1. 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.
    a. Use the appropriate R functions to answer the following questions:
        i. What are the observational units in this study?
        The observational units would be the students and the course sections.
        ii. Identify the variables mentioned in the narrative paragraph and determine which are categorical and quantitative?
        Looking solely at the narrative paragraph above, section would be classified as categorical because sports and regular are the only options under section. Course grades (A,B,C,etc) would be categorical as well. Total points on the other hand would be quantitative.
        iii. 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_sec <- filter(course_df, Section == "Regular")

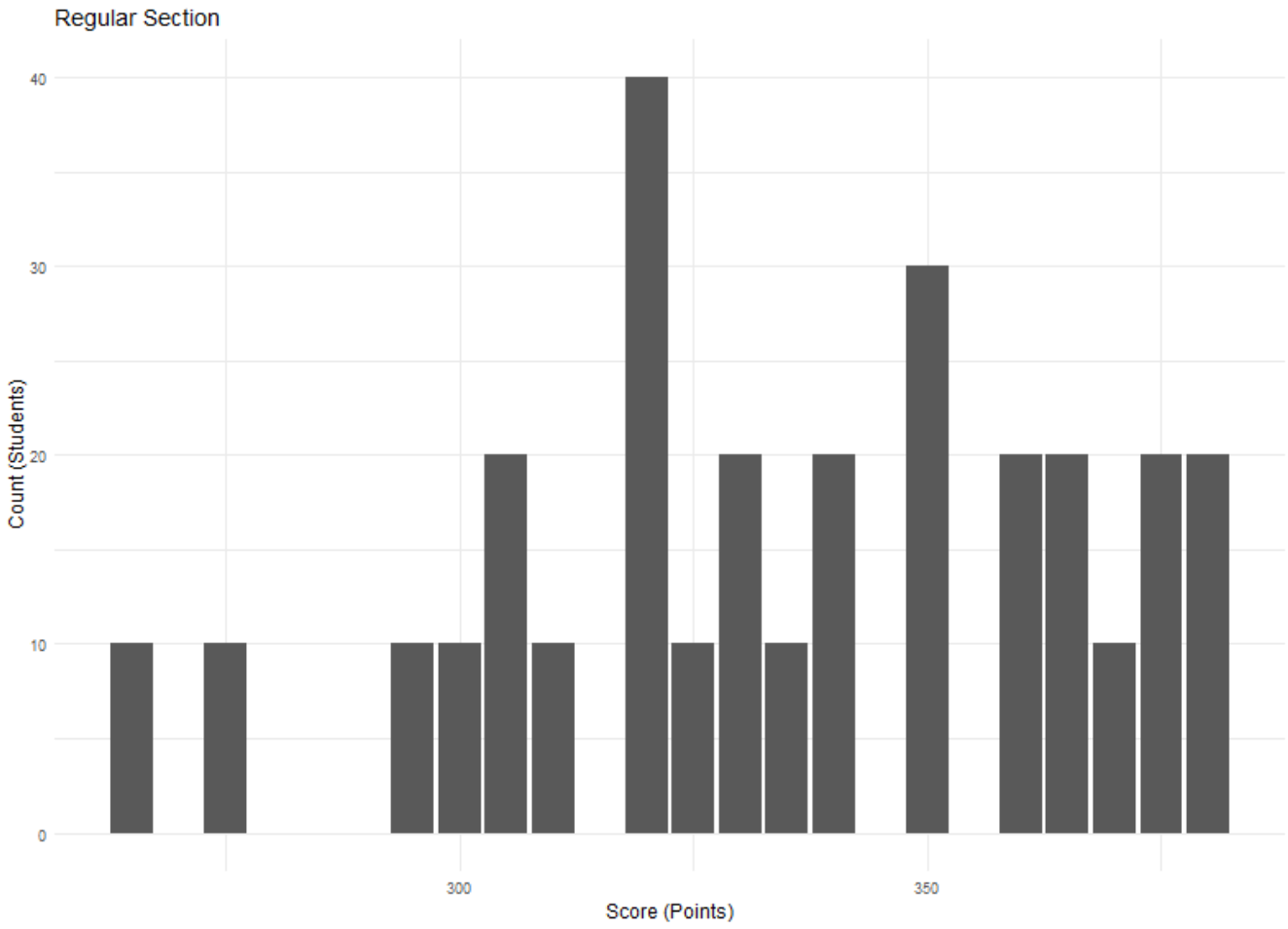
```
> ## Create one variable to hold a subset of your data set that contains only the Regular
> ## Section and one variable for the Sports Section.
> reg_sec <- filter(course_df, Section == "Regular")
> print(reg_sec)
   Count Score Section
1     10   265 Regular
2     10   275 Regular
3     10   295 Regular
4     10   300 Regular
5     10   305 Regular
6     10   310 Regular
7     20   320 Regular
8     10   305 Regular
9     20   320 Regular
10    10   325 Regular
11    20   330 Regular
12    10   335 Regular
13    20   340 Regular
14    30   350 Regular
15    20   360 Regular
16    20   365 Regular
17    10   370 Regular
18    20   375 Regular
19    20   380 Regular
```

        sports_sec <- filter(course_df, Section == "Sports")

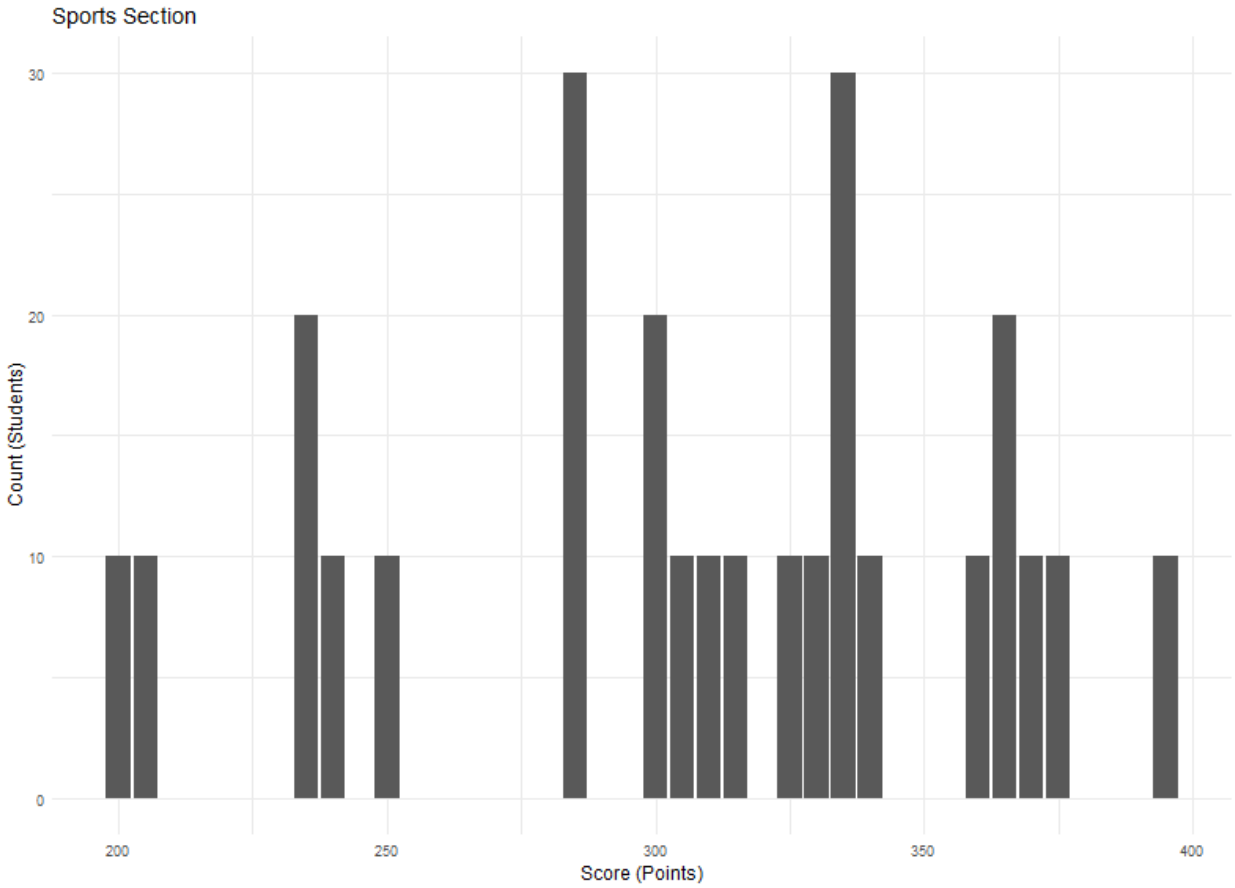
```
> sports_sec <- filter(course_df, Section == "Sports")
> print(sports_sec)
   Count Score Section
1     10   200  Sports
2     10   205  Sports
3     20   235  Sports
4     10   240  Sports
5     10   250  Sports
6     30   285  Sports
7     20   300  Sports
8     10   305  Sports
9     10   310  Sports
10    10   315  Sports
11    10   325  Sports
12    10   330  Sports
13    30   335  Sports
14    10   340  Sports
15    10   360  Sports
16    20   365  Sports
17    10   370  Sports
18    10   375  Sports
19    10   395  Sports
>
```

iv. 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:

ggplot(reg_sec, aes(x=Score, y=Count)) + geom_col() + ggtitle("Regular Section") + xlab("Score (Points)") + ylab("Count (Students)")


Regular Section
40
30
20
10
0
Count (Students)
300
350
Score (Points)

ggplot(sports_sec, aes(x=Score, y=Count)) + geom_col() + ggtitle("Sports Section") + xlab("Score (Points)") + ylab("Count (Students)")



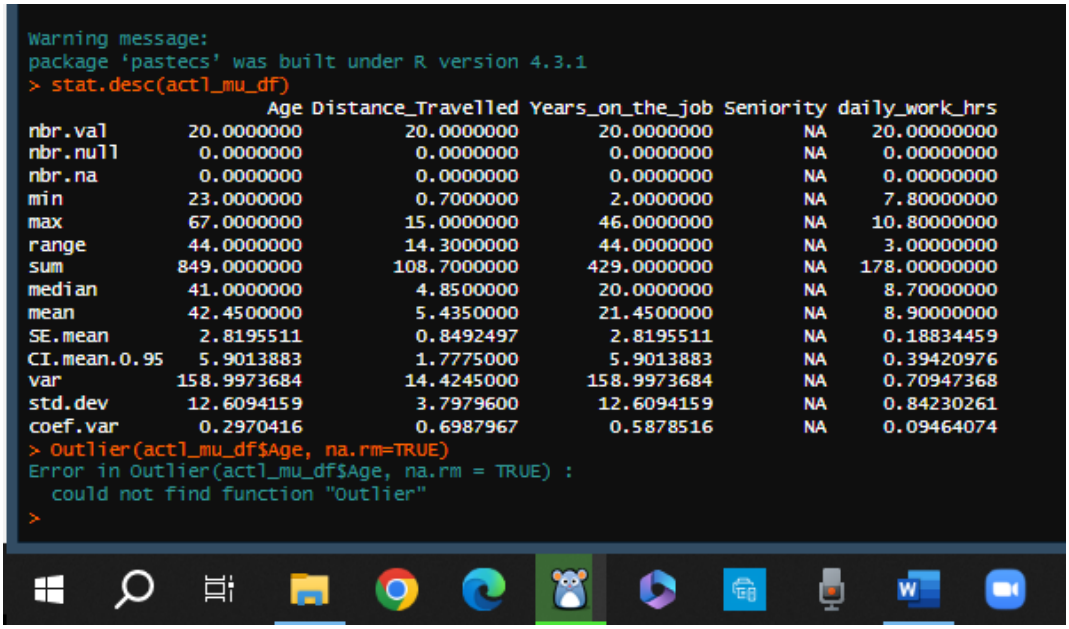
1. Comparing and contrasting the point distributions between the two section, looking at both tendency and consistency: Can you say that one section tended to score more points than the other? Justify and explain your answer.
   It looks like the regular section tended to score more points than the sports section. I based my answer on the fact that there are more students that scored above 300 points in the regular section than in the sports section.
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.
   No. Statistical tendency otherwise known as central tendency is when one value from the dataset is used to show the center of the distribution of the value for the entire dataset.
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?

Attendance. If a significant number of students from either section missed a significant number of classes. It is possible that that could influence their eventual score in the course.

2. We interact with a few datasets in this course, one you are already familiar with, the 2014 American Community Survey and the second is a Housing dataset, that provides real estate transactions recorded from 1964 to 2016. For this exercise, you need to start practicing some data transformation steps – which will carry into next week, as you learn some additional methods. For this week, using either dataset (or one of your own – although I will let you know ahead of time that the Housing dataset is used for a later assignment, so not a bad idea for you to get more comfortable with now!), perform the following data transformations:
   a. Use the apply function on a variable in your dataset.
      apply(mu_df[, c(1,2,3,4)], 2, mean)
   b. Use the aggregate function on a variable in your dataset.
      aggregate(Age~Seniority, mu_df, mean)
   c. Use the plyr function on a variable in your dataset – more specifically, I want to see you split some data, perform a modification to the data, and then bring it back together.

```
weekly_wrk_hrs <- mu_df[2]

new_mu_df <- mu_df[ , c(1,3,4,5)]

daily_wrk_hrs <- weekly_wrk_hrs/5

colnames(daily_wrk_hrs) = "daily_work_hrs"

actl_mu_df <- bind_cols(new_mu_df, daily_wrk_hrs)
```

   d. Check distributions of the data.
      stat.desc(actl_mu_df)



```
Warning message:
package 'pastecs' was built under R version 4.3.1
> stat.desc(actl_mu_df)
                Age Distance_Travelled Years_on_the_job Seniority daily_work_hrs
nbr.val      20.0000000       20.0000000      20.0000000        NA    20.00000000
nbr.null      0.0000000        0.0000000       0.0000000        NA     0.00000000
nbr.na        0.0000000        0.0000000       0.0000000        NA     0.00000000
min          23.0000000        0.7000000       2.0000000        NA     7.80000000
max          67.0000000       15.0000000      46.0000000        NA    10.80000000
range        44.0000000       14.3000000      44.0000000        NA     3.00000000
sum         849.0000000      108.7000000     429.0000000        NA   178.00000000
median       41.0000000        4.8500000      20.0000000        NA     8.70000000
mean         42.4500000        5.4350000      21.4500000        NA     8.90000000
SE.mean       2.8195511        0.8492497       2.8195511        NA     0.18834459
CI.mean.0.95  5.9013883        1.7775000       5.9013883        NA     0.39420976
var         158.9973684       14.4245000     158.9973684        NA     0.70947368
std.dev      12.6094159        3.7979600      12.6094159        NA     0.84230261
coef.var      0.2970416        0.6987967       0.5878516        NA     0.09464074
> Outlier(actl_mu_df$Age, na.rm=TRUE)
Error in Outlier(actl_mu_df$Age, na.rm = TRUE) :
  could not find function "Outlier"
>
```

e. Identify if there are any outliers.

According to the table that comes up after using the stat.desc function. Only daily work hours doesn't have outliers. This is because the standard deviation for daily work hours is less than one. Which means daily wok hours is closer to the mean. Age and Years on the job have the highest standard deviation. The standard deviation is 12.6 for both. Which tells us that the data is spread far from the mean. This means that both Age and years on the job contain outliers.

f. Create at least 2 new variables.

```
Pay_per_hr <- c(40,49,71,32,21,28,26,51,34,39,21,23,29,41,52,45,43,29,35,41)

vacation_hrs <- c(650,607,896,203,80,302,491,671,601,405,75,91,309,495,651,201,
523,26,222,591)

actl_mu_df <- bind_cols(new_mu_df, Pay_per_hr, vacation_hrs)

colnames(actl_mu_df)[5] = "Pay_per_hr"
colnames(actl_mu_df)[6] = "Vacation_hrs"
```