# Introduction

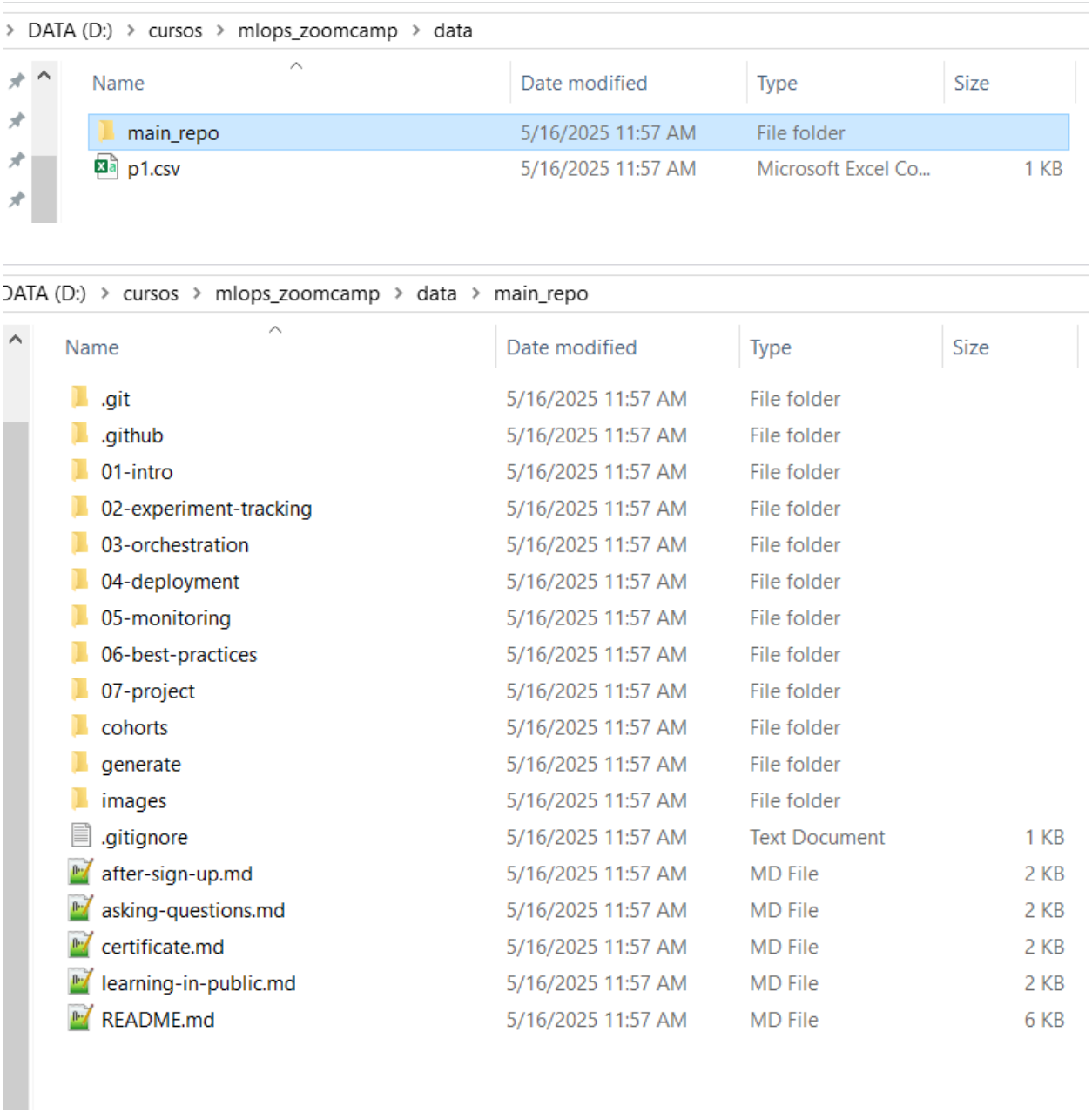
First, my attemp is to combine with docker, and make it all using docker and my personal PC.

Using docker desktop in windows 10, I managed to create a container running ALpine linux with python, running the requirements, and clonning the official github repo in a volume

I call the container tuto1-python and the volume was in my host/path: D:\cursos\mlops_zoomcamp\data

```
D:\cursos\mlops_zoomcamp>docker build -t tuto1-python .
[+] Building 85.5s (11/11) FINISHED
 => [internal] load build definition from dockerfile
 => => transferring dockerfile: 806B
 => [internal] load metadata for docker.io/library/python:3.12-alpine
 => [internal] load .dockerignore
 => => transferring context: 2B
 => [1/6] FROM docker.io/library/python:3.12-alpine@sha256:9c51ecce261773a684c8345b2d4673700055c513b4d54bc0719337d3e4ee552e
 => [internal] load build context
 => => transferring context: 665B
 => CACHED [2/6] RUN apk update && apk add --no-cache git
 => CACHED [3/6] RUN mkdir -p /app/data
 => CACHED [4/6] WORKDIR /app
 => [5/6] COPY . /app
 => [6/6] RUN pip install --no-cache-dir -r requirements.txt
 => exporting to image
 => => exporting layers
 => => writing image sha256:3c4dc42f3fbb15deb0d04848219124d6b8d16da98d32a1ec3917e32ade9e9cb7
 => => naming to docker.io/library/tuto1-python

View build details: docker-desktop://dashboard/build/default/default/nv5bj8fa0g7cvhkzyrr4s4vbu

What's Next?
  View a summary of image vulnerabilities and recommendations → docker scout quickview

D:\cursos\mlops_zoomcamp>docker run --rm -v "D:\cursos\mlops_zoomcamp\data":/app/data tuto1-python
HELLO FROM CONTAINER LUIS
2025-05-16 14:57:28.517 | INFO     | __main__:<module>:13 - path already created
2025-05-16 14:57:28.525 | INFO     | __main__:<module>:17 - data saved
2025-05-16 14:57:28.525 | INFO     | __main__:<module>:19 - clonning git repo
2025-05-16 14:57:40.319 | INFO     | __main__:<module>:22 - repo clonned

D:\cursos\mlops_zoomcamp>
```

We can see that the folder and csv (just fortesting) is created at the host-path




I am also trying loguru instead of regular print in python, and works great.

```python
print('HELLO FROM CONTAINER LUIS')

import pandas as pd
import os
from git import Repo
from  loguru import logger


if not os.path.exists('./data/'):
    logger.info('creatingi dir data')
    os.makedirs('./data')
else:
    logger.info('path already created')

df =  pd.DataFrame()
df.to_csv('./data/p1.csv')
logger.info('data saved')

logger.info('clonning git repo')

Repo.clone_from('https://github.com/DataTalksClub/mlops-zoomcamp.git', '/app/data/main_repo/')
logger.info('repo clonned')
~
```

Then, I explored the idea of downloading the data directly using python and running bash command in it to do so, so i figured out that it can ben possible with os.system(bash_command)

```
D:\cursos\mlops_zoomcamp>docker run --rm -v "D:\cursos\mlops_zoomcamp\data":/app/data tuto1-python
HELLO FROM CONTAINER LUIS
2025-05-16 21:43:52.217 | INFO     | __main__:<module>:13 - path already created
2025-05-16 21:43:52.225 | INFO     | __main__:<module>:17 - data saved
2025-05-16 21:43:52.226 | INFO     | __main__:<module>:19 - clonning git repo
2025-05-16 21:44:04.191 | INFO     | __main__:<module>:22 - repo clonned
2025-05-16 21:44:04.191 | INFO     | __main__:<module>:26 - downloading data
Connecting to d37ci6vzurychx.cloudfront.net (65.8.245.51:443)
saving to '/app/data/yellow_tripdata_2023-01.parquet'
yellow_tripdata_2023   0% |                                |  268k  0:02:52 ETA
yellow_tripdata_2023   8% |**                              | 3922k  0:00:21 ETA
yellow_tripdata_2023  17% |*****                           | 8075k  0:00:14 ETA
yellow_tripdata_2023  24% |*******                         | 11.2M  0:00:12 ETA
yellow_tripdata_2023  31% |*********                       | 14.2M  0:00:10 ETA
yellow_tripdata_2023  44% |*************                   | 20.0M  0:00:07 ETA
yellow_tripdata_2023  53% |***************                 | 24.1M  0:00:06 ETA
yellow_tripdata_2023  60% |******************              | 27.4M  0:00:05 ETA
yellow_tripdata_2023  70% |*********************           | 31.9M  0:00:03 ETA
yellow_tripdata_2023  77% |***********************         | 35.4M  0:00:02 ETA
yellow_tripdata_2023  88% |***************************     | 40.0M  0:00:01 ETA
yellow_tripdata_2023  92% |****************************    | 42.1M  0:00:00 ETA
yellow_tripdata_2023 100% |********************************| 45.4M  0:00:00 ETA
'/app/data/yellow_tripdata_2023-01.parquet' saved
2025-05-16 21:44:17.185 | INFO     | __main__:<module>:29 - downloaded..ok
```

Of course, dealing with an old laptop, that runs win10 and perhaps needs several upgrades, errors starting to pop up, especially while attempt to install scikit-learn and python 3.12,

```
 -Db_ndebug=if-release -Db_vscrt=md --native-file=/tmp/pip-install-wedzuuxa/scikit-learn_02396f26026a49638feea3989d58b56e/.mesonpy-zgulsgp0/meson-python-native-file.ini
94.51       The Meson build system
94.51       Version: 1.8.0
94.51       Source dir: /tmp/pip-install-wedzuuxa/scikit-learn_02396f26026a49638feea3989d58b56e
94.51       Build dir: /tmp/pip-install-wedzuuxa/scikit-learn_02396f26026a49638feea3989d58b56e/.mesonpy-zgulsgp0
94.51       Build type: native build
94.51       Project name: scikit-learn
94.51       Project version: 1.6.1
94.51
94.51       ../meson.build:1:0: ERROR: Unknown compiler(s): [['cc'], ['gcc'], ['clang'], ['nvc'], ['pgcc'], ['icc'], ['icx']]
94.51       The following exception(s) were encountered:
94.51       Running `cc --version` gave "[Errno 2] No such file or directory: 'cc'"
94.51       Running `gcc --version` gave "[Errno 2] No such file or directory: 'gcc'"
94.51       Running `clang --version` gave "[Errno 2] No such file or directory: 'clang'"
94.51       Running `nvc --version` gave "[Errno 2] No such file or directory: 'nvc'"
94.51       Running `pgcc --version` gave "[Errno 2] No such file or directory: 'pgcc'"
94.51       Running `icc --version` gave "[Errno 2] No such file or directory: 'icc'"
94.51       Running `icx --version` gave "[Errno 2] No such file or directory: 'icx'"
94.51
94.51       A full log can be found at /tmp/pip-install-wedzuuxa/scikit-learn_02396f26026a49638feea3989d58b56e/.mesonpy-zgulsgp0/meson-logs/meson-log.txt
94.51       [end of output]
94.51
94.51   note: This error originates from a subprocess, and is likely not a problem with pip.
94.80
94.80 [notice] A new release of pip is available: 25.0.1 -> 25.1.1
94.80 [notice] To update, run: pip install --upgrade pip
94.80 error: metadata-generation-failed
94.80
94.80 × Encountered error while generating package metadata.
94.80 ╰─> See above for output.
94.80
```

Finally, I managed to do so, changing the dockerfile to run python:3.9-slim instead of alpine.

```
 46450K .......... .......... .......... .......... .......... 99% 1.36M 0s
 46500K .......... .......... .......... .......... .......... 99% 5.01M 0s
 46550K .......... .......... .......... .......... .......... 99% 5.85M 0s
 46600K .......... .......... ........                        100% 9.77M=18s

2025-05-17 00:59:43 (2.47 MB/s) - '/app/data/yellow_tripdata_2023-02.parquet.2' saved [47748012/47748012]

2025-05-17 00:59:44.061 | INFO     | __main__:<module>:30 - downloaded..ok
2025-05-17 00:59:44.062 | INFO     | __main__:<module>:33 - INIT HOMEWORK-1
2025-05-17 00:59:44.862 | INFO     | __main__:<module>:42 - reading data
2025-05-17 01:00:16.311 | INFO     | __main__:<module>:45 - Data has 19 columns
2025-05-17 01:00:35.867 | INFO     | __main__:<module>:54 - The Std of the duration data is: 42.59435124195458
2025-05-17 01:00:36.142 | INFO     | __main__:<module>:60 - Percentage after filtering data: 98.1220282212598
/app/main.py:64: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
  df2[categorical] = df2[categorical].astype(str)
2025-05-17 01:01:19.411 | INFO     | __main__:<module>:68 - Number of columns of X_train: 515
2025-05-17 01:01:19.412 | INFO     | __main__:<module>:73 - fitting linear reg model and computing rmse on train set
2025-05-17 01:01:57.679 | INFO     | __main__:<module>:79 - RMSE-Train: 7.649261929201487
2025-05-17 01:01:57.679 | INFO     | __main__:<module>:97 - READING TRAIN AND VAL DATA
2025-05-17 01:04:48.740 | INFO     | __main__:<module>:114 - fitting a linear reg model and computing rmse on val-set
2025-05-17 01:05:53.616 | INFO     | __main__:<module>:122 - RMSE-val: 7.811819793542861
2025-05-17 01:05:53.617 | INFO     | __main__:<module>:123 - END

D:\cursos\mlops_zoomcamp>
```