**UNIVERSITY OF TEXAS AT ARLINGTON   
   
2228-BSTAT-5325-002   
   
Advanced Methods for Analytics**

**Professor: Nabil Raad   
   
BSTAT 5325 Course Project: Fall 2022  
AMES House Price Prediction and Analysis**



**Team 4**

|  |  |
| --- | --- |
| Hitesh Katari | 1002071317 |
| Rahul Kumar | 1002082638 |
| Safoora Naureen | 1002050531 |
| Liveesh Roy Ravipati | 1002037516 |
| Harish Sainath | 1002077949 |
| Susmitha Pattanshetty Shekhar | 1002026230 |

**AMES House Price Prediction and Analysis**

**Business Problem:**

The data analysis team has to come up with a reliable and significant model to *predict the house prices* in the city of Ames, Iowa for the real estate company named: Ames Real Estate Associates (AREA). In addition to predicting house prices with good accuracy, the data analysis team must also provide the real estate company with the *factors which are much likely to influence the prices* of houses. The *magnitude* by which the selected factors are influencing the house price is also to be determined so as to invest proportionately for improving the house quality. Also, the *prime time* at which the house is to be sold must also be determined to get the best price for it. These factors would then be advised by the AREA team to its customers to get a better price for the homes they are planning to sell.

**Motivation:**

After COVID 19, many people have become financially literate. Everyone is trying to look for investment opportunities. House and land areas are a great and sustainable form of investment in the highly developing world scenario. Thus, we have decided to know the current market trend for housing prices specifically to Ames area, unlike the online house pricing websites which are trained on larger areas for a wider customer group.

**Assumptions and Limitations:**

The data provided by AREA is a list of houses sold from 2006 to 2010. The house price prediction done after analysis will only be limited to the houses sold in that particular time period. Since information was gathered on home sales that took place in Ames, Iowa between 2006 and 2010, this research is an observational one rather than a control randomized one. There is no possible way to establish a causal connection between the explanatory factors and the response variable (SalePrice), just a correlation. Also, the house range for which the predictions can be done is only from the cheapest and the costliest house sold in the list. Any predictions outside the aforementioned range cannot be done as the algorithm has not seen features with such houses previously. It is assumed that all the houses sold in the Ames area have been covered by the dataset provided to avoid imbalanced data aspects during analysis.

**Business Questions the analysis will attempt to answer:**

* What is the expected selling price of a house with particular features?
* Which factors influence the price of a house?
* What is the magnitude of their influence?
* When is the best time to sell, buy a house?
* How much should I invest in improving the condition of my home in order to increase the expected price by more than the cost of improvements?
* Which homes should I compare my house to?
* Which neighbourhoods have houses with highest selling price?
* Are old homes costlier than new ones for their antiquity?

**Data Preparation, Exploration, and Understanding:**

***Data Description:***

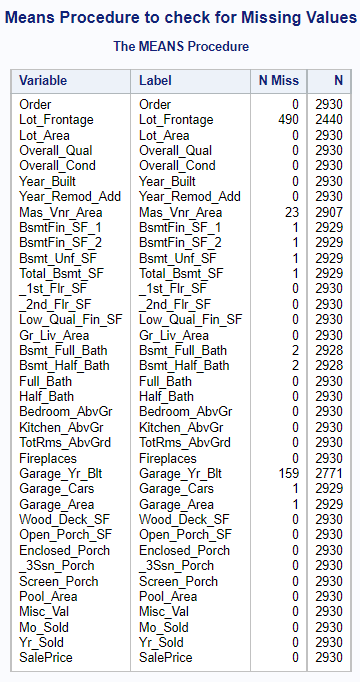
The Ames House Price Dataset consists of 81 columns which are different features of the house like Lot Shape, Number of bedrooms etc, and one target variable as Sale Price for 2930 houses. The distribution of the attributes depending on their type is as follows.

|  |  |
| --- | --- |
| Variable Type | *Count* |
| Discrete variables | 14 |
| Continuous variables | 20 |
| Nominal variables | 23 |
| Ordinal variables | 23 |

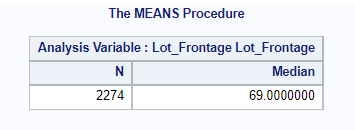
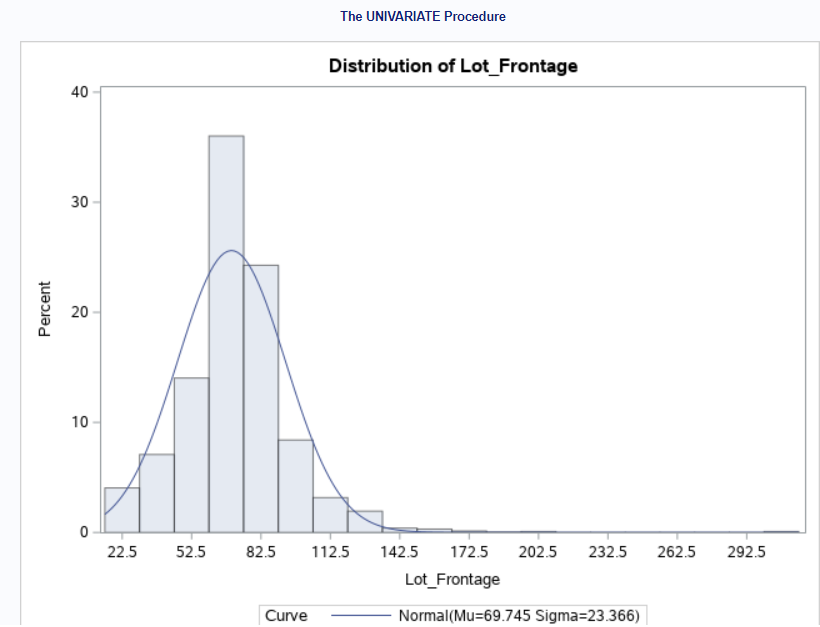
The analysis for the project is done in SAS Studio. The observations and their distributions are first viewed using proc contents in SAS. As the PID column gives just the individual denotation of a house, it is not much of any statistical significance. Thus, the column is dropped.

***Handling Missing Values:***

Statistical analyses are hampered by missing data. The findings of statistical tests will be biased if missingness is ignored as it might be associated with the outcome of interest. Hence, they are to be handled with appropriate measures. SAS reveals that some columns like Mas\_Vnr\_Area, BsmtFin\_SF\_1, BsmtFin\_SF\_2, Bsmt\_Unf\_SF, Total\_Bsmt\_SF, Bsmt\_Full\_Bath, Bsmt\_Half\_Bath, Garage\_Yr\_Blt, Garage\_Cars, Garage\_Area have missing values.



They mainly account for less than 6% of the entire data in the columns. Hence, the observations can be removed from the dataset. The Lot Frontage column has 490 rows which are missing which account for around 17% of the entire data. This value is too high to be deleted. The correlation of Lot Frontage with other variables is computed using Pearson’s Correlation Coefficient. It shows that there is not much correlation between lot frontage and other variables as the values were far away from 1. It can thus be concluded that the column is **MISSING COMPLETELY AT RANDOM**.

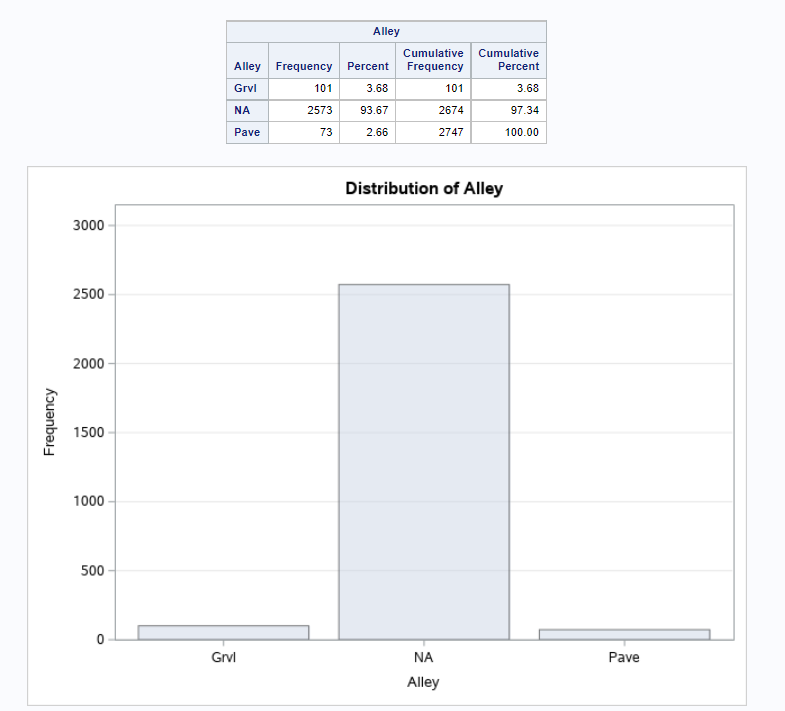
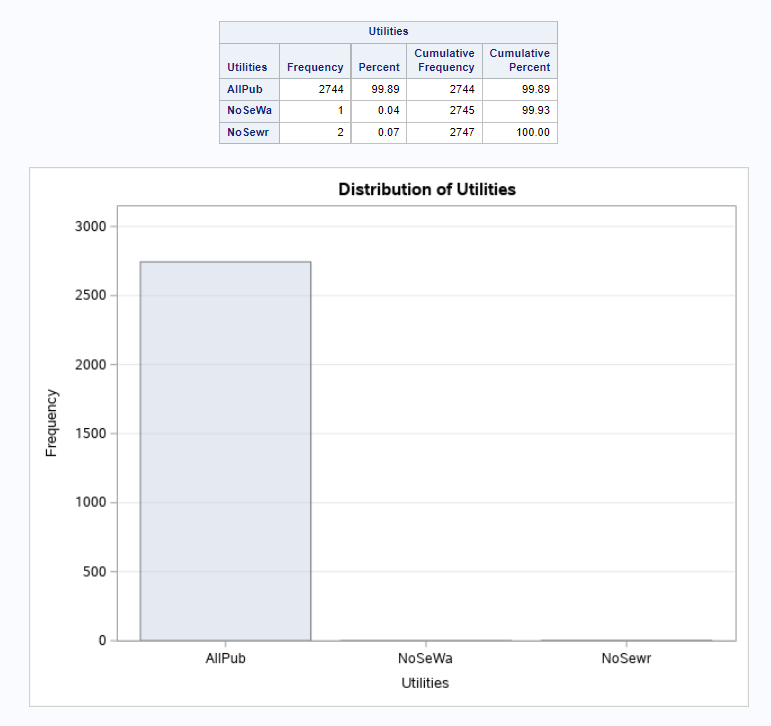


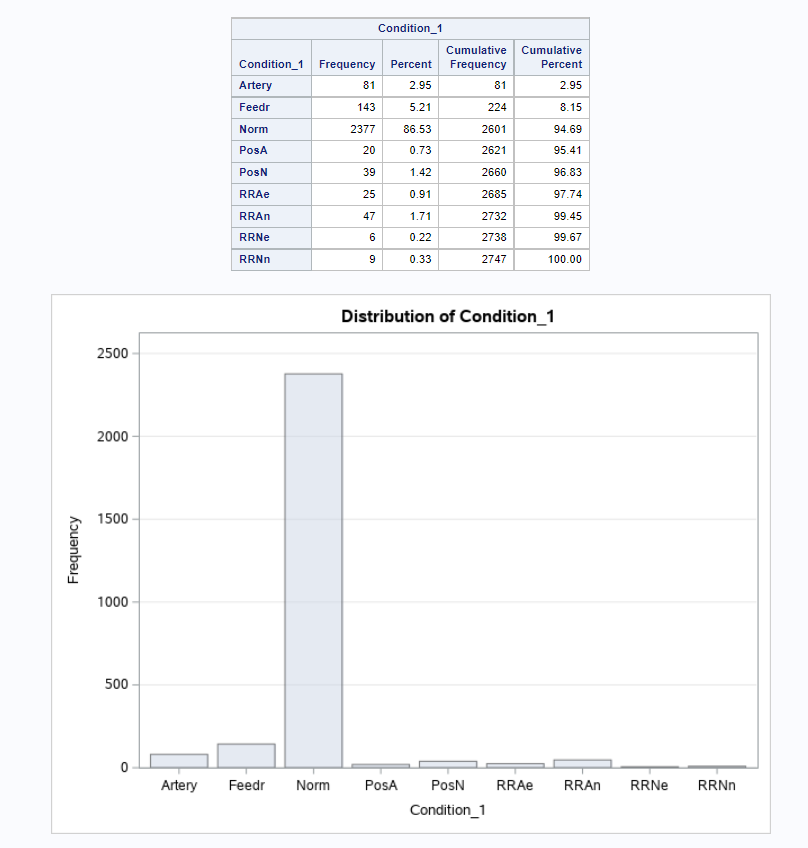
To impute the missingness, its distribution is checked which shows the variable Lot Frontage is skewed to the right. Mean will not be a proper method for imputation here. Median is computed and imputed in the missing values of Lot Frontage.

***Analysis of categorical variables:***

To ensure that the parsimony of the final model is maintained, variables with correlation to the target attribute must be considered. For calculating correlation, categorical variables which can be numerically encoded must be done. Distribution of the categorical variables is extracted using FREQ procedure.

Some columns such as Alley have more than 50% of the values as ‘NA’. Some other columns such as Utilities have one category with more than 90% of all values. This data is highly imbalanced. Hence such attributes are removed from the dataset.

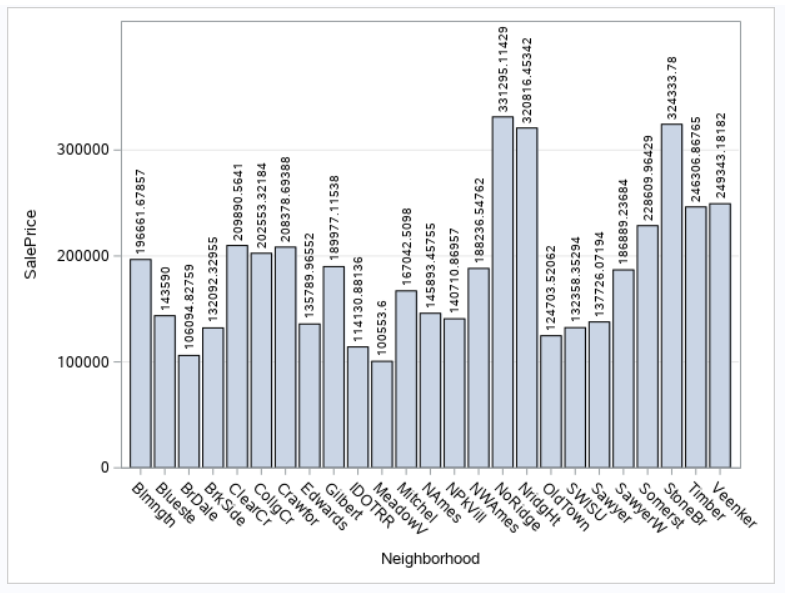


In addition to that, there are a few categories in the column which have a frequency less than 10. This seems very less data in comparision to the 2900 observations. Hence, such records are discarded from the dataset.

After removing less appropriate attributes and observations, encoding of ordinal categorical variables can be done. The encoding in SAS is done keeping in view, the order presented in the documentation for such attributes. The encoded variables are given a suffix ‘\_new’ at the end. After encoding, original ordinal categorical variables are removed.

**Relationships of a few categorical variables with SalePrice:**

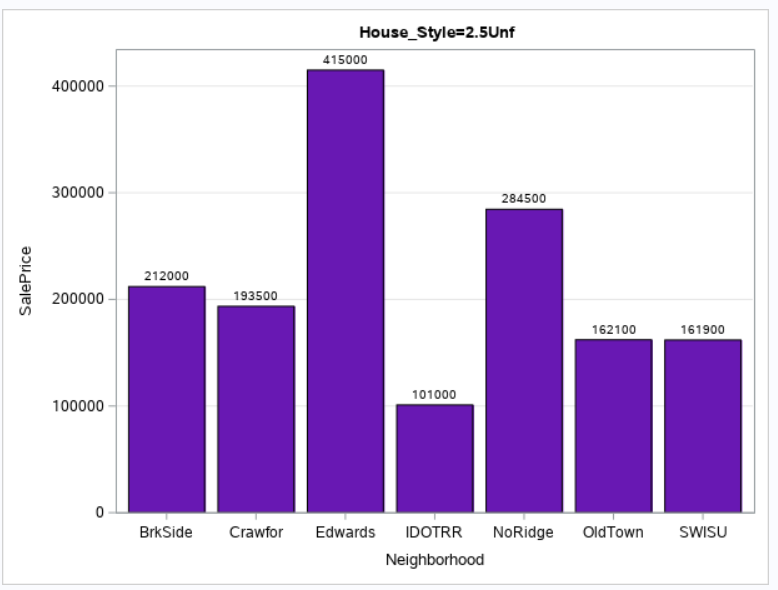
1. Average SalePrice Distribution for each neighborhood is as shown below.



* Which neighbourhoods have houses with highest selling price?

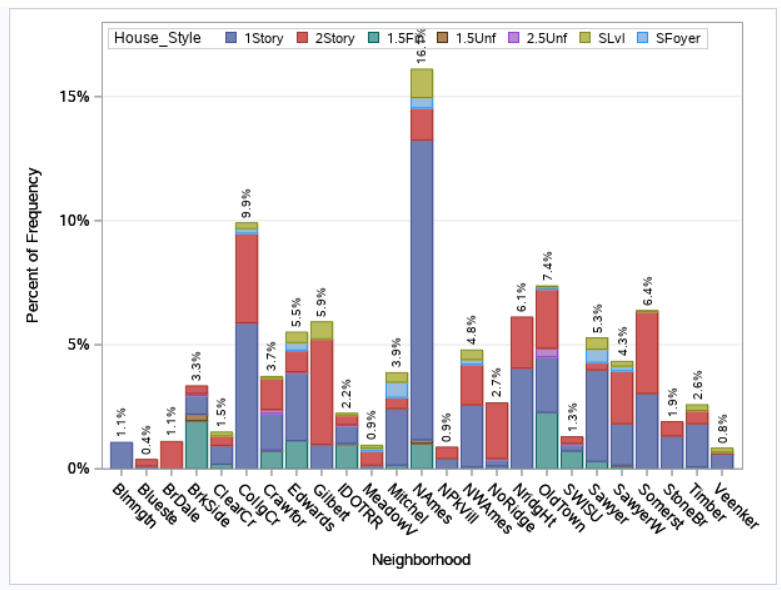
The average Sale Price of houses in NoRidge (North Ridge) and StoneBr (Stone Brook) is highest among all the neighbourhoods.

We can see that the Average SalePrice of homes in Edwards is much lesser than that of NoRidge (North Ridge). But, when we take a closer look by taking House Style into consideration, the average sale price of Edwards for 2.5Unf (Two and one-half story: 2nd level unfinished) is much higher than that of NoRidge.



We can conclude that, comparison of Averge homeprices must be done by mentioning the factors which are being taken into consideration else it would lead to ***Simpson’s Paradox***.

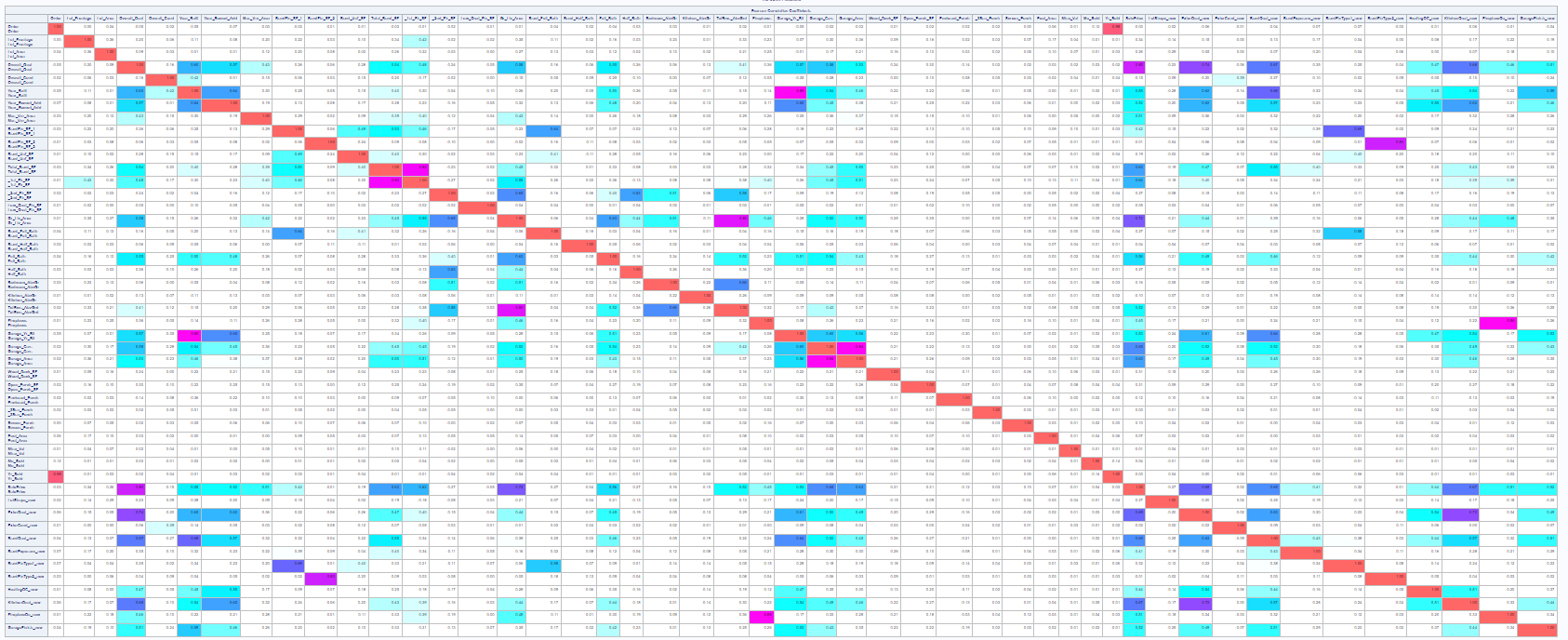
1. Distribution of houses in each neighborhood based on house style shows the distribution as follows



The data seems to be unevenly distributed for NAmes (North Ames) giving a clear example of **Imbalanced Data**.

***Correlation for all variables:***

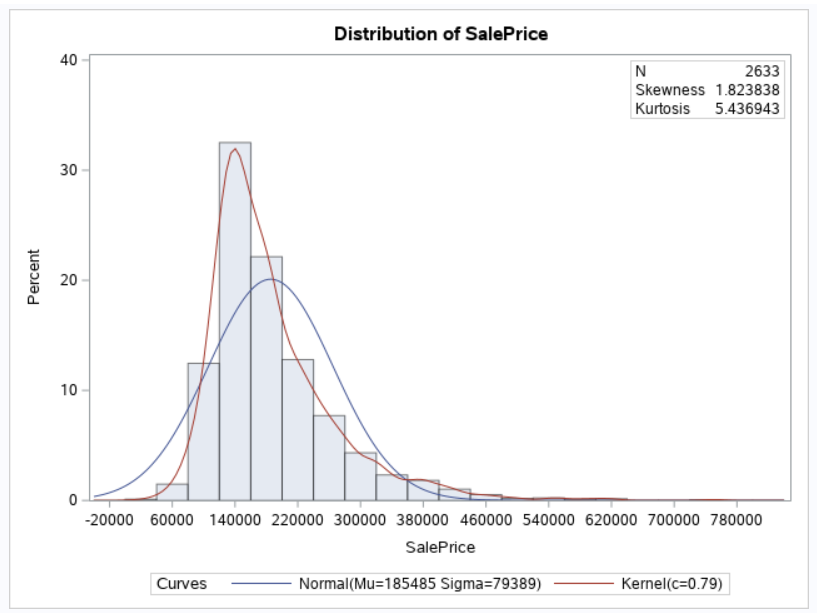
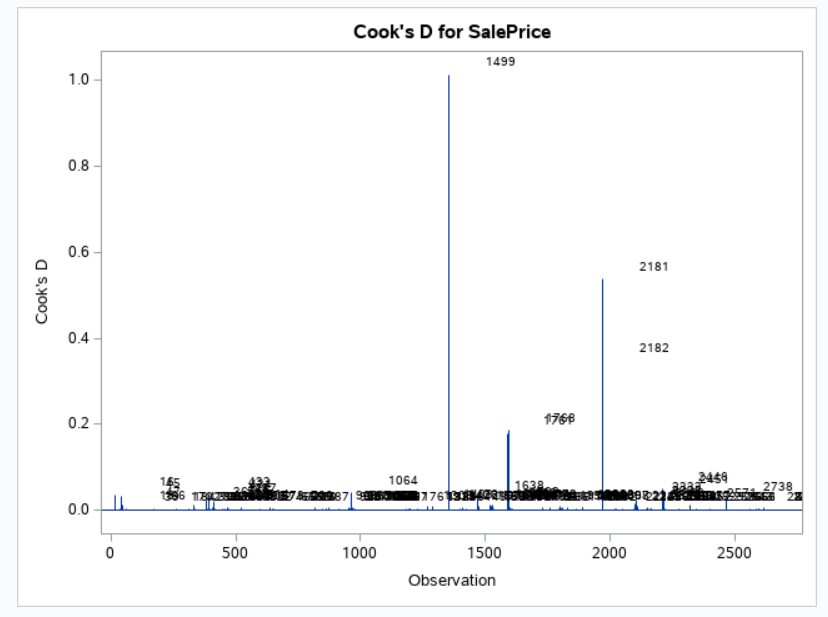
Correlation for all numerical variables is calculated using Pearson’s Correlation Coeffient. The heatmap is plotted with extremely higher or lower values with darker colors as follows.



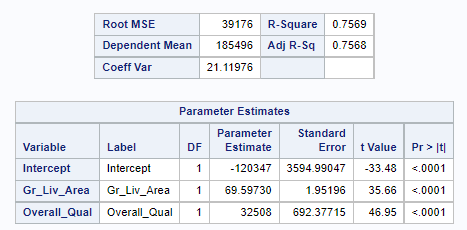
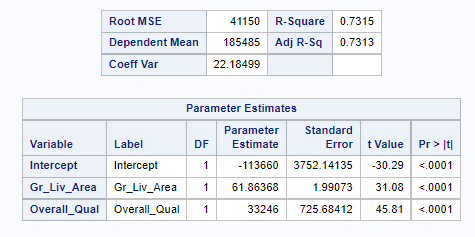
On closer look, SalePrice has greater than 0.7 correlation with Overall\_Qual and Gr\_Liv\_Area. In addition to that, Garage\_Yr\_Blt and Year\_Built, BsmtFinType2\_new and BsmtFin\_SF\_2, Total\_Bsmt\_SF and \_1st\_Flr\_SF, TotRms\_AbvGrd and Gr\_Liv\_Area, FireplaceQu\_new and Fireplaces, Garage\_Area and Garage\_Cars have greater than 0.8 correlation among themselves. One among the pair must be removed to avoid multicollenearity. Thus, variable having less correlation with SalePrice is removed from each pair.

***Analysis of Dependent Variable:***

On measuring the skewness and kurtosis of the Sale Price, we see that kurtosis which needs to be 3 for normal distribution is 5.437 concluding that the distribution is Leptokurtic. It shows that there are outliers in the distribution. The skewness is 1.82 which means the distribution is positively or rightly skewed.

To investigate the outliers from Kurtosis of SalePrice, the global influence measure of Cook’s D is computed. It shows that observation number 1499, 2181, 2182 have high Cook’s D values leading to the Adjusted R-Square value of 0.7313.



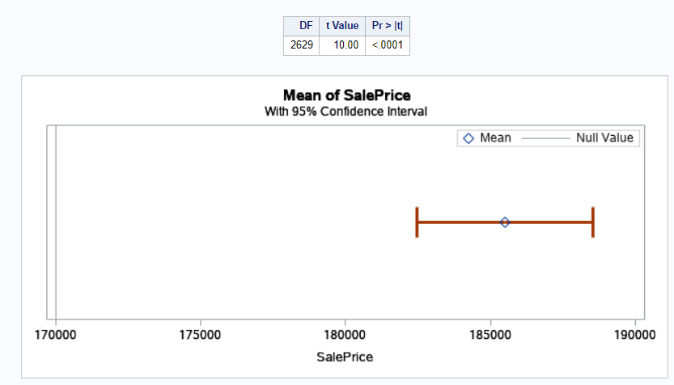
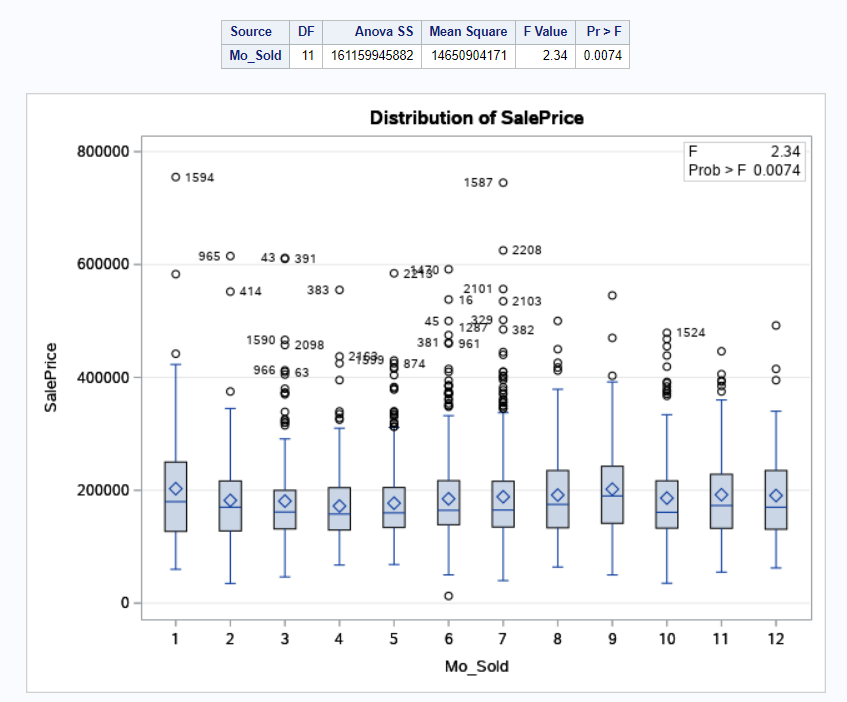
When the three observations are deleted, the Adjusted R-Square value becomes 0.7568.

***Hypothesis Testing:***

1. **Hypothesis testing to check whether the mean of SalePrice of homes is $170000.** With 95% confidence interval, we get p-value <0.0001. Hence we can reject the null hypothesis and accept the alternate hypothesis of mean SalePrice being a value other than 170k USD.
2. **H0: There is no difference between the mean SalePrice of homes sold in different months.**

**Ha: There is a difference between mean SalePrice of homes sold in different months.**

Because our p-value 0.0074 is less than .05, we reject our null hypothesis that there is no difference in the mean SalePrice of homes sold in different months

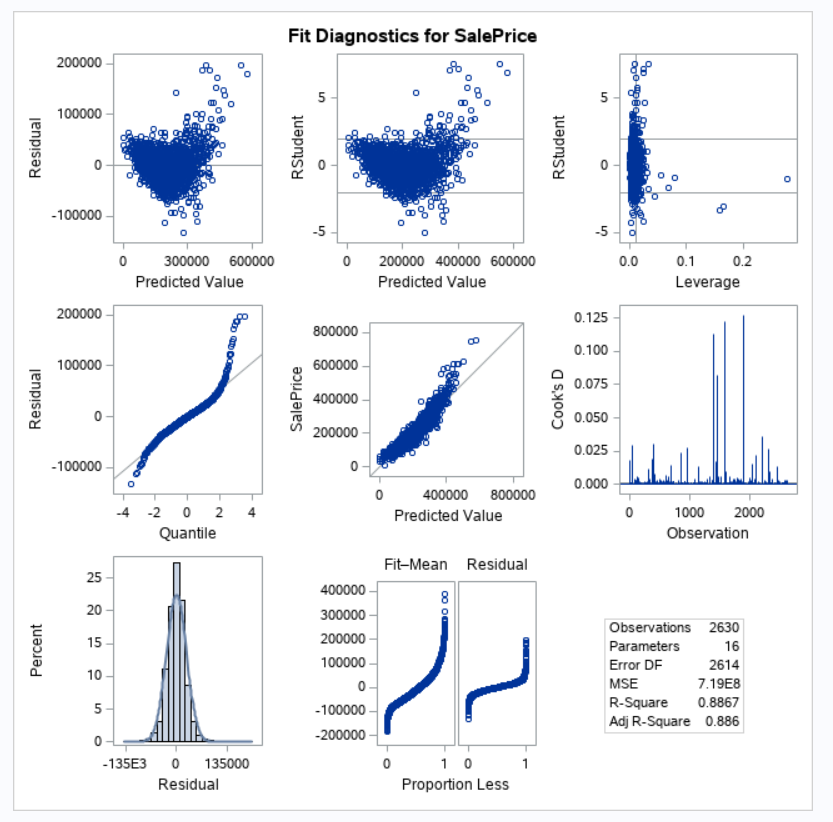
 

* When is the best time to sell, buy a house?

As the mean SalePrice is different for different months, the months with lowest price i.e., April is the time when a customer should buy. Similarly, January is the month when the customer should sell the house.

***Regression Analysis:***

After various analysis of variables and imputations, we are left with 2630 observations and 58 attributes in total with SalePrice being the dependent variable. Stepwise regression analysis is done as it is amodified version of both forward and backward selection adding and deleting variables where the F-Statistic is significant at SLENTRY or SLSTAY level. This gives us with a total of 15 statistically siginificant variables for predicting saleprice on 21st step with an adjusted R-square value of 88.6% and a Root Mean Square Error of 26816 USD.



The fit diagnostics for Sale Price reveal that the Residual is equally distributed on each side of 0 with a few outliers. The R-student values is mostly within the positive and negative ranges. The error e has a normal distribution giving a nearly straight line for the QQ-plot. There is linearity in prediction and a smaller number of Cook’s D values for outliers showing that all the assumptions for Linear Regression are satisfied.

The regression equation containing the intercepts and beta values becomes as follows.

**SalePrice** = -572285 +(11324\***Overall\_Qual**) +(67.62983\* **Gr\_Liv\_Area**) +(22.56078\* **BsmtFin\_SF\_1**) + (14845\* **ExterQual\_new**) + (23.13918\* **Total\_Bsmt\_SF**) +(0.57297\* **Lot\_Area**) +(9480.42707\* **KitchenQual\_new**) +(30.80669\* **Mas\_Vnr\_Area**) +(3898.20424\* **BsmtExposure\_new**) +(-23794\* **Kitchen\_AbvGr**) +(7750.13302\* **Garage\_Cars**) +(5508.71253\* **Overall\_Cond**) +(232.31876\* **Year\_Built**) +(213.26817\* **Lot\_Frontage**) +(-6538.97435\* **Bedroom\_AbvGr**)

* What is the expected selling price of a house?

When a random house’s variables are put into the regression equation, the predicted house price is computed as 173651.2765 USD where the acutal SalePrice was 180900 USD giving an error of 7k USD.

**SalePrice** = -572285 +(11324\*5) +(67.62983\*1629) +(22.56078\*791) +(14845\*2)+ (23.13918\*928)+ (0.57297\*13830)+ (9480.42707\*2)+ (30.80669\*0)+(3898.20424\*1)+(-23794\*1) +(7750.13302\*2) +(5508.71253\*5) +(232.31876\*1997) +(213.26817\*74) +(-6538.97435\*3)

* What factors influence the price of the house and what is the magnitude of their influence?

The factors Overall\_Qual Gr\_Liv\_Area, BsmtFin\_SF\_1, ExterQual\_new, Total\_Bsmt\_SF, Lot\_Area, KitchenQual\_new, Mas\_Vnr\_Area, BsmtExposure\_new, Kitchen\_AbvGr, Garage\_Cars, Overall\_Cond, Year\_Built, Lot\_Frontage, Bedroom\_AbvGr effect the SalePrice in descending order.

* How much should I invest in improving the condition of my home in order to increase the expected price by more than the cost of improvements?

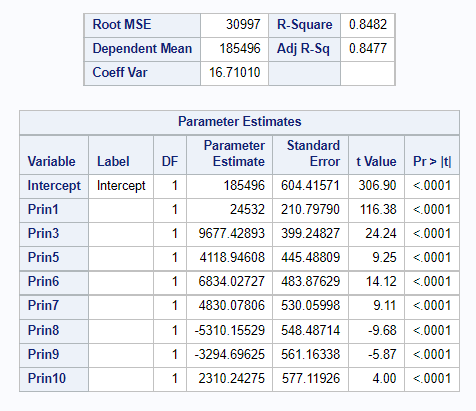
The investment amount on improving the condition of the house should be based on the magnitude by which they influence the Sale Price. The variables with higher influence must be provided higher revenue for improvement.

* Which homes should I compare my house to?

The variable having similar Overall\_Qual, Gr\_Liv\_Area, BsmtFin\_SF\_1, ExterQual\_new, Total\_Bsmt\_SF, Lot\_Area, KitchenQual\_new, Mas\_Vnr\_Area, BsmtExposure\_new, Kitchen\_AbvGr, Garage\_Cars, Overall\_Cond, Year\_Built, Lot\_Frontage, Bedroom\_AbvGr can be used to compare the house.

***Principal Compnent Analysis:***

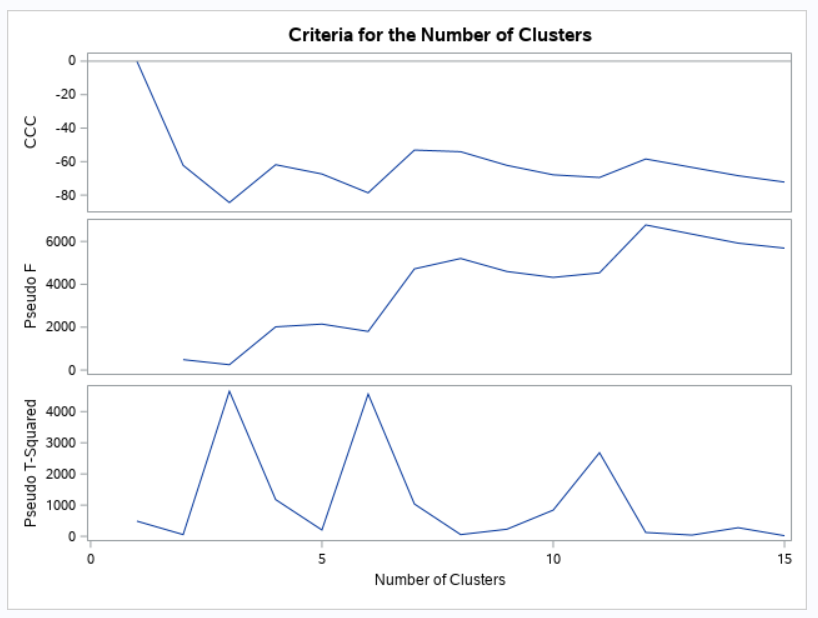
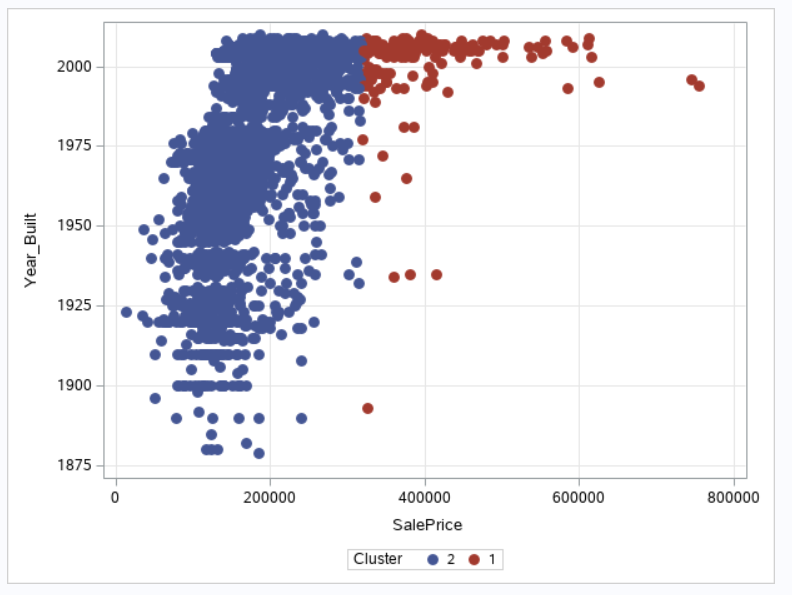
Principal Component Analysis is done for all the 58 attributes after normalizing them with a mean of 0 and standard deviation of 1. It reveals that 12 Principal Components have an Eigen Value greater than 1. When Multiple Regression for those 12 PCs is done it reveals 8 of them: Prin1, Prin3, Prin5, Prin6, Prin7, Prin8, Prin9, Prin10 are statistically signifcant with a p-value < 0.0001. Multiple Regression Analysis on statistically significant PCs finally gives a model with an Adjusted R-square of 84.77%.



***Cluster Analysis:***

We need to find out if the antiquity of houses effects their SalePrice. For that, cluster analysis is to be done on Year Built to the Sale Price of houses. From the CCC plot below, it can be seen that the elbow has dropped at three. Hence, the optimum cluster would be 2. The clusters are then formed using year in which the house was built and its corresponding saleprice.

* Are old homes costlier than new ones for their antiquity?

The graph below shows that there are very few houses which were build long time ago and have a high Sale Price, thus concluding that SalePrice of house is not effected by the antiquity of it.

***Conclusions and Recommendations:***

The analysis of the dataset revealed that there is a lot of variation in average SalePrice of homes among neighborhoods. North Ridge and Stone Brook have the costliest homes whereas BriarDale and Meadow Village have the cheapest houses. The average house price changes with the type of house taken into consideration which explains the uneven distribution of house types in a few neighborhoods. The month of April is best to buy a house whereas January and September are the best months to sell a house.

Sale Price analysis showed that very few houses have higher price and there are outliers in the distribution. Such observations were not taken into consideration. Houses with highest Overall Quality and bigger above ground living area constitute to costlier homes. Efforts can be made in improving the overall quality keeping the exterior part and kitchens in priority for best selling prices. Kitchen above ground and Bedroom above ground are negatively dependent on Sale Price meaning people prefer ground floors.

The year built is positively dependent with Sale Price showing new houses have higher sale values. This is further corroborated in cluster analysis where very few older houses fell in the higher price cluster. Hence, older houses can increase their sale price by improving other factors such as Overall condition including their Garage and Kitchen quality.