MAT 9102 - Probability and Statistical Inference
Assignment – III

24/11/2020

Submission guidelines:

- You will need to upload only one single R markdown **(.Rmd)** file.
- File name of your RMD file must be **regnumber_assignemntnumber**
- Do not upload given datasets (if any).
- Make use of R built in datasets (if mentioned in the question). If you have considered external dataset instead of R built in, upload the dataset without zipping it.
- Please use the following statement while installing any package. *if(!require(packageName))install.packages("packageName")*

General Instructions:

- Read the questions carefully and answer all parts to secure full marks.
- Post any queries in public channel or send a personal message
- Do not ask for direct solutions. This is a part of your assessment.
- Assignment will be penalized if you miss any of the submission guidelines.

PCA

1. What do the eigenvectors of the covariance matrix give us? [2 marks]
2. When can we decide to compress the data in PCA process? Explain the effects if any. [2 marks]
3. Read the glass identification data provided. Apply PCA algorithm to reduce the dimensions. Analyze your findings. [5 marks]

Note: Read glass dataset. Kindly save the read dataset in a variable called "glass". Do not change the variable name. For ex: glass <- read.csv("glassidentification.csv").

Difference

1. Are there any differences between patients having different chest pain to the angiographic disease status? Report your findings. [Hint: Consider variables ChestPain and AHD] [5 marks]

2. Is there any difference between cholesterol level and angiographic disease status? Report your findings. [Hint: Consider variables Chol and AHD] [5 marks]

Note: Read heartdisease dataset for Q1. and Q2. Kindly save the read dataset in a variable called "heartdisease". Do not change the variable name. For ex: heartdisease <- read.csv("heartdisease.csv").

3. Are there any differences between the free sulfur dioxide and quality of the wine? Report your findings. [Hint: Consider variables free_sulfur_dioxide and quality] [5 marks]

Note: Read winequality dataset. Kindly save the read dataset in a variable called "wine". Do not change the variable name. For ex: wine <- read.csv("winequality-red.csv").

Predictive statistics

1. Model the relationship between humidity and total rented bikes. How good is the model? [Hint: Consider the variables hum and cnt] [3 marks]

2. Include a dummy variable (working day) to the model and compare the relationship. Report your findings. [3 marks]

   [Note: Read bikesharing dataset. Kindly save the read dataset in a variable called "bike". Do not change the variable name. For ex: bike <- read.csv("bikesharing.csv").]