# Humann_Edx_Capstone_CYO

James Humann

April 30, 2019

# Introduction

This report analyses the New York City Property Sales dataset available from (https://www.kaggle.com/new-york-city/nyc-property-sales (https://www.kaggle.com/new-york-city/nyc-property-sales)), which contains over 200,000 records of property sales within New York City in the year 2017. The data set contains records of the sale price, square footage, borough, neighborhood, lot, tax class, and building class.

We analyze this data set in an effort to predict selling prices. This can be a very beneficial prediction, as property is inherently unique, and thus difficult to value objectively. Because of the immense prices (sales were routinely in the millions of dollars), this can be the most expensive purchase that a family makes in their lifetime, and thus overpaying even by a few percent can be disastrous.

To do our analysis, we perform data cleaning, then exploratory analysis to identify trends. Finally we split our data into a training set (90% of data) and test set (10% of data), so that we can train a linear model on the training set and evaluate its performance in the test set.

We focus our data analysis on predicting Price per Square Foot (PPSQFT), so that the results can be independent of property size, which is obviously a large determining factor in the final price. The final price can always be determined from the PPSQFT since the square footage of the property will be known before sale.

# Data Cleaning

The original data contains many nonsensical sales, such as $0 or $100 sales. It also includes many rows with missing data for the square footage or price. We remove all of these since our anlysis will be based on price per square foot. We also decide to filter out sales below $100,000 since it is not possible to buy a family home or business for these prices in New York City. The orginal and cleaned data sets are shown here.

```
#original
head(df)
```

```
## # A tibble: 6 x 22
##      X1 BOROUGH NEIGHBORHOOD `BUILDING CLASS~ `TAX CLASS AT P~ BLOCK   LOT
##   <dbl>   <dbl> <chr>        <chr>            <chr>            <dbl> <dbl>
## 1     4       1 ALPHABET CI~ 07 RENTALS - WA~ 2A                 392     6
## 2     5       1 ALPHABET CI~ 07 RENTALS - WA~ 2                  399    26
## 3     6       1 ALPHABET CI~ 07 RENTALS - WA~ 2                  399    39
## 4     7       1 ALPHABET CI~ 07 RENTALS - WA~ 2B                 402    21
## 5     8       1 ALPHABET CI~ 07 RENTALS - WA~ 2A                 404    55
## 6     9       1 ALPHABET CI~ 07 RENTALS - WA~ 2                  405    16
## # ... with 15 more variables: `EASE-MENT` <lgl>, `BUILDING CLASS AT
## #   PRESENT` <chr>, ADDRESS <chr>, `APARTMENT NUMBER` <chr>, `ZIP
## #   CODE` <dbl>, `RESIDENTIAL UNITS` <dbl>, `COMMERCIAL UNITS` <dbl>,
## #   `TOTAL UNITS` <dbl>, `LAND SQUARE FEET` <chr>, `GROSS SQUARE
## #   FEET` <chr>, `YEAR BUILT` <dbl>, `TAX CLASS AT TIME OF SALE` <dbl>,
## #   `BUILDING CLASS AT TIME OF SALE` <chr>, `SALE PRICE` <chr>, `SALE
## #   DATE` <dttm>
```

```
#cleaned
head(buildingSales)
```

```
## # A tibble: 6 x 8
##   BOROUGH NEIGHBORHOOD  CLASS   LAND_SQFT GROSS_SQFT  PRICE PPSQFT LOT_SIZE
##   <fct>   <fct>         <fct>       <dbl>      <dbl>  <dbl>  <dbl> <fct>
## 1 1       ALPHABET CITY 07 REN~      1633       6440 6.62e6  1029. Small
## 2 1       ALPHABET CITY 07 REN~      2272       6794 3.94e6   579. Medium
## 3 1       ALPHABET CITY 07 REN~      2369       4615 8.00e6  1733. Medium
## 4 1       ALPHABET CITY 07 REN~      1750       4226 3.19e6   756. Small
## 5 1       ALPHABET CITY 08 REN~      4489      18523 1.62e7   876. Large
## 6 1       ALPHABET CITY 08 REN~      3717      12350 1.03e7   838. Large
```

# Analysis

Here we perform exploratory analysis and then settle on a linear model to predict PPSQFT.

**Exploratory Analysis**

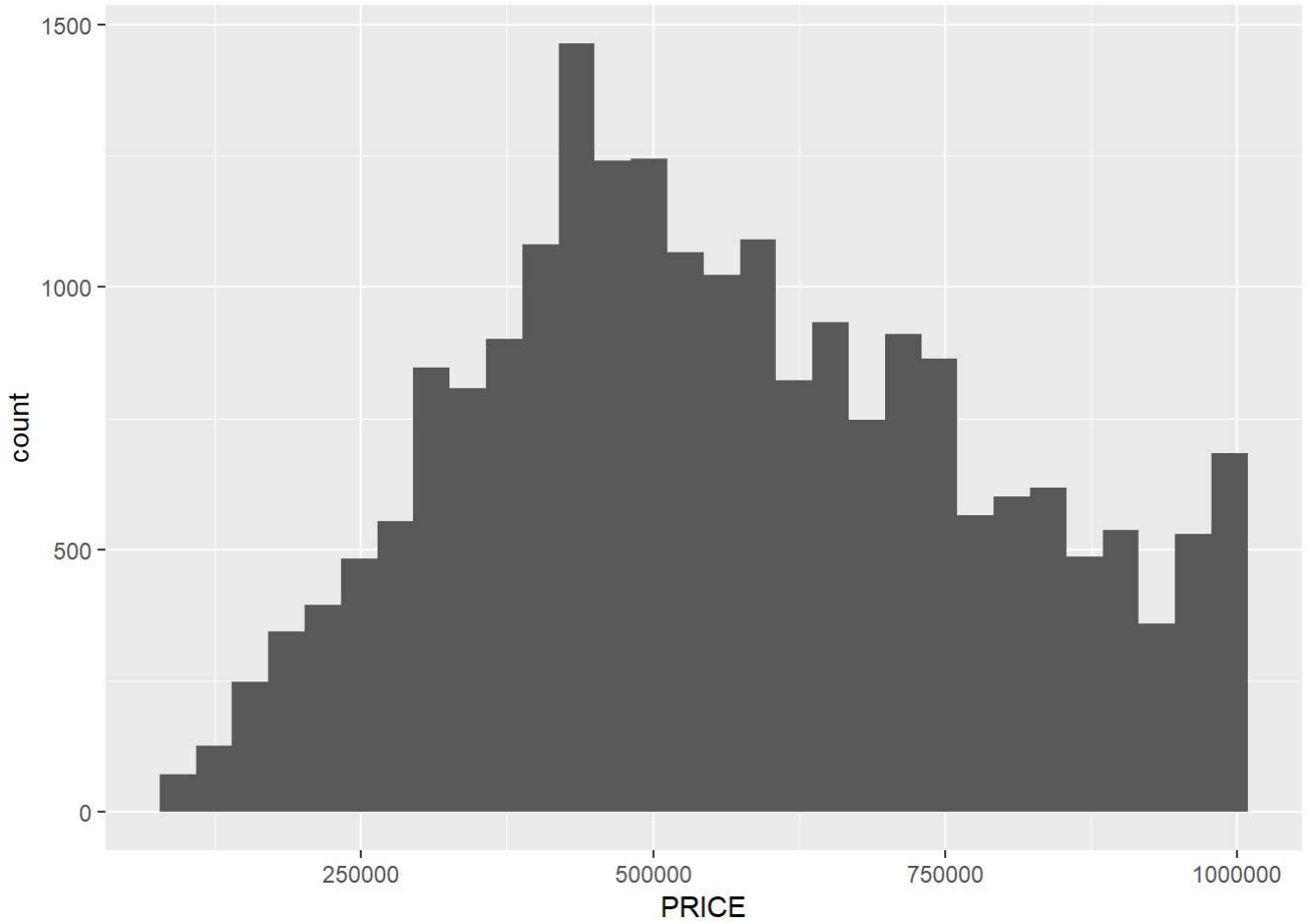
The distribution of sales prices is quite wide, making histograms hard to read. Here we calculate the maximum sale price and show a histogram of prices below $1,000,000:

```
max(buildingSales$PRICE)
```

```
## [1] 2.21e+09
```

```
price_hist_mil
```

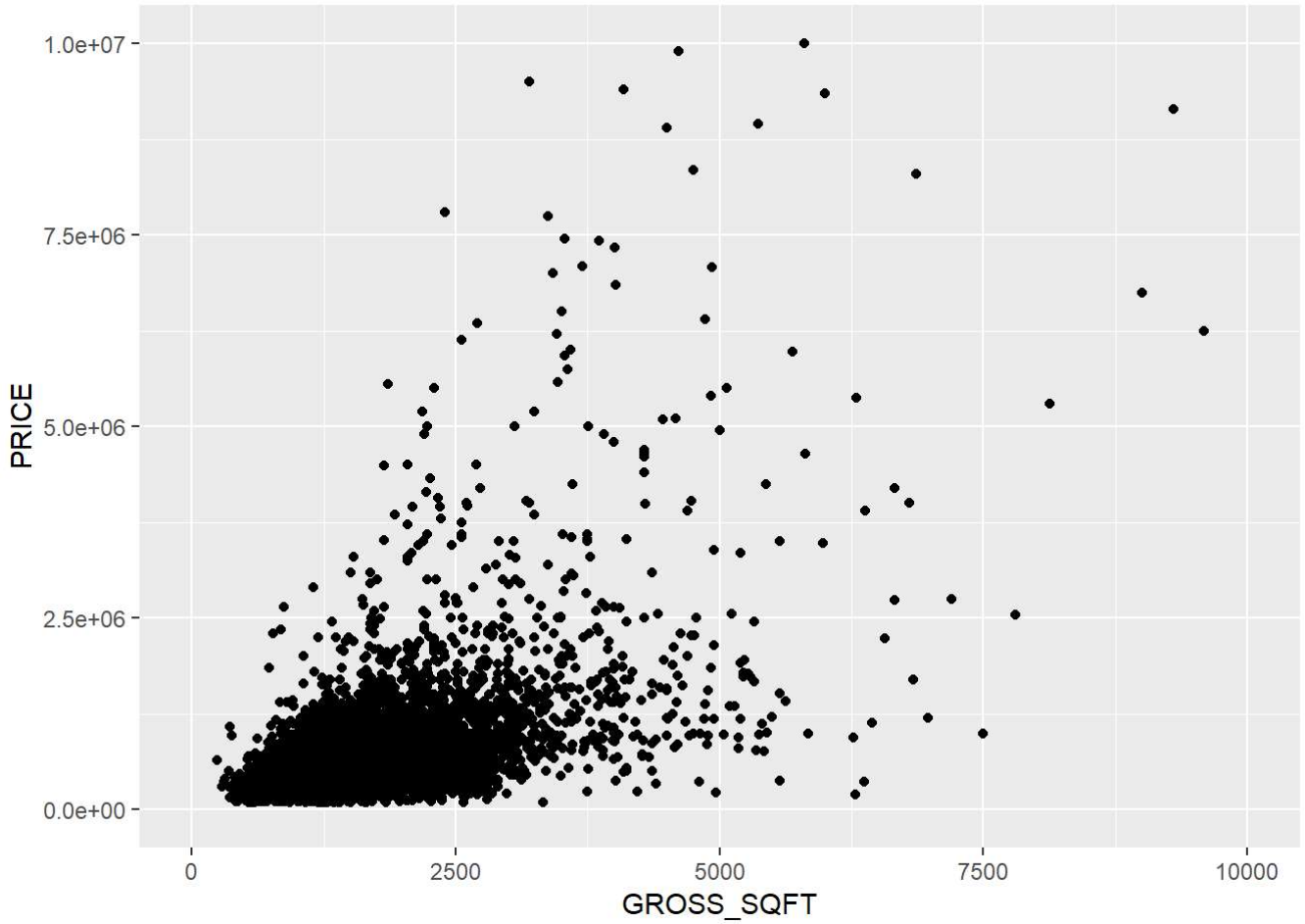
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




We test our assumption that we can focus on PPSQFT instead of the overall price by plotting them against each other for single-family homes below $10,000,000. We see a somewhat linear trend that is nonetheless very noisy so will need advanced machine learning techniques to be stratified by borough, neighborhood, and/or building class.

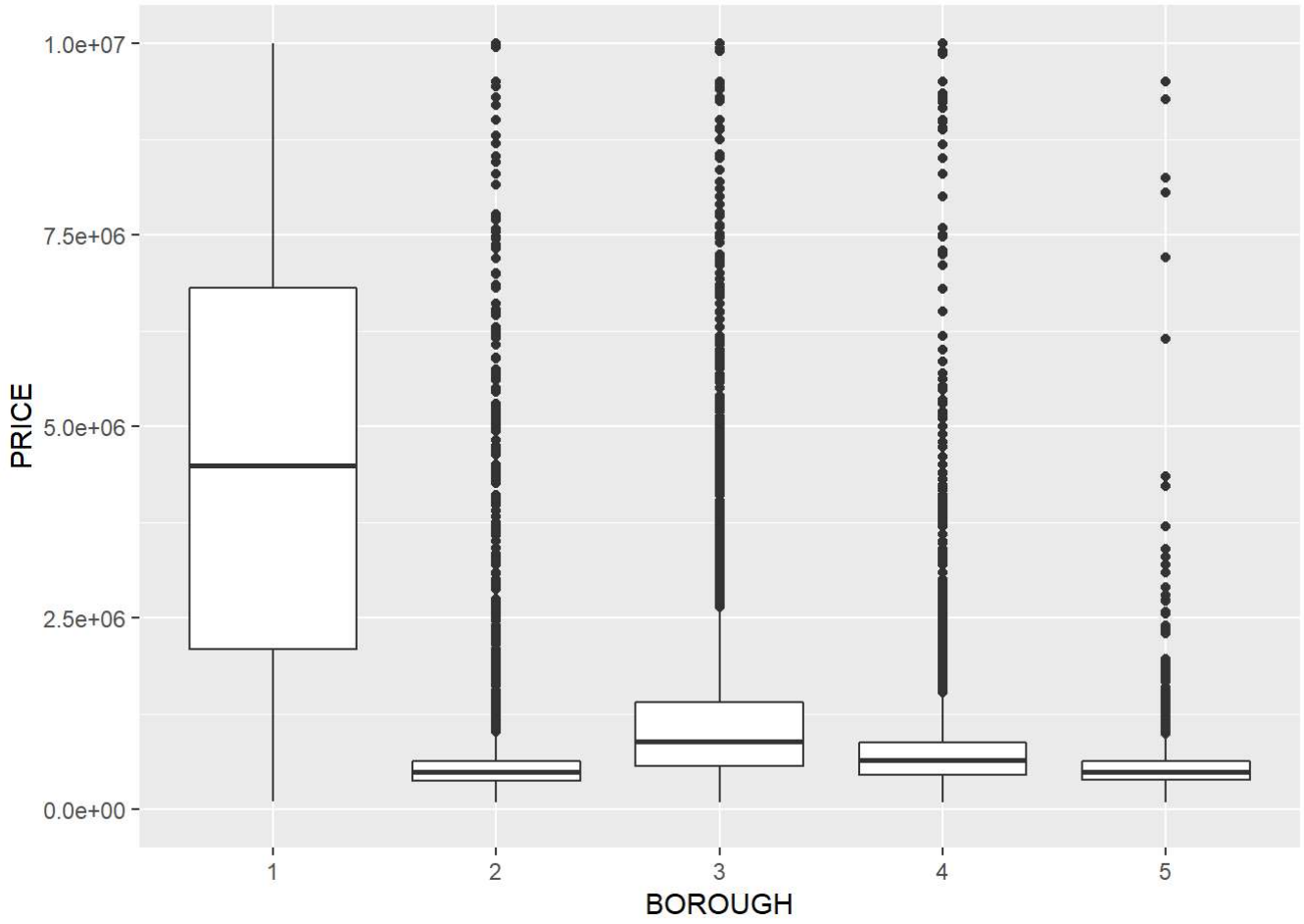
```
single_fam_sqft
```

```
## Warning: Removed 27 rows containing missing values (geom_point).
```

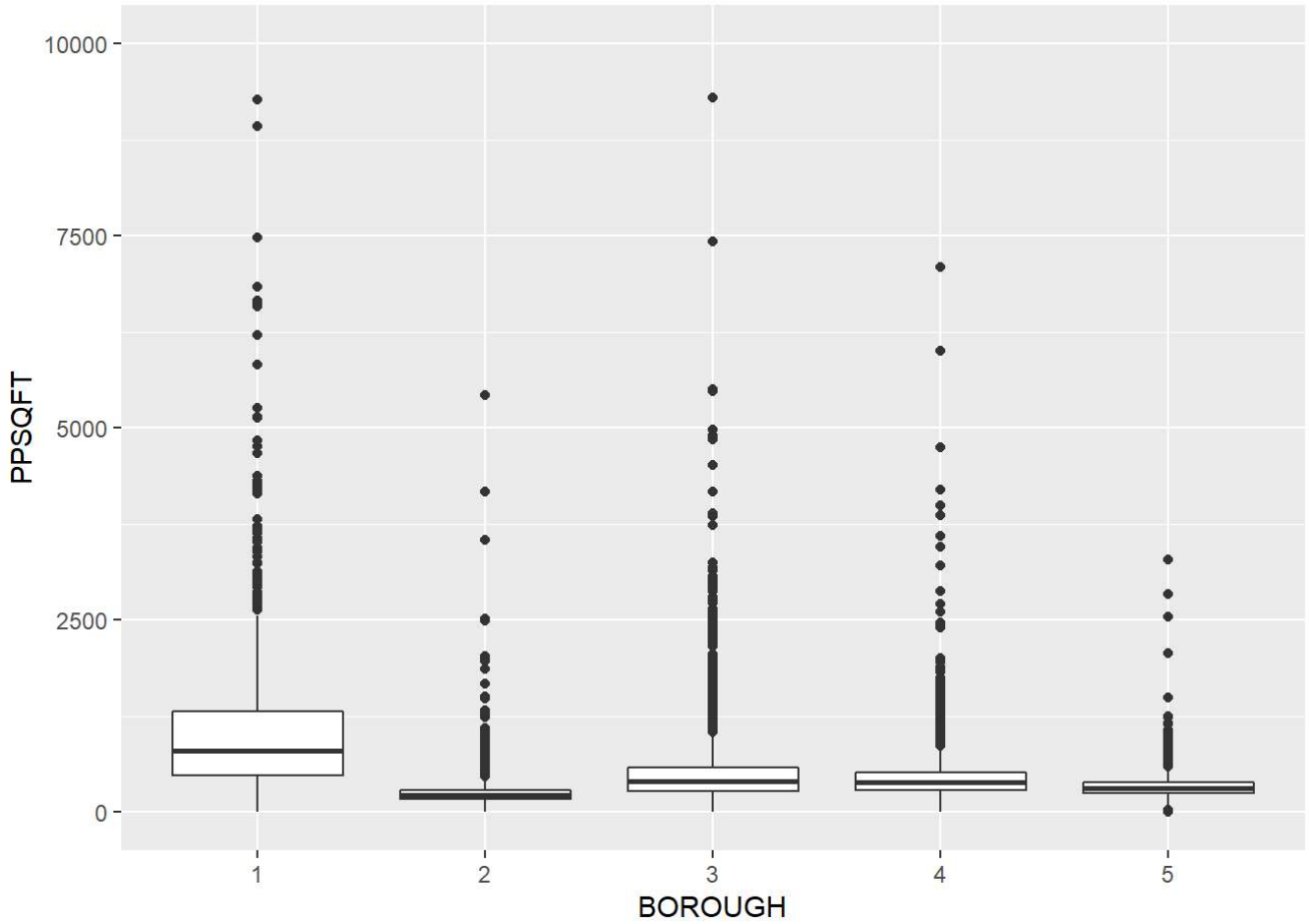



The following box plots show that borough and neighborhood (only top 10 by number of sales are shown) are good stratifiers for predicing price and price per square foot:

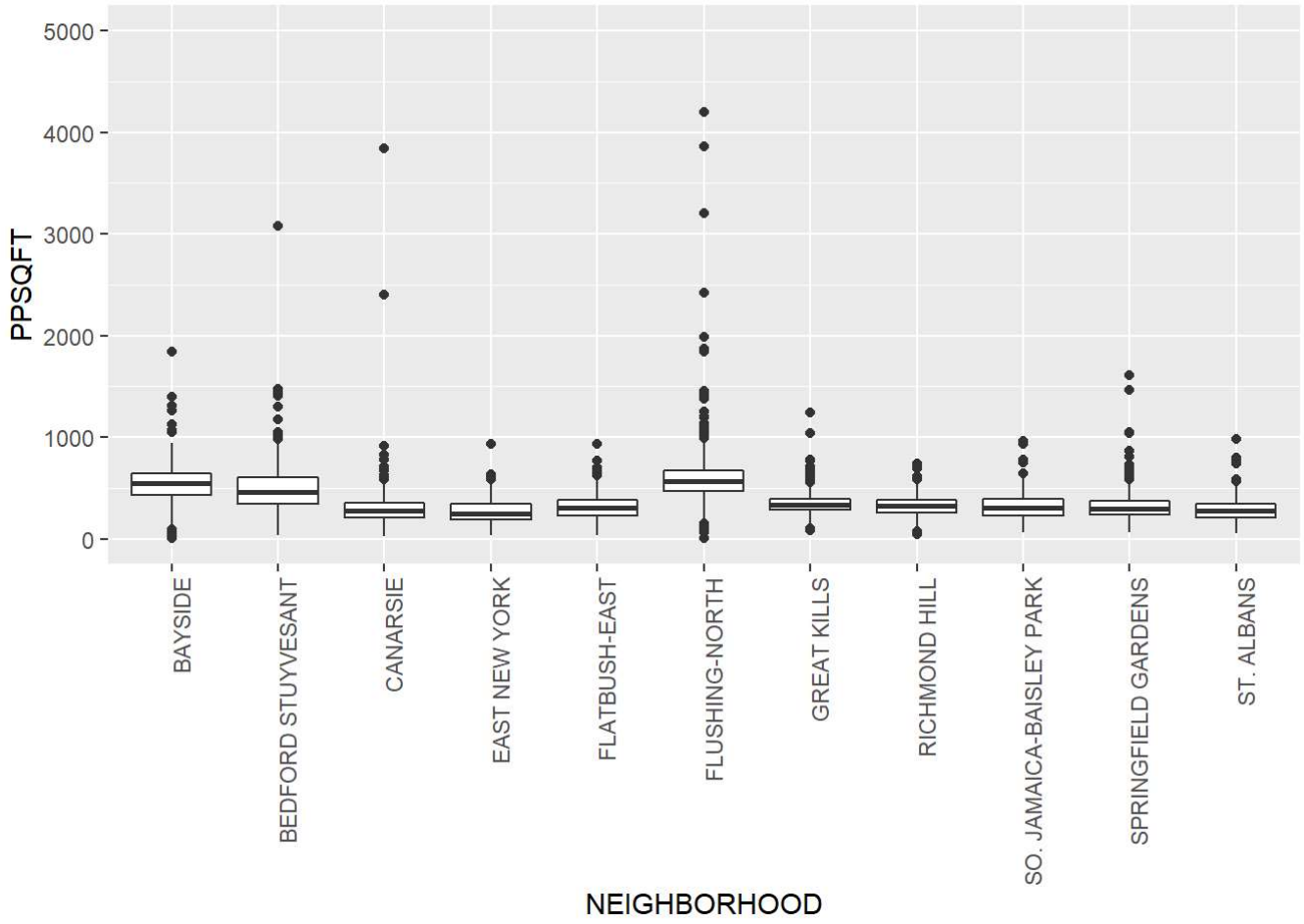
```
## Warning: Removed 538 rows containing non-finite values (stat_boxplot).
```




```
## Warning: Removed 11 rows containing non-finite values (stat_boxplot).
```




```
## Warning: Removed 1 rows containing non-finite values (stat_boxplot).
```




**Linear Model**

From our exploratory analysis, it appears that a linear model based on the boroughs and neighborhoods can give us predictive power, so we train a linear model to predict PPSQFT based on these two variables. We also attempted to refine the predictions using building class or lot size, but surprisingly these actually hindered the performance of our system.

# Results

The following RMSE values show the performance of the four linear models we tested:

```
results
```

```
##                                 fit              err
## 1                          BOROUGH 397.889966687128
## 2             BOROUGH + NEIGHBORHOOD 343.579086302918
## 3 BOROUGH + NEIGHBORHOOD + LOT_SIZE 344.556112752319
## 4    BOROUGH + NEIGHBORHOOD + CLASS 352.872712254953
```

We choose the linear model based on neighborhood and borough, as it gives the lowest RMSE of 343.579086302918. This RMSE corresponds to an error of about 18.5358864 in the price per square foot.

# Conclusion

In our analysis, simpler is better, as the most accurate linear model to predict PPSQFT used on the borough and neighborhood of the sale. Future work may be able to refine these predictions not with a linear model, but with a nearest-neighbors algorithm, which would find the average sales price of similar properties. This is akin to what home-buyers and agents currently do manually, comparing “comps,” or the sales price of comparable properties.

Buying and selling property, especially in global cities such as New York, will continue to be a major investment that could make or break a family or business’s finances. Using advanced machine learning tools to ensure that one is getting a fair price should become standard practice in the modern market.