Numerical :

Link : <https://www.kaggle.com/datasets/harlfoxem/housesalesprediction>

Name:

**House Sales in King County, USA**

The shape of dataset is 21613 row & 21 column

Add the column discribtion from this link in kaggel <https://www.kaggle.com/datasets/harlfoxem/housesalesprediction/discussion/207885>

The dataset is clean and the data types are correct

We can add some of information from the describe function in the notebook as the min & max & mean of price and the same in sqft\_living

Preprocessing:

### we see that

### Most Values of these Columns (yr\_renovated,view,waterfront) is zero ,so we gonna drop them[¶](https://kkb-production.jupyter-proxy.kaggle.net/static/dist/jupyterlab/jupyterlab-index-6cd47fde172ee47cc970.html?session=eyJhbGciOiJub25lIiwidHlwIjoiSldUIn0.#Most-Values-of-these-Columns-(yr_renovated,view,waterfront)-is-zero-,so-we-gonna-drop-them)

Add image drope 1 & drope 2

See the distribution

* Using histplot from seaborn Add the distribution\_before.png
* Using boxplot from seaborn Add the outlier\_before.png

# We scaling data with Standard Scaler

# To all column and ignore this column "grade","condition","floors","bathrooms","bedrooms", date

# Because values in this column is categorical

# transform data with power transform (method = yeo-johnson ) to make distribution of data normalized

# Remove outliers using IQR

After the preprocessing the shape of data is 20698 row , 17 column

See the distribution

* Using histplot from seaborn Add the distribution\_after.png
* Using boxplot from seaborn Add the outlier\_after.png
* We see the correlation and drop the column that have the multicollinearity This is the column 'sqft\_above','sqft\_living15','sqft\_lot15','date'

We split the data to 25% test and 75% train

We using LinearRegression with r2\_score is 73%

Add the mean\_squared\_error & mean\_absolute\_error & r2\_score in a table for each model (“I will give you the زتونه”)

Add cross validation for each model

Add thanks