BERLIN AIRBNB DATA ANALYSIS

Instructor : Prof. Daniel Acuna

Team Members: Hanlin Zhang, Isha Havaldar, Vidisha Badhe, Wei Mu

1. Problem and Objective

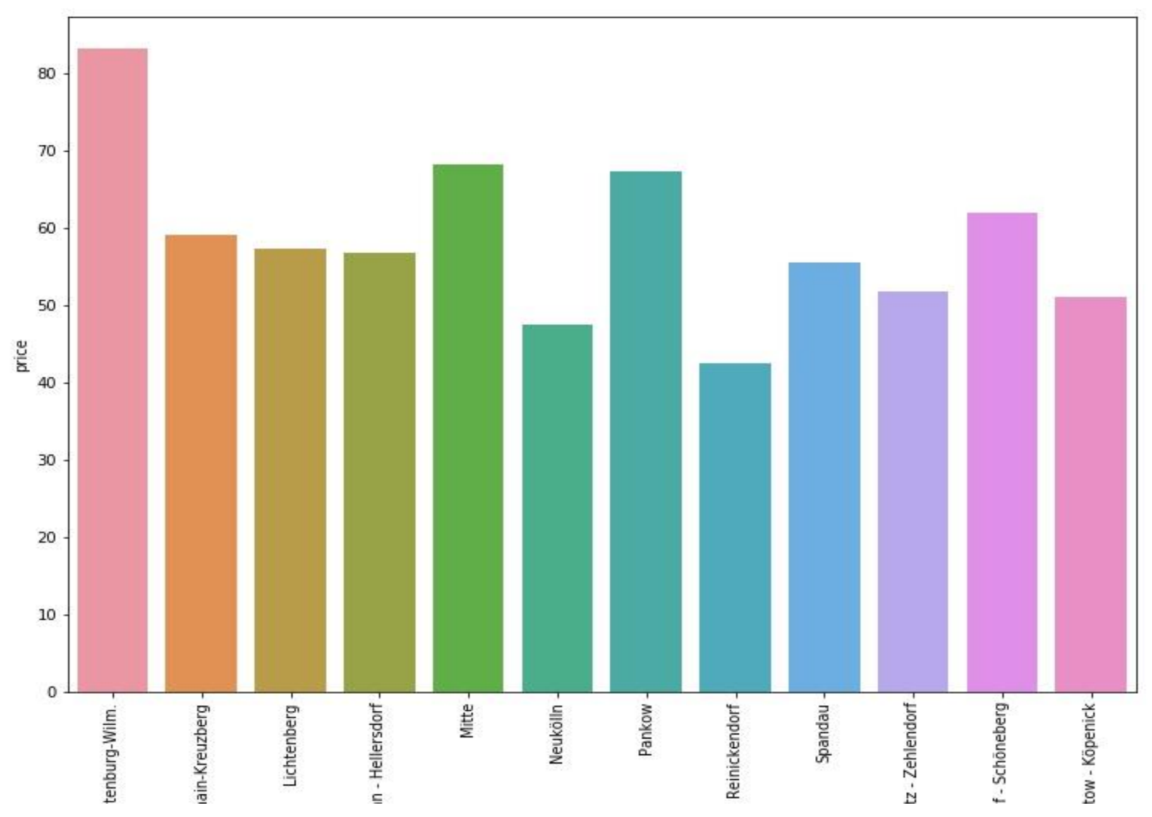
We are looking forward to identifying the problems and features of Airbnb service in Berlin. With this analysis, we could provide recommendations and suggestions to Airbnb to improve their service and provide a guideline for tourists. Our main objective is to gain insights and solutions with machine learning algorithms and models that will help us understand the factors which cause these problems and affect the Airbnb listings.

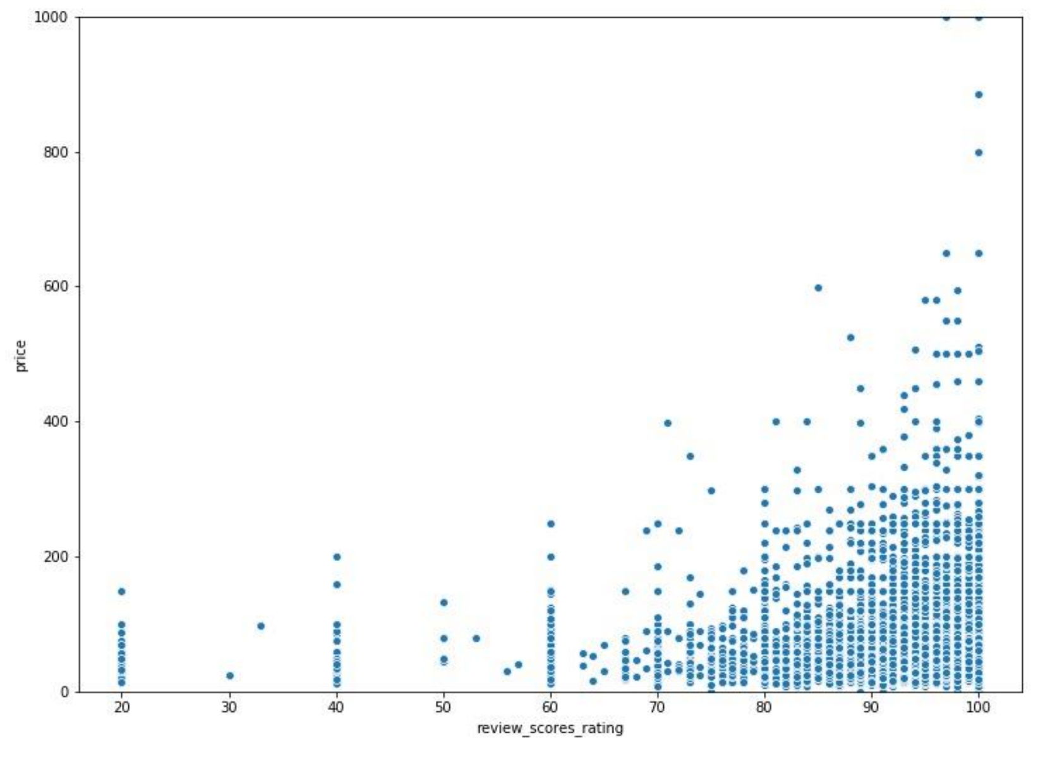
2. Goals

- To develop an analysis for customer ratings based on different features highlighting the most significant feature using PCA
- To see the trends of tourism in Berlin by implementing NLP models on customer reviews
- Building a regression model to classify the listings based on the rating score

3. Data Description

The dataset we chose is the Berlin Airbnb Dataset from Kaggle. It has 6 files which contain data about the various aspects of the Airbnb in Berlin, that is the Reviews, Neighborhood's, Listings, etc. The first file is the calendar summary which has 8.23 million rows and 4 columns. The next dataset is of the listings with 22.6k rows and 16 columns. The next dataset is of listings summary which has 22.6k rows and 96 columns. Further, the neighborhoods dataset has 139 rows and 2 columns, the reviews dataset has 402k rows and 2 columns and the last dataset i.e. the reviews summary dataset has 402k rows and 6 columns. The dataset will be used for drawing out insights and conclusions for the various issues the customers are facing based on the reviews, their tourism trends and patterns, etc.





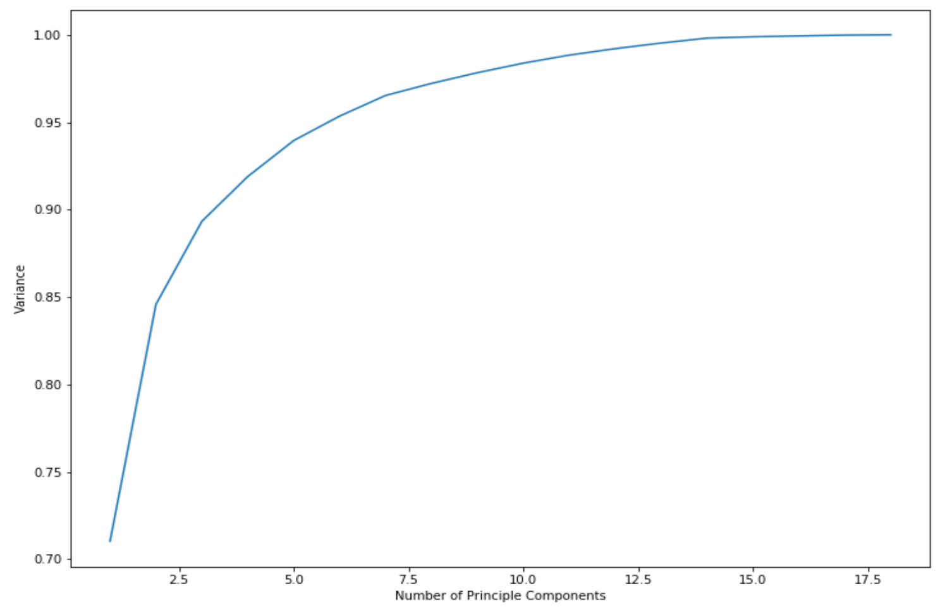
4. Model Description

Model	Features	Description	Evaluation
PCA	Availability, accommodation, bedrooms, review ratings	Recognizes most significant features in the dataset	N/A
Logistic Regression	Availability, accommodation, bedrooms, review ratings	To predict whether a listing is liked or disliked by the customers	AUC
NLP	Comments	Analysing the positive & Negative words to understand customer's reviews	AUC

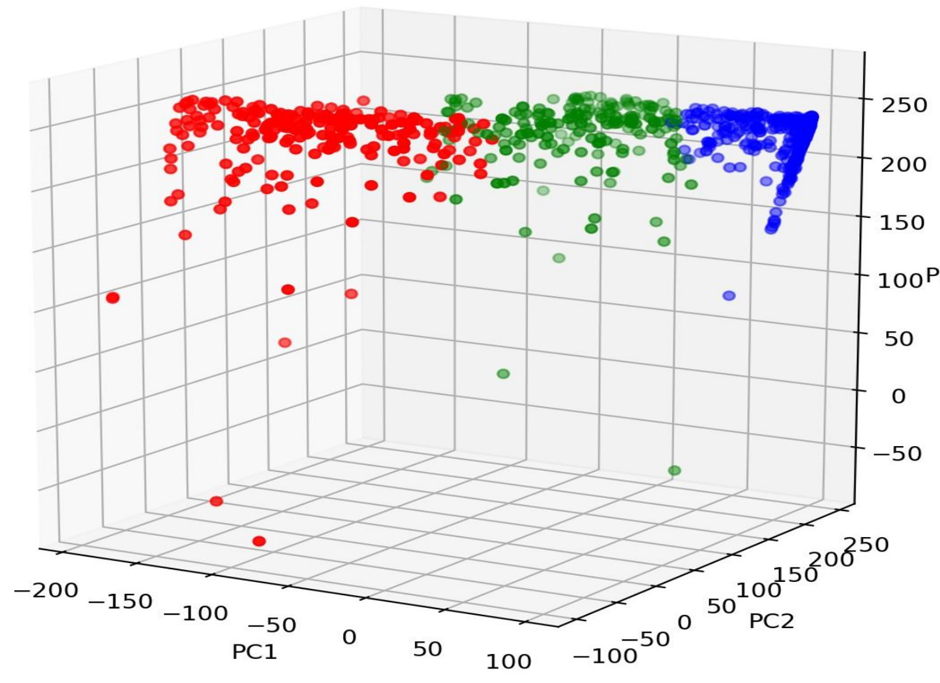
5. Model Results

PCA Evaluation

Elbow Graph



3D Graph for PCA



NLP Evaluation

English positive words

	word	weight
37231	well-ordered	0.190971
41306	unkomplizierte	0.183480
8243	deserved	0.152214
29571	matresses	0.149989
39602	6.30	0.146091
65255	kevins	0.128514
16134	vicky	0.124610
22640	center...	0.124608
46285	(~40	0.120748
103084	living-bedroom.	0.115547
34750	accomadate	0.113576

German positive words

	word	English	weight
61754	getummelt.	romped	0.249130
26354	holzfussboden,	Parquet floors	0.204283
38621	luxuriöses,	luxurious	0.204188
56505	nackten	naked	0.198136
18187	angemessenen	reasonable	0.195098
4739	moderne,	modern	0.193277
3630	hochwertige	quality	0.187295
48801	seminarwochenendes	seminar weekend	0.178860
20897	o.g.	above-mentioned	0.170245
26409	zögern,	hesitate	0.166438

Logistic Regression Evaluation

```
+------------------+    +---------------------+
|    en_lr_accuracy|    |en_lr_accuracy_testing|
+------------------+    +---------------------+
|0.6615898959881129|    |   0.6445974576271186|
+------------------+    +---------------------+
```

6. Performance and Interpretation

feature	PC1	PC2
availability_30	-0.470924	0.398138
availability_60	-0.500584	0.299259
availability_90	-0.548309	0.129563
availability_365	-0.471319	-0.849481

feature	PC3
accommodates	-0.728609
beds	-0.484469
guests_included	-0.273863
bedrooms	-0.287143

feature	PC4
availability_30	-0.693834
availability_90	0.676883

feature	PC5
review_scores_rating	0.989148

From the PCA, we got 5 Principle Components which give out the most significant features. Availability is the feature which comes in 3 PCA's and affects the dataset the most. After that, accommodation, number of bedrooms and guests included are the second most significant features followed by review scores.

For evaluating the performance of the NLP, we found out the Area under curve which is ~0.76 and tells us that it is 76% accurate in classifying the words. `Model AUC: 0.7537204899064261`

The accuracy scores (64.46%) of the logistic regression shows the effectiveness of our model in predicting the review rating score based on the numerical features.

7. Conclusion

After performing analysis on the Berlin dataset by using various models like NLP and PCA Analysis, we concluded that the most significant feature as per the PCA analysis is the availability of the room followed by accommodations, the number of bedrooms and bathrooms which is then followed by the review score. So, the idea here is that for people, availability of the room is the highest priority which is obvious as they will do further analysis of the room only after they discover whether that room is available or not. NLP shows that English and German customers are concerned about accommodations and the overall sentiment about listings is positive. Many keywords for preferred listing are related to facilities. Also, the AUC is 76.35% which defines the accuracy of our NLP model.

Data Source : https://www.kaggle.com/brittabettendorf/berlin-airbnb-data