**🏠House Price Prediction – Project Report**

**📌 Project Objective:**

The primary goal of this project is to **predict house prices** using various features such as the number of rooms, location coordinates, type of property, and others. The project follows a structured machine learning workflow that includes data preprocessing, visualization, model training, evaluation, and prediction.

**📁 Dataset Overview:**

* The dataset contains real estate listings from **Melbourne**.
* It includes features like:
  + Rooms, Bathroom, Car, BuildingArea, YearBuilt
  + Suburb, Type, Method, CouncilArea, Lattitude, Longtitude
  + The target variable is Price.

**🧹 Data Preprocessing:**

* **Missing Value Treatment**:
  + BuildingArea, YearBuilt, and Car columns were imputed using **IterativeImputer**.
  + CouncilArea was filled with 'Unknown' using **SimpleImputer**.
* **Verification**:
  + After imputation, it was confirmed that there were **no missing values** left in the dataset.

**📊 Data Visualization:**

* **Heatmap** was used to identify correlation between numeric features.
* A custom function visualize\_outliers() was used to display **boxplots and histograms**, especially to explore **outliers and skewness** in Price, Rooms, and other key features.
* **Count plots** helped visualize the distribution of categories like CouncilArea and Property Type.
* **Scatter plots** using both **Seaborn** and **Plotly** helped assess the relationship between:
  + Rooms, Price, and Type
  + Lattitude, Longtitude, and Price (geographical price distribution)

**🧠 Model Building:**

* **Target Variable**: Price
* The model was built using:
  + **Random Forest Regressor**
  + Dataset was split into training and testing sets using train\_test\_split.

**✅ Model Evaluation:**

* Evaluation metrics included:
  + **Mean Absolute Error (MAE)**
  + **Root Mean Squared Error (RMSE)**
  + **R² Score**
* **Visualizations**:
  + A **stacked bar plot** was created to compare y\_test vs y\_pred, along with the error magnitude.
  + This provided a clear visual understanding of where predictions were accurate and where large errors occurred.

**📌 Key Insights:**

* **Outliers** in the Price column were evident and may need special handling (e.g., log transformation or removal) to improve model performance.
* The Price distribution is **right-skewed**.
* Features like Rooms, BuildingArea, and Lattitude/Longtitude show strong correlation with house prices.

**🛠 Future Improvements:**

* Perform **feature engineering** to extract meaningful attributes.
* Experiment with **log transformation** or **scaling** on Price.
* Try **advanced models** like XGBoost or LightGBM.
* Perform **hyperparameter tuning** using GridSearchCV.