# CDS6214

## Data Science Fundamentals

**Lecture 4**

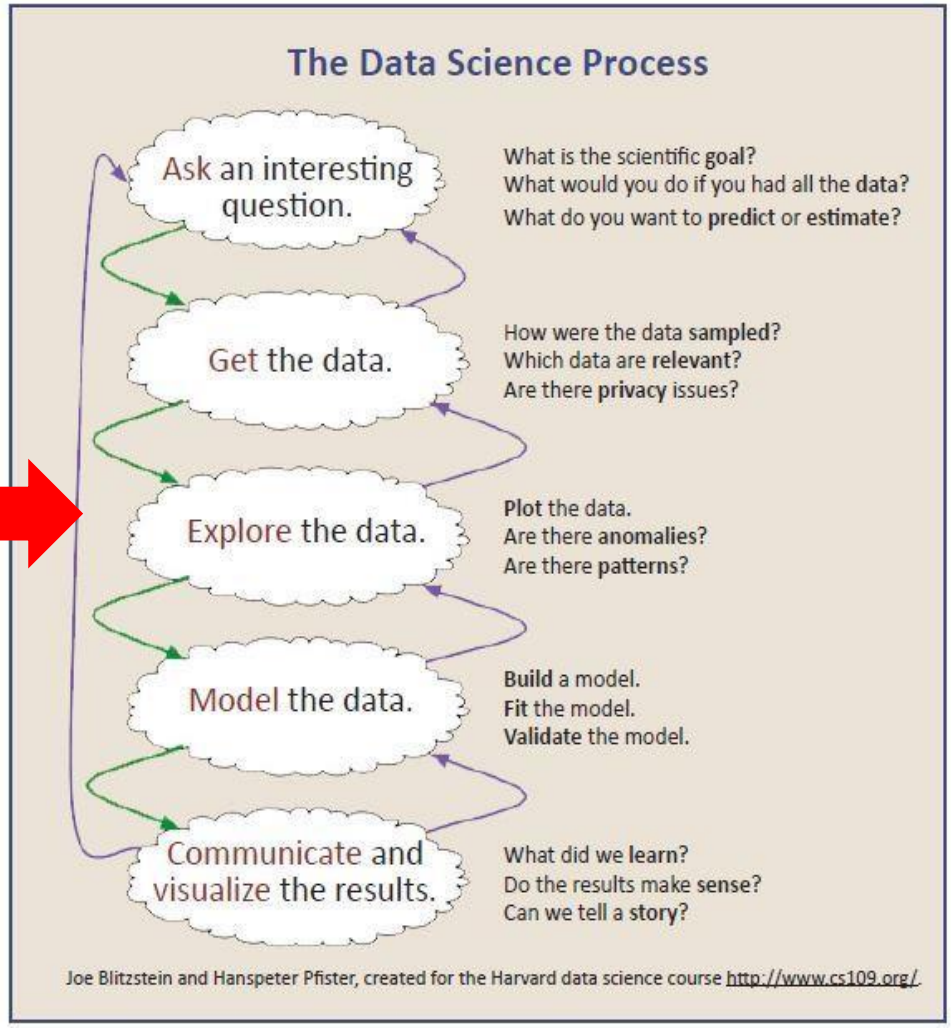
**Exploratory Data Analysis**

# Outline

- **Exploratory Data Analysis (EDA)**
  - Data Types (Numeric, Categorical)
  - Statistical Description of Data
    - Measures of Central Tendency
    - Measures of Dispersion of Data
    - Outliers
    - Graphical Representation of Data
  - Relationship between Attributes
    - Correlation Analysis
    - Simpson's Paradox (Correlation vs. Causality)

# exploratory data analysis

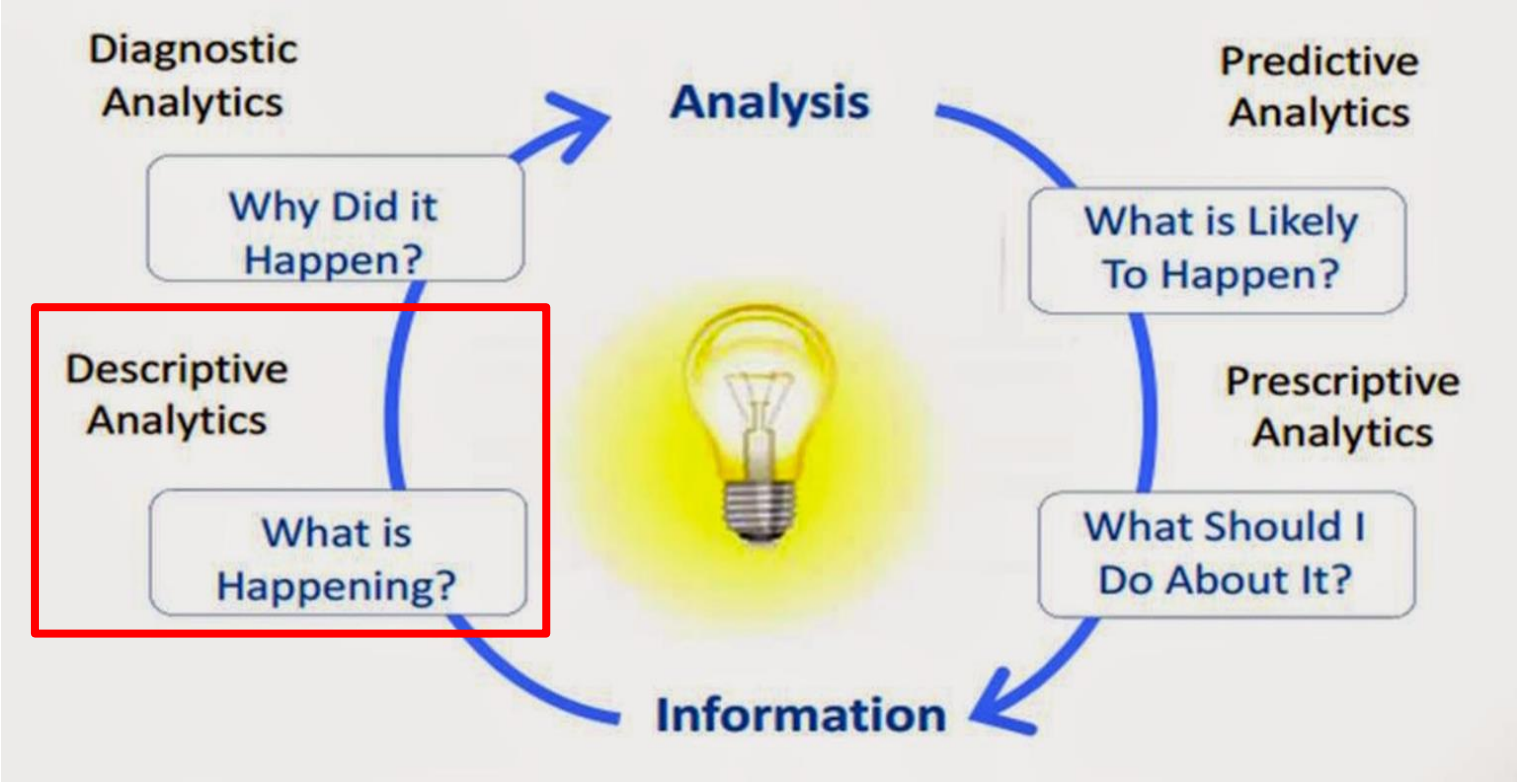
# Data Science Process




**What should we do next?**

- Identify the question
- Collect and pre-process the data
- Explore and analyze the data
- Model the data
- Infer and visualize results

# 4 Types of Analytics




Source: Gartner

# Exploratory Data Analysis

- To begin analysing data, we always start with the EDA step, and by examining some conditions (or properties) of the data that we have identified:

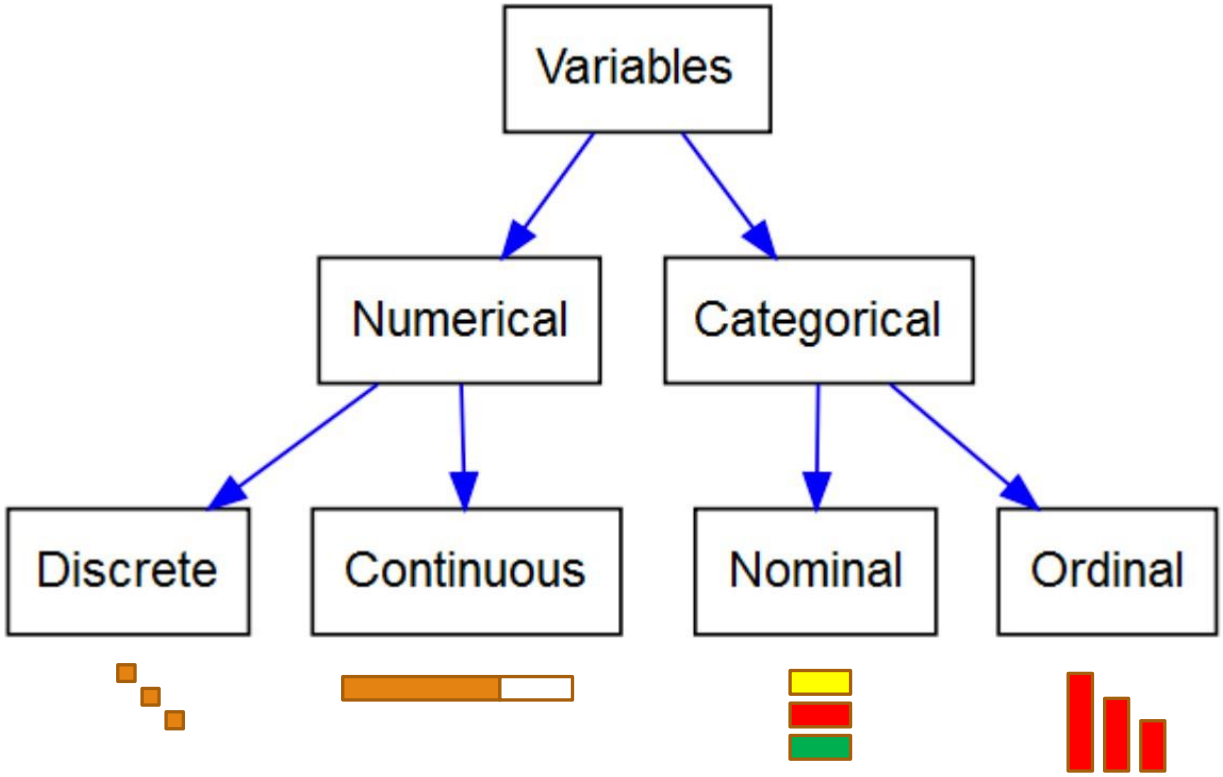

1. The data types of columns and the granularity of rows in the dataset.
2. The distributions of quantitative data and measures of center and spread.
3. Relationships between quantities in the dataset.

- Exploration can be done numerically and/or graphically – **descriptive analytics**

# Dimensionality of Data

- **Univariate** data: Measurement made on one variable per subject
- **Bivariate** data: Measurement made on two variables per subject
- **Multivariate** data: Measurement made on many variables per subject

# Variables/Data Types



# Data Types

- **Numerical data**
  - ***Discrete data***
    - Values are distinct and separate, i.e. they can be counted e.g. number of staff
  - ***Continuous data***
    - Values may take on any value within a finite or infinite interval e.g height and weight
- **Categorical data**
  - ***Nominal data***
    - Values can be assigned a code in the form of a number, where the numbers are simply labels e.g. gender: male, female
  - ***Ordinal data***
    - Values can be ranked or have a rating scale attached e.g. very unsatisfied, unsatisfied, neutral, satisfied, very satisfied.

# Categorical Data: Nominal

- **Nominal** attributes are attributes that are "related to names"
- Possible nominal values are symbols or names of things
- Values represent categories or codes (symbolic of the categories) but do not have any meaningful order
- Most of the time, categorical values are referring to nominal attributes.
- Can you give a few examples of nominal attributes?

# Categorical Data: Binary

- **Binary** attributes is a special case of **nominal** attributes, such that there are only two possible categories or "states"
  - Yes or No
  - 1 or 0
  - True or False
- Known as Boolean attributes for the case of True / False
- A binary attribute is symmetric if both states are equally valuable and carry the same weight
- A binary attribute is asymmetric if the outcome of the states are not equally important

# Categorical Data: Ordinal

- **Ordinal** attributes contains possible value that have a meaningful order or ranking among them but the magnitude between successive values are not known
- Example: Size of French Fries packs from McDonald's

- **Ordinal** attributes are useful
  - For registering subjective qualities that cannot be measure objectively. Common in survey for ratings.
  - Can be obtained from discretization of numeric quantities by range splitting

# Categorical Data

- Overall, **nominal, binary** and **ordinal** attribute are **qualitative** ⇒ they describe a feature of an object without giving the actual size or quantity
- The value of qualitative attributes are typically words representing categories
  - Attributes can be abbreviated or symbolically mapped to some discrete values if the attribute names are too long)
  - **Even if integers are used**, they are **NOT** meant as measurable quantities

# Numerical Data

- Numerical attributes are quantitative (i.e. measurable in quantity)
- Integer values (discrete) in two forms:

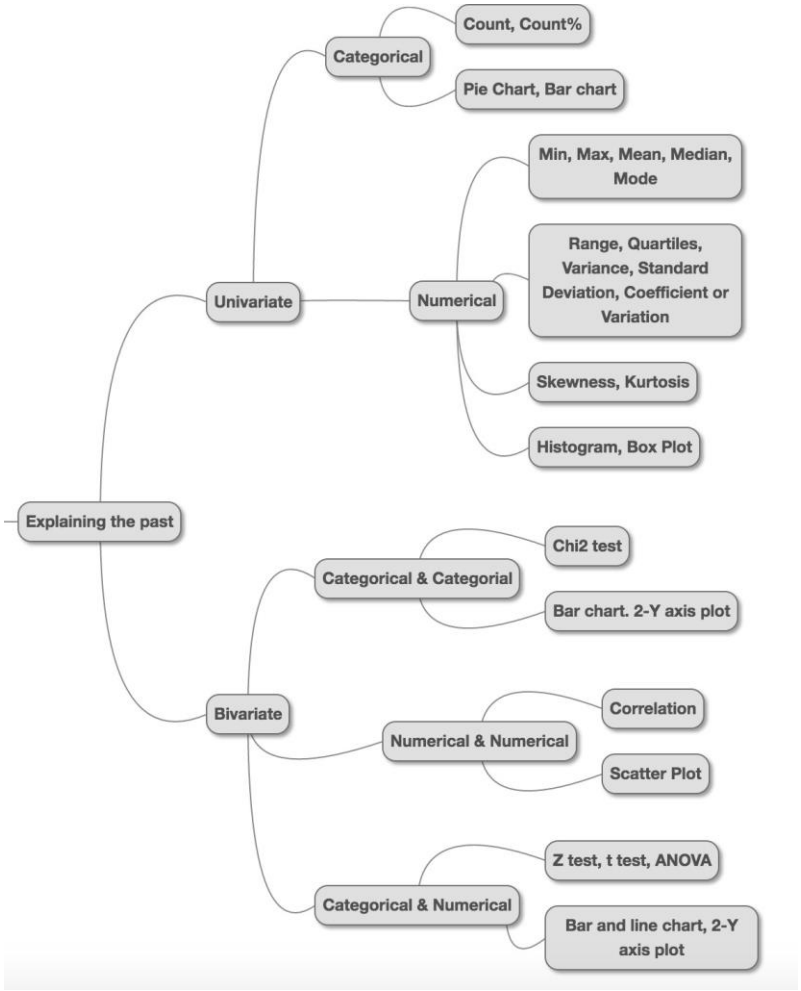
| Interval-Scaled Attributes | Ratio-Scaled Attributes |
|---|---|
| Measured on a scale of equal-size units | Having inherent zero point |
| Ordering / ranking is important | A value can be represented as multiple / ratio of another value |

- Can you give examples of interval-scaled and ratio-scaled attributes?

# Discrete vs. Continuous

- From Machine Learning perspective…
- Discrete attribute has finite / countably infinite set of values which may / may not be represented as integers
  - A set of possible values is infinite but the values can be put into a one-to-one correspondence with natural numbers
- Continuous attribute typically represented as floating-point variables

# Descriptive Analytics

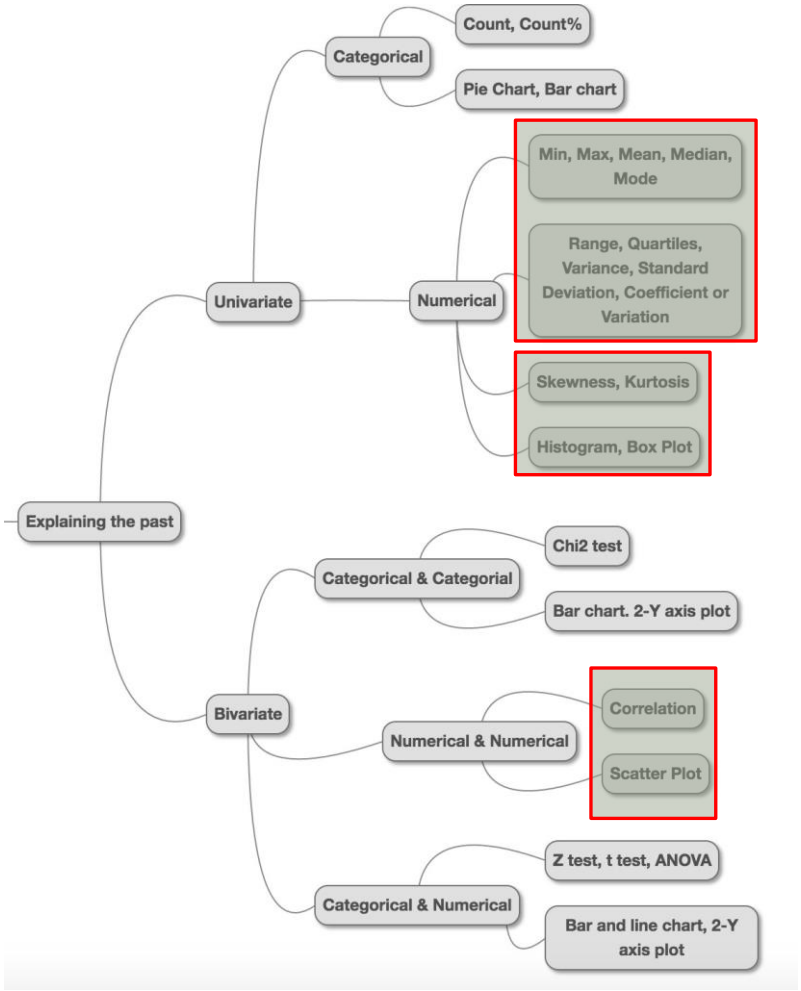


# Describing Data

Data can be described as follows:

- **Tabular**
  - Frequency Distributions
  - Relative Frequency Distributions
- **Graphical**
  - Bar Charts, Histograms
  - Stem and Leaf Plots
  - Scatter plots
  - Time-series plots
- **Summary Statistics**

# Describing Data




Numerical Data can be represented as Summary Statistics, Tabular and Graphical ways

# statistical description

# Statistical Description

| Measures of Central Tendency | Measures of Dispersion of Data | Graphical displays of basic statistical description to visual inspection of data |
|---|---|---|
| Mean | Range | |
| Median | Quantiles | |
| Mode | Quartiles | |
| | Interquartile range | |
| | Percentiles | |
| | Boxplots | |
| | Variance | |
| | Standard deviation | |

# Measures of Central Tendency

| Measures | Description |
|---|---|
| (Arithmetic) mean | The most common and effective measure of the "center" of a set of data. However, it is sensitive to extreme (e.g: outlier) values where even a small number of extreme values can corrupt the mean. |
| Median | The middle value in a set of ordered data values. Note that calculation for median differs if the total number of observations are odd or even. |
| Mode | Value that occurs most frequently in data. |

Given the following dataset, calculate the mean, median, and mode.
65 55 89 56 35 14 56 55 87 45 92

# Mean and sensitivity to outliers

- An outlier is a value that is very different from the other data in your data set. An outlier affects the mean and as a result do not provide a realistic picture of the data
- Consider the following two datasets, calculate the mean, median and mode.

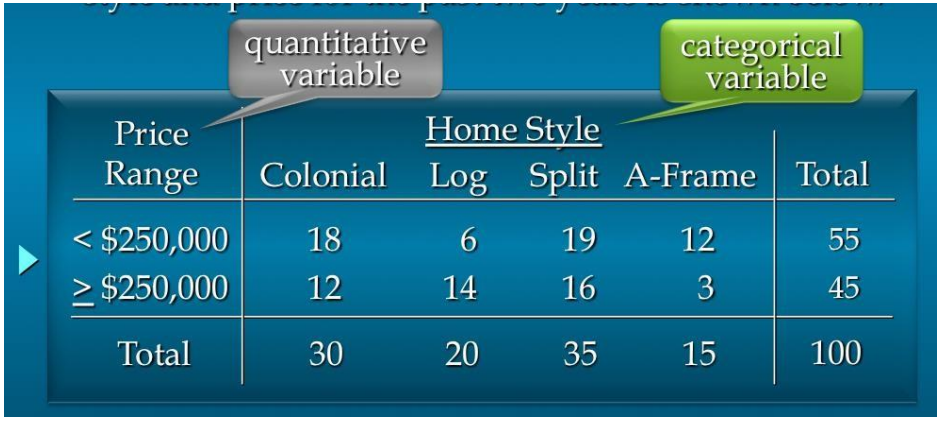
| | | Mean |
|---|---|---|
| **Dataset A** | 65 55 89 56 35 14 56 55 87 45 **920** | **115.27** |
| **Dataset B** | 65 55 89 56 35 14 56 55 87 45 92 | 59 |

# Tabular Representation

- Frequency Distribution and Relative Frequency Distribution Table

| Profit | Frequency | Relative Frequency | Found by |
|---|---|---|---|
| $ 200 up to $ 600 | 8 | .044 | 8/180 |
| 600 up to 1,000 | 11 | .061 | 11/180 |
| 1,000 up to 1,400 | 23 | .128 | 23/180 |
| 1,400 up to 1,800 | 38 | .211 | 38/180 |
| 1,800 up to 2,200 | 45 | .250 | 45/180 |
| 2,200 up to 2,600 | 32 | .178 | 32/180 |
| 2,600 up to 3,000 | 19 | .106 | 19/180 |
| 3,000 up to 3,400 | 4 | .022 | 4/180 |
| Total | 180 | 1.000 | |

- Cross Tabulation of Two Variables



# Statistical Description

| Measures of Central Tendency | Measures of Dispersion of Data | Graphical displays of basic statistical description to visual inspection of data |
|---|---|---|
| Mean | Range | |
| Median | Quantiles | |
| Mode | Quartiles | |
| | Interquartile range | |
| | Percentiles | |
| | Boxplots | |
| | Variance | |
| | Standard deviation | |

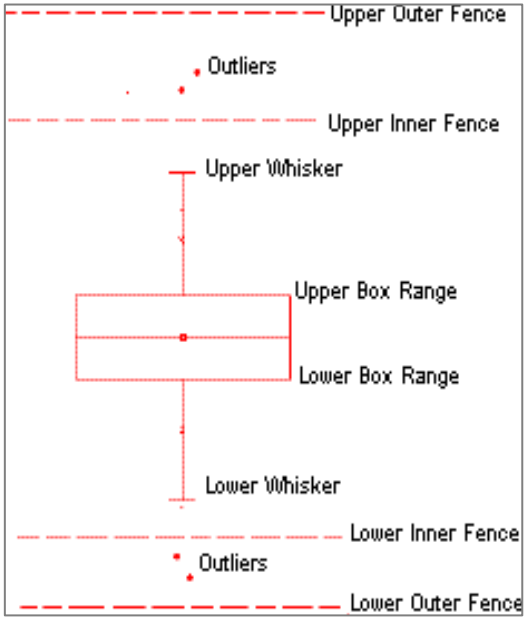
# Measures of Dispersion of Data

| Measures | Description |
|---|---|
| **Range** | Difference between largest and smallest values |
| **Quantiles** | Points taken at regular intervals of a data distribution, dividing it into essentially equal-size consecutive sets. |
| **Quartiles** | One-fourth of a data distribution |
| **Percentiles** | 100 quantiles |
| **Interquartile range** | Distance between first and third quartiles |

# Five-number summary, Boxplots, Outliers

- Five-number summary of a distribution consists of (in order): Minimum, Quartile Q1, Median, Quartile Q3 and Maximum.
- Boxplot incorporates five-number summary as follows:
  - The ends of the box are at the quartiles so that the box length is interquartile range (IQR)
  - The median is marked by a line within the box
  - Two lines (called whiskers) outside the box extend to the minimum and maximum observations

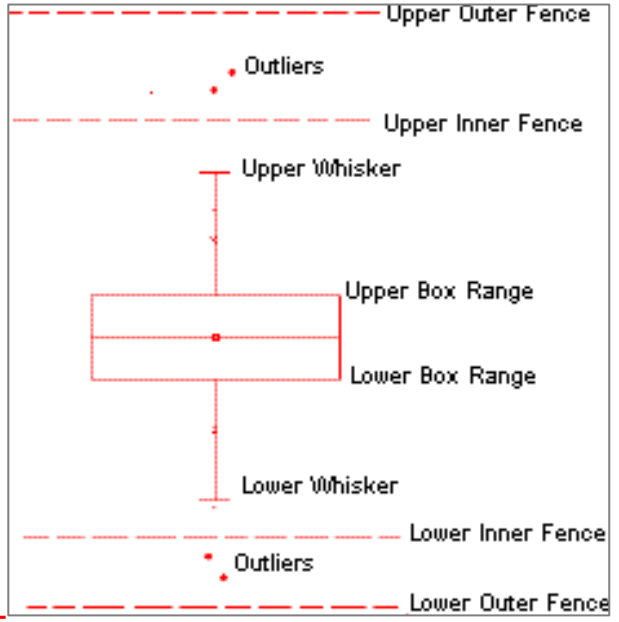
The anatomy of a boxplot



# Five-number summary, Boxplots, Outliers

- The following quantities (called fences) are needed for identifying extreme values in the tails of the distribution.
  - Lower inner fence: Q1 – (1.5*IQR)
  - Upper inner fence: Q3 + (1.5*IQR)
  - Lower outer fence: Q1 – (3*IQR)
  - Upper outer fence: Q3 + (3*IQR)

The anatomy of a boxplot




A point beyond an inner fence on either side is considered a mild outlier. A point beyond an outer fence is considered an extreme outlier.

# Case Study

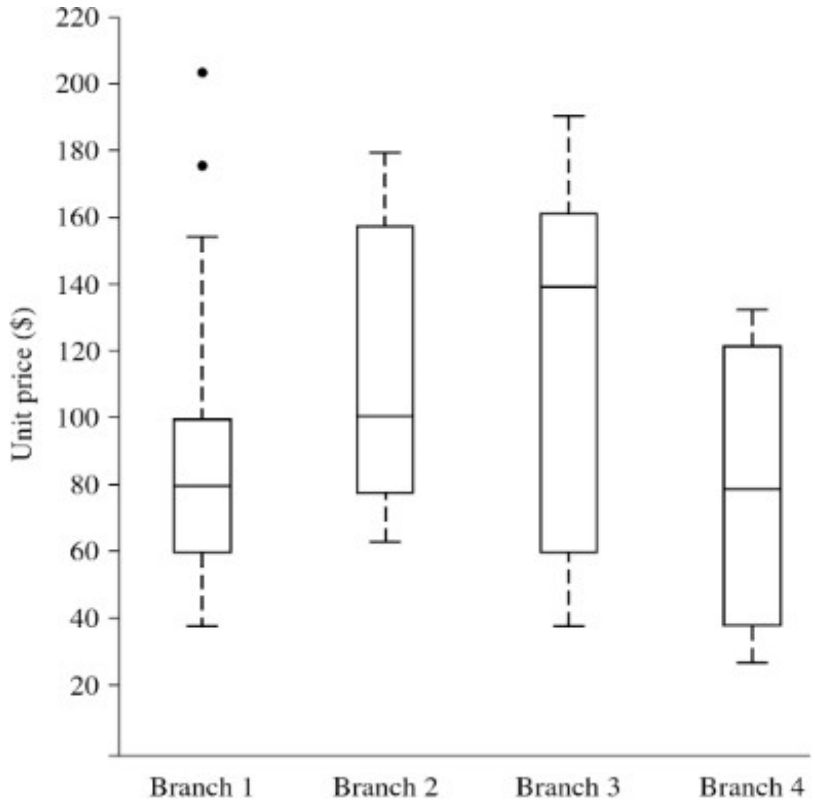



Figure shows the boxplots for unit price data for electronics items sold at 4 branches of Electronics4U Inc. during a fixed time period.

**Q1**:What can you conclude from Branch 1?

**Q2:**Compare the sales in Branch 2 and Branch 3. Contrast their sales.

# Finding outliers when the number of observations is even

30, 171, 184, 201, 212, 250, 265, 270, 272, 289, 305, 306, 322, 322, 336, 346, 351, 370, 390, 404, 409, 411, 436, 437, 439, 441, 444, 448, 451, 453, 470, 480, 482, 487, 494, 495, 499, 503, 514, 521, 522, 527, 548, 550, 559, 560, 570, 572, 574, 578, 585, 592, 592, 607, 616, 618, 621, 629, 637, 638, 640, 656, 668, 707, 709, 719, 737, 739, 752, 758, 766, 792, 792, 794, 802, 818, 830, 832, 843, 858, 860, 869, 918, 925, 953, 991, 1000, 1005, 1068, 1441

The computations are as follows:

Median = (n+1)/2 largest data point = the average of the 45th and 46th ordered points = (559 + 560)/2 = 559.5

Lower quartile = .25(N+1)th ordered point = 22.75th ordered point = 411 + .75(436-411) = 429.75

Upper quartile = .75(N+1)th ordered point = 68.25th ordered point = 739 +.25(752-739) = 742.25

Interquartile range = 742.25 - 429.75 = 312.5

Lower inner fence = 429.75 - 1.5 (312.5) = -39.0

Upper inner fence = 742.25 + 1.5 (312.5) = 1211.0

Lower outer fence = 429.75 - 3.0 (312.5) = -507.75

Upper outer fence = 742.25 + 3.0 (312.5) = 1679.75

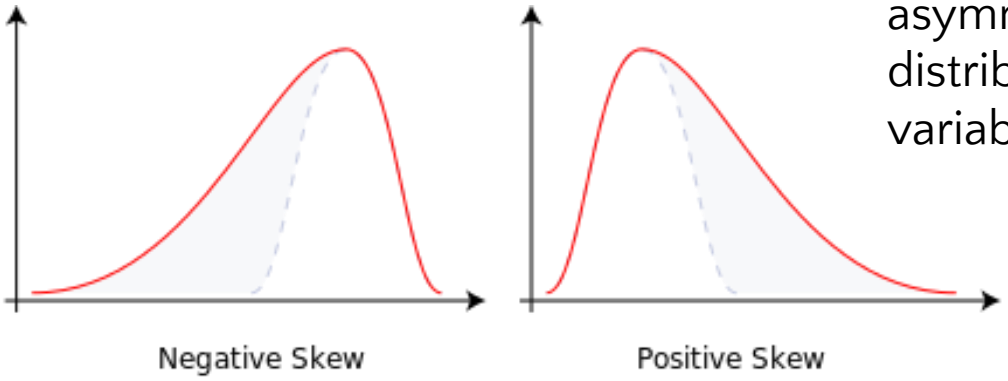
From an examination of the fence points and the data, one point (1441) exceeds the upper inner fence and stands out as a mild outlier; there are no extreme outliers.

# Variance and Standard Deviation

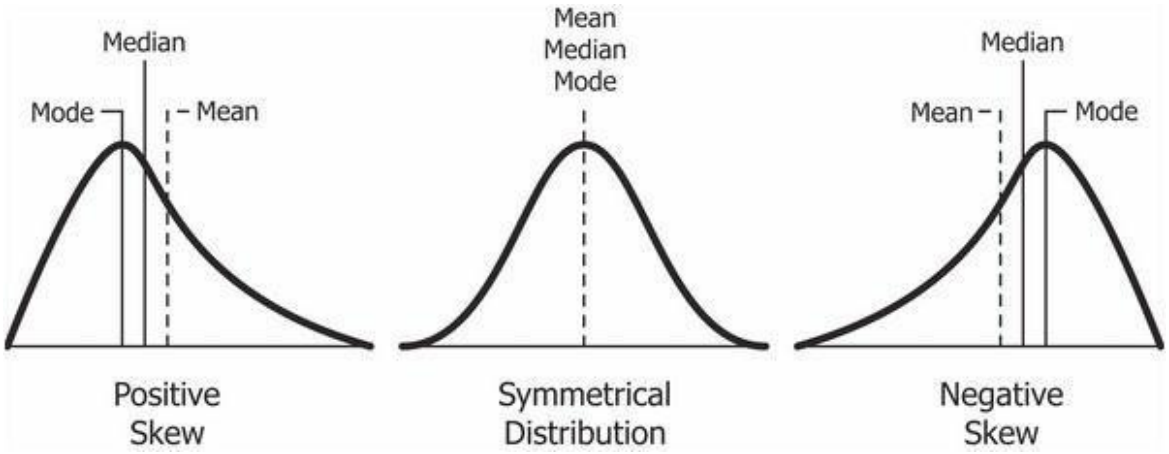
- Low standard deviation ⇒ data observations tend to be very close to the mean ⇒ small spread
- High standard deviation ⇒ data are spread out over a large range of values ⇒ large spread
- The **standard deviation** of the observation is the square root of the **variance**. It measures the dispersion of data.

*Both measures should be familiar to you from earlier math/stat-related subjects in your studies.*

# Skewness




**Skewness** is the measure of the asymmetry of the probability distribution of a real-valued random variable about its mean.




Source: https://codeburst.io/2-important-statistics-terms-you-need-to-know-in-data-science-skewness-and-kurtosis-388fef94eeaa

# Skewness

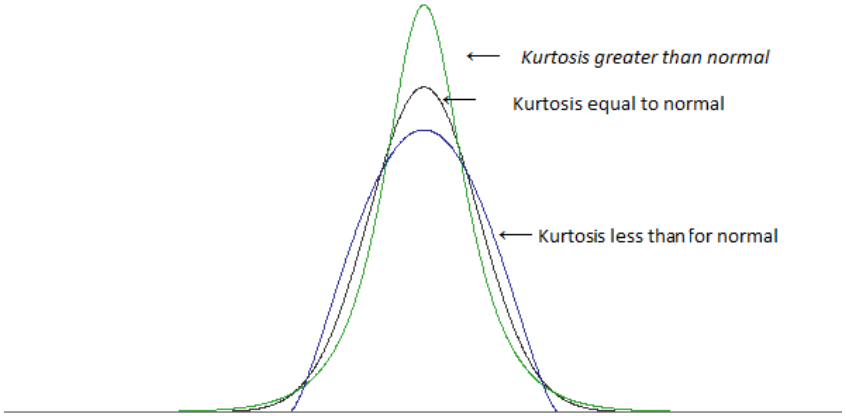
$$Sk = \frac{3(Median - Mean)}{SD}$$

**Rule of Thumb:**

- If the skewness is between -0.5 and 0.5, the data are fairly symmetrical.
- If the skewness is between -1 and -0.5(negatively skewed) or between 0.5 and 1(positively skewed), the data are moderately skewed.
- If the skewness is less than -1(negatively skewed) or greater than 1(positively skewed), the data are highly skewed.

Source: https://codeburst.io/2-important-statistics-terms-you-need-to-know-in-data-science-skewness-and-kurtosis-388fef94eeaa

# Kurtosis




**Kurtosis** is a measure of the "tailedness" of the probability distribution of a real-valued random variable.

**High kurtosis:** “Heavy” tails, or potential outliers
**Low kurtosis:** “Light” tails, or less likely having outliers

A standard normal distribution has a kurtosis of three (3)
Higher and sharper peaks have kurtosis > 3
Lower and broader peaks have kurtosis < 3

Source: https://stats.stackexchange.com/questions/84158/how-is-the-kurtosis-of-a-distribution-related-to-the-geometry-of-the-density-fun

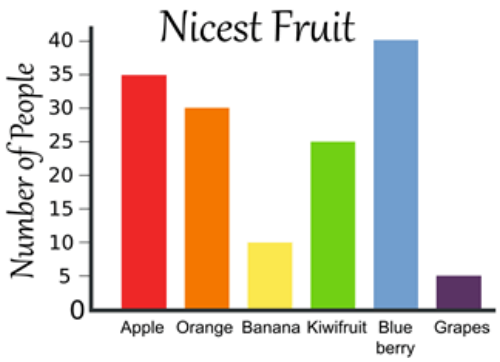
# Statistical Description

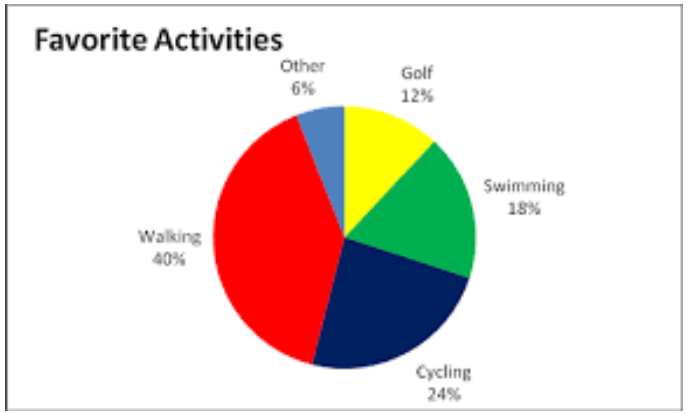
| Measures of Central Tendency | Measures of Dispersion of Data | Graphical displays of basic statistical description to visual inspection of data |
|---|---|---|
| Mean | Range | |
| Median | Quantiles | |
| Mode | Quartiles | |
| | Interquartile range | |
| | Percentiles | |
| | Boxplots | |
| | Variance | |
| | Standard deviation | |

# Graphical Representation (Showing Distribution)

Bar graph bars do not touch because the data is categorical. Histogram bars do touch because the categories are intervals of continuous numbers.

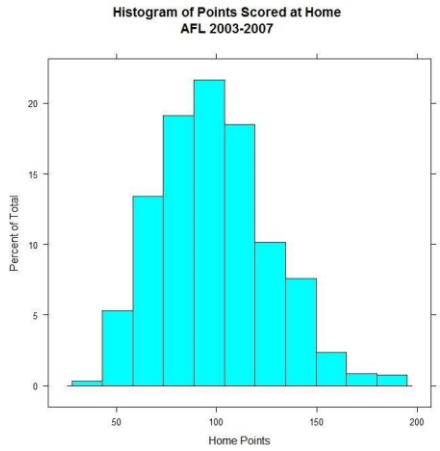
Bar graph

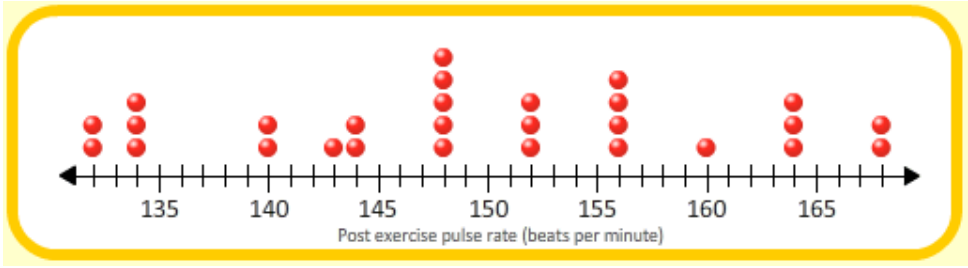



Pie Chart

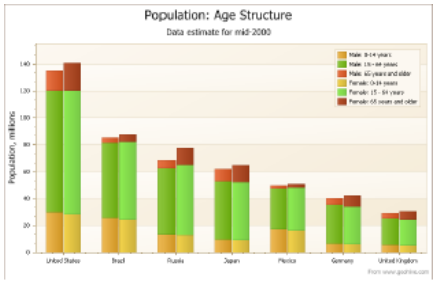
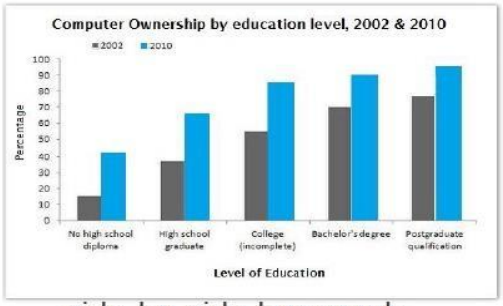
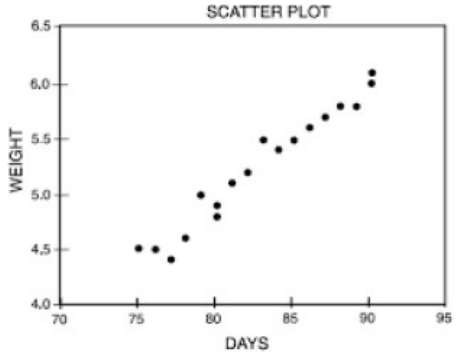
Histogram



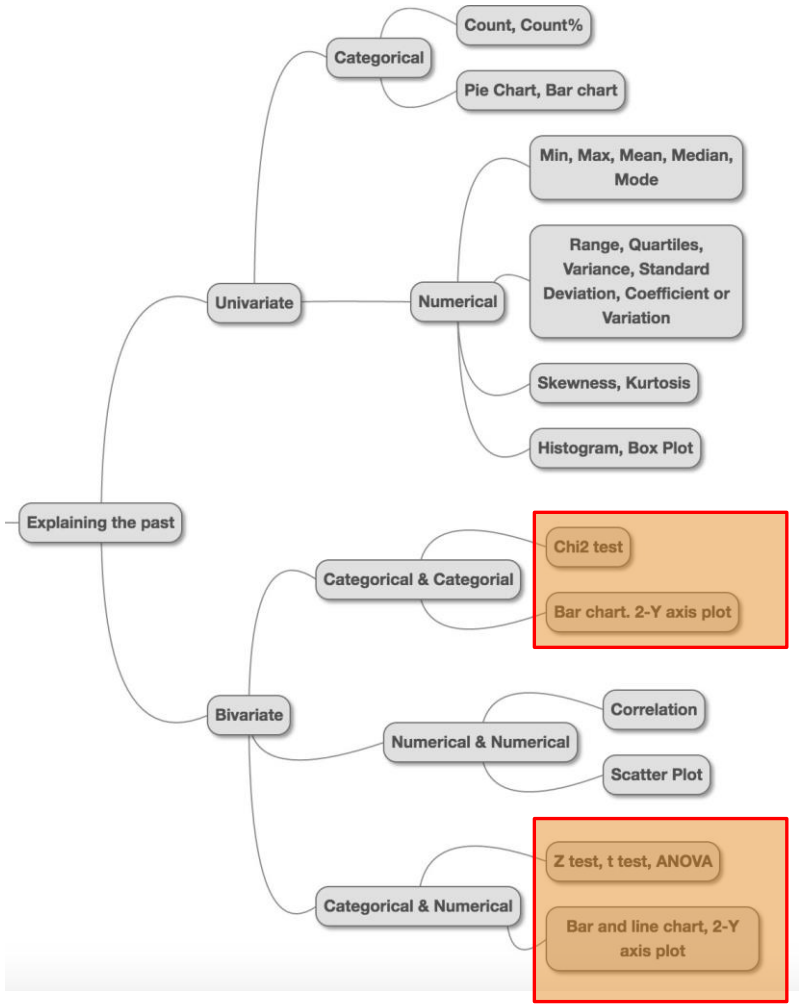




Dot plot

# Graphical Representation

| Showing Comparison | Showing Relation |
| --- | --- |
|  stacked bar chart<br>  side-by-side bar graph |  <br>  |

# Describing Data




Categorical Data can be represented using a variety of statistical tests and displays

# Describing Bivariate Data

| Variable 1 | Variable 2 | Display |
|---|---|---|
| Categorical | Categorical | Crosstabs<br>Stacked Box Plot |
| Categorical | Continuous | Boxplot |
| Continuous | Continuous | Scatterplot<br>Stacked Box Plot |

# Describing Bivariate Data

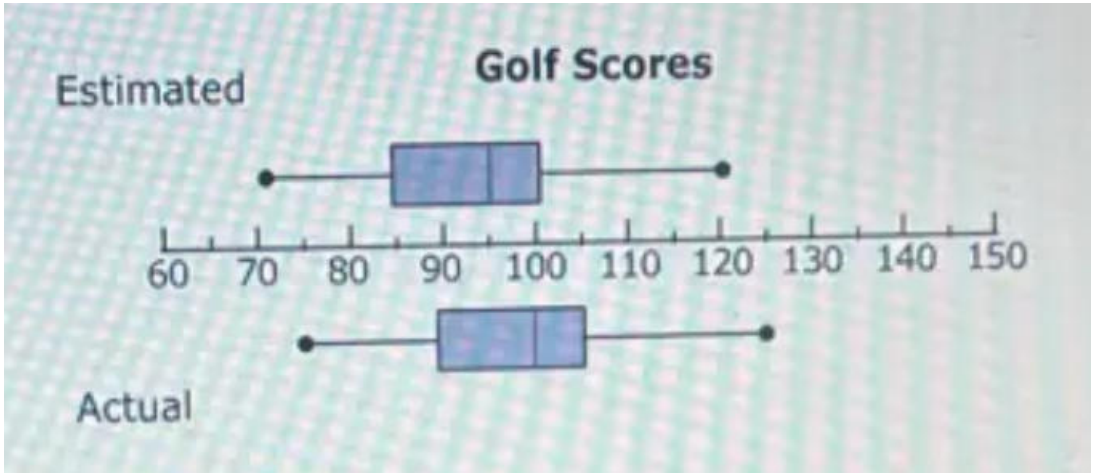
**Class rank * Do you live on campus? Crosstabulation**

Count

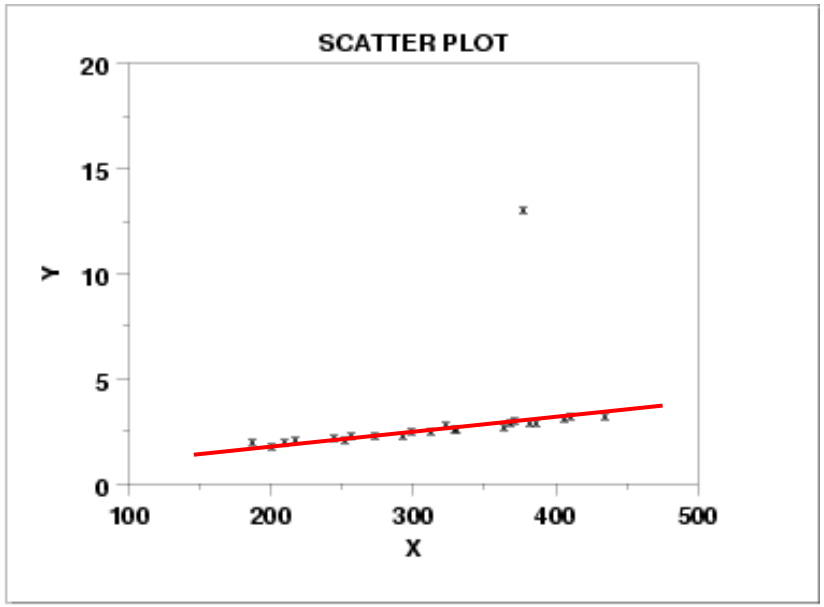
| | | Do you live on campus? | | |
|---|---|---|---|---|
| | | Off-campus | On-campus | Total |
| Class rank | Freshman | 37 | 100 | 137 |
| | Sophomore | 42 | 48 | 90 |
| | Junior | 90 | 8 | 98 |
| | Senior | 62 | 1 | 63 |
| Total | | 231 | 157 | 388 |

Crosstabs

Stacked Box Plot



# Using Scatterplot for Outlier Detection




The scatter plot here reveals a basic linear relationship between X and Y for most of the data, and a single outlier at X = 375

Note: Outliers should be investigated carefully. Often they contain valuable information about the process under investigation or the data gathering and recording process. Before considering the possible elimination of these points from the data, one should try to understand why they appeared and whether it is likely similar values will continue to appear.

relationship between quantities

# Exploratory Data Analysis

- EDA:
  1. The data types of columns and the granularity of rows in the dataset.
  2. The distributions of quantitative data and measures of center and spread.
  3. Relationships between quantities in the dataset.

  - How can we represent the relationship between variables (features / columns / attributes) ?
  - How many of these variables can we represent their relationship at one time?

# Correlation Analysis

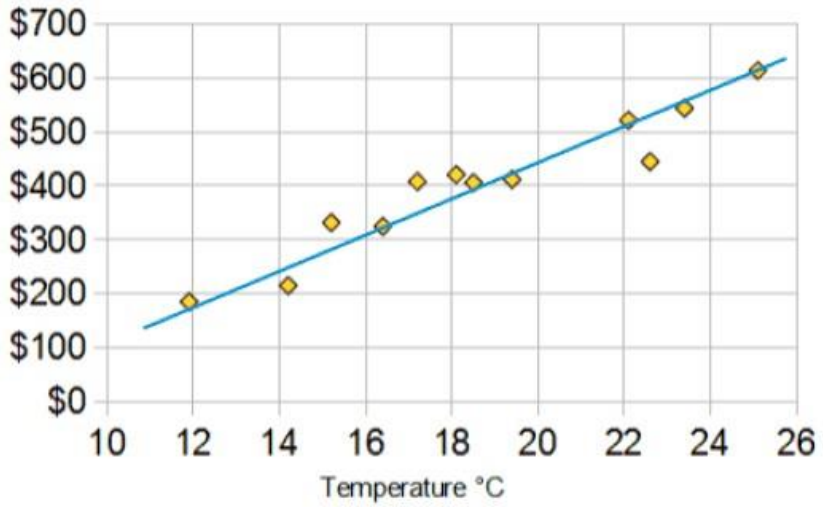
- Correlation analysis is to study the strength of a relationship between two, numerically measured, continuous variables (e.g. height and weight)

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

- Correlation can be positive or negative:
  - **Positive correlation**: one variable increases with the other
  - **Negative correlation**: one variable decreases when the other increases
  - **Correlation close to 0**: both variables have little influence over the other

# Correlation Analysis – Example

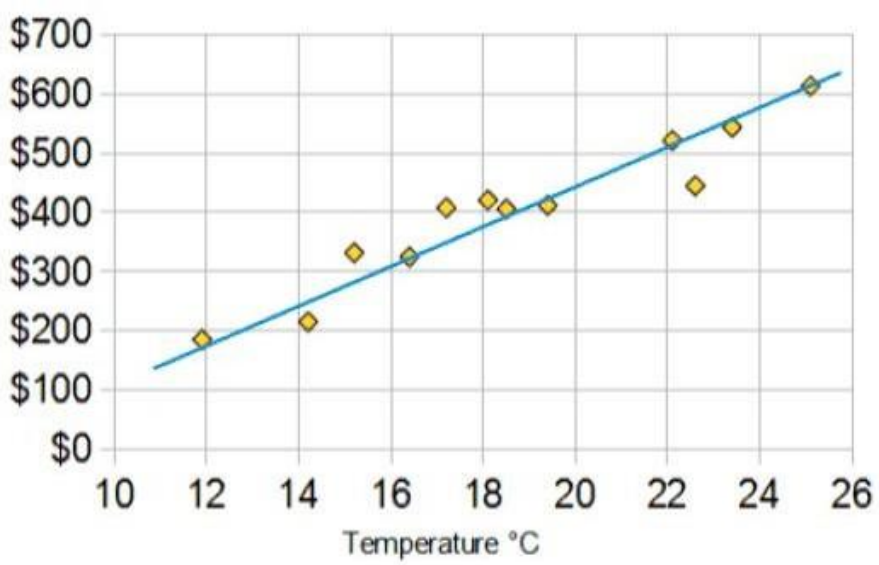
| *Ice Cream Sales vs Temperature* | |
|---|---|
| **Temperature °C** | **Ice Cream Sales** |
| 14.2° | $215 |
| 16.4° | $325 |
| 11.9° | $185 |
| 15.2° | $332 |
| 18.5° | $406 |
| 22.1° | $522 |
| 19.4° | $412 |
| 25.1° | $614 |
| 23.4° | $544 |
| 18.1° | $421 |
| 22.6° | $445 |
| 17.2° | $408 |

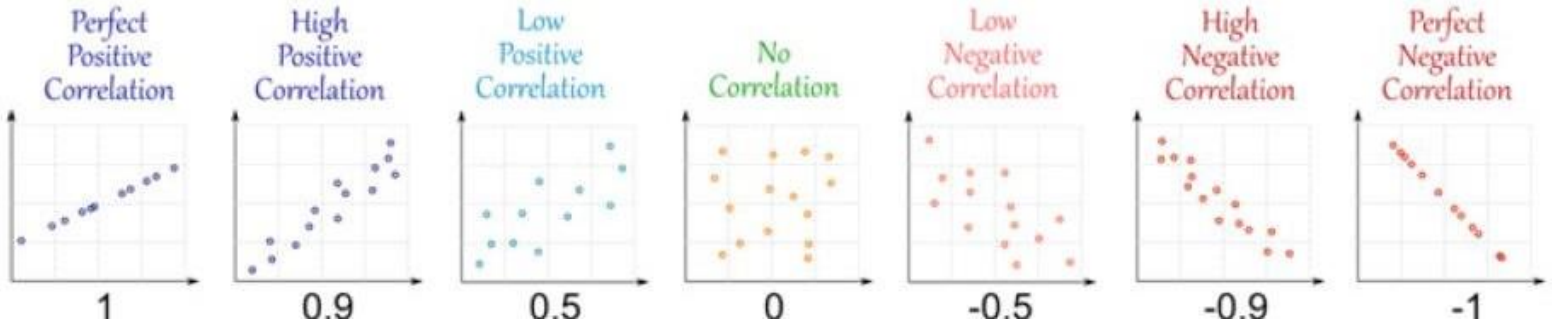



- Shows that warmer weather leads to more sales, but the relationship is not perfect.

Source: https://www.mathsisfun.com/data/correlation

# Correlation Analysis – Example




Source: https://www.mathsisfun.com/data/correlation

# Simpson's Paradox

- **Simpson's Paradox** → Correlations can be misleading when *confounding* variables are ignored – a common surprise when analysing data
- The key issue is that correlation is measuring the relationship between your two variables *all else being equal*. If your data classes are assigned at random, as they might be in a well-designed experiment, "all else being equal" might not be a terrible assumption. But when there is a deeper pattern to class assignments, "all else being equal" can be an awful assumption.
- The only real way to avoid this is by *knowing your data* and by doing what you can to make sure you've checked for possible confounding factors. This is not always possible.

# Simpson's Paradox

| coast | # of members | avg. # of friends |
|---|---|---|
| West Coast | 101 | 8.2 |
| East Coast | 103 | 6.5 |

- Imagine that you can identify all of your members as either East Coast data scientists or West Coast data scientists. You decide to examine which coast's data scientists are friendlier:
- It certainly looks like the West Coast data scientists are friendlier than the East Coast data scientists. Your co-workers advance all sorts of theories as to why this might be: maybe it's urbanized setting, or the coffee, or the possible network, or the infrastructure?

# Simpson's Paradox

| coast | degree | # of members | avg. # of friends |
|---|---|---|---|
| West Coast | PhD | 35 | 3.1 |
| East Coast | PhD | 70 | 3.2 |
| West Coast | no PhD | 66 | 10.9 |
| East Coast | no PhD | 33 | 13.4 |

**What is the confounding factor here?**

- When playing with the data you discover something very strange. If you only look at people without PhDs, the East Coast data scientists have more friends on average. And if you only look at people without PhDs, the East Coast data scientists also have more friends on average!
- Once you account for the users' degrees, the correlation goes in the opposite direction! Bucketing the data as East Coast/West Coast disguised the fact that the East Coast data scientists skew much more heavily toward PhD types.

# Correlation vs. Causation

- **Correlation** is a statistical measure (expressed as a number) that describes the size and direction of a relationship between two or more variables.
  - Correlation, however, does not automatically mean that the change in one variable is the **cause** of the change in the values of the other variable.
- **Causation** indicates that one event is the result of the occurrence of the other event; i.e. there is a causal relationship between the two events. This is also referred to as "cause-and-effect".

Source: www.abs.gov.au/websitedbs/a3121120.nsf/home/statistical+language+-+correlation+and+causation

# Reading Material

- Howard Seltman, CMU, "Exploratory Data Analysis" [PDF], 2018
- Yassien Shaalan (Growing Data), "A guided introduction to Exploratory Data Analysis (EDA) using Python", 2019
- [YouTube] Prof. Patrick Meyer, "Exploratory Data Analysis", 2015
- Karthikeya Boyini, Tutorialspoint, "Exploratory Data Analysis in Python", 2019