Applied Data Science Capstone

Zeiad Wael Sabra 4/5/2019

Analysis of New York Airbnb Prices

Week 4 Report

Table of Content

Introduction

New York City is a huge tourist attraction visited by millions each year, making finding a place to stay a very difficult and pricey endeavour. IsPriceRight is a website that offers recommendations for prices of Airbnb. Is the place you plan on staying at overpriced?

Head to RightPrice to check if the price is right or if you are being scammed.

Business Plan

RightPrice wants us to make a model to advise tourists visiting New York City on the optimal price for a place to stay in New York city. A tourist provides us with Information about the place and we provid him/her with the optimal price using our model. Our goal is to build a model that give an estimate of the rent of a place in New York City using available data.

The Desired outcomes are:

- A model for calculating rental prices.
- A description of the most relivant features of the model.
- Cluster the Neigbourhoods based on the Rent, Venues, and location.

Data Selection

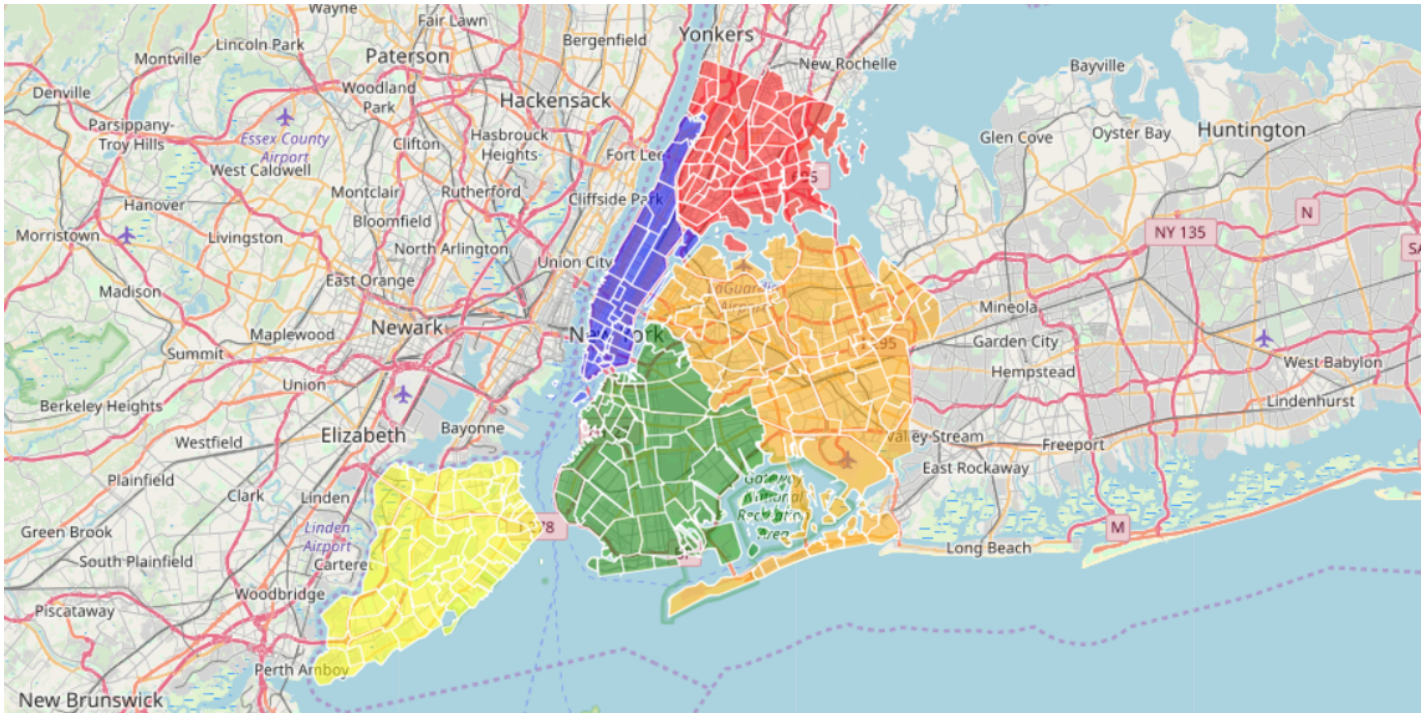
1. Geospatial Data

- The data consists of Neigbourhoods names, Boroughs names, Neigbourhood bountries and some other columns.

Here are the first 5 rows of the geospatial data

	neighborhood	boroughCode	borough	@id	geometry
0	Allerton	2	Bronx	http://nyc.pediacities.com/Resource/Neighborho...	POLYGON ((-73.84859700000018 40.87167000000012...
1	Alley Pond Park	4	Queens	http://nyc.pediacities.com/Resource/Neighborho...	POLYGON ((-73.74333268196389 40.7388830992604,...
2	Arden Heights	5	Staten Island	http://nyc.pediacities.com/Resource/Neighborho...	POLYGON ((-74.169827 40.56107800000017, -74.16...
3	Arlington	5	Staten Island	http://nyc.pediacities.com/Resource/Neighborho...	POLYGON ((-74.15974815874296 40.64141652579018...
4	Arrochar	5	Staten Island	http://nyc.pediacities.com/Resource/Neighborho...	POLYGON ((-74.06077989345394 40.59318800468343...

We also use Folim to plot the boundries of each Neighbourhood assigning then colors by their borough



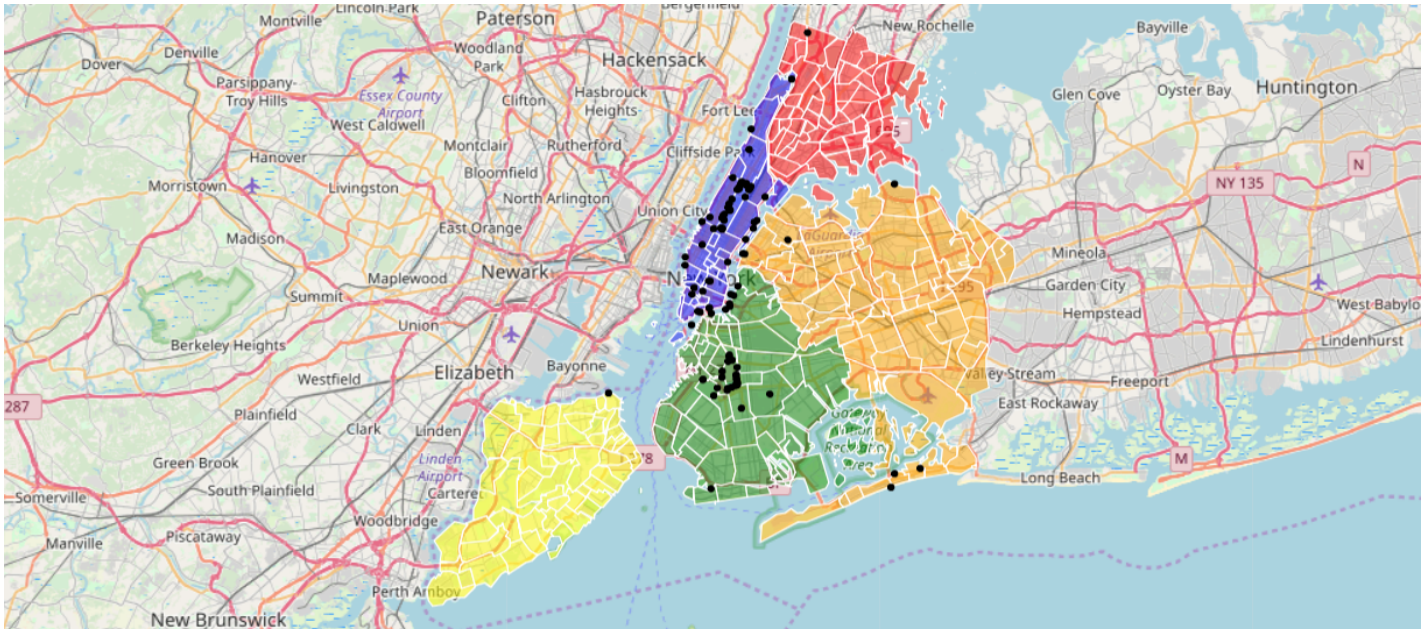
2. Airbnb Rental data

Here are the first 5 rows of the data after cleaning and removing unwanted columns

	address	name	price	review_scores_location	latitude	longitude	bedrooms	room_type	bathrooms	property_type	guests_included
0	Central Park, Manhattan	Charming Studio - Central Park	150.0	10.0	40.781561	-73.971238	0.0	Entire home/apt	1.0	Apartment	1
1	NaN	Rockaway Bungalow by the Bay	60.0	9.0	40.591061	-73.814242	1.0	Private room	1.0	Apartment	1
2	Central Park, Manhattan	Cozy Mexican Inspired Private Room	97.0	10.0	40.779410	-73.969830	1.0	Private room	1.0	Apartment	1
3	Prospect Park, Brooklyn	Modern 1BD with exposed brick	100.0	10.0	40.655026	-73.962212	1.0	Entire home/apt	1.0	Apartment	1
4	NaN	Manhattan Cozy 1BR Apartment $60	60.0	9.0	40.873336	-73.911239	1.0	Entire home/apt	1.0	Apartment	2

The data also contains price per night, latitude, longitude and many other features

Here are the first 100 Places in the data



3. Foursquare API

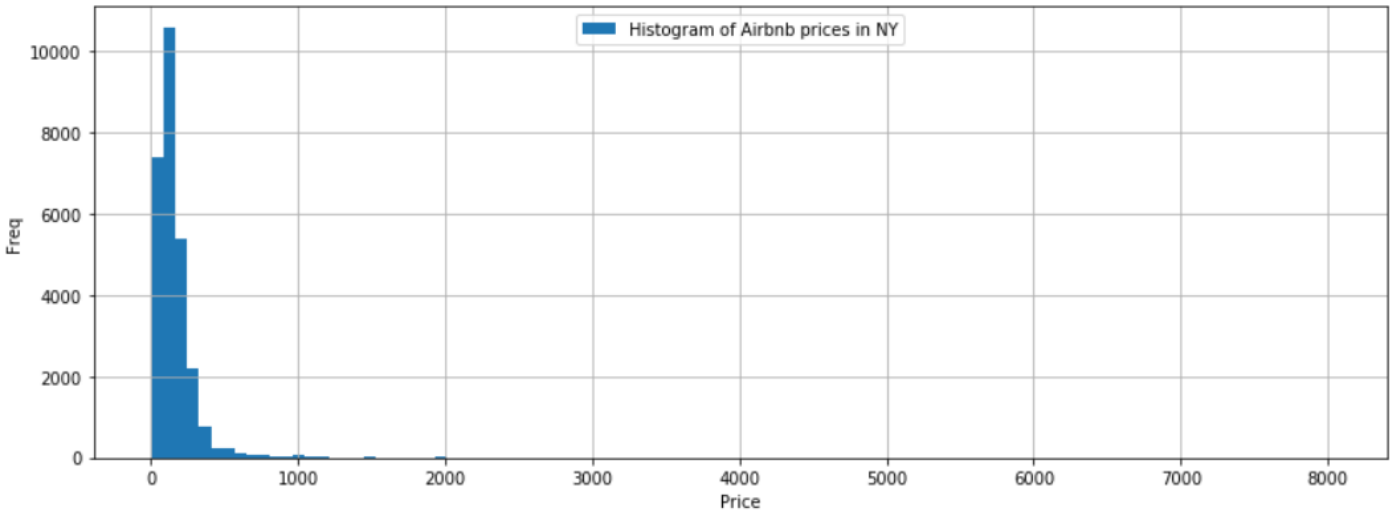
- We are going to use the Foursquare API to explore the nearby venues available around each listing of the Airbnb dataset and see how they affect the price of the listing.

Here are the list of venues near the first Place listed in the data

	Place	Place Latitude	Place Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Charming Studio - Central Park	40.781561	-73.971238	American Museum of Natural History	40.781282	-73.973238	Science Museum
1	Charming Studio - Central Park	40.781561	-73.971238	Hayden Planetarium	40.781718	-73.973239	Planetarium
2	Charming Studio - Central Park	40.781561	-73.971238	American Museum of Natural History Museum Shop	40.780973	-73.973028	Souvenir Shop
3	Charming Studio - Central Park	40.781561	-73.971238	Rose Center for Earth and Space	40.781741	-73.973127	Planetarium
4	Charming Studio - Central Park	40.781561	-73.971238	Shakespeare Garden	40.779755	-73.969976	Garden

Methodology

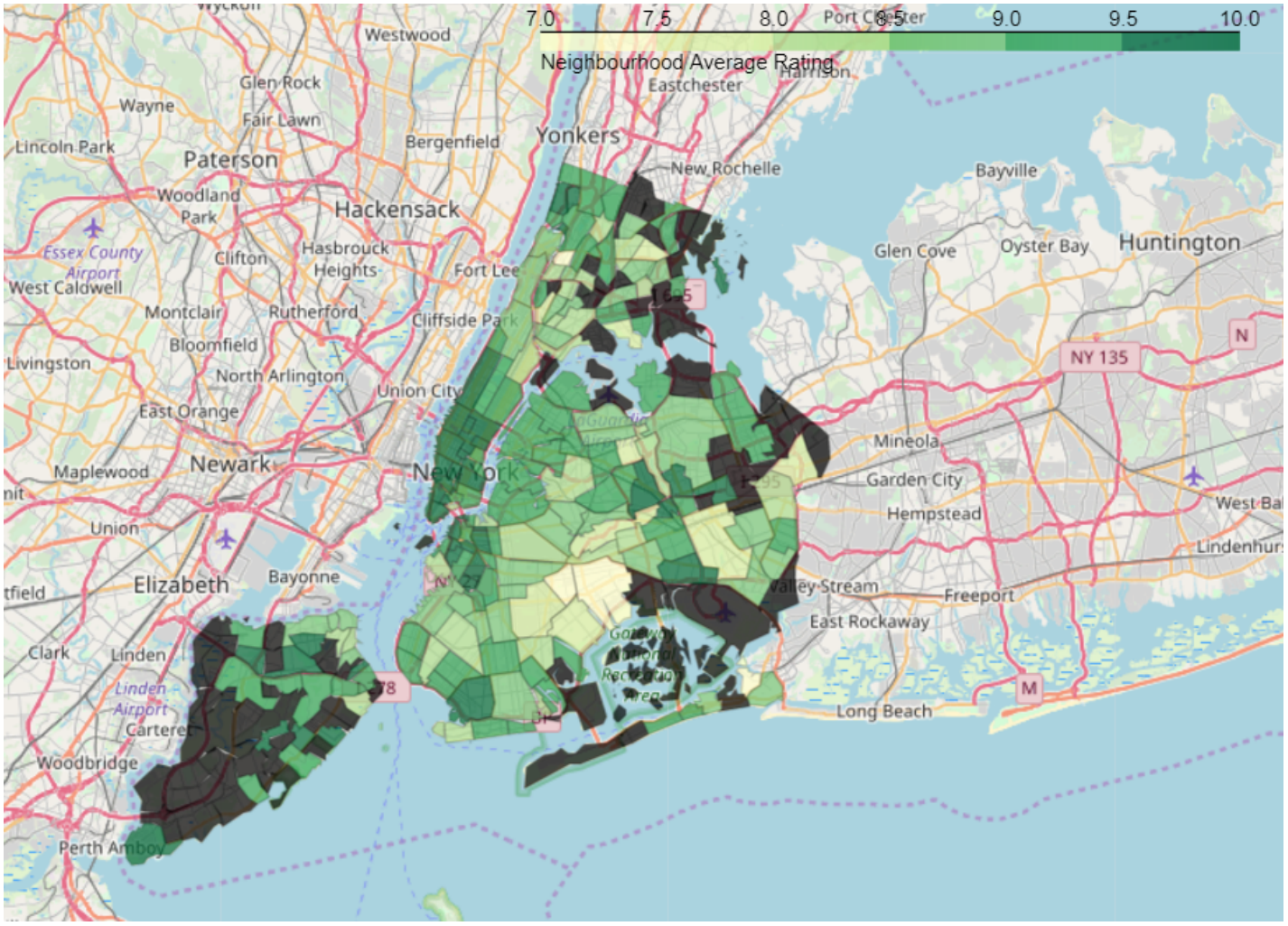
We start by looking at the prices, we immediately see that there are some outliers that skiew the prices which we need to handle



Our goal is predict the prices of Airbnb one night prices in New York, so we look the neighbourhoods of New York and see how are they thought of.

We look the average review_score_location for each Neighbour hood and we can see that Manhattan is ranked highest.

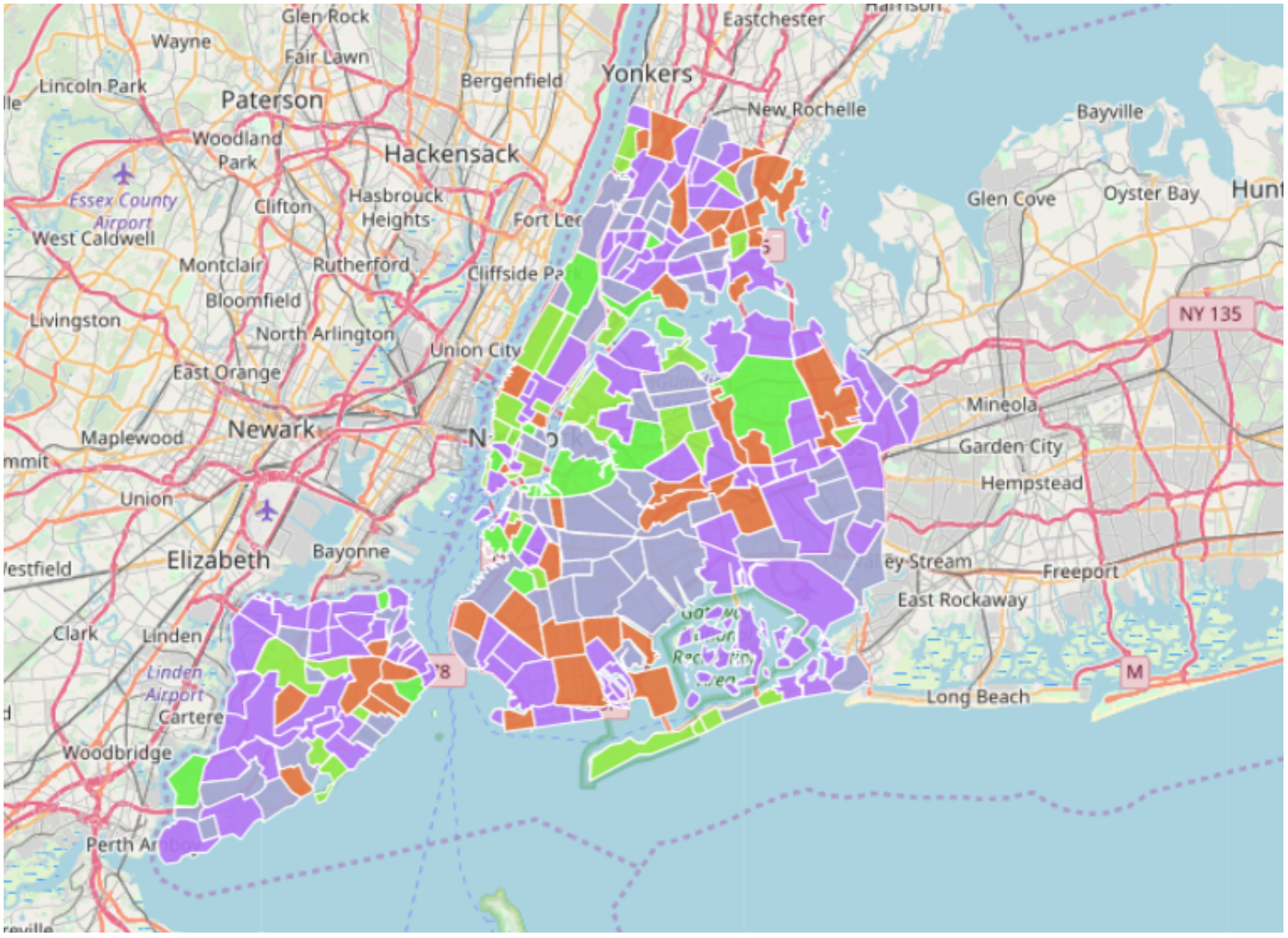
The Dark Areas are those without score in the data.



We then look at the Foursquare data and see the top venues for each neighbourhood.

	neighbourhood	ATM	Accessories Store	Acupuncturist	Adult Boutique	Adult Education Center	Advertising Agency	Afghan Restaurant	African Restaurant	Airport	...	Well	Whisky Bar	Wine Bar	Wine Shop	Winery
0	Allerton, Bronx	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0
1	Alley Pond Park, Queens	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0
2	Arden Heights, Staten Island	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0
3	Arlington, Staten Island	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0
4	Arrochar, Staten Island	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0

We then cluster the neighbourhoods using Kmeans into 5 clusters



Neighbourhoods next each other are clustered similary, however, one drawback, I have noticed, is that the similarity is really unfair, for example, the distance between a chinese restaurant and an indian restautant is the same as the distance between an office and a chinese restauarant. We could lump all restaurants together. but then we would lose the distinction.

We, then, combine the datasets of airbnb listings and Foursquare data so it can be ready for the regression algorithms. After combining the datasets, we split them into training, validation, and testing dataset.

```
X_t_v, X_test, y_t_v, y_test = train_test_split(X, y, test_size=0.2)
```

Split the Traing/validation data into traing and validation

```
X_train, X_val, y_train, y_val = train_test_split(X_t_v, y_t_v, test_size=0.33)
```

We start training our model using Linear Regression, however, model show clear

evidence of overfitting as the validation mean squared error is extremely large.

```
lm = LinearRegression()
lm.fit(X_train, y_train)
```

```
LinearRegression(copy_X=True, fit_intercept=True, n_jobs=None,
         normalize=False)
```

```
lm_mse_train = mean_squared_error(lm.predict(X_train), y_train)
print("Training Error: ", lm_mse_train)
```

```
Training Error:  28956.189363203102
```

```
lm_mse_val = mean_squared_error(lm.predict(X_val), y_val)
print("Validation Error: ", lm_mse_val)
```

```
Validation Error:  2.565207319572144e+20
```

We then look at the option of regularization using L1 Lasso Regression, we start of with L1

```
lss = Lasso()
lss.fit(X_train, y_train)
```

```
Lasso(alpha=1.0, copy_X=True, fit_intercept=True, max_iter=1000,
   normalize=False, positive=False, precompute=False, random_state=None,
   selection='cyclic', tol=0.0001, warm_start=False)
```

```
lss_mse_train = mean_squared_error(lss.predict(X_train), y_train)
print("Training Error: ", lss_mse_train)
```

```
Training Error:  30757.109686446474
```

```
lss_mse_val = mean_squared_error(lss.predict(X_val), y_val)
print("Validation Errot: ", lss_mse_val)
```

```
Validation Errot:  37797.071490769726
```

The validation error is much better than the unregulized linear regression.

Feature Selection

Lasso regression is a great method for feature selection as it zeros all but the most relevant features.

```
coef = pd.Series(lss.coef_, index = X.columns)
coef = coef[abs(coef) != 0]
coef
```

```
review_scores_location        16.663473
bedrooms                      67.694707
bathrooms                    100.748614
guests_included               -5.210580
cluster_0                    -34.035317
cluster_2                    -22.423239
cluster_4                     14.294990
room_type_Entire home/apt     98.683553
property_type_House          -11.154660
property_type_Loft             4.339545
dtype: float64
```

```
lss.intercept_
```

```
-209.93575034902625
```

Conclusion

Here we have found that the number of bathrooms and type of residence are the most important features for predicting the price

Next

We should work on a better clustering algorithms that uses a better distance metric.

Sources

We used data.beta.nyc as the source of our data.

The data for the Rental Data was download from
data.insideairbnb.com/united-states/ny/new-york-city/2015-05-01/data/listings.csv.gz

The data for the geolocations and boundries of New York's neighbourhoods was downloaded from:
data.beta.nyc/dataset/pediacities-nyc-neighborhoods

The venues data was aquired using Foursquare API at foursquare.com