

Homework 1

BUAN 6356

Read the instructions below before you start your analysis.

1. Create an R Markdown document to prepare your answers. You should upload **two (2)** files on eLearning: (i) an **.RMD** file; and (ii) a **.PDF** file that is generated using “knit” in the .RMD file. Both of these files should contain the required R code, R tables and charts, and all the required explanations and answers to the questions in the homework.
2. Include your group number in the name of the file you upload. For example, if your name is group number is 7, then name the file BUAN6356_HW1_Group7.
3. Write your answers outside of the code chunks in the markdown file.
4. **DO NOT** use an absolute directory path. I should be able to “knit” your R Markdown document to an .html/.pdf document without trying to find the input data in another directory. Test the “knit” process before uploading files on eLearning.
5. **DO NOT** change the dataset name before importing it into R. If you rename the dataset or any variable(s), use your R script to do that.
6. Label the charts and tables appropriately. A reader should be able to figure out what information a chart is providing by looking at the chart title and its labels.
7. Any assignment submitted after the deadline will be considered late and will not be graded.

Homework 1

Use “iris” data from the datasets package in RStudio. For answering questions below, use the entire data set (do not create training/test sets).

1. Generate a correlation plot using the data. Which variables have the highest correlation coefficient?
2. Describe two drawbacks of correlation coefficient. Feel free to use hypothetical numbers to explain your answer.
3. Calculate average (mean) values of the numeric variables in the data using data.table package. Which variable has the highest mean?
4. Create a scatterplot showing the relationship between Sepal.Length and Sepal.Width variable, using ggplot2 package. Color code the points using Species variable.
5. Create a scatterplot showing the relationship between Petal.Length and Petal.Width variables, using ggplot2 package. Color code the points using Species variable.
6. Which combination of variables (used in Question 4 or Question 5), creates better separation among records of different Species. Explain your answer.

Use the *Confusion Matrix* below to answer Questions 7-10. The confusion matrix was generated by running an algorithm that only included Setosa and Versicolor species.

Predicted	Actual	
	Setosa	Versicolor
Setosa	45	2
Versicolor	5	48

7. What is the overall accuracy of the model? How does it compare with the No Information Rate (NIR)? Explain your answer.
8. Calculate the sensitivity of the model, assuming the class of interest is Setosa. Show the calculations.
9. Calculate the specificity of the model, assuming the class of interest is Setosa. Show the calculations.
10. Calculate the precision of the model, assuming the class of interest is Setosa. Show the calculations.