House Prices Prediction

I. Introduction

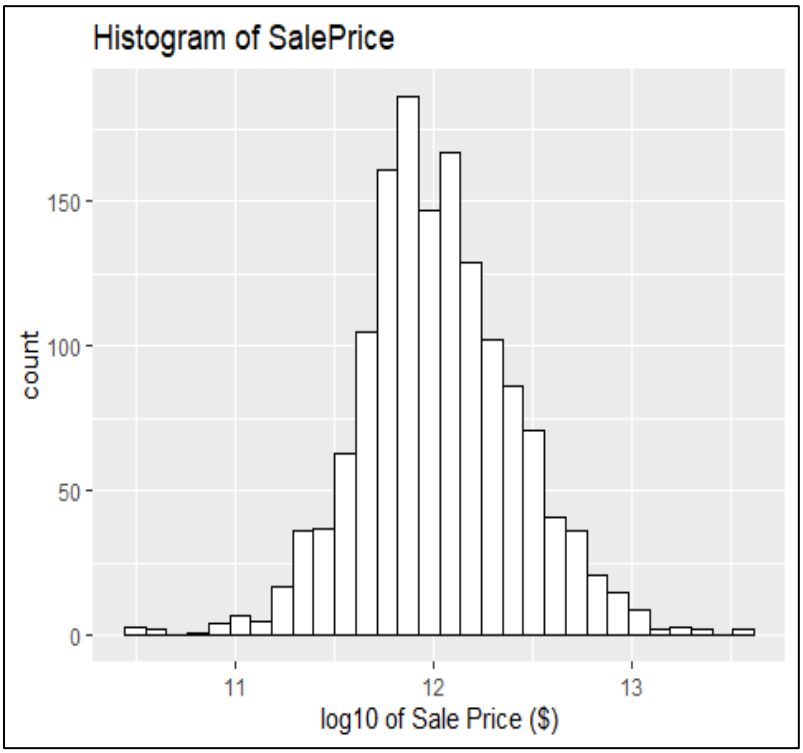
The data set chosen is 'House Price Prediction' from Kaggle[1]. This dataset involves 79 explanatory features that are used in predicting house prices in the US. It is a regression problem of predicting house prices of new data points given previous house prices in accordance with quantitative and categorical explanatory variables. For EDA purposes we have chosen 7 meaningful variables out of which one is the target variable of 'sale price'. The motive and goal of this project are to get insights and relationships amongst the variables which might prove to be useful in deciding factors responsible for deciding the house prices.

II. Data Description

Some of the variables chosen from the original data set that seems to be useful for prediction purposes are as follows.

a. Target Variable

Sale Price:
The property's sale price in dollars. This is used as the target variable for which the predictive model is being utilized.

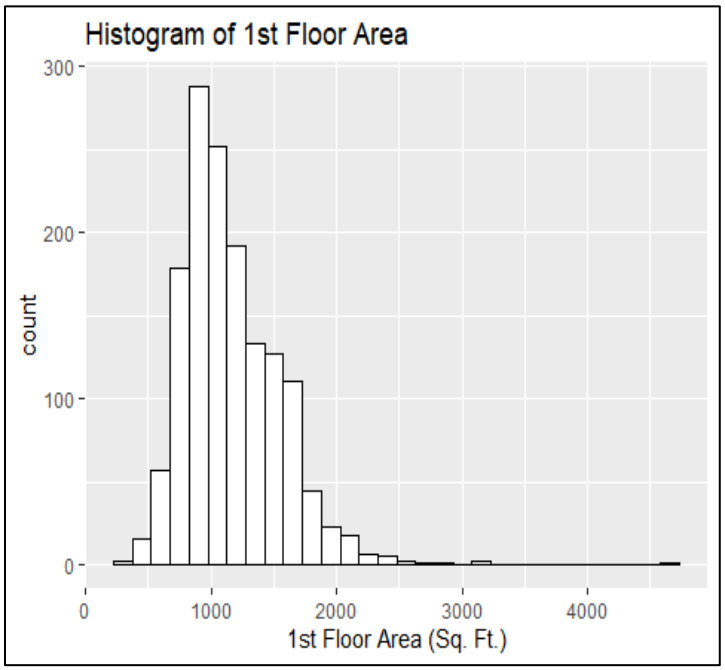


Above is the distribution of the target variable. The sale prices have been log-transformed which makes the distribution normal without adding excess complexities to the model.

[1] **https://www.kaggle.com/c/house-prices-advanced-regression-techniques/overview**

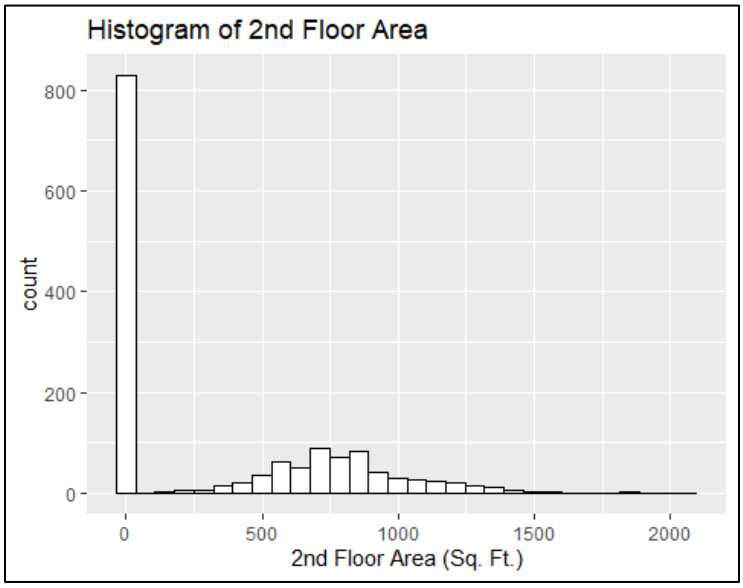
b. Quantitative Variables

1. *1stFlrSF:*
   This is the First-Floor area in square feet.



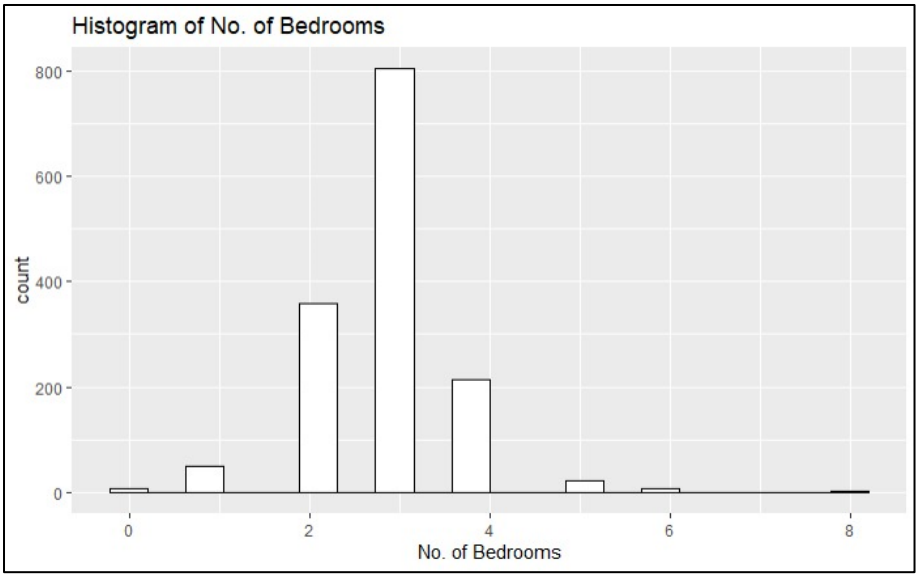
This is the distribution of the 1st floor area variable which shows that most of the houses' areas lie within the range of 500 - 2000 Sq. Ft.

2. *2ndFlrSF:*
   This is the second-floor area in square feet.

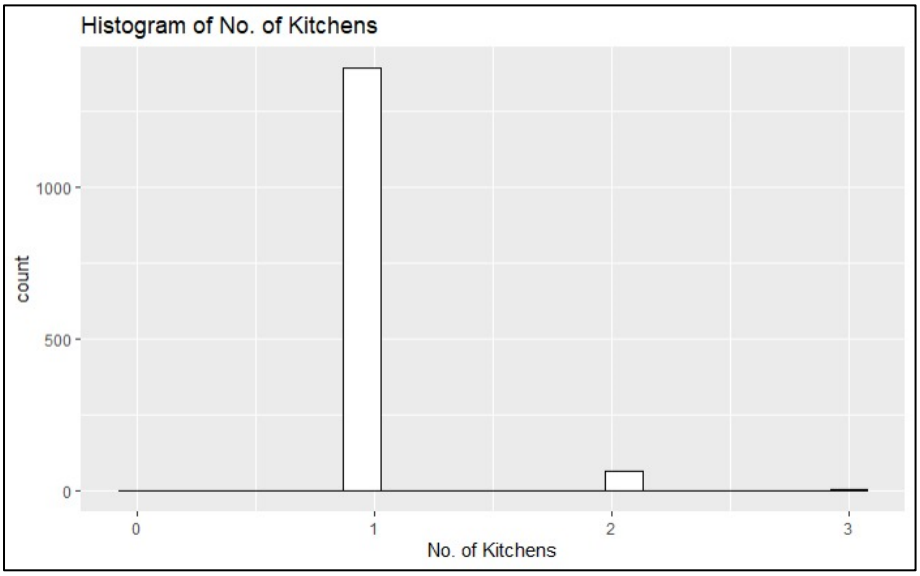


The above distribution shows that there are many houses with no 2nd floor (hence these houses have been filtered out in the analysis of 2nd floor area graphs). Apart from those houses, the distribution is somewhat normal.

3. *Bedroom:*
   This is a discrete quantitative variable explaining the number of bedrooms above the basement level in each house.

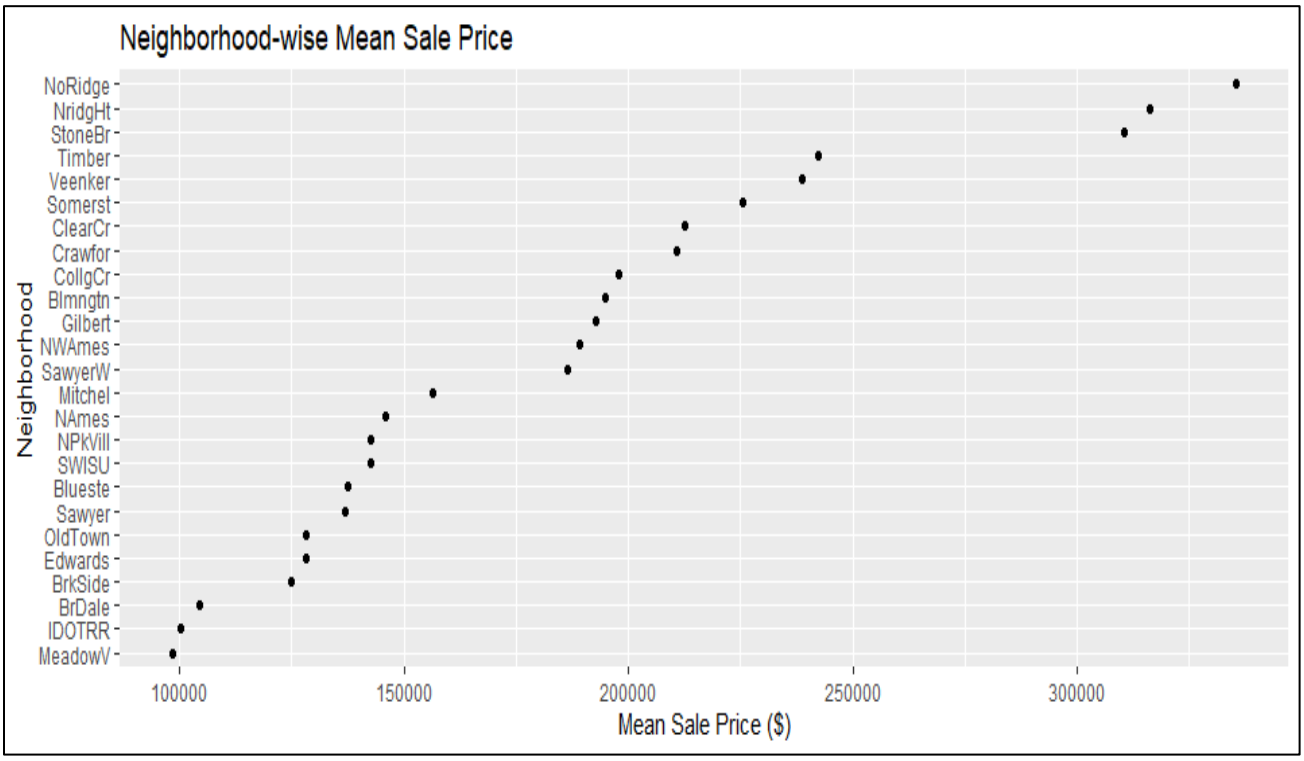


4. *Kitchen:*
   It tells the number of kitchens in each house. Its distribution though not very insightful did tell us that almost all the houses have only one kitchen. Only a few of the houses have 2 - 3 kitchens.

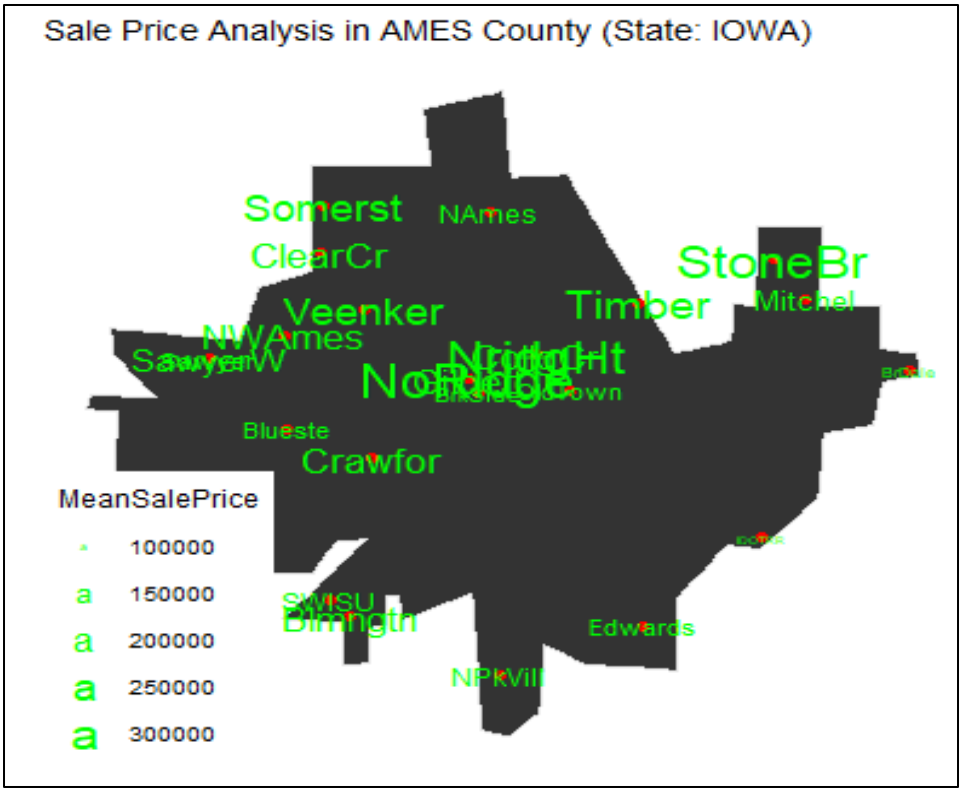


5. *YearBuilt:*
   This variable shows the original construction date of the house.

6. *Neighborhood:*
   This categorical variable defines the physical locations within Ames city limits.



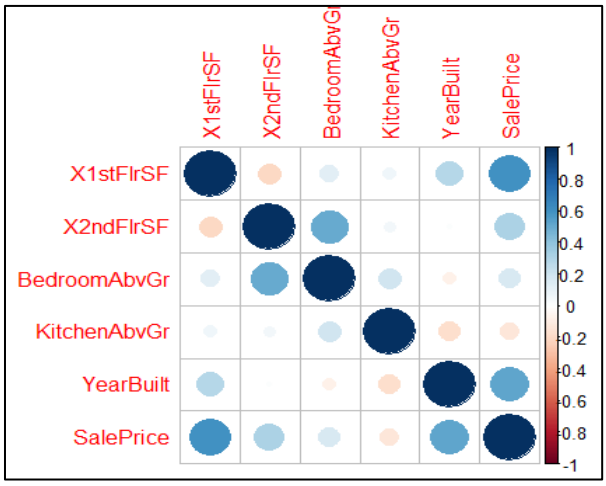
'NoRidge' neighborhood has the highest Mean sale price and 'MeadowV' has the lowest. The same information is visualized geo-spatially as seen in the next graph (map).



The county plot above depicts the distribution of neighborhood in the AMES county. The size of the Neighborhood names suggests how expensive are the houses present in that area.

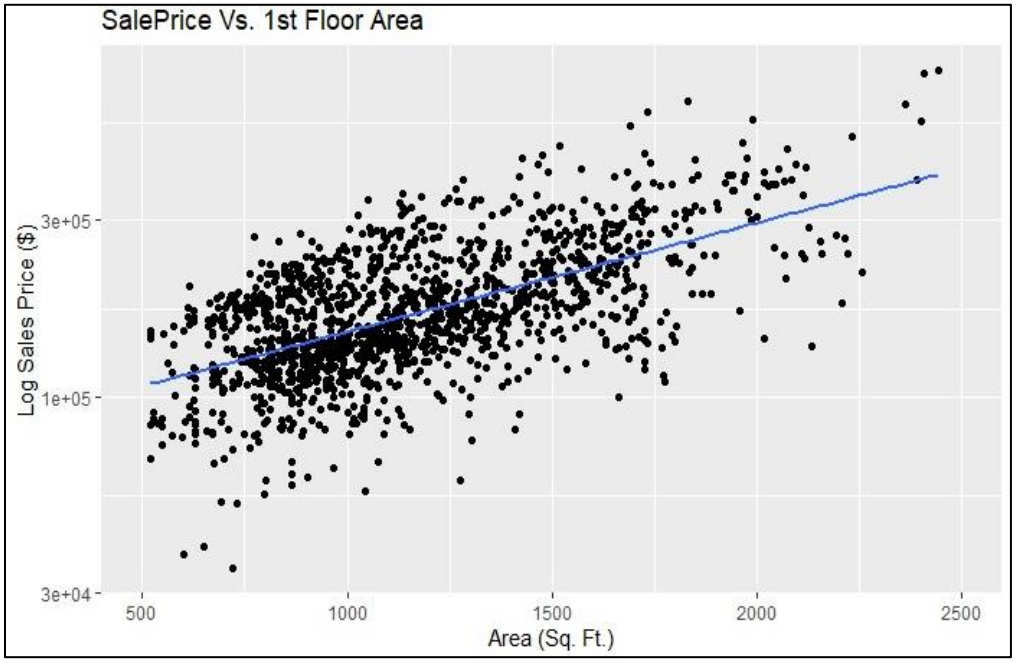
III. EDA and Data Modelling

The first step was to observe a correlation rank matrix amongst the explanatory variables to check for a high correlation amongst features.



It's observed that none of the features are significantly correlated hence no redundancy is there in Data.

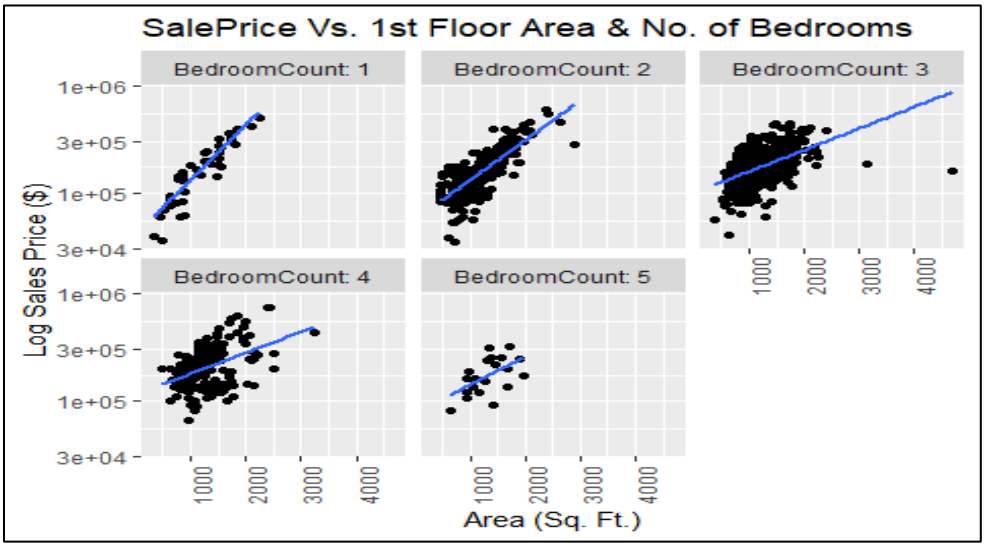
a. SalePrice Vs. 1st Floor Area



The 1st Floor Area and Sale Price shows a linear trend and hence linear model is used. From the plot, we can deduce that the sale price increases with the increase in the 1st floor area (excluding a few outliers).

Sale Price Vs. 2nd Floor Area showed a similar trend, therefore its plot has not been included in the report.

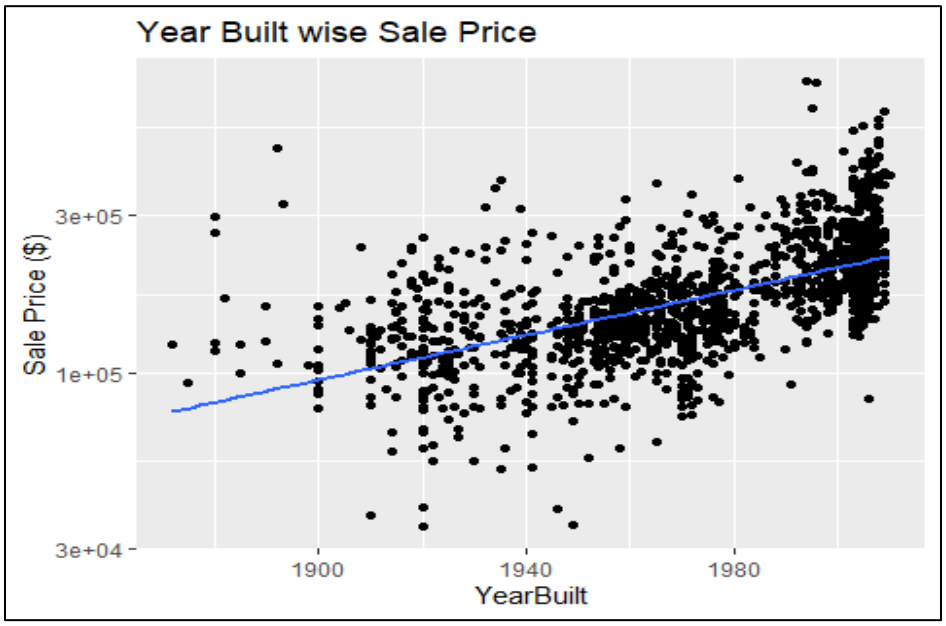
b. SalePrice Vs. 1st Floor Area and No. of Bedrooms



From the above graph, it is evident that Sale Price is linearly dependent on the 1st Floor Area. In this graph, we introduce another factor "Bedroom Count" where we observe the trend between each bedroom count. It also depicts the same story, where the trend is linearly increasing.

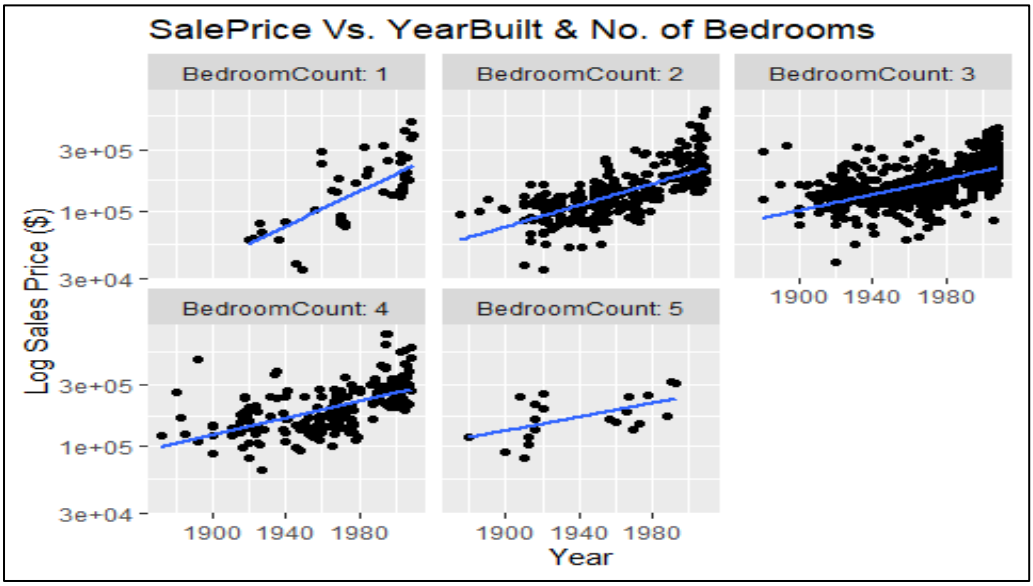
Graph for "Sale Price Vs. Floor Area & Kitchen Count" is not included because there are not many houses that have a kitchen on the 2nd floor. This is because most of the houses only have 1 kitchen. Hence, we cannot get any insights from the plot.

c. Sale Price Vs. Year Built



This graph explains the relationship between the 'Year Built' variable and 'Sale Price'. Though house prices have a wide range for a particular year, the big picture can be derived that sale prices increase with 'Year Built' with nearly a linear trend.

d. Relationship Between Year Built and No. of Bedrooms



This graph shows the trend of sale prices Vs 'Year Built' across the different number of bedrooms.

Irrespective of no. of bedrooms, the sale price shows an increase as the year progresses with a marginal difference across several bedrooms. Hence the number of bedrooms not that important of a feature while predicting the Sale price of houses.

Relationship Between Year Built and No. of Kitchens is not included since as aforementioned most of the data points (houses) contain only one kitchen.

e. Model fitting

1. *Linear Model*

```
## Coefficients:
##                Estimate
## BedroomAbvGr -0.139933
## KitchenAbvGr -0.199590
## YearBuilt     0.151870
## X1stFlrSF     0.810935
## X2ndFlrSF     0.272050

## Multiple R-squared:  0.7154, Adjusted R-squared:  0.7144
```

Linear Model is the first option for this dataset which can be deduced intuitively (since the target variable follows a linear relationship with most of the variables) and it's quite easy to interpret.

According to our EDA and model fitting it seems, 1st-floor area (followed by the 2nd-floor area) is the most important feature. Intuitively (and as per our EDA results) 1st-floor area shows linear relation with the highest slope when plotted against the target variable Sales Price. This makes sense as with larger areas house prices have increased considerably.

Statistically, it can be observed from the results of the linear model fitting that the 1st-floor area variable has the highest estimate (weight) in the linear model consisting of all the quantitative variables. After Area, 'Year Built' is the next important factor in deciding house prices.

Note:- To physically interpret the estimates (of R's Liner Model output) all the numeric features have been scaled to range between 0 and 1.

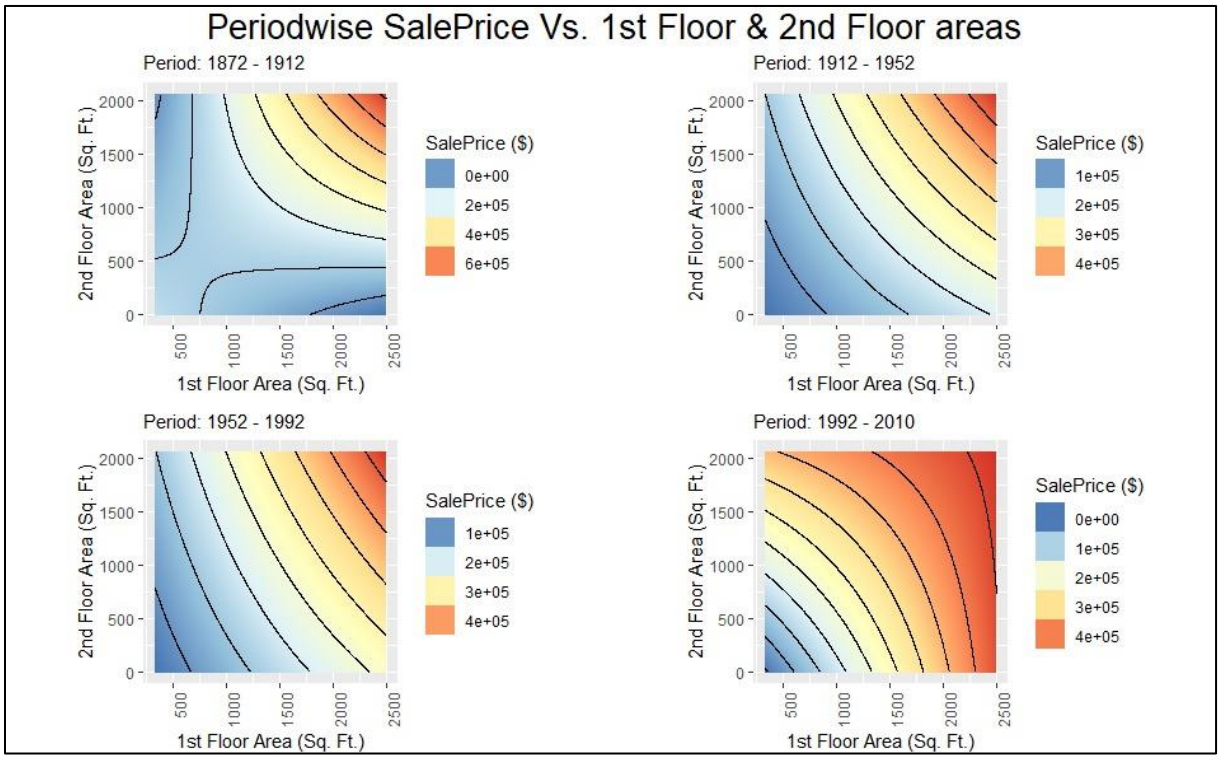
2. *Linear Model with interactions*

The previous model though very easy to interpret was quite simplistic. Hence adding non-linearity in the model, we introduce interaction terms to our pre-existing linear model. A new feature is introduced with interactions between the 1st-floor area, 2nd-floor area, and the Year-Built of houses. The intuition behind this is interacting the 3 major factors responsible for sale-price (which we obtained after seeing the estimates from the previous linear model) and then observe how the model behaves.

```
## YearBuilt:X1stFlrSF           2.080e+00
## YearBuilt:X2ndFlrSF           2.724e+00
## X1stFlrSF:X2ndFlrSF           4.494e+00
## YearBuilt:X1stFlrSF:X2ndFlrSF -2.274e-03
## Multiple R-squared:  0.7499, Adjusted R-squared:  0.7483
```

From the above results, we observe that overall R-squared has been improved by the new linear model (with interactions). This explains that this model fits the data even better. Hence this model fits the data best, and theirs is no specific use of implementing more complicated models.

f. Visualizing the final Model



Next, we decide to visualize our finalized model. The contour plots are developed using a linear model (lm) with interactions between the input variables (between 1st floor and 2nd-floor area). This shows that the target variable "Sale Price" is more dependent on the "1st Floor Area" as compared to "2nd Floor Area".

In the period 1872-1912, we can observe that most of the houses were in the low-mid price range. However post-1952, houses with low-mid price range & mid-high price range have equal proportions in the market. The recent years show that low price range houses have decreased with an increase in mid-high price range houses.

From the above density plot we observe that the predicted sale prices are almost equal to the actual sale prices, thus proving the credibility of our model and subsequently the conclusions that were claimed from the model.

IV. Conclusion

In this study, we analyzed Sale Prices of houses in Ames county (Iowa State) based on various predictors such as 1st Floor Area, 2nd Floor Area, No. of Bedrooms, etc. We found that Sale Price is largely dependent on the 1st Floor Area followed by 2nd Floor Area & Year Built. Hence, we predicted values by fitting a linear model by introducing interactions between the 3 mentioned important predictors. One major goal of the project was to get important deciding factors of house's sale prices. This has been achieved via two-fold process, first by EDA plots and secondly by modelling results. Area of the houses followed by year in which the houses were built seem to be important factors in deciding house prices for Ames county.

V. Limitations and Future Work

1. 'Neighborhood' variable though used for EDA purposes has not been used in building the predictive model. This variable can be added to the model in the future, as it can explain house prices variability due to location. This will though require some non-trivial treatment (like ***One-Hot Encoding)*** to this categorical variable.

2. If the data set would have had more observations, then some outlier treatment could have been done to finer tune the predictive results.

3. Some more features can be extracted like the name of the House Builder. This explains the good-will associated with the builder.