**Dataset Report and Documentation**

**Dataset Overview**

This dataset consists of **21 columns and 21,613 rows**, containing detailed information about **residential properties**. It is commonly used for **house price prediction**, **real estate market analysis**, and **location-based insights**.

**Column Overview:**

| **Column Name** | **Description** |
| --- | --- |
| **Date:** | Unique identifier for each employee |
| **Price:** | The sale price of the house (in USD). |
| **Bedrooms:** | Number of bedrooms in the house. |
| **Bathrooms:** | Number of bathrooms in the house (can include half-baths) |
| **Sqft\_living:** | Interior living space in square feet. |
| **Sqft\_lot:** | Number of years of professional experience |
| **Floors:** | Number of floors (levels) in the house. |
| **Waterfront** | Indicates if the house has a waterfront (0 = no, 1 = yes). |
| **View:** | Quality of the view from the house (0–4). |
| **Condition:** | Overall condition of the house (1–5). |
| **Grade:** | Overall grade of the house (based on King County grading system, 1–13). |
| **Sqft\_above:** | Square footage of the house above ground level (no basement). |
| **Sqft\_basement:** | Square footage of the basement. |
| **Year\_built:** | The year the house was originally built. |
| **Year\_renovated:** | The year the house was last renovated (0 if never renovated). |
| **Zipcode:** | The ZIP code where the house is located. |
| **Lat:** | Latitude coordinate of the house’s location. |
| **Long:** | Longitude coordinate of the house’s location. |
| **Sqft\_living15:** | Average interior living space (in sqft) of homes in the same neighborhood (based on 15 nearest neighbors). |
| **Sqft\_lot15:** | Average lot size (in sqft) of homes in the same neighborhood (based on 15 nearest neighbors). |
| **Year:** | Extracted sale year from the date column. |
| **Month:** | Extracted sale month from the date column (1–12). |
| **Day:** | Extracted sale day from the date column (1–31). |
| **Weekday:** | Extracted weekday from the date column (e.g., Monday=0, Sunday=6). |

**Data Preprocessing**

**Unique Values Handling**

| **Column** | **Missing Values Count** |
| --- | --- |
| Zipcode: | 70 unique values |
| Waterfront: | 2 unique values |
| View: | 5 unique values |
| Condition: | 5 unique values |
| Grade: | 12 unique values |

**Dropping Column**

We dropped certain columns that we thought unnecessary for the analysis and they are not directly connected to target column

Dropped Columns: Id , Data.

**Missing Values**

No columns have missing values in this dataset.

**Duplicate values**

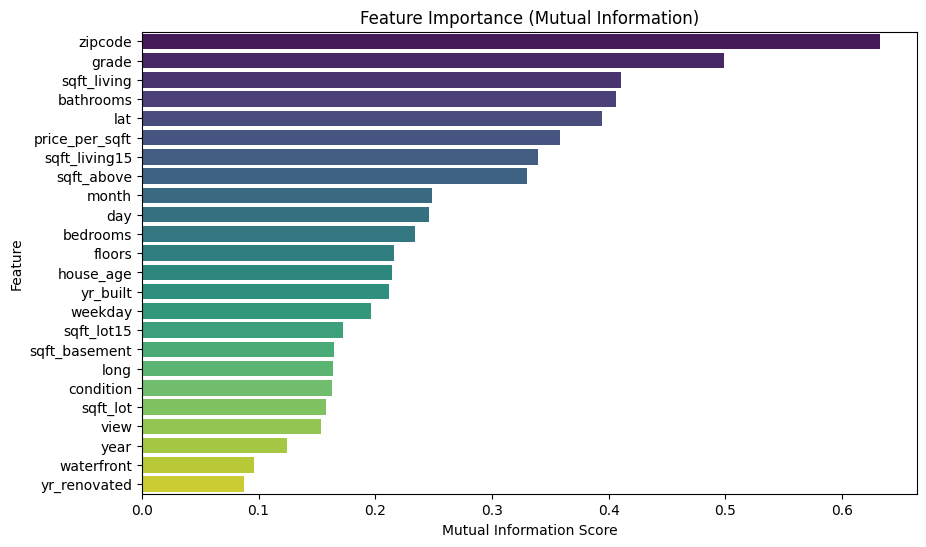
No columns have duplicate values too in this dataset.

**Encoding**

For handling categorical variables, **Label Encoding** was used for all categorical features to convert them into numerical values.

**Used Klib:** for data cleaning and reduced size of dataset

**Mutual Information:**



**Feature Enginering:**

**Created new features:**

| **Old feature** | **New = 1** | **New = 2** |
| --- | --- | --- |
| **house\_age** | **year** | **yr\_built** |
| **price\_per\_sqft** | **price** | **sqft\_living** |

**Scaling: Standard Scaling**

**1)First Method:** Only specific features were scaled, while others remained in their original form.

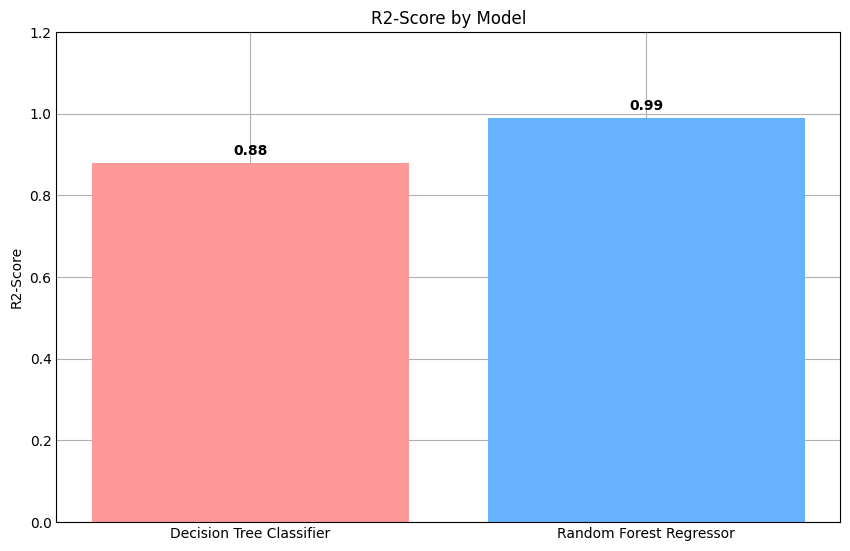
**2)Second Method:** All numerical columns, except price, were scaled to ensure consistency across features.

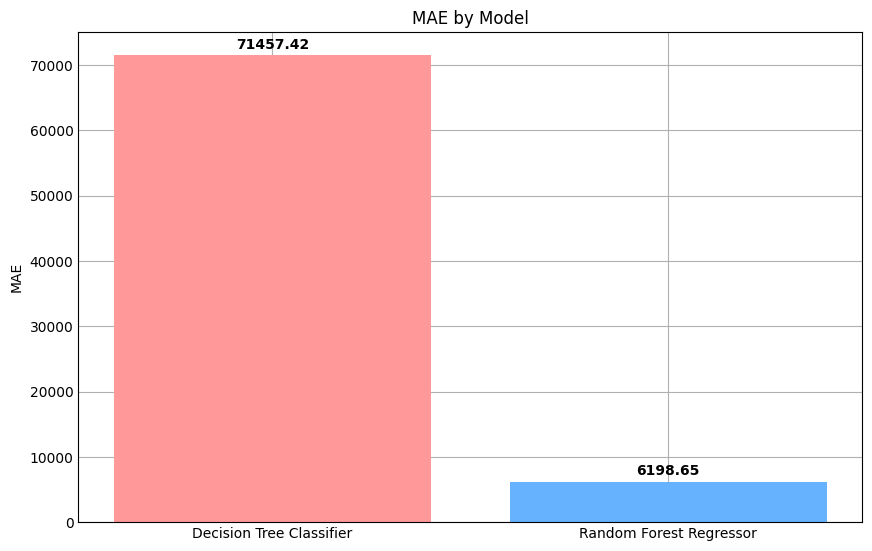
**Pipeline:**

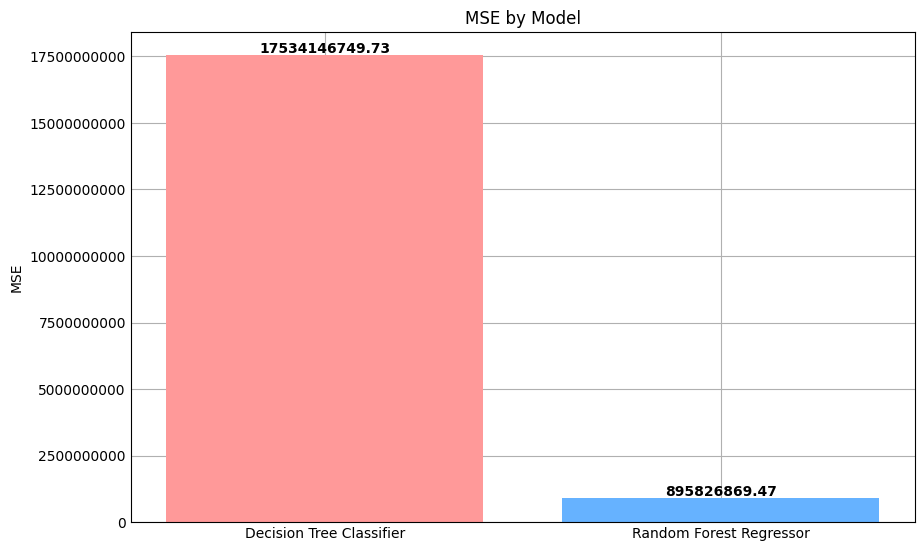
Pipeline built using **Standard Scaler** for feature scaling and **Linear Regression** for modeling.

**Model Performance Summary:**

| **Model** | **R2-score** | **MAE** | **MSE** | **RMSE** |
| --- | --- | --- | --- | --- |
| **Decision Tree Classifier** | 0.88 | 71457.42 | 17534146749.73 | - |
| **Random Forest Regressor** | 0.99 | 6198.65 | 895826869.47 | 29930.36 |







**Summary:**

**Random Forest Regressor** is the superior model across all metrics, offering higher accuracy and lower prediction errors.

1)**R**²-Score Comparison:****

* The **Random Forest Regressor (0.99)** significantly outperforms the **Decision Tree Classifier (0.33)** in explaining variance, indicating a much better fit.

**2)**Mean Absolute Error (MAE):****

* The **Random Forest Regressor (6,198.65)** has a much lower MAE than the **Decision Tree Classifier (71,457.42)**, meaning it makes significantly smaller average errors.

**3) **Mean Squared Error (MSE):****

* The **Random Forest Regressor (895,826,869.47)** has a much lower MSE compared to the **Decision Tree Classifier (17,534,146,749.73)**, showing that the individual errors are much smaller.