**House Price Prediction Data Analysis Report**

**1. Introduction**

House price prediction plays a crucial role in the real estate industry by assisting buyers, sellers, and investors in making informed decisions. Understanding the factors that impact property value helps in setting appropriate prices and predicting market trends.

This project focuses on analysing various features of residential properties to determine their impact on house prices. By leveraging **Python-based data analysis**, we explore how features such as **location, house condition, basement size, renovations, and categorical variables** influence house prices.

The project also investigates whether house prices can be predicted using only categorical features without numerical data, providing valuable insights into real estate price modelling.

**2. Methodology**

To conduct a structured data analysis, the following steps were followed:

1. **Data Exploration** – Understanding the dataset, feature types, and identifying missing values.
2. **Data Cleaning & Pre-processing** – Handling missing values, encoding categorical variables, and transforming numerical data.
3. **Feature Engineering** – Creating new features such as house age, years since renovation, and transformations for skewed variables.
4. **Visualization & Statistical Analysis** – Using various charts and correlation matrices to derive insights.
5. **Machine Learning Model** – Building a predictive model using only categorical features to assess their impact.
6. **Model Evaluation** – Measuring the model’s performance using metrics like **Mean Absolute Error (MAE)** and **R-squared Score**.

**3. Requirement Analysis**

**3.1 Tools and Libraries Used**

The following tools and libraries were used for data analysis and modelling:

* **Python Libraries:**
  + Pandas for data manipulation
  + NumPy for numerical computations
  + Matplotlib & Seaborn for data visualization
  + Scikit-learn for machine learning and evaluation metrics
* **Visualization Tools:**
  + Histograms, Boxplots, Scatter plots, Heatmaps, and Bar charts to analyse data.
* **Machine Learning Model:**
  + Linear Regression was used to predict house prices using categorical features.

**3.2 Dataset Overview**

The dataset includes various **categorical** and **numerical** features:

* **Categorical Features:** MS Zoning, BldgType, LotConfig, OverallCond, Exterior1st
* **Numerical Features:** LotArea, TotalBsmtSF, SalePrice, YearBuilt, YearRemodAdd
* **Target Variable:** SalePrice

**4. Data Cleaning and Feature Engineering**

**4.1 Handling Missing Values**

* **Numerical features**: Imputed using **mean** or **median**.
* **Categorical features**: Filled with the **mode** (most frequent value).
* **Columns with excessive missing values** were dropped to maintain dataset integrity.

**4.2 Encoding Categorical Variables**

* **One-Hot Encoding**: Converted categorical variables into binary dummy variables.
* **Label Encoding**: Used for ordinal categorical features like OverallCond.

**4.3 Scaling and Transformation**

* **Log Transformation** applied to skewed numerical variables.
* **Standardization (Z-score Scaling)** for normally distributed features.
* **Normalization (MinMaxScaler)** for feature scaling where necessary.

**4.4 Handling Temporal Variables**

* **House Age** = Current Year - YearBuilt
* **Years Since Last Renovation** = Current Year - YearRemodAdd

**5. Data Visualization & Insights**

**5.1 Sale Price Analysis**

* **Distribution of SalePrice** was examined to detect skewness.
* **Boxplots and histograms** were used to visualize price variations.

**5.2 Impact of Location on SalePrice**

* **Boxplots for** MSZoningandLotConfig showed significant variations in pricing based on location.

**5.3 Effect of House Condition**

* **Boxplot of** OverallCondvs**.** SalePrice revealed that well-maintained houses generally had higher prices.

**5.4 Relationship Between Basement Size and SalePrice**

* **Scatter plot of** TotalBsmtSFvs**.** SalePrice indicated a strong positive correlation.

**5.5 Effect of Renovation on SalePrice**

* **Scatter plot of** YearsSince Last Renovation vs. SalePrice showed that recently renovated homes tend to have better prices.

**6. Predictive Modelling**

**6.1 Model Selection**

* **Linear Regression** was used to predict SalePrice using only categorical features.
* **One-Hot Encoding** was applied to categorical variables before training the model.

**6.2 Model Evaluation**

* **Mean Absolute Error (MAE):** Measured prediction accuracy.
* **R-Squared Score:** Evaluated model fit.

**7. Insights and Key Findings**

1. **Location Matters:** Certain zoning classifications (MSZoning) had significantly higher Sale Prices.
2. **House Condition Affects Price:** Well-maintained homes (higher OverallCond) had better pricing.
3. **Basement Size is Important:** Larger basements (TotalBsmtSF) correlated positively with Sale Prices.
4. **Renovation Helps:** Recently renovated houses had improved SalePrices.
5. **Predicting House Prices with Only Categorical Features:** The model provided some predictive power, but numerical features would improve accuracy.

**8. Conclusion & Future Scope**

This project successfully analysed house prices and their key influencing factors. **Feature engineering, statistical visualization, and machine learning** were used to derive insights from the dataset.

While categorical features alone provided some predictive power, incorporating numerical variables like square footage, lot size, and year built could significantly improve model performance. **Future enhancements** could include:

* Using advanced models like **Random Forest, Decision Trees, or Neural Networks**.
* Implementing **feature selection techniques** to optimize model performance.
* Conducting **geospatial analysis** for better location-based pricing predictions.

This study provides a solid foundation for real estate pricing analysis and serves as a valuable reference for further exploration in predictive modelling.

**End of Report**