# Career Connections – Data Science Volunteer Project

## Introduction to Pennines team

Welcome to the Pennines team! This team is aimed at providing experience to IQVIAns who may be relatively new or inexperienced with python and/or machine learning. The Pennines team will be supported by some more experienced volunteers.

The aims of Pennines team are as follows:

* Support inexperienced volunteers in upskilling on python and machine learning fundamentals.
* Explore a variety of open-source datasets.
* Evaluate a set of tools on these different datasets.
* Share our learnings with the rest of the volunteer team, and with IQVIA more broadly through [our project SharePoint site](https://quintiles.sharepoint.com/sites/CareerConnectionsModelStory/).

Volunteers who demonstrate a combination of good technical and collaborative skills can be invited to join the more advanced volunteer project teams.

## Before You Start

You will need to have done IQVIA’s course on Data Privacy.

* The [IQVIA Global Data Privacy 2024](https://iqvia.sumtotal.host/core/pillarRedirect?relyingParty=LM&url=app%2Fmanagement%2FLMS_ActDetails.aspx%3FActivityId%3D1888322%26UserMode%3D0) course on Learning Edge will introduce you to the main concepts involved in working with sensitive healthcare data, which will be important as we progress further with the project. Please share a certificate of completion with Rehan Ali, the project lead, as soon as you’ve completed this course.

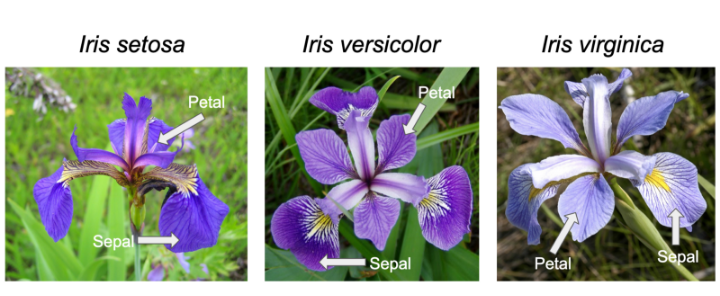
All datasets used by the Pennines team are classified as Level 1 (as they’re available to the public and have no identifiable patient information or other sensitive information within them).

## Getting Started with the Iris Dataset

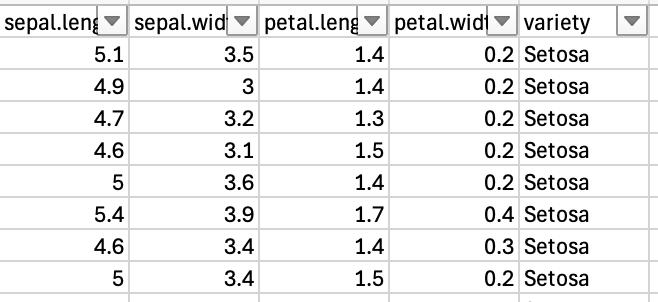
If you don’t have any experience with programming or machine learning, don’t worry. The aim of this part is to help you learn the basics.

* We will first learn by replicating what others have produced in the past. This way, you can see live examples of python code being used to do data science and machine learning, using a simple example dataset. You can play around with this and get familiar with the broad strokes of the techniques that we use.
* We’ll then apply these techniques to new healthcare datasets. This will be a chance for us all to learn something new.

Our first activity will be to work with a basic statistics training dataset called Iris. Iris is a flower classification dataset, containing measurements about three different types of Iris flower. The goal is to come up with a technique that can let you identify the flower species if you’re given those measurements up-front. Here are the different types of flower represented in the dataset:



Here is the [Iris dataset](https://gist.githubusercontent.com/netj/8836201/raw/6f9306ad21398ea43cba4f7d537619d0e07d5ae3/iris.csv) to download directly, if you’d like to explore it with a tool like Excel. Here’s what it looks like in Excel:



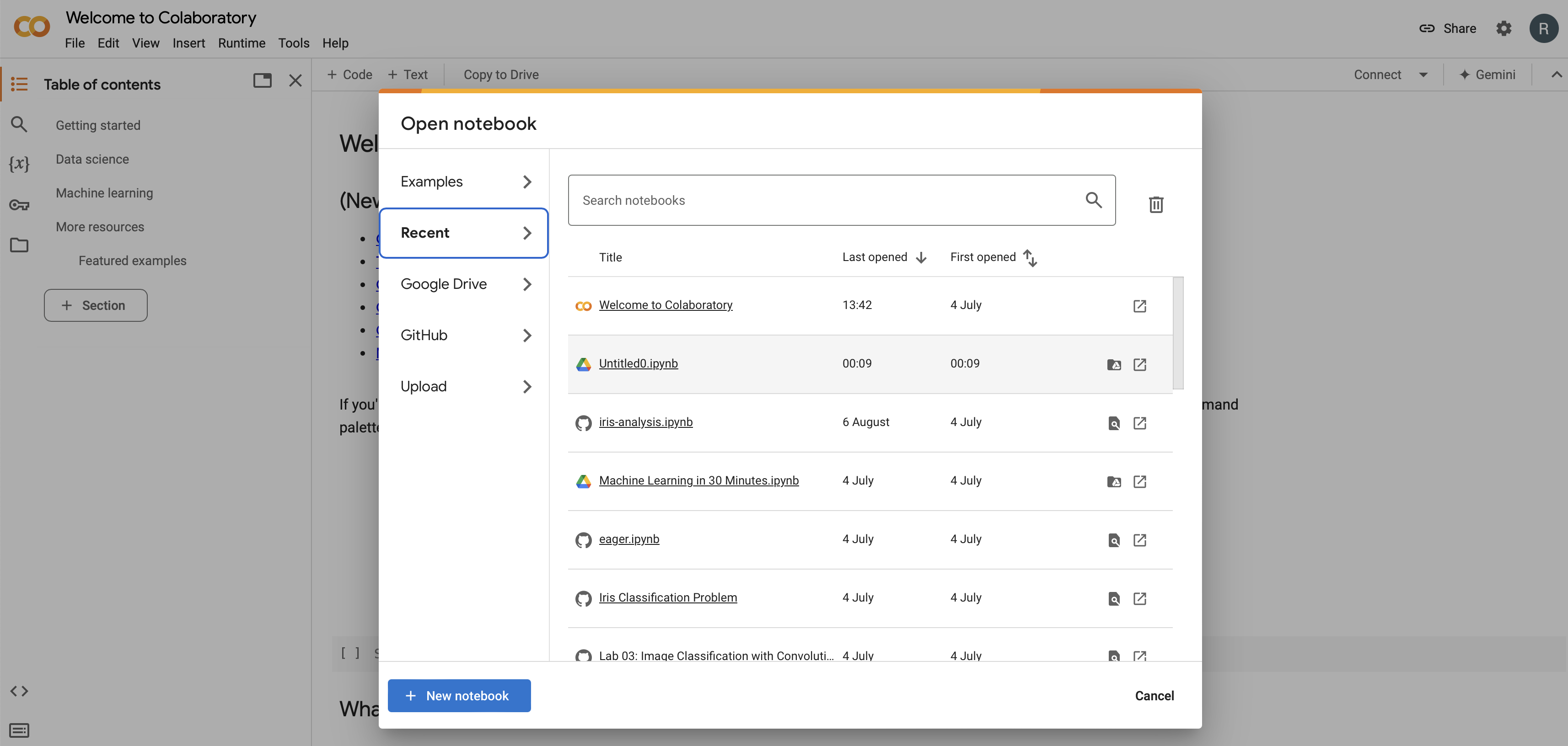
To start off, try and summarise the key properties of the three different varieties (setosa, versicolor and virginica) in Excel, using techniques that you’re already familiar with. What can you learn about them?

## Working with the Iris Dataset with Python

The next phase is to get some experience in working with this dataset using Python. For this, you will need a Google account. You can [create one here](https://support.google.com/accounts/answer/27441?hl=en).

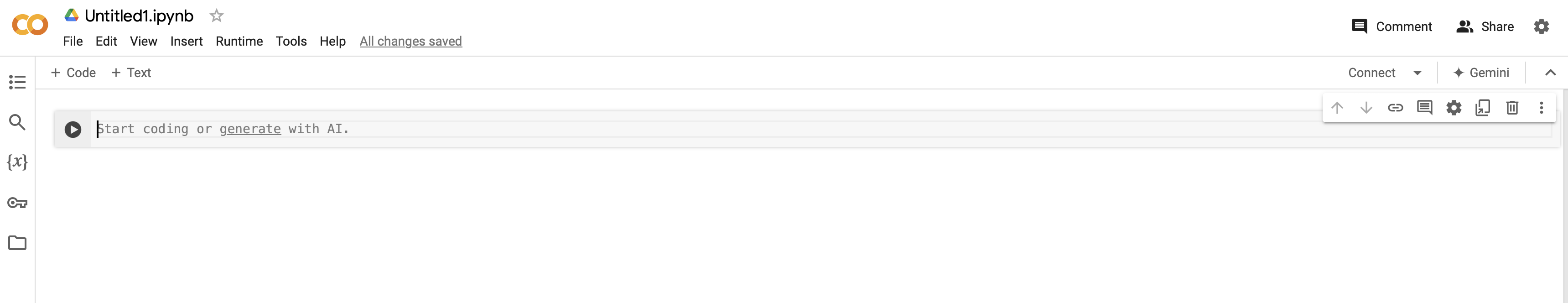
The reason for this is that we’ll be using an online tool called Google CoLab (which is short for Google Collaborative Laboratory). <https://colab.research.google.com>

CoLab is an amazing free tool that lets you run python code via your web browser on a cloud based computer. When you visit the page, you should see a page like this:



Click on the “New notebook” button at the bottom. This will launch a new browser window which contains a programming “notebook”. This is a special app that lets you mix text with code, so that you can run code and also write about it as you go along. (This is great for remembering what you did, and for letting other people understand your work.)

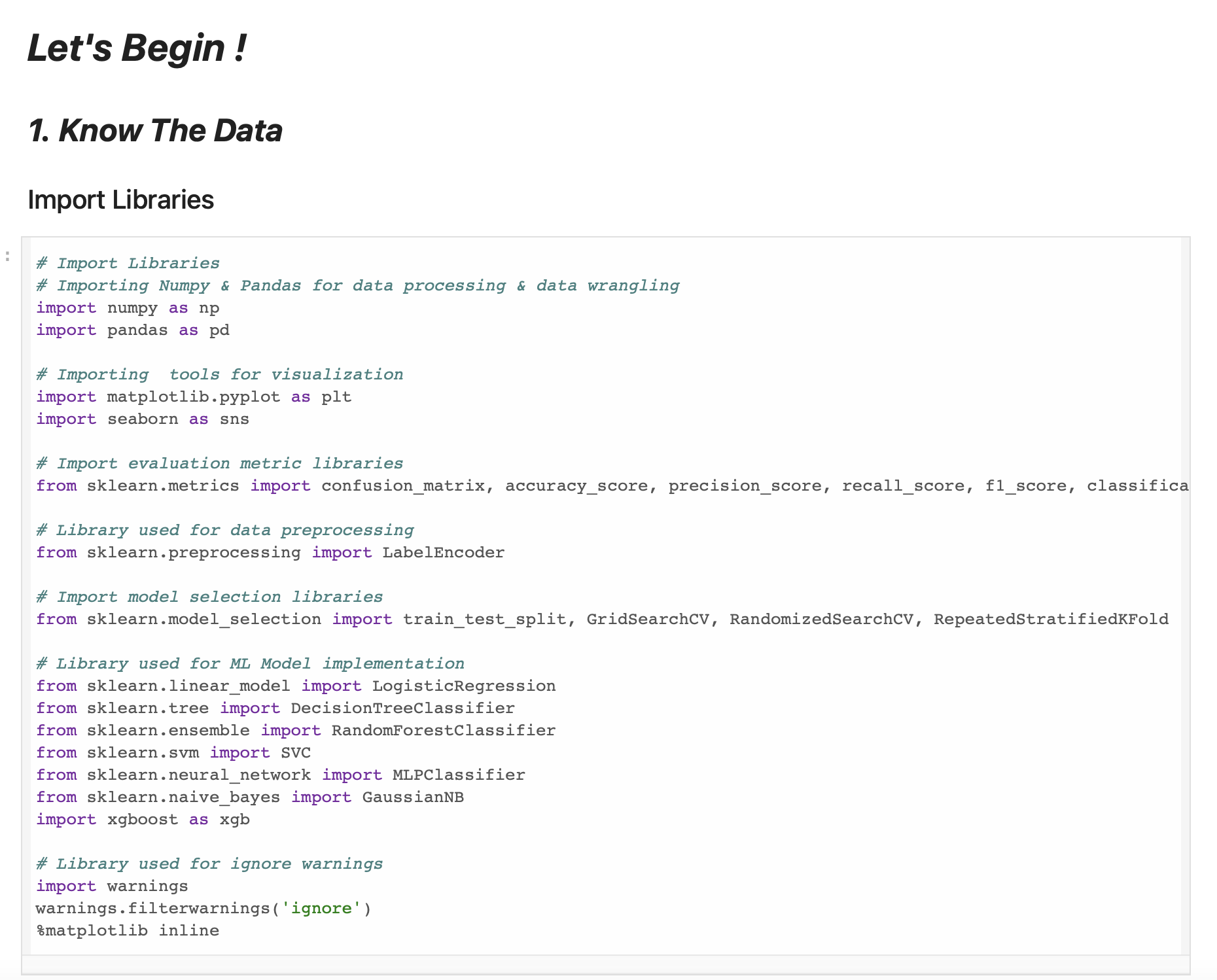
You’ll see a page like this:



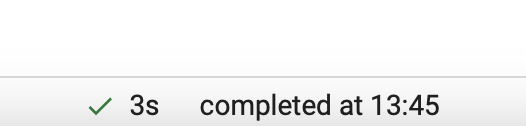
This is your notebook where you’ll be doing your work. Right now, you can type in text or code into “cells”, which are the grey shaded areas. You can create multiple cells, some with text and some with code.

Let’s put some code into our cell. We’re going to borrow some code from an existing notebook that someone has shared online. You can find this notebook here: <https://github.com/Apaulgithub/oibsip_taskno1/blob/main/Iris_Flower_Classification.ipynb>

Take the content of the first cell, shown below, and paste the code into your new notebook cell.

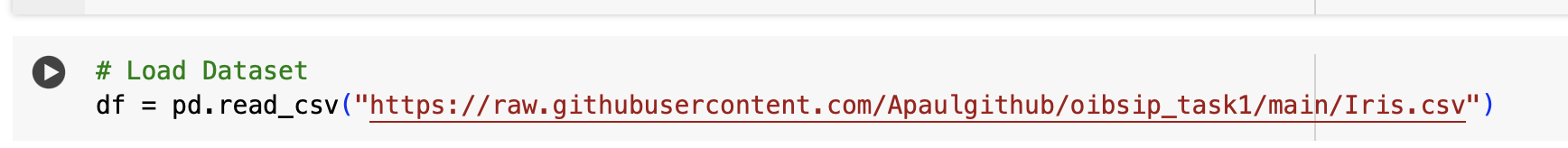


Now hit the run button on the left side of your cell. This will start a computer behind the scenes, and it’ll take some time to process. Eventually you’ll see a message at the bottom of your screen that says something like this:



If you see this, then great, the code ran successfully. Well done, you just ran your first python code for the project!

Now let’s take the next piece of code, paste it in a new code cell (by clicking the “+ Code” button on the top), and run it with the black icon on the left.



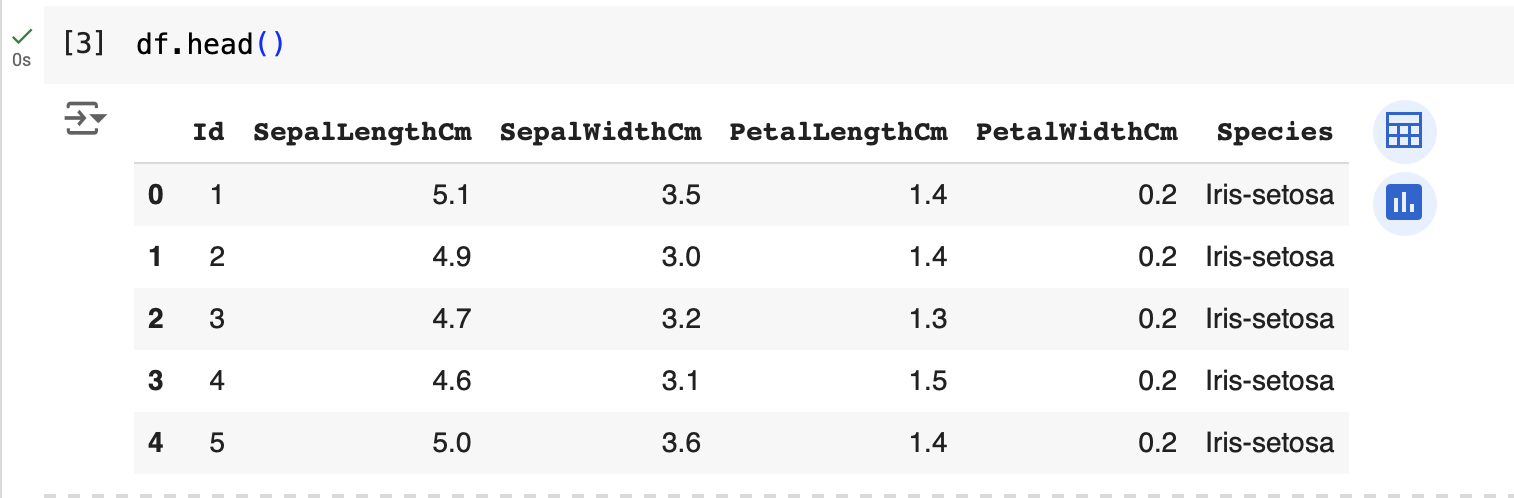
If that ran okay, you’ll see a little icon appear on the left of the cell, saying it ran successfully.



Now create a new cell below these two, and enter this code:

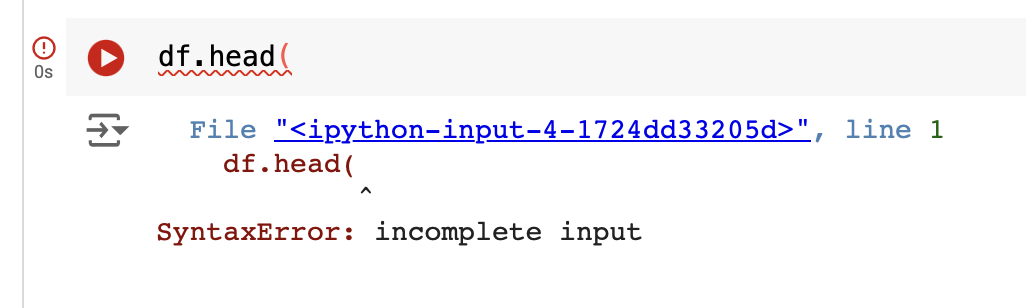
df.head()

When you run this, you should see a snippet (or the “head”) of the dataframe containing the Iris dataset.



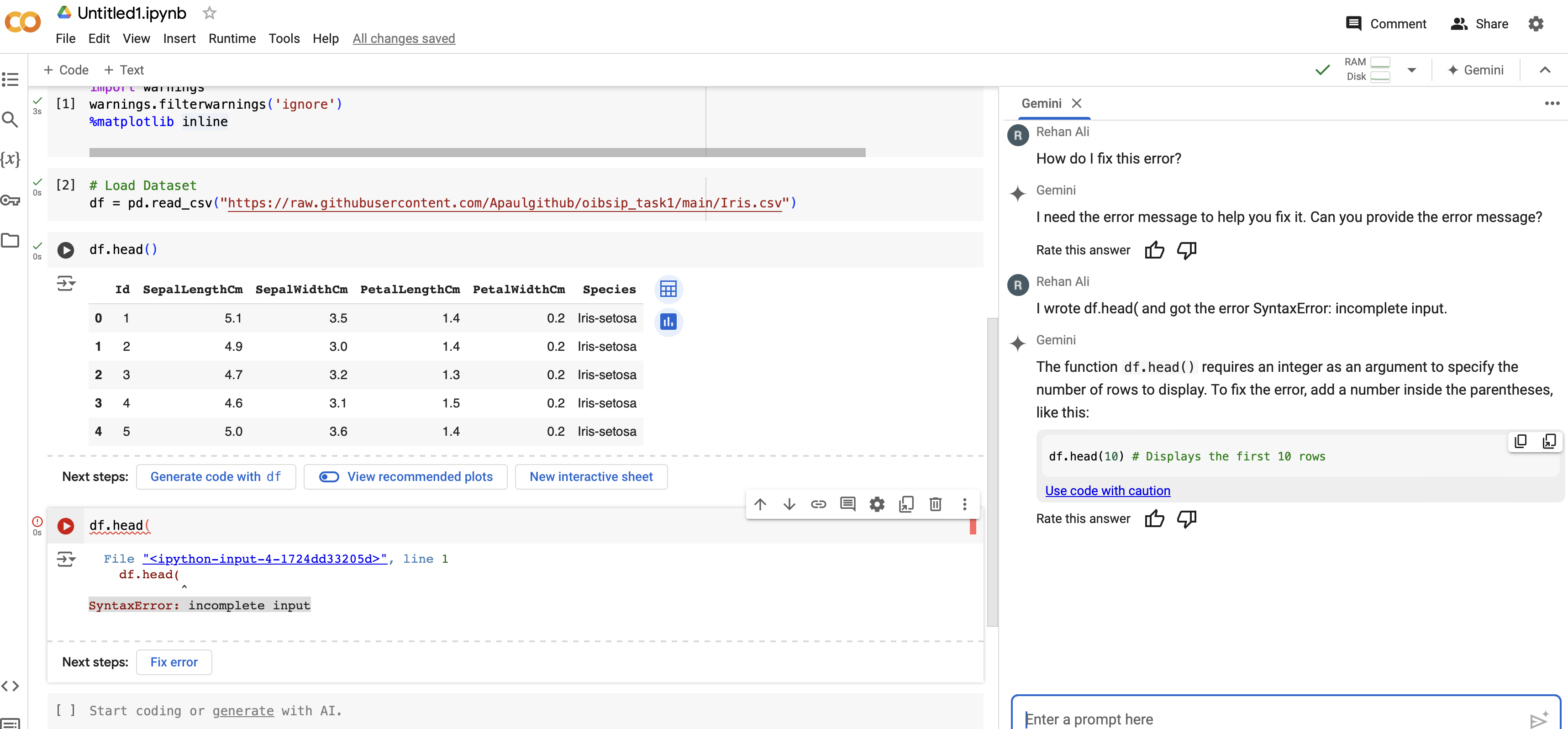
This means that the dataset is now available in a form that you can work with using python, namely a “pandas dataframe”. Once you’ve loaded the data into a dataframe, you can do a lot of different analyses on it, and that’s something you’ll learn as you continue with this project.

What if you get something wrong? CoLab will tell you. Here, we forgot to add the last bracket, and it resulted in an error.



If this happens, and you’re not sure how to fix this, there are lots of ways you can try to address this:

* Google the error message (in this case, “SyntaxError: incomplete input”), although this isn’t always helpful if you’re starting out.
* Ask a mentor on the project.
* Post a question on the project teams channel.
* Ask Google Gemini, the built in chat bot (which you can open by clicking on the Gemini button on the top right). Below, we’ve tried to do exactly this.



## Developing your python and ML skills

Keep working through the Iris notebook to get comfortable with exploring the data through python, and with running some simple machine learning algorithms on it. This notebook will walk you through how someone has already done Exploratory Data Analysis (EDA) on the Iris dataset.

Try and copy and paste the python code into a new CoLab notebook and run it, to see if you can get similar results.

* Feel free to experiment and play with the code, to get a better understanding of how it works. An important part of learning machine learning is through experimentation. Look up key techniques and read up on how they work and see what happens when you change the parameters for them.

A quick word of caution for later on in the project:

* Google CoLab must NEVER be used to process any IQVIA sensitive data (L2 or L3). For this volunteer project, it will only be used to process data that is already available online (L1).

Once you’ve gone through and understood this notebook, we’ll move to a more complex one. This notebook already exists on CoLab, and you can run it directly (or copy and paste the code into a fresh notebook). Here, the author has used more advanced visualisation and clustering settings, and you can play with these here. This second notebook touches upon some more complex topics, so if you’re finding parts of this confusing, please feel free to skip over them or ask someone for help.

<https://colab.research.google.com/github/wgova/kmeans-clustering/blob/master/notebooks/iris_analysis.ipynb#scrollTo=JrQJsTmHXznQ>

If you’re not sure about any of the concepts in these notebooks, please check the resources listed in the Appendix section. You can also ask questions on the Teams channel.

Learning tasks:

* Ask questions about anything you’re not clear about.
* Try to create different plots and statistics.
* Try to describe the trends in the outputs in your own words. What can we learn from these various analyses?

Once you’ve done these, we’d like you to try and do the following tasks:

Project tasks:

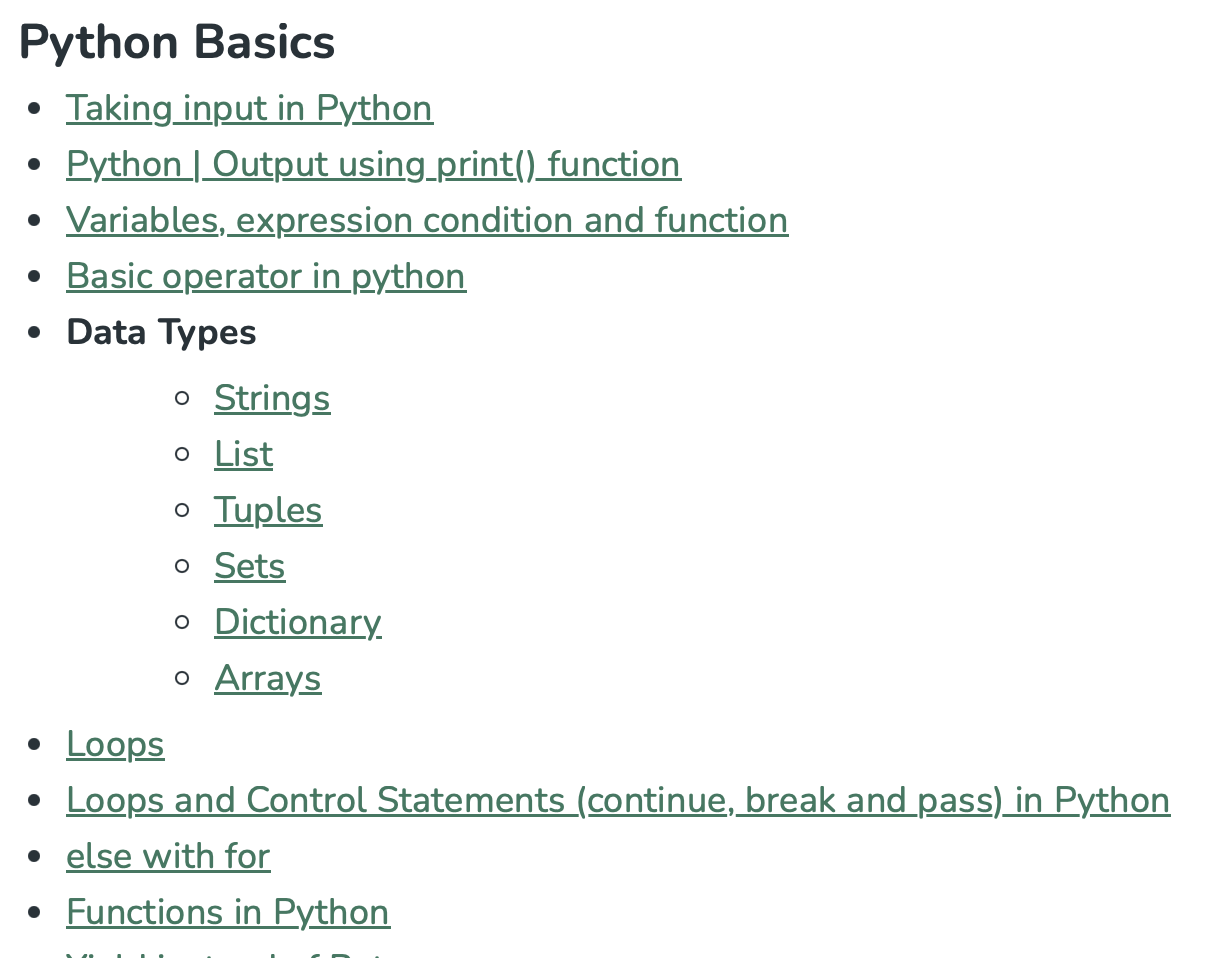
* Can you write a logic rule that can correctly classify a sample? How well does it perform? What is the best rule that you can create?
* Can we automatically generate rules that can do better?
  + Try out this package called [SkopeRules](https://skope-rules.readthedocs.io/en/latest/). Here’s a [blog where it’s applied to Iris](https://medium.com/game-of-bits/deriving-interpretable-rules-from-classification-models-3fda725cc5a3).
* How well do these methods (either creating rules manually, or automatically), do when compared to machine learning methods for classifying the Iris samples?
* Can you come up with a reasonable real life application of these algorithms being applied to the Iris data?

## Next Tasks

Now that you’ve gone through the basics, let’s apply these new skills to new data. On the channel, we’ve created tickets where we would like you to do an EDA and train a classification model on an open source healthcare dataset from the Kaggle website (see the “Tasks” tab on the Pennines channel). For each dataset, we’d like you to write up a summary of your results and post on the project Teams channel. This will help us to improve our understanding of these datasets, and how various ML methods perform on them, which will be helpful for our team.

## Appendix - Learning References

You may want to do some learning to prepare for this project, or have a guide to refer to as you make progress with the project activities. The two main skills needed for this team’s projects are to be comfortable with the fundamentals of Python 3 and machine learning basics. The resources below will be useful to brush up on these skills:

* Python for data science
  + All of our code is written in Python 3, so you should be comfortable with the core concepts including how to use *pandas*, *numpy*and *scikit-learn*.
  + <https://www.geeksforgeeks.org/data-science-with-python-tutorial/>
  + You should focus on the sections shown below and ensure you’re comfortable with these:
  + 
* Machine learning basics
  + It will help to have some understanding of the basic principles of machine learning, specifically classification.
  + <https://microsoft.github.io/ML-For-Beginners/#/> could be a useful resource, and you can focus on the “Classification” topics to begin with. These are shown below