

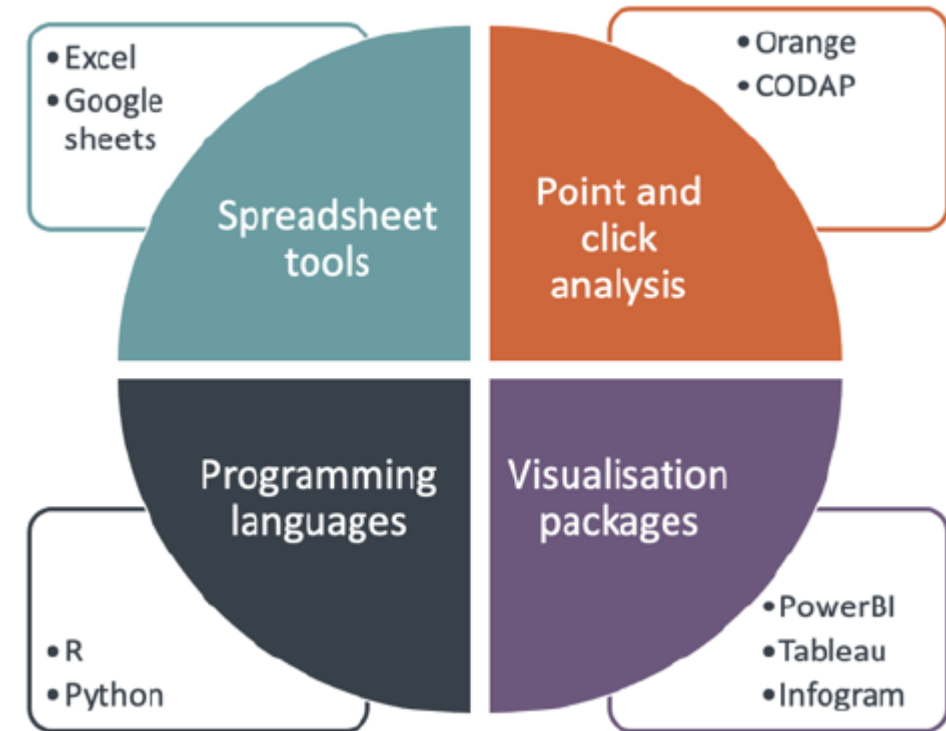


TOOLS AND LANGUAGES

Lecturer: Mark Gillan

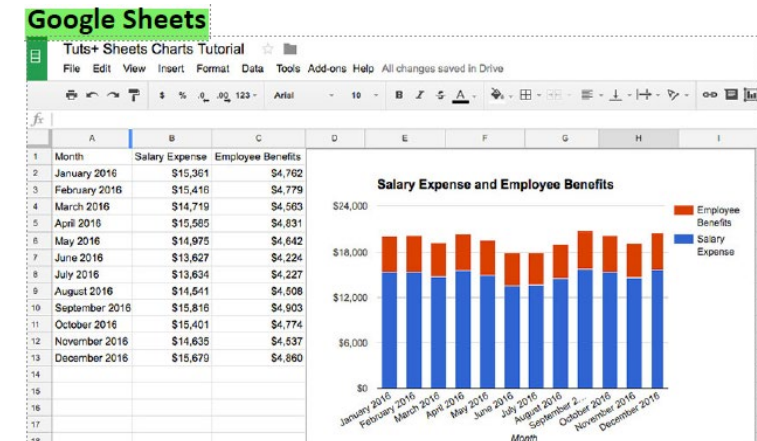
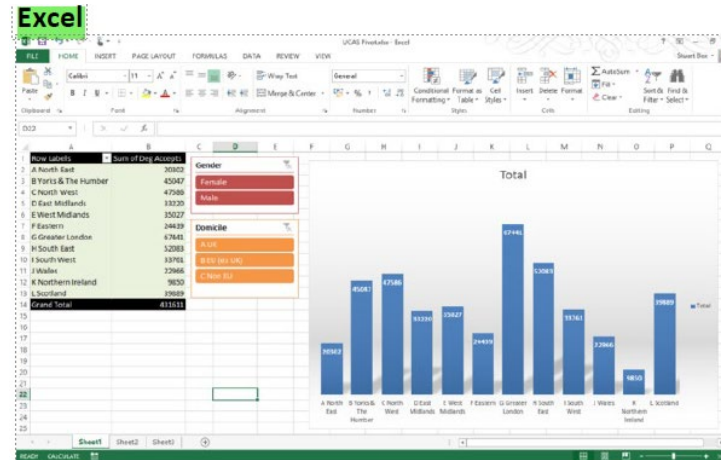
TOOLS AND LANGUAGES

- There are a number of different tools and languages that can be used for data analysis and visualisation.
- The choice of tool to use will depend on several different factors, such as access to the internet, requirements to keep data private, the need for version control, time available to learn to keep data private, the need for version control, time available to learn to use the tool and the programming experience of the user
- Some tools can be used for analysis and visualisation, whilst others specialise in visualisation only.



SPREADSHEETS

- Spreadsheet tools are suitable for quick ad-hoc data analysis and visualisation.
- The two most popular tools are Microsoft Excel and Google Sheets. Both tools have scripting languages which can be used to automate repetitive tasks.
- Challenges with all spreadsheet tools include auditability, error checking and reproducibility of analysis.



COMMERCIAL VISUALISATION PACKAGES

Power BI



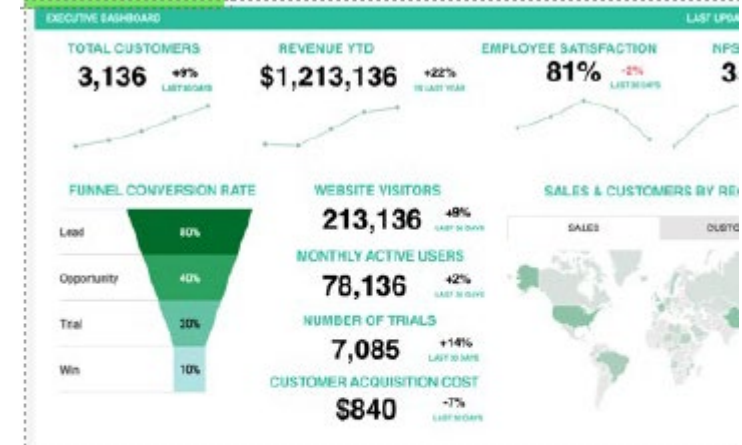
- powerbi.microsoft.com
- Part of the office suite
- Easily links to excel for data preparation

Tableau



- tableau.com
- Good array of training materials available

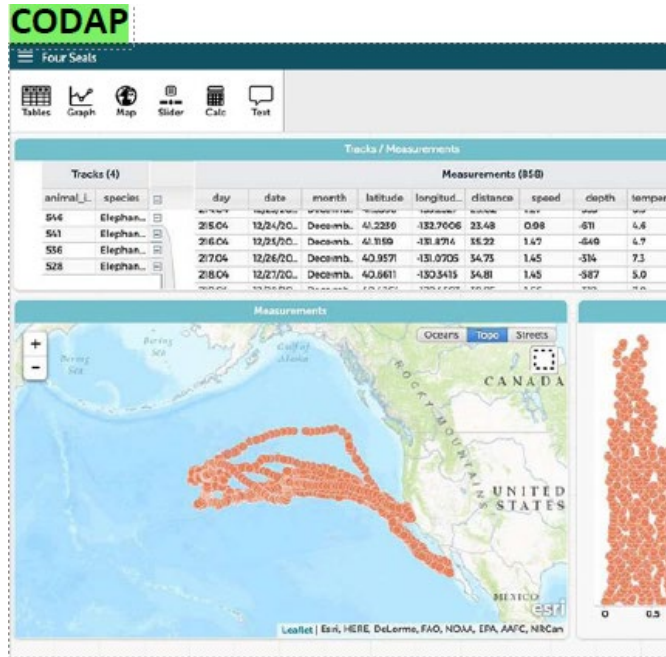
Infogram



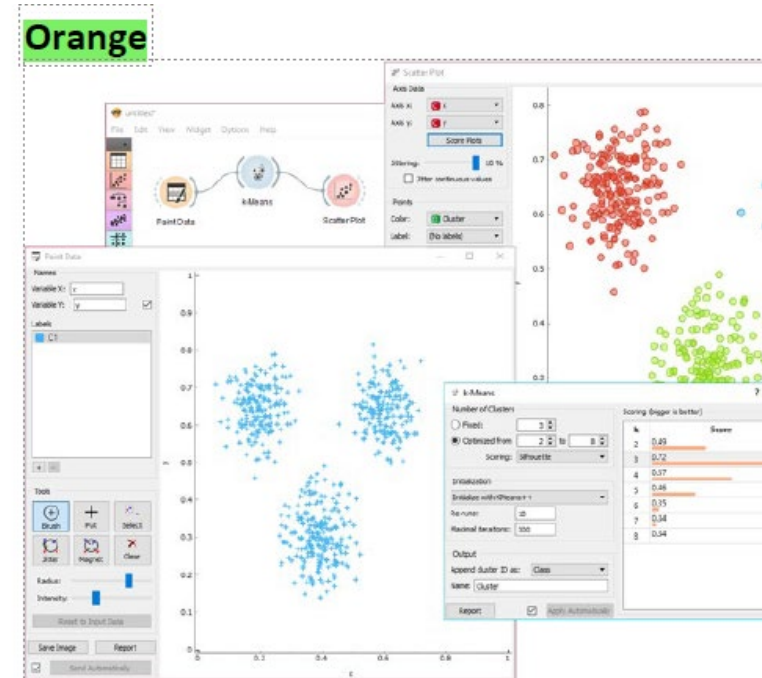
- infogram.com
- Basic version is free but with limited functionality



POINT AND CLICK



- codap.concord.org
- Designed for educational use online
- Intuitive data exploration
- Datasets are not secure
- No version control



- orange.biolab.si
- Must be downloaded locally
- Takes time to become proficient
- Enables advanced analysis such as predictive modelling



PROGRAMMING LANGUAGES

The advantages of using a programming language for analysis and visualisation are flexibility and reproducibility.

The main open-source languages that data scientists use are R and Python.

However, there are also commercial languages available. R and Python both have special data types designed for manipulating tidy tabular data.

They all have multiple packages for creating visualisations and more complex dashboards that can be implemented in web applications.

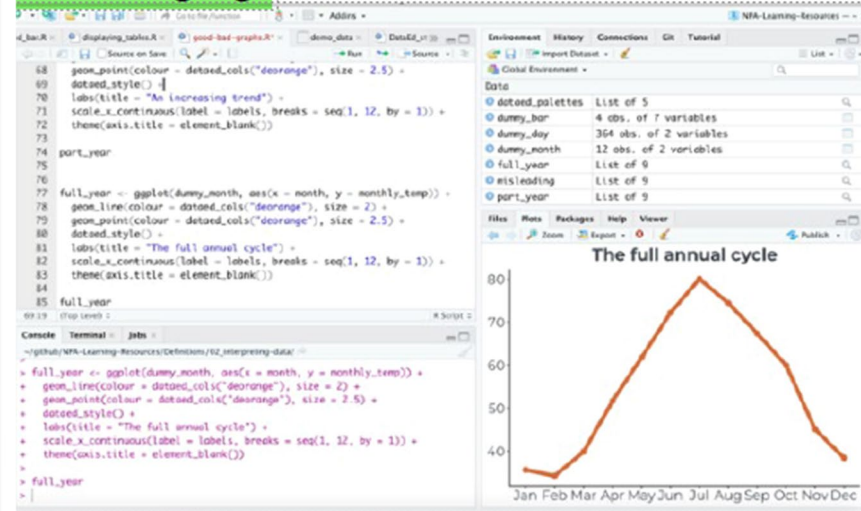
Both languages have IDEs (interactive development environments) for easy development and integration with source control tools such as git (github.com).

Alternatives to using an IDE include interactive notebooks, which can be used in either language. Popular notebooks include Jupyter (jupyter.org) or web-based Google Colab (dataed.in/colab)



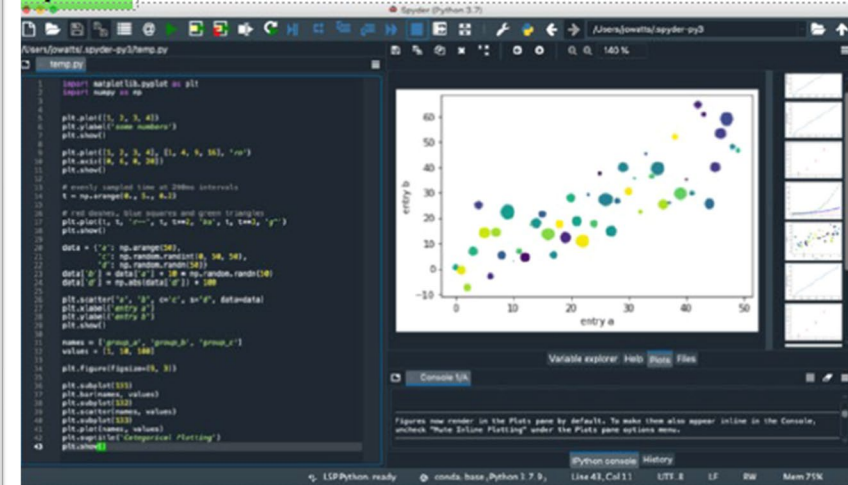
PROGRAMMING LANGUAGES OF R AND PYTHON

The R Language



- Language designed for statistical analysis and visualisation
- Has the RStudio development environment
- Popular packages:
 - **Tidyverse** for data manipulation
 - **ggplot** for visualisations
 - **RShiny** for Interactive web-based applications

Python



- A full object-oriented programming language
- A recommended IDE to use is Spyder, available as part of the Anaconda suite
- Popular packages for data science:
 - **Pandas** for data manipulation
 - **Numpy** for data functions
 - **Matplotlib** or **Seaborn** for visualisation
 - **Plotly** for interactive plotting

SCOTTISH QUALIFICATION AUTHORITY

- NPA Data Science Learner's Guide
- This Learner's guide has been produced with an accompanying Educator's guide. Both guides have been created to support the core units of the National Progression Award (NPA) in Data Science at SCQF levels 4, 5 and 6.
- Share Alike 4.0
- Attribution – Scottish Qualification Authority Pack for Data Science
- CC BY-NC-SA 4.0
- Downloaded from teachdata.science and edited for presentations to students.
- Additional material and edits by Mark Gillan
- The original pack developed thanks to the support and funding from the Scottish Government – Riaghaltas na h-Alba gov.scot



DEVELOPED THANKS TO SUPPORT AND INVOLVEMENT



THE UNIVERSITY
of EDINBURGH