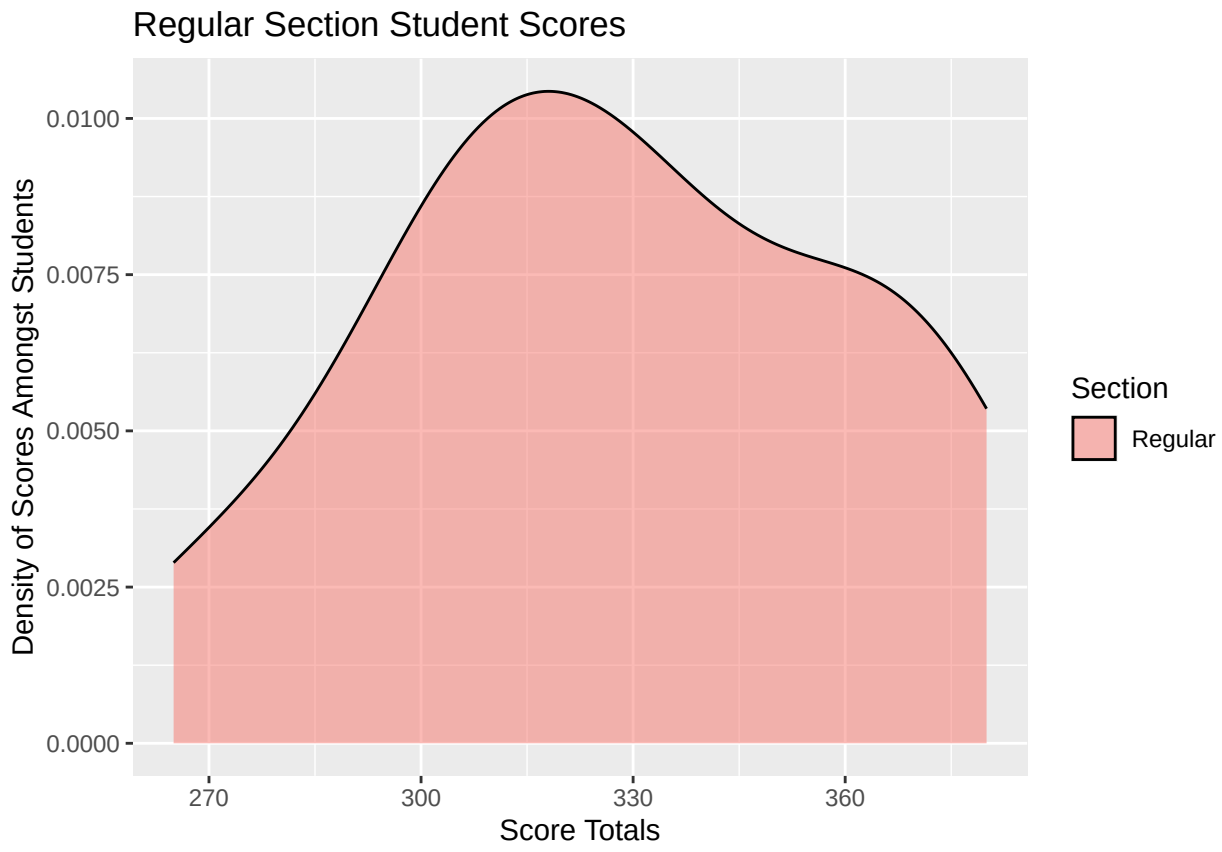
```r
# Test Scores Dataset Assignment
# DSC 520
# Week 4
# Statistics for Data Science Assignment Week 4
# David Berberena
# 1/7/2024

# Assignment Start

# Upload readr library for file importing

library(readr)

# set working directory for smooth file importing

setwd("C:/Users/dbzda/Documents/School/DSC 520 Statistics for Data Science")

# Import the scores.csv file to view its properties

scores <- read.csv("scores.csv")

## 1. What are the observational units in this study?

str(scores)
```

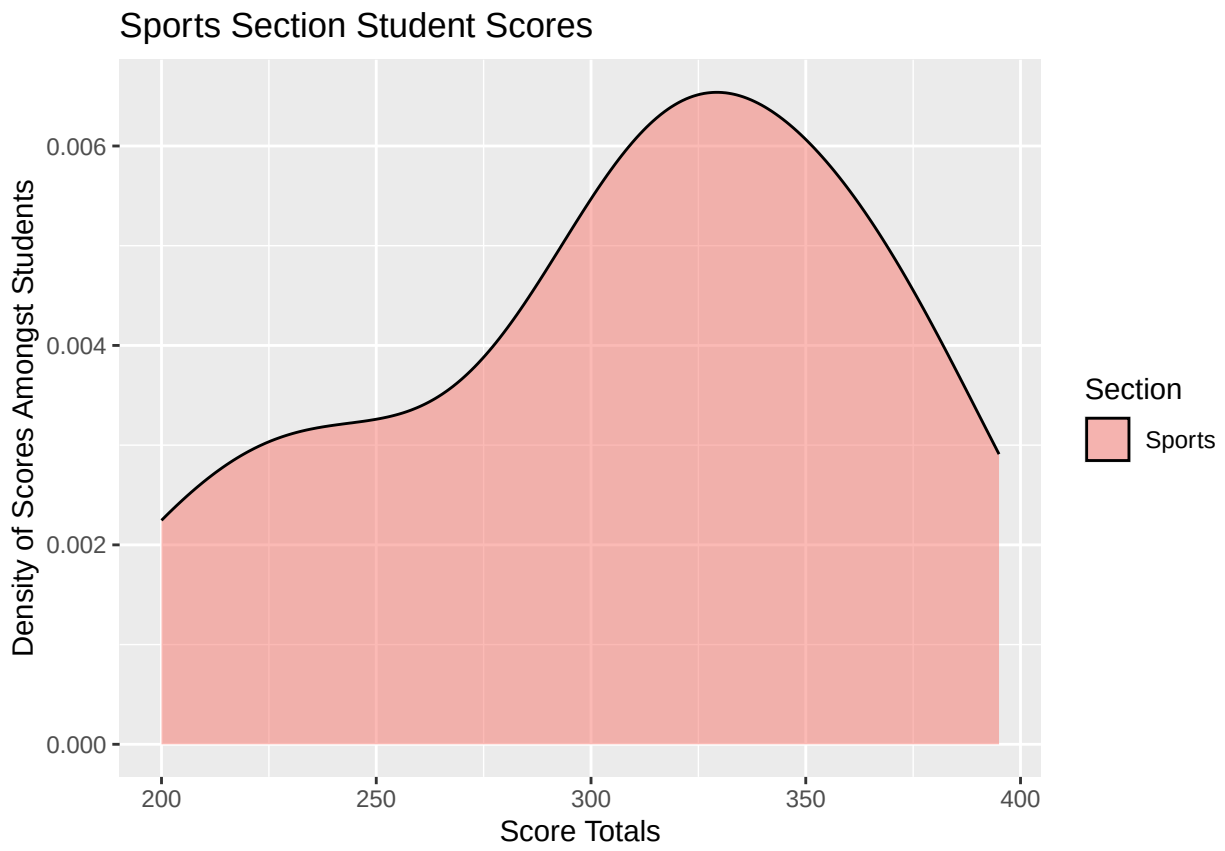
```
## 'data.frame':    38 obs. of  3 variables:
##  $ Count  : int  10 10 20 10 10 10 10 30 10 10 ...
##  $ Score  : int  200 205 235 240 250 265 275 285 295 300 ...
##  $ Section: chr  "Sports" "Sports" "Sports" "Sports" ...
```

```r
# The str() function tells us that the 38 observational units are integers

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

# Looking at the narrative, the variables are scores, which are
# quantitative (int), and section, which are categorical (chr).

## 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.

# This is accomplished using the built-in subset() function
# I rearranged the columns so I could understand the data better

regular_subset <- subset(scores, Section == "Regular",
                         select = c(Score, Count, Section))

sports_subset <- subset(scores, Section == "Sports",
                         select = c(Score, Count, Section))

## 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.
```

```r
# First plot for Regular scores

# Upload ggplot2 package to create plots

library(ggplot2)

ggplot(data = regular_subset, aes(x = Score, fill = Section)) +
  geom_density(alpha = 0.5) +
  ggtitle("Regular Section Student Scores") +
  xlab("Score Totals") +
  ylab("Density of Scores Amongst Students")
```




```r
# Second plot for Sports scores

ggplot(data = sports_subset, aes(x = Score, fill = Section)) +
  geom_density(alpha = 0.5) +
  ggtitle("Sports Section Student Scores") +
  xlab("Score Totals") +
  ylab("Density of Scores Amongst Students")
```




```
## 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.

# Summary statistics for both subsets of data can help to answer this question
# DescTools library is imported to use the sapply() function properly \

library(DescTools)

summary(regular_subset)
```

```
##      Score           Count         Section
##  Min.   :265.0   Min.   :10.00   Length:19
##  1st Qu.:305.0   1st Qu.:10.00   Class :character
##  Median :325.0   Median :10.00   Mode  :character
##  Mean   :327.6   Mean   :15.26
##  3rd Qu.:355.0   3rd Qu.:20.00
##  Max.   :380.0   Max.   :30.00
```

```
sapply(regular_subset, Mode)
```

```
## $Score
## [1] 305 320
## attr(,"freq")
## [1] 2
```

```
##
## $Count
## [1] 10
## attr(,"freq")
## [1] 10
##
## $Section
## [1] "Regular"
## attr(,"freq")
## [1] 19
```

```
summary(sports_subset)
```

```
##      Score           Count         Section
##  Min.   :200.0   Min.   :10.00   Length:19
##  1st Qu.:267.5   1st Qu.:10.00   Class :character
##  Median :315.0   Median :10.00   Mode  :character
##  Mean   :307.4   Mean   :13.68
##  3rd Qu.:350.0   3rd Qu.:15.00
##  Max.   :395.0   Max.   :30.00
```

```
sapply(sports_subset, Mode)
```

```
##    Score    Count  Section
##       NA     "10" "Sports"
```

```
# I can say that the regular section tended to score more points than the sports
# section. The three main measures of tendency (mean and median provided by
# summary() function and mode provided by sapply() function) were all higher for
# regular section than the sports section (there wasn't a mode present for the
# sports section at all). Regarding consistency of the scores, almost every
# summary statistic for the regular section is higher than the sports section
# except for the max value. Looking at the count means for both sections, the
# count mean for the regular section is higher than the sports section, which
# means that there are more people scoring higher grade totals in the regular
# section than in the sports section on average.

## 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.

# No, as indicated by the summary statistics, there were students in the sports
# section that did score higher than students in the regular section. In this
# context, a statistical tendency is the typical value around which data points
# in the dataset center around. This is more of a centralized average concept
# which looks at the dataset as a whole and minimizes the impact of outliers on
# the dataset. Mean, mode, and median are the standard values used to decide
# statistical tendency.

## 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?
```

```
# Another variable that could have influenced the point distribution between the
# two sections is the students' interest in sports related knowledge, news, and
# statistics. If a majority of the students in the sports section do not watch
# or care about sports related issues, they would be at a disadvantage in
# comparison to those students in the regular section. These disadvantaged
# students would perpetually have to sift through the course knowledge based on
# examples that they don't know about or understand and would do worse more
# consistently than those students who may not know much about sports, but be
# presented with only a few sports related examples and many more examples of
# other subjects they may know much better. If a student is highly interested in
# sports and were in the sports section, they would most likely do better than
# someone in the regular section. This variable could also be supported by the
# current dataset due to the maximum score totals being present in the sports
# section and not the regular section, even though the regular section tended
# to score higher at a more consistent level.
```