4BUIS014C Business Computing – Portfolio (2023-24)	
Module leader	Dr Deepika
Unit	Portfolio
Weighting:	30%
Qualifying mark	12%
Description	The portfolio will comprise a number of exercises that students will be allowed to attempt these exercises on the topic(s)); individualised feedback will be given to students to help them improve their understanding, learn from their performance, and prepare for the coursework.
Learning Outcomes Covered in this Assignment:	This assignment contributes towards the following Learning Outcomes (LOs): LO2 Utilise a programming language, software packages and tools to create small scale business applications to retrieve and manipulate data stored in a data repository (database, spreadsheet, etc); LO3 Utilise a programming language, software packages and tools to extract from, manipulate and load to data stored in different repositories.
Handed Out:	23rd October 2023
Due Date	16th November 2023 at 13:00
Expected deliverables	Submit on Blackboard a zip file the required documentation (either in rar or zip). All implemented codes should be included in your zip file, and a detailed report consisting of the screenshots of the python code and the output along with the explanation and analysis for the questions.
Method of Submission:	Electronic submission on BB via a provided link close to the submission time.
Type of Feedback and Due Date:	Feedback will be provided on BB within three weeks from the due date

4BUIS014W Business Computing, Portfolio – 2023-24

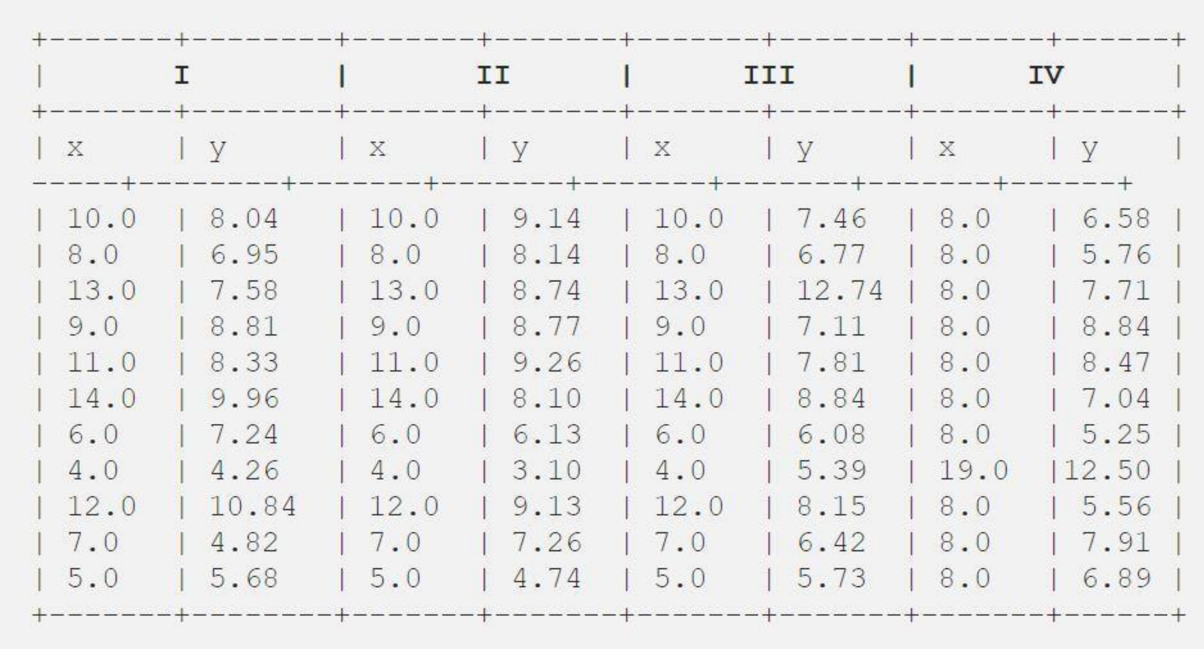
Value: 30% of Module Mark

Online Submission Date (via BB): 16/11/2023 at 13:00

Problem One: Anscombe's quartet comprises four datasets (I, II, III, IV) that have nearly identical simple statistical properties, yet appear very different when graphed. Each dataset consists of eleven (x,y) points. They were constructed in 1973 by the statistician Francis Anscombe to demonstrate both the importance of graphing data before analysing it and the effect of outliers on statistical properties.

Simple understanding:

Once Francis John "Frank" Anscombe who was a statistician of great repute found 4 sets of 11 datapoints in his dream and requested the council as his last wish to plot those points. Those 4 sets of 11 data-points are given below.



I		II		III		IV	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

After that, the council analysed them using only descriptive statistics between x and y for each data set ((**I, II, III, IV**)).

Please download the csv file from the module BB site under Section Assessment or from here.

Tasks:

1. Construct in Python four data frames (df1, df2, df3, df4) to store the four data sets (I, II, III, IV). Each dataset consists of eleven (x,y) points;

[1 Marks]

2. Find the basic descriptive statistics in Python using the method describe();

[1Marks]

3. Plot a scatterplot in Python using the plt.scatter(x, y) and plt.show() methods one for each dataset (I, II, III, IV)

[2 Mark]

4. Based on the plots obtained explain why descriptive statistics are not enough to describe the four data sets (I, II, III, IV). Provide a short answer 3-5 lines.

[1 Mark]

Problem Two: This story is interesting to young people about to join a college in USA. I'd like to be able to perform some analysis for myself using the data found on GitHub behind the story The Economic Guide To Picking A College Major.

data_url=("https://raw.githubusercontent.com/fivethirtyeight/data/master/college-majors/recent-grads.csv")

With the aid of plots answer the following questions

1. If you pick a major with higher median earnings, do you also have a lower chance, of unemployment? Explain if a correlation does exist between earnings and unemployment. Hint: use a scatter plot. [2 Marks]
2. Let's investigate all majors whose median salary is above $60,000. First, you need to filter these majors with the mask df[df["Median"] > 60000]. Then you can create another bar plot showing all three earnings columns: ["P25th", "Median", "P75th"]
   Is it true that petroleum engineering majors have by far the best-paid recent graduates? [2 Marks]

Dataset Description: You be working with a dataset on the job outcomes of students who graduated from college between 2010 and 2012. The original data on job outcomes was released by the American Community Survey, which conducts surveys and aggregates data.
Each row in the dataset represents a different major in college and contains information on gender diversity, employment rates, median salaries, and more. Here are some of the columns in the dataset:

1. Rank - Rank by median earnings (the dataset is ordered by this column).
2. Major_code - Major code.
3. Major - Major description.
4. Major_category - Category of major.
5. Total - Total number of people with major.
6. Sample_size - Sample size (unweighted) of full-time.
7. Men - Male graduates.
8. Women - Female graduates.
9. ShareWomen - Women as share of total.
10. Employed - Number employed.
11. Median - Median salary of full-time, year-round workers.
12. Low_wage_jobs - Number in low-wage service jobs.
13. Full_time - Number employed 35 hours or more.
14. Part_time - Number employed less than 36 hours.

Problem Three: Revisit **4BUIS014W_W3_Pandas under folder week 3 on BB;** more specifically revise the part on "Vectorisation - A more efficient way to perform Iterations with Pandas"

1. Reload with Pandas the Data Set user_reviews.csv available on BB under folder Week 3, in a data frame named df.
2. Add the column "len_text" as in **4BUIS014W_W3_Pandas.**
3. Create a new column, called "super category", based on multiple conditions and existing column values. In particular, a person should be classified as:
   - "expert reviewer", if the length of their review, "len_text", is greater than 1000 characters or if the review "grade" is greater than or equal to 9.
   - "opposed-reviewer", if the review "grade" is less than or equal to 1 AND the length of their review is greater than 1000 characters.
   - "neutral reviewer" otherwise.

   Implement the above the super category feature code using a

   - for loop [2 Marks]

- while loop [2Marks]
- vectorisation [2Marks]

4. You need to submit the source code in a single file for the above tasks.

Problem Four: Consider the following SQL DDL Statements

```
CREATE TABLE locations
(
            location_id                   INT(3),
            street_address                VARCHAR(50) unique not null,
            postal_code                   VARCHAR(10) not null,
city                          VARCHAR(50) not null,
state_province                VARCHAR(50) not null,     country
            VARCHAR(50) not null,
            constraint                    l_lid_pk PRIMARY KEY (location_id)
);


CREATE TABLE departments
(
            department_id                 INT(4),
            department_name               VARCHAR(20) unique not null,
location_id                   INT(3),              constraint
d_did_pk PRIMARY KEY (department_id),   constraint
d_lid_fk FOREIGN KEY (location_id)          references
locations(location_id)
);


INSERT INTO
locations (location_id, street_address, postal_code, city, state_province, country)
VALUES
(100, '2 Nice Road', 'N2 7TH', 'London', 'Greater London', 'UK'),
(200, '23 Pretty Road', 'BS1 8FD', 'Bristol', 'Bristol County', 'UK'),
(300, '26 Great Street', 'BN1 4BF', 'Brigthon', 'Sussex', 'UK'),
(400, '143 Lovely Road', 'CB1 2NV', 'Cambridge', 'Cambridgeshire', 'UK');

INSERT INTO departments (department_id, department_name, location_id) VALUES
(10, 'IT', 100),
(20, 'Operations', 200),
(30, 'Sales', 300),
(40, 'Marketing', 200),
(50, 'Management', NULL);
```

1. Write a Python program to
    a. create the listed tables (locations, departments)
    b. insert data to listed tables (locations, departments)

[3 Marks]

2. Display all departments and their locations, as well the locations with no departments.

[3 Marks]

Problem Five: Building a Module to Create and Manipulate a Database Using SQLITE3

What are the *important functions* of a Module named as **DBModule** that creates a Database Using SQLITE3?

Task 1

Important Functions to be implemented via a Module named **DBModule**:

1. Connect to Sqlite Server and create a new Database Space. You need to supply 2 arguments only, the database location and the database name

   [2Marks]

2. Read data from a csv file. You need to supply 2 arguments only the csv file location and the file name;

   [2Marks]

3. Write the csv Data to a table; You need to supply 2 arguments only the dataframe that holds the csv data and the name of the table for the data to be stored;

   [2Marks]

Task 2

1. Write a small programme to import functions 1-3 from the **DBModule and execute them.** You may use any of the csv files provided in the tutorials for the purpose of demonstrating the correctness of your implementation.

   [3Marks]

Assessment regulations

Refer to section 4 of the "How you study" guide for undergraduate students for a clarification of how you are assessed, penalties and late submissions, what constitutes plagiarism etc.

Penalty for Late Submission

If you submit your coursework late but within 24 hours or one working day of the specified deadline, 10 marks will be deducted from the final mark, as a penalty for late submission, except for work which obtains a mark in the range 40 – 49%, in which case the mark will be capped at the pass mark (40%). If you submit your coursework more than 24 hours or more than one working day after the specified deadline you will be given a mark of zero for the work in question unless a claim of Mitigating Circumstances has been submitted and accepted as valid. It is recognised that on occasion, illness or a personal crisis can mean that you fail to submit a piece of work on time. In such cases you must inform the Campus Office in writing on a mitigating circumstances form, giving the reason for your late or non-submission. You must provide relevant documentary evidence with the form. This information will be reported to the relevant Assessment Board that will decide whether the mark of zero shall stand. For more detailed information regarding University Assessment Regulations, please refer to the following website:http://www.westminster.ac.uk/study/current-students/resources/academic-regulations