Version control

+

Notebooks

The tale of a group project

Preamble

You are a mature student enrolled in a MSc. Online. Part-time.

You are an experienced professional during the day. You become a writer of essays during the night.

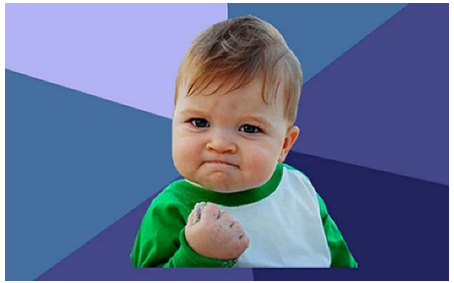
It’s week 5 of the new module and it’s time to start the group project …

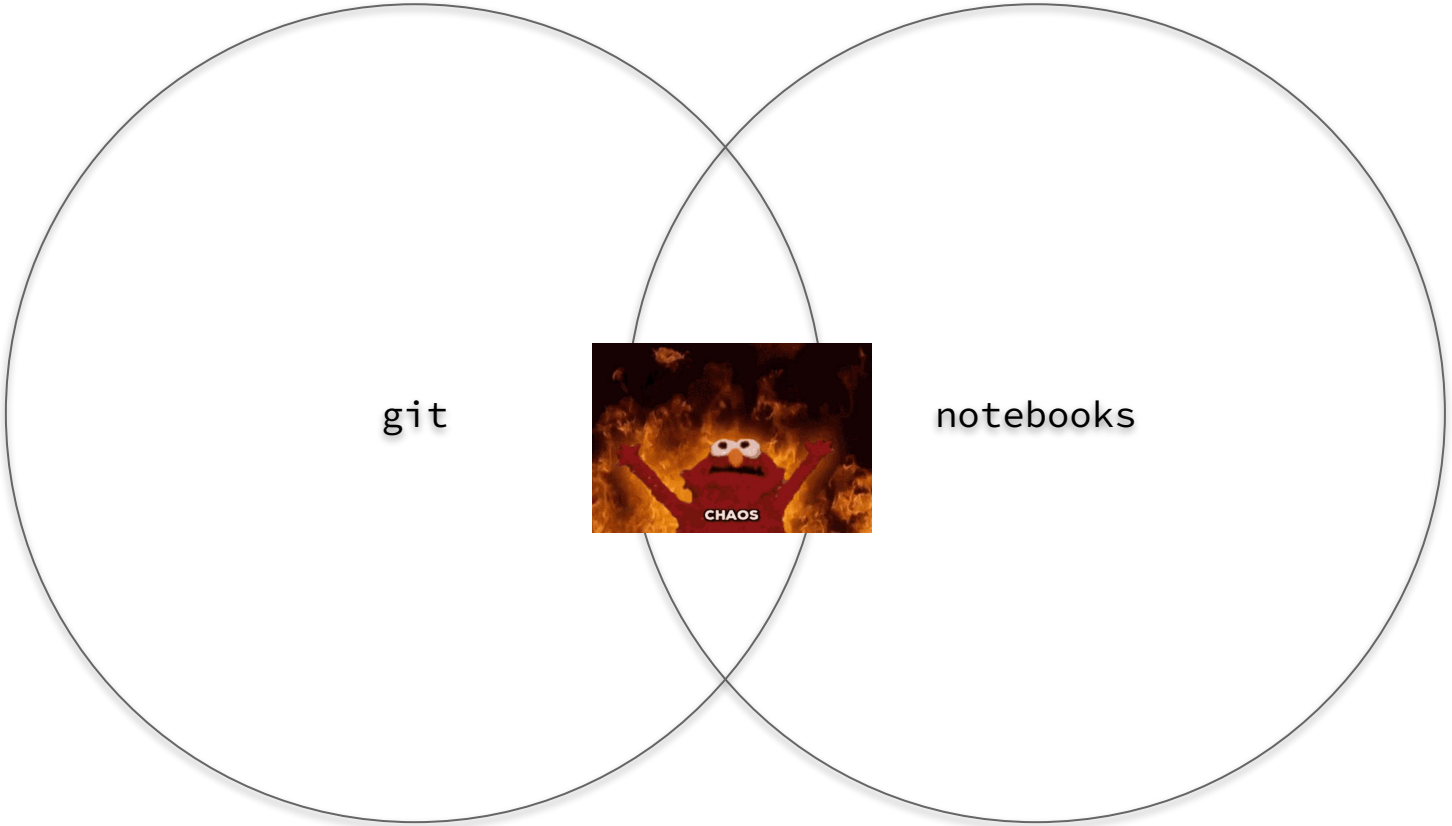

MUSIC BAND
HOW DO YOU DO, FELLOW KIDS?

The group project submission is a Jupyter notebook (*good*).

You have to coordinate with fellow students in 4 different time zones (*not so good*).

The module instructions says “*you can only use emails, the forum or Teams to collaborate, no other system is allowed*” (*... seriously?!*) but the group agrees to use ***git***




git
CHAOS
notebooks

```
CAgICBtYXhCb3VuZHM6IGJvdW5kcywKICAgICAgICAgICAgICAgICAgICAgICAgICAgICAgICAgIC
gICAgIGxheWVyczogW10sCiAgICAgICAgICAgICAgICAgICAgICAgICAgICAgICAgICAgICB3
3JsZENvcHlkdW1wOiBmYWxzZSwKICAgICAgICAgICAgICAgICAgICAgICAgICAgICAgICAgIC
gIGNyczogTC5DUlMuRVBTRzM4NTcKICAgICAgICAgICAgICAgICAgICAgICAgICAgICAgICAg
CAgfSk7CiAgICAgICAgICAgIAogICAgICAgICAgICAgICAgICAgICAgIHZhciB0aW
lX2xheWVyXzJiNzY2ODliYmYxNDQyYjk4MTEwN2NiN2VkZGRhYjBlID0gTC50aWxlTGF5
XIoCiAgICAgICAgICAgICAgICAnaHR0cHM6Ly97c30udGlsZS5vcGVuc3RyZWV0bWFwLm
yZy97en0ve3h9L3t5fS5wbmcnLAogICAgICAgICAgICAgICAgewogICJhdHRyaWJ1dGlv
iI6IG51bGwsCiAgImRldGVjdFJldGluYSI6IGZhbHNlLAogICJtYXhab29tIjogMTgsCi
gIm1pblpvb20iOiAxLAogICJub1dyYXAiOiBmYWxzZSwKICAic3ViZG9tYWlucyI6ICJh
mMiCn0KICAgICAgICAgICAgICAgICkuYWRkVG8obWFwX2U1ZjZiNjg5MGZkNzQzZmRiOT
jZWRiMGI5MmFkODA0KTsKICAgICAgICAgICAKPC9zY3JpcHQ+\" style=\"position:abso
ute;width:100%;height:100%;left:0;top:0;border:none !important;\" all
wfullscreen webkitallowfullscreen mozallowfullscreen></iframe></div><
div>"
```

```
91  +       "<div style=\"width:100%;\"><div style=\"position:relative;wi
    th:100%;height:0;padding-bottom:60%;\"><iframe src=\"data:text/html;c
    arset=utf-8;base64,PCFET0NUWVBFIGh0bWw+CjxoZWFkPiAgICAKICAgIDxtZXRhIG
    0dHAtZXF1aXY9ImNvbnRlbnQtdHlwZSIgY29udGVudD0idGV4dC9odG1sOyBjaGFyc2V0
    VVURi04IiAvPgogICAgPHNjcmlwdD5MX1BSRUZFUl9DQU5WQVMgPSBmYWxzZTsgTF9OT1
    UT1VDSCA9IGZhbHNlOyBMX0RJU0FCTEVfM0QgPSBmYWxzZTs8L3NjcmlwdD4KICAgIDxz
    3JpcHQgc3JjPSJodHRwczovL2Nkbi5qc2RlbGl2ci5uZXQvbnBtL2xlYWZsZXRAMS4yLj
    vZGlzdC9sZWFmbGV0LmpzIj48L3NjcmlwdD4KICAgIDxzY3JpcHQgc3JjPSJodHRwczov
    2FqYXguZ29vZ2xlYXBpcy5jb20vYWpheC9saWJzL2pxdWVyeS8xLjExLjEvanF1ZXJ5Lm
    pbi5qcyI+PC9zY3JpcHQ+CiAgICA8c2NyaXB0IHNyYz0iaHR0cHM6Ly9tYXhjZG4uYm9v
    HN0cmFwY2RuLmNvbS9ib290c3RyYXAvMy4yLjAvanMvYm9vdHN0cmFwLm1pbi5qcyI+PC
    zY3JpcHQ+CiAgICA8c2NyaXB0IHNyYz0iaHR0cHM6Ly9jZG5qcy5jbG91ZGZsYXJlLmNv
    S9hamF4L2xpYnMvTGVhZmxldC5hd2Vzb21lLW1hcmtlcnMvMi4wLjIvbGVhZmxldC5hd2
```

Yes, you could clear all the outputs, remove the metadata, use *nbconvert* and *nbstripout* in *pre-commit* …

Surely there must be a better solution.

… ***jupytext*** *has entered the chat* …

https://github.com/mwouts/jupytext

A quick look on GitHub, the project has 6.5k stars, good documentation and A WORKFLOW THAT IS PERFECT FOR ME …

In 4 easy steps

- Install Jupytext
  ```
  $ pip install jupytext
  ```
- Create *jupytext.toml*
  ```
  $ cat jupytext.toml
  # jupytext.toml at the root of the notebook directory
  formats = "ipynb,py:percent"
  ```
- Restart Jupyter
- Add **.ipynb* to your *.gitignore*

From now on …

- Whenever you save a Jupyter notebook, Jupytext will save a *.py* file with the same name
- If you have a *.py* file created with Jupytext, select it in Jupyter and click on *Open with -> Notebook*

Jupytext will keep the *notebook* and the *.py* file in sync automagically!

From now on …

… all you are going to commit on GitHub is a **Python file**, instead of the notebook.

But wait, there is more …

- As the notebook is now just a .py file you can re-use your pre-commit hooks to run linters, formatters, etc.
- It’s incredibly easy to diff / version control the code changes
- It’s incredibly easy to collaborate on the same notebook!

125 CSCK503-week-8-03-analysis.py

```
@@ -39,11 +39,15 @@
# # !pip install --upgrade pip setuptools
# # !pip install -r requirements.txt

- # %% [markdown] jp-MarkdownHeadingCollapsed=true
# ## Import packages

# %%
import pandas as pd
+ import seaborn as sns
+ import matplotlib.pyplot as plt
+ import sklearn
+ from sklearn.model_selection import train_test_split


# %% [markdown]
@@ -54,7 +58,7 @@

# %%
# Unzip the dataset if neededeed
- # ! unzip -jn data/CSCK503-week-8-data.zip  -d data

# %% [markdown]
# ## Load the dataset in Pandas
@@ -74,5 +78,120 @@
# %%
df.head()

+ # %% [markdown]
+ # # Analysis
+
+ # %% [markdown]
+ # ## Drop NAs, apply encoding
+
+ # %%
+ df.drop(labels=["Lts", "DotRef"], axis=1, inplace=True)
+
+ # %%
+ # df = pd.get_dummies(df, columns=["Location_ExactCut", "BoroughName_ExactCut", "Emissions", "Pollutant"])
+
+ # %%
+ df.shape
```

FIN

LinkedIn: giuseppecunsolo

GitHub: markgreene74

Stay in touch!