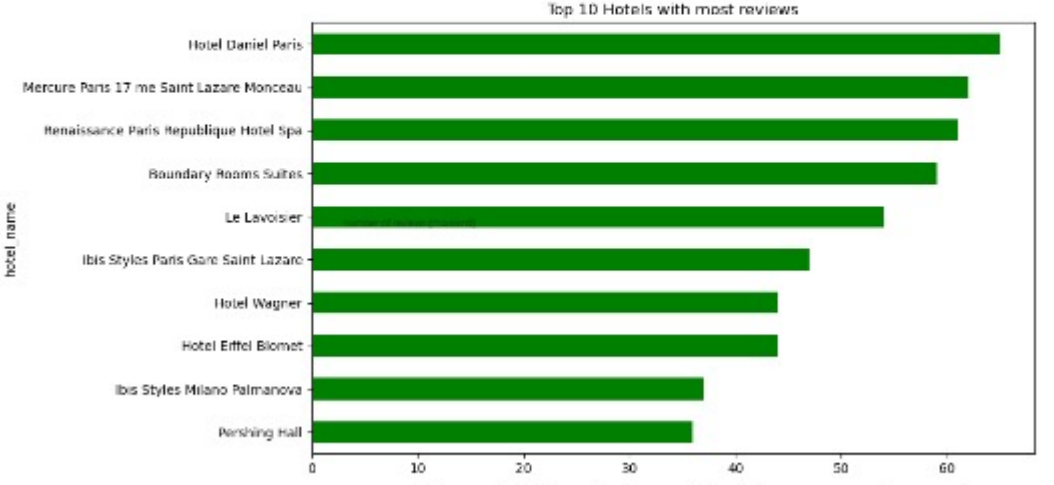
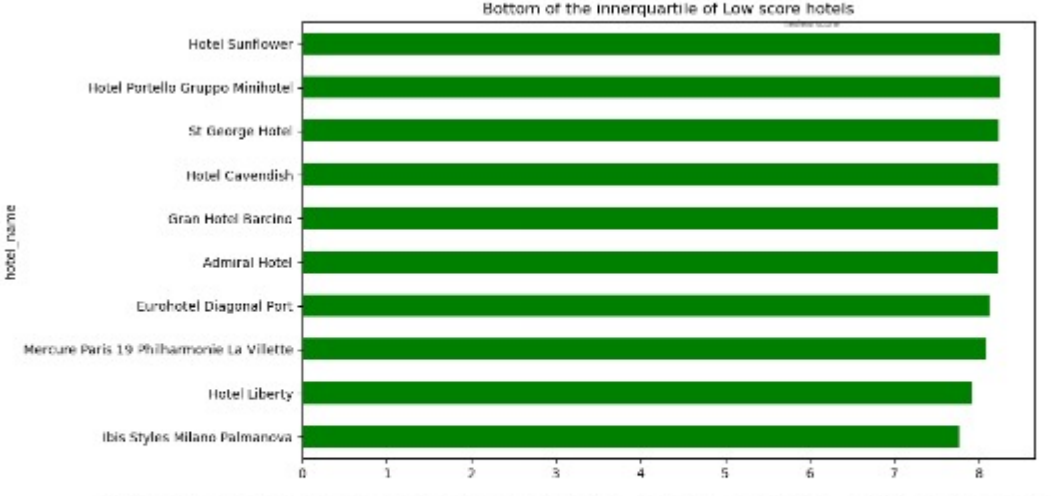
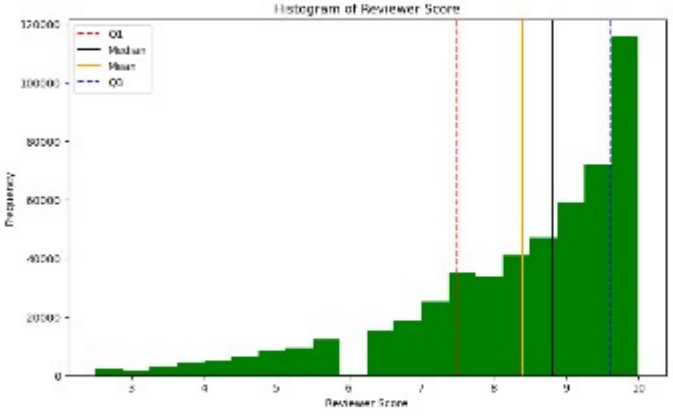
Part 1. Analysis



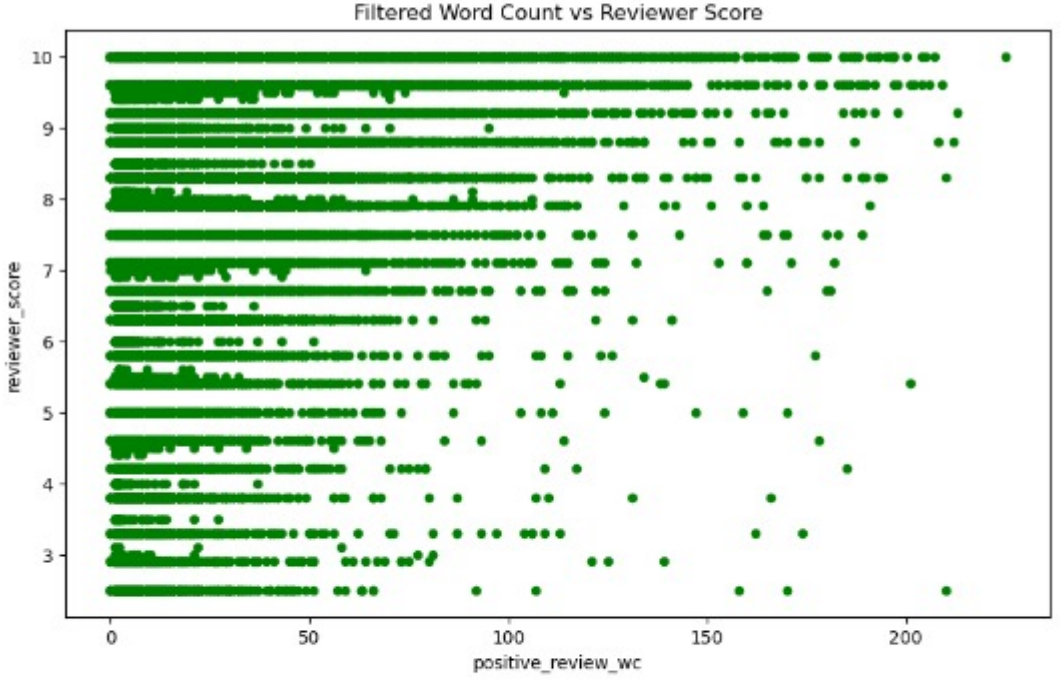
Top 10 hotels with the most reviews.



The least reliable hotels from the bottom of the inner quartile.



Upper and lower quartile have no threshold
The least reliable hotels are counted from the red line with a score of 8 with an average score between 8.0 and 9.0.



More words are said for hotels with high review scores. The chart displays higher density of plots in the upper ranges of each axis.

Part 2. Prediction

Random forest achieving the highest accuracy with a training time more than 5 times longer then logistic regression training. Predicting a score equal or greater than 9.

Data Engineering:

- Considering that the place in the world and hotel names do not make a difference influencing a higher review score, lng, lat and hotel name is removed as they along with address holds the same value throughout the dataset.
- Reviewer word count is added.
- Review Sentiment added.
- Review Topic added
- Accommodation type

Best Models

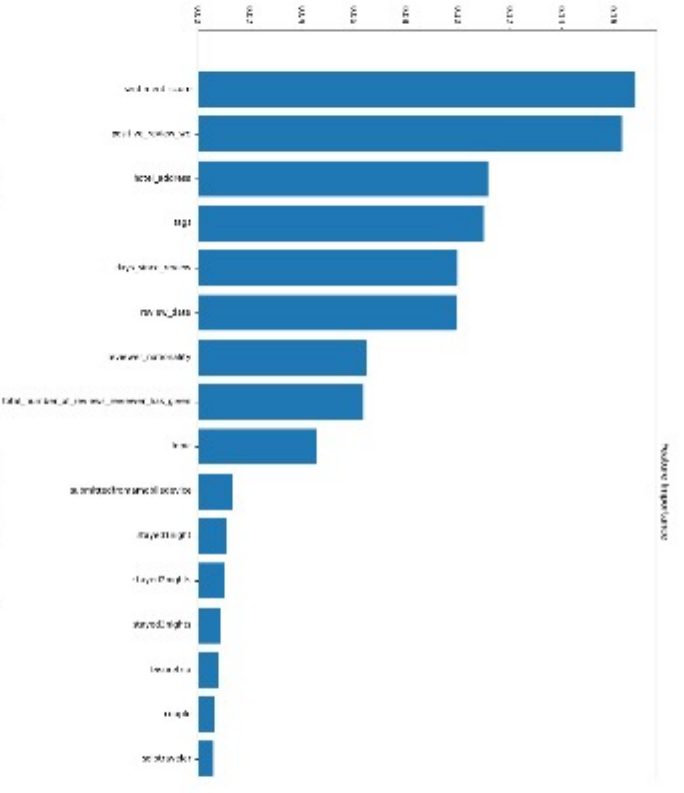
1: XGBOOST

Accuracy: 0.667
Precision: 0.652
Recall: 0.648
F1 Score: 0.65

2: Random Forest

Accuracy: 0.629
Precision: 0.616
Recall: 0.598
F1 Score: 0.607

The feature importance graph shows the strongest features



Optimization also included model parameter grid search, feature removing and adding features to gain a up to 10% accuracy increase from baseline to best model. The best features are review related. Assuming that the weight of the review depends on the type of review given more data was extracted from the sentences.