Assignment 2 Airbnb in New York City

Jiaqi Zhu jz2783

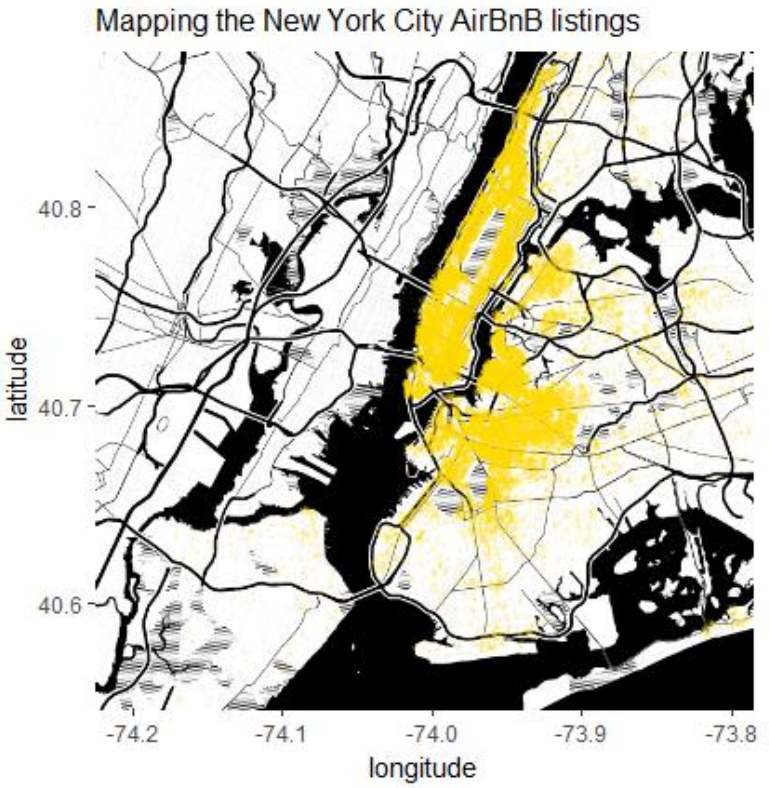
Nowadays, self-employment and rental market is booming with the momentum from the new technological platforms such as Airbnb. For now, the Airbnb is no doubt a domaining choice for travelers to seek accommodations. Many questions arose with the hope of better understanding this rule-breaker(maker) and emerging monopoly in real estate rental market. Why is there always available Airbnb rooms for the travel destination? How are the Airbnb rentals spatially distributed? Do availability days of the Airbnb rooms affect the price and income? Do the locations of Airbnb rentals correlated with public transportation access? Bearing these questions in mind, I am taking the New York City as an case to study the Airbnb in three folds: 1. Overall description of the Airbnb rental locations ; 2. Exploring the differences between sporadic rentals with (semi-)permanent rentals; 3. Zooming in Williamsburg neighborhood to investigate the subway access and other attributes of the Airbnb rentals.

Part I Overall location

1-1 Mapping the New York City Airbnb listings' location

To start with, I mapped all the Airbnb rentals in the New York City with yellow dots in Figure1.As we can see from Figure1, the Airbnb rentals are mainly distributed in Manhattan, south of Bronx, north west of Brooklyn and west part of Queens. Staten Island have Airbnb rentals but relatively sparse.

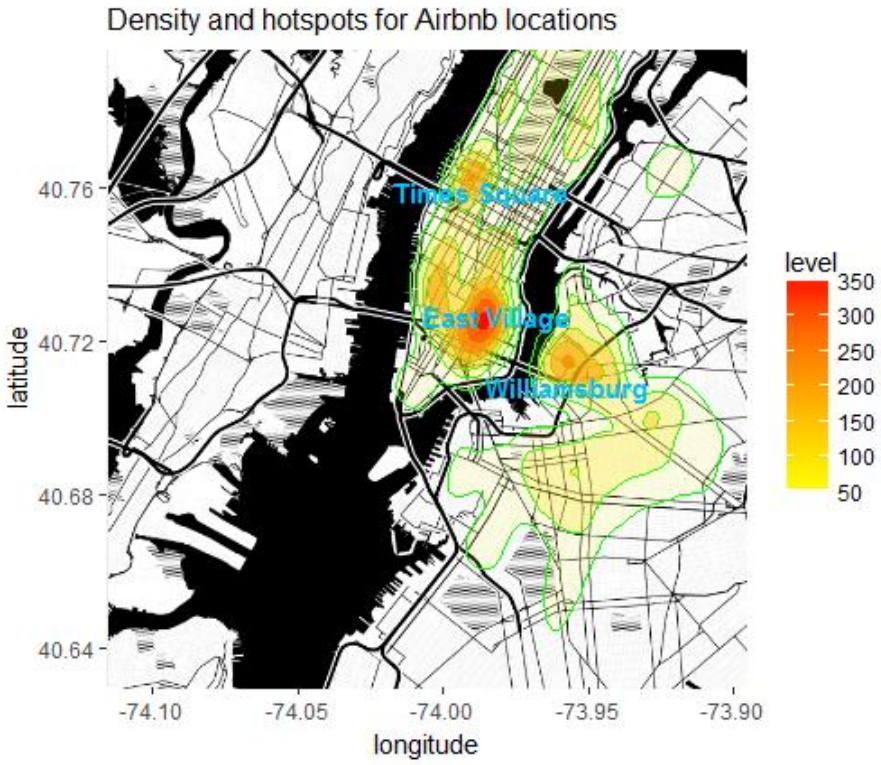
Figure 1



1-2 Mapping the density of Airbnb listings and hotspots for Airbnb locations

Location mapping can only tell us the geographical distribution of Airbnb with little information of other attributes. In Figure2, I mapped the density of Airbnb listings and highlighted three hotspots for Airbnb accommodation: Time Square, East Village and Williamsburg. Time Square and East Village are located in Manhattan while the Williamsburg locates in Brooklyn. It's no surprising that these three spots are dense with Airbnb rentals because they are all popular tourist destinations and landmarks in New York. From the density and hotspot map, we can conjecture that Airbnb targets on travelers who seeks temporary accommodations.

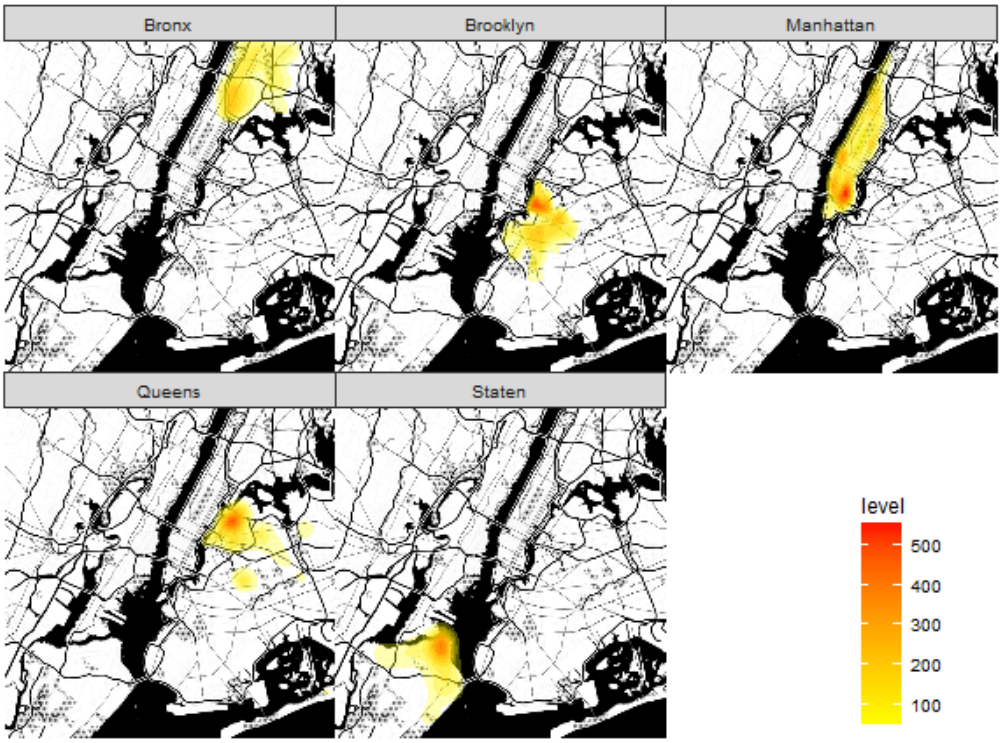
Figure 2



Not sure whether the other three boroughs have aggregation and density spots , I mapped separate density graphs for the five boroughs in New York . From Figure 3 I found that not only Manhattan and Brooklyn, the west part of Queens has a highest density spot and north part of Staten also has a dense spot. Comparing to the other four boroughs, Bronx seem to have no centralized Airbnb spot.

Figure 3

Airbnb density in each borough in New York



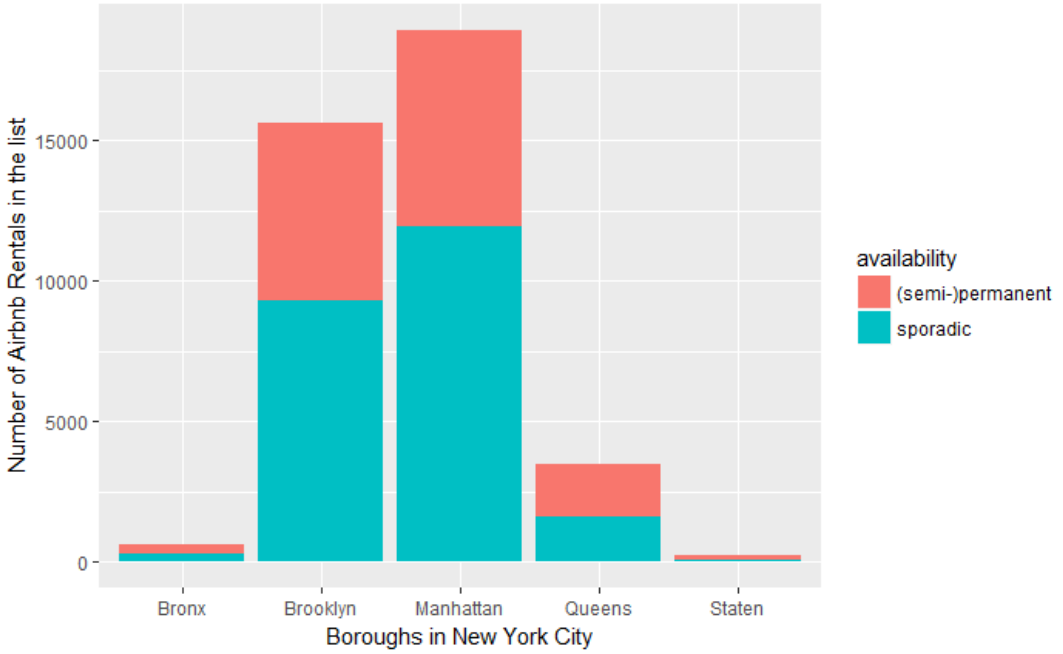
Part II Sporadic vs. (semi-)Permanent rentals

2-1 Non-map graph explore where in NYC listings are available sporadically vs. year round

By scanning through the Airbnb listing file I can not find many rentals that has 365 available days in a year ,which is also, in some case, illegal to have all-year round Airbnb rentals. After googling the cutline for sporadic and permanent, I found no standardized answer for this question. Some articles even say a room with more than 30 days available in a year should be registered and categorized into another rental system other than Airbnb. With little knowledge of the legislation issues in the rental market, I am using the median of number days in a year ---- 183---- as a cutline to separate Airbnb rentals into two categories: Sporadic rentals(available days in a year < 183) vs (semi-) Permanent rentals (available days in a year>= 183). In Figure 4, I graphed the number comparison between sporadic rentals and (semi-) permanent rentals in 5 boroughs with bars. From the Figure 4 we can not only see there are, generally, more sporadic rentals than (semi-)permanent rentals, but also tell gap of the Airbnb numbers among the five boroughs in New York which , in part, reflecting the tourist economy of each borough. Manhattan is the first and Brooklyn is the second. One thing worth mention is that, even though there are more sporadic rentals in the top two boroughs , the number of (semi-)permanent rentals equals or even outweighs the sporadic rentals in other three boroughs. This might be explained by there are more vacant housing with low usage in percentage in Bronx, Queens and Staten as their economies perform worse than the top two boroughs.

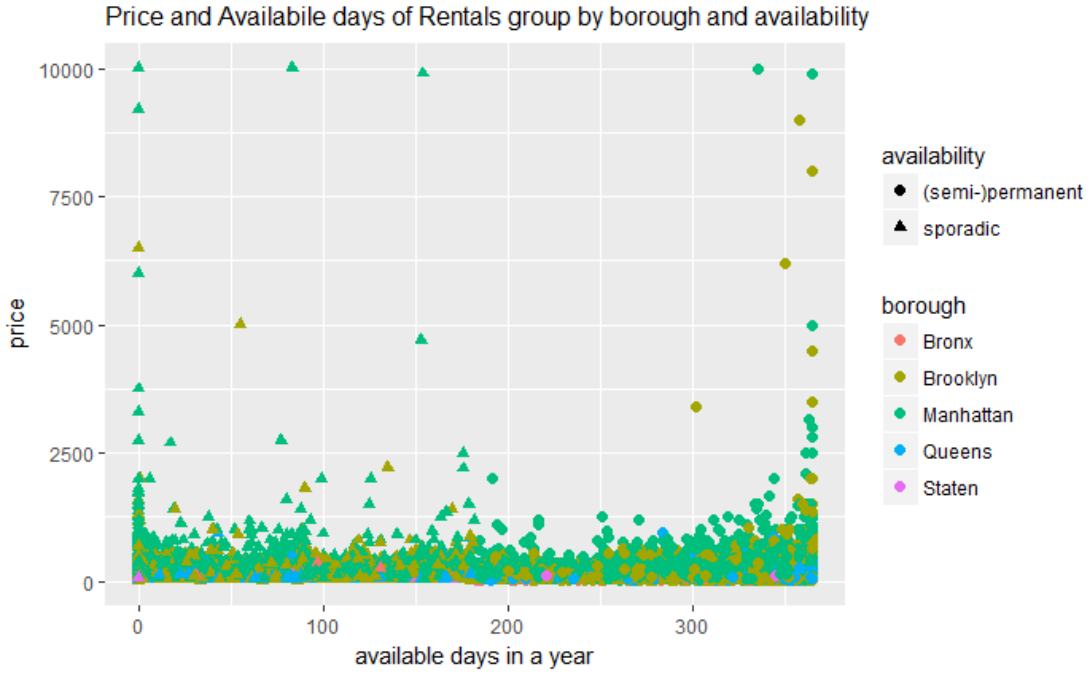
Figure 4

(semi-)Permanent vs. Sporadic Rentals in five boroughs in New York



Curious about whether the availability days in a year affects the price of the Airbnb rentals, I drew Figure 5 with scatterplots. However, from Figure 5, even though grouped the dots by boroughs, I couldn't see the difference made by availability.

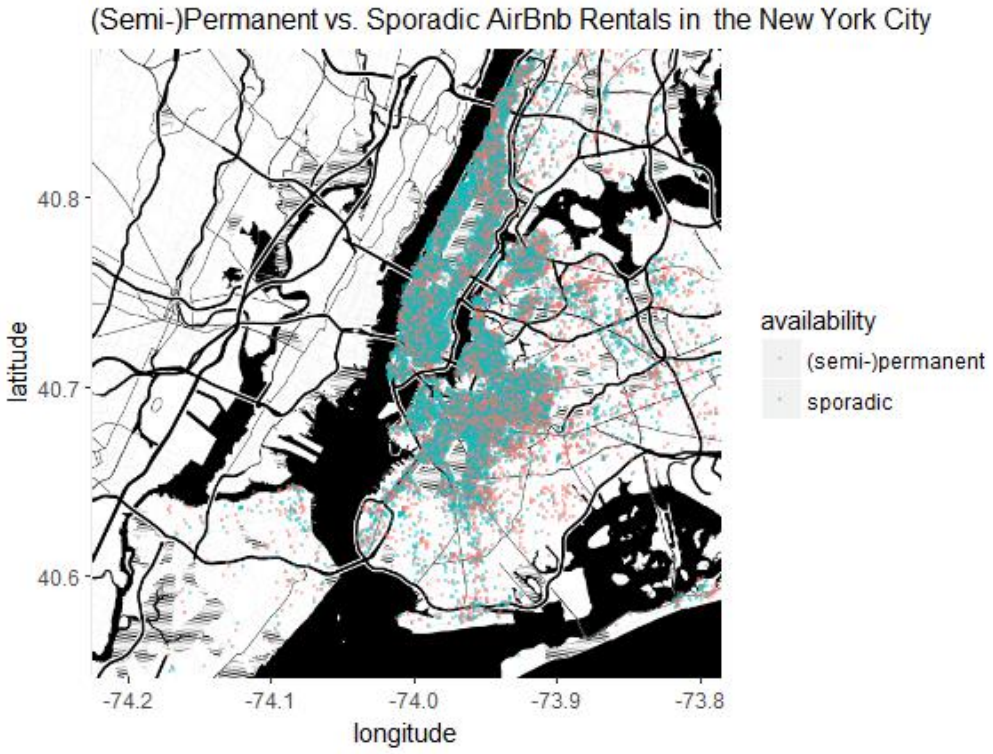
Figure 5



1-3 Map graph explore where in NYC listings are available sporadically vs year round

In Figure 6, I mapped the locations of (semi-)permanent rentals and sporadic rentals with different color. The blue ones are sporadic rentals. From the occupied area we can see that there is more blue dotted areas that indicates more sporadic rentals.

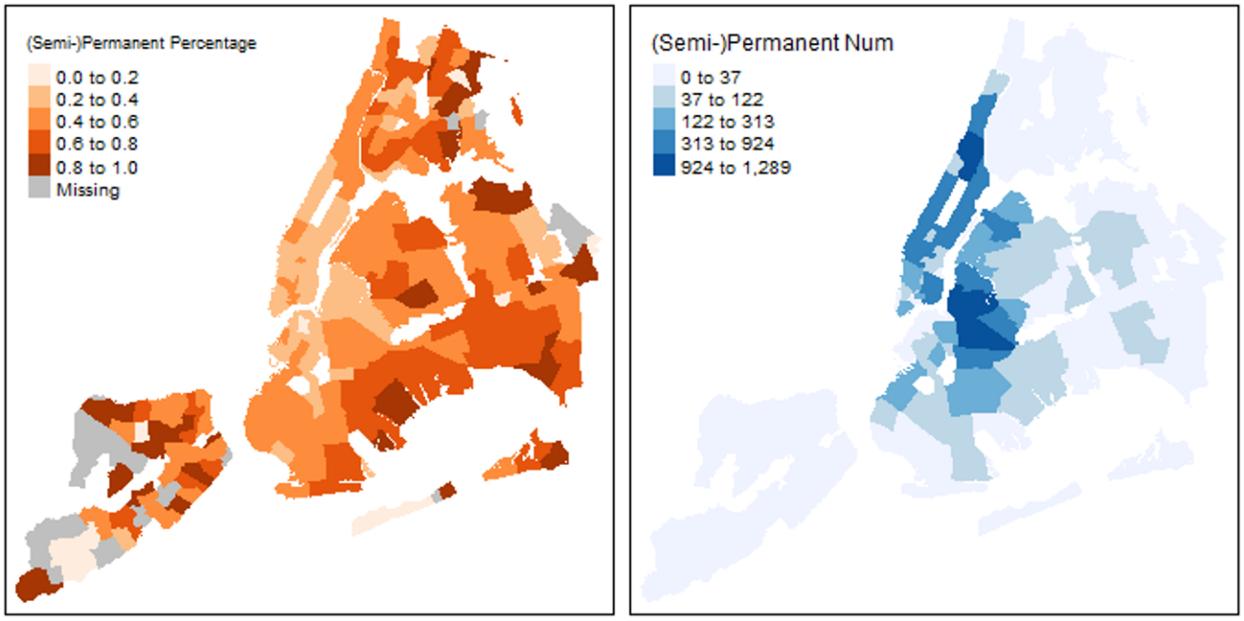
Figure 6



More interested in (semi-) permanent rentals, I tried to calculated the percentage and sum of numbers of (semi-)permanent rentals for each neighborhood in New York, and drew the statistics in Figure 7. I was uncertain about which quantity of the (semi-)permanent rentals to use. After making the Figure 7, I found that, even though the percentage of (semi-)permanent rentals among all the rentals in neighborhood sounds reasonable for analysis, it is not comparable between neighborhoods which may be misleading. The left map of Figure 7 is the percentage of (semi-)permanent rentals in each neighborhoods in New York. This orange map is consistent with the bar chart Figure 4's analysis results. There are more (semi-)permanent rentals in less developed boroughs in New York in percentage. However, the total number of Airbnb rentals between neighborhoods varies from borough to borough. Therefore, we should be more favor of the right hand side map of Figure 7 which is somewhat consistent with the density map of New York city as a whole.

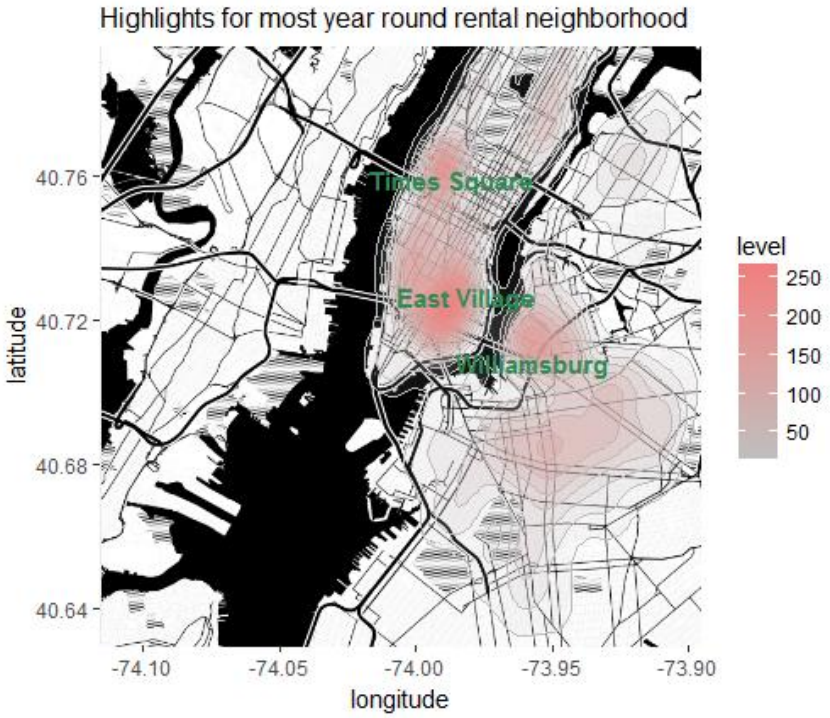
Figure 7

Percentage vs. Number of (semi-)permanent rentals in neighborhood



Therefore , I drew a density map for (semi-)permanent rentals in Figure 8 highlighted the same three hotspots for the (semi-)permanent rentals as the total density map.

Figure 8



2-3 How (semi-)permanent rentals differ from sporadic rentals

Before talking about the regression models and graphs, I would like to introduce a new variable created for the project analysis--- Average Monthly Income. The formula for calculating this variable is as followed:

Average monthly income = availability in 30 days * price * reviews_per_month / 30

This quantity would measure the average income in a monthly basis of each Airbnb host. Price is a quantity for the housing value however, the real income has the real impact on hosts and the economy. Therefore, I would like to explore the relationship between availability and average monthly incomes as well as other possible variables.

Here are my hypotheses :

My first hypothesis is that, the availability of rentals has correlation with average monthly income.

My second hypothesis is that, the more sporadic rentals are, the less average monthly income would be.

My third hypothesis is that, the higher review score is the higher average monthly income would be.

I am also including the boroughs, property_types, bathroom per person(calculated by the number of bathrooms divided by accommodates), beds per person(calculated by the number of beds divided by accommodates). These are seemed as control variables which would also be tested to see if there is statistical significance between the dependent variable.

I build the model 1 for hypothesis testing. The dependent variable for this regression model is the average monthly income, the independent variables are the variables discussed above. The statistical results are in Table 1.

From Table 1, we can see that my hypothesis 1 is examined to be true that the availability of rentals does correlate with the average monthly income with statistical significance. Holding other variables fixed, the sporadic rentals has on average 45 dollars less in average monthly income than (semi-)permanent rentals with statistical significance, which means my hypothesis 2 is also accepted. Net of other variables, one point increase in review score rating is statistically correlated with 0.23 dollar increase in average monthly income. My third hypothesis is also tested and accepted. The regression model results tells us that, to increase the income of Airbnb hosts, one should work on improving the review score rating as well as increase the days of the available rentals.

Table 1 . Model 1 for Average Monthly Income

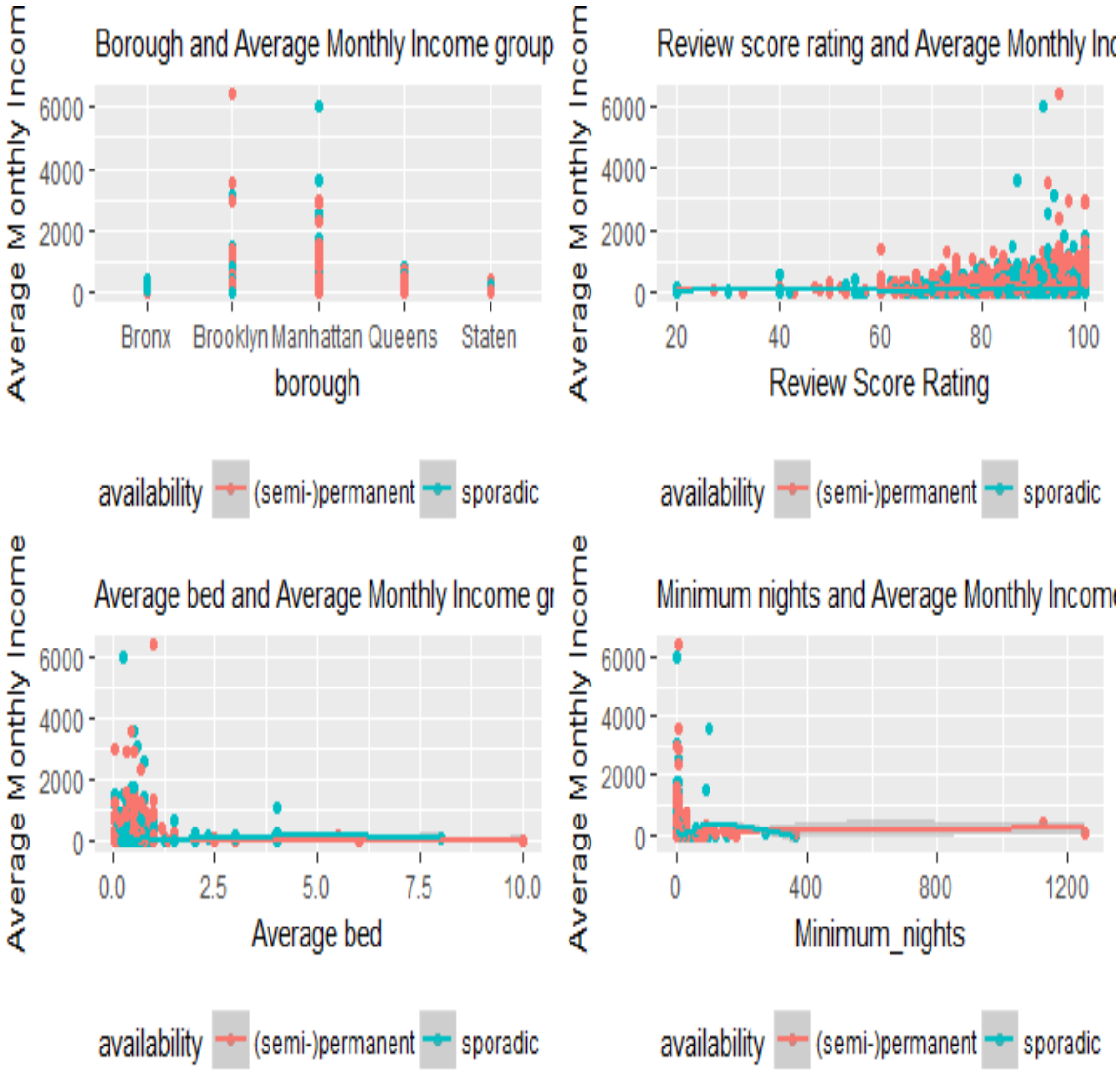
model1 <- lm(mon_income ~ review_scores_rating + availability + borough + as.factor(property_type) + ave_bath + ave_bed, data = airbnb)

```
Residuals:
   Min     1Q Median     3Q    Max
-190.1  -55.5  -30.0   15.7 6300.6

Coefficients:
                                      Estimate Std. Error t value Pr(>|t|)
(Intercept)                           71.74804   10.13407   7.080 1.47e-12 ***
review_scores_rating                   0.23219    0.08574   2.708 0.006773 **
availabilitysporadic                 -44.96494    1.50752 -29.827  < 2e-16 ***
boroughBrooklyn                       12.35456    6.18481   1.998 0.045773 *
boroughManhattan                      36.53685    6.20428   5.889 3.93e-09 ***
boroughQueens                         12.05024    6.55773   1.838 0.066137 .
boroughStaten                         -0.63616   11.97134  -0.053 0.957620
as.factor(property_type)Bed & Breakfast  19.62760   12.07442   1.626 0.104055
as.factor(property_type)Boat          97.22327   64.23084   1.514 0.130124
as.factor(property_type)Boutique hotel  -50.85130   90.79155  -0.560 0.575423
as.factor(property_type)Bungalow      57.50495   48.58886   1.184 0.236620
as.factor(property_type)Cabin        -21.25001   90.79247  -0.234 0.814947
as.factor(property_type)Camper/RV    -75.11269  128.42203  -0.585 0.558626
as.factor(property_type)Castle       145.68098   74.20748   1.963 0.049637 *
as.factor(property_type)Cave         -40.62206  128.38827  -0.316 0.751701
as.factor(property_type)Chalet        24.86771  128.38928   0.194 0.846420
as.factor(property_type)Condominium   55.99773    8.16237   6.860 7.00e-12 ***
as.factor(property_type)Dorm         -36.41935   25.72469  -1.416 0.156863
as.factor(property_type)Guesthouse    58.40316   24.82201   2.353 0.018635 *
as.factor(property_type)Hostel       -24.95143  128.39399  -0.194 0.845915
as.factor(property_type)House         18.86350    2.85210   6.614 3.81e-11 ***
as.factor(property_type)Hut           38.49461  128.39708   0.300 0.764325
as.factor(property_type)Island       -24.18149  128.38675  -0.188 0.850604
as.factor(property_type)Loft          49.26532    4.81362  10.235  < 2e-16 ***
as.factor(property_type)Other         25.93305    9.39748   2.760 0.005791 **
as.factor(property_type)Timeshare    644.59594  128.38627   5.021 5.18e-07 ***
as.factor(property_type)Townhouse     38.94145    6.40703   6.078 1.23e-09 ***
as.factor(property_type)Villa        -17.66693   45.42200  -0.389 0.697315
ave_bath                             -73.70840    2.99428 -24.616  < 2e-16 ***
ave_bed                               12.11495    3.00979   4.025 5.71e-05 ***
minimum_nights                        -0.21426    0.06178  -3.468 0.000525 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

To help better understand the regression results, I plotted several graphs in Figure 9. In Figure 9, we can see the Brooklyn and Manhattan have the higher average monthly income. The regression line of sporadic rentals and (semi-)permanent rentals under review score rating has little difference. Holding the average bed constant, the average monthly income of sporadic rental seem a little bit higher than (semi-)permanent rentals. Holding the minimum nights fixed, the difference between sporadic rentals and (semi-)permanent rentals is hard to decide. Therefore, to tailor out the true effect and differences, simple plots with two or three variables are not powerful than multivariable regression models which is not easy to visualized in the way as simple graphs.

Figure 9

Four graphs for Average Monthly Income

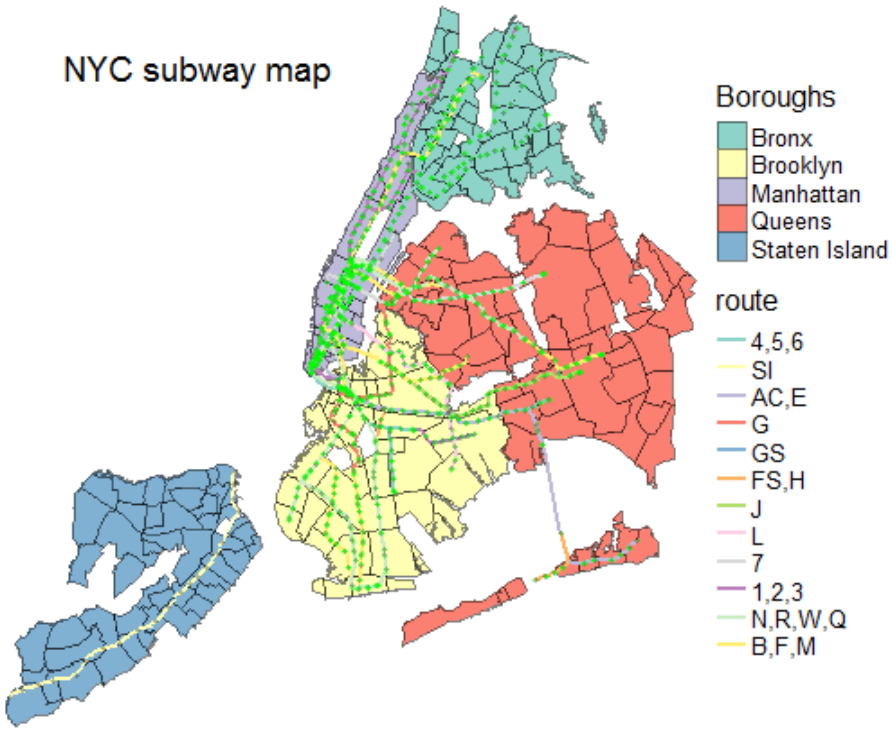


Part III Airbnb and Subway Access (zooming in Williamsburg neighborhood)

3-1 New York City Subway Overview (stops, entrances and routes mapping)

Before exploring and zooming in details, I would like to give an overview of the New York City's Subway system. In Figure 10, I plotted all the subway routes, stops and entrances. Stops are in black dots and entrances are in green dots.

Figure 10

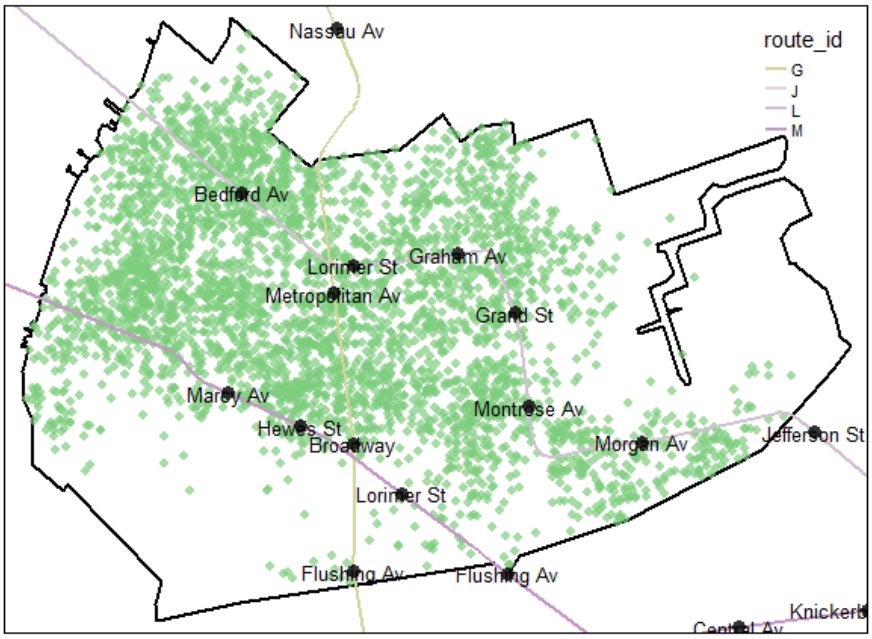


3-2 Zoom in Williamsburg neighborhood and see the locations of Airbnb rentals and subway stops

In Figure 11, I zoomed in the New York City Map into Williamsburg neighborhood in Brooklyn. The green dots are the Airbnb rental's location. The Black dots are the subway stops with names on them. There are mainly three lines of subway across the Wiiliamsburg neighborhood: G,L,M.

Figure 11

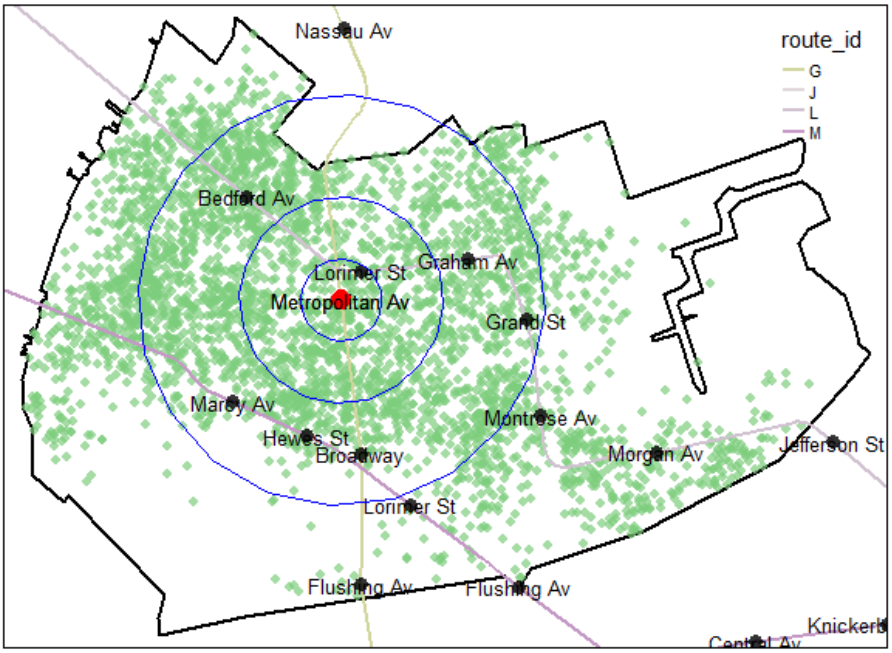
Zoom in Williamsburg neighborhood with Airbnb, subway stops and routes



3-3 Calculate and display how many listings are in different perimeter around Metropolitan Av station

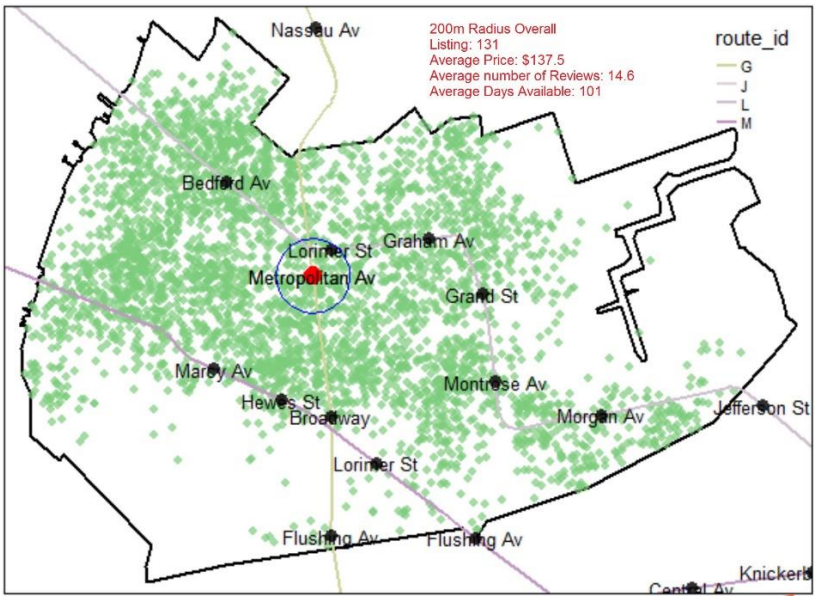
In the following graphs , I would us 200meter, 500meter and 1000meter radius with the central to be Metropolitan Av subway station to calculate the parameters of the Airbnb listings within the buffer. Figure 12 shows the size differences between different buffers.

Figure 12. Three buffers with central of Metropolitan Av stop



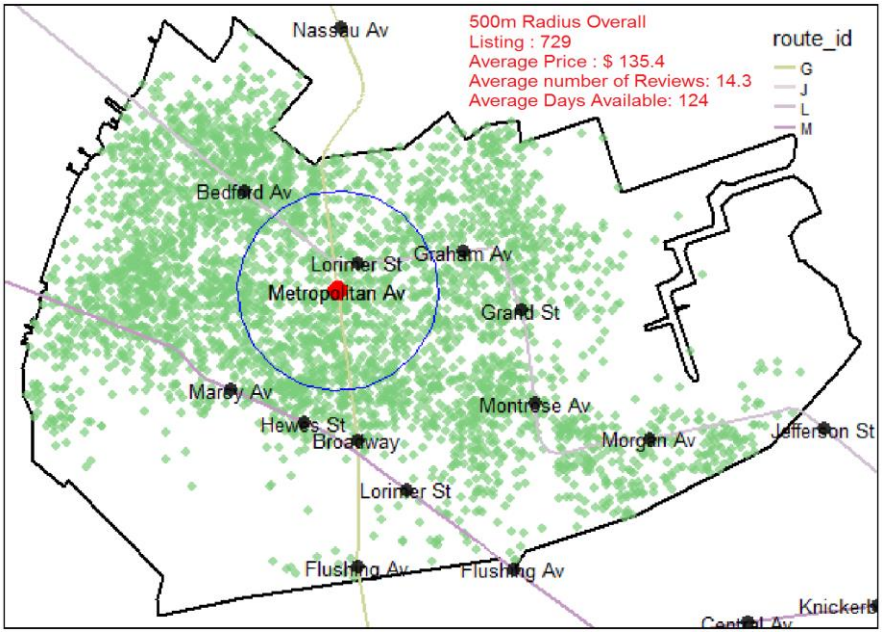
In Figure 13, we can see that there are , actually two subway stations within the 200m buffer. There are 131 listed Airbnb rentals within this buffer. The average price of them are 137.5 dollars per day.The number of their average reviews is 14.6. The average days available among these 131 listings is 101 which is within the range of sporadic rentals.

Figure 13. 200m Radius Buffer



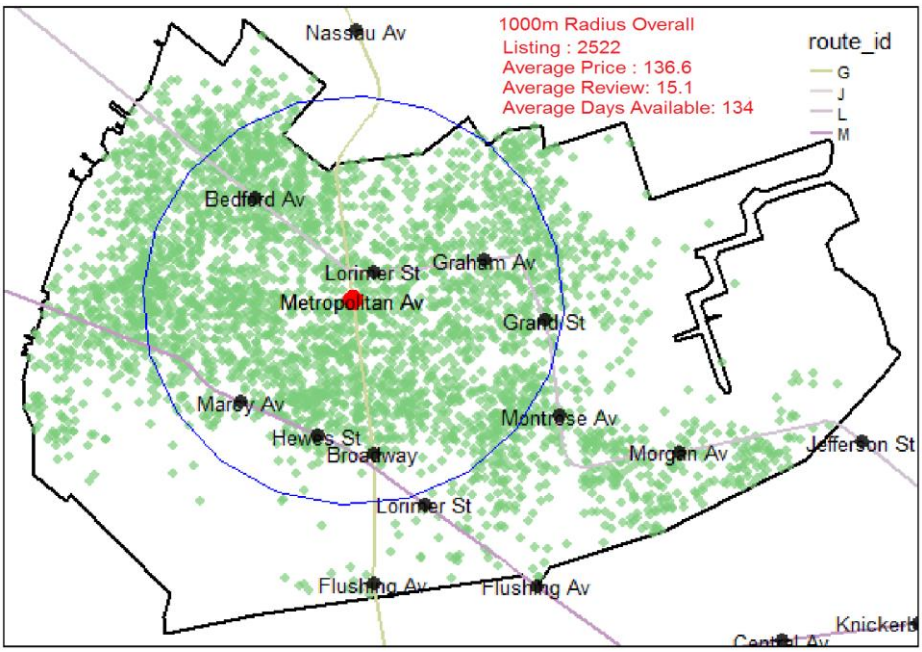
In Figure 14, we can see that there are , still only two subway stations within the 500m buffer. There are 729 listed Airbnb rentals within this buffer. The average price of them are 135.4 dollars per day. The number of their average reviews is 14.3. The average days available among these729 listings is 124 which is within the range of sporadic rentals.

Figure 14. 500m Radius Buffer



In Figure 15, we can see that there are , there are 8 subway stations within the 1000m buffer. There are 2522 listed Airbnb rentals within this buffer. The average price of them are 136.6 dollars per day. The number of their average reviews is 15.1. The average days available among these 2522 listings is 134 which is within the range of sporadic rentals.

Figure 15. 1000m Radius Buffer



Comparing the average price of the listings within different perimeters of the buffers, I found little variance. The 200m buffer has the highest average price. However as the perimeter going up, there are more subway stations involved which make it harder to tell the difference of the accessibility of the listings in the buffer. That's why, the 1000m buffer's average price is higher than 500m buffer. Generally speaking, as the subway stations in Williamsburg is relatively randomly distributed that make smaller variance of nearest subway station distance among listing Airbnb rentals, the average price of them has relatively smaller difference due to the similar accessibility of subway stations.

3-4 Exploring whether the price of listings is related to having access to the nearby subway

Zooming in and mapping the dots can not render precise statistical evaluation of the relationships between variables. Therefore I drew Figure 16 with the aim of figuring out the relationship between price of the rental and the proximity of the nearest subway station. From Figure 16, we can see that, the regression lines of the (semi-) permanent rentals and sporadic rentals are almost overlapping showing little difference between the availability which is consistent with the previous findings. However, the slope of lines are little negative which indicates that the farther the locations of the rental from the nearest subway station, the lower the price. This totally make sense because, the nearer subway stations would endow location advantage for public transportation convenience that would add favorability to the rentals. People are lazy and tired of commuting.Where ever is more convenient, is a better choice, especially in metropolis. Therefore, the better accessibility to subway stations the higher value of the rentals, in return , the higher price it can claim.

Figure 16

However, there are many other variables that would also affect the price such as property_types, room_type, availability and number of reviews. Therefore I built the second regression model with price as the dependent variable, distance to the nearest subway station(meter) as the independent variable, and the property_types, room_type, availability and number of reviews as the control variables. In Table2, the results showed that the farther the distance the lower the price , however , without statistical significance. This may because of the data is just the subset of the Williamsburg neighborhood whose subway system is relatively even distributed.

Table 2. Model 2 for price and distance to nearest subway stop

```
lm(formula = price ~ distance + property_type + room_type + availability +
    number_of_reviews, data = wl)

Residuals:
   Min     1Q Median     3Q    Max
-316.4  -41.9  -13.8   13.3 8852.5

Coefficients:
                              Estimate Std. Error t value Pr(>|t|)
(Intercept)                  2.217e+02  9.516e+00  23.299  < 2e-16 ***
distance                    -6.659e-03  9.338e-03  -0.713  0.47578
property_typeBed & Breakfast -9.968e-01  8.733e+01  -0.011  0.99089
property_typeCastle          -6.423e+01  2.131e+02  -0.301  0.76311
property_typeChalet           6.325e+01  2.144e+02   0.295  0.76798
property_typeCondominium      1.875e+01  3.847e+01   0.487  0.62596
property_typeDorm             4.709e+01  2.147e+02   0.219  0.82639
property_typeGuesthouse      -1.151e+02  2.132e+02  -0.540  0.58918
property_typeHouse            6.065e+01  2.262e+01   2.682  0.00736 **
property_typeLoft             5.363e+01  1.319e+01   4.066 4.88e-05 ***
property_typeOther            2.375e+02  4.398e+01   5.400 7.09e-08 ***
property_typeTownhouse        1.331e+02  5.923e+01   2.247  0.02471 *
room_typePrivate room        -1.190e+02  7.106e+00 -16.743  < 2e-16 ***
room_typeShared room         -1.445e+02  2.656e+01  -5.440 5.68e-08 ***
availabilitysporadic         -1.737e+01  7.571e+00  -2.294  0.02186 *
number_of_reviews            -3.130e-01  1.253e-01  -2.499  0.01251 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 213 on 3708 degrees of freedom
Multiple R-squared:  0.08706,   Adjusted R-squared:  0.08337
F-statistic: 23.57 on 15 and 3708 DF,  p-value: < 2.2e-16
```

Conclusion

From the mapping, graphs, regression model results above, we found that, the Airbnb rentals are densely distributed in the popular touristy areas in New York; there are more sporadic rentals than (semi-) permanent rentals as a whole; higher review score rating is correlated with higher average monthly income; (semi-)permanent rentals on average have higher average monthly income but not higher price; subway accessibility is a factor that may influence the price of the Airbnb listing, the closer to subway station the higher the price.

However, more work should be done to better understand the relationship between subway density and price of Airbnb such as generalize the analysis into other neighborhoods in New York.

Project Book

To start out this project I firstly outlined the framework for it:

1. Overall Location

1-1 mapping the New York City AirBnB listings' location

1-2 mapping the density of these listings and hotspots for AirBnb locations(annoate a few hotspots on the map)

2. Sporadic rentals vs. (semi-)permanent rentals

2-1 non-map graph explore where in NYC listings are available sporadically vs. year round

2-2 map explore where in NYC listings are available sporadically vs. year round

2-3 summary statistics / map/ non-map graph . how such permanent rentals differ from sporadic rentals in a consise format

3. AirBnB and Subway Access

3-1 Explore how the location, type and features of AirBnB listings are related to subway access . select a single neighborhood that lends itself to such an analysis(i.e. has multiple subway stations, different types of AirBnB listings etc.) you can complement with an analysis of the entire city or an entire borough

3-2 use the information about the location of subway stations and AirBnB listings to calculate distances from each listing to the next(nearest) subway.

3-3 calculate (and display) how many listings are in different perimenters around a subway station. make sure to map the subway stations(and lines) to give the reader an idea of what you are doing

3-4 explore whether the price of listings is related to having access to the subway nearby. try to control for some other obvious determinants ofr price: how many people the space sleeps whether it's an entire propoerty or a private room, the type of property(apartment boat hous loft) and the number of reviews display and describe your findings

After outlining, I spent a lot of time on the data recoding and merging.

After cleaning and reshaping the data, I began trying out the maps, graphs and models.

The graphs, maps and statistical results included in the project are selected decided after trying out many experiments. For Figure 13,14,15, I used Paint to add the information : listing, average price, average review, average days available as red text in the blank area of the map for I didn’t know how to add random text in tmap.

Codes for the graphs, maps and regression models

Figure 1

```
## 1-1 Mapping the New York City AirBnB listings' location
map_NYC <- get_map("New York City", zoom = 11, source = "stamen", maptype = "toner-background")
g1 <- ggmap(map_NYC)
g1 <- g1 +
      geom_point(aes(x = longitude, y = latitude), data = airbnb, size = 0.1, alpha = 0.1, color = "gold") +
      labs(x = "longitude",
           y = "latitude",
           title = "Mapping the New York City AirBnB listings" ) +
      theme(plot.title = element_text(size = 12))

g1
```

Figure 2

```
## 1-2 Mapping the density of these listings and hotspots for AirBnb locations(annoate a few hotspots on the map3
map_NYC <- get_map("New York City", zoom = 12, source = "stamen", maptype = "toner-background")
g2 <- ggmap(map_NYC)
g2 <- g2 +
      geom_density_2d(aes(x=longitude,y=latitude), color = "Green", data= airbnb, size=0.5) +
      stat_density2d(aes(x=longitude,y=latitude, fill=..level.., alpha = ..level..), data = airbnb ,geom="polygon") +
      scale_alpha(range = c(0.1, 0.9), guide = "none") +
      theme(legend.position = "right") +
      labs(x = "longitude",
           y = "latitude",
           title = "Density and hotspots for Airbnb locations" ) +
      theme(plot.title = element_text(size = 12)) +
      scale_fill_gradient(low = "yellow", high = "red")

(g2 <- g2 + annotate("text",x=-73.9571, y=40.7081, label="Williamsburg", color="deepskyblue",fontface=2, size=4) +
       annotate("text",x=-73.987325, y=40.758899, label="Times Square", color="deepskyblue",fontface=2, size=4) +
       annotate("text",x=-73.9815, y=40.7265, label="East Village",color="deepskyblue",fontface=2, size=4))
```

Figure 3

```
## Graph3 mapping the density in each borough
map_NYC <- get_map("New York City", zoom = 11, source = "stamen", maptype = "toner-background")
theme_map <- function(base_size=9, base_family="") {
    require(grid)
    theme_bw(base_size=base_size, base_family=base_family) %+replace%
    theme(axis.line=element_blank(),
          axis.text=element_blank(),
          axis.ticks=element_blank(),
          axis.title=element_blank(),
          panel.background=element_blank(),
          panel.border=element_blank(),
          panel.grid=element_blank(),
          panel.margin=unit(0, "lines"),
          plot.background=element_blank(),
          legend.justification = c(0,0),
          legend.position = c(0.85,0)
          )
}
ggmap(map_NYC) +
  stat_density2d(aes(x=longitude,y=latitude, fill=..level..,alpha = ..level.., group= borough),
                 data= airbnb,geom="polygon") +
  scale_alpha(range = c(0.3, 0.5), guide=FALSE) +
  scale_fill_gradient(low = "yellow", high = "red") +
  theme(legend.position = "right") +
  theme(strip.text.x = element_text(size = 10, face=2)) +
  facet_wrap(~ borough, ncol = 3) +
  theme_map()
```

Figure 4

```
## Graph4 bar chart for availability in each borough
airbnb$num <- NA
airbnb$num[airbnb$borough == "Bronx"] <- table(airbnb$borough)[1]
airbnb$num[airbnb$borough == "Brooklyn"] <- table(airbnb$borough)[2]
airbnb$num[airbnb$borough == "Manhattan"] <- table(airbnb$borough)[3]
airbnb$num[airbnb$borough == "Queens"] <- table(airbnb$borough)[4]
airbnb$num[airbnb$borough == "Staten"] <- table(airbnb$borough)[5]

sub3 <- with(airbnb, data.frame(borough = borough,
                                availability = availability,
                                property_type = property_type))
sub3 <- na.omit(sub3)
str(sub3)
g4 <- ggplot(sub3, aes(x = borough, fill = availability)) +
      xlab("Boroughs in New York City") +
      ylab("Number of Airbnb Rentals in the list") +
      theme(legend.position = "right")
g4 + geom_bar()
```

Figure 5

```
## Graph 5 Price and Availability
g5 <- ggplot(airbnb, aes(x = availability_365, y = price, shape = availability , color = borough)) +
      geom_point(size =2, alpha = 1) +
      labs(x = "available days in a year",
           y = "price",
           title = "Price and Availabile days of Rentals group by borough and availability" ) +
      theme(plot.title = element_text(size = 12))
g5
```

Figure 6

```
## Graph 6 Mapping the Availability
map_NYC <- get_map("New York City", zoom = 11, source = "stamen", maptype = "toner-background")
g6 <- ggmap(map_NYC)
g6 <- g6 +
      geom_point(data = airbnb,aes(x = longitude, y = latitude,color = availability),  size = 0.3, alpha = 0.3) +
      labs(x = "longitude",
           y = "latitude",
           title = "(Semi-)Permanent vs. Sporadic AirBnb Rentals in  the New York City" ) +
      theme(plot.title = element_text(size = 12))

g6
```

Figure 7

```
ny_nb@data$permanent_pct <- 0
ny_nb@data$permanent_num <- 0
for (i in 1:length(ny_nb@data$neighbourhood)) {
  a <- subset(airbnb, airbnb$neighbourhood_cleansed == ny_nb@data$neighbourhood[i])
  ny_nb@data$permanent_pct[i] <- mean(a$pp)
  ny_nb@data$permanent_num[i] <- sum(a$pp)
}

tm_shape(ny_nb) + layout +
    tm_fill(c("permanent_pct", "permanent_num"),
        style=c("pretty", "kmeans"),
        palette=list("Oranges", "Blues"),
        auto.palette.mapping=FALSE,
        title=c("(Semi-)Permanent Percentage", "(Semi-)Permanent Num")) +
tm_style_white()
```

Figure 8

```
## Graph 8 for (semi-)permanent rentals density
sub1 <- filter(airbnb, airbnb$availability == "(semi-)permanent")
map_NYC <- get_map("New York City", zoom = 12, source = "stamen", maptype = "toner-background")
g8 <- ggmap(map_NYC)
g8 <- g8 +
      geom_density_2d(aes(x=longitude,y=latitude), data= sub1, color = "grey") +
      stat_density2d(aes(x=longitude,y=latitude, fill=..level.., alpha = ..level..), data = sub1 ,geom="polygon") +
      scale_alpha(range = c(0.1, 0.5), guide = "none") +
      theme(legend.position = "right") +
      labs(x = "longitude",
           y = "latitude",
           title = "Highlights for most year round rental neighborhood" ) +
      theme(plot.title = element_text(size = 12)) +
      scale_fill_gradient(low = "grey", high = "lightcoral")
g8

(g8 <- g8 + annotate("text",x=-73.9571, y=40.7081, label="Williamsburg", color="seagreen4",fontface=2, size=4) +
       annotate("text",x=-73.987325, y=40.758899, label="Times Square", color="seagreen4",fontface=2, size=4) +
       annotate("text",x=-73.9815, y=40.7265, label="East Village",color="seagreen4",fontface=2, size=4))
```

Figure 9

```
g7 <- ggplot(subset(airbnb, airbnb$mon_income != 0), aes(x = borough, y = mon_income,color = availability)) +
  geom_point() +
  theme(legend.position = "bottom") +
  labs(x = "borough", y = "Average Monthly Income", title = "Average Monthly Income and borough group by availability") +
    theme(plot.title = element_text(size = 11)) +
    geom_smooth(aes(color = availability))

g7

g8 <- ggplot(airbnb, aes(x = ave_bed, y = mon_income, color = availability)) +
  geom_point() +
  theme(legend.position = "bottom") +
  labs(x = "Average bed", y = "Average Monthly Income", title = "Average bed and Average Monthly Income group by availability") +
    theme(plot.title = element_text(size = 11)) +
    geom_smooth(aes(color = availability))


g9 <- ggplot(airbnb, aes(x = review_scores_rating, y = mon_income, color = availability)) +
  geom_point() +
   theme(legend.position = "bottom") +
  labs(x = "Review Score Rating", y = "Average Monthly Income", title = "Review score rating and Average Monthly Income group by availability") +
    theme(plot.title = element_text(size = 11)) +
    geom_smooth(aes(color = availability)) +
    theme(axis.text.x = element_blank())


g10 <- ggplot(airbnb, aes(x = minimum_nights, y = mon_income, color = availability)) +
  geom_point() +
  theme(legend.position = "bottom") +
 # facet_grid(~ borough) +
  labs(x = "Minimum_nights", y = "Average Monthly Income", title = "Minimum nights and Average Monthly Income group by availability ") +
    theme(plot.title = element_text(size = 11)) +
    geom_smooth(aes(color = availability)) +
    theme(axis.text.x = element_blank())
 g10
multiplot(g7,g8,g9,g10, cols = 2)
```

Figure 10

```
## Graph 10
l <- levels(sb_routes@data$color)
levels(sb_routes@data$color) <- c(
  "4,5,6", "SI", "AC,E", "G", "GS", "FS,H", "J",
  "L", "7", "1,2,3", "N,R,W,Q", "B,F,M")
sb_routes@data$route <- sb_routes@data$color
g10 <- tm_shape(ny_nb) +
        tm_style_white() +
        tm_fill("neighbourhood_group",
                title = "Boroughs",
                 labels = c("Bronx","Brooklyn","Manhattan","Queens","Staten Island")) +
        tm_borders(col = "black", alpha = 0.5) +
        tm_shape(sb_routes) +
        tm_lines(col = "route", scale=2, legend.lwd.show = FALSE) +
        tm_shape(sb_stops) +
        tm_dots(col = "black", size = 0.05,alpha = 0.3) +
        tm_shape(sb_entrances) +
        tm_dots(col = "green", size = 0.05, alpha = 0.3) +
        tm_layout(title = "NYC subway map",
                  title.size = 1.2,
                  title.position = c(0.1, 0.9),
                  legend.outside = TRUE,
                  legend.title.size = 1.2,
                  legend.text.size = 0.8,
                  legend.position = c(0,0.1),
                  frame=FALSE,
                  bg.color = "white")
g10
```

Figure 11

```
g11<- tm_shape(neigh) +
       tm_borders(col = "black", lwd = 2) +
       tm_shape(wl) +
       tm_dots(size = 0.2, col = "palegreen3", alpha = 0.7) +
       tm_shape(sb_routes) +
       tm_lines(col = "route_id", scale=2, legend.lwd.show = FALSE) +
       tm_shape(wl_stops) +
       tm_dots(size = 0.4, col = "black", alpha = 0.8) +
       tm_text("stop_name", size = 0.8)
g11
```

Figure 12 - 15

```
cent_met <- g11 +  tm_shape(met_buffer) + tm_dots(col="red", size=0.8) +   tm_text("name", size=0.8)
cent_met
g12 <- cent_met + tm_shape(buffer_200) + tm_borders(col="blue") +
       tm_shape(buffer_500) + tm_borders(col="blue") +
       tm_shape(buffer_1000) + tm_borders(col="blue")

g12

g13 <- cent_met + tm_shape(buffer_200) + tm_borders(col="blue")
g13

g14 <- cent_met + tm_shape(buffer_500) + tm_borders(col="blue")
g14

g15 <- cent_met + tm_shape(buffer_1000) + tm_borders(col="blue")
g15
```

Figure 16

```
## Graph 16
ggplot(data = subset(wl,wl$price <=2500), aes(x = as.numeric(distance), y = as.numeric(price) ,color = availability)) +
  geom_point() +
  geom_smooth(method = lm, se = FALSE) +
  labs(x = "Distance to the nearest subway stop(meter)",
       y = "Price of the Airbnb listing",
       title = "Distance to the nearest subway stop and Price group by availability in Williamsburg Neighborhood") +
  theme(plot.title = element_text(size = 12))
```

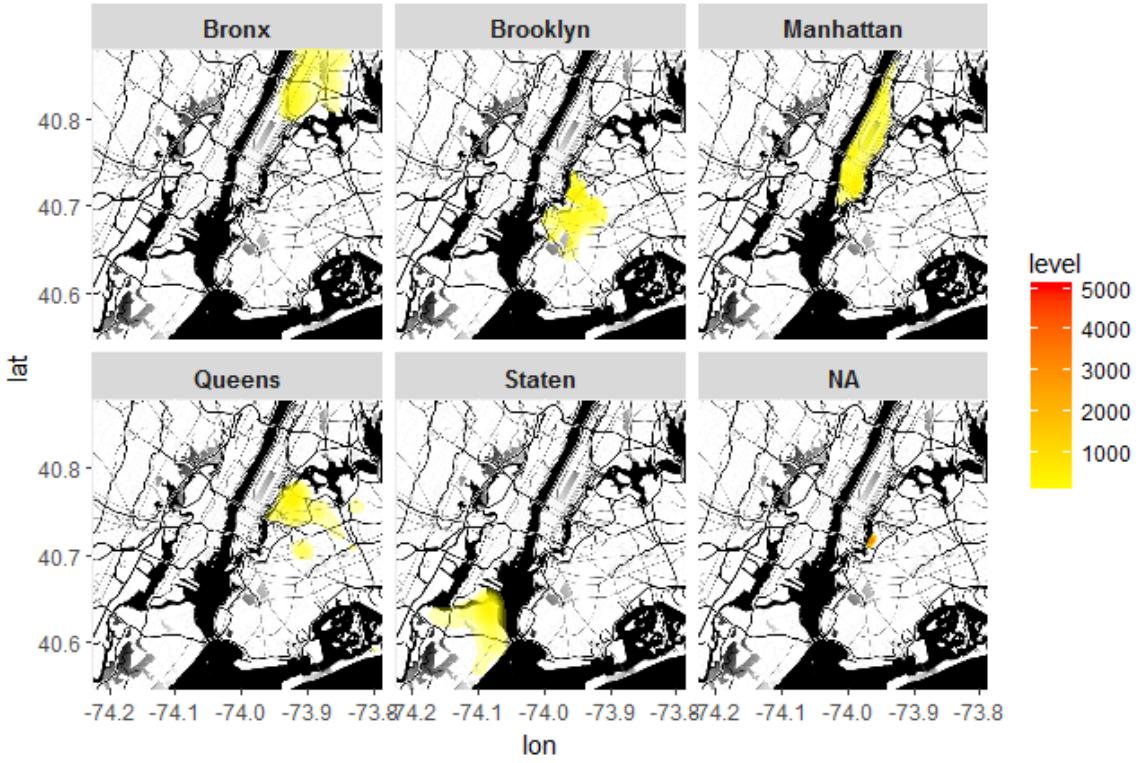
Table 1

```
model1 <- lm(mon_income ~ review_scores_rating + availability + borough + as.factor(property_type) + ave_bath + ave_bed + minimum_nights, data = airbnb )
summary(model1)
```

Table 2

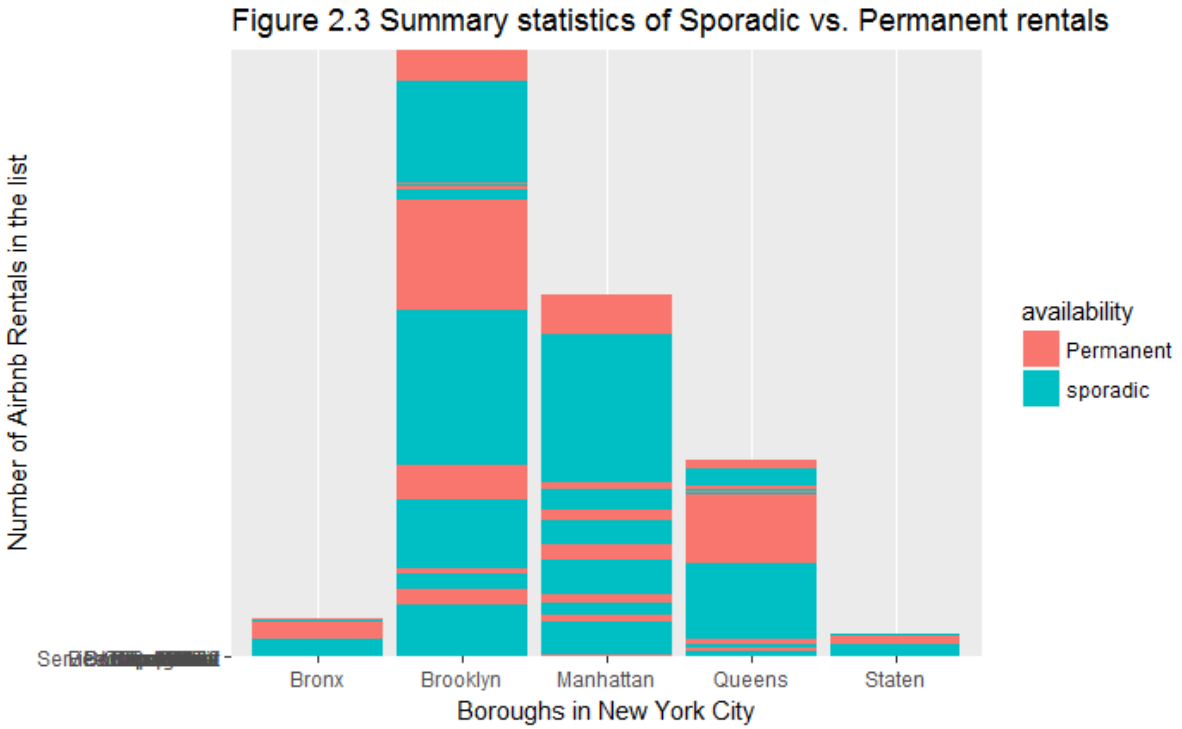
```
model2 <- lm(price ~ distance + property_type + room_type + availability + number_of_reviews, data = wl)
summary(model2)
```

Exploratory Graphs



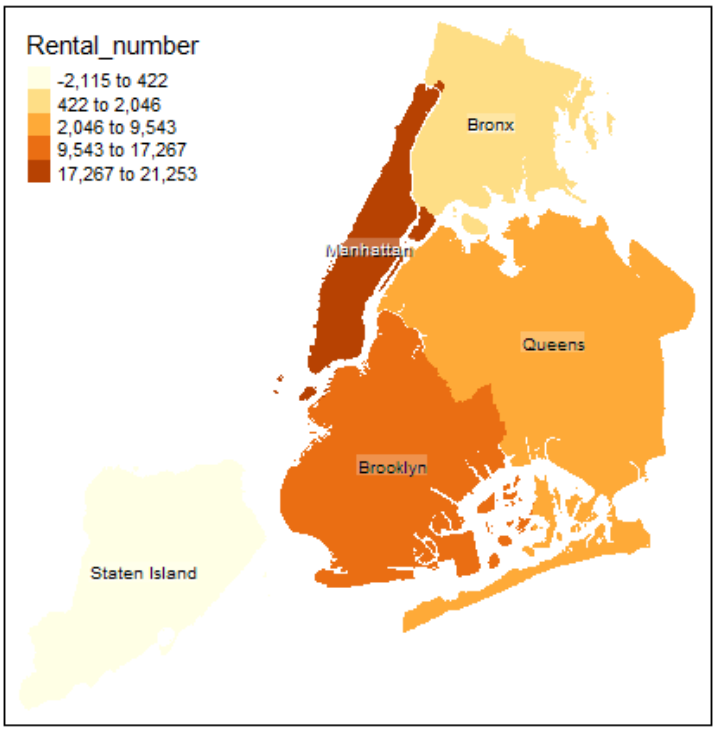

```
map_NYC <- get_map("New York City", zoom = 11, source = "stamen", maptype = "toner-background")
ggmap(map_NYC) +
  stat_density2d(aes(x=longitude,y=latitude, fill=..level..,alpha = ..level.., group= borough),
                 data= airbnb,geom="polygon") +
  scale_alpha(range = c(0.3, 0.5), guide=FALSE) +
  scale_fill_gradient(low = "yellow", high = "red") +
  theme(legend.position = "right") +
  theme(strip.text.x = element_text(size = 10, face=2)) +
  facet_wrap(~ borough, ncol = 3)
```

This exploratory graph helped me to identify the locations of NA rentals and help me with the missing data correction.



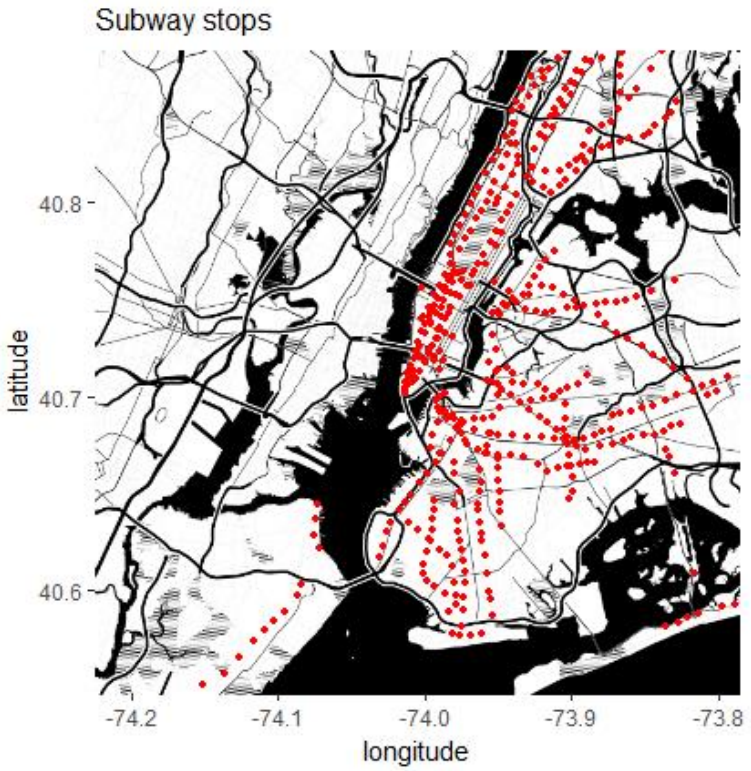

```
g5 <- ggplot(sub3, aes(x = borough, y = property_type, fill = availability)) +
      xlab("Boroughs in New York City") +
      ylab("Number of Airbnb Rentals in the list") +
      ggtitle("Figure 2.3 Summary statistics of Sporadic vs. Permanent rentals") +
      theme(legend.position = "right")

g5 + geom_bar(stat = "identity")
```

This is an ugly try but at least it tells me not to group them by property_type and make it fill in the bar.



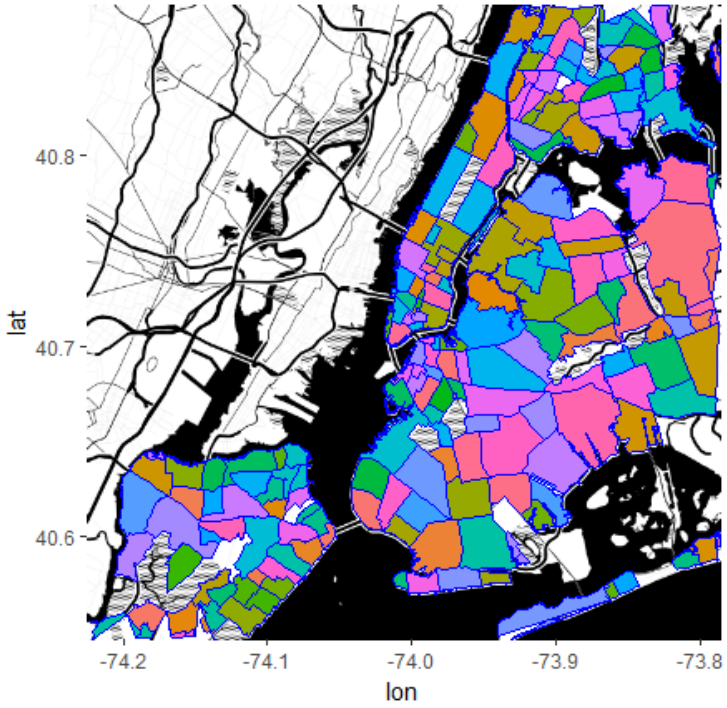

```
ny_b <- readOGR("C://Users/Jiaqi Zhu/Google Drive/QMSS/2017spring/Data Visualization/HW/Assignment2",layer =
"nybb")
a <- table(airbnb$borough)
b <- as.data.frame(a)
b$bn <- as.character(b$Var1)
b$bn[b$bn == "Staten"] <- "Staten Island"
b$Var <- as.factor(b$bn)
ny_b@data <- left_join(ny_b@data, b, by =c("BoroName" = "Var"))
ny_b@data$Rental_number <- ny_b@data$Freq
tm_shape(ny_b) + tm_borders()
tm_shape(ny_b) + tm_fill("Rental_number") + tm_text("BoroName", size=.6, shadow=TRUE,
  bg.color="white", bg.alpha=.25,
  remove.overlap=TRUE)
```

This is the graph I tried to make the borough shapefile useful. This is showing that the most dense area is Manhattan but that is also shown in the density map so it is somewhat redundant. What's more,this graph didn't really show the density of rentals in each borough and they are not in range but a point number. So I didn't choose this in the project.



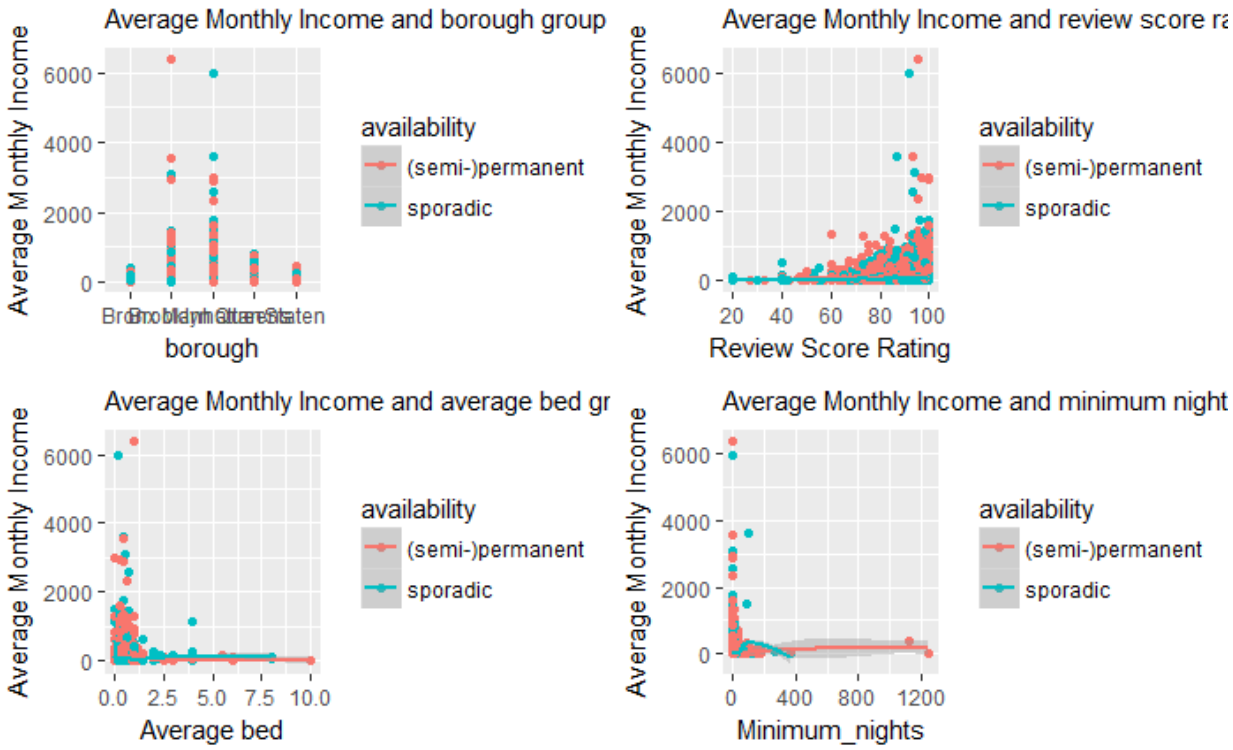

```
g7 <- ggmap(map_NYC) +
      geom_point(aes(x = stop_lon, y = stop_lat), data = as.data.frame(sb_stops), size = 1, alpha = 1, color =
"red") +
      labs(x = "longitude",
           y = "latitude",
           title = "Subway stops" ) +
      theme(plot.title = element_text(size = 12))

g7
```

I was trying to map the subway stops but it turned out to be ugly and part of New York is not included. Therefore I turned to the neighborhood geojson file.

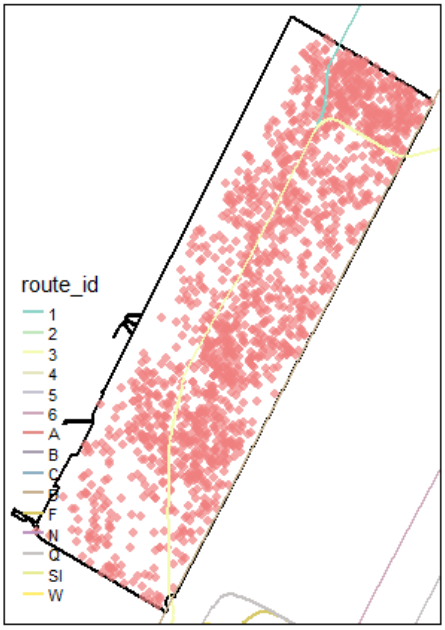



```
ny_nb2 <- spTransform(ny_nb, CRS("+proj=longlat +datum=WGS84"))
ny_nb2 <- fortify(ny_nb2)
map_poly <- ggmap(map_NYC) +    # our raster map from before
  geom_polygon(aes(x=long, y=lat, group=group, fill=id),
  size=0.5, color='blue', data=ny_nb2) + theme(legend.position = "none")
map_poly
```

This polygon is not illustrative for the content I want to emphasize.

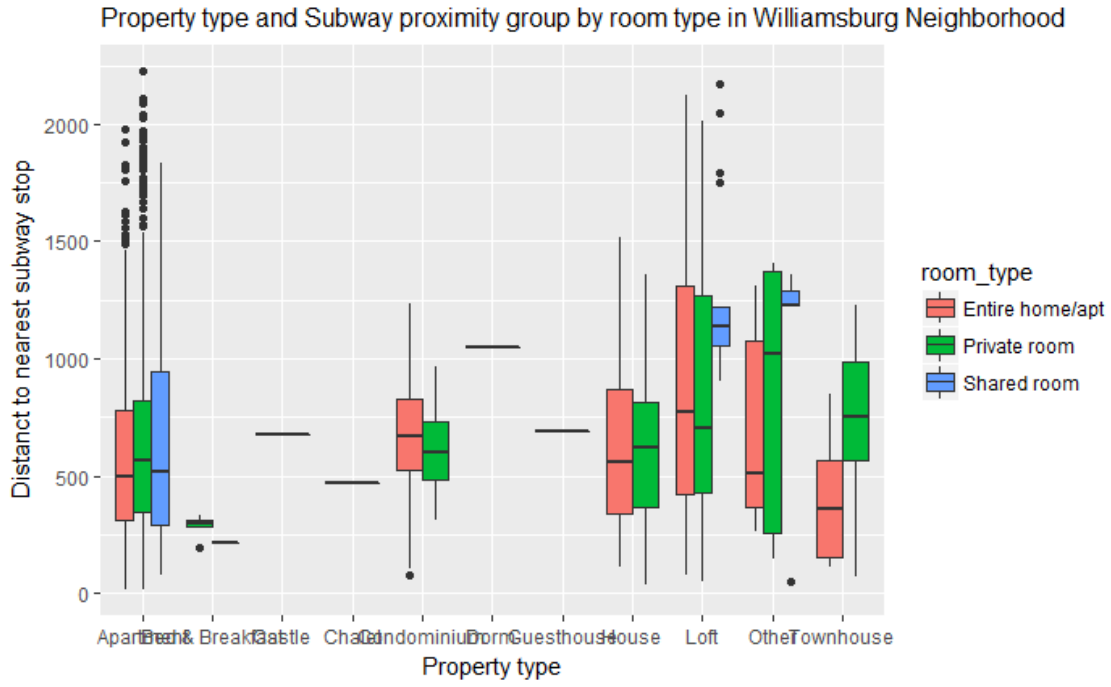


This one the legends took too much space therefore I posited them “bottom” instead in the project.



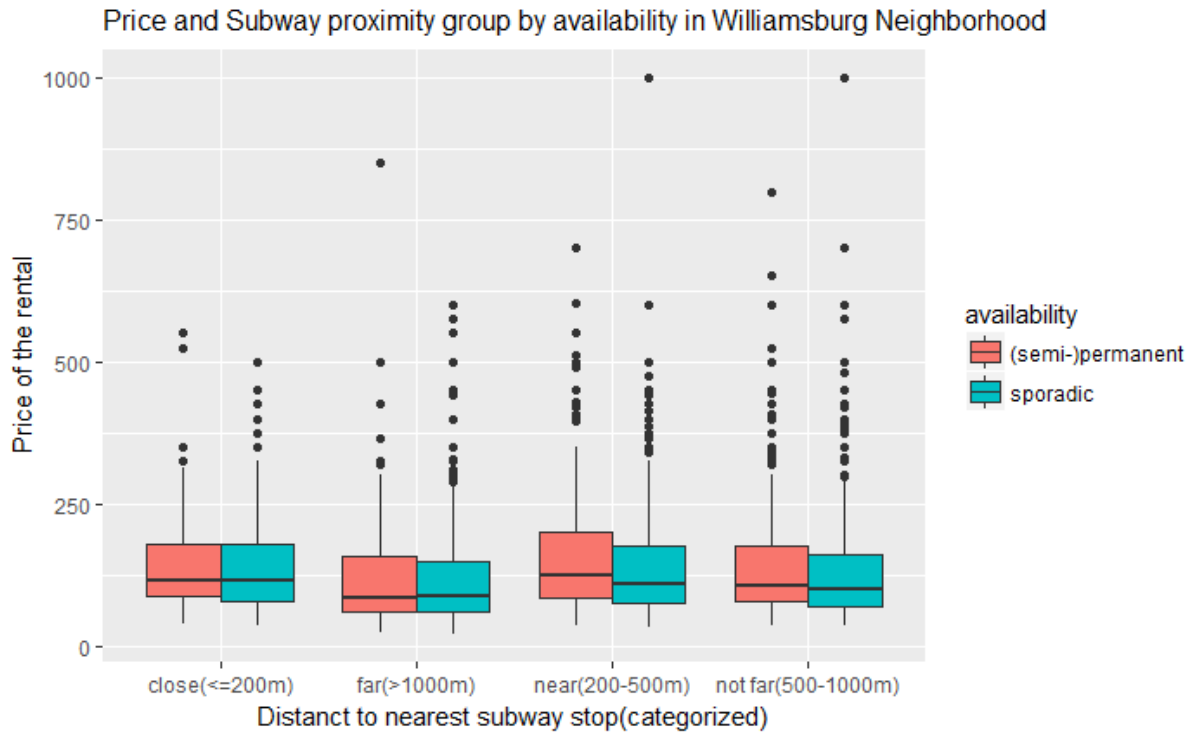
```
upw <- subset(airbnb, neighbourhood_cleansed == "Upper West Side")
coordinates(upw) <- ~ longitude + latitude
proj4string(upw) <- proj4string(ny_nb)

upw_stops <- sb_stops@data[, c(4, 3, 2)]
names(upw_stops) <- c("Longitude", "Latitude", "stop_names")
coordinates(upw_stops) <- ~ Longitude + Latitude
proj4string(upw_stops) <- proj4string(ny_nb)
n_stop <- over(upw_stops, ny_nb)
upw_stops <- sb_stops[c(462, 465, 466), ]

sel <- ny_nb$neighbourhood == "Upper West Side"
neigh <- ny_nb[sel, ]
proj4string(neigh) <- proj4string(ny_nb)
g3_2 <- tm_shape(neigh) +
       tm_borders(col = "black", lwd = 2) +
       tm_shape(upw) +
       tm_dots(size = 0.2, col = "lightcoral", alpha = 0.7) +
       tm_shape(sb_routes) +
       tm_lines(col = "route_id", scale=2, legend.lwd.show = FALSE) +
       tm_shape(upw_stops) +
       tm_dots(size = 0.4, col = "black", alpha = 0.8) +
       tm_text("stop_name", size = 0.8)
g3_2
```

I firstly tried Upper West Side neighborhood, however , it turned out to be , somewhat awkward and skinny. So I changed to Williamsburg.




```
g3_44 <- ggplot(data = subset(wl,wl$price <= 1000), aes(x = property_type, y = distance)) +
  geom_boxplot(aes(fill= room_type)) +
  geom_smooth(method = lm, se = FALSE) +
  labs(x = "Property type",
       y = "Distanct to nearest subway stop",
       title = "Property type and Subway proximity group by room type in Williamsburg Neighborhood") +
  theme(plot.title = element_text(size = 12))
g3_44
```

This graph has some missing data but I couldn't figure out how to filter them so I discarded.




```
g3_44 <- ggplot(data = subset(wl,wl$price <= 1000), aes(x = dist_ca, y = price)) +
  geom_boxplot(aes(fill= availability)) +
  geom_smooth(method = lm, se = FALSE) +
  labs(x = "Distanct to nearest subway stop(categorized)",
       y = "Price of the rental",
       title = "Price and Subway proximity group by availability in Williamsburg Neighborhood") +
  theme(plot.title = element_text(size = 12))
g3_44
```

This is a good comparison between sporadic and permanent rentals but the order of the distance is a mess and I don't know how to deal with it so I discarded it.