**OPIM 5606 – Predictive Modeling**

**TEAM 1**

**FINAL PROJECT WHITE SHEET**

**Topic: Airbnb in NYC 2019**

**Summary:**

Our project is about analyzing the Airbnb dataset for listings in New York City to identify trends and provide insights by using Modeling and Data Visualization. The dataset contains 16 columns that include information about price, neighbourhood, room type, reviews, availability, etc. Our goal is to predict the price based on the data available and then select the best fit model to predict house price and determine the factors that affect the same. This information will help businesses to make predictions for valuable business strategies.

**Problem Statement:**

Airbnb has been one of the most popular online platforms for listing and renting local homes. It provides a community-based customer-to-customer service that connects local hosts and travelers. Airbnb enables property owners to rent out their vacant rooms to the renters who are looking for suitable rental homes. This business model creates opportunities for property owners to make profits on the vacant spaces and offers easy access for travelers to local homes.

Since Airbnb is a platform for customer-to-customer services and transactions, the company has limited power in monitoring host service and room qualities. To ensure quality services and long-lasting business presence, Airbnb needs to identify the characteristics that most influence the price and to be able to predict the future price. This will enable the business to understand the unknown factors for more accurate pricing commensurate with the profit.

**Business Case:**

Currently, the company has the data which records the information of hosts in NYC before 2019. We would like to predict the future 2022 Airbnb performance in NYC. More specifically, we will try to determine relationships among rent prices, locations, number\_of\_reviews, room\_type, availability, etc, and then select the best fit model to predict house price.

**Sample**

This data was taken from kaggle and it contains information of the airbnbs that are present in the city of New york and are utilized by the people for the year 2019.

**Explore**

Columns selection:

For this dataset, we had 16 columns in total.

A picture containing text

Description automatically generated

Here is the data dictionary:

**id:** Airbnb's unique identifier for the listing

**host\_id**: Airbnb's unique identifier for the host

**latitude:** Uses the World Geodetic System (WGS84) projection for latitude and longitude.

**longitude:** Uses the World Geodetic System (WGS84) projection for latitude and longitude.

**price:** Daily price in local currency

**minimum\_nights:** Minimum number of night stay for the listing (calendar rules may be different)

**number\_of\_reviews:** The number of reviews the listing has

**last\_review:** The date of the last/newest review

**review\_per\_month:** The number of reviews the listing has over lifetime of the listing

**calculated\_host\_listings\_count:** The number of listings the host has in the current scrape, in the city/region geography.

**avalibility\_365:** The availability of the listing 365 days in the future as determined by the calendar. Note a listing may not be available because it has been booked by a guest or blocked by the host.

**name:** Name of the listing

**host\_name:** Name of the host. Usually just the first name(s).

**neighbourhood\_group:** The neighbourhood group as geocoded using the latitude and longitude against neighborhoods as defined by open or public digital shapefiles.

**neighbourhood:** The neighbourhood as geocoded using the latitude and longitude against neighborhoods as defined by open or public digital shapefiles.

**room\_type:** All homes are grouped into the following three room types:Entire place, Private room, Shared room and Entire place

Based on the data dictionary and what conclusions we are interested in, we choose to hide some of the columns which we thought of accordingly, which is id, name, host\_id, host\_name, neighbourhood, latitude, longitude, last\_review and reviews\_per\_month. Because when we considered these columns were increasing the complexity of the analy part as most of them were having unique values which would ultimately make the analysis part difficult. And here are the reasons for us to exclude every column:

**id:** We can see from the data dictionary and the dataset itself that id is only the unique identifier for the listing, which will not create any interesting conclusion for us.

**name:** According to the data dictionary, it is the name of each listing, every name is different, so they do not have much value.

**host\_id:** Same with id, we can see from the data dictionary and the dataset itself that host\_id is only the unique identifier for the host, which will not create any interesting conclusion for us.

**host\_name:** Same with name, it is the name of each host, every name is different, so they do not have much value.

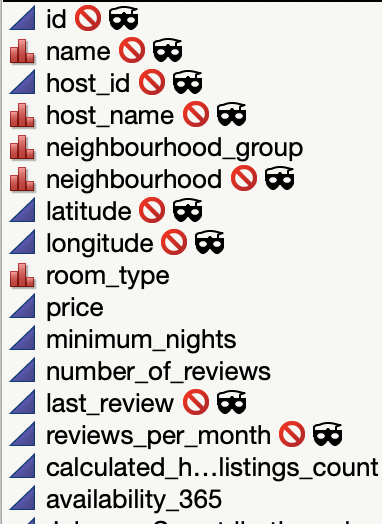
**neighbourhood**: There are too many different values of neighbourhood. Since the neighbourhood\_group have the same function with neighbourhood but less values, we choose to exclude this column

**latitude & longitude**: The latitude and the longitude of each listing tell us the location of each listing. Since the value of all of the listings is very similar and we already choose to keep neighbourhood\_group, we will exclude this column.

**last\_review**: The last\_review is the date of the last review, which is not valuable. And since this column has over 10000 missing values that can not be imputed, we decided to exclude this column.

**review\_per\_month**: This column is derived from the number\_of\_reviews column. So we decided to exclude this column and to keep the number\_of\_reviews.

After excluding columns, here are the left columns which we would like to keep.



Reasons for us to keep the following columns:

**neighbourhood\_group:** important geographic information that would help us to understand price, rating, and availability distributions at different areas.

**room\_type:** different types of room may vary on prices and popularities. It might be one of key indicators for ratings and more.

**price:** one of the most valuable variables in this dataset. We could use this variable to explore customers’ willingness to pay based on differences in other variables, like room\_type, and neighborhood\_group.

**minimum\_nights:** we may explore the relations between minimum\_nights requirement to the ratings and popularities.

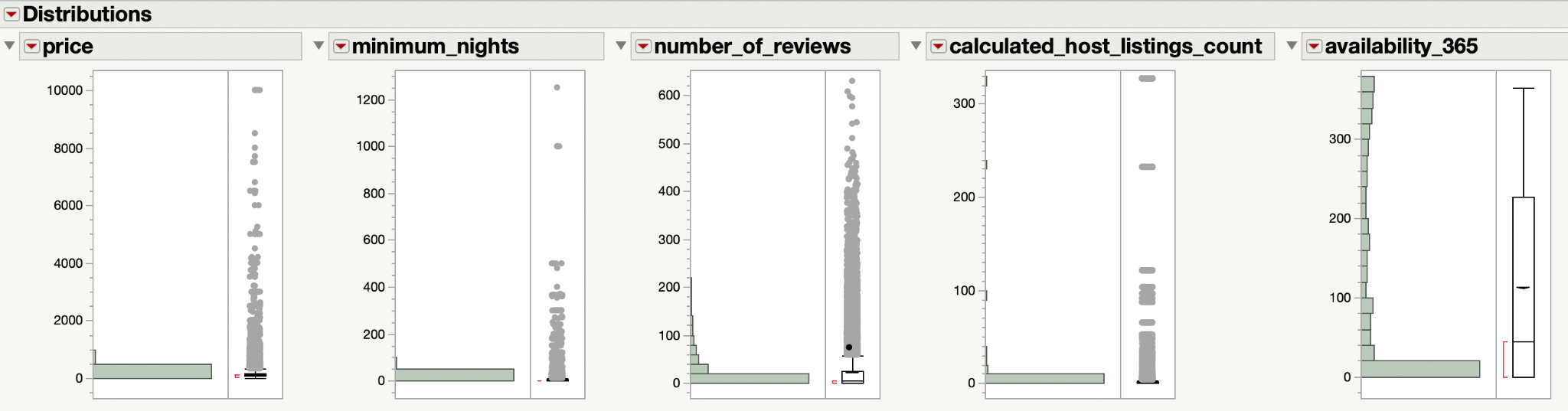
**number\_of\_reviews:** this might be a good indicator telling us which host, or what type of room, or what area is most popular.

**calculated\_host\_listings\_count:** we could try to explore if host\_listing\_counts has anything to do with its popularities, pricings, and its relation to geographical locations.

**availability\_365:** more days a room is available for customers, higher the chance it may receive reviews. And there might be correlation between pricings and room\_type to availability that we could work on to find out.

Outlier Exploration:

According to the distribution of the numerical columns we kept, we can see that price, minimum\_nights, number\_of\_review, caculated\_host\_listings\_count have outliers. So we need to modify them on the next step.



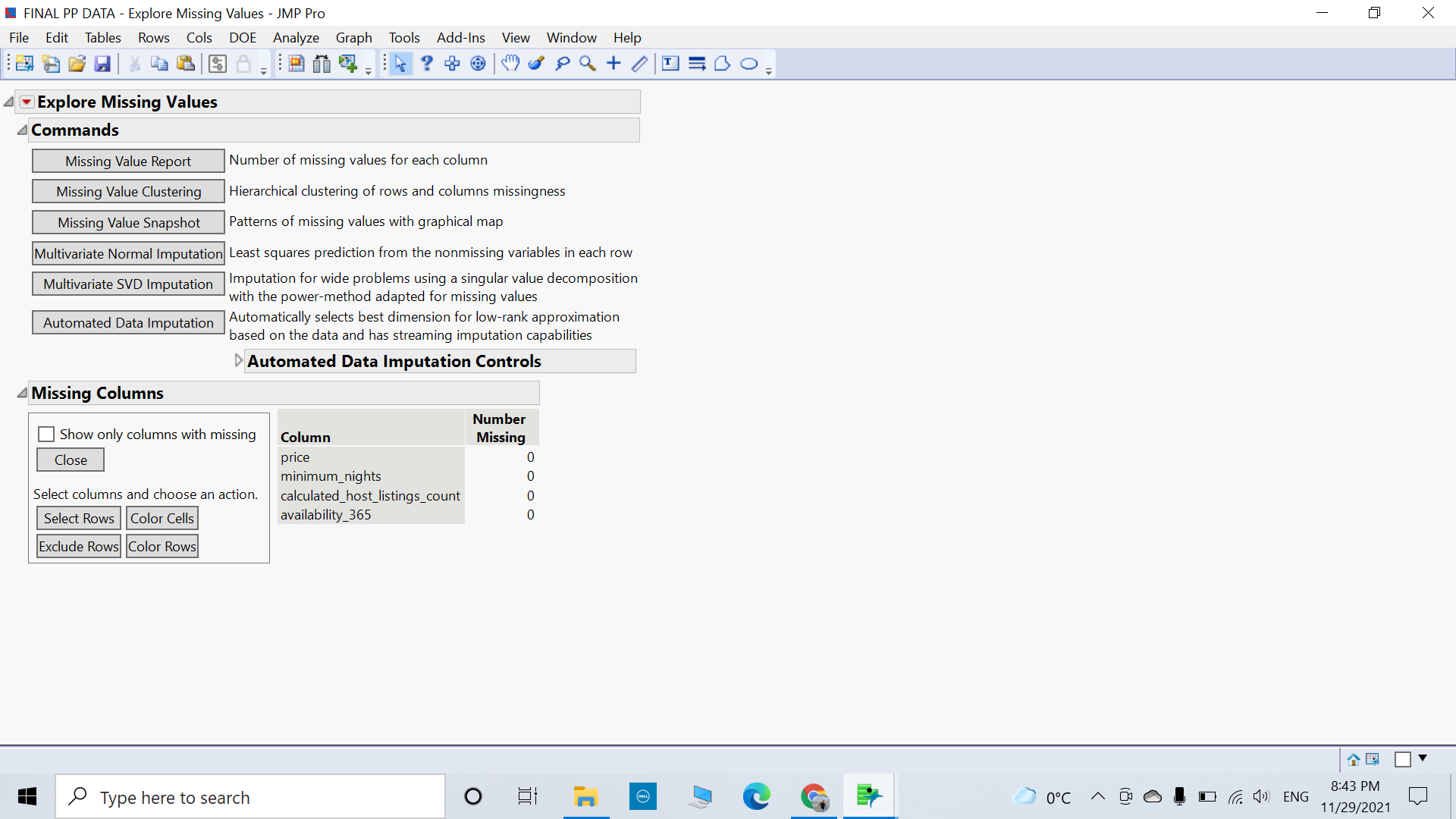
**Modify**

Modification of the data is done by performing the outlier analysis and exploring for missing values. After doing the above process, we apply principal component analysis between variables. Then we build graphs for different categories comparing hosts, a neighborhood group. prices, room type, and minimum nights

**Missing Values**

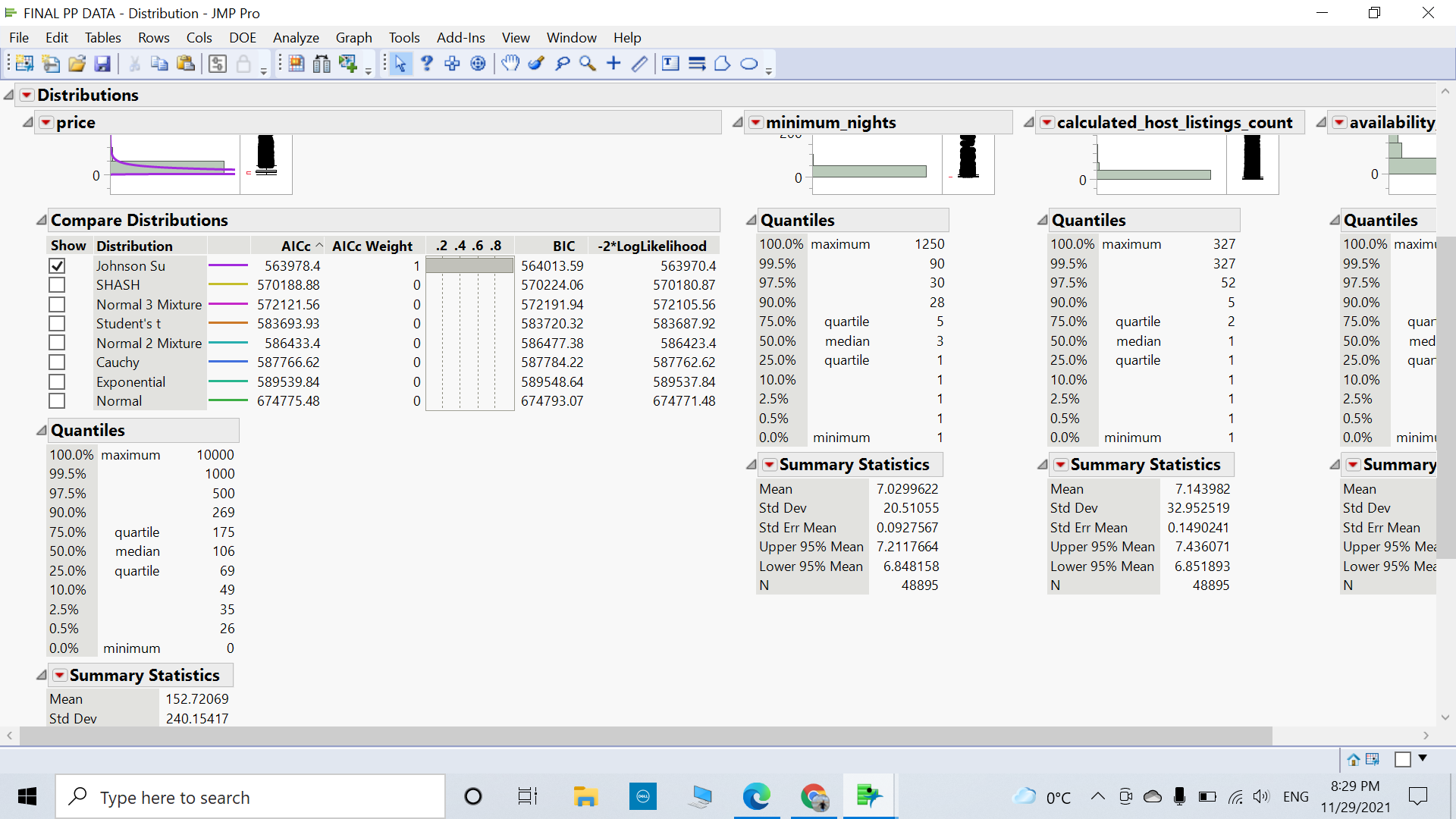
Before solving for the outliers we solved for the missing data values and found out that none of the columns had any missing values. To find out the missing values we took 4 variables that are price , minimum\_nights, calculated\_hosts\_listings\_count , availability\_365

And tried to find out for the missing values in these columns.

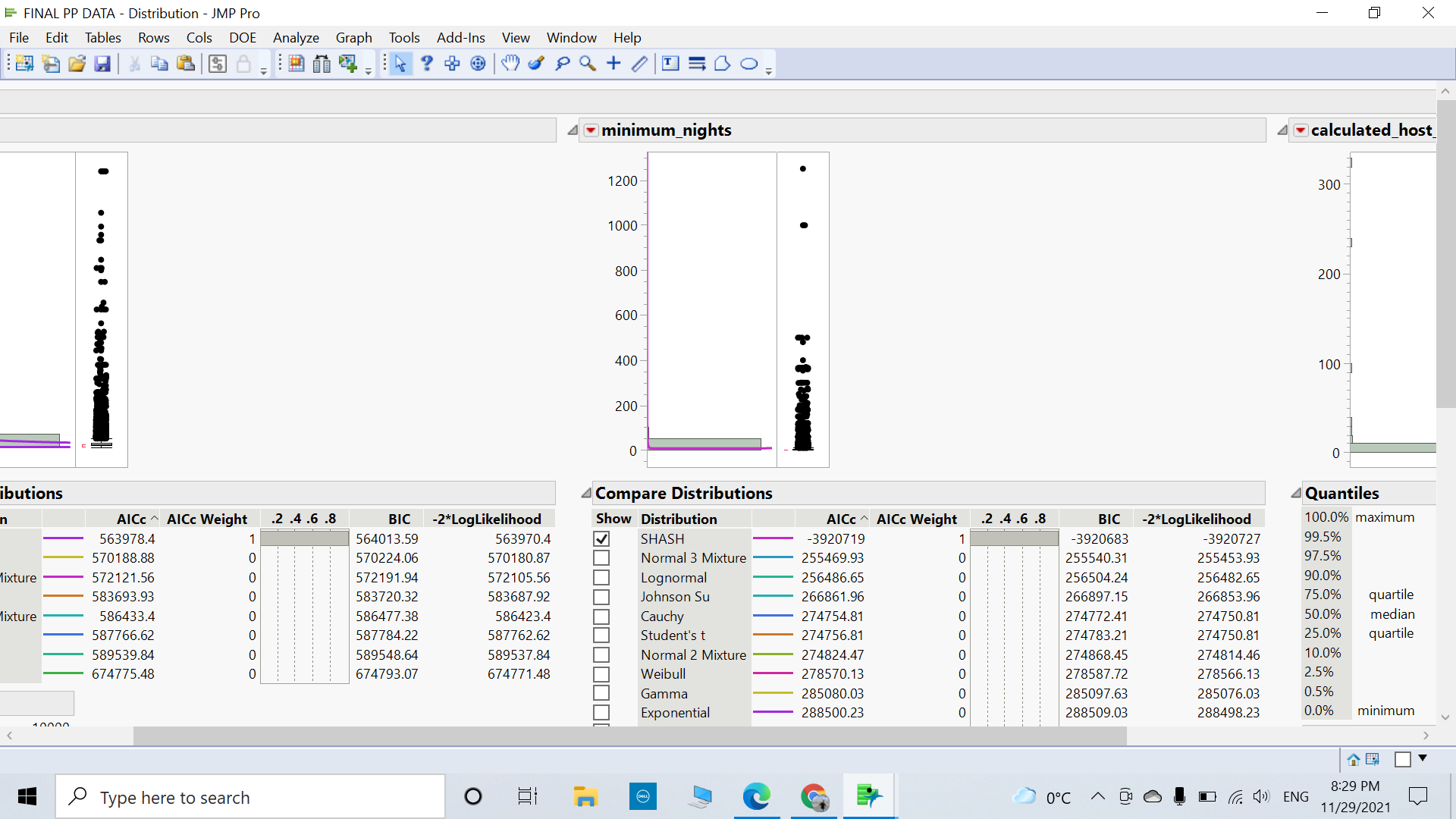


Outlier Analysis

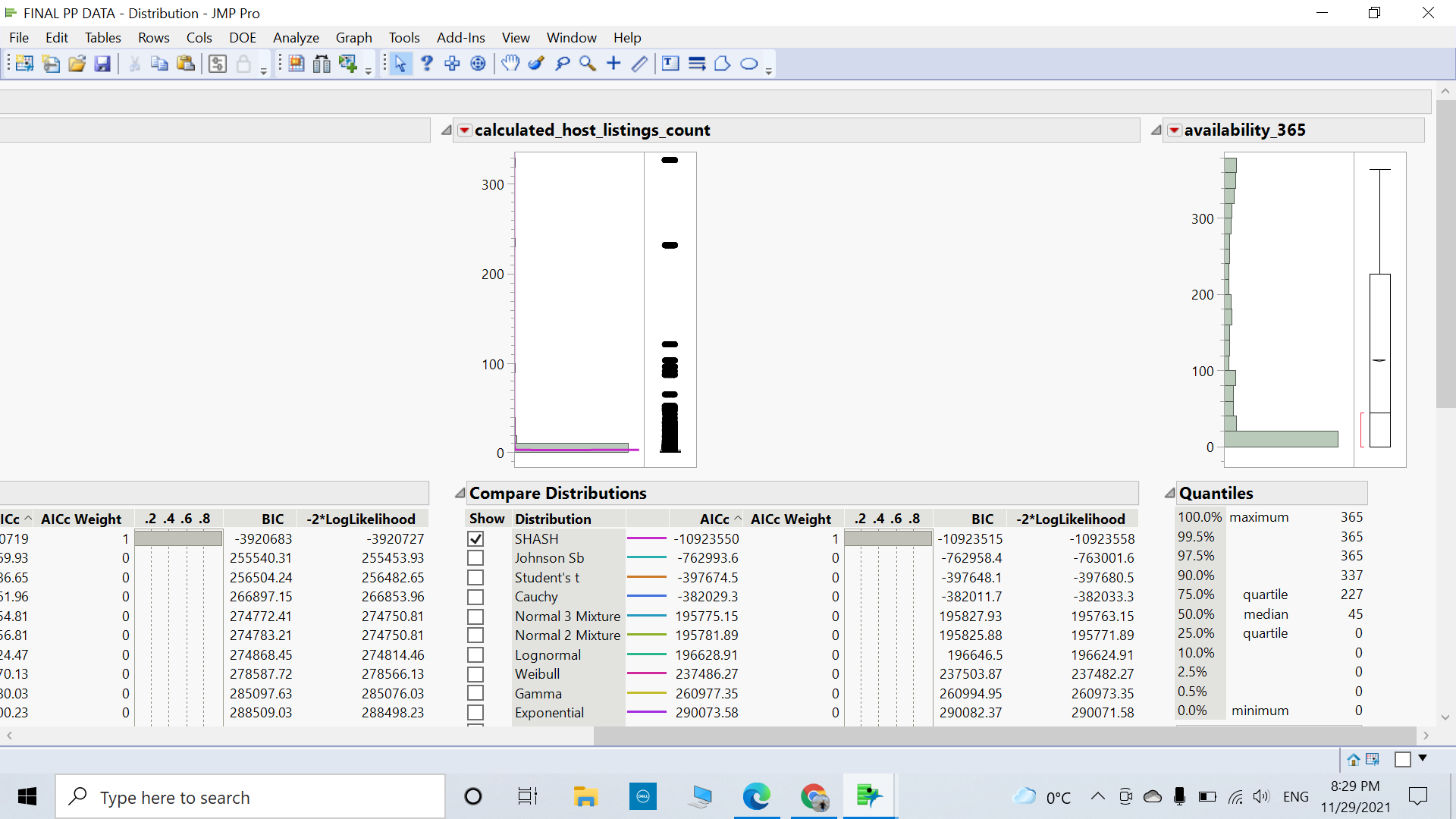
We took all the four continuous variables which are price, minimum\_nights, calculated\_hosts\_listings\_count , availability\_365 and did the outlier analysis.

**Price**

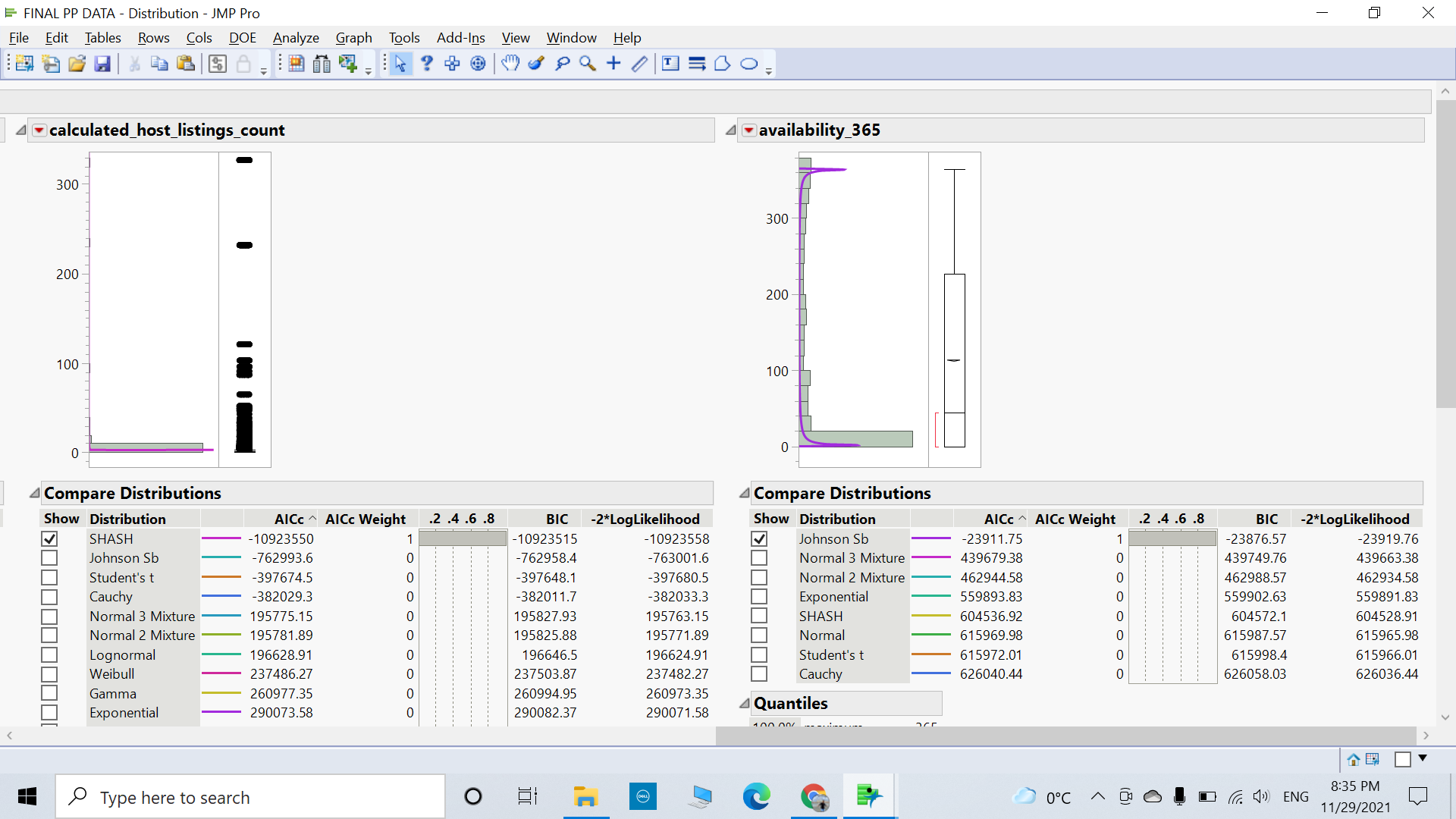
**Minimum\_nights**



**Calculated\_hosts\_listings\_count**

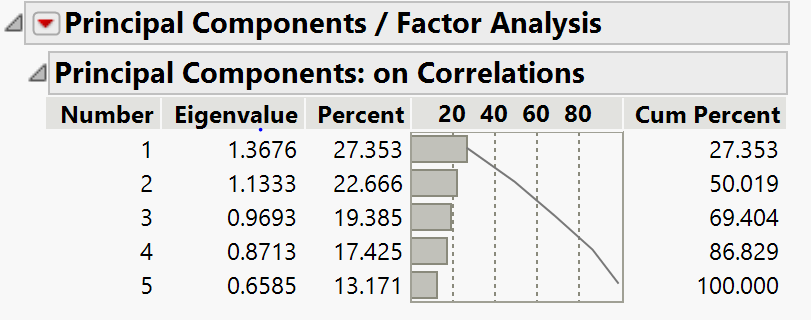


**availability\_365**



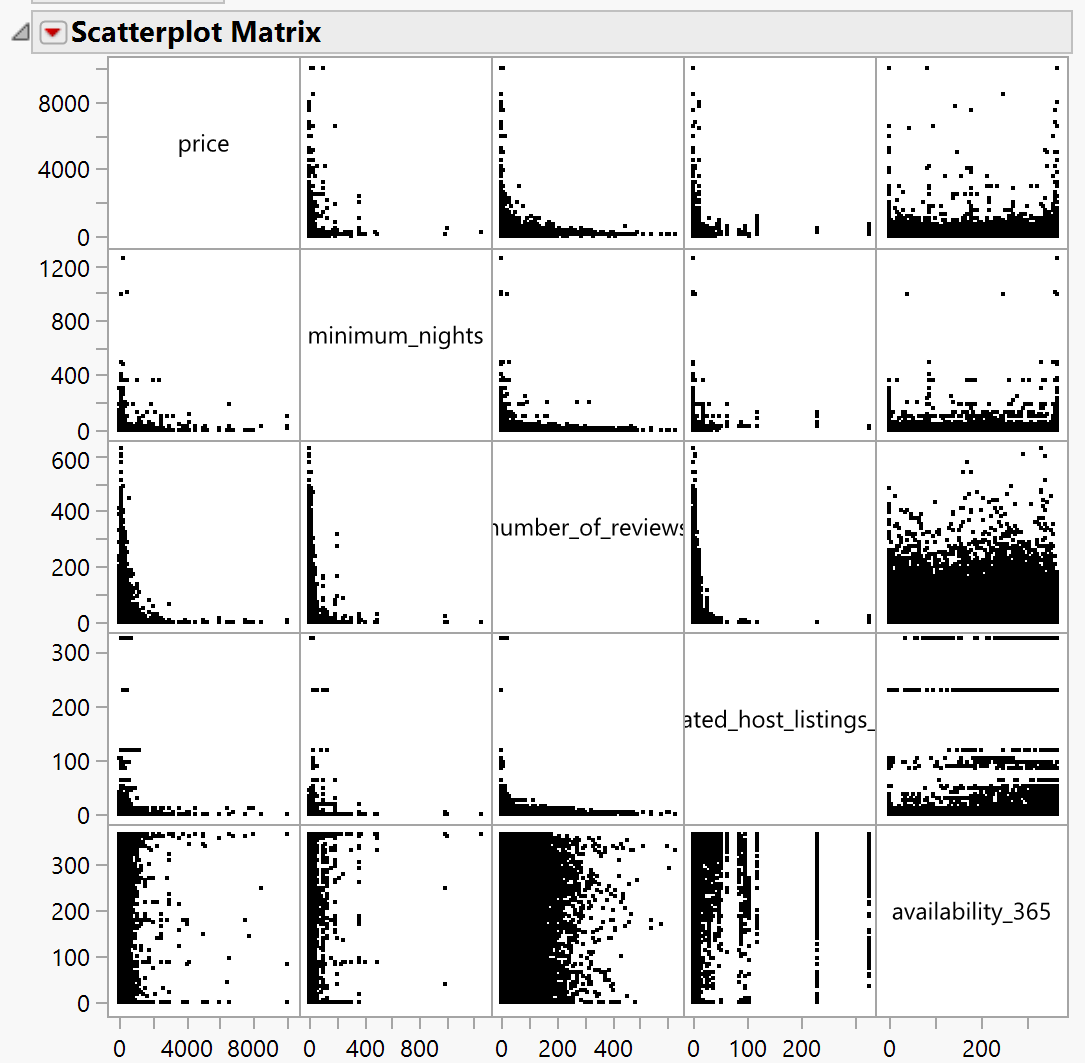
**Principal Component Analysis:**

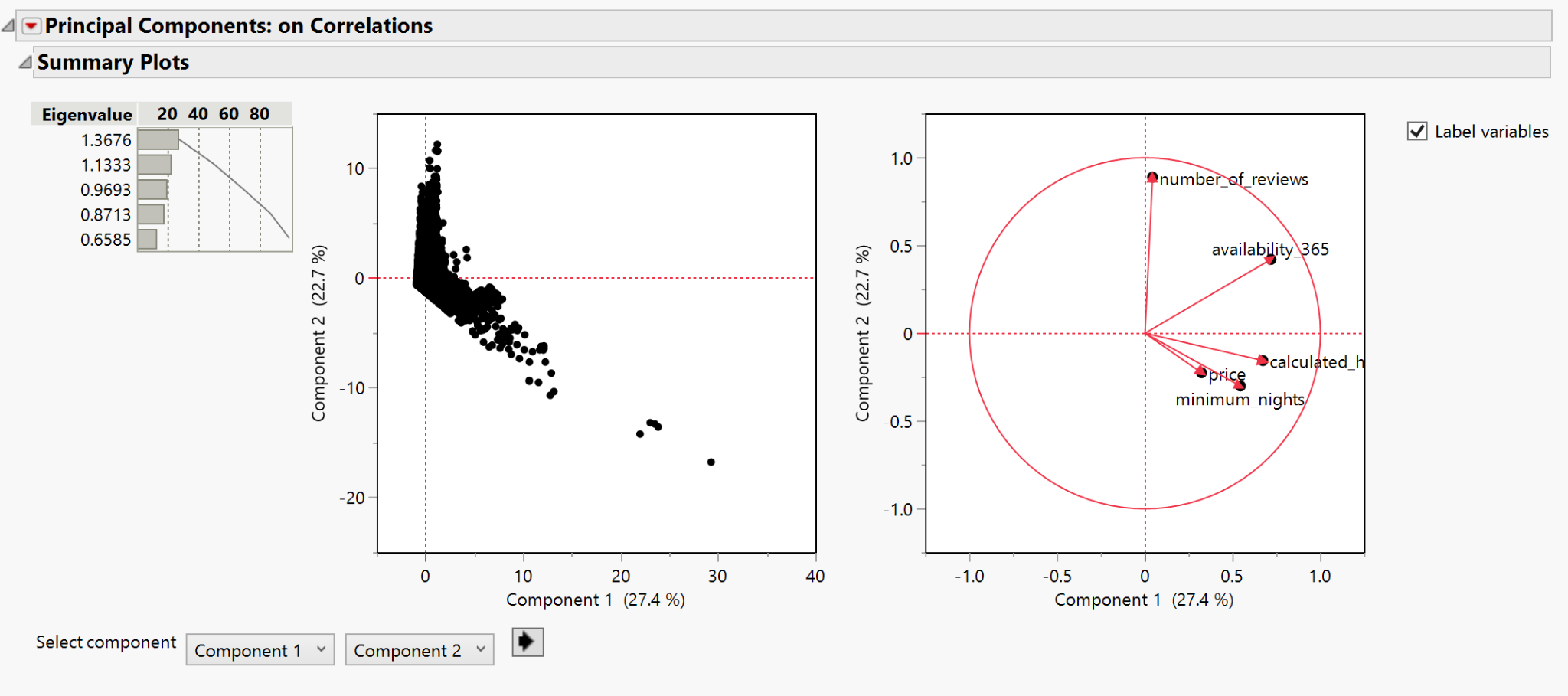
After doing all the above processes, we did principal component analysis(PCA) between the pair of variables to find correlations and decrease complexity.



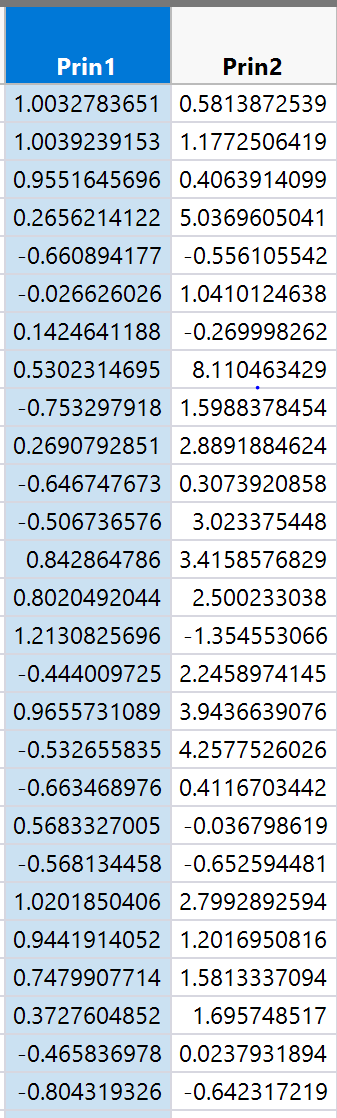
Here, the first three components show over 65% of the total variance.

**Correlations between continuous variables:**

****



After saving the principal component analysis by specifying 2 components to be created we have the following columns



Formula:



**Models**

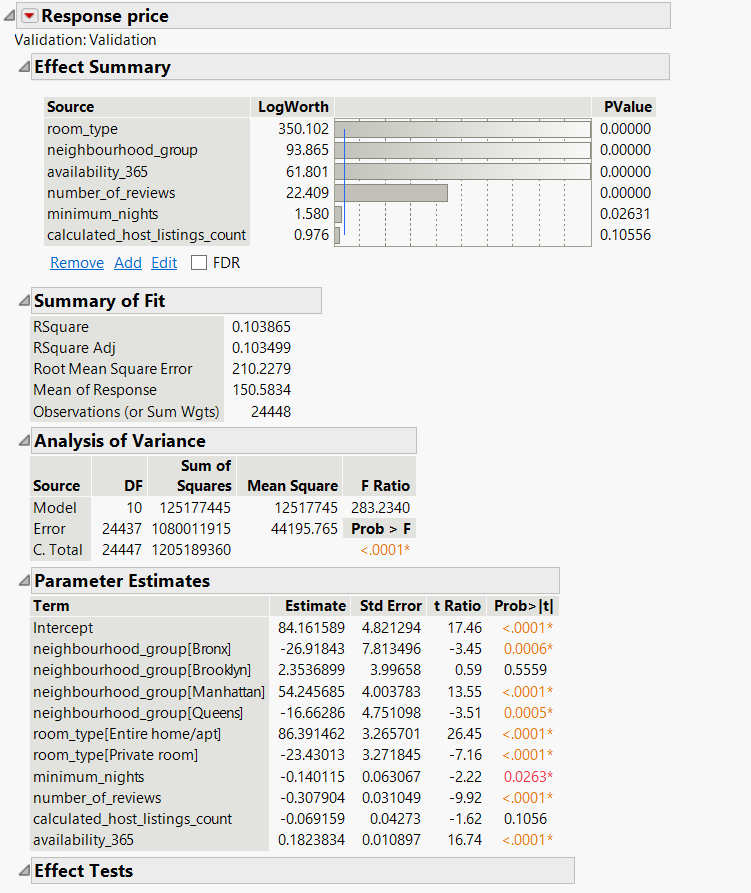
After we successfully completed the data modification, with those two steps we completed the pre-processing section. Now we start the “Model” part of the data, which is responsible for giving us all possible metrics of the data. We use this section to run and try out various models on our data based on the target variable that we have. We also use this section to see how each model is running against a specified validation set/column that we manually generate.   
  
For our data set, we generated a validation column of the following dimensions:

* Training: 50%
* Validation: 30%
* Testing: 20%

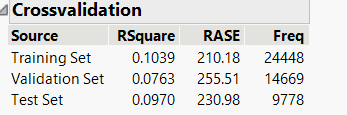
We ran multiple models on our data which included K-Means, Classification Trees, Boosted Trees, Bootstrap Trees, Neural Networks and Regressions.

REGRESSION

For running this, we use the target variable as Price and we add all the other columns to the variables list, such as Availability\_365, number\_of\_reviews, calculated\_host\_listings\_count, Minimum\_nights, etc. We also use the Validation column while using the dialog box for regressions. The regression method was Fit Least Squares. After running the regression, we had the following results:

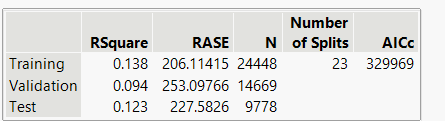


We also can observe the following statistics for each set specifically. We have attached a screenshot of the RASE, RSquare values as they are the values that give the highest meaning to data.

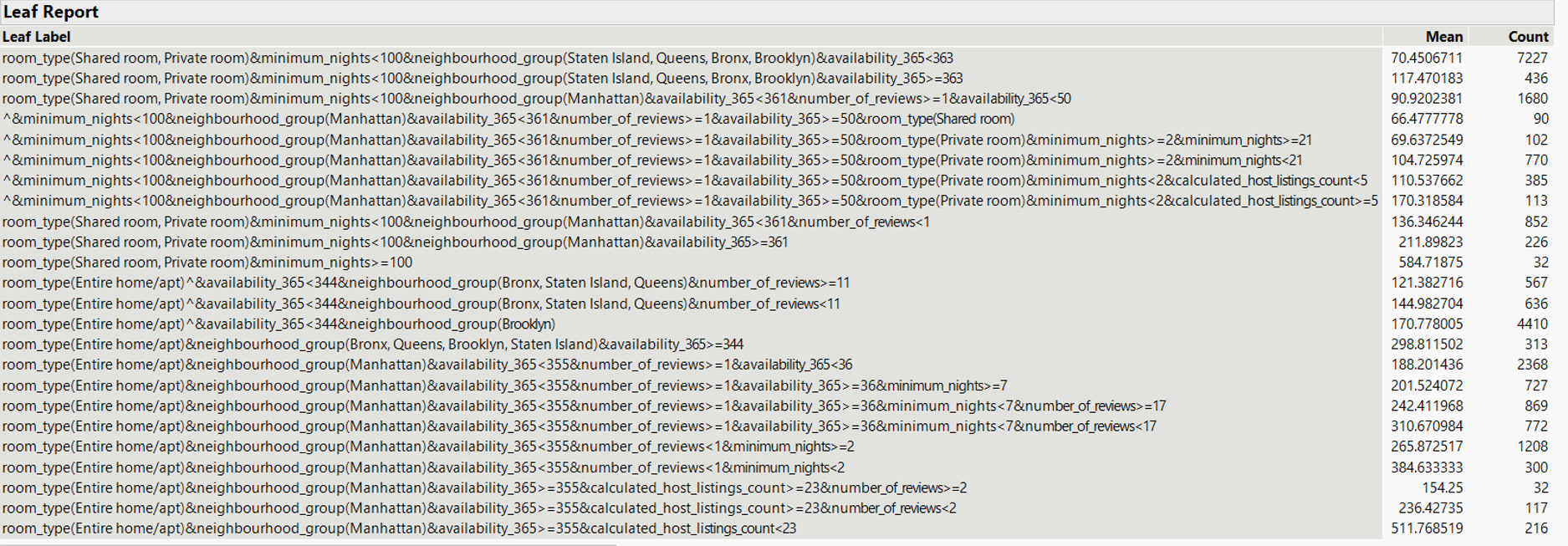


CLASSIFICATION TREE

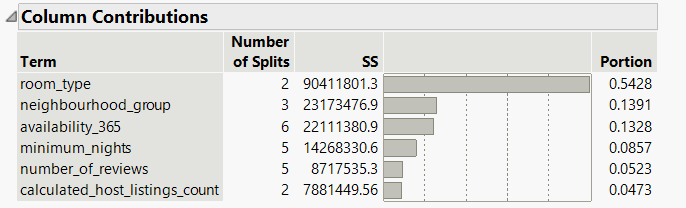
We performed the classification/partition tree model on the data. We used the same columns as above for this model. Our target variable was Price and our relatable variables included: Availability\_365, number\_of\_reviews, calculated\_host\_listings\_count, Minimum\_nights, room\_type and neighbourhood\_group. After inputting the validation column as well, we used the automated “go” button to let the computer estimate the ideal number of splits. After doing the same, we received the following results:

****

I am attaching a screenshot of the leaf report so you can better understand how the splits were created as we are unable to take a picture of the entire classification tree to fit it. The picture might be blurred, but please zoom in and you should be able to see all the details.



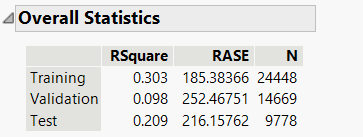
Obviously the most important insight we could gain from running this model is the column contributions as we could first-hand see which variable was influencing the price factor the most and we learned the following details about it:



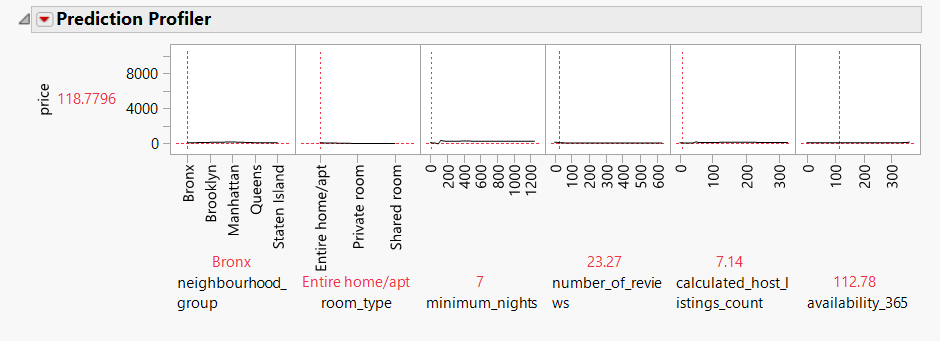
BOOSTED TREE

After running the partition tree, we next ran a model of the boosted tree to see how this model would fit and handle the data and provide results of the same. We used the same variables as mentioned above with price as our target variable and all the above mentioned columns in the other sections, using the validation column as well.

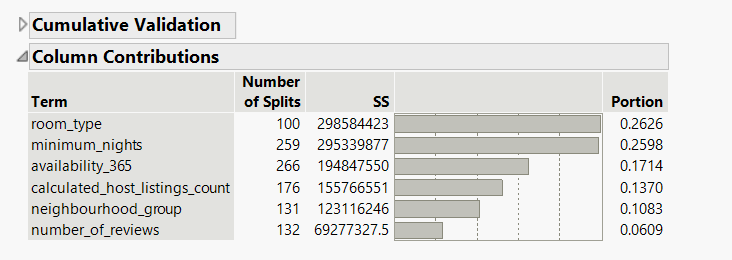
Below is a screenshot of the model that we could observe:



We also were curious to see and try out the prediction profiler so we checked that option to drag across a few values and figure out how every other variable was affecting the target variable as well.

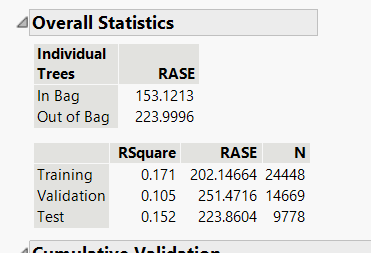


And finally, we used the column contributors option to see which variables were influencing the target variable for this particular model.

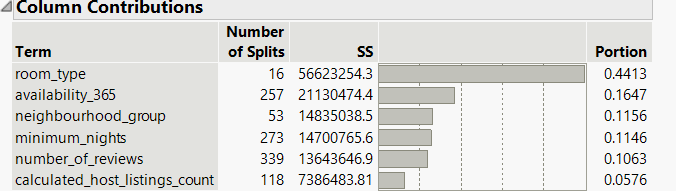


BOOTSTRAP FOREST

Our next model was to run a bootstrap forest and we used the same variables as above. I am attaching screenshots of the output that we received when we tried to run the following model for our data.

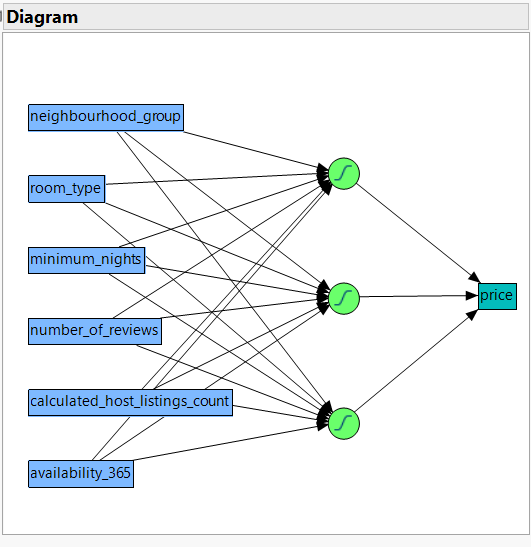


After this, we ran the column contributors check for this model too and this is what we observed:



NEURAL NETWORKS

Finally, we were left with just running the neural networks. We had the same input variables as mentioned above. Our target variable was price, however this time, we only gave in one layer of divisions as we did not want to overcomplicate the data with two divisions. So we input 3 in the TanH and 3 in Linear. We did not input anything in Gaussian. After running this, we received the below network to summarize our model for the data.

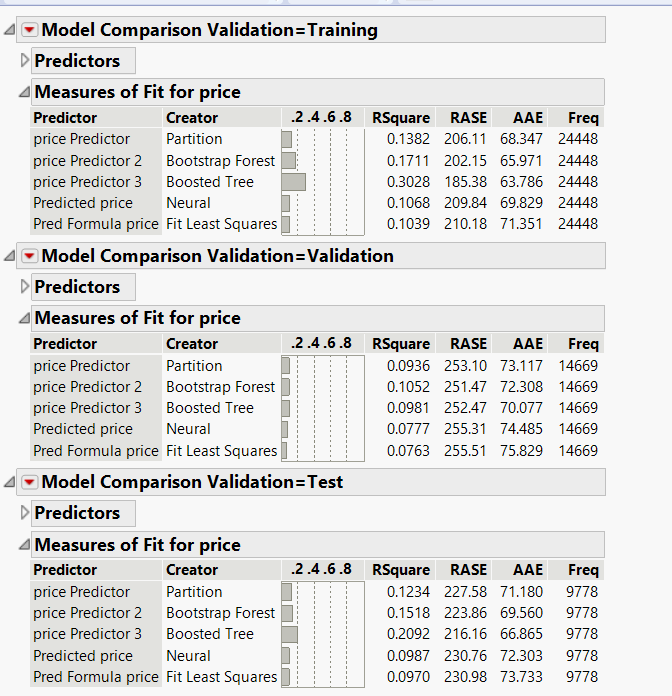


We also ran the check of the prediction profiler for the neural network model as well. You will find a screenshot referencing that below:



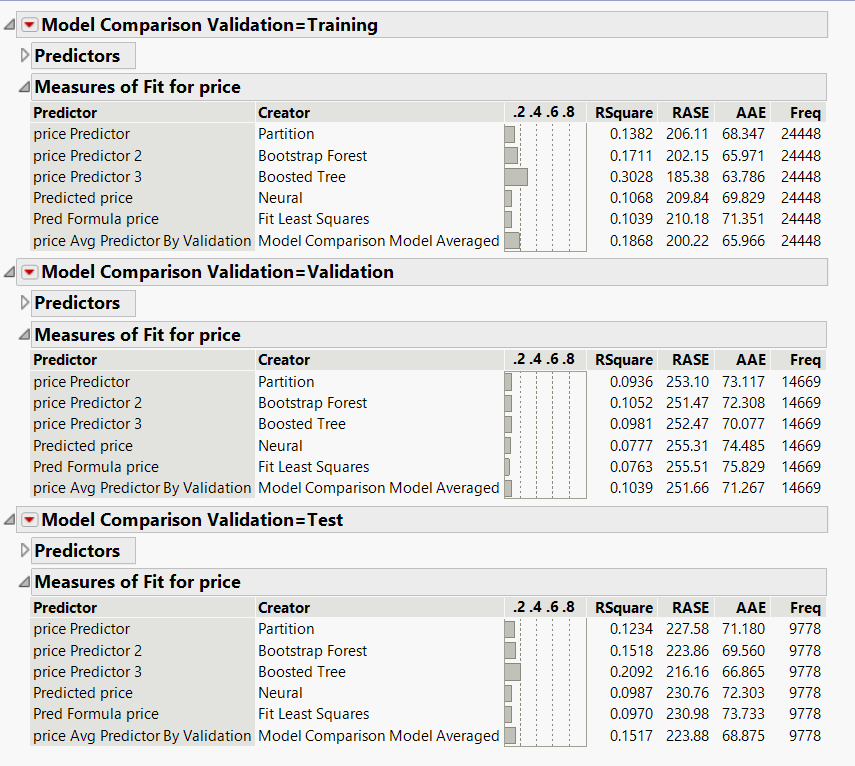
MODEL COMPARISON

After running all the above mentioned models exhaustively, and checking various aspects of it such as the leaf reports, profilers, their column contributions for that model, and looking at the overall performance of the model, our final step in the Model section of the data was to run a model comparison between all the prediction formulas. We saved all the prediction formulas for every model that we ran whenever we ran them. This allowed us to conveniently run a model comparison between all of the models and figure out which one was performing the best.   
  
You will find the screenshot of the output below:



For when the target variable is numerical, we have to prefer a model with RSquare as high as possible and RASE as low as possible. Since our target variable is price, we can look at the screenshot above and conclude that the model generated by running Bootstrap Forest is the best possible model with RSquare value of 0.1052 and RASE value of 251.47. From this we can see that, we are only looking at the value which is there in the validation column. This is because, though we might have RSquare values and RASE values for a better model, we still do not consider this because the validation set is our most important set. We do not consider the training and testing set to determine the best fitting model. We will look into this in more detail in the next section.

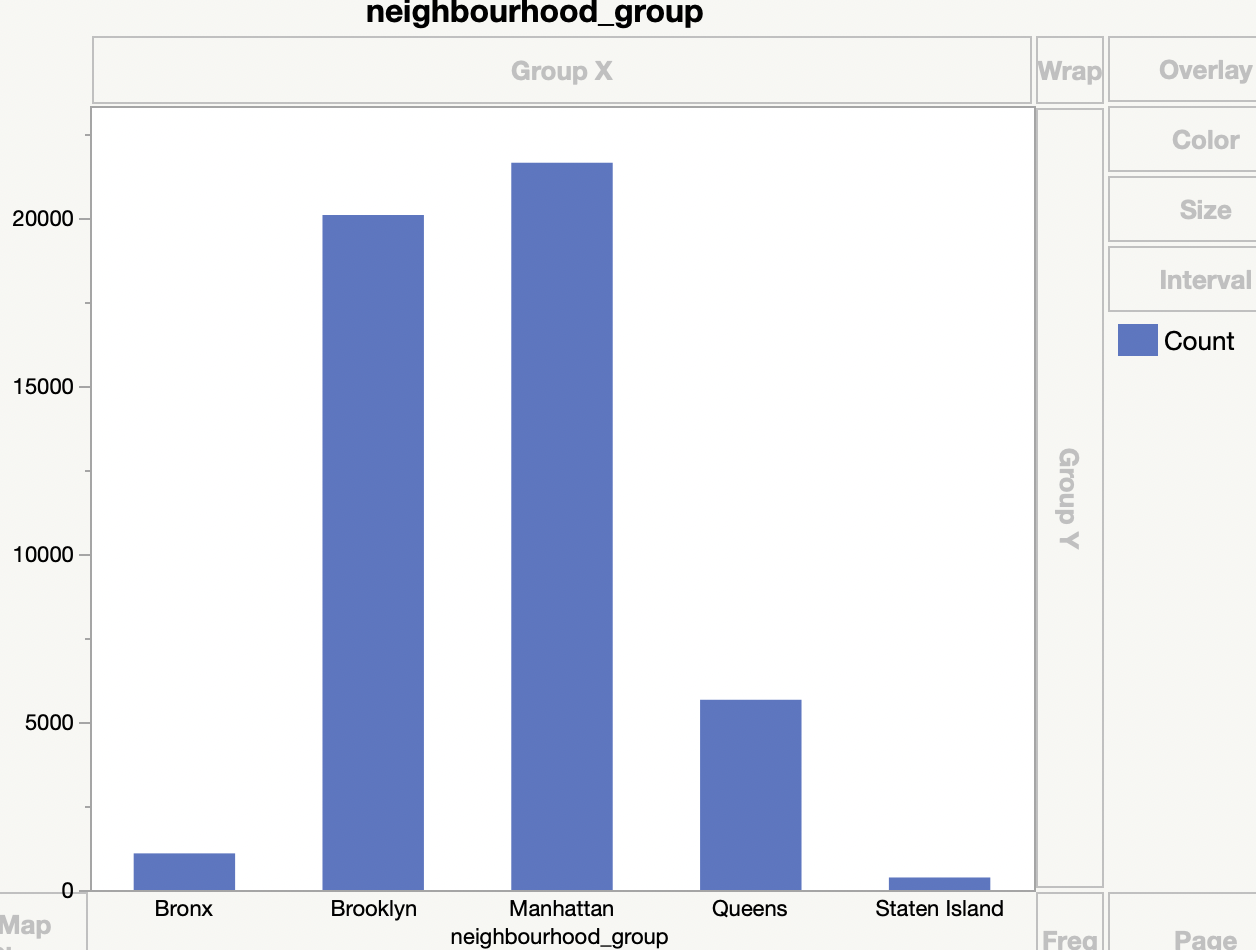
AFTER MODEL AVERAGING, THIS IS THE DATA WE RECEIVE



**Model Assessment:**

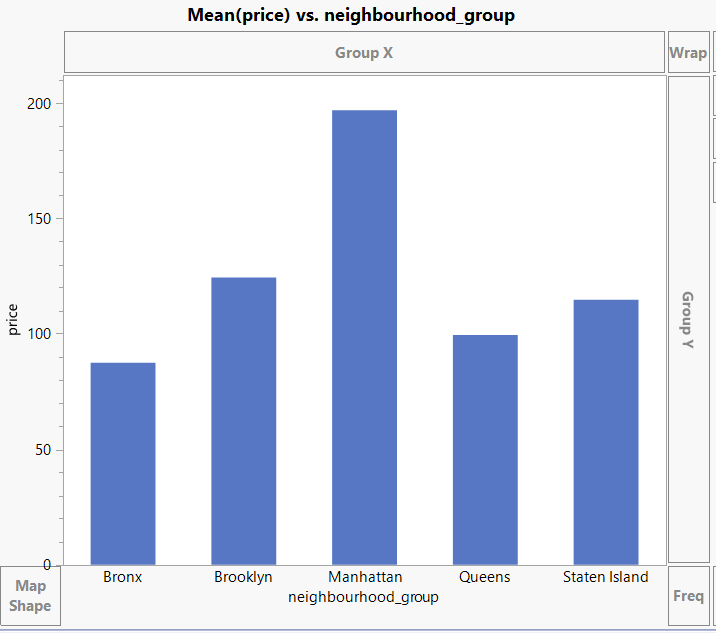
Looking at the models and the comparison, for the continuous target variable: Price, the best model is Bootstrap Forest with the lowest RASE of 251.47 and highest Rsquare value of 0.1052. We took the average of all the models and established that Bootstrap Forest was still the best model as the RASE for the average model is 251.66 that is slightly more than the lowest out of all. Whereas, the Regression model performed the worst with the lowest Rsquare 0.0763 and highest error 255.51 (RASE).

**Insights:**

**The number of hosts in areas:**

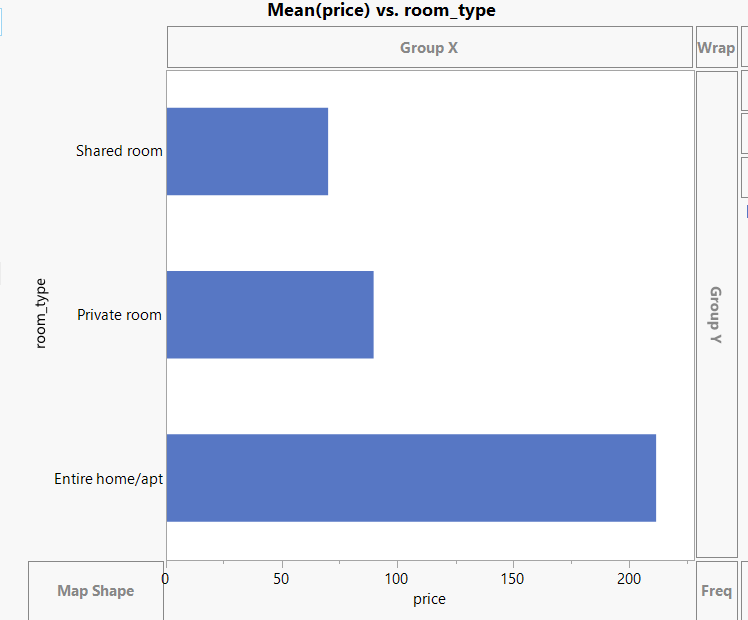
On the chart above we can see that most hosts are from Manhattan and Brooklyn. The significant level of housing supplies indicate strong customer demands in the neighbourhoods.

**Prices in different neighborhoods:**



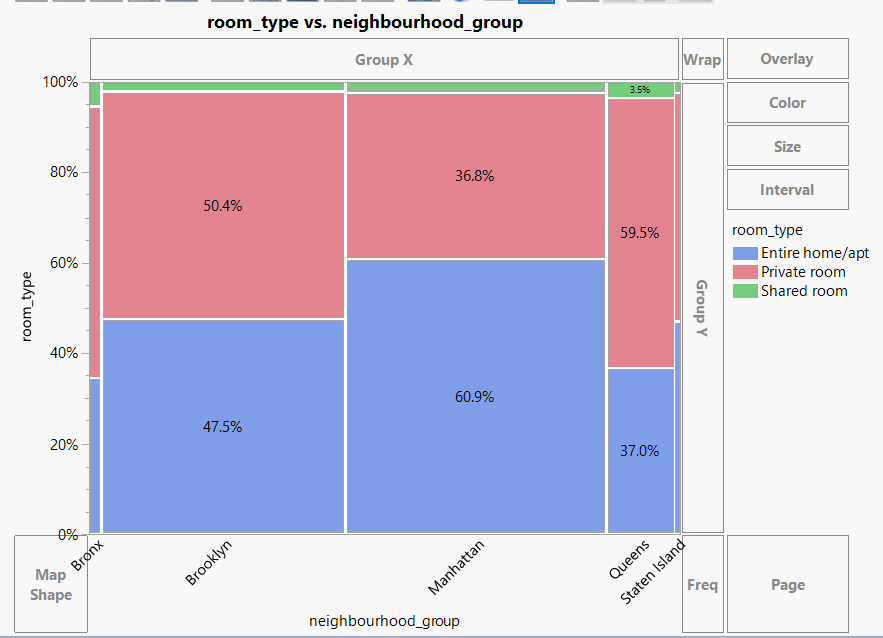
This chart reveals the mean prices for different neighborhoods. Manhattan has the highest average price for housing rents whereas the Bronx on average has the lowest prices.

**Prices of different room type:**



There are three different types of rooms: Entire home/apt, private room, and shared room. On average, the highest price is for the entire home/apt, which is more than $200, and the lowest is shared room, which is around $60.

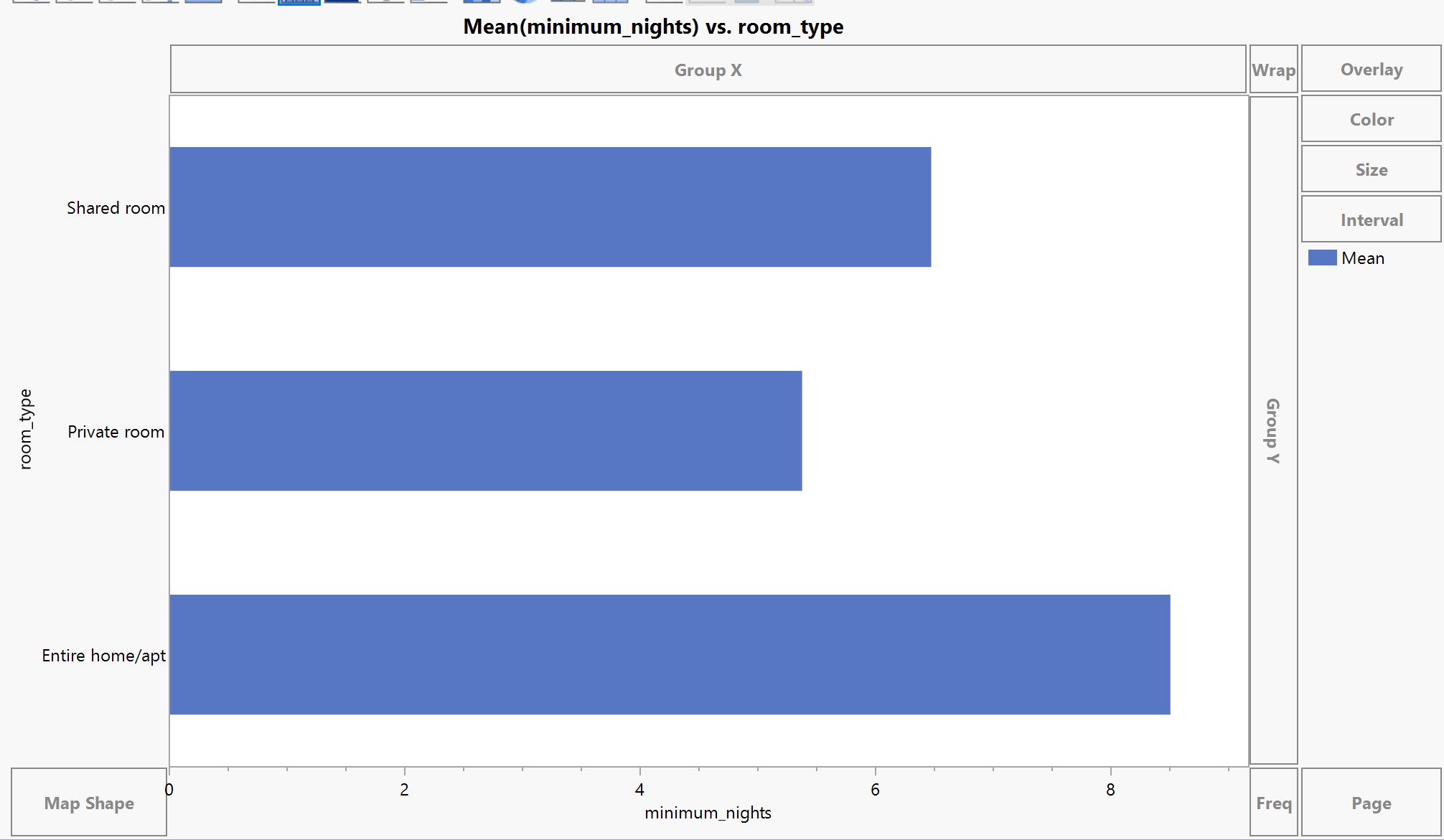
**Room type distribution in different neighborhoods:**



The percentage of Entire home/apt in Manhattan with 60.9%, Brooklyn with 47.5%, Queens has 37%, Staten Island has 47.18%, and the Bronx with 34.4%.

There is less number of shared rooms compared to a private room and Entire home/apt.

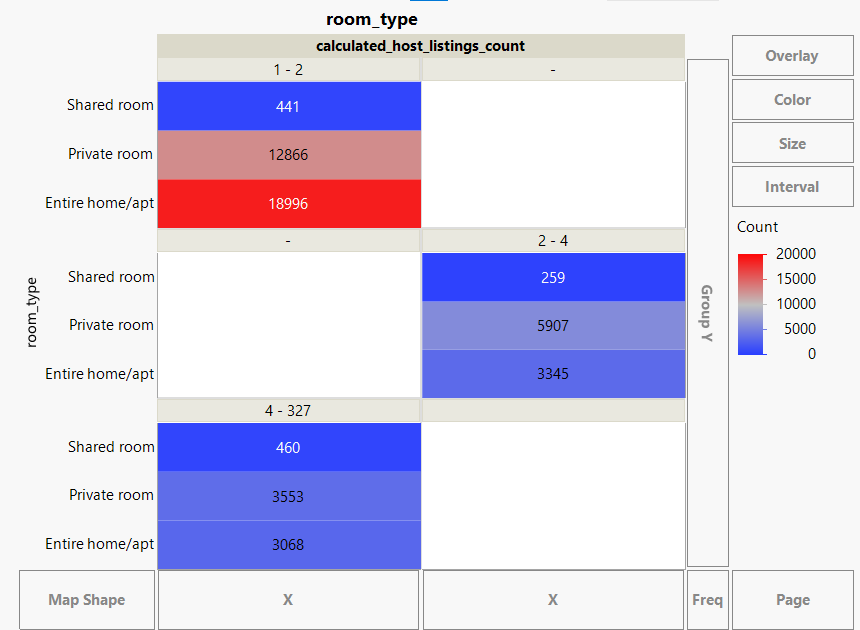
**Average prices of different room types:**



We have calculated the average for minimum nights and room type.

As shown on the above graph, the entire home/apt types on average require 8 minimum nights. Private rooms on average require 5 minimum nights. The graph shows that private room hosts are more flexible than other type hosts.

**Listing counts of three room types:**



For hosts with listings of 1- 2 counts, both the entire home/apt and private room are very popular, counting 18996 and 12866 respectively.

The hosts have 2 - 4 listings, most of the listings are private rooms.

For hosts listing 4 and more, the most common room types are private rooms and entire/apt.

From this graph we can observe that the most commonly offered room types are entire home/apt and private rooms. Majority of hosts are listing 1 - 2 rooms for rent.

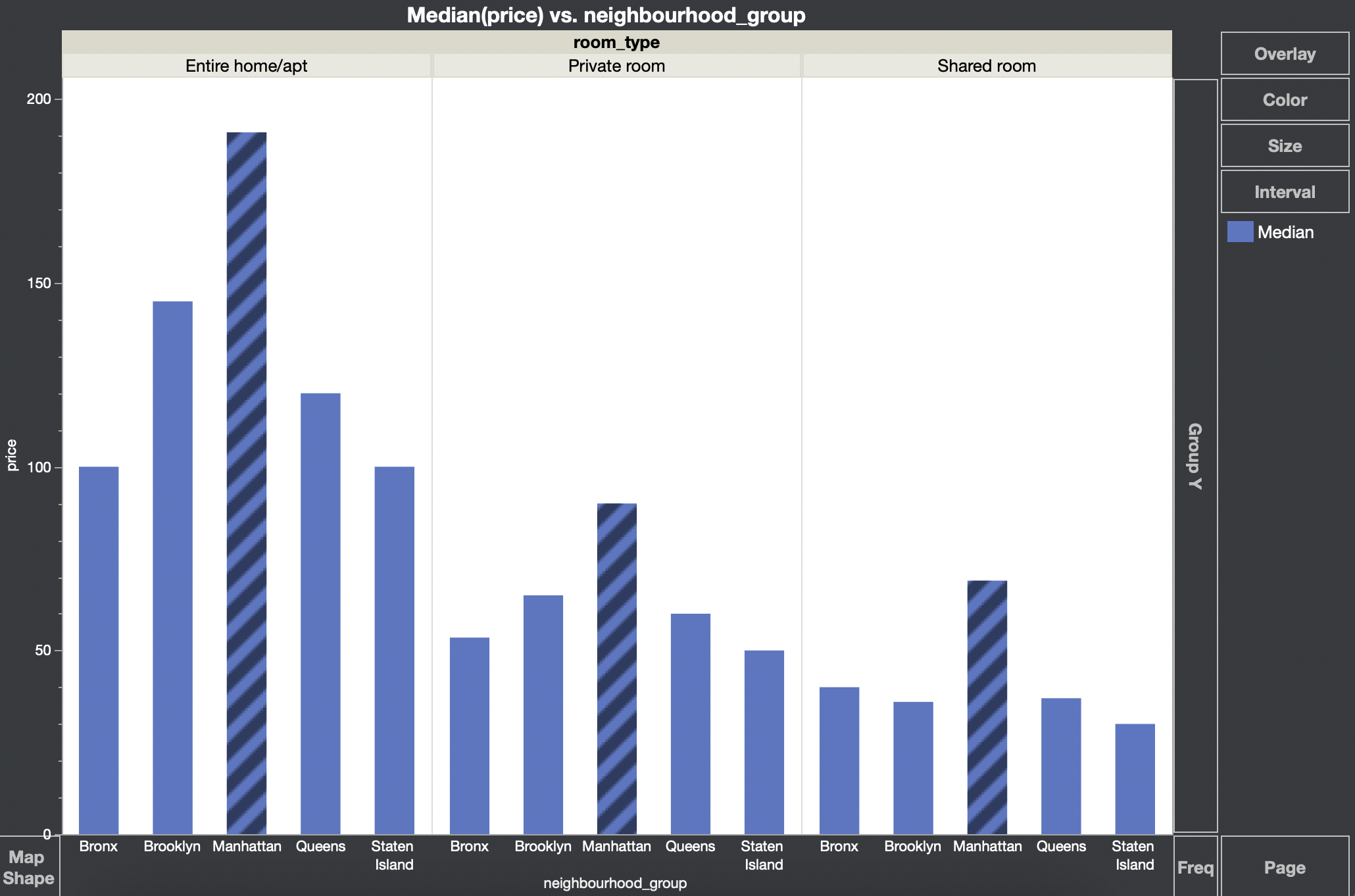
**Room supply distribution by geography with price ranges:**



From the graph above, we can see that private rooms in Brooklyn at $0 - 61 and $61 - 91 price ranges are most common, totalling more than 8000 listings in the area.

In Manhattan, private rooms and entire home/apt are very popular at $91 - 131 range. The most popular room type at this price range is entire home/apt in Brooklyn.

For price ranges of $131 - 201 and $201 above, Manhattan offers most entire home/apt listings.

**Prices’ relationship with room type and neighbourhood group:**

This graph shows on average Manhattan has the highest prices with all room types.

Among all three room types, entire home/apt is the most expensive one.

The graph exhibits neighbourhood group and room type as good predictors of room prices.

**Conclusion:**

Businesses can price the houses according to the trend to maintain the profits and avoid loss when there is no flexibility in the changes..

The factors that most influence the price are room\_type, neighbourhood\_group, and availability\_365 can be modified to improve the profits.

Data Visualization can help identify trends, draw insights ,and make informed business decisions.

**References:**

1. Dataset obtained from the URL below: https://www.kaggle.com/dgomonov/new-york-city-airbnb-open-data
2. Bruce, P., Patel, N., Shmueli, G., and Stephens, M., Data Mining for Business Intelligence: Concepts, Techniques, and Applications with JMP Pro, John Wiley and Sons, Inc, 2017
3. Lecture Notes and PPTs by Prof. Jose Cruz