

Introduction to Data Analytics with Python



Overview

- Explore the basics of **Python programming** for data analytics
- Explain how to work with **datasets**
- Explain basic skills needed for **data analysis** with **pandas**:
 - **Importing Data**
 - **Exploratory Data Analysis**
 - **Data Visualisation**
- Hands-On Example: Palmer Penguins

By the end of this session you will have: a basic grasp of Python 🐍 , pandas 🐼 , and ... Palmer Penguins 🐧 🐧 🐧 ! ✨

Learning to Program with Python



How do you feel it?

Let's find out ...

Click [here](#) or scan the QR code to vote!

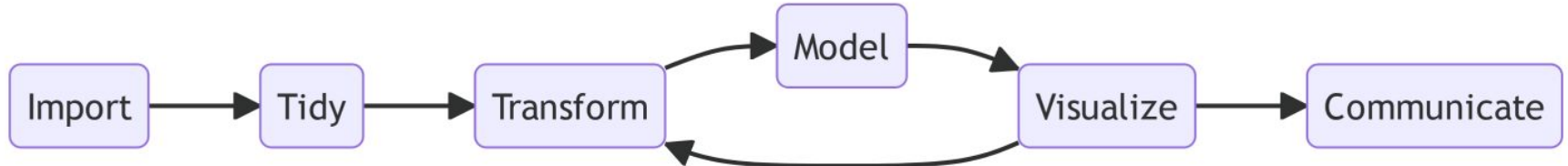
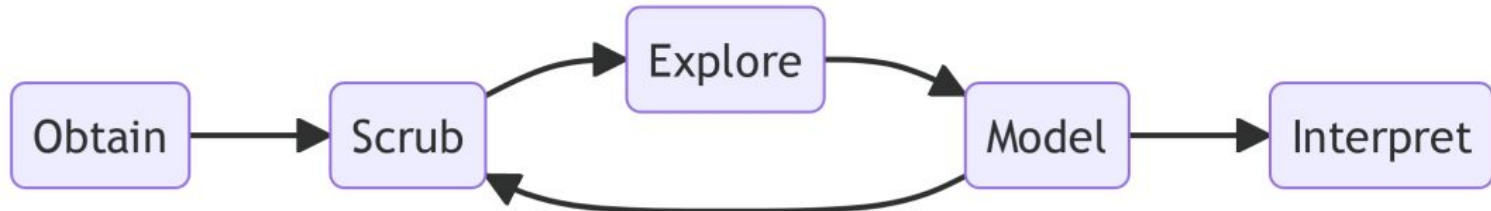
✨ What is Data Analytics? ✨

- Collect, process, analyse datasets using **statistics** and **programming**
- Identify trends, solve problems, produce actionable insights, and inform decision-making
- Science of analyzing data to reveal **patterns**, derive **insights** and inform **decisions**

Skills and knowledge of data analytics allows you to **solve real-world problems!**



What is Data Analytics?



What is Python? 🐍



Computer programming language

- Invented by Dutch programmer Guido van Rossum in Christmas 1989, first released 1991
- “Python” named after British comedy [Monty Python's Flying Circus](#)
- Now one of the world's **most popular** programming languages!

“Computer Programming for Everybody” (1999)

“We believe that Python is a good language for teaching [programming] to absolute beginners”

- Designed to develop computer literacy for coding “newbies”
- Simple syntax similar to the English language
- Free and open source software! 🙌

“Pythonic”: Zen of Python 🙏

“Beautiful is better than ugly.

Explicit is better than implicit.

Simple is better than complex ...

There should be one-- and preferably only one --obvious way to do it ...

Now is better than never”

by Tim Peters (1999)

TO DO: In a **Jupyter** code cell, try typing `import this`, and then click on the **Run Cell** arrow. What happens?

Note: Writing “Pythonic” code is related to writing “clean” code!

Why Python? 🐍

1. Free and Open Source Software Community 🤝

- Half-a-million free and open source software **packages** (or 'libraries') developed for [Python Package Index](#) (PyPi) (2024) (e.g. `numpy`, `pandas`, `matplotlib`)
- [Python Software Foundation](#) supported by global technology companies (e.g. Google, Meta, Microsoft)
- Conferences organised worldwide (e.g. PyCon, PyData, PyLadies, SciPy)! See [NumFOCUS](#)

Why Python? 🐍

2. Popular Programming Language 🙌

- Surveys show Python is **one of most popular programming languages!**
- My [Python Surveys](#) `streamlit` app summarises recent survey results 👁️

Why Python? 🐍

3. Core Tool for Data Analysis 🔍

- **SQL** (1974)
- **Bash** Command Line Interface (1988)
- **Python** (1989)
- **Git** (2005)

What is Jupyter?



- **Project Jupyter**: open source project for data analysis, data science, and scientific computing.
- “Jupyter”: named after three core scientific computing languages: **Julia**, **Python**, **R**.
- **Jupyter Notebooks**: interactive computing environments built for data analysis, which can be used in your web browser, or using Jupyter Notebook files in a local development environment (e.g. open `notebook.ipynb` in **PyCharm** or **Visual Studio Code**)

Note: Jupyter Notebooks are built on top of the **IPython** console - an interactive command line shell - another core tool for data analysts and data scientists!

Jupyter Notebooks: Strengths and Weaknesses



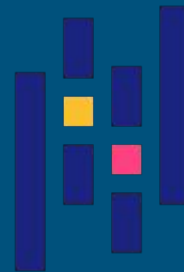
Good for ...

- Teaching!
- Drafting, exploring, experimenting
- Sharing your work with colleagues or stakeholders (e.g. convert Jupyter Notebook to a PDF or slideshow presentation)

... but not so good for:

- Developing applications
- Putting your applications into production
- Version control (e.g. [git](#))

What is Pandas? 🐼



Pandas: powerful Python package for **data analysis**

Advantages

- Developed for working with **structured** (tabular, relational, labeled) datasets (e.g. spreadsheet-style datasets)
- Developed for working with **time series** data (e.g. weather measurements, monthly sales, stock market prices, sensor data)
- Supports importing and exporting data from a wide range of common **file formats** used for data analysis (e.g. CSV, Excel, SQL, JSON, HTML)

Disadvantage

- **pandas** can have slow performance when using large datasets; consider using polars for larger datasets

What is Matplotlib?

- Matplotlib is a comprehensive package for creating static, animated, and interactive visualizations in Python. We can plot with `matplotlib` directly from `pandas`! 🙌
- Seaborn is a Python **data visualisation** package based on `matplotlib`.
- `seaborn` provides a high-level interface for drawing attractive and informative statistical graphics.



Hands-On Example



Dataset: Palmer Penguins



Hands-On Example



The [palmerpenguins](#) dataset by Allison Horst, Alison Hill, and Kristen Gorman was first made publicly available as an R package.

The goal of the Palmer Penguins dataset is to replace the highly overused [Iris](#) dataset for data exploration & visualization.

- 344 penguins
- 3 penguin species (Adélie, chinstrap, and gentoo)

Image: Dr. Kristen Gorman in the field, surrounded by penguins, at islands near Palmer Archipelago, Antarctica



Next Steps



Modeling Data with Python

- How to **model** data
- **Multivariate** modeling
- Build a simple **web app**

We'll build an **interactive web dashboard** like [this](#)!

👁️ Further Resources 👁️

Essential Resource

- [Pandas Cheat Sheet](#)

Basic Resources

- [Data Analysis with Python certificate with FreeCodeCamp.org](#): a free certification!
- [Computer Programming for Everybody](#): a manifesto by Guido van Rossum, inventor of Python
- [The Untold Story of Palmer Penguins](#): from the creators of [palmerpenguins](#)

Advanced Resources

- [The Python Tutorial](#): by the Python Software Foundation
- [Data Analysis Examples](#) from *Python for Data Analysis* by Wes McKinney, inventor of Pandas!