**HOUSING PRICE ANALYSIS AND PREDICTION**

**ALY 6040 DATA MINING APPLICATIONS**

**WEEK 2 ASSIGNMENT - PATTERNS AND EDA**

**NORTHEASTERN UNIVERSITY**

**SPRING 2019**

**Introduction**

The dataset that we found - ‘House Sales in King County USA’ comes from Kaggle. This data contains the house sale prices along with some different features. There are 21 house features and 21613 observations for King County in Seattle from May 2014 to May 2015. The following are our objectives:

* Find the trends and the patterns for different variables through plots, graphs, and map
* Identify how sales went over the years, months, and days in the past
* Predict the sale price of a house if the required parameters are given

**Dataset Description**

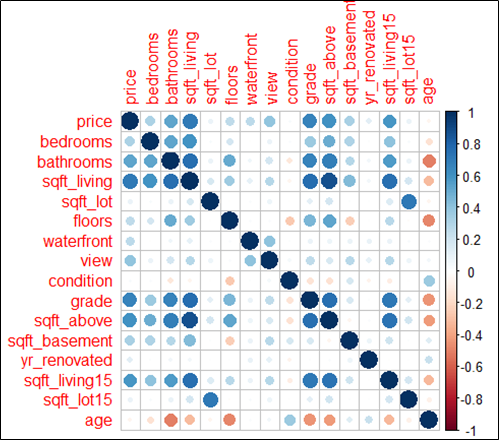
|  |  |
| --- | --- |
| id | A notation for the house |
| date | Date when the house was sold (yyyy-mm-dd) |
| price | Price for which the house was sold |
| bedrooms | Number of bedrooms in the house |
| bathrooms | Number of bathrooms in the house |
| sqft\_living | Square footage of the house |
| sqft\_lot | Square footage of the lot size |
| floors | Total floors (levels) in the house |
| waterfront | If the house has a view to a waterfront |
| view | If the house has been viewed before buying |
| condition | The overall condition of the house (higher the better) |
| grade | Overall grade given to the housing unit, based on King County grading system |
| sqft\_above | Square footage of house excluding the basement |
| sqft\_basement | Square footage of the basement |
| yr\_built | Year when the house was built |
| yr\_renovated | Year when house was renovated |
| zipcode | ZIP code for the house |
| lat | Latitude coordinate |
| long | Longitude coordinate |
| sqft\_living15 | Total area of the house in 2015 (implies some renovations) |
| sqft\_lot15 | Lot size area in 2015 (implies some renovations) |

**Data Cleaning**

The ‘date’ column in the dataset was not in a proper format. It is converted to the appropriate date format. In addition, three new columns from the ‘date’ column have been added to the dataset. The columns contain the year, month, and the days of the month when the apartment was sold. Moreover, the age of the houses has been added as another column.

**Analysis**

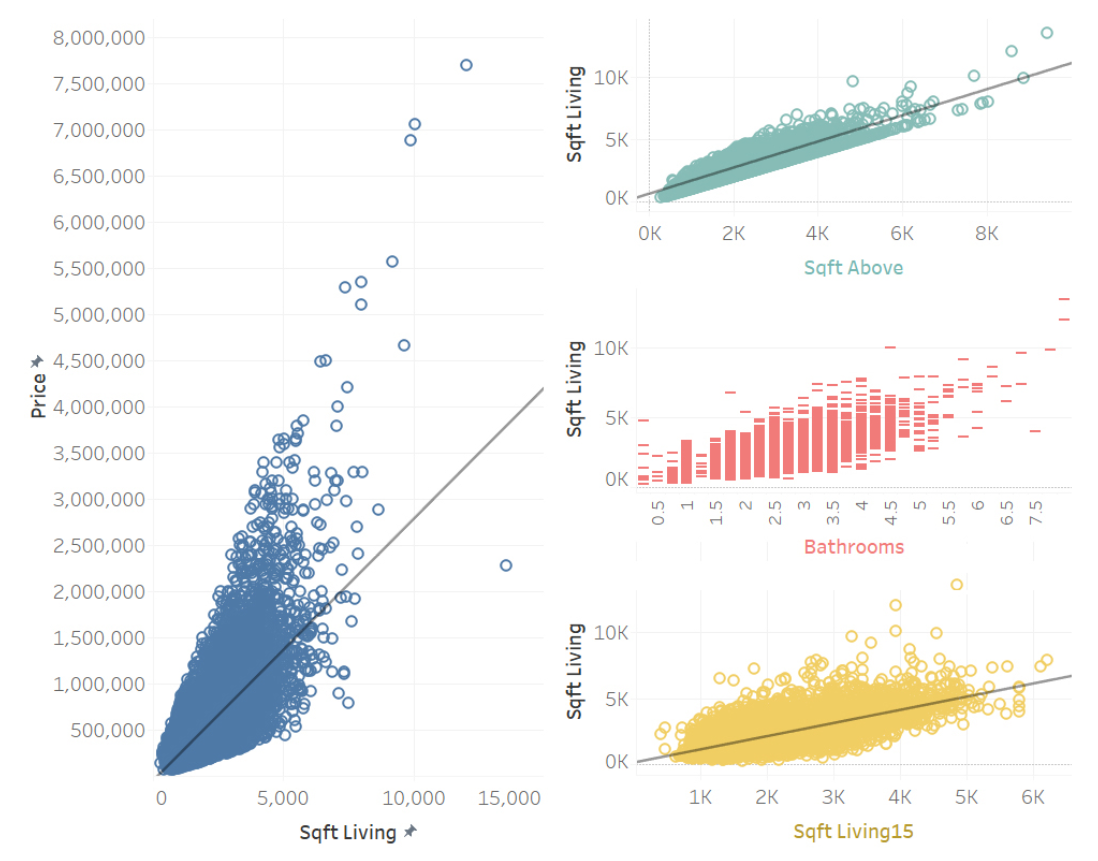
**General Correlation**

****

From the correlation plot, it can be observed that the sale price of houses has a strong positive correlation with the original area of houses, and the grade of houses. The prices are somewhat correlated to the area of houses excluding the basement, and the area of houses in 2015. The total area of the houses is strongly correlated to the area of houses excluding the basement. Moreover, the total area of houses has a positive correlation with the number of bedrooms, and the number of bathrooms. Further, the price of houses does not show a correlation with the area of the lot, the condition of houses, or the age of houses.

**Square Area**

From the previous part, we observe that the living area (sqft) has the highest correlation coefficient with the house prices, therefore, we decide to examine more about this feature. There are three variables which have the highest correlation coefficients with the square feet of living area: the above area of the house (sqft), the living area reported after 2015 (sqft), and the number of bathrooms.

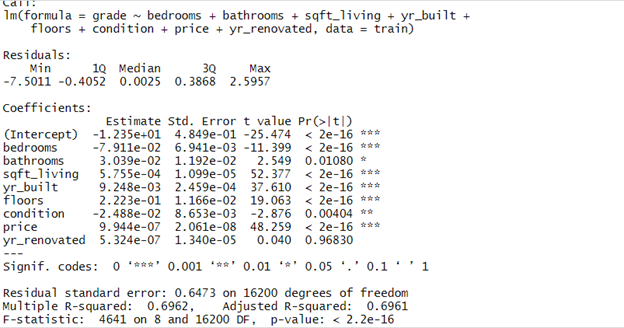
****

The dashboard illustrates the positive correlations between the living area and the other three variables. These results are easy to understand because there are strong relationships between them in real life. In fact, the living area contains the above area of the house as well as the living area reported after 2015 is also the same feature with the living area before that year. Moreover, houses having a large number of bathrooms can only exist when they have sufficient living spaces.

As we can see from the chart, there is a significant positive relationship between the house price and the living area. Because the above area of the house, the living area reported after 2015, and the number of bathrooms are correlated with the living area, we can conclude that there exists the correlations between those features and the house price.

**Grade**

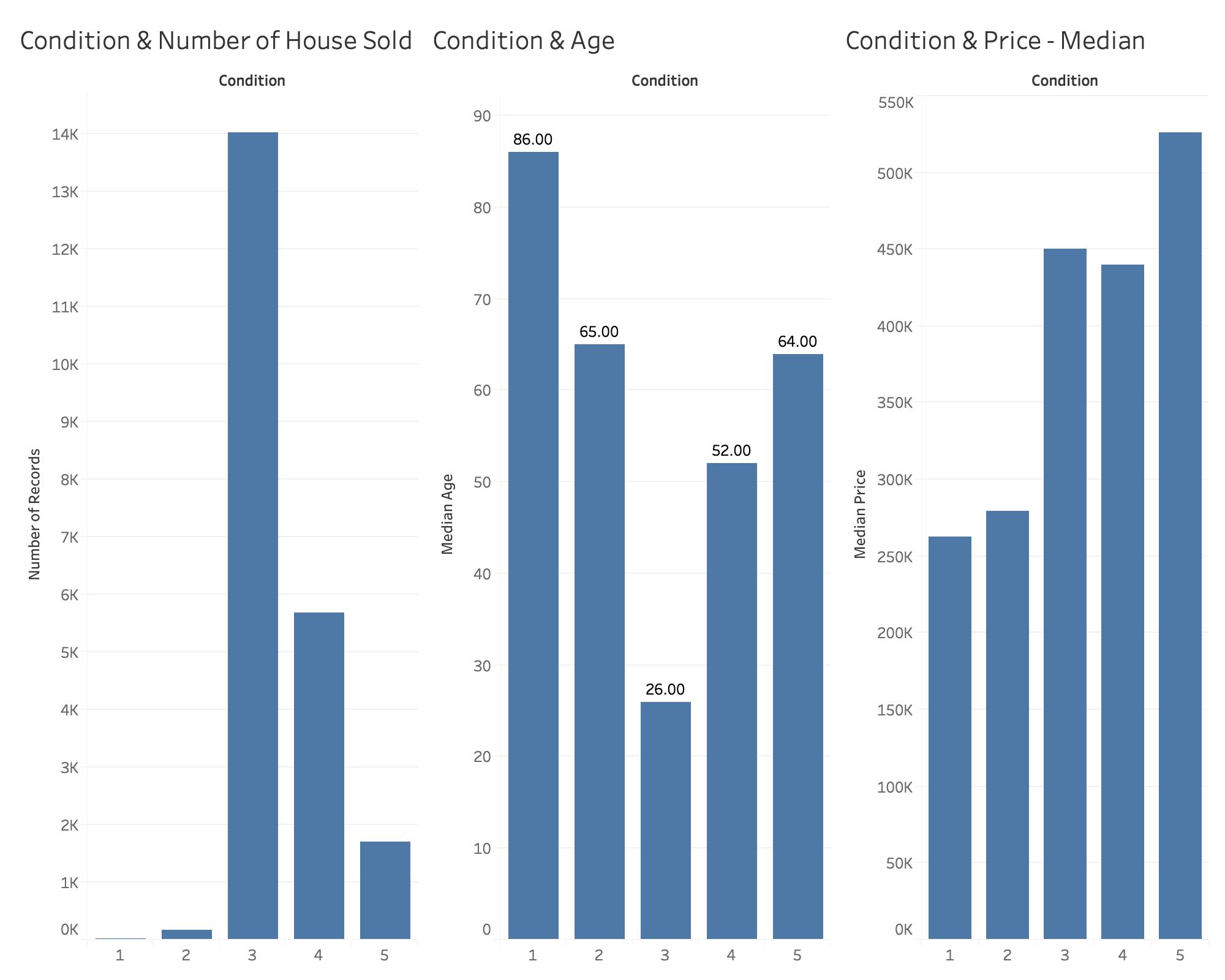
Next, we used linear regression analysis to see what factors determine the home rating. Linear regression analysis is a technique for analyzing the correlation between one independent variable, y, and several independent variables, x. For this process, eight independent variables were used, which are bedrooms, bathrooms, sqft\_living, yr\_built, floors, condition, price, yr\_renovated. I used the set.seed (123) function and divided the ratio of training data to test data by 0.75 to 0.25, using the “lm” function to perform a linear regression. Here is the result.

****

As you can see in the chart, all the variables came out as meaningful variables because their r value is lower than the confidence level of 0.05 and close to zero except for one variable, yr\_nenovated. In this chart, as the price of a house rises, the rating of a house also increases. We can see that it shares elements of raising the price of the house and increasing the rating. The R-squared of the model is also quite large at 0.6961, which means that the model explains the data by 69 percent.

We also examined the predictive performance of this function using the function “predict”. We tested the training data with the test data and the r-value was quite high at 0.71, but since mse is also a little high at 0.4, we cannot tell this model has accurate predictions

**Condition**

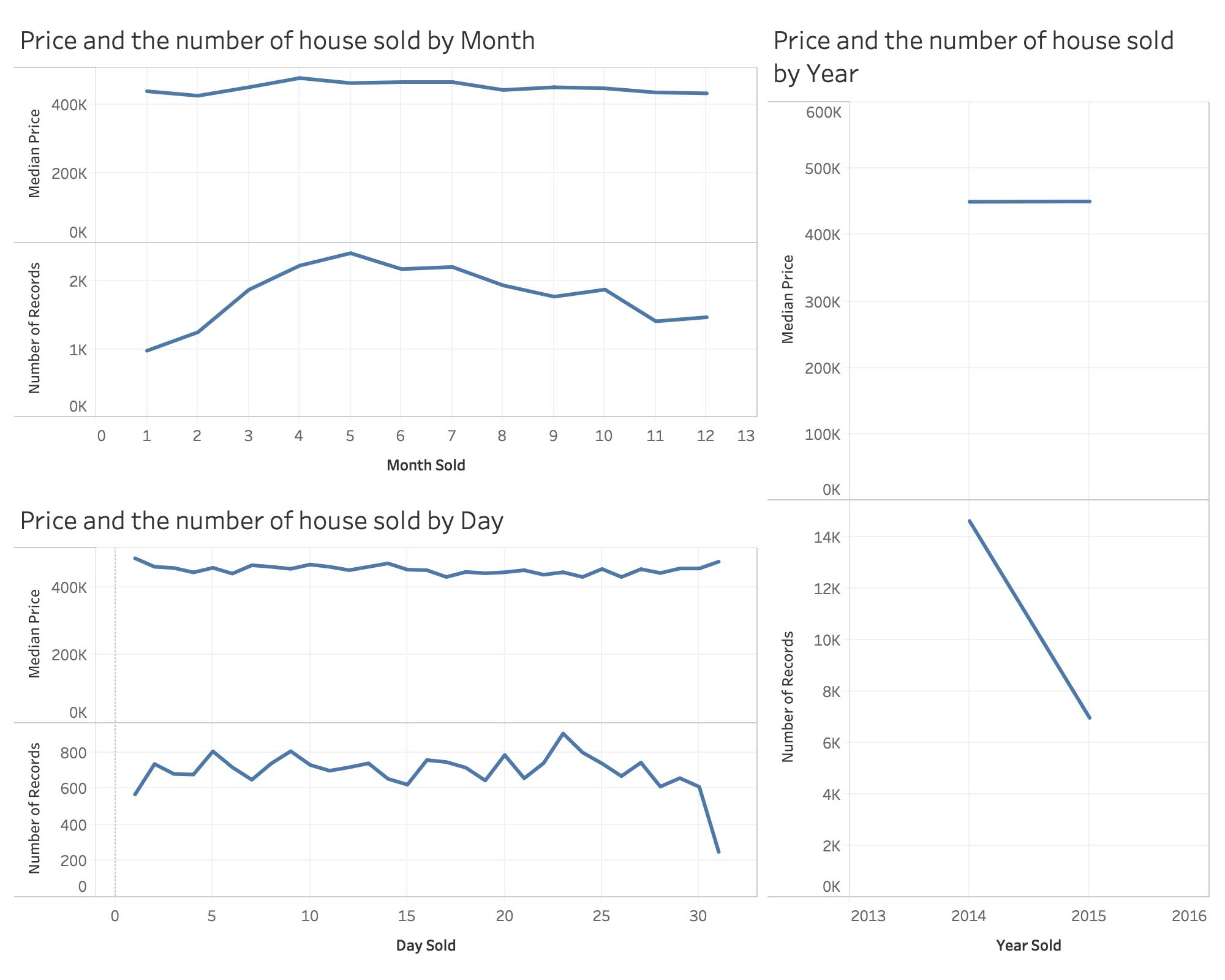
****

First of all, it can be clearly seen that most of the houses sold had condition 3. The total number of houses with condition 3 was sold twice as much as houses with other conditions. For the years 2014 and 2015, only 202 houses were sold which were in the worst condition. However, the number of sold houses in condition 4 and 5 are lower than the number of houses sold in condition 3. Thus, the condition of the house might not be a strong factor which impacts the house buying decision of customers.

Secondly, to avoid the outlier impact, these graphs were created to display the relationship between the condition of the house and the median age of the house as well as the median price of the house. It is obvious that the houses in condition number 3 are houses with the youngest age. The houses at condition 1 are the oldest house, thus it might be the reason they are hard to sell. However, the houses with condition 2 and condition 5 have quite similar median age. The house at condition 5 might be better maintained than the other one.

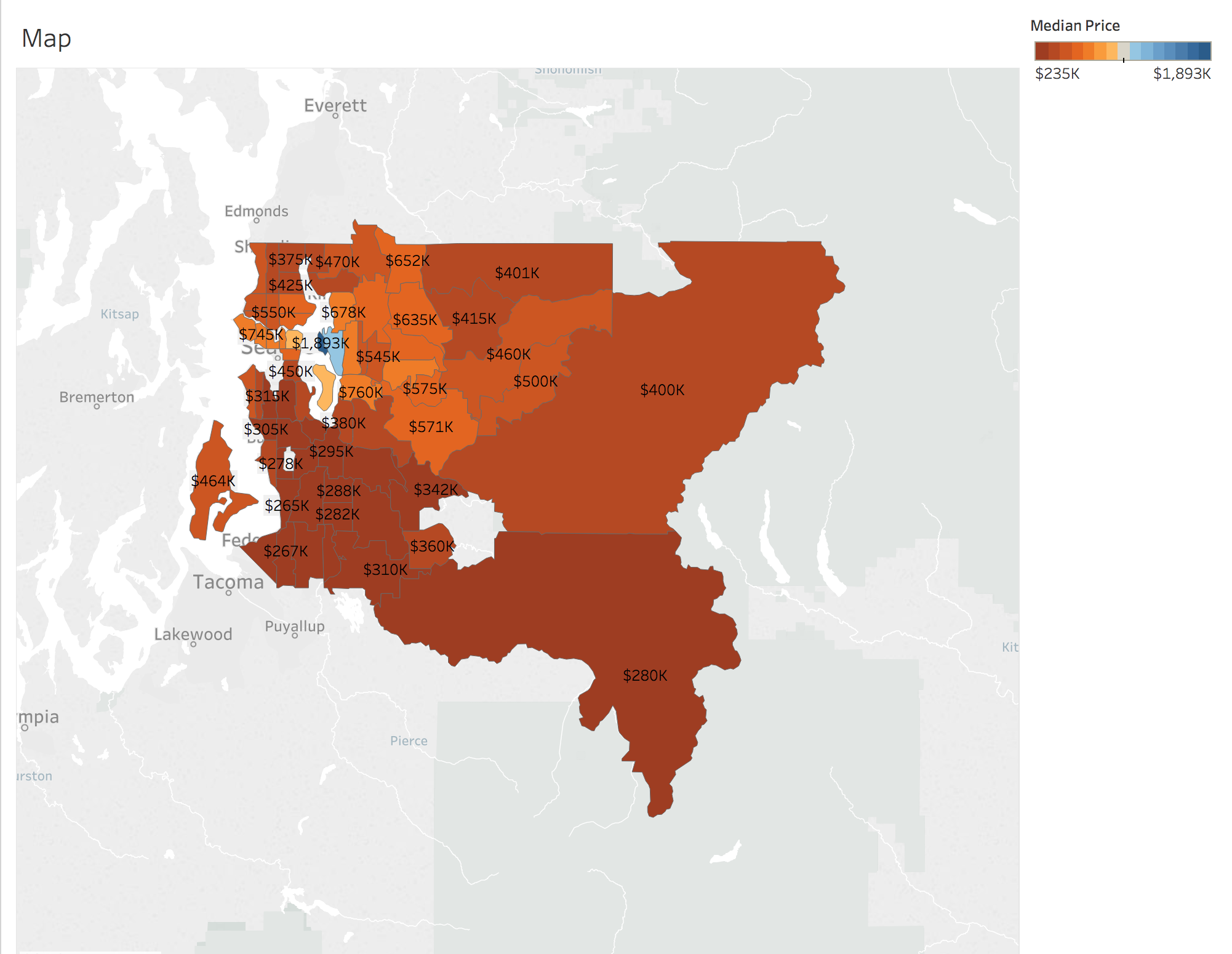
Moreover, it is obvious that the house with the highest condition has the highest price and house with the lowest condition has the lowest price of selling. However, the price of the house at condition 3 is slightly higher than the price of the house at condition 4. Thus, it could be understood that condition of the house is not really a strong factor which determines the price of the house and the number of the house sold.

**Month and Year**

****

As you can see that the lines show the median price of houses sold by the day, month, and year. The cost of a house in 2014 is same as the price of a house in 2015, about $450,000 per house even though the number of houses sold by 2014 is more than twice as much as the number of houses sold by 2015, that is 14,500 and 7,000 houses respectively. Besides, the price of a house is highest in April while in May, most of the houses were sold. The houses were sold at the highest price in the first and the end day of a month compared with other days. But the number of houses sold the most is on the 23rd day of a month. One interesting thing is that people seldom bought a house at the end of a month, especially the last day.

**Location**

****

The heat map above is for King County’s median house sale prices located in Seattle. We can tell that the expensive houses are in the northwestern area, to which the median price range are mostly between $500 thousand to $1,893 thousand. The median price range for the less expensive houses is approximately between $265 thousand to $400 in the south, north, and northeastern area.

**Conclusion**

We consider that the critical features are between square area, grade, condition, time, location, and price. After making the comprehensive analysis, we can tell that there are strong correlations between the square area, condition, and price. The more space and excellent condition a house has, the more expensive the house is. However, the number of houses sold with a decent condition is 6 times higher than the number of houses sold with excellent condition. People seemingly prefer to buy the kind of house that is with a decent condition, which may be affordable for most people. In addition, the number of sales in 2014 were doubled compared with sales in 2015. During the cycle, people would like to buy houses during the summer, particularly in May. January is the month that has the least sales. We think that is probably due to the severe weather in winter and other unknown factors. And because of that, the market has to lower the price, then the lowest price happens in February. After that, the sales are rising while the price is becoming steady. We will show further analysis and predictions for the sale price of a house in the upcoming weeks.

**References**

House Sales in King County, USA. (2019). Retrieved from https://www.kaggle.com/harlfoxem/housesalesprediction