Project Final

William Ma

May 22, 2019

Data Science Tutorial

Welcome to the world of data science! In this tutorial, we will go over important concepts that you will need to know as a data scientist. The topics we will go over includes: data wrangling (cleaning and unifying complex datasets for easy access and analysis), EDA (or Exploratory Data Analysis), Machine Learning (Linear Regression), and final discussion of the insights learned from analysis.

SAT Scores and Property Sales of New York City

This analysis will examine two datasets: SAT Scores of public schools and Property Sales, both located in New York City at relatively the same time. I would like to know if there is a direct correlation that can be established between these two attributes. This analysis will use cost of living (property price) within a borough as an indicator of wealth, and cumulative SAT scores (out of 2400) as an indicator of academic performance. I hope to prove with substantial data that there is a direct correlation between family wealth and academic performance.

Ingest Data

The following datasets that I will use are labeled, “scores.csv” and “nyc-rolling-sales.csv”, both extracted from Kaggle.

```
school_df <- read.csv(file="/Users/wma52/Downloads/scores.csv", header=TRUE, sep=",")
property_df <- read.csv(file="/Users/wma52/Downloads/nyc-rolling-sales.csv", header=TRUE, sep=
",")

head(school_df)
```

```
##   School.ID                                                School.Name
## 1    02M260                              Clinton School Writers and Artists
## 2    06M211   Inwood Early College for Health and Information Technologies
## 3    01M539 New Explorations into Science, Technology and Math High School
## 4    02M294                                          Essex Street Academy
## 5    02M308                                 Lower Manhattan Arts Academy
## 6    02M545                High School for Dual Language and Asian Studies
##     Borough Building.Code       Street.Address      City State Zip.Code
## 1 Manhattan          M933 425 West 33rd Street Manhattan    NY    10001
## 2 Manhattan          M052   650 Academy Street Manhattan    NY    10002
## 3 Manhattan          M022  111 Columbia Street Manhattan    NY    10002
## 4 Manhattan          M445     350 Grand Street Manhattan    NY    10002
## 5 Manhattan          M445     350 Grand Street Manhattan    NY    10002
## 6 Manhattan          M445     350 Grand Street Manhattan    NY    10002
##   Latitude Longitude   Phone.Number Start.Time End.Time Student.Enrollment
## 1 40.75321 -73.99786   212-695-9114                                     NA
## 2 40.86605 -73.92486 718-935-3660      8:30 AM  3:00 PM                 87
## 3 40.71873 -73.97943   212-677-5190    8:15 AM  4:00 PM               1735
## 4 40.71687 -73.98953   212-475-4773    8:00 AM  2:45 PM                358
## 5 40.71687 -73.98953   212-505-0143    8:30 AM  3:00 PM                383
## 6 40.71687 -73.98953   212-475-4097    8:00 AM  3:35 PM                416
##   Percent.White Percent.Black Percent.Hispanic Percent.Asian
## 1
## 2          3.4%         21.8%            67.8%          4.6%
## 3         28.6%         13.3%            18.0%         38.5%
## 4         11.7%         38.5%            41.3%          5.9%

## 5          3.1%         28.2%            56.9%          8.6%
## 6          1.7%          3.1%             5.5%         88.9%
##   Average.Score..SAT.Math. Average.Score..SAT.Reading.
## 1                       NA                          NA
## 2                       NA                          NA
## 3                      657                         601
## 4                      395                         411
## 5                      418                         428
## 6                      613                         453
##   Average.Score..SAT.Writing. Percent.Tested
## 1                          NA
## 2                          NA
## 3                         601          91.0%
## 4                         387          78.9%
## 5                         415          65.1%
## 6                         463          95.9%
```

```
head(property_df)
```

```
##   X BOROUGH  NEIGHBORHOOD                    BUILDING.CLASS.CATEGORY
## 1 4       1 ALPHABET CITY 07 RENTALS - WALKUP APARTMENTS
## 2 5       1 ALPHABET CITY 07 RENTALS - WALKUP APARTMENTS
## 3 6       1 ALPHABET CITY 07 RENTALS - WALKUP APARTMENTS
## 4 7       1 ALPHABET CITY 07 RENTALS - WALKUP APARTMENTS
## 5 8       1 ALPHABET CITY 07 RENTALS - WALKUP APARTMENTS
## 6 9       1 ALPHABET CITY 07 RENTALS - WALKUP APARTMENTS
##   TAX.CLASS.AT.PRESENT BLOCK LOT EASE.MENT BUILDING.CLASS.AT.PRESENT
## 1                   2A   392   6        NA                       C2
## 2                    2   399  26        NA                       C7
## 3                    2   399  39        NA                       C7
## 4                   2B   402  21        NA                       C4
## 5                   2A   404  55        NA                       C2
## 6                    2   405  16        NA                       C4
##                 ADDRESS APARTMENT.NUMBER ZIP.CODE RESIDENTIAL.UNITS
## 1           153 AVENUE B                    10009                5
## 2  234 EAST 4TH   STREET                    10009               28
## 3  197 EAST 3RD   STREET                    10009               16
## 4    154 EAST 7TH STREET                    10009               10
## 5 301 EAST 10TH   STREET                    10009                6
## 6 516 EAST 12TH   STREET                    10009               20
##   COMMERCIAL.UNITS TOTAL.UNITS LAND.SQUARE.FEET GROSS.SQUARE.FEET
## 1                0           5             1633              6440
## 2                3          31             4616             18690
## 3                1          17             2212              7803
## 4                0          10             2272              6794
## 5                0           6             2369              4615
## 6                0          20             2581              9730
##   YEAR.BUILT TAX.CLASS.AT.TIME.OF.SALE BUILDING.CLASS.AT.TIME.OF.SALE
## 1       1900                         2                              C2
## 2       1900                         2                              C7
## 3       1900                         2                              C7
## 4       1913                         2                              C4
## 5       1900                         2                              C2
## 6       1900                         2                              C4
##   SALE.PRICE           SALE.DATE
## 1    6625000 2017-07-19 00:00:00
## 2        -   2016-12-14 00:00:00
## 3        -   2016-12-09 00:00:00
## 4    3936272 2016-09-23 00:00:00
## 5    8000000 2016-11-17 00:00:00
## 6        -   2017-07-20 00:00:00
```

Data Wrangling

Tidy Property Data

NYC Property Sales dataset will be labeled ‘property_df’, which contains information about sales that had occurred in recent years, including the property price, boroughs the property is in, gross square feet, etc. The most important information we need in this dataset is the borough and property price. Afterwards, the data is manipulated as follows:

- First, we want to change the boroughs column to have actual names instead of numbers, so that we can join tables later for wrangling.
- Then, we will rename the variable for clarity.
- There are many '-' indicating missing values under property price and gross square feet. We will replace them with the average of all property price or all gross square feet.
- There are also many 0's in property prices. According to the datasource, $0 essentially means transfer of ownership, and this can harm our analysis. We will set all 0's to the average of all property prices.
- We will create a new attribute "property level", which categorizes the level of how much a property cost. This data will be used later in visualization.
- Many of the missing values are located property price and gross square feet. The best way to handle this data is to impute the empty slots with the average of their respective columns.

```
property_df <- property_df %>%
    mutate(Borough = ifelse(BOROUGH == 1, 'Manhattan', ifelse(BOROUGH == 2, 'Bronx', ifelse(BORO
UGH == 3, 'Brooklyn', ifelse(BOROUGH == 4, 'Queens', ifelse(BOROUGH == 5, 'Staten Island', NA
))))))  %>%
  select(Borough, GROSS.SQUARE.FEET, SALE.PRICE) %>%
  rename(
    gross_sqr_ft = GROSS.SQUARE.FEET,
    property_price = SALE.PRICE
  )
property_df$Borough <- as.factor(property_df$Borough)
suppressWarnings(property_df$property_price <- as.numeric(as.character(property_df$property_pric
e)))
suppressWarnings(property_df$gross_sqr_ft <- as.numeric(as.character(property_df$gross_sqr_ft)))

property_df[property_df == 0] <- NA

property_df <- property_df %>%
  mutate(property_level =
    ifelse(property_price >= 0 & property_price < 10000, "D", ifelse(property_price >= 10000 & p
roperty_price < 50000, "C", ifelse(property_price >= 50000 & property_price < 400000, "B", ifels
e(property_price >= 400000, "A", NA))))
  ) %>%
  arrange(property_price)

property_df <- property_df %>%
  replace_na(list(property_price=mean(.$property_price, na.rm = TRUE))) %>%
  replace_na(list(gross_sqr_ft=mean(.$gross_sqr_ft, na.rm = TRUE)))

head(property_df)
```

```
##     Borough gross_sqr_ft property_price property_level
## 1 Manhattan     5060.445              1              D
## 2 Manhattan     5060.445              1              D
## 3 Manhattan     5060.445              1              D
## 4 Manhattan     5060.445              1              D
## 5 Manhattan     5060.445              1              D
## 6 Manhattan     5060.445              1              D
```

Tidy Property Data

NYC Public School SAT Scores dataset will be labeled ‘school_df’, which contains information about SAT assessment performances that had occurred in recent years, including the reading, writing, and math scores, student demographic, borough, latitude, longitude, etc. The most important information we need in this dataset is the borough, SAT scores, and location (latitude, longitude). Afterwards, the data is manipulated as follows:

- First, we want to combine the average scores of each school’s SAT sections (reading, writing, and math) for our analysis. The cumulative score will be out of 2400.
- We will then rename for clarity
- We will create a new column that divides and categorizes a student’s SAT score. This will be used later in data visualization.
- All missing values are found in SAT scores, because some of these public schools do not require students to take the SAT, such as the school for the disabled. We will set their scores as the average.

```
school_df <- school_df %>%
  select(School.ID, School.Name, Borough, Latitude, Longitude, Average.Score..SAT.Math., Averag
e.Score..SAT.Reading., Average.Score..SAT.Writing.) %>%
  group_by(School.ID, School.Name, Borough, Latitude, Longitude) %>%
  summarize(sat_score = sum(Average.Score..SAT.Math., Average.Score..SAT.Writing., Average.Scor
e..SAT.Reading.)) %>%
  rename(
    school_id = School.ID,
    school_name = School.Name
  ) %>%
  mutate(score_level =
    ifelse(sat_score >= 0 && sat_score < 600, "D", ifelse(sat_score >= 600 && sat_score < 1200,
"C", ifelse(sat_score >= 1200 && sat_score < 1800, "B", ifelse(sat_score >= 1800 && sat_score <=
2400, "A", NA))))
  )
school_df <- school_df %>%
  replace_na(list(sat_score=mean(.$sat_score, na.rm = TRUE))) %>%
  replace_na(list(score_level='D'))
head(school_df)
```

```
## # A tibble: 6 x 7
## # Groups:   school_id, school_name, Borough, Latitude [6]
##   school_id school_name    Borough Latitude Longitude sat_score score_level
##   <fct>     <fct>          <fct>      <dbl>     <dbl>     <dbl> <chr>
## 1 01M292    Henry Street ~ Manhat~     40.7     -74.0      1197 C
## 2 01M448    University Ne~ Manhat~     40.7     -74.0      1144 C
## 3 01M450    East Side Com~ Manhat~     40.7     -74.0      1327 B
## 4 01M509    Marta Valle H~ Manhat~     40.7     -74.0      1245 B
## 5 01M539    New Explorati~ Manhat~     40.7     -74.0      1859 A
## 6 01M696    Bard High Sch~ Manhat~     40.7     -74.0      1914 A
```

Combining datasets

We will now combine both school_df and property_df via left_join. The join will matched by the ‘Borough’ column. Also, it is important to note that property prices have been filtered down to between 1,000 and 500,000 USD. The purpose for this is because of the wide range of housing cost that exists in New York City. As mentioned earlier, many property prices were $0 due to transferring ownership between family members. There were thousands of

that specific case. Also there were many outliers that existed beyond millions of dollars in property value. The difference in price is astronomical and will definitely harm the analysis. We are also taking a small sample of the dataset due to the sheer amount of entities that exists.

```
borough_df <- school_df %>%
  left_join(property_df, by='Borough') %>%
  group_by(Borough) %>%
  filter(property_price >= 1000 & property_price <= 500000) %>%
  sample_frac(.01)
borough_df
```

```
## # A tibble: 18,964 x 10
## # Groups:   Borough [5]
##    school_id school_name Borough Latitude Longitude sat_score score_level
##    <fct>     <fct>       <fct>      <dbl>     <dbl>     <dbl> <chr>
##  1 12X267    Bronx Latin Bronx       40.8     -73.9      1138 C
##  2 11X270    Academy fo~ Bronx       40.9     -73.9      1200 B
##  3 08X305    Pablo Neru~ Bronx       40.8     -73.9      1161 C
##  4 08X269    Bronx Stud~ Bronx       40.8     -73.9      1232 B
##  5 08X519    Felisa Rin~ Bronx       40.8     -73.9      1210 B
##  6 08X312    Millennium~ Bronx       40.8     -73.9      1204 B
##  7 10X243    West Bronx~ Bronx       40.9     -73.9      1165 C
##  8 10X368    In-Tech Ac~ Bronx       40.9     -73.9      1249 B
##  9 09X252    Mott Hall ~ Bronx       40.8     -73.9      1214 B
## 10 10X284    Bronx Scho~ Bronx       40.9     -73.9      1144 C
## # ... with 18,954 more rows, and 3 more variables: gross_sqr_ft <dbl>,
## #   property_price <dbl>, property_level <chr>
```

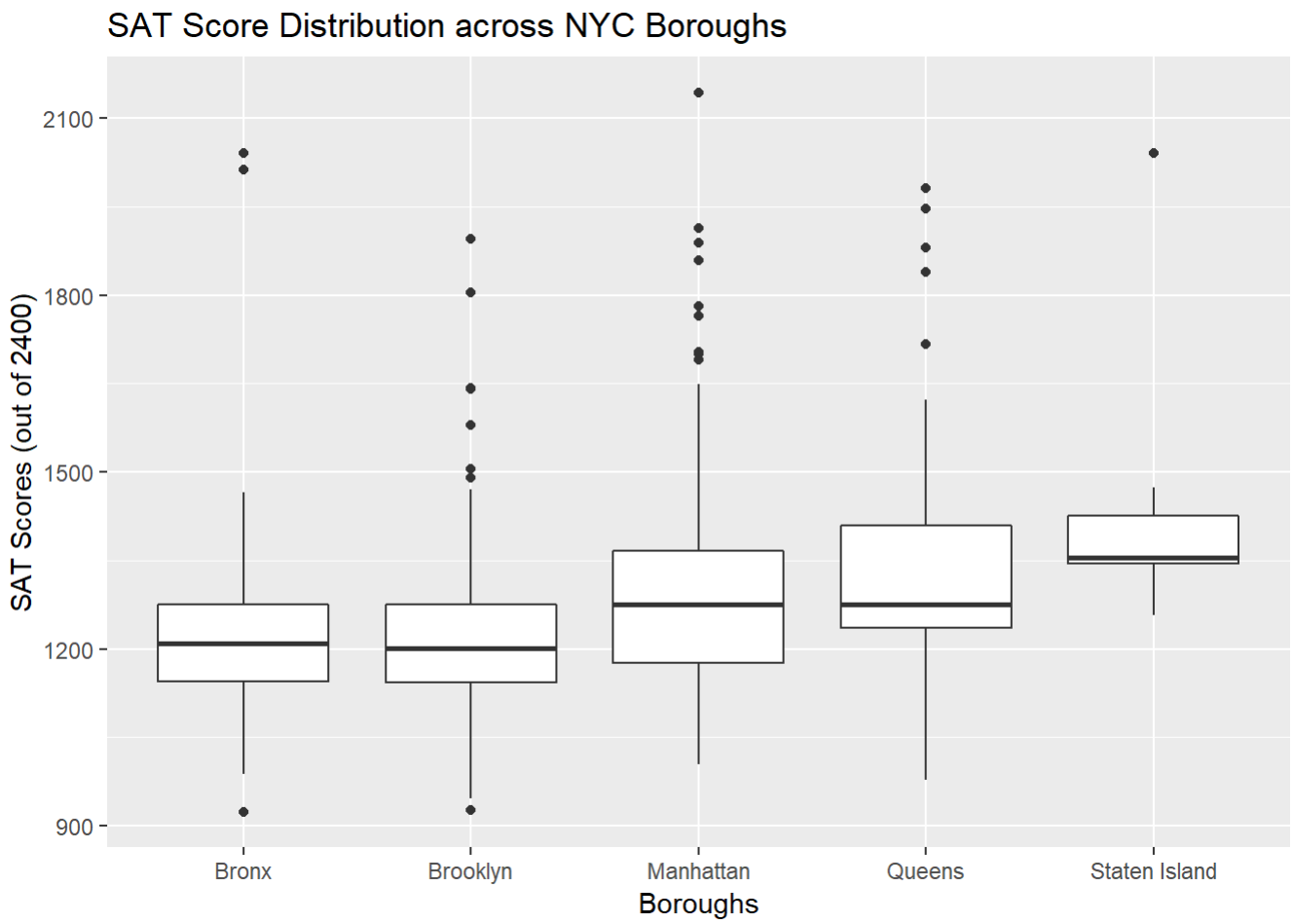
We are now ready for our EDA!

Exploratory Data Analysis (EDA)

SAT Score Distribution across NYC Boroughs

In this first plot, we want to see the data distribution across each NYC borough and observe difference in central tendency. The SAT score is scored out of 2400.

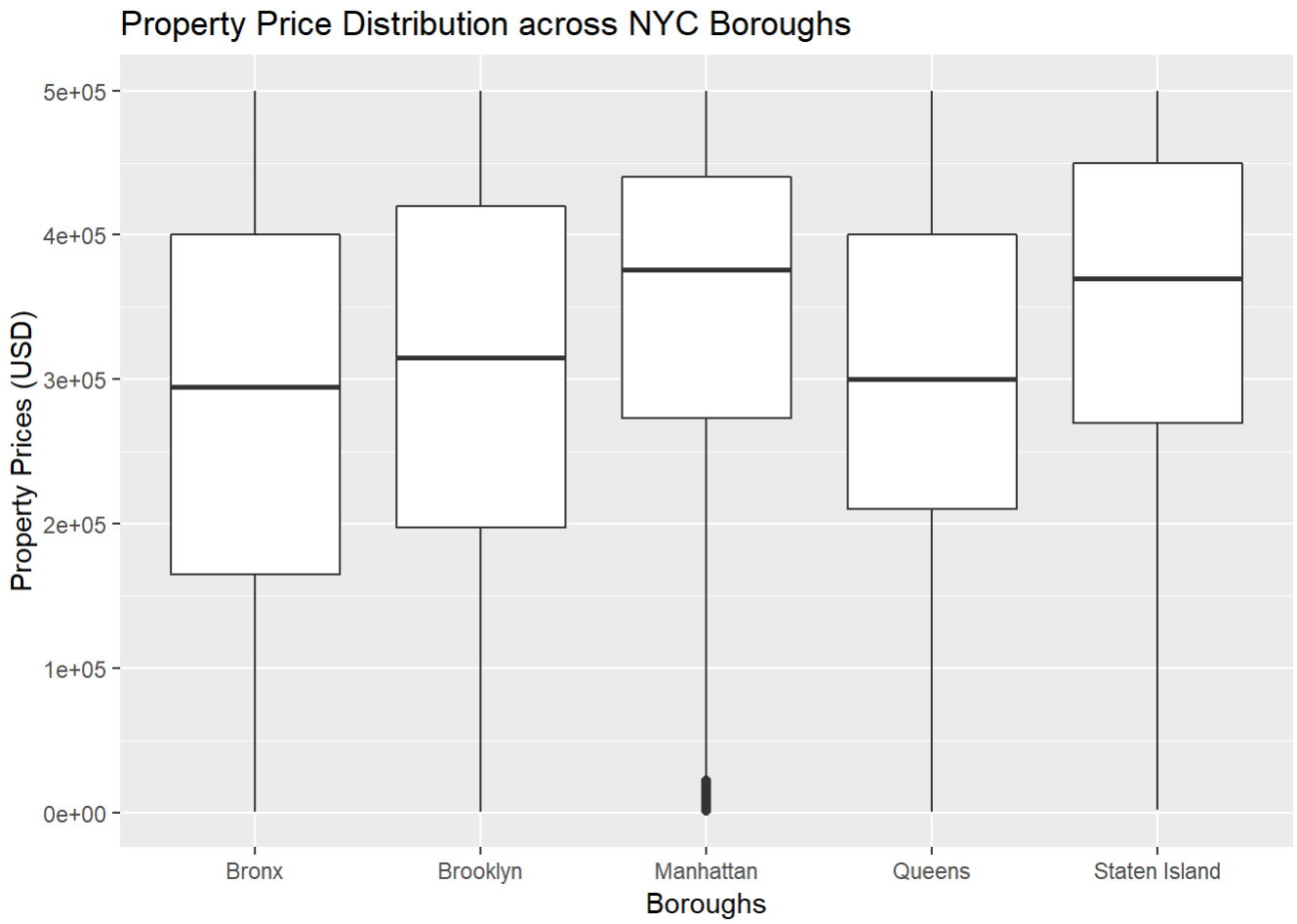
```
eda1_boxplot <- borough_df %>%
  rowid_to_column() %>%
  ggplot(mapping = aes(x=factor(Borough), y=sat_score)) +
  geom_boxplot() +
  labs(title="SAT Score Distribution across NYC Boroughs",
  x = "Boroughs",
  y = "SAT Scores (out of 2400)")
eda1_boxplot
```



We can see that there are outliers in every borough. In terms of highest scores when comparing the median, Staten Island comes first. The highest score overall goes to a school in Manhattan. Let us now look at the Property Prices Distribution across the NYC Boroughs…

Property Price Distribution across NYC Boroughs

```
eda1_boxplot <- borough_df %>%
  rowid_to_column() %>%
  ggplot(mapping = aes(x=factor(Borough), y=property_price)) +
  geom_boxplot() +
  labs(title="Property Price Distribution across NYC Boroughs",
  x = "Boroughs",
  y = "Property Prices (USD)")
eda1_boxplot
```

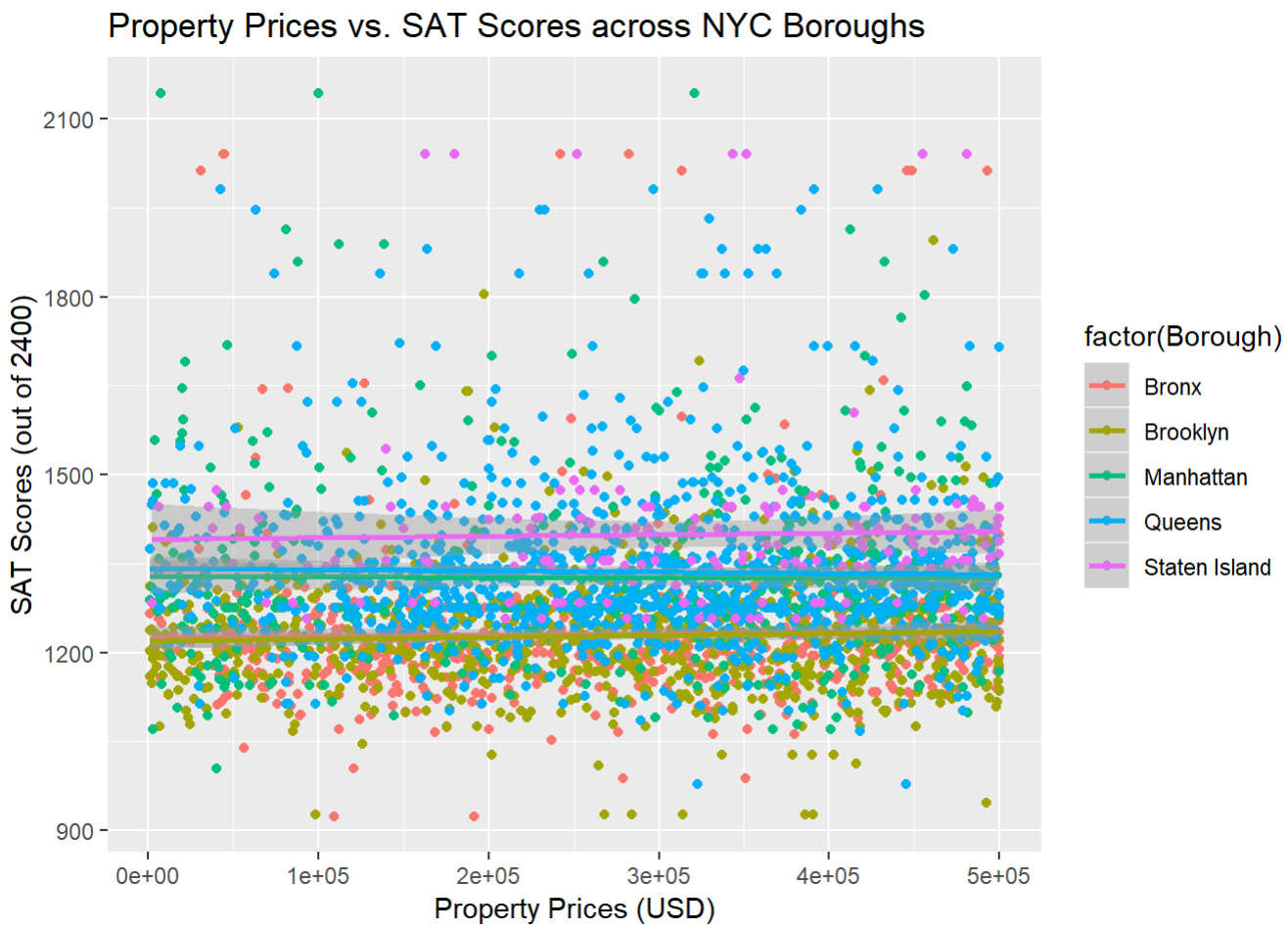


Based on these plots, we can see that the median property price of Manhattan is the highest compared to the rest, indicating that this borough is the wealthiest.

SAT Score Distribution vs. SAT Scores across NYC Boroughs

Now let's answer the million dollar question, is property price (wealth) directly correlated to SAT scores (academic performance)? We will now create a scatter plot for these two numeric attributes:

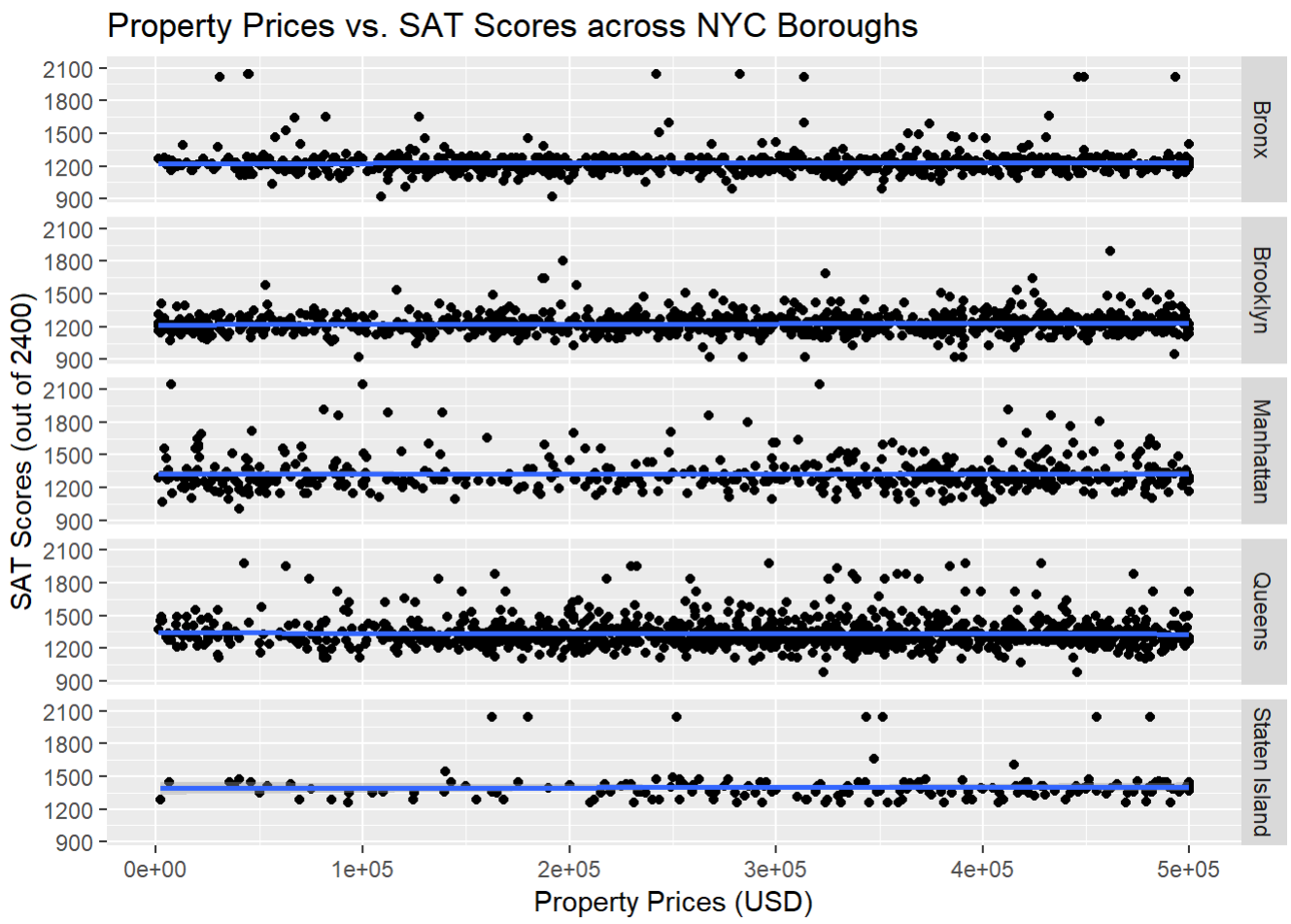
```
eda1_scatter <- borough_df %>%
  rowid_to_column() %>%
  group_by(Borough, property_price) %>%
  summarize(mean_score = mean(sat_score)) %>%
  ggplot(mapping=aes(y=mean_score, x=property_price, color=factor(Borough))) +
  geom_point() + geom_smooth(method=lm) +
  labs(title="Property Prices vs. SAT Scores across NYC Boroughs",
  x = "Property Prices (USD)",
  y = "SAT Scores (out of 2400)")
eda1_scatter
```



The graph above depicts all the data points of SAT scores and property prices, each point colored with their specific borough. This graph is ugly and overwhelming. Let's try to make it easier to look at and get a better grasp of the data we see.

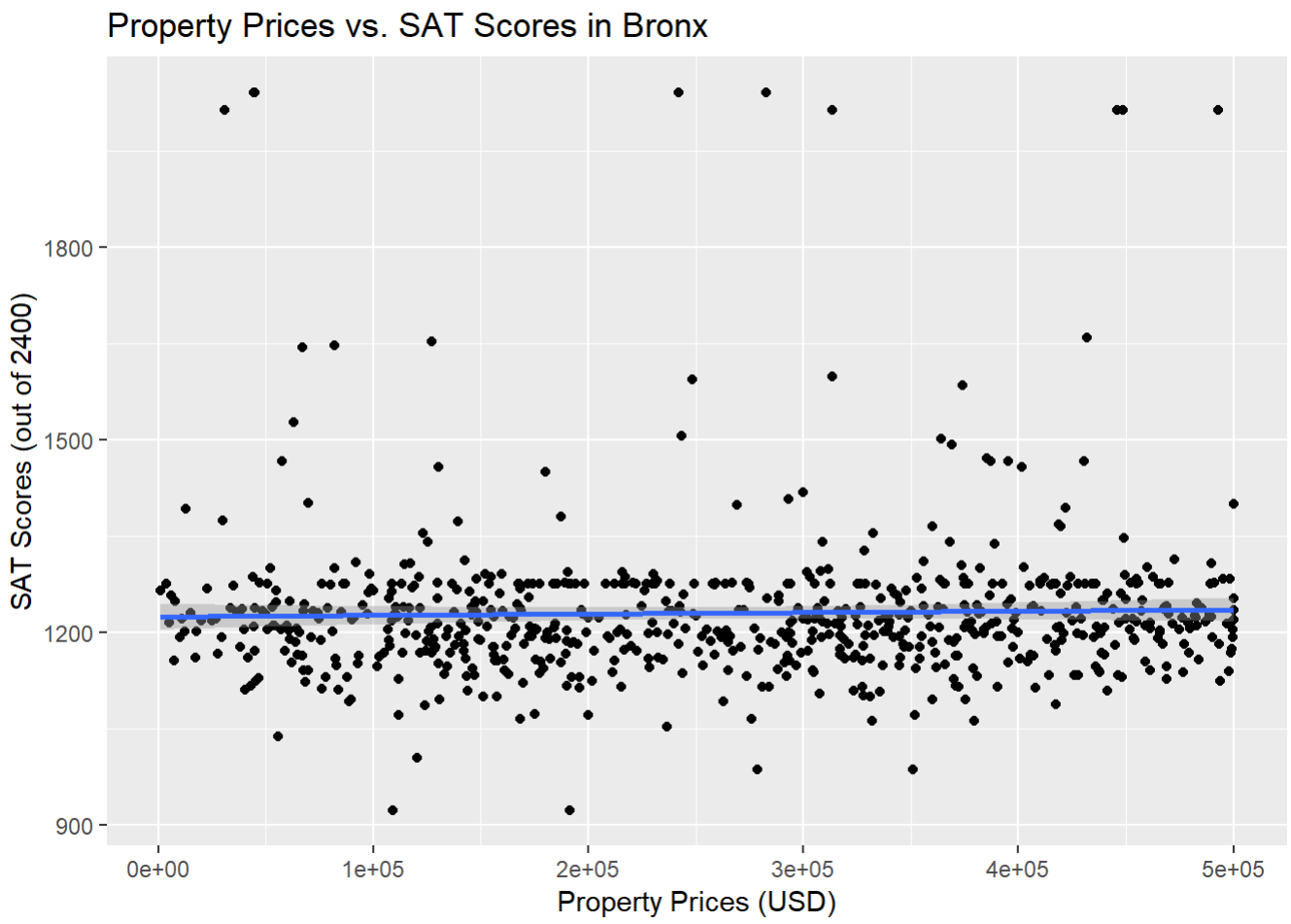
We will now plot the same graph. But this time, instead of color coding our boroughs, we will now facet our visualization. This means that we will create different graphs, splitting based on conditions we have set. In this case, we are splitting based on different boroughs.

```
eda1_scatter <- borough_df %>%
  rowid_to_column() %>%
  group_by(Borough, property_price) %>%
  summarize(mean_score = mean(sat_score)) %>%
  ggplot(mapping=aes(y=mean_score, x=property_price)) +
  facet_grid(Borough~.) +
  geom_point() + geom_smooth(method=lm) +
  labs(title="Property Prices vs. SAT Scores across NYC Boroughs",
  x = "Property Prices (USD)",
  y = "SAT Scores (out of 2400)")
eda1_scatter
```

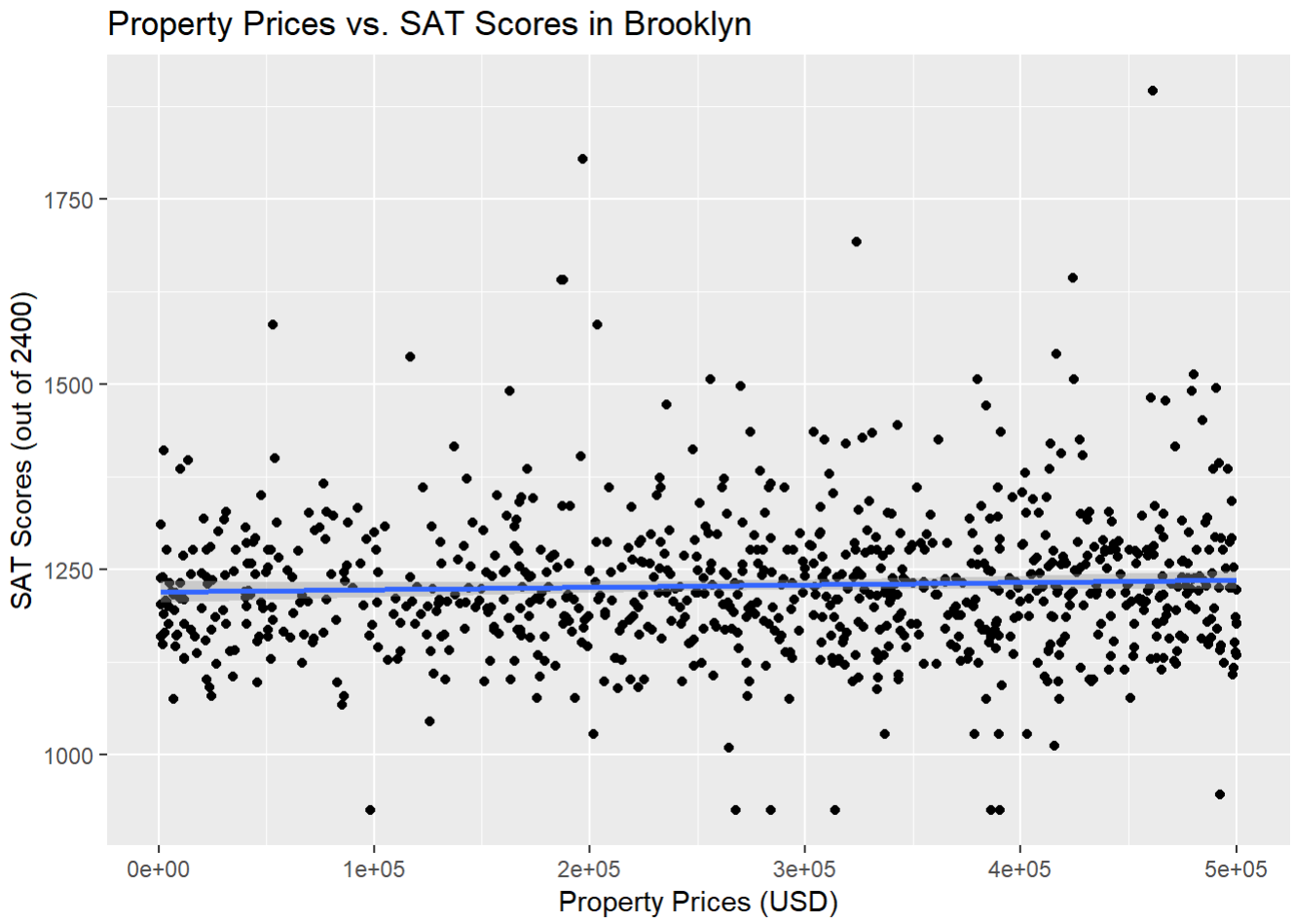


The graphs are now easier to look at. We will take it one step further and graph each borough individually. The reason is to examine the linear regression, and plot residual graphs to determine if the linear relationship is a good approximation.

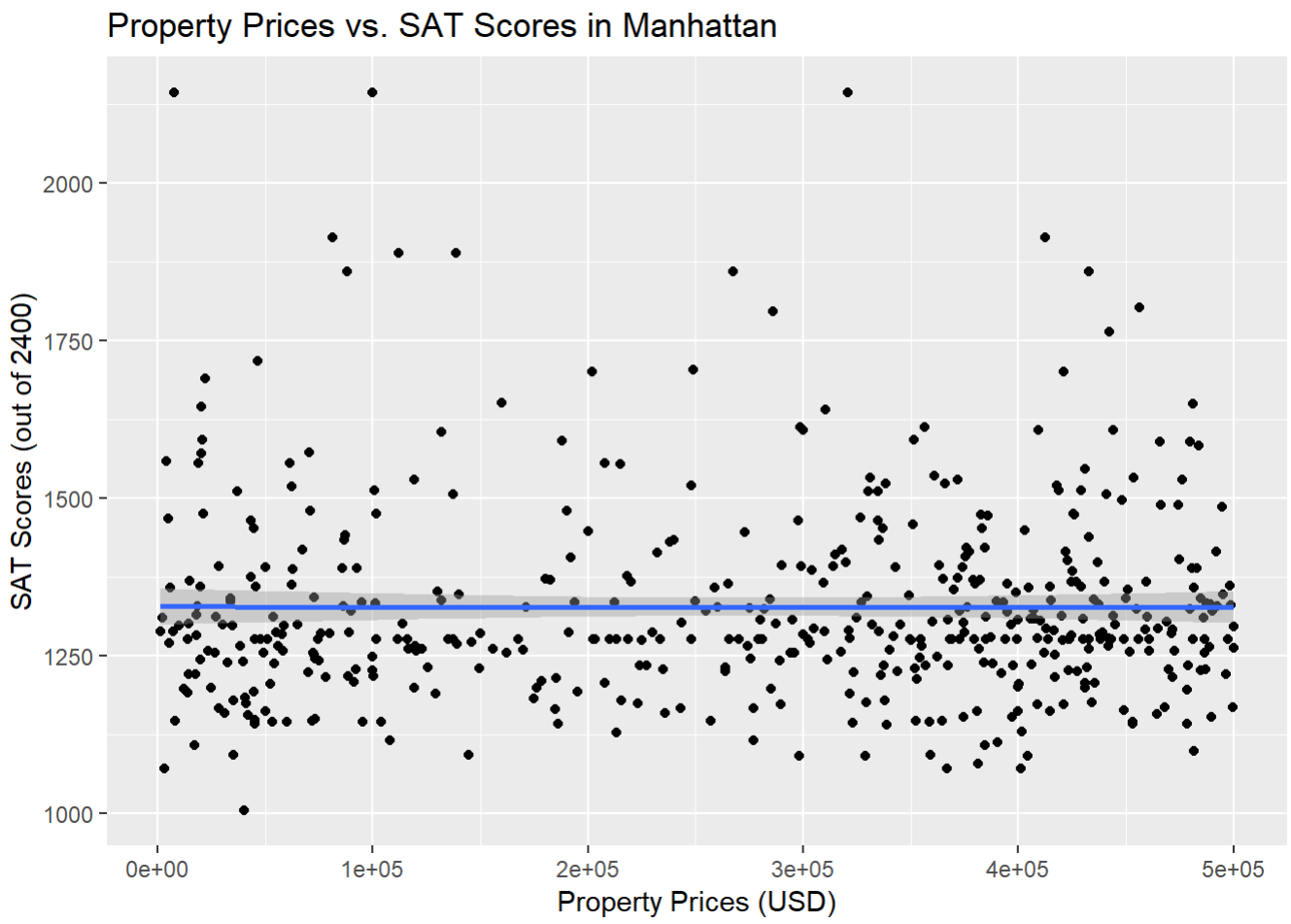
```
bronx_scatter <- borough_df %>%
  filter(Borough == 'Bronx') %>%
  rowid_to_column() %>%
  group_by(Borough, property_price) %>%
  summarize(mean_score = mean(sat_score)) %>%
  ggplot(mapping=aes(y=mean_score, x=property_price)) +
  geom_point() + geom_smooth(method=lm) +
  labs(title="Property Prices vs. SAT Scores in Bronx",
  x = "Property Prices (USD)",
  y = "SAT Scores (out of 2400)")
bronx_scatter
```



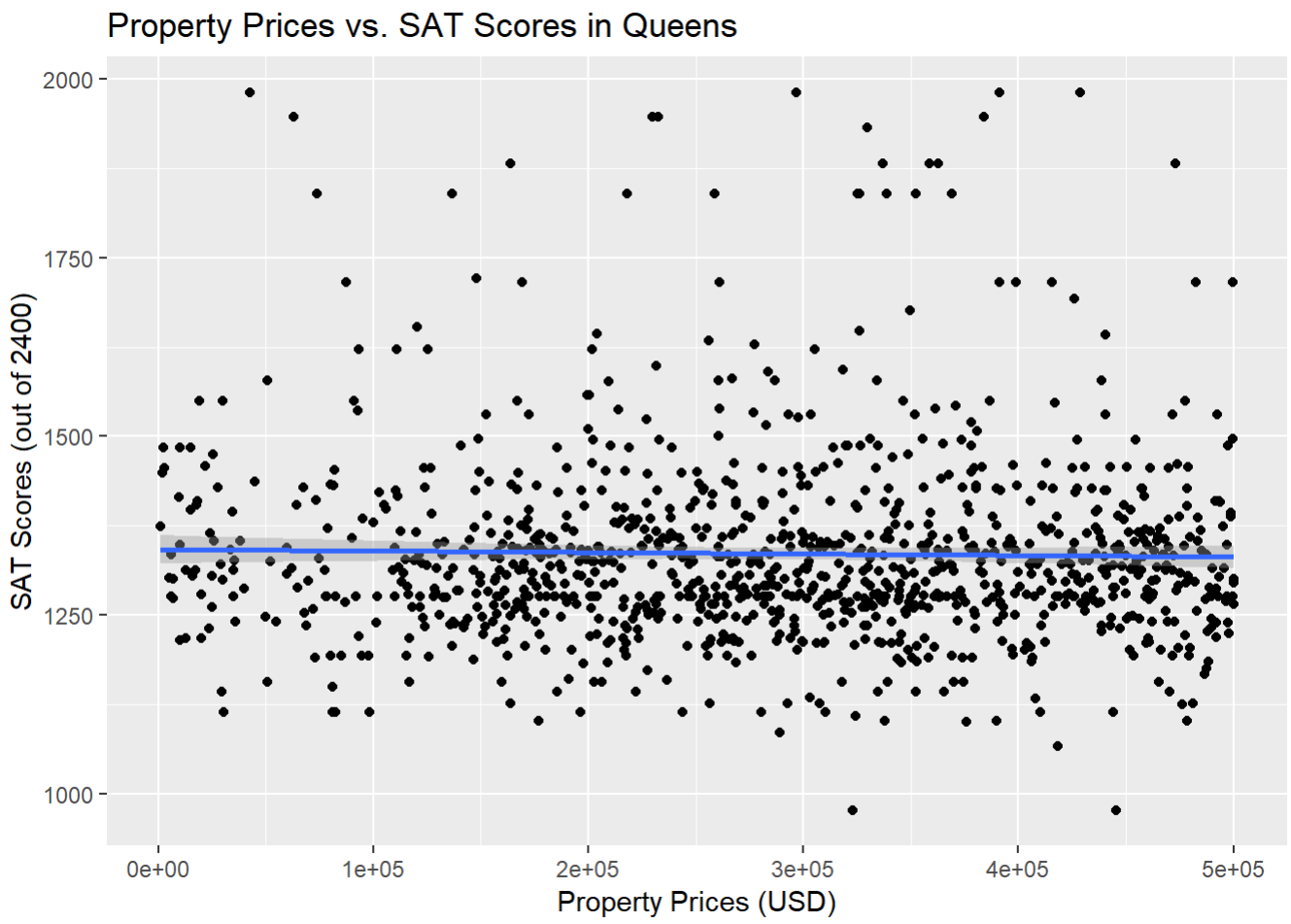

```
brooklyn_scatter <- borough_df %>%
  filter(Borough == 'Brooklyn') %>%
  rowid_to_column() %>%
  group_by(Borough, property_price) %>%
  summarize(mean_score = mean(sat_score)) %>%
  ggplot(mapping=aes(y=mean_score, x=property_price)) +
  geom_point() + geom_smooth(method=lm) +
  labs(title="Property Prices vs. SAT Scores in Brooklyn",
  x = "Property Prices (USD)",
  y = "SAT Scores (out of 2400)")
brooklyn_scatter
```



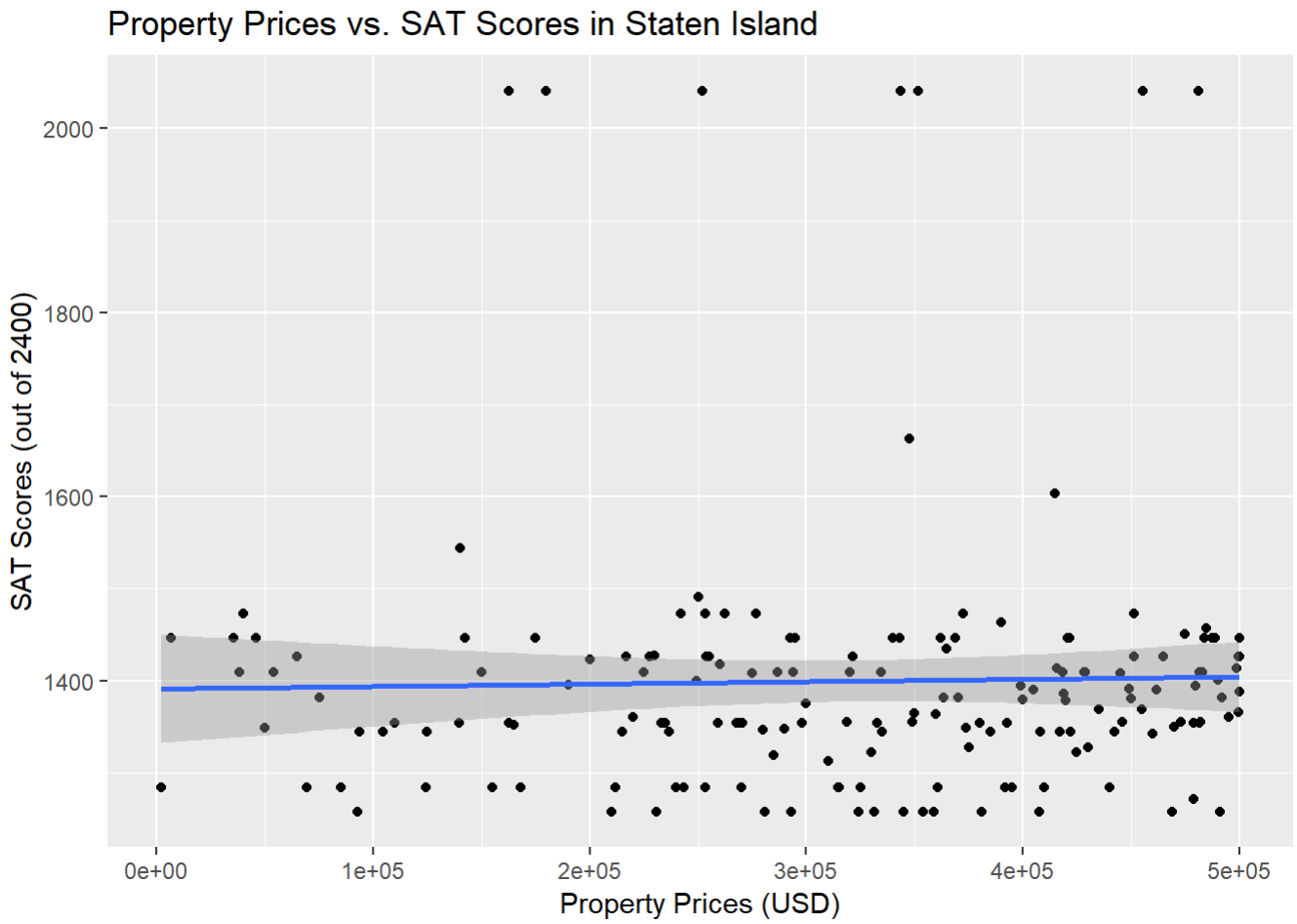

```
manhattan_scatter <- borough_df %>%
  filter(Borough == 'Manhattan') %>%
  rowid_to_column() %>%
  group_by(Borough, property_price) %>%
  summarize(mean_score = mean(sat_score)) %>%
  ggplot(mapping=aes(y=mean_score, x=property_price)) +
  geom_point() + geom_smooth(method=lm) +
  labs(title="Property Prices vs. SAT Scores in Manhattan",
  x = "Property Prices (USD)",
  y = "SAT Scores (out of 2400)")
manhattan_scatter
```




```
queens_scatter <- borough_df %>%
  filter(Borough == 'Queens') %>%
  rowid_to_column() %>%
  group_by(Borough, property_price) %>%
  summarize(mean_score = mean(sat_score)) %>%
  ggplot(mapping=aes(y=mean_score, x=property_price)) +
  geom_point() + geom_smooth(method=lm) +
  labs(title="Property Prices vs. SAT Scores in Queens",
  x = "Property Prices (USD)",
  y = "SAT Scores (out of 2400)")
queens_scatter
```




```
staten_scatter <- borough_df %>%
  filter(Borough == 'Staten Island') %>%
  rowid_to_column() %>%
  group_by(Borough, property_price) %>%
  summarize(mean_score = mean(sat_score)) %>%
  ggplot(mapping=aes(y=mean_score, x=property_price)) +
  geom_point() + geom_smooth(method=lm) +
  labs(title="Property Prices vs. SAT Scores in Staten Island",
  x = "Property Prices (USD)",
  y = "SAT Scores (out of 2400)")
staten_scatter
```



There appears to be a gradual linear relationship. The SAT score moves slightly upward the higher the property prices go. There seems to be high residuals in all the plots.

Residual Graphs

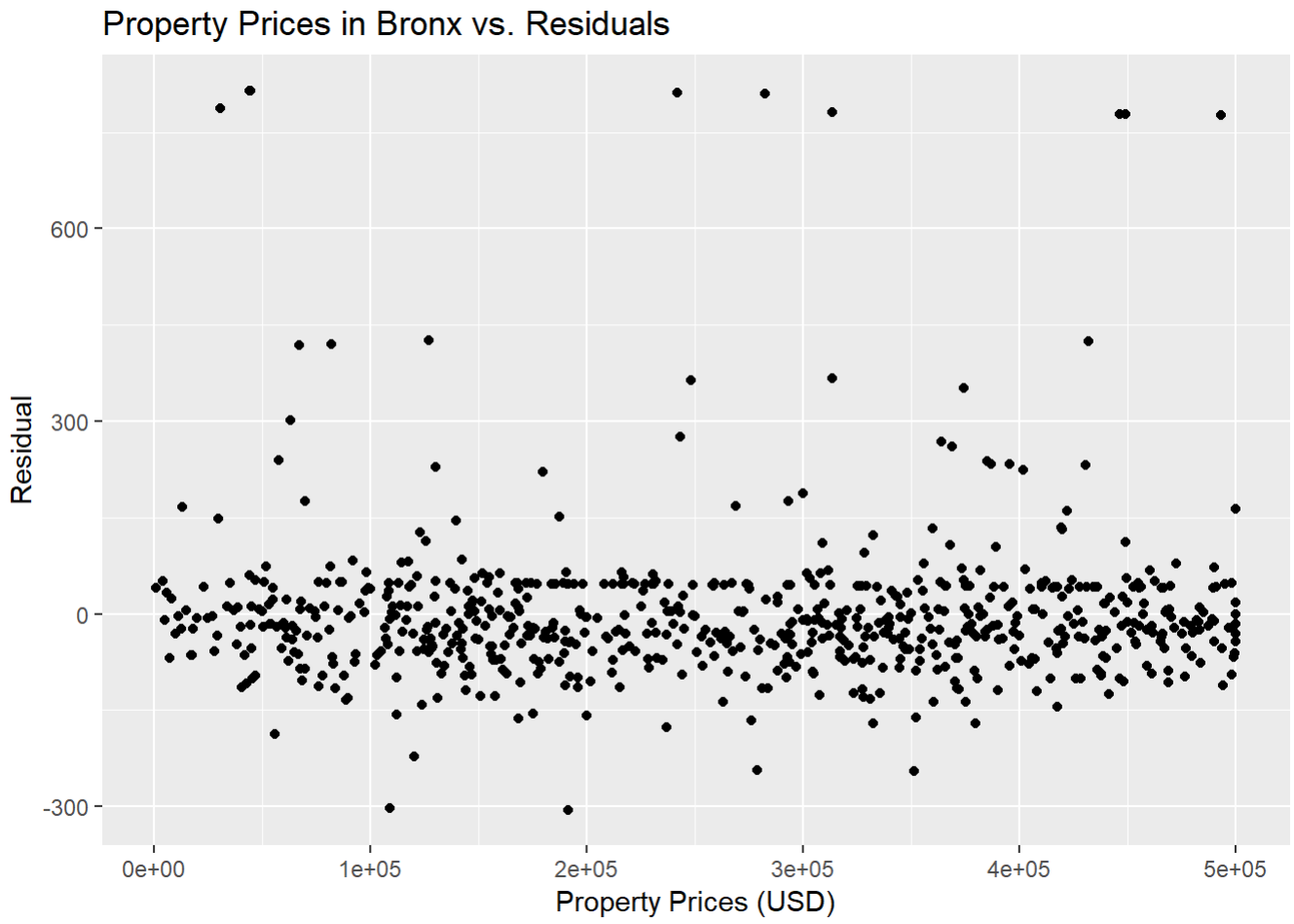
We will confirm that the linear relationship is a good approximation by plotting the residuals against property prices for each borough.

```
bronx_data <- borough_df %>%
  filter(Borough == 'Bronx') %>%
  rowid_to_column() %>%
  group_by(Borough, property_price) %>%
  summarize(mean_score = mean(sat_score))

bronx_res <- lm(formula = mean_score~property_price, data = bronx_data)
tidy(bronx_res)
```

```
## # A tibble: 2 x 5
##   term              estimate  std.error statistic p.value
##   <chr>                <dbl>      <dbl>     <dbl>   <dbl>
## 1 (Intercept)    1225.        10.00        123.      0
## 2 property_price    0.0000219  0.0000338     0.648   0.517
```

```
bronx_res <- bronx_res %>%
  augment() %>%
  ggplot(mapping = aes(x=property_price, y=.resid)) +
  geom_point() +
  labs(title="Property Prices in Bronx vs. Residuals",
  x = "Property Prices (USD)",
  y = "Residual")
bronx_res
```

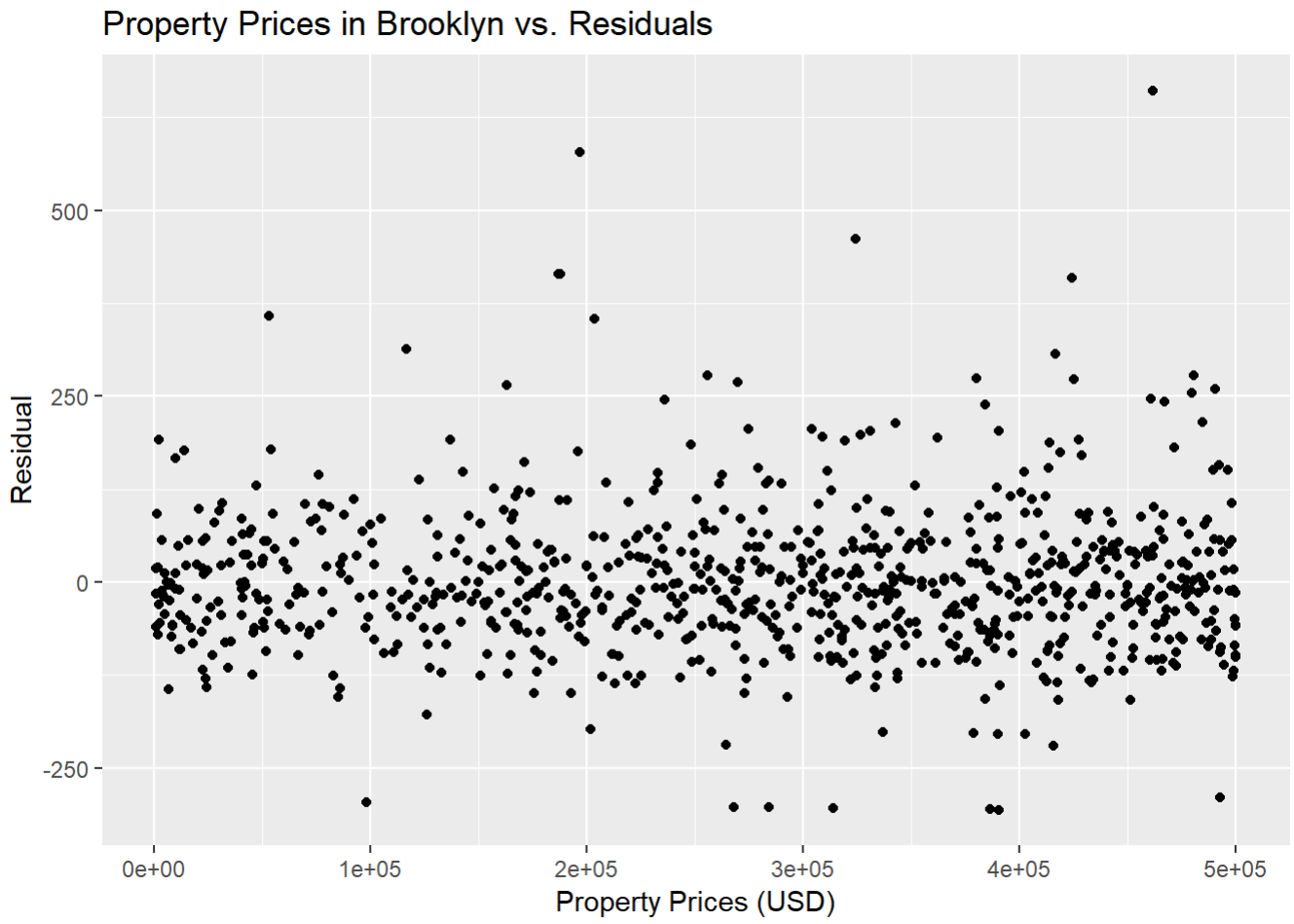



```
brooklyn_data <- borough_df %>%
  filter(Borough == 'Brooklyn') %>%
  rowid_to_column() %>%
  group_by(Borough, property_price) %>%
  summarize(mean_score = mean(sat_score))

brooklyn_res <- lm(formula = mean_score~property_price, data = brooklyn_data)
tidy(brooklyn_res)
```

```
## # A tibble: 2 x 5
##   term              estimate std.error statistic p.value
##   <chr>                <dbl>     <dbl>     <dbl>   <dbl>
## 1 (Intercept)    1220.        7.26         168.     0
## 2 property_price    0.0000323 0.0000233      1.39   0.165
```

```
brooklyn_res <- brooklyn_res %>%
  augment() %>%
  ggplot(mapping = aes(x=property_price, y=.resid)) +
  geom_point() +
  labs(title="Property Prices in Brooklyn vs. Residuals",
  x = "Property Prices (USD)",
  y = "Residual")
brooklyn_res
```

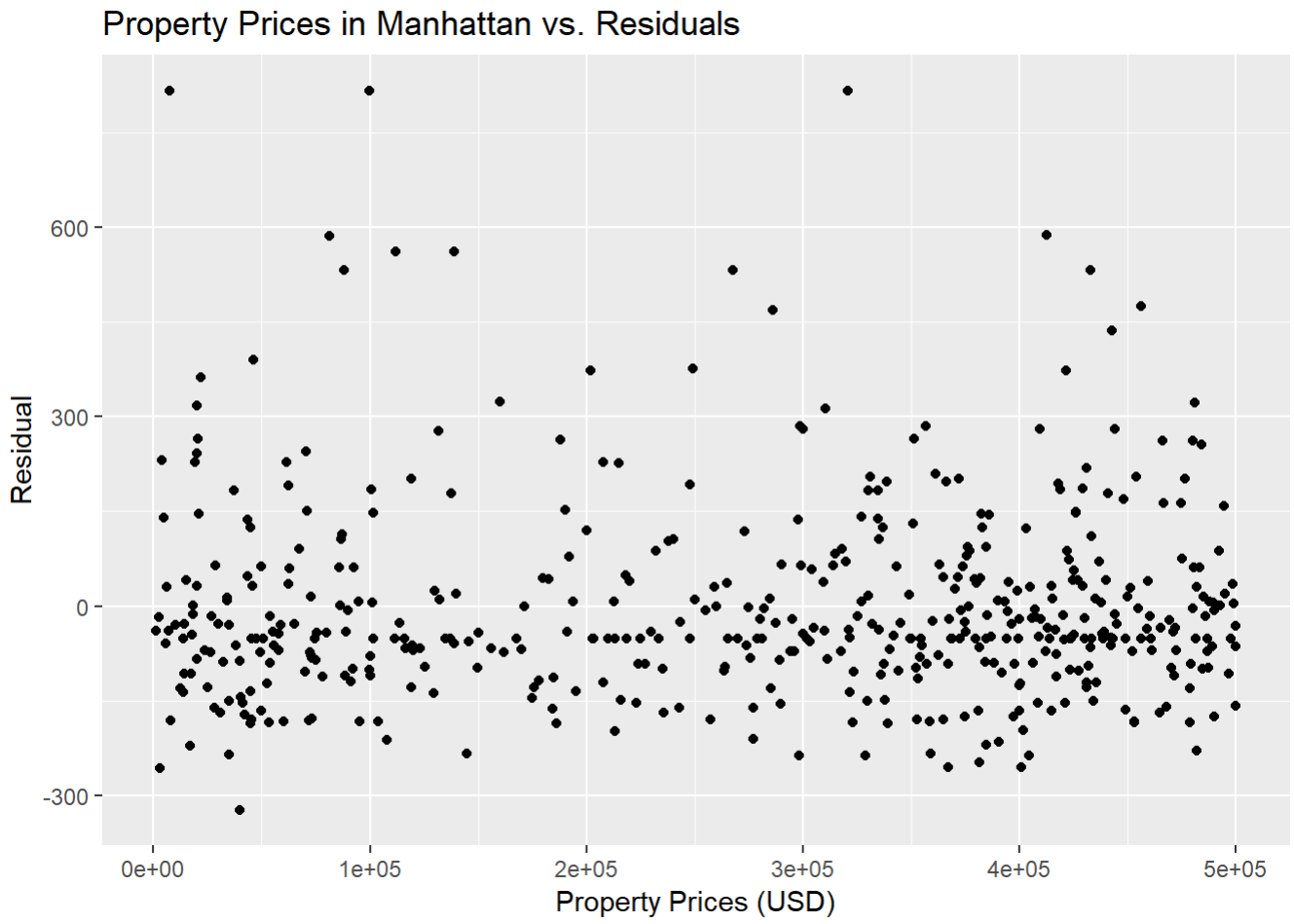



```
manhattan_data <- borough_df %>%
  filter(Borough == 'Manhattan') %>%
  rowid_to_column() %>%
  group_by(Borough, property_price) %>%
  summarize(mean_score = mean(sat_score))

manhattan_res <- lm(formula = mean_score~property_price, data = manhattan_data)
tidy(manhattan_res)
```

```
## # A tibble: 2 x 5
##   term           estimate  std.error statistic   p.value
##   <chr>             <dbl>      <dbl>     <dbl>     <dbl>
## 1 (Intercept)     1.33e+3 14.6         91.1    7.88e-306
## 2 property_price -1.65e-6  0.0000462   -0.0358 9.71e-  1
```

```
manhattan_res <- manhattan_res %>%
  augment() %>%
  ggplot(mapping = aes(x=property_price, y=.resid)) +
  geom_point() +
  labs(title="Property Prices in Manhattan vs. Residuals",
  x = "Property Prices (USD)",
  y = "Residual")
manhattan_res
```

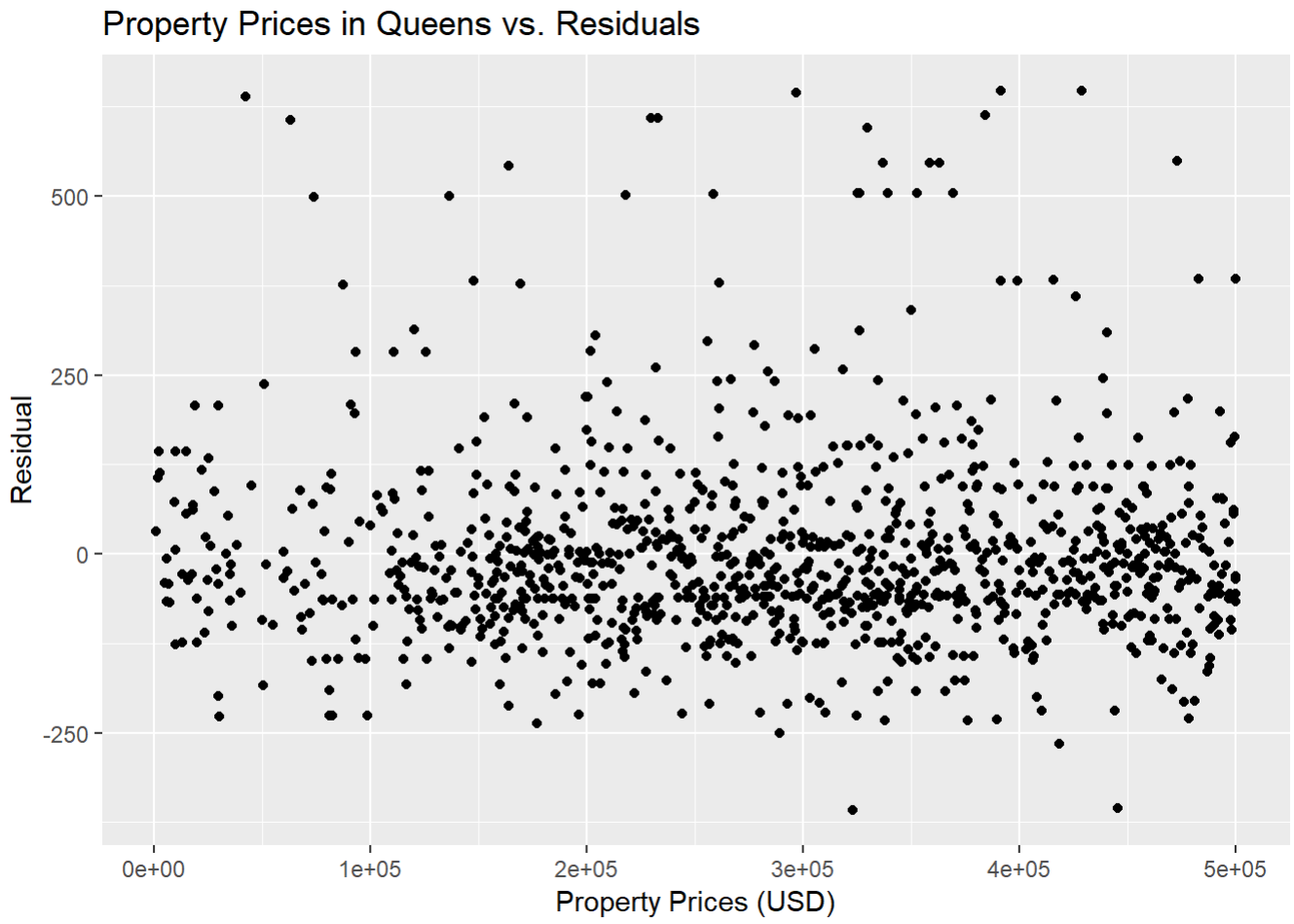



```
queens_data <- borough_df %>%
  filter(Borough == 'Queens') %>%
  rowid_to_column() %>%
  group_by(Borough, property_price) %>%
  summarize(mean_score = mean(sat_score))

queens_res <- lm(formula = mean_score~property_price, data = queens_data)
tidy(queens_res)
```

```
## # A tibble: 2 x 5
##   term                estimate  std.error statistic p.value
##   <chr>                  <dbl>      <dbl>     <dbl>   <dbl>
## 1 (Intercept)     1342.        10.0          134.      0
## 2 property_price    -0.0000210  0.0000320    -0.656   0.512
```

```
queens_res <- queens_res %>%
  augment() %>%
  ggplot(mapping = aes(x=property_price, y=.resid)) +
  geom_point() +
  labs(title="Property Prices in Queens vs. Residuals",
  x = "Property Prices (USD)",
  y = "Residual")
queens_res
```

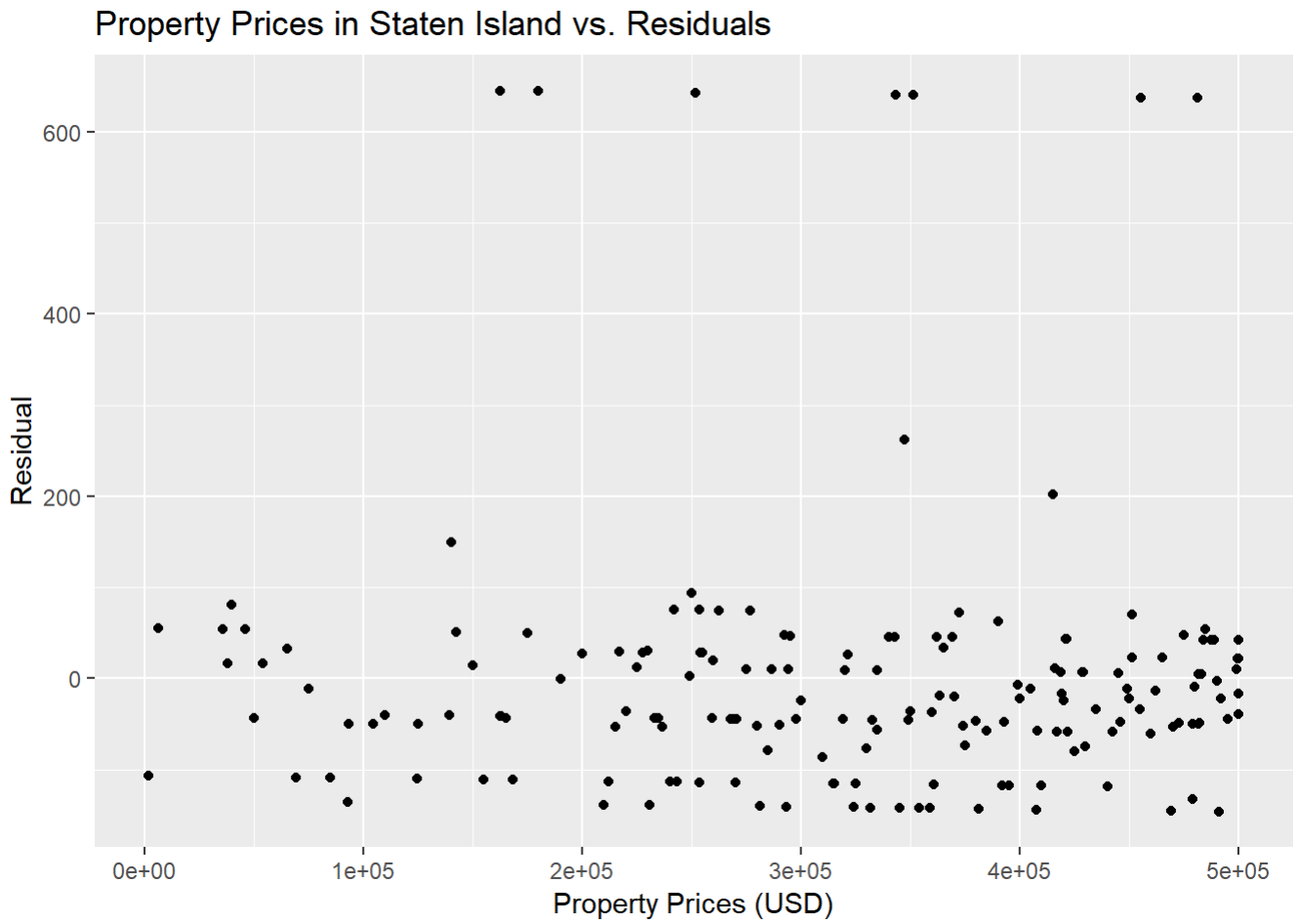



```
staten_data <- borough_df %>%
  filter(Borough == 'Staten Island') %>%
  rowid_to_column() %>%
  group_by(Borough, property_price) %>%
  summarize(mean_score = mean(sat_score))

staten_res <- lm(formula = mean_score~property_price, data = staten_data)
tidy(staten_res)
```

```
## # A tibble: 2 x 5
##   term              estimate  std.error statistic  p.value
##   <chr>                <dbl>      <dbl>     <dbl>    <dbl>
## 1 (Intercept)    1391.        30.0          46.4   3.25e-99
## 2 property_price    0.0000261  0.0000871     0.300 7.65e- 1
```

```
staten_res <- staten_res %>%
  augment() %>%
  ggplot(mapping = aes(x=property_price, y=.resid)) +
  geom_point() +
  labs(title="Property Prices in Staten Island vs. Residuals",
  x = "Property Prices (USD)",
  y = "Residual")
staten_res
```

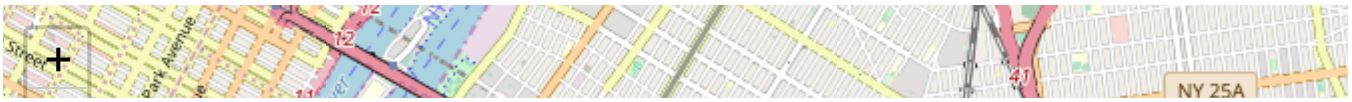


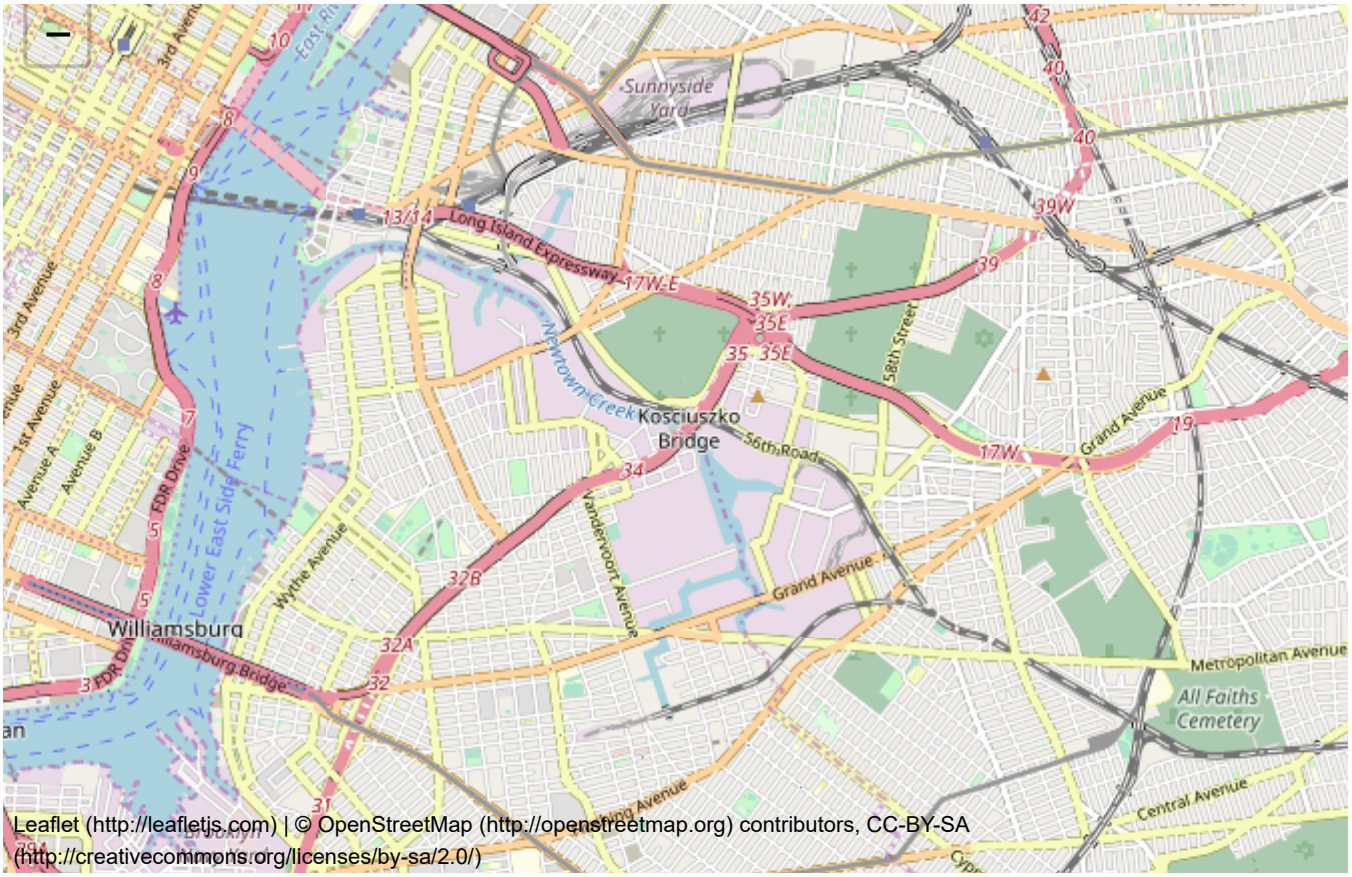
At a glance, we can see that the value is centered around 0 and no correlation between residual and property prices. We can conclude that the linear model is appropriate. Also we can see that Bronx and Queens boroughs experiences very slight increase, while the other boroughs experiences very slight decreases in property price.

Leaflet

In this section, we will create visualizations of our data point using Leaflet. First we will set up our latitude and longitude on NYC.

```
nyc_map <- leaflet(borough_df) %>%
 addTiles() %>%
 setView(lat=40.73, lng=-73.93, zoom=13)
nyc_map
```

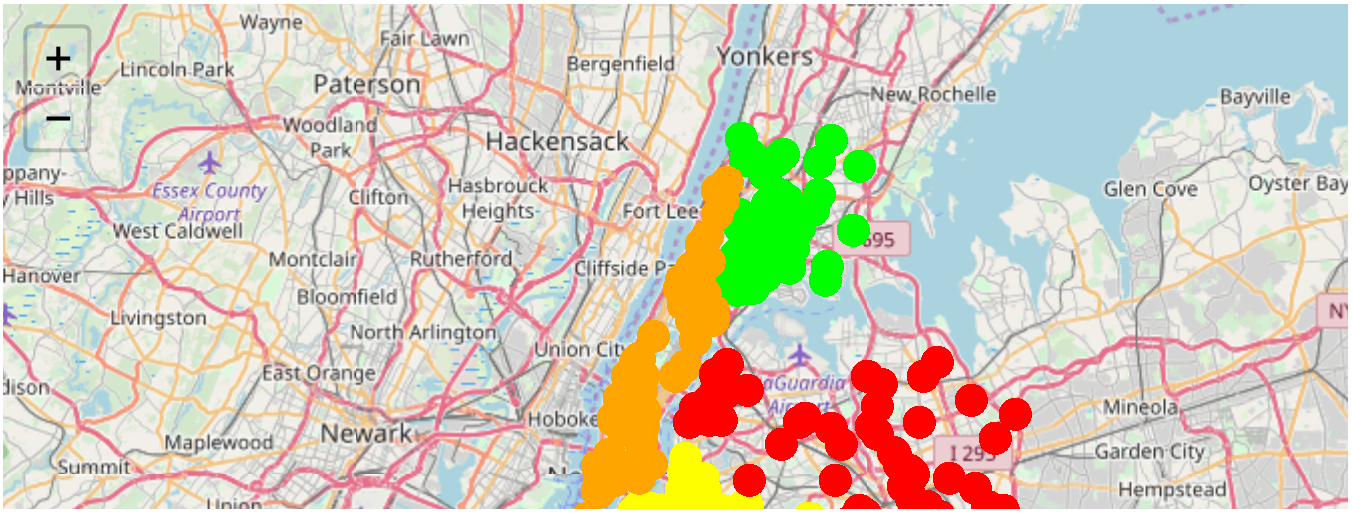


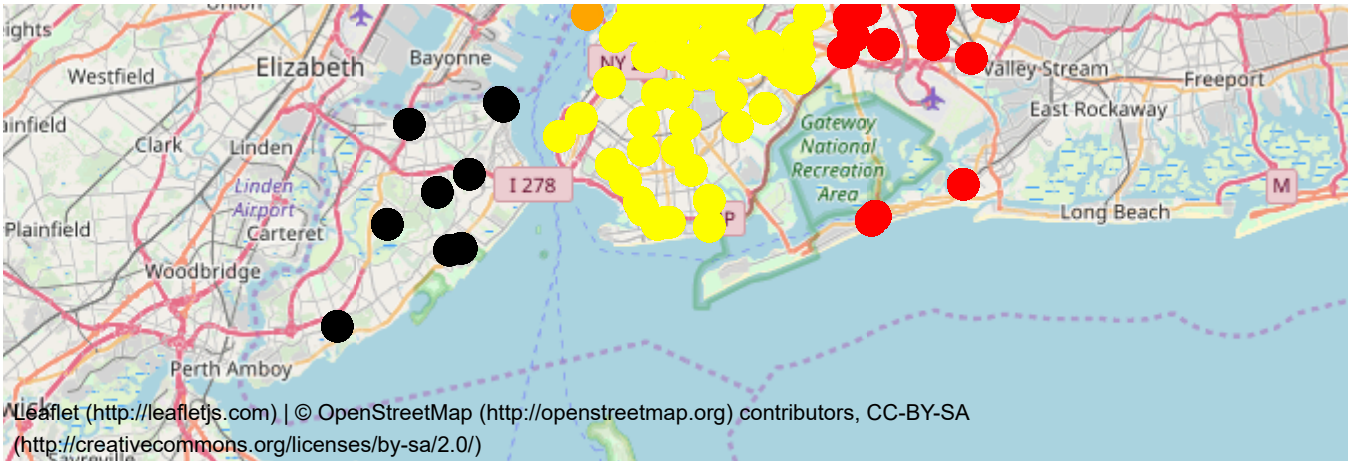


We are now ready to input data. In this first implementation, we will simply mark all data points based on boroughs.

```
pal <- colorFactor(c("green","yellow", "orange", "red", "black"), domain = c("Bronx", "Brooklyn"
, "Manhattan", "Queens", "Staten Island"))
leaflet(borough_df) %>% addTiles() %>%
 addCircleMarkers(
 radius = 8,
 color = ~pal(Borough),
 stroke = FALSE, fillOpacity = 0.8
 )
```

```
## Assuming "Longitude" and "Latitude" are longitude and latitude, respectively
```

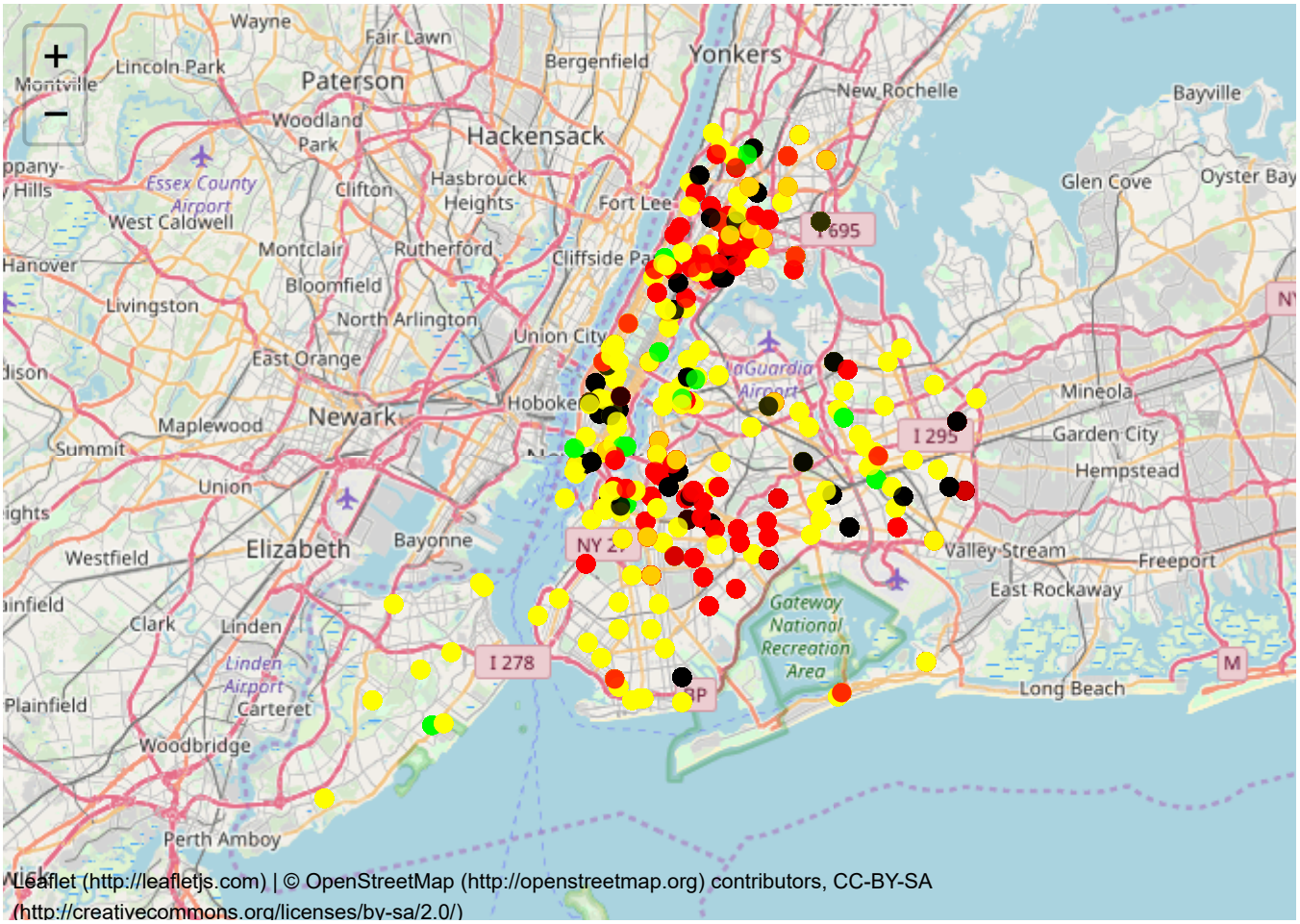




Remember the score_level we made in the beginning of the tutorial? We will use that column to color code the various SAT scores in NYC.

```
pal <- colorFactor(c("green","yellow", "red", "black"), domain = c("A", "B", "C", "D"))
leaflet(borough_df) %>% addTiles() %>%
 addCircleMarkers(
 radius = 5,
 color = ~pal(score_level),
 stroke = FALSE, fillOpacity = 0.8
 )
```

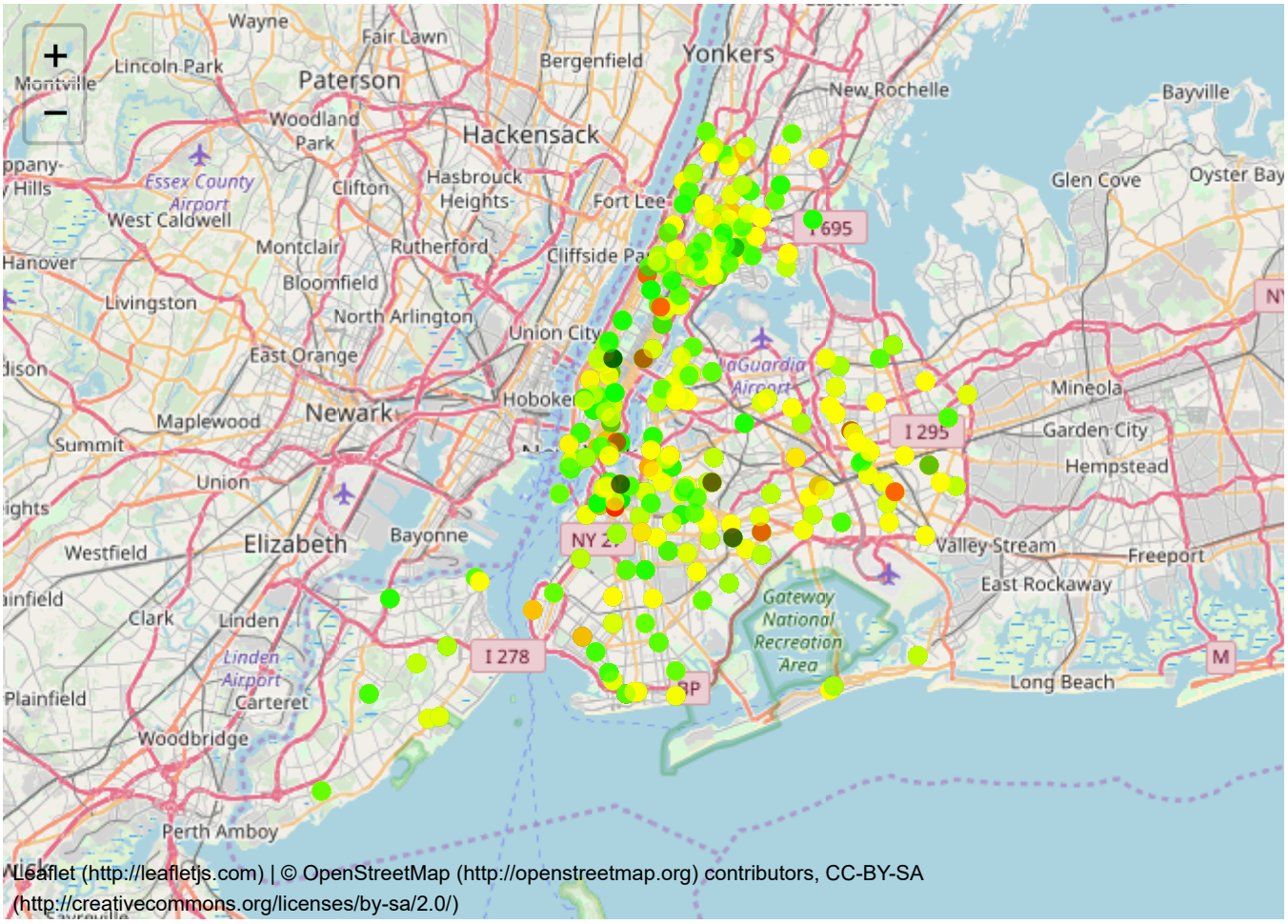
```
## Assuming "Longitude" and "Latitude" are longitude and latitude, respectively
```



Just like the score_level, property_level will be used for visualization, categorizing the properties based on grades determined at the beginning of this tutorial.

```
pal <- colorFactor(c("green","yellow", "red", "black"), domain = c("A", "B", "C", "D"))
leaflet(borough_df) %>% addTiles() %>%
 addCircleMarkers(
 radius = 5,
 color = ~pal(property_level),
 stroke = FALSE, fillOpacity = 0.6
 )
```

```
## Assuming "Longitude" and "Latitude" are longitude and latitude, respectively
```



Discussion

OVerall, the correlation between property price and SAT score is very slight. After plotting different graphs, such as facets, individual plots, and data visualizations with leaflet, many of these plots indicate little to no correlation. However, the residual graph confirms that the linear model is a good fit. There are many aspects involved with the conclusion as to why property prices has little impact to academic performance. One possibility is the cost of living in New York City is relatively higher than other places in the country. This might mean that the public school system in a relatively wealthy area can provide further educational assistance.

Another possible reason as to the small trends we observed could be the property price. Many of the prices were 0 USD, meaning that there were shifts in ownership free of charge. Also, there was a huge disparity in living expense. Some of the prices go up in the millions. To prevent harm in the analysis, we filtered the price and set all 0 values to the average of the filtered price. By removing parts of the price range, we might not be seeing the complete picture.

Any further research that would like to find the correlation between individual wealth and academic performance should take these possibilities into consideration.

This concludes the analysis between SAT scores and Property Sales in NYC.

Datasets Credits:

[Average SAT Scores for NYC Public Schools dataset] was compiled and published by the New York City Department of Education, and the SAT score averages and testing rates were provided by the College Board. https://www.kaggle.com/nycopendata/high-schools (https://www.kaggle.com/nycopendata/high-schools)

[NYC Property Sales dataset] is a concatenated and slightly cleaned-up version of the New York City Department of Finance's Rolling Sales dataset. https://www.kaggle.com/new-york-city/nyc-property-sales (https://www.kaggle.com/new-york-city/nyc-property-sales)