CS24200: Homework 4

Due date: Friday October 11, 11:59 PM EST

Use R instead of python for this homework. You might need to install RStudio to complete the homework.

Make sure to annotate all graphs, i.e., your graphs should include a title, a legend, and labels on the axes. Make sure that all font sizes are legible. Include all the plots and report your observations in a pdf file 'Homework4.pdf'. Submit your code along with the pdf file in ***gradescope*** *whose link is on blackboard in* ***Contents*** *section).*

In this assignment you will use the data in **aneurysm_data.csv** for analysis.

1 Basic plots in R (12 pts)

1.1 **Scatter Plot**: Create a scatter plot showing the relationship between the BloodPressure and Age. Fit a line to the points using the `lowess` function and add the line to the plot.

1.2 **Box Plot**:

(a) Create a box plot for the Aneurysms_q1 (a discrete integer variable) vs. age (a continuous variable).

(b) Create a second box plot for Adneurysms_q2 vs. Age.

1.3 **Histogram**: Create a plot that contains two histograms, one for the Aneurysms_q3, and one for Aneurysms_q4. Make sure that you choose x and y ranges that ensure both histograms are fully visible. Color each histogram a different color to differentiate, and add a legend to the plot.

2 Using ggplot2 (12 pts)

2.1 Use the `ggplot2` library to create a scatter plot as in Q1.1. Make sure to add the fitted line as well, using the `loess` method in `geom_smooth`.

2.2 The `quantile()` function in R gives us the rank of the order of the values in a numerical dataset. E.g. if a dataset has been read in the variable called `data`, then `quantile(data, 0.25)` would return the member in data whose rank is the 25th percentile (25% of all elements of the dataset are less than or equal to it). The median therefore is the 50th percentile, and can be computed via `quantile(data, 0.5)`.
Furthermore, the `subset()` function in R can be used to select specific rows from a dataset. Again, say a dataset has been read into the variable `data`, and if one wishes to select the rows corresponding to the second quartile of a specific numeric column,

say `val`, one can use

```
quart_2 <- subset(data, data$val > quantile(data$val, 0.25) &
           data$val <= quantile(data$val, 0.50))
```

Use the `ggplot2` library to plot the density of Age for each of the first, second, third, and fourth quartile of Aneurysms_q1. Include all the densities in a single plot. Make sure that you choose x and y ranges that ensure all densities are fully visible. Use different colors to differentiate the densities and add a legend to the plot.

2.3 Use the `ggplot` library to create a histogram plot as in Q1.3. Make sure to include both histograms.

2.4 Discuss which plotting library (ggplot vs. basic R) produces more "beautiful" plots and why.

2.5 Identify a way to extend/enhance each plot in ggplot (i.e., to make the plot more aesthetically pleasing and/or more informative). Include the new plots for comparison.

3 Using plotly (6 pts)

3.1 Use the `plotly` library to plot Aneurysms_q1 vs. Age as in Q1.2b. However, use a violin plot instead of a box plot.

3.2 Report any differences that you observe between the violin plot and the box plot made in Q1.2b.

3.3 Publish your violin chart to the web with Plotly's web service. Include a link to the online chart in your results.