Fitting aggregation function

Week | 3

Understanding relationship

Correlation

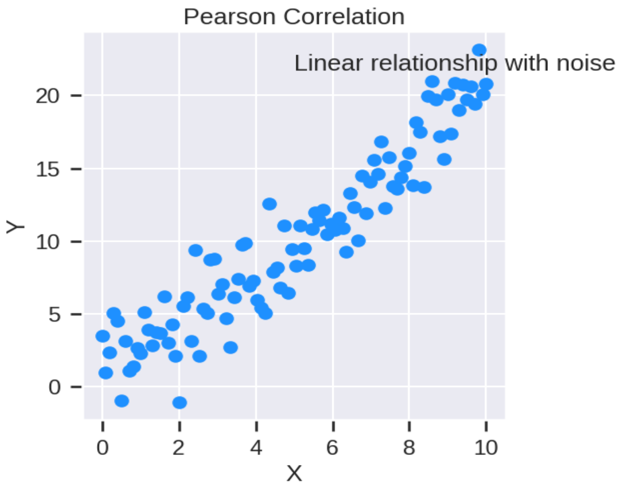
It is a statistical technique used to measure the strength and direction of the relationship between **two variables**

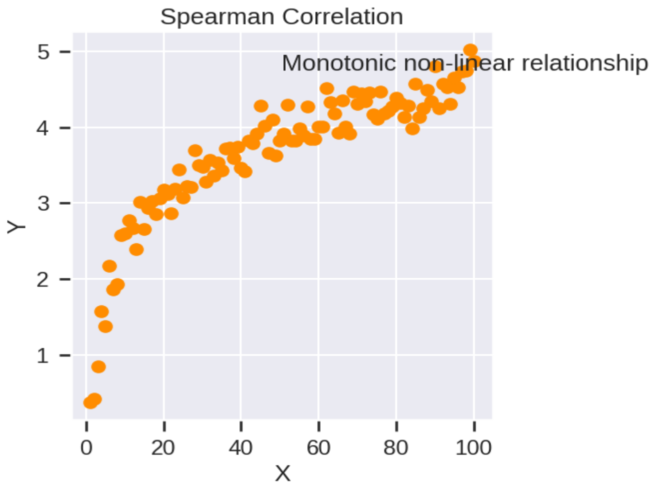
Methods to find relationship

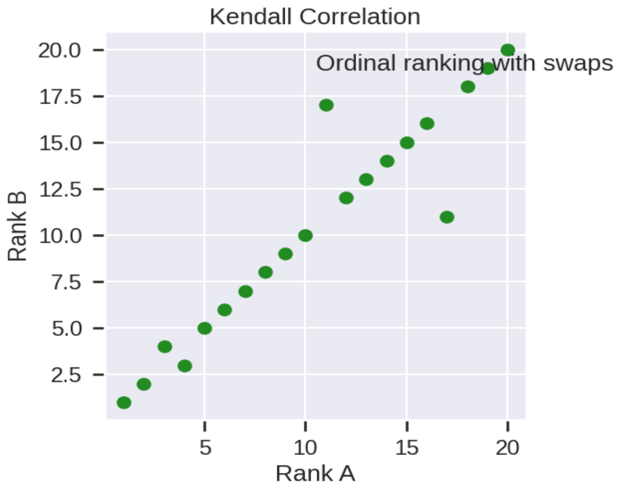
	Pearson Correlation	Spearman Correlation	Kendall Correlation
Symbol	r	ρ (rho)/ r_s	τ (Tau)
Scale	[-1,1]	[-1,1]	[-1,1]
Relationship	Linear (Continuous variables)	Monotonic (Continious or ordinal)	Monotonic (Continious or ordinal)
	As one increases other variable increases[1]. As one variable increases other decreases[-1]	As one variable increases other may not decrease [1]. If one variable increases other may not increase[-1]	As one variable increases other may not decrease [1]. If one variable increases other may not increase[-1]. Robust to outliers
Distribution	Normal (Sample size>=30)	Non-Parametric Sample size <30 or any value	Non-Parametric Sample size <30 or any value
Assumption	Homoscadascity	Variables are ranked	Concordant and Discordant pairs

When to use ?

Pearson	Spearman	Kendall
Best when to use: Continuous variables with a linear relationship	**Best when to use:** Variables have a monotonic relationship (consistently increasing or decreasing, not necessarily linear)	**Best when to use:** Ordinal or ranked data, small samples, or when robustness is needed
Example: Temperature vs electricity bill — as temperature rises, electricity bills rise in a roughly straight line (more AC use)	**Example:** Age vs number of wrinkles — as age increases, wrinkles increase, but not in a straight line (the rate changes)	**Example:** Ranking of movies by critics vs ranking by audience — comparing two ordered lists
Example: Height vs weight — taller people generally weigh more in a straight-line fashion	**Example**: Age vs blood pressure — as age increases, blood pressure tends to rise, but not in a straight line	**Example**: Ranking of students by math scores vs ranking by science scores — comparing two ordered lists







Pearson correlation

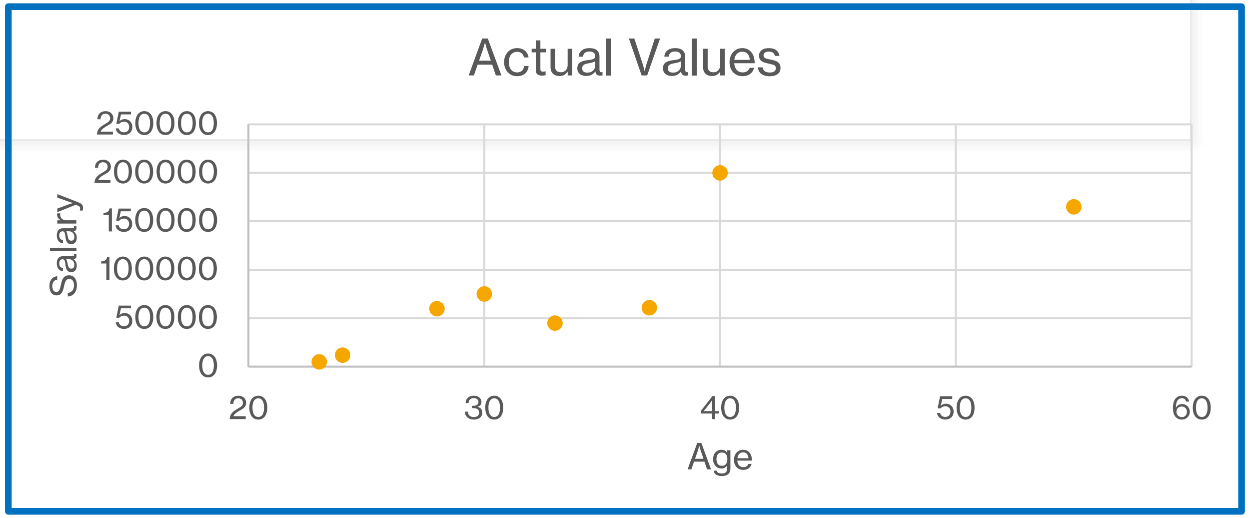
$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

Where,

r = Pearson Correlation Coefficient

x_i = x variable samples y_i = y variable sample

$\bar{x}$ = mean of values in x variable $\bar{y}$ =mean of values in y variable



Age (x)	Salary (y)
23	5000
24	12000
33	45000
30	75000
28	60000
37	61000
55	165000
40	200000

$\bar{x}$ = 33.75

$\bar{y}$ = 77875

	Age	*Salary*
Age	1	
Salary	0.81331	1

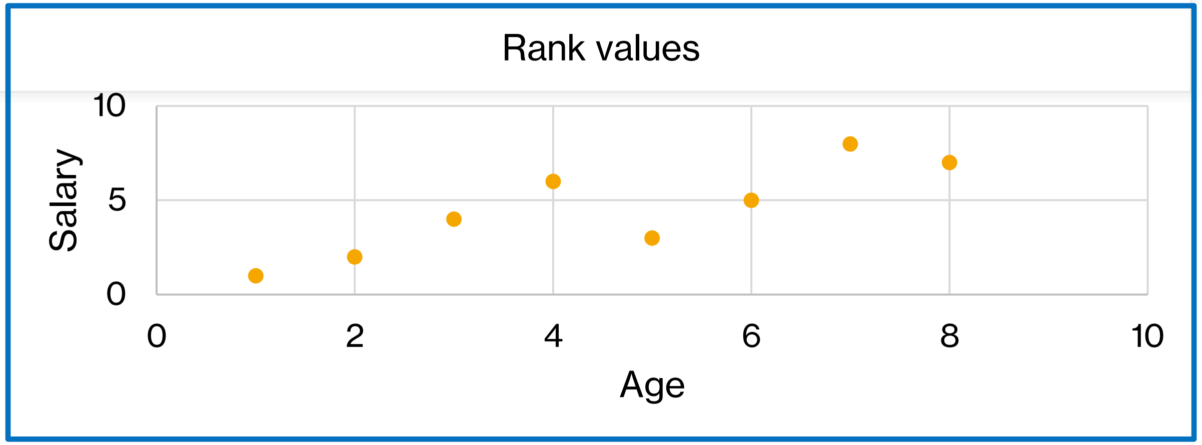
Spearman correlation

$$\rho = 1 - \frac{6\sum d_i^2}{n(n^2-1)}$$

ρ = Spearman's rank correlation coefficient

d_i = difference between the two ranks of each observation

n = number of observations



Values

Age	Salary
23	5000
24	12000
33	45000
30	75000
28	60000
37	61000
55	165000
40	200000

Rank

Age	Salary
1	1
2	2
5	3
4	6
3	4
6	5
8	7
7	8

Age	Salary	d	d^2
1	1	0	0
2	2	0	0
5	3	2	4
4	6	-2	4
3	4	-1	1
6	5	1	1
8	7	1	1
7	8	-1	1
			12

1- $\frac{(6*12)}{8*(64-1)}$

Spearman coeff 0.857143

Kendall correlation

- **Kendall correlation** is a statistical measure of the ordinal association between two measured quantities.
- Also commonly known as “Kendall’s tau coefficient”.
- Formula:

$$\tau = \frac{n_c - n_d}{n_c + n_d}$$

n_c - Concordant pairs

n_d - Discordant pairs

- The original values of the data is first ranked.
- The Kendall correlation coefficient depends only the order of the pairs, and it can always be computed assuming that one of the rank order serves as a reference point (e.g., with N = 4 elements we assume arbitrarily that the first order is equal to 1234). Therefore, with two rank orders provided on N objects, there are N! different possible outcomes (each corresponding to a given possible order) to consider for computing the sampling distribution of τ.

When should one use it ?

- When the sample size of data is small.
- When a dataset is non-parametric (If means it can be considered when data doesn’t follow a normal distribution)
- When the relationship between the variables is non-linear.
- When there are ties in the data

If we change the order of variables the coefficien

Kendall's Tau relies on comparing pairs of observations to determine if they are concordant (both variables increase or decrease together) or discordant (one variable increases while the other decreases). We should not sort data

Values

Age	Salary
23	5000
24	12000
33	45000
30	75000
28	60000
37	61000
55	165000
40	200000

Rank

Age	Salary
1	1
2	2
5	3
4	6
3	4
6	5
8	7
7	8

Same inputs – This should be right approach

Rank

Age	Salary							
1	1							
2	2	+						
5	3	+	+					
4	6	+	+	+				
3	4	+	+	+	-			
6	5	+	+	+	-	+		
8	7	+	+	+	+	+	+	
7	8	+	+	+	+	+	+	+

$\frac{26-2}{26+2}$ $\frac{24}{28}$ 0.857

Changing input X –Age

Rank

Age	Salary							
1	1							
2	2	+						
3	4	+	+					
4	6	+	+	+				
5	3	+	+	-	-			
6	5	+	+	+	-	+		
7	8	+	+	+	+	+	+	
8	7	+	+	+	+	+	+	-

$\frac{24-4}{24+4}$ $\frac{20}{28}$ 0.71

Changing input Y –Salary

Rank

Age	Salary							
1	1							
2	2	+						
5	3	+	+					
3	4	+	+	+				
6	5	+	+	+	+			
4	6	+	+	+	+	+		
8	7	+	+	+	+	+	+	
7	8	+	+	+	+	+	+	+

$\frac{28-0}{28+0}$ $\frac{28}{28}$ 1.0

An alternative formula to calculate Kendall Coefficient

$$\tau = \frac{n_c - n_d}{\frac{n(n-1)}{2}}$$

n_c - Concordant pairs, n_d - Discordant pairs, n –Sample Size

Rank

Age	Salary							
1	1							
2	2	+						
5	3	+	+					
4	6	+	+	+				
3	4	+	+	+	-			
6	5	+	+	+	-	+		
8	7	+	+	+	+	+	+	
7	8	+	+	+	+	+	+	+

$\frac{26-2}{[8*(8-1)]/2}$ $\frac{24}{56/2}$ $\frac{24}{28}$ 0.857

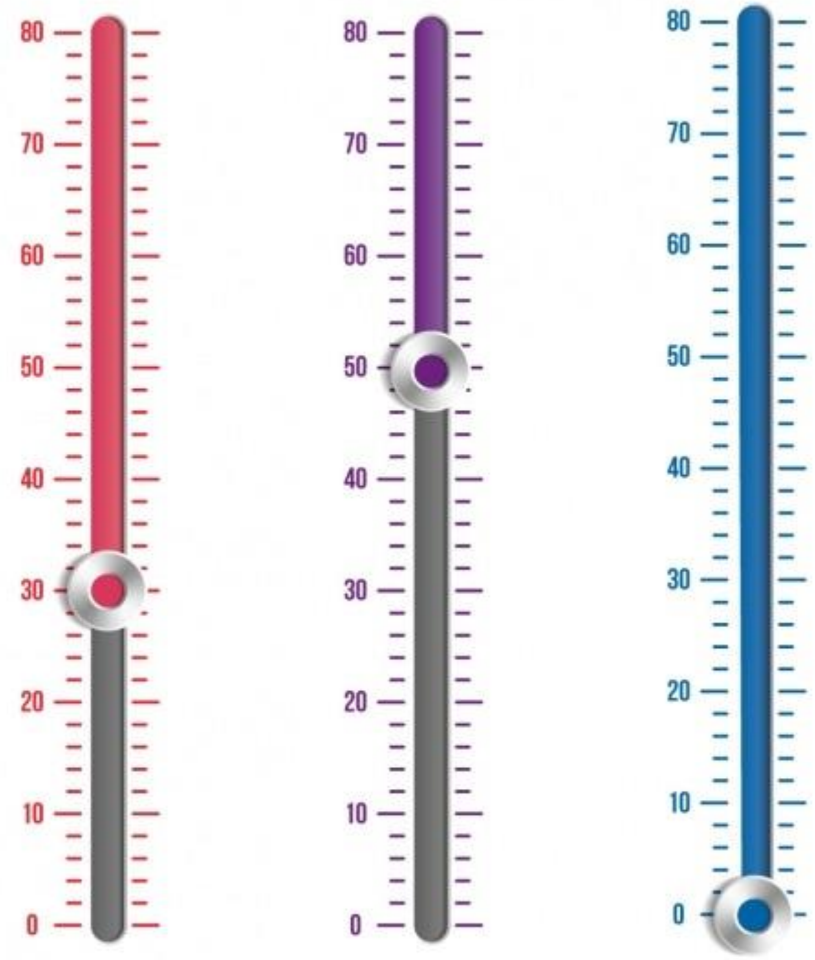
Errors Measures

Formulas for calculating errors

Technique	Abbreviation	Error Calculation
Sum of Absolute Difference/ Sum of Absolute Error	SAD/ SAE	$\Sigma(\lvert A-P \rvert)$
Mean Absolute Error/ Average Absolute Error	MAE/ Av.AE	$\frac{1}{n}\Sigma(\lvert A-P \rvert)$
Sum of Squared Difference	SSD	$\Sigma(A-P)^2$
Mean Squared Error	MSE	$\frac{1}{n-k-1}\Sigma(A-P)^2$
Root Mean Square Error	RMSE	$\sqrt{\frac{1}{n-k-1}\Sigma(A-P)^2}$
Mean Absolute Percent Error	MAPE	$\frac{1}{n}\Sigma\left[\frac{(A-P)}{A}\right]\times 100$

A- Actual data
P- Predicted output
n – Number of observations
k- Number of columns/features

Knowing Similarities



Distance measurement

- Distance between points (smaller the difference = similar, higher the difference = dissimilar)
- Examples of common distance measures
  - **Manhattan Distance** = |8|+|4| = 12
  - For m variables

$$|x_2-x_1|+|y_2-y_1|+|z_2-z1|+.....+m$$

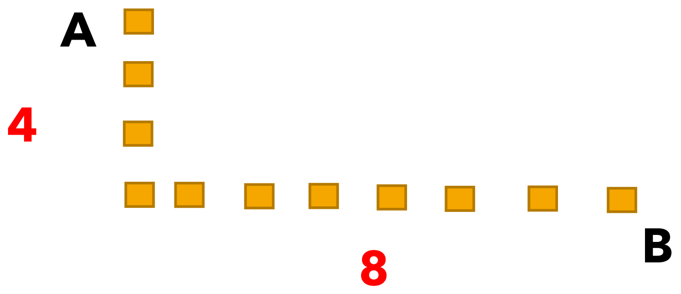
  - **Euclidean Distance** = Sqrt(8^2+4^2) = 8.94
  - For m variables

$$\sqrt{(x_2-x_1)^2+(y_2-y_1)^2+(z_2-z_1)^2_{\ldots}+m}$$

  - **Chebyshev Distance** (Chess board distance) = Max(8,4) = 8

For m variables

$$Max(|x_2-x_1|+|y_2-y_1|+|z_2-z1|+.....+m)$$



Thank you