**Linear Regressions on Home Sales Prices**

**Leonardo Ji**

**BIA 6309 Linear and Multivariate Models**

**Table of Contents**

[**Table of Contents** 2](#_Toc476574548)

[**Introductions** 3](#_Toc476574549)

[**Data Description** 4](#_Toc476574550)

[**Analysis** 9](#_Toc476574551)

[**Linear Regression using living area, overall quality, neighborhood on sales price** 9](#_Toc476574552)

[**Beta Regression using quantitative variables** 11](#_Toc476574553)

[**Multiple Linear Regression using both categorical and quantitative variables** 11](#_Toc476574554)

[**An interaction term Overall Quality and Living Area** 11](#_Toc476574555)

[**More interaction terms** 12](#_Toc476574556)

[**Reduce overfitting and simplify model** 13](#_Toc476574557)

[**Results and Conclusions** 16](#_Toc476574558)

[**References** 16](#_Toc476574559)

[**Appendix** 17](#_Toc476574560)

# **Introductions**

In this class we learned two important supervised statistic learning methods: linear regression, and logistic regression. Linear regression is the most important statistical learning technique and very good at predicting quantitative results. Logistic regression uses logistic function predicting binomial result (yes or no answer).

Linear regression assumes linear relationship between variables and result of the linear function. (Trevor Hastie, 2001) The result of linear function is equals sum of intercept and each coefficient multiply by variable. (James, Daniela , Hastie, & Tibshirani, 2013)

*Y = β0 + β1X*

Because the variables are known in order to solve linear regression, we want to estimate coefficients β0 and β1 on the right hand side. It becomes a residual sum of square (RSS) minimization problem. Using calculus, we can estimate coefficient from taking partial derivative on the RSS term. Coefficient is solved by linear equation. If there are two variables on the right hand side, then the linear regression fit is a line. But If there are multiple variables on the right hand side then the regression fitted becomes a plain or space in multiple dimensions. The quality of a linear regression fit is typically assessed using two related quantities: the residual standard error (RSE) and the R2 statistic. (James, Daniela , Hastie, & Tibshirani, 2013)

*RSS = .*

*R2 = (TSS – RSS) / TSS = 1− RSS*

The hard part is to decide which variables to pick on the right hand side to come up with a model with highest R-squared and low P value. The smaller P value for the selected term means the selected term is more statistically significant. In R studio if P value is smaller than 0.001 then it has three stars which means most significant with high confidence interval.

Data scientist uses CRISP-DM process (Sharda, 2010), an iterative data mining process to experimenting and gain knowledge from data. In my final project I follows this data mining process to find a better linear regression model to predict Ames, Iowa house sales price using several house attributes in dataset. In the end if we find a model with 0.9 adjusted R-squared then we can submit the model to the completion to use on testing data.

We use R, R studio, ggplot2, Tableau software for this final project. R is an open source statistical programming language to do statistical analysis and machine learning. R studio is an open source IDE for R, it also manages packages installation and software updates. ggplot2 is a software package within R for data visualization. Tableau is a commercial software product that quickly produce data visualization without writing a single line of code. Both ggplot2 and Tableau create beautiful charts and graphs that follow principles of good visualization that improves understanding the data (Yao, 2013). The following sections we show graphs and charts created by ggplot2 and Tableau.

# **Data Description**

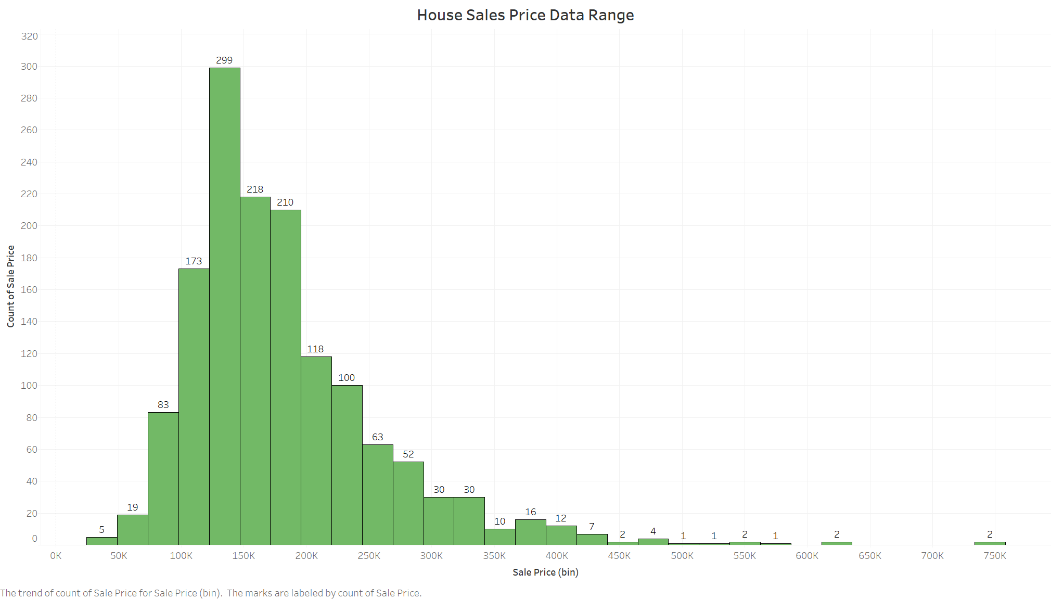
This data is provided by a Kaggle competition (Kaggle house price advanced regression technique, 2017) called “Housing prices advanced regression technique”. Kaggle hosts many data science competition including famous one million dollars Netflix Price competition for any team come up with a predictive analytics algorithm that is better than current model by 10% when top teams merged together and an ensemble method won (Siegel, 2016). This dataset includes Ames, Iowa house sales price from 2006 to 2010. The dataset includes 82 attributes of for each houses sold. This dataset contains both quantitative and categorical variables. Like any Kaggle competition there is a training dataset and a testing dataset. This training data set contains 1460 rows and 81 columns. Each column is an attribute such as lot size, garage car size, years built, kitchen grade, number of bedrooms, basement size…etc. The attributes names are listed on Kaggle competition site (Kaggle house price advanced regression technique, 2017). The sales price is the y value regression model tries to predict. The sales price unit is USD. The average sales price is $180900, and the price range is from $34900 to $755000.

**Table 1 Ames Iowa House Sale Prices (in USD)**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Min. | 1st Qu. | Median | Mean | 3rd Qu. | Max. |
| 34900 | 130000 | 163000 | 180900 | 214000 | 755000 |

Figure 1 bar chart is created in Tableau. By visualize the sales price data and occurance of the data at each price point we can see it is uniform distributed lightly skewed to the higher price.

**Figure 1 House Sales Price Data Range**

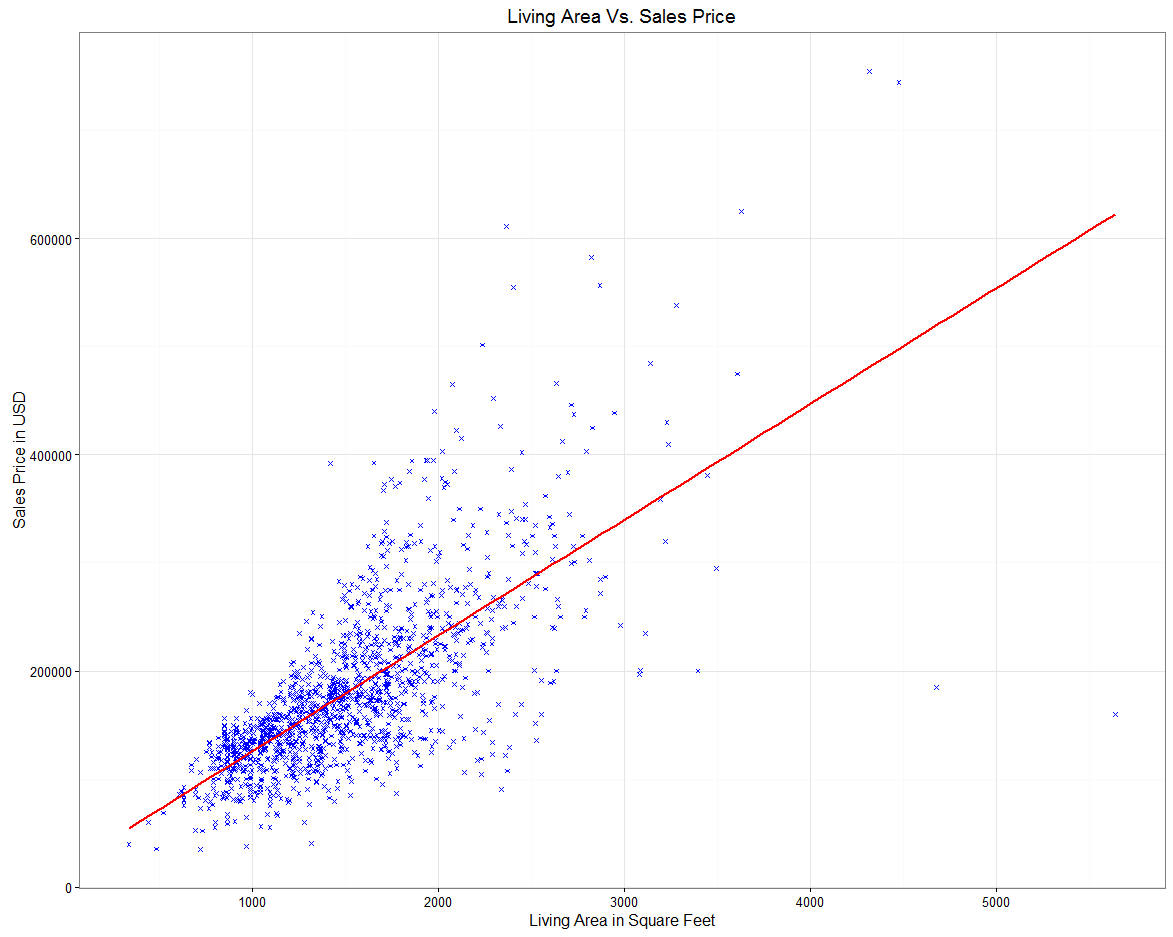


In R studio we can calculate the correlation matrix of those quantitative attributes to understand if there are any linear correlations between attributes. This information would help us select which variables are useful to add to the model. From Table 2 it appears Overall Quality, Living Area have the highest correlation to sales price. Next we will create data visualization on those variables.

**Table 2 Correlation Matrix of Quantitative Attributes**

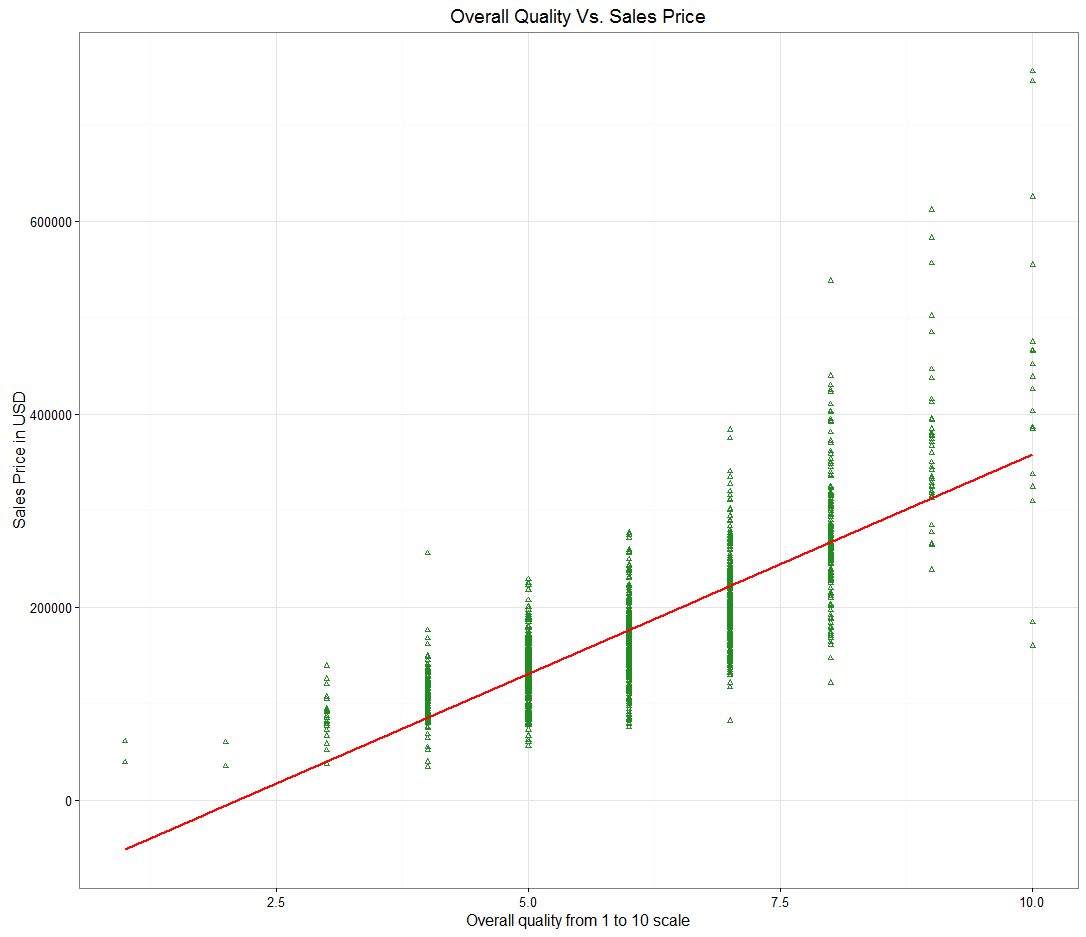
|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | SalePrice | OverallQual | LotArea | GrLivArea | GarageCars | BedroomAbvGr | KitchenAbvGr | TotalBsmtSF | YearBuilt |
| SalePrice | 1 | 0.7909816 | 0.26384335 | 0.7086245 | 0.64040920 | 0.16821315 | -0.13590737 | 0.61358055 | 0.52289733 |
| OverallQual | 0.7909816 | 1 | 0.10580574 | 0.5930074 | 0.60067072 | 0.10167636 | 0.18388223 | 0.53780850 | 0.57232277 |
| LotArea | 0.2638434 | 0.1058057 | 1 | 0.2631162 | 0.15487074 | 0.11968991 | 0.01778387 | 0.26083313 | 0.01422765 |
| GrLivArea | 0.7086245 | 0.5930074 | 0.26311617 | 1 | 0.46724742 | 0.52126951 | 0.10006316 | 0.45486820 | 0.19900971 |
| GarageCars | 0.6404092 | 0.6006707 | 0.15487074 | 0.4672474 | 1 | 0.08610644 | -0.05063389 | 0.43458483 | 0.53785009 |
| BedroomAbvGr | 0.1682132 | 0.1016764 | 0.11968991 | 0.5212695 | 0.08610644 | 1 | 0.19859676 | 0.05044996 | -0.07065122 |
| KitchenAbvGr | -0.1359074 | -0.1838822 | -0.01778387 | 0.1000632 | -0.05063389 | 0.19859676 | 1 | -0.06890064 | -0.17480025 |
| TotalBsmtSF | 0.6135806 | 0.5378085 | 0.26083313 | 0.4548682 | 0.43458483 | 0.05044996 | -0.06890064 | 1 | 0.39145200 |
| YearBuilt | 0.5228973 | 0.5723228 | 0.01422765 | 0.1990097 | 0.53785009 | -0.07065122 | -0.17480025 | 0.39145200 | 1 |

**Figure 2 Linear Relationship between living area and sales price**



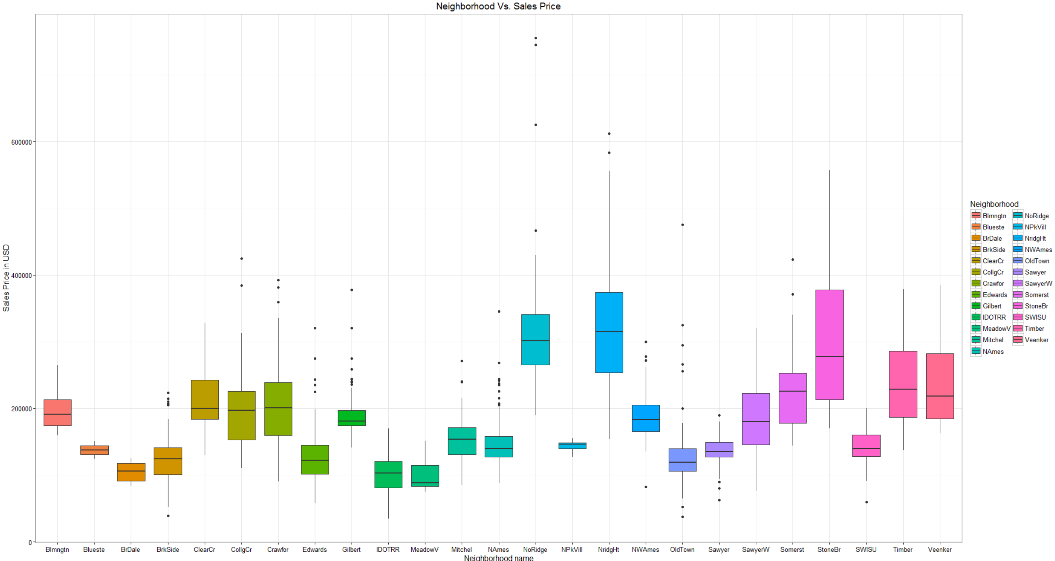
Using ggplot2 (ggplot doc, n.d.) we can create a scatter plot of living area vs sales price. The living area (attribute name: GrLivArea) is measured in square feet. The red line is a linear regression fitted line. We can see house sales price are generally increasing with bigger living area. This knowledge is common sense but it is still good to verify the data by visualization. Another thing we see from this graph is there are a lot of houses from 1000 square feet to 2000 square feet. There are fewer hours above 3000 square feet and they are above the regression line. There are a few data points higher than 4000 square feet but below red regression line. We could remove outliers to improve model accuracy. The Intro to Statistical Learning book listed outlier is one of the potential problems for linear regression. Other problems are Non-linearity of the response-predictor relationships, correlation of error terms, non-constant variance of error terms, high-leverage points, collinearity (James, Daniela , Hastie, & Tibshirani, 2013). In the next analysis section, we will focus on issue of collinearity.

**Figure 3 Linear Relationships between overall quality and sales price**

****

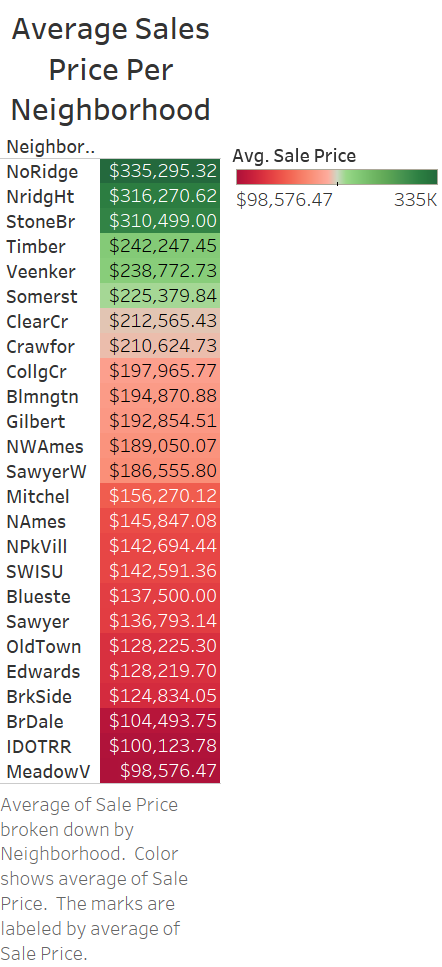
The high quality houses have higher sales price.This chart shows clearly the linear relationship between house overall quality and sales price. We can see from Table 1 the correlation is 0.79 between overall quality and sales price. It is the highest correlation between attributes and sales price. A red regression line is added to show linear relationship. The overall quality attribute would be a very good candidate for our linear regression model.

**Figure 4 Neighborhoods and sales price**

****

The box plot of neighborhood vs sales price is created from ggplot2. The nice thing about box plot is that it clearly shows the average and the range. The houses’ sales prices are higher in good neighborhoods. For example, the Northridge Heights area house has a mean sales price $3163000, while Northpark Villa has a mean sales price of $1427000. Figure 5 Tableau heat map is sorted by mean sales price per neighborhoods. We can see neighborhood attribute made a big difference on sales prices.

**Figure 5**

****

# **Analysis**

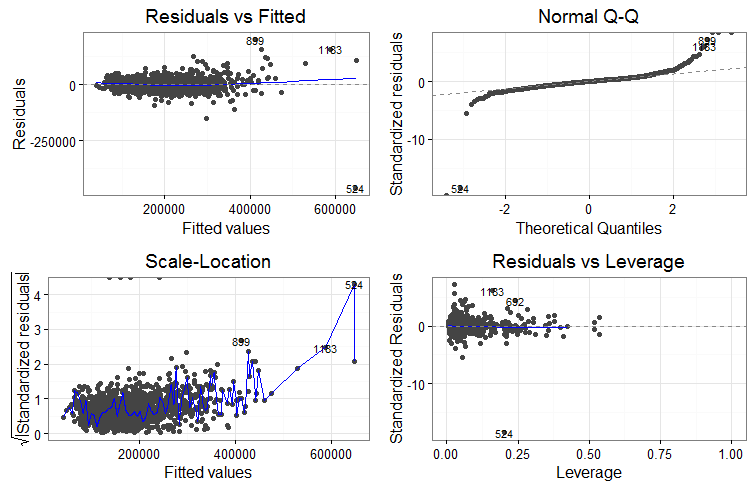
## **Linear Regression using living area, overall quality, neighborhood on sales price**

In this previous section we created some charts to show strong correlations between living area, neighborhood, and overall quality to sales price. We can create a simple linear regression model to estimate sales price. We use R to run it then test this model and rely on analysis of variance (ANOVA) table to assess model accuracy. For this first model the summary of the model shows R-squared equals to 0.7868 and adjusted R-squared is 0.7976. Standard error is 34270. Adjusted R-squared takes into account the number of terms in the model. We will always use adjusted R-squared for this project. The result shows 79% of the sales price can be explained by this simple model of three terms: overall quality, neighborhood, and living area. The other 21% are not explained by the model. Figure 6 shows residuals using R ggfortify library. We will use this linear regression model as our base model. In the next few sections we will try to improve on it.

**Table 3 Analysis of Variance Table**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Response:SalePrice | Df | Sum Sq | Mean Sq | F value | Pr(>F) |
| OverallQual | 1 | 5760947152122 | 5760947152122 | 4206.204 | < 0.00000000000000022  \*\*\* |
| Neighborhood | 24 | 807047644756 | 33626985198 | 24.552 | < 0.00000000000000022  \*\*\* |
| GrLivArea | 1 | 677235591413 | 677235591413 | 494.466 | < 0.00000000000000022  \*\*\* |
| Residuals | 1433 | 1962680946320 | 1369630807 |  |  |

**Figure 6**

****

## **Beta Regression using quantitative variables**

How do we know we used the best quantitative variables out of all variables in the dataset? We can scale the data down and use Beta regression (z scale). Then, we can use R leaps library’s selection algorithm to select best group of variables for the regression model by highest R-squared value.

**Table 4 Leaps selection algorithm R-squared**

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
| R2 | 0.6253951 | 0.7137809 | 0.7415355 | 0.7571912 | 0.7659163 | 0.7716466 | 0.7757506 | 0.7786099 | 0.7816831 | 0.7837320 | 0.7836072 | 0.7834982 |

The Leaps selection algorithm shows increasing R-squared value. When using all 12 variables the R-squared is the highest at 0.783. This is a fairly good result. But if we compare it to result from the last section when both categorical variable (neighborhood) and quantitative variables (overall quality and living area) are used this result is not that good.

## **Multiple Linear Regression using both categorical and quantitative variables**

Looks like we have to add both categorical and quantitative variables. We can analyze data in Tableau and find those terms that produce higher sales prices with increasing values. Appendix shows data visualization for those selected variables and explains why they are chosen. If we add all seventeen categorical and quantitative variables (OverallQual, Neighborhood, LotArea, GrLivArea, GarageCars, MasVnrArea, GarageArea, BedroomAbvGr, KitchenAbvGr, TotalBsmtSF, YearBuilt, YearRemodAdd, SaleType, SaleCondition, HouseStyle, RoofMatl, MSZoning, BldgType, Foundation) to the model, we can get a far better result. In R it shows the new model’s adjusted R-squared is now 0.8621. Standard error is 29450. Both R-squared and standard error are much better than first model that use three variables. P values are all significant except sales condition which we can remove in the next model.

## **An interaction term Overall Quality and Living Area**

We know interaction term results from two variables together create synergic effect (James, Daniela , Hastie, & Tibshirani, 2013). Figure 7 is a box plot of overall quality and living area size over sales price created in Tableau. Figure 7 shows the bigger houses in excellent overall quality sold at much higher prices. Recall figure 2 shows houses bigger than 3000 square feet sold at higher prices. We can create an interaction term by multiply overall quality to living area if living area is greater than 3000 square feet.

**Figure 7**

****

If we add interaction term to the previous model, then can we get a better R-squared result? Yes, the adjusted R-squared increases to 0.8747. Residual standard error is 28070. Both number are slightly better than the last model. The very small p value is 0.00001358512 means it is statistically significant.

## **More interaction terms**

What if we add more interaction terms to the model? Would we get a much better results? In this next model we add 16 interaction terms. Enter the following terms into R, the adjusted R-squared is 0.9066, standard error is 24220. We have hit our 0.90 adjusted R-squared target.

*SalePrice ~ largeGoodQualityHome+Neighborhood\*OverallQual+Neighborhood\*BldgType+KitchenAbvGr\*KitchenQual+RoofStyle\*RoofMatl+MasVnrType\*MasVnrArea+LotArea\*LotShape+BsmtUnfSF\*TotalBsmtSF+SaleType\*SaleCondition+YearBuilt\*YearRemodAdd+BsmtQual\*BsmtCond+GrLivArea+GarageCars+BedroomAbvGr+MSZoning+Foundation*

## **Reduce overfitting and simplify model**

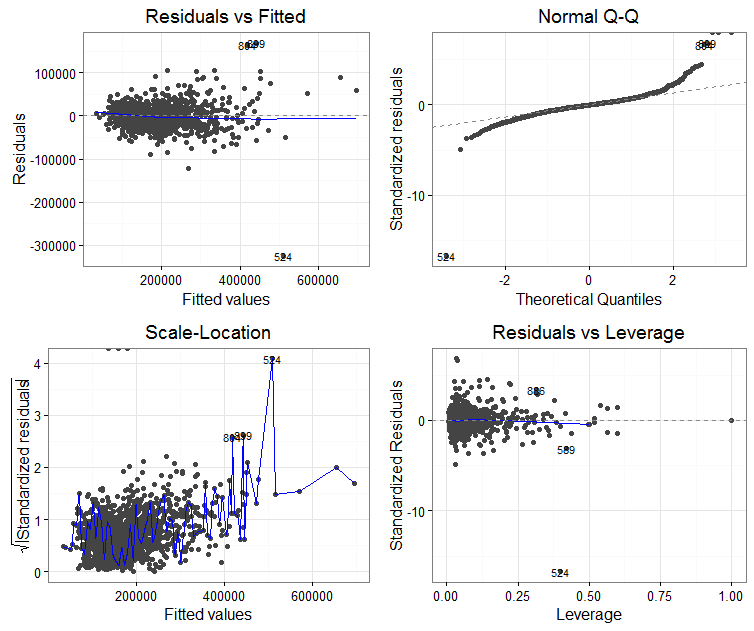
There are sixteen terms on the last model. A good model usually has less than that. We want to reduce number of terms to prevent overfitting. When we apply the model to real data, overfitting model would give us biased result. This is something we need to consider. Can we still get a good R-squared value by reduce number of terms? By trial and error, we reduced the number of variables and still able to get adjusted R-squared at 0.90, and standard error at 25170. These numbers are only slightly lower than our last model. All terms have smaller p value < 0.001. gives 95% confidence actual sales prices are within the range. Table 5 is analysis of variance table; figure 6 is residual graphs. We can compare them to previous models. We have a good model.

*SalePrice~GrLivArea\*OverallQual+Neighborhood\*OverallQual+BldgType+KitchenAbvGr+KitchenQual+RoofStyle+RoofMatl+LotArea+BsmtUnfSF+TotalBsmtSF+SaleType+SaleCondition+YearBuilt+YearRemodAdd*

**Table 5 Analysis of Variance Table**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Response: SalePrice | Df | Sum Sq | Mean Sq | F value | Pr(>F) |
| GrLivArea \* | 1 | 4623740248497 | 4623740248497 | 7304.2201 | < 0.00000000000000022 \*\*\* |
| OverallQual | 1 | 1952303470545 | 1952303470545 | 3084.0950 | < 0.00000000000000022 \*\*\* |
| Neighborho d | 24 | 677235591413 | 677235591413 | 494.466 | < 0.00000000000000022 |
| BldgType | 4 | 126831271291 | 31707817823 | 50.0895 | < 0.00000000000000022 \*\*\* |
| KitchenAbvGr | 1 | 19465256261 | 19465256261 | 30.7497 | 0.0000000351692287 \*\*\* |
| KitchenQual | 3 | 166543687717 | 55514562572 | 87.6975 | < 0.00000000000000022 \*\*\* |
| RoofStyle | 5 | 23864032976 | 4772806595 | 7.5397 | 0.0000005384959917 \*\*\* |
| RoofMatl | 7 | 170660114362 | 24380016337 | 38.5136 | < 0.00000000000000022 \*\*\* |
| LotArea | 1 | 44976730153 | 44976730153 | 71.0507 | < 0.00000000000000022 \*\*\* |
| BsmtUnfSF | 1 | 22623695072 | 22623695072 | 35.7391 | 0.0000000028720413 \*\*\* |
| TotalBsmtSF | 1 | 183330162403 | 183330162403 | 289.6105 | < 0.00000000000000022 \*\*\* |
| SaleType | 8 | 47708840172 | 5963605021 | 9.4208 | 0.0000000000009714 \*\*\* |
| SaleCondition | 5 | 15735999420 | 3147199884 | 4.9717 | 0.0001618 \*\*\* |
| YearBuilt | 1 | 32144044600 | 32144044600 | 50.7786 | 0.0000000000016652 \*\*\* |
| YearRemodAdd | 1 | 12488965596 | 12488965596 | 19.7291 | 0.0000096400697462 \*\*\* |
| GrLivArea:OverallQual | 1 | 94640417162 | 94640417162 | 149.5055 | < 0.00000000000000022 \*\*\* |
| OverallQual:Neighborhood | 22 | 133159998394 | 6052727200 | 9.5616 | < 0.00000000000000022 \*\*\* |
| Residuals | 1372 | 868507730740 | 633023127 |  |  |

**Figure 8**



# **Results and Conclusions**

The multivariate linear regression model did a good job at estimating house sales prices. It is pretty clear from data visualization that relationship between house sales prices and some of the attributes like living area, neighborhood, and overall quality resembled linear relationships. Both quantitative and categorical attributes were needed for the regression model to get good results. The difficult part was finding which attributes to use and how many attributes to use. P-value and R-squared number are used to assess model accuracy. Data visualizations in Tableau and ggplot helped us understand data better and also find good variables to add to the model. We learned ways to gain better accuracy by using interaction terms, and remove outliers. We used CRISP-DM process to iteratively look for better model experimenting different terms. The last reduced model achieved 0.9 R-squared with all remaining terms’ p-value still statistically significant less than 0.001. The standard error is small. We can further reduce number of terms to prevent overfitting. Over-fitted model from training data does not produce accurate result on actual testing data. There are other advanced regression techniques to improve on linear regression model but that will be the future work.

# **References**

*ggplot doc*. (n.d.). Retrieved from ggplot doc: http://docs.ggplot2.org/current/geom\_abline.html

James, G., Daniela , W., Hastie, T., & Tibshirani, R. (2013). *An Introduction to Statistical Learning with Applications in R.* Springer.

*Kaggle house price advanced regression technique*. (2017). Retrieved from Kaggle: https://www.kaggle.com/c/house-prices-advanced-regression-techniques

Sharda, E. T. (2010). *Business Intelligence: A Managerial Approach.* Prentice Hall.

Siegel, E. (2016). *Predictive Analytics: The Power to Predict Who Will Click, Buy, Lie, or Die.* Wiley.

Trevor Hastie, J. H. (2001). *The Elements of Statistical Learning Data Mining, Inference, and Prediction.* Springer.

Yao, N. (2013). *Data Points.* Wiley.

# **Appendix**

Figure 9 High Kitchen quality and Kitchen above grade have higher mean sales prices.

K



Figure 10 Good Roof Material and certain Roof style have higher sales prices.

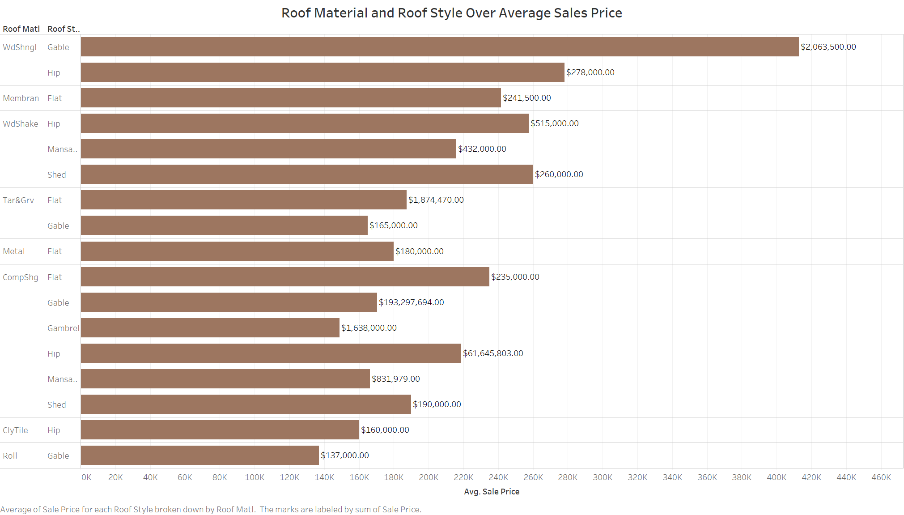


Figure 11 New houses and Newly remodeled houses have higher sales prices. Year sold has not impact on sales prices.

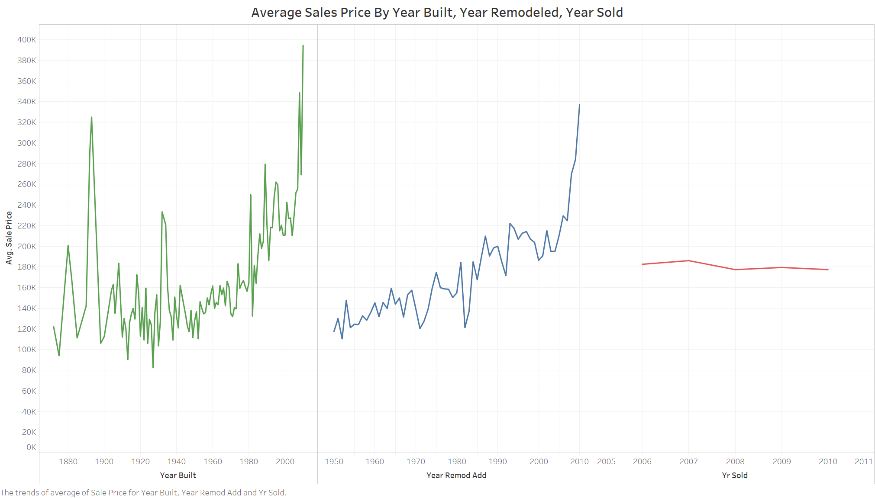


Figure 12 Cash payment or conventional 15% payment has higher sales prices.

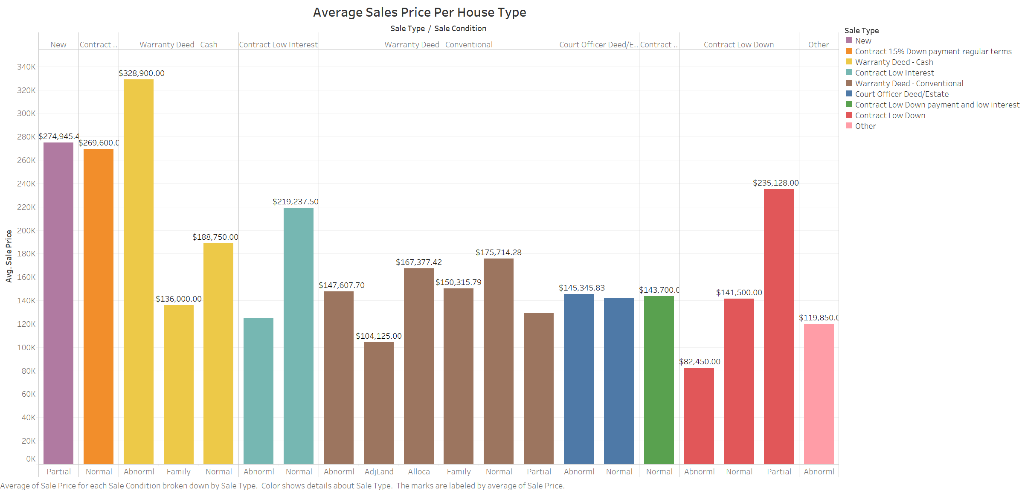
ffffff

Figure 13 Good neighborhood and good quality houses have high sales price.

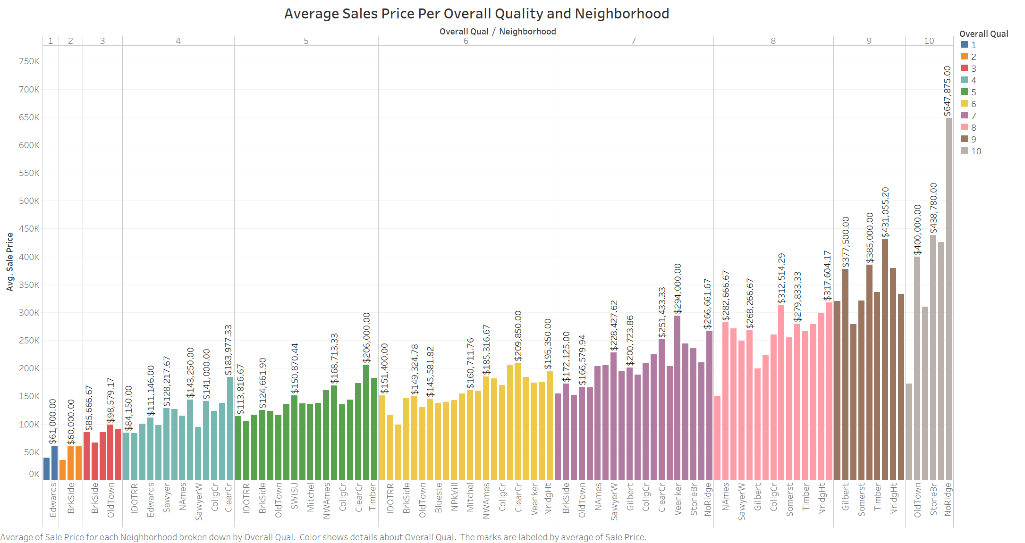


Figure 14 Higher unfinished basement size and higher total basement size has higher sales prices.

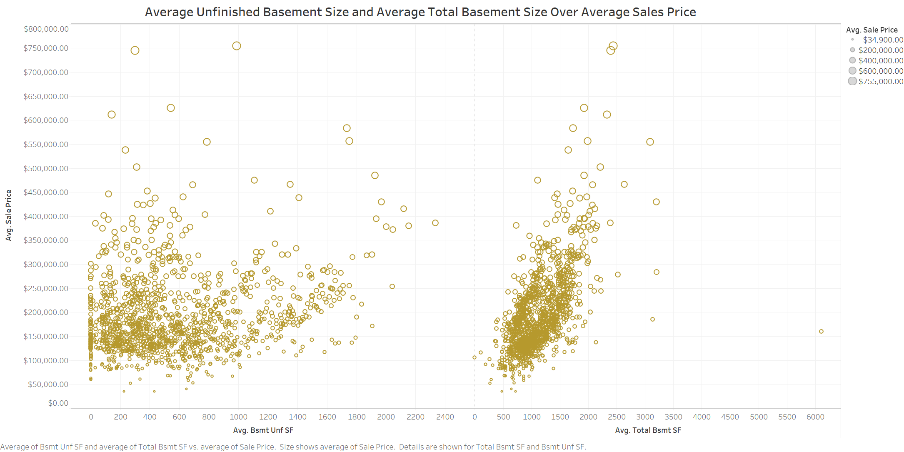


Figure 15 Higher Lot Area has higher sales price.

