Supervised Learning

DAY2 AGENDA

- Linear Regression with One variable
- Evaluation Metrics in Regression Models
- Train/Test splitting of data
- Cross Validation
- Linear Regression with Multiple Variables

TYPE OF DATA IN THE REAL-WORLD

- Structured and unstructured, ordered and unordered, Indexed nd unindexed data
- Numerical, Decimal, strings

Ordered/Structured

Excel, csv, DB, RDBMS, Tabular

unordered/unstructured

PPT, PDFs, Images, Videos, Audio

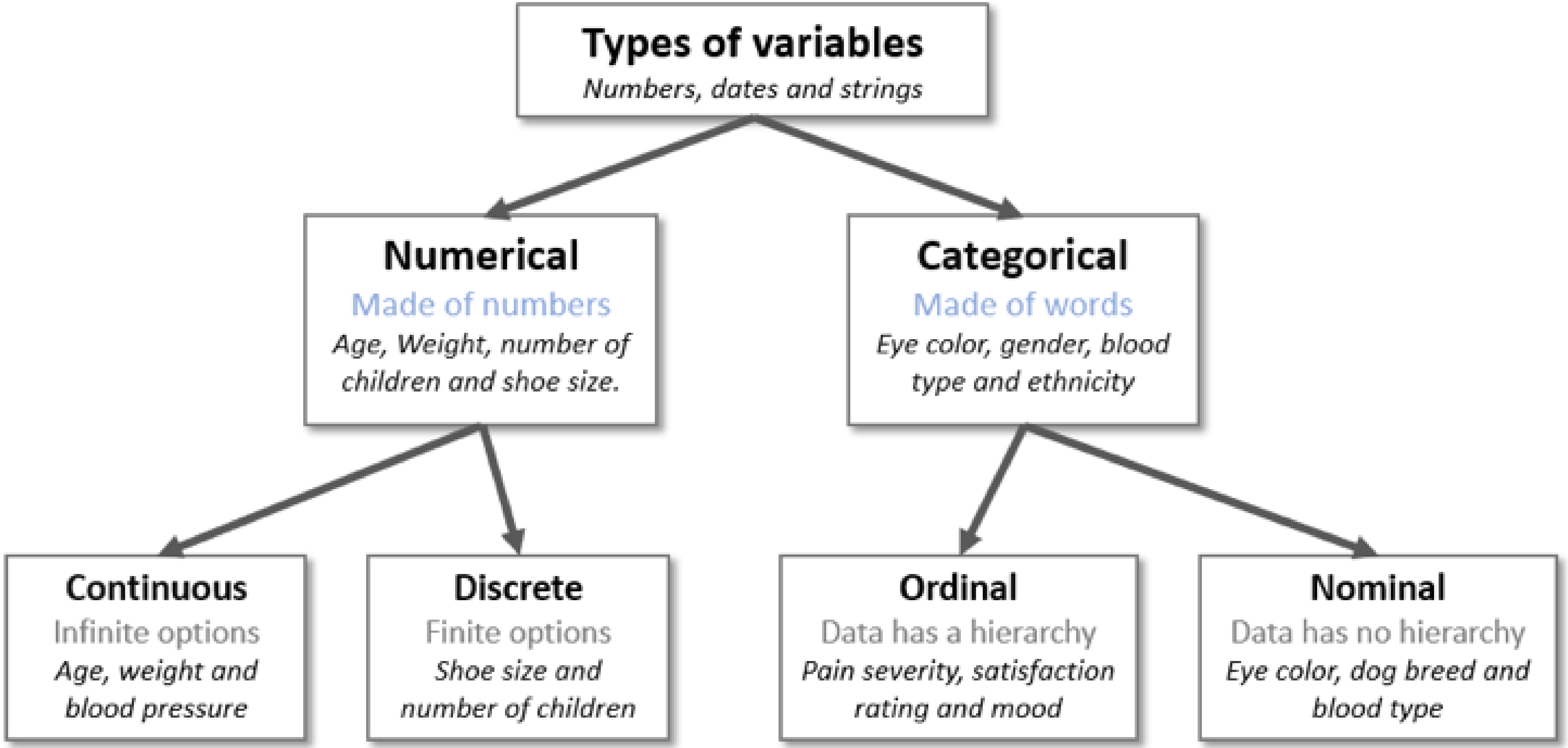
Semi-structured

- XML, JSON, HTML

SUPERVISED LEARNING TYPES

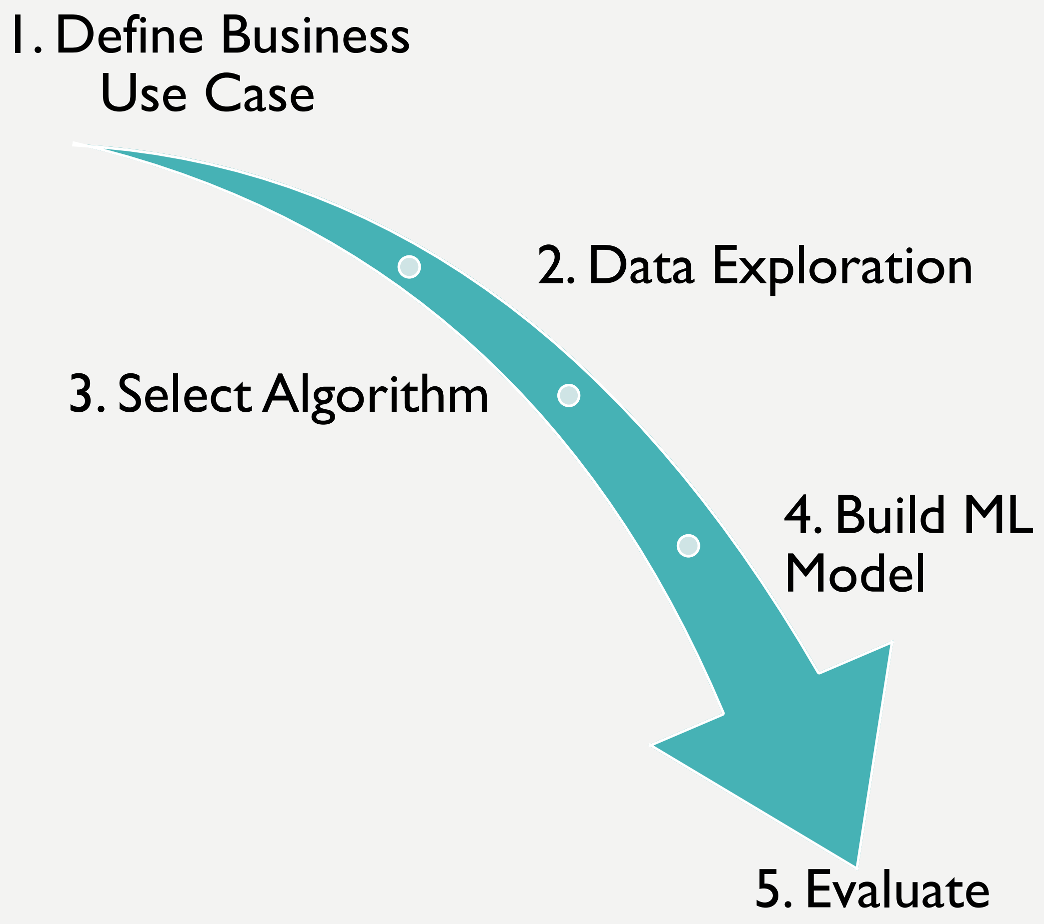
- Regression
- Classification

TYPES OF STATISTICAL DATA



https://en.wikipedia.org/wiki/Statistical_data_type

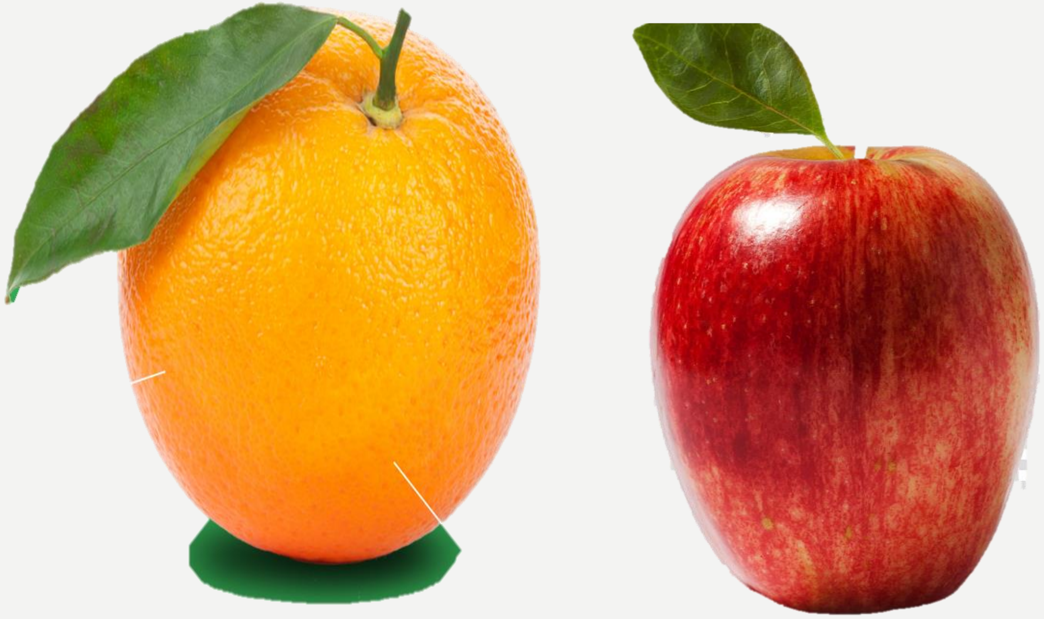
ML MODEL DEVELOPMENT LIFE CYCLE



FEATURES / ATTRIBUTES

- **Features (aka attributes) are used to train an ML system.**

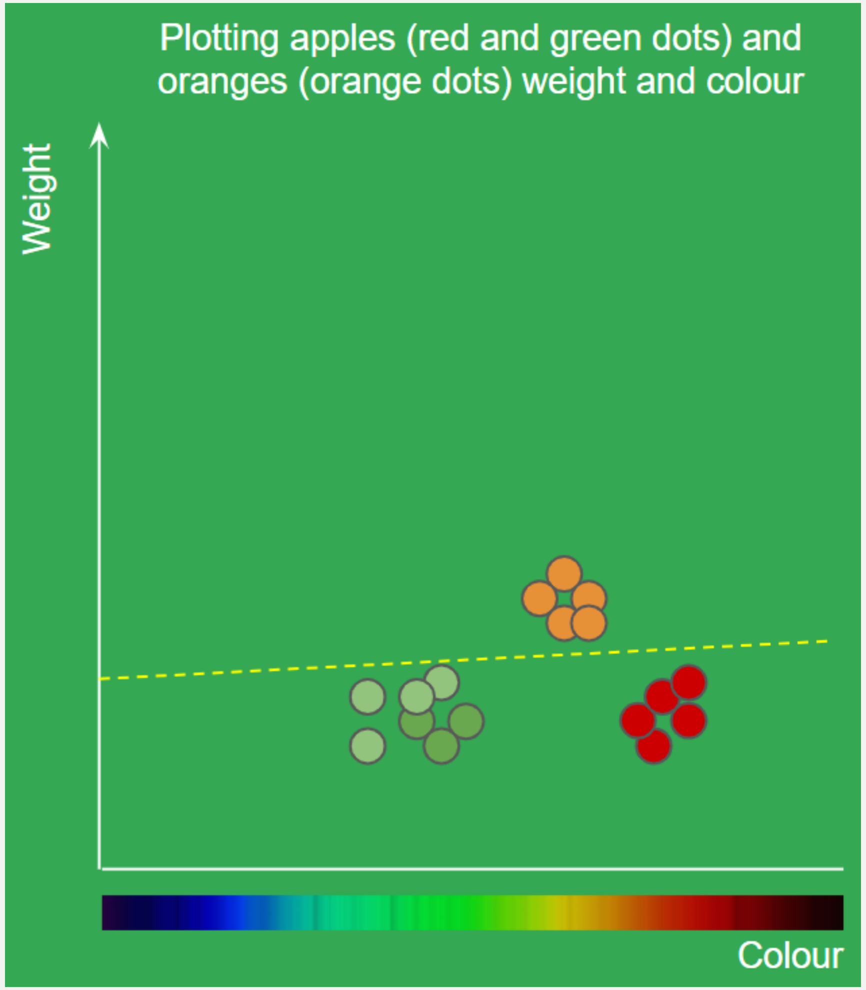
They are the properties of the things you are trying to learn about.



FEATURES / ATTRIBUTES

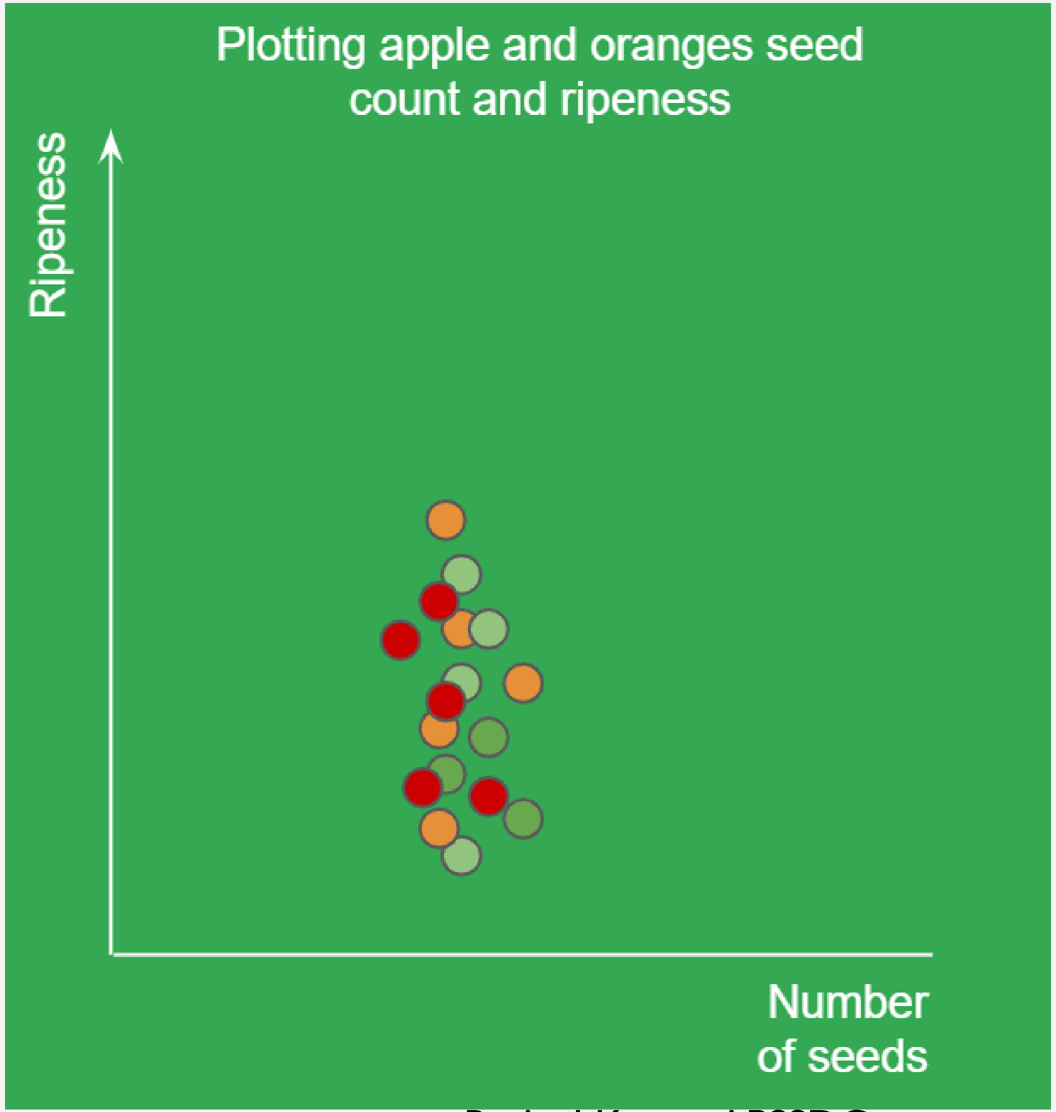
Taking fruit as an example. Features of a fruit might be weight and color. 2 features, would mean there are 2 dimensions. A 2D system may be plotted on a graph if features are represented in a numerical way.

In the plot on the right, the ML system can learn to split the data up with a line to separate apples from oranges. This **can now be used to make future classifications when we plot new points the system has not seen (anything above is orange, below is apple)**

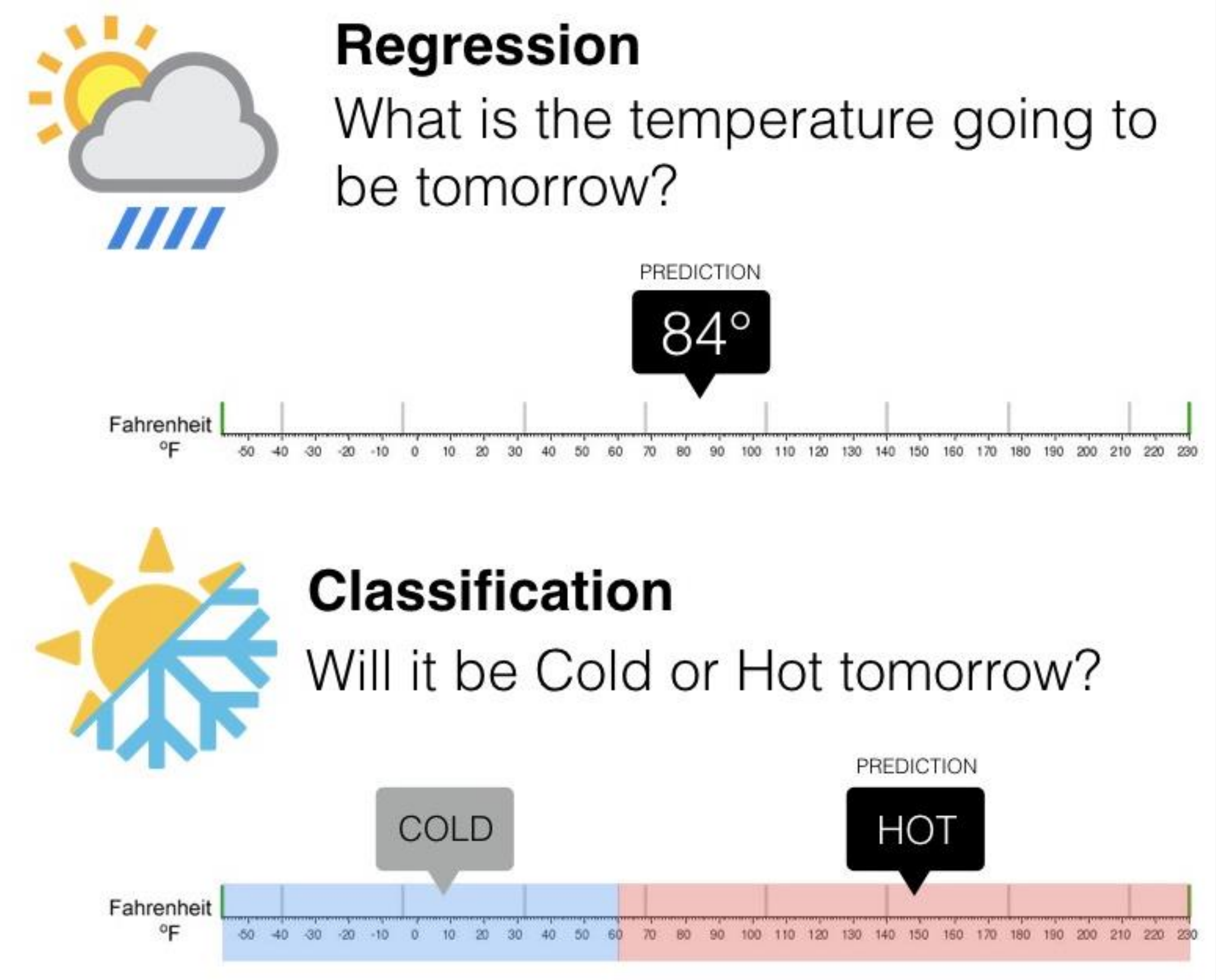


FEATURES / ATTRIBUTES

- **Choosing useful features can have a big impact on the quality of the ML system.** Some features may not be useful enough to separate the data points.
- In this example we take bad features of fruits(ripeness and seed count) that do not allow us to learn any distinguishing factors for the fruit.



REGRESSION VS CLASSIFICATION



Linear Regression in Machine Learning

By Anil Kumar
APSSDC

What is Regression?

- Function: a mathematical relationship enabling us to predict what values of one variable (Y) correspond to given values of another variable (X).
- Y: is referred to as the ***dependent variable***, **the *response variable* or the *predicted variable*.**
- X: is referred to as **the *independent variable***, the *explanatory variable* or the ***predictor variable***.

Thus Regression analysis is a form of predictive modelling technique which investigates the relationship between a dependent and independent variable.

Example

- Finding relationship between the features and target
- **Humidity**, **moisture**, **light**, → Temperature

Regression

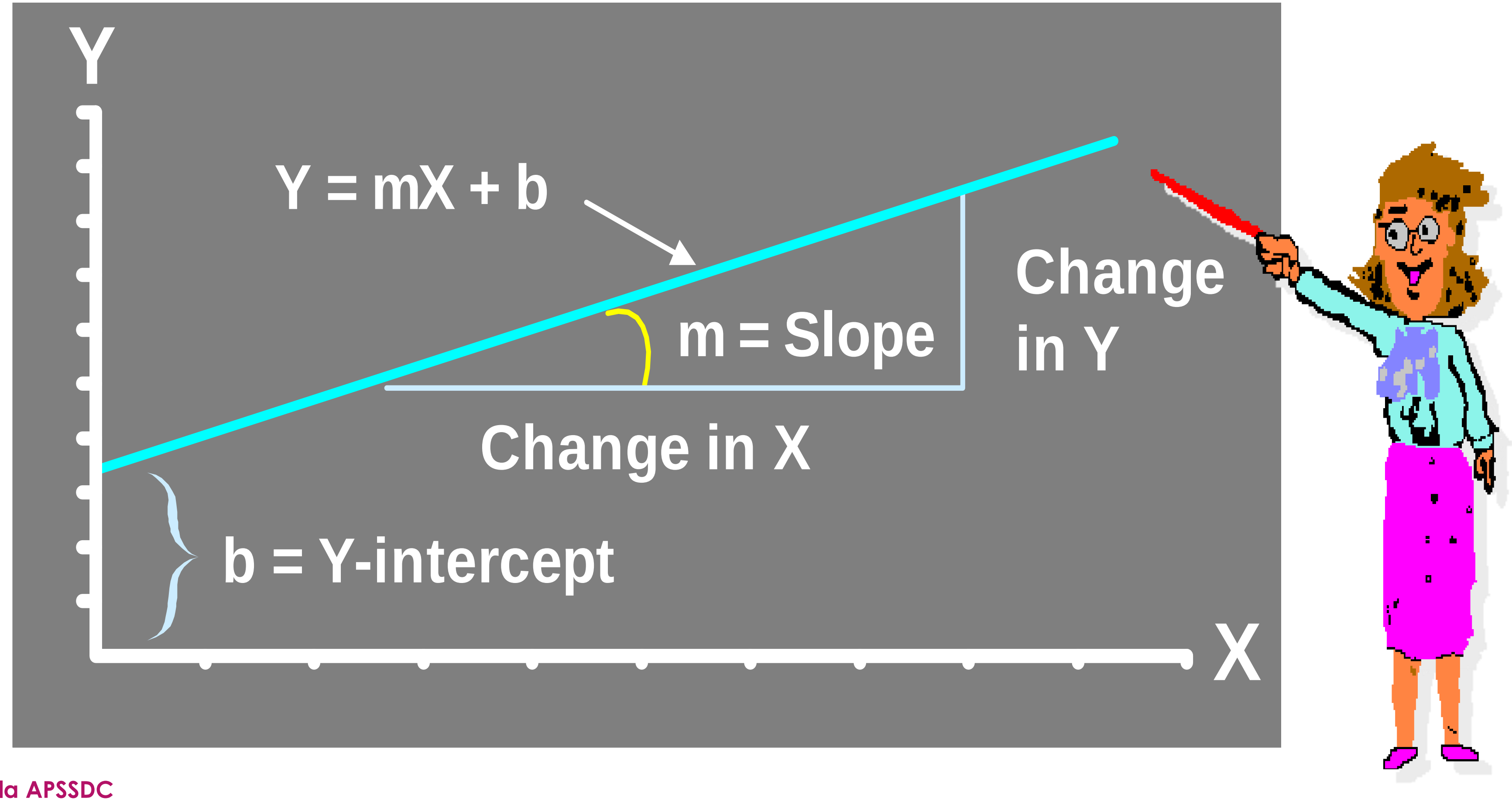
- **Linear Regression**
  - **Linear Regression with one variable**
  - **Linear Regression with multiple variable**
- Non-Linear Regression/Polynomial Regression
  - Non-Linear Regression with one variable
  - Non-Linear Regression with multiple variables
- SGD
- Ridge
- Lasso
- Elastic Net

Contd.

A typical Linear Regression model can be represented in the form :

y = b1x + b0 where b1 is slope and b0 is the intercept.

Linear Equations



Linear Regression Model

Relationship Between Variables Is a Linear Function

Y-Intercept

Slope

Random Error

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

Dependent (Response) Variable

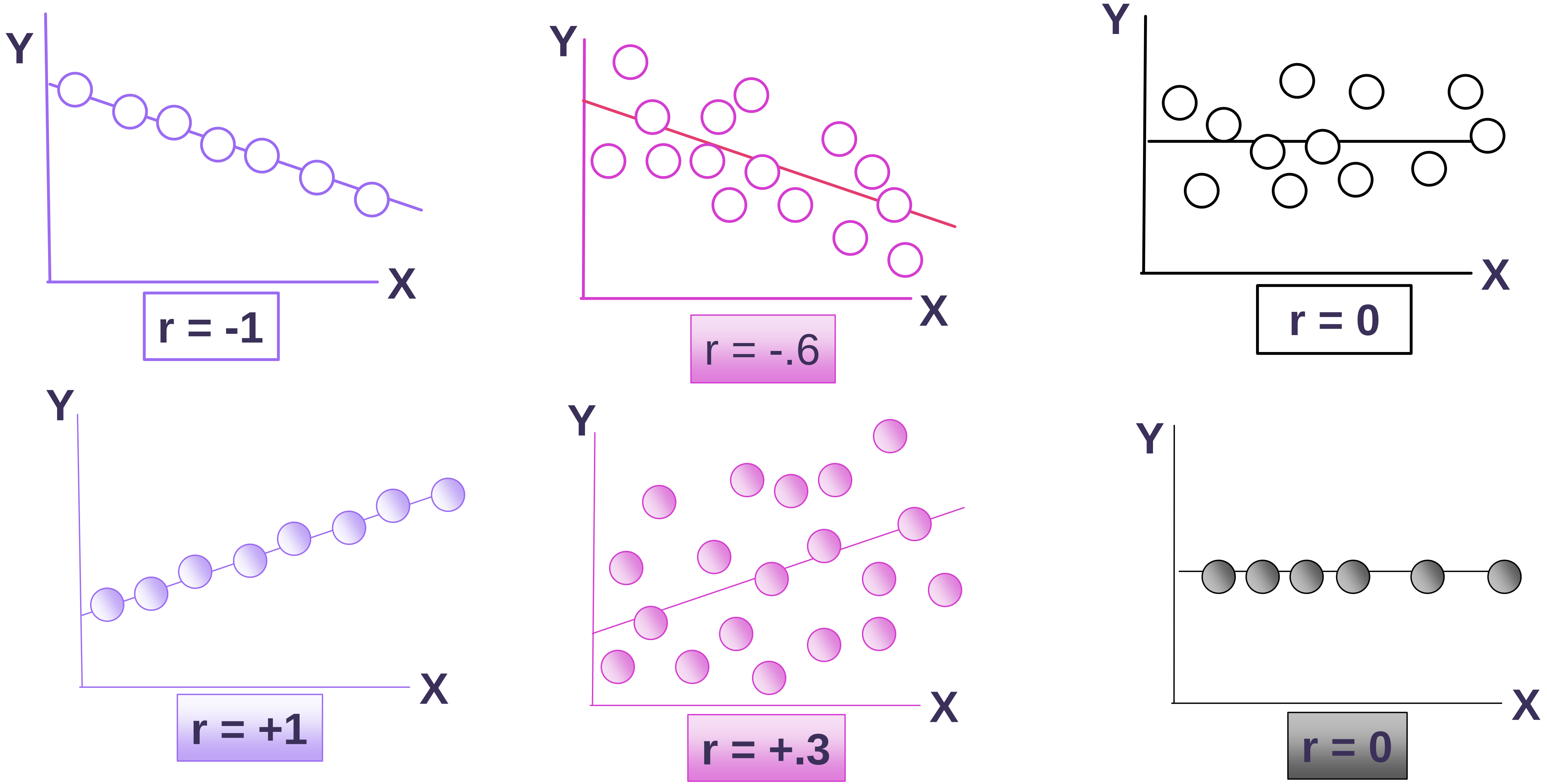
Independent (Explanatory) Variable

- $\beta_1 > 0 \Rightarrow$ Positive Association
- $\beta_1 < 0 \Rightarrow$ Negative Association
- $\beta_1 = 0 \Rightarrow$ No Association

Correlation

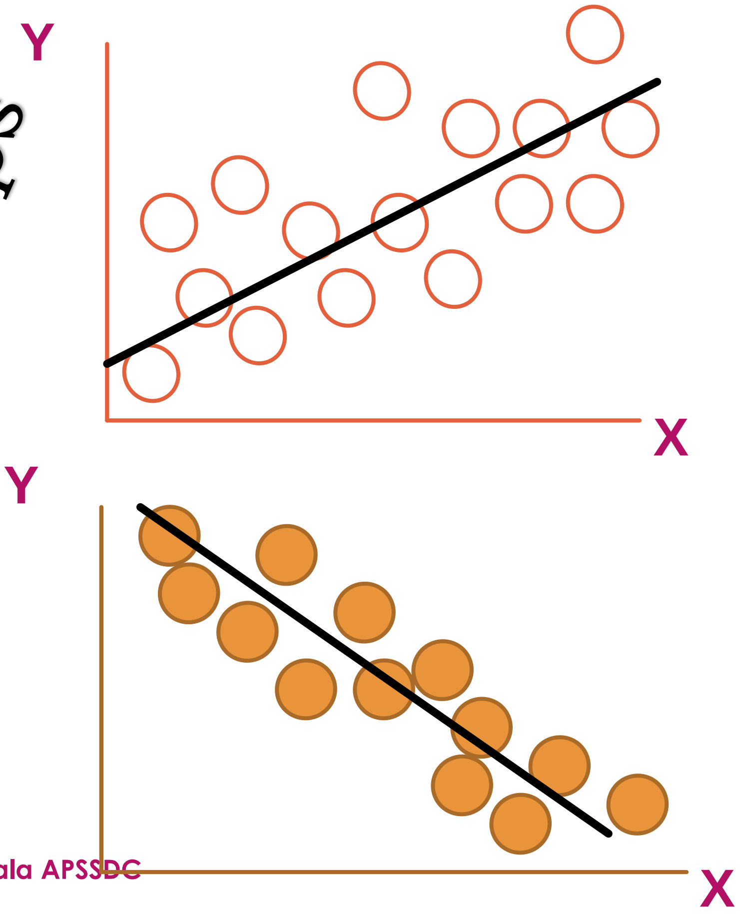
- Measures the relative strength of the *linear* relationship between two variables Unit-less
- Ranges between –1 and 1
- The closer to –1, the stronger the negative linear relationship
- The closer to 1, the stronger the positive linear relationship
- The closer to 0, the weaker any positive linear relationship

Scatter Plots of Data with Various Correlation Coefficients

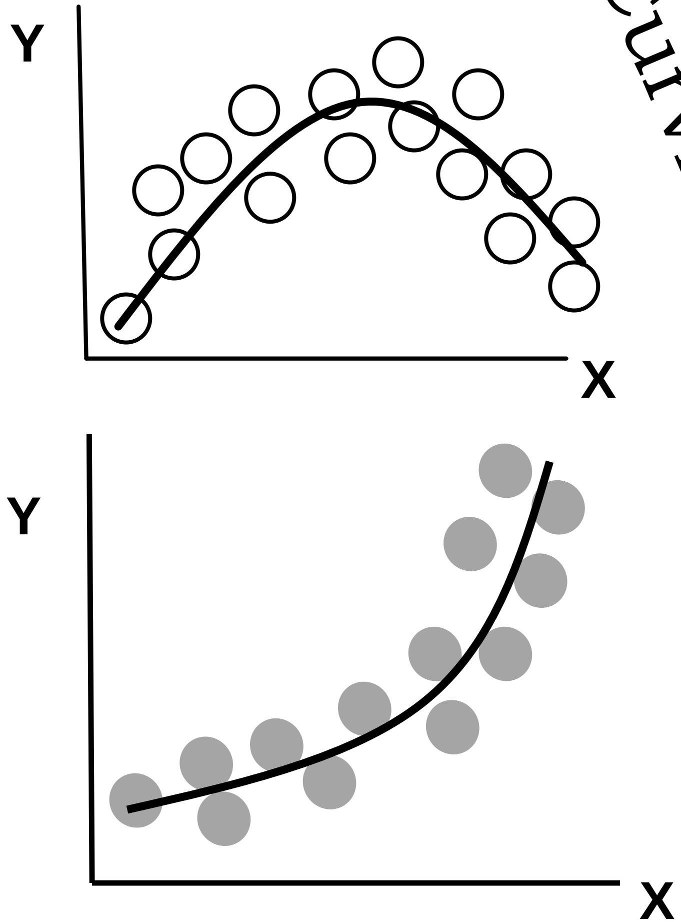


Linear Correlation

Linear relationships

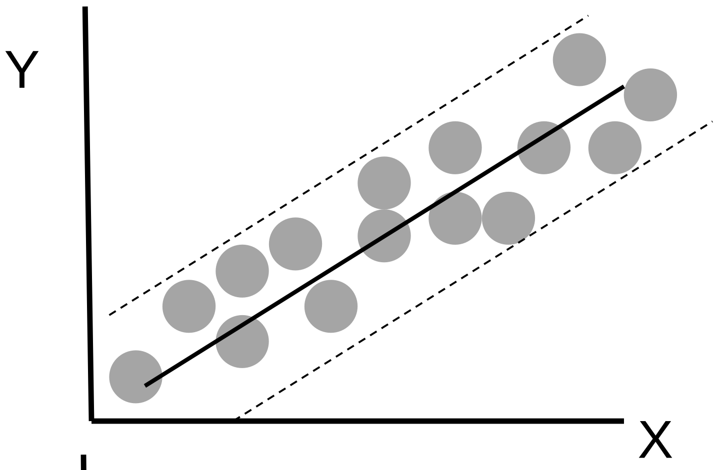


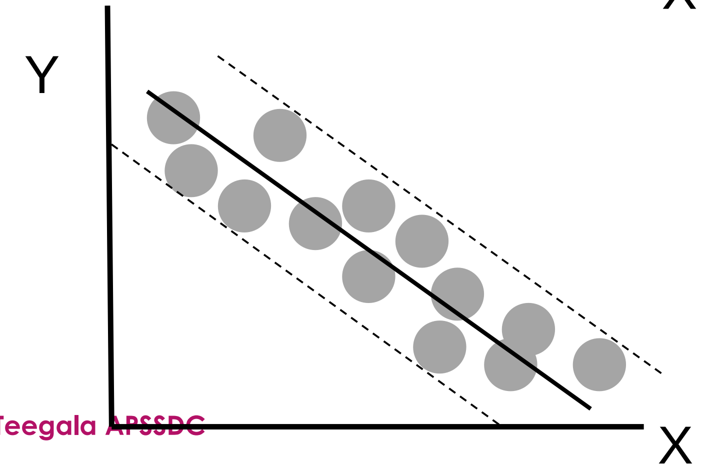
Curvilinear relationships



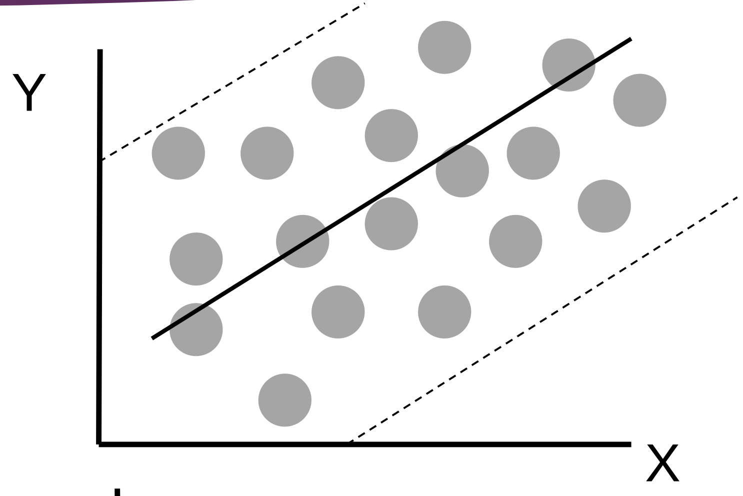
Linear Correlation

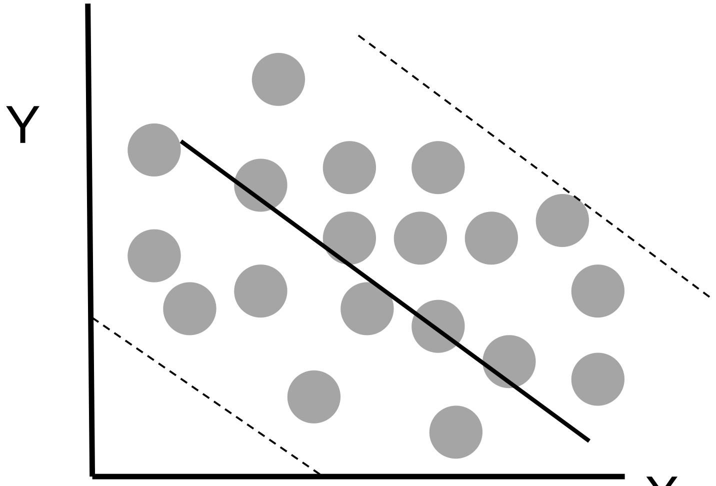
Strong relationships





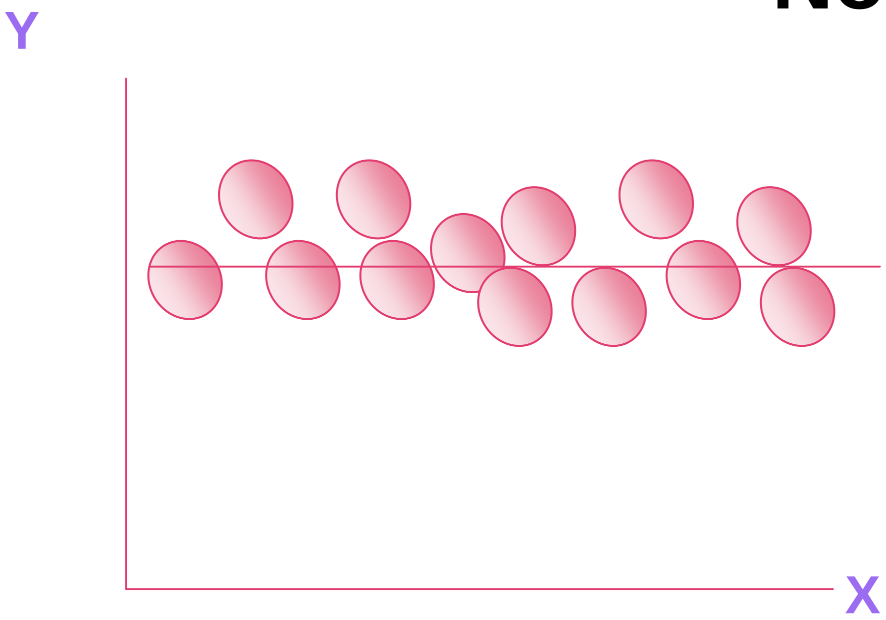
Weak relationships

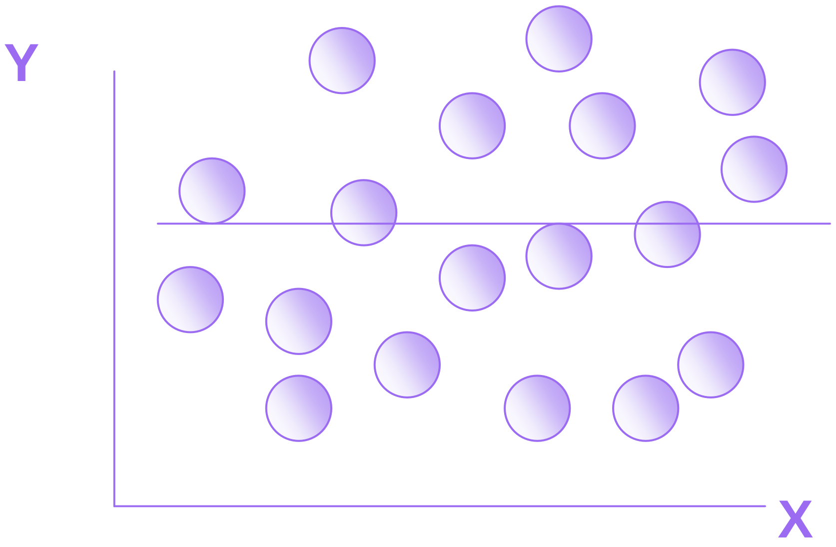




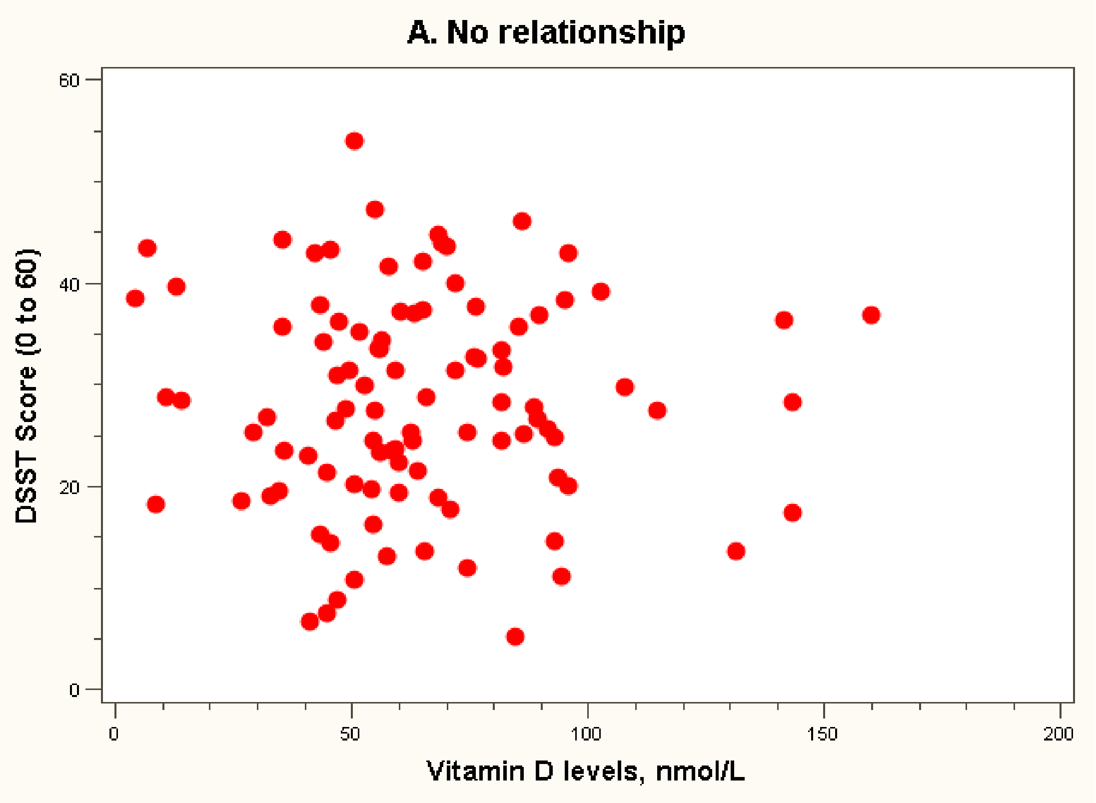
Linear Correlation

No relationship

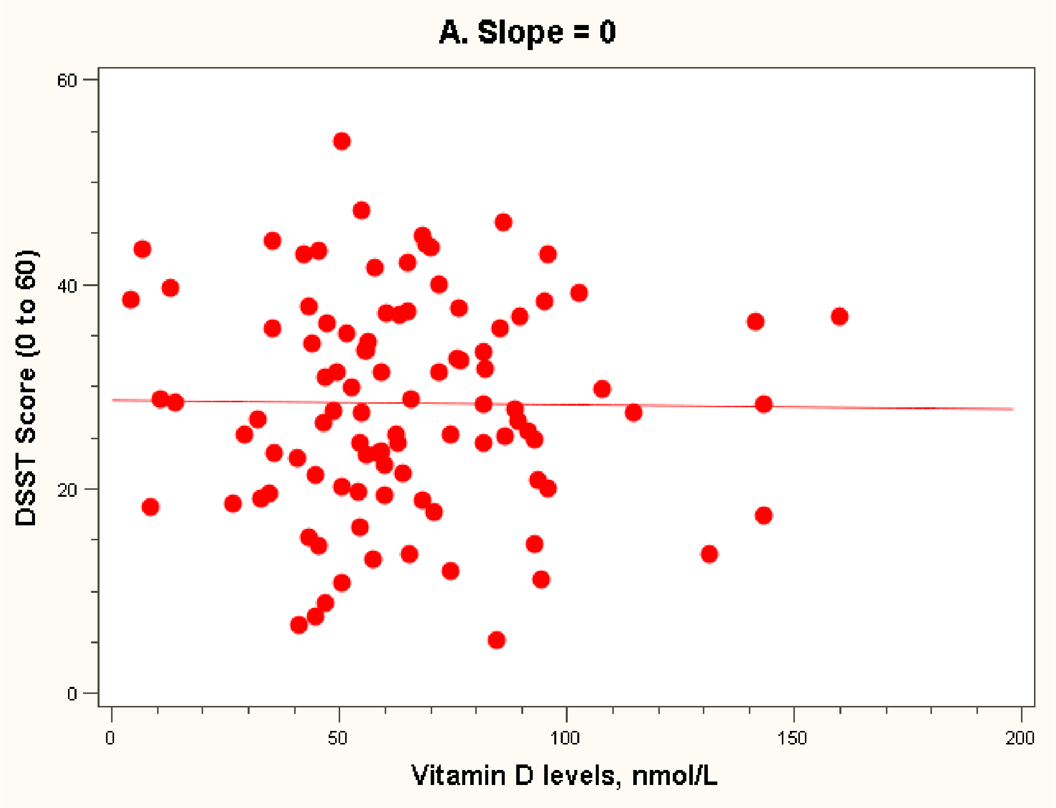




Dataset 1: No Relationship



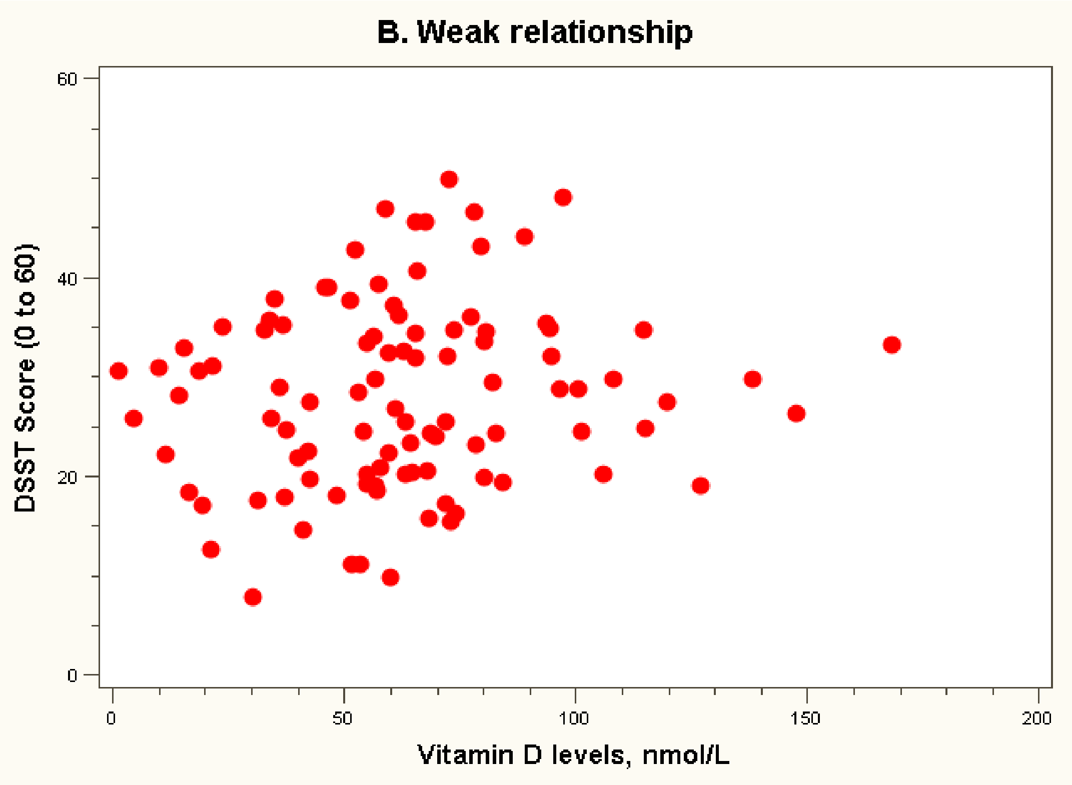
The "Best fit" line



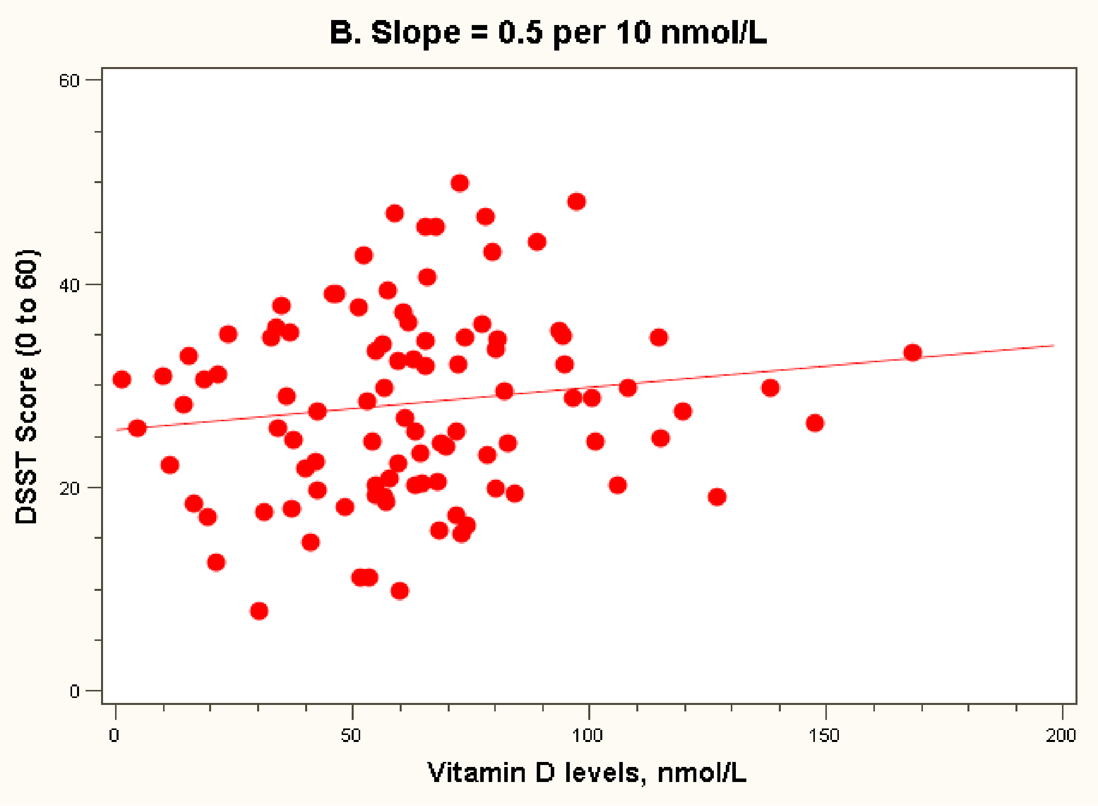
Regression equation:

E(Yi) = 28 + 0*vit Di (in 10 nmol/L)

Dataset 2: weak relationship



The “Best fit” line

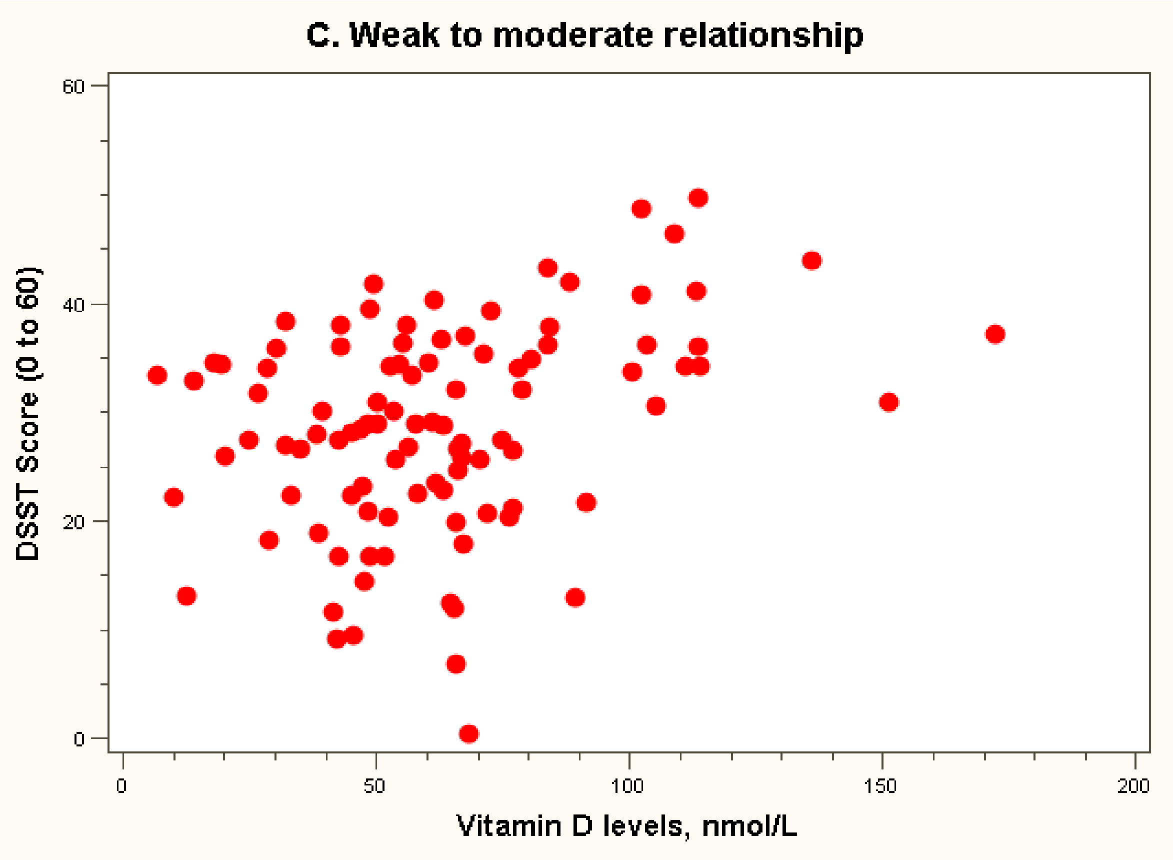


Note how the line is a little deceptive; it draws your eye, making the relationship appear stronger than it really is!

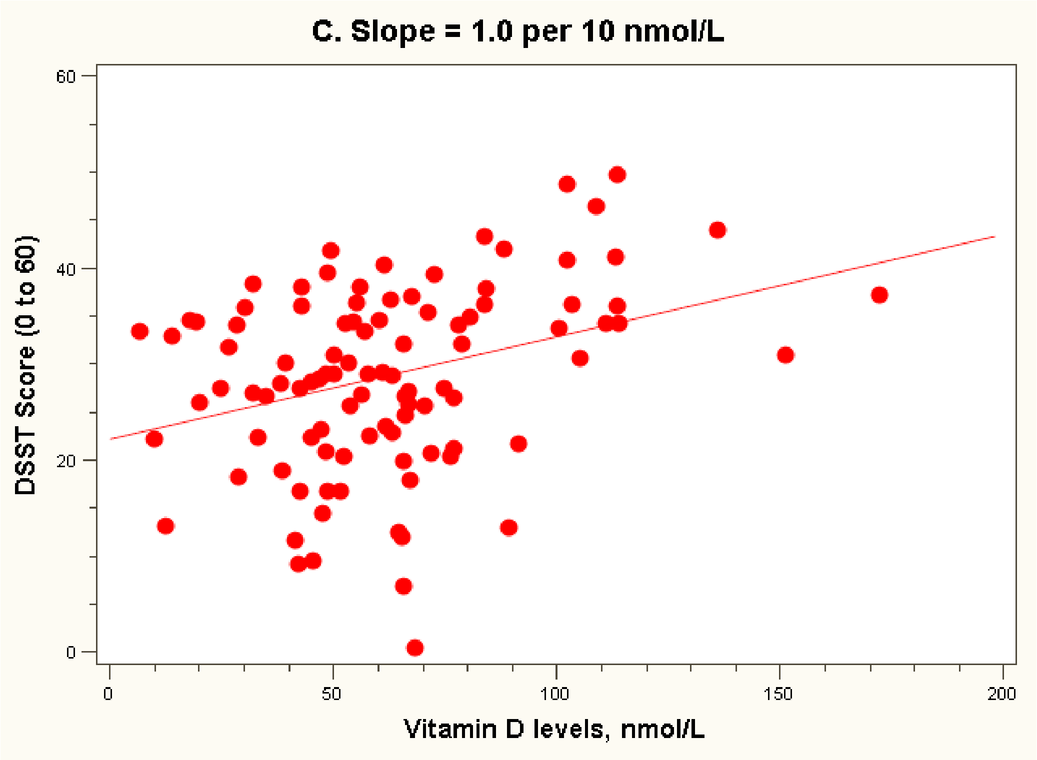
Regression equation:

E(Yi) = 26 + 0.5*vit Di (in 10 nmol/L)

Dataset 3: weak to moderate relationship



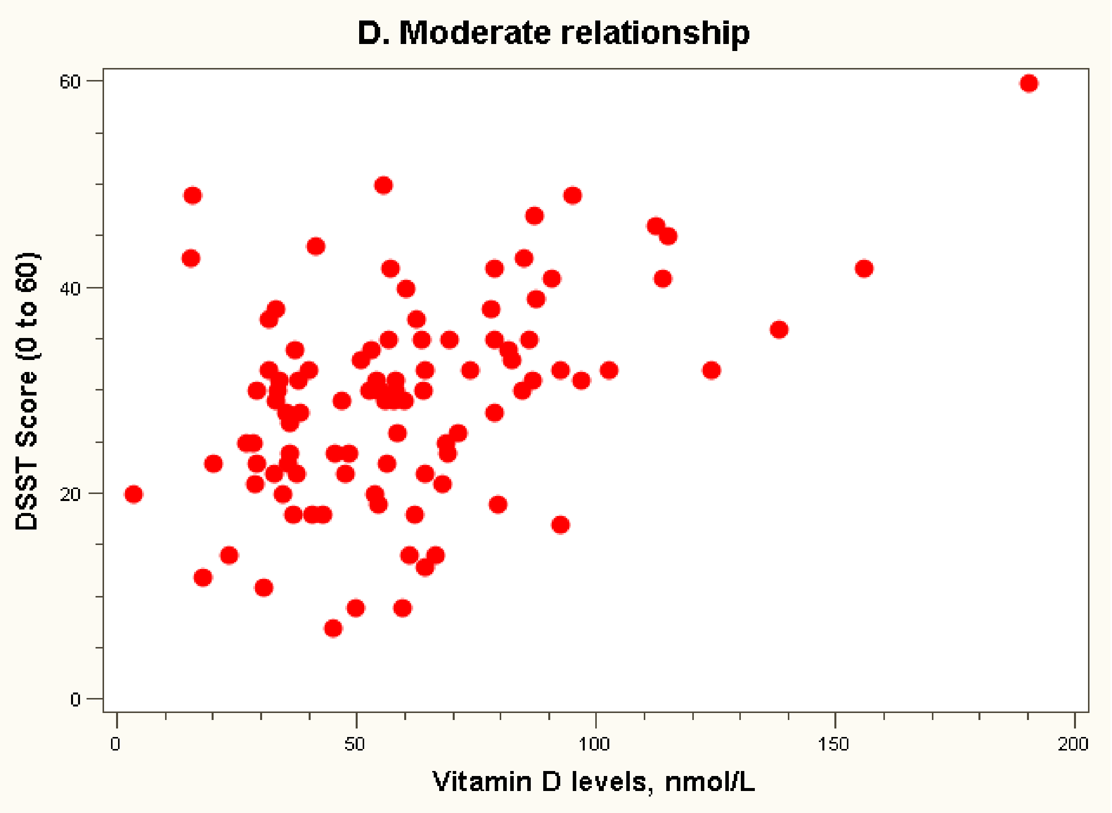
The “Best fit” line



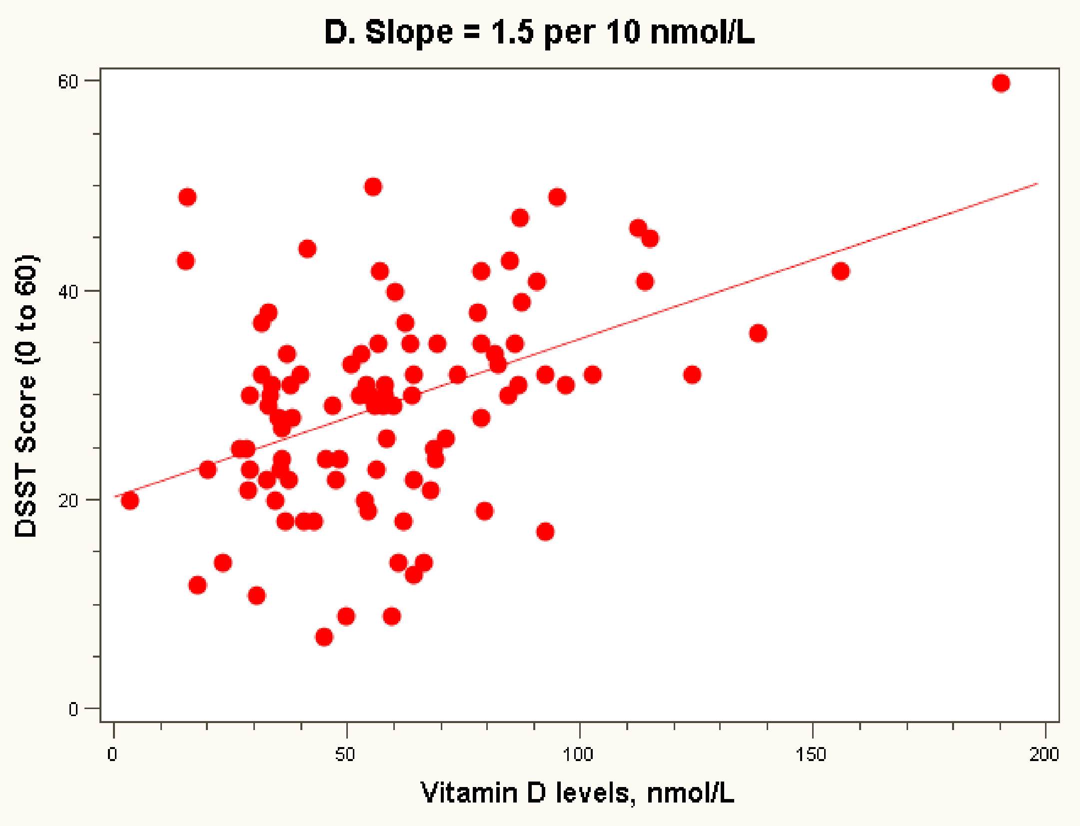
Regression equation:

$E(Y_i) = 22 + 1.0 \cdot \text{vit } D_i$ (in 10 nmol/L)

Dataset 4: moderate relationship



The “Best fit” line



Regression equation:

E(Yi) = 20 + 1.5*vit Di (in 10 nmol/L)

Note: all the lines go through the point (63, 28)!

Steps in Regression Analysis

- Examine the scatterplot of the data.
  I. Does the relationship look linear?
  II. Are there points in locations they shouldn't be?
  III. Do we need a transformation?
- Assuming a linear function looks appropriate, estimate the regression parameters.
  I. How do we do this? (Method of Least Squares)
- If there is a significant linear relationship, estimate the response, Y, for the given values of X, and compute the residuals

Regression Analysis

- Thus we have the regression formula as :
  Y = MX + C + error(e).

Initially we calculate the value for slope and predict the values of Y for any given X values we have.

$$\text{Slope(M)} = \sum_{i=0}^{len(X)} \frac{(X_i - X_{mean}) * (Y_i - Y_{mean})}{(X_i - X_{mean})^2}$$

Thus we calculate the C value and find out the “Line of Regression”.

Regression Analysis

- Next our job is to reduce the distance between the actual value and the predicted value or in other words reduce the error between the actual and predicted value. Thus the line with least error will be the **"Best Fit Line"**.
- In order to check it out we calculate the "Coefficient of Determination".

$$Mean\ Squared\ value\ (R^2) = \sum_{i=0}^{len(X)} \frac{(Y_{pred} - Y_{mean})^2}{(Y - Y_{mean})^2}$$

- Thus our ultimate aim is to reduce the error i.e. distance between the actual and predicted values.

Evaluation Metrics

1. Mean Absolute Error

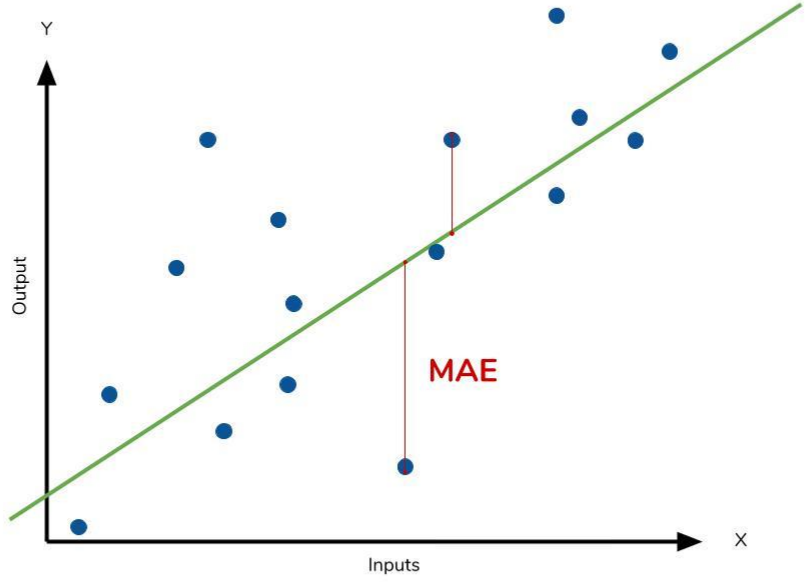
Divide by the total number of data points

Predicted output value

Actual output value

$$MAE = \boxed{\frac{1}{n}} \sum \underbrace{\left| \boxed{y} - \boxed{\hat{y}} \right|}_{\text{The absolute value of the residual}}$$

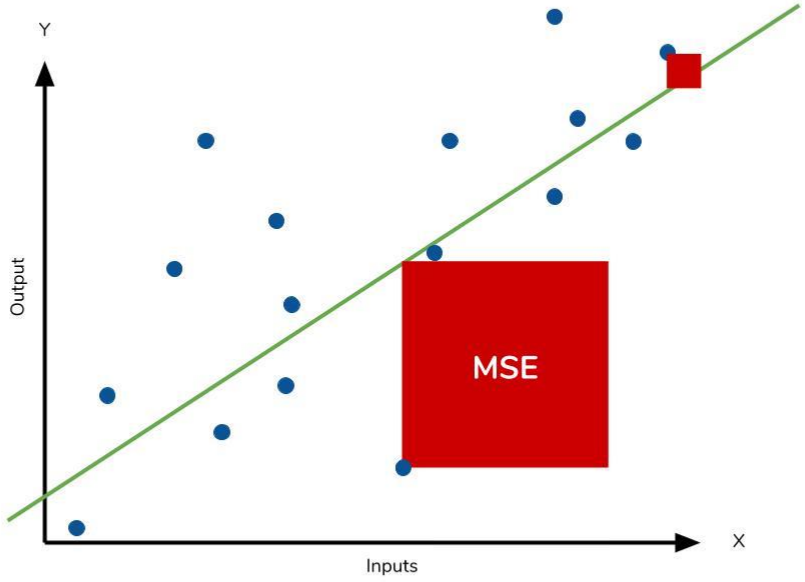
Sum of



Contd..

2. Mean Square Error

$$MSE = \frac{1}{n} \Sigma \underbrace{\left(y - \widehat{y} \right)^2}_{\text{The square of the difference between actual and predicted}}$$



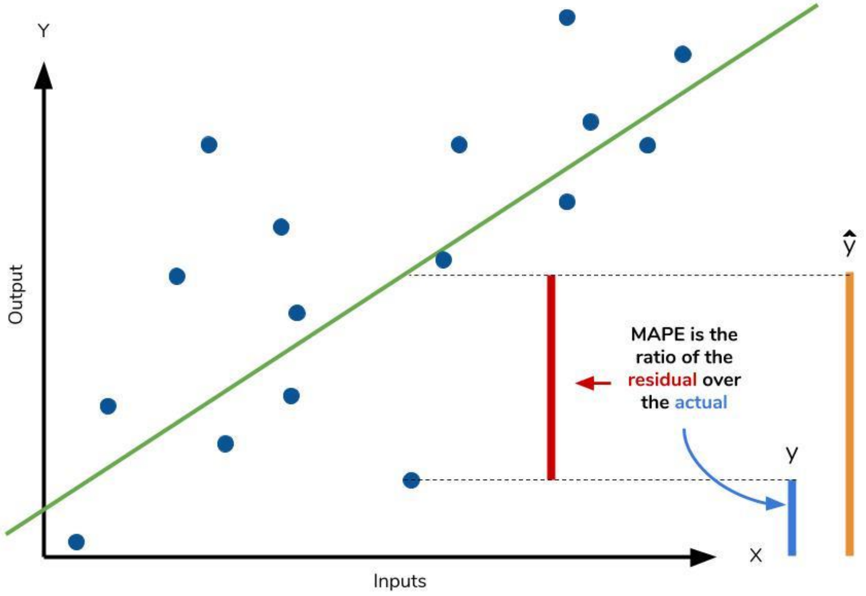
Contd..

3. Mean Absolute Percentage Error

Multiplying by 100% converts to percentage

$$MAPE = \frac{\boxed{100\%}}{n}\sum\left|\frac{\overbrace{y-\hat{y}}^{\text{The residual}}}{\boxed{y}}\right|$$

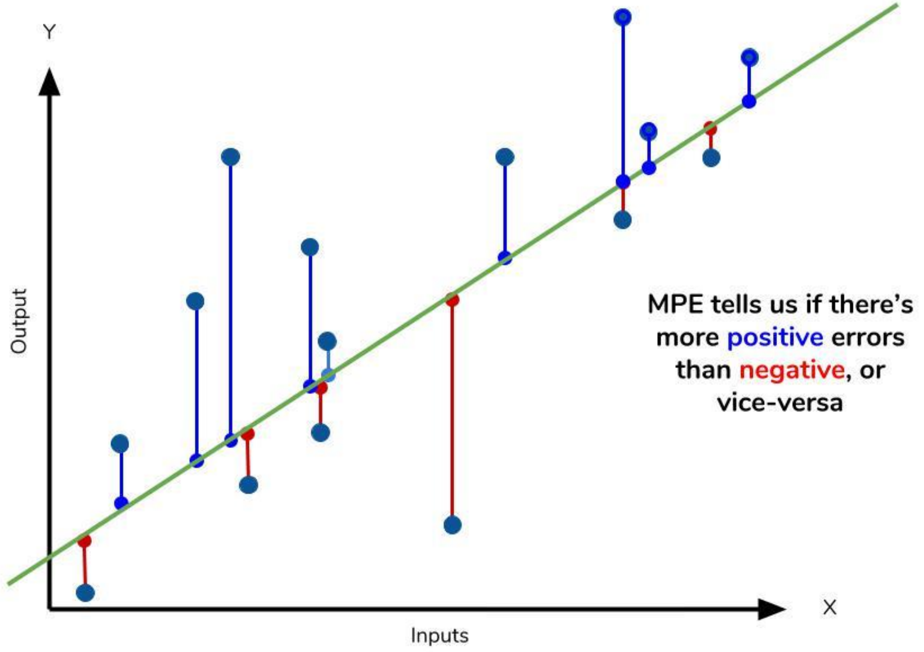
Each residual is scaled against the actual value



Contd..

3. Mean Percentage Error

$$MPE = \frac{100\%}{n} \sum \left(\frac{y - \hat{y}}{y} \right)$$



Conclusion

Acroynm	Full Name	Residual Operation?	Robust To Outliers?
MAE	Mean Absolute Error	Absolute Value	Yes
MSE	Mean Squared Error	Square	No
RMSE	Root Mean Squared Error	Square	No
MAPE	Mean Absolute Percentage Error	Absolute Value	Yes
MPE	Mean Percentage Error	N/A	Yes

https://www.dataquest.io/blog/understanding-regression-error-metrics/