MATH 9102 - Probability and Statistical Inference

Assignment 3

Due 6pm, Sunday 14th April, 2024

15 marks

Submission guidelines:

- You will need to upload only **single** R markdown (.Rmd) file.
- File name of your Rmd file must be **YourName_StudentID_CA3**
- Do not upload given datasets (or use zip).
- 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.
- Use the following statement if installing any package: if(!require(packageName))install.packages(”packageName”)

General Instructions:

- Read the questions carefully and answer all parts to secure full marks.
- Do not ask for direct solutions. This is part of your assessment.
- Assignment will be penalized if you miss any of the submission guidelines.
- Please complete assignment individually and avoid plagiarism as it will lead to penalties and negatively affect your overall grade.
- Including comments/markup in your code to explain what you did and provide answers to all questions.

1. **PCA**

   a) What do the eigenvectors of the covariance matrix give us? **[1 mark]**

   b) When can we decide to compress the data in PCA process? Explain the effects if any. **[1 mark]**

   c) Read the glass identification data provided. Apply PCA algorithm to reduce the dimensions. Analyze your findings. **[2.5 marks]**

   Reading the glass dataset: Save the read dataset in a variable called “glass”. Do not change the variable name.
   For example: glass <- read.csv(”glassidentification.csv”).

2. **Difference**

   a) Are there any differences between patients having different chest pain to the angiographic disease status? Report your findings. [Hint: Consider variables ChestPain and AHD] **[2.5 marks]**

   b) Is there any difference between cholesterol level and angiographic disease status? Report your findings. [Hint: Consider variables Chol and AHD] **[2.5 marks]**

   Reading the heartdisease dataset for Q1. and Q2: Save the read dataset in a variable called “heartdisease”. Do not change the variable name.
   For example: heartdisease <- read.csv(”heartdisease.csv”).

   c) 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] **[2.5 marks]**

   Reading the winequality dataset: Save the read dataset in a variable called “wine”. Do not change the variable name.
   For example: wine <- read.csv(”winequality-red.csv”).

3. **Predictive statistics**

   a) Model the relationship between humidity and total rented bikes. How good is the model? [Hint: Consider the variables hum and cnt] **[1.5 marks]**

   b) Include a dummy variable (working day) to the model and compare the relationship. Report your findings. **[1.5 marks]**

   Reading the bikesharing dataset: Save the read dataset in a variable called “bike”. Do not change the variable name.
   For example: bike <- read.csv(”bikesharing.csv”).]