For the correlation test and distribution in this study, the following variables have been selected:

- SalePrice: Sale Price
- LotArea: Lot Size in square feet
- YearBuilt: Original construction date
- GrLivArea: Above grade (ground) living area square feet
- GarageArea: Garage Area in square feet
- TotalBsmtSF: Total square feet of basement area

In this study, we want to conduct a Correlation Analysis and the distribution between the variables mentioned. For that, we use statistics and plots to obtain the results.

Sale Price Distribution Analysis:

Majority of Lower-Priced Homes: Most of the data points are concentrated towards the lower end of sale prices. This suggests that a larger proportion of the properties have lower sale prices.

Few High-Priced Homes: There are relatively fewer properties with very high sale prices, and these data points are spread out towards the right tail of the distribution.

Long Right Tail: The histogram's tail on the right side is longer and thinner, indicating that there are a few extreme values (very high sale prices) that extend far from the main concentration of data.

Positive Skewness: The skewness value (a measure of the asymmetry of the distribution) is positive, indicating a right skew. Positive skewness suggests that the mean is greater than the median, and the distribution is stretched to the right.

Potential Outliers: The right-skewed distribution may contain potential outliers, which are the high-priced properties that are distant from the main cluster of data.

Lot Area Distribution Analysis

Slight Positive Skewness: The histogram's shape is slightly right-skewed, indicating that the majority of the data points are clustered towards the lower end of lot areas. However, this skewness is not very pronounced, and there is a somewhat more even distribution of lot areas compared to a strongly right-skewed distribution.

Mean and Median Relationship: The fact that the median is very close to, but slightly less than, the mean suggests that the skewness is not extreme. In a perfectly symmetric distribution, the mean and median are equal. When the median is slightly less than the mean, it indicates a small degree of right skew, but not to the extent where the mean is significantly pulled to the right by a few extreme values.

Spread of Lot Areas: The spread of lot areas is somewhat balanced, with a slight tendency for larger lot areas. This means that while there are properties with larger lot areas, they are not significantly outnumbered by smaller lot areas.

Potential Outliers: There may be a few properties with very large lot areas, but their impact on the mean is not substantial, given that the median is close to the mean.

Correlation between Sale Price and Lot Area

The distribution of sale prices is right-skewed, as indicated by the following observations:

The mean (180,921.20) is greater than the median (50th percentile or 163,000.00).

The 25th percentile (129,975.00) is closer to the median than the 75th percentile (214,000.00), which is characteristic of right-skewed distribution.

The presence of several outliers on the right side of the distribution contributes to the right-skewness.

There is a wide range in sale prices, as suggested by the large standard deviation (79,442.50). This indicates that sale prices vary considerably.

The minimum sale price is 34,900.00, the maximum is 755,000.00, and there are several values that exceed the upper quartile value (outliers).

Lot Area:

Slight Right Skewness:

The distribution of lot areas is slightly right-skewed, as indicated by the mean (10,516.83) being slightly greater than the median (50th percentile or 9,478.50).

The presence of outliers on both the right and left sides of the distribution contributes to the slight right-skewness.

Central Tendency:

The median is 9,478.50, which suggests that about half of the lot areas fall below this value and half above.

The mean (average) is slightly higher, indicating that the distribution is pulled to the right by the presence of larger lot areas.

Variability:

The standard deviation (9,981.26) is relatively large, suggesting that the lot areas vary considerably.

Range of Lot Areas:

The minimum lot area is 1,300 square feet, and the maximum is 215,245 square feet, indicating a wide range of lot sizes in the dataset.

Outliers:

The presence of outliers on both sides of the distribution indicates that there are properties with very small lot areas (outliers to the left) and properties with very large lot areas (outliers to the right).

Interpreting the Outliers:

Outliers to the right may represent properties with exceptionally large lots, while outliers to the left may represent properties with very small lots.

Conclusion:

In summary, both variables, "Sale Price" and "Lot Area," exhibit right-skewed distributions with the majority of data concentrated towards the lower end. However, the degree of skewness varies, with "Sale Price" showing a more pronounced right-skew and a wider range of values. Both variables also have outliers that contribute to their respective skewness and suggest the presence of properties with both small and large values.

The association between both variables are not strong as is described in the Scatterplot and the Correlation Coefficient.

2) Analysis of the relationship between Sale Price and Year Built

Analysis of the distribution and relationship between Price Sales and Year Built

Year Built Histogram (Left Skewed):

The left-skewed histogram for "YearBuilt" indicates that the majority of properties in the dataset were constructed more recently, with fewer properties having earlier construction dates.
This means that there is a concentration of properties with relatively newer construction dates and a tail extending to the left with fewer older properties.
Correlation Coefficient (0.523):

The Pearson correlation coefficient of 0.523

Suggests a moderately strong positive linear relationship between "SalePrice" and "YearBuilt."
A positive correlation coefficient indicates that, in general, as the year of construction increases (properties are newer), the sale price tends to be higher.
The positive correlation coefficient aligns with the left-skewed histogram, as the majority of higher-priced properties are relatively more recently constructed.

Year Built Whisker Plot (Left Skewed with Outliers to the Left):

The left-skewed whisker plot and the presence of outliers to the left confirm the left-skewed nature of the "YearBuilt" variable.
The outliers to the left suggest the existence of older properties with either very high or very low sale prices compared to the majority of properties in the dataset.

3) Analysis of the relationship between Sale Price and Ground Living Area

Analysis of the distribution and relationship between Price Sales and Ground Living Area

Ground Living Area Histogram (Slightly Right Skewed):

The slightly right-skewed histogram for "GrLivArea" indicates that the majority of properties have ground living areas concentrated towards the lower end, with a tail extending to the right for larger living areas.
This suggests that most properties have relatively smaller ground living areas, with some properties having significantly larger living spaces.

Correlation Coefficient (0.708):

The Pearson correlation coefficient of 0.708 indicates a relatively strong positive linear relationship between "SalePrice" and "GrLivArea."
A positive correlation coefficient implies that, on average, as the ground living area increases, the sale price tends to be higher.
The strong positive correlation aligns with the slightly right-skewed histogram, as the majority of higher-priced properties tend to have larger living areas.

Ground Living Area Whisker Plot (Slightly Right Skewed with Outliers to the Right):

The slightly right-skewed whisker plot and the presence of outliers to the right confirm the slightly right-skewed nature of the "GrLivArea" variable.
The outliers to the right suggest the existence of properties with very large ground living areas, which contribute to the right skewness.

Analysis of the distribution and relationship between Price Sales and Garage

Garage Area Histogram (Right Skewed):

The right-skewed histogram for "GarageArea" indicates that the majority of properties have garage areas concentrated towards the lower end, with a tail extending to the right for larger garage areas.
This suggests that most properties have relatively smaller garage areas, with some properties having significantly larger garage spaces.

Correlation Coefficient (0.623):

The Pearson correlation coefficient of 0.623 indicates a moderately strong positive linear relationship between "SalePrice" and "GarageArea."
A positive correlation coefficient implies that, on average, as the garage area increases, the sale price tends to be higher.
The moderately strong positive correlation aligns with the right-skewed histogram, as the majority of higher-priced properties tend to have larger garage areas.

Garage Area Whisker Plot (Right Skewed with Several Outliers to the Right):

The right-skewed whisker plot and the presence of outliers to the right confirm the right-skewed nature of the "GarageArea" variable.
The outliers to the right suggest the existence of properties with very large garage areas, which contribute to the right skewness.

Analysis of the distribution and relationship between Price Sales and Total Basement Area

Total Basement Area Histogram (Slightly Right Skewed):

The slightly right-skewed histogram for "TotalBsmtSF" indicates that the majority of properties have total basement areas concentrated towards the lower end, with a tail extending to the right for larger basement areas.
This suggests that most properties have relatively smaller basement areas, with some properties having significantly larger basement spaces.

Correlation Coefficient (0.613):

The Pearson correlation coefficient of 0.613

Suggests a moderately strong positive linear relationship between "SalePrice" and "TotalBsmtSF."
A positive correlation coefficient implies that, on average, as the total basement area increases, the sale price tends to be higher.
The moderately strong positive correlation aligns with the slightly right-skewed histogram, as the majority of higher-priced properties tend to have larger basement areas.

Total Basement Area Whisker Plot (Slightly Right Skewed with Outliers):

The slightly right-skewed whisker plot and the presence of some outliers to the right and a few to the left confirm the slightly right-skewed nature of the "TotalBsmtSF" variable.
The outliers to the right suggest the existence of properties with very large basement areas, which contribute to the right skewness. The presence of outliers to the left may represent properties with smaller basement areas compared to the majority.