

# CHPC & NITheCS Coding Summer School Probability & Statistics

Data Handling & Exploratory Data Analysis

René Stander

Department of Statistics

February 2025



UNIVERSITEIT VAN PRETORIA  
UNIVERSITY OF PRETORIA  
YUNIBESITHI YA PRETORIA

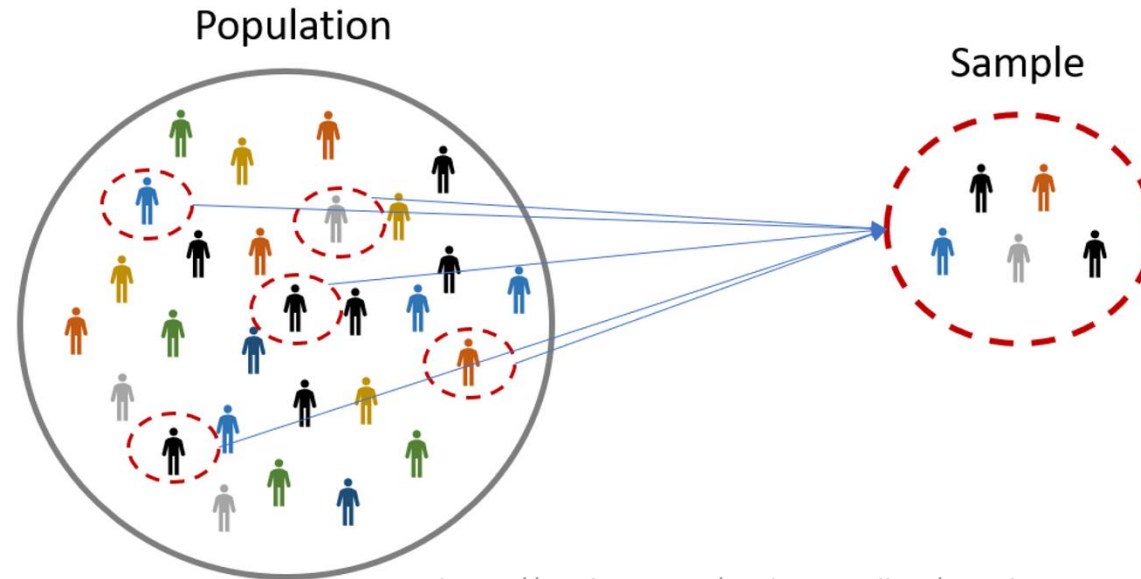
**NITheCS**  
National Institute for  
Theoretical and Computational Sciences

# 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.



# Population vs Sample



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

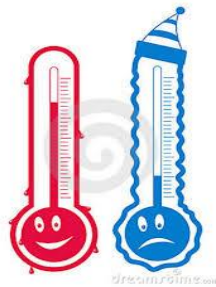
- **Population:** Collection of objects about which information is sought.
- **Sample:** Part of the population that is observed.

# Data

- Data are the **raw facts** that are collected, analysed and summarised for presentation and interpretation in order to make informed decisions.
- We are usually interested in certain characteristics of objects.
- A **variable** is any characteristic whose value may change from one object to another.

## Types of data:

- Categorical
- Numerical (Quantitative)



<https://pixy.org/4154488/>



[https://www.pngitem.com/middle/hwmRbx\\_clipart-art-red-car-clipart-picture-of-car/](https://www.pngitem.com/middle/hwmRbx_clipart-art-red-car-clipart-picture-of-car/)



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



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



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



[https://www.clipartmax.com/middle/m2H7d3N4N4G6H7K9\\_kid-measuring-clip-art-measuring-height-clipart/](https://www.clipartmax.com/middle/m2H7d3N4N4G6H7K9_kid-measuring-clip-art-measuring-height-clipart/)

# Data is everywhere!

Data is collected when you:

- Make a purchase online (such as Takealot)
- Click an advertisement
- Like or comment on someone's social media post
- Stream music or movies online (using platforms such as Spotify and Netflix)
- Review our experience with a product or service online (such as our stay at an Airbnb)
- Engage in physical activity and even while we are sleeping! (when we are wearing a fitness tracker such as a Fitbit or Garmin).



Spotify®

NETFLIX





airbnb

GARMIN™

# Data set structure

Variables in the columns

Elements in the rows



Province	Coastal (1) or Inland (2)	Population size	HDI	% of agricultural households	Land area (sq km)	Population density (per sq km)	% of households with no internet access	Sex ratio	Median age
Eastern Cape (EC)	1	7230204	Medium	20	168966	43	34.3	90	27
Free State (FS)	2	2964412	High	6	129825	23	20.8	90.4	28
Gauteng (GP)	2	15099422	High	11	18178	831	13.6	101.8	30
Kwa-Zulu Natal (KZN)	1	12423907	High	22	94361	132	18	91	28
Limpopo (LP)	2	6572720	High	21	125754	52	31.9	89.2	26
Mpumalanga (MP)	2	5143324	Medium	10	76495	67	22	92.4	27
Northern Cape (NC)	1	1355946	High	1	372889	4	28.8	93	27
North West (NW)	2	3804548	Medium	7	104882	36	27.6	98.2	27
Western Cape (WC)	1	7433019	High	2	129462	55	16.1	94	31

Observation

# Data Handling

- Import datasets
- Extracting variables
- Removing variables
- Subset the dataset
- Removing observations
- Creating new variables
- Removing duplicates
- Sort the data

# Exercise: Data Handling

The data used for this example is based on the online Racer game from the Statistics Department at Grinnel College, Iowa, USA

<https://www.stat2games.sites.grinnell.edu/>

For this example, follow the instructions in the following Python script: `Data-Handling.py`





# On your own: Data Handling



The data used for this example is based on the online CoffeeTruck game from the Statistics Department at Grinnell College, Iowa, USA  
<https://www.stat2games.sites.grinnell.edu/>

Use the **CoffeeTruck.csv** data and answer the following questions in a Python script:

1. How many rows and columns does the data set have?
2. List all the variables in the data set.
3. Find the unique values in the 'Location' variable. How many unique locations are present in the data set?
4. Subset the data to include only the observations of sales at the Zoo.
5. How many duplicated rows are there in the data? Remove all the duplicated rows from the data set. How many rows are left in the data set?
6. Sort the data set first by profit made (from smallest to largest) and then by music played (from Z to A). (Make use of the data set where the duplicates are removed.)
7. Create a new variable called 'Indicator' that contains the word 'Loss' if the profit for the day is less than 0, and 'Profit' if the profit for the day is greater than 0. (Make use of the sorted data set.)

# 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

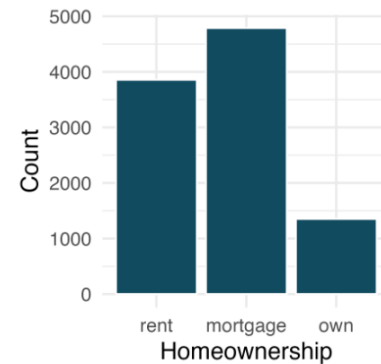
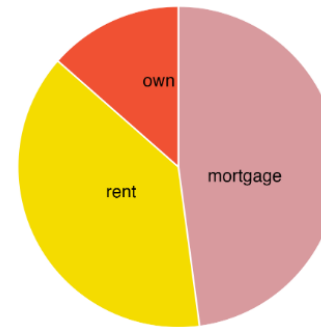
- **Descriptive Statistics:**

- Frequency tables

- **Visualisation:**

- Bar plot
- Pie chart

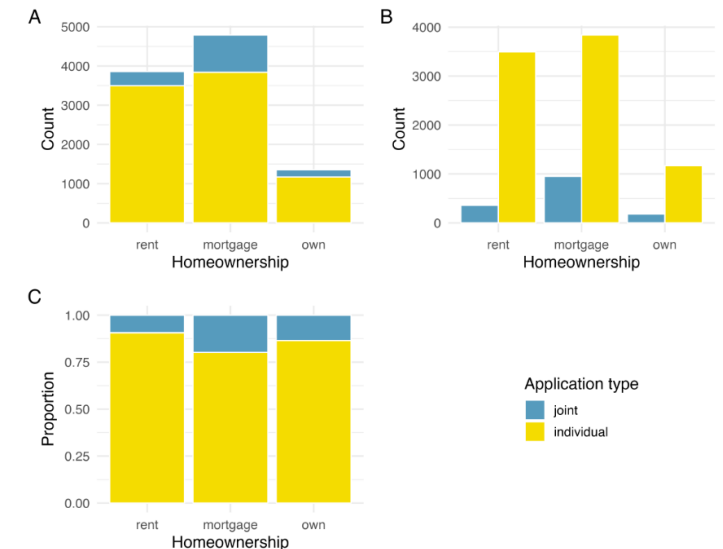
Homeownership



<https://pixabay.com/ve ctors/amazon-stars-star-ratings-5094895/>



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



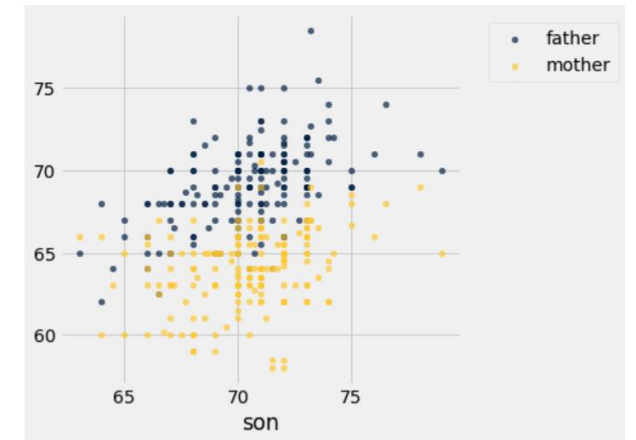
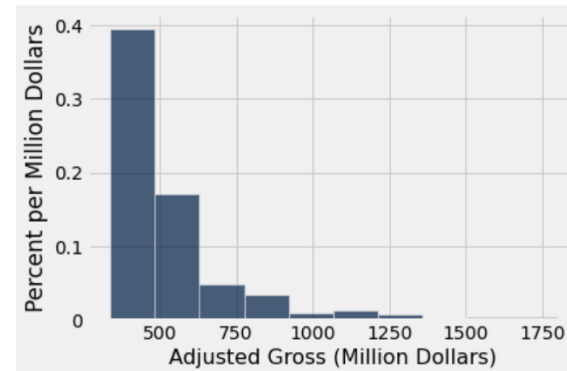
# Numerical Data

- **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/>

# On your own: EDA



The data used for this example is based on the online CoffeeTruck game from the Statistics Department at Grinnell College, Iowa, USA  
<https://www.stat2games.sites.grinnell.edu/>

Use the **CoffeeTruck.csv** data and answer the following questions in a Python script:

1. Draw a frequency table of the 'Location' variable. At which location did the coffee truck stop most frequently?
2. Create a new variable called 'Indicator' that contains the word 'Loss' if the profit for the day is less than 0, and 'Profit' if the profit for the day is greater than 0.
3. Draw a frequency table of the 'Music' variable and the new "Indicator" variable.
4. Draw a histogram of the profits made. Create appropriate labels for the plot.
5. Draw a barplot of the 'Music' variable.
6. Draw a side-by-side boxplot of the profit per location. Which location has the most outliers?
7. Obtain numerical summaries of the profit for each location.
8. BONUS: What is the mean number of sales at the Zoo when the coffee truck made a profit for the day?