HOTEL REVIEWS: A LOOK AT EUROPE'S LUXURY HOTELS

Farahin Choudhury

Clayton Monroe

Objectives

Data analysis methods are as follows:

- Exploratory Data Analysis
- Visualizations (with ggplot)
- Clustering Analysis
- Association rule mining
- Text mining

Business Case

We run a website that offers travel advice and we do extensive research on various destinations across Europe. We provide a wealth of information about popular destinations and part of the maintaining the business involves sharing reviews from booking.com and compiling ratings to help people decide where to travel to.

We want visitors to our site to be able to trust our expertise and we want to direct them in an informed manner. We will have snippets of the reviews of various hotels but we will also connect the booking.com reviews on our site so that folks can read them for themselves and rest assured the reviews are real.

Overview of the data

1492 different hotels

There are reviews from 1,492 different hotels across the continent. Many properties appear to have similar names but they each have reviews from individuals from around the world.

515,738 total reviews

There are reviews of varying lengths (all in English). The data set includes both positive and negative reviews, along with corresponding average score for each hotel (each record is an individual review).

Hardcoded coordinates

Not only were the hotel addresses provided, the latitude and longitude coordinates were also provided in the data set for each hotel.

Data preparation and exploration methods

Counting values

For both character and integer data

The number of words in each review were already counted, however the data has a "Tags" column which provides further insight into the ratings and could help with identifying patterns

Identifying missing values

With over a half million observations, it is crucial to find variables that don't have any data

Frequency of data

Many questions to explore

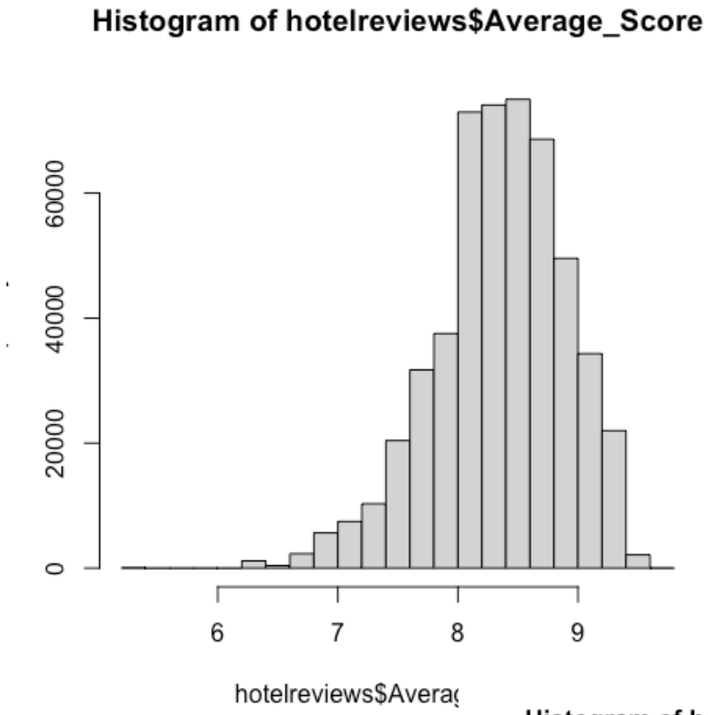
Where are the most popular or least popular hotels? Which hotel is rated the highest? How many words does the review associated with the lowest rating have?

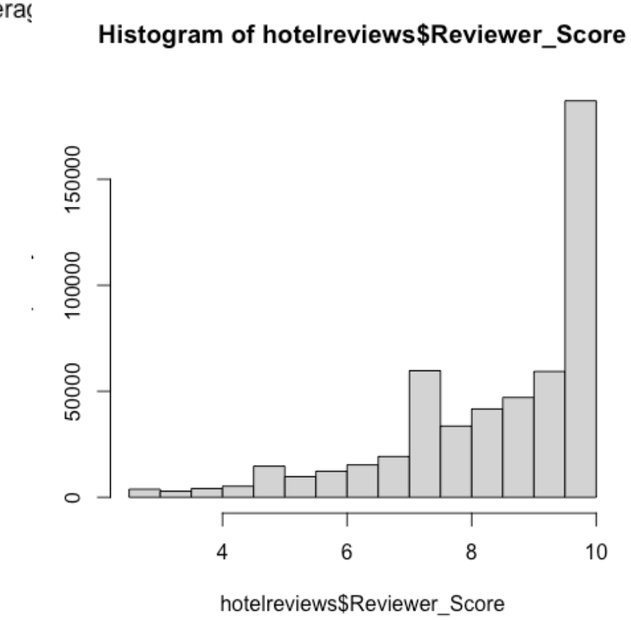
Average ratings

for all hotels

The histogram shows that the data is skewed to the right, which corresponds to higher ratings among the reviewers.

According to this histogram, the data shows the most observations for ratings of between 8 and 9 (out of 10).





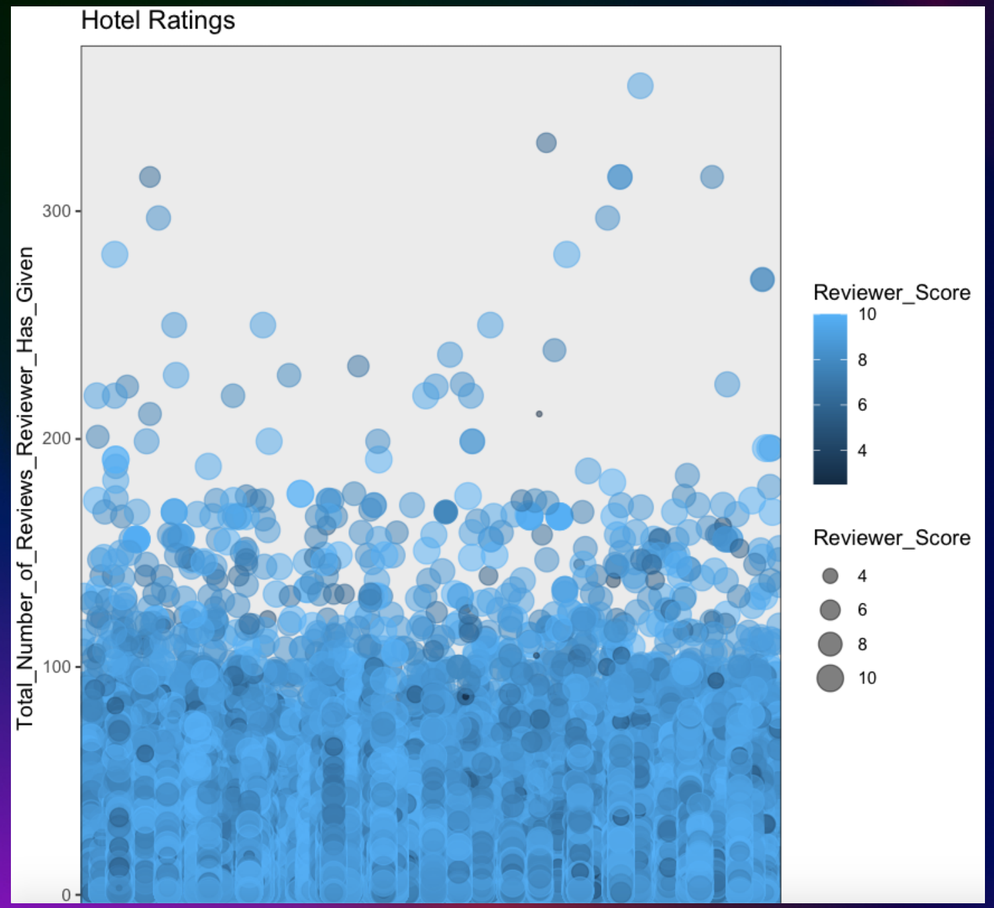
w	lat	lng	Business_Travel	Leisure_Travel	Solo_Trip	Couples_Trip	Family_Trip	Group_Trip	With_Pet
	52.36058	4.915968	FALSE	TRUE	FALSE	TRUE	FALSE	FALSE	FALSE
	52.36058	4.915968	FALSE	TRUE	FALSE	TRUE	FALSE	FALSE	FALSE
	52.36058	4.915968	FALSE	TRUE	FALSE	FALSE	TRUE	FALSE	FALSE
	52.36058	4.915968	FALSE	TRUE	TRUE	FALSE	FALSE	FALSE	FALSE
	52.36058	4.915968	FALSE	TRUE	FALSE	TRUE	FALSE	FALSE	FALSE
	52.36058	4.915968	FALSE	TRUE	FALSE	FALSE	FALSE	TRUE	FALSE
	52.36058	4.915968	FALSE	TRUE	FALSE	FALSE	FALSE	TRUE	FALSE
	52.36058	4.915968	FALSE	TRUE	FALSE	TRUE	FALSE	FALSE	FALSE
	52.36058	4.915968	FALSE	TRUE	FALSE	TRUE	FALSE	FALSE	FALSE
	52.36058	4.915968	FALSE	TRUE	FALSE	TRUE	FALSE	FALSE	FALSE
	52.36058	4.915968	FALSE	TRUE	FALSE	FALSE	FALSE	TRUE	FALSE
	52.36058	4.915968	TRUE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE
	52.36058	4.915968	FALSE	TRUE	TRUE	FALSE	FALSE	FALSE	FALSE
	52.36058	4.915968	TRUE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE
	52.36058	4.915968	FALSE	TRUE	FALSE	FALSE	TRUE	FALSE	FALSE
	52.36058	4.915968	FALSE	TRUE	FALSE	TRUE	FALSE	FALSE	FALSE
	52.36058	4.915968	FALSE	TRUE	FALSE	FALSE	TRUE	FALSE	FALSE
	52.36058	4.915968	FALSE	TRUE	FALSE	TRUE	FALSE	FALSE	FALSE
	52.36058	4.915968	TRUE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE
	52.36058	4.915968	FALSE	TRUE	FALSE	TRUE	FALSE	FALSE	FALSE
	52.36058	4.915968	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE
	52.36058	4.915968	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE
	52.36058	4.915968	FALSE	TRUE	FALSE	FALSE	TRUE	FALSE	FALSE
	52.36058	4.915968	FALSE	TRUE	FALSE	TRUE	FALSE	FALSE	FALSE
	52.36058	4.915968	FALSE	TRUE	FALSE	TRUE	FALSE	FALSE	FALSE
	52.36058	4.915968	FALSE	TRUE	FALSE	TRUE	FALSE	FALSE	FALSE
	52.36058	4.915968	FALSE	TRUE	FALSE	FALSE	FALSE	TRUE	FALSE
	52.36058	4.915968	TRUE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE
	52.36058	4.915968	FALSE	TRUE	FALSE	TRUE	FALSE	FALSE	FALSE

Data manipulation

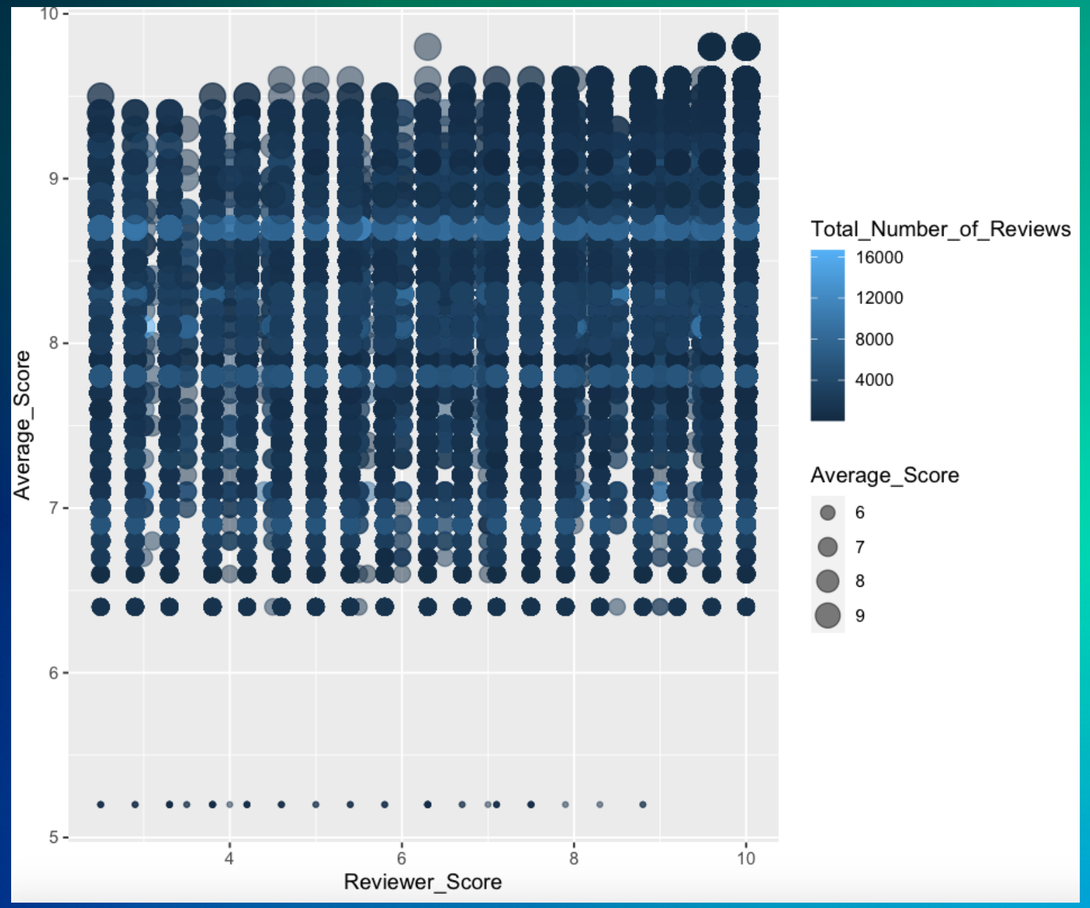
Taken from the "Tags" column in the data set

The tags/keywords associated with the reviews have been assigned values of either true or false. This is a good way to categorize ratings and to see patterns in positive/negative reviews

ggplot viz



rating vs # of reviews by reviewer



rating vs avg. score per hotel

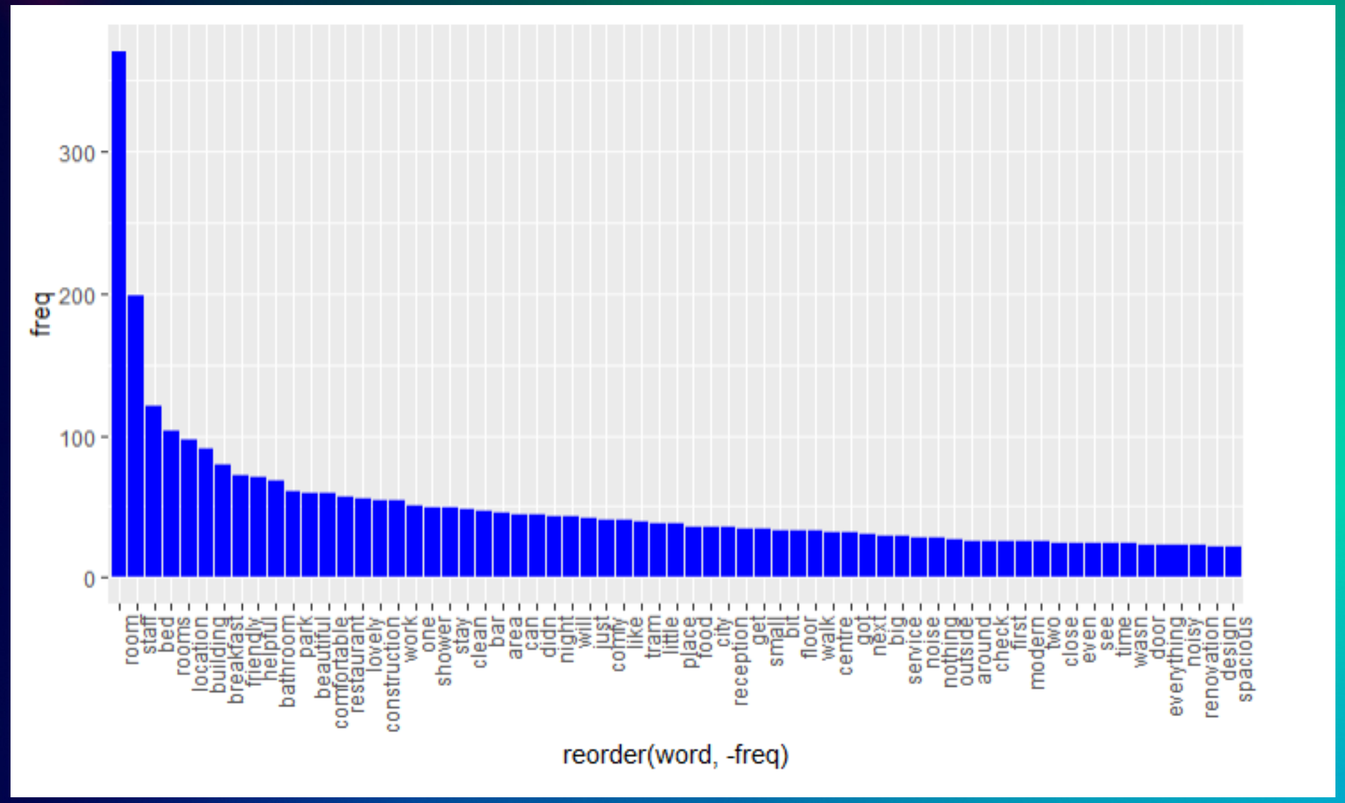
Text mining

This dataset has text from actual reviews from users of booking.com

A goal for this analysis with regards to text mining and text analytics was to find frequently occurring words for both positive and negative reviews

Clustering analysis

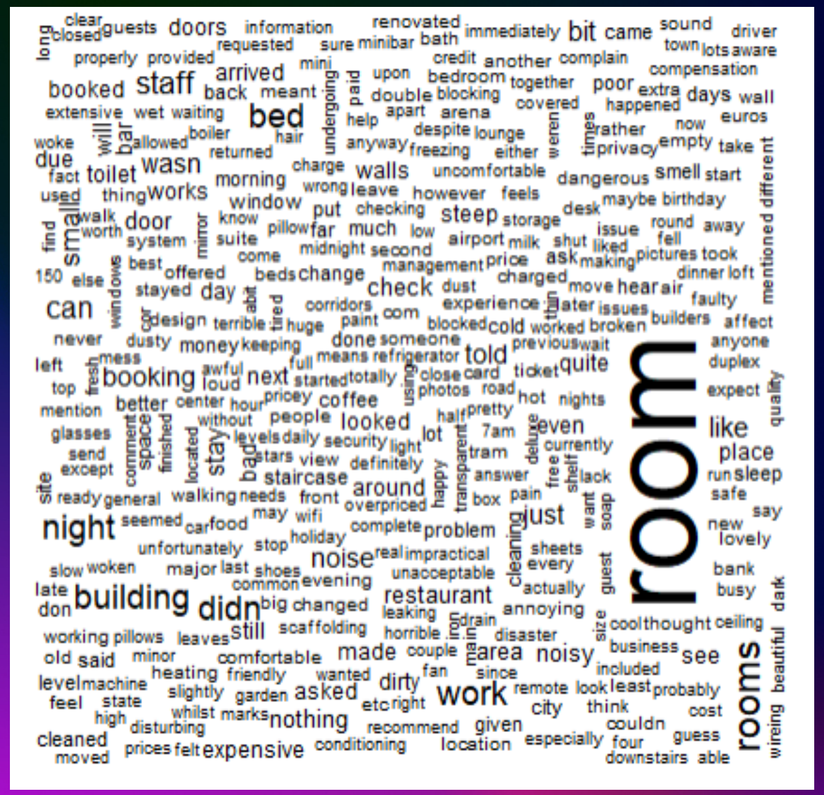
chart showing the frequency of words that appear in reviews for **Hotel Arena**



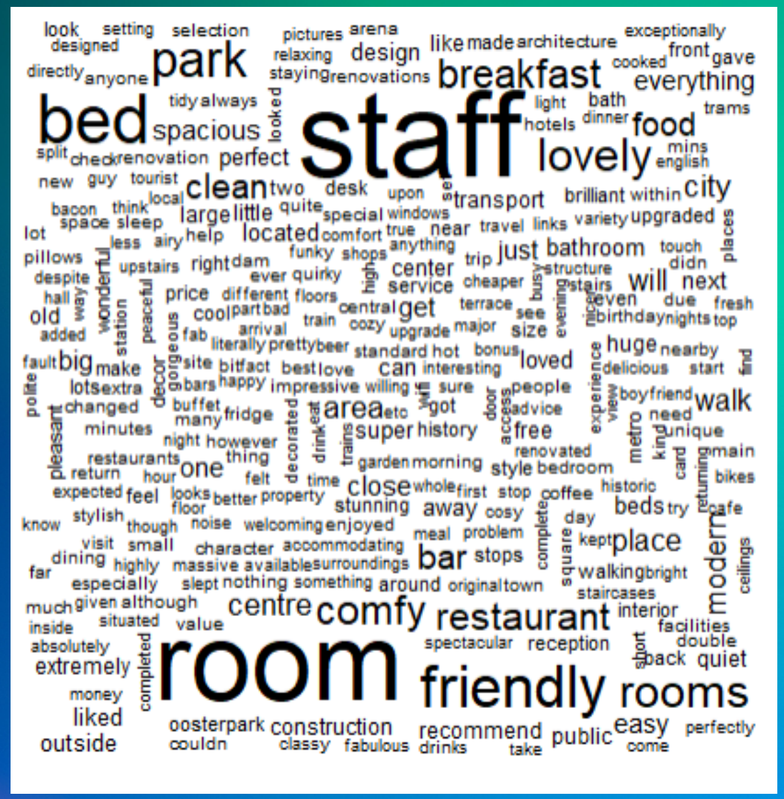
Wordclouds

for a specific hotel

Negative



Positive



Challenges and setbacks

1.

Unnecessary data

There were attributes not relevant to the analysis

2.

"Too much" data

There were over half a million observations, and there seemed to be little differentiation between each record

3.

Text mining

With unstructured text as individual observations, it makes the analysis tedious and harder to classify or categorize into positive or negative

4.

Missing data

Even after removing all "NA"s from the dataset, some code didn't seem to work (svm model, length(which.max())

Next steps/how to take this analysis further

More data exploration

While there is a balance of text and numerical data, it requires further analysis beyond identifying vectors of data. With such a large dataset, the exploration doesn't quite end

API for map data

The coordinates in the dataset were hardcoded but there might need to be an API key in order to plot a map of the continent of Europe. There are packages that exist that allow for US map data to be plotted (which is different).

NLP

Natural language processing: while this topic falls outside the scope of the objectives and analysis for this project, there might be more interesting insights with the use of NLP

Thank you!