**TASKS**

**1. Hypothesis Building**

* When looking for a house the living area of the house seems like the most important feature to look for in a house and hence is expected to have a high correlation with SalePrice
* Other important features include the overall condition and quality of the house, the number of bathrooms, parkings in garage, bedrooms, basement and hence these features are also expected to have high correlation
* The SalePrice feature is expected to be right skewed as few houses are mansion like and have high price and are huge
* Similarly other features such as area are also expected to be right skewed
* Also the GrLivArea feature is expected to have a high positive correlation with SalePrice which would be apparent from a scatter plot
* There will be null values wherever features such as pool, fireplace or garage is missing

**2. EDA**

* Doing some EDA as shown in notebook verifies most of the hypotheses

**3. Data Cleaning**

* The cleaning is explained in the notebook
* Null values are delt by one of the following ways
  + Filling with ‘None’ - For most categorical variables
  + Filling mode – for some categorical variables
  + Filling median or 0 – For most quantitative variables

**4. Model Building**

1. Baseline model was made and removed
2. Various types of linear regression models were fit such as lasso, elasticnet and ridge regression
3. Residual Vs Fitted plot  is plotted at end and we find that the model overshoots the target variable as the residuals are skewed towards negative side
4. Polynomial regression of several higly target variable correlated features is used
5. K-fold cross validation is used

**5. Final Prediction**

Final predicted values are stored in submission.csv file along with ID of the house

**Rest of the code is explained as markdown in the notebook itself**