Homework 4 *(due: Feb 17)*
Machine Learning - COSC 4360

Department of Computer Science and Electrical Engineering

Spring 2025

Exercises

*Create a **New Project** for every exercise. Take a screenshot of the source code along with its output and place the **source code** and the **screenshot** in a **zipped folder** named **LastNameFirstName_HW4***

Exercise 1

Given the following dataset: *avgHigh_jan_1895-2018.csv*, perform **Simple Linear Regression** using the first two columns of the dataset. **Predict** temperatures for the following *three* dates: Jan 2019, Jan 2023, Jan 2024. You may use any *built-in* functions you wish. Your output should resemble Fig. 1 below.

Note: Ignore the column *Anomaly*. It is the difference between the temperature for the given date and the average temperatures for all dates.

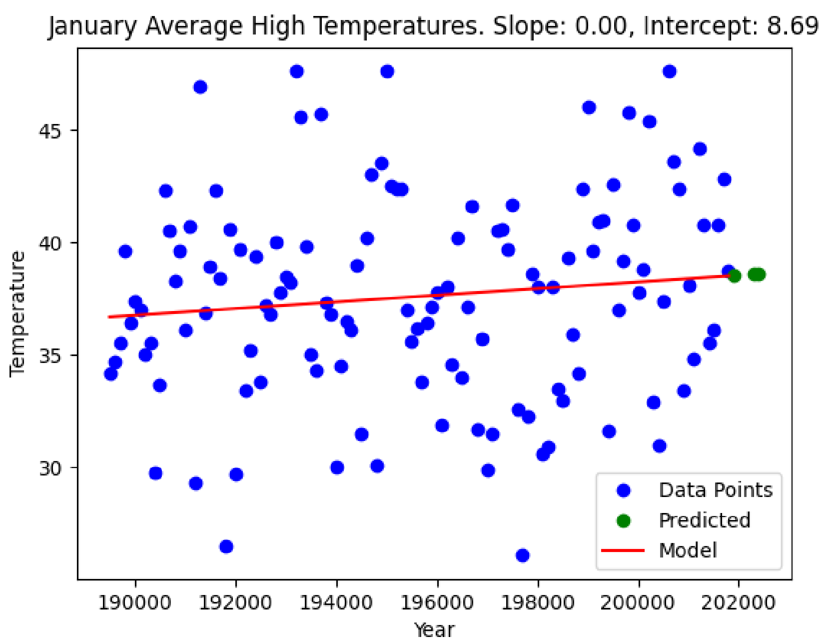

Figure 1: Simple Linear Regression between dates of the year and temperature.

Exercise 2

Given the following dataset: *avgHigh_jan_1895-2018.csv*, **split** your data into **training** and **test**. Perform **Simple Linear Regression** on the **training** dataset and use it to **predict** temperature values for the **test** dataset. The **test size** should be provided as input by the user. Print the **actual** temperatures from the **test** dataset as well as the **predicted** values. Then, compute the **Root Mean Square Error** (RMSE) between the **actual** temperatures and the **predicted** values from the **test** dataset. Your output should resemble Fig. 2 below.

Note 1: Ignore the column *Anomaly*. It is the difference between the temperature for the given date and the average temperatures for all dates.

Note 2: You may use any *built-in* functions you wish except from functions to *split* the dataset.

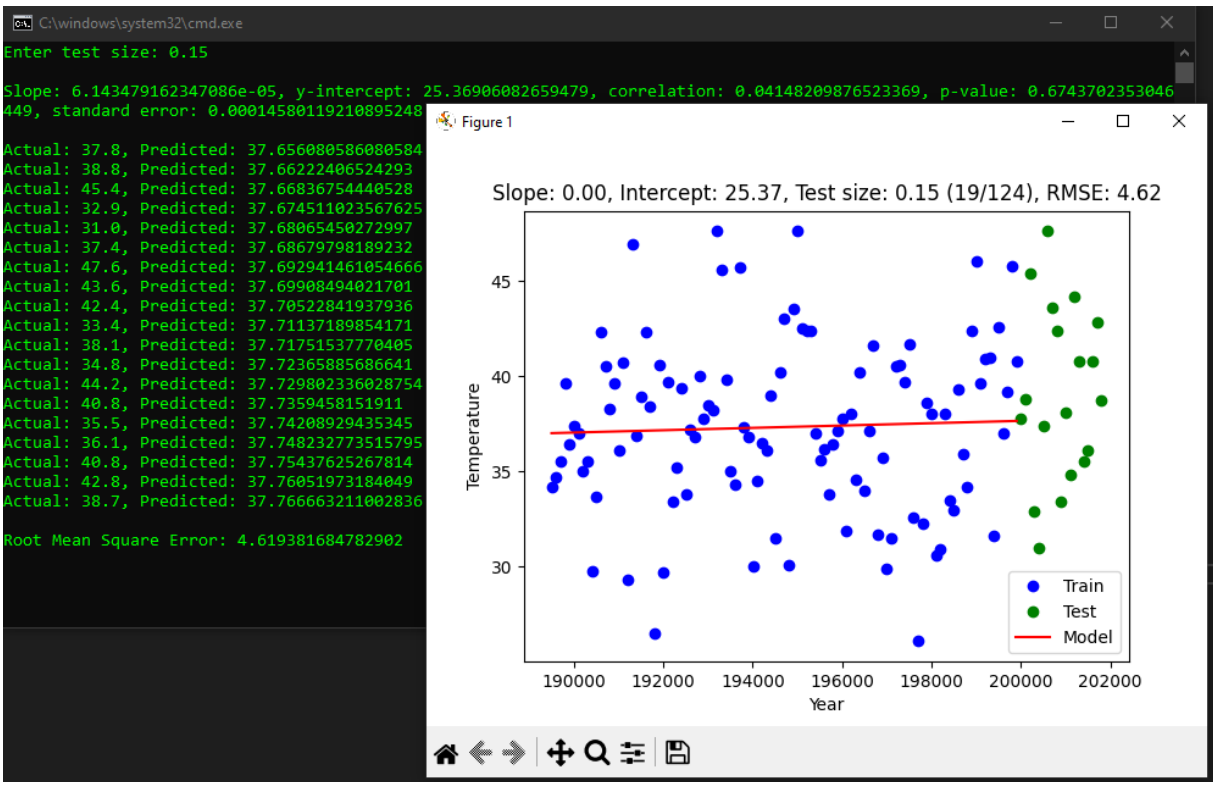

Figure 2: Simple Linear Regression between dates of the year and temperature.

Exercise 3

Given the following dataset: *materials.csv*, compute the **correlation coefficient**, **r**, between the response variable, that is, **Strength**, and each of the **predictor** variables. Then, using **Multiple Linear Regression**, predict **Strength** for the following two data points for *Time, Pressure, Temperature*, respectively:

32.1, 37.5, 128.95

36.9, 35.37, 130.03

Note: For prediction, do not use any *built-in* functions and do not *hard-code* the coefficients (you can, for example, use a loop). Also, do *not* scale your data.

Exercise 4

Using the same dataset, and after retaining only the **two** features that correlate the most with the response variable, from Ex. 1, perform **Multiple Linear Regression** with those two features. Then, generate a **3D (meshgrid) plot** to visualize the relationship between the two features and the response variable. On the same figure, also create a **3D scatter** plot of the same independent and dependent variables, as shown in Fig. 3 below.

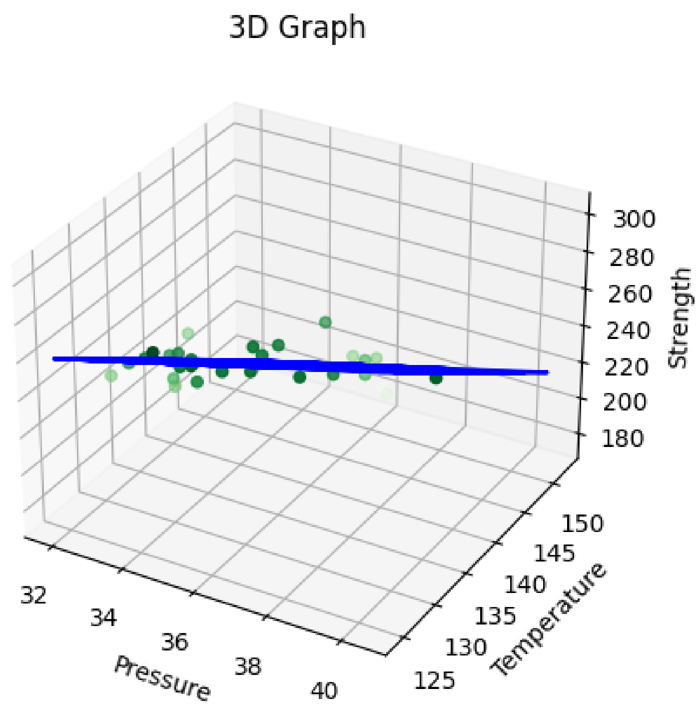

Figure 3: Multiple Linear Regression and 3D scatter plot between two predictor variables and the response variable.

Exercise 5

Given the following dataset: *materialsOutliers.csv*, use **RANSAC** to detect and remove **outliers** between each one of the independent variables and the dependent variable. Remove rows that contain outliers and perform **Multiple Linear Regression**.

Note 1: You will need to swap x with y and then apply **RANSAC**. You should also use the following two values in the **RANSACRegressor()** method: *residual_threshold=15, stop_probability=1.00*

Note 2: For more information, please refer to: RANSAC Regressor

Exercise 6 *(Optional)*

Implement your own version of the **RANSAC** algorithm.

Note: Submit through **Canvas**