Assignment 1.1

Instructions – Final Submission Format

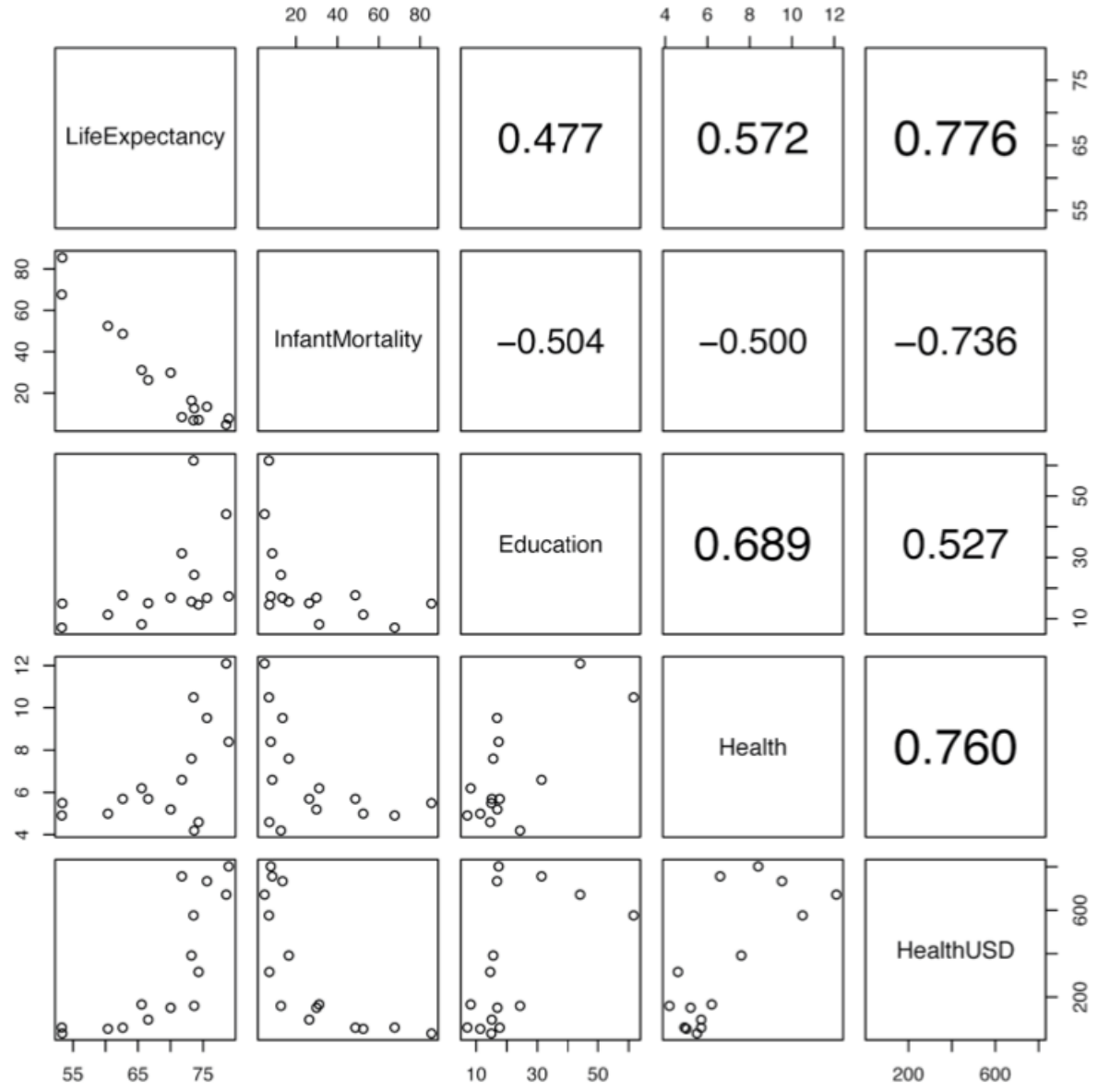
- Submit one assignment file that combines the answers to all questions:
- The file must be either a **Microsoft Word** or a **PDF** document file.
- **Do not** combine and submit files into a zipped compressed folder.
- Please use the following naming conventions for your **Word** or **PDF** assignment file:
  File Naming: **`LastName_FirstName_Assignment<Number>.pdf`**
  Example: **`Smith_James_Assignment1.pdf`**
- Answer all parts of a question in one place and answer questions in the order they appear in the assignment.
- For programming answers using **R**, it is recommended that the answers are written in R Markdown and 'knitted' to a Word/PDF file.
  - Do not print data frames in your submission, if you want to make a point about data you can use head(df) to print the first few rows.
  - Submit the code used to answer the questions in the assignment with your name on it, answers without code and appropriate results will not get full credit.
  - It is not a professional practice, but in case of difficulty, you can take screenshots of code and outputs and submit them in a Word/PDF file.
- Maximum number of pages should be 15. Any submissions that exceed 15 pages will **not be graded.**
- Use common sense to gauge the expectations of the answer to the number of points assigned to the question.

Questions

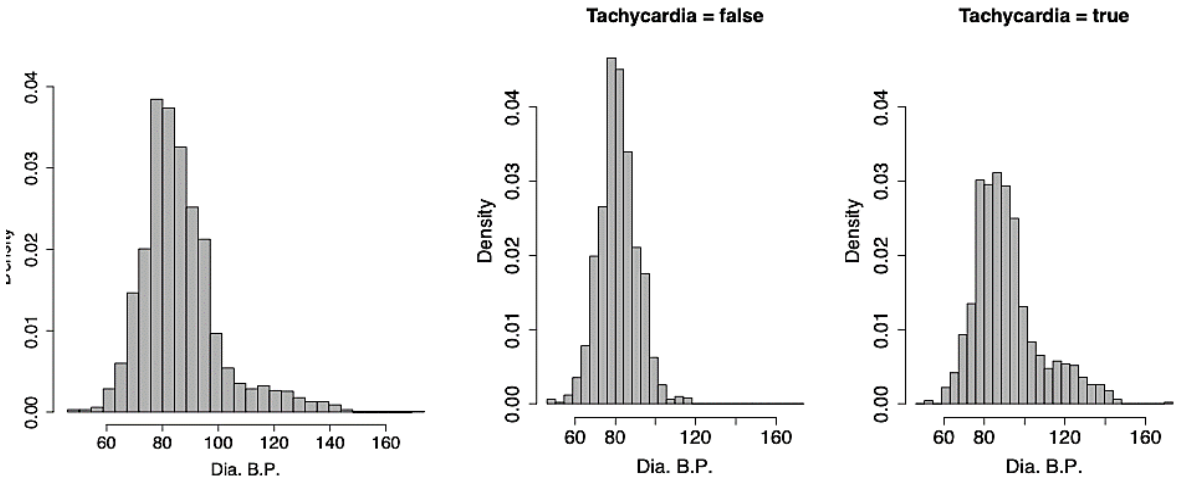
1. (30 points) The UC Irvine Machine Learning Repository contains a Glass Identification Data Set [https://archive.ics.uci.edu/ml/datasets/Glass+Identification]. The data consist of 214 glass samples labeled as one of seven class categories. There are nine predictors, including the refractive index and percentages of eight elements: Na, Mg, Al, Si, K, Ca, Ba, and Fe. The data can be accessed via:
   library(mlbench)
   data(Glass)
   a. Using visualizations, explore the predictor variables to understand their distributions as well as the relationships between predictors. Explore the relationship between predictors and response. (10 points)
   b. Are there any outliers in the data? Are any predictors skewed? (10 points)
   c. Are there any relevant transformations of one or more predictors that

might improve the classification model? Assume the model requires the predictors to have approximately symmetric distribution. Apply relevant transformations to the predictors and observe the changes to the distributions of predictors. (10 points)

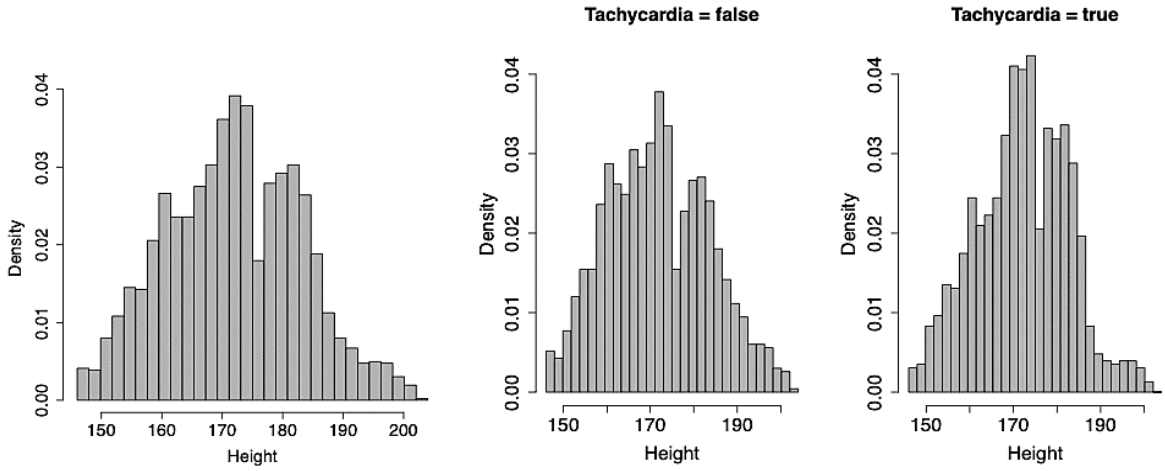
2. (20 points) The image below shows a scatter plot matrix of the continuous features of a dataset. Discuss the relationships between the features in the dataset that this scatter plot highlights. Make sure to discuss relationships between all pairs.


LifeExpectancy
InfantMortality
Education
Health
HealthUSD
0.477
0.572
0.776
−0.504
−0.500
−0.736
0.689
0.527
0.760

3. (10 points) Discuss the relationships between the variables shown in below visualizations:
    a. The visualization below illustrates the relationship between Diastolic BP and Tachycardia, left most plot has data where Tachycardia = true and false. (5 points)



    b. The visualization below illustrates the relationship between Height and Tachycardia, left most plot has data where Tachycardia = true and false. (5 points)



4. (30 points) Use the HCV Data Set at the UCI Machine Learning Repository (or download the "hcvdat0.csv" file in Blackboard) and **pick the numeric predictors (you can do this by excluding columns "X" , "Category", "Age" and "Sex")** to perform the following analysis in R:
    a. Are there any missing data in the predictors? Identify all the predictors with missing values (5 points)
    b. Summarize the missing data by each predictor. (5 points)
    c. Plot the histograms of predictors and visually identify predictors with skewed distributions. (5 points)
    d. Compute skewness using the skewness function from the e1071 package. Are the skewness values aligning with the visual interpretations from part c. (5 points)
    e. Apply box-cox transformations to the data and then recompute the skewness metrics and report the differences; does box-cox transformation help mitigate skewness? (5 points)
    f. Plot histograms of transformed predictors to observe changes to skewness visually. (5 points)