Assignment: Week 4 Exercise 4.2
Name: Sharma, Dipika
Date: 2020-04-11

1. Test Score

```
score_df <- read.csv("/Users/dipikasharma/R_Projects/DSC520/data/scores.csv")
score_df
# 1. What are the observational units in this study?
dim(score_df)
```

Ans It shows 38 observations and 3 variables.

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

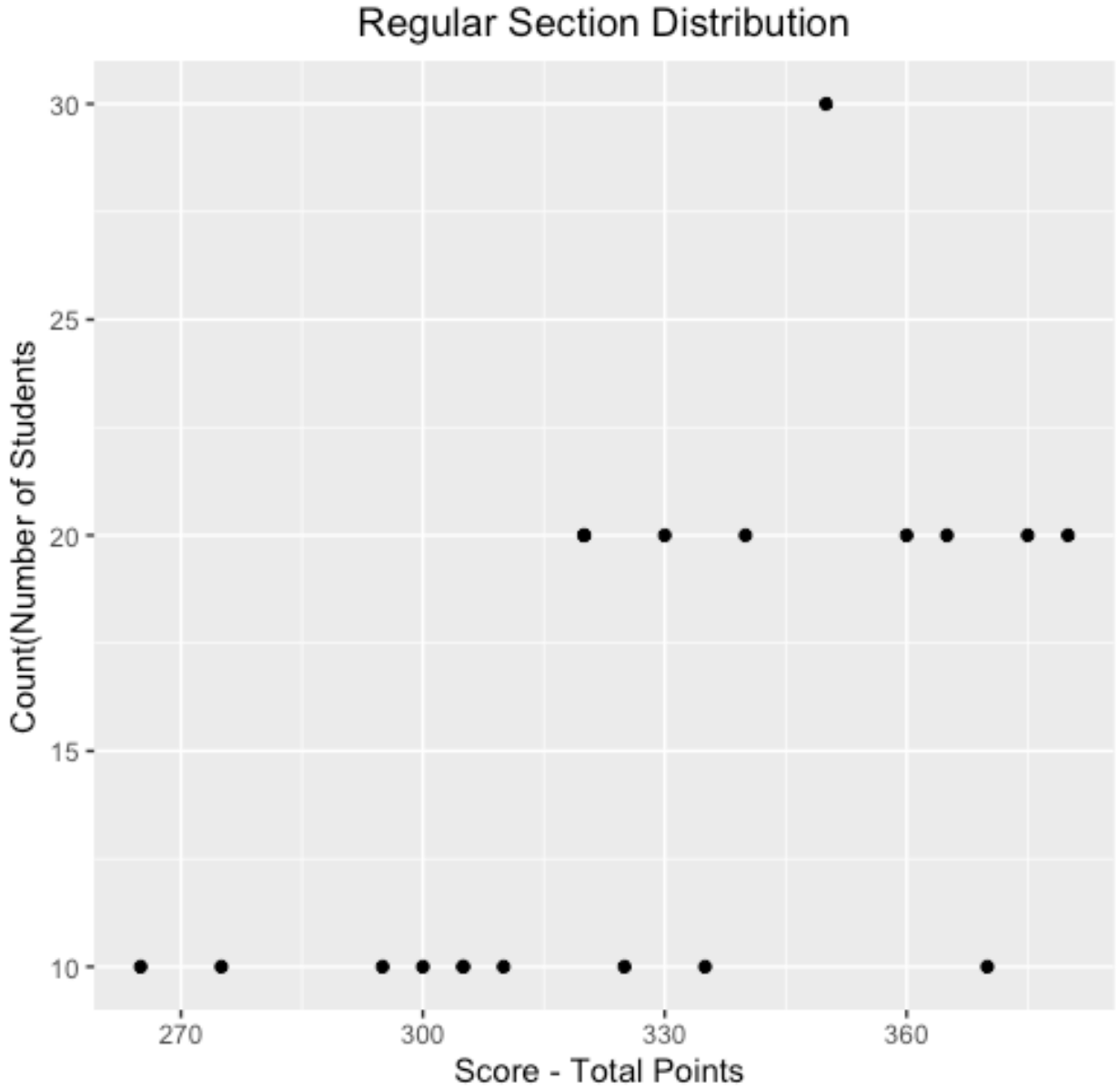
Ans 2. Variables mentioned in the narrative paragraph are section, course grades and total points earned in the course.

- # section is categorical
- # Assuming course grades to be character like 'A', 'B', so it is categorical.
- # Total Points is quantitative

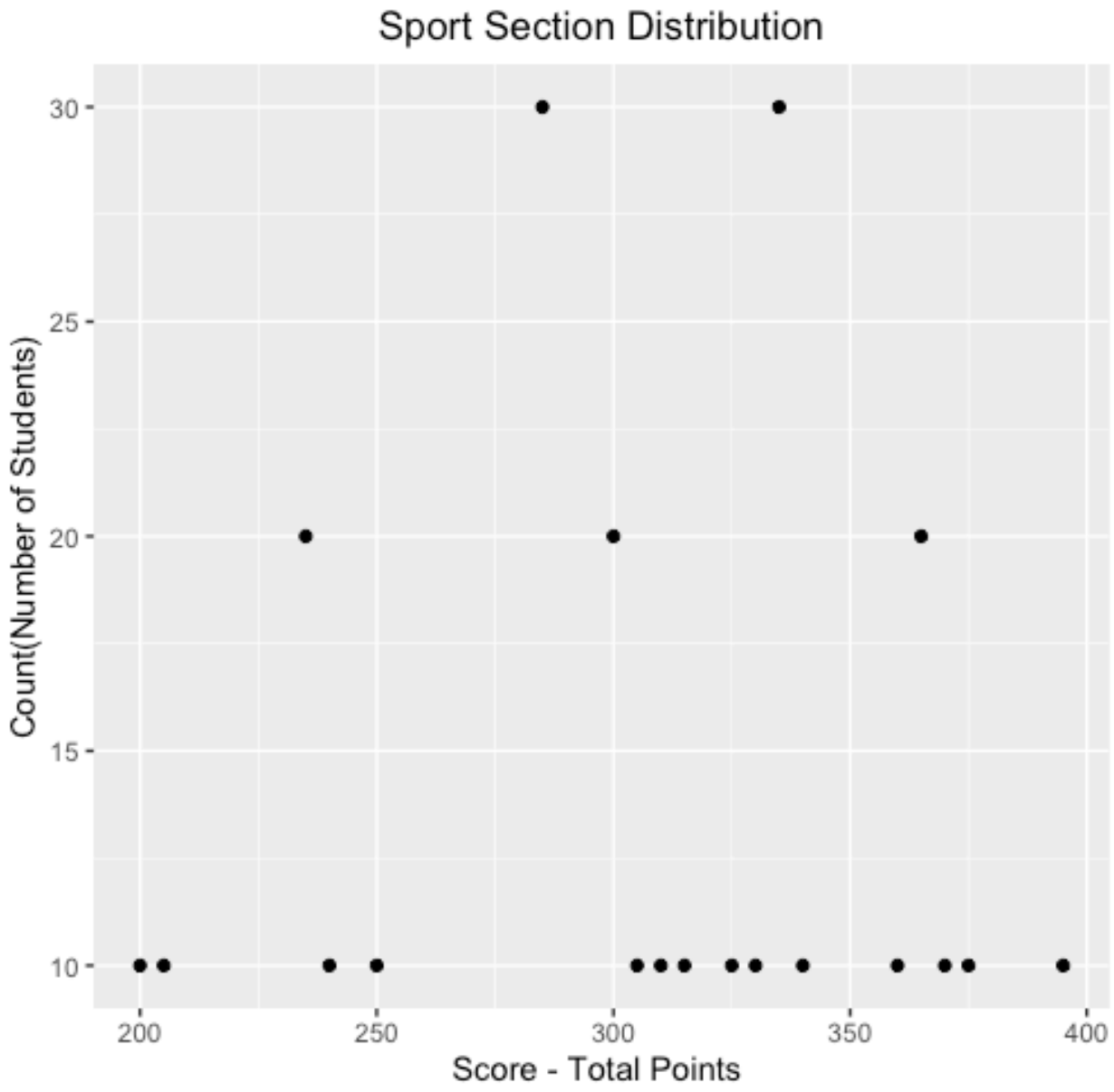
Looking at the score.csv file, I found count, score and section variable.
Where Count and score is quantitative and section variable is categorical.

```
# 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.
Ans
reg_df <- subset(score_df, score_df$Section == "Regular")
reg_df

sport_df <- subset(score_df, score_df$Section == "Sports")
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.

Ans
install.packages("ggplot2")
library(ggplot2)

ggplot(reg_df, aes(x=Score, y=Count)) + geom_point()
ggplot(sport_df, aes(x=Score, y=Count)) + geom_point()
```

```
ggplot(reg_df, aes(x=Score, y=Count)) + geom_point() + ggtitle("Regular Section
Distribution") +
  theme(plot.title = element_text(hjust = 0.5)) + xlab("Score - Total Points ") +
ylab("Count(Number of Students")
```



```
ggplot(sport_df, aes(x=Score, y=Count)) + geom_point() + ggtitle("Sport Section
Distribution") +
  theme(plot.title = element_text(hjust = 0.5)) + xlab("Score - Total Points ") +
ylab("Count(Number of Students)")
```



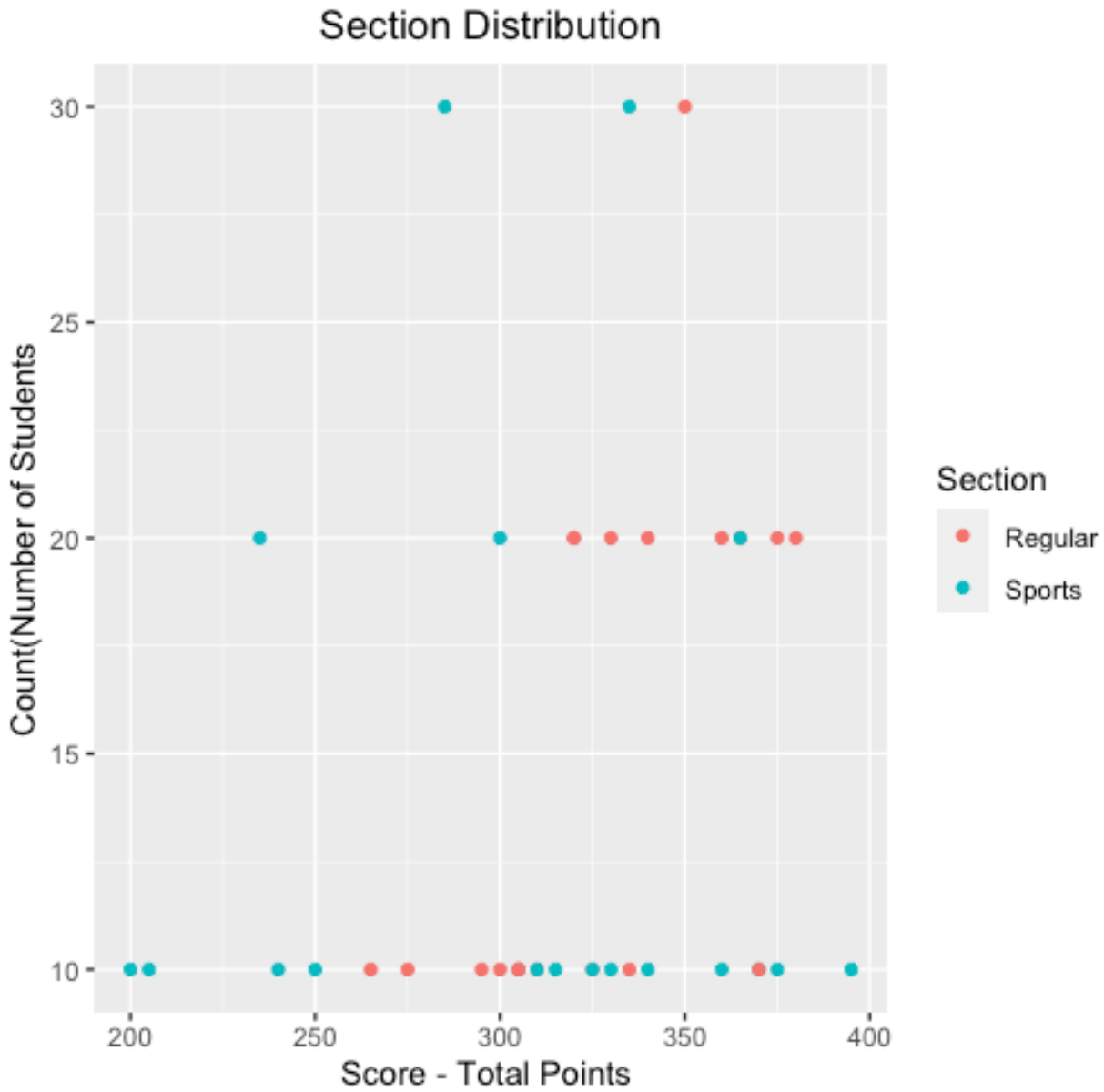

```
# Once you have produced your Plots answer the following questions:
# 4.a. Comparing and contrasting the point distributions between the two section,
```

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.

Ans 4.a. Both regular and sports sections are doing good to score more points, but sport section scored more grade points of 395 whereas regular section scored 380 as highest-grade point. Also, we can see there are 10 students who scored 395 in sports section. where as 20 students in regular section scored 380 grade points.

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.

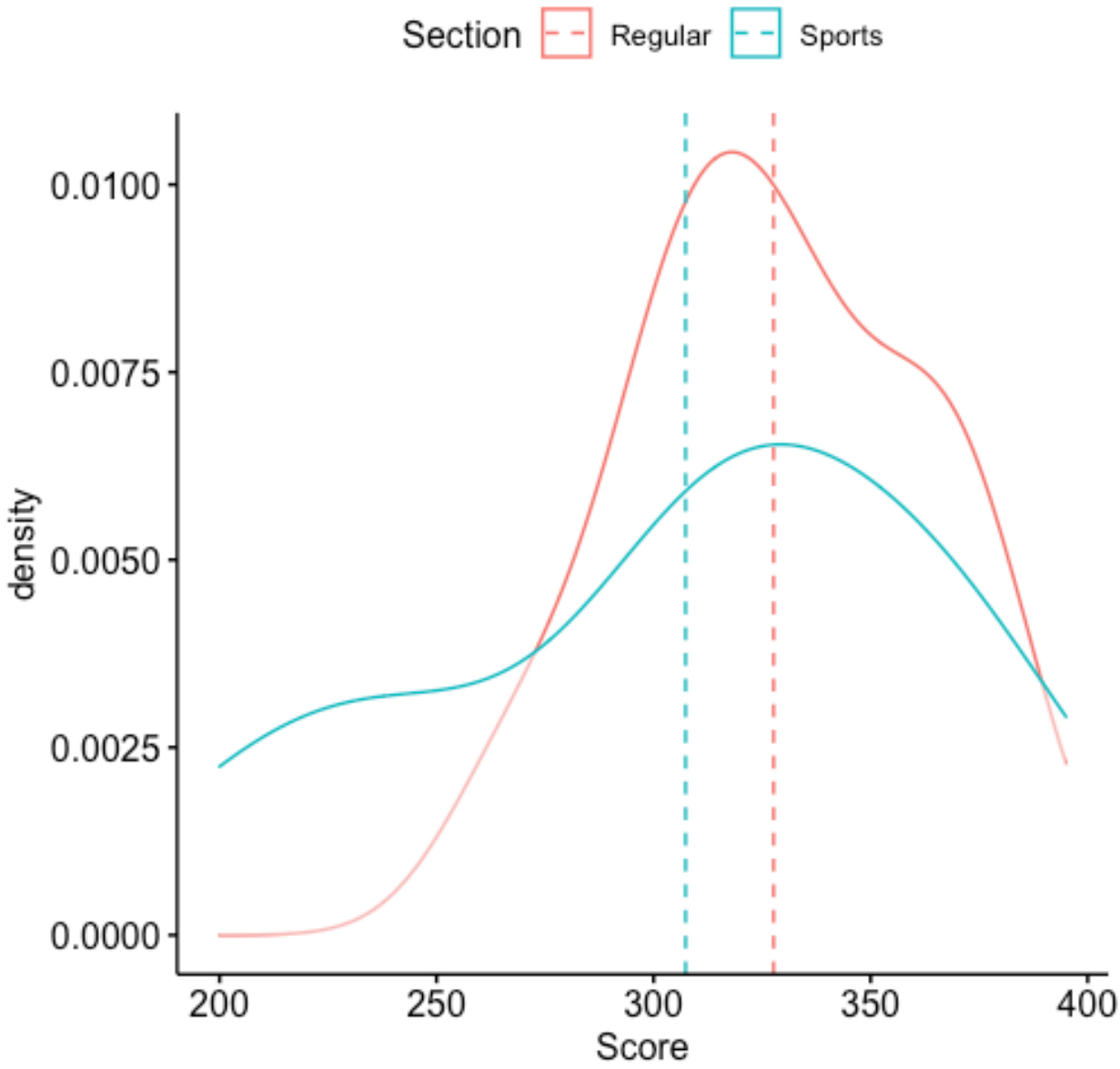
```
ggplot(score_df, aes(x=Score, y=Count, col=Section)) + geom_point() + ggtitle("Section Distribution") +
  theme(plot.title = element_text(hjust = 0.5)) + xlab("Score - Total Points ") +
ylab("Count(Number of Students") +
  stat_summary(fun.data = "mean_sdl", geom = "linerange", colour = "red", size = 2, mult = 1)
```



Ans. We can clearly see in above graph that not every student of one section have scored more point than the students of other section. at some places sport section

student scored highest and, in some places, regular section student scored highest grade points.

```
install.packages("ggpubr")
library(ggpubr)
ggdensity(score_df, "Score", color = "Section") +
  stat_central_tendency(aes(color = Section), type = "mean", linetype = 2)
```



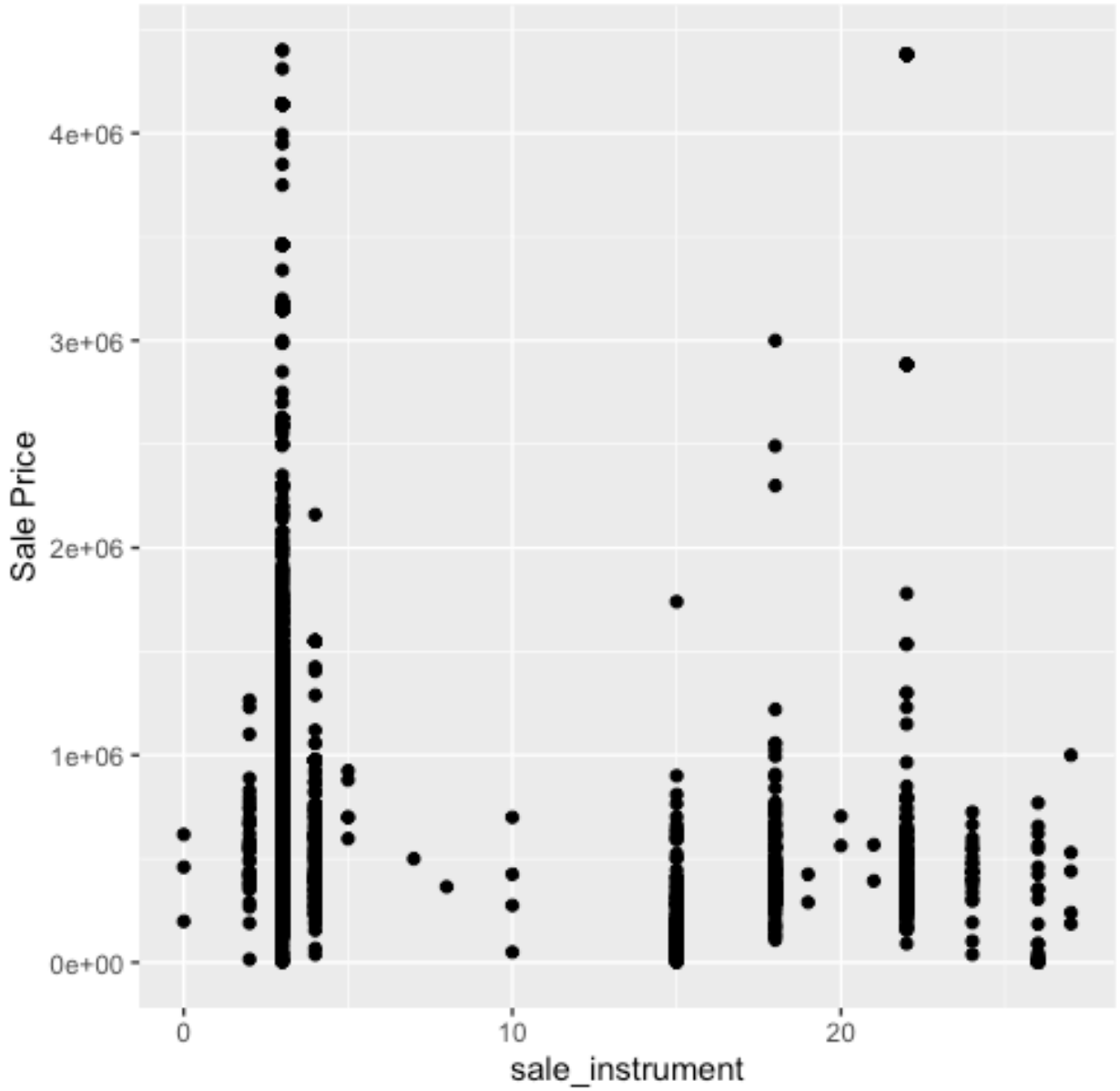

```
#    The plotted graph is showing the mean of regular section in red dashed line
#    and mean of sports section in green dashed line.
```

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?
Ans. The only variable that was not discussed in narrative is count which is the number of students scoring same grade or total points. This variable is important as it is showing how many students scored highest grade point and which section student grade is consistent or which section student scored highest grades.

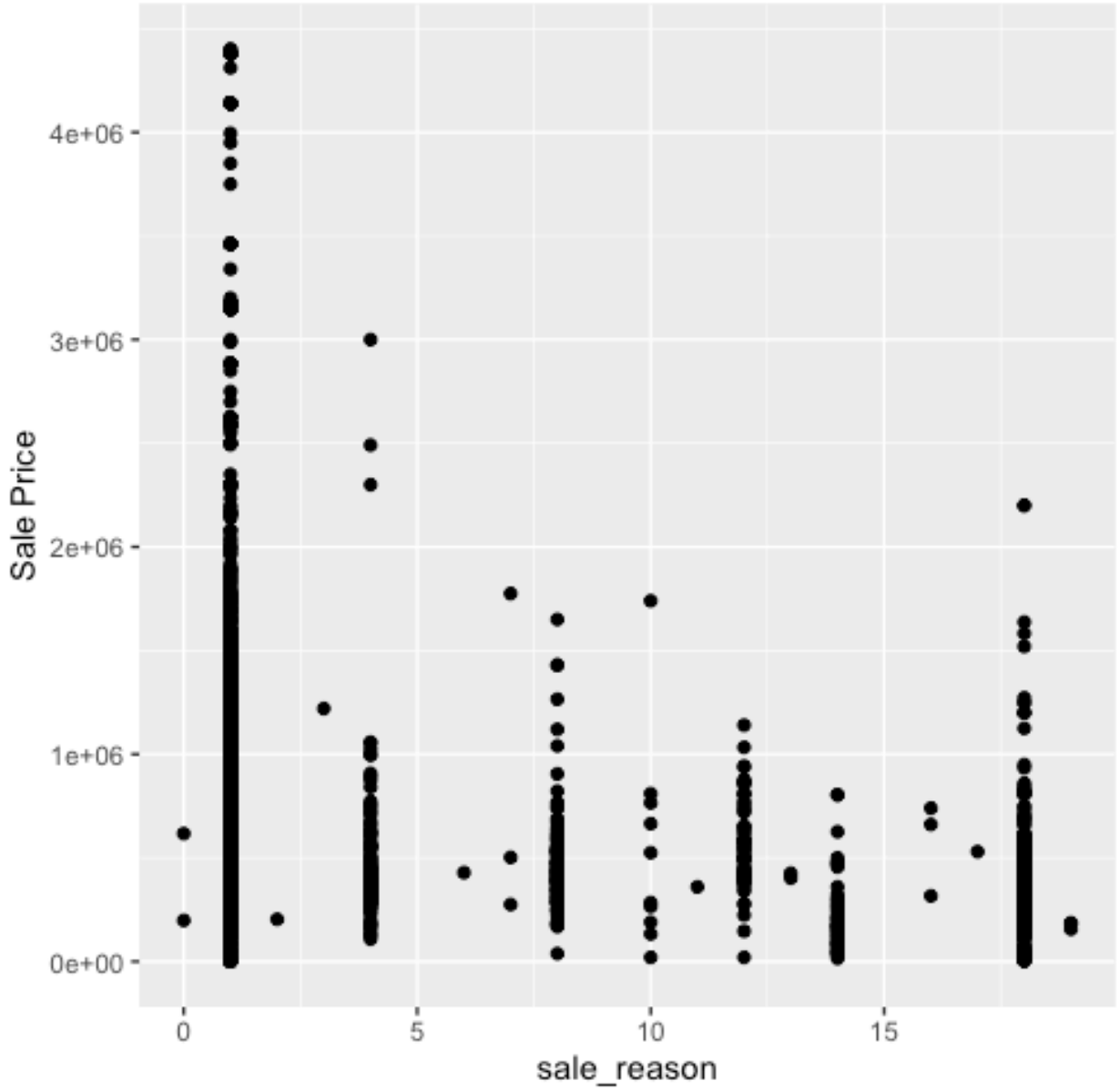
2.

```
install.packages("readxl")
library("readxl")
housing_df <- read_excel("/Users/dipikasharma/R_Projects/DSC520/data/week-7-housing.xlsx")
housing_df
dim(housing_df)

# a. Use the apply function on a variable in your dataset
Ans
apply_saleprice <- apply(housing_df[,2,drop=F],2,sum)
apply_saleprice

# b. Use the aggregate function on a variable in your dataset
Ans
agg_by_SR <- aggregate(housing_df$`Sale Price`,
by=list(housing_df$sale_reason),FUN=sum)
agg_by_SR

agg_by_SI <- aggregate(housing_df$`Sale Price`,
by=list(housing_df$sale_instrument),FUN=sum)
agg_by_SI

# 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
Ans
install.packages("plyr")
library(plyr)

ddply(housing_df, .(housing_df$sale_instrument), transform,
        total.saleprice = sum(`Sale Price`))

# d. Check distributions of the data

library(ggplot2)
```

ggplot(housing_df, aes(x=sale_instrument, y=`Sale Price`)) + geom_point()



ggplot(housing_df, aes(x=sale_reason, y=`Sale Price`)) + geom_point()



e. Identify if there are any outliers

Ans

Plotting sale_instrument vs sale price, we can see for sale_instrument 22, there are two values which lies at some distant compare to other observation data point. Also, for sale_instrument 4 we have increase in sale price, but this does not seem to be an outlier as we see gradual increase in sale price.

Plotted sale_reason vs sale price, I see gradual increase for sale reason 1 and others, I do not see outlies for this plot.

f. Create at least 2 new variables
Ans

```
ddply(housing_df, .(sale_reason), mutate, sumby_SaleReason = sum(`Sale Price`),
    meanby_SaleReason = mean(`Sale Price`))
```