**Discovering Statistics Using R – Tasks**

Akila Selvaraj

Masters of Data Science,

Bellevue University

DSC 520 T301

January 08,2022

# Assignment: ASSIGNMENT 4

# Name: Selvaraj, Akila

# Date: 2022-01-07

install.packages("ggplot")

library("ggplot2")

getwd()

setwd("G:/Users/a162940/Akila/Work/R/Projects/dsc520-master")

list.files(R.home())

ls()

list.files (path = "G:/Users/a162940/Akila/Work/R/Projects/dsc520-master/data")

Scores\_df <- read.csv("data/scores.csv")

str(Scores\_df)

**#1. What are the observational units in this study?**

Observational Units in this study is Section which says Regular or Sports. Variable is Score.

**#2. Identify the variables mentioned in the narrative paragraph and determine which are categorical and quantitative?**

Type of section, Course grades and total points earned in the course are the variables. Categorical variable is type of section. Course grades and total points earned are quantitative variable.

**#3. Create one variable to hold a subset of your data set that contains only the Regular Section and one variable for the Sports Section.**

Sports\_Score\_df <- subset(Scores\_df, Section=="Sports")

Sports\_Score\_df

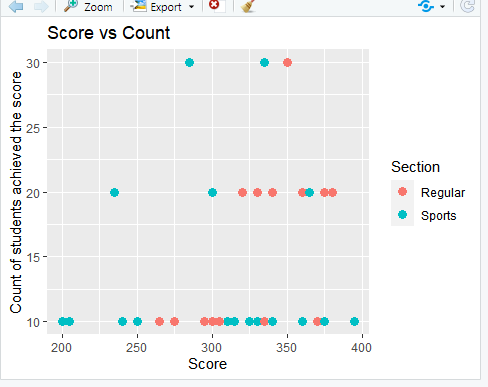
Regular\_Score\_df = subset(Scores\_df, Section=="Regular")

Regular\_Score\_df

**#Use the Plot function to plot each Sections scores and the number of students achieving that score.**

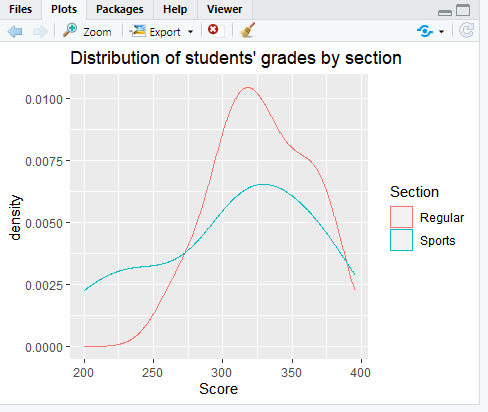
**#Use additional Plot Arguments to label the graph and give each axis an appropriate label.**

plot(Scores\_df$Score,Scores\_df$Count,xlab = "Scores", ylab = "Count", col = Scores\_df$Color)



ggplot(Scores\_df, aes(Score,Count,color=Section)) + geom\_point(size =3) + ggtitle("Score vs Count") + xlab("Score") + ylab("Count of students achieved the score")

ggplot(Scores\_df, aes(Score)) + geom\_density(aes(color=Section)) + ggtitle("Distribution of students' grades by section")



**#Once you have produced your Plots answer the following questions:**

**a.Comparing and contrasting the point distributions between the two section,**

**looking at both tendency and consistency: Can you say that one section tended to score more points than the other? Justify and explain your answer.**

In the scatterplot, relationship between Scores gained and the count of students achieved that score are shown. For each and every score, count of students achieved that score is plotted. Highest score is achieved by students in Sports section. At the same time, there are students in Regular section who scored more than students in Sports section. There is a mix of scores distribution in both the classes. So, It is hard to determine which section tended to score more points than the other with the data given.

**b. Did every student in one section score more points than every student in the other section? If not, explain what a statistical tendency means in this context.**

By looking at the plot, it is not easy to come to a conclusion, that every student in one section scored more points than every other student in other section.

mean(Sports\_Score\_df$Score)

[1] 307.3684

mean(Regular\_Score\_df$Score)

[1] 327.6316

median(Sports\_Score\_df$Score)

[1] 315

median(Regular\_Score\_df$Score)

[1] 325

Measure of statistical tendency which describes a whole set of data with a single value in this case is mean and median scores of each section. If we look at the mean of scores in both the sections, average score earned by Regular Section is greater than the average scored by Sports section.

Middle value of the dataset measured by median is 315 for Sports section whereas median scored by regular section is 325. By looking at the central tendency, points earned by Regular section is slightly more than the points earned by Sports section.

**c. What could be one additional variable that was not mentioned in the narrative that could be influencing the point distributions between the two sections?**

Grade earned by individual student could influence the point distribution between two sections. Also, there are various factors which can be added to the dataset like hours each student spent in each class influence the individual student's score more.