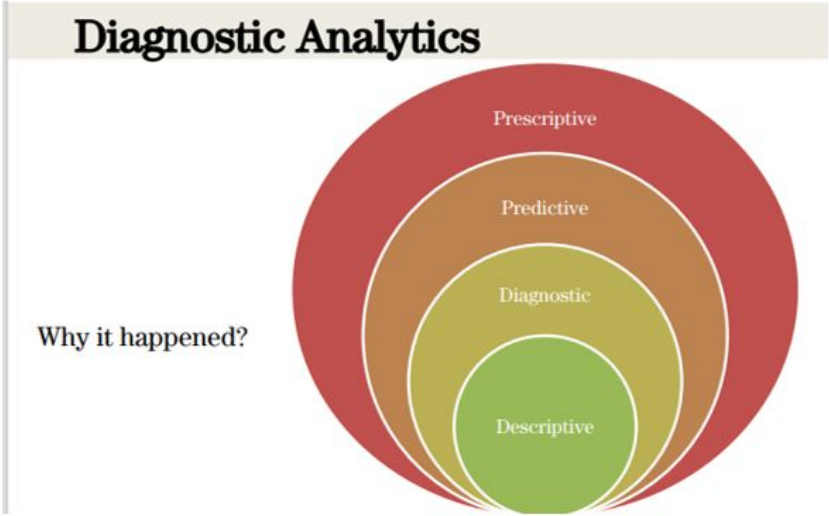
Session5 and 6

RV, Correlation, covariance and outliers

Data Analytics life cycle

Answer: Diagnostic analytics is identifying why "X" happened

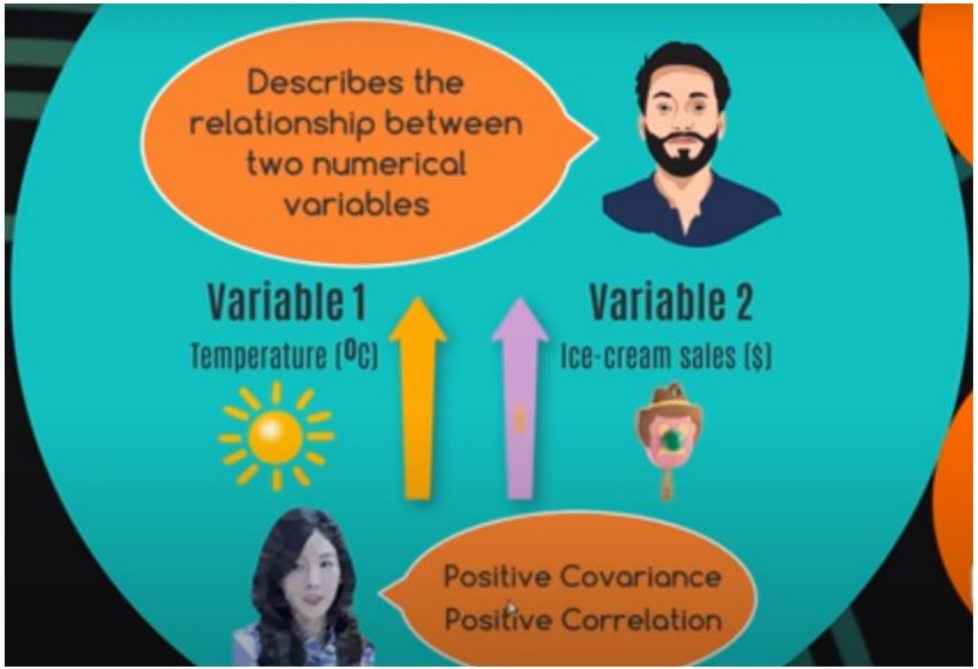
Example: For example, if we have student's attendance, mid-term marks and their final marks, and we want to know why few students scored less than 40?



Parameters of diagnostics analytics

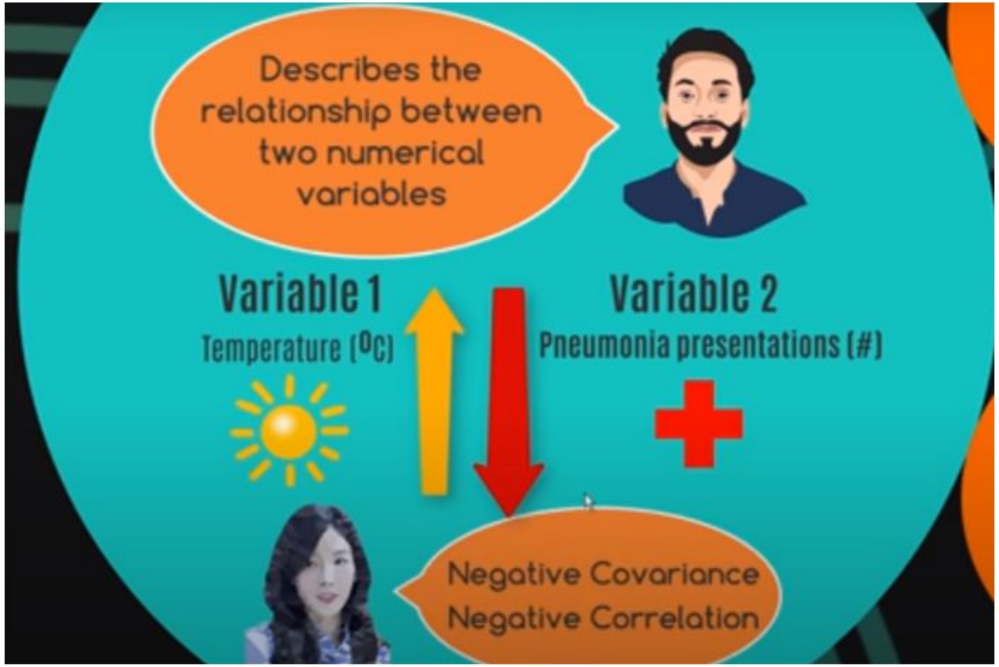
- Correlation
- Covariance
- Outliers etc

Introduction to covariance and correlation

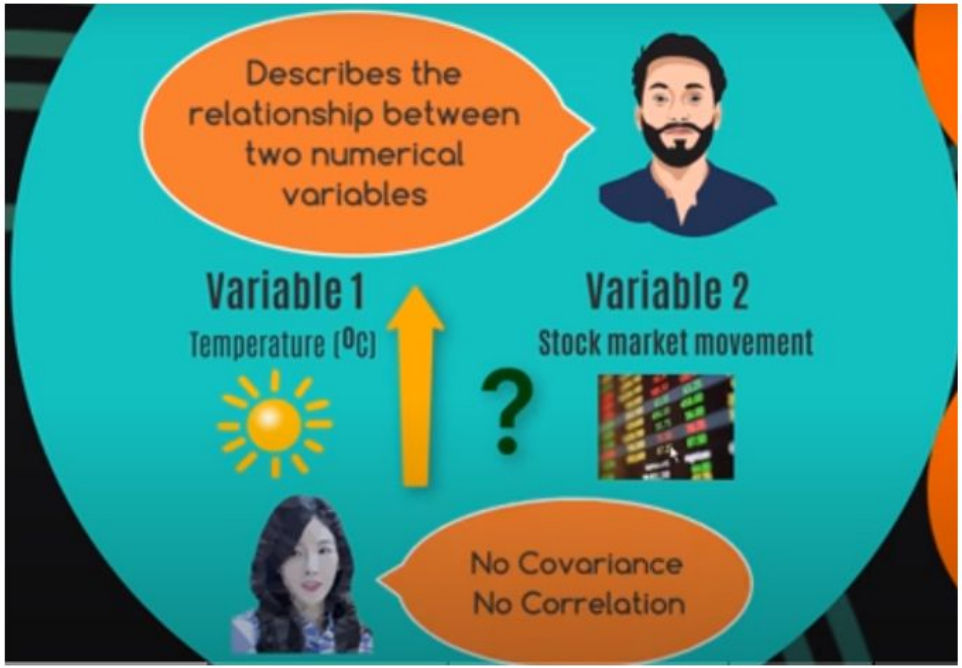


Introduction to covariance and correlation…

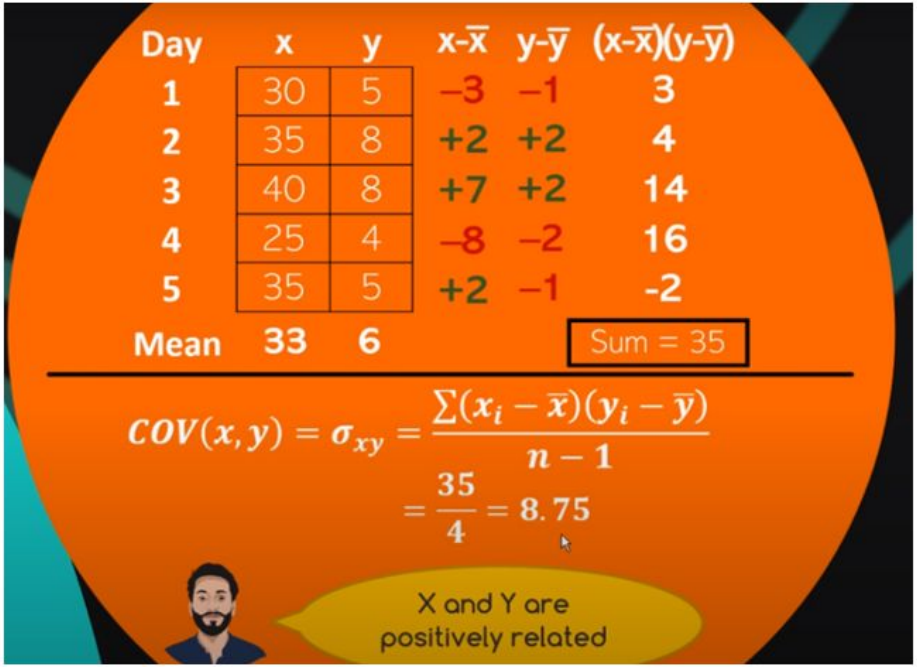
- Example2: -ve covariance and correlation



Introduction to covariance and correlation…



Covariance: Calculate covariance between 2 stock market data



Correlation

- Different correlation coefficients
    - Pearson correlation coefficient “r”
    - Spearman’s rank correlation ‘√’

Pearson correlation

- Assumes both X and Y are linear
    - -1 strong negative correlation
    - +1 strong positive correlation

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

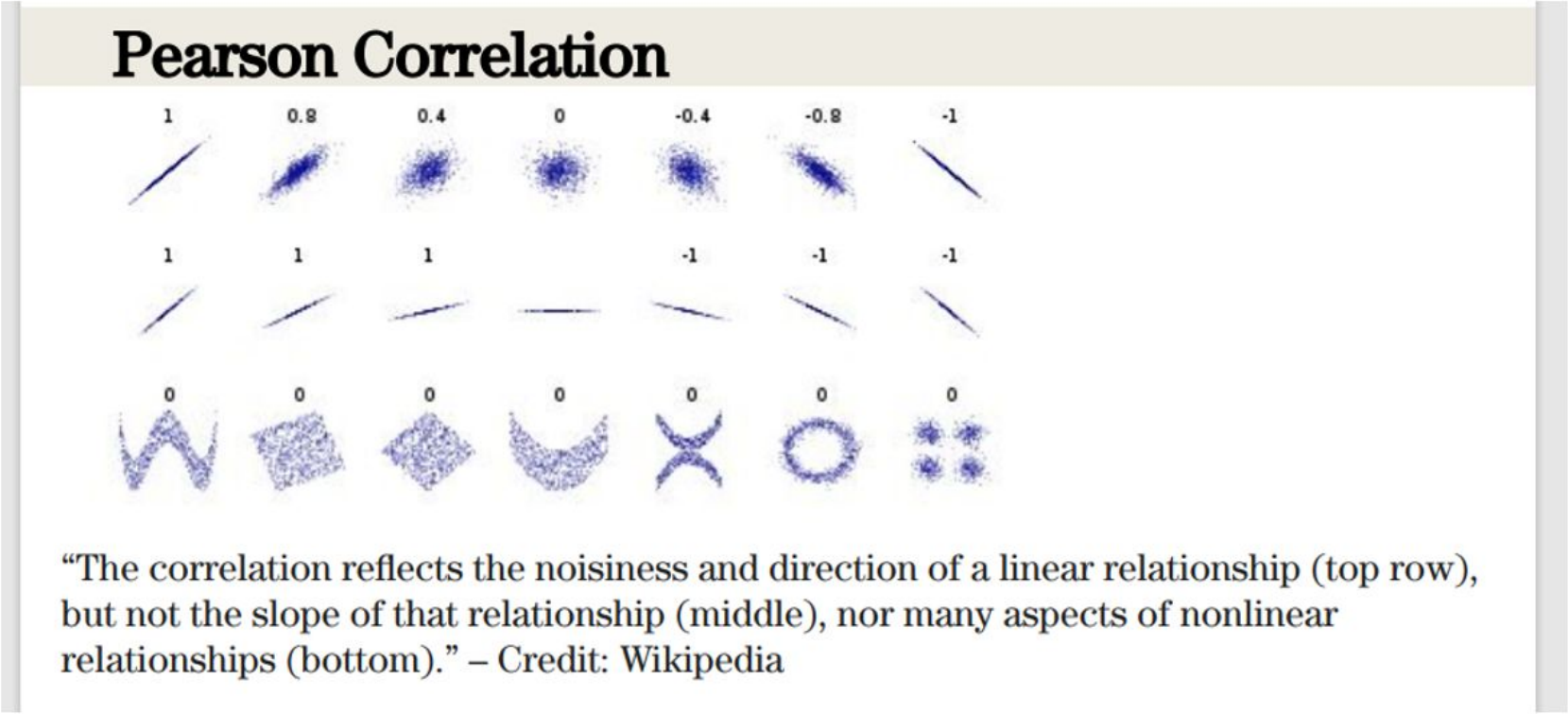
Find Pearson Correlation for

- Type equation here.

Day	x	y	$x-\bar{x}$	$y-\bar{y}$	$(x-\bar{x})(y-\bar{y})$
1	30	5	−3	−1	3
2	35	8	+2	+2	4
3	40	8	+7	+2	14
4	25	4	−8	−2	16
5	35	5	+2	−1	-2
Mean	33	6			Sum = 35

Note: it is giving both direction(+V) and strength(0.824) of relationship

Pearson correlation



"The correlation reflects the noisiness and direction of a linear relationship (top row), but not the slope of that relationship (middle), nor many aspects of nonlinear relationships (bottom)." – Credit: Wikipedia

Limitations of Pearson correlation

- Slope doesn’t indicate relationship
- Non-linearity of X and Y is not considered
- For this we can use Spearman’s Rank method

Spearman’s rank correlation

- Pearson’s correlation coefficient is not sensitive to nonlinear relationships
- We can use other correlation coefficients like Spearman’s or Kendall rank

Find Spearman's rank correlation between maths and physics marks

Student ID	Maths	Physics
1	35	30
2	23	33
3	47	45
4	17	23
5	10	8
6	43	49
7	9	12
8	6	4
9	28	31

Find the Spearman Rank Correlation Coefficient up to one decimal place.

- 0.6
- 0.4
- 0.9
- 0.3

Assignment: Try the same problem

Day	x	y
1	30	5
2	35	8
3	40	8
4	25	4
5	35	5

Outliers

- Outlier is a data object that deviates significantly from the rest of the data objects and behaves in a different manner.
- Example: 150years as the age of a human
- Can be seen on box plot or scatter plot