

Python for Data Science

Project

|  |  |  |  |
| --- | --- | --- | --- |
| **Faculty** | Information Technology | | |
| **Module Name** | Python for Data Science | **Module Code** | ITPPA0/ITPFA0 |
| **Project Number** | 1 | **Copy Editor** | Ms Nicole Stern |
| **Student Name** | Sopumelela | **Date Submitted** | 16/03/2024 |
| **Student Surname** | Sandekela | **Student Number** | CON-1474940-C5L6 |
| **Marker Name** |  | **Date Marked** |  |
| **Mark** | /170 | **Percentage** | % |
| **Moderator Name** |  | **Date Moderated** |  |
| **Mark** | /170 | **Percentage** | % |

|  |  |  |  |
| --- | --- | --- | --- |
| **Question Number** | | **Total** | **Mark Achieved** |
| Question 1 | | 30 |  |
| Question 2 | | 40 |  |
| Question 3 | | 30 |  |
| Question 4 | | 70 |  |
| **Lecturer** |  | **Total Mark** |  |

# ITPFA0 –**Project** – Python for Data Science 2024 | V3.0 Page 1 of 6

|  |  |  |
| --- | --- | --- |
| **Criteria** | **Marks** | **Achieved** |
| Student made use of coded Python methods to generate the  dataset and export to a file. | 5 |  |
| Student number should be generated using any of the following:  range(), for loop or while loop. | 2 |  |
| Student age should be an integer. It should be generated using a  suitable random() function. | 3 |  |
| Average hours spent studying on campus may be an integer and  should be generated using a suitable random() function. | 3 |  |
| Student mark should be generated using a suitable random()  function. (2 marks)  Furthermore, a list should be appended with the corresponding percentage, where the total mark is assumed to be 130. (3 marks) | 5 |  |
| Time taken on examination (minutes) should be generated using a  suitable random() function. | 3 |  |
| Generated datasets are exported in a CSV format. | 4 |  |
| Data generated within the specified ranges. | 3 |  |
| 150 rows present in each dataset. | 2 |  |
| **Other Deductions** | | |
|  |  |  |
|  |  |  |
|  |  |  |
| **Total** | **30** |  |

|  |  |  |
| --- | --- | --- |
| **Criteria** | **Marks** | **Achieved** |
| Headings of the Frequency Tables should be as follows:   * Average hours spent on campus * Student age * Student mark   (1 mark per Frequency Table) | 2 |  |
| Students made use of an empty dictionary with the suitable iterator and conditional statement to generate the Frequency Tables.  Note to Marker:   * Students are expected to use the given ranges as the “key” and number of occurrences as the “value”. * Each student mark is the “key” and the number of its   occurrences is the “value”. | 15 |  |
| Bar chart showing student ages and numbers of students correctly  plotted using Matplotlib. Ensure that the x-axis represents the student age groups. | 5 |  |
| Line graph showing correlation between higher marks and time  spent on campus correctly plotted using Matplotlib. | 5 |  |
| Scatter plot chart showing each student’s mark and the time taken  on the examination using Matplotlib. | 5 |  |
| Scatter plot chart showing the relationship between time spent on  campus and the age of the student using Matplotlib. | 5 |  |
| Code commented sufficiently. | 3 |  |
| **Other Deductions** | | |
|  |  |  |
|  |  |  |
|  |  |  |
| **Total** | **40** |  |

|  |  |  |
| --- | --- | --- |
| **Criteria** | **Marks** | **Achieved** |
| Report formatted neatly and professionally | 2.5 |  |
| Source code included | 2.5 |  |
| Frequency Table: Average hours included | 2 |  |
| Conclusion drawn | 1 |  |
| Frequency Table: Student age included | 2 |  |
| Conclusion drawn | 1 |  |
| Frequency Table: Student marks included | 2 |  |
| Conclusion drawn | 1 |  |
| Bar Chart ages/student numbers included | 3 |  |
| Conclusion drawn | 1 |  |
| Line Graph marks/time on campus included | 3 |  |
| Conclusion drawn | 1 |  |
| Scatter Plot Chart mark/time taken included | 3 |  |
| Conclusion drawn | 1 |  |
| Scatter Plot Chart time on campus/student age included | 3 |  |
| Conclusion drawn | 1 |  |
| **Other Deductions** | | |
|  |  |  |
|  |  |  |
|  |  |  |
| **Total** | **30** |  |

|  |  |  |
| --- | --- | --- |
| **Criteria** | **Marks** | **Achieved** |
| Percent-bachelors-degrees-women-usa.csv file successfully read  into a list of lists. | 2 |  |
| Result assigned to variable BDW*.* | 2 |  |
| First five (5) rows of BDW successfully displayed on different lines. | 2 |  |
| Header row removed; remainder of dataset assigned as BDW1*.* | 2 |  |
| 1st, 2nd and 3rd row of BDW1 shown using the slicing technique. | 2 |  |
| Dictionary indexcount\_year created using a for loop. | 2.5 |  |
| Dictionary indexcount\_year has the correct keys and values as per  the specification. | 2.5 |  |
| Dictionary indexpercent\_year created using a for loop. | 2.5 |  |
| Dictionary indexpercent\_year has the correct keys and values as  per the specification. | 2.5 |  |
| Maths\_Stats variable correctly obtained using a for loop and  assigned as per the specification. | 5 |  |
| Physic\_Sci variable correctly obtained using a for loop and  assigned as per the specification. | 5 |  |
| Engine variable correctly obtained using a for loop and assigned as  per the specification. | 5 |  |
| Comp\_Sci variable correctly obtained using a for loop and  assigned as per the specification. | 5 |  |
| Year variable correctly obtained and assigned as per the  specification. | 5 |  |
| Numpy array Selected4Majors correctly created as per the  specification (each element in the list must be converted to a float before creating the Numpy array). | 5 |  |
| Dictionary “Majors” correctly created as per the specification. | 2 |  |
| Python function correctly created to plot the graph  Year/Selected4Majors. This should be done on one graph. | 5 |  |
| Legend correctly displayed in upper left corner. | 2 |  |
| Title correctly set as per the specification. | 1 |  |

|  |  |  |
| --- | --- | --- |
| x-axis correctly labelled as per the specification. | 1 |  |
| y-axis correctly labelled as per the specification. | 1 |  |
| Conclusions drawn from the graph make logical sense and are  relevant. | 8 |  |
| **Other Deductions** | | |
|  |  |  |
|  |  |  |
|  |  |  |
| **Total** | **70** |  |

**Feedback:**

**Table of Contents**

**QUESTION 38**

**Frequency Table8**

Average hours8

Student age8

Student marks8

Table8

Source code9-10

**Bar Graph11**

Conclusion11

Graph12

Source code13

**Line Graph14**

Conclusion14

Graph15

Source code16

**Scatter Chart-** **Student Marks vs. Time Taken 17**

Conclusion17

Chart17

Source code18

**Scatter Chart-** **Time Spent vs. Student Ages**19

Conclusion19

Graph20

Source code21

**QUESTION 4.6**22

Line Graph Analysis22

Graph23

Source code24

**Question 3**

## Frequency Table

In Data Science, frequency tables play a crucial role in summarizing and organizing data.

In this case we will focus on the following criteria then draw conclusions from the data reflected.

Average hours:

Hours are grouped in ranges 1-2, 2-3 and 4-5 hours now for us to get the average we needed to make an average of all the ranges mixed in one column, now the frequent average hour reflected seem to be the range 1-2 hours which might mean most people spend 1-2 hours in campus.

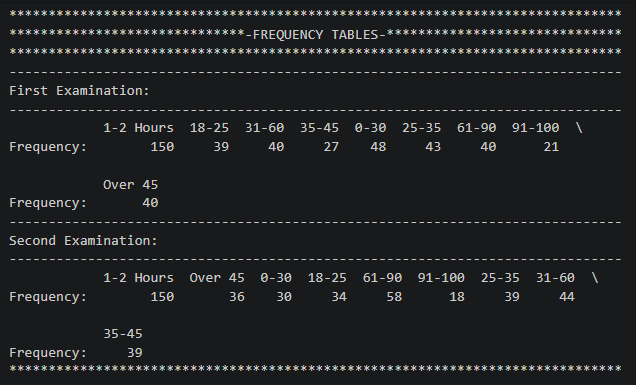
Student ages:

Student ages are grouped in ranges 18-25, 25-35, 35-45 and over 45. We noticed that most frequent age group on first table are the age group 25-35 while on the second table the frequent age groups are 25-35 and 35-45 age groups. From this insight we can conclude that most populations of student in campus are between age 25 and 35.

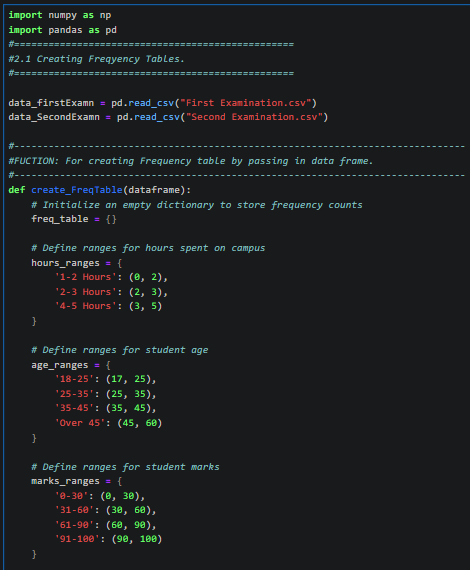
Student marks:

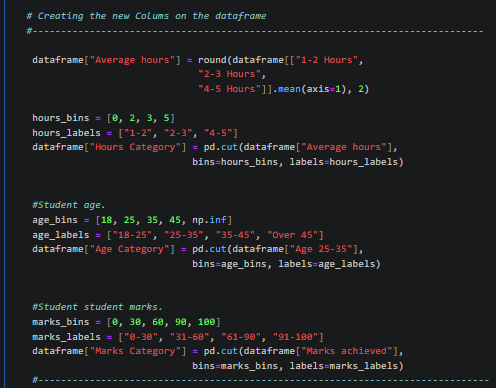
Student marks are grouped in 0-30%, 31-60%, 61-90% and 91-100%. For the first table the most frequently obtained marks are between 0-30 this is not a good insight as it lowers the pass rate of the campus. On the second table the most obtained marks are between 61-90% this is a better insight than the previous table as it shows improvement. We can conclude that the contribution of the age group 34-45 did help to raise the average.

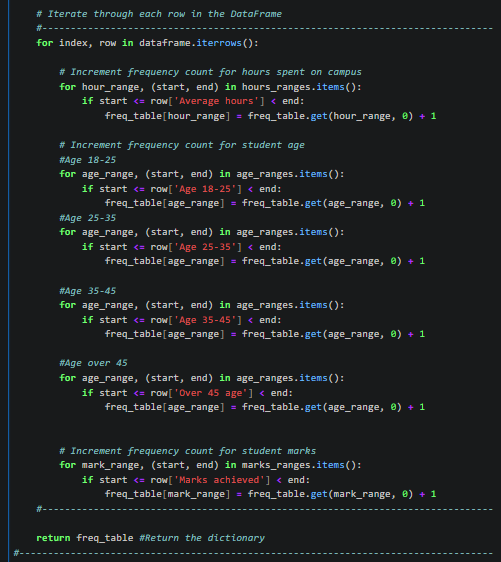
Table:

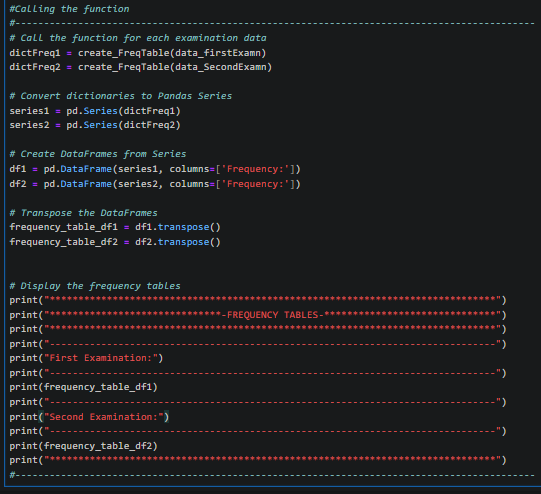


Source code:









## Bar Graph

Age / Student numbers analysis

This line graph has Number of Students on y-axis and Age Groups on x-axis.

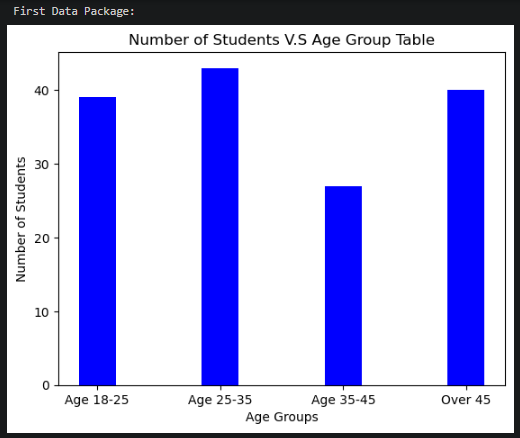
First Table:

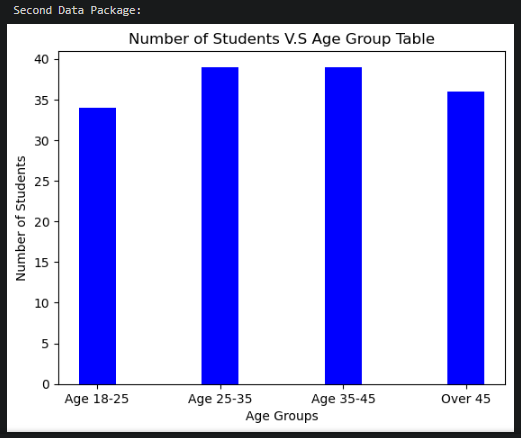
The first bar graph shows that the most student numbers are from the age group 25-35 and the least comes from the age group 35-45. This makes sense that 35-45 age was the least frequent age group when we make comparisons with the frequency table. Another insight I notice it’s that the greatest number of students are age group 25-35.

Second Table:

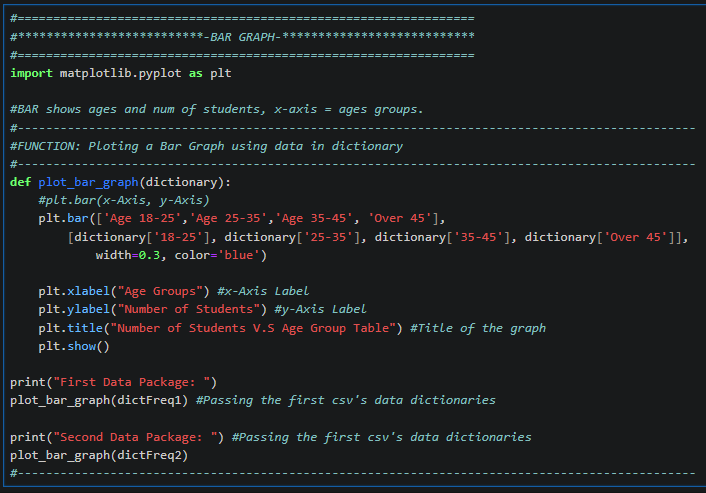
The second bar graph shows a great and equal number of students on the age groups 25-35 & 35-45. This insight we can conclude that the increase in student marks to 61-90 average being the highest student mark average the increase in number of students on 35-45 age range had an influence.

Graph:





Source code:



## Line Graph

Marks/Time on campus analysis

This line graph has Higher marks on y-axis and Time spent on Campus on x-axis.

First graph:

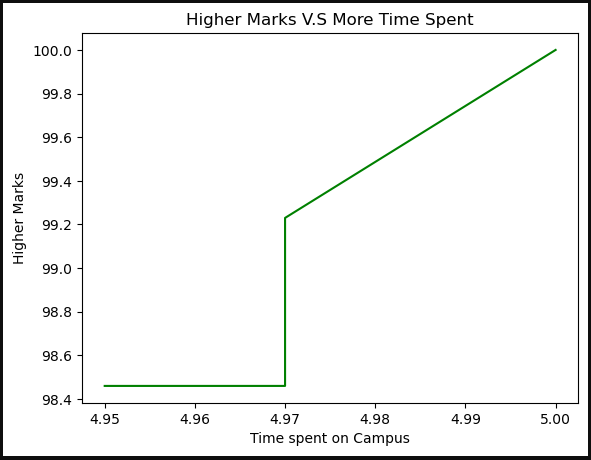
On the first line graph we can see that from 4.95 hours to 4.97 hours there is no rate of change just straight 98.43% Marks. Suddenly we notice steep rate of change increase up to 99.20% marks, then from there the is a good even exponential rate from 4.97 hours to 5 hours. Now we can conclude that students who spend most time on campus obtain better marks.

Second graph:

On the second line graph we can observe that from 4.97 hours to 4.98 hours there is no change in rate.

Then suddenly between 4.98 hours to 5 hours there is an even exponential rate of change from 97.7% to 98.3% marks. Finally, from 98.30% to 99.20% marks there is a steep rate of change.

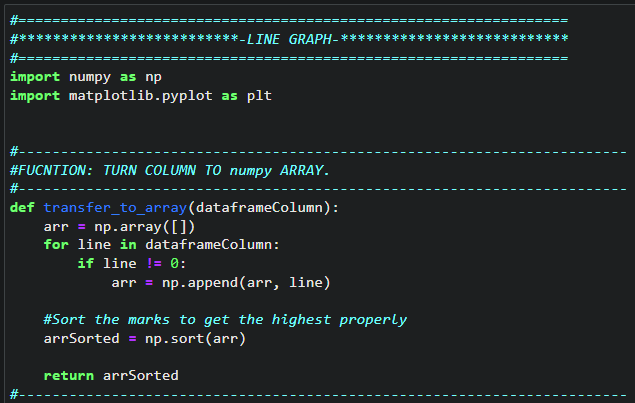
Graph:



A graph with a green line

Description automatically generated

Source code:



A screen shot of a computer program

Description automatically generated

## Scatter Chart- Student Marks vs. Time Taken

Mark/Time taken.

This scatter chart has Student's Mark on y-axis and Time taken on x-axis.

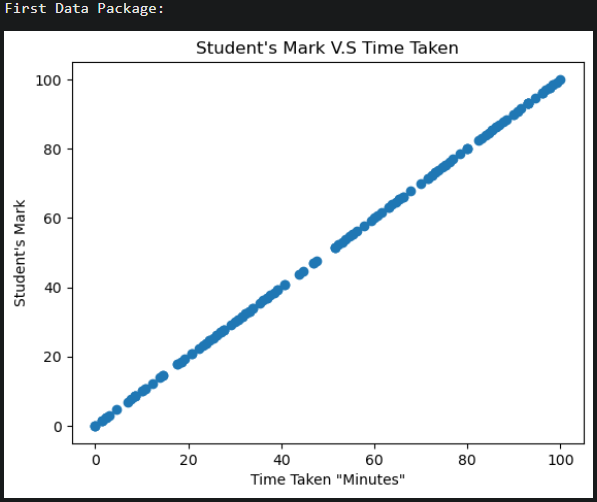
First scatter chart:

On the first scatter chart we can observe a positive correlation on a high strength as they are tightly clustered

Second scatter chart:

On the second scatter chart we see the same clustered positive correlation

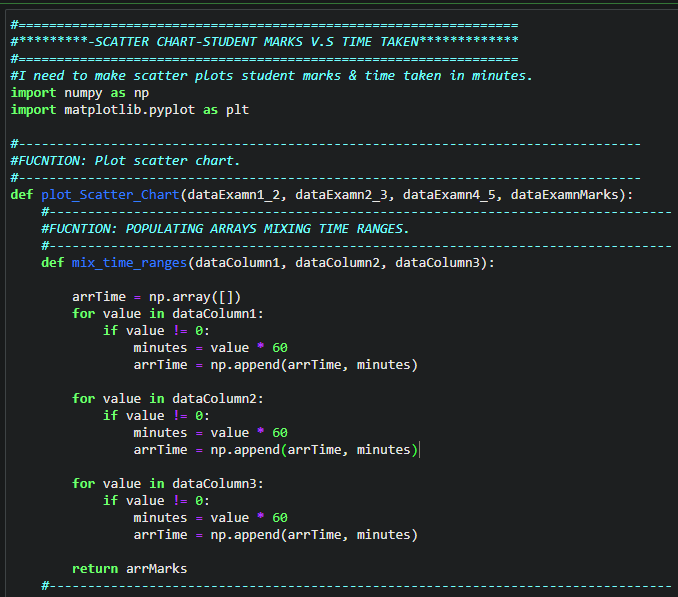
Chart:



A graph with blue dots

Description automatically generated

Source code:



A computer screen with white text

Description automatically generated

A computer screen shot of a computer code

Description automatically generated

## Scatter Chart- Time Spent vs. Student Ages

Time on Campus/Student age

This scatter chart has Time Spent on Campus on y-axis and Student Ages on x-axis.

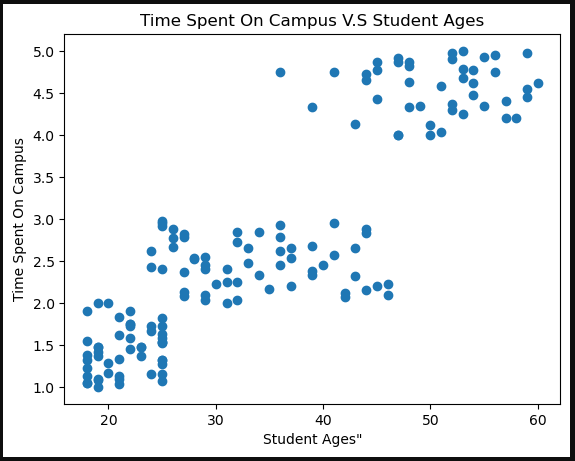
First scatter plot:

This chart shows a good strength of clustered data points between 1.0 to 2.0 Time spent on y-axis and 18 to 25 ages. From this we can conclude that the youth 18-25 age students spend less time on campus. From 3-4 hours we can see that least students spend that time on campus. Then we can observe that from age group 35 to 60 there is a medium strength of clustered data points.

Second scatter plot:

This chart shows a good strength of clustered data points between 1.0 to 2.0 Time spent on y-axis and 18 to 25 ages as the first. Same as the first chart from age group 35 to 60 there is a gathered strength but now a stronger and more clustered data points.

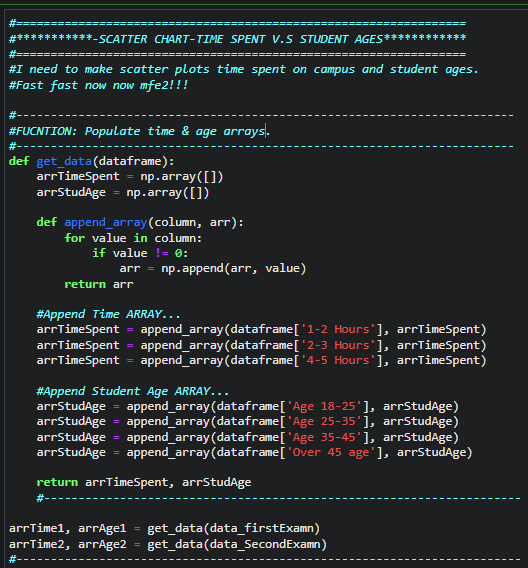
Chart:



A graph with blue dots

Description automatically generated

Source code:



A screen shot of a computer program

Description automatically generated

**QUESTION 4.6**

Line Graph analysis

Line graphs are a powerful tool in data science for visualizing trends and relationships within numerical data. In this case we are examining” Percentage of Selected4 Degree Awarded per year”, where we have “Year” as our x-axis and “Selected4Degrees” as our y-axis. These “Selected4Degrees” are:

Math and Statistics, Physical Science, Engineering, Computer Science

Math and Statistics:

Math and Statistics line is denoted as the blue line. Observing the trend of this line throughout from start to end we can say it had a slight increase compared to the starting point. The line seemed promising from the year: 1975 – 1985 as it showed upward trend but started a downward trend from year: 2000 – 2010. This shows not many women improved from Math and Statistics.

Physical Science:

Physical Science line is denoted as the red line. Looking at this line it can be easily seen as the line with the most upward trend showing that Physical Science is the most major with improved women from 1975 – 1910.

Physical Science is the major with the most upward trend by far.

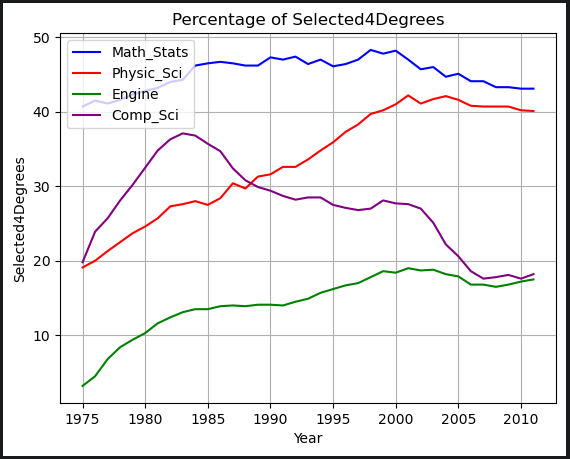
Engineering:

Engineering line is denoted as the green line. Observing the trend of this line we can say it is the second best in upward trend looking at it throughout from starting point to the ending point. What affected its upward trend it’s the resistance it got from year: 1985 – 1990. Now we can conclude that Engineering is the second best major that women improved on.

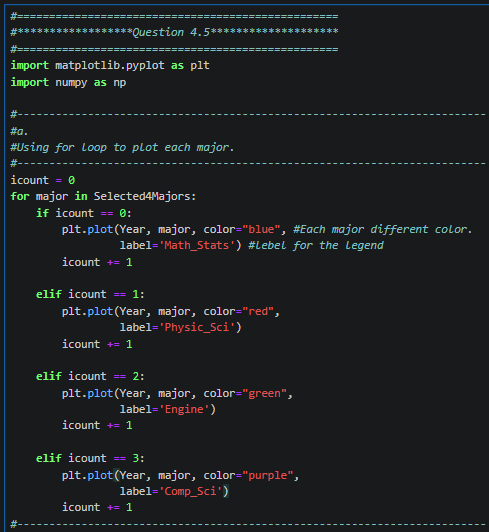
Computer Science:

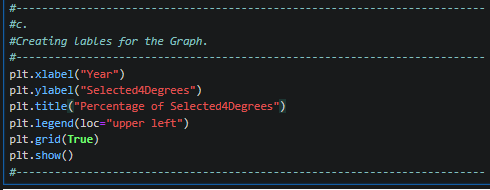
Computer Science line is denoted as the purple line. Computer Science is the worst performing major looking at its starting point to its end point. Starting with the most pretty good rate in change of upward trend from year: 1975 - 1983 but ending with a bad downward trend from year: 1983 – 2006 even though there was resistance from year: 1990 – 1997 but it did not help at the end. Now we can conclude that Computer Science is the wors performing major from year: 1975 – 2010.

Graph



Source Code:

****

****