EDA on Ames Housing Data Analysis.

Student_Name

04/05/2024

1. Title and abstract:

Among data scientists, the Ames Housing Data Set is well-liked. The goal for each person is to minimize the RMSE value on the test set. Similar to that, in this paper we use a variety of exploratory data analysis techniques to analyze the information and develop an attractive linear model with a train and test data-set that includes 81 features that describe a wide range of attributes of 1,460 Ames residences sold between 2006 and 2010. In addition, five issues are included in this study that are recognized as components of the SalePrice prediction process based on different explanatory characteristics. With the data gathered from multiple websites, we are able to pinpoint here the salient characteristics that are actually influencing SalePrice. At the conclusion of the project, a list of all the websites that were referenced.To further guarantee the robustness and dependability of the model, we carry out a comprehensive assessment of its performance, which includes residual analysis and cross-validation. Our findings support educated decision-making by offering insightful information to real estate agents, legislators, and potential homeowners alike.

2. Problem Identification Part 1:

Loading and understanding the data.

```
Ames_train <- read.csv("data/train.csv",na.strings=c("","  ","NA"))
Ames_test < - read.csv("data/test.csv",na.strings=c("","  ","NA"))
```

The Ames Housing dataset comprises 80 variables, with SalePrice as our target feature of interest. Among these, 79 explanatory variables primarily focus on quantifying and assessing numerous physical attributes of the properties. These attributes typically include factors such as construction date, property size, living area square footage, parking availability, bathroom and bedroom count, flooring and roofing materials, and property location. Several continuous variables provide data on various dimensions of the properties, such as LotArea, PoolArea, and GarageArea. Additionally, categorical variables describe the quality and type of amenities, materials used in construction or renovation, street or neighborhood characteristics, and nearby amenities.Discrete variables detail the number and location of amenities, bedrooms, bathrooms, and kitchens within each property. Furthermore, temporal variables indicate the year of renovation, garage construction, and property construction

The data types of the columns are mixed:
we have integers, numric data & factors (levels). So, it's clear that features come in fundamentally different types :

1. **Some features are inherently NUMERICAL. They are quantities that we can measure or count. Some of these are continuous, such as the total living area (GrLivArea), while others are discrete, such as the number of rooms (TotRmsAbvGrd).**

2. **Other features are CATEGORICAL. They are qualitative or descriptive in nature. For example, this includes the neighbourhood in which the house is located (Neighborhood), and the type of foundation the house was built on (Foundation). There is no inherent ordering to these features.**

3. **Yet others are ORDINAL. They comprise categories with an implicit order. Examples of this include the overall quality rating (OverallQual) or the irregularity of the lot (LotShape). We can think of them as representing values on an arbitrary scale**

Thorough examination, several key features significantly influence SalePrice. These can be summarized as the size of the property, the number of rooms, location, available amenities, construction materials, overall age of the property, and the condition of both the property and itsamenities

3. Problem Identification Part 2:

These are the few DataScience problems, we came across while studying dataset and affect SalePrice based on the key features. We will find solution to each one of them once we finish data Analysis based on several questions which we think of.

1. **Problem 1: Identify which suburb/location had the biggest growth in SalePrice by plotting and examining the sale prices cross different suburbs.Has there been a trend on the type of house bought and had big hike in Sale price from 2006 to 2010**
2. **Problem 2: Analyze a possible pattern of SalePrice vs YrSold/MoSold, LotArea and/or some other variables which can reasonably be included considering Totl_Area instead of LotArea,SeasonSold instead of MonthSold here.**
3. **Problem 3: Whether SaleCondition has any impact on the SalePrice, Explain with Data Analysis and give insights on whether this feature needs to be considered.**
4. **Problem 4 : Any change in SalePrice over the period from 2006 to 2010 based on GarageQual**

5. **Problem 5:Over the years.How the SalePrice changed based on Neighborhood and BldgType Explain**

6. **Problem 5(b): Use predictions from your final model to compare suburbs which have shown varying growth. Or, to identify which suburbs have been growing the most over the last few years.**

4. Data Preprocessing:

Fitting a model that could indicate that there is zero error in the training data appears to be quite simple. That kind of model, however, would be extremely subpar since it fails to define the relationship between the explanatory features and the target variable, sales price. Fitting a linear model with the lowest RMSE and highest R-squared value is the primary goal here.

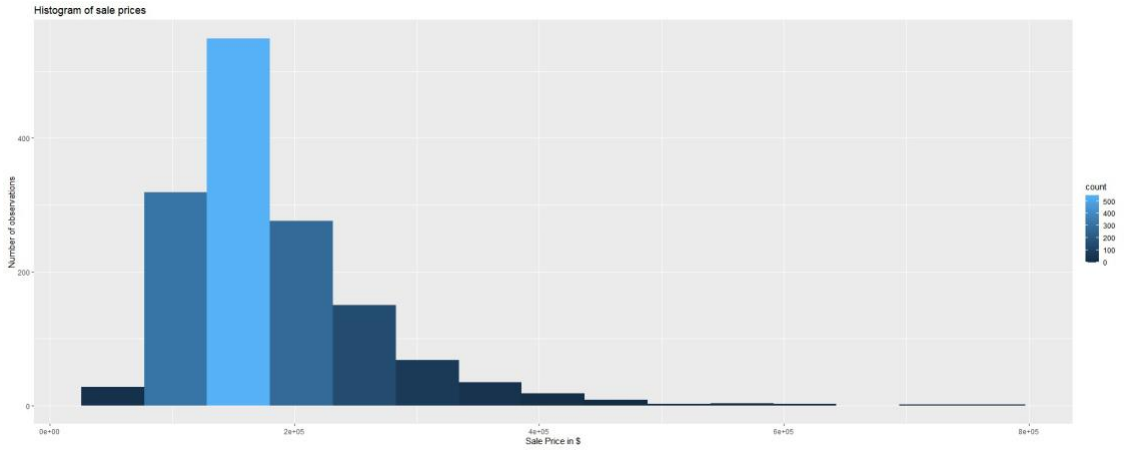
Here is the cyclic process wherein started with data exploration followed by exploratory data analysis and missing value detection and imputation. Basically, during data exploration process, one need to thoroughly study the data and relationship to the target feature. Once finding linear relationship, next step is to check the correlation values of numeric features. Here this stage answers to the question, which variables are most strongly correlated with the response. These set of numeric features will be the strongest predictor of the SalePrice. Correlation heatmap is used here to capture correlation values.
Deep hues in the heatmap typically indicate a strong correlation. The linear modeling is directly impacted by missing values. The model's quality decreases with the amount of missing values. A significant percentage of missing values indicates that the feature is essentially absent from the property. The missing value proportion of the characteristics is also displayed in the plots. For instance, there will be differences in the correlation to the SalePrice because the houses with and without pools undoubtedly account for the missing values. These NAs are imputed with the median value for numerical attributes and are marked as "Missing" in categorical features.

In the feature engineering stage, Saleprice is trained using a logarithmic transformation. Otherwise, SalePrice had a couple outiliers, which was a reasonable tilt. While some features were ordinal, others were strings. These were transformed into numerical levels so that further data could be added to the models. Next, using temporal features, it was possible to impute all year-based columns—aside from YearSold—in order to indicate the property's and garage's ages. This demonstrated the linearity of SalePrice. In order to verify the relationship between the SalePrice and the MonthSold, a SeasonSold column was also developed. In addition, a new column is established that includes the total area of the basement plus the ground living area. With this new feature, it is possible to find linearity with SalePrice once more. The 5 problems which are defined above are also solved later on. Linear modeling with the best features selected by deeper study into the data set using Exploratory data analysis helped further to choose and find the best model.

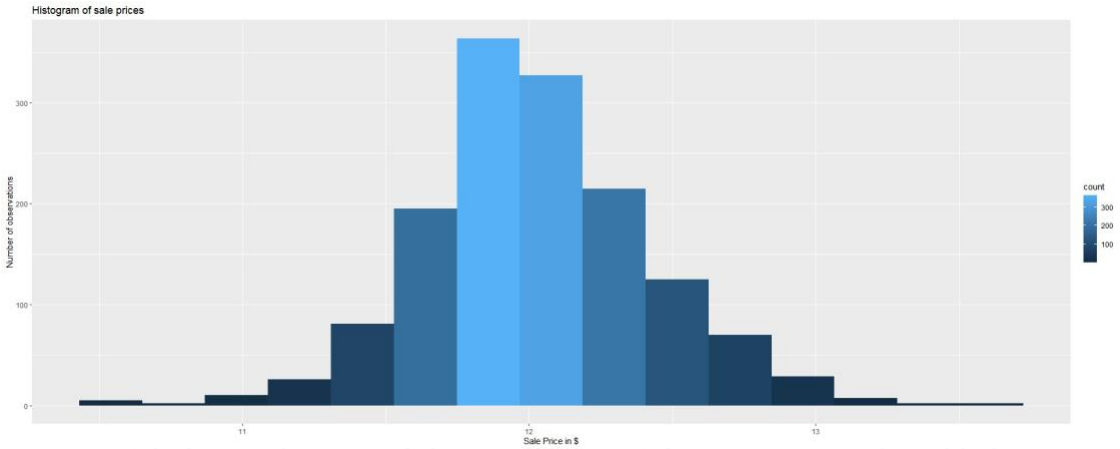
I was able to explore four models with varying but very near high R-squared values around 0.8 thanks to the knowledge I gained during the data analysis process. I primarily chose the features for modeling based on the good variability of the explanatory variables and the high correlation values.and lastly fitted the curve and trained the model. The prediction is examined using the test data set, and it is discovered that, out of the three models with the lowest R-squared value, the selected model is the best.

5. EDA:

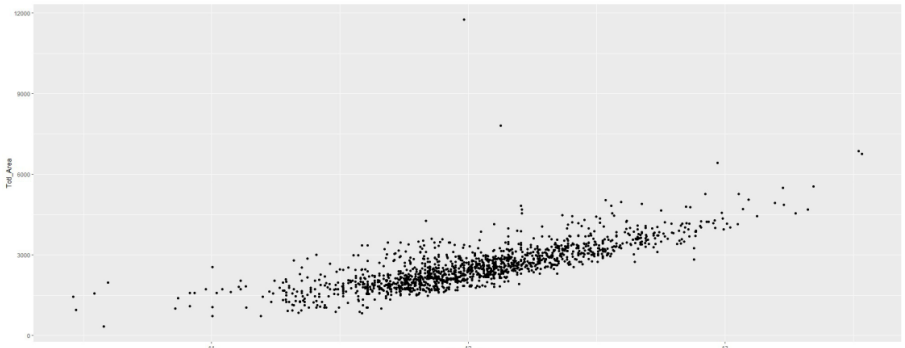
Every data analysis starts with checking on the distrubution statics of the target feature. The distribution of SalePrice looks skewed positive, and have outliers. We could consider log transformation in such a case.



The log transformation of SalePrice looks like normally distributed now.

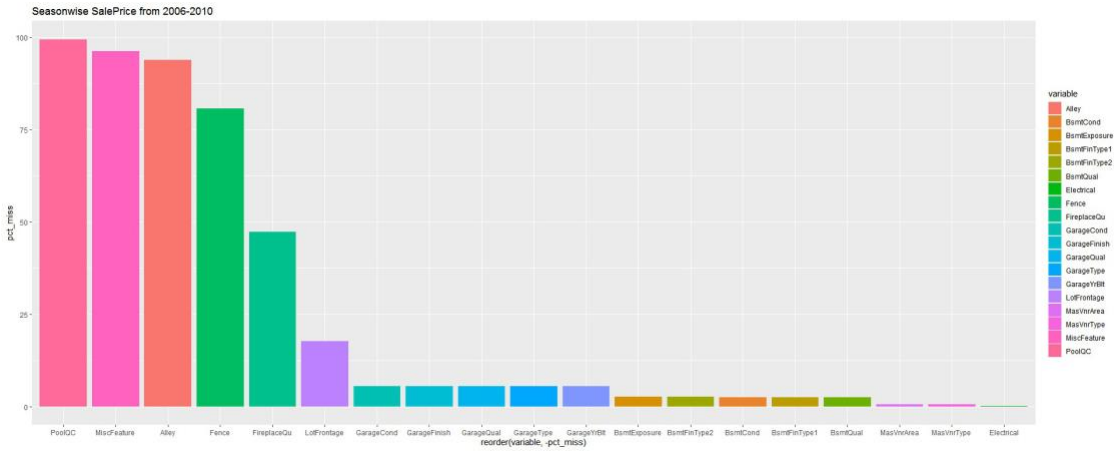


After careful observation, found that Total Area of the property can be added as new feature by adding features TotalBsmtSF and GrLivArea,



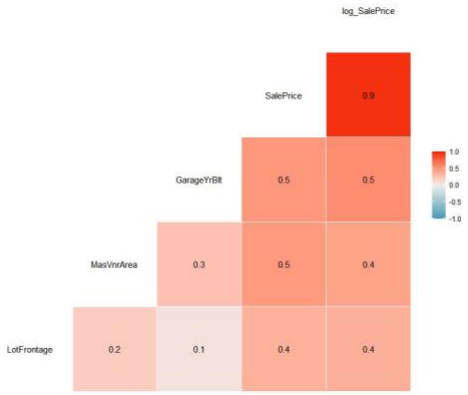
There is medium to strong linear positive relation between Totl_Area and log_SalePrice.

Here is the missing value representation of numerical columns.

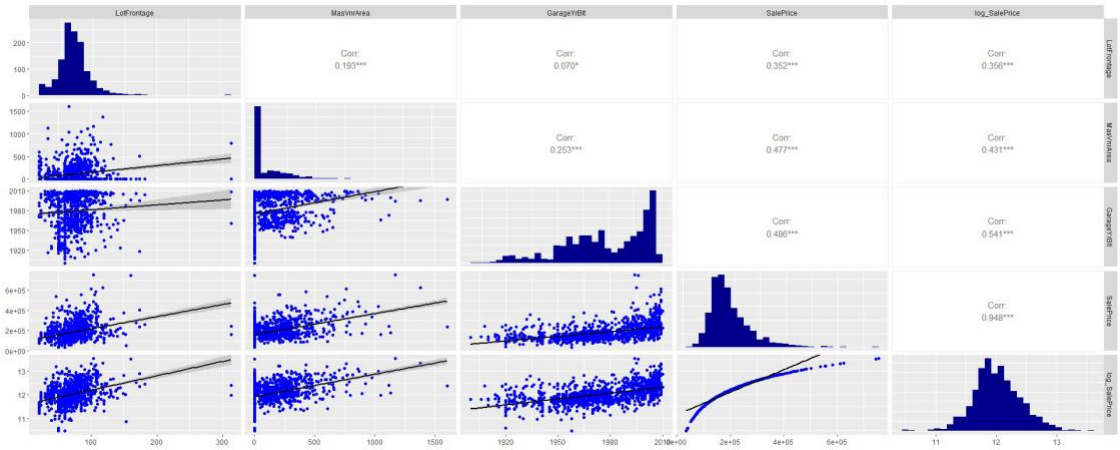


For the numeric columns with missing values, comparing correlation value here with target feature.

Considering to remove LotFrontage after careful observation. MasVnrArea and GarageYrBlt having good correlation with Saleprice.



FALSE `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
FALSE `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
FALSE `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
FALSE `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
FALSE `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.

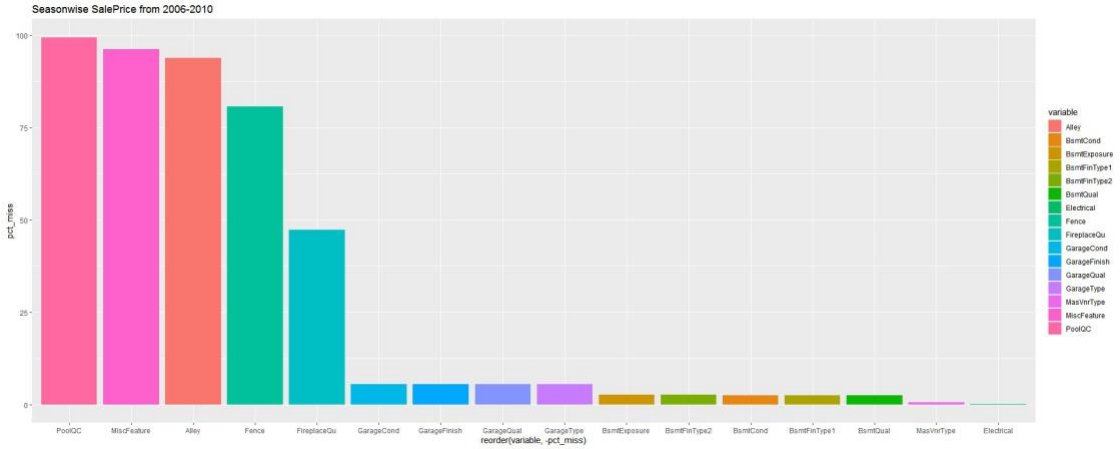


LotFrontage shows a moderate to weak positive connection with SalePrice, a low correlation, and a large number of outliers.This step also includes imputation of missing values for MasVnrArea. Since GarageYrBlt is associated with the Time feature, we may want to investigate further using additional temporal features.

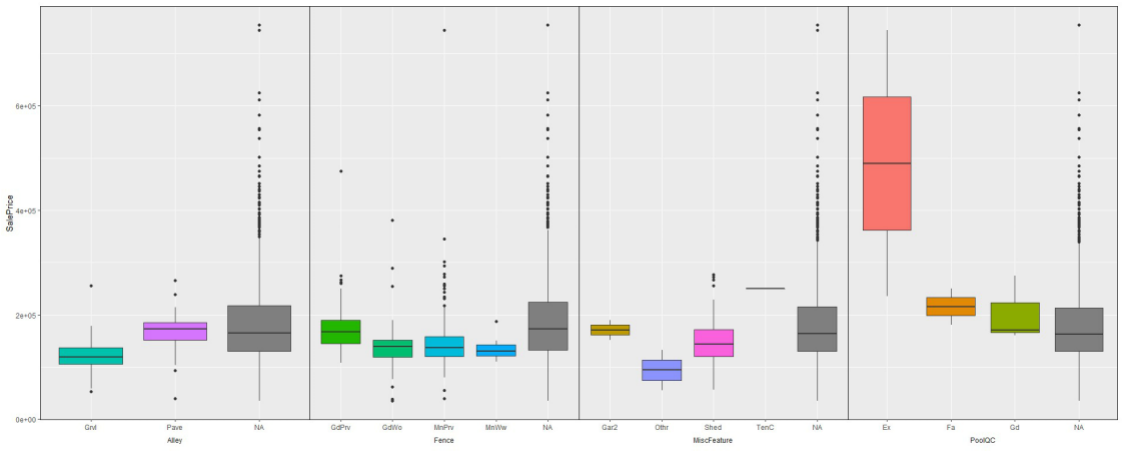
Here we complete missing values imputation for Numeric columns with NA's.

Here comes the bigger categoric feature with missing values

Considering only categorical columns which have missing values
Barplot of missing percentage for each of the categoric feature is shown here. There are 3 categorical features which shows more than 80% missing value

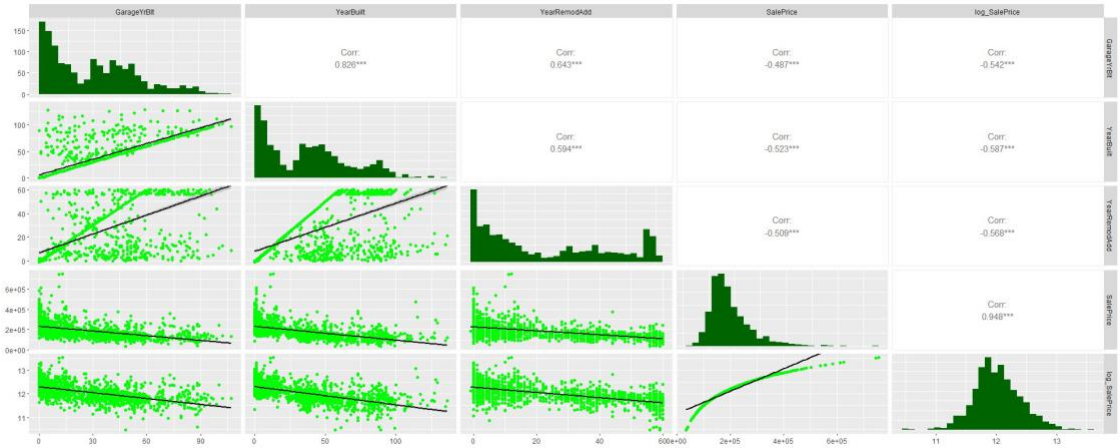


FALSE [1] "Alley" "PoolQC" "Fence" "MiscFeature"



Alley, PoolQC, Fence, MiscFeature, and FireplaceQu exhibit outliers and have NA values that are equal to or greater than 50%. Additionally, there isn't much variation in SalePrice. We can eliminate these variables immediately

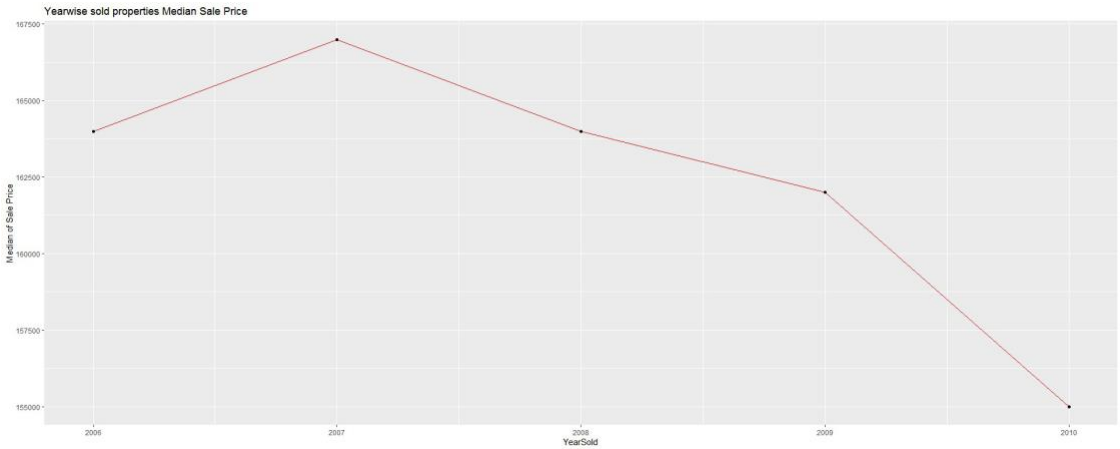
collecting all Temporal(Time/Year) related features for analysis here.After careful observation of all temoral variable, decided to convert GarageYrBuilt,Yrbuilt,YearRemod so that they represent the age of the garage and the property. As the YearSold already showing for which and all year we have data, taking difference between YrSold respectively with each feature here gives the age of the property and Garage area.



The Sale Price is declining as the property matures in terms of the garage or the building itself.

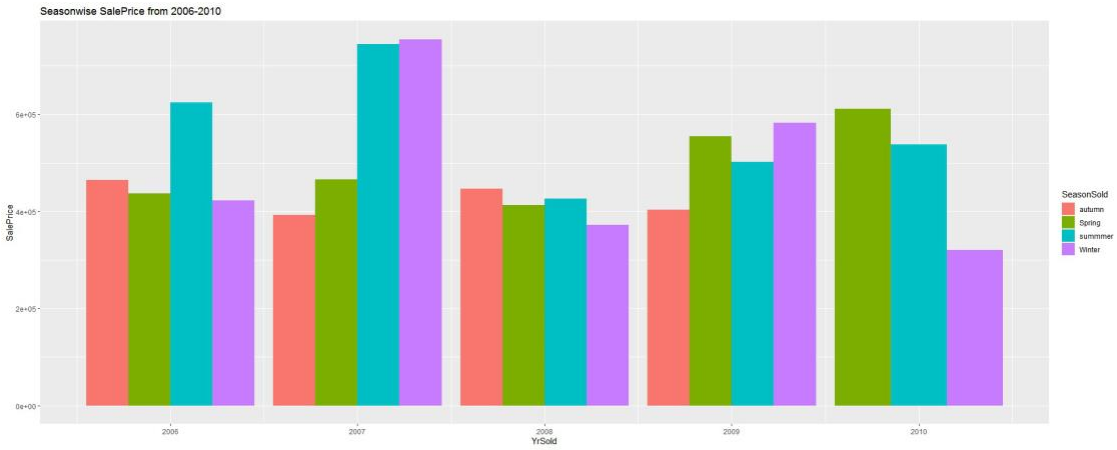
Given that the three yearwise columns' correlation values are all quite high, it appears that they all have a negative, strong link with sale price and are therefore significant in forecasting sale price.

Here checking for YrSold and SalePrice relationship.



Sale Price declining between 2006 and 2010 is another noteworthy finding to take into account. Property values typically increase as a result of the community and amenities. However, the building's general exterior and interior quality also deteriorates with increasing property age.

We might also make the intriguing column SeasonSold in this instance. Weather-related possibilities must exist; perhaps SalePrice is a determining factor. While we build this SeasonSold column here, MonthSold may be removed.

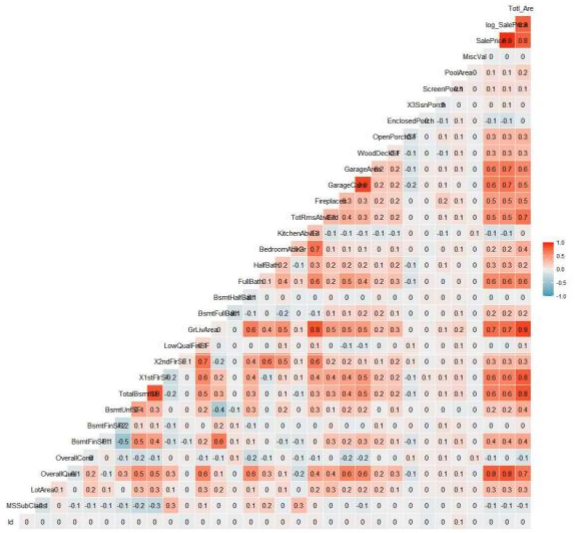


SeasonSold has very big impact on the SalePrice this was particularly true in 2007, when winter and spring had the highest sale prices for the houses sold, followed by autumn and spring.With the exception of 2007, every season when properties were sold between 2006 and 2010 generally followed the same pattern.

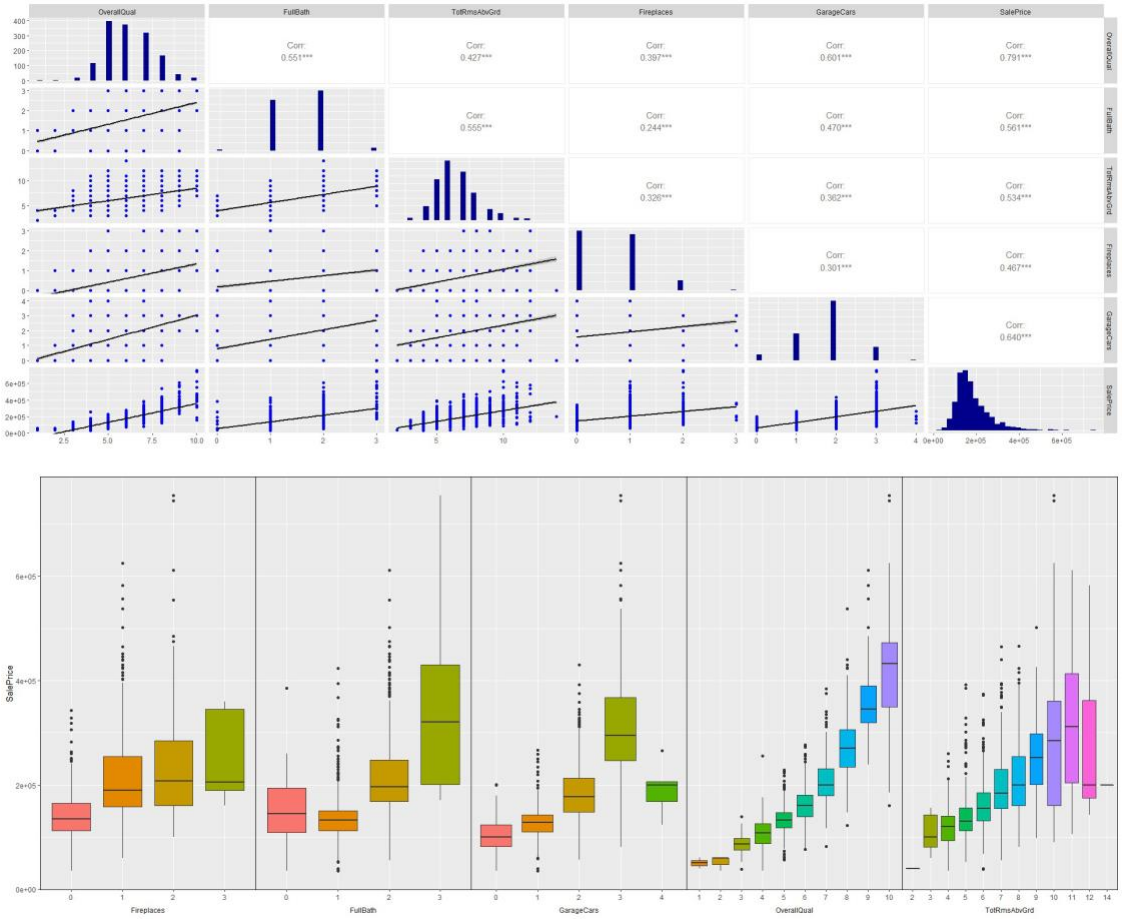
Till now explored with whether any new features can be created. And also dealt with missing values.

Imputation of Temporal variable. Filling missing values with median is done later.

Here analyzing numerical columns without missing values. Whether any feature could be dropped or not is decided based on correlation values with SalePrice.

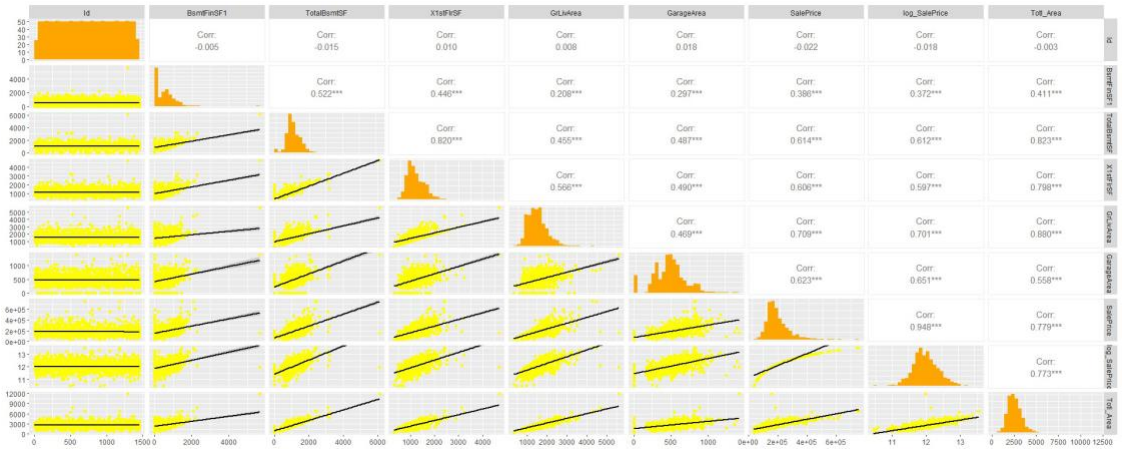


As we could see the feature having less than 0.5 correlation SalePrice shows clear info on the scatterplot which is not printed here. When Analysed further segregating them to discrete and continuous features, got more insights on them.



Complete bathroom, garage cars, and general quality all have a high, positive correlation with sale price. Fireplaces have a favorable increase in SalePrice, and TotRmsAbvGrd likewise shows a positive strong relation and much variability with SalePrice

Now dealing with continuous variables with exploratory data analysis.



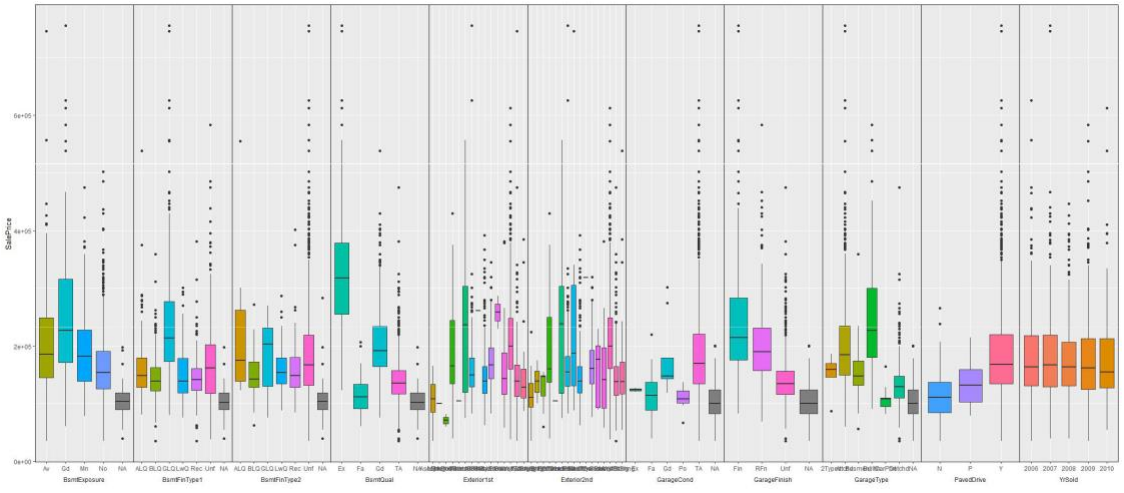
All these continuous features has very strong positive linear relation with SalePrice.
All the numerical columns analysis completes here.

Now considering all the categorical variables which are not having missing values.And even those with missing values are imputed here in this step.

As Given the abundance of category features, they are divided into internal and external key categories as well as miscellaneous categories. This is purely for the purpose of visualization.

Utilities could be removed because of no variability with SalePrice and many outliers when checked with boxplot which is not printed here.

Since there is little variation in this case when compared to SalePrice, four columns—BsmtCond, ExterCond, RoofStyle, and RoofMatl—can be eliminated at this time.once the boxplot, which is not printed here, has been checked.



In comparison to other features here, TotRmsAbvGrd has very good variability.
Here, I'm dropping everything but TotRmsAbvGrd.once the boxplot, which is not printed here, has been checked.

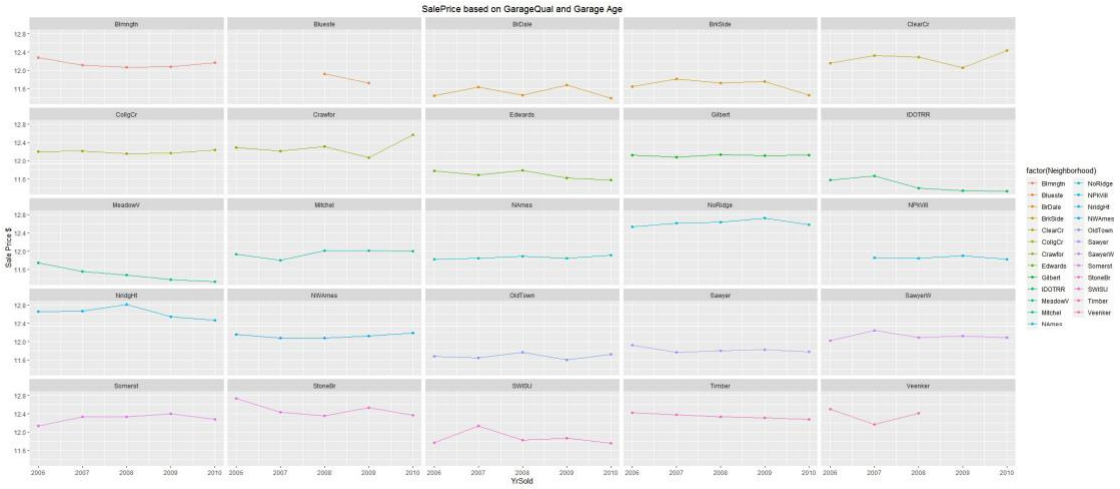
BsmtQual looks better related compared to all other features here. Removing all other except BsmtQual here. Once we select features from categorical features, Imputation for all the categorical features done at this stage.

6. Further Preprocessing:

Further Preprocessing and Exploratory data analysis done at this step. Deeper understanding and finding some of the solutions to the problems identified at the beginning of this project.

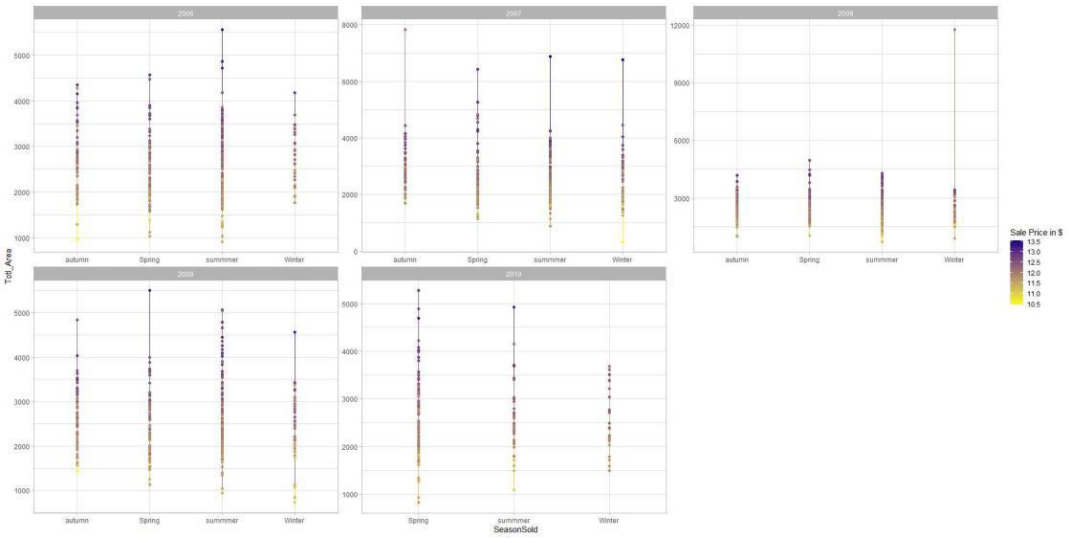
During the first stage of data analysis , the decissions are taken carefuly. so that it doesnt take much of the effort here.

Problem 1: Identify which suburb/location had the biggest growth in SalePrice by plotting and examining the sale prices cross different suburbs.Has there been a trend on the type of house bought and had big hike in Sale price from 2006 to 2010.



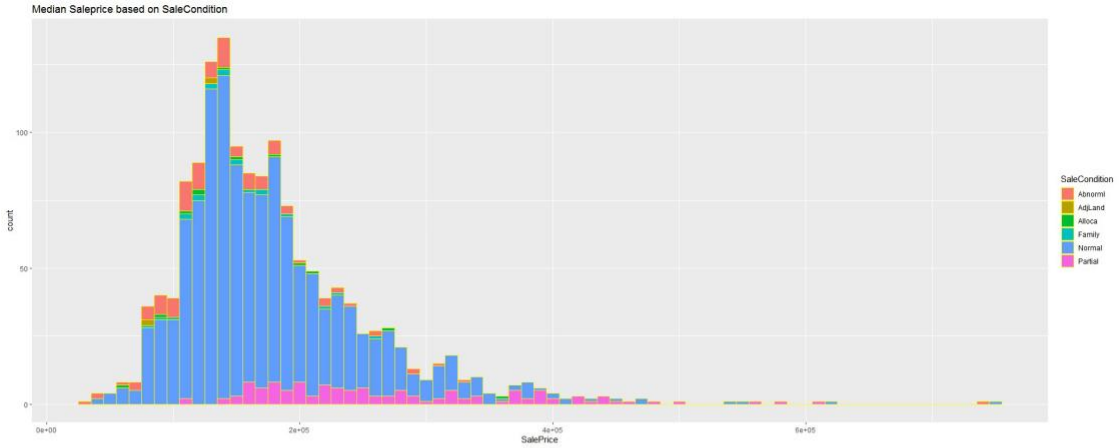
1. The price of the Somerst suburb decreased in 2010 after rising until 2009.
2. Although Noridge's sale price has been rising as a trend, it has declined in 2009.
3. NAmes shows a yearly tendency of minor increase. So there must be other features which are affecting SalePrice here along with Neighborhood(suburb/location) of the property.

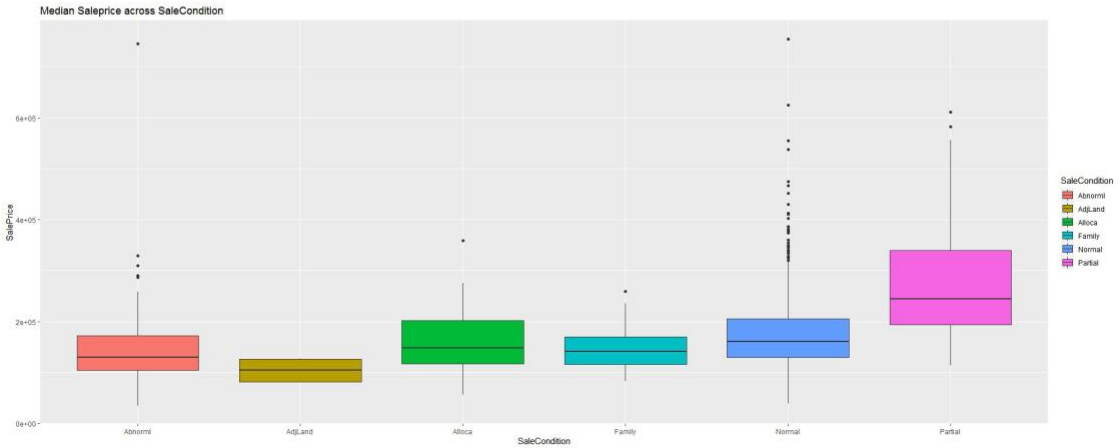
Problem 2: Analyze a possible pattern of SalePrice vs YrSold/MoSold, LotArea and/or some other variables which can reasonably be included considering Totl_Area instead of LotArea, SeasonSold instead of MonthSold



1) The summer of 2006 saw an exceptional surge in the sale price that appeared to be an outlier; nonetheless, the maximum sale price for all houses sold with larger total area (total basement area and ground living) was reached throughout the season at $4000.
2) In 2008, all seasons saw extremely low sale prices, especially for large estates in Totl Area. The average log_sale price was approximately $4,000.
3) In 2009, all seasons saw maxima for the properties with larger Totl_Area of over $5,000.

Problem 3: Whether SaleCondition has any impact on the SalePrice,Explain with Data Analysis and give insights on whether this feature needs to be considered.





1) SalePrice and SaleCondition unquestionably have a relationship.
2) Typical Transaction a lot of outliers and varied.
3) The sale prices of family sales between relatives decreased significantly.
4) In a similar vein, an abnormally large sale could result from a swap, short sale, or foreclosure.

Problem 4 : Any change in SalePrice over the period from 2006 to 2010 based on GarageQual

1) The SalePrice did, in fact, vary over the years. From 2006 to 2009
2) The average quality garage properties' log_SalePrice was approximately $3,500.
3) In 2010, the Sale Price of even very good grade garages unexpectedly dropped; this could have been brought on by the garage's advanced age.

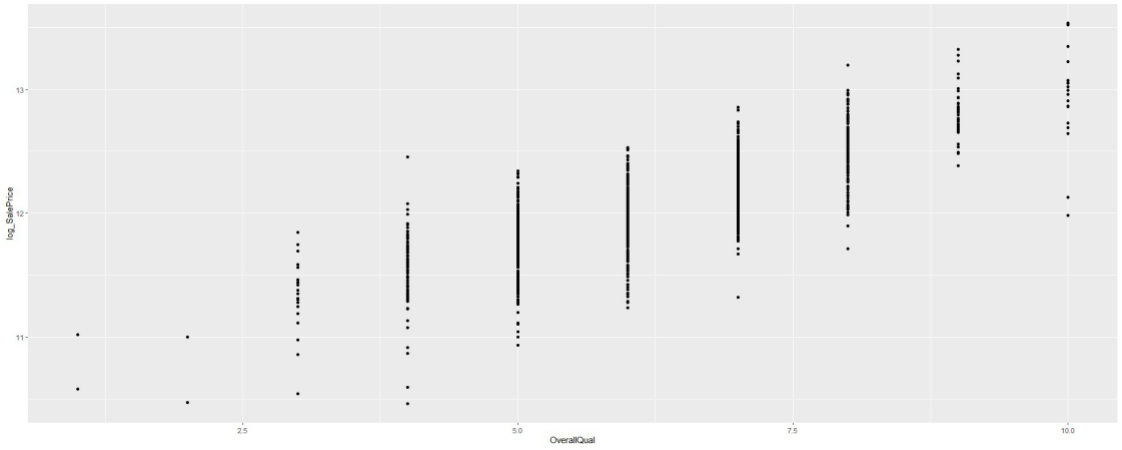
Problem 5(a) :Over the years.How the SalePrice changed based on Neighborhood and BldgType .Explain

1) A number of neighborhoods, including BrkSide, ClearCr, Gilbert, IDOTRR, NoRidge, NWAmes, Timber, NRidgHt, StoneBr, and others, have sold dwellings of type 1Fam.

2) From 2006 to 2010, the only 1Fam building type in the ClearCr Gilbert, Timber NoRidge location seemed stylish. Every year, the sale price is essentially the same.

3) The building type TwnhsE, 1Fam in Suburban Blmngtn CollgCr StoneBr, Somerst, appears to be fashionable. From 2006 to 2010, the sale price for both types of buildings is about the same.

Problem 5(b):Do you think we could get good linear model with just considering Overall quality as a stand alone parameter for Sale Price prediction? Give thoughts




```
model <- lm(Ames_train_tidied$log_SalePrice ~ Ames_train_tidied$OverallQual,
            Ames_train_tidied)
summary(model)
```

Even though RMSE is 0.2303 the R-squared: 0.6678, which cannot be considered as a good linear model just by considering the OverallQuality.

The problem 5(b) will be solved once fitting the model is completed

7. Modeling:

Here considered features with strong correlation and good variability to SalePrice.

I was able to obtain the R square=0.8599 and RMSE=0.1519 values using the first fitted model.

```
summary(model_Ames_train_tidied)
```

It was possible to obtain the values of R square = 0.8772 and RMSE = 0.1431 using the second fitted model.

I was able to obtain the R square = 0.8507 and RMSE = 0.1567 values using the third fitted model.

```
summary(mode3_Ames_train_tidied)
```

mode4_Ames_train_tidied is the best fit with RMSE=0.1401 # and R^2=0.884 compared to other fitted models for test data set-> lm

There are few outliers as following residual shows in the fourth model.

```
outlierTest(mode4_Ames_train_tidied)
```

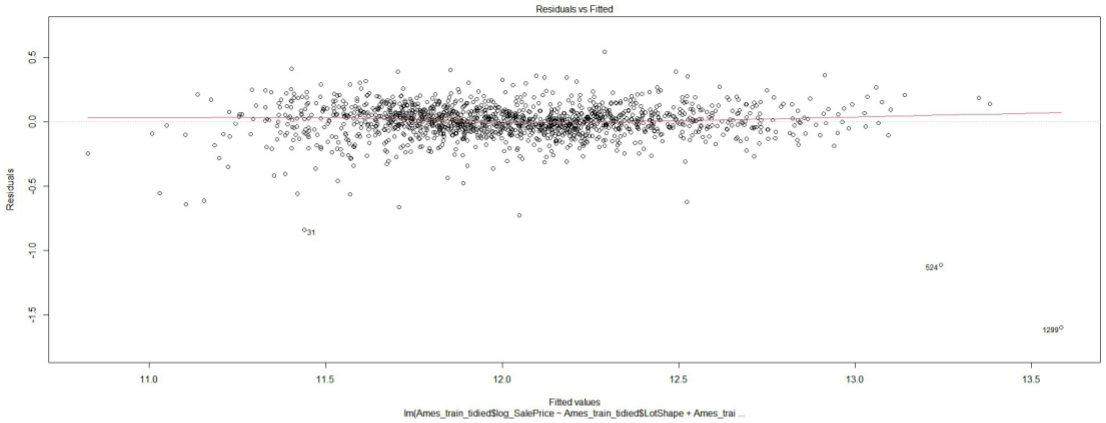
	rstudent	unadjusted p-value	Bonferroni p
FALSE			5.4678e-46
FALSE 1299	-15.341654	3.7451e-49	1.0486e-14
FALSE 524	-8.731117	7.1822e-18	9.5339e-07
FALSE 31	-6.221352	6.5301e-10	8.6348e-05
FALSE 633	-5.451498	5.9142e-08	1.9638e-03
FALSE 463	-4.854481	1.3450e-06	2.8316e-03
FALSE 496	-4.780196	1.9394e-06	6.8505e-03
FALSE 1325	-4.596458	4.6921e-06	1.0184e-02
FALSE 969	-4.511834	6.9754e-06	2.8903e-02
FALSE 314	4.282135	1.9796e-05	

There are few outliers.

With the third fitted model, was able to achieve the R square=0.884 and RMSE=0.1401 value.

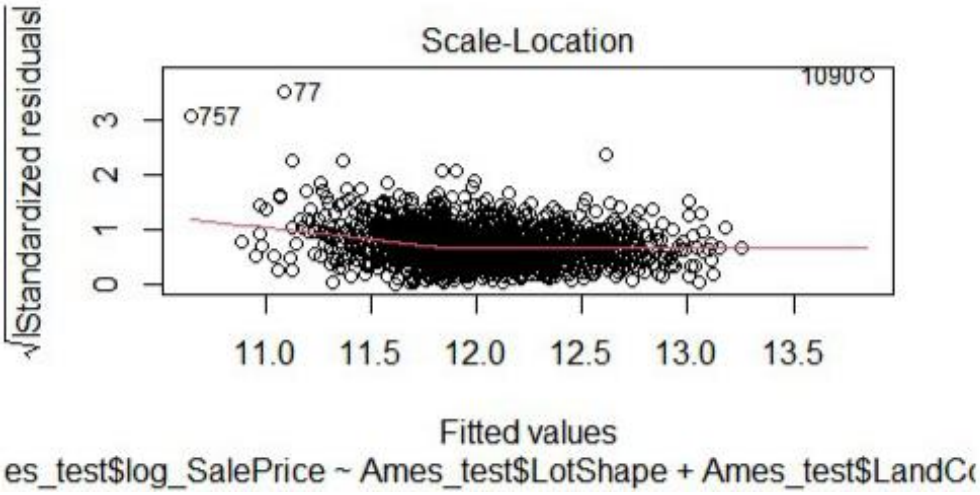
```
summary(mode4_Ames_train_tidied)
```

Here is the linear regression,residual and cook's distance plot for the best fitted model.



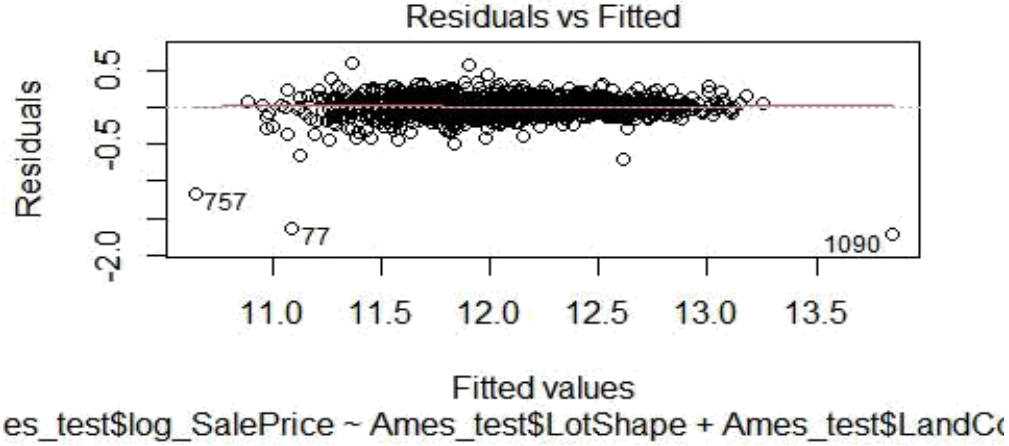
When predicted with test dataset, I got the R square-0.8973 and RMSE=0.1375 value which is very near to the fitted model.

```
summary(model_Ames_test)
plot(model_Ames_test,1)
plot(model_Ames_test,3)
```



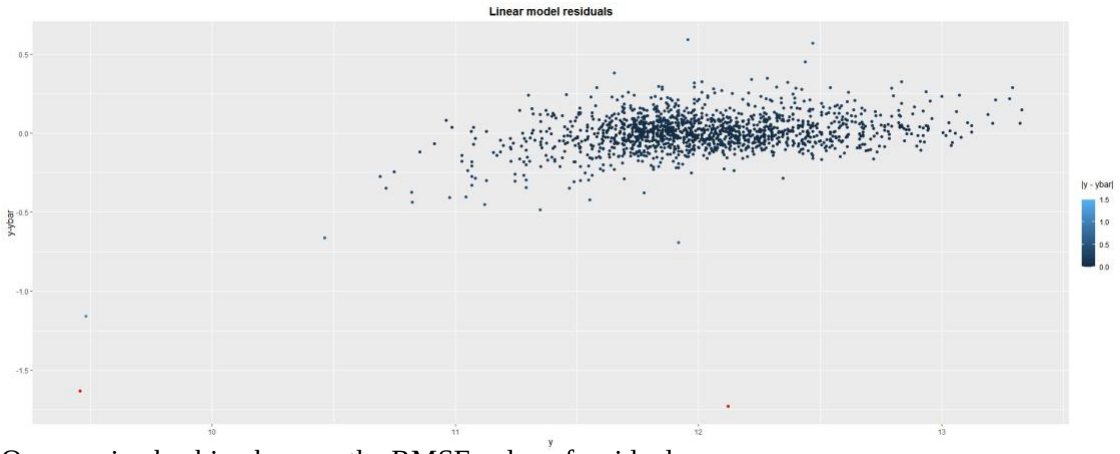
Here is the comparision of RMSE values of both fitted model and test dataset.

```
rmse(mode4_Ames_train_tidied,Ames_train_tidied)
FALSE [1] 0.1359736
rmse(model_Ames_test,Ames_test)
FALSE [1] 0.1333493
```



8. Evaluation:

predicting test dataset and plotting residuals



Once again checking here on the RMSE value of residuals.

FALSE [1] 0.1333493

Problem 5(b): Use predictions from your final model to compare suburbs which have shown varying growth. Or, to identify which suburbs have been growing the most over the last few years.

According the fitted final model the formula for prediction of SalePrice is as follows.

*SalePrice = (-3.479e-02)*NeighborhoodBlueste+1.318e+01*
*SalePrice = (-9.795e-02)*NeighborhoodBrDale+1.318e+01*
*SalePrice = (1.237e-01)*NeighborhoodVeenker+1.318e+01*
and so on

So Based on the Neighbourhood value in test dataset SalePrice will be predicted. Similary the formulas could be derived with each and every dependent feature to predict SalePrice. The final formula would be as follows.

as Multiple linear models will follow the general form

y = a1x1 + a2x2 + . + b
SalePrice = a1Neighborhood + a2OverallQual + b….and so on

while a1= feature1*coefficient+/-intercept and so on wheere feature=Neighborhood and coefficient and intercepts are the respective estimations using linear modeling.

9. Recommendation and Conclusion:

The relationship between house prices and the economy is an important factor for predicting house prices. As per buyer and sellers concern Housing prices trends are very important to study before making an investment, Hence it is directly or indirectly related to current economic situation. Therefore it is important to predict housing prices without bias to help both buyers and sellers make their decisions

Through conducting data collection, including various data processing methods, and applying analytic approaches, I have identified a multiple linear regression model that is suitable for the dataset.

First, I cleaned up the train dataset and used it to generate 4 models. I could make four models based on the attributes' qualities.

The first model having R square value 0.8599 and RMSE: 0.1519.

The second model having R square value 0.8772 and RMSE: 0.1431.

The third model having R square value 0.8507 and RMSE: 0.1567.

The fourth model having R square value 0.884 and RMSE: 0.1401

Essentially, I chose the attributes based on their significant relationship to the SalePrice and their variability within the SalePrice distribution. In the end, I determine that the fourth model fits and predicts the values the best. Plotting residuals and predicting using the test dataset with the best model later, to see if the prediction based on the fitted model is close to the absolute value. Considered here is log_SalePrice rather than SalePrice. Using SalePrice's log transformation is the best option, since SalePrice's consideration for modeling yielded RMSE=30020 and R-squared=0.8654.

If there is any more research done, I would prefer to focus on multiple regression methods rather than just linear regression. That would assist me in examining and taking into account the remaining features that I neglected to include in my linear regression calculations.

10. References:

Big vote of thanks to all the references mentioned below here. Without which I would have not successfully able to complete linear modelling for multivariable data set.

http://jse.amstat.org/v19n3/decock.pdf

http://stackoverflow.com/

http://www.cran.r-project.org

http://www.stackoverflow.com

http://www.edx.org

http://www.rapidtables.com

https://rpubs.com/RobbyS/622233

https://scholarworks.calstate.edu/downloads/fx719m836

https://www.youtube.com/watch?v=wR4Xfwjr-3Y&list=LL&index=14

https://nycdatascience.com/