STAT 425 Applied Regression and Design

Regression Analysis for AmesHousing Data

Group Member: Lijun Zhang, Chuanyue Shen, Lingzhu Gu

University of Illinois at Urbana-Champaign

Executive Summary

In this project, we analyze the AmesHousing data, which includes housing price in Ames, Iowa from 2006 to 2010, and other 81 variables. Sale price is treated as response, while others are used as possible predictors. There are 2930 observations initially, but includes some missing data and false data, thus data preprocessing is performed first. The features with the same value exceeding 80% is deleted, and the observation with missing categorical feature data is also discarded. The observation with missing numerical feature data is still reserved, but the value of numerical feature is replaced with its median. 2880 observations are left after preprocessing, with 42 categorical features and 35 numerical features.

Then collinearity is performed on the 35 numerical features with the criteria as 0.9, and no feature is deleted in this step. Using "boxcox" function, we find no transformation needed for the response. After that, variable selection is performed by both AIC and BIC methods, and forward, both, as well as backward models are all considered. 10 common numerical variables and 11 common categorical variables are observed in the six selected models, where the variables used for visualization analysis is chosen. Then AIC forward model is decided as our final model by ANOVA, which contains 22 numerical variables and 27 categorical variables to predict sale price. We also do diagnostics for our data, including high leverage points, outliers, and high influential points. 15 outliers and 12 high influential points are deleted. In addition, it is found that our data satisfies linear assumption but violates constant variance assumption and normality assumption.

In visualization, we explore the relationship between sale price and year built, first floor area, neighborhood, latitude/longitude, as well as sale condition/basement exposure/kitchen quality respectively. Three interesting findings are listed as follows.

1) The sale price of house increases more rapidly in latest years than past. And based on the trend line and forecasting, we have 95% confidence that the average sale price will stay stable around $260,000 in next five years.
2) NridgHt neighborhood is noticed with the highest sale price, while the BrDale has the lowest sale price. Also, the sale price in NirdgHt is the most sensitive to first floor area. Therefore, we suppose that BrDale is the most economical location for buying house.
3) The numbers of house available in different conditions are limited, and most houses have kitchen quality as TA, sale condition as Normal with no basement exposure. The houses with excellent (Ex) kitchen, Abnormal sale condition and good (Gd) basement exposure are noticed with the highest average sale price. Therefore, we can save money if we could tradeoff between basement and kitchen when selecting house.

Table of Contents

Table of Figures

1. Introduction

Housing price is an important indicator of the economy. The fluctuation of housing prices attracts a broad attention from the public. An analysis of real estate data is conducted in this term project. R package [AmesHousing] is used as the data set for analysis, which describes the sale of individual residential property in Ames, Iowa from 2006 to 2010 [1]. In the package, ames_raw and ames_geo are merged into one data set, resulting in 2930 observations and 80 variables including SalePrice and geo-info. The SalePrice will be used as the response.
The 80 explanatory variables include 23 nominals, 23 ordinals, 14 discrete, and 20 continuous [1]. The 23 nominal variables are used to identify various types of dwellings, materials, garages, and environmental conditions. The 23 ordinal variables are used to represent Streets (gravel or paved) and Neighborhoods (areas within the Ames city limits). The 14 discrete variables are used to quantify the number of kitchens, bedrooms, bathrooms, garage spaces and their respective location in each house. The 20 continuous variables are used to quantify the area of total dwelling square footage, total lot size, living area, and room dimensions of each house.
In this project, it is expected to find an appropriate model to explain the relationship between the SalePrice and the selected features. Interesting findings are also expected through the visualization in Tableau. The project is conducted with the following major steps: 1) data preprocessing, 2) variable selection and boxcox transformation, 3) model selection and diagnostics, and 4) visualization.

2. Data Preprocessing

A large amount of NA is observed in the data set. Data pre-processing is performed to obtain clean data before selections of variables and models.

The first step is to summarize each variable and filter out insignificant ones. The insignificant feature is assumed to have a same value exceeding 80%. After summarizing, if the same value (including NA) exceeds 80%, the feature is deleted due to its insignificance. In this step, four features, i.e. `Alley`, `Pool Qu`, `Fence`, and `Misc Feature` are deleted.

For the remaining features, NA is handled with. If the NA means no typical feature, the NA is replaced with “None”. Otherwise, NA is possibly a missing data. In this case, for categorical variables, the corresponding variable is deleted, while for numerical variables, the NA is replaced with median. The whole process of examining the remaining variables are elaborate in the following paragraphs.

The 9 variables related to basement are examined together. First, for each observation, check if NA exists simultaneously in `Bsmt Qual`, `Bsmt Cond`, `Bsmt Exposure`, `BsmtFin Type 1` and `BsmtFin Type 2`. If it is true, the NA is replaced with “None” and 0 is assigned to the corresponding values in `Bsmt Unf SF` and `Total Bsmt SF`. If the `BsmtFin Type1` (`BsmtFin Type 2`) is none, 0 is assigned to corresponding `BsmtFin SF1` (`BsmtFin SF2`). For the rest of NA, replace them with median for `Bsmt Unf SF` and `Total Bsmt SF`, and delete the observation for the other 7 basement related variables.

The 7 variables related to garage are examined together. For each observation, if NA exists simultaneously in `Garage Type`, `Garage Finish`, `Garage Yr Blt`, `Garage Qual` and `Garage Cond`, median is assigned to NA for `Garage Yr Blt` while “None” is assigned to the NA for other 4 features. For NA in `Garage Cars` and `Garage Area`, replace median with NA. For the rest of NA, the corresponding observations are deleted.

For `Electrical` as a categorical variable, only one observation has NA and is deleted. For `LotFrontage` as a numerical variable, NA is replaced with median in each observation. For `Fireplace Qu`, the NA is replaced with “None”.

After data preprocessing, a clean data set is obtained with a total of 2880 observations and 77 features. The 77 features are divided into 42 categorical variables and 35 numerical variables for further analysis.

The categorical variables are `MS SubClass`, `MS Zoning`, `Street`, `Lot Shape`, `Land Contour`, `Utilities`, `Lot Config`, `Land Slope`, `Neighborhood`, `Condition 1`, `Condition 2`, `Bldg Type`, `House Style`, `Overall Qual`, `Overall Cond`, `Roof Style`, `Roof Matl`, `Exterior 1st`, `Exterior 2nd`, `Mas Vnr Type`, `Exter Qual`, `Exter Cond`, `Foundation`, `Bsmt Qual`, `Bsmt Cond`, `Bsmt Exposure`, `BsmtFin Type 1`, `BsmtFin Type 2`, `Heating`, `Heating QC`, `Air`, `Electrical`, `Kitchen Qual`, `Functional`, `Fireplace Qu`, `Garage Type`, `Garage Finish`, `Garage Qual`, `Garage Cond`, `Paved Drive`, and `Sale Type`.

The 35 numerical variables are `Lot Frontage`, `Lot Area`, `Year Built`, `Year Remod/Add`, `Mas Vnr Area`, `BsmtFin SF 1`, `BsmtFin SF 2`, `Bsmt Unf SF`, `Total Bsmt SF`, `1st Flr SF`, `2nd Flr SF`, `Low Qual Fin SF`, `Gr Liv Area`, `Bsmt Full Bath`, `Bsmt Half Bath`, `Full Bath`, `Half Bath`, `Bedroom AbvGr`, `Kitchen AbvGr`, `TotRms AbvGrd`, `Fireplaces`, `Garage Yr Blt`, `Garage Cars`, `Garage Area`, `Wood Deck SF`, `Open Porch SF`, `Enclosed Porch`, `3Ssn Porch`, `Screen Porch`, `Pool Area`, `Misc Val`, `Mo Sold`, `Yr Sold`, `Longitude`, and `Latitude`.

3. Model Selection

3.1 Collinearity

Collinearity is performed on 35 numerical variables using the function “cor” in R. The criteria for correlation between two variables is set as 0.9. That is, if the correlation between two variables exceeds or equals to 0.9, only one of them will be reserved for further variable selection. In our project, no variable is deleted due to the failure of collinearity.

3.2 Response Transformation

We use “boxcox” function in R to make transformation of the response, sale price. Figure 1 shows the log-likelihood along different λ. The λ to maximize the likelihood of data is 0.22. As λ is usually round to 0.5, we select $\lambda = 0$ in our project. In other words, no transformation is required for our response.

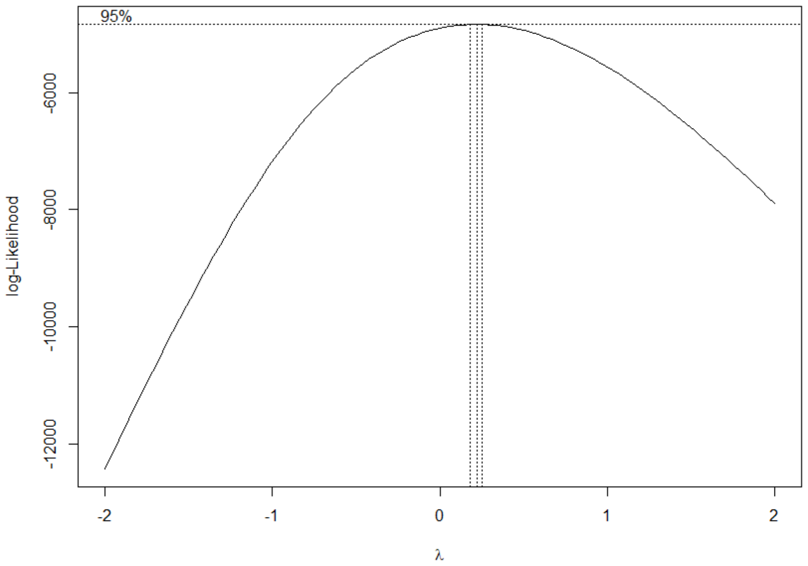


Figure 1. λ vs. Log-Likelihood

3.3 Variable Selection

For variable selection, “step” function in R is used. Addictive model is determined as the biggest model as the interaction model contains too many terms thus “step” function works inefficient. Since no interaction is considered in our model, transformation of variables, that is polynomial model, will also be neglected in our project.

In “step” function, both AIC and BIC methods are applied. And, all of the cases, forward, both, and backward, are done for each method respectively. Therefore, we obtain six models in this step. In AIC forward model, that is, 49 variables is selected, including 22 numerical variables and 27 categorical variables. In AIC both model, there is 51 variables, 23 numerical variables and 28 categorical variables. And the backward model from AIC method is the same as AIC both model. Using BIC, the variables selected is much less. BIC forward model contains 27 variables, among which there is 15 numerical variables and 12 categorical variables. And both BIC both model and BIC backward model contain 26 variables, 14 numerical variables and 12 categorical variables. The six models are shown as follows (the italic variables refer to numerical variables).

AIC forward model

SalePrice ~ `Overall Qual` + *`Gr Liv Area`* + Neighborhood + *`BsmtFin SF 1`* + `Roof Matl` + `MS SubClass` + `Bsmt Exposure` + `Overall Cond` + *`Year Built`* + *`Misc Val`* + `Condition 2` + `Sale Condition` + `Kitchen Qual` + *`Garage Area`* + *`Lot Area`* + `Screen Porch` + `Bsmt Qual` + `Exterior 1st` + `Condition 1` + *Fireplaces* + `Land Contour` + Functional + *`BsmtFin SF 2`* + *`Bsmt Unf SF`* + *`Pool Area`* + *`Mas Vnr Area`* + `Land Slope` + *`Year Remod/Add`* + `Lot Config` + `Mas Vnr Type` + *`2nd Flr SF`* + `Exter Qual` + *`Bsmt Full Bath`* + *`Full Bath`* + *Latitude* + `BsmtFin Type 2` + `BsmtFin Type 1` + *`Bedroom AbvGr`* + *`Garage Yr Blt`* + Street + *`Half Bath`* + `House Style` + `Lot Shape` +

`Garage Cars` + `Fireplace Qu` + `Roof Style` + `Garage Qual` + `Mo Sold` +
`Yr Sold`

AIC both model & backward model

SalePrice ~ `Lot Area` + `Year Built` + `Year Remod/Add` + `Mas Vnr Area` + `BsmtFin SF 1` +
`BsmtFin SF 2` + `Bsmt Unf SF` + `1st Flr SF` + `2nd Flr SF` + `Low Qual Fin SF`
+ `Bsmt Full Bath` + `Full Bath` + `Half Bath` + `Bedroom AbvGr` +
`Kitchen AbvGr` + Fireplaces + `Garage Cars` + `Garage Area` + `Wood Deck SF`
+ `Enclosed Porch` + `Screen Porch` + `Pool Area` + `Misc Val` + `Mo Sold` +
Latitude + Street + `Lot Shape` + `Land Contour` + `Lot Config` + `Land Slope` +
Neighborhood + `Condition 1` + `Condition 2` + `Bldg Type` + `House Style` +
`Overall Qual` + `Overall Cond` + `Roof Style` + `Roof Matl` + `Exterior 1st` +
`Mas Vnr Type` + `Exter Qual` + `Bsmt Qual` + `Bsmt Exposure` +
`BsmtFin Type 1` + `BsmtFin Type 2` + `Kitchen Qual` + Functional +
`Fireplace Qu` + `Garage Qual` + `Sale Condition`

BIC forward model

SalePrice ~ `Overall Qual` + `Gr Liv Area` + Neighborhood + `BsmtFin SF 1` + `Roof Matl` +
`MS SubClass` + `Bsmt Exposure` + `Overall Cond` + `Year Built` + `Misc Val` +
`Garage Cars` + `Screen Porch` + `Sale Condition` + `Lot Area` + `Condition 2` +
`Kitchen Qual` + `Total Bsmt SF` + Fireplaces + `Bsmt Qual` + `BsmtFin SF 2` +
`Land Contour` + `Garage Yr Blt` + `2nd Flr SF` + `Pool Area` + `Bedroom AbvGr`
+ `Mas Vnr Area` + `Full Bath`

BIC both model & backward model

SalePrice ~ `Lot Area` + `Year Built` + `BsmtFin SF 1` + `BsmtFin SF 2` + `Bsmt Unf SF` +
`1st Flr SF` + `2nd Flr SF` + `Full Bath` + `Bedroom AbvGr` + Fireplaces +
`Garage Yr Blt` + `Garage Area` + `Screen Porch` + `Pool Area` + `Misc Val` +
`Land Contour` + Neighborhood + `Condition 2` + `Bldg Type` + `Overall Qual` +
`Overall Cond` + `Roof Matl` + `Bsmt Qual` + `Bsmt Exposure` + `Kitchen Qual`
+ `Sale Condition`

In these six models,10 common numerical variables and 11 common categorical variables are observed, where the variables used for visualization analysis is selected. The common numerical variables are BsmtFin SF 1, Year Built, Misc Val, Lot Area, Fireplaces, BsmtFin SF 2, 2nd Flr SF, Pool Area, Bedroom AbvGr, Full Bath. And the common categorical variables include Overall Qual, Neighborhood, Roof Matl, Bsmt Exposure, Overall Cond, Screen Porch, Sale Condition, Condition 2, Kitchen Qual, Bsmt Qual, Land Contour.

Then, by ANOVA, we select the best model as the AIC forward model.

3.4 Diagnostic

3.4.1 High Leverage Points

The criteria to judge high leverage point is $2 \times p/n$, where p refers to the degree of freedom of predictors, and n is the sample size. In this project, the criteria is 0.054. Function “influence” in R is used to calculate the leverage for each sample, and 1156 high leverage points are observed. However, as high leverage points are those whose x value is far away from the center of the whole sample, not all the high leverage points are adverse for regression. Therefore, we keep the high leverage points.

3.4.2 Outliers

The criteria of outlier is determined by Bonferroni Correction. In this project, it is 4.204. The studentized residual can be calculated by function "rstudent" in R for each sample. Then 15 outliers are found and deleted in our project.

3.4.3 High Influential Points

High influential points are those with cook distance higher than 1. They can be outliers as well as high leverage points. Figure 2 shows the cook distance for all the samples. It is noticed that the cook distance of the majority of the sample points is pretty small. However, there are still 12 high influential points found and deleted.

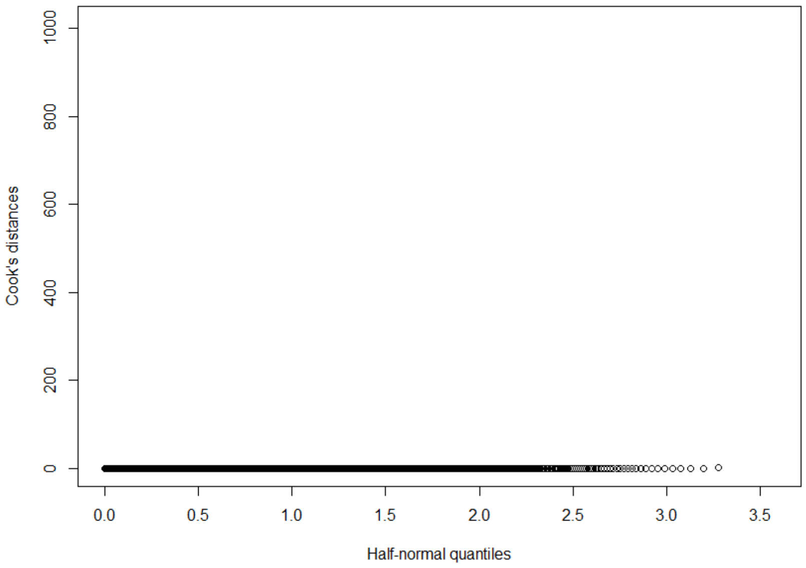


Figure 2. Cook distance

After deleting 15 outliers and 12 high influential points, 2853 observations remains for regression. We run the AIC forward model again using the 2853 observations, and obtain the final model, whose R^2 is 0.94, and mean square error (MSE) is 3.29×10^8.

3.5 Assumption Check

3.5.1 Linearity & Constant Variance Assumption

As Figure 3 depicts, Fitted values vs. Residuals Plot can be used to check linearity and constant variance assumptions. For any fitted value, the residuals seem roughly centered at 0, indicating that the linearity assumption is satisfied. However, it is noticed that the spread of the residuals non-uniform. It is wider when fitted value is larger than 3×10^5, and fewer points are observed for very small and large fitted values. Therefore, the constant variance assumption is violated here.

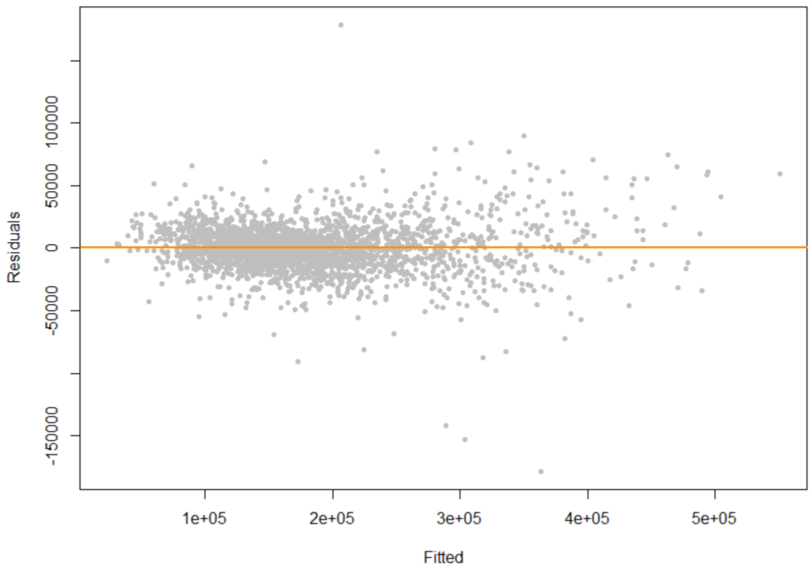


Figure 3. Fitted values vs. Residuals

Although Fitted values vs. Residuals Plot can give an idea about homoscedasticity, Breusch-Pagan Test is a more formal test for homoscedasticity assumption check. The null hypothesis and alternative hypothesis of BP test is as follows.

H_0: Homoscedasticity. The errors have constant variance about the true model.

H_a: Heteroscedasticity. The errors have non-constant variance about the true model

The p-value for our final model is 7.83044×10-198, much lower than 0.05, thus we reject null hypothesis. The similar conclusion to that from Fitted values vs. Residuals Plot is obtained. That is, the variance of our data is not constant.

3.5.2 Normality Assumption

Histogram is a common method to check normality assumption. Figure 4 shows the histograms for residuals. It is observed that the histogram does have a rough bell shape, but has a very sharp peak. Thus, the histogram is insufficient to check normality assumption, and more powerful tools should be used.

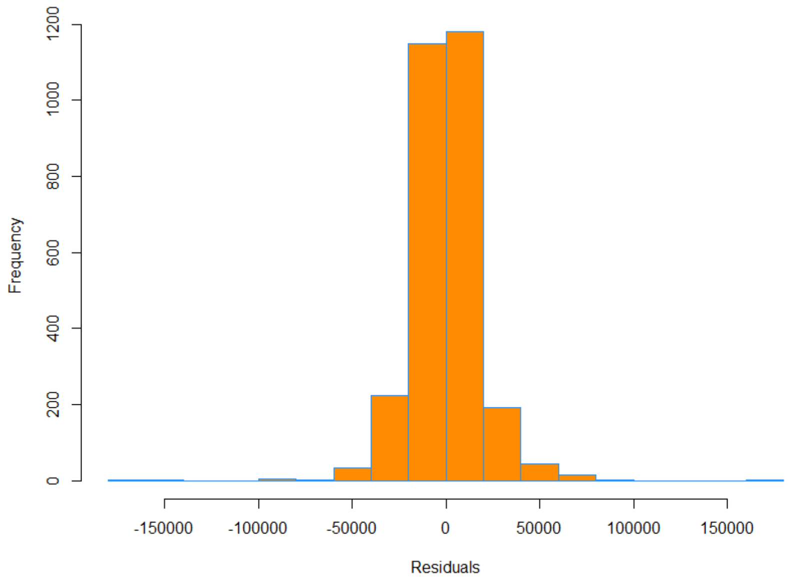


Figure 4. Histogram

Then we generates a Q-Q plot and also perform Shapiro-Wilk test to assess the normality of errors. As Figure 5 shows, the points in Q-Q plot do not perfectly follow the straight line which represents normal distribution, suggesting that the errors may not follow a normal distribution. Similar conclusion is obtained from Shapiro-Wilk test. The small p-value means that the null hypothesis, normality assumption, should be rejected. In other words, there is only a small probability that the data is sampled from a normal distribution.

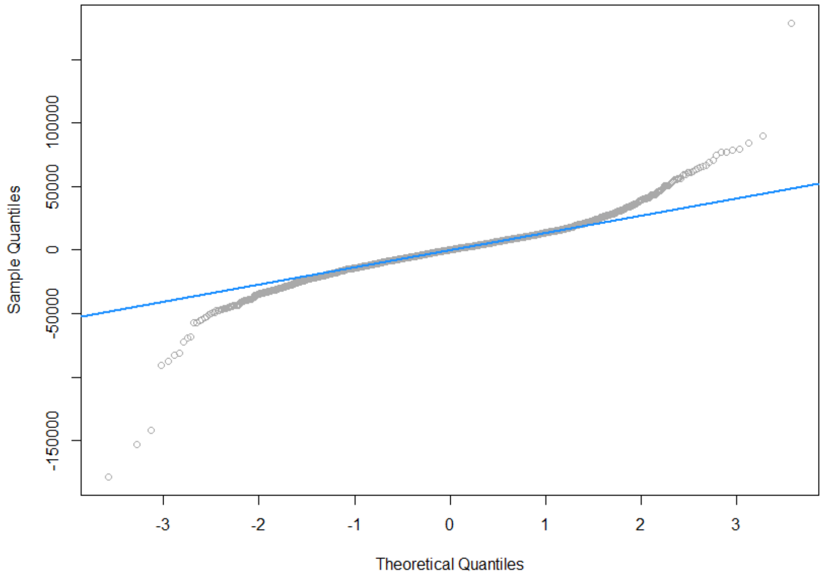


Figure 5. Q-Q Plot

4. Visualization Findings

In this part, we use Tableau to create a user interface based on a subset of variables we selected in the prediction model. Visualization helps us to explore data more intuitionally and demonstrates trend explicitly.

We use simple scatter plot for individual numerical variable analysis, and matrix scatter plot for one numerical variable and one categorical variable. We also do time-series data plot to figure out the sale price trend when time goes by. Furthermore, mosaic plot is used to show the differences of sale price based on multiple categorical variables.

We create multiple tab boxes on the bottom of the dashboard, so it makes it easy to switch plots showing subcategory of the data. We also include slider on the right side, which makes it possible to select the range of the data for the graphical display.

4.1 Sale Price vs. Year Built

Figure 6 shows the house sale prices over year. The data range in the left one is 1872-2010 using the whole dataset, while the data range in the right one is 1950-2010. Judging from the plot, sale price generally increases when time goes by, but the slope of the right plot is steeper than the left one, which indicates that the sale price increases more in latest years than past. Moreover, based on the trend line and forecasting, we have 95% confidence that the average sale price will stay stable around $260,000 in next five years.

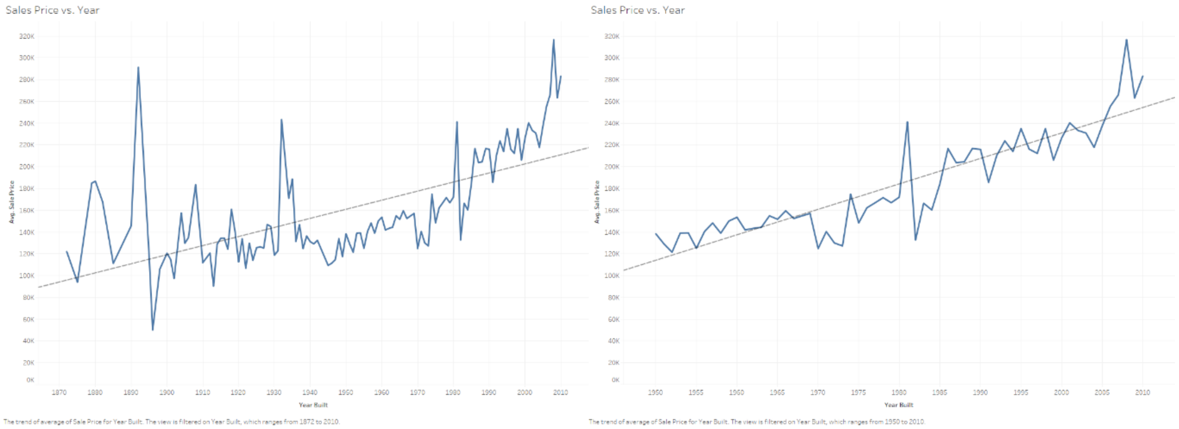


Figure 6. Sale Price vs. Year Built. Left: from 1872 to 2010; Right: from 1950 to 2010.

4.2 Sale Price vs. First Floor Area

The relationship between sale price and a numerical predictor, first floor area is investigated. Figure 7 suggests that there may be a linear relationship between sale price and first floor area. We also add the trend line, whose p-value is less than 0.0001 and R-square equals to 0.405. Therefore, first floor area is possibly a significant variable for sale price prediction model.

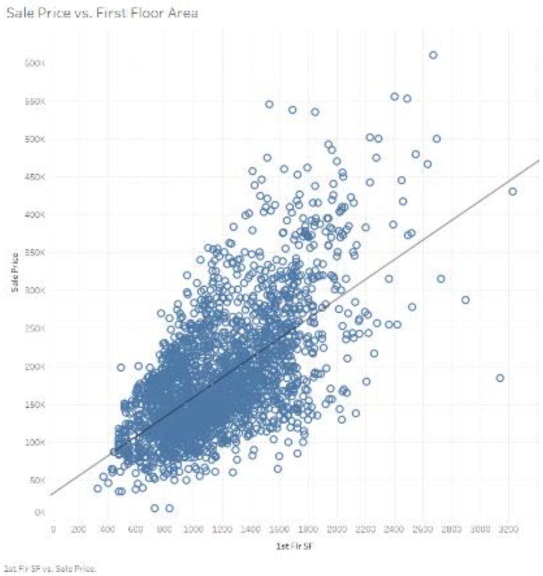


Figure 7. Sale Price vs. First Floor Area

4.3 Sale Price vs. Neighborhood

A box plot related to sale price and a categorical predictor, neighborhood is shown in Figure 8. From the figure, we can easily see the median, upper/lower hinge, and upper/lower whisker for each neighborhood. It's obvious that the sale price and its distribution in different neighborhood are different, so we suppose that neighborhood is a significant variable for sale price analysis. In addition, NridgHt neighborhood is noticed with the highest sale price, while the BrDale has the lowest sale price. Also, the sale price in NirdgHt is the most sensitive to first floor area. Therefore, we argue that BrDale is the most economical location for buying house.

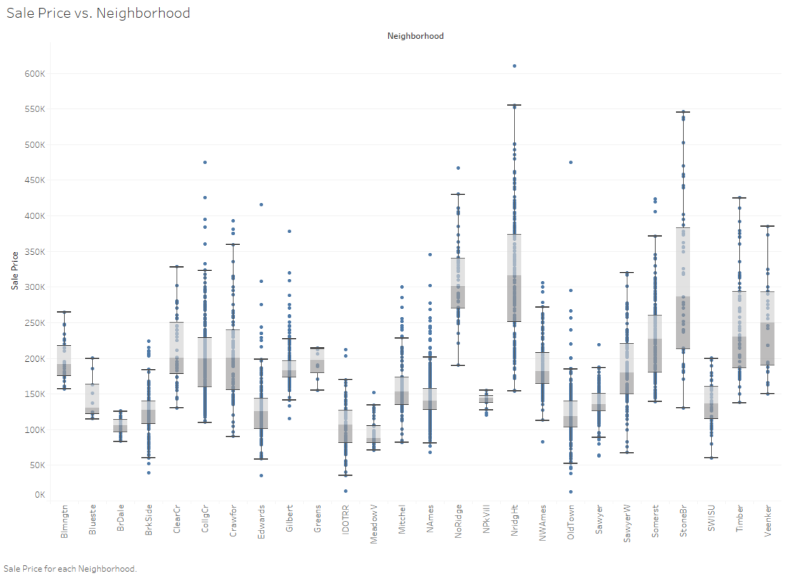
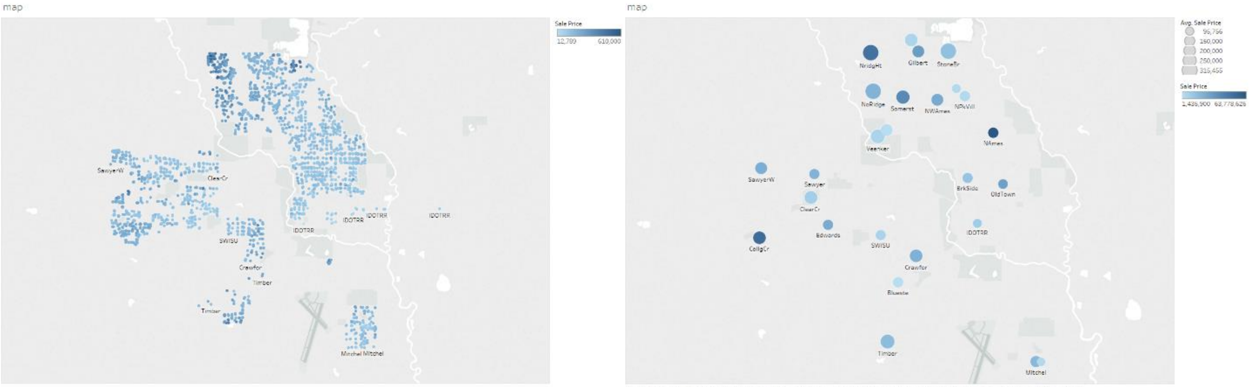


Figure 8. Sale Price vs. Neighborhood

A similar conclusion can be obtained from Figure 9, which is related to sale price and latitude and longitude. These two figures demonstrate the geo-mapping layout for sale price. The left one shows all the data points we have, and the shade of color indicates the magnitude of sale price. To make the figure more clearly, the right figure shows the average sale price in each neighborhood, both the color and the size of the points are based on the magnitude of sale price. In Figure 9, we can also find the locations with the highest and lowest average sale price.



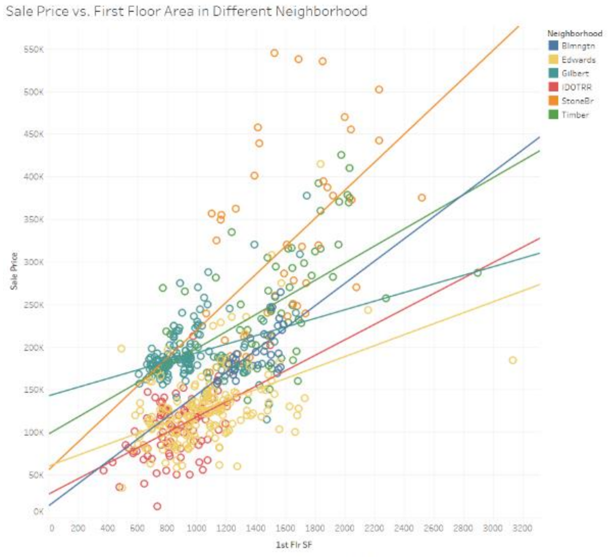
Map based on Longitude and Latitude. Color shows Sale Price. The marks are labeled by Neighborhood.

Map based on average of Longitude and average of Latitude. Color shows sum of Sale Price. Size shows average of Sale Price. The marks are labeled by Neighborhood.

Figure 9. Sale Price vs. Latitude/Longitude

4.4 Sale Price vs. First Floor Area/Neighborhood

To avoid showing all the data which may be messy in plot, we choose a subset randomly from neighborhood. Figure 10 indicates that in different neighborhood, the linear relationship between sale price and first floor area might be different. Therefore, we could include both first floor area and neighborhood in an additive model for sale price prediction.

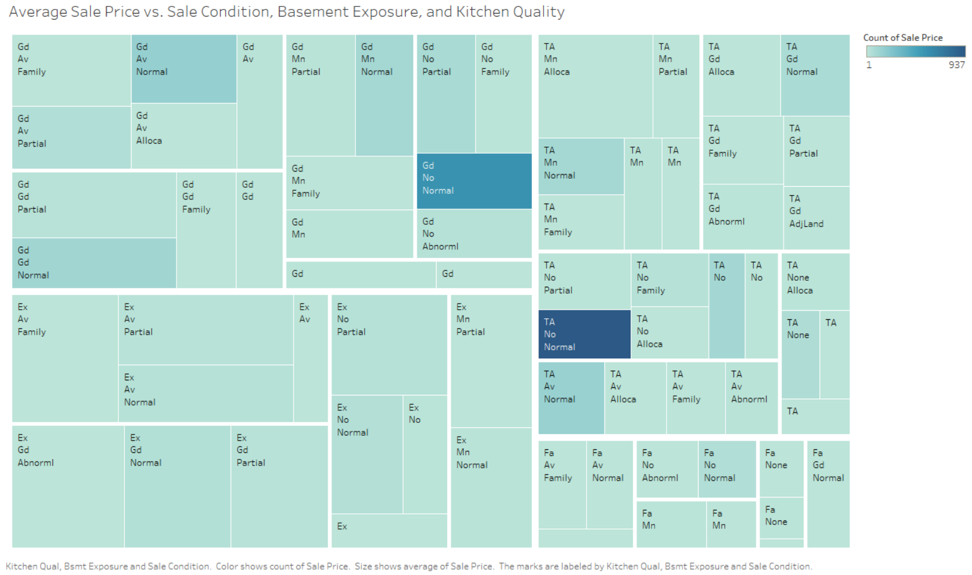


1st Flr SF vs. Sale Price. Color shows details about Neighborhood. The view is filtered on Neighborhood, which keeps 6 of 26 members.

Figure 10. Sale Price vs. First Floor Area/Neighborhood

4.5 Sale Price vs. Sale Condition/Basement Exposure/Kitchen Quality

Figure 11 is a Mosaic plot for multiple categorical variables. Our model chooses 11 significant categorical variables, and here we only pick three of them for demonstration. This figure depicts the sale price under different sale condition, basement exposure and kitchen quality, the size of each square is proportional to the average sale price, while the shade of color is based on the number of observations. We notice that there are limited numbers of house available in different conditions, and most houses have kitchen quality as TA, sale condition as Normal with no basement exposure. The houses with excellent (Ex) kitchen, Abnormal sale condition and good (Gd) basement exposure have the highest average sale price, and it will be cheaper if we could tradeoff between basement and kitchen when selecting house.



Kitchen Qual, Bsmt Exposure and Sale Condition. Color shows count of Sale Price. Size shows average of Sale Price. The marks are labeled by Kitchen Qual, Bsmt Exposure and Sale Condition.

Figure 11. Sale Price vs. Sale Condition/Basement Exposure/Kitchen Quality

5. Conclusion

After performing collinearity on variables, Boxcox transformation on response, variable selection by both AIC and BIC methods, model selection by ANOVA, as well as diagnostic and assumption checking, we select AIC forward model which contains 22 numerical variables and 27 categorical variables to predict sale price. In the six models by AIC and BIC, there are 10 common numerical variables and 11 common categorical variables, where the variables used for visualization analysis is selected. In visualization, we explore the relationship between sale price and year built, first floor area, neighborhood, latitude/longitude, as well as sale condition/basement exposure/kitchen quality respectively. Three interesting findings are listed as follows.

1) The sale price of house increases more rapidly in latest years than past. And based on the trend line and forecasting, we have 95% confidence that the average sale price will stay stable around $260,000 in next five years.
2) NridgHt neighborhood is noticed with the highest sale price, while the BrDale has the lowest sale price. Also, the sale price in NirdgHt is the most sensitive to first floor area. Therefore, we suppose that BrDale is the most economical location for buying house.
3) The numbers of house available in different conditions are limited, and most houses have kitchen quality as TA, sale condition as Normal with no basement exposure. The houses with excellent (Ex) kitchen, Abnormal sale condition and good (Gd) basement exposure are noticed with the highest average sale price. Therefore, we can save money if we could tradeoff between basement and kitchen when selecting house.

Reference

[1] D. De Cock, "Ames, Iowa: Alternative to the Boston Housing Data as an End of Semester Regression Project", *Journal of Statistics Education*, vol. 19, no. 3, 2011.