# Data Mining final Report
# Datasets: Bank Client Attributes and AIRBNB Property Sales

Data Mining BUAN201

Prof: Swapnajit Chakraborti and Deepak Srivastav

Project by: Kunth Shah

Due Date: 25th April 2024

# Bank Client Attributes and Marketing Outcomes

*Report by: Kunth Shah*

**Research Aim:** This part of the report aims to understand the relationship between multiple attributes of specific Clients and analyse whether they would accept the marketing campaign or not.

**Objective and Application:** Stakeholders may use the following analysis performed in this report to target specific customers with these attributes with their marketing campaigns, to reduces costs and maximize acceptance of the campaigns.

## Data Set Explanation

**Link for Dataset:**
https://www.kaggle.com/datasets/ara001/bank-client-attributes-and-marketing-outcomes

This is a categorical dataset with 18 attributes. The specifics include: 5 categorical (more than 2 categories), 4 binary (yes - no), 8 numerical and 1 Unique ID, type of attributes.

Independent variables to be used: Day, Education, Housing, Job, loan, default, marital, age, duration, campaign and acceptance percentage (New field created) against the dependent variable 'y'. This variable y in the dataset reveals whether the customer accepted("yes") or rejected("no") the marketing campaign.

## Process and approach

The dataset was first cleaned using python code and pandas library to remove null values. Then R language in the IDE "R Studio" was used for performing a logistic regression model, which was then tested and followed by creating-testing for predictions. Lastly Tableau was used to perform analysis on some of the remaining variables (Majorly other categorical attributes) and for visualising trends to obtain insights.

## Data Cleaning

The data was cleaned using the python code "Null Data Remover", this code checked for each row and removed all rows which had a single null instance for any attribute.

## Exploratory Analysis using Tableau

Tableau was used to plot general trends between numerical values and the variable – y. This exploration revealed some positive correlations between balance, age, day, and duration with the variable y. The research was then moved to R, to perform specific analysis and understanding of this numerical fields. Tableau would later be used for data visualization.

## Exploratory Analysis using R

The exploration of data was done using R by plotting box-whisker graphs among multiple variables. This revealed that there were multiple outliers, within this data set and required cleaning.

## Data Cleaning and selection in R

The first part of the code required me to convert the values which were being recognized as characters to factors. Over here I used the *factor* function and specified *levels* so that the data can be interpreted as binary values.

The next step involved plotting the box-whisker graphs. After plotting these graphs, it was inferred that the data lacked consistency and involved a lot of outliers.

**Addressing Outliers:**
The data was filtered for outliers, using functions such as *boxplot.stats* to extract the outliers and then strip the original dataset of these values. Once this was done, the data was re-plotted, which generated better graphs.

**Example Changes:**
This is a demo of the change in output which was noticed in the box-whisker graph, pre and post outliers cleaning.

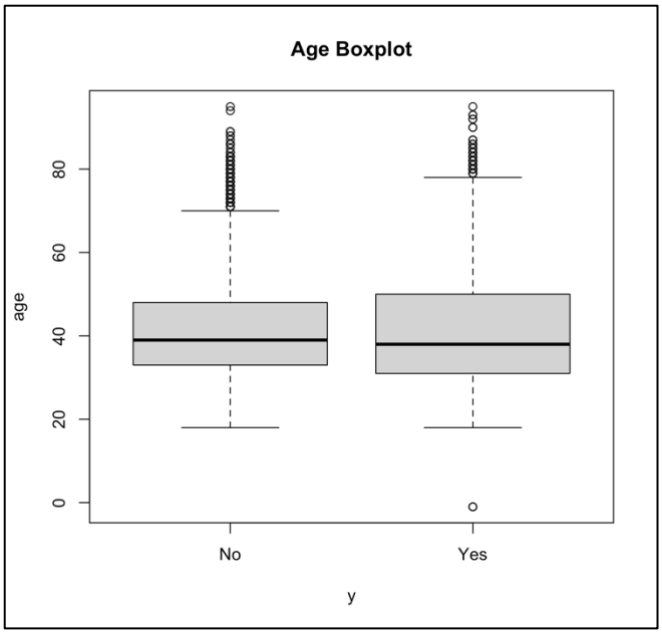


*Figure 1 Age Graph before Outlier removal*

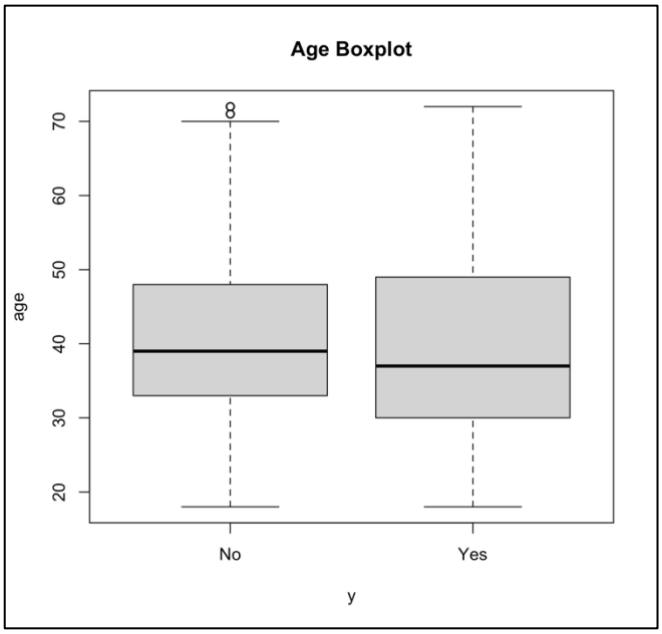


*Figure 2 Age After Outlier removal*

This exact same process was repeated for other variables: Balance, duration and day.

## Developing Generalized Linear Models

After the dataset was fixed with the better values, the next part of the process involved developing generalized linear models to observe the strength of correlation between y and all other variables.

**Summary of GLM models:**

| Variable | Strength of Correlation | Pr(>\|z\|) | AIC |
|---|---|---|---|
| balance | 99.99 | <2e-16 | 20370 |
| duration | 99.99 | <2e-16 | 18493 |
| day | 99.99 | 4.16e-14 | 20589 |
| age | 90 | 0.0999 | 20644 |

After conducting this research, a drill down approach was taken to form the prediction model. The chosen model was best fitted with all these values and was chosen for final analysis, as this had the least AIC value and gave the best predicts seen later.

**Final Model Summary:**

Coefficients:

| | Estimate | Std. Error | z value | Pr(>\|z\|) |
|---|---|---|---|---|
| (Intercept) | -3.788e+00 | 9.911e-02 | -38.218 | < 2e-16 *** |
| balance | 3.534e-04 | 2.125e-05 | 16.627 | < 2e-16 *** |
| duration | 5.659e-03 | 1.235e-04 | 45.836 | < 2e-16 *** |
| day | -1.479e-02 | 2.437e-03 | -6.068 | 1.29e-09 *** |
| age | -3.257e-03 | 1.991e-03 | -1.636 | 0.102 |

---

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 20642 on 37287 degrees of freedom
Residual deviance: 18193 on 37283 degrees of freedom
AIC: 18203

Number of Fisher Scoring iterations: 6

**The Hosmer-Lemeshow goodness-of-fit test**

This test was performed on the model, and it revealed the following data:

Statistic = 179.2898
degrees of freedom = 8
p-value = < 2.22e-16

This indicates that the model is very well fitted with negligible p-value and a strong statistical score.

## Prediction using the model

The model was used for prediction by first creating a contrast for “Yes” and “No” values in the model. Then an empty vector was created with 10,000 values of “No”, only for those where the

threshold was more than 0.5, the values would be put as yes.

This model revealed the following confusion matrix:

| | Actual Values | |
|---|---|---|
| **Predicted Values** | No | Yes |
| No | 9887 | 94 |
| Yes | 58 | 38 |

The total error rate was only 0.0152. However, we want the model to make as low predictions for the False Positive as possible, where the error rate for that was only 0.00583. This indicates that the model is well fit for predicting the values.

Lastly the ROC curve was plotted:

The ROC curve is plotted by using the best threshold using the *coords* function. Once this threshold was found, this was used to plot the graph.

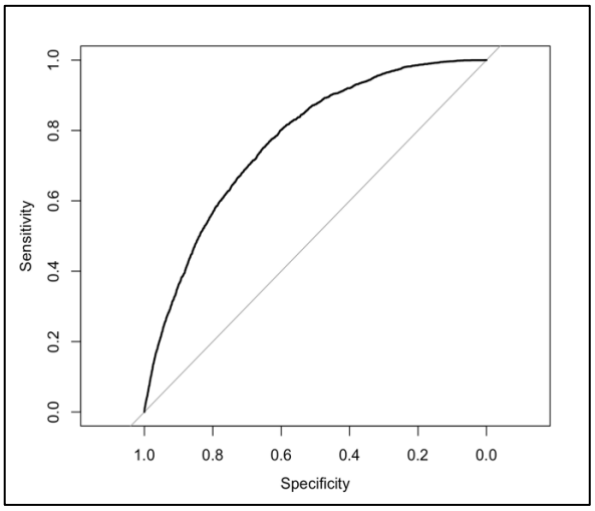


*Figure 3 ROC Curve*

The ROC curve on the other hand did not cover a major area under the Sensitivity vs Specificity plot. The graph only covers 77.1% of the data frame.

## Data Analysis using Tableau

Withing tableau, analysis not only include exploratory, but involved multiple solid findings as well. The "Roll-Up" approach for analysis was taken, where in-depth analysis of multiple variables was done, which then went higher up in the abstraction level, to show common patterns and draw out conclusions, using summaries of these findings.

**Balance and Education (Refer story page 1):** This revealed that people with a higher balance in their account are usually connected to higher education levels (Tertiary and Secondary). Additionally, this relationship was plotted against specific months to identify in which months, the acceptance rate was highest.

To find the acceptance rate, a custom field was created with the following formula:

$$SUM(IF\ [Y]\ =\ 'Yes'\ THEN\ 1\ ELSE\ 0\ END)\ /\ SUM(IF\ [Y]\ =\ 'No'\ THEN\ 1\ ELSE\ 0\ END)$$

This formula finds the relationship between Yes vs No answers. A general trend revealed that campaigns ran in the month of March, December, September, and October are associated with a higher acceptance rate of the campaign. The specifics can be revealed as filters in the tableau story.

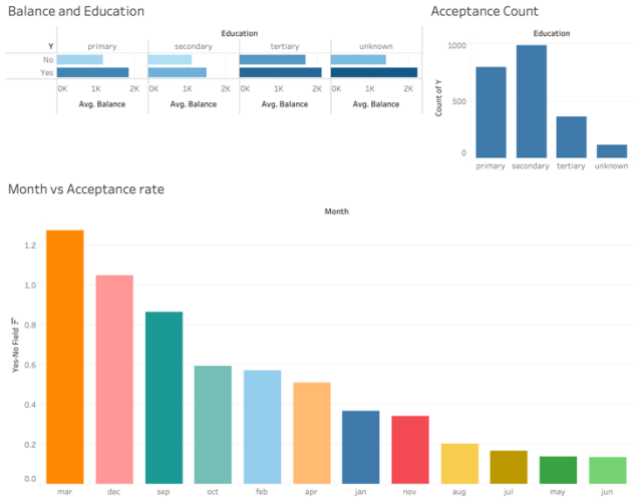


*Figure 4 Balance and Education level story screenshot - Story page 1 (Only for reference – not for inference)*

**Jobs combined with Months and Education Levels (Refer story page 2)**

The next part of the analysis involved using a heated map type of graph, which helped map out the jobs of people who would most likely accept the campaign.

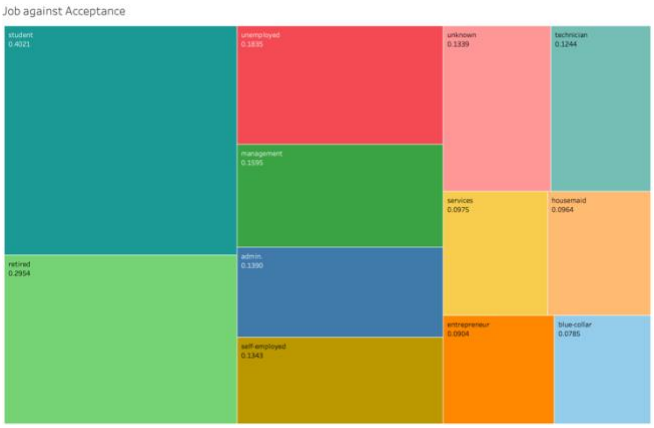


*Figure 5 Jobs with highest acceptance rate*

This data was then combined with the previous findings to reveal specific insights for each demographic of these top Jobs who would accept the campaign. The specific roles include: Student, Retired, Unemployed, management, admin and self-employed, in that particular order.

Combining this with the previous data helps create filters for specific months for each of these employment status to accept the campaign or not.

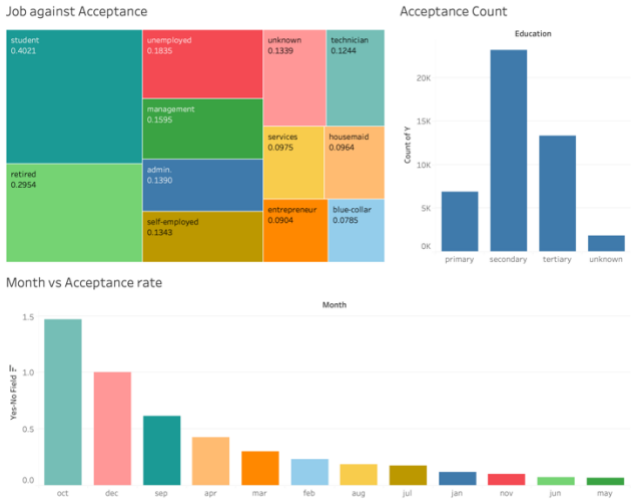


*Figure 6 Job combined story screenshot - Story page 2 (Only for reference – not for inference)*

**Housing for top Job accepters (Refer story page 3)**

The data is continued to approached in a Roll-up approach as now these top job acceptors are matched, if they owned a house or not. The tableau analysis revealed that generally these individuals do not own homes, as the count for not home owners was more.

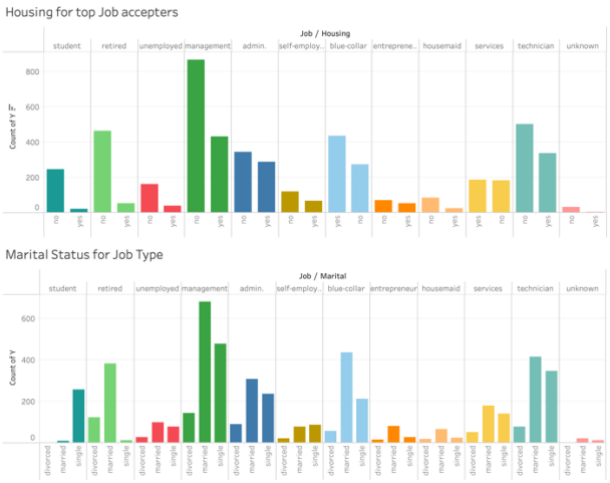


*Figure 7 Housing for top Job accepters - Story page 3 (Only for reference – not for inference)*

## Overall Conclusion and improvements

This sums the analysis of the report, where the R Studio logistic regression revealed the following model to be the best fit:

$$p(y = 1 \mid balance, duration, day, age) = \frac{1}{1 + e^{-(-3.788 + 0.0003534 \times 0.005659 \times duration - 0.01479 \times day - 0.003257 \times age)}}$$

This analysis further looked at other variables such as top Job types which would accept the campaign and their corresponding month along with their housing presence. This was modelled using tableau.

Some areas of further research could include testing with multiple binary against binary variables and incorporating them in R. Additionally performing further regression based analysis on some numeric values and using classification trees as well for further research.

# AIRBNB Property Sales Dataset

**Research aim:** The aim of this study is to precisely analyse the correlation patterns among variables including duration of stay, occupancy capacity, specific amenities, and their impact on the revenue generated by three distinct property types: "Professional," "2-5 Units," and "Single Owners."

**Objective and application:** This research objective is to understand the factors contributing to the high value and revenue generation potential of properties. By doing so, stakeholders may use these insights to make informed business decisions regarding optimal room allocation and strategic investments in amenities, thereby maximizing revenue generating opportunities for each property type.

## Data Set Explanation

The dataset was taken from Kaggle.
Link:
https://www.kaggle.com/datasets/computingvictor/zillow-market-analysis-and-real-estate-sales-data

This dataset comprises a comprehensive collection of information pertaining to the Airbnb rental market and property sales in two distinct areas within California: Big Bear and Joshua Tree.

This includes the following files:
• Amenities
• Geo Location
• Market Analysis (combined and 2019)
• Sale Properties with Zip Code (4 files)
• Sale Properties with Pool and Zip Code

Out of all these files, the research involves using the Amenities, Geo Location and Market Analysis 2019 files. These files are selected based on the correlational need between different features with what amount of revenue each particular type of property is generating.

## Process and Approach

The analysis approach for this was to first observe general features and feature-type amongst different csv files to choose the right files for this analysis.

Post general observational analysis, the data was then cleaned for multiple different reasons, using python code. Once the cleaned data was obtained, this was then connected to form a relational structure using the “unified-id” value, in tableau to observe general trends.

To find strong correlations and draw out conclusion, this analysis was done by generating a Linear Regression model using R, for each property owner type, followed by analysis of the findings.

## Data Cleaning

The following problems were identified in the dataset.

1. **Amenities.csv**
   - Duplicate Data
2. **Geo Location.csv**
   - Duplicate Data
   - Missing values in street name
   - Values in Latitude and Longitude separated by comma (‘,’), instead of a full-stop (‘.’).
3. **Market Analysis 2019.csv**
   - Missing Values
   - Values in Revenue are also separated by comma (‘,’), instead of a full-stop (‘.’).

This data was fixed by writing multiple python codes using the pandas library to clean this data. The final step was to combine these files into one file, named as “market2019FINAL.csv”

## Exploratory Analysis using Tableau

Some initial trends and analysis was plotted using tableau. These tableau dashboards look like the following:

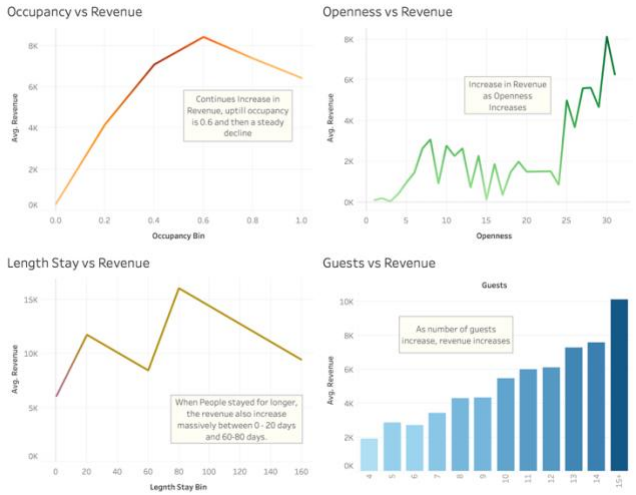


*Figure 8 Professional Properties Exploration (Only for reference – not for inference)*

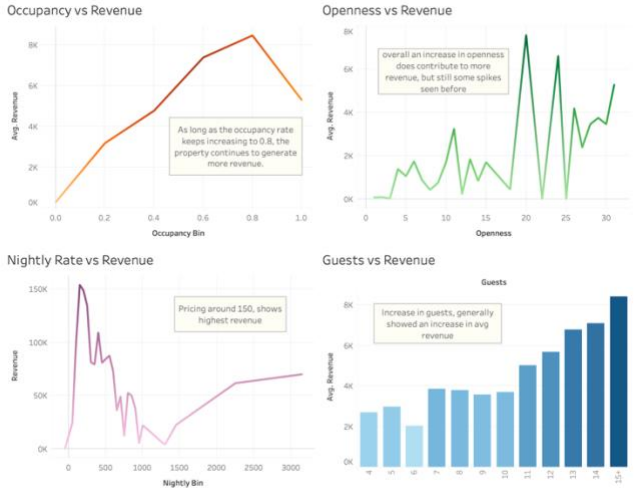


*Figure 9 '2-5 Units' Properties Exploration (Only for reference – not for inference)*

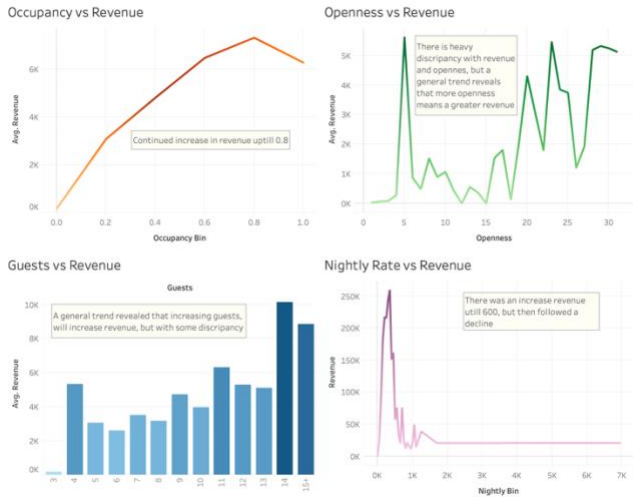


*Figure 10 'Single Owner' Properties Exploration(Only for reference – not for inference)*

The general understanding obtained from these graphs was the presence of these specific numerical values that had an impact on revenue of the property. This is a summary of some of the values such as: Openness, Occupancy, Length-stay, Guests, and Nightly-rate on revenue (Please refer to Story boards 1-3 for Inference).

## Refinement and preparing data in R

The process followed in refinement, Modelling and Analysis was repeated for all 3 property types.

Once the data was imported in R, the dataset required further refinement, these steps include first converting the data-types into the right category. This was done for the guests (numeric), month (date), Hot.tub (logical) and Pool (logical).

This was followed by one-hot encoding approach where the host_type was one hot encoded, so that each categorical variables transforms to a binary format where each category becomes a separate binary feature.

## Modelling and Analysis in R

The analysis begins by first observing the general structure and values of the dataset, to check for value types.

```
'data.frame':   2151 obs. of  20 variables:
 $ unified_id   : chr  "AIR10000347" "AIR10052559" "AIR10178668" "AIR10204420" ...
 $ month        : Date, format: "2019-01-01" "2019-01-01" "2019-01-01" ...
 $ zipcode      : int  92315 92315 92315 92252 92315 92315 92314 92315 92314 92314 ...
 $ city         : chr  "Big Bear Lake" "Big Bear Lake" "Big Bear Lake" "Joshua Tree" ...
 $ bedrooms     : int  3 3 3 3 3 5 3 3 3 5 ...
 $ bathrooms    : num  2 2.5 2.5 2 3 2 1 2 2 4 ...
 $ guests       : num  10 8 7 12 9 12 5 8 8 NA ...
 $ revenue      : num  13949 9947 6027 6549 8610 ...
 $ openness     : int  31 31 31 31 31 31 31 31 31 31 ...
 $ occupancy    : num  1 0.581 0.484 0.452 0.645 ...
 $ nightly.rate : num  450 553 402 468 430 ...
 $ lead.time    : num  8 10.6 7.1 38 43.1 ...
 $ length.stay  : num  65 2.57 1.8 2.43 2.33 ...
 $ Street.Name  : chr  "Cienega Road" "Heavenly Valley Road" "Round Drive" "NULL_STREET" ...
 $ Latitude     : num  34.2 34.2 34.3 34.2 34.2 ...
 $ Longitude    : num  -117 -117 -117 -116 -117 ...
 $ Hot.Tub      : logi  FALSE TRUE FALSE TRUE TRUE FALSE ...
 $ Pool         : logi  FALSE FALSE FALSE TRUE FALSE FALSE ...
 $ Professionals: int  0 1 1 0 1 0 0 1 0 0 ...
 $ Single Owners: int  0 0 0 0 0 1 0 0 0 0 ...
```

*Figure 11 Data Frame of the Market2019 File*

The data was then plotted as a histogram, where it was noticed that the data was skewed, mainly to the left, therefore to **normalize** the values, log of the values were taken for interpretation. The updated data for One of the properties look as follows:

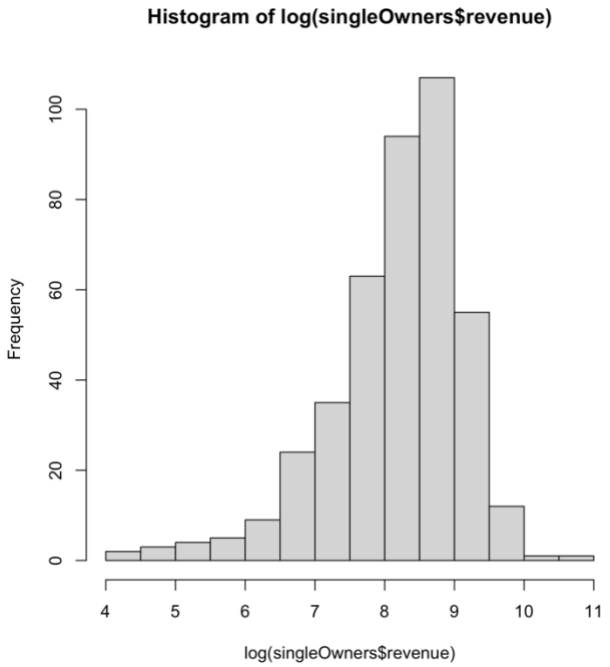


*Figure 12 Histogram for Single Owners, post normalization. This process was repeated for other graphs as well regarding other property owner types.*

The next step involved performing a logistic regression function. Over here the "Roll up" approach was taken, where the model was experimented with one variable and kept on increasing the number of variables for the best fit.

The best model was chosen on the basis of Adjusted R^2 value, as it shows the correct measure of the quality of the model, despite adding more variables. Additionally F Statistic was also observed, while plotting.

## Interpretation for "Single Owner" Property Type

```
Residuals:
     Min       1Q   Median       3Q      Max
-2.45753 -0.21334  0.06614  0.30361  1.08824

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   3.403e+00  1.469e-01  23.165   <2e-16 ***
guests        1.045e-01  1.103e-02   9.475   <2e-16 ***
openness      8.512e-02  3.555e-03  23.943   <2e-16 ***
occupancy     2.031e+00  9.714e-02  20.911   <2e-16 ***
nightly.rate  9.335e-04  6.462e-05  14.445   <2e-16 ***
lead.time    -1.237e-03  7.073e-04  -1.748   0.0812 .
length.stay  -3.630e-03  2.230e-03  -1.628   0.1044
Hot.TubTRUE   1.374e-01  4.773e-02   2.880   0.0042 **
PoolTRUE     -1.529e-01  1.055e-01  -1.449   0.1482
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4625 on 387 degrees of freedom
  (125 observations deleted due to missingness)
Multiple R-squared:  0.7761,    Adjusted R-squared:  0.7714
F-statistic: 167.6 on 8 and 387 DF,  p-value: < 2.2e-16
```

*Figure 13 LR Model for Single Owner Types*

**Technical Analysis:**
Strong correlations between the variables in this regression model—particularly the number of bedrooms, bathrooms, openness, occupancy, and guests—indicate that people prefer large residences, was observed. length of stay has little bearing and suggests shorter stays. When compared to other features, the relationships between certain amenities—like pools or hot tubs—are weaker.

Adjusted R-squared: At 77.14%, this value indicates a good model fit, though around 1/3rd of the data remains unexplained.
F-statistic: With a very low p-value (< 2.2e-16), the model is statistically significant.
Residuals: Small, symmetric residuals around zero suggest no major violations of linear regression assumptions.

**Plotted Model Analysis using GOF**
The visual representation of the model suggests that the Residual vs fitted graph has majority of the data points residing on the trend line, indicating a strong accuracy of the linear regression model.

Similarly the QQ Residuals graph shows a steady diagonal line with data points residing on the trend line, indicating that the data is normally distributed. This has some limitations for values above and below the 2nd quartiles generated. The standardized residuals also show that there aren't many data points which have a large impact on the model.

The Scale location on the other hand does not show equal variance of residuals, as there is a parabolic curve, but most data sits under the 0 to 10,000 data point fitted value.

## Business Application for" Single Owner" Property Type

For Single Owner family homes, the data revealed that most of the customers don't really care about the length of stay and the type of amenities that they might get, but instead they do mainly care about all other features such as number of bedrooms, bathrooms, guest occupancy, etc.

This indicates a very fair model, that for these Single Owner properties, it might be more worth it to spend on the overall development and maintenance of the properties rather than relying on expensive specific amenities such as pools and hot tubs.

## Interpretation for "Professional" Property Type

```
Residuals:
    Min      1Q  Median      3Q     Max
-3.5766 -0.1084  0.0904  0.2251  0.8034

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   3.732e+00  1.033e-01  36.118  < 2e-16 ***
bedrooms     -5.870e-02  3.161e-02  -1.857 0.063679 .
bathrooms     3.141e-02  2.518e-02   1.247 0.212617
guests        3.176e-02  8.417e-03   3.773 0.000172 ***
openness      8.058e-02  2.108e-03  38.228  < 2e-16 ***
occupancy     2.522e+00  6.597e-02  38.234  < 2e-16 ***
nightly.rate  2.169e-03  6.541e-05  33.158  < 2e-16 ***
lead.time    -2.098e-03  6.168e-04  -3.401 0.000701 ***
length.stay  -5.281e-04  2.127e-03  -0.248 0.803980
Hot.TubTRUE   6.791e-02  2.866e-02   2.370 0.018019 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4115 on 904 degrees of freedom
  (311 observations deleted due to missingness)
Multiple R-squared:  0.8344,    Adjusted R-squared:  0.8328
F-statistic: 506.2 on 9 and 904 DF,  p-value: < 2.2e-16
```

*Figure 14 LR Model for Professional Property Types*

**Technical Analysis:**
The linear regression analysis highlights strong correlations between variables such as Openness, occupancy, nightly rate, length of stay and lead. Time, with a notable 99.99% significance.

Conversely, Bedrooms and bathrooms show weak correlations, challenging the assumption that more bedrooms equate to higher profits. Properties with more guests and longer stays tend to generate higher revenue, while shorter lead times correlate with increased earnings. Among amenities, only Hot Tub presence shows significant correlation, at 95%. The model's adjusted R-squared value of 0.8328 indicates strong explanatory power, with nearly 83% of revenue variance explained by the independent variables.

This is supported by negligible p-values and a robust F Statistic, affirming the model's strength.

**Plotted Model Analysis using GOF**
Relations are balanced in the ideal model, with most falling between 0 and 10,000 and in the area of the median quartiles. Still, there's space for development. Variable variance is visible on the scale-location graph, especially in the range of 0 to 10,000. With a few outliers, the Q-Q Residuals graph primarily shows a normal distribution. Most of the points on the residual vs. leverage graph are grouped around 0, indicating that the model is generally accurate with very few significant outliers.

## Business Application for "Professional" Property Type

**Location Matters:** Properties in areas with higher openness, occupancy rates, and nightly rates tend to generate more revenue.

**Lead Time Consideration:** Shorter lead times might be preferable for maximizing revenue.

**Amenities Boost Revenue:** Properties with amenities like Hot Tub generally command higher revenue.

**Room Configuration:** While the number of bedrooms and bathrooms seems to positively influence revenue, their significance levels are relatively low in this model. Meaning more rooms does not necessarily mean more revenue.

## Interpretation for "2-5 Units" Property Type

```
Residuals:
     Min       1Q   Median       3Q      Max
-2.89979 -0.16825  0.07881  0.26516  0.98933

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)   3.7997392  0.2152054  17.656  < 2e-16 ***
bedrooms     -0.1136456  0.0625697  -1.816 0.070374 .
bathrooms     0.0369880  0.0516888   0.716 0.474830
guests        0.0507016  0.0143155   3.542 0.000464 ***
openness      0.0798561  0.0037304  21.407  < 2e-16 ***
occupancy     2.0072141  0.1025331  19.576  < 2e-16 ***
nightly.rate  0.0019527  0.0001151  16.967  < 2e-16 ***
lead.time    -0.0013137  0.0007074  -1.857 0.064351 .
length.stay   0.0011432  0.0025975   0.440 0.660205
Hot.TubTRUE   0.2968417  0.0593452   5.002 9.94e-07 ***
PoolTRUE     -0.1257491  0.0898077  -1.400 0.162540
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4597 on 285 degrees of freedom
  (109 observations deleted due to missingness)
Multiple R-squared:  0.8134,    Adjusted R-squared:  0.8069
F-statistic: 124.3 on 10 and 285 DF,  p-value: < 2.2e-16
```

*Figure 15 LR Model for 2-5 Units Property Type*

**Technical Analysis:**

The analysis suggests that factors such as the number of guests accommodated, property openness, occupancy rates, nightly rate, and the presence of a hot tub significantly influence the prices of the "2-5 Units" property type.

However, the number of bedrooms, bathrooms, lead time, length of stay, and the presence of a pool do not have significant effects on property prices in this model.

The model performs effectively, accounting for about 81.34% of the variation in values. Its adjusted R-squared value of 80.69% suggests a strong match. The model's overall fit for the data is confirmed by the highly significant F-statistic ($p < 2.2e-16$).

All of these measures point to the model being able to accurately account for a significant portion of the variation in property prices for AIRBNB listings that fall under the "2-5 Units" property classification, providing trustworthy information about the dynamics of pricing in this market.

**Plotted Model Analysis using GOF:**

The accuracy of the model could be well visualized by looking at the generated plots, as it was clearly observed that the Residuals vs fitted graph shows that majority of the values follow the trend line, indicating that the model represents a larger number of values from the dataset. Similarly on QQ Residuals chart, even though it has a slightly upwards sloping line, has majority of the values between -1 to 1 to follow the trend line, making the model accurate for those predictions. The residuals vs leverage graph indicates that the majority of the values are near 0, with only a few exceptions.

## Business Application for "2-5 Units" Property Type

**Optimize Pricing:** Adjust nightly rates based on demand, seasonality, and local events to maximize revenue without sacrificing occupancy. Since pricing seems to be inelastic, an increase in price shows a stronger increase in revenue. Therefore the customer base is not that price conscious.

**Maximize Occupancy:** Implement marketing strategies to attract more guests and minimize unoccupied nights. As properties with more occupancy and number of guests generate more revenue.

**Property Improvement:** Enhance property amenities such as adding hot tubs or upgrading bedrooms and bathrooms to justify higher rates and attract more guests.

## Overall Conclusion and improvements

Each of the model generated within this part of the research indicated a strong fit, for predicting the prices of the properties, with more than 70% Adjusted R^2 values for all property types.

The specific business applications to maximize revenue for each property type is given in sub-sections. Areas of improvement could involve the following:

1. Further normalizing the data
2. Provide more analysis and tests of the models, this will include performing cross-validation techniques.
3. Detect more outliers and clean data to only work with the best sub-set dataset.