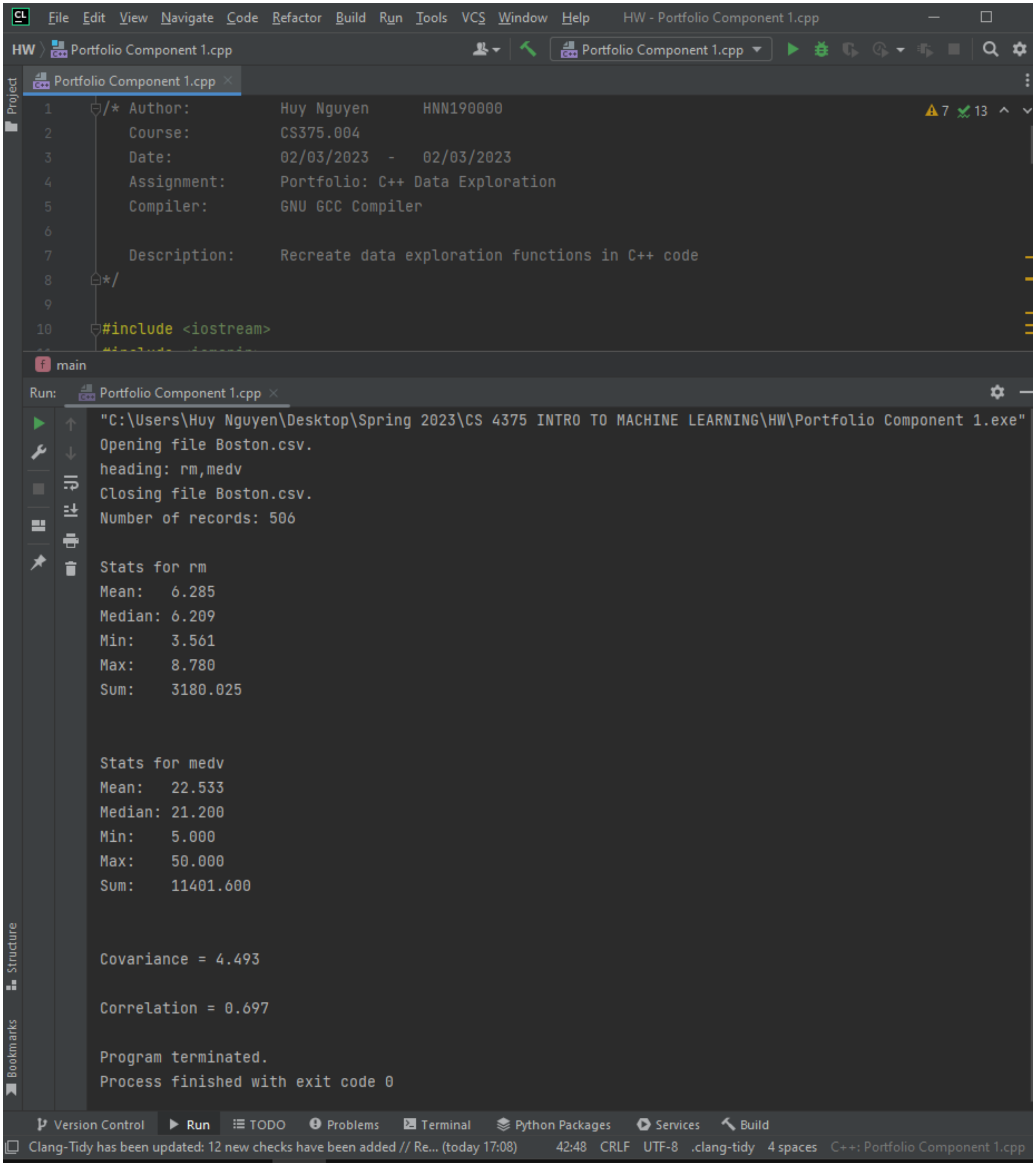
Huy Nguyen
HNN190000
CS 4375.004

Portfolio Component 1: Data Exploration

a. copy/paste runs of your code showing the output.


File
Edit
View
Navigate
Code
Refactor
Build
Run
Tools
VCS
Window
Help
HW - Portfolio Component 1.cpp
HW
Portfolio Component 1.cpp
Portfolio Component 1.cpp
Portfolio Component 1.cpp
Project
⚠7 ✓13
main
Run:
Portfolio Component 1.cpp
Structure
Bookmarks
Version Control
Run
TODO
Problems
Terminal
Python Packages
Services
Build
Clang-Tidy has been updated: 12 new checks have been added // Re... (today 17:08)
42:48
CRLF
UTF-8
.clang-tidy
4 spaces
C++: Portfolio Component 1.cpp

```
/*  Author:         Huy Nguyen      HNN190000
    Course:         CS375.004
    Date:           02/03/2023  -   02/03/2023
    Assignment:     Portfolio: C++ Data Exploration
    Compiler:       GNU GCC Compiler

    Description:    Recreate data exploration functions in C++ code
*/

#include <iostream>
```

```
"C:\Users\Huy Nguyen\Desktop\Spring 2023\CS 4375 INTRO TO MACHINE LEARNING\HW\Portfolio Component 1.exe"
Opening file Boston.csv.
heading: rm,medv
Closing file Boston.csv.
Number of records: 506

Stats for rm
Mean:   6.285
Median: 6.209
Min:    3.561
Max:    8.780
Sum:    3180.025


Stats for medv
Mean:   22.533
Median: 21.200
Min:    5.000
Max:    50.000
Sum:    11401.600


Covariance = 4.493

Correlation = 0.697

Program terminated.
Process finished with exit code 0
```

Huy Nguyen
HNN190000
CS 4375.004
Portfolio Component 1: Data Exploration

b. describing your experience using built-in functions in R versus coding your own functions in C++

In R studio, all I need to do was load the csv and run summary, cor, cov functions. Creating my own functions is much longer and tedious. I spent more time formatting the print statements than actually coding the sum, mean, median, and range. I had to take some time to refer to the textbook to refresh on the formula for covariance and correlation.

c. describes the descriptive statistical measures mean, median, and range, and how these values might be useful in data exploration prior to machine learning

Mean is the average of the data set. Median is the middle of the data. Range shows bounds of the data.

d. describes the covariance and correlation statistics, and what information they give about two attributes. How might this information be useful in machine learning?

Correlation is how closely related the data with another. Covariance is how much of a correlation there is between the data. This can be used to find trends to help predict future values.