

Mini Project: Setting up and Go

The learning outcomes of this mini project are:

- A. Understand the important role of data preprocessing and Exploratory Data Analysis (**EDA**) in data mining;
- B. Set up the **website** for your team project;
- C. Learning basics of setting up and using **R and Python** as tools of data mining;
- D. Review following concepts learned in the class through a case study: data preprocessing, linear regression, prediction problem, statistical inferences, and feature selection.

Team Project Website: You should (1) sign up as a team at Canvas; (2) construct a simple website for your team project; (3) enter the URL of the website from student entry.

R learning: R is very useful in Exploratory Data Analysis (EDA) which plays important role in data mining. Each student is required to study the following and ***produce a unique EDA using R and capture your process using RMD for movie review data sets given.***

- Basics of R (download R and R Studio, learning from examples and online resources)
- Using R for linear regression, and EDA with statistics
- The demo code on this case study given in the class
- Writing your own case study using the same movie review data sets but remember to make a different EDA + prediction experiment. This is a hands-on experience using R and RMD. Enjoy.

Python learning: Python has many powerful data mining libraries which can help us to carry out data preprocessing and data mining tasks. Each of you is required to do the following:

- Set up a Python programming environment
- Solve a simple data mining problem with Python from data preprocessing to classification model building. You may choose a data set and problem of your own, or use sample data sets available.

Submission:

Using Canvas, submit following files before the due date:

- (1) Your Movie prediction and EDA case study results – two files (a) RMD file, (b) *HTML file* generated by your RMD on the given movie review data sets.
- (2) A brief summary of your learning experience on Python including data mining problem description, computation steps, and results. Please also include at least two data mining learning resources/tools you enjoy and would like to recommend to the class. A pdf file for summary and 1-2 supporting files.