Executive Summary

This dataset involves 79 explanatory features (out of which 6 are used for EDA and modelling purposes) which are used in predicting house prices in IOWA state (AMES county) of US. The motive and goal of this project is to get insights and relationships amongst the variables which might prove to be useful in deciding factors responsible for deciding the house prices.

1. Relationship between the Variables and the target variable of Sale Price

The relationship was studied between 1st floor Area, 2nd floor area, Year in which house was built, Number of bedrooms and kitchens in order to predict the target variable, Sales Price of the house. The areas of the house were directly proportional to sales price, i.e. with increase in area of 1st and 2nd floors there was an increasing trend of house prices. This similar trend was showed by year built, where the newer houses followed an increasing sale prices as compared to their older counterparts. Number of bedrooms and kitchens did not show any significant trend. Moreover, most of the houses had a single kitchen and very few houses had more than 2 kitchens.

2. Any useful patterns found while studying tri-variate relationships?

Sale Price Vs Area and number of bedrooms- It also depicts the same story, where the trend is linearly increasing. For each bedroom count, sale price was almost linearly dependent (directly proportional) on the 1st Floor Area.

Sale Price Vs number of bedrooms and Year Built- Irrespective of no. of bedrooms the sale price shows an increase as year progresses with marginal difference across number of bedrooms. Hence number of bedrooms (and number of kitchens) do not seem to be a deciding factor in the house prices.

3. Which amongst these seem to be the most important feature in predicting the house prices.

The process of deciding the most important factor for house's sale prices was two-fold. As per EDA plots it seemed that 1st floor area was most strongly correlated with house sale price (followed by 2nd floor area). Even Year built showed an increasing relation, but the spread of the data was much more as compared to area. This showed that for the same year there was a wide range of house prices, though it could be said that mean house prices were increasing as year progressed.

Using this fact and the linear model results, another linear model was fitted with interactions amongst these 3 important features. It increased the R-squared hence better fitting the model and bolstering our claim that Area followed by Year Built of houses is most important deciding factor of houses' sale prices.