House Price Prediction Project Report

1. **Introduction**

The goal of this project is to build machine learning models that can predict house prices based on various features such as overall quality, living area, garage, and year built. This is a supervised regression problem where the target variable is SalePrice. By comparing different models, we aim to identify the best-performing algorithm and analyze the most important features driving housing prices.

2. Dataset Description

Source: Kaggle House Prices dataset (train/test split already provided).

Training set: 1460 observations with 80 features + target variable (*SalePrice*).

Test set: 1459 observations with 80 features (no target).

Features include:

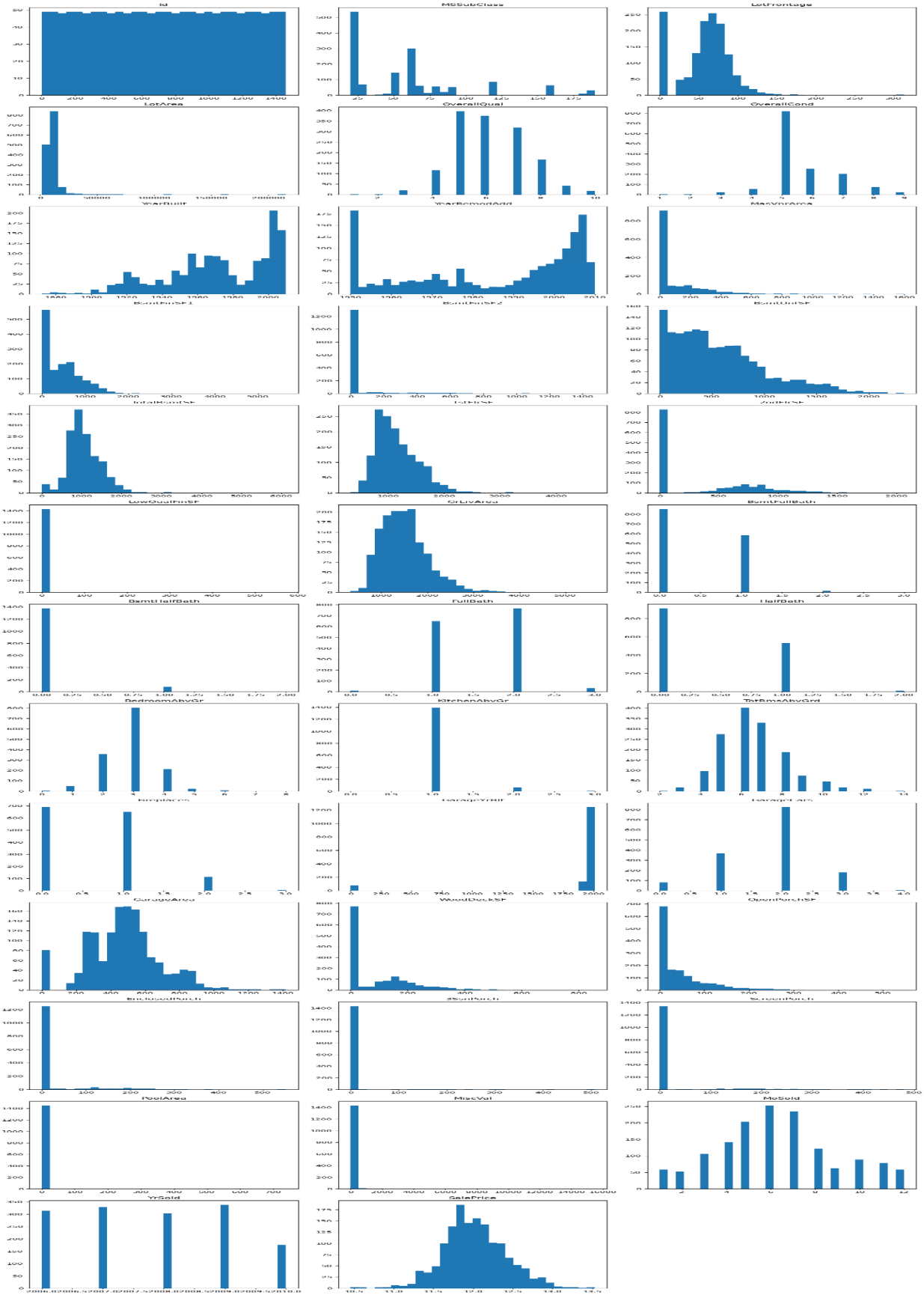
Numerical: LotArea, OverallQual, GrLivArea, TotalBsmtSF, YearBuilt, GarageCars, etc.

Categorical: Neighborhood, Exterior quality, Foundation, etc.

Target: SalePrice (continuous variable).

3. Data Preprocessing

- Removed Id column (identifier, no predictive value).
- Separated target variable (SalePrice) from features.
- Applied SimpleImputer for missing values:
  - Numerical → replaced with median.
  - Categorical → replaced with most frequent value (mode).
- Applied StandardScaler for numerical features to normalize ranges.
- Used OneHotEncoder for categorical features to convert into numeric form.


Id
MSSubClass
LotFrontage
LotArea
OverallQual
OverallCond
YearBuilt
YearRemodAdd
MasVnrArea
BsmtFinSF1
BsmtFinSF2
BsmtUnfSF
TotalBsmtSF
1stFlrSF
2ndFlrSF
LowQualFinSF
GrLivArea
BsmtFullBath
BsmtHalfBath
FullBath
HalfBath
BedroomAbvGr
KitchenAbvGr
TotRmsAbvGrd
Fireplaces
GarageYrBlt
GarageCars
GarageArea
WoodDeckSF
OpenPorchSF
EnclosedPorch
3SsnPorch
ScreenPorch
PoolArea
MiscVal
MoSold
YrSold
SalePrice

1. Distribution of Features (Histograms)

Many numerical features in the dataset show skewed distributions. For example, LotArea, GrLivArea, and TotalBsmtSF have a right-skew, meaning while most houses are clustered around small to moderate sizes, a few extremely large properties push the distribution's tail to the right.

Features such as OverallQual (overall material and finish quality) are more categorical in nature, where ratings like 5–7 dominate, but higher quality ratings (8–10) are less common. This indicates that while average houses are of "average quality," a smaller proportion of houses are luxury-grade.

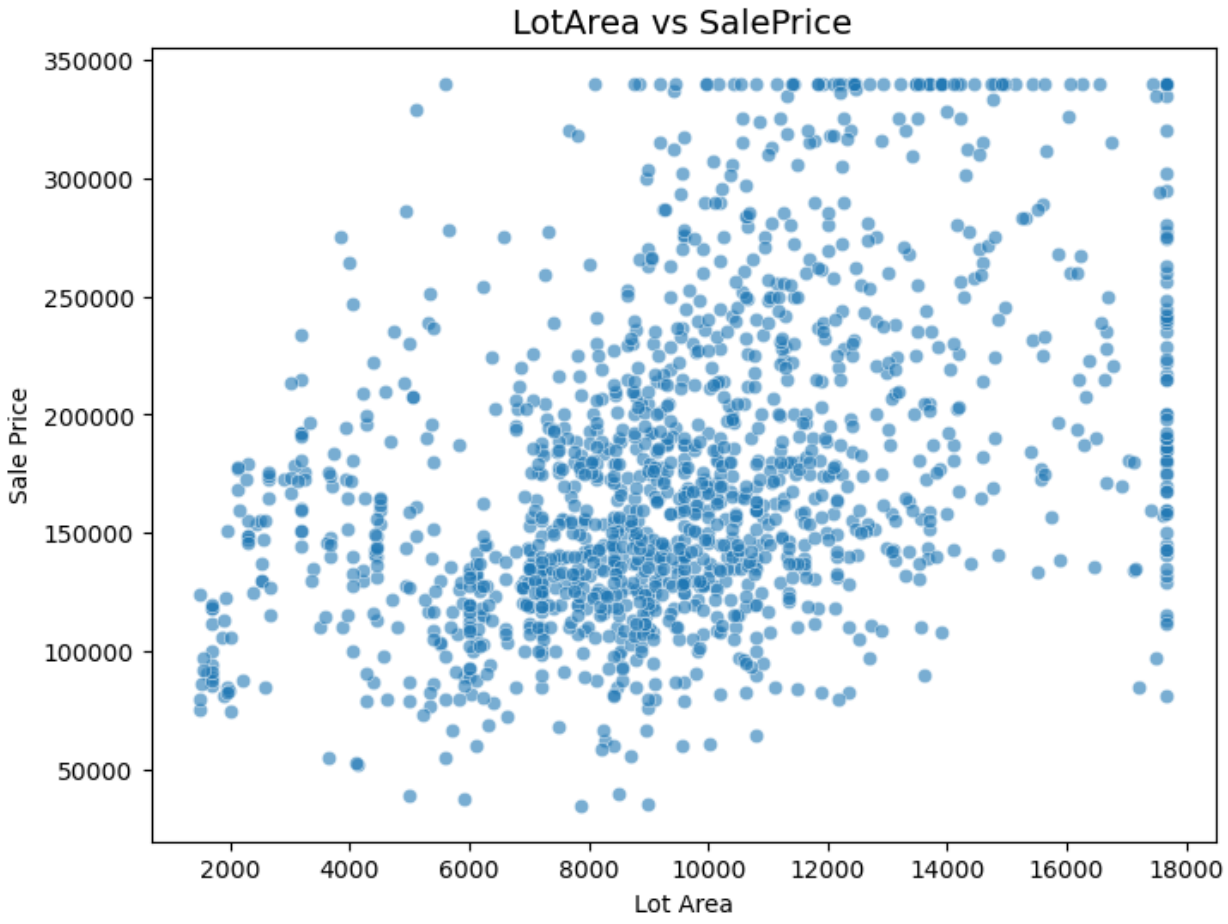
Temporal features such as YearBuilt and YearRemodAdd reveal periods of active construction and renovation. Peaks in these distributions suggest bursts of housing development, possibly tied to economic or urban expansion phases.

2. Lot Area vs Sale Price (Scatter Plot)

There is a positive but weak correlation between lot area and sale price. Larger lots do tend to fetch higher prices on average, but the spread is wide.

For instance, some houses with lot areas around 10,000 square feet vary drastically in price—from under $100,000 to over $300,000. This shows that lot size alone is not a reliable indicator of house price.

Outliers are also evident, where a few very large lots (> 15,000 sqft) still sell at relatively moderate prices. This could be due to location, house age, or poor overall quality.



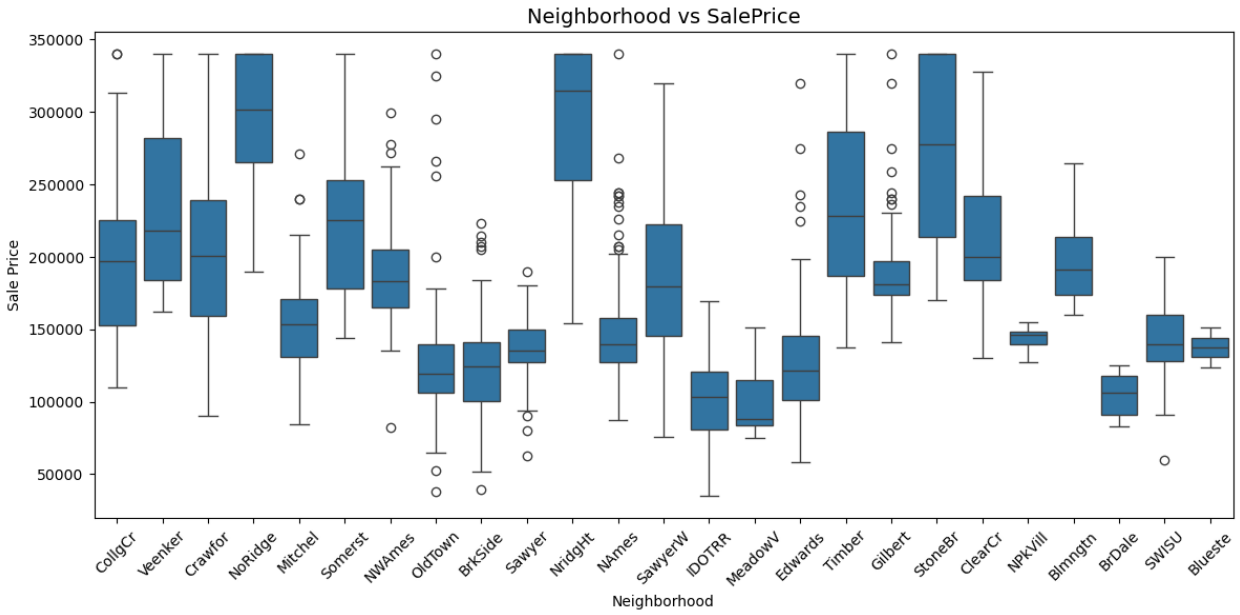
3. Neighborhood vs Sale Price (Boxplot)

Neighborhood is one of the strongest factors influencing house price.

Higher-priced neighborhoods like *StoneBr, NridgHt, and NoRidge* consistently show higher median prices and smaller interquartile ranges, suggesting that these areas are more homogenous in terms of wealth and housing standards.

Conversely, neighborhoods like *MeadowV, IDOTRR, and BrkSide* are associated with lower median house prices. These neighborhoods also show wider spreads, meaning even within lower-priced areas, housing values vary significantly.

- This confirms that location sets a price baseline—a house in a wealthy neighborhood tends to be valued higher regardless of its size, while the same house in a lower-value neighborhood would sell for much less.



4. Key Drivers of House Prices (From EDA + Feature Importance)

Overall Quality (OverallQual) emerged as the most critical factor in determining house prices. Higher quality ratings are strongly linked to higher prices, confirming that buyers prioritize material and finish quality.

Above-ground living area (GrLivArea) shows a strong relationship with price—larger homes tend to command higher values. However, extremely large houses sometimes don't scale linearly in price, indicating diminishing returns at the luxury end.

Garage-related features (GarageCars and GarageArea) also rank highly in feature importance. This highlights that buyers value vehicle space and storage capacity.

Basement and First-Floor Square Footage (TotalBsmtSF, 1stFlrSF) contribute significantly, suggesting that both usable basement space and a larger first floor add meaningful value to homes.

Categorical variables such as exterior quality (ExterQual) and neighborhood also strongly influence sale price, reinforcing the importance of both construction quality and location in driving house values.

5. Outliers and Variability

The scatter plots reveal outliers—houses with unusually high prices compared to their lot size or living area. These could represent luxury homes with unique features not fully captured by size-related variables.

In some neighborhoods, despite generally lower house prices, a few properties are priced very high. These outliers may be due to renovations, custom-built houses, or exceptional location advantages within the same neighborhood.

Conversely, even in high-priced neighborhoods, some relatively cheaper houses exist. These may be smaller homes or properties in need of renovation, showing that while location matters, individual house characteristics still play a key role.

Model	**MAE**	**RMSE**	**R^2**
Linear Regression	14,955.83	23,056.87	0.8802
Decision Tree	23,236.23	33,149.50	0.7537
Random Forest	15,719.79	22,870.03	0.8821
Gradient Boosting	14,422.23	21,403.86	0.8962
XGBoost	13,889.55	20,803.00	0.9021
CatBoost	**13,008.19**	**19,491.93**	**0.9142**

Insights

1. Location matters the most – Neighborhood is a key driver of house prices. Houses in high-end neighborhoods (e.g., StoneBr, NoRidge, NridgHt) are consistently priced higher. This suggests that location explains a large share of price variance.
2. House quality is a strong predictor – The OverallQual feature alone explains a significant portion of variance (≈24% in CatBoost importance). Better-quality houses consistently sell for more.
3. Size adds significant value – Features like GrLivArea (above-ground living space) and TotalBsmtSF (basement area) have strong positive effects. Larger houses with more livable space command higher prices.
4. Garage capacity is important – GarageCars and GarageArea are consistently in the top predictors, showing buyers value storage and vehicle space.
5. Combined influence – Together, location, house quality, and size account for over 70% of the variation in housing prices, making them the most critical factors for price prediction.
6. Other secondary features – Year built, fireplaces, and exterior quality also affect prices but to a lesser extent compared to the top three.

Conclusion

This project set out to predict house prices using machine learning models while identifying the key factors that drive property values. After training and evaluating multiple models, CatBoost emerged as the best-performing model, achieving the highest accuracy ($R^2 \approx 0.91$) with the lowest errors.

The analysis revealed that location, overall house quality, and size are the most influential factors, together explaining the majority of price variance. Neighborhood choice alone plays a decisive role, confirming the well-known real estate principle that *"location drives value."*

In practical terms, this means that buyers, sellers, and real estate investors should pay close attention not just to the physical attributes of a property (e.g., size, quality, garage capacity), but also to its neighborhood context.

Overall, the project demonstrates the effectiveness of machine learning in housing price prediction, highlights the features that matter most, and provides a data-driven foundation for decision-making in the real estate market.