EDA on House Price Dataset

Introduction
The goal of this analysis is to predict **Sale Price** of houses in Ames, Iowa, using **79 explanatory variables** describing structural, locational, and neighborhood aspects of the properties. The dataset contains **1,460 houses and 80 features**.

Feature Overview
Numerical Features (38)

- **Property & Structural Details:** LotFrontage, LotArea, YearBuilt, YearRemodAdd
- **Basement & Living Area:** BsmtFinSF1, TotalBsmtSF, 1stFlrSF, GrLivArea
- **Rooms & Amenities:** TotRmsAbvGrd, Fireplaces, FullBath, BedroomAbvGr
- **Garage & Outdoor Features:** GarageCars, GarageArea, WoodDeckSF
- **Target Variable: SalePrice (House sale price)**

Categorical Features (43)

- **Zoning & Location:** MSZoning, Neighborhood, Condition1
- **House Design & Construction:** MSSubClass, HouseStyle, RoofStyle
- **Basement & Garage Features:** BsmtQual, BsmtCond, GarageType, GarageFinish
- **Utilities & Sales Info:** HeatingQC, Electrical, SaleType, SaleCondition

Note: Some numerical columns represent ordinal data and were converted into categorical, leading to **25 numerical** and **56 categorical** features.

Missing Values:

- Here, PoolQC, FireplaceQu, Fence, MiscFeature, Alley are with the missing percentage >30%. So we can consider to drop off the columns as they're not giving potential valuess.
- Here, LotFrontage, GarageType, GarageYrBlt, GarageFinish,GarageQual, GarageCond are having the missing percentage > 5%. So we will impute the values either median (as skewed) and mode for categorical.
- Otherwise the columns with less than 5% missing percentage, we will drop off the rows with missing values.

Key Insights from Outliers Boxplot:

- Right-skewed trends – LotArea, TotalBsmtSF, GrLivArea, GarageArea, and SalePrice all show right-skewed distributions with outliers, meaning most homes fall within a typical range, but a few luxury or larger properties push the extremes.
- Older vs. Newer Homes – YearBuilt shows a mix of historic and modern homes, with a mid-20th century peak, indicating steady development over time.
- Common vs. Premium Features – Many homes lack masonry veneer, and moderate-sized basements, garages, and living areas are the norm. However, outliers suggest some high-end properties with extensive features.

Key Insights from Histograms

1. **Right-Skewed Distributions:** Most numerical variables (e.g., Lot Area, TotalBsmtSF, GrLivArea) have **right-skewed** distributions with a few large values.
2. **Property Size Trends:** Most homes have **LotFrontage (50-100 ft)** and **LotArea (5,000-15,000 sq. ft.)**.

3. **House Age & Renovations:**
    - Most homes were **built between 1950-2000**, peaking in the **1970s-1980s**.
    - Remodels surged between **1980-2010**, peaking in the **early 2000s**.
4. **Living Space Trends:**
    - Most homes have **1,000-2,500 sq. ft. of living space**.
    - Single-story homes are common, but **two-story houses have varying sizes**.
5. **Seasonality & Sales:**
    - **Peak sales occur in June & July**, while winter months (January & December) have **low sales**.
6. **Pricing Insights:**
    - Most homes are priced **between $100,000 - $300,000**, peaking at **$150,000 - $200,000**.

Key Insights from Bar Charts

1. **Market Trends & Home Design**
    - **Single-family homes (MSSubClass 20)** dominate the dataset.
    - Most homes are **one or two-story**, with **gable roofs** and **vinyl siding**.
2. **Structural & Functional Features**
    - **3-bedroom homes** are most common, with **1-2 full bathrooms**.
    - Most homes have **average (TA) kitchen & heating quality**.
3. **Basement & Garage Insights**
    - Majority have **attached garages (2-car)** with **unfinished or partially finished basements**.
4. **Outdoor Features**
    - **Most homes lack pools, decks, or porches**, but paved driveways are **common**.
5. **Sale & Pricing Trends**
    - Most homes are sold through **Warranty Deeds (WD)** under **normal market conditions**.
    - **High sale prices are associated with summer sales.**

Key Insights from Correlation Matrix

1. **House Quality is the Best Predictor of Sale Price:**
    - **OverallQual vs. SalePrice (0.8):** Higher quality materials increase sale price.
2. **Bigger Homes Sell for More:**
    - **GrLivArea (0.7), TotalBsmtSF (0.6), and TotRmsAbvGrd (0.6)** all show positive correlations with SalePrice.
3. **Garages & Basements Matter:**
    - **GarageCars (0.64) and GarageArea (0.62)** show that larger garages add significant value.
4. **Renovations & Newer Homes Sell for Higher Prices:**
    - **YearBuilt (0.6) & YearRemodAdd (0.5)** show that newer and remodeled homes attract higher prices.
5. **Seasonality Impacts Sale Price:**
    - Homes sold in **summer (June & July)** tend to sell for **higher prices**.

Key Insights from Bivariate Analysis:

- **Scatterplot - Sales price vs living area showing that** larger house tends to have higher sale price.
- **Bar plot for average SalePrice by OverallCond showing that** poor-condition houses (1-3) have the lowest prices, houses with average condition (5-6) sell for prices similar to those in good condition (7-8). This suggests that buyers prioritize other factors (like house quality, size, and location) over condition alone.

Hypothesis Testing

1. **Houses with higher quality ratings (OverallQual) tend to have significantly higher sale prices.**
    - **Bar Chart Interpretation:** Quality ratings **5-8 show a positive trend** in average sale prices, supporting this hypothesis.
    -
2. **Houses sold in summer months (May - July) tend to have higher average sale prices compared to winter months (December - February).**
    - **Bar Chart Interpretation: June & July show higher average sale prices**, while **December & January have lower prices**, confirming seasonality impacts.

Conclusion

- **Sale prices are driven by house quality, size, garage & basement spaces, and renovations.**
- **Market seasonality affects pricing**, with summer months having higher sales prices.
- **Future modeling should focus on OverallQual, GrLivArea, and Garage features** as key predictors of SalePrice.