UNIVERSITÀ
DEGLI STUDI
DI PADOVA

Machine Learning
2024/2025

AMCO
ARTIFICIAL INTELLIGENCE, MACHINE
LEARNING AND CONTROL RESEARCH GROUP

Lecture #04
Correlation & Data Visualization

Gian Antonio Susto

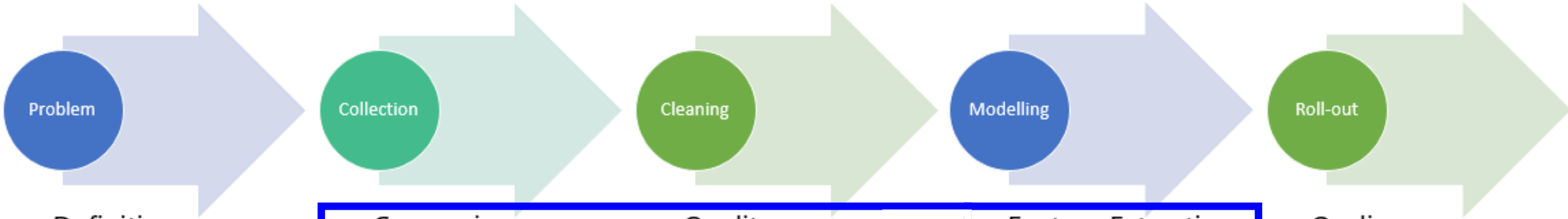
Recap – Tabular Data (the 'design matrix') - *x*

p attributes (variables, features) potentially related to the phenomenon under examination

n observations: the number of times the phenomenon we need to 'describe' is available in our data through historical examples

39	State-gov	77516	Bachelors	13	Never-married	Adm-clerical	Not-in-family	White	Male	2174	0	40	United-States	<=50K
50	Self-emp-not-inc	83311	Bachelors	13	Married-civ-spouse	Exec-managerial	Husband	White	Male	0	0	13	United-States	<=50K
38	Private	215646	HS-grad	9	Divorced	Handlers-cleaners	Not-in-family	White	Male	0	0	40	United-States	<=50K
53	Private	234721	11th	7	Married-civ-spouse	Handlers-cleaners	Husband	Black	Male	0	0	40	United-States	<=50K
28	Private	338409	Bachelors	13	Married-civ-spouse	Prof-specialty	Wife	Black	Female	0	0	40	Cuba	<=50K
37	Private	284582	Masters	14	Married-civ-spouse	Exec-managerial	Wife	White	Female	0	0	40	United-States	<=50K
49	Private	160187	9th	5	Married-spouse-absent	Other-service	Not-in-family	Black	Female	0	0	16	Jamaica	<=50K
52	Self-emp-not-inc	209642	HS-grad	9	Married-civ-spouse	Exec-managerial	Husband	White	Male	0	0	45	United-States	>50K
31	Private	45781	Masters	14	Never-married	Prof-specialty	Not-in-family	White	Female	14084	0	50	United-States	>50K
42	Private	159449	Bachelors	13	Married-civ-spouse	Exec-managerial	Husband	White	Male	5178	0	40	United-States	>50K
37	Private	280464	Some-college	10	Married-civ-spouse	Exec-managerial	Husband	Black	Male	0	0	80	United-States	>50K
30	State-gov	141297	Bachelors	13	Married-civ-spouse	Prof-specialty	Husband	Asian-Pac-Islander	Male	0	0	40	India	>50K
23	Private	122272	Bachelors	13	Never-married	Adm-clerical	Own-child	White	Female	0	0	30	United-States	<=50K
32	Private	205019	Assoc-acdm	12	Never-married	Sales	Not-in-family	Black	Male	0	0	50	United-States	<=50K
40	Private	121772	Assoc-voc	11	Married-civ-spouse	Craft-repair	Husband	Asian-Pac-Islander	Male	0	0	40	?	>50K
34	Private	245487	7th-8th	4	Married-civ-spouse	Transport-moving	Husband	Amer-Indian-Eskimo	Male	0	0	45	Mexico	<=50K
25	Self-emp-not-inc	176756	HS-grad	9	Never-married	Farming-fishing	Own-child	White	Male	0	0	35	United-States	<=50K
32	Private	186824	HS-grad	9	Never-married	Machine-op-inspct	Unmarried	White	Male	0	0	40	United-States	<=50K
38	Private	28887	11th	7	Married-civ-spouse	Sales	Husband	White	Male	0	0	50	United-States	<=50K
43	Self-emp-not-inc	292175	Masters	14	Divorced	Exec-managerial	Unmarried	White	Female	0	0	45	United-States	>50K
40	Private	193524	Doctorate	16	Married-civ-spouse	Prof-specialty	Husband	White	Male	0	0	60	United-States	>50K
54	Private	302146	HS-grad	9	Separated	Other-service	Unmarried	Black	Female	0	0	20	United-States	<=50K
35	Federal-gov	76845	9th	5	Married-civ-spouse	Farming-fishing	Husband	Black	Male	0	0	40	United-States	<=50K
43	Private	117037	11th	7	Married-civ-spouse	Transport-moving	Husband	White	Male	0	2042	40	United-States	<=50K
59	Private	109015	HS-grad	9	Divorced	Tech-support	Unmarried	White	Female	0	0	40	United-States	<=50K

Recap – The Machine Learning pipeline



- Definition
- Expected Impact
- Evaluation metric

- Conversion
- Parsing
- Aggregation
- Alignment

- Quality
- Reconciliation
- Missing data handling
- Denoising
- Outlier detection

- Feature Extraction
- Building
- Evaluation/ Comparison

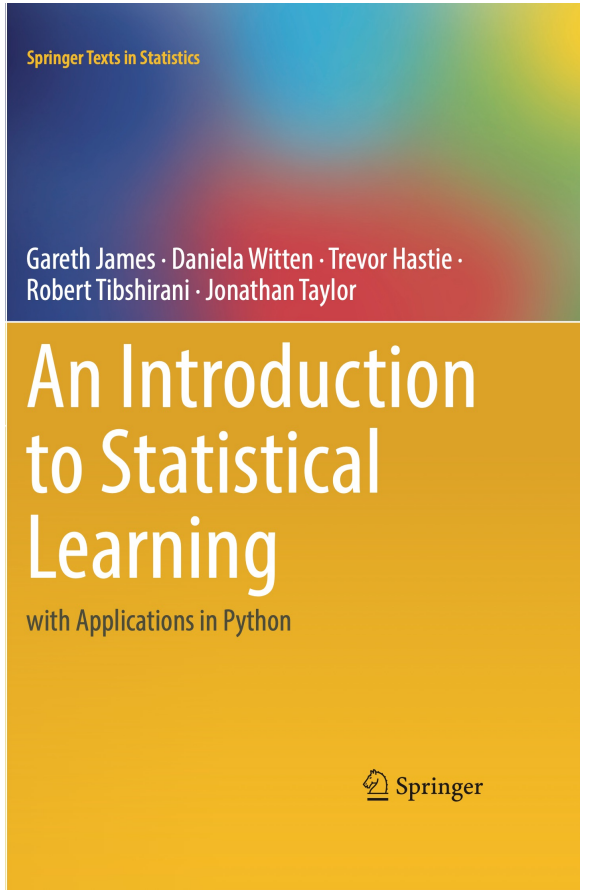
- On-line implementation
- Business outcome
- Improvement

Preprocessing

Recap – Why use Statistics in ML?

Why Use Statistics in ML?

1. [Preprocessing] Data Understanding – Descriptive statistics (mean, variance, distributions) help explore and clean data, identifying patterns and outliers.
2. [Preprocessing] Feature Engineering – Techniques like correlation analysis, PCA, and scaling rely on statistical principles.
3. [Building] Probability & Uncertainty – ML often deals with probabilistic models (e.g., Naïve Bayes) and uncertainty estimation.
4. [Building] Generalization & Inference – Concepts like overfitting, hypothesis testing, and bias-variance tradeoff come from statistics.
5. [Evaluation] Model Evaluation – Metrics like MSE, MAE, accuracy, precision, and recall are rooted in statistical concepts.



Recap – Statistical moments in ML

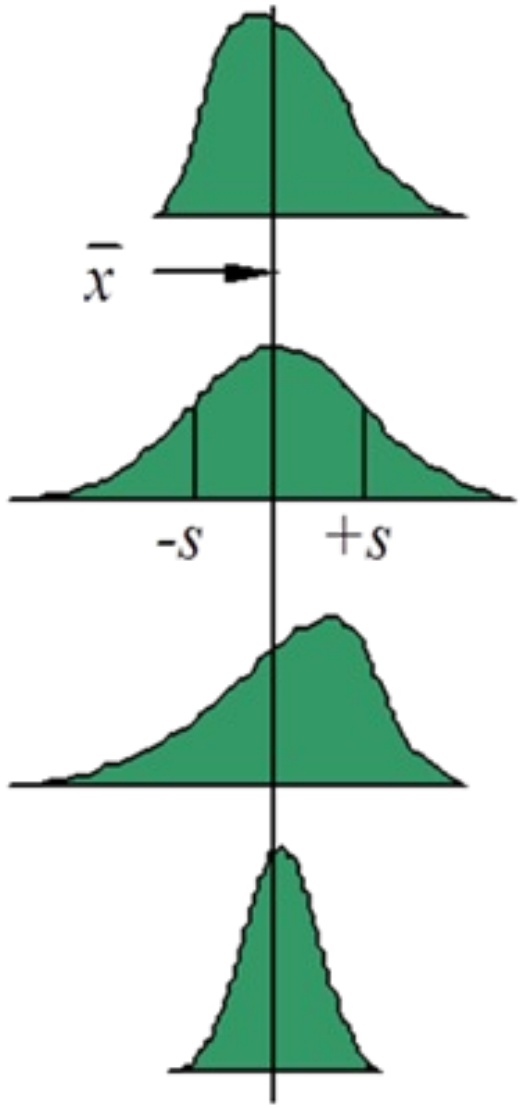
Moments of a Random Variable X

1.First Moment (Mean) – Central Tendency
$E[X] = \mu$
The expected value of X, representing the average outcome.

2.Second Moment (Variance) – Spread of Data
$Var(X) = E[(X - \mu)^2] = \sigma^2$
Measures how far values of X deviate from the mean.

3.Third Moment (Skewness) – Asymmetry of Distribution
$Skew(X) = E[(X - \mu)^3] / \sigma^3$
Indicates whether the distribution leans right (negative skew) or left (positive skew).

4.Fourth Moment (Kurtosis) – Tailedness of Distribution
$Kurt(X) = E[(X - \mu)^4] / \sigma^4$
Measures how heavy or light the tails of the distribution are compared to a normal distribution.

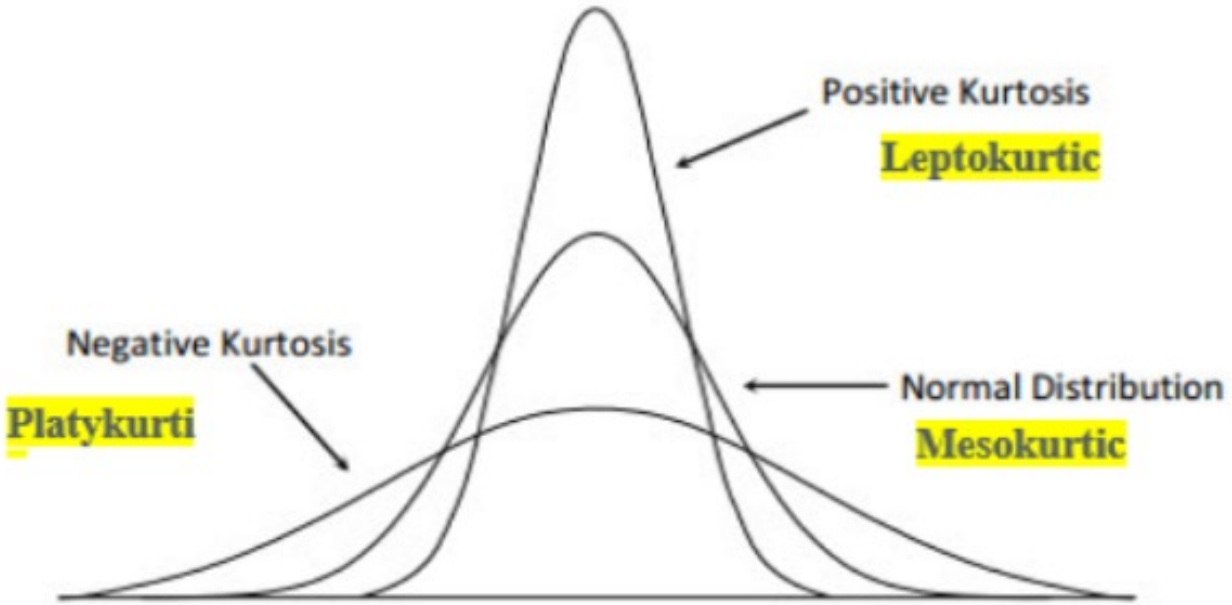


‘Recap’ – Statistical moments in ML

Normal Kurtosis (K = 3) - Mesokurtic
•The distribution has the same shape as the normal distribution.

High Kurtosis (K > 3) - Leptokurtic
•Heavier tails than the normal distribution, meaning more extreme values (outliers).
•The distribution is more "peaked" in the center and has longer tails.

Low Kurtosis (K < 3, K_excess < 0) - Platykurtic
•Lighter tails than the normal distribution, meaning fewer extreme values.
•The distribution is "flatter" in the center with shorter tails.



4.Fourth Moment (Kurtosis) – Tailedness of Distribution
$Kurt(X) = E[(X - \mu)^4] / \sigma^4$
Measures how heavy or light the tails of the distribution are compared to a normal distribution.

A numerical example

A *n = 10* dataset
X=[10,15,20,25,30,35,40,45,50,55]

Let’s compute the kurtosis:

$$\text{Kurt}(X) = \frac{\sum(X_i - \mu)^4}{n \cdot \sigma^4}$$

A numerical example

A *n = 10* dataset
X=[10,15,20,25,30,35,40,45,50,55]

Let's compute the kurtosis:

$$\text{Kurt}(X) = \frac{\sum(X_i - \mu)^4}{n \cdot \sigma^4}$$

$$(10 - 32.5)^4 = 3013025.5625$$

$$(15 - 32.5)^4 = 938906.25$$

$$(20 - 32.5)^4 = 244140.625$$

$$(25 - 32.5)^4 = 31640.625$$

$$(30 - 32.5)^4 = 39.0625$$

$$(35 - 32.5)^4 = 39.0625$$

$$(40 - 32.5)^4 = 31640.625$$

$$(45 - 32.5)^4 = 244140.625$$

$$(50 - 32.5)^4 = 938906.25$$

$$(55 - 32.5)^4 = 3013025.5625$$

A numerical example

A *n = 10* dataset
X=[10,15,20,25,30,35,40,45,50,55]

Let's compute the kurtosis:

$$\text{Kurt}(X) = \frac{\sum(X_i - \mu)^4}{n \cdot \sigma^4}$$

The kurtosis is less than 3, so the distribution is platykurtic (fewer "tails" compared to the normal distribution).

$$(10 - 32.5)^4 = 3013025.5625$$

$$(15 - 32.5)^4 = 938906.25$$

$$(20 - 32.5)^4 = 244140.625$$

$$(25 - 32.5)^4 = 31640.625$$

$$(30 - 32.5)^4 = 39.0625$$

$$(35 - 32.5)^4 = 39.0625$$

$$(40 - 32.5)^4 = 31640.625$$

$$(45 - 32.5)^4 = 244140.625$$

$$(50 - 32.5)^4 = 938906.25$$

$$(55 - 32.5)^4 = 3013025.5625$$

$$3013025.56 + 938906.25 + 244140.62 + 31640.62 + 39.06 + 39.06 + 31640.62 + 244140.62 + 938906.25 + 3013025.56 = 8282825$$

$$\text{Kurt}(X) = \frac{8282825}{10 \cdot (206.25)^2} = \frac{8282825}{42514.0625} = 1.8$$

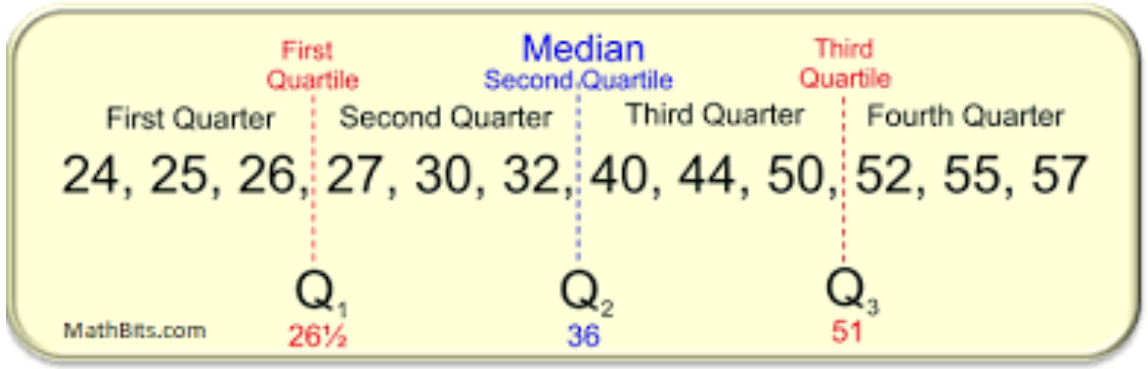
Recap – Quartiles

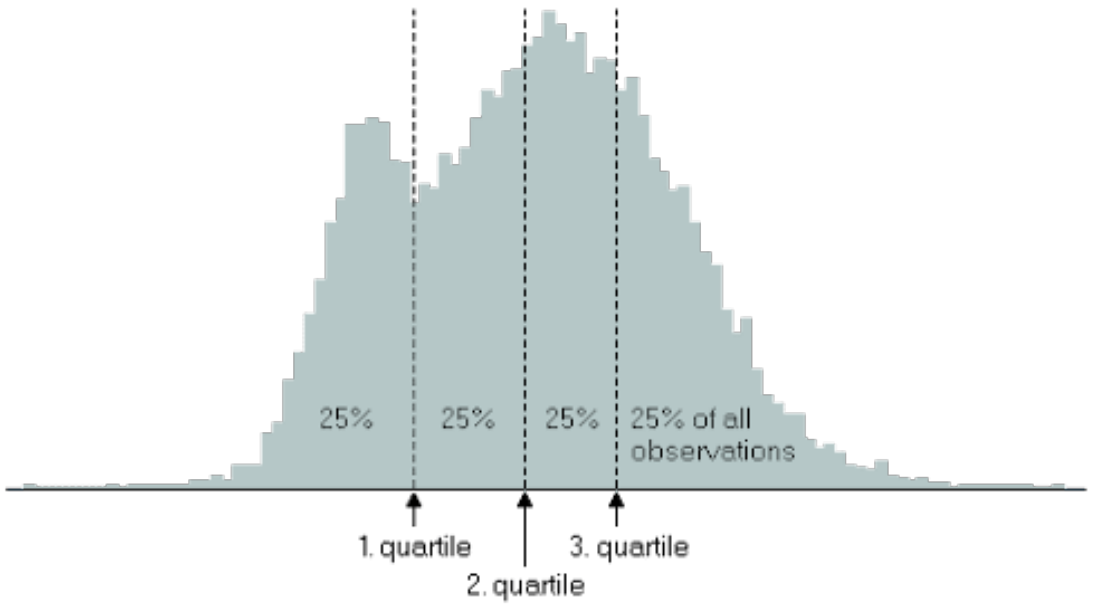
Quartiles divide a dataset into **four equal parts**, helping to understand the distribution and spread of the data.

Quartile Definitions:
- *Q1* (First Quartile, 25%) – The value below which 25% of the data falls.
- *Q2* (Second Quartile, 50%) – The median, the value that splits the data into two equal halves.
- *Q3* (Third Quartile, 75%) – The value below which 75% of the data falls.
- Interquartile Range (*IQR*) – The range between Q1 and *Q3*, measuring the spread of the middle 50% of the data:

$$IQR = Q3 - Q1$$

The median is used many times instead of the mean, as it is a robust quantity w.r.t. 'outliers' (strange data)





Common convention (for the theoric part of the exam): no interpolation!

Procedure:

- Order the array
- Compute the exact percentage
- Always take the closest number which percentage is AT LEAST 25%/50%/75% for Q1/Q2/Q3 respectively

Common convention (for the theoric part of the exam): no interpolation!

Procedure:

- Order the array
- Compute the exact percentage
- Always take the closest number which percentage is AT LEAST 25%/50%/75% for Q1/Q2/Q3 respectively

$$[10, 15, 20, 25, 30, 35, 40, 45, 50, 55]$$

$$[10\%, 20\%, 30\%, 40\%, 50\%, 60\%, 70\%, 80\%, 90\%, 100\%]$$

Common convention (for the theoric part of the exam): no interpolation!

Procedure:

- Order the array
- Compute the exact percentage
- Always take the closest number which percentage is AT LEAST 25%/50%/75% for Q1/Q2/Q3 respectively

$$[10, 15, 20, 25, 30, 35, 40, 45, 50, 55]$$

$$[10\%, 20\%, 30\%, 40\%, 50\%, 60\%, 70\%, 80\%, 90\%, 100\%]$$

Common convention (for the theoric part of the exam): no interpolation!

Procedure:

- Order the array
- Compute the exact percentage
- Always take the closest number which percentage is AT LEAST 25%/50%/75% for Q1/Q2/Q3 respectively

$$[5, 10, 15, 20, 25, 30, 35, 40, 45, 50, 55]$$

$$[9.09\%, 18.18\%, 27.27\%, 36.36\%, 45.45\%, 54.55\%, 63.64\%, 72.73\%, 81.82\%, 90.91\%, 100\%]$$

Common convention (for the theoric part of the exam): no interpolation!

Procedure:

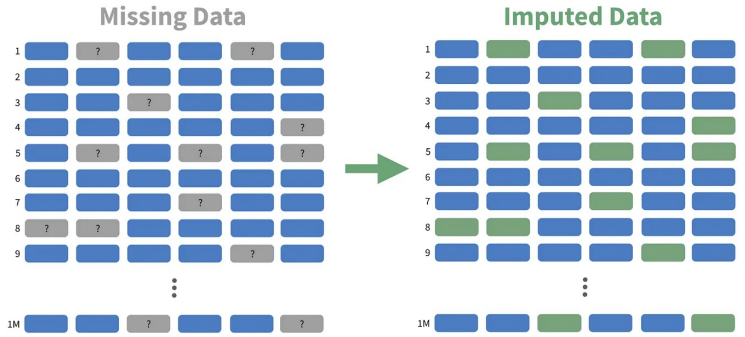
- Order the array
- Compute the exact percentage
- Always take the closest number which percentage is AT LEAST 25%/50%/75% for Q1/Q2/Q3 respectively

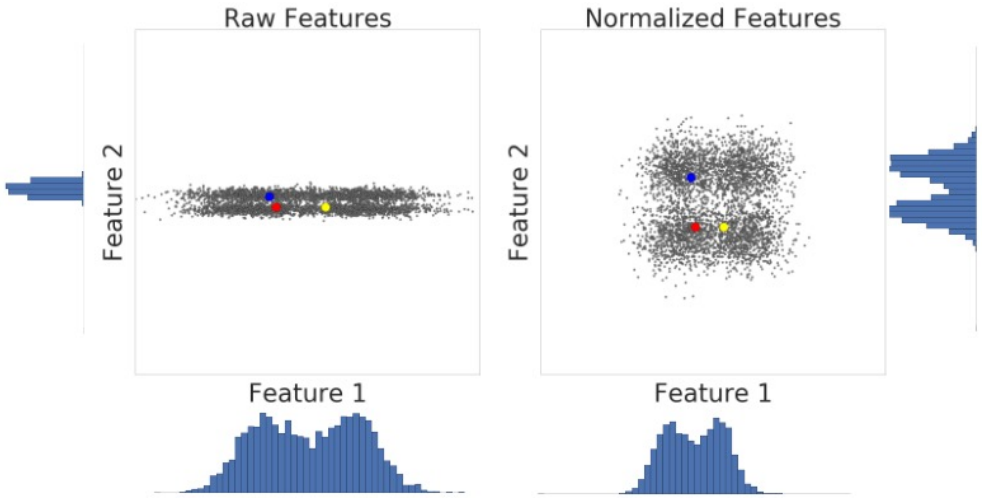
$$[5, 10, 15, 20, 25, 30, 35, 40, 45, 50, 55]$$

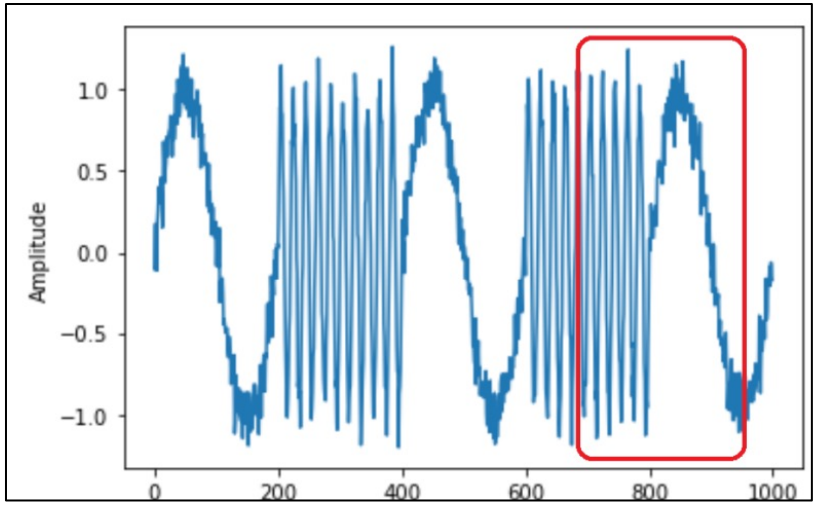
$$[9.09\%, 18.18\%, 27.27\%, 36.36\%, 45.45\%, 54.55\%, 63.64\%, 72.73\%, 81.82\%, 90.91\%, 100\%]$$

Recap - Statistical quantities in ML

- Understand a dataset
- Correct a dataset (filling missing data)
- Feature Engineering (extract quantities for example from a time-series sensor)
- Reduce the dimensionality of a dataset (by excluding variables with 0 variance)
- Optimize a dataset for ML (standardization)







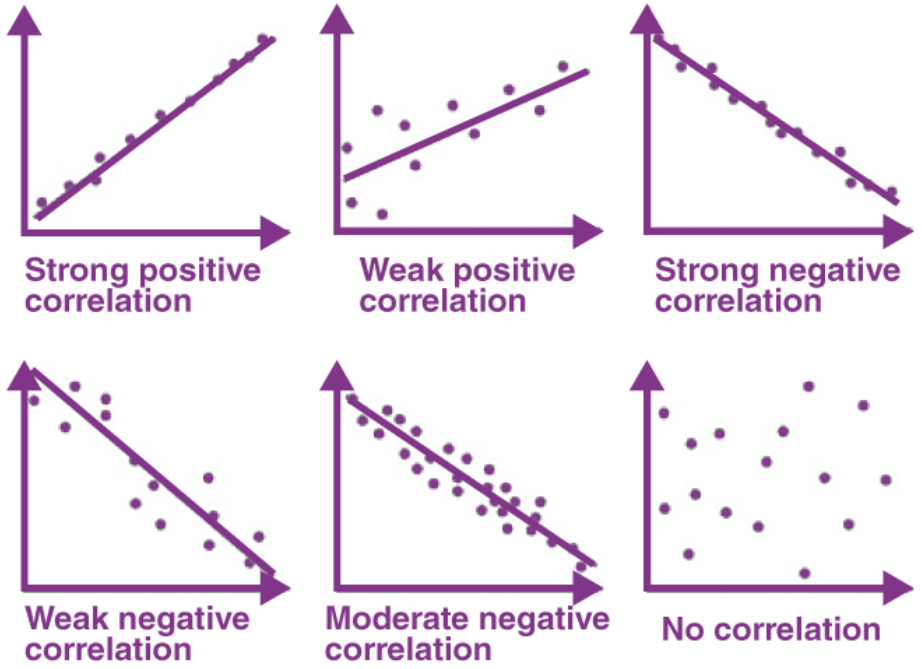
Correlation

Correlation

Correlation is a statistical measure that expresses the degree to which two variables move in relation to each other. It quantifies the strength and direction of their relationship.

Correlation values range from **-1 to 1**.

- **+1**: Perfect positive correlation (when one variable increases, the other increases proportionally).
- **0**: No correlation (no relationship between the variables).
- **-1**: Perfect negative correlation (when one variable increases, the other decreases proportionally).



Pearson Correlation

The **Pearson correlation coefficient** (denoted as *r*) measures the **linear relationship** between two variables *X* and *Y*. It quantifies how strongly and in which direction two variables are related.

$$r = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum(X_i - \bar{X})^2} \cdot \sqrt{\sum(Y_i - \bar{Y})^2}}$$

- The numerator represents the covariance between *X* and *Y*.
- The denominator is the product of the standard deviations of *X* and *Y*.

Pearson Correlation: a numerical example #01

$$X = [1, 0, 2], \quad Y = [2, -1, 5]$$

$$r = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum (X_i - \bar{X})^2} \cdot \sqrt{\sum (Y_i - \bar{Y})^2}}$$

Pearson Correlation: a numerical example #01

$$X = [1, 0, 2], \quad Y = [2, -1, 5]$$

$$r = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum (X_i - \bar{X})^2} \cdot \sqrt{\sum (Y_i - \bar{Y})^2}}$$

$$\bar{X} = \frac{1 + 0 + 2}{3} = \frac{3}{3} = 1$$

$$\bar{Y} = \frac{2 + (-1) + 5}{3} = \frac{6}{3} = 2$$

Pearson Correlation: a numerical example #01

$$X = [1, 0, 2], \quad Y = [2, -1, 5]$$

$$r = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum (X_i - \bar{X})^2} \cdot \sqrt{\sum (Y_i - \bar{Y})^2}}$$

$$\bar{X} = \frac{1 + 0 + 2}{3} = \frac{3}{3} = 1$$

$$\bar{Y} = \frac{2 + (-1) + 5}{3} = \frac{6}{3} = 2$$

X_i	$X_i - \bar{X}$	Y_i	$Y_i - \bar{Y}$
1	$1 - 1 = 0$	2	$2 - 2 = 0$
0	$0 - 1 = -1$	-1	$-1 - 2 = -3$
2	$2 - 1 = 1$	5	$5 - 2 = 3$

$$\sum (X_i - \bar{X})(Y_i - \bar{Y}) = 0 + 3 + 3 = 6$$

Pearson Correlation: a numerical example #01

$$X = [1, 0, 2], \quad Y = [2, -1, 5]$$

$$\bar{X} = \frac{1 + 0 + 2}{3} = \frac{3}{3} = 1$$

$$\bar{Y} = \frac{2 + (-1) + 5}{3} = \frac{6}{3} = 2$$

X_i	$X_i - \bar{X}$	Y_i	$Y_i - \bar{Y}$
1	$1 - 1 = 0$	2	$2 - 2 = 0$
0	$0 - 1 = -1$	-1	$-1 - 2 = -3$
2	$2 - 1 = 1$	5	$5 - 2 = 3$

$$\sum (X_i - \bar{X})(Y_i - \bar{Y}) = 0 + 3 + 3 = 6$$

$$r = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum (X_i - \bar{X})^2} \cdot \sqrt{\sum (Y_i - \bar{Y})^2}}$$

$$\sum (X_i - \bar{X})^2 = (0)^2 + (-1)^2 + (1)^2 = 0 + 1 + 1 = 2$$

$$\sqrt{2} \approx 1.41$$

$$\sum (Y_i - \bar{Y})^2 = (0)^2 + (-3)^2 + (3)^2$$

$$= 0 + 9 + 9 = 18$$

$$\sqrt{18} = 4.24$$

Pearson Correlation: a numerical example #01

$$X = [1, 0, 2], \quad Y = [2, -1, 5]$$

$$\bar{X} = \frac{1 + 0 + 2}{3} = \frac{3}{3} = 1$$

$$\bar{Y} = \frac{2 + (-1) + 5}{3} = \frac{6}{3} = 2$$

X_i	$X_i - \bar{X}$	Y_i	$Y_i - \bar{Y}$
1	$1 - 1 = 0$	2	$2 - 2 = 0$
0	$0 - 1 = -1$	-1	$-1 - 2 = -3$
2	$2 - 1 = 1$	5	$5 - 2 = 3$

$$\sum (X_i - \bar{X})(Y_i - \bar{Y}) = 0 + 3 + 3 = 6$$

$$r = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum (X_i - \bar{X})^2} \cdot \sqrt{\sum (Y_i - \bar{Y})^2}}$$

$$\sum (X_i - \bar{X})^2 = (0)^2 + (-1)^2 + (1)^2 = 0 + 1 + 1 = 2$$

$$\sqrt{2} \approx 1.41$$

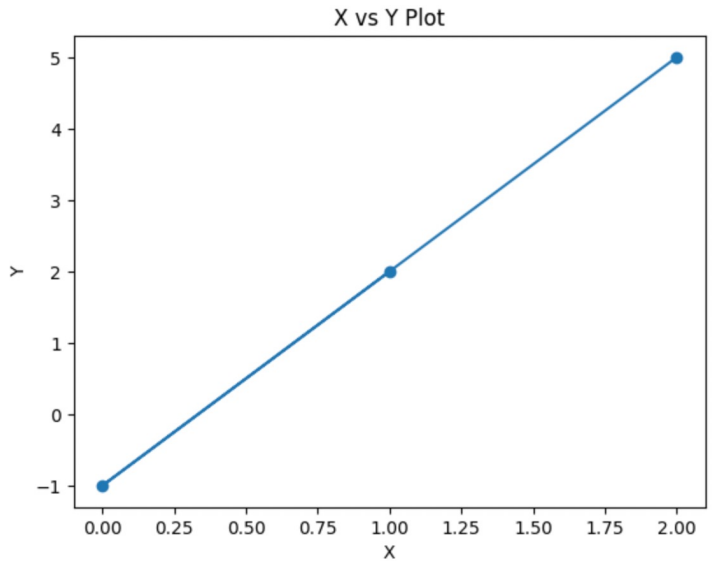
$$\sum (Y_i - \bar{Y})^2 = (0)^2 + (-3)^2 + (3)^2$$

$$= 0 + 9 + 9 = 18$$

$$\sqrt{18} = 4.24$$

$$r = \frac{6}{(1.41 \times 4.24)}$$

$$r = \frac{6}{5.98} \approx 1.0$$



Pearson Correlation: a numerical example #02

$$X = [0, 1, 2, 5], \quad Y = [4, 1, 3, 0]$$

Pearson Correlation: a numerical example #02

$$X = [0, 1, 2, 5], \quad Y = [4, 1, 3, 0]$$

$$\bar{X} = \frac{0 + 1 + 2 + 5}{4} = \frac{8}{4} = 2$$

$$\bar{Y} = \frac{4 + 1 + 3 + 0}{4} = \frac{8}{4} = 2$$

X_i	$X_i - \bar{X}$	Y_i	$Y_i - \bar{Y}$
0	$0 - 2 = -2$	4	$4 - 2 = 2$
1	$1 - 2 = -1$	1	$1 - 2 = -1$
2	$2 - 2 = 0$	3	$3 - 2 = 1$
5	$5 - 2 = 3$	0	$0 - 2 = -2$

$$\sum (X_i - \bar{X})(Y_i - \bar{Y}) = -4 + 1 + 0 - 6 = -9$$

Pearson Correlation: a numerical example #02

$$X = [0, 1, 2, 5], \quad Y = [4, 1, 3, 0]$$

$$\bar{X} = \frac{0 + 1 + 2 + 5}{4} = \frac{8}{4} = 2$$

$$\bar{Y} = \frac{4 + 1 + 3 + 0}{4} = \frac{8}{4} = 2$$

X_i	$X_i - \bar{X}$	Y_i	$Y_i - \bar{Y}$
0	$0 - 2 = -2$	4	$4 - 2 = 2$
1	$1 - 2 = -1$	1	$1 - 2 = -1$
2	$2 - 2 = 0$	3	$3 - 2 = 1$
5	$5 - 2 = 3$	0	$0 - 2 = -2$

$$\sum (X_i - \bar{X})(Y_i - \bar{Y}) = -4 + 1 + 0 - 6 = -9$$

$$\sum (X_i - \bar{X})^2 = (-2)^2 + (-1)^2 + (0)^2 + (3)^2$$

$$= 4 + 1 + 0 + 9 = 14$$

$$\sqrt{14} \approx 3.74$$

$$\sum (Y_i - \bar{Y})^2 = (2)^2 + (-1)^2 + (1)^2 + (-2)^2$$

$$= 4 + 1 + 1 + 4 = 10$$

$$\sqrt{10} \approx 3.16$$

Pearson Correlation: a numerical example #02

$$X = [0, 1, 2, 5], \quad Y = [4, 1, 3, 0]$$

$$\bar{X} = \frac{0 + 1 + 2 + 5}{4} = \frac{8}{4} = 2$$

$$\bar{Y} = \frac{4 + 1 + 3 + 0}{4} = \frac{8}{4} = 2$$

X_i	$X_i - \bar{X}$	Y_i	$Y_i - \bar{Y}$
0	$0 - 2 = -2$	4	$4 - 2 = 2$
1	$1 - 2 = -1$	1	$1 - 2 = -1$
2	$2 - 2 = 0$	3	$3 - 2 = 1$
5	$5 - 2 = 3$	0	$0 - 2 = -2$

$$\sum (X_i - \bar{X})(Y_i - \bar{Y}) = -4 + 1 + 0 - 6 = -9$$

$$\sum (X_i - \bar{X})^2 = (-2)^2 + (-1)^2 + (0)^2 + (3)^2$$

$$= 4 + 1 + 0 + 9 = 14$$

$$\sqrt{14} \approx 3.74$$

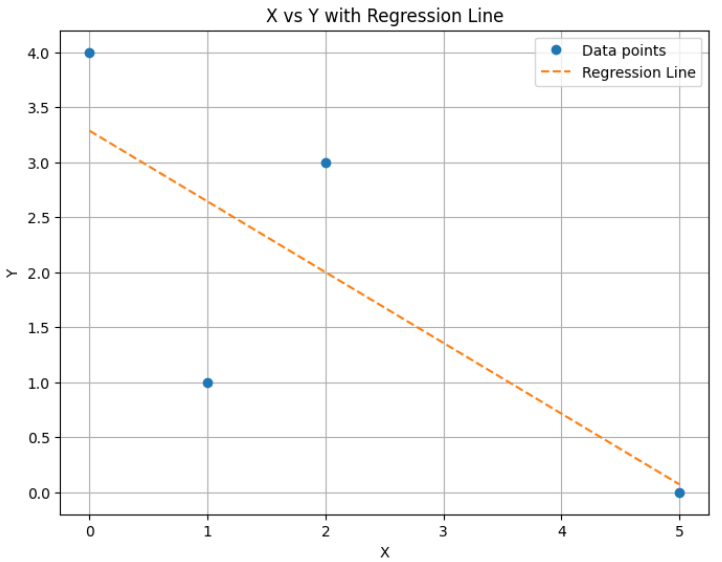
$$\sum (Y_i - \bar{Y})^2 = (2)^2 + (-1)^2 + (1)^2 + (-2)^2$$

$$= 4 + 1 + 1 + 4 = 10$$

$$\sqrt{10} \approx 3.16$$

$$r = \frac{-9}{(3.74 \times 3.16)}$$

$$r = \frac{-9}{11.83} \approx -0.761$$



Other Correlations

There are other correlations, for example:

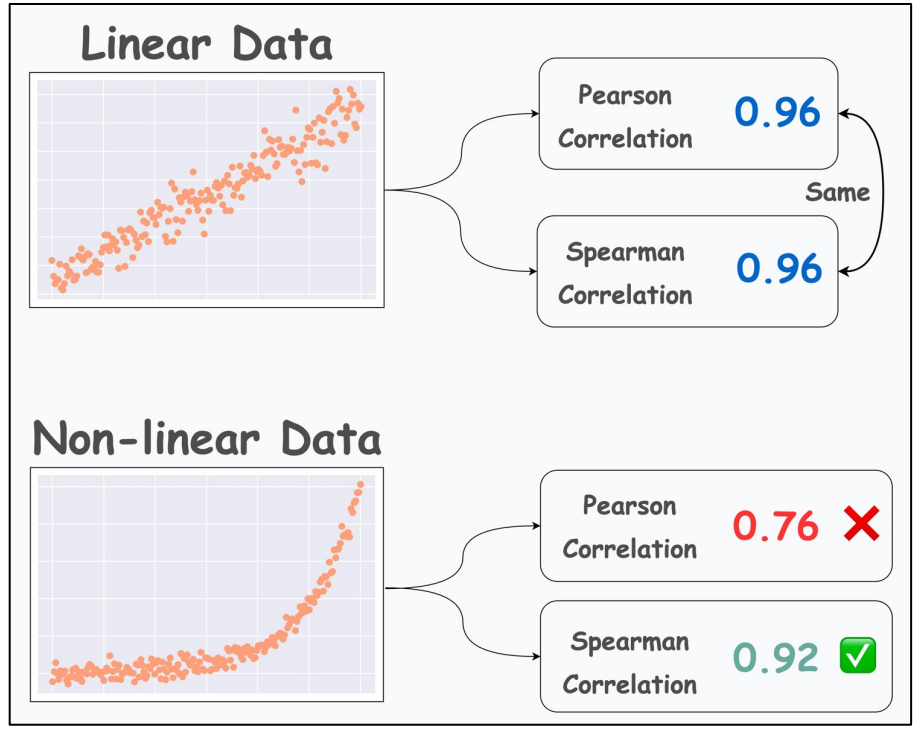
Spearman's Rank Correlation (ρ)

- Measures the monotonic relationship between two variables (not necessarily linear).
- Instead of using actual values, it ranks the data and calculates Pearson's correlation on the ranks.

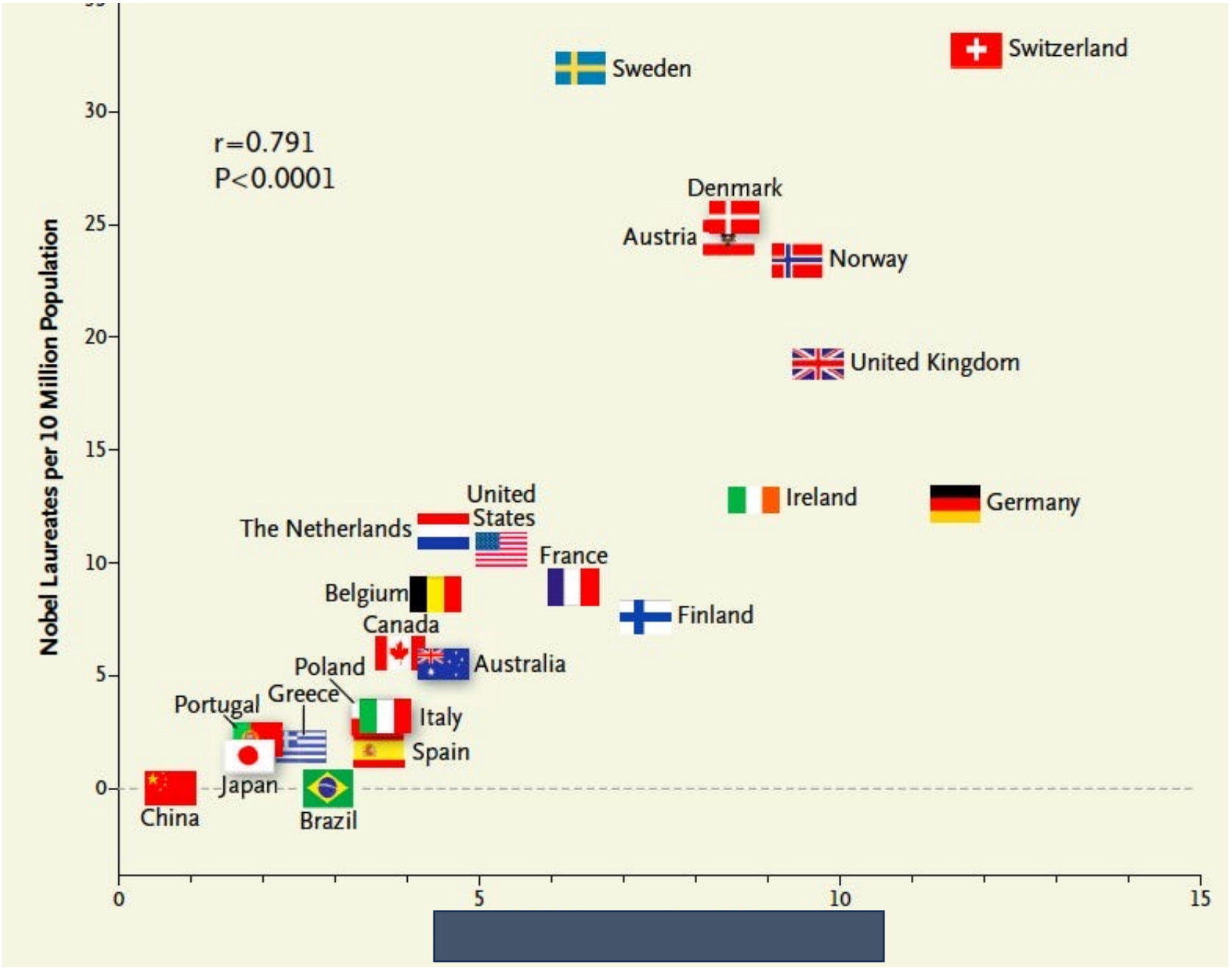
Kendall's Tau (τ)

- Measures the degree of agreement between two rankings.
- Based on concordant and discordant pairs rather than numerical differences.

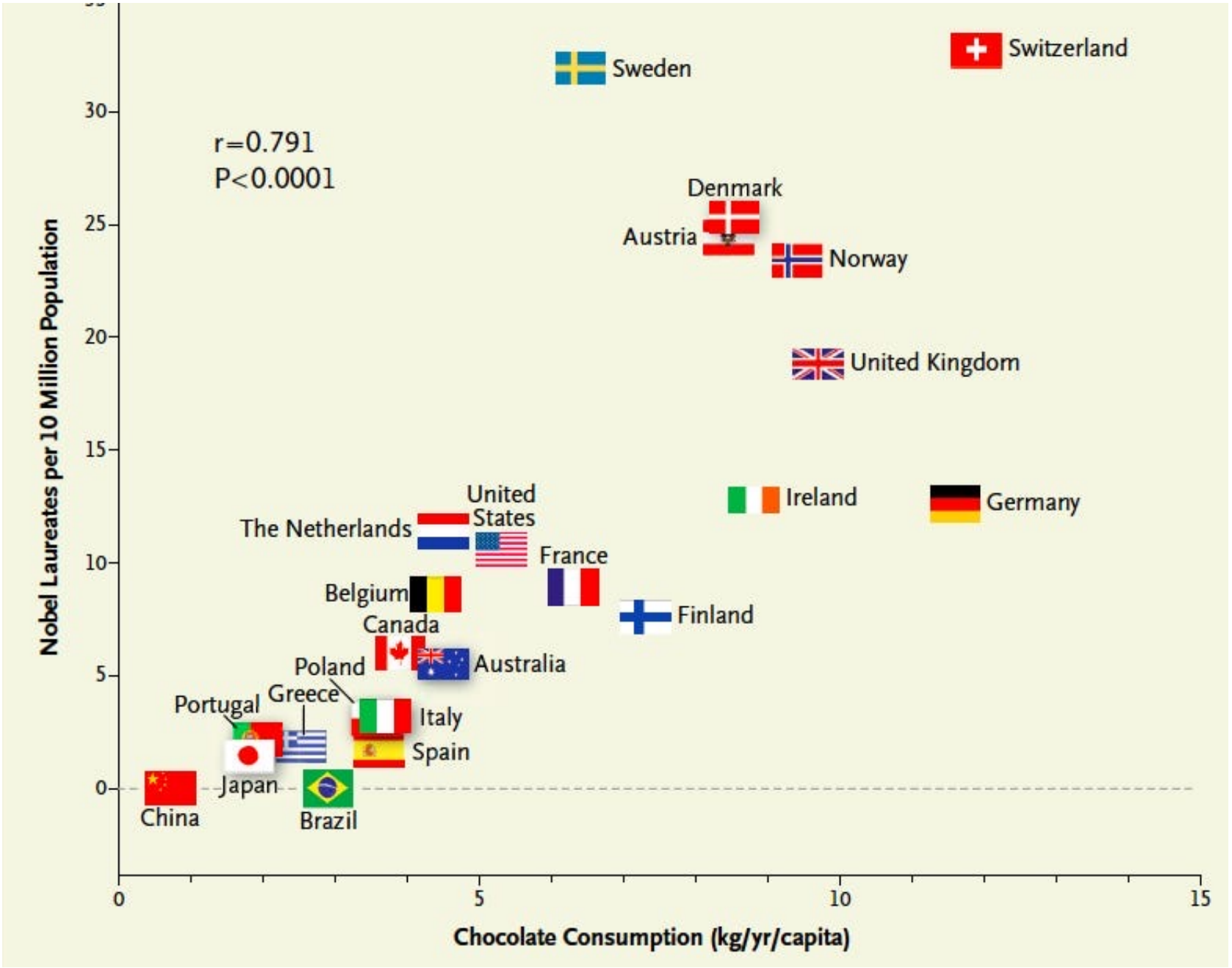
In this course, correlation = Pearson correlation!



A real example of strong positive correlation

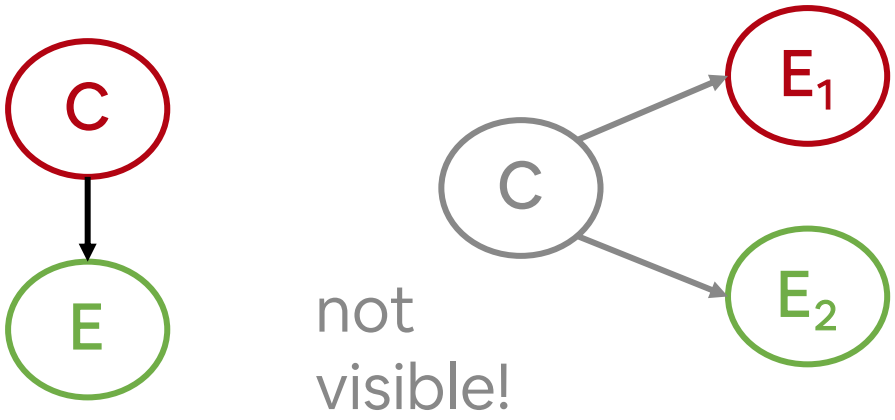


A real example of strong positive correlation



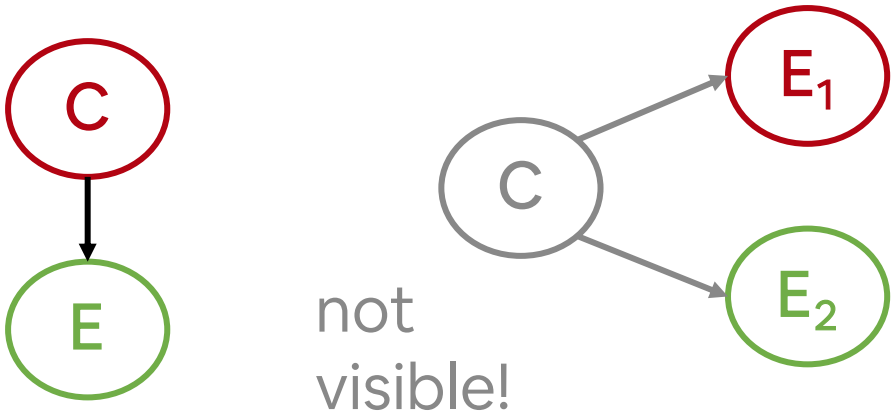
Common Misconception: Correlation does not imply causation

We think that - The truth may be:



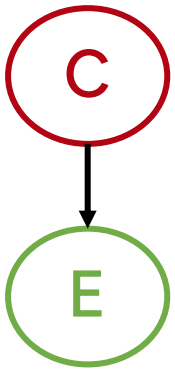
Common Misconception: Correlation does not imply causation

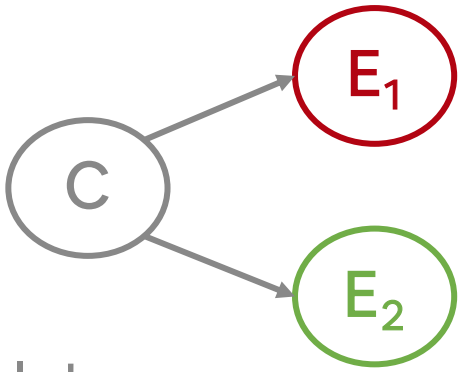
We think that - The truth may be:



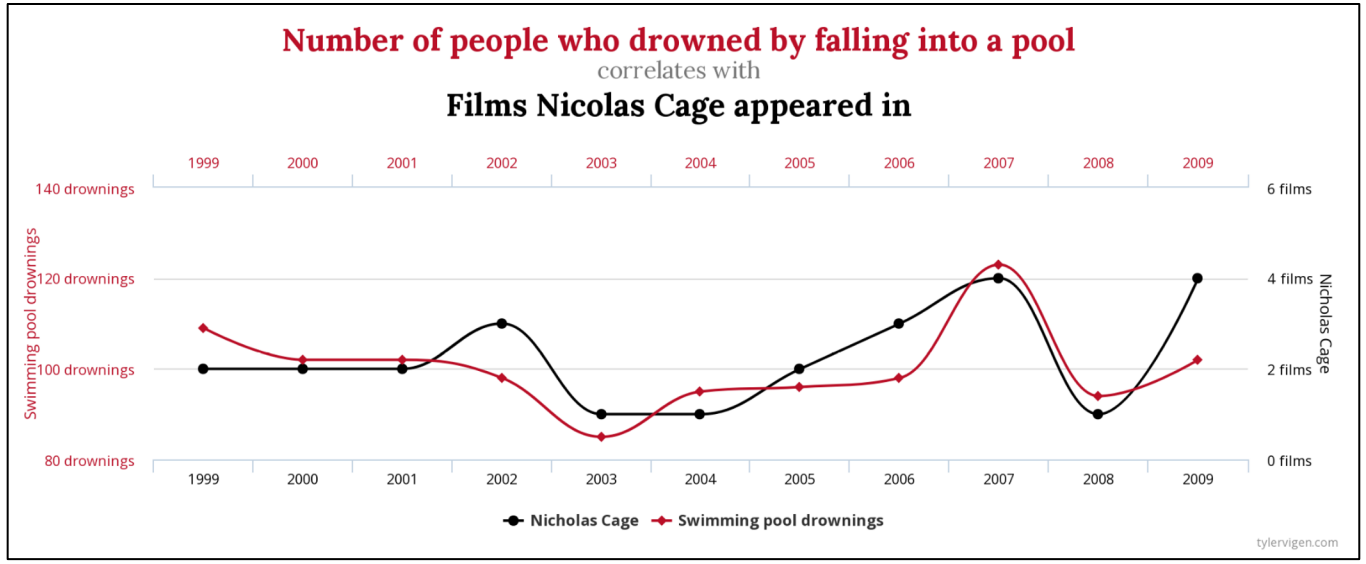
Common Misconception: Correlation does not imply causation

We think that - The truth may be:



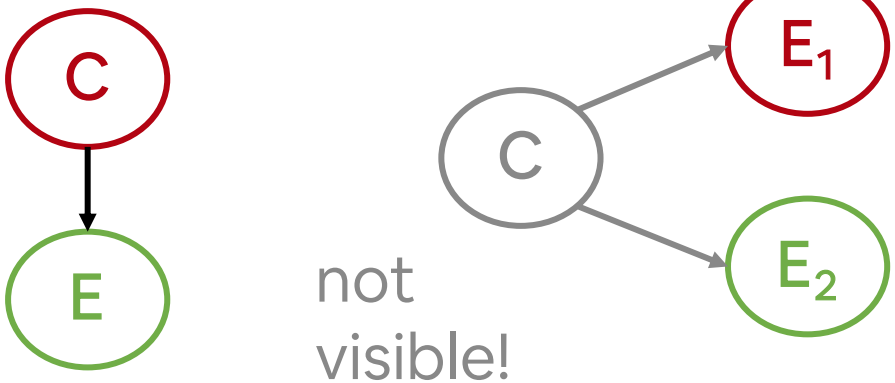


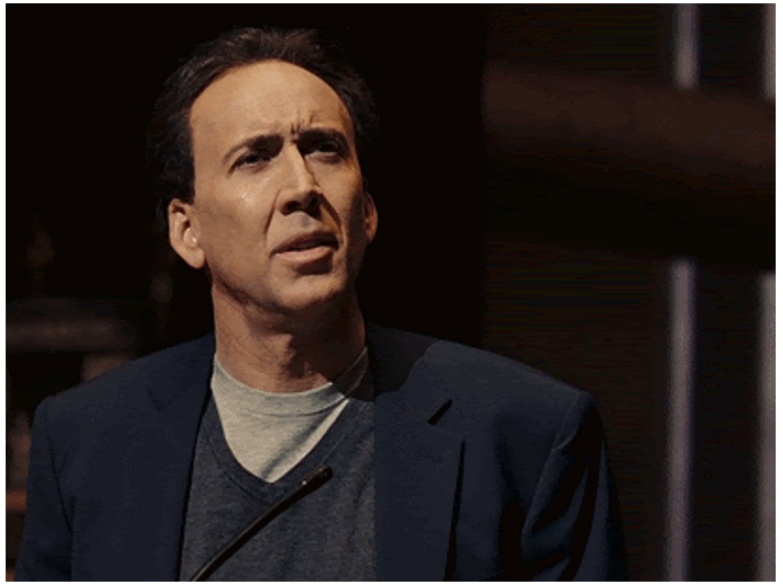
not visible!

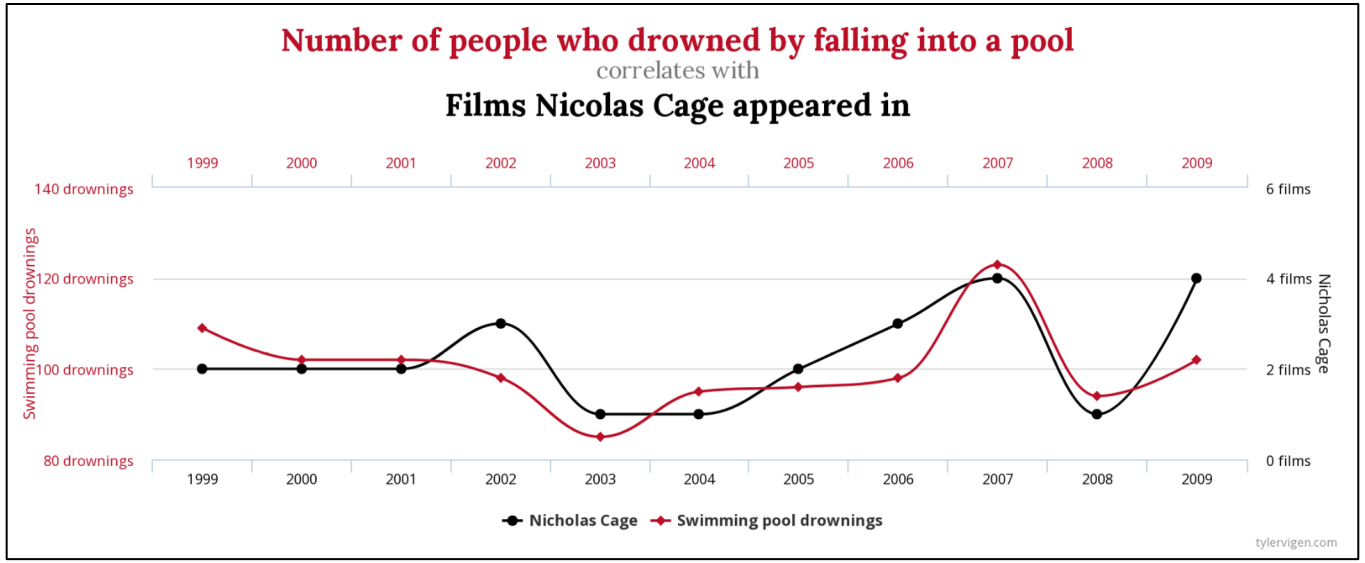


Common Misconception: Correlation does not imply causation

We think that - The truth may be:



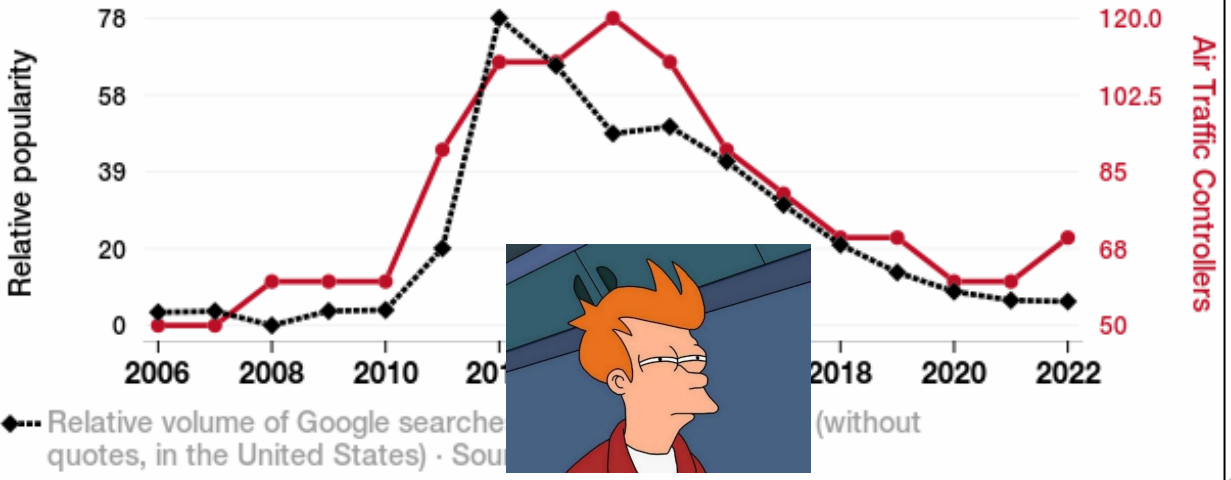




Popularity of the 'not sure if' meme

correlates with

The number of air traffic controllers in Montana



Relative volume of Google searche (without quotes, in the United States) · Sou

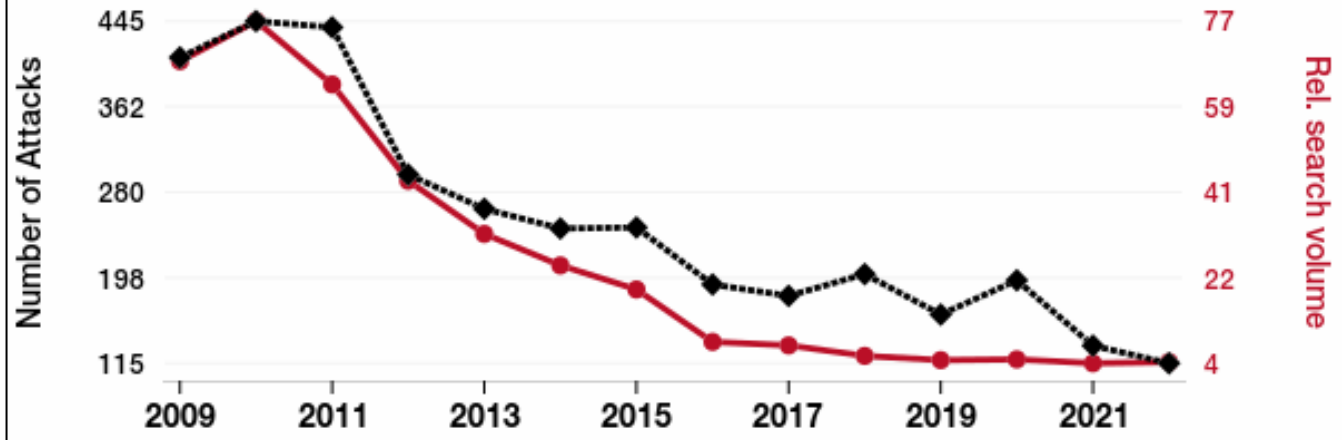
BLS estimate of air traffic controllers in Montana · Source: Bureau of Larbor Statistics

2006-2022, r=0.917, r²=0.842, p<0.01 · tylervigen.com/spurious/correlation/5957

Pirate attacks globally

correlates with

Google searches for 'download firefox'



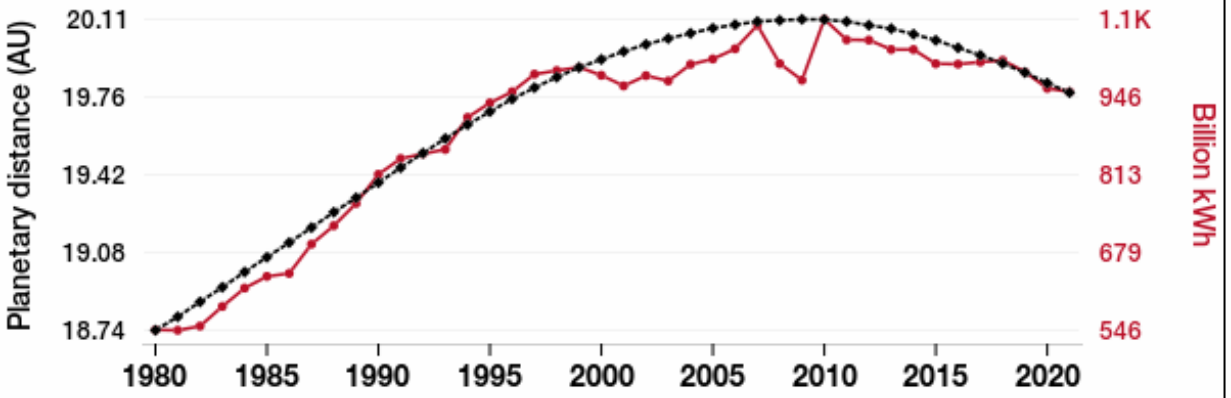
Global Pirate Attack Count · Source: Statista

Relative volume of Google searches for 'download firefox' (Worldwide, without quotes) · Source: Google Trends

The distance between Uranus and the moon

correlates with

Electricity generation in Japan

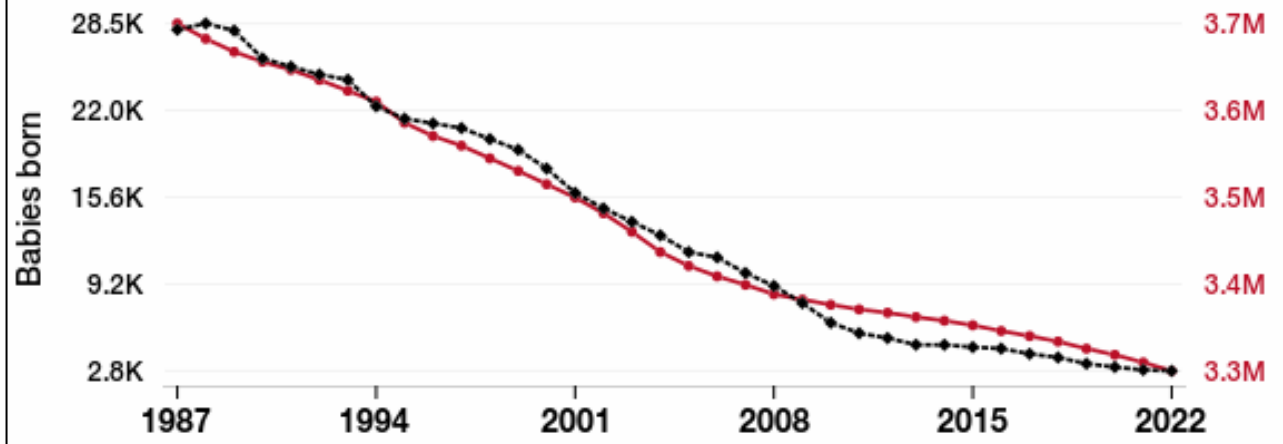


The average distance between Uranus and the moon as measured on the first day of each month · Source: Caclculated using Astropy

Popularity of the first name Sarah

correlates with

Remaining Forest Cover in the Brazilian Amazon

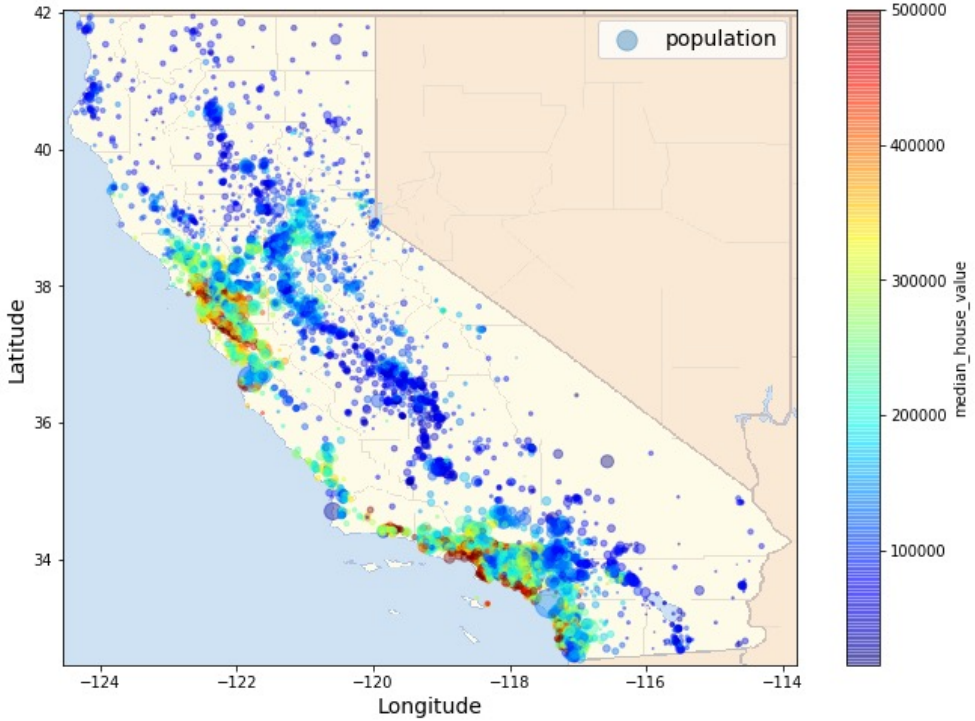


Babies of all sexes born in the US named Sarah · Source: US Social Security

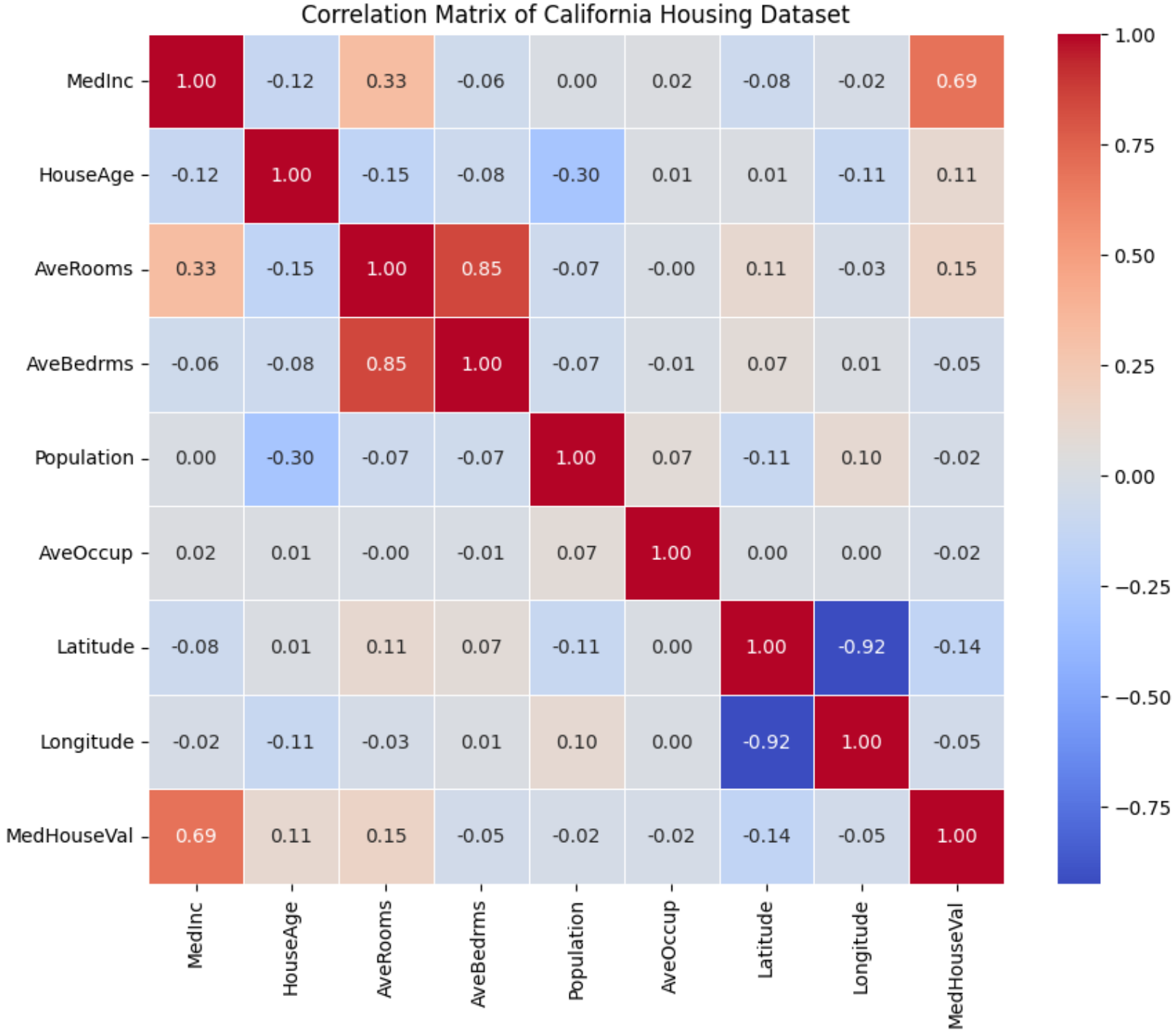
Benefit of correlation in ML: understanding ‘important’ variables

- In supervised tasks we can understand which features are relevant

- Ex. California Housing



Variable	Description
MedInc	Median household income in the area (in tens of thousands of dollars).
HouseAge	Median age of houses in the area (in years).
AveRooms	Average number of rooms per dwelling in the area.
AveBedrms	Average number of bedrooms per dwelling in the area.
Population	Total population in the area.
AveOccup	Average number of people per household in the area.
Latitude	Geographic latitude of the area.
Longitude	Geographic longitude of the area.
MedHouseVal	Median house value in the area (in hundreds of thousands of dollars). **This is the target variable** in the dataset.


Correlation Matrix of California Housing Dataset
MedInc
HouseAge
AveRooms
AveBedrms
Population
AveOccup
Latitude
Longitude
MedHouseVal
1.00 -0.12 0.33 -0.06 0.00 0.02 -0.08 -0.02 0.69
-0.12 1.00 -0.15 -0.08 -0.30 0.01 0.01 -0.11 0.11
0.33 -0.15 1.00 0.85 -0.07 -0.00 0.11 -0.03 0.15
-0.06 -0.08 0.85 1.00 -0.07 -0.01 0.07 0.01 -0.05
0.00 -0.30 -0.07 -0.07 1.00 0.07 -0.11 0.10 -0.02
0.02 0.01 -0.00 -0.01 0.07 1.00 0.00 0.00 -0.02
-0.08 0.01 0.11 0.07 -0.11 0.00 1.00 -0.92 -0.14
-0.02 -0.11 -0.03 0.01 0.10 0.00 -0.92 1.00 -0.05
0.69 0.11 0.15 -0.05 -0.02 -0.02 -0.14 -0.05 1.00
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75

Benefit of correlation in ML: reduce dataset size!

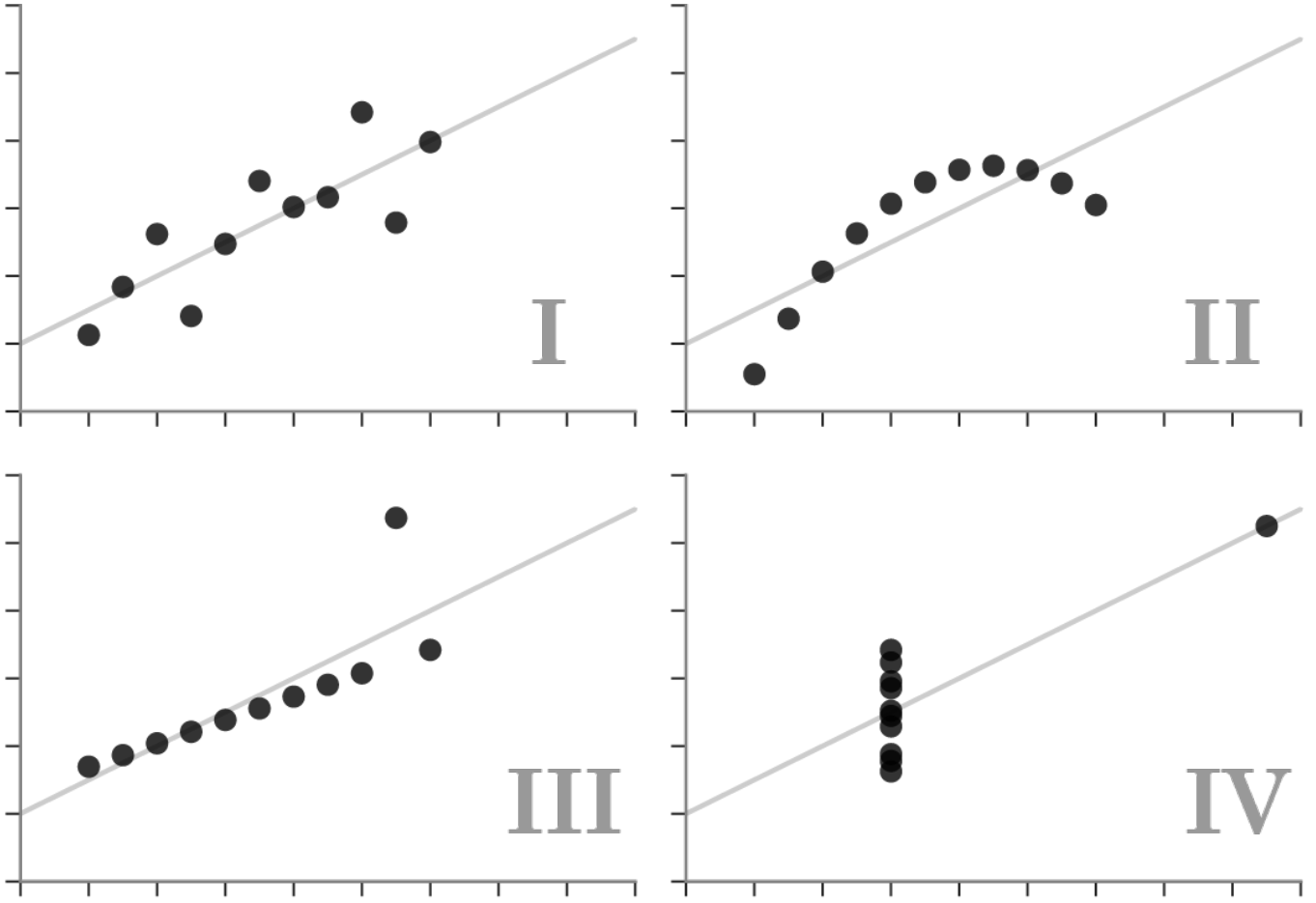
- If two variables have really high correlation (in absolute value, ie. also really high negative correlation) they are containing the same informative content!
- For the task, just one of the two variables should be kept for efficiency and for better ‘engineering’ of a productive solutions

	Feature_1	Feature_2	Constant_Var
1	54.9671415301123 3	23.636499344142 013	100
2	48 5698 15	23 90197 67	100
3	56.4 53 6 92	26.0 48 0 753	100
4	65.23 80 26	30.49 44 757	100
5	47.6584 2766 644	28.638 2842 315	100
6	47.658 508 2	25.824 960 837	100
7	65.71 89 1	32.21 7 59	100
8	57 729 0 9	22 721 83	100
9	45.305256140650 48	25.842892970704 362	100

Now we have statistical moments and correlation to understand a dataset... is that enough?

We should still be careful about summary indicators

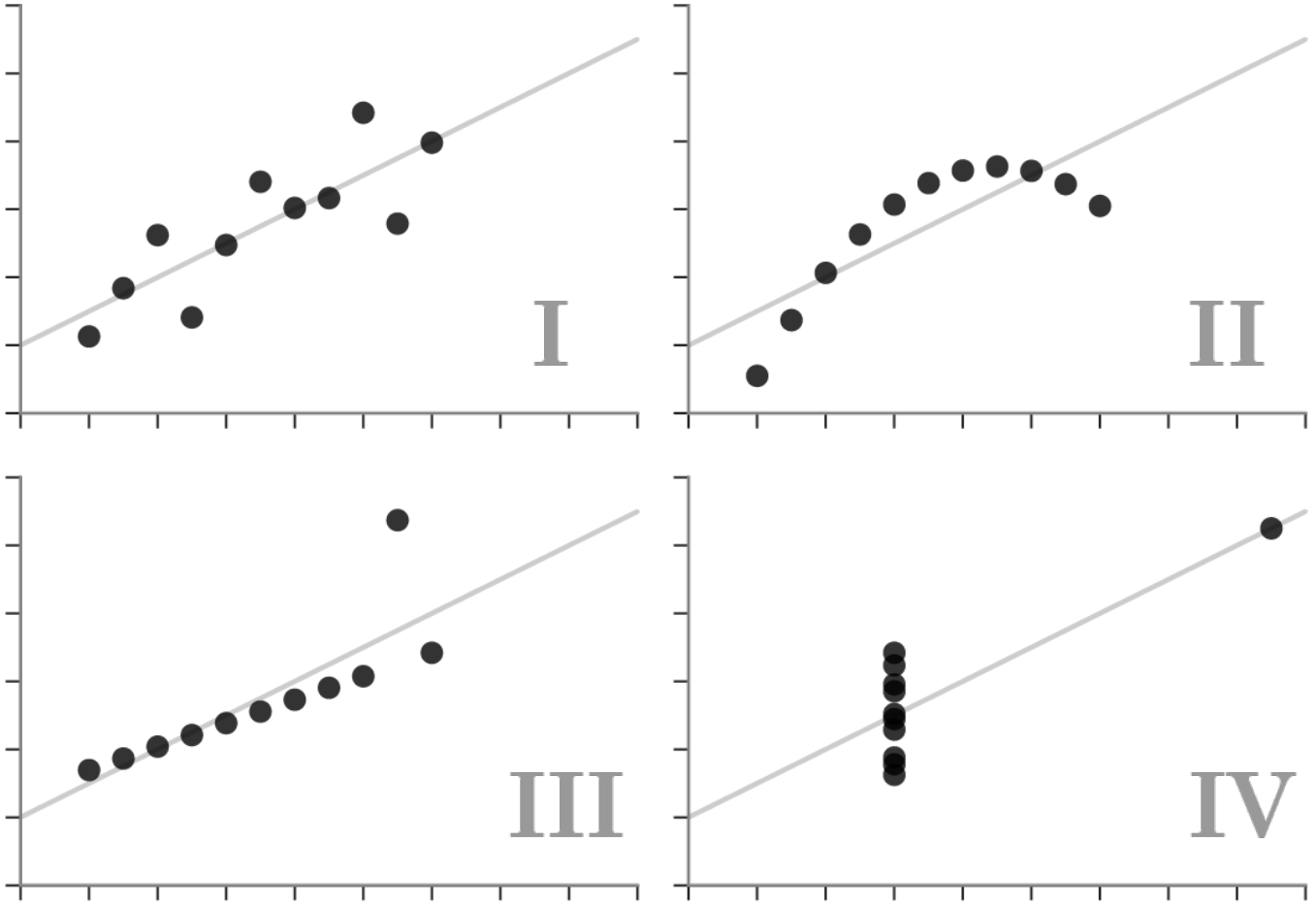
Anscombe's Quartet



We should still be careful about summary indicators

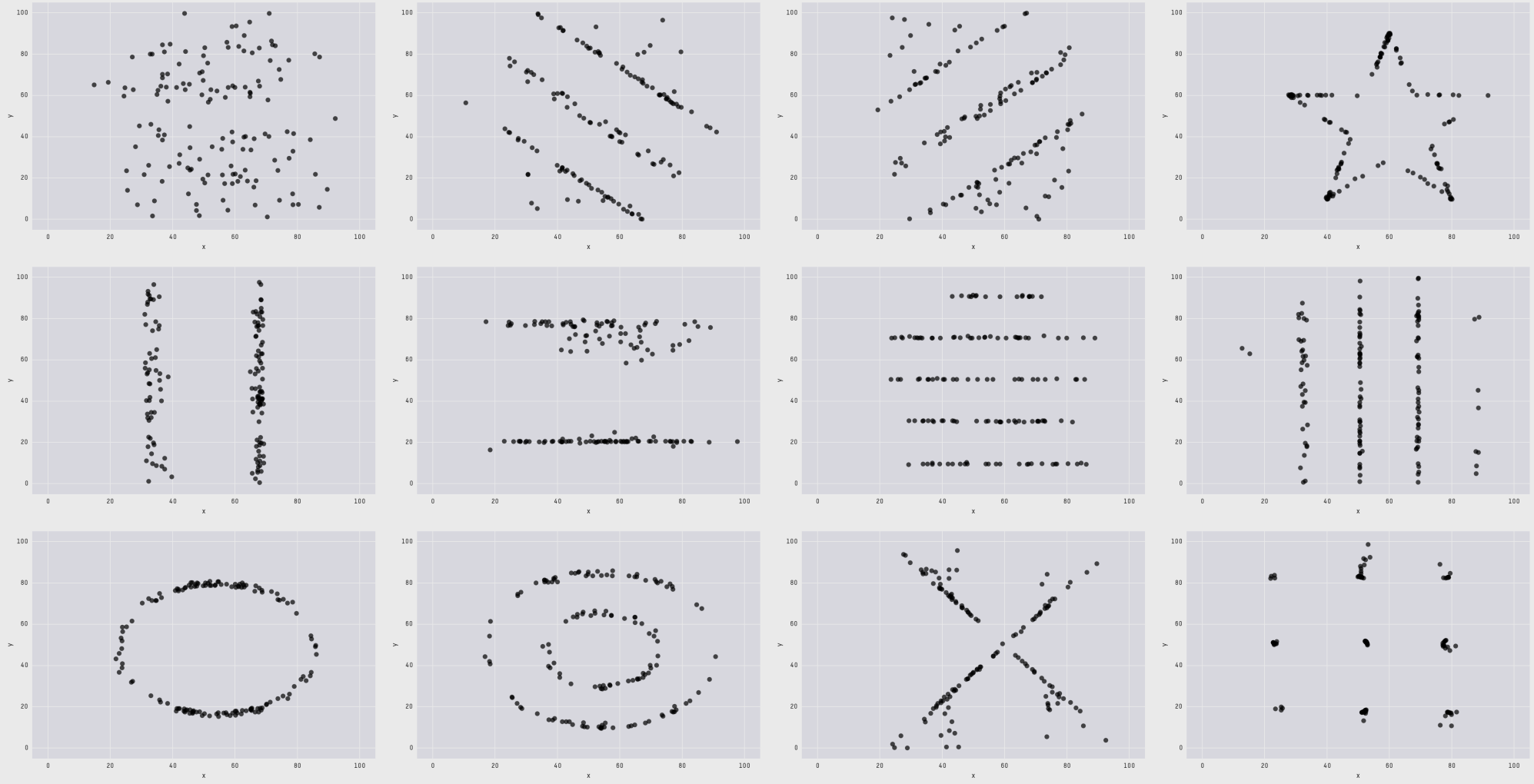
Anscombe's Quartet

Each dataset has the same summary statistics (mean, standard deviation, correlation), and the datasets are *clearly different*, and *visually distinct*.

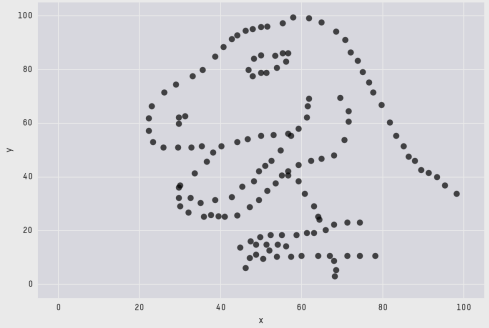


The Datasauros Dozens

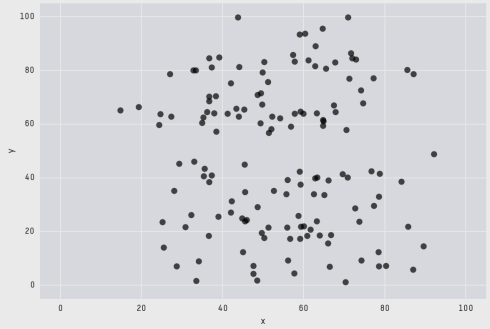
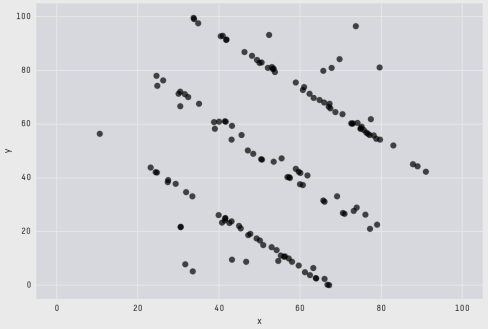
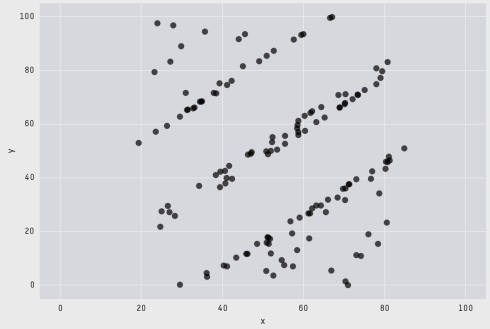
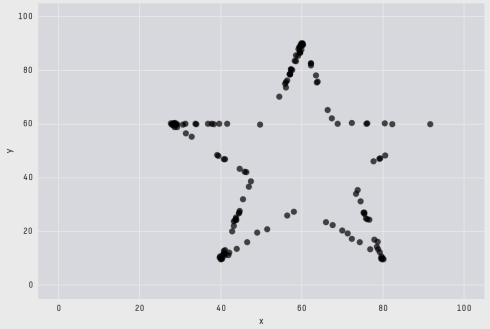
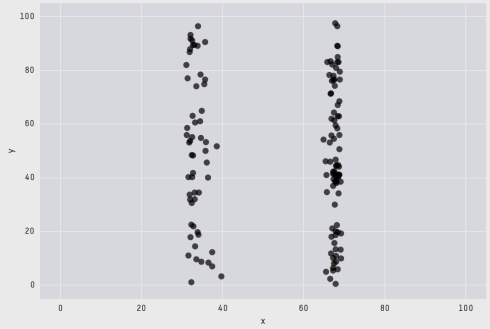
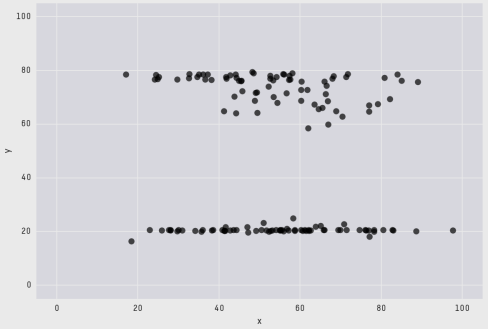
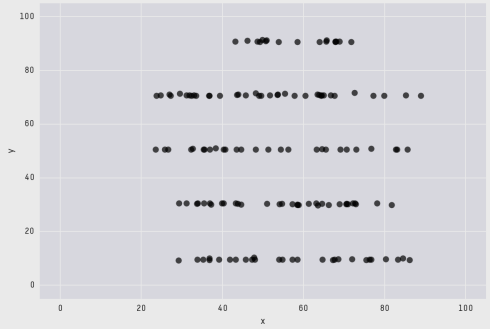
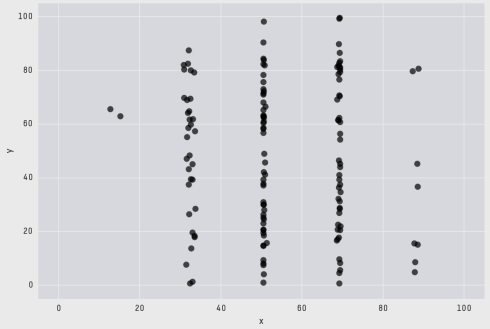
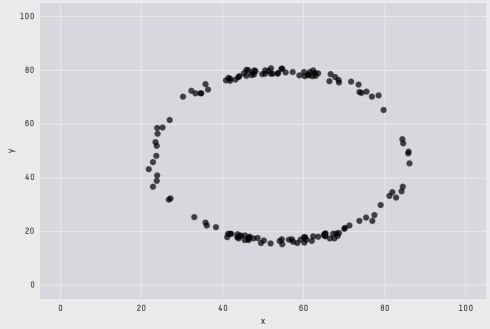
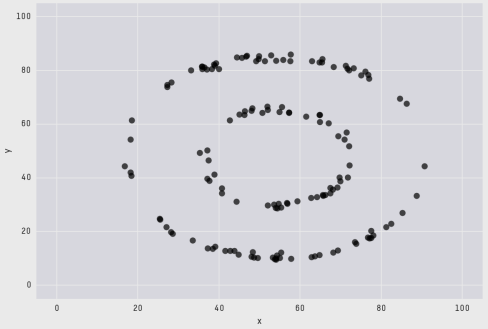
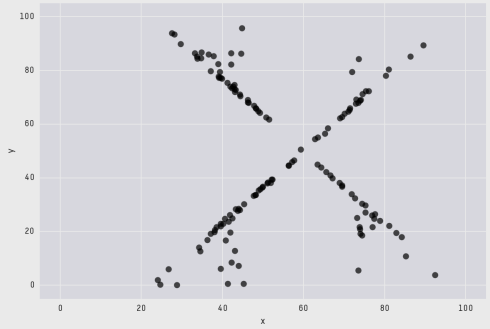
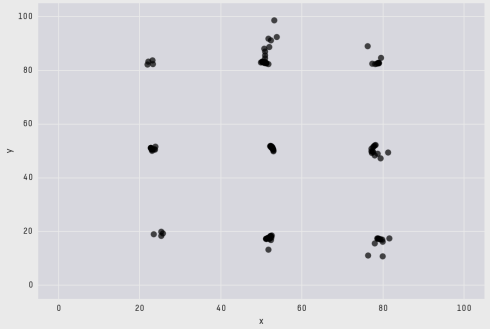
X Mean: 54.26
Y Mean: 47.83
X SD : 16.76
Y SD : 26.93
Corr. : -0.06



The Datasauros Dozens



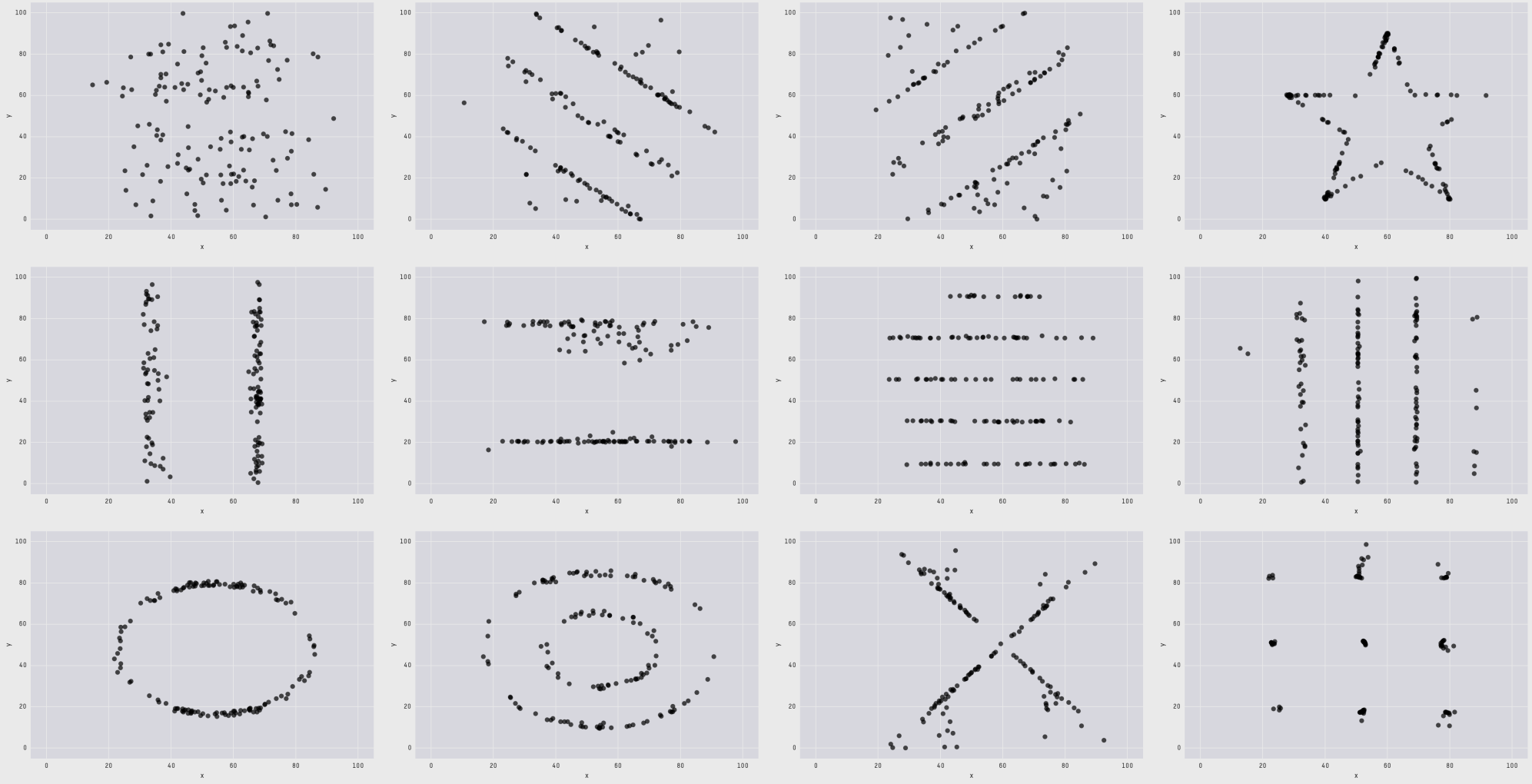
X Mean: 54.26
Y Mean: 47.83
X SD : 16.76
Y SD : 26.93
Corr. : -0.06

The Datasauros Dozens

X Mean: 54.26
Y Mean: 47.83
X SD : 16.76
Y SD : 26.93
Corr. : -0.06

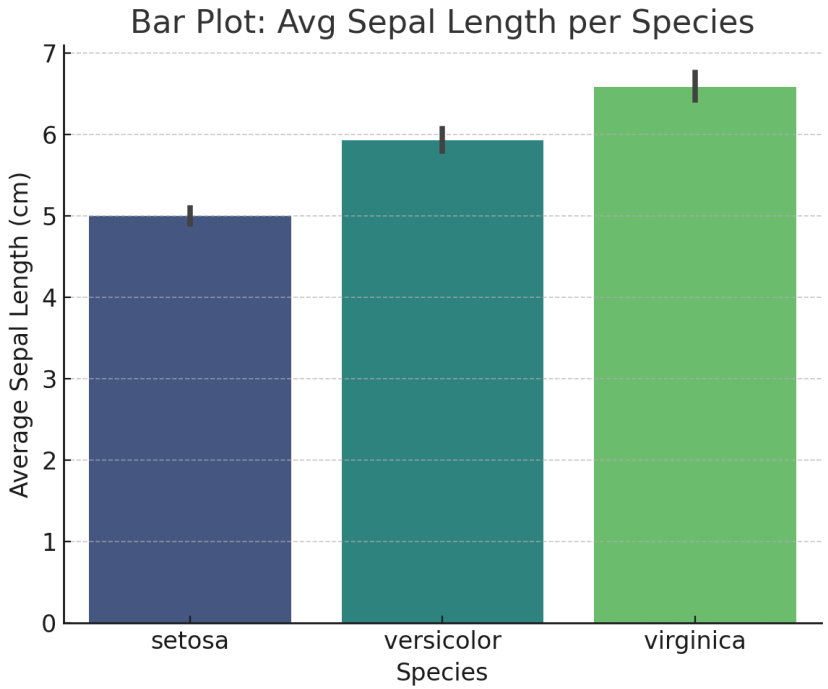
Visualizations (smart ones), may be really helpful

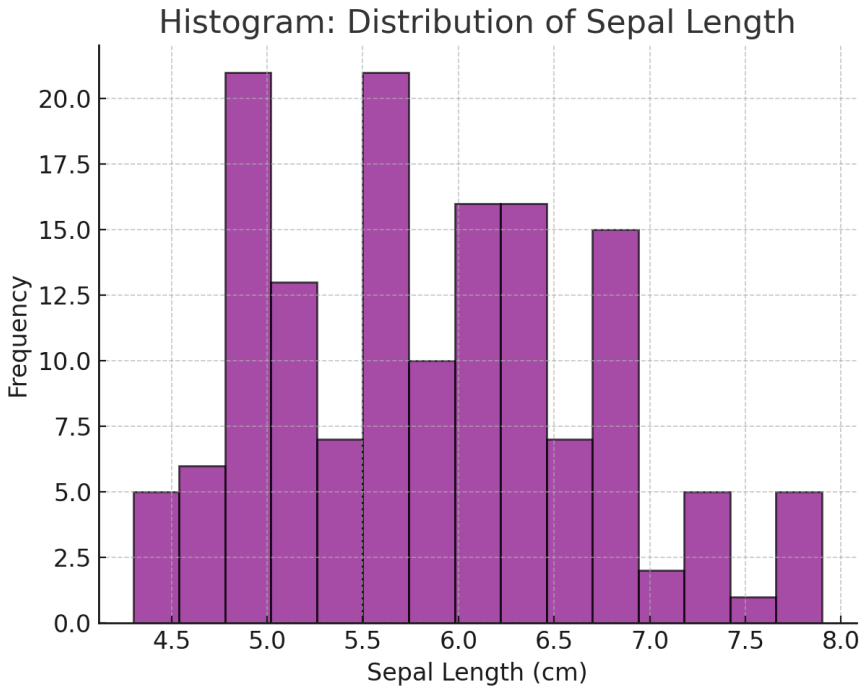


[1D Plots] Bar plots & Histograms

Bar Plot: A graph that represents categorical data using bars, where the height of each bar corresponds to the count, mean, or another statistic of the category. Bars are separate.

Histogram: A graph that represents the distribution of continuous data by grouping values into bins and showing their frequency. Bars are touching since data is continuous.



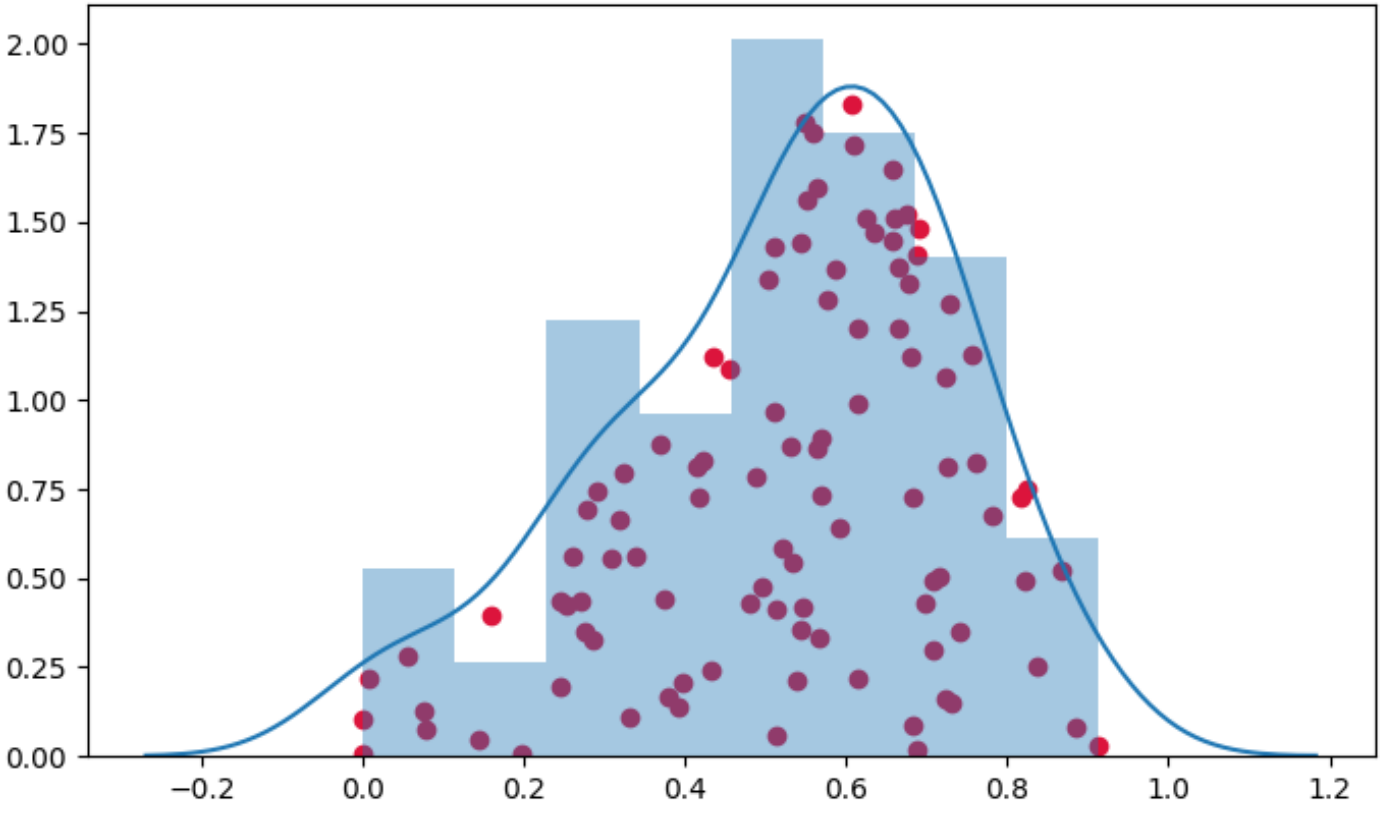


[1D Plots] Kernel Density Estimate (KDE)

KDE is a non-parametric technique for estimating the probability density function (PDF) of a continuous random variable. It provides a smooth representation of the data distribution, unlike histograms, which use discrete bins.

How KDE works:

- Places a kernel function K (e.g., Gaussian) at each data point
- Sums the contributions of all kernels to estimate the density
- The bandwidth (h) controls how spread out each kernel is.



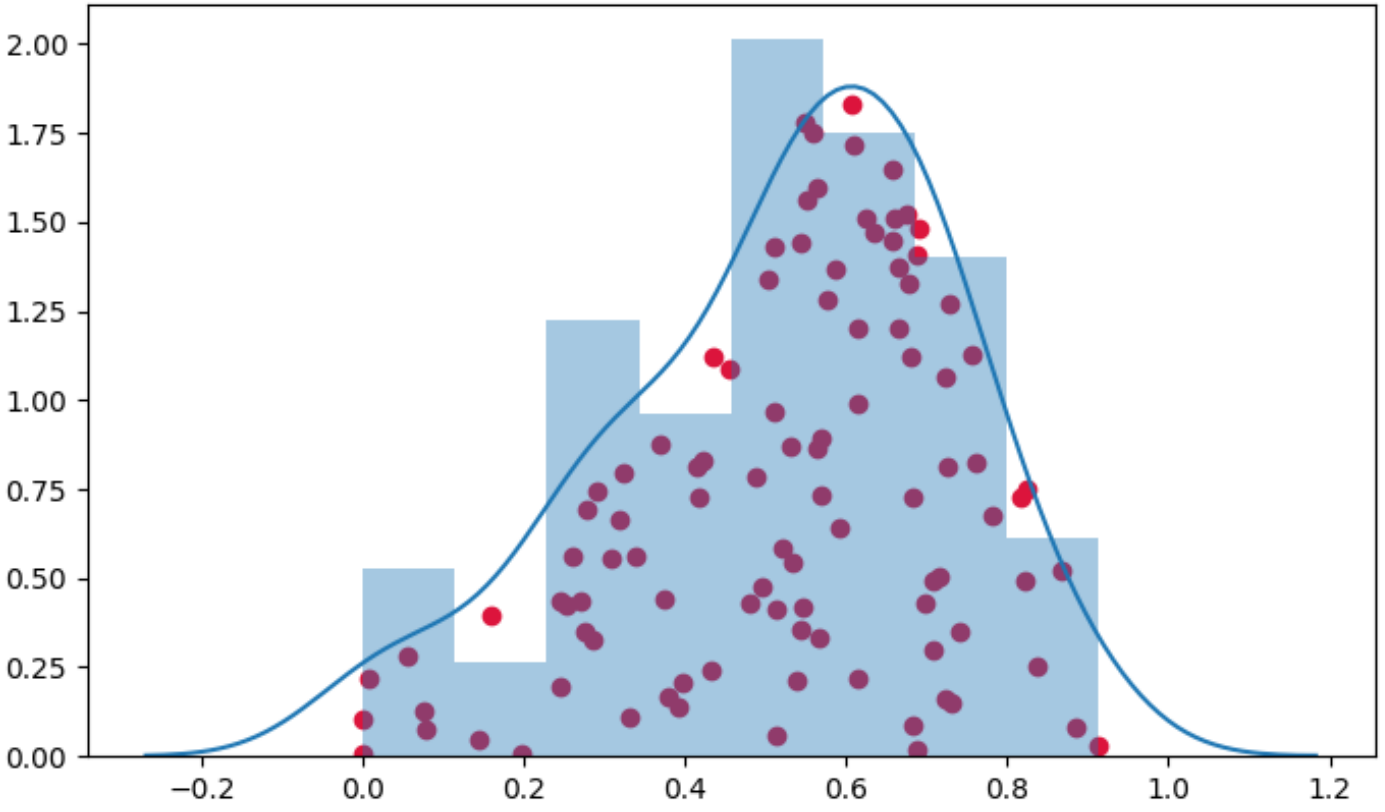
$$\hat{f}(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right)$$

[1D Plots] Kernel Density Estimate (KDE)

KDE is a non-parametric technique for estimating the probability density function (PDF) of a continuous random variable. It provides a smooth representation of the data distribution, unlike histograms, which use discrete bins.

How KDE works:

- Places a kernel function K (e.g., Gaussian) at each data point
- Sums the contributions of all kernels to estimate the density
- The bandwidth (h) controls how spread out each kernel is.



When to Use KDE?

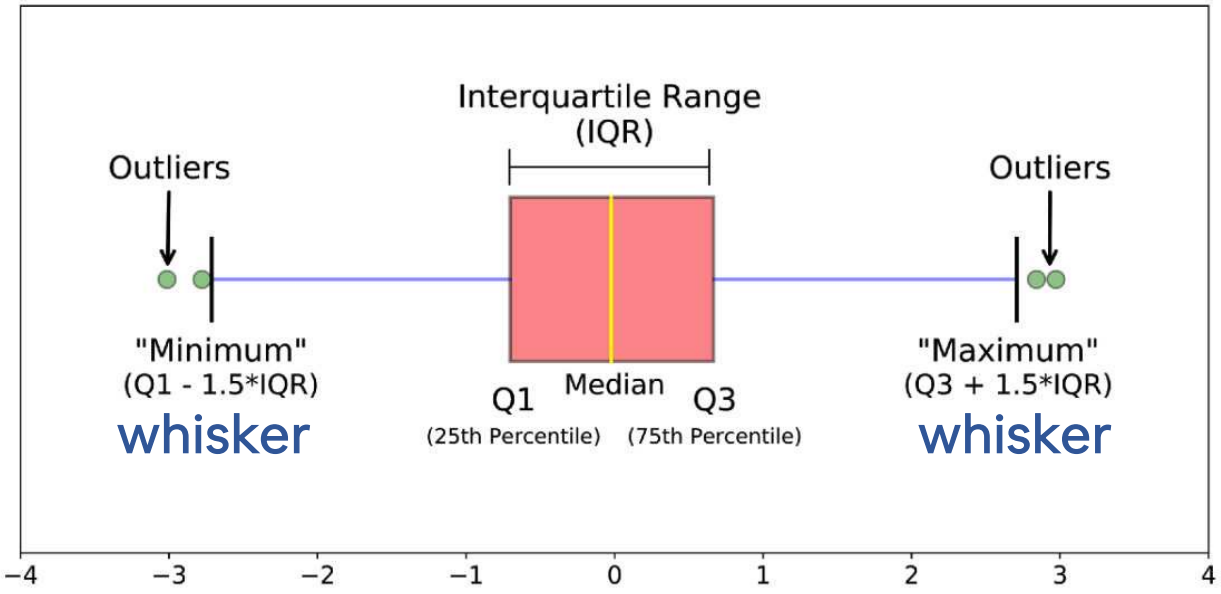
- When you want to **estimate and visualize the true distribution** of data.
- When a **histogram is too coarse or misleading.**
- When comparing **multiple distributions smoothly.**

[1D Plots] Box Plot (Box-and-Whisker Plot)

A Box Plot (or Box-and-Whisker Plot) is a graphical representation of the distribution of a dataset based on five key summary statistics:

- Minimum (Q0) – The smallest data point (excluding outliers).
- First Quartile (Q1, 25th percentile) – The median of the lower half of the dataset.
- Median (Q2, 50th percentile) – The middle value of the dataset.
- Third Quartile (Q3, 75th percentile) – The median of the upper half of the dataset.
- Maximum (Q4) – The largest data point (excluding outliers).

Outliers are plotted as individual points beyond the "whiskers", aka 1.5 x IQR from Q1 and Q3.

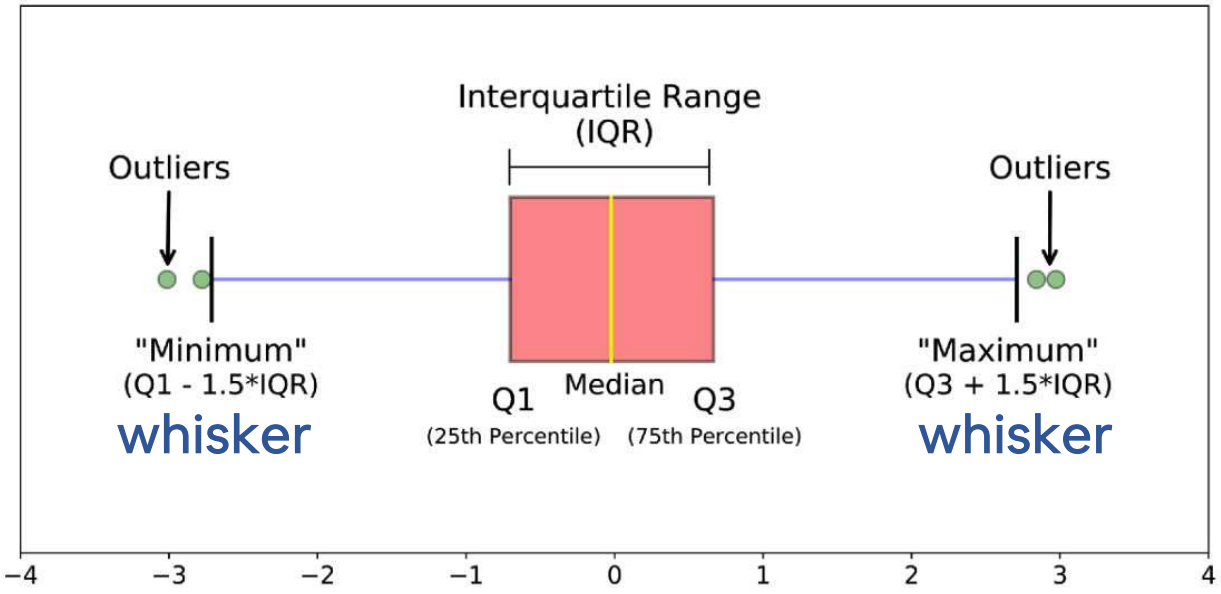


[1D Plots] Box Plot (Box-and-Whisker Plot)

A Box Plot (or Box-and-Whisker Plot) is a graphical representation of the distribution of a dataset based on five key summary statistics:

- Minimum (Q0) – The smallest data point (excluding outliers).
- First Quartile (Q1, 25th percentile) – The median of the lower half of the dataset.
- Median (Q2, 50th percentile) – The middle value of the dataset.
- Third Quartile (Q3, 75th percentile) – The median of the upper half of the dataset.
- Maximum (Q4) – The largest data point (excluding outliers).

Outliers are plotted as individual points beyond the "whiskers", aka 1.5 x IQR from Q1 and Q3.



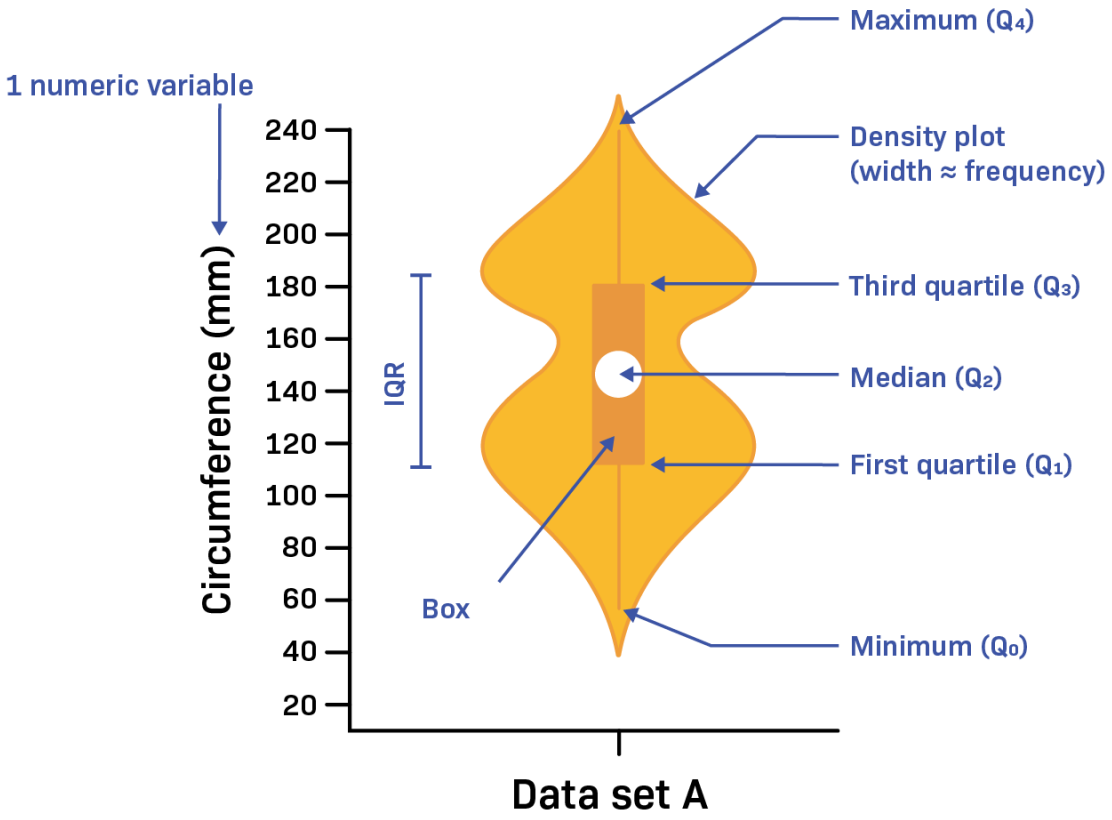
When to Use Box Plot?

- Detect (simple) outliers in your dataset.
- Compare distributions between multiple groups*.
- Understand spread, skewness, and central tendency in data.
- Works well for non-normal and skewed data.

* We will use this a lot when evaluating model performances

[1D Plots] Violin Plot (KDE+Box Plot)

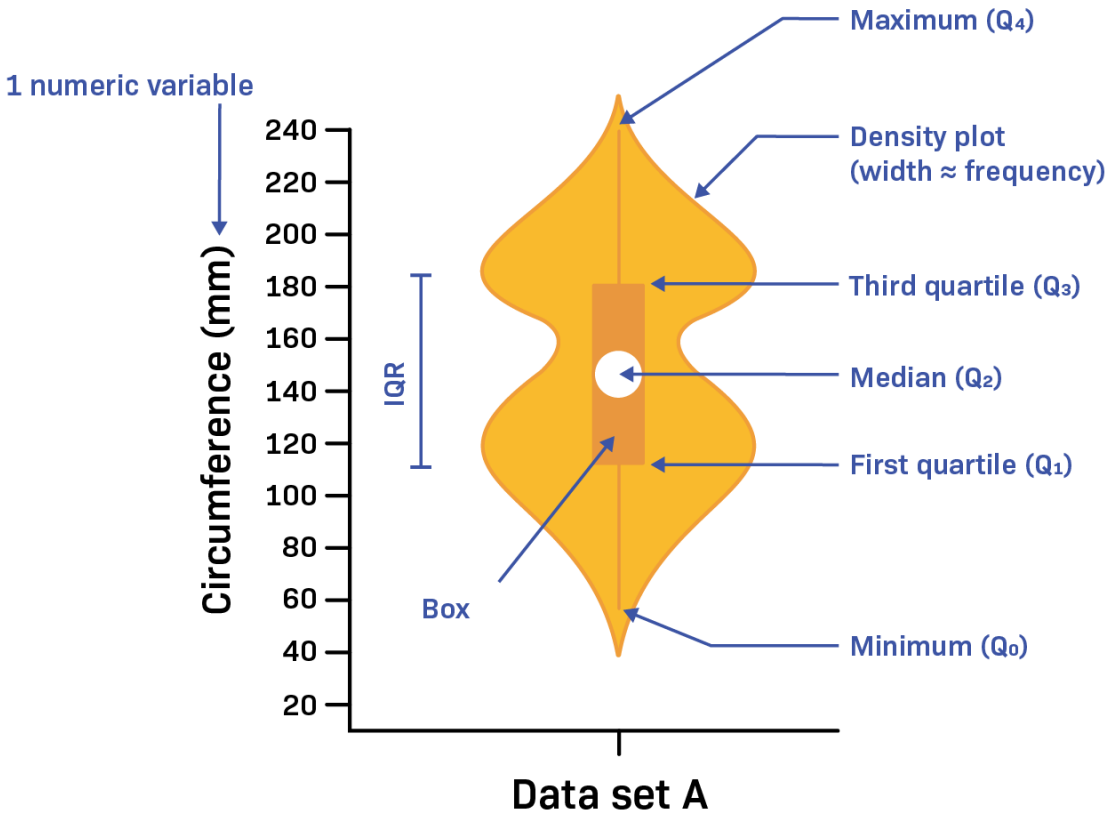
Combines the benefits of box plots and KDE plots (shows both summary statistics and distribution).

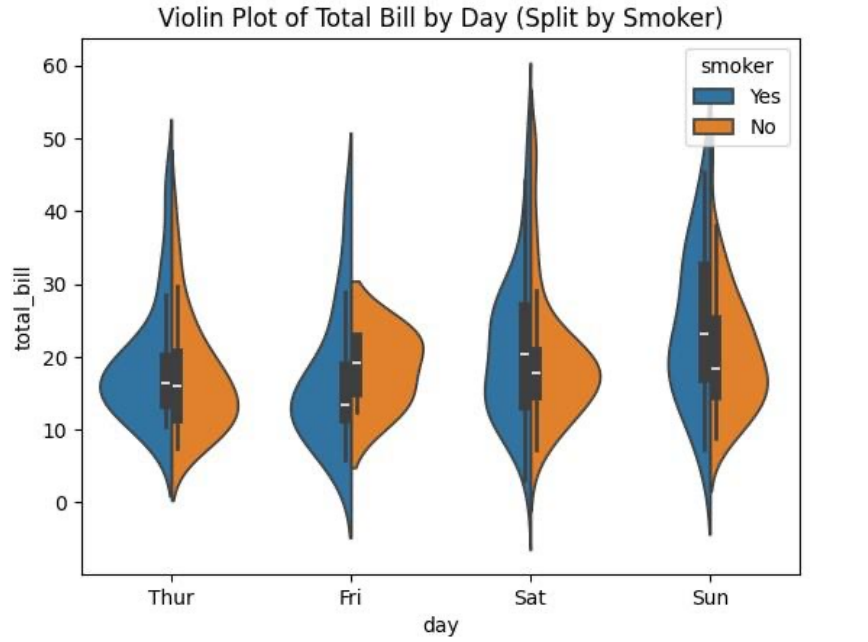


[1D Plots] Violin Plot (KDE+Box Plot)

Combines the benefits of box plots and KDE plots (shows both summary statistics and distribution).

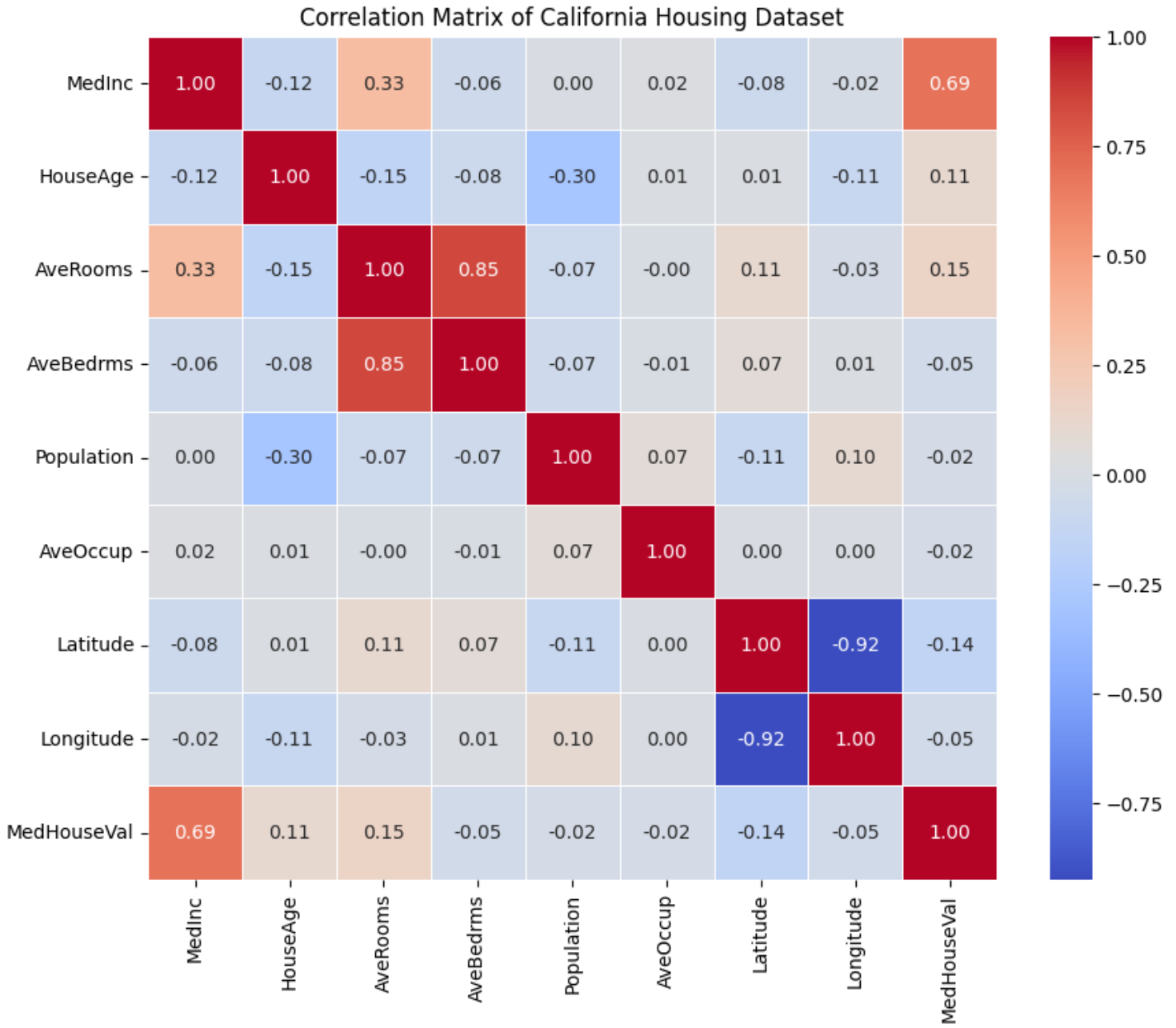
Works well for comparing multiple categories.





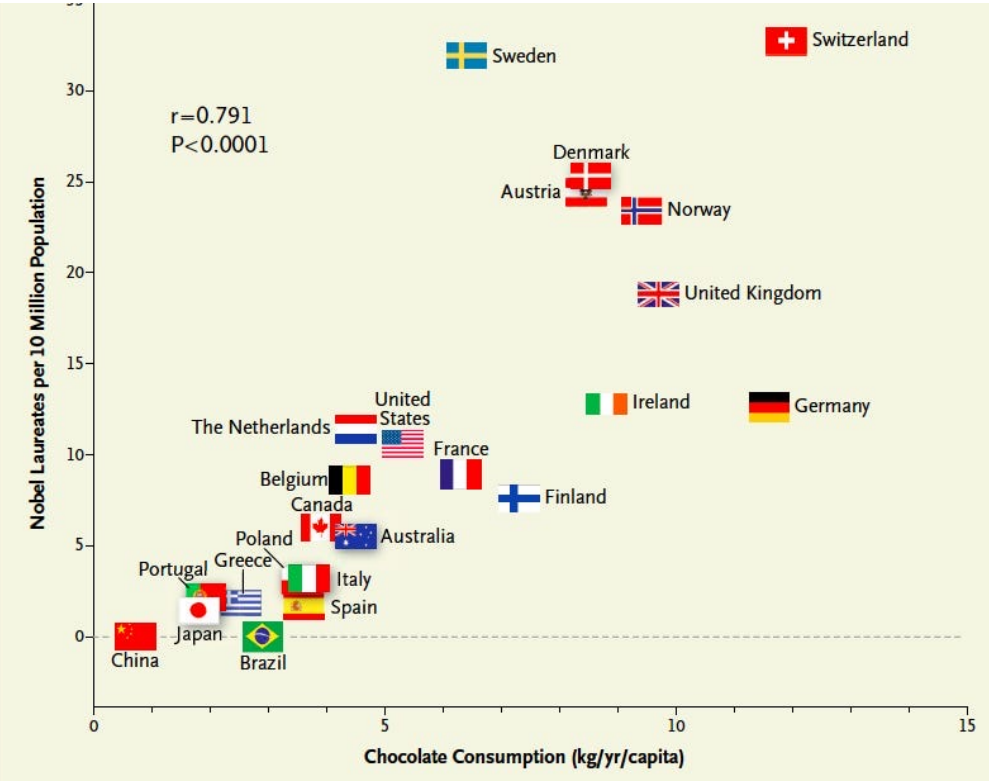
[2D Plot] Correlation heatmap

A correlation heatmap is a visual representation of the correlation matrix between numerical variables in a dataset. It helps to quickly identify relationships between variables using color intensity.



[2D Plot] Scatter Plot

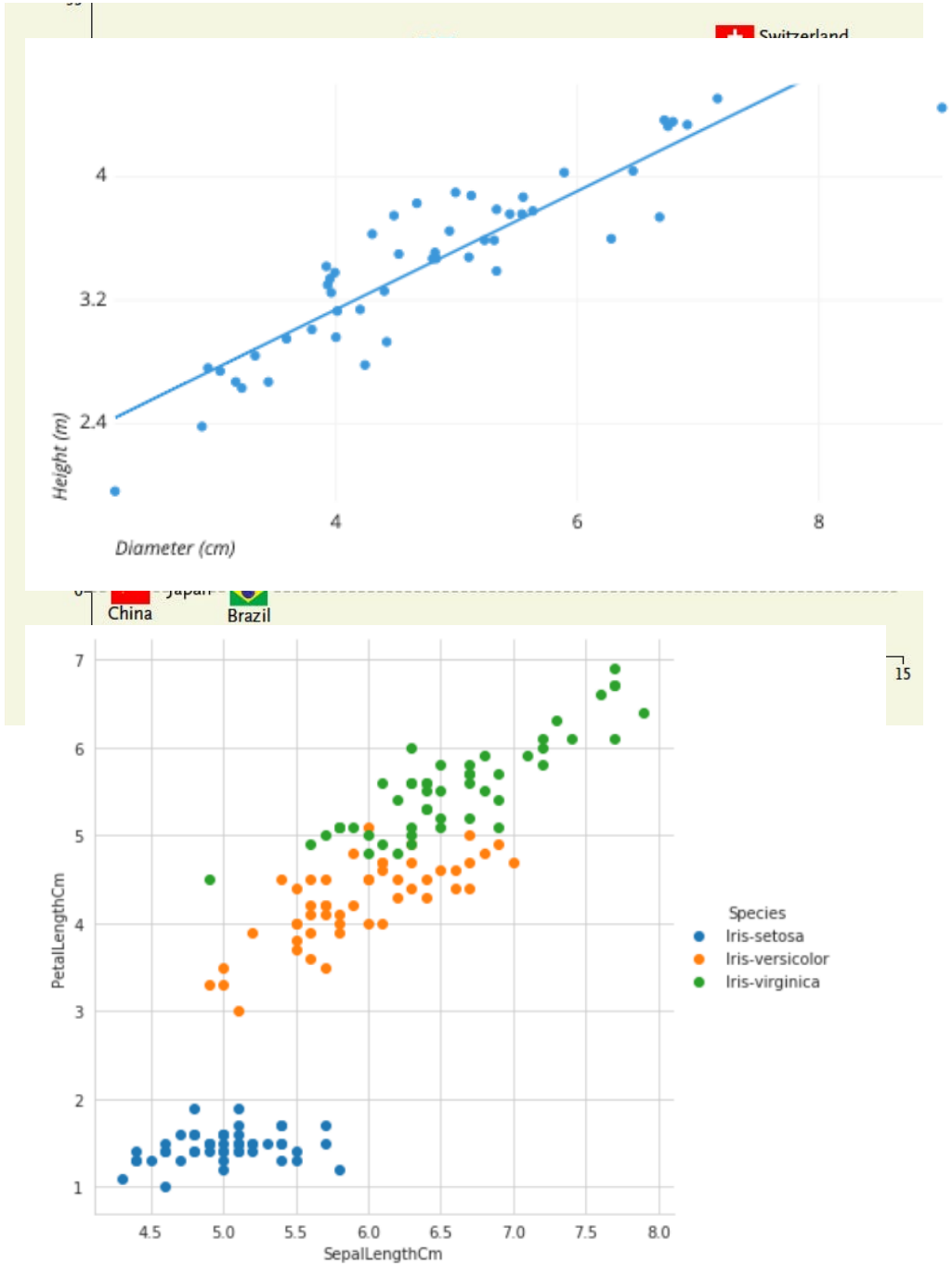
A scatter plot is a visualization that shows the relationship between two numerical variables. Each point represents an observation, with one variable on the X-axis and the other on the Y-axis.



[2D Plot] Scatter Plot

A scatter plot is a visualization that shows the relationship between two numerical variables. Each point represents an observation, with one variable on the X-axis and the other on the Y-axis.

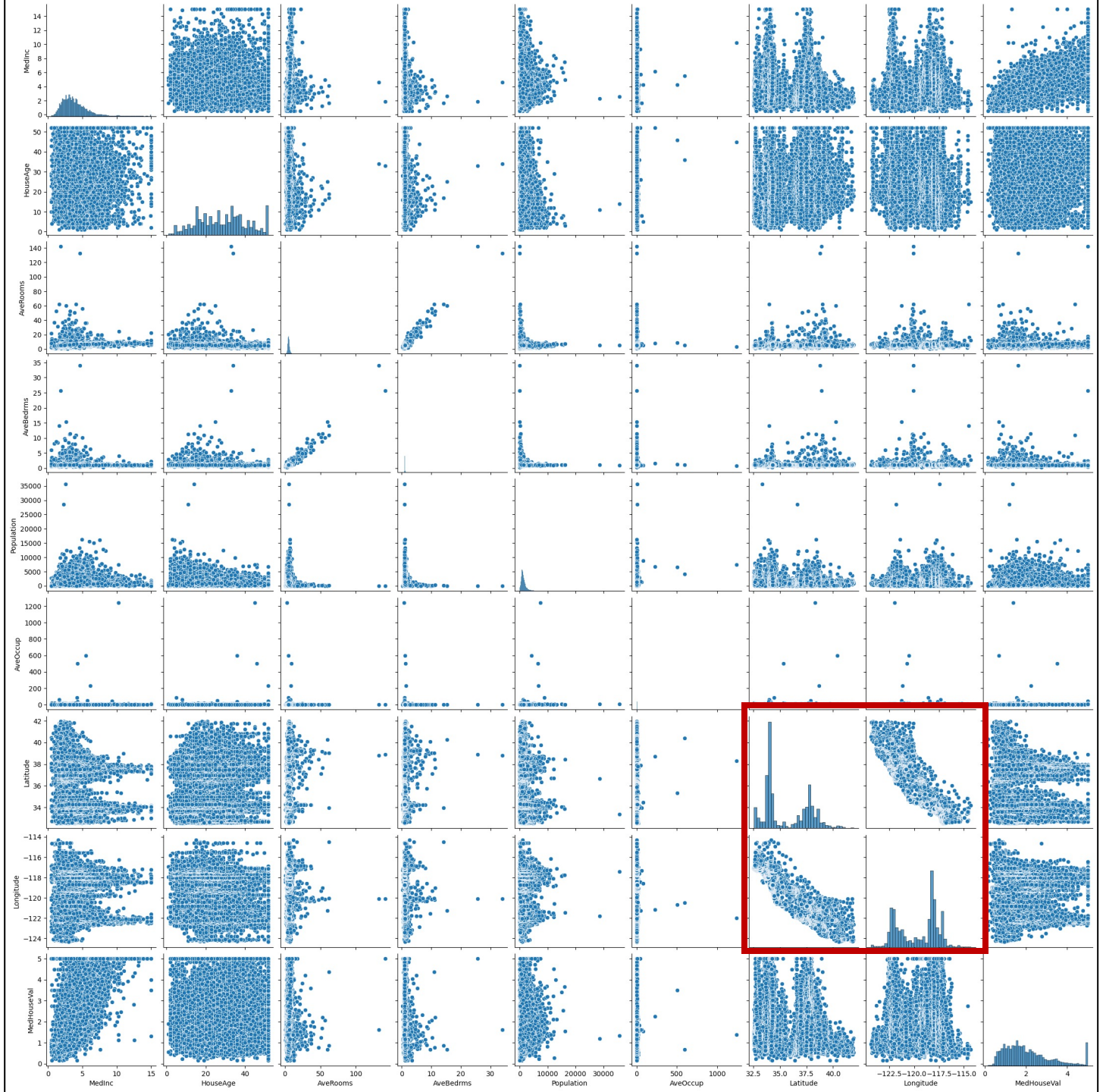
- Shows trends, correlations, and patterns between two variables.
- Reveals outliers that deviate from the general trend.
- Can show linear, non-linear, or no relationships.
- Can be enhanced with color, size, or grouping to represent additional variables.



[2D Plot] Scatterplot Matrix

A scatterplot matrix (also called a pair plot) is a grid of scatter plots that displays pairwise relationships between multiple numerical variables in a dataset:

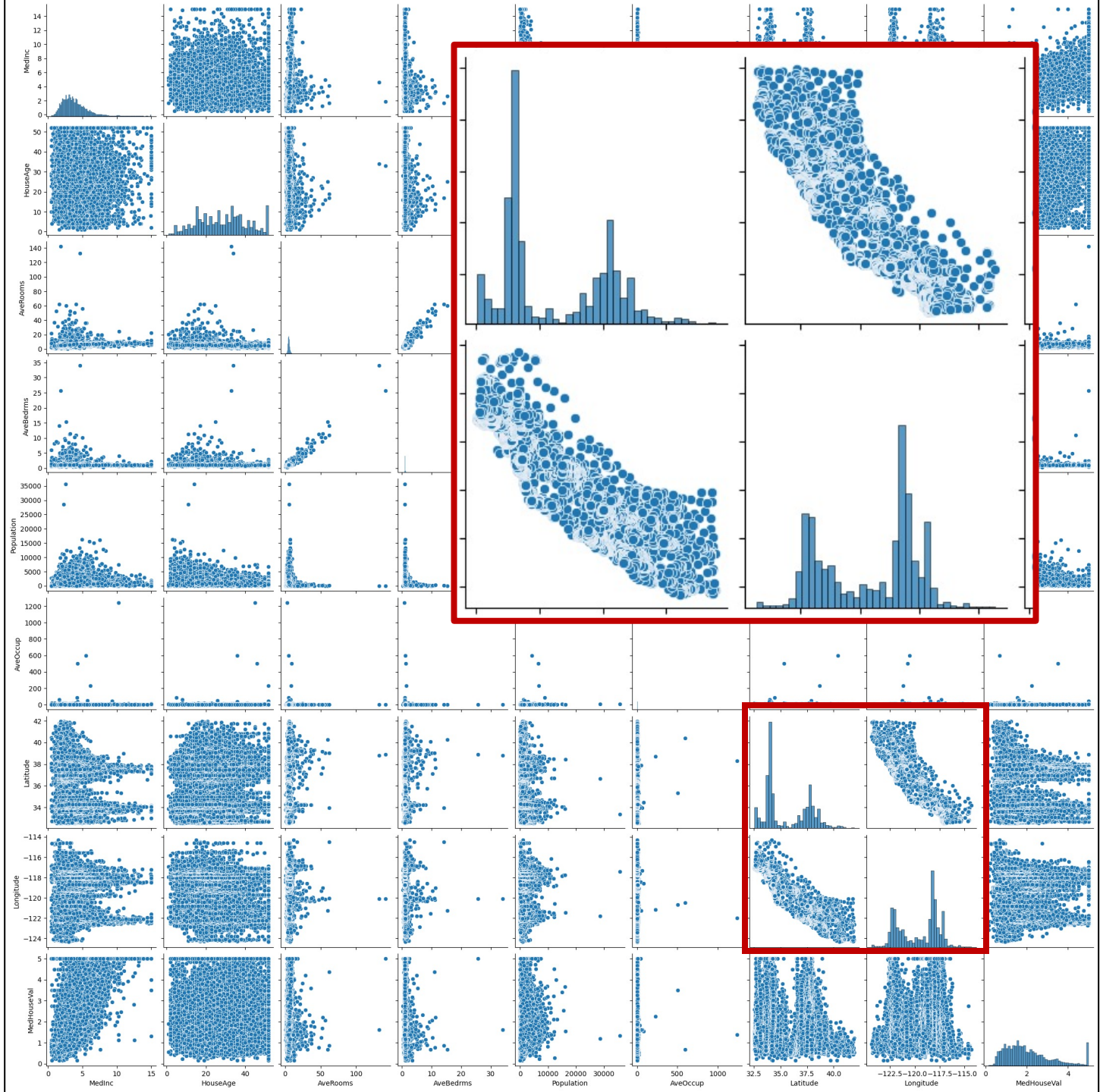
- Each cell in the matrix shows a scatter plot for two variables
- The diagonal often contains histograms or density plots of individual variables
- Useful for detecting correlations, trends, clusters, and outliers in multivariate data.



[2D Plot] Scatterplot Matrix

A scatterplot matrix (also called a pair plot) is a grid of scatter plots that displays pairwise relationships between multiple numerical variables in a dataset:

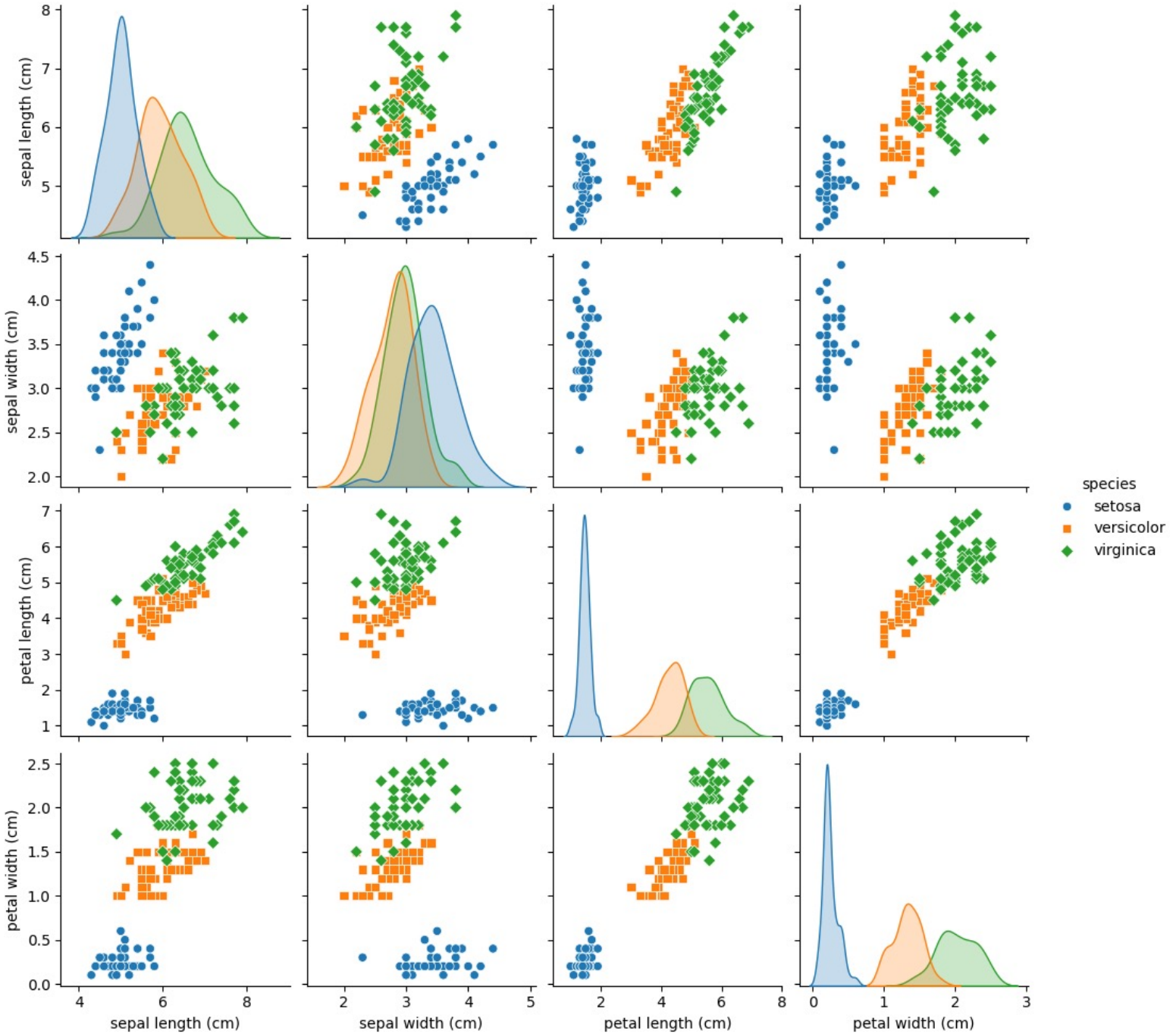
- Each cell in the matrix shows a scatter plot for two variables
- The diagonal often contains histograms or density plots of individual variables
- Useful for detecting correlations, trends, clusters, and outliers in multivariate data.



[2D Plot] Scatterplot Matrix

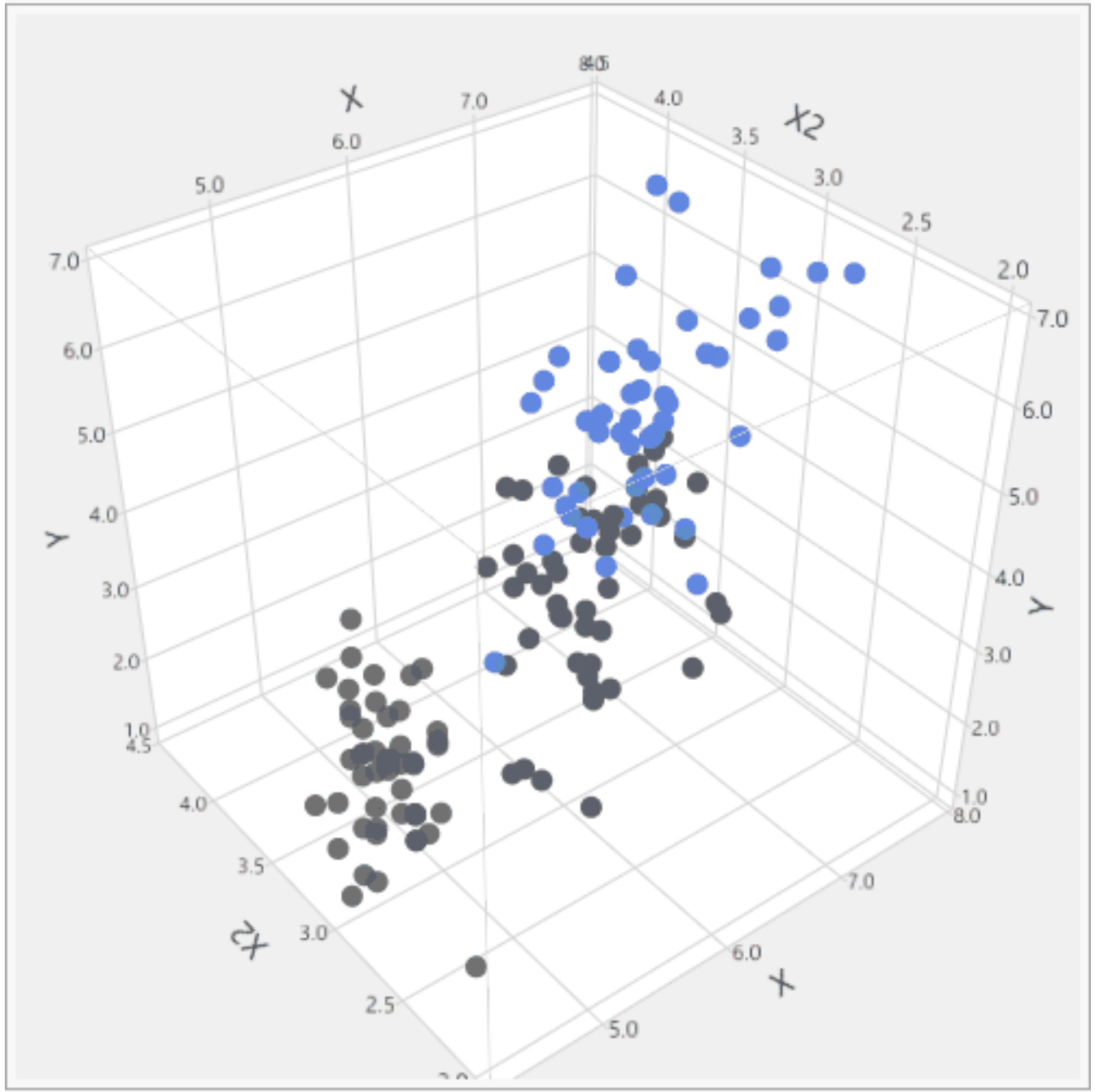
A scatterplot matrix (also called a pair plot) is a grid of scatter plots that displays pairwise relationships between multiple numerical variables in a dataset:

- Each cell in the matrix shows a scatter plot for two variables
- The diagonal often contains histograms or density plots of individual variables
- Useful for detecting correlations, trends, clusters, and outliers in multivariate data.



[3D Plot] Scatterplot Matrix

Scatterplot can be extended to 3D data.



[3D Plot] Scatterplot Matrix

Scatterplot can be extended to 3D data.

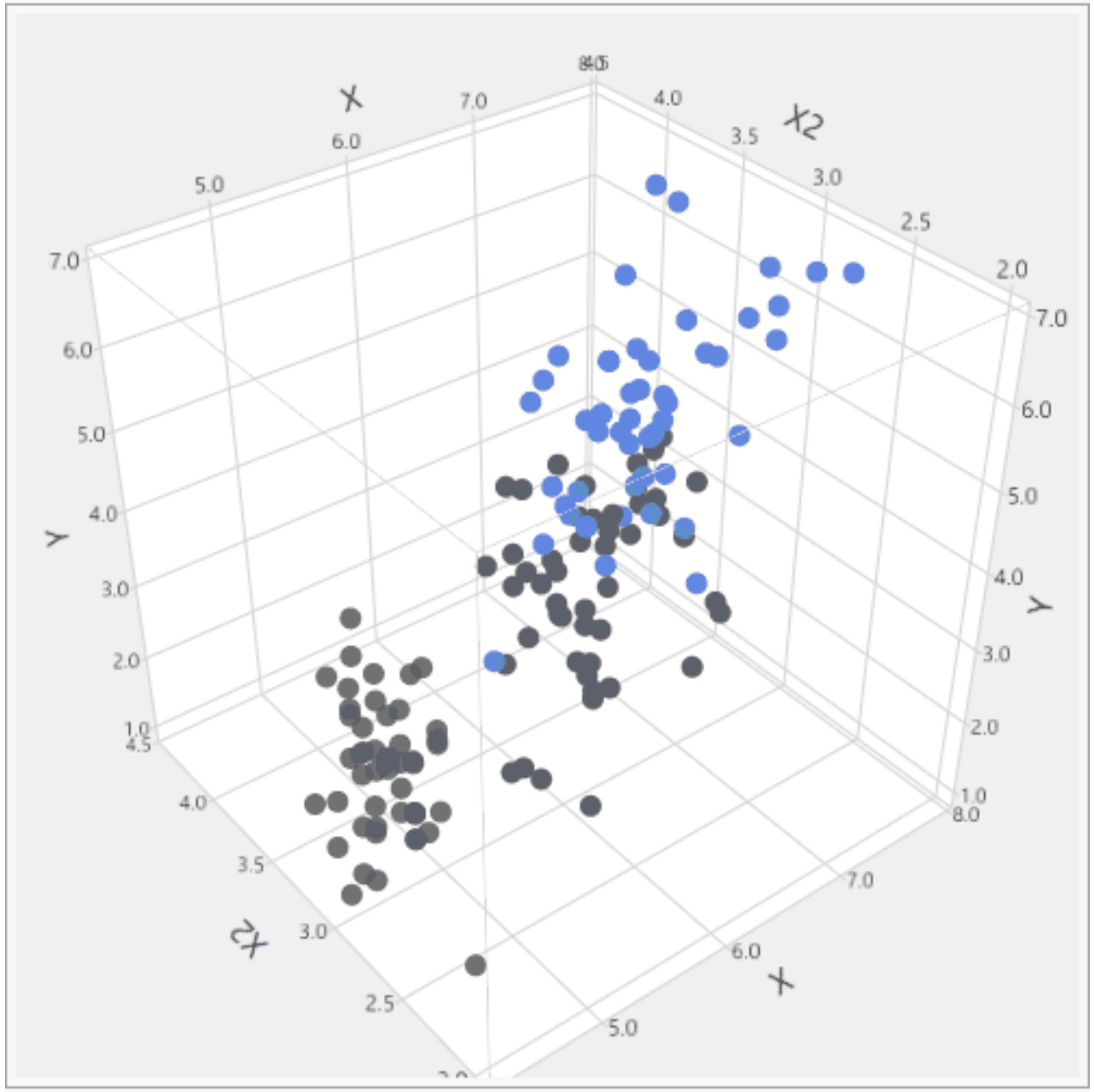
However:

<u>Hard to Interpret</u>

- When plotted in 2D, relationships are easier to see.
- In 3D, overlapping points and perspective distortions make analysis harder.

<u>Difficult to Compare Data Points</u>

- In 2D, distances between points are easy to measure.
- In 3D, points might look closer or farther than they really are due to the viewing angle.



What else?

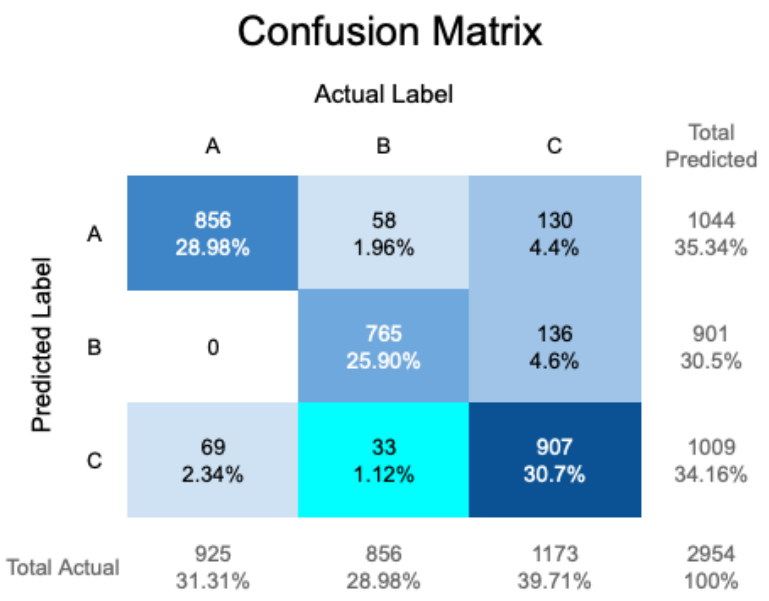
Multi-dimensional plots (lec. 05): we are dealing with multi-dimensional data, and ML is to find multi-dimensional relationships (not pairwise ones)

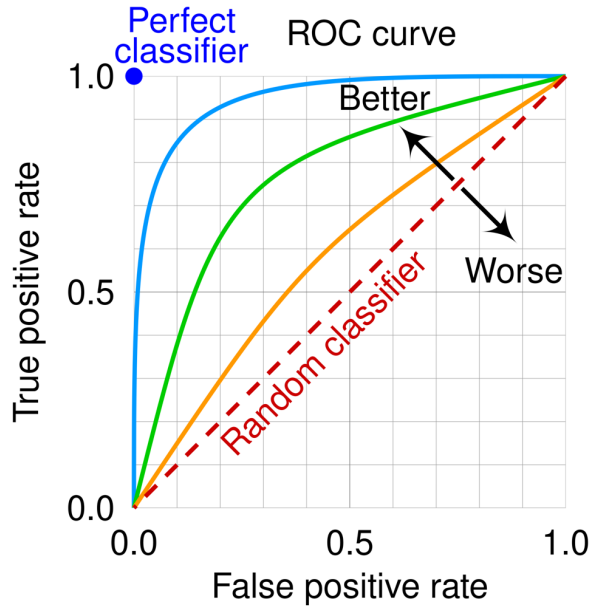
Plots for evaluation (later in the course): for examples

- Confusion matrix

- Area Under the Curve (AUC)

39	State-gov	77516	Bachelors	13	Never-married	Adm-clerical	Not-in-family	White	Male	2174	0	40	United-States	<=50K
50	Self-emp-not-inc	83311	Bachelors	13	Married-civ-spouse	Exec-managerial	Husband	White	Male	0	0	13	United-States	<=50K
38	Private	215646	HS-grad	9	Divorced	Handlers-cleaners	Not-in-family	White	Male	0	0	40	United-States	<=50K
53	Private	234721	11th	7	Married-civ-spouse	Handlers-cleaners	Husband	Black	Male	0	0	40	United-States	<=50K
28	Private	338409	Bachelors	13	Married-civ-spouse	Prof-specialty	Wife	Black	Female	0	0	40	Cuba	<=50K
37	Private	284582	Masters	14	Married-civ-spouse	Exec-managerial	Wife	White	Female	0	0	40	United-States	<=50K
49	Private	160187	9th	5	Married-spouse-absent	Other-service	Not-in-family	Black	Female	0	0	16	Jamaica	<=50K
52	Self-emp-not-inc	209642	HS-grad	9	Married-civ-spouse	Exec-managerial	Husband	White	Male	0	0	45	United-States	>50K
31	Private	45781	Masters	14	Never-married	Prof-specialty	Not-in-family	White	Female	14084	0	50	United-States	>50K
42	Private	159449	Bachelors	13	Married-civ-spouse	Exec-managerial	Husband	White	Male	5178	0	40	United-States	>50K
37	Private	280464	Some-college	10	Married-civ-spouse	Exec-managerial	Husband	Black	Male	0	0	80	United-States	>50K
30	State-gov	141297	Bachelors	13	Married-civ-spouse	Prof-specialty	Husband	Asian-Pac-Islander	Male	0	0	40	India	>50K
23	Private	122272	Bachelors	13	Never-married	Adm-clerical	Own-child	White	Female	0	0	30	United-States	<=50K
32	Private	205019	Assoc-acdm	12	Never-married	Sales	Not-in-family	Black	Male	0	0	50	United-States	<=50K
40	Private	121772	Assoc-voc	11	Married-civ-spouse	Craft-repair	Husband	Asian-Pac-Islander	Male	0	0	40	?	>50K
34	Private	245487	7th-8th	4	Married-civ-spouse	Transport-moving	Husband	Amer-Indian-Eskimo	Male	0	0	45	Mexico	<=50K
25	Self-emp-not-inc	176756	HS-grad	9	Never-married	Farming-fishing	Own-child	White	Male	0	0	35	United-States	<=50K
32	Private	186824	HS-grad	9	Never-married	Machine-op-inspct	Unmarried	White	Male	0	0	40	United-States	<=50K
38	Private	28887	11th	7	Married-civ-spouse	Sales	Husband	White	Male	0	0	50	United-States	<=50K
43	Self-emp-not-inc	292175	Masters	14	Divorced	Exec-managerial	Unmarried	White	Female	0	0	45	United-States	>50K
40	Private	193524	Doctorate	16	Married-civ-spouse	Prof-specialty	Husband	White	Male	0	0	60	United-States	>50K
54	Private	302146	HS-grad	9	Separated	Other-service	Unmarried	Black	Female	0	0	20	United-States	<=50K
35	Federal-gov	76845	9th	5	Married-civ-spouse	Farming-fishing	Husband	Black	Male	0	0	40	United-States	<=50K
43	Private	117037	11th	7	Married-civ-spouse	Transport-moving	Husband	White	Male	0	2042	40	United-States	<=50K
59	Private	109015	HS-grad	9	Divorced	Tech-support	Unmarried	White	Female	0	0	40	United-States	<=50K





Thank you!

Gian Antonio Susto