Summary of Assignment 3 Part 3

Exploratory data analysis using House Price Dataset from Kaggle

Introduction:

For the analysis on **House Price Dataset**, I have analyzed the correlation between the variables and further analyzed the distribution of the dataset. Thus, I am submitting my findings below:

Steps:

Initially, the required libraries were imported in Jupyter notebook such as NumPy, Pandas, Matplotlib and Seaborn for statistics and plots. Then, house price dataset was loaded, moving on the ".head()" was used to observe the first few rows of the dataset.

EDA to check if there is any correlation between the variable:

The correlation matrix is visualized as a heatmap using the seaborn library. By analyzing the heatmap, the correlation between variables can be observed. For instance, the variables with the **highest correlation** to SalePrice (the target variable) are OverallQual, GrLivArea, and GarageCars. This suggests that these variables might have a **significant impact** on the sale price of a property. Additionally, there are several variables that are highly correlated with each other, such as GarageArea and GarageCars, TotalBsmtSF and 1stFlrSF, and YearBuilt and GarageYrBlt.

Analyzing the distribution of the dataset:

To analyze the distribution of the dataset, I have used histogram of five variables to observe the variety in the dataset. Hereby, I am sharing the insights I have gathered from the histograms below:

1. Histogram for LotFrontage variable: This histogram represents the linear feet of street connected to the property. The histogram shows a slightly **right-skewed distribution**, with the majority of the properties having a lot frontage between 50 to 100 feet.
2. Histogram for LotArea variable: This histogram represents the lot size in square feet. The histogram shows a heavily **right-skewed distribution**, with the majority of the properties having a lot of area less than 20,000 square feet.
3. Histogram for YearBuilt variable: This histogram represents the original construction date of the property. The histogram shows a **left-skewed distribution**, with the majority of the properties being built after 1950.
4. Histogram for GrLivArea variable: This histogram represents the above grade (ground) living area square feet. The histogram shows a **normal distribution**, with the majority of the properties having a living area between 1,000 to 2,000 square feet.
5. Histogram for SalePrice variable: This histogram represents the property's sale price in dollars. The histogram shows a heavily **right-skewed distribution**, with the majority of the properties sold for less than $400,000.