Biostatistics 6640
Python and R in Data Science
Python Data Analysis Project: due December 12 by 11:59 PM

For this project, you will be analyzing the following dataset: `https://archive.ics.uci.edu/ml/datasets/Diabetes`. The data files, description of the data and references are all available on the website. You can download all files via a zip archive. Each file is a patient's data. There are 70 patients. This is a time series dataset on diabetic patients' glucose levels.

Using Jupyter Notebook and only Python code, please carry out the following analyses and answer the following questions.

1. Create one analytic dataset from the 70 files, creating a unique file ID column to keep track of each dataset in the full dataframe. Please use concise code. Print the first 5 lines and the data frame dimensions.

2. Table 1: create a table that summarizes the median and interquartile range (IQR: 25th percentile, 75th percentile) of glucose level by pre-breakfast, post-breakfast, pre-lunch, post-lunch, pre-supper and post-supper.

3. Table 2: Create a table that summarizes the codes 65-72. Assume that if the code was recorded, then it was an occurrence of the event (i.e., ignore 'value' for these codes). To summarize, create frequencies and proportions by approximate time of day for the following three times: 1) morning = 00:01 AM to 11:59 AM; 2) afternoon = 12:00-16:59 PM; 3) evening: 17:00 PM to 24:00. For example, if a patient had at least one recording of a 65 between 00:01 AM and 11:59 AM, this is coded as a 1 for the morning period. The denominator is 70 for each time window. Therefore, if 12 patients had a 65 in the first time window, then you would report 12 (17%) in that cell of the table.

4. Create side-by-side boxplots of insulin doses (the values of codes 33-35) grouped by the insulin type and time windows in question 3. If a patient has more than one of each insulin type in a window on the same day, then sum the total doses for each type in that window. Here is what the figure should look like: `https://python-graph-gallery.com/34-grouped-boxplot/`

5. Plot the time series of glucose values across the entire time series for all subjects (points). Fit a spline through the points and add this to the plotted points. Create

three versions of this plot: 1) Pre and Post breakfast (two lines, different color for pre and post); 2) Pre and Post lunch (two lines different colors); 3) Pre and Post supper (two lines different colors)

6. Create density plots for each set of pre/post meal glucose readings. There should be three density plots with two densities in each plot - one for pre and one for post. The three are 1) breakfast, 2) lunch, 3) supper. For an example, see: `https://python-graph-gallery.com/74-density-plot-of-several-variables/`.

Note: This project should be completed by all individuals in the class. You may work in groups, but if you do, each person needs to submit their own write-up and please include the names of the individuals with whom you worked.