HOUSE PRICE PREDICTION

PREDICTING HOUSE PRICES USING ADVANCED REGRESSION TECHNIQUES

Abdulazeez Saliu

Table Of Content

1.0 Introduction:

The real estate market is complex and ever-changing, with various factors impacting home prices. Accurately predicting these prices is essential for buyers, sellers, real estate agents, and investors. The House Price Prediction Model is designed to offer reliable price estimates based on a comprehensive set of features, enabling real estate companies to accurately value properties for sale.

1.1 Problem Statement:

Accurately predicting house prices is a significant challenge in the real estate market, shaped by numerous factors like dwelling types, zoning classifications, lot features, property conditions, and sale conditions. Traditional property valuation methods often depend on limited data and subjective judgments, leading to inconsistencies and inaccuracies. This uncertainty can impact various stakeholders, including home buyers, sellers, real estate agents, and investors, potentially causing financial losses and inefficiencies for the real estate company.

1.2 Objective:

To overcome these challenges, our objective is to create a robust House Price Prediction Model using advanced machine learning techniques to analyze a wide range of property features and market data. We aim to achieve an overall accuracy of 85%, with a maximum difference of $25,000 between actual and predicted prices. This model is designed to deliver precise, data-driven price predictions, empowering stakeholders to make well-informed decisions in the real estate market.

1.3 In this notebook,we would make some assumptions and aim to answer several Key Questions for Gaining Insights into House prices:

Assumptions

- Sale Price is influeced by Location

- House Features affect the Sale Price

- Age of A building Affects its Sale Price

General Questions

a. **What is the distribution of house prices?**

- Understanding the overall spread, central tendency, and any outliers in house prices.

Location and Proximity

b. **How does the physical location (Neighborhood) within the city and Zoning influence house prices?**

- Comparing house prices across different neighbourhoods to identify high and low-value areas.

- Investigating the relationship between different zoning classifications (e.g., residential, commercial) and house prices.

c. **How do proximity features (Condition1, Condition2) affect house prices?**

- Determining how proximity to various conditions (e.g., arterial streets, railroads, parks) impacts house prices.

Feature-Specific Questions

d. **What is the impact of dwelling type (MSSubClass) on house prices?**

- Analyzing how different types of dwellings (e.g., 1-story, 2-story, duplex) affect house prices.

e. **How does the condition and quality of the house (OverallCond) impact house prices?**

- Assessing how the overall quality and condition ratings of a house influence its price.

f. **How do different Foundation type relate to house prices?**

- Examining how foundation types affect house prices.

g. **What is the impact of basement features (e.g., TotalBsmtSF, BsmtQual, BsmtCond) on house prices?**

- Analyzing how basement size, quality, and condition correlate with house prices.

h. **How do living area features (e.g., GrLivArea) influence house prices?**

- Investigating the relationship between the total living area and house prices.

i. **How do amenities such as garages (GarageType) impact house prices?**

- Determining the value added by amenities like garages.

Sale type and Sale Conditions

j. **How do sale type (SaleType) and sale condition (SaleCondition) affect house prices?**

- Analyzing the influence of different sale types (e.g., warranty deed, cash sale) and sale conditions on house prices.

Time specific Questions

k. **What is the effect of the year built (YearBuilt) and year remodeled (YearRemodAdd) on house prices?**

- Analyzing whether newer houses fetch higher prices.

l. **What is the impact of the time of sale (MoSold) on house prices?**

- Investigating seasonal trends and changes in house prices over time.

By answering these questions through visualizations and statistical analyses during the EDA, i can uncover important insights and relationships that will help inform the house price prediction model.

Methodology

- Exploratory Data Analysis using Univariate and Bivariate analysis

- Feature Engineering

-Predictive Modeling with Linear Regresion , RandomForest and XgBoost

-Model Evaluation with Root mean square and R-scored

1.4 Data Sources:

The Ames Housing dataset was compiled by Dean De Cock for use in data science education. It's an incredible alternative for data scientists looking for a modernized and expanded version of the often-cited Boston Housing dataset.

Kaggle

This project also serves as a capstone project for the Data Science Diploma at AltSchool.

AltSchool Africa

GitHub Project Repository

Github

1.5 Summary :

Linear Regression, Random Forest Regressor, and XGBoost models were developed for the dataset. Among them, XGBoost outperformed the others, achieving the lowest RMSE of $23,000 and the highest R-squared of 0.91.

2 Libraries & Configurations

2.1 Libraries

List of libraries to be used in the Exploratory data analysis and Model development:

pandas for data manipulation

numpy as for data computation

matplotlib for 2D data visualization

seaborn for 2D data visualization

scipy for statistics

import seaborn as sns for visualization

StandardScaler for standardization

train_test_split for splitting the data

LinearRegression for base linear regression model

RandomForestRegressor for Esemble linear regression model

xgboost for ensemble machine learning model

mean_squared_error, mean_absolute_error, r2_score for evaluation metrics

GridSearchCV for hyperparameter selection

2.2 Configurations

configurations used for the analysis.

SEED = 42

3. Data Wrangling

This dataset was loaded and discovered to have 1460 rows and 81 columns consisting of both numeric and categorical features.

	Id	MSSubClass	MSZoning	LotFrontage	LotArea	Street	Alley	LotShape
0	1	60	RL	65.0	8450	Pave	NaN	Reg
1	2	20	RL	80.0	9600	Pave	NaN	Reg
2	3	60	RL	68.0	11250	Pave	NaN	IR1
3	4	70	RL	60.0	9550	Pave	NaN	IR1
4	5	60	RL	84.0	14260	Pave	NaN	IR1

Fig 1.0 preview of the dataset

3.1 Data Validation

The id colunm was dropped for being a unique identifer and a brief description of the dataset was previewed .

	Id	MSSubClass	LotFrontage	LotArea	OverallQual	OverallCond
count	1460.000000	1460.000000	1201.000000	1460.000000	1460.000000	1460.000000
mean	730.500000	56.897260	70.049958	10516.828082	6.099315	5.575342
std	421.610009	42.300571	24.284752	9981.264932	1.382997	1.112799
min	1.000000	20.000000	21.000000	1300.000000	1.000000	1.000000
25%	365.750000	20.000000	59.000000	7553.500000	5.000000	5.000000
50%	730.500000	50.000000	69.000000	9478.500000	6.000000	5.000000
75%	1095.250000	70.000000	80.000000	11601.500000	7.000000	6.000000
max	1460.000000	190.000000	313.000000	215245.000000	10.000000	9.000000

Fig 1.1 Statistical summary of the dataset

3.2 Data Cleaning

The data had lots of missing values with some columns having 99% missing values.

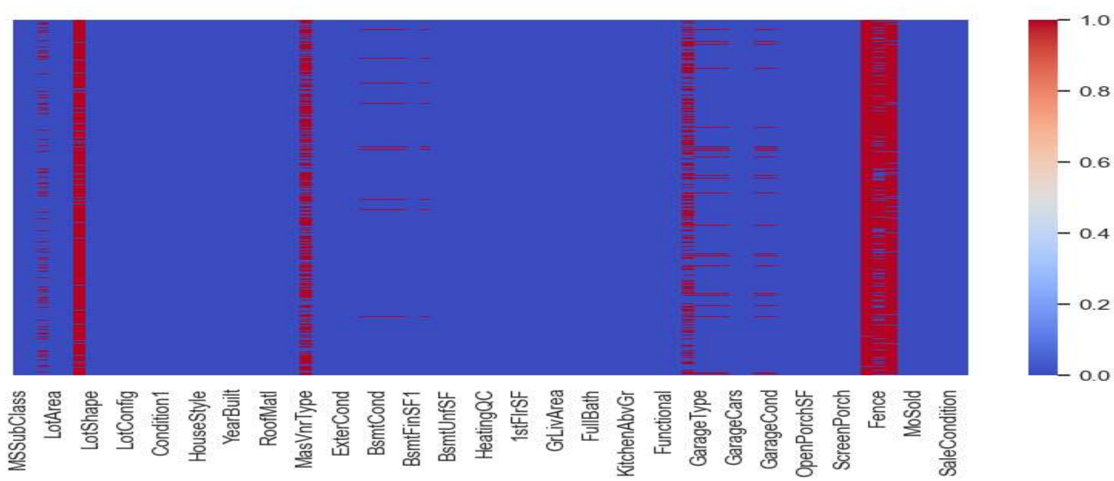


Fig 1.2Visualization of missing values

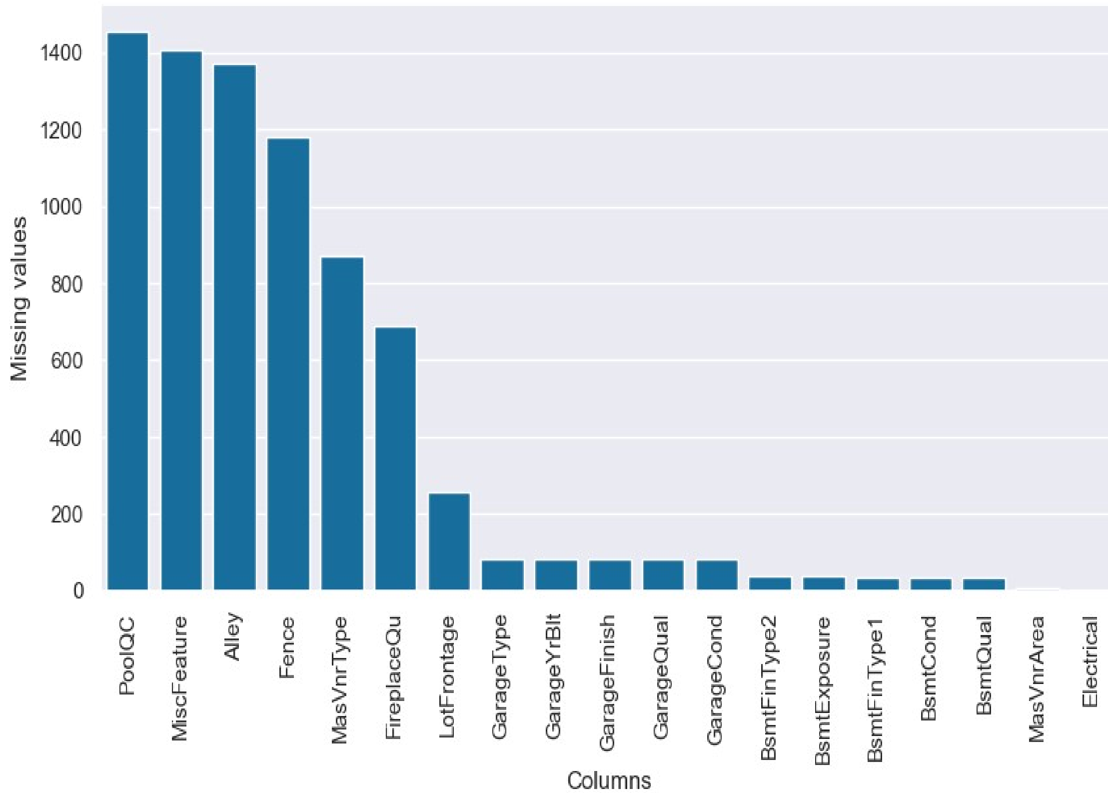


Fig1.4 Visualization of missing values

	Columns	Missing values	missing value percentage
16	PoolQC	1453	99.52
18	MiscFeature	1406	96.30
1	Alley	1369	93.77
17	Fence	1179	80.75
2	MasVnrType	872	59.73
10	FireplaceQu	690	47.26
0	LotFrontage	259	17.74
11	GarageType	81	5.55
12	GarageYrBlt	81	5.55
13	GarageFinish	81	5.55

Fig 1.5 Table showing missing values

3.2.1 Filling missing values

Columns with Missing values more than 60% of the total column number were dropped while other columns were filled with None, median and mode statistical values.

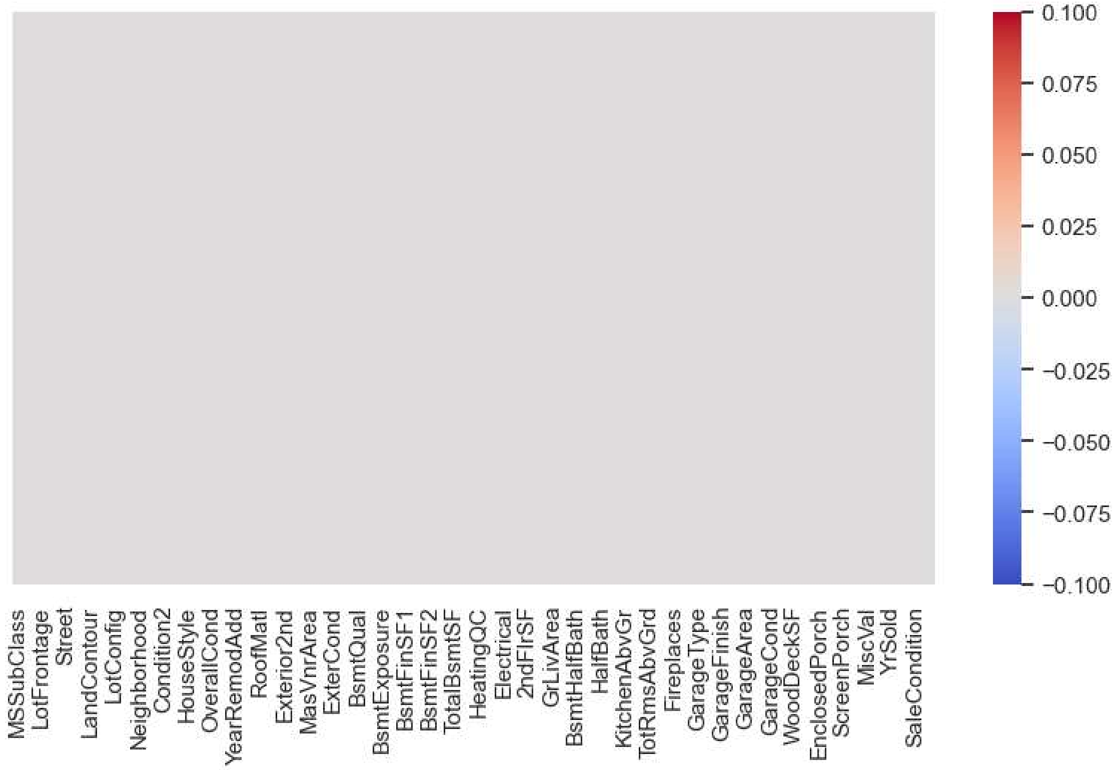


Fig 1.6 Visualization of missing values post cleaning

The shape of the dataset after cleaning is 1,455 rows and 76 columns from a previous 1460 rows and 81 columns

4.0 Exploratory Data Analysis

Exploratory data analysis was performed to uncover certain insights about the dataset. Univariant and Bivariant analysis were performed.

4.1 Univariant Analysis

Explored the Sale Price column in Isolation

4.1.1 Target Variable- Sale Price

The variable 'Sale Price' is our target variable for predictive analysis. It represents the price of buildings that have been sold having various features

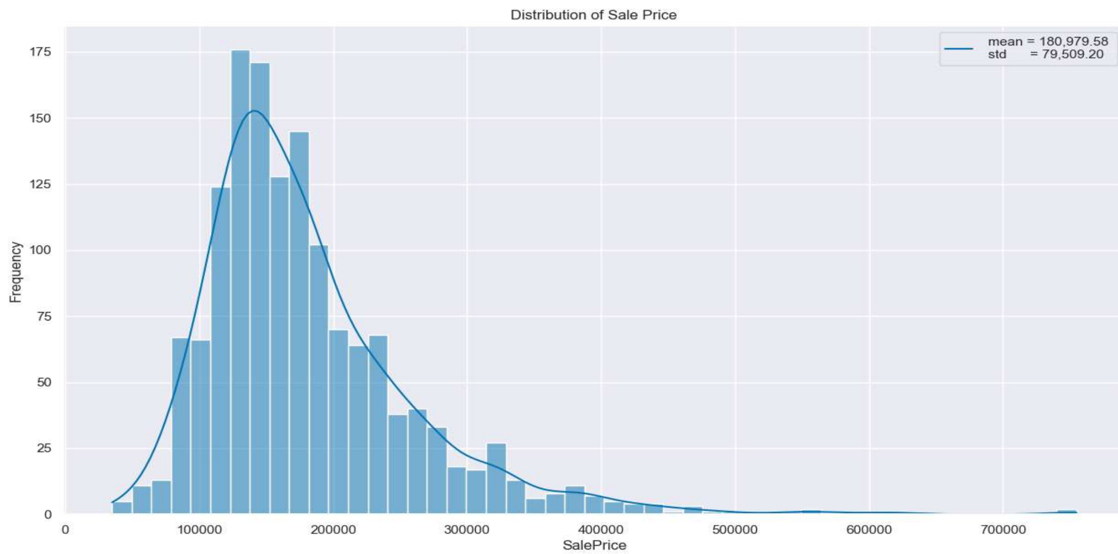


Fig 1.7 Distribution of Sale Price showing its skewness to the right

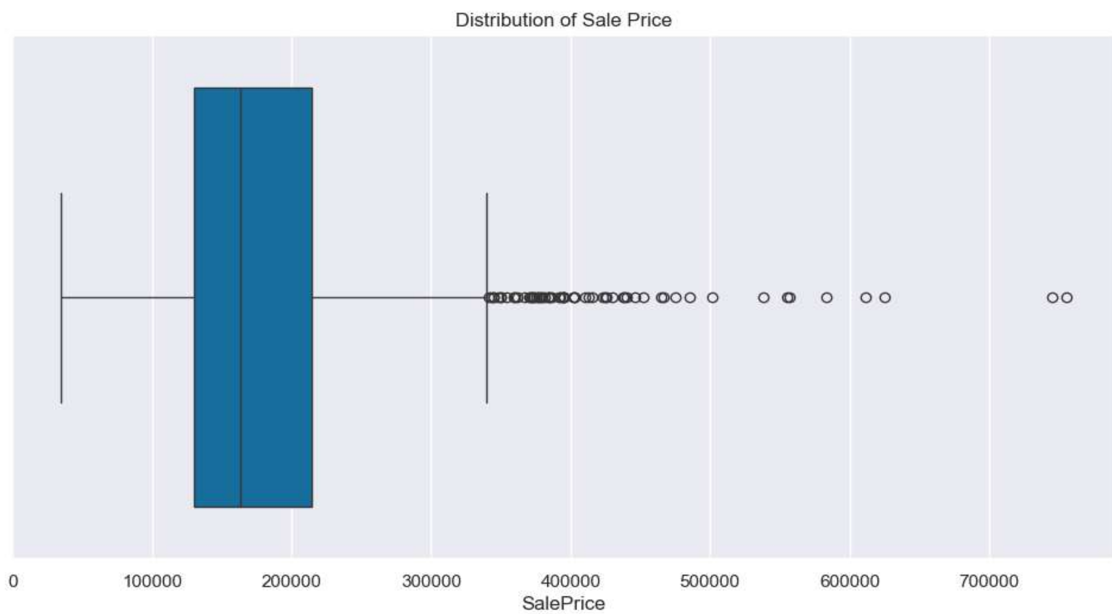


Fig 1.8 Box plot showing outliers in the Sale Price column

The sale prices range from $34,900 to $755,000, with a mean of $180,979 and a median of $163,000. The difference between the mean and median arises from the right-skewed distribution of the data. In this context, the median offers a more accurate measure of central tendency since it is less affected by extreme values.

4.2 Bivariant and MultiVariant Analysis

Bivariant analysis of the feature columns , explored their relationship with the Sale Price column

4.2.1 Categorical Features

Location and Proximity Analysis

- Neighbourhoods

Compared house prices across different neighborhoods and zones to identify high and low-value areas and also analysed the effect of proximity to certain features (rail roads, parks,greenbelt etc) on the sale price.

Fig 1.9 Distribution Of Sale Price by Neighborhood

Fig 2.0 Box plot of Sale Price by Neighborhood

High-Price Neighborhoods: NridgHt, NoRidge, and StoneBr consistently appear as high-price neighborhoods in both plots.

Low-Price Neighborhoods: MeadowV, IDOTRR, and BrDale are consistently lower in price.

Price Variability: Some neighborhoods show a wide range of house prices (e.g., NridgHt and Timber), while others have more consistent prices (e.g., SWISU and Blueste).

-Zones

compared Sale prices across zones

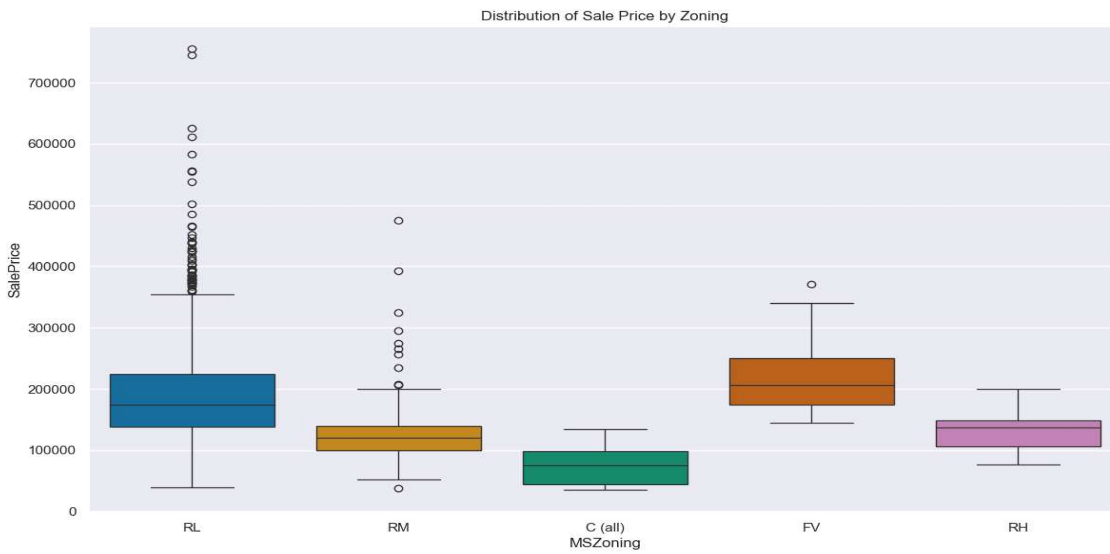


Fig 2.1 Box plot showing the distribution of Sale price by Zones

The figures revealed FV(Floating Village Residential) has the highest median sale price of $205,950 with C(commercial) having the least sale price of $74,700.

- Proximity

explored the relationship between the proximity to certain geographical features and Sale Price

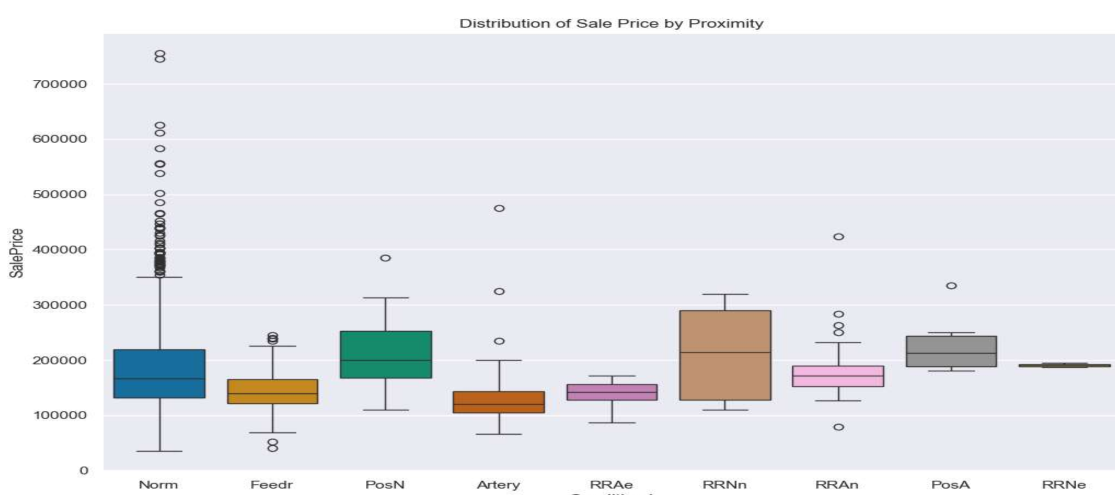


Fig 2.2 Showing the distribution of Sale price based on proximity to certain features

Properties near positive features (PosN, PosA) and within 200' of North-South Railroad (RRNn) tend to have higher median prices above $200,000.

Properties adjacent to feeder and arterial streets (Feedr, Artery), and adjacent to railroads (RRAn, RRAe, RRNe) have lower median prices.

Properties in a "normal" condition (Norm) have a wide range of prices with many high-price outliers.

Feature-Specific Analysis

Analyzed the impact of housing features on Sale prices. Features like house class,lot size, lot area,lot shape, land contour were explored.

- Dwelling Type

Explored the relationship between housing dwelling classifications and sale prices.

	DwellingTypeDescription	SalePrice
0	2-STORY 1946 & NEWER	215200.0
1	1-STORY PUD (Planned Unit Development) - 1946 ...	192000.0
2	SPLIT OR MULTI-LEVEL	166500.0
3	2-1/2 STORY ALL AGES	163500.0
4	1-STORY 1946 & NEWER ALL STYLES	159500.0

Fig 2.3 showing the median Sale price based on Building Dwelling type

- Building Quality Rating

explored how the Quality rating of a building affects its Sale price

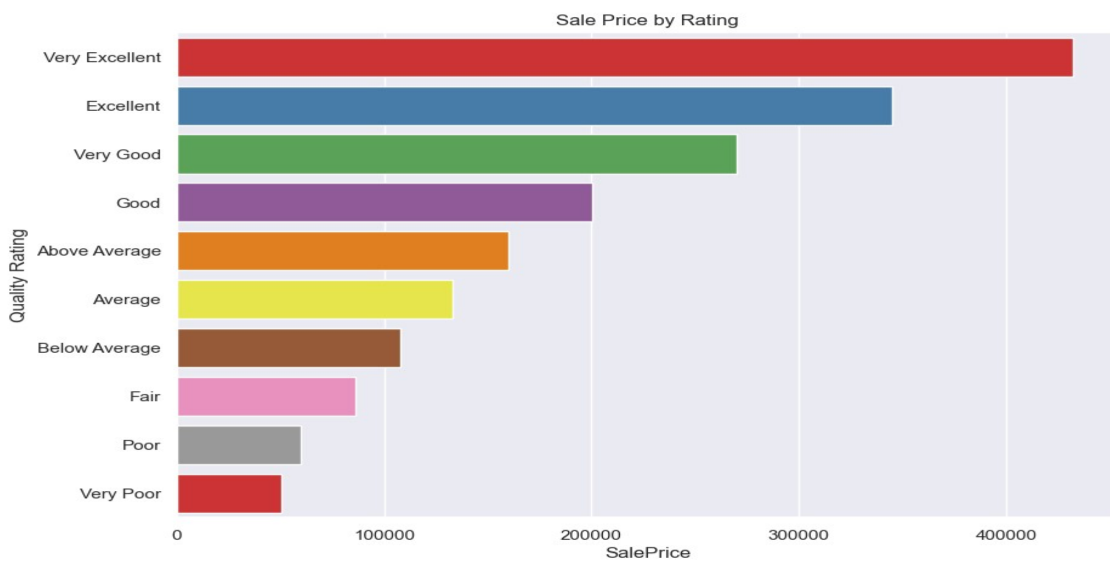


Fig 2.4 showing the median price by Building Quality rating

As expected buildings with high quality rating have a higher median sale price.

- Foundation type

Explored the effects of the foundation type used for the building and its sales price.

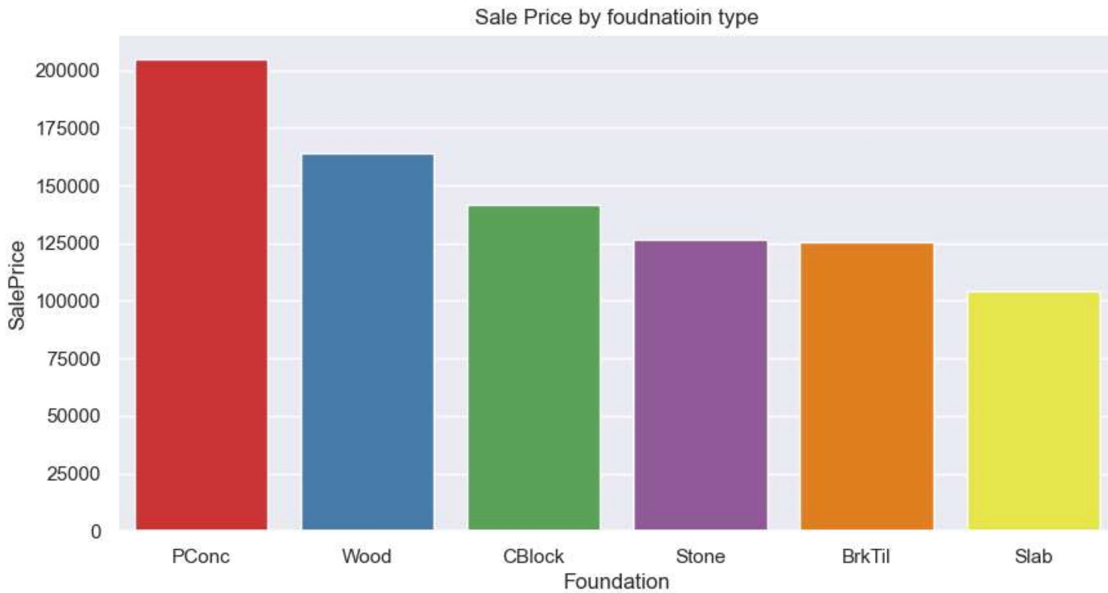


Fig 2.5 Bar chat of Median Price Of Buildings based on the foundation type

- Basement Height

Explored the effect of Basement features on the Sale price.

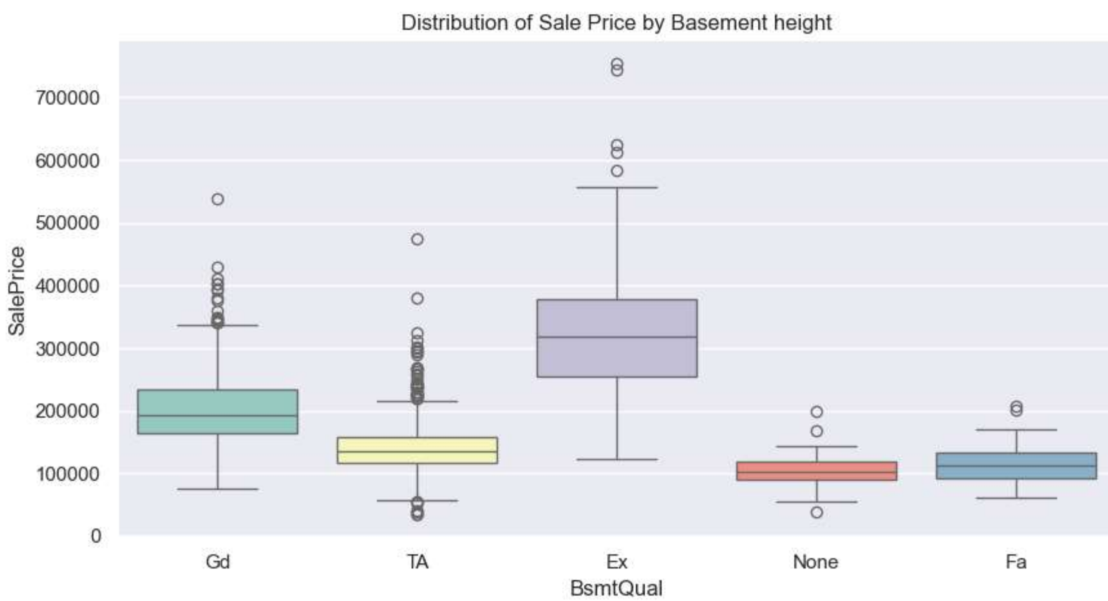


Fig 2.6 Box plot Sale price by Basement Height

Houses with basements having an "excellent" height (100+ inches) have a median sale price of $318,000, while those without a basement have a much lower median sale price of $101,800. This indicates a significant correlation between basement height and house sale price.

- Central Air Conditioning

Explored the effect of Air conditioning type of the building and its sales price

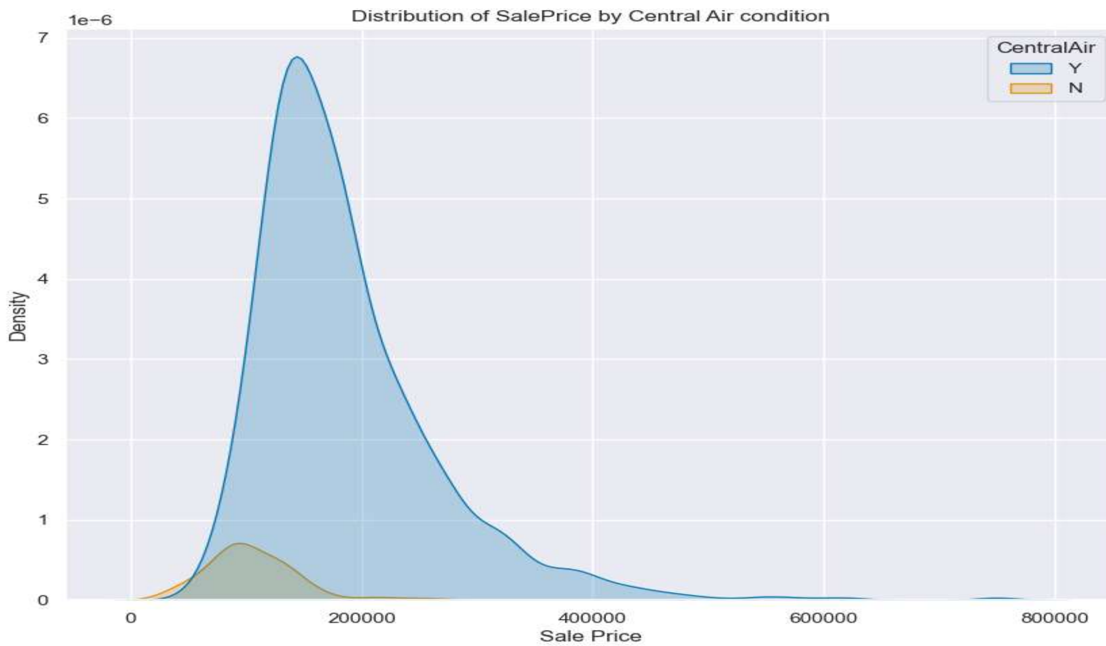


Fig 2.7 KDE plot of Sale price by Centra air conditioning

The KDE plot shows that homes with central air conditioning tend to have higher sale prices, with most clustered around $200,000, compared to homes without central air, which are generally priced lower

- Roof type

Explored the effect of Roof type of the building and its sales price.

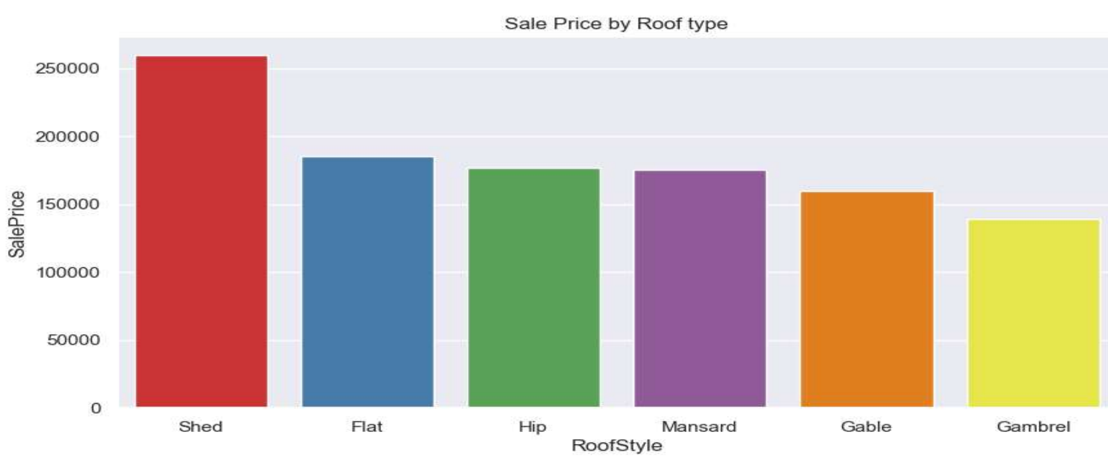


Fig 2.8 Bar chart showing median price of Houses based on Roofing type

Houses with Shed roofs have the highest median sale price at $260,000, followed by those with Flat, Hip, Mansard, Gable, and Gambrel roofs. This suggests that Shed roof houses generally command higher prices compared to other roof styles.

- Garage Type

Explored the effect of garage type of the building and its sales price.

	GarageType	SalePrice
0	BuiltIn	230000.0
1	Attchd	185000.0
2	2Types	159000.0
3	Basment	148000.0
4	Detchd	129500.0
5	CarPort	108000.0
6	None	100000.0

Fig 2.9 Showing the median price of buildings baed on Garage typ

Houses without a garage have the lowest median price while houses with a built in Garage have a median price of $230,000.

-Sale Type

Analyzed the effect of Sale typeon the Sale Price of houses.

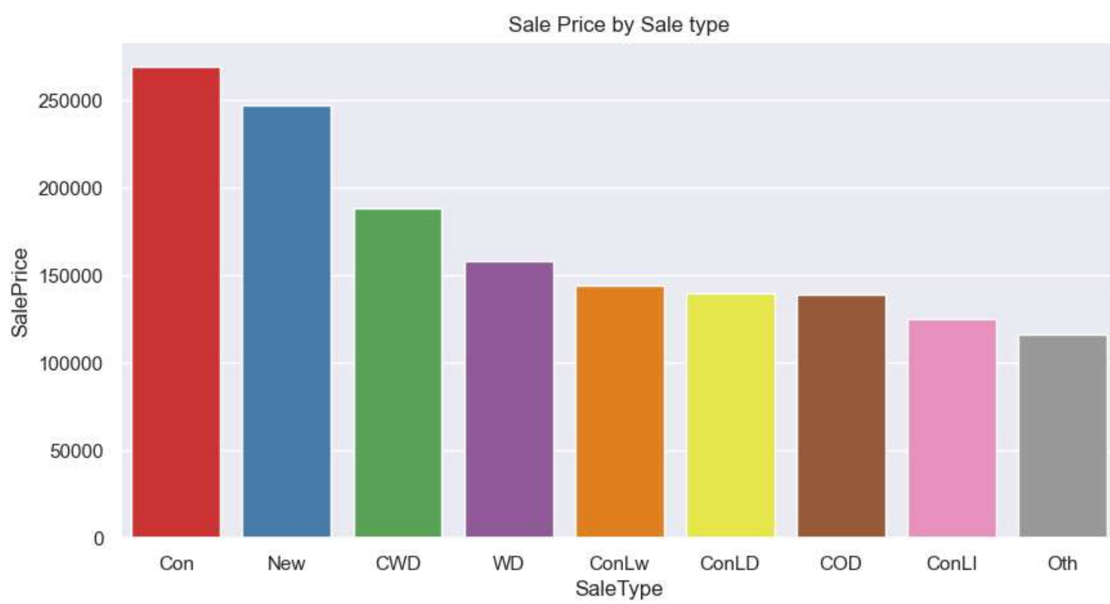


Fig 3.o showing median sale price by Sale type

Houses with a 'Contract with 15% Down payment with regular terms' have a median sale price oof $269,600 closely followed by Newly built houses with a median sale price of $247,453.

4.2.2 Numeric Variable

Explored the relationship between the numeric variables and sales price

- Corrolation

Exlored correlation amongs the numeric features using the Pearson correlation

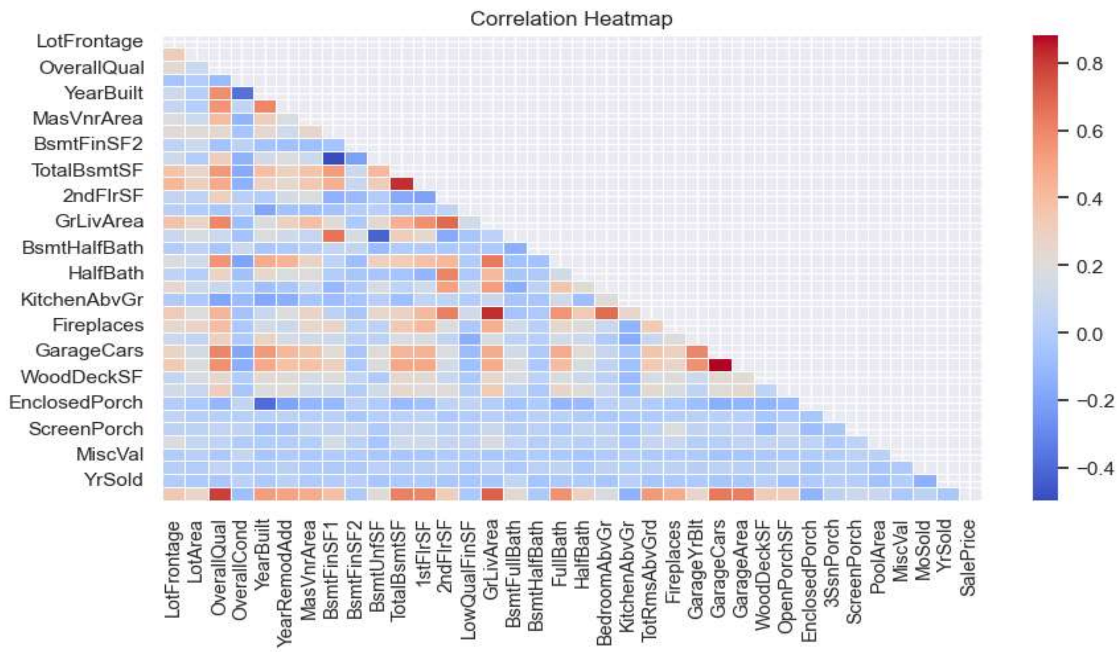


Fig 3.1 Correlation heatmap showing correlation between various numeric features

	Column	Coefficient
0	OverallQual	0.790999
1	GrLivArea	0.709451
2	GarageCars	0.640529
3	GarageArea	0.623354
4	TotalBsmtSF	0.613649
5	1stFlrSF	0.605796
6	FullBath	0.560223
7	TotRmsAbvGrd	0.535754
8	YearBuilt	0.522736
9	YearRemodAdd	0.506520
10	MasVnrArea	0.473558
11	Fireplaces	0.466442
12	BsmtFinSF1	0.386838

Fig 3.2 Table showing correlation values between top numeric features and SalePrice

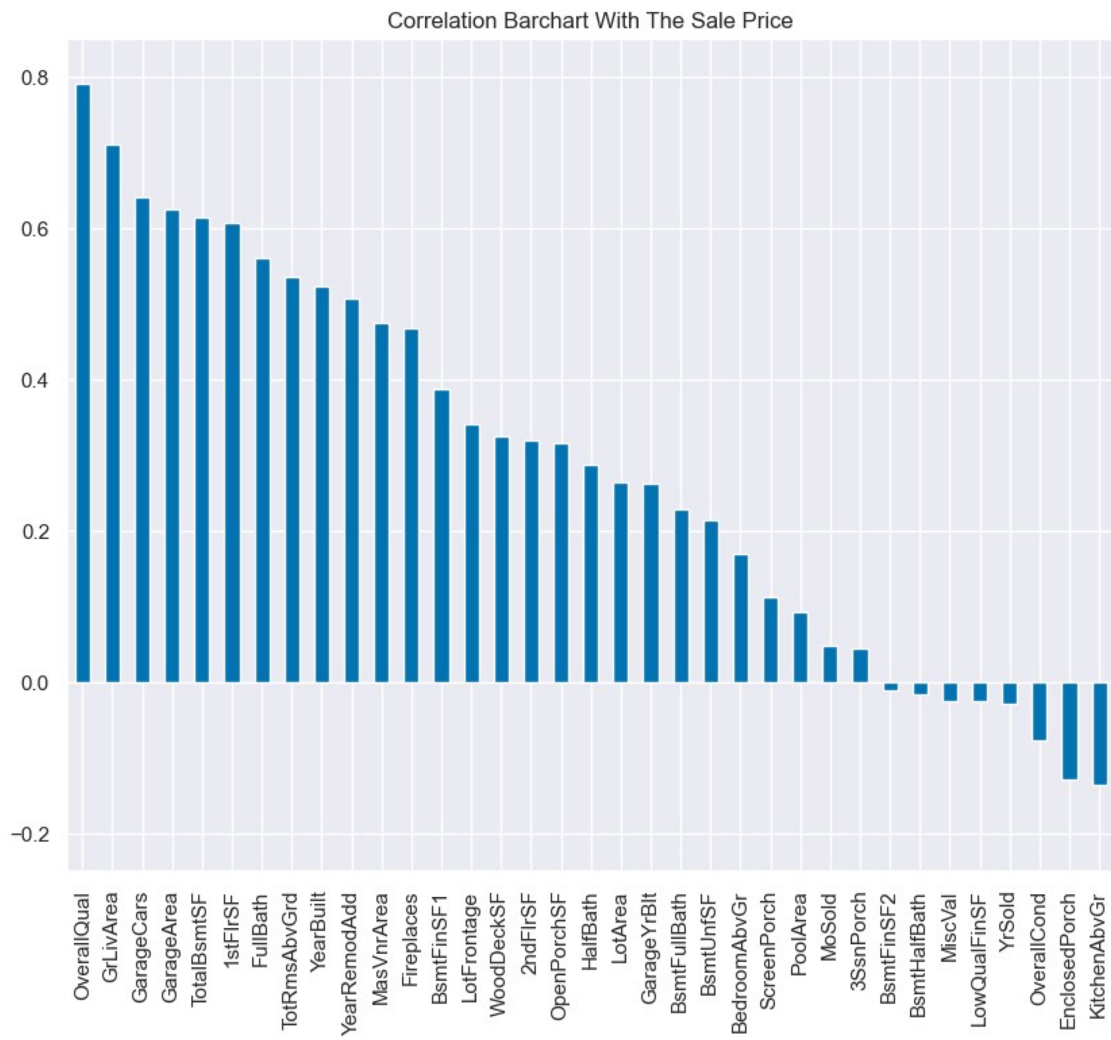


Fig 3.3 Correlation bar chart between features and SalePrice

High Positive Correlations:

OverallQual, GrLivArea, and GarageCars show strong positive correlations with SalePrice. This suggests that higher quality ratings, larger living areas, and more garage space are associated with higher sale prices.

There are strong correlations between similar features, such as GarageArea and GarageCars, 1stFlrSF and TotalBsmtSF, as well as TotRmsAbvGrd and GrLivArea highlighting potential multicollinearity issues among the features.

Moderate to Weak Correlations:

Features like YearBuilt, YearRemodAdd, and FullBath show moderate positive correlations with SalePrice.

Some features, such as LotFrontage and LotArea, have weaker correlations with SalePrice

Negative Correlations:

There are very few negative correlations, and they tend to be weak. Features like EnclosedPorch have slight negative correlations with SalePrice.

Redundant Features:

The heatmap reveals features that are highly correlated with each other (e.g., GarageArea and GarageCars). These redundant features would be combined or one of them could be dropped to simplify the model.

- OverallQual

The heatmap revealed a strong positive correlation between OverallQual and Sale price.

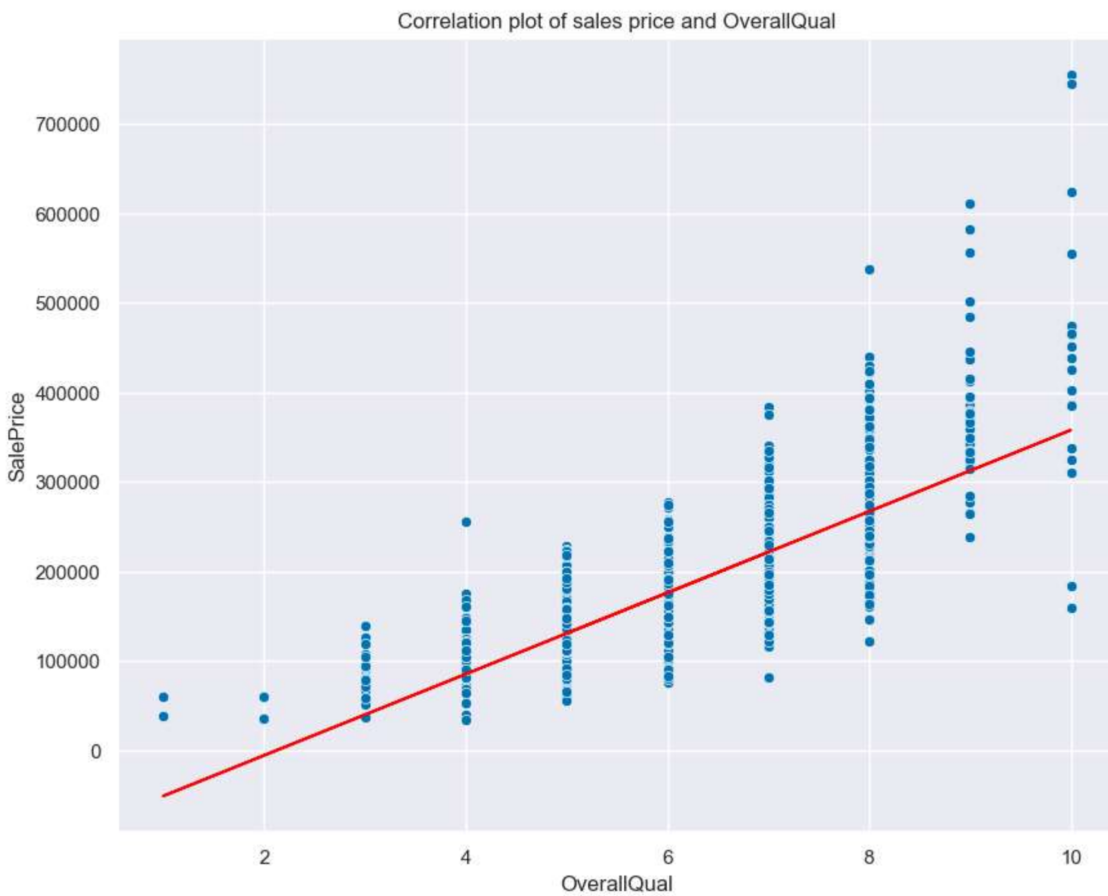


Fig 3.4 Scatter plot of Overall quantity and Sale Price

As the overall quality rating increases, the sale price tends to increase as well, indicating that higher quality houses generally sell for higher price evident in the trend of the scatter plot and alos the corrolation value of 0.79

-GrLivArea: Above grade (ground) living area square feet

Explored the relationship between the Total house living area (minus basement) and Sale price

Fig 3.5 Scatter plot of Sale Price and Living Area

The scatterplot revealed that buildings with larger living area tend to have higher sales price evident with the Pearson correlation coefficient of 0.71.

- Garage Area

Explored further the positive correlation between Sale price and garage area

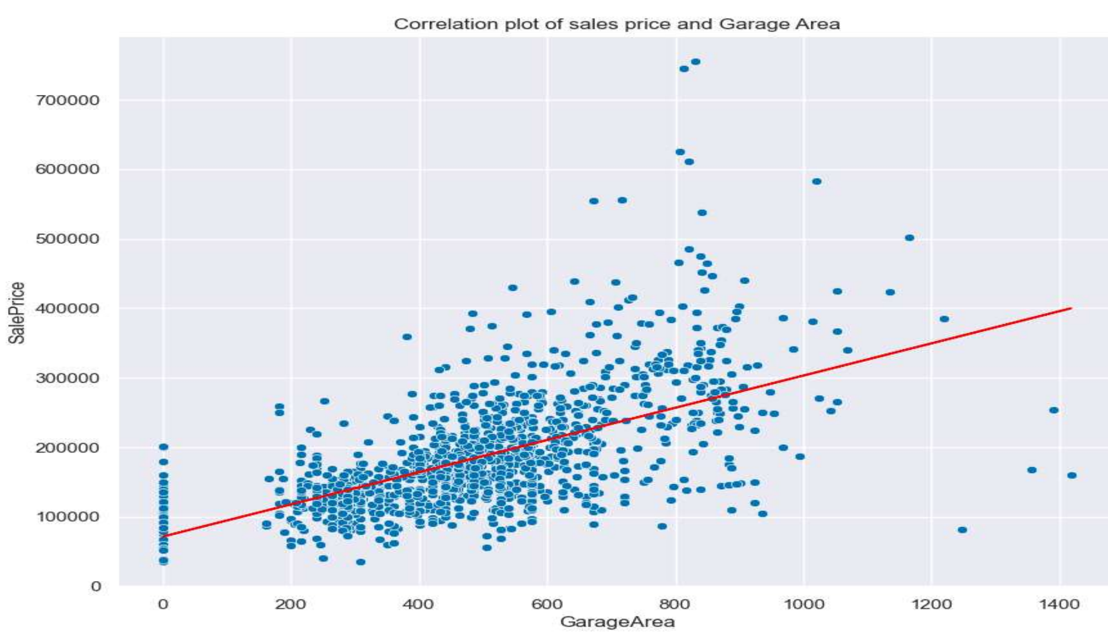


Fig 3.6 Scatter plot of Garage Area and Sale price

Buildings with large garage area tend to have a higher Sale price evident with the corrolation coefficient of 0.62

- Year Built

Explored the effect of year of construction on sale price

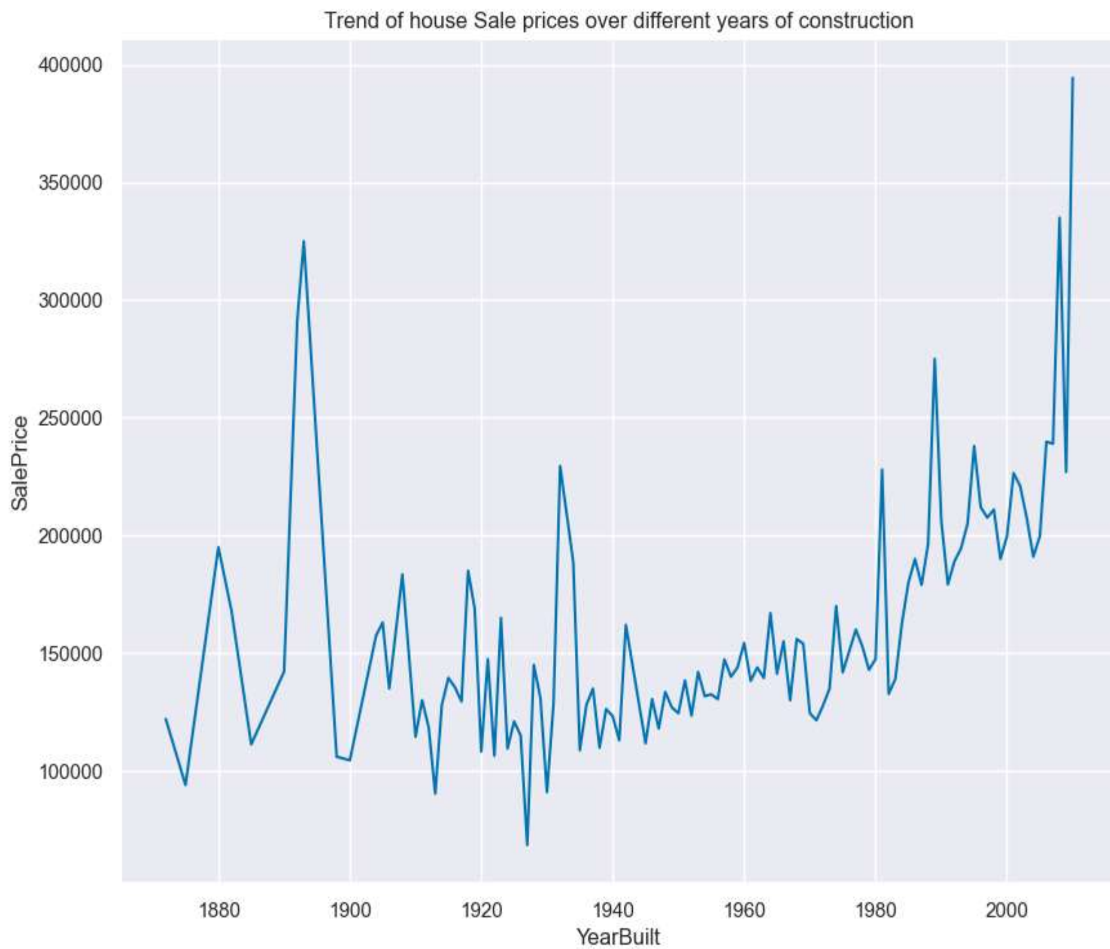


Fig 3.7 Showing trend of Sale price for building year of construction

The line graph illustrates a general increase in house sale prices over the years, with distinct peaks in the early 1900s and a significant upward trend as the year 2000 approaches. This pattern suggests that newer houses tend to command higher sale prices, reflecting historical market trends and likely improvements in construction quality and amenities over time. The correlation coefficient between SalePrice and YearBuilt is 0.52, indicating that while there is a positive relationship, other factors also play a significant role in determining sale prices.

- Month Sold

explored the effect of Month of sale on sale price

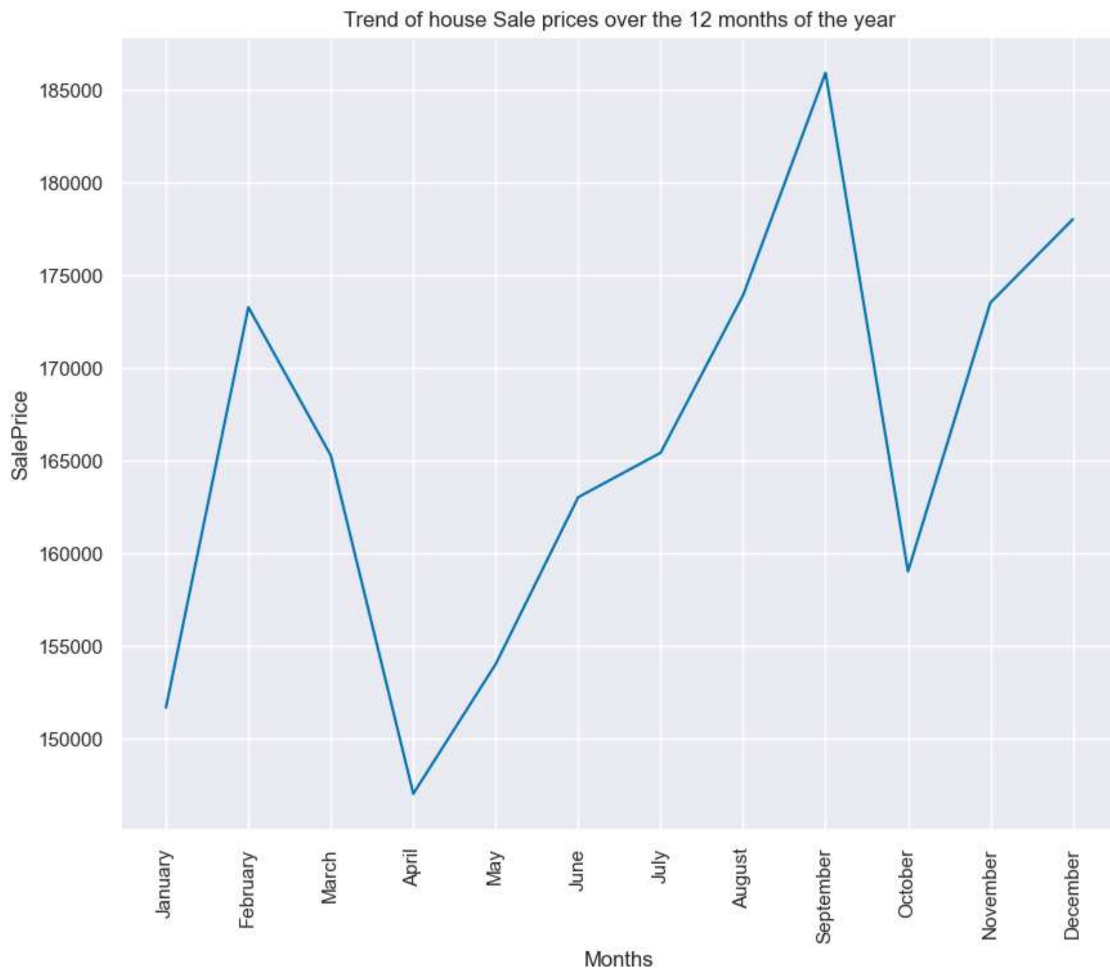


Fig 3.8 Showing trend of Sale price over the 12 months of the year

The line plot shows the trend of house sale prices over the 12 months of the year. Sale prices start lower in January, peak around Febuary, dip in April, and then rise again, reaching the highest point in September. This indicates seasonal variations in house sale prices, with the highest prices typically occurring towards the end of the year. This plots helps both buyers and sellers to determine the best time of the year to buy or sell buildings.

4.3 Insights

Summary of Insights from EDA on Sale Prices

- Sale Price Distribution

The sale prices of houses range from $34,900 to $755,000, with an average price of $180,979 and a median of $163,000. Most houses are priced below $200,000, although there is a significant number of outliers at higher prices.

- Neighborhood Influence

Neighbourhoods such as NridgHt, NoRidge, and StoneBr consistently appear as high-price areas, while MeadowV, IDOTRR, and BrDale are associated with lower prices.

- Zoning Classification

Houses in the FV (Floating Village Residential) zoning classification have the highest median sale price at $205,950, whereas those in the C (Commercial) zoning classification have the lowest median sale price at $74,700.

- Impact of Location and Accessibility

Properties located near feeder streets, arterial streets, and railroads tend to have lower median sale prices.

- Building Type

Among different building types, 2-STORY 1946 & NEWER buildings have the highest median sale price, with a value of $215,200.

- Building Quality

As expected buildings with high quality rating have a higher median sale price. The rating is has good corrolation with sale price

- Basement Features

Houses with basements that have a ceiling height of 100+ inches have a median sale price of $318,000. In contrast, houses without basements have a median price of $101,800. Interestingly, houses with poorly rated basements tend to have lower prices than those without basements at all.

- Central Air Conditioning

Properties with central air conditioning generally command higher and more variable sale prices, with a median price of $168,250 compared to those without central air conditioning.

- Roof Style and Material

Houses with Shed-style roofs have the highest median sale price at $260,000, followed by other roof styles. The type of roofing material also affects the sale price, with wood shingle roofs having the highest prices and roll roofs the lowest.

- Garage Type

Houses with built-in garages have a higher median sale price of $230,000, whereas those without garages tend to have lower prices.

- Sale Conditions

Properties sold under contracts with a 15% down payment have a median sale price of $269,600, closely followed by newly built houses with a median price of $247,453. On the other hand, houses sold with adjoining land purchases have the lowest median sale price of $104,000.

- Overall Quality

As the overall quality rating of houses increases, the sale price tends to increase as well, indicating that higher quality houses generally sell for higher prices.

- Living Area

The analysis shows that buildings with larger living areas tend to have higher sale prices.

- Garage Area

Similarly, buildings with larger garage areas are associated with higher sale prices.

- Historical and Seasonal Trends

House sale prices have generally increased over time, with noticeable peaks around the early 1900s and a significant upward trend toward the year 2000. This reflects historical market trends and possibly improvements in construction quality and amenities. Additionally, house prices show seasonal variations, with lower prices in January, a peak in February, a dip in April, and the highest prices typically occurring in September. These trends can help both buyers and sellers determine the best time of year to buy or sell properties.

5.0 Model Fitting & Evaluation

Predicting the Sale Price is a regression problem that was effectively tackled using a Linear Regression model, which I chose as my foundational approach. To compare performance, I used the Random Forest Regressor and XGBoost, both of which excel at handling outliers and capturing complex, non-linear patterns.

For model evaluation, I focused on RMSE, aiming to keep the maximum difference between predicted and actual sale prices within $25,000.

5.1 Pre-Processing

The data was prepared before it was fed into the models

5.1.1 Feature Engineering

A new column Age and Renovated were created to enhance the targets relationship with the features.

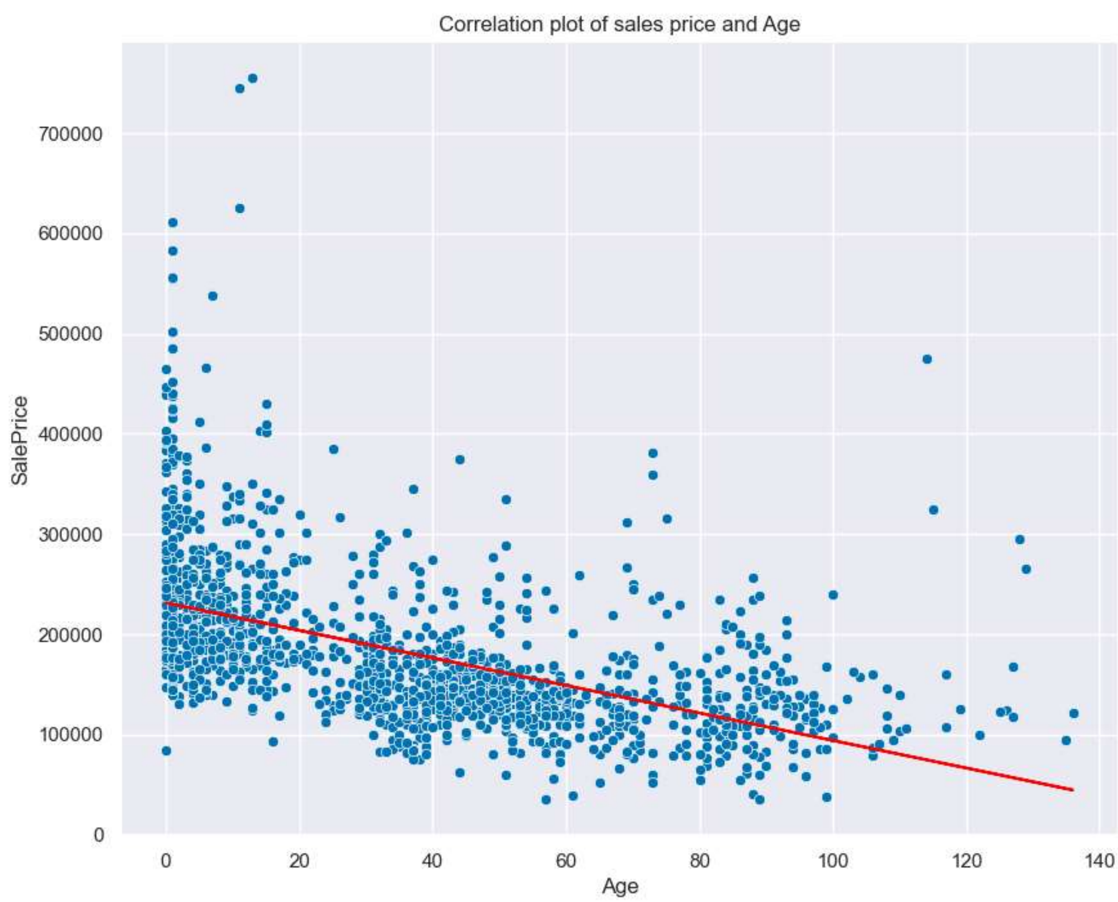


Fig 3.9 Scatter plot of Building Sale Price and Age

The scatter plot reinforced our earlier observation that newly built houses generally attract higher sale prices, while older houses tend to be less expensive. However, with a moderate negative correlation coefficient of -0.52, it's clear that other factors significantly influence sale prices.

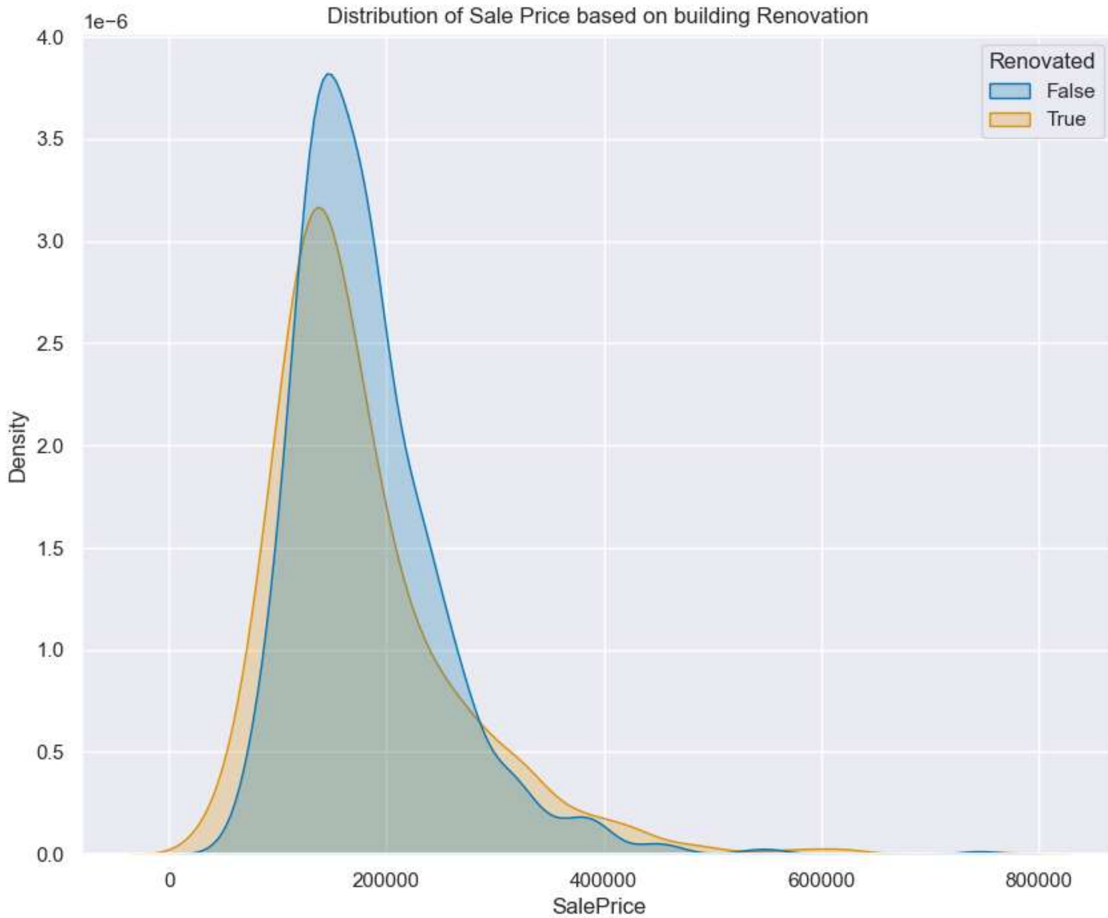


Fig 4.0 Showing Distribution of Sale price based on Building Renovation

The plot indicates that while renovation might contribute to a higher sale price, the effect is not overwhelmingly strong, as the distributions for renovated and non-renovated homes are quite similar, with a moderate difference towards higher prices for renovated homes.

- Feature Selection

Selecting Numeric features with above 0.4 corrolation with the Target column and Categorical features with p-value below 0.05 this would reduce noise in the mode

5.1.2 Outliers

Using the Inter Quantile range, outliers were revealed in the dataset on the Sale price column. Upon further analysis, it was revealed that the outliers represent luxury homes. The outlier were capped with the values of the upper quantile.

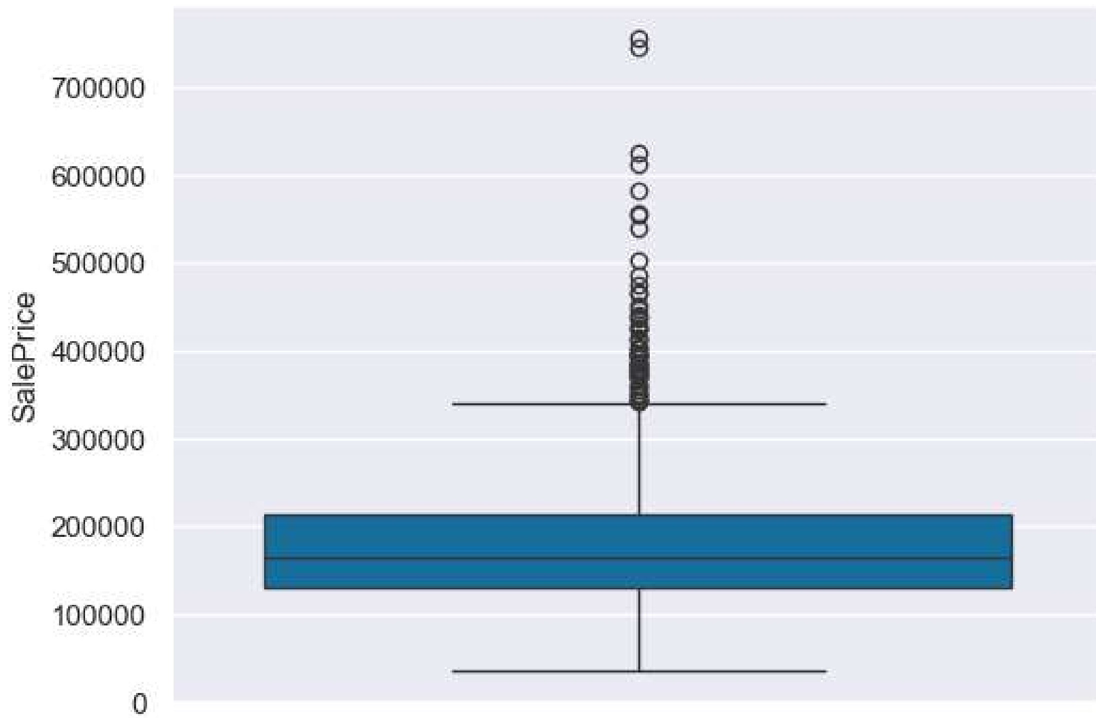


Fig 4.1 BoxPlot showing outliers in the SalePrice column

The boxplot of "SalePrice" revealed that while the majority of home prices range from approximately $150,000 to $250,000, The Outliers start to appear beyond the upper whisker of the boxplot. In this case, the upper whisker extends up to approximately $350,000., with some sale prices reaching up to $700,000, indicating a significant number of homes sold at much higher prices compared to the typical range.

5.1.3 One Hot encoding

The categorical variables were encoded into binary values using the pandas get_dummies method due to machine learnig models ability to only deal with numeric values.

5.1.4 Train-Test Split

The data was split into Train and test set to allow for evaluation on the test set (unseen data). The data was split 80/20 in favor of the train set

5.1.5 Normalization

Normalization was performed to bring all features to the same scale to avoid bias. The normalization was done using the sklearn StandardScaler method.

5.2 Linear Regression Model

The Linear Regression model was chosen as the base model to establish a foundation for other ensemble models to build upon and enhance.

-Model Evaluation

R-squared (R^2) = 0.75

The R^2 value of 0.75 means that approximately 75% of the variance in the target variable (SalePrice) is explained by the independent variables in the model. While 0.75 is not extremely low, it suggests that the model may not be capturing all the relevant patterns in the data. There might be other factors influencing the target variable that are not included in the model, or the relationship between the features and the target might be more complex than what a linear model can capture.

RMSE = 36,634.57

The Root Mean Squared Error (RMSE) is a measure of the average magnitude of the prediction error in the same units as the target variable. An RMSE of 36,634.57 means that, on average, your model's predictions are off by about $36,634.57.This relatively high error indicates that the model's predictions are not very accurate. It suggests that there could be significant discrepancies between the actual and predicted values.

MSE = 1,342,091,467.32

The Mean Squared Error (MSE) is the average of the squared differences between actual and predicted values. It is more sensitive to outliers because errors are squared. An MSE of 1.3 billion indicates substantial variance in the errors.

Such a high MSE further supports the notion that the model is not performing well.

5.3 Random Forest Model

The Random Forest model is an ensemble method chosen for its robustness and ability to withstand the impact of extreme values in the dataset.

-Model Evaluation

R-squared (R^2) = 0.85

The R^2 value of 0.85 means that approximately 85% of the variance in the target variable (SalePrice) is explained by the independent variables in the model. This is a significant improvement from the base linear model of 0.75 due to the random forest model being an emsemble model.

RMSE = 27,896.05

An RMSE of $27,896.05 indicates that, on average, the model's predictions deviate from the actual values by approximately $27,896.05. While this represents a notable improvement over the baseline linear model, it still falls short of the key performance indicator (KPI) target of $25,000.

MSE = 778,189,697.68

An MSE of 778,189,697.68 in comparison to the base linear model suggest the models ability to explain the variance in the dataset

5.4 xgboost Model

XGBoost is a powerful ensemble model selected for its ability to handle complex, non-linear relationships and improve on the baseline performance established by the base Linear Regression and random forest

-Model Evaluation

R-squared (R^2) = 0.90

The R^2 value of 0.90 means that approximately 91% of the variance in the target variable (SalePrice) is explained by the independent variables in the model. This is a good improvement from the random forest model of 0.85.

RMSE = 22,915.99

An RMSE of 22,915.99 indicates that, on average, the model's predictions deviate from the actual values by approximately 22,915.99.64. this meets the key performance indicator (KPI) target of $25,000.

MSE = 525,142,786.69

An MSE of 525,142,786.69 indicates our model is adept at dealing with the variance in our dataset

- Hyperparameter Tuning

The Xgboost hyperparameters were tuned using GridsearchCv module with an aim of finding the best hyperparamters and improving models performance

The below hyper parameters were obtained :

Best Parameters: {'colsample_bytree': 1.0, 'gamma': 0, 'learning_rate': 0.1, 'max_depth': 3, 'n_estimators': 150, 'subsample': 0.8}

Best Score (MSE): 566478581.7864558

When applied the below model performance was recorded:

Mean Squared Error (MSE): 494140876.72

Root Mean Squared Error (RMSE): 22229.28

R-squared (R^2): 0.91

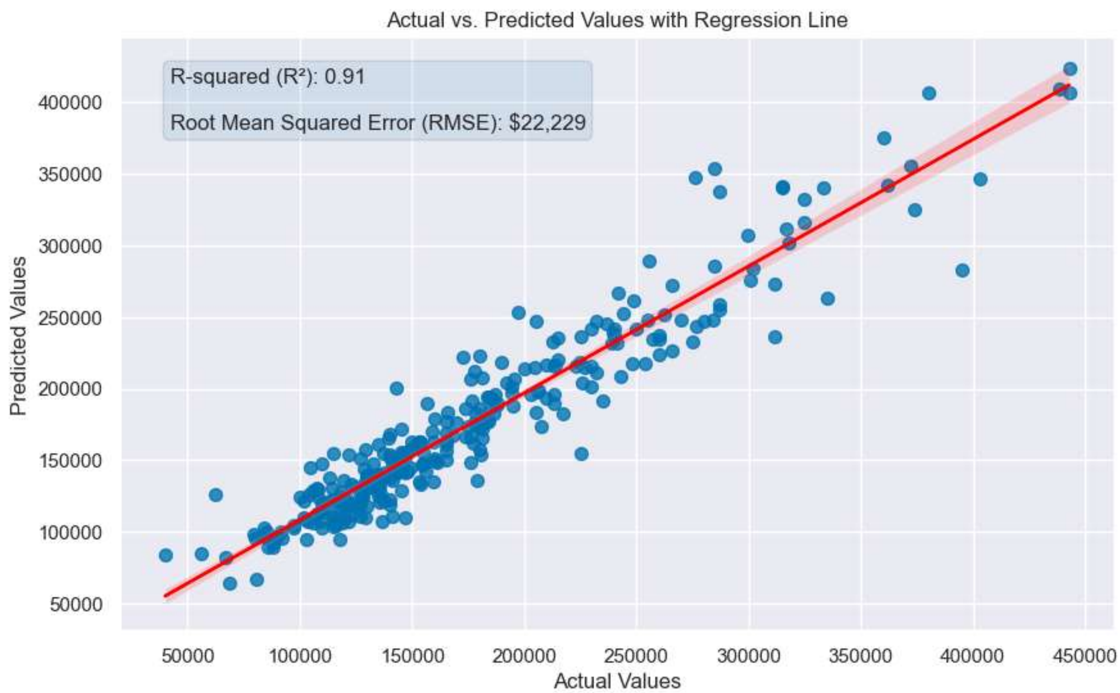


Fig 4.2 visualizing the models performance

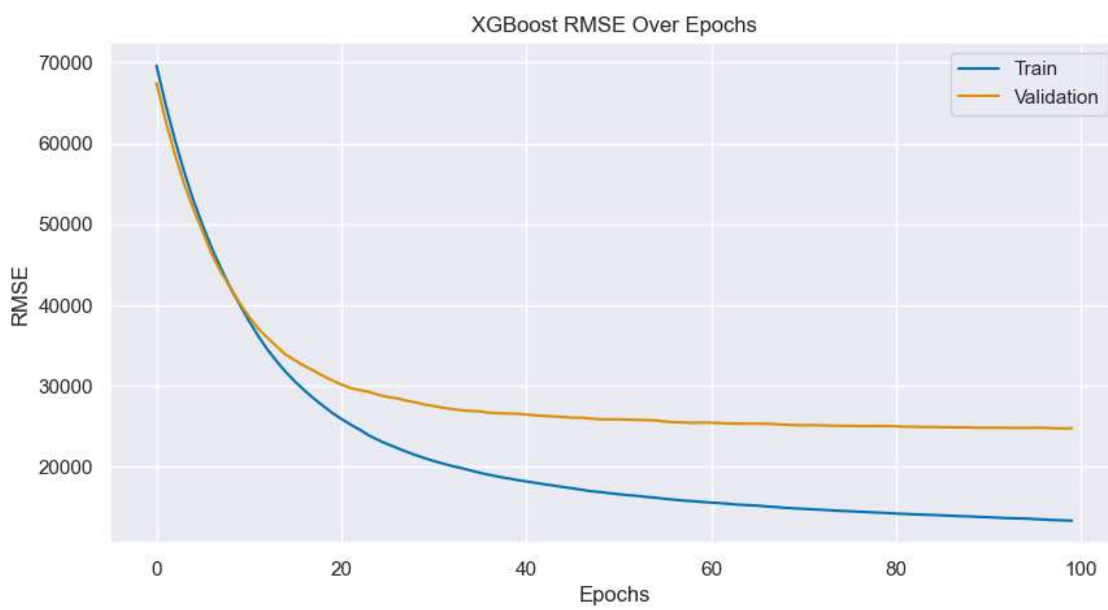


Fig 4.3 visualizing the models performance against overfitting

The plot shows that the XGBoost model's RMSE decreases for both the training and validation sets over the first 100 epochs, with the training RMSE continuing to decrease more sharply than the validation RMSE, which begins to plateau, indicating potential overfitting as the model continues to train.

Feature Importance

Checking the features that contiributed the most to the model performanceChecking the features that contiributed the most to the model performance

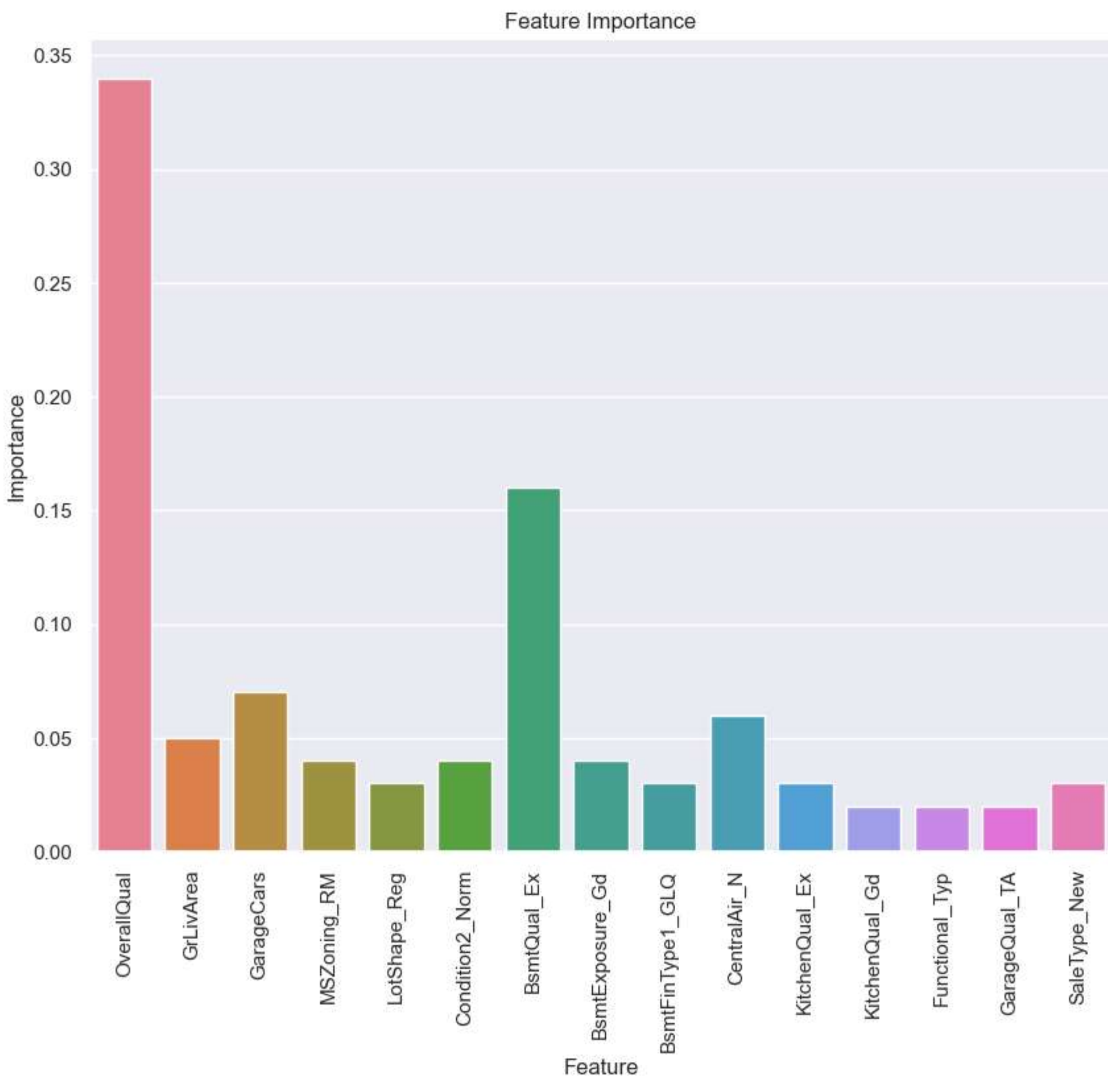


Fig 4.4 visualizing the important features

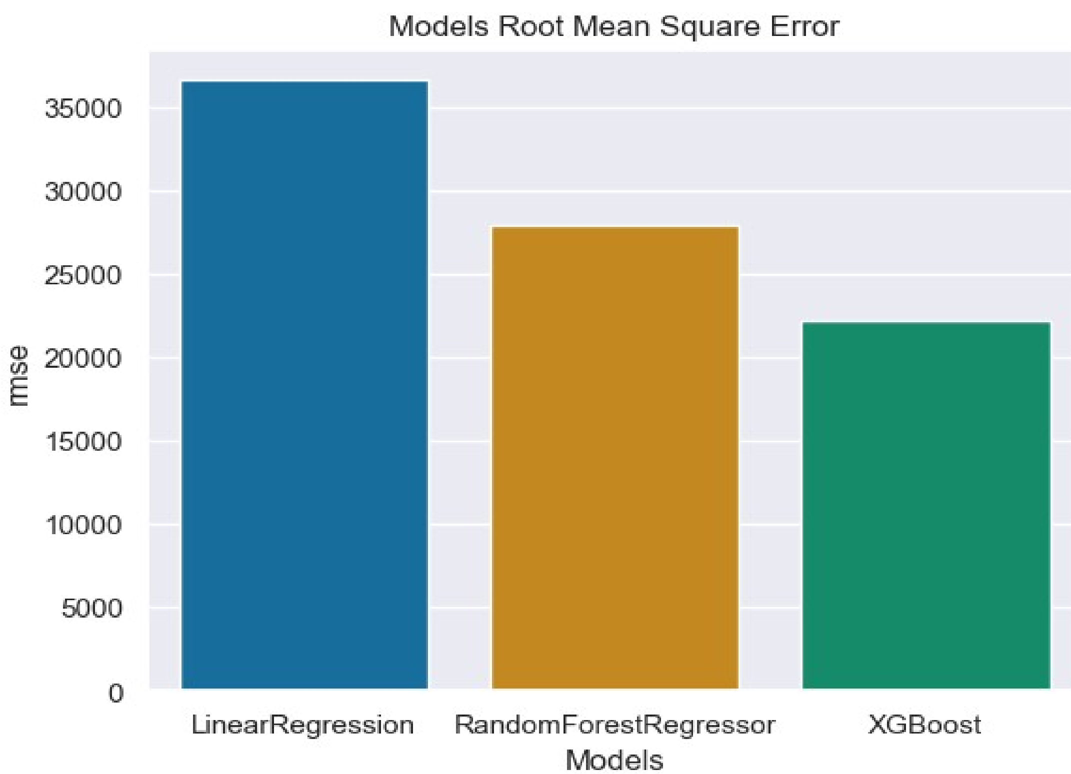
The model identified Overall Quality as the most important feature, confirming our earlier assumption about the significant impact of building features on Sale Price.

Zoning also plays a key role in the model's performance, validating our assumption that location influences Sale Price.

However, Age did not contribute significantly to the model, contradicting our earlier belief that age would affect Sale Price.

5.5 Model Summary

	Models	rmse	R2
0	LinearRegression	36634.566564	0.748403
1	RandomForestRegressor	27896.051650	0.854115
2	XGBoost	22229.279717	0.907365



The bar chart compares the Root Mean Square Error (RMSE) across three models: Linear Regression, Random Forest Regressor, and XGBoost. The RMSE values indicate the model's prediction error, with lower values representing better model performance.

Overall, XGBoost is the most effective model in this comparison, followed by Random Forest, with Linear Regression trailing behind.

6.0 Business Recommendations

To efectively predict house prices and achieve a minimum accuracy of 85% and RMSE of $25,000, the real estate company should focus on the following recommendations for model deployment:

- Track Model Accuracy:

Regularly monitor the accuracy score of the model's predictions to assess the variance between the predicted and actual sale prices. This will help in managing the error in quoted sale prices.

- Set a Target RMSE:

Given that the XGBoost model currently has an RMSE of $22,229 and r-squared of 91%, it's advisable to set a target RMSE of around $25,000. This target accounts for potential overfitting and changes in data trends once the model is in use.

- Adjust Based on Performance:

As data collection continues and the model's performance is evaluated in real-world scenarios, adjust the RMSE target as needed. If the model consistently meets or exceeds the current target, consider raising the target to align with advancements in model accuracy or evolving business needs.

Appendix

MSSubClass: Identifies the type of dwelling involved in the sale.

20 1-STORY 1946 & NEWER ALL STYLES

30 1-STORY 1945 & OLDER

40 1-STORY W/FINISHED ATTIC ALL AGES

45 1-1/2 STORY - UNFINISHED ALL AGES

50 1-1/2 STORY FINISHED ALL AGES

60 2-STORY 1946 & NEWER

70 2-STORY 1945 & OLDER

75 2-1/2 STORY ALL AGES

80 SPLIT OR MULTI-LEVEL

85 SPLIT FOYER

90 DUPLEX - ALL STYLES AND AGES

120 1-STORY PUD (Planned Unit Development) - 1946 & NEWER

150 1-1/2 STORY PUD - ALL AGES

160 2-STORY PUD - 1946 & NEWER

180 PUD - MULTILEVEL - INCL SPLIT LEV/FOYER

190 2 FAMILY CONVERSION - ALL STYLES AND AGES

MSZoning: Identifies the general zoning classification of the sale.

A Agriculture

C Commercial

FV Floating Village Residential

I Industrial

RH Residential High Density

RL Residential Low Density

RP Residential Low Density Park

RM Residential Medium Density

LotFrontage: Linear feet of street connected to property

LotArea: Lot size in square feet

Street: Type of road acces s to property

Grvl Gravel

Pave Paved

Alley: Type of alley access to property

Grvl Gravel

Pave Paved

NA No alley access

LotShape: General shape of property

Reg Regular

IR1 Slightly irregular

IR2 Moderately Irregular

IR3 Irregular

LandContour: Flatness of the property

Lvl Near Flat/Level

Bnk Banked - Quick and significant rise from street grade to building

HLS Hillside - Significant slope from side to side

Low Depression

Utilities: Type of utilities available

AllPub All public Utilities (E,G,W,& S)

NoSewr Electricity, Gas, and Water (Septic Tank)

NoSeWa Electricity and Gas Only

ELO Electricity only

LotConfig: Lot configuration

Inside Inside lot

Corner Corner lot

CulDSac Cul-de-sac

FR2 Frontage on 2 sides of property

FR3 Frontage on 3 sides of property

LandSlope: Slope of property

Gtl Gentle slope

Mod Moderate Slope

Sev Severe Slope

Neighborhood: Physical locations within Ames city limits

Blmngtn	Bloomington Heights
Blueste	Bluestem
BrDale	Briardale
BrkSide	Brookside
ClearCr	Clear Creek
CollgCr	College Creek
Crawfor	Crawford
Edwards	Edwards
Gilbert	Gilbert
IDOTRR	Iowa DOT and Rail Road
MeadowV	Meadow Village
Mitchel	Mitchell
Names	North Ames
NoRidge	Northridge
NPkVill	Northpark Villa
NridgHt	Northridge Heights
NWAmes	Northwest Ames
OldTown	Old Town
SWISU	South & West of Iowa State University
Sawyer	Sawyer
SawyerW	Sawyer West
Somerst	Somerset
StoneBr	Stone Brook
Timber	Timberland
Veenker	Veenker

Condition1: Proximity to various conditions

Artery	Adjacent to arterial street
Feedr	Adjacent to feeder street

Norm	Normal
RRNn	Within 200' of North-South Railroad
RRAn	Adjacent to North-South Railroad
PosN	Near positive off-site feature--park, greenbelt, etc.
PosA	Adjacent to postive off-site feature
RRNe	Within 200' of East-West Railroad
RRAe	Adjacent to East-West Railroad

Condition2: Proximity to various conditions (if more than one is present)

Artery	Adjacent to arterial street
Feedr	Adjacent to feeder street
Norm	Normal
RRNn	Within 200' of North-South Railroad
RRAn	Adjacent to North-South Railroad
PosN	Near positive off-site feature--park, greenbelt, etc.
PosA	Adjacent to postive off-site feature
RRNe	Within 200' of East-West Railroad
RRAe	Adjacent to East-West Railroad

BldgType: Type of dwelling

1Fam	Single-family Detached
2FmCon	Two-family Conversion; originally built as one-family dwelling
Duplx	Duplex
TwnhsE	Townhouse End Unit
TwnhsI	Townhouse Inside Unit

HouseStyle: Style of dwelling

1Story	One story
1.5Fin	One and one-half story: 2nd level finished
1.5Unf	One and one-half story: 2nd level unfinished
2Story	Two story
2.5Fin	Two and one-half story: 2nd level finished
2.5Unf	Two and one-half story: 2nd level unfinished
SFoyer	Split Foyer

SLvl	Split Level

OverallQual: Rates the overall material and finish of the house

10	Very Excellent

9	Excellent

8	Very Good

7	Good

6	Above Average

5	Average

4	Below Average

3	Fair

2	Poor

1	Very Poor

: Rates the overall condition of the house

10	Very Excellent

9	Excellent

8	Very Good

7	Good

6	Above Average

5	Average

4	Below Average

3	Fair

2	Poor

1	Very Poor

YearBuilt: Original construction date

YearRemodAdd: Remodel date (same as construction date if no remodeling or additions)

RoofStyle: Type of roof

Flat	Flat

Gable	Gable

Gambrel	Gabrel (Barn)

Hip	Hip

Mansard	Mansard

Shed Shed

RoofMatl: Roof material

ClyTile Clay or Tile

CompShg Standard (Composite) Shingle

Membran Membrane

Metal Metal

Roll Roll

Tar&Grv Gravel & Tar

WdShake Wood Shakes

WdShngl Wood Shingles

Exterior1st: Exterior covering on house

AsbShng Asbestos Shingles

AsphShn Asphalt Shingles

BrkComm Brick Common

BrkFace Brick Face

CBlock Cinder Block

CemntBd Cement Board

HdBoard Hard Board

ImStucc Imitation Stucco

MetalSd Metal Siding

Other Other

Plywood Plywood

PreCast PreCast

Stone Stone

Stucco Stucco

VinylSd Vinyl Siding

Wd Sdng Wood Siding

WdShing Wood Shingles

Exterior2nd: Exterior covering on house (if more than one material)

AsbShng	Asbestos Shingles
AsphShn	Asphalt Shingles
BrkComm	Brick Common
BrkFace	Brick Face
CBlock	Cinder Block
CemntBd	Cement Board
HdBoard	Hard Board
ImStucc	Imitation Stucco
MetalSd	Metal Siding
Other	Other
Plywood	Plywood
PreCast	PreCast
Stone	Stone
Stucco	Stucco
VinylSd	Vinyl Siding
Wd Sdng	Wood Siding
WdShing	Wood Shingles

MasVnrType: Masonry veneer type

BrkCmn	Brick Common
BrkFace	Brick Face
CBlock	Cinder Block
None	None
Stone	Stone

MasVnrArea: Masonry veneer area in square feet

ExterQual: Evaluates the quality of the material on the exterior

Ex	Excellent
Gd	Good
TA	Average/Typical
Fa	Fair
Po	Poor

ExterCond: Evaluates the present condition of the material on the exterior

Ex Excellent

Gd Good

TA Average/Typical

Fa Fair

Po Poor

Foundation: Type of foundation

BrkTil Brick & Tile

CBlock Cinder Block

PConc Poured Contrete

Slab Slab

Stone Stone

Wood Wood

BsmtQual: Evaluates the height of the basement

Ex Excellent (100+ inches)

Gd Good (90-99 inches)

TA Typical (80-89 inches)

Fa Fair (70-79 inches)

Po Poor (<70 inches

NA No Basement

BsmtCond: Evaluates the general condition of the basement

Ex Excellent

Gd Good

TA Typical - slight dampness allowed

Fa Fair - dampness or some cracking or settling

Po Poor - Severe cracking, settling, or wetness

NA No Basement

BsmtExposure: Refers to walkout or garden level walls

Gd Good Exposure

Av Average Exposure (split levels or foyers typically score average or above)

Mn Mimimum Exposure

No	No Exposure

NA	No Basement

BsmtFinType1: Rating of basement finished area

GLQ	Good Living Quarters

ALQ	Average Living Quarters

BLQ	Below Average Living Quarters

Rec	Average Rec Room

LwQ	Low Quality

Unf	Unfinshed

NA	No Basement

BsmtFinSF1: Type 1 finished square feet

BsmtFinType2: Rating of basement finished area (if multiple types)

GLQ	Good Living Quarters

ALQ	Average Living Quarters

BLQ	Below Average Living Quarters

Rec	Average Rec Room

LwQ	Low Quality

Unf	Unfinshed

NA	No Basement

BsmtFinSF2: Type 2 finished square feet

BsmtUnfSF: Unfinished square feet of basement area

TotalBsmtSF: Total square feet of basement area

Heating: Type of heating

Floor	Floor Furnace

GasA	Gas forced warm air furnace

GasW	Gas hot water or steam heat

Grav	Gravity furnace

OthW	Hot water or steam heat other than gas

Wall	Wall furnace

HeatingQC: Heating quality and condition

Ex Excellent

Gd Good

TA Average/Typical

Fa Fair

Po Poor

CentralAir: Central air conditioning

N No

Y Yes

Electrical: Electrical system

SBrkr Standard Circuit Breakers & Romex

FuseA Fuse Box over 60 AMP and all Romex wiring (Average)

FuseF 60 AMP Fuse Box and mostly Romex wiring (Fair)

FuseP 60 AMP Fuse Box and mostly knob & tube wiring (poor)

Mix Mixed

1stFlrSF: First Floor square feet

2ndFlrSF: Second floor square feet

LowQualFinSF: Low quality finished square feet (all floors)

GrLivArea: Above grade (ground) living area square feet

BsmtFullBath: Basement full bathrooms

BsmtHalfBath: Basement half bathrooms

FullBath: Full bathrooms above grade

HalfBath: Half baths above grade

Bedroom: Bedrooms above grade (does NOT include basement bedrooms)

Kitchen: Kitchens above grade

KitchenQual: Kitchen quality

Ex Excellent

Gd Good

TA Typical/Average

Fa Fair

Po Poor

TotRmsAbvGrd: Total rooms above grade (does not include bathrooms)

Functional: Home functionality (Assume typical unless deductions are warranted)

Typ Typical Functionality

Min1 Minor Deductions 1

Min2 Minor Deductions 2

Mod Moderate Deductions

Maj1 Major Deductions 1

Maj2 Major Deductions 2

Sev Severely Damaged

Sal Salvage only

Fireplaces: Number of fireplaces

FireplaceQu: Fireplace quality

Ex Excellent - Exceptional Masonry Fireplace

Gd Good - Masonry Fireplace in main level

TA Average - Prefabricated Fireplace in main living area or Masonry Fireplace in basement

Fa Fair - Prefabricated Fireplace in basement

Po Poor - Ben Franklin Stove

NA No Fireplace

GarageType: Garage location

2Types More than one type of garage

Attchd Attached to home

Basment Basement Garage

BuiltIn Built-In (Garage part of house - typically has room above garage)

CarPort Car Port

Detchd Detached from home

NA No Garage

GarageYrBlt: Year garage was built

GarageFinish: Interior finish of the garage

Fin Finished

RFn Rough Finished

Unf Unfinished

NA No Garage

GarageCars: Size of garage in car capacity

GarageArea: Size of garage in square feet

GarageQual: Garage quality

Ex Excellent

Gd Good

TA Typical/Average

Fa Fair

Po Poor

NA No Garage

GarageCond: Garage condition

Ex Excellent

Gd Good

TA Typical/Average

Fa Fair

Po Poor

NA No Garage

PavedDrive: Paved driveway

Y Paved

P Partial Pavement

N Dirt/Gravel

WoodDeckSF: Wood deck area in square feet

OpenPorchSF: Open porch area in square feet

EnclosedPorch: Enclosed porch area in square feet

3SsnPorch: Three season porch area in square feet

ScreenPorch: Screen porch area in square feet

PoolArea: Pool area in square feet

PoolQC: Pool quality

Ex Excellent

Gd Good

TA Average/Typical

Fa Fair

NA No Pool

Fence: Fence quality

GdPrv Good Privacy

MnPrv Minimum Privacy

GdWo Good Wood

MnWw Minimum Wood/Wire

NA No Fence

MiscFeature: Miscellaneous feature not covered in other categories

ElevElevator

Gar2 2nd Garage (if not described in garage section)

Othr Other

Shed Shed (over 100 SF)

TenC Tennis Court

NA None

MiscVal: $Value of miscellaneous feature

MoSold: Month Sold (MM)

YrSold: Year Sold (YYYY)

SaleType: Type of sale

WD Warranty Deed - Conventional

CWD Warranty Deed - Cash

VWD Warranty Deed - VA Loan

New Home just constructed and sold

COD Court Officer Deed/Estate

ConContract 15% Down payment regular terms

ConLw Contract Low Down payment and low interest

ConLI Contract Low Interest

ConLD Contract Low Down

Oth Other

SaleCondition: Condition of sale

Normal Normal Sale

Abnorml Abnormal Sale - trade, foreclosure, short sale

AdjLand Adjoining Land Purchase

Alloca Allocation - two linked properties with separate deeds, typically condo with a garage unit

Family Sale between family members

Partial Home was not completed when last assessed (associated with New Homes)