- ***-> working with data analysis and data processing with Python***
  - -> exploring what the tools do and what the best way to use them is
  - -> the data is in a CSV file
    - -> the data is a sales data spreadsheet
    - -> we want to understand this
  - -> read the csv file into Python in one line of code
  - -> then we have a data frame <- a CSV representation
- ***-> there is a lot of data (100,000 rows) -> so when you process the entire data frame, you can't see all of the data***
  - -> then using .info() in Python to print summary statistics
  - -> the summary statistics help you a lot more than looking at the individual data to try and understand the patterns will
  - -> you can look at, for example the median age - to see what makes sense given the context
  - -> you don't need to go and look at the data (you can just look at the summary statistics)
- ***-> then plotting the data using matplotlib***
  - -> this is a box and whiskers plot
  - -> this shows outliers
  - -> another example is a density plot
  - -> the mean and the median in the plots
  - -> the age groups / age of customers
  - -> you can also enter ... .head to show a specific database's information
  - -> in this case here are four different age categories
  - -> you can also create a pie chart -> this shows the amount of people in each age range relative to each other
- ***-> correlation***
  - -> you can make a correlation matrix
  - -> profit and loss
  - -> calculating the profit correlation with quantity
  - -> the more people spend the more the company makes
  - -> you can also analyse the age with the profit -> there is a linear dependency
    - -> profit per age group
- -> https://github.com/ine-rmotr-curriculum/FreeCodeCamp-Pandas-Real-Life-Example.git
  - -> these are the notebooks from this section of the course
- ***-> question***
  - What does the shape of our dataframe tell us?
    - The size in gigabytes the dataframe we loaded into memory is.
    - How many rows and columns our dataframe has <- This one
    - How many rows the source data had before loading.
    - How many columns the source data had before loading.