CYOS_Capstone

Esther Sienkiewicz

27/12/2021

Introduction

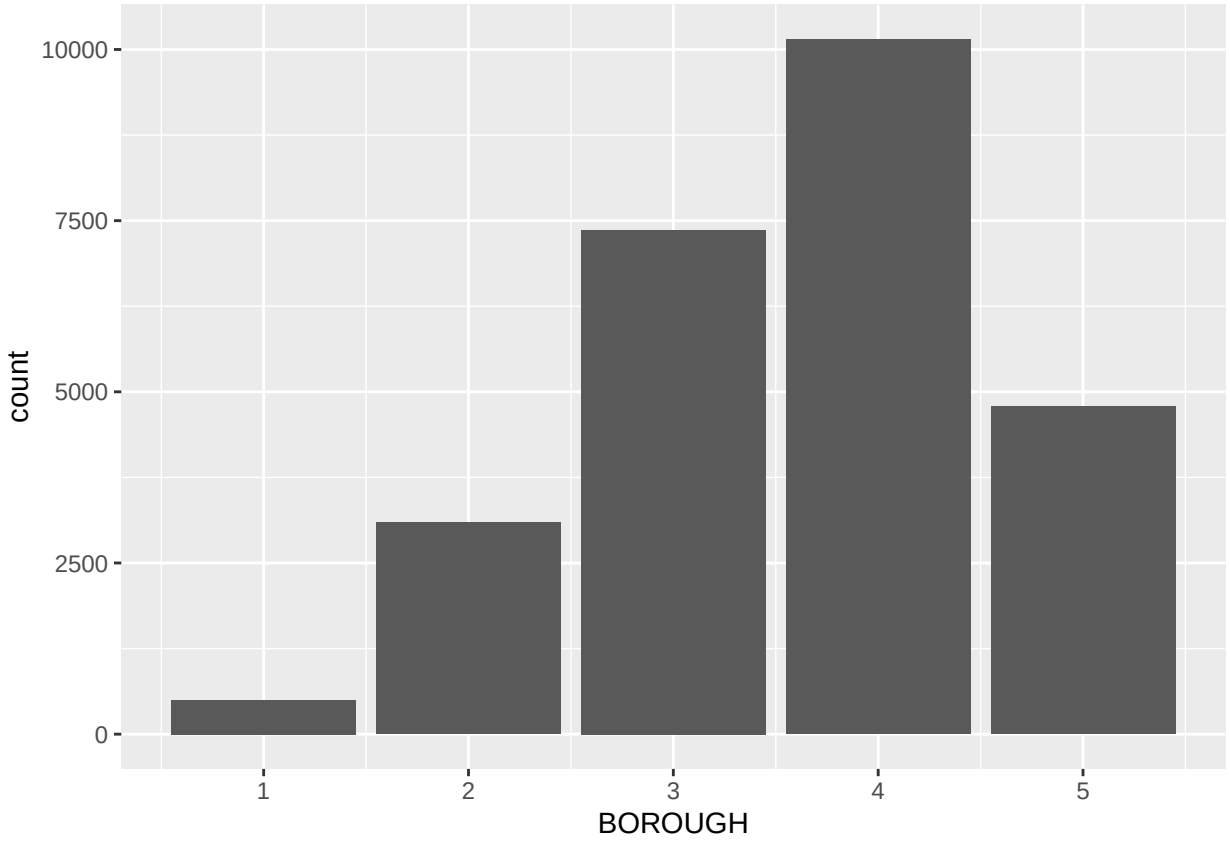
With the latest in news reports about rising interest rates and how this pandemic has caused for many to look into upgrading their dwellings the decision to look further into a data set relating to housing seemed timely. The data set chosen is a record of every building or building unit (apartment, condo etc) sold in the New York City property market over a 12-month period.
https://www.kaggle.com/new-york-city/nyc-property-sales

With this data set, the analysis focused on whether there was a way to find any predictions with pricing, as mentioned earlier there has been a lot of focus in the recent news about housing and affordability.

Analysis

The first curiosity that the data brings is where is the most sales happening. Since the data is broken down by location of boroughs using the code: Manhattan (1), Bronx (2), Brooklyn (3), Queens (4), and Staten Island (5). What's the guess? As Manhattan seems the most desirable or well known location one would have thought that there would be more sales of homes in this area, turns out Queens is the more up and coming area.

It was surprising!



Data Cleaning

When looking at the column data there seems to be a collection of both commercial and residential units in the data set.

```
colnames(df)
```

```
##  [1] "X"                 "BOROUGH"           "LOT"
##  [4] "RESIDENTIAL.UNITS" "COMMERCIAL.UNITS"  "TOTAL.UNITS"
##  [7] "LAND.SQUARE.FEET"  "GROSS.SQUARE.FEET" "YEAR.BUILT"
## [10] "SALE.PRICE"        "SALE.DATE"
```

Since it's difficult to really compare the commercial and residential properties together it's better to separate them out and focus the analysis on only residential properties.

```
df <- filter(df, COMMERCIAL.UNITS == 0)
df <- filter(df, RESIDENTIAL.UNITS > 0)
```

Next, the sales prices seem to have some " - " in the cell meaning that this was not a sell but a transfer of deeds between parties. So to help make the values recognizable the " - " was replaced with "0". But the problem still persists that this is not a sale and could alter the overall statistics of the data. Original number of entries in the data set was 84548, after only filtering to have residential entries only this left only 57049 entries. Next, removing the "0" sale price and those entries with gross square feet of "-" gave a grand total of 38796 entries.

It looks like still many sales of nonsensical prices of 10 dollars, so question becomes what is the threshold that should be placed on the value amount to make sense? What is the square footage that should be considered nonsensical? Is 10 square feet a realistic residential unit? Thresholds chosen where such that only units of sales price greater than 100 dollars and residential units greater than 40 square feet. That would lead to a total of 25894 entries.

Model 1

Further data cleaning was the removal of all categories which would not be necessary in order to achieve the goal to understand the relationship between size (GROSS.SQUARE.FEET) of the residential unit and the sale price. Building a machine learning algorithm that predicts the sale price of a residential unit Y using the gross square foot of the unit X. It is then needed to generate testing and training sets:

```
set.seed(1)
index <- createDataPartition(y = df$SALE.PRICE, p = 0.5, list = FALSE)
test <- df[index,]
train <- df[-index,]
```

It's suspected that the two variables are of normal distribution, the condition expectation is equivalent to the regression line stated:

$$f(x) = E(Y|X = x)$$

Therefore we only need to find the slope and intercept in order to fulfill the conditional expectation:

$$\hat{f}(x) = 778242.05 + 62.04x$$

An interesting phenomenon happens when looking at the portioning of the training and testing data. If the portion of the test data is too large (or training set too small) it becomes a higher root mean squared error. In this case it was noticed the best results given us the p = 0.5 to give the correct balance. Additonally as the data is very skewed towards small sale price (as seen in the figure) there is difficulty to have too small or too large of the train, test data. This would either not given enough data entries for training or testing would be inaccurate. It was observed using root mean squared error $(\hat{y} - y)^2$ is indeed improved compared to the simple guessing approach.

```
m <- mean(train$SALE.PRICE)
sqrt(mean((m - test$SALE.PRICE)^2))
```

```
## [1] 2007829
```

```
lm_model1 <- lm(sale.price.test ~ GROSS.SQUARE.FEET, data = train)
y_hat <- predict(lm_model1, test)
sqrt(mean((y_hat - test$SALE.PRICE)^2))
```

```
## [1] 1912946
```

The RMSE has improved in predicting the residential home sale price compared to guessing. There's more information in the data set that could help, it's known that location is one key aspect in real estate. If we can consider X as the variable of the gross square feet, and Y as the variable to represent the borough.

$$f(x) = E(Y|X = x, Y = y)$$

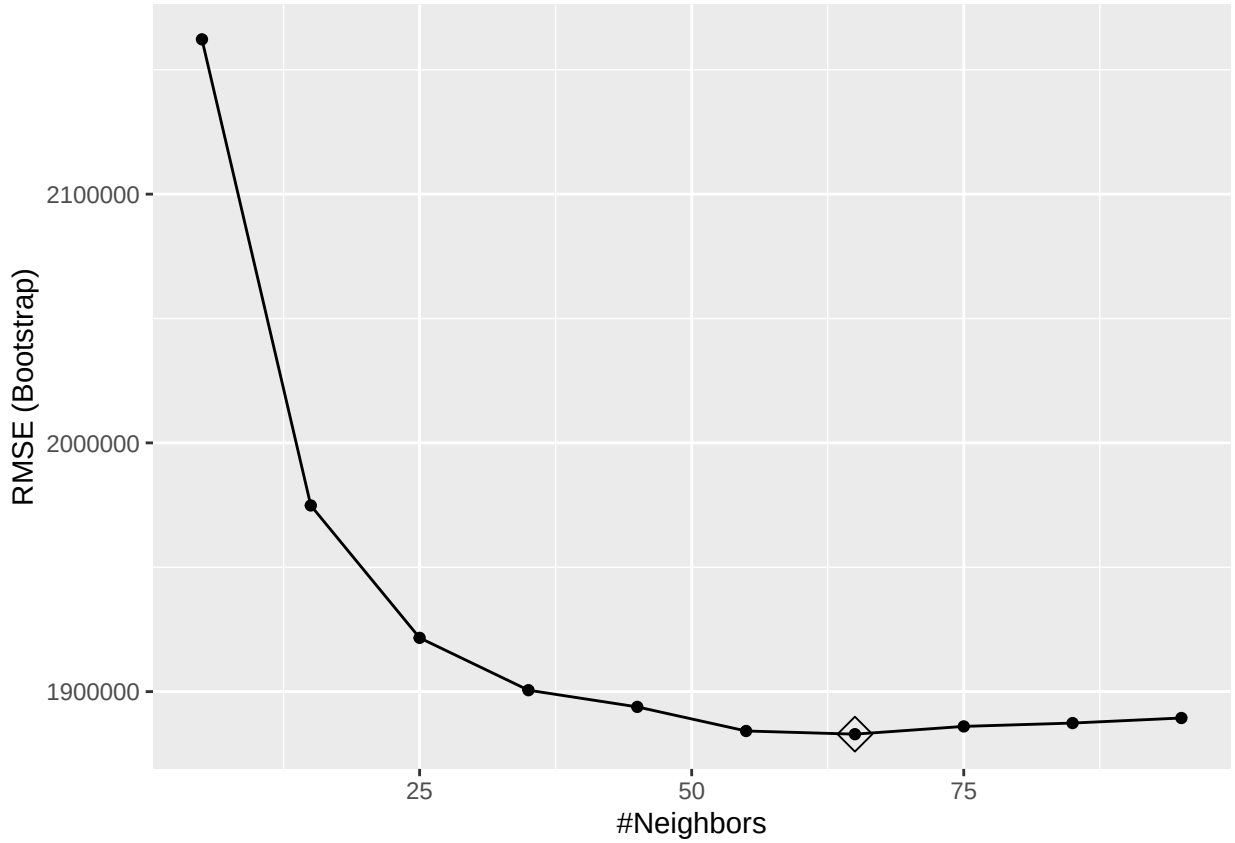
```
lm_model2 <- lm(sale.price.test ~ GROSS.SQUARE.FEET + BOROUGH, data = train)
y_hat <- predict(lm_model2, test)
sqrt(mean((y_hat - test$SALE.PRICE)^2))
```

```
## [1] 1871887
```

Model 2

In this model, we use the kNN as a second model to help further reduce the RMSE error. Since our initial data was very clustered together it seemed appropriate to apply the K-Nearest Neighbors to further the prediction of residential sales price estimated from the gross square footage.

First, some cross validation is applied to find the closest neighor which provides the optimum results. We initially cast a wide net with $k = 5, 15, 25...100$ from there it was determined that when $k = 65$ is the best performing model, visualized below.



Now to understand the performance of the model with the evaluation of the RMSE:

```
y_hat <- predict(train_knn, test)
sqrt(mean((y_hat - test$SALE.PRICE)^2))
```

```
## [1] 1510303
```

Conclusion

This report has shown the performance between linear regression and prediction outcomes and the nearest neighbor algorithm. The following comparisons between models can be made:

Model	RMSE
Baseline	2007829
Linear regression (sqt ft)	1912946
Linear regression (sqt ft + borough)	1871887
kNN	1510303

Perhaps the obvious limitations of these models is predicting the future prices with only historical data, without any other environmental factors such as inflation or zoning. Making such predictions in a vacuum can be limiting. However, it's still interesting to know that with the best outcome (kNN) the average sales price expected to pay in New York City is $943250.8

Since the RMSE is also so large, it can be noted that perhaps the dataset (n) is not large enough to have any effect, or have a correlation between the square footage, and location to the price is not enough.