**Viktoryia Strylets Lab5 Report**

Introduction:

My data set is the sample of homes from the Ames home sale data that includes 2930 property sales from 2006 to 2010 in Ames, Iowa. In order to build a model that would predict the sales prices of the typical home in Ames, Iowa, during 2006 and 2010 I needed to familiarize with data to detect potential issues that might affect the prediction ability of my model and choose the variables that might have a great power in forecasting SalePrice of Ames houses.

I will compare several visualization techniques for high-D data and will analyze the interesting findings after exploring the visualizations.

I will define my sample of “typical” houses for Ames, Iowa. Bivariate analysis should give me idea of multicollinearity among factor variables. The PCA and biplot visualizations will help to detect the two attributes that might be strong predictors of house values in Ames, Iowa, for period from 2006 to 2010.

Section 1: Sample Definition

I obtained descriptive summary statistics from mydata such as distributional implications including potential skewness and outliers which helped me to define the sample. The examination of summary statistics indicated that the majority of houses locate in the residential area, namely 2273 in Residential Low Density (Rl), 462 in Residential Medium Density and 139 in Floating Village Residential areas with few houses in Residential High Density(RH) zone. Based on this observation, I concluded that the typical zoning type of Ames houses is residential and dropped observations that belong to Industrial, Commercial and Agriculture zones. Moreover, after inspection SalePrice outliers for all zoning types, I discovered that Floating Village Residential zone produces a disproportionate number of Saleprice outliers which also justifies imputation of that zone (5 out of 137 SalePrice outliers belong to Floating Village Residential zone).

Single family houses prevail the population, 2425 out of 2930 observations belong to single family building type, which makes this type of dwelling typical for Ames and justifies dropping houses of other building types. Moreover, the second largest group of Town End Unit building type with 233 observations has a disproportionate presence in SalePrice outliers (7 out of 137).

The majority of houses have two types of roof which are Gable (2321) and Hip (551) suggesting my sample of “typical” Ames houses to include only observations with those two types of roofing. However, since many SalePrice outliers have Hip type roofs, out of 137 SalePrice outliers 88 are of Hip roofing type, the sample might consist only of Gable roof style houses.

The SaleCondition column’ inspection revealed that the largest number of houses have normal sale condition (2413 out of 2930) and in addition many SalePrice outliers have partial sale condition which rationalizes the exclusion of all houses from the sample with sale condition other than normal. Out of 137 SalePrice outliers 64 have non-normal sale condition and out of those 64 houses 58 have partial sale condition.

I also inspected Roof Matl column and detected that houses with Standard (Composite) Shingle

roofs are prevalent in population (1914 out of 1943 observations) suggesting that a “typical” Ames house has Standard (Composite) Shingle roofing.

As a result of above-mentioned data transformation, houses with partial utilities availability were automatically eliminated from the sample and the majority of houses are revealed to be heated by GasA type of heating (1504 out of 1524) with only few observations having other type of heating, GasW - 12 observations, Grav - 5, Other - 2, Wall -1, Floor - 0). Three GasW heated houses belong to Saleprice outliers which justifies including only GasA types of houses in the sample of “typical” IOWA houses.

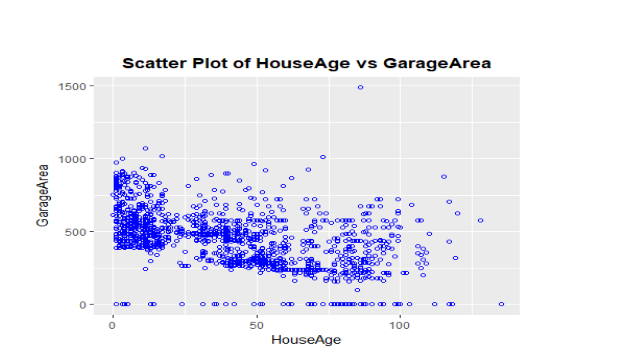
In order for my sample data to be representative of the population that I want to model I identified the scope of the model to the ‘single family’ homes with ‘normal’ sales that located in residential area, have all public utilities, Gas forced warm air furnace type of heating and Gable style roof made out of Standard (Composite) Shingle. My sample has 87 variables and 1504 observations. Out of 137 SalePrice outliers only 46 have left in my sample.

**Bivariate scatterplot**

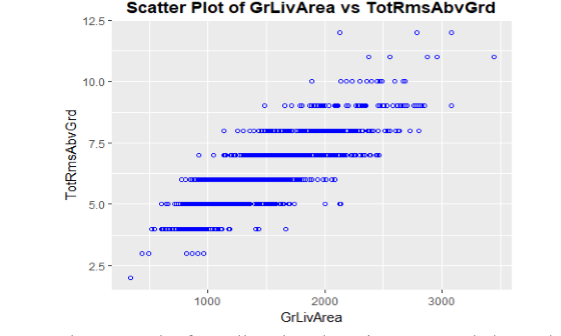
A bivariate scatter plot is very useful in visual representation of the relationship between two numerical variables (attributes). The resulting pattern shows the type (linear or non-linear) and strength of the relationship between two variables. More information can be added to a two-dimensional scatter plot, for example, we might label points with a code to indicate the level of a third variable. The downside of such a technique is it fails to visually represent the relationship between all the variables or visualize the data that have more than three dimensions. If we are dealing with many variables in a data set, a way of presenting all possible scatter plots of two variables at a time is in a *scatter plot matrix*. To encode the data with more than three dimensions we need to use other visual techniques which discussed later. However, the major cause of problems with scatterplots is discretization of values. This happens when decimal places are rounded off, measurements are not accurate enough, or a data field is categorical.

Interesting findings:

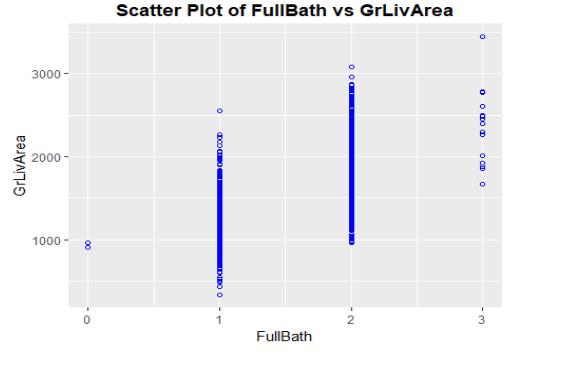
The scatterplot of GarageArea and Houseage revealed a slight to moderate negative correlation . But there is a lot of variation observed in GarageArea variable itself. From the scatterplot is also seen that data points are sort of clustered in three age group, < 25, >25<75 and >75 which suggests creating less categories for HouseAge variable and, thus potentially reduce variation.



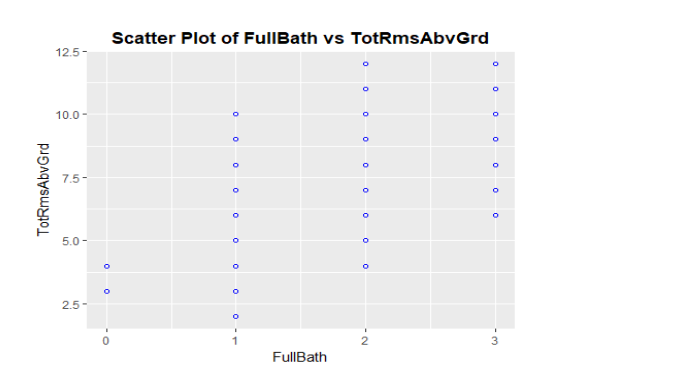
2. The scatterplot for GrLivArea and TotRmsAbvGrd revealed a strong positive correlation between variables.



3. The scatterplot for Full Bath and GrLivArea revealed a moderate positive correlation.

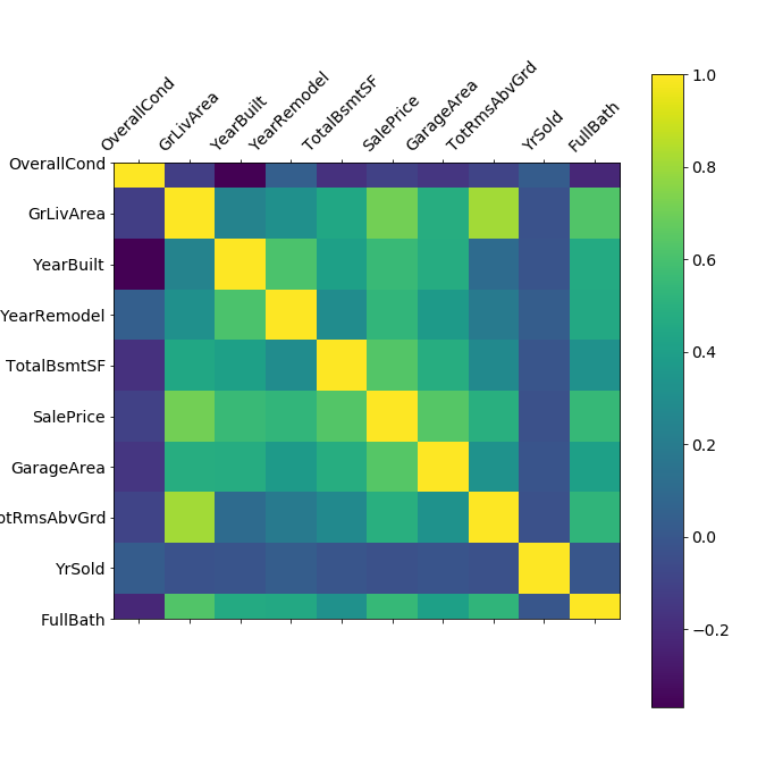


4. The scatterplot of Full Bath vs TotRmsAbvGrd showed some slight to moderate positive correlation between variables. Interesting observation is that outliers with no FullBath have not less than 3 rooms.



**Correlation matrix:**

A correlation matrix is a very useful tool to show correlation coefficients between many variables. Each cell in the table shows the correlation between two variables. A correlation matrix is used to summarize data, as an input into a more advanced analysis, and as a diagnostic for advanced analyses. So it is useful in Initial Exploratory Data Analysis. However this technique fails to show other very important data anomalies like outliers (which can be a very useful in the data analysis),

****

Here some interesting findings:

1. FullBath with cor.coef. 0.654 indicating relatively strong positive correlation between Ful lBath and SalePrice

2. TotalBsmtSF with cor.coef. 0.579 indicating moderate positive correlation between Total BsmtSF and SalePrice

3. GrLivArea with cor.coef. 0.783 indicating relatively strong positive correlation between GrLivArea and SalePrice

4. GarageArea with cor.coef. 0.65 indicating relatively strong positive correlation between GarageArea and SalePrice

5. TotRmsAbvGrd with cor.coef. 0.618 indicating relatively strong positive correlation betw een TotRmsAbvGrd and SalePrice