McMaster University ENGINEERING

W BOOTH SCHOOL OF ENGINEERING PRACTICE AND TECHNOLOGY

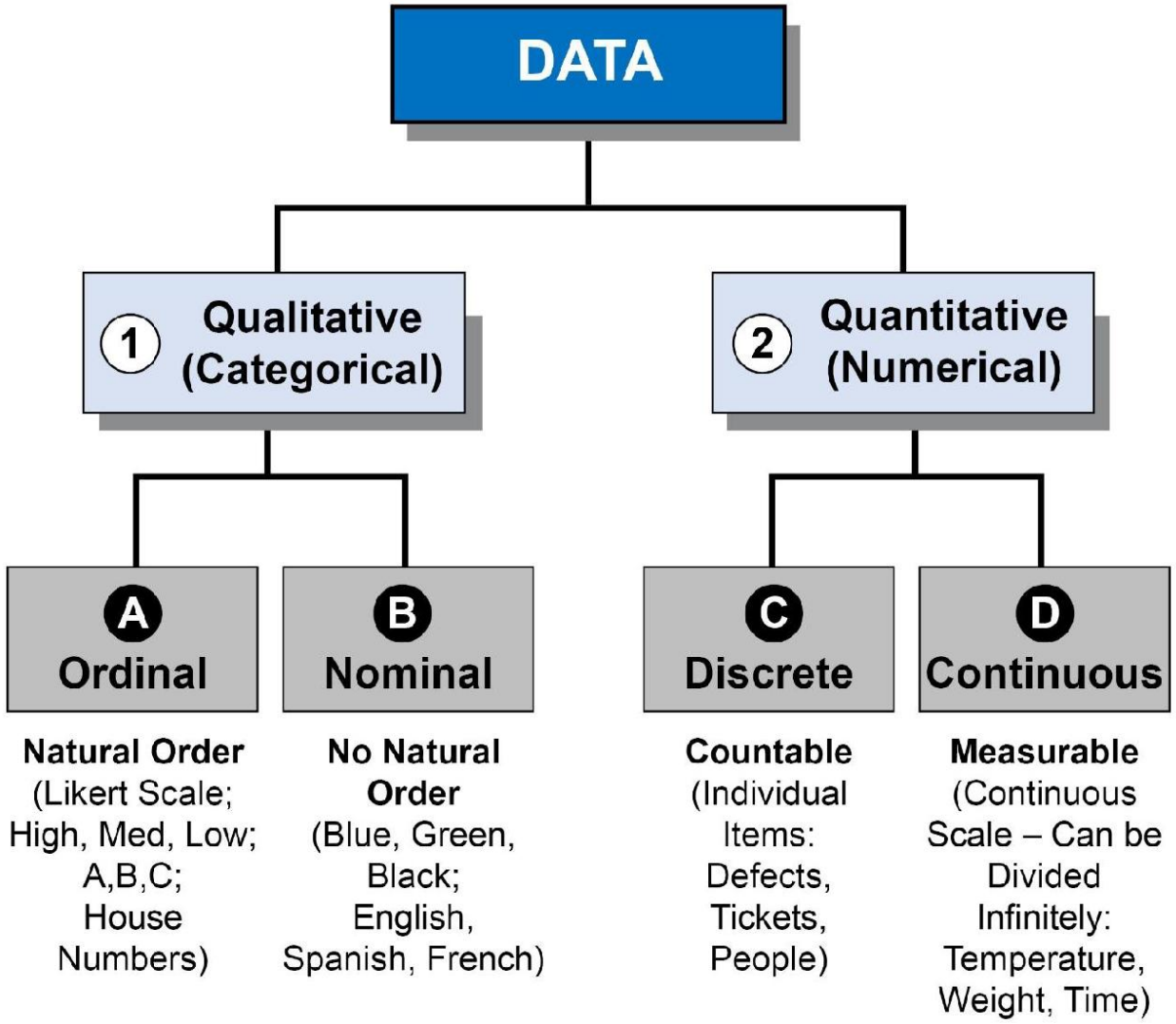
# SEP 785: Machine Learning

**Lecture 2:** Data Mining

Instructor: Dr. Dalia Mahmoud, PhD
(Mechanical Engineering, McMaster University )
Email: mahmoudd@mcmaster.ca

# Recap




DATA
1 Qualitative (Categorical)
2 Quantitative (Numerical)
A Ordinal
B Nominal
C Discrete
D Continuous
Natural Order (Likert Scale; High, Med, Low; A,B,C; House Numbers)
No Natural Order (Blue, Green, Black; English, Spanish, French)
Countable (Individual Items: Defects, Tickets, People)
Measurable (Continuous Scale – Can be Divided Infinitely: Temperature, Weight, Time)

# Lecture Intended Learning Outcomes

McMaster University
ENGINEERING
W Booth School of Engineering Practice and Technology

Recognize effective strategies for examining and interpreting data.

1. Utilize fundamental statistical methods to analyze data.
2. Select suitable data visualization techniques based on the characteristics of the data.
3. Understand foundational concepts of data scaling, normalization, and standardization.
4. Apply appropriate data encoding techniques.
5. Overcome challenges associated with imbalanced datasets and missing data.

# Lecture Contents



- Exploratory Data Analysis
  - Summary Statistics
  - Data Visualization
- Data Preprocessing
  - Scaling, normalization and standardization
  - Data encoding
  - Handling missing data and imbalance data

# Lecture Contents



- Exploratory Data Analysis
  - Summary Statistics
  - Data Visualization
- Data Preprocessing
  - Scaling, normalization and standardization
  - Data encoding
  - Handling missing data and imbalance data

# Exploratory Data Analysis



- You should always “look” at the data first.
- But how do you “look” at features and high-dimensional examples?
  - Summary statistics.
  - Visualizations.
  - ML + DM.

# Summary Statistics



- Measures of location for continuous features:
  - Mean: average value.
  - Median: value such that half points are larger/smaller.
  - Quantiles: value such that ‘k’ fraction of points are larger.
- Measures of spread for continuous features:
  - Range: minimum and maximum values.
  - Variance: measure of how far values are from mean.
- Square root of variance is “standard deviation”.
  - Interquartile ranges: difference between quantiles.

**Population by year, by province and territory (Number)**

| | 2014 |
|---|---|
| **Canada** | **35,540.4** |
| Newfoundland and Labrador | 527.0 |
| Prince Edward Island | 146.3 |
| Nova Scotia | 942.7 |
| New Brunswick | 753.9 |
| Quebec | 8,214.7 |
| Ontario | 13,678.7 |
| Manitoba | 1,282.0 |
| Saskatchewan | 1,125.4 |
| Alberta | 4,121.7 |
| British Columbia | 4,631.3 |
| Yukon | 36.5 |
| Northwest Territories | 43.6 |
| Nunavut | 36.6 |

# Population and Sample

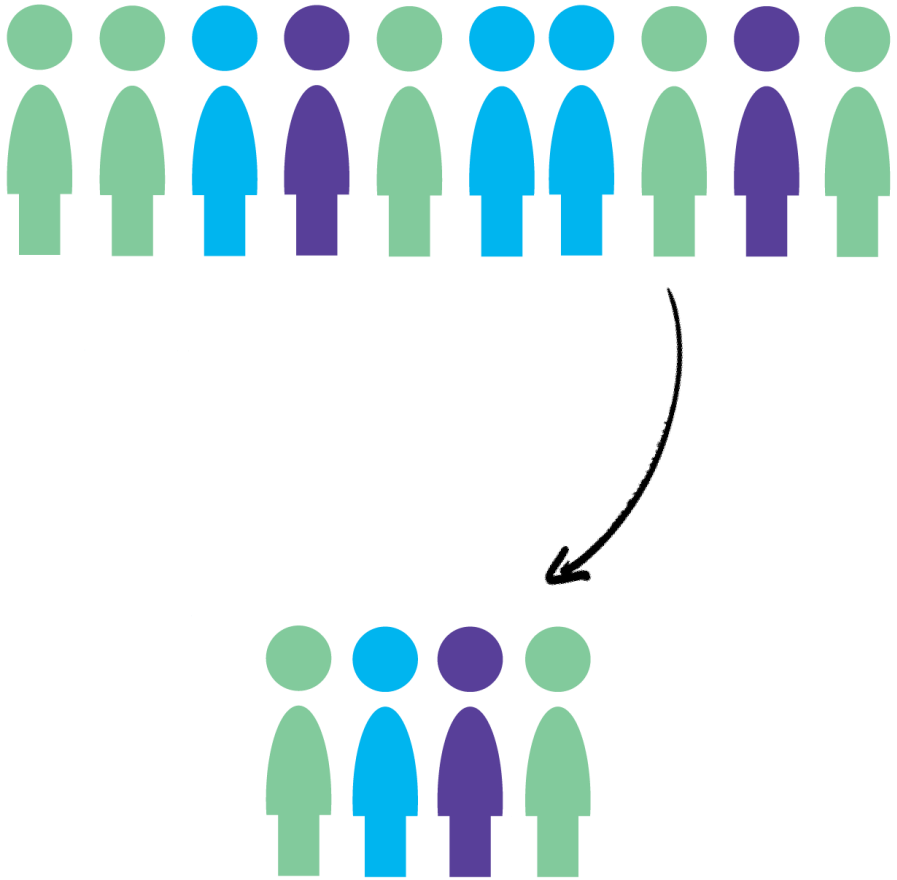


Population:

- In statistics, the population comprises all observations (data points) about the subject under study.
  - An example of a population is studying the voters in an election.

Sample:

- In statistics, a sample is a subset of the population. It is a small portion of the total observed population.
  - An example of a sample is analyzing the first-time voters for an opinion poll.



# Measures of Central Tendency



- Measures of central tendency are the measures that are used to describe the distribution of data using a single value.
- Mean, Median and Mode are the three measures of central tendency.

| | Name | Salary |
|---|---|---|
| 0 | Adeola | 50000 |
| 1 | Minh | 54000 |
| 2 | Riya | 50000 |
| 3 | Sofia | 189000 |
| 4 | Hana | 55000 |
| 5 | Gao | 40000 |
| 6 | Hiroshi | 59000 |

# Measures of Central Tendency

McMaster University
ENGINEERING
W Booth School of Engineering Practice and Technology

- Mean: The arithmetic mean is the average of all the data points.

```
print(df['Salary'].mean())
```

```
71000.0
```

- Median: is the middle value that divides the data into two equal parts once it sorts the data in ascending order.

```
print(df['Salary'].median())
```

```
54000.0
```

- Mode: is the observation (value) that occurs most frequently in the data set. There can be over one mode in a dataset.

```
print(df['Salary'].mode())
```

```
0    50000
Name: Salary, dtype: int64
```

| | Name | Salary |
|---|---|---|
| 0 | Ahmed | 50000 |
| 1 | Minh | 54000 |
| 2 | Riya | 50000 |
| 3 | Sofia | 189000 |
| 4 | Hana | 55000 |
| 5 | Gao | 40000 |
| 6 | Hiroshi | 59000 |

# Range:



- The Range in statistics is the difference between the maximum and the minimum value of the dataset.

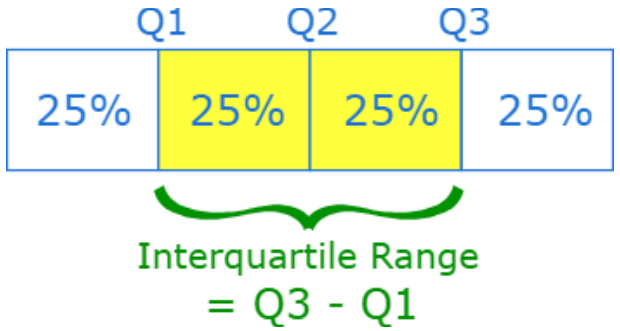
```
import numpy as np

data = [10, 12, 13, 16, 16, 20, 25, 31, 34, 41, 43, 45, 50, 54, 60, 61, 68,
        75, 87, 91, 95]

range = max(data) - min(data)

print(range)
```

```
85
```

# Interquartile Range (IQR) :



- The IQR is a measure of the distance between the 1st quartile (Q1) and 3rd quartile (Q3).




```python
import numpy as np

data = [10, 12, 13, 16, 16, 20, 25, 31, 34, 41, 43, 45, 50, 54, 60, 61, 68,
        75, 87, 91, 95]

# First quartile (Q1)
Q1 = np.percentile(data, 25, interpolation = 'midpoint')

# Third quartile (Q3)
Q3 = np.percentile(data, 75, interpolation = 'midpoint')

# Interquaritle range (IQR)
IQR = Q3 - Q1

print(IQR)
```

```
41.0
```

# Variance and Standard Deviation

McMaster University
ENGINEERING
W Booth School of Engineering Practice and Technology

Variance is used to measure the variability in the data from the mean.

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

```
print(df['Grade'].var())
```

```
685.6190476190476
```

| | Name | Salary | Hours | Grades |
|---|---|---|---|---|
| 0 | Ahmed | 50000 | 41 | 50 |
| 1 | Minh | 54000 | 40 | 50 |
| 2 | Riya | 50000 | 36 | 46 |
| 3 | Sofia | 189000 | 17 | 95 |
| 4 | Hana | 55000 | 35 | 50 |
| 5 | Gao | 40000 | 39 | 5 |
| 6 | Hiroshi | 59000 | 40 | 57 |

# Variance and Standard Deviation

McMaster University
ENGINEERING
W Booth School of Engineering Practice and Technology

Standard deviation in statistics is the square root of the variance. Variance and standard deviation represent the measures of fit, meaning how well the mean represents the data.

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \mu)^2}{n}}$$

```
print(df['Grade'].std())
```

```
26.184328282754315
```

| | Name | Salary | Hours | Grades |
|---|---|---|---|---|
| 0 | Ahmed | 50000 | 41 | 50 |
| 1 | Minh | 54000 | 40 | 50 |
| 2 | Riya | 50000 | 36 | 46 |
| 3 | Sofia | 189000 | 17 | 95 |
| 4 | Hana | 55000 | 35 | 50 |
| 5 | Gao | 40000 | 39 | 5 |
| 6 | Hiroshi | 59000 | 40 | 57 |

# Covariance



- Covariance is a statistical term that refers to a systematic relationship between two random variables in which a change in the other reflects a change in one variable.

$$Covri(x,y) = \frac{\sum_{i=1}^{n}(x_i - x')\,(y_i - y')}{n}$$

?????

$$Covari(x,y) = \frac{\sum_{i=1}^{n}(x_i - x')\,(y_i - y')}{n-1}$$

- The covariance value can range from -∞ to +∞, with a negative value indicating a negative relationship and a positive value indicating a positive relationship.

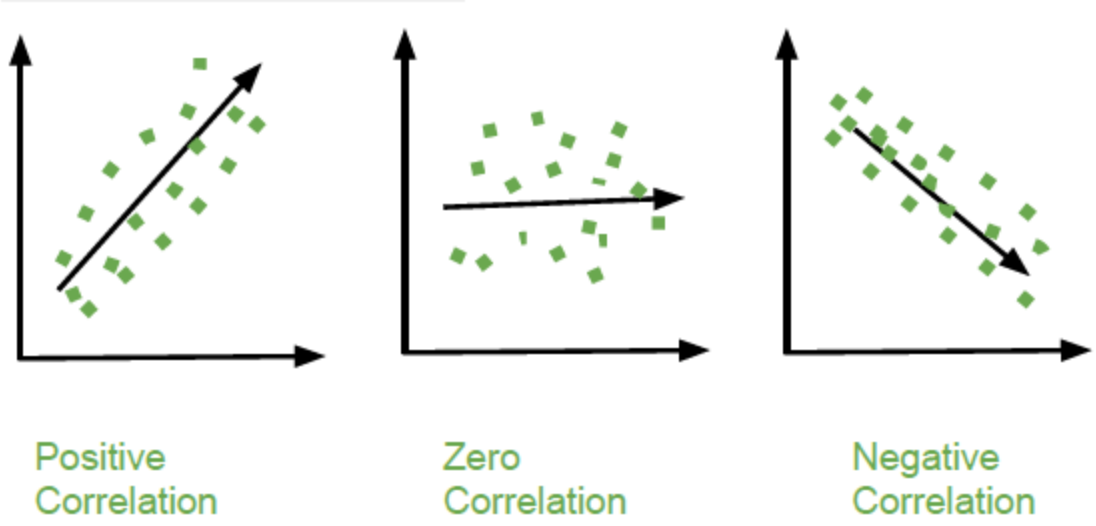
Large Negative Covariance | Nearly Zero Covariance | Large Positive Covariance

# Correlation



- The statistical relationship between two variables is referred to as their correlation.
- A correlation could be presented in different ways:
- Positive Correlation: both variables change in the same direction.
- Neutral Correlation: No relationship in the change of the variables.
- Negative Correlation: variables change in opposite directions.

$$Corr(x,y) = \frac{\sum_{i=1}^{n}(x_i - x')\,(y_i - y')}{\sqrt{\sum_{i=1}^{n}(x_i - x')^2 \sum_{i=1}^{n}(y_i - y')^2}}$$



# Covariance Vs Correlation

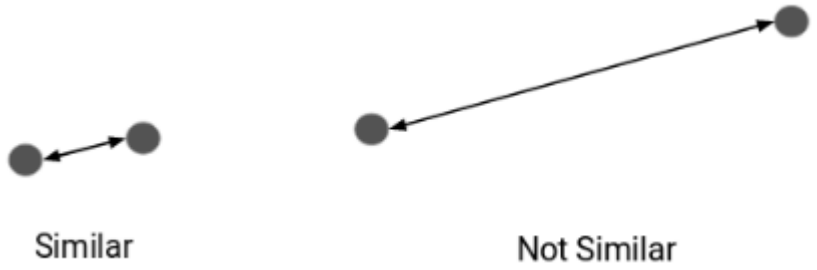


| Covariance | Correlation |
| --- | --- |
| Covariance is a measure of how much two random variables vary together | Correlation is a statistical measure that indicates how strongly two variables are related. |
| Involves the relationship between two variables or data sets | Involves the relationship between multiple variables as well |
| Lie between -infinity and +infinity | Lie between -1 and +1 |
| Measure of correlation | Scaled version of covariance |
| Provides direction of relationship | Provides direction and strength of relationship |
| Dependent on scale of variable | Independent on scale of variable |
| Have dimensions | Dimensionless |

# Distance Metrics



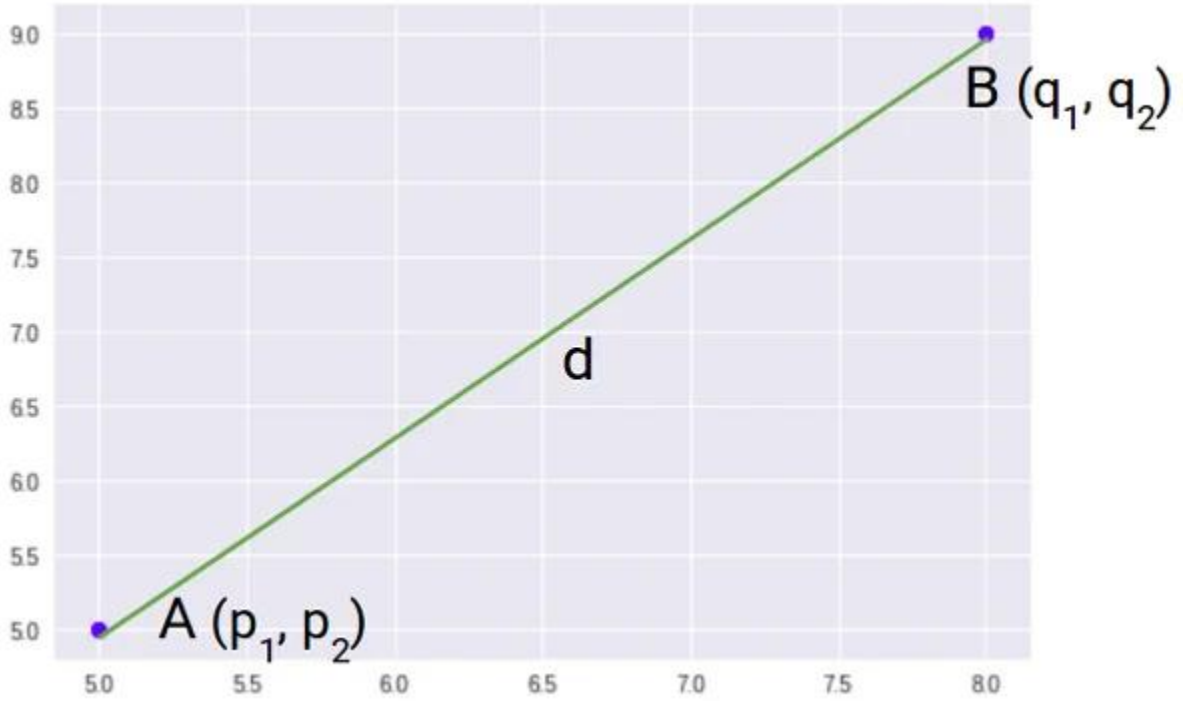
- Distance metrics deal with finding the proximity or distance between data points and determining if they can be clustered together.
  - Euclidean Distance
  - Manhattan Distance
  - Mahalanobis Distance
  - Hamming Distance



# Euclidean Distance



- Euclidean Distance represents the shortest distance between two vectors. It is the square root of the sum of squares of differences between corresponding elements.

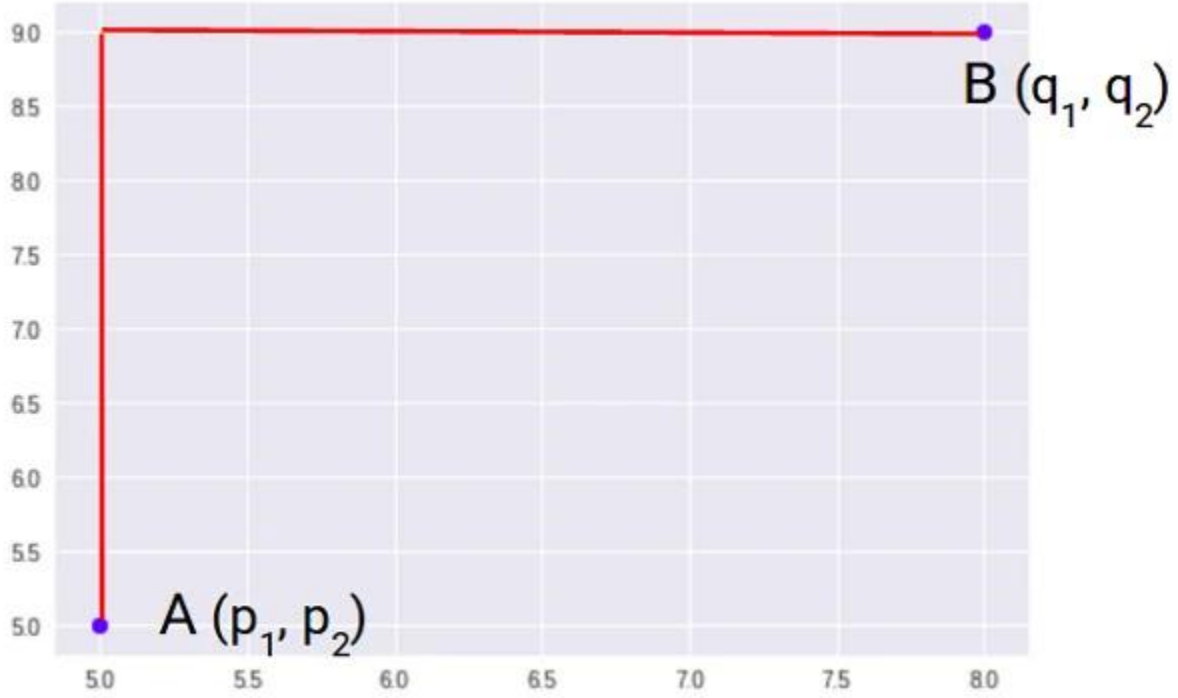


# Manhattan Distance



- Manhattan Distance is the sum of absolute differences between points across all the dimensions

$$d = |p_1 - q_1| + |p_2 - q_2|$$

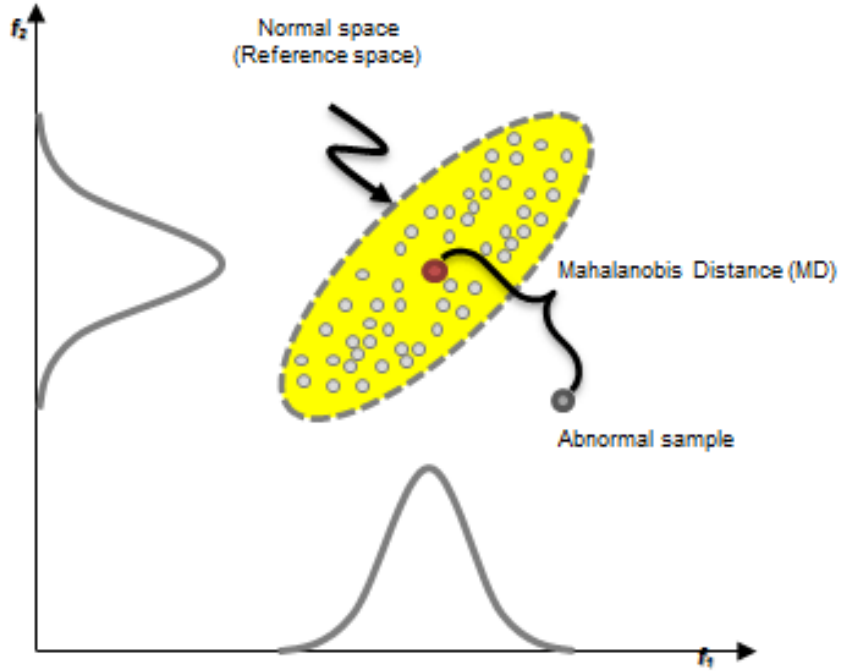


# Mahalanobis distance



- Mahalonobis distance is the distance between a point and a distribution. And not between two distinct points. It is effectively a multivariate equivalent of the Euclidean distance.

$$D^2 = (x - m)^T \cdot C^{-1} \cdot (x - m)$$

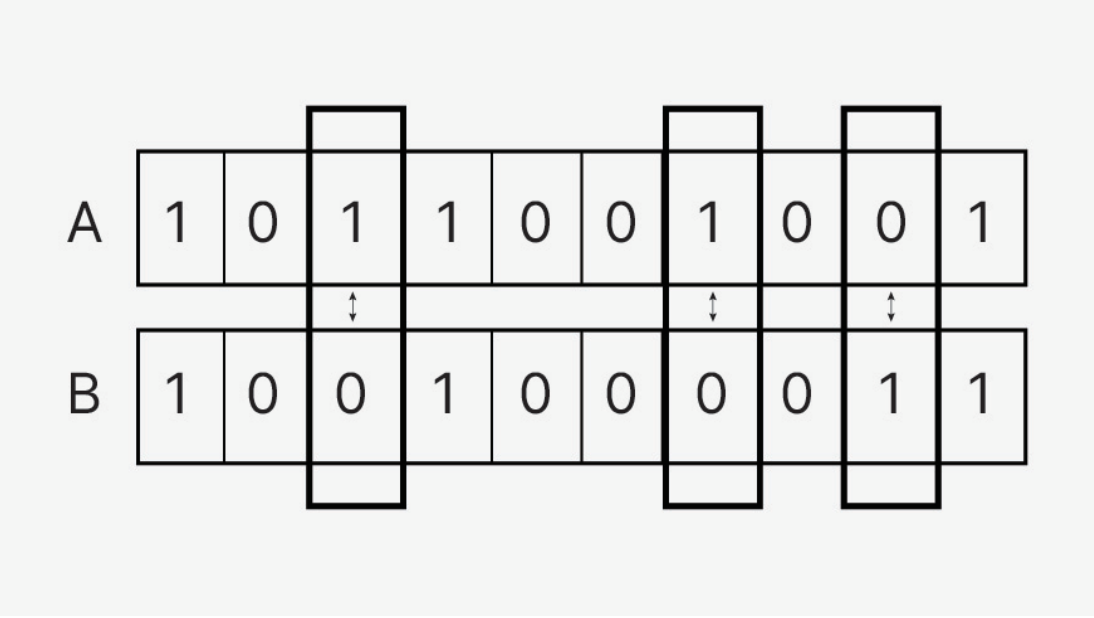



where,
D^2 is the square of the Mahalanobis distance.
x is the vector of the observation (row in a dataset)
m is the vector of mean values of independent variables (mean of each column),
C^(-1) is the inverse covariance matrix of indepen variables.

# Hamming Distance

McMaster University
ENGINEERING
W Booth School of Engineering Practice and Technology

- The Hamming Distance Algorithm is a fundamental tool for measuring the dissimilarity between two pieces of data, typically strings or integers.




```python
def hamming_distance(str1, str2):
    if len(str1) != len(str2):
        raise ValueError("Strings must be of equal length")
    return sum(ch1 != ch2 for ch1, ch2 in zip(str1, str2))
# Example
string1 = "karolin"
string2 = "kathrin"
print(hamming_distance(string1, string2))
```

3

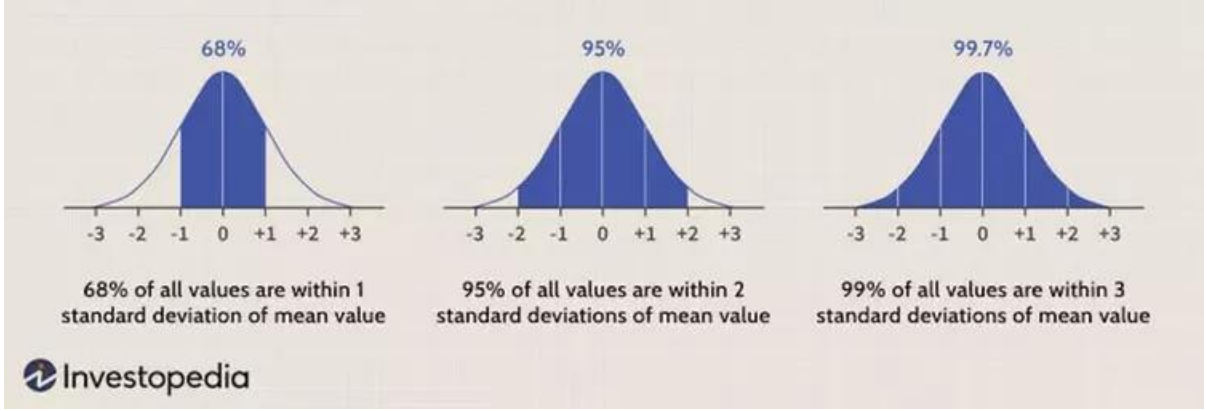
```python
def hamming_distance(x, y):
    return bin(x ^ y).count('1')
# Example
num1 = 2
num2 = 7
print(hamming_distance(num1, num2))
```

2

# Normal Distribution



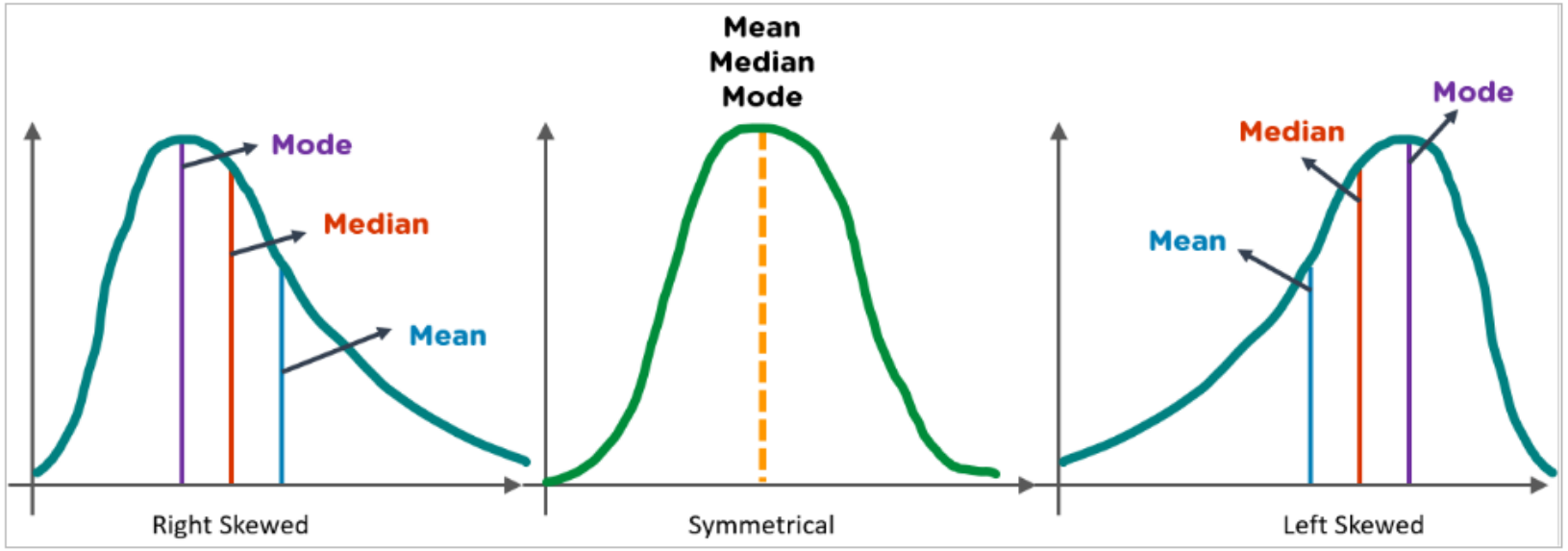
- The standard normal distribution has two parameters: the mean and the standard deviation. In a normal distribution, mean (average), median (midpoint), and mode (most frequent observation) are equal.
- For all normal distributions, 68.2% of the observations will appear within plus or minus one standard deviation of the mean; 95.4% will fall within +/- two standard deviations; and 99.7% within +/- three standard deviations.



# Skewness:



- Skewness measures the shape of the distribution. A distribution is symmetrical when the proportion of data at an equal distance from the mean (or median) is equal. If the values extend to the right, it is right-skewed, and if the values extend left, it is left-skewed.



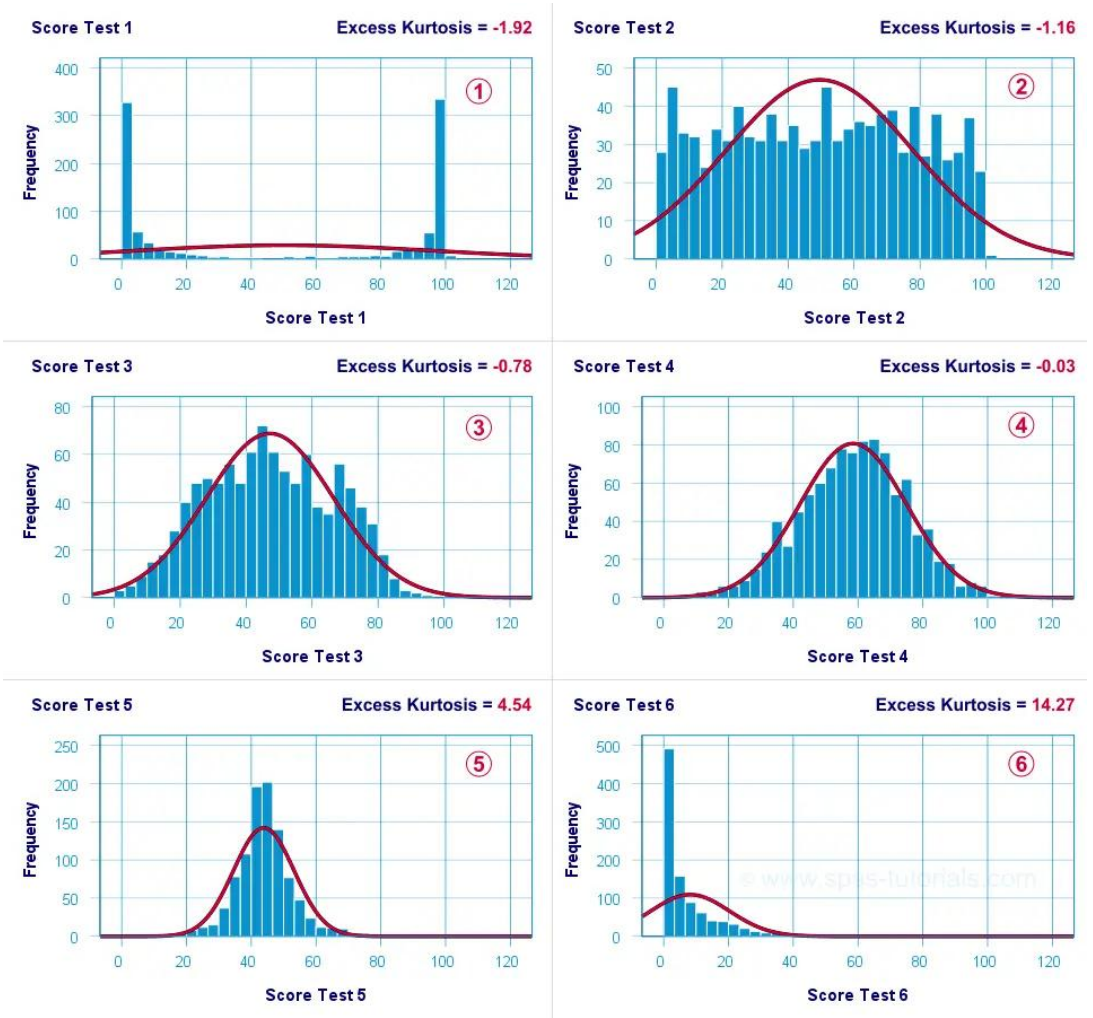
# Kurtosis



- Kurtosis in statistics is used to check whether the tails of a given distribution have extreme values. It also represents the shape of a probability distribution.

$$K_p = \frac{M_4}{M_2^2}$$

$$M_2 = \frac{\sum_{i=1}^{N}(X_i - \overline{X})^2}{N} \qquad M_4 = \frac{\sum_{i=1}^{N}(X_i - \overline{X})^4}{N}$$



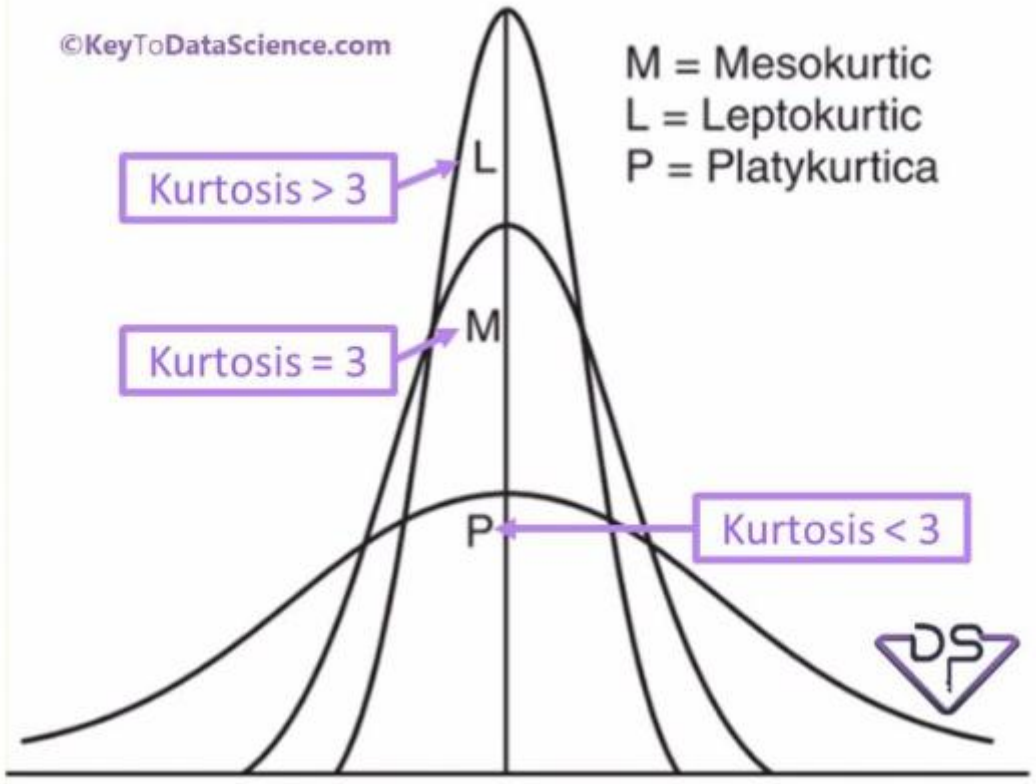
# Excess Kurtosis



A normally distributed variable has a kurtosis of 3.0. Since this is undesirable, population excess kurtosis $EK_p$ is defined as

$$EK_p = K_p - 3$$

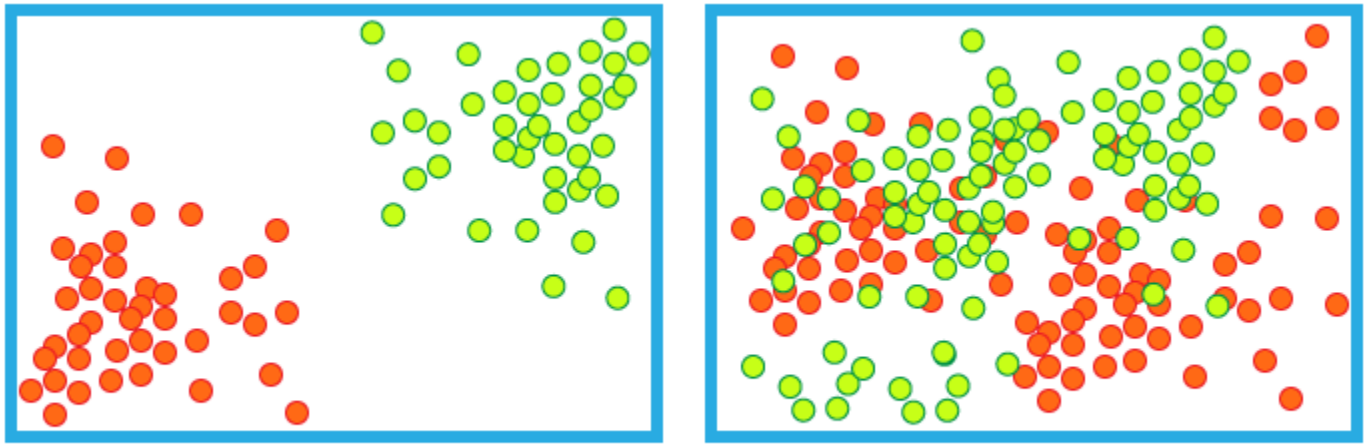
so that excess kurtosis is 0.0 for a normally distributed variable.



# Entropy



- Entropy is the measurement of disorder or impurities in the information processed in machine learning. It determines how a decision tree chooses to split data.



# Entropy



- So how do we calculate Entropy? Well, Entropy is denoted with the variable “H”, and uses the following formula:

$$H = -\sum_i p_i(\log_2 p_i)$$

- So, lets say we have the following set of Heads (H) and Tails (T):
- S = {H, H, H, H, H, T, T, T, T, T, T, T, T}
- Here we have 5H’s and 8T’s This gives us a total of 13. If we substitute these values in we get:

$$H(S) = -\left[\underbrace{\left(\frac{8}{13}\log_2\frac{8}{13}\right)}_{\text{For the T's}} + \underbrace{\left(\frac{5}{13}\log_2\frac{5}{13}\right)}_{\text{For the H's}}\right] = 0.96124$$

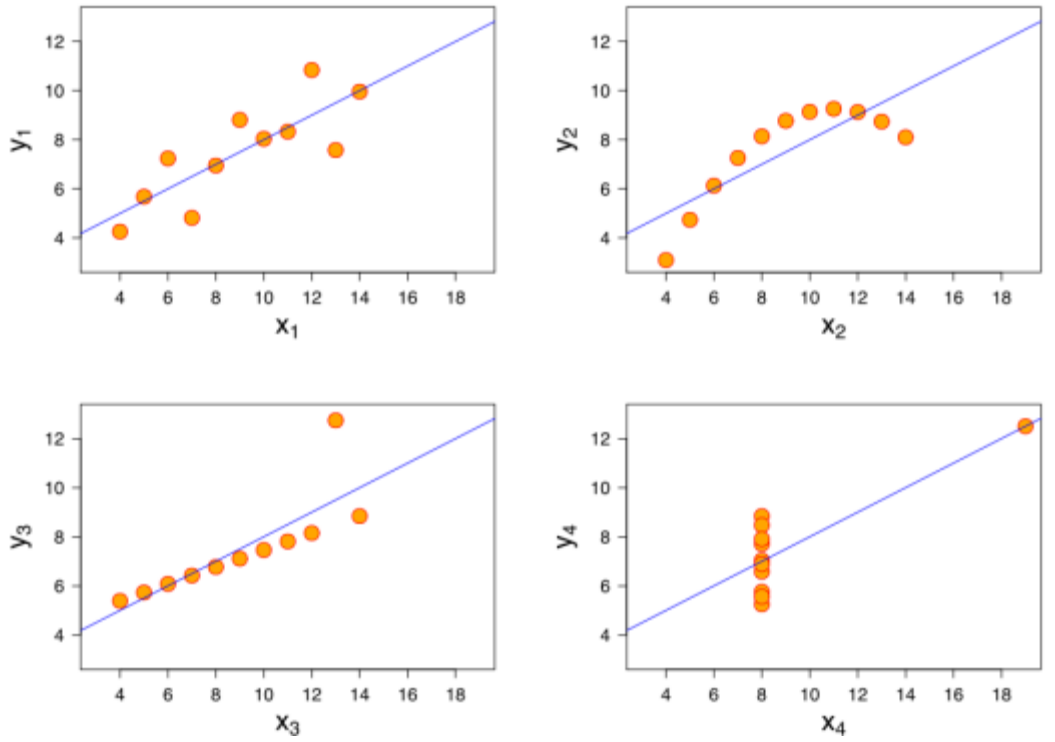
- So, as you can see. “Pi” is the probability of that event happening. So for Heads say, the probability of a Heads appearing is P(head)=5/13.

- And, we have a 0.96124 uncertainty as to whether the side will be a Heads or a Tails.

# Limitations of Summary Statistics

McMaster University
ENGINEERING
W Booth School of Engineering Practice and Technology

- On their own summary statistic can be misleading.
- Why not to trust statistics
- Amcomb's quartet:
  - Almost same means.
  - Almost same variances.
  - Almost same correlations.
  - Look completely different.
- Datasaurus Dozon



# Lecture Contents



- Exploratory Data Analysis
  - Summary Statistics
  - Data Visualization
- Data Preprocessing
  - Scaling, normalization and standardization
  - Data encoding
  - Handling missing data and imbalance data

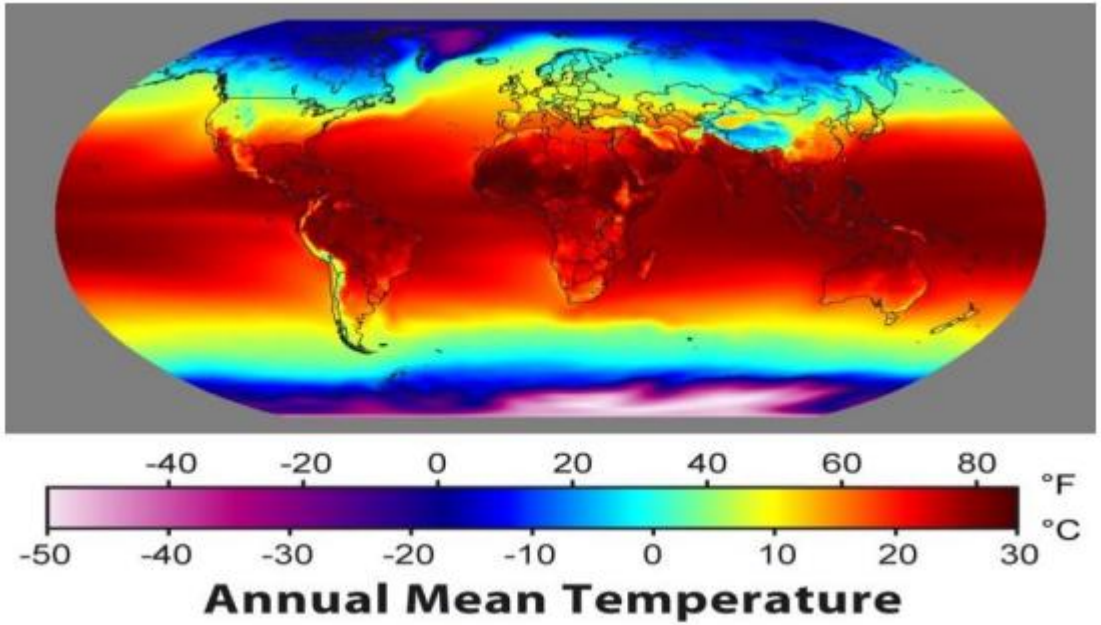
# Visualization



You can learn a lot from 2D plots of the data: – Patterns, trends, outliers, unexpected behavior.

| Lat | Long | Temp |
|---|---|---|
| 0 | 0 | 30.1 |
| 0 | 1 | 29.8 |
| 0 | 2 | 29.9 |
| 0 | 3 | 30.1 |
| 0 | 4 | 29.9 |
| … | … | … |

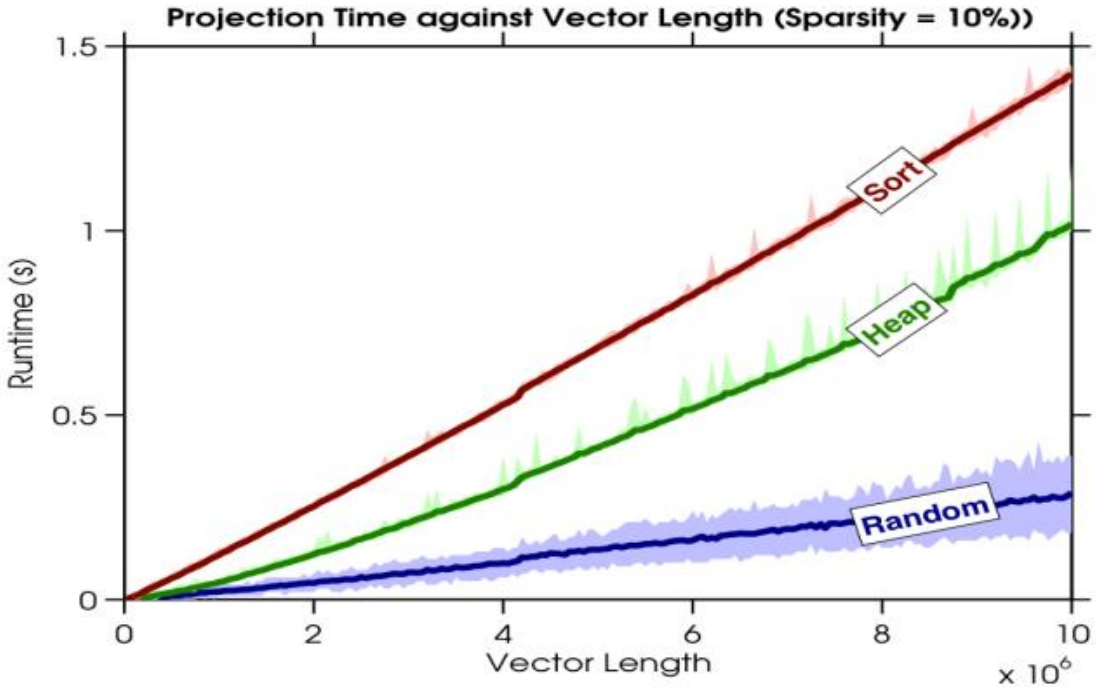
vs.

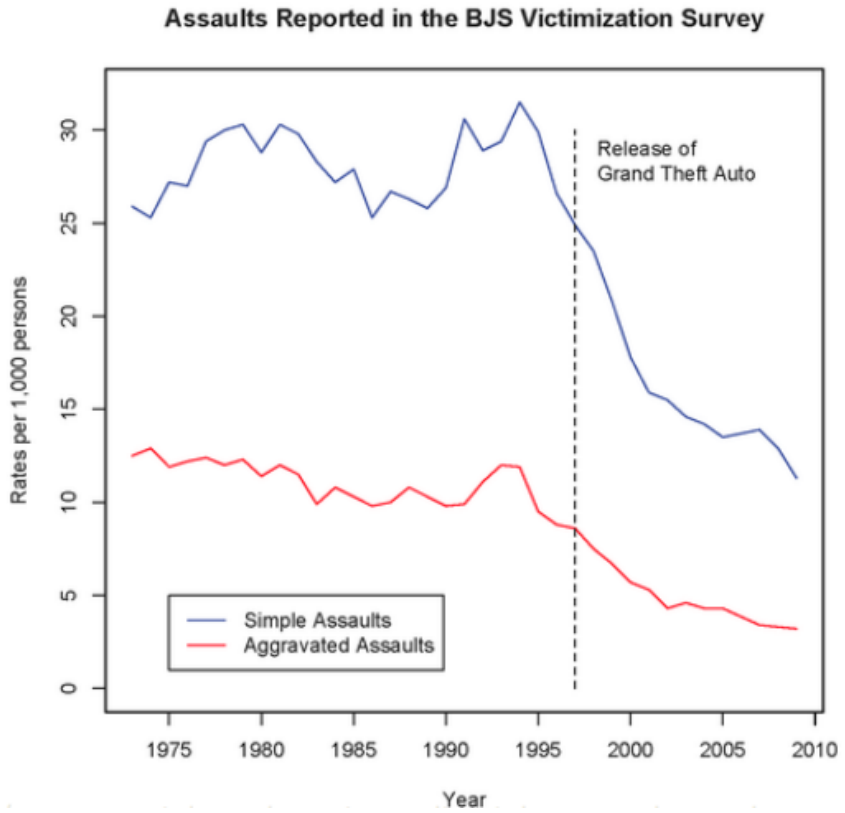


# Basic Line Plot

McMaster University
ENGINEERING
W Booth School of Engineering Practice and Technology

Visualize one variable as a function of another

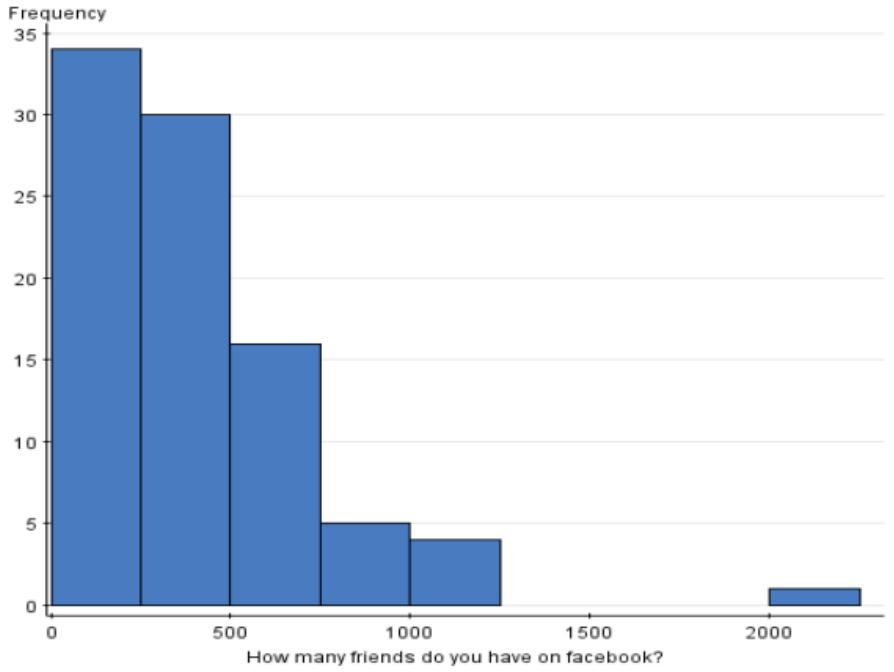


# Histogram

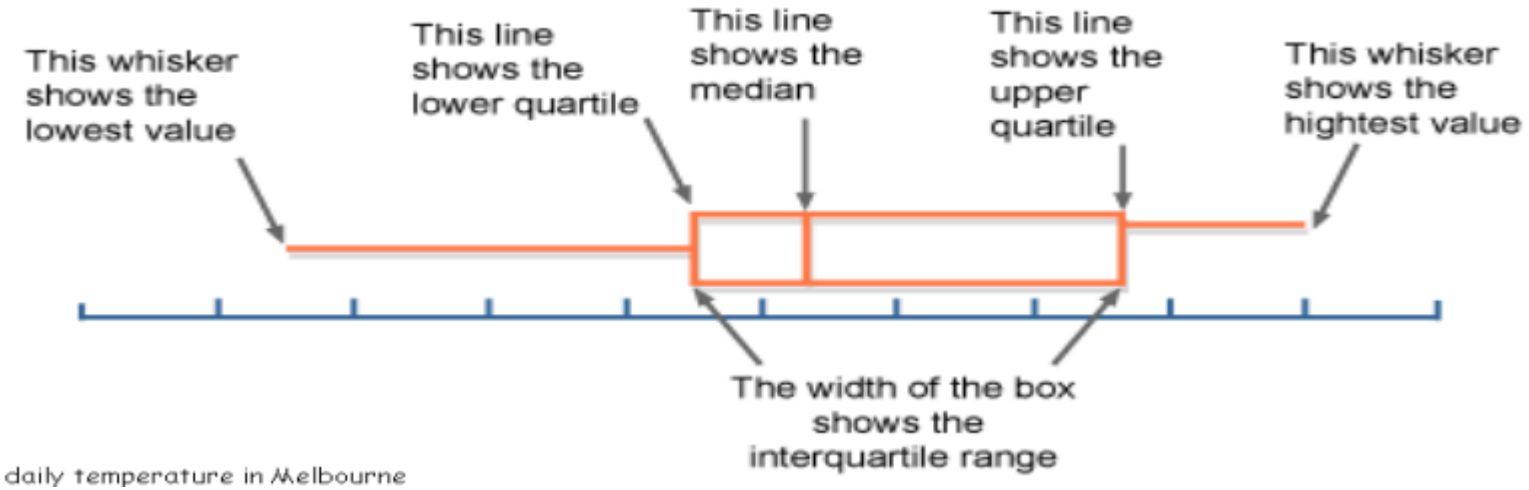


Histograms display counts a variable, split into “bins”

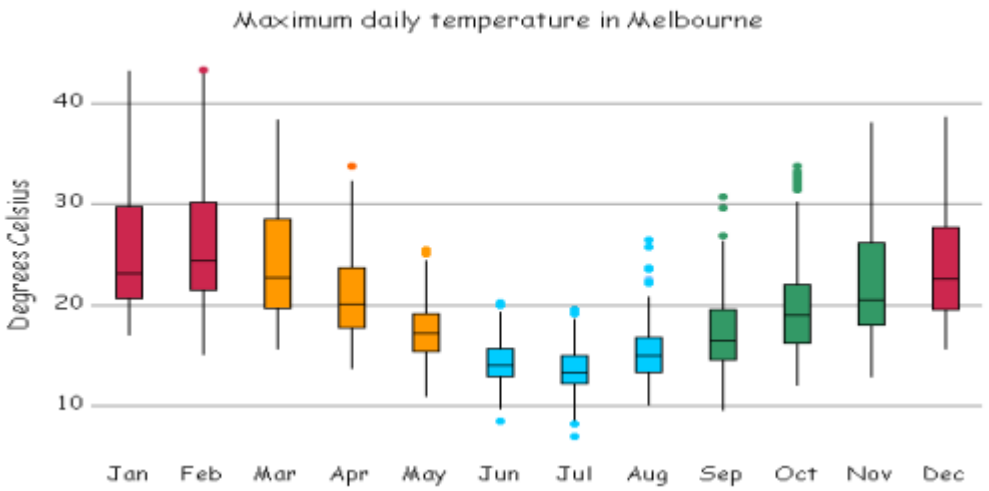


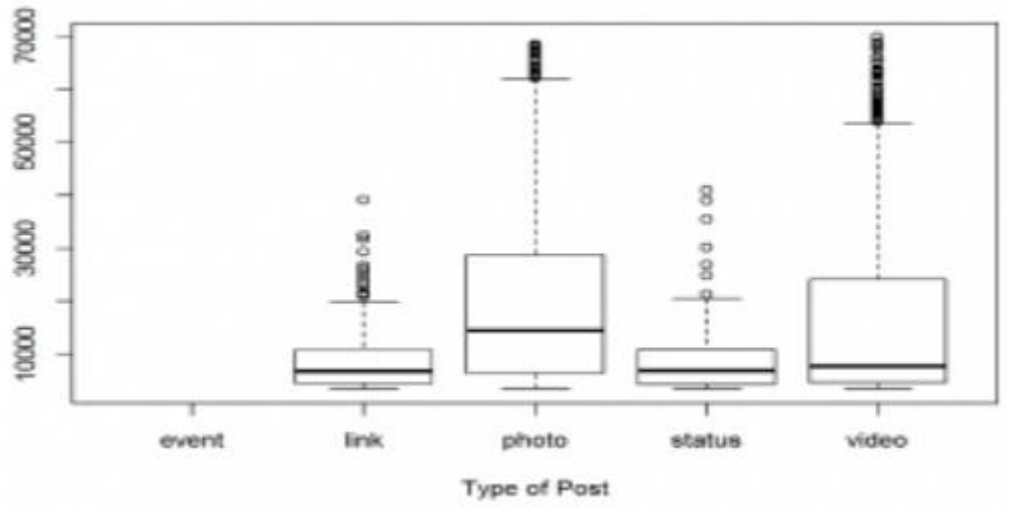
# Box Plot






num daily temperature in Melbourne



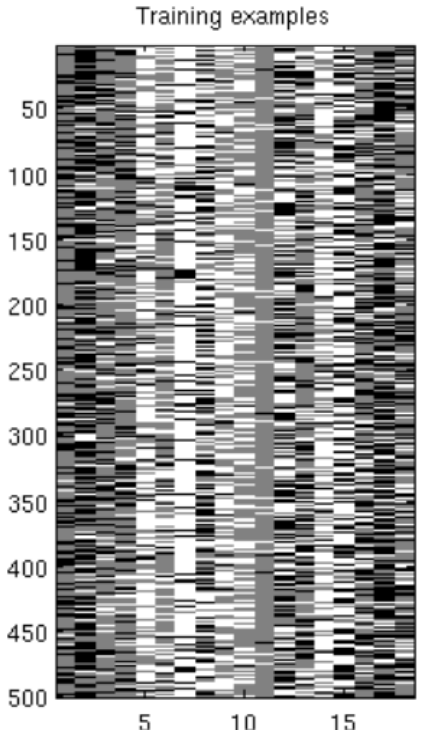
# Matrix Plot

McMaster University
ENGINEERING
W Booth School of Engineering Practice and Technology

A matrix plot of all similarities (or distances) between features: Color reflects distance/similarity.

| | BTC | ETH | XRP | XEM | ETC | LTC | DASH | XMR |
|---|---|---|---|---|---|---|---|---|
| BTC | 1.00 | 0.61 | 0.36 | 0.51 | 0.60 | 0.56 | 0.55 | 0.66 |
| ETH | 0.61 | 1.00 | 0.28 | 0.49 | 0.68 | 0.43 | 0.70 | 0.64 |
| XRP | 0.36 | 0.28 | 1.00 | 0.48 | 0.08 | 0.35 | 0.40 | 0.44 |
| XEM | 0.51 | 0.49 | 0.48 | 1.00 | 0.40 | 0.43 | 0.47 | 0.52 |
| ETC | 0.60 | 0.68 | 0.08 | 0.40 | 1.00 | 0.47 | 0.56 | 0.53 |
| LTC | 0.56 | 0.43 | 0.35 | 0.43 | 0.47 | 1.00 | 0.59 | 0.67 |
| DASH | 0.55 | 0.70 | 0.40 | 0.47 | 0.56 | 0.59 | 1.00 | 0.74 |
| XMR | 0.66 | 0.64 | 0.44 | 0.52 | 0.53 | 0.67 | 0.74 | 1.00 |

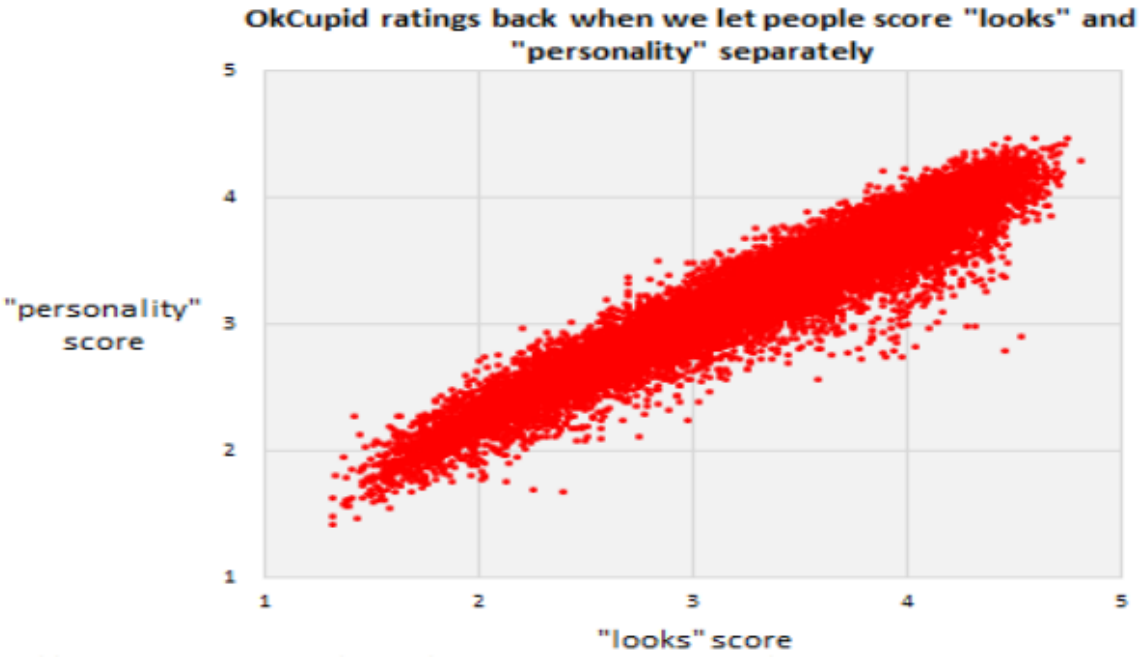
We can view (examples) x (features) data table as a picture to visualize trends



# Scatterplot



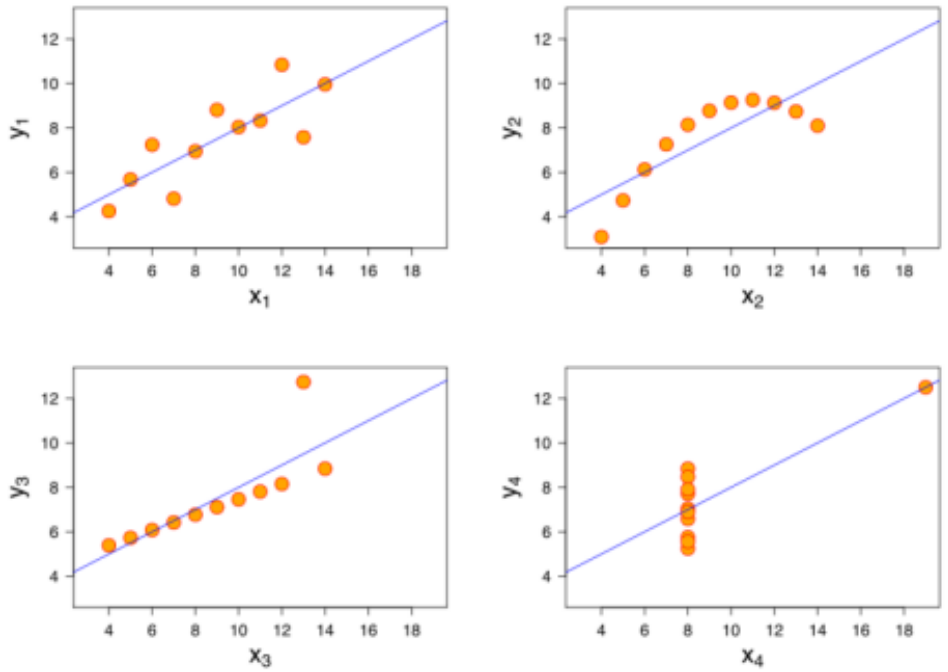
- Look at distribution of two features:
  - Feature 1 on x-axis.
  - Feature 2 on y-axis.
  - Basically a “plot without lines” between the points.
- Shows correlation between “personality” score and “looks” score.



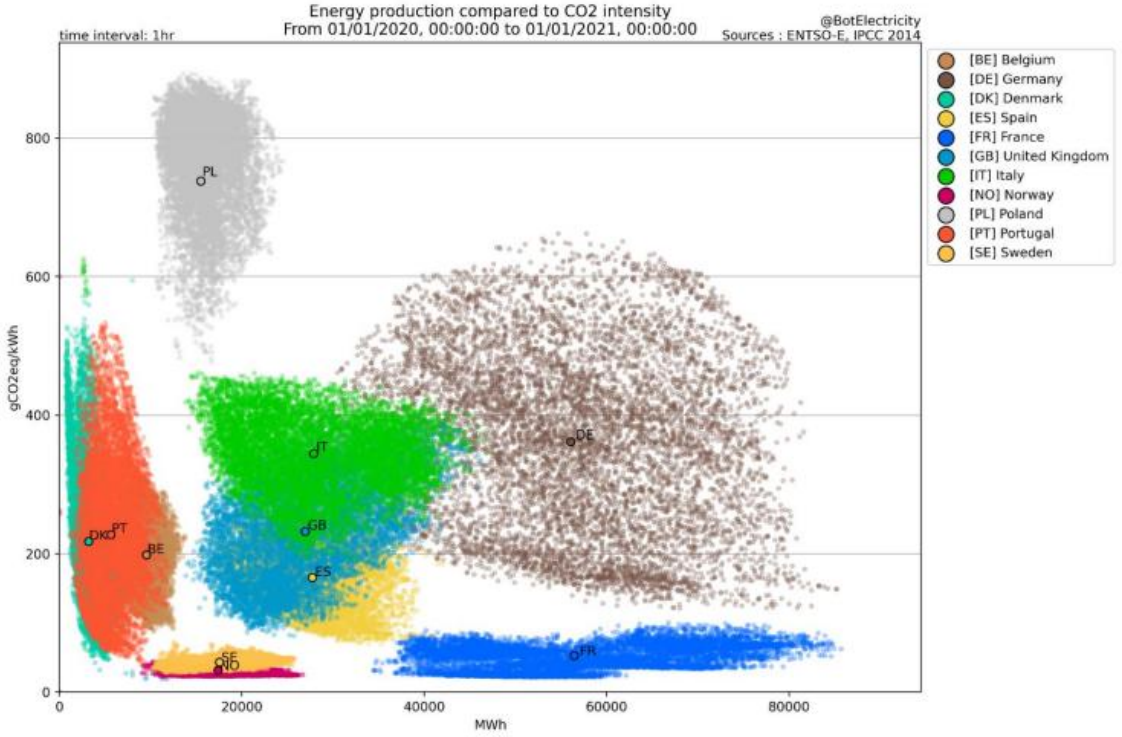
# Scatterplot



Shows correlation between “personality” score and “looks” score. - But scatterplots let you see more complicated patterns.




You can add color to display a 3rd variable:



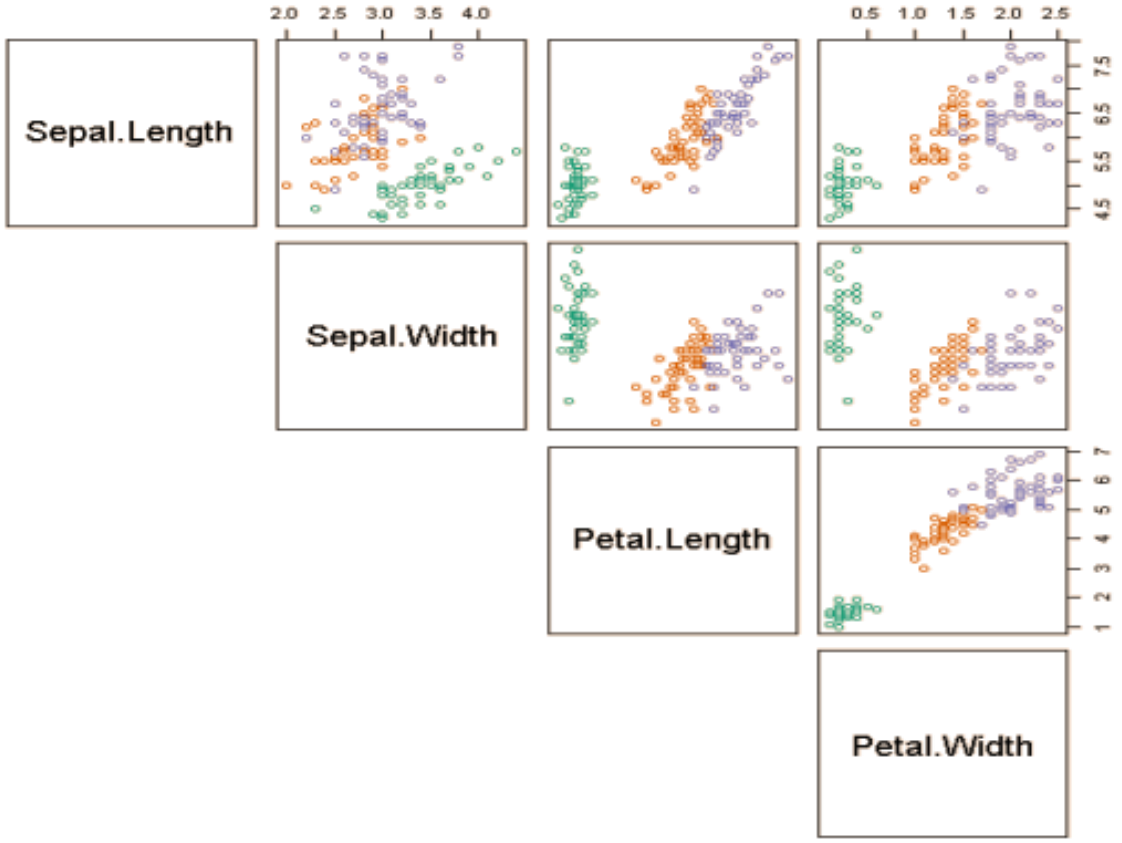
# Scatterplot Arrays

McMaster University
ENGINEERING
W Booth School of Engineering Practice and Technology

For multiple variables, you can use scatterplot arrays.

**Fisher's *Iris* Data** [hide]

| Sepal length | Sepal width | Petal length | Petal width | Species |
|---|---|---|---|---|
| 5.0 | 2.0 | 3.5 | 1.0 | *I. versicolor* |
| 6.0 | 2.2 | 4.0 | 1.0 | *I. versicolor* |
| 6.2 | 2.2 | 4.5 | 1.5 | *I. versicolor* |
| 6.0 | 2.2 | 5.0 | 1.5 | *I. virginica* |
| 4.5 | 2.3 | 1.3 | 0.3 | *I. setosa* |
| 5.0 | 2.3 | 3.3 | 1.0 | *I. versicolor* |
| 5.5 | 2.3 | 4.0 | 1.3 | *I. versicolor* |
| 6.3 | 2.3 | 4.4 | 1.3 | *I. versicolor* |
| 4.9 | 2.4 | 3.3 | 1.0 | *I. versicolor* |
| 5.5 | 2.4 | 3.7 | 1.0 | *I. versicolor* |
| 5.5 | 2.4 | 3.8 | 1.1 | *I. versicolor* |
| 5.1 | 2.5 | 3.0 | 1.1 | *I. versicolor* |



# Why Not to Trust Plots



- We've seen how summary statistics can be mis-leading.
- Note that plots can also be mis-leading, or can be used to mis-lead.

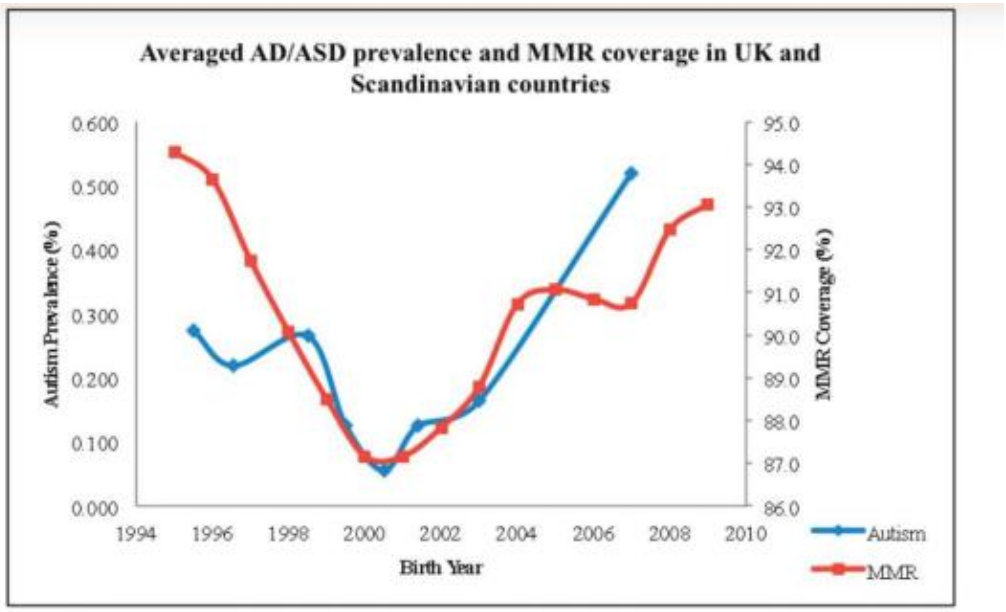
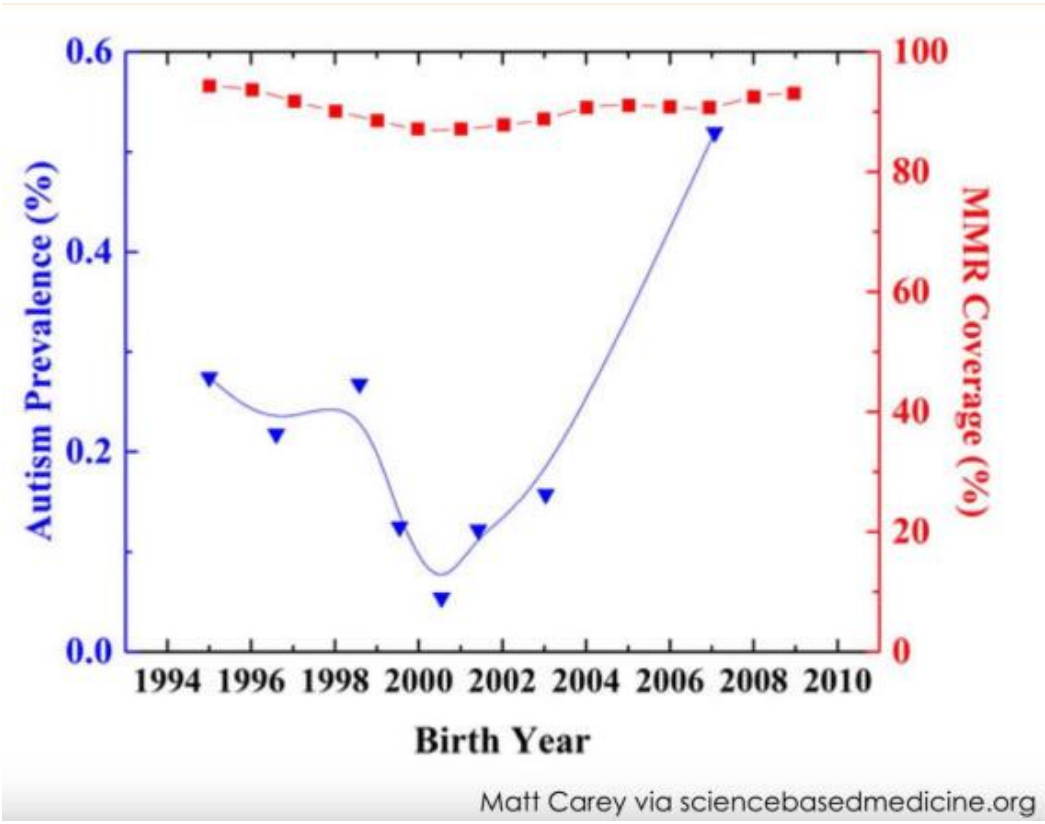



**Figure 1**-Averaged AD/ASD prevalence and MMR coverage in UK, Norway and Sweden. Both MMR and AD/ASD data are normalized to the maximum coverage/prevalence during the time period of this analysis.

Diesher *et al.* 2015 *Issues in Law and Medicine*




Matt Carey via sciencebasedmedicine.org

Mis-Leading Axes

# Lecture Contents

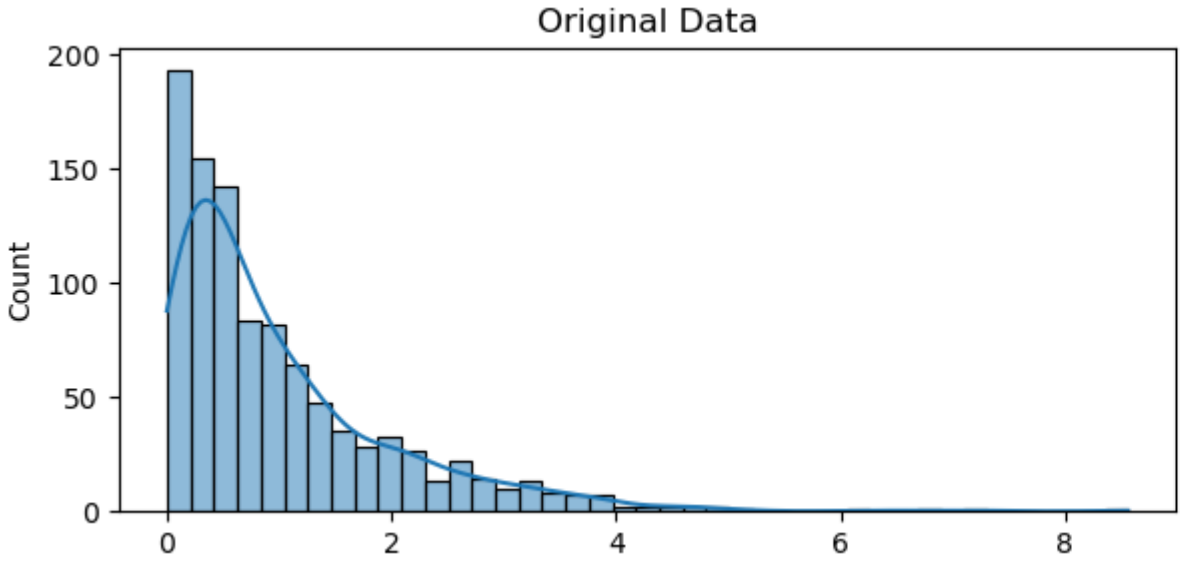


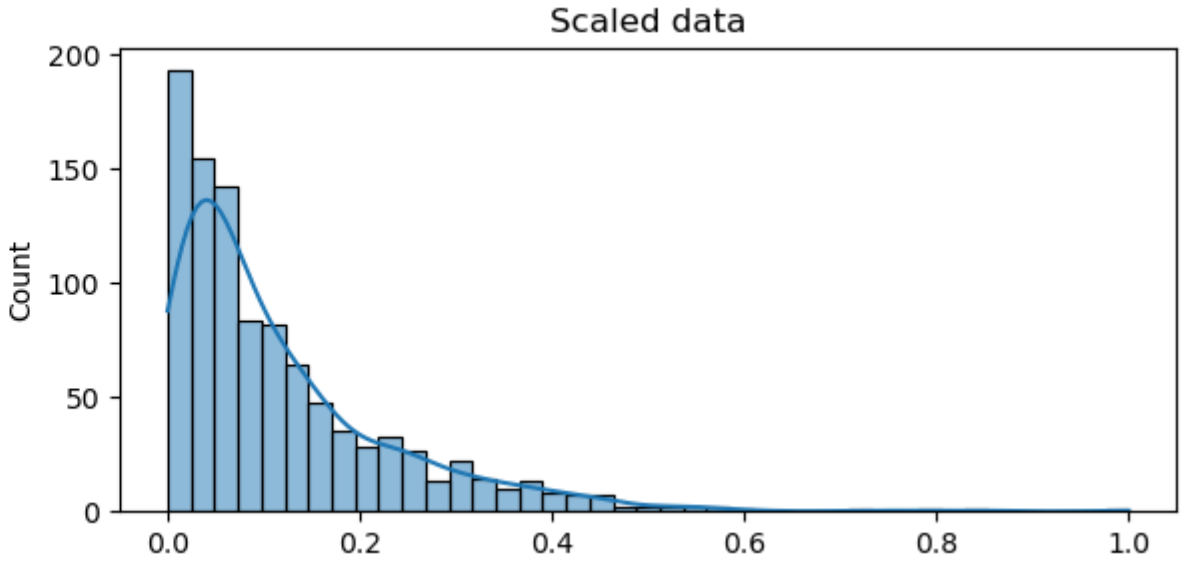
- Exploratory Data Analysis
  - Summary Statistics
  - Data Visualization
- Data Preprocessing
  - Scaling, normalization and standardization
  - Data encoding
  - Handling missing data and imbalance data

# Scaling



- In scaling, you're changing the range of your data
- Methods like KNN and SVM compute distances (e.g., Euclidean distance) to determine similarity or decision boundaries.
- Without scaling, features with larger ranges disproportionately affect these distances.

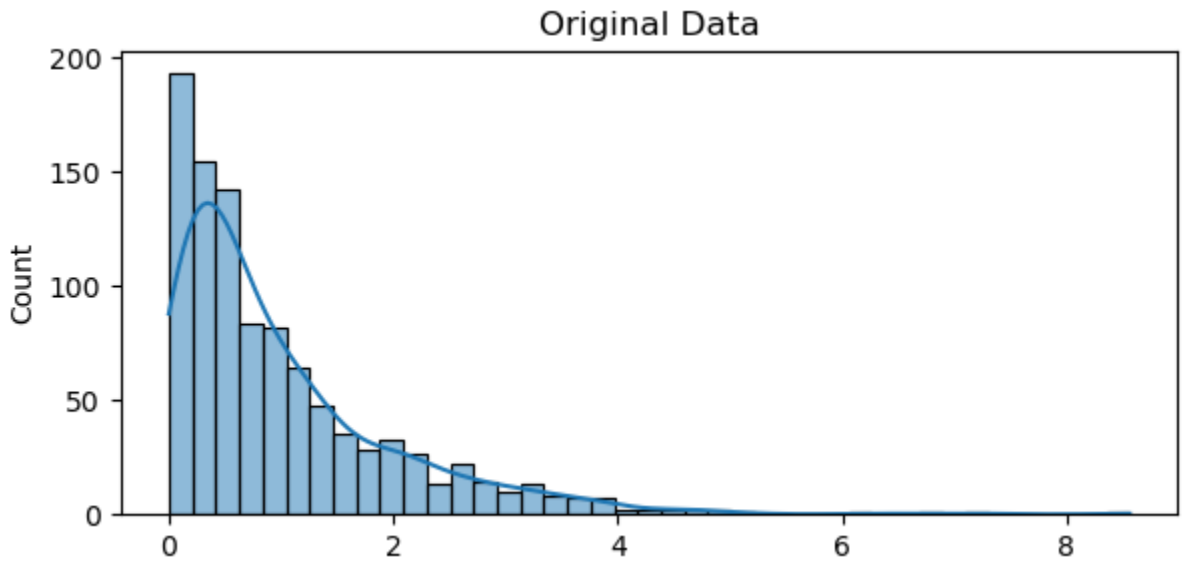


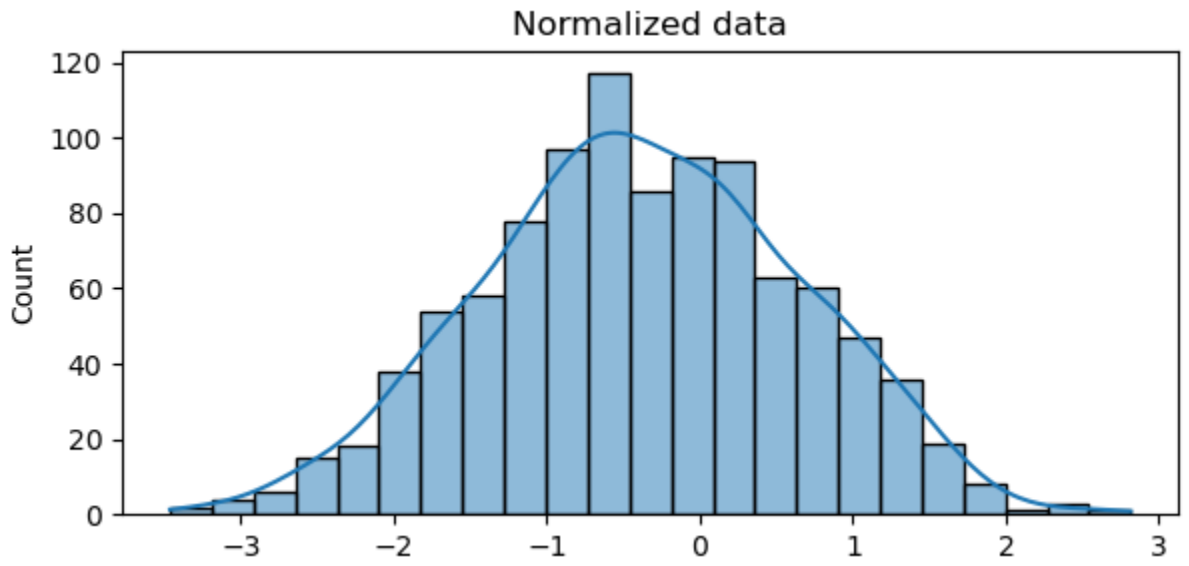



# Normalization

McMaster University
ENGINEERING
W Booth School of Engineering Practice and Technology

- In normalization, you're changing the shape of the distribution of your data.
- Scaling just changes the range of your data.
- The point of normalization is to change your observations so that they can be described as a normal distribution.

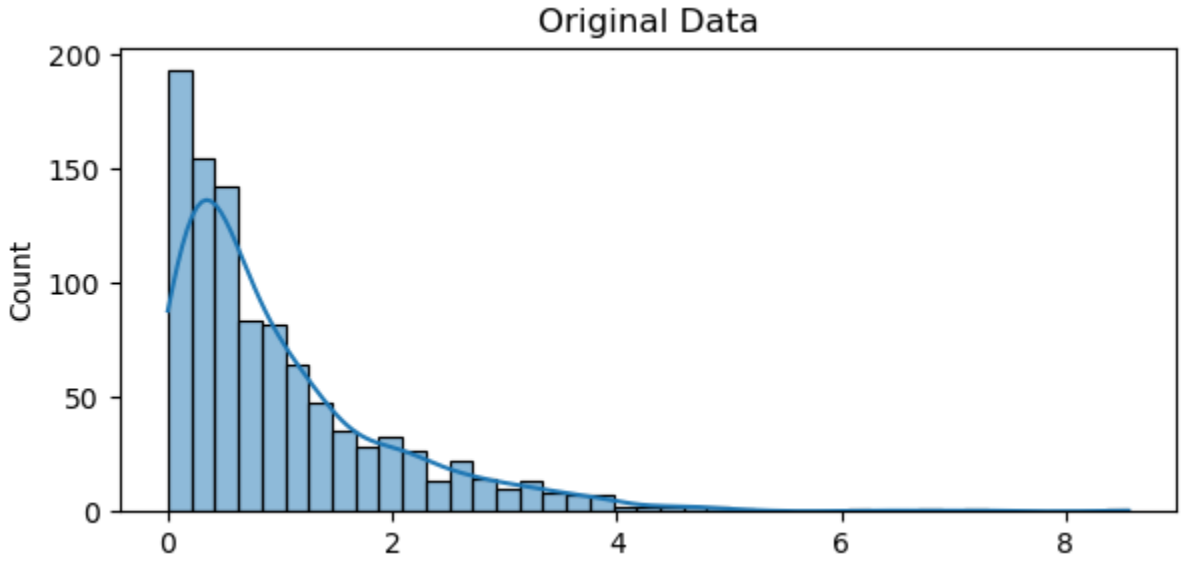


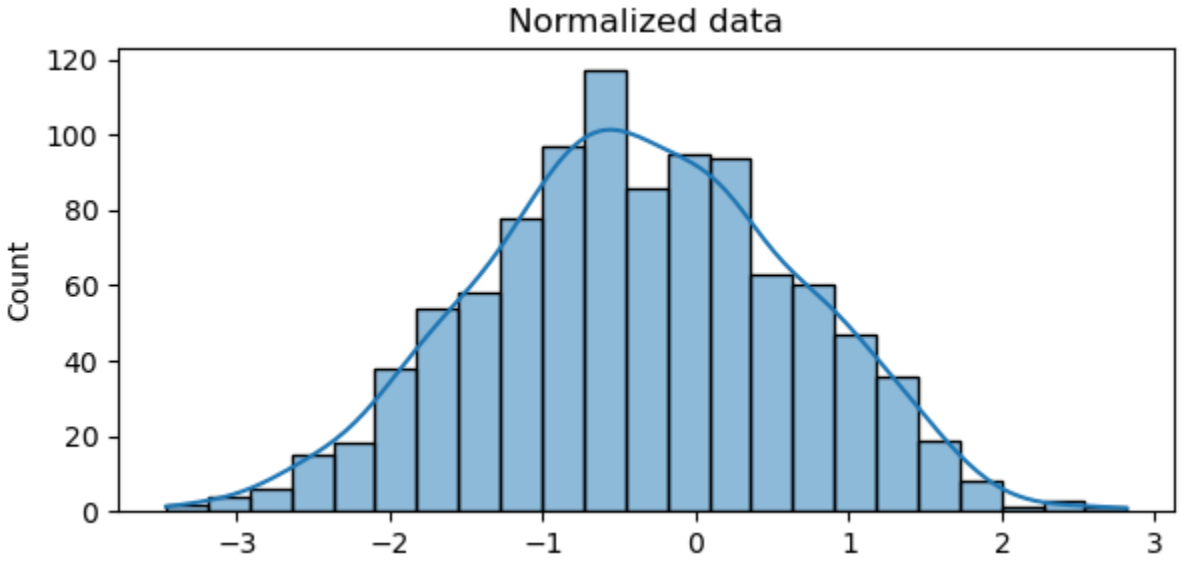



# Normalization



- In general, you'll normalize your data if you're going to be using a machine learning or statistics technique that assumes your data is normally distributed.
- Some examples of these include linear discriminant analysis (LDA) and Gaussian naive Bayes.

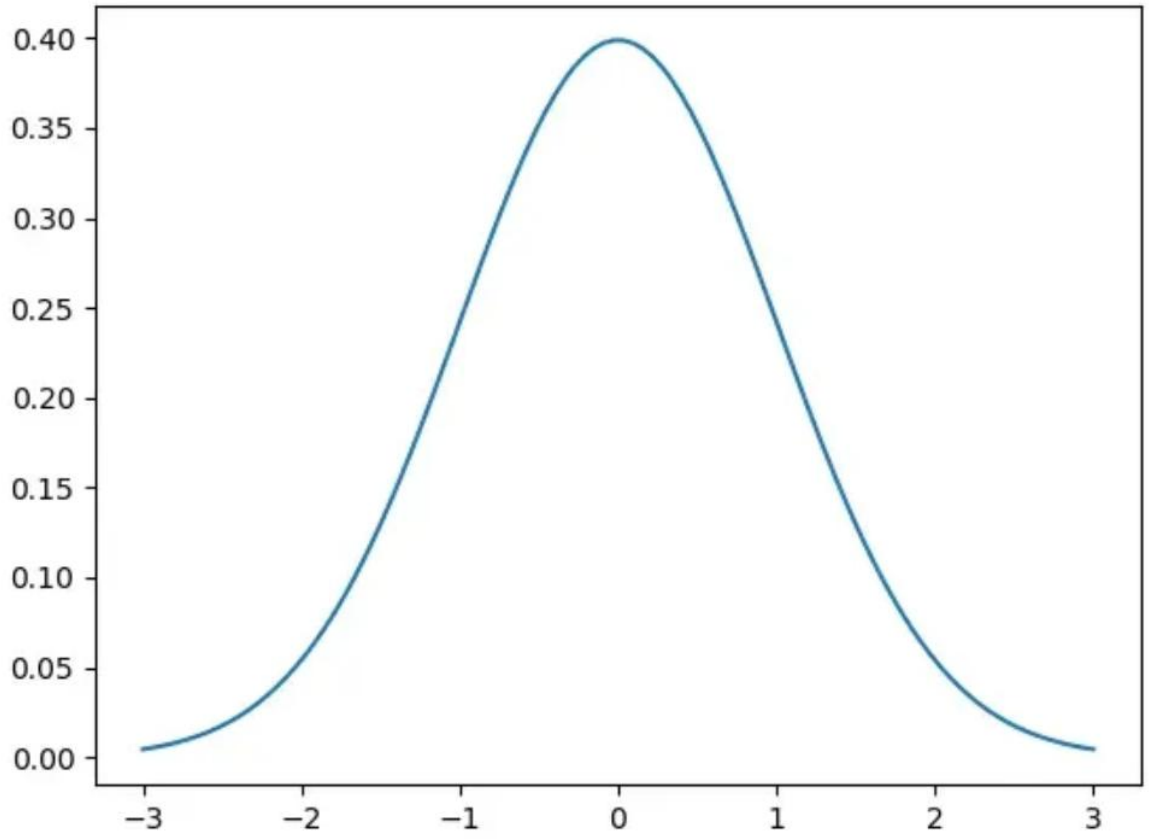


# Standardization



- Unlike Normalization, it is important to know that Standardization will make assumptions about your raw dataset:
- Your raw set of values display a Gaussian Distribution; and
- The Mean and Standard deviation are non-anomalous.
- If the dataset has **outliers**, these values can be distorted, leading to incorrect scaling.



# Normalization and Standardization Example



| # | Emp | Age | Salary |
|---|---|---|---|
| 1 | Emp1 | 44 | 73000 |
| 2 | Emp2 | 27 | 47000 |
| 3 | Emp3 | 30 | 53000 |
| 4 | Emp4 | 38 | 62000 |
| 5 | Emp5 | 40 | 57000 |
| 6 | Emp6 | 35 | 53000 |
| 7 | Emp7 | 48 | 78000 |

Age Range: 27-48

Salary Range: 47000-78000

# Normalization and Standardization Example



| # | Emp | Age | Salary |
|---|---|---|---|
| 1 | Emp1 | 44 | 73000 |
| 2 | Emp2 | 27 | 47000 |
| 3 | Emp3 | 30 | 53000 |
| 4 | Emp4 | 38 | 62000 |
| 5 | Emp5 | 40 | 57000 |
| 6 | Emp6 | 35 | 53000 |
| 7 | Emp7 | 48 | 78000 |

Use 47 and 73 not 47000 and 73000

Distance between Emp2 and Emp1 = $\sqrt{(27-44)^2+(47000-73000)^2} = 31.06$

Distance between Emp2 and Emp3 = $\sqrt{(30-27)^2+(53000-47000)^2} = 6.70$

# Normalization and Standardization Example



$$x_{\text{norm}} = \frac{x - \min(x)}{\max(x) - \min(x)}$$

| # | Emp | Age | Salary |
|---|---|---|---|
| 1 | Emp1 | 44 | 73000 |
| 2 | Emp2 | 27 | 47000 |
| 3 | Emp3 | 30 | 53000 |
| 4 | Emp4 | 38 | 62000 |
| 5 | Emp5 | 40 | 57000 |
| 6 | Emp6 | 35 | 53000 |
| 7 | Emp7 | 48 | 78000 |

Normalization

| Age | Normalized Age | Salary | Normalized Salary |
|---|---|---|---|
| 44 | 0.80952381 | 73000 | 0.838709677 |
| 27 | 0 | 47000 | 0 |
| 30 | 0.142857143 | 53000 | 0.193548387 |
| 38 | 0.523809524 | 62000 | 0.483870968 |
| 40 | 0.619047619 | 57000 | 0.322580645 |
| 35 | 0.380952381 | 53000 | 0.193548387 |
| 48 | 1 | 78000 | 1 |
| | **Range 0-1** | | **Range 0-1** |

How to calculate Normalized value?
X = 35, min = 27, max = 48 for column Age.
Xnorm(for 35) = $\frac{35-27}{48-27} = 0.3809$

**After Normalization**

Distance between Emp2 and Emp1 = $\sqrt{(0-.80)^2+(0-.83)^2} = 1.15$
Distance between Emp2 and Emp3 = $\sqrt{(.14-0)^2+(.19-0)^2} = 0.23$

**Comparison will be more significant**

# Normalization and Standardization Example



$$x_{\text{stand}} = \frac{x - \text{mean}(x)}{\text{standard deviation } (x)}$$

| # | Emp | Age | Salary |
|---|---|---|---|
| 1 | Emp1 | 44 | 73000 |
| 2 | Emp2 | 27 | 47000 |
| 3 | Emp3 | 30 | 53000 |
| 4 | Emp4 | 38 | 62000 |
| 5 | Emp5 | 40 | 57000 |
| 6 | Emp6 | 35 | 53000 |
| 7 | Emp7 | 48 | 78000 |
| | | Mean = 37.42857 | Mean = 60428.5714 |
| | | Std. Dev. = 6.883876 | Std. Dev = 10499.7570 |

Standardization

How to calculate Standardized value?
X = 35, mean = 37.42, Std. Dev. = 6.88
for column Age.
$X_{std}(\text{for } 35) = \frac{35 - 37.42}{6.88} = -0.3527$

| Age | Standardized Age | Salary | Standardized Salary |
|---|---|---|---|
| 44 | 0.954611636 | 73000 | 1.197306616 |
| 27 | -1.514927162 | 47000 | -1.278941158 |
| 30 | -1.079126198 | 53000 | -0.707499364 |
| 38 | 0.083009708 | 62000 | 0.149663327 |
| 40 | 0.373543684 | 57000 | -0.326538168 |
| 35 | -0.352791257 | 53000 | -0.707499364 |
| 48 | 1.535679589 | 78000 | 1.673508111 |
| | **Mean = 0 Std. dev. = 1** | | **Mean = 0 Std. dev. = 1** |

**After Standardization**

Distance between Emp2 and Emp1 = $\sqrt{(-1.51 - 0.95)^2 + (-1.27 - 1.19)^2} = 3.47$

Distance between Emp2 and Emp3 = $\sqrt{(-1.07 + 1.51)^2 + (-0.70 + 1.27)^2} = 0.71$

**Comparison will be more significant**

# Lecture Contents



- Exploratory Data Analysis
  - Summary Statistics
  - Data Visualization
- Data Preprocessing
  - Scaling, normalization and standardization
  - Data encoding
  - Handling missing data and imbalance data

# Encoding



- Encoding categorical data is an essential step in preparing data for machine learning models.
- Choosing the right encoding technique depends on the type of categorical data and the model's requirements:
  - Label Encoding is suitable for nominal data with no order.
  - One-Hot Encoding works best for nominal data where categories have no ranking.
  - Ordinal Encoding preserves the order of ordinal data.
  - Target Encoding is effective when there's a relationship between the categorical feature and the target variable.
  - Frequency Encoding is useful for handling high-cardinality features.

# One-Hot Encoding



- The idea behind one-hot encoding is to represent each category as a binary vector. Here's how it works:

| Index | Animal |
|---|---|
| 0 | Dog |
| 1 | Cat |
| 2 | Sheep |
| 3 | Horse |
| 4 | Lion |

One-Hot code

| Index | Dog | Cat | Sheep | Lion | Horse |
|---|---|---|---|---|---|
| 0 | 1 | 0 | 0 | 0 | 0 |
| 1 | 0 | 1 | 0 | 0 | 0 |
| 2 | 0 | 0 | 1 | 0 | 0 |
| 3 | 0 | 0 | 0 | 0 | 1 |
| 4 | 0 | 0 | 0 | 1 | 0 |

# Label Encoding

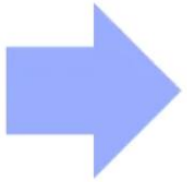


- Label Encoding assigns a unique integer to each category. However, it does not respect the order of the categories, making it more suitable for nominal data where the order doesn’t matter.

Original Data:

| Color | Size | Price |
|---|---|---|
| Blue | L | 100 |
| Green | M | 150 |
| Red | S | 200 |
| Green | XL | 120 |
| Red | M | 180 |

Label Encoding



Label Encoded Data:

| Color | Size | Price |
|---|---|---|
| 0 | 0 | 100 |
| 1 | 1 | 150 |
| 2 | 2 | 200 |
| 1 | 3 | 120 |
| 2 | 1 | 180 |

# Ordinal Encoding



- Ordinal Encoding is used for ordinal data, where categories have a natural order. It converts categorical values into numeric values, preserving the inherent order.

| Original Encoding | Ordinal Encoding |
|---|---|
| Poor | 1 |
| Good | 2 |
| Very Good | 3 |
| Excellent | 4 |

# Target Encoding



- A target encoding is any kind of encoding that replaces a feature's categories with some number derived from the target.
- This kind of target encoding is sometimes called a mean encoding.

| Animal Type | Year 1 Cost | Year 2 Cost | Year 3 Cost |
|---|---|---|---|
| Cat | $80 | $90 | $110 |
| Dog | $210 | $200 | $190 |
| Bird | $40 | $50 | $60 |

| Animal Type | Year 1 Cost | Year 2 Cost | Year 3 Cost |
|---|---|---|---|
| $93.33 | $80 | $90 | $110 |
| $200 | $210 | $200 | $190 |
| $50 | $40 | $50 | $60 |

# Frequency Encoding



- It gives each category a number based on how often it appears in the data.
- The numbers add up to 1.

| Animal Type | Frequency Encoding |
|---|---|
| Dog | 0.6 |
| Cat | 0.3 |
| Bird | 0.1 |

# Lecture Contents



- Exploratory Data Analysis
  - Summary Statistics
  - Data Visualization
- Data Preprocessing
  - Scaling, normalization and standardization
  - Data encoding
  - Handling missing data and imbalance data

# Handling missing data




| PassengerId | Survived | Pclass | Sex | Age | SibSp | Parch | Ticket | Fare | Cabin | Embarked |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 0 | 3 | male | 22 | 1 | 0 | A/5 21171 | 7.25 | | S |
| 2 | 1 | 1 | female | 38 | 1 | 0 | PC 17599 | 71.2833 | C85 | C |
| 3 | 1 | 3 | female | 26 | 0 | 0 | STON/O2. 3101282 | 7.925 | | S |
| 4 | 1 | 1 | female | 35 | 1 | 0 | 113803 | 53.1 | C123 | S |
| 5 | 0 | 3 | male | 35 | 0 | 0 | 373450 | 8.05 | | S |
| 6 | 0 | 3 | male | | 0 | 0 | 330877 | 8.4583 | | Q |

| | Height | Weight | Country | Place | Number of days | Some column |
|---|---|---|---|---|---|---|
| 0 | 12.0 | 35.0 | India | Bengaluru | 1.0 | NaN |
| 1 | NaN | 36.0 | US | New York | 2.0 | NaN |
| 2 | 13.0 | 32.0 | UK | London | NaN | NaN |
| 3 | 15.0 | NaN | France | Paris | 4.0 | NaN |
| 4 | 16.0 | 39.0 | US | California | 5.0 | 12.0 |
| 5 | NaN | NaN | NaN | Mumbai | NaN | NaN |
| 6 | NaN | NaN | NaN | NaN | 6.0 | NaN |

**NaN (Not a Number)**
**NULL or None:** if data is coming from SQL
**Empty Strings**:
**Empty Strings**: -999, 9999
**Blanks or Spaces**:

# How to Handle Missing Data



- **Deletion**: This involves removing rows or columns with missing values. This is a straightforward method, but it can be problematic if a significant portion of your data is missing. Discarding too much data can affect the reliability of your conclusions.
- **Imputation**: This replaces missing values with estimates. There are various imputation techniques, each with its strengths and weaknesses. Here are some common ones:
  - Mean/Median/Mode Imputation
  - K-Nearest Neighbors (KNN Imputation)
  - Model-based Imputation
- Interpolation: interpolation estimates the value of missing values based on the surrounding trends and patterns. This approach is more feasible to use when your missing values are not scattered too much.

# Handling imbalanced data



- Imbalanced data refers to datasets where the target class has an uneven distribution of observations, i.e., one class label has a very high number of observations, and the other has a deficient number of observations.

# Handling imbalanced data

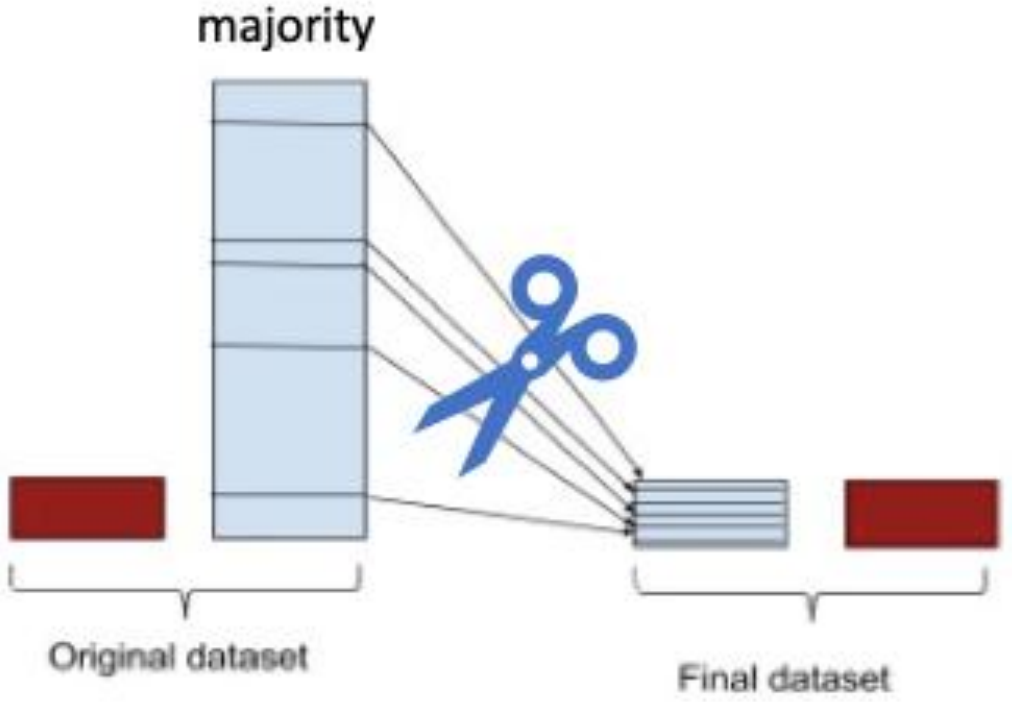


- There are quite a few ways to handle imbalanced data in machine classification problems.
  - Random under-sampling
  - Random over-sampling
  - Synthetic over-sampling: SMOTE
  - Choose the algorithm wisely
  - Play with the loss function
  - Solve an anomaly detection problem

# Random under-sampling



- This approach is generally used when you have a huge amount of training data with you. The random under-sampling technique works by randomly eliminating the samples from the majority class until the classes are balanced in the remaining dataset.

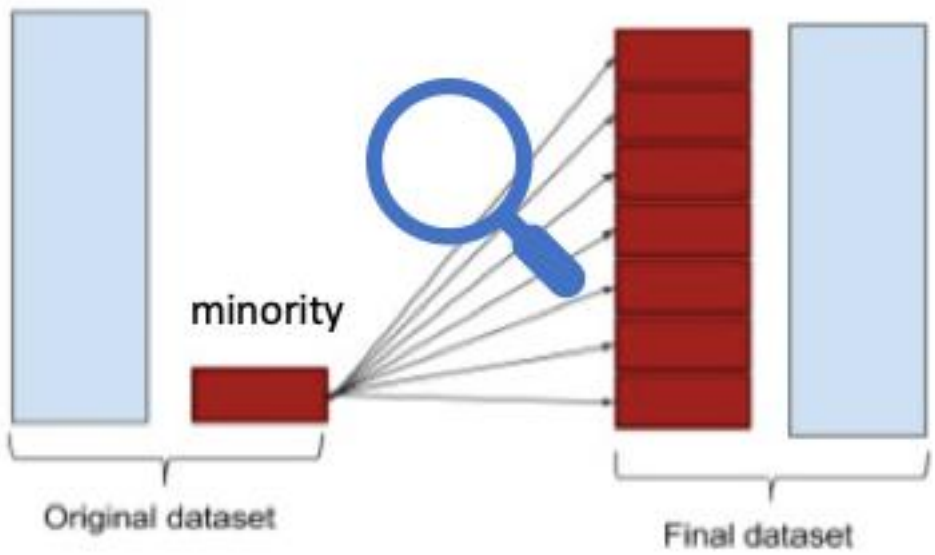



https://medium.com/dataman-in-ai/sampling-techniques-for-extremely-imbalanced-data-part-i-under-sampling-a8dbc3d8d6d8

# Random over-sampling



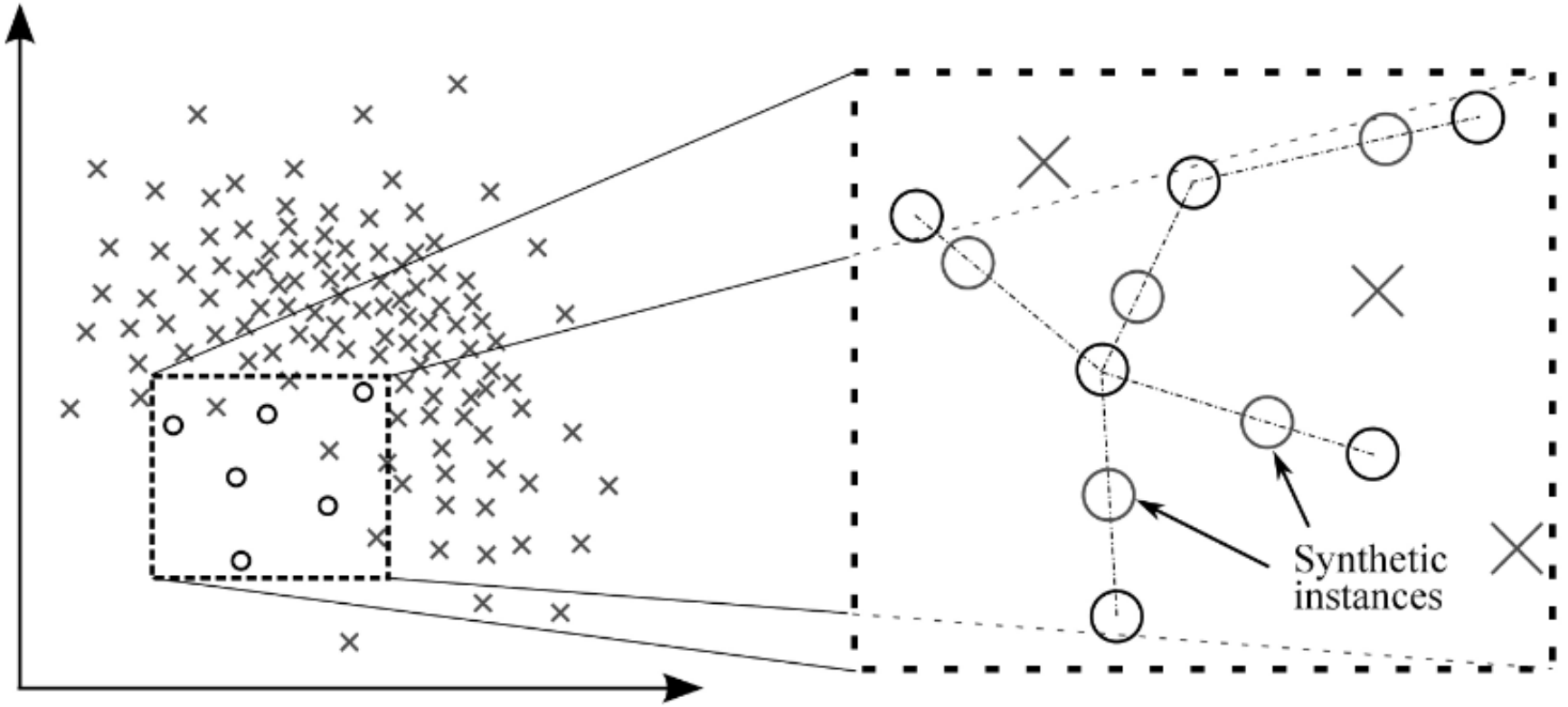
- In this technique, we try to increase the instances of the minority class by random replication of the already present samples.
- This technique is better than under-sampling as there is no information loss here and it outperforms the under-sampling in practice.



# Synthetic over-sampling: SMOTE



- In SMOTE, a subset of minority class is taken, and new synthetic data points are generated based on it. These synthetic data points are then added to the original training dataset as additional examples of the minority class.



# References and Reading




- https://ocw.mit.edu/courses/15-062-data-mining-spring-2003/pages/lecture-notes/
- https://openlearninglibrary.mit.edu/courses/course-v1:MITx+HST.953x+3T2020/courseware/e1d2de6025b742a68bb0ca2da4106eb9/300ac6a981114fe8a7b8a46dae429009/?child=first
- https://www.cs.ubc.ca/~schmidtm/Courses/340-F22/L2.pdf

McMaster University ENGINEERING

W BOOTH SCHOOL OF ENGINEERING PRACTICE AND TECHNOLOGY

# SEP 785: Machine Learning

**Lecture 2:** Data Mining

Instructor: Dr. Dalia Mahmoud, PhD
(Mechanical Engineering, McMaster University )
Email: mahmoudd@mcmaster.ca

Thank you !!
Questions ??