PROJECT I ECON 494 F20

Dr. Levkoff

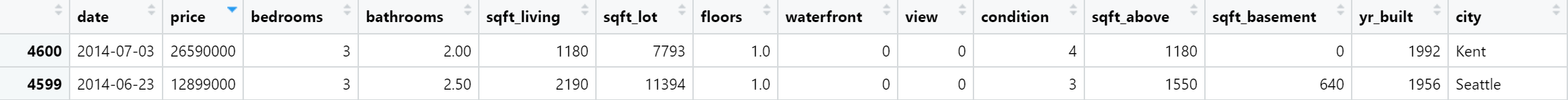
Xi Zhang(Simon)

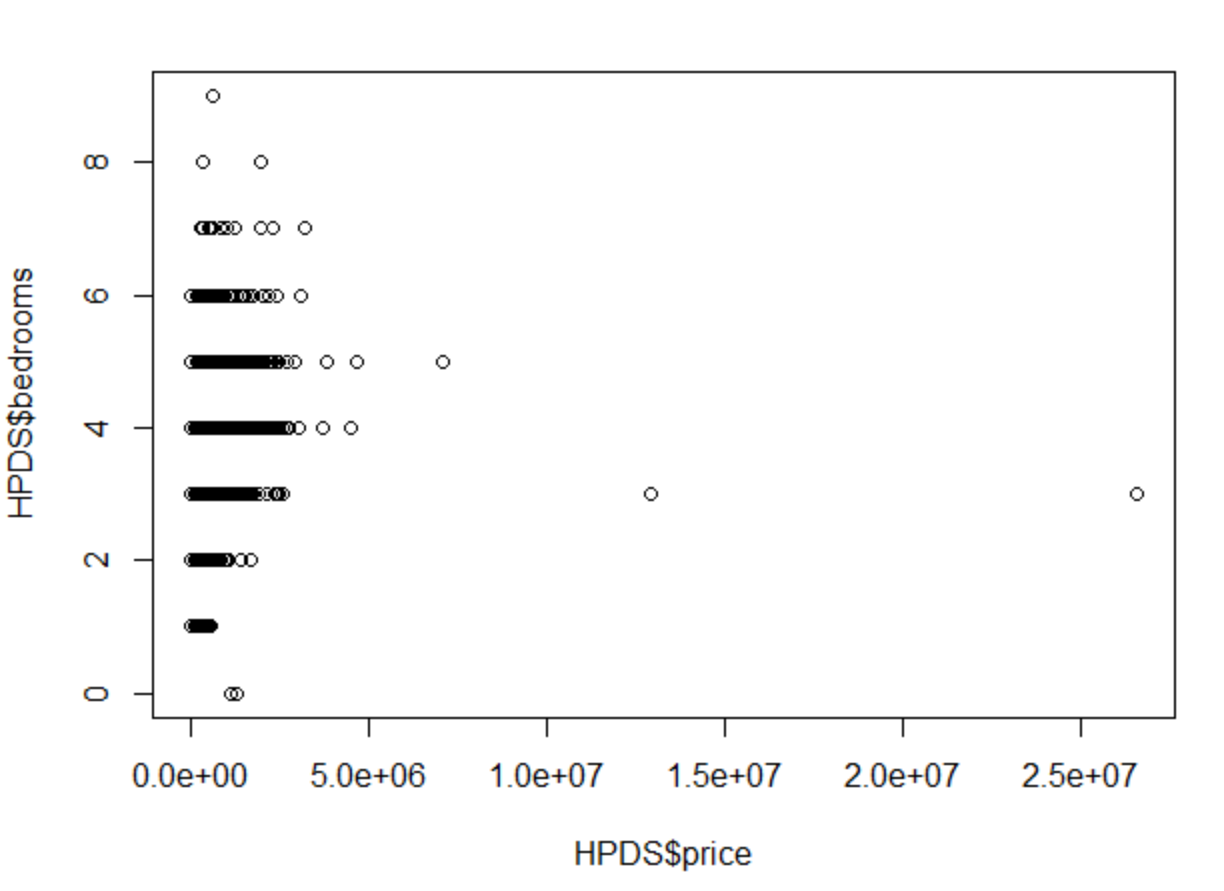
10/25/2020

**Preprocessing**

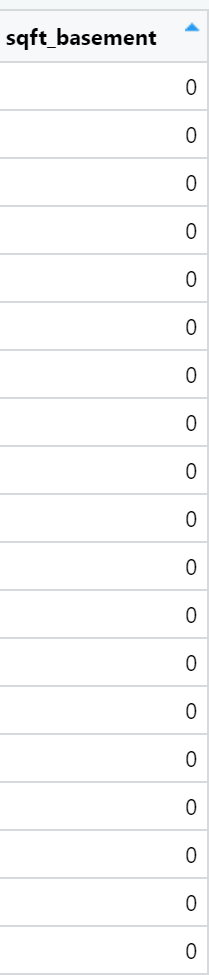
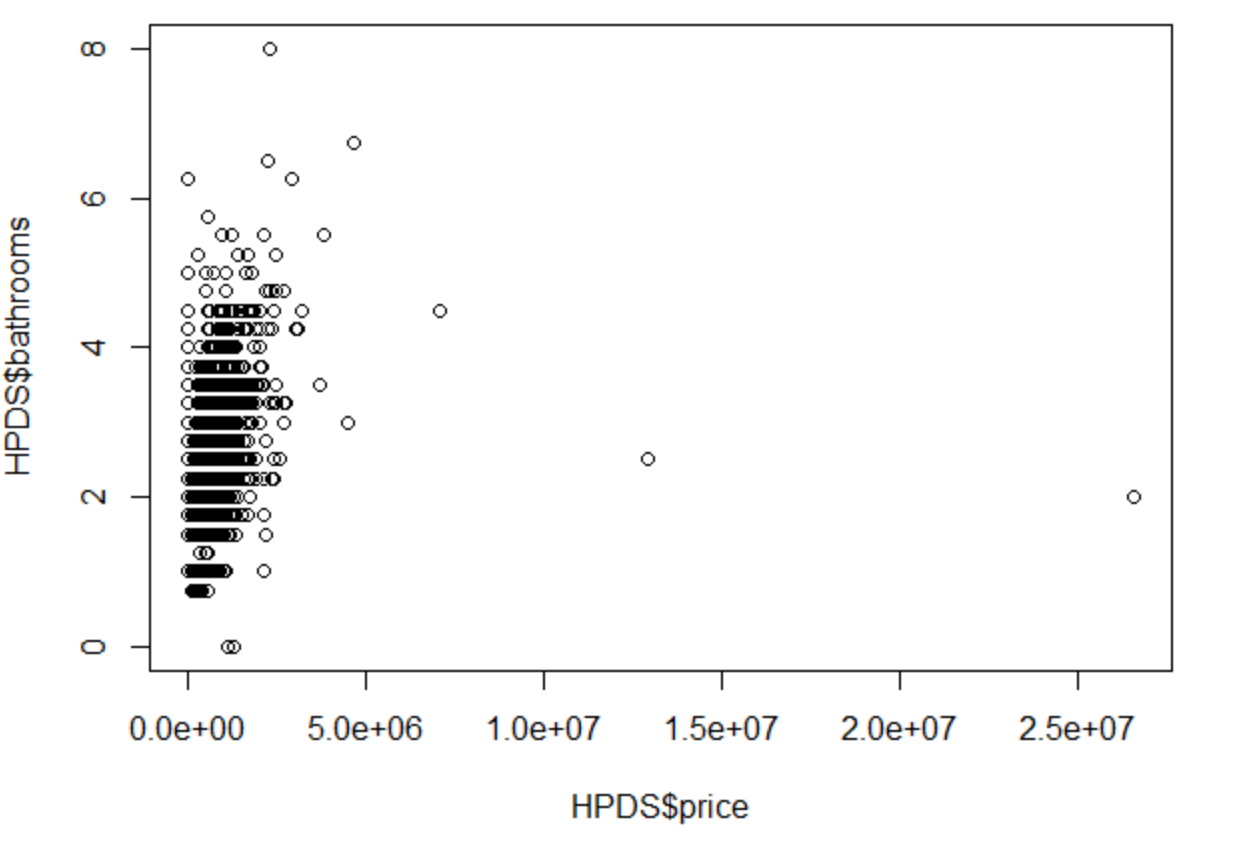
My first step to this project was to go on to kaggle.com and search for business-related datasets by filtering only CSV files. Then I ended up with this dataset called housing price prediction. Since the variable of interest in the dataset we want to study is the housing price, I will construct multiple plots to study the relationship between each variable and the housing price variable. This dataset is a typical cross-sectional dataset having quantitative variables: “date”, “housing price”, “living square footage”, “lot square footage”, “above ground square footage”, and “basement square footage”. Also, there are categorical variables such as “number of bedrooms”, “numbers of bathrooms”, “floors”, “view”, “condition”, “build year”, “city”, and a dummy variable “waterfront”.

**Cleaning**

In terms of cleaning data, the first thing was to import the excel file into Rstudio. So I installed the “readxl” package and imported the data as a new variable called HPDS(housing price dataset) using the “read\_excel” function. Then I did a quick check on every single variable related to the housing price variable by using the “plot” function. The first thing I did was to create a new variable called HPDS\_b to compare to the original dataset after the cleaning steps. Then I found out that the price variable had some observations with non-logical data. For example, the observations as shown: . Even though some of the prices were super high, it didn’t have many rooms nor living space. Plus there are also many observations with zero price, which would be false data. So I set all observations with a price of more than 4 million and less than 80,000 as NAs to avoid that outlier because there are only 5 observations with a higher price than 4 million. The next thing I noticed in the testing plot was that many observations had a bedroom number of either 0 or more than 7 as shown below.



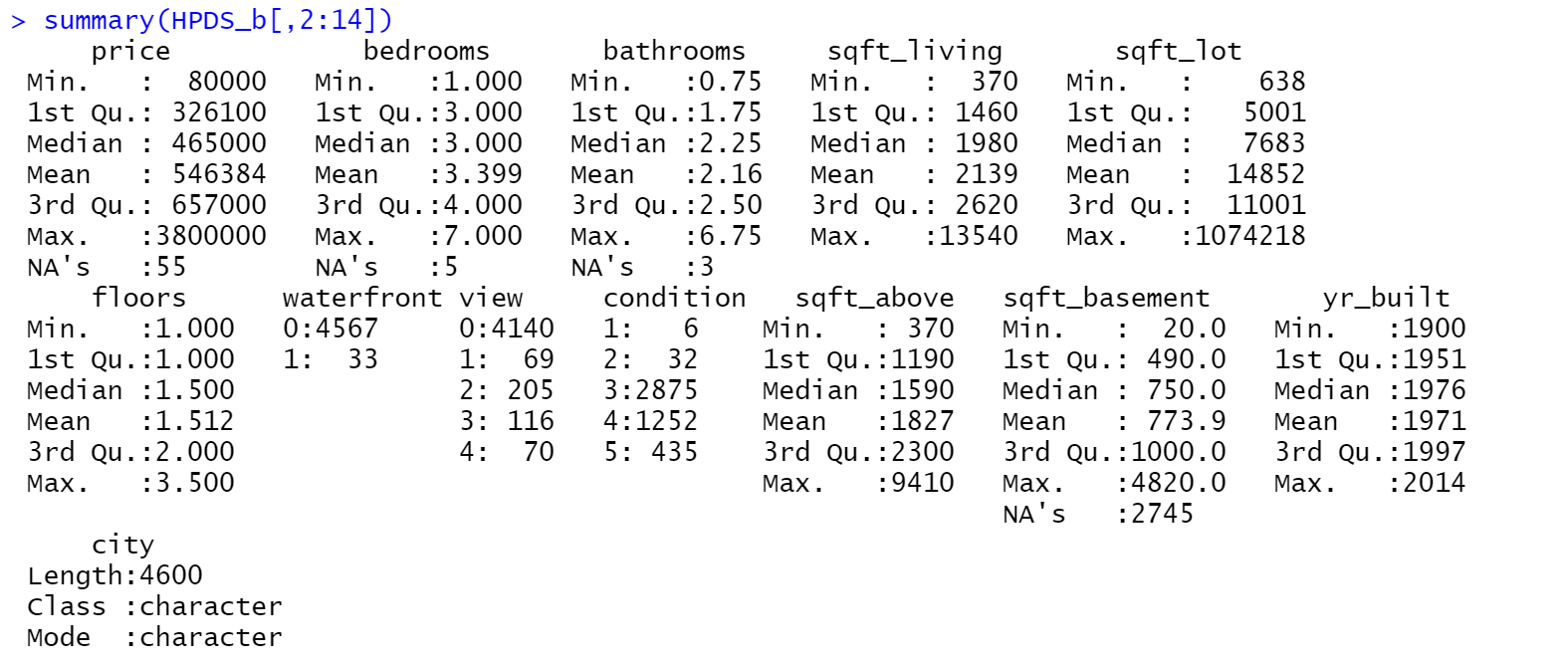
To make it logical, I set all the non-logical bedrooms numbers as NAs. Together with bedroom variables, the variable bathrooms and basement square footage also have the same problem.



So similarly I set bathrooms and basement square footage with a value equal to 0 as NAs, where for bathrooms I also set any value larger than 7 as NAs. For variable “sqft\_lot”, I discovered in the further section with scatter plots that there are a couple outliers with values larger than 400,000. Thus I set all values greater than 400,000 as NA’s for the lot square footage variable, and we will be able to see the detailed plot in the next section. Since I wanted to show how categorical variables impacted price, I set variables including waterfront, condition, and views as factors, because I cannot plot with categorical variables unless I change it into factors variables. When I was messing with ggplot2, since the best way I found to explain categorical variables such as bedroom, bathrooms, floors, condition, and views was to use the line plot using the average of each category for all these variable, I added a new variable for each of these variables using the mean price for every category in the variable(as I will show in the next section how it will plot the graph). The last new variable I created was the mean value of the variable “build year” in terms of waterfront using the ddply function. I also used function view() and dim() to check my HPDS\_b if any observation was lost during my work. Then at the end, I finished this part by checking each of the new variables that I created with the mean() function to see if the NA’s were successfully created(explicitly for variables: “price”, “bathrooms”, “bedrooms”, “basement square footage”). I was lucky that the dataset came pretty clean, but there will be more stuff to talk about in the exploratory part.

**Exploratory analysis**

The first thing I did for my exploratory analysis was to create the summary statistics for all the variables in my dataset using a summary(). The output is shown below:



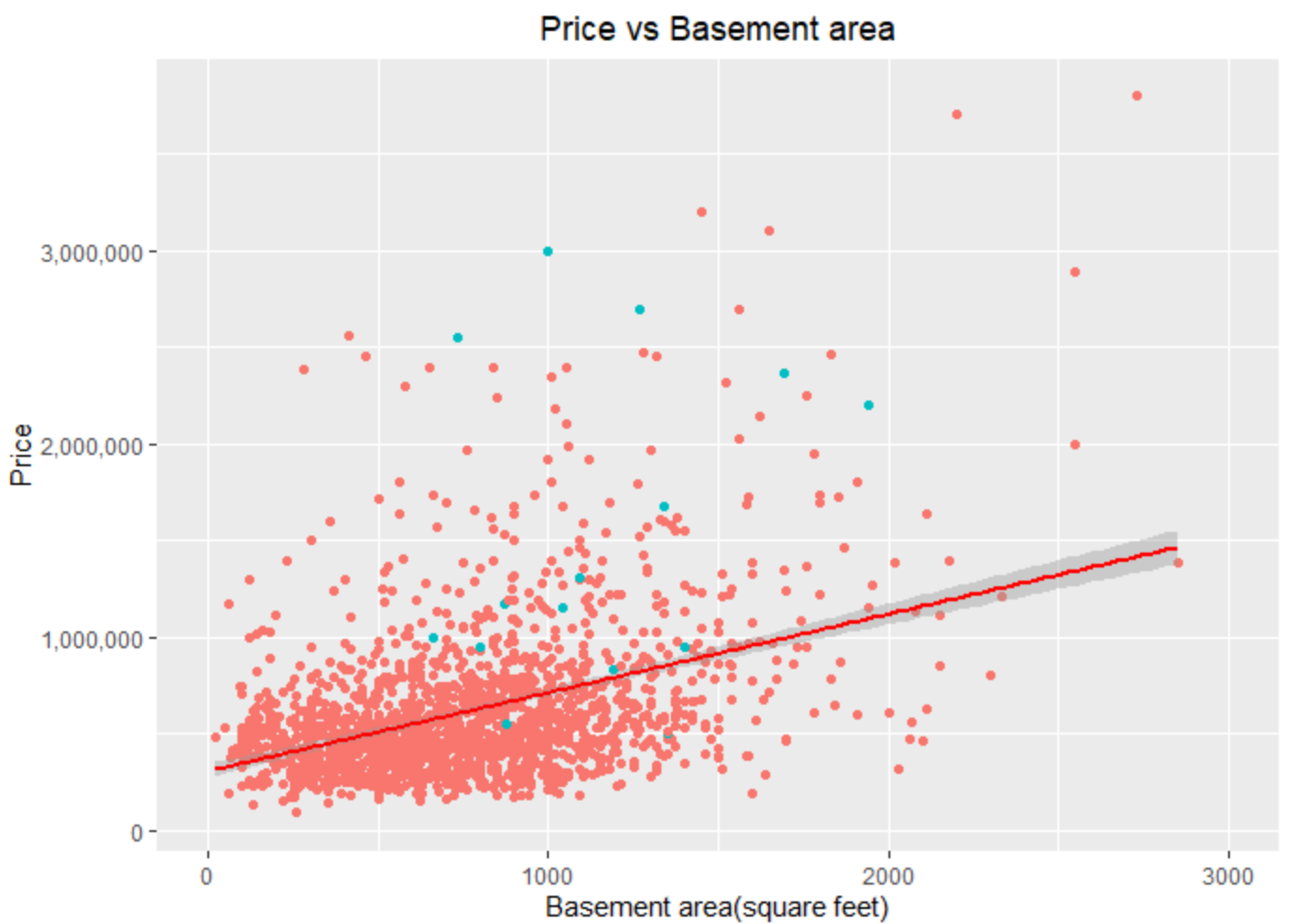
We could see that the mean value of each variable has been shown. Also, we could see how many NA’s were created for each variable. It is interesting to see that more than half of the house sold had no basement or didn’t report.

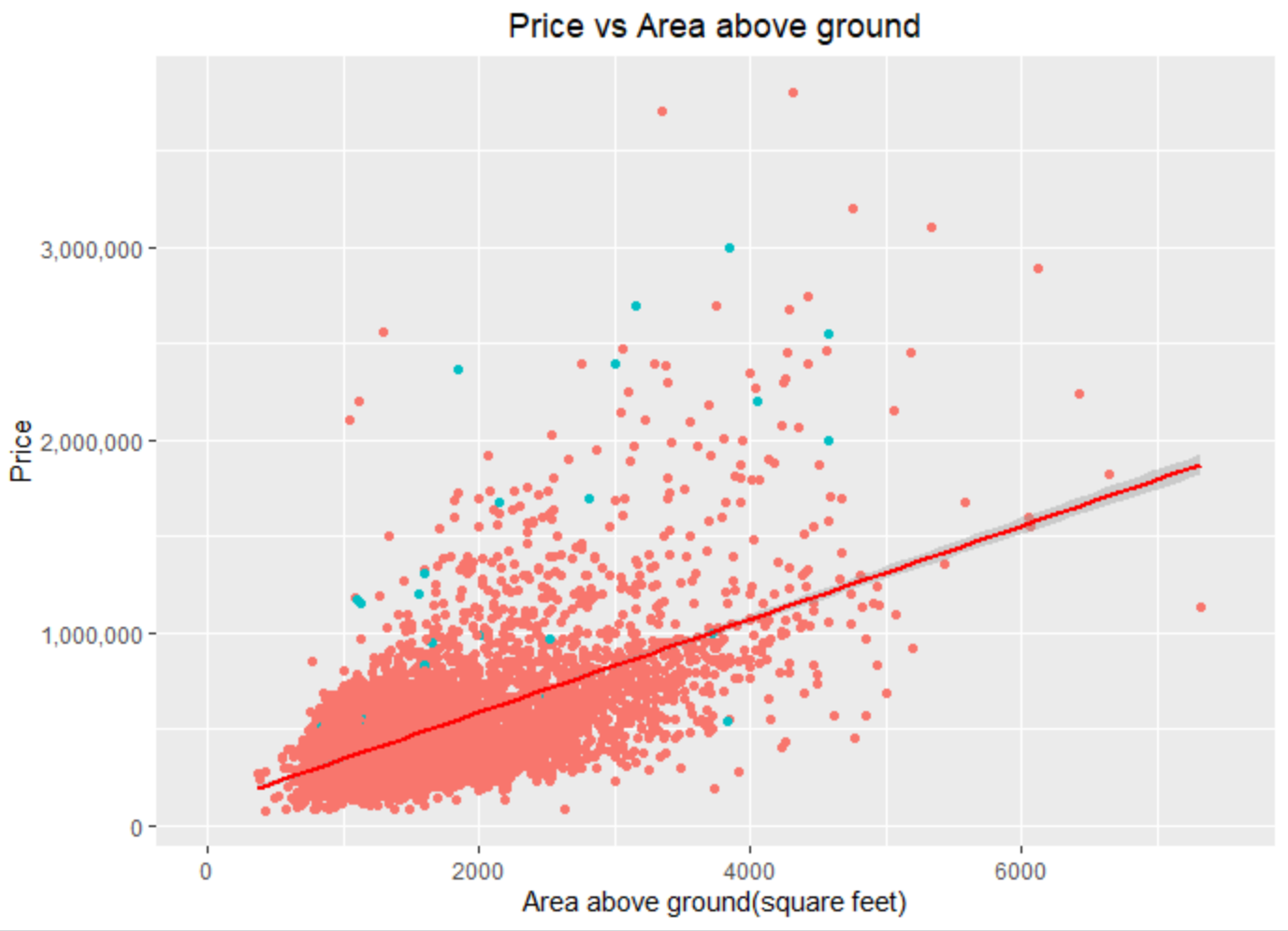
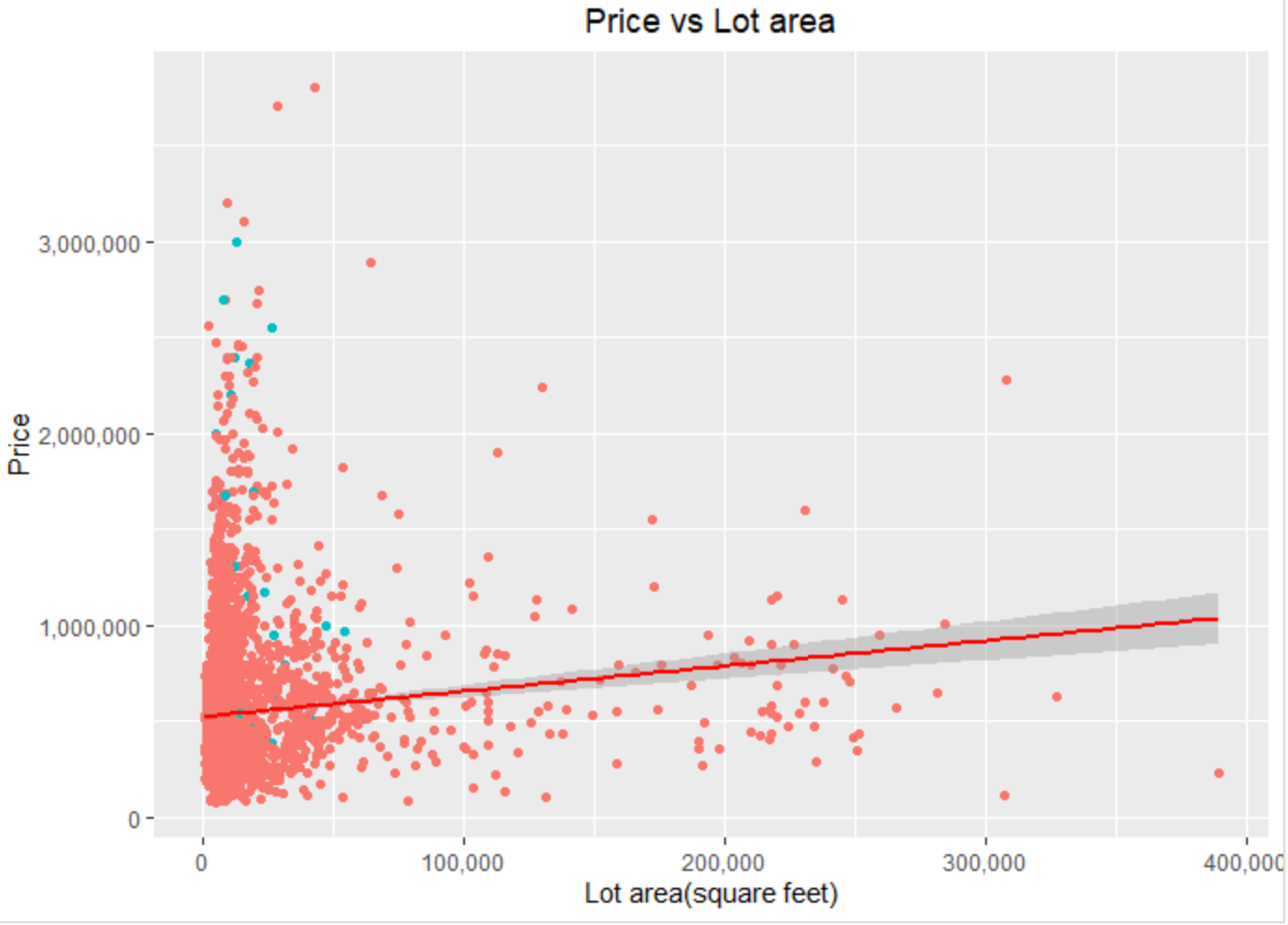
Next, I installed the ggplot2 package, and the first part of my plot will be focusing on quantitative variables. The first plot was to explore the relationship between price and living area base using a scatter plot.



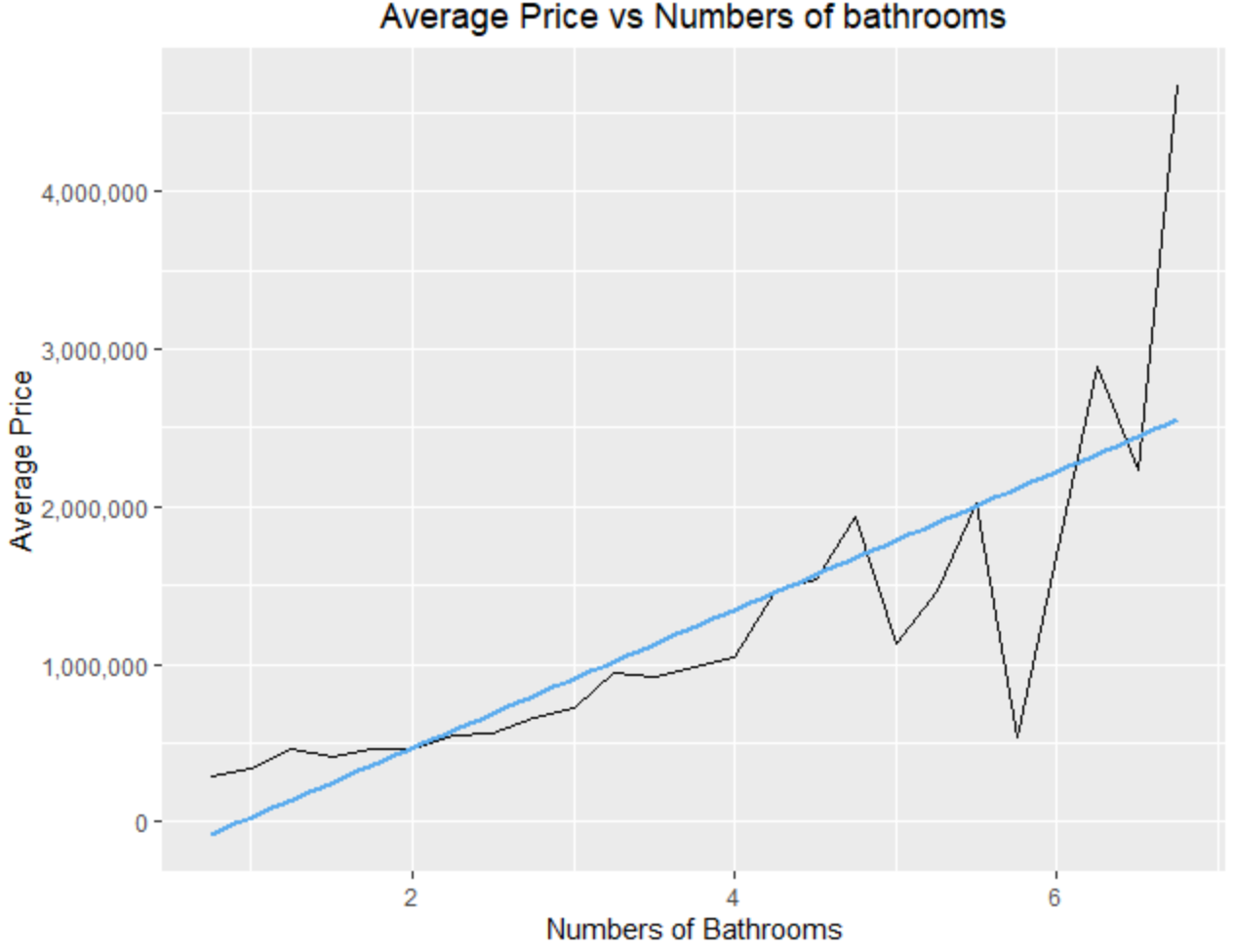
In this plot, we can see a positive linear relationship between the price and living area as the trend line shown to us. Also, it seems like most of the houses sold were under 4000 sqft living area and 1 million dollars. In addition to that, I have changed the color of the houses that are waterfront. So we can also see that waterfront houses are very uniformly distributed in all ranges of price and living areas instead of clustered at the lower price/living area.

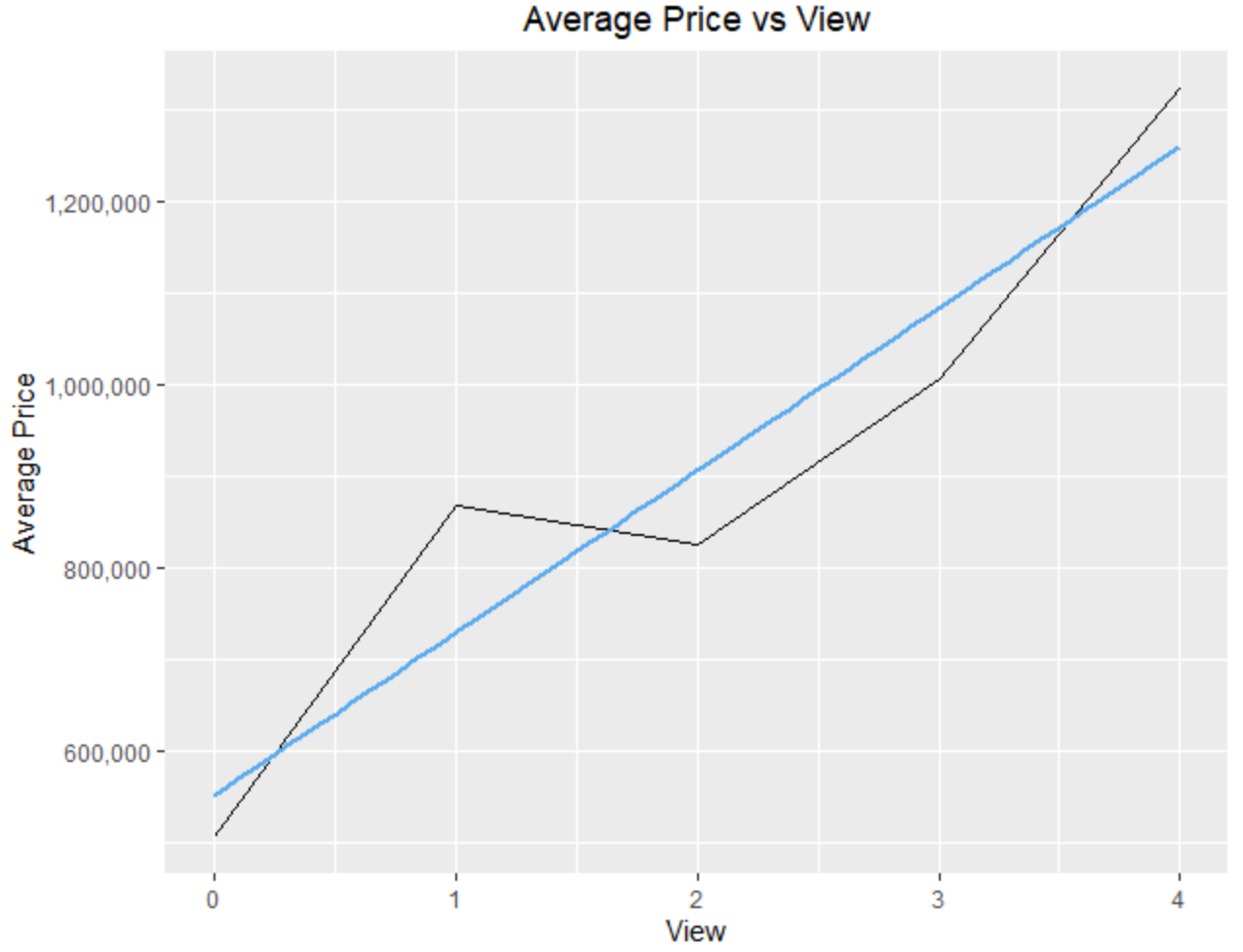
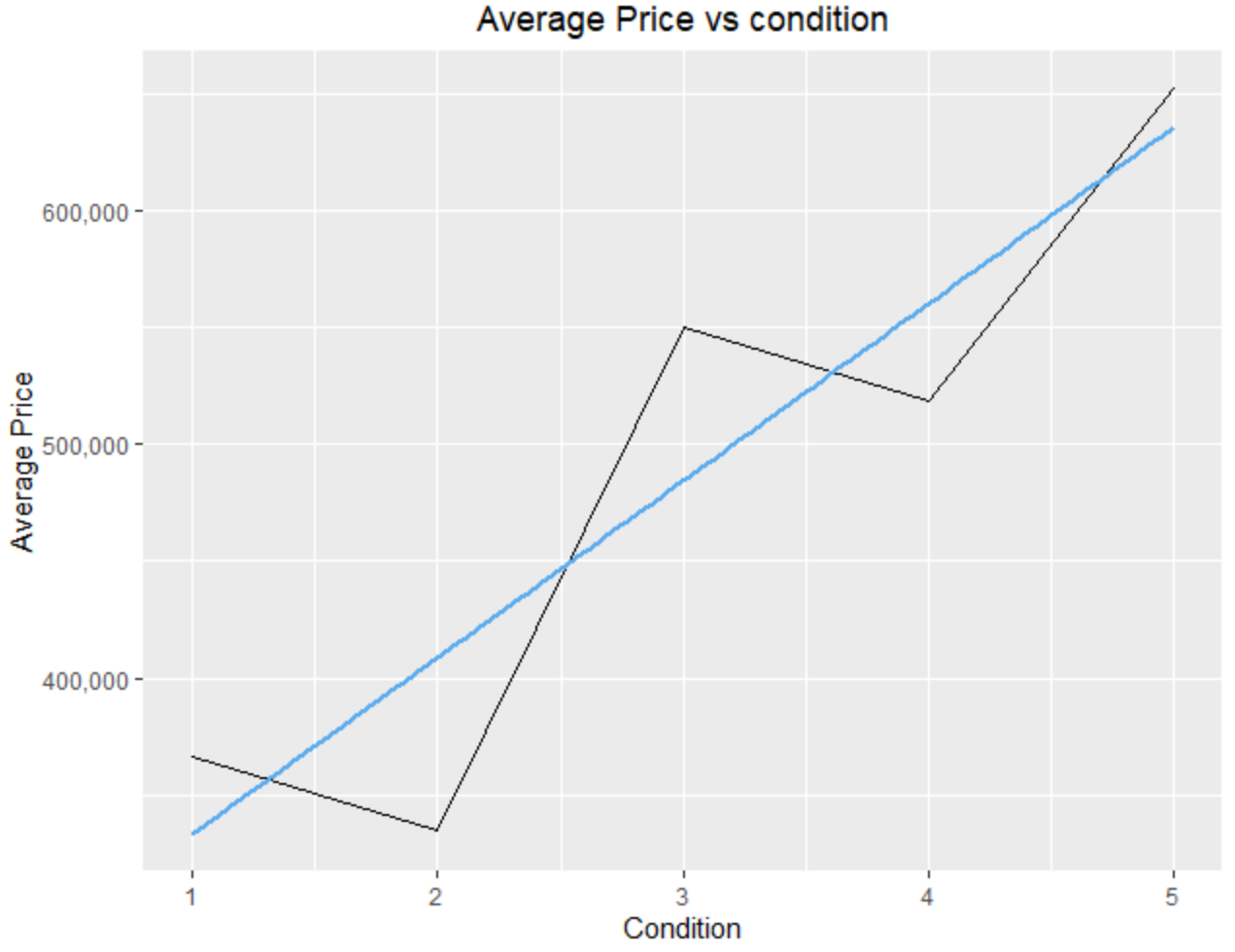
Similarly, I also did the same scatter plot for variables “lot area”, “areas above ground”, and “basement area” all respect to price. Not surprisingly, they all have a positive linear relationship with price. The area above ground has a stronger positive linear relationship relative to the basement area and lot area.





In the part, I will be focusing on category variables including “numbers of bathrooms”, “numbers of bedrooms”, “condition” and “view” with respect to the average price of houses sold(the variable we created in the last section). From the graphs below we could see that every variable had a strong positive linear relationship with price. We could also conclude that the numbers of bedrooms and conditions had a stronger positive relationship with price than the numbers of bathrooms and views.

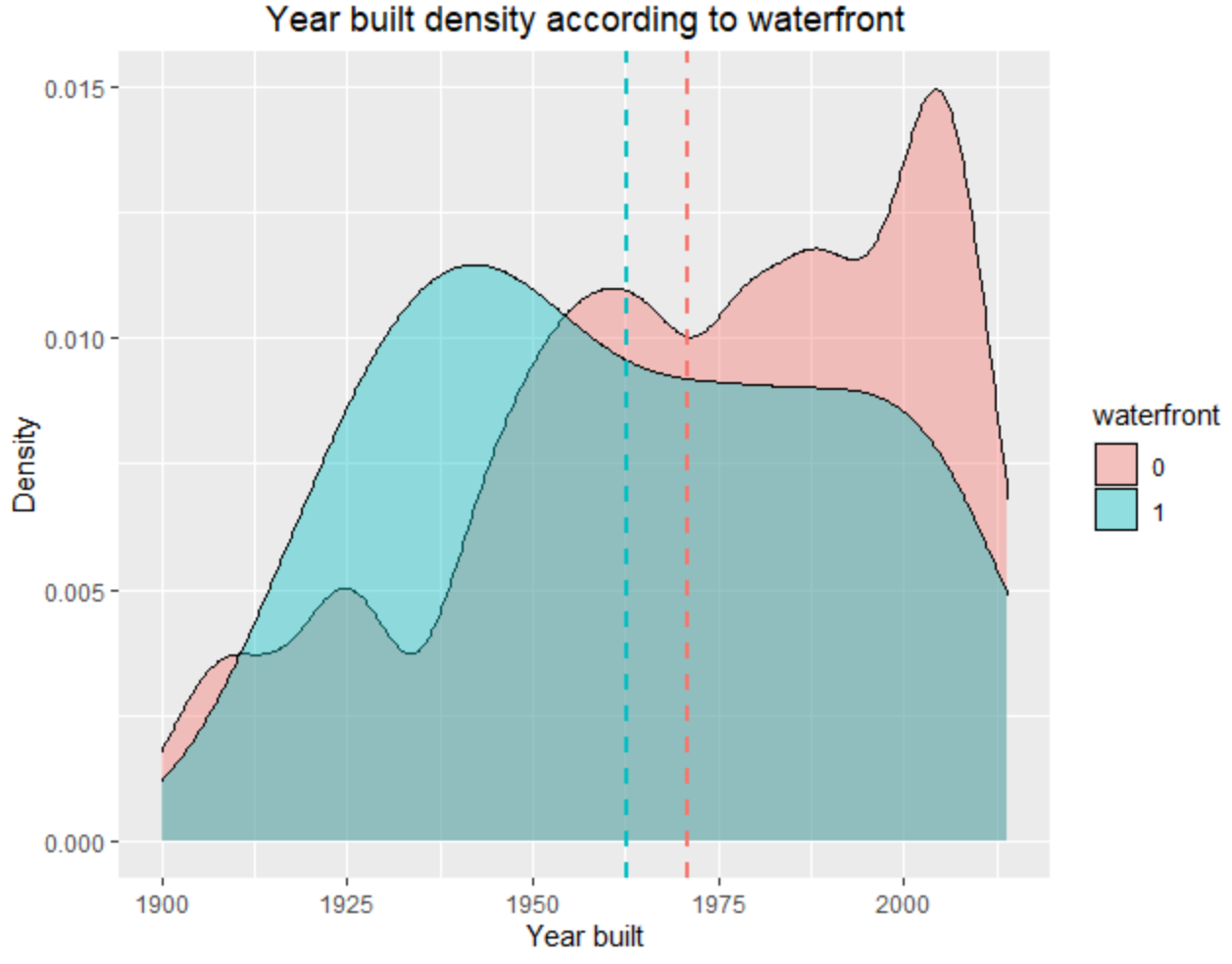




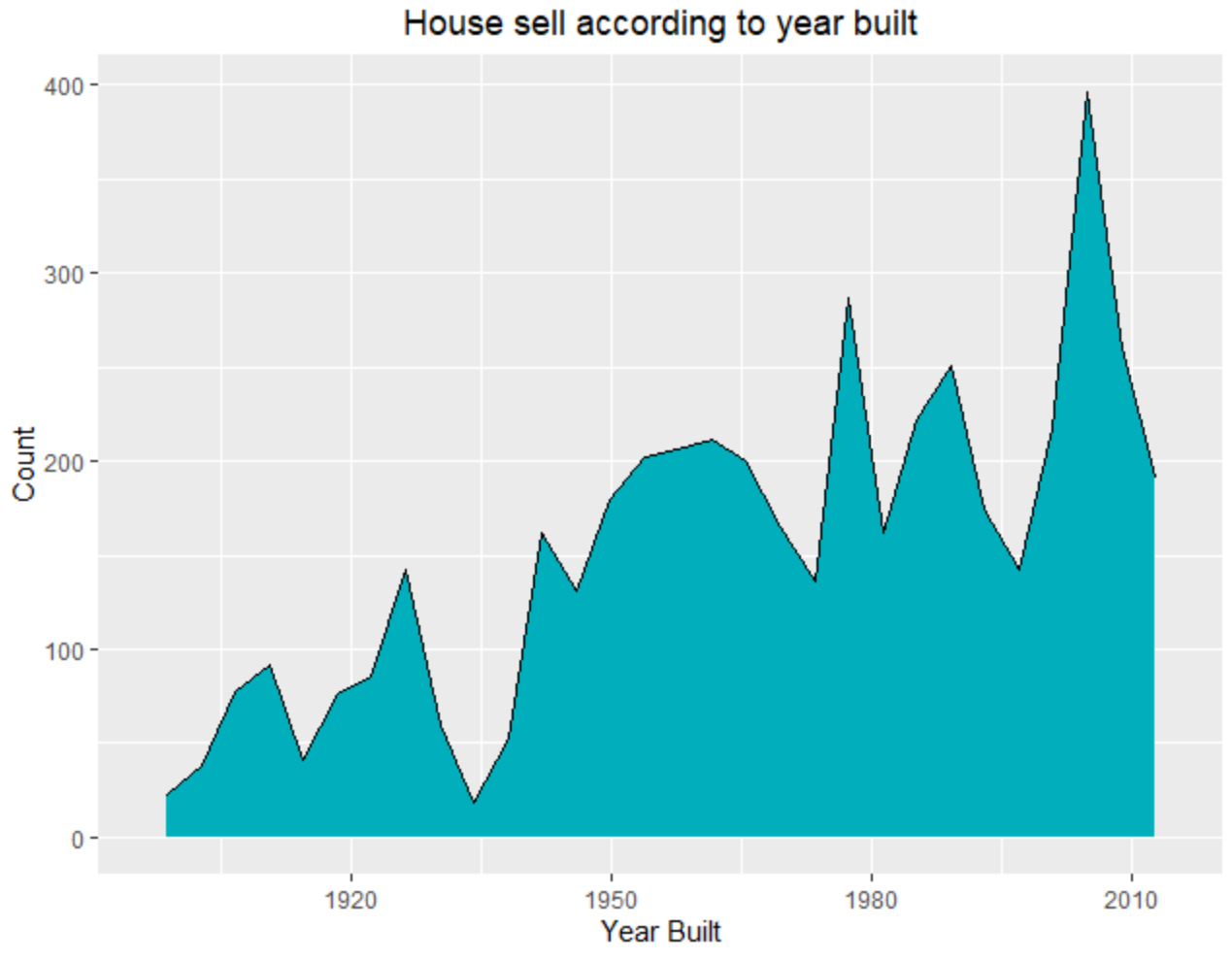
In terms of the dummy variable “waterfront”, I have constructed a violin plot with no waterfront(0’s) versus waterfront house(1’s). From the graph, we can see that being waterfront has a significantly higher average price. Additionally, that, being waterfront also has a higher 25% tail and higher outliers, which shows that the distribution is going towards higher prices rather than gathered at lower prices like not having waterfront.



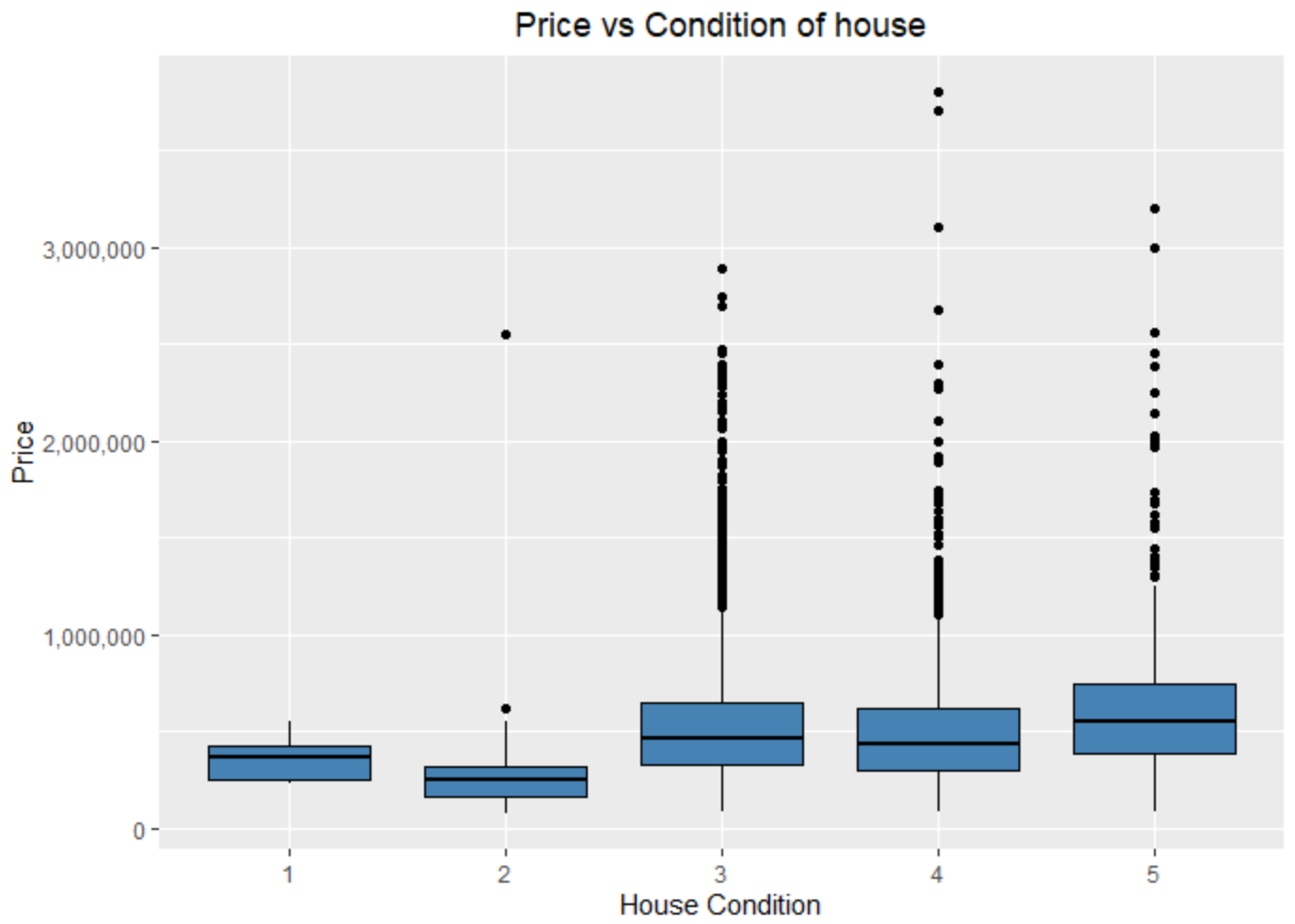
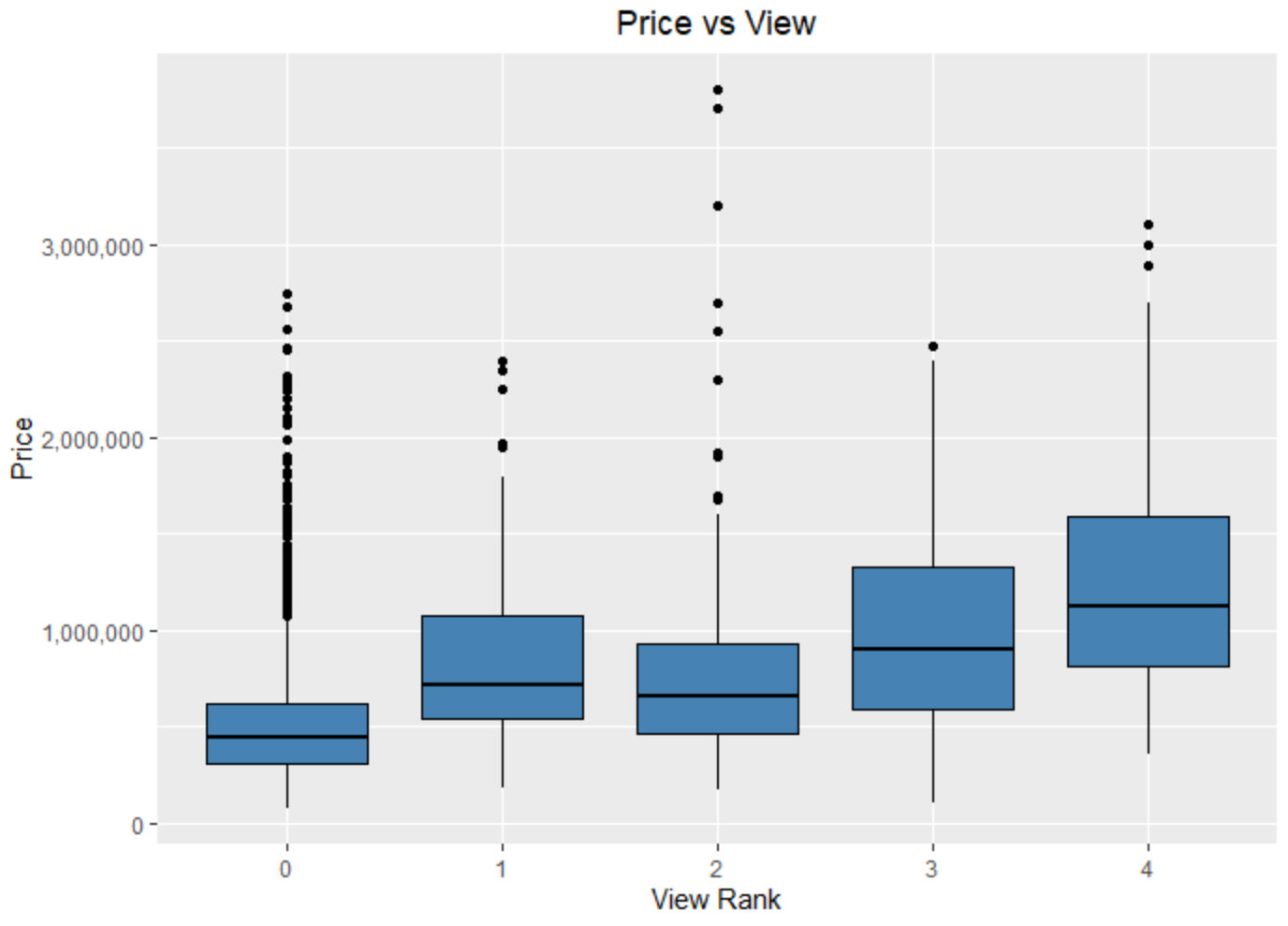
The next graph I wanted to show here is a density plot with a building year on the x-axis grouped by the waterfront. We can see with the vertical mean line that the average built year of waterfront houses are older than houses that are not waterfront. We could also see it from the shaded area where the area of not waterfront house is skewed to the left(newer) and the area of the waterfront house is skewed to the right(older).



The graph shown below is the year built distribution of all the houses sold. We can tell from the graph that the data in our dataset is mostly newer houses. So when we are doing regression analysis in the next project we can use this to conclude that the price won’t be affected by the historical value so that we don’t have to add that variable to our model.



The last two graphs I wanted to show here are the box plots for categorical variables “view” and “conditions” with respect to price. As we can see in the graph, as the rank gets better, the mean value and the range of price will go up. So they have some positive linear relationship with price, which is relatively not too strong but detectable.



**Detailed describe**

From the previous plots, we could discover that all the independent variables have a positive linear relationship with respect to price. We also discovered that the dummy variable does have an effect on housing prices as well as all the categorical variables. The dataset is a perfect shape for linear regression

By doing this project, I have practiced and learned three main abilities: Collecting data, cleaning the data, and exploring the data using ggplot2. It taught me a great deal about using different functions on ggplot2 for desired plots as well as understanding the plot. Even though the class has taught me enough knowledge to do this project, one additional important thing I learned from this project is about using Google. I have learned that no matter how good you learned about coding, you will always face some unexpected problems while doing tasks. And the best way to solve it is to use Google and search for StackOverflow or GitHub answers. It might take some time but as long as you understand what the code does, it will help you through this.