New York City Property Sale Price Prediction

Cyrus Cai

5/16/2020

1 Introduction

With the assumption that historical data contains inights for the future, we attempt to predict human behavior. The research is done in various areas, such as retail, movies, polls etc.

1.1 The Challenge

In this project, I will build a regression model for property sale. The model will predict the sale price for an exsiting property in New York City. I will only look at the properties which is between 50 thousand and 5 millions and is one family dwelling.

1.2 Dataset: Properties Sold in New York City

I will use **nyc-rolling-sales** as my dataset. The data set is generated by New York City Department of Finance. You can find and download the daata through the link [https://www.kaggle.com/new-york-city/nyc-property-sales].

```
##     X1 BOROUGH           NEIGHBORHOOD BUILDING.CLASS.CATEGORY
## 1 4176       1 GREENWICH VILLAGE-WEST 01 ONE FAMILY DWELLINGS
## 2 4177       1 GREENWICH VILLAGE-WEST 01 ONE FAMILY DWELLINGS
## 3 4804       1         HARLEM-CENTRAL 01 ONE FAMILY DWELLINGS
## 4 4805       1         HARLEM-CENTRAL 01 ONE FAMILY DWELLINGS
## 5 4808       1         HARLEM-CENTRAL 01 ONE FAMILY DWELLINGS
##   TAX.CLASS.AT.PRESENT BLOCK LOT EASE.MENT BUILDING.CLASS.AT.PRESENT
## 1                    1   585  69        NA                        A5
## 2                    1   585  69        NA                        A5
## 3                    1  1942  58        NA                        A4
## 4                    1  1960  41        NA                        A9
## 5                    1  2024  50        NA                        A5
##                ADDRESS APARTMENT.NUMBER ZIP.CODE RESIDENTIAL.UNITS
## 1         2 GROVE COURT            <NA>    10014                 1
## 2         2 GROVE COURT            <NA>    10014                 1
## 3   288 W. 137TH STREET            <NA>    10030                 1
## 4   307 WEST 136 STREET            <NA>    10030                 1
## 5 238 WEST 139TH STREET            <NA>    10030                 1
##   COMMERCIAL.UNITS TOTAL.UNITS LAND.SQUARE.FEET GROSS.SQUARE.FEET YEAR.BUILT
## 1                0           1              384              1152       1901
## 2                0           1              384              1152       1901
## 3                0           1             1549              3036       1910
```

```
## 4                0           1             1665             3200        1910
## 5                0           1             1699             3620        1910
##   TAX.CLASS.AT.TIME.OF.SALE BUILDING.CLASS.AT.TIME.OF.SALE SALE.PRICE
## 1                         1                             A5    1375000
## 2                         1                             A5    1375000
## 3                         1                             A4    2300000
## 4                         1                             A9    1510000
## 5                         1                             A5    3050000
##    SALE.DATE   Prop_ID Building.Class
## 1 2016-10-07  1_585_69              A
## 2 2016-10-07  1_585_69              A
## 3 2016-11-30 1_1942_58              A
## 4 2017-01-03 1_1960_41              A
## 5 2017-01-31 1_2024_50              A
```

1.3 Goal

I will find a consice and yet accurate model. The project is aimed at adjusted R square of 50% and RMSE of 100K. The project is also aimed to produce an easy-to-follow and reproducible report.

2 Analysis

Before modeling, I create two dataset: **edx** and **validation**. The **validation** is 10% of the full dataset, and is only used for final assessement of the model. It won't be used for anywhere else in the project. The **edx** data is going to be split into two subset, **train_set** and **test_set**.

2.1 The Simplest Model

The simplest model, {Sale Price}=b0+b1x{Gross Square Feet}+e_i.I only consider the gross square feet in this model.

```
fit_simplest<- lm(SALE.PRICE~GROSS.SQUARE.FEET,data = train_set)

tidy(fit_simplest)
```

```
## # A tibble: 2 x 5
##   term              estimate std.error statistic       p.value
##   <chr>                <dbl>     <dbl>     <dbl>         <dbl>
## 1 (Intercept)         61686.  10185.        6.06 0.00000000145
## 2 GROSS.SQUARE.FEET     367.      6.21     59.0   0
```

```
summary(fit_simplest)[9]
```

```
## $adj.r.squared
## [1] 0.2835926
```

```
b0<-fit_simplest$coefficients[1]
b1<-fit_simplest$coefficients[2]
```

So the model 1 can be expressed as: {Sale.Price}=61686+367*{GROSS.SQUARE.FEET}.

The model predicts the base price for a property is 61686 dollars and every extra square feet is 367 dollars.

Performance

```
rmse_results
```

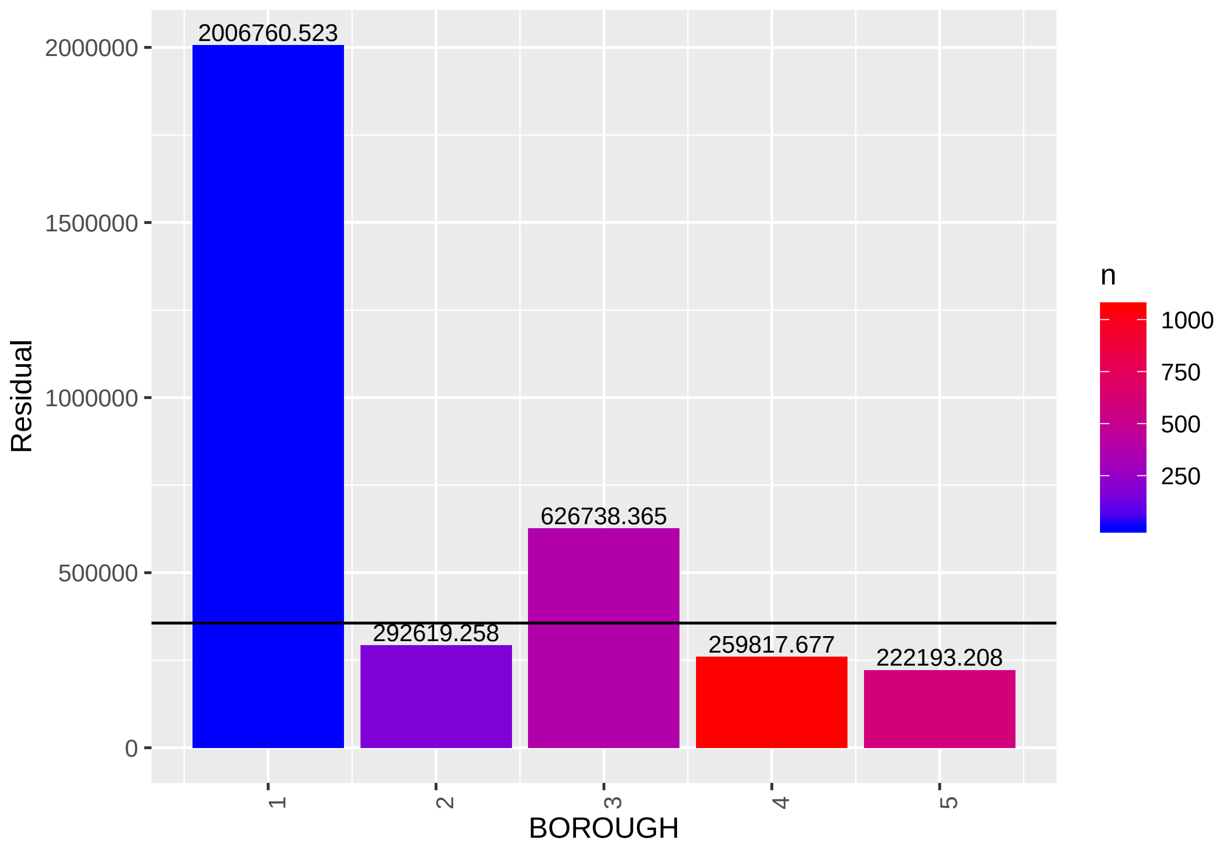
```
##        method RMSE_in_thousand MED_in_thousand k adj.r.squared
## 1 gross sq ft              356             163 2     0.2835926
```

Pros

This modle is the simplest linear regression model.

Cons

The adjusted r square is only 28%. And the RMSE for the properties in Manhattan (Borough 1) is very high.



Therefore, I need to consider borough as a factor for the next model.

2.2 Model 2

In the 2nd Model,I consider GROSS.SQUARE.FEET as well as whethr the property is in Manhattan.The regression expression is y=b0+b1xGSF+b2x(I(Manhattan)xGSF.

I(Manhattan) is identity funcition. It equals to 1 if the property is in Manhattan; otherwise zero.

```r
fit_IsManhattan<- lm(SALE.PRICE~GROSS.SQUARE.FEET+GSFandManhattan,data = temp)
tidy(fit_IsManhattan)
```

```
## # A tibble: 3 x 5
##   term              estimate std.error statistic  p.value
##   <chr>                <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)         79050.  10198.        7.75 1.01e-14
## 2 GROSS.SQUARE.FEET     354.      6.25     56.6  0.
## 3 GSFandManhattan       271.     22.1      12.3  2.47e-34
```

```r
summary(fit_IsManhattan)[9]
```

```
## $adj.r.squared
## [1] 0.2955594
```

So the model can be expressed as: {Sale.Price}=79050+354* {GROSS.SQUARE.FEET} +271 * {IS.MANHATTAN} * {GROSS.SQUARE.FEET}.

The model predicts the base price for a property is 79050 dollars.If the property is not in Manhattan,every extra square feet is 354 dollars. If the property is in Manhattan, every extra square feet is 625 dollars.

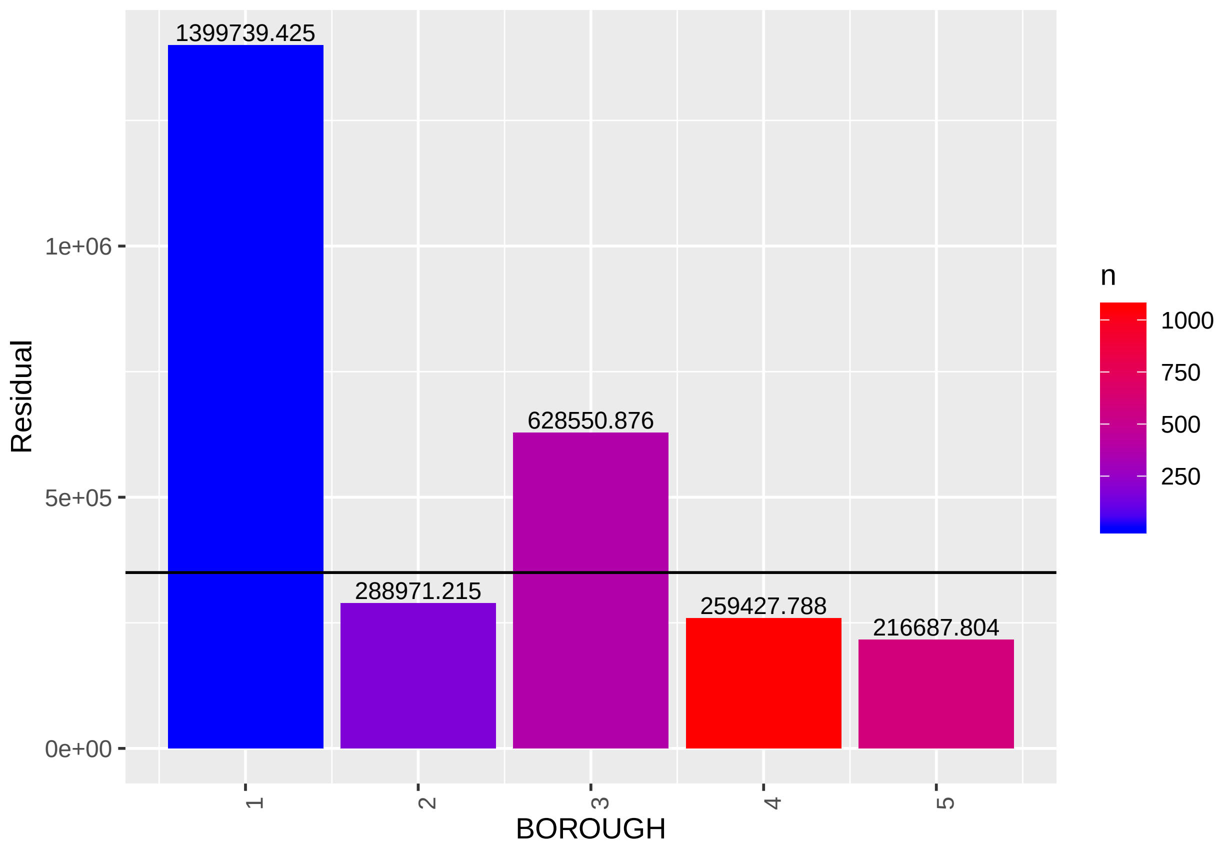
Performance

```r
rmse_results
```

```
##          method RMSE_in_thousand MED_in_thousand k adj.r.squared
## 1  gross sq ft              356             163 2     0.2835926
## 2 Is Manhattan              350             163 2     0.2955594
```

Pros

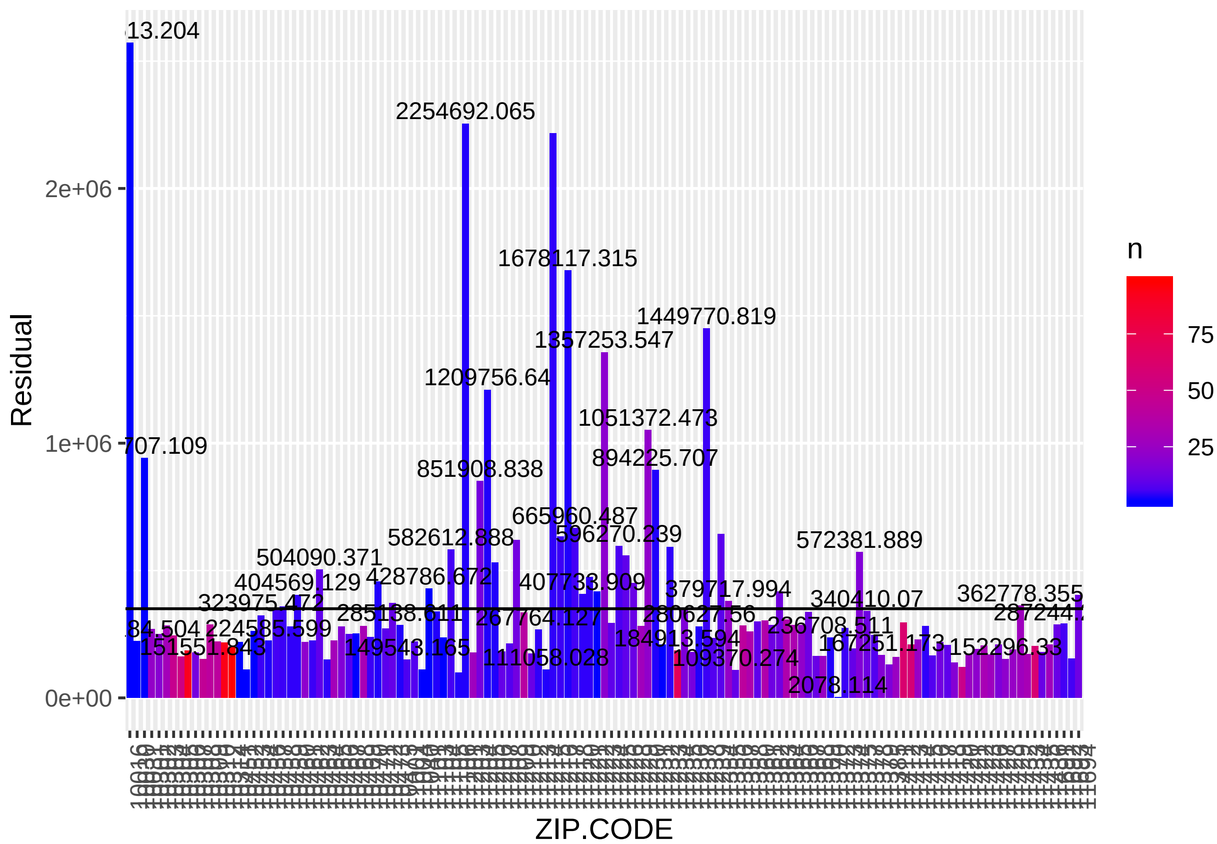
The model quantify the price difference bettween Manhattan and other boroughs in New York. The RMSE reduced and the Adjsuted R Square increased.



The RMSE in Manhattan siginificantly reduced from 2 miliion to 1.4 milition.

Cons

The adjusted r square is 30%. And the RMSE for the properties is high for some ZIP codes.




```
i=12
d<-temp_p%>%dplyr::select(i,len-1) %>%
        rename(A=1) %>%
        group_by(A) %>%
        summarise(r=sqrt(mean(residul)),n=NROW(A))
sum(d$n[d$r>300000])
```

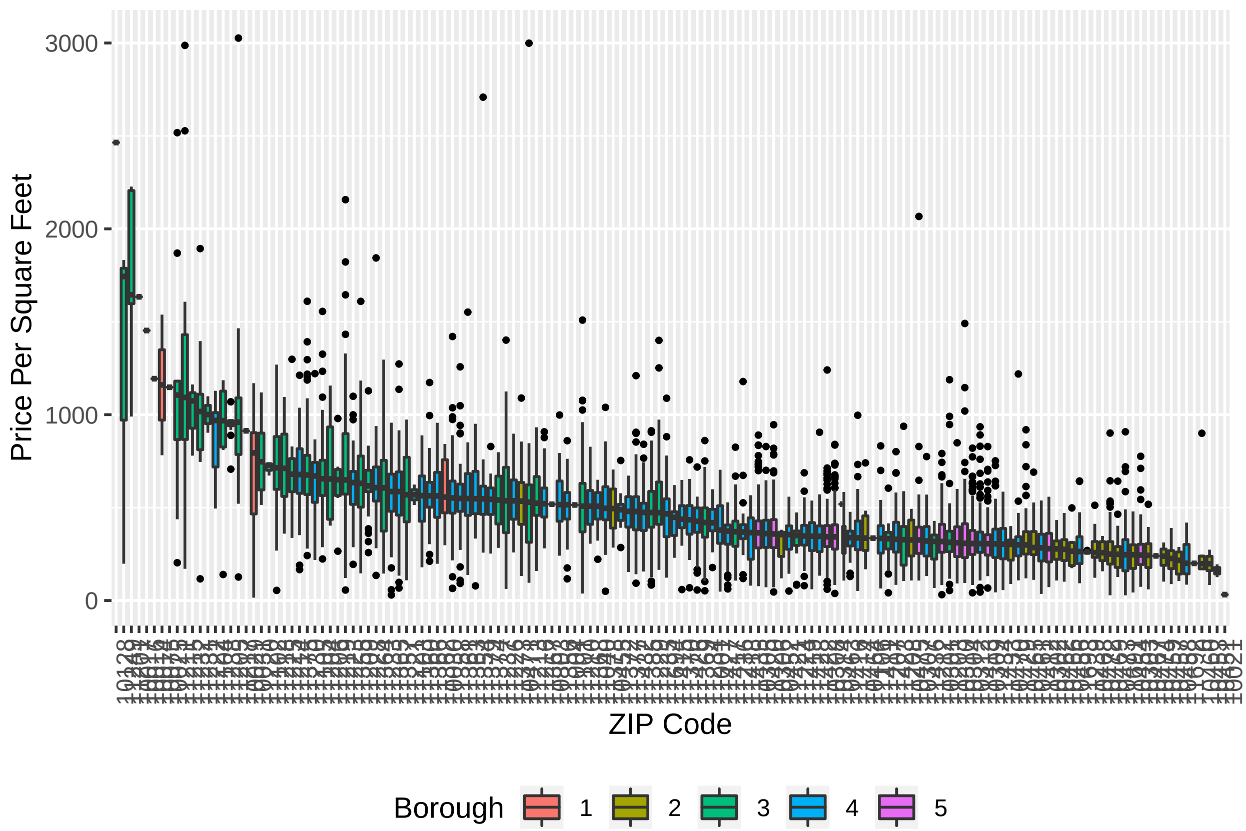
```
## [1] 426
```

There are 426 properties within the the high RMSE ZIP Code area. We try to reduce this number in model 3.

Therefore, let's zoom in and consider ZIP code in our next model.

2.3 Model 3

I added a column "price per square feet", which is calulated as {SALE.PRICE}/{GROSS.SQUARE.FEET}.The following boxplot shows the how the price per square feet varies by ZIP Code.



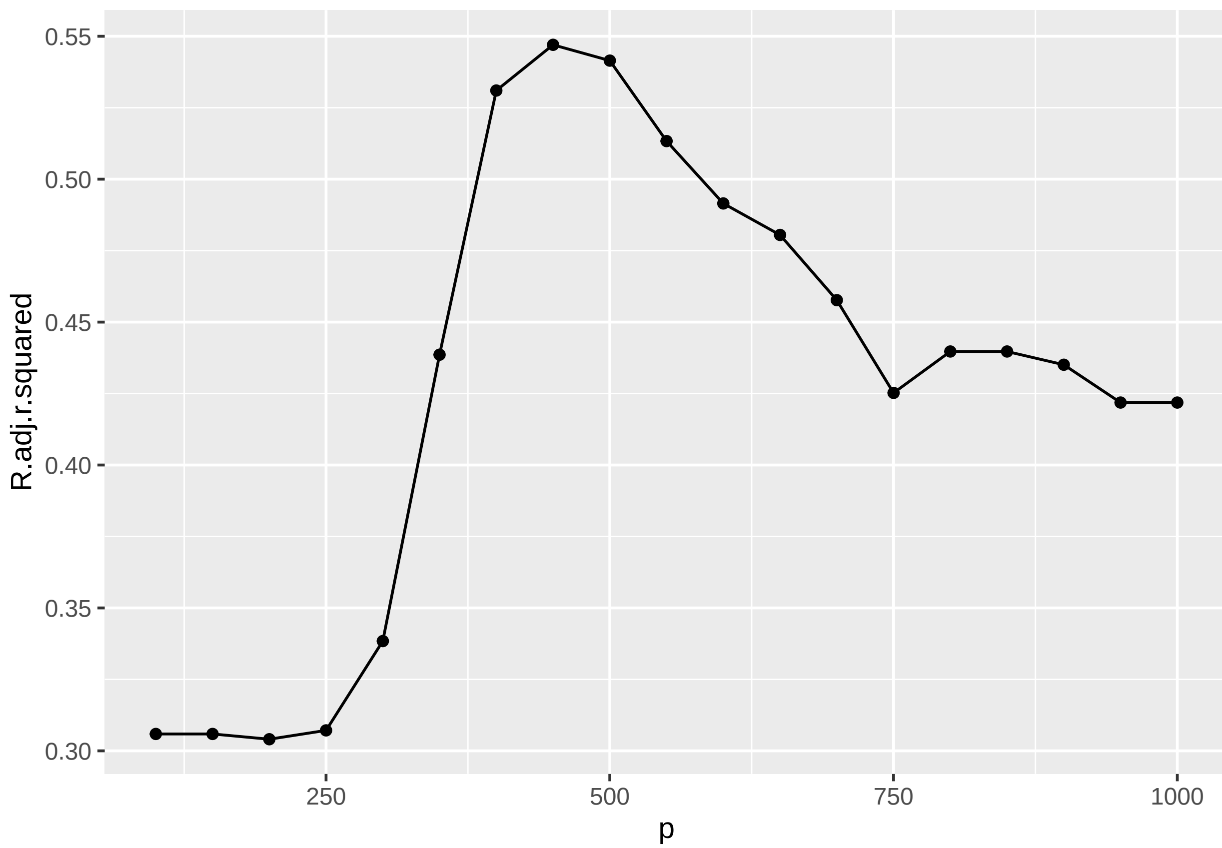
Since the means of price per squaer feet by ZIP Code vary between 2000 and 500 roughly, I am going to pick a number M. If the price per square feet is higher than M, it is in a High Pricing Area.

Then how to pick a proper number? I will run a loop and use the number that gives best adjusted R square as the M.

```
df[which.max(df$R.adj.r.squared),]
```

```
##     p R.adj.r.squared
## 8 450       0.5470054
```

```
df %>% ggplot(aes(x=p,y=R.adj.r.squared))+geom_point()+geom_line()
```



So M=450.

In the 3rd Model,I consider whether the property is in a High Pricing Area.The regression expression is y=b0+b1xGSF+b2x(I(Manhattan)xGSF.+b3x(I(PPSF>450)xGSF.

I(PPSF>450) equals to 1 if the price per square feet of the property is higher than 450; otherwise zero.

```
fit_IsManhattanAndHighPricingArea<- lm(SALE.PRICE~GROSS.SQUARE.FEET+GSFandManhattan+ GSFandHIGH.PRICING

tidy(fit_IsManhattanAndHighPricingArea)
```

```
## # A tibble: 4 x 5
##   term                     estimate std.error statistic   p.value
##   <chr>                       <dbl>     <dbl>     <dbl>     <dbl>
## 1 (Intercept)               166580.   8273.        20.1  3.49e- 88
## 2 GROSS.SQUARE.FEET            203.      5.46      37.2  2.43e-281
## 3 GSFandManhattan              163.     17.8        9.19 4.79e- 20
## 4 GSFandHIGH.PRICING.AREA      276.      3.95      69.9  0.
```

```
summary(fit_IsManhattanAndHighPricingArea)[9]
```

```
## $adj.r.squared
## [1] 0.5470054
```

So the model can be expressed as: {Sale.Price}=166580+203* {GROSS.SQUARE.FEET} +163 * {IS.MANHATTAN} * {GROSS.SQUARE.FEET}+276 * {IS.HIGH.PRICING.AREA} * {GROSS.SQUARE.FEET}.

The model predicts the base price for a property is 166580 dollars.Every extra square feet is:

If the property is not in Manhattan and not in high pricing area, 203 dollars;

If the property is in Manhattan and not in high pricing area, 366 dollars;

If the property is not in Manhattan and in high pricing area, 479 dollars;

If the property is in Manhattan and in high pricing area, 642 dollars.

Performance

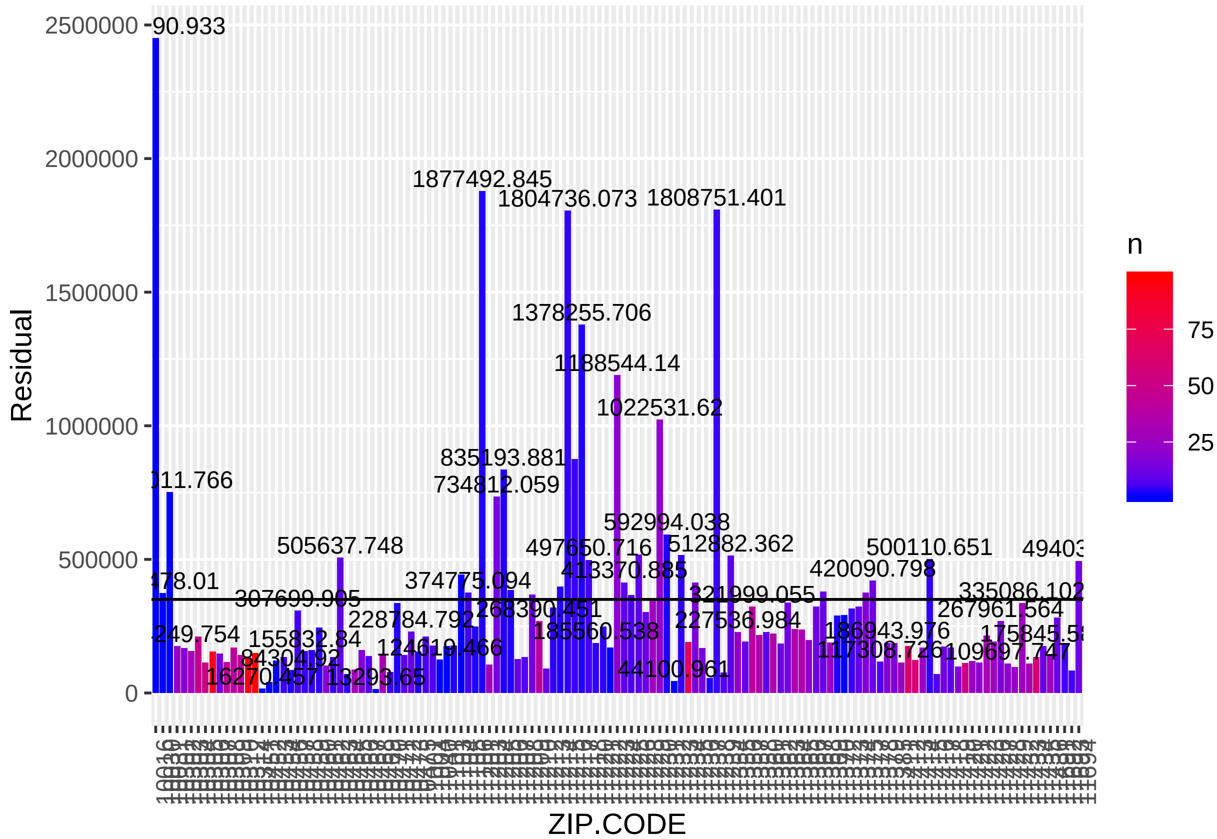
```
rmse_results
```

```
##                               method RMSE_in_thousand MED_in_thousand k
## 1                        gross sq ft              356             163 2
## 2                       Is Manhattan              350             163 2
## 3 Is Manhattan and High Pricing Area              300             100 3
##   adj.r.squared
## 1     0.2835926
## 2     0.2955594
## 3     0.5470054
```

Pros

The Adjusted R Square increase from 30% (model 2) to 55%. The RMSE is reduced from 350K to 300K. The median of the residual is reduced from 163K to 100K.

Also, the RMSEs by ZIP Code varies less than model 2.



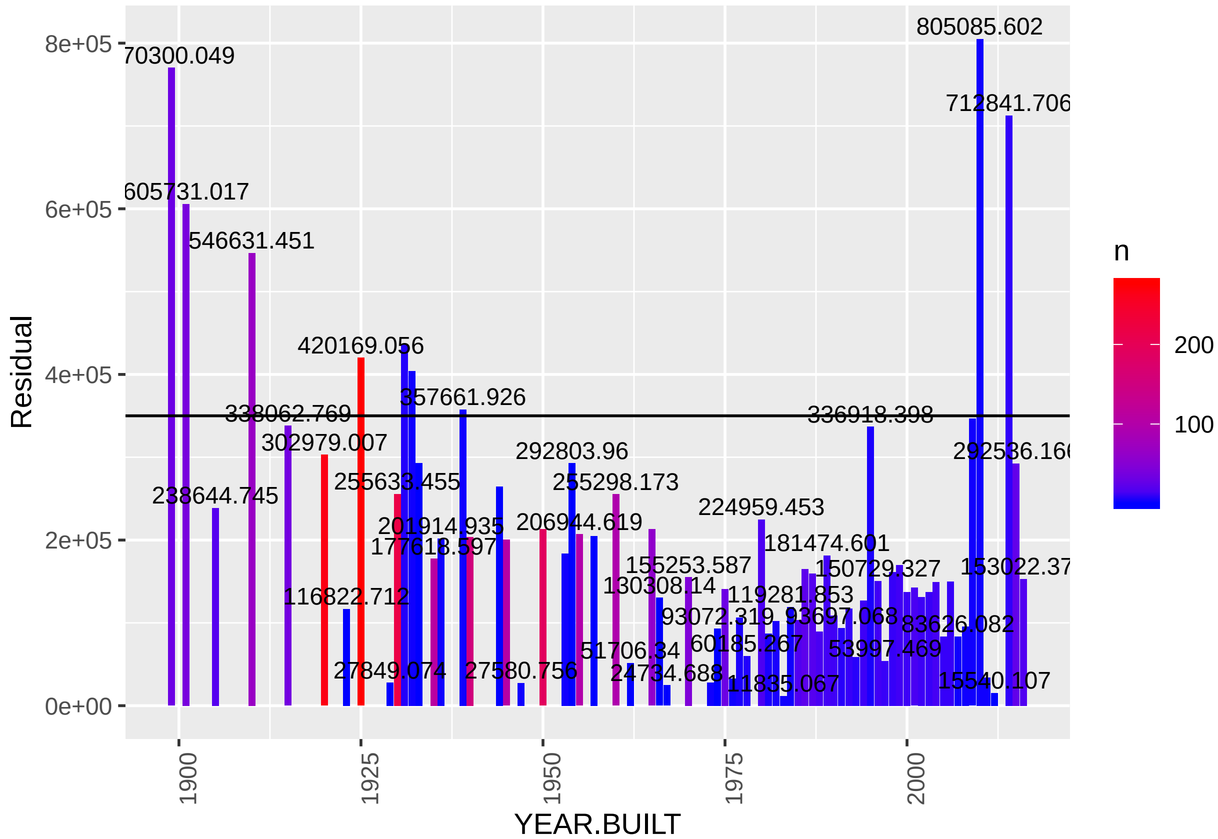
```
i=12
d<-temp_p%>%dplyr::select(i,len-1) %>%
        rename(A=1) %>%
        group_by(A) %>%
        summarise(r=sqrt(mean(residul)),n=NROW(A))
sum(d$n[d$r>300000])
```

```
## [1] 377
```

There are 377 properties within the the high RMSE ZIP Code area. It is siginificantly lower than model 2.

Cons

Let's look at the RMSE by year built. The RMSE decreases as year built is newer. It indicates the model doesn't account for the change of value when the property gets old.




```
i=18
d<-temp_p%>%dplyr::select(i,len-1) %>%
        rename(A=1) %>%
        group_by(A) %>%
        summarise(r=sqrt(mean(residul)),n=NROW(A))
sum(d$n[d$r>300000])
```

```
## [1] 739
```

There are 739 properties built in the years of higher RMSEs.

3 Results

```
rmse_validation
```

```
## [1] 320688.4
```

The RMSE for the validation data set is 320688.

Model Summary

The simplest model considers only gross square feet and is used as a baseline model. It has an adjusted R square of 28%.

The model 2 considers the properties are more expensive in Manhattan. It reduces RMSE of the properties within Manhattan. However, the overall model doesn't improve very much. The adjusted R square increases to 30%.

The model 3 considers the price per square feet as a factor. It classifies the properties by ZIP Code. The RMSE reduces from 350K to 300K, which falls short of the goal of 100K. However, the adjusted R square jumps siginificantly to 55%, which reaches the project goal of 50%.

4 Conclusion

Summary of the report

The report starts from the simpleset model and is developed by adding factors to the previous model. And I use the RMSE histogram, boxplot to as a guidance of what factors should be added to the model.

I like the ZIP Code Model, because it takes into price per square feet into consideration.

Limitations

The model 3 performs poorly with older properties. It doesn't consider that the age of the properties has impacts on the property value.

It also does not consider differenct building classes.

Future Works

In order to improve the model, I will include year built in my next model. Also, I will look into the impact from building class.

5 Reference

1 R Markdown Cheat Sheet [(https://rstudio.com/wp-content/uploads/2015/02/rmarkdown-cheatsheet.pdf)]

2 Data Science Courses from HarvardX [(https://courses.edx.org/)]

3 Boxplot in R [http://www.sthda.com/english/wiki/ggplot2-box-plot-quick-start-guide-r-software-and-data-visualization]