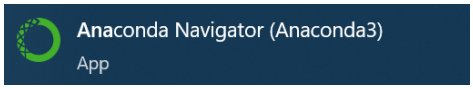
CS989: Big Data Fundamentals
CS982: Big Data Technologies

Before you start

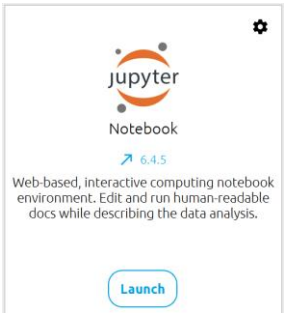
For all labs related to this module, you can use any Python editor/compiler that you are familiar with. However, it's highly recommended to use the Jupyter Notebook provided by Anaconda. The software is already installed on all university computers.

If you would like to install Anaconda on your own Personal Computer, please use the following link: https://www.anaconda.com/download

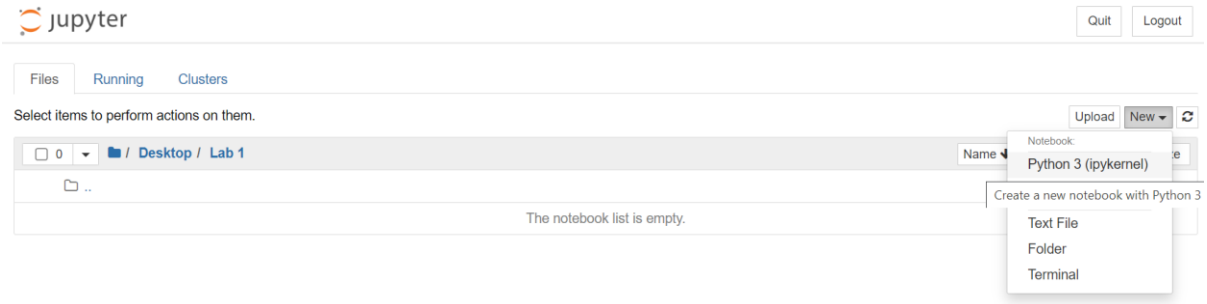
If you are using the university computers, remember you don't need to install Anaconda. Therefore, you need to only start the Anaconda Navigator application.



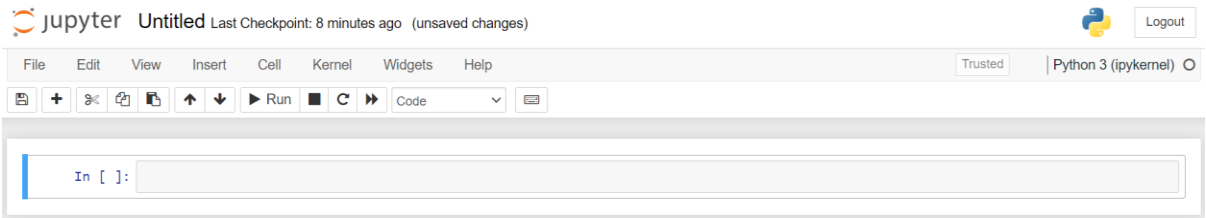
Then launch the Jupyter Notebook within Anaconda:



Once started, click on Desktop, locate a specific folder where you want to save all your lab materials or create a new folder if needed. Then click on New → Python 3 to create a new Python file to write Python code.



Now you are ready to write Python!! So, let's start.



CS989: Big Data Fundamentals
CS982: Big Data Technologies

Lab 3 - Visualisation

Before you start this lab, you need to get familiar with some instructions. We know that programming can be daunting, so we have some extra materials to help you. Make sure to download the notebooks available on MyPlace in week 3. There are 10 different notebooks, each notebook will focus on one specific type of graphs. You can either download all as a folder or download each and every file by itself. But remember to put the Jupyter notebook in the same folder as the dataset(s) so it can work properly.

In addition, filling null (missing) values in a dataset can be challenging and it requires a deep understanding of the dataset itself. Below is a link to a notebook on Kaggle showing an implementation of filling missing values for a specific dataset, Spaceship Titanic, using Python Language, based on some observations:

(1) https://www.kaggle.com/code/josephelgemayel/spaceship-titanic-fillna

The ideas implemented in the notebook are based on a discussion about some rules to fill NaNs by Vincent Debout:

(2) https://www.kaggle.com/competitions/spaceship-titanic/discussion/315987

For a better understanding, it's is recommended to explore first the discussion (2) and then explore the implementation (1) of the ideas.

CS989: Big Data Fundamentals
CS982: Big Data Technologies

Part I – Visualisations with a single variable

Question 1

Import the necessary libraries (numpy, pandas, matplotlib and seaborn).

Question 2

In this lab, we are going to work with the dataset Titanic available on Kaggle. For this reason, you need first to download the dataset from Kaggle by following this link:

https://www.kaggle.com/c/titanic/data (it is train.csv that you want)

Once downloaded, make sure to put the dataset in the same folder as the notebook, then read the dataset using the method .read_csv(…).

Question 3

Produce some box plots for numeric values in the dataset.

Question 4

Plot a bar chart showing the number of survivors and fatalities. Include a title on the chart.

Question 5

Produce a horizontal bar chart showing all passenger classes, ordered with smallest number at top and largest at bottom.

Question 6

Produce a density plot for number of siblings (SibSp).

Question 7

Try to personalise the graphs that you've produced.

Question 8

Try to produce other graphs that you find relevant.

CS989: Big Data Fundamentals
CS982: Big Data Technologies

Part II – Visualisations with more than one variable

For this part, you should create another notebook.

Question 1

Import the necessary libraries (numpy, pandas, matplotlib and seaborn).

Question 2

Import the Titanic dataset that you have downloaded in Part I.

Question 3

Produce a stacked bar chart showing the number of each gender in each passenger class.

Question 4

Produce a heatmap showing the correlation between each numerical variable. What shows a strong correlation?

Question 5

Produce a single scatter plot showing age and passenger class as well as age and number of siblings, different symbols should be used to represent the two different comparisons.

Question 6

Play around with different visualisations of this dataset.