PODS Lab 5: Correlation

Hamza Alshamy

Center for Data Science, NYU
ha2486@nyu.edu

02/21/25

1. **Covariance:** Relationship between Variables
2. **Linear Correlation**
   - Pearson Correlation $(\rho_{x,y})$
3. **Non-linear Correlation**
   - Spearman Correlation $(p_{x,y})$
   - Chatterjee's Correlation $(\xi_{x,y})$
4. **Handling Missing Data**

Motivation: Quantify the relationship between variables

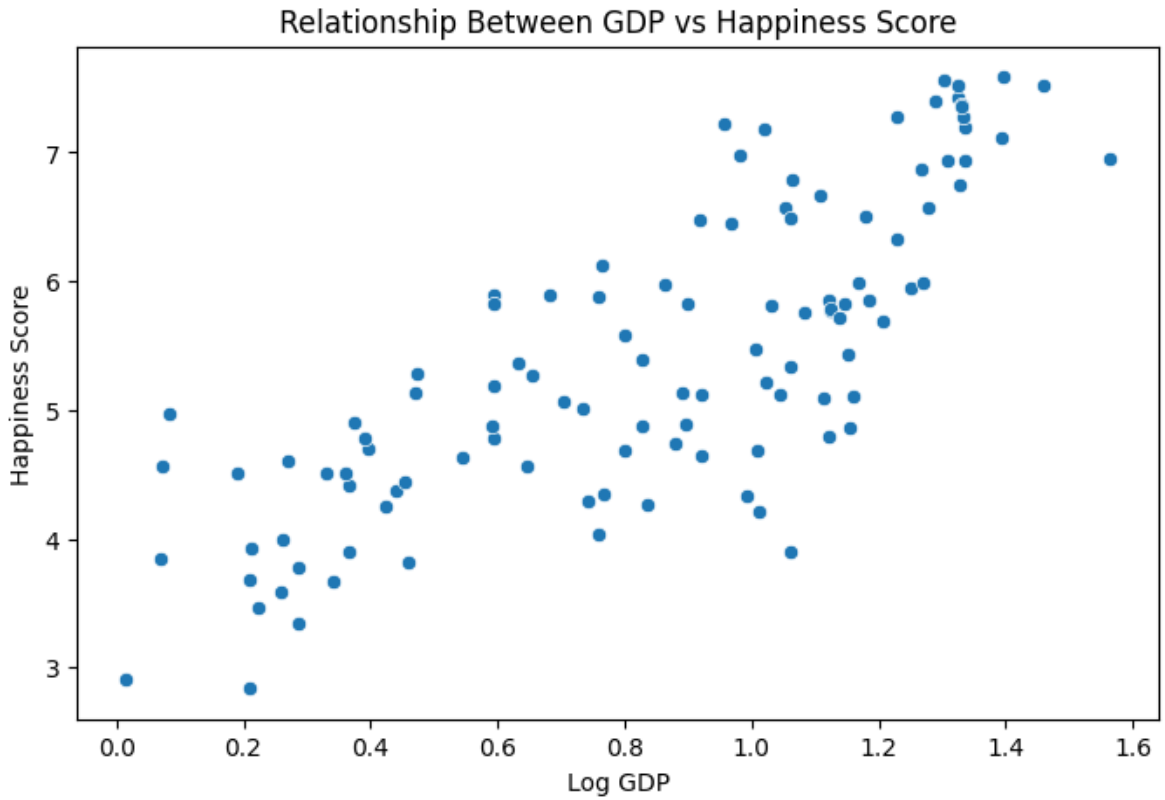
- **Motivation:** Quantify relationship between variables in a number.
- **Ideally** in a way that allows us to compare different relationships.

Motivation: Quantify the relationship between variables

- **Motivation:** Quantify relationship between variables in a number.
- **Ideally** in a way that allows us to compare different relationships.
- **Example:** Can we compare the relationship between height and IQ V.S. income and happiness?

Example: GDP (income) and Happiness

- You want to **quantify** this relationship.
- **How?**



Possible Way: Covariance

- One way to quantify the relationship between two variables is through **Covariance**.
- Start by recalling how standard deviation is defined.

$$\sigma_x = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \mu_x)^2} \qquad \text{(Standard Deviation)}$$

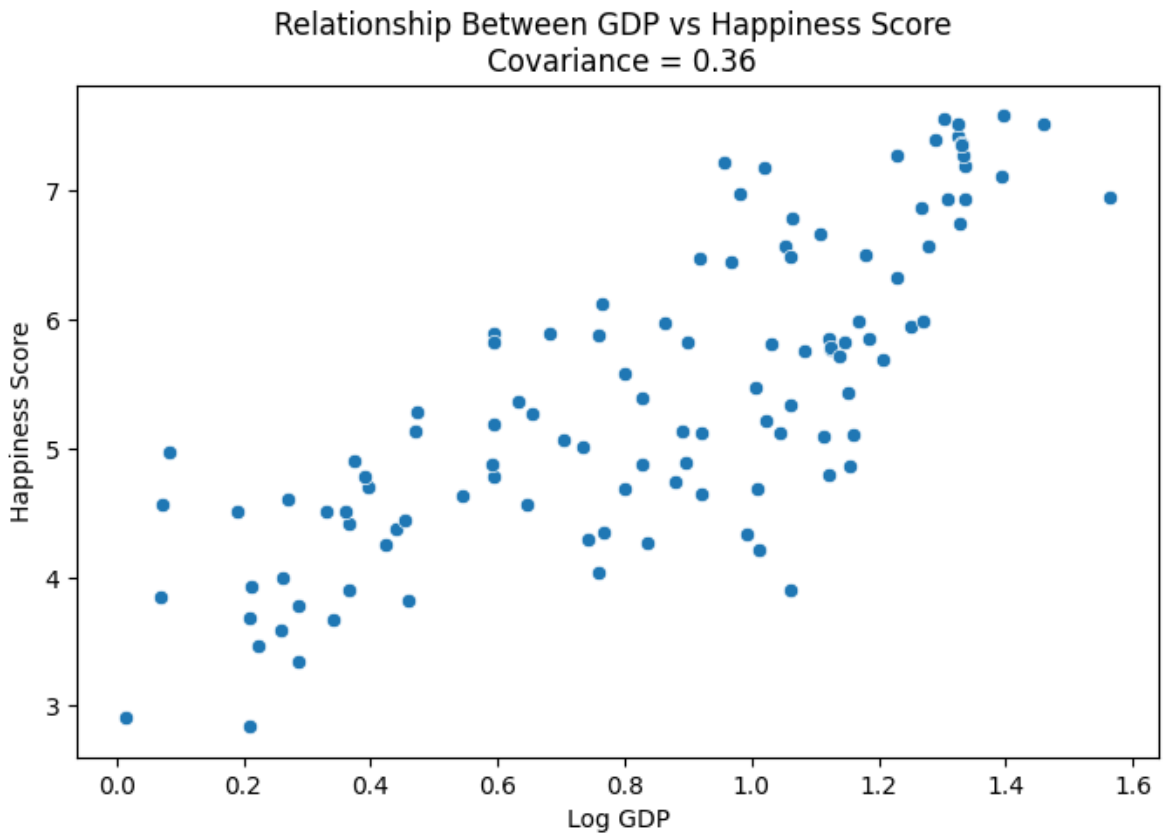
$$\sigma_x^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu_x)^2 \qquad \text{(Variance)}$$

$$\sigma_x^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu_x)(x_i - \mu_x)$$

- Instead of having $(x_i - \mu_x)$ twice, generalize it to two variables:

$$\boxed{\text{Cov}(X,Y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu_x)(y_i - \mu_y)} \qquad \text{(Covariance)}$$

Example Revisited: GDP (income) and Happiness Covariance



- A **positive** covariance means that both log GDP and happiness move together.
- Covariance values are **unstandardized**, meaning they depend on the units of the variables

Covariance

$$\text{Cov}(X,Y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu_x)(y_i - \mu_y)$$

- **Covariance** looks at how variable **vary** together.
- **Issue:** Not standardized
  - Variables x and y might be measured on different scales
  - **Covariance** does NOT take that into account
  - How to interpret? Can we compare different covariances?
  - **Range:** Cov(X,Y) $\in [-\infty, \infty]$

From Covariance to (Pearson) Correlation

The issue: Covariance depends on the units of its variables. Can we remove the units (standardize)? – **Yes!**

- We standardize by dividing the covariance by the standard deviation of each variable.

Remember that:

$$\text{Cov}(X,Y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu_x)(y_i - \mu_y)$$

Pearson Correlation Formula

$$r_{X,Y} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \mu_x)(y_i - \mu_y)}{\sigma_X \sigma_Y} = \boxed{\frac{\text{Cov}(X,Y)}{\sigma_X \sigma_Y}}$$

Example Revisited: GDP and Happiness Correlation

- **Covariance (0.36):** Positive relationship, but its magnitude is difficult to interpret.
- **Pearson Correlation (0.79):** Standardizes covariance, making it **unit-free** and easier to interpret.

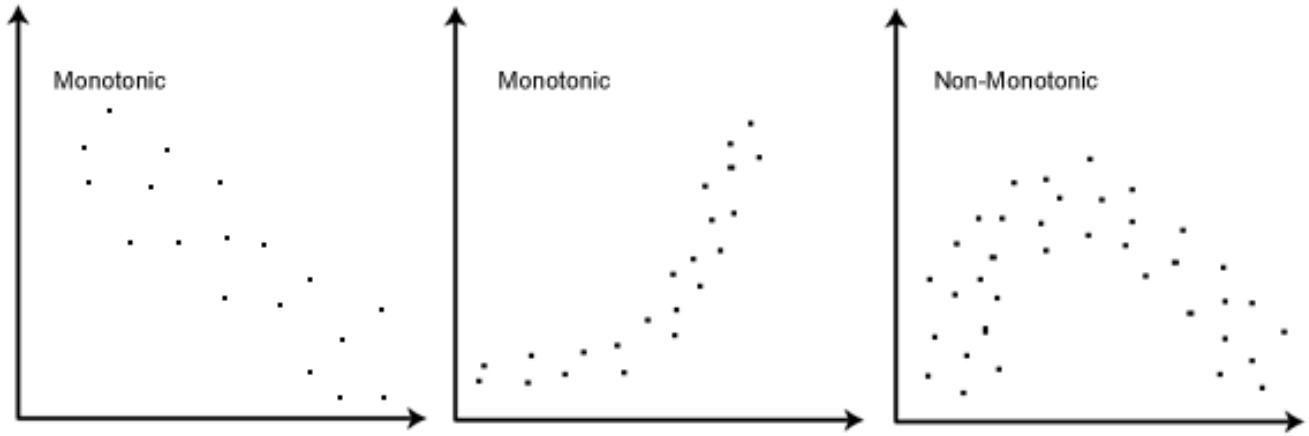
Pearson Correlation – Linear

$$r_{X,Y} = \frac{\text{Cov}(X,Y)}{\sigma_X \sigma_Y} \qquad \text{(Pearson Correlation)}$$

Properties:

- **Range:** $r_{X,Y} \in [-1, 1]$.
  - $r = 1$: Perfect positive **linear** relationship.
  - $r = -1$: Perfect negative **linear** relationship.
  - $r = 0$: No **linear** relationship (but other forms of association may exist).
- **Linear Association Only:** Nonlinear relationships may exist even if $r \approx 0$.
- **Unit-Free:** Since it is standardized, correlation does not depend on the units of X and Y.
- **Not Causation:** Correlation does not imply causation – only association.

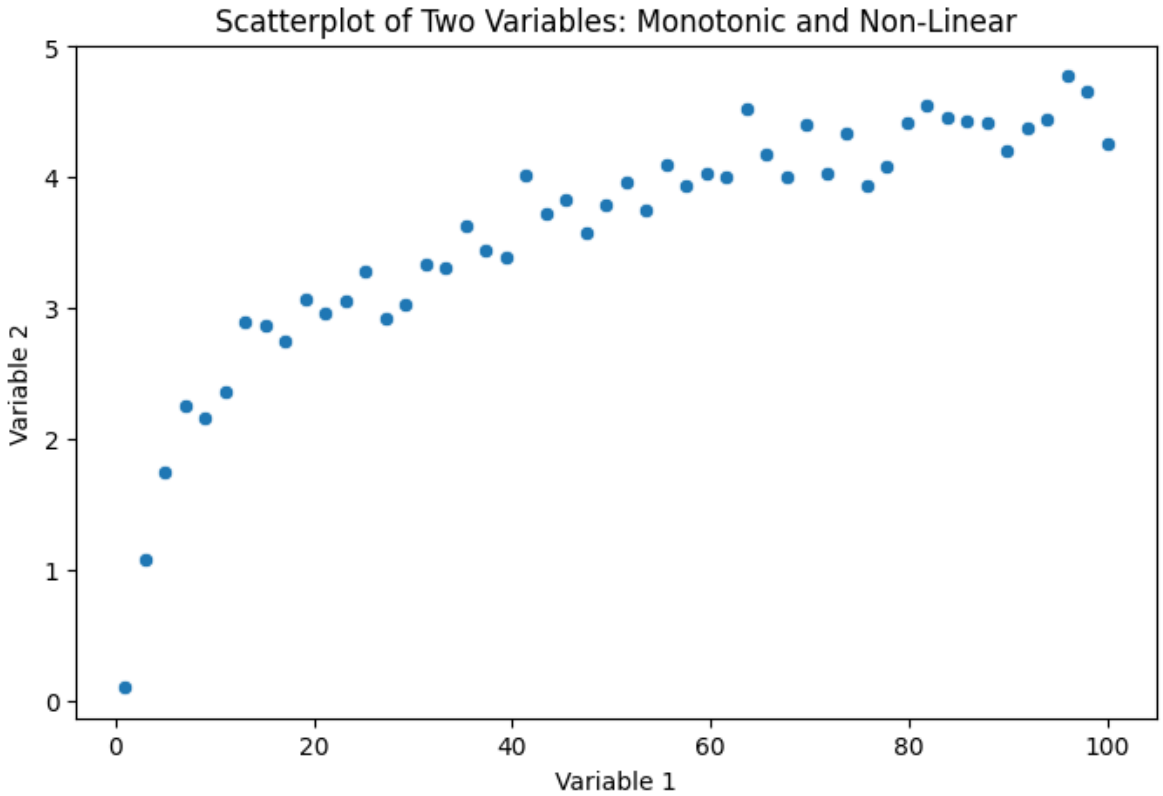
Non-linear Correlation



1. **Linear:** Increases/decreases at a constant rate.
   - Pearson Correlation
2. **Monotonic:** Increases/decreases at a varying rate.
   - Spearman Correlation
3. **Non-monotonic:** Changes direction.
   - Chatterjee Correlation

Spearman (rank) Correlation ρ – Monotonic

- **Suppose you want to quantify the relationship between these two variables:**



- **Issue:** Pearson correlation $r_{X,Y}$ only captures linear associations.

Spearman (Rank) Correlation ρ – Monotonic

- Introducing **Spearman Correlation** (ρ)!
- Captures **monotonic** relationships – detects trends even if they are nonlinear.
- Based on the **rank function**, replacing raw data values with their ranks.
- **Properties:**
  - **Range:** $\rho \in [-1, 1]$, just like Pearson correlation.
  - **Monotonicity:** Measures **monotonic association**
  - **Resistant to Outliers:** Since it is based on ranks, extreme values have less influence than in Pearson correlation.

Spearman Correlation – Rank Function

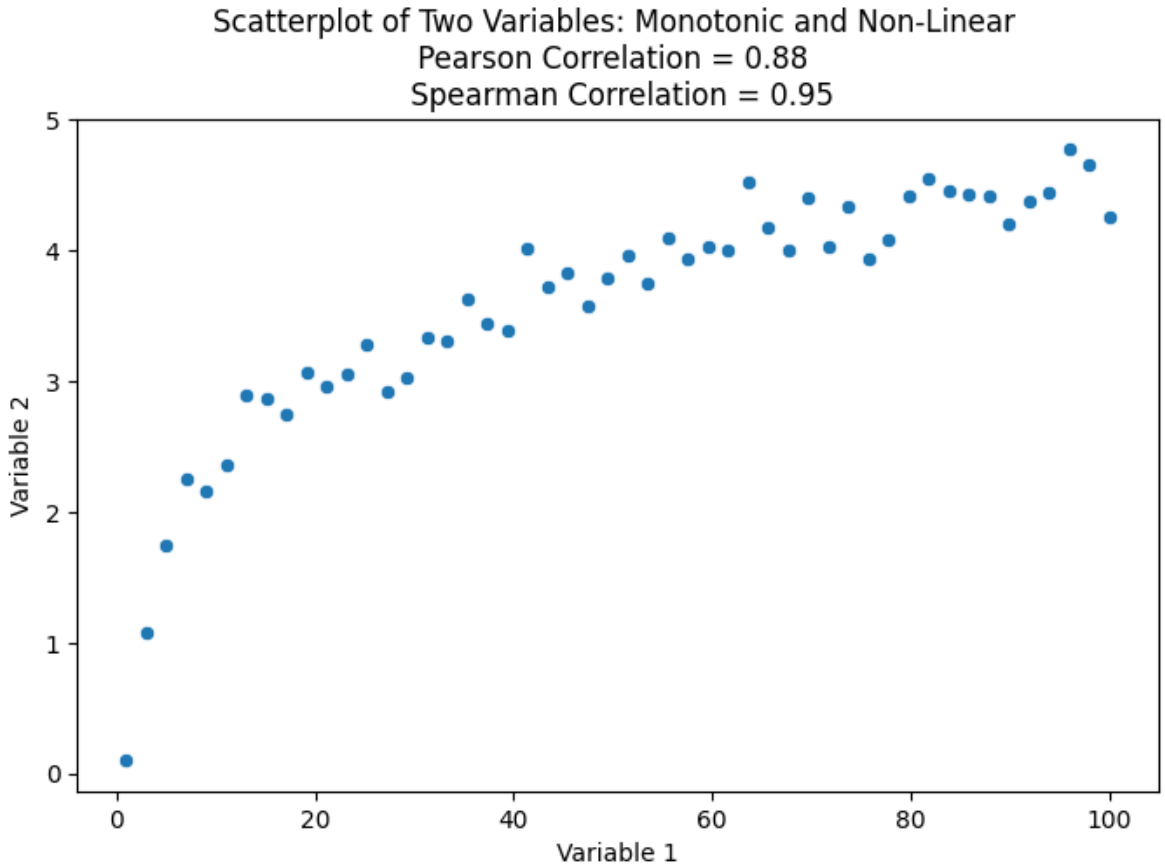
Example: IQ and Height Ranks

IQ (Score)	Rank (IQ)	Height (in")	Rank (Height)	d
135	1	74	3	-2
120	2	78	1	1
115	3	70	5	-2
105	4	76	2	2
95	5	72	4	1

Spearman Correlation Formula

$$\rho_{X,Y} = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)}$$

Spearman (rank) Correlation ρ – Monotonic

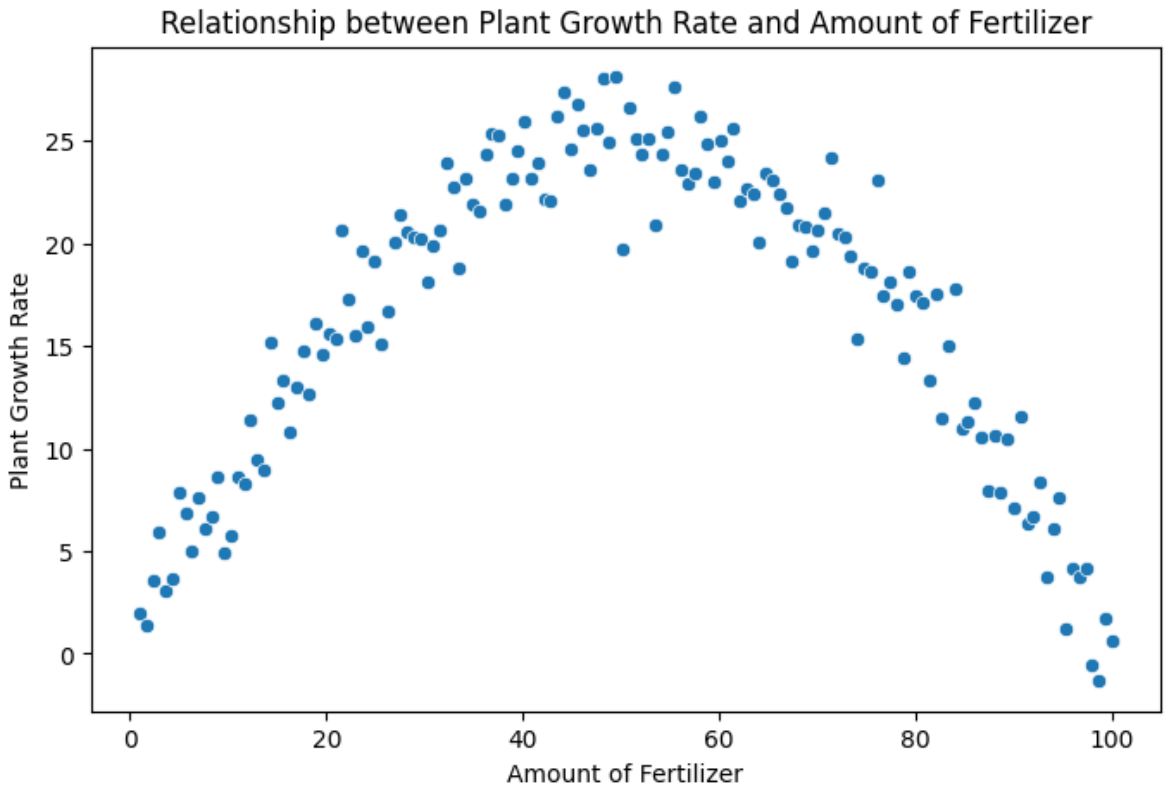


- Notice how Spearman correlation $\rho_{X,Y}$ captures the non-linearity in the relationship better than Pearson correlation $r_{X,Y}$.

Chatterjee Correlation

Example: Plant Growth Rate and Amount of Fertilizer

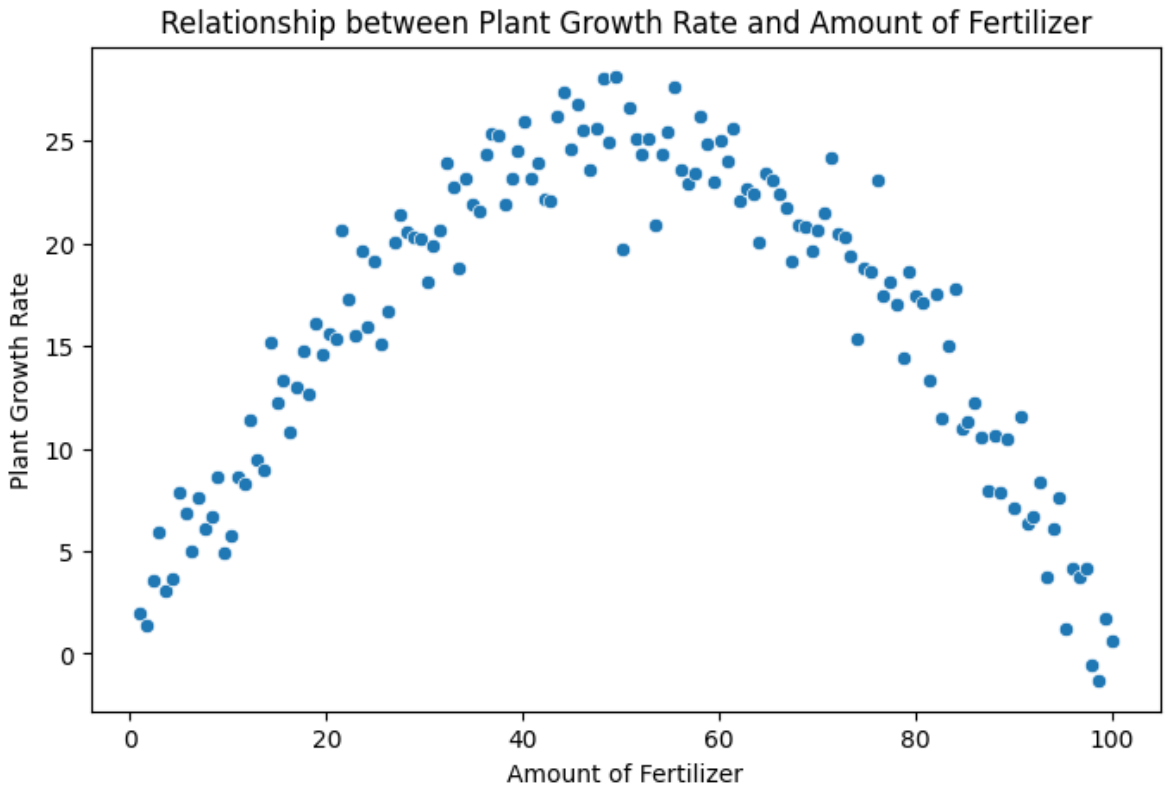
- **Question:** What if the relationship between the two variables are **non-linear** AND **non-monotonic**?



Chatterjee Correlation

Example: Plant Growth Rate and Amount of Fertilizer

- **Question:** What if the relationship between the two variables are **non-linear** AND **non-monotonic**?



- **Answer:** Chatterjee Correlation $\xi_{X,Y}$!

Chatterjee Correlation ξ

Chatterjee Correlation Formula

$$\xi := 1 - \frac{3\sum_{i=1}^{n-1} |r_{i+1} - r_i|}{n^2 - 1}$$

where:

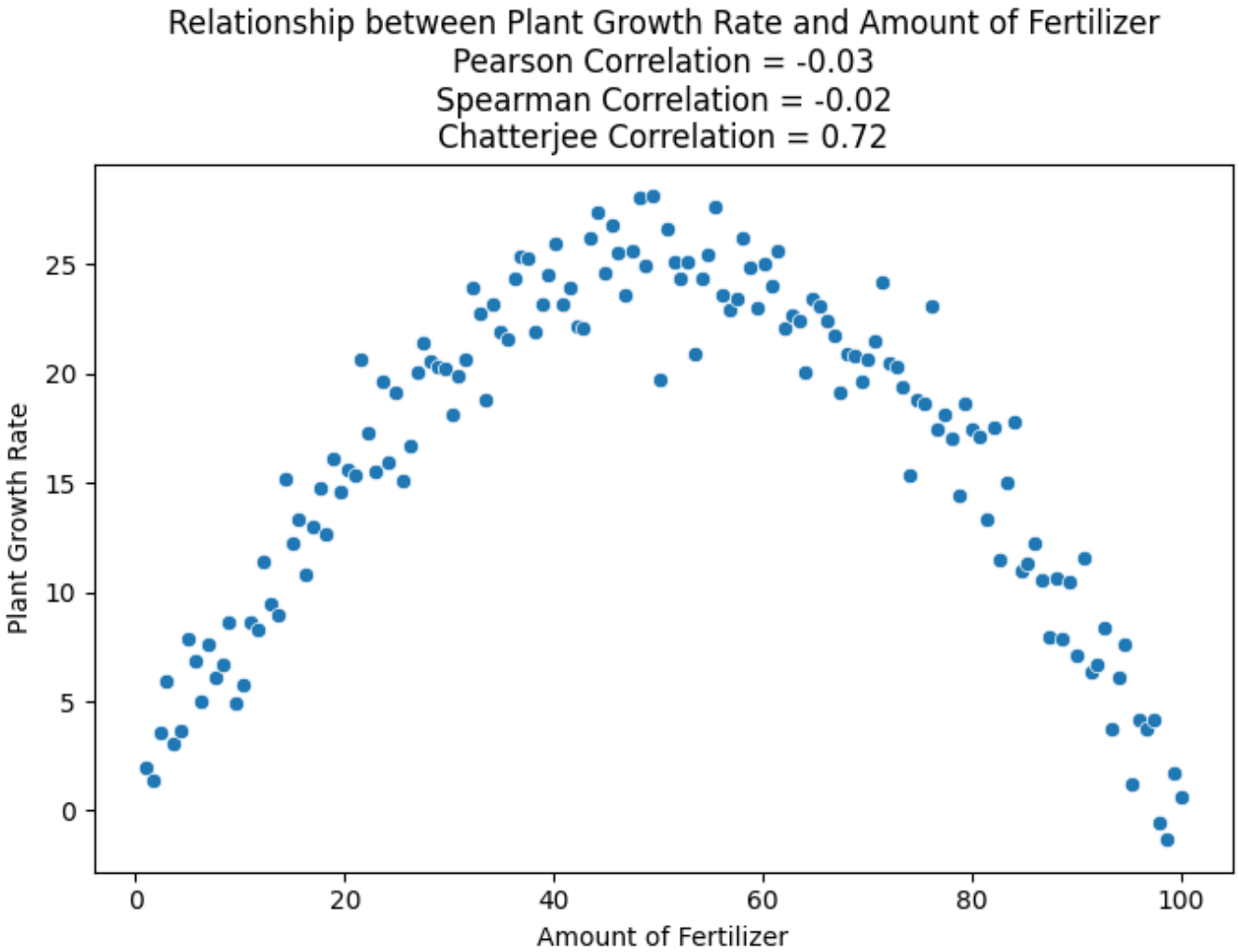
- ξ: Chatterjee Correlation.
- r_i: The rank of X_i in $X_1, X_2, ..., X_n$.
- n: Total number of observations.

Properties

- Captures **non-monotonic** relationships.
- **Range:** $0 \leq \xi \leq 1$

Chatterjee Correlation

Example Revisited: Plant Growth Rate and Amount of Fertilizer



Summary of Correlation

(1) Pearson Correlation ($r_{X,Y}$)

- **Context:** Linear Relationship
- **Range:** $r \in [-1, 1]$

(2) Spearman Correlation ($\rho_{X,Y}$)

- **Context:** Measures monotonic relationships (captures some non-linear trends).
- **Range:** $\rho \in [-1, 1]$

(3) Chatterjee Correlation ($\xi_{X,Y}$)

- **Context:** Captures general dependence (non-monotonic & non-linear)
- **Range:** $\xi \in [0, 1]$

How to Handle Missing Data (NaN)?

Three Ways to Handle Missing Data:

1. **Dropping missing entries:** Row-wise
2. **Dropping missing entries:** Element-wise
3. Imputing missing data

Handling Missing Data

(1) Row-wise Removal

Harry Potter and the Chamber of Secrets (2002)	Harry Potter and the Sorcerer's Stone (2001)
	3
	1.5
4	3
2	
2	3

- **In row-wise removal**, drop the entire row when a user has not rated all movies.
- Reduces sample size.

Handling Missing Data

(2) Element-wise Removal

Harry Potter and the Chamber of Secrets (2002)	Harry Potter and the Sorcerer's Stone (2001)
	3
	1.5
4	3
2	
2	3

- **In element-wise removal**, drop only the entries that have missing values.
- Larger sample size **but** may introduce survivorship bias
- **Example:** Users who enjoyed the first Harry Potter are more likely to enjoy the second Harry Potter, whereas individuals who did not enjoy the first Harry Potter are less likely to watch the second Harry Potter

Handling Missing Data

(3) Imputing

Harry Potter and the Chamber of Secrets (2002)	Harry Potter and the Sorcerer's Stone (2001)
	3
	1.5
4	3
2	
2	3

*Predicted value for entry

- **In imputation**, we predict the values of the missing entries via some predecitve method (e.g. mean/median of the column).