**Linear Regression Model on Airbnb Pricing for Clark County Nevada**

Heidi L. Petersen

College of Information Technology, Western Governors University

C772: Data Analytics Graduate Capstone

Dr. William Sewell

November 28, 2021

**Objectives of Analysis**

Airbnb, Inc. was created in 2008 as Air Bed and Breakfast Inc. as an online marketplace for people to rent and book accommodations (Bloomberg, 2018). The Center for Business and Economic Research at the University of Nevada- Las Vegas states that “the driving force” of the Southern Nevada economy historically has been tourism as it “supports 32% of jobs in the region and generates more than $9.7 billion in gaming revenue to Clark County's economy” (UNLV, 2021). Tourism is a vital part of the economy in Clark County. The ability to properly price and anticipate future pricing of rentals will contribute to residents continuing to support the yearly tourism lodging needs.

This study will examine factors that are of most significance to pricing Airbnb rentals and will provide a model for predicting the price of a rental based on the most significant factors. In the study, Modelling Apartment Prices with the Multiple Linear Regression Model, Gustafsson and Wogenius use multiple linear regression to price apartments in Stockholm City Centre in Sweden. The factors they include in their analysis are “address, area, balcony, construction year, elevator, fireplace, floor number, maisonette, monthly fee, penthouse and number of rooms” (Gustafsson, 2014). Their analysis provides a basis for this study’s contributory factors. This study will include factors such as neighborhood, room type, bathrooms, bedrooms, beds, price per guest, minimum night stay, maximum night stay, availability, and instantly bookable.

The contribution of this study to the field of Data Analytics and the MSDA program will be a model and study to base other similar studies and data set questions on. It will provide the basis for additional studies on similar topics.

The analysis will use multiple linear regression. Linear regression is an appropriate method for this analysis because it can forecast trends in data that fulfill certain assumptions (which this data set does, as will be demonstrated). This technique uses many predicator variables to predict the value of a specific response variable, which is what type of variables contained in the data set.

Multiple linear regression requires a few assumptions: linearity (there exists a linear relationship between the dependent and independent variables), independence (the independent variables do not have a high correlation with each other), normality of residuals (the residuals are normally distributed), and homoscedasticity (the residuals in the model have constant variance).

This analysis will use R to run a univariate and bivariate analysis of the remaining variables after cleaning the dataset. It will also be used to verify that the resulting data will fit the assumptions needed to create a linear regression model to estimate correct pricing on rentals.

R is free and open source. It is a good resource for statistical analysis. It has the features needed to achieve the goals of this analysis quickly and efficiently. The main reason for using R is the author’s familiarity with its usefulness (R Advantages, 2021).

**Cleaning the Data Set:**

Inside Airbnb updates by scraping the Airbnb website annually (Cox, 2014). All information in the datasets is publicly displayed on the Airbnb website. The site contains approximately 650 data sets for use. It was last updated July 13, 2021 (Cox, 2014). The datasets are licensed under public domain. Data can be viewed at the following link: <http://insideAirbnb.com/get-the-data.html>.

The data set for Clark County, Nevada contains has 8,903 entries with 74 variables. The existing dataset required substantial cleaning. The data sparsity is 11.6%. 76,528 out of 658,822 had missing values. The price, bed, bedrooms, and bathrooms will have to be adjusted by number of guests to make the data more homogeneous. A new variable, the target variable, will have to be created as price per guest for each of the listings. Also, variables for the number of bedrooms, bathrooms, and beds will be created for each rental. The original dataset contains a considerable number of variables, many of which are not relevant to this analysis. Numerous variables needed to be removed before beginning a more detailed analysis. In the study, Modelling Apartment Prices with the Multiple Linear Regression Model, Gustafsson and Wogenius use multiple linear regression to price apartments in Stockholm City Centre in Sweden. The factors they include in their analysis are “address, area, balcony, construction year, elevator, fireplace, floor number, maisonette, monthly fee, penthouse and number of rooms” (Gustafsson, 2014). Their analysis provides a basis for this study’s contributory factors. This study includes factors such as neighborhood, room type, bathrooms, bedrooms, beds, price, minimum night stay, maximum night stay, availability, and instantly bookable.

A screenshot of a computer

Description automatically generated with medium confidence

**Univariate Graphs**

Chart, bar chart

Description automatically generatedChart, histogram

Description automatically generatedChart, histogram

Description automatically generatedChart, histogram

Description automatically generatedChart, histogram

Description automatically generatedChart, histogram

Description automatically generatedChart, histogram

Description automatically generatedChart, bar chart, histogram

Description automatically generatedChart, histogram

Description automatically generatedChart, histogram

Description automatically generatedChart, histogram

Description automatically generatedChart, waterfall chart

Description automatically generatedChart, histogram

Description automatically generatedChart, histogram

Description automatically generatedChart, histogram

Description automatically generated

**Gross Multiple Linear Regression Model**

After the data was cleaned, an initial model was created. The resulting model had a few variables that were linearly related to each as well as had a large number of independednt variables. A reduced model would be more likely to fulfull the assumptions of linear regression, reduce the variables one would need to use, and could still maintain an acceptable level of accuracy.

Text

Description automatically generated

**Reduced Linear Regression Model**

To create a reduced linear regression model, variables with no predictive value and those that were highly correlated with others were dropped. After a primary component analysis to detemine the most important factors, regression was repeated to reduce variables until R2 was acceptably close to adjusted R2.

The initial model, gave the following information:

­­Text

Description automatically generated

First, SharedRoom and North\_LV were linearly related to other predictor variables. Those were dropped.

Next, a principal component analysis was used to determine the most important variables.

A picture containing text

Description automatically generated

A screenshot of a computer

Description automatically generated with medium confidence

Seven of the primary components have eigenvalues above 1 but it takes ten components to get a cumulative variance above 0.8. The next step was identifying the components that correspond to those values.

Text

Description automatically generated with medium confidence

The most important factors are number of people the rental accommodates, number of bathrooms per guest, number of beds per guest, number of bedrooms per guest, rental type being entire facility or private room, location in Las Vegas city, availibility per year, location in an unicorporated area of Clark County, and the minimum number of nights required.

Chart, histogram

Description automatically generated

A screenshot of a computer

Description automatically generated with medium confidence

After a biplot to see correlation between variables, the variables that were highly correlated were dropped. That left 4 independent variables that formed the final reduced equation: number of beds per guest, rental type being entire facility, availibility per year, and the minimum number of nights required.

Table

Description automatically generated

**Price\_per\_guest** ≈ **33.6897 +22.0998X1 – 10.6294X2 + 0.0837X3– 0.4084X4**

(Where X1=Number of Beds per guest, X2=Entire Renatl, X3=Number of nights available per year and X4=Minimum required night stay)

The graphs of the residuals showed that the residulas were varied and were approximately normal. This final reduced equation met the assumptions and had aceptable R2 .

Chart

Description automatically generatedChart, histogram

Description automatically generated

**Conclusion**

This analysis can help provide an estimate for pricing on Airbnb rentals in Clark county. There are conditions that have greater significance than others on price: number of people the rental accommodates, number of bathrooms per guest, number of beds per guest, number of bedrooms per guest, rental type being entire facility or private room, location in Las Vegas city or in an unincorporated area of Clark county, availibility per year, and the minimum number of nights required for rental. There are a number of limitations of this analysis. Some of the limitations of this study include that there is only limited data on the rentals. The study examines more of a snapshot of the rental availability at the time of web scraping and not how their price has changed over time. Another study at another point in time would reveal further accuracy in a model. This study includes Clark County, Nevada instead of a larger sample of cities. Partly this choice is to focus on an area where tourism is the biggest portion of the economy, Las Vegas and the surrounding areas. Narrowing the focus to the area specifically being analyzed would provide a narrower model. Also, this study does not include anything related to reviews as it will be focused on early parts of the process contributing to pricing. Pictures and host demographic information may have influence on the price, but this study will be restricted to more quantifiable measurements.

**Works Cited**

"Company Overview of Airbnb, Inc". Bloomberg L.P. January 7, 2018. Archived from the original on January 8, 2018. Retrieved August 21, 2021 from <https://www.bloomberg.com/research/stocks/private/snapshot.asp?privcapId=115705393>.

Cox, Murray. 2014. Inside Airbnb. Retrieved August 21, 2021 from <http://insideAirbnb.com/>.

Gustafsson, Alexander, and Sebastian Wogenius. Modelling Apartment Prices with the Multiple Linear Regression Model. Royal Institute of Technology School of Engineering Sciences, 2014, <https://www.divaportal.org/smash/get/diva2:725045/FULLTEXT01.pdfMultiple>.

R advantages and disadvantages. www.javatpoint.com. (n.d.). Retrieved October 26, 2021, from <https://www.javatpoint.com/r-advantages-and-disadvantages> .

University of Nevada-Las Vegas (UNLV) Center for Business and Economic Research. (2021). “Southern Nevada Economy” [Webpage]. Retrieved from <https://cber.unlv.edu/SNBDI/economy.html> .