**Linear Regression Model on Airbnb Pricing**

**for Clark County Nevada**

**Executive Summary**

**Overview:**

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 examined factors that are of most significance to pricing Airbnb rentals and provides a model, using R and multiple linear regression, for predicting the price of a rental based on the most significant factors. This analysis provides an estimate for pricing on Airbnb rentals in Clark county.

**Cleaning the Dataset:**

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, was created as price per guest for each of the listings. Also, variables for the number of bedrooms, bathrooms, and beds were 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.

**Primary Component Analysis:**

PCA revealed seven of the primary components have eigenvalues above 1 but it takes ten components to get a cumulative variance above 0.8.

Chart, histogram

Description automatically generated

These ten variables were: 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.

**Reduced 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.

Variables with no predictive value and those that were highly correlated with others were dropped. Regression was repeated to reduce variables until R2 was acceptably close to adjusted R2.

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. **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

**Limitations of Study and Suggested Actions:**

* 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 combined with the data from other web scraping dates 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.
* 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.

**Sources**

"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> .