**Department of Electrical Engineering and   
Computer Science**

**Faculty Member:** Dr. Ahmad Salman **Dated:** 06/10/2023

**Semester:** 7th **Section:** BEE 12C

**CS-471 Machine Learning**

Lab 4: Introduction to Pandas and Matplotlib

**Group Members**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  |  | **PLO4 - CLO4** | | **PLO5 -CLO5** | **PLO8 -CLO6** | **PLO9 -CLO7** |
| **Name** | **Reg. No** | **Viva / Quiz / Lab Performance** | **Analysis of Data in Lab Report** | **Modern Tool Usage** | **Ethics and Safety** | **Individual and Teamwork** |
|  |  | **5 Marks** | **5 Marks** | **5 Marks** | **5 Marks** | **5 Marks** |
| Afif Arif Siddiqi | 344504 |  |  |  |  |  |
| Muhammad Ali Farooq | 331879 |  |  |  |  |  |
| Muhammad Ahmed Mohsin | 333060 |  |  |  |  |  |
| Danial Ahmad | 331388 |  |  |  |  |  |
| Ehtishaam Tanveer | 333074 |  |  |  |  |  |
| Muhammad Umer | 345834 |  |  |  |  |  |

**Table of Contents**

[2 Introduction to Pandas and Matplotlib 3](#_Toc147745848)

[2.1 Introduction 3](#_Toc147745849)

[2.2 Objectives 3](#_Toc147745850)

[2.3 Theory 3](#_Toc147745851)

[3 Lab Tasks 4](#_Toc147745852)

[3.1 Task 1 – Pandas Series and Dataframes 4](#_Toc147745853)

[3.2 Task 2 – CSV Files 4](#_Toc147745854)

[3.3 Task 3 – Dataset Cleaning 5](#_Toc147745855)

[3.4 Task 4 – Line and Scatter Plots 6](#_Toc147745856)

[3.5 Task 5 – Dataset Batches 9](#_Toc147745857)

[3.6 Task 6 – Your Own Dataset 11](#_Toc147745858)

[4 Conclusion 13](#_Toc147745859)

# Introduction to Pandas and Matplotlib

## Introduction

This laboratory exercise is focused on handling and visualizing datasets for machine learning purposes. In any machine learning task, we are working with data. For dataset handling, we use the Pandas library which can load .csv files into a data frame. During machine learning, we also need to make plots. For this, we make use of the PyPlot submodule in the Matplotlib library.

## Objectives

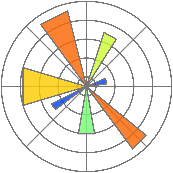
The following are the main objectives of this lab:

* Learn how to load and handle datasets for machine learning using the Pandas library.
* Identify and remove outliers and inconsistencies in datasets using data visualization.
* Create and interpret plots of machine learning data using the PyPlot submodule in the Matplotlib library.
* Use data visualization to improve the performance of machine learning models.

## Theory

Pandas (panel data) is a library that can load tabular data from .csv files and store them into a NumPy compatible table known as a “Pandas Data Frame”. Each column in a data frame is of the “Pandas Series” type. Aside from loading datasets, pandas also enable us to perform basic mean, mode, median operations as well as clean up incomplete or duplicate data.

Matplotlib is another library focused on data visualization. It contains many functions for displaying plots, subplots, scatter plots etc. Line plots are used widely for monitoring training accuracies and losses. Scatter plots are used mainly for modeling the feature space of the dataset.

# Lab Tasks

## Task 1 – Pandas Series and Dataframes

1. Create a Pandas series using a dictionary and display the output.
2. Create a Pandas dataframe using a dictionary and display the output.

Provide all the codes and screenshots of the final outputs.

### TASK 1 CODE STARTS HERE ###

series = pd.Series({"a": 1, "b": 2, "c": 3})

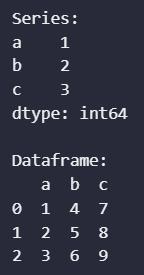
print("Series:\n", series)

dataframe = pd.DataFrame({"a": [1, 2, 3], "b": [4, 5, 6], "c": [7, 8, 9]})

print("\nDataframe:\n", dataframe)

### TASK 1 CODE ENDS HERE ###

### TASK 1 OUTPUT SCREENSHOT STARTS HERE ###



### TASK 1 OUTPUT SCREENSHOT ENDS HERE ###

## Task 2 – CSV Files

Load dataset 1 into a dataframe and perform the following

1. Print the dataset using the head and tail functions
2. Print any 3 rows from the dataset
3. Print any 5 elements from the dataset
4. Use the mean, mode and median functions for each column in the dataset

Provide all the codes and screenshots of the final output.

### TASK 2 CODE STARTS HERE ###

df = pd.read\_csv(path\_data1)

print("Head:\n", df.head())

print("\nTail:\n", df.tail())

print("\n3 rows:\n", df.iloc[0:3])

print("\n5 elements:\n", df.iloc[1:6, 1:6])

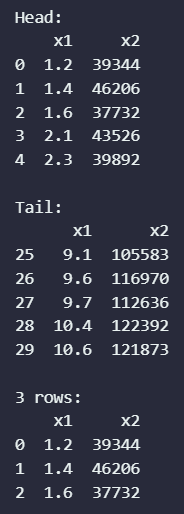
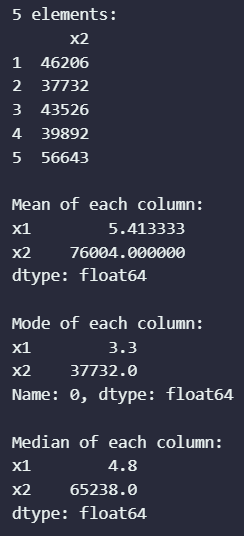
print("\nMean of each column:\n", df.mean())

print("\nMode of each column:\n", df.mode().iloc[0])

print("\nMedian of each column:\n", df.median())

### TASK 2 CODE ENDS HERE ###

### TASK 2 OUTPUT SCREENSHOTS START HERE ###

### TASK 2 OUTPUT SCREENSHOTS END HERE ###

## Task 3 – Dataset Cleaning

Load dataset 2 into a dataframe.

1. Write code to remove the incomplete rows from the dataset
2. Write code to remove the duplicated rows from the dataset
3. Save the cleaned dataset into a dataframe. You need to attach this cleaned dataset file (renamed to lab5\_task3.csv) in your lab submission.

### TASK 3 CODE STARTS HERE ###

df = pd.read\_csv(path\_data2)

df = df.dropna()

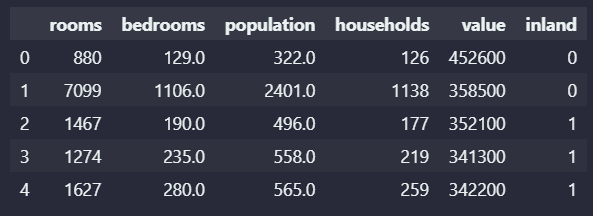
df = df.drop\_duplicates()

df.to\_csv("lab4\_task3.csv", *index*=False)

df.head()

### TASK 3 CODE ENDS HERE ###

### TASK 3 OUTPUT SCREENSHOT STARTS HERE ###



### TASK 3 OUTPUT SCREENSHOT ENDS HERE ###

## Task 4 – Line and Scatter Plots

For this task, you will need to use datasets 1 and 2. You will also require the matplotlib.pyplot module for plotting. Perform the following.

1. Make line plots of the following equations for x = 1 to 100. You will need to make use of NumPy arrays for this part.

y = 2x + 1

y = 3x2

y = cos(x)

1. Load dataset 1 and make a scatter plot (axes x1 and x2)
2. Load the cleaned dataset 2 and make a scatter plot (axes x1 and x2). You need to use markers for the labels (y) such that 0 corresponds to a red circle and 1 corresponds to a blue square. The label y is the “inland” column. For x1 and x2, choose any 2 columns from the dataset and also mention the columns that you are using.
3. Load the cleaned dataset 2 and make a 3-D scatter plot between any three features in the dataset (axes x1, x2, x3)

### TASK 4 CODE STARTS HERE ###

*# a)*

x = np.linspace(1, 100, 1000)

y1 = 2 \* x + 1

y2 = 3 \* x \*\* 2

y3 = np.cos(x)

plt.plot(x, y1)

plt.title("y = 2x + 1")

plt.xlabel("x")

plt.ylabel("y")

plt.show()

plt.plot(x, y2)

plt.title("y = 3x^2")

plt.xlabel("x")

plt.ylabel("y")

plt.show()

plt.plot(x, y3)

plt.title("y = cos(x)")

plt.xlabel("x")

plt.ylabel("y")

plt.show()

*# b)*

df = pd.read\_csv(path\_data1)

plt.scatter(df["x1"], df["x2"])

plt.title("Scatter Plot ~ Dataset 1")

plt.xlabel("x1")

plt.ylabel("x2")

plt.show()

*# c*

df = pd.read\_csv("lab4\_task3.csv")

x1 = df["value"] # selected column 1

x2 = df["bedrooms"] # selected column 2

y = df["inland"]

fig, ax = plt.subplots()

markers = {0: "o", 1: "s"}

colors = {0: "red", 1: "blue"}

for label in y.unique():

    mask = y == label

    ax.scatter(

        x1[mask], x2[mask], *marker*=markers[label], *color*=colors[label], *label*=label

    )

ax.legend(*loc*="upper left")

plt.title("Scatter Plot ~ Dataset 2")

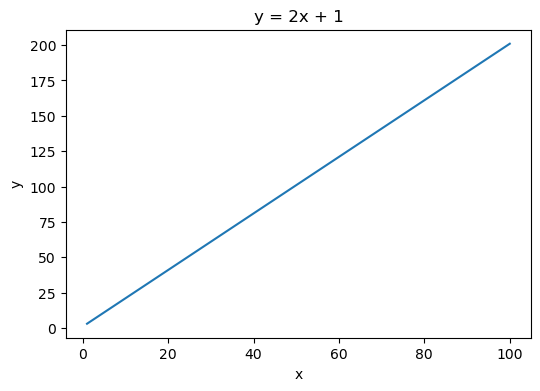
plt.xlabel("value")

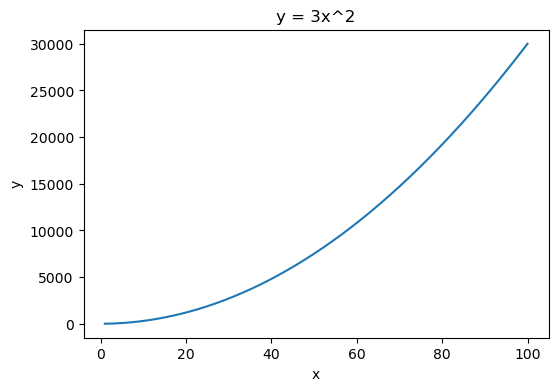
plt.ylabel("bedrooms")

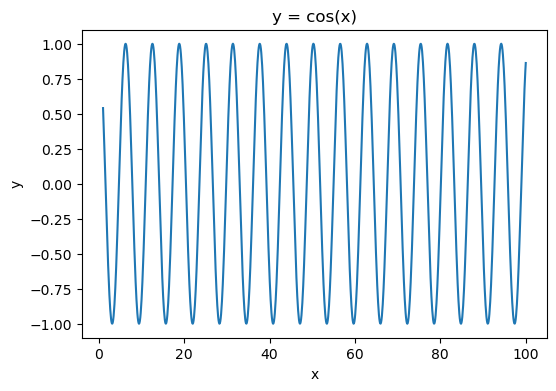
plt.show()

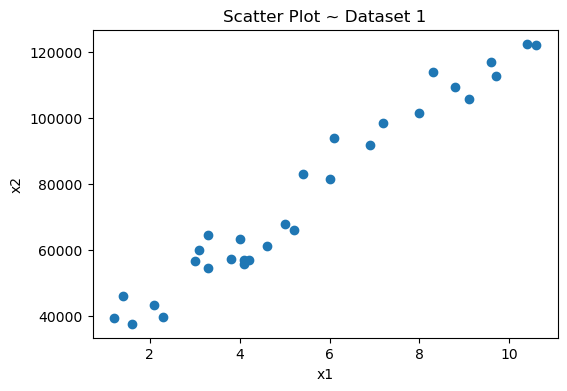
### TASK 4 CODE ENDS HERE ###

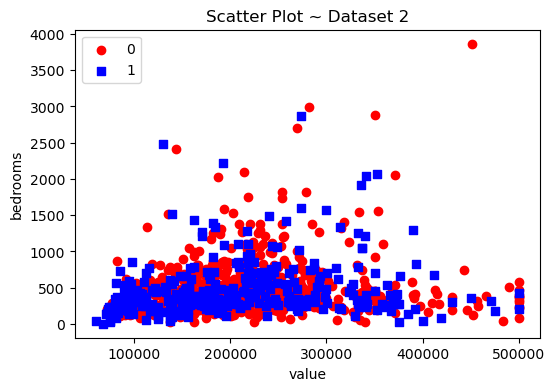
### TASK 4 OUTPUT SCREENSHOTS START HERE ###











### TASK 4 OUTPUT SCREENSHOTS END HERE ###

## Task 5 – Dataset Batches

Load the cleaned dataset 2 into as a dataframe. For this task, you will divide the dataset into 10 batches. For each individual batch, calculate the mean, mode and median for any 3 feature columns. Finally, make line plots showing the batch number on the x-axis and the mean, mode and median on the y-axis.

### TASK 5 CODE STARTS HERE ###

*def* split\_dataframe(*df*, *chunk\_size*):

    chunks = *list*()

    num\_chunks = len(df) // chunk\_size + 1

    for i in range(num\_chunks):

        chunks.append(df[i \* chunk\_size : (i + 1) \* chunk\_size])

    return chunks

df = pd.read\_csv(path\_data2)

df = df.dropna()

df = df.drop\_duplicates()

batch\_size = 10

batches = split\_dataframe(df, len(df) // batch\_size + 1)

*# Features: "rooms", "households", "population"*

features = ["rooms", "households", "population"]

for feature in features:

    mean = []

    mode = []

    median = []

    for batch in batches:

        mean.append(batch[feature].mean())

        mode.append(batch[feature].mode().iloc[0])

        median.append(batch[feature].median())

    plt.plot(range(1, len(batches) + 1), mean, *label*="Mean")

    plt.plot(range(1, len(batches) + 1), mode, *label*="Mode")

    plt.plot(range(1, len(batches) + 1), median, *label*="Median")

    plt.title(feature)

    plt.xlabel("Batch")

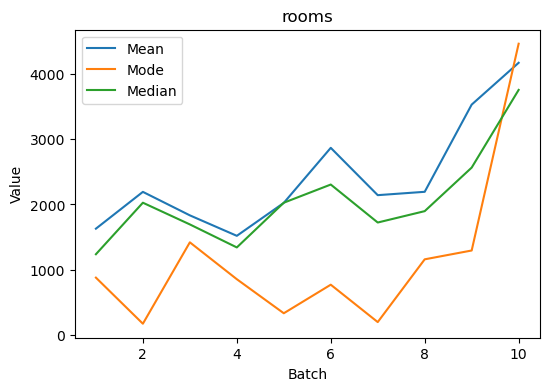
    plt.ylabel("Value")

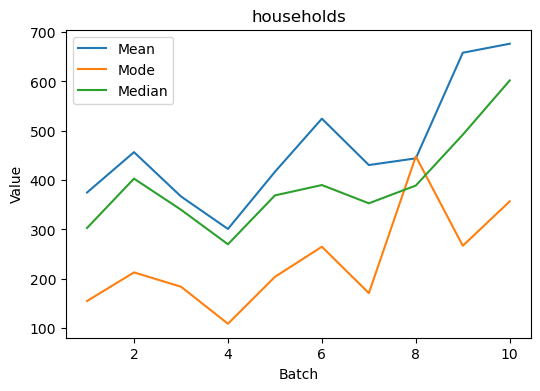
    plt.legend()

    plt.show()

### TASK 5 CODE ENDS HERE ###

### TASK 5 OUTPUT SCREENSHOT STARTS HERE ###







### TASK 5 OUTPUT SCREENSHOT ENDS HERE ###

## Task 6 – Your Own Dataset

Download your own CSV dataset from the internet (e.g., Kaggle). Your dataset must have at least 500 rows and at least 2 feature columns. Your dataset must also have a labels column with classification data (0/1). Make a scatter plot between the feature axes and show the labels with different markers. Provide all the codes and screenshots of the plots. You will also need to submit the downloaded dataset with your report (renamed as lab4\_task6.csv). Note that no two submitted datasets must be the same.

### TASK 6 CODE STARTS HERE ###

*# https://www.kaggle.com/datasets/azeembootwala/titanic*

df = pd.read\_csv("titanic.csv")

df.head()

df.to\_csv("lab4\_task6.csv", *index*=False)

x1 = df["Age"]

x2 = df["Fare"]

y = df["Survived"]

fig, ax = plt.subplots()

markers = {0: "o", 1: "s"}

colors = {0: "red", 1: "blue"}

labels = {0: "Died", 1: "Survived"}

for label in y.unique():

    mask = y == label

    ax.scatter(

        x1[mask],

        x2[mask],

*marker*=markers[label],

*color*=colors[label],

*label*=labels[label],

    )

ax.legend(*loc*="upper left")

plt.title("Scatter Plot ~ Dataset 2")

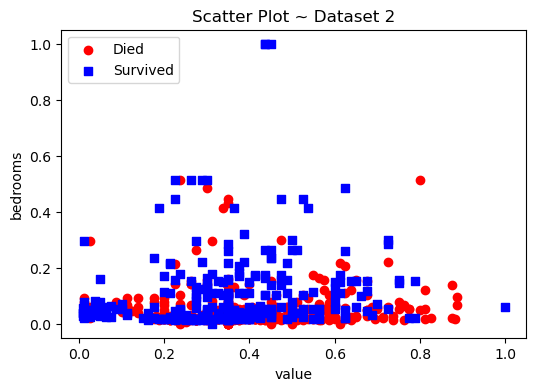
plt.xlabel("value")

plt.ylabel("bedrooms")

plt.show()

### TASK 6 CODE ENDS HERE ###

### TASK 6 SCREENSHOTS START HERE ###



### TASK 6 SCREENSHOTS END HERE ###

# Conclusion

In this laboratory exercise, we learned how to handle and visualize datasets for machine learning purposes using the Pandas and Matplotlib libraries. We were able to load and handle datasets into data frames, identify and remove outliers and inconsistencies, and create and interpret plots of machine learning data. We also learned how to use data visualization to improve the performance of machine learning models.