# Marking Scheme (25 Marks)

1. **Data Preparation (12 marks)**
2. Load citi bike data and preprocess the data
   1. Download and unzip data correctly (2 marks)
   2. Combine all the csv files into a spark dataframe (1 mark)
   3. Aggregate the data into daily records (2 mark)
3. Load weather data and preprocess the data
   1. Load data from DBFS filestore (1 mark)
   2. Preprocess the missing value (1 mark)
4. Combine the data and prepare the data ready for modeling
5. Combine all the data into a joined table with trip\_count & other feature columns (2 mark)
6. Save the aggregated table into DBFS and read the table from DBFS before running the rest of steps (2 marks)
7. Filter the data into train and test (1 mark)
   * It’s ok if they download train & test separately, so no need to split/filter.
8. **Modeling Building (including data encoding) (5 marks)**
9. Data Preparation (2 mark)
10. Build the 3 steps ML Pipeline (2 mark)
11. Fit the Pipeline on the training data (1 mark)
12. **Model Evaluation: (8 marks)**
13. Correct Codes (1 mark)
14. A valid evaluation with reasonable performance (3 mark)
    * MAE <15000
15. Better performance (4 marks)

* MAE <12000

**FAQ**

For students using wrong features (i.e. the features cannot be obtained or forecasted in advance), we have deducted 4 marks from section 3c.