ASSINGMENT 4

Name: Manasa Vathumilli

ID: 700745467

Question 1 :

a. Read the provided CSV file ‘data.csv’.

b. https://drive.google.com/drive/folders/1h8C3mLsso-R- sIOLsvoYwPLzy2fJ4IOF?usp=sharing

c. Show the basic statistical description about the data. d. Check if the data has null values. i. Replace the null values with the mean

e. Select at least two columns and aggregate the data using: min, max, count, mean. f. Filter the dataframe to select the rows with calories values between 500 and 1000.

g. Filter the dataframe to select the rows with calories values > 500 and pulse < 100. h. Create a new “df\_modified” dataframe that contains all the columns from df except for “Maxpulse”.

i. Delete the “Maxpulse” column from the main df dataframe j. Convert the datatype of Calories column to int datatype.

k. Using pandas create a scatter plot for the two columns (Duration and Calories).

import numpy as np

import pandas as pd

# 1(a) Import the given “Data.csv”

dst\_Data = pd.read\_csv('/Users/manasavathumilli/Downloads/data.csv')

dst\_Data.info()

A screenshot of a computer

Description automatically generated

#(c) Show the basic statistical description about the data.

dst\_Data.head()

A screen shot of a computer

Description automatically generated

#(d)Check if the data has null values.

dst\_Data.isnull().any()

A white rectangular object with a black border

Description automatically generated with medium confidence

dst\_Data.fillna(dst\_Data.mean(), inplace=True)

dst\_Data.isnull().any()

A screenshot of a computer

Description automatically generated

#d(i)Replace the null values with the mean

column\_means = dst\_Data.mean()

print(column\_means)

dst\_Data = dst\_Data. fillna(column\_means)

print(dst\_Data.head(20))

A screenshot of a computer

Description automatically generated

#(e)Select at least two columns and aggregate the data using: min, max, count, mean.

res = dst\_Data.agg({'Calories': ['mean', 'min','max', 'count'],'Pulse': ['mean', 'min', 'max', 'count']})

print(res)

A screen shot of a computer

Description automatically generated

#(f)Filter the dataframe to select the rows with calories values between 500 and 1000

filter\_dst\_Data1=dst\_Data[(dst\_Data['Calories'] > 500) & (dst\_Data['Calories'] < 1000)]

print(filter\_dst\_Data1)

#(g)Filter the dataframe to select the rows with calories values > 500 and pulse < 100.

filter\_dst\_Data2=dst\_Data[(dst\_Data['Calories'] > 500) & (dst\_Data['Pulse'] < 100)]

print(filter\_dst\_Data2)

A screenshot of a computer

Description automatically generated

#(h)Create a new “df\_modified” dataframe that contains all the columns from dst\_data except for

#“Maxpulse”.

df\_modified = dst\_Data.loc[:, dst\_Data.columns != 'Maxpulse']

print(df\_modified)

A screenshot of a computer

Description automatically generated

#(i). Delete the “Maxpulse” column from the main dst\_data dataframe

dst\_Data.drop('Maxpulse', inplace=True, axis=1)

print(dst\_Data.dtypes)

A screenshot of a computer

Description automatically generated

#(j). Convert the datatype of Calories column to int datatype

dst\_Data["Calories"] = dst\_Data["Calories"].astype(float).astype(int)

print(dst\_Data.dtypes)

A screen shot of a computer

Description automatically generated

#(k)Using pandas create a scatter plot for the two columns (Duration and Calories).

as1 = dst\_Data.plot.scatter(x=‘Duration',y='Calories')

print(as1)

A screen shot of a computer screen

Description automatically generated

Question 2:

2. Linear Regression

a) Import the given “Salary\_Data.csv”

b) Split the data in train\_test partitions, such that 1/3 of the data is reserved as test subset.

c) Train and predict the model.

d) Calculate the mean\_squared error

e) Visualize both train and test data using scatter plot.

# 2(a) Import the given “Salary\_Data.csv”

dst\_Sal = pd.read\_csv('/Users/manasavathumilli/Downloads/Salary\_Data (2).csv')

dst\_Sal.info()

dst\_Sal.head()

A screenshot of a computer

Description automatically generated

A = dst\_Sal.iloc[:, :-1].values #excluding last column i.e., years of experience column

B = dst\_Sal.iloc[:, 1].values #only salary column

# (b) Split the data in train\_test partitions, such that 1/3 of the data is reserved as test subset.

from sklearn.model\_selection import train\_test\_split

A\_train, A\_test, B\_train, B\_test = train\_test\_split(A, B, test\_size=1/3, random\_state=0)

# (c) Train and predict the model.

from sklearn.linear\_model import LinearRegression

reg = LinearRegression()

reg.fit(A\_train, B\_train)

B\_Pred = reg.predict(A\_test)

B\_Pred

A screenshot of a computer

Description automatically generated

# (d) Calculate the mean\_squared error

S\_error = (B\_Pred - B\_test) \*\* 2

Sum\_Serror = np.sum(S\_error)

mean\_squared\_error = Sum\_Serror / B\_test.size

mean\_squared\_error

A screenshot of a computer

Description automatically generated

# (e) Visualize both train and test data using scatter plot.

import matplotlib.pyplot as plt

# Training Data set

plt.scatter(A\_train, B\_train)

plt.plot(A\_train, reg.predict(A\_train), color='red')

plt.title('Training Set')

plt.show()

# Testing Data set

plt.scatter(A\_test, B\_test)

plt.plot(A\_test, reg.predict(A\_test), color='red')

plt.title('Testing Set')

plt.show()

A screenshot of a computer

Description automatically generated

Video Link: <https://drive.google.com/file/d/1H858VN1kGQtC9TVImxGlI-WI1CrCJfTm/view?usp=sharing>

GitHub Link: <https://github.com/Manasav17/NNDL_ICP4>

SourceCode Link: <http://localhost:8888/notebooks/Documents/Neural/NNDL_ICP_4.ipynb>