

**School of InfoComm Technology**

**Machine Learning**

Diploma in Data Science (DS)

Diploma in Information Technology (IT)

October 2022 Semester

**INDIVIDUAL ASSIGNMENT 1**

(30% of Machine Learning Module)

**Deadline for Submission:**

**17th Dec 2022 (Saturday), 2359 Hours**

|  |  |  |
| --- | --- | --- |
| Student Name | : | Loke Kai Fong |
| Student Number | : | S10204160J |
| Video Presentation Link | : | <https://youtu.be/ZiNfbTqv-4o> |

**Penalty for late submission:**

10% of the marks will be deducted every day after the deadline.

**NO** submission will be accepted after 24th Dec 2022, 23:59.

**Table of Contents**

[**1. Overview 2**](#_Toc122209947)

[**2. HR Analytics 4**](#_Toc122209948)

[**2.1 Problem Understanding 4**](#_Toc122209949)

[**2.1.1 Context 4**](#_Toc122209950)

[**2.1.2 Load & Explore 4**](#_Toc122209951)

[**2.2 Data Exploration 5**](#_Toc122209952)

[**2.2.1 Univariate Analysis 5**](#_Toc122209953)

[**2.2.2 Exploring Data with Visualizations 6**](#_Toc122209954)

[**Does number of training affect promotion? 6**](#_Toc122209955)

[**Does education affect promotion? 7**](#_Toc122209956)

[**Does length of service/age affect promotion? 8**](#_Toc122209957)

[**Which gender has more promotions? 9**](#_Toc122209958)

[**Top 10 region that has the most promotions? 10**](#_Toc122209959)

[**Does Training Score Impact Promotion? 11**](#_Toc122209960)

[**Correlation of HR Features (Before Transformation) 12**](#_Toc122209961)

[**2.3 Data Cleansing and Transformation 13**](#_Toc122209962)

[**2.3.1 Stratified Sampling 13**](#_Toc122209963)

[**2.3.2 Dropping Irrelevant Features 14**](#_Toc122209964)

[**2.3.3 Missing Value Handling 14**](#_Toc122209965)

[**2.3.4 Outlier Handling 14**](#_Toc122209966)

[**2.3.5 Numerical Transformation 15**](#_Toc122209967)

[**2.3.6 Categorical Transformation 15**](#_Toc122209968)

[**2.4 Correlation Analysis 16**](#_Toc122209969)

[**2.5 Modelling 17**](#_Toc122209970)

[**3. Airbnb Analytics 18**](#_Toc122209971)

[**3.1 Problem Understanding 18**](#_Toc122209972)

[**3.1.1 Context & Problem Statement 18**](#_Toc122209973)

[**3.1.2 Load & Explore 19**](#_Toc122209974)

[**3.2 Data Exploration 19**](#_Toc122209975)

[**3.2.1 Univariate Analysis 19**](#_Toc122209976)

[**3.2.1 Exploring Data with Visualizations 20**](#_Toc122209977)

[**Which room types are more expensive? 20**](#_Toc122209978)

[**What is the distribution of prices across neighbourhood regions? 21**](#_Toc122209979)

[**What are the popular Room Types for airbnbs? 22**](#_Toc122209980)

[**Reviews per Month & Minimum Nights vs Price 23**](#_Toc122209981)

[**Airbnb locations ('latitude' & 'longitude') by region 25**](#_Toc122209982)

[**Correlation of Airbnb Features (Before Transformation) 25**](#_Toc122209983)

[**3.3 Data Cleansing and Transformation 26**](#_Toc122209984)

[**3.3.1 Filter data for a specified model 26**](#_Toc122209985)

[**3.3.2 Dropping Irrelevant Features 27**](#_Toc122209986)

[**3.3.1 Missing Value Handling 27**](#_Toc122209987)

[**3.3.3 Outlier Handling 27**](#_Toc122209988)

[**3.3.4 Numerical Transformation 28**](#_Toc122209989)

[**3.3.5 Categorical Transformation 28**](#_Toc122209990)

[**3.4 Correlation Analysis 28**](#_Toc122209991)

[**3.5 Modelling 29**](#_Toc122209992)

[**4. Summary 31**](#_Toc122209993)

[**5. Further Improvements 31**](#_Toc122209994)

# 1. Overview

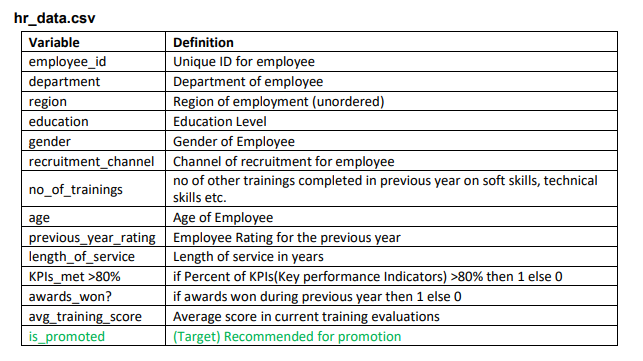
This report aims to cover the process involved in preparing datasets for machine learning modelling and document any key findings found during data preparation, exploration, and analysis.

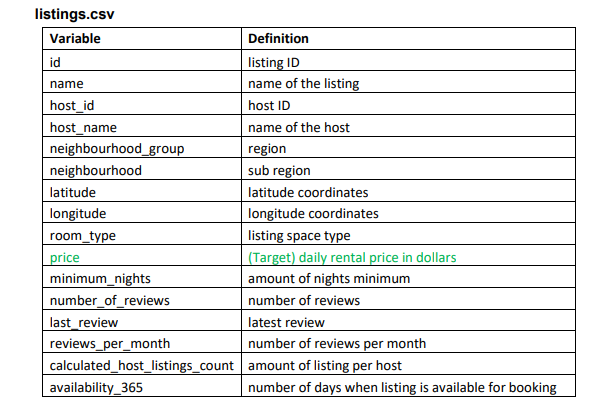
The preparation of the two datasets, “hr\_data.csv” and “listings.csv” will first consist of data exploration and analysis through visualizations and statistical analysis. Then, identifying the relationships, trends, and anomalies within the features to form a problem statement as the basis for our feature transformation and the construction of our machine learning model. Next, we will explore and analyse the correlation between the transformed features of the dataset to determine which features are useful as predictors for our target variable. Finally, we will export new newly transformed datasets into new .csv files for future machine learning modelling.

Exploratory Data Analysis will be done using a combination of univariate and bivariate analysis techniques. This is done to help us learn more about the characteristics, patterns, relationships, and even potential errors within the dataset. It is important to perform exploration without making any prior assumptions about the data, as it can affect the bias and interpretation of the data. This will help us determine which transformation and encoding methods to implement to better prepare our data for modeling.

Correlation Analysis will be done by using visualization such as heatmap visualization to observe the correlation of features. Model testing will also be done to determine feature importance and provide greater understanding of the relevance of each feature, which will be helpful in further transforming the dataset.

The first section of the report will cover the exploratory analysis and transformations used for the dataset “hr\_data.csv”, as well as the construction of the problem statement. The target variable for this model will be “is\_promoted”, which is a binary feature consisting of 0 and 1. Therefore, we will prepare and transform the data to be fitted into a binary classification model such as a logistic regression model as we are only interested in a specific outcome.



The next section will cover the analysis and transformation of the dataset “listing.csv”, and the construction of the problem statement. For this dataset, we are more interested in predicting the price of an Airbnb in Singapore. As we are predicting the value of a numerical value, “price”, a linear regression model would be best suited for predicting the target variable as it predicts the value of the target variable based on given predictor features. Thus, we will be preparing and transforming the features in “listing.csv” based on linear regression.

# 2. HR Analytics

## 2.1 Problem Understanding

Before exploring and transforming the dataset, we must first understand the context behind the data to form our problem statement and load it into the notebook for the actual processing and transformation.

### 2.1.1 Context

The “hr\_data.csv” dataset is a compilation of a company’s employee information such as their education, past performance department, etc. The aim is to utilize the data to create a machine learning model to identify the employees who are likely to get promoted based on their information. Therefore, the prediction problem statement for this dataset is to “Predict if an employee will get promoted based on their records.”

### 2.1.2 Load & Explore

The data is loaded into the jupyter notebook environment from the .csv file using pd.read\_csv() and saving it as hr\_data. From here, the dimensions and information regarding the features in the dataset can be observed with .shape and .info(). (show img) Based on the information, hr\_data consists of 14 columns with 54808 rows. The data types in hr\_data consist of 1 float64, 8 int64, and 5 object.

The categorical features in the dataset consist of “department”, “region”, “education”, “gender”, and “recruitment\_channel”. For numerical features, it consists of “employee\_id”, “no\_of\_trainings”, “age”, “previous\_year\_rating”, “length\_of\_sevice”, “KPI\_met >80%”, “awards\_won?”, “avg\_training\_score”, and the target variable “is\_promoted”.

Lastly, we can observe the statistical values of the numerical values with the .describe() function. This will show us the count, mean, standard deviation, and the distribution of the values in each feature. This can be useful for identifying and handling features with outliers later.

## 2.2 Data Exploration

Exploratory Data Analysis is done to understand the trends, relationships, and traits of the features. It is also done to identify outliers and missing values within the dataset. Data exploration will enable us to choose the appropriate methods for encoding and transformation to better prepare our data.

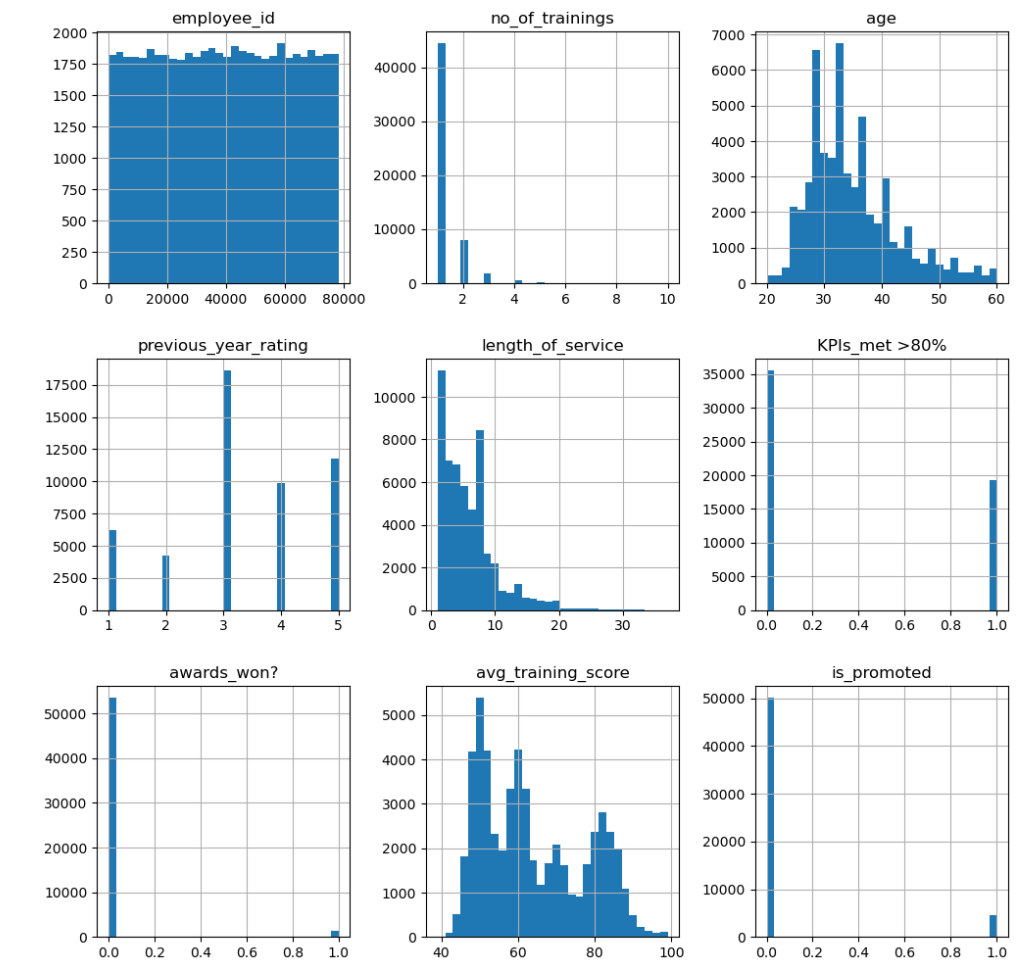
### 2.2.1 Univariate Analysis

First, we explore each feature individually using univariate analysis. This is done to provide us with more context and descriptions for each of the features. This analysis will allow us to narrow down on which features and the types of visualizations to be used in bivariate and multivariate analysis.

For the categorical features, we will be identifying the number of unique values for each feature. This allows us to see the data that we will be working with and identify any features with irrelevant or unnecessary values that will not be useful for prediction. Some key findings are,

* There are 3 recruitment channels, “other”, “soucing”, and “referred”. Most employees are from “other” sources.
* There are 3 levels of education, Bachelor's, Master's & above, and Below Secondary. Most employees are at Bachelor's.
* There are a total of 34 regions, with “region\_2” having the most employees
* There are 9 departments. The Sales & Marketing department has the highest number of employees.
* There are more males than females in this company

For the numerical features, we can use box plots and histogram visualizations to observe the statistical distribution of each numerical feature. This lets us see which features are normally distributed and allows us to identify the numerical values that are binary, discrete, or continuous. From the histograms, it can be observed that the features, “KPI\_met >80%”, “awards\_won?”, and “is\_promoted” are binary features. While the rest appear to be discrete or continuous variables. Outliers are present in “age”, “no\_of\_trainings”, “length\_of\_sevice”, “previous\_year\_rating”, “awards\_won?”, and “is\_promoted”.

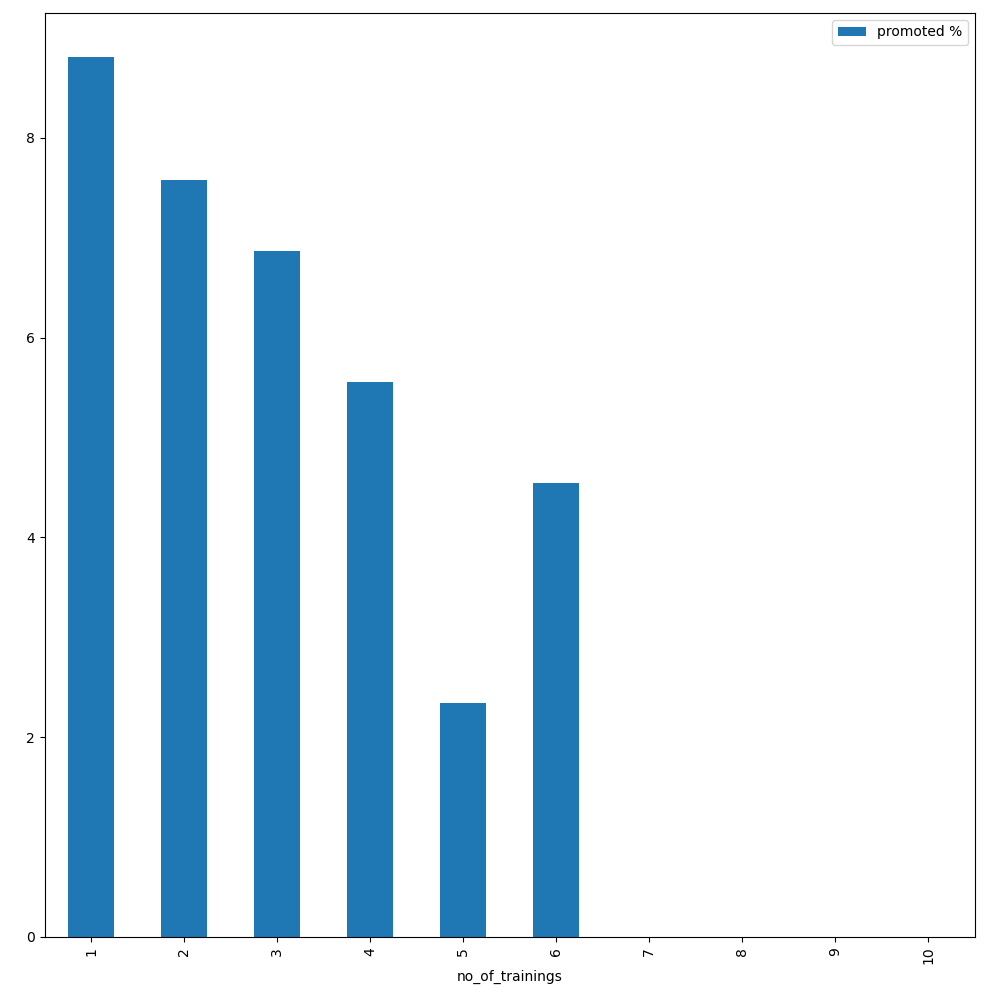


### 2.2.2 Exploring Data with Visualizations

After performing univariate analysis, bivariate and multivariate analysis will be done to further understand the data. The analysis will be done using visualizations and will mainly focus on the trends and relationships between the predictor features in relation to the target variable “is\_promoted.

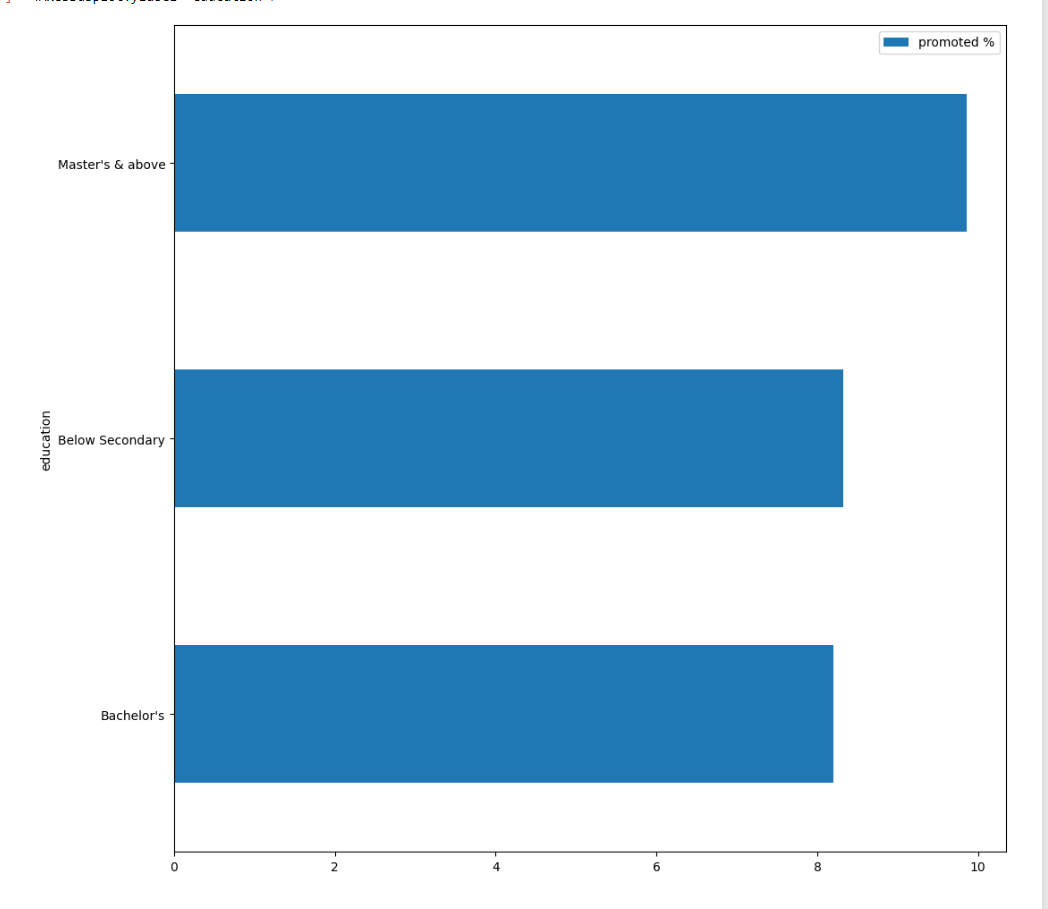
### Does number of training affect promotion?

The first visualization aims to find out if the number of trainings an employee takes affects the likelihood of a promotion. This is done by using the .groupby() and .sum() functions, to group the sum of 1s in “is\_promoted” by the number of trainings every employee has went through, then divided by the total number of employees to obtain the percentage of promoted, by the number of trainings. Based on the bar chart, it can be observed that the number of trainings is not a key criterion for promotion as many of the employees who were promoted only went through 1 training.



### Does education affect promotion?

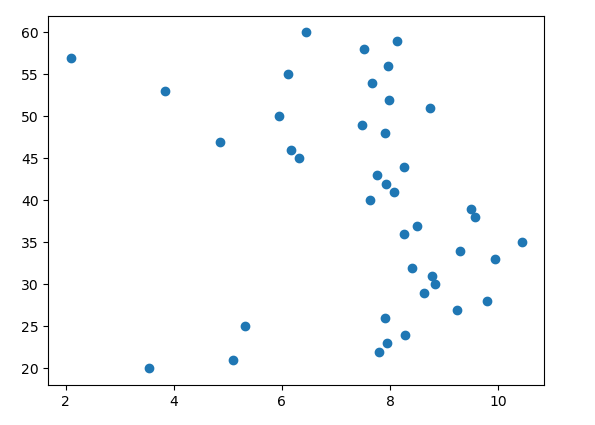
The second visualization aims to determine if the education level is a factor for promotion. This is done similarly to the previous visualization, using the .groupby() function to group the sum of those who were promoted by their education level, then dividing by the total employees for each group to obtain the percentage. Based on the bar chart, we can see that there is a slightly higher percentage of people who were promoted if they have a Master’s or above.



### Does length of service/age affect promotion?

The next visualization determines if the length of service of each employee influences the likelihood of a promotion. This visual uses a scatter plot to determine if there is a correlation between the two numerical variables. The scatter plot shows that there is no correlation between the two variables. This indicates that the length of service is not a factor when it comes to promotion. The following scatter plot visual plots age by promoted %. Employees at age 35 have the highest chance of getting a promotion. However, based on the scatter plot, there is little relation between age and promotion.

Length of Service



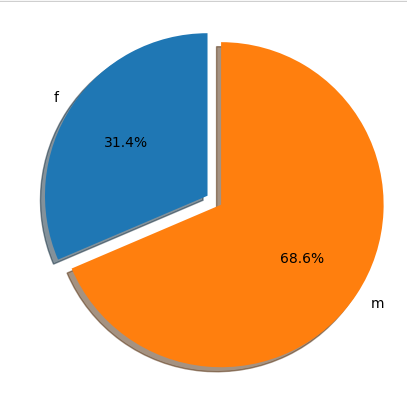
Age

Chart, scatter chart

Description automatically generated

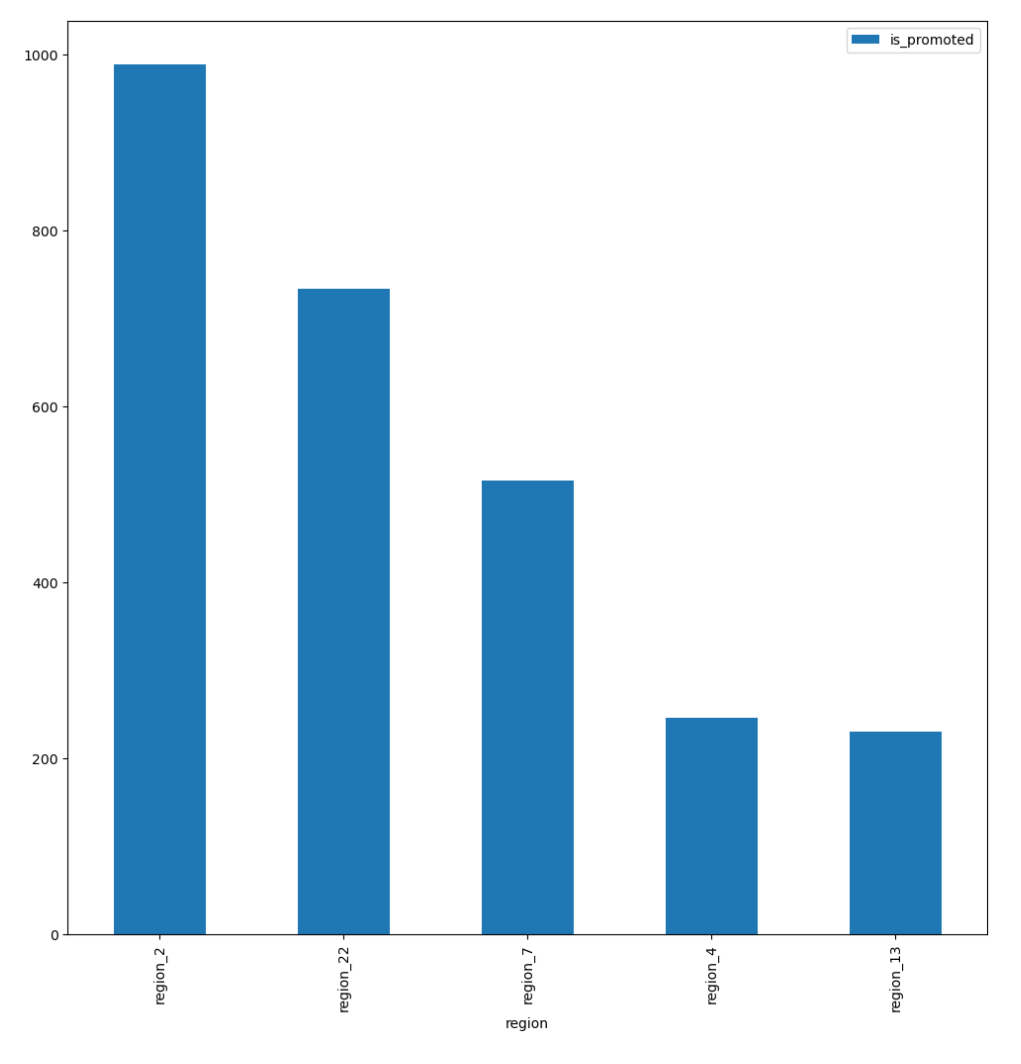
### Which gender has more promotions?

The next visualization shows a pie chart that indicates the number of employees who were promoted by gender. As a whole, there are more males than females in this company, the number of males who obtained a promotion is greater than females.



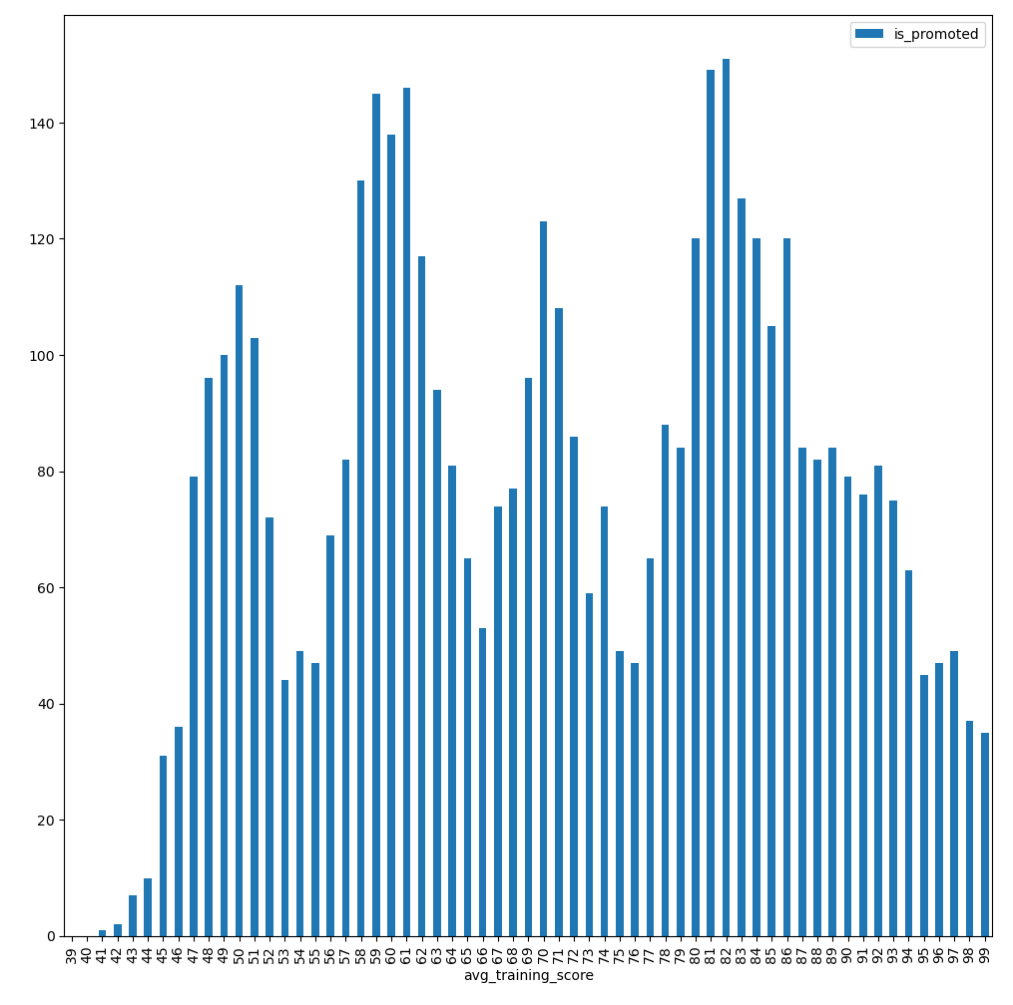
### Top 10 region that has the most promotions?

The next visualization focuses on the top 10 regions with the most promotions. This is done to see if the employees of certain regions are performing better and obtaining promotion compared to the other region. Based on the bar chart, “region\_2” has the most promotions among the other regions, which may indicate that region 2 has better performing employees or more employees than the other regions.



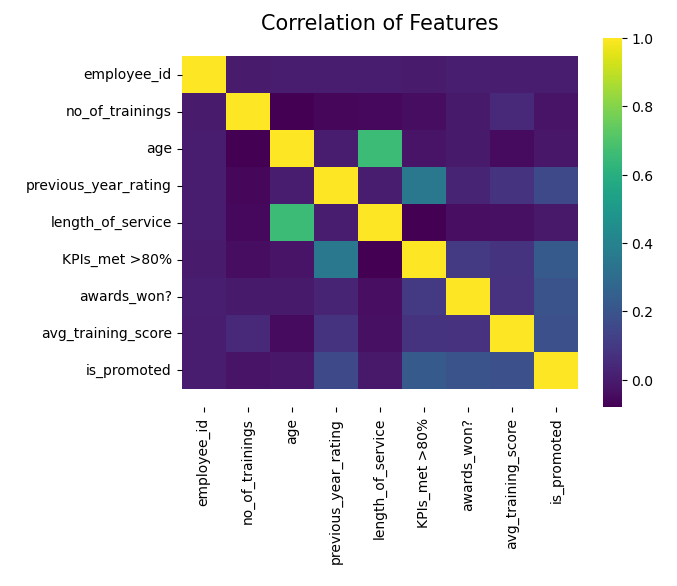
### Does Training Score Impact Promotion?

The next visualization shows the average training score by promotion. This visualization determines if the average training score has an impact on the likelihood of promotion. Based on the bar chart, we can see that there are typically more promotions for employees who have a higher average training score. This indicates that the feature will be helpful as a predictor on the target.



### Correlation of HR Features (Before Transformation)

The next visualization is a heatmap showing the correlation of the features before any transformation is done. This is done as preliminary step to identify and remove any unnecessary features before transformation, as well as to observe the correlation values between the predictor features and the target variable. From the heatmap, one key observation is that length of service and age have a strong correlation with each other. This may mean that they provide similar value to the machine learning model. It is unnecessary to have 2 similar predictors. Therefore, we drop “length\_of\_service” and keep “age” as it is a better feature for prediction based on the exploratory data analysis.



## 2.3 Data Cleansing and Transformation

After exploring the dataset, data transformation has to be done to in order to convert data into a format or structure that is better suited for the model. This involves handling missing values and outliers, as well as transformation techniques such as numerical transformation and categorical encoding.

### 2.3.1 Stratified Sampling

During the exploratory data analysis, it was found that there is a much higher proportion of people who were not promoted compared to people who were promoted. This may cause issues with modeling as it may not be the best representation of the dataset, which may make our model results trivial. Therefore, stratified sampling is performed to take an even number of people who were promoted and not promoted. This gives us a better distribution of the target variable, and ensures we properly represent data from both groups. This makes our prediction results more useful.

### 2.3.2 Dropping Irrelevant Features

Before performing transformation, we have to drop the features identified as irrelevant or insignificant as a predictor. The features identified are, “employee\_id” and “length\_of\_service”.

### 2.3.3 Missing Value Handling

Missing values in the dataset were found in the features “previous\_year\_rating” and “education”. The values will be imputed with another value to prevent the loss of any data. For previous year rating, the missing values were mainly found for employees that just joined the company, where their length of service is equal to 1. For this scenario, it is best to replace the missing values with the median value. Replacing the value with an arbitrary value such as 0 would indicate that new hires have a poor rating from the beginning, which may be unrepresentative of the data. For education, the null values were replaced with a string “missing” to reduce the risk of introducing falsified data into the data which can affect the bias of the model.

### 2.3.4 Outlier Handling

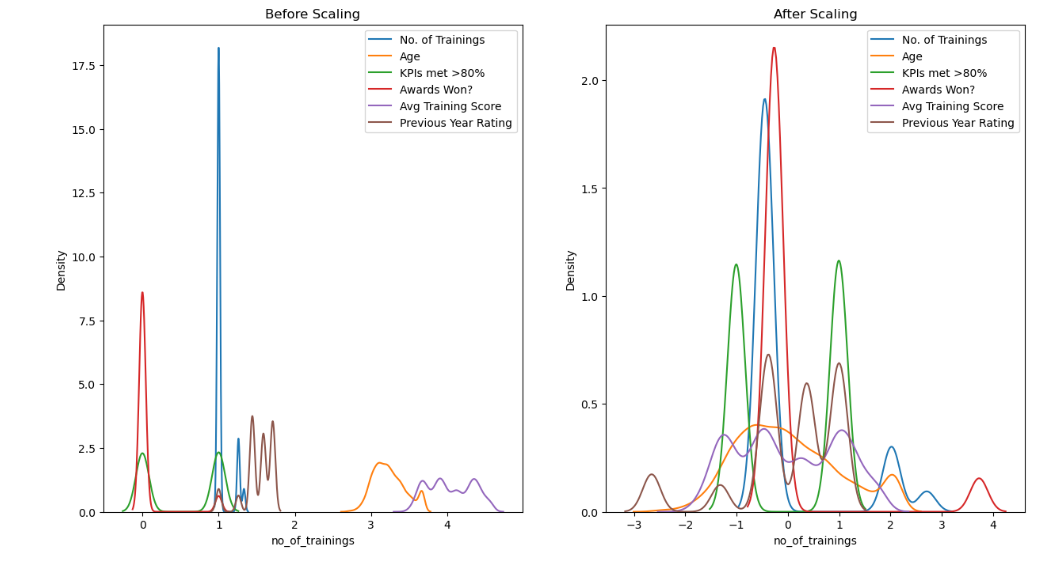
Outliers in the features also have to be handled as they can affect the skew of our model and is unrepresentative of the dataset as a whole. A user defined diagnostic plot function was defined to identify outliers present in the model. Outliers were found in every numerical feature except “is\_promoted”. However, outlier handling will only be performed on "age", "avg\_training\_score", "previous\_year\_rating", and "no\_of\_trainings". As “is\_promoted, awards\_won?, KPIs\_met >80%” are binary numerical features, therefore, there is no reason to handle outlier for those features. To handle outliers, the Winsorizer function with a capping method of gaussian, and fold = 2 on the right tail was used. This caps the values by the right tail based on the standard deviation of the feature. Right tail was specified as the outliers in the selected features had outliers that were above the maximum value.

### 2.3.5 Numerical Transformation

Before any transformation, the dataset is split into train and test datasets. This is to ensure that the datasets are transformed separately to prevent data leakage.

For numerical transformation, Power transformer was applied to all the numerical predictor features using feature engine. Numerical transformation is done to scale the numerical features down to ensure that they have similar magnitude. This is to prevent bias from being introduced when training the model. The distribution of the data becomes more "gaussian" or "normal" after using power transformation. I chose to utilize this transformer because, in contrast to other transformation techniques, it can be used on positive, negative, and zero values.

Feature scaling was also applied to further scale the numerical values to reduce the magnitude between the features. The feature scaling method used is standard scaler as it removes the mean and scales the data to the unit variance.

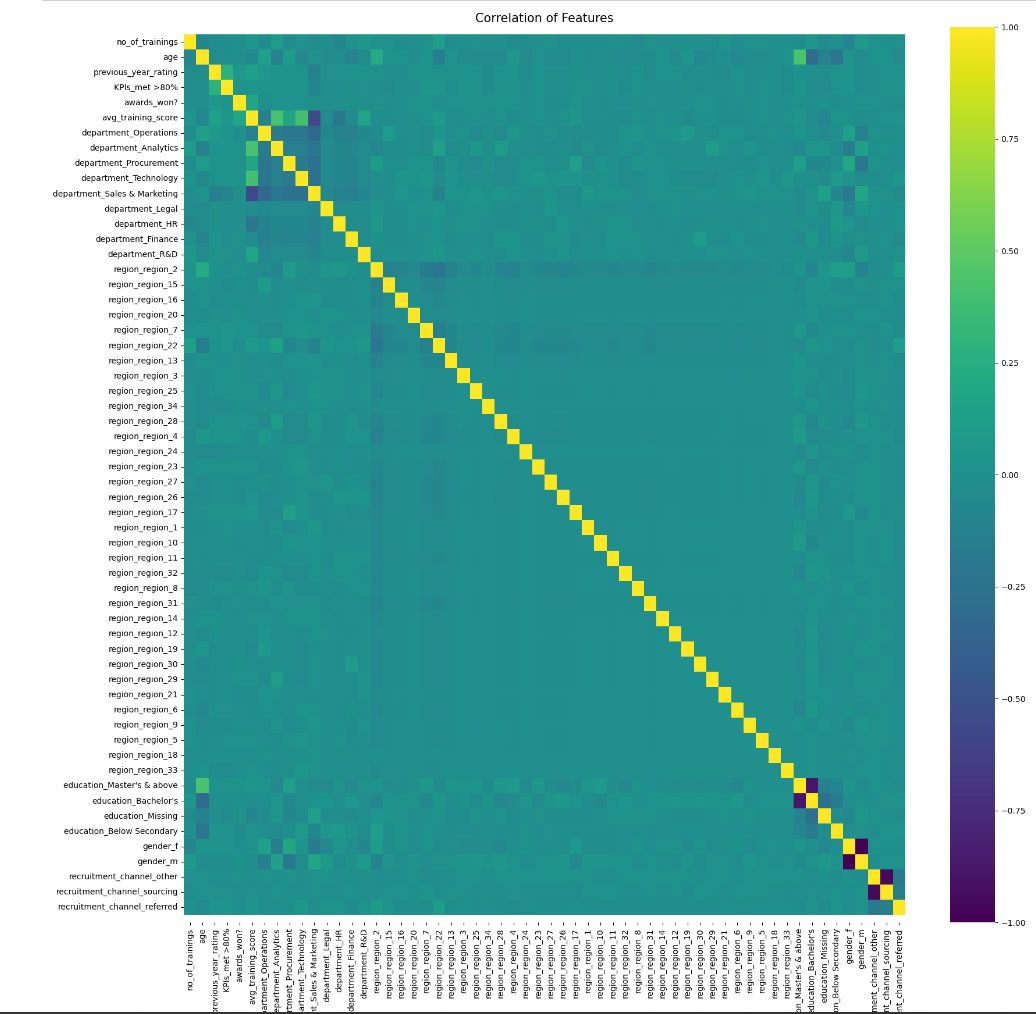


### 2.3.6 Categorical Transformation

For categorical encoding, One-Hot encoding was employed as it converts the categorical values into numerical values that can be read by the model. One-Hot encoding was used as it encodes the categorical values into binary values of 0s and 1s. This ensures that the model does not assume that higher numbers are more important, which is an issue with other encoders such as Ordinal encoding.

## 2.4 Correlation Analysis

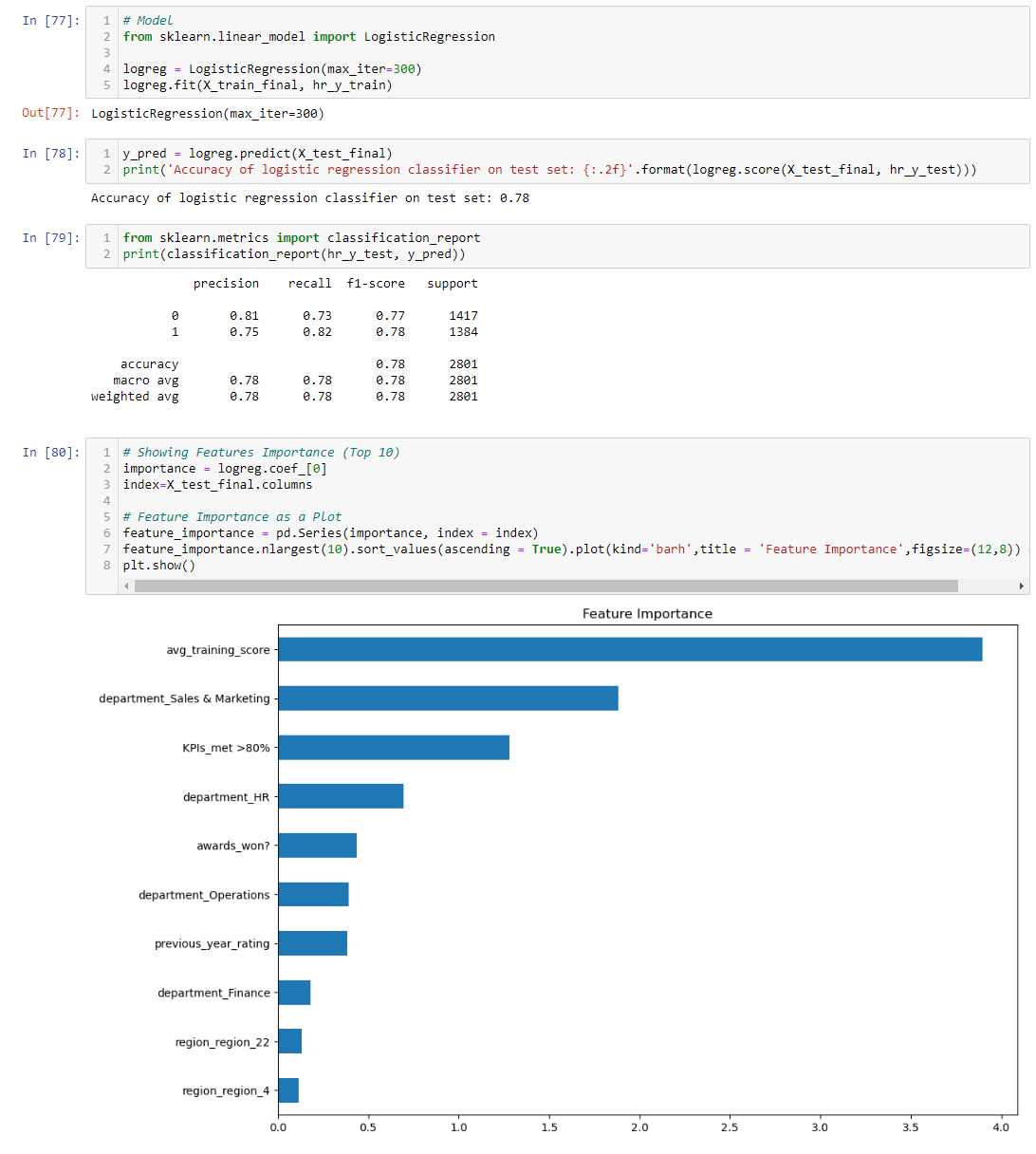
Correlation Analysis is method used to determine or discover the relationship between two features. This is done through the bivariate analysis done during data exploration. The best way to observe correlation is using a heat map. This is a heat map that shows the correlation between the features after transformation and encoding. The target variable is not present in this heatmap as it was dropped during the train test split.



This heat map shows the correlation between the predictor features. Observing the heat map, it is apparent that many of the predictor features do not correlate with one another with a few exceptions. For these exceptions, we may want to drop these features as they provide a similar value to the model. Dropping these features may further optimize the model’s performance.

## 2.5 Modelling

Lastly, to check if our transformed features are useful for prediction, the data can be fitted to a testing machine learning model to obtain the feature importance of each feature, as well as evaluate the accuracy of our predictive model to see if the transformations used were sufficient. The test accuracy of the logistic regression model on the data is 0.78, which indicates a good accuracy. Next, we can use the model coefficients to obtain the feature importance. Based on the bar chart, we can see that the features with the highest importance are avg\_training\_score, department\_Sales and Marketing and KPIs\_met>80%. These are the values that affect the likelihood of promotion greatly.



# 3. Airbnb Analytics

## 3.1 Problem Understanding

Similar to the previous dataset, we have to understand the context behind the data to form our problem statement before exploring and transforming the dataset.

### 3.1.1 Context & Problem Statement

The “listings.csv” file shows the listing activity of airbnbs from 2013 to 2019. The file contains details of the hosts, information on the homes and the reviews. Based on the features available, we can construct a model to predict the price of an Airbnb based on information available. Thus, the prediction problem statement for this dataset is “Predict the prices for airbnbs in the central region based on its characteristics.”.

### 3.1.2 Load & Explore

The steps for the initial exploration are the same as the previous dataset. First using pd.read\_csv() to load the data, saving it as bnb\_data. Then .shape and .info() is used to obtain a brief overview of the dataset. This dataset consists of 16 features with 7907 rows. Null values are present as there are features with less rows compared to the majority. The data types include float64(3), int64(7), and object(6).

The categorical features are “name”, “host\_name”, “neighbourhood\_group”, “neighbourhood”, “room\_type”, “last\_review”. Numerical features in the dataset are, “id”, “host\_id”, latitude', 'longitude', 'minimum\_nights', 'calculated\_host\_listings\_count', 'reviews\_per\_month', "price", 'availability\_365', and “reviews\_per\_month”.

Lastly, .describe() was used provide a brief overview of the statistical values of the numerical features.

## 3.2 Data Exploration

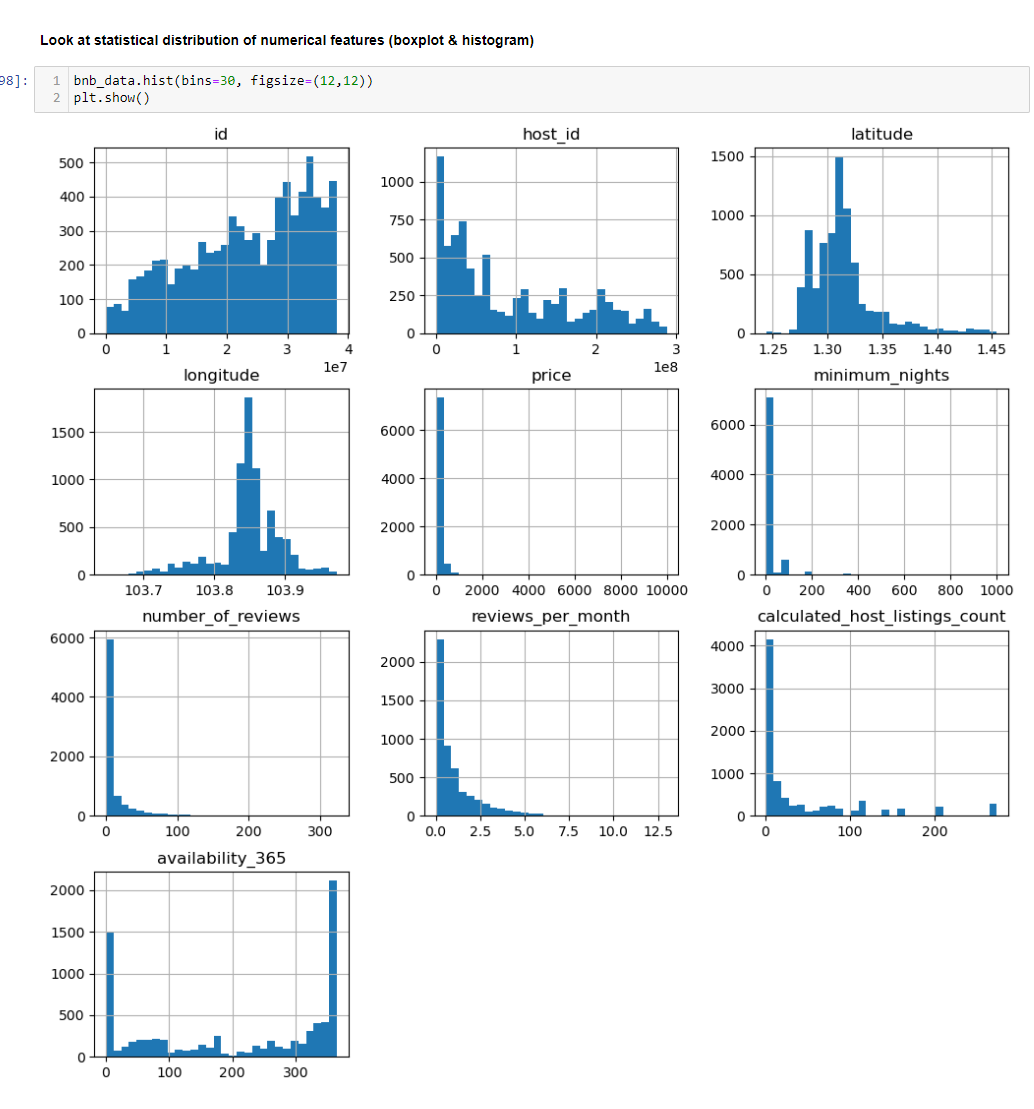
Exploratory data analysis is conducted next to identify trends and relationships between the features. The exploration of the data will be conducted in relation to the target variable “price”.

### 3.2.1 Univariate Analysis

For univariate analysis on the categorical features, the key findings are,

* There are 7454 Airbnb listings with unique names, and there are 1833 hosts with unique names.
* Central Region has the most number of Airbnb at 6309.
* Kallang neighbourhood has 1043 Airbnb while Lim Chu Kang only has 1.
* There are 3 room types and Entire home/apt is the most popular room type.

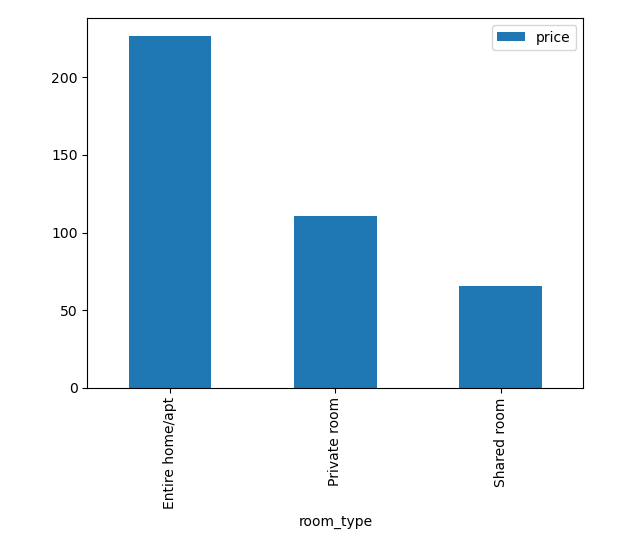
For univariate analysis on numerical values, histograms were used to show numerical distribution of the features. The histograms for the features “price”, “minimum\_nights”, “number\_of\_review” and “reviews\_per\_month” indicate the presence of extreme outliers. These outliers will have to be handled to prevent it from affecting the model’s performance. We can also observe that longitude and latitude are normally distributed.



### 3.2.1 Exploring Data with Visualizations

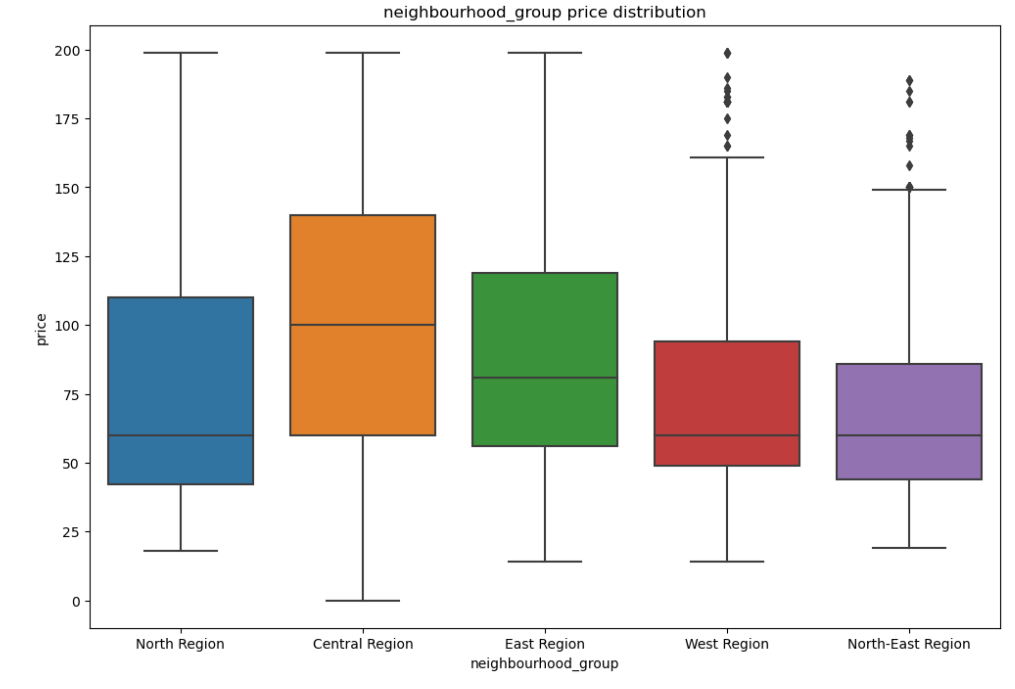
### Which room types are more expensive?

The first visualization aims to understand which room types are the most expensive based on average price. Based on the bar chart, it is observed that “Entire home/apt” is the most expensive room type on average.



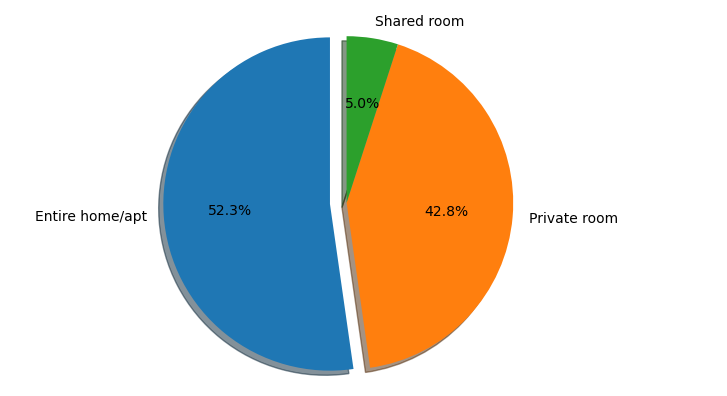
### What is the distribution of prices across neighbourhood regions?

The next visualization utilizes multiple box plots to show the distribution of prices across the different neighbourhood regions that are less than the average. This is to identify which neighbourhood region has the high price distribution on average. Central Region is shown to have the highest price distribution among the neighbourhood groups. This will be helpful for building a specialized model for predicting price in the central region as it is more effective to focus on a specific region due to the large price distributions across the country. This will allow us to build a model that better represents the subset of data we are predicting.



### What are the popular Room Types for airbnbs?

The next visual is a pie chart that showcases the percentage of room types available as a percentage of a whole. This visual shows which room type is the most popular in Singapore. Based on the pie chart, Entire home/apt has the highest share of listings on airbnb and is the preferred room type. Conversely, Shared room is the leased prefered, with the lowest share among the room types.



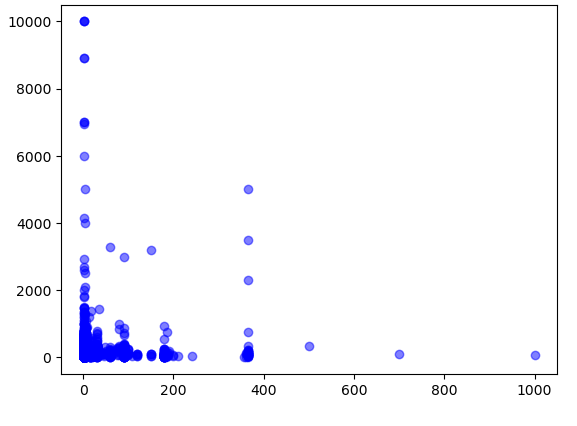
### Reviews per Month & Minimum Nights vs Price

The next visualizations are scatter plots that plot reviews per month and minimum nights against price. This is done to see if there is any correlation between the two numerical values to determine if these features are useful for predicting the target variable. For reviews per month against price, there seems to be a slight positive correlation, and for minimum nights and price, there also appears to be a slight positive correlation ignoring the outliers.

Reviews per monthChart

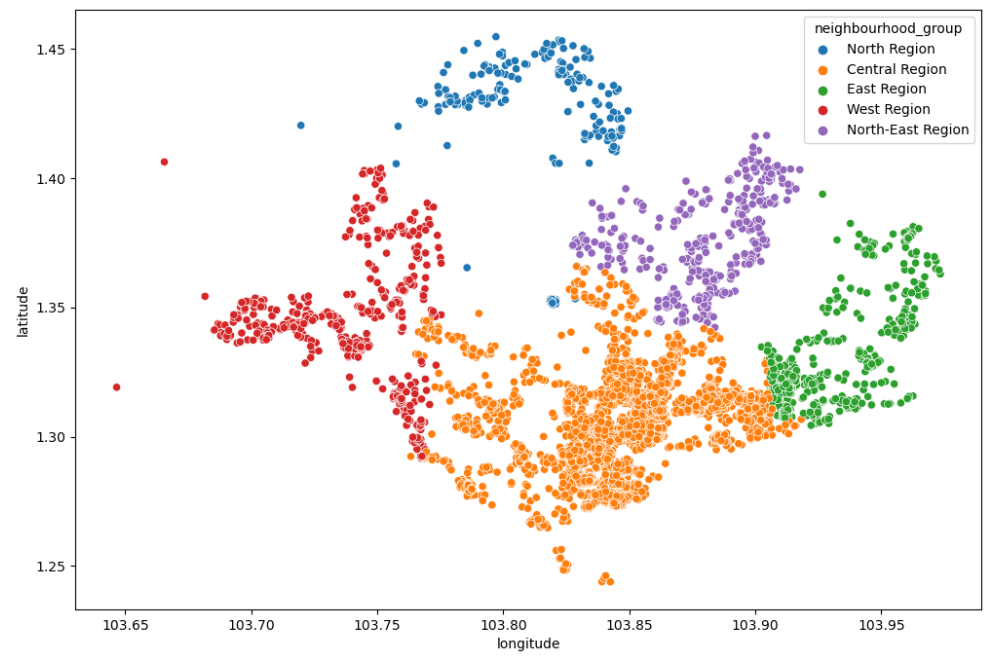
Description automatically generated

Minimum nights



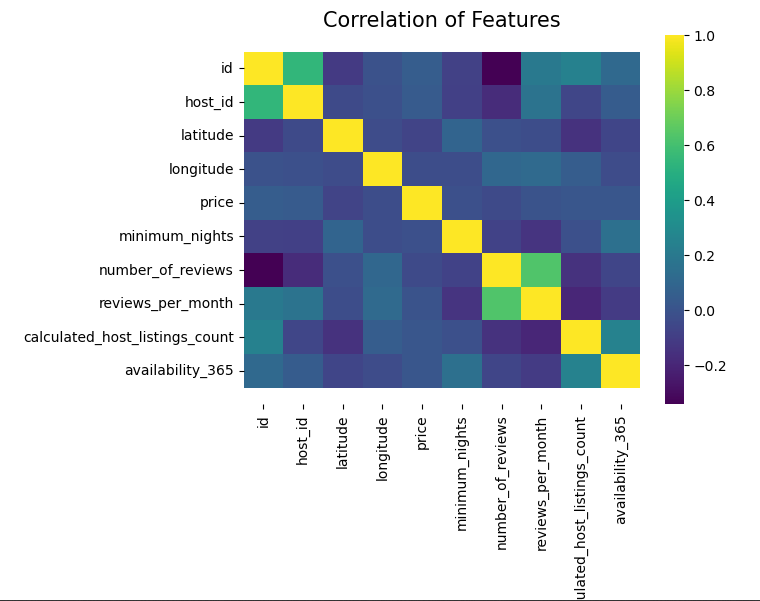
### Airbnb locations ('latitude' & 'longitude') by region

The next visualization is a scatter plot of the locations of the Airbnb by neighbourhood groups. This visual allows us to observe the distribution of the Airbnb listings available across Singapore, as the plot shows a visible map of Singapore using longitude and latitude values. From the visual, we can see that many of the Airbnb listings are concentrated in the central region. This is another indicator that it is more effective to build a model to predict prices in the central region only.



### Correlation of Airbnb Features (Before Transformation)

The last visualization is a heatmap showing the correlation of the features before transformation. This is a preparatory stage that looks for and excludes any features that are not necessary before transformation and checks the correlation between the predictor features and the target variable. From the heatmap, host id and id, as well as reviews per month and number of reviews have a strong correlation. It is unnecessary to have 2 similar predictors. Therefore, we will drop “reviews\_per\_month” and keep “number of reviews” as it is a better feature for prediction based on the exploratory data analysis. Both host id and id will be dropped as id values are generally not useful as predictor features.

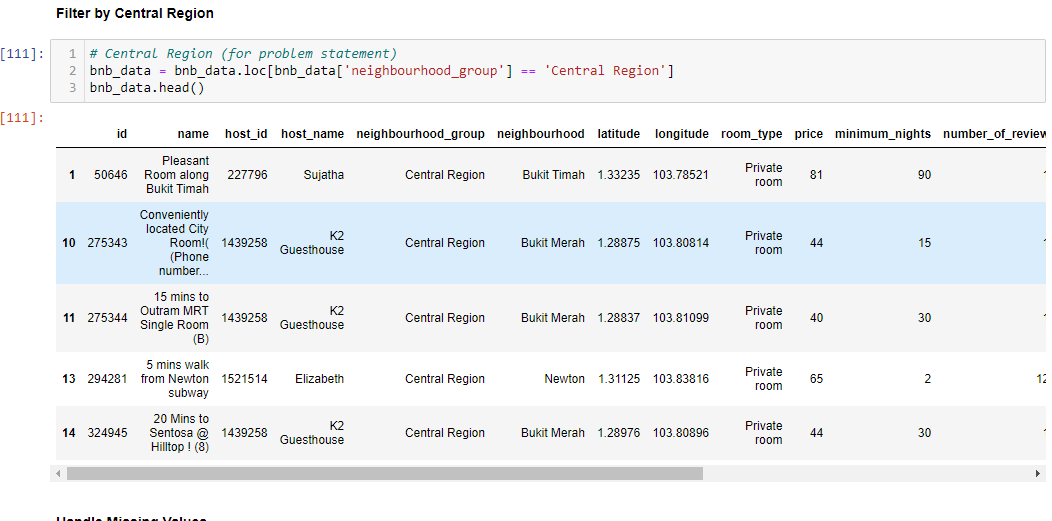


## 3.3 Data Cleansing and Transformation

Data transformation is necessary to transform data into a format or structure that is better appropriate for the model once the dataset has been explored.

### 3.3.1 Filter data for a specified model

During the exploratory data analysis, it was found that the distribution of the prices across the entire country were extremely varied. This can cause the model to be less effective as we would be predicting on a much wider range of values, which may make our model results trivial. Therefore, we will use Boolean masking to filter the values in the dataset, to only contain values from the Central Region. Then the neighborhood group column can be dropped as it would only contain a single value. This will make model more effective and allow our prediction results to be more useful.



### 3.3.2 Dropping Irrelevant Features

Before performing transformation, we have to drop the features identified as irrelevant or insignificant as a predictor. The features identified are, 'name', 'id', 'host\_id', 'last\_review', 'reviews\_per\_month', 'host\_name', 'neighbourhood\_group'.

### 3.3.1 Missing Value Handling

Missing values in the dataset were found in the features 'reviews\_per\_month', 'last\_review', and 'name'. The categorical features were imputed with missing, while the numerical feature was imputed with an arbitrary number as mean would be less accurate as it would be based on other listings.

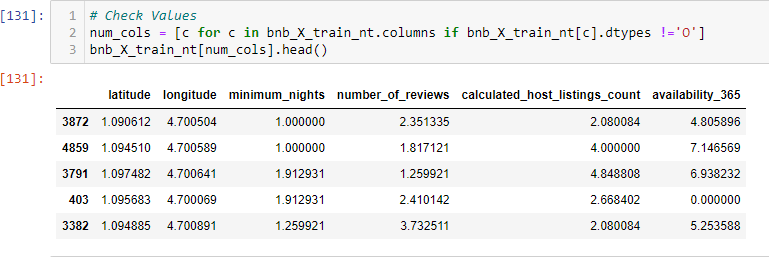
### 3.3.3 Outlier Handling

Outliers must handle as they can affect the skew and bias of our model. Using the same method as the previous dataset, a user defined diagnostic plot function was defined to identify outliers. Outlier handling was performed on "price", 'minimum\_nights', 'latitude', 'longitude', and 'calculated\_host\_listings\_count' as outliers in these values would greatly affect the bias of our final model, therefore it is best to handle the, therefore, there no reason to handle outlier for those features. To handle outliers, the Winsorizer function with a capping method of quantiles, and fold = 0.1 on the right tail was used. This caps the values by the right tail based on the 90th and 10th percentile of the feature. Right tail was specified as the outliers in the selected features tend to have outliers that were above the maximum.

### 3.3.4 Numerical Transformation

Before any transformation, the dataset is split into train and test datasets. This is to ensure that the datasets are transformed separately to prevent data leakage.

For numerical transformation, Power transformer was applied to all the numerical predictor features using feature engine. I choose to use this transformer because it can be applied to positive, negative, and zero values, unlike other transformation methods. Furthermore, when testing other transformation methods, the values were scaled to infinity which caused issues. Feature scaling was not applied as the values were already similar in scale and performing too many transformations can increase the complexity of the data, causing overfitting.



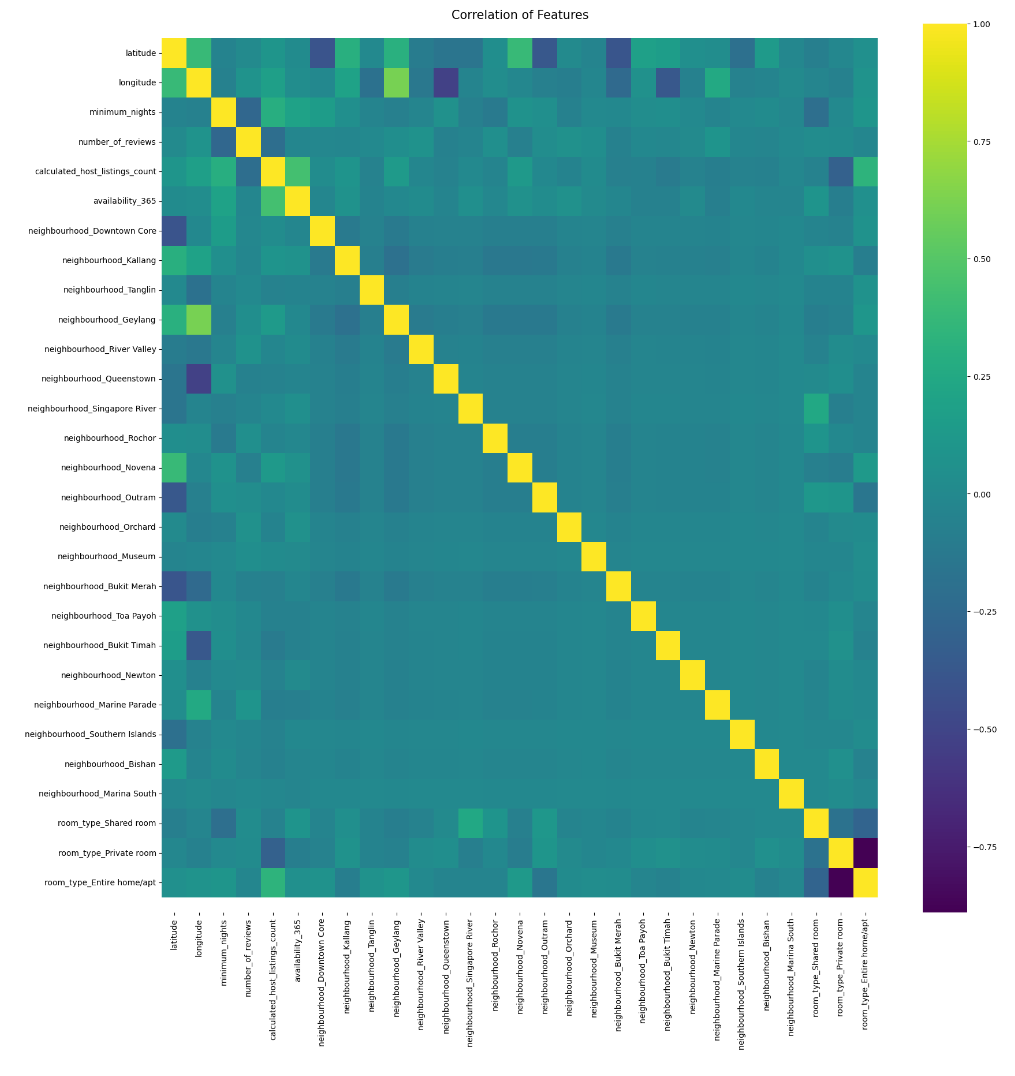
### 3.3.5 Categorical Transformation

One-Hot encoding, which converts categorical values into binary values of 0s and 1s, was employed for categorical encoding. By doing this, the model is prevented from assuming that greater values are more significant.

## 3.4 Correlation Analysis

Correlation Analysis is method used to determine and discover relationships between features. Once again, this has been done once during data exploration, before applying transformation to identify irrelevant features for removal.

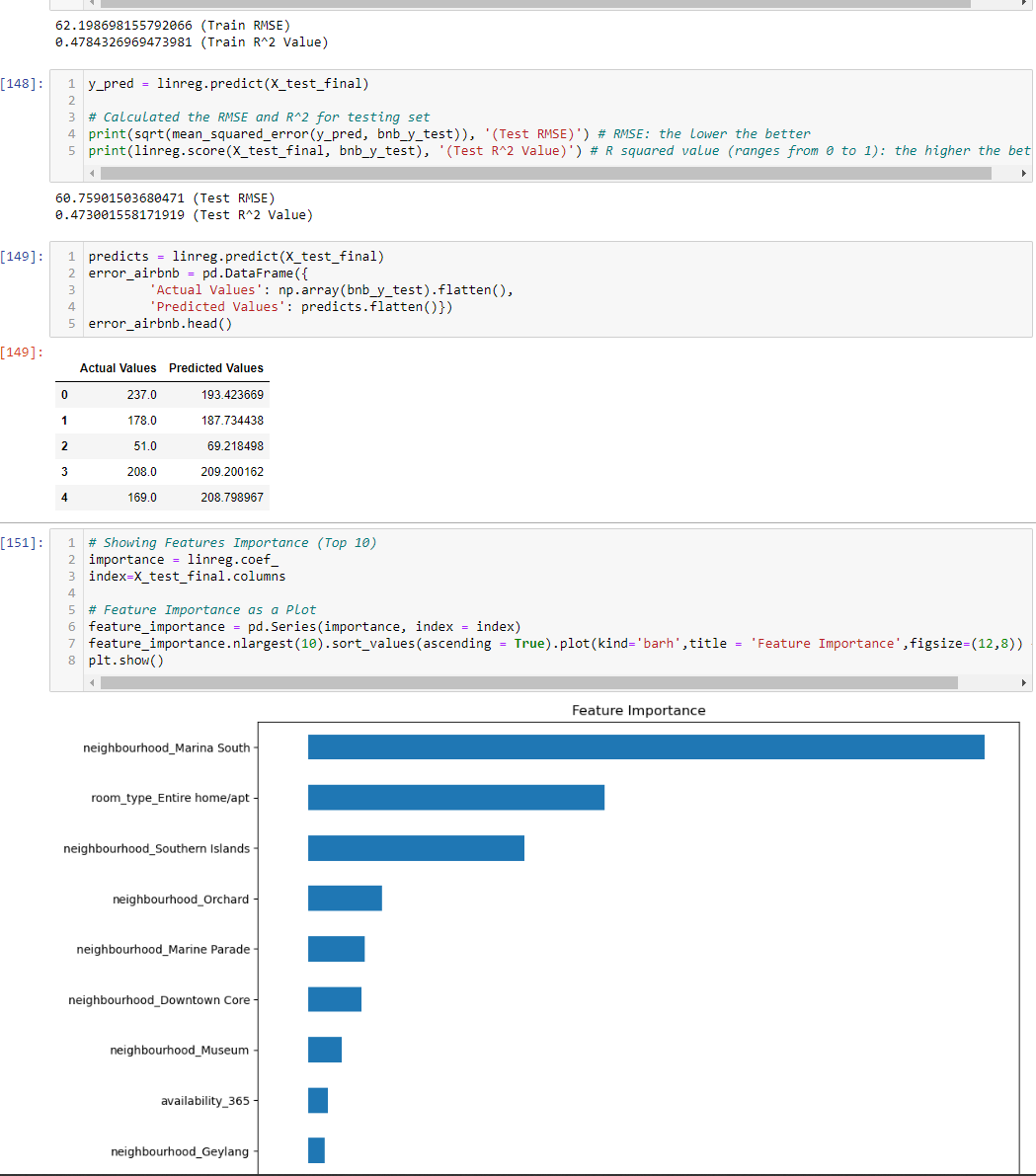
This heat map shows the correlation between the features after transformation and encoding. The target variable is not present in this heatmap as it was dropped during the train test split. Similar to the previous dataset, with a few notable exceptions, it is clear that many of the predictor features do not correlate with one another. We might want to remove these features for these exceptions to further optimise the training process and performance of our model.



## 3.5 Modelling

Once again, we will want to test the accuracy of our dataset on a model to evaluate the effectiveness of our transformations and encoding. The dataset is fitted onto a linear regression model.

A linear regression model would be the most appropriate because we are predicting the value of a numerical value, "price," and it predicts the value of the target variable based on specified predictor features. Using the metrics RMSE and R^2, we can evaluate the performance of our model. The RMSE and R^2 values of the model on the test data is 60.75901503680471and 0.473001558171919 respectively. This means that the model is able to explain around 47% of the variance in the model, making this model a mediocre predictor of prices. Further transformation or encoding has to be tested to improve the model’s performance.



# 4. Summary

This section will summarise this report’s findings, after conducting exploratory data analysis and performing data transformation on the two datasets.

For the hr\_data.csv dataset, the exploratory data analysis was able to yield useful information that assisted in defining the prediction problem statement as well as further optimizing the data wrangling and transformation of the data. The final features for prediction after transforming the data are, “no\_of\_trainings”, “age”, “previous\_year\_rating”, “KPIs\_met >80%”, “awards\_won?”, “avg\_training\_score”, “department”, “education”, “gender”, and “recruitment\_channel\_referred”. By testing the final transformed features onto a logistic regression model, the data was able to achieve an accuracy of 0.78, and the features with the highest importance are avg\_training\_score, department\_Sales & Marketing and KPIs\_met>80%.

For the listings.csv dataset, the exploratory analysis done helped to further narrow down the scope of our prediction, preventing the influence of extreme outlier values. Thus, allowing our model performance and results to be better for prediction. The final features for prediction after transforming the data are, “latitude”, “longitude”, “minimum\_nights”, “number\_of\_reviews”, “calculated\_host\_listings\_count”, “availability\_365”, “neighbourhood”, and “room\_type”. For the linear regression model, the model was able to achieve 60.75 RMSE and 0.47 R^2, which means our model is an adequate predictor for “price”. “neighbourhood\_Marina South”, “room\_type\_Entire home/apt”, and “neighbourhood\_Southern Islands” are the features that affect the value of price strongly.

# 5. Further Improvements

Improvements that can be considered to further improve the transformation of the two datasets include more research and more meaningful data exploration to deepen my understanding of the data.

Adding more data from other sources or creating new features that can further explain the target variable can also improve the data. Features such as whether an employee has leadership roles for the hr\_data or features of bad/good reviews for the listings data.

Lastly, testing other machine learning models should be explored, as the model used in the report is just a baseline model for evaluating the quality of the transformations. Other models such as Random Forest Classifiers or Support Vector Regression may be better suited for the data.