# CHPC & NITheCS Coding Summer School

# Probability & Statistics

Exploratory Data Analysis

René Stander
Department of Statistics
February 2024

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

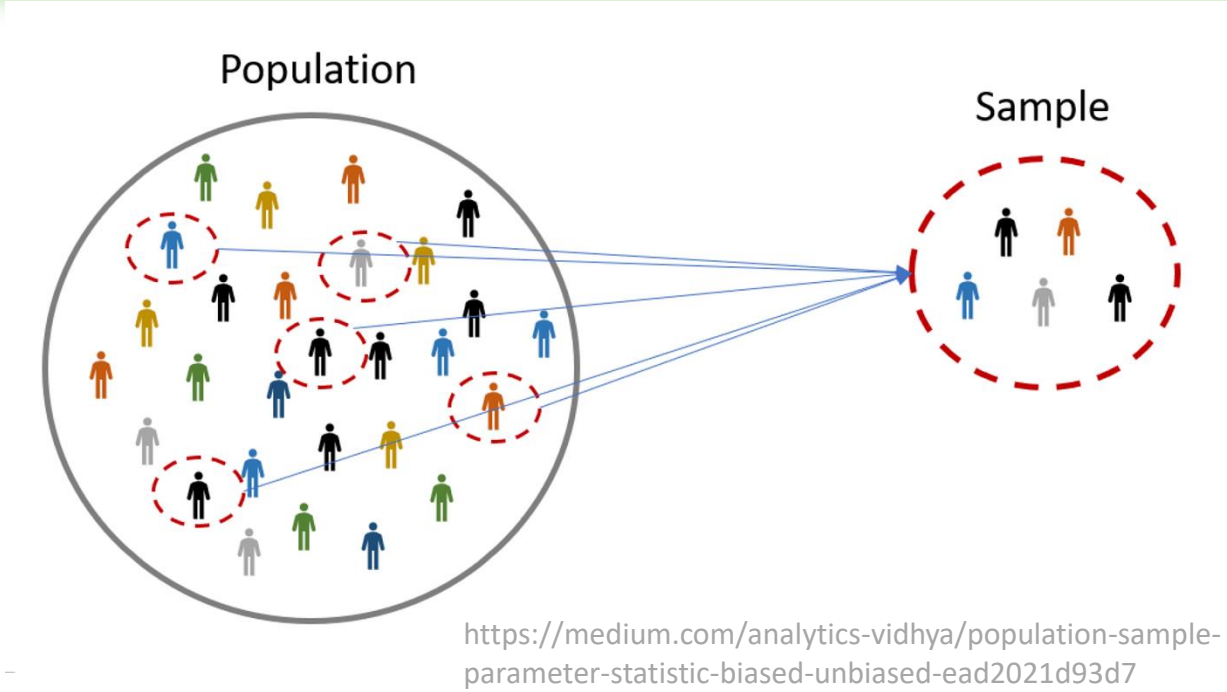
NITheCS
National Institute for
Theoretical and Computational Sciences

# What is Statistics?

- It is **making sense** of numbers.
- Making **informed decisions** in the presence of uncertainty and variation.
- Organising and summarising data to **draw conclusions** based on the information contained in the data.


https://medium.com/@yennhi95zz/the-importance-of-statistics-concepts-in-data-science-and-machine-learning-e84451c7b67

# Population vs Sample




https://medium.com/analytics-vidhya/population-sample-parameter-statistic-biased-unbiased-ead2021d93d7

- **Population:** Collection of objects about which information is sought.
- **Sample:** Part of the population that is observed.

# Data

- We are usually interested in certain characteristics of objects.
- A **variable** is any characteristic whose value may change from one object to another.

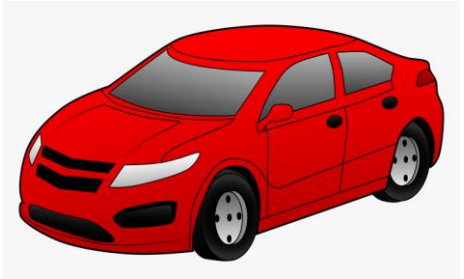
https://pixabay.com/vectors/amazon-stars-star-ratings-5094895/

https://www.zmescience.com/science/why-eyes-colored-04322/

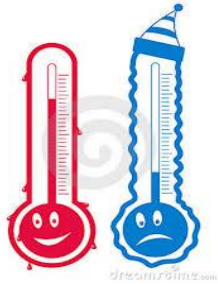
https://www.clipartmax.com/middle/m2H7d3N4N4G6H7K9_kid-measuring-clip-art-measuring-height-clipart/

Types of data:

- Categorical
- Numerical (Quantitative)



https://www.pngitem.com/middle/hwmRbx_clipart-art-red-car-clipart-picture-of-car/



https://pixy.org/4154488/


https://www.shutterstock.com/search/marital+status

# Exploratory Data Analysis (EDA)

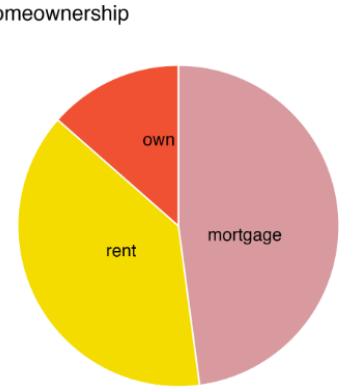
- simplify large amounts of data in a sensible way.
- do not draw conclusions beyond data we are analysing.
- do not reach conclusions regarding hypothesis.
- do not try to infer characteristics of the population.
- present quantitative descriptions of the data in a manageable form.
- simply describe the data.
- basis of every quantitative analysis.

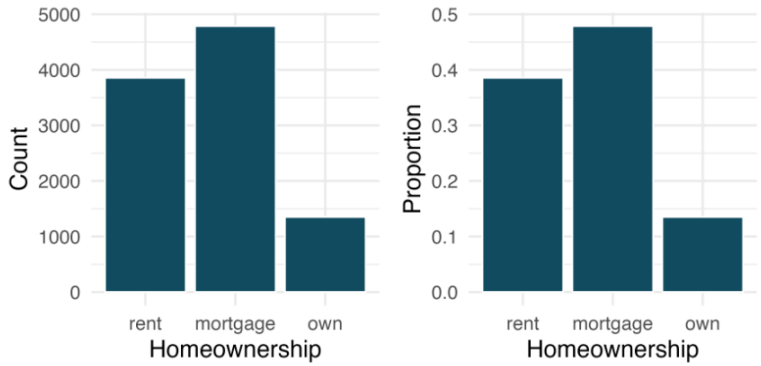
# Categorical Data

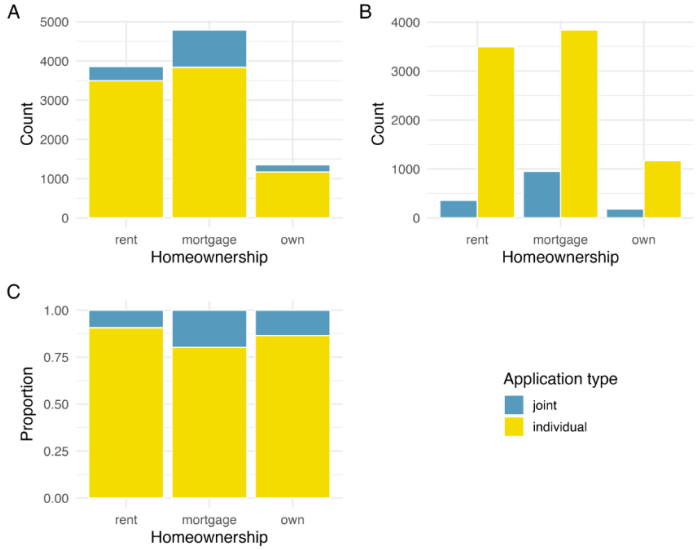
https://pixabay.com/vectors/amazon-stars-star-ratings-5094895/

https://www.zmescience.com/science/why-eyes-colored-04322/

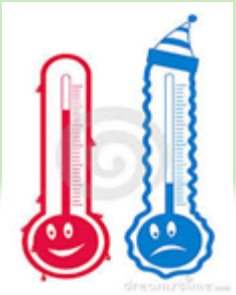
- **Descriptive Statistics:**
  - Frequency tables
- **Visualisation:**
  - Bar plot
  - Pie chart



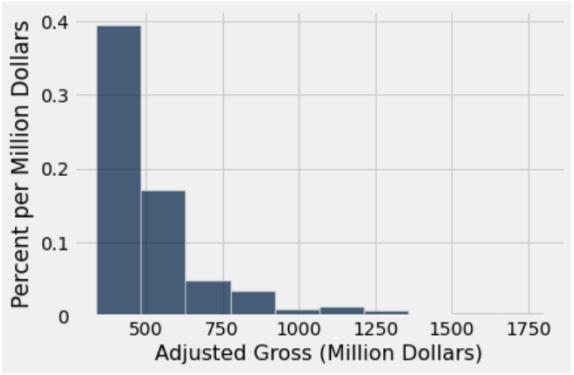
# Numerical Data

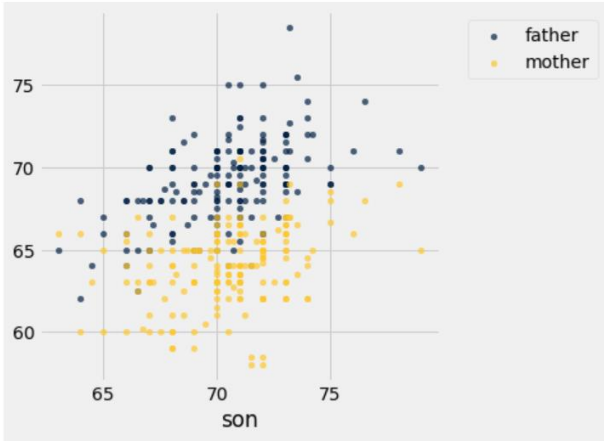


https://pixy.org/4154488/

https://www.clipartmax.com/middle/m2H7d3N4N4G6H7K9_kid-measuring-clip-art-measuring-height-clipart/

- **Descriptive Statistics:**
  - Location
    - Mean
    - Five-number summaries (Minimum, Q1, Q2, Q3, Maximum)
    - Median
  - Variability
    - Standard deviation
    - Interquartile range
- **Visualisation:**
  - Histogram
  - Box-and-whisker plot
  - Scatterplot




# Exercise: Pick n Pay animal cards


https://www.seamonster.co.za/portfolios/super-animals-1/

- Complete the assignments in the Python script.


https://htxt.co.za/2017/05/15/heres-a-scan-of-every-single-pick-n-pay-super-animals-2-card/


https://htxt.co.za/2017/05/15/heres-a-scan-of-every-single-pick-n-pay-super-animals-2-card/