

Privago: Hotels Search System

INFORMATION PROCESSING AND
RETRIEVAL – PRI
MILESTONE #1: DATA PREPARATION

André Ávila – up202006767@up.pt
André Costa – up201905916@up.pt
Fábio Morais – up202008052@up.pt
Fábio Sá – up202007658@up.pt

Project Overview



Theme



Datasets



Data Extraction and Preparation



Data Domain Conceptual Model



Data Characterization

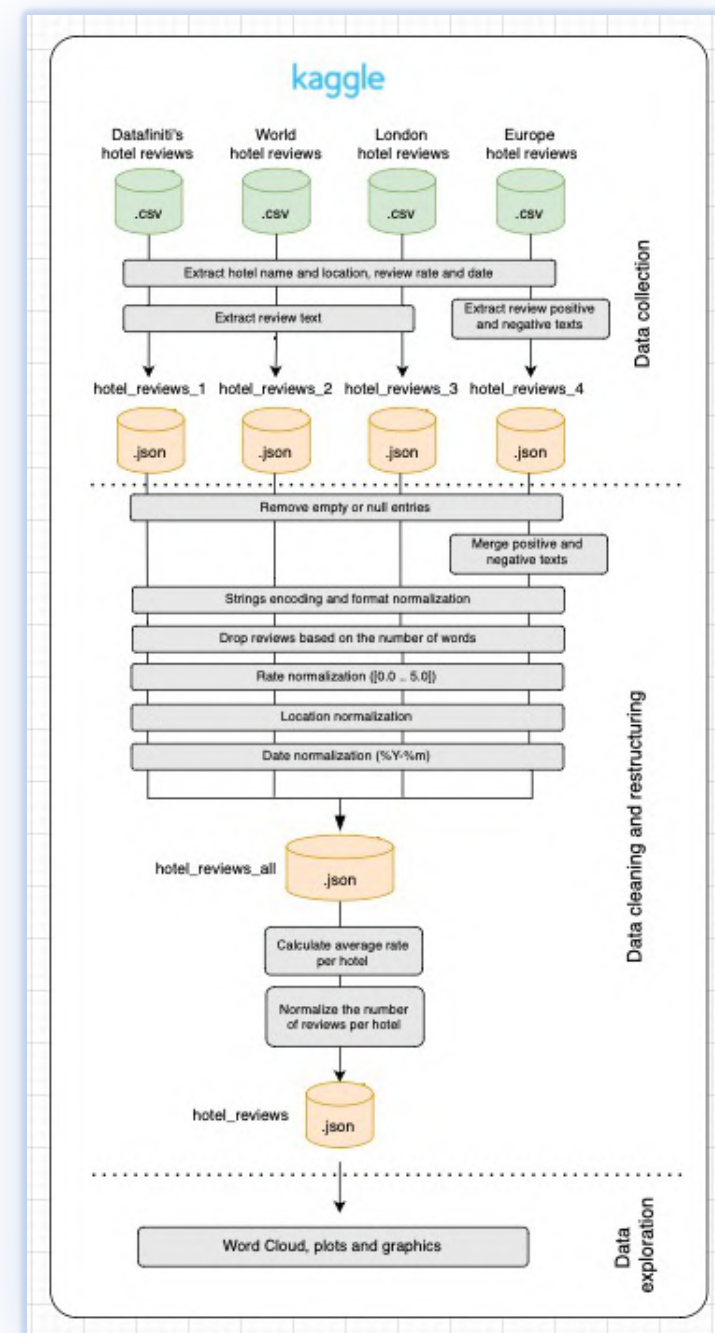


Search Queries



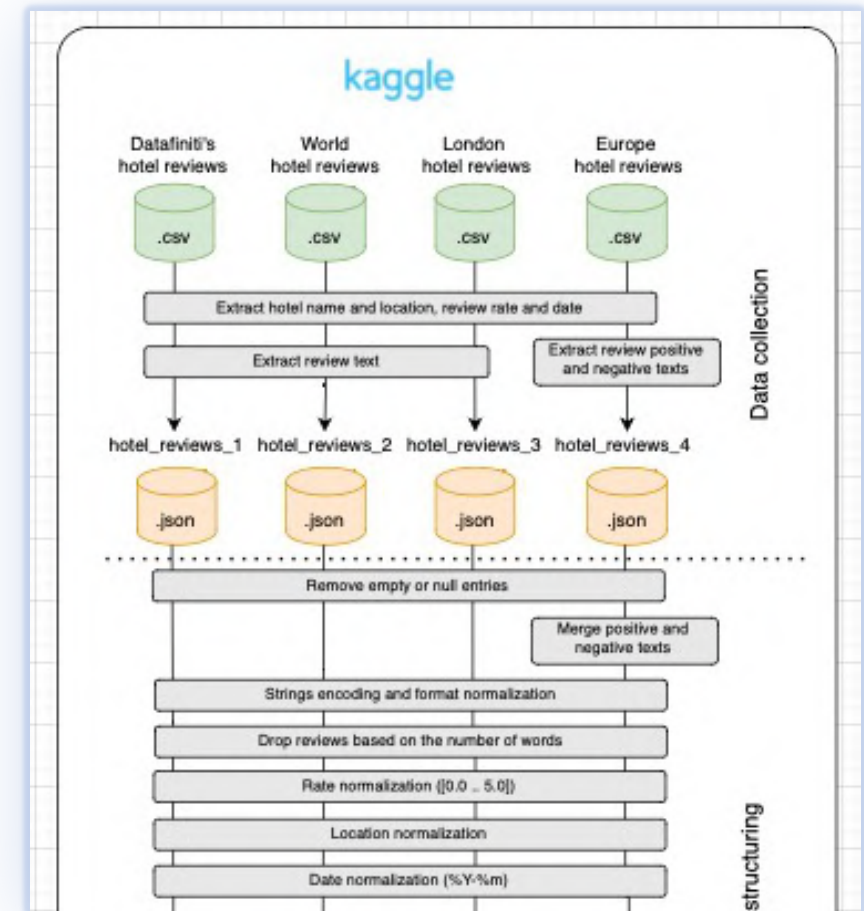
Conclusion on the topic

Data Extraction and Preparation



Data Extraction and Preparation

Hotels Names Normalization

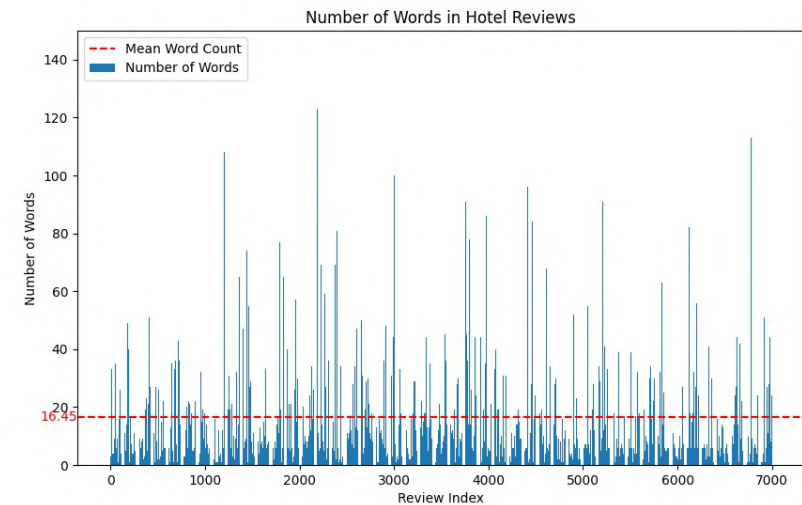
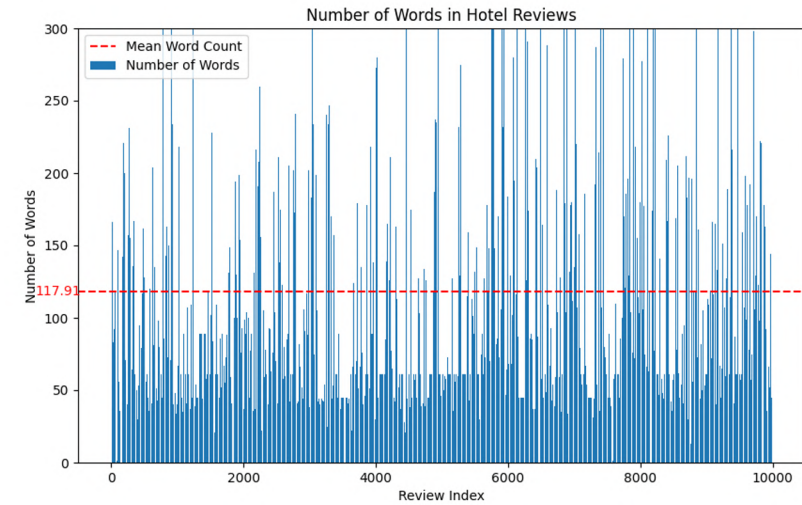


"45 Park Lane - Dorchester Collection"

"45 Park Lane Dorchester Collection"

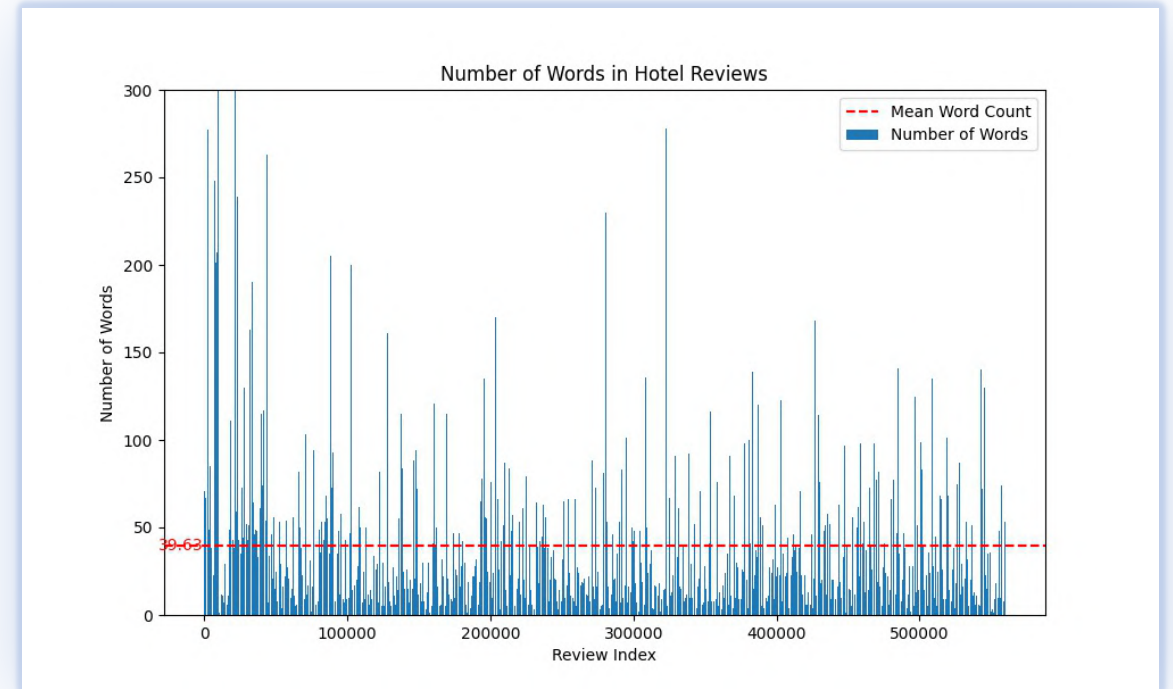
Data Extraction and Preparation

Number of Words per Review
per Dataset



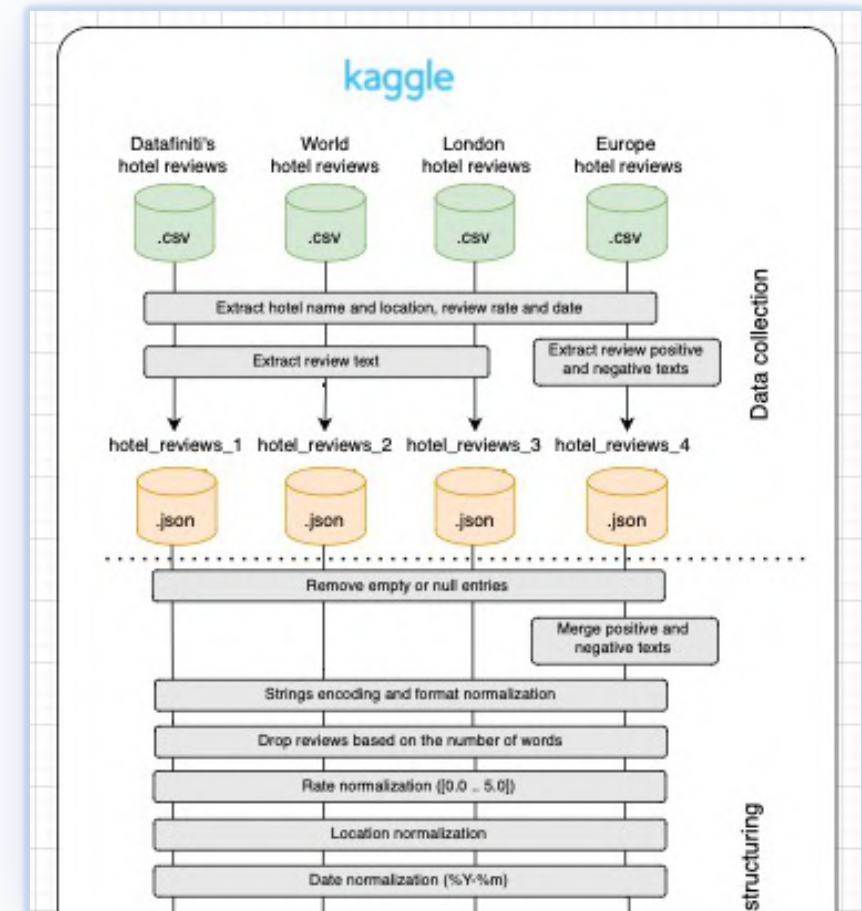
Data Extraction and Preparation

Number of Words per Review
per Dataset



Data Extraction and Preparation

Dates and Rates
Normalization

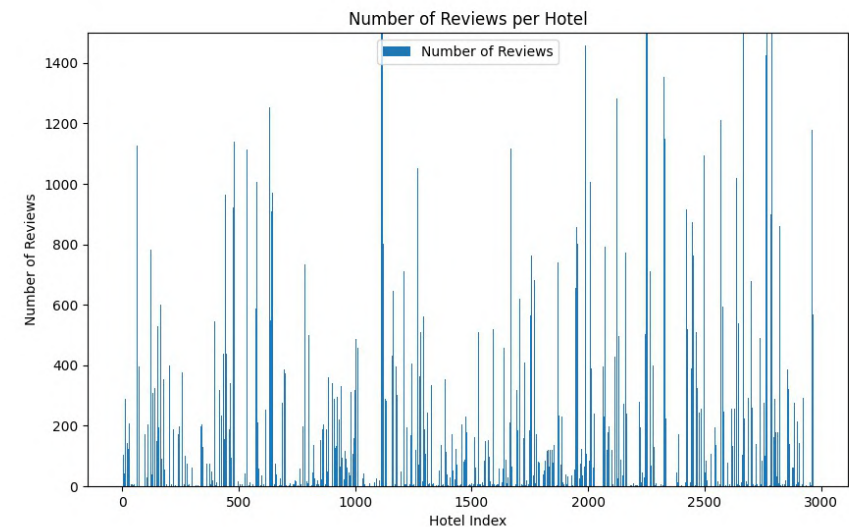
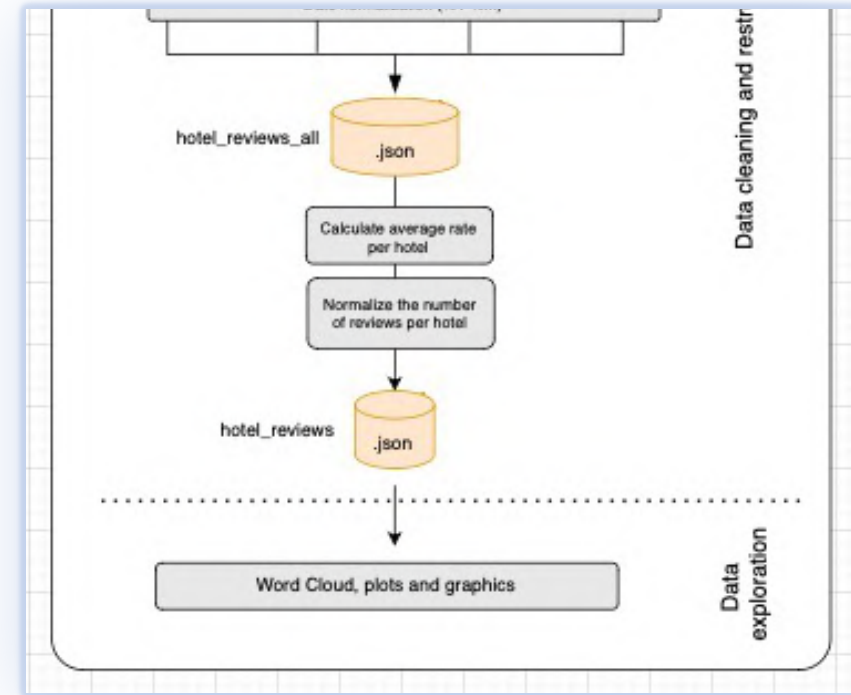


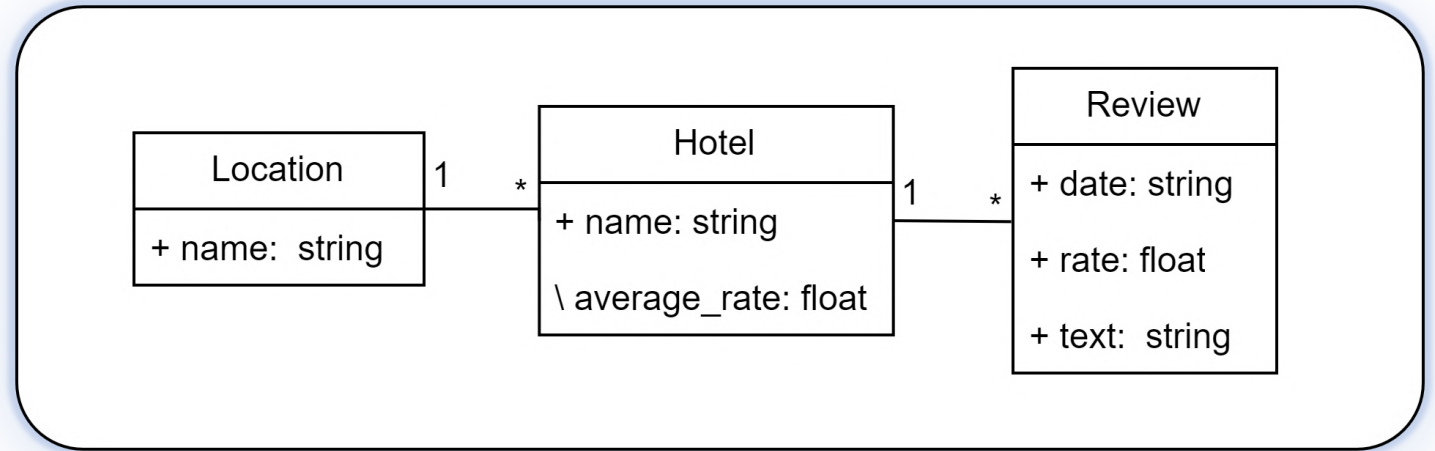
Review_Date ▾
Jul-23
Aug-23

Date Of Review ▾
10/20/2012
3/23/2016

Data Extraction and Preparation

Reviews per hotel

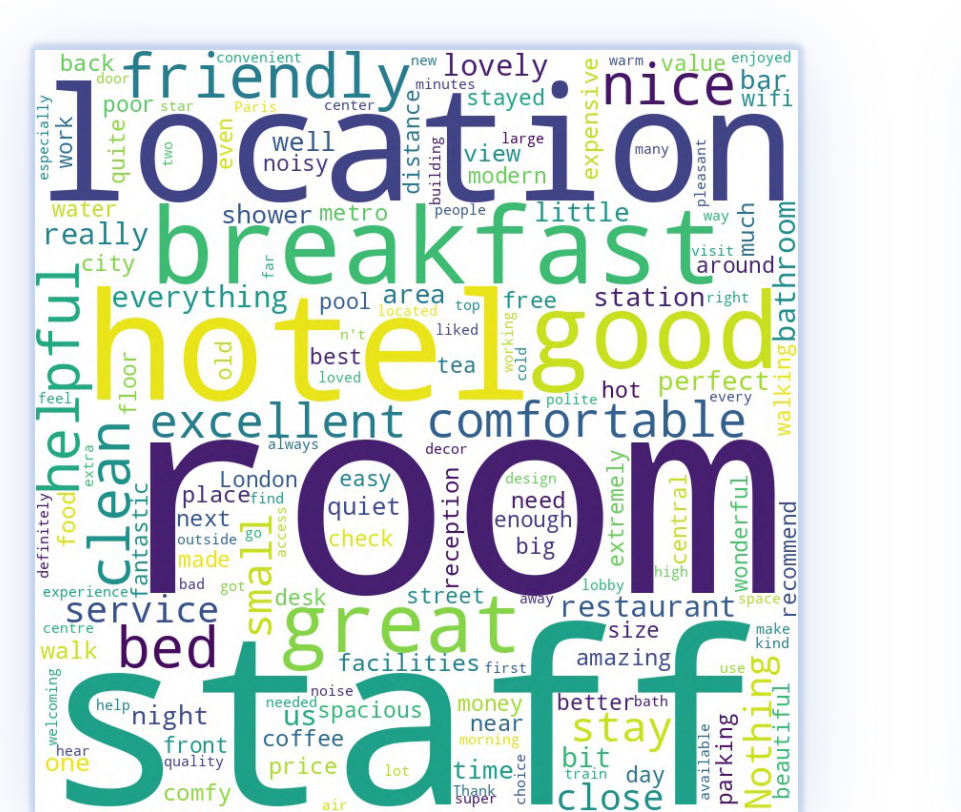




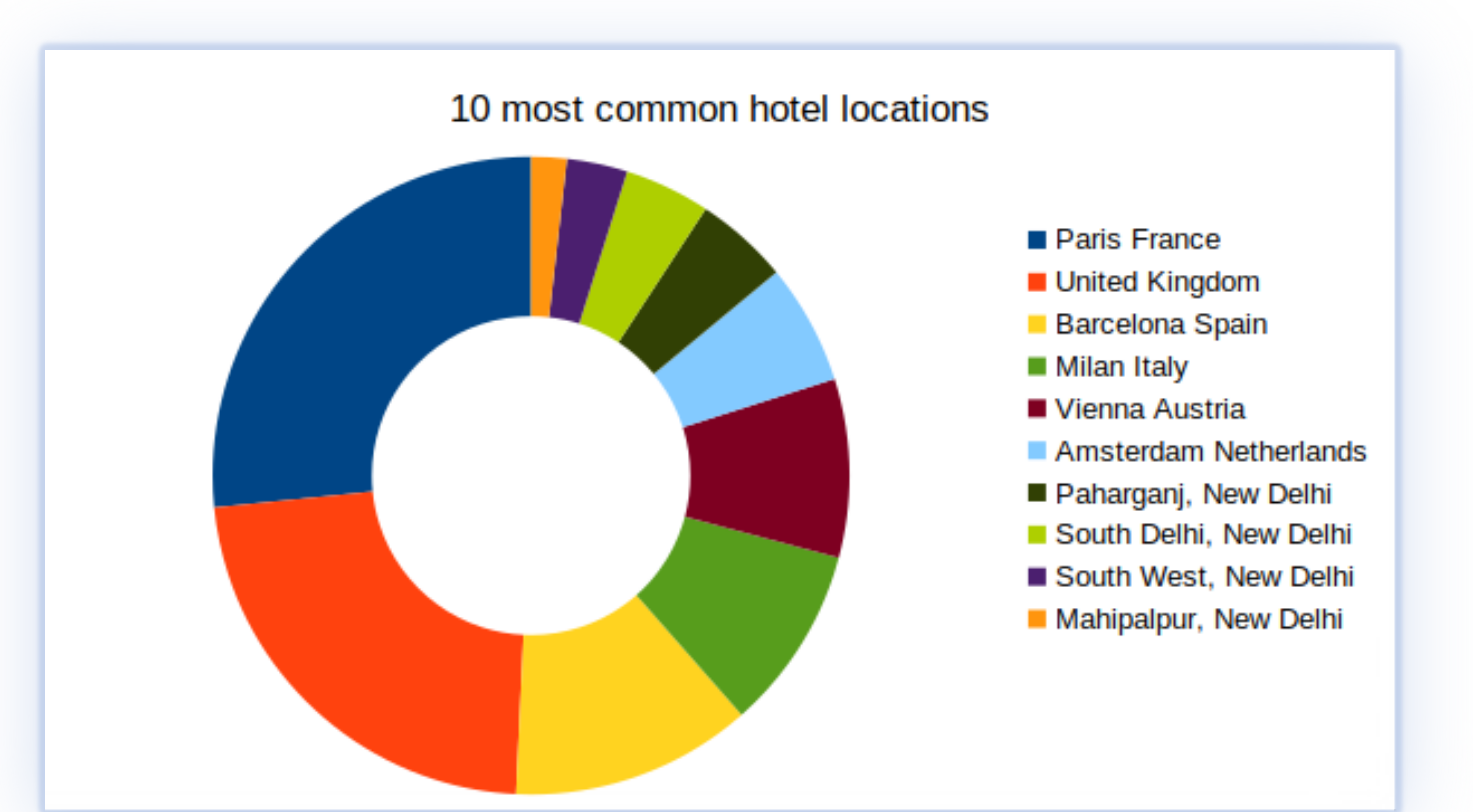
Data Domain Conceptual Model

Data Characterization

Word Cloud

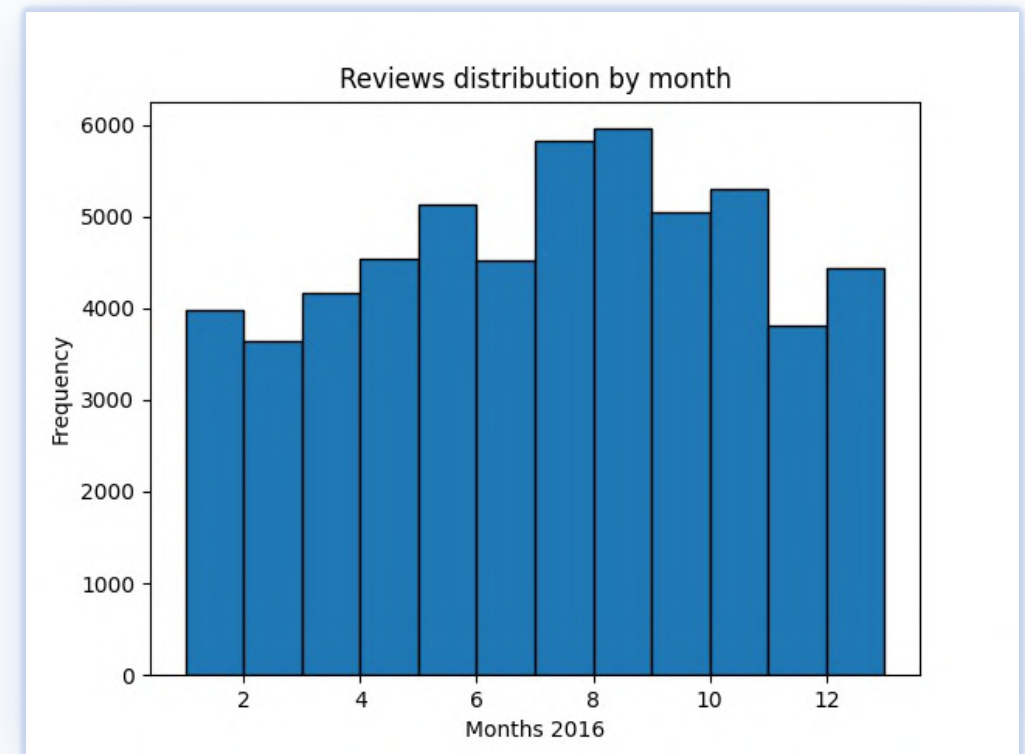
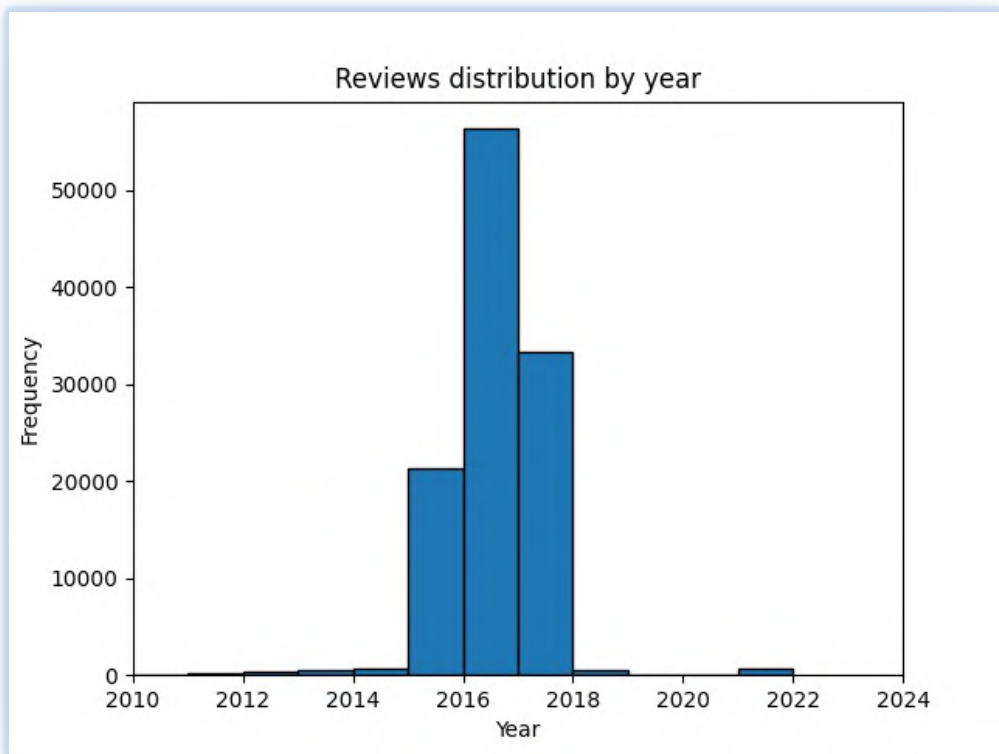


Location Distribution



Data Characterization

Date Distribution:

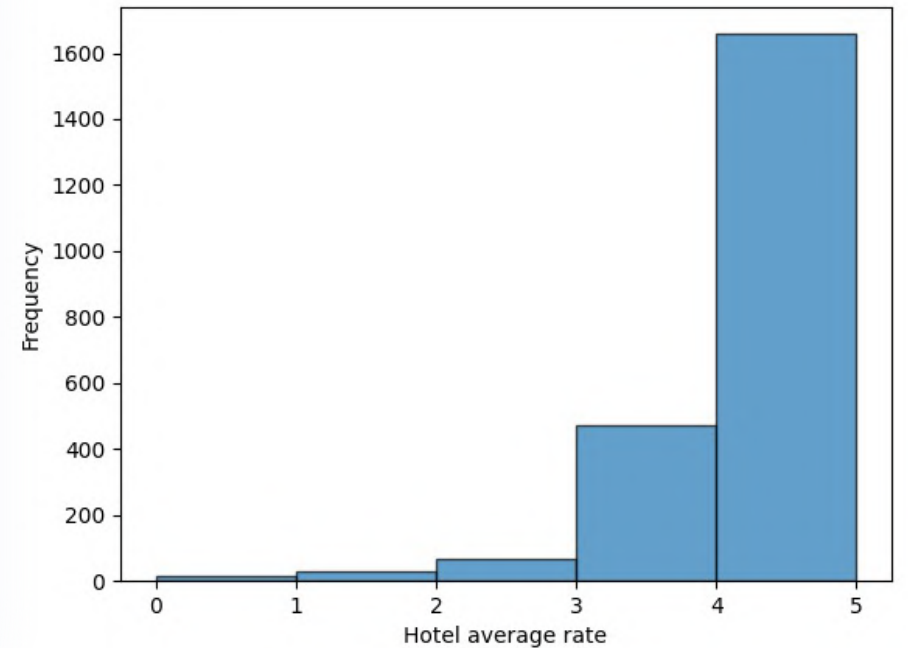


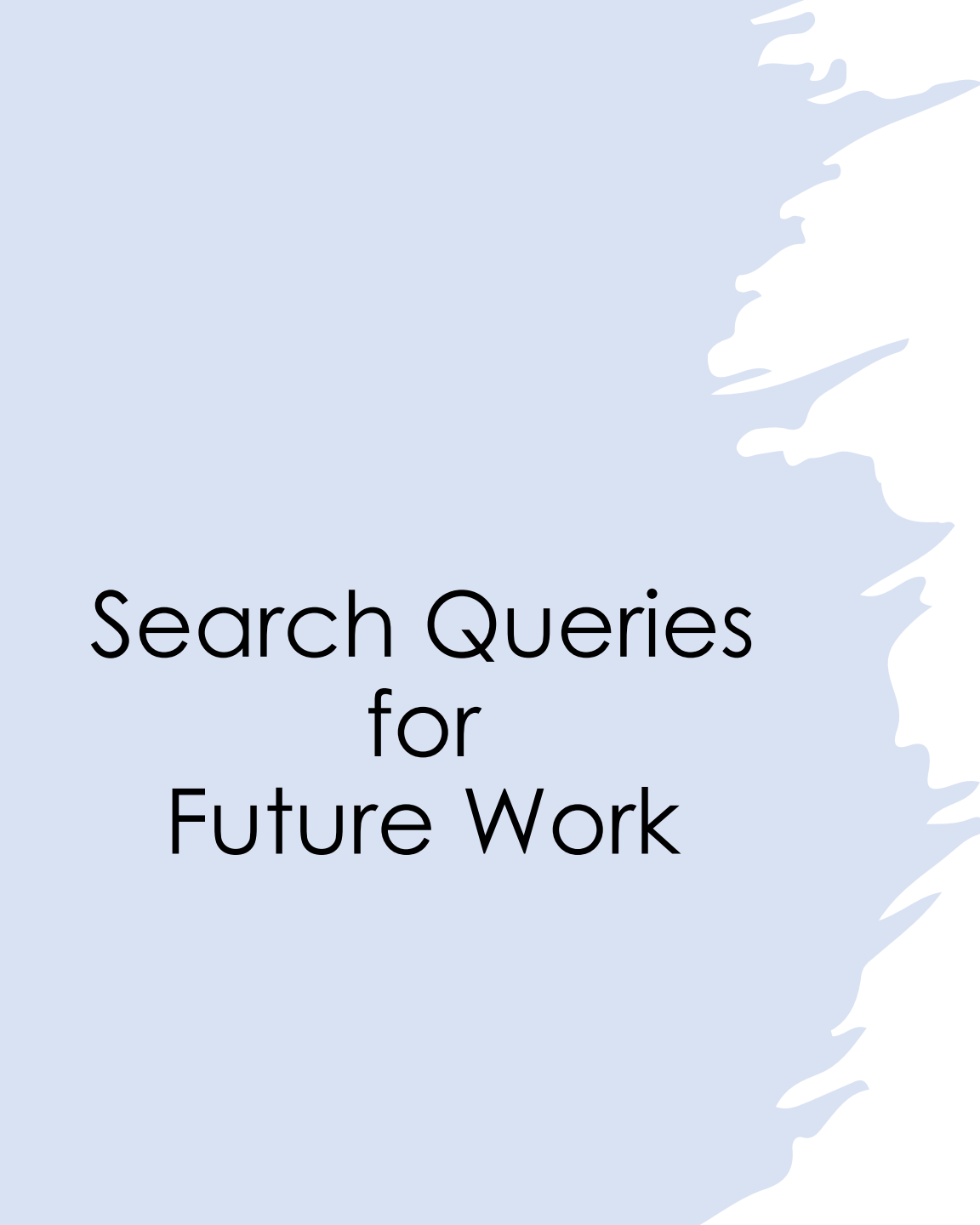
Data Characterization

Document Structure:

```
[
  {
    "name": "11 Cadogan Gardens",
    "location": "United Kingdom",
    "average_rate": 4.34,
    "reviews": [
      {
        "date": "2017-07",
        "rate": 4.8,
        "text": "Lovely hotel. I thought I had booked the ref
      },
      {
        "date": "2017-07",
        "rate": 5.0,
        "text": "Customer service was above and beyond from a
      }
    ]
  }
]
```

Hotels Rate Distribution:





Search Queries for Future Work

- Topics we found relevant within the data characterization phase:
 - Location:
 - Best hotels in [City/Region/Country].
 - Near the airport or landmark.
 - Breakfast:
 - Hotels with good breakfast.
 - Room service:
 - Affordable room service.
 - Staff:
 - Hotel with a helpful staff.
 - Room quality:
 - Comfortable bed and clean bathroom.

Conclusion: Milestone Achievements



We have successfully accomplished all tasks set for this milestone.



The most challenging aspect of our project was developing effective strategies to handle the high volume of reviews without impacting the dataset.

References

-
- [1] - [Kaggle](<https://www.kaggle.com>)
 - [2] - [Datafiniti's Hotel Reviews](<https://www.kaggle.com/datasets/datafiniti/hotel-reviews>)
 - [3] - [Datafiniti's Business Database](<https://www.datafiniti.co>)
 - [4] - [Hotel Review Insights](<https://www.kaggle.com/datasets/juhibhojani/hotel-reviews>)
 - [5] - [London Hotel Reviews](<https://www.kaggle.com/datasets/PromptCloudHQ/reviews-of-londonbased-hotels>)
 - [6] - [DataStock](<https://datastock.shop>)
 - [7] - [Europe Hotel Reviews](<https://www.kaggle.com/datasets/jiashenliu/515k-hotel-reviews-data-in-europe>)
 - [8] - [Booking](<https://www.booking.com>)
 - [9] - [Pandas](<https://pandas.pydata.org>)
 - [10] - [Matplotlib](<https://matplotlib.org>)
 - [11] - [Numpy](<https://numpy.org>)
 - [12] - [NLTK](<https://www.nltk.org>)