EDA on Advance House Price

The "shape" of the dataset shows that it has 1460 rows/instances, with data from 80 attributes. Out of the 81 attributes, one is the target variable (SalePrice) that the model should predict.   
Hence, there are 80 attributes that may be used for feature selection/engineering. Out of the 80 remaining attributes the dataset shows that there are 39 numerical attributes, with 42 categorical features. But from the distribution graphs of all initial numerical attributes:

The following has no linear relationship with the target variable SalesPrice: so they are converted to categorical variables MSSubClass, OverallCond,BsmtFullBath,BsmtHalfBath,FullBath,HalfBath,FirePlaces,GarageCars.so finally we have say like 31 numerical attributes which even depict size/measurement.

In order to get a better exploration of the dataset the following hypothesis was developed

* All numerical features are continuous variables
* The distribution of the target variable is normal.
* The distribution of other numerical features is not skewed
* All the numerical variables are correlated to the target variable only.
* There is no outliers found in the features of the dataset

The plot Distribution of Sales Price (FIG.1) shows that the SalePrice has a left skewed distribution and a high kurtosis which means most of the data falls to the lower range values.

Considering the distributions of other features in questions. features like BsmtUnfSf ,GrlVArea, BsmtSFinSF1,1StFlrSf data lie to the positive side of the distribution.Forexample YearBlt and GarageYearBlt are left skewed this makes sense for both features since most garage will be built the same time the house is built.Other features like LotFrontage ,LotArea ,MasVnrArea, GrlvArea shows possibility of outliers from the distribution graphs of all features(FIG.2) this could be confirmed using boxplot or scatterplots. Also features like MSSubClass, OverallCond,BsmtFullBath,BsmtHalfBath,FullBath,HalfBath, FirePlaces,GarageCars.I suspects the individual features values of the mention features are not linear with the target SalesPrice.The above feature could be divided into buckets(binning)to learn something different about housing Price values for each bucket individual buckets. Example we could check how group/categories(buckets) of BsmtFullBath,FullBath or both affects Salesprice.So those features will automatically threated as categorical varriables.

To check how numeric features are related to SalePrice.We start with the features that represent size.First we calculate the correlation coefficients between all these size related features and SalePrice, next these values are visualized in a heatmap.

From (Fig 3) SalePrice shows strong correlations with the features GrLivArea (0.71), GarageArea (0.62), 1stFlSF (0.61), TotalBsmtSF (0.61). MasVnrArea (0.48) shows a moderate correlation with Saleprice . The remaining features show weak () or very weak (PoolArea, ScreenPorch, 3SsnPorch, EnclosedPorch, BsmtFinSF2) correlations with SalePrice.

The scatterplot (Fig 4) above shows outliers in majority of the numerical features. The BOXPLOT(FIG 5) is a good way to see the outliers.Based on a first viewing of the scatter plots against SalePrice, there appears to be: A few outliers on the LotFrontage (say, >200) and LotArea (>100000) data. BsmtFinSF1 (>4000) and TotalBsmtSF (>6000),1stFlrSF (>4000) GrLivArea (>4000 AND SalePrice <300000), LowQualFinSF (>550).