Exploratory Data Analysis for Housing Prices Data (Kaggle)

(\*\*Please refer to the data story named EDA\_House\_Prices in DataVisualisation.tbm file.)

**Feature Engineering:** To start with EDA, I engineered two variables using the following formula:

1. **Age\_House = IF [Yr Sold] - [Year Remod Add] > 0**

**THEN [Yr Sold] - [Year Remod Add]**

**ELSE [Yr Sold] - [Year Built] END**

1. **Total\_Area = [Gr Liv Area] + [BsmtFinSF1] + [BsmtFinSF2]**

**Bucketing Variables:** The dataset has close to 80 independent variables which comprise of different domain knowledge about various aspects of a house. In order to understand them more accurately, I created following buckets comprising of different variable types.



**Fig 1.1: Variables categorized into buckets representing the house characteristics**

Using the above buckets, I tried to find patterns in the data by using following dashboards and created a story around it to support my modelling techniques. The EDA was very crucial for my model as the dataset had less number of rows and imputing using mice package was not really helping me out with my accuracy.

Dashboards:

1. **Time trends:**
   1. The first thing about prices to know is how the prices vary over the years and time. The first dashboard says the SalePrice declines as the age of the house increase which means there is a negative correlation. This variable also played an important part in the linear regression model.
   2. Second graph shows about the monthly trends. It is evident from the graph that June and July have peak number of houses sold. Summer season seems to the peak season for buying houses may be because of the extreme weather conditions during winters.
2. **House Quality Index**:
   1. Bubble plot intuitively shows that with increase quality rank, the average house price increases indicating a positive linear relationship. Because of this relationship I used OverallQual and OverallCond as a numeric variable in my analysis and it improved my score.
3. **Exterior Conditions**:
   1. Neighborhood bubble plot shows how the market has moved over the past. What kind of societies have developed and reduced their prices over the past. In the beginning houses built in Old town had more average Sale Price but later Northridge height started developing and had more average sale price, but today stone brook is the only neighborhood getting developed with higher average sales price.
   2. OpenPorchSF and Enclosed Porch have less linear variation with the Sales Price but still I included them in the study as it has some economic significance.
   3. Areas with No fence has higher average sales prices. This may be because these are the house built in multistory societies and have no physical fences around their house. More information about the type of house based on urban or rural location would give a better picture.
4. **Room Types**:
   1. In a house there are mainly three most usable types of rooms namely bedrooms, kitchen and bathrooms. Price linearly and positively varies with total number of bed rooms above ground. Bedroom number 14 looks to be an outlier as the price is much lower at higher number of bedrooms.
   2. More houses have one bathroom and they tend to sell at a higher price.
   3. More houses have number of kitchens as 1.
5. **Basement Condition**:
   1. Total basement area has positive linear relationship with the Sale Price, but the data is highly skewed and hence box cox transformation is required to make the error normally distributed. Economically it makes sense as with more basement area, the house become a bit expensive.
   2. The data for basement quality looks very consistent as the average price for excellent basement condition is the highest.
   3. Basement Cond and Quality data richness is quite poor as most of the data points are around typical category.
6. **Garage Condition**:
   1. Average Sales Price increases for garages which were built recently. This trend makes sense as new garages comes with all sorts of automation which might have led to increase in price.
   2. Garage Condition and Garage Type are heavily skewed on typical category. Special attention would be required while replacing null values in this category.
7. **Aesthetic Appeal**:
   1. Price variation based on house styles are summarized in the tree map showing data richness as well. Mostly houses are either one story or two story.
   2. Price variation by Lot shape revealed that Regular and Slightly Irregular Lot shaped houses are sold more which economically makes sense as no one would like to buy an irregular house. I paid specially attention while imputing the missing values in variables bucketed under neighborhood in order to maintain consistencies across variables.
   3. Hip and Gabel are the two most famous roof style house being sold according to the data with wooden shakes as the primary roof material used.
8. **Correlation Analysis**:
   1. Correlation plot was plotted between numerical variables like LotFrontage, LotArea, GrvLivArea, Age\_House, TotalBsmtSF and observed good correlation to the target variables. Used all of them to make the model more accurate.

**Analytical Approach to the problem:**

1. Instead of using mice package, manually observed and imputed missing values for all variables except LotFrontage. Tried to maintain data consistency across each bucket as mentioned in the figure 1.1. For Ex: There were three entries in MasVnrArea which were marked as 1 which is not possible so imputed them with mean Area according to MasVnrType.
2. Saw data inconsistency in GarageYrBuilt as it has one value = 2207 which is not possible and hence imputed it with 2207.
3. Feeded many numerical variables as categorical like Garage Condition, Garage Quality etc.
4. Imputed values for LotFrontage using Mice Package (Ref: Prof Lanham’s Code shared through email).
5. With all the above techniques and applying simple linear regression, I achieved 968 Rank on Kaggle with an RMSE error of 0.12135.
6. To advance better in the accuracy I used the kernel: <https://www.kaggle.com/nafisur/top-10-0-10943-stacking-mice/notebook>. The code helped me achieve an RMSE of 0.10941.
7. Algorithms Used: Ensembling between xgboost, lightgbm and ridge regression, used stacked regression to reduce the errors by building models stacks and feeding errors back to the model with finetune coefficient of 15.