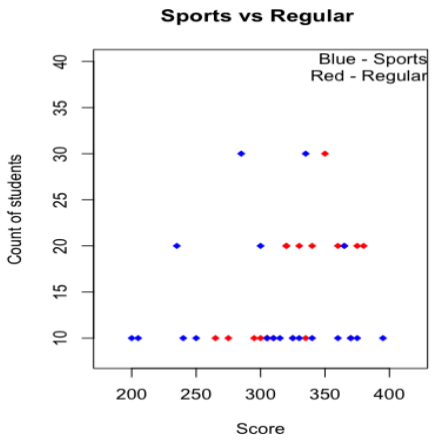
```
> # Assignment: Week4, Assignment 1
> # Name: Koppula, Veera
> # Date: 2010-06-21
>
> ## Load the ggplot2 package
> library(ggplot2)
Keep up to date with changes at https://www.tidyverse.org/blog/
> theme_set(theme_minimal())
>
> ## Set the working directory to the root of your DSC 520 directory
> ##setwd("/home/jdoe/Workspaces/dsc520")
>
> ## Load the `data/r4ds/heights.csv` to
> scores_df <- read.csv("data/scores.csv")
> scores_df
   Count Score Section
1     10   200  Sports
2     10   205  Sports
3     20   235  Sports
4     10   240  Sports
5     10   250  Sports
6     10   265 Regular
7     10   275 Regular
8     30   285  Sports
9     10   295 Regular
10    10   300 Regular
11    20   300  Sports
12    10   305  Sports
13    10   305 Regular
14    10   310 Regular
15    10   310  Sports
16    20   320 Regular
17    10   305 Regular
18    10   315  Sports
19    20   320 Regular
20    10   325 Regular
21    10   325  Sports
22    20   330 Regular
23    10   330  Sports
24    30   335  Sports
25    10   335 Regular
26    20   340 Regular
27    10   340  Sports
28    30   350 Regular
```

```
29    20   360 Regular
30    10   360  Sports
31    20   365 Regular
32    20   365  Sports
33    10   370  Sports
34    10   370 Regular
35    20   375 Regular
36    10   375  Sports
37    20   380 Regular
38    10   395  Sports
>
> #Use the appropriate R functions to answer the following questions:
> ##1.1. What are the observational units in this study?
>   #Count of students or Course grades and total points earned in the course are the
obervational units
> ##1.2. Identify the variables mentioned in the narrative paragraph and determine which are
categorical and quantitative?
>   #Section or type of examples used is Categorical
>   #total score of the students in course is Quantitative
>   #Count of students or total grades is Quantitative
> ##1.3. Create one variable to hold a subset of your data set that contains only the Regular
Section and one variable for the Sports Section.
> scores_df_Regular <- subset(scores_df,Section == "Regular")
> scores_df_Regular
   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
16    20   320 Regular
17    10   305 Regular
19    20   320 Regular
20    10   325 Regular
22    20   330 Regular
25    10   335 Regular
26    20   340 Regular
28    30   350 Regular
29    20   360 Regular
31    20   365 Regular
34    10   370 Regular
35    20   375 Regular
37    20   380 Regular
```

```
> scores_df_Sports <- subset(scores_df,Section == "Sports")
> scores_df_Sports
   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
11    20   300  Sports
12    10   305  Sports
15    10   310  Sports
18    10   315  Sports
21    10   325  Sports
23    10   330  Sports
24    30   335  Sports
27    10   340  Sports
30    10   360  Sports
32    20   365  Sports
33    10   370  Sports
36    10   375  Sports
38    10   395  Sports
> ##1.4 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(scores_df_Regular$Score, scores_df_Regular$Count, pch=18,
+      main= "Sports vs Regular ",xlab= "Score",ylab="Count of students",
+      xlim= c(180,420) , ylim= c(8,40), col.main = "black", col.lab="black", col = "Red")
> points(scores_df_Sports$Score, scores_df_Sports$Count, pch = 18, col = "Blue")
> mtext(paste(" Blue - Sports\nRed - Regular"), side= 3, line =-2, adj=1)
```




```
>
>
```

> ##1.4.1 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.
> #Comparing and contrasting the point distributions between both the plots it is clear that the section where sports applications examples were taken the scored highest compared to the regular section.
> #Set of students in sports section scored 395 whereas the maximum score in regular category is 380.
> #Also we can observe that many students under Sports category scored under 250, whereas the least scores of regular categories is around 260.
>
> ##1.4.2 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.
> #No. As we can see both the plots are almost equally distributed,from the plot it is evident that not every student in any one section scores more points than every student in the other section.
> #The plot also shows that students under regular category were consistent. As we could see the scores range from 260 <regular_scores<380. On the other hand, sports category didnt show the consistency (200<sports_scores<395).
>
> ##1.4.3 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 feel that the data provided in the csv file is self-sufficient. If asked for improvements I would suggest adding an additional variable describing the students if they are athletic with values as yes or no.
> #In the summary if the question it was already mentioned about advertising the sports category explicitly.
> #I believe that athletes and sports enthusiasts would have different experience and the results could be skewed with more athletes in the section that covers sports explicitly.