# Housing Sales Prices & Venues Data Analysis of New York

Vishnudev K

February-6-2020

# Introduction

## Background

New York is one of the largest metropolises in the world where over 9 million people live, and it has a population density of 10,194 people per square kilometer. The city is divided many neighborhoods. As part of this project I am analyzing the variation of price of properties based on the facilities nearby.

As we all know price of a property depends on direct factors such as area, features, type of property etc. But in this analysis, we are studying the properties dependency on nearby presence of Collages, Shopping malls, Hospitals etc.

## Problem

The price of properties varies lot based on factors that we might not be considering or its hard to be quantifiable. As part of this project we are doing a predictive analysis based on the location data.

# Data acquisition and cleaning

## Data sources

I found one Property sales data set from New York city large enough to do predictive analysis in Kaggle (https://www.kaggle.com/new-york-city/nyc-property-sales)

It contains data around 80k property transactions.

Neatly arranged Price, Area, type of property, zip, Address etc. information etc.

This data is missing the neighborhood information. Which we will get it from venue API of foursquare.

## Data cleaning

Data downloaded or scraped from multiple sources were combined into one table. There were a lot of missing values. because of lack of record keeping. There are many places where even the depending variable – Price is missing. I have decided to remove these records because it makes no sense on keeping them.

I later removed few rows having extreme outlying price/area. Which can be have a negative impact on the predictive analysis.

## Feature selection

After data cleaning, there were 20,000 samples and 21 features in the data. Upon examining the meaning of each feature, it was clear that there was some redundancy in the features and some features that has not impact on the outcome.

I have removed the columns such as row number, Apartment number, Building number etc.

Some features seem to have obvious impact on the property price later proved no impact on the predictive analysis.

Given the plot of relation of Age of the property with respect to price of the property.

According to the Sea Born plots. The price is scattered around and there is no relation with age of property. The correction value was close to 0.

So, we need to ignore that field.

Based on the location data we added 12 more columns to the original data set based on the zip code’s nearest places.

We choose the number of following types of venues from FourSquare venues API.

Category id is in brackets,

Arts & Entertainment (4d4b7104d754a06370d81259)

College & University (4d4b7105d754a06372d81259)

Food (4d4b7105d754a06374d81259)

Professional & Other Places (4d4b7105d754a06375d81259)

Office (4bf58dd8d48988d124941735)

Residence (4e67e38e036454776db1fb3a)

Shop & Service (4d4b7105d754a06378d81259)

Shopping Mall (4bf58dd8d48988d1fd941735)

Travel & Transport (4d4b7105d754a06379d81259)

Airport (4bf58dd8d48988d1ed931735)

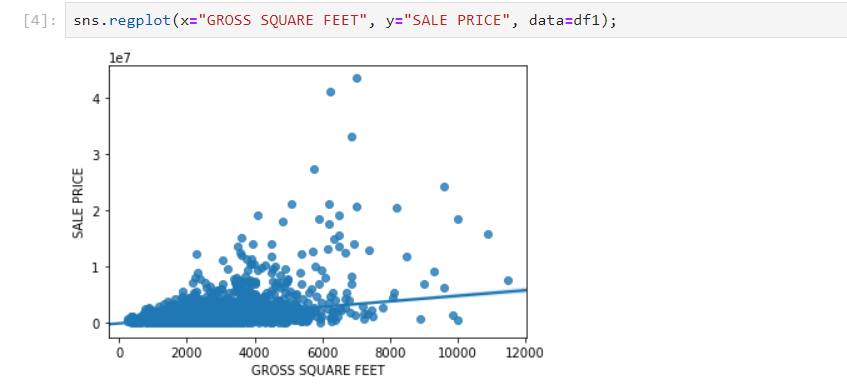
Bus Station (4bf58dd8d48988d1fe931735)

Train Station (4bf58dd8d48988d129951735)

# Exploratory Data Analysis

## Relationship between price vs area

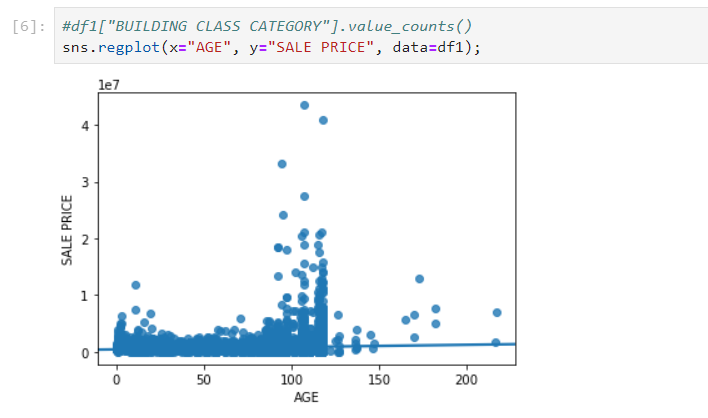
Price of area with sales price is plotted below. As expected, we can conclude a positive correlation on price with square feet area.



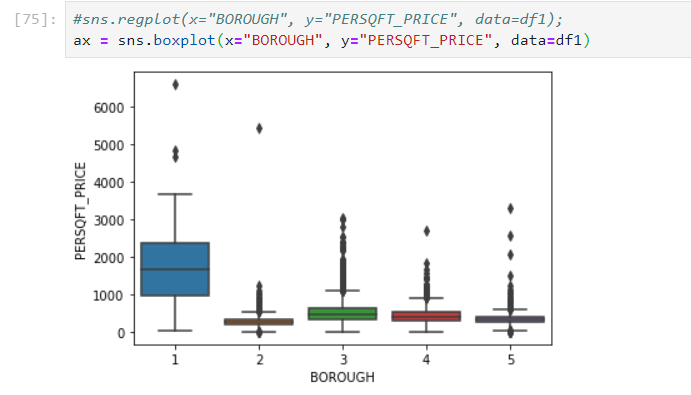
## Relationship between age of the property with price

I had an intuition that sales price will drop as the age of the property increases.

But plotting the regression plot proved there is no relationship between age of the property with price.



## Relationship between various borough with rate per SqFt.



# Predictive Modeling

There are two types of models, regression and classification, that can be used to predict price of a property. Regression models can provide additional information on the amount of property. The underlying algorithms are similar between regression and classification models, but different

audience might prefer one over the other. We are choosing regression model here to predict the price of the property.

## Regression models

* + 1. **Applying standard algorithms and their problems**

I applied linear models, linear regression and deep neural network gradient descent models to the dataset, using mean squared error (MSE) as the tuning and evaluation metric.

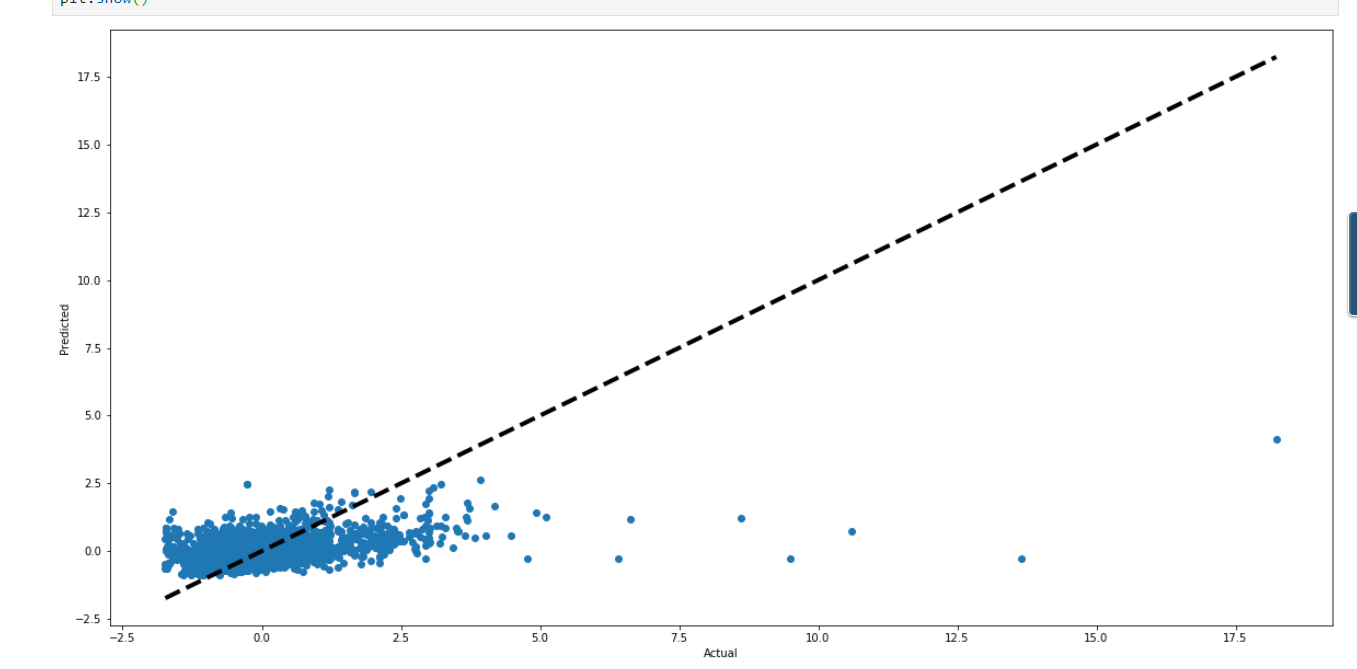
We are applying the same algorithms and evaluate the performance on original data set and enriched data set.

Finally, the prediction is compared to check if enriched data has any impact on the dataset.

## Analysis on original data

Upon applying the rules, predictive model on original data we got a MSE error of 0.73 and 0.72 on Linear and Deep Neural Network models

Following is the plot of actual value vs predicted value for cross validation sets.

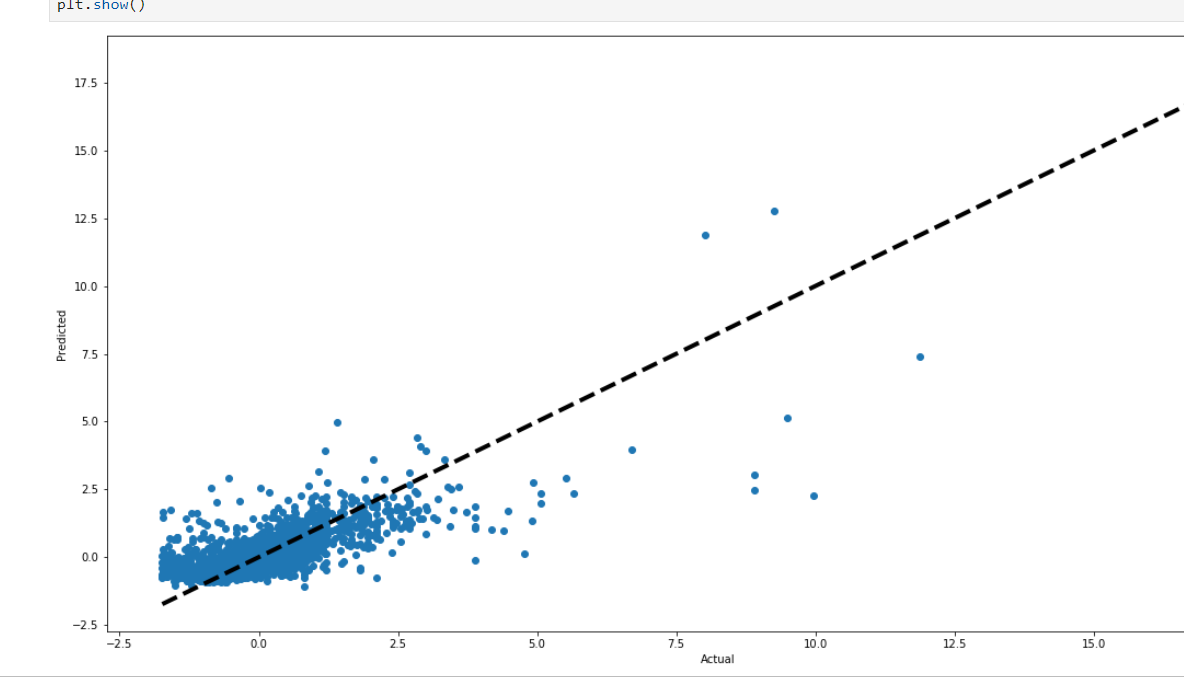


## Analysis on enriched data.

Based on predictive analysis we could come up with a better model having a better MSE error.

Linear and Deep Neural Network models gave a lower error of 0.64 and 0.46 resp.

The plot of actual price vs predicted price is as given below.



# Conclusions

In this study, I analyzed the relationship between property pricing with neighborhood data.

We can conclude that neighborhood has a +ve correlation with the pricing of a property.

There was a **35%** improvement on the accuracy of the price prediction by using neighborhood data.

# Future directions

Even though we could improve on the accuracy of our prediction model (**35% improvement**) with neighborhood data, our predictive model was still not having a better overall prediction accuracy. We can improve the prediction accuracy by taking following steps

1. Getting address specific neighborhood data instead of zip-based neighborhood data. Since there is a free limitation of Foursquare API per day, I’d to rely on zip code to get the data. With an enterprise account its easy to get data for each address in a smaller area of search.
2. Modify tweak the neighborhood data. Currently we selected 12 venues randomly based on which prediction model is formed. We will have to increase the venue categories and do a comprehensive analysis on which all fields contribute more to the price of the property.
3. More data. We had mere 20k data to train which go cut to 10k transaction when we did a grouping with zip. If we can get hands on a full data set. The model will have better accuracy.
4. In this analysis we found that increasing the number of hidden layer parameter had a positive impact on the prediction. But this complexity increase is making the system to train slowly. We can tweak these parameters along with optimum learning rate to improve on the accuracy of prediction