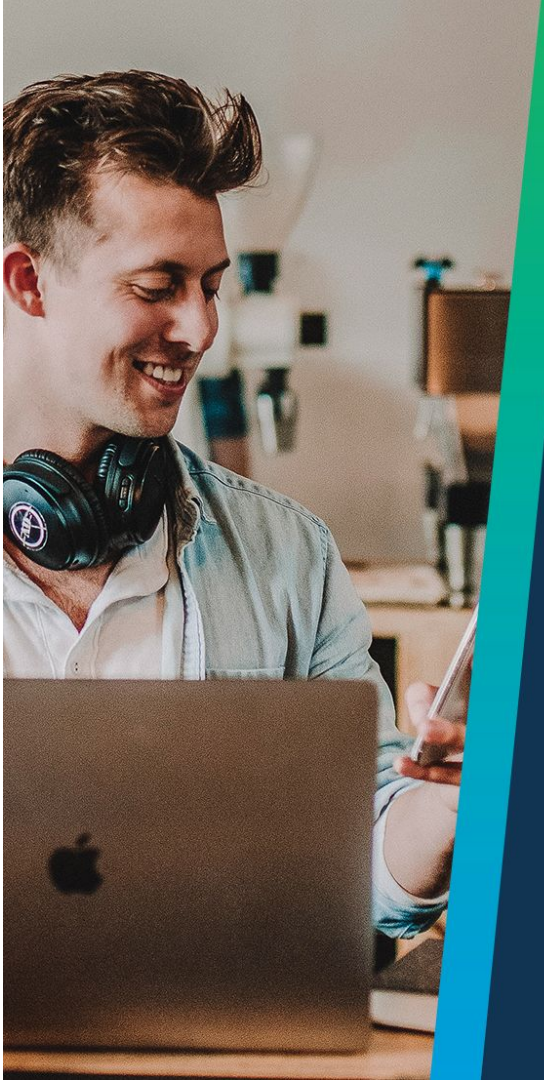


Data Science Bootcamp

Hyperiondev

Working with Datasets

WELCOME TO THE EVENT HANDLING TASK

Your Lecturer for This Session

Christiaan Joubert

Lecture – Housekeeping

- The use of disrespectful language is prohibited in the questions, this is a supportive, learning environment for all - please engage accordingly.
- No question is daft or silly - **ask them!**
- There are Q/A sessions midway and at the end of the session, should you wish to ask any follow-up questions.
- You can also submit questions here: hyperiondev.com/sbc4-ds-questions
- For all non-academic questions, please submit a query: hyperiondev.com/support
- Report a safeguarding incident: hyperiondev.com/safeguardreporting
- We would love your feedback on lectures: https://hyperionde.wufoo.com/forms/zsgv4m40ui4i0g/

Lecture - Code Repo

Go to: github.com/HyperionDevBootcamps

Then click on the "**C4_DS_lecture_examples**" repository, do view or download the code.

Objectives

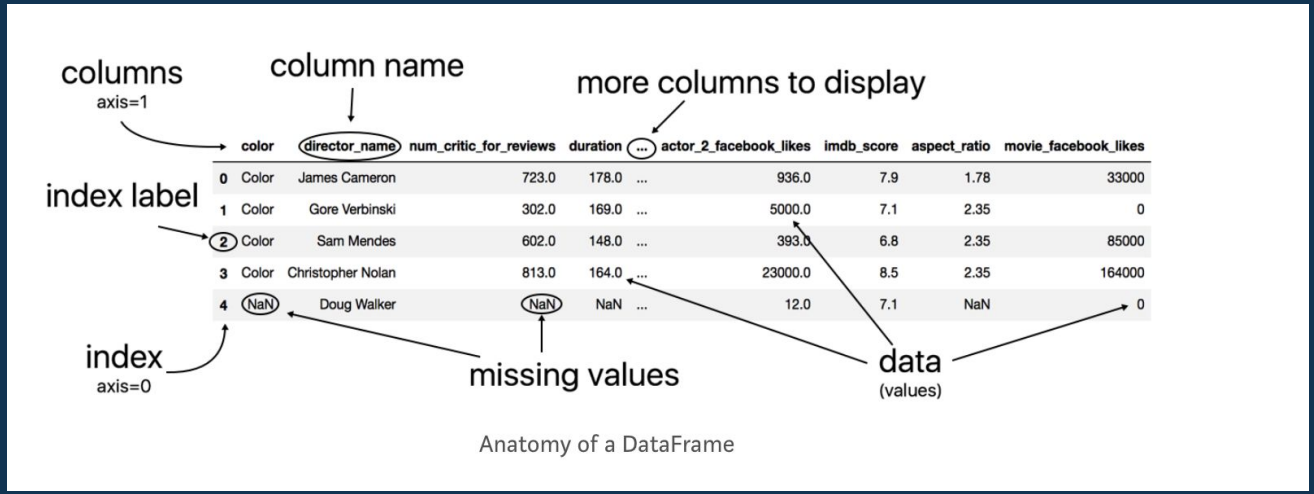
1. Learn how to read and manipulate data using Pandas
2. Generate graphs in Python using Matplotlib and Seaborn

How to Access a Dataset in Python

- **Pandas**
  - Basically SQL for Python (without actually using a database).
  - Entire dataset is stored in RAM.
  - pip install pandas
  - import pandas as pd

Dataframes

- Very fancy word for what is basically a 2D array.
- Pandas notation makes navigating large datasets very easy.



	color	director_name	num_critic_for_reviews	duration	...	actor_2_facebook_likes	imdb_score	aspect_ratio	movie_facebook_likes
0	Color	James Cameron	723.0	178.0	...	936.0	7.9	1.78	33000
1	Color	Gore Verbinski	302.0	169.0	...	5000.0	7.1	2.35	0
2	Color	Sam Mendes	602.0	148.0	...	393.0	6.8	2.35	85000
3	Color	Christopher Nolan	813.0	164.0	...	23000.0	8.5	2.35	164000
4	NaN	Doug Walker	NaN	NaN	...	12.0	7.1	NaN	0

Anatomy of a DataFrame

Reading Data into a Dataframe

- To read from a CSV file:
  - pd.read_csv('credit.csv', delimiter = ',')
- However, we are not limited to CSV files:
  - We also have read_excel(), read_sql(), among others.

Selecting Columns

```
import pandas as pd
import seaborn as sns

iris_df = sns.load_dataset('iris')
print(iris_df.columns)  #  ['sepal_length',  'sepal_width',  'petal_length',  'petal_width',
'species']
just_the_species = iris_df['species']
sepal_and_petal_info = iris_df[['sepal_length', 'sepal_width', 'petal_length',
'petal_width']]
```

(NOTE: we are just loading a dataset from Seaborn. You will learn more about this module later)

- Just like with a dictionary, you can specify keys.
- To select multiple columns, use a list of strings.

Advanced Searches in Pandas

```
small_sepal_length = iris_df[iris_df['sepal_length'] < 4.2]
```

- This is basically a way of saying “Give me all rows where the sepal_length is less than 4.2”.
- Similar to the following SQL statement (which can be read in Human Language):
- ```
  SELECT * FROM Iris
  WHERE sepal_length < 4.2
  ```
  ```

In-built Dataframe Methods

- **mean()**
  - Mean for a particular column
- **min()**
  - Minimum value for a particular column
- **max()**
  - Maximum value for a particular column
- **std()**
  - Standard Deviation
- **var()**
  - Variance
- **nunique()**
  - Number of unique values in a column

Grouping Data

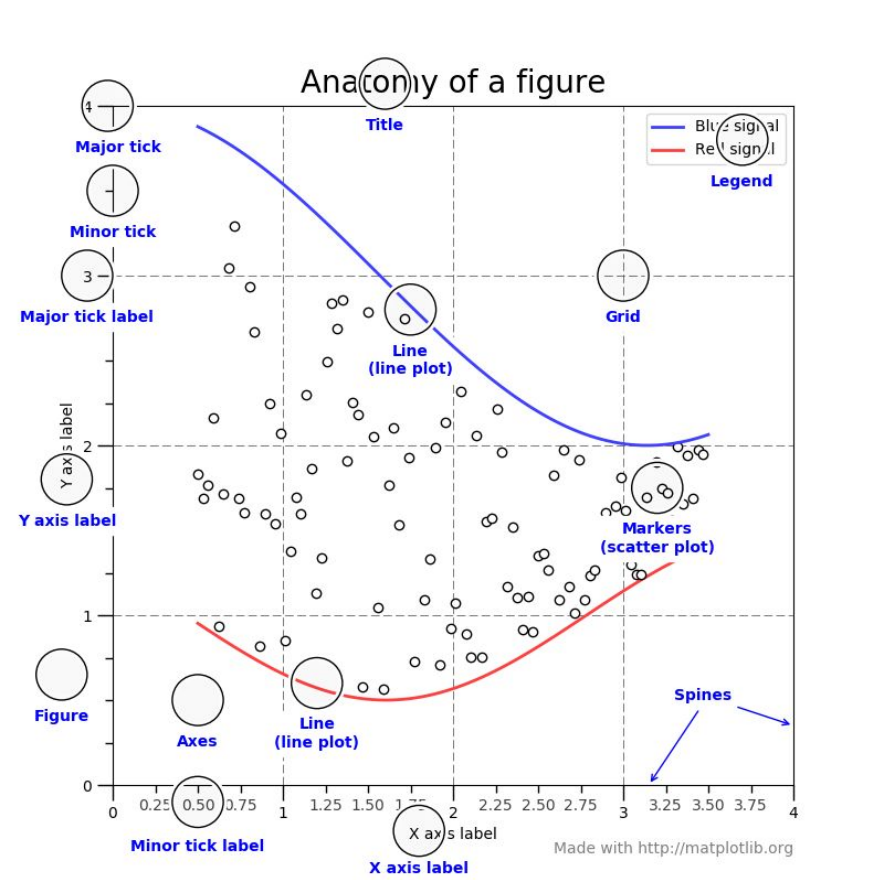
- Sometimes, grouping up data can be valuable.
- For example, “Show me the average insurance charge per age group”.
- This could end up being a lot of lines of code.
- However, in Pandas, you only need one line of code:
  - `insurance_df.groupby('age')['charges'].mean()`

How to Make Graphs in Python

- **Matplotlib**
  - Python package for data visualisation
  - Use code to create graphs
- To install: pip install matplotlib
- To use: import matplotlib.pyplot as plt

Terminology of a Graph

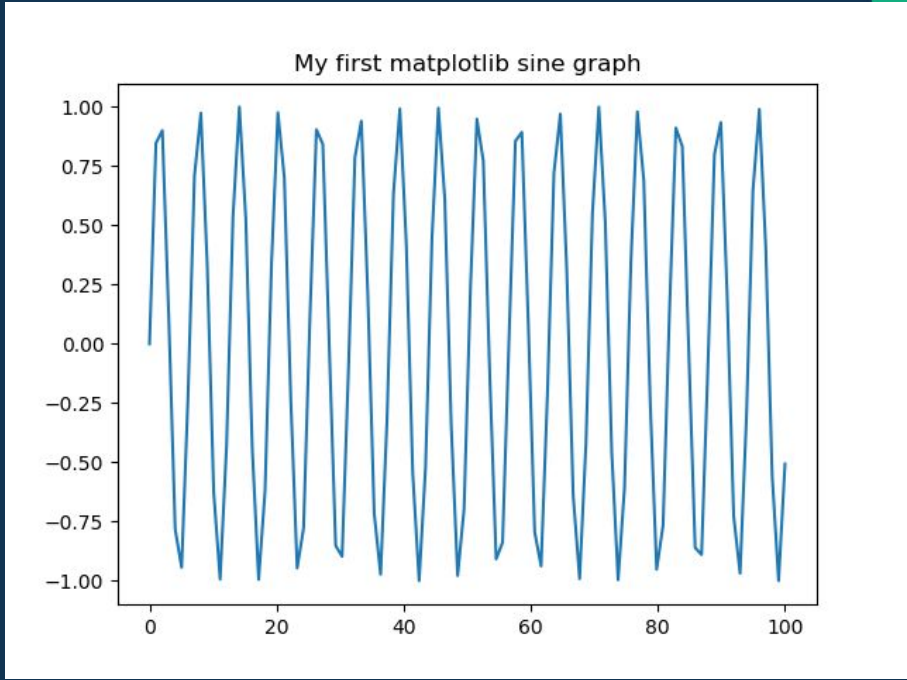
- Of importance:
  - Line
  - Ticks
  - Axes
  - Labels
  - Title
  - Legend



Let's Create a Sin Graph

```
import matplotlib.pyplot as plt
import numpy as np

#Prepare the data
x = np.linspace(0, 100, 100) #x axis
y = np.sin(x) # y values
#Plot the data
plt.plot(x, y, label="sine")
#Create a title
plt.title("My first matplotlib sine graph")
#Show the plot
plt.show()
```



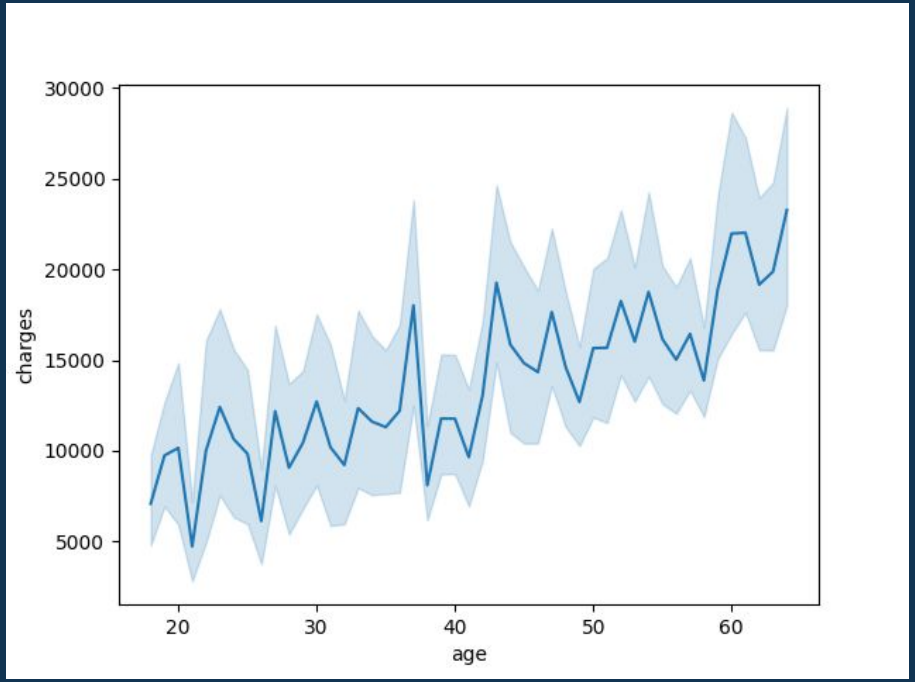
Seaborn

- Basically Matplotlib, but better.
- Built on Matplotlib.
- Allows you to create complex visualisations with extreme ease.
- Integrates well with Pandas.
- pip install seaborn
- import seaborn as sns

Simple Seaborn Example: Lineplot

```
plt.figure()
sns.lineplot(x='age', y='charges', data=ins_df)
plt.savefig('sns.png')
plt.close()
```

- Note: The lineplot takes a Dataframe as the data argument, then column names in x and y arguments.



Other Seaborn Graphs

- **histplot()**
  - Histogram
- **barplot()**
  - A bar chart (different to a histogram)
- **boxplot()**
  - Box plot showing median, range, etc.
- **Many others**
  - The list just goes on and on . . .

Hyperiondev

Q & A Section

Please use this time to ask any questions relating to the topic, should you have any.

Hyperiondev

Thank You for Joining Us