#Description

The task of this assignment is to predict the price of the Airbnb listings in NYC.

The first step was to examine the dataset, and that does not mean visual inspection. Using the head and dataset.type function in pandas, variables are displayed to investigate possible candidates for the prediction model. At first glance, there are 16 variables in the dataset. Excluding come obvious identifier: id, name, host\_id, host\_name, longitude, and longitude, there are 10 variables left.

From these variables, I decided to investigate if there is any missing within the data or if any recode is needed. Using function of sum and is.null I was able to get a sense of if there is any missing values we are dealing with for potential IVs. There are no missing for 8 variables. And the two variables that have missing values are reviews\_per\_month and last\_review. They have the same number of missing values and conventionally we can just get rid of the missing values. I thought about the logic behind the missing, they have missing values not because they are random missing but they are simply new listings that no one has lived/reviewed yet. I decided to recode reviews\_per\_month’s missing into zero. And since the last\_review is an time-stamped variable, I decided to drop it. Future attempt can be made to investigate review month/year’s effect relative to the impact on price.

By using the value\_counts and subsetting with index of neighbourhood\_group I identified the distribution of the location. There are five individual groups within the dataset, and we have a significant amount of listings in Brooklyn and Manhattan. I think these count numbers are ok as we have an ok number of distributions among different neighborhoods. On the other hand, there are many neighourhood so I decided to split up my experiment into 2 groups. I decided to use Lasso for neighbourhood and Ridge for neighbourhood\_group.

The price variable’s distribution was investigated by using histogram and quantiles. The 3rd quantile was 175 dollars while the maximum was at 10000. I decided to use 300 as a cutoff point for extreme values as this is approx. twice as the mean of the price variable. This makes the analysis to just focus on listings within reasonable range and we are not going to run into problems where the data on higher end has way less observations than these on the lower end. The 3357 rows of prices greater than 275 are 9.3% of the observations.

Similar operation was conducted on the minimum\_nights variable to see if there are any outliers that seems unreasonable. The max of 1250 is obviously unreasonable. I decided to use 14 days as this is approximately the twice of the mean value and also two weeks as the ceiling of this variable.

Experiment 1. Normal regression

[ 1.67713708e-07 -1.50148318e+00 1.63065194e-02 -1.87655530e+00

1.90378307e-01 1.90661910e+01 4.71183779e+01 1.06495918e+01

1.77187767e+00 -7.40311390e+01 -9.58796002e+01]

Ridge regression

[ 1.66169496e-07 -1.49483675e+00 1.60627182e-02 -1.86921470e+00

1.90375022e-01 1.84044448e+01 4.64439817e+01 9.98770153e+00

1.12525210e+00 -7.39547861e+01 -9.57559556e+01]

First experiment: Using Ridge Regression for a small number of IVs

Second experiment: Using Lasso Regression for a small number of IVs

Third experiment: Using Ridge Regression for a big number of IVs

Third experiment: Using Lasso Regression for a big number of IVs