PROJECT

ON

Exploratory Data Analysis (EDA) for Real Estate Pricing: Unveiling the Dynamics of House Valuation in a Dynamic Market

- **NAME: MINAL DEVIKAR**
- **INSTITUTION: DIGICHROME ACADEMY**

PROJECT OVERVIEW

Uncovering Pricing Dynamics in the Real Estate Market

- The real estate industry is a complex ecosystem influenced by myriad factors that collectively determine the pricing of residential properties.
- As analytics professionals, our mission is to navigate through this complexity and extract actionable insights from the available data.

PROBLEM STATEMENT

1. **Dynamic Real Estate Landscape:**
    1. Highlight the ever-changing nature of the residential real estate market.
    2. Emphasize the challenge of determining an optimal and competitive price for a house.
2. **Analytical Task:**
    1. Clarify the role as a key member of the analytics team in a leading real estate company.
    2. Describe the task as conducting a comprehensive analysis to identify and understand variables influencing house prices.
3. **Goal of Analysis:**
    1. Use advanced data analytics techniques and visualization tools.
    2. Uncover patterns, correlations, and trends within the dataset.
4. **Benefits for the Company:**
    1. Enable informed decision-making.
    2. Strategically position properties for better business opportunities.
5. **Approach:**
    1. Employ advanced data analytics techniques such as regression analysis, machine learning, etc.
    2. Utilize visualization tools to explore and communicate insights effectively.
6. **Focus Areas:**
    1. Identify key variables influencing house prices (e.g., location, size, amenities).
    2. Analyze the dataset to understand relationships and trends.

DATA OVERVIEW

1. Dataset Description:

The dataset used for this analysis consists of detailed information on residential properties, encompassing 81 columns and various attributes such as lot size, zoning, utilities, and sale prices.

2. Source of the Data:

Dataset Download: "Housing Data.csv

3. Key Variables:

•Examples of important variables: LotArea, MSZoning, SalePrice, etc.

Exploratory Data Analysis (EDA)

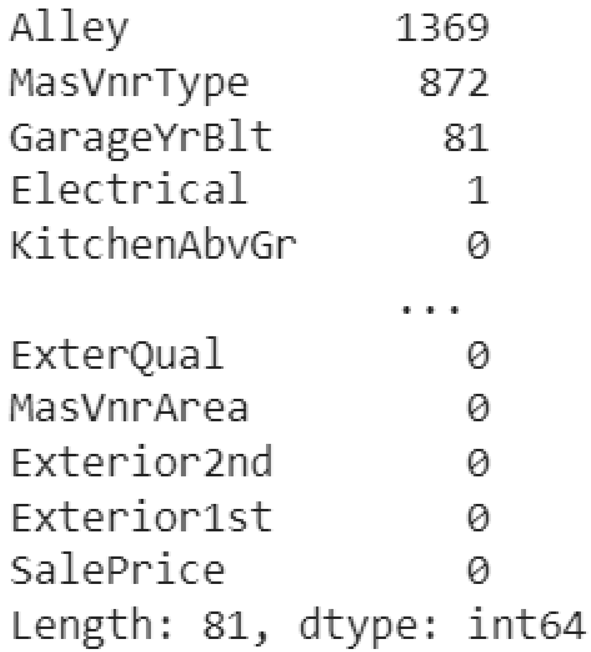
During EDA, the main issues explored were:

- Null values
- Skewed distributions
- Correlations and collinear features
- Features with possible linear relationship with saleprice

DATA CLEANING

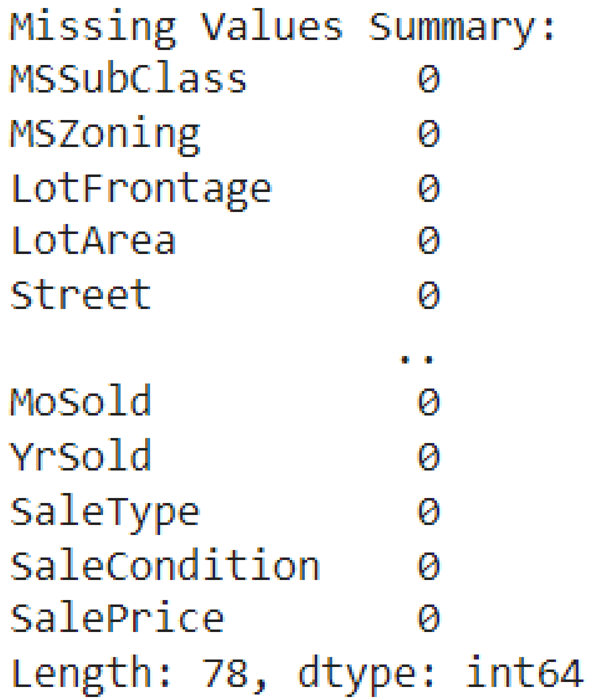
1. Handling Missing Values

- **Identify Missing Values.**



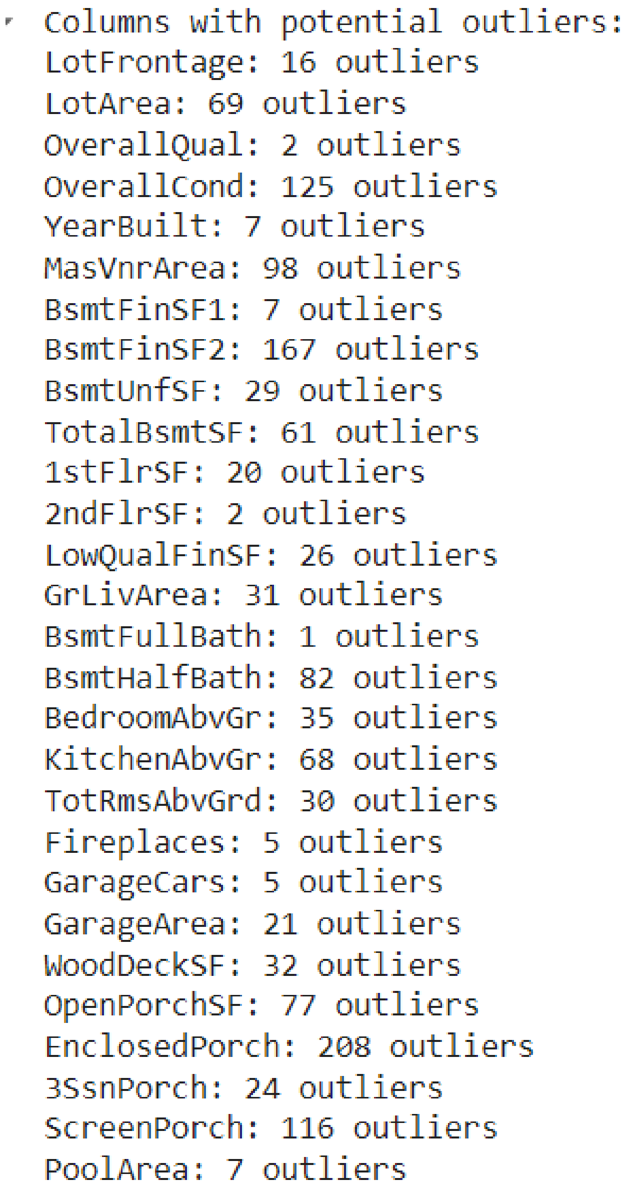
```
Alley             1369
MasVnrType         872
GarageYrBlt         81
Electrical           1
KitchenAbvGr         0
                ...
ExterQual            0
MasVnrArea           0
Exterior2nd          0
Exterior1st          0
SalePrice            0
Length: 81, dtype: int64
```

Replace the missing values by imputation, mean ,median ,mode, etc.

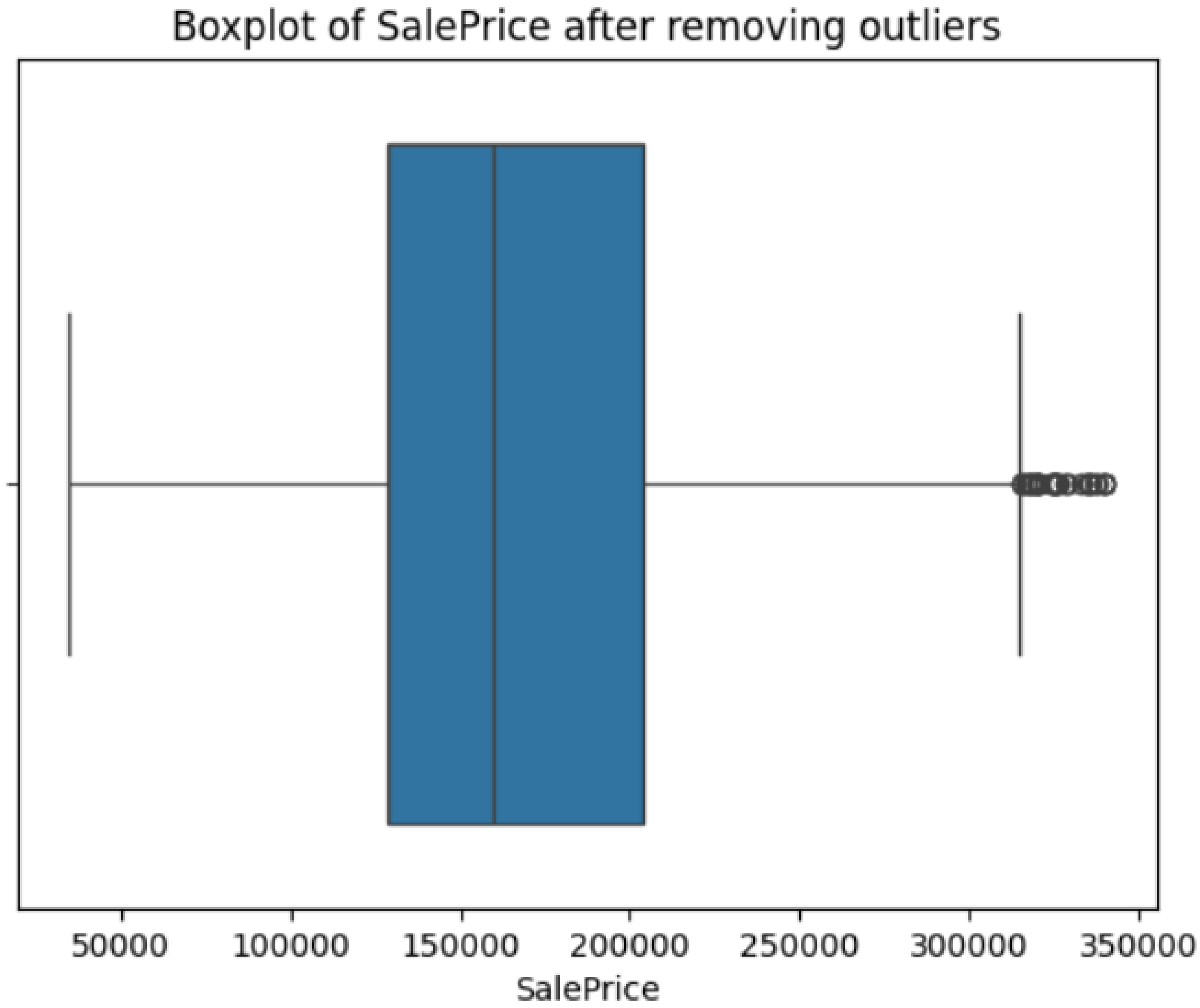


```
Missing Values Summary:
MSSubClass       0
MSZoning         0
LotFrontage      0
LotArea          0
Street           0
                ..
MoSold           0
YrSold           0
SaleType         0
SaleCondition    0
SalePrice        0
Length: 78, dtype: int64
```

Treating with Outliers

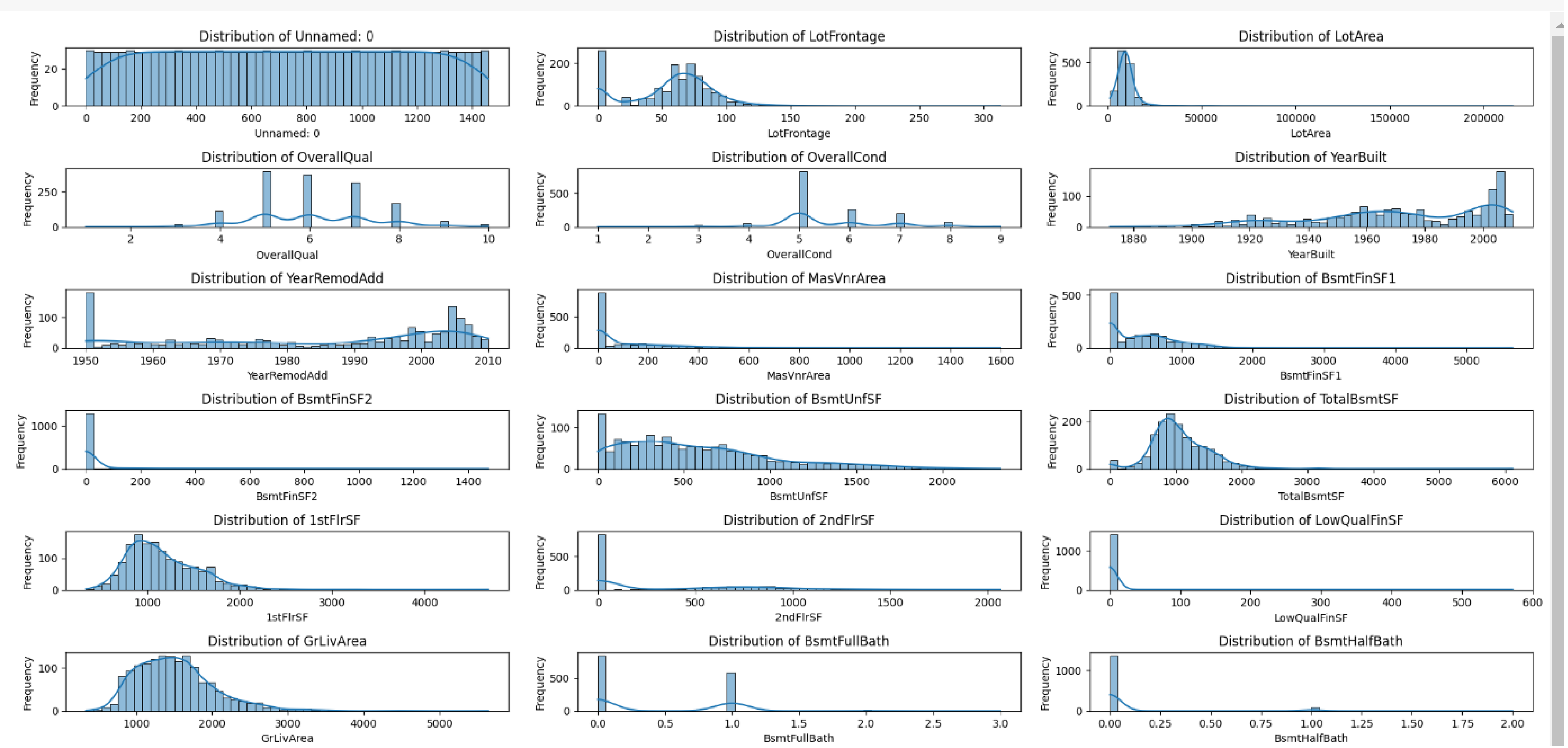


```
Columns with potential outliers:
LotFrontage: 16 outliers
LotArea: 69 outliers
OverallQual: 2 outliers
OverallCond: 125 outliers
YearBuilt: 7 outliers
MasVnrArea: 98 outliers
BsmtFinSF1: 7 outliers
BsmtFinSF2: 167 outliers
BsmtUnfSF: 29 outliers
TotalBsmtSF: 61 outliers
1stFlrSF: 20 outliers
2ndFlrSF: 2 outliers
LowQualFinSF: 26 outliers
GrLivArea: 31 outliers
BsmtFullBath: 1 outliers
BsmtHalfBath: 82 outliers
BedroomAbvGr: 35 outliers
KitchenAbvGr: 68 outliers
TotRmsAbvGrd: 30 outliers
Fireplaces: 5 outliers
GarageCars: 5 outliers
GarageArea: 21 outliers
WoodDeckSF: 32 outliers
OpenPorchSF: 77 outliers
EnclosedPorch: 208 outliers
3SsnPorch: 24 outliers
ScreenPorch: 116 outliers
PoolArea: 7 outliers
```



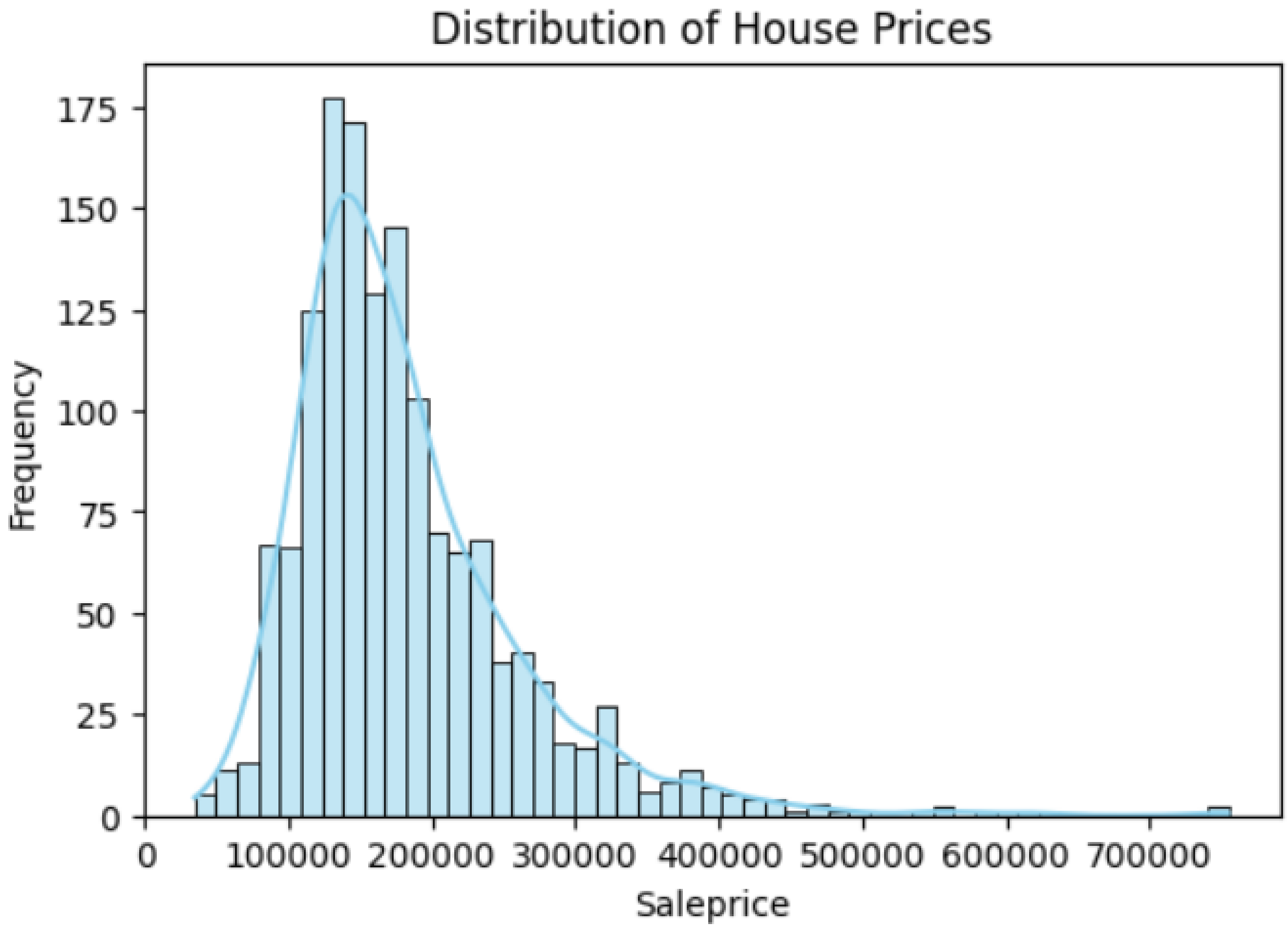
Univarient Analysis

Some columns are heavily skewed towards one value



- These are examples of distribution plots of numerical features, and histograms of categorical features. They show the frequency of each value and each category so one can get a sense of the distribution.
- Above grade living area (grlivarea) and saleprice are right skewed and have some houses that were bigger and more expensive than others. Most houses had no miscellaneous features, and almost all houses had centralair.

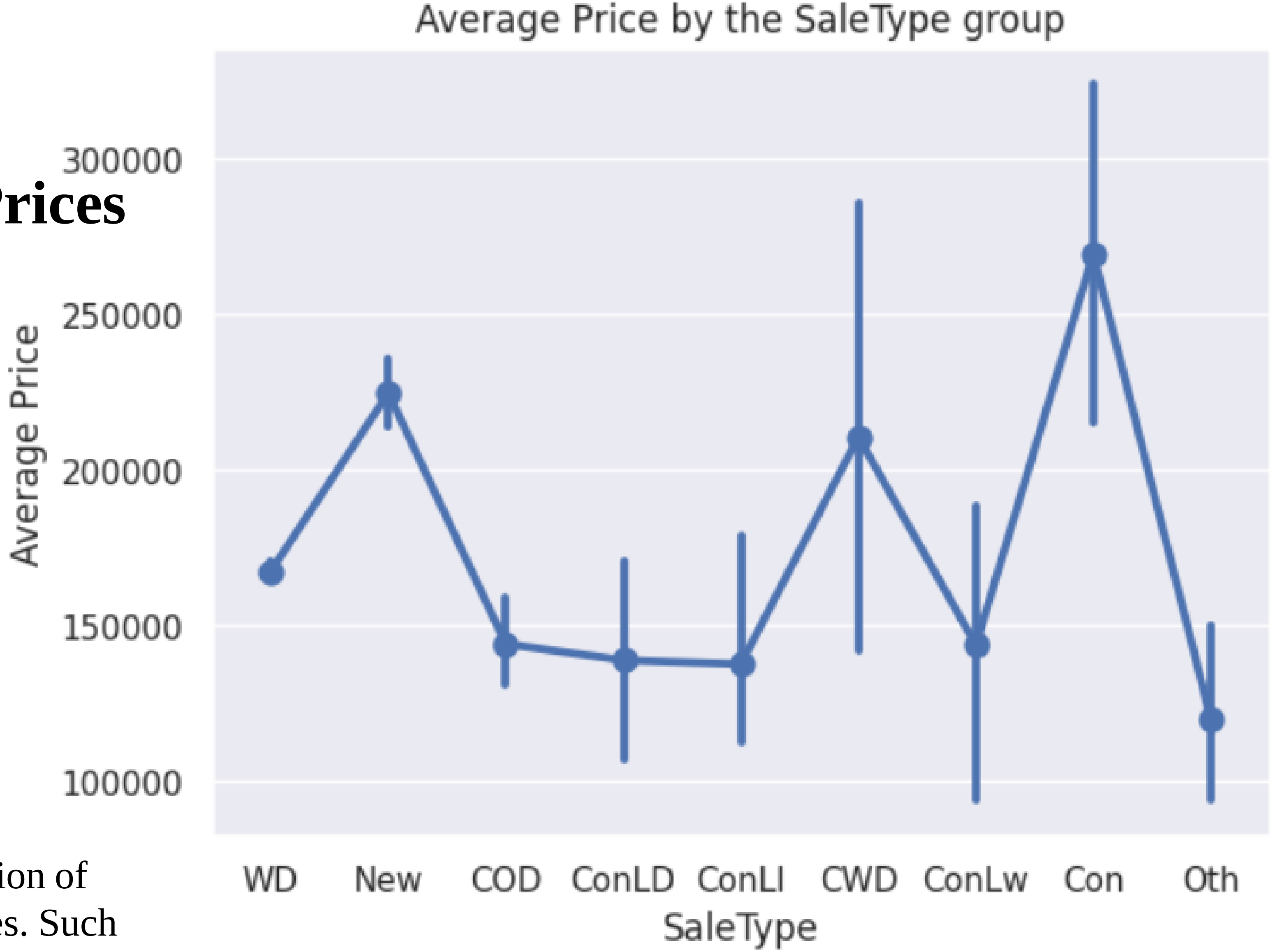
Data Distribution of Target



The distribution appears to be right-skewed, with a majority of house prices clustered towards the lower end. The presence of a kernel density estimation (KDE) curve suggests a smooth approximation of the underlying distribution. This visualization aids in understanding the spread and central tendencies of house prices within the dataset, guiding further analysis and decision-making processes.

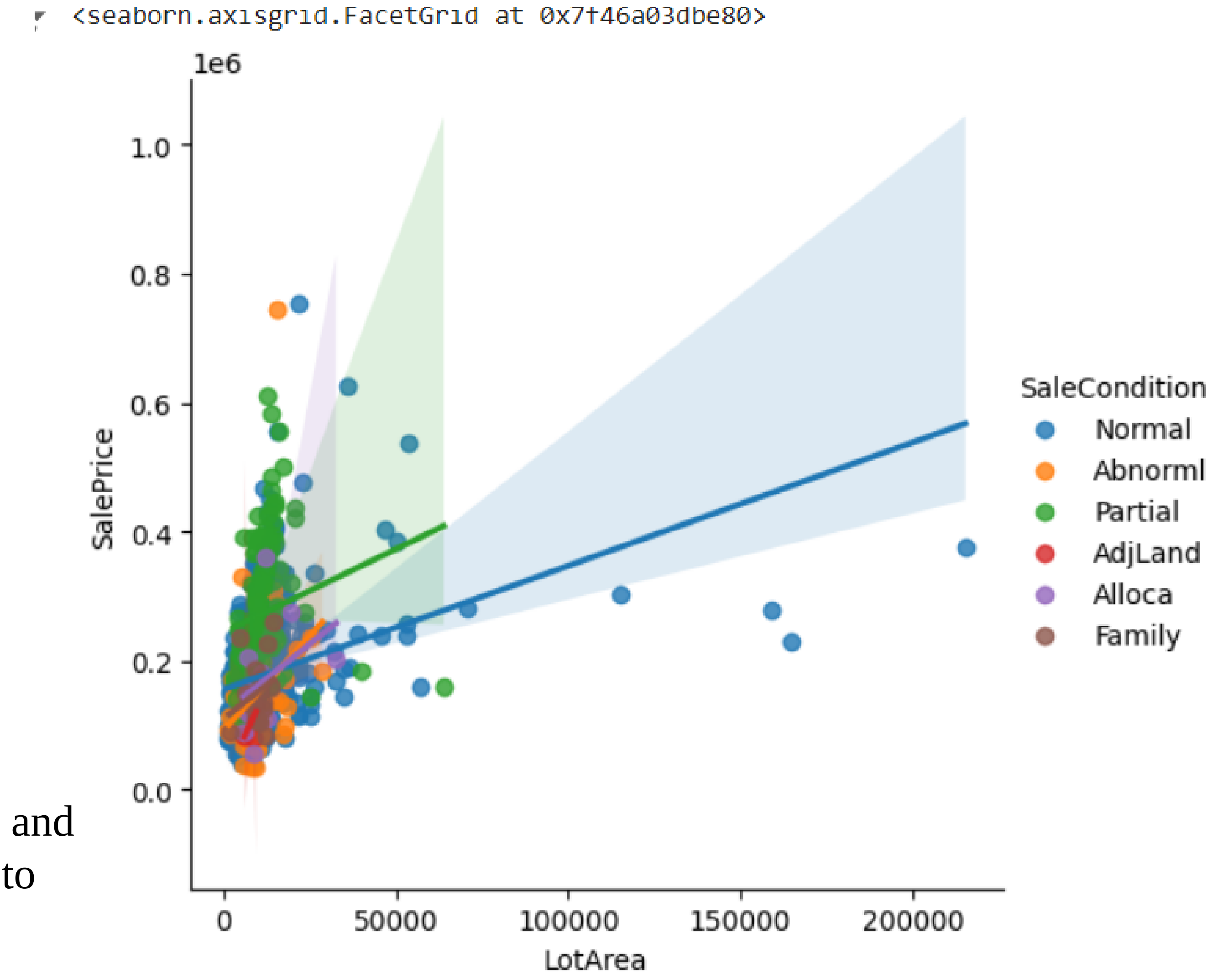
Byvarient Analysis

Analysis of Average Sale Prices by Sale Type



This allows for comparison and identification of pricing trends across various sale categories. Such insights can inform decision-making processes in real estate, aiding in pricing strategies and market positioning.

Pearson Correlation Coefficient: The Pearson correlation coefficient between LotArea and SalePrice is a quantitative measure of this relationship.



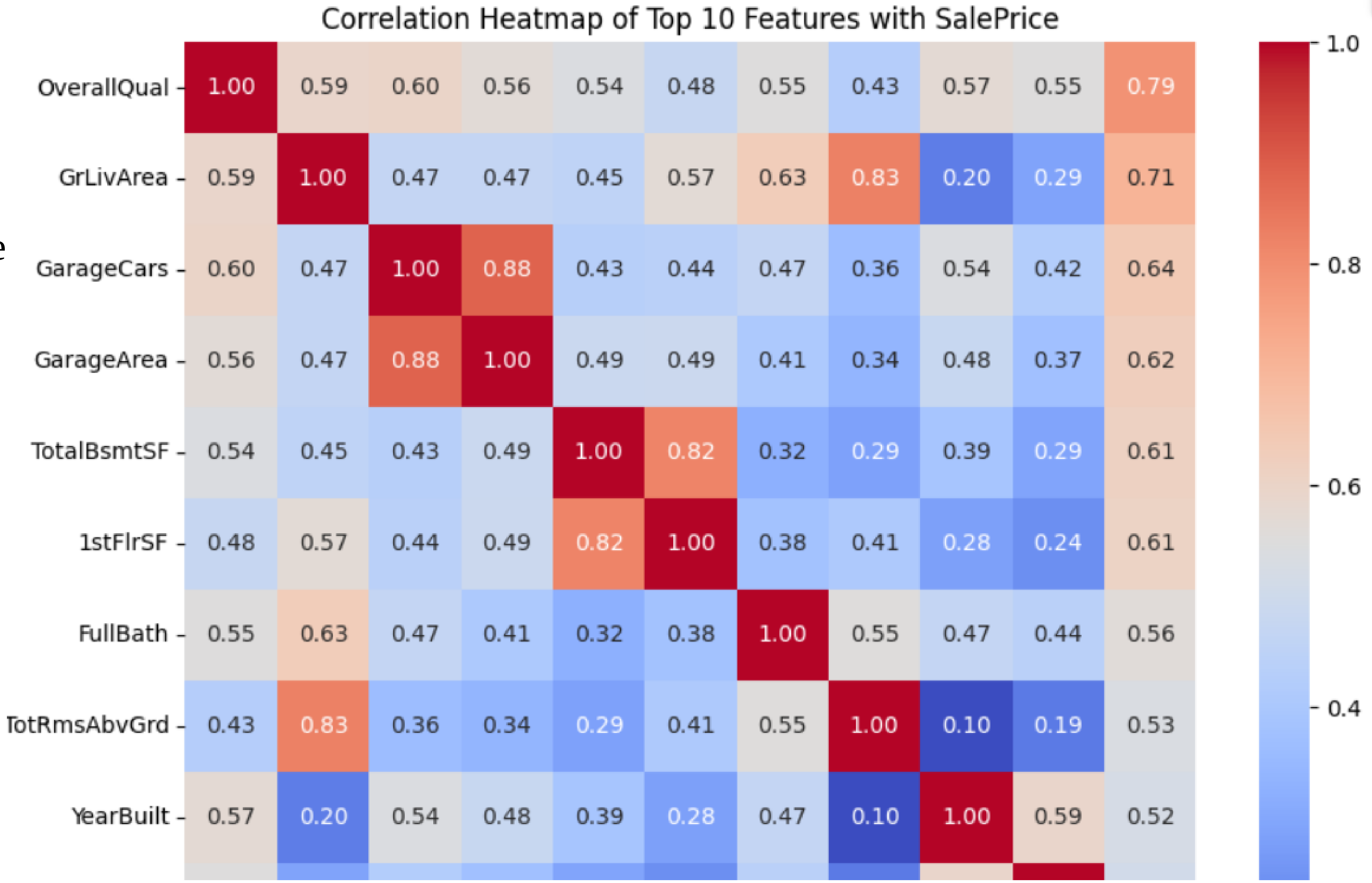
```
<seaborn.axisgrid.FacetGrid at 0x7f46a03dbe80>
```

There is a general positive correlation between LotArea and SalePrice. As the lot area increases, the sale price tends to increase as well.

Multivarient Analysis

Correlation Analysis

- The heatmap shows strong positive correlations between SalePrice and features like Overall Quality, GrLivArea, and Garage Cars, indicating that higher values in these features increase SalePrice.
- Negative correlations, though less common, suggest undesirable attributes.
- High inter-correlations among features may indicate multicollinearity, affecting model performance.
- The coolwarm color gradient aids in quickly identifying the strength and direction of these relationships.

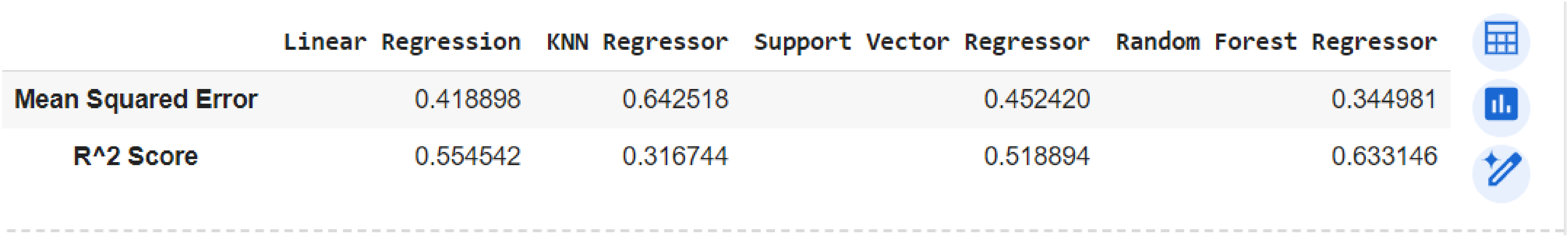


Correlation Heatmap of Top 10 Features with SalePrice

OverallQual	1.00	0.59	0.60	0.56	0.54	0.48	0.55	0.43	0.57	0.55	0.79
GrLivArea	0.59	1.00	0.47	0.47	0.45	0.57	0.63	0.83	0.20	0.29	0.71
GarageCars	0.60	0.47	1.00	0.88	0.43	0.44	0.47	0.36	0.54	0.42	0.64
GarageArea	0.56	0.47	0.88	1.00	0.49	0.49	0.41	0.34	0.48	0.37	0.62
TotalBsmtSF	0.54	0.45	0.43	0.49	1.00	0.82	0.32	0.29	0.39	0.29	0.61
1stFlrSF	0.48	0.57	0.44	0.49	0.82	1.00	0.38	0.41	0.28	0.24	0.61
FullBath	0.55	0.63	0.47	0.41	0.32	0.38	1.00	0.55	0.47	0.44	0.56
TotRmsAbvGrd	0.43	0.83	0.36	0.34	0.29	0.41	0.55	1.00	0.10	0.19	0.53
YearBuilt	0.57	0.20	0.54	0.48	0.39	0.28	0.47	0.10	1.00	0.59	0.52

RESULT

Test performance



	Linear Regression	KNN Regressor	Support Vector Regressor	Random Forest Regressor
Mean Squared Error	0.418898	0.642518	0.452420	0.344981
R^2 Score	0.554542	0.316744	0.518894	0.633146

- **Linear Regression** shows moderate performance with an MSE of 0.4189 and an R^2 of 0.5545.
- **KNN Regressor** has the poorest performance, with the highest MSE (0.6425) and lowest R^2 (0.3167).
- **Support Vector Regressor** performs similarly to Linear Regression, with an MSE of 0.4524 and an R^2 of 0.5189.
- **Random Forest Regressor** is the best model, with the lowest MSE (0.3247) and highest R^2 (0.6547).

CONCLUSION

1.**Price Distribution**: The distribution of house prices is right-skewed, indicating that a significant portion of houses falls within lower price ranges, while fewer properties are priced higher.
2.**Sale Type Trends**: Average sale prices vary across different sale types, suggesting distinct market segments with varying price dynamics. Understanding these trends can help tailor marketing strategies and target specific buyer demographics effectively.
3.**Feature Correlations**: Certain features, such as overall quality, living area, and garage capacity, demonstrate strong positive correlations with sale prices. This underscores the importance of these attributes in determining property values.
4.**Market Positioning**: Analysis of average sale prices by sale type provides insights into market demand and preferences, enabling businesses to strategically position their offerings and capitalize on emerging trends.
Overall, this dataset analysis equips stakeholders with valuable information for making informed decisions in real estate investment, marketing, and pricing strategies.

REFERENCES

- Seabold, Skipper, and Josef Perktold. "Statsmodels: Econometric and Statistical Modeling with Python." *Proceedings of the 9th Python in Science Conference*, vol. 57, 2010, pp. 61-66.
- McKinney, Wes. "Data Analysis in Python with pandas." *Proceedings of the 9th Python in Science Conference*, vol. 445, 2010, pp. 51-56.
- Hunter, John D. "Matplotlib: A 2D Graphics Environment." *Computing in Science & Engineering*, vol. 9, no. 3, 2007, pp. 90-95.