Data Set Description:

**Data Set name**: House price prediction

The dataset consists of 36 columns of house features such as number of rooms, area, location, price, and more.

**Types of Data**: It contains both **numerical** and **categorical** data.

1**. Numerical Data**: price (price), bedrooms (number of bedrooms), bathrooms (number of bathrooms), sqft\_living (living area), and more.

2. **Categorical Data:** Street, city, date, and more**.**

**Total Number of samples: 4600 sample.**

**Number of Testing samples: 20% (920 sample).**

**Number of training samples: 80% (3680 sample).**

**Missing Data:**

The data set has missing data in some columns (bedrooms, bathrooms, floors, street, city, date).

The Imputation technique was used to handle these missing data

(Mean for numerical data and Mode for Categorical data).

Comparison between Algorithms:

1. Linear regression:

Linear Regression assumes a **linear relationship** between input features (X) and the target variable (Y).

it minimizes the sum of squared errors (**Mean Squared Error**).

1. KNN:

KNN assumes no specific relationship between input features and the target variable. It predicts the target value for a test point based on the **average of the k-nearest points** in the training data.

So, the efficient algorithm in our case (house price prediction) is

The linear Regression algorithm.

**Evaluation Metrics**

|  |  |  |
| --- | --- | --- |
| **Metric** | **Linear regression** | KNN |
| R-squared | 0.05 | -0.09 |
| MSE | 2.03 | 2.31 |
| MAE | 0.49 | 0.57 |

A graph of a graph with different colored bars

Description automatically generated with medium confidence