01-setting-dataset

March 24, 2025

1 Setting Dataset Exercise

Import necessary libraries

[1]:
```
import pandas as pd
```

Download and save into google drive Display the dataset

[2]:
```
# downloaded dataset from https://www.kaggle.com/datasets/abdulmoiz12/
 ↪amazon-stock-data-2025
# AMZN stock 2000 - 2025

amzn = pd.read_csv("../../data/amzn-2000-2025.csv")
amzn.head()
```

[2]:

	date	open	high	low	close	adj_close	volume
0	2000-01-03 00:00:00-05:00	4.075000	4.478125	3.952344	4.468750	4.468750	322352000
1	2000-01-04 00:00:00-05:00	4.268750	4.575000	4.087500	4.096875	4.096875	349748000
2	2000-01-05 00:00:00-05:00	3.525000	3.756250	3.400000	3.487500	3.487500	769148000
3	2000-01-06 00:00:00-05:00	3.565625	3.634375	3.200000	3.278125	3.278125	375040000
4	2000-01-07 00:00:00-05:00	3.350000	3.525000	3.309375	3.478125	3.478125	210108000

Import directly from UCI Machine learning repository

- https://archive.ics.uci.edu/datasets

Install UCI library (ucimlrepo)

[3]:
```
# installing ucimlrepo using pip

!pip install ucimlrepo
```

```
Requirement already satisfied: ucimlrepo in
c:\users\jihit\appdata\roaming\python\python312\site-packages (0.0.7)
Requirement already satisfied: pandas>=1.0.0 in
c:\users\jihit\appdata\roaming\python\python312\site-packages (from ucimlrepo)
(2.2.3)
Requirement already satisfied: certifi>=2020.12.5 in
c:\users\jihit\.conda\envs\test-env\lib\site-packages (from ucimlrepo)
(2025.1.31)
Requirement already satisfied: numpy>=1.26.0 in
c:\users\jihit\appdata\roaming\python\python312\site-packages (from
pandas>=1.0.0->ucimlrepo) (2.2.4)
Requirement already satisfied: python-dateutil>=2.8.2 in
c:\users\jihit\appdata\roaming\python\python312\site-packages (from
pandas>=1.0.0->ucimlrepo) (2.9.0.post0)
Requirement already satisfied: pytz>=2020.1 in
c:\users\jihit\appdata\roaming\python\python312\site-packages (from
pandas>=1.0.0->ucimlrepo) (2025.1)
Requirement already satisfied: tzdata>=2022.7 in
c:\users\jihit\appdata\roaming\python\python312\site-packages (from
pandas>=1.0.0->ucimlrepo) (2025.2)
Requirement already satisfied: six>=1.5 in
c:\users\jihit\appdata\roaming\python\python312\site-packages (from python-
dateutil>=2.8.2->pandas>=1.0.0->ucimlrepo) (1.17.0)
```

Fetch the dataset

[4]:
```
from ucimlrepo import fetch_ucirepo

# fetch dataset
bank_marketing = fetch_ucirepo(id=222)

# data (as pandas dataframes)
X = bank_marketing.data.features
y = bank_marketing.data.targets

bank = pd.concat([X,y], axis=1)
```

Display the dataset

[5]:
```
bank.head()
```

[5]:

	age	job	marital	education	default	balance	housing	loan	\
0	58	management	married	tertiary	no	2143	yes	no	
1	44	technician	single	secondary	no	29	yes	no	
2	33	entrepreneur	married	secondary	no	2	yes	yes	
3	47	blue-collar	married	NaN	no	1506	yes	no	
4	33	NaN	single	NaN	no	1	no	no	

	contact	day_of_week	month	duration	campaign	pdays	previous	poutcome	y

```
0     NaN           5   may       261         1     -1         0      NaN  no
1     NaN           5   may       151         1     -1         0      NaN  no
2     NaN           5   may        76         1     -1         0      NaN  no
3     NaN           5   may        92         1     -1         0      NaN  no
4     NaN           5   may       198         1     -1         0      NaN  no
```

1.0.1 Import directly from Kaggle

- https://www.kaggle.com/datasets

Install opendatasets library (opendatasets)

[6]:
```
!pip install opendatasets
!pip install legacy-cgi
```

```
Requirement already satisfied: opendatasets in
c:\users\jihit\appdata\roaming\python\python312\site-packages (0.1.22)
Collecting tqdm (from opendatasets)
  Using cached tqdm-4.67.1-py3-none-any.whl.metadata (57 kB)
Requirement already satisfied: kaggle in
c:\users\jihit\appdata\roaming\python\python312\site-packages (from
opendatasets) (1.7.4.2)
Requirement already satisfied: click in
c:\users\jihit\appdata\roaming\python\python312\site-packages (from
opendatasets) (8.1.8)
Requirement already satisfied: colorama in
c:\users\jihit\appdata\roaming\python\python312\site-packages (from
click->opendatasets) (0.4.6)
Requirement already satisfied: bleach in
c:\users\jihit\appdata\roaming\python\python312\site-packages (from
kaggle->opendatasets) (6.2.0)
Requirement already satisfied: certifi>=14.05.14 in
c:\users\jihit\.conda\envs\test-env\lib\site-packages (from
kaggle->opendatasets) (2025.1.31)
Requirement already satisfied: charset-normalizer in
c:\users\jihit\.conda\envs\test-env\lib\site-packages (from
kaggle->opendatasets) (3.4.1)
Requirement already satisfied: idna in c:\users\jihit\.conda\envs\test-
env\lib\site-packages (from kaggle->opendatasets) (3.10)
Requirement already satisfied: protobuf in
c:\users\jihit\appdata\roaming\python\python312\site-packages (from
kaggle->opendatasets) (6.30.1)
Requirement already satisfied: python-dateutil>=2.5.3 in
c:\users\jihit\appdata\roaming\python\python312\site-packages (from
kaggle->opendatasets) (2.9.0.post0)
Requirement already satisfied: python-slugify in
c:\users\jihit\appdata\roaming\python\python312\site-packages (from
kaggle->opendatasets) (8.0.4)
Requirement already satisfied: requests in c:\users\jihit\.conda\envs\test-
```

```
env\lib\site-packages (from kaggle->opendatasets) (2.32.3)
Requirement already satisfied: setuptools>=21.0.0 in
c:\users\jihit\.conda\envs\test-env\lib\site-packages (from
kaggle->opendatasets) (75.8.0)
Requirement already satisfied: six>=1.10 in
c:\users\jihit\appdata\roaming\python\python312\site-packages (from
kaggle->opendatasets) (1.17.0)
Requirement already satisfied: text-unidecode in
c:\users\jihit\appdata\roaming\python\python312\site-packages (from
kaggle->opendatasets) (1.3)
Requirement already satisfied: urllib3>=1.15.1 in
c:\users\jihit\.conda\envs\test-env\lib\site-packages (from
kaggle->opendatasets) (2.3.0)
Requirement already satisfied: webencodings in
c:\users\jihit\appdata\roaming\python\python312\site-packages (from
kaggle->opendatasets) (0.5.1)
Using cached tqdm-4.67.1-py3-none-any.whl (78 kB)
Installing collected packages: tqdm
Successfully installed tqdm-4.67.1
Requirement already satisfied: legacy-cgi in
c:\users\jihit\appdata\roaming\python\python312\site-packages (2.6.2)
```

Import necessary libraries

[7]:
```
import pandas as pd
import os
import opendatasets as od
```

Fetch the dataset and enter your username and kaggle API key

[8]:
```
# I'm using AAPL Stock Data "https://www.kaggle.com/datasets/umerhaddii/
  ↪apple-stock-data-2025"
# Putting it into the data folder for better organization

url = "https://www.kaggle.com/datasets/umerhaddii/apple-stock-data-2025?
  ↪select=apple_stock.csv"
target_folder = "../../data"

od.download(url, data_dir=target_folder)
```

```
Skipping, found downloaded files in "../../data\apple-stock-data-2025" (use
force=True to force download)
```

Display the dataset

[9]:
```
aapl = pd.read_csv("../../data/apple_stock.csv")
aapl.head()
```

[9]:

	Unnamed: 0	Adj Close	Close	High	Low	Open	Volume
0	1980-12-12	0.098834	0.128348	0.128906	0.128348	0.128348	469033600

1	1980-12-15	0.093678	0.121652	0.122210	0.121652	0.122210	175884800
2	1980-12-16	0.086802	0.112723	0.113281	0.112723	0.113281	105728000
3	1980-12-17	0.088951	0.115513	0.116071	0.115513	0.115513	86441600
4	1980-12-18	0.091530	0.118862	0.119420	0.118862	0.118862	73449600

Renaming the column

[10]:
```
# Renaming only the 1st column to "date"

aapl.columns.values[0] = "Date"
```

[11]:
```
aapl.head()
```

[11]:

	Date	Adj Close	Close	High	Low	Open	Volume
0	1980-12-12	0.098834	0.128348	0.128906	0.128348	0.128348	469033600
1	1980-12-15	0.093678	0.121652	0.122210	0.121652	0.122210	175884800
2	1980-12-16	0.086802	0.112723	0.113281	0.112723	0.113281	105728000
3	1980-12-17	0.088951	0.115513	0.116071	0.115513	0.115513	86441600
4	1980-12-18	0.091530	0.118862	0.119420	0.118862	0.118862	73449600