STAT 6348
Applied Multivariate Analysis
Spring 2024 Project 1

This project is individual work. So do not consult with anybody in or out of class. You can ask me or TA questions. You are encouraged to google about R commands and packages.

Sign below on this page and attach with your project. Your project will not be graded without it.

This project is entirely my work. I have not discussed about this project with anybody in or out of class. I understand and have complied with the academic integrity policies written in the *Handbook of Operating Procedures* **of UT Dallas** https://policy.utdallas.edu/utdsp5003**.**

I understand that if any academic misconduct is suspected, it will be referred to the Office of Community Standards and Conduct https://conduct.utdallas.edu/.

YOUR NAME ______________________________

DATE ______________________________

YOUR SIGNATURE ______________________________

Directions:

- Arrange your report in the following order: (1) Answers, (2) Outputs, and (3) Code.
- Type your answers including comments. These should not exceed more than 3 pages.
- Attach only the part of output that has been asked for in the question.
- At the end, attach your R script (just the commands – not the whole output) with comments included to indicate the parts. You may use R Markdown if you prefer.
- Your plots should be properly labeled and in a presentable format (e.g., properly labeled axes and legend). Do not present very small plots where it is difficult to see the pattern.

Question 1

Consider the HCV data. The data set contains laboratory values of blood donors and Hepatitis C patients (including its progress - 'just' Hepatitis C, Fibrosis, Cirrhosis) and demographic values age and sex. The variables in the dataset are the following:

The laboratory data are the attributes 5-14.

1) X (Patient ID/No.)
2) Category (diagnosis): '0=Blood Donor', '1=Hepatitis', '2=Fibrosis', '3=Cirrhosis'
3) Age (in years)
4) Sex (f, m)
5) ALB
6) ALP
7) ALT
8) AST
9) BIL
10) CHE
11) CHOL
12) CREA
13) GGT
14) PROT

a) Calculate Mahalanobis distances of each observation vector (with laboratory variables only) to the center in terms of the covariance matrix. Report only the summary statistics of the distances. Which patients are nearest and furthest to the center and how do they compare? How would these results change if Euclidean distance was used instead.
b) What are the directions and half-lengths of the two longest axes of the ellipsoid?

c) For laboratory variables, calculate sample mean vector, covariance matrix, and correlation matrix. Create a correlation heatmap in which intensity of colors reflects the strength of correlation. Add a legend to the heatmap.

d) Make a scatterplot matrix of laboratory variables by labeling the points using different colors/symbols for different categories (add a legend). In the diagonal cells of this matrix, add QQ plot of each variable. Comment on the pairwise relationships between variables based on the scatterplots and heatmap. Which pairs of variables are useful in distinguishing between different categories and how?

e) For each laboratory variable, make side-by-side box plots for each category. Put the plots for all attribute variables in one page. Comment on the marginal distributions based on the different levels of the category variable.

f) Create a panel plot consisting of the following three plots:

   (i) Scatterplot of CHE vs. CHOL with different colors/symbols for males and females. Add marginal plots for CHE and CHOL. Also, add circles to the same plot based on the Age variable (larger circles for larger ages). Carefully choose symbol sizes and line widths to minimize the crowded look.

   (ii) Box plot of Age grouped by Category. Label outliers by the value of Sex.

   (iii) Histogram of Age for category 1 superimposed by histogram of Age for category 3 (choose colors carefully to ensure that both histograms are visible).

   Note: Divide the plotting area into three parts in such a manner that none of the plots get squeezed/distorted. Make efficient use of the plotting area in order to minimize white space. Comment on any pattern observed in the plots.

g) Make a 3D scatterplot of CHE, CHOL, and PROT. Use different colors to represent different categories. Make another version of the plot by adding vertical lines that connect the points to the floor. Comment on the relationship between the three variables and whether it varies by category.

h) For this and the remaining parts, consider a subset of variables Age, ALB, CHE, CHOL, and PROT. Check univariate normality assumption (using all plots and measures discussed in class). If normality appears to be violated, explore transformations that may help. For each variable, include only one transformation and the corresponding plot (after transformation) that appears to be the most helpful.

i) Retain the transformations found above and check the multivariate normality assumption and find univariate and multivariate outliers. Comment about the normality assumption based on previous and this part.

Helpful links:
https://www.rdocumentation.org/packages/car/versions/1.2-0/topics/scatterplot.matrix
https://www.statmethods.net/graphs/scatterplot.html

http://www.sthda.com/english/wiki/scatterplot3d-3d-graphics-r-software-and-data-visualization
https://r-charts.com/correlation/pairs/
https://www.rdocumentation.org/packages/stats/versions/3.6.2/topics/heatmap
https://r-graph-gallery.com/ggplot2-package.html