Faculty of Natural Sciences

School of Computer Science and Mathematics

Assessment Brief

Module	CSC-40054
Assessment Component	Coursework
Weighting	50%
Deadline	13:00 pm on 22nd March 2024
Module Leader	Wenjuan Zhou
Office Hours/Meeting Booking Link	Monday and Tuesday during semester 2 teaching. Email: w.zhou@keele.ac.uk to arrange a meeting.

1. What is the task for this assessment?

One report should be submitted that covers the following two tasks:

- Part I: Database management
- Part II: Data analytics

2. What is required of me in this assessment?

Guidelines	Detailed guidelines for the tasks are listed below.
Self- assessment checklist Make sure that you…	• The required components and weightings of each part are detailed below. Use these lists to structure and focus the content of your report. • Use the materials discussed in the lectures and tutorials to complete the tasks. • Ensure that your codes run correctly.
Three key pieces of advice based on the feedback given to the previous cohort who completed this assignment	• Read the task description thoroughly and ask for clarification. • Read the formatting guidelines and stick to them. • Start working on your coursework early. Ensure that you understand the tasks and the submission process.
Formatting Guidelines	• A report (maximum 3000 words) on the accessing, storage, manipulation and analysis of data available from an internet-based data repository. The code

	needs to be submitted as an appendix. The appendix does not count for the word count. • Your submission should be in the form of a single zipped file. This zipped file should be named with your Student Number. • The zipped file should contain your report in two formats MS Word and PDF. So, it should contain ONLY two files. Any other submitted file will not be marked. • For each task, you should include the screenshot of the output after you run the code (when its execution finishes successfully). Failing to include the output screenshot will cause you to lose marks. • The codes should be included as the appendix at the end of the report. You should clearly specify which code is related to which task. Otherwise, they will be marked zero. • Screenshots of codes are not accepted in the appendix and will be marked zero. You should include the original code with the correct indentation so anyone can copy and paste the code into an editor and run it. • If you have written the code, but it doesn't run correctly, mention it in your report. • An example of the accepted appendix is shown at the end of the coursework.
Referencing Style	Harvard referencing style
Assessment Criteria/ Markscheme:	The assessment criteria and mark scheme are listed separately below

3. What is the purpose of this assessment?

The following table shows which of the module learning outcomes are being assessed in this assignment. Use this table to help you see where and how to transfer feedback from one assignment to another. Note that your feedback may mention some of these outcomes, but that you will not receive a 'mark' against each one.

Module Learning Outcomes assessed
1. Evaluate available data and determine how best to analyse the information available to provide required outcomes. 2. Evaluate machine learning methods in the context of statistical analysis of data representing social or natural systems. 3. Develop advanced applications of statistical data analytics techniques using an advanced specialist programming language (e.g. Python). 4. Assess the options of storing, managing and manipulating very large volumes of data in the context of research or business organisations. 5. Assess a range of statistical approaches and apply the correct statistical approaches to extract information from a set of data typically available in a modern business or research organisation.

Rationale	• Parts I assess your ability to access and manipulate databases

	• Part II assesses your ability to critically evaluate and apply big data applications, advanced analytics, and statistical modelling techniques appropriate to different types of problems.

4. What resources might I use to get started?

Lecture slides and your completed in-class activities. However, when needed, you are expected to research and use different resources to complete the tasks.

Guidelines for Tasks

For the coursework assignment, you will be working on a dataset from a car-sharing company. The dataset contains information about the customers' demand rate between January 2017 and August 2018. The data were collected on an hourly basis and included the time data such as date, hour, and season as well as weather data such as the weather condition, temperature, humidity, and wind speed. The 'demand' column represents the customer's willingness for renting a car for a specific time. Higher demand rates show that customers are more willing to rent a car and vice versa. A complete description of the data is also shown in Table I.

IMPORTANT NOTICE: You should complete the Part I (database management) tasks using only the sqlite3 python module and SQL statements. You shouldn't use any other python modules for these tasks. However, you can use any modules for Part II (data analytics) tasks or for importing and exporting data.

Download the dataset "CarSharing" from the KLE and complete the tasks.

Part I: Database Management (50%)

1. Create an SQLite database and import the data into a table named "CarSharing". Create a backup table and copy the whole table into it. **[5%]**
2. Add a column to the CarSharing table named "temp_category". This column should contain three string values. If the "feels-like" temperature is less than 10 then the corresponding value in the temp_category column should be "Cold", if the feels-like temperature is between 10 and 25, the value should be "Mild", and if the feels-like temperature is greater than 25, then the value should be "Hot". **[5%]**
3. Create another table named "temperature" by selecting the temp, temp_feel, and temp_category columns. Then drop the temp and temp_feel columns from the CarSharing table. **[5%]**
4. Find the distinct values of the weather column and assign a number to each value. Add another column named "weather_code" to the table containing each row's assigned weather code. **[5%]**
5. Create a table called "weather" and copy the columns "weather" and "weather_code" to this table. Then drop the weather column from the CarSharing table. **[5%]**
6. Create a table called time with four columns containing each row's timestamp, hour, weekday name, and month name (Hint: you can use the strftime() function for this purpose). **[5%]**
7. Assume it's the first day you have started working at this company and your boss Linda sends you an email as follows: "

   Hello, welcome to the team. I hope you enjoy working at this company. Could you please give me a report containing the following information:

   a) Please tell me which date and time we had the highest demand rate in 2017. **[4%]**
   b) Give me a table containing the name of the weekday, month, and season in which we had the highest and lowest average demand rates throughout 2017. Please include the calculated average demand values as well. **[4%]**
   c) For the weekday selected in (b), please give me a table showing the average demand rate we had at different hours of that weekday throughout 2017. Please sort the results in descending order based on the average demand rates. **[4%]**

d) Please tell me what the weather was like in 2017. Was it mostly cold, mild, or hot? Which weather condition (shown in the weather column) was the most prevalent in 2017? What was the average, highest, and lowest wind speed and humidity for each month in 2017? Please organise this information in two tables for the wind speed and humidity. Please also give me a table showing the average demand rate for each cold, mild, and hot weather in 2017 sorted in descending order based on their average demand rates. **[4%]**

e) Give me another table showing the information requested in (d) for the month we had the highest average demand rate in 2017 so that I can compare it with other months. **[4%]**

NOTICE: Full marks for task 7 will be given to solutions that use the CarSharing table after all changes in tasks 1-6 have been made to it.

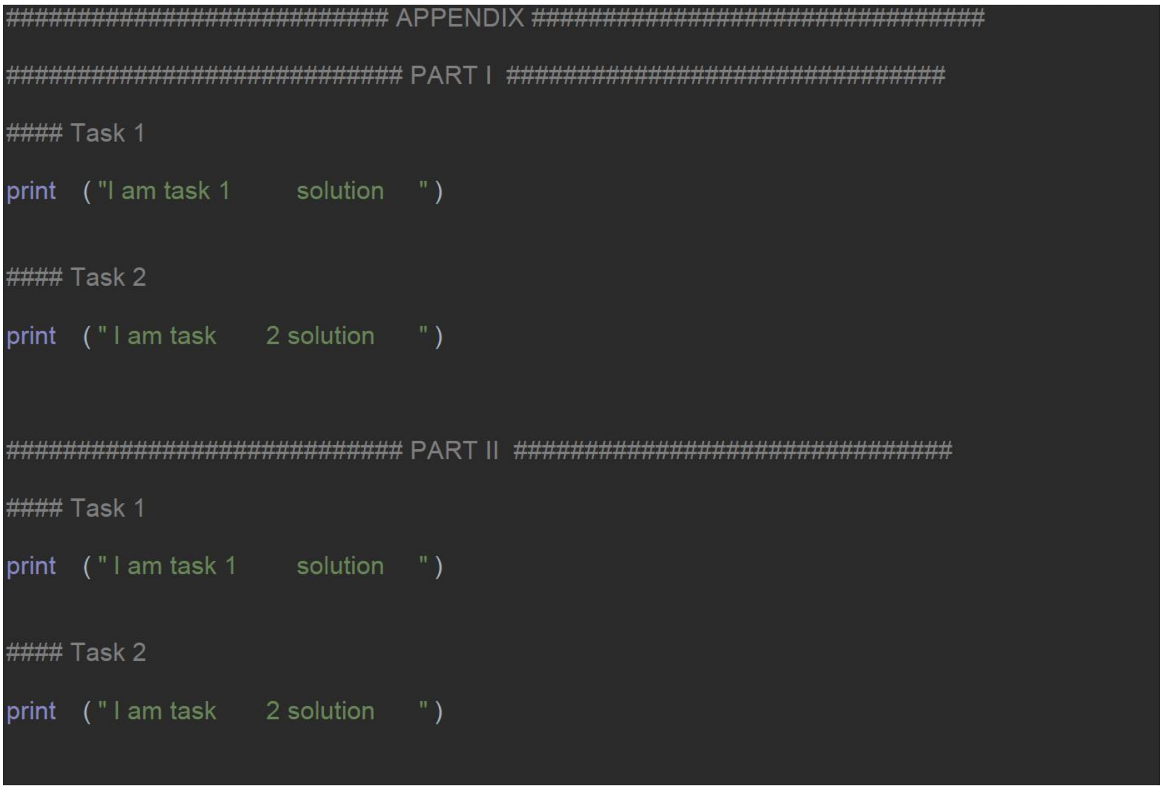
Part II: Data Analytics (50%)

1. Import the CarSharing table into a CSV file and preprocess it with python. You need to drop duplicate rows and deal with null values using appropriate methods. **[5%]**
2. Using appropriate hypothesis testing, determine if there is a significant relationship between each column (except the timestamp column) and the demand rate. Report the tests' results. **[5%]**
3. Please describe if you see any seasonal or cyclic pattern in the temp, humidity, windspeed, or demand data in 2017. Describe your answers. **[5%]**
4. Use an ARIMA model to predict the weekly average demand rate. Consider 30 percent of data for testing. **[7.5%]**
5. Use a random forest regressor and a deep neural network to predict the demand rate and report the minimum square error for each model. Which one is working better? Why? Please describe the reason. **[10%]**
6. Categorize the demand rate into the following two groups: demand rates greater than the average demand rate and demand rates less than the average demand rate. Use labels 1 and 2 for the first and the second groups, respectively. Now, use three different classifiers to predict the demand rates' labels and report the accuracy of all models. Use 30 percent of data for testing. **[10%]**
7. Assume k is the number of clusters. Set k=2, 3, 4, and 12 and use 2 methods to cluster the temp data in 2017. Which k gives the most uniform clusters? (Clusters are called uniform when the number of samples falling into each cluster is close.) **[7.5%]**

NOTICE: You should use the original CarSharing dataset for Part II, not the dataset you have modified in Part I.

Guidelines for Preparing the Appendix

The following example shows a well-formatted appendix.

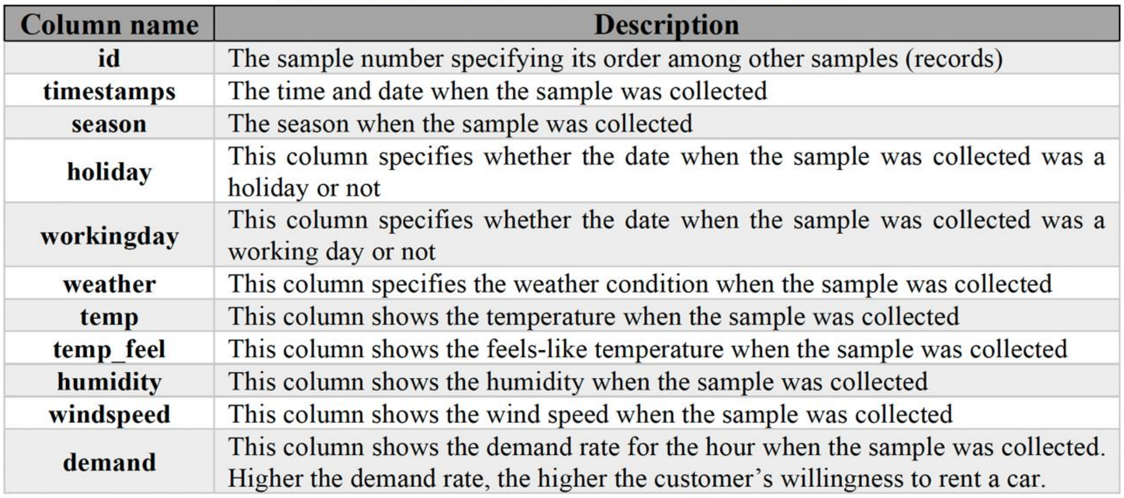


```
########################### APPENDIX ################################

############################ PART I  ###############################

#### Task 1

print   ("I am task 1        solution     " )


#### Task 2

print   (" I am task      2 solution      " )



############################ PART II  ###############################

#### Task 1

print   (" I am task 1       solution     " )


#### Task 2

print   (" I am task      2 solution      " )
```

As you can see, PART I and PART II are clearly separated, and the code can be copied and pasted. The following steps show how you can copy your code and paste it into your report in the same way as the above example:

1. Select the code in your IDE using the mouse.
2. Use CTRL+C to copy the code.
3. Use CTRL+V to paste the code into your report.

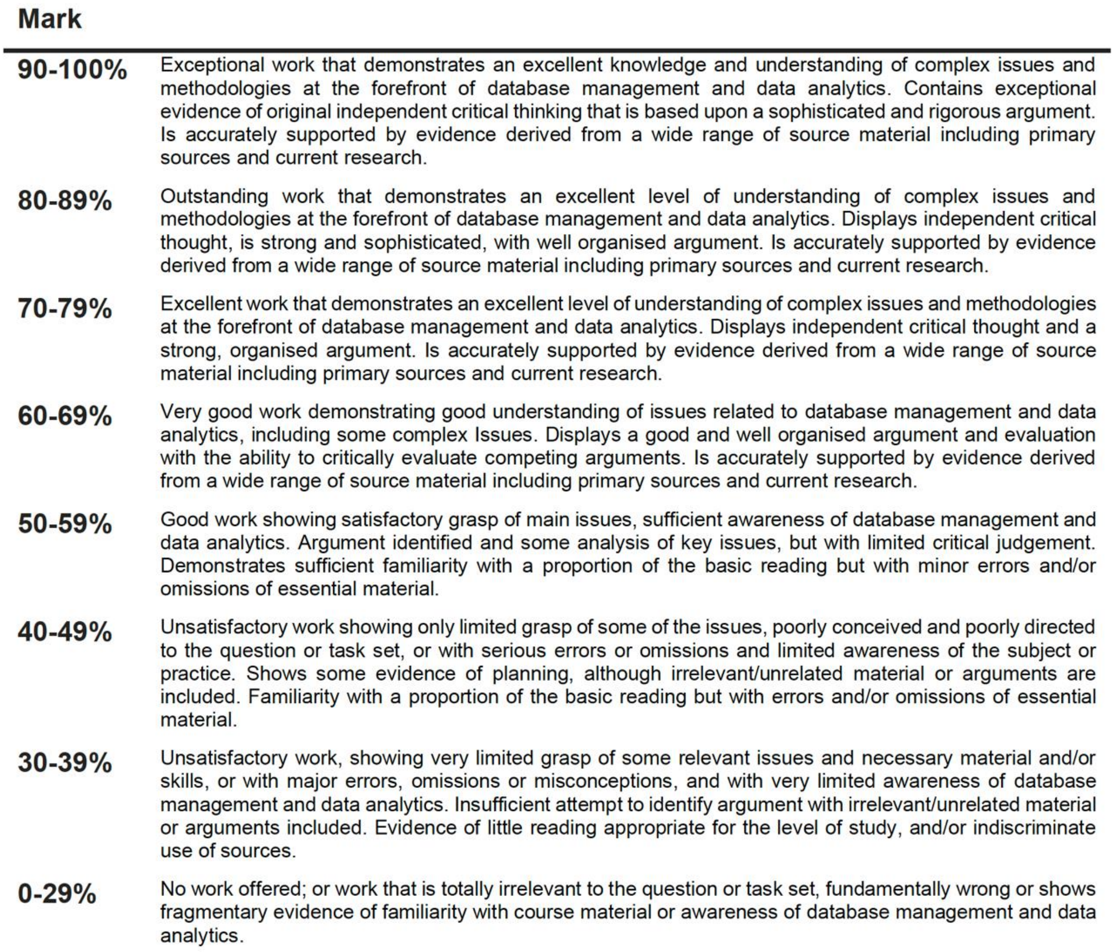
TABLE I. THE DATASET'S COLUMNS DESCRIPTION



Column name	Description
id	The sample number specifying its order among other samples (records)
timestamps	The time and date when the sample was collected
season	The season when the sample was collected
holiday	This column specifies whether the date when the sample was collected was a holiday or not
workingday	This column specifies whether the date when the sample was collected was a working day or not
weather	This column specifies the weather condition when the sample was collected
temp	This column shows the temperature when the sample was collected
temp_feel	This column shows the feels-like temperature when the sample was collected
humidity	This column shows the humidity when the sample was collected
windspeed	This column shows the wind speed when the sample was collected
demand	This column shows the demand rate for the hour when the sample was collected. Higher the demand rate, the higher the customer's willingness to rent a car.

Assessment Criteria

The criteria are adapted from the University assessment criteria and relate to all parts of this assessment. Within each element of the marking scheme, marking will be according to the assessment criteria below.



Mark	
90-100%	Exceptional work that demonstrates an excellent knowledge and understanding of complex issues and methodologies at the forefront of database management and data analytics. Contains exceptional evidence of original independent critical thinking that is based upon a sophisticated and rigorous argument. Is accurately supported by evidence derived from a wide range of source material including primary sources and current research.
80-89%	Outstanding work that demonstrates an excellent level of understanding of complex issues and methodologies at the forefront of database management and data analytics. Displays independent critical thought, is strong and sophisticated, with well organised argument. Is accurately supported by evidence derived from a wide range of source material including primary sources and current research.
70-79%	Excellent work that demonstrates an excellent level of understanding of complex issues and methodologies at the forefront of database management and data analytics. Displays independent critical thought and a strong, organised argument. Is accurately supported by evidence derived from a wide range of source material including primary sources and current research.
60-69%	Very good work demonstrating good understanding of issues related to database management and data analytics, including some complex Issues. Displays a good and well organised argument and evaluation with the ability to critically evaluate competing arguments. Is accurately supported by evidence derived from a wide range of source material including primary sources and current research.
50-59%	Good work showing satisfactory grasp of main issues, sufficient awareness of database management and data analytics. Argument identified and some analysis of key issues, but with limited critical judgement. Demonstrates sufficient familiarity with a proportion of the basic reading but with minor errors and/or omissions of essential material.
40-49%	Unsatisfactory work showing only limited grasp of some of the issues, poorly conceived and poorly directed to the question or task set, or with serious errors or omissions and limited awareness of the subject or practice. Shows some evidence of planning, although irrelevant/unrelated material or arguments are included. Familiarity with a proportion of the basic reading but with errors and/or omissions of essential material.
30-39%	Unsatisfactory work, showing very limited grasp of some relevant issues and necessary material and/or skills, or with major errors, omissions or misconceptions, and with very limited awareness of database management and data analytics. Insufficient attempt to identify argument with irrelevant/unrelated material or arguments included. Evidence of little reading appropriate for the level of study, and/or indiscriminate use of sources.
0-29%	No work offered; or work that is totally irrelevant to the question or task set, fundamentally wrong or shows fragmentary evidence of familiarity with course material or awareness of database management and data analytics.