KEELE UNIVERSITY
SCHOOL OF COMPUTING AND MATHEMATICS
CSC-40054
COURSEWORK ASSIGNMENT

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 II.

Your assignment consists of two parts. The first part includes two sets of tasks corresponding to database management and data analytics. For the second part, you need to present a report containing a description of your work and the codes you have written. You should comment on your code specifying which code is related to which question and what task each piece of code is doing. Your coursework marks comprise 50% of your final marks.

<u>IMPORTANT NOTICE</u>: You should complete the 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 data analytics tasks or for importing and exporting data.

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

Part 1. 1

Database Management

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.

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".

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.

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- 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.

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 surftime() function for this purpose).

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.

(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.

(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.

(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.

(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.

Please don't rush. You can prepare your report by the end of the 12th of January 2023. However, I may be busy by that time. So, please send your report to Amin, and he will check it.

Kind regards,
Linda"

Please prepare the information Linda requested and send it to me as she mentioned 😉

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 1. 2

Data Analytics

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.

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.

3- Please describe if you see any seasonal or cyclic pattern in the temp, humidity, windspeed, or demand data in 2017. Describe your answers.

4- Use an ARIMA model to predict the weekly average demand rate. Consider 30 percent of data for testing.

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.

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.

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)

Part 2

Report

The report should not be more than 3000 words. The codes should be commented on and added to the report as appendices. The appendix does not count for the word count. **You should specify which code is related to which task**. The report should contain the description of your solution for each task including the answers to the questions asked in different tasks. The report should be submitted as a **single** ZIP file which should be named using your name and Keele's user ID. The report must be uploaded to the KLE **by the end of 12th of January 2023.**

TABLE I. THE MARKING SCHEMA

Marking Schema		
Section name	**marks**	**Criteria**
Part 1.1 – Database Management	**30**	An excellent solution should use only SQL statements and generate the expected output. Only the sqlite3 python module is allowed in this section. You can also use other python modules for importing or exporting data. Extra marks will be given to the novelty in writing the queries.
Part 1.2 – Data Analytics	**30**	The solution should generate the expected output, be logically related to the related questions, and provide compelling answers to them. For prediction tasks, higher marks will be given to the solutions producing higher accuracies or lower mean square errors.
Part 2 - Report	**40**	The report should be structured well complying with academic report standards. It should properly explain the solutions provided in part 1 while critically evaluating them. Charts and appropriate visualization should be used where applicable.
Total	**100**	

TABLE II. 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.