University of Windsor
Faculty of Science

COMP 8157 Advanced Database Topics
University of Windsor, School of Computer Science
Lab 1
Weight: 3.75 %

Aim: This assignment will assess your understanding of data Mining concepts. Submission:

A report to answer all question and R file

Due:

To accomplish this assignment, you need to:

- Download the "**vehicles**" dataset from Brightspace.
- Use R studio for analysing and visualizing the dataset.
- Grades are given for each question.

Part 1: Data Exploration (12 marks)

1. Import the **Vehicle** dataset, summarize it and explain the output **(2 marks).**
2. Show the structure and dimension of the dataset and explain it **(2 marks)**.
3. Show the column names of the **Vehicle** dataset and the first 3 rows and the last 6 rows of it **(2 marks)**.
4. Show the average Kms_Driven for each type of car (Car_Name) in the dataset **(1 mark).**
5. What is the average Selling_Price of the cars in each year? **(1 mark)**
6. Show the unique combinations of Car_Name, Fuel_Type, Seller_Type, and Transmission in the **Vehicle** dataset. **(2 marks)**
7. What are the different combinations of Car_Name, Fuel_Type, Seller_Type, and Transmission in the **Vehicle** dataset, and how many times does it occur? (Display all such in both ascending and descending orders) **(2 marks)**

Part 2: Data Pre-Processing (28 marks)

8. Find if there are any missing values in the **Vehicle** dataset **(1 mark)**.
9. Find which columns contain missing values in the **vehicles** dataset. What are the total missing values for each column? **(3 marks)**
10. Replace the missing values in the dataset with the most repeated value of that field. Check if the missing values were replaced successfully **(4 marks)**.
11. Find if the dataset has duplicate rows. Remove them, if exist **(4 marks).**
12. Replace the values of the following attributes:

    a Fuel_Type: “Petrol”: 0, “Diesel”: 1, “CNG”: 2

    b Seller_Type: “Dealer”: 0, “Individual”: 1

    c Transmission: “Manual”: 0, “Automatic”: 1

    Show the conversion output of the specific attribute **(6 marks)**.

13. Add a new field called **‘Age’**, and input the values by using the field **Year.** Show the output **(4 marks).**
14. Create a new dataset by selecting only the columns “Car_name”, “Selling_Price”, “Present_Price”, and “Kms_Drive”. Show the output of the new dataset **(4 marks).**
15. Shuffle the rows of the Vehicle dataset randomly and show the output **(2 marks)**.

Part 3: Data Visualization (60 marks)

16. Import the **Vehicle** dataset. Create a scatter plot of the Selling_Price Vs Present_Price. Colour code the points based on the Transmission **(5 marks)**.
    a. Add labels, title and colour to the plot. The colour should be red for Transmission type ‘0’ and blue for ‘1’.
    b. Add open triangles to the plot.
    c. What do you understand from the output (5 marks)?

17. Create a box plot of the Selling_Price Vs Transmission and Fuel_Type **(5 marks)**.
18. Create a scatter plot of the Selling_Price Vs Kms_Driven, and use k-means clustering to cluster the points into 4 clusters. Colour-code based on the cluster they belong to **(10 marks)**.

19. Create a scatter plot of the Selling_Price Vs Present_Price, and use hierarchical clustering to cluster the points into 3 clusters? Colour-code the points based on the cluster they belong to **(10 marks)**.

20. Add a new field called **'Age',** and calculate it using the field **'Year'.** Create a barplot for the following fields of the dataset: **(10 marks)**

   a. 'Age', 'Year', 'Transmission', 'Seller_Type', 'Fuel_Type' and 'Owner'
   b. Add labels, titles, and colours to the plot.

21. Create a correlation plot of the whole dataset variables and explain the output. Do not forget to convert some of the variable's datatype if required and possible **(10 marks)**.

22. Create a scatter plot of the Selling_Price Vs Kms_Driven, and use DBSCAN clustering to cluster the points into 3 clusters. Color-code based on the cluster they belong to. Add a legend to the plot **(5 marks)**.