Feature Selection Vs Feature Extraction

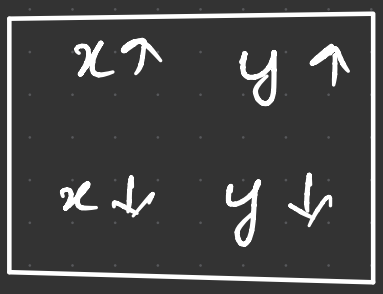
① Why Dimensionality Reduction?

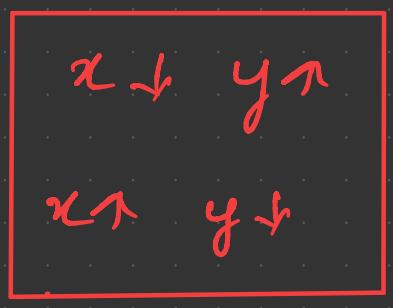
* Prevent curse of dimensionality
* Improve the performance of model
* Visualize the data → to understand the data.

Feature Selection

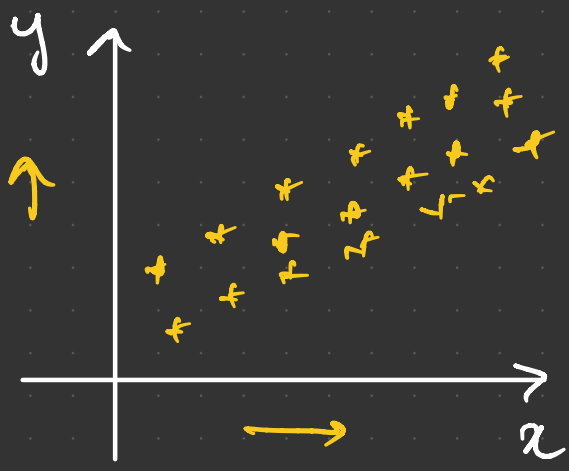
Feature selection is the process which helps to select the most important features which actually help us to predict the output.

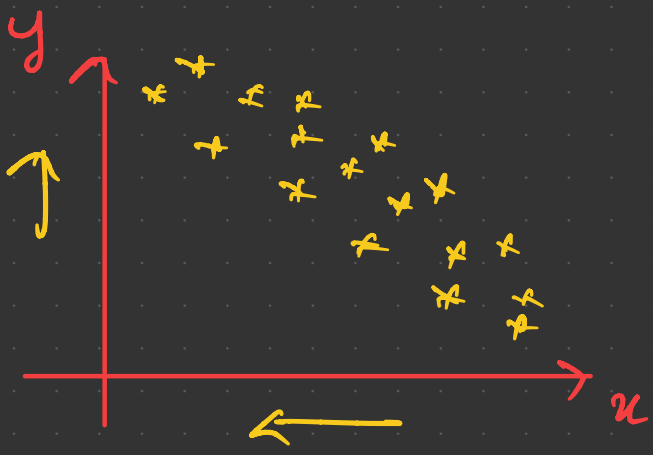
I/P	O/P
x	y
-	-
-	-
-	-





these are the 2 possible relation I can have with x and y





If you see the above plot we have the linear relationship. x value will be helpful to find out the output y Mathematically we can quantify the relationship. the technique we use in this specific case is covariance.

$$cov(x, y) = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{N - 1}$$

the outcome of this either can be positive or negative.

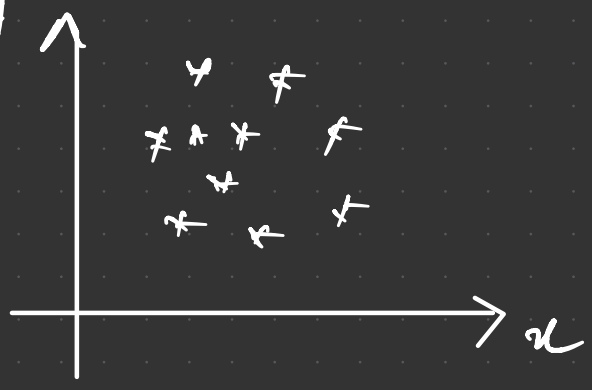
If the outcome is positive then it follows linear relationship,

If the outcome is negative then it follows inverse linear relationship

If the covariance is $\simeq 0$, then there is no relation between x and y.

⇓

If I try to plot, it will be in a circular way. so that we can remove such features from the dataset.



Pearson correlation

$$\frac{cov(x, y)}{\sigma_x \, \sigma_y}$$

correlation ranges from

-1 to $+1$

the more the value of $+1$ the more $+ve$ correlation x and y.

the more the value of -1 the more $-ve$ correlation of x and y.

If the correlation is $\simeq 0$, then this means no relationship.

housing Dataset

Independent features

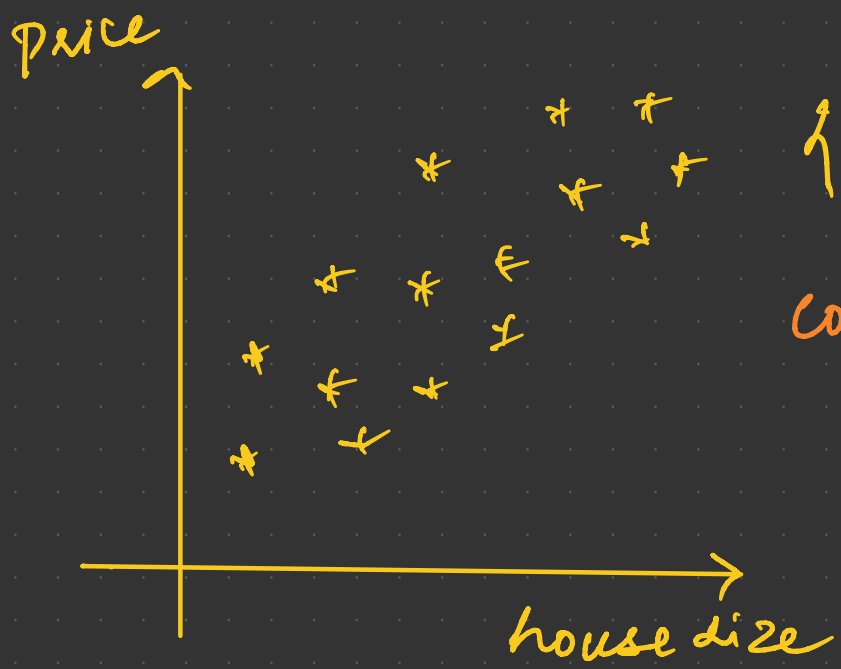
o/p features.

house size

Fountain size

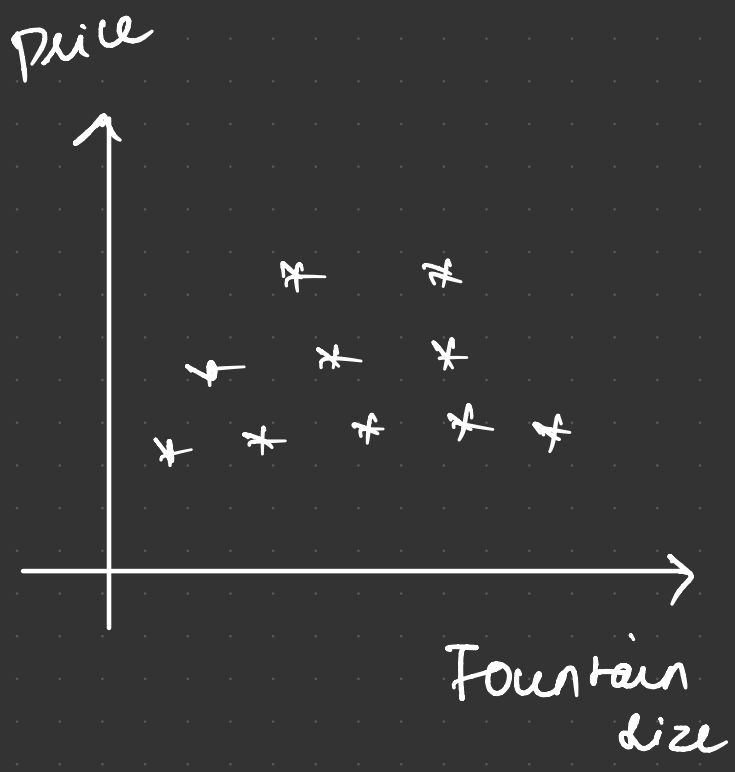
price

this mayn't play a role in prediction



Linear relationship

Covariance { +ve, -ve }



If you look at this plot eventhough the fountain size increases, the price of the house is stagent at one region In this case we can consider this feature is not essential is prediction

We can drop the feature that are not relavent, this process is called as feature selection.

Feature Extraction

room size	no of rooms	price

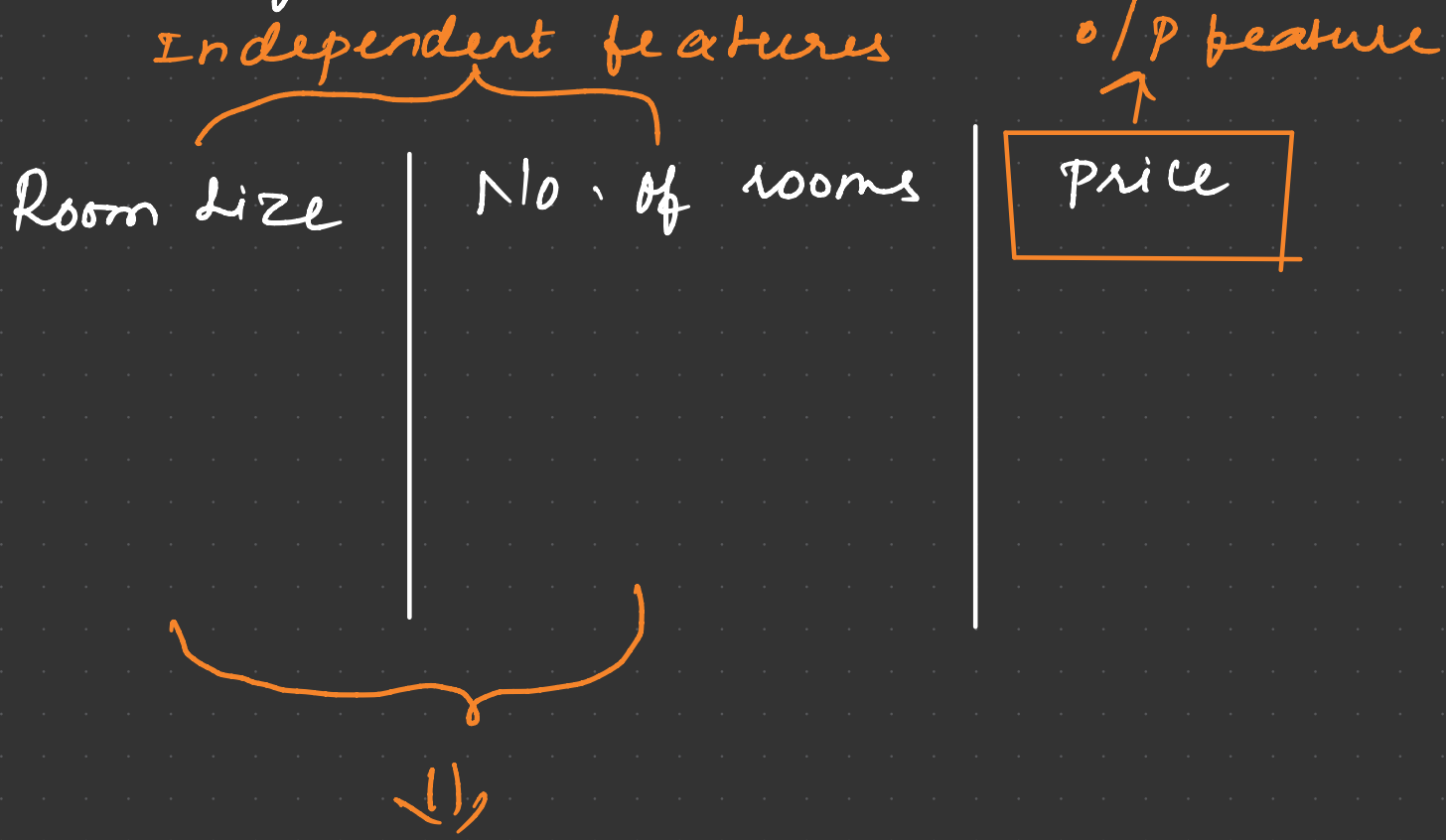
2 features ⇒ 1 feature

⇓

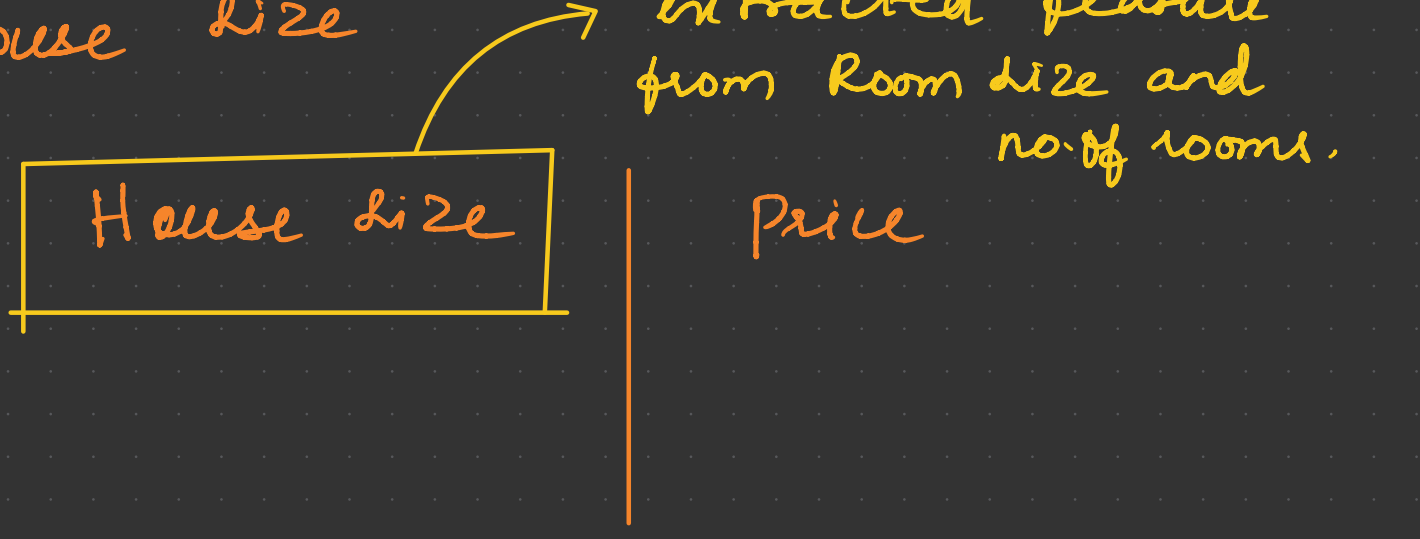
Dimensionality reduction

In the above dataset i cannot simply drop one feature and consider one feature for predicting the price of the house. If you look at the data set are super important for predicting the price of the house.

In this case we perform something called feature extraction.



we apply some transformation to extract the new feature called house size



with this new feature that is extracted from two independent features, we can predict the price of the house.