#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. Set data as: ['minimum_nights', 'number_of_reviews', 'reviews_per_month', 'calculated_host_listings_count', 'nbhdg_Brooklyn', 'nbhdg_Manhattan', 'nbhdg_Queens', 'nbhdg_Staten Island', 'rt_Private room', 'rt_Shared room'] Notice that I have n-1 level created for the dummy.

Normal regression

R squared = 0.46

```
Coefficient = [-1.47982074e+00 -9.85665662e-03 -1.26876267e+00  1.96964059e-01

  1.71853557e+01  4.54610234e+01  9.10120386e+00  2.01873491e+00

 -7.38385328e+01 -9.54593425e+01]
```

Ridge regression

R squared = 0.46

```
Coefficient = [-1.41637198e+00 -9.51312045e-03 -1.26235083e+00  1.96311757e-01
  1.20482248e+01  4.02023493e+01  3.95028101e+00 -3.01815729e+00
 -7.30906773e+01 -9.42129144e+01]
```

Lasso

R square = 0.45

```
Coefficient = [ -0.          -0.          -0.12363113   0.11466472   0.
  27.27827173  -3.25280071  -0.          -69.55263865 -81.47983883]
```

Not much difference with a small number of variables between ridge and regular linear. I tested alpha level variation for both ridge and lasso and found smaller alpha leads to smaller difference in parameters between the regular regression. When alpha = 0.01, lasso was able to identify some parameter estimates that are suppose to be zero. The R2 across all models are similar. I should be using the adjusted R2 as the number of variable increases the R2 will always be improving. Lasso helped me to see that minimum nights, number of reviews, neighborhood in Brooklyn, and Staten Island are not that useful in predicting prices.

Experiment 2. Use a large variable pool by creating dummy from neighborhood variable. And I decided to drop the neighborhood_group. Since they are essentially describing the same thing and one neighborhood variable has better resolution.

Regular regression: R2 = 0.52

Ridge regression: R2 = 0.52

Lasso regression: R2 = 0.47

Not much information can be obtained other than a very long coefficient list. I decided to calculate adjusted R2 so that I can compare across the experiments.

X being the x_train and y being the y_train.

The equation I used is: 1 - (1-machine.score(X, y))*(len(y)-1)/(len(y)-X.shape[1]-1)

	Exp1 reg	Exp1 ridge	Exp1 Lasso	Exp2 reg	Exp2 ridge	Exp2 Lasso
R squared	0.46	0.46	0.45	0.528	0.528	0.477
Adjusted R squared	0.46	0.46	0.45	0.525	0.525	0.472

It seems that by adding neighborhood dummies we are able to explain more about the dataset. Using adjusted r2 as a performance measure, regular regression and ridge both won. More work can be done experimenting adding one variable at a time and then we will see ridge regression will win because of more stabilized parameter estimates. If in the future I have more time, I will experiment with elastic net. I think that captures the advantage of both ridge and lasso regression.

-Updated March 4th

Running with N level with N number of dummies

	Exp1 reg	Exp1 ridge	Exp1 Lasso	Exp2 reg	Exp2 ridge	Exp2 Lasso
R squared	0.46	0.46	0.45	0.528	0.528	0.477
Adjusted R squared	0.46	0.46	0.45	0.525	0.525	0.472

I obtained the same result for performance measure as expected.