WHERE ARE WE NOW ?

Module 1: Explore Data using Python 6 Hrs
Understanding Characteristics of the dataset – Size, Shape, value counts – Data Types and change data types

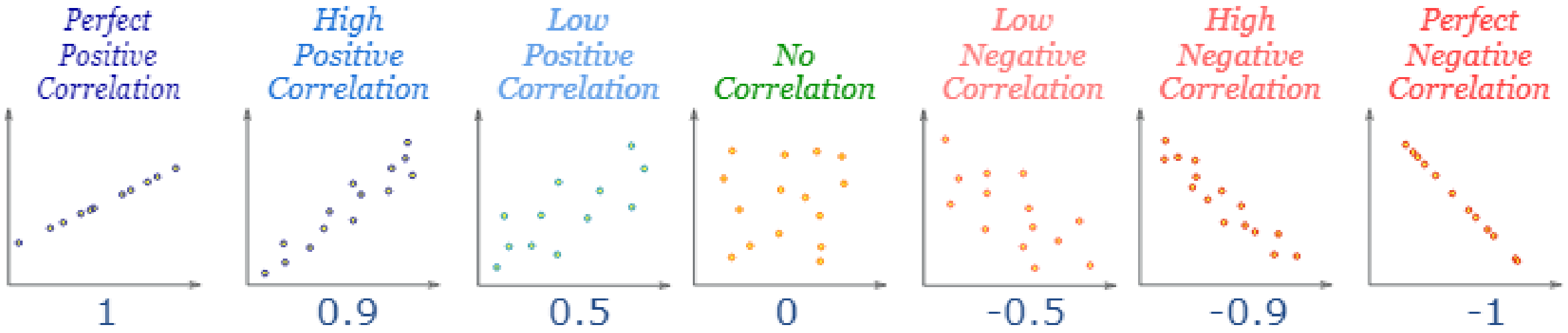
Module 2: Measures of Central Tendency 5 Hrs
Arithmetic Mean, Median, Mode – Relationship between mean, median and mode – Computation of the measures for grouped and ungrouped data

Module 3: Measures of dispersion 6 Hrs
Range, mean deviation and standard deviation – coefficient of variation and its use – Quartiles and Inter quartile range – Quintiles, Deciles and Percentiles — Skewness and Kurtosis and their uses

Module 4: Understanding Correlation 3 Hrs
Understanding correlation – calculating correlation using Pandas – Interpreting a correlation matrix

Module 5: Statistical Quality Control 3 Hrs
Nature of Control Limits – Purpose of Control Charts – Control Charts for Variables – Control Charts for Attributes

Module 6: Conducting EDA using Python 6 Hrs
Data Analysis using Python – Handling missing data – Computing metrics – Analysis & Interpretation of the data and connected visualization



PYTHON FOR DATA SCIENCE — MODULE 4

- Understanding correlation
- Calculating correlation using Pandas
- Interpreting a correlation matrix

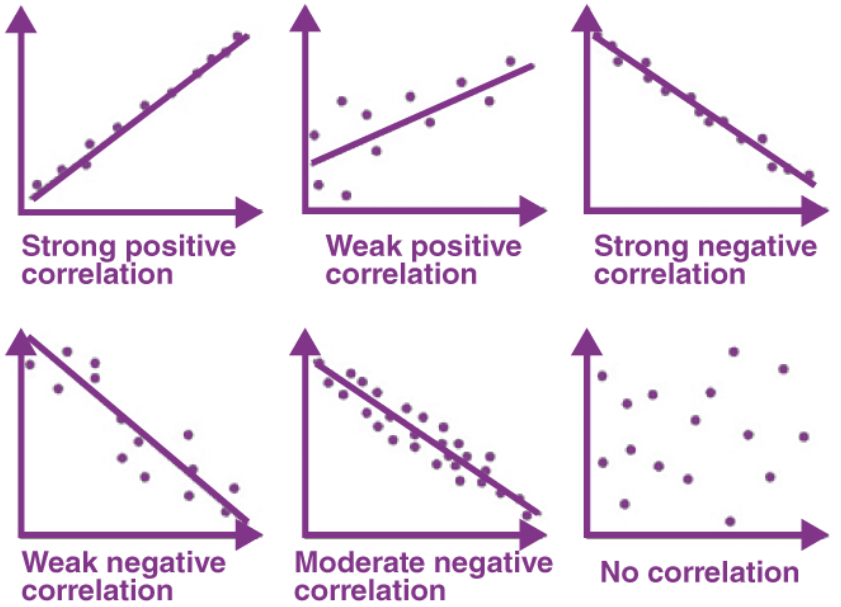
PYTHON FOR DATA SCIENCE – MODULE 4
CORRELATION

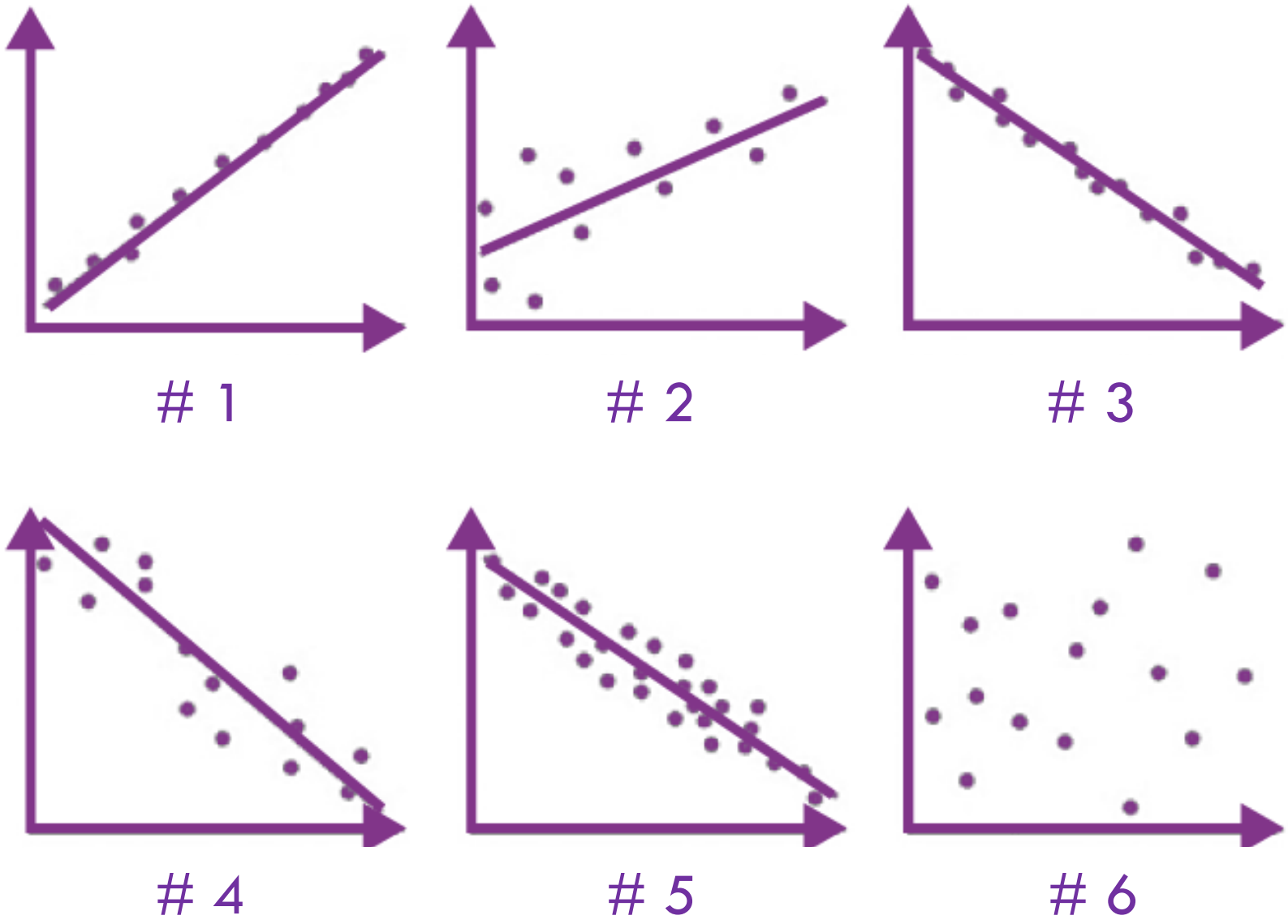
- Correlation describes the strength of an association between two variables
- It is completely symmetrical, correlation between A & B is same as correlation between B & A
- Two variables are linearly related (meaning they change together at a constant rate)
- It's a common tool for describing simple relationships without making a statement about cause and effect
- The sample correlation coefficient, r, quantifies the strength of the relationship. Correlations are also tested for statistical significance
- Correlation can't look at the presence or effect of other variables outside of the two being explored
- Importantly, correlation doesn't tell us about cause and effect
- Correlation also cannot accurately describe curvilinear relationships

PYTHON FOR DATA SCIENCE – MODULE 4
CORRELATION

$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}}$$

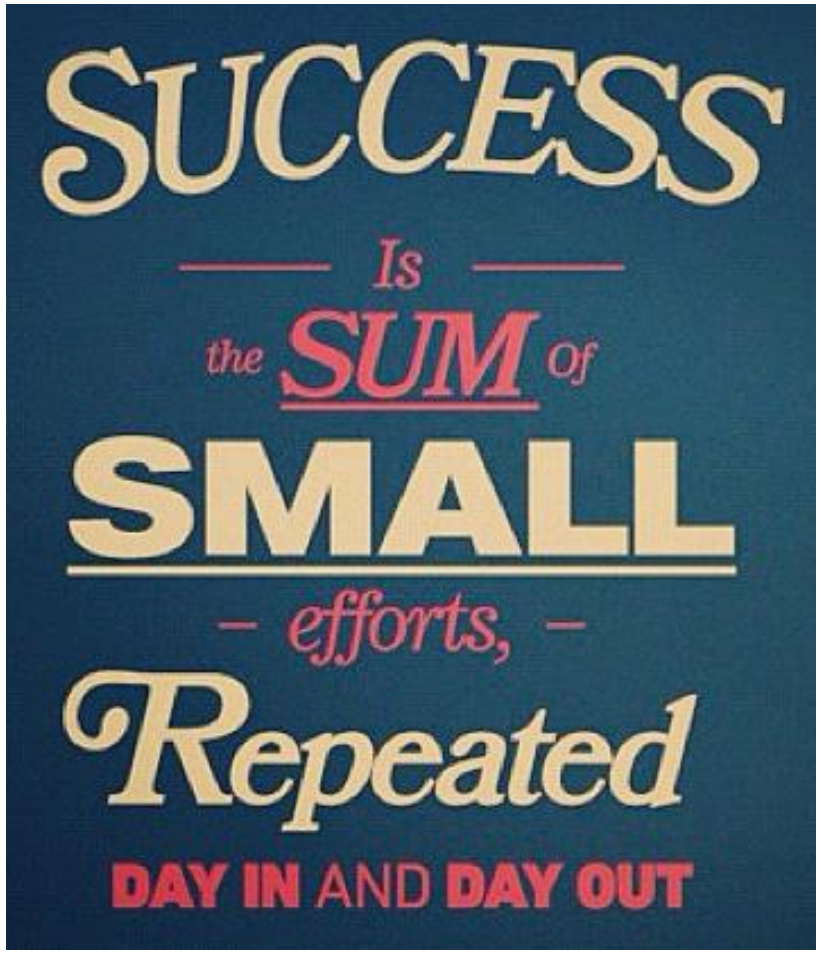
- It is a unit-free measure called the correlation coefficient which ranges from -1 to +1 and is denoted by r
  - *The closer **r is to zero**, the weaker the linear relationship*
  - ***Positive r** values indicate a positive correlation, where the values of both variables tend to increase together*
  - ***Negative r** values indicate a negative correlation, where the values of one variable tend to increase when the values of the other variable decrease*




1
2
3
4
5
6


the MORE
YOU PRACTICE
THE BETTER
YOU GET


Practice
Habit
Practice
Practice


SUCCESS
Is
the SUM of
SMALL
efforts,
Repeated
DAY IN AND DAY OUT

NEW DATA SENT – CORRELATION

Load data, perform shape, dtypes etc.,

1. Orders in UK and USA how similar or different they are, please comment.
2. What is your insight about Rohit & Steve Smith's batting style is there a Correlation ?
3. Is there a pattern in the way India bats? Compare when they bat first and chase
4. What is relationship between Revenue & Income and Revenue & # of Children. Write your observations.