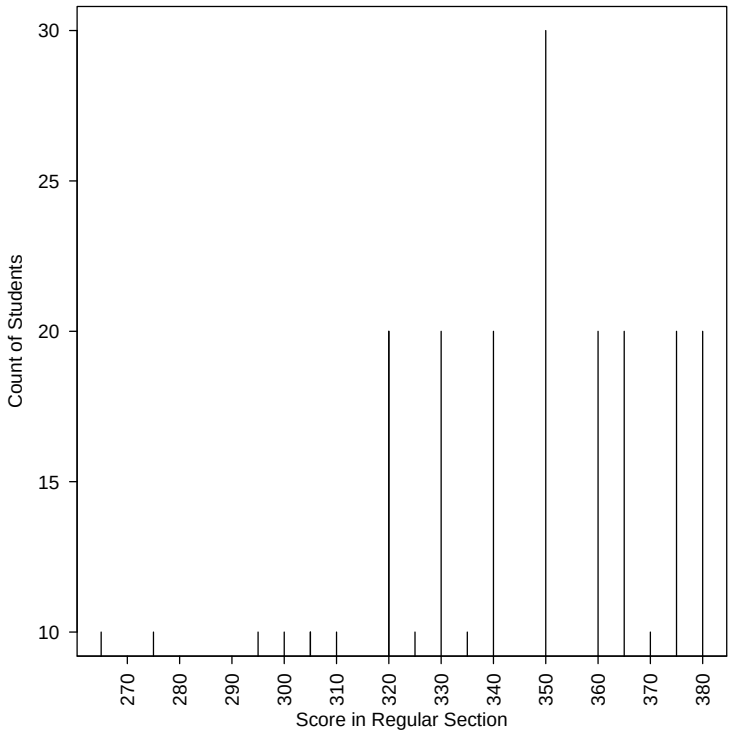
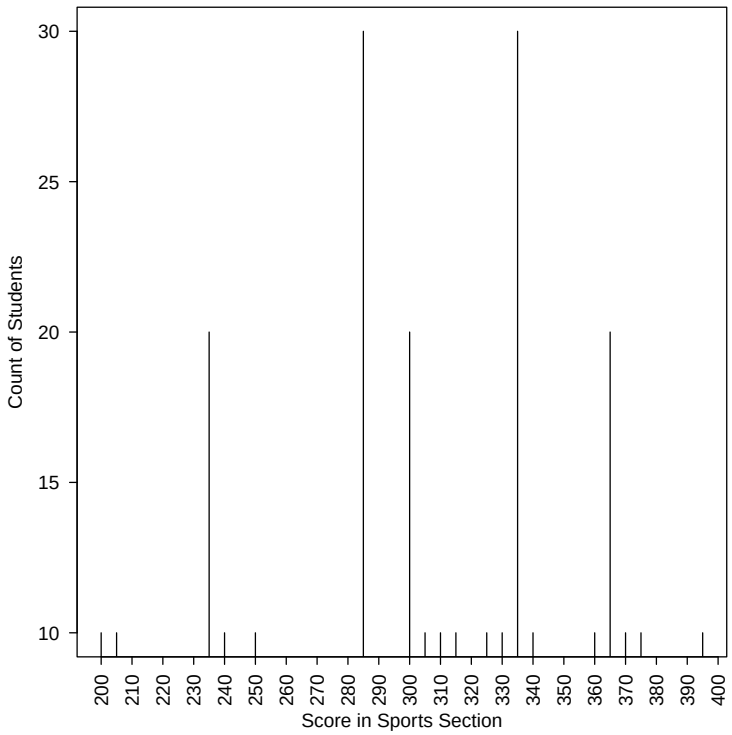
The results below are generated from an R script.

```
# Assignment: ASSIGNMENT 4.1 Scores Exercise
# Name: Ghanta, Madhavi
# Date: 2023-04-06

## Load the ggplot2 package
library(ggplot2)
theme_set(theme_minimal())

## Load the pastecs package
library(pastecs)

##If the current directory does not contain the data directory, set the working
## directory to project root folder.(the folder should contain  data directory )
## Set the working directory to the root of your DSC 520 directory
setwd("C:/Users/mghan/Documents/dsc520")

## Load the `data/scores.csv` to Scores_df using read.csv
Scores_df <- read.csv("data/scores.csv")

##A professor has recently taught two sections of the same course with only one
##difference between the sections. In one section, he used only examples taken
##from sports applications, and in the other section, he used examples taken
##from a variety of application areas. The sports themed section was advertised
##as such; so students knew which type of section they were enrolling in. The
##professor has asked you to compare student performance in the two sections
##using course grades and total points earned in the course. You will need to
##import the Scores.csv dataset that has been provided for you.

## Examine the structure of Scores_df using str()
## 1.What are the observational units in this study?
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" ...

# We have 38 observations with three variables.
## score and count are observational units in this study.

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

##      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 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.
View(Scores_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

View(reg_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

View(sport_df)

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



```r
# 4.1. 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 marks > 300.

# 4.2. 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)
```

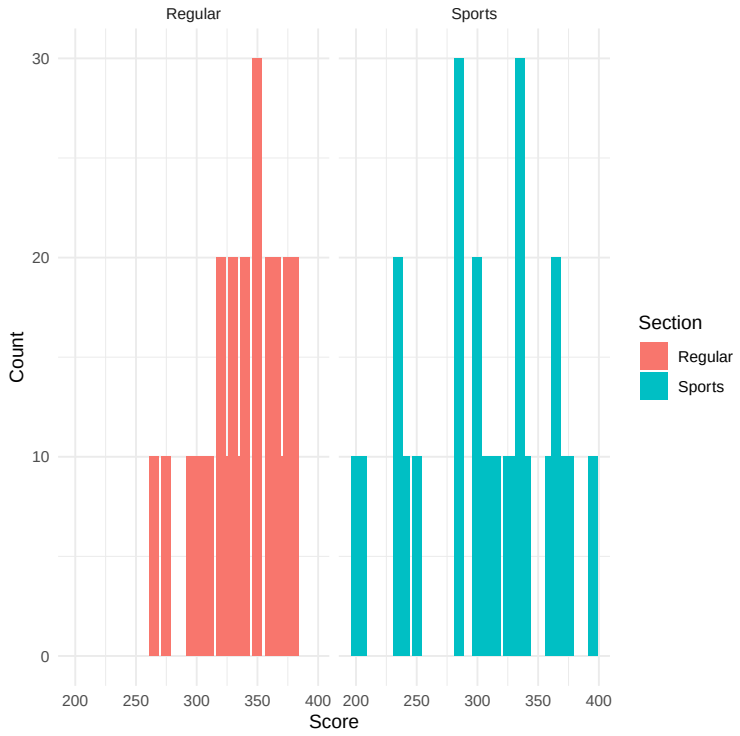
```
##                    Count        Score
## nbr.val       19.0000000   19.0000000
## nbr.null       0.0000000    0.0000000
## nbr.na         0.0000000    0.0000000
## min           10.0000000  265.0000000
## max           30.0000000  380.0000000
## range         20.0000000  115.0000000
## sum          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
```

```r
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
```

```r
bar <- ggplot(Scores_df, aes(Score,Count, fill = Section))
bar + stat_summary(fun = mean, geom = "bar", position="dodge",width = 8)+ facet_wrap( ~ Section)
```

```
## Warning:  `position_dodge()` requires non-overlapping x intervals
## `position_dodge()` requires non-overlapping x intervals
```




```
#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 sect

# 4.3. 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 'size of classes in each section' will be an additional variable could be influencing the point
#distributions between the two sections
```

The R session information (including the OS info, R version and all packages used):

```
sessionInfo()

## R version 4.2.2 (2022-10-31 ucrt)
## Platform: x86_64-w64-mingw32/x64 (64-bit)
## Running under: Windows 10 x64 (build 22621)
##
## Matrix products: default
##
## locale:
## [1] LC_COLLATE=English_United States.utf8  LC_CTYPE=English_United States.utf8
## [3] LC_MONETARY=English_United States.utf8 LC_NUMERIC=C
## [5] LC_TIME=English_United States.utf8
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base
##
## other attached packages:
## [1] pastecs_1.3.21 ggplot2_3.4.1  tidyr_1.3.0
```

```
##
## loaded via a namespace (and not attached):
##  [1] rstudioapi_0.14  knitr_1.42       magrittr_2.0.3   tidyselect_1.2.0 munsell_0.5.0
##  [6] colorspace_2.1-0 R6_2.5.1         rlang_1.1.0      fansi_1.0.4      highr_0.10
## [11] dplyr_1.1.1      tools_4.2.2      grid_4.2.2       gtable_0.3.3     xfun_0.38
## [16] utf8_1.2.3       cli_3.6.1        withr_2.5.0      tibble_3.2.1     lifecycle_1.0.3
## [21] farver_2.1.1     purrr_1.0.1      vctrs_0.6.1      evaluate_0.20    glue_1.6.2
## [26] labeling_0.4.2   compiler_4.2.2   pillar_1.9.0     generics_0.1.3   scales_1.2.1
## [31] boot_1.3-28      pkgconfig_2.0.3

Sys.time()

## [1] "2023-04-07 22:44:17 PDT"
```