Homework 1

(Advanced) Data Mining: Algorithms and Applications-Winter 2022

Due on Jan 25, 11.59PM

Datasets for this homework can be found at the link below and feel free to add this folder to your Google Drive:
`https://drive.google.com/drive/folders/1xc2T1q1SAl-KpuxWR0bz7RZSnSXM2ybo?usp=sharing`

1. Use "Su_raw_matrix.txt" for the following questions (30 points).
   (a) Use `read.delim` function to read `Su_raw_matrix.txt` into a variable called `su`. (Notice that `su` has become a data frame now)
   (b) Use `mean` and `sd` functions to find mean and standard deviation of `Liver_2.CEL` column.
   (c) Use `colMeans` and `colSums` functions to get the average and total values of each column.
2. Use `rnorm(n, mean = 0, sd = 1)` function in R to generate 10000 numbers for the following `(mean, sigma)` pairs and plot histogram for each, meaning you need to change the function parameter accordingly. Then comment on how these histograms are different from each other and state the reason. (20 points)
   (a) `mean=0, sigma=0.2`
   (b) `mean=0, sigma=0.5`

   Please save your figures as image from RStudio. (Hint: to see the difference in plots you may need to set the `xlim` parameter in plot function to `c(-5,5)`)
3. Perform the steps below with "dat" dataframe which is just a sample data for you to observe how each plot function (3b through 3e) works. Notice that you need to have `ggplot2` library installed on your system. Please refer slides how to install and import a library. Installation is done only once, but you need to import the library every time you need it by saying `library(ggplot2)`. Then Run the following commands and observe how the plots are generated. (40 points)
   (a)
   ```
   dat <- data.frame(cond = factor(rep(c("A","B"), each=200)),
   rating = c(rnorm(200),rnorm(200, mean=.8)))
   ```
   (b)
   ```
   # Overlaid histograms
   ggplot(dat, aes(x=rating, fill=cond)) +
   geom_histogram(binwidth=.5, alpha=.5, position="identity")
   ```
   (c)
   ```
   # Interleaved histograms
   ggplot(dat, aes(x=rating, fill=cond)) + geom_histogram(binwidth=.5, position="dodge")
   ```
   (d)
   ```
   # Density plots
   ggplot(dat, aes(x=rating, colour=cond)) + geom_density()
   ```
   (e)
   ```
   # Density plots with semitransparent fill
   ggplot(dat, aes(x=rating, fill=cond)) + geom_density(alpha=.3)
   ```
   (f) Read "diabetes_train.csv" into a variable called `diabetes` and apply the same functions 3b through 3e for the `mass` attribute of `diabetes` and save the images. (Hint: instead of `cond` above, use the `class` attribute to color your groups. When you have fill option, your plots should show same type of chart for both groups in different colors on the same figure. Keep in mind that `diabetes` and `dat` are both DataFrames)
4. By using `quantile()`, calculate $10^{th}, 30^{th}, 50^{th}, 60^{th}$ percentiles of skin attribute of diabetes data. (10 points)

Important

- Please put all images and their explanations in a single pdf file and submit it along with an R script which has all R commands that you used.
- Please leave comments in your R script
- Only submissions through Canvas will be accepted.
- If you have a GitHub page and want to submit an R notebook, you can give a link to your work on GitHub