NORTHEASTERN VS HARVARD UNIVERSITY HOUSING DATA STATISTICS

Megha Rao and Hiu Ching(Tracy) Law

Group Name: LawRao

Course Number: DS4100 Spring 2019

DESCRIPTION:

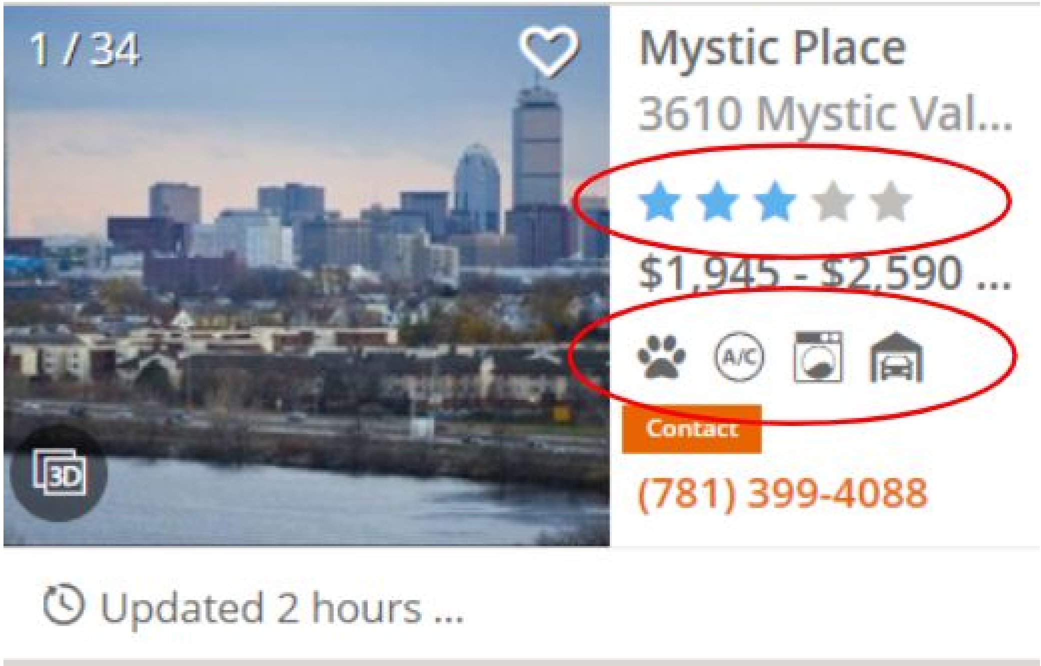
As Northeastern students who struggled with finding decently priced housing close enough to campus, we were curious to see what apartments were available in the area. The Northeastern community lacked a good database with convenient housing options. Through conversation with our peers, we realized this was a high demand prospect. As a result, we got together to compile a list of this database and analyzed it. We wanted to visualize where the most apartments were near campus, how highly priced they were and whether or not they were worth the ratings they were given.

Since we wanted to be able to compare the Northeastern data to another set, we chose to compare it to Harvard University data, since it was far away enough but still in the Higher Boston area. After analysis and visualization, we hope to see where the better housing options are: Northeastern or Harvard?

COLLECTION OF DATA:

Our first step included collecting the data. This was probably where we struggled the most. Originally, our idea was to compare Northeastern housing data vs. Boston housing statistics. We were to collect the data using surveys of our peers who lived off campus and see how they compared to the general Boston population. However, collecting data through surveys seemed like a difficult task. We were not sure if we could get enough valuable data by the time we expected to start the project. As a result, we had to quickly change our project idea. This was when we realized we would rather webscrape the data off a reputable source. As a result, we changed our idea to NU vs Harvard housing data statistics.

We collected our data off a website known as Apartment Finder[1]. This allowed us to look for housing options near both schools we were targeting. Our next struggle came with the actual webscraping. This step probably took us the longest since there were some unanticipated hurdles. Although the actual scraping using the CSS selector tool of text on the website went smooth, we struggled because all their ratings and amenities were in a picture format on the screen.



Mystic Place

3610 Mystic Val...

★★★☆☆

$1,945 - $2,590 ...

(781) 399-4088

Updated 2 hours ...

[1] https://www.apartmentfinder.com/Off-Campus-Housing/Massachusetts/Apartments-Near-Harvard-University-l146f7d

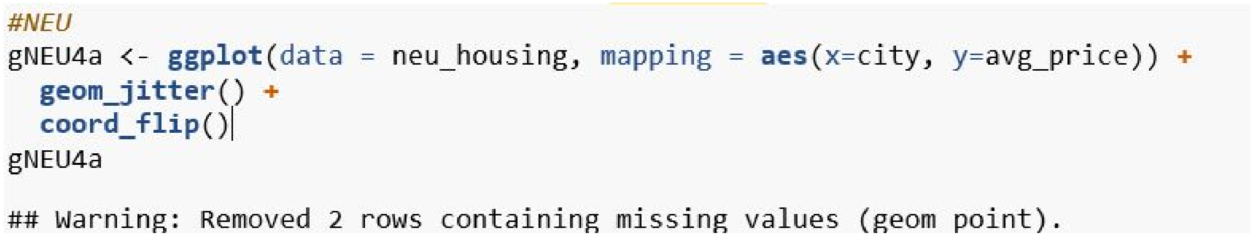
After experimenting with the Inspect HTML code built in on Google chrome, we figured out how we could extract the rating data. However, it was not easy for us to extract the amenities data since it was also in a picture format. Given the time constraint we had, we did not have enough time to figure out how to resolve this issue by the deadline. Given more time, we would definitely look into this, because this could have been valuable information that helped us analyze the data better and with more accuracy.

After webscraping all the required data, we made sure to convert any <fctr> data to <chr> and converting some of the <chr> to numeric, namely minimum price, maximum price (, and hence the average price). This extra step made it easier for enabling the process for us to make the data visual for analysis and deeper understanding.

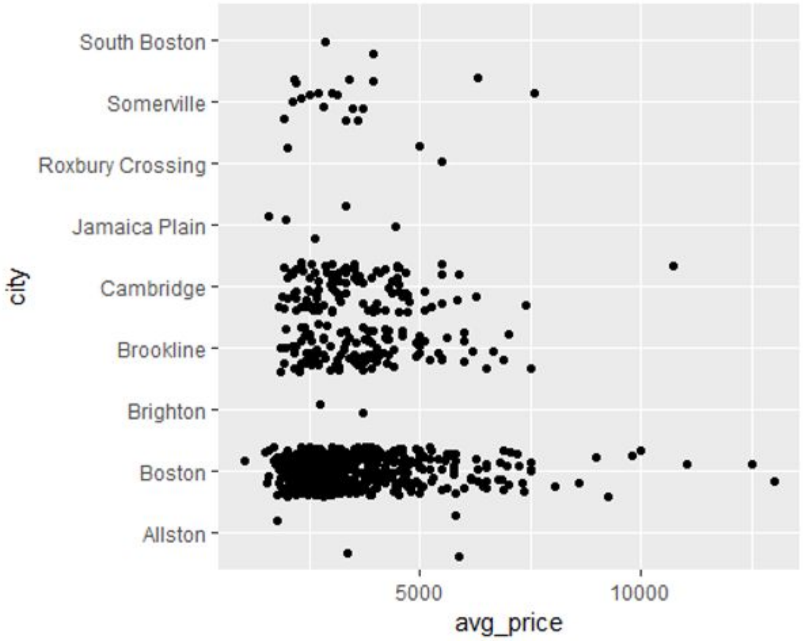
DATA VISUALIZATION AND ANALYSIS:

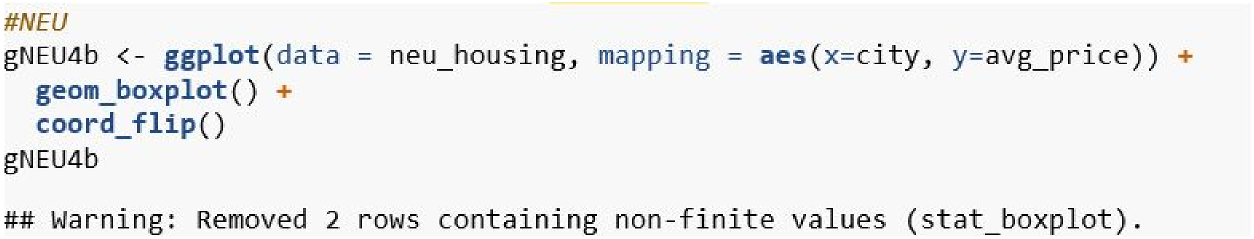
Now that we had the data, we wanted to visualize it in a meaningful way. We proceeded to answer these questions by using dplyr and ggplot2 to extract the information we wanted and then plotting graphs in R. We chose to ask the following question:

1) Is there a relationship between the average price and the city?

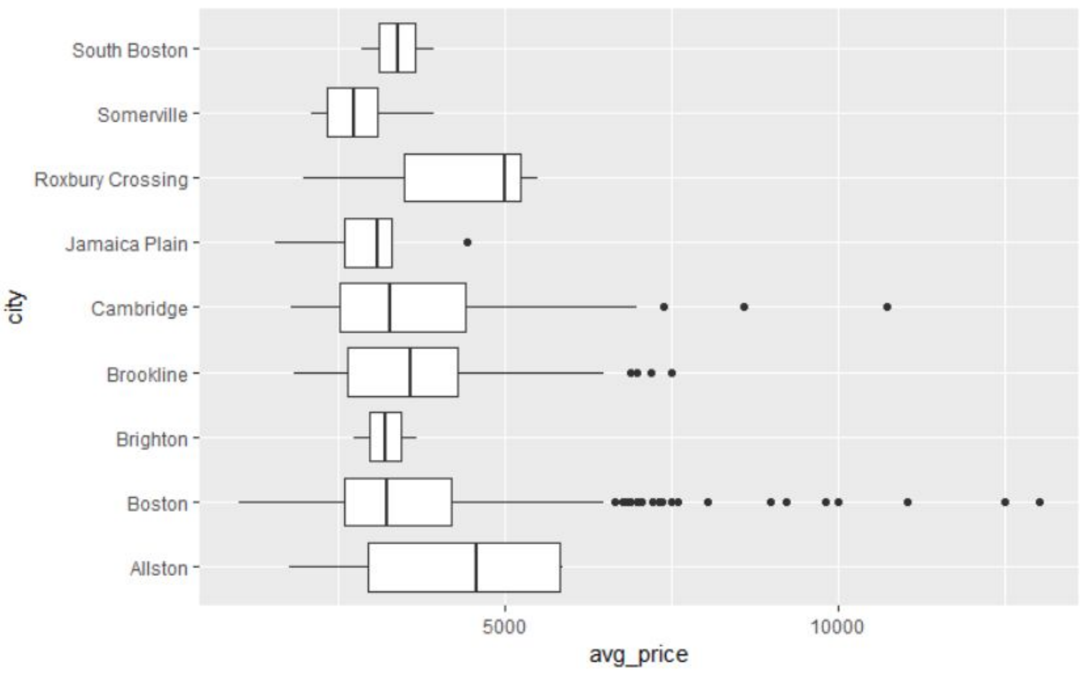


```
#NEU
gNEU4a <- ggplot(data = neu_housing, mapping = aes(x=city, y=avg_price)) +
  geom_jitter() +
  coord_flip()
gNEU4a

## Warning: Removed 2 rows containing missing values (geom point).
```

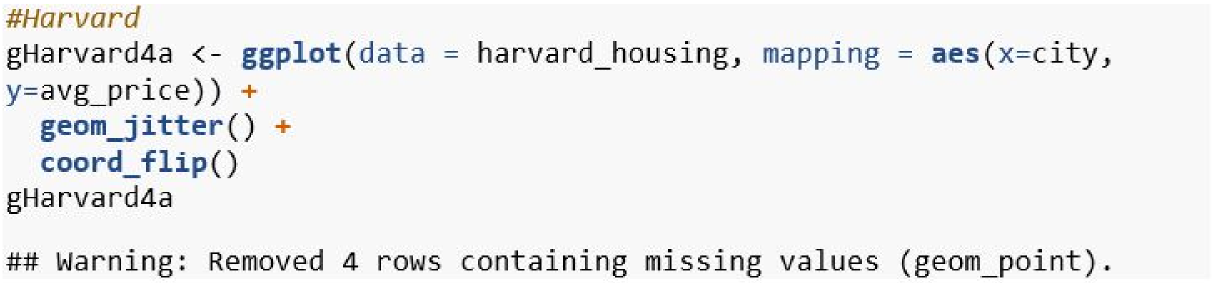




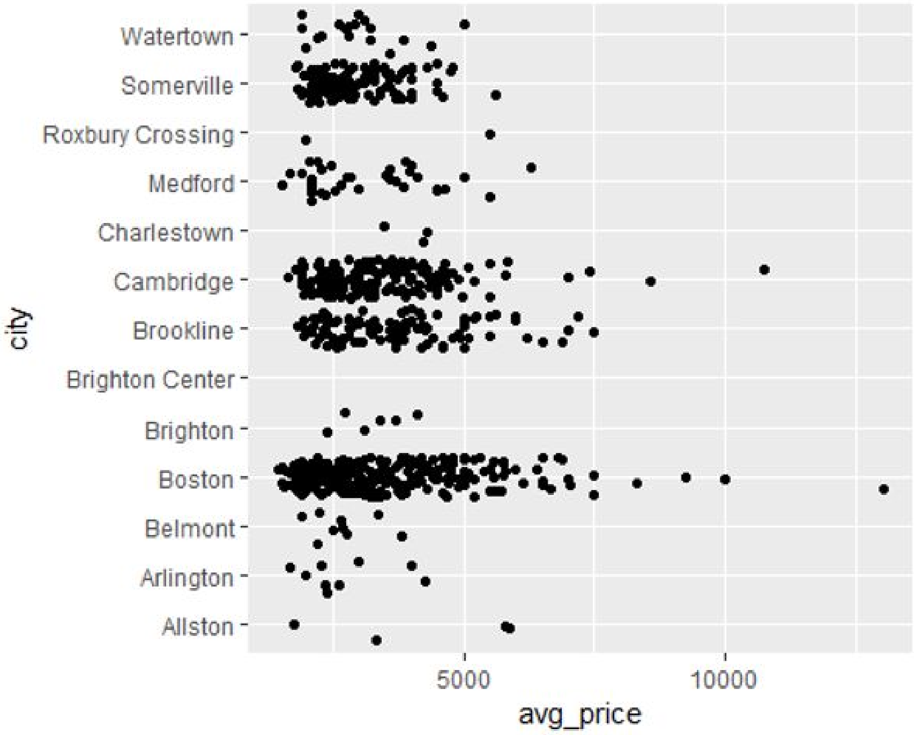
```
#NEU
gNEU4b <- ggplot(data = neu_housing, mapping = aes(x=city, y=avg_price)) +
  geom_boxplot() +
  coord_flip()
gNEU4b

## Warning: Removed 2 rows containing non-finite values (stat_boxplot).
```

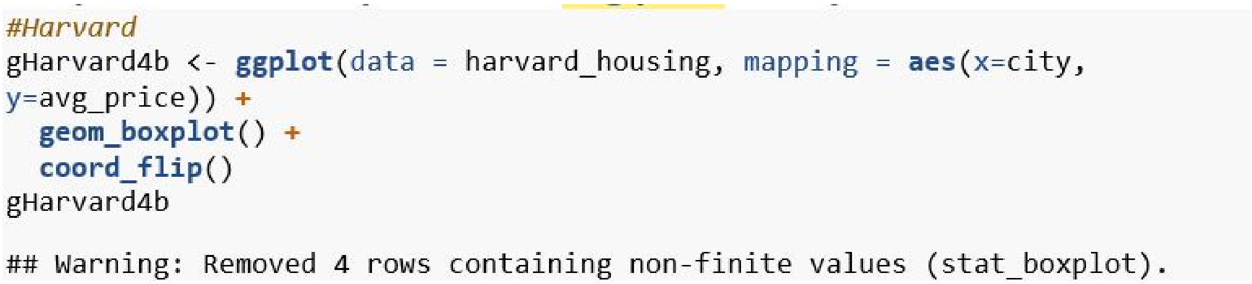


HARVARD

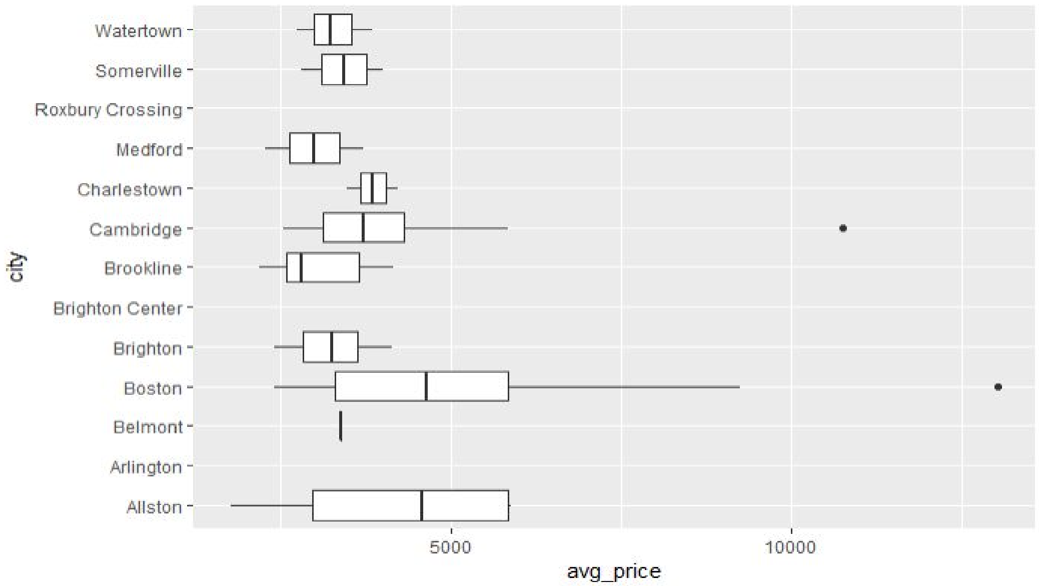


```
#Harvard
gHarvard4a <- ggplot(data = harvard_housing, mapping = aes(x=city,
y=avg_price)) +
  geom_jitter() +
  coord_flip()
gHarvard4a

## Warning: Removed 4 rows containing missing values (geom_point).
```





```
#Harvard
gHarvard4b <- ggplot(data = harvard_housing, mapping = aes(x=city,
y=avg_price)) +
  geom_boxplot() +
  coord_flip()
gHarvard4b

## Warning: Removed 4 rows containing non-finite values (stat_boxplot).
```



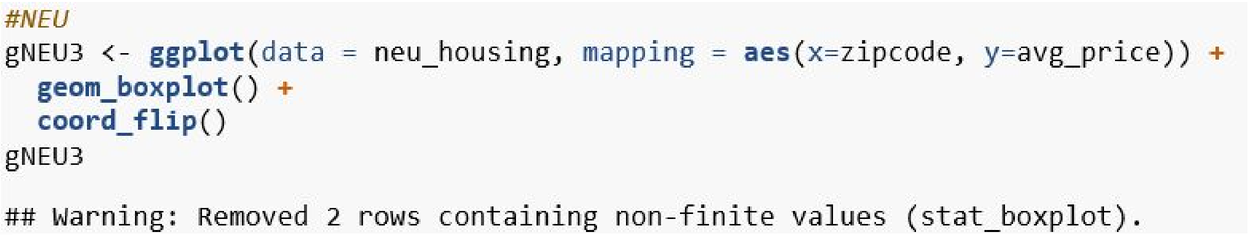
a) Based on the plots we saw from visualizing both Harvard and Northeastern datasets, it remained unclear as to whether there exists a correlation between average price and city/location. Nonetheless, we were able to observe that in certain areas, the range/interquartile range is much larger than those of other areas.

b) Boston has the largest range of average prices of apartments out of all cities, with apartments of prices concentrated near the $1250-$5000 range; then followed by Cambridge, which has the second largest range of average prices of apartments. The fact that Allston apartments have the largest interquartile range is also interesting to note.

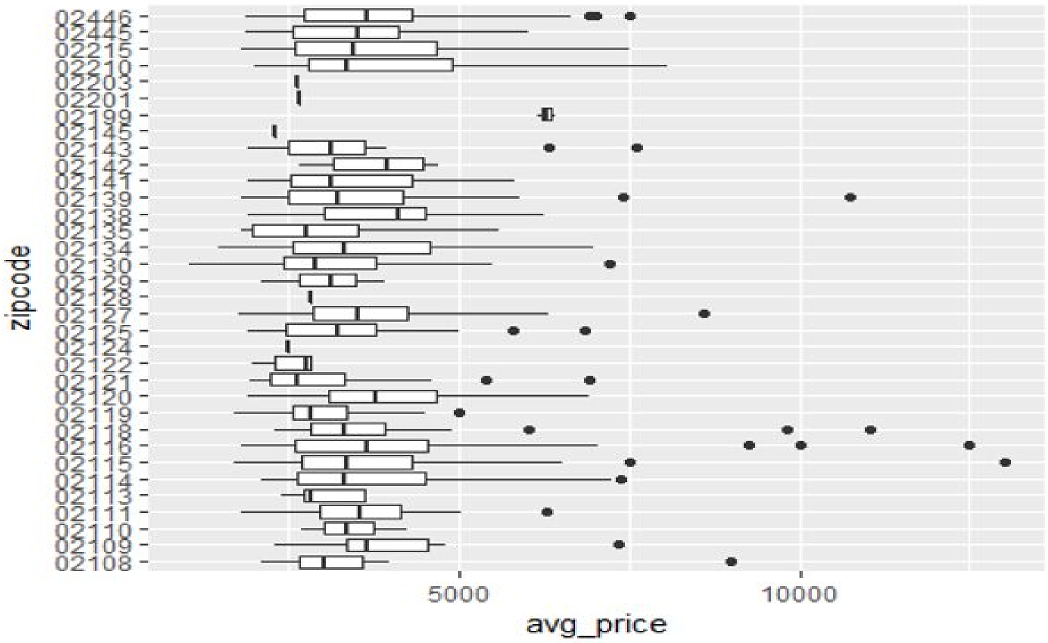
   Northeastern and Harvard: As you can see living in the Boston area is at a much higher rate than living in the Brookline area.

2) Does the zip code affect the price?

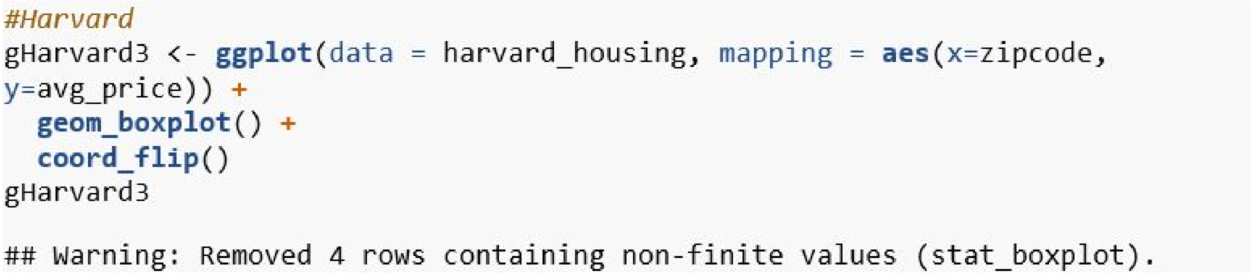
NEU



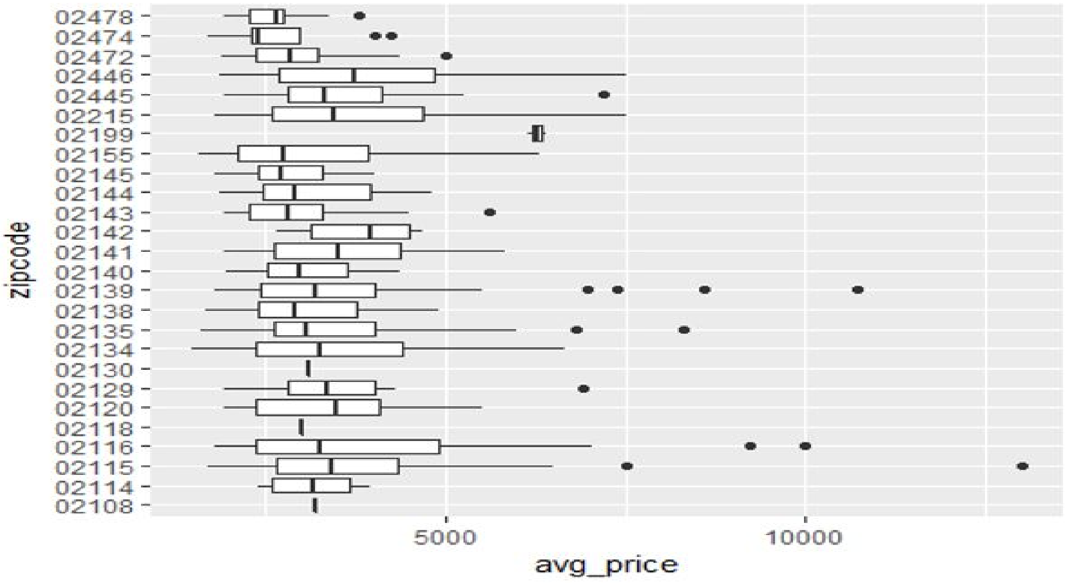
```
#NEU
gNEU3 <- ggplot(data = neu_housing, mapping = aes(x=zipcode, y=avg_price)) +
  geom_boxplot() +
  coord_flip()
gNEU3

## Warning: Removed 2 rows containing non-finite values (stat_boxplot).
```



HARVARD



```
#Harvard
gHarvard3 <- ggplot(data = harvard_housing, mapping = aes(x=zipcode,
y=avg_price)) +
  geom_boxplot() +
  coord_flip()
gHarvard3

## Warning: Removed 4 rows containing non-finite values (stat_boxplot).
```



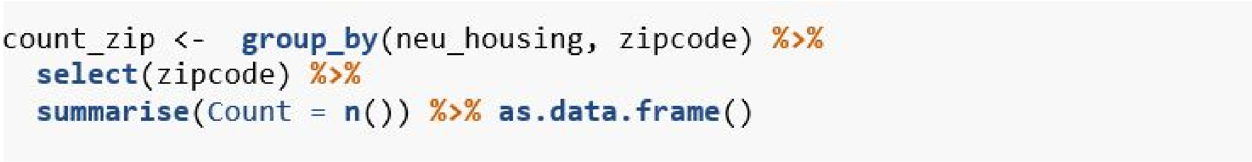
a) We plotted this by plotting the zip codes with the min price, max price and average price. This gave us an idea of what zip codes had the highest prices vs which didn't.

Northeastern: Disregarding the outliers(the dots on the graph), it can be noted that the area with zip code 02446(Brookline) has the largest range of average price of apartments.

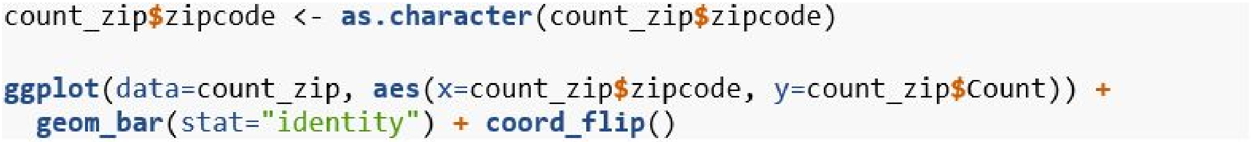
Harvard: Disregarding the outliers(the dots on the graph), it can be concluded that the area with zip code 02215(Boston) has the largest range of average price of apartments, followed by areas with zip code 02446(Brookline), 02134(Brighton), 02116(Boston). Again, exceptions include 02108(Boston), 02130(Boston) and 02118(Boston) which have a very small range. However, this might be again caused by the lack of housing data in those areas.

3) How many apartments are available in each zip code area?

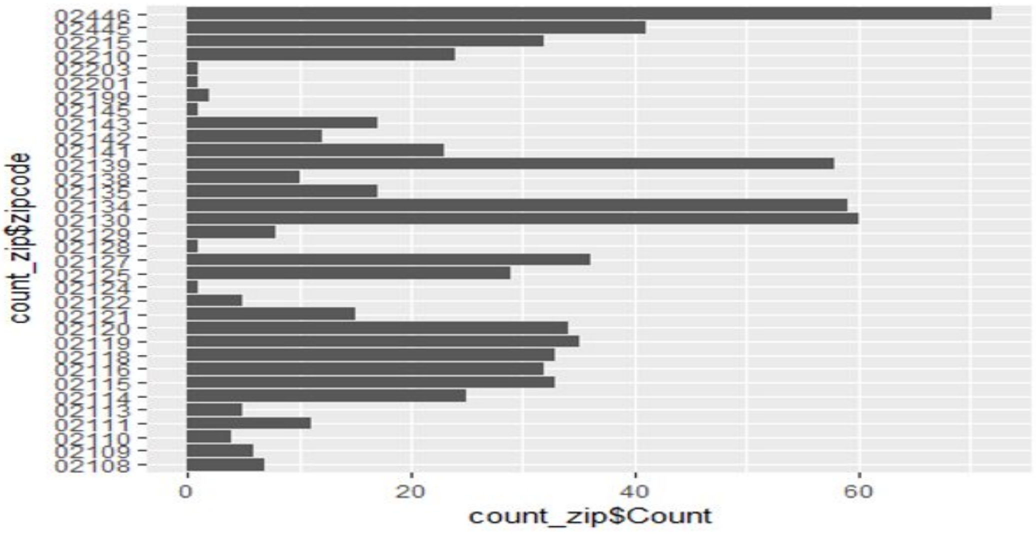
NEU



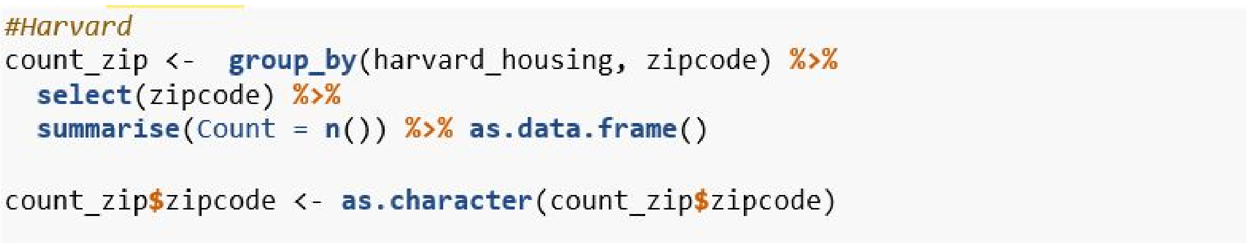
```r
count_zip <-  group_by(neu_housing, zipcode) %>%
  select(zipcode) %>%
  summarise(Count = n()) %>% as.data.frame()
```



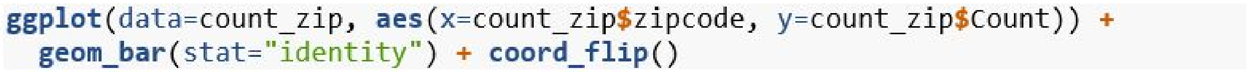
```r
count_zip$zipcode <- as.character(count_zip$zipcode)

ggplot(data=count_zip, aes(x=count_zip$zipcode, y=count_zip$Count)) +
  geom_bar(stat="identity") + coord_flip()
```



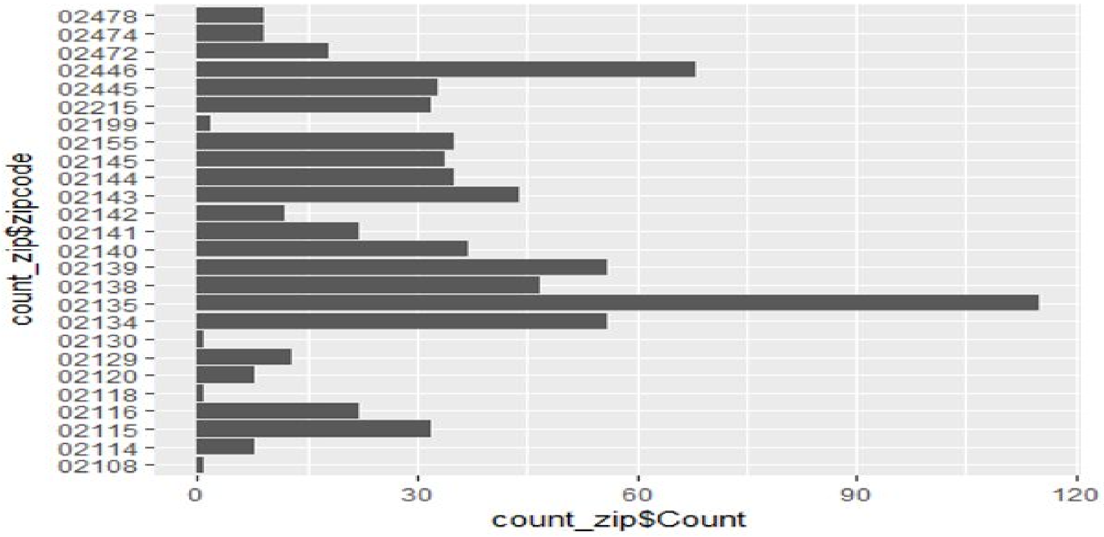
HARVARD



```
#Harvard
count_zip <-  group_by(harvard_housing, zipcode) %>%
  select(zipcode) %>%
  summarise(Count = n()) %>% as.data.frame()

count_zip$zipcode <- as.character(count_zip$zipcode)
```



```
ggplot(data=count_zip, aes(x=count_zip$zipcode, y=count_zip$Count)) +
  geom_bar(stat="identity") + coord_flip()
```



a) This was done by first writing some dplyr code to create a separate dataframe with the required information and then using that dataframe to plot the required data using a bar chart.

Northeastern: This shows us that the most number of apartments in the NEU area are in the 02446 zip code The least number of apartments in the NEU area are in the 02203/02201/02145/02128/02124

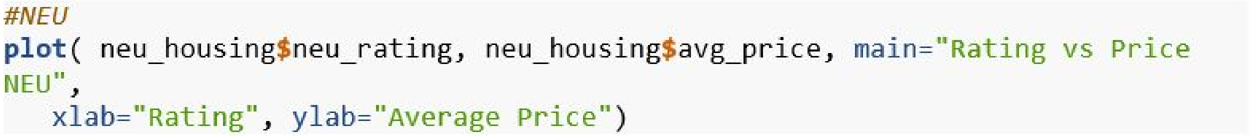
Harvard: As you can see, most number of apartments in the Boston area

02135 - The exact opposite of NEU! Least number of apartments in the zip code is 02130/02118 and 02108

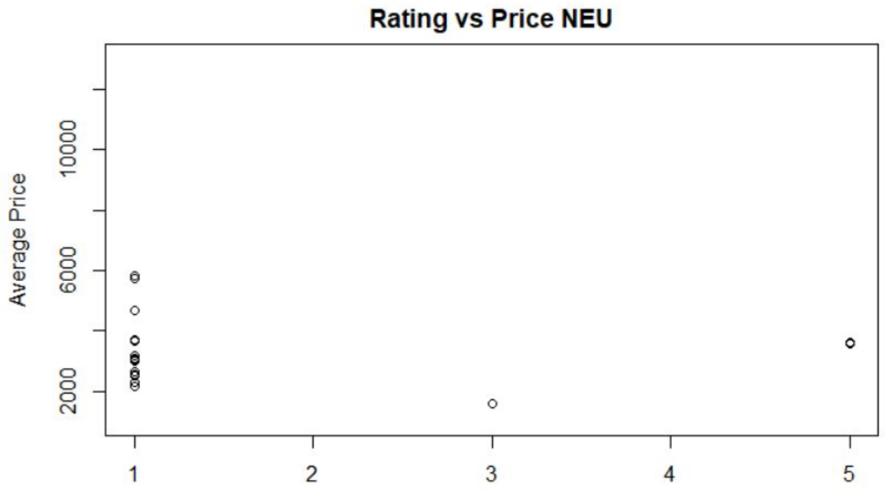
4) Does the price/city affect the rating of the apartment?

   a) Price:

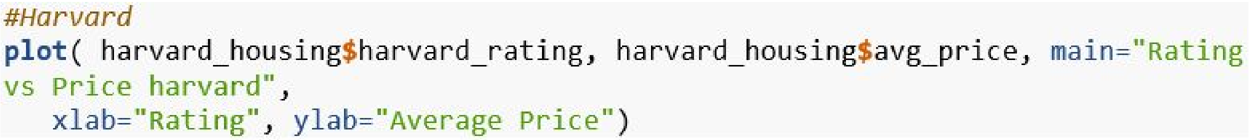
NEU



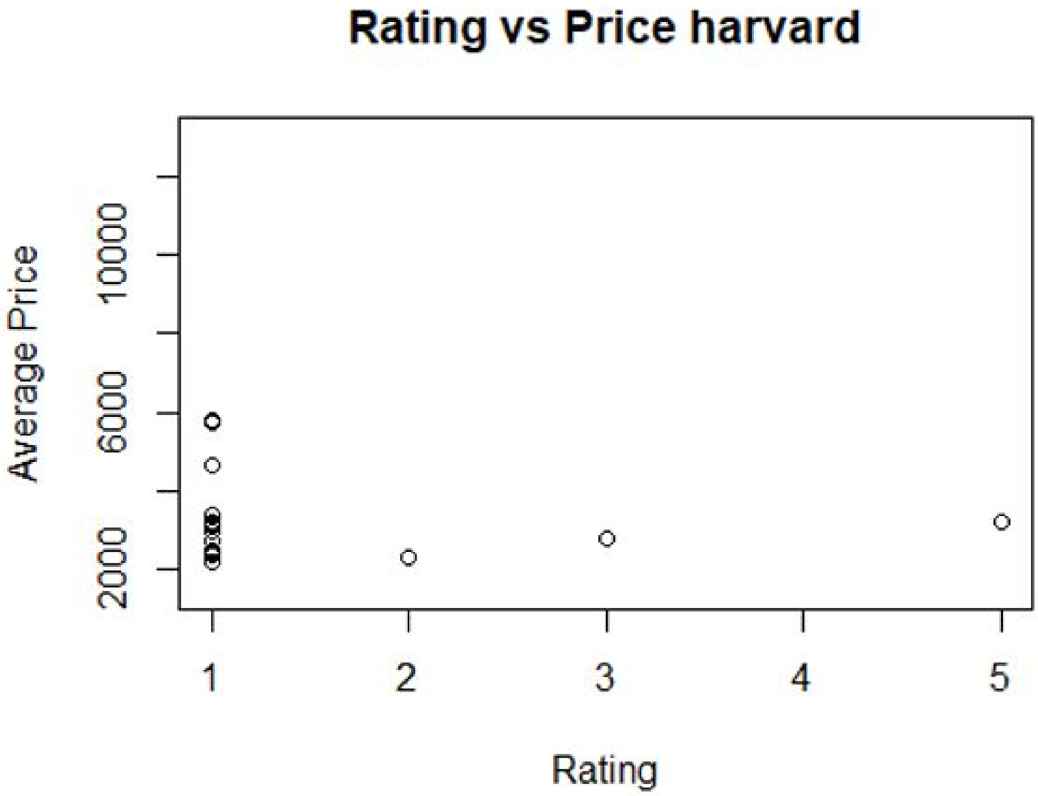
```r
#NEU
plot( neu_housing$neu_rating, neu_housing$avg_price, main="Rating vs Price NEU",
   xlab="Rating", ylab="Average Price")
```



HARVARD:



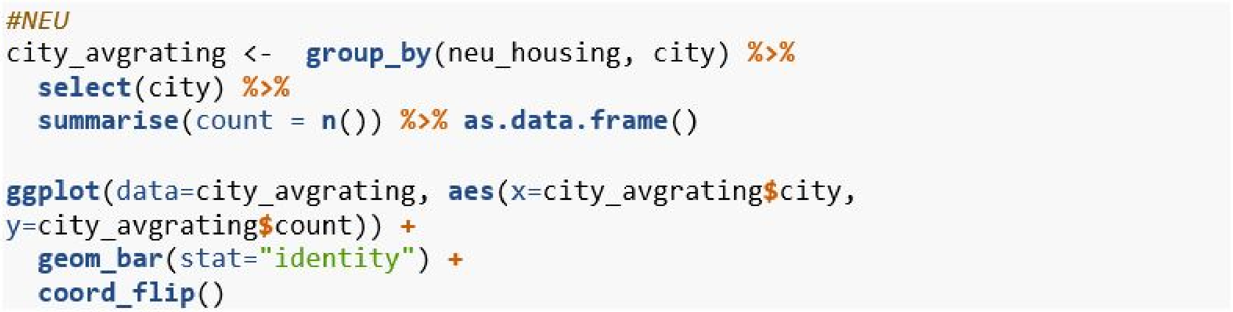
```
#Harvard
plot( harvard_housing$harvard_rating, harvard_housing$avg_price, main="Rating
vs Price harvard",
   xlab="Rating", ylab="Average Price")
```



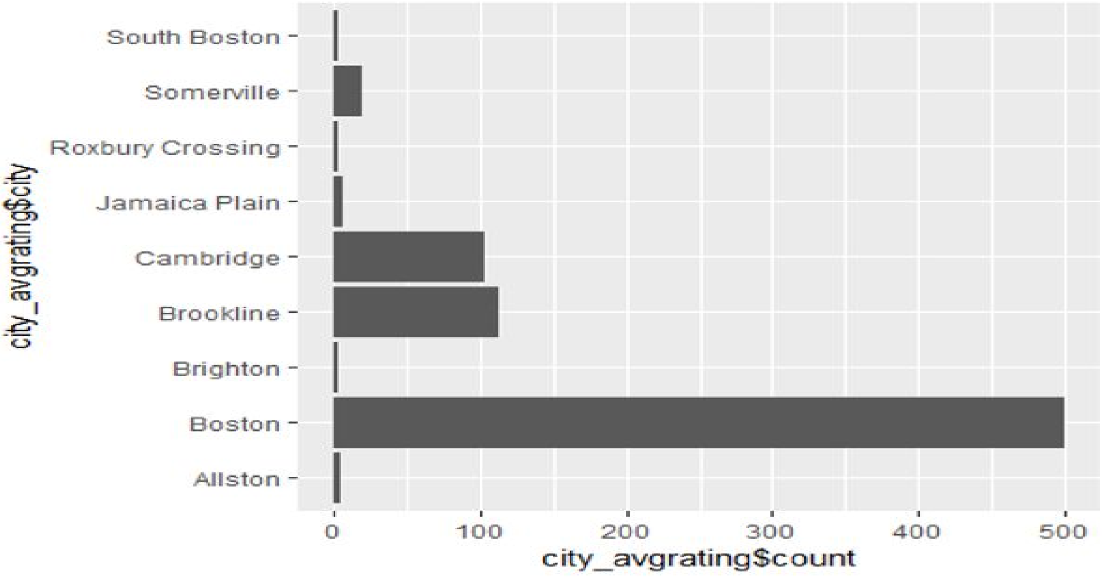
i) Using a scatter plot to plot the data points using the average price and the rating, it was unclear as to whether rating had a direct correlation with price since there were a lot of missing data points.. This was due to the fact of missing data.

b) City:

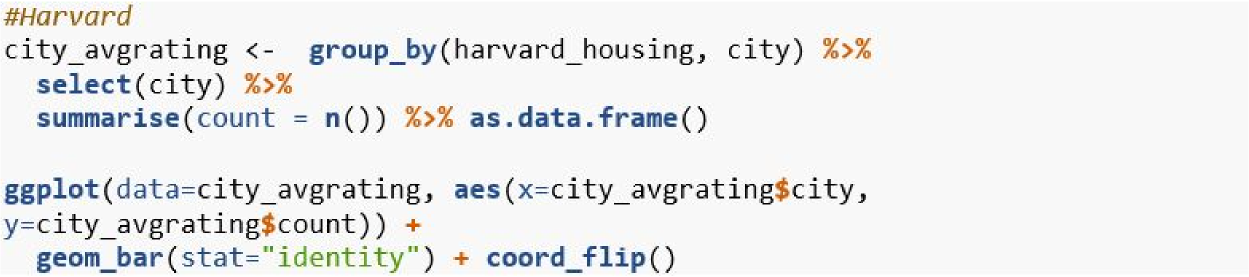
NEU



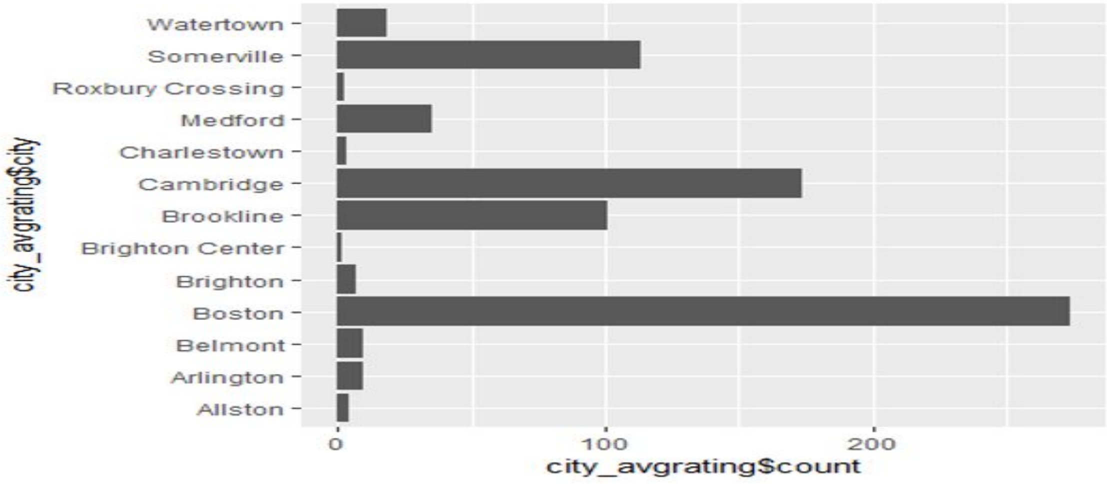
```
#NEU
city_avgrating <-  group_by(neu_housing, city) %>%
  select(city) %>%
  summarise(count = n()) %>% as.data.frame()

ggplot(data=city_avgrating, aes(x=city_avgrating$city,
y=city_avgrating$count)) +
  geom_bar(stat="identity") +
  coord_flip()
```



HARVARD



```
#Harvard
city_avgrating <-  group_by(harvard_housing, city) %>%
  select(city) %>%
  summarise(count = n()) %>% as.data.frame()

ggplot(data=city_avgrating, aes(x=city_avgrating$city,
y=city_avgrating$count)) +
  geom_bar(stat="identity") + coord_flip()
```



i) Using a bar chart to plot the data, it was apparent to see that the data showed that certain cities had higher ratings as compared to others. For the Northeastern dataset, it showed that the Boston and Brookline area had the highest ratings while the Roxbury Crossing area had the lowest. As for Harvard, Boston area had the highest ratings while the Brighton Center area had the lowest.

CONCLUSIONS:

Based on price, rating, proximity and location, it seems like the Northeastern University has better housing statistics as compared to Harvard University. While Harvard students have more options, living in the city of Boston has its own perks.

FUTURE PLANS:

Based on the data and analysis we have done, a few future ideas to better our project include:

1) Getting the amenities to better analyze the data
2) Getting data from websites like Yelp/Google to analyze proximity of restaurants, gyms and stores. This would further help us analyze if this is an important factor for college students when considering where to live.
3) Actually collecting data from students who go to Northeastern and Harvard through surveys. This way the data is more accurate.