Module 03
Data Discovery
Lecture 2

Wan Fang

Southern University of Science and Technology

Agenda

- Descriptive vs. Inferential
  - Descriptive Statistics
  - Inferential Statistics

- Exploratory Data Analysis (EDA)

DS363: Design and Learning with Data

Spring 2023

Descriptive vs. Inferential

Wan Fang

Southern University of Science and Technology

[Adapted from Statistics by Scribbr]

A Beginner's Guide to Statistical Analysis

- Investigating trends, patterns, and relationships using quantitative data.
  - An important research tool used by scientists, governments, businesses, and other organizations.
- Statistical analysis planning.
  - Specify your hypotheses and make decisions about your *research design*, *sample size*, and *sampling procedure*.
  - **After collecting data** from your sample, you can organize and summarize the data using ***descriptive statistics***.
  - **Then**, you can use ***inferential statistics*** to formally test hypotheses and make estimates about the population.
  - **Finally**, interpret and generalize your findings.

Example: Causal research question

- Can meditation improve exam performance in teenagers?

Example: Correlational research question

- Is there a relationship between parental income and college grade point average (GPA)?

Step 1: Write your hypotheses and plan your research design

Step 2: Collect data from a sample

Step 3: Summarize your data with descriptive statistics

Step 4: Test hypotheses or make estimates with inferential statistics

Step 5: Interpret your results

Types of Descriptive Statistics

- The **distribution** concerns the frequency of each value.
- The **central tendency** concerns the averages of the values.
- The **variability** or dispersion concerns how spread out the values are.

Research example

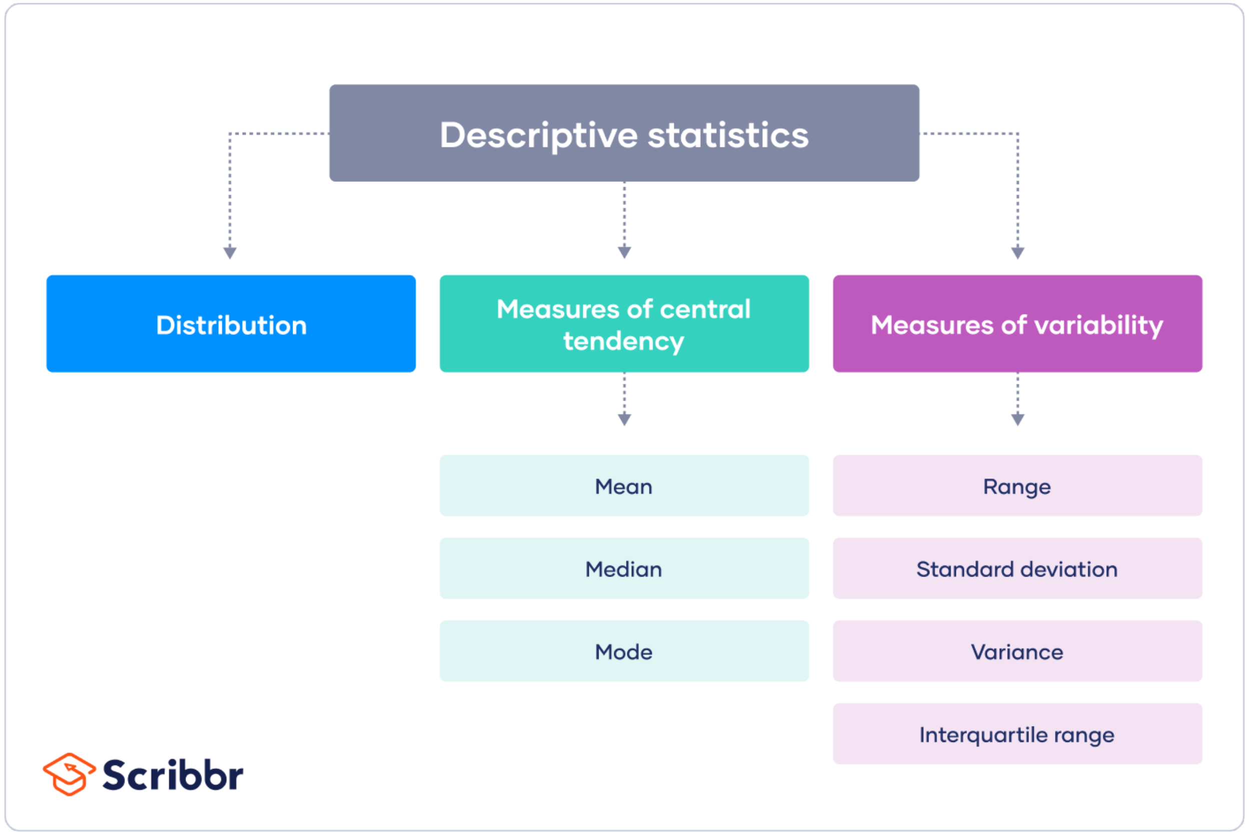
You want to study the popularity of different leisure activities by gender.

You distribute a survey and ask participants how many times they did each of the following in the past year:

- Go to a library
- Watch a movie at a theater
- Visit a national park

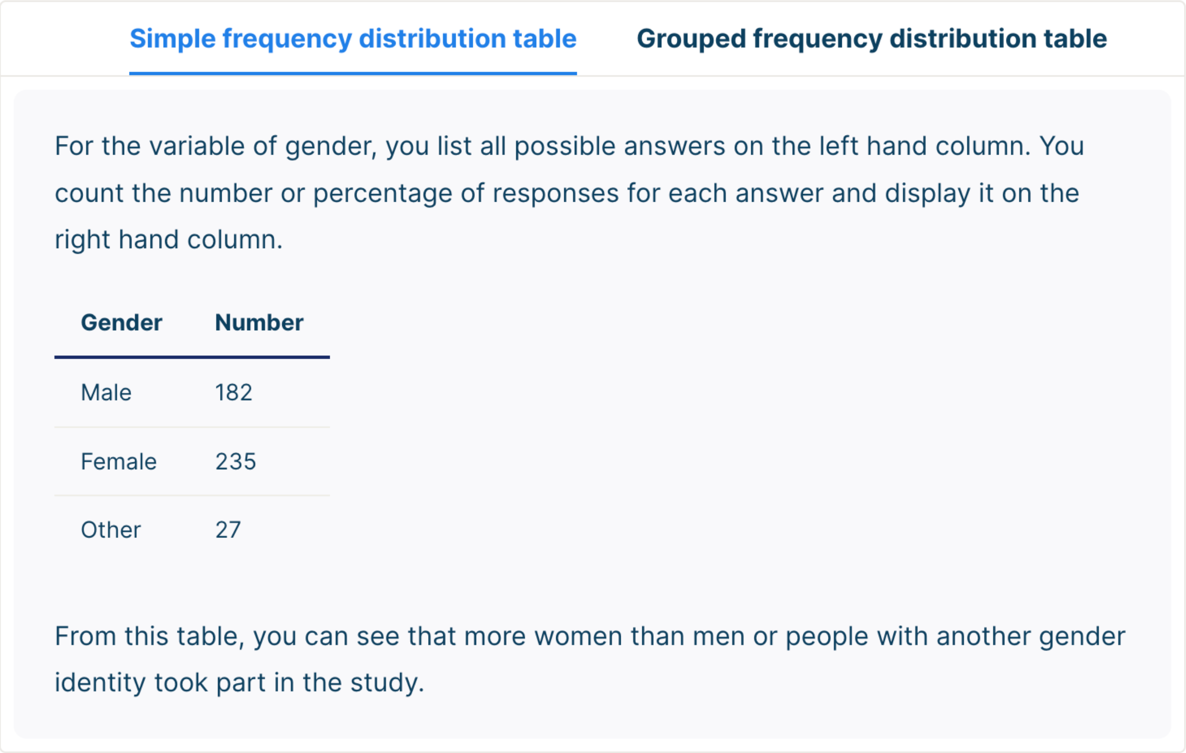
Your data set is the collection of responses to the survey.

Now you can use descriptive statistics to find out the overall frequency of each activity (distribution), the averages for each activity (central tendency), and the spread of responses for each activity (variability).



Frequency Distribution

- A data set is made up of a distribution of values, or scores.
  - In tables or graphs, you can summarize ***the frequency of every possible value of a variable in numbers or percentages***.


Simple frequency distribution table
Grouped frequency distribution table

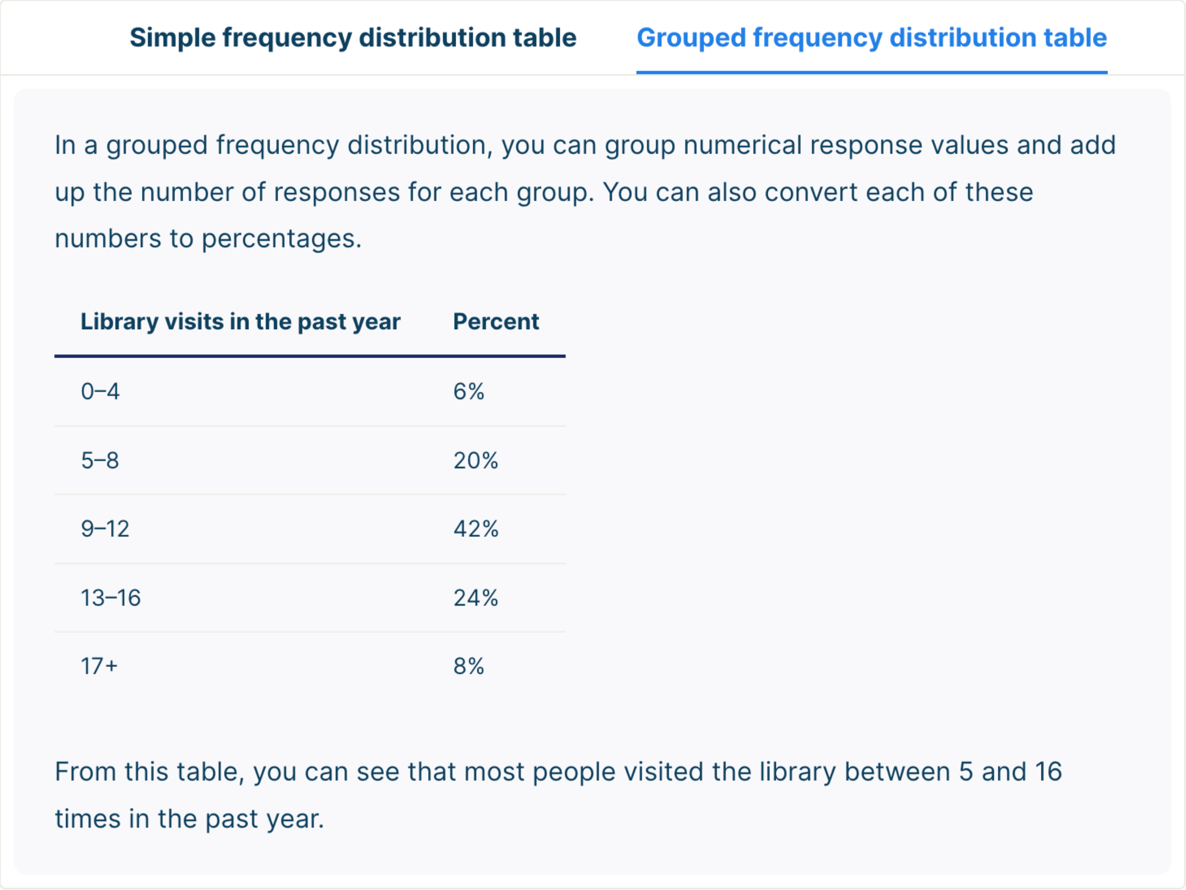
For the variable of gender, you list all possible answers on the left hand column. You count the number or percentage of responses for each answer and display it on the right hand column.

Gender	Number
Male	182
Female	235
Other	27

From this table, you can see that more women than men or people with another gender identity took part in the study.

Frequency Distribution

- A data set is made up of a distribution of values, or scores.
  - In tables or graphs, you can summarize ***the frequency of every possible value of a variable in numbers or percentages***.


Simple frequency distribution table
Grouped frequency distribution table

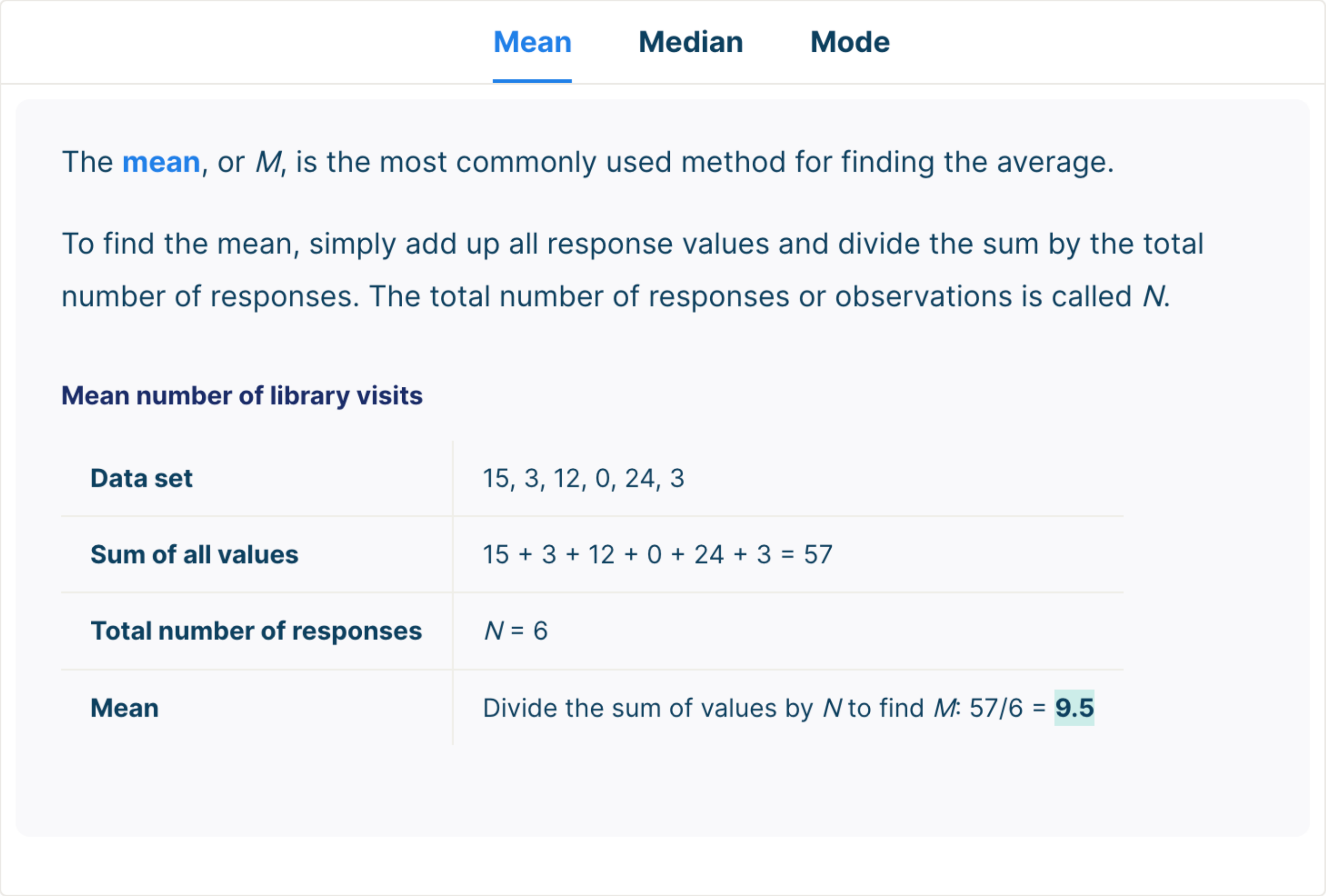
In a grouped frequency distribution, you can group numerical response values and add up the number of responses for each group. You can also convert each of these numbers to percentages.

Library visits in the past year	Percent
0–4	6%
5–8	20%
9–12	42%
13–16	24%
17+	8%

From this table, you can see that most people visited the library between 5 and 16 times in the past year.

Measures of Central Tendency

- Estimate the center, or average, of a data set.
  - The **mean**, median and mode are 3 ways of finding the average.



The **mean**, or *M*, is the most commonly used method for finding the average.

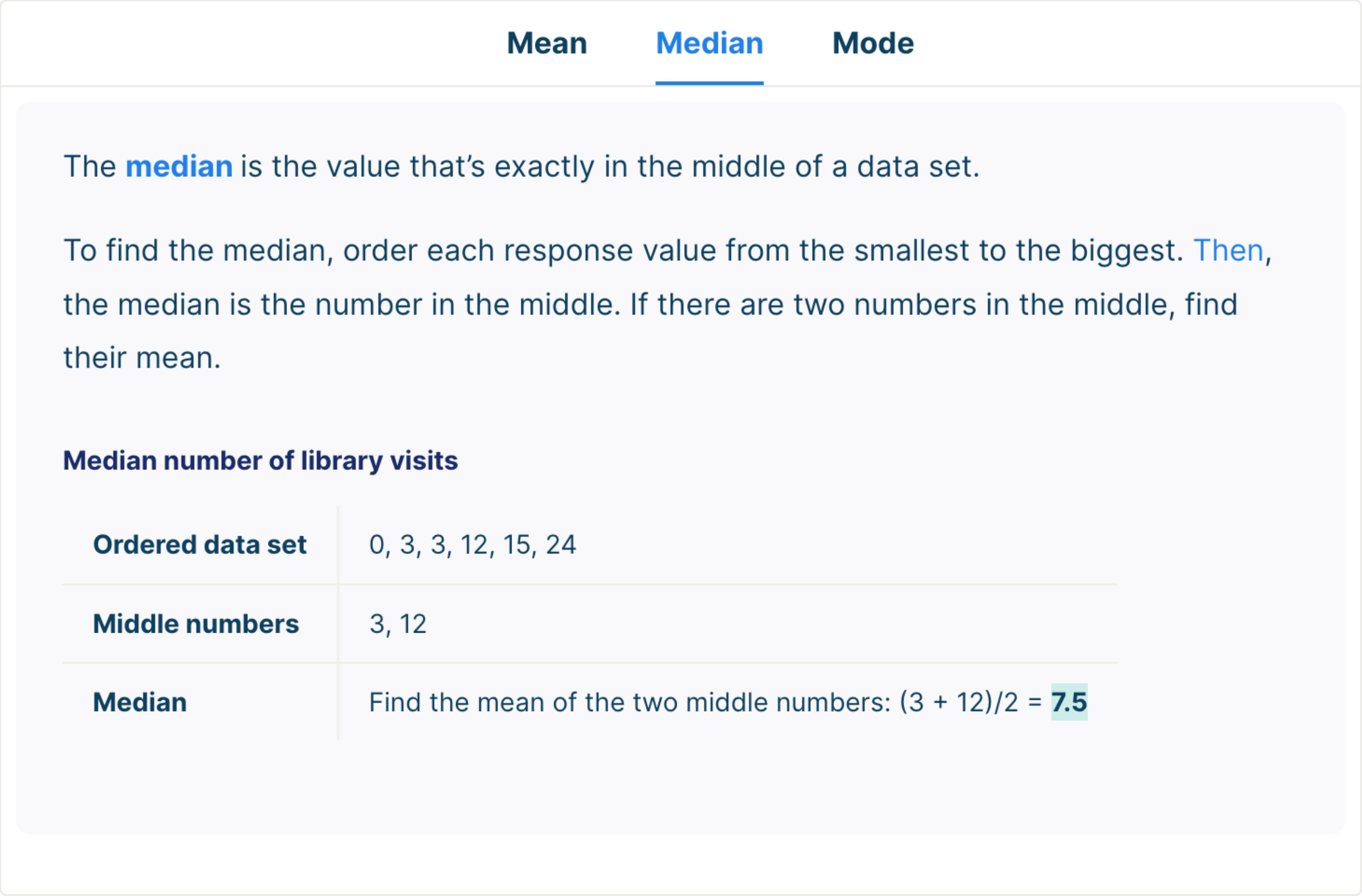
To find the mean, simply add up all response values and divide the sum by the total number of responses. The total number of responses or observations is called *N*.

Mean number of library visits

Data set	15, 3, 12, 0, 24, 3
Sum of all values	15 + 3 + 12 + 0 + 24 + 3 = 57
Total number of responses	*N* = 6
Mean	Divide the sum of values by *N* to find *M*: 57/6 = **9.5**

Measures of Central Tendency

- Estimate the center, or average, of a data set.
  - The mean, **median** and mode are 3 ways of finding the average.



The **median** is the value that's exactly in the middle of a data set.

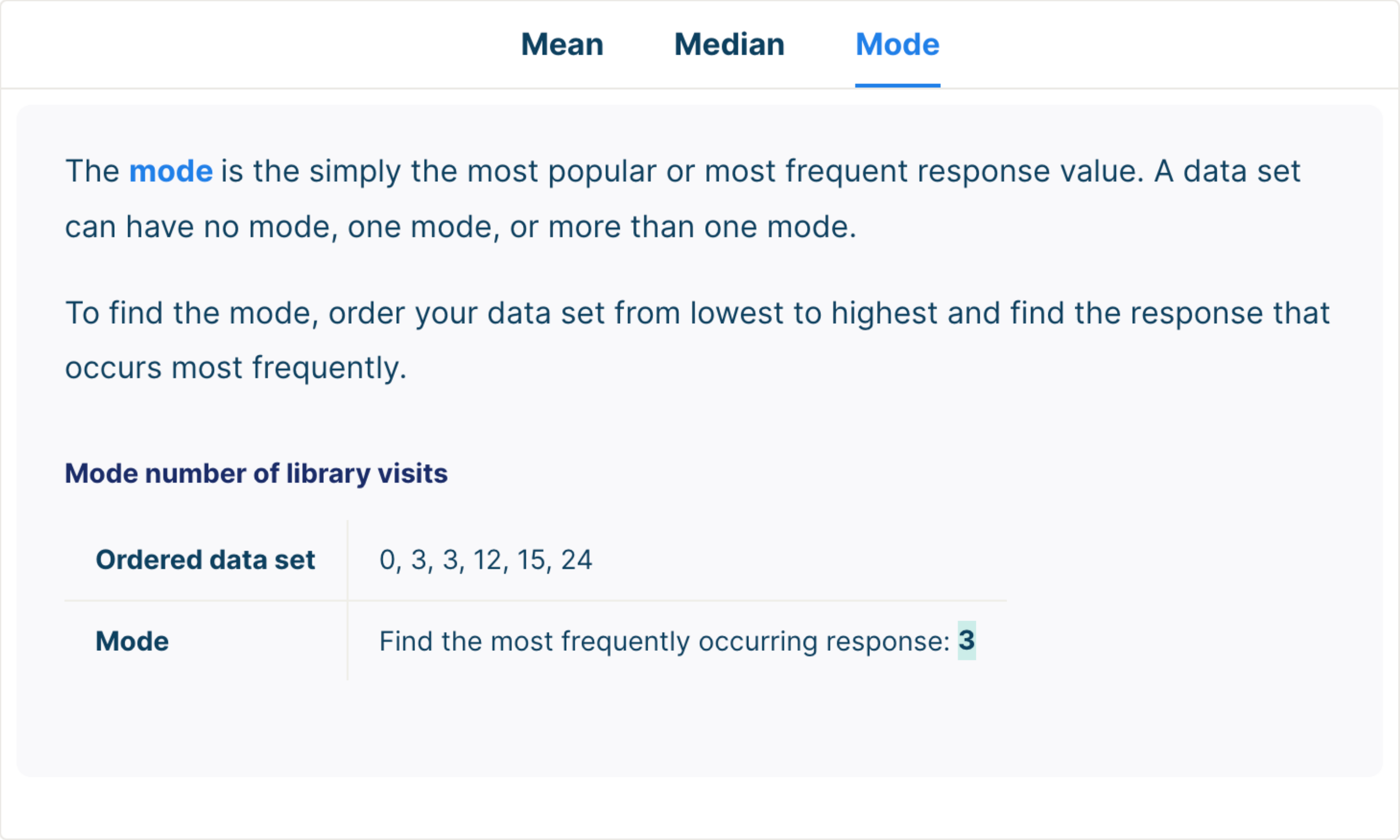
To find the median, order each response value from the smallest to the biggest. Then, the median is the number in the middle. If there are two numbers in the middle, find their mean.

Median number of library visits

Ordered data set	0, 3, 3, 12, 15, 24
Middle numbers	3, 12
Median	Find the mean of the two middle numbers: (3 + 12)/2 = **7.5**

Measures of Central Tendency

- Estimate the center, or average, of a data set.
  - The mean, median and **mode** are 3 ways of finding the average.



The **mode** is the simply the most popular or most frequent response value. A data set can have no mode, one mode, or more than one mode.

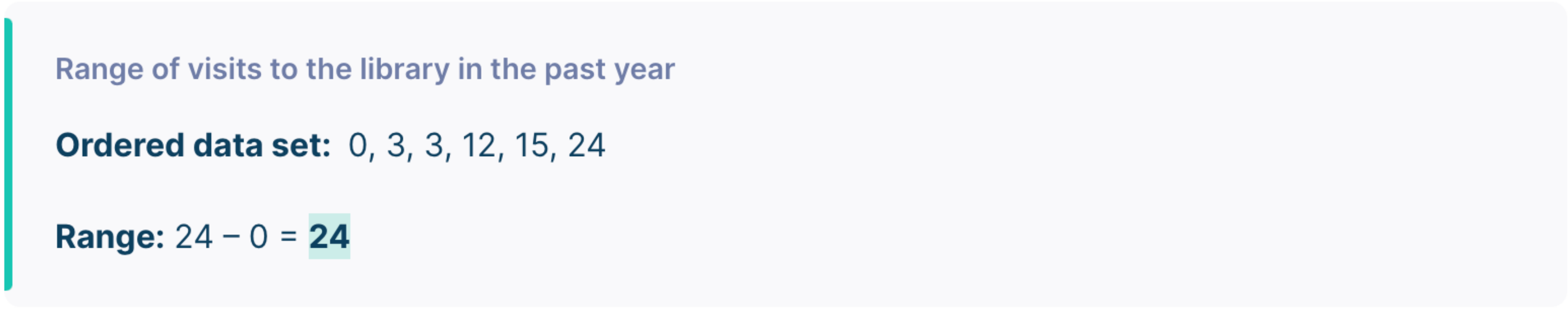
To find the mode, order your data set from lowest to highest and find the response that occurs most frequently.

Mode number of library visits

Ordered data set	0, 3, 3, 12, 15, 24
Mode	Find the most frequently occurring response: **3**

Measures of Variability

- Give you a sense of how spread out the response values are.
  - The range, standard deviation and variance each reflect different aspects of spread.
- **Range**
  - The range gives you an idea of how far apart the most extreme response scores are.
  - To find the range, simply subtract the lowest value from the highest value.



Range of visits to the library in the past year

Ordered data set: 0, 3, 3, 12, 15, 24

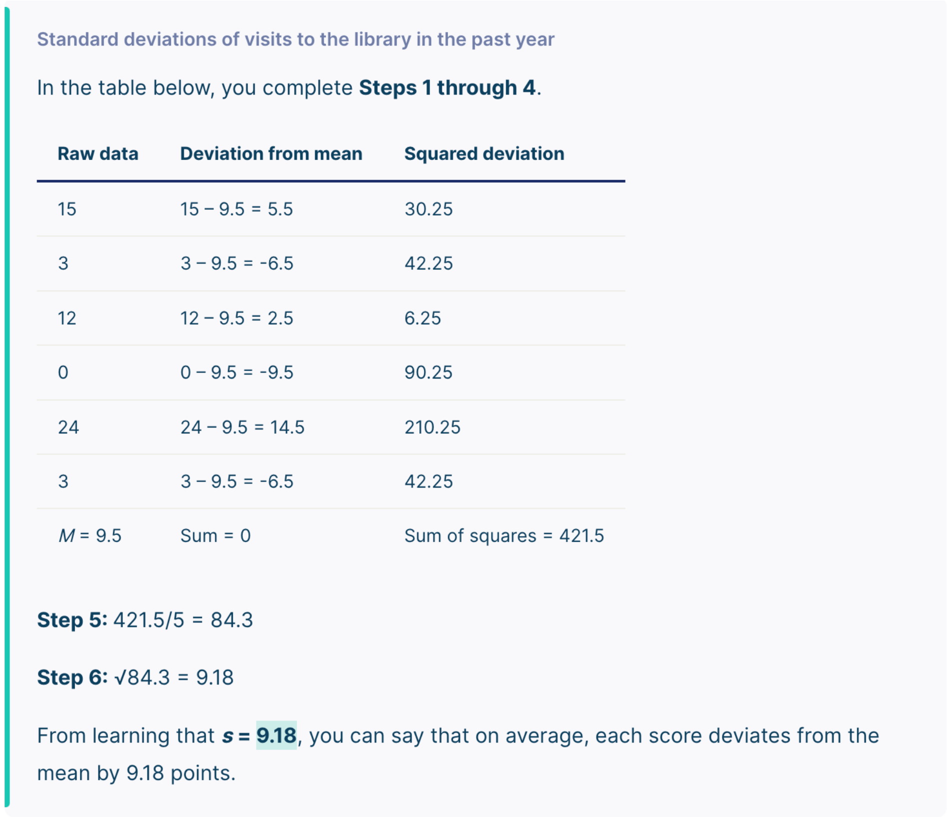
Range: 24 − 0 = **24**

Measures of Variability

- **Standard deviation**
  - The standard deviation (*s* or *SD*) is the average amount of variability in your dataset.
    - It tells you, on average, how far each score lies from the mean.
    - The larger the standard deviation, the more variable the data set is.

Six steps for finding the standard deviation:

1. *List each score and find their mean.*
2. *Subtract the mean from each score to get the deviation from the mean.*
3. *Square each of these deviations.*
4. *Add up all of the squared deviations.*
5. ***Divide the sum of the squared deviations by N – 1.***
6. ***Find the square root of the number you found.***



Standard deviations of visits to the library in the past year

In the table below, you complete **Steps 1 through 4**.

Raw data	Deviation from mean	Squared deviation
15	15 – 9.5 = 5.5	30.25
3	3 – 9.5 = -6.5	42.25
12	12 – 9.5 = 2.5	6.25
0	0 – 9.5 = -9.5	90.25
24	24 – 9.5 = 14.5	210.25
3	3 – 9.5 = -6.5	42.25
M = 9.5	Sum = 0	Sum of squares = 421.5

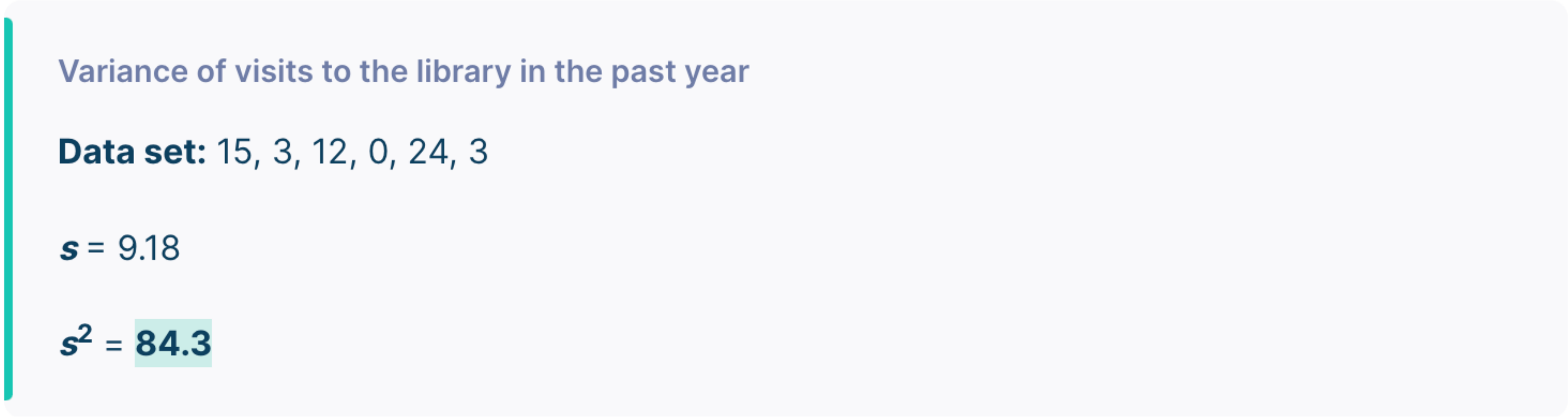
Step 5: 421.5/5 = 84.3

Step 6: √84.3 = 9.18

From learning that ***s* = 9.18**, you can say that on average, each score deviates from the mean by 9.18 points.

Variance

- The average of squared deviations from the mean.
  - Variance reflects the degree of spread in the data set.
  - The more spread the data, the larger the variance is in relation to the mean.
  - To find the variance, simply square the standard deviation.
  - The symbol for variance is s^2.



Variance of visits to the library in the past year

Data set: 15, 3, 12, 0, 24, 3

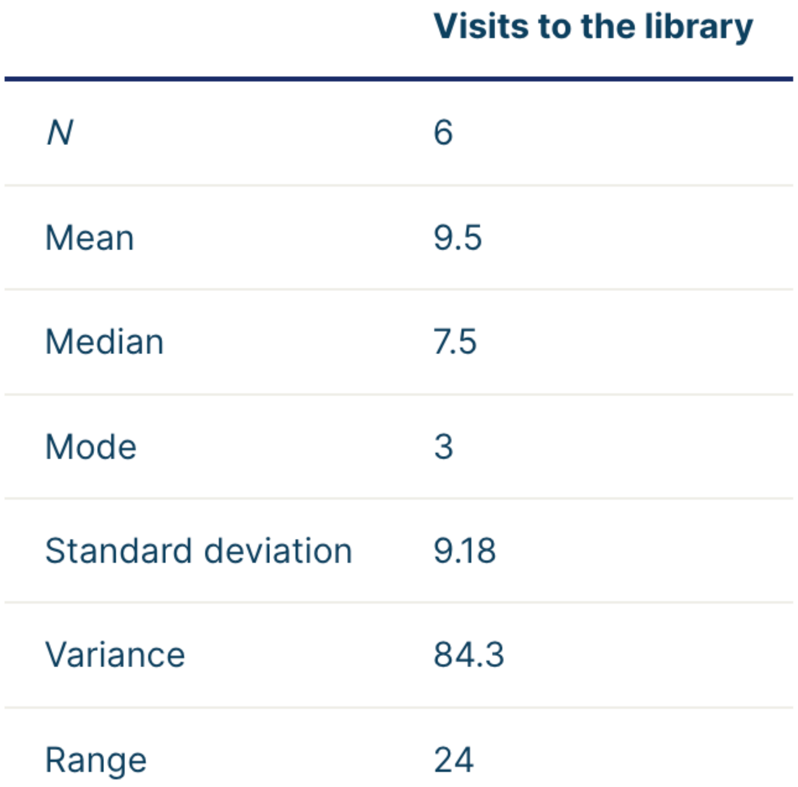
s = 9.18

s^2 = **84.3**

Univariate Descriptive Statistics

- Focus on only one variable at a time.
  - It’s important to examine data from each variable separately using multiple measures of distribution, central tendency and spread.
  - Programs like SPSS and Excel can be used to easily calculate these.

  - *If you were to only consider the mean as a measure of central tendency, your impression of the “middle” of the data set can be skewed by outliers, unlike the median or mode.*
  - *Likewise, while the range is sensitive to outliers, you should also consider the standard deviation and variance to get easily comparable measures of spread.*



	Visits to the library
N	6
Mean	9.5
Median	7.5
Mode	3
Standard deviation	9.18
Variance	84.3
Range	24

Inferential Statistics

- While descriptive statistics summarize the characteristics of a data set, **inferential statistics** help you *come to conclusions and make predictions based on your data*.

- When you have collected data from a sample, you can *use* **inferential statistics** *to understand the larger population from which the sample is taken*.

- Inferential statistics have two main uses:
  - **Making estimates** about populations
    - (for example, the mean SAT score of all 11th graders in the US).
  - **Testing hypotheses** to draw conclusions about populations
    - (for example, the relationship between SAT scores and family income).

Descriptive versus Inferential Statistics

to *describe* vs. to *make inferences*

Descriptive statistics

- Using descriptive statistics, you can report characteristics of your data:
  - The **distribution** concerns the frequency of each value.
  - The **central tendency** concerns the averages of the values.
  - The **variability** concerns how spread out the values are.
- In descriptive statistics, there is no uncertainty – the statistics precisely describe the data that you collected.
  - If you collect data from an entire population, you can directly compare these descriptive statistics to those from other populations.

Example: Descriptive statistics

- You collect data on the SAT scores of all 11th graders in a school for three years.
- You can use descriptive statistics to get a quick overview of the school's scores in those years. You can then directly compare the mean SAT score with the mean scores of other schools.

Inferential statistics

- Most of the time, you can only acquire data from samples, because it is too difficult or expensive to collect data from the whole population that you're interested in.

- Inferential statistics **use your sample to make reasonable guesses about the larger population.**

- With inferential statistics, it's important to **use random and unbiased sampling methods**.
  - If your sample isn't representative of your population, then you can't make valid statistical inferences or generalize.

Example: Inferential statistics

- You randomly select a sample of 11th graders in your state and collect data on their SAT scores and other characteristics.
- You can use inferential statistics to make estimates and test hypotheses about the whole population of 11th graders in the state based on your sample data.

Hypothesis Testing

- Hypothesis testing is ***a formal process of statistical analysis using inferential statistics***.
  - The goal is to compare populations or assess relationships between variables using samples.

- Statistical tests can be **parametric** or **non-parametric**.
  - Parametric tests are considered more statistically powerful because they are more likely to detect an effect if one exists.

- **Parametric tests** make assumptions that include the following:
  - the population that the sample comes from follows a normal distribution of scores
  - the sample size is large enough to represent the population
  - the variances, a measure of variability, of each group being compared are similar

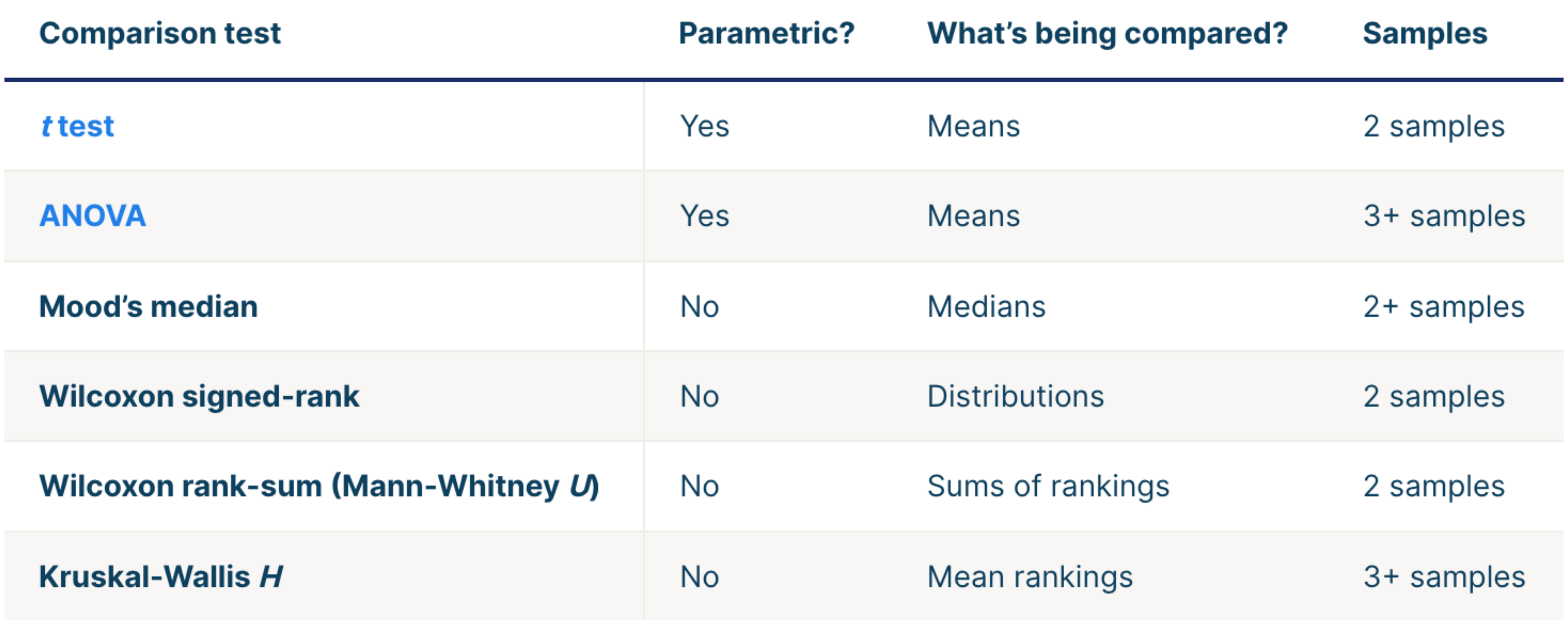
- When your data violates any of these assumptions, **non-parametric tests** are more suitable.
  - Non-parametric tests are called "distribution-free tests" because they don't assume anything about the distribution of the population data.

- Statistical tests come in three forms: *tests of comparison*, *correlation* or *regression*.

Comparison Tests

- Comparison tests assess whether there are differences in means, medians or rankings of scores of two or more groups.
  - To decide which test suits your aim, consider whether your data meets the conditions necessary for parametric tests, the number of samples, and the levels of measurement of your variables.

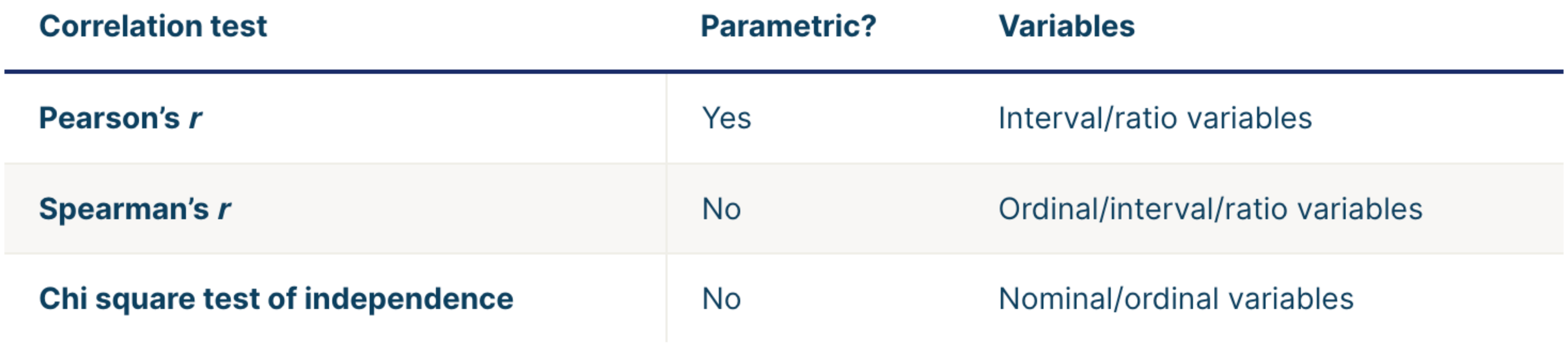
$$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}} \qquad t = \frac{(\bar{x}_1 - \bar{x}_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$



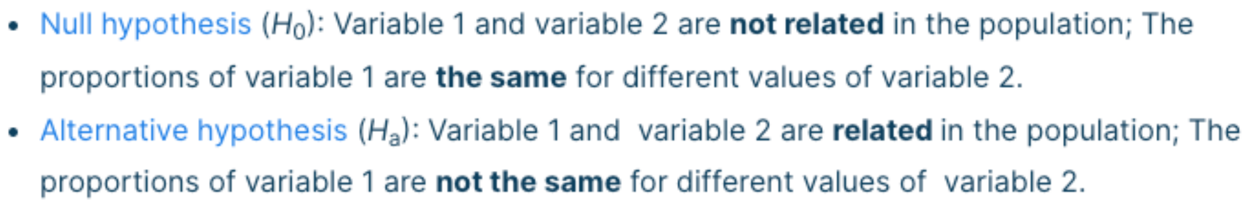
Comparison test	Parametric?	What's being compared?	Samples
t test	Yes	Means	2 samples
ANOVA	Yes	Means	3+ samples
Mood's median	No	Medians	2+ samples
Wilcoxon signed-rank	No	Distributions	2 samples
Wilcoxon rank-sum (Mann-Whitney *U*)	No	Sums of rankings	2 samples
Kruskal-Wallis *H*	No	Mean rankings	3+ samples

Correlation Tests

- Correlation tests determine the extent to which two variables are associated.
  - Although **Pearson's *r*** is the most statistically powerful test, Spearman's r is appropriate for interval and ratio variables when the data doesn't follow a normal distribution.
  - The chi square test of independence is the only test that can be used with nominal variables.



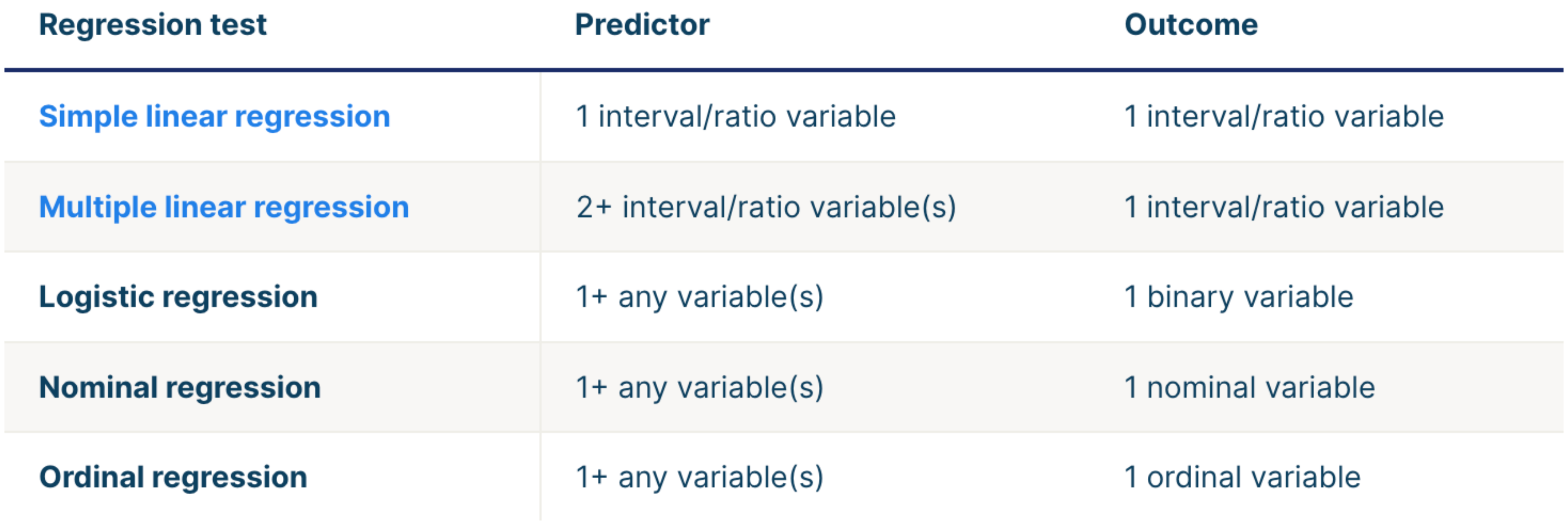
Correlation test	Parametric?	Variables
Pearson's *r*	Yes	Interval/ratio variables
Spearman's *r*	No	Ordinal/interval/ratio variables
Chi square test of independence	No	Nominal/ordinal variables



- Null hypothesis (H_0): Variable 1 and variable 2 are **not related** in the population; The proportions of variable 1 are **the same** for different values of variable 2.
- Alternative hypothesis (H_a): Variable 1 and variable 2 are **related** in the population; The proportions of variable 1 are **not the same** for different values of variable 2.

Regression Tests

- Regression tests demonstrate whether changes in predictor variables cause changes in an outcome variable.
- Most of the commonly used regression tests are parametric.
  - If your data is not normally distributed, you can perform data transformations.



Regression test	Predictor	Outcome
Simple linear regression	1 interval/ratio variable	1 interval/ratio variable
Multiple linear regression	2+ interval/ratio variable(s)	1 interval/ratio variable
Logistic regression	1+ any variable(s)	1 binary variable
Nominal regression	1+ any variable(s)	1 nominal variable
Ordinal regression	1+ any variable(s)	1 ordinal variable

Exploratory vs. Explanatory Data Analysis

Wan Fang

Southern University of Science and Technology

Exploratory vs. Explanatory Analysis

🧭 Exploratory data analysis is …

- ... the **"herding cats"** 😸 stage of working with data. It is a chaotic, often solitary, exercise requiring persistence in search of insights.
- ... finding **what matters in the data** by connecting data sources, determining relationships within the data, and understanding what measures and dimensions are most important.
- ... the **starting point** for working with data. Without exploring and understanding your data, you cannot move on to explaining it to others.

📖 Explanatory data presentation is …

- ... the **"herding cows"** 🐮 stage of working with data. It is about pointing the data in a direction that will connect with other people.
- ... gathering insights and building a logical structure for **communicating those insights**. It requires understanding the audience for the data and knowing what matters to them in their role.
- ... the **"last mile"** of working with data to encourage people to take action on the insights.

Types of exploratory data analysis

Univariate non-graphical

- This is simplest form of data analysis, where the data being analyzed consists of just one variable. The main purpose of univariate analysis is to describe the data and find patterns that exist within it.

Multivariate nongraphical

- Multivariate data arises from more than one variable. Multivariate non-graphical EDA techniques generally show the relationship between two or more variables of the data through cross-tabulation or statistics.

Univariate graphical

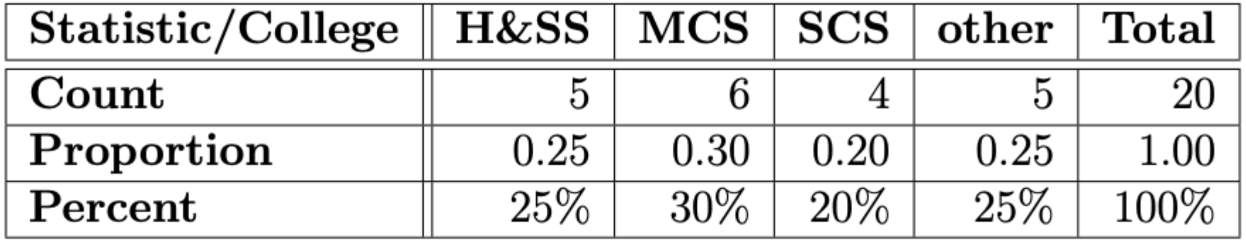
- Graphical methods provide a full picture of the data.

Multivariate graphical

- Multivariate data uses graphics to display relationships between two or more sets of data.

Univariate non-graphical

- **Frequency for categorical data**

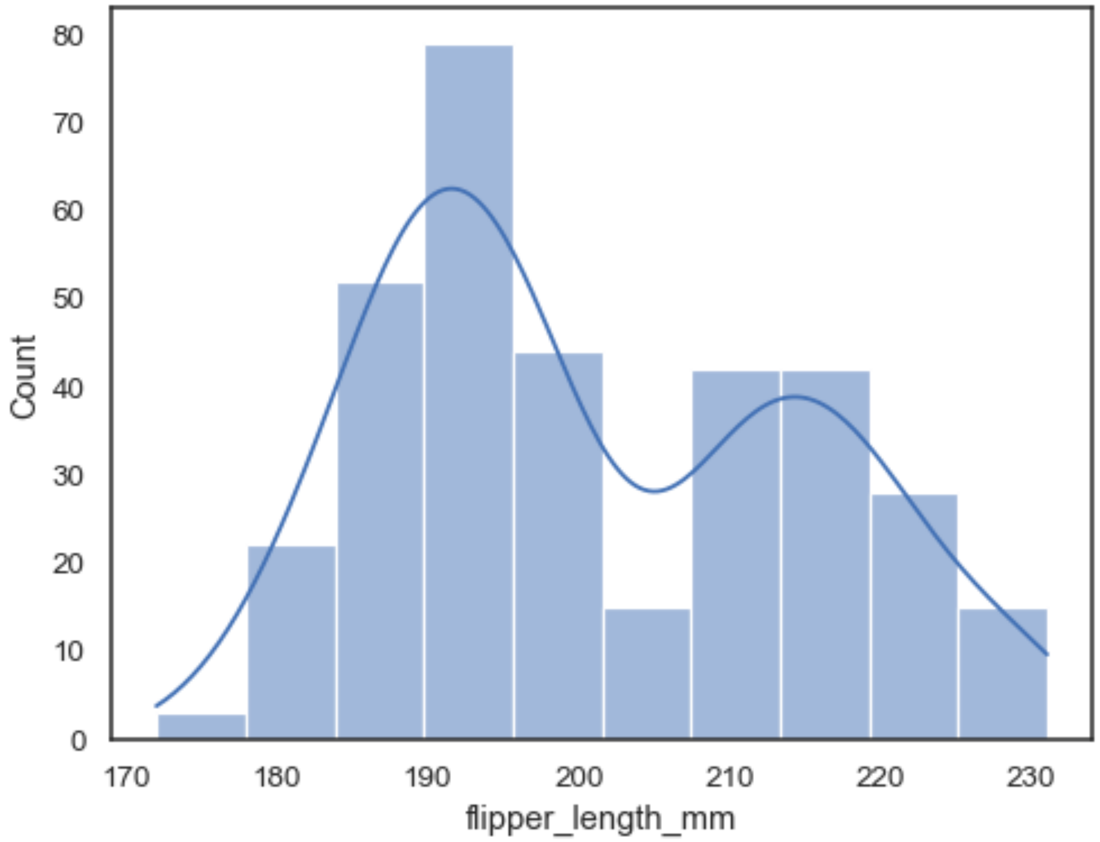


Statistic/College	H&SS	MCS	SCS	other	Total
Count	5	6	4	5	20
Proportion	0.25	0.30	0.20	0.25	1.00
Percent	25%	30%	20%	25%	100%

- **Central Tendency**
  - The three generally estimated are mean, median, and mode.
- **Range**
  - The range is the difference between the maximum and minimum value in the data.
- **Variance and Standard Deviation**
  - indicates the spread of all data points in a data set.
- **Skewness, Outliers**

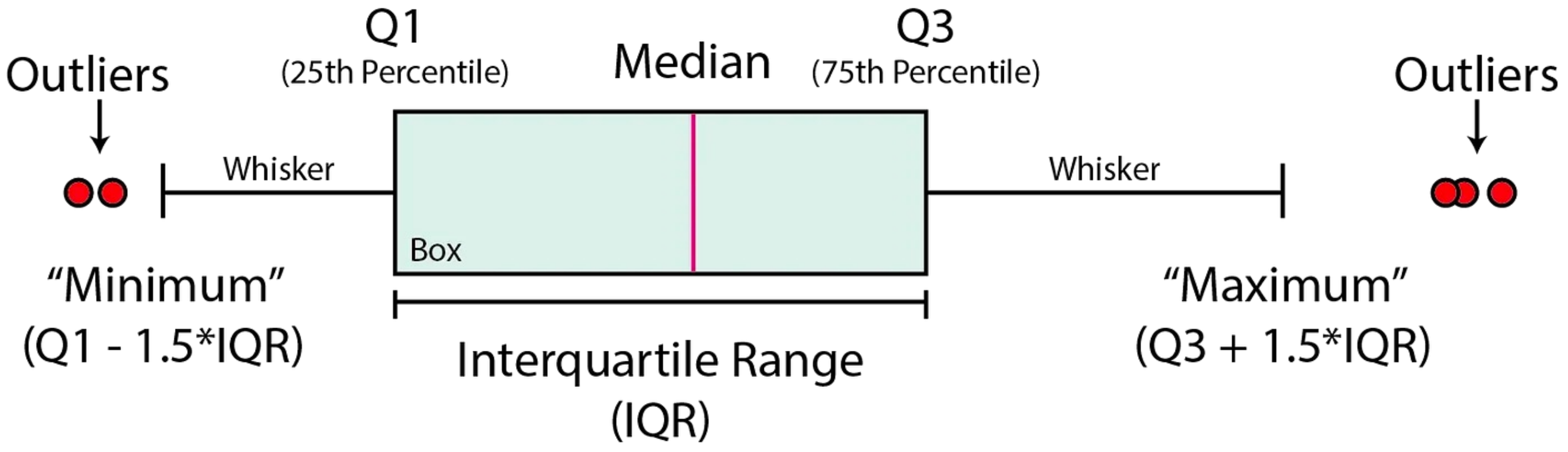
Univariate Graphical

- **Histograms**
  - A bar plot in which each bar represents the frequency (count) or proportion (count/total count) of cases for a range of values.



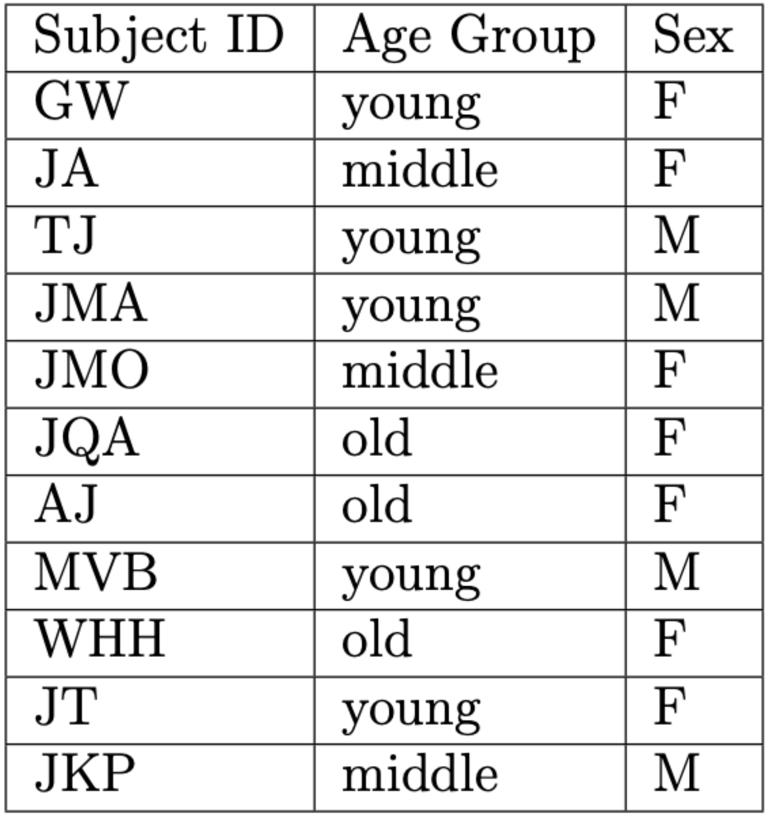
Univariate Graphical

- **Box plots**
  - graphically depict the five-number summary of minimum, first quartile, median, third quartile, and maximum.



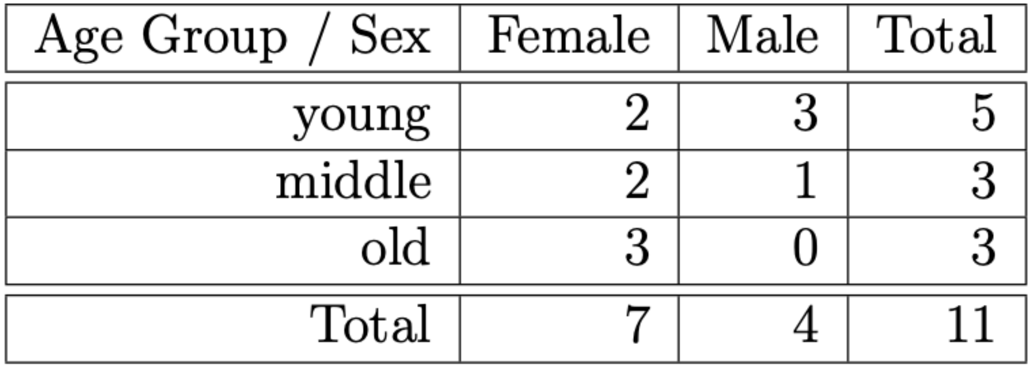
Multivariate nongraphical

- **Cross-tabulation**
  - The basic bivariate non-graphical EDA technique



Subject ID	Age Group	Sex
GW	young	F
JA	middle	F
TJ	young	M
JMA	young	M
JMO	middle	F
JQA	old	F
AJ	old	F
MVB	young	M
WHH	old	F
JT	young	F
JKP	middle	M

Table 4.1: Sample Data for Cross-tabulation



Age Group / Sex	Female	Male	Total
young	2	3	5
middle	2	1	3
old	3	0	3
Total	7	4	11

Table 4.2: Cross-tabulation of Sample Data

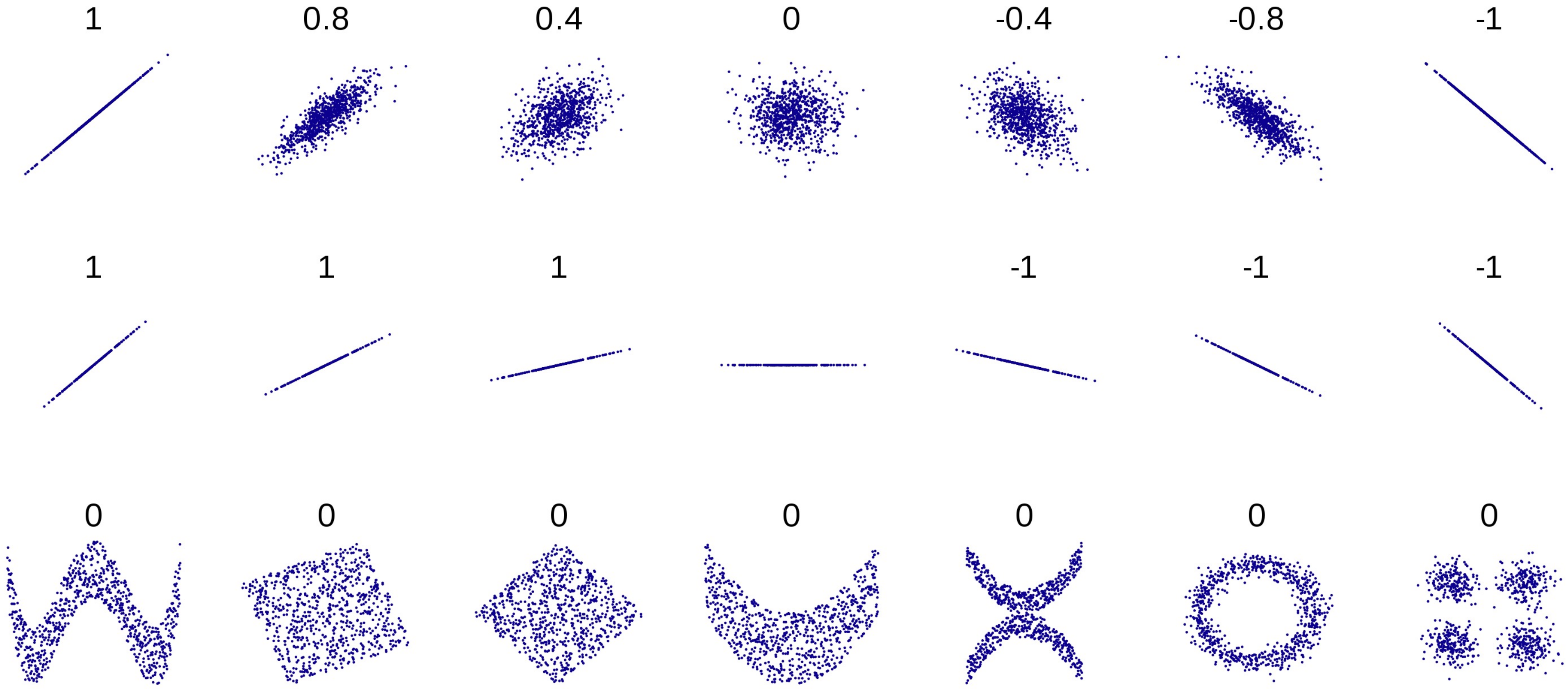
Multivariate nongraphical

- **Correlation coefficient**
  - The correlation between two random variables is a number that runs from -1 through 0 to +1 and indicates a strong inverse relationship, no relationship, and a strong direct relationship, respectively.

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

Multivariate nongraphical

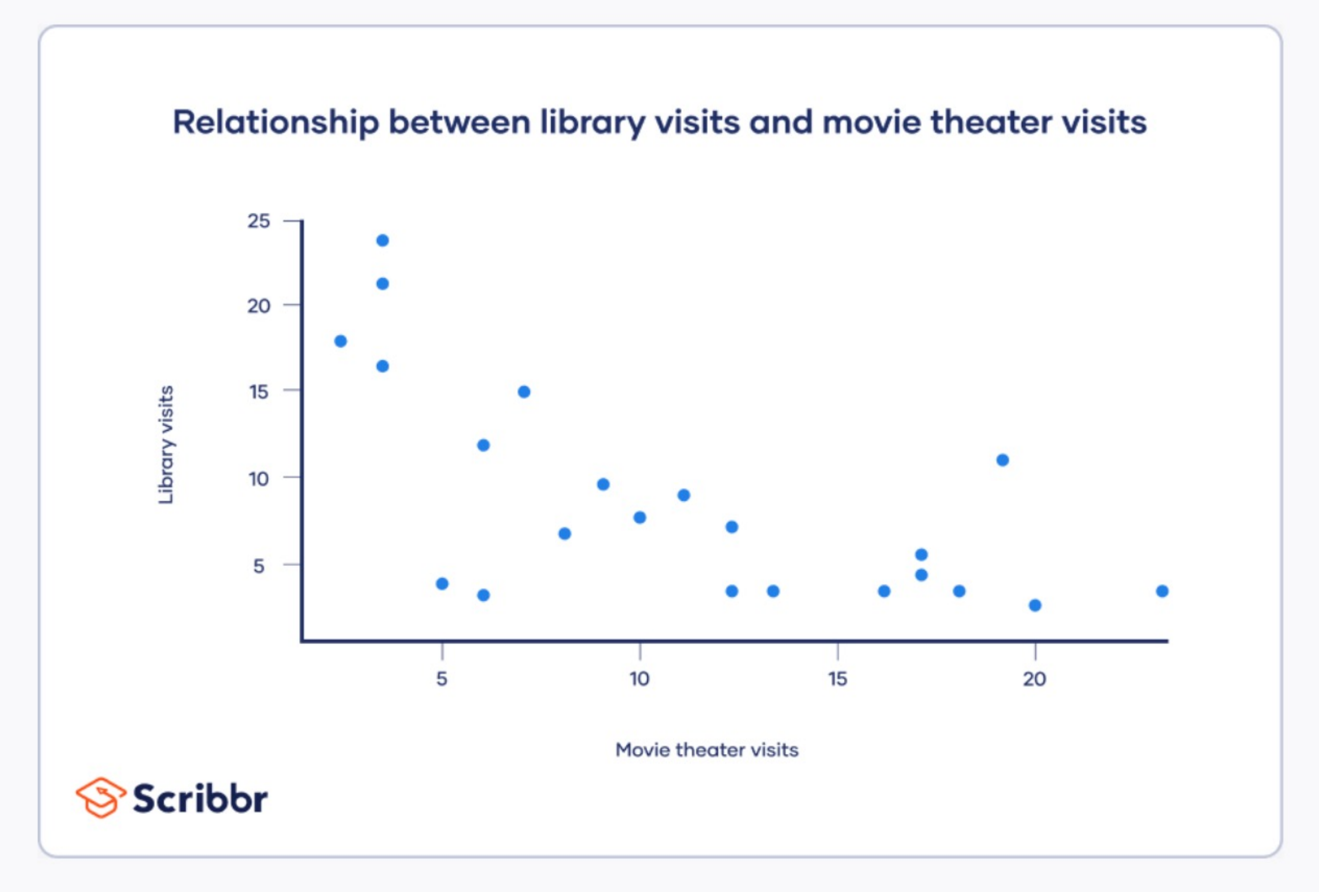
- **Correlation coefficient**



https://en.wikipedia.org/wiki/Pearson_correlation_coefficient

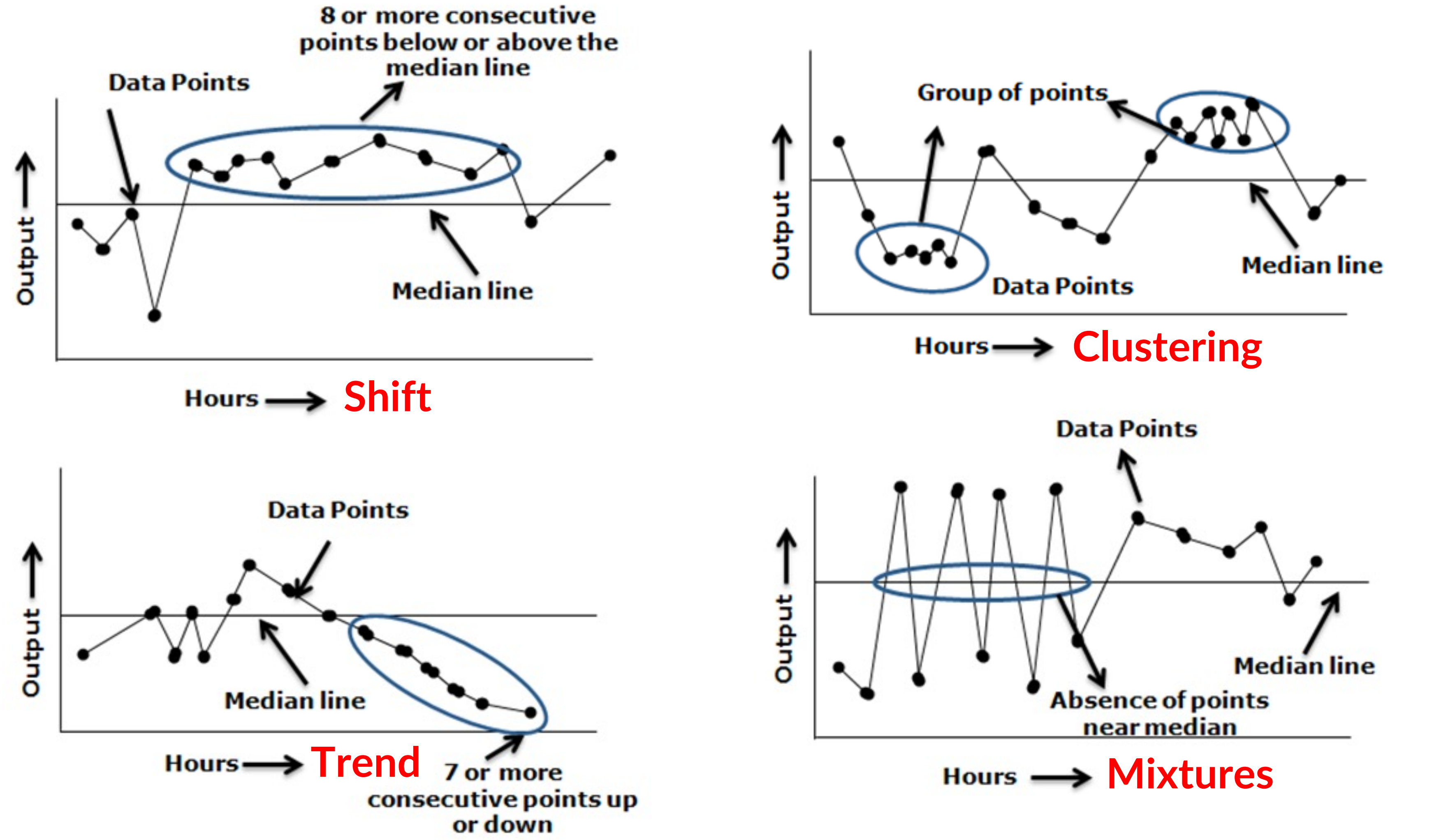
Multivariate Graphical

- **Scatter plot**, plot data points on a horizontal and a vertical axis to show how much one variable is affected by another.



Multivariate Graphical

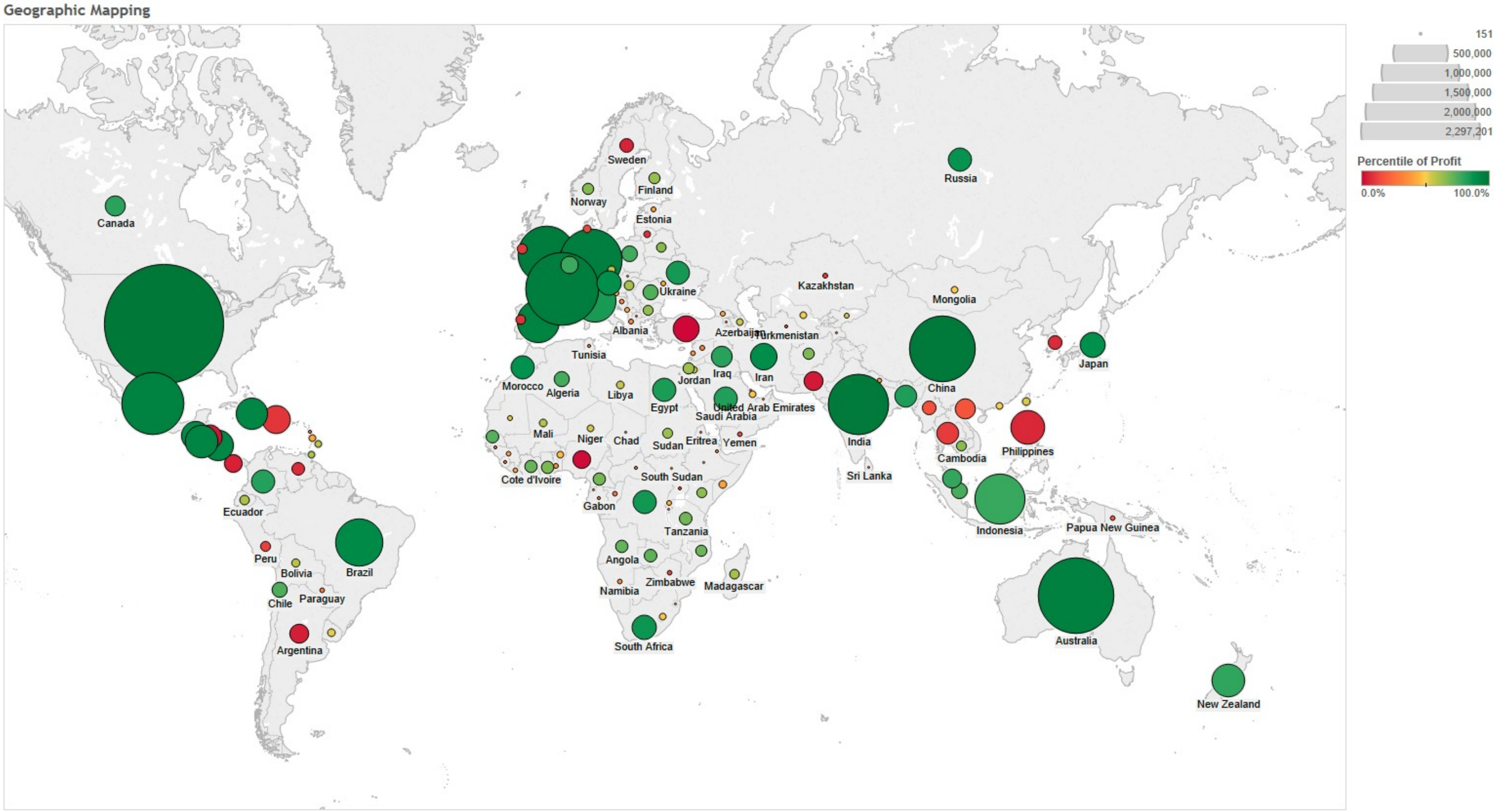
- **Run chart**, which is a line graph of data plotted over time.



https://sixsigmastudyguide.com/run-chart/

Multivariate Graphical

- **Bubble chart**, which is a data visualization that displays multiple circles (bubbles) in a two-dimensional plot.



https://community.tableau.com/s/idea/0874T000000HAgfQAG/detail

Multivariate Graphical

- **Heat map**, which is a graphical representation of data where values are depicted by color.

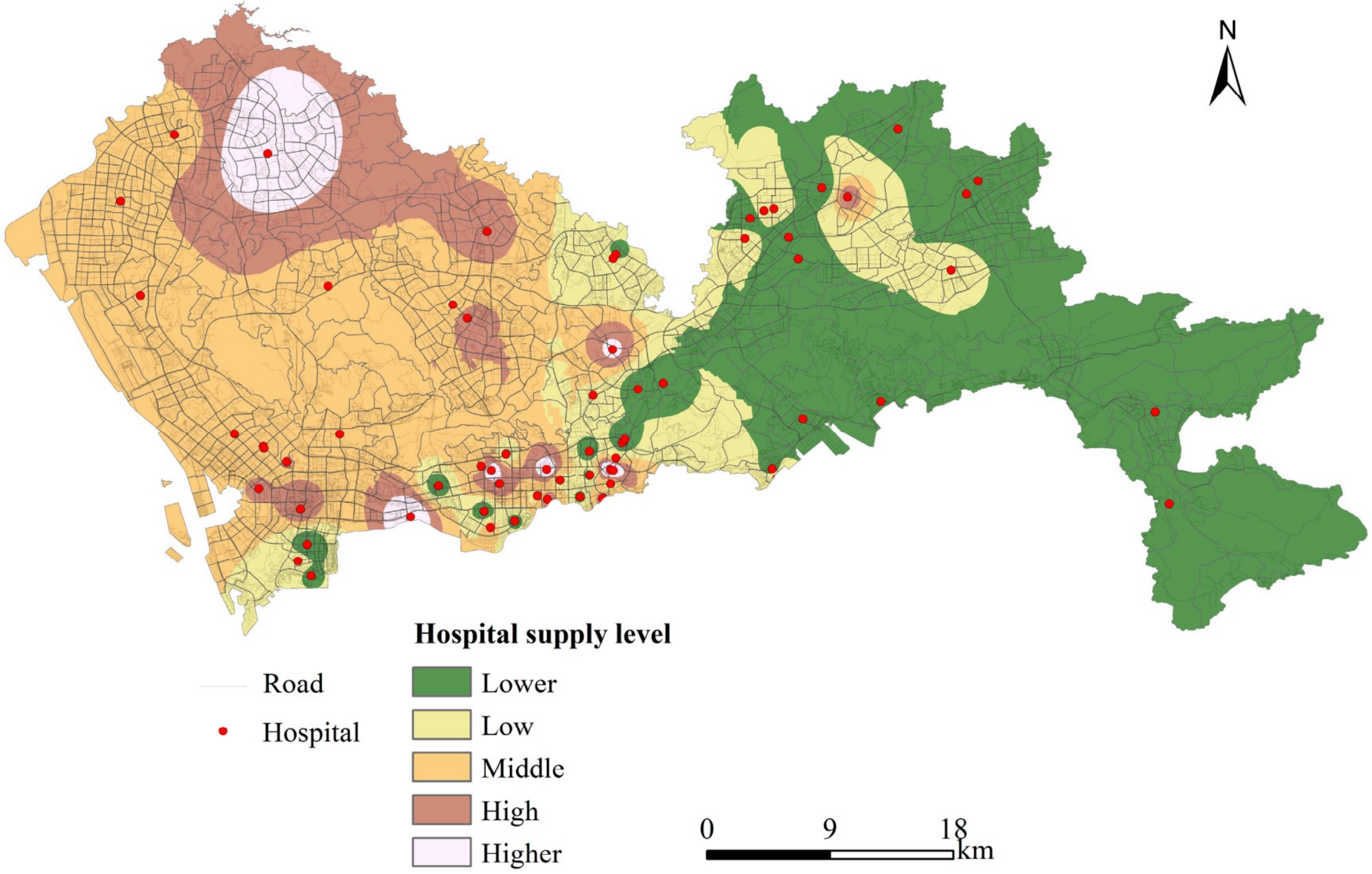


Figure 4. Spatial distribution of road network and hospital supply level in Shenzhen.

Identifying Spatial Matching between the Supply and Demand of Medical Resource and Accessing Carrying Capacity: A Case Study of Shenzhen, China

Summary of EDA

- You should always perform appropriate EDA before further analysis of your data.
- Perform whatever steps are necessary to become more familiar with your data,
  - check for obvious mistakes,
  - learn about variable distributions, and
  - learn about relationships between variables.
- EDA is not an exact science – it is a very important art!

Example

Source

- Data related white variants of the Portuguese "Vinho Verde" wine.
- Input variables (based on physicochemical tests):

  1 - fixed acidity
  2 - volatile acidity
  3 - citric acid
  4 - residual sugar
  5 - chlorides
  6 - free sulfur dioxide
  7 - total sulfur dioxide
  8 - density
  9 - pH
  10 - sulphates
  11 - alcohol

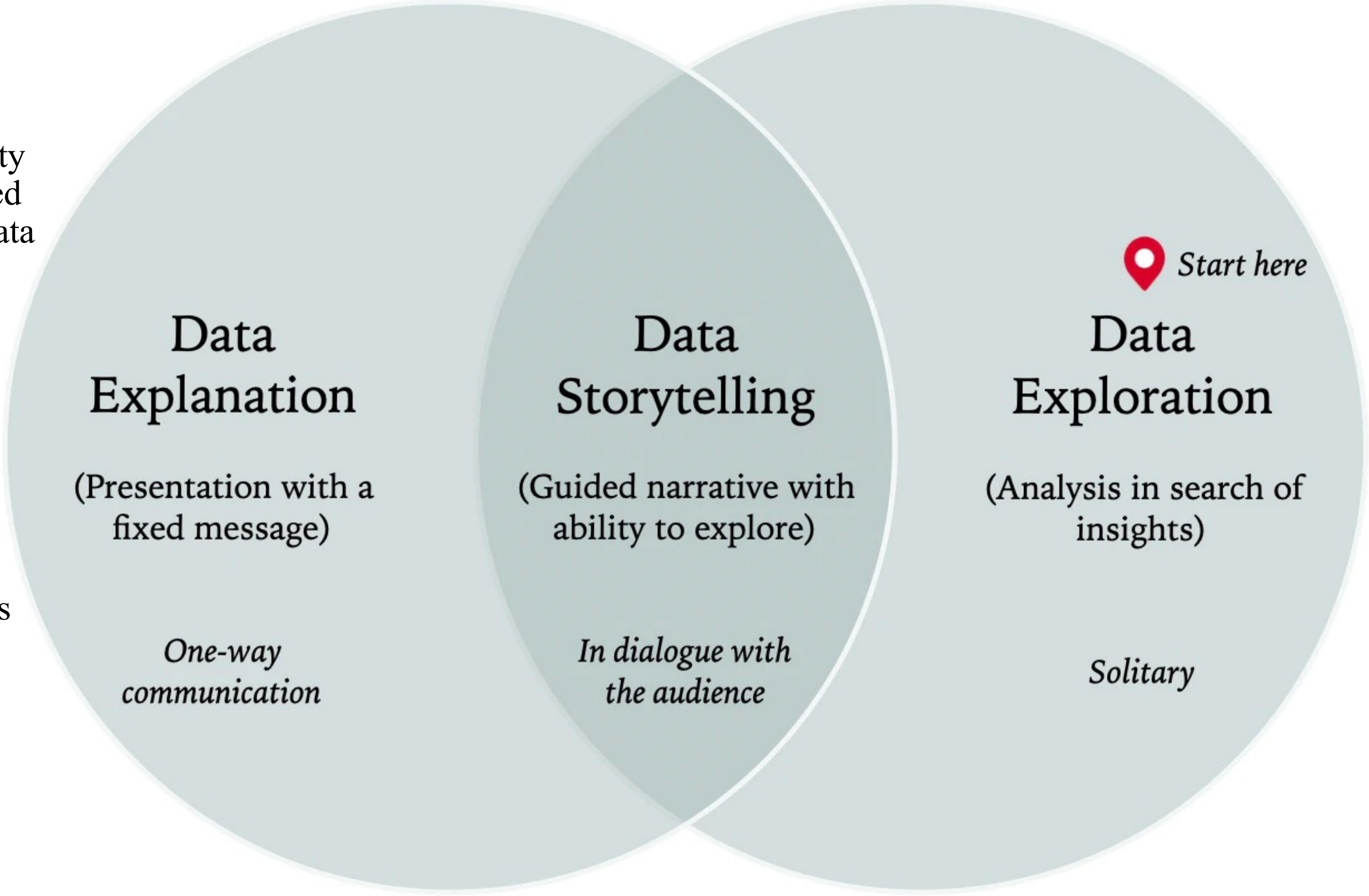
- Output variable (based on sensory data):
  - quality (score between 0 and 10)

Can we combine exploratory and explanatory?

- Sure. There is a middle ground that combines data explanation and data exploration. We can call it **interactive data storytelling**.

- At this intersection, there is an opportunity to combine the guided narrative nature of data explanation with the ability to find new insights through exploration.

- Some examples of these three categories on the right.


Start here
Data Explanation
(Presentation with a fixed message)
One-way communication
Data Storytelling
(Guided narrative with ability to explore)
In dialogue with the audience
Data Exploration
(Analysis in search of insights)
Solitary

Thank you~

Wan Fang
Southern University of Science and Technology