Statistical Methods for Data Science (Fall 2019)

Project 2

Instructions:

- Due date: Sep 30, 2019.
- Total points = 20.
- Submit a typed report.
- Justify all steps and provide all relevant explanations.
- You can work on the project either individually or in a group of no more than two students. In case of the latter, submit only one report for the group, and include a description of the contribution of each member.
- It is OK to discuss the project with other students in the class (even those who are not in your group), but each group must write its own code and answers. If the submitted report (including code and answer) is similar (either partially or fully) to someone else's, this will be considered evidence of academic dishonesty, and you will referred to appropriate university authorities.
- Start early.
- You must use the following template for your report:

  Project #
  Name
  Names of group members (if applicable)
  Contribution of each group member
  Section 1. Answers to the specific questions asked
  Section 2: R code. Your code must be annotated. No points may be given if a brief look at the code does not tell us what it is doing.

---

1. (12 points) Consider the dataset `roadrace.csv` posted on eLearning. It contains observations on 5875 runners who finished the 2010 Beach to Beacon 10K Road Race in Cape Elizabeth, Maine. You can read the dataset in `R` using `read.csv` function.

   (a) Create a bar graph of the variable `Maine`, which identifies whether a runner is from Maine or from somewhere else (stated using `Maine` and `Away`). You can use `barplot` function for this. What can we conclude from the plot? Back up your conclusions with relevant summary statistics.

   (b) Create two histograms the runners' times (given in minutes) — one for the `Maine` group and the second for the `Away` group. Make sure that the histograms on the same scale. What can we conclude about the two distributions? Back up your conclusions with relevant summary statistics, including mean, standard deviation, range, median, and interquartile range.

   (c) Repeat (b) but with side-by-side boxplots.

   (d) Create side-by-side boxplots for the runners' ages (given in years) for male and female runners. What can we conclude about the two distributions? Back up your conclusions with relevant summary statistics, including mean, standard deviation, range, median, and interquartile range.

2. (8 points) Consider the dataset `motorcycle.csv` posted on eLearning. It contains the number of fatal motorcycle accidents that occurred in each county of South Carolina during 2009. Create a boxplot of data and provide relevant summary statistics. Discuss the features of the data distribution. Identify which counties may be considered outliers. Why might these counties have the highest numbers of motorcycle fatalities in South Carolina?