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

Aleksei Ostlund

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.

Data Information

Column names:

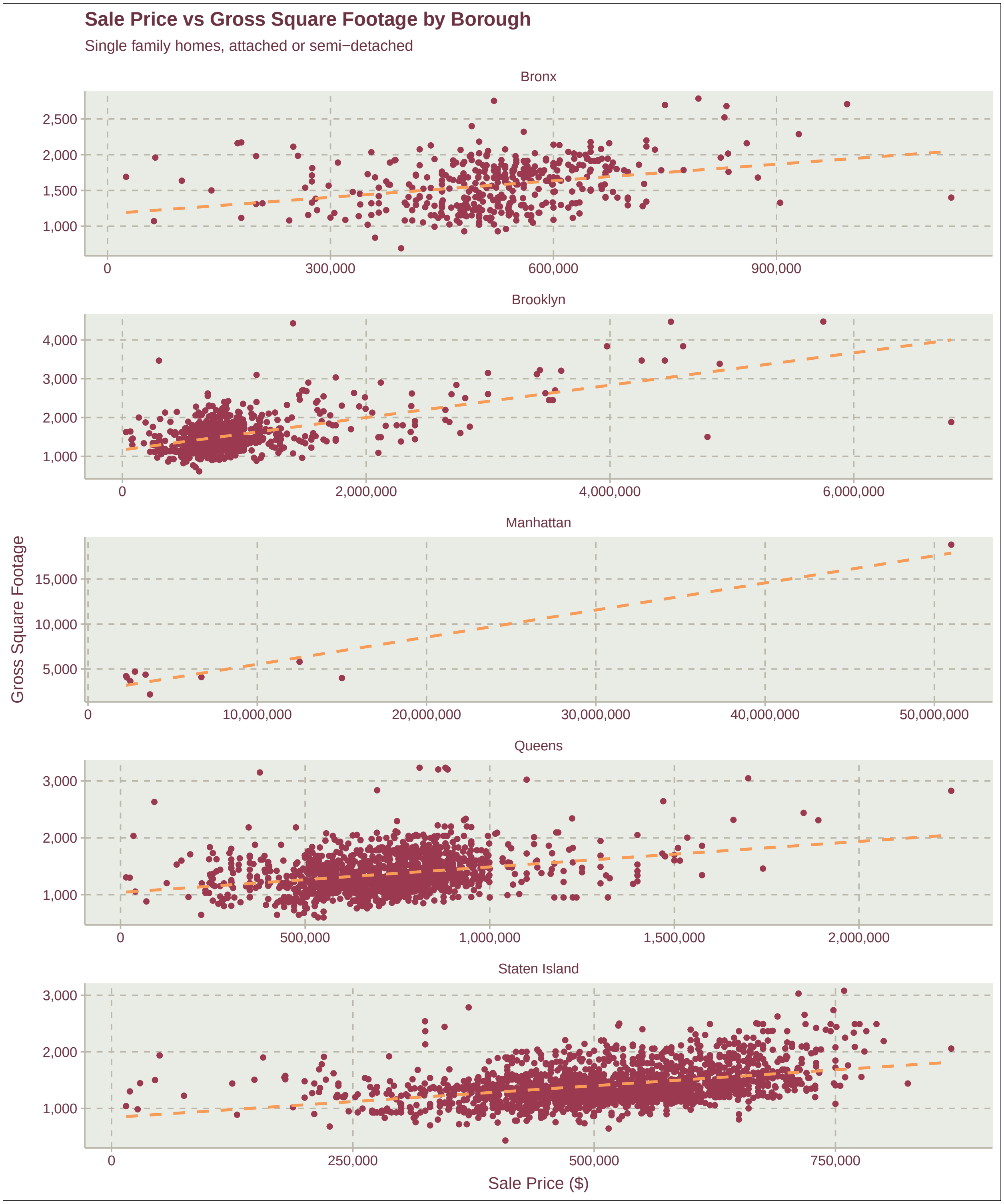
```
##  [1] "borough"                        "neighborhood"
##  [3] "building_class_category"        "tax_class_at_present"
##  [5] "block"                          "lot"
##  [7] "building_class_at_present"      "address"
##  [9] "apartment_number"               "zip_code"
## [11] "residential_units"              "commercial_units"
## [13] "total_units"                    "land_square_feet"
## [15] "gross_square_feet"              "year_built"
## [17] "tax_class_at_time_of_sale"      "building_class_at_time_of_sale"
## [19] "sale_price"                     "sale_date"
```

Row count:

```
## [1] 87212
```

The data has been filtered to only include data for class A5 properties.

Data Visualization



The top sale in Manhattan is the Herbert M Strouse House. This is likely the largest private residence in Manhattan so it has a celebrity factor attached to it.It was not considered an outlier as it still represented a legitimate property sale.

Data Distribution

```
## # A tibble: 5 x 2
##   borough           n
##   <chr>         <int>
## 1 Bronx           409
## 2 Brooklyn        782
## 3 Manhattan        10
## 4 Queens         1447
## 5 Staten_island  1873
```

As Manhattan had at least 10 data points it was still included in the model.

Linear Regression Model

```
##
## 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 0.05) and large enough t value(above an absolute value of 2). One can conclude that square footage is a predictor for the sale price of a property. This is as the t value exceeds

However, the r squared score is a weak to moderate 44%. This implies that a significant proportion of sale price is explained by other variables.

Linear Regression by Borough

```
## # A tibble: 5 x 7
## # Groups:   borough [5]
##   borough       estimate std.error statistic  p.value conf.low conf.high
##   <chr>            <dbl>     <dbl>     <dbl>    <dbl>    <dbl>     <dbl>
## 1 Bronx             116.     18.3       6.31 7.20e-10     79.6      152.
## 2 Brooklyn          806.     40.6      19.9  2.41e-71    726.       885.
## 3 Manhattan        3041.    326.        9.32 1.44e- 5   2288.      3793.
## 4 Queens            199.     16.7      11.9  3.00e-31    166.       232.
## 5 Staten_island     144.      7.57     19.0  9.86e-74    129.       159.
```

Price by Borough

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

We can see that Manhattan has the biggest price per square foot at $3040. Brooklyn comes in second with the other boroughs noticeable behind.

R Squared by Borough

```
## # A tibble: 5 x 2
## # Groups:   borough [5]
##   borough       r_squared
##   <chr>             <dbl>
## 1 Manhattan        0.916
## 2 Brooklyn         0.336
## 3 Staten_island    0.162
## 4 Queens           0.0893
## 5 Bronx            0.0892
```

Looking at the r-squared values for all 5 boroughs it appears that Manhattan has a strong value and Brooklyn has a low-moderate one. The other 3 have low values. This indicates that in these three boroughs other factors, such as location, proximity to transport, or the quality of the interior much have a greater effect on the price of single family homes. Manhattan also only had 10 sales of this type in the year that was analyzed. More historical data could be used to evaluate this correlation.

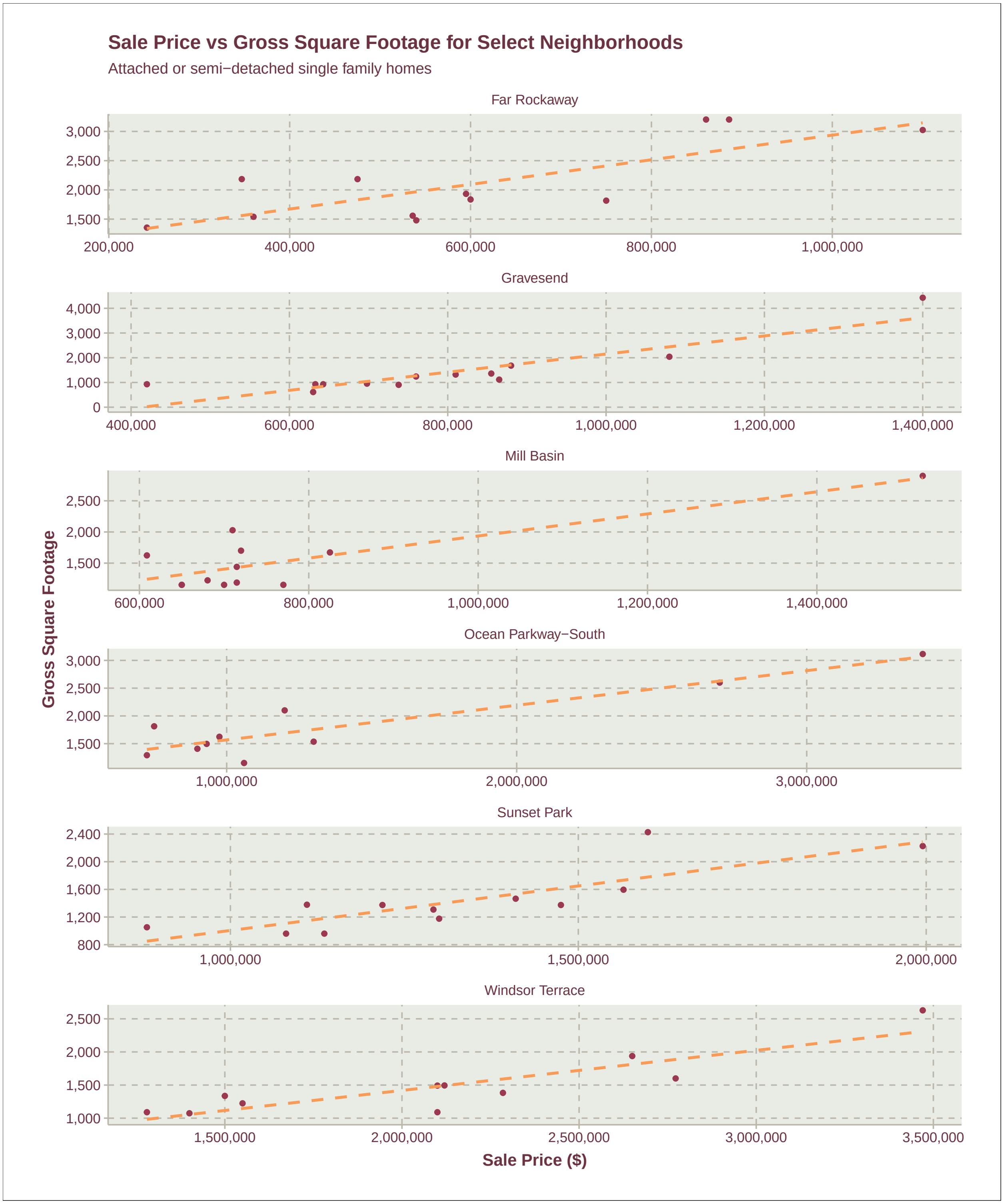
Neighborhoods with Best Fit

```
## # A tibble: 6 x 4
## # Groups:   borough, neighborhood [6]
##   borough  neighborhood        r.squared   p.value
##   <chr>    <chr>                   <dbl>     <dbl>
## 1 Brooklyn Ocean Parkway-South     0.825 0.000277
## 2 Brooklyn Gravesend               0.812 0.0000261
## 3 Brooklyn Windsor Terrace         0.781 0.000308
## 4 Brooklyn Mill Basin              0.690 0.000824
## 5 Brooklyn Sunset Park             0.686 0.000876
## 6 Queens   Far Rockaway            0.613 0.00259
```

The above neighborhoods had an r squared above 0.5 (50%) and a p value below 0.05. Most of them are in Brooklyn, reinforcing the idea that square footage is a helpful predictor in that borough.

Neighborhoods Visualization

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.



Conclusion

New York City as a whole had a moderate-weak correlation between gross square footage 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 highest cost per square foot. Brooklyn was second while the other 3 boroughs were significantly lower.

The findings indicate that in the Bronx, Queens, and Staten Island other factors have a larger impact on pricing of attached/semi-detached single family homes.

Additionally, the reliability of the model varied by neighborhood. Gross footage was a particularly strong predictor in a select number of neighborhoods in Brooklyn.