Assignment: ASSIGNMENT 4.2.1

Name: Anjale, Jiteshwar

Date: 2021-04-06

#Analysis of Test Scores

Load the ggplot2 package

library(ggplot2)

theme_set(theme_minimal())

Load the pastecs package

library(pastecs)

Set the working directory to the root of your DSC 520 directory

setwd('C:/Users/anjal/OneDrive/Desktop/MS/DSC520/dsc520')

Load the `data/scores.csv` to Scores

Scores_df <- read.csv("data/scores.csv")

1.What are the observational units in this study?

str(Scores_df)

```
> str(Scores_df)
'data.frame':   38 obs. of  3 variables:
 $ Count  : int  10 10 20 10 10 10 10 30 10 10 ...
 $ Score  : int  200 205 235 240 250 265 275 285 295 300 ...
 $ Section: chr  "Sports" "Sports" "Sports" "Sports" ...
```

#There are 2 (score and count) observational units in the given study.

#There are 38 observations in the study.

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

str(Scores_df)

```
> str(Scores_df)
'data.frame':   38 obs. of  3 variables:
 $ Count  : int  10 10 20 10 10 10 10 30 10 10 ...
 $ Score  : int  200 205 235 240 250 265 275 285 295 300 ...
 $ Section: chr  "Sports" "Sports" "Sports" "Sports" ...
```

summary(Scores_df)

```
> summary(Scores_df)
     Count           Score         Section
 Min.   :10.00   Min.   :200.0   Length:38
 1st Qu.:10.00   1st Qu.:300.0   Class :character
 Median :10.00   Median :322.5   Mode  :character
 Mean   :14.47   Mean   :317.5
 3rd Qu.:20.00   3rd Qu.:357.5
 Max.   :30.00   Max.   :395.0
```

#Section is the categorical variable for the study.

#Count and Score are quantitative variables for the study.

#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.

reg_df <-Scores_df[which(Scores_df$Section=='Regular'),]

head(reg_df)

```
> reg_df <-Scores_df[which(Scores_df$Section=='Regular'),]
> head(reg_df)
   Count Score Section
6     10   265 Regular
7     10   275 Regular
9     10   295 Regular
10    10   300 Regular
13    10   305 Regular
14    10   310 Regular
```

sport_df<-Scores_df[which(Scores_df$Section=='Sports'),]

head(sport_df)

```
> sport_df<-Scores_df[which(Scores_df$Section=='Sports'),]
> head(sport_df)
  Count Score Section
1    10   200  Sports
2    10   205  Sports
3    20   235  Sports
4    10   240  Sports
5    10   250  Sports
8    30   285  Sports
>
```

4.Use the Plot function to plot each Sections scores and the number of

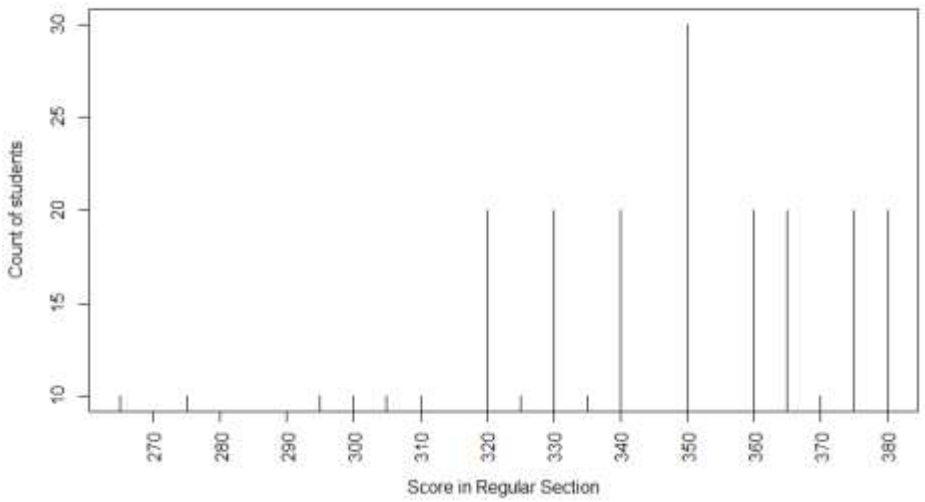
students achieving that score. Use additional Plot Arguments to label the

graph and give each axis an appropriate label. Once you have produced your

Plots answer the following questions:

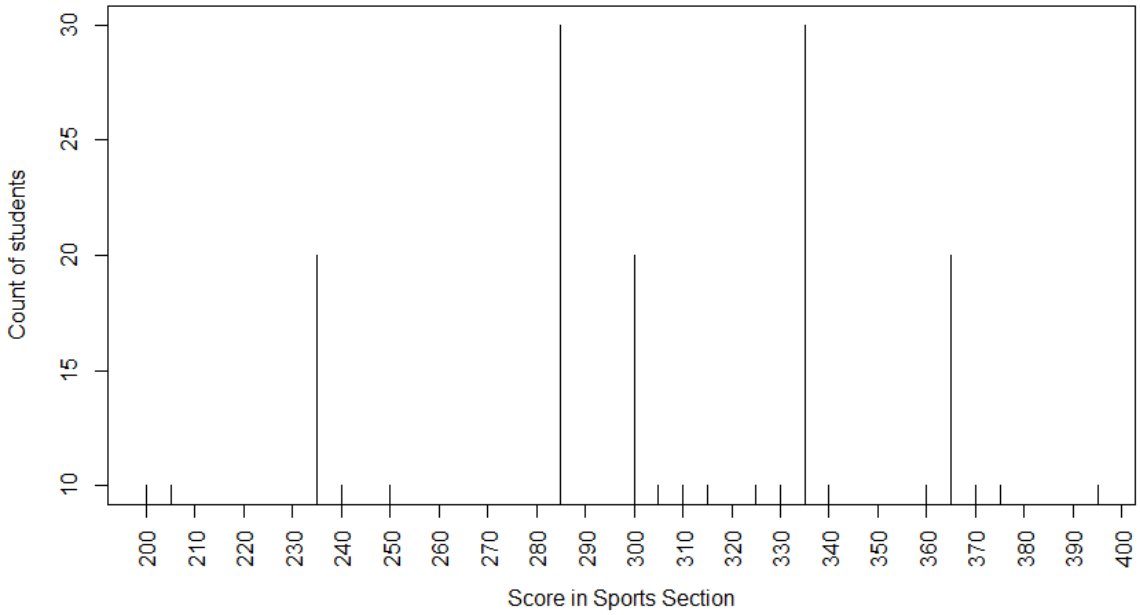
plot(reg_df$Score,reg_df$Count,type='h',xaxt="n",xlab="Score in Regular Section",ylab="Count of students")

axis(1, at = seq(200, 400, by = 10), las=2)



plot(sport_df$Score,sport_df$Count,type='h',xaxt="n",xlab="Score in Sports Section",ylab="Count of students")

axis(1, at = seq(200, 400, by = 10), las=2)



4.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.

#By looking at the two histograms plots, it seems that sports section students

#scored more higher makes > 300.

4.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.

stat.desc(reg_df[,1:2], basic=TRUE, desc=TRUE, norm=FALSE, p=0.95)

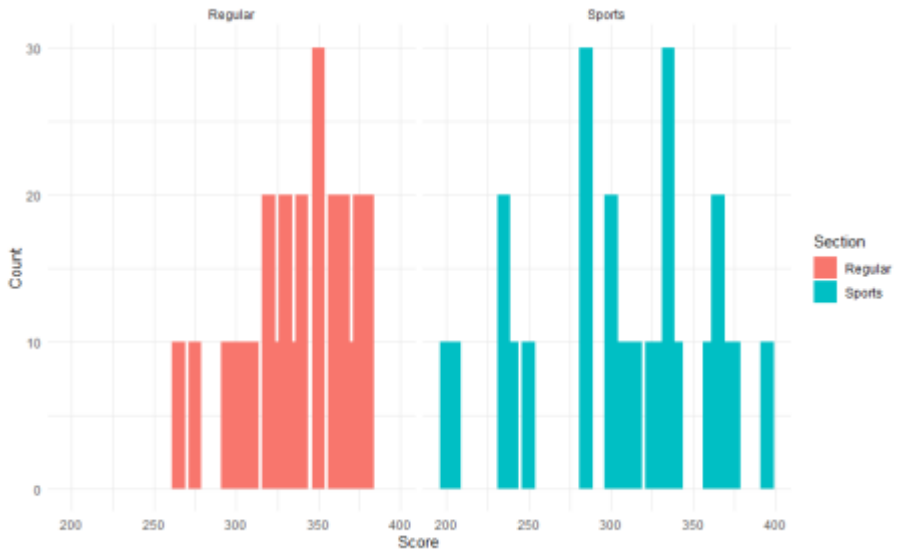
```
> stat.desc(reg_df[,1:2], basic=TRUE, desc=TRUE, norm=FALSE, p=0.95)
                  Count        Score
nbr.val      19.0000000   19.0000000
nbr.null      0.0000000    0.0000000
nbr.na        0.0000000    0.0000000
min          10.0000000  265.0000000
max          30.0000000  380.0000000
range        20.0000000  115.0000000
sum         290.0000000 6225.0000000
median       10.0000000  325.0000000
mean         15.2631579  327.6315789
SE.mean       1.4035088    7.6315789
CI.mean.0.95  2.9486625   16.0333524
var          37.4269006 1106.5789474
std.dev       6.1177529   33.2652814
coef.var      0.4008183    0.1015326
```

stat.desc(sport_df[,1:2], basic=TRUE, desc=TRUE, norm=FALSE, p=0.95)

```
> stat.desc(sport_df[,1:2], basic=TRUE, desc=TRUE, norm=FALSE, p=0.95)
                  Count        Score
nbr.val      19.0000000   19.0000000
nbr.null      0.0000000    0.0000000
nbr.na        0.0000000    0.0000000
min          10.0000000  200.0000000
max          30.0000000  395.0000000
range        20.0000000  195.0000000
sum         260.0000000 5840.0000000
median       10.0000000  315.0000000
mean         13.6842105  307.3684211
SE.mean       1.5691705   13.3134085
CI.mean.0.95  3.2967049   27.9704333
var          46.7836257 3367.6900585
std.dev       6.8398557   58.0318021
coef.var      0.4998356    0.1888021
```

bar <- ggplot(Scores_df, aes(Score,Count, fill = Section))

bar + stat_summary(fun = mean, geom = "bar", position="dodge",width = 8)+ facet_wrap(~ Section)



#Total number of students in regular section is 290 and their mean score is 327.63

#Total number of students in sports section is 260 and their mean score is 307.37

#It looks like not every student in sports section score more points than every student in regular section.

4.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?

#I think 'gender' will be an additional variable could be influencing the point

#distributions between the two sections