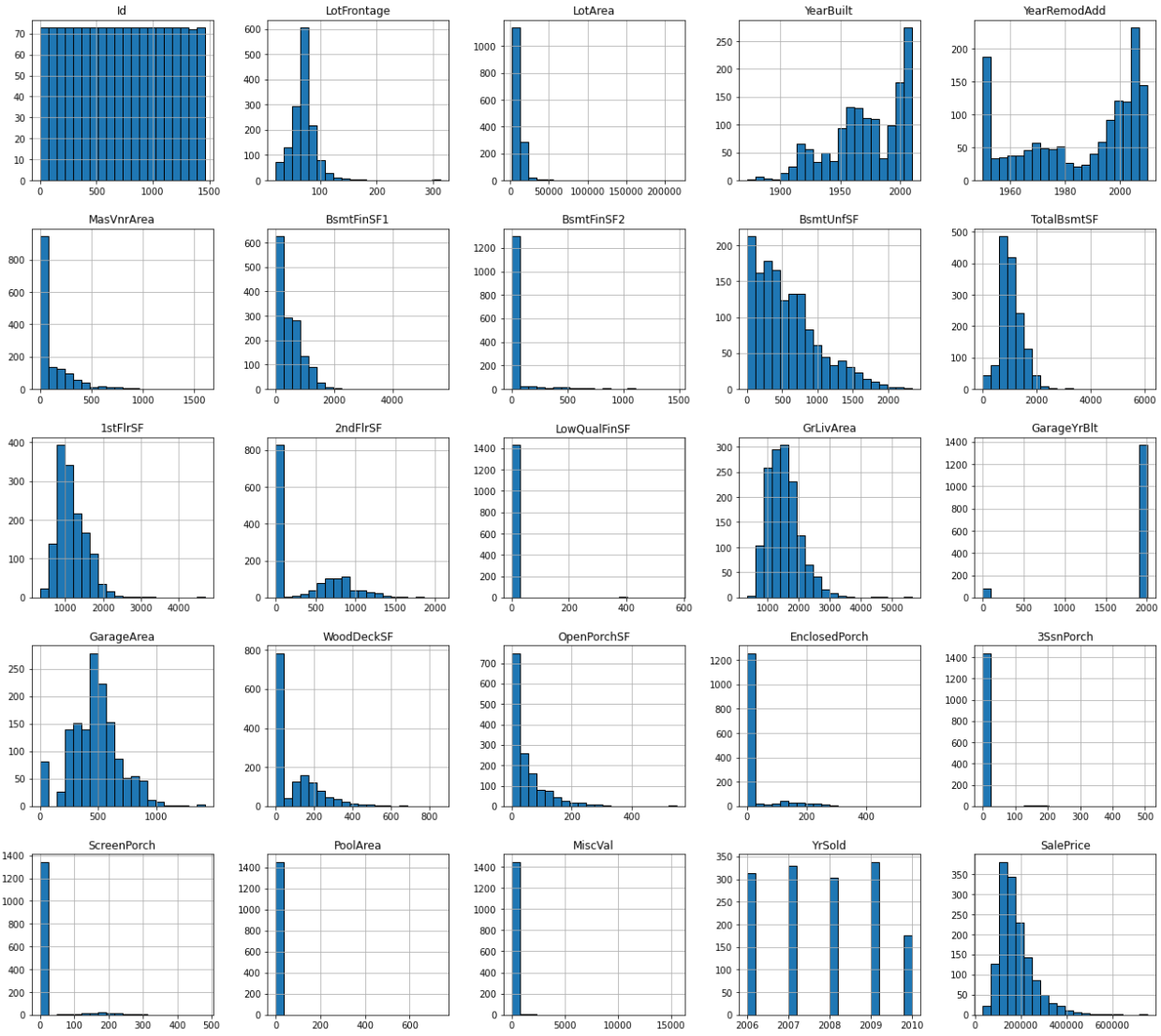
# EDA Results on House Price Dataset

## Histogram of Numerical Variables:

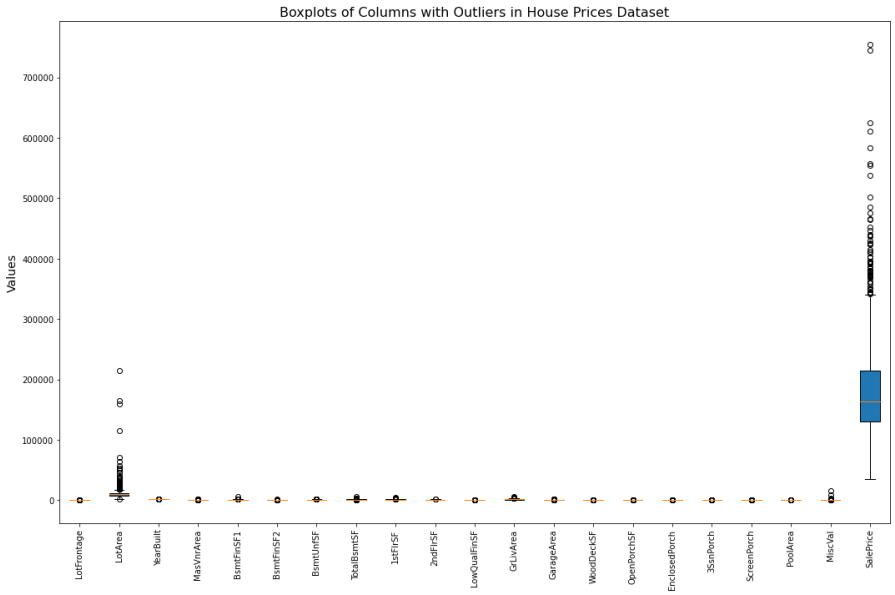



## Interpretation:

1. **Right-Skewed Distributions**
   - Most numerical variables (e.g., Lot Frontage, Lot Area, TotalBsmtSF, GrLivArea, GarageArea) are right-skewed, meaning most values are concentrated at the lower end, with fewer extreme values at the higher end.
   - This suggests that larger properties, bigger garages, and expansive living areas are relatively rare.

2. **Lot & Property Size Trends**
    - Lot Frontage: Most houses have 50-100 feet frontage.
    - Lot Area: Majority fall between 5,000-15,000 sq. ft., peaking around 10,000 sq. ft.
    - TotalBsmtSF & 1stFlrSF: Most houses have a basement between 500-1,500 sq. ft. and a first floor between 800-2,000 sq. ft.
3. **Age of Houses & Remodeling Trends**
    - Most houses were built between 1950 and 2000, peaking in the 1970s-1980s.
    - Remodeling activity surged between 1980 and 2010, peaking in the early 2000s.
4. **Basement Areas & Utilization**
    - BsmtFinSF1: Many houses have finished basement spaces of 0-1,000 sq. ft..
    - BsmtFinSF2: Most houses either have no second finished basement area or very small ones.
    - BsmtUnfSF: Many homes have some unfinished basement space, but mostly less than 1,000 sq. ft.
5. **Garage & Driveway Characteristics**
    - Garage Area: Most garages are 200-800 sq. ft., peaking at 400-600 sq. ft..
    - Garage Year Built: Many garages were constructed around 2000, with older garages decreasing in frequency.
    - Driveway: Paved driveways are the standard feature.
6. **Porches, Decks, and Miscellaneous Features**
    - Most houses do not have wood decks, screen porches, enclosed porches, or pools.
    - For houses with these features, their sizes tend to be modest (100-200 sq. ft.).
7. **Living Space Trends**
    - Above-Ground Living Area (GrLivArea): Most houses have 1,000-2,500 sq. ft., with a peak at 1,500-2,000 sq. ft.
    - Second Floors (2ndFlrSF): Many houses are single-story with no second floor. Among those with a second story, sizes vary widely.
8. **Seasonality & Sales Trends**
    - Houses are most frequently sold in June & July and least in January & December.
    - Most sales occur under normal conditions (non-distressed transactions).
    - Warranty Deed (WD) is the most common sale type, indicating traditional home sales.
9. **Pricing & Market Trends**
    - Sale Prices are right-skewed, with most houses priced between $100,000 - $300,000.
    - A peak occurs around $150,000-$200,000, indicating a concentration of homes in this price range.
    - Higher-priced homes are much less frequent.

**Boxplots of Outliers of Numerical Variables:**




**Interpretation:**

1. **Lot Area:** Right-skewed with several outliers. This indicates that most houses have lot areas around the median, but there are a few properties with significantly larger lot areas.
2. **LotFrontAge**: Right-skewed with many outliers (very large lots).Most houses are on smaller lots, but there are a few properties with very large lot areas, which could be estates or rural properties.
3. **YearBuilt**: Range:Typically 1870 to 2020. Slightly left-skewed (more newer homes). The dataset includes a mix of older and newer homes, with a concentration of homes built in the mid-20th century.
4. **MasVnrArea (Masonry Veneer Area):** Median: Around 0 (many homes have no masonry veneer). Right-skewed with outliers. Most homes have little to no masonry veneer, but a few have significant amounts, which could indicate higher-end construction.
5. **TotalBsmtSF (Total Basement Square Feet):** Most homes have moderate-sized basements, but a few have very large basements, which could indicate larger or more luxurious homes with distribution of Right-skewed with outliers
6. **GrLivArea (Above-Grade Living Area):** Most homes have moderate living areas, but a few have very large living spaces (right skewed with outliers), which could indicate larger or more luxurious homes.
7. **GarageArea:** Most homes have moderate-sized garages, but a few have very large garages (right skewed with outliers), which could indicate luxury homes or homes with multiple cars.
8. **SalePrice:** Most homes are moderately priced, but there are a few high-priced outliers, which could indicate luxury properties.

**Barcharts for Categorical Variables:**

**1. Property Characteristics & Zoning**

- Single-family homes (MSSubClass 20) dominate the dataset, with most properties falling under Residential Low-Density zoning (RL).
- The majority of houses are built on level land (LandContour: Lvl) and regularly shaped lots (LotShape: Reg).

**2. Structural Features & Home Design**

- One-story (1Story) and two-story (2Story) homes are the most common.
- Houses predominantly have gable or hip-style roofs, with composite shingle (CompShg) as the most common roofing material.
- The most frequently used exterior siding materials include vinyl and wood siding.

**3. Basement, Garage & Parking Trends**

- Most homes have unfinished or partially finished basements with standard basement conditions and limited exposure (BsmtExposure: No).
- Attached garages are the most common type, and most homes have 2-car garages.
- Garage quality and condition are mostly rated as average (TA).

**4. Interior Features & Functional Spaces**

- 3-bedroom homes are the most common, with kitchens and functional spaces rated as average or good.
- The above-ground living area (GrLivArea) is mostly between 1,000 and 2,500 sq. ft., meaning homes have moderate living spaces.
- Most houses have 1-2 full bathrooms, with half-baths being less common.

**5. Heating, Cooling & Utility Access**

- Most homes have gas heating (GasA) and central air conditioning (CentralAir: Y).
- All houses have access to public utilities (AllPub), ensuring standard infrastructure availability.

**6. Outdoor Features & Landscaping Trends**

- Most homes do not have pools, wood decks, or enclosed porches.
- When present, porches and decks are relatively small, typically ranging between 100-200 sq. ft.
- Most houses have paved driveways (PavedDrive: Y).

**7. Market Trends & Sales Patterns**

- Peak home sales occur in June and July, while winter months (January, February, December) see fewer transactions.
- Most homes are sold through normal sale conditions (SaleCondition: Normal), meaning no foreclosures or distress sales.
- The most common sale type is a Warranty Deed (WD), reflecting standard home transactions.

**8. House Quality & Condition Ratings**

- Most homes have an overall quality rating between 5-7, meaning moderate to good construction.
- Overall condition ratings are centered around 5-6, with very few homes in excellent or poor condition.
- Exterior and heating quality ratings are mostly average (TA).

**9. Lot & Land Features**

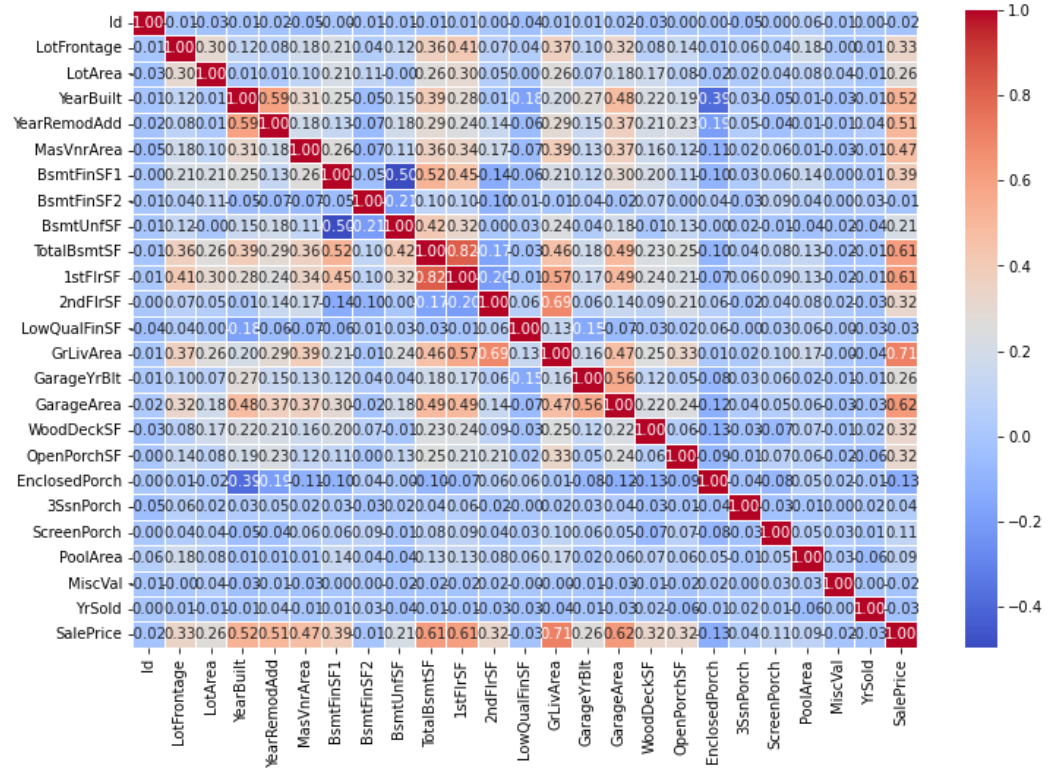
- Most properties have a lot frontage between 50-100 feet, indicating moderate-sized properties.
- Lot area distribution is right-skewed, with most properties ranging between 5,000-15,000 sq. ft.
- Corner and cul-de-sac lots are less common, with most houses being in inside lots (LotConfig: Inside).

**10. Pricing & Home Value Insights**

- Sale prices are right-skewed, with most homes priced between $100,000 and $300,000.
- The most common price range is between $150,000 and $200,000, indicating that affordable and mid-range homes are more frequent.
- Fewer homes are in the high-end luxury segment, suggesting that the dataset primarily includes middle-class properties

**Note**: **56** barcharts are being done. So only sharing the interpretations based on the bar charts.

**Correlation Matrix insights:**




**Strong Positive Correlations:**

- OverallQual and SalePrice (0.8): Higher quality construction and materials significantly increase the sale price of houses.
- GrLivArea and TotRmsAbvGrd (0.8): Larger above-ground living areas are associated with more rooms. Properties with more rooms typically have higher living areas, and this can be a key indicator of house size.
- GarageCars and GarageArea (0.85): The size of the garage and the number of cars it can accommodate are closely related. Both features contribute significantly to property value and can be used interchangeably in certain analyses.
- TotalBsmtSF and 1stFlrSF (0.8): Larger basements are often paired with larger first floors, indicating that homes with spacious basements typically have more substantial ground floors as well.

**Weak Correlations:**

- YrSold and SalePrice (0.0): This suggests that sale prices are more influenced by other factors such as quality and size, rather than the year of sale.
- LowQualFinSF and GrLivArea (0.04): Low-quality finished square footage has a minimal correlation with the overall ground living area, indicating that low-quality finishes do not substantially impact the total living area.

- BsmtFinSF2 and TotalBsmtSF (0.18): The second finished basement area has a weak correlation with the total basement area, suggesting that not all houses utilize their entire basement space for finished areas.

**Negative Correlations:**

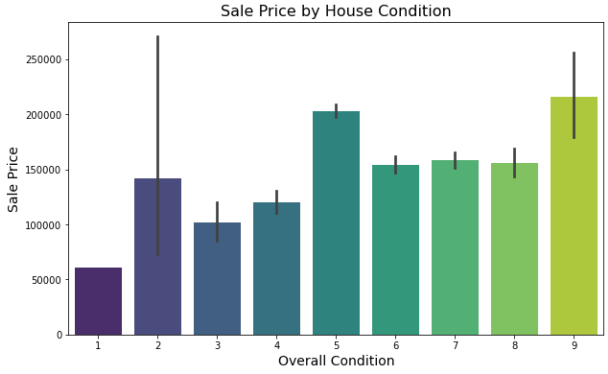
- OverallQual and Age (-0.2): Newer houses tend to have higher overall quality ratings, indicating that construction standards and materials have likely improved over time.
- GarageYrBlt and Age (-0.4): Newer garages are typically associated with newer houses, and older garages correlate with older houses. This relationship reflects the improvements in garage construction over the years.

**Bivariate Analysis:**

- **Scatterplot - Sales price vs living area**


**Interpretation:** It's showing that larger house tends to have higher sale price.

- **Bar plot for average SalePrice by OverallCond**



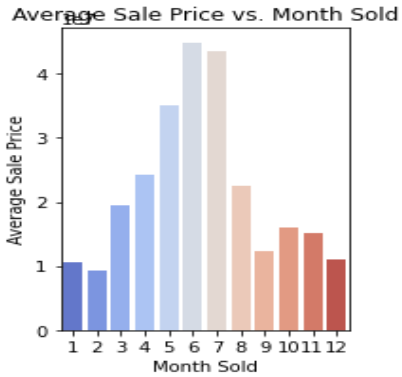

**Interpretation:** It shows that poor-condition houses (1-3) have the lowest prices, houses with average condition (5-6) sell for prices similar to those in good condition (7-8). This suggests that buyers prioritize other factors (like house quality, size, and location) over condition alone.

**Hypothesis Testing:**

1. **Houses with higher quality ratings (OverallQual) tend to have significantly higher sale prices**


Interpretation from Bar chart:
House with quality ratings of 5,6,7,8 have the positive trends for average sales which aligns with the above hypothesis.

2. **Houses sold in summer months (May - July) tend to have higher average sale prices compared to houses sold in winter months (December - February).**




Interpretation from Bar chart:

Chart is showing that Month with the number of 5, 6, 7 - May, June, and July which are summer months tends to have higher sale prices for the houses. Months with the number of 9-12 - September, October, November and December which are winter months tends to have lower sale prices. So Sales price is affected by seasonality.