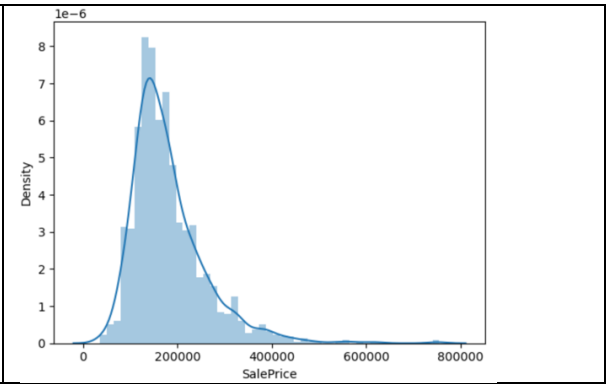
EDA Summary on Housing Price Dataset

Observations:

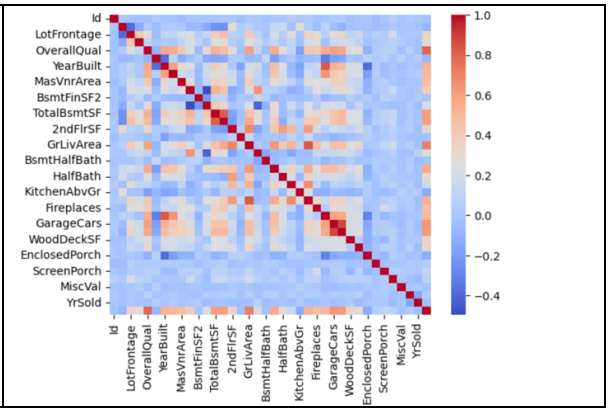
- The dataset contains 1460 observations with 81 variables.
- The target variable 'SalePrice' has a right-skewed distribution, indicating that most houses were sold for less than the median sale price.

	SalePrice
count	1460.000000
mean	180921.195890
std	79442.502883
min	34900.000000
25%	129975.000000
50%	163000.000000
75%	214000.000000
max	755000.000000



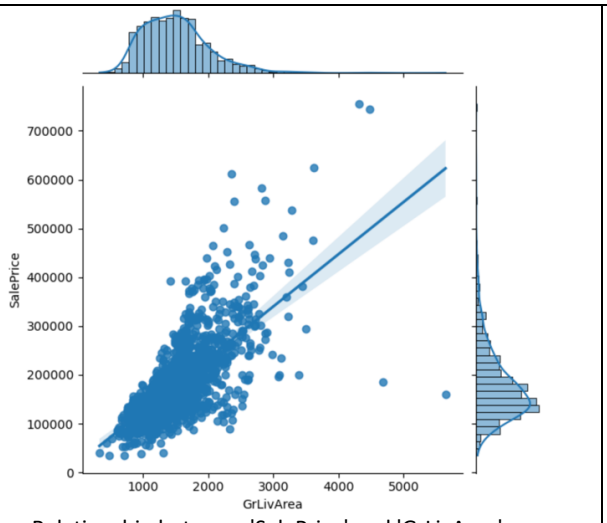
- 'OverallQual', and 'GrLivArea' are strongly correlated with 'SalePrice'.

	SalePrice
OverallQual	0.790982
GrLivArea	0.708624
SalePrice	1.000000

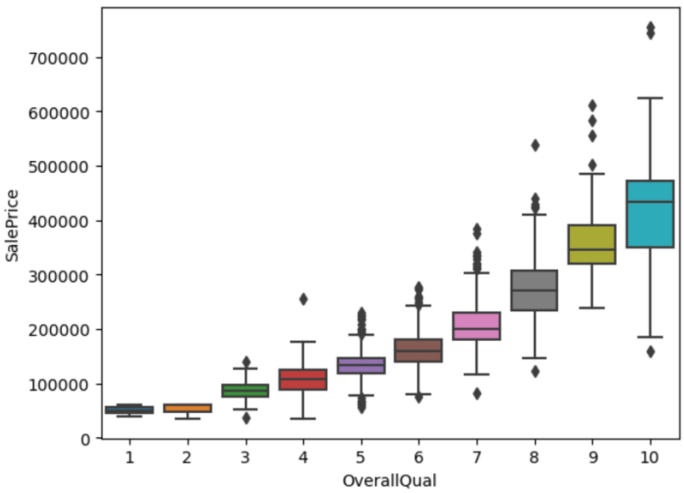


- There are some missing values in the dataset, particularly in the variables 'Alley', 'PoolQC', 'Fence', and 'MiscFeature'.

Visualizations



Relationship between 'SalePrice' and 'GrLivArea'



Relationship between 'SalePrice' and 'OverallQual'

Linear Regression

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

```
Coefficients: [107.13035897]
Intercept: 18569.025856487133
```

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.