Boston Airbnb Listings Project

TEAM MEMBERS:

Nadja Jouhari

Vaishnavi Soundararajan

Phase 1: Data Set Source and Description, Value Proposition, Data Cleaning and Preparation, and Data Dictionary

SOURCE OF TEAM'S VIABLE DATA SET

Full Name: Boston Airbnb Listings Dataset

To perform in depth analysis on the features of Airbnb's listings specifically for the densely populated city, Boston.

Root URL: https://data.world/data

Specific URL: https://data.world/markdias/listingsairbnb

GENERAL DESCRIPTION OF THE PRE-CLEANED DATA SET

<What is the pre-cleaned data set's unit of analysis, i.e., what does each observation represent?>

Each observation represents an Airbnb's listing and the features the listing has. The unit of analysis is the listing.

<How many observations does the pre-cleaned data set include?>

The amount of observations for the pre-cleaned data is 5100 observations in total

<If the data are cross-sectional, which period (e.g., a certain year) do they represent? If the data are longitudinal, what range of time is represented (e.g., 1 July 2018 through 30 June 2019)?>

The data is cross-sectional. The data has the listing details from 2008 till 2019.

<How many variables does the pre-cleaned data set include? How many of these are *non-binary* numeric? How many are *non-binary* categorical? How many are binary (i.e., Y/N, 1/0)? How many are date-based? Details about each variable that was kept for analysis will be provided in the data dictionary (below), so just be general here>

Boston Airbnb Listings Project

The number of variables in this dataset is 35. Out of the 35, 15 are non-binary numeric. Then 9 are non-binary categorical variables. 5 are binary and the remaining 4 are date-based variables.

<What was the intended type of analysis at the outset? If MLR, which variable will likely serve as the dependent variable? (It must be a non-binary numeric variable.) If time series, what time intervals will appear in the plot's *x*-axis (e.g., days, weeks, months)? What is the time-series variables (or classes)?

The intended type of analysis is Simple Linear Regression (SLR) for this dataset. For SLR analysis, the dependent variable (DV) would be Price of the listings and the important IV's would be host_neighborhood, room_type and property_type.

VALUE PROPOSITION

<Who might be interested in the results and findings of this investigation? Refer to specific roles, positions, and job titles (e.g., higher education administrators, health care administrators, middle managers, and operational staff, etc.). How do they stand to potentially benefit from this investigation? >

The hosts of the listings and the guests or customers are the people who these findings would be for. Our goal is to analyze the different property type, room type and host neighborhood features of the listings and determine how it influences the price of the listings. The guests can utilize this information by looking into the features of the listing which contributes to a higher price, and which features has a lower price. The guests and customers will know which type of listing to be chosen efficiently which is a crucial factor when choosing to book a place. They could plan their budget and book a suitable place which caters to their needs.

Variables with Incorrect Values, Inscrutable Variables, and Variables Deemed Unimportant

1. Variables with Incorrect values – Zip code
2. Variables with Inscrutable values – None
3. Variables Deemed Unimportant– last_scraped and name variables were considered as unimportant variables, and these were removed from the dataset during data prep and cleaning process. There are other variables tend to not work well as predictors and those are not removed from the dataset.

Boston Airbnb Listings Project

Variables with Missing Values

<Which variables, if any, contain many missing values? Identify the number and percentage of missing values for each of these variables.

There are six variables which had larger number of missing observations and these are discussed below:

1. **host_neighbourhood –** 285 missing observations – removed: Number = 285, Percentage = (285/5100) * 100 = 5.58%
2. **Host_location –** 1847 observations – removed locations other than Boston, Massachusetts, United States: Number = 1847, Percentage = (1847/5100) * 100= 36.21%. Since, we have more than 10% of missing observations, we performed the listwise deletions and retained the variable.
3. **Security_deposit –** 898 missing observations – removed those missing observations: Number = 898, Percentage = (898/5100) * 100 = 17.60%. Since, we have more than 10% of missing observations, we performed the listwise deletions and retained the variable.
4. **Cleaning_fee –** 43 missing observations – removed those missing observations: Number = 43, Percentage = (43/5100) * 100 = 0.84%
5. **First_review –** 105 missing observations – removed those missing observations: Number = 105, Percentage = (105/5100) * 100 = 2.05%
6. **Price –** 3 observations marked 0$ for price of listing – removed those three observations: Number = 3, Percentage = (3/5100) * 100 = 0.05%

Variables with too few missing observations: There were nine variables with too few missing observations which are discussed below:

1. **Zipcode –** 8 missing observations – assigned values for those observations
2. **Property_type –** 1 missing observation – assigned value to that observation
3. **Room_type –** 1 missing observation – assigned value to that observation
4. **Accommodates –** 1 missing observation – assigned value to that observation
5. **Bathrooms –** 8 missing observations – assigned values for those observations
6. **Bedrooms –** 6 missing observations – assigned values for those observations
7. **Beds –** 2 missing observations – assigned values for those observations
8. **Bed_type –** 1 missing observation – assigned value to that observation
9. **Price –** 1 missing observation – assigned value to that observation

Boston Airbnb Listings Project

In the pre-cleaned dataset, there were 5100 total number of observations and after the data prep and cleaning process, the dataset has 1918 total number of observations.

Variables Requiring Transformation

<Identify here each categorical variable that may be made numeric. If the variable is nominal, identify which classes will be transformed into "dummy" variables. (Remember that transforming many classes into dummy variables will lead to variable proliferation and, in turn, many variables to choose amongst for MLR modeling.) If the categorical variable is ordinal, identify the scale (e.g., 1 through 5).>

The categorical ordinal variable which can be ordered in scale of 1 through 4 is the cancellation_policy.

The scales for this variable are –

Flexible = isFlexible = 1

Moderate = isModerate =2

strict_14_with_grace_period = isStrict14 = 3

super_strict_30 = isStirct30 = 4

Scale value 1 = 257 number of observations

Scale value 2 = 621 number of observations

Scale value 3 = 1039 number of observations

Scale value 4 = 1 observation

The nominal variables such as property_type, host_neighborhood and room_type are changed into dummy variables for different categories.

Property_type: There are 13 categories for the property_type and the nominal variable is transformed into 13 categories of dummy variables shown below:

Boston Airbnb Listings Project

Figure 1: Property type dummy variable creation table

Variable Creation

	Label
pt_1	property_type =Apartment
pt_2	property_type =Bed and breakfast
pt_3	property_type =Boutique hotel
pt_4	property_type =Chalet
pt_5	property_type =Condominium
pt_6	property_type =Guest suite
pt_7	property_type =Guesthouse
pt_8	property_type =Hotel
pt_9	property_type =House
pt_10	property_type =Loft
pt_11	property_type =Other
pt_12	property_type =Serviced apartment
pt_13	property_type =Townhouse

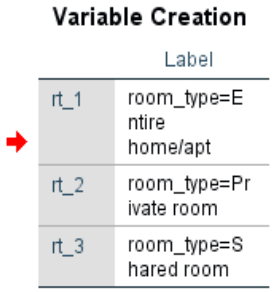
If pt_1 = 1, then property type belongs to apartment.

If pt_1 = 0, then the property type is not apartment.

The property_type with various categories will be used for the SLR analysis vis-à-vis the price of the listings.

Room_type: There are 3 categories for room_type and this nominal variable is transformed into 3 dummy variables as shown below:

Figure 2: Room type dummy variable creation table



Variable Creation

	Label
rt_1	room_type=Entire home/apt
rt_2	room_type=Private room
rt_3	room_type=Shared room

The room_type with three categories will be used for the SLR analysis vis-à-vis the price of the listings.

Host_neighborhood: There are 32 categories in the host neighborhood which is an important IV for our analysis. Since there are too many categories for host neighborhood, we have identified the top 5 major neighborhoods of Boston and created dummy variables for those categories as shown below:

Dummy Variable	Label
n_1	Host_neighborhood = Cambridge
n_2	Host_neighborhood = North End
n_3	Host_neighborhood = West End
n_4	Host_neighborhood = East Boston
n_5	Host_neighborhood = South End

These 5 major neighborhoods of Boston will be used in our SLR analysis for predicting the price per night of the listing.

<Identify each numeric variable that will be made categorical. Also identify the classes that will constitute the variable (e.g., Low, Medium, High, Very High.>

The numeric variable that will be made categorical is cleaning_fee. Since we are looking at a range of numbers for the cleaning fee, we have divided the cleaning_fee variable into six classes each consisting of a range of fees in US dollars. Each range is separated by 90$ difference for each of the classes.

The six ranges and their respective dummy variables are mentioned below:

0 – 90: firstFee = 1354 number of observations; firstFee consists of the cleaning fees range between 0$ to 90$.

91 – 181: secondFee = 504 number of observations; secondFee consists of the cleaning fees range between 91$ to 181$.

182 – 272: thirdFee = 49 number of observations; thirdFee consists of the cleaning fees range between 182$ to 272$.

273 – 363: fourthFee = 9 number of observations; fourthFee consists of the cleaning fees range between 273$ to 363$.

364 – 454: fifthFee = 1 observation; fifthFee consists of the cleaning fees range between 364$ to 454$.

455 – 545: sixthFee = 1 observation; sixthFee consists of the cleaning fees range between 455$ to 545$.

THE DATA DICTIONARY FOR THE FINAL ANALYTIC FILE

<Once you have cleaned and prepared the data set for analysis, the resulting *final analytic file* will require a data dictionary.>

<Before completing the table below, identify the unit of analysis and state the number of observations in the final analytic file.>

The unit of analysis is the listing and the final number of observations is 1918.

Guidelines for filling out the data dictionary (below):
> Each variable name should match the name in the final analytic file
> Each description should be as brief as possible yet informationally sufficient (it's an art!)
> For Antecedent Variable(s), identify (*only for derived variables*) the variable from the pre-cleaned data set from which the variable was transformed. Thus, any row in the data dictionary which includes a derived variable (i.e., a variable engineered by the team) must identify the Antecedent Variable(s).
> For Data Type, indicate whether the variable is binary, categorical, or numeric
> In the final column, indicate with a "Y" if the variable is a potential target variable

Boston Airbnb Listings Project

<If the variable has a categorical data type, identify its constituent classes (if there are 8 or fewer classes) within its Data Type cell.>

Variable Name	Brief Description	Antecedent Variable(s)?	Data Type	TV?
Id	Listing identifier (unique Id)		Numeric	
Host_Id	Each Airbnb host has a unique host_id value		Numeric	
Host_Name	To identify host name		Categorical	
Host_Since	The date of registration of the host		Date	
Host_Location	Location of the host listing		Categorical	
Host_is_Superhost	Describe the highly rated and reliable hosts (**ordinal**)		Binary	
Host_Neighborhood	Location by neighborhood	Y	Categorical	
n_1	Host_neighborhood= Cambridge			
n_2	Host_neighborhood= North End			
n_3	Host_neighborhood= West End			
n_4	Host_neighborhood= East Boston			
n_5	Host_neighborhood= South End			

Boston Airbnb Listings Project

Host_Listings_Count	The count of the number of listings a host has		Numeric	
Host_Has_Profile_Picture	A binary value stored as text where t denotes TRUE and f denotes FALSE. Profiles with pictures are more credible		Binary	
Host_Identity_Verified	A binary value stored as text where t denotes TRUE and f denotes FALSE.		Binary	
Latitude	The geographical latitude of the apartment		Numeric	
Longitude	The geographical longitude of the apartment		Numeric	
Property_Type	The type of property the listing can be categorized as	Y	Numeric	
Pt_1	Property_type = Apartment		Numeric	
Pt_2	Property_type = Bed and breakfast		Numeric	
Pt_3	Property_type = Boutique hotel		Numeric	
Pt_4	Property_type = Chalet		Numeric	

Pt_5	Property_type = Condominium		Numeric	
Pt_6	Property_type = Guest suite		Numeric	
Pt_7	Property_type = Guesthouse		Numeric	
Pt_8	Property_type = Hotel		Numeric	
Pt_9	Property_type = House		Numeric	
Pt_10	Property_type = Loft		Numeric	
Pt_11	Property_type = Other		Numeric	
Pt_12	Property_type = Serviced Apartment		Numeric	
Pt_13	Property_type = Townhouse		Numeric	
Room_Type	Whether the listing will be an entire property, private room, or shared room	Y	Numeric	
Rt_1	Room_type = Entire home/ apt		Numeric	
Rt_2	Room_type = Private room		Numeric	
Rt_3	Room_type = Shared room		Numeric	
Accommodates	Number of people the listings accommodates		Numeric	
Bathrooms	Number of bathrooms		Numeric	
Bedrooms	Number of bedrooms		Numeric	
Beds	The total number of beds		Numeric	

Boston Airbnb Listings Project

Bed_Type	Type of bed/beds provided		Categorical	
Price	Price per night for number of guests		Numeric	Y
Security_Deposit	Security deposit cost		Numeric	
Cleaning_Fee	Fee charged by host to cover the cost of cleaning the listing post stay	Y	Categorical	
firstFee	firstFee = 0$ – 90$ range of cleaning fee			
secondFee	secondFee = 91$ – 181$ range of cleaning fee			
thirdFee	thirdFee = 182$ – 272$ range of cleaning fee			
fourthFee	fourthFee = 273$ – 363$ range of cleaning fee			
fifthFee	fifthFee = 364$ – 454$ range of cleaning fee			
sixthFee	sixthFee = 455$ – 545$ range of cleaning fee			
Guests_Included	Number of guests		Numeric	
Minimum_Nights	Minimum nights the listing can be reserved for		Numeric	
Number_of_Reviews	Number of reviews for the apartment		Numeric	
First_Review	Date of the first review left		Date	

Boston Airbnb Listings Project

Last_Review	Date of the last review left		Date	
Cancellation_Policy	4 categories can be ordered from flexible to strict	Y	Numeric	
isFlexible	isFlexible = 1			
isModerate	isModerate =2			
isStrict14	isStrict14 = 3			
isStirct30	isStirct30 = 4			
Require_Guest_Profile_Picture	True/false if the host requires profile picture of guests		Binary	
Require_Guest_Phone_Verification	True/False if the host requires cell phone verification		Binary	
Calculated_Host_Listings_Count	Number of listings per host		Numeric	
Reviews_Per_Month	Number of reviews per month		Numeric	

Boston Airbnb Listings Project

Phase 2: Data Exploration Through Data Visualizations

Figure 3: Descriptive statistics for the Price variable

	price
Mean	193.8045
Standard Error	9.483569
Median	135
Mode	99
Standard Deviation	415.3327
Kurtosis	119.057
Skewness	10.5602
Range	4990
Minimum	10
Maximum	5000
Count	1918

The most often price per nice is $99.00. Since there are outliers of $4,000 and $5,000 per night, the standard deviation and range are very high. The statistics in this chart are important the analyze before creating the simple linear regression to get a full understanding of the dependent variable.

Boston Airbnb Listings Project

Figure 4: Count of prices per night

Out of all the Airbnb's, most of them have the cost range between $10-$200. The most expensive being around $5000 and those comprise of too few listings. The bar graph conveys the information here showing the different price ranges along with the number of listings.

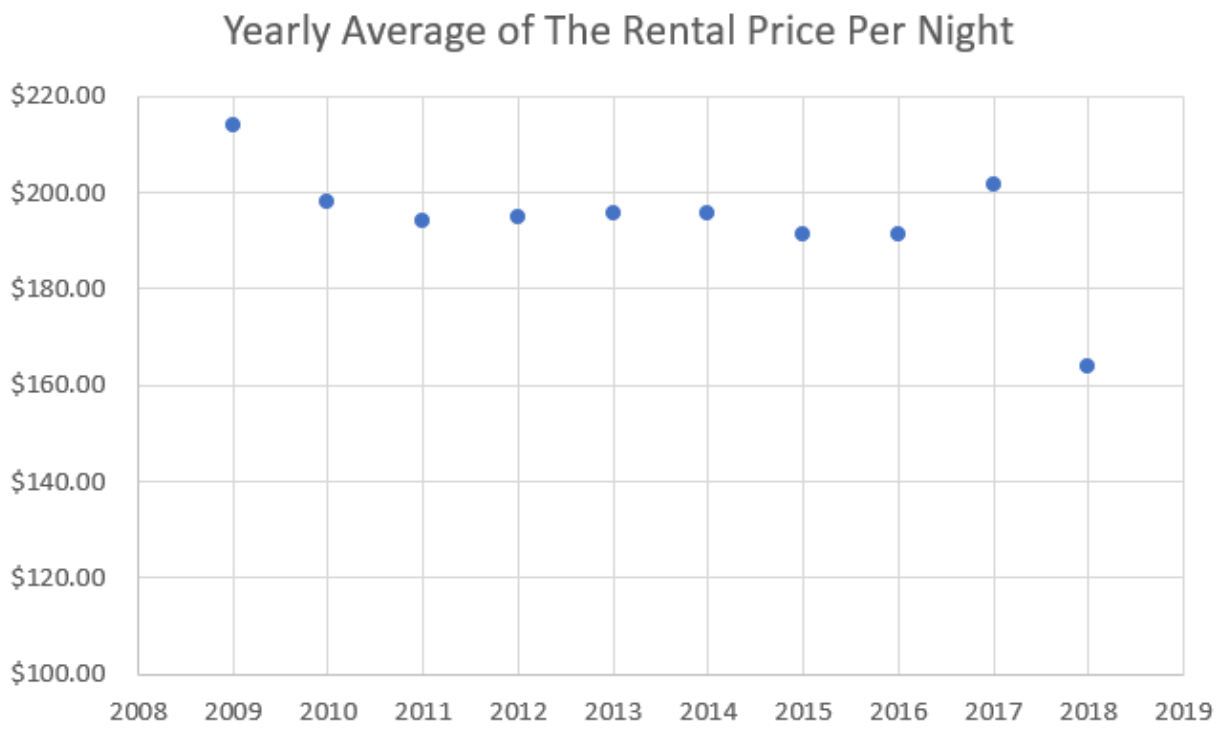


Figure 5: Yearly Average rental price per night from for the years 2008 - 2019

This graph has the average rental price per night from 2009 – 2018 only. This graph shows that overall, the price for an Airbnb stays between $180-$210 for most of the years regardless of the property type or neighborhood. In the dataset, the property types range from a private room to an entire house. So, what can be taken from this graph is that the price will hover around $190 per night anywhere in Boston and the property type does not affect the price. As the host gets to choose how much the Airbnb will go for, the price is to their choosing completely.

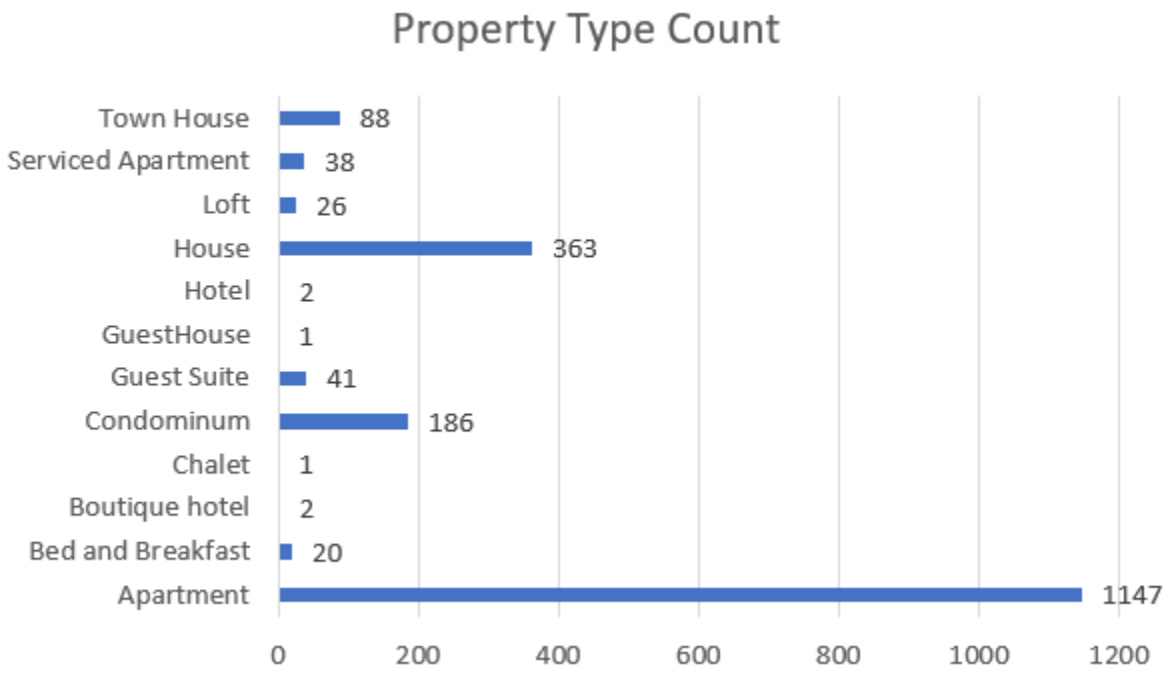


Figure 6: Count of listings for each property type

The chart above shows how many listings of each property type were in the dataset. This graph shows the variety of property types in the Boston area that were listed to be Airbnb's. Property type is one of the important predictors to the price per night of the listings and we have found from the graph that most of the property type are apartments. This is expected in a large city and densely populated city like Boston.

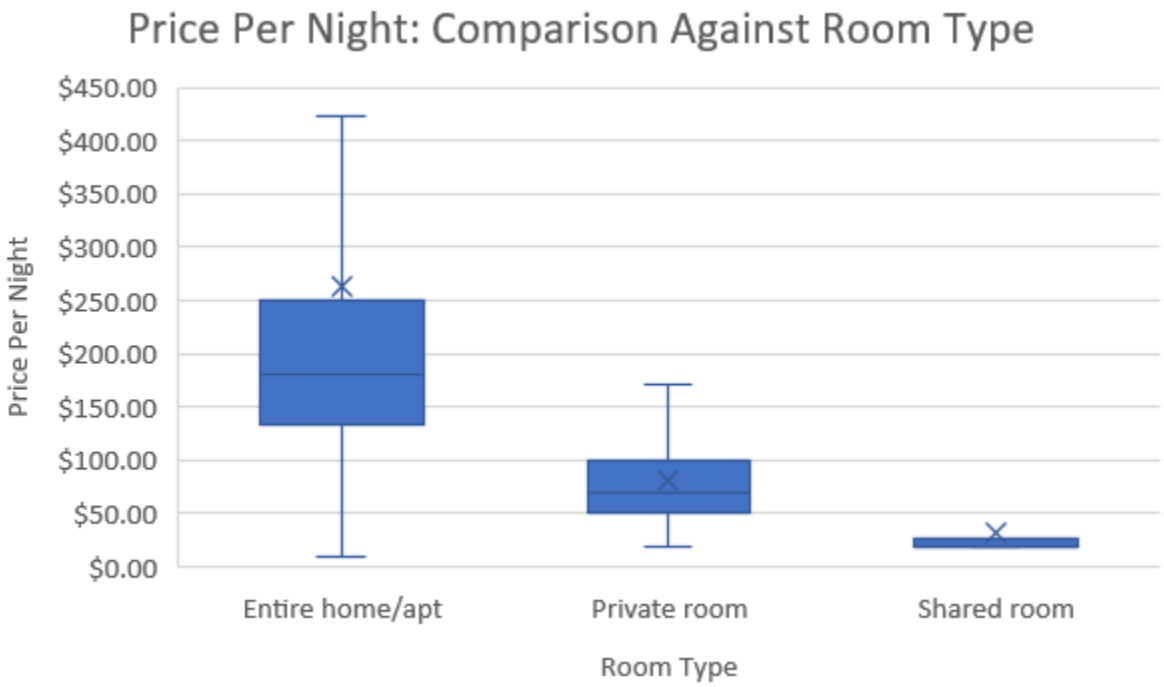


Figure 7: Price per night for each Room type category

The Box and whisker plot above show the variety of prices per night for each room type. This graph shows that the entire house or apartment has a wide range of pieces per night. Whereas the private room has a much smaller range of prices per night. This can be explained due to the dataset combining a whole house with an apartment. The outliers have been removed to better visualize the data.

Figure 8: Map showing neighborhoods of Boston

The map is used here to know the different neighborhoods and how it contributes to a higher price of the listing. We have 32 host neighborhood categories and therefore, we split them into top 5 major host neighborhoods based on this visualization such as the north, south, east, west, and central regions of Boston. This geographical map helps us visualize the proximity to the main downtown. The closer to the main hub, the more expensive the Airbnb's can get. This map also shows some of Boston's attractions, which could also influence the Airbnb's price per night.

Phase 3: MLR Analysis or Time-Series Analysis

1. Our chosen dependent variable for the dataset of Airbnb's in Boston is the price per night. We have chosen to analyze price per night against property type, room type, and host neighborhood which are important predictor variables. Price per night can be affected by many different variables, so we are

running simple linear regressions on the mentioned chosen variables to see which effects the price of a night in an Airbnb the most.

2.

a) The variables that were not included due to having no meaningful relationship to the price per night are the variables that contain information about the host, how many people the Airbnb accommodates, required visitor information, and cleaning fee range. The variables that were not chosen that pertain to the host information are Host ID, Host Name, Host Since, Host Location, Host Is Super host, Host Listings Count, Host Has Profile Pic, Host Identity Verified. Those variables which contain information about guest such as the Require guest profile picture, require guest phone verification were also not included in the analysis as it does not provide any information to Airbnb price per night. These variables do not pertain the actual Airbnb or its price per night, so they were not chosen.
b) The confidence level is 99%, so the alpha level is .01. The variables that were not chosen due to being not statistically significant at this level are the number of bedrooms, bathrooms, and beds. This has been deemed not significant because all the Airbnb's had similar amounts for these variables.
c) Most of the variables are unique and were eliminated for other reasons. The variables that have a multicollinearity impact would be the location variables because we chose to use neighborhood as once of the independent variables. Choosing another location bases variable would create similar results, which would be unnecessary. Another two that would cause multicollinearity if they were chosen would be the counts of beds and bedrooms. These two variables have most of their variables matching each other, so if they were chosen the results would also be too similar.
d) To make the model simpler, we eliminated 27 neighborhood categories from the host neighborhood predictor variable for the analysis. Too many IV's would make the model overfit. Therefore, we have used the top five neighborhood regions for our analysis such as the Cambridge, North End, South End, East Boston, and West End as the categories of host neighborhoods. These neighborhoods have significant impact on the price of the listings. Therefore, to ensure the model fitness and parsimony, we run a simple linear regression with chosen alpha level 0.01 and run 21 different LR analysis for each variable individually with the price. After the regression results, we took the top 8 variables which are statistically significant to the Price of the listings based on their p-values.

The Adjusted R^2 value is compared to each of those 8 variables and we displayed the model comparison with the help of a line chart as shown below:

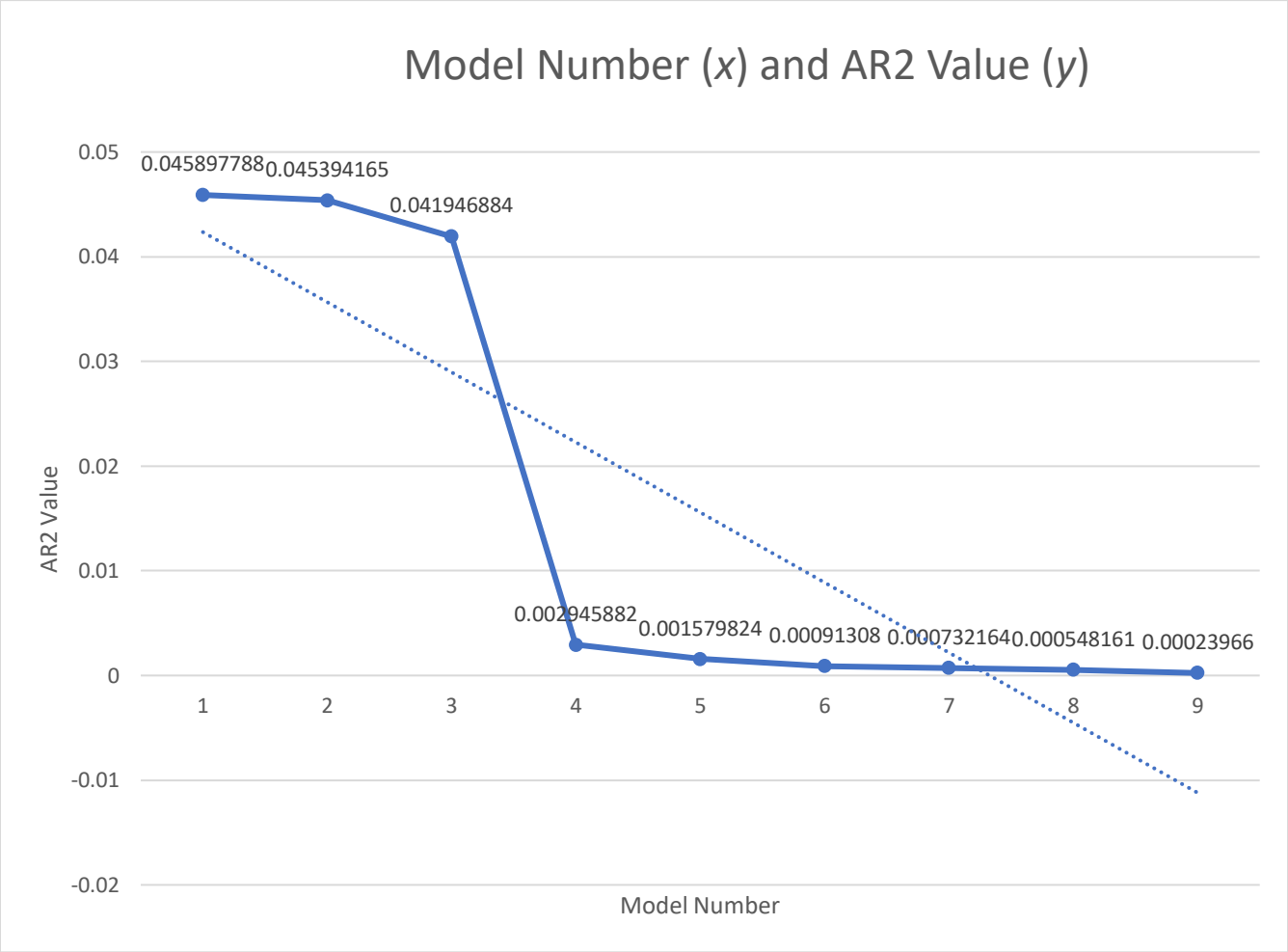


Figure 9: Line chart represents the different models with their AR2 Value

From Figure 9, we can say that there was a steep downward fall in the curve from model 3 to model 4 transition. This line chart shows a downward trendline with each of the AR2 Value for these models.

From the model comparison table shown below, we can conclude that the best fit model would be SLR Rt_2 as we do not lose many IVs in this and it has higher AR2 Value. Once we go beyond SLR Rt_2, we are losing way too much AR2 Value and the Inter-Delta R^2 value is doubled. Therefore, the SLR Rt_2, which is Private room type, would be the best model fit to our analysis. From this we can say that private room type category has the most significant importance to the price of the listing.

MODEL COMPARISON:

	modelNum	AR2Value	ModelSelection								
Entire home/apt	SLR Rt_1	0.045897788	Predictor Variable	SLR Rt_1	SLR n_2	SLR Rt_2	SLR Pt_9	SLR Rt_3	SLR Pt_5	SLR Pt_1	SLR n_4	SLR Pt_6
North End	SLR n_2	0.045394165	rt_1									
Private room	SLR Rt_2	0.041946884	n_2									
House	SLR Pt_9	0.002945882	rt_2									
Shared room	SLR Rt_3	0.001579824	pt_9									
Condominium	SLR Pt_5	0.00091308	rt_3									
Apartment	SLR Pt_1	0.000732164	pt_5									
East Boston	SLR n_4	0.000548161	pt_1									
Guest Suite	SLR Pt_6	0.00023966	n_4									
			pt_6									
			Adj R-Squared	0.045897788	0.045394165	0.041946884	0.002945882	0.00157982	0.0009131	0.00073216	0.00054816	0.00023966
			Inter-Delta R^2		-0.000503623	-0.003950904	-0.04295191	-0.044318	-0.0449847	-0.0451656	-0.0453496	-0.0456581

Figure 10: Model Comparison with AR2 Value and Inter-Delta R^2 Value

Phase 3: SLR Analysis

The simple linear regression screenshots of 21 potential IVs with the price of the listing are provided below:

Figure 11: Property Type: Apartment Simple Linear Regression

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.035403814
R Square	0.00125343
Adjusted R Square	0.000732164
Standard Error	415.1805964
Observations	1918

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	414490.3341	414490.3341	2.404585979	0.121145279
Residual	1916	330270361.3	172374.9276		
Total	1917	330684851.7			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 99.0%*	*Upper 99.0%*
Intercept	175.8741894	14.95236395	11.76229993	6.85313E-31	146.54957	205.1988088	137.3210477	214.4273311
pt_1	29.98282894	19.33536843	1.55067275	0.121145279	-7.937751524	67.90340941	-19.8714417	79.83709959

Figure 12: Property Type: Bed and Breakfast Simple Linear Regression

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.005711896
R Square	3.26258E-05
Adjusted R Square	-0.000489278
Standard Error	415.4342641
Observations	1918

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	10788.84445	10788.84445	0.062512995	0.802594055
Residual	1916	330674062.8	172585.6278		
Total	1917	330684851.7			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 99.0%*	*Upper 99.0%*
Intercept	194.0479452	9.535735787	20.34955137	1.67527E-83	175.3464326	212.7494578	169.4610254	218.634865
pt_2	-23.34794521	93.38207347	-0.250025988	0.802594055	-206.4891377	159.7932473	-264.1240742	217.4281838

Figure 13: Property Type: Boutique Hotel Simple Linear Regression

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.00301937
R Square	9.11659E-06
Adjusted R Square	-0.000512799
Standard Error	415.4391475
Observations	1918

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	3014.719538	3014.719538	0.017467553	0.894867763
Residual	1916	330681837	172589.6853		
Total	1917	330684851.7			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 99.0%*	*Upper 99.0%*
Intercept	193.8449896	9.490949575	20.42419339	4.78318E-84	175.2313118	212.4586673	169.3735464	218.3164327
pt_3	-38.84498956	293.9131177	-0.132164872	0.894867763	-615.2682457	537.5782666	-796.6699138	718.9799347

Boston Airbnb Listings Project

Figure 14: Property Type: Chalet Simple Linear Regression

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.007690015
R Square	5.91363E-05
Adjusted R Square	-0.000462753
Standard Error	415.4287572
Observations	1918

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	19555.48947	19555.48947	0.113312	0.736441484
Residual	1916	330665296.2	172581.0523		
Total	1917	330684851.7			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 99.0%*	*Upper 99.0%*
Intercept	193.8774126	9.488236472	20.43345075	4.09E-84	175.2690558	212.4857694	169.412965	218.3418603
pt_4	-139.8774126	415.5370969	-0.336618352	0.736441	-954.8299685	675.0751432	-1211.297332	931.5425072

Figure 15: Property Type: Condominium Simple Linear Regression

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.037871523
R Square	0.001434252
Adjusted R Square	0.00091308
Standard Error	415.1430107
Observations	1918

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	474285.4917	474285.5	2.751974331	0.097298087
Residual	1916	330210566.2	172343.7		
Total	1917	330684851.7			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 99.0%*	*Upper 99.0%*
Intercept	188.6512702	9.975249952	18.91193	2.83611E-73	169.0877812	208.2147593	162.9311081	214.3714323
pt_5	53.13905237	32.03255762	1.658908	0.097298087	-9.683292242	115.961397	-29.45362243	135.7317272

Figure 16: Property Type: Guest Suite Simple Linear Regression

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.02758956
R Square	0.000761184
Adjusted R Square	0.00023966
Standard Error	415.2828976
Observations	1918

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	251712	251712	1.459539	0.227152551
Residual	1916	3.3E+08	172459.9		
Total	1917	3.31E+08			

	Coefficients	*andard Err*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 99.0%*	*Upper 99.0%*
Intercept	195.4976026	9.585437	20.39527	7.78E-84	176.6986161	214.296589	170.7825337	220.2126714
pt_6	-79.20491963	65.56081	-1.20811	0.227153	-207.7829615	49.37312227	-248.2467545	89.8369152

Figure 17: Property Type: Guest House Simple Linear Regression

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.00543479
R Square	2.95369E-05
Adjusted R Square	-0.000492368
Standard Error	415.4349057
Observations	1918

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	9767.418528	9767.419	0.056594	0.81198729
Residual	1916	330675084.3	172586.2		
Total	1917	330684851.7			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 99.0%*	*Upper 99.0%*
Intercept	193.856025	9.488376902	20.43089	4.27E-84	175.2473928	212.4646573	169.3912153	218.3208348
pt_7	-98.85602504	415.5432471	-0.2379	0.811987	-913.8206426	716.1085925	-1170.291802	972.5797523

Boston Airbnb Listings Project

Figure 18: Property Type: Hotel Simple Linear Regression

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.001540983
R Square	2.37463E-06
Adjusted R Square	-0.000519545
Standard Error	415.4405479
Observations	1918

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	785.2539852	785.254	0.00455	0.946228743
Residual	1916	330684066.4	172590.8		
Total	1917	330684851.7			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 99.0%*	*Upper 99.0%*
Intercept	193.8251566	9.490981569	20.42203	4.96E-84	175.2114161	212.4388971	169.353631	218.2966822
pt_8	-19.82515658	293.9141085	-0.06745	0.946229	-596.2503559	556.6000427	-777.6526354	738.0023223

Figure 19: Property Type: House Simple Linear Regression

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.058872689
R Square	0.003465994
Adjusted R Square	0.002945882
Standard Error	414.7204583
Observations	1918

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	1146152	1146152	6.663941	0.009912001
Residual	1916	3.3E+08	171993.1		
Total	1917	3.31E+08			

	Coefficients	*andard Err*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 99.0%*	*Upper 99.0%*
Intercept	205.6154341	10.51696	19.55084	9.29E-78	184.9895398	226.2413284	178.4985254	232.7323427
pt_9	-62.40606769	24.17471	-2.58146	0.009912	-109.8175765	-14.99455884	-124.7380828	-0.074052602

Figure 20: Property Type: Loft Simple Linear Regression

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.000140255
R Square	1.96714E-08
Adjusted R Square	-0.000521901
Standard Error	415.4410371
Observations	1918

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	6.505027	6.505027	3.77E-05	0.995102259
Residual	1916	3.31E+08	172591.3		
Total	1917	3.31E+08			

	Coefficients	*andard Err*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 99.0%*	*Upper 99.0%*
Intercept	193.8113108	9.551	20.29225	4.38E-83	175.0798627	212.5427588	169.1850348	218.4375868
pt_10	-0.503618475	82.0326	-0.00614	0.995102	-161.3861828	160.3789459	-212.0162801	211.0090431

Figure 21: Property Type: Serviced Apartment Simple Linear Regression

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.008306156
R Square	6.89922E-05
Adjusted R Square	-0.000452892
Standard Error	415.4267098
Observations	1918

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	22814.68	22814.68	0.132198	0.716203764
Residual	1916	3.31E+08	172579.4		
Total	1917	3.31E+08			

	Coefficients	*andard Err*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 99.0%*	*Upper 99.0%*
Intercept	193.9409922	9.493143	20.42959	4.37E-84	175.3230125	212.5589718	169.4638933	218.418091
pt_11	-87.2743255	240.0345	-0.36359	0.716204	-558.0307176	383.4820666	-706.1787908	531.6301398

Figure 22: Property Type: Townhouse Simple Linear Regression

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.015393718
R Square	0.000236967
Adjusted R Square	-0.00028483
Standard Error	415.3918155
Observations	1918

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	78361.25	78361.25	0.454136	0.500458866
Residual	1916	3.31E+08	172550.4		
Total	1917	3.31E+08			

	Coefficients	*andard Err*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 99.0%*	*Upper 99.0%*
Intercept	192.8957447	9.580298	20.13463	6.09E-82	174.1068369	211.6846525	168.1939262	217.5975632
pt_12	45.86741321	68.06307	0.673896	0.500459	-87.61807019	179.3528966	-129.6262458	221.3610722

Figure 23: Property Type: Other Simple Linear Regression

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.01093326
R Square	0.000119536
Adjusted R Square	-0.000402322
Standard Error	415.4162103
Observations	1918

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	39528.80189	39528.8	0.229059	0.632277284
Residual	1916	330645322.9	172570.6		
Total	1917	330684851.7			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 99.0%*	*Upper 99.0%*
Intercept	194.8	9.710864509	20.06001	2.11E-81	175.7550244	213.8449756	169.7615287	219.8384713
pt_13	-21.69772727	45.33575577	-0.4786	0.632277	-110.6103426	67.21488801	-138.5913384	95.19588383

Figure 24: Room Type: Entire house/Apartment Simple Linear Regression

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.215396133
R Square	0.046395494
Adjusted R Square	0.045897788
Standard Error	405.6892924
Observations	1918

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	15342287	15342287	93.21869	1.43793E-21
Residual	1916	3.15E+08	164583.8		
Total	1917	3.31E+08			

	Coefficients	*andard Erı*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 99.0%*	*Upper 99.0%*
Intercept	78.30876217	15.12966	5.175845	2.51E-07	48.63643391	107.9810904	39.29848686	117.3190375
rt_1	184.754624	19.13568	9.654983	1.44E-21	147.2256808	222.2835672	135.4152397	234.0940082

Figure 25: Room Type: Private Room Simple Linear Regression

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.20602585
R Square	0.042446651
Adjusted R Square	0.041946884
Standard Error	406.5283973
Observations	1918

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	14036464	14036464	84.9329	7.86923E-20
Residual	1916	3.17E+08	165265.3		
Total	1917	3.31E+08			

	Coefficients	*andard Erı*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 99.0%*	*Upper 99.0%*
Intercept	258.1477551	11.6151	22.22519	1.49E-97	235.3681932	280.927317	228.1994149	288.0960954
rt_2	-178.081377	19.32327	-9.2159	7.87E-20	-215.9782218	-140.1845323	-227.9044423	-128.2583118

Figure 26: Room Type: Shared Room Simple Linear Regression

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.045832831
R Square	0.002100648
Adjusted R Square	0.001579824
Standard Error	415.004464
Observations	1918

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	694652.5925	694652.6	4.033315	0.044750314
Residual	1916	329990199.1	172228.7		
Total	1917	330684851.7			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 99.0%*	*Upper 99.0%*
Intercept	196.0354123	9.540962793	20.54671	6.07E-85	177.3236484	214.7471761	171.4350152	220.6358094
rt_3	-164.5738738	81.94639064	-2.00831	0.04475	-325.287372	-3.860375561	-375.8642638	46.71651624

Figure 27: Neighborhood: Cambridge Simple Linear Regression

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.002777462
R Square	7.71429E-06
Adjusted R	-0.000514202
Standard E	415.4394388
Observatic	1918

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regressior	1	2550.999807	2550.999807	0.014781	0.90324765
Residual	1916	330682300.7	172589.9273		
Total	1917	330684851.7			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 99.0%*	*Upper 99.0%*
Intercept	193.7672234	9.490956229	20.41598535	5.49E-84	175.1535326	212.3809	169.2957631	218.2386837
n_1	35.73277662	293.9133238	0.121575899	0.903248	-540.6908837	612.1564	-722.0926789	793.5582322

Figure 28: Neighborhood: North End Simple Linear Regression

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.214224494
R Square	0.045892134
Adjusted R Square	0.045394165
Standard Error	405.7963499
Observations	1918

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	15175833	15175833	92.15869	2.39669E-21
Residual	1916	3.16E+08	164670.7		
Total	1917	3.31E+08			

	Coefficients	*andard Erı*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*ower 99.0%*	*pper 99.0%*
Intercept	172.0762431	9.538253	18.04065	2.53E-67	153.3697933	190.7827	147.4828	196.6697
n_2	385.8774606	40.19586	9.599932	2.4E-21	307.0452333	464.7097	282.2366	489.5184

Figure 29: Neighborhood: West End Simple Linear Regression

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.005267279
R Square	2.77442E-05
Adjusted R Square	-0.000494162
Standard Error	415.4352781
Observations	1918

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	9174.596749	9174.597	0.053159	0.817678721
Residual	1916	330675677.1	172586.5		
Total	1917	330684851.7			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*ower 99.0%*	*pper 99.0%*
Intercept	193.5480973	9.550867219	20.26498	6.91E-83	174.8169088	212.2792857	168.9222	218.174
n_3	18.91344121	82.03145878	0.230563	0.817679	-141.9668929	179.7937753	-192.596	230.4232

Figure 30: Neighborhood: East Boston Simple Linear Regression

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.03270357
R Square	0.001069523
Adjusted R Square	0.000548161
Standard Error	415.2188198
Observations	1918

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	353675.213	353675.2	2.051401	0.152229318
Residual	1916	330331176.5	172406.7		
Total	1917	330684851.7			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*ower 99.0%*	*pper 99.0%*
Intercept	197.0888521	9.754340979	20.20525	1.87E-82	177.9586104	216.2191	171.9383	222.2394
n_4	-59.42847474	41.49247125	-1.43227	0.152229	-140.8036293	21.94668	-166.413	47.55562

Figure 31: Neighborhood: South End Simple Linear Regression

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.007141146
R Square	5.0996E-05
Adjusted R Square	-0.000470898
Standard Error	415.4304481
Observations	1918

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	16863.5935	16863.59	0.097713	0.754625287
Residual	1916	330667988.1	172582.5		
Total	1917	330684851.7			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 99.0%*	*Upper 99.0%*
Intercept	193.0521643	9.786354213	19.72667	5.21E-79	173.8591381	212.2451904	167.8190505	218.285278
n_5	12.43921505	39.79387557	0.312591	0.754625	-65.60464876	90.48307886	-90.16522456	115.0436547

Figure 32: Side by Side Linear Regression

	Coefficient	Standard Err	t Stat	P-value	Lower 95%	Upper 95%	Lower 99.0	Upper 99.0
rt_1	184.754624	19.13567604	9.654982848	1.43793E-21	147.2256808	222.2835672	135.4152397	234.0940082
n_2	385.8774606	40.19585515	9.599931615	2.39669E-21	307.0452333	464.7096879	282.2365577	489.5183635
rt_2	-178.081377	19.32326579	-9.215904754	7.86923E-20	-215.9782218	-140.1845323	-227.9044423	-128.2583118
pt_9	-62.40606769	24.17470883	-2.581460986	0.009912001	-109.8175765	-14.99455884	-124.7380828	-0.074052602
rt_3	-164.5738738	81.94639064	-2.00831144	0.044750314	-325.287372	-3.860375561	-375.8642638	46.71651624
pt_5	53.13905237	32.03255762	1.658907572	0.097298087	-9.683292242	115.961397	-29.45362243	135.7317272
pt_1	29.98282894	19.33536843	1.55067275	0.121145279	-7.937751524	67.90340941	-19.8714417	79.83709959
n_4	-59.42847474	41.49247125	-1.432271276	0.152229318	-140.8036293	21.94667985	-166.4125697	47.55562019
pt_6	-79.20491963	65.56080581	-1.208113882	0.227152551	-207.7829615	49.37312227	-248.2467545	89.8369152
pt_12	45.86741321	68.06306681	0.673895776	0.500458866	-87.61807019	179.3528966	-129.6262458	221.3610722
pt_13	-21.69772727	45.33575577	-0.478600762	0.632277284	-110.6103426	67.21488801	-138.5913384	95.19588383
pt_11	-87.2743255	240.0345187	-0.363590728	0.716203764	-558.0307176	383.4820666	-706.1787908	531.6301398
pt_4	-139.8774126	415.5370969	-0.336618352	0.736441484	-954.8299685	675.0751432	-1211.297332	931.5425072
n_5	12.43921505	39.79387557	0.312591193	0.754625287	-65.60464876	90.48307886	-90.16522456	115.0436547
pt_2	-23.34794521	93.38207347	-0.250025988	0.802594055	-206.4891377	159.7932473	-264.1240742	217.4281838
pt_7	-98.85602504	415.5432471	-0.237895877	0.81198729	-913.8206426	716.1085925	-1170.291802	972.5797523
n_3	18.91344121	82.03145878	0.230563268	0.817678721	-141.9668929	179.7937753	-192.5962883	230.4231708
pt_3	-38.84498956	293.9131177	-0.132164872	0.894867763	-615.2682457	537.5782666	-796.6699138	718.9799347
n_1	35.73277662	293.9133238	0.121575899	0.90324765	-540.6908837	612.156437	-722.0926789	793.5582322
pt_8	-19.82515658	293.9141085	-0.067452211	0.946228743	-596.2503559	556.6000427	-777.6526354	738.0023223
pt_10	-0.503618475	82.03259595	-0.006139248	0.995102258	-161.3861828	160.3789459	-212.0162801	211.0090431

Phase 4: Advice to Decision Makers

In conclusion, the top variables that were the most significant for the dependent variable were the room types entire room/apartments and private room along with the North End Neighborhood of Boston. The property type House was the most significant of its category as well. When analyzing these variables against price per night of each Airbnb, the room type was the variable that had the most impact on the price per night. Sharing a room will obviously be less expensive than booking an entire house, so the results are very predictable. Nevertheless, this analysis can still be useful for both the host of the Airbnb and the visitor booking an Airbnb.

As the host advertising their Airbnb, the results can help them choose their price by understanding what room types and neighborhoods have the higher prices. The room type of entire house/apartment had then largest range of prices, so an option for the host is to advertise a lower than average cost to have a competitive advantage.

Referring to the map provided, the North End of Boston consist of suburbs that are spread out. This means that most of the North End will be entire houses. Figure 8 shows that the prices per night for an entire house mostly range form $140-250, which is the highest for the room types. With that information, as an individual choosing an Airbnb to book, I would avoid the North End to cut the potential to costs.

Boston Airbnb Listings Project

The decisions in what to advertise and how to price the host's Airbnb will change due to this information. Seeing which variable to consider for the price per night and which to leave out of the calculation is very valuable. If I were the host of an Airbnb in Boston, I would use these results as a tool to decide how to price each aspect of the Airbnb to be the most desirable cost to the customer.

If I were to obtain better data for this analysis, I would like to have information on the booked Airbnb's as well. This dataset pertains to all the Airbnb's in Boston. If we had the information about the same variables and which ones were booked, an analysis over which variables were the most desirable to the customer would bring everything to another level.