Data Science and Analytics

Semester Project - House Price Prediction

Team C

Jane(Yu-Chen) Su, Leon Tan, Yuxiang Zhao,

Kushagra Patel, Anisha Vimalan, Yangyang Yan

December 5, 2019

A. Framing the Problem

a. Problem Recognition

Back in 1931, when James Truslow Adams first introduced American Dream, he defined it as “that dream of a land in which life should be better and richer and fuller for everyone, with opportunity for each according to ability or achievement”. Moving along, that dream has gradually changed to a picture of a family and their dog standing in front of their own house with a white picket fence and a private yard in the last century. People start having conversations about their dream houses. But, since different people have different pictures of their dream houses, is there any general pattern that we can trace? What usually makes people more willing to pay for their dream house? What makes a house have a higher price than others’?

Having these questions, we want to evaluate different factors that will affect houses sales prices and come up with a model that can analyze and forecast the house sale prices.

b. Review of Previous Findings

In daily life, we can easily use our past life experience to speculate that the housing cost of large cities is higher than the housing cost of small towns and it is normal to us. In China, house prices in Beijing, Shanghai, Guangzhou, and Shenzhen are very high, the same thing also happens to Seoul in South Korea, New York in the United States, and Tokyo in Japan. We could say an inch of land, an ounce of gold. These cities focus on the national best-quality political resources, business resources, education and medical resources, and human resources. People will naturally gather in places with a high concentration of resources, and cities with a high concentration of resources will also bid for price increases. We have conjecture. House prices are affected by the concentration of resources. In areas where there are no political resources, no financial resources, no medical resources, and no human intellectual resources, such as backward rural areas, house

prices will only become lower; in areas with high resource concentration and the net inflow of population, the housing prices will rise.

When we try to find and review the stories that have happened in the past, we find that most of the variables that affect house prices are the types of information that typical home buyers want to understand. When was it built? How big is the room? How many square feet of living space is in this house? Is there a basement? How many bathrooms are there? These questions all point to the direction of the house hardware conditions. Such a discovery is actually no surprise. The house as a commodity, its physical attributes (housing area, infrastructure, convenient transportation, greening degree, old or new housing, etc.) is what each purchaser concerns for any consumption purpose, and they will be carefully examined. We believe that there is a particular inevitable connection between the hardware facilities of the house and the house price.

In the end, our team narrowed the investigation by studying previous models and analysis, ignoring external economic and political factors. We only treat houses as commodities. Then as a commodity, its building materials and the state of its final presentation are regarded as the overall quality of the house, which is also the value of the house, which will be our variable one. First Floor Square Feet and the number of bedrooms are related to the physical properties of the house we studied as two other variables.

B. Solving the Problem

a. Modeling and Variable Selection

House price depends on various factors such as location, land size, economy, supply and demand for houses, interest rates, new amenities, parking, basement, etc. From previous studies, we already know that the aforementioned variables are significant and have a positive linear relationship with the price. Our job was to look for other variables that might play an important role in determining the house price. After extensive research and going through more than 300 variables, we selected three variables and assumed that people will consider them while purchasing a house. (See Figure 1)

1) Overall Quality - This variable is an ordinal data with a discrete attribute. It has ten levels, including very excellent, excellent, very good, good, above average, average, below average, fair, poor, and very poor. House prices might depend on what type of material is used by the developer during the construction of the house along with the interior finish. If the premium quality material is used, then the sale price should be high and vice versa. An old house with high-end finishing and excellent overall quality may command the same or high price as compared to a newly constructed house with poor overall quality. High quality means maintenance costs will be significantly lower. House buyers will consider the overall material specifications and decoration of the house. (See Figure 2-3)

2) First-floor square feet - This variable is numerical attribute and ratio-scaled data, which is between 334 and 4692. Usually, the first floor in a house has a living room, kitchen, dining room, 1 or 2 bedrooms, laundry room, etc. People spent the majority of their time on the first floor as compared to the other floors in the house. Most of the house buyers are looking for a house that has at least one guest bedroom on the first floor as it will be convenient for their guests. So bigger

first floor area means a bigger kitchen area, more bedrooms, and plenty of additional space for gym equipment, storage, etc. Buyers mostly form their opinion about a house based on the first-floor area, so we think that the floor area of the first floor affects the price of the house more than the second floor and the basement area. (See Figure 4-6)

3) Number of bedrooms - This variable is numeric attribute and ratio-scaled data, which is between 0 and 8. The number of bedrooms is an important criterion for most of the potential house buyers. Typically, Buyers who are married or have a family will look for a house with more than 2 bedrooms while bachelors will look for a 1-2 bedroom house. More bedrooms imply bigger lot area, more occupants, more house levels, etc. So, we think that the number of bedrooms will significantly affect the sale price of a house. (See Figure 7-8)

Variable	**Notes**
Overall Quality	Ratings of the overall material and finish of the house
First Floor Square Feet	The area of the first floor in square feet
Number of Bedroom	The number of bedrooms above the ground. Basement bedrooms are not included

b. Data Collection

Firstly, our team chose a database provided by the Kaggle website. Because Kaggle provides amount of real-world data, as well as the actual problems encountered by companies. Besides, there is the community on Kaggle, that is, we can learn how to think about problems and solve problems from experts in various fields on Kaggle. The notebooks in Kaggle provided by

these experts become the nutrients of our problem-solving skills, further let us think about how to decide variables and models to predict the housing price more precisely.

Next, we selected a suitable database based on our current learning ability as our project. This Ames Housing database was compiled by Dean De Cock, the professor of statistics at Truman State University. In fact, the initial original database collected by the Ames City Assessor's Office was used for tax assessment purposes, containing 113 variables which were a combination of nominal, ordinal, continuous, and discrete data and describing 3, 970 property sales. However, it lends itself to the prediction of house selling prices.

The professor removed the variables which are used for the previous calculation or required domain knowledge, and also deleted the variables that were related to adjustment factors used in Ames City's modeling system. In addition, the professor studied about Boston's housing data during his graduate studies, so we believe that he has related experience and ability in the house industry, and trust the Ames Housing database he compiled are factual and credible.

c. Data Analysis

After the data collection, we filtered the dataset by only selecting the three required variables. First, we used descriptive statistics to summarize the data in a meaningful way so that we can identify some patterns from the data. One general pattern that can be observed is that the house price increases with the increase in first-floor area, number of bedrooms and the overall quality. But we cannot draw the conclusions for our hypotheses just based on the descriptive statistics as it simply describes our data. Since we are trying to examine the relationship between two or more variables of interest, regression analysis would be an ideal method as it mathematically sorts out which independent variables may have a significant impact on the dependent variable. We developed a linear regression model where the value of the house price

variable depends on the value of overall quality, first-floor area and the number of bedrooms variables. After running the model, we found that all the variables are significant and have a positive linear relationship with the house price. The number of bedrooms has a positive linear relationship if the value is equal or less than 4 and has a negative relationship if the value is greater than 4.

SalePrice = -133,055 + 37,128 * OverallQuality + 59.53 * FirstFloor + 6,387 * NumberOfBedrooms

Based on our descriptive statistics results, we decided to perform a Shapiro-Wilk normality test to determine if the dataset is well-modeled by a normal distribution or not. Also, a Non-constant variance score test was executed to assess the equality of variances for the sale price variable. The model suggests that the sale price increases if the area of the first floor, number of bedrooms and overall quality increases.

d. Quantitative Analysis and Creativity

i. Quantitative analysis

Descriptive statistics and linear regression model are used for data analysis. The standard deviation of house price is $ 79,442.5 which is very high and indicates more heterogeneous or dissimilar spread of the data. The linear regression model was used to determine the relationship between the four variables. After running the model, the following observations were made.

Each level of overall quality increase leads to a predicted $37,128 increase in sale price. One square foot increase in the first-floor area leads to a predicted $59.53 increase in sales price. The sale price of a house increases by $6387 on average by an increase in one bedroom. R-squared for the model is 0.6978 which means 69.78% of the variance in the sale price variable is explained by the independent variables used in the regression. The p-value is 2.2e-16 which is less than the

significance level (0.05 or 0.01). Hence, we can reject the null hypothesis and the result is statistically significant. (See Figure 9)

When establishing a linear regression, we must also confirm whether the residual meets the normality, independence, and homogeneity of variance assumptions. The test for residual (normality, independence, homogeneity of variance) shows that normality and homogeneity of variance failed while independence held. Since normally distributed residual is one of the assumptions for regression analysis, we cannot rely on the results as the confidence interval is not accurate. But our analysis is still reliable because the sample size is greater than 15 and we didn't use confidence interval in our analysis. Therefore, normality has no impact on our results. Similarly, the homogeneity assumption can be ignored because the estimators are still unbiased. (See Figure 10)

Additionally, we would like to see if there is multicollinarity. Through the VIF consequence, we can conclude that there is no multicollinearity problem because all VIF are smaller than 10. The correlation between Saleprice and FirstFloor is negligibly positive because the correlation coefficient is 0.606. The correlation between Saleprice and Bedroom is negligibly positive because the correlation coefficient is 0.169. The correlation between Saleprice and OverallQuality is negligibly positive as the correlation coefficient is 0.791. (See Figure 11)

ii. Creativity

We were being creative by selecting the overall quality as one of the independent variables. People, typically, forget to take into account the quality of material used in the house during construction and decoration while purchasing the house. If the developer has used premium material for the house, then naturally the sale price will be higher. Also, selecting the first floor area variable was creative. As we can see from previous findings, people usually use basement

area or lot area to determine the house price. The first floor is the place where residents spend the majority of their time, and has various areas such as kitchen, living room, guest bedroom, etc. People usually take the decision of buying a house based on kitchen area size, guest bedroom, etc. which ultimately converges to first floor area. We were also being creative by first performing the descriptive statistics to form our hypotheses and then develop a linear regression model to test our hypothesis. Use of correlation and t-test to check the independent variables similarity and relation with the dependent variables.

Communicating on Results

a. Result Presentation and Applications

Based on regression mentioned before, the result of linear regression model is:

SalePrice = -133055 + 37128 * OverallQuality + 59.53 * FirstFloor + 6387 * Bedroom with all coefficients are significantly different from 0 and R square is 0.69878.

Due to coefficients' significance and value, following explanations are derived from model:

On average, each level of overall quality increase leads to $37,128 increase in predicted sales price. On average, predicted sales price increases $59.53 with one square feet more in first floor area. On average, one more bedroom raises $6,387 in predicted sales price. Sales price is significantly influenced by three independent variables: overall quality, first floor square feet and number of bedrooms. Three independent variables explained 69.78% change of dependent variable (sales price).

Following are two applications of our model and analysis.

i. Application 1

Scenario：A real estate firm is asked by a client to sell a house. There are several data about the house: pretty good quality and is rated as 8; first floor square feet is 1,325 sq; 3 bedrooms total above ground. To better pricing the house and provide a reasonable market price to client, our model is used to estimate sales price. Plus, overall quality is 8, first floor square feet is 1,325 and number of bedrooms is 3 into SalePrice = -133,055 + 37,128 * OverallQuality + 59.53 * FirstFloor + 6,387 * Bedroom. We get sales price =$262,007.25.

Conclusion: The market price for this house is $262,007.25. If the client wants a better deal, sales price should be higher than $262,007.25; if the client wants to sell the house quickly, the sales price should be lower than $262,007.25. Also, $262,007.25 is the most reasonable price for this house. The final price should not be too far away from this number.

i. Application 2

Scenario: A portfolio manager invests in real estate. The manager needs to know whether a house is currently overvalued or undervalued in order to make buy-or-sell decision. A house, which is prices at $199,000, has average quality (overall quality = 6), 1520 square feet in first floor area and 4 bedrooms in total. Based on model, this house is worth $205,746.6, which is higher than the current price.

Conclusion: The portfolio manager should buy this house because the house is undervalued currently. The manager will get $6,746.6 return immediately.

b. Limitations

We test three assumptions for residual (normality, independence, homogeneity of variance) and results shows that normality and homogeneity of variance are failed, and independence is held. Although some assumptions are failed, our model and analysis are still significant and estimators are unbiased.

Failed normality means that residual is not distributed normally. If sample size is larger than 15, failed normality has no influence on estimators' unbiasedness (The Minitab Blog, 2014). Since normality is failed, it should be noted that confidence interval is not accurate. Any analysis involved estimators' standard error is biased.

Heteroskedasticity means variance of residual is not constant and bias significance test, which probably challenge our regression model. Current independent variables have some relationships with residual. From scatter plots, it is obvious that variance of residual increases with increasing first floor square feet. In future regression analysis, some factors in residual now should be treated as independent variables to better predict sales prices.

Exhibition

Figure 1 – Scatter plot of all variables

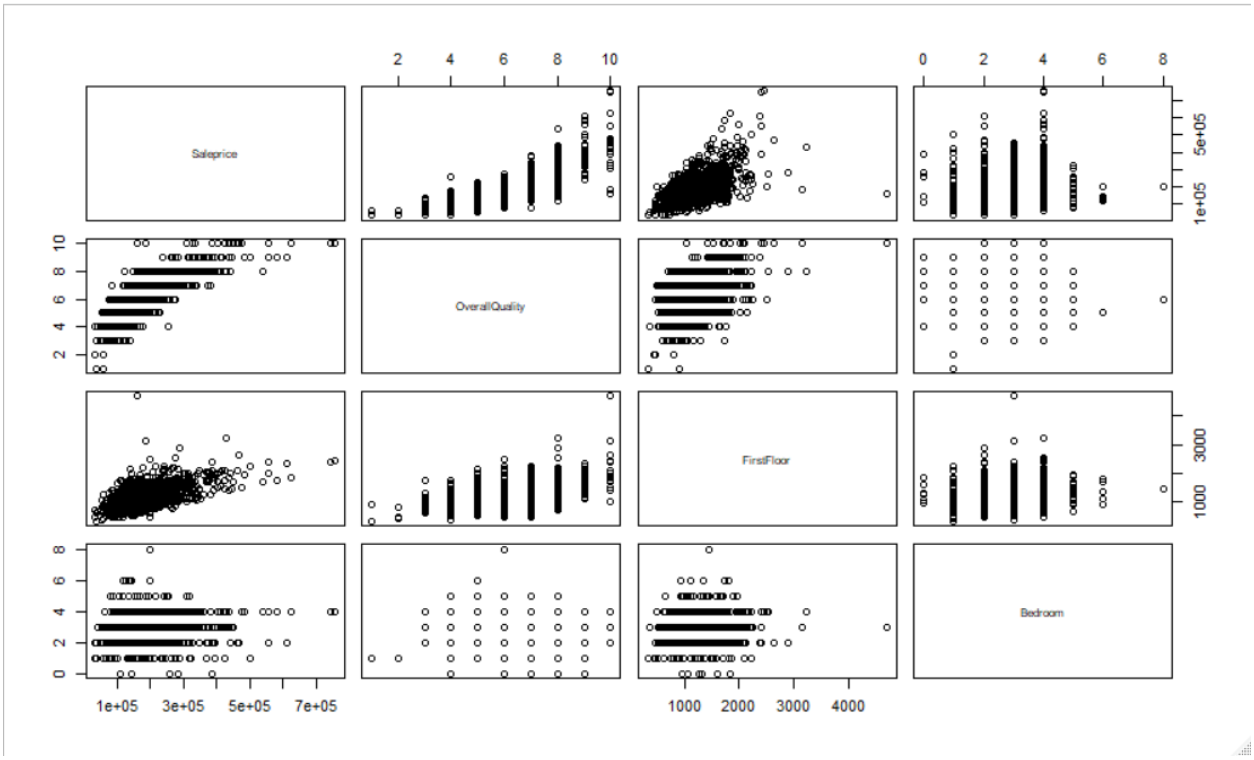


Figure 2 – Scatter plot of Overall Quality by Sale Price

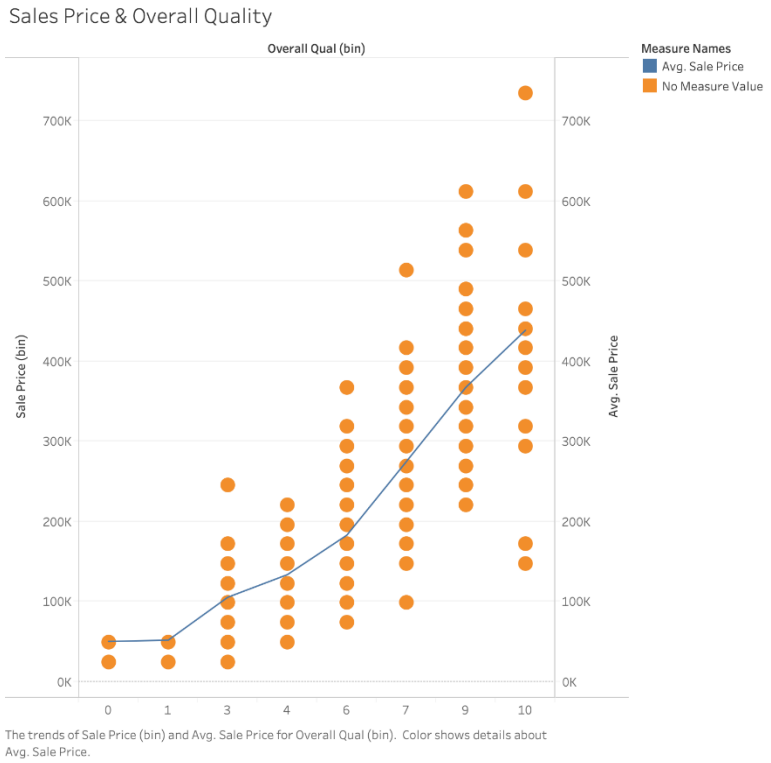


Figure 3 – Box plot of Overall Quality by Sale Price

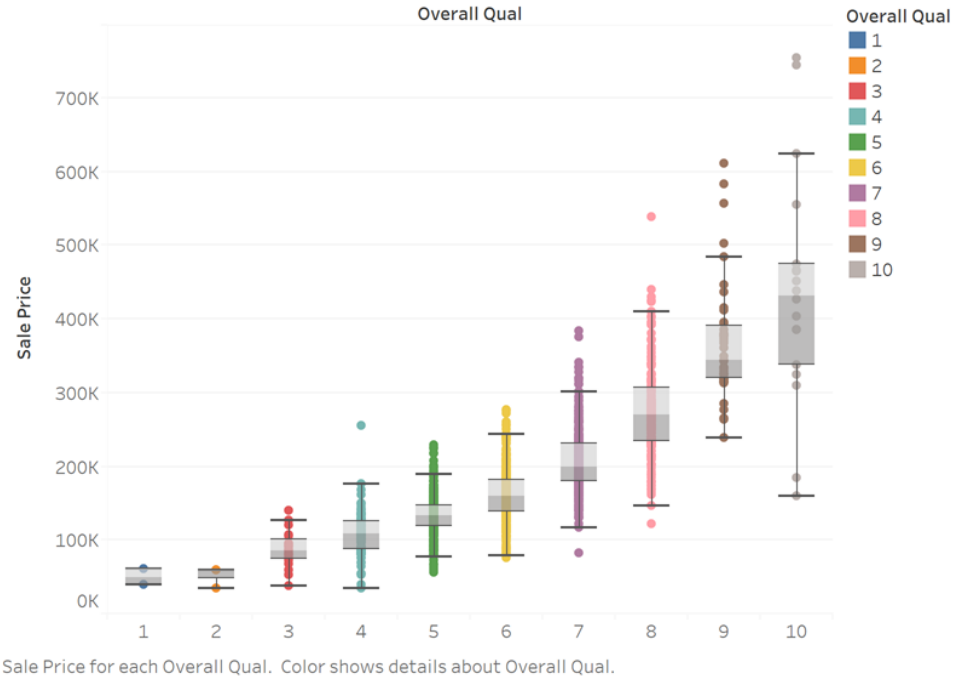


Figure 4 – Scatter plot of First Floor Square Feet y by Sale Price

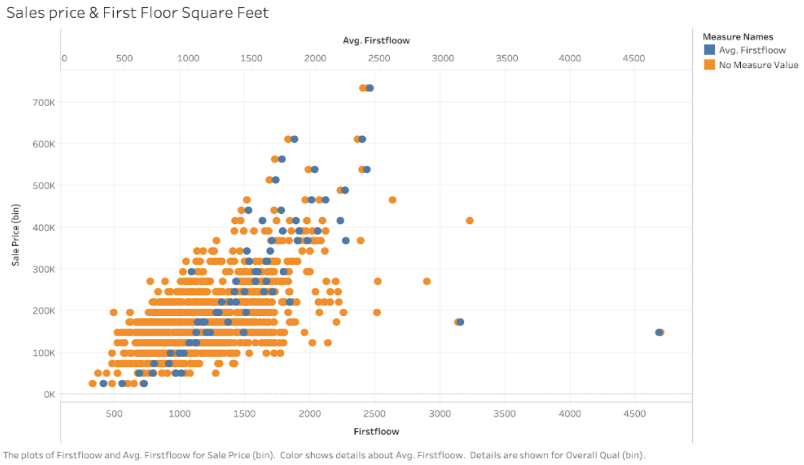


Figure 5 – Box plot of First Floor Square Feet y by Sale Price (1)

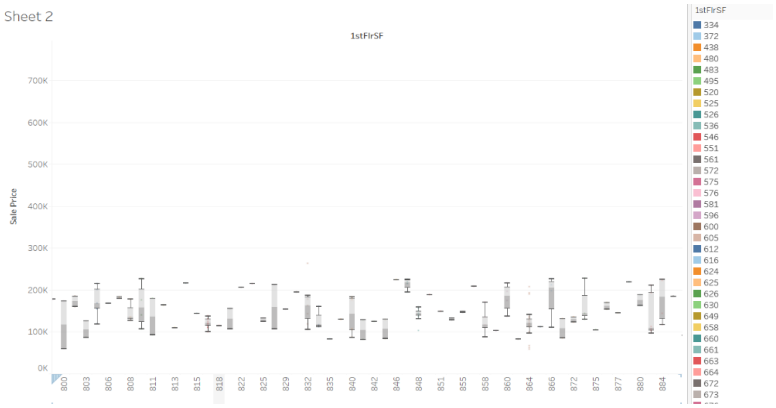
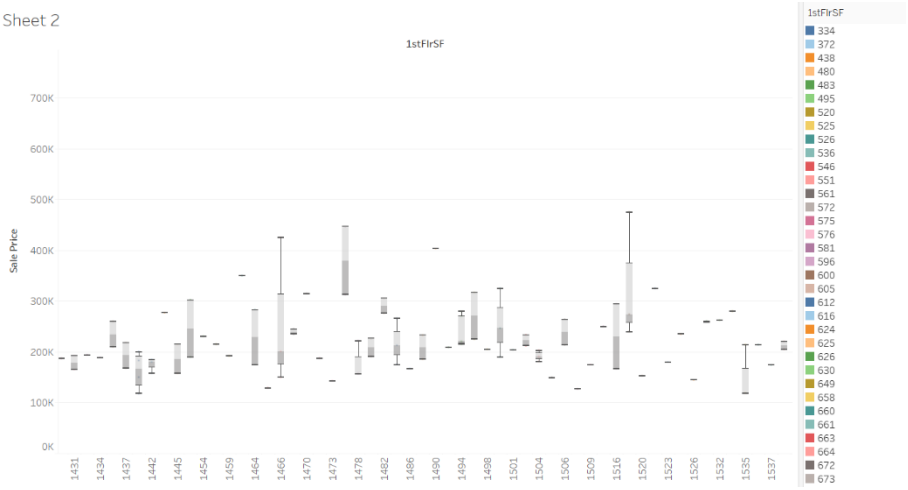


Figure 6 – Box plot of First Floor Square Feet y by Sale Price (2)



(Note: The first floor square feet is below 900, we can see the sale price is also below 200K, however, the first floor square feet is above 1,300, we can observe that variance of the sale price is larger.)

Figure 7 – Scatter plot of Number of Bedroom by Sale Price

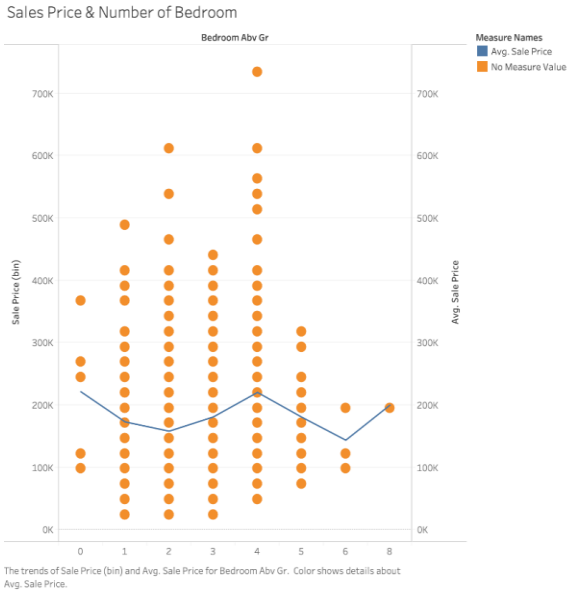


Figure 8 – Box plot of Number of Bedroom by Sale Price

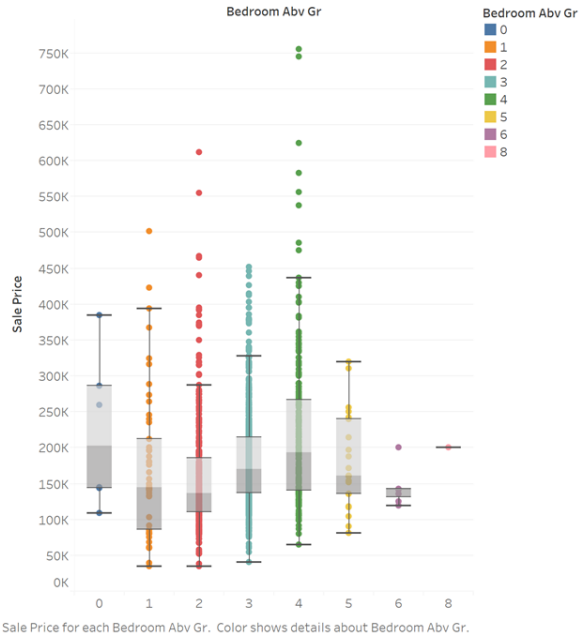


Figure 9 – Linear Regression Model

```
Call:
lm(formula = Saleprice ~ OverallQuality + FirstFloor + Bedroom)

Residuals:
    Min      1Q  Median      3Q     Max
-376701  -25398   -1581   20704  345734

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)    -1.331e+05  6.322e+03  -21.05  < 2e-16 ***
OverallQuality  3.713e+04  9.422e+02   39.41  < 2e-16 ***
FirstFloor      5.953e+01  3.381e+00   17.61  < 2e-16 ***
Bedroom         6.387e+03  1.416e+03    4.51 6.99e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 43720 on 1456 degrees of freedom
Multiple R-squared:  0.6978,    Adjusted R-squared:  0.6972
F-statistic:  1121 on 3 and 1456 DF,  p-value: < 2.2e-16
```

Figure 10 – Residual Test for Normality, Independence and Homogeneity of Variance

```
> shapiro.test(Model$residual)

        Shapiro-Wilk normality test

data:  Model$residual
W = 0.88454, p-value < 2.2e-16

> # 2. test the independence
> require(car)
> durbinWatsonTest(Model)
 lag Autocorrelation D-W Statistic p-value
   1   -0.0004803672      2.000913   0.918
 Alternative hypothesis: rho != 0

> # 3. test the Homogeneity of Variance
> require(car)
> ncvTest(Model)
Non-constant Variance Score Test
Variance formula: ~ fitted.values
Chisquare = 1195.846, Df = 1, p = < 2.22e-16
```

Figure 11 – Test for Multicollinarity

```
> M <- lm(houseprice$SalePrice~.,data=houseprice)
> vif(Model)
OverallQuality     FirstFloor        Bedroom
      1.296173       1.303938       1.018751
> vif(M)
 OverallQual         FlrSF BedroomAbvGr
    1.296173      1.303938      1.018751
>
```

Reference

The documentation file explaining details of the Ames Housing Dataset

http://jse.amstat.org/v19n3/decock/DataDocumentation.txt

The Ames Housing Dataset

http://www.amstat.org/publications/jse/v19n3/decock/AmesHousing.xls

The text file version of the Ames Housing Dataset

http://www.amstat.org/publications/jse/v19n3/decock/AmesHousing.txt

Ames, Iowa: Alternative to the Boston Housing Data as an End of Semester Regression Project

http://jse.amstat.org/v19n3/decock.pdf

How Important Are Normal Residuals in Regression Analysis?

https://blog.minitab.com/blog/adventures-in-statistics-2/how-important-are-normal-residuals-in-regression-analysis