1st SIT COURSEWORK QUESTION PAPER: **Autumn Semester 2021**

Module Code:	**CC7182NI**
Module Title:	**Programming for Data Analytics**
Module Leader:	**Sukrit Shakya (Islington College)**

Coursework Type:	**Individual**
Coursework Weight:	This coursework accounts for **100%** of your total module grades.
Submission Date:	**Week 13**
When Coursework is given out:	**Week 10**
Submission Instructions:	Submit the following to Islington College RTE department • Report in PDF format • Source code
Warning:	London Metropolitan University and Islington College takes Plagiarism seriously. Offenders will be dealt with sternly.

CC7182 Programming for Data Analytics
Autumn Semester 2021

Coursework Assignment

The coursework is an **individual** assessment weighted **100%** of the marks for the module. It is primarily an exercise in applying programming knowledge and skills to data analysis tasks, demonstrating your skills for problem-solving and critical thinking/evaluation.

The assignment consists of 2 parts. The first part involves the analysis of a student performance dataset from two Portuguese schools. The second part involves the analysis of livestock data of Nepal.

You are expected to write Python code and a technical report on data understanding, preparation, exploration, analysis and visualization.

Coursework Submission

The course work is due on **Week 13.** You need to submit the following items via Google Classroom:

- Technical report as a single PDF file
- Associated Python files as a ZIP file

Please note that plagiarism is a serious academic offence, for which penalties are severe. All suspected cases of plagiarism will be reported.

Part 1. Analysis of a Student Performance Dataset

Data Description

The data given here is of student achievements in secondary education of two Portuguese schools namely *“Gabriel Pereira”* and *“Mousinho da Silveira”*. The data attributes include student grades (in mathematics), demographic, social and school related features and it was collected by using school reports and questionnaires.

The detailed data description of the data file is given below:

*Filename****: student.csv***

The data set contains 395 student records. Each record consists of 33 variables, which includes information about the students.

Variable 33, G3 – final grade (numeric: 0 - 20), is the target variable.

The dataset is in CSV format with following attributes:

1. ***school*** - student's school ("GP" - Gabriel Pereira or "MS" - Mousinho da Silveira)
2. ***sex*** - student's sex
3. ***age*** - student's age
4. ***address*** - student's home address type ("U" - urban or "R" - rural)
5. ***famsize*** - family size ("LE3" - less or equal to 3 or "GT3" - greater than 3)
6. ***Pstatus*** - parent's cohabitation status ("T" - living together or "A" - apart)
7. ***Medu*** - mother's education

8. ***Fedu*** - father's education
9. ***Mjob*** - mother's job ("teacher", "health" care related, civil "services" (e.g. administrative or police), "at_home" or "other")
10. ***Fjob*** - father's job ("teacher", "health" care related, civil "services" (e.g. administrative or police), "at_home" or "other")
11. ***reason*** - reason to choose this school (close to "home", school "reputation", "course" preference or "other")
12. ***guardian*** - student's guardian ("mother", "father" or "other")
13. ***traveltime*** - home to school travel time
14. ***studytime*** - weekly study time
15. ***failures*** - number of past class failures
16. ***schoolsup*** - extra educational support
17. ***famsup*** - family educational support
18. ***paid*** - extra paid classes within the course
19. ***activities*** - extra-curricular activities
20. ***nursery*** - attended nursery school
21. ***higher*** - wants to take higher education
22. ***internet*** - Internet access at home
23. ***romantic*** - with a romantic relationship
24. ***famrel*** - quality of family relationships

25. ***freetime*** - free time after school

26. ***goout*** - going out with friends

27. ***Dalc*** - workday alcohol consumption

28. ***Walc*** - weekend alcohol consumption

29. ***health*** - current health status

30. ***absences*** - number of school absences

31. ***G1*** - first period grade (from 0 to 20)

32. ***G2*** - second period grade (from 0 to 20)

33. ***G3*** - final grade (from 0 to 20, target variable)

The tasks are as follows:

1. ***Data Understanding***

   - Understand what your data resources are and what the characteristics of those resources are. Write down your findings including the characteristics of the different columns in the dataset.

   (4 marks)

2. ***Data Transformation***

   - Write code to transform variables according to the following guidelines:

     a) ***school***, ***sex***, ***address, famsize, Pstatus, schoolsup, famsup, paid, activities, nursery, higher, internet*** and ***romantic*** into binary; 0 or 1 (create new columns without overwriting the existing ones)

     (1+1+1+1+1+1+1+1+1+1+1+1+1 = 13 marks)

b) ***Medu***, ***Fedu***, ***Mjob***, ***Fjob***, ***reason, guardian, traveltime, studytime, famrel, freetime, gout, Dalc, Walc*** and ***health*** into ordinal numbers based on number of cases in the data set (create new columns without overwriting the existing ones).

(1+1+1+1+1+1+1+1+1+1+1+1+1+1 = 14 marks)

c) create a new column named ***age_category*** whose values should be based on the values in the ***age*** column, divide the values into 3 ordinal numbers; **1** – 15 to 17, **2** – 18 to 20, **3** – 21 and over

(1 mark)

d) create a new column named ***passed (yes or no)*** whose values should be based on the values present in the ***G3*** column (>=8 – yes, <8 – no)

(1 mark)

3. Initial Data Analysis

- Write code to show the summary statistics (sum, mean, median, standard deviation, max and min) of the variables ***age, absences, G1, G2*** and ***G3***.

(1+1+1+1+1 = 5 marks)

- Write code to calculate and show the correlation between the variables ***absences, failures, G1, G2*** and ***G3***. Present the result using a heatmap and interpret the results.

(2 marks)

4. Data Exploration and Visualization

- Write code to show histogram plots and boxplots to visualize the distribution of the variables ***age, absences*** and ***G3***. Interpret the results and comment about the distribution of each variable.

(2 + 2 + 2 = 6 marks)

- Write code to show a bar graph of the total number of students who passed the final term grouped according to the school that they belong to. Use proper labels in the graph and interpret the results.

(2 marks)

- Write code to show a bar graph of the total number of students who failed the final term grouped according to their weekly study time. Use proper labels in the graph and interpret the results.

(2 marks)

5. Further Analysis

- For this task, you need to further explore the given dataset on your own by using different analysis and visualization techniques and then present the insights that you have gained. Be creative and come up with interesting insights and draw conclusions from the data.

(20 marks)

Part 2. Analysis of Livestock Data of Nepal

Data Description

The data given here is about livestock raised across Nepal according to different districts/regions and the commodities produced by them. The overall data is spread across multiple files. The list of files are as follows:

1. horseasses-population-in-nepal-by-district.csv
2. milk-animals-and-milk-production-in-nepal-by-district.csv
3. net-meat-production-in-nepal-by-district.csv
4. production-of-cotton-in-nepal-by-district.csv
5. production-of-egg-in-nepal-by-district.csv
6. rabbit-population-in-nepal-by-district.csv
7. wool-production-in-nepal-by-district.csv
8. yak-nak-chauri-population-in-nepal-by-district.csv

You are required to study the data and understand its structure and properties. Then using python you should clean and merge all the data sources and perform EDA (Exploratory Data Analysis) on it.

The tasks are as follows:

1. ***Data Understanding***
    - Understand what your data resources are and what the characteristics of those resources are. Write down your findings including the characteristics of the different columns in the dataset.

(5 marks)

2. ***Data Merging and Cleaning***
    - Read all the data sources and merge them into a single data source. Do necessary data cleaning and pre-processing.

(5 marks)

3. Exploratory Data Analysis

- Explore the data using different analysis and visualization techniques and then present the insights that you have gained. Be creative and come up with interesting insights

(20 marks)

The technical report should have screen shots of the code. The results achieved and the interpretations of the results should also be included in the technical report. Python code should include adequate comments as well.

For ***task 5*** in ***part 1*** and ***task 3*** in ***part 2***, 20 marks are allocated according to the following categories

- creativity (5 marks)
- quality of analysis performed/interpretations of the results/presentation of insights (10 marks)
- programming style/use of tools (5 marks)

END