TOINE BOGERS
AALBORG UNIVERSITY COPENHAGEN

USER STUDIES & INFORMATION BEHAVIOR

# DATA ANALYSIS: DESCRIBING & VISUALIZING DATA

# OUTLINE

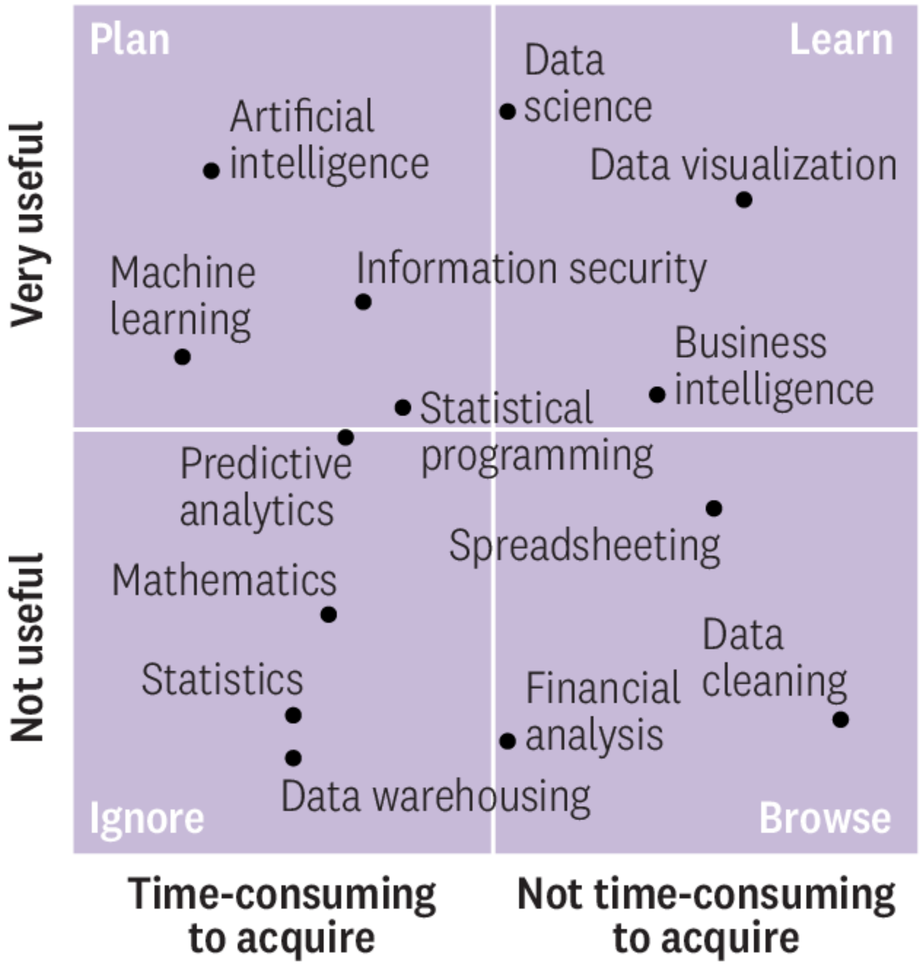


- Introduction
- Measuring variables
- Descriptive statistics
  - Measures of central tendency
  - Measures of dispersion
  - Why data visualization?
- Data visualization
  - What are our options and when do we use them?
  - How do we visualize data using `ggplot2` in R? (Florian's part)
- Pitfalls

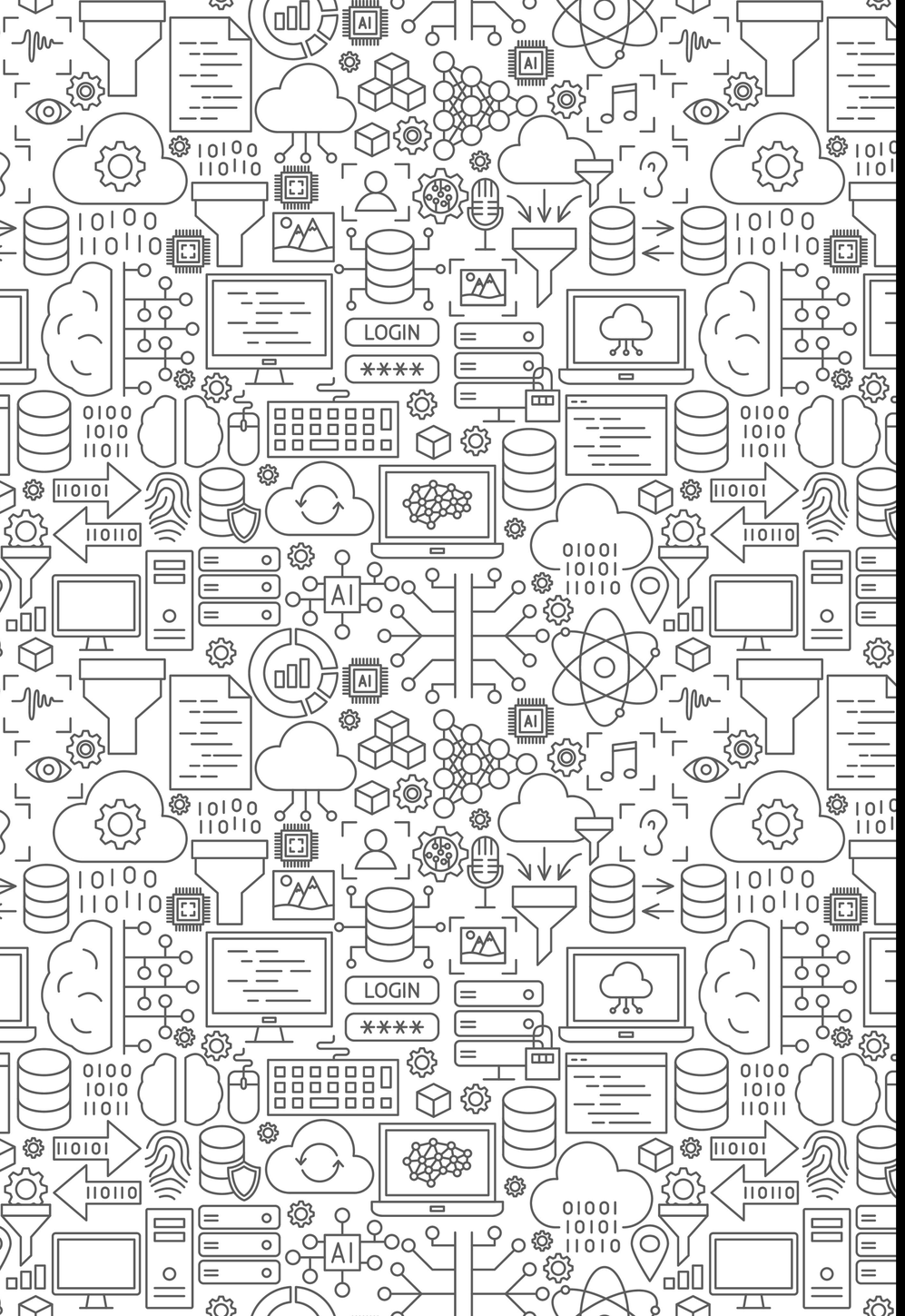
# WHY IS THIS RELEVANT?



- Data visualization skills are among the most high-priority skills to acquire




Littlewood (2018)




PART 1

# INTRODUCTION

# DESCRIPTIVE STATISTICS



- A descriptive statistic is a **summary statistic** that quantitatively describes or summarizes part(s) of our data set
  - Descriptive statistics is the process of using and analyzing those statistics to help us…
    - …**summarize** our data
    - …identify **outliers, anomalies, and human errors**
    - …understand **which statistical tests are appropriate**
- Tools at our disposal
  - Measures of central tendency
  - Measures of dispersion
  - Graphical & non-graphical representations

# INFERENTIAL STATISTICS

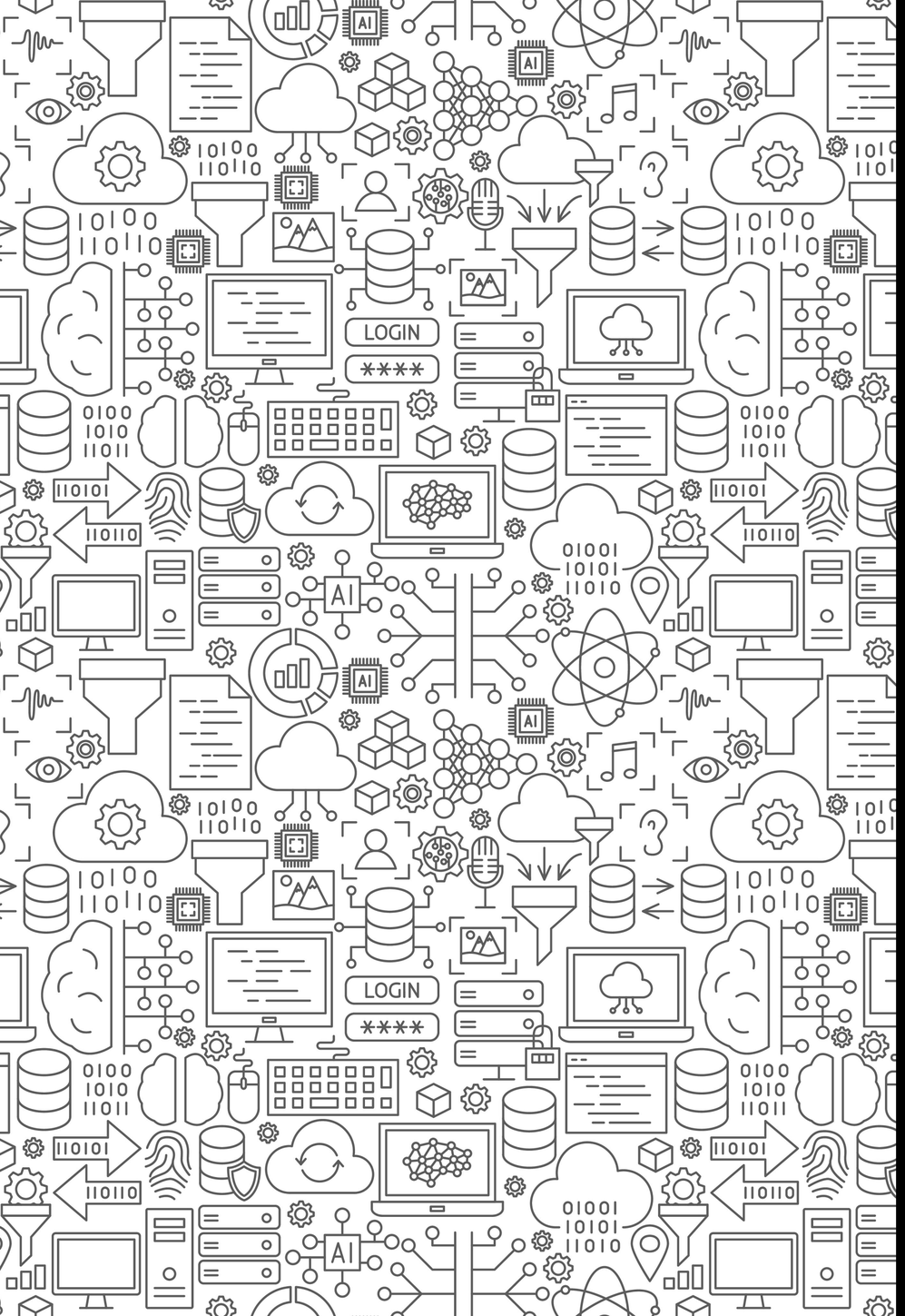


- Descriptive statistics is only focused on **describing properties** of our data **sample**
- Inferential statistics allows us to make **inferences (= predictions)** about the **population** using statistics from the sample
  - How likely is it that what we observed in our sample also holds in the real world?
  - Helps us reject or accept research hypotheses based on our cross-sectional or experimental results
  - Also known as **inductive statistics**
  - Basics of inferential statistics are covered in the '*Introduction to Data Science*' elective

# STATISTICAL LITERACY



- **Statistical literacy** is your ability to **understand and reason with statistics and data**
- Why is this important?
  - During your time at AAU, it will help you analyze the quantitative data you collect
  - In your professional life, it will be beneficial in helping you analyze behavioral, marketing or social media data
  - Statistical literacy is necessary for citizens to understand and critically evaluate material presented in publications such as newspapers, television, and the Internet




PART 2

# MEASURING VARIABLES

# TYPES OF VARIABLES



- We can distinguish between different **types** of variables (Bryman, 2016, p. 42)
  - What **role** do they play in our study?
    - Independent, dependent and hidden variables

    } covered in 'Research designs' lecture
  - At what **level** do we **measure** them (i.e., how do we operationalize them?)
    - Refers to the **relationship between the values** of a variable
    - Influences how the data can be interpreted
    - Influences which statistical measures and analysis methods are allowed

# LEVELS OF MEASUREMENT



- **Discrete** variables (a.k.a. qualitative or categorical variables)
  - Discrete variables contain a **fixed number of categories or distinct groups** and are **selected**, not measured
  - **Nominal**
  - **Ordinal**
- **Continuous** variables (a.k.a. quantitative variables)
  - Continuous variables contain an **unlimited number of possible values** and are **measured**, not selected
  - **Interval**
  - **Ratio**

# NOMINAL LEVEL



- **Nominal** variable
  - Each value represents a **unique** category (or name) from a fixed set
    - If there are only two values, it is be called a **dichotomous** variable
    - Even though Bryman (2016) sees them as two different types of variable
  - There is **no relative ordering** between the different categories
  - Objects have something in common if they share the same value
    - E.g., same gender, same country of origin

# NOMINAL LEVEL



- Examples
  - Gender
  - Hair color
  - Religion
  - Jersey numbers
  - Person names
  - Hash tags


# ORDINAL LEVEL

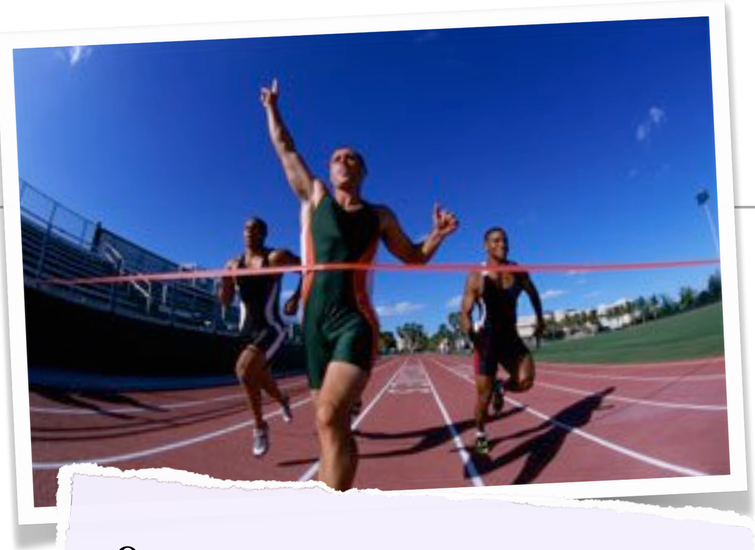


- Ordinal variable
  - Each value represents a **unique category** (or name) from a fixed set
  - Values are part of an **ordered set** and can be ordered by
    - Superiority
    - Intensity
    - Temporal position
    - Etc.

# ORDINAL LEVEL



- Examples
  - Restaurant ratings (from ⭐☆☆☆☆ to ⭐⭐⭐⭐⭐)
  - Outcome of a race
  - Level of education
  - Likert scales
    - Depending on the scale and on whom you ask!



0 = less than high school
1 = some high school
2 = high school degree
3 = some college
4 = college degree
5 = post college

Orange is the best color ever!

| Strongly disagree | Disagree | Neutral | Agree | Strongly agree |
|---|---|---|---|---|
| O | O | O | O | O |

# INTERVAL LEVEL

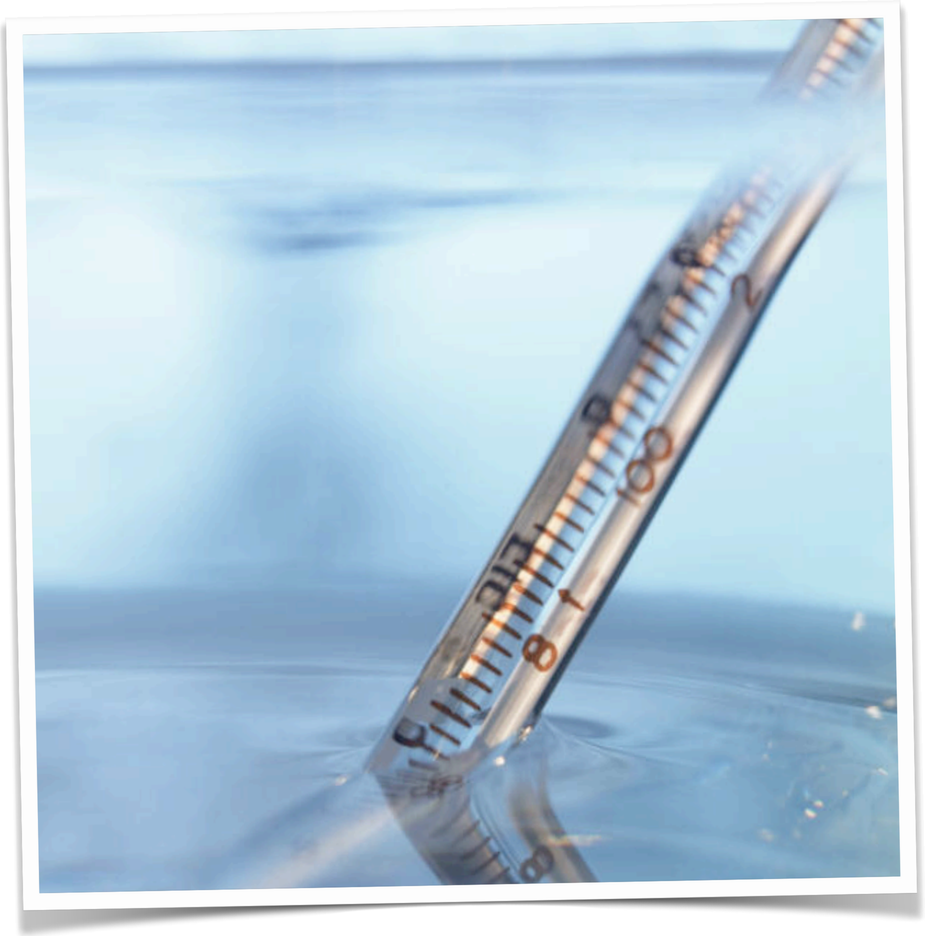


- Interval variable
  - **Measure** values along an **ordered scale** (and thus continuous)
  - Each position is **equidistant** from one another
  - Multiplication/division is **not allowed**
  - Zero value has been chosen arbitrarily
  - Fairly rare level of measurement

# INTERVAL LEVEL



- Examples
  - Temperature in Celsius
  - IQ scores
  - Dates
  - Likert scales
    - Likert scales can be treated as interval/ continuous if they have **seven levels or more** (Tabachnick & Fidell, 2013, p. 7)



# INTERVAL LEVEL



**Q1: I sometimes worry about how much time I spend on Instagram**

- ○ Strongly agree
- ○ Moderately agree
- ○ Agree a little
- ○ Neither agree nor disagree
- ○ Disagree a little
- ○ Moderately disagree
- ○ Strongly disagree

seen as same distance by people

→ **interval variable**

**Q2: How much pain are you in right now?**

- ○ No pain at all
- ○ A tiny bit
- ○ A little
- ○ Quite a bit
- ○ A lot
- ○ An excruciating amount!
- ○ This is unbearable!

not seen as the same step size!

→ **ordinal variable**

# RATIO LEVEL



- Ratio variable
  - Values on ordered, continuous scale, with equidistant spacing
  - Values can be compared as **multiples** of one another
  - **Zero value has meaning**

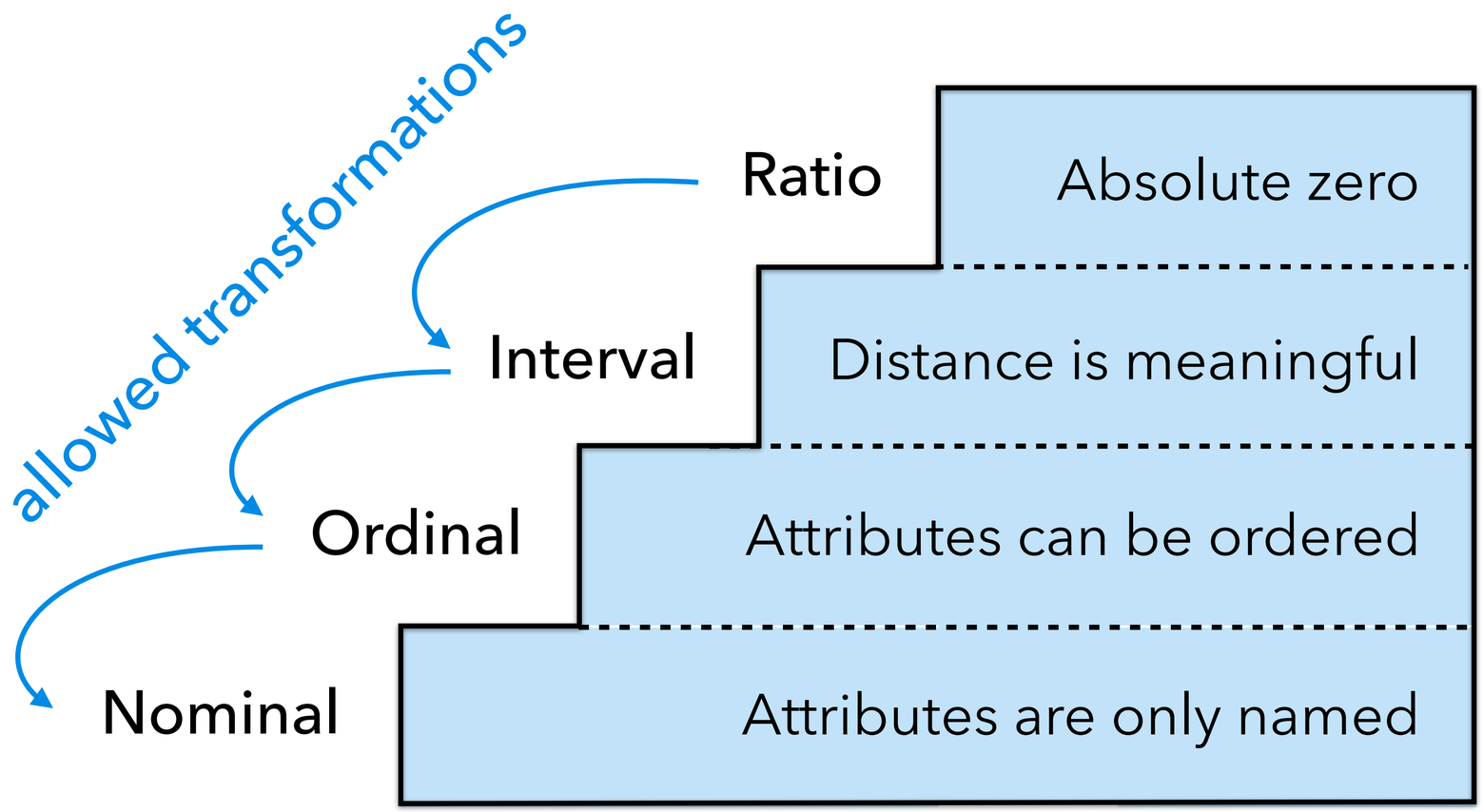
# RATIO LEVEL



- Examples
  - Time
  - Temperature in Kelvin
    - 0 means no temperature (or rather no energy in the system)
  - Weight
  - Height
  - Most count variables
    - Money in your wallet
    - Number of documents viewed
    - Number of beers you can drink before vomiting

# LEVEL TRANSFORMATIONS





# LEVELS OF MEASUREMENT



- Temperature
  - **Ratio** Kelvin
  - **Interval** Celsius
  - **Ordinal** Hot / Warm / Lukewarm / Chilly / Cold
  - **Nominal** not applicable

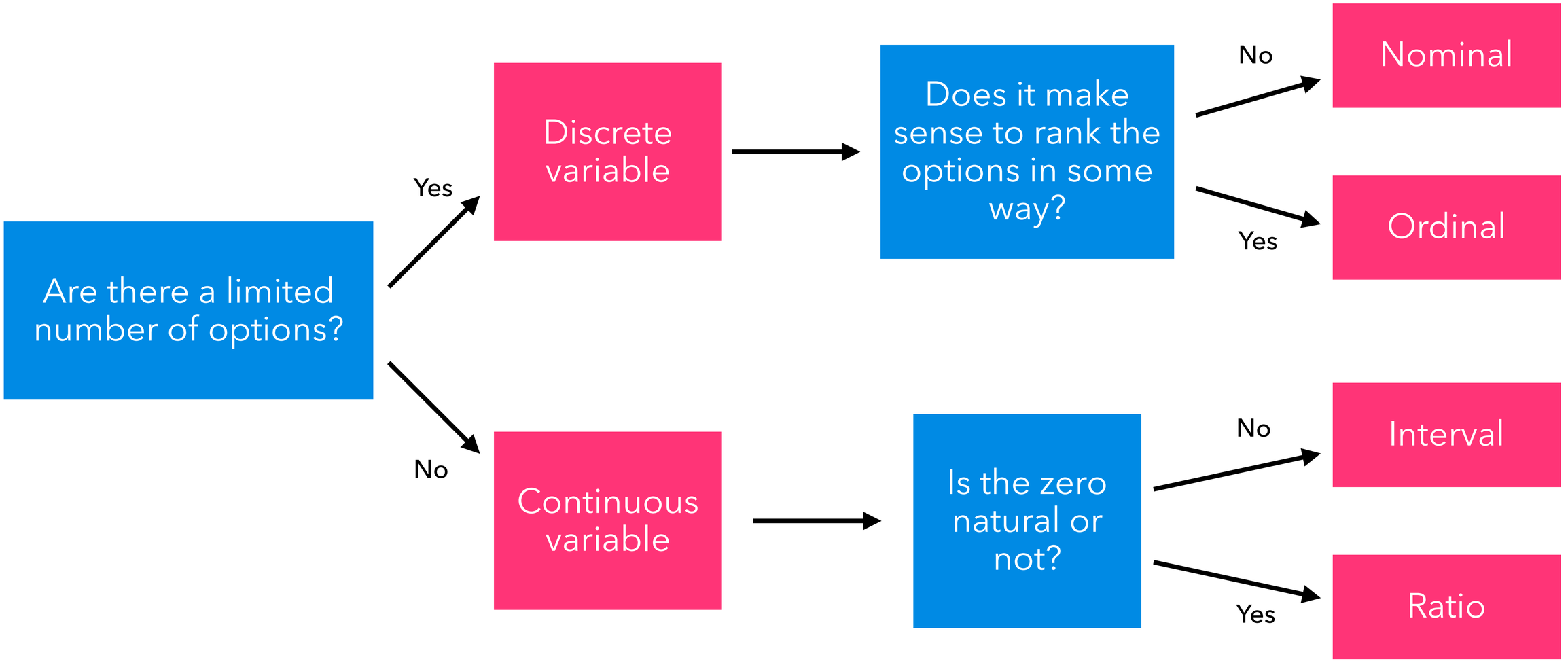
# LEVELS OF MEASUREMENT

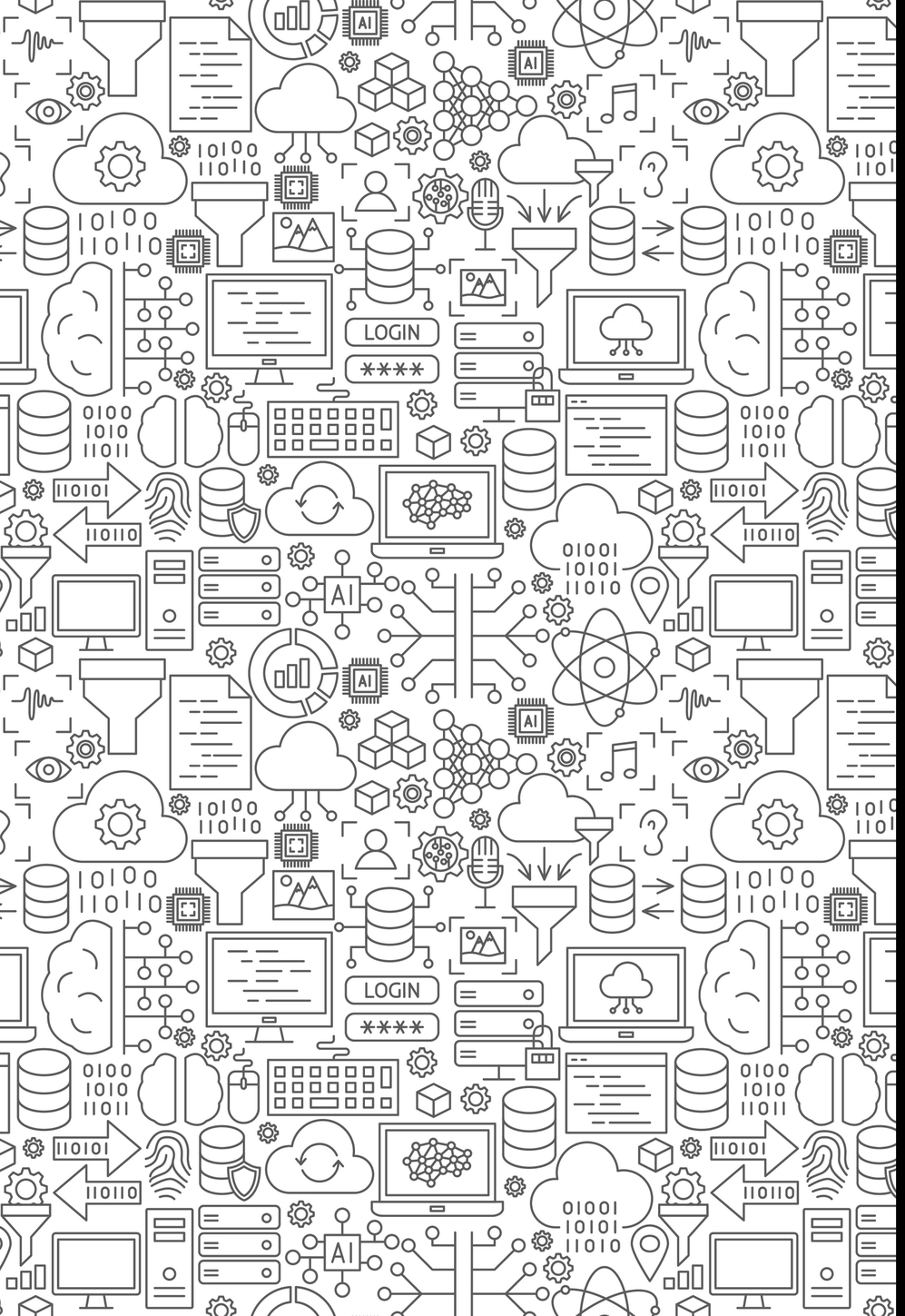


- Task completion
  - **Ratio** Seconds
  - **Interval** not applicable
  - **Ordinal** Long / Medium / Short
  - **Nominal** Complete / Incomplete

# HOW TO DECIDE WHICH LEVEL OF MEASUREMENT A VARIABLE IS








PART 3

# MEASURES OF CENTRAL TENDENCY

# CENTRAL TENDENCY



- Describe how observations (scores) **cluster around the center of the distribution**
  - Where is the center of our distribution located?
  - Replaces displaying the entire frequency table
  - Allows us to easily compare two or more different groups on the same variable
  - Three different measures
    - (Arithmetic) **Mean** (= Average)
    - **Median**
    - **Mode**

# MEAN



- Most common of the measures of central tendency
  - Also (narrowly) referred to as the **average**
  - Arithmetic mean $\bar{x}$ (or *M)* is sum of all values for variable *x* divided by the number of values *n*

$$M = \bar{x} = \frac{\sum x}{n}$$

$$\sum x = the\ summation\ of\ x$$

$$n = the\ number\ of\ cases$$

# CALCULATING THE MEAN IN R



- Create a new RStudio project
- Download `instagram-questionnaire.xlsx` from Moodle and move it to the directory belonging to that project
- Read in the dataset

- Inspect the dataset

# CALCULATING THE MEAN IN R



- You can calculate the mean in R using the `mean()` function
  - The `mean()` function only works on vectors
  - It is a good habit to always explicitly tell `mean()` to disregard `NA` values!

```r
# Mean time spent on Instagram in minutes.
mean(survey$daily_time_in_mins, na.rm = TRUE)
```

  - R is flexible, so there are multiple ways of doing the same thing

```r
select(survey, daily_time_in_mins) %>% pull() %>% mean(na.rm = TRUE)

summarize(survey, mean_minutes = mean(daily_time_in_mins, na.rm = TRUE))

survey[, "daily_time_in_mins"] %>% pull() %>% mean(na.rm = TRUE)
```

# CALCULATING THE MEAN IN R



- You can use `group_by()` to get means by gender (or age group)

```
1  survey %>%
2    group_by(gender) %>%
3    summarize(mean = mean(daily_time_in_mins, na.rm = TRUE))
```

```
survey %>%
  group_by(gender) %>%
  summarize(mean = mean(daily_time_in_mins, na.rm = TRUE))
# A tibble: 2 × 2
  gender  mean
  <chr>  <dbl>
1 Female  62.8
2 Male    37.8
```

# CALCULATING THE MEAN IN R



- This is much more efficient than using filters:

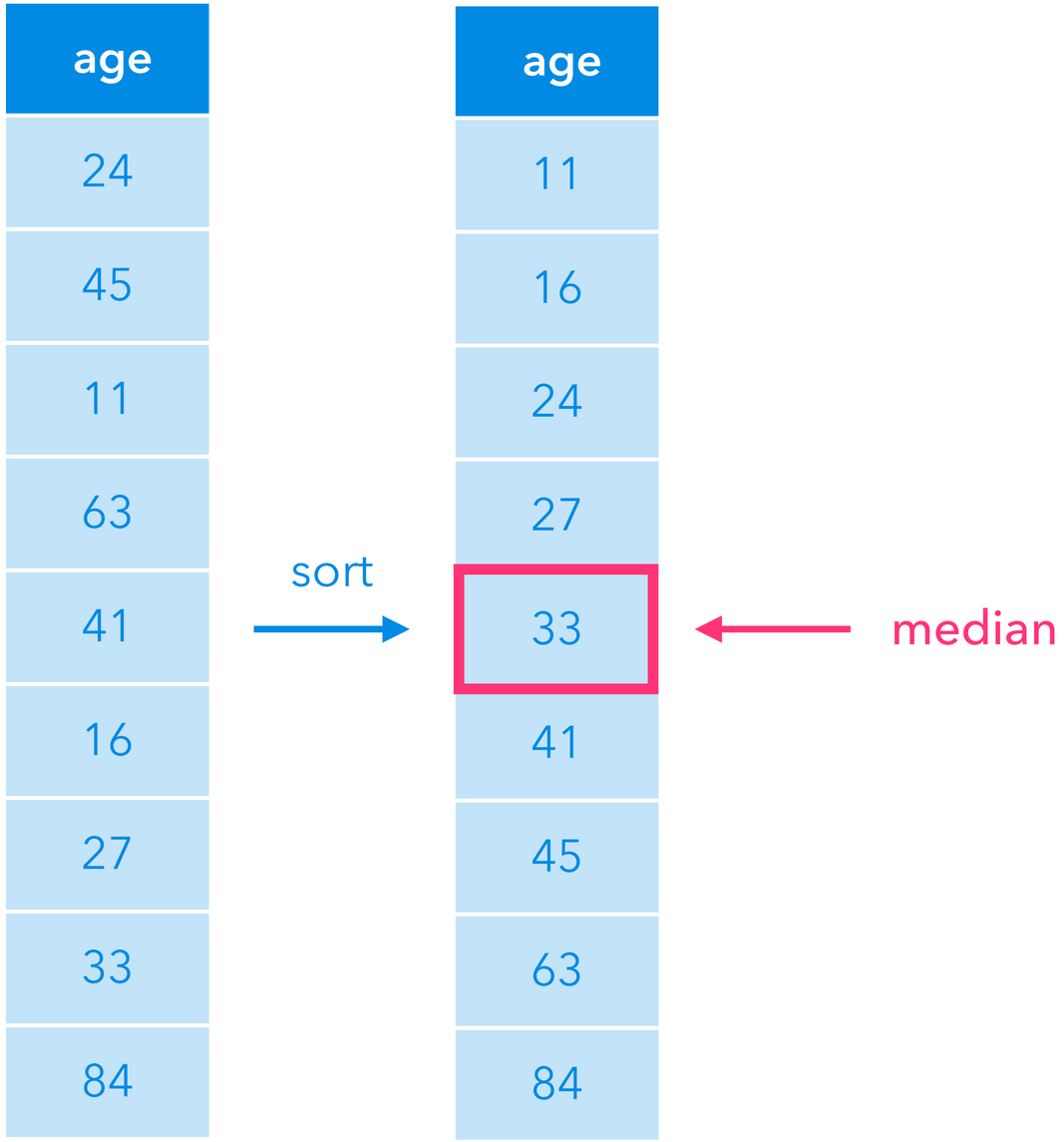
```
survey %>%
  filter(gender == “Female") %>%
  pull(daily_time_in_mins) %>% mean(na.rm = TRUE)

survey %>%
  filter(gender == "Male") %>%
  pull(daily_time_in_mins) %>% mean(na.rm = TRUE)
```

# MEDIAN



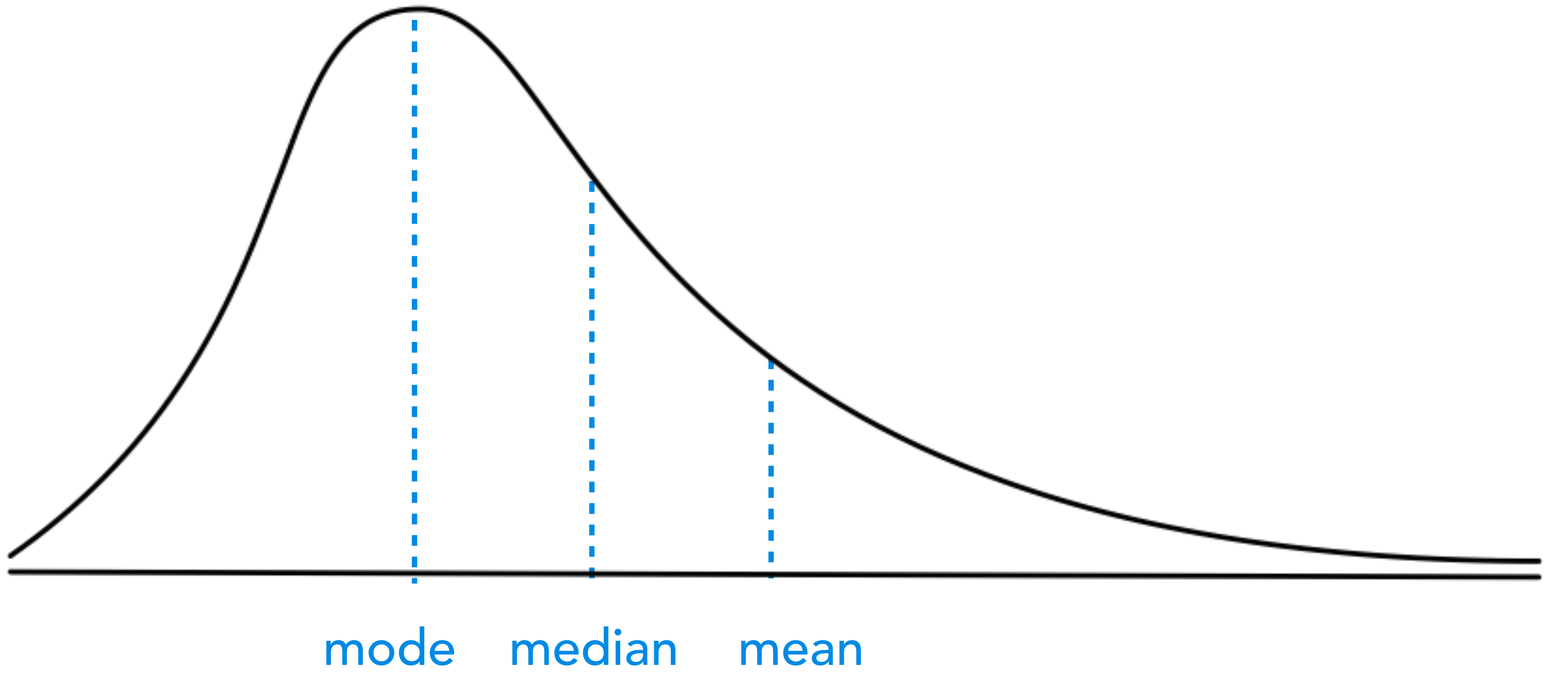
- Middle score in a **sorted** list of scores
  - How to calculate?
    - Sort all values of *x* in ascending (or descending order)
    - If odd no. of values, take the middle value
    - If even no. of values, take the average of the two middle values



# MEDIAN



- Why is the median relevant?
  - If your data distribution is **skewed** or if you have some **extreme outliers**, the median is less affected by this than the mean



# CALCULATING THE MEDIAN IN R



- You can calculate the median in R using the `median()` function

```
survey %>%
  group_by(gender) %>%
  summarize(median = median(daily_time_in_mins, na.rm = TRUE))
```

```
survey %>%
  group_by(gender) %>%
  summarize(median = median(daily_time_in_mins, na.rm = TRUE)
# A tibble: 2 × 2
  gender median
  <chr>   <dbl>
1 Female     51
2 Male       30
```

# MODE



- Simply put: the **most frequent** value of x
- More accurately
  - Value of a variable that contains more cases than in either value adjacent to it
    - So the value of the peak in the distribution
  - Means that there can be **more than one mode**!
    - This is the case in so-called **bimodal** or **multimodal** distributions
  - Means that values should remain sorted

# CALCULATING THE MODE IN R



- There is no built-in function for calculating the statistical mode in R
  - The following custom function calculates the (unimodal) mode for you (the `mode()` function already exists in R but does something else)

```
calculate_mode <- function(v) {
   uniqv <- unique(v)
   uniqv[which.max(tabulate(match(v, uniqv)))]
}

survey %>%
  group_by(gender) %>%
  summarize(mode = calculate_mode(daily_time_in_mins))
```

  - As you can see we can use our own functions in `summarize()` as well (without the `NA` removal)

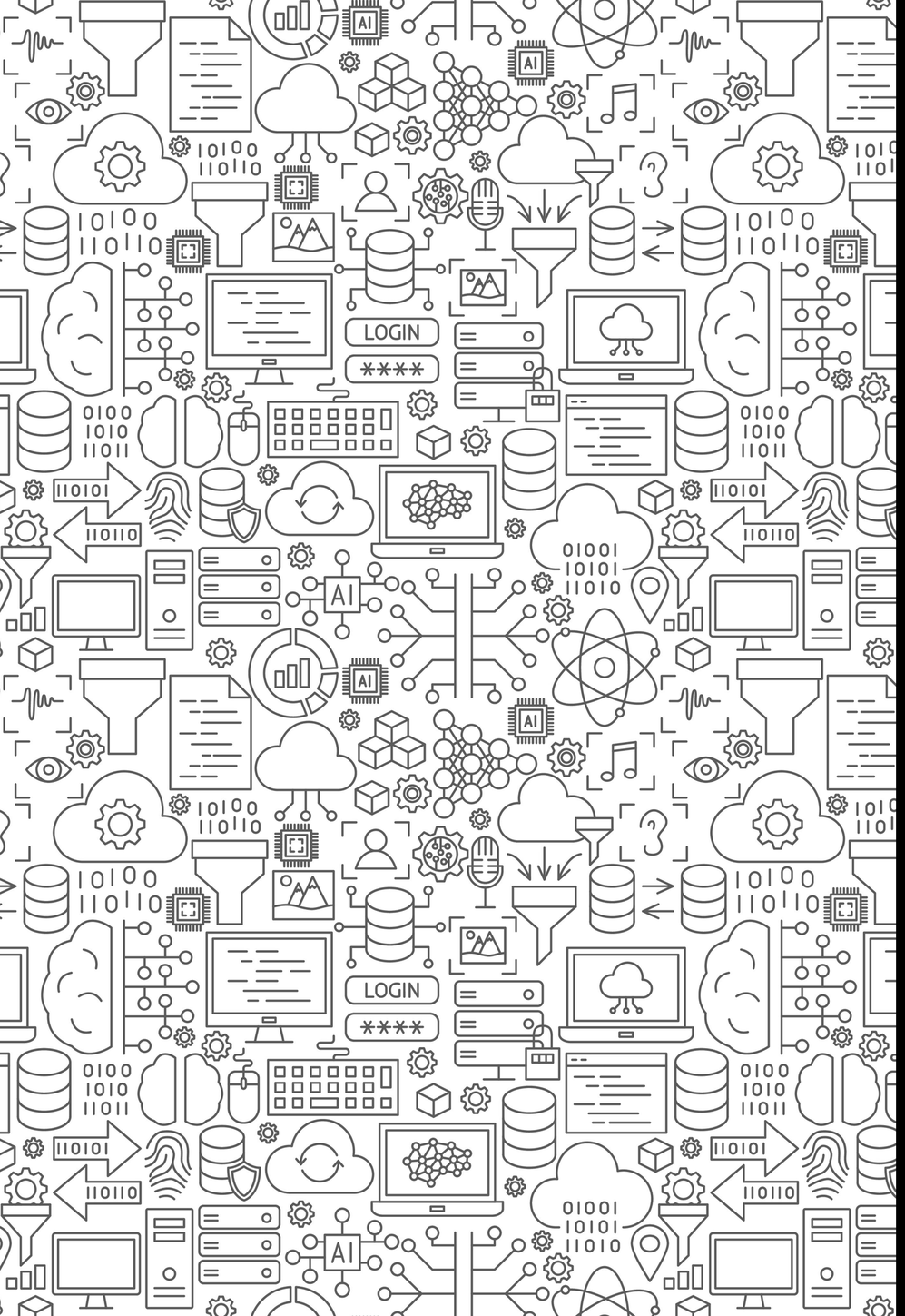
# CALCULATING THE MEDIAN IN R



- And remember we can combine as many of them as we want

```r
survey %>%
  group_by(gender) %>%
  summarize(mean = mean(daily_time_in_mins, na.rm = TRUE),
            median = median(daily_time_in_mins, na.rm = TRUE),
            mode = calculate_mode(daily_time_in_mins),
            max = max(daily_time_in_mins, na.rm = TRUE),
            min = min(daily_time_in_mins, na.rm = TRUE))
```

```
# A tibble: 2 × 6
  gender  mean median  mode   max   min
  <chr>  <dbl>  <dbl> <dbl> <dbl> <dbl>
1 Female  62.8     51    60   420     0
2 Male    37.8     30    60   105     1
```




PART 4

# MEASURES OF DISPERSION

# MEASURES OF DISPERSION

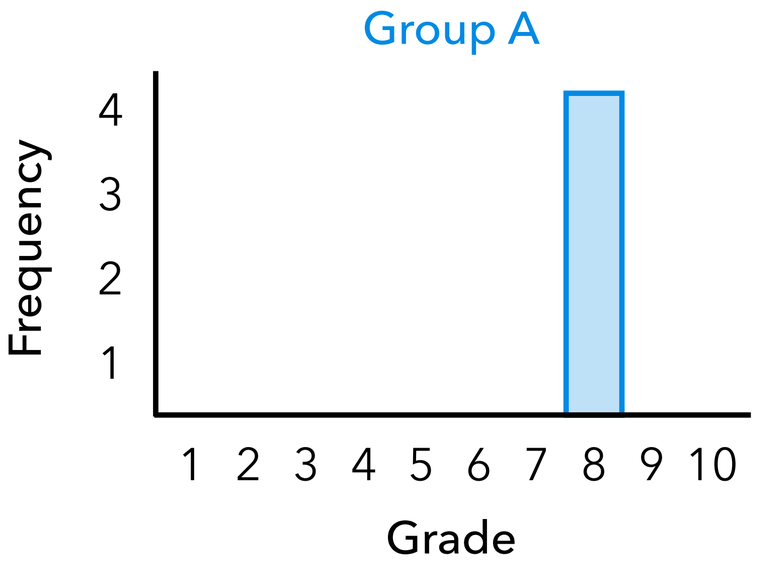


- Describe the **degree of clustering around the mean**
  - Are most scores relatively close to the mean?
  - Are they scattered over a wider interval?
  - Also known as **measures of variability**
    - Range
    - Mean deviation
    - Variance
    - Standard deviation

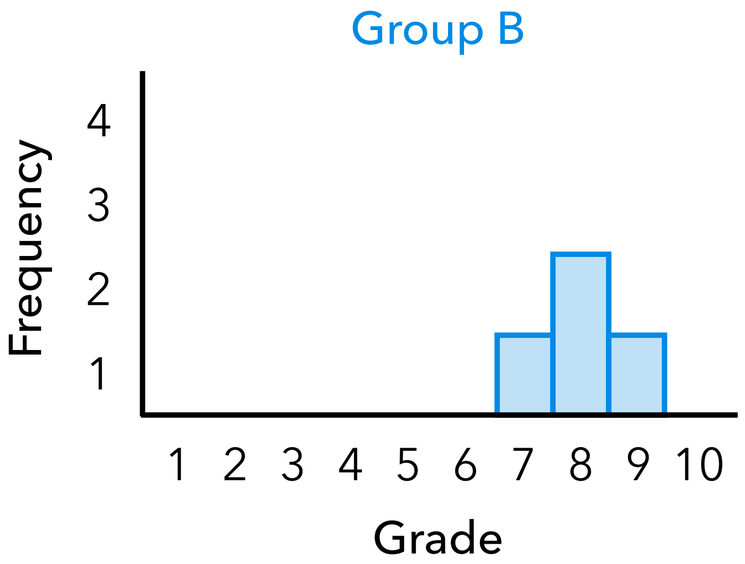
# MEASURES OF DISPERSION



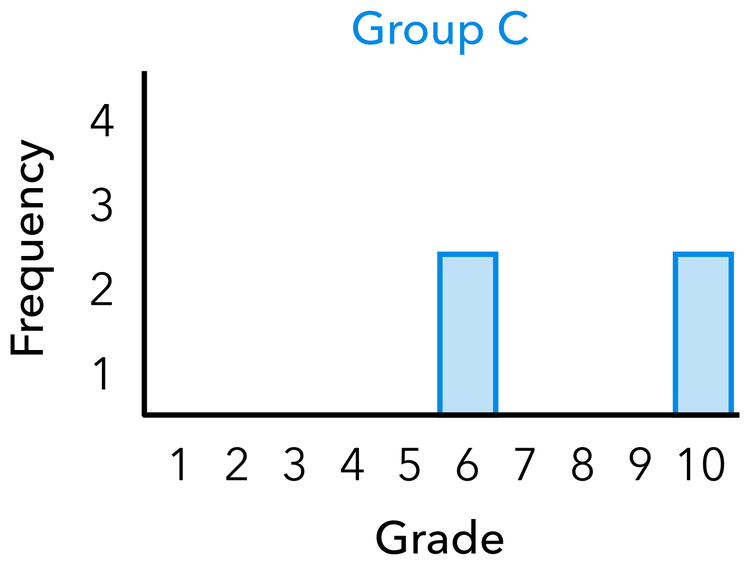
- Why are these necessary?
  - Central tendency measures do not always provide a complete picture




mean = 8




mean = 8



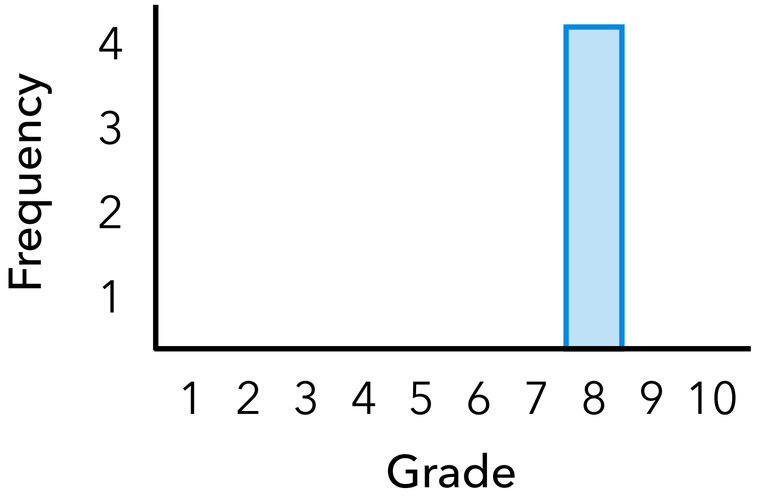

mean = 8

# RANGE



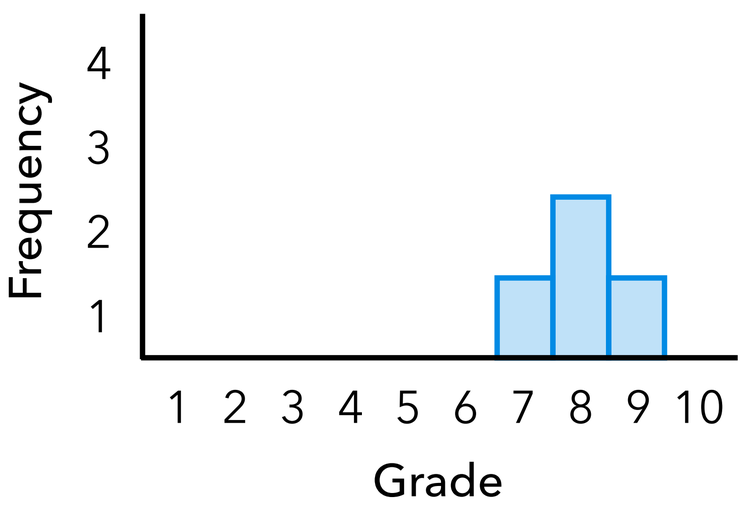
- Difference between the highest and lowest scores
  - Simplest measure of dispersion
  - Problem is that only two scores have an impact and that **outliers** destroy the measure's usefulness

Group A



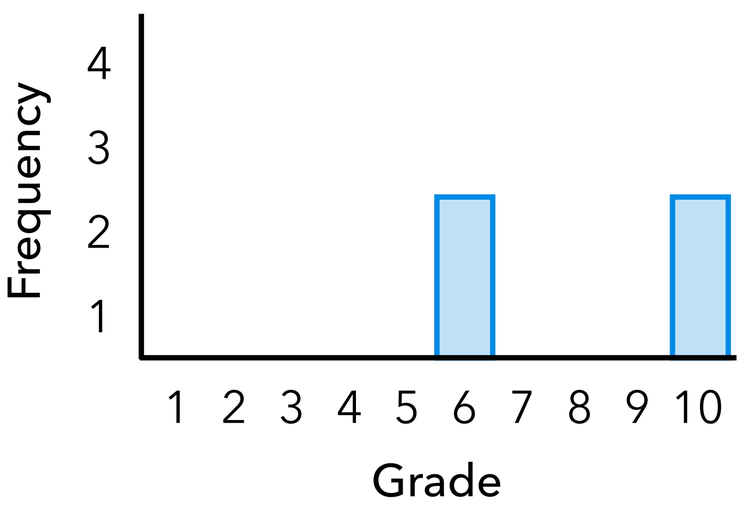

range = 8 – 8 = 0

Group B




range = 9 – 7 = 2

Group C




range = 10 – 6 = 4

# CALCULATING THE RANGE IN R



- We can calculate the range in R using the built-in `range()` function
  - However, this shows you the max and min values, which is not exactly what we want

```
range(survey$daily_time_in_mins, na.rm = TRUE)
```

```
> range(merged_data$product_views, na.rm = TRUE)
[1]  1 13
```

  - We would still need to subtract these two values from each other to get the true range, so it is more efficient to just use the `max()` and `min()` functions in `dplyr`

```
survey %>%
  group_by(gender) %>%
  summarize(range = max(daily_time_in_mins, na.rm = TRUE) -
min(daily_time_in_mins, na.rm = TRUE))
```

# CALCULATING THE RANGE IN R



```
> survey %>%
+   group_by(gender) %>%
+   summarize(range = max(daily_time_in_mins, na.rm = TRUE) - min(daily_time_in_mins, na.rm = TRUE))
# A tibble: 2 × 2
  gender range
  <chr>  <dbl>
1 Female   420
2 Male     104
```

# MEAN DEVIATION



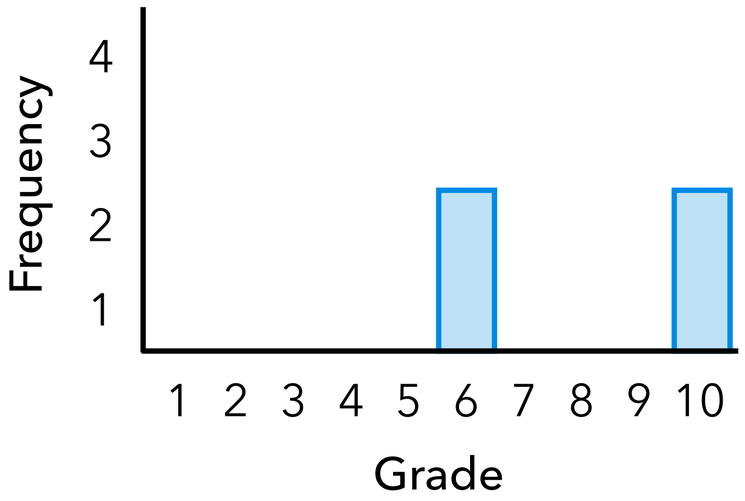
- Represents the **average deviation** of all scores from the mean
  - For each score, calculate the difference with the average
  - Calculate the mean of all absolute differences

$$mean\ deviation = \frac{\sum |x - \bar{x}|}{n}$$

# MEAN DEVIATION



Group C




range = 10 – 6 = 4

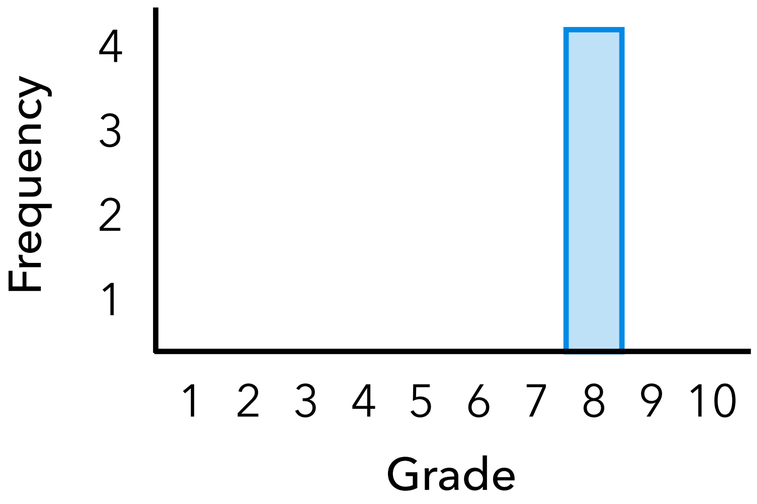
| x | $\bar{x}$ | x - $\bar{x}$ | \|x - $\bar{x}$\| |
|---|---|---|---|
| 10 | 8 | 2 | 2 |
| 10 | 8 | 2 | 2 |
| 6 | 8 | -2 | 2 |
| 6 | 8 | -2 | 2 |

$$mean\ deviation = \frac{\sum |x - \bar{x}|}{n} = \frac{2+2+2+2}{4} = \frac{8}{4} = 2.0$$

# MEAN DEVIATION

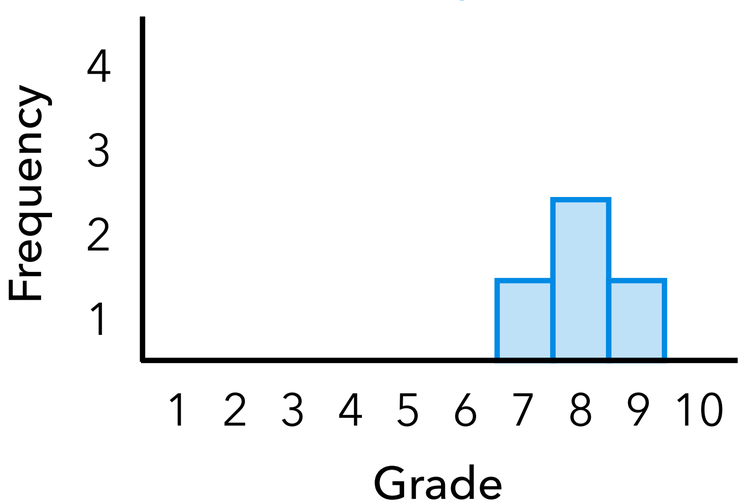


Group A



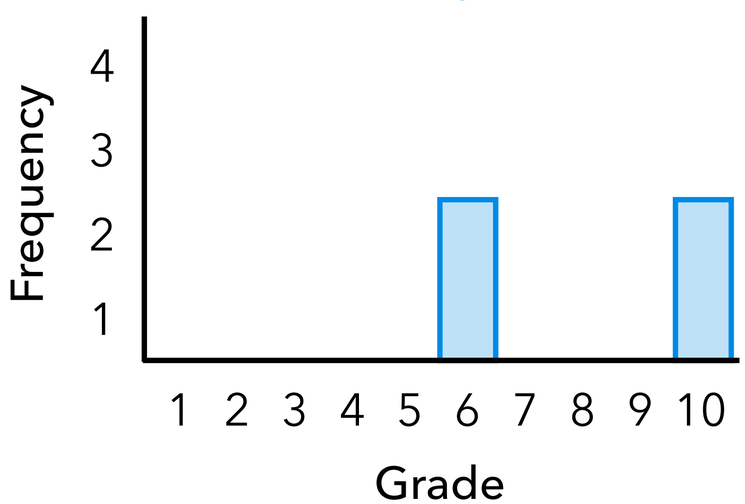

mean deviation$_A$ = 0

Group B




mean deviation$_B$ = 0.5

Group C




mean deviation$_C$ = 2.0

# VARIANCE



- Represents the **average squared deviation** of all scores from the mean
  - For each score, calculate the difference with the average
  - Calculate the mean of the squared differences

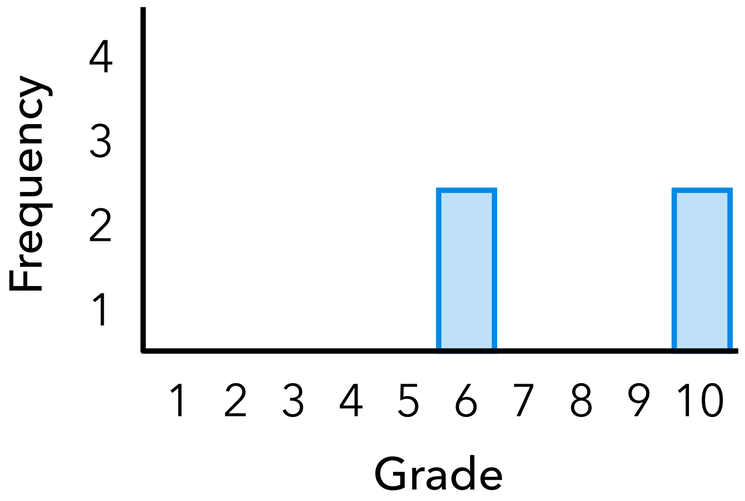
$$s^2 = \frac{\sum(x - \bar{x})^2}{n}$$

  - More sensitive to big differences (so usually larger than mean deviation)

# VARIANCE



Group C




range = 10 – 6 = 4

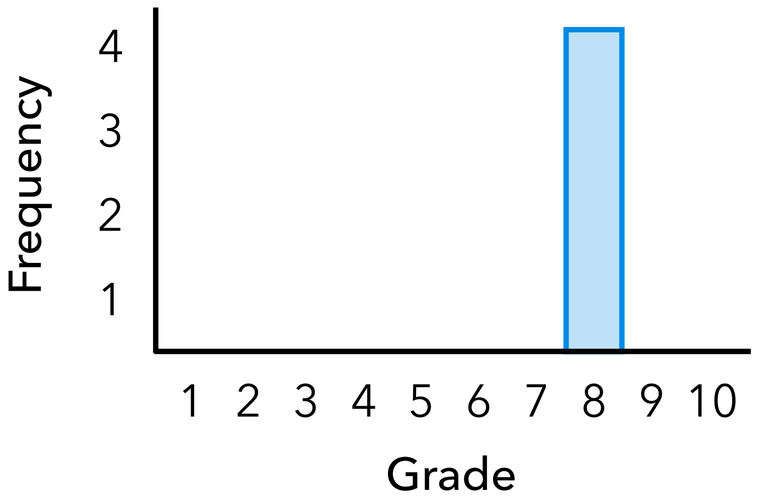
| x | $\bar{x}$ | $x - \bar{x}$ | $(x - \bar{x})^2$ |
|---|---|---|---|
| 10 | 8 | 2 | 4 |
| 10 | 8 | 2 | 4 |
| 6 | 8 | -2 | 4 |
| 6 | 8 | -2 | 4 |

$$s^2 = \frac{\sum(x - \bar{x})^2}{n} = \frac{4+4+4+4}{4} = \frac{16}{4} = 4$$

# VARIANCE

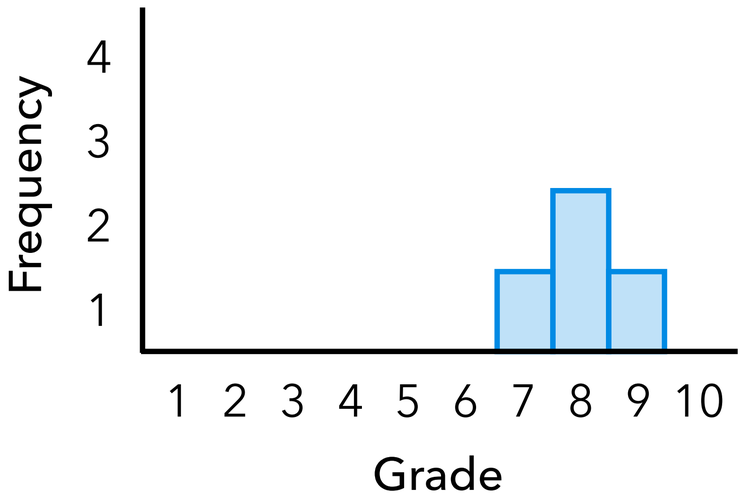


Group A



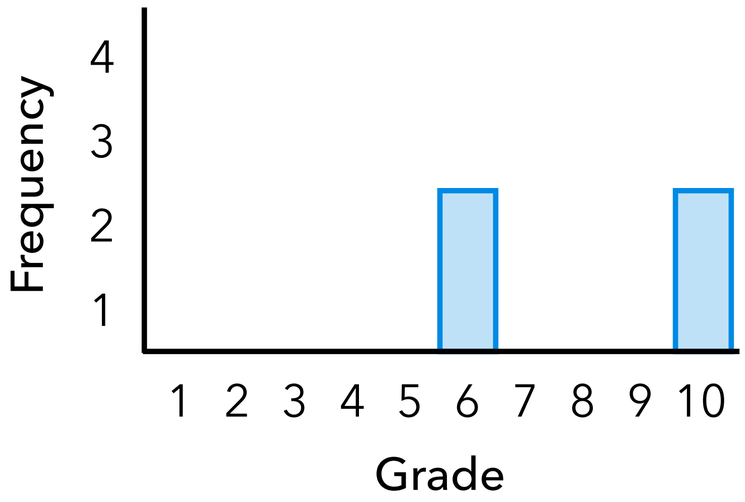

$s^2_A = 0$

Group B




$s^2_B = 0.5$

Group C




$s^2_C = 4.0$

# STANDARD DEVIATION



- Problem with variance
  - Not in the same unit of measurement as scores
- Solution
  - Take the square root of the **variance** ($s^2$)
  - **Standard deviation** (**s**)

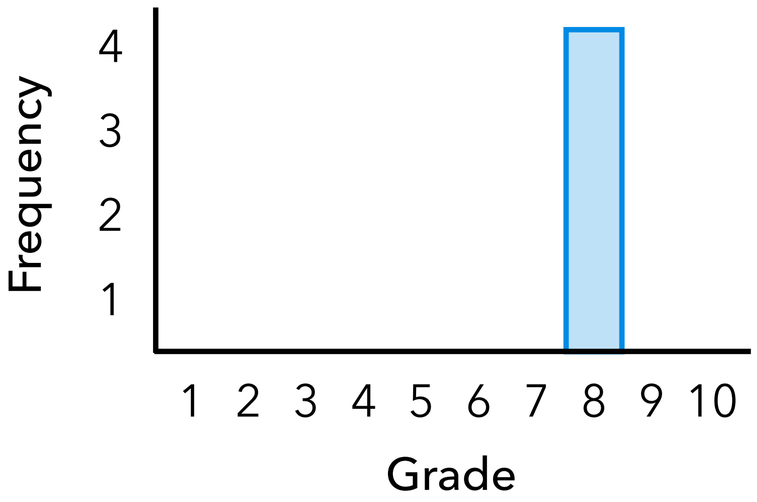
$$s = \sqrt{\frac{\sum(x - \bar{x})^2}{n}}$$

  - Is in the **same unit of measurement** as the variable scores themselves!

# STANDARD DEVIATION

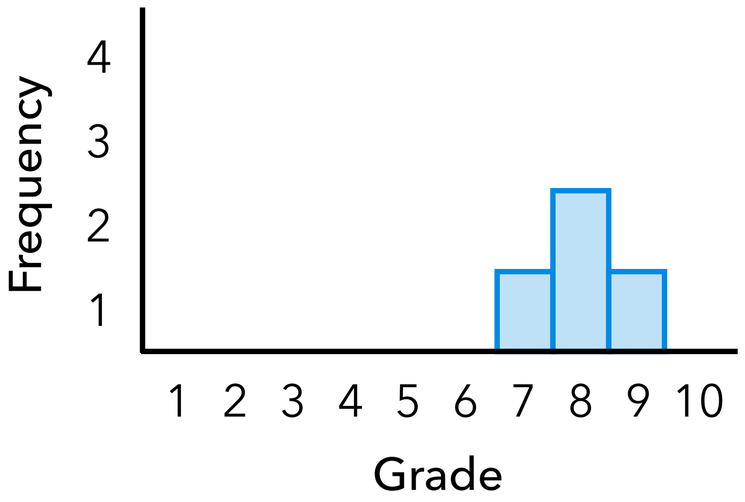


Group A



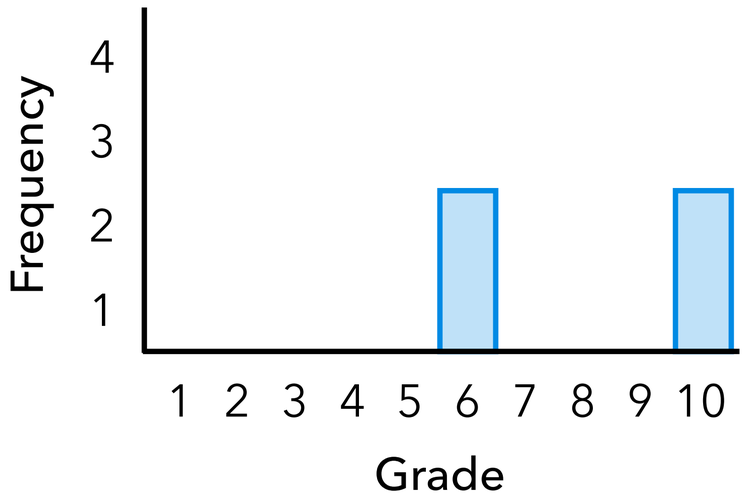

$s_A = 0$

Group B




$s_B = 0.707$

Group C

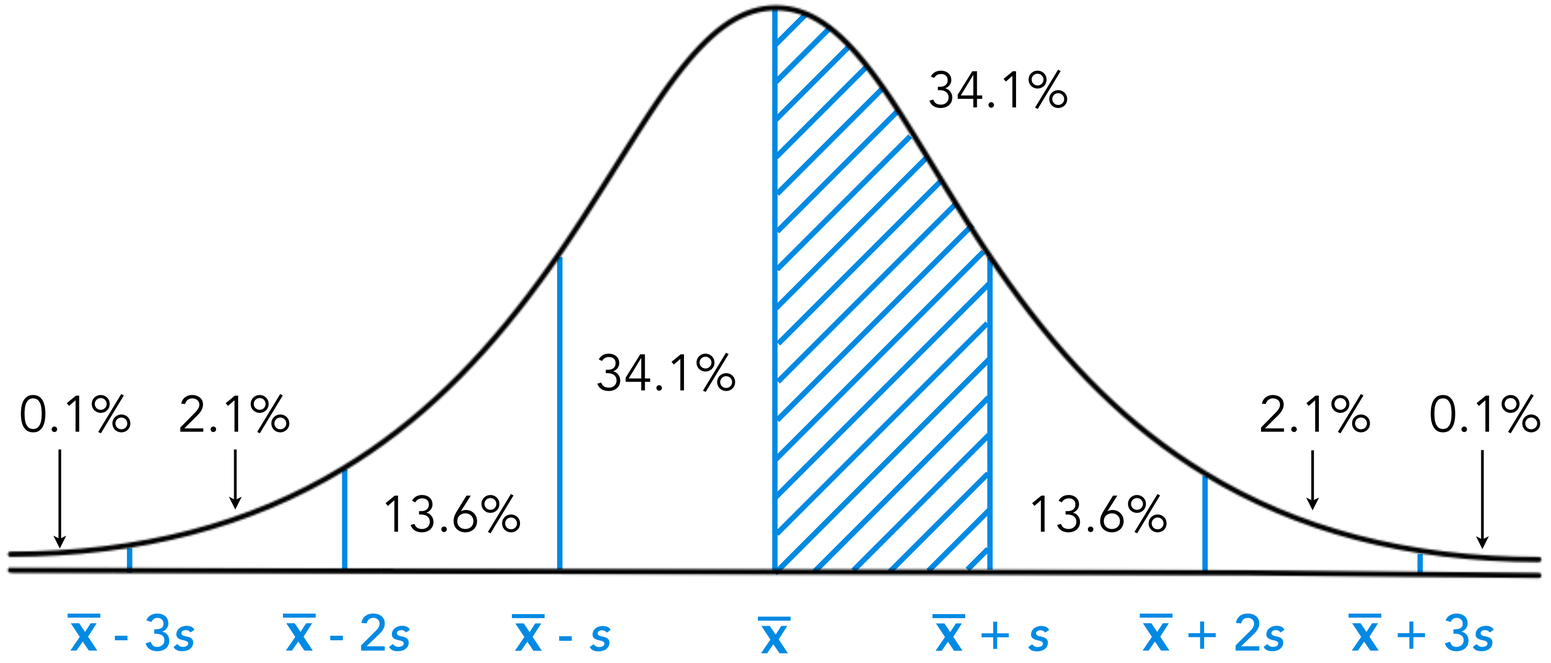



$s_C = 2.0$

# NORMAL DISTRIBUTION



- The standard deviation is especially useful, because we can use it to calculate the probability of observing a particular value
  - For example, if the data is normally distributed



# CALCULATING THE STANDARD DEVIATION IN R



- We can calculate the standard deviation in R using the built-in `sd()` function

```
survey %>%
  group_by(gender) %>%
  summarize(range = max(daily_time_in_mins, na.rm = TRUE) -
min(daily_time_in_mins, na.rm = TRUE),
            stdev = sd(daily_time_in_mins, na.rm = TRUE))
```

```
# A tibble: 2 × 3
  gender range stdev
  <chr>  <dbl> <dbl>
1 Female   420  57.8
2 Male     104  25.7
```

# LEVELS OF MEASUREMENT

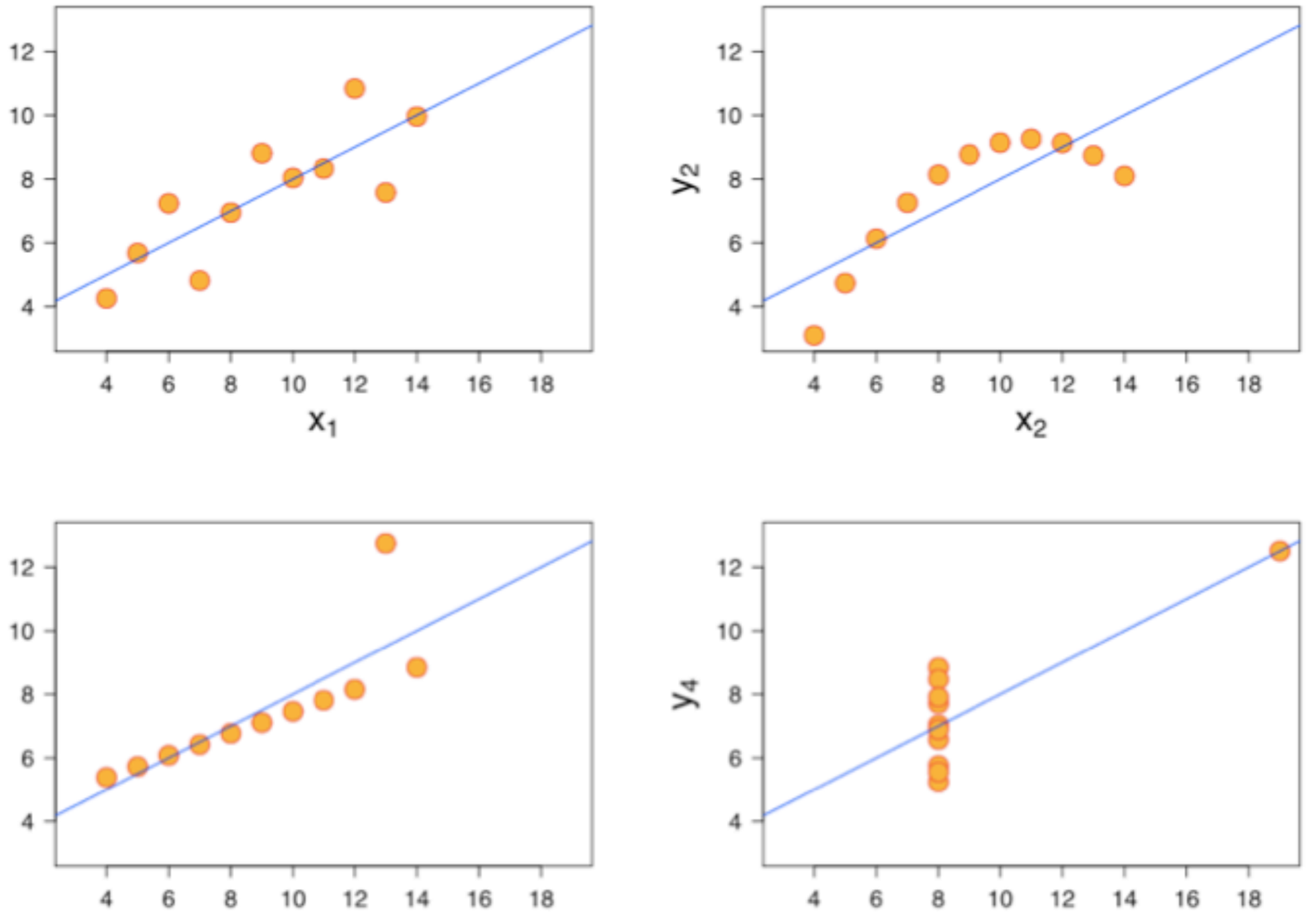


| Level of measurement | Measures of central tendency | | | Measures of dispersion | | |
|---|---|---|---|---|---|---|
| | Mode | Median | Mean | Range | Variance | Standard deviation |
| Nominal | ✓ | | | | | |
| Ordinal | ✓ | ✓ | ? | ✓ | | |
| Interval/Ratio | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |

# WHY DATA VISUALIZATION?



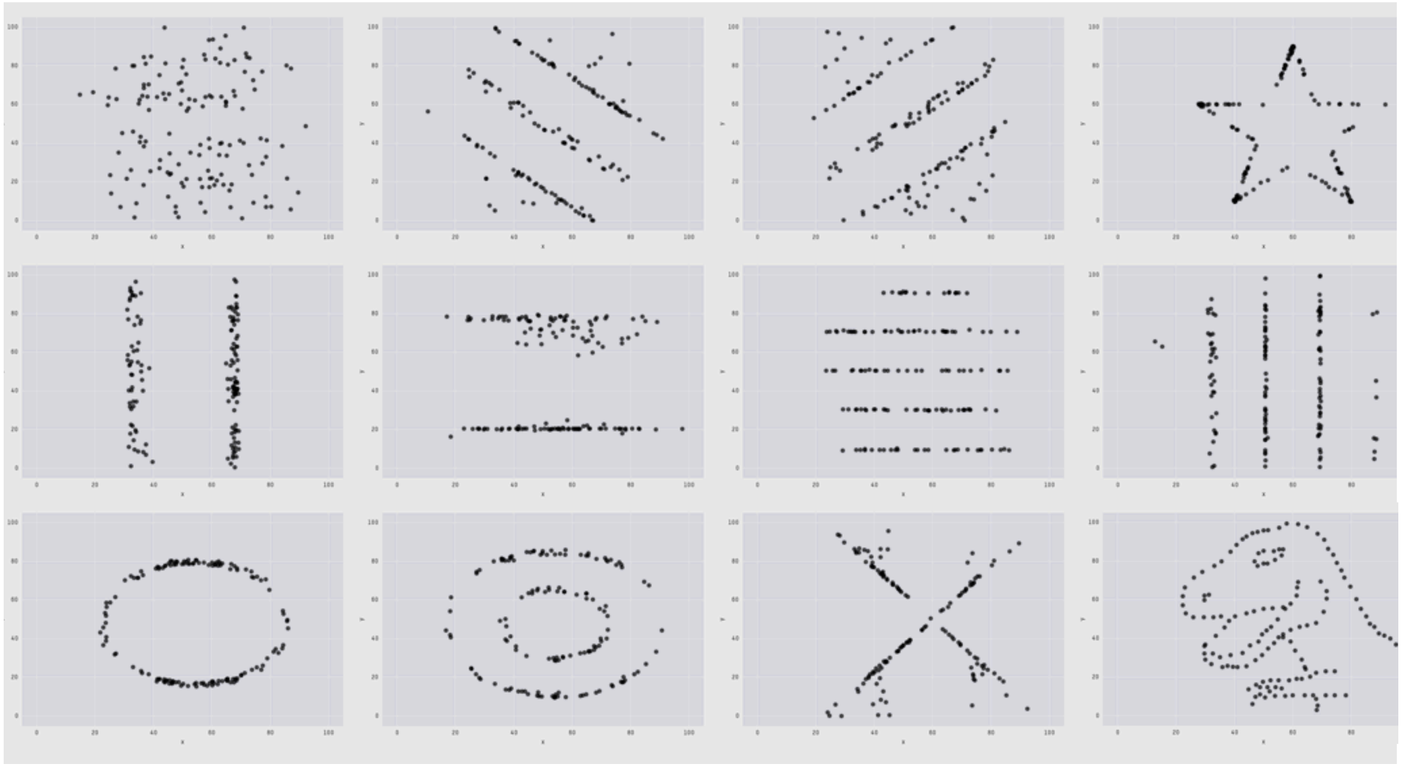
- Because numbers alone do not tell the whole story!
  - Example: Anscombe's quartet is made up of four data sets that all have the **same mean, standard deviation, and correlation**

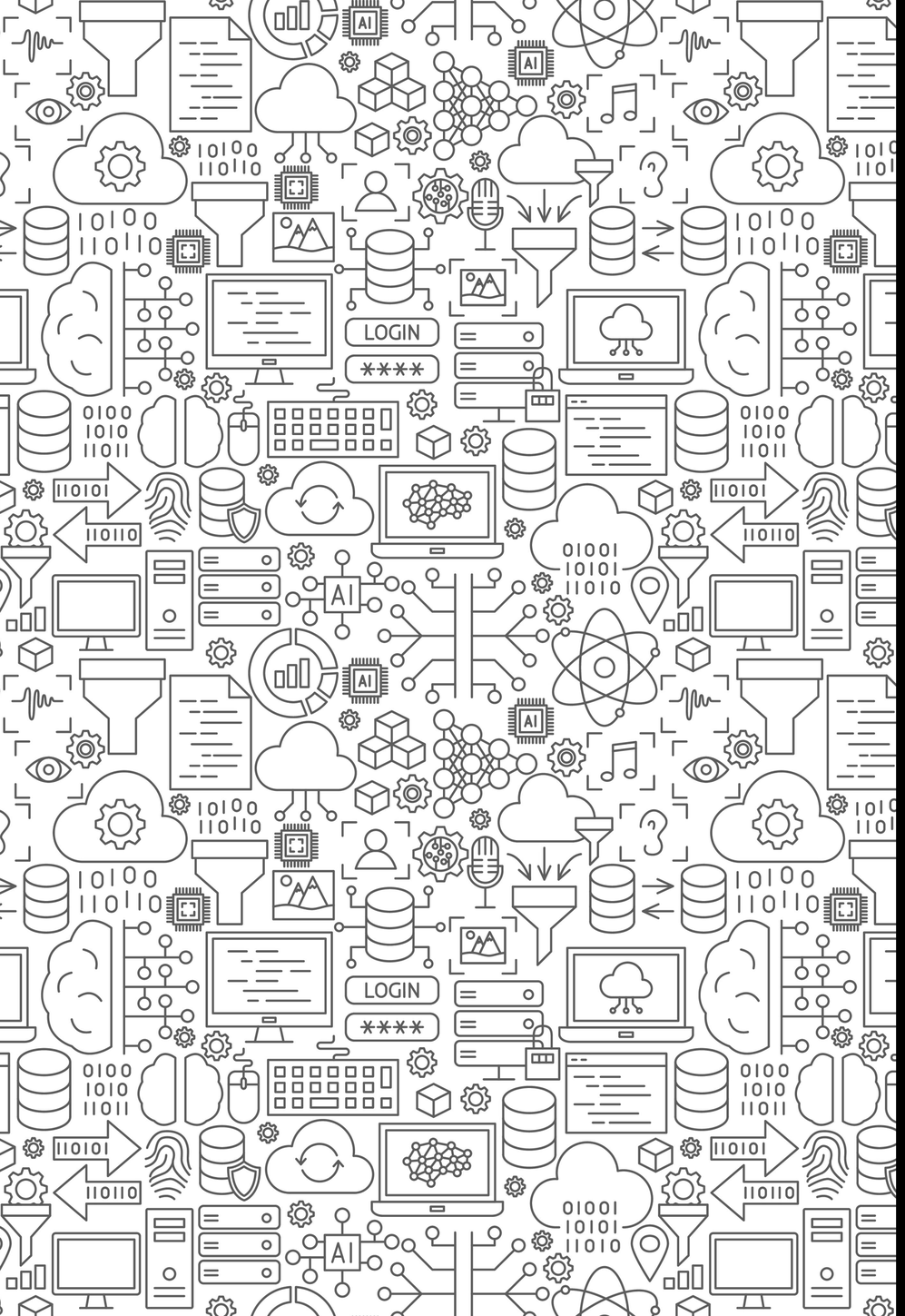


# SO WHY DO WE NEED TO VISUALIZE OUR DATA?



- Because numbers alone do not tell the whole story!
  - Example: Anscombe's quartet is made up of four data sets that all have the **same mean, standard deviation, and correlation**
  - Example: **Datasaurus Dozen** is made up of twelve such data sets (Matejka & Fitzmaurice, 2017)
  - Rule of thumb: make both calculations (**summary statistics**) and **visualizations**






PART 5

# DATA VISUALIZATION

# CHOOSING AN APPROPRIATE VISUALIZATION



- We have two types of visualizations at our disposal
  - **Non-graphical visualizations** (a.k.a. tables)
  - **Graphical visualizations** (a.k.a. figures)

- When should we use what?
  - **Measures** → If you can represent the data in question in 1-2 sentences
  - **Tables** → If you want to show the make-up of your data
  - **Figures** → When you want to communicate a particular trend/difference in your data

# FREQUENCY TABLE



- A **frequency table** counts the absolute/relative **frequencies** of a single **discrete** variable

**Think of self as liberal or conservative**

| | Frequency | Percent |
|---|---|---|
| Extremely liberal | 53 | 3.8 |
| Liberal | 149 | 10.6 |
| Slightly liberal | 153 | 10.9 |
| Moderate | 598 | 42.4 |
| Slightly conservative | 199 | 14.1 |
| Conservative | 217 | 15.4 |
| Extremely conservative | 40 | 2.8 |
| **Total** | **1409** | **100.0** |

# CONTINGENCY TABLE / CROSSTAB



- A **contingency table** tabulates **frequencies** of **two** discrete variables
  - Can easily be created using Tableau or using pivot tables in Excel

| | | **Sex before marrage** | | | | |
|---|---|---|---|---|---|---|
| | | Always wrong | Almost always wrong | Sometimes wrong | Not wrong at all | **Total** |
| **Political party affiliation** | Democrat | 70 | 32 | 75 | 134 | **311** |
| | Independent | 113 | 35 | 86 | 190 | **424** |
| | Republican | 45 | 15 | 25 | 28 | **113** |
| | **Total** | **228** | **82** | **186** | **352** | **848** |

# CHOOSING AN APPROPRIATE VISUALIZATION

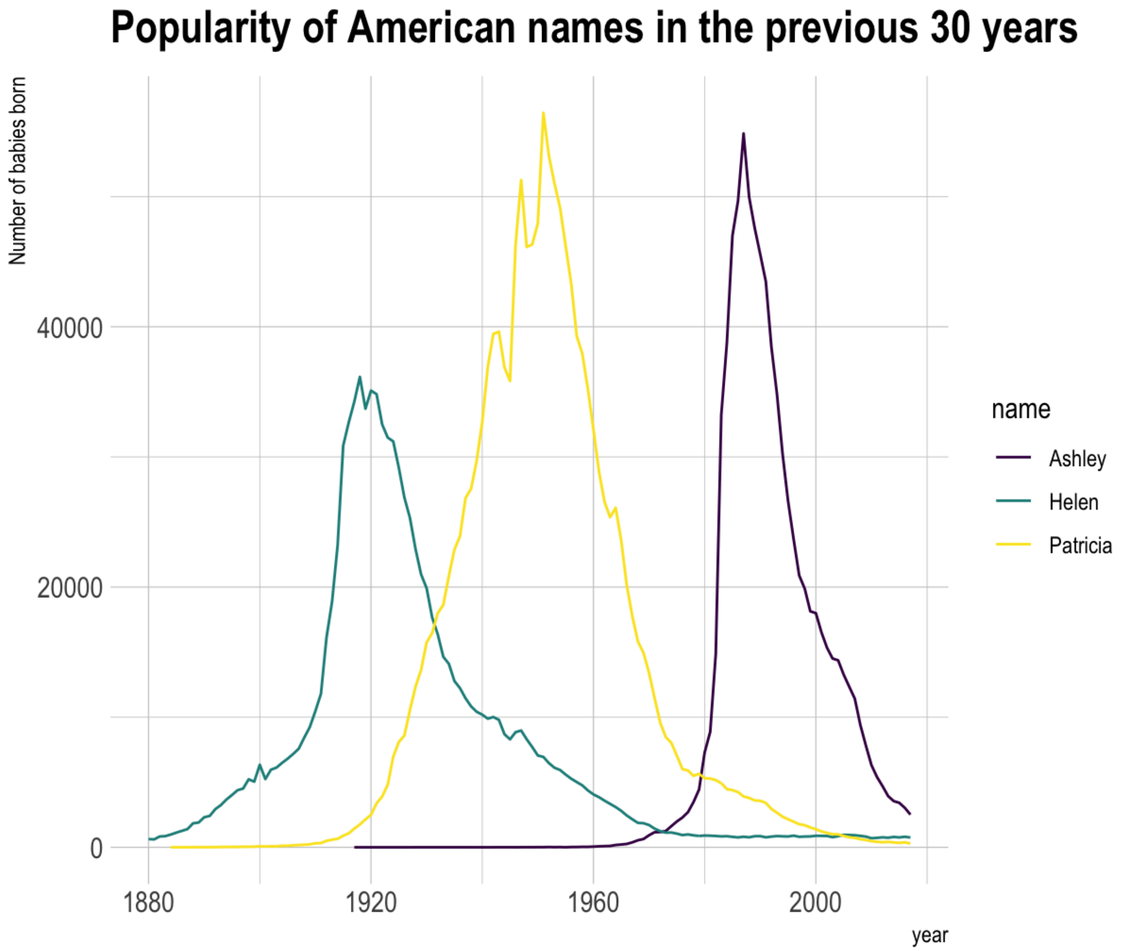


- Which visualization is best depends on **what we are trying to show** (Smith et al., 2019)
  - **Change over time** → Give emphasis to changing trends
  - **Magnitude** → Show size comparisons
  - **Part-to-whole** → Show how an entity can be broken down into its components
  - **Ranking** → Show size comparisons where the ordering is important
  - **Distribution** → Show values in a dataset and how often they occur
  - **Correlation** → Show the relationship between two or more variables
  - **Spatial** → Show precise locations or geographical patterns in data
  - **Flow** → Show the volumes or intensity of movement between two or more conditions

# CHANGE OVER TIME



- When?
  - If you want to emphasize changing trends for one or more **continuous** variables
- Types of visualizations
  -  **Line chart**
    - Shows information as a series of data points with straight lines connecting the points



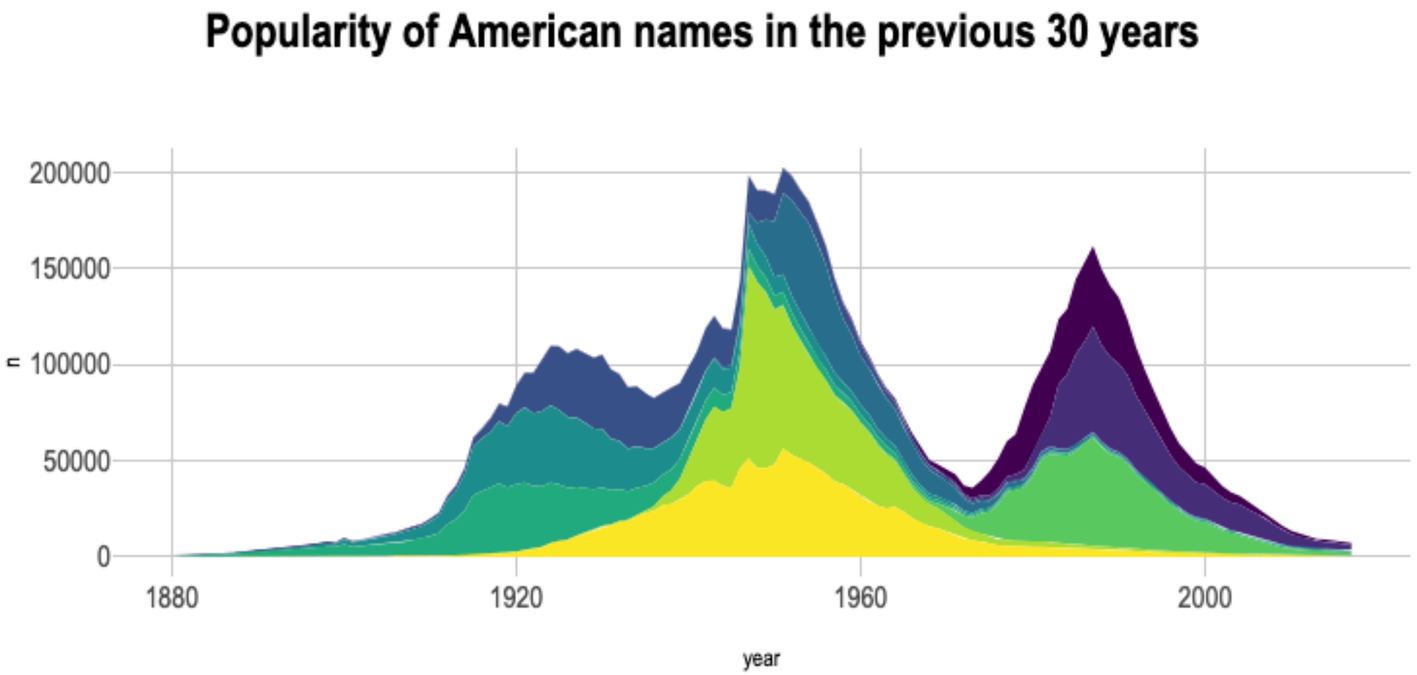

https://www.r-graph-gallery.com/line-chart-several-groups-ggplot2.html

# CHANGE OVER TIME



## Stacked area chart

- Displays the evolution of a numeric variable for several groups
- Good for showing changes to total, but seeing change in components can be difficult

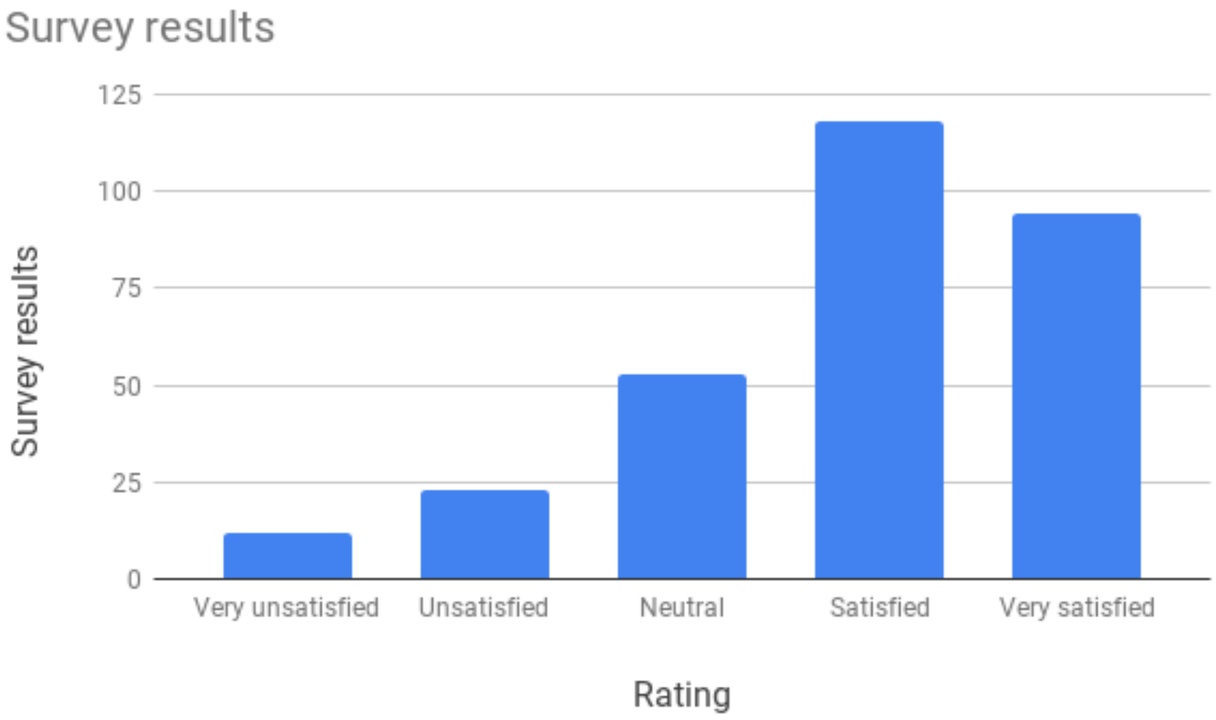



https://www.r-graph-gallery.com/stacked-area-graph.html

# MAGNITUDE



- When?
  - If you want to show **size comparisons** of **discrete** data (**nominal/ordinal**)
- Types of visualizations
  -  **Bar chart**
    - Shows information as rectangular bars with lengths proportional to the values that they represent
    - Bars can be vertical bars (a.k.a. column charts in that case) or horizontal bars (if you have long labels)

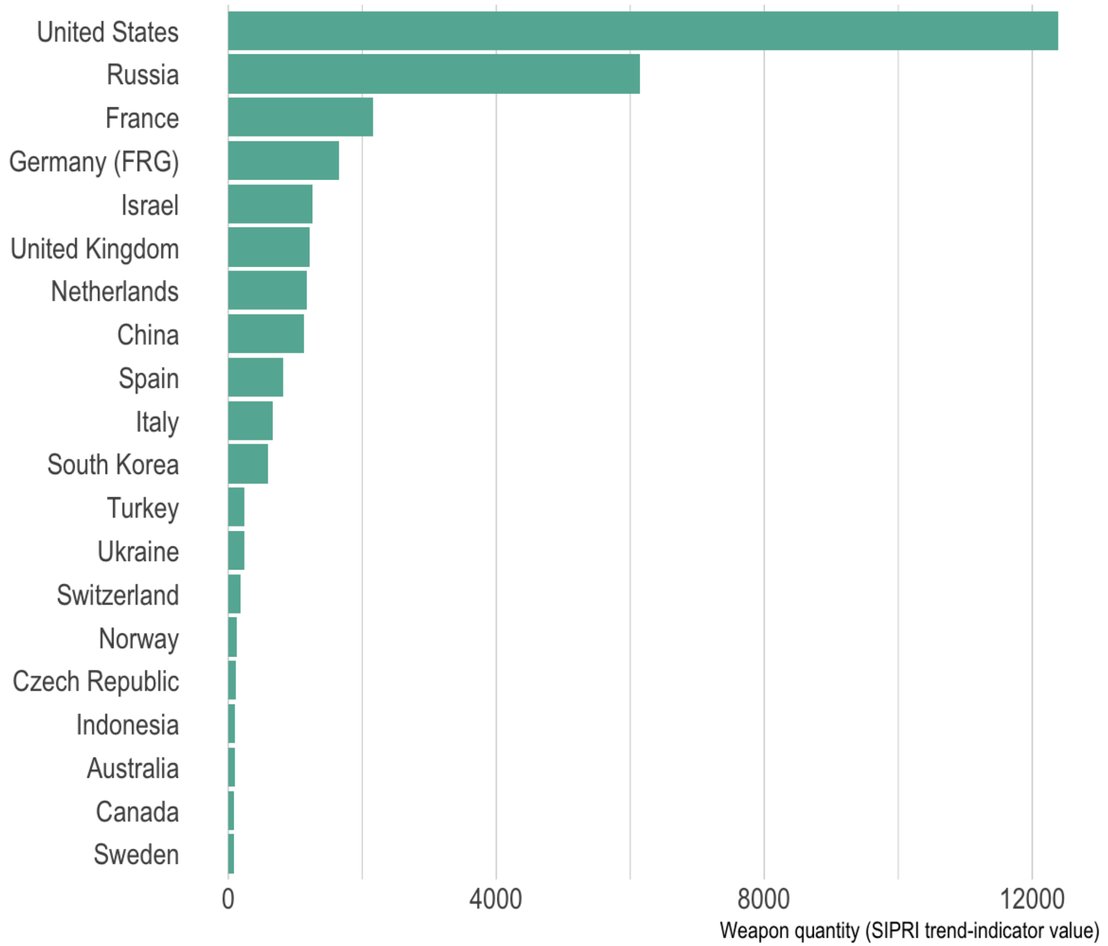



https://www.r-graph-gallery.com/barplot.html

# MAGNITUDE



- When?
  - If you want to show **size comparisons** of **discrete** data (**nominal/ordinal**)
- Types of visualizations
  - **Bar chart**
    - Shows information as rectangular bars with lengths proportional to the values that they represent
    - Bars can be vertical bars (a.k.a. column charts in that case) or horizontal bars (if you have long labels)



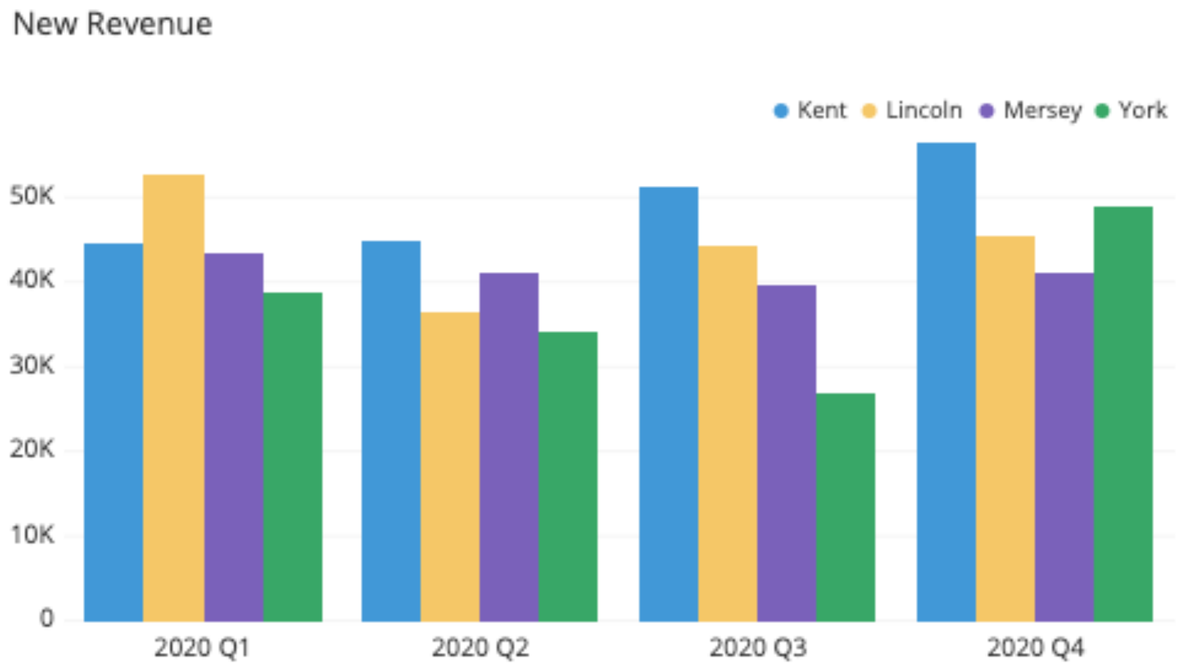

https://www.r-graph-gallery.com/barplot.html

# MAGNITUDE



## – Paired bar chart

- Paired bar chars (a.k.a. **grouped bar chart**) shows **two (or more) discrete variables at once**
- Be careful about adding too many variables; more than two is usually a bad idea



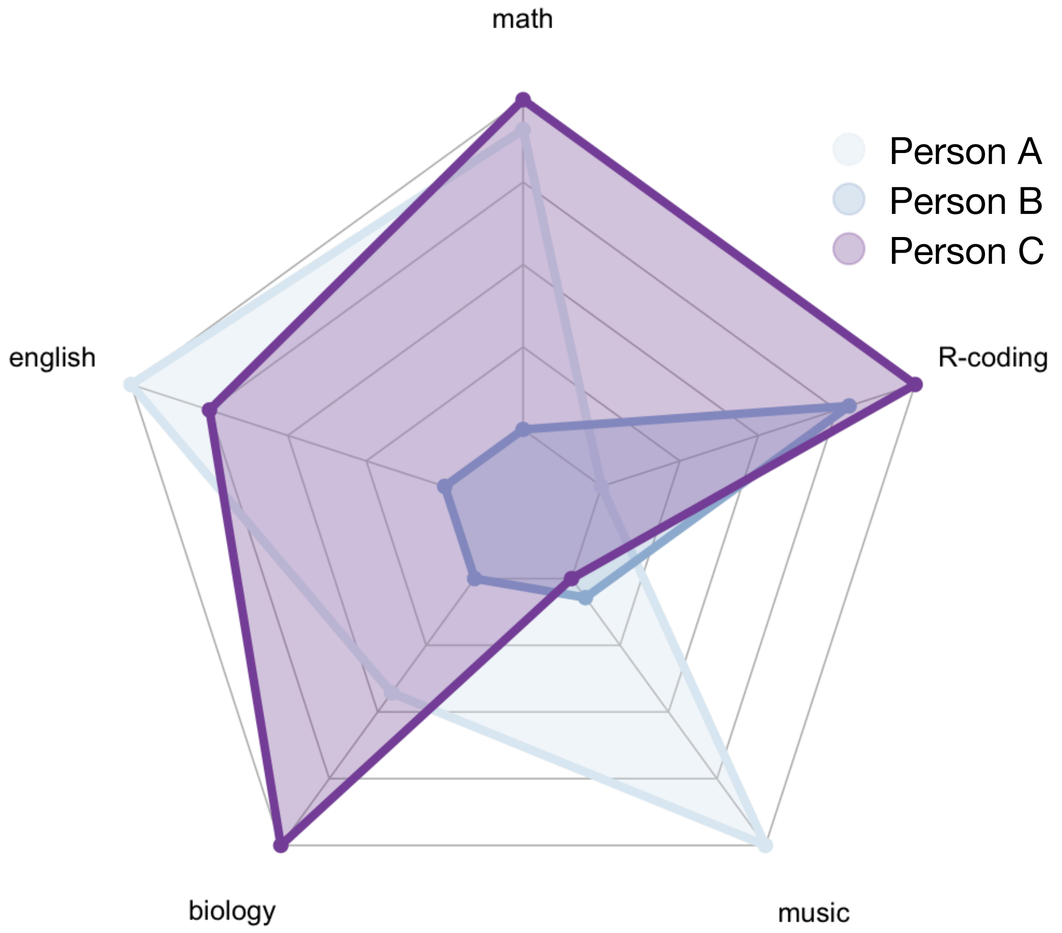

https://www.r-graph-gallery.com/stacked-barplot.html

# MAGNITUDE



## Radar plot

- A radar plot (a.k.a. spider chart) is a two-dimensional visualization designed to show **one or more series of values** over **multiple (ratio, interval or ordinal) variables**
- Radar plots are a space-efficient way of showing value of multiple variables, but make sure the organization makes sense!

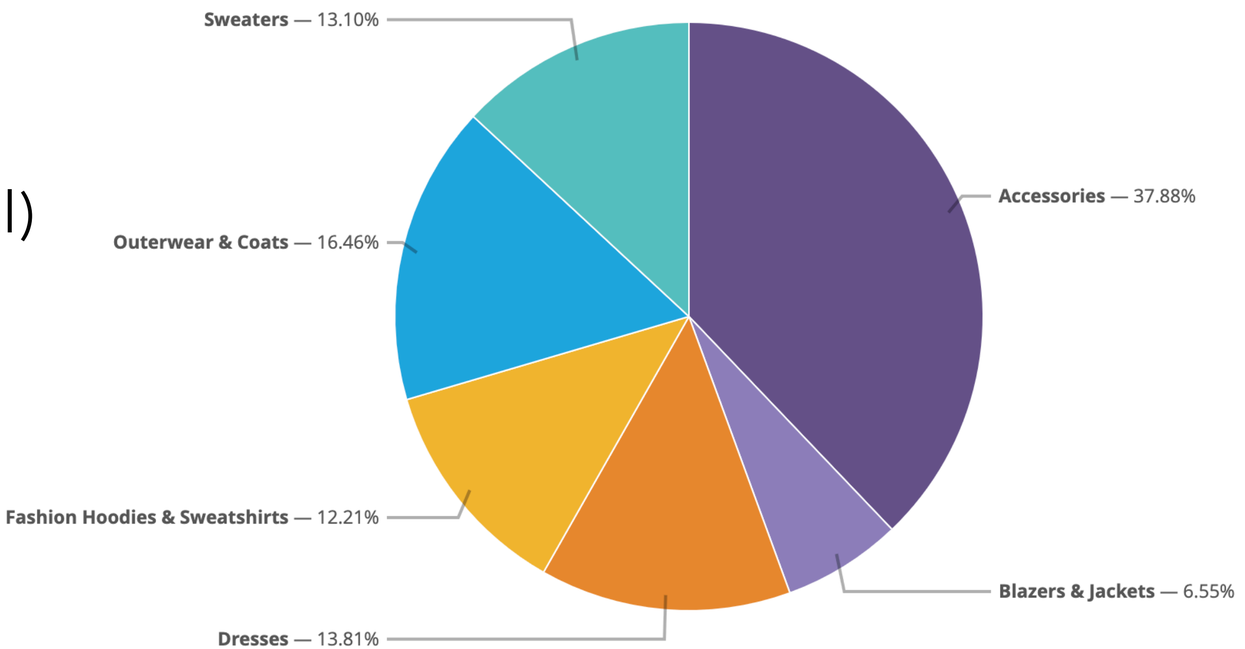


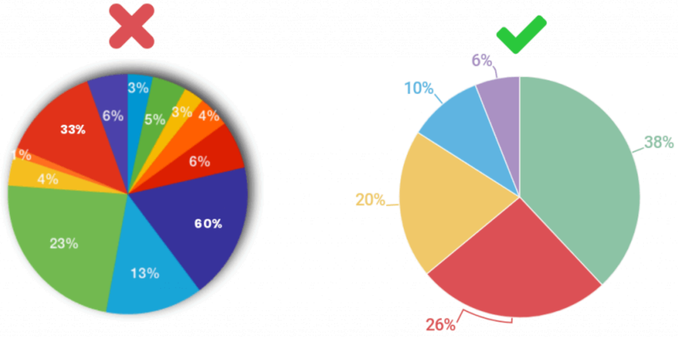

https://www.r-graph-gallery.com/spider-or-radar-chart.html

# PART-TO-WHOLE



- When?
  - If you want to how how a single entity represented by **discrete** (nominal/ordinal) data can be **broken down into its component elements**
- Types of visualizations
  -  **Pie chart**
    - Percentages determine the size of the pie slices
    - Works well when there is a **small number of distinct values**



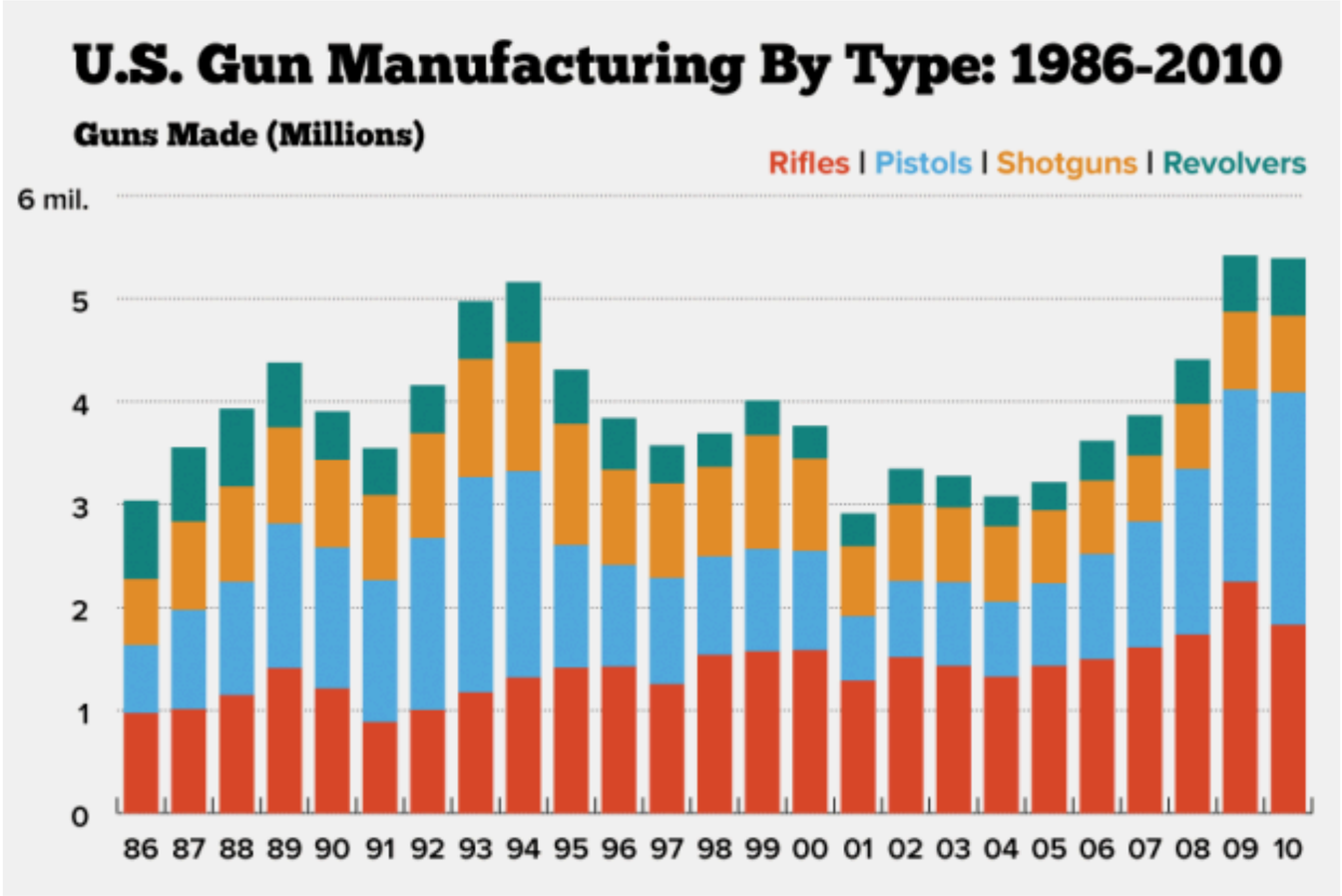




https://www.r-graph-gallery.com/pie-plot.html

# PART-TO-WHOLE



-  **Stacked bar chart**
  - Stacked bar charts show the composition of one **discrete** variable compared to another variable
  - Length can be **proportional to the total frequencies** or **normalized to add up to 100%**




https://www.r-graph-gallery.com/stacked-barplot.html

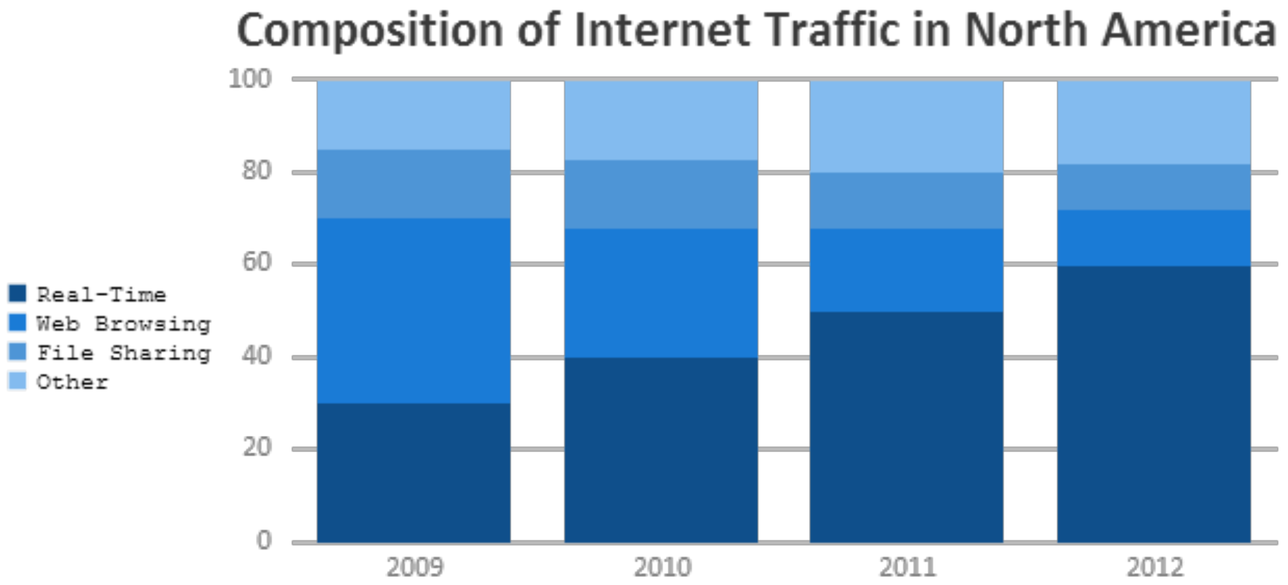
# PART-TO-WHOLE



-  **Stacked bar chart**

  - Stacked bar charts show the composition of one **discrete** variable compared to another variable

  - Length can be **proportional to the total frequencies** or **normalized to add up to 100%**

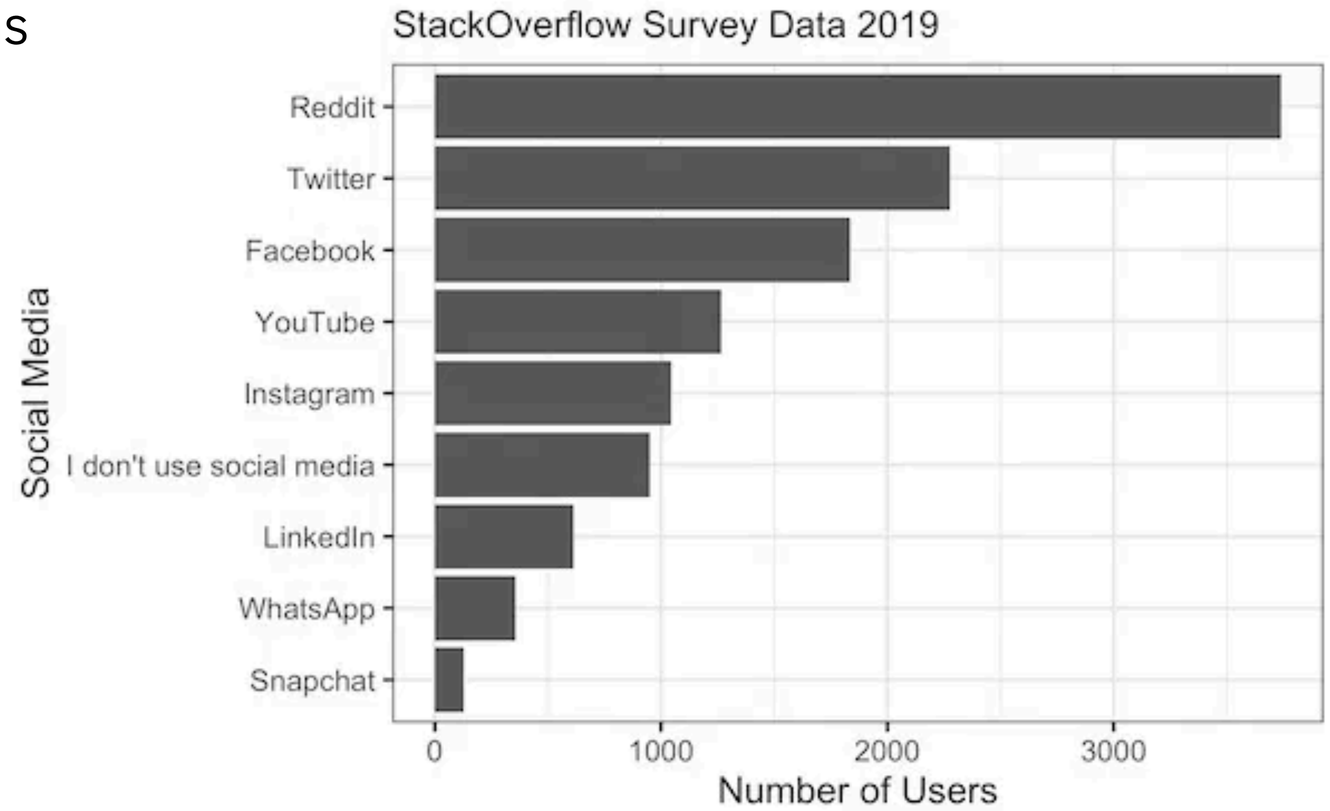



https://www.r-graph-gallery.com/stacked-barplot.html

# RANKING



- When?
  - If you want to show size comparisons where the **ordering is important**
- Types of visualizations
  -  **Ordered bar chart**
    - Same as a regular bar chart, but ordered by value
    - Stacked bar charts can also be ordered, either by their total or by one of its components

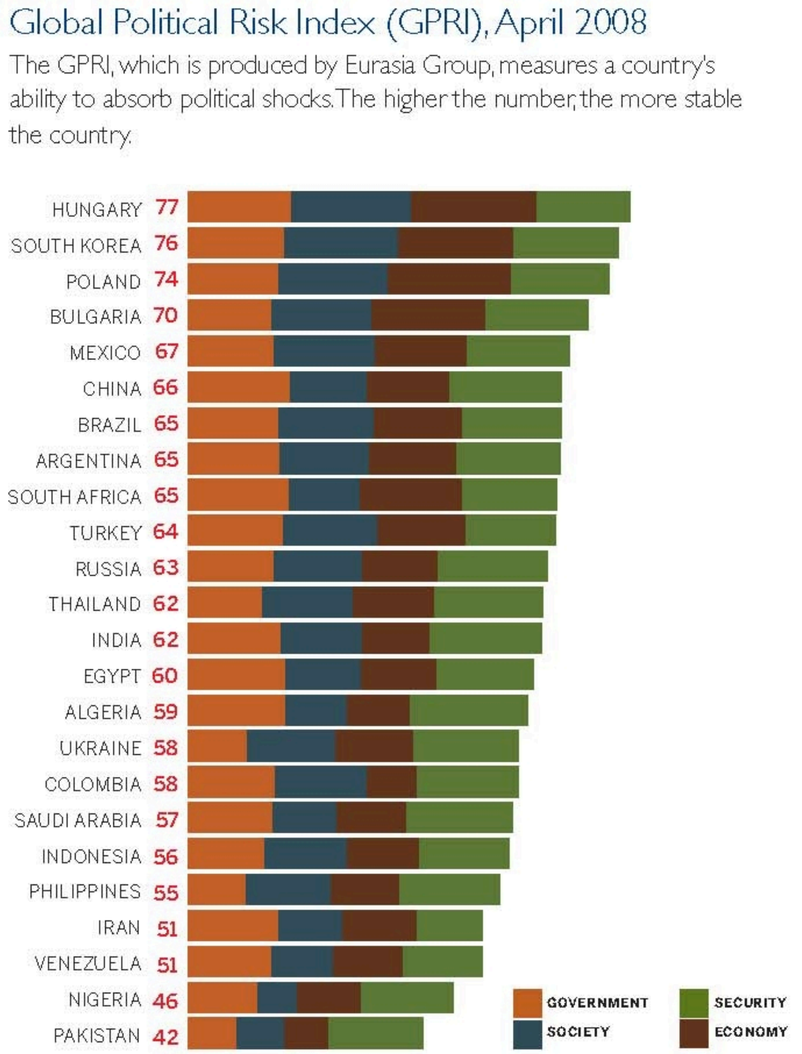



https://www.r-graph-gallery.com/barplot.html

# RANKING



- When?
  - If you want to show size comparisons where the **ordering is important**
- Types of visualizations
  -  **Ordered bar chart**
    - Same as a regular bar chart, but ordered by value
    - Stacked bar charts can also be ordered, either by their total or by one of its components




https://www.r-graph-gallery.com/barplot.html

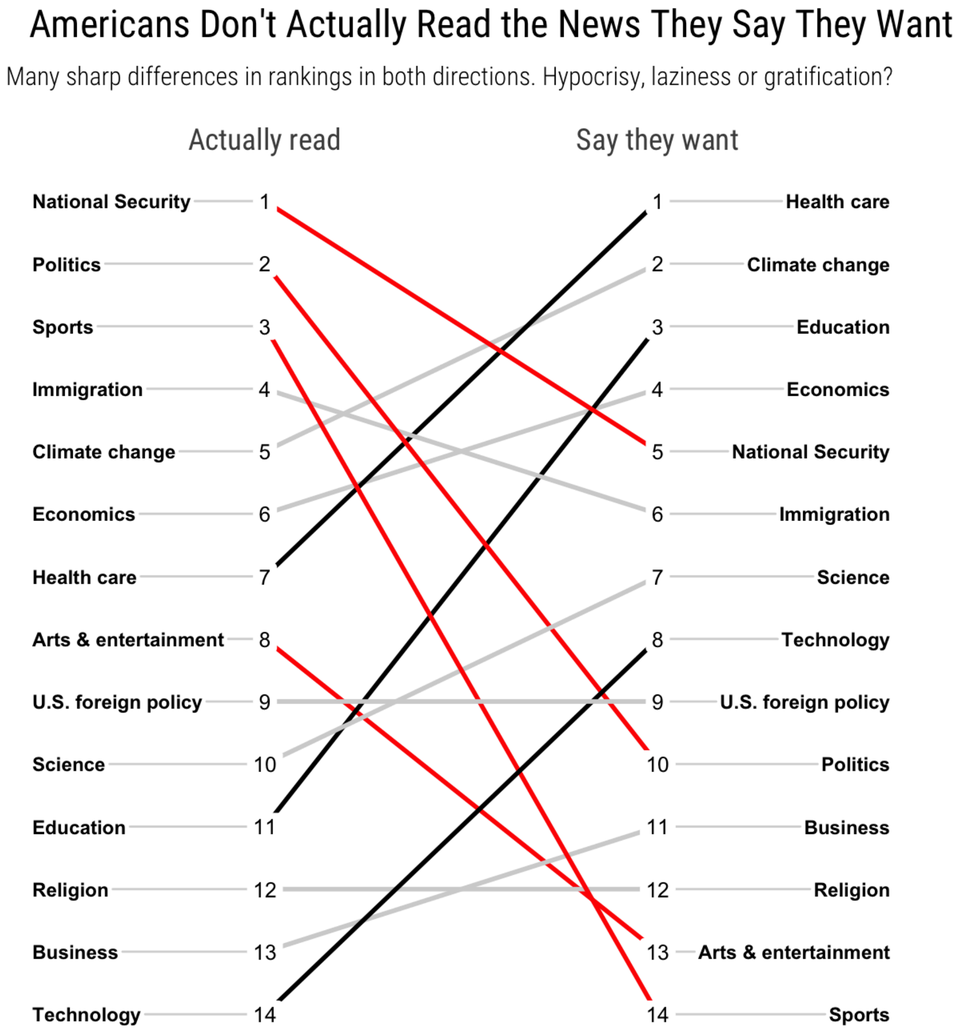
# RANKING



- 

## Slope

- Perfect for showing how ranks have changed over time or vary between categories

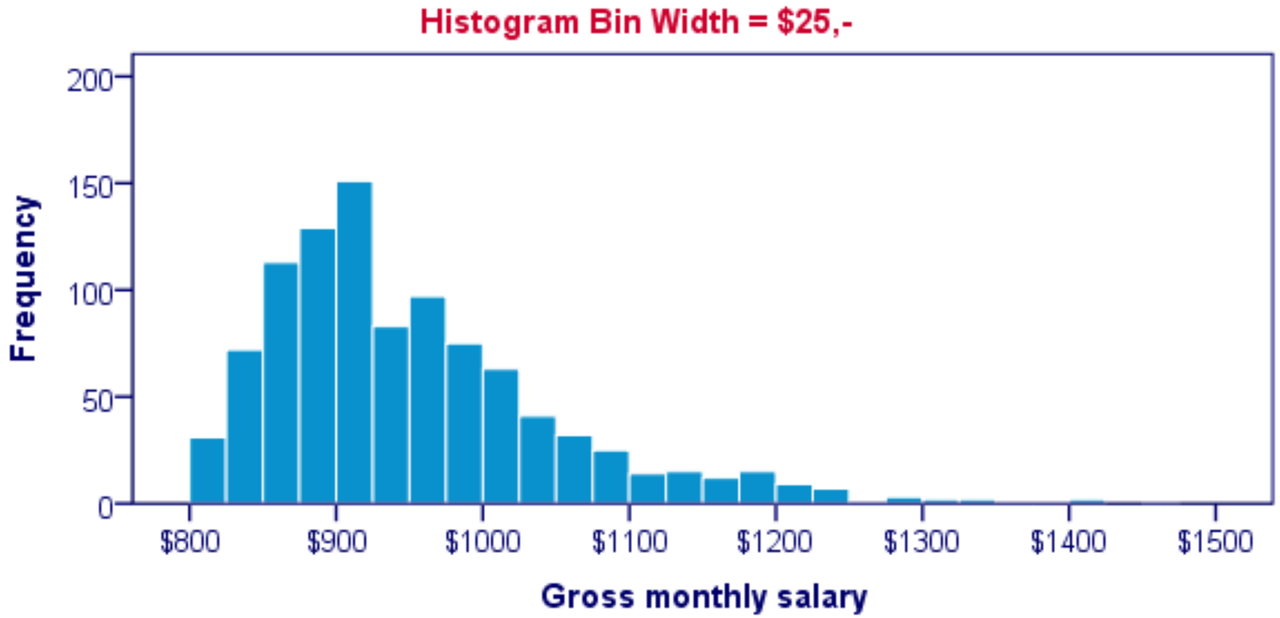



https://www.r-bloggers.com/2018/06/creating-slopegraphs-with-r/

# DISTRIBUTION



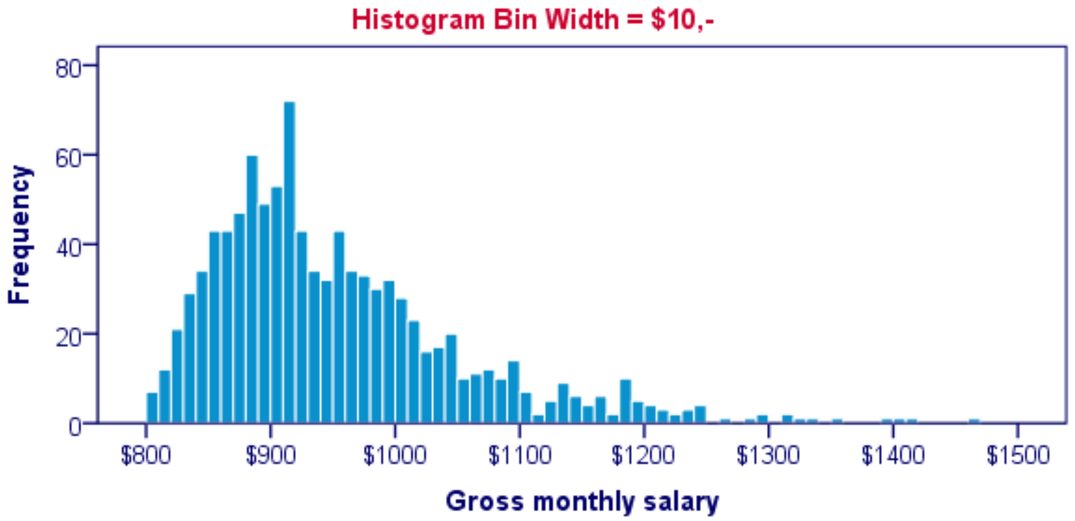
- When?
  - If you want to show values of a **continuous** variable and **how often they occur**
- Types of visualizations
  -  **Histogram**
    - Data plotted as adjacent rectangles for a set of **consecutive, non-overlapping bins**
    - Each bin covers a specific interval
    - Length of rectangle represents frequency or % of cases in that bin
    - Size of bins can be changed (but all equal-sized!)
    - Useful for visual appraisal of how your data is distributed

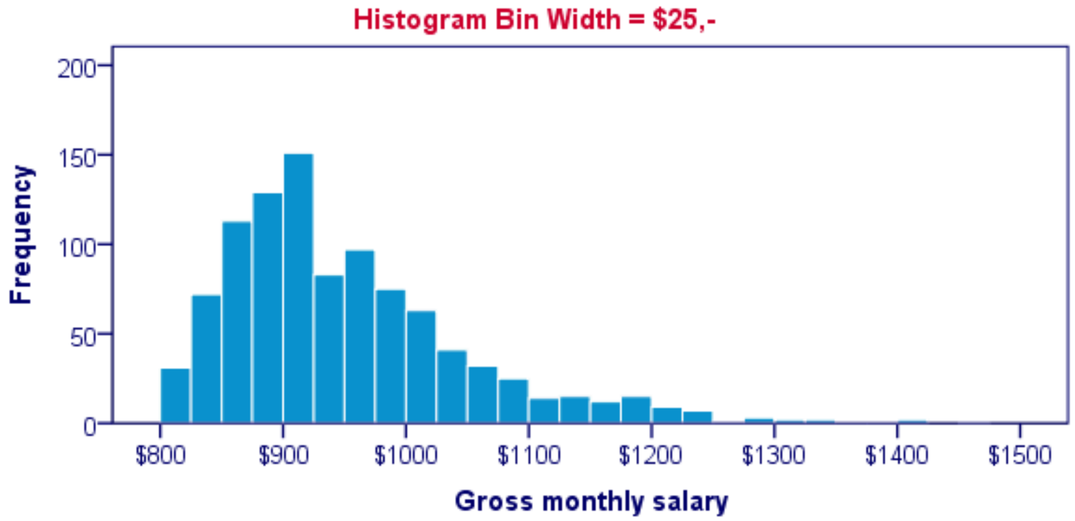


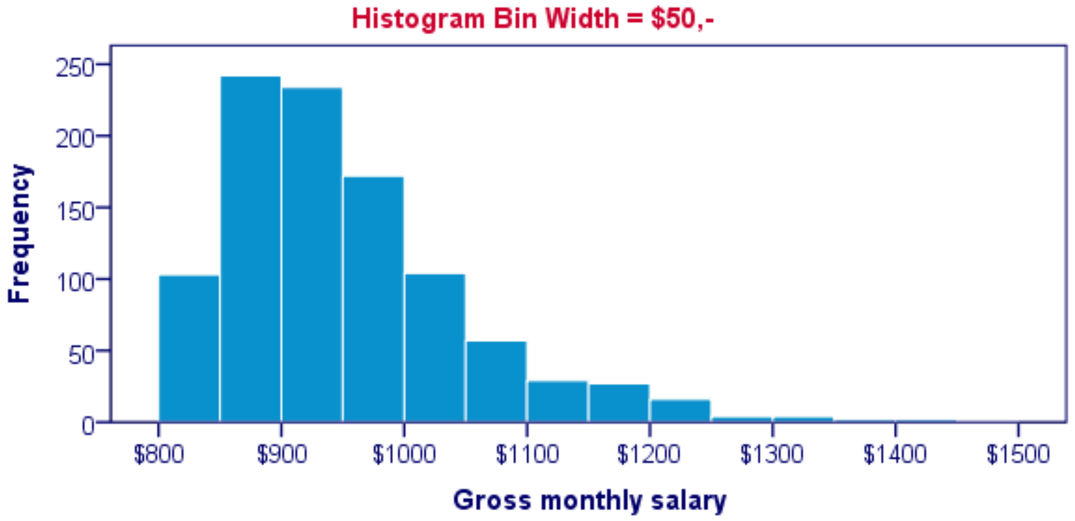

https://www.r-graph-gallery.com/histogram.html

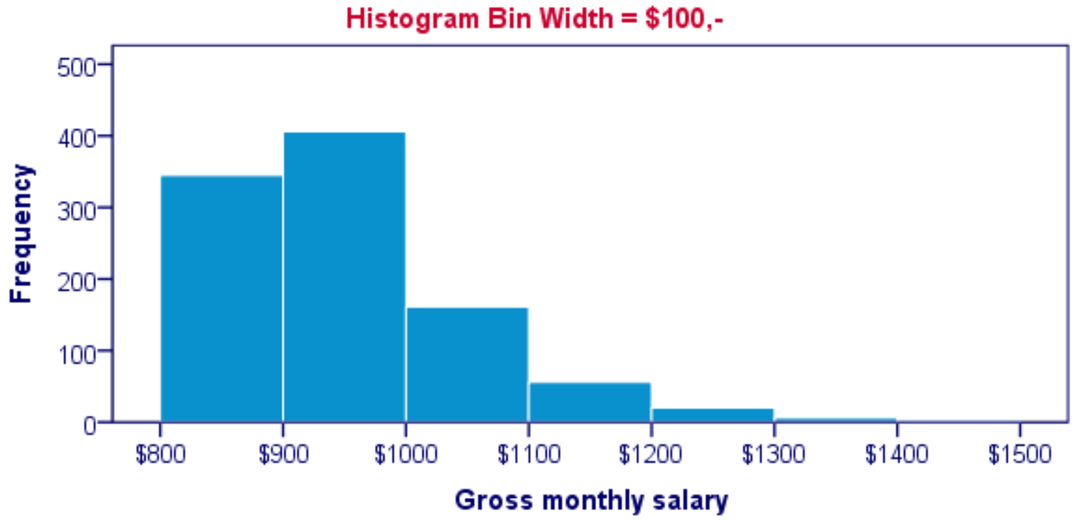
# DISTRIBUTION





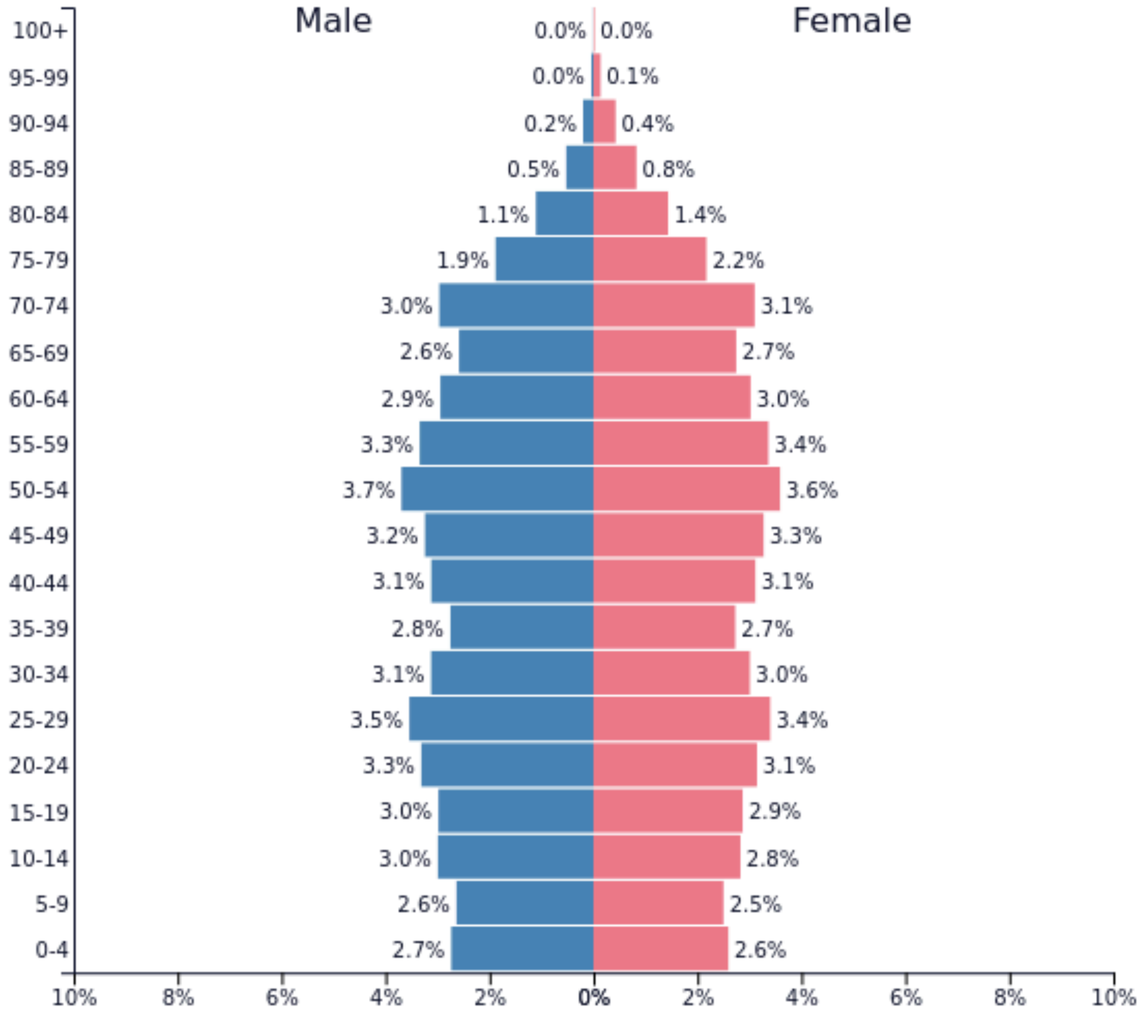




https://www.r-graph-gallery.com/histogram.html

# DISTRIBUTION



-  **Population pyramid**
  - A standard way for showing the age and sex breakdown of a population distribution
  - Basically two histograms with the same variable and bin size glued to each other

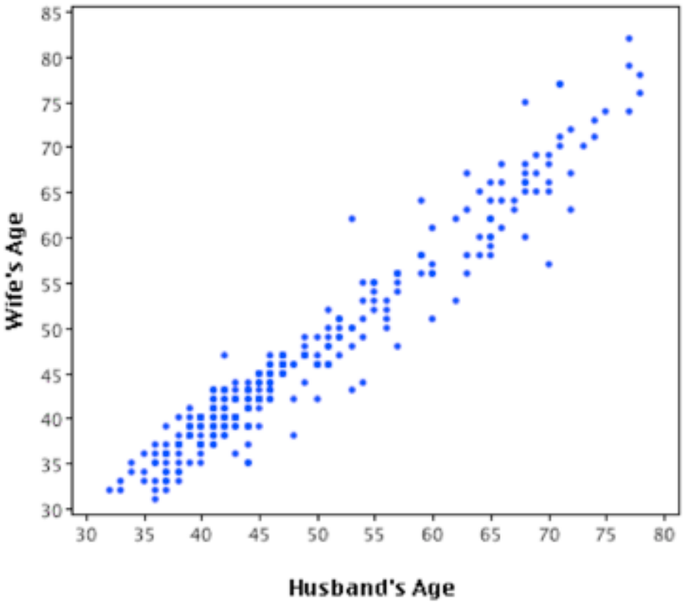


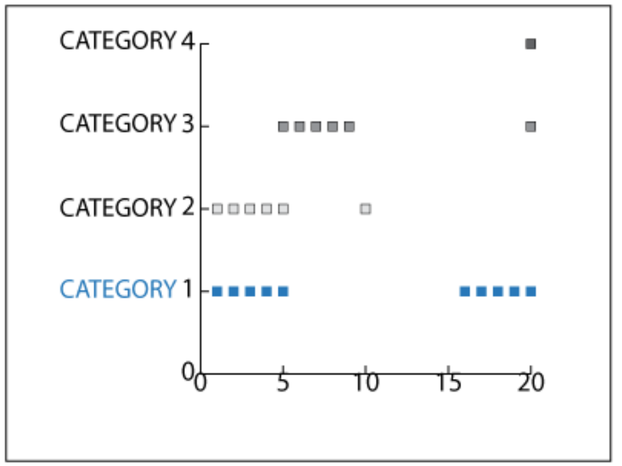

https://rpubs.com/walkerke/pyramids_ggplot2

# CORRELATION



- When?
  - If you want to show the **relationship between two or more (continuous) variables**
- Types of visualizations
  -  **Scatterplot**
    - Each variable has its own axis
    - Advantages
      - ★ Good overview of the **distribution** of values and their **relation**
      - ★ Good for detecting **outliers**
    - Disadvantages
      - ★ Not very useful for visualizing **discrete** variables



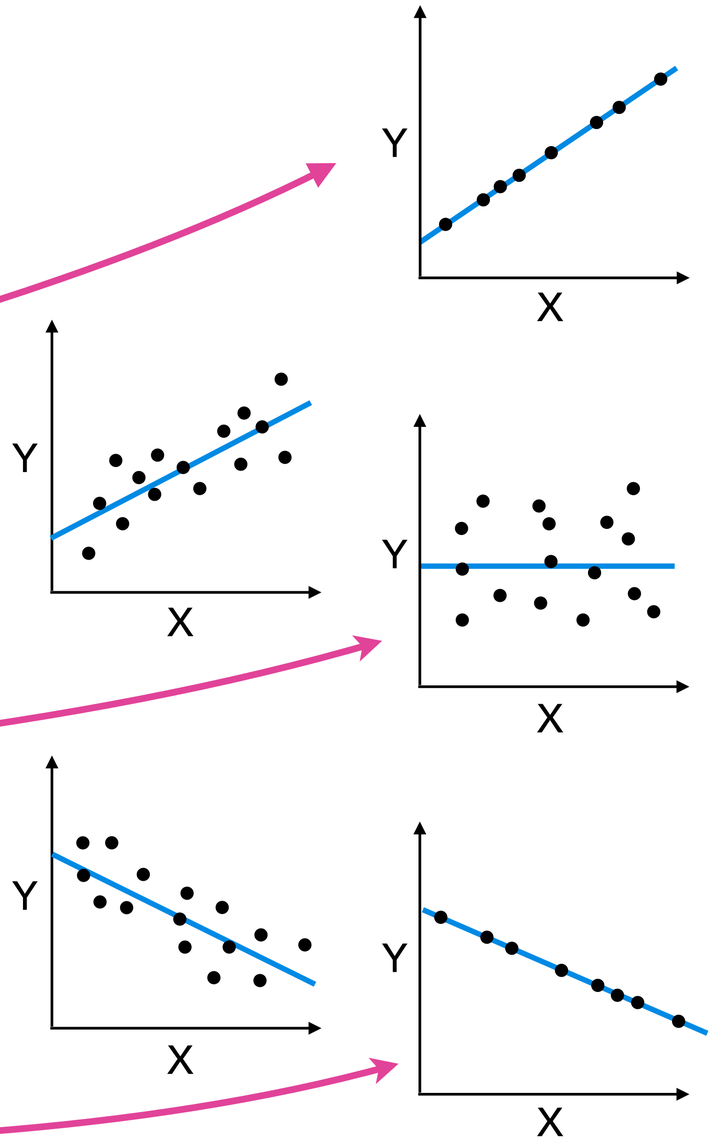




https://www.r-graph-gallery.com/scatterplot.html

# CORRELATION COEFFICIENT



- Scatterplots are closely related to another descriptive statistic: the **correlation coefficient**
  - A correlation coefficient is a **numeric summary** of the relationship we can see in the scatterplot
  - Correlation coefficients typically range from +1 to 0 to -1
    - +1 is perfect positive relationship
    - Between +1 and 0 is an imperfect positive relationship
      - ★ X increases → Y tends to increase as well
    - 0 is no relationship
    - Between 0 and -1 is an imperfect negative relationship
      - ★ X increases → Y tends to decrease
    - -1 is perfect negative relationship



# CALCULATING THE CORRELATION COEFFICIENT IN R



- You can calculate the correlation coefficient in R using the `cor.test()` function
  - The two variables to correlated should be provided as **two vectors**
  - The `method` parameter allows you to choose between different correlation coefficients
    - `pearson` for Pearson's *r*
    - `spearman` for Spearman's $\rho$ (rho)

# PICKING THE RIGHT CORRELATION COEFFICIENT



| Variable 2 / Variable 1 | Nominal | Ordinal | Interval/Ratio |
|---|---|---|---|
| **Nominal** | Cramer's *V* | Cramer's *V* | Not covered |
| **Ordinal** | Cramer's *V* | Spearman's $\rho$ | Spearman's $\rho$ |
| **Interval/Ratio** | Not covered | Spearman's $\rho$ | Pearson's *r* |

# PICKING THE RIGHT CORRELATION COEFFICIENT



| Variable 1 \ Variable 2 | Nominal | Ordinal | Interval/Ratio |
|---|---|---|---|
| **Nominal** | Cramer's *V* | Cramer's *V* | Not covered |
| **Ordinal** | Cramer's *V* | Spearman's $\rho$ | Spearman's $\rho$ |
| **Interval/Ratio** | Not covered | Spearman's $\rho$ | Pearson's *r* |

# PICKING THE RIGHT CORRELATION COEFFICIENT



| Variable 1 \ Variable 2 | Nominal | Ordinal | Interval/Ratio |
|---|---|---|---|
| **Nominal** | Cramer's *V* | Cramer's *V* | Not covered |
| **Ordinal** | Cramer's *V* | Spearman's $\rho$ | Spearman's $\rho$ |
| **Interval/Ratio** | Not covered | Spearman's $\rho$ | Pearson's *r* |

# PICKING THE RIGHT CORRELATION COEFFICIENT



| Variable 1 \ Variable 2 | Nominal | Ordinal | Interval/Ratio |
|---|---|---|---|
| **Nominal** | Cramer's *V* | Cramer's *V* | Not covered |
| **Ordinal** | Cramer's *V* | Spearman's $\rho$ | Spearman's $\rho$ |
| **Interval/Ratio** | Not covered | Spearman's $\rho$ | Pearson's *r* |

# CALCULATING THE CORRELATION COEFFICIENT IN R



- What is the correlation between no. of followers and no. of accounts followed?
  - Which correlation coefficient should we use?

```
1  cor.test(survey$follower_count, survey$accounts_followed, method = "pearson")
```

```
> cor.test(survey$follower_count, survey$accounts_followed, method = "pearson")

   Pearson's product-moment correlation

data:  survey$follower_count and survey$accounts_followed
t = 9.7079, df = 265, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.4178360 0.5956142
sample estimates:
      cor
0.5121905
```

# CORRELATION COEFFICIENT



- How do we interpret imperfect correlations?

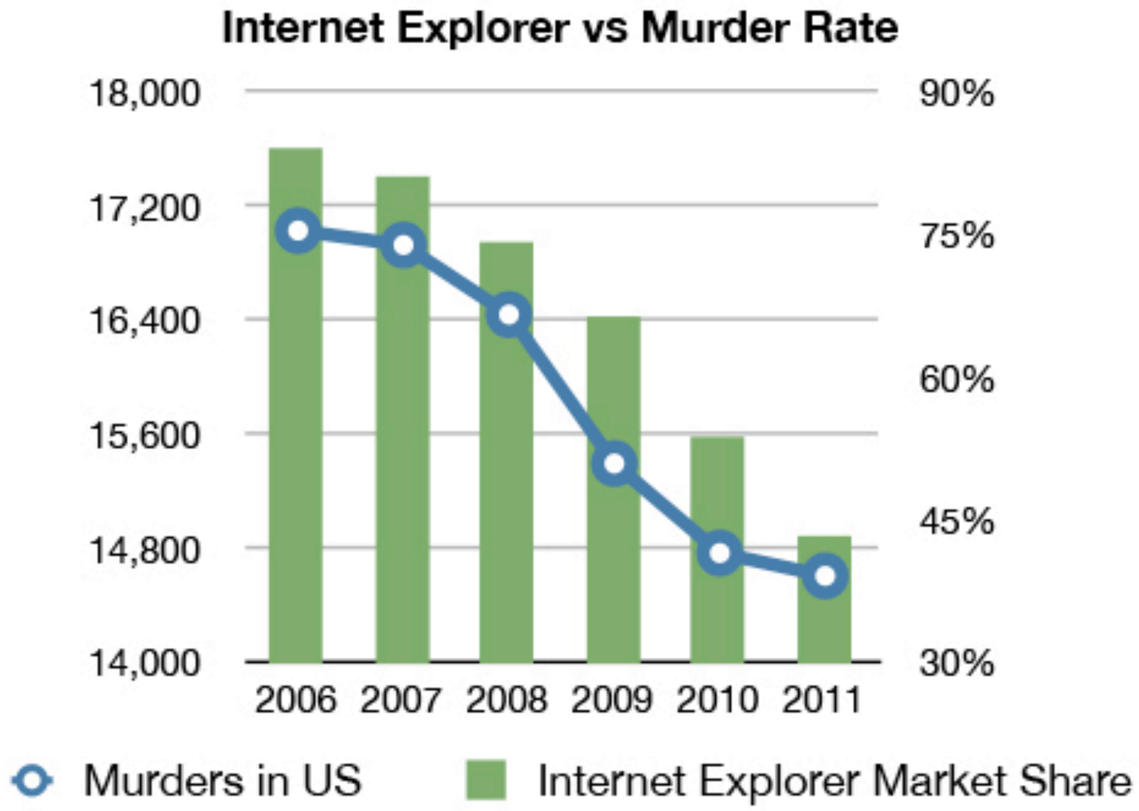
| Value | Interpretation |
|---|---|
| 0 | No relationship |
| 0.01-0.19 | Negligible relationship |
| 0.20-0.29 | Weak relationship |
| 0.30-0.39 | Moderate relationship |
| 0.40-0.69 | Strong relationship |
| 0.70-0.99 | Very strong relationship |
| 1 | Perfect relationship |

# SPURIOUS CORRELATIONS



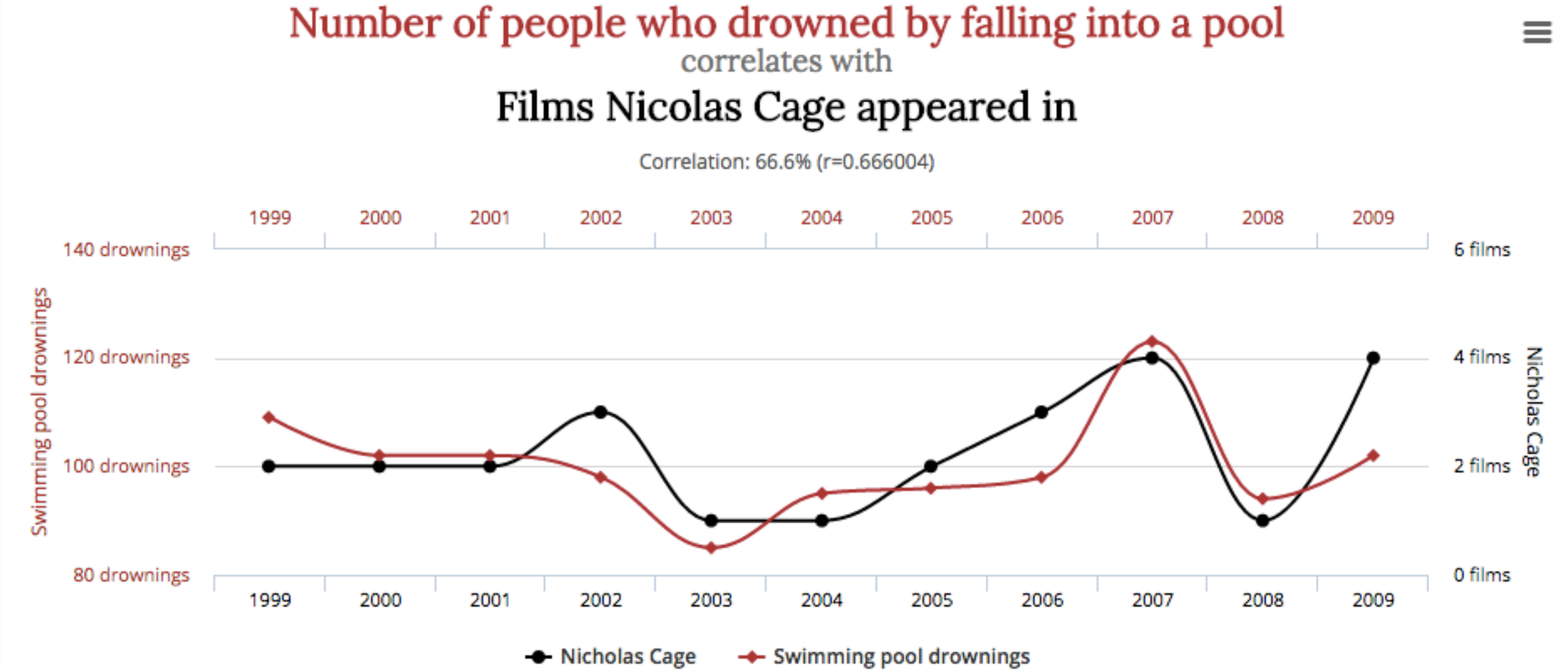
spurious = fake

- **Correlation ≠ causation**
  - Just because two variables are correlated that does not mean one causes the other!



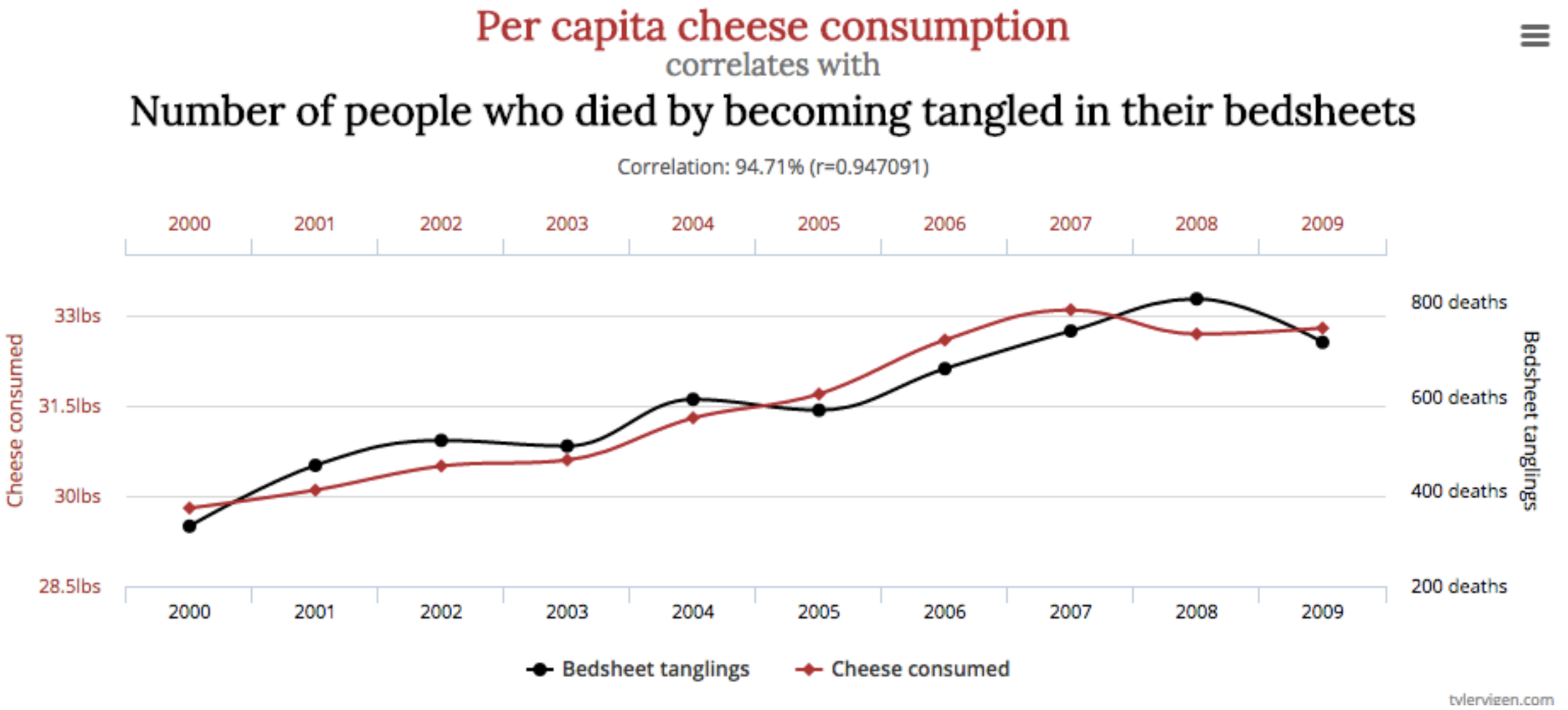
# SPURIOUS CORRELATIONS



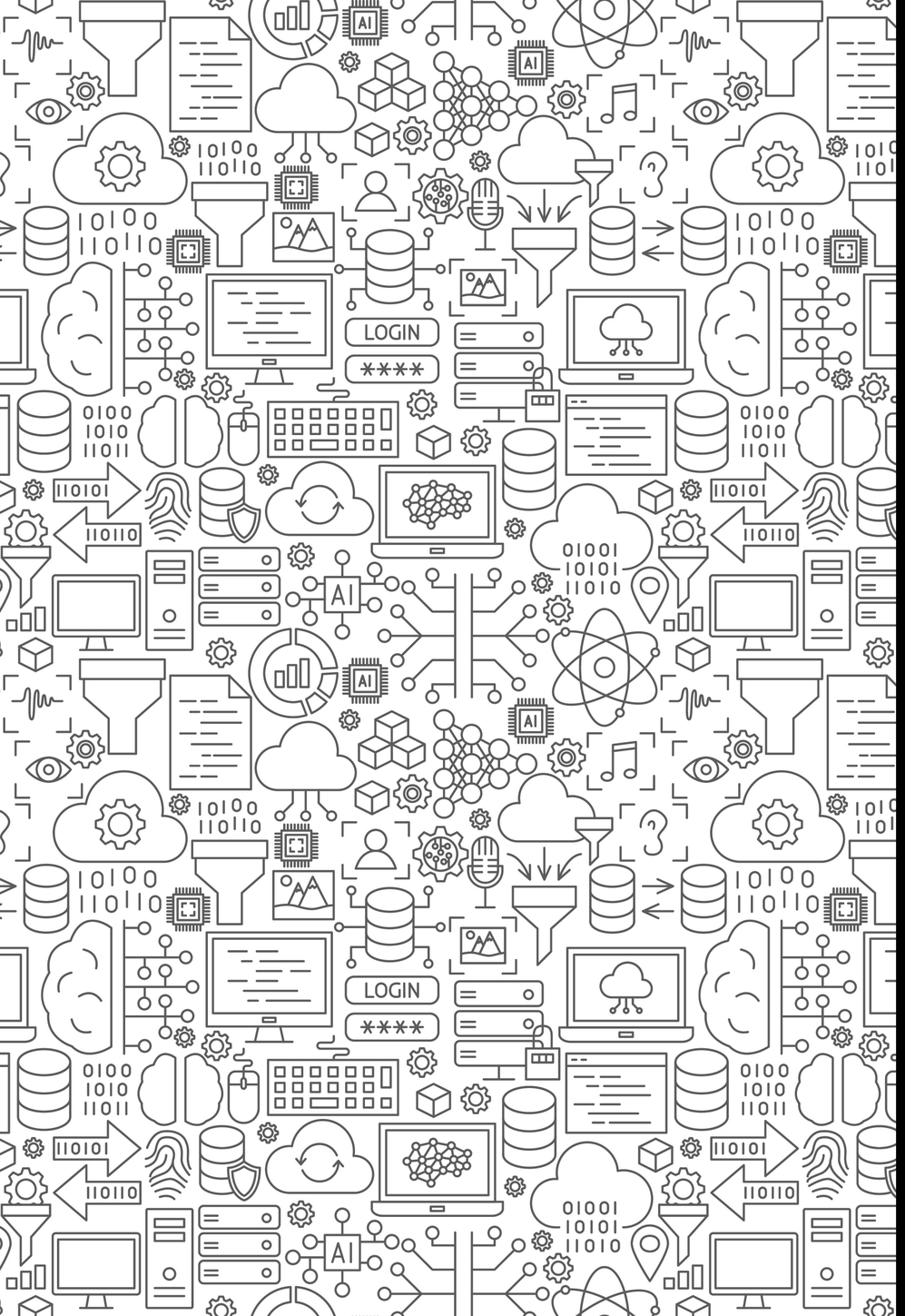


# EXAMPLES OF SPURIOUS CORRELATIONS






- Find more at http://www.tylervigen.com/spurious-correlations

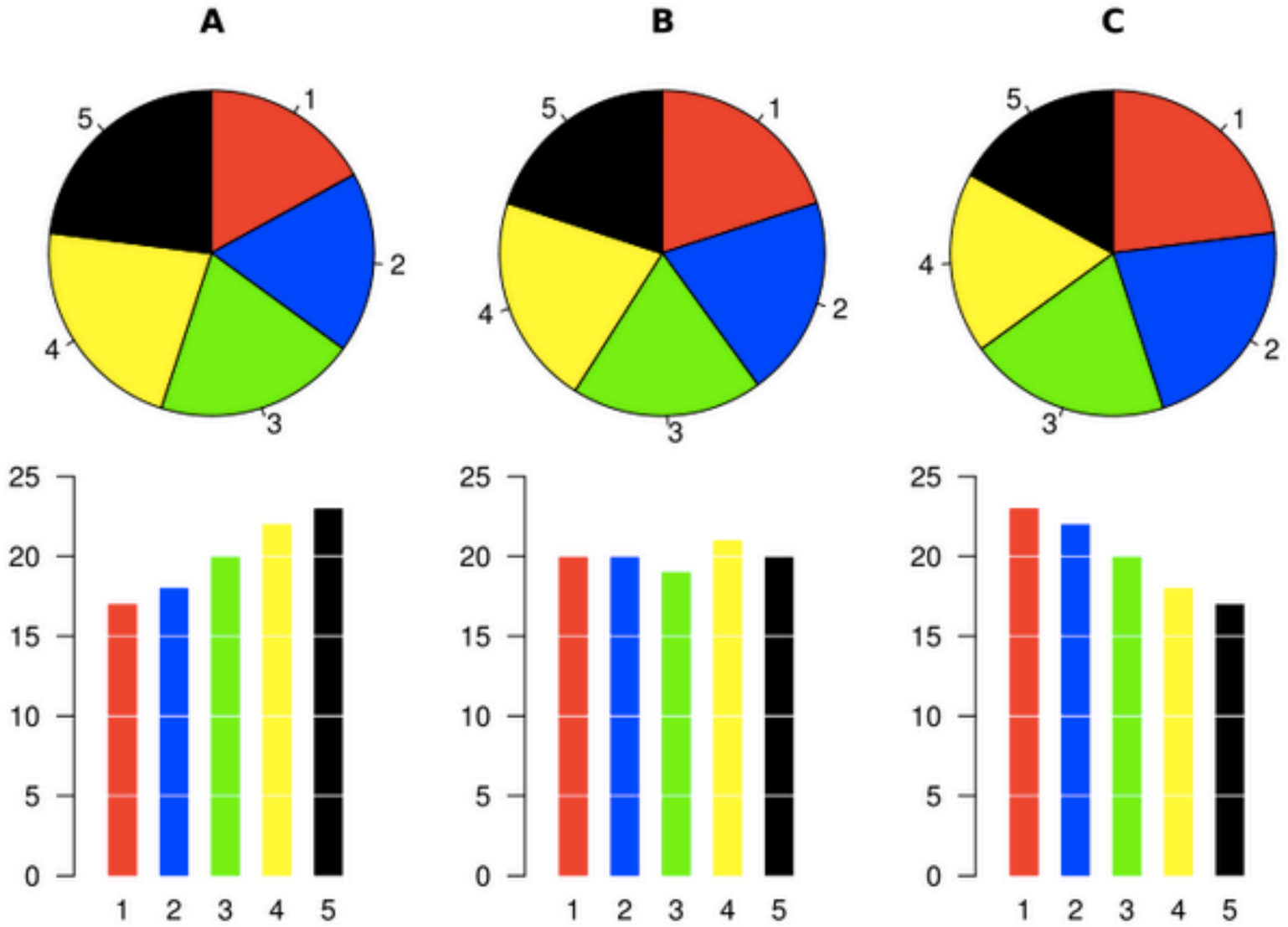



PART 6

# PITFALLS

# AVOID PIE CHARTS IF YOU CAN



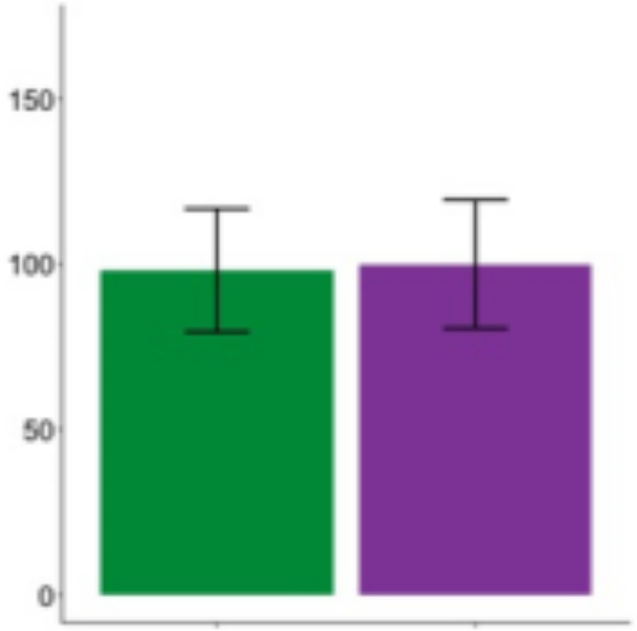


# BAR CHARTS AREN'T ALWAYS THE SOLUTION EITHER

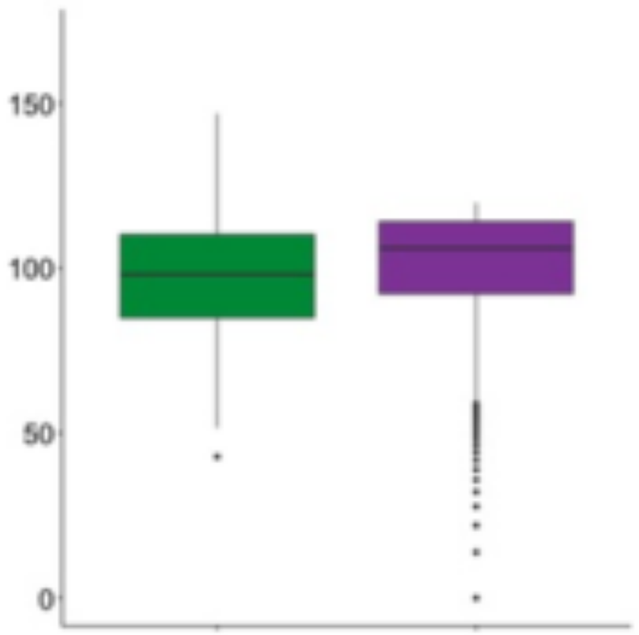


**Friends don't let friends make unnecessary bar plots.**

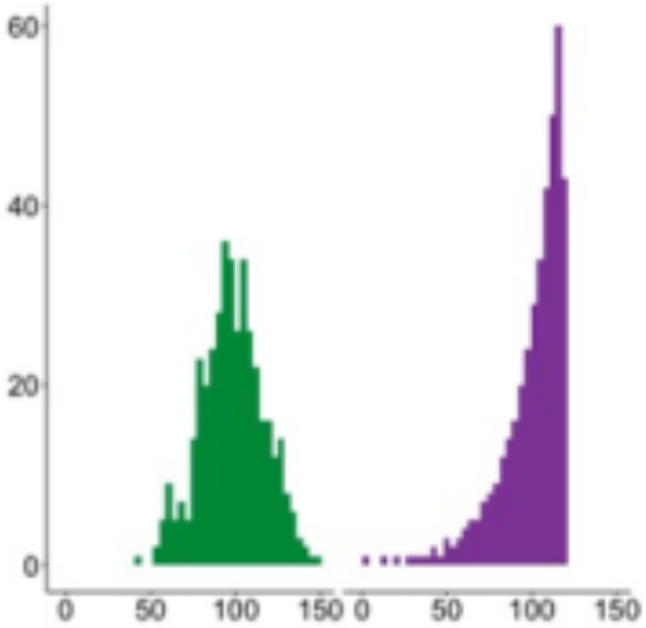
These look the same!




Wait a minute!



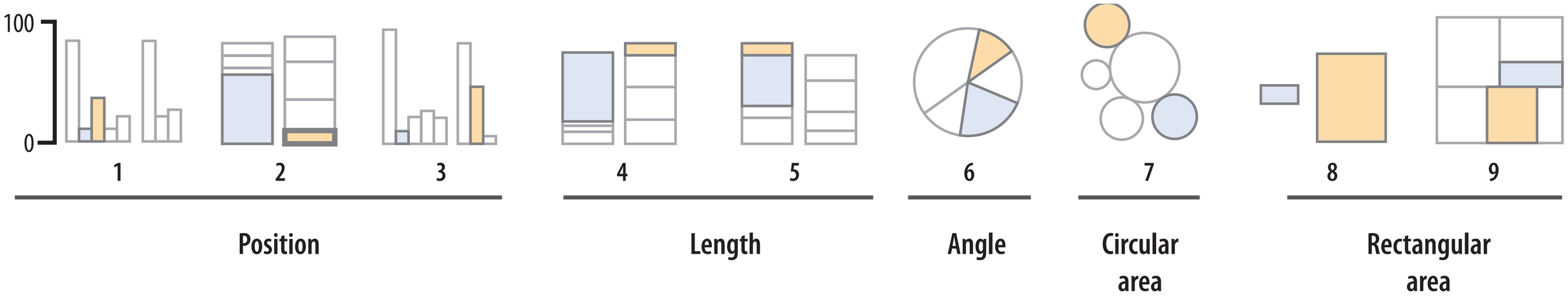

Oooh!



# INTERPRETABILITY OF VISUALIZATION TYPES



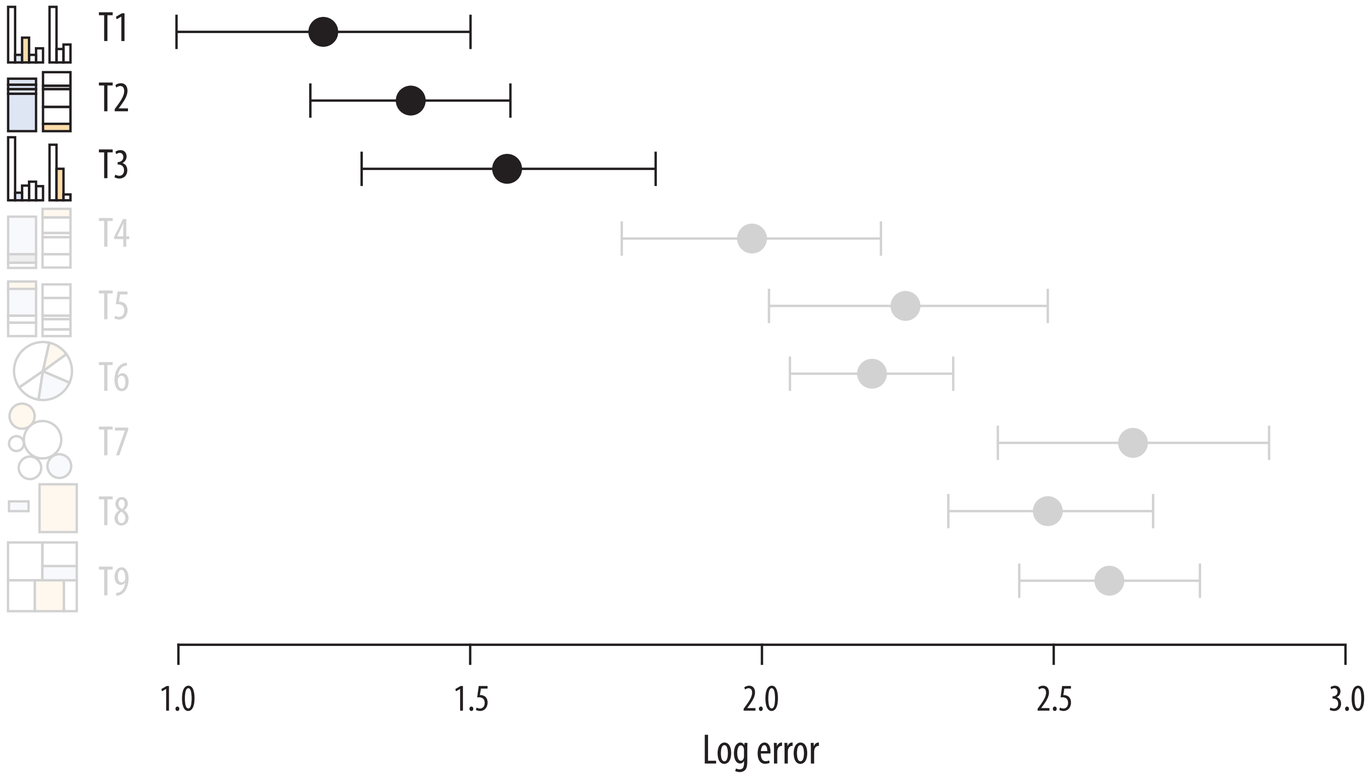
- These examples show us that some visualizations are easier to interpret than others!
  - Cleveland & McGill (1984) and Heer & Bostock (2010) have studied how good people are inferring information from different types of visualizations
  - Participants were asked to (1) identify the smaller of two marked segments, and (2) estimate what percentage the smaller one was of the larger (Heer & Bostock, 2010)



# INTERPRETABILITY OF VISUALIZATION TYPES



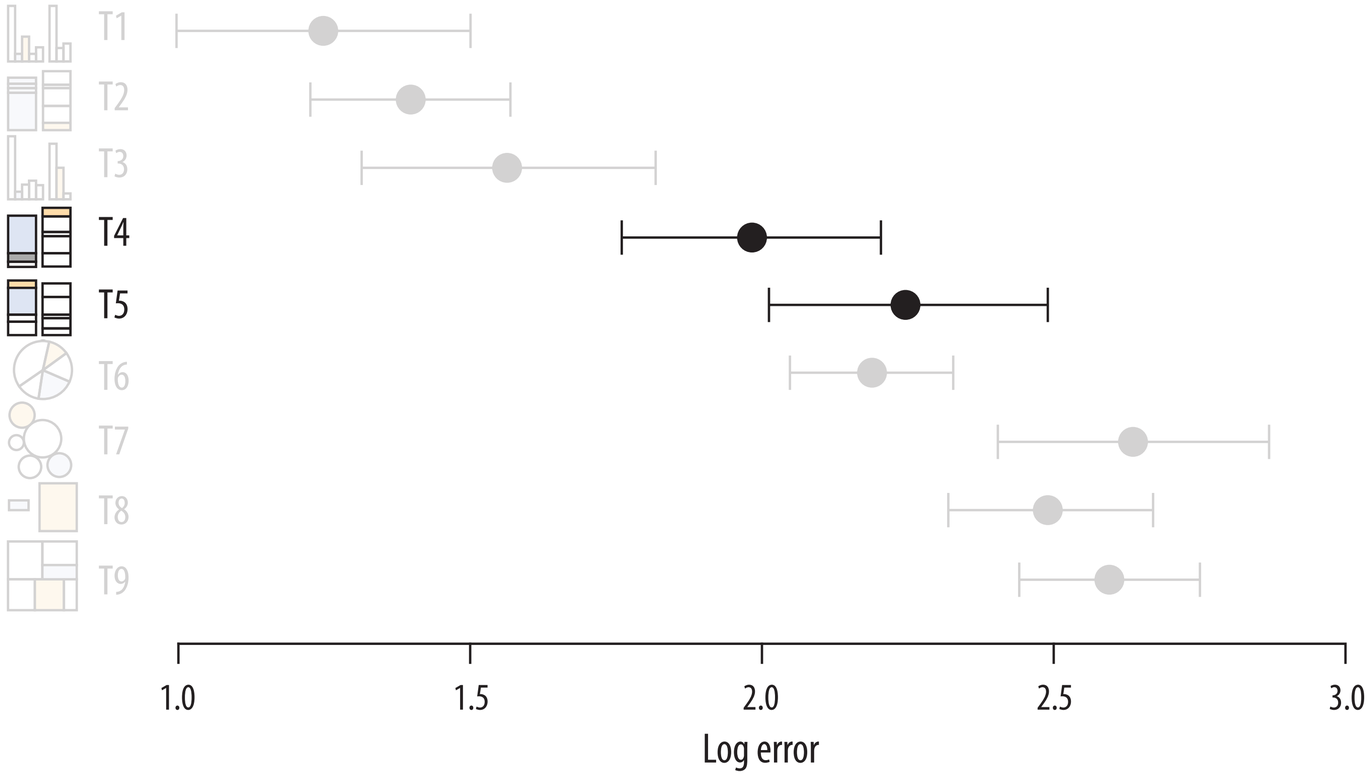
- Their results showed that there are **better and worse ways** of visualizing data
  - **Relative position (T1-T3)** is the easiest because of the **common scale** and **alignment** of elements



# INTERPRETABILITY OF VISUALIZATION TYPES



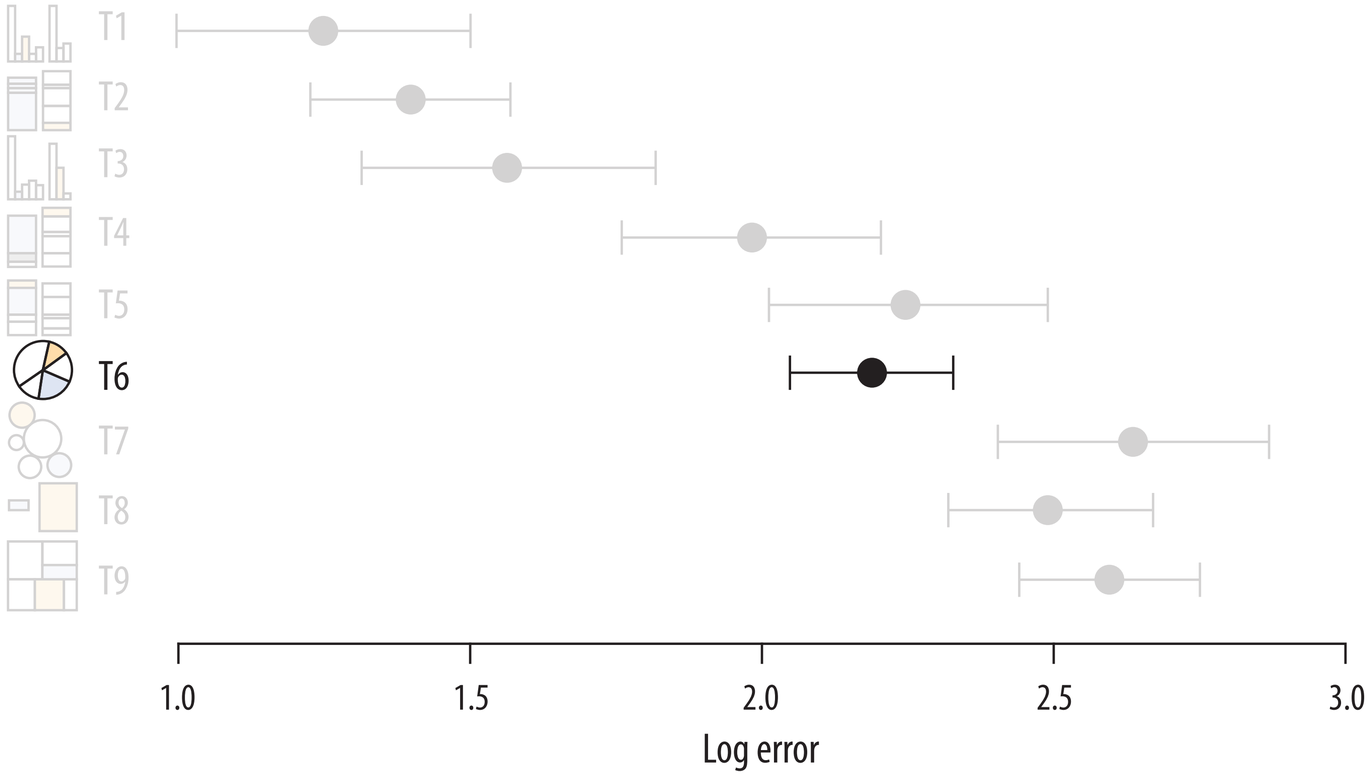
- Their results showed that there are **better and worse ways** of visualizing data
  - **Relative position (T1-T3)** is the easiest because of the **common scale** and **alignment** of elements
  - If we only have the **length (T4-T5)** on a common scale, things get more difficult



# INTERPRETABILITY OF VISUALIZATION TYPES



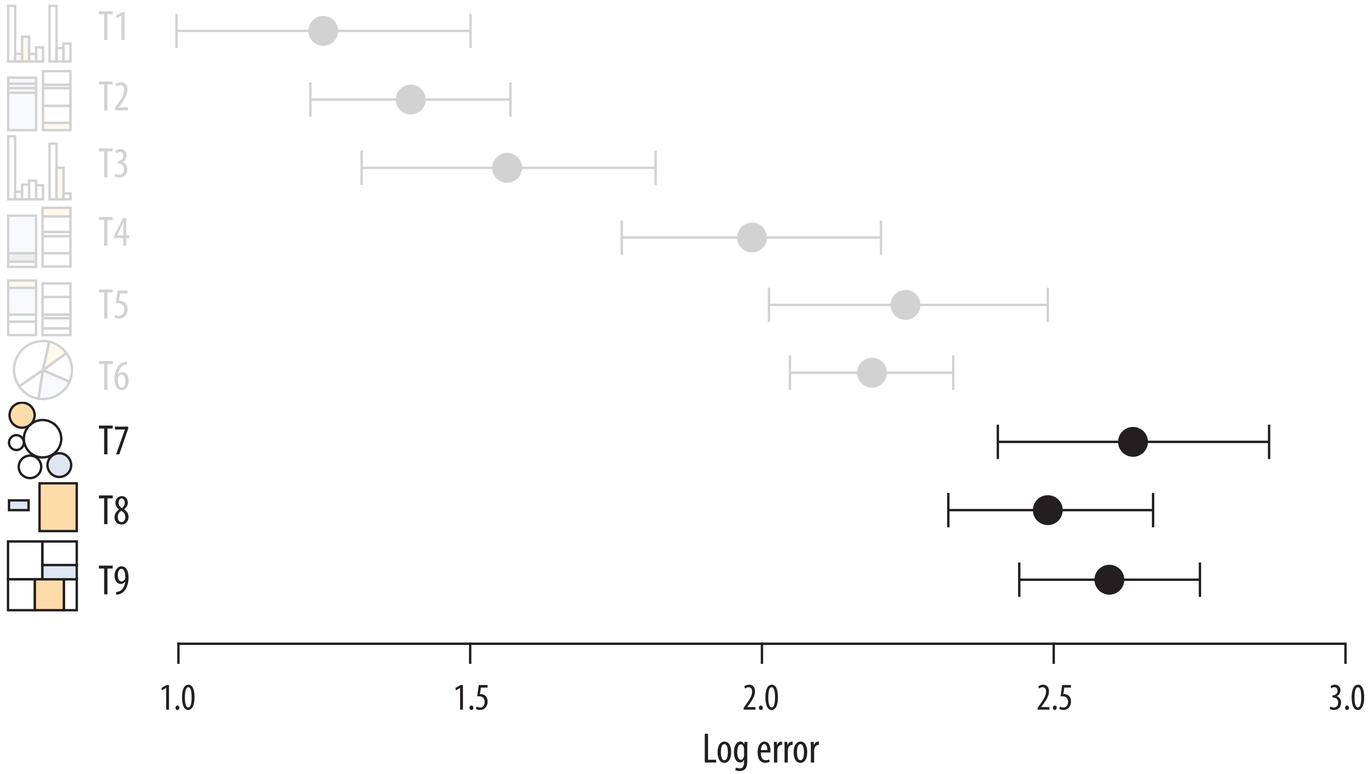
- Their results showed that there are **better and worse ways** of visualizing data
  - **Relative position (T1-T3)** is the easiest because of the **common scale** and **alignment** of elements
  - If we only have the **length (T4-T5)** on a common scale, things get more difficult
  - Judging **angles (T6)** is even more difficult



# INTERPRETABILITY OF VISUALIZATION TYPES



- Their results showed that there are **better and worse ways** of visualizing data
  - **Relative position (T1-T3)** is the easiest because of the **common scale** and **alignment** of elements
  - If we only have the **length (T4-T5)** on a common scale, things get more difficult
  - Judging **angles (T6)** is even more difficult
  - **Area**-based comparisons (T7-T9) perform the worst



# AVOIDING BAD VISUALIZATIONS



- Some more good tips on how to avoid making bad visualizations
  - https://www.data-to-viz.com/caveats.html


# QUESTIONS?

# REFERENCES




- Bryman, A. (2016). Quantitative data analysis (chapter 15). In *Social Research Methods* (5th ed., pp. 329-351). Oxford: Oxford University Press.
- Fekete, J. D., Van Wijk, J., Stasko, J., & North, C. (2008). The Value of Information Visualization. *Information Visualization*, 1-18.
- Healy, K. (2018). Look at Data (chapter 1). In *Data Visualization: A Practical Introduction* (1st ed., pp. 1-31). Princeton University Press.
- Hinton, P. R. (2014). Descriptive statistics (chapter 2). In: *Statistics Explained* (4th ed.), New York, NY: Routledge.
- Kirk, A. (2012). Chapter 4: Conceiving and Reasoning Visualization Design Options. In *Data Visualization: A Successful Design Process* (pp. 79-117). Birmingham: Packt Publishing.
- Kirk, A. (2012). Chapter 5: Taxonomy of Data Visualization Methods. In *Data Visualization: A Successful Design Process* (pp. 119-158). Birmingham: Packt Publishing.
- Lazar, J., Feng, J.H., and Hochheiser, H. (2010). Statistical analysis (chapter 4). In: *Research Methods in Human-Computer Interaction* (2nd ed., pp. 71-104), Hoboken, NJ: John Wiley & Sons Ltd.

# REFERENCES




- Littlewood, C. (2018, December 18). *Prioritize Which Data Skills Your Company Needs with This 2×2 Matrix*. Retrieved September 30, 2020, from https://hbr.org/2018/10/prioritize-which-data-skills-your-company-needs-with-this-2x2-matrix
- Matejka, J., & Fitzmaurice, G. (2017). Same Stats, Different Graphs: Generating Datasets with Varied Appearance and Identical statistics through simulated Annealing. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (pp. 1290-1294).
- Shneiderman, B. (1996). The Eyes have it: A Task by Data Type Taxonomy for Information Visualizations. In *Proceedings of the IEEE Symposium on Visual Languages* (pp. 336-343).
- Sirkin (1999). Statistics for the Social Sciences. SAGE Publications
- Smith, A., Campbell, C., Bott, I., Faunce, L., Parrish, G., Ehrenberg-Shannon, B., . . . Stabe, M. (2019). *Financial Times Visual Vocabulary*. Retrieved October 01, 2020, from https://github.com/ft-interactive/chart-doctor/tree/master/visual-vocabulary
- Wilke, Claus O. (2019). *Fundamentals of Data Visualization*. O'Reilly Media.