MSDS 6306: Introduction to Data Science

Case Study 2 (Group Project)

Due: No late acceptance

6:30pm, April 24 (Monday) for both Monday class (Section: 404) and Wednesday class (Section: 405):

Deliverables by answering the 3 questions below:

1. **You need to create a repository for Case Study 2 GitHub. Submit a word file that has your group name with group members and the link to the file in GitHub via the submission link for Case Study 2 in 2DS.**
    a. It also includes the codes of question 1 and the screenshot of the result
2. **Deliverable for Questions 2 and 3: Markdown file uploaded to the GitHub containing the following, which is what you have done at Case Study 1**
    a. **Introduction to the project.** The introduction should not start with “For my project II …”. The introduction needs to be written as if you are presenting the work to someone who has given you the data to analyze and wants to understand the result. In other words, pretend it’s not a case study for a course. Pretend it’s a presentation for a client.
    b. **Code** for downloading, tidying, and merging data in a R Markdown file. The code should be in a make file style, meaning that the source RMD document pulls in separate files for importing data, cleaning the data, and data analysis.
    c. **Brief explanations** of the purpose of the code. The explanations should appear as a sentence or two before or after the code chunk. Even though you will not be hiding the code chunks (so that I can see the code), you need to pretend that the client can’t see them.
    d. **Code** to answer the questions 2 and 3 below (plus the answers) in the same R Markdown file.
    e. **Clear answers** to the questions. Just the code to answer the questions is not enough, even if the code is correct and gives the correct answer. You must state the answer in a complete sentence outside the code chunk.
    f. **Conclusion** to the project. Summarize your findings from this exercise.
    g. **Important:** The file must be readable in GitHub – **20** points off if I have to download the file to read it! In other words, don’t forget to keep the **md** file!!

Question 1 (15 points)

Create the X matrix and print it from SAS, R, and Python.

$$X = \begin{pmatrix} 4 & 5 & 1 & 2 \\ 1 & 0 & 3 & 5 \\ 2 & 1 & 8 & 2 \end{pmatrix}$$

SAS code **(5 points)**

R code **(5 points)**

Python Code **(5 points)**

Question 2 (30 points)

The built-in data set called ***Orange*** in R is about the growth of orange trees. The Orange data frame has 3 columns of records of the growth of orange trees.

Variable description

Tree: an ordered factor indicating the tree on which the measurement is made. The ordering is according to increasing maximum diameter.

age: a numeric vector giving the age of the tree (days since 1968/12/31)

circumference: a numeric vector of trunk circumferences (mm). This is probably "circumference at breast height", a standard measurement in forestry.

a) Calculate the mean and the median of the trunk circumferences for different size of the trees. (Tree)
b) Make a scatter plot of the trunk circumferences against the age of the tree. Use different plotting symbols for different size of trees.
c) Display the trunk circumferences on a comparative boxplot against tree. Be sure you order the boxplots in the increasing order of maximum diameter.

Question 3 (55 points)

Download “Temp” data set at box.com

(i) Find the difference between the maximum and the minimum monthly average temperatures for each country and report/visualize top 20 countries with the maximum differences for the period **since 1900**.

(ii) Select a subset of data called “UStemp” where US land temperatures from 01/01/1990 in Temp data. Use UStemp dataset to answer the followings.

a) Create a new column to display the monthly average land temperatures in Fahrenheit **(°F)**.
b) Calculate average land temperature by year and plot it. The original file has the average land temperature by month.
c) Calculate the one year difference of average land temperature by year and provide the maximum difference (value) with corresponding two years.

(for example, year 2000: add all 12 monthly averages and divide by 12 to get average temperature in 2000. You can do the same thing for all the available years. Then you can calculate the one year difference as 1991-1990, 1992-1991, etc)

(iii) Download “CityTemp” data set at box.com. Find the difference between the maximum and the minimum temperatures for each major city and report/visualize top 20 cities with maximum differences for the period since 1900.

(iv) Compare the two graphs in (i) and (iii) and comment it.