A FINAL REPORT: DATA MINING PROJECT ON AIRBNB LISTINGS AND METRICS IN NEW YORK CITY, USA

PREPARED BY:

GROUP 11

Student Name	Student ID
SABNAM SHRESTHA	11446896
YAMINI MANGARAI	11535594
SAI DEEPAK JALARU	11509954
ROHITH KOMARAGIRI	11526545
MAHIMA YALAVARTHI	11522517

Submitted to:

Dr Mahdi Fathi

Data Mining and Machine Learning for Business

DSCI 5240

Fall 2021

11/29/2021

TABLE OF CONTENTS **Page No**

EXECUTIVE SUMMARY

As per the initial project proposal, we have decided to work on the information from Airbnb in New York City. Founded in 2008, as a niche site providing accommodation for certain events, the Airbnb website connects hosts who want to rent out their homes, apartments and other properties to visitors who are considering a place to stay. The simple idea has developed into a billion-dollar business that processes a vast quantity of data. Firstly, we have obtained the data from Kaggle which consists of 48895 unique rows and 16 columns. After our preliminary analysis, we have decided that we must decrease the dataset and only be concerned about data for one year, that is 2019. We are interested in studying the New York City Airbnb data and evaluating how the data generated on this online marketplace can help determine the profitability, prices and improve host listings utilizing the Data Mining techniques from the course DSCI 5240.
For this purpose, we will be utilizing New York City Airbnb dataset acquired from Kaggle to run data mining algorithms such as Linear Regression, Principal Component Analysis, Clustering and Decision Tree analysis to produce further insights into our project objectives.

The dataset focuses on listing activity and metrics for Airbnb in New York City for the year 2019. Data has 25209 Rows and 16 Columns containing variables such as Host ID, Neighborhood Group, Price, Minimum Number of Nights, Number of Reviews etc., which are further demonstrated with their type and description in a table shown below in the appendix. This data file has the data needed to generate insights about host popularity, listings per geographical location, and estimations about prices for a certain property. Our tentative primary target variables are Price, Calculated Host Listings Count and Reviews per Month. However, for this status report, we will be focusing on the most important numeric target variable i.e., Price.

PROJECT MOTIVATION/BACKGROUND

Among myriad of data present on the internet, we have selected to create our data mining project with New York City Airbnb data. Our primary motivation for completing this project comes from learning interesting data mining techniques in the DSCI 5240 Data Mining and Machine learning for Business course. During our class, we were exposed to several powerful data mining tools that apply to almost any type of data. We are especially thankful to Dr Mahdi for encouraging us to apply the concepts learnt in the class through different software and analytical tools like SAS, Python and R. This project is an attempt to incorporate the hands-on application into our learning in this course.

The Airbnb Listings and Metrics in New York data allows us to explore the variation in price for properties listed in New York City Airbnb. The main reason behind choosing this dataset is its large volume which can be used for this kind of project, and it has a lot of information that has been presented in a detailed format, and which makes it easier to run several regression models. Professor Dr Mahdi has also encouraged us to apply our theoretical knowledge in real-time applications to get hands-on experience and the data on which we are working with i.e., Airbnb Listings and Metrics in New York in the year 2019. It helps us to various create regression models with ease as our major aim of this project is to focus on Price as our primary target variable and we found that except Name and ID all other variables are important for predicting our target variable ‘Price’. So, we are trying to apply different types of techniques such as Linear Regression, Principal Component Analysis, Clustering, Regression Models using Principal Components and Decision trees.

Airbnb is a thriving business in a location like New York City. Like all leading businesses that do not own private assets like Uber and Amazon, Airbnb relies on its hosts to provide rental space for its patrons. However, price is volatile and tends to change as per neighbourhood, seasonality, availability, and this can impact host listings and the number of people who would like to work with Airbnb as hosts. Thus, we want to ensure that the hosts receive the optimum price and know which factors allow them to set the best price for their property. Likewise, we can help visitors to compare prices more efficiently without any issues and hence, as we are making it easier for the customers and hosts to be confident about the property’s price and thus helping the business to grow.

DATA DESCRIPTION (SUMMARY STATISTICS)

1. Secondary Data:

Our data New York City Airbnb has been obtained from Kaggle, a site that provides its users with open data sets and metrics for various uses like publishing, exploring, building models, and entering competitions.

Initially, the data contained 48895 unique rows and 16 columns. For simplicity purposes, we have sorted and selected the data that had been reviewed within the year 2019. It allowed us to omit rows with missing data in multiple columns. The summary statistics for the reduced data with 25209 unique values in rows and 16 columns were obtained in the SAS Enterprise Miner using the StatsExplore node. In the Preliminary data exploration, we eliminated two nominal variables 'Host Name,' which included the first name of the hosts of the property and 'Name,' which included the names of the listed properties.

2. **Data Exploration**:

In the Summary Statistics section, we are interested in describing the attributes of the interval input variables. As per the results from our StatExplore node, specifically, the summary statistics for the target variable "Price" are as follows:

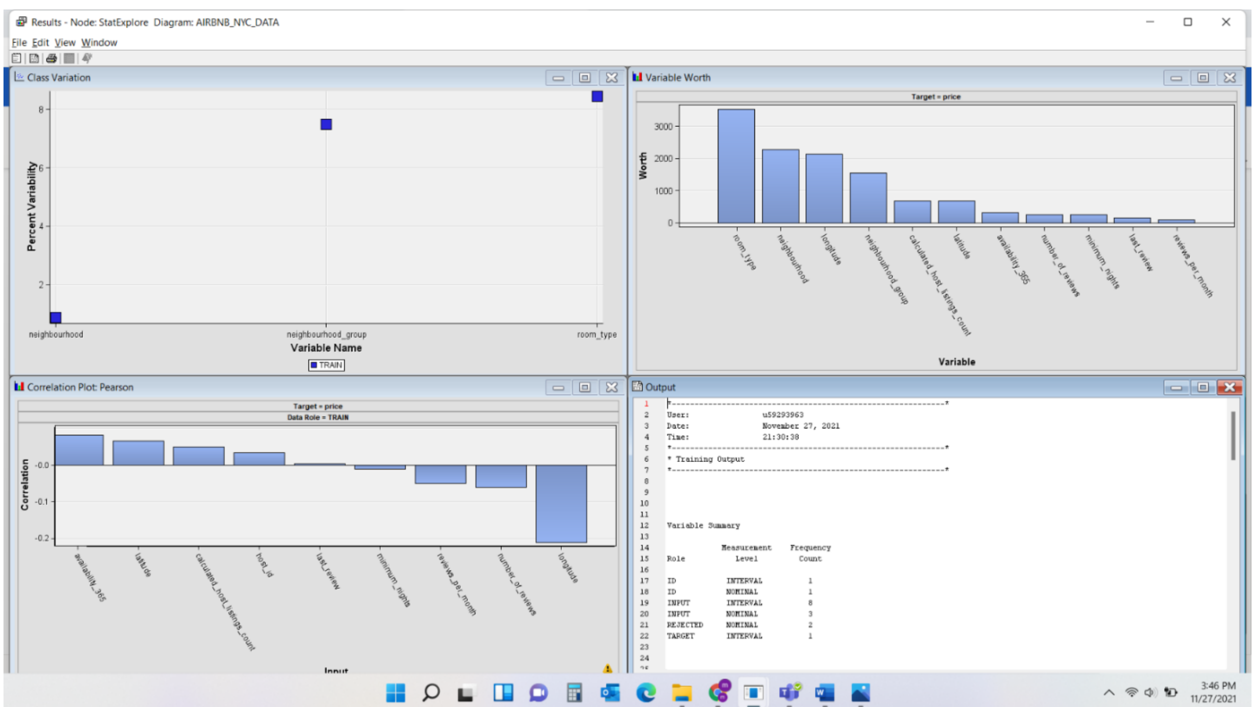


Figure: Results StatExplore Node

The mean price for a property listed in New York Airbnb was $141.2428. The Standard Deviation of the prices listed for 2019 was 147.54.

The total number of missing values was 141 and the total number of non-missing values was 25209. For this reason, we will choose to ignore the missing values in our target variable. Subsequently, the minimum price was found to be 0. This is due to the 141 missing values that do not include the Price information.
The maximum price for a property rented was $7500.

Likewise, with a skewness measure of 12.49286 approximately 12.5, the Prices are highly positively skewed, which is clearly explained by the difference between the mean Price141.2428 and the maximum price at $7500.

As for the independent interval variables, the summary statistics were as observed in the table below:

Variable	Mean	Standard Deviation	Non-Missing Values	Missing Values	Minimum	Maximum	Skewness
Host_id	78432000	84363548	25088	262	2571	2.7364E6	-0.62641
Availability_365	146.2492	127.1315	25088	262	0	363	-1.36023
Calculated_host_listings_count	6.989409	34.57075	25209	141	0	363	72.77469
Last_review	21699.3	46.26353	25088	262	21550	21738	3.213264
Latitude	40.17675	7.924663	25209	141	-74.1625	40.9130	203.3638
Longitude	-73.9481	0.050885	25088	262	-74.2444	-73.713	3.335599
Minimum_nights	5.058511	11.09068	25209	141	1	365	288.9714
Number_of_reviews	40.22824	55.35843	25088	262	1	629	10.89658
Reviews_per_month	1.973492	1.805231	25209	141	0.02	58.5	48.26374

Figure: Table of Summary Statistics for Independent Variables

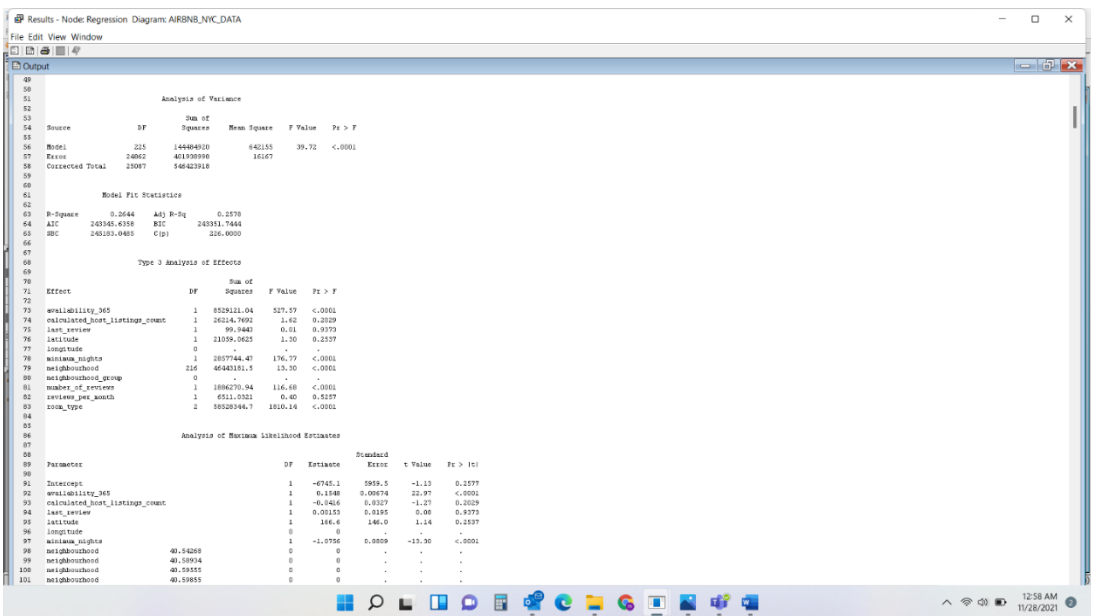
We can notice that the Host_id is included as an input variable as it seems like an interval variable, but it will not be notable in predicting the price changes. In our data mining models, the Host_id variable is likely to be eliminated. Ignoring the Host_id variable from the interval variables, we can notice that 90% of our data is positively skewed. The number of missing values is also repeating itself which reflects that there are multiple rows with missing values, but this is negligible in our analysis since all of them are less than 2% of the total non-missing values. Thus, we choose to ignore these missing values in our data.
Also, to note that in the variable worth measure of StatExplore node the highest worth is given to the variables Room_type, Last_review, Neighborhood, Longitude and Neighbourhood_group.

DATA PREPARATION ACTIVITIES
MODELS/ENTERPRISE MINER & PYTHON DIAGRAMS

The general purpose of our project is to generate meaningful insights from our data. Specifically, we want to condense our input variables to those that can give us more information about the price and how hosts can increase their listings and receive the best price for the listed property. From our data exploration, we have seen that there is a price difference of $7358.7572 between the maximum listed price and the mean price of Airbnb properties in New York City. We can further find out what factors are responsible for the high price and increase host listings accordingly that can maximize the price for their rental properties. We will briefly discuss the models that we are using to analyze our data and further explain the outputs and results of these models generated in the SAS Enterprise Miner Software, the detailed documentation of the output of our model results will be attached in the Appendix section of this report.

a. Linear Regression (from the status report) :

Our first data mining model is linear regression. Although linear regression is the simplest model, it is a powerful tool in assessing the significance of the independent variables in predicting the response/ target variables. We will use linear regression for analyzing the factors reasonable for the high/ low price of the Airbnb properties in the New York City area in the year 2019. For our analysis, we will be taking all the interval variables in our data as our input/ independent variables and setting Price as the target variable.



Results - Node: Regression Diagram: AIRBNB_NYC_DATA

File Edit View Window

Output

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	225	144484920	642155	39.72	<.0001
Error	24862	401938998	16167		
Corrected Total	25087	546423918			

Model Fit Statistics

R-Square	0.2644	Adj R-Sq	0.2578
AIC	243345.6358	BIC	243351.7444
SBC	245183.0485	C(p)	226.0000

Type 3 Analysis of Effects

Effect	DF	Sum of Squares	F Value	Pr > F
availability_365	1	8529121.04	527.57	<.0001
calculated_host_listings_count	1	26214.7692	1.62	0.2029
last_review	1	99.9443	0.01	0.9373
latitude	1	21059.0625	1.30	0.2537
longitude	0	.	.	.
minimum_nights	1	2857744.47	176.77	<.0001
neighbourhood	216	46443181.5	13.30	<.0001
neighbourhood_group	0	.	.	.
number_of_reviews	1	1886270.94	116.68	<.0001
reviews_per_month	1	6511.0321	0.40	0.5257
room_type	2	58528344.7	1810.14	<.0001

Analysis of Maximum Likelihood Estimates

Parameter		DF	Estimate	Standard Error	t Value	Pr > |t|
Intercept		1	-6745.1	5959.5	-1.13	0.2577
availability_365		1	0.1548	0.00674	22.97	<.0001
calculated_host_listings_count		1	-0.0416	0.0327	-1.27	0.2029
last_review		1	0.00153	0.0195	0.08	0.9373
latitude		1	166.6	146.0	1.14	0.2537
longitude		0	0	.	.	.
minimum_nights		1	-1.0756	0.0809	-13.30	<.0001
neighbourhood	40.54268	0	0	.	.	.
neighbourhood	40.58934	0	0	.	.	.
neighbourhood	40.59555	0	0	.	.	.
neighbourhood	40.59855	0	0	.	.	.

Figure: Initial Linear Regression Model

From our initial regression, from our, we obtained an R-square of 0.2644 or 26.44% and an adjusted R-square of 0.2578 or 25.78%, which tells us that almost 27% of the variations in our dependent variables is explained by all the independent variables we have used in our linear regression Model 1.

The input variables that are statistically significant based on the t-value statistics and associated probabilities (Pr >|t|) are the variables that have their p-values of less than 0.05 and hence, are better able to explain the variations in Price. The statistically significant variables are listed below:

Availability_365

Minimum Nights

Neighbourhood

Number of Reviews

Room_type

Initially, we had selected 11 input variables with numeric values to check their impact on our target variable ‘Price’ but after running our initial regression model in SAS Enterprise Miner, we were able to identify 5 statistically significant input variables. This helps us to reduce unwanted data that are not useful in providing any information on the changes in price for our data and focus our analysis further on just the statistically significant variables.

Regression analysis lets us know the strength of the relationship between variables. By using statistical measurements like R-squared or adjusted R-squared, regression analysis can tell us how much of the total variability in the data is justified by our model.

b. Clustering:

Clustering in data mining is an unsupervised technique for classification with no pre-defined classes. Clustering depends on the degree of similarity in data, in which there is a high similarity among objects within the same cluster and dissimilarity among objects in different clusters. It is one of the preprocessing steps applied for predictive model building.

In our New York City Airbnb data, we want to know how many clusters we can form from our data and what each cluster means for our data. Since we have a large data set for the Airbnb price and listings for a single year and the range for the price is extremely high i.e., equal to the maximum price of $7500. We seek to look for clusters that may give us some general idea about how we can group our data. In clustering, a group of like data objects are categorized as similar objects. Data sets are separated into various groups in cluster analysis, which is based on the likeness of the data.

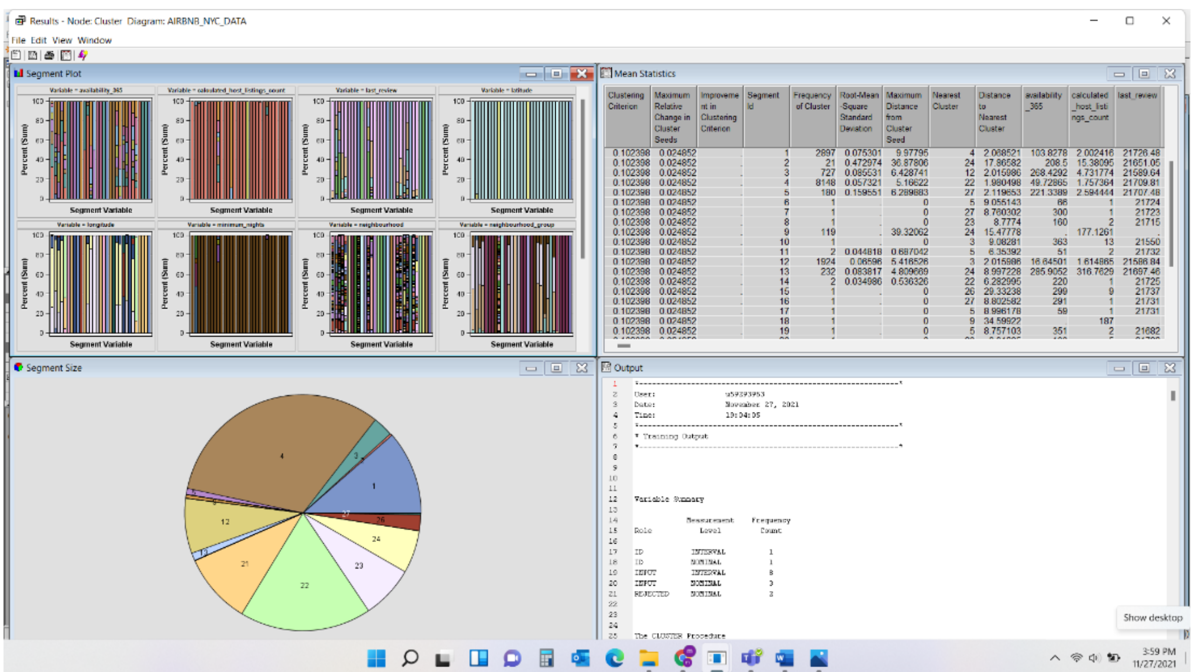


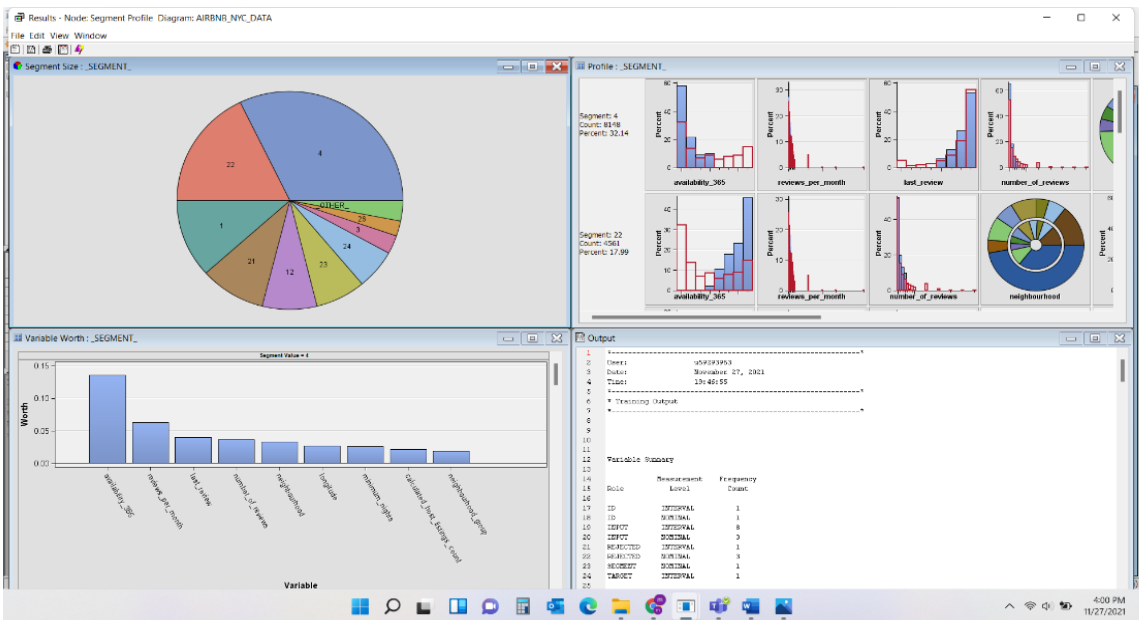
Figure: Cluster Diagram Airbnb NYC

The observation from running a cluster node in SAS Enterprise Miner for our data is discussed further. For clustering we had set as input, 8 interval variables, 3 nominal variables and we had also kept the ID variables in this analysis, rejecting 3 nominal variables such as Hostname and Name that did not add much information for our target variable price.

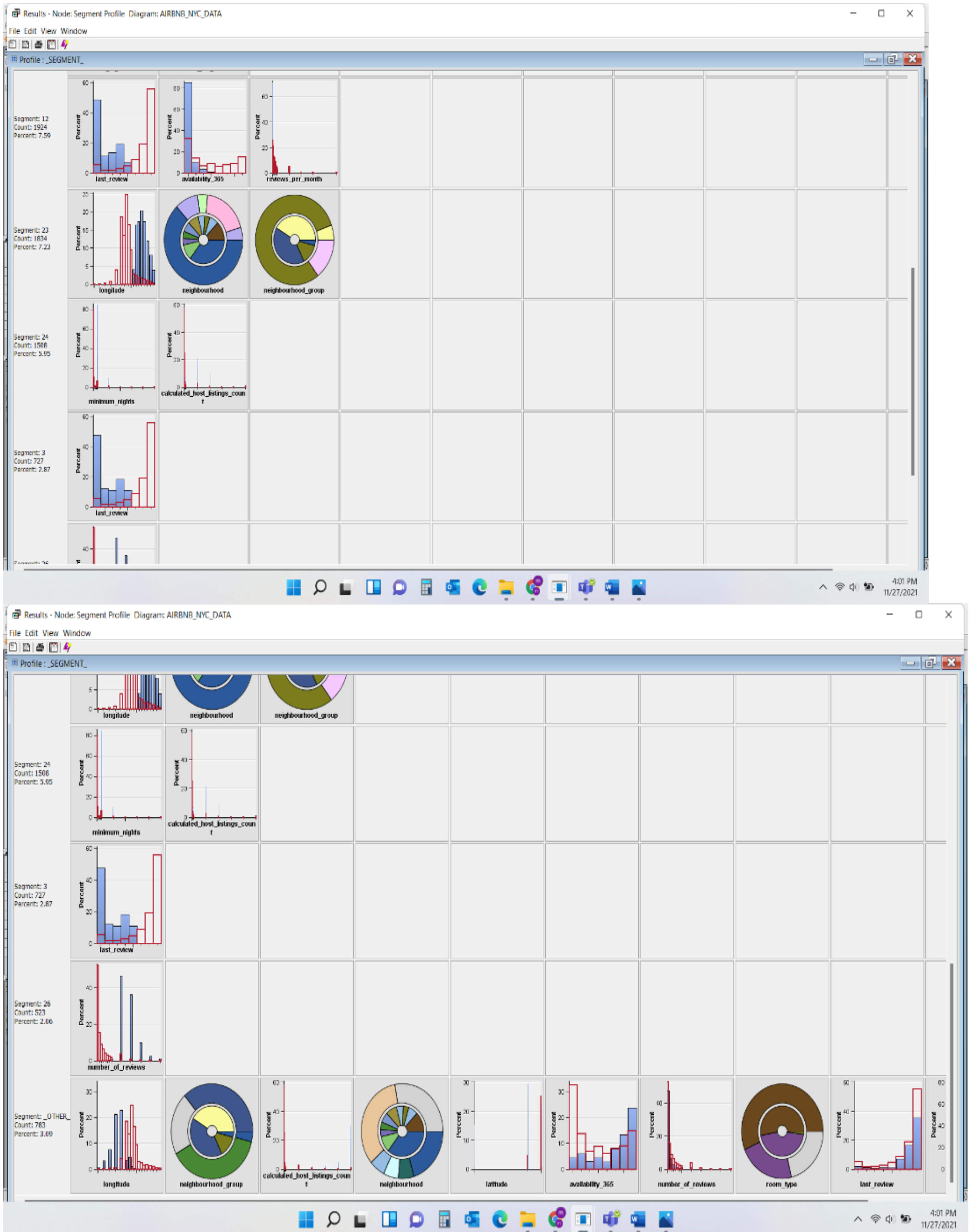
We have also selected the Internal Standardization in the Cluster Node, by standardizing our data we are taking into consideration the data within exceptionally higher ranges as well as the smaller ranges. The underlying concept is clustering the equation of distance among the data points. Thus, if we do not standardize our inputs, the data points within higher ranges will influence the information in our cluster in comparison to the data points within a smaller range and may not provide the actual cluster information due to the impact of exceptionally large distances between these two extremes.

In the results node from our Cluster Diagram, we can generate a segment size window, which shows that we have 13 different segments in our cluster, and Segment 4 represents the highest number of data points for our data. This tells us that a vast majority of the price information for New York City's Airbnb properties fall within this cluster. We also believe that given many of our datasets, 13 clusters help us identify the distinctness of a cluster from another and further improve our classification to know how our data can be segregated in the measure of similarity in how they affect our target variable 'Price'.

We have also added a Segment Profile node to our Cluster node to see how our clusters are formed and how ed and how individual clusters are distributed. The results from running the segment profile node are as follows:



Interestingly, it can be noted that previously in our StatExplore node results, the variable with the highest worth was Room Type, now it has changed to availability_365, this provides us further information about how the price changes as per the time of availability and seasonality. Then, it is worth finding and incorporating more data about property availability and duration in which there is a high demand for Airbnb in New York City to maximize the price for the hosts.



Figures: Segment Profile Diagram for each Cluster 1.2.3.

From our results from the Segment Profile, we can further learn how our data behaves independently versus when the data is clustered and grouped under segments. In each of the segments, there is a significant difference observed in the distribution of the variables after clustering and the segment node is added. Specifically, we can see that in Segment 12, which represents 1924 data points and important variables like availability_365 and last_review, the distribution under the segment node represented by the blue node shows how the data is highly positively skewed as depicted by the blue histograms whereas, in the original data these variables re represented to be negatively skewed as seen in the red histograms.

Finally, after adding a segment node the 13 clusters are further condensed into 10 unique clusters, where the highest data points are still represented by Segment 4 with 8148, and the 3 segments are merged into one, this gives us more clarity about the measure of similarity for our data and thus we can conclude that for the variables taken into account for clustering, the New York City Airbnb data has a high degree of similarity where 25209 unique data points are sufficiently grouped among 10 different clusters. It makes our analysis easier and gives us the impression that our data is valuable in predicting prices for Airbnb properties.

c. Principal Component Analysis:

Principal Component Analysis is a dimensionality reduction technique based on feature extraction, which means that PCA will transform high-dimensional data into lower dimensional sub-space by using a linear transformation. The main use of PCA is to signify a multivariate data table as a smaller set of variables to examine trends, jumps, clusters and outliers. PCA is a technique used to decrease the number of variables in the data by extracting important ones from a large pool while retaining the accuracy of the model.

Since PCA only involves predictors/ independent variables, we will run a second regression by using the resulting principal components that are statistically significant from our Principal Component Analysis. We will make a comparison of our initial regression and the resulting significant predictors with subsequent regression models build using the Principal Components as our predictor variables for our dataset Airbnb Listings and Metrics in New York and measure whether any changes in dimension is identified and how will that reflect in the significance of the principal component predictors, given by individual P-values and their explanation of variation in our target variable given by the R- square and Adjusted R-square values.

For our PCA Analysis, we have set an Eigenvalue cutoff of 0.85 per cent i.e., we are willing to sacrifice 15% of our data to reduce the dimensions of our data and retain only the required data to predict the response variable “Price”.

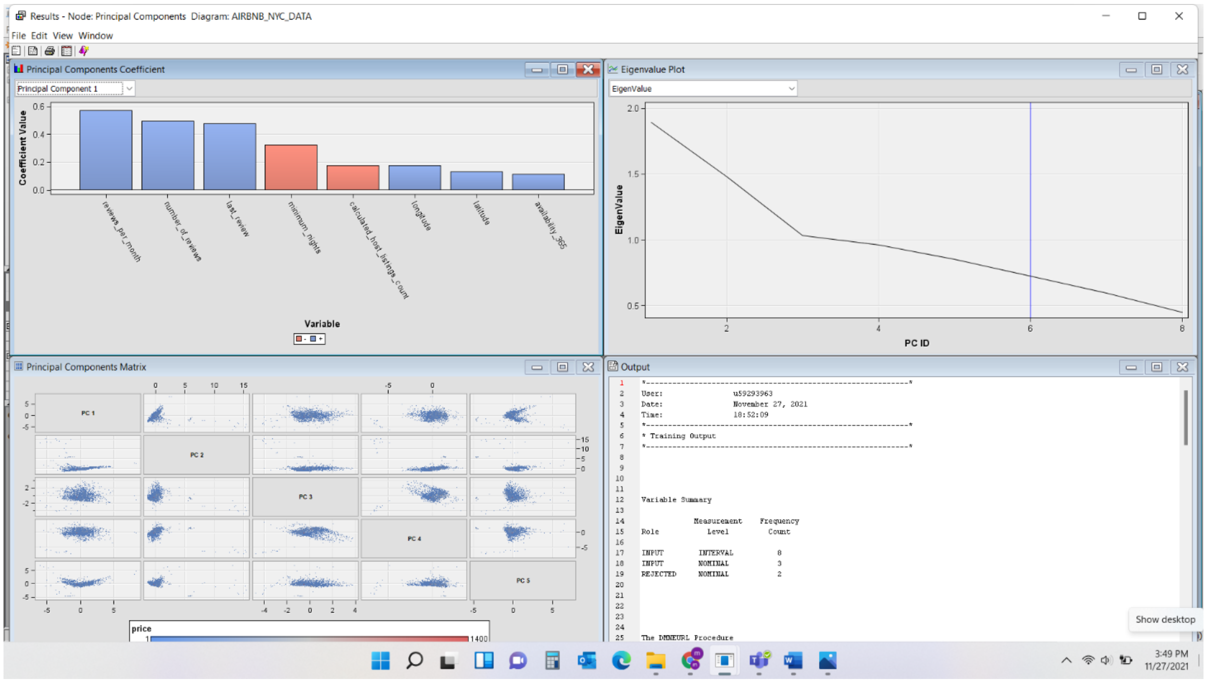


Figure: Principal Components Analysis Result

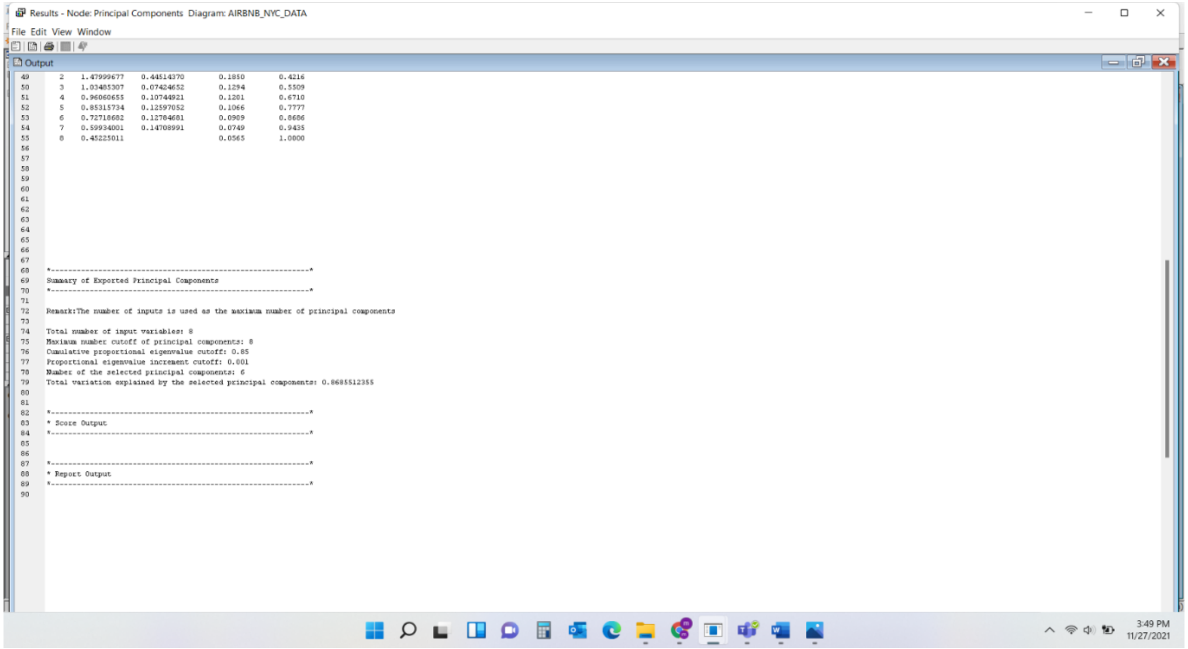
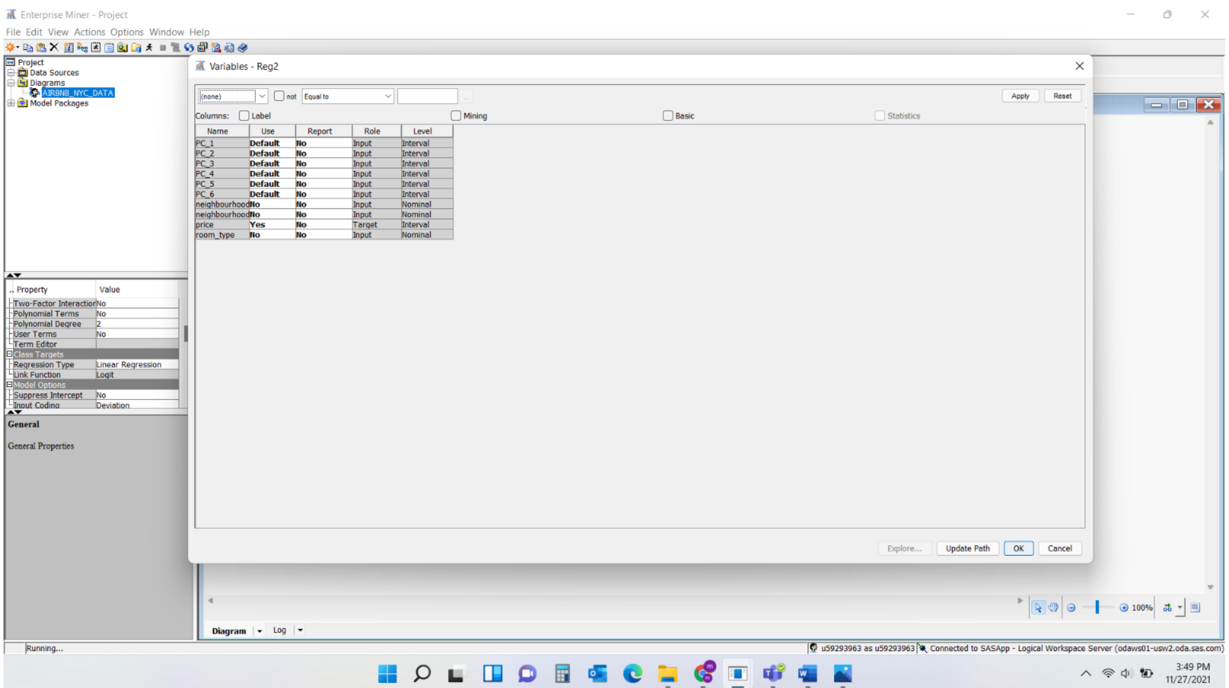


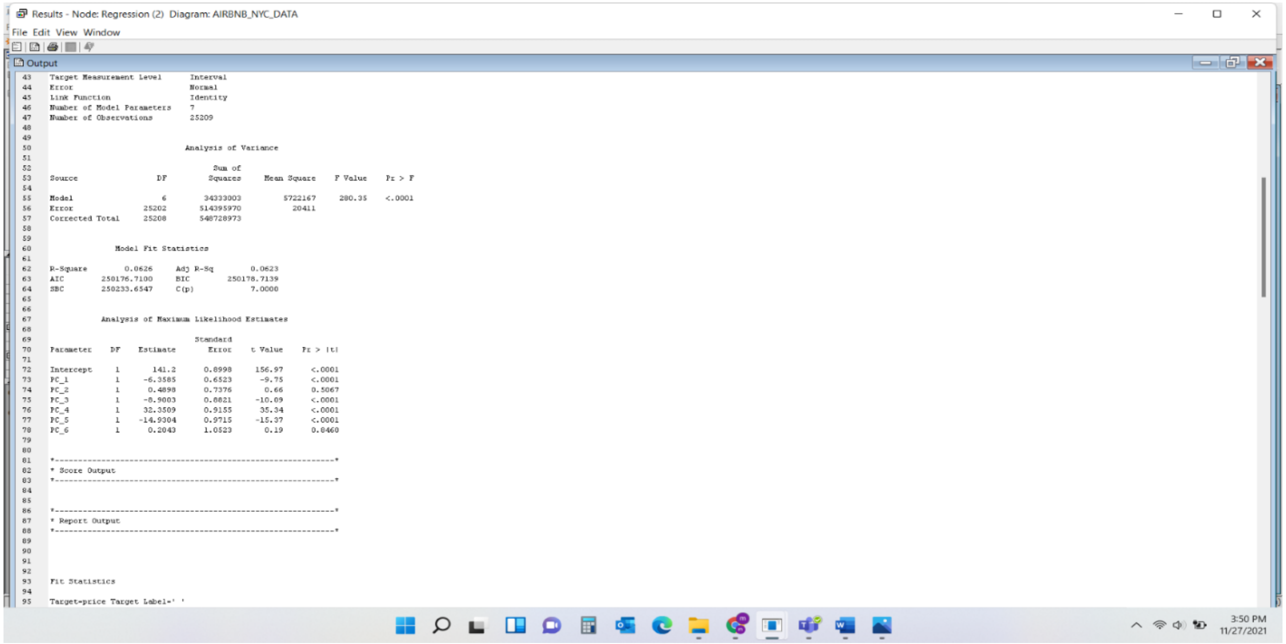
Figure: PCA Node results

Initially, we included a total of 8 interval variables in our initial regression, after applying PCA, we can see in the results that 6 principal components are selected from our initial 8 interval variables. This implies that at an Eigenvalue cut-off of 0.85, we were able to reduce 25% (2/8 = 0.25) of our data by eliminating 15% of our original data. Also, the total variation explained by the selected principal components is given to be 0.8686 or approximately 87% which suggests that the resulting Principal Components help explain the variation in our data.

Figure: Using Principal Components for Second Regression Model



We will further use these six principal components to perform additional linear regression to see whether the models with the Principal Components perform better than the initial regression model.



Results - Node: Regression (2) Diagram: AIRBNB_NYC_DATA

File Edit View Window

Output

```
Target Measurement Level      Interval
Error                         Normal
Link Function                 Identity
Number of Model Parameters    7
Number of Observations        25209


                         Analysis of Variance

                                   Sum of
Source              DF            Squares     Mean Square    F Value    Pr > F

Model                6           34333003         5722167     280.35    <.0001
Error            25202          514395970           20411
Corrected Total  25208          548728973


             Model Fit Statistics

R-Square          0.0626     Adj R-Sq          0.0623
AIC          250176.7100     BIC          250178.7139
SBC          250233.6547     C(p)              7.0000


          Analysis of Maximum Likelihood Estimates

                                    Standard
Parameter     DF     Estimate          Error     t Value     Pr > |t|

Intercept      1        141.2         0.8998      156.97       <.0001
PC_1           1      -6.3585         0.6523       -9.75       <.0001
PC_2           1       0.4898         0.7376        0.66       0.5067
PC_3           1      -8.9003         0.8821      -10.09       <.0001
PC_4           1      32.3509         0.9155       35.34       <.0001
PC_5           1     -14.9304         0.9715      -15.37       <.0001
PC_6           1       0.2043         1.0523        0.19       0.8460


*------------------------------------------------------------*
* Score Output
*------------------------------------------------------------*


*------------------------------------------------------------*
* Report Output
*------------------------------------------------------------*




Fit Statistics

Target=price Target Label=' '
```

Figure: Results from Second Regression Model

For our second regression model, we will use PC_1, PC_2, PC_3, PC_4, PC_5 and PC_6 as our input variable and select 'Price' as our target variable and check for the performance of our second regression model. As we can see from the results of the second regression model using the Principal Components that our model has suffered, specifically the R-square value is 0.0626 and the adjusted R-square value is 0.0623, which means the principal components selected explained only 6.26% of the variation in the data.

Further among the 6 principal components used as our predictors, only 4 principal components are significant as they have their corresponding P-values at <0.0001, however, principal components PC_2 and PC_6 is insignificant with P-values greater than 0.05. It could mean that dimension reduction will not necessarily add value to our data analysis.

To confirm this view, we will re-run a third regression model with just the significant principal components PC_1, PC_3, PC_4 and PC_5 to check whether our regression model will perform better than our initial regression model.

From the results node of our third regression model, we can see that the four principal components PC_1, PC_3, PC_4 and PC_5 can explain only 26.43% of variations in the price and the principal component PC_4 is shown to be statistically insignificant in this model.

Looking at the R-square of 0.2643 and Adjusted R-square value of 0.2542 for the third model and R-square of 0.2644 or 26.44% and an adjusted R-square of 0.2578 or 25.78% for our initial regression model, we conclude that reducing dimension using PCA for our New York Airbnb data is not worthwhile.

Figure: Selecting PC_1, PC_3, PC_4 and PC_5 for Model 3

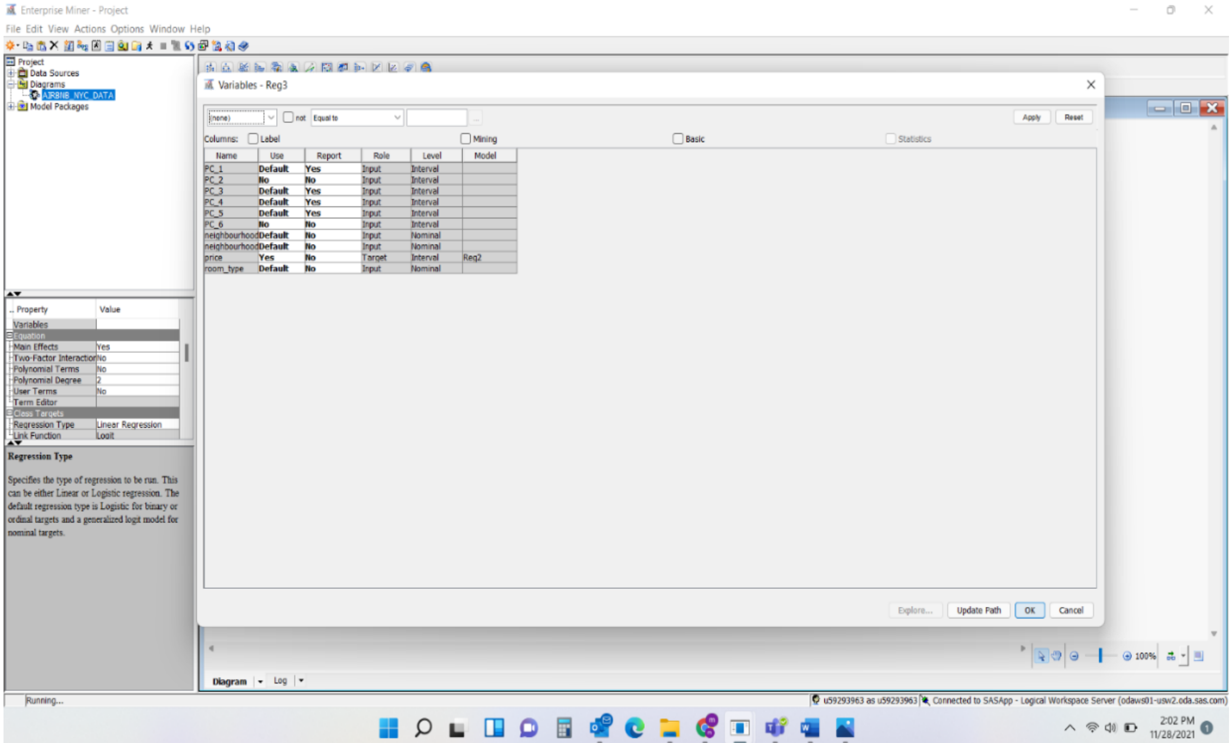
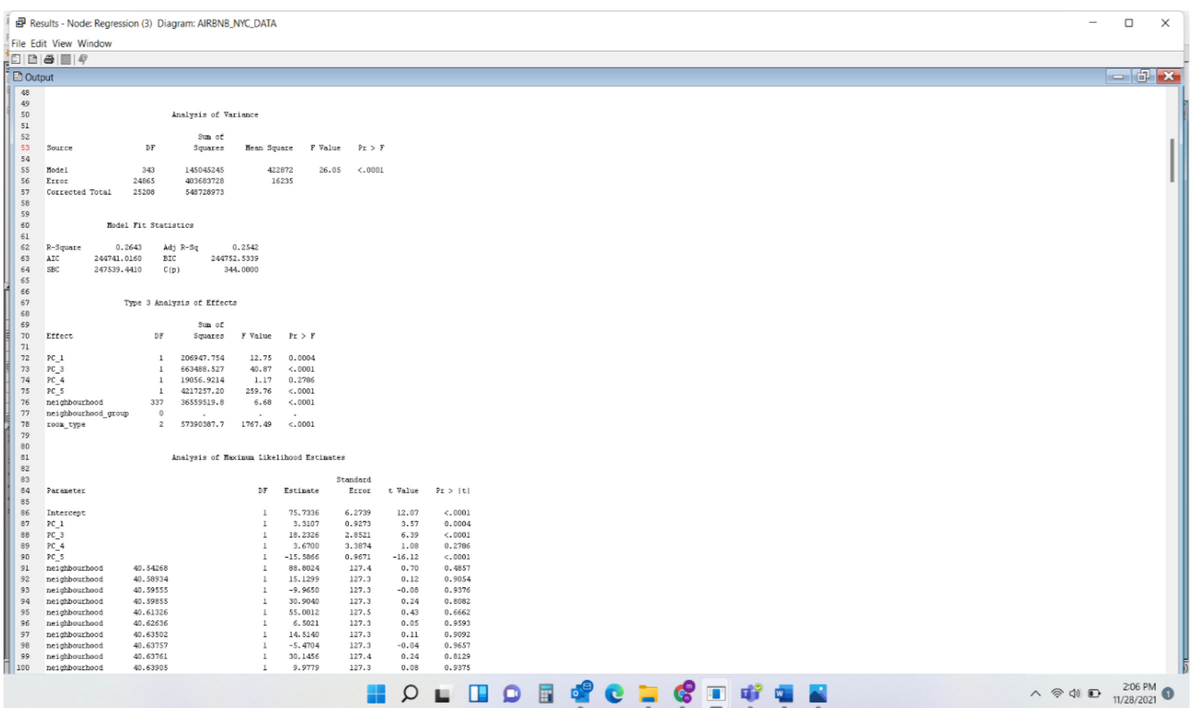


Figure: Regression Model 3 Output



d. Decision Tree:

Decision Tree algorithms are sets of nested tests that utilize top-down, recursive tests which apply the divide and conquer approach to splitting the decisions into several decisions in the form of splits in the tree until finally coming to rest at a leaf. The decision tree follows a greedy algorithm that focuses on locally optimal decisions. It is specifically helpful for analyzing categorical data.

We will also attempt to run a Decision Tree model in Python for our data. For simplicity, we will be using Jupyter notebook and interpreting the results from the output. First, we will import all the libraries in Scikit learn that is required to run our regression model.

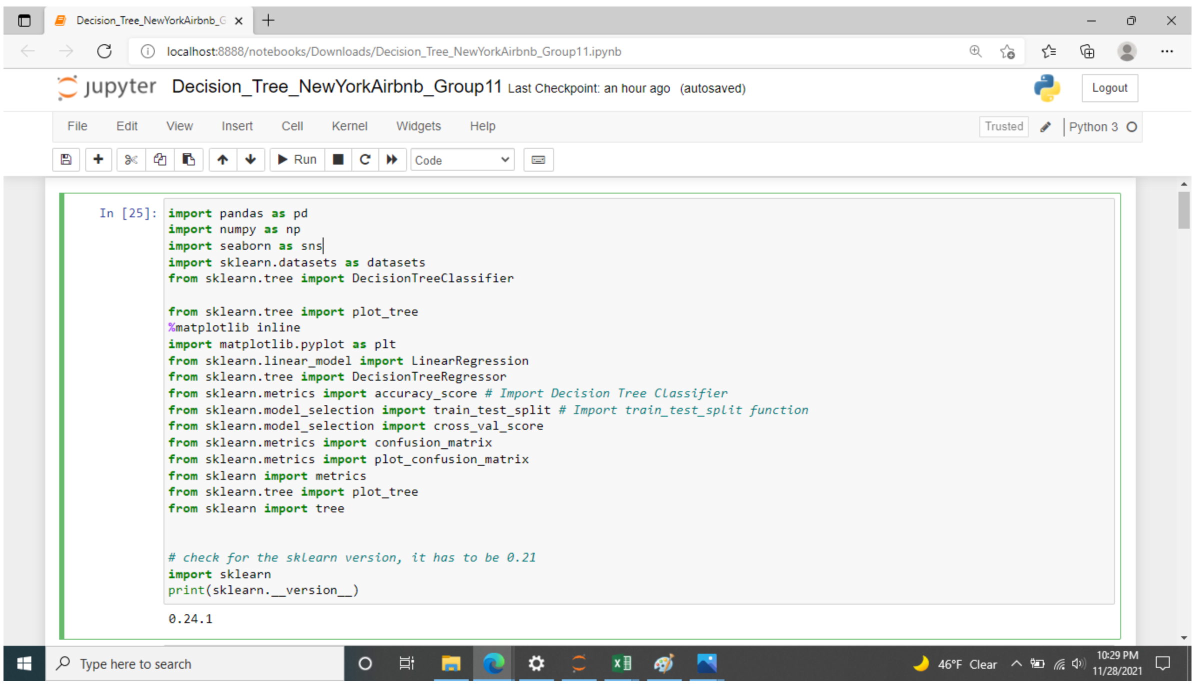


Figure: Importing Libraries for Decision Tree Analysis

Next, we will load our data using the pandas dataframe's .pd.read_csv( ) function and upload our data in the pandas data frame. We will also check the head for displaying all the variable names of our dataset. We have set our target variable as 'Price' and then we will select the features for prediction, the selected features will be our independent variables.

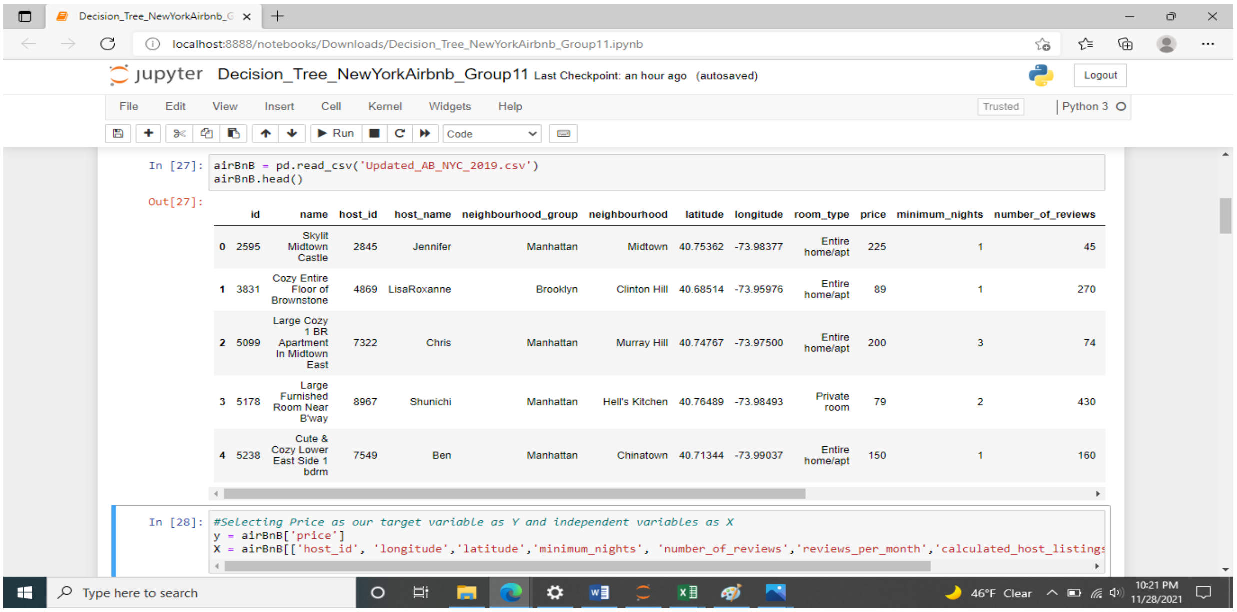


Figure: Reading the Airbnb data using pandas.read_csv function

Now, we will introduce a training and testing split for our data and at this point, we have used 67% of our data for training and 33% for testing. There is more data in the training set which might affect the accuracy of the model. Once we predict Y, we will check for the accuracy of the model. At this point, we can notice that the accuracy of our initial decision tree is less than 5 per cent.

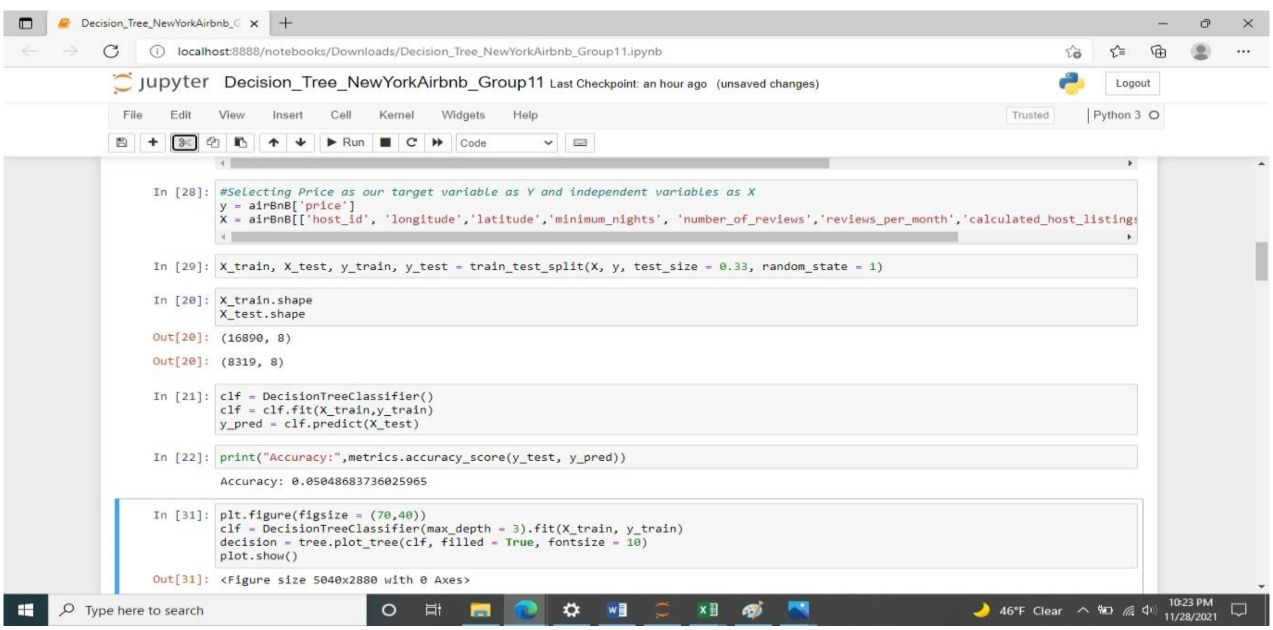


Figure: Decision Tree Training and Testing

We have generated a tree plot using sci-kit-learn, which shows that there are 8 leaves in the optimal tree. However, the accuracy of this model is very less to be considered important in our analysis. This also suggests that we need to incorporate more categorical data into the New York City Airbnb data to increase the accuracy of the optimal decision tree.

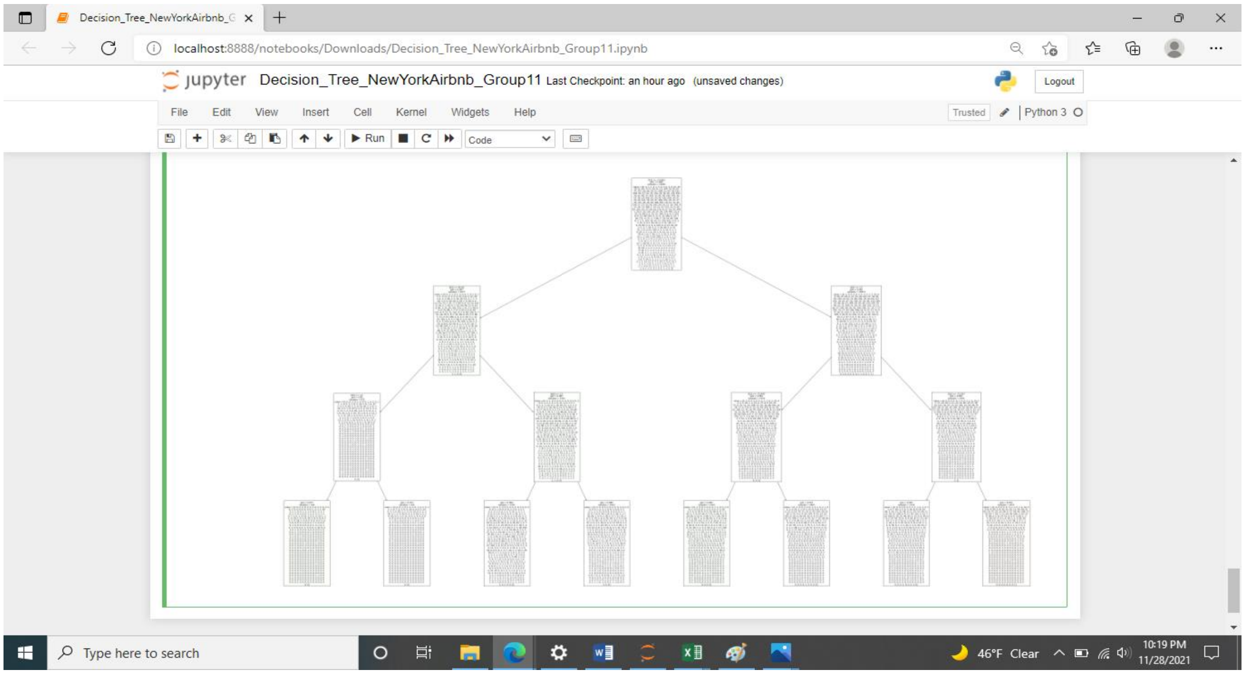


Figure: Decision Tree

MANAGERIAL IMPLICATIONS/CONCLUSION

After the analysis of our data from different perspectives like data exploration and data mining algorithms, we have observed that as analysts we must be familiar with our data on various levels, from its type, usage, significance and classes, to how it will behave under different types of algorithms being run. Data preparation and preprocessing is of prime importance for datasets like the Airbnb Listings and Metrics in New York City in 2019 data, which incorporates different variable types like IDs, categorical variables like host names, neighborhood, room type, property name and interval variables like price, reviews per month, minimum nights of stay, 365 availability and object type variables like review date.The summary statistics of our data provided us with the big picture of our data. Focusing on our target variable "Price", we can say that the price has high variations among the Airbnb rental properties within the different neighborhoods of New York City with a range of $7500 and a mean of $141.2428 for a single property booking. The missing values for our overall data including price were less than 2%, where missing values for the target variable "Price" was 0.56%. The price for Airbnb rentals is found to be highly positively skewed, with a skewness measure of 12.5. Thus, it is highly important to find out the factors that are responsible for price variation.

From our initial regression, we found that the independent variables; Availability 365, Minimum Nights, Number of Reviews, Neighborhood and Room type to have the most influence over price. However, the R-square value for our regression was just 25.88%, which means that we need additional variables data to explain the price variation. Subsequently, we performed a principal component analysis at an eigenvalue cutoff of 0.85 to find out the principal components that might better explain the price variations and performed regression analysis again taking into consideration only the significant principal components. However, the R-square value only increased slightly at 26.43%. This further confirms that we need additional data that will incorporate all significant factors responsible for the price variation.

We also performed clustering analysis, in which we obtained 13 different segments in clusters. The segments are justifiable provided a large number of our dataset and the clusters were able to classify the data well. A vast majority of the price information was represented by segment 4 with 8148 data points and more than 75% of our data were represented by 7 clusters. This tells us that majority of our price information data can be categorized into these 7 clusters and further analysis is required to gain more information on these clusters to appropriately explain the variations in the price. Using python we created a decision tree for our data, we used a train-test split of 67% and 33% which resulted in the optimal tree with 8 leaves and an accuracy of 5.05%, thus as with the Principal Component Analysis, decision tree analysis also requires more data and data preprocessing must be performed diligently to make further inferences using decision tree analysis for the New York City Airbnb data.

Lastly, our data is comprehensive data incorporating data types like identifiers, categorical variables, interval variables and object variables that require separate treatment for data analysis at each step. At present, focusing on significant predictors like room type, availability, neighborhood, reviews per month and minimum nights of the stay will better help optimize price for Airbnb hosts in New York City.

REFERENCES

Data Source: https://www.kaggle.com/dgomonov/new-york-city-airbnb-open-data

Data Interpretation: Lecture Slides by Professor Dr Mahdi Fathi

APPENDIX:

The column variables description table.

Field Name	**Order**	**Type**	**Description**
ID	1	Interval	Identification Number of the Guest
Name	2	String	Name of the Guest
Host ID	3	Interval	Identification Number of the Host
Host Name	4	String	Name of the Host
Neighbourhood Group	5	String	Neighbourhood Group
Neighbourhood	6	String	Neighbourhood
Latitude	7	Interval	Location
Longitude	8	Interval	Location
Room Type	9	String	Type of the Room
Price	10	Interval	Price of the Room
Number of nights	11	Interval	Number of nights guests booked a room
Number of Reviews	12	Interval	Total number of Reviews
Last Review	13	Interval	Date when the last review received
Reviews per month	14	Interval	Total number of reviews received per month
Calculated host listings Count	15	Interval	Count of host listings
Availability	16	Interval	Number of Days of availability