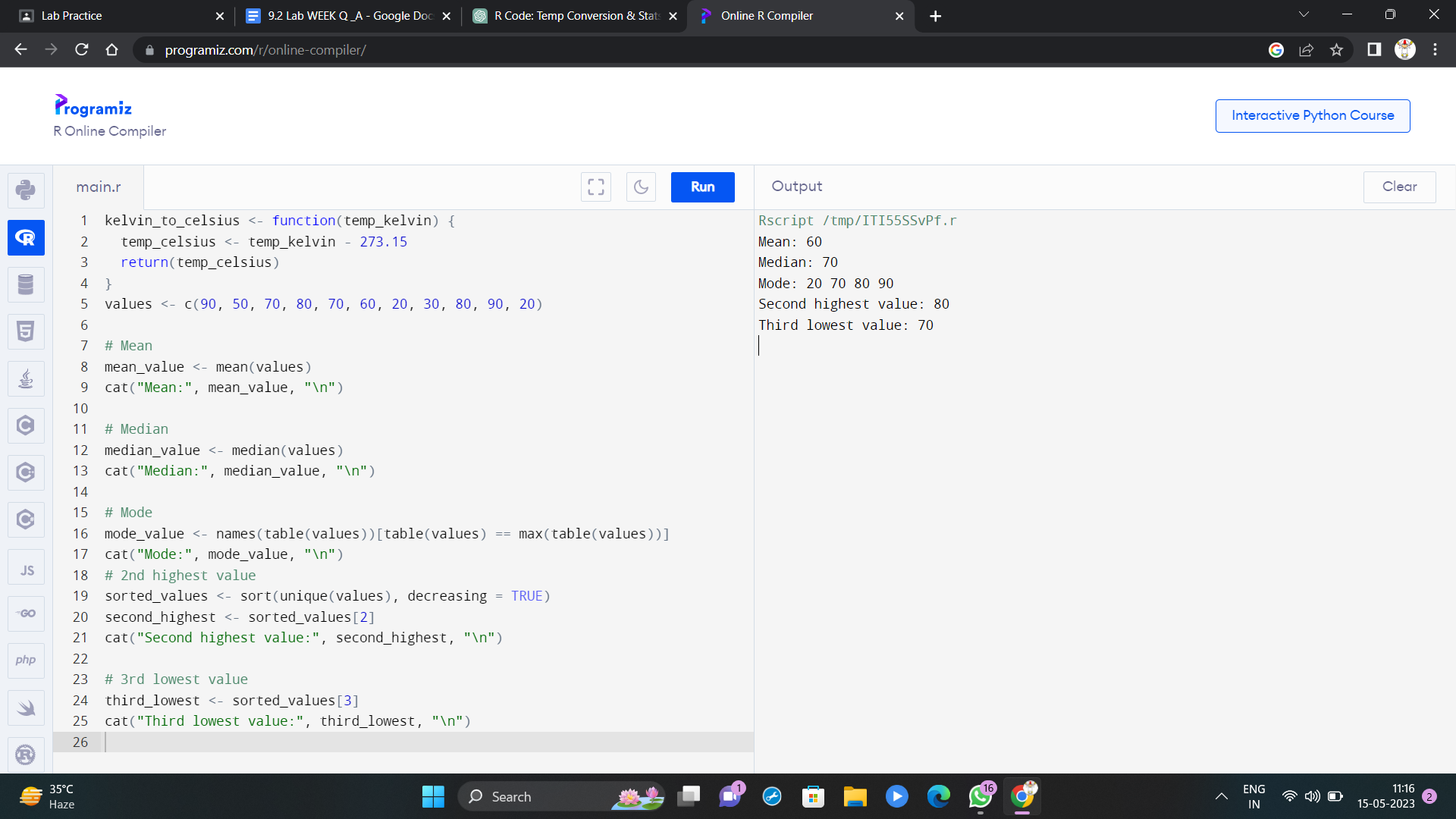
Set-I

1. .(i) Write a function called kelvin\_to\_celsius() that takes a temperature in Kelvin and returns that temperature in Celsius (**Hint:** To convert from Kelvin to Celsius you subtract 273.15) (ii) Write suitable R code to compute the mean, median ,mode of the following values c(90, 50, 70, 80, 70, 60, 20, 30, 80, 90, 20)

(iii) Write R code to find 2nd highest and 3rd Lowest value of above problem.



2. Explore the airquality dataset. It contains daily air quality measurements from New York during a period of five months:

• Ozone: mean ozone concentration (ppb),

• Solar.R: solar radiation (Langley),

• Wind: average wind speed (mph),

• Temp: maximum daily temperature in degrees Fahrenheit,

• Month: numeric month (May=5, June=6, and so on),

• Day: numeric day of the month (1-31).

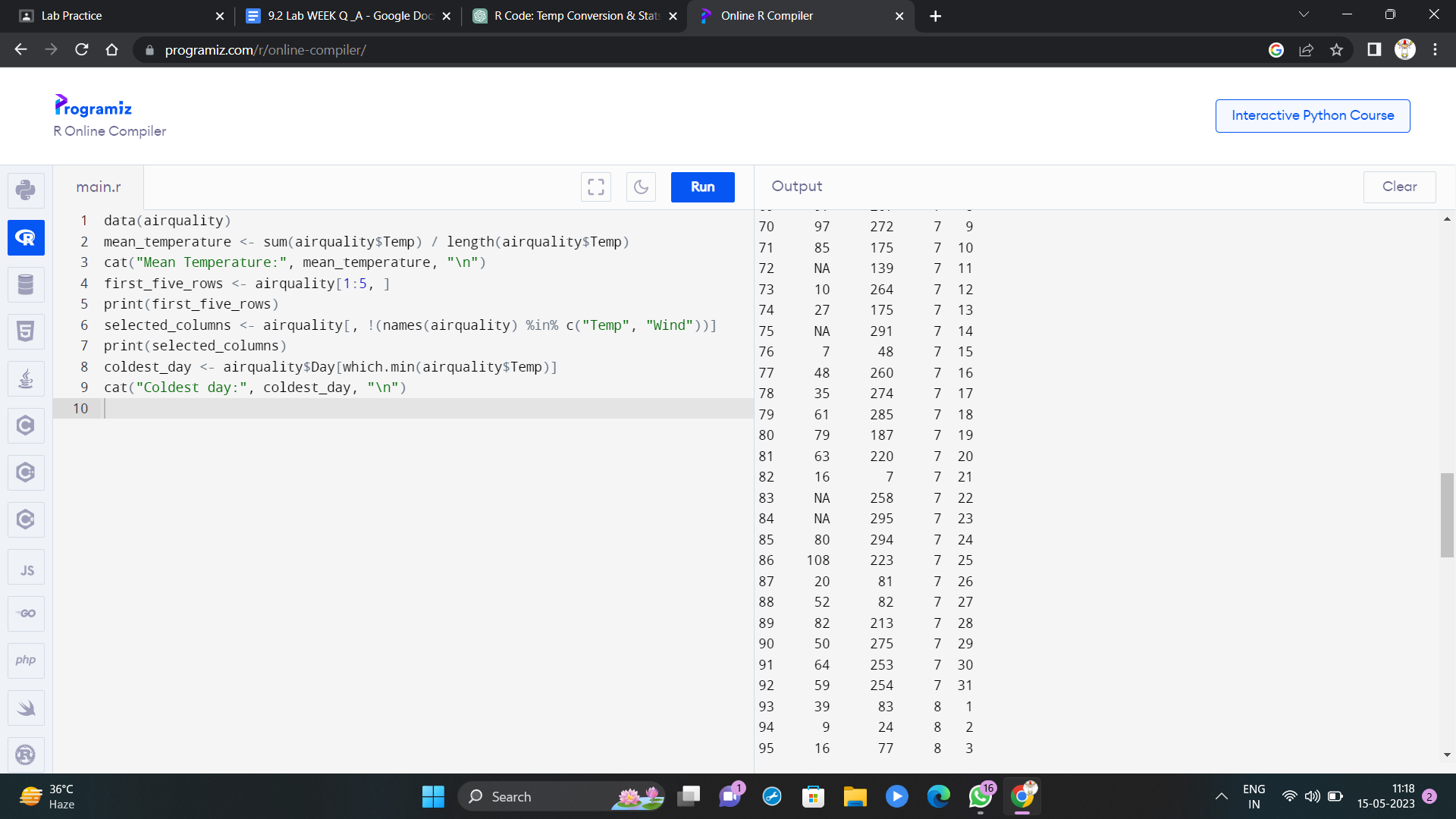
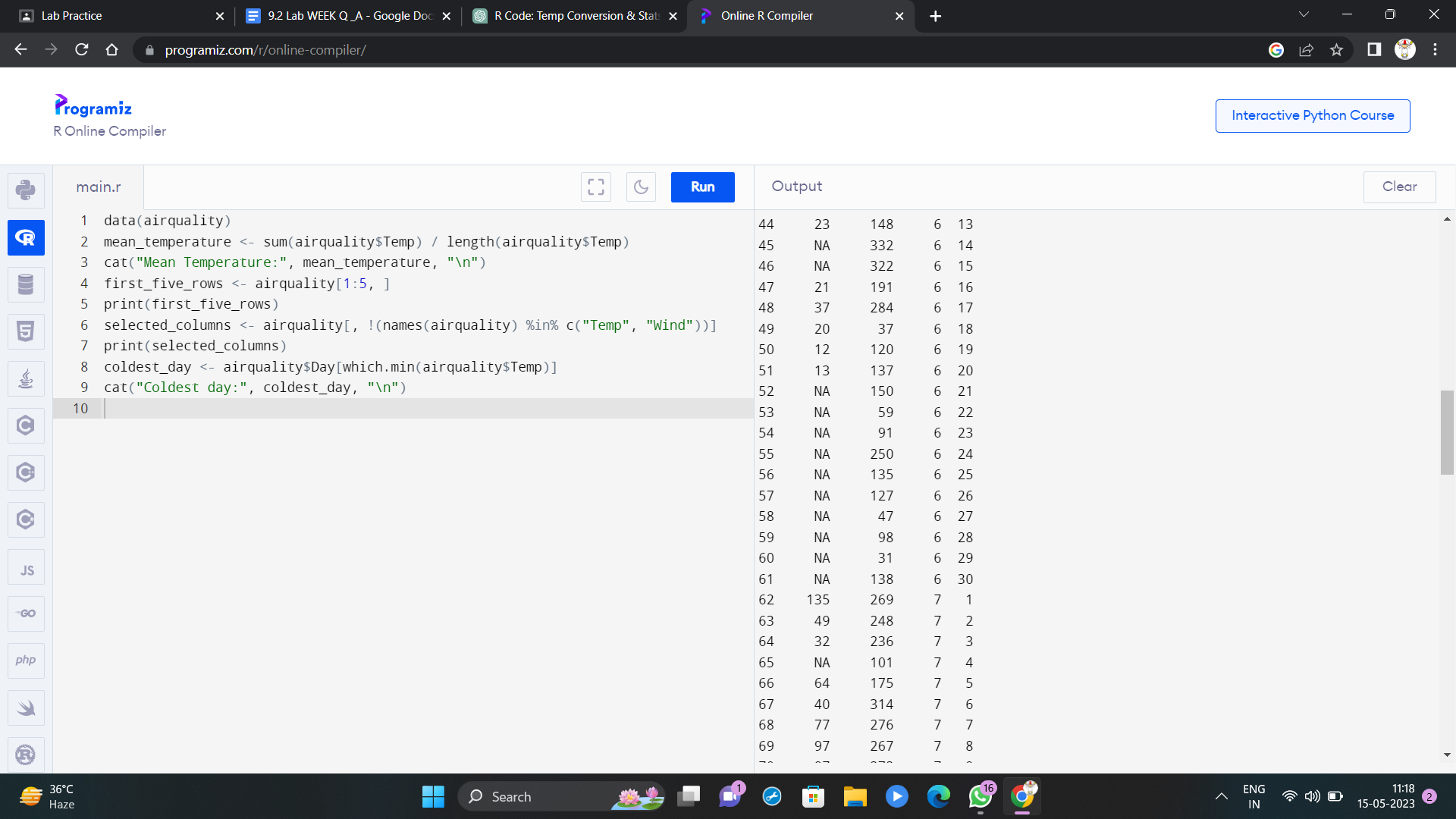
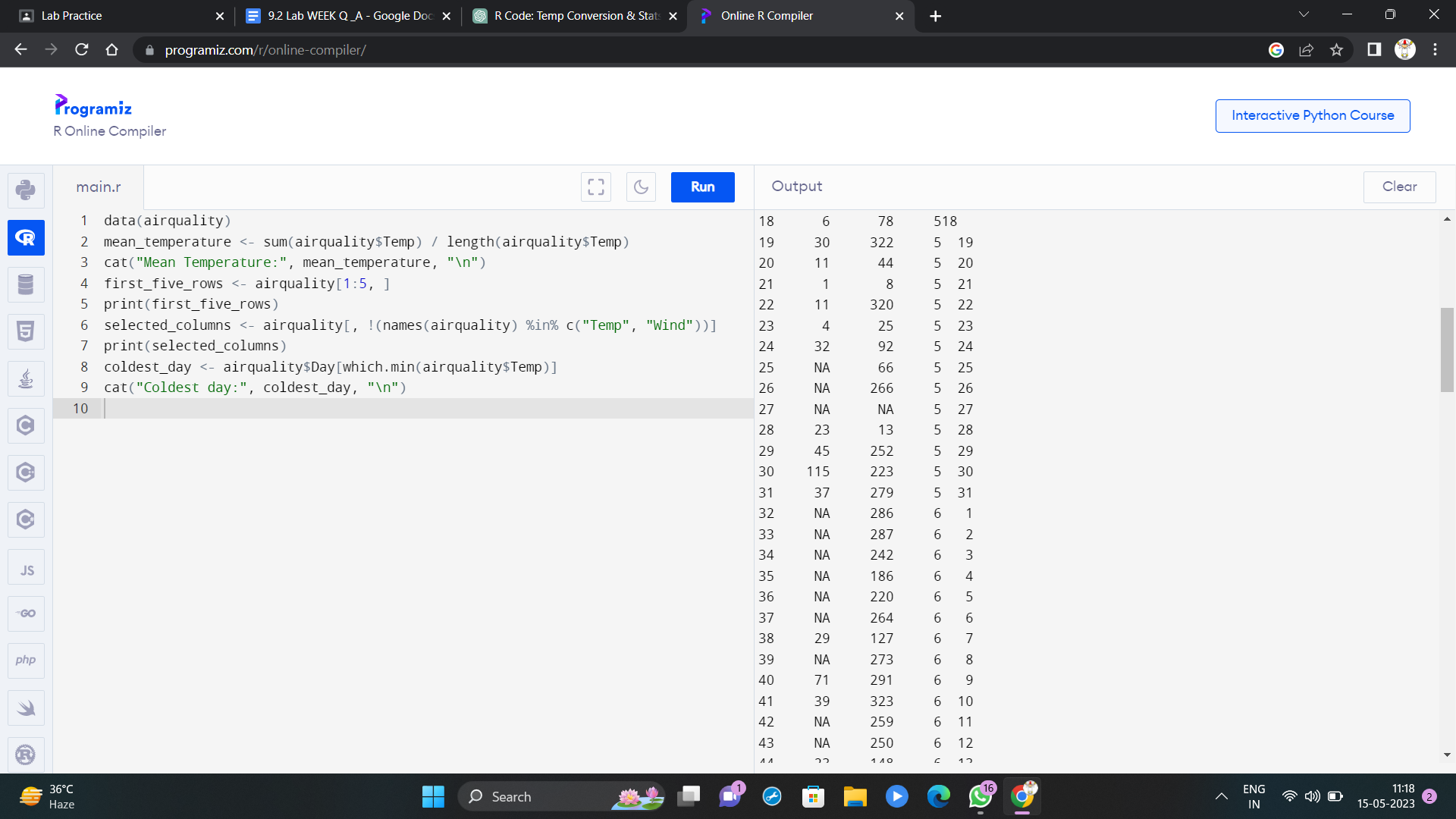
i. Compute the mean temperature(don’t use build in function)

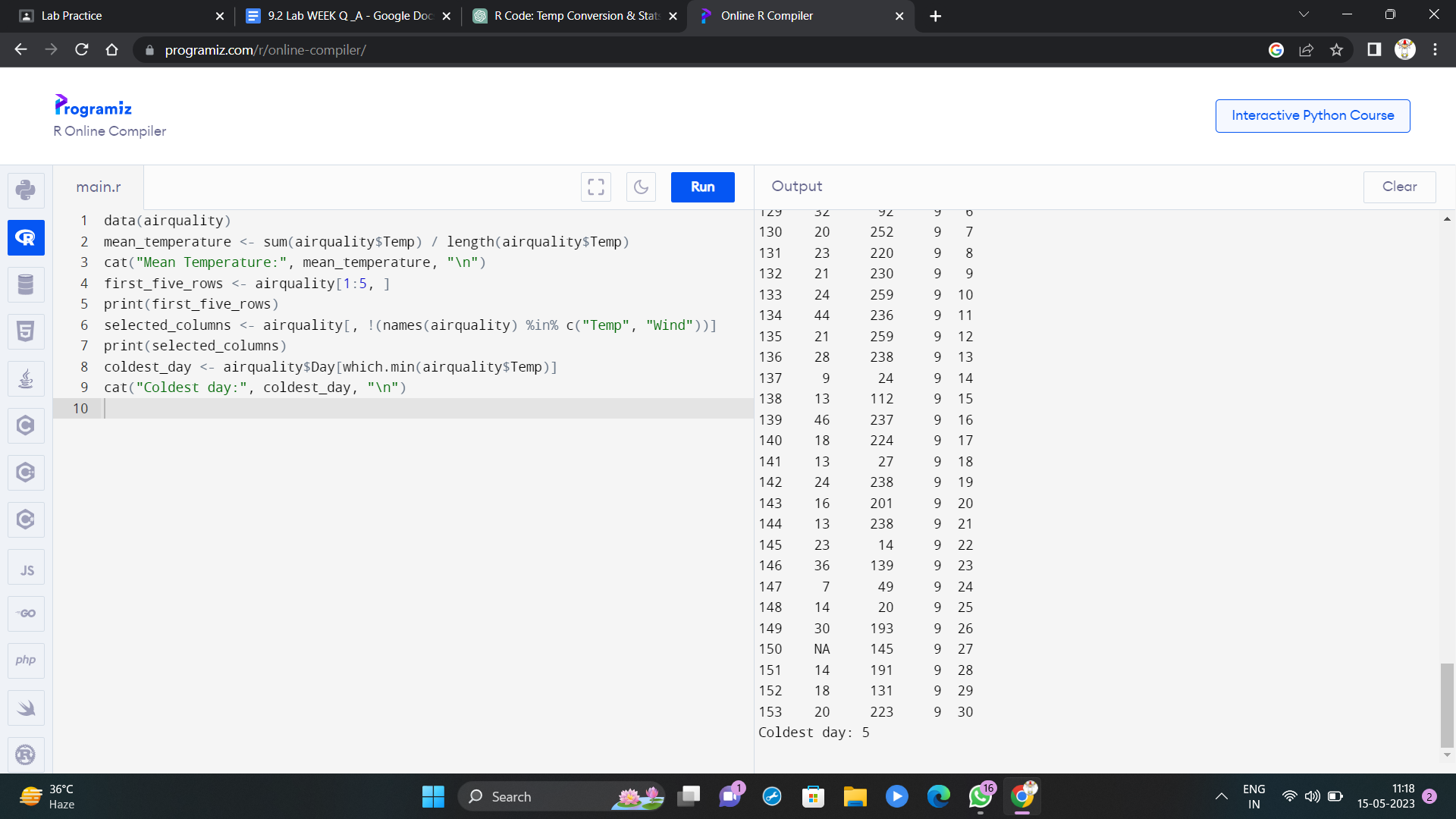
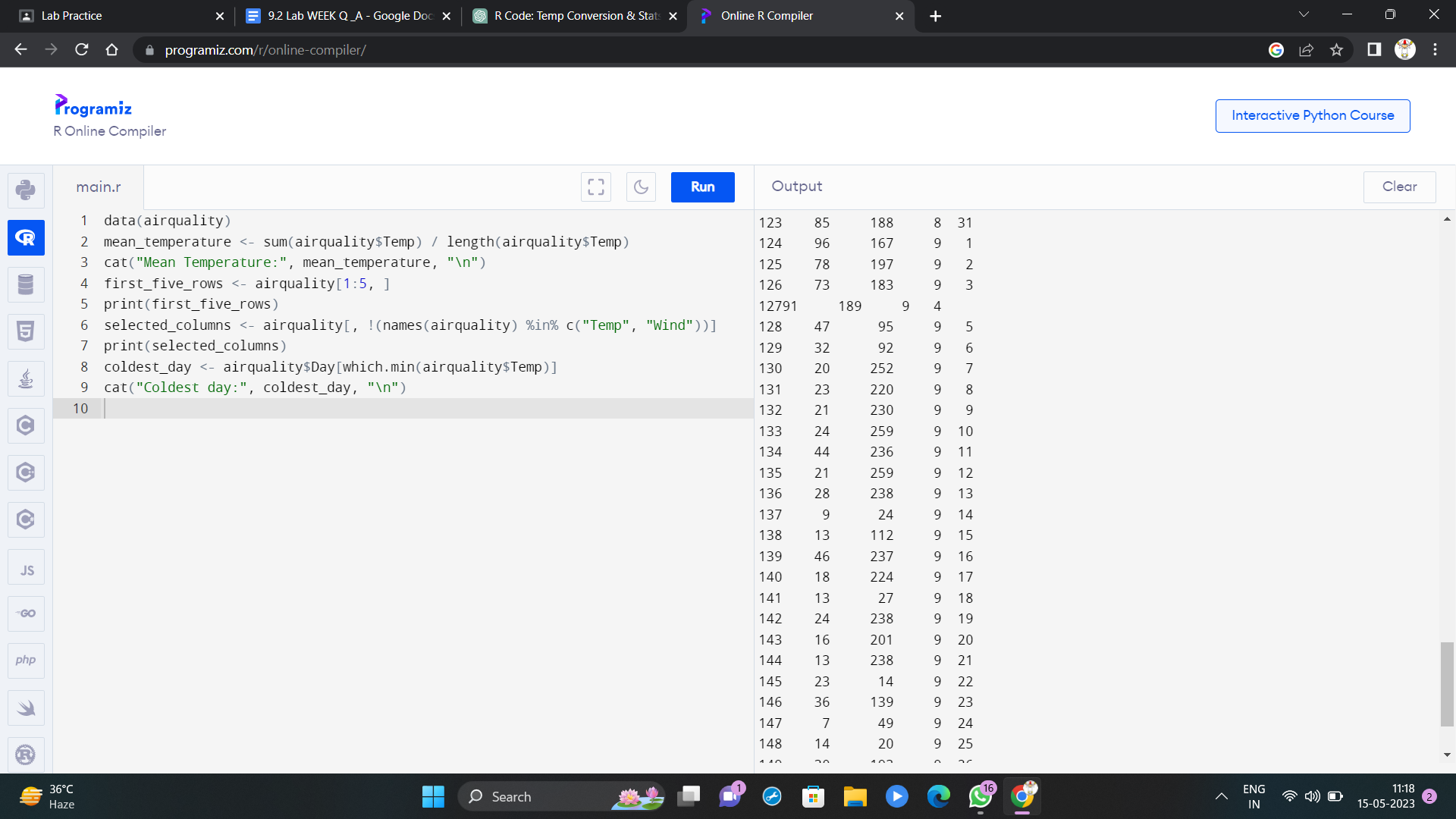
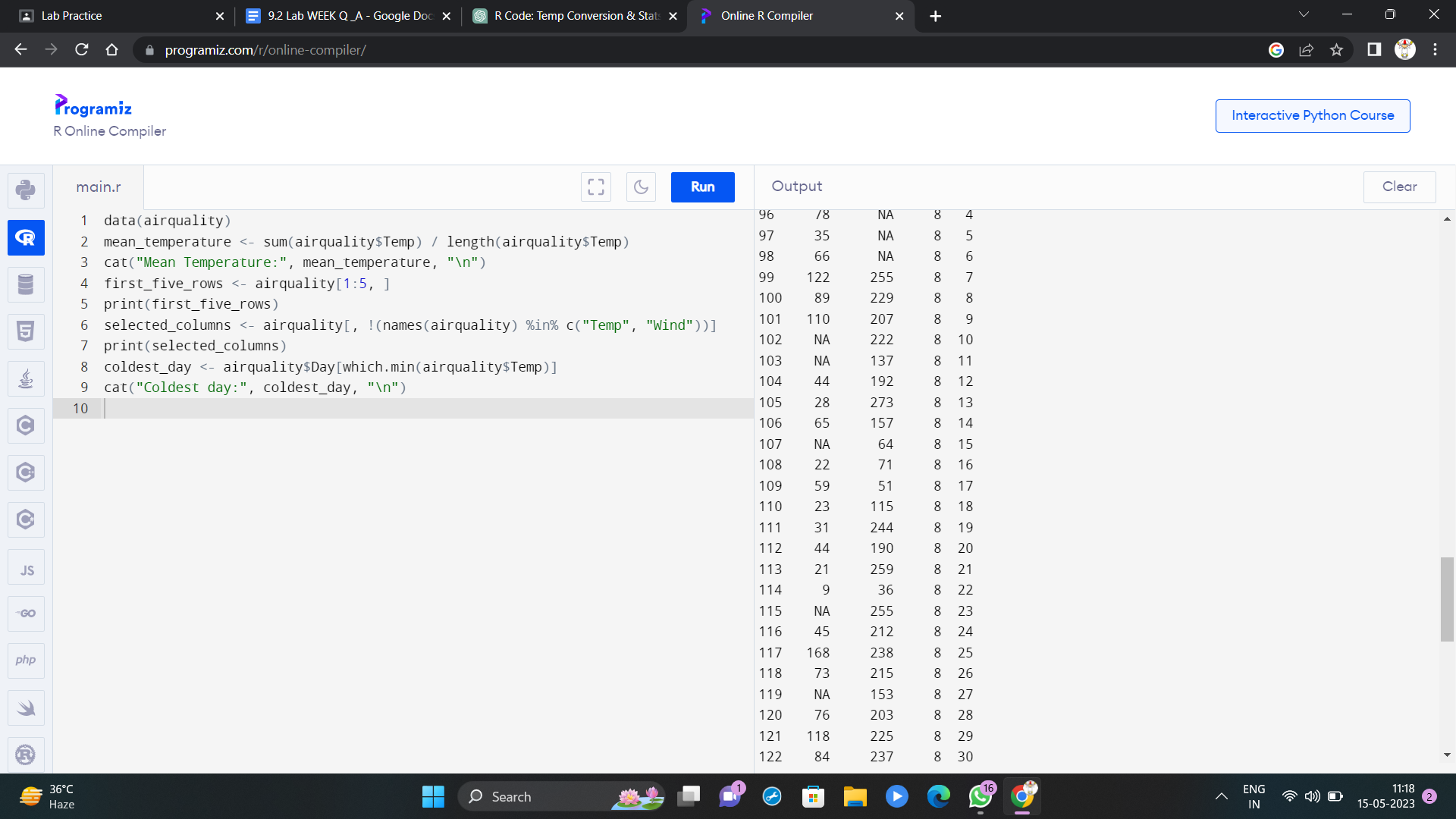
ii. Extract the first five rows from airquality.

iii. Extract all columns from airquality *except* Temp and Wind

iv. Which was the coldest day during the period?

v. How many days was the wind speed greater than 17 mph?





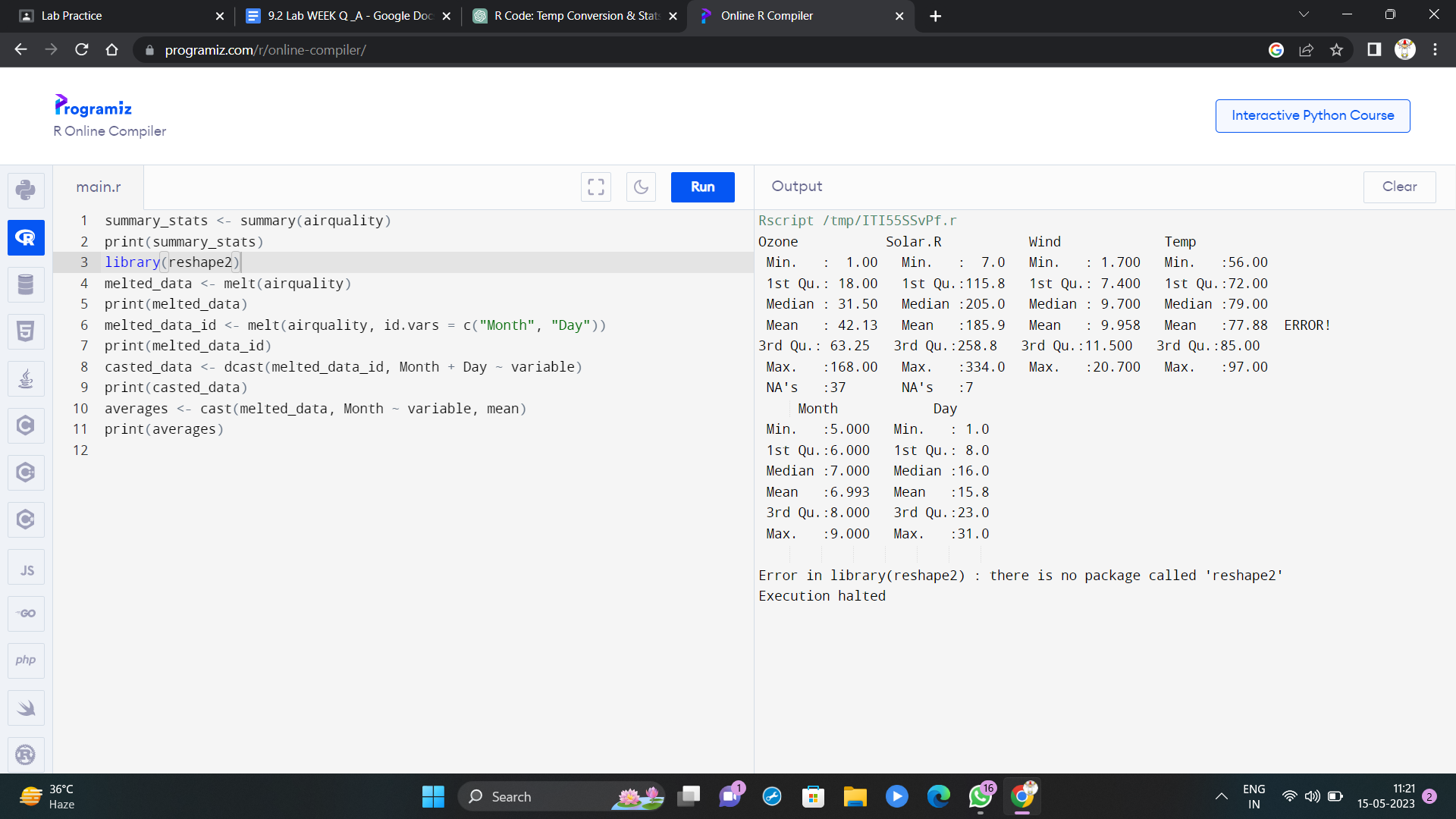
3. (i) Get the Summary Statistics of air quality dataset

(ii)Melt airquality data set and display as a long – format data?

(iii)Melt airquality data and specify month and day to be “ID variables”?

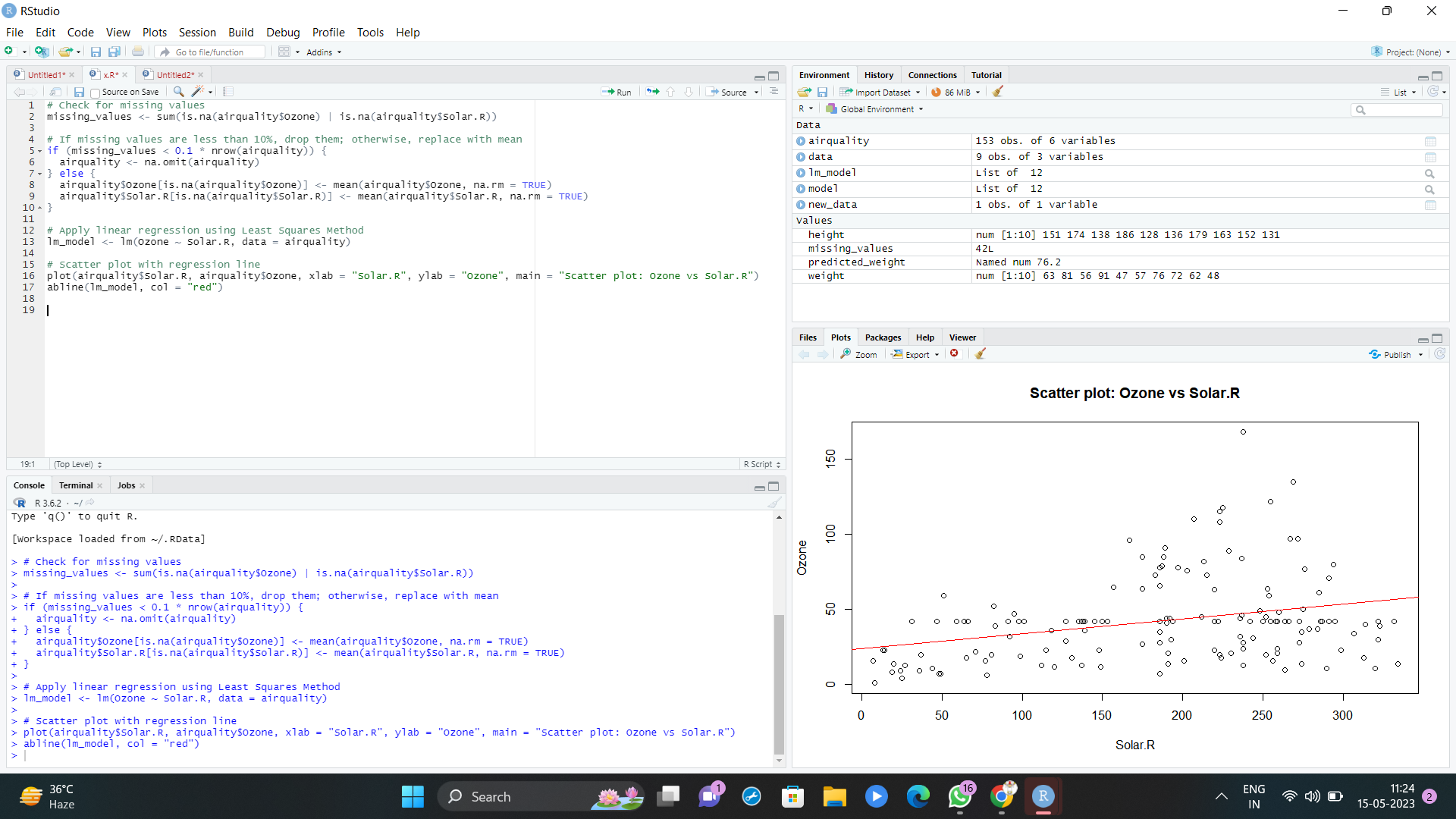
(iv)Cast the molten airquality data set with respect to month and date features

(v) Use cast function appropriately and compute the average of Ozone, Solar.R , Wind and temperature per month?



4.(i) Find any missing values(na) in features and drop the missing values if its less than 10% else replace that with mean of that feature.

(ii) Apply a linear regression algorithm using Least Squares Method on “Ozone” and “Solar.R” (iii)Plot Scatter plot between Ozone and Solar and add regression line created by above model

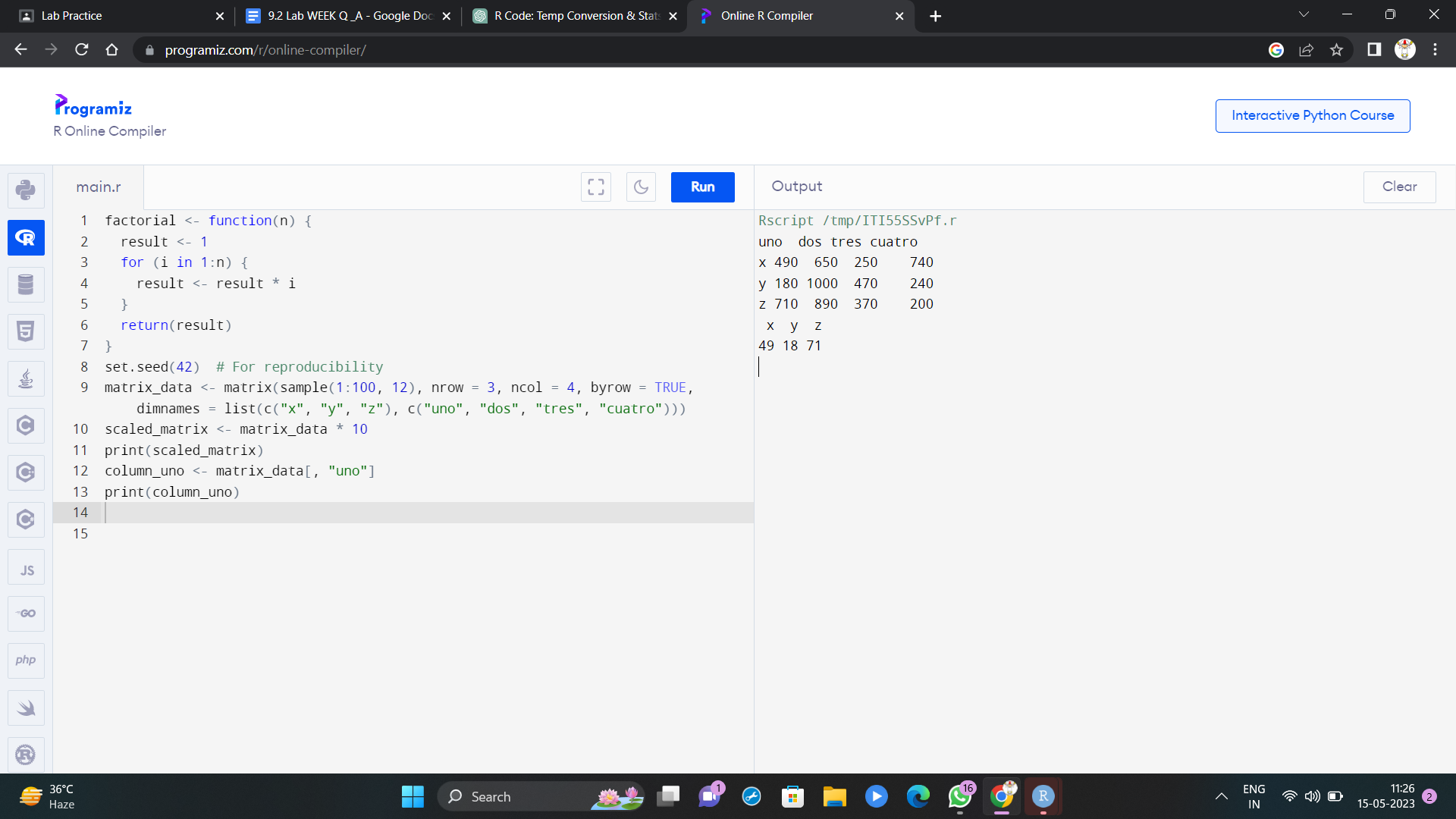


Set-II

1. (i)Write a function to find the factorial of a given number using “for” Loop

(ii) Create a 3x4 matrix with 12 random numbers between 1-100; have the matrix be filled our row by-row, instead of column-by-column. Name the columns of the matrix *uno*, *dos*, *tres*, *cuatro*, and the rows *x*, *y*, *z*. Scale the matrix by 10 and save the result.

(iii) Extract the column called “uno” as a vector from the original matrix and save the result



2. In 1936, Edgar Anderson collected data to quantify the geographic variations of iris flowers. The data set consists of 50 samples from each of the three sub-species ( *iris setosa*, *iris virginica,* and *iris versicolor*).Four features were measured in centimeters (cm): the lengths and the widths of both sepals and petals

(i)Find dimension, Structure, Summary statistics, Standard Deviation of all features. (ii)Find mean and standard deviation of features groped by three species of Iris flowers (Iris setosa, Iris virginica and Iris versicolor)

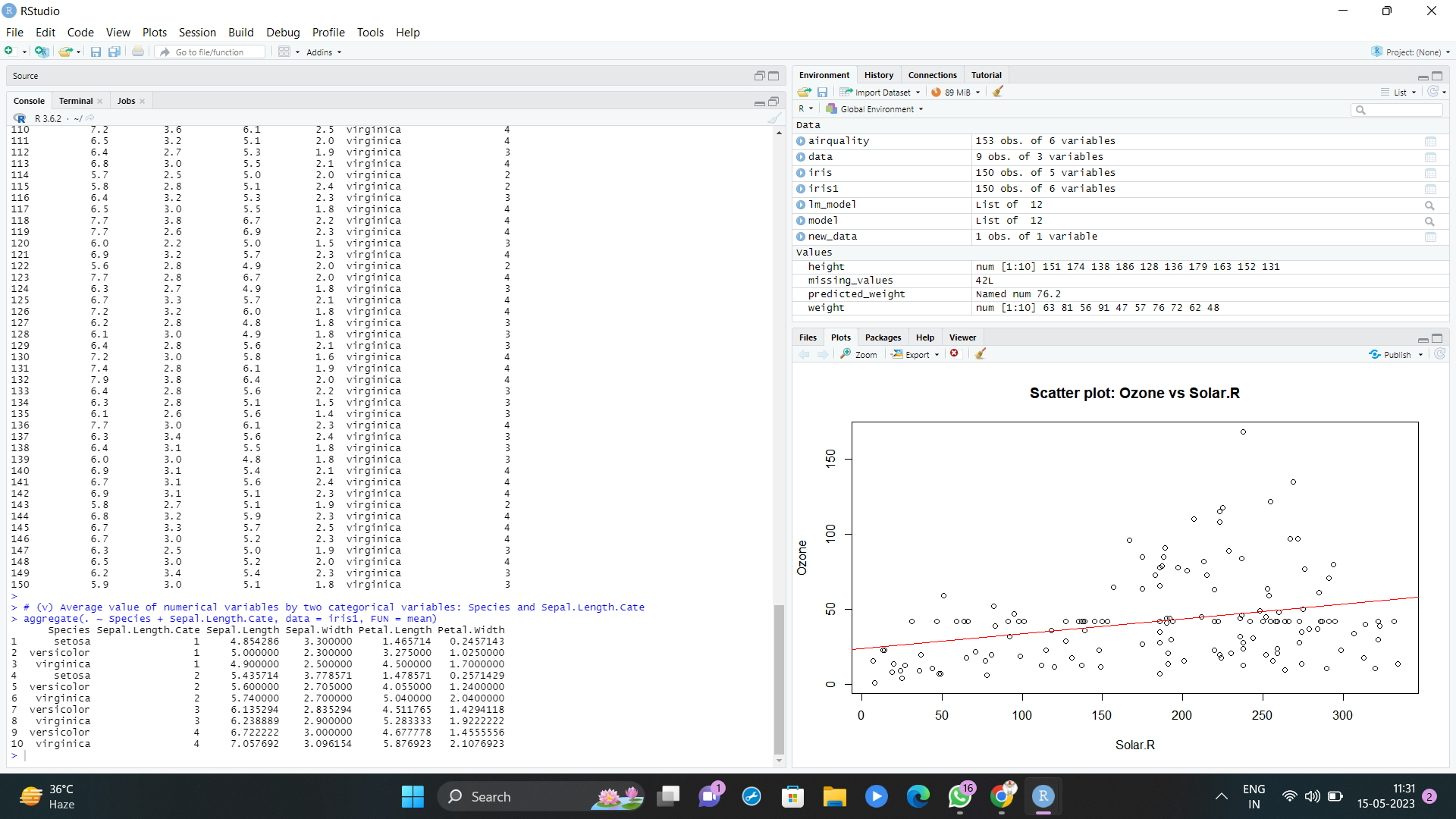
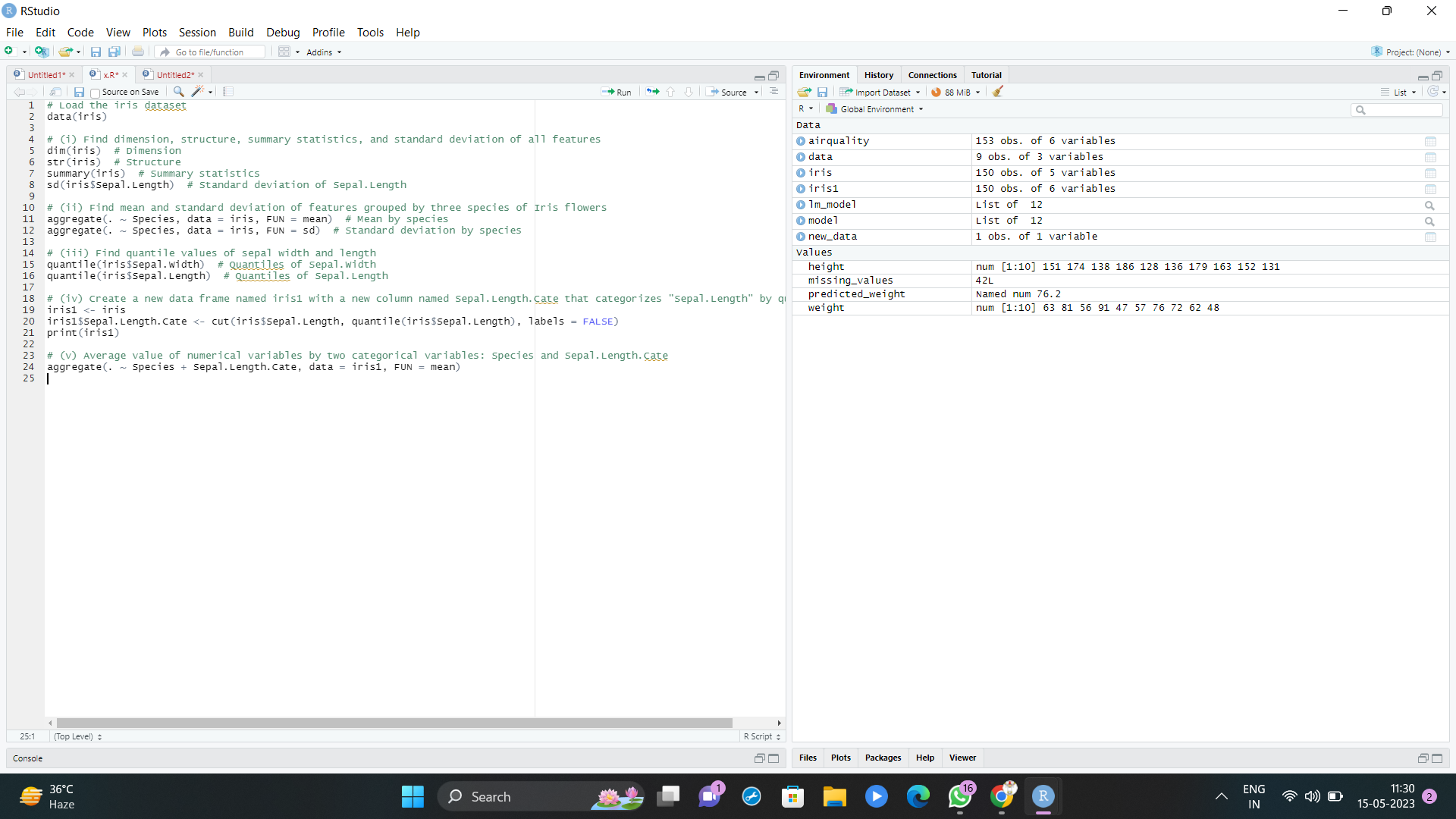
(iii)Find quantile value of sepal width and length

(iv)create new data frame named iris1 which have a new column name

**Sepal.Length.Cate** that categorizes “Sepal.Length” by quantile

(v) Average value of numerical varialbes by two categorical variables: Species and

Sepal.Length.Cate.



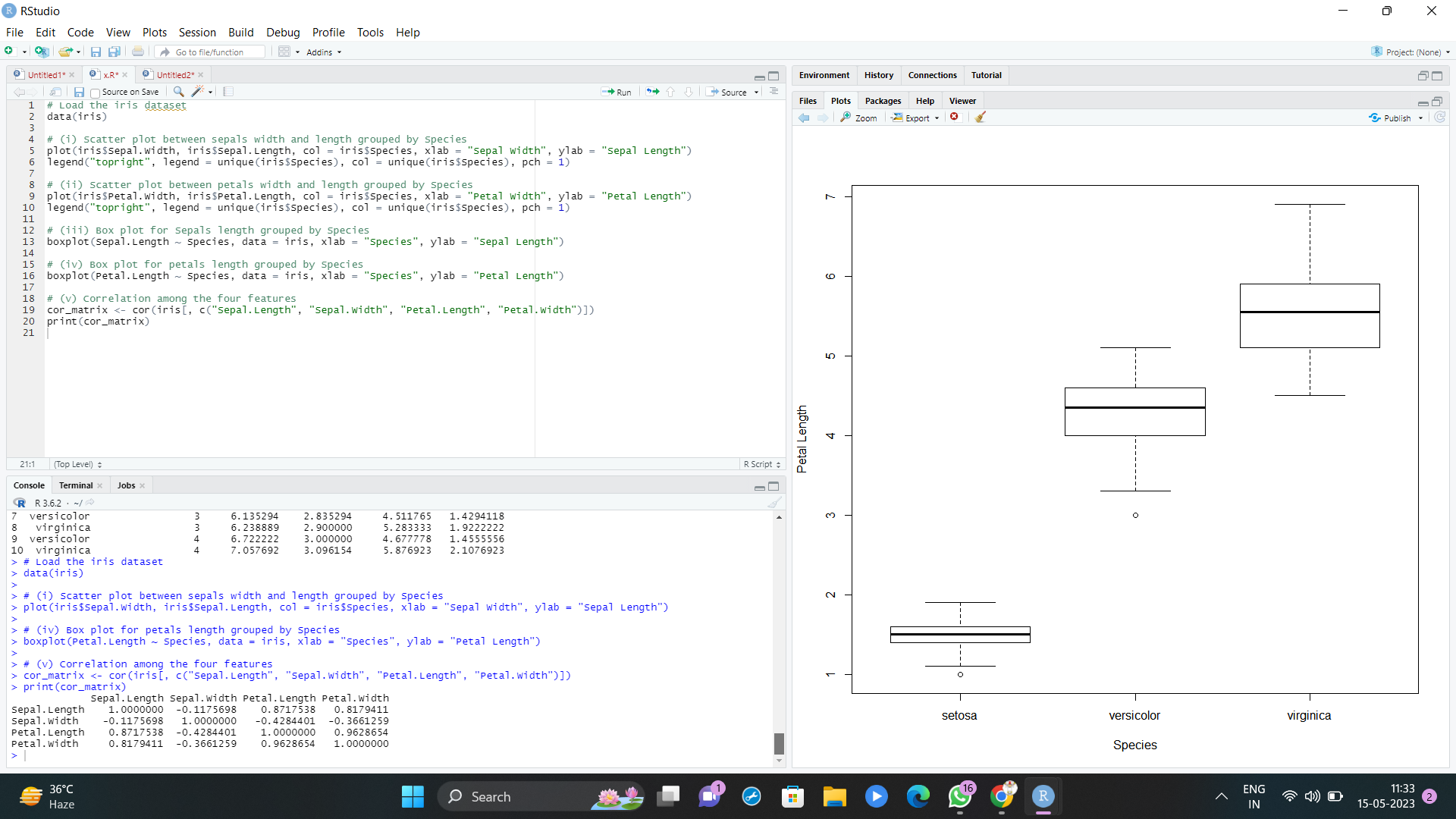
3. (i)Plot Scatter plot between sepals width and length grouped by Species

(ii) Plot Scatter plot between petals width and length grouped by Species

(iii)Draw the Box plot for Sepals length grouped by Species

(iv) Draw the Box plot for petals length grouped by Species

(v)Find the correlation among the four features

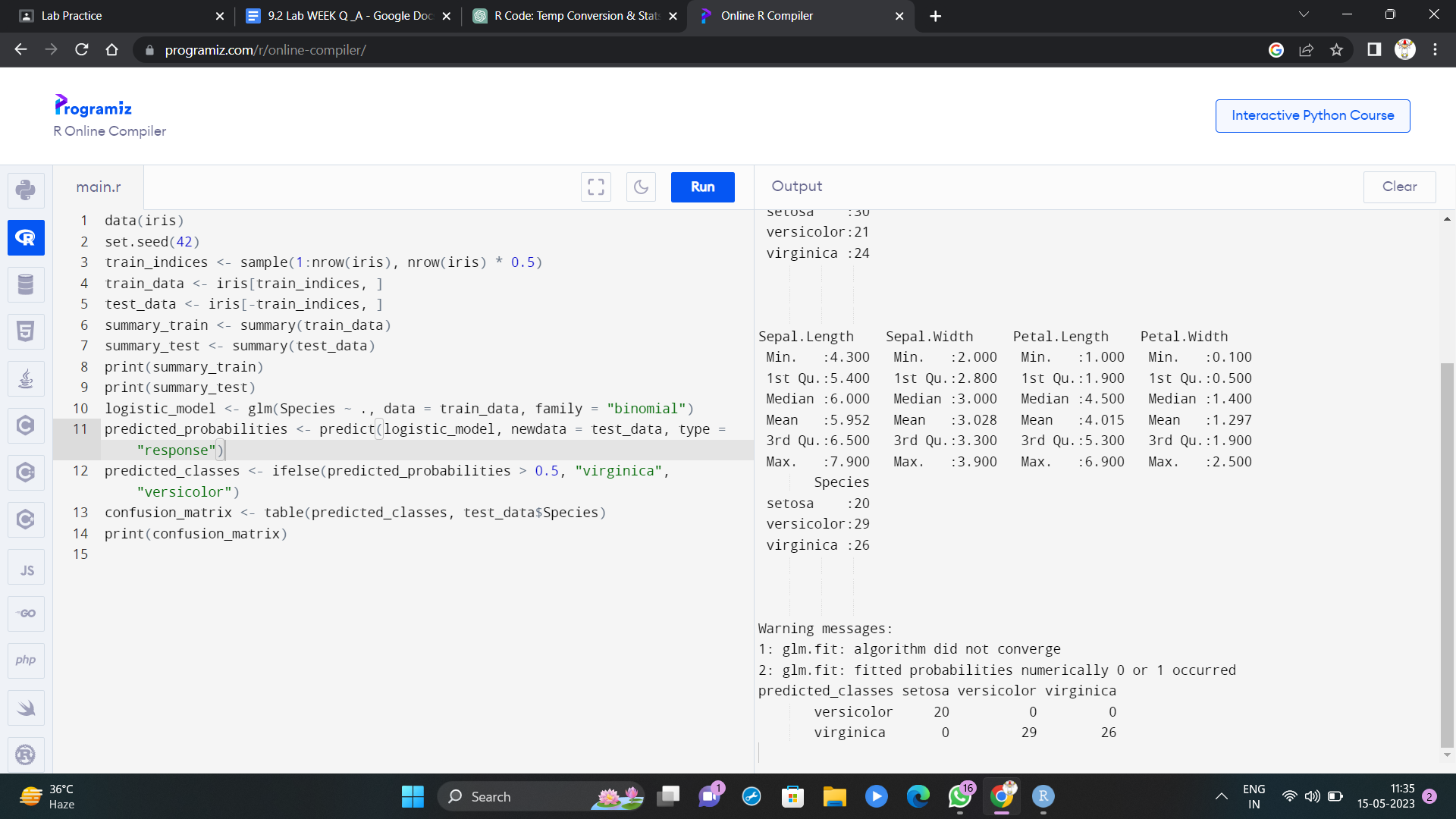
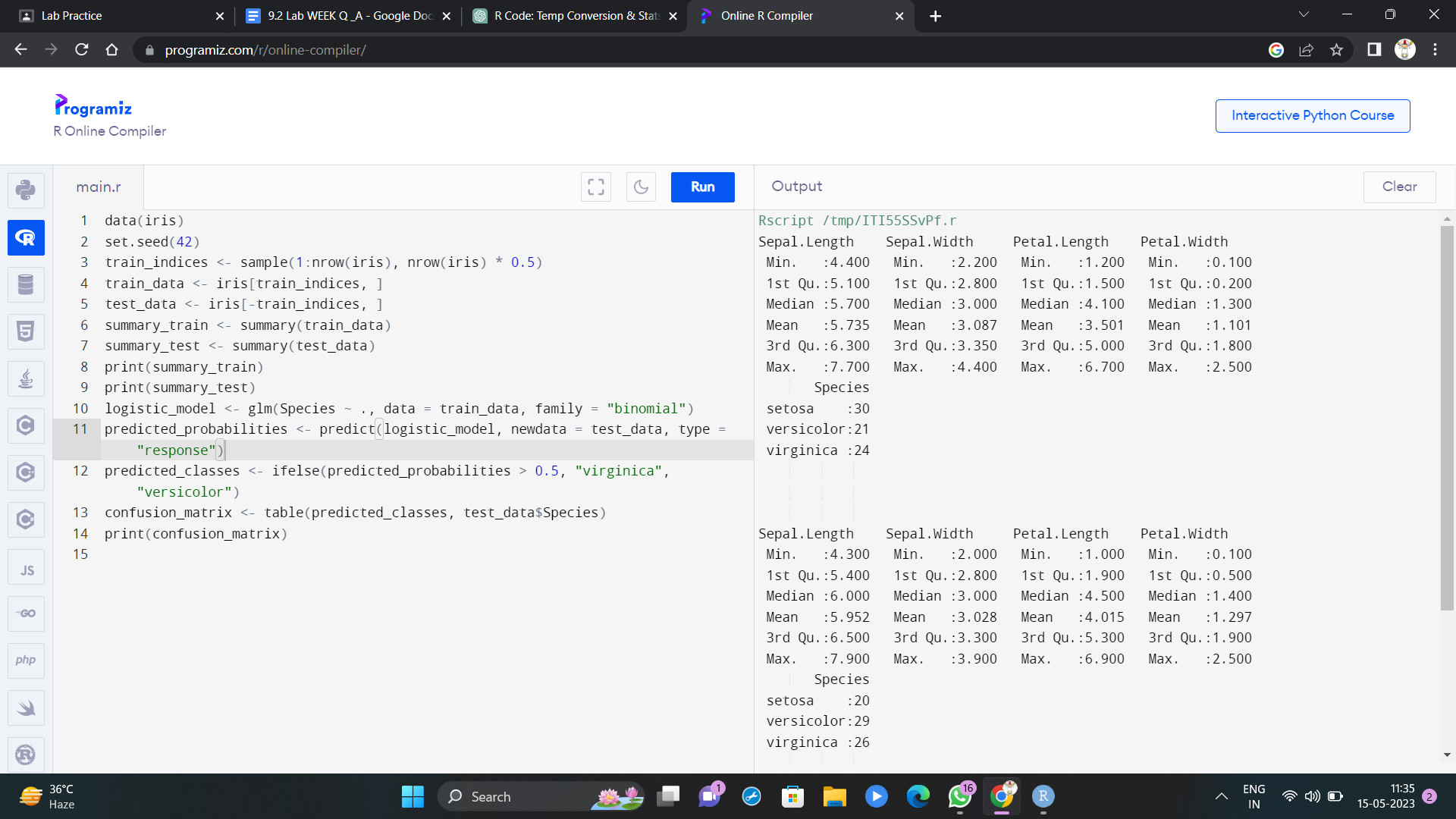


4.(i) Randomly Sample the iris dataset such as 50% data for training and 50% for test (ii)find summary statistics of above train and test dataset.

(iii)Create Logistics regression with train data

(iv)Predict the probability of the model using test data

(v)Create Confusion matrix for above test model



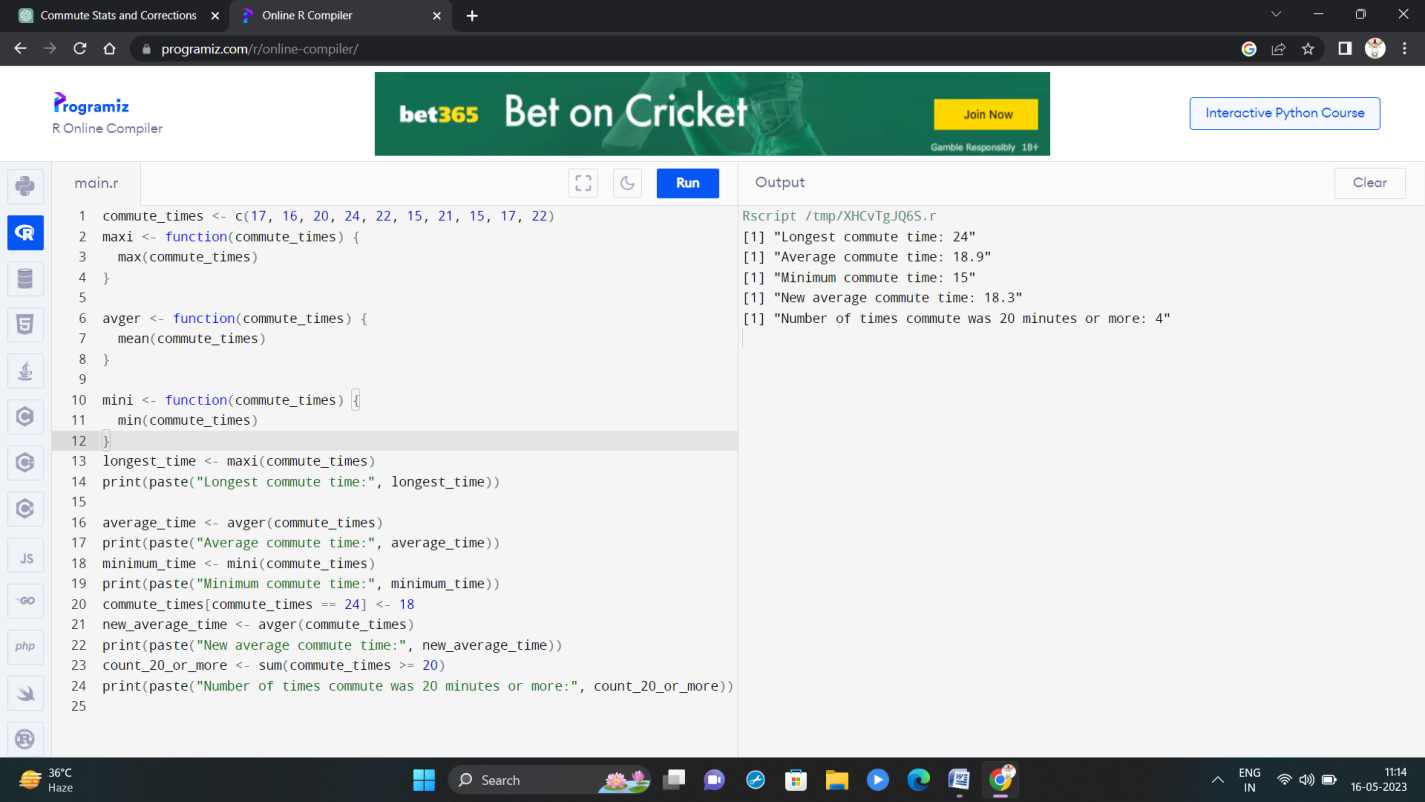
Set-III

1. Suppose you track your commute times for two weeks (10 days) and you find the following times in minutes 17 16 20 24 22 15 21 15 17 22 Enter this into R as vector data type.

(i)create function maxi to find the longest commute time, the function avger to find the average and the function mini to find the minimum.

(ii)Oops, the 24 was a mistake. It should have been 18. How can you fix this? Do so, and then find the new average using above functions.

(iii)How many times was your commute 20 minutes or more?



2. There is a popular built-in data set in R called "**mtcars**" (Motor Trend Car Road Tests), which is retrieved from the 1974 Motor Trend US Magazine.

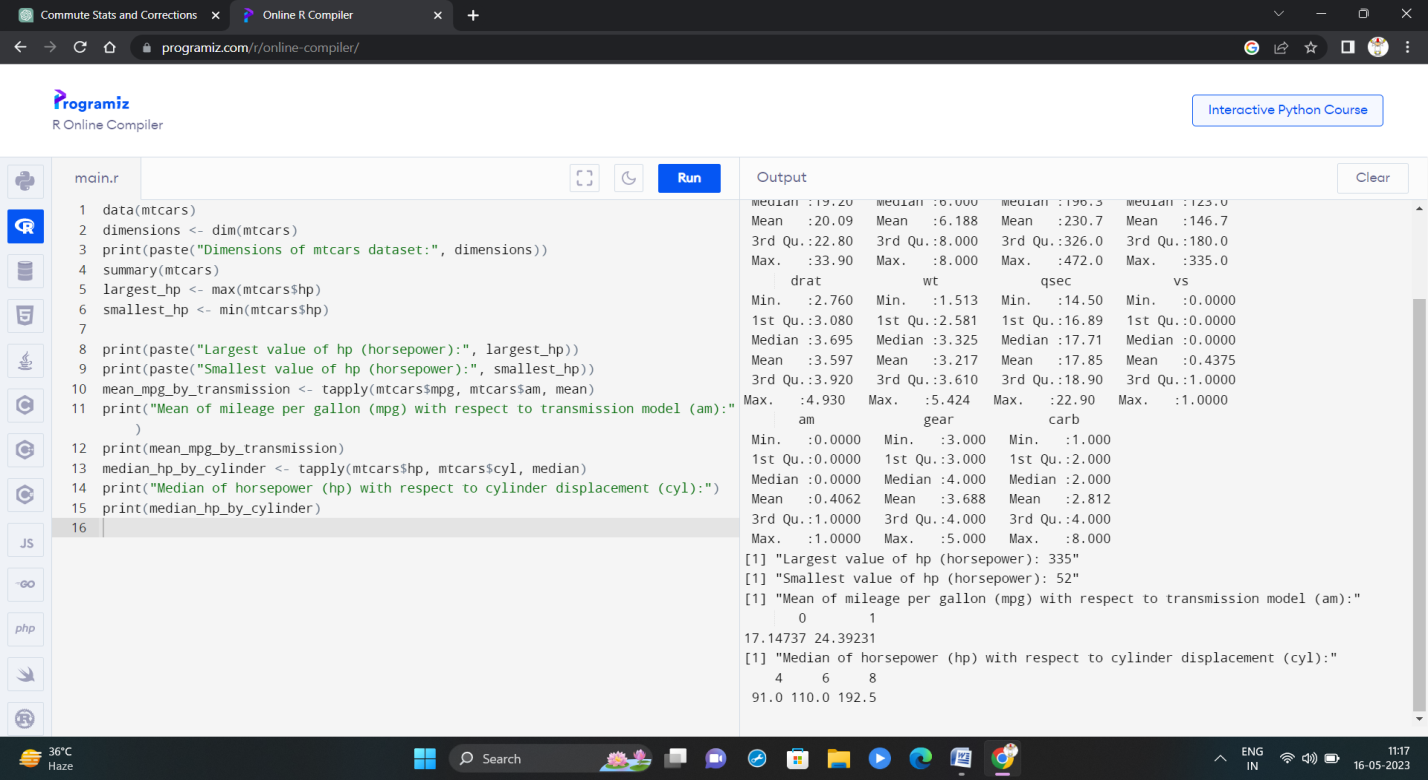
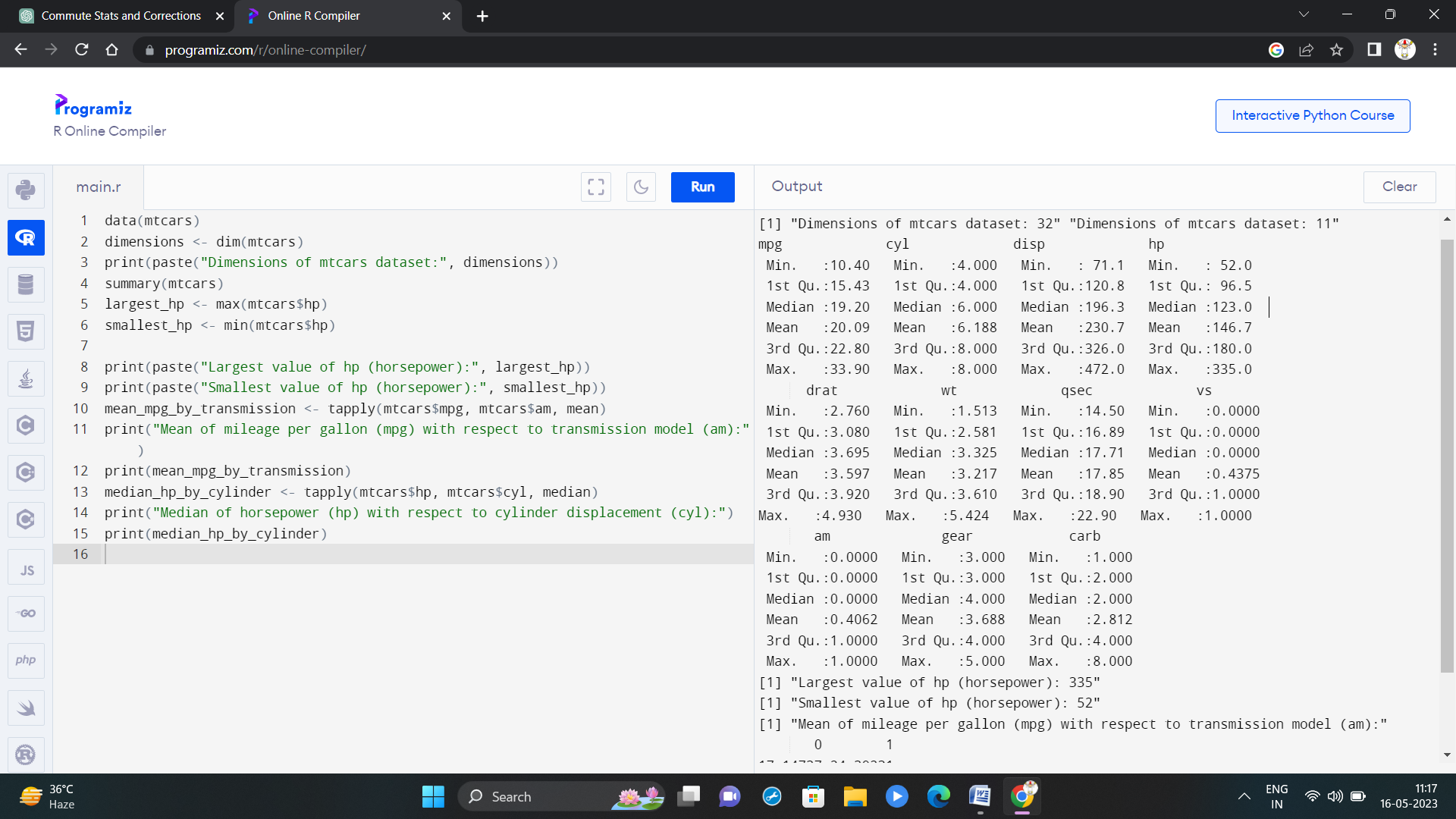
(i)Find the dimension of the data set

(ii)Give the statistical summary of the features.

(iii)Find the largest and smallest value of the variable hp (horsepower).

(iv)Give the mean of mileage per gallon (mpg) with respect to transmission model (feature named as ‘am’)

(v)Give the median of horsepower (hp) with respect to cylinder displacement(cyl)



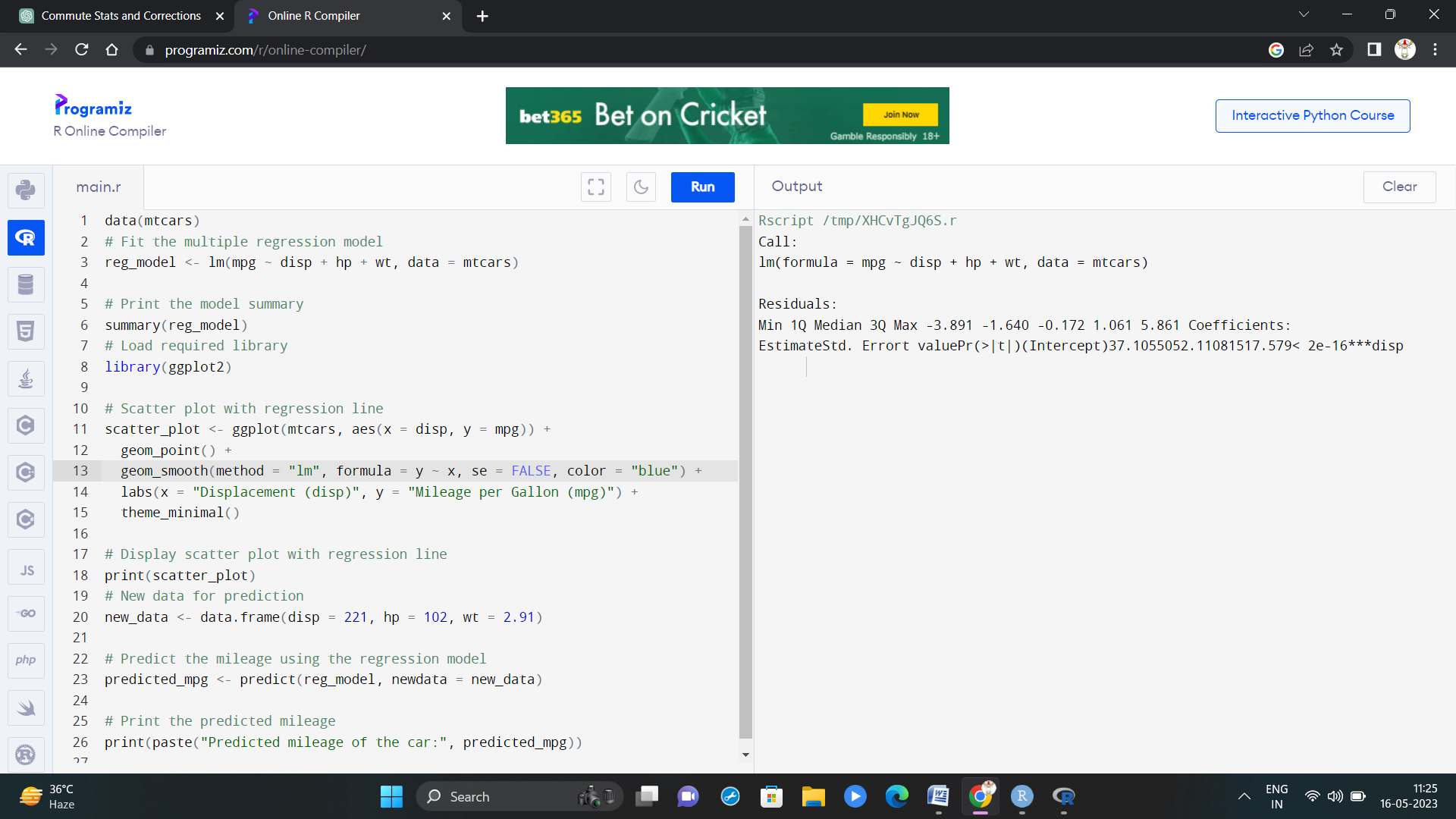
3.(i)Create Scatter plot mpg vs hp, grouped by transmission model (feature named as ‘am’) (ii)Create Box plot for mpg with respect to transmission model (feature named as ‘am’) (iii)Create histogram plot which shows statistical distribution of hp

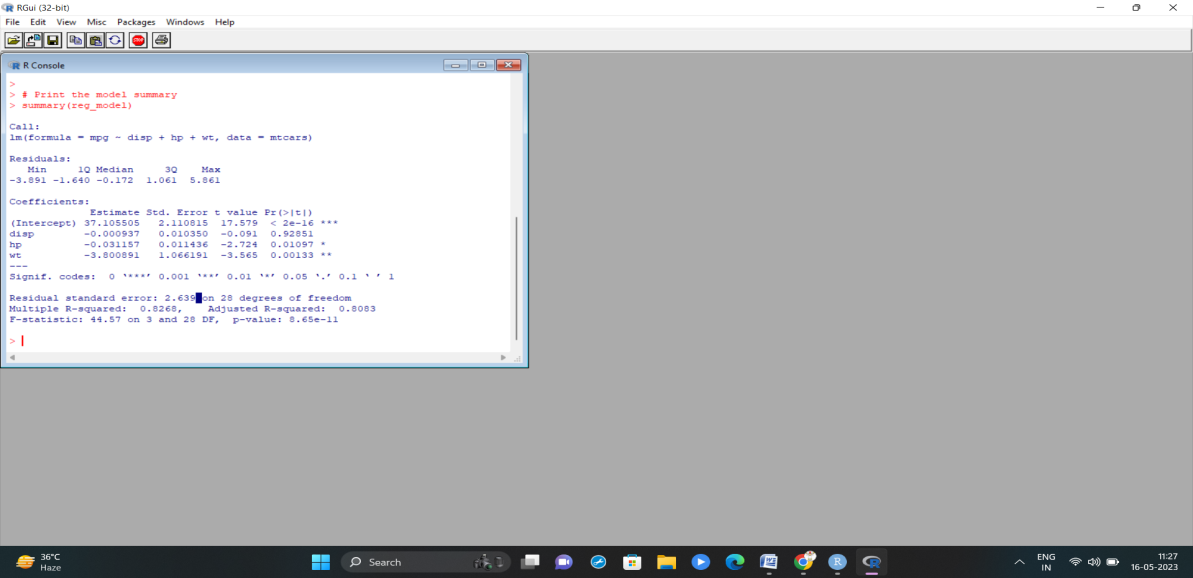
(iv)Draw the Bar Chart to show car distribution with respect to number of gears grouped by cylinder.(Grouped or multiple bar chart)

(v)Draw Pie chart which shows the percentage of distribution by number of gears.

4. (i)Generate a multiple regression model using the built-in dataset mtcars. Establish the relationship between "mpg" as a response variable with "disp","hp" and "wt" as predictor variables . (ii)Plot the multiple regression line model with above model parameters.

(iii) Predict the mileage of the car with dsp=221, hp=102 and wt=2.91





Set IV

1. (i) Write a function in R programming to print generate Fibonacci sequence using Recursion in R .

(ii) Find sum of natural numbers up-to 10, without formula using loop

statement.

(iii) create a vector 1:10 and Find a square of each number and store that in a separate list.

2. **mtcars**(motor trend car road test) comprises fuel consumption, performance and 10 aspects of automobile design for 32 automobiles. It comes pre-installed with **dplyr** package in R.

(i)Find the dimension of the data set

(ii)Give the statistical summary of the features.

(iii)Print the categorical features in Dataset

(iv)Find the average weight(wt) grouped by Engine shape(vs)

(v)Find the largest and smallest value of the variable weight with respect to Engine shape

3.Use ggplot package to plot below EDA questions label the plot accordingly (i)Create weight(wt) vs displacement(disp) scatter plot factor by Engine Shape(vs) (ii) Create horsepower (hp) vs mileage (mgp) scatter plot factor by Engine Shape(vs) (iv)In above(ii) plot , Separate columns according to cylinders(cyl) size

(v) Create histogram plot for horsepower (hp) with bin-width size of 5

4. Performing Logistic regression on dataset to predict the cars Engine shape(vs) . (i)Do the EDA analysis and find the features which is impact the Engine shape and use this for model.

(ii) Split the data set randomly with 80:20 ration to create train and test dataset and create logistic model

(iii)Create the Confusion matrix among prediction and test data.

Set-V

1.(i) Write a R program to extract the five of the levels of factor created from a random sample from the LETTERS (Part of the base R distribution.)

(ii)Write R function to find the range of given vector. Range=Max-Min

Sample input, C<-(9,8,7,6,5,4,3,2,1), output=8

(iii)Wirte the R function to find the number of vowels in given string

Sample input c<- “matrix”, output<-2

2.Load inbuild dataset “ChickWeight” in R

(i) Explore the summary of Data set, like number of Features and its type. Finds the number of records for each features

(ii)Extract last 6 records of dataset

(iii) order the data frame, in ascending order by feature name “weight” grouped by feature “diet” (iv)Perform melting function based on “Chick","Time","Diet" features as ID variables (v)Perform cast function to display the mean value of weight grouped by Diet

3.(i)Get the Statistical Summary of “ChickWeight” dataset

(ii)Create Box plot for “weight” grouped by “Diet”

(iii)Create a Histogram for “Weight” features belong to Diet- 1 category

(iv) Create a Histogram for “Weight” features belong to Diet- 4 category

(v) Create Scatter plot for weight vs Time grouped by Diet

4.(i) Create multi regression model to find a weight of the chicken , by “Time” and “Diet” as as predictor variables

(ii) Predict weight for Time=10 and Diet=1

(iii)Find the error(MAE) in model for same