Data preprocessing

Necessary to make sure the models work accurately

Will be using independent variables to predict dependent variables

**Data preprocessing**

Three essential libraries we are going to use

Pandas

Matplotlib

Numpy

You need to break data into vectors

X vector= independent variable/s

Y vector = dependent variable/s

**Class**

The model of what we want to build .. ex construction plan to build a house

**object**

The instance of a class

Example would be a house. The object(house) was build by following instructions of the class

**Method**

Tool we use on the object to complete a specific action

**Dealing with missing data**

It is dangerous to remove rows where the data is simply missing.

What you can do is simply take the mean of the columns and put it in for missing data

**Catagorical data**

You will need to convert categorical data to numerical data. ML is mathematical based and you will have issues if not all the data is mathematical

**Feature scaling**

Some variables will be completely out of scale vs other variables.

For example, age ranges from say 20-80.

Salary varies from 30k -100k.

These variables are not on the same scale and will cause problems in our model

Many of the models are based on the Euclidean distance. This is the distance between two data points.

This is the square root of the sum of squared coordinates

In our example above, the Euclidean distance would be dominated by salary since the distance between the points is greater than age.

If you don’t put them on scale with one another, the model will be dominated by salary. Essentially rendering age irrelevant when it comes to influencing the Y variable.

Two ways to scale the data,

**Normalization and standardization**

Decision trees are not based on eucledian distance, but it is best practice to always scale your data as it helps the algorithm run faster