**EDA Summary on Housing Price Dataset**

**Observations:**

* The dataset contains 1460 observations with 81 variables.
* The target variable 'SalePrice' has a right-skewed distribution, indicating that most houses were sold for less than the median sale price.

|  |  |
| --- | --- |
| Table  Description automatically generated | Chart, histogram  Description automatically generated |

* 'OverallQual', and 'GrLivArea' are strongly correlated with 'SalePrice'.

|  |  |
| --- | --- |
| Table  Description automatically generated | Chart  Description automatically generated |

* There are some missing values in the dataset, particularly in the variables 'Alley', 'PoolQC', 'Fence', and 'MiscFeature'.

**Visualizations**

|  |  |
| --- | --- |
| Chart, scatter chart  Description automatically generated  Relationship between 'SalePrice' and 'GrLivArea' | Chart, box and whisker chart  Description automatically generated  Relationship between 'SalePrice' and 'OverallQual' |

**Linear Regression**

Using the LinearRegression() function from sklearn linear model, we have,

|  |
| --- |
| A picture containing text  Description automatically generated |

Hence, we can write the formula:

SalePrice = 107.13\*GrLivArea + 18569.03

This means for every square feet of above grade (ground) living area, ‘SalePrice’ goes up by 107.13 dollar, vice versa.

The interpretation of y-intercept does not make any sense in this case because we do not have any data with 0 ‘GrLivArea’.

**Conclusion**

This EDA gives a good understanding of the Housing Price dataset and the relationships between the dependent variable (‘SalePrice’) and other independent variables. A basic predictive model for the housing prices is also created using linear regression.