EXERCISE-3

Exploratory data analysis and see if there is any correlation between the variables and analyze the distribution of the dataset.

1- To analyze the distribution of the variables YearBuilt and SalePrice

We can see that the mean sale price is around $180,921, with a standard deviation of $79,442. The minimum sale price is $34,900, and the maximum sale price is $755,000. The mean year of construction is 1971, with the oldest house being built in 1872 and the most recent house being built in 2010.

Histogram for 'YearBuilt': The histogram is created with 20 bins, with each bin representing a range of values for the 'YearBuilt' variable. The x-axis is labeled 'YearBuilt'. Depending on the shape of the histogram, as per observation it may suggest that the 'YearBuilt' variable is Left-Skewed.

Histogram for 'SalePrice': The histogram is created with 20 bins, with each bin representing a range of values for the 'SalePrice' variable. The x-axis is labeled 'SalePrice. Depending on the shape of the histogram, as per observation it may suggest that the 'SalePrice variable is Right Skewed.

Scatter plot between 'SalePrice' and SalePrice:

1- **The direction of the relation:** dots are concentrating or spreading from lower left to upper right, then the relation has a positive direction.
2- **The strength of the relation:** The dots are close (relatively) in the vertical direction, so it has a strong relation
3- **Presence of outliers:** here is one or more dots that are located significantly far away (in any direction) from the rest of the dots so yes, there is presence of outliers.
4- **Form:** As per the observation it is curvilinear relation.

Correlation coefficient for 'YearBuilt' and 'SalePrice' and used heatmap for visualization: The resulting correlation coefficient is a value between -1 and 1, where a value of 1 indicates a perfect positive correlation, a value of -1 indicates a perfect negative correlation, and a value of 0 indicates no correlation between the two variables. In this we get 2 values:

0.52 – It means low Positive Correlation.

1 - It means perfect Positive Correlation.

2- To analyze the distribution of the variables GarageArea and SalePrice

Histogram for 'GarageArea': The histogram is created with 20 bins, with each bin representing a range of values for the GarageArea variable. The x-axis is labeled GarageArea. Depending on the shape of the histogram, as per observation it may suggest that the GarageArea variable is Symmetrical.

Scatter plot between GarageArea and SalePrice:

1- **The direction of the relation:** dots are concentrating or spreading from lower left to upper right, then the relation has a positive direction.
2- **The strength of the relation:** The dots are close (relatively) in the vertical direction, so it has a strong relation.
3- **Presence of outliers:** here is one or more dots that are located significantly far away (in any direction) from the rest of the dots so yes, there is presence of outliers.
4- **Form:** As per the observation it is linear relation(as we can draw a straight line through it).

Correlation coefficient for 'GarageArea and 'SalePrice' and used heatmap for visualization: The resulting correlation coefficient is a value between -1 and 1, where a value of 1 indicates a perfect positive correlation, a value of -1 indicates a perfect negative correlation, and a value of 0 indicates no correlation between the two variables. In this we get 2 values:

0.62 – it means low Positive Correlation

1 - it means perfect Positive Correlation