AxxendCorp Data Report

HOUSE PRICE PREDICTION

Introduction (Project Goal)

The goal of this project is to build a machine learning model to predict house prices based on various features such as size, location, and number of rooms. The project demonstrates the full pipeline from data preprocessing, exploratory data analysis (EDA), model training, evaluation, and reporting of insights.

Dataset Description

The dataset used contains 1460 observations with 81 features related to house attributes. These features include numerical, ordinal, and nominal categorical variables describing aspects such as overall quality, living area size, garage capacity, basement area, and more. The dataset contains missing values in several features which required cleaning and imputation.

Data Preprocessing

- Missing values were handled by imputing median values for numerical features (e.g., **LotFrontage**) and filling categorical features with the most frequent values or 'None' where appropriate.

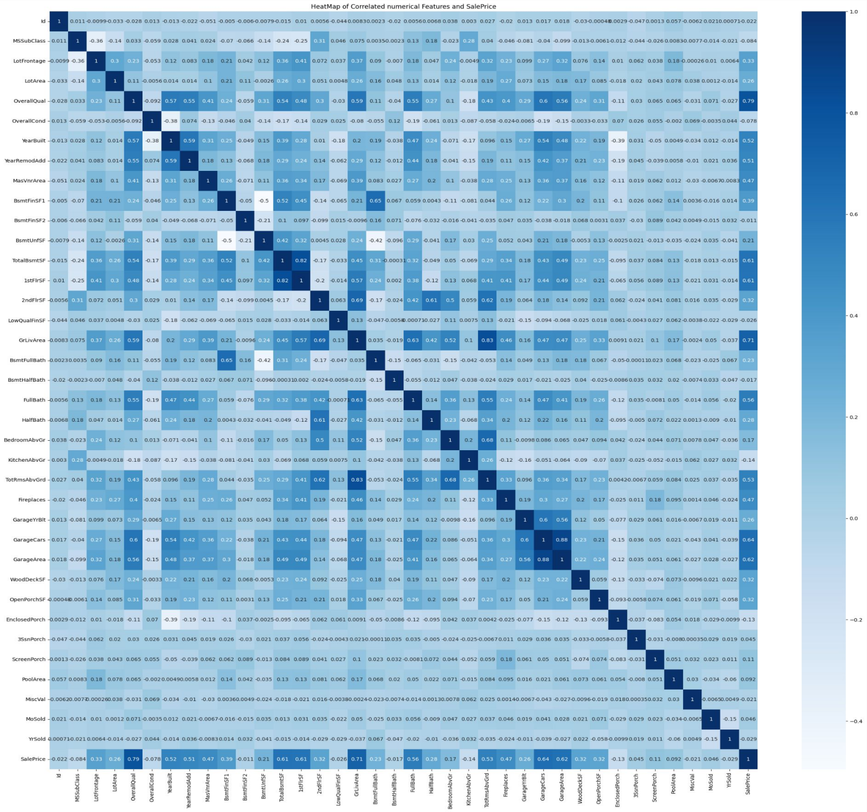
- Categorical features were separated into ordinal and nominal types. Ordinal features were encoded using **OrdinalEncoder** based on meaningful rank, while nominal features were **one-hot encoded.**

- Features with low correlation to the target variable (**SalePrice**) were dropped to reduce dimensionality.
- Outliers in numerical and ordinal features were capped using the **Interquartile Range (IQR) method** to reduce skewness.
- Feature engineering was performed by creating new features such as **HouseAge, RemodAge, SinceRemod,** and **TotalRooms** to capture additional information about the properties.

Correlation of SalePrice to Numeric Features

House prices are most strongly influenced by **overall quality**, **living area size**, and **garage capacity**, showing that buyers prioritize quality and functional space. **Modernity and renovations** also increase value, as newer or recently updated homes sell for higher prices. Amenities such as bathrooms, fireplaces, and outdoor spaces add moderate value, while features like enclosed porches, extra kitchens, or sale year have little to no effect. Overall, quality, space, and updates explain most of the variation in house prices.

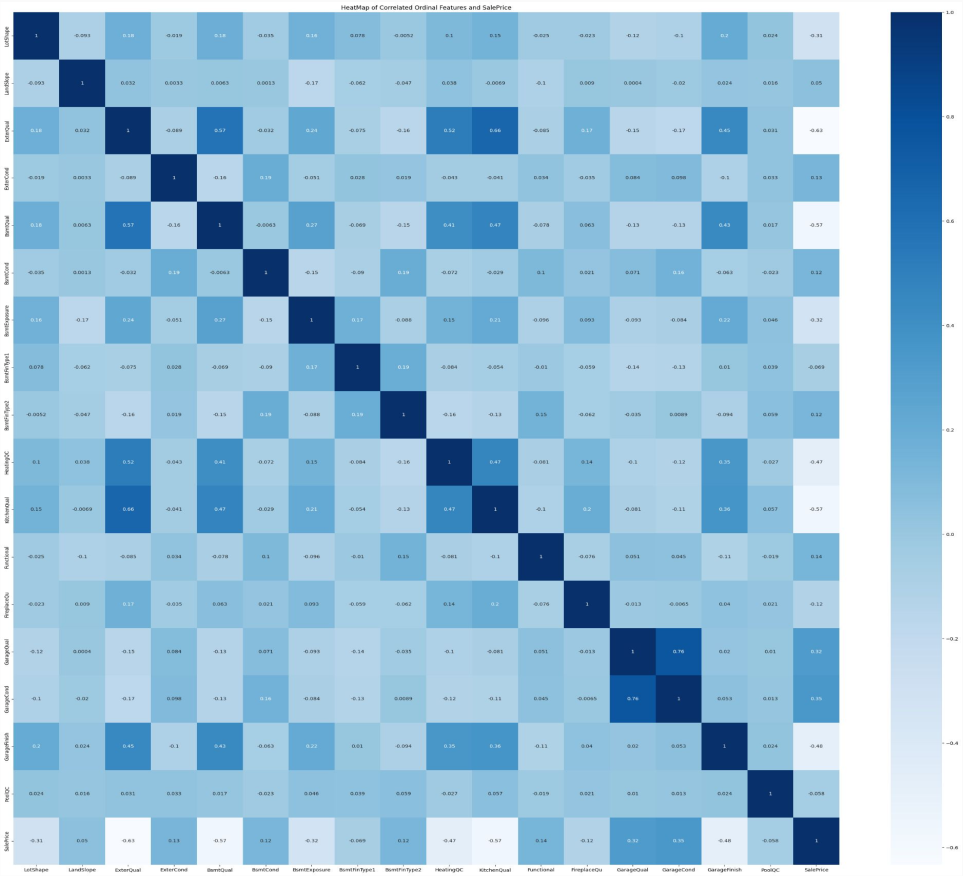
Heatmap of Numeric features and SalePrice



Correlation of SalePrice to Ordinal Features

From the extended correlation analysis, **garage condition/quality**, **basement condition**, and **exterior condition** show weak-to-moderate positive influence, confirming that structural soundness adds value.

Heatmap of Ordinal features and SalePrice



EDA Findings

- Heatmaps of correlations showed strong positive correlations between **SalePrice** and features like **OverallQual, GrLivArea, GarageCars, GarageArea,** and **TotalBsmtSF.**

- Boxplots revealed the distribution and presence of **outliers** in numerical features, which were addressed during preprocessing.

- Feature importance analysis from the **Random Forest model** highlighted the most influential features in predicting house prices.

(Visuals such as heatmaps, boxplots, and feature importance charts are also included in the notebooks.)

Model Performance Table

Model	RMSE	MAE	| R^2 Score |
Linear Regression	*33450.00*	*21068.67*	*0.85*
Decision Tree	*44337.39*	*28976.40*	*0.74*
| Random Forest	*29375.16*	*18401.12*	*18401.12*

Model Performance Insights

We tested three models to predict house prices.

- **Random Forest's results (R^2 = 0.89, MAE ≈ $18K)** are actually strong — it means the model is right about **9 times out of 10**, with relatively small mistakes compared to the full house value.

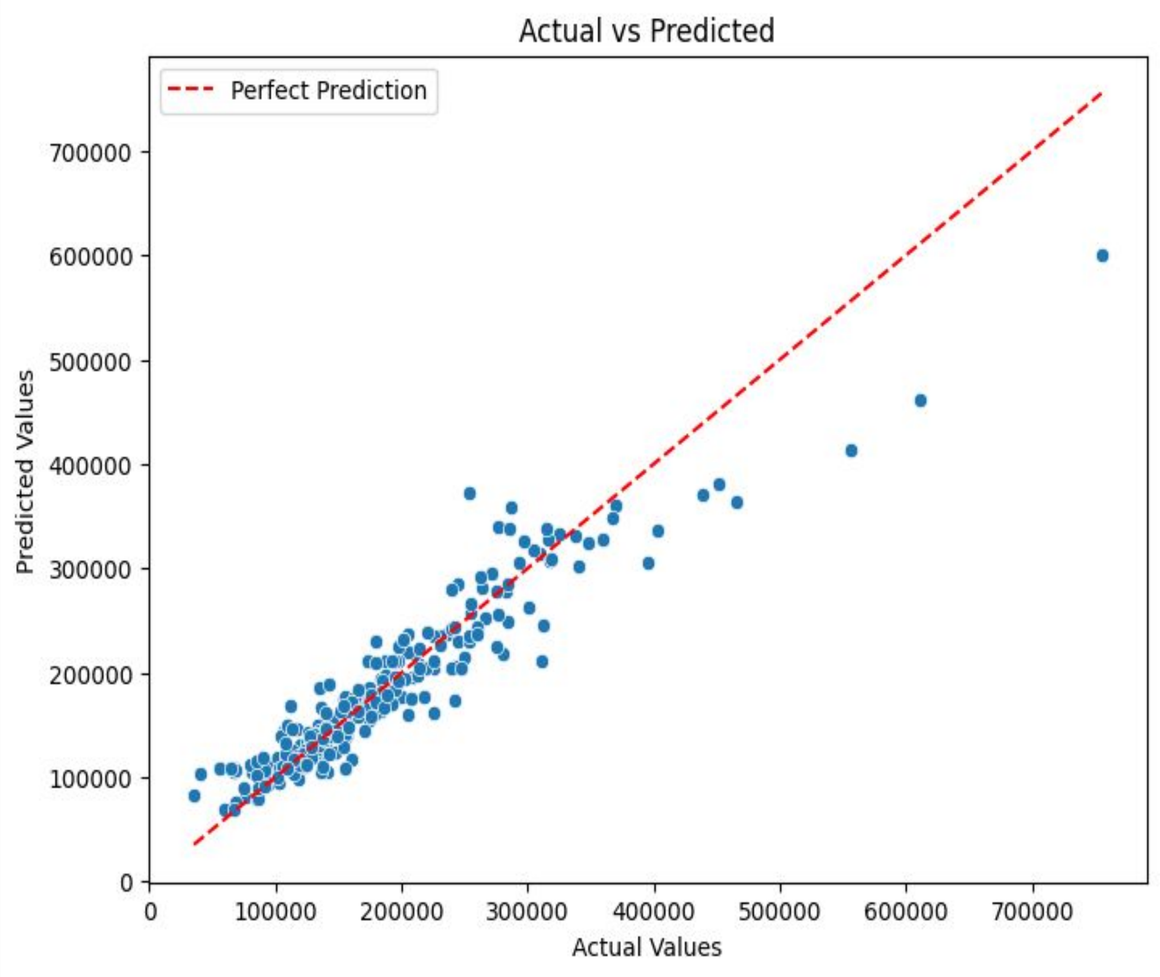
- **Linear Regression** is okay (85% accuracy).

- **Decision Tree** isn't great on its own, but it's often used as a building block for stronger models like Random Forest.

Overall, Random Forest gave the most reliable predictions, while the Decision Tree was the weakest performer.

Model Performance on Actual Values vs Predicted Values

This graph illustrates the performance of the best model on the test data. The predicted house prices align closely with the actual values, forming a tight cluster along the diagonal line. This indicates that the model has successfully learned the underlying patterns in the data and is making reliable predictions.



Overall Insights

- **house quality** and other major features explain a significant portion of the variance in house prices.

- **OverallQual** is the most correlated feature with SalePrice, indicating that the overall material and finish quality strongly influences price.

- Larger living areas (**GrLivArea**), garage capacity (**GarageCars**), garage area (**GarageArea**), and basement size (TotalBsmtSF) are positively correlated with higher sale prices.

- **The Random Forest model,** after hyperparameter tuning, achieved the best performance with the highest R^2 score and lowest RMSE and MAE, indicating strong predictive capability.

The top 3 features affecting the price of the houses are **Overall Quality**, **Above-Ground Living Area** and **Garage Capacity**

Thank you !

AMEND

YOUR PARTNER FOR SUCCESS