Kaggle late submission - **OK**

# Key Points I will take note of

* Location (e.g gr\_liv\_area), size (e.g garage\_area), time (e.g year\_built) – based on metrics/columns from dataframe
* Group them according to correlations

# EDA

* Numerical columns against SalePrice with Scatter Plot
  + Including identify potential categorical columns that can be changed to numerical format
  + Columns like ‘Exter Qual’ can be converted to numbers (train\_df['exter\_qual'].value\_counts() – can see distinctly 4 different categories)
* Categorical columns
  + Drop features like ‘utilities’ which only has 1 significant category.

In addition:

* Identify null values
* Identify lone rows
* Heatmap
* Subplots
* Grouping related variables
* Identify outliers

# Features and Regression

* Upon identifying potential X features with plots, conduct feature engineering
  + Use\_dummies for categories
  + Conduct outside research to identify key factors for sale price in housing at sale
* Perform Ridge, Lasso and ElasticNet with optimal alphas to discover correlation
* Lasso can be used to eliminate and reduce features to optimize X features
* PolynomialFeatures to further optimize X features to identify in depth correlation
* Identify which model is suitable and optimized