SPRINGBOARD - DATA SCIENCE CAREER TRACK

# EXPLORATORY DATA ANALYSIS

# HOUSING PRICES- ADVANCED REGRESSION TECHNIQUES

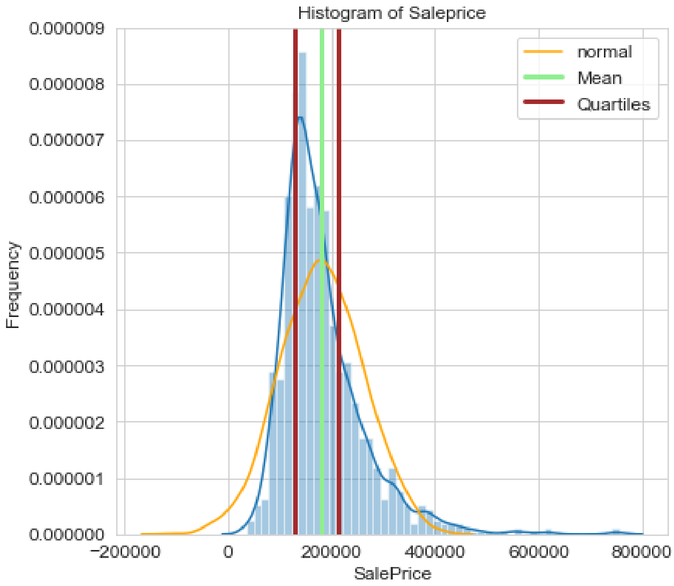
## INTRODUCTION

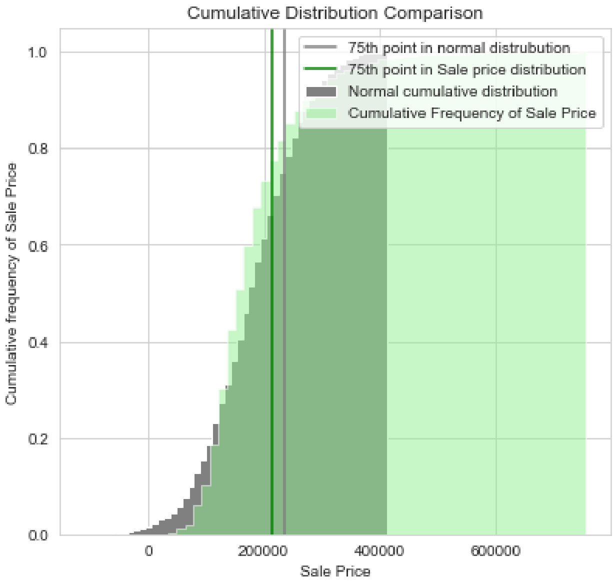
In this notebook, we will analyse the target variable and its relationship with independent variables. This will help in selecting predictive variables.

## TARGET VARIABLE - Sale Price

We will look at the following:

1. Distribution plot of SalePrice alongside kde plot of a normal distribution.
2. Develop a cumulative distribution of Sale Price alongside cumulative normal distribution.





Observations:

1.The distribution is not normal. Mean Sale Price and median Sale Price, both have lower probabilities of occurrence than the mode.

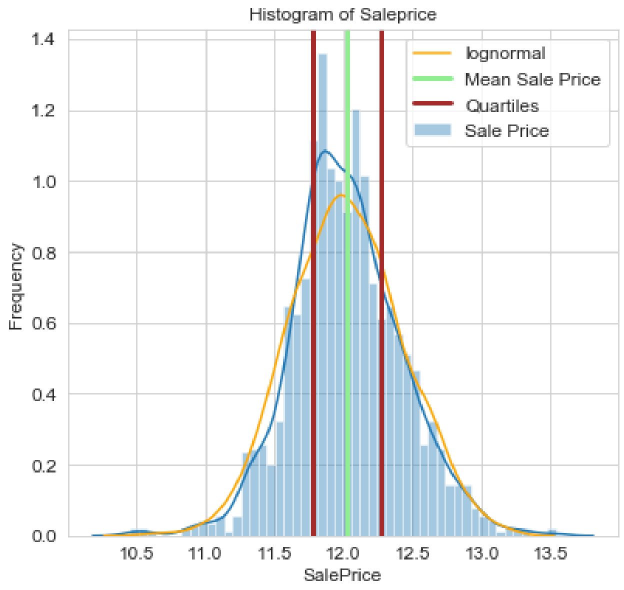
2. Distribution of SalePrice is leptokurtic. An example of a leptokurtic distribution is the Laplace distribution, which has tails that asymptotically approach zero more slowly than a Gaussian.

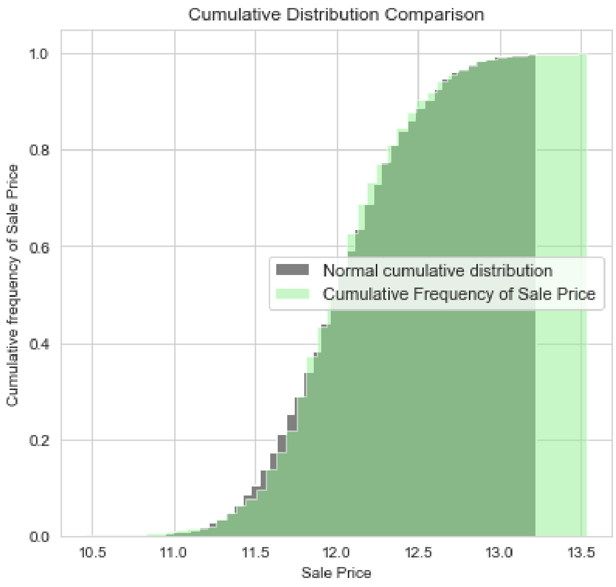
3. The distribution is right skewed. The range of upper 25% of data is around 60000 which is 3 times more than the Interquartile range.

4. Mean Sale Price is not a good representation and there are quite a number of outliers.

5. The cumulative distribution comparison makes it very clear that SalePrice does not conform to a normal pattern. There is also a greater chance of SalePrice in the range 100000 to 300000 occuring in comparison to what a normal distribution should have. We can also see that the range of upper 25% of Sale Price is abnormally greater than normal pattern.

## Log transformation of Target Variable 'SalePrice'





Observations:

The distribution of 'SalePrice' is very close to lognormal distribution.

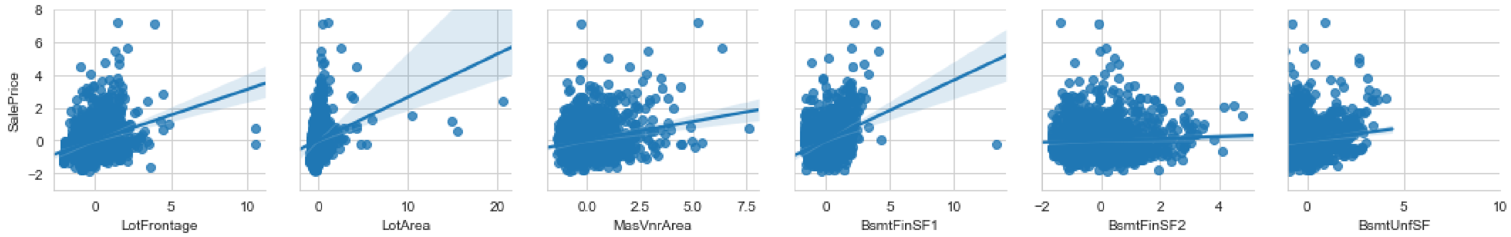
The tails are matching, though 'SalePrice' appears to be bimodal.

The range above the upper quartile has normalised to quite an extent.

The cumulative distribution of Sale Price is very close to cumulative lognormal distribution.

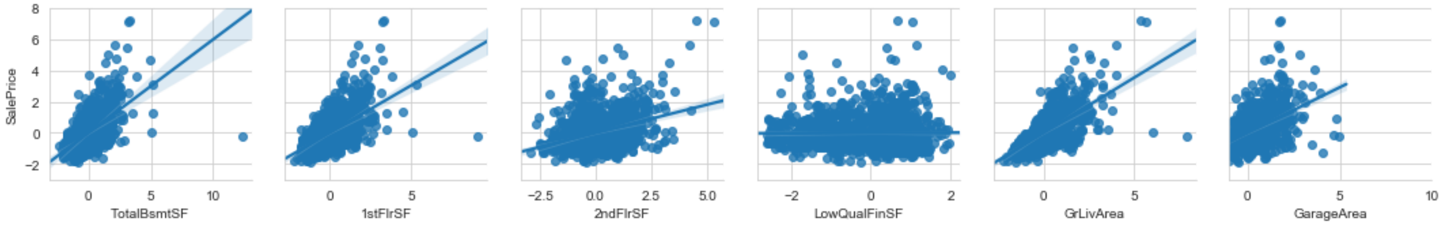
# 'SalePrice' Relationship with Independent numerical variables

## Pair Grid of Scatter plots of Numeric variable vs Sale Price:



**Observations:**

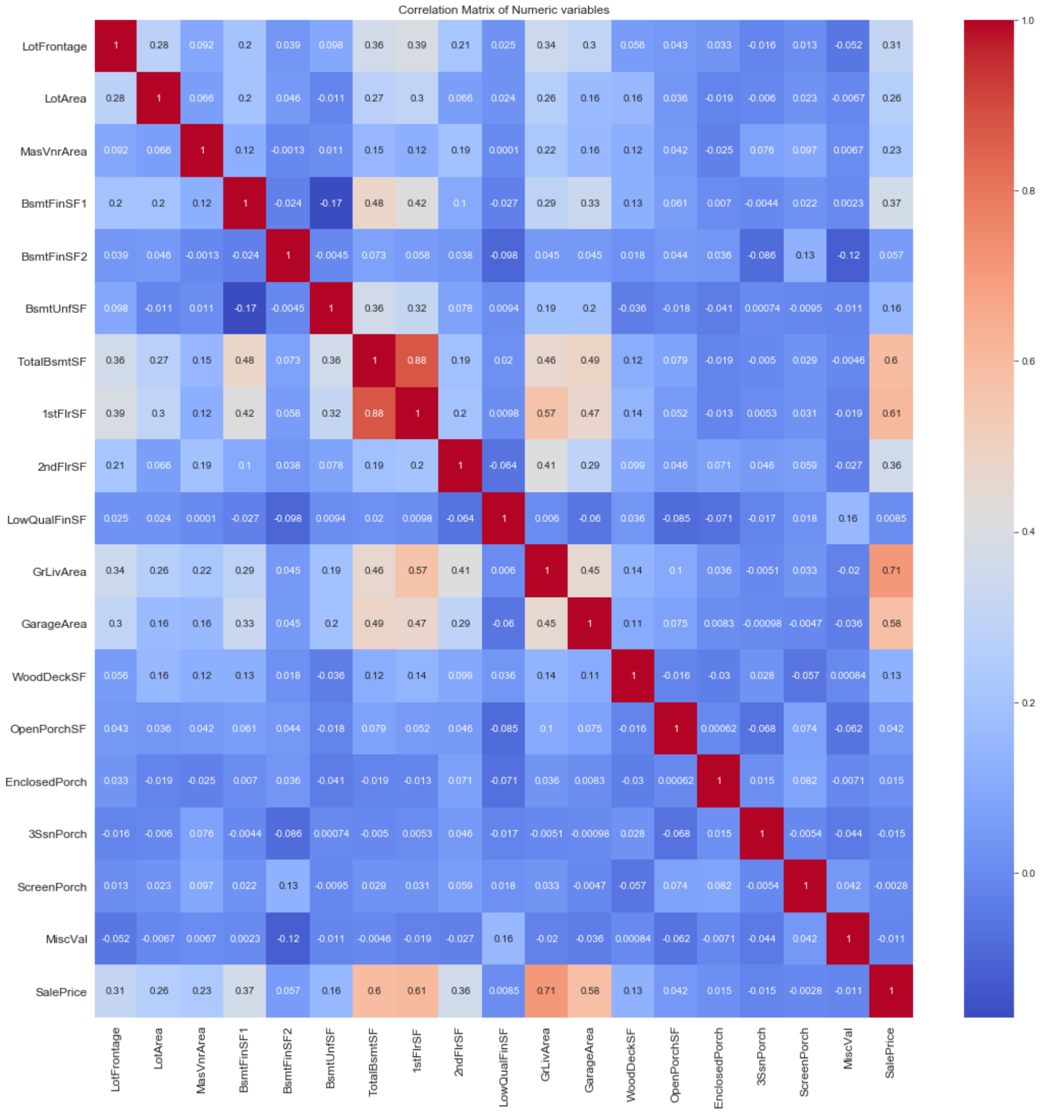
1. SalePrice has non constant variance with Variables LotFrontage, LotArea, MasVnrArea and BsmtSF1. Linear relationships cannot be clearly established.
2. BsmtUnfSF and BsmtFinSF2 both are weakly correlated with SalePrice.
3. SalePrice shows weak linear relationship with BsmtFinSF1 and has non constant variance.



**Observations**

1. TotalBsmtSF and FirstFlrSF have a correlation of .6 with SalePrice though relationship is somewhat linear. Variance in SalePrice increases as these independent variables increase , for eg we see a range from 100000 to 400000 with both these variables.Both these variables impact the average SalePrice. There is an outlier in both cases.
2. The variance in SalePrice is non constant with SecondFlrSF with a weak linear relationship. It has a .3 correlation with SalePrice.
3. GrLivArea and GarageArea have .7 and .58 correlation with SalePrice. SalePrice variance increases with higher values of GrLivArea and eventually spreads out of the group at few very high values. SalePrice has a constant variance, with GarageArea and here too it has outliers. GrLivArea shows some linearity with SalePrice.

## Multicollinearity between numeric variables as observed from Correlation Heatmap:



Correlation Matrix of Numeric variables

| | LotFrontage | LotArea | MasVnrArea | BsmtFinSF1 | BsmtFinSF2 | BsmtUnfSF | TotalBsmtSF | 1stFlrSF | 2ndFlrSF | LowQualFinSF | GrLivArea | GarageArea | WoodDeckSF | OpenPorchSF | EnclosedPorch | 3SsnPorch | ScreenPorch | MiscVal | SalePrice |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| LotFrontage | 1 | 0.28 | 0.092 | 0.2 | 0.039 | 0.098 | 0.36 | 0.39 | 0.21 | 0.025 | 0.34 | 0.3 | 0.056 | 0.043 | 0.033 | -0.016 | 0.013 | -0.052 | 0.31 |
| LotArea | 0.28 | 1 | 0.066 | 0.2 | 0.046 | -0.011 | 0.27 | 0.3 | 0.066 | 0.024 | 0.26 | 0.16 | 0.16 | 0.036 | -0.019 | -0.006 | 0.023 | -0.0067 | 0.26 |
| MasVnrArea | 0.092 | 0.066 | 1 | 0.12 | -0.0013 | 0.011 | 0.15 | 0.12 | 0.19 | 0.0001 | 0.22 | 0.16 | 0.12 | 0.042 | -0.025 | 0.076 | 0.097 | 0.0067 | 0.23 |
| BsmtFinSF1 | 0.2 | 0.2 | 0.12 | 1 | -0.024 | -0.17 | 0.48 | 0.42 | 0.1 | -0.027 | 0.29 | 0.33 | 0.13 | 0.061 | 0.007 | -0.0044 | 0.022 | 0.0023 | 0.37 |
| BsmtFinSF2 | 0.039 | 0.046 | -0.0013 | -0.024 | 1 | -0.0045 | 0.073 | 0.058 | 0.038 | -0.098 | 0.045 | 0.045 | 0.018 | 0.044 | 0.036 | -0.086 | 0.13 | -0.12 | 0.057 |
| BsmtUnfSF | 0.098 | -0.011 | 0.011 | -0.17 | -0.0045 | 1 | 0.36 | 0.32 | 0.078 | 0.0094 | 0.19 | 0.2 | -0.036 | -0.018 | -0.041 | 0.00074 | -0.0095 | -0.011 | 0.16 |
| TotalBsmtSF | 0.36 | 0.27 | 0.15 | 0.48 | 0.073 | 0.36 | 1 | 0.88 | 0.19 | 0.02 | 0.46 | 0.49 | 0.12 | 0.079 | -0.019 | -0.005 | 0.029 | -0.0046 | 0.6 |
| 1stFlrSF | 0.39 | 0.3 | 0.12 | 0.42 | 0.058 | 0.32 | 0.88 | 1 | 0.2 | 0.0098 | 0.57 | 0.47 | 0.14 | 0.052 | -0.013 | 0.0053 | 0.031 | -0.019 | 0.61 |
| 2ndFlrSF | 0.21 | 0.066 | 0.19 | 0.1 | 0.038 | 0.078 | 0.19 | 0.2 | 1 | -0.064 | 0.41 | 0.29 | 0.099 | 0.046 | 0.071 | 0.046 | 0.059 | -0.027 | 0.36 |
| LowQualFinSF | 0.025 | 0.024 | 0.0001 | -0.027 | -0.098 | 0.0094 | 0.02 | 0.0098 | -0.064 | 1 | 0.006 | -0.06 | 0.036 | -0.085 | -0.071 | -0.017 | 0.018 | 0.16 | 0.0085 |
| GrLivArea | 0.34 | 0.26 | 0.22 | 0.29 | 0.045 | 0.19 | 0.46 | 0.57 | 0.41 | 0.006 | 1 | 0.45 | 0.14 | 0.1 | 0.036 | -0.0051 | 0.033 | -0.02 | 0.71 |
| GarageArea | 0.3 | 0.16 | 0.16 | 0.33 | 0.045 | 0.2 | 0.49 | 0.47 | 0.29 | -0.06 | 0.45 | 1 | 0.11 | 0.075 | 0.0083 | -0.00098 | -0.0047 | -0.036 | 0.58 |
| WoodDeckSF | 0.056 | 0.16 | 0.12 | 0.13 | 0.018 | -0.036 | 0.12 | 0.14 | 0.099 | 0.036 | 0.14 | 0.11 | 1 | -0.016 | -0.03 | 0.028 | -0.057 | 0.00084 | 0.13 |
| OpenPorchSF | 0.043 | 0.036 | 0.042 | 0.061 | 0.044 | -0.018 | 0.079 | 0.052 | 0.046 | -0.085 | 0.1 | 0.075 | -0.016 | 1 | 0.00062 | -0.068 | 0.074 | -0.062 | 0.042 |
| EnclosedPorch | 0.033 | -0.019 | -0.025 | 0.007 | 0.036 | -0.041 | -0.019 | -0.013 | 0.071 | -0.071 | 0.036 | 0.0083 | -0.03 | 0.00062 | 1 | 0.015 | 0.082 | -0.0071 | 0.015 |
| 3SsnPorch | -0.016 | -0.006 | 0.076 | -0.0044 | -0.086 | 0.00074 | -0.005 | 0.0053 | 0.046 | -0.017 | -0.0051 | -0.00098 | 0.028 | -0.068 | 0.015 | 1 | -0.0054 | -0.044 | -0.015 |
| ScreenPorch | 0.013 | 0.023 | 0.097 | 0.022 | 0.13 | -0.0095 | 0.029 | 0.031 | 0.059 | 0.018 | 0.033 | -0.0047 | -0.057 | 0.074 | 0.082 | -0.0054 | 1 | 0.042 | -0.0028 |
| MiscVal | -0.052 | -0.0067 | 0.0067 | 0.0023 | -0.12 | -0.011 | -0.0046 | -0.019 | -0.027 | 0.16 | -0.02 | -0.036 | 0.00084 | -0.062 | -0.0071 | -0.044 | 0.042 | 1 | -0.011 |
| SalePrice | 0.31 | 0.26 | 0.23 | 0.37 | 0.057 | 0.16 | 0.6 | 0.61 | 0.36 | 0.0085 | 0.71 | 0.58 | 0.13 | 0.042 | 0.015 | -0.015 | -0.0028 | -0.011 | 1 |

Observations on correlation:

There are a lot of variables which have a correlation with each other.

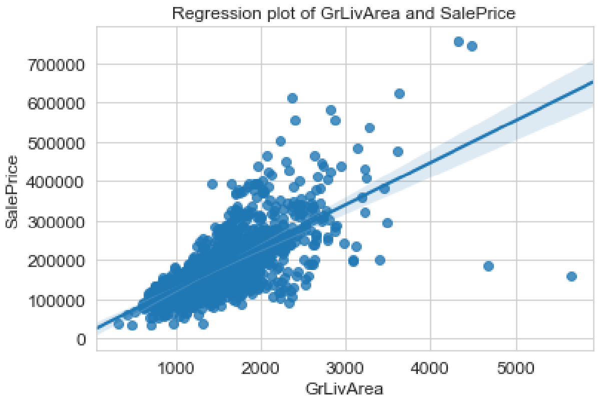
1. TotalBsmtSF nad 1stFlrSF have a high correlation with each other i.e .88.

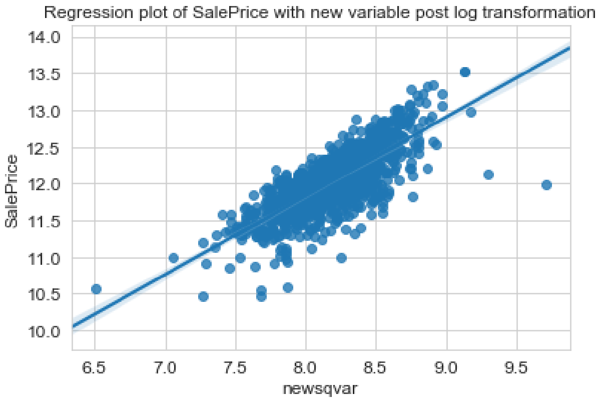
2. GrLivArea and  TotalBsmtSF have correlation of .61.

3. There is correlation between multiple variables- MasVnrArea, TotalBsmtSF, 1stFlrSF, GrLivArea and GarageArea

### Independent variable GrLivArea and its transformation:

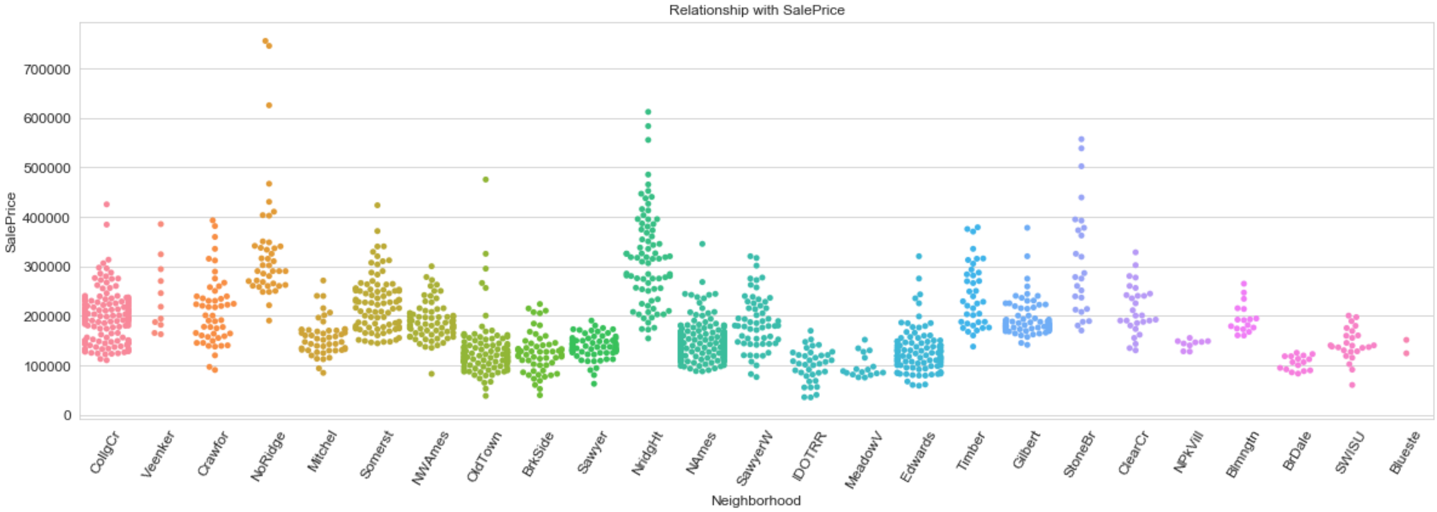
We have seen earlier that some variables have high correlation with each other and linear relation with SalePrice is also similar. On this basis we create a new variable- 'newsqvar' which is a combination of 'TotalBsmtSF','1stFlrSF' and 'GrLivArea'. Let's take the log value of this variable and see the regression plot with SalePrice. Great! This new variable has a strong linear relation with SalePrice and a constant variance as well. The correlation also has improved from .70 to .76

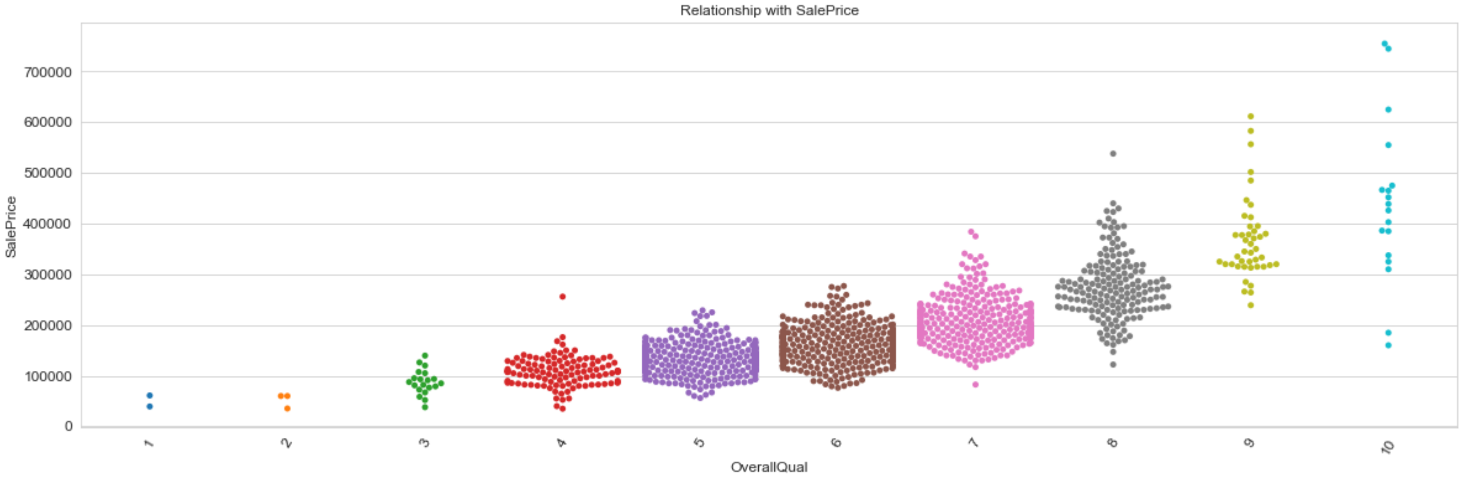




## <u>'SalePrice' Relationship with Independent categorical variables</u>

### Swarmplots of categorical variables with Sale Price:





**Observations with some categorical variables of interest:**

**Neighborhood**: This variable has 25 categories. Overall it looks like each category has a differentiated range in SalePrice.Most categories also have an almost equally distributed count. Few categories look similar in spread and average SalePrice like OldTown, BrkSide and Edwards.

**ExterQual**: The categories TA and Gd almost have an equal counts. The average SalePrice of Gd is higher than TA.

**ExterCond**: Category TA dominates in counts and has a higher average SalePrice, compared to Gd.

**Foundation**: The range in SalePrice and average SalePrice is differentiated in the 3 categories-PConc, CBlock, and BrkTil. The counts vary, with PConc count leading and followed by CBlock.

**BsmtQual**:The SalePrice range in the 3 categories- Gd,TA and Ex differs. The count of Gd category is the highest, followed by TA. The average SalePrice of Ex category is highest, followed by Gd category.

**BsmtFinType1**: The Average SalePrice and range differs in the 3 categories GLQ, ALQ and Unf, with GLQ leading and having almost equal counts with Unf.

**KitchenQual**: The categories Gd, TA and Ex have differentiated SalePrice range and Average price with Ex having the highest Average SalePrice and Gd having the highest counts.

**GarageFinish**:The 3 categories RfN, Unf and Fin have differentiated Avg Sale price, different price range and almost equal in counts.

**OverallQual**: The Average SalePrice and its range is increasing as the parameters increase from 1 to 10.Counts are mainly spread in categories 4 to 8.

**FullBath**: This variable has different Sale Price range and average SalePrice in 3 categories- 1,2, and 3, with 3 having the highest Avg SalePrice followed by category 2. Counts are highest for category 1 followed by category 2.

**HalfBath**: This variable has the highest category count of 0 and average Sale Price is higher for 1.

**BedroomAbvGr**: This variable has maximum counts spread between categories 2,3,4 with 3 being highest in count. Spread in Sale Price and Average SalePrice varies slightly between these categories.

**TotRmsAbvGrd**: The categories in this variable are differentiated in terms of range in Sale Price and Average Sale Price.

**GarageCars**: The categories 0,1,2,3 are well differentiated in terms of Average Sale Price and range in SalePrice. Category 2 has highest counts followed by 1.

**YearBuilt**: There is steady increase in Average Sale Price, as well as counts from the middle half of YearBuilt.

**YearRemodAdd**: There is an increasing trend in Average SalePrice over the years.

ZeroMasVnrArea, ZeroWoodDeckSF, ZeroBsmtFinSF1, ZeroOpenPorchSF and ZeroScreenPorch show lower Average Sale price with category '1'. Whereas ZeroEnclosedPorch and ZeroBsmtFinSF2 have higher Average Sale Price with category '1'.

## Transforming labels of categorical variables.

Basis our observations we find that there are categories, sensitive to Average Sale Price. For including these variables in the prediction models we need to assign them numerical labels. Here, have assigned the labels based on average SalePrice. There are also easy options available with OneHotEcoder and LabelEncoder for this job.