BTMA 531 Assignment 1 w/ Answers

Due January 31, 2024 by noon on D2L

Instructions

- You should create a single R script (or R Markdown if you wish) called {firstname}_{lastname}_Asgn1.r, which has the required code for all parts of the assignment. For Q9 (Power BI), please submit the Power BI file (.pbix file) which includes the requested data manipulation and dashboard.
- Make sure to use commenting (#) so that your R script file can be understood by someone else. You do not need to comment on what you are trying to do in each line, but it should be clear where the answer for each question is. Provide the written answers in comments (#) within the R script file.
- Make sure that your R script is executable from top to bottom on another computer. Make sure all requirements of questions are done using R (e.g. do not use calculator, excel, . . . to calculate things).
- The assignment submission should be done through D2L dropbox. Upload the files to the assigned dropbox folder in D2L.
- The purpose of the assignments is to help you learn through practice. You may want to use R help or search online or use other tools to get help for answers. However, note that this is an individual assignment. The work you submit should be 100% yours. Do not copy, share, or ask for files, chunks of code, or answers with other students. Refer to the course outline for some examples of what to do and what not to do, and to Code of Student Conduct for more information on cheating and plagiarism. If you are note sure about a behavior, please ask.
- The focus is on the response to the question, and not on the approach you use to get there. However, unless otherwise stated, you should use only code to get to the response (that is, you cannot use a calculator or excel for example to do some intermediate calculations, everything needs to be done in R).
- You are allowed to use any online resource to find answers or chunks of code. However, make sure that you reference the source in your script using commenting (#). This is generally a good practice in coding, and for the purpose of this course, removes any confusion when several students use the same online source and therefore have similar codes in their scripts.

Questions

1 [15]

a) [2.5] Create a sequence of numbers starting from 100 and going up to 10,000 by increments of 25. Assign this sequence to an object called "object000". What is the answer to following command?

```r
object000[29]
```

b) [2.5] List out all objects in the environment. Remove "object000" from the environment.

```r
ls()
```

```
## character(0)
```

```r
rm(object000)
```

```
## Warning in rm(object000): object 'object000' not found
```

c) [2.5] What do you need to do to use the functions in “tidytext” package using code only in your current environment? Please provide the answer as a comment rather than line of code.

d) [5] Create a randomly generated matrix called “matrix1” with two variables (columns) and 25 observations (rows). The observations from the first variable are normally distributed with $\mu = 10$ and $\delta = 3$ and the observations from the second variable are uniformly distributed in the range [1,20]. Derive the standard deviation of each of the columns. Are the results what you expected?

e) [2.5] Plot the two variables in “matrix1” as a scatter plot (variable 2 on x axis and variable 1 on y axis). What is the correlation between the two columns and what does that mean?

2 [10]

a) [5] Load the package “ISLR” which includes a dataset called “Carseats”. How many variables does this dataset have? How many of these variables are qualitative/categorical and how many are quantitative (answer programmatically)?

b) [5] What would be the variable of interest that you would want to predict in this dataset? What would be three relevant questions that you could ask and answer using this dataset? Explain.

3 [5] In the dataset “Carseats” within the “ISLR” package, calculate the average Advertising of car seats for which shelf location is “Medium”. Calculate the average Advertising of car seats in the shelf location “Bad”.

4 [10] From the dataset “Carseats” within the “ISLR” package:

a) [2.5] Graph a scatter plot of carseat Advertising and Sales. Properly name the x and y axes.

b) [2.5] Add a linear estimate of the relationship to the graph.

c) [2.5] Find the correlation between these two variables. Explain what this correlation implies.

d) [2.5] Could one state that price causes sales to increase/decreases solely based on the above analysis? Why? If yes, explain why; if no, explain what else is needed to make such a statement?

5 [15] From the dataset “Carseats” within the ISLR package:

a) [5] Create frequency tables for “ShelveLoc” and “Urban” variables. How many entries are there for each shelf location type and each location type (Urban/Not Urban), respectively?

b) [5] Create two plots: one bar plot of the “Urban” variable, and one bar plot of the “ShelveLoc” variable, and put them side by side (both plots in one figure).

c) [5] create a histogram of the variable “CompPrice” (one single graph filled by the figure).

6 [10]

a) [2.5] Import the attached “ad.csv” file to R as an object called “AdData”, once using read.csv() and once using fread(). Add libraries if needed.

b) [2.5] Create a simple linear regression model for sales using TV advertising variable. Does TV advertising have any impact on sales? Predict the sales for when TV advertising is 145.

c) [2.5] Create a multiple linear regression model for sales using TV, radio, and newspaper advertising variables.

d) [2.5] Draw the normality plot of residuals for the multiple linear regression. What does it say?

7 [10] From dataset “Wage” within the “ISLR” package:

a) [5] Create a multiple linear regression model (with categorical variables) that uses “year”, “age”, “health_ins”, and “health” to predict “wage”.
b) [5] Predict wage of a person in year 2006 who is 40 years old without health insurance and very good health.

8 [10] Considering the Carseats dataset within the “ISLR” package:

a) [2] From a business point of view, which variable is the desired outcome?

b) [4] What aspects of the business does this data capture, and what aspects are missing in your opinion?

c) [4] What questions can you think of to ask from this dataset, the answer to which may be helpful to the business?

9 [15] Use the attached database “BookstoreDemo.accdb” for this problem. This is the simplified database for the school library.

a) [3] When loading the data, perform the following transformation: Split the “Term” column in the course table (“CourseTBL”) to two columns: 1. a column named “Term” which has the term in that year (W, F, or P) and 2. a column named “Year” which has the year for the term.

Then create a dashboard containing the following visuals in a single page:

b) [3] A “potential sales” matrix with “Publisher” as rows and “InstLastName” on columns showing the enrollment for each instructor-publisher pair.

c) [3] A line chart showing the trend of “Enrollment” as well as book “OrderSize” over the years.

d) [3] A stacked column chart showing the “Enrollment” for each “InstLastName” which is stacked by (legend) the instructors’ “CourseName”.

e) [3] Another visual which reveals something interesting/important to the library managers. Explain what it shows and why it is interesting/important.