Contents

[1. Airbnb 5](#_Toc155046249)

[1.1 Problem Understanding 6](#_Toc155046250)

[1.2 Subset Data 6](#_Toc155046251)

[1.3 Exploratory Data Analysis 6](#_Toc155046252)

[1.4 Feature Creation 14](#_Toc155046263)

[1.5 Missing Value Imputation 15](#_Toc155046265)

[1.6 Handling Outliers 16](#_Toc155046266)

[1.7 Numerical Transformation 17](#_Toc155046267)

[1.8 Categorical Data Encoding 19](#_Toc155046270)

[1.9 Correlation Analysis 20](#_Toc155046271)

[1.10 Feature Selection 21](#_Toc155046273)

[1.11 Feature Scaling 22](#_Toc155046277)

[3.12 Polynomial Expansion 22](#_Toc155046278)

## 

# 1. Airbnb

## 1.1 Problem Understanding

The Airbnb dataset includes various attributes such as neighborhood details, room types, amenities. The primary objective is to understand the influential factors that affect pricing within this specific region, exploring categorical and numerical features to uncover patterns and relationships. The aim is to prepare the dataset in order to develop a predictive model that accurately estimates property prices based on key attributes, providing valuable insights for both property owners and potential guests seeking accommodation.

## 1.2 Subset Data

I have decided to only focus on the **Central Region Neighbourhood** as this region contributes to close to **80%** of the data collected. The reason for doing this is because focusing on a particular demographic might be more relevant for the research.

Some benefits of subsetting data:

* + Large datasets contain irrelevant or noisy information. Focusing on a subset relevant to the analysis reduces unnecessary noise and complexity, allowing for a clearer understanding of the factors that directly impact the scope.
  + Businesses often have targeted objectives or decisions related to specific customer segments, product categories, or geographic areas. Analyzing a subset corresponding to these segments aids in making more informed and targeted decisions.

In my case, when subsetting data to only 1 region which is central, this will be helpful as usually, tourists would like to book an Airbnb in central areas and therefore I can target these audiences when creating my model.

The figure below shows the shape of the dataset after subsetting.



Shape of Dataset

By subsetting the dataset, I have reduced the samples from 7907 to 6309. However, for the price range column, it is still very diverse as the minimum is at 0 while the maximum is at 10000.

## 1.3 Exploratory Data Analysis

Categorical Data Analysis

## A graph of a number of blue rectangular objects Description automatically generated

**Some Background Context:**

**Private Room**: Guests have exclusive access to the bedroom/sleeping area of the listing. Other parts area such as the living room, kitchen, and bathroom are likely open either to the host even to other guests.

**Entire home/apt**: Guests have the whole place for themselves. It usually includes a bedroom, bathroom, and kitchen.

**Shared Room**: Guest sleep in a bedroom or a common area that could be shared with others.

* Entire home/apt has the most demand followed by Private Room and lastly Shared Room.
* Usually those who use Entire home/apt are families and this shows that in Singapore, the largest user of Airbnb will be families.
* Shared Room is not popular as customers may find it uncomfortable to sleep with other people, but still few people use it as it can be very cheap.

## A graph of a number of blue rectangular objects Description automatically generated

* Entire home/apt is the most expensive followed by Private Room and lastly Shared room.
* As Entire home/apt gives the most amenities it is expected that it would be the most expensive compared to other room types.

A screen shot of a graph

Description automatically generated

* Entire home/apt has the largest price range, a few of this rooms can be cheap but mostly the price of this room ranges between 100 to 300. There are cases where the price can be very expensive and go up till 8k
* For Private room, the price range is usually below 200 but there are outliers where the price for a private room went up till 10k which is much more than the most expensive Entire home/apt room type.
* Shared room price range is usually below 100.

## A map of singapore with many dots Description automatically generated

## A map of singapore with many hexagons Description automatically generated

* Most of the listings are more expensive in more central areas of Singapore like Orchard.
* The listings in Sentosa are very expensive as well
* When tourists come to Singapore, they usually visit very popular places in Singapore like Sentosa and Orchard and therefore these areas tend to be more expensive compared to other areas.

## A graph with numbers and letters Description automatically generated with medium confidence

* Kallang, Southern Islands and Outram has outliers that shows that the price going above 6k.
* However, Orchard does not have any extreme outliers and the price range lies between 36 to 900 which is a very surprising as other towns like Southern Islands have a greater range compared to Orchard.

## A graph with blue squares Description automatically generated

* Southern Islands has the highest average price followed by Marian South and Orchard.
* Southern Islands stands out compared to the other neighbourhoods as the avg price of Southern Islands is 1,8k while the other neighbourhoods are below 500.

## A close-up of words Description automatically generated

* The most repeated words in the name column is 'mrt', 'room', 'city'.
* I would be exploring more about 'mrt' later on.

Numerical Data Analysis

## A graph with a blue line Description automatically generated

* price is very heavily right skewed.
* Many outliers are present.

## A graph with a blue line Description automatically generated

* minimum\_nights is right skewed.
* Many outliers are present.

## A graph with a line and a red line Description automatically generated

* number\_of\_reviews is right skewed
* Many outliers are present

A graph with a blue line

Description automatically generated

* reviews\_per\_month is right skewed.
* Some outliers are present.

A graph with a red line

Description automatically generated

* calculated\_host\_listings\_count is right skewed.
* Few outliers are present.

A graph with a line drawn on it

Description automatically generated

* availability\_365 is right skewed
* No outliers are present

## 1.4 Feature Creation

Based on the word cloud I created, it showed that ‘MRT’ was a very popular word used in listing names. Therefore, I wanted to create some features related to that.

Whenever possible, property developers, would if they could advertise their properties as being in walking distance from a train station. With this, it can be reasonable to extrapolate that listings near train stations can command a higher-than-average price. I found the nearest MRT/LRT station for each property and as well calculated the distance between each listing and the MRT/LRT.

I have created 4 features nearest\_stn, nearest\_stn\_lat, nearest\_stn\_lng, distance\_from\_stn(m). The figure below shows a sample of the data. This data was created with the help of the mrt\_lrt\_data.csv dataset which I will also be uploading in Brightspace.

A screenshot of a white table with numbers and text

Description automatically generated

With the nearest\_stn\_lat, nearest\_stn\_lng, I was able to create a chart where I can visualise all the listings and the train stations together.

A map of singapore with many dots

Description automatically generated

The black spots show the train station points.

## A graph of a number of people Description automatically generated

* If a Airbnb listing is close to Harbourfront, then that listing price would be very expensive. This can be because the closest station for Sentosa is Harbourfront and therefore the average price of listings near Harbourfront MRT station is expensive.
* 2nd is City Hall and this can be due to the fact that there are many tourist spots at City Hall and it is a very touristy spot therefore listings near this MRT station is expensive.

## 1.5 Missing Value Imputation

There are 3 columns in the dataset that has missing values, and I will show how I have imputated the missing values.

A screenshot of a computer program

Description automatically generated

This is how I carried out **Missing Value Imputation**:  
  
**name:**

name will not be able to provide any significant value to the model as it is a categorical column that has long sentences and it will not be easy to convert this column to a numerical column as since the model only understands numbers.

**last\_review:**

last\_review is a date and it is not possible to give a date datatype to the model. Additionally, this feature does not seem helpful to predict prices.

**reviews\_per\_month:**

This column has null values as listings have never been reviewed and therefore I replaced the null values with 0 to indicate that the listing has never received reviews before.

## 1.6 Handling Outliers

I have used winsorization to handle the outliers. I choose winsorisation because:

* The biggest issue with this dataset is the outliers and they make the dataset very noisy.
* Removing outliers is an option but that will lead to a huge data loss.
* A screenshot of a computer screen

  Description automatically generatedWinsorizing retains the information from outliers while reducing their impact on statistical measures or models. It adjusts extreme values to more plausible values within a certain range.

By using Winsorisation, the feature's extreme values which are beyond a certain point are replaced or set to a specified threshold value and this technique was effective in fixing the outliers. Another positive is that the skewness of the features has also reduced.

## 1.7 Numerical Transformation

I have used numerical transformation to further reduce the skewness in the features.

## A screenshot of a graph Description automatically generated

## A screenshot of a website Description automatically generated

These transformations have reduced the skewness of the features further which is very beneficial.

## 1.8 Categorical Data Encoding

**Encoding methods chosen:**

**neighbourhood:**  
Ordered Ordinal Encoding

There seem to be a natural order to the neighbourhood. Due to this nature forming a ordinal relationship, I decided to use ordinal encoding for this feature. Moreover, doing one hot encoding could lead to problems as it will result in curse of dimensionality as this column has a high cardinality.

**nearest\_stn:**  
Ordered Ordinal Encoding

Similar to the feature above, I decided to use ordinal encoding for this feature as this has high cardinality and by doing dummy encoding, it will lead to curse of dimensionality.

**room\_type:**  
Dummy Encoding

There seem to be a natural order to the room type. The order will be Shared room, Private room then Entire home/apt. I have customised the pipeline for this feature to follow this custom order.

## 1.9 Correlation Analysis

## A screenshot of a graph Description automatically generated

Some significant correlation between the features and the target variable:

* neighbourhood
* room\_type
* nearest\_stn

## Feature Selection

I have done 2 tests to identify important features:

## A table with numbers and letters Description automatically generated

* From the above model summary, we can see that host\_id and reviews\_per\_month have a higher p-value compared to other features. It means these features has lower confidence on the coefficients of these features and these features may not be good indicators to predict price.
* The other features have p-value close to 0.

Now, I will F-scores to identify important features and not important features:

## A graph with blue bars Description automatically generated

## 

Based on the 2 tests done, I removed 'host\_id', 'reviews\_per\_month', 'distance\_from\_stn(m)', 'calculated\_host\_listings\_count' as these columns will not help the target variable based on the tests I did abo

## Feature Scaling

The reason to do scaling for this dataset is because:

* Linear Regression heavily depend on distances between data points, and features with larger scales can disproportionately impact distance computations which adds bias.
* Scaling features to a similar range can improve the stability of calculations within the algorithms, especially when features have very different scales. This can improve the stability of the model.

A comparison of a graph

Description automatically generated

## 1.12 Polynomial Expansion

Polynomial Expansion is a feature engineering technique that is used to create new features by raising the existing features to various powers. It is usually used to improve models and it does it in several ways:

* Polynomial expansion allows the model to capture and learn more complex, non-linear relationships between the features. This enables the model to better represent and fit non-linear patterns in the data.
* Polynomial expansion increases the flexibility of the model by introducing a wider range of functional forms. It allows the model to fit the data more closely and make more accurate predictions.

For the polynomial expansion, I have used a degree of 2. Then, I filtered the newly created features based on their correlation with the price target column. After filtering the dataframe, I created the correlation heatmap of how the final dataframe. The reason why I did polynomial expansion is that it is able to create more features that is correlated to the target column.

## A screenshot of a graph Description automatically generated