Chavez, Stanley Kurt Mickael B.	Prof. Eleazear Repil
IT32S2	March 10, 2025

Exercise 3 | Data Preprocessing in Jupyter Notebook

Objective:

By the end of this activity, students should be able to:

1. Load and explore a dataset using Pandas in Jupyter Notebook.
2. Perform data preprocessing techniques such as handling missing values, duplicate data, and data normalization.
3. Visualize preprocessed data using Matplotlib and Seaborn.

Activity Instructions

Kindly download the csv file: students_data.csv Download

students_data.csv

d and the worksheet here Download

here

.

Colab loading dataset: ColabIntro.ipynb - Colab.pdf Download

ColabIntro.ipynb - Colab.pdf

May consider the following below:

About setting Up the Environment

1. Open Jupyter Notebook.

2. Create a new Python 3 Notebook.
3. Install required libraries (if not installed) by running:
   !pip install pandas matplotlib seaborn
4. Import necessary libraries:
   import pandas as pd import
   numpy as np
   import matplotlib.pyplot as plt
   import seaborn as sns

About loading and Exploring the Dataset

1. Download or create a sample dataset (e.g., students_data.csv with missing values and duplicate records).
2. Load the dataset into a Pandas DataFrame: df =
   pd.read_csv("students_data.csv")
3. Display the first few rows:
   df.head()
4. Get basic information about the dataset:
   df.info()
5. Check for missing values:
   df.isnull().sum()

About Data Visualization

1. Plot a histogram of students' ages:
   plt.hist(df["Age"], bins=5, color='skyblue', edgecolor='black')
   plt.xlabel("Age")
   plt.ylabel("Frequency") plt.title("Distribution
   of Students' Ages") plt.show()
2. Create a boxplot to identify outliers in grades:
   sns.boxplot(x=df["Grade"])
   plt.title("Boxplot of Grades")
   plt.show()
3. Display a correlation heatmap:
   sns.heatmap(df.corr(), annot=True, cmap="coolwarm", linewidths=0.5) plt.title("Correlation Matrix")
   plt.show()

OUTPUT

INSTALL JUPYTER

```
C:\Users\Admin>pip install jupyter
Collecting jupyter
  Downloading jupyter-1.1.1-py2.py3-none-any.whl.metadata (2.0 kB)
Requirement already satisfied: notebook in c:\users\admin\appdata\local\programs\python\python313\lib\site-packages (fro
m jupyter) (7.3.2)
Collecting jupyter-console (from jupyter)
  Downloading jupyter_console-6.6.3-py3-none-any.whl.metadata (5.8 kB)
Requirement already satisfied: nbconvert in c:\users\admin\appdata\local\programs\python\python313\lib\site-packages (fr
om jupyter) (7.16.6)
Requirement already satisfied: ipykernel in c:\users\admin\appdata\local\programs\python\python313\lib\site-packages (fr
om jupyter) (6.29.5)
Collecting ipywidgets (from jupyter)
  Downloading ipywidgets-8.1.5-py3-none-any.whl.metadata (2.3 kB)
Requirement already satisfied: jupyterlab in c:\users\admin\appdata\local\programs\python\python313\lib\site-packages (f
rom jupyter) (4.3.5)
Requirement already satisfied: comm>=0.1.1 in c:\users\admin\appdata\local\programs\python\python313\lib\site-packages (
from ipykernel->jupyter) (0.2.2)
Requirement already satisfied: debugpy>=1.6.5 in c:\users\admin\appdata\local\programs\python\python313\lib\site-package
s (from ipykernel->jupyter) (1.8.13)
Requirement already satisfied: ipython>=7.23.1 in c:\users\admin\appdata\local\programs\python\python313\lib\site-packag
es (from ipykernel->jupyter) (9.0.2)
Requirement already satisfied: jupyter-client>=6.1.12 in c:\users\admin\appdata\local\programs\python\python313\lib\site
-packages (from ipykernel->jupyter) (8.6.3)
Requirement already satisfied: jupyter-core!=5.0.*,>=4.12 in c:\users\admin\appdata\local\programs\python\python313\lib\
site-packages (from ipykernel->jupyter) (5.7.2)
Requirement already satisfied: matplotlib-inline>=0.1 in c:\users\admin\appdata\local\programs\python\python313\lib\site
-packages (from ipykernel->jupyter) (0.1.7)
Requirement already satisfied: nest-asyncio in c:\users\admin\appdata\local\programs\python\python313\lib\site-packages
(from ipykernel->jupyter) (1.6.0)
```

OPENING JUPYTER NOTEBOOK

```
C:\Users\Admin>jupyter notebook
Fail to get yarn configuration. [WinError 193] %1 is not a valid Win32 application
[I 2025-03-10 01:06:23.958 ServerApp] jupyter_lsp | extension was successfully linked.
[I 2025-03-10 01:06:23.963 ServerApp] jupyter_server_terminals | extension was successfully linked.
[I 2025-03-10 01:06:23.970 ServerApp] jupyterlab | extension was successfully linked.
[I 2025-03-10 01:06:23.976 ServerApp] notebook | extension was successfully linked.
[I 2025-03-10 01:06:24.484 ServerApp] notebook_shim | extension was successfully linked.
[I 2025-03-10 01:06:24.534 ServerApp] notebook_shim | extension was successfully loaded.
[I 2025-03-10 01:06:24.537 ServerApp] jupyter_lsp | extension was successfully loaded.
[I 2025-03-10 01:06:24.537 ServerApp] jupyter_server_terminals | extension was successfully loaded.
[I 2025-03-10 01:06:24.541 LabApp] JupyterLab extension loaded from C:\Users\Admin\AppData\Local\Programs\Python\Python3
13\Lib\site-packages\jupyterlab
[I 2025-03-10 01:06:24.541 LabApp] JupyterLab application directory is C:\Users\Admin\AppData\Local\Programs\Python\Pyth
on313\share\jupyter\lab
[I 2025-03-10 01:06:24.542 LabApp] Extension Manager is 'pypi'.
[I 2025-03-10 01:06:24.933 ServerApp] jupyterlab | extension was successfully loaded.
[I 2025-03-10 01:06:24.940 ServerApp] notebook | extension was successfully loaded.
[I 2025-03-10 01:06:24.941 ServerApp] Serving notebooks from local directory: C:\Users\Admin
[I 2025-03-10 01:06:24.941 ServerApp] Jupyter Server 2.15.0 is running at:
[I 2025-03-10 01:06:24.941 ServerApp] http://localhost:8888/tree?token=d7af9cc0c3614b06ecf0c3a23809ee93b2be94768eb6edbf
[I 2025-03-10 01:06:24.942 ServerApp]     http://127.0.0.1:8888/tree?token=d7af9cc0c3614b06ecf0c3a23809ee93b2be94768eb6e
dbf
[I 2025-03-10 01:06:24.942 ServerApp] Use Control-C to stop this server and shut down all kernels (twice to skip confirm
ation).
[C 2025-03-10 01:06:24.980 ServerApp]
```

PIP INSTALL

```
!pip install pandas matplotlib seaborn
```

```
Requirement already satisfied: pandas in c:\users\admin\appdata\local\programs\python\python313\lib\site-packages (2.2.3)
Requirement already satisfied: matplotlib in c:\users\admin\appdata\local\programs\python\python313\lib\site-packages (3.10.1)
Requirement already satisfied: seaborn in c:\users\admin\appdata\local\programs\python\python313\lib\site-packages (0.13.2)
Requirement already satisfied: numpy>=1.26.0 in c:\users\admin\appdata\local\programs\python\python313\lib\site-packages (from pandas) (2.2.3)
Requirement already satisfied: python-dateutil>=2.8.2 in c:\users\admin\appdata\local\programs\python\python313\lib\site-packages (from pandas) (2.9.0.po
st0)
Requirement already satisfied: pytz>=2020.1 in c:\users\admin\appdata\local\programs\python\python313\lib\site-packages (from pandas) (2025.1)
Requirement already satisfied: tzdata>=2022.7 in c:\users\admin\appdata\local\programs\python\python313\lib\site-packages (from pandas) (2025.1)
Requirement already satisfied: contourpy>=1.0.1 in c:\users\admin\appdata\local\programs\python\python313\lib\site-packages (from matplotlib) (1.3.1)
Requirement already satisfied: cycler>=0.10 in c:\users\admin\appdata\local\programs\python\python313\lib\site-packages (from matplotlib) (0.12.1)
Requirement already satisfied: fonttools>=4.22.0 in c:\users\admin\appdata\local\programs\python\python313\lib\site-packages (from matplotlib) (4.56.0)
Requirement already satisfied: kiwisolver>=1.3.1 in c:\users\admin\appdata\local\programs\python\python313\lib\site-packages (from matplotlib) (1.4.8)
Requirement already satisfied: packaging>=20.0 in c:\users\admin\appdata\local\programs\python\python313\lib\site-packages (from matplotlib) (24.2)
Requirement already satisfied: pillow>=8 in c:\users\admin\appdata\local\programs\python\python313\lib\site-packages (from matplotlib) (11.1.0)
Requirement already satisfied: pyparsing>=2.3.1 in c:\users\admin\appdata\local\programs\python\python313\lib\site-packages (from matplotlib) (3.2.1)
Requirement already satisfied: six>=1.5 in c:\users\admin\appdata\local\programs\python\python313\lib\site-packages (from python-dateutil>=2.8.2->pandas)
(1.17.0)

[notice] A new release of pip is available: 24.3.1 -> 25.0.1
[notice] To update, run: python.exe -m pip install --upgrade pip
```

IMPORT

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
```

LOAD THE DATASET AND DISPLAY FIRST FEW ROWS

```
df = pd.read_csv("students_data.csv")
```

```
df.head()
```

	Student_ID	Name	Age	Gender	Grade	Attendance
0	101	Alice	20.0	F	85.0	95.0
1	102	Bob	21.0	M	78.0	88.0
2	103	Charlie	22.0	M	92.0	92.0
3	104	David	20.0	M	65.0	80.0
4	105	Eve	23.0	F	88.0	97.0

GET BASIC INFORMATION ABOUT DATASET

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 20 entries, 0 to 19
Data columns (total 6 columns):
 #   Column      Non-Null Count  Dtype  
---  ------      --------------  -----  
 0   Student_ID  20 non-null     int64  
 1   Name        20 non-null     object 
 2   Age         19 non-null     float64
 3   Gender      19 non-null     object 
 4   Grade       18 non-null     float64
 5   Attendance  19 non-null     float64
dtypes: float64(3), int64(1), object(2)
memory usage: 1.1+ KB
```

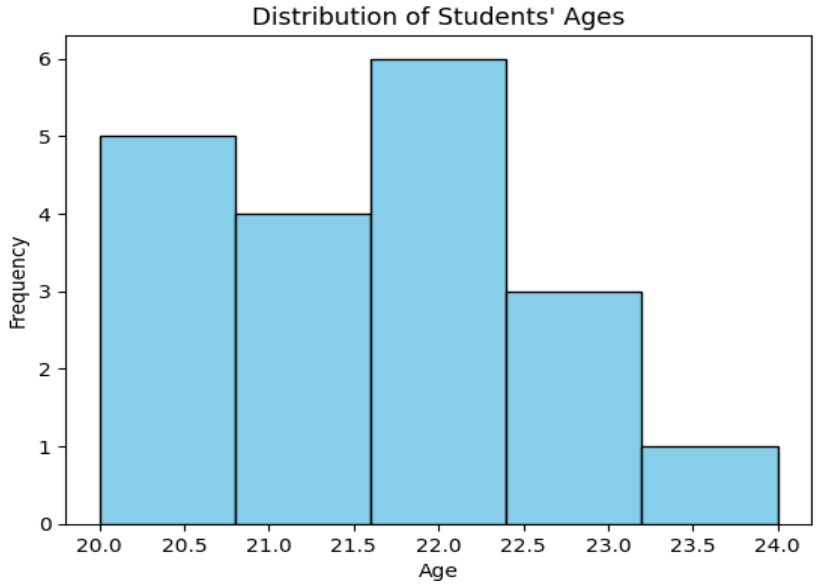
CHECK FOR MISSING VALUE

```
[6]: df.isnull().sum()
```

```
[6]: Student_ID    0
     Name          0
     Age           1
     Gender        1
     Grade         2
     Attendance    1
     dtype: int64
```

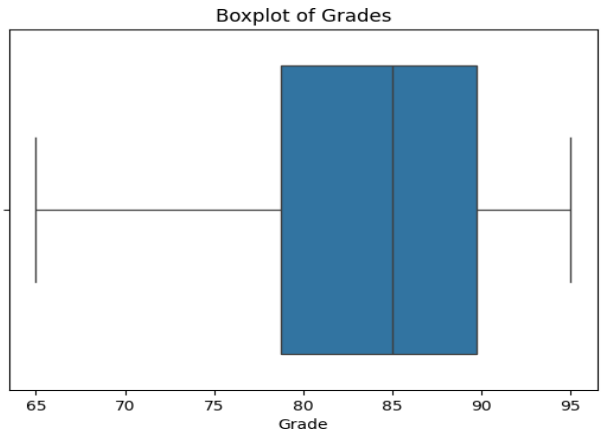
DATA VISUALIZATION

```
[10]: plt.hist(df["Age"], bins=5, color='skyblue', edgecolor='black')
      plt.xlabel("Age")
      plt.ylabel("Frequency")
      plt.title("Distribution of Students' Ages")
      plt.show()
```



CREATE A BOXPLOT TO IDENTIFY OUTLIERS IN GRADES

```
[11]: sns.boxplot(x=df["Grade"])
      plt.title("Boxplot of Grades")
      plt.show()
```



DISPLAY A CORRELATION HEATMAP

[18]:
```
# Select only numeric columns
numeric_df = df.select_dtypes(include=['number'])

# Now generate the heatmap
sns.heatmap(numeric_df.corr(), annot=True, cmap="coolwarm", linewidths=0.5)
plt.title("Correlation Matrix")
plt.show()
```