Hyperiondev

TASK

Working with Datasets

Visit our website

Introduction

WELCOME TO THE WORKING WITH DATASETS TASK!

In the context of this task, when we refer to a dataset, we are referring to a collection of related data. This data can be manipulated in various ways programmatically. In this task, you will be using Pandas DataFrames to manipulate data.

Get in touch

Connect for support

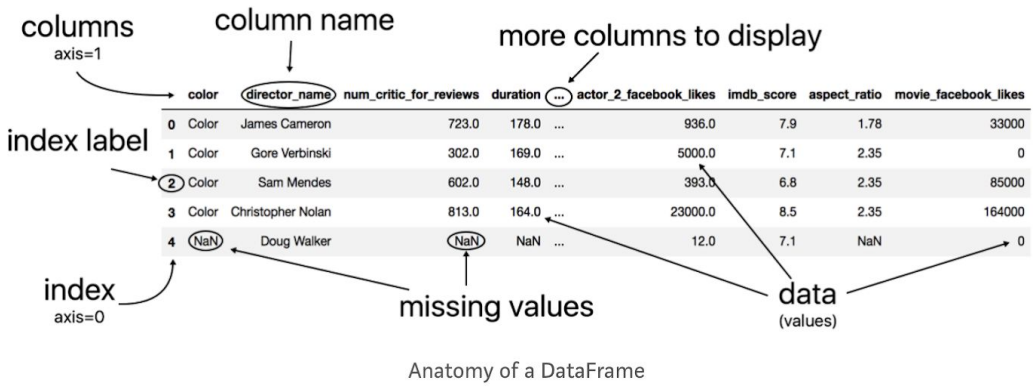
Remember that with our courses, you're not alone! You can contact your mentor to get support on any aspect of your course.

The best way to get help is to login to **www.hyperiondev.com/portal** to start a chat with your mentor. You can also schedule a call or get support via email.

Your mentor is happy to offer you support that is tailored to your individual career or education needs. Do not hesitate to ask a question or for additional support!

WHAT IS A PANDAS DATAFRAME?

As you learnt in the previous task, the **pandas' library documentation** defines a DataFrame as a "two-dimensional, size-mutable, with labelled rows and columns".



Anatomy of a DataFrame

Image source: (Petrou, 2017)

In simple terms, think of a DataFrame as a table of data with the following characteristics (Lynn, 2018):

- "There can be multiple rows and columns in the data.
- Each row represents a sample of data,
- Each column contains a different variable that describes the samples (rows).
- The data in every column is usually the same type of data – e.g. numbers, strings, dates.
- Usually, unlike an excel data set, DataFrames avoid having missing values, and there are no gaps and empty values between rows or columns."

In the previous task you also learnt to read data from a .csv file into a DataFrame using the `read_csv()` function as shown below:

```
pd.read_csv('credit.csv', delimiter = ',')
```

However, there are also other functions that can be used to read data from different sources into DataFrames. For example, **`read_excel()`** can be used to read data from a spreadsheet file into a DataFrame, and **`read_sql()`** can be used to load data from a SQL database. Sometimes, as was done in the previous task, it is easier to extract data from other sources into a .csv file and reading it into a DataFrame.

JUPYTER

In this compulsory task, you will be using the Jupyter Notebook. This tool is **described as follows**: "The Jupyter Notebook is an open-source web application that allows you to create and share documents that contain live code, equations, visualisations and narrative text. Uses include data cleaning and transformation, numerical simulation, statistical modelling, data visualisation, machine learning, and much more."

To use this tool, do the following:

1. **Install Jupyter**
    - **Option 1: Installing Jupyter with pip**

      First, ensure that you have the latest pip; older versions may have trouble with some dependencies:

      ```
      pip3 install --upgrade pip
      ```

      Then install the Jupyter Notebook using:

      ```
      pip3 install jupyter
      ```

    - **Option 2: Installing Jupyter using Anaconda**
        - Download **Anaconda**. We recommend downloading Anaconda's latest Python 3 version.
        - Install the version of Anaconda which you downloaded, following the instructions on the download page.

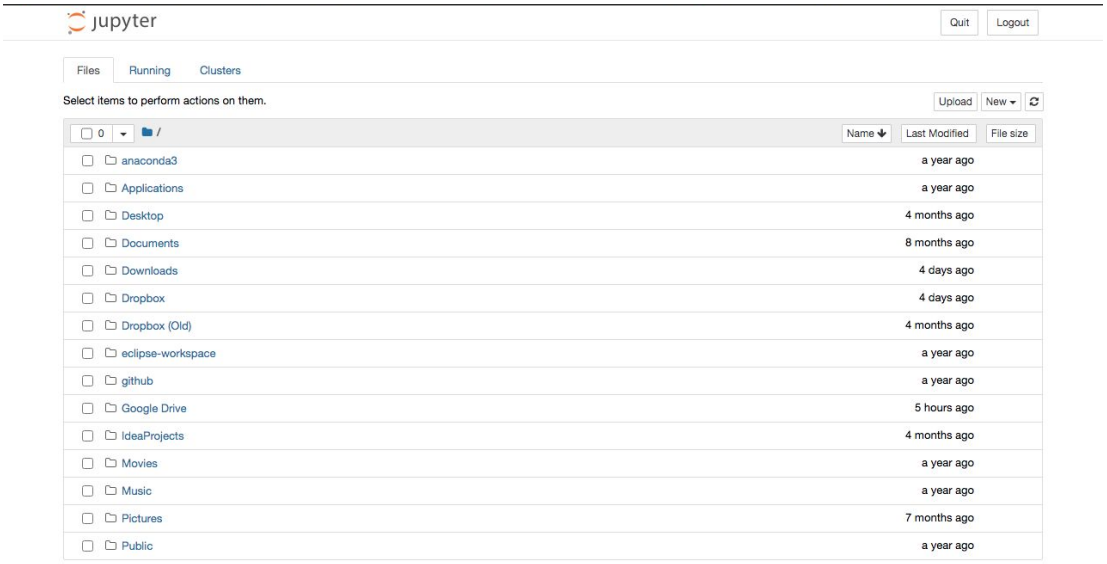
2. **Run the Jupyter notebook**

   Once you have installed Jupyter, you can start the notebook server from the command line:

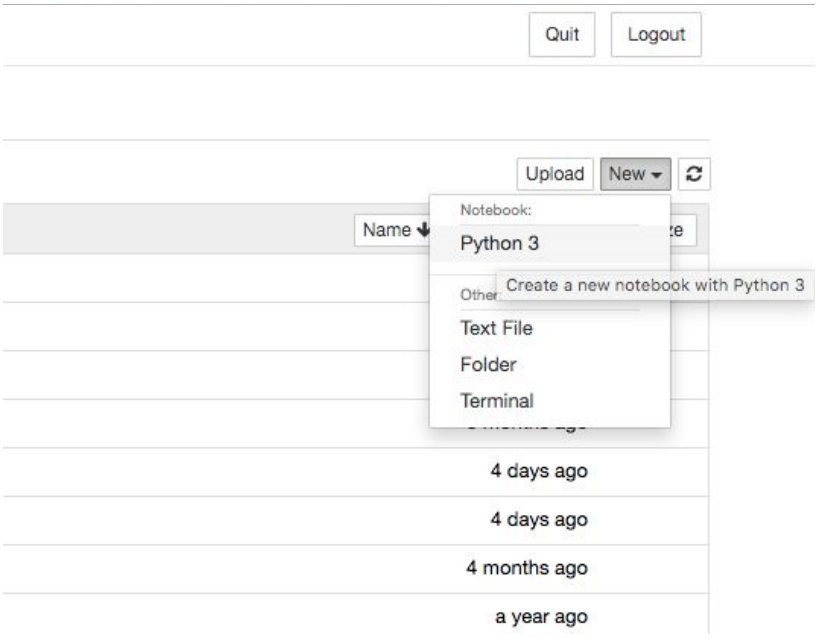
   ```
   jupyter notebook
   ```

   This will print some information about the notebook server in your terminal, including the URL of the web application. The notebook will then open in your browser.

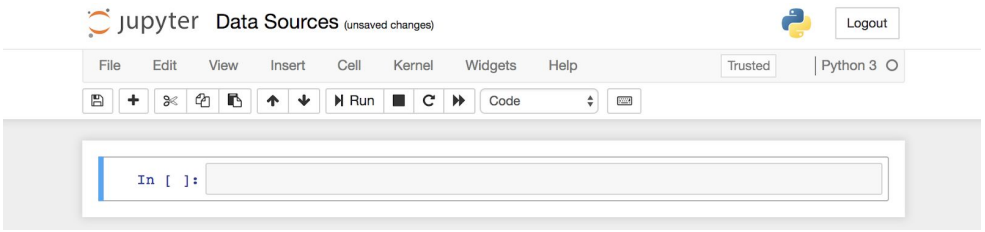
Once the notebook has opened, you should see the dashboard showing the list of notebooks, files and subdirectories in the directory you've opened. You can see an example of a Jupyter notebook below:



To start a new notebook, click the *New* drop-down menu and click on *Python 3*.



A new Jupyter notebook will look like the screenshot below. Make sure to change the name of the notebook to Data Sources.



A note from our coding mentor

Ridhaa

Additional reading: *For more information about working with Jupyter, please consult the first chapter ("**IPython: Beyond Normal Python**") in the book entitled, "**Python Data Science Handbook**" by Jake VanderPlas.*

Compulsory Task 1

Follow these steps:

- Follow the instructions in this task to install Jupyter Notebook.
- In your command line interface, change directory (`cd`) to the Dropbox folder that contains this task.
- Open Jupyter notebook by typing: `jupyter notebook`
- Within this task folder, you will find a Jupyter Notebook named **datasources.ipynb**. You can open it by going to Jupyter's home screen and double-clicking on the notebook. The notebook will contain the rest of the content for this Task.

Compulsory Task 2

Open and run the example file for this task in IDLE before attempting this task. Follow these steps:

- Create a new Python file in this folder called **Report.py**.
- Create a DataFrame that contains the data in **balance.txt**.
- Write the code needed to produce a report that provides the following information:
    - Compare the average income based on ethnicity.
    - On average, do married or single people have a higher balance?
    - What is the highest income in our dataset?
    - What is the lowest income in our dataset?
    - How many cards do we have recorded in our dataset? (Hint: use `sum()`)
    - How many females do we have information for vs how many males? (Hint: use `count()`. For a list of all methods for computation of descriptive stats, see **here**.)

Completed the task(s)?

Ask your mentor to review your work!

Review work

Rate us
Share your thoughts

HyperionDev strives to provide internationally-excellent course content that helps you achieve your learning outcomes.

Think that the content of this task, or this course as a whole, can be improved? Do you think we've done a good job?

Click here to share your thoughts anonymously.

References:

Lynn, S. (2018). The Pandas DataFrame – loading, editing, and viewing data in Python. Retrieved from Shane Lynn: Pandas Tutorials: **https://www.shanelynn.ie/using-pandas-dataframe-creating-editing-viewing-data-in-python/**

Jupyter Team. (2015). Running the Notebook — Jupyter Documentation 4.1.1 alpha documentation. Retrieved 18 August 2020, from **https://test-jupyter.readthedocs.io/en/latest/running.html**

Petrou, T. (2017, October 27). Dissecting the anatomy of a DataFrame. (Packt>) Retrieved from Pandas Cookbook: Recipes for Scientific Computing, Time Series Analysis and Data Visualization using Python: **https://subscription.packtpub.com/book/big_data_and_business_intelligence/9781784393878/1/ch01lvl1sec12/dissecting-the-anatomy-of-a-dataframe**