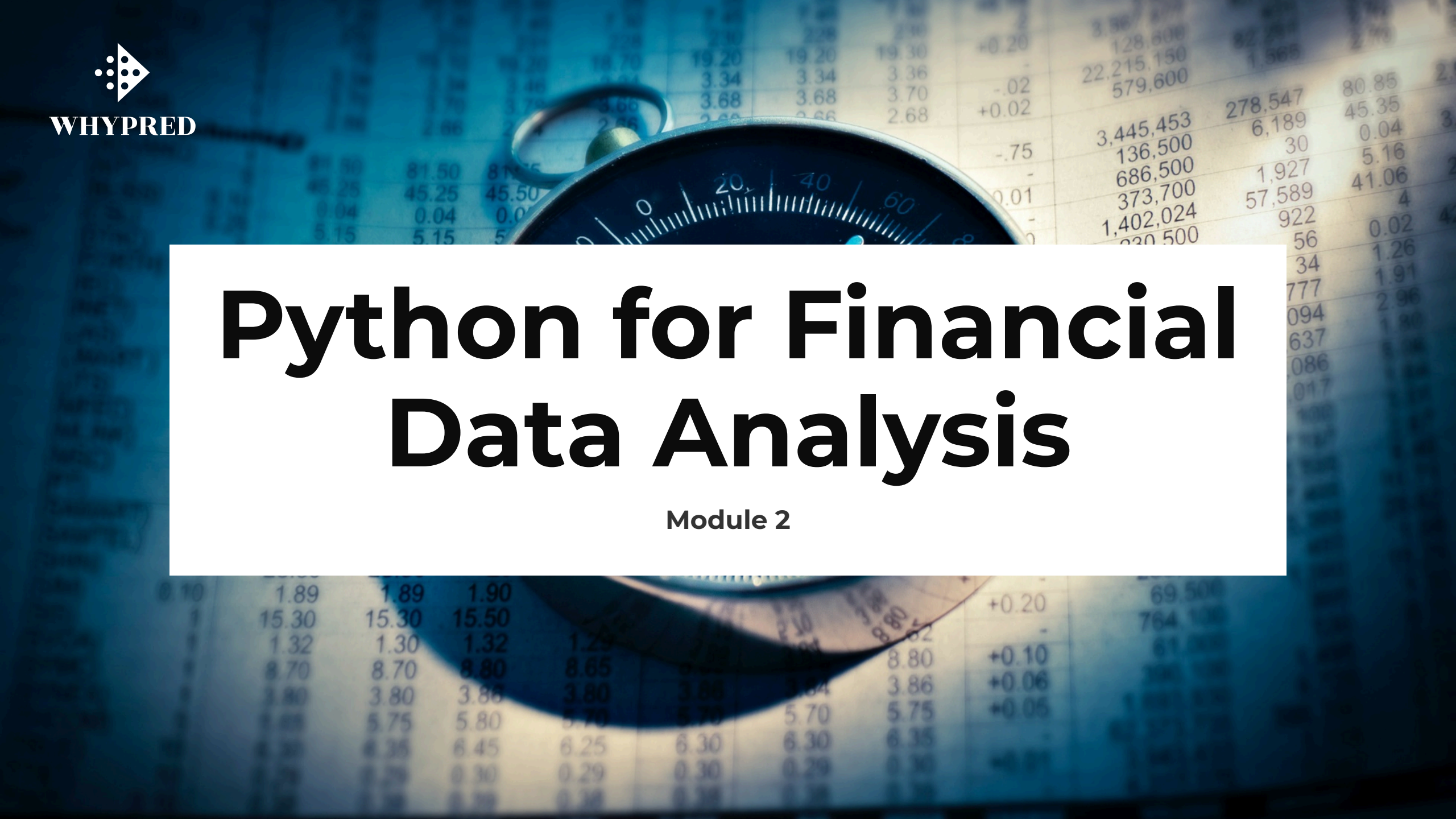

WHYPRED
Python for Financial Data Analysis
Module 2

Session Map

WHYPRED

1 **Recap**

Colab + Python fundamentals

2 **Data Formats**

csv + txt + json + yaml + api + xlsx

3 **The Dataframe**

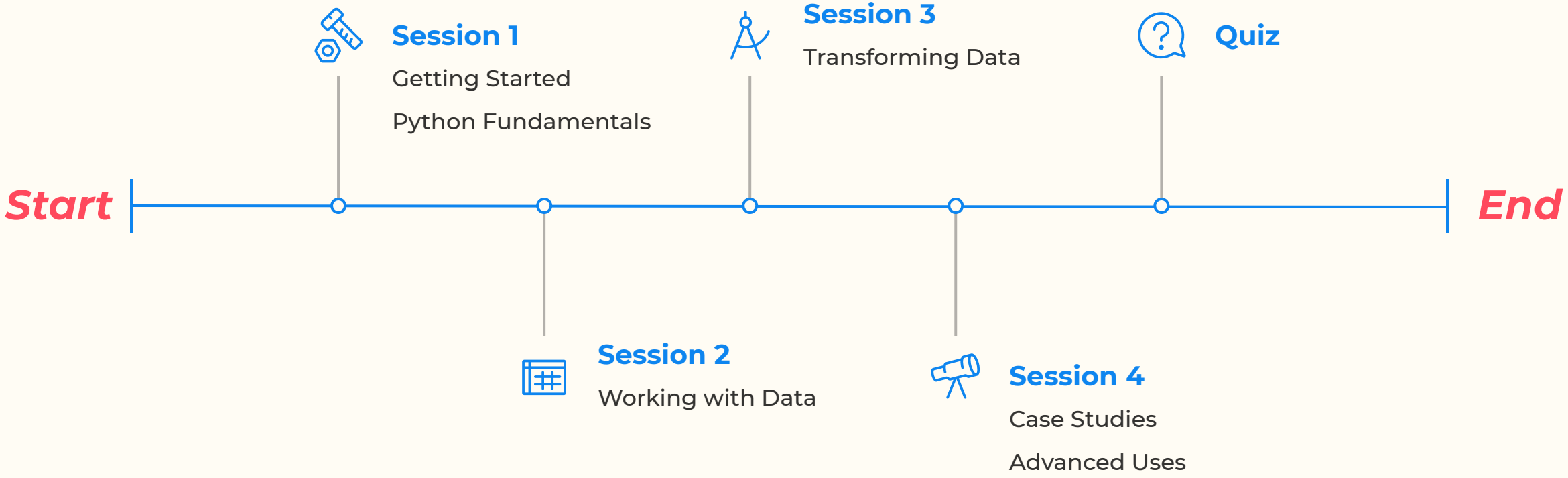
Tabular Data in Python

4 **Dataframe Operations**

Working with data frames

Course Outline

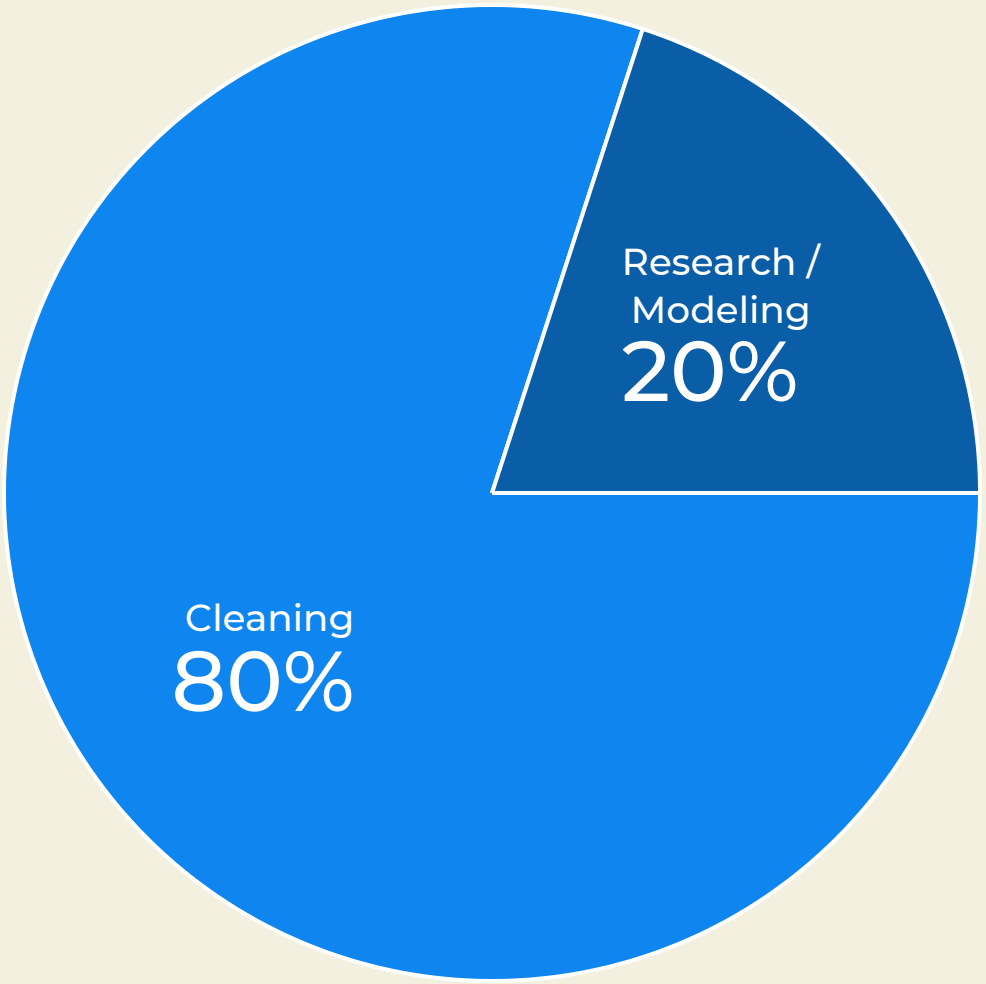
WHYPRED



Financial Data Analysis

WHYPRED

Analyst Effort



Recap

WHYPRED

- Python Quirks
- Variables
- Data Structures
- Control Flows
- Functions

WHYPRED

Data Formats

Data Storage Formats

WHYPRED

There are different ways that data can be stored

The most common is of course the 'csv' file or the spreadsheet

Other file formats include

- XLSX
- JSON
- YAML
- Parquet

Comma Separated Values (CSV)

WHYPRED

- It is a simple file format used to store tabular data, such as a spreadsheet or database.
- Each line in a CSV file corresponds to a row in the table, and each field in that row is separated by a comma
- the contents of a csv file might look like this but it can be opened in spreadsheets
- *Date,Transaction,Amount*

  2023-01-01,Application,1000

  2023-01-02,Redemption,-200

  2023-01-03,Application,500

Comma Separated Values (CSV)

WHYPRED

Features

- **Simplicity**: CSV files are straightforward to create and read. They can be opened with any text editor or spreadsheet software like Microsoft Excel or Google Sheets.
- **Compatibility**: CSV is a widely accepted format and can be used across different platforms and software applications. This makes it easy to share financial data between different systems.
- **Human-Readable**: Since CSV files are plain text, they can be easily read and understood by humans. This is particularly useful for quick inspections and debugging.

Comma Separated Values (CSV)

WHYPRED

Drawbacks

- **Lack of Data Types**: CSV files do not store data types. All data is treated as text, which can lead to issues when importing into systems that require specific data types (e.g., dates, numbers).
- **No Support for Complex Data**: CSV files are not suitable for storing complex data structures like nested records or hierarchical data.
- **Limited Metadata**: CSV files do not contain metadata (data about data), which means additional context about the data (e.g., units of measurement, data source) must be documented separately.
- **Inefficiency with Big Data**: CSV are not are ideal for the storage and querying of large datasets.

JavaScript Object Notation (JSON)

WHYPRED

- JSON (JavaScript Object Notation) is a lightweight data interchange format that is easy for humans to read and write,
- it is easy for machines to parse and generate. I
- it is recognisable by the use of curly brackets. JSON is built on two structures:
    1. **A collection of key/value pairs**
    2. **An ordered list of values**

JavaScript Object Notation (JSON)

WHYPRED

Does this sound familiar?

it should because it is exactly like python's dictionary data structure! JSON is one of the most commonly used formats in programming and for sending data over the internet.

*{ **"fund"**: "Jane Doe Capital",*

* **"horizon"**: 7,*

* **"is_open"**: true,*

* **"fum"**: 4000000,*

* **"address"**: { **"street":** "123 Main St", **"city"**: "Anytown", **"state"**: "CA" },*

* **"phone_numbers"**: ["123-456-7890", "987-654-3210"]*

* }*

JavaScript Object Notation (JSON)

WHYPRED

Features

Structured Data Representation: JSON's ability to represent hierarchical data

Interoperability: JSON is language-independent including making it a versatile format for exchanging data between different systems and applications.

Ease of Use: JSON is easy to read and write, which simplifies the process of data entry and review.

APIs and Web Services: Many financial services and APIs use JSON as their primary data format for communication.

Yet Another Markup Language (YAML)

WHYPRED

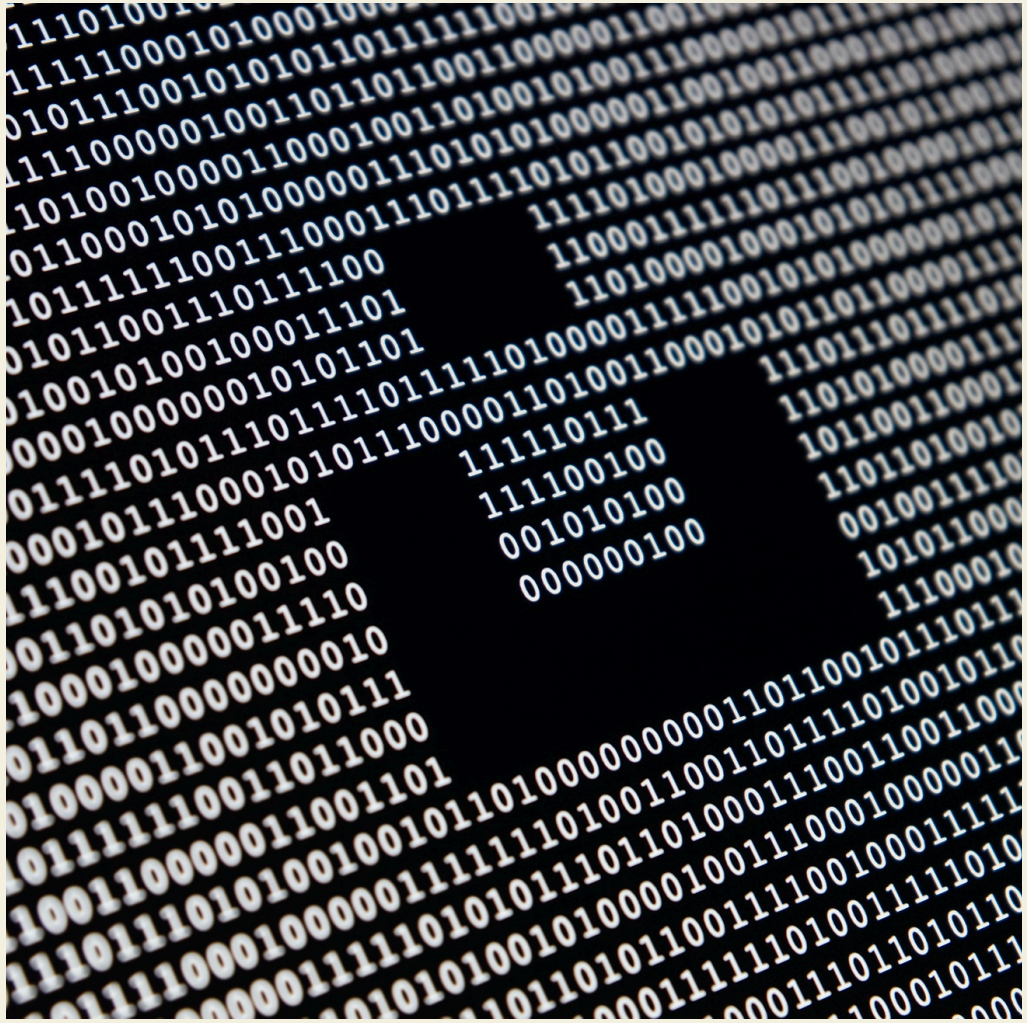
- YAML is a human-readable data serialization standard that is commonly used for configuration files in software applications, defining API and web service specifications, storing and organizing metadata.
- Their human-readable format, flexibility, and widespread adoption across various domains make YAML a versatile language.
- ***baseCurrency: USD***

  exchangeRates:
  - currency: EUR
  rate: 0.92
  description: Euro

  - currency: GBP
  rate: 0.81
  description: British Pound

Parquet

WHYPRED



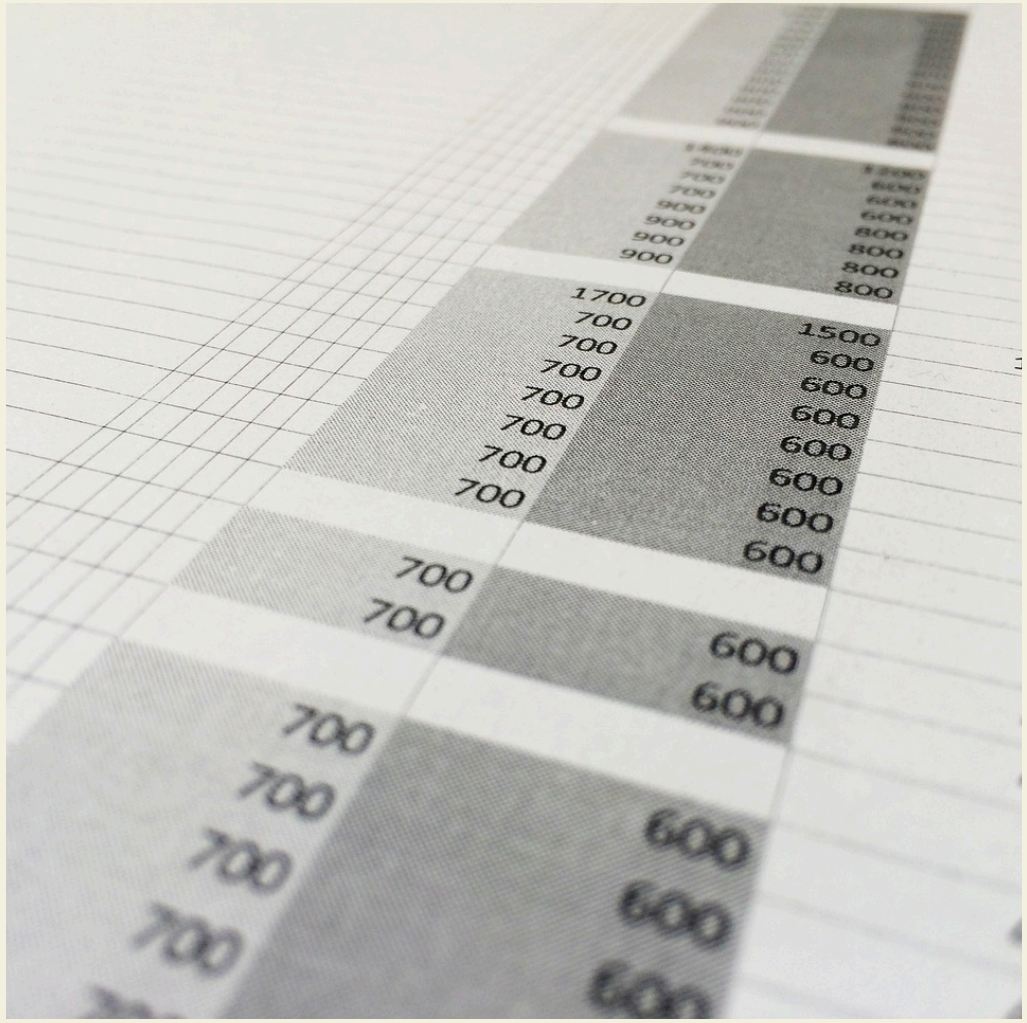
- Parquet is a columnar storage file format optimized for use with big data processing frameworks.
- It is designed to bring efficiency compared to traditional row-based file formats like CSV and JSON.
- What this means is that in parquet the file is read column by column instead of row by row, this saves a lot of time especially you know what columns you want to work with
- The parquet file format is not human readable and is stored in binary format.

WHYPRED

Dataframes

What's a Dataframe

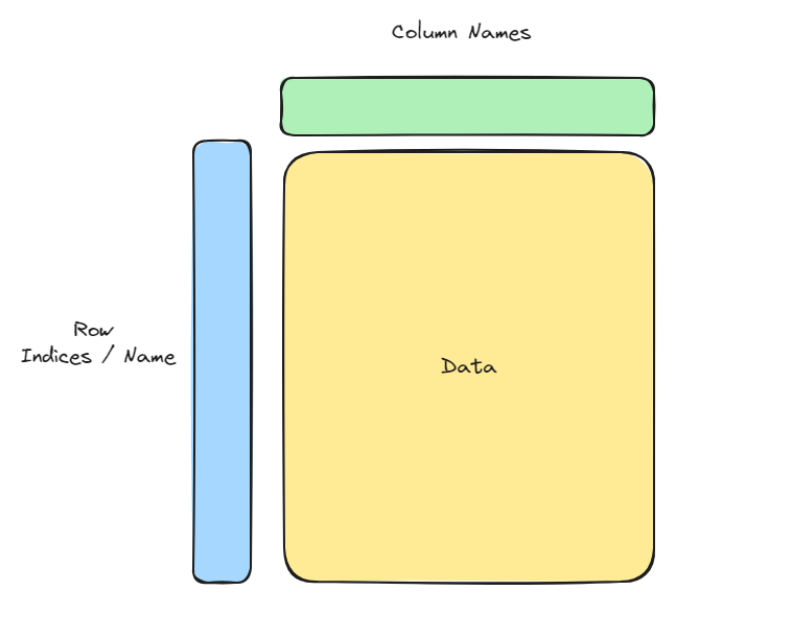
WHYPRED



- in Python dataframes are a powerful data structure provided by the pandas library.
- They are used for storing and manipulating tabular (table-like) data
- A DataFrame is a two-dimensional labeled data structure with columns of potentially different types, similar to a table in a database or an Excel spreadsheet.

What's a Dataframe

WHYPRED



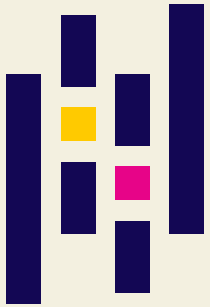
	ticker	sector	mer	fum	one_month_return
2	A200	Equity - Australia	0.04	4,296.27	1.48
3	IOZ	Equity - Australia	0.05	5,122.88	1.34
4	ILC	Equity - Australia	0.24	596.46	1.88
5	MVW	Equity - Australia	0.35	2,098.95	0.36
6	QOZ	Equity - Australia	0.4	500.09	2.24
7	STW	Equity - Australia	0.05	4,987.77	1.19

Pandas Package

WHYPRED

A package is a collection of modules and that provide pre-written code to help analysts perform common tasks. Reducing the need to write code from scratch. We will use the pandas library to help us work with dataframes.

We can use package in our code by calling the ***'import <library name>'*** command



Pandas

Pandas is a powerful library for data manipulation and analysis. It provides data structures like DataFrames and Series, which make it easy to work with structured data, such as financial time series data.

Financial Data Analysis

WHYPRED

Some common functions to wrangle/clean dataframes include:

- Summarise data - ***df.sum(), df.count(), df.max()***
- Explore Data - ***df.size, df.shape, df.dtypes, df.columns***
- Slicing data - ***df.loc()*** and ***df.iloc()***
- Selecting columns with column names - ***df[["colname1", "colname2"]]***
- Filling missing values - ***df.fillna()***
- Filtering data using values - ***df[df["colname" > 100]]***

Notebook Exercise

WHYPRED

2-1_working_with_datafrmaes.ipynb

- All notebooks and slides are available here
- Remember Google Colab is a shared cloud service, everyone is looking at the same notebook!
- To prevent accidental changes the notebooks are read only, at the beginning of each exercise make sure you create a copy so that you can edit your own copy!

Notebook Exercise

WHYPRED

Functions are reusable chunks code of code that are defined using the ***def*** keyword

e.g. imagine we want to get the average monthly sales of business over 3 months, in python the code might look something like this:

```
total_sales = 100 + 200 + 300

average_sales = total_sales / 3
```

But we have to repeat the code for each quarter, solution:

```
def avg_sales(sales_m1, sales_m2, sales_m3):
    result = (sales_m1 + sales_m2 + sales_m3) / 3
    return(result)
```

That's it for module 2!

Disclaimer

WHYPRED

The information contained in this slide pack has been prepared by WhyPred Pty Ltd ("WhyPred") for informational purposes only. The data, analysis, and opinions expressed herein are based on sources believed to be reliable and provided in good faith, but no representation or warranty, express or implied, is made as to its accuracy, completeness, or correctness.

In addition, sample data has been used in this presentation and should not be relied upon for accuracy. This presentation does not constitute investment advice, nor is it an offer or solicitation to buy, hold, or sell any securities or financial instruments. Any projections, forecasts, or estimates herein are indicative only and subject to change without notice. Past performance is not indicative of future results.

Investors should seek their own independent financial, legal, and tax advice before making any investment decision. WhyPred and its affiliates, directors, employees, or agents accept no liability whatsoever for any loss or damage arising from any use of this document or its contents. By accessing and using this slide pack, you agree to the terms of this disclaimer.