Assignment 1

Ashish Anand, Akshay Parekh
CS595: Data Visualization
Due Date: January 24, 2020

January 17, 2020

1 Outline

Objective of the assignment is the following:

- Understand the role of Data Visualization in Exploratory Data Analysis (EDA)
- How to summarize a tabular data using different types of plots
- Relevant plots for the assignment: barplot, histogram, smoothed density, quantile-quantile plots
- Understand the advantages and limitations of those plots
- Understand the relation between these plots
- Familiarize yourself with ggplot2 (R library for creating visualizations) features

Reference: **Chapter 8: https://rafalab.github.io/dsbook/**

2 Questions

Question 1. [30 points] Variables in a tabular data are either `categorical` or `numerical`. Consider the *heights data* available with `R`. You can follow the steps given below to get the access of data.

- library(dslabs) //If it is not available, please install
- data("heights")

Answer the following questions.

I. Which graph you will use to plot data for gender distribution and height distribution? Plot and Justify. Do we need any plot to understand gender distribution?

II. Show using plot, what percentage of students have heights between 66 inches and 72 inches?

III. What range of heights contains 95% of the data? Which graph is effective to show such analysis? Plot and give justification.

Question 2. [10 points] Use smoothed density curve to plot the height of Male students, highlighting the students with height between 66 and 72 inches. For the same data, plot and justify how is smoothed density graph different from histogram.

Question 3. [10 points] Does male height follow normal distribution? Justify your answer with suitable smoothed density plots. Also, answer what percentage of values lies within 1.5 standard deviation of the mean.

Question 5. [10 points] Plot a quantile-quantile plot (QQ Plot) to check whether the Male height distribution is well approximated by the normal distribution.