Lingnan University 嶺南大學 香港 Hong Kong

School of Data Science 數據科學學院

CDS 533
Statistics for Data Science

Instructor: Lisha Yu

Division of Artificial Intelligence

School of Data Science

Lingnan University

Fall 2024

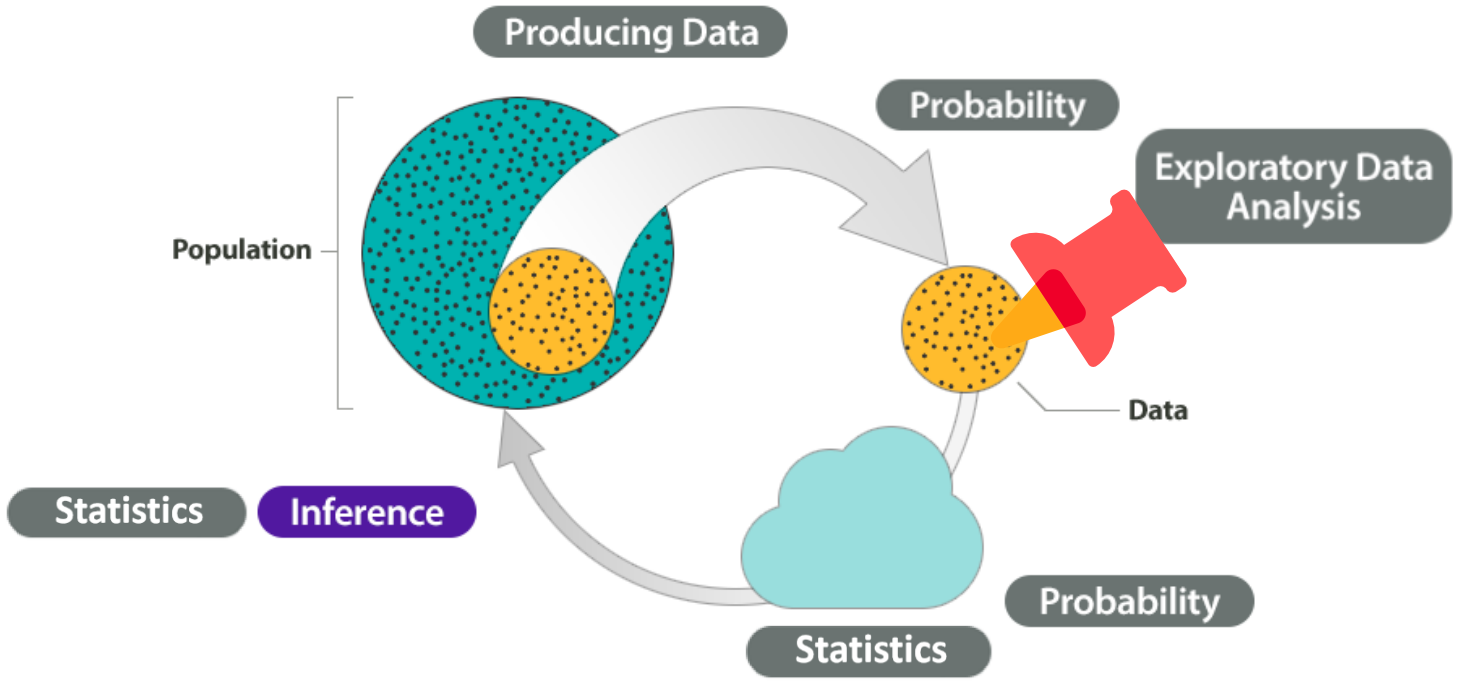
Survey Example

Doctor of Artificial Intelligence

https://forms.office.com/Pages/ResponsePage.aspx?id=nReQ_geCskyINAzifuAWLbxijFLlNFFOrsXVJM7sZEBUNFZEM0JHVVRXRUxOQ0VRQ0ZHNldPUUQ4MS4u

Big Picture of Statistics

Exploratory Data Analysis



Motivation Question

Recall from the last lecture that data have been collected by sampling method or design of experiment,

Data is ready. What's NEXT?

Key Characteristics of a Data Set

Every data set is accompanied by important background information. In a statistical study, **always ask the following questions**:

- **Who**? What **cases** do the data describe? **How many** cases does a data set have?

Descriptive Statistics

Allows to describe or summarize the data. This is typically seen at the beginning of a results section. It gives an idea of the sample size, the characteristics of the study (e.g. baseline characteristics, ect.)

Example: *A total of 235 students participated in this study, 163 women (69.4%) versus 72 men (30.6%). On average the female students (81.3 ± 19.4) had a slightly higher score on exam 2 in comparison to the male students (80.7 ± 18.1).*

Who? How many? Enough?

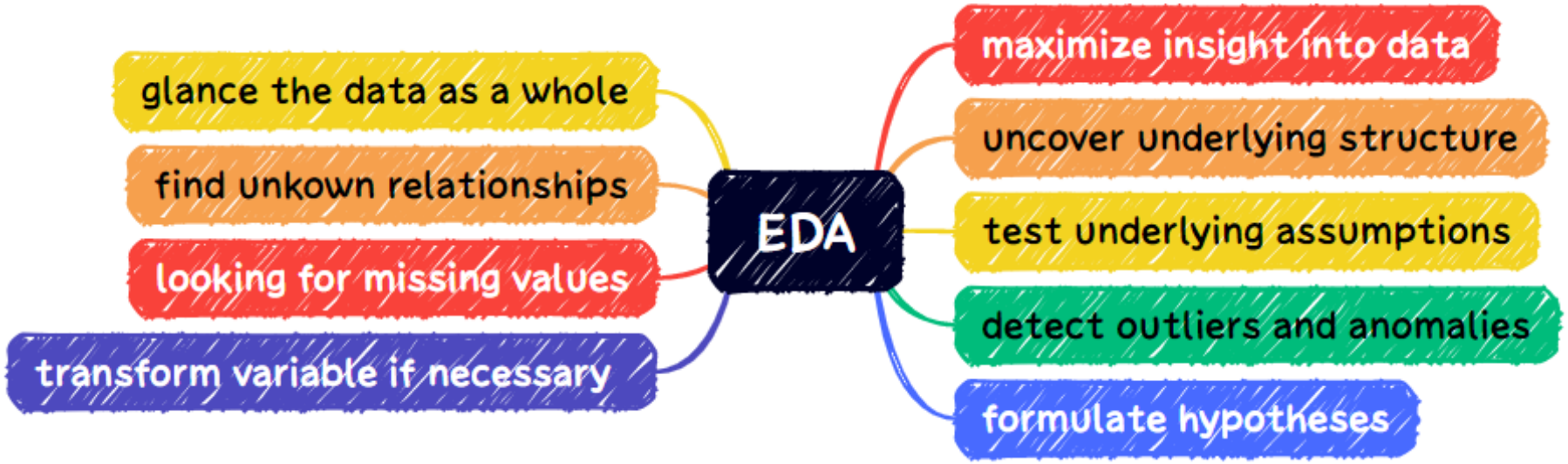
Key Characteristics of a Data Set

Every data set is accompanied by important background information. In a statistical study, **always ask the following questions**:

- **Who**? What **cases** do the data describe? **How many** cases does a data set have?
- **What**? How many **variables** does the data set have? What are the **exact definitions** of these variables? What are the units of measurement for each quantitative variable?
- **Why**? **What purpose** do the data have? Do the data contain the information needed to answer the questions of interest?

Exploratory Data Analysis (EDA)

Exploratory Data Analysis (EDA) is an approach/philosophy for data analysis that employs a variety of techniques (**most graphical**) to

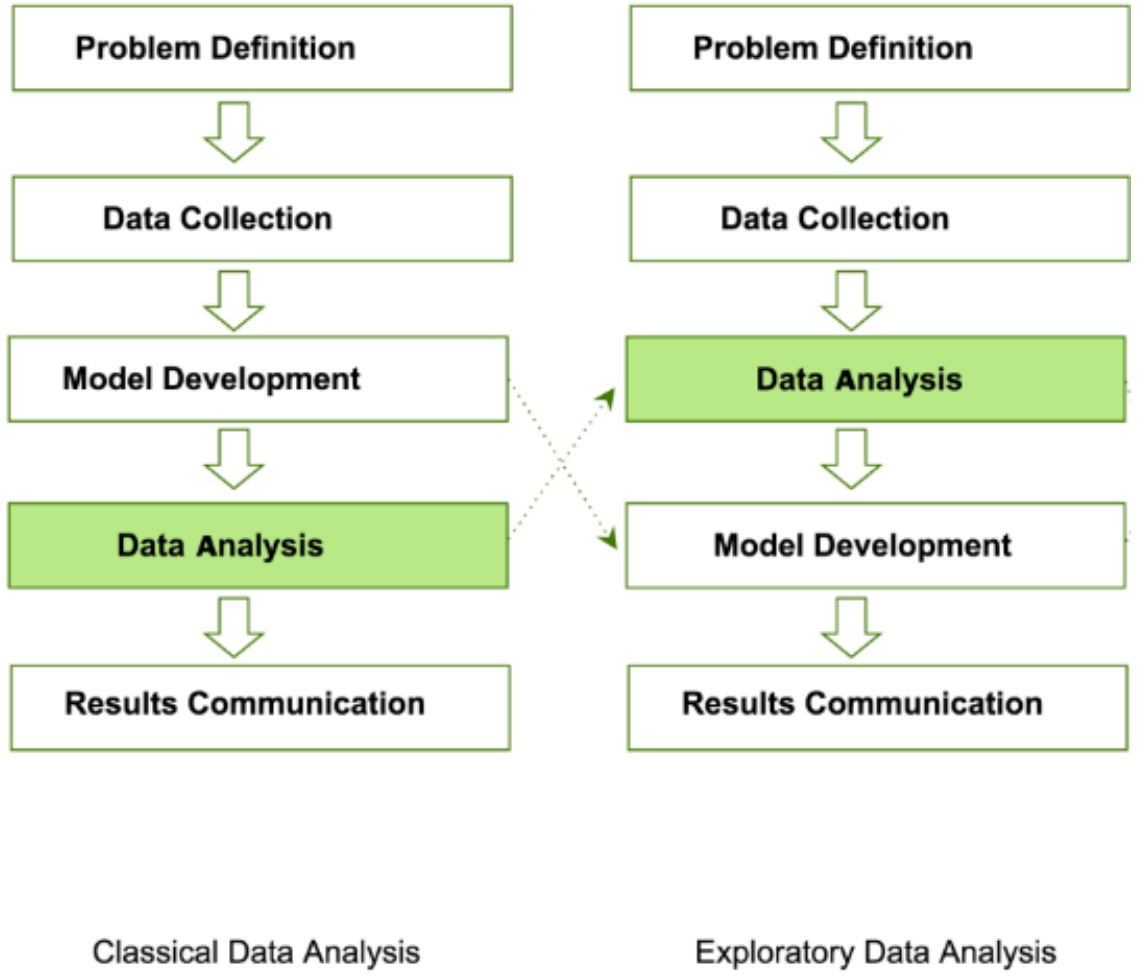


Exploratory Data Analysis (EDA)

"The role of the data analyst is to listen to the data in as many ways as possible until a plausible "story" of data is apparent."

--- Behrens, 1997

Paradigms for Analysis Techniques



The main focus of EDA is on the data, its structure, outliers, models, and visualizations. Generally, we do not impose any deterministic or probabilistic models on the data.

DEA Differ from Summary Analysis

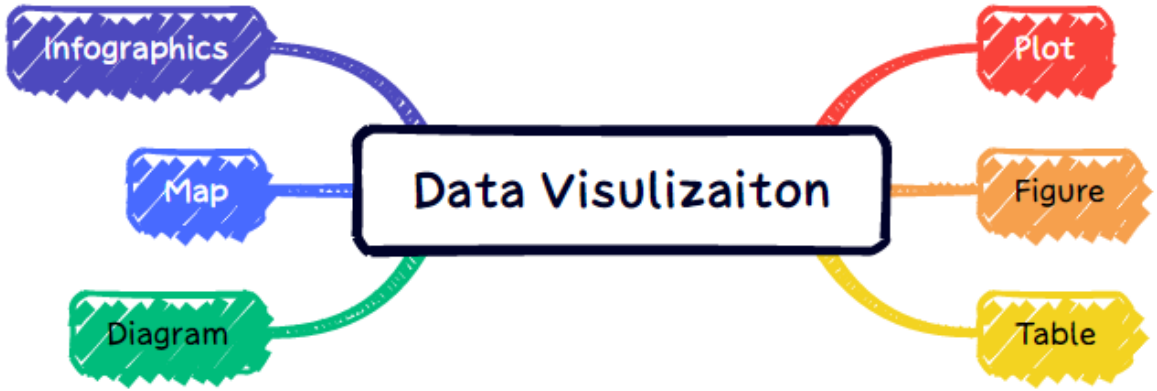
Summary

- Numerical reduction of dataset
- Passive; arrive at a few key statistics
- Focus is in the past
- Replace/add to the data

Exploratory

- Gain insight into scientific process
- Active and futuristic
- Use data as a “window” to peer into question behind
- Enormously large role for EDA approach

Tips: Data Visualization



(not limit to)

- Data visualization can help to get a handle on what's going on in the dataset. **(exploratory plots)**
  - They are made quickly.
  - You'll make a large number of them.
  - The axes and legends are cleaned up.

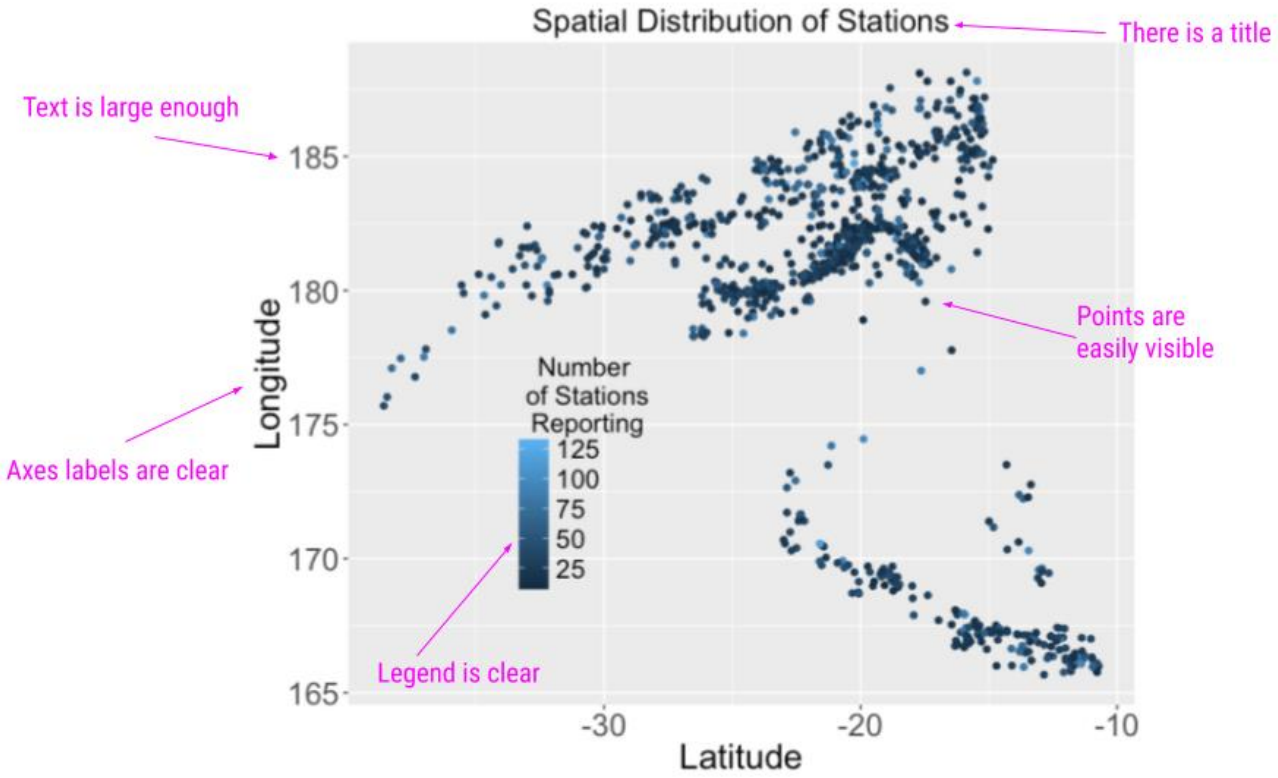
- Similarly, once we have completed our analysis and are ready to present our findings, data visualizations are a highly effective way to communicate our results to others. **(explanatory plots)**
  - They take a while to make.
  - There are only a few of these for each project.
  - You've spent time making sure the colors, labels, and sizes are all perfect for your needs.

Tips: Data Visualization Guide

Good plots have a number of features. While not exhaustive, good plots have:

- The Subtitle
- Clearly-labeled axes
- Text that is large enough to see
- Axes that are not misleading
- The layout and color of outliers
- Data that are displayed appropriately considering the type of data you have

(Perceptual Hierarchy of Visual Cues)



Doing EDA is not merely about plotting graphs. It is about making informative graphs.

EDA Steps

	A	B	C	D	E	F
1	ID	Type	Name	Item	RegPrice	DiscPrice
2	1	Italian	Domo's	Pizza	20	10
3	2	Italian	Mama Rita's	Pizza	20	12
4	3	BBQ	Smokey McSween's	Barbecue	30	17
5	4	BBQ	Smokey Grill	Ribs	20	11
6	5	Mexican	Dos Amigos	Tacos	16	8
7	6	Mexican	Holy Guacamole	Steak fajitas	13	8
8	7	Seafood	Sea Grille	Shrimp platter	20	11

- **Begin** by examining **each variable by itself**.
- **Then** move on to study the **relationships among the variables**.

- **Begin** with a graph or **graphs**. (Graphical summary to inspect the shape of the distribution: symmetry, modality, heaviness of tails)
- **Then** add **numerical summaries** of specific aspects of the data.(Descriptive statistics: classical measures of location and spread)

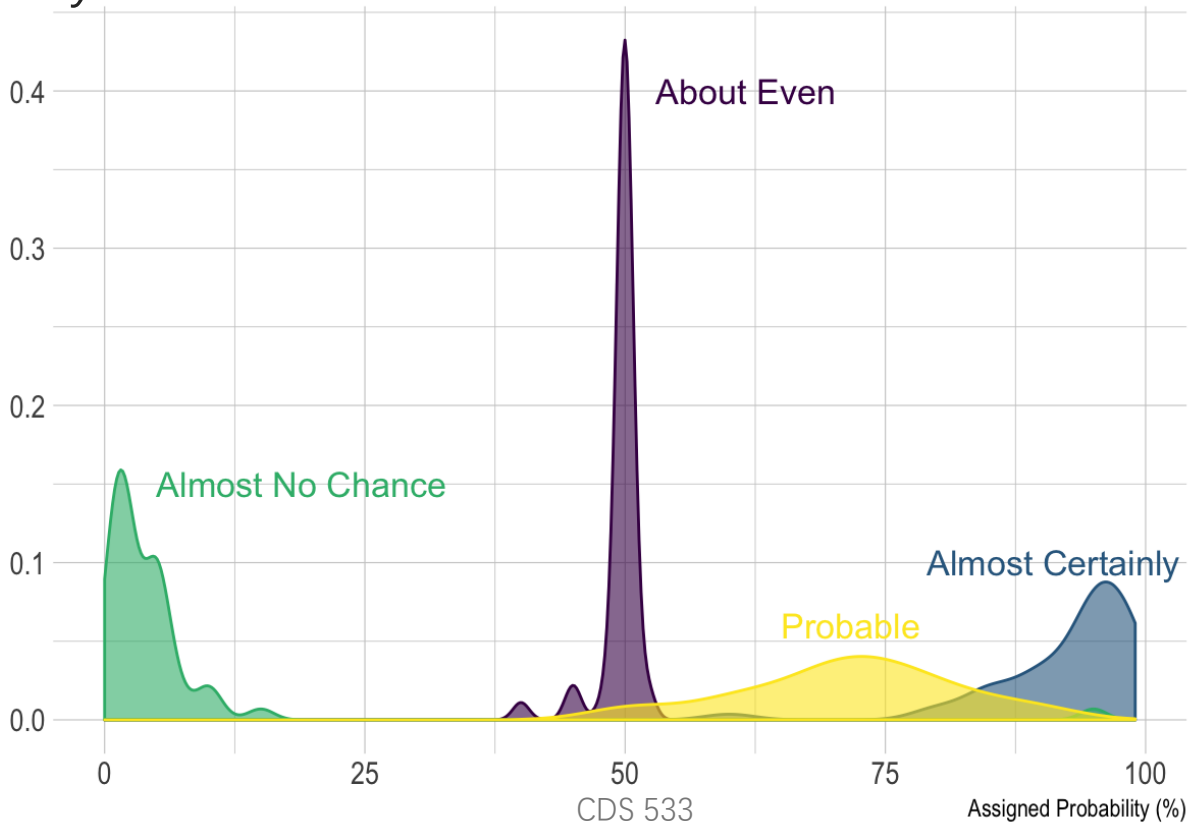
Tips: EDA

EDA is open ended and can be considered as an attitude toward the data.

Data Distribution Assumption

Some common data distribution assumptions when applying statistical analysis

- Normality
- Skewness
- Homogeneity

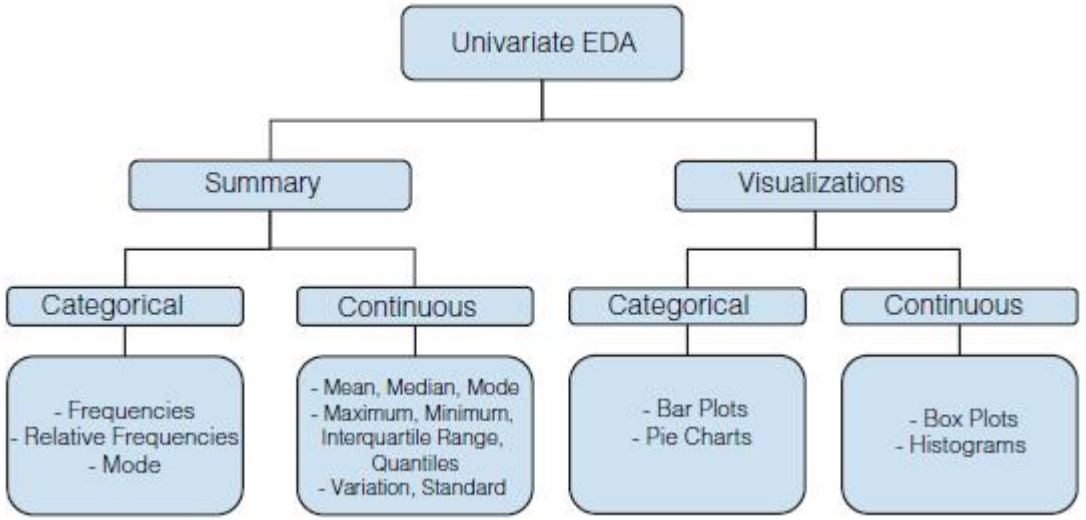


Looking at Data – Univariate Analysis

Univariate Analysis focuses on a single variable at a time, examining its distribution, summary statistics, and visual representation. It is a fundamental step before diving into more complex multivariate analyses.

Univariate Analysis

- Graphic Tools
  - Categorical variables
  - Numerical variables
- Descriptive Statistics
  - Categorical variables
  - Numerical variables
- Normality



Graphic - Univariate - Pie

Categorical Variables

How do you do online research? A study of 552 first-year college students asked about their preferences for online resources. One question asked them to pick their favorite.[3] Here are the results:

Resource	Count (*n*)
Google or Google Scholar	406
Library database or website	75
Wikipedia or online encyclopedia	52
Other	19
Total	552

Resource	Percent(%)
Google or Google Scholar	73.6
Library database or website	13.6
Wikipedia or online encyclopedia	9.4
Other	3.4
Total	100.0

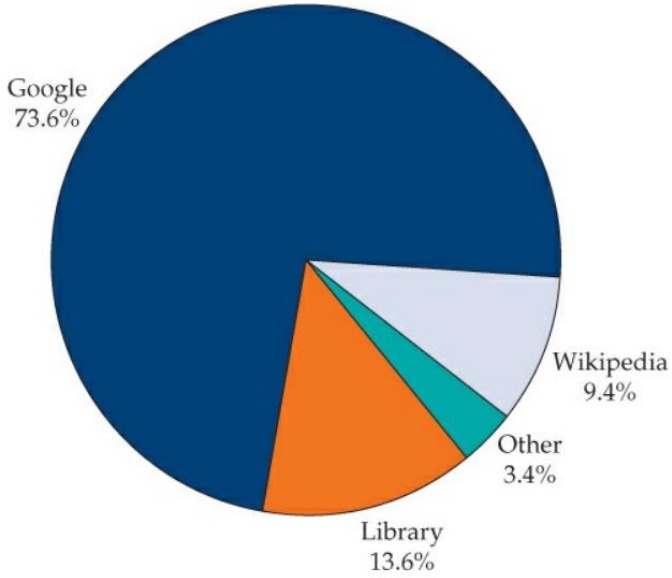
Univariate Analysis
- Graphic Tools
  - Categorical variables
  - Numerical variables
- Descriptive Statistics
  - Categorical variables
  - Numerical variables
- Normality

Graphic - Univariate - Pie

Categorical Variables

Pie charts show the distribution of a categorical variable as a "pie" whose slices are sized by the counts or percentages for the categories.

To make a pie chart, you must include all the categories that make up a whole.



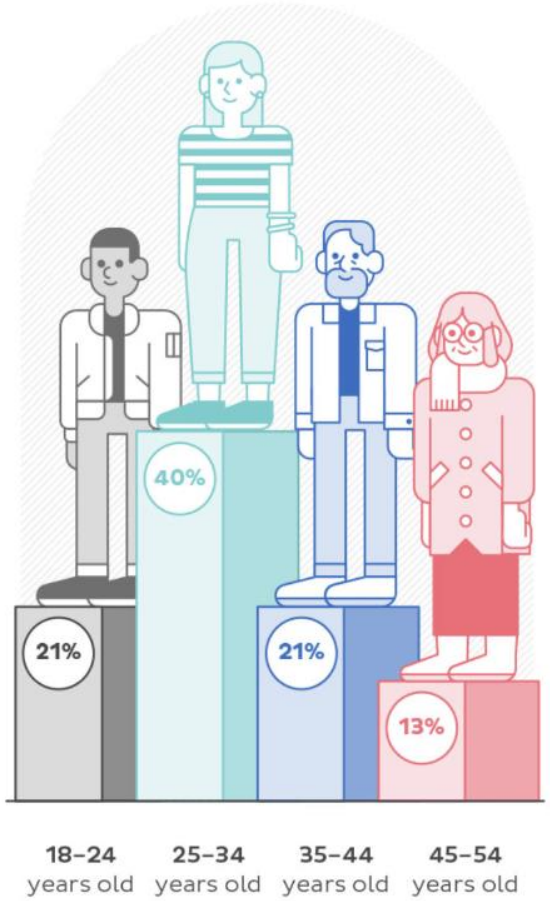
Univariate Analysis
- Graphic Tools
  - **Categorical variables**
  - Numerical variables
- Descriptive Statistics
  - Categorical variables
  - Numerical variables
- Normality

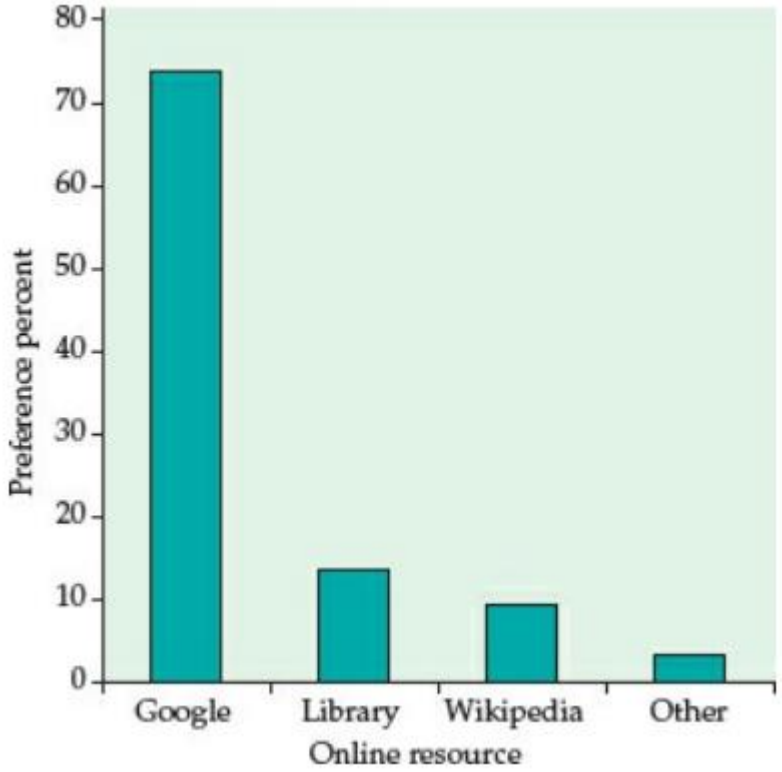
Graphic - Univariate – Bar Graph

Categorical Variables

Bar graphs represent categories as bars whose heights show the category counts or percentages.

You should always consider the best way to order the values of the categorical variable in a bar graph.





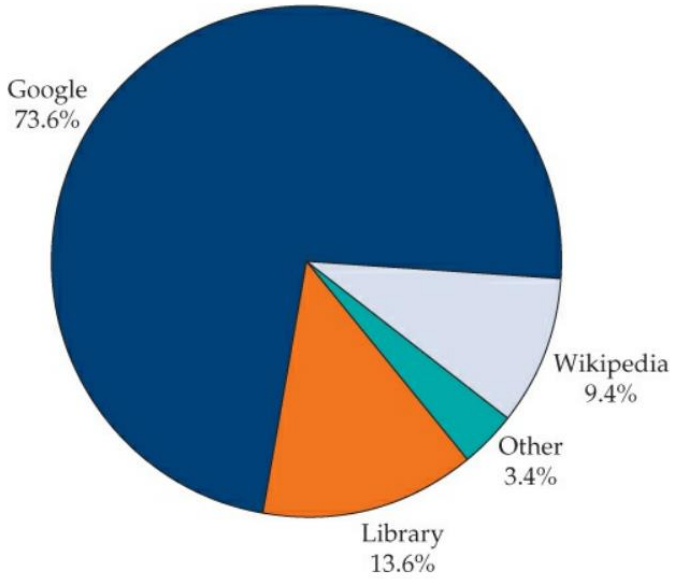
Univariate Analysis
- Graphic Tools
  - **Categorical variables**
  - Numerical variables
- Descriptive Statistics
  - Categorical variables
  - Numerical variables
- Normality

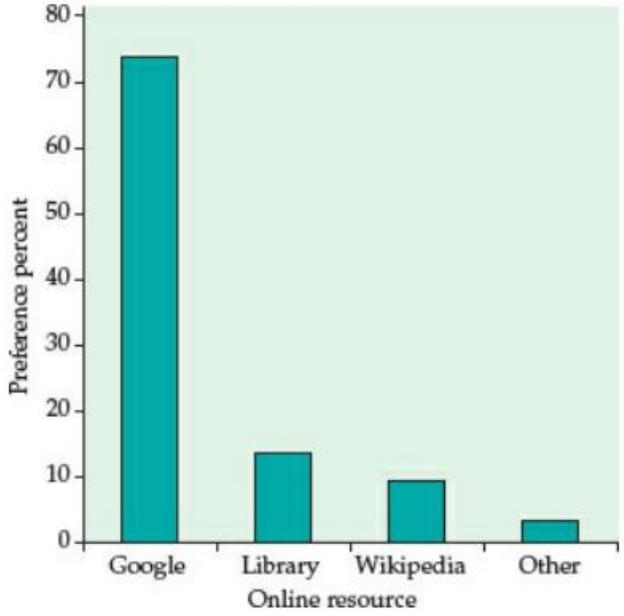
Graphic - Univariate – Bar Graph

Categorical Variables

Question: Compare the bar graph with the pie chart.

Refer to the online resource preference data. Which graphical display does a better job of describing the data? Give reasons for your answer.





Univariate Analysis
- Graphic Tools
  - **Categorical variables**
  - Numerical variables
- Descriptive Statistics
  - Categorical variables
  - Numerical variables
- Normality

Graphic - Univariate - Histogram

Numerical Variables

Distribution of IQ scores. You have probably heard that the distribution of scores on IQ tests is supposed to be roughly "bell-shaped." Let's look at some actual IQ scores. Table 1.1 displays the IQ scores of 60 fifth-grade students chosen at random from one school.

1. Divide the range of the data into classes of equal width. Let's use

$$75 \le \text{IQ score} < 85$$

$$85 \le \text{IQ score} < 95$$

$$\vdots$$

$$145 \le \text{IQ score} < 155$$

TABLE 1.1 IQ Test Scores for 60 Randomly Chosen Fifth-Grade Students

145	139	126	122	125	130	96	110	118	118
101	142	134	124	112	109	134	113	81	113
123	94	100	136	109	131	117	110	127	124
106	124	115	133	116	102	127	117	109	137
117	90	103	114	139	101	122	105	97	89
102	108	110	128	114	112	114	102	82	101

Univariate Analysis
- Graphic Tools
  - Categorical variables
  - **Numerical variables**
- Descriptive Statistics
  - Categorical variables
  - Numerical variables
- Normality

Graphic - Univariate - Histogram

Numerical Variables

frequency frequency table

2. Count the number of individuals in each class. These counts are called **frequencies**, and a table of frequencies for all classes is a **frequency table**.

Class	Count
$75 \leq$ IQ score < 85	2
$85 \leq$ IQ score < 95	3
$95 \leq$ IQ score < 105	10
$105 \leq$ IQ score < 115	16
$115 \leq$ IQ score < 125	13
$125 \leq$ IQ score < 135	10
$135 \leq$ IQ score < 145	5
$145 \leq$ IQ score < 155	1

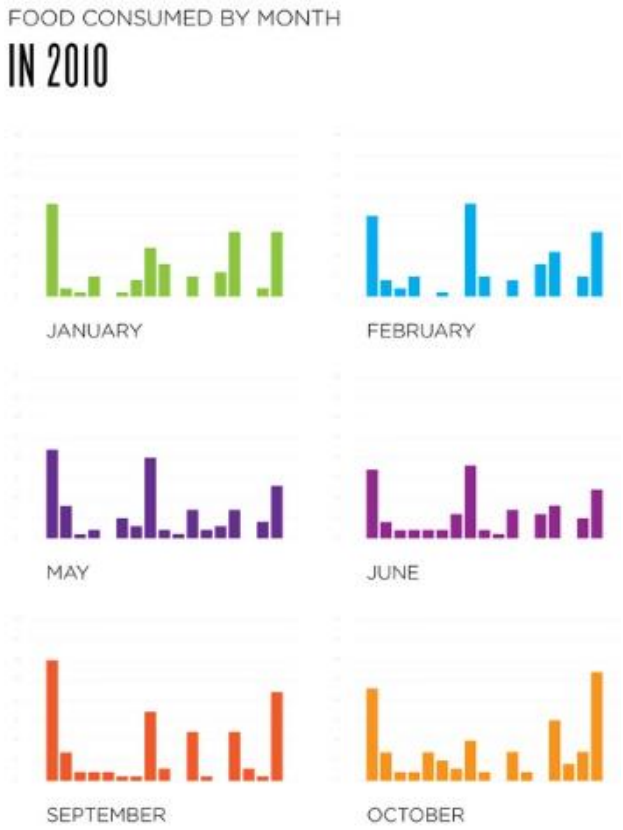
Univariate Analysis
- Graphic Tools
  - Categorical variables
  - Numerical variables
- Descriptive Statistics
  - Categorical variables
  - Numerical variables
- Normality

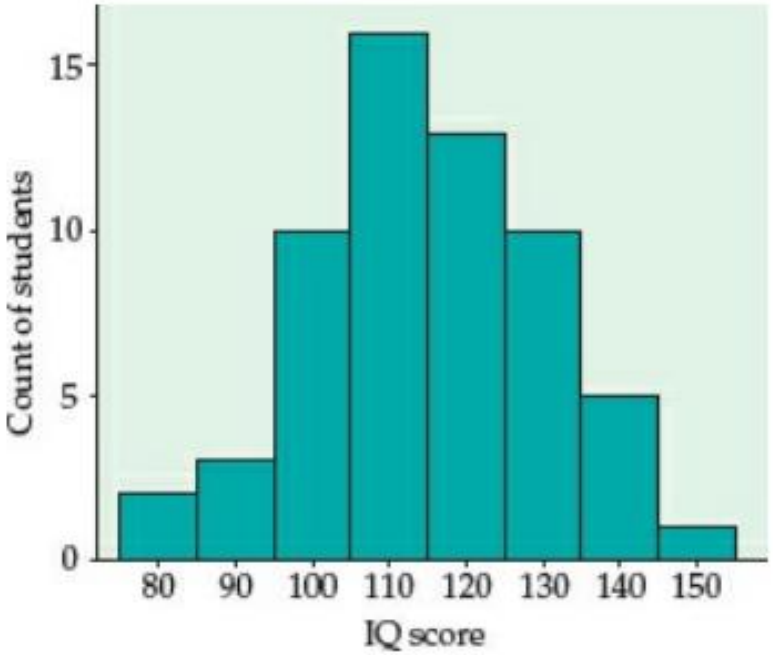
Graphic - Univariate - Histogram

Numerical Variables

Histograms show the distribution of a quantitative variable by using bars. The height of a bar represents the number of individuals whose values fall within the corresponding class.

You should be aware that the appearance of a histogram can change when you change the classes





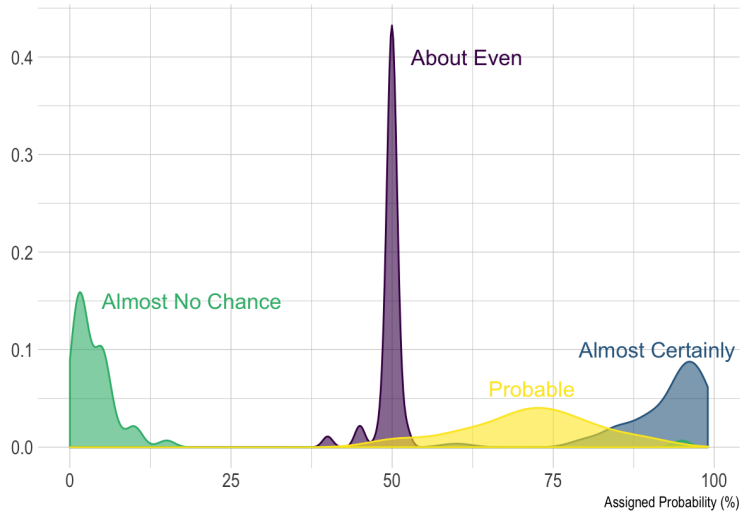
Univariate Analysis
- Graphic Tools
  - Categorical variables
  - **Numerical variables**
- Descriptive Statistics
  - Categorical variables
  - Numerical variables
- Normality

Graphic - Univariate

Examining a distribution

Making a statistical graph is not an end in itself. The purpose of the graph is to help us understand the data. After you make a graph, always ask, "What do I see?"

- In any data graph, look for the **overall pattern** and for striking **deviations** from that pattern.
- You can describe the overall pattern by its **shape, center, and spread**.
- An important kind of deviation is an **outlier**, an individual that falls outside the overall pattern.

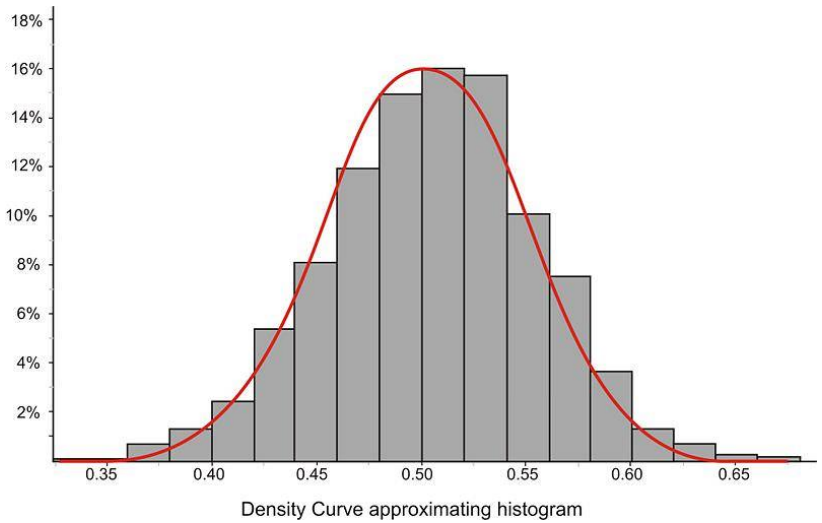


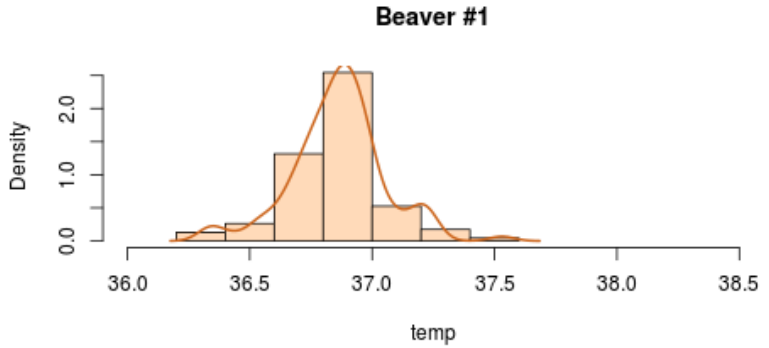
Univariate Analysis
- Graphic Tools
  - Categorical variables
  - **Numerical variables**
- Descriptive Statistics
  - Categorical variables
  - Numerical variables
- Normality

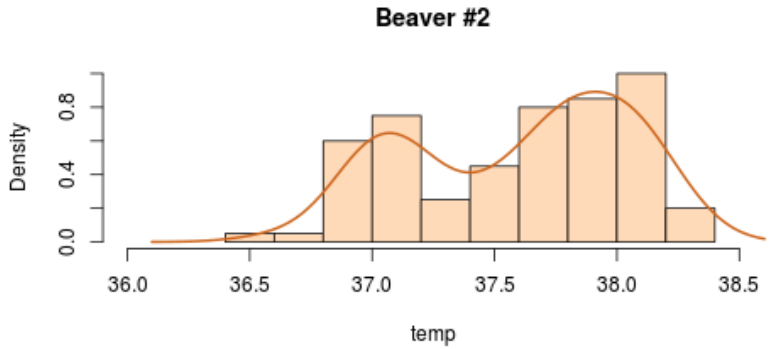
Graphic - Univariate – Density Plot

Numerical Variables

Density plots are smoothed versions of histograms, visualizing the distribution of a continuous variable. These plots effectively visualize the distribution shape and are, unlike histograms, are not sensitive to the number of bins chosen for visualization.







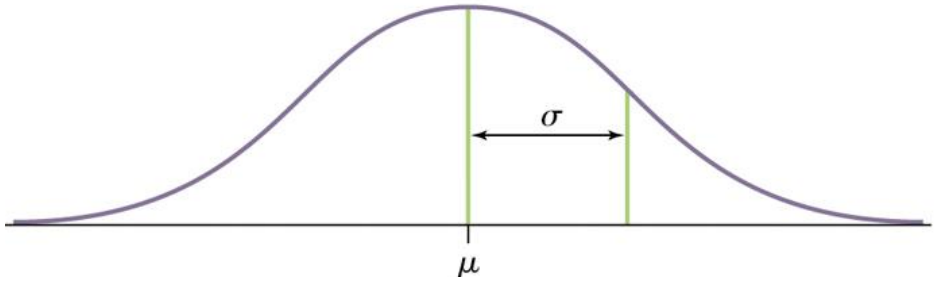
Univariate Analysis
- Graphic Tools
  - Categorical variables
  - **Numerical variables**
- Descriptive Statistics
  - Categorical variables
  - Numerical variables
- Normality

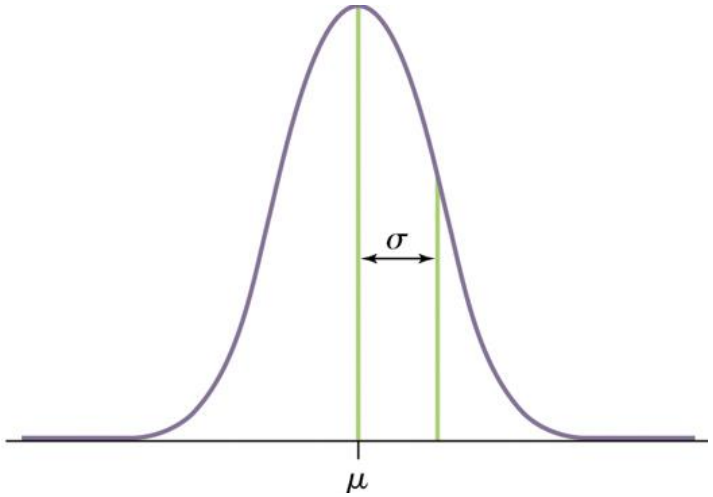
Graphic - Univariate – Normality

Numerical Variables

Normal distribution

- All Normal curves are symmetric, single-peaked, and bell-shaped.
- A specific Normal curve is described by $N(\mu, \sigma)$.





Univariate Analysis
- Graphic Tools
  - Categorical variables
  - **Numerical variables**
- Descriptive Statistics
  - Categorical variables
  - Numerical variables
- Normality

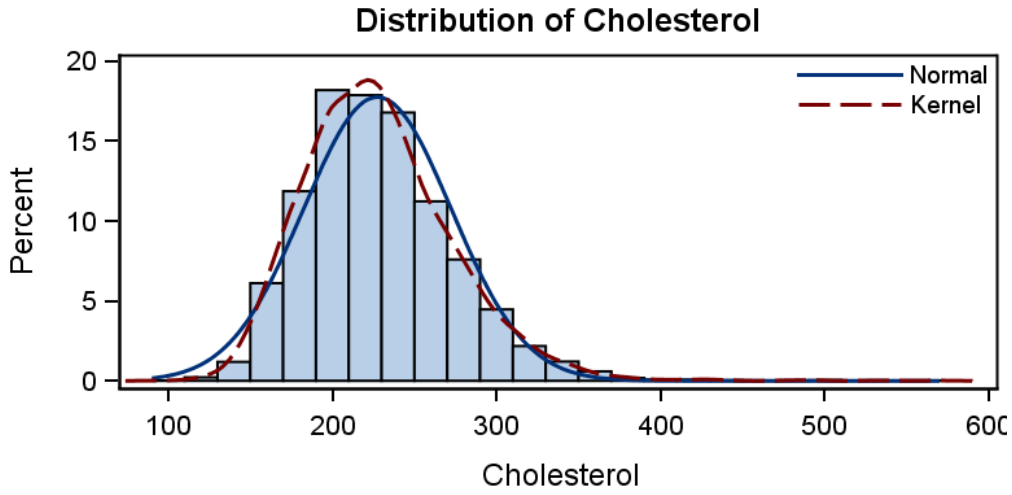
$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$

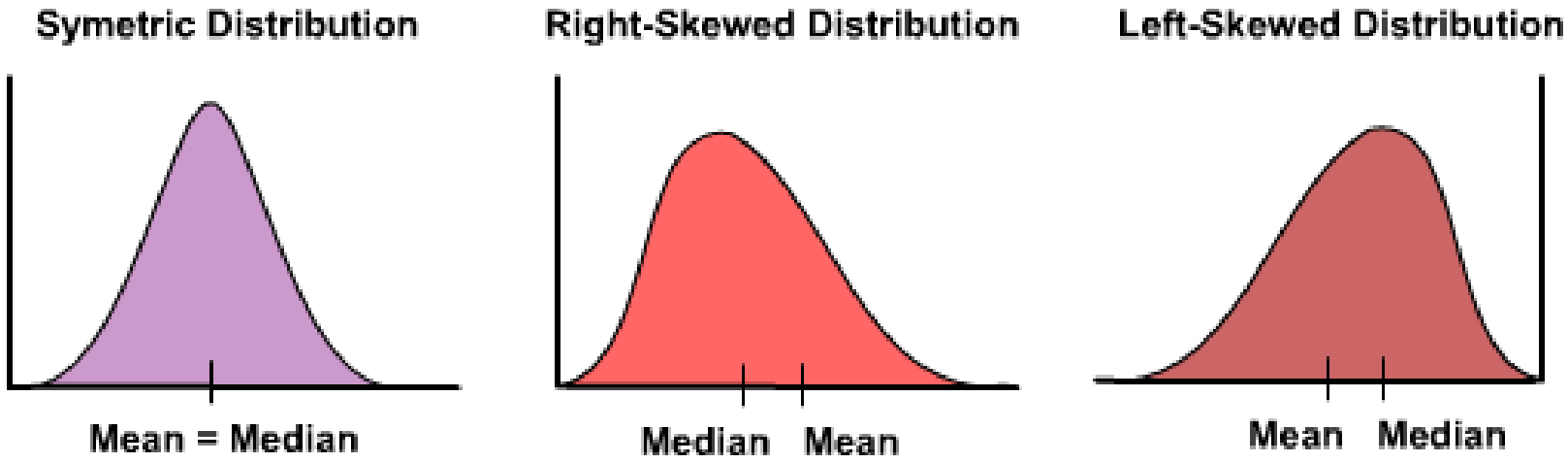
Graphic - Univariate – Skewness

Numerical Variables

Skewness

- Symmetric
- right-skewed
- left-skewed





Univariate Analysis
- Graphic Tools
  - Categorical variables
  - Numerical variables
- Descriptive Statistics
  - Categorical variables
  - Numerical variables
- Normality

Graphic - Univariate – QQ Plot

Numerical Variables

Normal quantile plot

One way to assess if a distribution is indeed approximately Normal is to plot the data on a Normal quantile plot (quantile-quantile (QQ) plot).

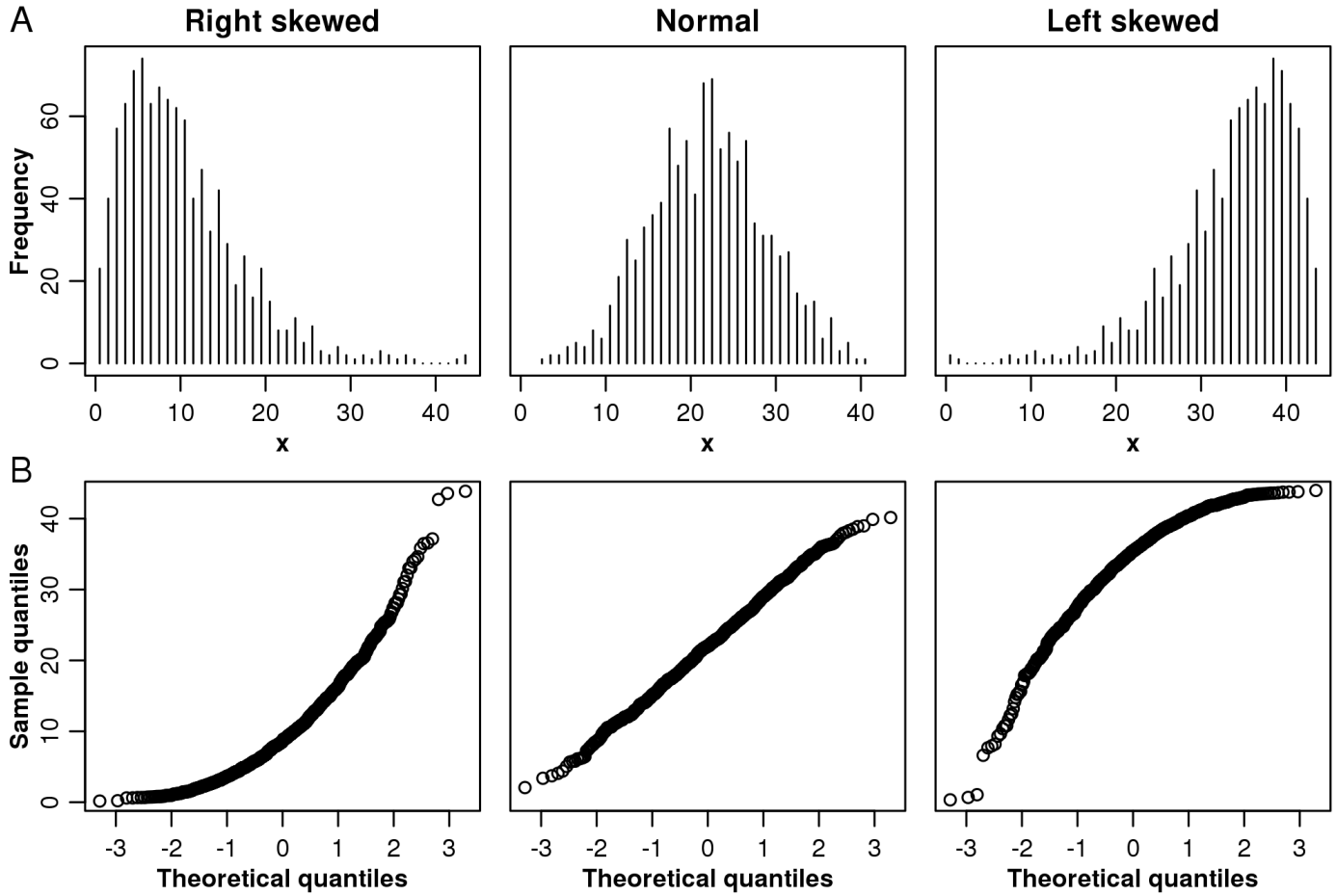
The z-scores are used for the x axis against which the data are plotted on the y axis.

- If the observed and the normal distribution are identical, points are expected to lie on a straight line.

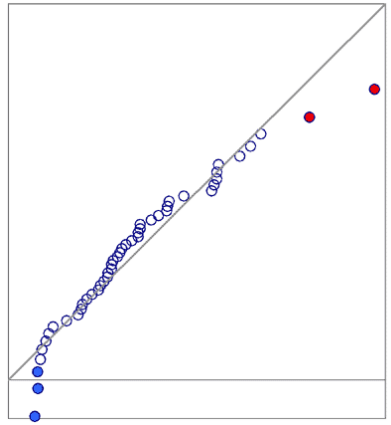
- Systematic deviations from a straight line indicate a non-normal distribution. **Outliers** appear as points far away from the overall pattern of the plot.

Univariate Analysis
- Graphic Tools
  - Categorical variables
  - Numerical variables
- Descriptive Statistics
  - Categorical variables
  - Numerical variables
- **Normality**

Graphic - Univariate – QQ Plot



Outliers

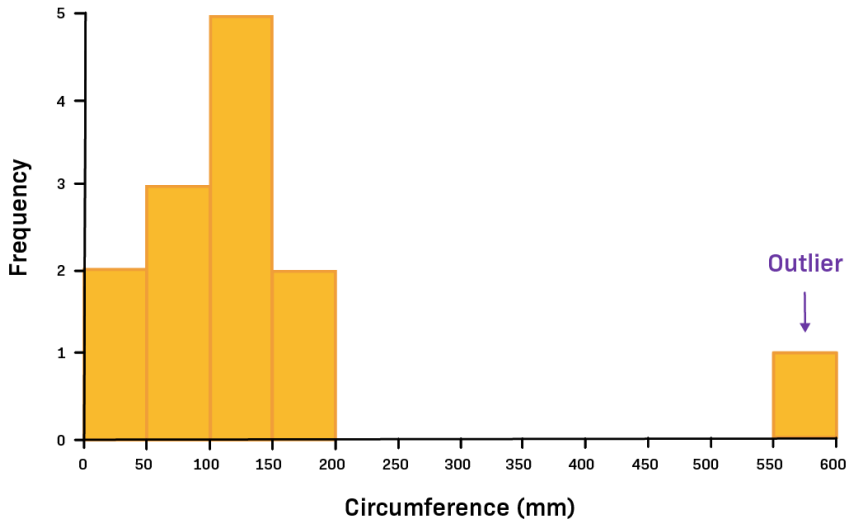


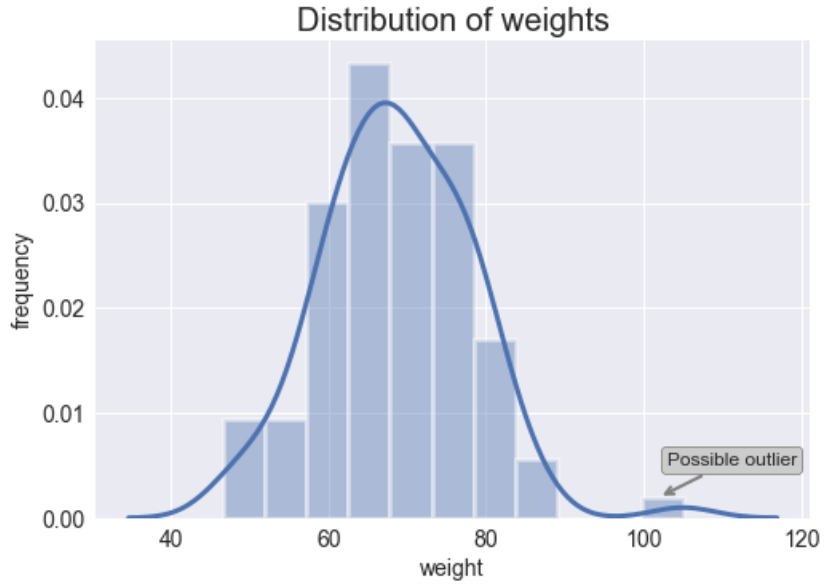
Univariate Analysis
- Graphic Tools
  - Categorical variables
  - Numerical variables
- Descriptive Statistics
  - Categorical variables
  - Numerical variables
- Normality

Graphic - Univariate - Outliers

Numerical Variables

Outliers are observations that lie outside the overall pattern of a distribution. Always look for outliers and try to explain them.



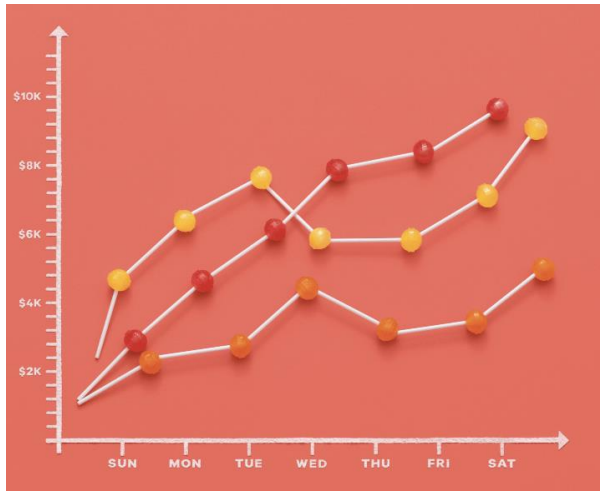


Univariate Analysis
- Graphic Tools
  - Categorical variables
  - **Numerical variables**
- Descriptive Statistics
  - Categorical variables
  - Numerical variables
- Normality

Graphic - Univariate - Line Plot

Numerical Variables

Line plots are most effective at showing a quantitative trend over time.



Univariate Analysis

- Graphic Tools
  - Categorical variables
  - **Numerical variables**
- Descriptive Statistics
  - Categorical variables
  - Numerical variables
- Normality

Descriptive - Univariate

Describing Distributions with Numbers

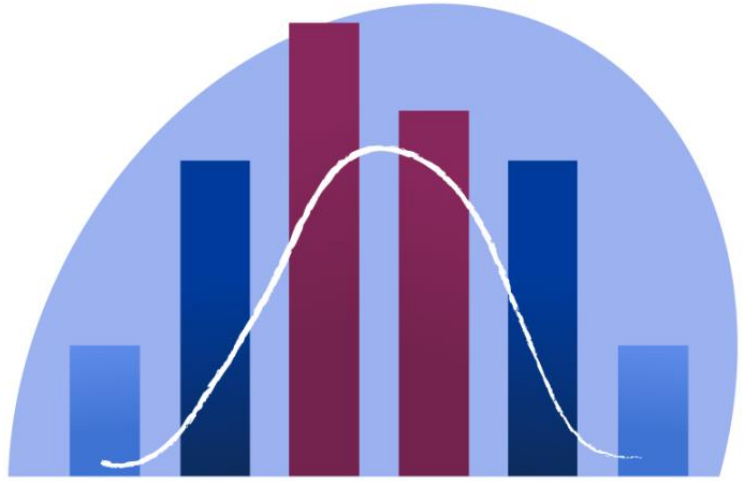
Type	Characteristic	Example	Descriptive statistic	Information content
Categorical	the set of all possible values can be enumerated			
• Nominal	Unordered categories	Gender, race	Counts, proportions	Lower
• Ordinal	Ordered categories	Degree of pain	Median	Intermediate
Continuous or ordered discrete	can take all possible values within some interval of real numbers (continuous) or limited to integers (discrete)	Weight, number of cigarettes per day	Mean, standard deviation	Higher

Univariate Analysis
- Graphic Tools
  - Categorical variables
  - Numerical variables
- **Descriptive Statistics**
  - Categorical variables
  - Numerical variables
- Normality

Descriptive - Univariate

Numerical summary

- Central tendency: mean; median; mode (useful for nominal data)
- Measuring dispersion: Five-number summary and boxplot
- Measuring variability: standard deviation
- Choosing among summary statistics

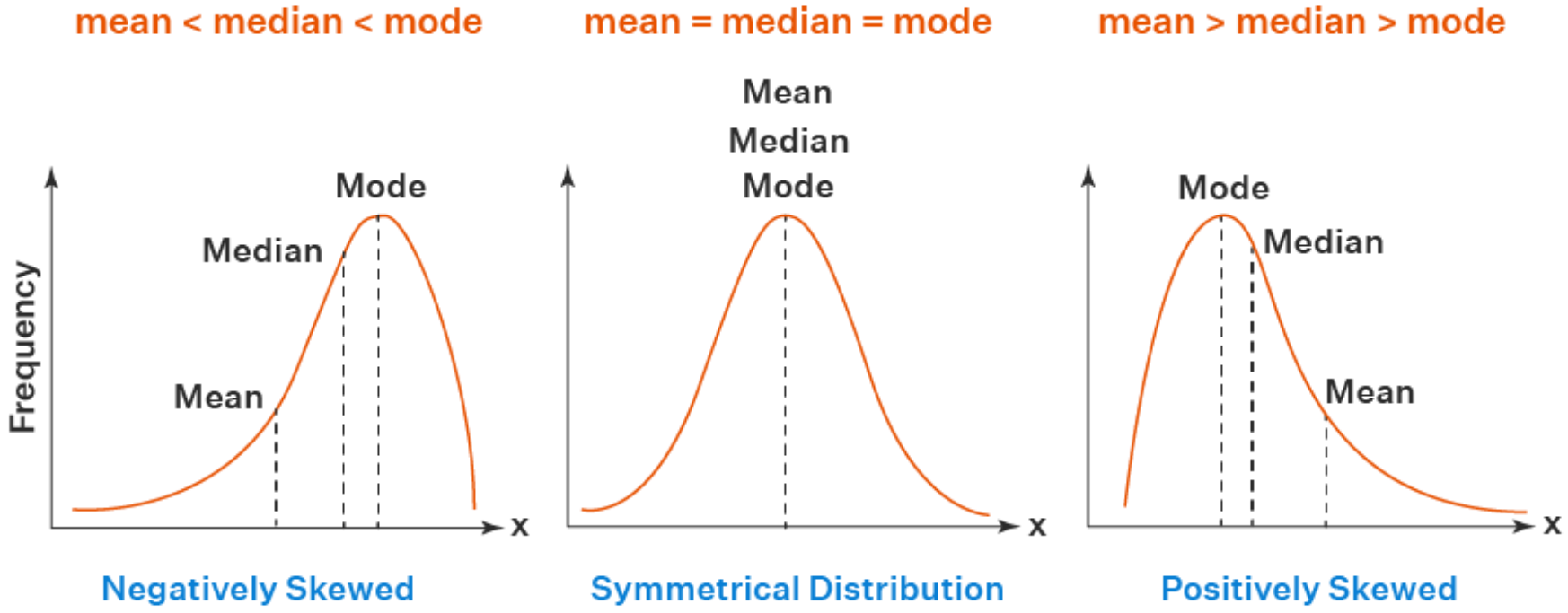


Univariate Analysis
- Graphic Tools
  - Categorical variables
  - Numerical variables
- **Descriptive Statistics**
  - Categorical variables
  - Numerical variables
- Normality

Descriptive - Univariate - Central

Central Tendency

- **Nominal level**: can only use the **mode** to find the most frequent value
- **Ordinal level or ranked data**: can use the **median** to find the middle value
- **Interval or ratio levels**: **mode and median**, and use the **mean** to find the average value



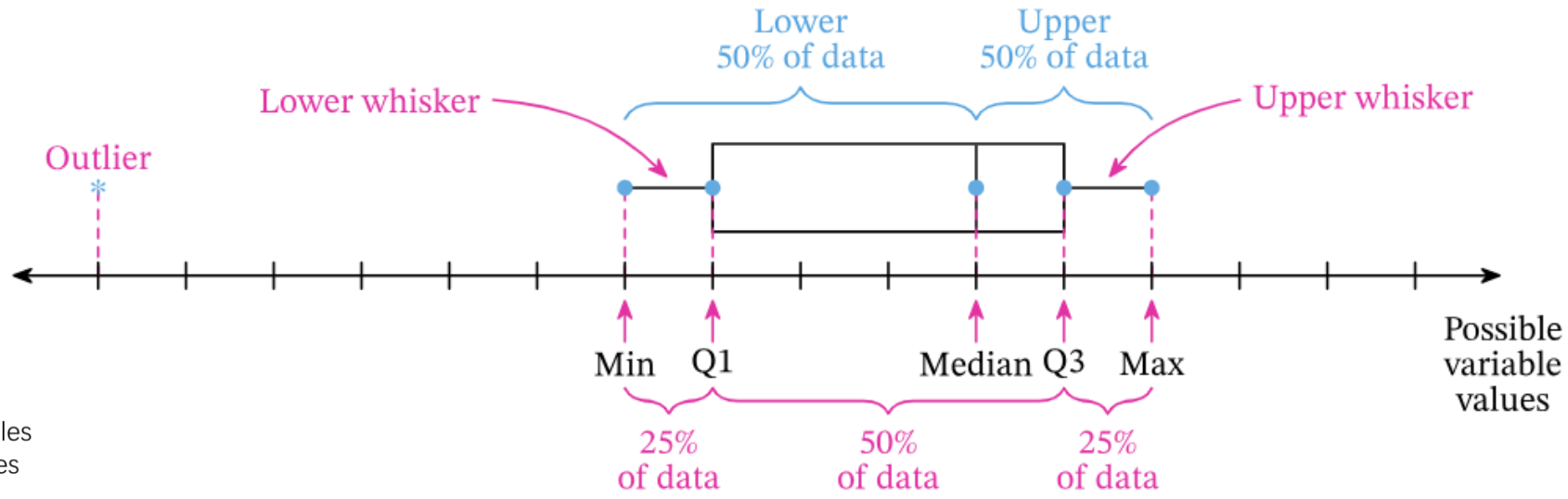
Univariate Analysis
- Graphic Tools
  - Categorical variables
  - Numerical variables
- **Descriptive Statistics**
  - Categorical variables
  - Numerical variables
- Normality

Descriptive - Univariate - Spread

Numerical Variables

The Five-Number Summary, Interquartile Range (IQR) and Boxplots

- The interquartile range (IQR) is defined as IQR = Q3 – Q1.
- A Boxplot is a graphic representation of those statistics.



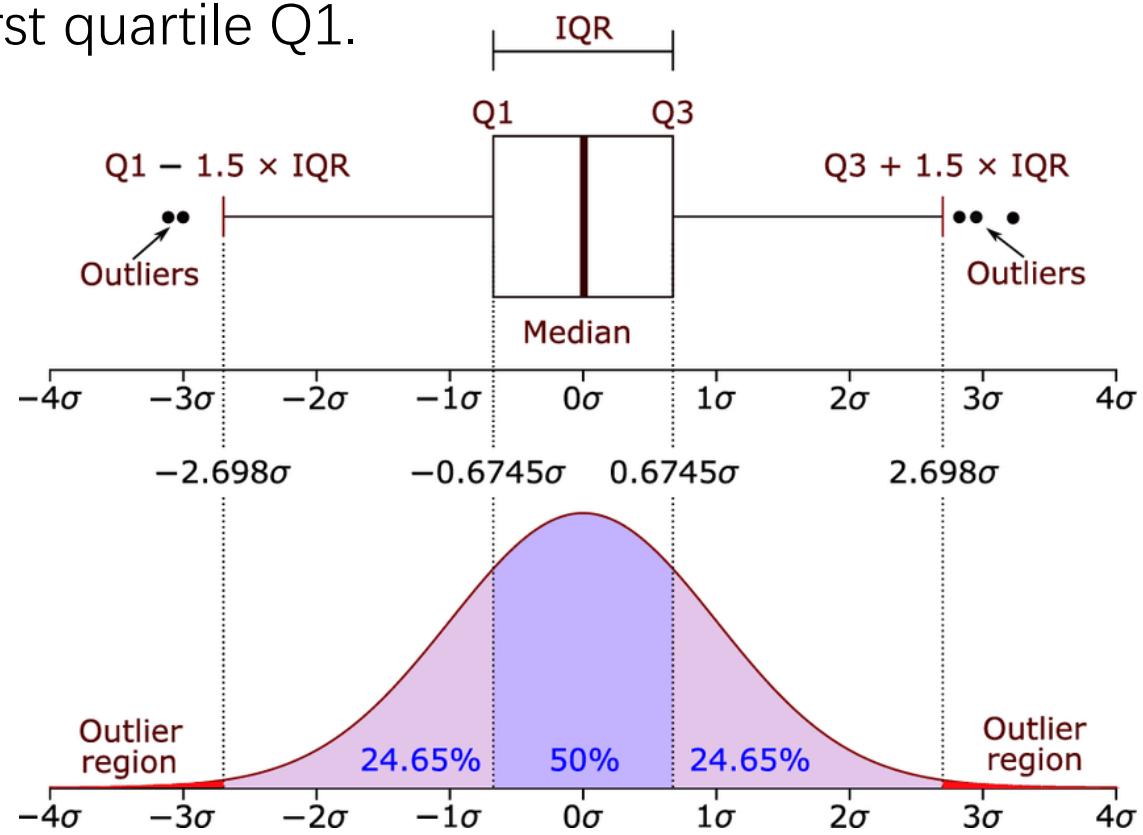
Univariate Analysis
- Graphic Tools
  - Categorical variables
  - Numerical variables
- Descriptive Statistics
  - Categorical variables
  - **Numerical variables**
- Normality

Descriptive - Univariate - Spread

Numerical Variables

Suspected Outliers: 1.5×IQR Rule

Call an observation an outlier if it falls more than 1.5 × IQR above the third quartile Q3 or below the first quartile Q1.



Univariate Analysis
- Graphic Tools
  - Categorical variables
  - Numerical variables
- Descriptive Statistics
  - Categorical variables
  - Numerical variables
- Normality

Descriptive - Univariate

Numerical Variables

Choosing Measures of Center and Spread

We now have a choice between two descriptions for center and spread:

- Mean and standard deviation
- Median and interquartile range

Use **median and IQR**: Skewed distribution or a distribution with outliers:
Use **mean and standard deviation**: Reasonably symmetric distributions that do not have outliers.

NOTE: Numerical summaries do not fully describe the shape of a distribution.

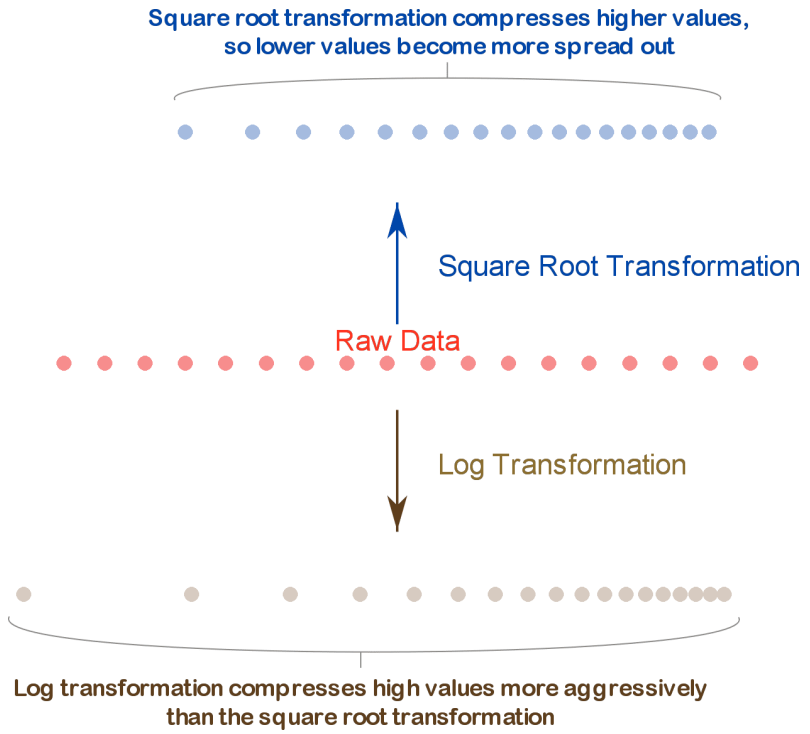
ALWAYS PLOT YOUR DATA!

Univariate Analysis
- Graphic Tools
  - Categorical variables
  - Numerical variables
- Descriptive Statistics
  - Categorical variables
  - **Numerical variables**
- Normality

Descriptive - Univariate - Transformation

Transformations

- One of the most common ways to deal with violations of normality
- Logarithmic or square root transformations, can help to achieve normality
- For example, if the data is skewed, taking the logarithm of the data may help to achieve a more normal distribution.



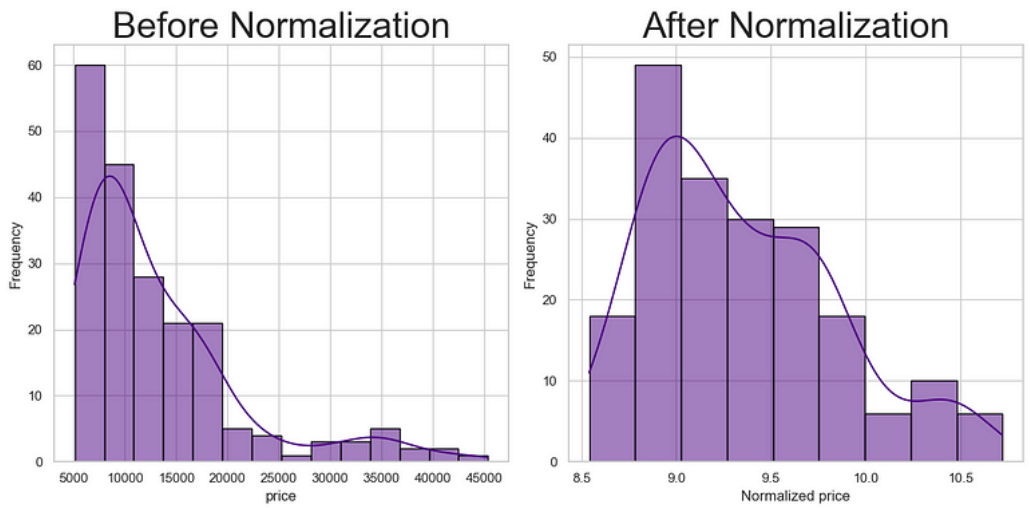
Univariate Analysis
- Graphic Tools
  - Categorical variables
  - Numerical variables
- Descriptive Statistics
  - Categorical variables
  - Numerical variables
- Normality

Descriptive - Univariate - Transformation

Log Transformations

- used to reduce the skewness of data, especially for data with a long tail. It's particularly effective when dealing with positively skewed data.

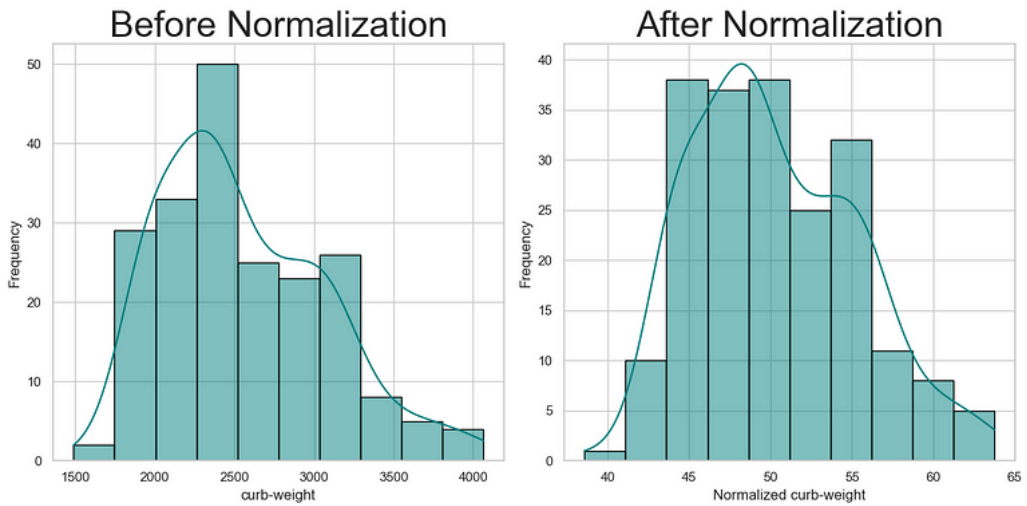
$$x_{log\ transformed} = log_b(x)$$



Square Root Transformations

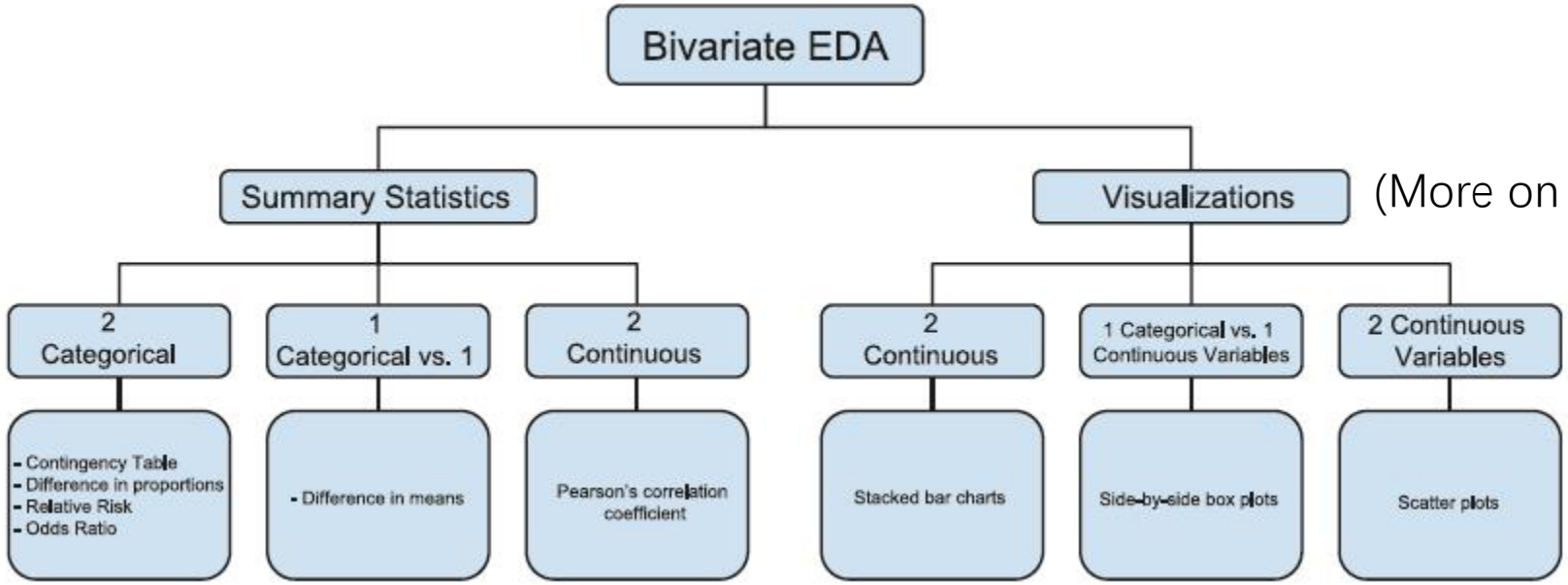
- milder than log transformation and is used when data includes zeros.

$$x_{square-root\ transformed} = \sqrt{x}$$



Looking at Data – Bivariate Analysis

- Bivariate Analysis
  - Graphic Tools
  - Descriptive Statistics
    - Numerical variables



(More on next page)

Graphic – Bivariate - Tools

Type of Variables (Vs.)	Categorical (incl. discrete numerical)	Continuous

Categorical (incl. discrete numerical)	• Frequency of the two categories/ other continuous variables' range
	  • Crosstab
	  • Heatmaps
	  • Stacked bar charts
		• Range of continuous variable with respect to each category
		  • Boxplots
		  • Violin plots
		  • Swam plots
		  • Count plots
		  • Bar plot

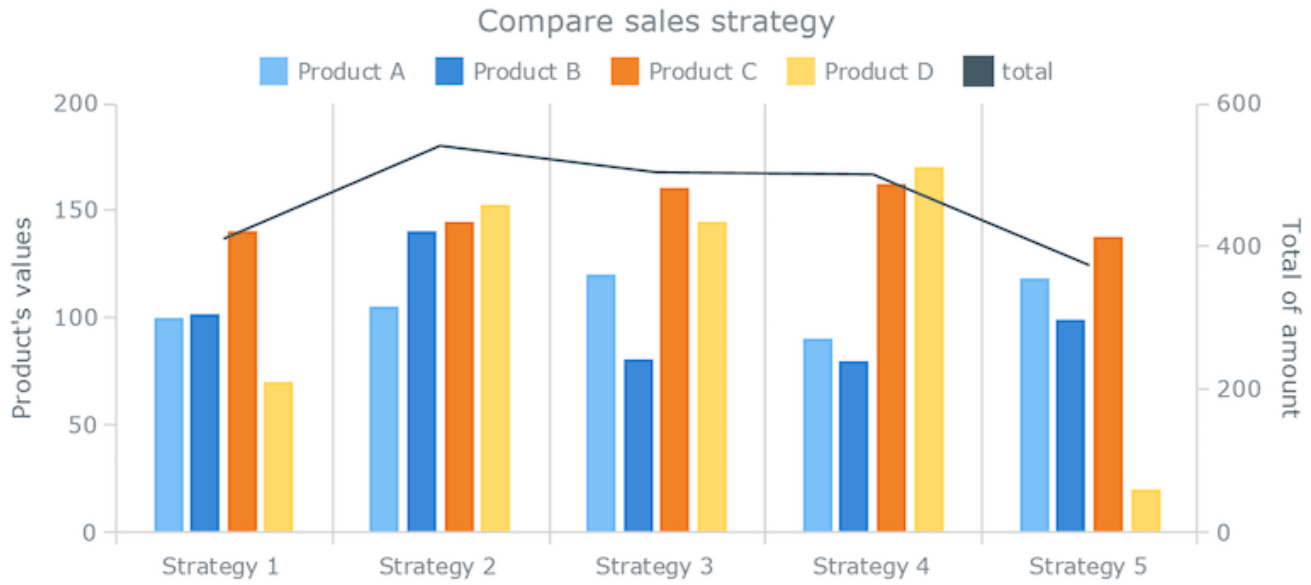
Continuous	• Range of continuous variable with respect to each category
	  • Boxplots
	  • Violin plots
	  • Swam plots
	  • Count plots
	  • Bar plot
		• How the increase or decrease in one variables changes with the other
		  • Scatterplot
		  • Line plots

- Bivariate Analysis
  - **Graphic Tools**
  - Descriptive Statistics
    - Numerical variables

Graphic – Bivariate – Multiseries Bar

Categorical Variables vs Categorical Variables

This multi-series bar chart displays sales of each product within each sales strategy.



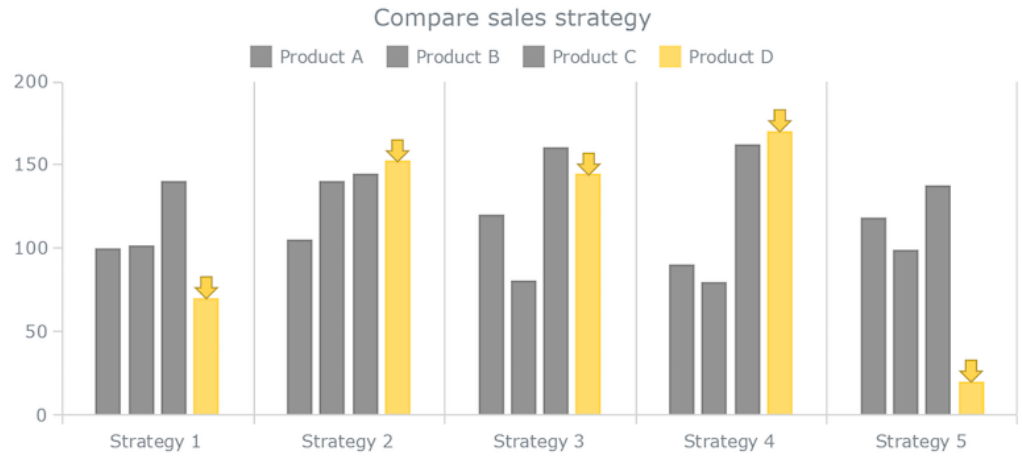
- Bivariate Analysis
  - **Graphic Tools**
  - Descriptive Statistics
    - Numerical variables

Graphic – Bivariate – Multiseries Bar

Categorical Variables vs Categorical Variables

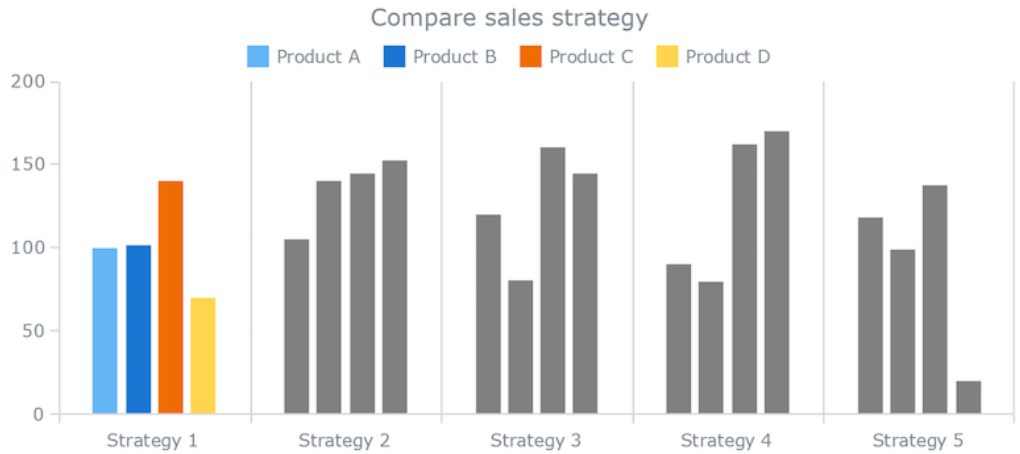
Answer the following questions:

- How did the products perform individually within a given strategy?



Product D was most successful in Strategy 4 and the least successful in Strategy 5

- Which strategy generated the most sales of every single product?



We also see that the biggest sales within Strategy 1 are attributable to Product C, whereas Product D finished last, and Products A and B are almost identical.

- Bivariate Analysis
  - **Graphic Tools**
  - Descriptive Statistics
    - Numerical variables

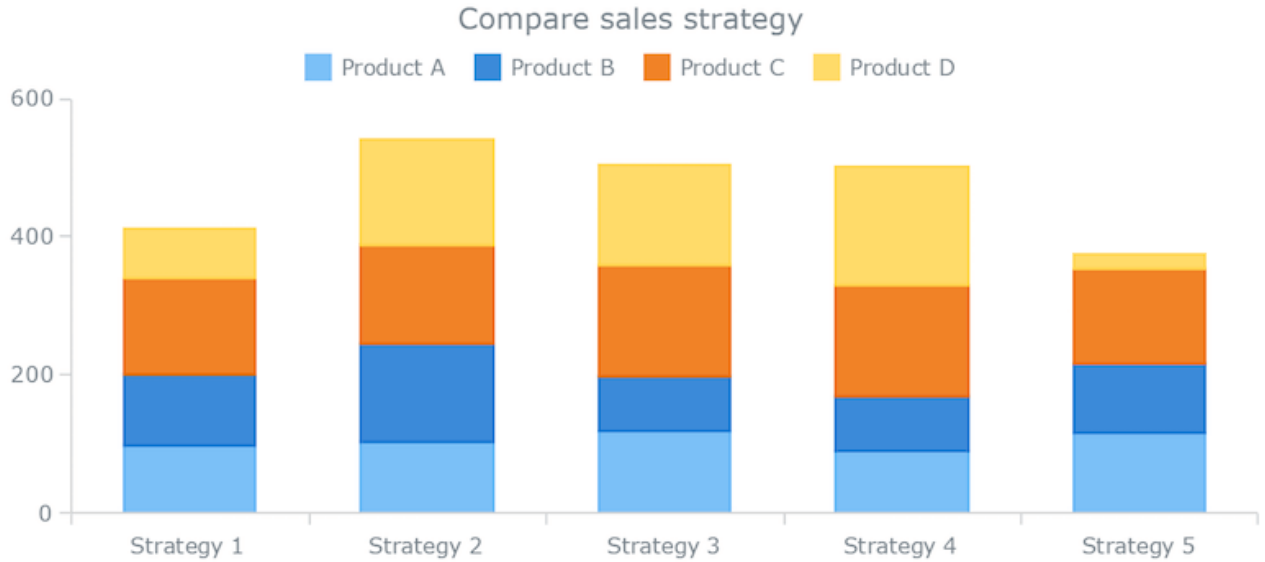
Graphic – Bivariate – Stacked Bar

Categorical Variables vs Categorical Variables:

Answer the following questions:

- What was the performance of a product within one strategy in relation to its sales within other strategies?

Stacked bar charts are designed to help you simultaneously compare totals and notice sharp changes at the item level that are likely to have the most influence on movements in category totals.



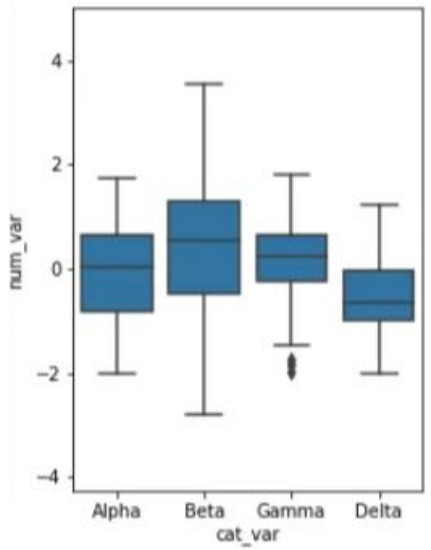
- Bivariate Analysis
  - **Graphic Tools**
  - Descriptive Statistics
    - Numerical variables

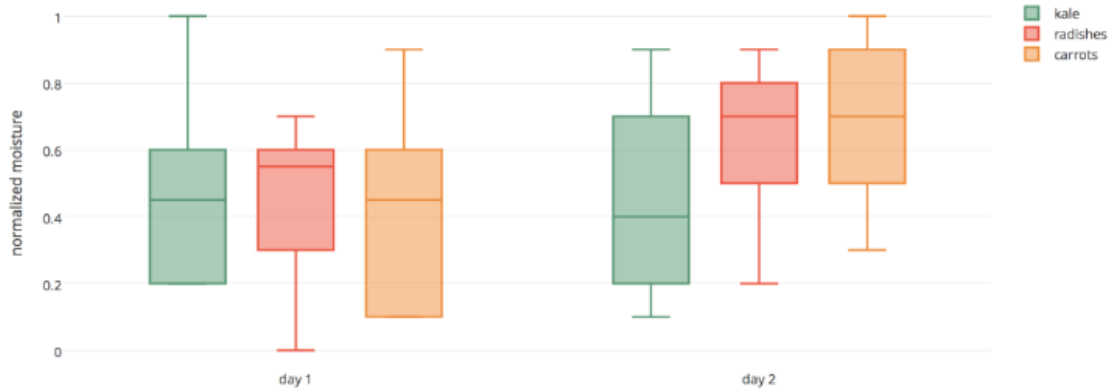
Better understanding of the big picture

Graphic – Bivariate – Boxplot

Categorical Variables vs Continuos Variables:

It summarize **numerical values across a category**; however, instead of just comparing the heights of the bar, they give us an idea of the range of values that each category can take.



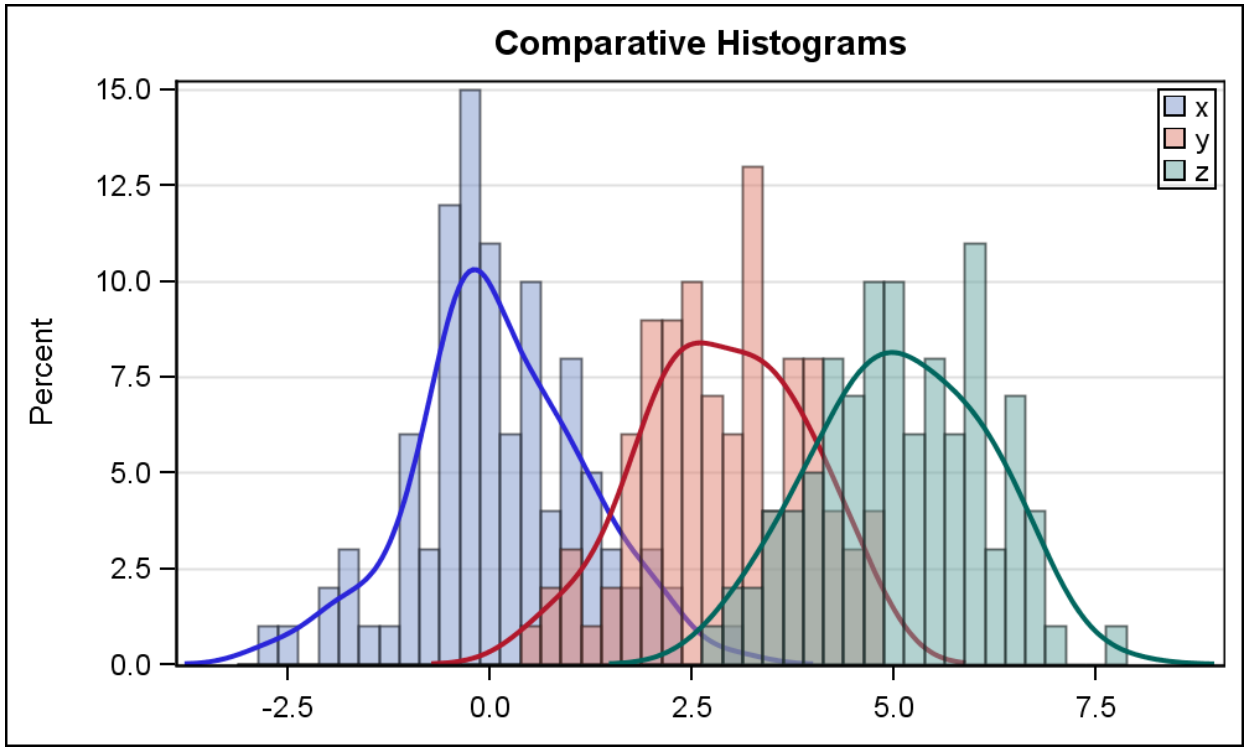


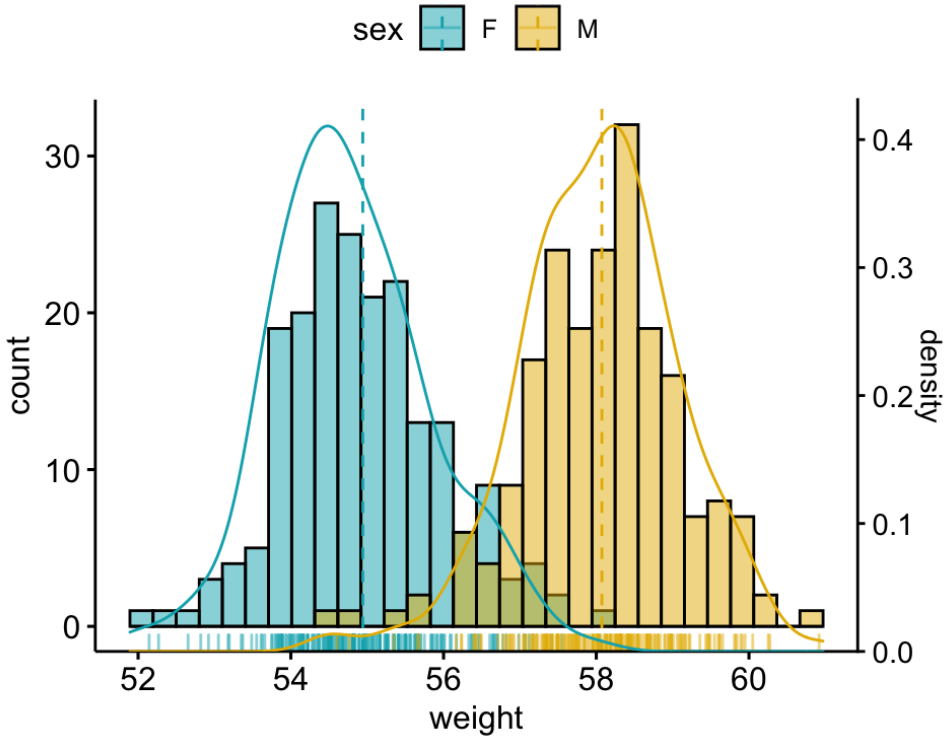
- Bivariate Analysis
  - **Graphic Tools**
  - Descriptive Statistics
    - Numerical variables

Box Plot: Summarizes data range
Swarm Plot: Shows data point positions
Violin Plot: Visualizes shape and density

Graphic – Bivariate – Density

Categorical Variables vs Continuos Variables:





- Bivariate Analysis
  - **Graphic Tools**
  - Descriptive Statistics
    - Numerical variables

Graphic – Bivariate – Scatterplot

Continuous Variables vs Continuos Variables:

Scatterplots are helpful when have **numerical values for two different pieces of information** and want to understand the **relationship** between those pieces of information.

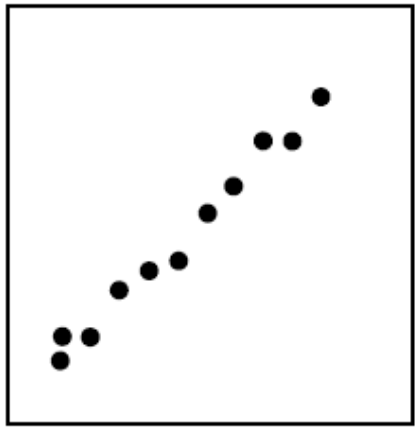
- Bivariate Analysis
  - **Graphic Tools**
  - Descriptive Statistics
    - Numerical variables

Graphic – Bivariate – Scatterplot

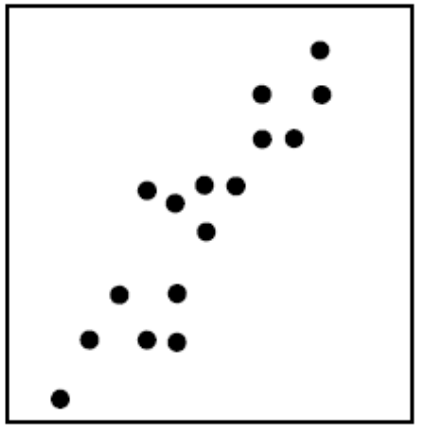
Describe the overall pattern by:

1. **Form** – a clear form such as **linear**
2. **Direction** – **positive**, **negative**
3. **Strength** – how **closely** the points follow a clear form

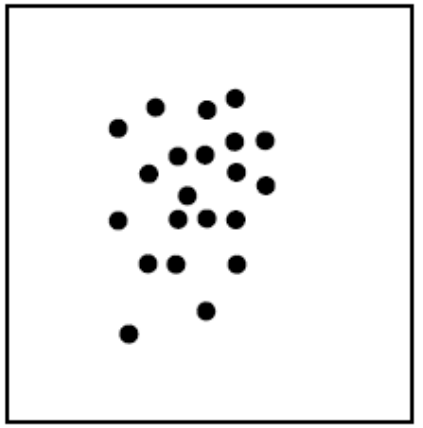
- Another important kind is an **outlier** that falls outside the overall pattern of the relationship.



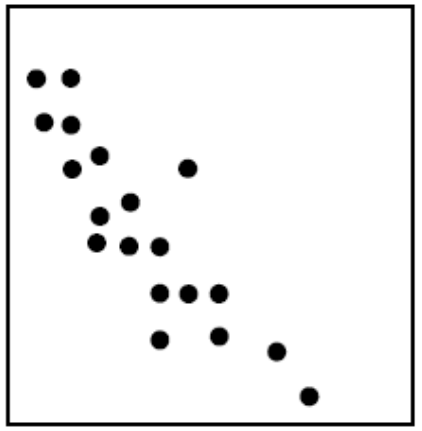
Strong positive correlation



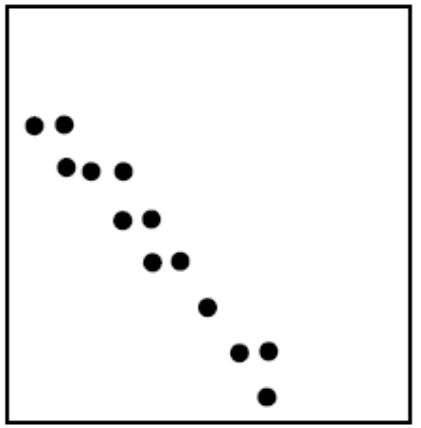
Moderate positive correlation



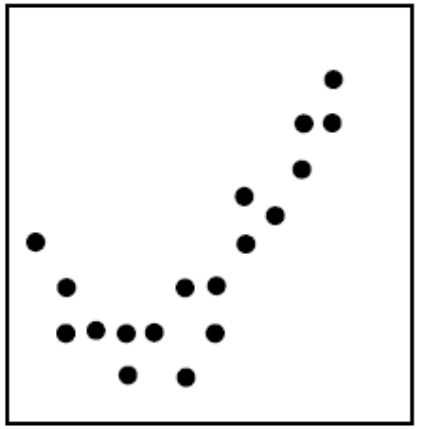
No correlation



Moderate negative correlation



Strong negative correlation



Curvilinear relationship

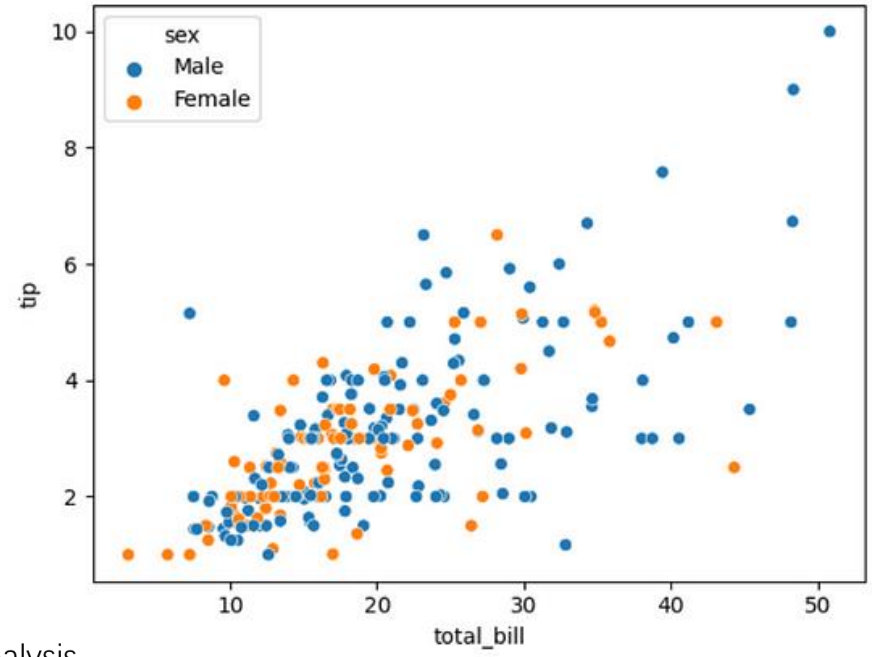
- Bivariate Analysis
  - **Graphic Tools**
  - Descriptive Statistics
    - Numerical variables

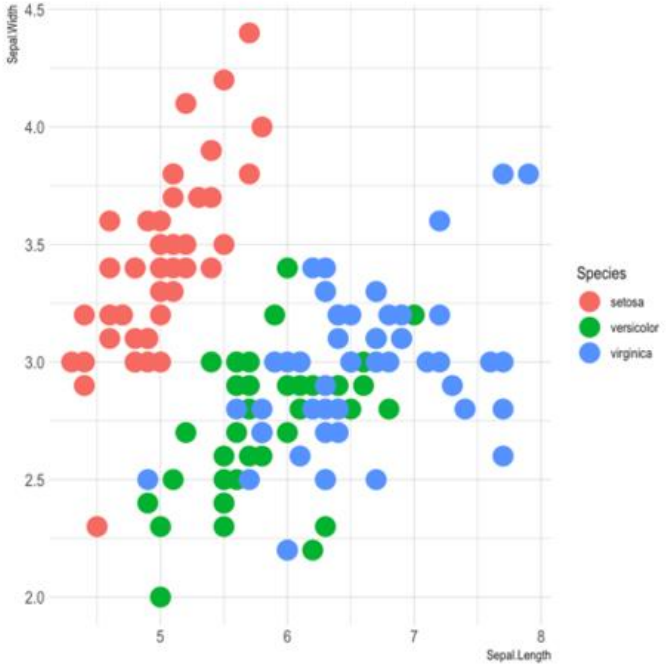
Graphic – Bivariate – Scatterplot

Contious Variables vs Continuos Variables:

Adding Categorical Variables

To add a categorical variable, use a different plot color or symbol for each category.





- Bivariate Analysis
  - **Graphic Tools**
  - Descriptive Statistics
    - Numerical variables

Descriptive – Bivariate – Table

Categorical Variables vs Categorical Variables

Two-way table describes two categorical variables, organizing counts according to a row variable and a column variable. Each combination of values for these two variables is called a cell.

Two-way table for "met requirement" and age

Met requirement	Age (years) 5 to 10	Age (years) 11 to 13	Total
No	194	557	751
Yes	861	417	1278
Total	1055	974	2029

- Bivariate Analysis
  - Graphic Tools
  - **Descriptive Statistics**
    - Numerical variables

Descriptive – Bivariate – Correlation

Continuous Variables vs Continuos Variables:

The **Pearson's correlation *r*** measures the **direction and strength** of the **linear relationship** between two quantitative variables. (-1 and 1)

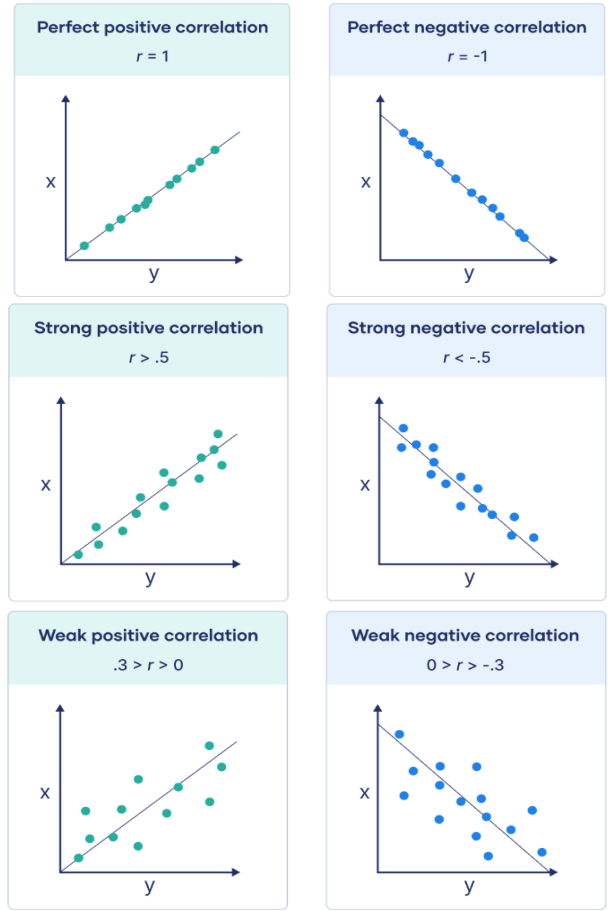
For variables x and y, for n individuals, the correlation r is defined as:

Pearson correlation coefficient (*r*) value	Strength	Direction
Greater than .5	Strong	Positive
Between .3 and .5	Moderate	Positive
Between 0 and .3	Weak	Positive
0	None	None
Between 0 and –.3	Weak	Negative
Between –.3 and –.5	Moderate	Negative
Less than –.5	Strong	Negative

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \cdot \sum_{i=1}^{n}(y_i - \bar{y})^2}} = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}},$$

where:

- $Sxy = \sum(x_i - x)(y_i - y) = \sum(xy) - \frac{\sum x \sum y}{n}$,
- $Sxx = \sum(x_i - x)^2 = \sum(x_i - x)(x_i - x) = \sum(x^2) - \frac{(\sum x)^2}{n}$,
- $Syy = \sum(y_i - y)^2 = \sum(y_i - y)(y_i - y) = \sum(y^2) - \frac{(\sum y)^2}{n}$.



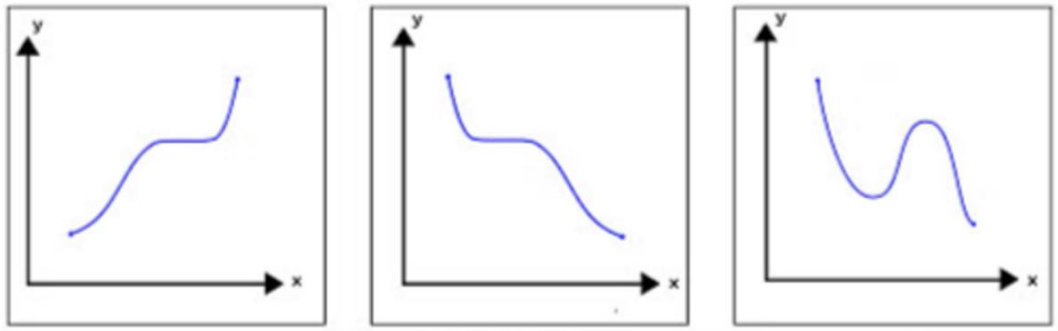
- Bivariate Analysis
  - Graphic Tools
  - **Descriptive Statistics**
    - Numerical variables

Descriptive – Bivariate – Correlation

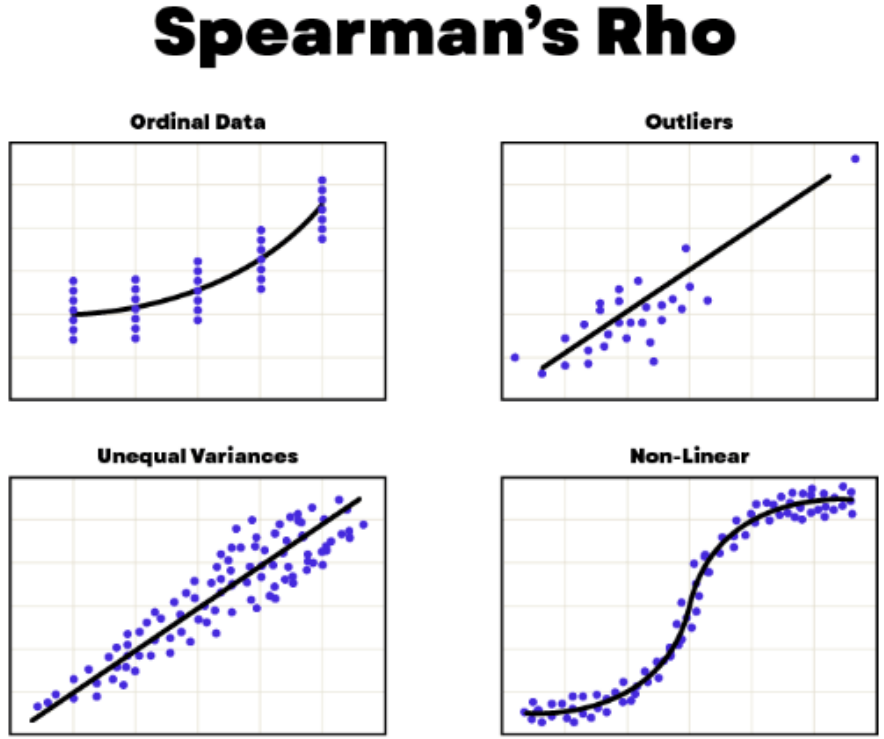
Ordinal Variables vs Ordinal Variables:

The **Spearman's Rank correlation** *r* measures of **monotone association** (extent to which as one variable increases, the other variable tends to increase or decrease) (-1 and 1)

- No assumption on linearity



Monotonic function



- Bivariate Analysis
  - Graphic Tools
  - **Descriptive Statistics**
    - Numerical variables

Descriptive – Bivariate – Correlation

Ordinal Variables vs Ordinal Variables:

The **Spearman's Rank correlation** *r*

d is the **difference in rank** between the two variables.

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

n is the **number of cases**

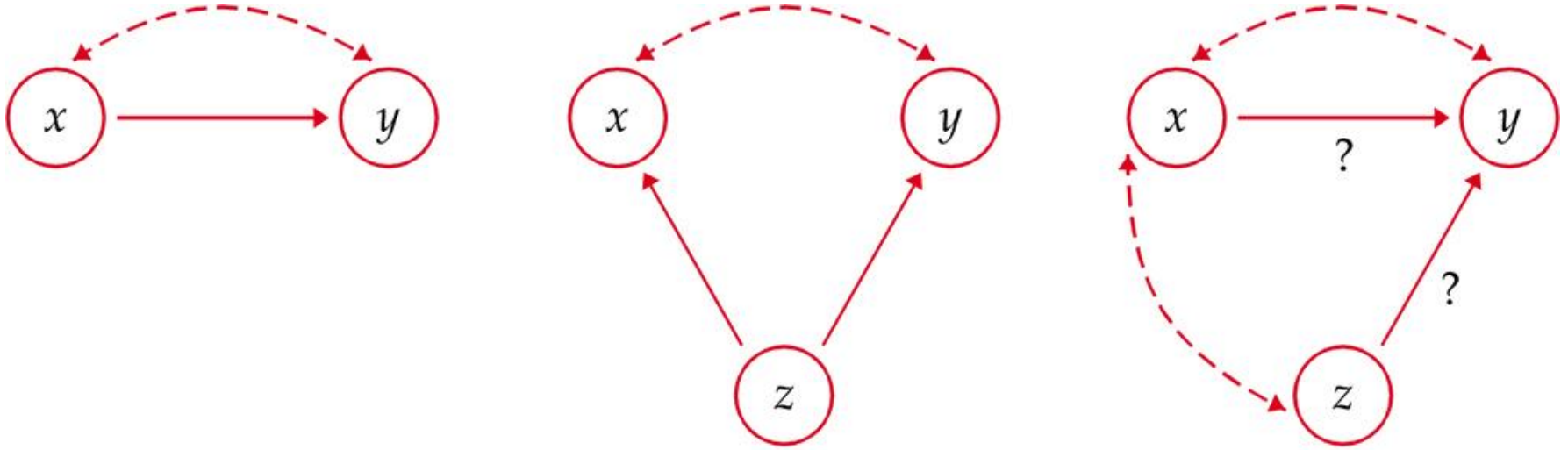
History	Rank	Geography	Rank	d	d square
35	3	30	5	2	4
23	5	33	3	2	4
47	1	45	2	1	1
17	6	23	6	0	0
10	7	8	8	1	1
43	2	49	1	1	1
9	8	12	7	1	1
6	9	4	9	0	0
28	4	31	4	0	0

- Bivariate Analysis
  - Graphic Tools
  - **Descriptive Statistics**
    - Numerical variables

Descriptive – Bivariate – Causation

Association, however strong, does NOT imply causation.

The dashed lines show an association. The solid arrows show a cause- and-effect link. x is explanatory, y is response, and z is a lurking variable.



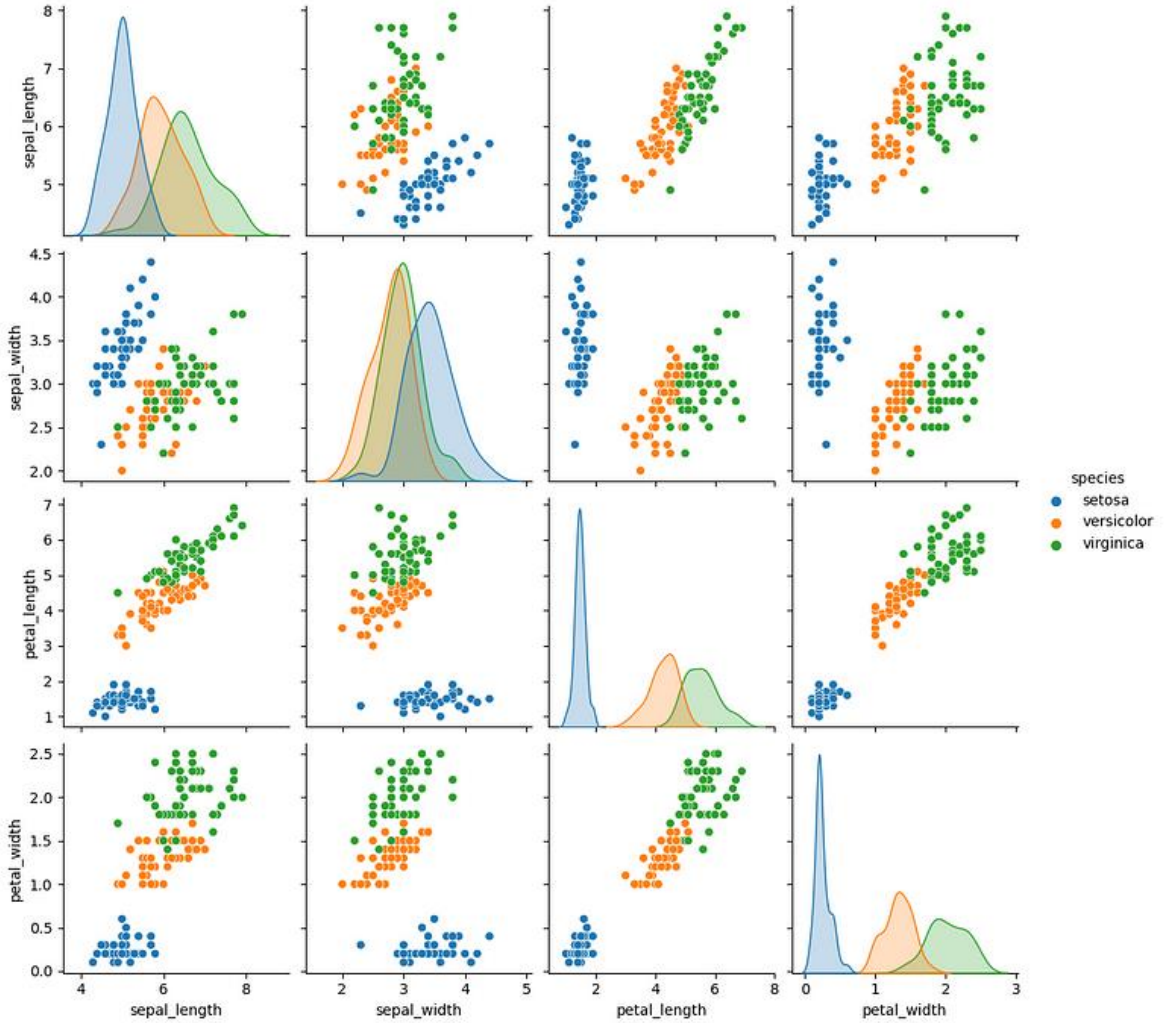
Causation | Common response | Confounding

"Beware the lurking variable"

- **Bivariate Analysis**
  - Graphic Tools
  - Descriptive Statistics
    - Numerical variables

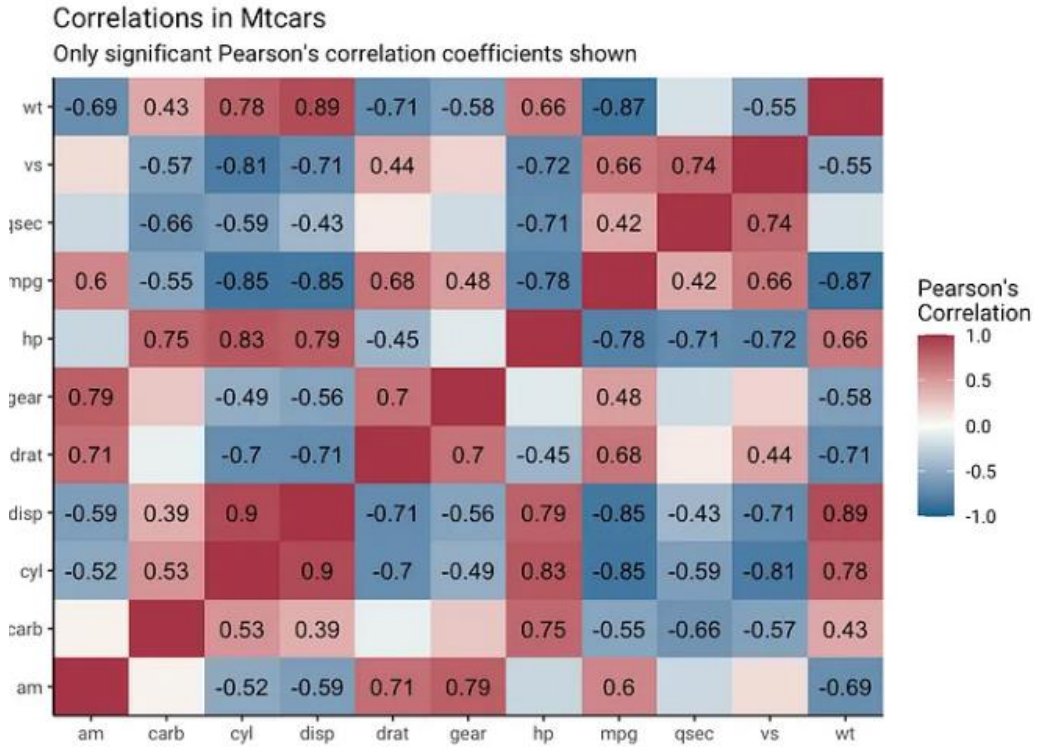
Looking at Data - Multivariate

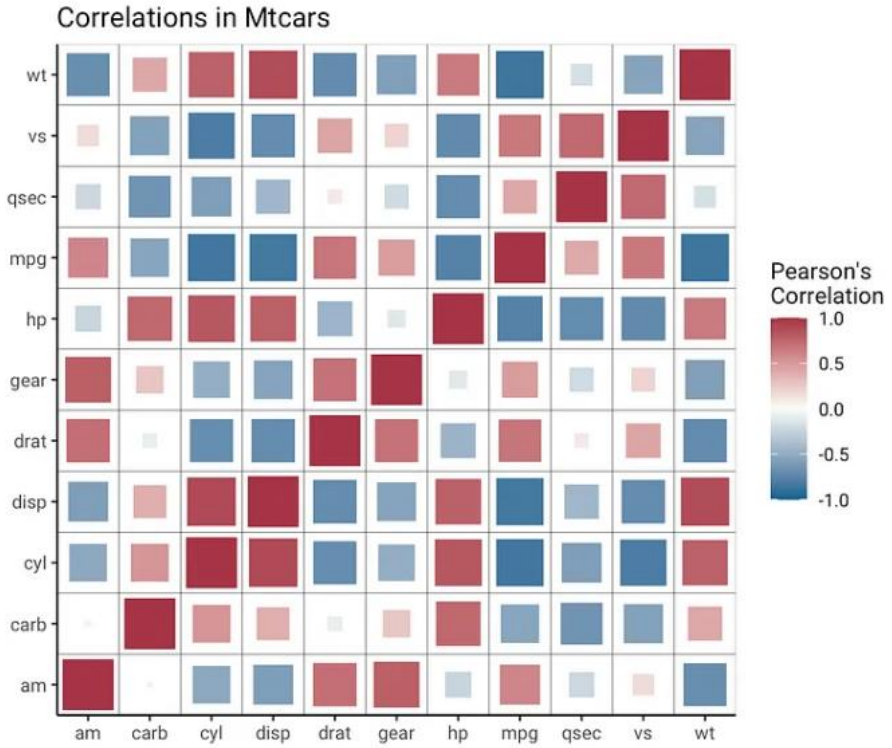
When dealing with multiple numerical variables, **pair plots** come in handy. They create a matrix of scatter/density plots automatically displaying all pairwise relationships between your variables.



Looking at Data - Multivariate

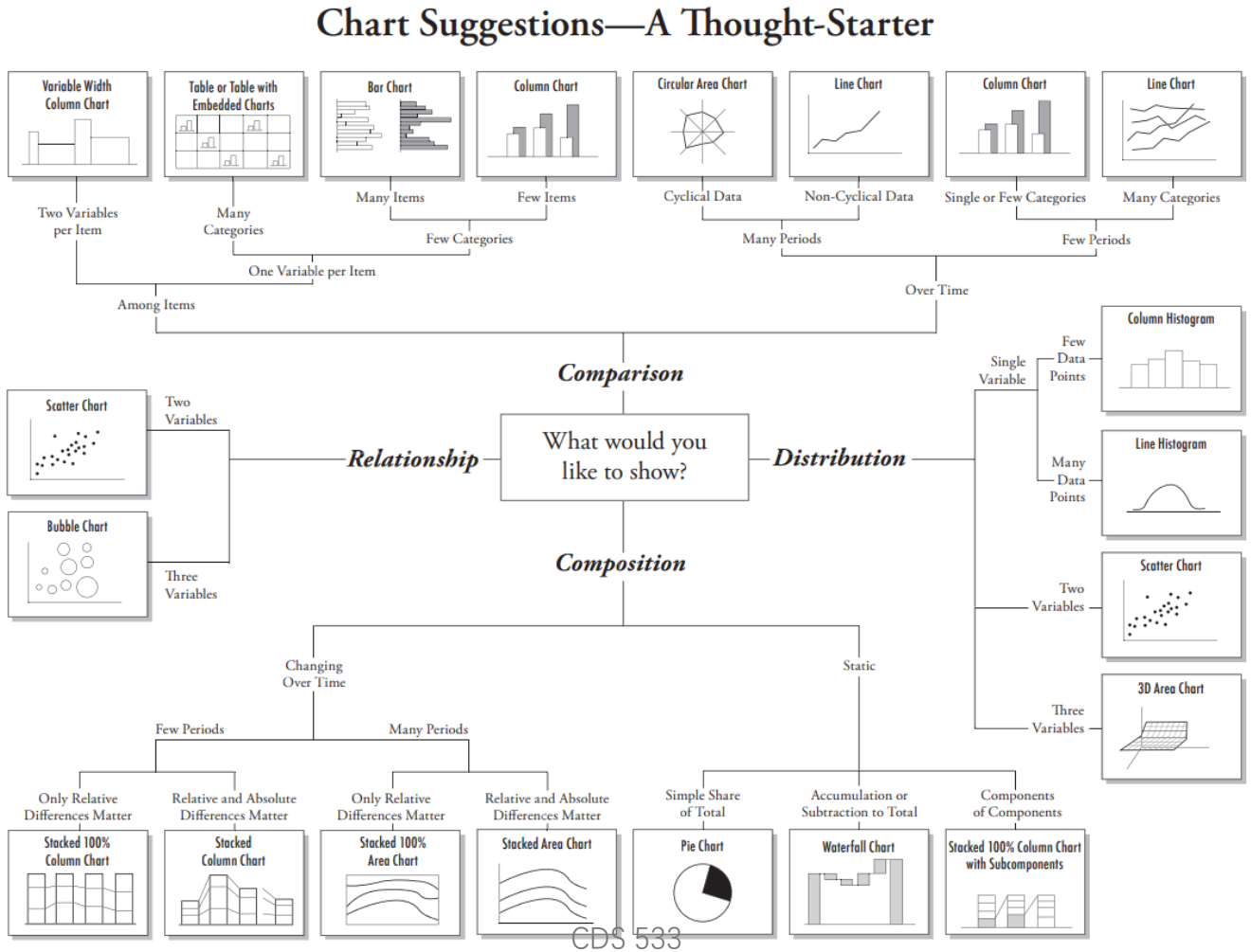
Heatmaps offer a visually striking way to explore these relationships, highlighting **strong positive and negative correlations** at a glance. This technique is particularly useful for large datasets with many variables, helping us navigate and identify potential areas of interest for further analysis.





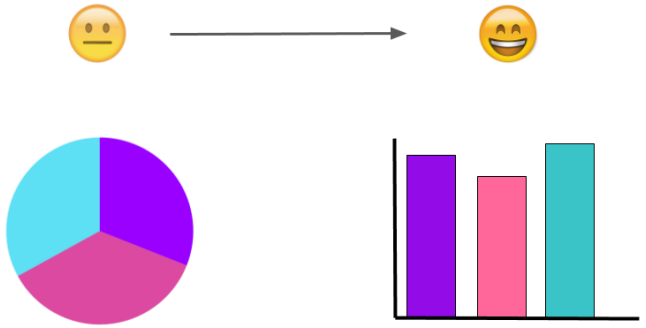
EDA- Data Visualization Tools

R Graph Gallery: https://r-graph-gallery.com/
Data Visualization Inspiration: https://datavizproject.com/



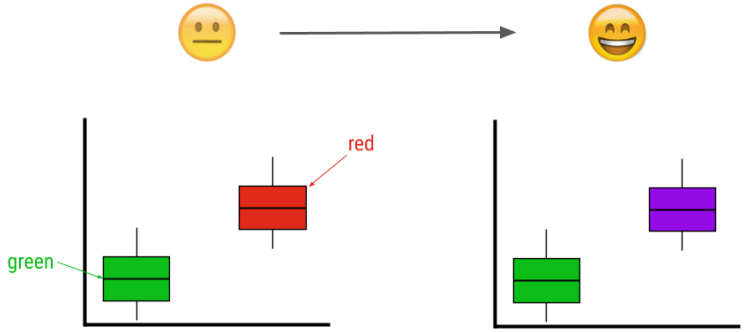
EDA- Data Visualization Tips

Choose the right type of plot



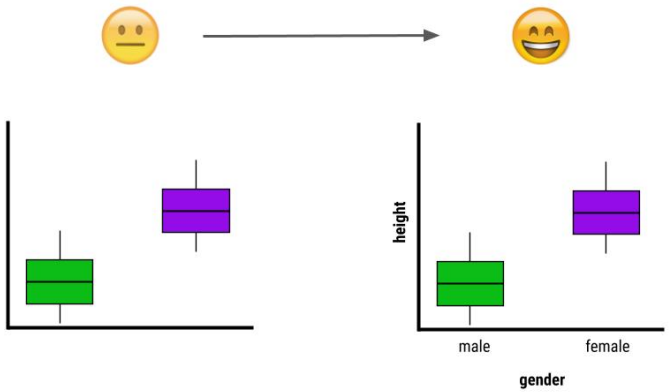
When looking at values, bar charts make it much easier to see the difference between groups!

Be mindful when choosing colors

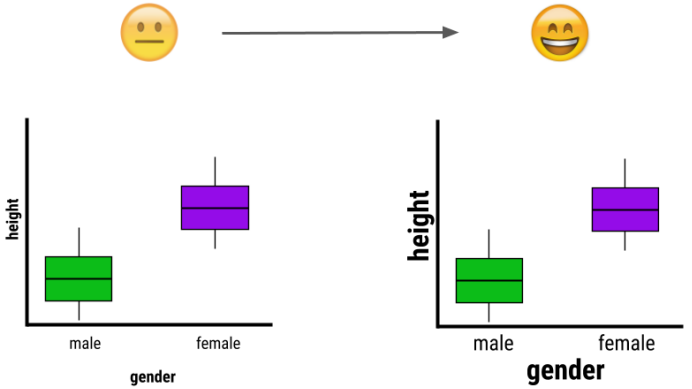


Many color-blind individuals cannot see the difference between red and green.

Label your axes!

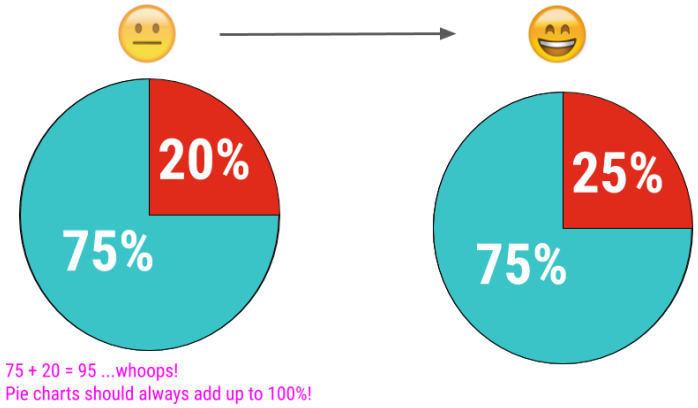


Make sure the text size is big enough!

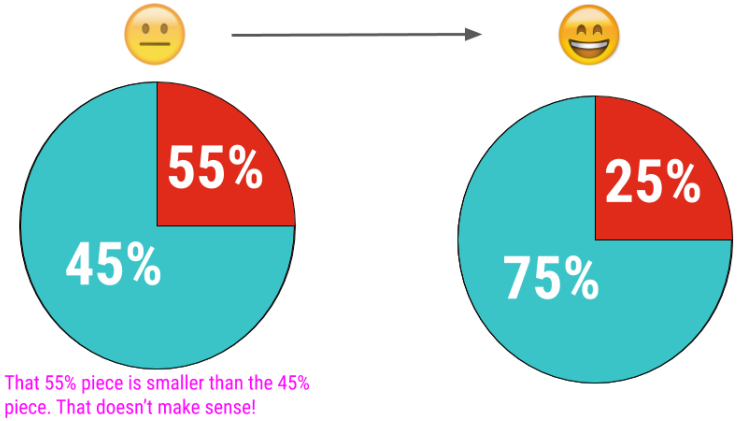


EDA- Data Visualization Tips

Make sure your numbers add up!

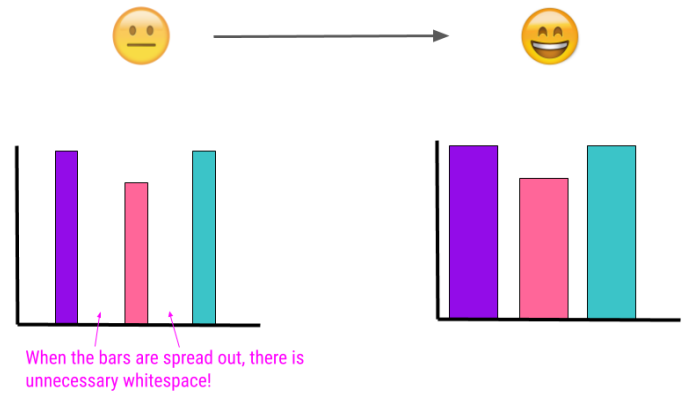


Make sure the numbers and graphic represent the data

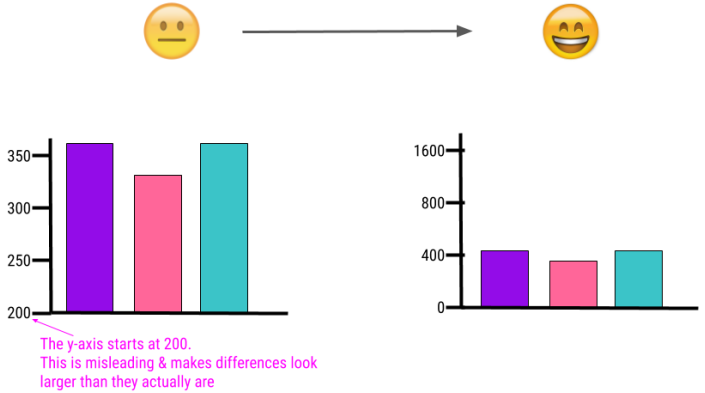


Make comparisons easy on your readers

Avoid unnecessary whitespace

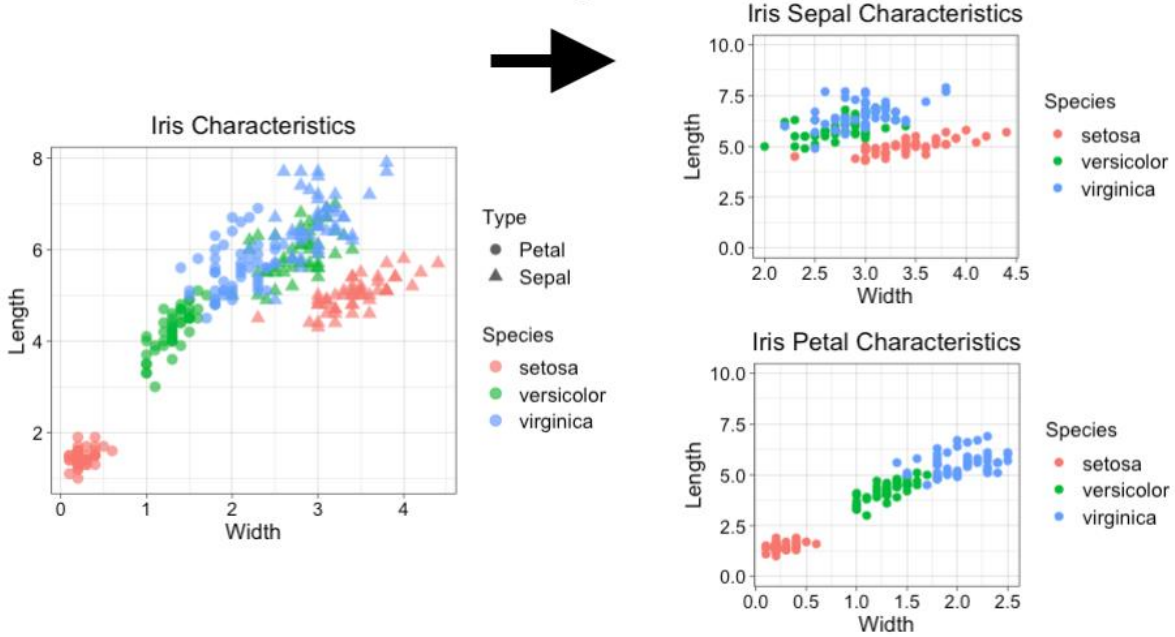


Use y-axes that start at 0 for barplots

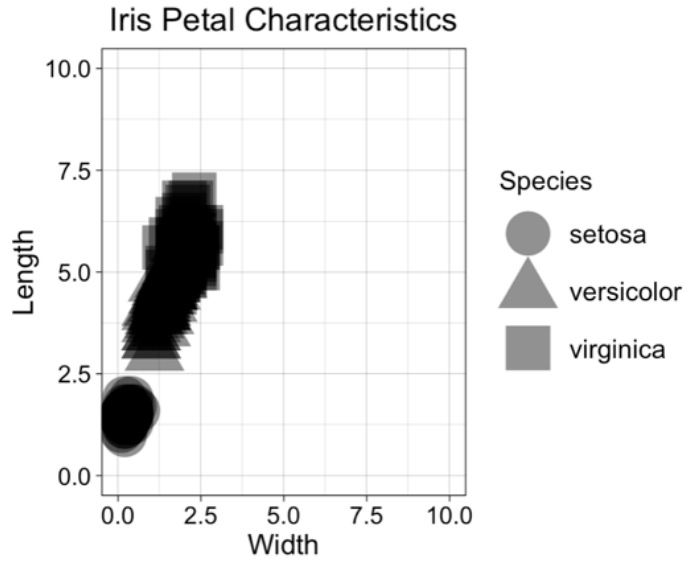


EDA- Data Visualization Tips

Keep it Simple



Don't Mislead!



Review

www.kahoot.it

Lab Time