



PRIFYSGOL  
**BANGOR**  
UNIVERSITY

**YSGOL CYFRIFIADUREG A PHEIRIANNEG**  
**SCHOOL OF COMPUTER SCIENCE AND ENGINEERING**

***ICE-4006 Data Science***

**Semester 1    2025/26**

**Case Study**

**Submission Deadline:**

**21 Dec. 2025**

**40% of ICE-4006**

**Total marks: 100 points**

## Work Brief:

The dataset ([bike.csv](#)) represents the number of shared bike rentals in two cities during specific hours under different influencing factors. We will use this dataset to build a regression prediction model that predicts the number of bike rentals during a given hour based on certain attributes in the data.

All values in the table are integers, and the table structure is as follows:

- *id*: Record ID, no other meaning
- *city*: City code, 0 = city-A, 1 = city-B
- *hour*: Hour of the day
- *is\_workday*: Whether it is a workday, 0 = No, 1 = Yes
- *temp\_air*: Air temperature in Celsius
- *temp\_body*: Feels-like temperature in Celsius
- *weather*: Weather code, 1 = Sunny, 2 = Cloudy/Overcast, 3 = Rainy/Snowy
- *wind*: Wind level, larger values indicate higher wind speed
- *y*: Number of shared bikes rented during that hour

## Please complete the task following the steps below:

### 1. Data Acquisition:

The dataset has already been saved into the file ([bike.csv](#)). Please use the [pandas](#) library to read this file. Hint: The original dataset contains 10,000 records.

### 2. Data Preprocessing I:

The *id* attribute does not help in building the regression prediction model, so please remove this column. Hint: Use [.drop\(..\)](#) method.

### 3. Data Preprocessing II:

We will not consider the effect of different cities on bike rentals. Please filter out all data for *city-B* (where *city* = 1), and then remove the *city* column. Hint: At this stage, 4,998 records remain.

### 4. Data Preprocessing III:

To simplify the data, please transform the *hour* column as follows:

- From 06:00 to 18:00, set the value to 1
- From 19:00 to 05:00 (next day), set the value to 0

### 5. Data Preprocessing IV:

The *y* column represents the number of bike rentals and will serve as our prediction target (*label*). Please extract this column and convert it into a [NumPy](#) column vector, then remove the original *y* column from the dataset. Hint: Use [.to\\_numpy\(\)](#) method.

### 6. Data Preprocessing V:

Please convert the dataset from a [DataFrame](#) into a [NumPy](#) array to

facilitate subsequent operations.

7. Dataset Splitting:

Please split the dataset into a training set and a testing set in an **8:2** ratio.

Hint: Use `train_test_split(...)` method.

8. Data Preprocessing VI:

Please normalize training set data, training set labels, testing set data, and testing set labels separately. Hint: Use `MinMaxScaler()` and `.fit_transform()` method.

9. Model Construction:

First, build a **Linear Regression model (*multivariate linear function*)**, and then train the model using the training set. Hint: Use `LinearRegression()` method.

10. Model Testing:

Use the test set to evaluate the trained model. Hint: Use `.predict(...)` method with the testing set as input.

11. Model Evaluation:

Please use **Root Mean Squared Error (RMSE)** as the evaluation metric and use `print(...)` method to output the RMSE value.

Root Mean Squared Error (RMSE) is defined as:

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - x_i)^2}$$

Where  $y_i$  is the true value,  $x_i$  is the predicted value, and  $n$  is the number of samples.

Hint: RMSE can be obtained directly by taking the square root of the MSE, and MSE can be implemented by `mean_squared_error(...)` method.

Note: Please calculate the RMSE in the scaled [0–1] range, not in real bike rental counts. So do not convert both predictions and ground truth back to the original scale.

**Marking Scheme Guidance:**

Step No.	Marks Available
1	5
2 ~ 10	10 for each step
11	5