parker\_dsc520\_week4

2022-12-19

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. Use the appropriate R functions to answer the following questions: What are the observational units in this study?

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

sports applications:categorical  
application areas: categorical course grades: quantitative total points: quantitative

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

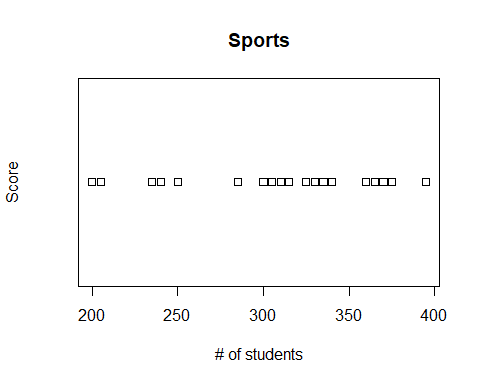
sports = subset(score\_data,Section == "Sports")  
   
regular = subset(score\_data,Section == "Regular")

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:

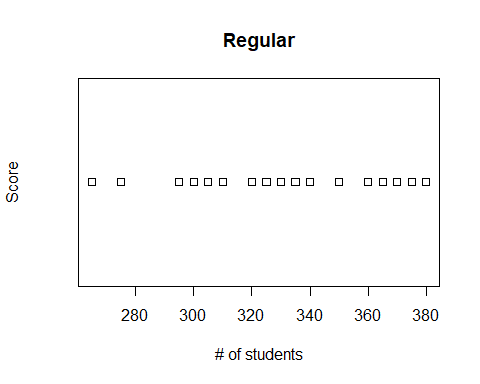
sports\_scores <- sports["Score"]  
regular\_scores <- regular["Score"]  
print(sports\_scores)

## # A tibble: 19 × 1  
## Score  
## <dbl>  
## 1 200  
## 2 205  
## 3 235  
## 4 240  
## 5 250  
## 6 285  
## 7 300  
## 8 305  
## 9 310  
## 10 315  
## 11 325  
## 12 330  
## 13 335  
## 14 340  
## 15 360  
## 16 365  
## 17 370  
## 18 375  
## 19 395

#ggplot(sports\_scores,aes(x=students, y=score)) + geom\_point()  
plot(sports\_scores, xlab="# of students", ylab="Score", main="Sports")



plot(regular\_scores, xlab="# of students", ylab="Score", main="Regular")



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.

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.

What could be one additional variable that was not mentioned in the narrative that could be influencing the point distributions between the two sections?

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: Use the apply function on a variable in your dataset

Use the aggregate function on a variable in your dataset

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

Check distributions of the data

Identify if there are any outliers

Create at least 2 new variables

## R Markdown

summary(cars)

## speed dist   
## Min. : 4.0 Min. : 2.00   
## 1st Qu.:12.0 1st Qu.: 26.00   
## Median :15.0 Median : 36.00   
## Mean :15.4 Mean : 42.98   
## 3rd Qu.:19.0 3rd Qu.: 56.00   
## Max. :25.0 Max. :120.00

## Including Plots

plot(pressure)

