The review is positive.
```

Recommendation Based on sentiment analysis

City: Paris

Rating: 9.20

```
Recommended Hotel:
Name: Le Narcisse Blanc Spa
Address: 19 Boulevard De La Tour Maubourg 7th arr 75007 Paris France
Average Score: 9.5
Reason for Recommendation: This hotel has a high average score and positive sentiment based on customer reviews.
```

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%)

Team Contributions: GitHub