

TANGLATEC SCHOOLSTRAAT 8 9883, BELLEM, BELGIUM +32486257049 info@tanglatec.com www.tanglatec.com VAT: BE0644.780.972

DETAILED WEEKLY REPORT

MACHINE LEARNING AND DATA SCIENCE WEEK 1

Submitted by Nyuysemo Calglain

Under the supervision of

TANGLATEC

26/7/2024

**Task**

**with the provided data set create a model that will plot the two D data in c++**

Table of **Contents**

[**1** **Introduction** 3](#_Toc172927816)

[**2** **Daily activities summarized** 3](#_Toc172927817)

[**2.1** **Day 1: Research and material download** 3](#_Toc172927818)

[2.1.1 **Procedure**: 3](#_Toc172927819)

[2.1.2 **Success**: 3](#_Toc172927820)

[**2.1.3** **Difficulties** 3](#_Toc172927821)

[**2.2** 4](#_Toc172927822)

[**2.3** **Day 2: More module downloads and environment set-up and familiarization.** 4](#_Toc172927823)

[**2.3.1** **Procedure** 4](#_Toc172927824)

[**2.3.2** **Problems** 4](#_Toc172927825)

[**2.4** 4](#_Toc172927826)

[**2.5** **Day 3: Importation of csv file and learning some c++ concepts (classes, vectors, constructors)** 4](#_Toc172927827)

[**2.5.1** **Procedure:** 4](#_Toc172927828)

[**2.6** **Day 4: Data description** 7](#_Toc172927829)

[**2.6.1** **Procedure** 7](#_Toc172927830)

[**2.7** **Day 5: Report writing, graphs plotting, code arrangement and upload to GitHub.** 7](#_Toc172927831)

[**2.7.1** **Procedure** 8](#_Toc172927832)

[8](#_Toc172927833)

[**3** **General Conclusion** 9](#_Toc172927834)

[**4** **Difficulties** 10](#_Toc172927835)

[**5** **Sources** 10](#_Toc172927836)

**Table of figures**

[**Figure 1 ETL.h file day 3** 6](#_Toc172929237)

[**Figure 2 ETL.cpp file day 3** 7](#_Toc172929238)

[**Figure 3 ETL.cpp day 4** 8](file:///C:\Users\CarlStorm\Desktop\internships2024\Tangla%20Tech%20internship\weekly%20report%20week2.docx#_Toc172929239)

[**Figure 4 main.cpp final** 9](#_Toc172929240)

[**Figure 5 output** 10](#_Toc172929241)

[**Figure 6 Regression code** 11](file:///C:\Users\CarlStorm\Desktop\internships2024\Tangla%20Tech%20internship\weekly%20report%20week2.docx#_Toc172929242)

[**Figure 7 Regression curve** 11](file:///C:\Users\CarlStorm\Desktop\internships2024\Tangla%20Tech%20internship\weekly%20report%20week2.docx#_Toc172929243)

[**Figure 8 Histogram code** 11](file:///C:\Users\CarlStorm\Desktop\internships2024\Tangla%20Tech%20internship\weekly%20report%20week2.docx#_Toc172929244)

[**Figure 9 Histogram** 12](file:///C:\Users\CarlStorm\Desktop\internships2024\Tangla%20Tech%20internship\weekly%20report%20week2.docx#_Toc172929245)

[**Figure 10 Gnuplot terminal execution** 12](file:///C:\Users\CarlStorm\Desktop\internships2024\Tangla%20Tech%20internship\weekly%20report%20week2.docx#_Toc172929246)

# **Introduction**

This report contains a detailed weekly process of how I worked, the difficulties encountered, success and failures and lessons learned.

The task that was assigned to me for internship week 2 was to be able to use the identified data set provided during week 1 to come out with a visual (graphical) representation of the data in c++.

# **Daily activities summarized**

* **Day 1:** Research and material download
* **Day 2:** More module downloads and environment set-up
* **Day 3:** Importation of csv file and learning some c++ concepts (classes, vectors, constructors)
* **Day 4**: Data description
* **Day 5**: Report writing, graphs plotting, code arrangement and upload to GitHub.

## **Day 1: Research and material download**

### **Procedure**:

I was able to gather up some few materials for machine learning with c++ and also found out the steps in setting up my environment with c++. The software and libraries I need are as follows;

* A c++ IDE (gcc)
* IDE (visual studio code)
* c++ machine learning libraries (Eigen, Gnuplot-cpp,boost)
* Build tools (CMake)

### **Success**:

Was able to download gcc, visual studio code, Eigen, CMake.

### **Difficulties**

* There is limited material on machine learning with cpp on the web but after some research I got my hands on something to start with.
* I could not completely set up my environment due to network issues so I had to wake up at midnight the following day when the internet speed Is usually good for the complete set up of the environment and familiarization with the environment.

## 

## **Day 2: More module downloads and environment set-up and familiarization.**

### **Procedure**

Due to difficulties of network speed I had to continue downloading the modules that I couldn’t download on day 1 which are gnuplot, boost and then went ahead to set up my environment on vs code by first downloading c ++ extensions; c/c++ extension pack, c/c++, c/c++ themes, CMake, Code Runner, CMake tools, gcc. I also installed the gnuplot preview to be able to see my plots on vs code, and Gnuplot to be able to type the commands with no error messages.

To be able to type gnut commands or any library I downloaded I had to include them to my path when installing.

### **Problems**

At first, I did not include to path and faced many errors. I then followed videos on YouTube to be able to add these extensions to my path.

## 

## **Day 3: Importation of csv file and learning some c++ concepts (classes, vectors, constructors)**

### **Procedure:**

I created a folder called ETL then

proceeded by creating **the ETL class** which stands for Export Transform and Load.

This class will permit us to read data from a csv file.

In each file we create we will import the necessary packages we are going to use some of them being **iostream, string,**

**boost**: used for input and output functions,

**Eigen**: it’s a library for matrices, linear algebra and vectors,

**Vectors**: permits us to declare and use vector arrays that is arrays where we don’t know the size yet

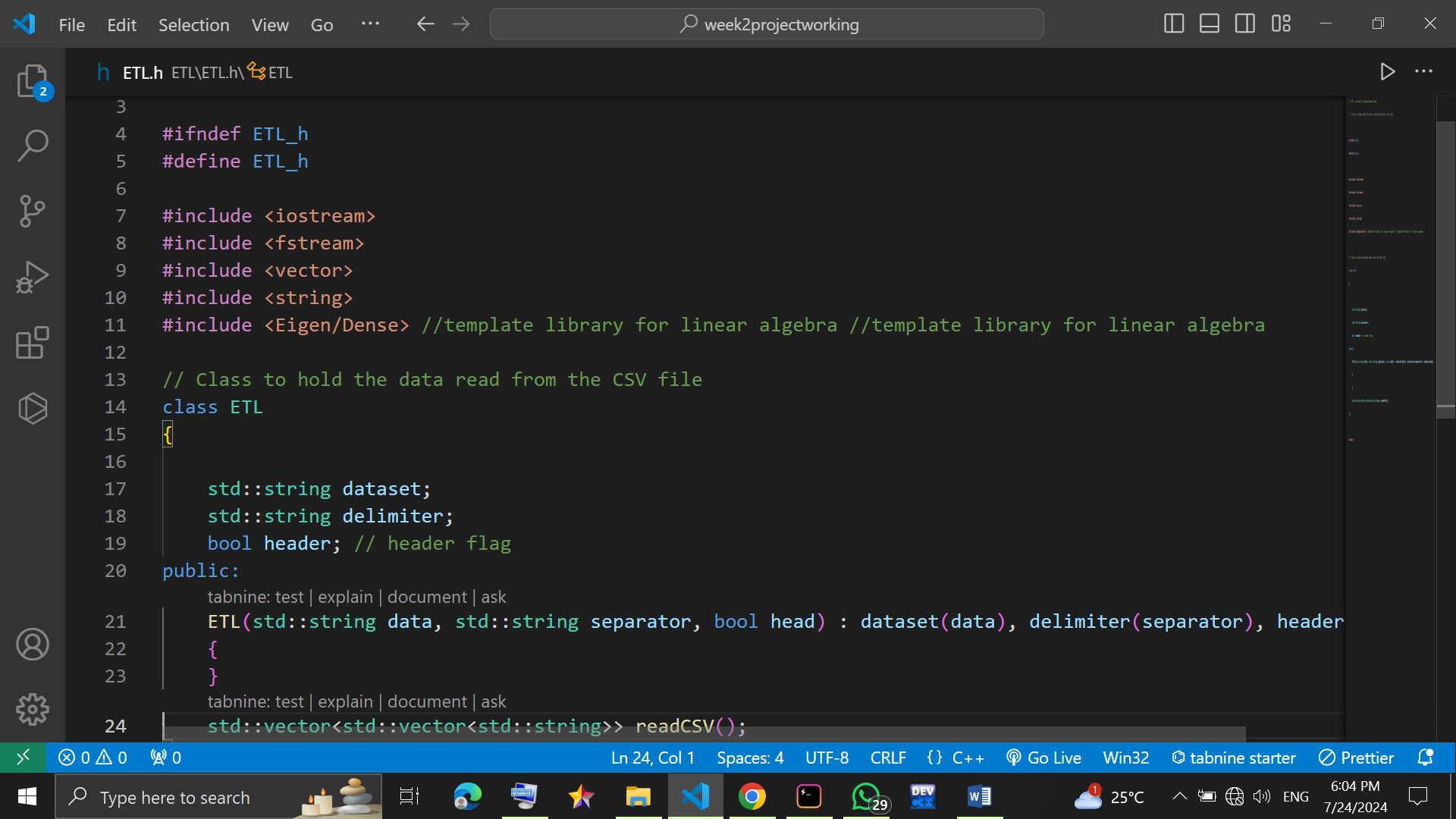
**Fstream:** used for file input and output.

Our ETL class has 3 attributes. The dataset that will collect our dataset in csv format, the delimiter or simply the separator of the x and y components and the header (header flag) to determine whether our data has a header or not. **We then create a constructor for this class**

In this class we also declare an instance of our function readCSV ()

That will read the csv file. as shown on the screenshot below.

**Screenshot:**



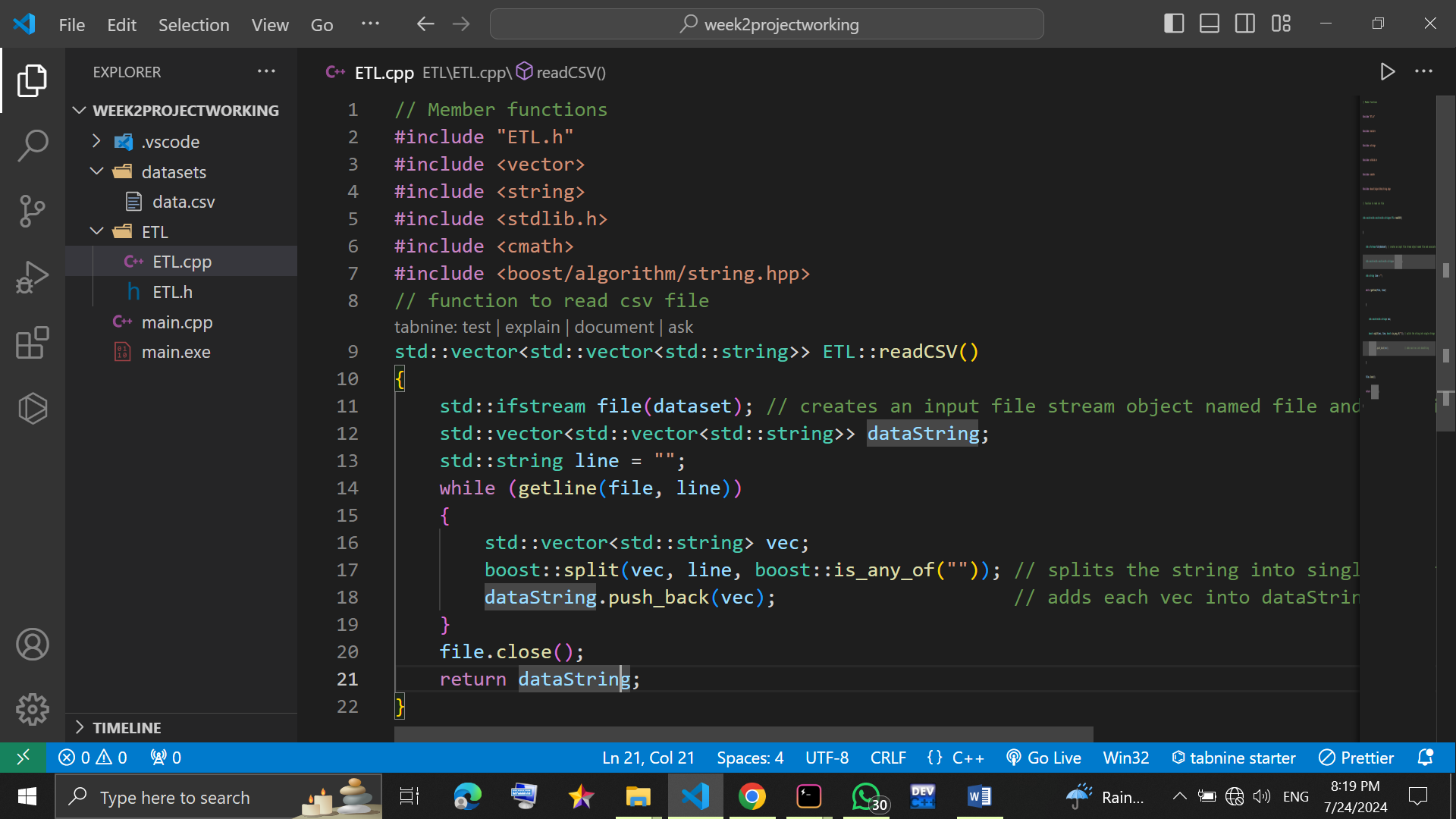
**Figure 1ETL.h file day 3**

We proceed by creating the file ETL.cpp in which we will build the functions for our ETL member class.

We import “ETL.h “header file and all the necessary libraries. Then we call an instance on the readCSV() function from the ETL library which returns a 2D vector string (data set) in csv format.

In this function we declare the various parameters that we will use such as dataString parameter which returns the actual data set in csv format. The code implementation is seen in the implementation below.

Screenshot:



**Figure 2ETL.cpp file day 3**

I created another folder called datasets and pasted my data set in the csv format as can be seen on the screenshot above.

## **Day 4: Data description**

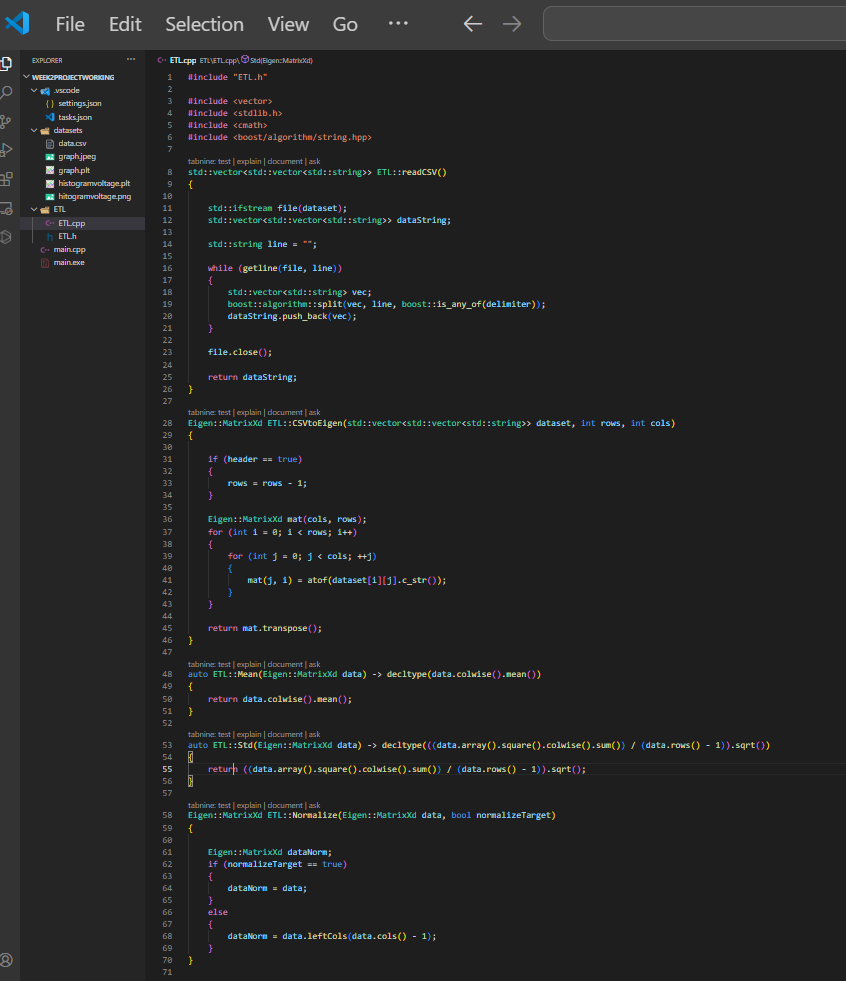
I was able to do some data description functions like the mean, variance and standard deviation. We never had light in the evening so I did not make a report.

### **Procedure**

To calculate the mean, standard deviation, Normalization I created an Eigen matrix instance in the ETL.h file and

Then create the implementation of the function in ETL.cpp as seen on the screenshots below.

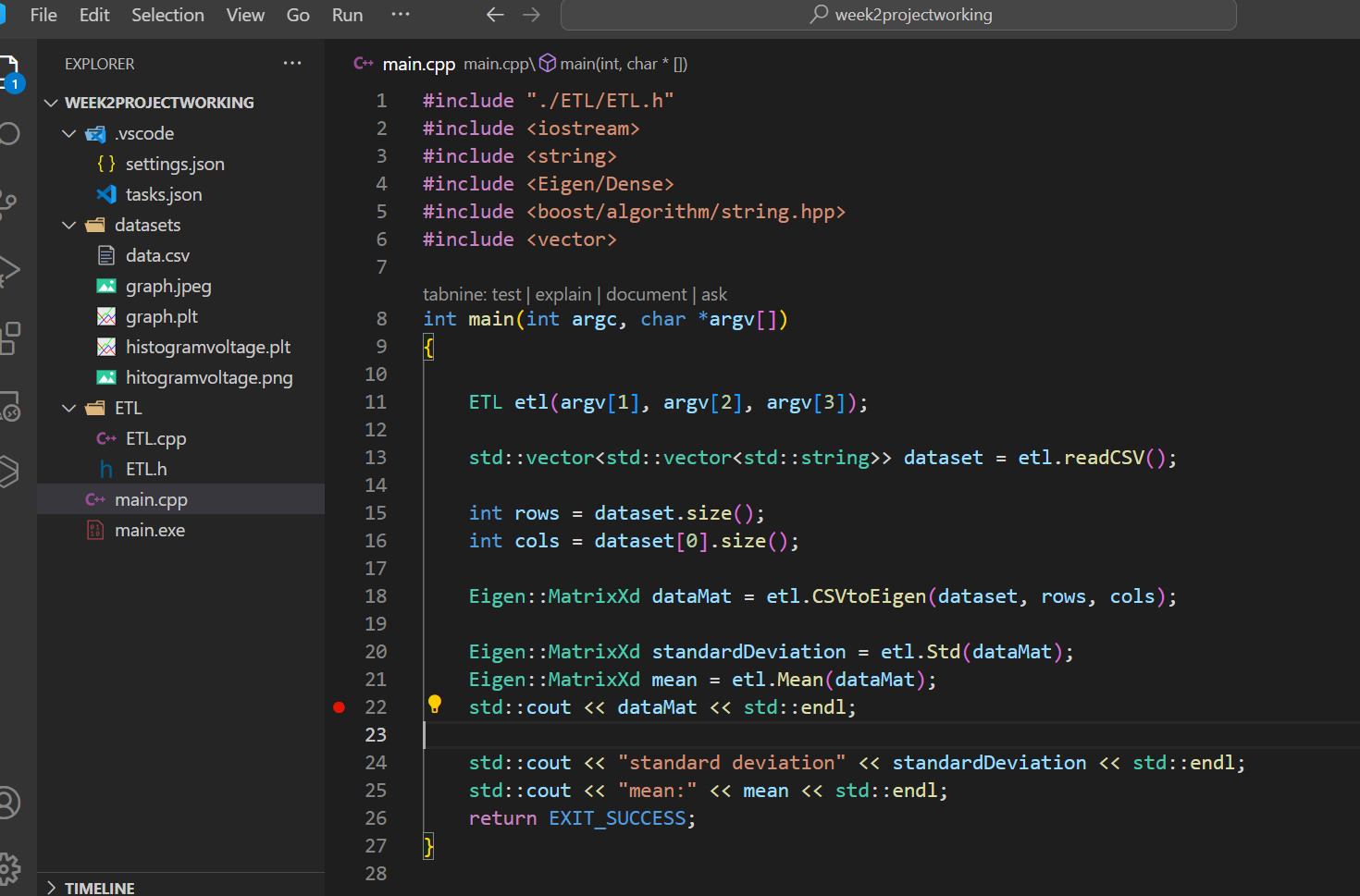
To test these functions, I called them in the main.cpp.

**screenshot:**

**Figure 3 ETL.cpp day 4**

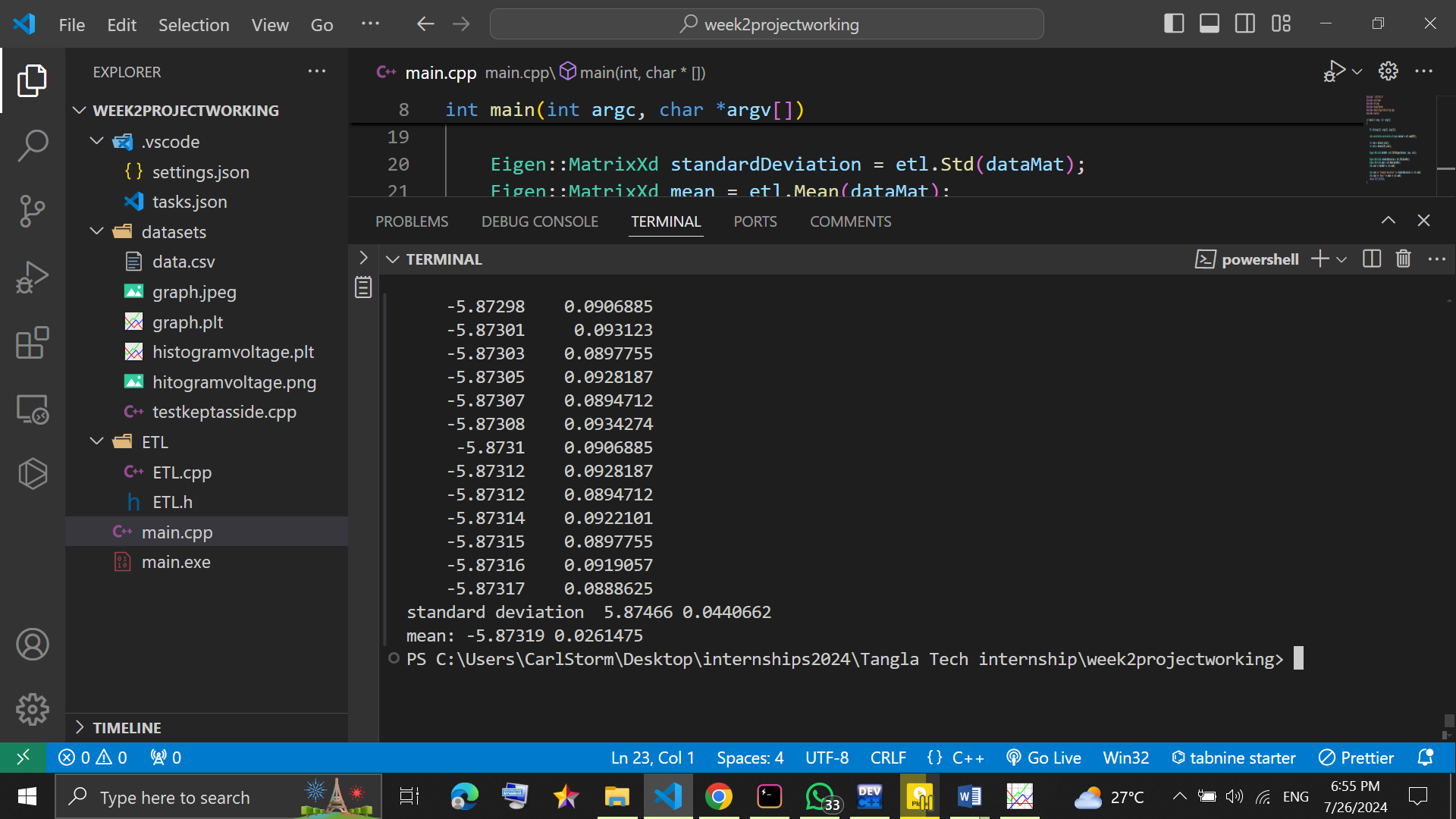
To test the implementation I create a main.cpp folder that will make reference to the functions we create like the readCSV() as seen on the screenshot below

Screenshot main.cpp file:



**Figure 4 main.cpp final**

Output:



**Figure 5 output**

## **Day 5: Report writing, graphs plotting, code arrangement and upload to GitHub.**

This so far was a busy day as I had to learn and apply plotting on gnuplot and also write a report for the whole week

### **Procedure**

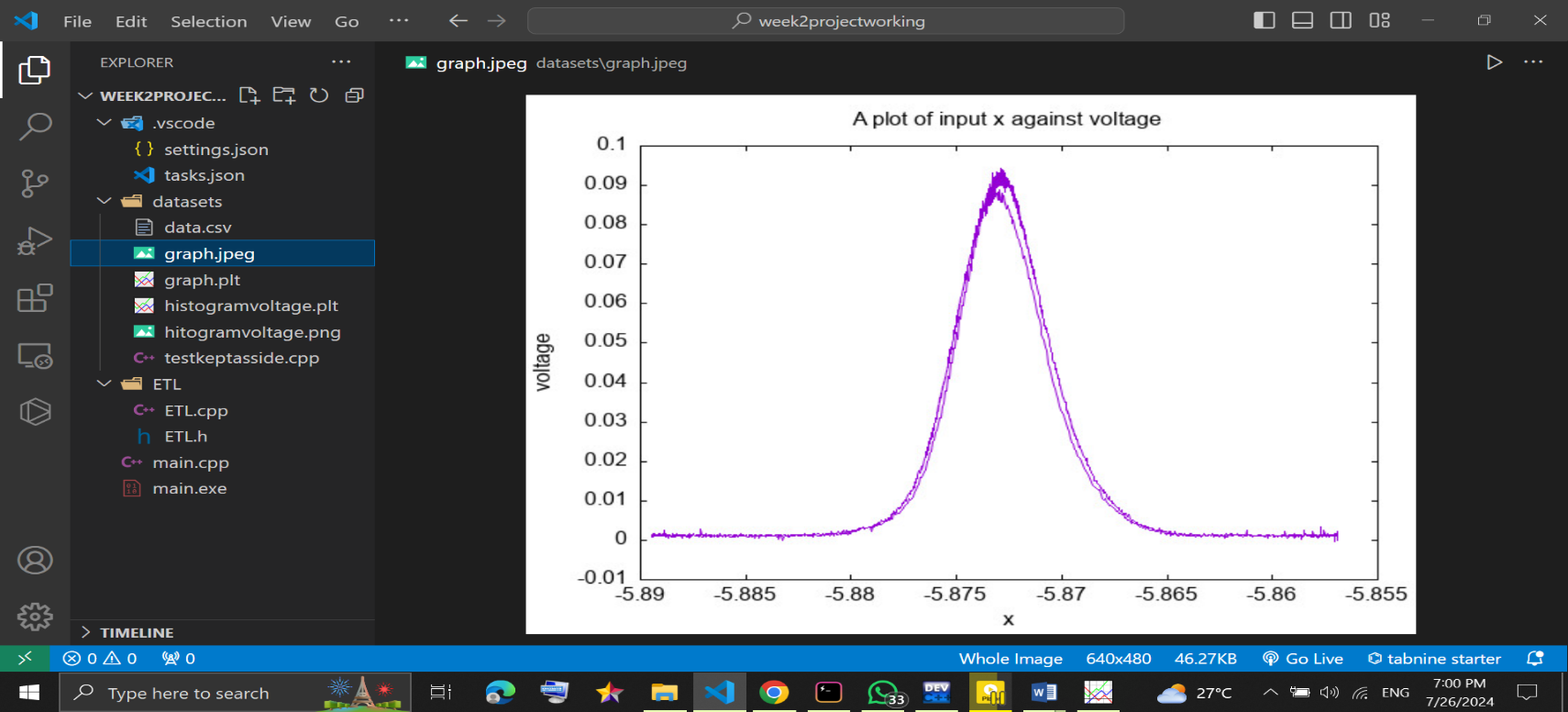
To plot on gnuplot I had earlier downloaded the application and the extensions on vs code I created a file with the extension of “.plt” so the compiler will recognize it’s a gnuplot file . I then proceed by typing the commands in the file and when am done I open the gnuplot terminal on my windows computer and load the file with the “.plt” extension.

I had two .plt files one to plot a histogram which did not go quite well and the order to plot the regression curve as explained by the screenshots below.

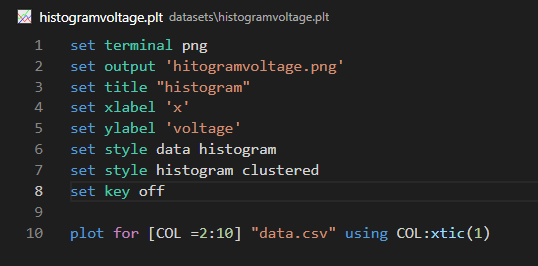
**Screenshot Regression curve:**

# 

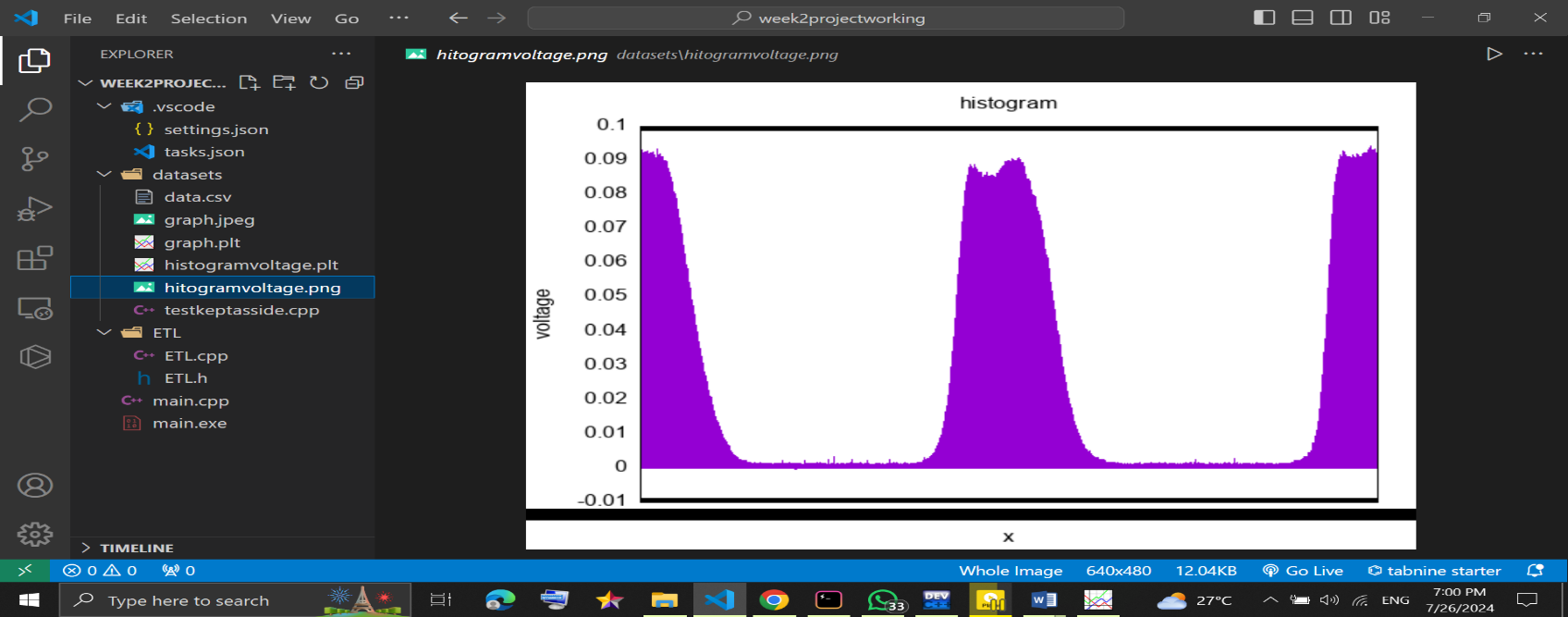
**Figure 6 Regression code**



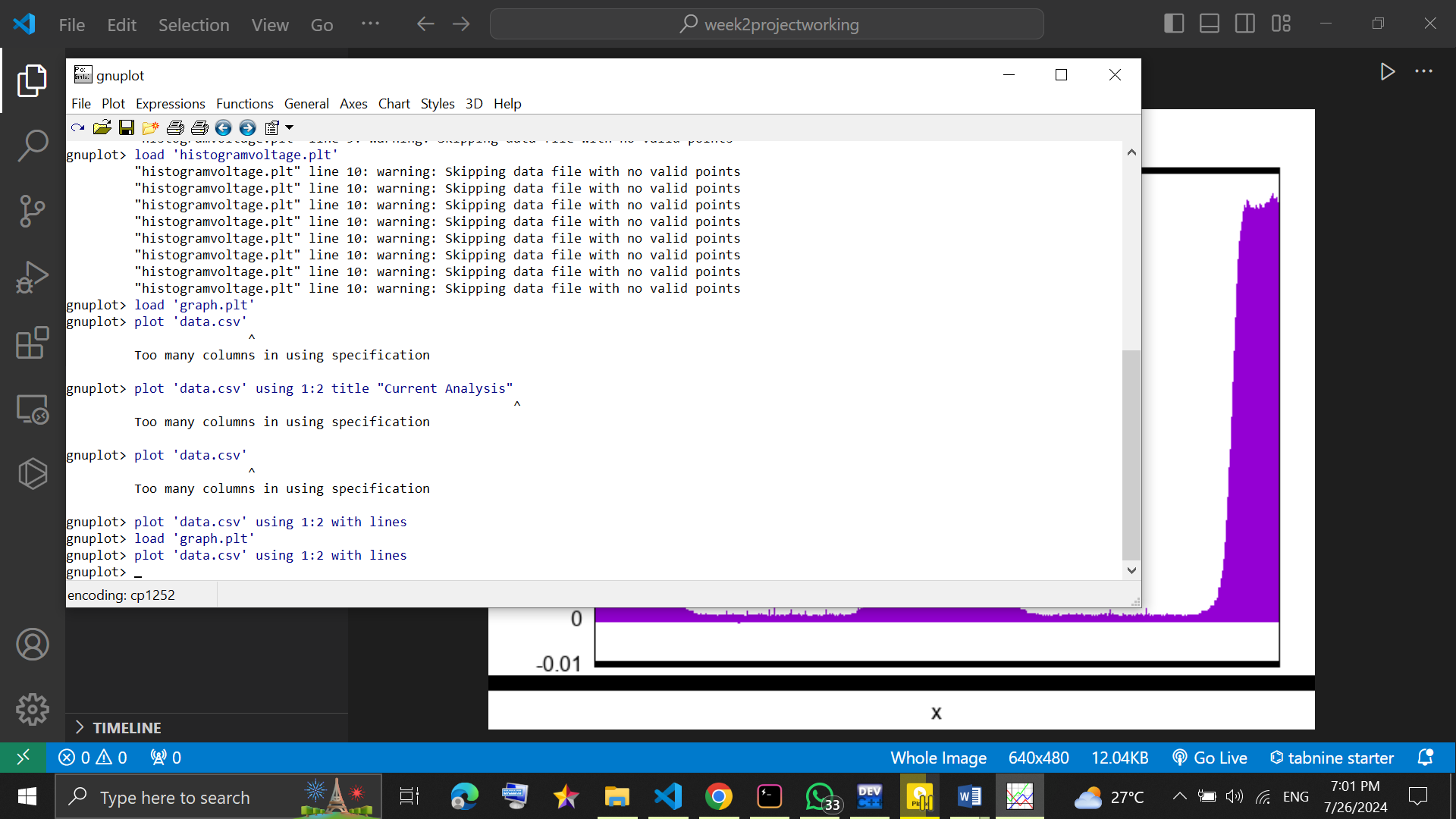
**Figure 7 Regression curve**

**Screenshot Histogram:**

**Figure 8 Histogram code**



**Figure 9 Histogram**

**Gnuplot execution screenshot**

**Figure 10 Gnuplot terminal execution**

# **General Conclusion**

During this week I was able to learn basics of c++ and other advance concepts like classes, constructors, functions and basics in using some c++ libraries to achieve my machine learning objectives it’s been a good but tiring process and I’ll continue to give in my all to become a machine learning and data scientist.

# **Difficulties**

I generally face difficulties of shortage of resources and slow internet connection. We had light seizure as well be it was minimal. Getting familiar with the environment has not been easy and I still have a long way to go. I was unable to build a model same as last week which am not proud to say but I will try to take the courses I need to be ready for that.

# **Sources**

* sourceForge (to download libraries)
* YouTube
* Google
* Chatgpt
* Friends
* Stack overflow