**Group 1-Airbnb Project Final Paper**

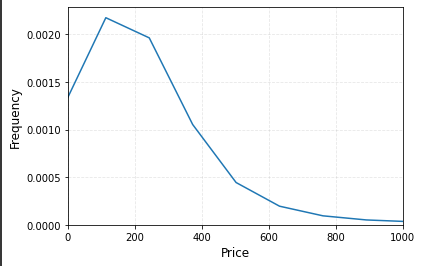
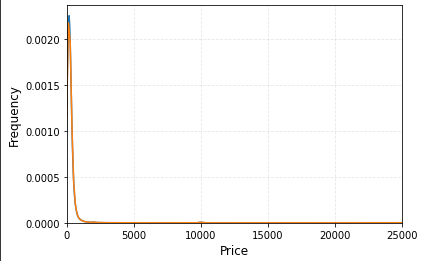
Airbnb is a popular online marketplace where users can rent short or long term stays at properties around the world. Founded in 2007 in San Francisco with just 3 properties, the company has grown to be in 220 countries around the world and over 100,000 cities worldwide. The app has over 4 million hosts who have welcomed more than 1 billion guests. Airbnb’s hosts and guests generated over $117 billion in 2019 alone. With so many properties and potential to make considerable revenue, our group was interested in understanding how to price Airbnb listings based on their characteristics. Would guests value amenities over space? What is the average price per night by neighborhood? What amenities are most sought after? After creating questions to answer, we set our dependent variable as ‘Price’, as we assumed that the higher the price per night the more successful the Airbnb property was.

We downloaded our data from <http://insideairbnb.com/get-the-data/>. After some discussion, we decided to analyze San Francisco’s data and downloaded the listings.csv and reviews.csv datasets. Considering the volume of each dataset (approximately 7,000 rows of data), we started by cleaning the datasets. First, we decided to drop 24 columns that we determined would not be useful in analysis. The columns were URLs for profile pictures for hosts and host information (such as their bio). Once the columns were dropped, we checked how many null values were in each of the columns. We caught that in the ‘Bedrooms’ column, there were many null values as many San Francisco properties are studio apartments (e.g no dedicated bedroom), so we replaced all null values in that column with 0. After transforming that column, we dropped all rows with null values. Next, we transformed the ‘Price’ column, as in the original data it had a ‘$’ which makes analysis ineffective. We discarded the ‘$’, double checked it was removed and finished our data cleaning. For the cleaned listings dataset, we had 3,573 rows and 57 columns. We also created a combined dataset with both the listings and reviews files; however, we ended up not utilizing this in our final model building.

Early on our team set a strong foundation of communication, both combining weekly meetings and a text group chat to communicate with each other. All team members were easy to get in touch with, which made any issues that came up easy to resolve. There was a shared drive that all members had access to, which allowed for seamless sharing of datasets and code files. We collaborated effectively, despite dealing with multiple time zones and many of us getting sick throughout the semester. After submitting our first iteration of this project, we received feedback that let us know that we needed more depth to our project. We swarmed the issue and had discussions on what exactly to focus on. Combining group discussions with direction from the professor, we were able to get back on track quickly.

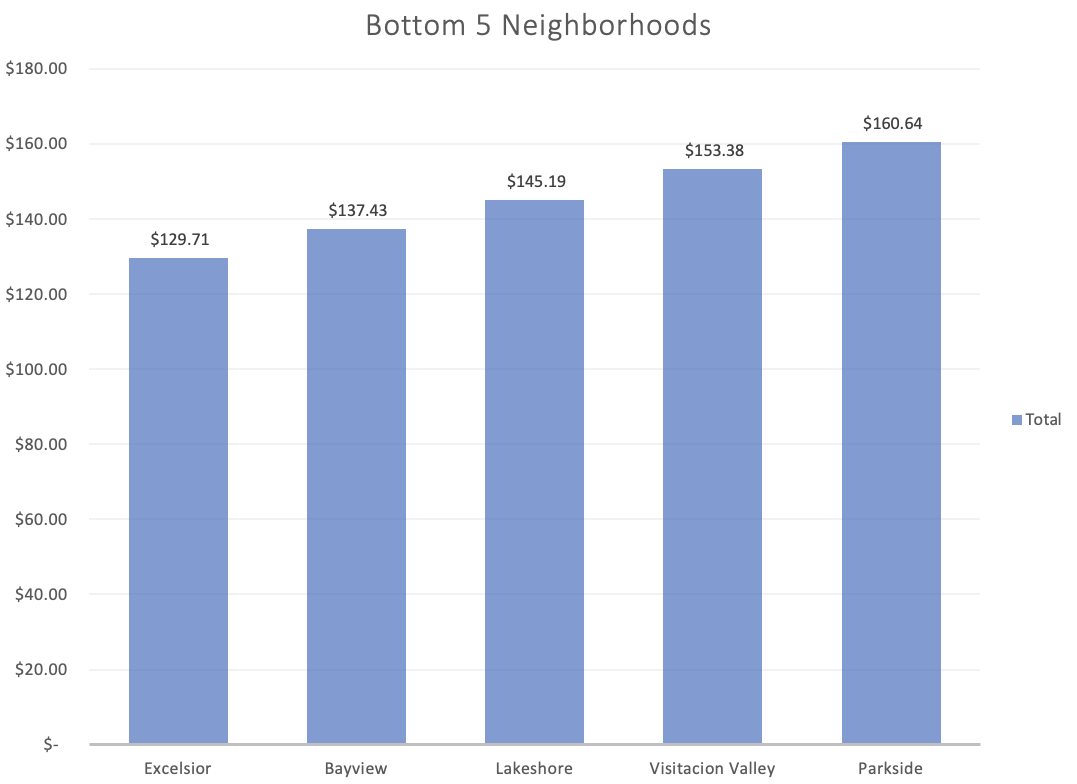
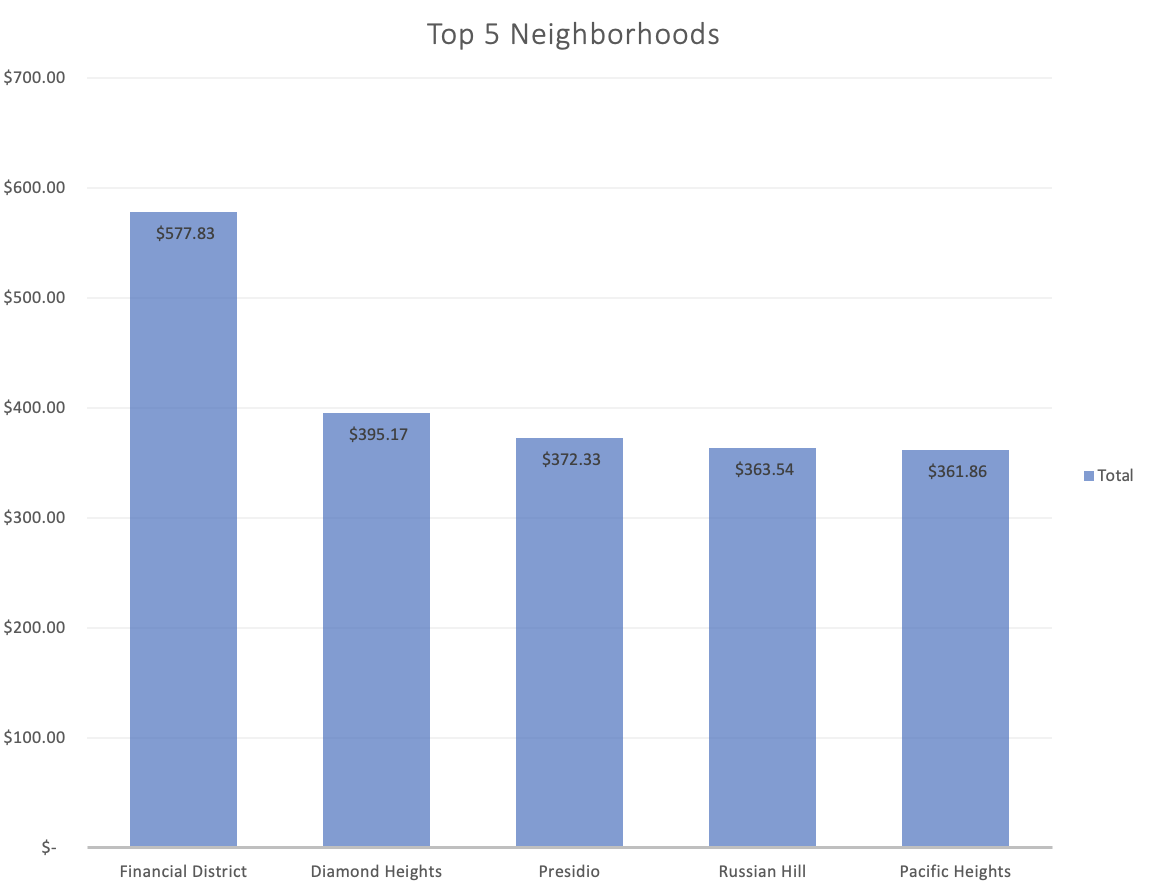
**Beginning Analysis**

As a preliminary step in our project, we wanted to see the price distribution of the Airbnb listings in San Francisco. We found that we were working with price points up to $25,000, despite this we can account for 98.08% of our price distribution under $1,000. This gives us an expectation that we will see some higher price points and may experience outliers in our data. We don’t need to take any further analysis into the values above $1,000, as this sets the foundation for our project going forward. The graphs for the price distribution are shown below:



The first analysis step we took was looking at neighborhoods and seeing how their average price point will vary. Using the cleaned neighborhood column in our dataset we were able to calculate the average price points desired. We wanted to see the average price between all neighborhoods so we can see how much the low and high price points vary from the average. The average price per night of all 36 neighborhoods was $226.77. The top five neighborhoods by price per night are: Financial District at $577.83, Diamond Heights at $395.17, Presidio at $372.33, Russian Hill at $363.54, and Pacific Heights at $361.86. All 5 of the top neighborhoods had a higher price per night than the average for all 36 neighborhoods. Finding the 5 neighborhoods with the lowest average price point, we see a smaller price distribution, all significantly lower than the average price of all neighborhoods. The five lowest are: Excelsior at $129.71, Bayview at $137.43, Lakeshore at $145.19, Visitacion Valley at $153.38 and Parkside at $160.64.

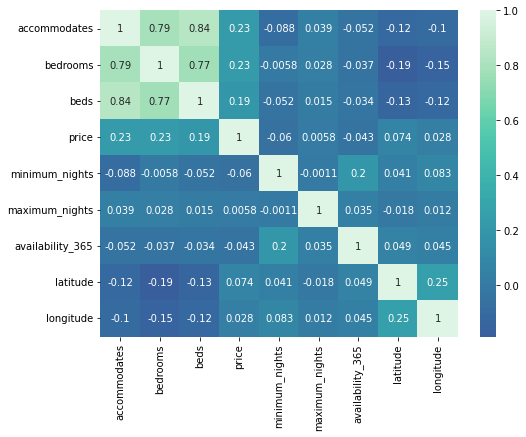
**Graph 1 Graph 2**



**Further Analysis**

Our group knew we wanted to build a model going forward to predict price based on other values. We imagine this model being used when a potential host is acquiring a property for Airbnb, they can input different characteristics of the property (number of bedrooms/bathrooms, amenities, etc) and get an estimated price per night. To assist us in building this model, we wanted to create a correlation matrix to see what columns have the highest correlation to our dependent variable “price”. Using seaborn we can easily get a graphical representation of the correlation values for each column. This visualization is shown below:

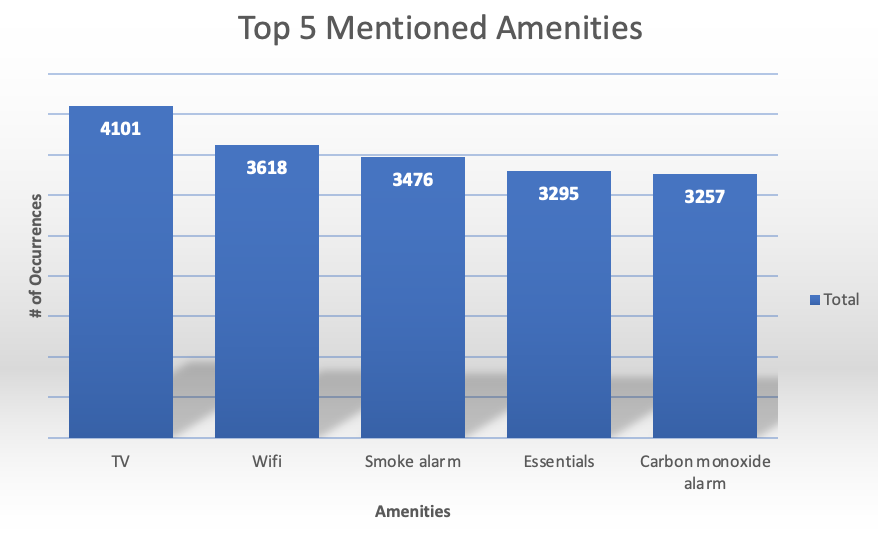
**Table 1**



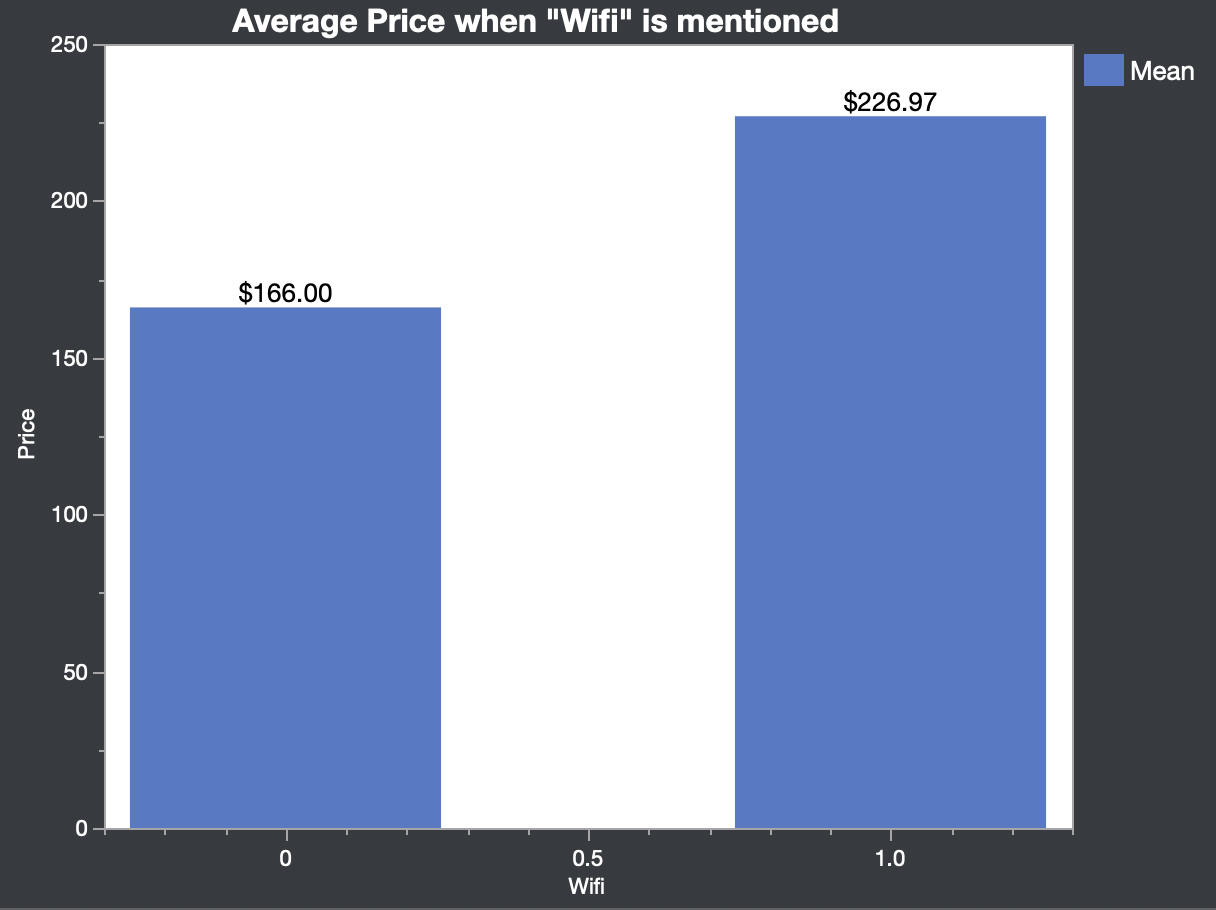
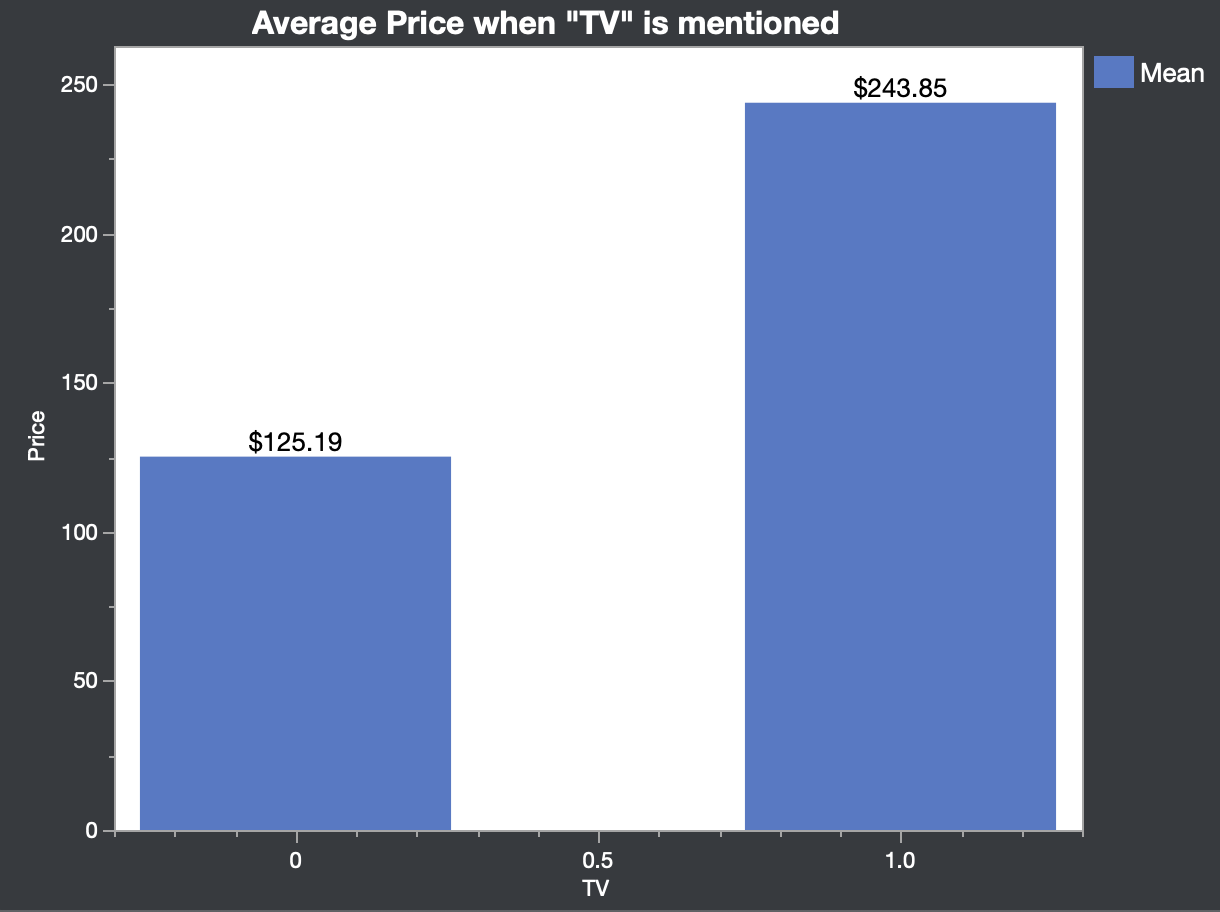
Here we can see that the lighter colors represent a higher correlation, and that the highest correlation to price are the columns: Accommodates (.23), Bedrooms (.23), and Beds (.19). We see that all other columns in this matrix are substantially low, so our best approach to a model will be using these variables.

We also wanted to take a look into the amenities column of our data to see the most frequently mentioned amenity and how those influence price. Since we are working with text, we need to run a text analysis, which is setting up a for loop to count how many times each amenity was mentioned for the listings. The top five amenities found are: TV, WiFi, smoke alarm, essentials, and carbon monoxide alarms. Looking at how this impacted the price listing, we only found that the higher priced properties had the amenities mentioned, but no other correlation was found, showing that there isn’t a statistically significant correlation between these amenities and price. Our visuals for this information is below on the following page:

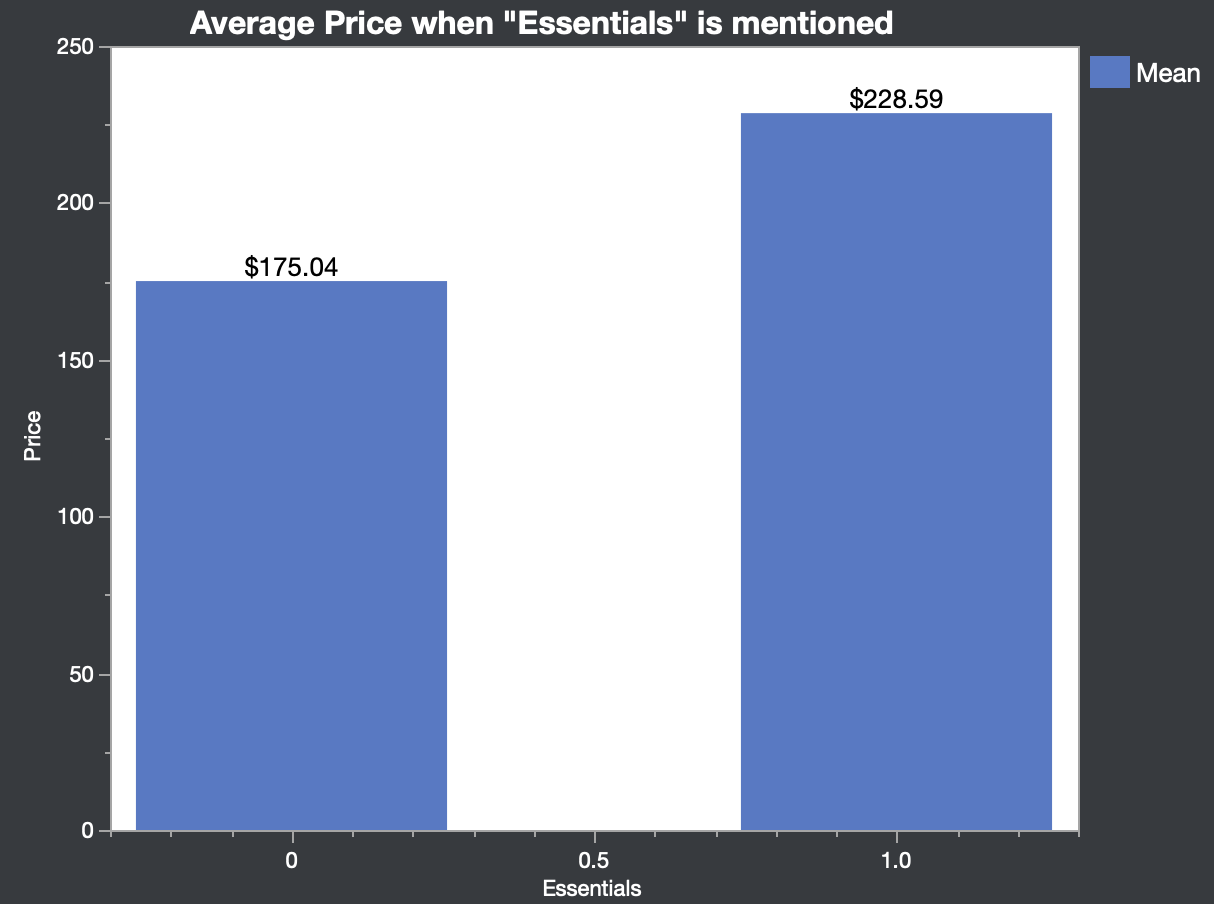
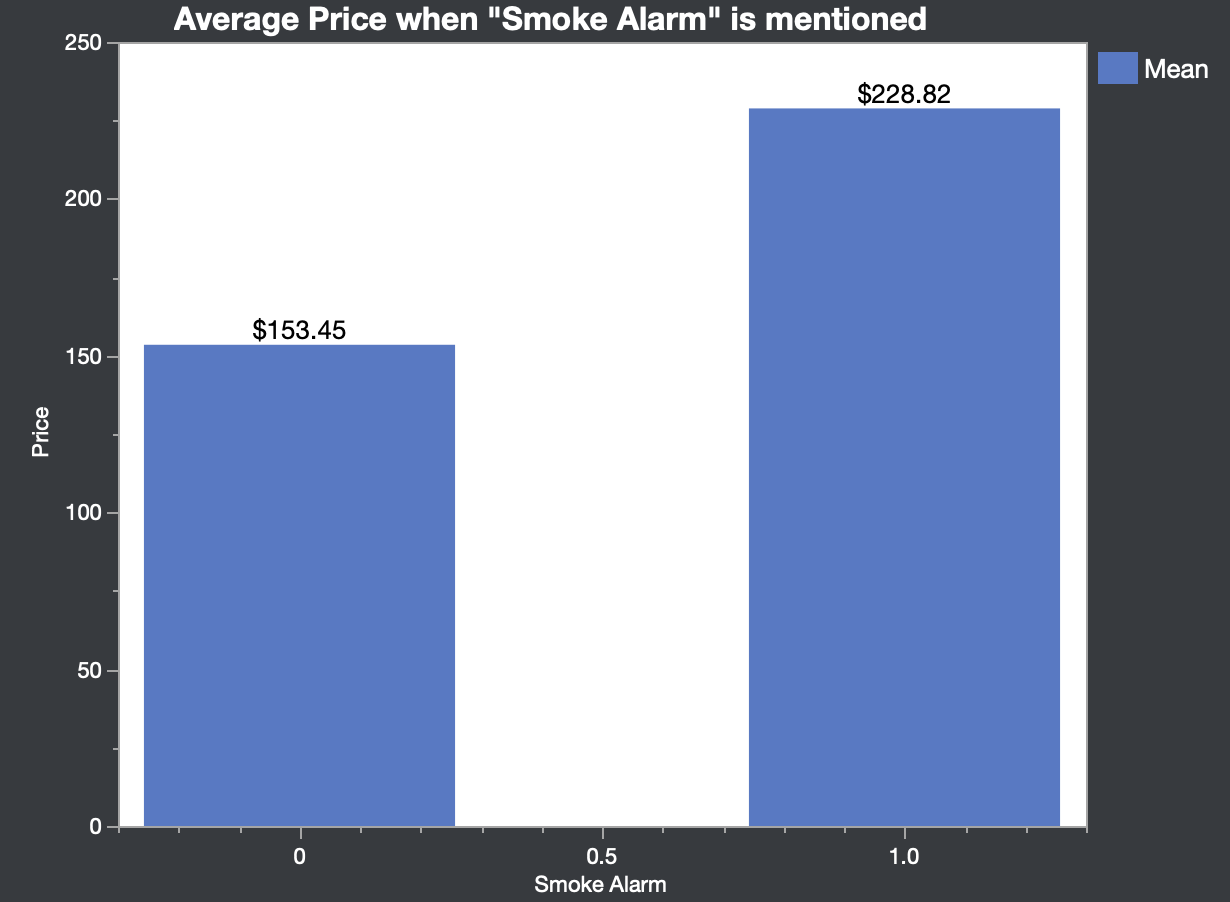
**Table 2**



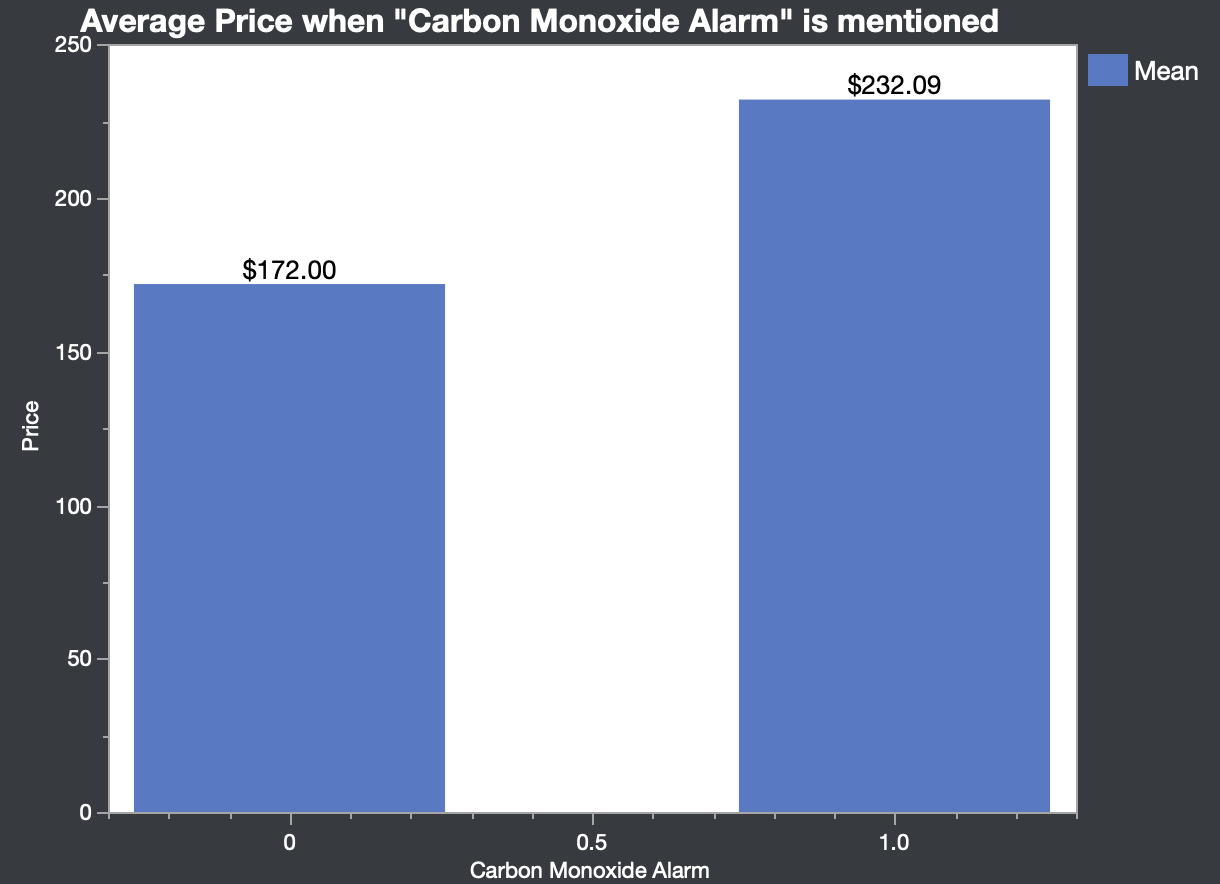
**Table 3 Table 4**



**Table 5 Table 6**



**Table 7**

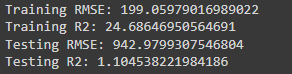


**Predictive Modeling**

The next step for our group was now seeing if we can predict prices based on variables found to be significant to our variable price. We did two models, both with linear regression methods, one done in Google Collab, and one done in Orange. Nathan and Saba made sure to have slightly varied approaches to see the accuracy comparison, as dealing with the results thus far, we were expecting to see poor results in making any type of model.

Starting with Saba’s model, this was done using only 3 variables: bedrooms, availability\_365, and minimum\_nights. This model used the same listings dataset that was cleaned more accurately. This was then split into an 80% training set, and 20% test set. Poor accuracy was found in both sets as shown below:

**Table 8**



The takeaway here is that with the data provided and cleaned, we can predict price with a 25% accuracy, but when we begin predicting based on new data, we see a 1.1% accuracy. The RMSE values show how poor our accuracy is. Our group was expecting to see this based on the results of analysis prior to modeling.

Next, we will look at Nathan’s model to see if there is a difference in results. Nathan’s approach used a different cleaned data set (one that was not as accurately cleaned) to see if that impacts our accuracy. The method for testing the data sample will be using a random forest, as with signs of overfitting from the linear regression, which could increase accuracy. More columns were used to provide more average values when pulling from the random forest. The columns analyzed were accommodates, bedrooms, beds, minimum\_nights, maximum\_nights, number\_of\_reviews, review\_scores\_rating, host\_response\_rate, and host\_acceptance\_rate. This was split into a 66% training set, and a 34% test set. When testing with random sampling we see the accuracy metrics below:

**Table 9**



We can see a huge improvement in this analysis method with an r squared score of 77.7%, and a RMSE value of 350.987. Seeing increased accuracy leads us to believe linear regression is not a good analysis method and predictive model for this data. We would like to see more variables compared for linear regression; however, this was still leading to low accuracy due to low coefficient correlations to our variable price. Random forest will pull averages which helps with analysis steps that lead to overfitting data. This is a huge reason we see such a dramatic difference in accuracy between the two methods.

**Conclusions**

Dealing with low accuracy throughout our findings prior to random forest modeling could be a result of data cleaning methods or how data is kept. The biggest disadvantage is that the “bathrooms” column of our data is blank. Seeing how bedrooms correlate high to price, we would expect bathrooms to appear high on the correlation coefficient as well. Anybody with mediocre knowledge of the housing market would know that this would be a great asset to have, as well as square footage. If we could have had this information, we assume that we would have higher model accuracy.

Our result of 77% accuracy gives a strong foundation for a predictive model. Our group was content seeing that random forest gave a significant accuracy increase, as prior to this method we were struggling to produce a model with high accuracy. This shows that incorporating and each working on different methods is a strong suit to ensure poor accuracy across multiple methods if we have poor data or find a method that gives consistent results.

.