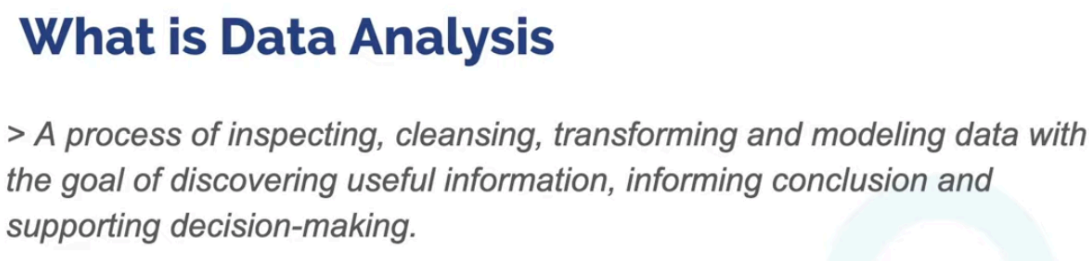
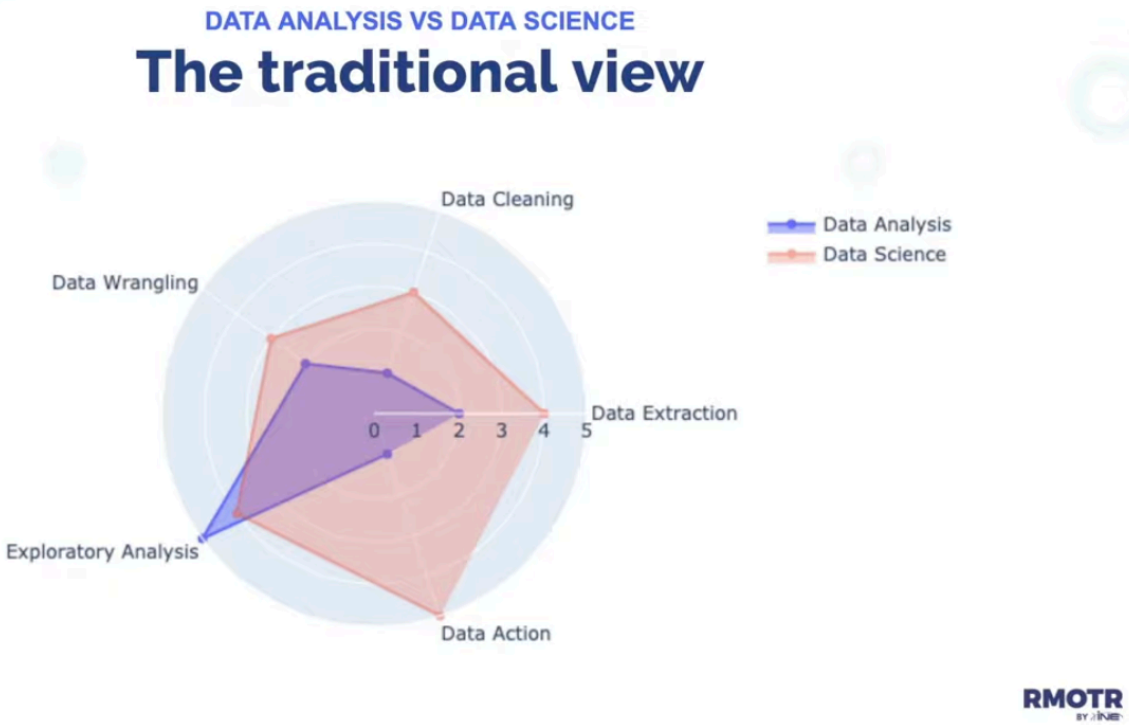
- ***Course outline***
  - data analysis with Python <- freeCodeCamp, remoter collaboration
  - reading data from multiple sources
  - cleaning and transforming data <- statistical functions
  - pandas, matplotlib, seaborne
  - managing data with Python and traditional data analytics
- ***About this tutorial***
  - what is data analysis
  - in the context of Python
  - SQL and pandas
  - SQL ('sequel')
  - an example / demonstration of this
  - explaining the tools in detail
  - Jupyter tutorials
  - Python in under 10 minutes <- Python recap
- ***What is data analysis***



  - -> gathering and cleaning data for analysis <- Pandas, Python
  - -> modelling data <- adapting real life scenarios to information systems
    - -> using inferential statistics
    - -> Pandas and seaborn visualisations
  - -> then discovering useful information -> we are trying to take the data and come out with patterns
    - -> providing evidence of the findings and visualising the patterns found
- ***Data analysis tools***

  - Ones which are open source and then ones which are sold by vendors
  - one is open source, one is closed source
  - one is more expensive, one is cheaper
  - one is powerful and the other is more limited
  - open source projects are harder to learn, but they are cheaper and more powerful
  - tools which come from vendors are also dependent on the company which made them
  - Python is free and open source -> many people contribute to it
  - advantages of Python over R and Julia
    - -> simple, "correct", powerful libraries, free and open source, conferences and docs
    - -> Python is preferred over R, since it is more general

        - -> R <- for statistical functions
- ***Process***
    - -> data extraction <- get the data
    - -> data cleaning <- put the data into the right form
    - -> data wrangling <- merge the tables etc
    - -> analysis <- perform statistical analysis on the data
    - -> action <- suggest an action given the suggested analysis
- ***Data analysis vs data science***



    - -> data scientists have more maths skills
    - -> data analysts have better communication skills
    - -> data scientists are more prestigious than analysts
- ***Libraries***
    - -> Python & PyData ecosystem
    - -> pandas
    - -> matplotlib <- visualisation
    - -> seaborn <- visualisation
- ***Python data analysts***
    - -> you are used to having a constant visual representation of the data
    - -> you can't see all of the data when you are using Python
- ***Which is not part of data analysis?***
    - Building statistical models and data visualizations.
    - Picking a desired conclusion for the analysis. <- this one
    - Fixing incorrect values and removing invalid data.
    - Transforming data into an appropriate data structure.