Tam Nguyen
Applied Statistical Methods

Exploratory Data Analysis

I. *General information about Ames Housing dataset*

The Ames Housing dataset was compiled by Dean De Cock for data science education. It describes the sale of individual residential property in Ames, Iowa from 2006 to 2010. The data set contains 2930 observations and a large number of categorical variables (23 nominal, 23 ordinals, 14 discrete, and 20 continuous) involved in home value assessments.

The purpose of this project is to predict the housing sale prices based on the group of most important predictors. Hence, detecting any strange relationship between explanatory variables and response variable if possible. Multiple linear regression is believed to be a good model in this project.

II. *Cleaning dataset:*

By reading the data description provided (Cock 2011) to better understand the information represented in the dataset, there are many data was labeled as "NA" to indicate absence or null value for a majority of the predictor variables.

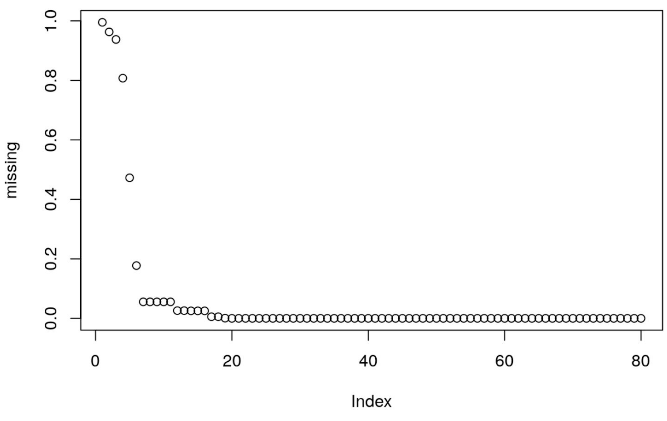
The explanatories include four types of variables which are nominal, ordinal, discrete, and continuous. However, there are some exception to this explanation was found in the case of; LotFrontage (where there were no explicit mention of NA); MasVnrType (where an explicit None level was defined); Electrical (where no explicit mention of NA was provided). This indicates that a majority of NA in the data set are missing not at random and may be replaced by reasonable choices. Of the 80 variables in the data set, nineteen (19) had missing values and eighteen (18) had missing values above five (5) percent. The analysis that follows utilized these R libraries to assist in the data management and model selection phase.

III. Preliminary feature selection

a. Missing value: (deleted 4 variables)

As mentioned above, there are many missing value in this dataset. I decided to delete variables that missing more than 80% of their total value.

The plot show that there are nineteen of the original 80 variables have some degree of missing value. Variables that I deleted are PoolQC(pool quality), MiscFeature(Miscellaneous feature not covered in other categories), Alley(Type of alley access to property), and Fence(Fence quality).



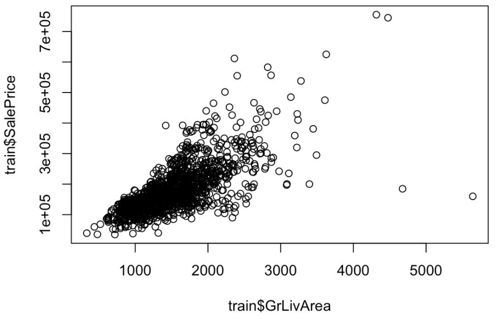
b. Low variance predictors (deleted 8 variables)

Removing low variance predictors is an important step in cleaning this dataset. According to Max Kuhn(2013), a rule for detecting near-zero variance predictors is: the fraction of unique values over the sample size is low (say 10 %); and the ratio of the frequency of the most prevalent value to the frequency of the second most prevalent value is large (say around 20).

c. Multicollinearity (deleted 3 variables)

It is also possible to have relationships between multiple predictors at once (called multicollinearity). Hence, after detecting multicollinearity predictors by Max Kuhl (2013), I delete all variables that have pairwise correlations less than 0.75.

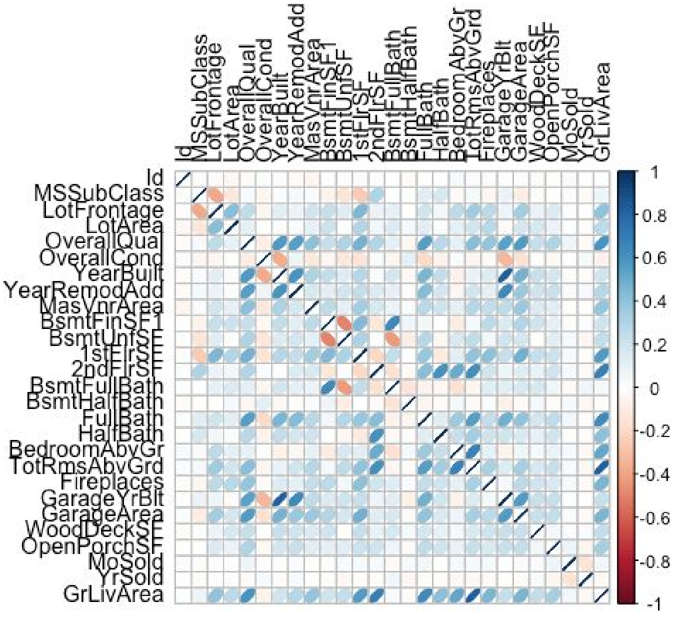
His algorithm suggests removing 3 predictor variables which are GarageCars (Size of garage in car capacity), GrLivArea (Above grade (ground) living area square feet), and TotalBsmtSF(Total square feet of basement area). However, I still keep the above grade living area as one of predictors because I will delete Total basement area variable and other sub-area variable, which will be reduce the multicollinearity of grade living area variable. We can look at the strong linear relationship between GrLivArea with Sale Price in the graph below:



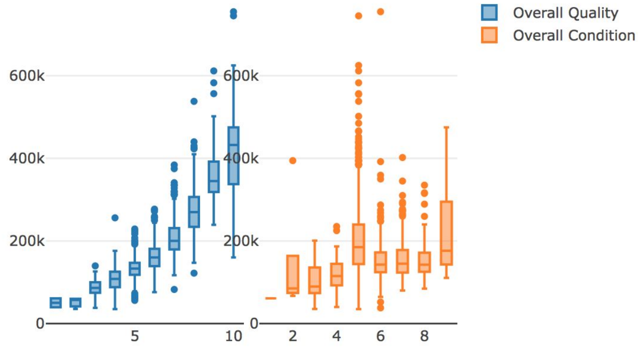
d. Intuitional deletion

After finding level of missing value for each predictor, low variance predictors, and multicollinearity, I still have 65 predictors to investigate. Then, I used my own intuition about what is important, and recognize that homeowner cares about lot size, overall quality, utilities, location, garage (or not), pool (or not) more than fireplace, or garage year built.

The figure belows is the correlation matrix for all continuous predictors after we remove low variance and multicollinear explanatory variables.



A closer look at the matrix plot of the predictor vectors clearly demonstrates that there remain predictor variables that may be better classified as factor variables. Half Bath, Fireplaces and BsmtHalfBath are examples of such predictors.



It comes to my attention while exploring the relationship between Overall Quality and Overall Condition to Sale Price.While the Overall Quality correlation with the Price is monotonic, the Overall Condition correlation is less clear. Those are two predictors that I believe are significant. Hence, I might threat ordinal variables as categorical or do some transformation in the future model.

The group of continuous predictor variables that I select in my future model are: Lot Area, Overall Quality, Overall Condition, Year Modition date, Total Room above ground. Otherwise, I classified Garage and its related variables, Bathroom and its related variables, Wood Deck and Open Porch, and Fire and its related variables as factor variables to answer “Yes or No” question. For example, if the house has Fireplace, I will encond as 1, otherwise 0.

When it comes to buy a house, I believe that location is extremely important. Even when driving is not a problem for American, the necessity around house like railroad, supermarket, school, and living standard is one of the priorities of homeowner.

In my future model, by encoding the same method as mentioned above, I have 26 variables to further investigate in future model. More description about these predictors and how I encoded are included in the Annotated Appendix.

IV. *Methodology for Multiple Linear Regression Model Creation*

Multiple linear regression model will be run with Sale Price as the response variable, and other 31 variables will be the explanatory and factor variables. In further investigation, there might be some potential interaction terms with the variables of interest to reduce the number of variables, and keep the model concentrate and simple. Lot Size and Overall Quality/Overall Condition are three variable that I believe will have a significant impact in the future model P-values for the Chi-sq test will be used to determine whether there is an association between the two variables. R-square is the criteria to choose which is the best model.