Assignment 1

Due: Sunday Sept 29 11:59pm

Goal: Our goal for this assignment is to familiarize ourselves with creating a Python **notebook**. A tangent goal is to learn how to create a **video,** which you will need to do for your final project. Today, videos are a primary ways for distributing content, so it is an essential skill.

Description. Find an interesting dataset from the Internet. It is better if the dataset is relatively new, e.g., after year 2000, and it should not contain only a few datapoints. You can find many at

- Kaggle
- UCI Machine Learning Repository.
- https://vincentarelbundock.github.io/Rdatasets/datasets.html

I would also urge you to find data about Switzerland! Many data about Switzerland can be found at https://opendata.swiss/en/.

The dataset may be in CSV, JSON, XLS or other format. Prepare a **Python notebook** that should answer in some narrative form the following:

Task 1:

Describe the dataset. Why did you choose it? What is it about? Put a **link** where one can download it. What is the **dimensionality** of the dataset (number of columns)? What is the **cardinality** of the dataset (number of rows/objects)?

To answer the above you have to read the readme file, and look at the data. For this assignment you can use Excel (or even Pandas if you are familiar with it; more about this next week!).

If there is a direct link where one can download the dataset you can include some code as follows, so that one can download the dataset directly from the notebook!

```
!mkdir -p data
# Fetches the dataset and stores it under the folder data
!curl 'the web link here' -o data/my_dataset.csv
```

If the dataset is not too big, you may also upload it together with the notebook, in which case you will upload a zip file of the notebook and the data file.

Task 2:

What kind of **questions** would one be interested to ask/answer about this dataset? For example, predict something about…, etc. Is this a regression, a classification, a clustering, an anomaly detection problem?

Task 3:

Prepare a short video (maximum 1min long) with the following content:
•Your name and student ID.
•What is the dataset about and a link where you found it.
•What questions one can ask about the dataset? It this a regression, classification, etc problem?

The video should include a voice-over and the face of the narrator (at least) in the beginning ("my name is Y, I will present the dataset Z."). To create the video, you may use Camtasia or Quicktime Player and iMovie on Mac. You could also your phone to create this video!

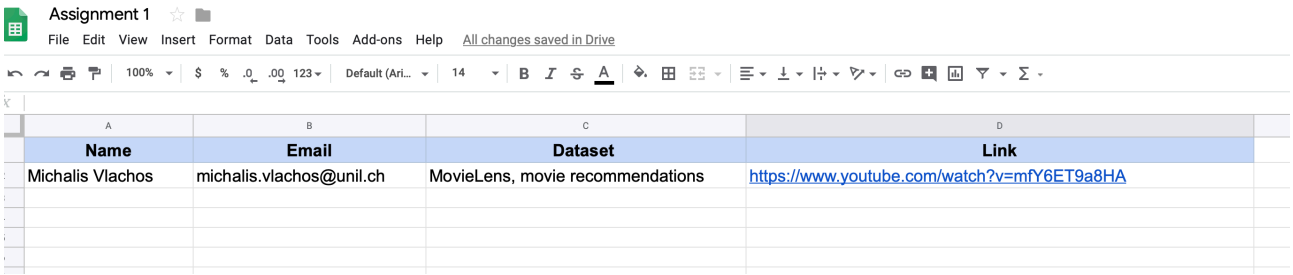
Upload the video to youtube and embed the link in the notebook (as in the Lab assignments).

```
%%HTML
<iframe width="560" height="340" src="https://www.youtube.com/
embed/inN8seMm7UI"></iframe>
```

In case you haven't done it before, there is a separate pdf file in Moodle that describes how to upload a video to YouTube and get its link.

Task 4:

4.1. Make a post in the slack channel of **week 1**, with the name of the dataset and the link to your video.
4.2: Add the same content in this file so that we have everything in one place (see below for a snapshot).



Name	Email	Dataset	Link
Michalis Vlachos	michalis.vlachos@unil.ch	MovieLens, movie recommendations	https://www.youtube.com/watch?v=mfY6ET9a8HA

Grade (Total 10 Points): 5 points for creating a good narrative in your notebook, with proper links, description of the dataset, and good command of the markdown language. **5 points** for the video.

Final note: I will select a couple of these videos to show in class…so show your best!