CE/CZ4073 : Data Science for Business

2017-2018, Semester 2 | Nanyang Technological University, Singapore

Assignment 1

Posted on 5 February 2018 · Clarify by 13 February 2018 · Submit by 19 February 2018

Submission : Submit a single `R` file for the assignment – `assign1_FullName.R` – containing all the steps and comments in connection with the exploratory data analysis and model building for both the problems, where `FullName` is your full name, as per the official records at NTU.

Your submission should be emailed to `sg.sourav@ntu.edu.sg` by midnight of 19 February 2018 from your official NTU Email Address (do not use your personal email address). Late submissions will be penalized with lower grades, and submissions after 23 February 2018 will not be graded.

Problem 1 [10 points]

Background : Consider the attached dataset `assign1_WineData.csv`, which has 12 variables:

```
"FixedAcidity"        "VolatileAcidity"     "CitricAcid"          "ResidualSugar"
"Chlorides"           "FreeSulphurDioxide"  "TotalSulphurDioxide" "Density"
"pH"                  "Sulphates"           "Alcohol"             "Quality"
```

The target is to fit an *optimal* linear regression model to predict the `"Quality"` of the wine.

Task : Import the dataset, perform exploratory data analysis on the variables, and construct the *best model*, in your opinion, to predict the response variable, that is, the `"Quality"` of the wine, in case of the given dataset `assign1_WineData.csv`. Briefly comment (within the code) on your observations and on the choices you make in the process of building the *best model*.

Problem 2 [10 points]

Background : Consider the attached dataset `assign1_CarData.csv`, which has 6 variables:

```
"cylinders"    "displacement"    "horsepower"    "weight"    "acceleration"    "mpg"
```

The target is to fit an *optimal* linear regression model to predict fuel efficiency `"mpg"` of the car.

Task : Import the dataset, perform exploratory data analysis on the variables, and construct the *best model*, in your opinion, to predict the response variable, that is, the `"mpg"` of the car, in case of the given dataset `assign1_CarData.csv`. Briefly comment (within the code) on your observations and on the choices you make in the process of building the *best model*.

This is an individual assignment. Properly acknowledge every source of information that you refer to, including discussions with your fellow students, if any. Verbatim copy from any source is strongly discouraged, and plagiarism will be heavily penalized. It is strongly recommended that you write the codes completely on your own. Feel free to write the codes in Python if you want.