Project Part I

Mengchen Wang (Veronique)

Data And Data Description

The dataset, “House price prediction” is from Kaggle. It is uploaded by a user named Shree. The data contains a large set of property sales records in 2014 in the state of Washington, which originally used to analyze and predict the future property prices in the real estate market.

The data is shown in a csv file with 18 columns and 4600 rows. The rows represent the number of properties this data set included, which means that it has 4600 data totally, and each data contains the information of a certain property. The columns represent the features/ information of these properties, which means that each property has 18 features.

The 18 columns are date, price, bedrooms, bethrooms, sqft_living, sqft_lot, floors, waterfront, view, condition, sqft_above, sqft_basement, yr_built, yr_renovated, street, city, statezip, and country. Date represents the date the data is collected; price represents in sale price of this property in dollars; bedrooms and bathrooms represent the number of bedrooms and bathrooms in this property; sqft_living and sqft_lot are the size of the living area and the total area of this property in square feet; floors represents the number of floors the property has; waterfront is if the property is waterfront or not; view represents the number of views; condition is scored from 1 to 5 as rating; sqft_above and sqft_basement describe the area above and below the ground in square feet; yr_built and yr_renovated describe the time this property built and renovated; and street, city, statezip, and country tell the location of this property.

The data is a sample. It is collected from 5/2/14 to 7/10/14 among different selling properties in different cities in the state of Washington, USA. The potential issue of this data is that the owner of this data does not mention how the sample is selected on the website. As a result, we do not know if this is a random sample or if this data set is a good representation of the population. The shortcoming is that the conclusion we get in this data set cannot reflect the features in the population; we can only use the results to predit inside this data set. Also, all the data were collected within two and half months, which cannot represent the whole year of 2014. The conclusion we get can only reflect in this time range.

Since there are too many features in this data set, I choose to highlight the relationships between price and the sqft_living, number of bedrooms, floors, condition in this project.

Subset Data

```
library(ggplot2)
library(dplyr)
library(ggpmisc)
## Reference 3: Not show the message
```

```
sales.sub <- dplyr::select(sales, price,sqft_living, bedrooms,
                           floors, condition)
## Reference 1: Kaggle
```

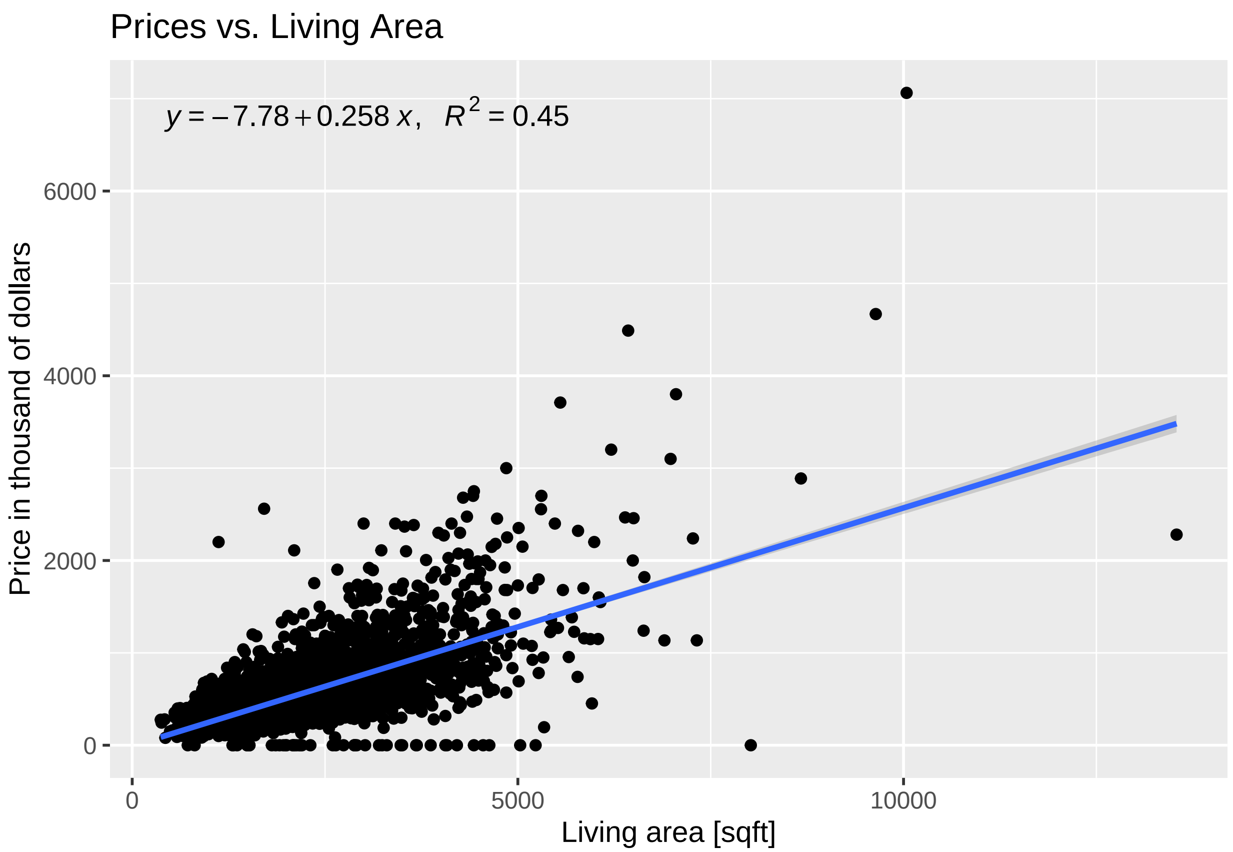
I use select in the dplyr package to subset the columns I need, including price, sqft_living, number of bedrooms, floors, condition in this project.

Data Summary And Plots

Price vs. sqft_living

To check the relationship between price and sqft_living, I use ggplot to make a scatterplot. Since there are two extreme outliers in this data set, I use filter to remove them. Also, to make the data easier to read, I make price unit from dollars to thousand of dollars.

```
data1 <-filter(sales.sub, price<1e7)
data1$price <- data1$price/1000
g1<-ggplot(data1, aes(x=sqft_living,y=price))+geom_point()+geom_smooth(
  method =lm)+stat_poly_eq(formula = y ~ x, aes(label = paste(..eq.label..,
    ..rr.label.., sep = "*plain(\",\")~~~")), parse = TRUE)+
  labs(title="Prices vs. Living Area",
       x="Living area [sqft]", y="Price in thousand of dollars")
g1
```



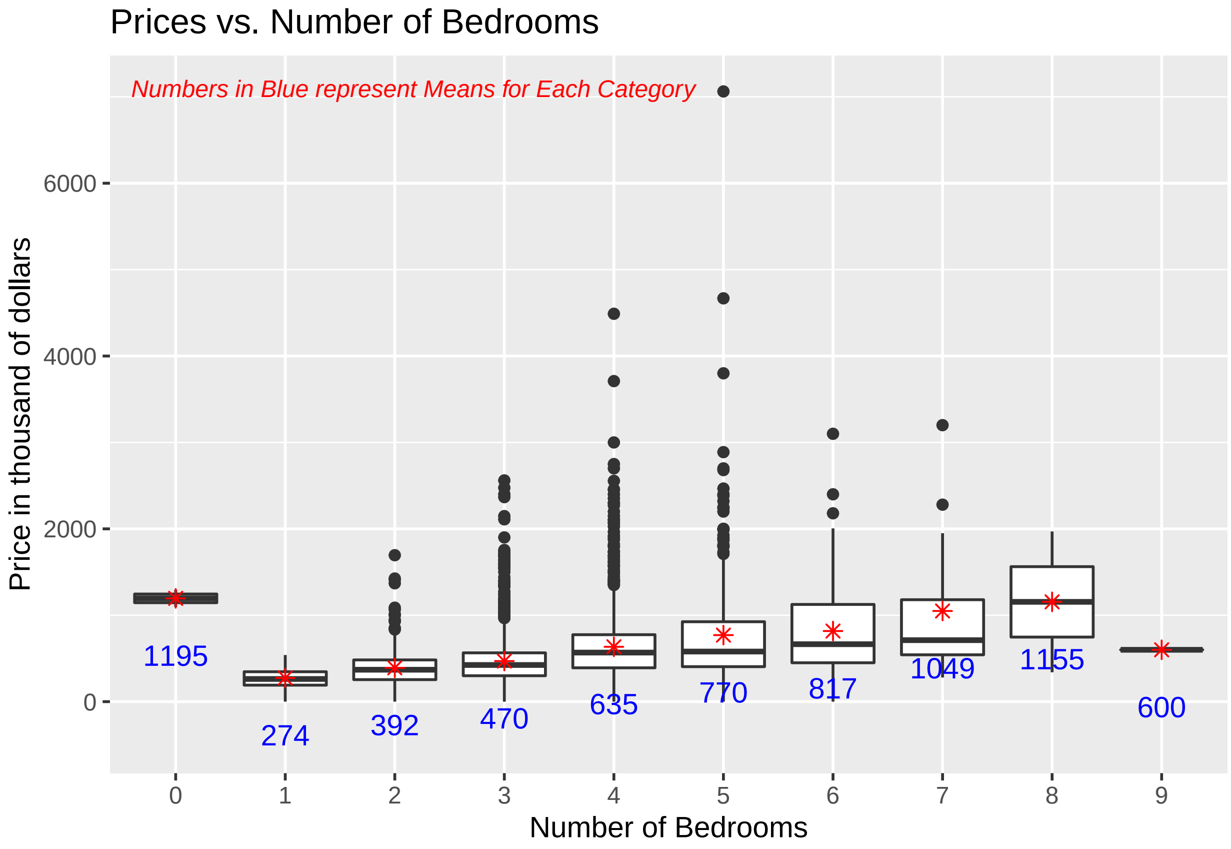
```
## Reference 2
```

The plot shows a positive relationship between price and sqft_living, and most data cluster at the left corner. I add a linear regression line to the plot as well. We can see that this linear regression line does not fit the points very well since the R^2 value is only 0.45.

Price vs. Bedrooms

I use boxplot to see the how the prices variance among different number of bedrooms and check means of price for each number of bedrooms as numerical representations.

```r
data1$bedrooms <- factor(data1$bedrooms)
g3<-ggplot(data1, aes(x=bedrooms,y=price))+geom_boxplot() +
  stat_summary(fun.y="mean", geom="point", shape=8, size=2,color ="red") +
  labs(title="Prices vs. Number of Bedrooms",
       x="Number of Bedrooms", y="Price in thousand of dollars") +
    stat_summary(fun.y=mean, geom="text", aes(label=round(..y..)),
      vjust=3.2, position=position_dodge(0.9), color="blue")+
  scale_y_continuous(limits = c(-450,7100)) + annotate(geom="text",
  x=3.17, y=7100, label="Numbers in Blue represent Means for Each Category",
  size =3.13, fontface = 'italic', color="red")
g3
```



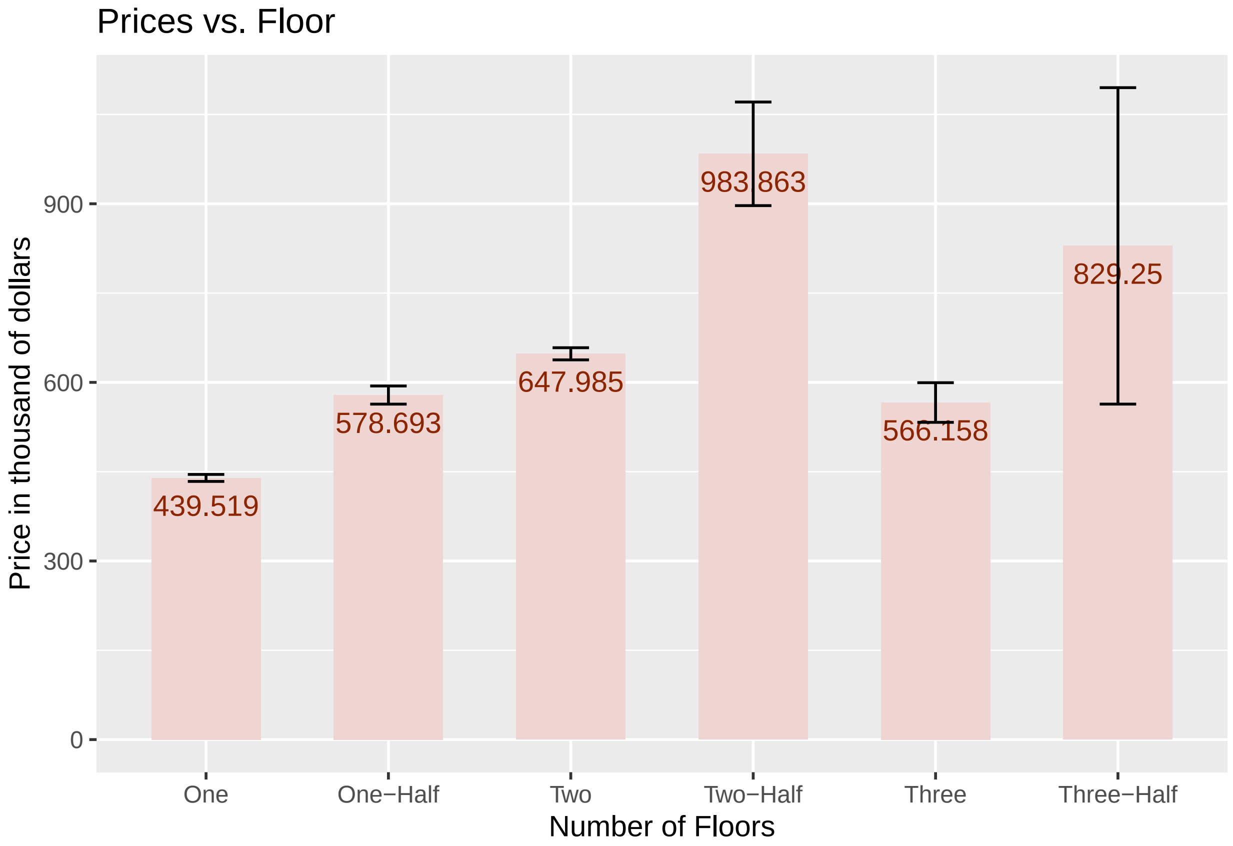
```
## Reference 4
```

We can tell from the plot that, except for zero and nine bedrooms (which is the rare case), there more bedrooms there are, the higher the price the property sales. We can tell it from the mean as well. Also, except for zero and nine bedrooms, regardless of outliers, the variation of prices increases (the box gets larger) as the number of bedrooms increases. We can also tell that there are more outliers when the number of bedrooms is between 3 and 5.

Price vs. Floors

I use stat_summary() to draw bar graphs to see the how the mean of prices differ among different number of floors both as plot and as numerical representations.

```
data1$floors<- factor(data1$floors, labels=c("One","One-Half","Two",
                    "Two-Half", "Three","Three-Half"))
g4<-ggplot(data1, aes(x=floors,y=price)) + stat_summary(fun.y=mean,
            geom="bar",fill = "mistyrose2", width =0.6) +
  labs(title="Prices vs. Floor", x="Number of Floors",
  y="Price in thousand of dollars") + stat_summary(fun.y=mean, geom="text",
    aes(label=round(..y..,3)), vjust=1.8, color ="orangered4",
    position=position_dodge(0.9))+
  stat_summary(fun.data=mean_se, geom="errorbar", width=0.2)
g4
```

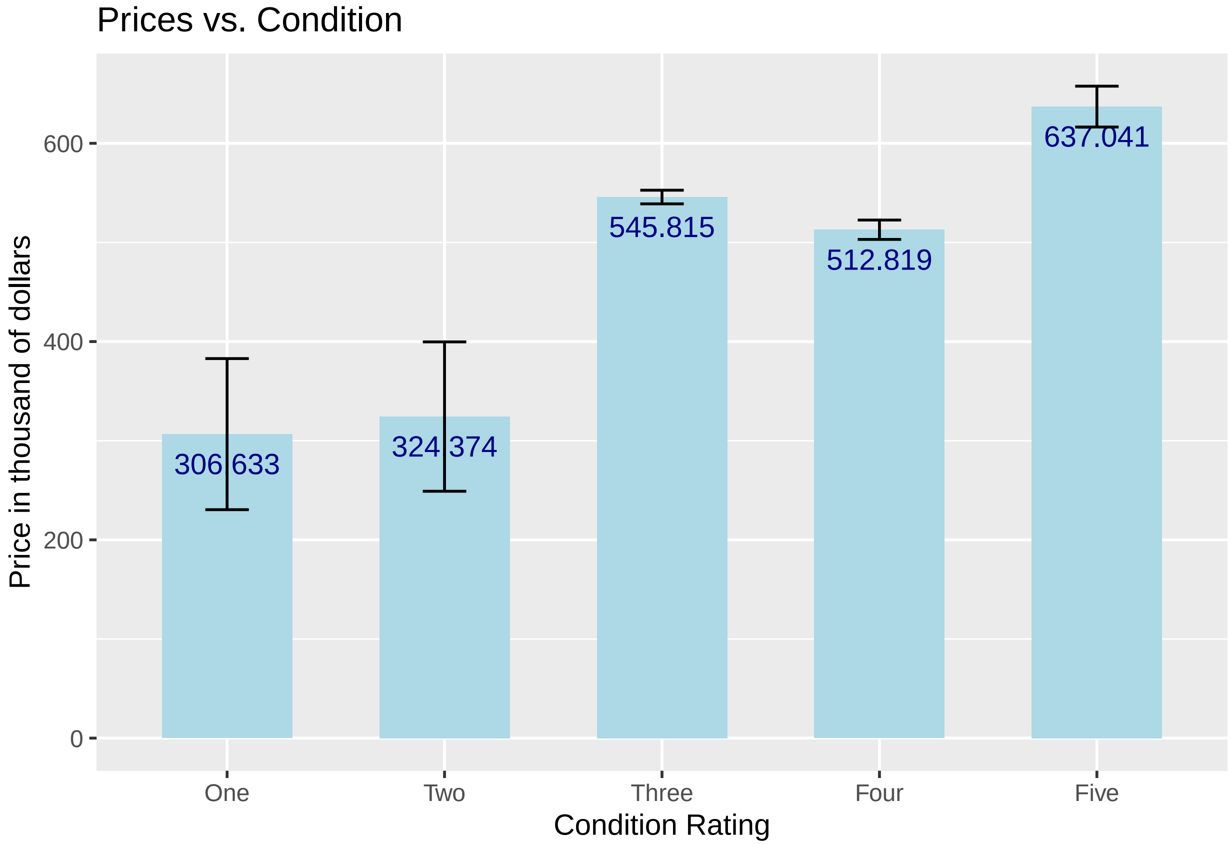


There is a positive relationship before three floors property. However, the means of prices of three and three-half floors drop. The mean price of Two-half floors reach the highest. By the one-standard error errorbar, we can see that the more the floor is, the wider the errorbar is, which means the variance of the price increases.

Price vs. Condition

I use stat_summary to add a bar plot to see how the mean of prices differ among different condition ratings both as plot and as numerical representations.

```
data1$condition <- factor(data1$condition, labels=c("One","Two",
                       "Three","Four","Five"))
g5<-ggplot(data1, aes(x=condition,y=price)) + stat_summary(fun.y=mean,
            geom="bar",fill = "lightblue", width =0.6) +
  labs(title="Prices vs. Condition",
            x="Condition Rating", y="Price in thousand of dollars") +
    stat_summary(fun.y=mean, geom="text", aes(label=round(..y..,3)),
    vjust=1.9, position=position_dodge(0.9), color = "navyblue") +
  stat_summary(fun.data=mean_se, geom="errorbar", width=0.2)
g5
```



We can tell from the bar plot that, in general, higher the condition rating is, the higher the mean of the price is. However, this relation is not absolute since the mean of condition rating three is slightly higher than that of condition rating four. We can tell from the one standard error errorbar, the variance of prices is larger when the condition rating is low.

Conclusion based on Data Summary

Based on the data summary and the plot, we can see that there are some clear relationships between prices and some factors. However, there are some factors that I cannot give a clear conclusion and future analysis is needed.

First of all, there is a positive relationship between price and sqft_living and bedrooms, based on the scatterplot and the boxplot. The positive linear regression line in the scatterplot (price vs. living area) and the equation y = -7.78 + 0.258x indicate the positive relationship as the coefficient before x is positive. However, this equation is only useful within the x range of the given data, and the y- intercept (-7.78) does not have an actual interpretation since even if the living area is zero, the price of the propperty will not be negative. Additionally, the R^2 (0.45) shows that the linear regression line does not fit the points very well; as a result, future analysis can use other factors, for example, conditions, to color the point and check if there is a clearer linear regression relationship between prices and living area with based on certain factors. Secondly, the boxplot between prices and number of bedrooms shows that besides the rare cases (zero and nine) and regardless of outliers, the more bedrooms the properties have, the higher the mean price is and the more variance on the prices. The range of prices is the largest when the number of bedrooms is 5.

The relationship between prices and floors and conditions are not very clear. For price versus number of floors, the bar plot shows that there is a positive trend from one to two-half floors, but not three and three-half. Additionally, we can tell that the prices are less uniform (varies more) when the number of floors is two and half and more with the errorbar. This might because most families do not need three and three-half house since these might be too large so the willingness of paying more for these property varies more. For mean prices versus condition ratings, there is a positive relationship, but it fluctuates when the condition is four. This might because that when the condition rating reaches three, most people think it is good enough, so that the rating does not affect the price that much. Also, the prices varies more when the condition rating is low. Future analysis may focus on the using other plot and separate bars with different factors.

Lastly, due to the shortcoming of the data (do not know how the sample is selected), the conclusions cannot reflect to the population. We can only use these conclusions within this data file and predict the prices of properties within the ranges of all independent factors.

References

1. https://www.kaggle.com/shree1992/housedata
2. https://stackoverflow.com/questions/7549694
3. https://stackoverflow.com/questions/44520686
4. https://stackoverflow.com/questions/30673470/ggplot-format-italic-annotation
5. https://stackoverflow.com/questions/6046263
6. Other codes are all from notes.