Project Specification

Faculty	Information Technology	**Module Name**	Python for Data Science
Module Code	ITPPA0/ITPFA0	**Project Number**	1
Total Marks	170	**Copy Editor**	Ms Nicole Stern

This project contributes 20% towards the final mark.

Instructions to Student

1. All work, including draft notes, must be submitted with the completed project.
2. The program must be operational with as few faults as possible.
3. 10% will be deducted from this project if it is returned for resubmission due to plagiarism.
4. This project will require the student to demonstrate core skills required in Python for Data Science.

Resource Requirements

- The module's learning manual and the prescribed textbook may be referenced.

Delivery Requirements (evidence to be presented by students)

The project submission must include:

- Submission consists of neatly formatted documentation, including your source code and a neatly and professionally formatted report, as well as copies of your original datasets in .csv format and your cleaned datasets.
- Your name, student number, project number and date of submission must be included on the document's cover page

Plagiarism and Referencing

Consult the section at the end of this document, which outlines how negative marking will be applied as well as the way in which it will affect the assignment mark.

Section A

Question 1 30 Marks

Study the scenario and complete the question(s) that follow:

> You have been contacted by a large tertiary institution that has just completed their first round of examinations. There are 150 students who wrote two examinations each. You will need to capture the data from these examinations into a .csv file to use and provide the institution with a neatly formatted report, providing key insights into the data groups.
>
> Source: Munnik, P.C. (2019)

Create Model Datasets

You need to create two datasets, one for each examination, with 150 rows each. These datasets will require the headings below. It is suggested that you generate your datasets thoughtfully, in order to have meaningful patterns.

- Student number
- Student age
  - 18 – 25, 25 – 35, 35 – 45, over 45 (as age groups)
  - Average hours spent studying on campus
  - 1 – 2, 2 – 3, 4 – 5 hours (as categories)
  - Student mark achieved
  - The maximum mark is 130; you need to calculate the percentage
- Time taken on examination (minutes)
  - The maximum time is three hours (180 min)

End of Question 1

Question 2 40 Marks

Please complete the following questions, based on your datasets created in Question 1. You need to meet each of the criteria for each of the questions asked.

Before performing operations on your data, you need to clean your data, remove any zero values and document the process thoroughly in your report.

2.1 Create frequency tables based on the following headings, and using the following criteria:

- Average hours spent on campus
- Student age
- Student mark

2.2 Using Matplotlib, complete the following:

- Use a bar chart to show the ages and numbers of students. Your x-axis should represent the age groups used in the previous question.
- Use a line graph to show if there is a correlation between higher marks and more time spent on campus.
- Use a scatter chart to plot each student’s mark and the time taken on the examination, in minutes.
- Use a scatter chart to plot the relationship between time spent on campus and the student’s age.

End of Question 2

Question 3 30 Marks

This section will require you to document the process and the conclusions you have drawn from the graphs and diagrams you created in the previous question.

You will need to include all the graphs and diagrams you have created in your report and a section explaining the conclusion you drew from that specific diagram or graph, as well as the factors that led to your conclusion.

This needs to be neatly and professionally documented and include your commented source code for all the graphs you created, as well as any operations you needed to perform on your data before you made use of it in graph format.

Question 4 70 Marks

Study the scenario and complete the question(s) that follow:

> As a data analyst, you have been provided with a dataset called "percent-bachelors-degrees-women-usa.csv". The dataset is a collection of data that describes the percentage of women who earned a degree in certain majors over certain years in the United States of America. Using this dataset, provide answers to the questions below.
>
> Source: Ajayi, M. (2020)

4.1 Using Python, execute the following requests:

a. Read the percent-bachelors-degrees-women-usa.csv file as a list of lists. (2 Marks)

b. Assign the result to the variable BDW. (2 Marks)

c. Display the first five rows of BDW separately on different lines. (2 Marks)

d. Remove the header row (column names) of the dataset and assign the rest of the dataset to BDW1. (2 Marks)

e. Using slicing, display the first, second and third row of BDW1. (2 Marks)

4.2 Using for loops in Python, from BDW1, create two dictionaries as instructed below:

a. A dictionary called "Indexcount_year" where the years are the keys and the values are the frequencies of occurrence of each year. (5 Marks)

b. A dictionary called "Indexpercent_year" where the years are the keys and the values are the indices of each item in the list. (5 Marks)

> **Note to Student**
>
> The index and frequencies should be integers and not strings.

4.3 Using for loops in Python, from BDW1, read the following as a list:

a. The percentages of women who earned a degree per year in the following majors:

- Math and Statistics. Assign the value to the variable Maths_Stats. (5 Marks)
- Physical Sciences. Assign the value to the variable Physic_Sci. (5 Marks)
- Engineering. Assign the value to the variable Engine. (5 Marks)
- Computer Science. Assign the value to the variable Comp_Sci. (5 Marks)

b. The year captured in the dataset as to when these degrees were earned. Assign it to the variable Year. (5 Marks)

4.4 Using the list obtained and its orientation (i.e. I-IV) in 4.3 a:

a. Create a Numpy array called “Selected4Majors”. **Each element in the list must be converted to a float before creating the Numpy array*.*** (5 Marks)

b. Create a dictionary called “Majors” where the major is the key and the values are the indices of each major, in regards to their orientation. (2 Marks)

4.5 Write a Python function that will accept two arguments, data and majorlist. Using this function, do the following:

a. With a for loop plot the data in the variable Year, obtained in 4.3 b (x-axis) against the data in the variable Selected4Majors (y-axis). This should be done on the same graph. (5 Marks)

b. Display the legend of the majors for each plot in the upper left corner. (2 Marks)

c. Set the title to “Percentage of Selected4Degrees Awarded per Year”, label the x-axis “Year” and the y-axis “Selected4Degrees”. (3 Marks)

4.6 Based on the plots in 4.5, what can you deduce from the graph? **You must provide any relevant comments.** (8 Marks)

Note to Student

For each question in question 4, describe what you want to achieve and provide comments for each line of code.

Section B

Plagiarism and Referencing

Eduvos places high importance on honesty in academic work submitted by students, and adopts a policy of zero tolerance on cheating and plagiarism. In academic writing, any source material e.g. journal articles, books, magazines, newspapers, reference material (dictionaries), online resources (websites, electronic journals or online newspaper articles), must be properly acknowledged. Failure to acknowledge such material is considered plagiarism; this is deemed an attempt to mislead and deceive the reader, and is unacceptable.

Eduvos adopts a zero tolerance policy on plagiarism, therefore, any submitted assessment that has been plagiarised will be subject to severe penalties. Students who are found guilty of plagiarism may be subject to disciplinary procedures and outcomes may include suspension from Eduvos or even expulsion. Therefore, students are strongly encouraged to familiarise themselves with referencing techniques for academic work. Students can access the Guide to Referencing on *my*LMS.

Negative Marking

- At the discretion of the marker, if a student has committed plagiarism, an immediate 0% will be awarded for the project and 10% will be deducted from their next submission.