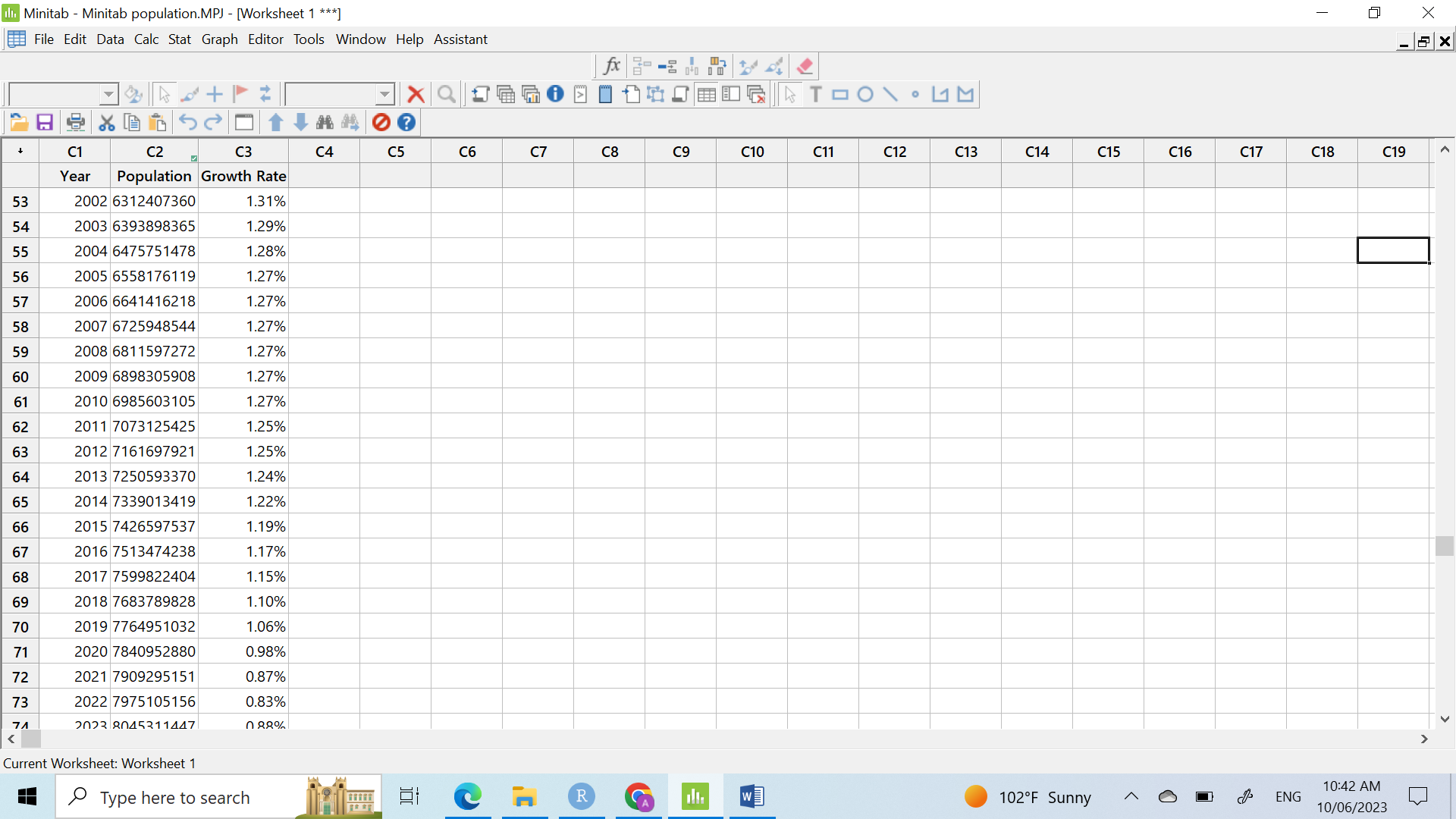
Introduction to Data Science Supervised learning (Working with data)

Exercise 1:

Q1)

Dataset :

The data set is about the year (as independent variable) and the population (as depend variable) we want to detemine if their any relation between these variables by using regression analysis .





This is a scatter diagram for this data. As can be seen from the scatter diagram, there does appear to be a linear relationship. It appears that as the age year increases, the number of population increase . We can use regression analysis to quantify that relationship. This will also allow us to predict the population based on the year.



As can be seen from the scatter diagram, there may be relationship between the year and growth rate . It appears that as the year increases, the growth rate decreases. But we can see some outliers appear which can affect the mean and the accuracy of the model.

Regression Analysis: Population versus Year

|  |  |  |  |
| --- | --- | --- | --- |
| S | R-sq | R-sq(adj) | R-sq(pred) |
| 99020507 | 99.67% | 99.66% | 99.64% |

R sq(adj) shows that about 99.66 % of the total variation in the response variable population growth is reduced or explained by this simple linear regression model having year as x variable. Only, 0.44% of the total variation in population growth was not explained

Coefficients

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Term | Coef | SE Coef | T-Value | P-Value | VIF |
| Constant | -1.52148E+11 | 1070585908 | -142.12 | 0.000 |  |
| Year | 79153155 | 538900 | 146.88 | 0.000 | 1.00 |
|  |  |  |  |  |  |

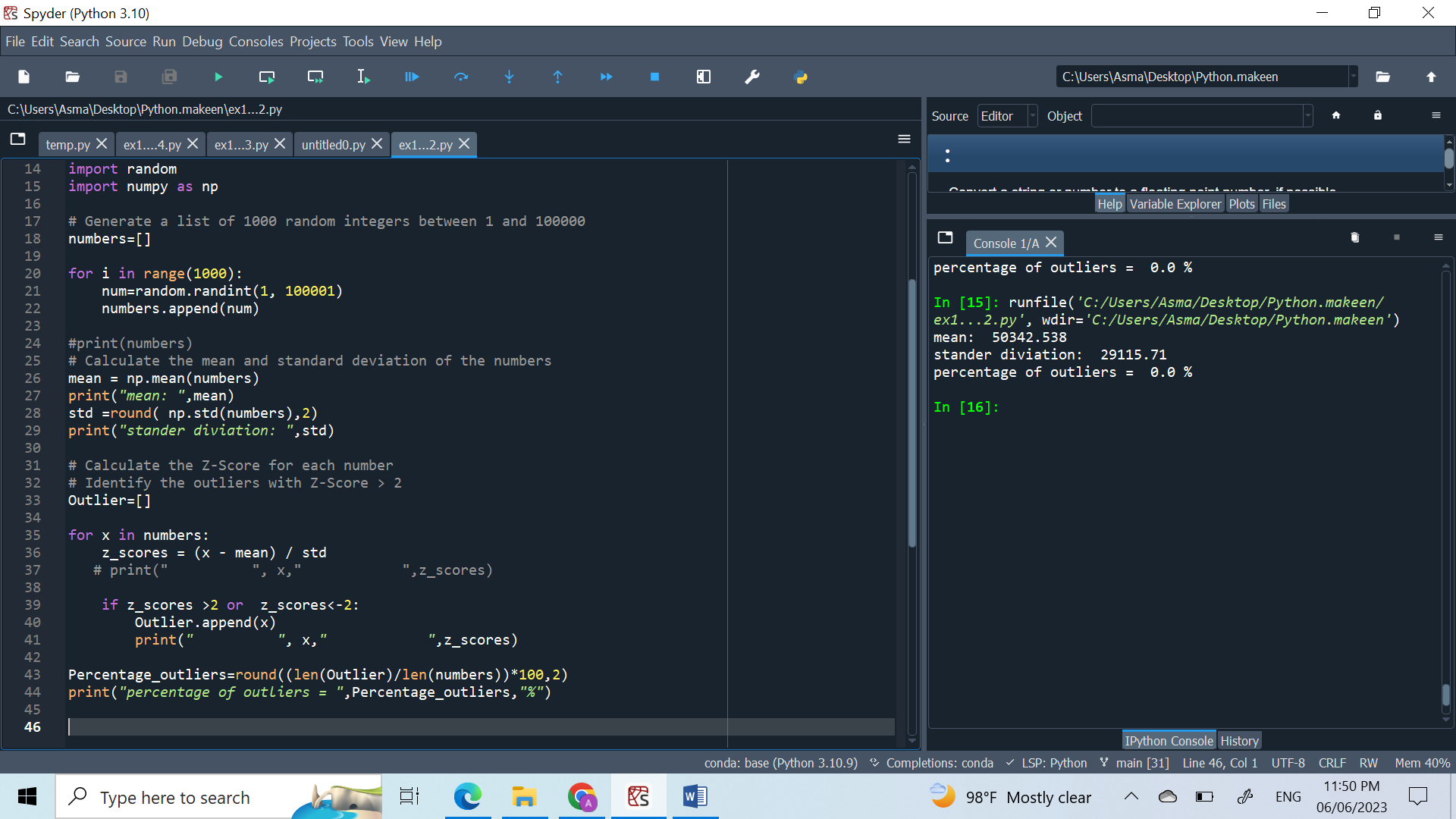
The table shown that year is significant while p value equal to 0 so less than 0.05

Regression Equation

|  |  |  |
| --- | --- | --- |
| Population | = | -152148121488 + 79153155 Year |

The simple regression analysis techniques have been used to forecast the population growth. simple linear regression is a highly flexible method for examining the relationship of an independent variable to a single dependent variable .

Q2)

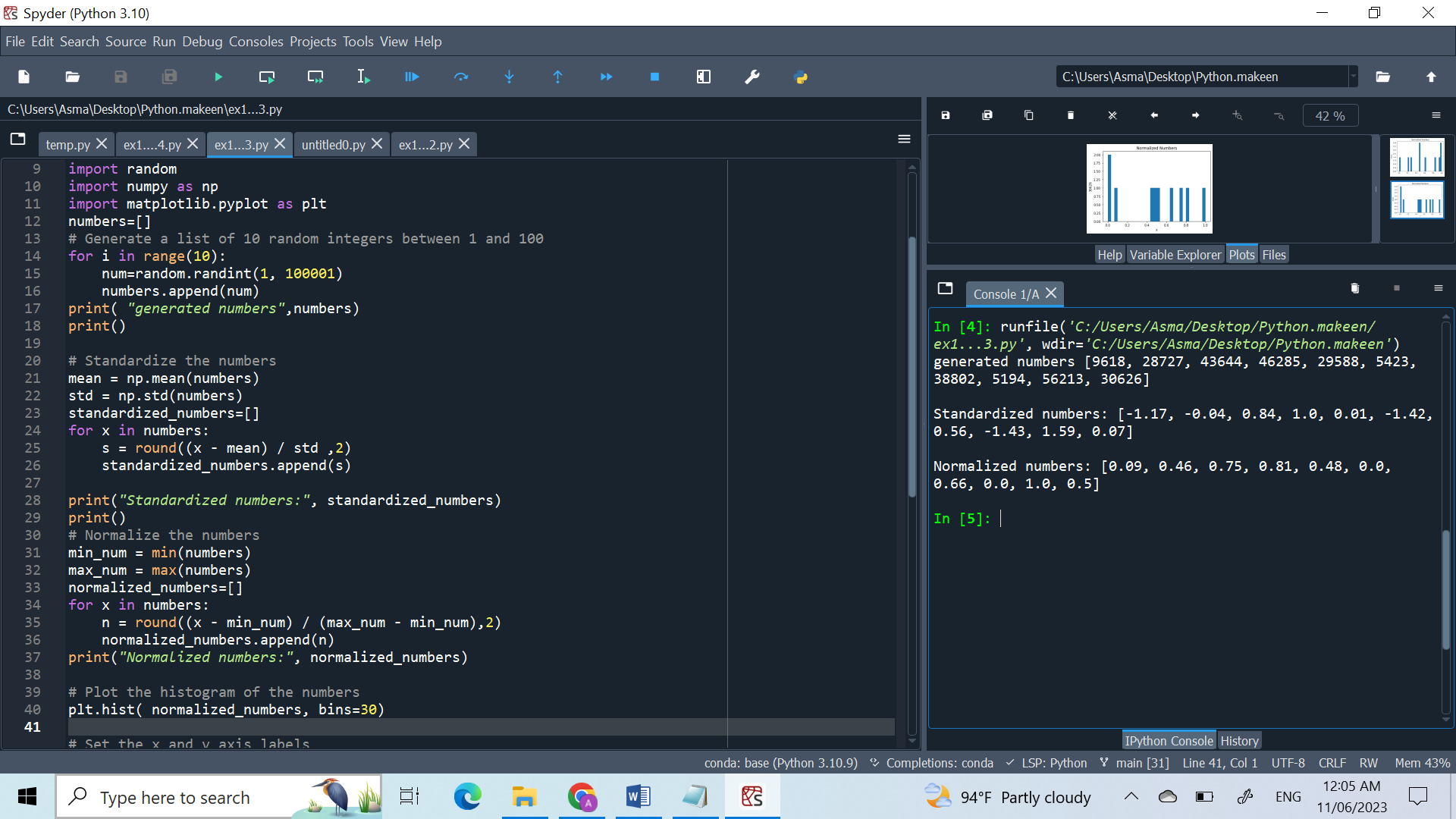


coment:

From the output that there are no outliers in the dataset when k equal to 2 . An outlier is a data point that is significantly different from other data points in the dataset, and its presence can affect the accuracy of statistical analyses. So If the percentage of outliers is 0, it indicates that the dataset is free of any significant deviations and can be analyzed without concern for outlier effects

But that percentage may be change if we change k

Q3)



Standardizing a dataset means transforming the data so that it has a mean of 0

and a standard deviation of 1. from the output we saw that standrized number is between -1.5

to 1.5 Standardizing a dataset makes it easier to compare different variables because they

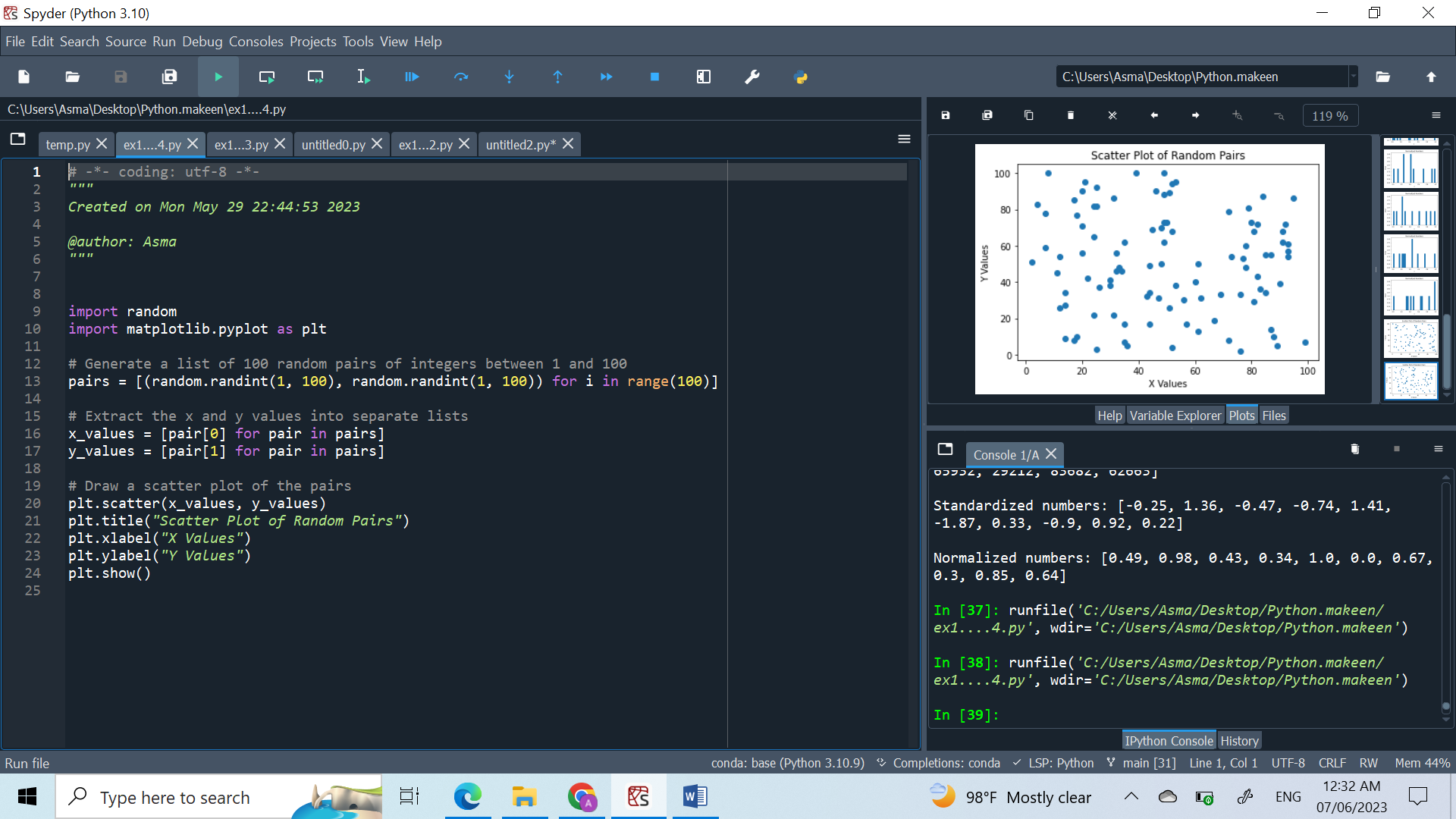
are all on the same scale.

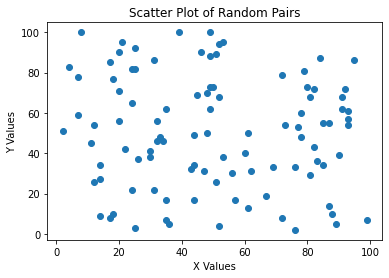
Normalized numbers are a type of standardized data that are scaled to have a range between

0 and 1. Normalization is the process of rescaling the data so that it falls within a specific

range, typically 0 and 1. and that what we recognize from the output

Q4)





This is a scatter diagram for this data. As can be seen from the scatter diagram,there does appear to be no relationship.