## Linear Regression of California Airbnb Data

### D214: Capstone

### Shawn Wheeler

### 000364855

Table of Contents

[Part I: Research Question 3](#_Toc171941293)

[A1: Research Question 3](#_Toc171941294)

[A2: Justification 3](#_Toc171941295)

[A3: Context 3](#_Toc171941296)

[A4: Hypothesis 3](#_Toc171941297)

[Part II: Data Collection 4](#_Toc171941298)

[B1: Data Collected 4](#_Toc171941299)

[B2: Methodology Advantages and Disadvantages 4](#_Toc171941300)

[B3: Challenges 4](#_Toc171941301)

[Part III: Data Extraction and Preparation 5](#_Toc171941302)

[C1: Data Extraction and Preparation Process 5](#_Toc171941303)

[C2: Tools and Techniques 25](#_Toc171941304)

[C3: Justification 25](#_Toc171941305)

[Part IV: Analysis 25](#_Toc171941306)

[D1: Description of Technique 25](#_Toc171941307)

[D2: Calculations and Output 26](#_Toc171941308)

[D3: Justification of Analysis Technique 38](#_Toc171941309)

[Part V: Data Summary and Implications 38](#_Toc171941310)

[E1: Results 38](#_Toc171941311)

[E2: Limitations 39](#_Toc171941312)

[E3: Recommended Course of Action 39](#_Toc171941313)

[E4: Approach for Future Study 39](#_Toc171941314)

[Part VI: Sources 40](#_Toc171941315)

[F: Sources 40](#_Toc171941316)

## Part I: Research Question

### A1: Research Question

Within the California Airbnb data set, to what extent do the independent variables ‘accommodates’, ‘beds’, ‘bathrooms\_text’, ‘bedrooms’, ‘amenities’, ‘room\_type’, and ‘city’ affect the dependent variable ‘price’?

### A2: Justification

It can be difficult to know how to price a rental listing. Utilizing multiple types of regression and comparison of models, this analysis will attempt to create an explanation of variables which affect price based on Airbnb data in California. Creating regression models on Airbnb data “provides interesting insights that can benefit a host looking to maximize their profits.” (Tersakyan, 2019). Therefore, one looking to enter the market or already in the market could theoretically use the model to accurately list their property without having to suffer through price experimentation which can have catastrophic effects.

### A3: Context

For the purposes of this study, only variables that would be able to be determined prior to listing a property were chosen to be included. Those variables were ‘accomodates’, ‘beds’, ‘bathrooms\_text’, ‘amenities’, ‘room\_type’, and ‘city’ in addition to the dependent variable ‘price’. It will be determined if the independent variables have enough correlation to be able to predict the price of a listing.

### A4: Hypothesis

Below are the null and alternate hypotheses for this study. The goal of the analysis is to determine significance between any independent variables and the dependent variable. In other words, it will be testing which factors are affecting the variable ‘price’ and to what degree. The alpha level is set at 0.05 meaning that if any variable shows a p-value of greater than 0.05, then it is not statistically significant with relation to the dependent variable. To determine whether the null hypothesis is accepted or rejected, the analysis will determine whether significance exists via evaluation.

**Null hypothesis**- H0: The independent variables ‘accomodates’, ‘beds’, ‘bathrooms\_text’, ‘bedrooms’, ‘amenities’, ‘room\_type’, and ‘city’ do not affect the dependent variable ‘price’ in a statistically significant way based on an alpha value of 0.05.

**Alternate Hypothesis**- HA: The independent variables ‘accomodates’, ‘beds’, ‘bathrooms\_text’, ‘bedrooms’, ‘amenities’, ‘room\_type’, and ‘city’ affect the dependent variable ‘price’ in a statistically significant way based on an alpha value of 0.05.

## Part II: Data Collection

### B1: Data Collected

The data set for this analysis was a compilation of data provided by Airbnb on property listings over a few months in the state of California. (Get the Data, n.d.) This compilation, however, was sourced from Kaggle. (Codify, 2024). There were 247,829 rows and 62 columns of original data. For the 62 columns, they were what one would expect from an Airbnb listing, such as, ratings, price, bathrooms, amenities, and many others. Below is a table of the variables that were selected for analysis. These were selected due to their relevance in attempting to fulfill the criteria of this analysis which requires information prior to listing a property.

*Table of Variables:*

|  |  |  |
| --- | --- | --- |
| **Variable** | **Description** | **Data Type** |
| room\_type | One of private/shared/hotel room or entire home/apt | Categorical |
| accommodates | Maximum number of occupants | Continuous |
| beds | Number of beds | Continuous |
| price | Price of a single night | Continuous |
| bathrooms\_text | Text of number of bathrooms | Categorical |
| bedrooms | Number of bedrooms | Continuous |
| amenities | Json text of amenities included, such as, wifi | Categorical |
| city | Location of property | Categorical |

### B2: Methodology Advantages and Disadvantages

One advantage of the methodology used for data collection is that the data set is large enough to accommodate a great deal of manipulation while still providing ample data. Furthermore, acquiring the data set through Kaggle as opposed to compiling it via Airbnb saved a great deal of time. Having a lack of input as to the variables themselves which appear in the data set is a disadvantage of having acquired the already compiled data set, however.

### B3: Challenges

Unfortunately, due to the way the data was scraped, it is not in an ideal format in all cases leading to complications. For instance, the ‘bathrooms\_text’ variable is an unnecessary list of strings, e.g. ‘1.5 bathrooms,’ instead of simply a list of numbers alone. There were also issues with the amenities column due to it being formatted as json and appearing as a list of strings which required unraveling into separate columns. Furthermore, he data for non-continuous columns had to be altered as linear regression “accepts only numeric values as inputs.” (kfollis, 2023).

## Part III: Data Extraction and Preparation

### C1: Data Extraction and Preparation Process

Using Python within JupyterLab, the California Airbnb data set was loaded for manipulation after importing libraries and packages as needed for the analyses. Next, the dataframe was examined for shape, rows, data types, and columns. At this point, all columns except for the independent variables and the dependent variable to be used were dropped. Afterwards, data was examined for missing values. Five columns were shown to have missing data; the most being ‘bedrooms’ at 130,031. The missing values for ‘bedrooms’ and ‘beds’ were then replaced with median values. Here, the only missing data remaining was ‘price’ with 2708, ‘amenities’ with 2, and ‘bathroom\_texts’ with 153. All of those rows were dropped.

*Screenshot of initial dataframe info():*

*A black and white list of text

Description automatically generated with medium confidence*

*Screenshot of missing data:*

*A screenshot of a computer

Description automatically generated*

Duplicates were then tested for revealing 2854 instances. Those rows were also then dropped. Then, the data was examined individually for outliers and any other aberrations. During this, ‘price’ was set to only include values between 19 and 400 (dollars), dropping the remaining rows. The ‘price’ column was then replaced with a logged version of each price in order to enhance the usefulness of the data. There were other outliers detected in the ‘accommodates’, ‘beds’, ‘bathrooms’, and ‘bedrooms’ variables which were replaced with median values.

*Screenshots of duplicates before and after:*

*A close up of numbers

Description automatically generated*

**

Variable ‘bathrooms\_text’ was converted from strings of text to a numerical value, and the column was renamed ‘bathrooms.’ The text for the ‘amenities’ column was pulled from to create individual columns indicating whether a property contains a particular amenity. An example of this would be ‘wifi’ or ‘self check-in.’ These columns were then recast as ints. Following, a column that tallied the number of amenities for a particular property was created and added to the data set as ‘number\_amenities.’ After extracting the desired data, the ‘amenities’ column was dropped.

*Screenshots of bathrooms\_text before and after:*

A screenshot of a computer

Description automatically generated

A table of numbers with a white background

Description automatically generated with medium confidence

*Screenshot of amenities:*

*A close up of a text

Description automatically generated*

*Dataframe at this point:*

*A screenshot of a computer

Description automatically generated*

Next came univariate and bivariate analysis of the data set. Each column was visualized alone and then as compared to ‘price.’ Below are the univariate and bivariate graphs for the data set at this point.

*Univariate analyses followed by bivariate analyses:*

*room\_type:*

*A graph with numbers and a bar

Description automatically generated*

*A diagram of a relationship between price and room type

Description automatically generated*

*accommodates:*

*A graph of blue rectangular bars

Description automatically generated*

*A graph of blue dots

Description automatically generated*

*beds:*

*A graph of a number of beds

Description automatically generated*

*A graph of blue dots

Description automatically generated*

*price:*

*A diagram of a distribution of price

Description automatically generated*

*bathrooms:*

*A graph of a number of bathrooms

Description automatically generated*

*A graph of blue dots

Description automatically generated*

*bedrooms:*

*A graph of a distribution of bedroom

Description automatically generated*

*A graph of blue dots

Description automatically generated*

*City:*

*A pie chart with different colored sections

Description automatically generated*

*A graph of a number of cities

Description automatically generated with medium confidence*

*wifi:*

*A green circle with a blue triangle and a purple triangle

Description automatically generated*

*A diagram of a relationship between price and wifi

Description automatically generated*

*tv:*

*A pie chart with text on it

Description automatically generated*

*A diagram of a relationship between tv and tv

Description automatically generated*

*water:*

*A purple and green pie chart

Description automatically generated*

*A diagram of a couple of rectangular objects

Description automatically generated*

*exercise:*

*A pie chart with numbers and a triangle

Description automatically generated*

*A diagram of a diagram

Description automatically generated*

*oven:*

*A pie chart with text on it

Description automatically generated*

*A chart with green rectangles and black lines

Description automatically generated*

*dryer:*

*A pie chart with text overlay

Description automatically generated*

*A diagram of a relationship between price and dryer

Description automatically generated*

*grill:*

*A pie chart with a pink and brown circle

Description automatically generated*

*A chart of a diagram

Description automatically generated with medium confidence*

*self\_check\_in:*

*A pie chart with a purple and green circle

Description automatically generated*

*A diagram of a relationship between price and self check

Description automatically generated*

*workspace:*

*A green and grey circle with text

Description automatically generated*

*A diagram of a relationship between price and workspace

Description automatically generated*

*view:*

*A pie chart with text on it

Description automatically generated*

*A diagram of a relationship between price and view

Description automatically generated*

*parking:*

*A pie chart with a green triangle and a purple triangle

Description automatically generated*

*A diagram of a parking lot

Description automatically generated with medium confidence*

*number\_amenities:*

*A graph of a number of amenity

Description automatically generated*

*A graph of numbers and lines

Description automatically generated with medium confidence*

After visualizing the data, dummy variables were created for ‘room\_type’ and ‘city’ and the first column dropped for each in order to combat multicollinearity. The dummy variables also were converted to int. (kfollis, 2023). All columns except for ‘price’ were then scaled using StandardScaler() to better suit the analysis. Penultimately, the data was split with an 80% training set and 20% testing set to prepare for analysis. The training set contained 166,374 rows while the test had 41,594.

*Screenshot of scaled data:*

*A screenshot of a table

Description automatically generated*

*Screenshot of final data frame:*

*A screenshot of a computer

Description automatically generated*

*Screenshot of shapes for split data (X\_train, X\_test, y\_train, and y\_test respectively):*

*A number and text on a white background

Description automatically generated*

### C2: Tools and Techniques

Python was selected for this task as opposed to R or SAS. JupyterLab was also used due to its ease of use and reproducibility.

Exploratory data analysis was implemented using univariate and bivariate analysis. This allowed for a visual representation of what the data could mean or how it may or may not correlate with the dependent variable. This assists with determining outliers, shape of the distribution of data, and many other aspects necessary for data preparation for linear regression. (Qualtrics, 2017).

For the initial data preparation, a number of libraries and packages were used. Seaborn and matplotlib were used for visualizations. Pandas was used to load the data set for manipulation and to then assist with determining aberrations and other potential issues that need to be addressed prior to the analysis portion of this project. Numpy was also used for data manipulation, especially with regards to correcting outliers Data was finally split using train\_test\_split from sklearn.model\_selection.

### C3: Justification

Python was chosen over SAS and R due to its ability to handle large datasets and the fact that SAS requires a license to use. (Anon, n.d.). As mentioned above, the exploratory data analysis was utilized due to its ability to assist with aberrant data. This would prove to be quite an advantage in performing this project. (Numeracy, n.d.). A disadvantage to the methodology used for extraction and preparation is that the ‘amenities’ column was broken down into specific, but not entirely inclusive, amenities themselves leading to some potential for variables that may have been useful for this analysis which are hidden within the amenities text list.

## Part IV: Analysis

### D1: Description of Techniques

Initially, data will be evaluated for correlation and normalcy using Variance Inflation Factors (VIF) and Shapiro-Wilk. Three techniques will then be used to construct regression models in an attempt to predict ‘price.’ First, linear regression using a stepwise approach with OLS (Ordinary Least Squares) to narrow down a model. Secondly, the LinearRegression model from sklearn.linear\_model will be created. Thirdly, RandomForestRegressor from sklearn.ensemble will be instantiated, fit, and evaluated. Evaluation of the models will be performed using various factors, such as R-squared, RMSE, and MSE, and compared to determine their overall and comparative usefulness.

### D2: Calculations and Output

The next step in the analysis process was to investigate correlation between variables which might damage the quality of the produced models and to perform a Shapiro-Wilk test. Therefore, a heatmap was created to visualize the correlations with red being a higher correlation and blue less of one. Next, VIFs were tested for and any variables with values over five were removed. The Shapiro-Wilk test for normality was performed, as well. The results were a statistic score of 0.987 and a p-value of 0.0. A statistic score closer to one describes data that follows a normal distribution. (Humaizi, 2024).

*Screenshot of heatmap:*

*A screen shot of a computer screen

Description automatically generated*

*Screenshot of VIFs:*

*A screenshot of a phone

Description automatically generated*

*Screenshot of Shapiro-Wilk:*

**

Below are the three aforementioned models in sequence.

**Stepwise OLS Model**

The initial model for the multiple regression is shown below.

*Initial stepwise OLS model:*

A screenshot of a computer

Description automatically generated

*Screenshot of RSE:*

**

As can be seen, there are many initial p-values above the threshold set initially of 0.05. There are also many other useful data points shown in these results. R-squared measures the explanation of changes between the independent and dependent variables. This indicates that 49.3% of the change is explainable through this model. The F-statistic is used to determine if the variables themselves are statistically significant and is shown to be 8782. Whereas the Prob (F-statistic) refers to the accuracy of the null hypothesis. Other values of note shown here include AIC, 2.302e+05, which will be used to compare linear regression models and omnibus score, 3110.290, for which a higher value is worse for this model. Relative Squared Error (RSE) was also calculated which describes differences between observed and predicted values. A low RSE indicates high precision and can be compared across models, as well. (Qualtrics, 2017). The initial model had an RSE of .4208.

Removing the variables with the highest p-value one at a time until only values less than 0.05 remained took seven additional iterations leaving the final model as shown below.

*Screenshot of final stepwise model:*

A screenshot of a computer

Description automatically generated

*Screenshot of RSE:*



Looking again at the final model shown above, the list of coefficients is shown in the “coef” column for each variable. These coefficients represent the magnitude and direction of the correlation between the variable and the dependent variable. The values are the result of the data having been scaled and also the ‘price’ variable logged. In order to understand these, accommodates has a coefficient of 0.1043, meaning a 1 point increase in accommodates leads to a .1043 increase in the log price. According to these, city\_pacific-grove has the highest positive correlation, meaning if a property is located in Pacific Grove, the log price increases by .3351. Negatively, the largest value was shown to be “room\_type\_Shared room” where if a room is shared, the value of the log price decreases by 1.0979. Const refers to a constant which is the intercept of the line (or the base value). Meaning that each calculation begins at 4.6178 based on this model.

Residuals were then plotted to check for normality of data. For fit, the data should be randomly distributed with zero being the mean. A Q-Q plot was also created.

*Screenshot of Residual Plot for Reduced Model:*

*A green triangle with blue line

Description automatically generated*

*Screenshot of Q-Q Plot for Reduced Model:*

*A graph of a red line

Description automatically generated*

Scores were calculated for this model and are shown below. These scores were mean score, RMSE (0.42), MSE (0.18), and R-squared (0.49). As mentioned, these will be used to compare models.

*Screenshots of evaluation scores for reduced model:*

**

*A number and numbers on a white background

Description automatically generated*

A graph was created to visualize test data with regards to the predictions created by the OLS model. It is a snapshot of the first 100 values for each the test and predicted values.

*Screenshot of Test vs Predicted Snapshot:*

*A graph with blue and orange lines

Description automatically generated*

**LinearRegression Model**

Once the training data was fit to the LinearRegression model, its evaluation scores were extracted for the training and testing data.

*Screenshot of evaluation scores for LinearRegression:*

*A screenshot of a computer

Description automatically generated*

For this model, the intercept was shown to be 4.89. The highest positively correlated coefficient was for “accommodates” at 0.21. Negatively correlated, again, “room\_type\_Private room” has the largest magnitude at -0.21.

*Screenshots of intercept and coefficients:*

**

*A screenshot of a cell phone

Description automatically generated*

Residuals for this model were visualized once again, as well as another Q-Q plot.

*Screenshot of Residual Plot for LinearRegression:*

*A blue dotted line graph

Description automatically generated*

*Screenshot of Q-Q Plot for LinearRegression:*

*A graph with a red line

Description automatically generated*

*Screenshot of mean score:*

**

Another graph comparing the test values and predicted values using the LinearRegression model was constructed.

*Screenshot of Test vs Predicted – LinearRegression:*

*A graph of a test

Description automatically generated*

**RandomForestRegressor Model**

Lastly, RandomForestRegressor() was instantiated and trained on the training set after having used GridSearchCV to find ideal parameters. The final parameters used were n\_estimators=500, max\_features=10, and max\_depth=10. Then, the model was evaluated for its performance on the training and testing data. This was evaluated on the test and training data. Although the test data values will be used for comparison to other models. Therefore, this model shows an RMSE of .41, an MSE of .17, and an R-squared of 0.52. The R-squared values for the training and testing data shows a lack of overfitting.

*Screenshot of evaluation scores:*

*A number on a white background

Description automatically generated*

*A number and numbers on a white background

Description automatically generated*

Using the random forest model, the feature importances were collected and visualized. Importances refer to the percentage an independent variable is shown to affect the dependent variable with regards to the other independent variables. As shown below, “accommodates” once again rose to the top with almost 44% of the total importance while “room\_type\_Hotel room” showed the lowest with 0.02% importance. A chart visualizing the comparison is shown below, also.

*Screenshot of features importances:*

*A screenshot of a computer code

Description automatically generated*

*Screenshot of Features Importance of Price graph:*

*A graph with red lines

Description automatically generated*

Residuals were then plotted for the random forest model along with its Q-Q plot. Mean score was also calculated. Finally, another chart showing the test vs predicted data using the random forest model is included for the first 100 values.

*Screenshot of Residual Plot for Random Forest:*

A graph of a plot

Description automatically generated

Screenshot of Q-Q Plot for Random Forest:

A graph of a graph with a red line

Description automatically generated

*Screenshot of mean score:*

**

*Screenshot of Test vs Predicted – Random Forest:*

*A graph showing a blue and orange line

Description automatically generated*

### D3: Justification of Analysis Technique

Stepwise OLS was selected initially because it allows for improving the model by reducing the amount of variables that cause the model to perform worse. However, it only performs as well as the variables given and the order they were added or removed causing potential discrepancies. (GeeksforGeeks, 2022).

LinearRegression from sklearn.linear\_model was chosen as a more automated linear regression than the previously manually performed stepwise model. It did provide slightly differing results than the manual method showing that there are advantages to using it; however, overall it seems to have performed about as well as the previous model. The advantage would be a great deal of time saving in performing the sklearn automated version while a disadvantage would be not being able to select the order of variables being removed prior to evaluation.

RandomForestRegressor then was selected as an alternative to the other versions of linear regression. It is from sklearn.ensemble. Random Forest is much more highly adaptable than the previous methods. For instance, it is non-parametric, meaning that it makes no assumptions regarding the distribution of data and handles outliers very well. However, with these advantages, it takes a great deal of computing power leading to slow processing times. (GeeksforGeeks, 2024).

## Part V: Data Summary and Implications

### E1: Results

The purpose of the study was to determine the statistical correlation between variables and to also determine which model is the most accurate and if any of them are worth using in the future. Looking at the scores below, lower values for RMSE and MSE are superior. A higher value for R-squared indicates a better explanation, as well. The mean scores are all very low, also. Therefore, based on this information, the random forest model is shown to be superior, although only slightly. With its R-squared value of .52 that again indicates that 52% is explained by the model. The random forest model itself is not an incredible predictor and therefore is a poor model overall even if it is the best of the three. (Beheshti, 2020).

*Scores for the Three Models:*

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Stepwise** | **LinearRegression** | **Random Forest** |
| RMSE | 0.42 | 0.42 | 0.41 |
| MSE | 0.18 | 0.18 | 0.17 |
| R-squared | 0.49 | 0.49 | 0.52 |
| Mean Score | 0.0 | 0.0018 | 0.0023 |

Regarding the initial hypothesis, the null hypothesis states that the selected variables are not statistically significant with regards to the dependent variable. Since the hypothesis listed all of the initial variables, and it has been shown that there are those that were not statistically significant (p-value > 0.05 for instance). Therefore, the null hypothesis must be accepted. However, there are variables that were shown to be statistically significant.

Of note, there were many similarities in which variables had the greatest positive and negative effects shown across models. “Accommodates” had an importance of 44% via random forest, and it had the highest positive coefficient for both the OLS and LinearRegression models. Shared rooms were shown to have a large negative coefficient and did not make the list of importances for the random forest model. Another notable difference is that large difference in location with regards to the prediction of price. Pacific Grove had the highest correlation for the initial stepwise model, but the other two models presented showed San Francisco as having the highest prices among the cities evaluated. Few amenities seemed to have much of an impact on price which was unexpected. The largest factors shown were how many people a property can accommodate, the number of beds, the location, and whether the space is private or shared.

### E2: Limitations

Linear regression relies on a set of assumptions which can heavily skew results if they aren’t accounted for. These assumptions include a linear relationship between the independent and dependent variables, a lack of multicollinearity, a lack of autocorrelation, outliers must be accounted for accurately, and the “error term is the same across all values of the independent variables.” (GeeksforGeeks, 2017). Further, a larger set of variables from which to pull from would have been greatly helpful, as well.

### E3: Recommended Course of Action

The recommended course of action based on this study is to compile more data points, additional variables, and to then repeat the experiment in the future with theoretically superior data. While the best model created in this project is not suitable for accurately predicting price, it does show possibility. Another consideration would be to use similar Airbnb data for another area and see how the models perform to gain additional insight as other studies have shown it to be a possibility as referenced below.

### E4: Approach for Future Study

Two approaches for future study based on this analysis are to study price changes over time with the Airbnb properties to determine correlation as a way of determining whether the price itself is more important than other potential factors when it comes to renting a property and to also survey clients in order to narrow down most important factors to then see if they would be able to be used to create a better, more usable model. The results would help with deciding whether to entirely abandon creating a model in these ways or if it is possible to pivot. The alternative hypothesis was proven for another set of Airbnb data showing the potential for a successful outcome (Wang, 2023).

## Part VI: Sources

### F: Sources

Anon. (n.d.). Is Predictive Modelling easier with R or with Python? (n.d.). ProjectPro. https://www.projectpro.io/article/is-predictive-modelling-easier-with-r-or-with-python/245

Beheshti, N. (2022, March 2). Random Forest Regression. Medium. https://towardsdatascience.com/random-forest-regression-5f605132d19d

Codify. (2024). AirBnB Listings in California. Kaggle.com. https://www.kaggle.com/datasets/setseries/airbnb-listings-in-california

GeeksforGeeks. (2017, March 19). *Linear Regression (Python Implementation) - GeeksforGeeks*. GeeksforGeeks. https://www.geeksforgeeks.org/linear-regression-python-implementation/

GeeksforGeeks. (2022). *Stepwise Regression in Python*. (2022, December 28). GeeksforGeeks. https://www.geeksforgeeks.org/stepwise-regression-in-python/

GeeksforGeeks. (2023). SAS vs R vs Python. (2023, June 12). GeeksforGeeks. https://www.geeksforgeeks.org/sas-vs-r-vs-python/

GeeksForGeeks. (2024, February 15). What are the Advantages and Disadvantages of Random Forest? GeeksforGeeks. https://www.geeksforgeeks.org/what-are-the-advantages-and-disadvantages-of-random-forest/

Get the Data. (n.d.). Insideairbnb.com. https://insideairbnb.com/get-the-data/

Humaizi, A. (2024, January 4). 10.0 Shapiro-Wilk test - Ahmad Humaizi - Medium. Medium; Medium. https://medium.com/@maizi5469/10-0-shapiro-wilk-test-5be38fd3c2a6

kfollis. (2023, October 31). Specify a Column to Use as Regressor in a Model. Learn.microsoft.com. https://learn.microsoft.com/en-us/analysis-services/data-mining/specify-a-column-to-use-as-regressor-in-a-model?view=asallproducts-allversions

Numeracy. (n.d.). Numeracy, Maths and Statistics - Academic Skills Kit. (n.d.). Www.ncl.ac.uk. https://www.ncl.ac.uk/webtemplate/ask-assets/external/maths-resources/statistics/regression-and-correlation/assumptions-of-regression-analysis.html

Qualtrics. (2017). Interpreting Residual Plots to Improve Your Regression - Qualtrics Support. (2017, July 19). Qualtrics Support. https://www.qualtrics.com/support/stats-iq/analyses/regression-guides/interpreting-residual-plots-improve-regression/

Tersakyan, A. (2019, October 18). Airbnb Price Prediction Using Linear Regression (Scikit-Learn and StatsModels). Medium. https://towardsdatascience.com/airbnb-price-prediction-using-linear-regression-scikit-learn-and-statsmodels-6e1fc2bd51a6

Wang, Haoqian. (2023). Predicting Airbnb Listing Price with Different models. Highlights in Science, Engineering and Technology. 47. 79-86. 10.54097/hset.v47i.8169.