Assignment	**2021 - 2022**
Title	**Bike Sharing Dataset**
Training data file	**bike_##. csv**
Test data file	**bike_test.csv**

Background information

Bike sharing systems are new generation of traditional bike rentals where the whole process from membership, rental and return back has become automatic. Through these systems, user is able to easily rent a bike from a particular position and return it back at another position. Currently, there are about over 500 bike-sharing programs around the world which is composed of over 500 thousands bicycles. Today, there exists great interest in these systems due to their important role in traffic, environmental and health issues.

Apart from interesting real world applications of bike sharing systems, the characteristics of data being generated by these systems make them attractive for the research. Opposed to other transport services such as bus or subway, the duration of travel, departure and arrival position is explicitly recorded in these systems. This feature turns bike sharing system into a virtual sensor network that can be used for sensing mobility in the city. Hence, it is expected that most of the important events in the city could be detected via monitoring these data.

Aim: Understanding what influences bike rental count hourly and also predict it in order to satisfy demand.

The data

Bike-sharing rental process is highly correlated to the environmental and seasonal settings. For instance, weather conditions, precipitation, day of week, season, hour of the day, etc. can affect the rental behaviors. The core data set is related to the two-year historical log corresponding to years 2011 and 2012 from Capital Bikeshare system, Washington D.C., USA which is publicly available in http://capitalbikeshare.com/system-data. We aggregated the data on hourly basis and then extracted and added the corresponding weather and seasonal information. Weather information are extracted from http://www.freemeteo.com.

Dataset characteristics

All datasets are random subsamples of 1500 hour occasions and have the following fields:

- **instant**: record index
- **dteday**: date
- **season** : season (1:springer, 2:summer, 3:fall, 4:winter)
- **yr** : year (0: 2011, 1:2012)
- **mnth** : month (1 to 12)
- **holiday** : weather day is holiday or not
- **weekday** : day of the week
- **workingday** : if day is neither weekend nor holiday is 1, otherwise is 0.
- **weathersit** : Possible outcomes
    - 1: Clear, Few clouds, Partly cloudy, Partly cloudy
    - 2: Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist
    - 3: Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds
    - 4: Heavy Rain + Ice Pallets + Thunderstorm + Mist, Snow + Fog
- **temp** : Normalized temperature in Celsius. The values are divided to 41 (max)
- **atemp**: Normalized feeling temperature in Celsius. The values are divided to 50 (max)
- **hum**: Normalized humidity. The values are divided to 100 (max)
- **windspeed**: Normalized wind speed. The values are divided to 67 (max)
- **casual**: count of casual users
- **registered**: count of registered users
- **cnt**: count of total rental bikes including both casual and registered (response)

Source of the data

- Fanaee-T, Hadi, and Gama, Joao, "Event labeling combining ensemble detectors and background knowledge", *Progress in Artificial Intelligence* (2013): pp. 1-15, Springer Berlin Heidelberg, doi:10.1007/s13748-013-0040-3.

The assignment

Each student will receive a random sub-sample of 1500 observations to use it for training their model and for inference. All students will use a common evaluation/test dataset of 500 observations. Please respond to the following tasks.

1) You should first perform some descriptive data analysis and visualization. Visualizing the data should give you some insight into certain particularities of this dataset. Pairwise comparisons will help you also learn about the association implied by the data.

2) The main aim is to identify the best model for predicting the number of bike rentals per hour (variable ***cnt***).
   a) Implement Lasso in order to select the covariates of your model.
   b) Select the appropriate features (after implementing lasso) using stepwise methods in order to select your final model. Be careful, your model should not be over-parameterized.
3) Check the assumptions of the model and revise your procedure.
4) Interpret the parameters and the predicting performance of the final model.
5) Use the test dataset to assess the out-of-sample predictive ability and compare the models selected in Q2. Also include the full and the null models in your comparison.
6) Describe the typical profile of a day for each season (autumn, winter, spring, summer).
7) Write a report summarizing your results (see attached directions for this).