Group Members: Elizabeth Binkina, Riya Trikha , Caleb Yu

# Research Question

We want to know if we can use specific features of an Airbnb to predict the nightly price in the greater Los Angeles area. We will be using two datasets for this project:

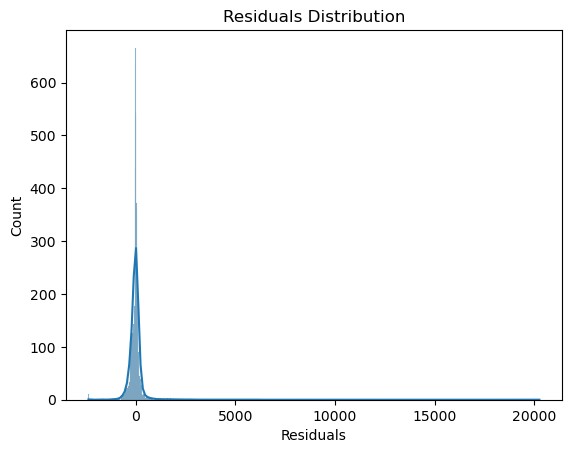
Dataset 1: <https://www.kaggle.com/datasets/oscarbatiz/los-angeles-airbnb-listings>

Dataset 2 : <https://insideairbnb.com/get-the-data/> (listings.csv) (neighbourhoods.csv)

The first dataset is from ‘Kaggle’ and it includes 25 columns that have data about Airbnb’s in Los Angeles. The second dataset is from ‘Inside Airbnb’ and it includes multiple cities but we will be focusing on the Los Angeles files that include detailed Airbnb listings and neighbourhoods. Specific features in an Airbnb that we are looking at for this project are the number of bathrooms, number of bedrooms, and the neighborhood it is located in.

# Preprocessing Steps

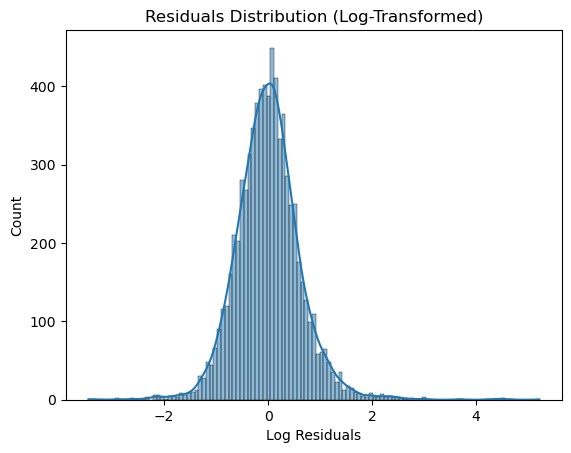
Some necessary data preprocessing steps that needed to be completed before we started anything was choosing relevant columns that we wanted to analyze since the dataset contained many columns that we cannot handle at once. We also needed to remove rows within the columns we selected since it includes missing or invalid values. Next, we converted the price variable into numeric values as the original dataset contained them as values containing ‘$” and commas. We fit a model using bathrooms, bedrooms, and accommodates as X and price as Y. This model will check whether or not we need to apply transformations to the predictor to create a normal distribution.



Skewed Model Performance:

Skewed R²: 0.27

Skewed RMSE: 636.27



Log-Transformed Model Performance:

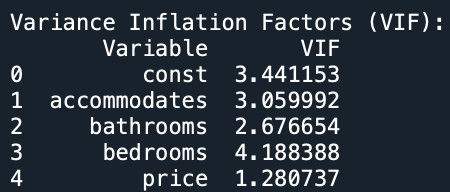
Log R²: 0.49

Log RMSE: 0.62

The first plot above shows the initial model using just price as a predictor. It is clear that it is a right-skewed residuals distribution so the model is not as accurate as we would expect it to be. To improve the model, we perform a log transformation on the predictor variable, price.

The second graph above shows the improved residuals plot as it is normally distributed now. This concludes that the model with log(price) as predictor is better fit for our data and we proceed with using log(price).

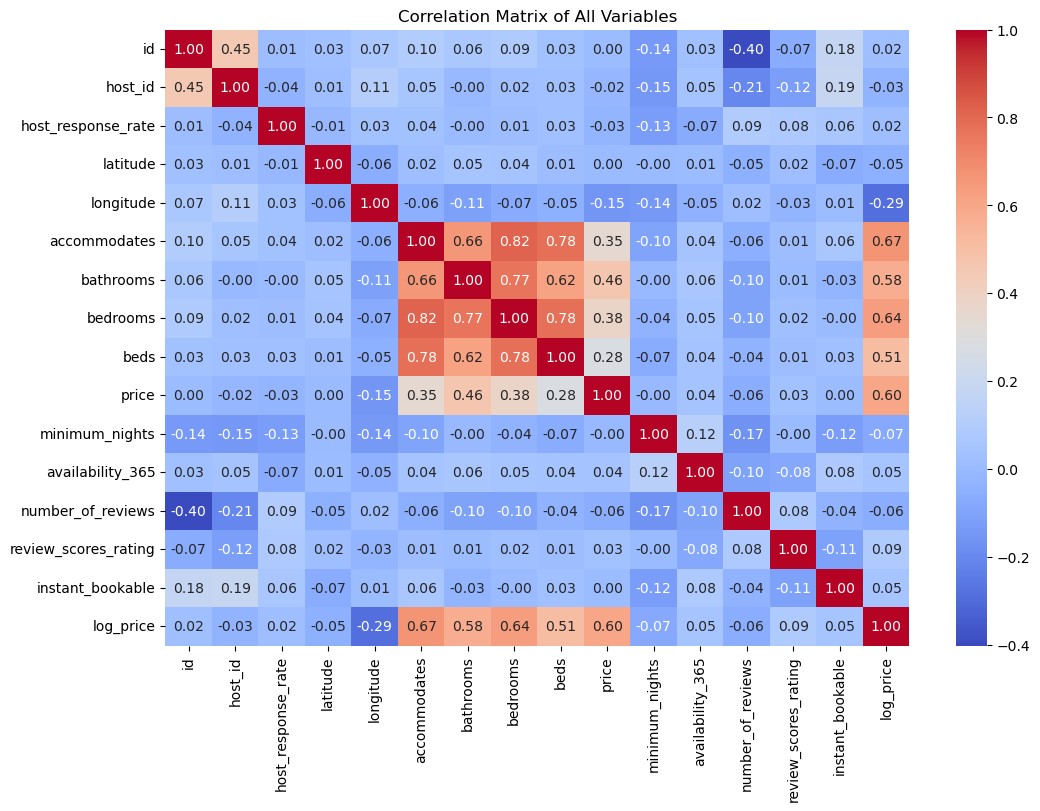
Another preprocessing step that needs to be completed before we can begin our modeling and analysis is to check the variation inflation factor for all the features we are interested in and confirm that no multicollinearity exists. We check the variance inflation factors for the accommodates, bathrooms, bedrooms, and price variables.



The output above shows most of the VIF values to be between 1-5 which indicates potential correlation between the variables but not enough evidence to imply the multicollinearity exists. However, bedrooms has the highest variance influence factor score at 4.188 which we will look into further in this analysis as to why this may be. For now, we can proceed with our modeling and analysis.

# Correlation Analysis

We want to find the highest correlated variables with log(price) as the predictor variable. In order to do this, we find the correlation matrix and plot a correlation heat map to find the highest correlated variables.



The correlation heat map above shows the four variables with the strongest correlation with log\_price are accommodates, bathrooms, bedrooms, and beds. These four variables are also highly correlated with price if no log transformation was applied. Next to view how a specific variable affects price when other variables are held constant, we fit the model and find the coefficients for the model.

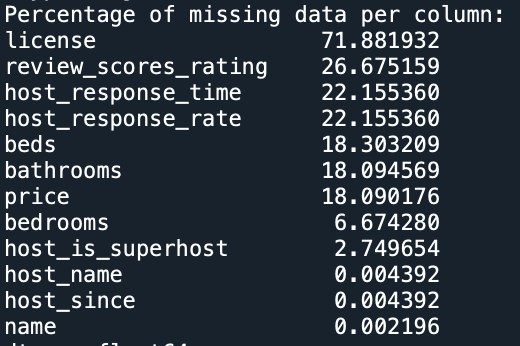




Based on the coefficients, we have an idea of how each variable affects the price. For every unit increase, accommodates will increase the price by $231 and bathrooms will increase the price by $22.The bedrooms coefficient seems to bring up a negative coefficient which could be explained by a possible relationship with accommodates which we will go into in the Hypothesis Test.

# Exploratory Data Analysis

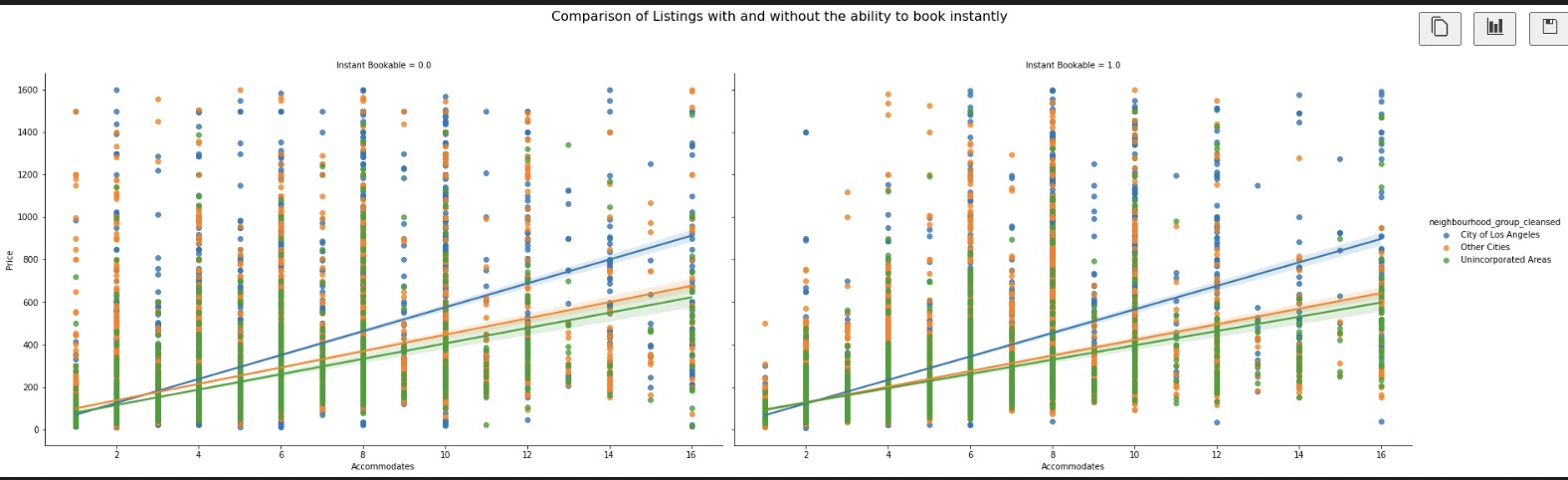
We checked the columns with the most missing data and got the total percentage.



The output above shows the column license, to have the greatest percentage of missing values which we would expect to have poor model accuracy. None of the features we are interested in are in the highest percentage of missing data so we can continue.

## Accommodates

One of the features we identified that we wanted to look into was accommodates as it was highly correlated with price.

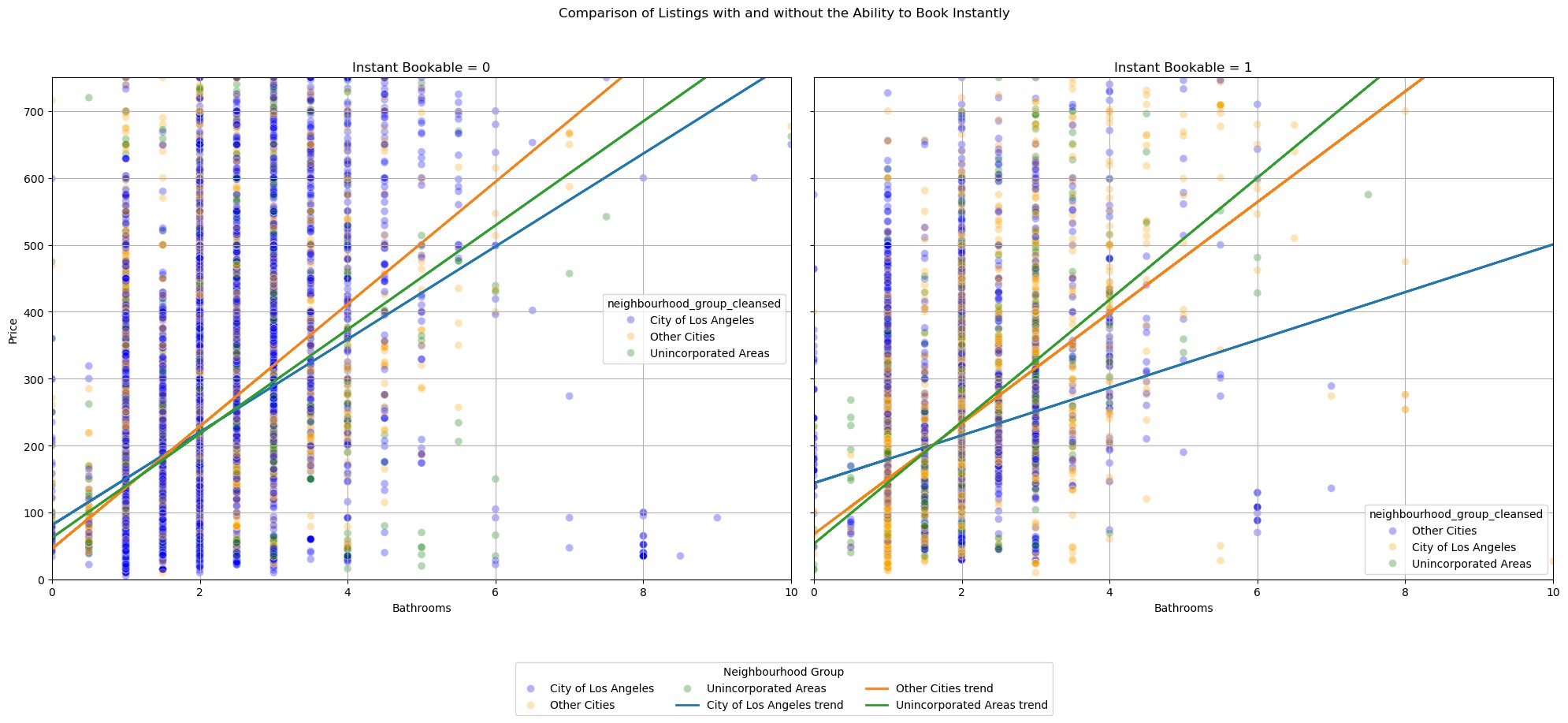


The plot above shows accommodates to have a strong linear relationship with price when we split based on Instant Bookability and group by neighborhood\_group\_cleansed. This indicates that the variable, accommodates, has a strong influence on predicting the price of an

Airbnb.

## Bathrooms

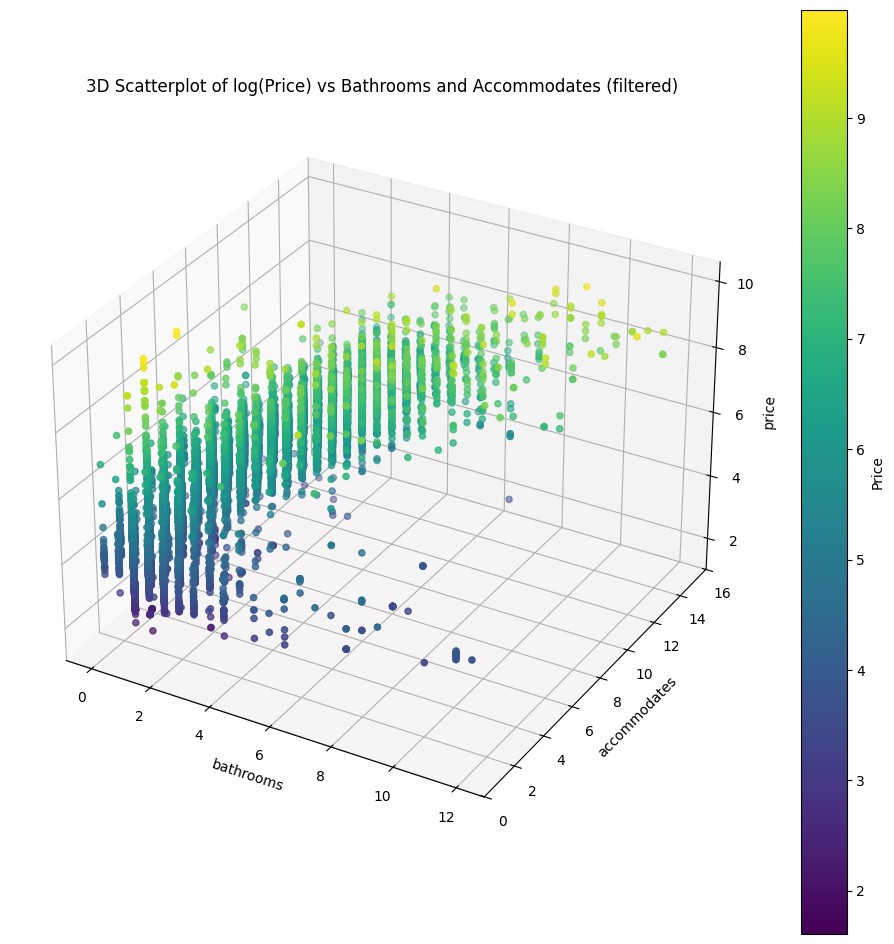
Another of the features we identified that we wanted to look into was bathrooms as it was highly correlated with price.



The plot above shows bathrooms to have a strong linear relationship with price and could indicate that bathrooms has an influence on predicting the price of an Airbnb.

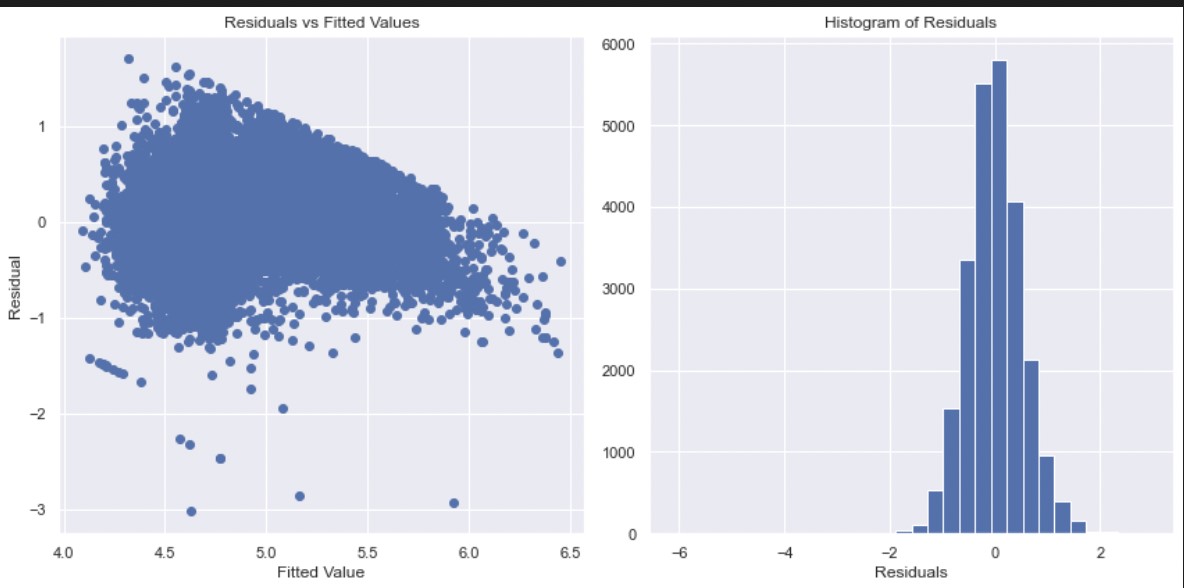
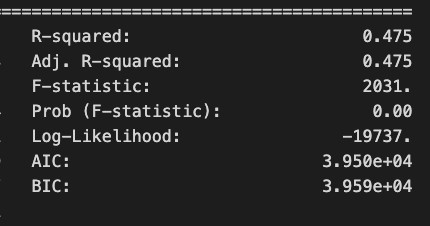
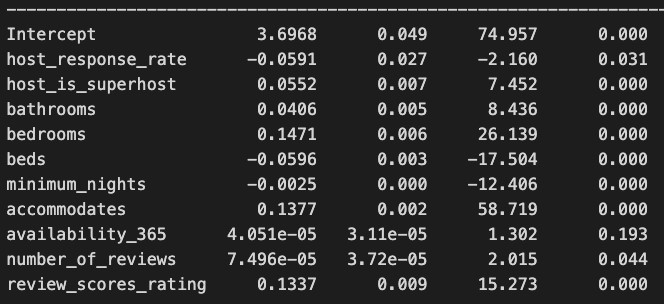
## 3D Plot

Another way we wanted to visualize our data is by comparing log(price) to bathrooms and accommodates. Which can be expressed by the plot below:



We can see that most of the data points are in the 1 - 6 bathroom range while accommodates have a wide range. This is expected since most houses will usually not have more then 6 bathrooms and accommodates can range based on various factors. We can see that there is a gradual increase to price as we increase the number of bathrooms and accommodates. This helped us to believe that both bathrooms and accommodates may have a significant effect on price and may be good predictor variables for our model.

# Logistic Regression



**Equation :** log(Price) = 3.697 - 0.0591(host\_responce\_rate) + 0.0552

(host\_is\_superhost) + 0.0406(bathrooms) + 0.1471(bedrooms) - 0.0596(beds) - 0.0025

(minimum\_nights) + 0.1377 (accommodates) + 0.1337(review\_score\_ratings)

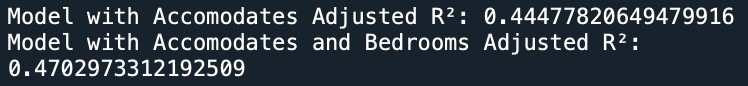
We removed all statistically insignificant attributes that had p-values lower than alpha (0.05) . We removed the Availability (how many days of the year the airBnb is available for) feature and the number of reviews feature, as they added no value to the model. According to the logistic regression model, the most significant predictors are accommodation, number of bedrooms, reviews, and ratings.

# Hypothesis Test

The bedrooms and accommodates variables are conceptually related as the amount of people an Airbnb accommodates typically will rise with the number of bedrooms an Airbnb has.

These two factors were also highly correlated at 0.818, indicating a strong positive relationship. The variance influence factor value for bedrooms was also the highest at 4.188. Thus, we create a hypothesis to check whether there is a significant difference in the model with accommodates and a model with accommodates and bedrooms.

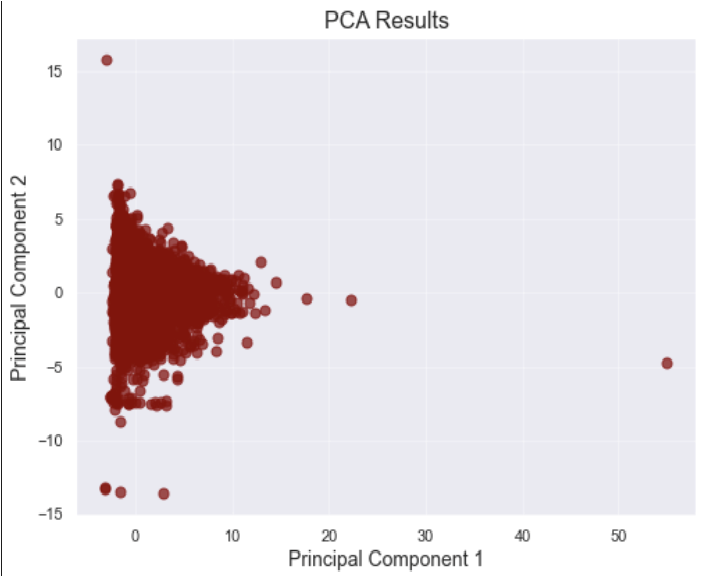
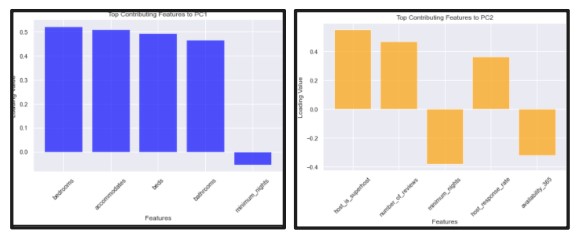




The output above shows that the adjusted R^2 score between the two models is slightly different from 0.445 to 0.470. This means that the model with the bedrooms variable included is a slightly better fit for the data than just accommodates but since it is a very slight increase, further analysis needs to be done to confirm the actual effect of bedrooms on the predictor

variable.

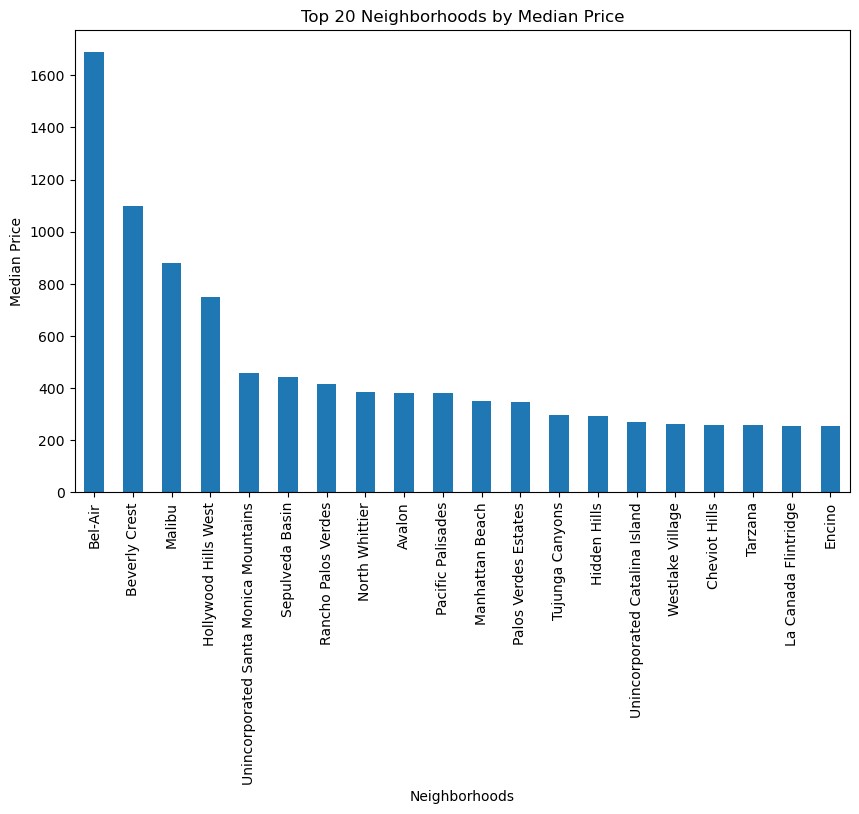
# Principal Component Analysis



PC 1 captured variation in features that relate to property size and capacity. Thus, it contained properties with a higher number of bedrooms, beds, bathrooms, and can accommodate more people. PC 2 captures variation in features that relate to host responsiveness and booking activity. Thus, it contains properties with hosts that have a higher response rate, more reviews, more flexible booking terms , and a lower minimum number of nights. Therefore, the main features explaining the variation in price is the physical attributes of the property and the host engagement and booking flexibility.

# Cluster Analysis - Joined Datasets

One of the features we wanted to research about to see if it is influential in predicting AIrbnb prices in Los Angeles is the neighborhood the Airbnb is located in. In order to measure if neighborhoods are significant or not, we need to merge the two datasets we found by a column that is similar on both datasets. After analyzing both datasets, ‘id’ is the similarity between them so we decide to merge based on Airbnb's id, which is an unique id that is assigned to every Airbnb property listed. Once we have merged the datasets, we want to see the top neighborhoods with the highest Airbnb price. However, in the datasets there are some extreme outliers and zero values so we cannot use the mean price per neighborhood as it is skewed. We opt to use the median house price instead to have a more accurate representation of neighborhood pricing for Airbnbs.



The plot above shows the top 20 neighborhoods by median Airbnb price. Bel-Air is listed as the neighborhood with the highest median priced Airbnbs. Once we explored the highest median price per neighborhood, we apply the log(price) regression model to each neighborhood in order to see how accurately the model can predict the price based on the identified features

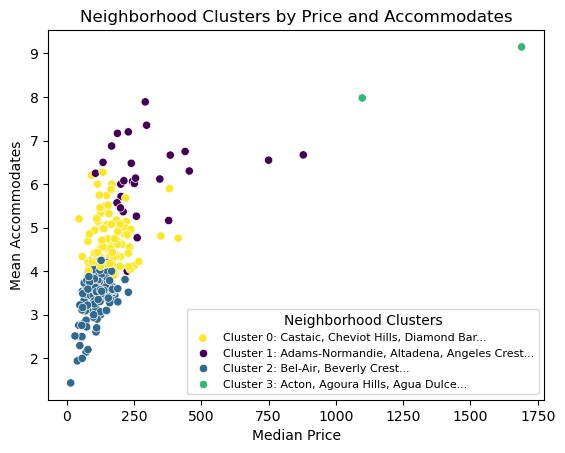
(bedrooms, bathrooms, accommodates).

R Output:

Highest R² Score: Glassell Park with R²= 0.92

Lowest R² Score: Veterans Administration with R²= -3.78

Based on the R output for the neighborhood with the highest and lowest R² value, we can see that the model does well in predicting the price for neighborhoods like Glassell Park at 0.92. Other neighborhoods like Veterans Administration, we can see that the model did very poorly in predicting the pricing. The range of R² values is pretty large, suggesting that neighborhoods may not be a significant influential feature on predicting prices for Airbnb’s in Los Angeles. Additionally to check the influence of neighborhoods, we can perform a k-means cluster analysis using the features we selected.



The plot above shows the 4 neighborhood clusters based on the median price and mean accommodates. Cluster 3 consists of a few outliers which could be inferred as luxury Airbnb listings. A common theory may be that a more luxurious neighborhood would mean higher Airbnb pricing since the houses are also more expensive. So in order to determine if neighborhoods influence price, we took a sample neighborhood from cluster 0 and cluster 1 and checked on ‘Zillow’ to see if their average house price were different. We picked Cheviot Hills from Cluster 0 and Agoura Hills from Cluster 3.

Zillow Findings:

Cheviot Hills: Average House Price ~ $2.2 millions <https://www.zillow.com/home-values/403142/cheviot-hills-los-angeles-ca/>

Agoura Hills: Average House Price ~ $1.2 million <https://www.zillow.com/home-values/9840/agoura-hills-ca/>

Based on a random sample from cluster 0 and cluster 3, it displays that the average house cost in that specific neighborhood may not be correlated with pricing of an Airbnb. This can conclude that the variable, neighborhoods, is not as significant of an influence on the predictor variable, price, as accommodates is. Further testing can be completed for this analysis like having a neighborhood specific model in order to make it more accurate in predicting pricing. Other confounding variables that may not be included in this model include safety and proximity to attractions.

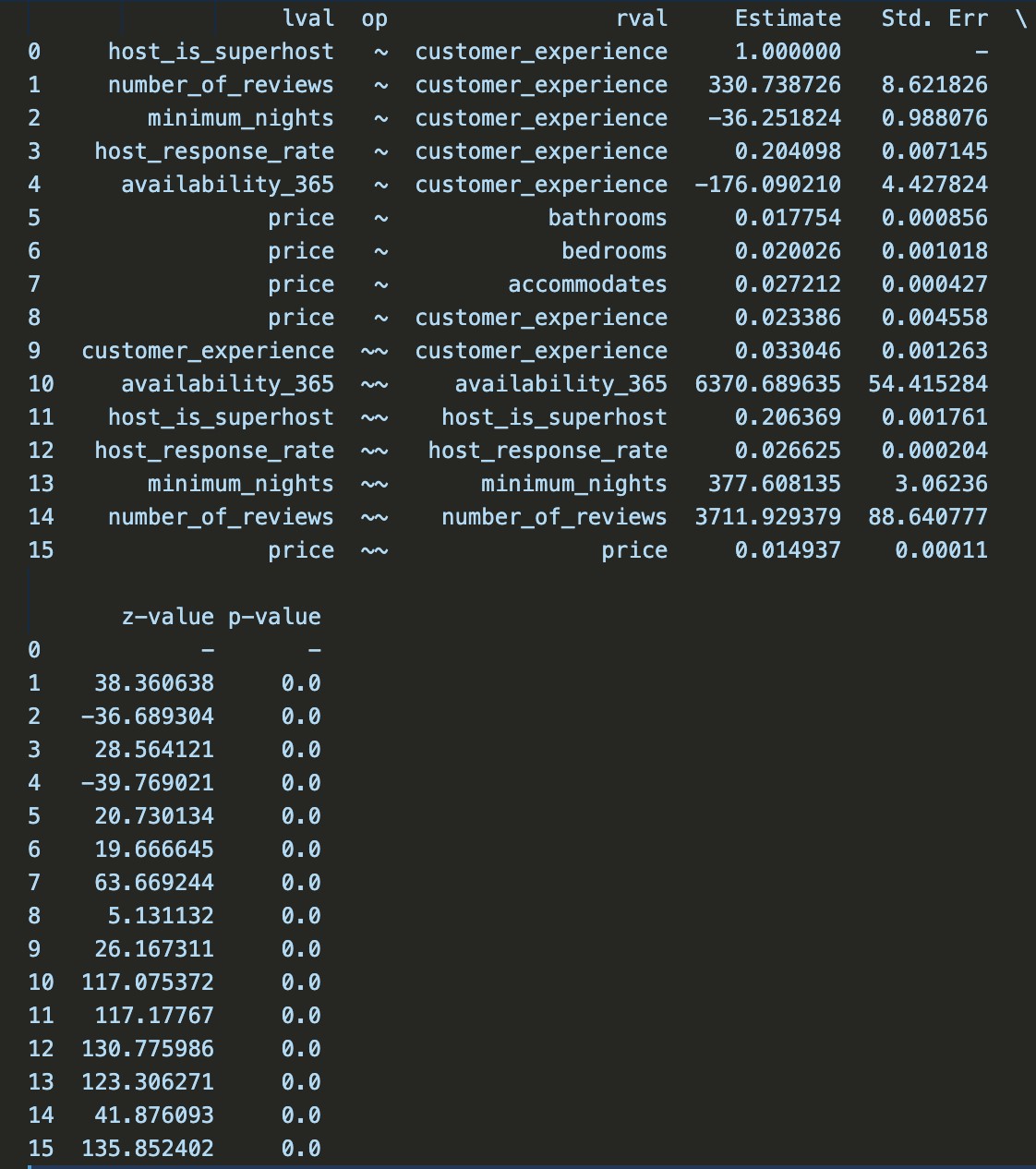
# Structural Equation Modeling

In our Structural Equation Modeling analysis, we wanted to see how might unobserved data be affecting the price of an Airbnb. To do this, we needed to create a latent variable. A latent variable is a variable that is unseen and not measured in our dataset. In many cases, the latent variable is created by utilizing many similar observed, measurable variables from our dataset. We decided to utilize the PCA results to get a good idea of what measurable variables were contributing to most of the variance in the data. We ended up creating a latent variable called customer\_experience. This variable consists of the observed variables; host\_is\_superhost (indicates whether the host is a Superhost (True/False)), number\_of\_reviews (total number of reviews received for the property), minimum\_nights (minimum number of nights required for a booking), host\_response\_rate (percentage of guest inquiries that the host responded to), and availability\_365 (number of days the property is available for booking in the next 365 days). To use this latent variable in our model, we needed to ensure that the variables were numerical. Fortunately, only the host\_is\_superhost variable was non-numeric so we converted it to either 0 (false) or 1 (true). Using this latent variable, we created the following model:

**price ~ bathrooms + bedrooms + accommodates + customer\_experience**

And by using the semopy library, we created the model and printed the following summary

results:



With a significance level of ɑ = 0.05, we can see that bathrooms, bedrooms, accommodates, and customer\_experience have a p-value < 0.05, therefore, they are significant to the model. From this result we can create the following equation for our latent variable:

**customer\_experience = (1) \* host\_is\_superhost + (330.738726) \* number\_of\_reviews + (-36.251824) \* minimum\_nights + (0.204098) \* host\_response\_rate + (-176.090210) \***

## availability\_365

From this equation, we can see that increasing number\_of\_reviews, host\_response\_rate, and being a super host will result in a higher customer\_experience value. While, increasing minimum\_nights and availability\_365 will result in a lower customer\_experience value. Also from this result, we can express our model in the following equation:

**price = (0.017754) \* bathrooms + (0.020026) \* bedrooms + (0.027212) \* accommodates +**

## (0.023386) \* customer\_experience

We can see that increasing the number of bathrooms, bedrooms, accommodates, and customer\_experience will result in a higher priced Airbnb, while decreasing them results in a lower price. It seems that out of these variables, that increasing accommodations will increase the price most as its coefficient value is the highest at **0.027212.**

# Conclusion

Based on our results, it is evident that variables such as accommodates, bathrooms, and bedrooms are strong predictors of price, with accommodates being the most influential predictor. This suggests that Airbnb listings with many bathrooms, bedrooms, and high accommodation will result in a higher cost.

This information may be important and useful for many different groups of people. For example, it could be used for customers that would like to budget estimates of living cost for their vacations. It could also help Airbnb hosts determine a good market price for their listing.

Our results can help with prediction of price, but are also somewhat limited. For example, the datasets are limited to the greater Los Angeles area, so it would not be a great predictor of an Airbnb's price around the world. Therefore, although these results can aid people it is not exhaustive and further research will be needed in order to predict the nightly price to stay at an Airbnb.