The Effect of Gross Square Footage on the Sale Prices of Single Family Homes in New York City

Aleksei Ostlund

04/10/2021

Introduction

The data sets used in this project are made publicly available by the City of New York. The data contained represents property sales for a 12-month rolling time period. For the purposes of this analysis, the time frame was September 2020 - August 2021.

The purpose of this project is to measure how well the size of housing explains the sale price across New York City and its boroughs. The housing type chosen for this analysis is A5: attached or semi-detached single family homes.

The question of how well gross square footage predicts the sale price in each borough will be addressed by using linear regression modeling.

Load packages

```
library(magrittr)
library(stringr)
library(dplyr)
library(readr)
library(ggplot2)
library(hrbrthemes)
library(tidyverse)
```

Data import

```
library(readxl)
manhattan_data <- read_xlsx("rollingsales_manhattan.xlsx", skip = 4)
bronx_data <- read_xlsx("rollingsales_bronx.xlsx", skip = 4)
brooklyn_data <- read_xlsx("rollingsales_brooklyn.xlsx", skip = 4)
queens_data <- read_xlsx("rollingsales_queens.xlsx", skip = 4)
si_data <- read_xlsx("rollingsales_statenisland.xlsx", skip = 4)
```

Combining rows from all 5 tables

```
nyc_property_data <- bind_rows(bronx_data, brooklyn_data, manhattan_data,
                               queens_data, si_data)
# drop other tables
remove(bronx_data, brooklyn_data, manhattan_data, queens_data, si_data)
```

```
# Change borough names from numbers to words
nyc_property_data <- nyc_property_data %>%
  mutate(BOROUGH=case_when(BOROUGH == 1 ~ 'MANHATTAN',
                           BOROUGH == 2 ~ 'BRONX',
                           BOROUGH == 3 ~ 'BROOKLYN',
                           BOROUGH == 4 ~ 'QUEENS',
                           BOROUGH == 5 ~ 'STATEN_ISLAND',
                           TRUE ~ 'na'))
```

Convert columns to lower case, remove spaces, duplicates, unneeded columns

```
colnames(nyc_property_data) %<>% str_replace_all("\\s", "_") %>% tolower()

# Convert capitalized fields to title case
nyc_property_data <- nyc_property_data %>%
  mutate(borough = str_to_title(nyc_property_data$borough, locale = 'en')) %>%
  mutate(neighborhood = str_to_title(nyc_property_data$neighborhood,
                                     locale = 'en')) %>%
  mutate(building_class_category = str_to_title(nyc_property_data$building_class_category, locale = 'en
  mutate(address = str_to_title(nyc_property_data$address, locale = 'en'))

# Remove possible duplicates
nyc_property_data <- nyc_property_data %>%
  distinct()
# Drop irrelevant columns
nyc_property_data <- subset(nyc_property_data, select = -easement )
```

Drop sales within families (assumed as under 10k), <150 sqft entries (The NYC Building Code requires all dwellings to be >150), and NA entries for sales/gross square feet columns

```
nyc_property_data <- nyc_property_data %>%
  filter(sale_price>10000) %>%
  filter(gross_square_feet>=150) %>%
  drop_na(sale_price, gross_square_feet) %>%
  arrange(borough, neighborhood)

# Export to CSV
write_csv(nyc_property_data, file='nycsales_cleaned.csv')
```

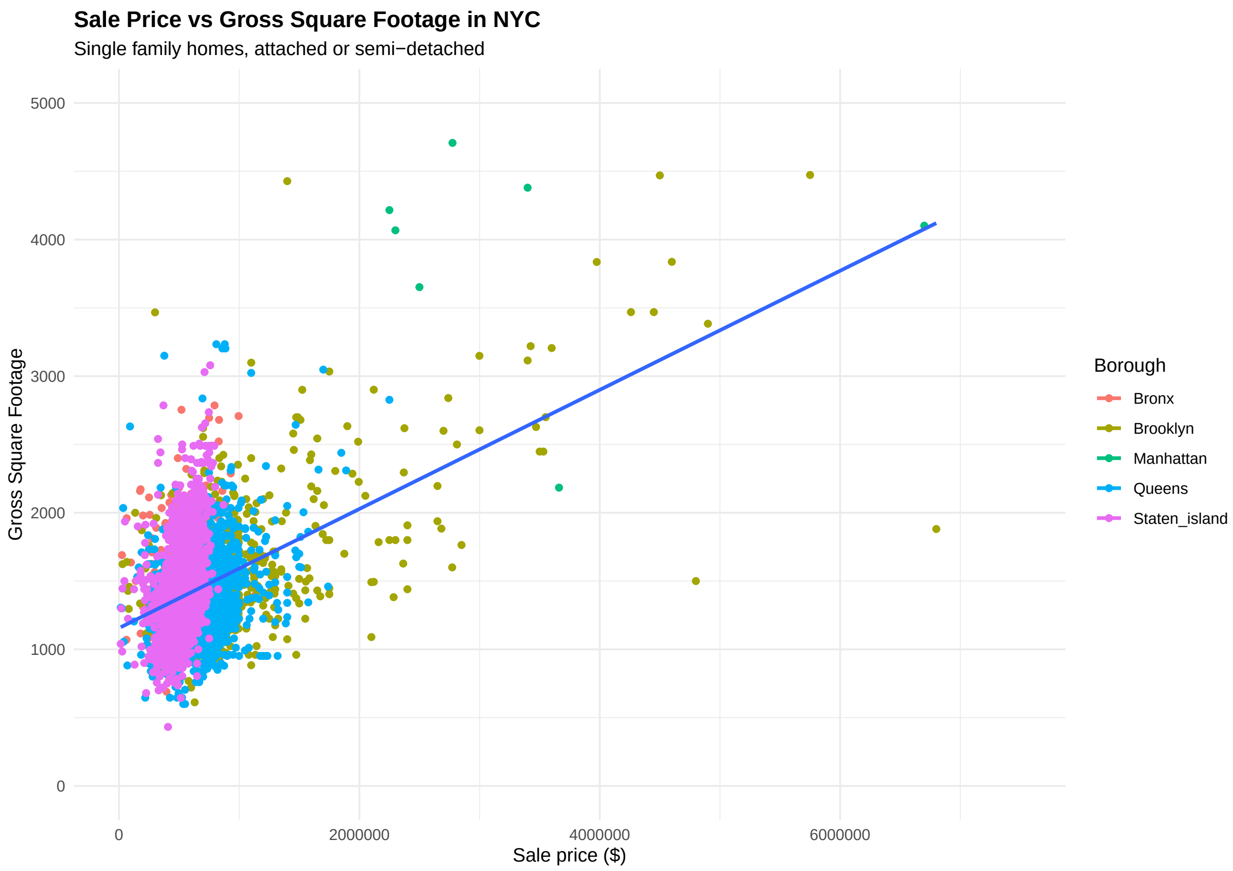
Filter for one family attached or semi-detached housing

```
one_family_attached <- nyc_property_data %>%
  filter(building_class_at_time_of_sale == 'A5')
```

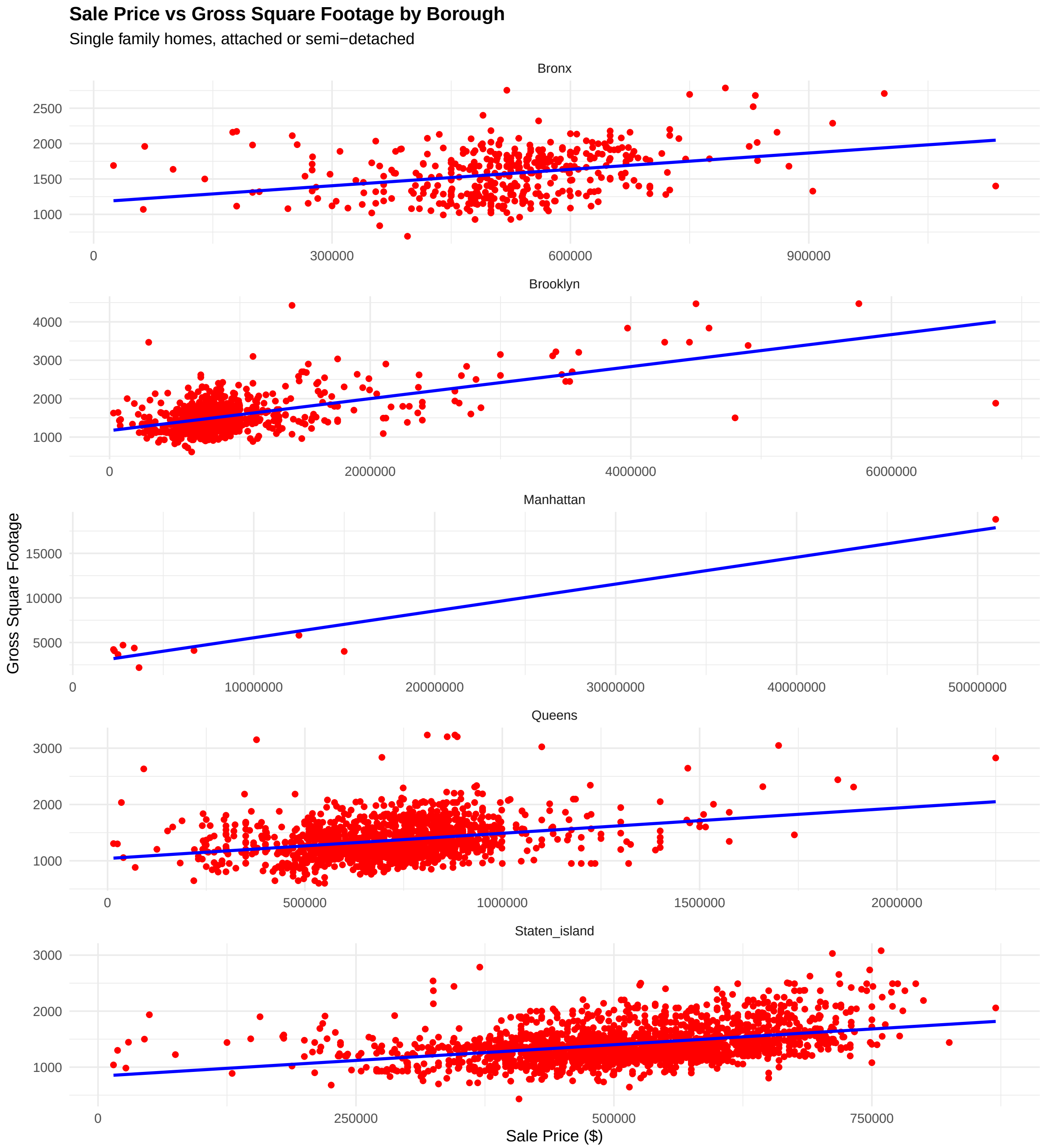
Scatterplots

```
options(scipen = 100)

ggplot(data= one_family_attached, mapping = aes(x= sale_price,
                                                 y= gross_square_feet,
                                                 color=borough))+
  geom_point()+
  xlim(0,7500000)+
  geom_smooth(method = 'lm', se= FALSE, aes(group = 1)) + theme(axis.title = element_text(size = 12),
    plot.title = element_text(face = "bold")) +labs(title = "The effect of gross square footage on sale
    x = "Sale price ($)", y = "Gross Square Footage",
    colour = "Borough") +
  theme_minimal()+
  ylim(0,5000) + theme(plot.title = element_text(face = "bold")) +
  labs(title = "Sale Price vs Gross Square Footage in NYC")+
  labs(subtitle = "Single family homes, attached or semi-detached")
```



```
# Scatterplot by borough
options(scipen = 100)
library(hrbrthemes)
ggplot(data= one_family_attached, mapping = aes(x= sale_price,
                                                y= gross_square_feet))+
  geom_point(color = 'red')+
  facet_wrap(~borough, ncol = 1, scales = 'free')+
  scale_y_continuous()+
  geom_smooth(method = 'lm', se= FALSE, color = 'blue') +
  theme_minimal() + theme(plot.title = element_text(face = "bold")) +
  labs(title = "Sale Price vs Gross Square Footage by Borough",
    x = "Sale Price", y = "Gross Square Footage")+labs(subtitle = "Single family homes, attached or semi
  labs(x = "Sale Price ($)")
```



Top sale in Manhattan is the Herbert M Strouse House, likely the largest private residence in Manhattan. It has a celebrity factor. I did not consider this an outlier as it is still a legitimate data point.

Generate linear regression models

```
one_family_lm <- lm(sale_price ~ gross_square_feet,data = one_family_attached)
summary(one_family_lm)
```

```
##
## Call:
```

```
## lm(formula = sale_price ~ gross_square_feet, data = one_family_attached)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -3017340  -269011    26648   269444 28508737
##
## Coefficients:
##                     Estimate  Std. Error t value            Pr(>|t|)
## (Intercept)       -1145799.41    31882.37  -35.94 <0.0000000000000002 ***
## gross_square_feet     1256.35       20.99   59.87 <0.0000000000000002 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 656000 on 4519 degrees of freedom
## Multiple R-squared:  0.4423, Adjusted R-squared:  0.4422
## F-statistic:  3584 on 1 and 4519 DF,  p-value: < 0.00000000000000022
```

The model has a sufficiently low p value (below 0.001) and high t value so that we can say the two variables are related. The r squared score for the entirety of New York is a moderate 0.4422.

Generate separate linear regression models for each borough.

```r
library(broom)
library(tidyr)
library(purrr)
borough_lms <- one_family_attached %>%
  group_by(borough) %>%
  nest() %>%
  mutate(linear_model = map(.x = data,
                            .f = ~lm(sale_price ~ gross_square_feet,
                            data = .)))
```

Generate tidy coefficients and regression summaries

```r
borough_lms <- borough_lms %>%
  mutate(tidy_coefficients = map(.x = linear_model,
                                 .f = tidy,
                                 conf.int = TRUE))

#Unnest
tidy_unnested <- borough_lms %>%
  select(borough, tidy_coefficients) %>%
  unnest(cols = tidy_coefficients)

tidy_unnested %>%
  filter(term == 'gross_square_feet') %>%
  rename(slope = estimate) %>%
  select(borough, term, slope) %>%
  arrange((desc(slope))) %>%
  print()
```

```
## # A tibble: 5 x 3
## # Groups:   borough [5]
##   borough       term              slope
##   <chr>         <chr>             <dbl>
## 1 Manhattan     gross_square_feet 3041.
## 2 Brooklyn      gross_square_feet  806.
## 3 Queens        gross_square_feet  199.
## 4 Staten_island gross_square_feet  144.
## 5 Bronx         gross_square_feet  116.
```

When looking at the slopes, we can see that Manhattan has the biggest price increase per sq foot followed by Brooklyn. Other boroughs are noticeable behind.

Generate regression summaries

```
borough_regressions <- one_family_attached %>%
  group_by(borough) %>%
  nest() %>%
  mutate(linear_regression = map(.x = data,
                                 .f = ~lm(sale_price ~ gross_square_feet,
                                          data =.
                                      ))) %>%
  mutate(tidy_regressions = map(.x = linear_regression,
                                .f = glance,
                                conf.int = TRUE)) %>%
  select(borough, tidy_regressions)%>%
  unnest(cols = tidy_regressions)

borough_regressions %>% arrange(desc(adj.r.squared)) %>%
  select(borough, adj.r.squared) %>%
  print()
```

```
## # A tibble: 5 x 2
## # Groups:   borough [5]
##   borough       adj.r.squared
##   <chr>                 <dbl>
## 1 Manhattan            0.905
## 2 Brooklyn             0.335
## 3 Staten_island        0.161
## 4 Queens               0.0887
## 5 Bronx                0.0869
```

Neighborhoods with best fit for linear model

```
# Remove neighborhoods with less than 10 sales
neighborhoods_count <- one_family_attached %>%
  count(neighborhood) %>%
  filter(n>9)
# New dataset
neighborhoods_cleaned <- one_family_attached %>%
  filter(neighborhood %in% neighborhoods_count$neighborhood)

# Linear regressions by neighborhood
neighborhood_regressions <- neighborhoods_cleaned %>%
  group_by(borough, neighborhood) %>%
  nest() %>%
  mutate(linear_model = map(.x = data,
             .f = ~lm(sale_price ~ gross_square_feet, data = .))) %>%
  mutate(tidy_regressions = map(.x = linear_model,
                               .f = glance,
                               conf.int = TRUE)) %>%
  select(borough, neighborhood, tidy_regressions) %>%
  unnest(cols=tidy_regressions)

# Remove high p values and filter for adj r squared
significant_neighborhood_regressions <- neighborhood_regressions %>%
  filter(p.value <0.05) %>%
  filter(adj.r.squared > 0.5) %>%
  arrange(desc(adj.r.squared))

print(significant_neighborhood_regressions)
```

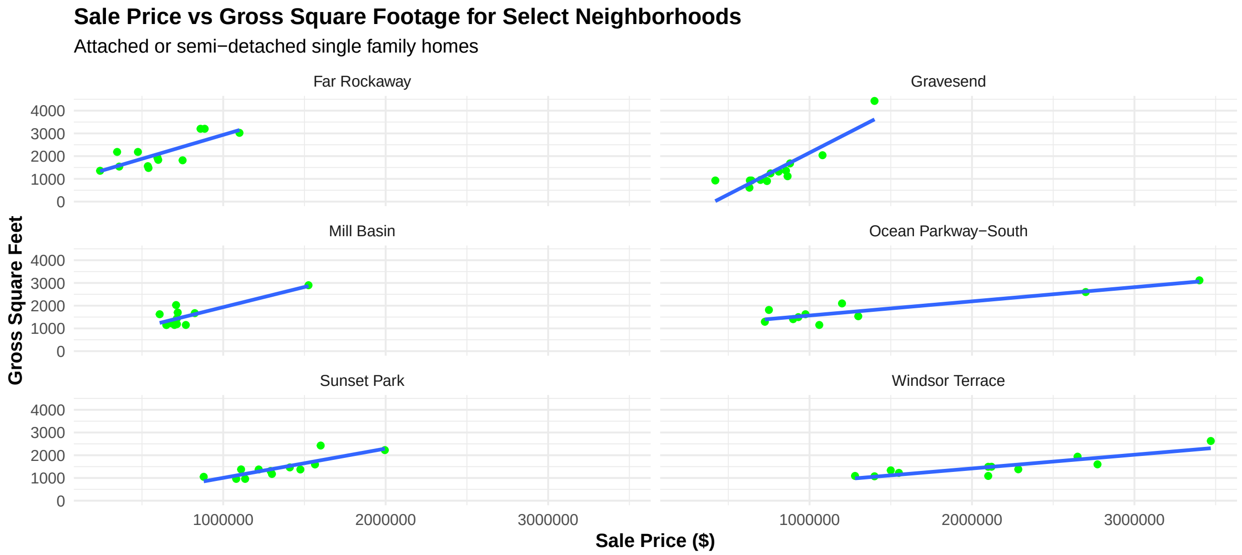
```
## # A tibble: 6 x 14
## # Groups:   borough, neighborhood [6]
##   borough  neighborhood    r.squared adj.r.squared  sigma statistic p.value    df
##   <chr>    <chr>               <dbl>         <dbl>  <dbl>     <dbl>   <dbl> <dbl>
## 1 Brooklyn Ocean Parkway~      0.825         0.803 4.02e5      37.7 2.77e-4     1
## 2 Brooklyn Gravesend           0.812         0.795 1.09e5      47.5 2.61e-5     1
## 3 Brooklyn Windsor Terra~      0.781         0.757 3.31e5      32.1 3.08e-4     1
## 4 Brooklyn Mill Basin          0.690         0.659 1.42e5      22.2 8.24e-4     1
## 5 Brooklyn Sunset Park         0.686         0.655 1.74e5      21.8 8.76e-4     1
## 6 Queens   Far Rockaway        0.613         0.575 1.64e5      15.9 2.59e-3     1
## # ... with 6 more variables: logLik <dbl>, AIC <dbl>, BIC <dbl>,
## #   deviance <dbl>, df.residual <int>, nobs <int>
```

Graph of neighborhoods where gross square footage is the best predictor of sale price

```
# Graph these neighborhoods

neighborhood_graphs <- one_family_attached %>%
  filter(neighborhood == 'Ocean Parkway-South' |
           neighborhood == 'Gravesend' |
           neighborhood == 'Windsor Terrace' |
           neighborhood == 'Mill Basin' |
           neighborhood == 'Sunset Park' |
           neighborhood == 'Far Rockaway')

ggplot(data=neighborhood_graphs, mapping= aes(x=sale_price,
                                              y=gross_square_feet))+
  geom_point(color='green') +
  theme_minimal() +
  geom_smooth(method = 'lm', se=FALSE) +
  facet_wrap(~neighborhood, ncol = 2) + theme(axis.title = element_text(face = "bold"),
    plot.title = element_text(face = "bold")) +labs(title = "Sale Price vs Gross Square Footage for Sele
    x = "Sale Price ($)", y = "Gross Square Feet",
    subtitle = "Attached or semi-detached single family homes")
```



Conclusion

New York City as a whole had a moderate-weak correlation between gross square feet and sale price. Looking at the r-squared values for all 5 boroughs individually it appears that Manhattan has a strong value and Brooklyn has a low-moderate one. The other 3 have low values. Furthermore, Manhattan had the largest sale price increase per square foot. Brooklyn was second while the other 3 boroughs did not have such large increases.

These values indicate that in the Bronx, Queens, and Staten Island, other factors have a larger impact on pricing of attached/semi-detached single family homes. These could be location, proximity to transport, or the quality of the interior. Finally, the values for Manhattan needs to be investigated more as there are only 10 sales of this property type in the past year. Data from previous years could be used to compare r squared values.

When picking out specific neighborhoods, there were 6 that had an r squared value greater than 0.5 while having suitable p and t values. When looking at sales for the chosen housing class in these select neighborhoods gross square footage can be seen as a fairly reliable indicator of sale price.