Amber Pasiak

INFO – 640

Predictive Data Analysis

11/17/19

For this project I used Street Easy Data about apartments in New York City. The street easy data had 20 different variables to choose from. These ranged from information about the apartment, such as age of the building, number of rooms, square footage, and amenities such as door man, patio, washer/dryer and elevator. Street Easy keeps a record of such data publicly available via their website. I first ran a summary of the data to get a better idea of the information. The data was split up into the three New York City boroughs of Brooklyn, Manhattan and Queens, with over 10,000 listings making up the dataset. I found some things interesting from this summary, such as an apartment on the 83 floor of a building, there was only 1 no-fee listing and only 1 apartment with an elevator.

This obviously is a limitation to the dataset, as when you do a simple search on the Street Easy website there are many options that have an elevator or are no-fee. This dataset, similar to others, is a snapshot of a specific time. I was unable to find the year or amount of time with which this data was collected, but that could be an interesting way to compare rent over time.

I wanted to see if there was a correlation between the price of rent and the distance to a subway stop. I choose to narrow this down by restricting the data to residents only in Brooklyn. From my personal experience apartment hunting, I had felt that the closer to a subway stop, the more expensive the apartment was. I wanted to see if this was true from the data available. I assume that most people also believe this to be true. Before diving deeper into the dataset, I did a quick search to see what the popular opinion was. I was shocked to see that the opposite was actually being argued. In a 2018 article from Metro-News, Kristin Toussaint claims that rent is going down closer to subway stops. She backs up this claim by using RentHop data that shows the average mean of one-bedroom apartments located within .6 miles of a subway stop. As I do not know when the street easy data was collected, any correlation I find cannot accurately dispute this. There is also the issue relating to distance, the Street Easy data uses walking time to a subway, which is a subjective variable.

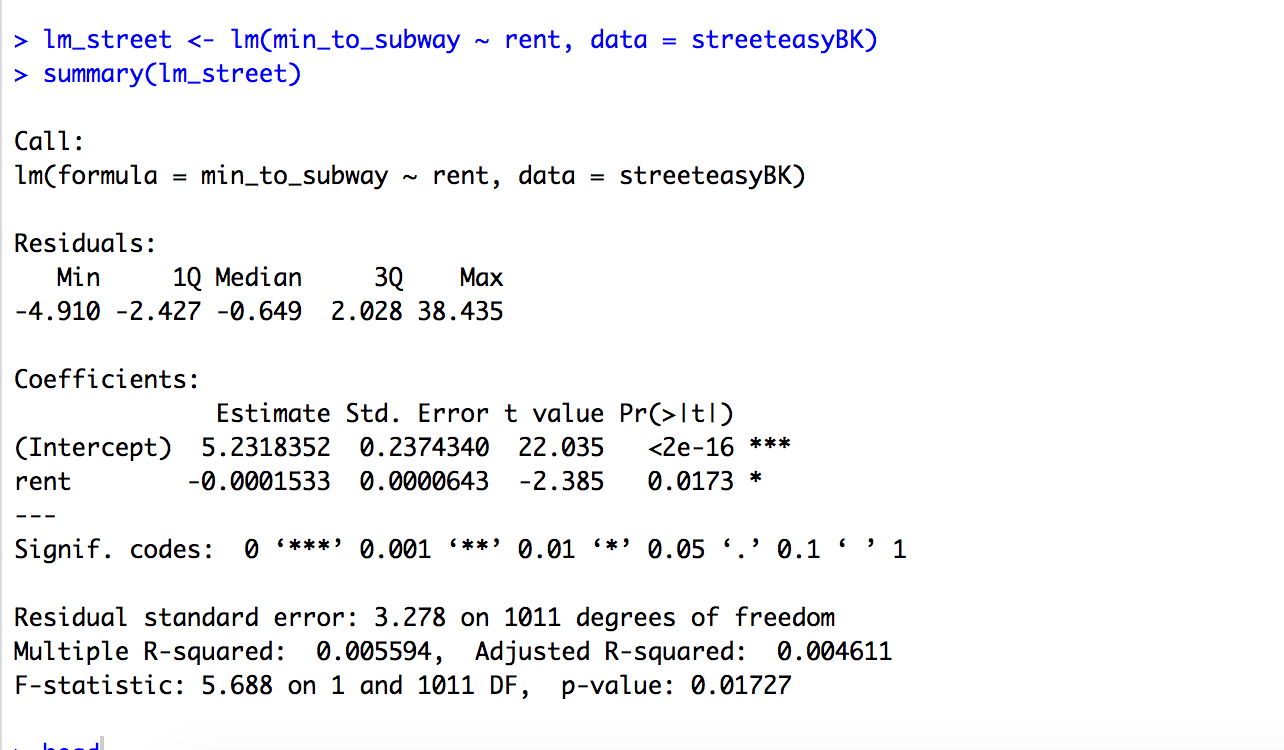
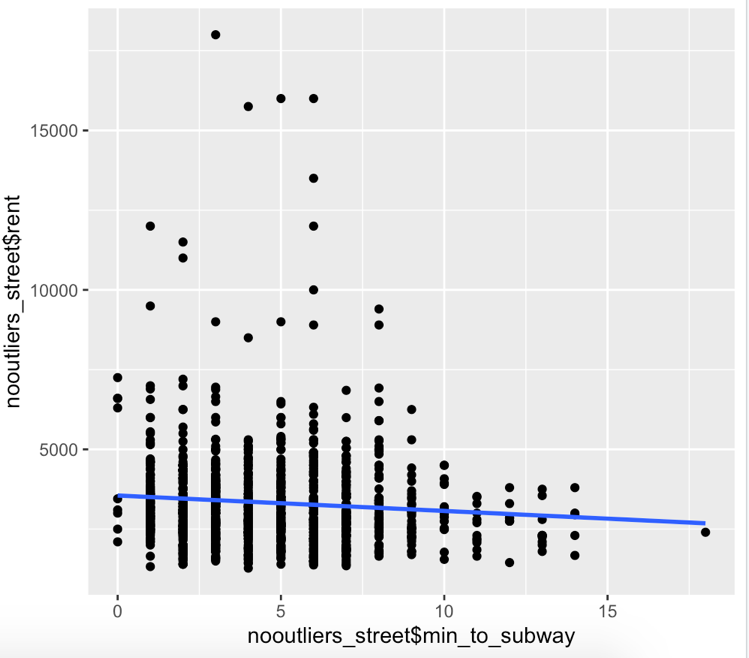
For this project though, walking distance to subway was fine. The average time for all of the boroughs was 5 minutes. And the average rent was $4,537. When specifically looking at Brooklyn, these variables changed slightly. In Brooklyn, the average walk time to the subway was 4 minutes and the longest was 43 minutes. The mean rent was $3,327 and the lowest rent was $1,275 with the highest being $3,850. From this information it is hard to see if there is a correlation between rent and distance to subway.

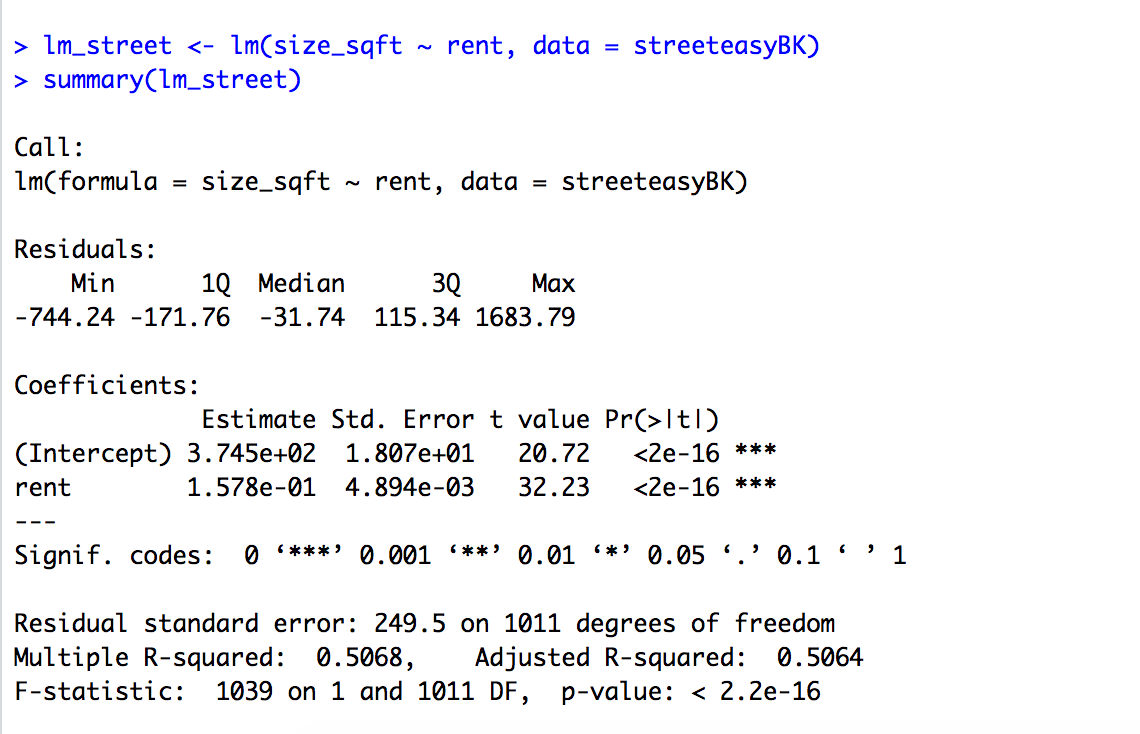
The next thing I did was run a scatter plot with a linear model regression line. There was a visible correlation shown between rent and distance, however it was not as drastic as I expected. The outliers of a 43-minute walk also appeared to skew the plot. Next I removed the outliers and ran the plot again. I then ran a summary on the linear regression line to find the residuals standard error, the R-squared and the p-value. The residual standard error was 2.803 on 1009 degrees of freedom and the R-squared was 0.007217. The p-value was 0.006879, which indicates a low relationship between variables. By looking at the coefficients, I can see that estimate is that rent does go down 1.49 units as distance increases.

Out of curiosity I ran this same process again with number of bedrooms and rent price. The correlation between these was actually higher than distance. R-Squared for this was 0.5064 and the p-value was 2.2. This indicates a much stronger relationship.

By looking at the scatter plot and examining the summary of the regression line I was able to confirm my hypothesis that rent decreased as you moved further away from the subway. However, the relationship was much lower than I expected. There seems to be other factors that more accurately relate to the price of rent.

**Images:**





**Code:**

library(lubridate)

library(dplyr)

library(tidyverse)

library(broom)

library(GGally)

install.packages("data.table")

library(data.table)

streeteasy <- read.csv ("../Desktop/Data Analysis - 640/info640-master/DataSets/streeteasy.csv")

summary(streeteasy)

glimpse (streeteasy)

head(streeteasy)

tail(streeteasy)

streeteasyBK <- streeteasy %>% filter (borough == "Brooklyn")

summary(streeteasyBK)

glimpse (streeteasyBK)

head(streeteasyBK)

tail(streeteasyBK)

ggplot(streeteasyBK, aes(x=streeteasyBK$min\_to\_subway, y=streeteasyBK$rent)) + geom\_point()+

labs(title = "Price and Distance ")

streeteasyBK %>% arrange(desc(rent))

ggplot(streeteasyBK, aes(x=streeteasyBK$min\_to\_subway, y=streeteasyBK$rent)) + geom\_point()+

stat\_smooth(method ="lm", se=FALSE)

labs(title = "Price by Distance")

lm\_street <- lm(min\_to\_subway ~ rent, data = streeteasyBK)

summary(lm\_street)

lm\_street <- lm(size\_sqft ~ min\_to\_subway, data = streeteasyBK)

summary(lm\_street)

lm\_matrix\_street <- broom::augment(lm\_street)

head(lm\_matrix\_street)

lm\_matrix\_street %>%

arrange(desc(.resid)) %>%

head()

nooutliers\_street <- streeteasyBK %>%

filter (min\_to\_subway < 43)

head(nooutliers\_street)

ggplot(nooutliers\_street, aes(x=nooutliers\_street$min\_to\_subway, y=nooutlier\_street$rent)) + geom\_point()+

labs(title = "Price and Distance ")

streeteasyBK %>% arrange(desc(rent))

ggplot(nooutliers\_street, aes(x=nooutliers\_street$min\_to\_subway, y=nooutliers\_street$rent)) + geom\_point()+

stat\_smooth(method ="lm", se=FALSE)

labs(title = "Price by Distance")

lm\_street <- lm(min\_to\_subway ~ rent, data = nooutliers\_street)

summary(lm\_street)

Resources:

<https://www.metro.us/news/local-news/new-york/nyc-rent-cheaper-near-subway-stops>

<https://www.renthop.com/studies/nyc/is-new-york-city-finally-getting-cheaper>