

**School of InfoComm Technology**

**Predictive Analytics**

Diploma in Financial Informatics (FI)

Diploma in Information Technology (IT)

YR 2/3 (2020/21), Semester 4/6

**INDIVIDUAL ASSIGNMENT 1**

(30% of Predictive Analytics Module)

**Deadline for Submission:**

**Presentation: Week 9, 14th – 18th Dec 2020**

**Report: 27th Dec 2020 (Sunday), 2359 Hours**

|  |  |  |
| --- | --- | --- |
| Student Name | : | Tan Zixu |
| Student Number | : | S10185168 |

**Penalty for late submission:**

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

**NO** submission will be accepted after 03rd Jan 2021, 23:59.

Contents

[Overview 3](#_Toc59974788)

[HR Analytics 4](#_Toc59974789)

[Problem Understanding 4](#_Toc59974790)

[Data Exploration 4](#_Toc59974791)

[General Information 4](#_Toc59974792)

[Data assessment & remarks 5](#_Toc59974793)

[Charts for HR columns 8](#_Toc59974794)

[Data Cleansing and Transformation 15](#_Toc59974795)

[Null Values 15](#_Toc59974796)

[Categorical to numerical data 16](#_Toc59974797)

[Column dropping 16](#_Toc59974798)

[Outlier 17](#_Toc59974799)

[Stratified sampling 17](#_Toc59974800)

[Correlation Analysis 18](#_Toc59974801)

[Airbnb Singapore 19](#_Toc59974802)

[Problem Understanding 19](#_Toc59974803)

[Data Exploration 19](#_Toc59974804)

[General information 19](#_Toc59974805)

[Data assessment & remarks 20](#_Toc59974806)

[Charts for Airbnb columns 24](#_Toc59974807)

[Data Cleansing and Transformation 30](#_Toc59974808)

[Null Values 30](#_Toc59974809)

[Outliers 31](#_Toc59974810)

[Region Locking 31](#_Toc59974811)

[Column Dropping 31](#_Toc59974812)

[Categorical to Numerical 32](#_Toc59974813)

[Correlation Analysis 33](#_Toc59974814)

[Summary and Further Improvement 34](#_Toc59974815)

# Overview

This is a report for Predictive Analytics Assignment 1. The report’s coverage includes two different topics, with one classification problem and one regression problem. The classification problem is on HR Analytic and a summary is to use predictive analytic to identify promotion candidates. The other problem is for Airbnb Singapore, with the goal to predict the rental price of listed properties. Both topics have provided a data set each, both of which will be explored and analysed in this assignment. For each topic, data exploration, cleansing, transformation and correlation analysis will be conducted. Findings on the datasets will be documented separately in their respective report, but datasets will undergo mostly similar steps with slightly differing action taken for the overall data preparation.

# HR Analytics

## Problem Understanding

Human Resource (HR) is a department that can be found in any business, the role of HR departments that manage employees and employee-related operations. One of the various operations that HR is responsible for is the promotion of employees within the company. However, analytics done to identify candidates for promotion can be rather laborious to process. Promotion screening is a process of great importance as HR cannot afford to hand out promotion to undeserving employees. The screening process also takes into account of various factors. Hence, HR needs to do extensive and stringent processing of key data in their analytics, placing a huge amount of constraint on the department. Which presents an opportunity to elevate the constraints on the HR by using predictive analytics to help identify potential promotion candidates.

## Data Exploration

To determine promotion candidates, understanding of each variable and data available in the dataset is vital as the first step of this analysis.

In this section, I will first list some general information about the dataset, followed by understanding each variable along with any notable finding and conclude the exploration with any initial thoughts on the data. To read the dataset, I created a data frame from reading the dataset that was in a csv file.

### General Information

Using .info() function on the data frame, I have found 54808 entries or rows across 14 different columns. Upon separating the columns based on their data type, 5 columns contain categorical data and the remaining 9 are numeric data. Counting the number of columns with the object data type, the identified categorical columns include department, region, education, gender and recruitment channel.

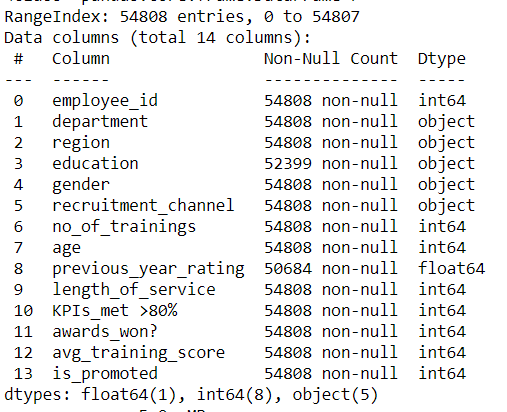


Figure . General Information of the Data Frame

### Data assessment & remarks

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Column Name | Data Type | Description | Remarks |
| 1 | employee\_id | Numerical | Employee’s unique id | This variable column role is to only uniquely identify each employee.  It does not serve any purpose for the correlation analysis nor the modelling in this classification problem. |
| 2 | department | Categorical | Department that the employee was from | First of the five categorical data columns, there are 9 different departments in this data frame.  The column could be a biased factor to determining which employee gets to be promoted. As the number of employees is not distributed evenly among each department.  Referring to Figure 2, we can see a drastically different number of employees working in various departments. Hence, a department could have more promotion candidates due to the bigger pool of the manpower in it. |
| 3 | region | Categorical | Region of the employee | The second categorical data column with the regions labelled as “region\_<number>”, the number spans from 1 to 34.  Similar to department, the number of employees could be unevenly distributed in particular regions, making it a biased factor.  As seen in Figure 3, some of the regions have a concentrated number of employees while others have a relatively low number in contrast. |
| 4 | education | Categorical | Education level of the employee | The third categorical data column with three different categories of education level.  While education level could be a determining factor for starting pay, it would be different in the context of the problem. This is because if the promotion criteria are based on actual performance, past qualification would be meaningless to the promotion screening. Further analysis will be able to affirm the actual performance assumption by proving education has a weaker correlation with the target.  Referring to Figure 4, most of the employees have at least a bachelor’s degree. I think phrasing for “below secondary” is a bit off. As the intention seems to be to grouping people that classify as secondary & below. But the original does not seem to include secondary. |
| 5 | gender | Categorical | Gender of the employee | The fourth categorical data column that strictly identifies the employees as either male or female. Referring to Figure 5, the data frame consists of more male over female employees. |
| 6 | recruitment\_channel | Categorical | Channel the employee was recruited from | The last categorical data column that represents the three recruitment channels an employee could come from.  Referring to Figure 6, most of the employees were either from sourcing or other channels. Only a selected few in the context of this data frame were referred by the company. |
| 7 | no\_of\_trainings | Numerical | Number of training the employee completed | This data column represented the number of training the employee had with the company. All employees have undergone at least 1 training within their service under the company.  Referring to Figure 7, the mode is at 1. It would seem that only a handful of employees are given the opportunity to undergo more than 1 training. |
| 8 | age | Numerical | Age of the employee | This data column reflected the age of the employee.  The age of an employee could represent various things such as the experiences the physical capability of the worker. There is no definite significance that can be derived from age to be used to determine promotion candidate. However, this could be proven right or wrong after the affirmation on the actual performance assumption.  Referring to Figure 8, the age range is a bit diverse considering there are outliers. |
| 9 | previous\_year\_rating | Numerical | Previous year rating of the employee | This data column seems to reflect the employee’s rating of previous year work performance from a scale of 1 to 5.  This could be a strong indicator for identifying promotion candidate. As this data reflects the historical performance of the employee, representing actual performance in the past. It is likely that the consistency of an employee is also evaluated as part of the promotion criteria.  Referring to Figure 9, the sizable number of employees scored 4 or 5 in this column. These employees are likely to be reviewed for promotion. |
| 10 | length\_of\_service | Numerical | How long the employee has served | This data column reflected the number of years that the employee has served.  Following up on the remark in the “age” column, by committing a long time to serve a company does not seem like a strong reason to promote an employee. It sounds unreasonable and it is not proven until correlation analysis is done.  Referring to Figure 10, there are a lot more outliers in this length\_of\_service as compared to age. It could mean that there are a handful of employees that dedicated their lives to the career within this company. |
| 11 | KPIs\_met>80% | Numerical | If an employee’s key performance index (KPI) percent is greater than 80% | This data column indicates if an employee’s key performance index is above a threshold of 80%.  This is like a strong indicator to identify promotion candidate. As it reflects the actual performance of the employee within the current year.  Referring to Figure 11, it is estimated that only one-third of the employees had met the KPI condition. |
| 12 | award\_won? | Numerical | If an employee won an award | This data column indicates if the employee has won an award. It is unknown if the indicator is before the screening or it is a performance-related award.  Referring to Figure 12, there are only a few employees that had won an award. |
| 13 | avg\_training\_score | Numerical | Employee’s average training score | This data column represents the training score of the employee on average.  Referring to Figure 13, the majority of the employee scored over the 50 percentiles. |
| 14 | is\_promoted | Numerical | Recommended as a promotion candidate | This data column is the targeted variable for the prediction, it reflects whether an employee is recommended for promotion.  Referring to Figure 14, there are more people who are not recommended for promotion. Stratified sampling might need to be deployed to ensure there is an evaluation for other variables in the correlation analysis. |

### Charts for HR columns

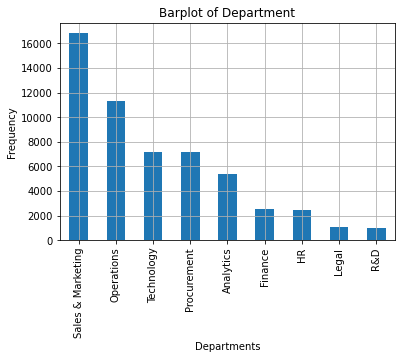


Figure . Bar chart of Department

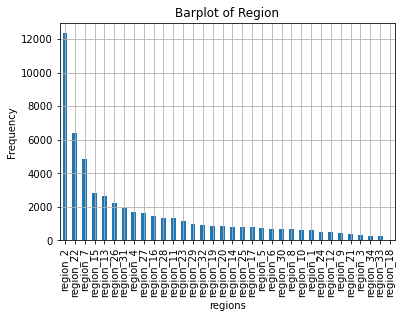


Figure . Bar chart of Region

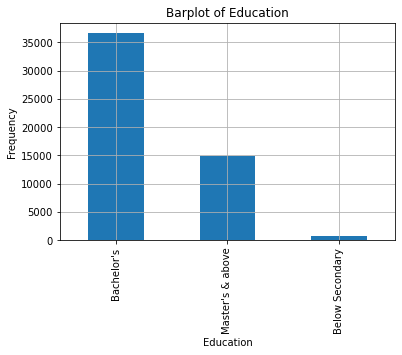


Figure . Bar chart of Education

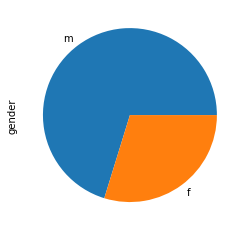


Figure . Pie chart of Gender

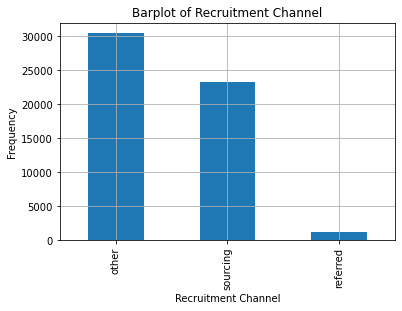


Figure . Bar chart of Recruitment Channel

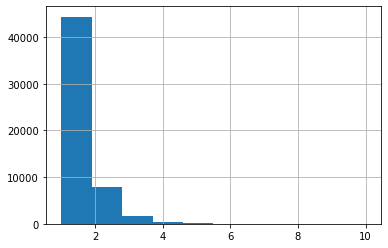


Figure . Histogram of Number of Training

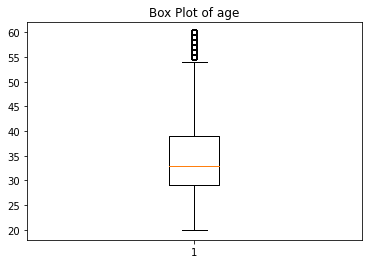


Figure . Box plot of Age

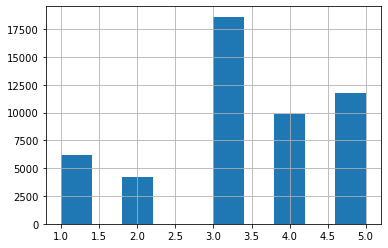


Figure . Histogram of Previous Year Rating

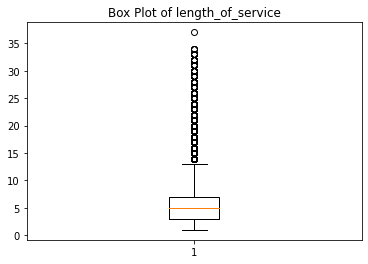


Figure . Box plot of Length of Service

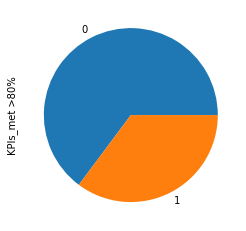


Figure . Pie chart of KPIs Met > 80%

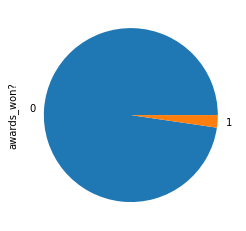


Figure . Pie chart of Awards Won?

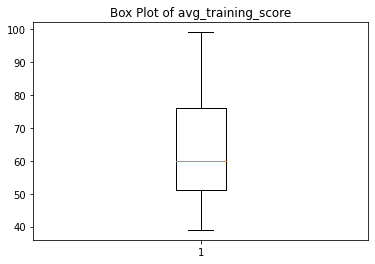


Figure . Box plot of Average Training Score

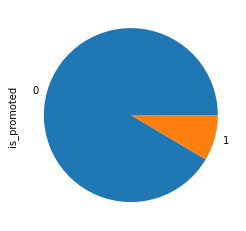


Figure . Pie chart of Is Promoted

## Data Cleansing and Transformation

### Null Values

With further inspection, there are 2 columns with missing values namely the education and previous year ratings. By using the .isnull() and .sum() function, I have identified the numbers of missing values at 2409 for education and 4124 for previous year rating. Null values will be either dropped or transformed in this section. However, it is preferred to keep dropping off any data rows to a minimum as it would be best to avoid limit the size of the data frame.

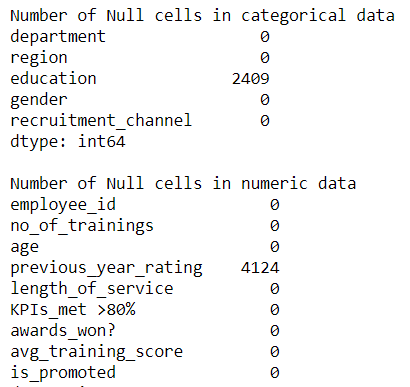


Figure . Number of NULL values

For the education column, null values will be transformed into one of the three categories in this data column. In this case, a possible course of action is to replace the null values with the mode value, which is “Bachelor’s”. However, I think it is unreasonable to give out bachelor’s easily, it would make more sense to assign the “below secondary”. As that category covers employee with an education level below secondary, making it justified to assigned at least the bare minimum. Hence, null values in education column will be assigned “below secondary” as a mean to clean the data.

For the previous\_year\_rating column, rows with a null value will, unfortunately, have to be dropped. Reason being the column itself is set context that it is for employees who have been in service since the previous year. Meaning that employees with a null value in this column are likely to be new recruits and it is not appropriate to assign a value from the scale or a new one. Thus, the null values in previous\_year\_rating will be dropped from the data frame.

### Categorical to numerical data

To do correlation analysis, the categorical data columns have to be transformed into numerical data columns. This transformation phase may differ for each data column based on the context to determine the appropriate changes.

For the gender column, the values are mapped to binary outputs in this data transformation phase. As the column outputs only contain male or female, the output will be mapped to reflect 0s and 1s after the transformation.



For the education column, the data will be mapped with ordinal rankings based on the education level. Such that education level will be reflected as numbers in ascending order, where the higher number would represent high education level in this ordinal scale. For example, employees with “Below Secondary” will be assigned “1”, “Bachelor’s” to “2” and “Master’s & above” to “3”.



For the department and recruitment\_channel column, the values will be mapped to numbers that correspond with the number of unique values in their columns. Such that department values will be mapped from 1 to 9 and recruitment\_channel will be from 1 to 3. The mapping for the department was done differently from recruitment\_channel. Due to space constraint, a similar picture of the code and thought process can be found in the Airbnb report for the mapping of the neighbourhood.



For the region column, a different way to map the values in the column will be deployed. As the naming of each region is already following a format and assigned a number as part of a string. The transformation in this column will use the str.split() function. By separating the string at “\_”, we will have the word “region” and whatever number that was attached to it. Keeping the number and transforming the column as an integer data type, the region column is now a numerical data column.



### Column dropping

For column dropping, there is only one data column to be dropped, “employee\_id” column. Reason being the column serves as only an identification of the employee, serving no purpose towards the identifying potential promotion candidate. One does not simply look at the employee id and blindly choose the employee as the promotion candidate. Hence, this column was deemed useless and to be dropped during the data transformation phase.

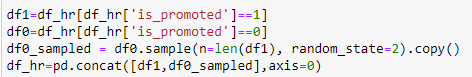


### Outlier

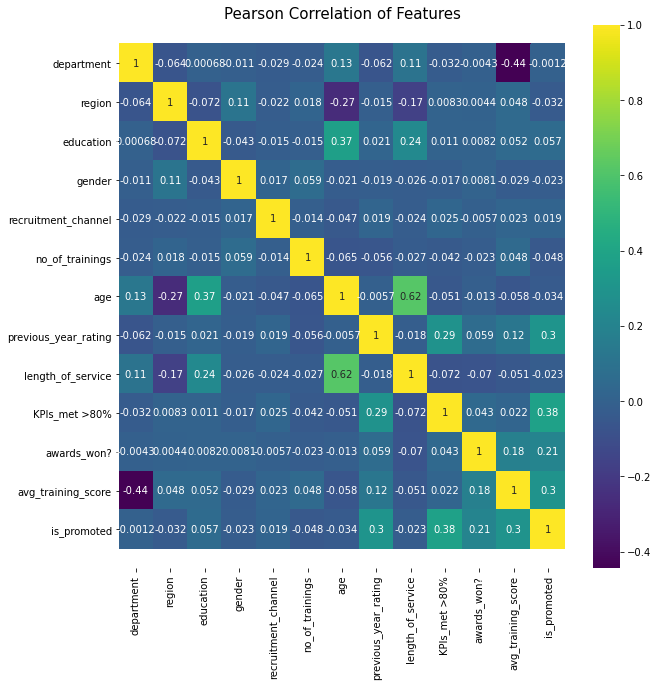
As mentioned during the assessment of each data column in the data exploration above, there are indeed outliers in a couple of data columns. However, age and length\_of\_service is not a reliable indicator for promotion screening. Such that being the oldest or longest-serving employee would not guarantee a promotion. Hence, it is determined that outliers in both of these columns would not influence the correlation with the target.

### Stratified sampling

This is a necessary procedure for this data transformation. As the problem is of a binary classification issue, the target column needs to have an even amount of both responses for a fair evaluation of the other column during correlation analysis. Hence, stratified sampling was deemed necessary to eliminate any biases during the correlation analysis.



## Correlation Analysis



This is the correlation heatmap of this data frame after data cleansing & transformation. Narrowing the focus towards the target column, the range of scores between the other variables and the target is between -0.048 to 0.38. From this heatmap alone, we can see that although 0.38 is not a strong correlation score, it is relatively high under the current context with the other variables.

In general, variables related to actual performance such as “KPIs\_met>80%”, “previous\_year\_rating” and “awards\_won?” seems to have a better score as compared to the other variables. With an exception of “avg\_training\_score”, this column did better than expected for the correlation with the target.

To summarize the findings, variables that are related to actual work performances are more closely correlated to the being recommended for promotion. Whereas variables that provide basic information on the employee scored poorly. Meaning that regardless of the employee’s background, qualification or affiliation, promotion recommendation is only for employees who have worked hard and produced results. Additional findings include that consistency of an employee may be another key indicator based on how “avg\_training\_score” correlation score was with the target column.

# Airbnb Singapore

## Problem Understanding

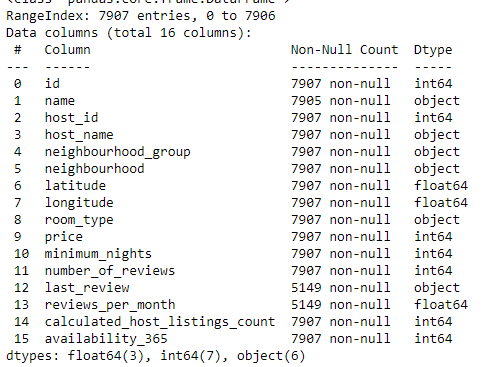
Airbnb is widely known as an online platform that connects travellers with host looking to rent out their properties. It is a popular tool for travellers or organisers that are looking for accommodation during their trip. While providing an opportunity for travellers to have a unique and personalized experience, the host gets to earn some rental revenue, making it a win-win situation for both travellers and hosts alike. In this problem, the task is to predict the rental price of the listing in Airbnb Singapore.

## Data Exploration

Similar to what was done in the HR analytic problem, I will start with some basic data exploration. Which includes general findings of the dataset and assessment of each data column to understand the data that is being worked with for this problem. With that goal in mind, the dataset is read from a csv file and change to a data frame for analysis.

### General information

Using .info() function on the data frame, 7907 entries across 16 different columns are found. Upon separating the columns based on their data type, there are 6 columns that contain categorical data and the remaining 10 are numeric data. Counting the number of columns with the object data type, the identified categorical columns include id, name, host\_id, host\_name, neighbourhood\_group, neighbourhood, room\_type and last\_review.



### Data assessment & remarks

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Column Name | Data Type | Description | Remarks |
| 1 | id | Numerical | Listing’s unique id | This variable column role is to only uniquely identify each listing.  It does not serve any purpose for the correlation analysis nor the modelling in this regression problem. |
| 2 | name | Categorical | Listing’s name | This variable column role is to only identify each listing using a name.  It does not serve any purpose for the correlation analysis nor the modelling in this regression problem. |
| 3 | host\_id | Numerical | Listing host’s unique id | This variable column role is to only uniquely identify the listing’s host.  It does not serve any purpose for the correlation analysis nor the modelling in this regression problem. |
| 4 | Host\_name | Categorical | Listing host’s name | This variable column role is to only identify the listing’s host name.  It does not serve any purpose for the correlation analysis nor the modelling in this regression problem. |
| 5 | Neighbourhood\_group | Categorical | Region of the listing | This data column represented where region each listing is based at. The column has classified Singapore into 5 different regions.  With listing being localised to the different regions, it might be better to focus on a specific region for the analysis. Having an area centric analysis will remove any biases in the form of any region-based dynamic pricing. Such that a particular region will have a higher price over the others based on the location.  Referring to Figure 16, the majority of the listings are in the central region, which will likely be the focus of this analysis. |
| 6 | neighbourhood | Categorical | The sub-region of the listing | This data column showed where the listing is specifically located at, in terms of sub-region.  The sub region of the listing seems like an unlikely strong variable that correlates with the target. As these sub-regions would still be generalised to the same region. That is if the analysis is based on 1 region.  Referring to Figure 17, there is not much to say other than Kallang holding the greatest number of listings. |
| 7 | latitude | Numerical | Latitude coordinate | This data column represents one of the two coordinate values of the listing. Which is will be used with longitude to find the exact geographical coordinates of the listing.  There isn’t much to do with this data column and it is uncertain if it can be used for any analysis purpose. |
| 8 | longitude | Numerical | Longitude coordinate | This data column represents one of the two coordinate values of the listing. Which is will be used with latitude to find the exact geographical coordinates of the listing.  Similar to latitude, there are uncertainties surrounding this data column. However, there was an interesting find. When this column is used to create a scatterplot with the latitude, it creates a rough sketch of the Singapore map. Refer to Figure 18, to see the map. |
| 9 | Room\_type | Categorical | Listing’s space type | This data column represents the type of sizes for each listing. Which are categorized into “Entire home/apt”, “Private room” and “Shared room”.  This is likely a strong variable that helps to determine the rental price. It would make sense that the rental price might scale with the size of the property.  Referring to Figure 19, the majority of the listings are offering an entire home/apt or private room. There are only a handful number of listings that offers shared rooms. |
| 10 | price | Numerical | Daily rental price of the listing | This data column is the targeted variable for prediction, it reflects the rental price of the listing.  Referring to Figure 20, there are a lot of outliers, with one instance of a listing priced at $10000. Changes need to be made for this data column. As it could affect the analysis and eventual modelling for this target column. |
| 11 | minimum\_nights | Numerical | Number of nights at a minimum | This data column represents the number of nights the guest needs to stay at a minimum.  While the minimum night is unlikely a direct variable that determines the price. I think the minimum nights will scale inversely with the price. Where higher the minimum night is, the lower the price is.  Referring to Figure 21, there a lot of instances that are outliers. It seems to have more outliers than price. |
| 12 | number\_of\_reviews | Numerical | Number of listing’s review | This data column reflects the total number of reviews that the listings have.  If the number of reviews can be interpreted as the popularity of the listing. It may be equivalent to the minimum number of hosting it has carried out. However, the number of reviews does not specify good or bad reviews. Hence, it is uncertain if the number of reviews affects the price but it is unlikely.  Referring to Figure 22, there are a lot of outliers in this column. An inference can be made that this column is biased or unfair. As there are listings with no reviews, this column is weakening as a possible variable that correlates with the target. |
| 13 | last\_review | Categorical | When the latest review was made | This data column reflects the total number of reviews that the listings have.  From the general information, there are already some missing entries. It is likely the missing entries are null value, whereas the other cell contains a value.  Using the .describe() function, it is found that the value in the cells only contain dates. With this information, I can infer that entries with null value are likely to new listings. Although, it is uncertain how it will correlate with the target. Perhaps more affirmation can be found after the correlation analysis. |
| 14 | review\_per\_month | Numerical | Number of reviews for each month on average | This data column reflects the number of reviews but it is an average value by the number of the months since the listing was made.  Similar to the number of review column, there are uncertainties surrounding this variable. Referring to Figure 23, a large number of outliers found with the boxplot. |
| 15 | calculated\_host\_listings\_count | Numerical | Number of other listings the host has | This data column shows the number of other listings the host have.  I think that regardless of how many other listings the host has, it does not necessarily determine the rental price. Suggesting that this column may not be a strong variable. Hence, making a note for modelling to try dropping.  Referring to Figure 24, there only a few outliers found in this boxplot. An interesting find is that some hosts have more than 1 listing. Such that there are players thriving in this accommodation market by investing in a few properties. While there really is no monopoly, it seems to have some competition. |
| 16 | availability\_365 | Numerical | Number of days the listing is available for throughout the year | This data column reflects the availability of the hosting in terms of days throughout the year long.  This is a column I suspect to have a strong influence over the pricing. As it can be seen as a limitation put on the listing. Meaning that the price could vary based on availability. However, there is no affirmation or any context given about this possibility. Hence, leaving it as an assumption till the correlation analysis.  Referring to Figure 25, it seems there are no dirty data. All entries seem to be within the limit of 365 days. |

### Charts for Airbnb columns

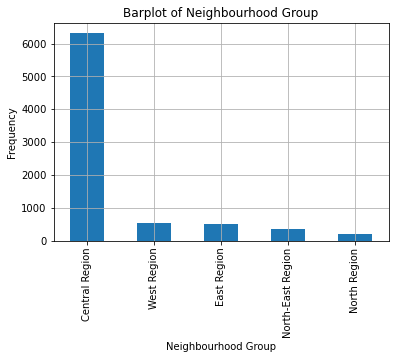


Figure . Bar chart of Neighbourhood Group

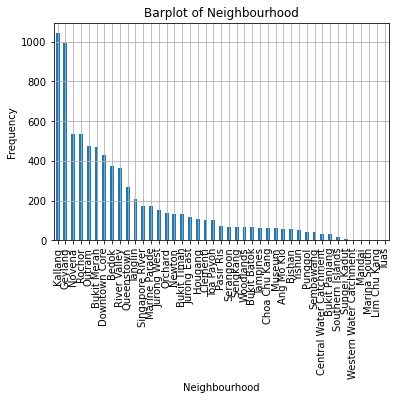


Figure . Bar chart of Neighbourhood

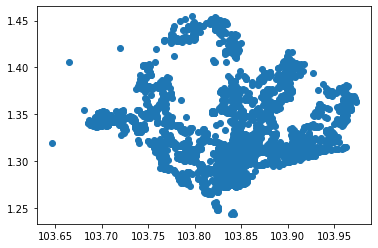


Figure . Scatter plot of Latitude & Longitude

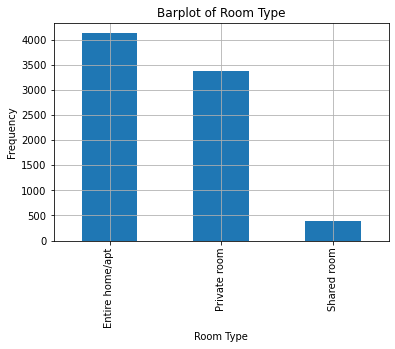


Figure . Bar chart of Room Type

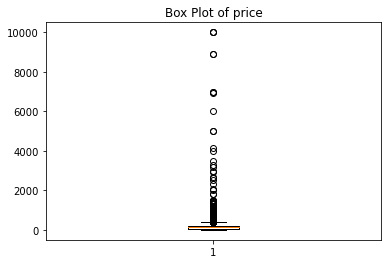


Figure . Box plot of Price

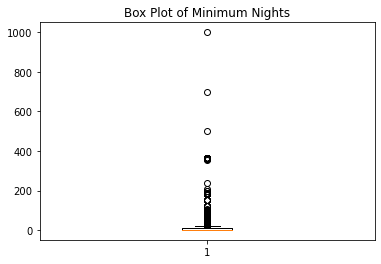


Figure . Box plot od Minimum Nights

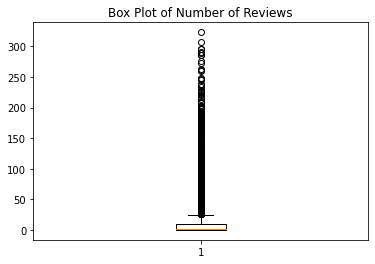


Figure . Box plot of Number of Reviews

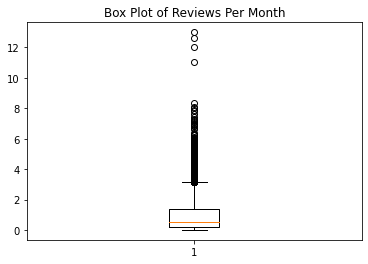


Figure . Box plot of Reviews Per Month

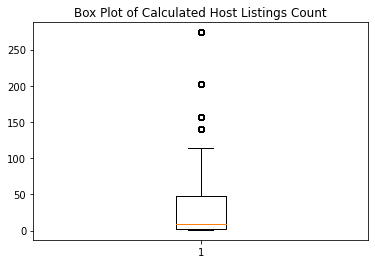


Figure . Box plot of Calculated Host Listings Count

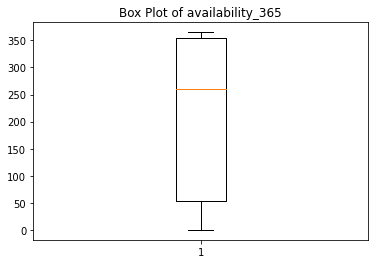
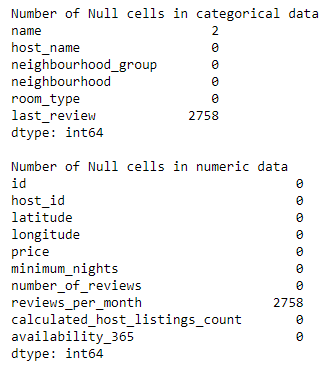


Figure . Box plot of Availability 365

## Data Cleansing and Transformation

### Null Values

Digging deeper into the data frame, two categorical and one numerical data column was found to have null values within them. In the name column, there are two entries with no empty cells in the column for those listings. For both reviews\_per\_month and last\_review, there were 2578 null values with those columns. Null values will be either dropped or transformed in this section. However, it is preferred to keep dropping off any data rows to a minimum as it would be best to avoid limit the size of the data frame.



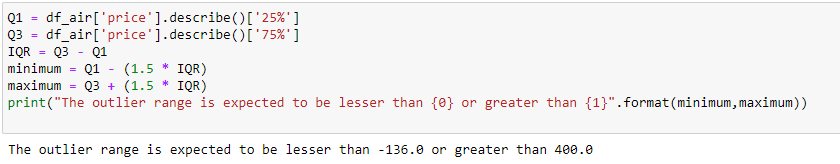
For name, the null entries will be dropped as the most appropriate choice at this point. There is justification in providing a temporary name for the analysis. However, removing the null values at that point of transformation had a priority over assigning values. The reason was also partly that I wanted to refrain from interacting with the data too much. Reflecting back at this decision, it would not have mattered too much since there were only two null values. But I do acknowledge there was another way to handle this transformation situation and I will look to apply the alternative methods for future data transformation.

For last\_review, the action taken would be to assign “0” to null values and at the same time transform the dates to “1”. This is because the context of this column is suggesting that it is an indicator if the listing had any reviews. If we look at the number of null values in this column and the reviews\_per\_month, both had the same amount. Implying that null entries were new listings, acting as an indicator based on the existence of input in that column. Hence, the values were transformed regardless of whether it is null.

For reviews\_per\_month, the null values will be assigned with a value, “0”. The decision is based on a similar line of thought that is found in the null value transformation of education column in the HR data set. To retain as much data entries as possible, a value that is safe and reasonable should be assigned. Hence, the minimum value will be used when dealing with the null values in this data column.

### Outliers

For outliers, it was found in many columns as seen in the data assessment above. The frequency of outliers appearing varies between the columns, but only one column stood out that require necessary action to be dealt with. The price column is set to have its entries with null values to be dropped. As it is the target column in this problem, the outliers found in here would affect the correlation analysis and the eventual predictive modelling. It is hence determined to be removed as measures against any internal influence from the data column itself. With the action taking place, the outlier range was found for the price column.



### Region Locking

This section was mentioned in the data assessment, with the analysis moving towards more area focused. The central region will be chosen as the focus as it holds the majority of the listing at around 75% of the data size. It is unfortunate that more than 20% of the data will be dropped in this section but it will be worth to cover the popular region. As it will provide meaningful information for various audiences to learn more about what is going on in the Airbnb market of a region that has the most listings.



### Column Dropping

In this section, there will be a lot more columns being dropped before the correlation analysis as compared to the HR report. There will be 5 different columns being dropped in total namely the id, name, host\_id, host\_name and neighbourhood\_group.

The first four columns that were listed will be dropped due to similar reasons found in the dropping of employee\_id in the HR data transformation. Where all the columns serve no purpose towards the analysis and the eventual modelling. These columns will be dubbed as useless variables.



As for neighbourhood\_group, the column itself was rendered redundant after locking the region, limiting it to only the central region. This effectively made all entries from the central region and the column will no longer serve any purpose. Hence, it is dropped in this section.



Lastly, a side note will be made about latitude and longitude. Both of these columns on the surface seems to serve no purpose like the useless variables. However, there are some uncertainties about their uses, as not much can be found at this point. Hence, these columns were spared and not dropped like the others.

### Categorical to Numerical

In this section, the remaining categorical data will be transformed into numerical data. It will be done by mapping categorical values with numeric values, in a similar fashion of the HR report. This is a necessary step to enable correlation analysis in the next step. The remaining two data columns will be transformed using mapping function and an automated loop that assign labels in ascending order.

For the neighbourhood, there 19 different sub-regions that are within the central region. Instead of finding the value count of the neighbourhood and manually mapped the values in ascending order, I created a couple of loops. First, I found the value count in ascending order and store it as a separate data frame to record the position of each neighbourhood in the order. Next, I create the new intended order list based on the length of the separated data frame. Lastly, I will go through the neighbourhood column in the original data frame, do some matching and map the appropriate value.

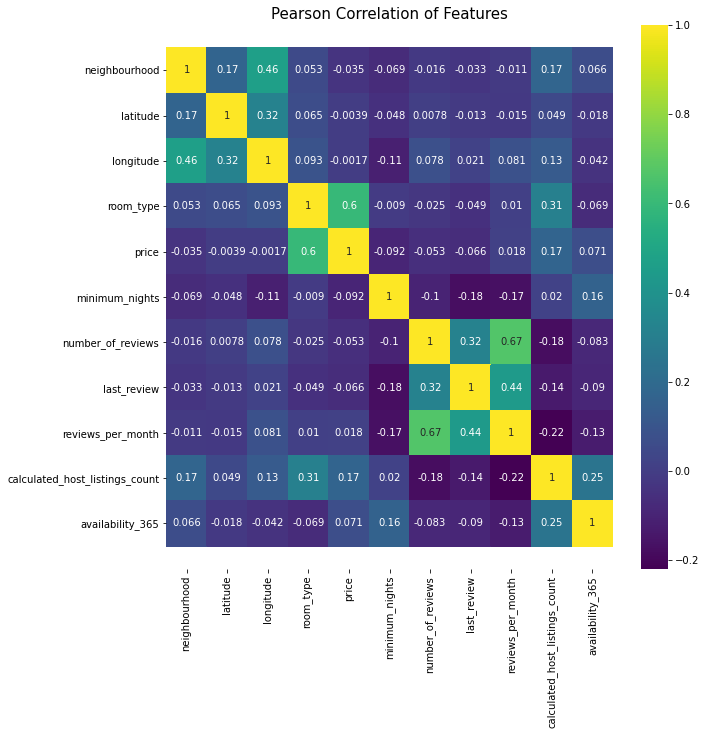


After testing around, the approach I made was found to be very strenuous on the system especially with big data frame or columns with a lot of unique values. While this approach was more inefficient as compared to the already built-in mapping function, it was an interesting experience to try out different approaches and I had fun creating this. Thus, I left it in and committed to this method when doing the mapping. This method was also applied in the mapping of the department column in the HR report.

For room\_type, I kept to the traditional mapping function since it only had three unique values. The mapping was made to be ascending as well.



## Correlation Analysis



This is the correlation heatmap of this data frame after data cleansing & transformation. Narrowing the focus towards the target column, the range of scores between the other variables and the target is between -0.092 to 0.6. From this heatmap alone, it was refreshing to see a high score of 0.6, but the other variables had a relatively poor score across the data columns.

The data column with the highest score was room\_type but the gap between it and other variables was abysmally huge. From this, we can affirm that on the room\_type is the only variable that holds a strong influence over determining the price. Other affirmation includes that minimum\_nights inversely scales with the price, as it had the highest negative correlation with the target. However, it was not a strong negative correlation under the context that -0.092 is extremely low. As for surprising finds, availability\_365’s score was disappointing as I had an expectation it might be a variable with a decent correlation with the target.

# Summary and Further Improvement

This section will be covering the summary of the report from each problem. The content will include objective findings of each data set and any interesting findings or notable inferences. Lastly, further improvement will be generalised at the end for both data set.

In HR Analytics problem, a brief recap of the goal is to identify promotion candidates. In this problem, various analysis has been done to understand the data that is being worked with. In the pre-assessment of the data set, an assumption was made that promotion candidates are recommended for their exemplary work performance. That variable such as the employee’s background or qualification should not a strong determining factor. Hence, data columns were identified and classified into a weak or strong variable before the analysis.

It was a preliminary assessment to get a general understanding of the dataset and what it offers. After the correlation analysis was made, some objective findings were found that the variable that strongly correlates with the targets were identified. It had given affirmation of the assumption made as variables like KPIs\_met>80% and previous\_year\_rating had a better correlation across the dataset. So, more inferences were made such that the promotion practices meritocracy, where employees are rewarded for their efforts in their performances. Another inference made was that consistency also might be a factor in choosing candidates. As the avg\_training\_score column had a decent correlation with the target. Meaning that employees who were recommended are consistent in their work ethic regardless if it was training or actual work. This couple of inferences were an interesting find and there is a better understanding of the problem, where the general idea of what the HR wants in the identified candidate.

For the Airbnb Singapore problem, the objective was to predict the rental price of the listed properties. A similar preliminary assessment was also done to identify variables that are likely strong in determining the price. Only 2 columns were identified in room\_type and availability\_365 to be strong determining factors. A scenario or assumption was also given some thoughts, where the minimum\_nights would inversely scale with the price. Such that the higher number of minimum nights the guest has to stay would result in lower price, noting that the price is the daily rate.

After the correlation analysis, some thoughts were proven to be true or false. Firstly, the room\_type had a really strong correlation with the target as the sole highest correlated column. Next, availability\_365 was not proven to be a strong indicator. Which was surprising to me, considering that the thought of availability would affect the price was something that I was fairly certain. Finally, the inverse scaling was proven to be true. The minimum\_nights and price had a negative correlation, proving the inverse scaling although the score was not big in any context.

Another point to summarise for both problems is that the handling of each data set can be unique and different. Despite having a similar process being applied for data preparation in both datasets, the transformation phase is still dependent on the situation. There are standard steps for the transformation phase such as removing null values or changing categorical to numerical data. However, the techniques apply may differ and based on the context, the dataset could have additional steps. An example is the stratified sampling applied in the HR Analytics problem and the region locking for Airbnb Singapore.

As for further improvements, both data sets may need another round of assessment and consider a few more scenario to develop a higher understanding of the data. As it is now, the understanding is still shallow in some ways and further column dropping might be needed.