DSC520_Week4_Guruprasad_Velikadu_Krishnamoorthy_Part_4.1

Guruprasad Velikadu Krishnamoorthy

2023-01-08

Assignment Week 4 Part 1

Loading the required Packages

```
library(ggplot2)
library(psych)
```

```
##
## Attaching package: 'psych'
```

```
## The following objects are masked from 'package:ggplot2':
##
##     %+%, alpha
```

```
library(qqplotr)
```

```
##
## Attaching package: 'qqplotr'
```

```
## The following objects are masked from 'package:ggplot2':
##
##     stat_qq_line, StatQqLine
```

```
library(pastecs)
```

Set the working directory to the root of your DSC 520 directory and initial settings

```
knitr::opts_knit$set(root.dir ="C:/Users/Gurup/GURU/Learning/Masters/Term_2/DSC520_T302_Statistics_for_
knitr::opts_chunk$set(tidy.opts = list(width.cutoff = 60), tidy = TRUE)
```

4.Test Scores

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.

4.1. What are the observational units in this study?

Solution :

```
scores_df <- read.csv(file = "data/scores.csv", header = TRUE,
    stringsAsFactors = FALSE)
str(scores_df)
```

The Observational unit is defined as the unit on which the statistics are collected for the dataset. Using the str function we can see there are 38 observations of 3 variables- Count, Score and Section which are the Observational units.

```
## '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" ...
```

```
dim(scores_df)
```

```
## [1] 38  3
```

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

Solution:

```
summary(scores_df)
```

The Categorical variable is a variable based on which the data can be split into various groups. Based on the scores.csv data, The Categorical Variable will be the "Section" that can categorize the data into "Sports" and "Regular" sections. The Quantitative variable provides the measurement and in our dataset, the Quantitative variable are "Score" that gives the measurement of the score by students and also the "Count" which is the number of students that received the score for each category. The results of SUmmary and Table function shows that there are 2 Unique values in the Section which can be the Categorical variable and the fields Score and Count are Numeric that are Quantitative.

```
##      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
```

```
table(scores_df$Section, scores_df$Score)
```

```
##
##           200 205 235 240 250 265 275 285 295 300 305 310 315 320 325 330 335
##   Regular   0   0   0   0   0   1   1   0   1   1   2   1   0   2   1   1   1
##   Sports    1   1   1   1   1   0   0   1   0   1   1   1   1   0   1   1   1
##
##           340 350 360 365 370 375 380 395
##   Regular   1   1   1   1   1   1   1   0
##   Sports    1   0   1   1   1   1   0   1
```

```
table(scores_df$Section, scores_df$Count)
```

```
##
##           10 20 30
##   Regular 10  8  1
##   Sports  14  3  2
```

```
table(scores_df$Score, scores_df$Count)
```

```
##
##       10 20 30
##   200  1  0  0
##   205  1  0  0
##   235  0  1  0
##   240  1  0  0
##   250  1  0  0
##   265  1  0  0
##   275  1  0  0
##   285  0  0  1
##   295  1  0  0
##   300  1  1  0
##   305  3  0  0
##   310  2  0  0
##   315  1  0  0
##   320  0  2  0
##   325  2  0  0
##   330  1  1  0
##   335  1  0  1
##   340  1  1  0
##   350  0  0  1
##   360  1  1  0
##   365  0  2  0
##   370  2  0  0
```

```
##   375  1  1  0
##   380  0  1  0
##   395  1  0  0
```

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

```
regularData <- subset(scores_df, scores_df$Section == "Regular")
head(regularData)
```

```
##    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
```

```
sportsData <- subset(scores_df, scores_df$Section == "Sports")
head(sportsData)
```

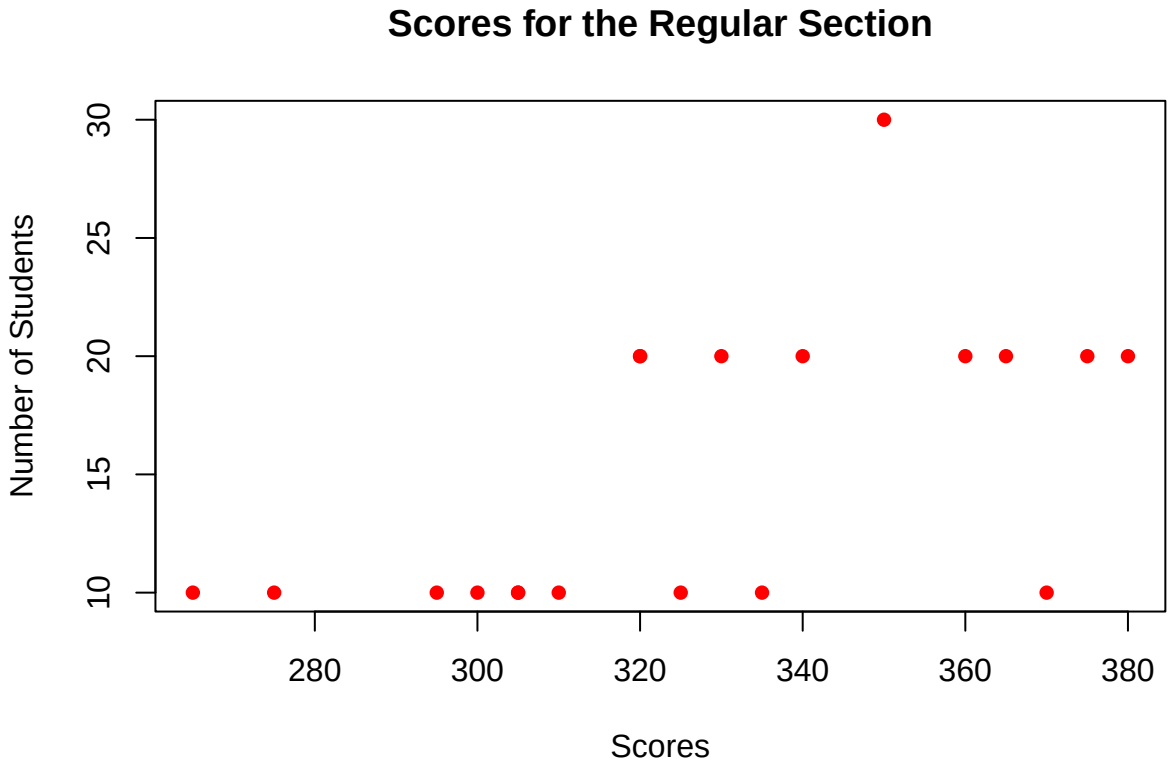
```
##   Count Score Section
## 1    10   200  Sports
## 2    10   205  Sports
## 3    20   235  Sports
## 4    10   240  Sports
## 5    10   250  Sports
## 8    30   285  Sports
```

4.4 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.

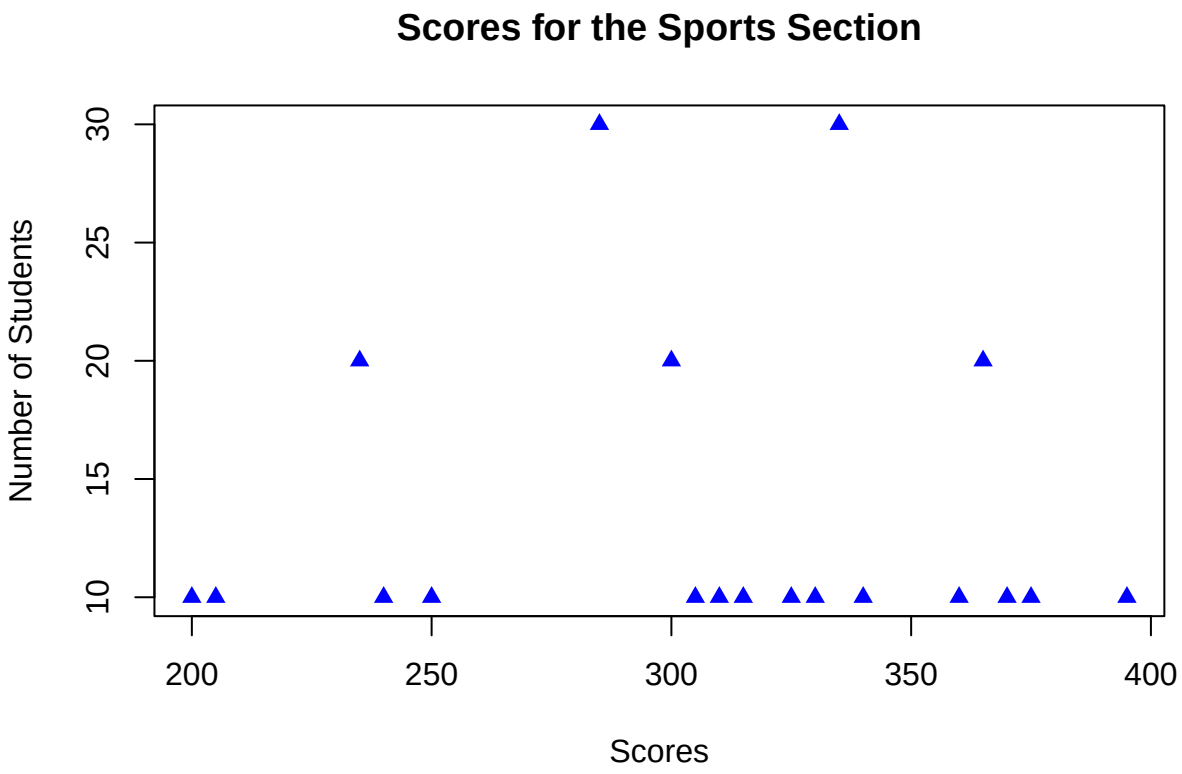
Solution:

```
# Scatter Plot for Regular section scores
plot(regularData$Score, regularData$Count, col = "Red", pch = 16,
    main = "Scores for the Regular Section", xlab = "Scores",
    ylab = "Number of Students")
```

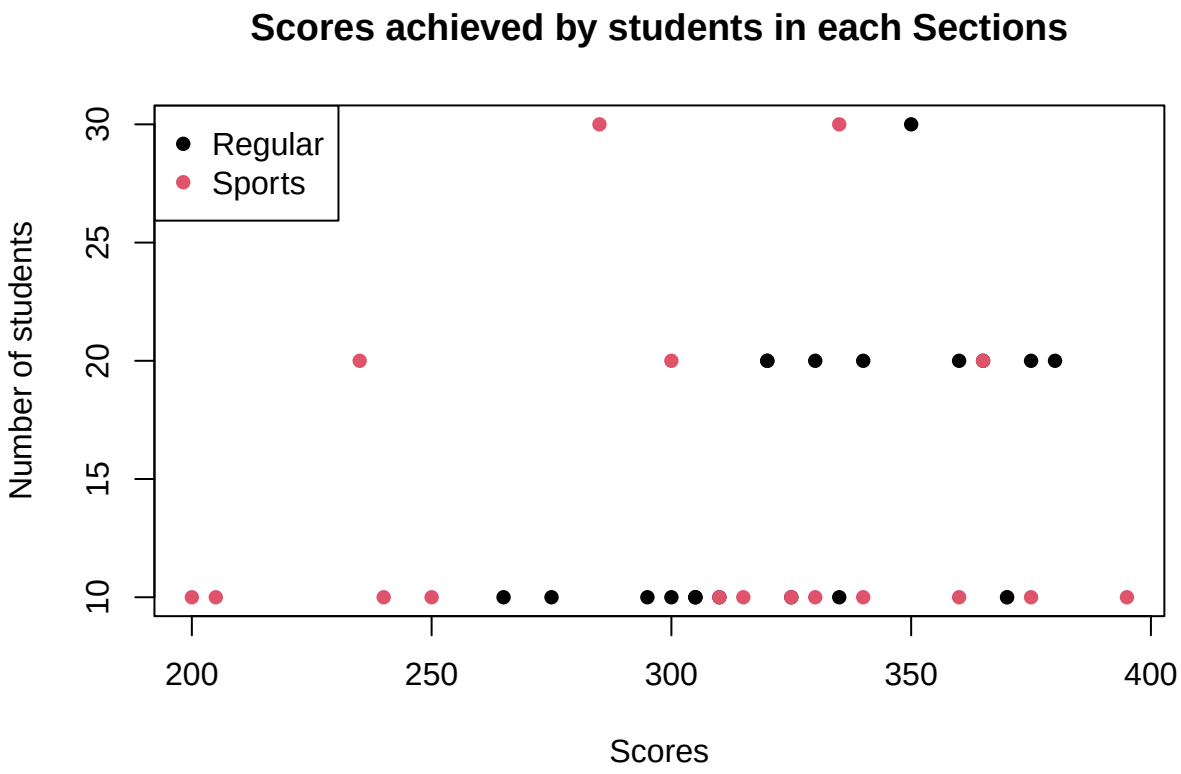
This section is answered by plotting Scatter plots using both Plot function and ggplot.



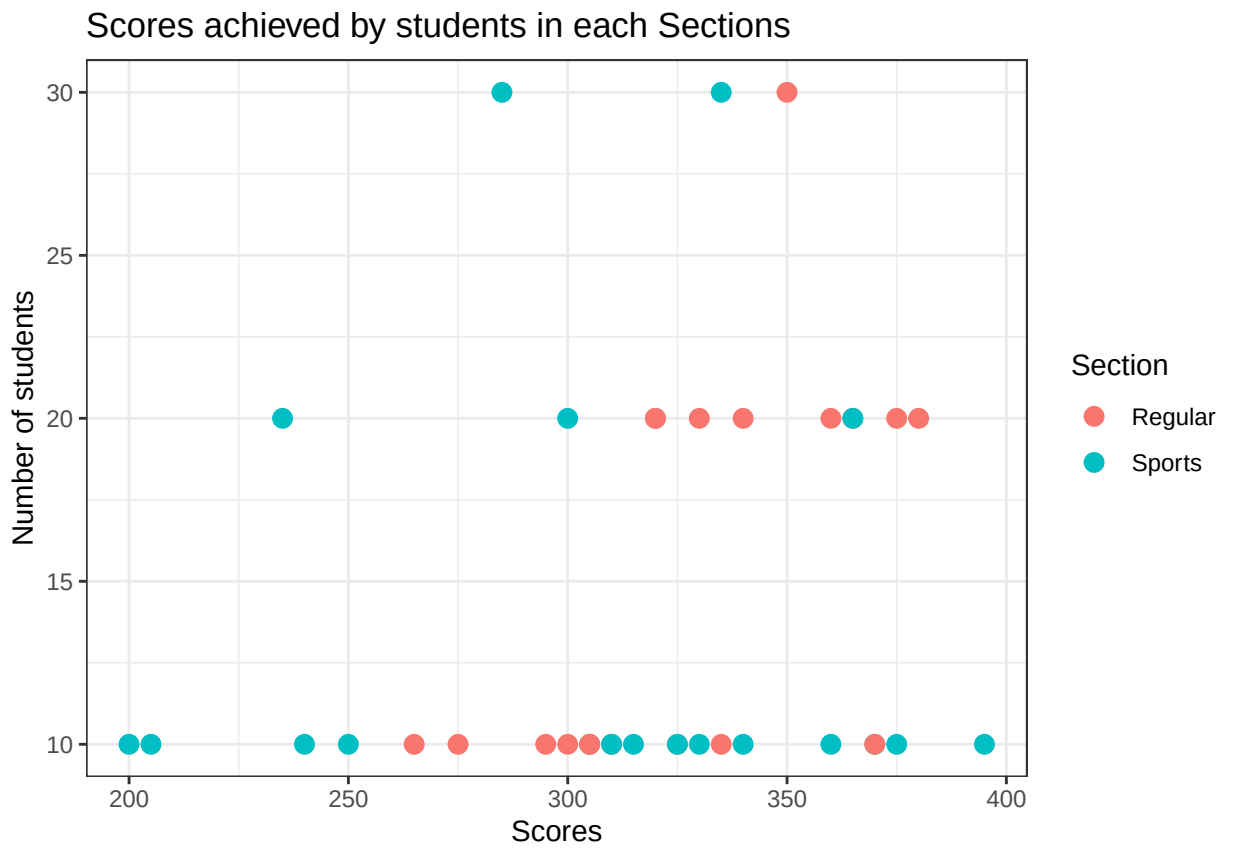

```
# Scatter Plot for Sports section scores
plot(sportsData$Score, sportsData$Count, col = "Blue", pch = 17,
    main = "Scores for the Sports Section", xlab = "Scores",
    ylab = "Number of Students")
```




```r
# creating groups to be used in plots
group <- as.factor(scores_df$Section)
# Scatter plot to show the two sections in one plot
plot(scores_df$Score, scores_df$Count, col = group, pch = 16,
    main = "Scores achieved by students in each Sections", xlab = "Scores",
    ylab = "Number of students")
legend("topleft", legend = c("Regular", "Sports"), col = 1:2,
    pch = 16)
```




```
# Using ggplot
ggplot(data = scores_df, aes(x = Score, y = Count)) + geom_point(aes(col = Section),
    size = 3) + labs(title = "Scores achieved by students in each Sections",
    x = "Scores", y = "Number of students") + theme_bw()
```



4.4.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.

```
# creating a new dataframe to compare the consistency
# between scores
new_scores_df = data.frame(regularData$Score, sportsData$Score)
alpha(new_scores_df[, c("regularData.Score", "sportsData.Score")])
```

Solution: From the plot above ,it appears the students scored more in the Regular section than the Sports section. The range of scores in Sports section(195) is wider than the range in Regular(115) section. The Minimum score in Sports section is 200 while the minimum in Regular is 265.

```
## Number of categories should be increased  in order to count frequencies.

##
## Reliability analysis
## Call: alpha(x = new_scores_df[, c("regularData.Score", "sportsData.Score")])
##
##   raw_alpha std.alpha G6(smc) average_r S/N   ase mean sd median_r
##       0.91      0.98    0.97      0.97  63 0.016  318 45     0.97
```

```
##
##     95% confidence boundaries
##          lower alpha upper
## Feldt     0.77  0.91  0.97
## Duhachek  0.88  0.91  0.94
##
##  Reliability if an item is dropped:
##                    raw_alpha std.alpha G6(smc) average_r S/N alpha se var.r
## regularData.Score       0.56      0.97    0.94      0.97  32       NA     0
## sportsData.Score        1.69      0.97    0.94      0.97  32       NA     0
##                    med.r
## regularData.Score   0.97
## sportsData.Score    0.97
##
##  Item statistics
##                     n raw.r std.r r.cor r.drop mean sd
## regularData.Score  19  0.99  0.99  0.98   0.97  328 33
## sportsData.Score   19  1.00  0.99  0.98   0.97  307 58
```

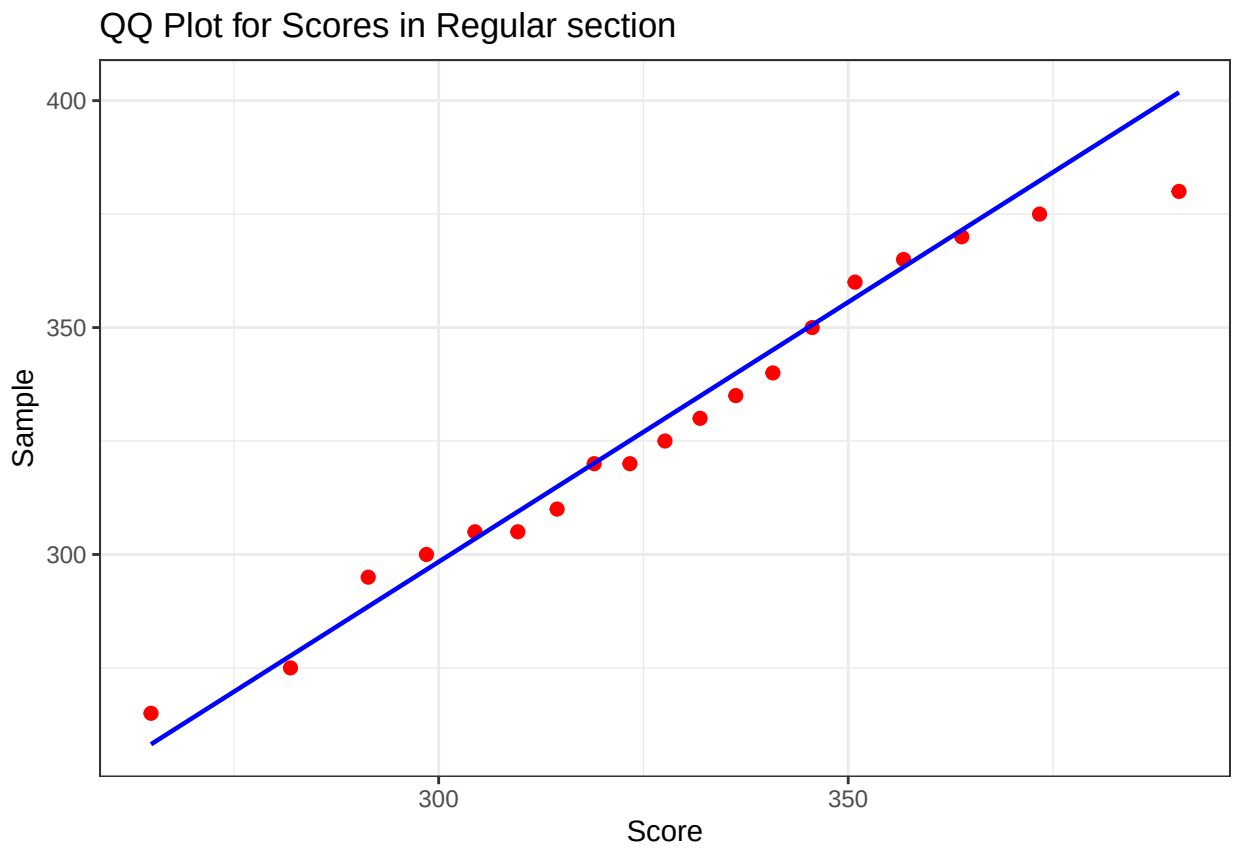
From the output of the Consistency alpha function, the value of raw_data is 0.91 which represents strong consistency between the scores in the 2 sections. The tendency results based on mean shows that students scored better in Regular section that Sports.

4.4.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.

```
# qqplot for Scores in Regular Section
ggplot(data = regularData["Score"], aes(sample = Score)) + stat_qq_point(size = 2,
    color = "red", distribution = "norm") + stat_qq_line(color = "blue") +
    labs(x = "Score", y = "Sample", title = "QQ Plot for Scores in Regular section") +
    theme_bw()
```

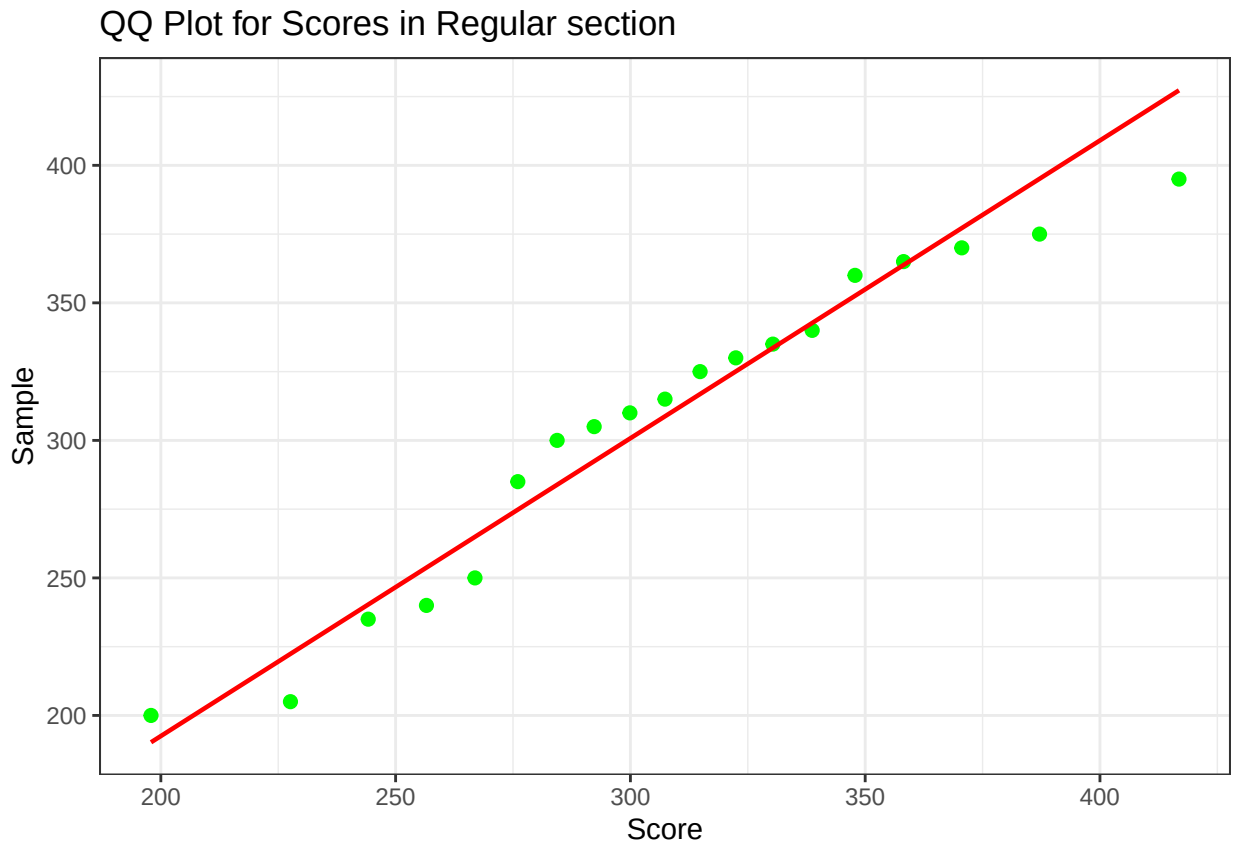
Solution : To answer this question, we have to compare the tendency(Mean, Median) of Scores from each section. QQPlots and Normal Distribution curves were plotted to justify the solution with the plots.The Mean and Median of Regular scores is greater than Sports scores(Refer output of stat.desc function). Though it is not possible to compare one score against another score, overall in terms of comparing them against the Mean and median of each section and from the qq-plots, we can say Regular section fared better than Sports overall.

```
## Warning: The following aesthetics were dropped during statistical transformation: sample
## i This can happen when ggplot fails to infer the correct grouping structure in
##   the data.
## i Did you forget to specify a ‘group‘ aesthetic or to convert a numerical
##   variable into a factor?
```



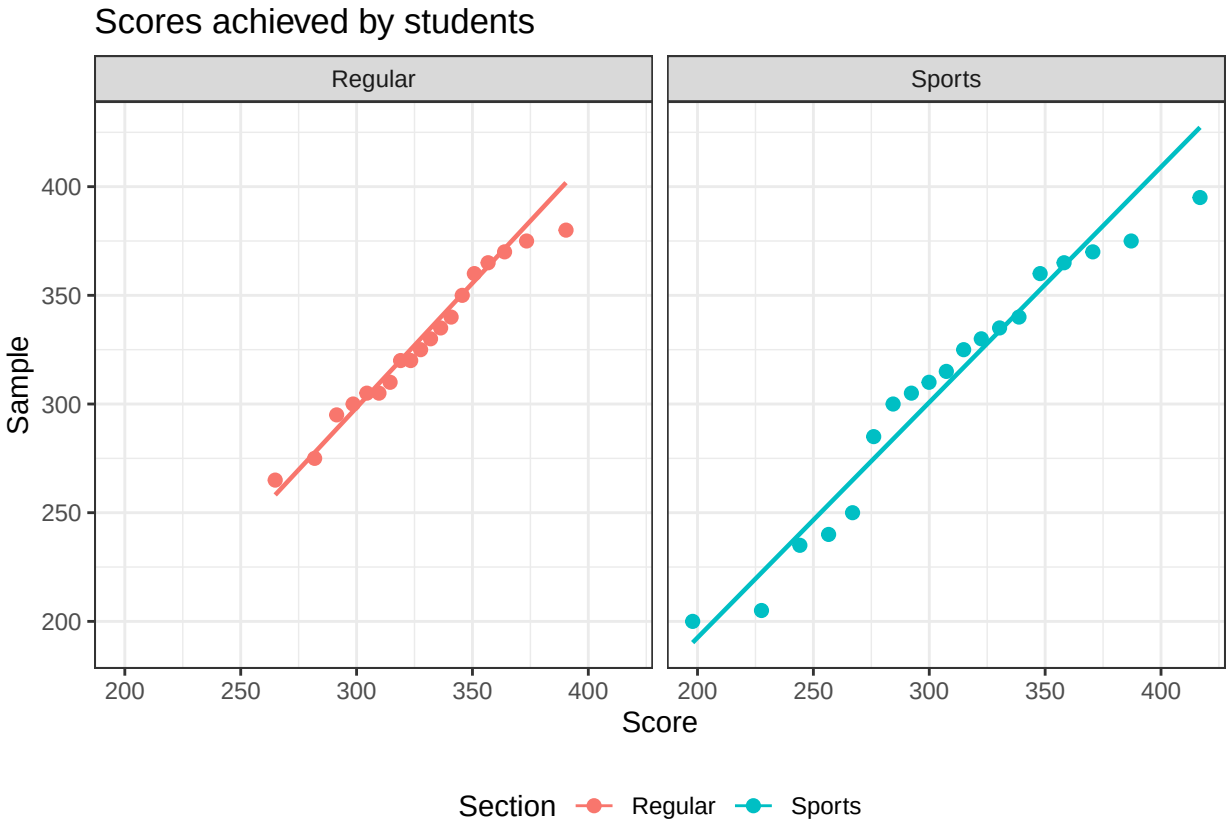

```
# qqplot for Scores in Sports Section
ggplot(data = sportsData["Score"], aes(sample = Score)) + stat_qq_point(size = 2,
    color = "green", distribution = "norm") + stat_qq_line(color = "red") +
    labs(x = "Score", y = "Sample", title = "QQ Plot for Scores in Regular section") +
    theme_bw()
```

```
## Warning: The following aesthetics were dropped during statistical transformation: sample
## i This can happen when ggplot fails to infer the correct grouping structure in
##   the data.
## i Did you forget to specify a ‘group‘ aesthetic or to convert a numerical
##   variable into a factor?
```

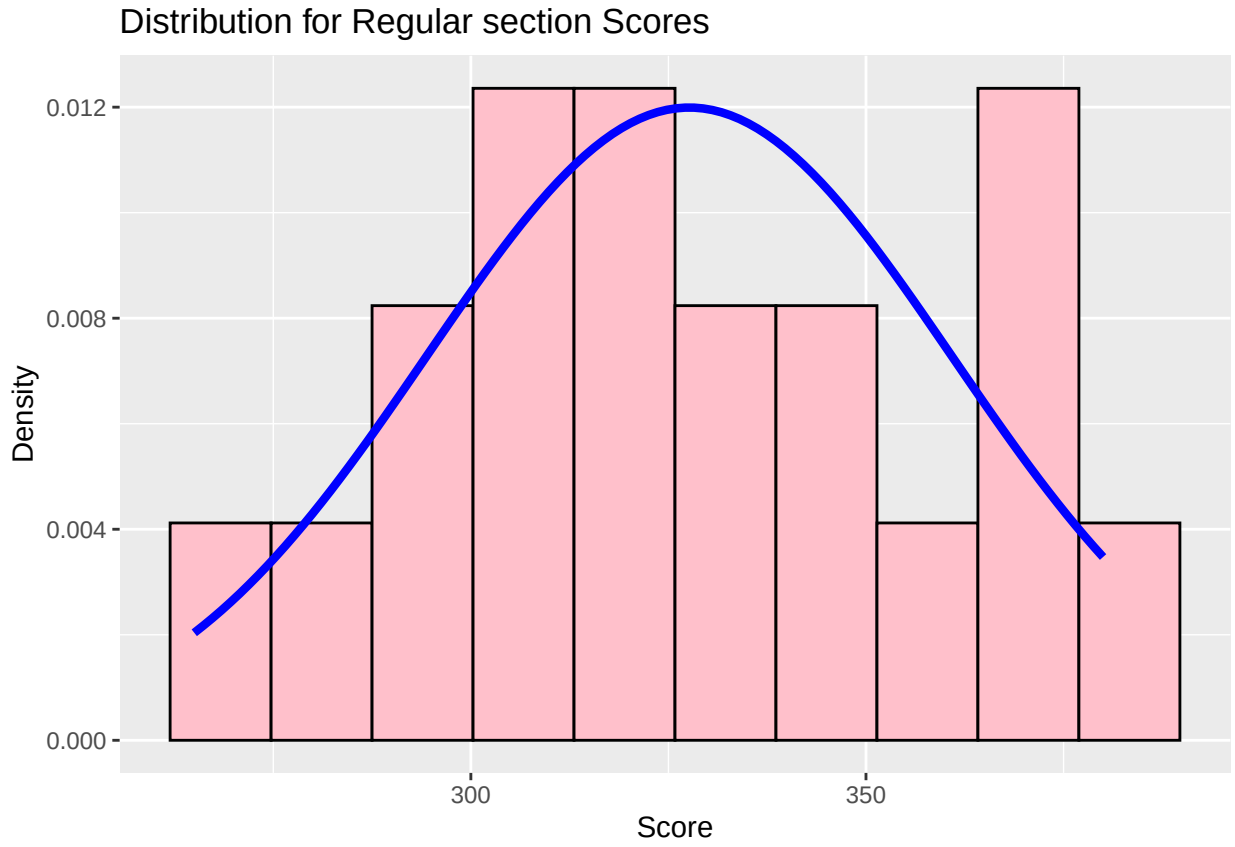



```
# qqplot for Scores for each Section
ggplot(data = scores_df, aes(sample = Score, color = Section)) +
    stat_qq_point(size = 2, distribution = "norm") + stat_qq_line() +
    facet_wrap(~scores_df$Section) + labs(x = "Score", y = "Sample",
    title = "Scores achieved by students") + theme_bw() + theme(legend.position = "bottom")
```

```
## Warning: The following aesthetics were dropped during statistical transformation: sample
## i This can happen when ggplot fails to infer the correct grouping structure in
##   the data.
## i Did you forget to specify a ‘group‘ aesthetic or to convert a numerical
##   variable into a factor?
## The following aesthetics were dropped during statistical transformation: sample
## i This can happen when ggplot fails to infer the correct grouping structure in
##   the data.
## i Did you forget to specify a ‘group‘ aesthetic or to convert a numerical
##   variable into a factor?
```

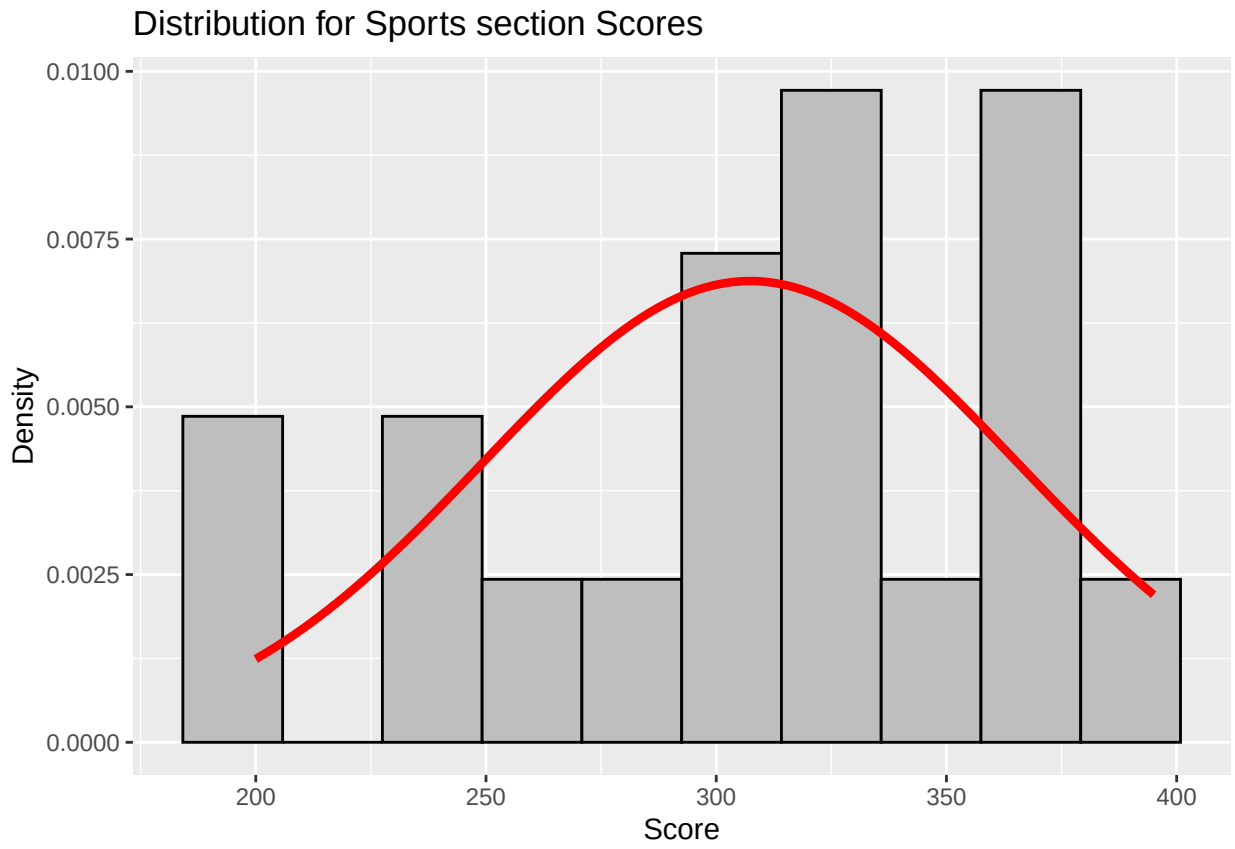



```r
# Histogram and Normal Curve for Regular Data
hist.regular <- ggplot(regularData, aes(Score)) + geom_histogram(bins = 10,
    aes(y = after_stat(density)), color = "black", fill = "pink") +
    labs(x = "Score", y = "Density", title = "Distribution for Regular section Scores")
hist.regular + stat_function(fun = dnorm, args = list(mean = mean(regularData$Score,
    na.rm = TRUE), sd = sd(regularData$Score, na.rm = TRUE)),
    color = "blue", linewidth = 1.5)
```




```r
# Histogram and Normal Curve for Sports Data
hist.sports <- ggplot(sportsData, aes(Score)) + geom_histogram(bins = 10,
    aes(y = after_stat(density)), color = "black", fill = "grey") +
    labs(x = "Score", y = "Density", title = "Distribution for Sports section Scores")

hist.sports + stat_function(fun = dnorm, args = list(mean = mean(sportsData$Score,
    na.rm = TRUE), sd = sd(sportsData$Score, na.rm = TRUE)),
    color = "Red", linewidth = 1.5)
```




```
# Statistics for Scores from each section
stat.desc(regularData$Score, norm = TRUE)
```

```
##       nbr.val      nbr.null        nbr.na           min           max
##   19.00000000    0.00000000    0.00000000  265.00000000  380.00000000
##         range           sum        median          mean       SE.mean
##  115.00000000 6225.00000000  325.00000000  327.63157895    7.63157895
##  CI.mean.0.95           var       std.dev      coef.var      skewness
##    16.03335241 1106.57894737   33.26528141     0.10153259   -0.07341545
##       skew.2SE      kurtosis      kurt.2SE    normtest.W    normtest.p
##   -0.07008412   -1.08697079   -0.53583906     0.96952149    0.76684878
```

```
stat.desc(sportsData$Score, norm = TRUE)
```

```
##      nbr.val     nbr.null       nbr.na          min          max        range
##   19.0000000    0.0000000    0.0000000  200.0000000  395.0000000  195.0000000
##          sum       median         mean      SE.mean CI.mean.0.95          var
## 5840.0000000  315.0000000  307.3684211   13.3134085   27.9704333 3367.6900585
##      std.dev     coef.var     skewness     skew.2SE     kurtosis     kurt.2SE
##   58.0318021    0.1888021   -0.4193351   -0.4003071   -1.0607215   -0.5228991
##   normtest.W   normtest.p
##    0.9445561    0.3179747
```

4.4.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?

Solution: The Student's grade in the dataset would have influences the point distributions in the plots. Instead of plotting the performance based on the scores which is hard to compare between each section, the Student's grade would have been an easy measure to compare the performane of students.For instance, with the additional data of Grades, we can tell how many students scored A in Regular vs Sports and that can tell how students fared in each section. With just the scores, there is nothing to compare with.

Session info

```
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
## [2] LC_CTYPE=English_United States.utf8
## [3] LC_MONETARY=English_United States.utf8
## [4] 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 qqplotr_0.0.5  psych_2.2.9    ggplot2_3.4.0
##
## loaded via a namespace (and not attached):
##  [1] highr_0.9         formatR_1.12      DEoptimR_1.0-11   pillar_1.8.1
##  [5] compiler_4.2.2    tools_4.2.2       boot_1.3-28       digest_0.6.30
##  [9] nlme_3.1-160      lattice_0.20-45   evaluate_0.18     lifecycle_1.0.3
## [13] tibble_3.1.8      gtable_0.3.1      pkgconfig_2.0.3   rlang_1.0.6
## [17] cli_3.4.1         DBI_1.1.3         rstudioapi_0.14   parallel_4.2.2
## [21] yaml_2.3.6        xfun_0.34         fastmap_1.1.0     withr_2.5.0
## [25] stringr_1.4.1     dplyr_1.0.10      knitr_1.41        generics_0.1.3
## [29] vctrs_0.5.0       grid_4.2.2        tidyselect_1.2.0  robustbase_0.95-0
## [33] glue_1.6.2        R6_2.5.1          fansi_1.0.3       rmarkdown_2.18
## [37] farver_2.1.1      magrittr_2.0.3    MASS_7.3-58.1     scales_1.2.1
## [41] htmltools_0.5.3   assertthat_0.2.1  mnormt_2.1.1      colorspace_2.0-3
## [45] labeling_0.4.2    utf8_1.2.2        stringi_1.7.8     munsell_0.5.0
```