A study of Amsterdam Airbnb rentals

1. Business Problem

Amsterdam is the capital and most populous city of Netherlands. Colloquially referred to as the "Venice of the North", attributed by the large number of canals which form a UNESCO World Heritage Site, it is also home to numerous tourist attractions. As a result, it is one of the most visited places in Europe, receiving more than 4.63 million international visitors annually. This results in a demand for places for a short term stay.

Airbnb, Inc. is an online marketplace for arranging or offering lodging, primarily homestays, or tourism experiences. The company does not own any of the real estate listings, nor does it host events; it acts as a broker, receiving commissions from each booking.

Anyone planning to host a short term vacation rental in Amsterdam should be aware of the most desirable places to put up listings, what type of listings are in demand and also what percentage of occupancy they can expect from the rental. On the other hand, any traveler traveling to the city should also be able to make an informed decision about where to stay based on their budget and interests.

This study will attempt to answer the following questions to benefit Amsterdam's airbnb hosts and travelers:

- What are the factors that drive the listing price?
- How to maximize the revenue earned from a rental?
- What are the factors that drive the popularity of a vacation home?
- How to increase the popularity of a rental?
- When are airbnbs most likely to be available?

2. Data

2.1. Data Collection

The data for this project has been sourced from mainly 2 sources:

- Foursquare API
    1. Venue Categories - Hierarchical list of categories applied to venues. This data has been used to group venues in each neighborhood
    2. Venue Search - List of venues near the current location. All venues have been rolled-up to these categories for analysis - Shop & Service, Food, Outdoors & Recreation, Travel & Transport, Nightlife Spot, Arts & Entertainment, Professional & Other Places
- Publicly available Airbnb listing data that includes:
    1. Information and metrics for listings in Amsterdam
    2. Detailed Calendar Data for listings in Amsterdam
    3. GeoJSON file of neighbourhoods of Amsterdam

Here are the relevant details of the listings dataset:

Name	Description
listing_id	Id of the airbnb listing
neighbourhood	Neighborhood of the listing
latitude/longitude	Coordinates of the listing
room_type	The type of listing. Private room, Shared room etc.
price	Price per night
minimum_nights	Minimum number of nights required to book
number_of_reviews	Number of reviews available for the listing
last_review	Last review date
reviews_per_month	Average number for reviews posted per month
availability_365	Number of days available in the next 365 days

Listing calendar dataset contains:

Name	Description
listing_id	Id of the airbnb listing
date	Date
available	Flag to indicate if available for booking
price	Listing price
adjusted_price	Booking price
minimum_nights	Minimum number of nights required to book
maximum_nights	Maximum number of nights the rental can be booked for

All sources of data and other information have been mentioned in the Reference section.

2.2. Data Cleaning

The following data cleaning measures were taken before analysis:

- The focus of this study will be on short term rentals. So, all rentals requiring more than 6 months of minimum stay have not been considered for the analysis.
- The listing type of 'hotel rooms' were removed as hotels are not considered as a traditional Airbnb hosting. Note that hotel listings only accounts for only 1.3% of the original data set.
- The listing data contained some $0 listings that have been removed as erroneous entries.

In addition to the above steps, some of the missing values were imputed. For example, the rentals with value NA 'reviews_per_month' have been updated to 0. This column is a measure of popularity of a listing. Finally, during data analysis and modeling, outliers have been removed from the relevant columns wherever necessary, so that the analysis results do not get skewed.

3. Methodology

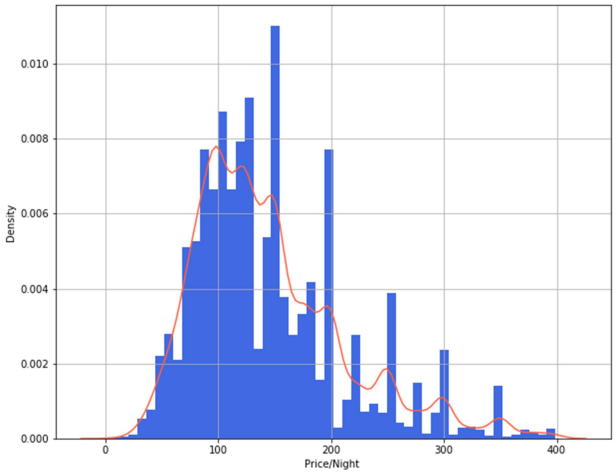
This study will assess the rentals on three aspects - Pricing, Popularity and Availability. The analysis will be conducted by doing data exploration first, to identify and visualize the most important features for each aspect. In the second step, the identified features will be used for clustering and modeling, to answer the questions in the business problem section. Finally, we will compare the results and draw a conclusion based on all the findings.

3.1. Price Analysis

Price is one of the main deciding factors for travelers when they pick where to stay. If a room is overpriced, then it is more likely to remain empty. Lowering the price too much will cause a loss of revenue for the host. So, it is essential for hosts to decide the optimum price when they put up a house or room for rent. In this section, the right conditions to have a high priced rental will be researched.

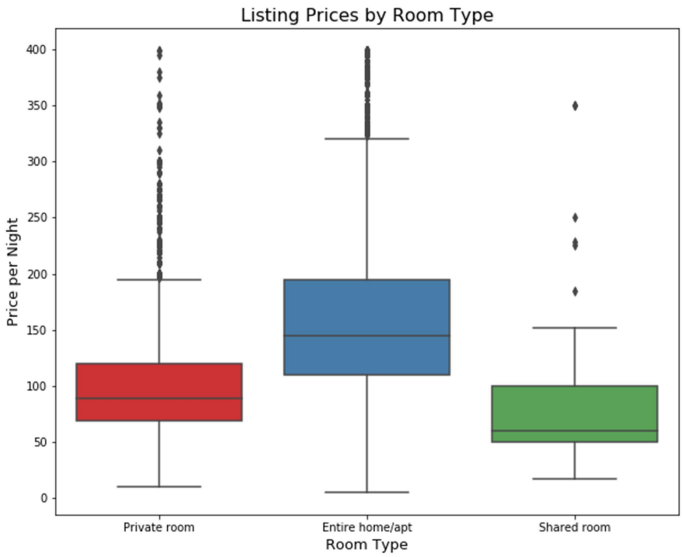
3.1.1. Data Exploration

Before a look at the features, here is a look at the price distribution between the rentals. It was observed that the number of rentals that cost over USD 400 are really few, so these will be considered as outliers. All analysis in the rest of this section will be done with rentals that cost less than USD 400 per night. Here is what the distribution looks like for rental listings under USD 400:



Relation between Room Type and Price

For feature selection, first the relationship between listing type and price will be analyzed. For this, a boxplot is used to find out the maximum, minimum and median prices for each listing type - Entire Home, Private Room and Shared Room.



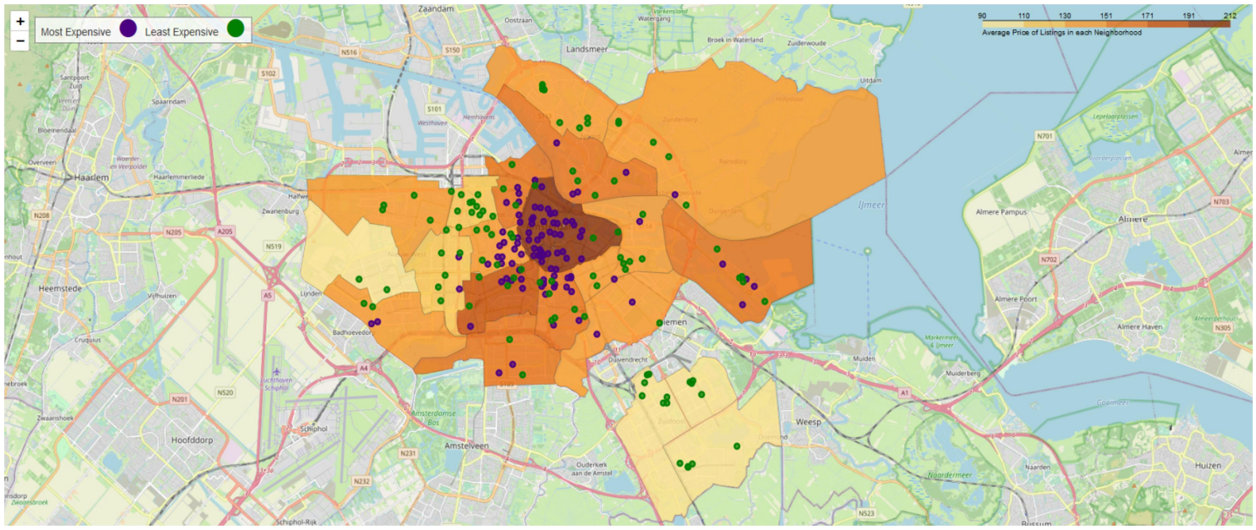
No surprises here, the entire apartments are priced higher than the private rooms, while shared rooms are the cheapest. The median price of entire house/apt listings are almost USD 150/night, the median price of private rooms are around USD 80/night, followed by shared rooms with a median price of USD 60/night. It seems that room type is a major decider of price and so will be considered as a feature for the later part of this section.

Relation between Location and Price

The next important feature to be looked into is the location. Here, the locations of the most expensive rentals are plotted on the map of Amsterdam to see if the higher priced neighborhoods vs. the lower priced ones can be identified.

The map visualization has been created in three steps:

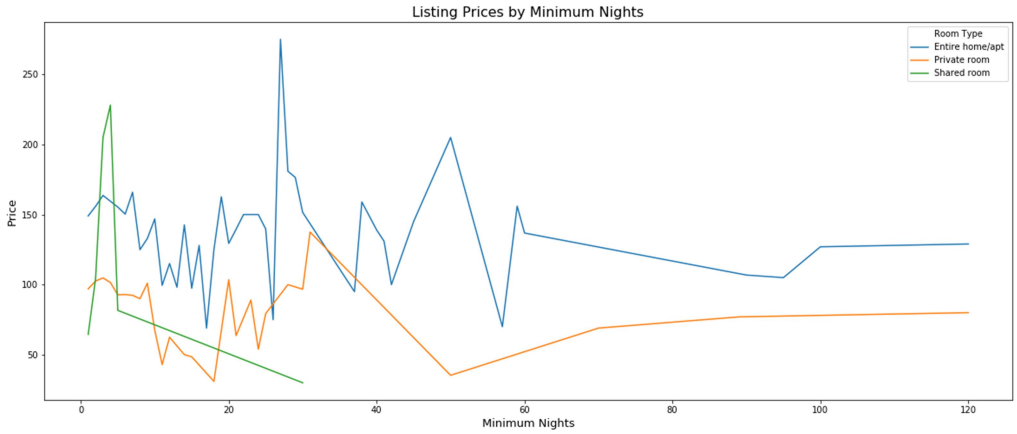
- A choropleth map is first created with the average listing price in each neighborhood
- The top 100 most expensive rentals are overlaid on the map in purple
- The top 100 cheapest rentals are overlaid on the map in green



It is evident from the map that the central part is the most expensive part of the city as there is a high density of the most expensive rentals (purple dots) in this area. The cheaper rentals are spread out in the outer neighborhoods, away from the city center.

Relation between Minimum Nights and Price

The last feature that will be explored in this section is minimum nights required to book a rental. To visualize how this feature relates to price, a line plot of price vs. minimum nights for each room type is drawn.



In the above plot, it is seen that for entire house/apts and private rooms the curves have certain similarities. Between the 0-20 day period, the price of a rental has a negative correlation with the minimum nights. There is a major spike at around the 25-30 day mark, following which the curve tapers down. A probable cause could be that the hosts who mandate a minimum stay of more than 25-30 days can also offer discounts. There is another dip in the price of entire house/apts at around the 60-day

mark, but private rooms do not share this trend. After 60 days, both the graphs flatten out. Shared rooms follow a completely different pattern. Here we see that the price is highest if the minimum stay is around 5 days and then the price keeps declining. There are no shared listings that require a stay of more than 30 days.

3.1.2. Clustering

Now that the top features that influence the price of a rental have been identified, the listing dataset will be clustered to examine the highest priced cluster. Before performing the clustering, one hot encoding on the 'room_type' variable is performed, since it is a categorical variable. Three dummy variables were created to represent each of the room types. Then 6 clusters are created from the dataset by k-means clustering method.

These 4 features were used for the clustering: Room Type (represented by 3 dummy variables), Location (represented by latitude and longitude), Minimum Nights and Price. Here are the results of the clustering:

Labels	Entire home/apt	Private room	Shared room	price	latitude	longitude	minimum_nights
0	1.000000	0.000000	0.0	129.741807	52.373278	4.867098	2.949218
1	1.000000	0.000000	0.0	133.317551	52.355204	4.920094	3.042653
2	0.000000	1.000000	0.0	99.159091	52.364840	4.892059	2.253575
3	0.993598	0.006402	0.0	261.626685	52.364478	4.887065	2.981806
4	0.000000	0.000000	1.0	93.813953	52.366988	4.893266	2.511628
5	0.823944	0.176056	0.0	137.732394	52.360547	4.889716	41.922535

- **Cluster 0** - Cluster of entire home/apt listings with its centroid near West Amsterdam with a minimum night requirement of around 3 nights
- **Cluster 1** - This cluster is very similar to Cluster 0 in every way except this cluster is centered in the eastern side of Amsterdam
- **Cluster 2** - Cluster of low value private room listings with a minimum night requirement of around 2.2 nights around the De Pijp - Rivierenbuurt neighborhood
- **Cluster 3** - This cluster contains the highest value listings and comprises almost entirely of entire home/apts with a minimum night requirement of around 3 nights between Centrum-Oost and Centrum-West neighborhood
- **Cluster 4** - Lowest value cluster of shared rooms with a minimum night requirement of 2.5 nights
- **Cluster 5** - Cluster contains a mixture of entire home and private rooms that have a very high minimum night requirement

If the clusters are mapped into price slabs, then the distribution looks like this:

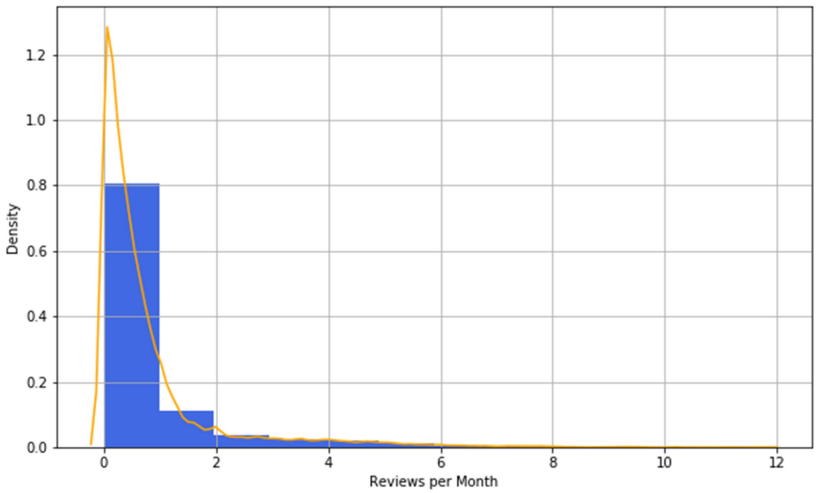
Price Slab	Cluster
High (250 and above)	3
Medium (101-249)	5, 1, 0
Low (up to 100)	2, 4

3.2. Popularity Analysis

Price is not the only factor to consider when hosting airbnbs. Popularity of a rental is another important factor to have a steady revenue stream. Proximity to restaurants, art/entertainment venues, transit locations is a major driving factor of popularity for rentals. For predicting the popularity of a listing, Foursquare's Places API will be utilized to get a list of venues close to the rental location.

3.2.1. Data Exploration

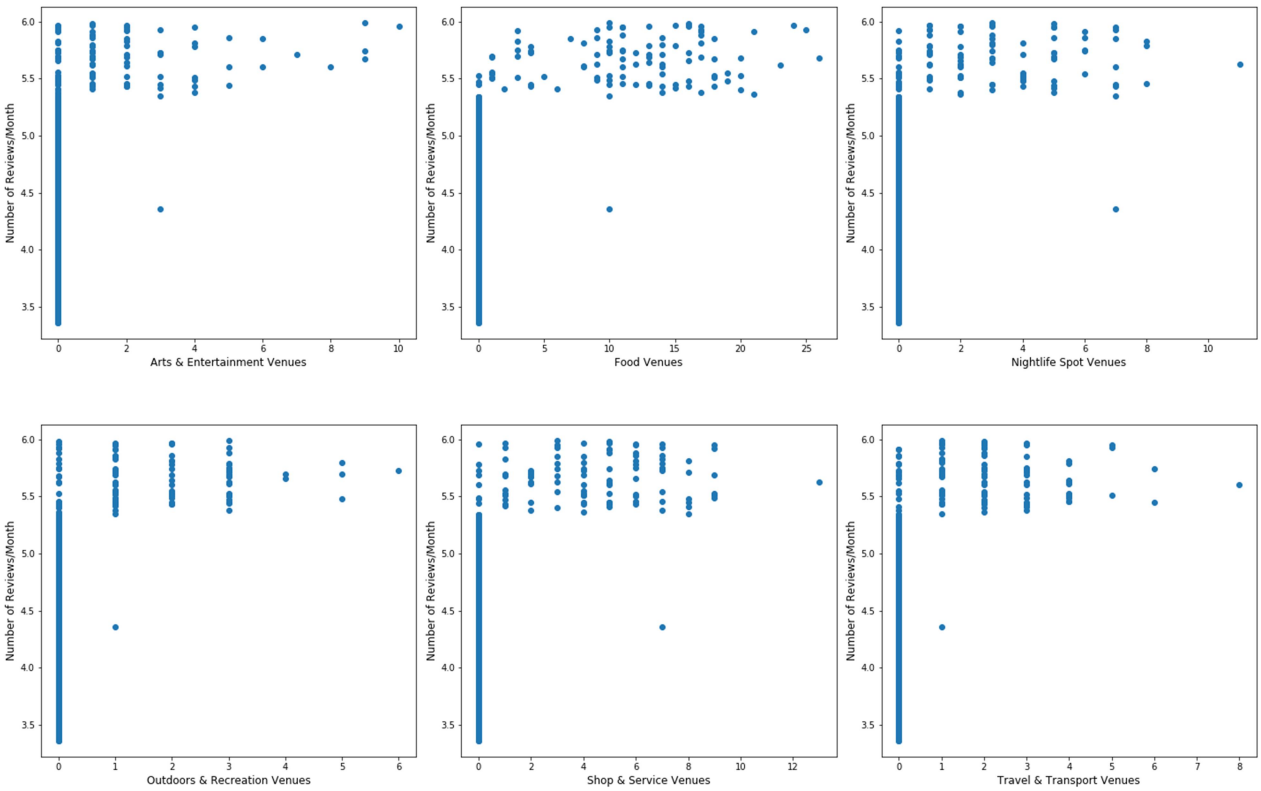
As a rental grows in popularity, it gets more bookings and subsequently more reviews. In this section, the number of reviews received per month by a rental will be used as a measure of popularity. Before starting any analysis on the popularity of airbnbs, it is important to look at the distribution of reviews per month to eliminate outliers.



From the histogram it is clear that there are not many rentals that get more than 6 reviews a month. So, these listings will not be considered for analysis and modeling in this section. From the remaining dataset the top 800 most popular listings will be used, since Foursquare api has restrictions on the number of calls made per day. In the next step, the explore endpoint of the foursquare api is called for each record to get all the venues within a 0.5 km radius. The number of venues near each listing is then categorized based on their type e.g. Food, Nightlife Spot etc. The resulting dataframe looks like this:

reviews_per_month	Arts & Entertainment	Food	Nightlife Spot	Outdoors & Recreation	Shop & Service	Travel & Transport
5.99	9	10	3	3	3	1
5.98	1	16	5	0	5	2
5.97	1	16	1	2	5	2
5.97	0	24	1	1	1	1
5.97	2	15	3	2	4	3

Now that the data is available, it will be possible to visualize the correlation between each type of venue and the rental popularity. To achieve this, the number of venues near each airbnb listing will be plotted against the number of reviews it receives per month. Remember that the reviews per month is the measure of popularity. Scatter plots have been used to plot the correlation.



In each of the scatter plots, it can be seen that rentals that are near one or more venues receive more reviews than those near 0 venues, irrespective of the venue type. So, it is clear that the nearness of venues increase the popularity of a rental.

3.2.2. Classification

In this section, two models will be built to classify whether or not a rental will be popular, based on its nearness to venues. The first model will be built using a k-nearest neighbor classifier and the second model will be a logistic regression classifier. Then the models will be compared to find the better classifier model. Before the classifications are done, the ‘reviews_per_month’ value is used to create the 'Popular' flag. The listings that get more than 5 reviews a month will be marked as Popular=True, the others will marked as Popular=False. This will be used as the target variable for the modeling part. The resulting dataframe looks like this:

reviews_per_month	Arts & Entertainment	Food	Nightlife Spot	Outdoors & Recreation	Shop & Service	Tı
5.99	9	10	3	3	3	
5.98	1	16	5	0	5	
5.97	1	16	1	2	5	
5.97	0	24	1	1	1	
5.97	2	15	3	2	4	
...	...	...	...	...	...	

The features and target are extracted from the dataset. Both the feature and target datasets are split into training and testing datasets using a 80/20 split. After the split, the training set contains 640 records and the testing set contains 160 records. A k-nearest neighbor classification model was built by fitting the feature and target training sets using values of k from 1 to 100. These models were then used to predict the target variable in the test set. It was noted that the best accuracy score was achieved for k=2. So, this value was used to build the final knn model. A logistic regression model is then built with c=.1 and 'liblinear' as the optimization algorithm using the training sets. The model is run on the test set to get the predicted values and probablities. Finally, the two models are compared using their jaccard index and f1 score.

Jaccard is the size of the intersection divided by the size of the union of two label sets. If the entire set of predicted labels for a sample strictly match with the true set of labels, then the subset accuracy is 1.0; otherwise it is 0.0. The F1 score is the harmonic average of the precision and recall, where an F1 score reaches its best value at 1 (perfect precision and recall) and worst at 0. It is a good way to show that a classifier has a good value for both recall and precision.

KNN Model Jaccard Index = 0.8562 KNN Model F1-score = 0.8335

Logistic Regression Jaccard Index = 0.8438 Logistic Regression F1-score = 0.8156

A higher score indicates a better prediction model for both the Jaccard Index and the F1-score. So, the K nearest neighbor classification model performs better than the logistic regression model and is the recommended classifier in this case.

A confusion matrix of the knn model looks like this:

	Actual Positive	**Actual Negative**
Predicted Positive	**17**	**23**
Predicted Negative	**0**	**120**

3.3. Availability Analysis

The final aspect of airbnbs that will be explored is the availability. For availability analysis, the future one year's booking data will be analyzed to identify patterns.

3.3.1. Data Exploration

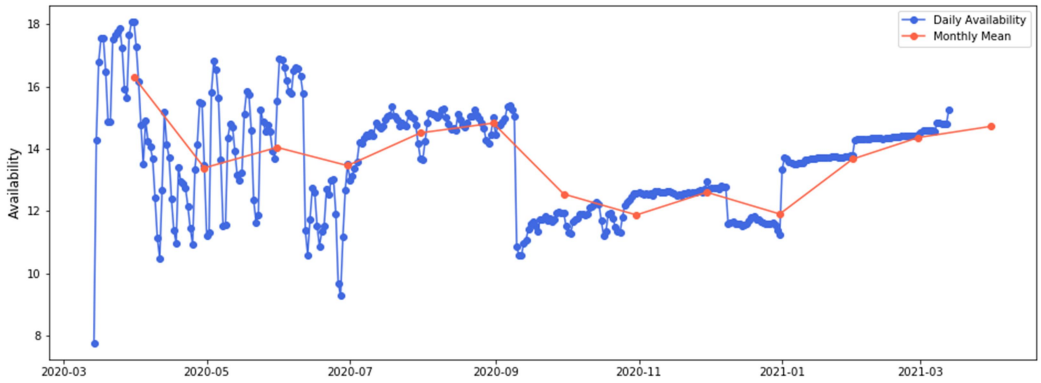
The data that will be used for the availability analysis is the booking data by listing over the next one calendar year. The data is formatted as a list of dates for each listing, with the availability marked with a flag. This data is converted to a time series data of percentage occupancy for each listing over a calendar year. Then, mean occupancy percentage is calculated for all the listings for each date. Finally, the dataframe that looks like this:

	date	occupancy
0	2020-03-14	7.75146
1	2020-03-15	14.2704
2	2020-03-16	16.7914
3	2020-03-17	17.5503
4	2020-03-18	17.5503
...	...	...
360	2021-03-09	14.8256
361	2021-03-10	14.8154
362	2021-03-11	14.8154
363	2021-03-12	14.8154
364	2021-03-13	15.261

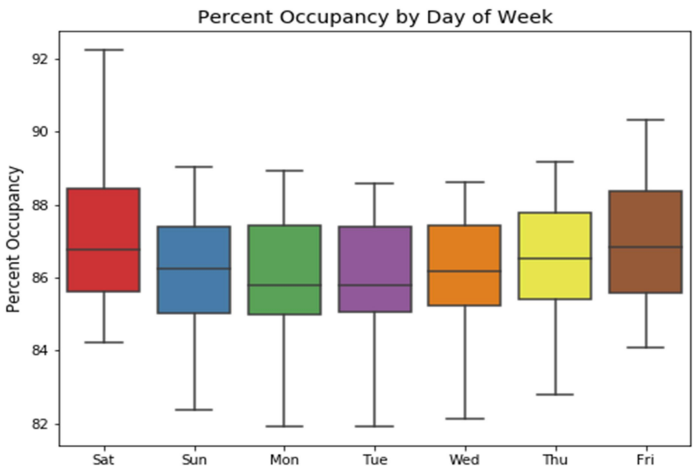
365 rows × 2 columns

Availability for a Calendar Year

Most time series data have trend and seasonality components. Trend is indicated by varying mean over time. Seasonality is the short term deviations in the data. To observe, how the occupancy varies over time, here is a time plot of the average availability over time. Since, there are a lot of ups and downs in the daily plot, the daily data is aggregated to a lower frequency (by each month) and is overlaid on the daily timeseries data. This method is called downsampling, since the timescale has been lowered.



Next, a boxplot has been generated to visualize the trends by each day of week.



From the plot, it can be observed that weekends have higher occupancy than the weekdays. This makes sense as people go out on weekends and are more likely to rent an airbnb. As a result, Saturdays have the highest median occupancy and Mondays have the lowest.

Unfortunately, historical reservation details of Amsterdam's airbnb rentals was not available in the dataset. So, a timeseries forecasting of rental availability could not be performed.

4. Results and Discussion

The biggest influencer of price and popularity when it comes to airbnbs is unsurprisingly their location. In Amsterdam, there is a lot of difference in airbnb pricing between the higher prices and lower priced neighborhoods as seen below:

Highest priced neighborhoods

Neighborhood	Average Price
Centrum-Oost	171.053273
Centrum-West	167.996933
Zuid	156.900548

Lowest priced neighborhoods

Neighborhood	Average Price
Bijlmer-Centrum	90.973913
Bijlmer-Oost	92.675926
Gaasperdam - Driemond	94.358333

The ideal mix of factors has been derived using a k-means clustering technique. It has been predicted that the entire home/apt listings with a minimum 3-night requirement between Centrum-Oost and Centrum-West neighborhood have the highest potential in terms of price per night.

In addition to the neighborhoods, it is also essential for the rental to be close to places of interest and restaurants to be popular. A classification model has been developed using k-nearest neighbor technique to predict the popularity of a rental based on its closeness to places of business like restaurants, nightlife spots etc.

Finally, the availability trend of the city's rentals for the next year show that a lot of bookings have already been made. Also, it is harder to book a rental over the weekends. But, the month of September in the current year has the highest availability. So, it will be easier to get a booking during that time.

5. Conclusion

This report should give potential airbnb hosts in the city of Amsterdam enough information to get a projection of the revenue they can expect from their listing. Existing hosts who are looking to maximize their profits should also get some valuable inputs. Tourists looking to book short term rentals should also get a fair idea of when they are most likely to get accommodation.

6. References

- https://en.wikipedia.org
- https://www.airbnb.com/
- http://insideairbnb.com/
- https://matplotlib.org/3.2.1/api/pyplot_summary.html
- https://pandas.pydata.org/docs/reference/index.html
- https://python-visualization.github.io/folium/
- https://developer.foursquare.com/docs/places-api/