Prediction of Toronto Airbnb Prices
Capstone 1- Slide Deck Presentation

Presenter: Puneeth Nagarajaiah

Date: 2020/05/15

Summary Data

Total listings 22,425

Mean Price is $143.53

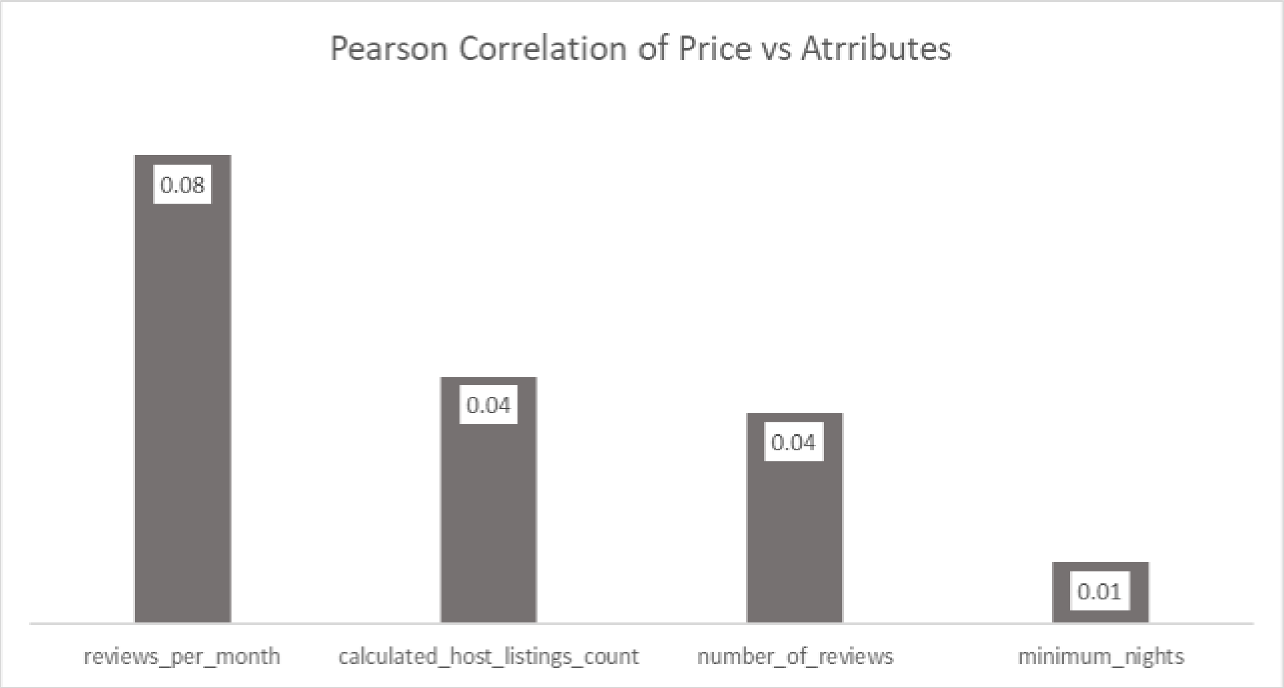
Median Price is $99 (difference indicates presence of outliers)

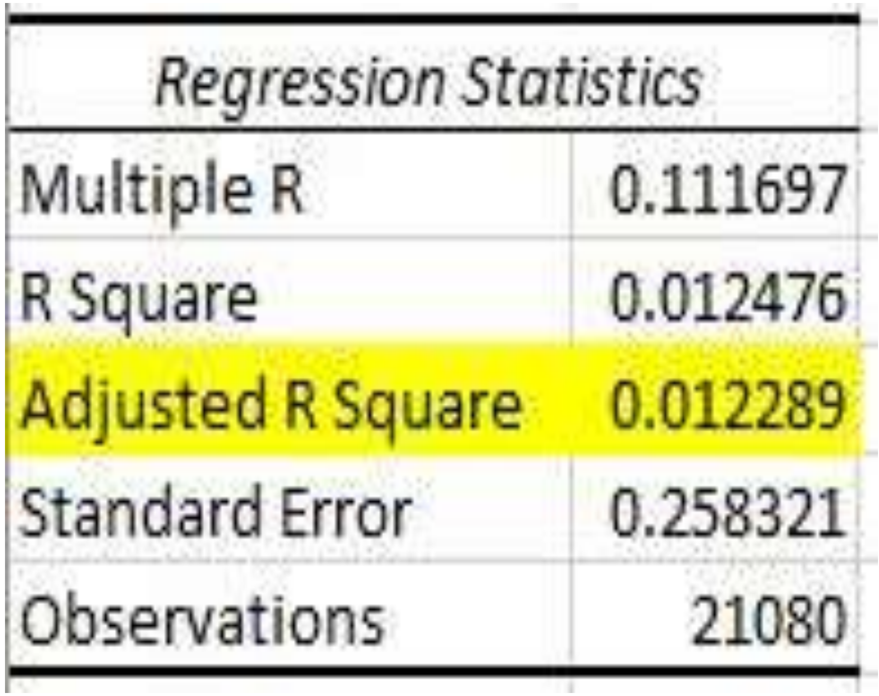
Maximum price is $14,008 per night (Penthouse)

Minimum Price is $0 (indicates property is now unlisted)

Mean > Mode: Indicates distribution of Price data is right skewed

The model can only predict 1% variation in prices





Regression Statistics	
Multiple R	0.111697
R Square	0.012476
Adjusted R Square	0.012289
Standard Error	0.258321
Observations	21080

- The Summary Data released for analysis by our source insideairbnb.com has only 14 attributes
- Of this, there are only four numerical variables that we can consider for analysis
- Since this model is very weak, we will try considering the full raw data set from our source

Raw Data

Total listings 23,397

Mean Price is $148.70

Median Price is $99 (difference indicates presence of outliers)

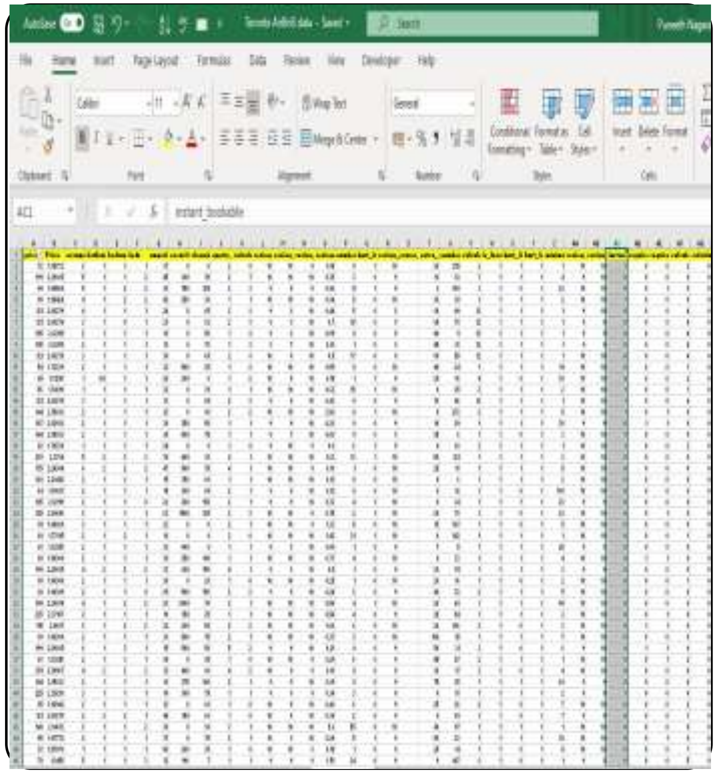
Maximum price is $13,244 per night (Penthouse)

Minimum Price is $0 (indicates property is now unlisted)

Mean > Mode: Indicates distribution of Price data is right skewed

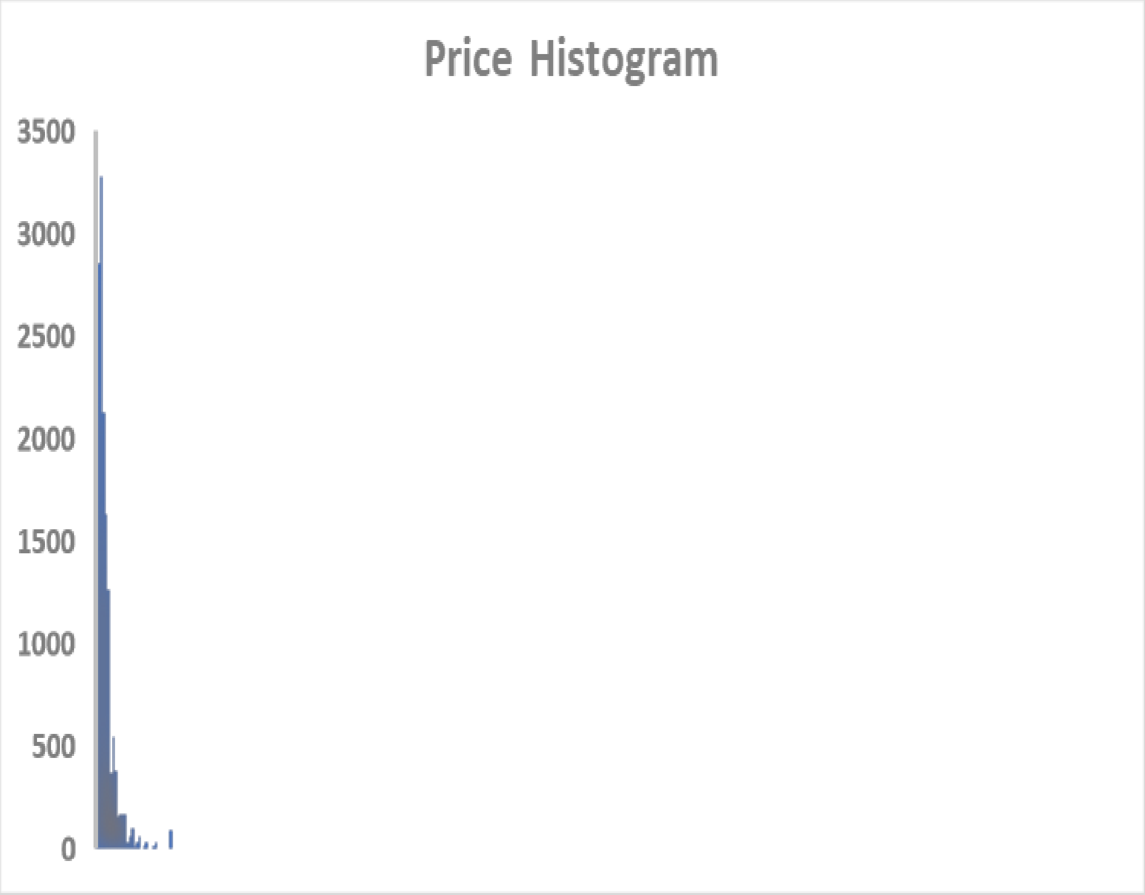
Data Cleaning

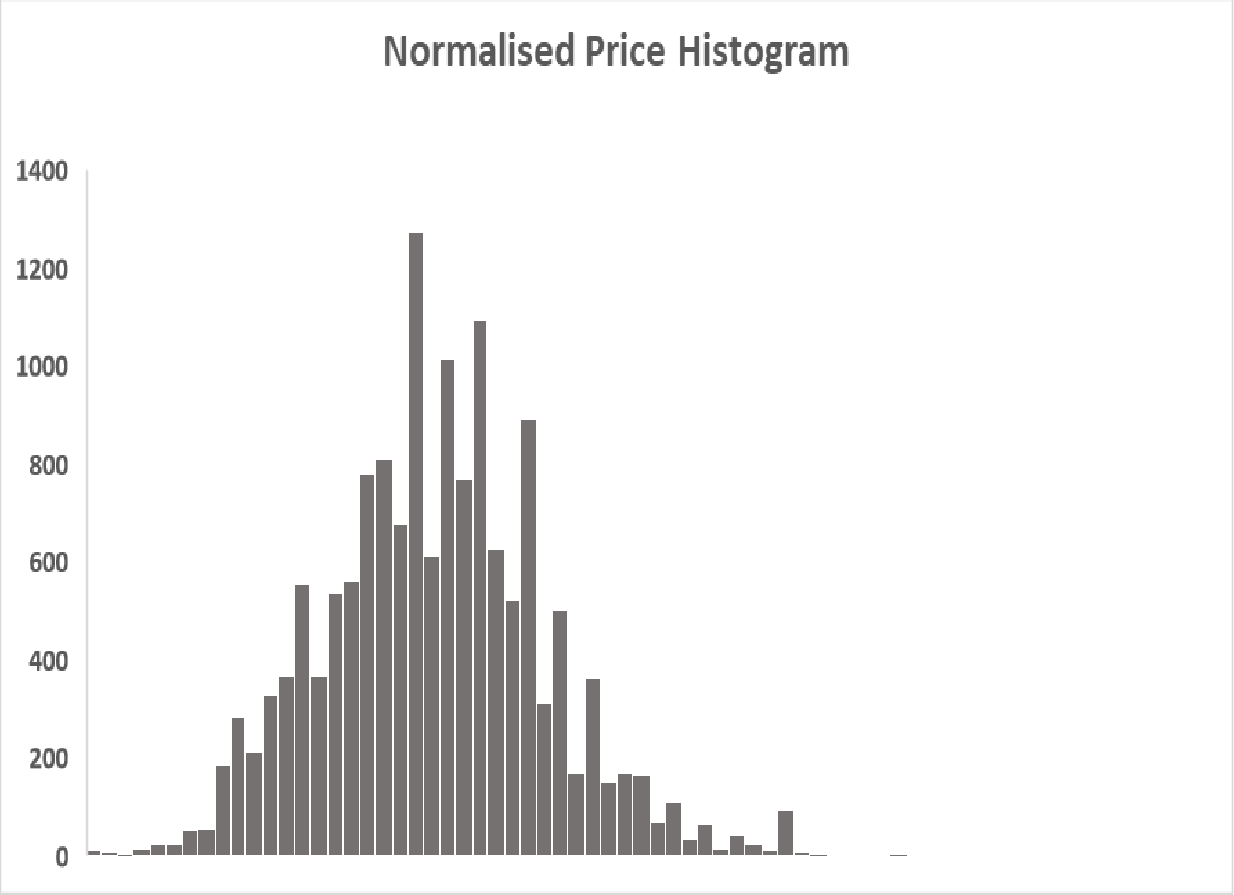
- Total attributes 72
- We will only consider numerical and categorical data with only t/f options
- sqft, weekly_price and monthly_price is mostly blank; availability, max_nights and other futuristic data will not be considered
- Blank rows will be deleted; imputation through mean/median or constant value could lead to bias
- We are now left with 31 attributes for analysis



AC1 instant_bookable

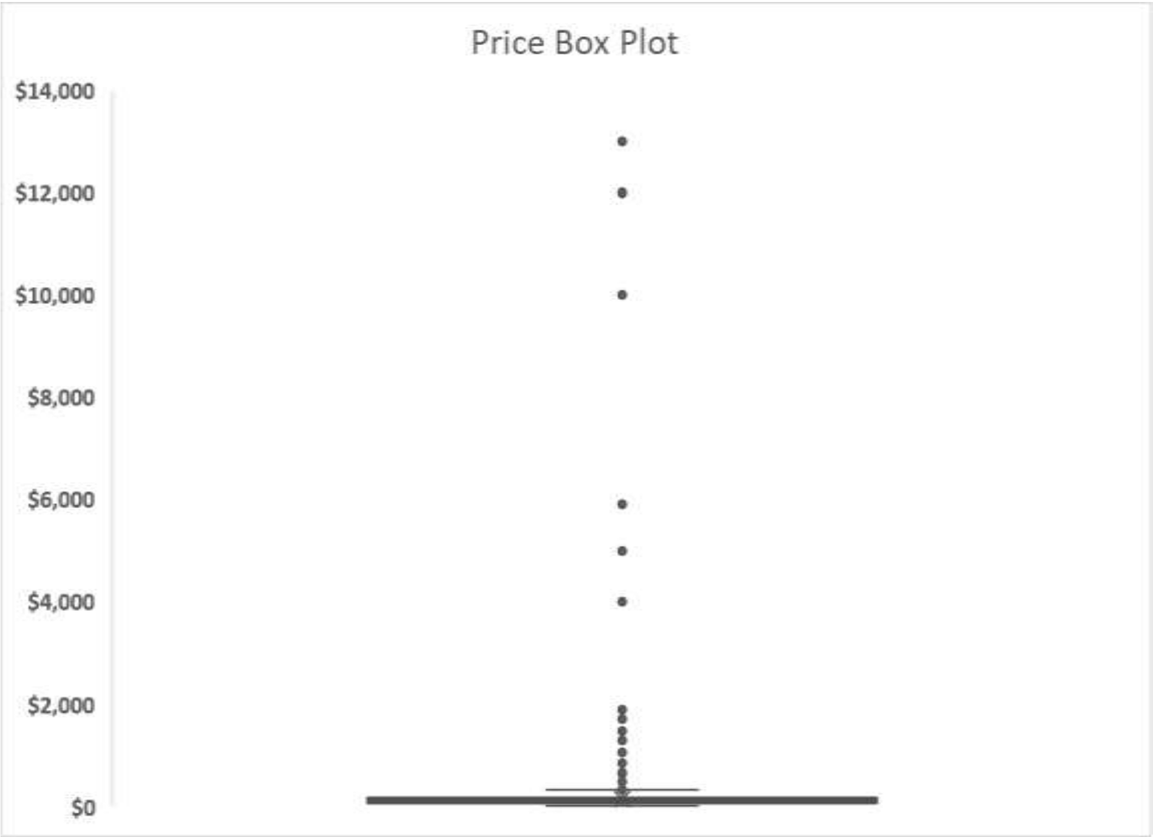
‘Price’ data distribution

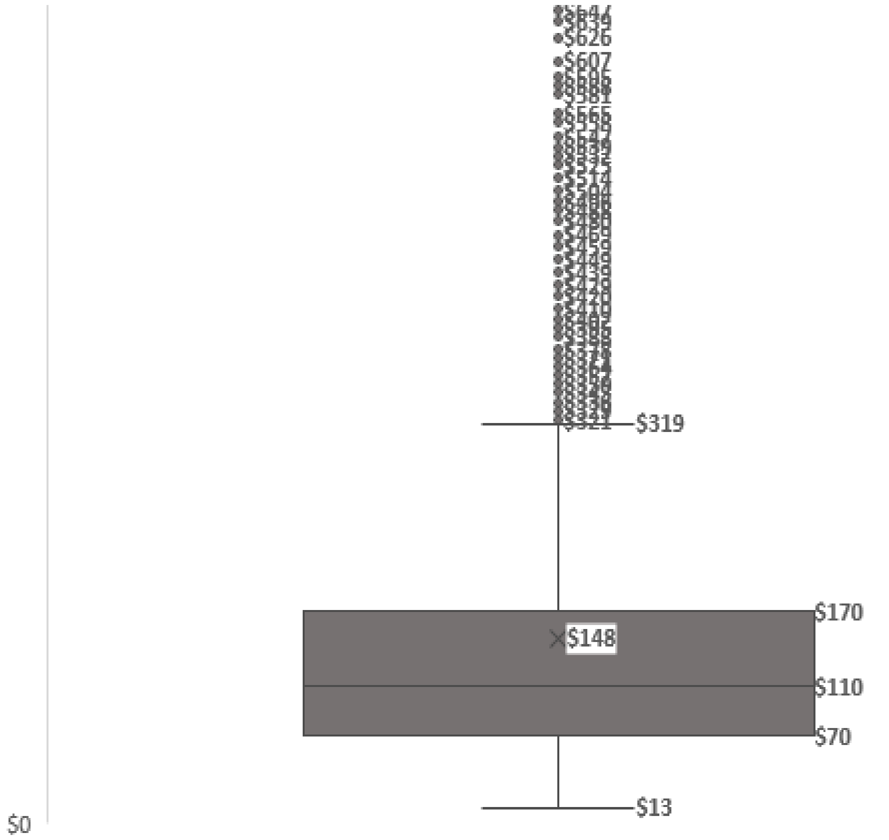




- From the chart on the left, we can observe that ‘price’ data is right skewed
- To Normalise the distribution, we take Log 10 of the values. This is represented in the chart on the right

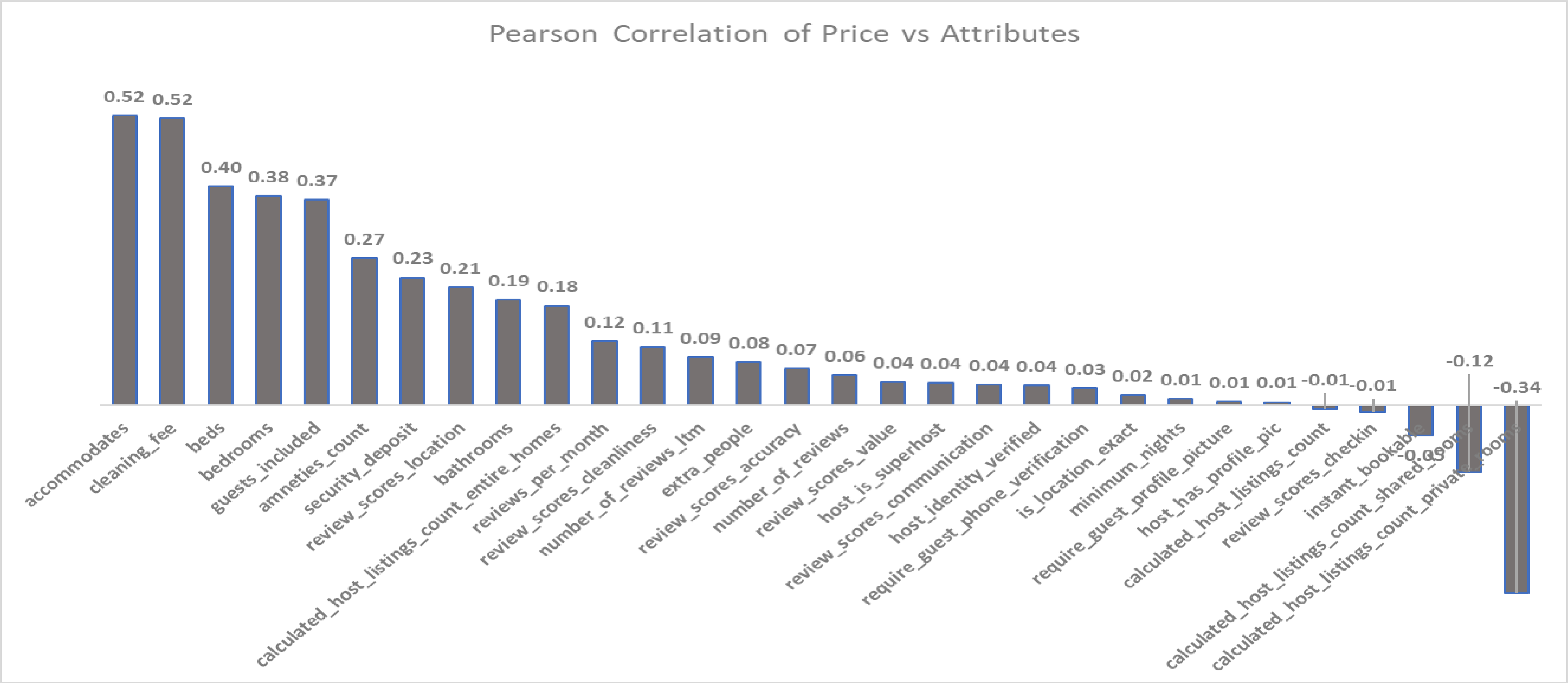
Identifying ‘Price’ outliers





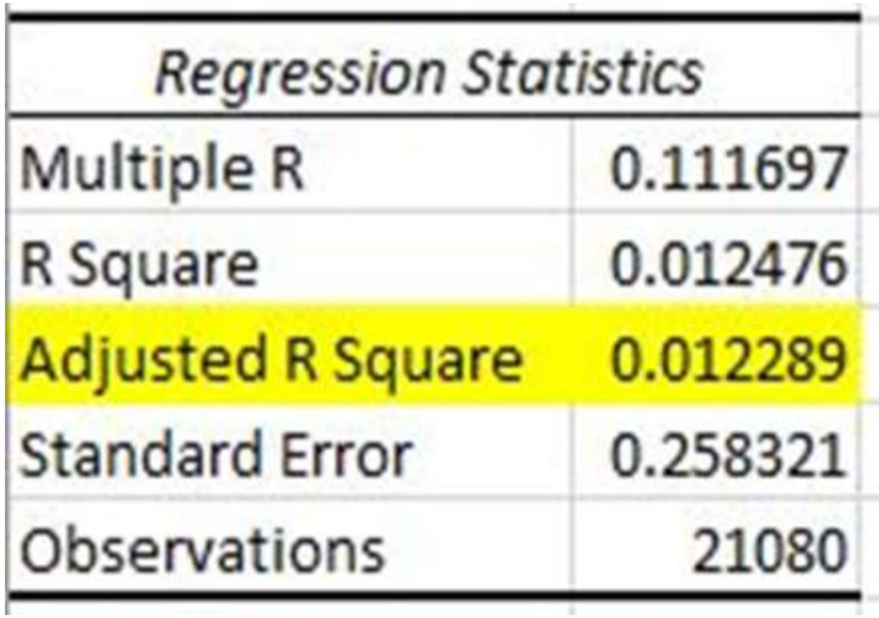
- For analysis we will not consider Listings with Price ‘0’ – Total 4 listings
- From the Box Plot, it is evident that the Upper Bound is $319
- All listings beyond this price point will be considered ‘Outliers’ and will not be included in our analysis – 1500 Listings

Pearson Correlation

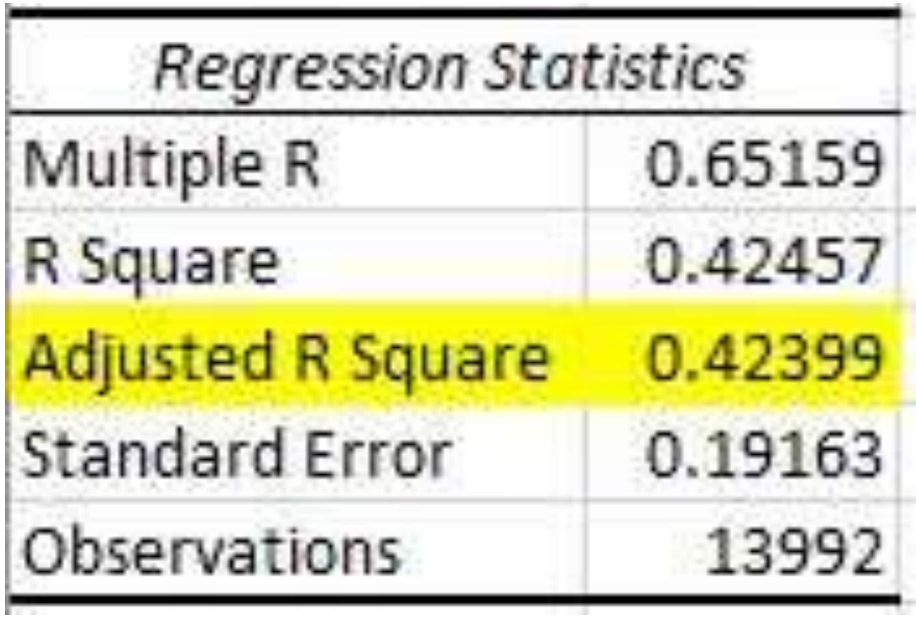

Pearson Correlation of Price vs Attributes

- There is weak to no correlation between Price and the attributes
- For analysis we will consider those attributes which show linearity

The model can now predict prices with 42.39% accuracy



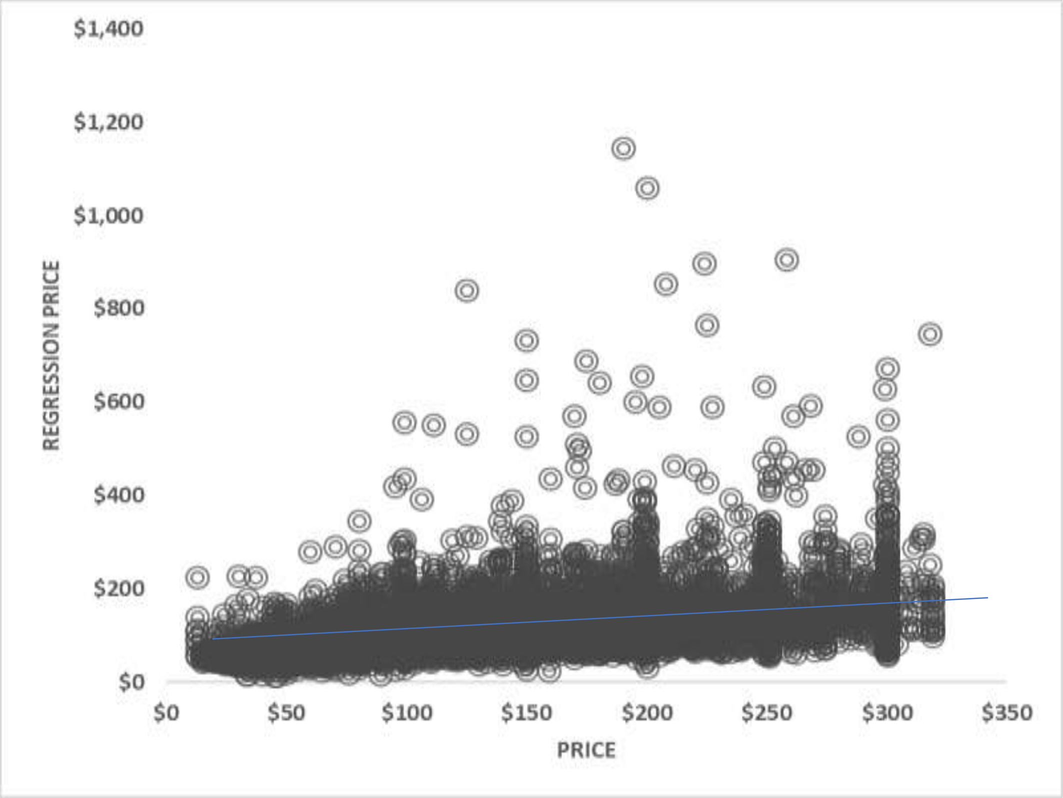
Regression Statistics	
Multiple R	0.111697
R Square	0.012476
Adjusted R Square	0.012289
Standard Error	0.258321
Observations	21080

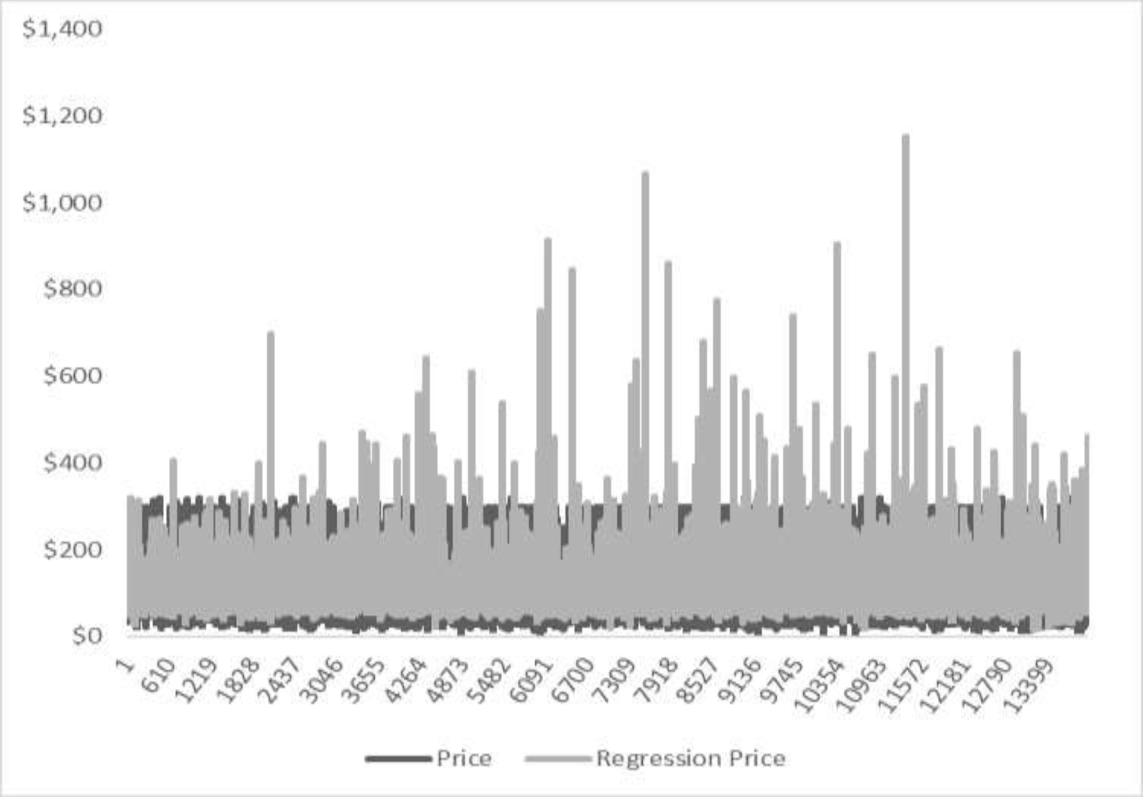


Regression Statistics	
Multiple R	0.65159
R Square	0.42457
Adjusted R Square	0.42399
Standard Error	0.19163
Observations	13992

- After running Regression Analysis on Excel, we get a model with 42.3921% accuracy
- The model is optimized by excluding statistically insignificant attributes (p>0.05) like extra_people and number_of_reviews
- accommodates, cleaning_fee and beds are the top three attributes

The model predicts mostly high values till price point $33 and range of high to low values thereafter





- The model is only 42.39% accurate
- To increase accuracy, we could do the following:

1. Use Python as a tool for Regression Analysis, as we can factor in more categorical data
2. Use more sophisticated statistical techniques, like Decision Trees, Random Forest etc.,