Special Topics: Data Analytics and Visualization in Healthcare
CSCI-GA.3033-096 (19635)
Lab assignment 3

Instructions: Find a healthcare dataset and perform the following exercises. The dataset should be different from the one(s) that you will use in your final project.

Exercise 1

- Perform a correlation analysis using (a) Tableau and (b) R or Python:
    - Choose two variables.
    - Which variable is the dependent variable?
    - Which variable is the independent variable?
    - Calculate the Pearson correlation coefficient with the independent and dependent variables.
    - Explain the correlation type (strong, moderate, or weak) and if it is a positive or negative correlation.

Exercise 2

- Create scatter plots with (a) Tableau and (b) R or Python:
    - Visualize the relationship between the independent variable and the dependent variable using scatterplots.
    - Color code the points for at least 2 different groups (for example, non-obese with gray and obese with brown).

Exercise 3

- Perform Linear Regression Analysis using R or Python:
    - Perform Simple Linear Regression between the dependent variable and the independent variable.
    - Obtain the best-fit line equation.
    - Visualize the regression line in the scatter plot.
- Perform Multiple Linear Regression Analysis using R or Python:
    - Perform Multiple Linear Regression with the dependent and the independent variables.
    - According to your multiple linear regression analysis, are your independent variables good predictors of your dependent variable?

Exercise 4

- Add a trend line in Tableau for the scatter plot created in Exercise 2.

Notes:

- This assignment is individual.
- Submit the following files:
    - The Tableau packaged workbook file (.twbx).

- The .R file or the .ipynb file
- Dataset file
- Write-up word/PDF file:
  - Describe the fields of your dataset.
  - Add a reference to your dataset.
  - Add your conclusions.

Due date: Thursday, November 3 at 11:59 PM.