جامعة بيرزيت
BIRZEIT UNIVERSITY

Data Science and Analytics
Comp 4381

Statistics For Data Science

References

- **Books**:
  - Python for Data Analysis 3rd edition - Wes McKinney – O'RIELLY (Ch 2-10)
  - Python data science handbook 2nd edition - Jake VanderPlas – O'RIELLY (Ch 37-40)
  - Statistics unplugged 4th edition – Sally Cardwell - Wadsworth: (Ch 1, 2)
- **Material & Notebooks**:
  - Mr. Hussein Soboh.
- **Additional Resources:**
  - Computational and Inferential Thinking: The Foundations of Data Science 2nd Edition by Ani Adhikari, John DeNero, David Wagner. Link
  - https://www.w3schools.com/python

Foundational Statistics

Understanding Core Statistical Principles

Understanding core statistical principles enables data scientists to interpret data more accurately, identify trends, and make predictions. We will dive into the following topics:

- Central Tendency Measures
- Measures of Spread
- Data Distribution
- Correlation Analysis
- Sampling Techniques

Dataset Context

- The dataset we will use in this section is the ***World Happiness Score for 2023***
- World Happiness Report dataset for 2023 offers a comprehensive examination of happiness metrics and the factors influencing well-being on a global scale
- This dataset is designed to provide valuable insights for
  - Policymakers
  - Researchers
  - Individuals interested in understanding the dynamics of happiness and well-being worldwide

Dataset Context

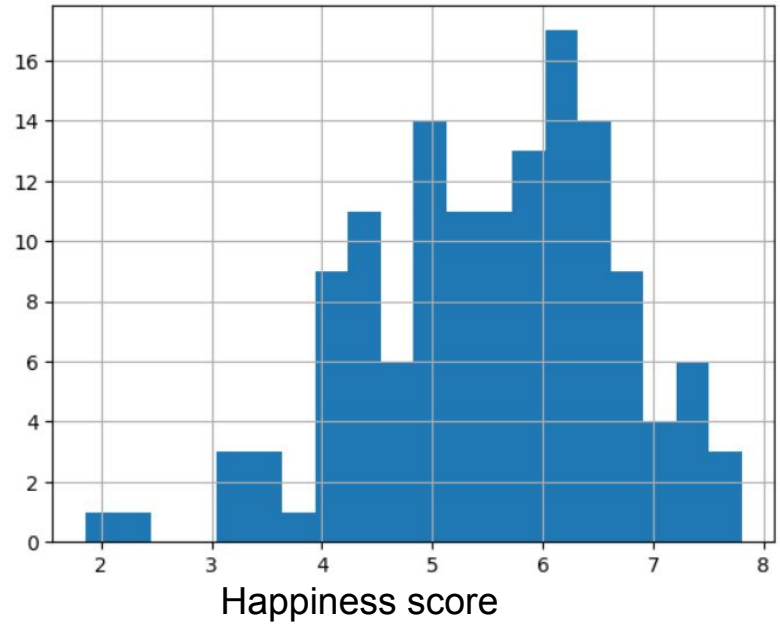
- ***World Happiness Score for 2023***

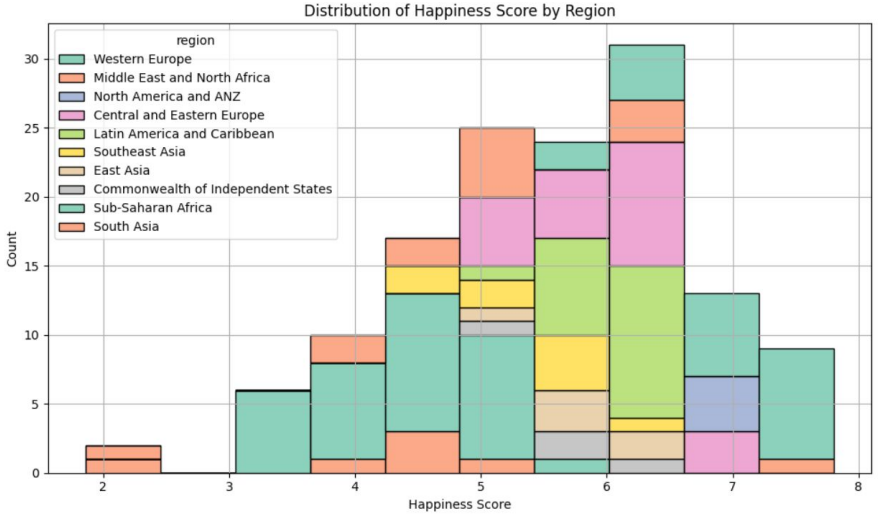
	country	region	happiness_score	gdp_per_capita	social_support	healthy_life_expectancy	freedom_to_make_life_choices	generosity	perceptions_of_corruption
0	Finland	Western Europe	7.804	1.888	1.585	0.535	0.772	0.126	0.535
1	Denmark	Western Europe	7.586	1.949	1.548	0.537	0.734	0.208	0.525
2	Iceland	Western Europe	7.530	1.926	1.620	0.559	0.738	0.250	0.187
3	Israel	Middle East and North Africa	7.473	1.833	1.521	0.577	0.569	0.124	0.158
4	Netherlands	Western Europe	7.403	1.942	1.488	0.545	0.672	0.251	0.394

Dataset Exploration

- **Let's start with a simple question:** Describe the distribution of happiness scores in the world
- We can plot a **histogram** that can show the distribution of a Pandas column.

```
df['happiness_score'].hist(bins=20)
```





How can we accurately describe its distribution using statistical measures?

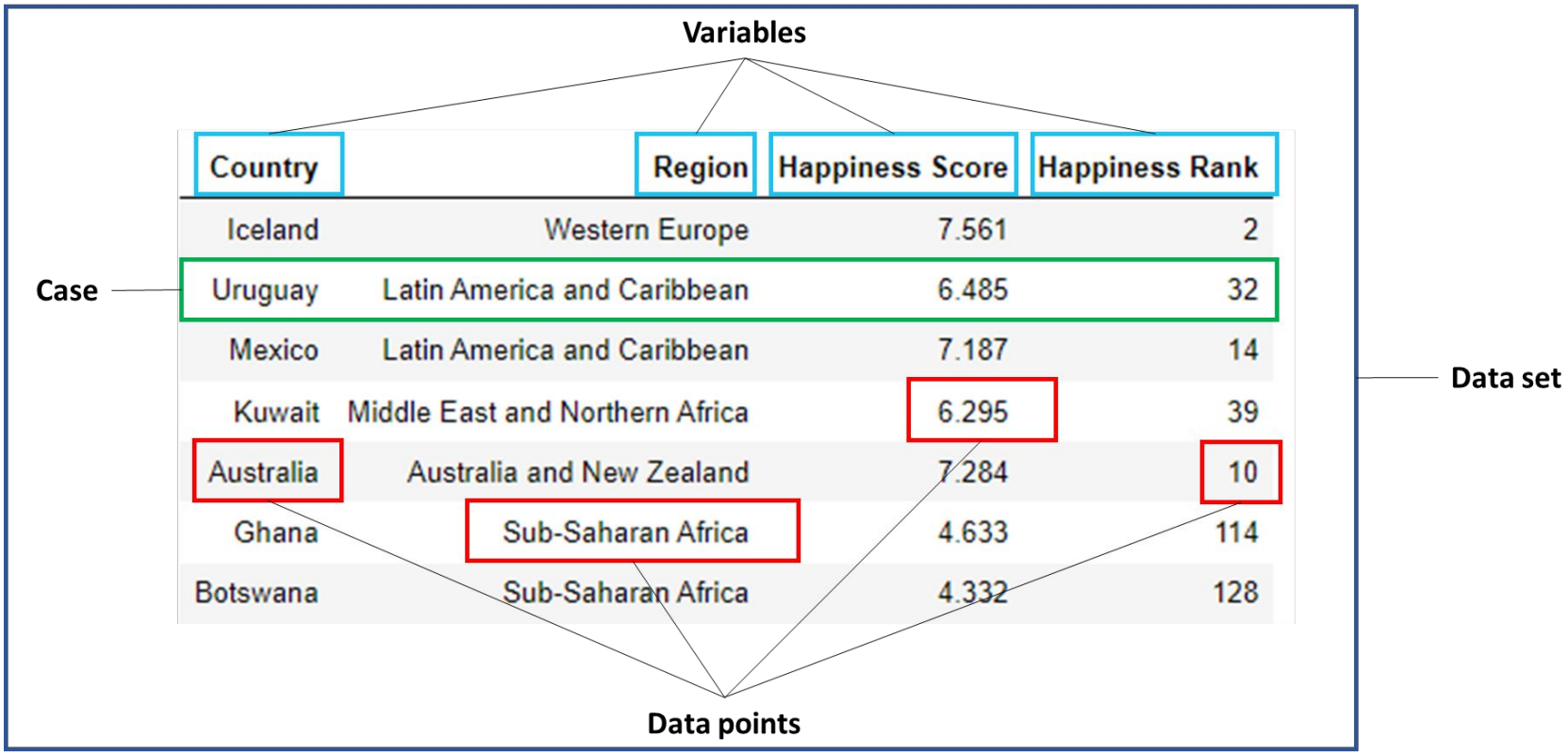
Definitions

- A **variable** is anything that can vary; it's anything that can take on a different quality or quantity.
  - The information about different variables is referred to as **data**, a term that's at the center of statistical analysis
- **Statistical Analysis:** Collection, organization, and interpretation of data according to well-defined procedures.
  - When the data relative to some specific variables are assembled, we refer to the collection or bundle of information as a **data set.**
  - The individual pieces of information are referred to as **data points**
- **Case (Observation):** The data points for the same subject

Variable Examples

- Age of students
- Attitudes toward a particular social issue
- The number of hours people spend on social media
- The crime rates in different cities
- The levels of air pollution in different locations

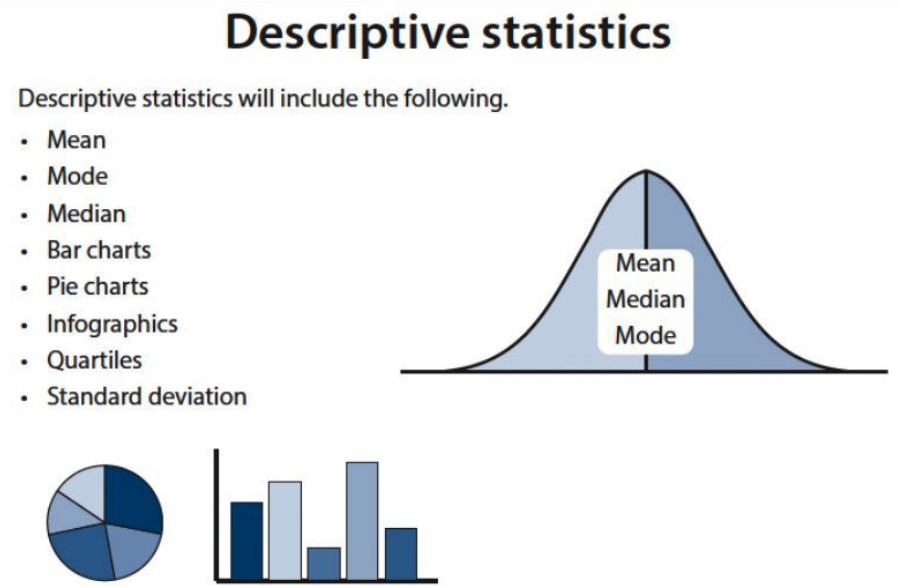
Examples

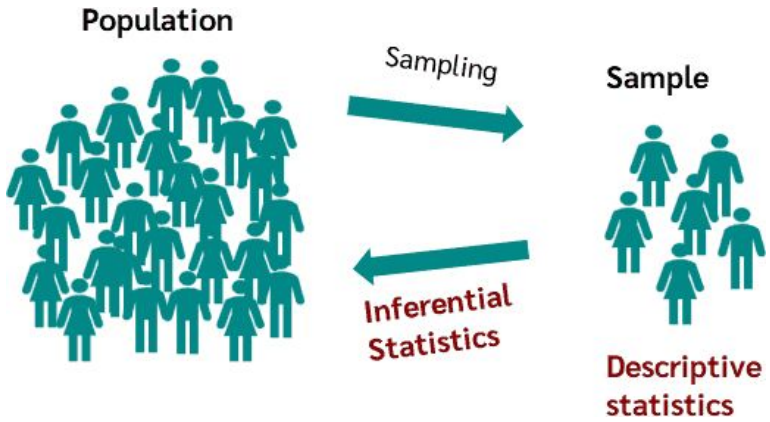


Country	Region	Happiness Score	Happiness Rank
Iceland	Western Europe	7.561	2
Uruguay	Latin America and Caribbean	6.485	32
Mexico	Latin America and Caribbean	7.187	14
Kuwait	Middle East and Northern Africa	6.295	39
Australia	Australia and New Zealand	7.284	10
Ghana	Sub-Saharan Africa	4.633	114
Botswana	Sub-Saharan Africa	4.332	128

Statistics

- **Statistics** is the science of **collecting**, **analyzing**, and **interpreting** data. It is a way to make sense of information.





Types of statistics

Descriptive statistics

Summarize and describe the data

50% Happy
25% Sad
25% Neutral

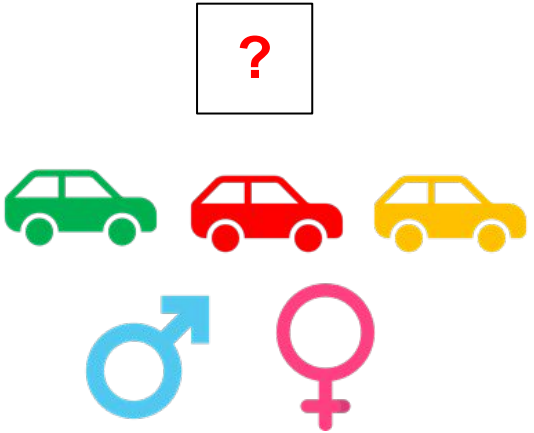
Inferential statistics

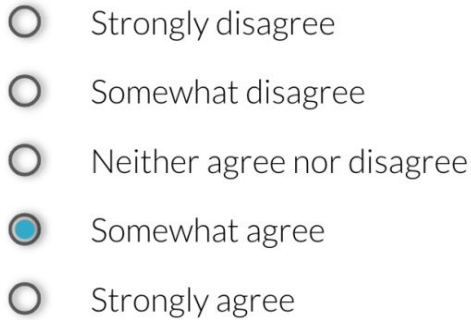
Make inferences about a population based on a sample of data

What percent of people are happy?

Types of data

- **Numeric (Quantitative)**
  - **Continuous**
    - Person height
    - Time it takes to complete a task
  - **Discrete**
    - Number of students in a class
    - Number of pages in a book
- **Categorical (Qualitative)**
  - Nominal (Unordered)
    - Married/unmarried
    - Country of residence
  - Ordinal





Exercise

Variable	Numeric		Categorical	
	Continuous	Discrete	Nominal	Ordinal
Temperature of a city	✓			
Marital status			✓	
Number of siblings		✓		
Favorite ice cream flavor			✓	
Education level				✓
Percentage of battery remaining	✓			
Year of birth		✓		
ID number			✓	
Number of steps taken in a day		✓		
Blood type			✓	

Why does data type matter ?

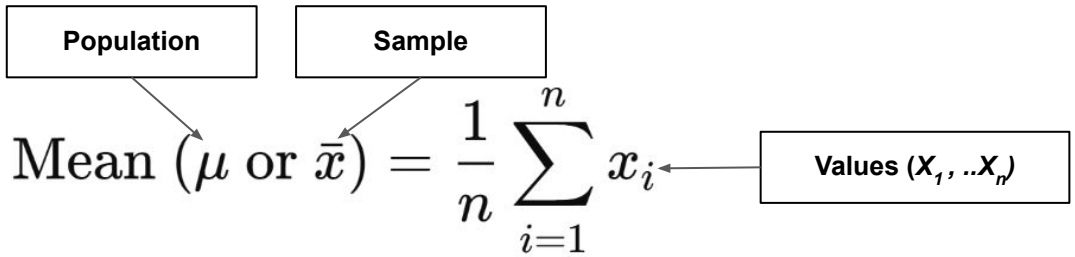
- Determines which **summary** statistics are appropriate
- Guides the choice of **visualizations** (e.g., scatter plot vs. bar chart)
- Ensures accurate interpretation of data
- Prevents misuse of statistical methods (e.g., mean on categorical data)

Describe Data Distribution

- **Central tendency measures**
  - The purpose behind any measure of central tendency is to get an idea about the **center**, or typicality, of a distribution
  - **Mean**
  - **Median**
  - **Mode**
- **Measure of spread**
  - Spread/Variability is an expression of the extent to which the data points are spread out in a distribution
  - **Range**
  - **Variance & Standard deviation**
  - **Quantiles**

Central Tendency Measures : Mean

- The mean is a measure of central tendency, calculated by **summing all data values and dividing by the count of observations**. It is commonly referred to as the **arithmetic mean** to distinguish it from other types of averages, such as **geometric** and **harmonic** means.
- **Arithmetic mean :**



- Types of **arithmetic** means:
  - **Population mean :** Calculated when data includes every member of a population
  - **Sample mean:** Calculated from a subset (sample) of the population. The sample mean is used to estimate the population mean

Central Tendency Measures : Mean

- **Geometric Mean (GM)**
- For multiplicative growth (e.g., investment returns, population growth)

$$\sqrt[n]{x_1 \cdot x_2 \cdot \ldots \cdot x_n}$$

- **Example**: Investment returns over 3 years: 10%, 20%, 30%
  - (converted to growth factors: 1.1, 1.2, 1.3)

$$\text{GM} = \sqrt[3]{1.1 \times 1.2 \times 1.3} \approx \sqrt[3]{1.716} \approx 1.197$$

- Average annual return: 19.7% (since 1.197−1=0.197)

Central Tendency Measures : Mean

- **Harmonic Mean (HM)**
  - Rates of the same quantity (e.g., speed)

$$\mathrm{HM} = \frac{n}{\frac{1}{x_1} + \frac{1}{x_2} + \cdots + \frac{1}{x_n}}$$

Remember ?

$$F1 = \frac{2 \times (\mathrm{Precision} \times \mathrm{Recall})}{\mathrm{Precision} + \mathrm{Recall}}$$

- **Example**: Driving 120 km at 60 km/h and 120 km at 40 km/h

$$\mathrm{HM} = \frac{2}{\frac{1}{60} + \frac{1}{40}} = \frac{2}{\frac{1}{24}} = 48 \text{ km/h}.$$

Central Tendency Measures: Arithmetic Mean

- **Advantage**: Easy to calculate and understand
- **Disadvantage**: Highly influenced by **extreme values,** potentially giving a misleading central tendency in skewed distributions
- **Example :**
  - Company A: A small tech startup where salaries are more evenly distributed.
    - Salaries: [3000, 3200, 3500, 4000, 4500]$
    - Mean: 3650
  - Company B: A large corporation with some very high executive salaries.
    - Salaries: [3000, 3200, 3500, 4000, 50000]$
    - Mean: 12740

Central Tendency Measures: Arithmetic Mean

- **Example Cont.**
- In **Company A**, all employees earn between $3000 and $4500, reflecting a more uniform salary structure. In **Company B**, four employees earn typical salaries, but one executive has a significantly higher salary of $50000, creating an extreme value
- **Company B** has mean salary of $12740, which is significantly higher than the majority of salaries and doesn't accurately reflect what most employees earn. This difference is due to the outlier (the $50000 salary).
- The inflated mean in Company B might give a misleading impression of higher typical salaries if we only report the mean.

Central Tendency Measures: Arithmetic Mean

- **Happiness dataset**

```
df['happiness_score'].mean()
```

```
5.539795620437956
```

- How can we interpret a mean world happiness score of 5.5?
  - Since the mean provides us with a central or "average" of the data then a mean happiness score of 5.5 (on a scale from 0 to 10) suggests that the overall happiness level is **slightly above the midpoint,** indicating that, on average, people around the world rate their happiness as **moderate**.

Practical Example

Continue in Jupyter Notebook