Data Analysis with Python

Main Concepts: Numpy, Pandas, Matplotlib, Seaborn

1. What is Data Analysis
2. **Real Example Data Analysis with Python**
3. How to use Jupyter Notebooks
4. Intro to NumPy (exercises included)
5. Intro to Pandas (exercises included)
6. Data Cleaning
7. Reading Data SQL, CSVs, APIs, etc

Part 1: Intro

What is Data Analysis?

- Data Analysis is the process of **investigating, organizing, transforming** and **modelling data** with the goal of **discovering useful information**, supporting decision-making and determining significant results and trends found within the data itself
    - Investigating, organizing, transforming
        - This part is usually tedious
        - It starts by gathering the data, cleaning it and transforming it for further analysis
        - This is where Python and PyData tools excel
        - Will use Pandas to read, clean and transform data
    - Modelling data
        - This means adapting real life scenarios to information systems
        - Using inferential stats to see if any patterns or models arise
        - For this, I'll use stat analysis features of Pandas and visualizations from Matplotlib and Seaborn
    - Discovering useful info
        - Once we've processed the data and modelled it, aim: derive conclusions from it
            - Patterns or anomalies
        - Information is key
        - Aim: transform data into information
            - Ex: data: list of purchases in Walmart last year; Info: pop tarts sell a lot on Tues
    - Informing conclusion and supporting decision-making
        - Final objective of data analysis
        - Need: provide evidence of findings, create readable reports & dashboards, aid other ppl/ departments with the information gathered

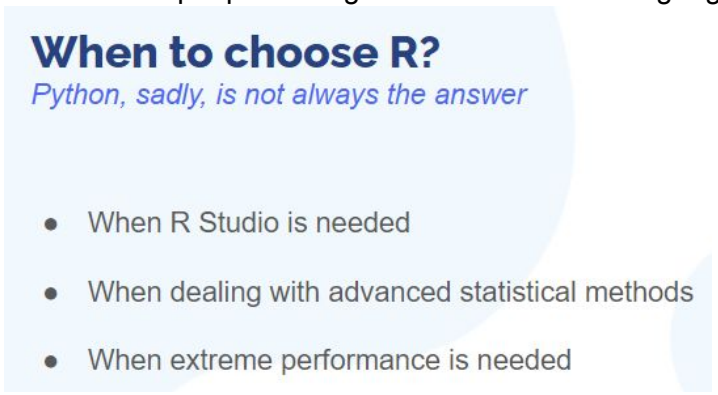
Auto-managed closed tools vs. Programming languages

Auto-managed Closed Tools	Programming Languages
Qlik	python
tableau	R
looker	julia
ZOHO Analytics	

- Auto managed tools are closed products
    - Tools you can buy and start using right out the box
    - Ex: excel
        - Tableau, looker are most popular for data analysis
        - Advantages: easy to learn
        - Disadvantage: scope of tool is limited, can't cross the boundaries of it
- In contrast, using Python and the universe of PyData tools, gives you amazing flexibility. Do you need to read data from a closed API using secret key authentication? You can do it. Do you need to consume data directly from AWS Kinesis? You can do it. A programming language is the most powerful tool you can learn. Another important advantage is the general scope of a programming language. What happens if Tableau, for example, goes out of business? Or if you just get bored from it and feel like your career is stuck? Learning how to process data using a programming language gives you freedom.
- The main disadvantage of a programming language is that it's not as simple to learn as with a tool. You need to learn the basics of coding first, and it takes time.

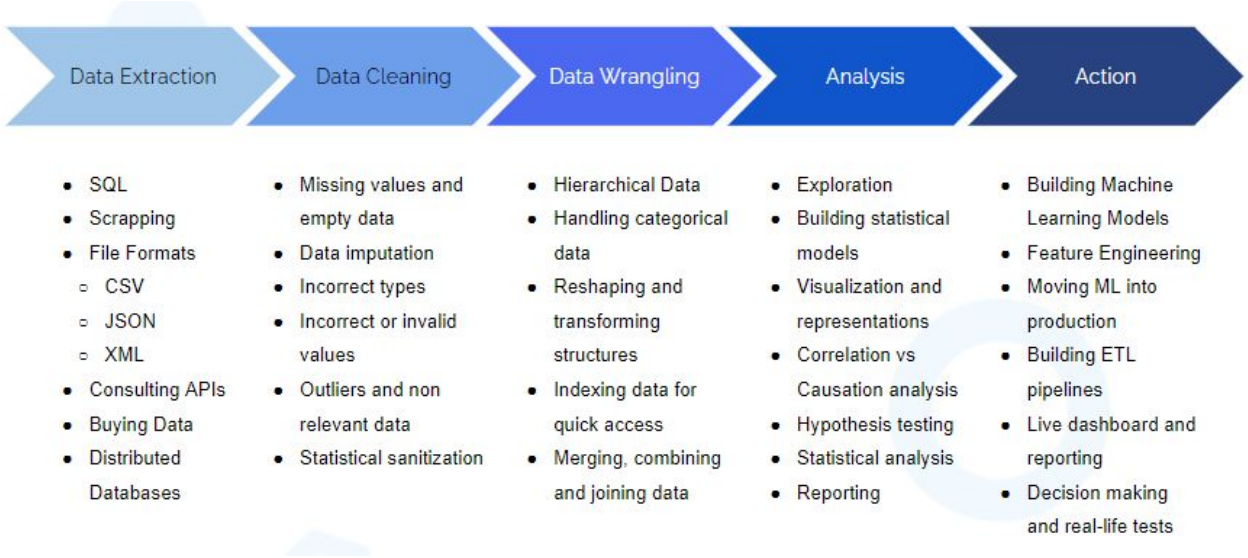
Why Python for Data Analysis?

- Python is the best programming language to learn to code. It’s simple, intuitive and readable. It includes thousands of libraries to do virtually anything, from cryptography to IoT.
- Python is free and open source. That means that there are thousands of eyes, very smart people seeing the internals of the language and the libraries



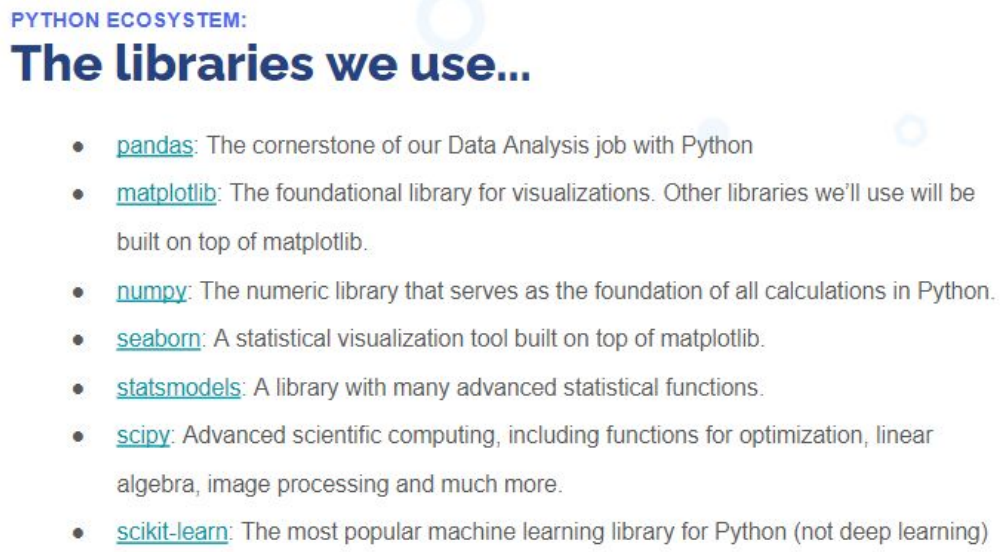
- R is also a great programming language.
- We prefer Python because it’s easier to get started, and more “general” in the libraries and tools it includes. R has a huge library of statistical functions

The Data Analysis Process



- The process starts by getting the data. Where is your data coming from? Usually, it's in your own database. But it could also come from files stored in different formats or web APIs.
- Once we've collected the data we'll need to clean it. If the source of the data is your own database, then it's probably already in shape. If you're using more extreme sources, like web scraping, then the process will be more tedious.
- With our data cleaned, we'll now need to rearrange and reshape the data for better analysis. Transforming fields, merging tables, combining data from multiple sources, etc. The objective of this process is to get the data ready for the next step.
- The process of analysis involves extracting patterns from the data that is now clean and in shape. Capturing trends or anomalies. Statistical analysis will be fundamental in this process.
- Finally, it's time to do something with that analysis. If this was a Data Science project, we could be ready to implement Machine Learning models. If we focus strictly on Data Analysis, we'll probably need to build reports, communicate our results and support decision making.

Data Analysis vs. Data Science



- The boundaries between Data Analysis and Data Science are not very clear. The main differences are that Data Scientists usually have more programming and math skills. They can then apply these skills in Machine Learning and ETL processes.
- Data Analysts, on the other hand, have better communication skills, creating better reports, with stronger storytelling abilities.
- Source: https://notebooks.ai/santiagobasulto/radar-chart-data-science-vs-data-analysis-ad638c75

- The most important libraries we'll be using are Pandas for Data Analysis, and Matplotlib and Seaborn for visualizations. But the ecosystem is large, and there are many useful libraries for specific use cases.

How Python Data Analysts Think

...

Visual tools: Excel, Tableau, etc

References

https://www.youtube.com/watch?v=r-uOLxNrNk8

https://docs.google.com/presentation/d/1fDpjlyMiOMJyuc7_jMekcYLPP2XlSl1eWw9F7yE7byk/edit#slide=id.g6fe1465eda_0_215