**HOUSING DATASET**

**USING ADVANCE PYTHON**

**N. PANDIDURAI**

**DA &DS-MAY 2025**

**REINFORCEMENT PROJECT-1**

**23-07-2025**

**(OFFLINE)**

**\*\*\*\*\*\***

**Table of Content**

1**. Introduction**

**2. Aim**

**3. Problem Statement**

**4. Project Workflow**

**5. Data Understanding**

**6. Data Cleaning**

**7. Obtaining Derived Metrics**

**8. Filtering Data for Analysis**

**9. Statistical Analysis**

**10. Exploratory Data Analysis (EDA)**

* **Univariate Analysis**
* **Bivariate Analysis**
* **Multivariate Analysis**

**11Data Visualization**

**12. Overall Insights from Analysis**

**13. Conclusion**

**1.Introduction:**

The housing dataset provides a comprehensive snapshot of real estate properties with detailed information on their physical characteristics, location, and pricing. It includes attributes such as the number of bedrooms and bathrooms, total living area, lot size, presence of waterfront views, condition, renovation status, and geographic information like city, zip code, and country.

In this data set contains column likes:

* 'bathrooms',
* 'bedrooms',
* 'city',
* 'condition',
* 'country',
* 'date',
* 'floors',
* 'price',
* 'Sqft\_above',
* 'Sqft\_basement',
* 'Sqft\_living',
* 'Sqft\_lot',
* 'statezip',
* 'Street',
* 'view',
* 'waterfront',
* 'Yr\_built',
* 'Yr\_renovated'.

With this data are contain in my data set reinforcement project in python. With a mix of numerical, categorical, and temporal variables, the dataset offers a real-world opportunity to apply data science Like Python we used to build code.

**2.Aim:**

The aim of this project is to analyze and interpret housing data to identify the key factors that influence property prices, and to provide actionable insights through like,

* **Data cleaning,**
* **Exploratory analysis,**
* **Visualization, and**
* **Feature engineering.**

This analysis supports better decision-making for real estate stakeholders such as investors, buyers, sellers, and developers.

**3.Problem statement:**

In the real estate industry, determining the fair market value of a property is a complex task influenced by multiple factors such as location, size, condition, and amenities. Without data-driven insights, property valuation often relies on subjective judgment, leading to inaccurate pricing and missed investment opportunities.

This project aims to analyze a comprehensive housing dataset to uncover the most significant factors affecting property prices, detect trends and anomalies, and support better forecasting and decision-making through exploratory data analysis, feature engineering, and visualizations.

**4. Data Collection:**

Import the housing dataset (housing.csv) into the Python environment using Pandas.

**5.Project Workflow:**

**Step 1**: Data Loading – Load dataset using pandas.

**Step 2:** Data Cleaning and Preprocessing – Handle missing values, correct inconsistencies, encode categorical variables.

**Step 3**: Exploratory Data Analysis (EDA) – Univariate, bivariate, multivariate analysis.

**Step 4**: Outlier Detection and Handling – Use boxplots and IQR method.

**Step 5**: Feature Engineering – Create age of property, price per square foot.

**Step 6**: Visualization – Scatter plots, histograms, heatmaps, trend lines.

**Step 7:** Analysis and Interpretation – Identify price drivers and key patterns.

**6.Data cleaning:**

**Handle Missing Values** – Identify and fill or remove missing data.

**Fix Data Types** – Ensure each column has the correct type (e.g., dates, numbers).

**Remove Duplicates** – Eliminate duplicate rows.

**Normalize Categorical Values** – Standardize formats (e.g., uppercase/lowercase).

**Correct Outliers or Invalid Entries** – Identify extreme or incorrect values.

**7.Exploratory Data Analysis (EDA):**

EDA is the process of analyzing datasets to summarize their main characteristics using statistical and visual methods. It helps uncover:

* **Univariate Analysis**
* **Bivariate Analysis**
* **Multivariate Analysis**
* **Univariate Analysis:**
* Distribution of individual variables.
* Univariate analysis involves examining one variable at a time to understand its distribution, central tendency (mean, median, mode), spread (variance, standard deviation), and frequency.
* **Bivariate Analysis**
* Relationships like Price vs Sqft\_living, Bedrooms vs Bathrooms
* It is comparing of two columns
* **Multivariate Analys**
* Multivariate analysis examines three or more variables simultaneously to understand interactions, patterns, and relationships among them.
* Its comparing more then three data.

**8.Data Visualization:**

Data Visualization is the graphical representation of data to:

* **Make patterns and trends more understandable**
* **Highlight relationships between variables**
* **Simplify complex data**
* **Support decision-making**

**9.Overall Insights from Analysis**

* **Properties near waterfront or with better view/condition tend to be priced higher.**
* **Larger living space (sqft\_living) shows strong positive correlation with price.**
* **Renovated houses fetch better prices.**
* **Certain cities consistently show higher property values.**

**10.conclusion:**

This analysis reveals the significant factors influencing housing prices and helps stakeholders make informed decisions. The Python-based workflow is modular and

can be extended for predictive modeling or real-time price estimation in future projects.

**Tools and Libraries Used**

python (Jupiter Notebook)

* + Pandas, NumPy – for data handling
  + Matplotlib, Seaborn – for visualization