```
# Assignment: Test Scores
# Name: Koby-Hercsky, Theodore
# Date: 2021-04-07

# Use the appropriate R functions to answer the following questions:
View(scores)

# 1. What are the observational units in this study?
summary(Test_Score)
Test_Score
# Answer: observational units in this study are the course grades and the total points to compare them to the two sections to determine the students performance.

# 2. Identify the variables mentioned in the narrative paragraph and determine which are categorical and quantitative?
# Answer: The two variables Count and Section are known as categorical and Score is quantitative

# 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.
Regular_Section <- scores[scores$Section == "Regular",]
Sports_Section <- scores[scores$Section == "Sports",]

# 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:
#Answer: plot function shows both regular and sports class section in different colors
ggplot(scores, aes(x = Count, y = Score, shape = 'color'))+geom_point(aes(color = Section))+theme(legend.position = "bottomright")+labs(title = "Legend")+labs(title = "Test Score Analysis")

# 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.
# Answer: I will say that in the comparison in our plot function we see that in the Sports class we have higher scores with a lower count of student than we are seeing with the Regular class section. As we see a higher count in the 20s of students that received higher scores.
# Answer: In conclusion Regular students had higher count of students that received higher scores than the Sports section.

# 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.
# Answer: No every student in either section did not score more points than every student in the other section. As the mean in the regular section is 327.6 and the mean in sports is 307.4.

# 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?
# Answer: Another variable that could influence the point between the two sections would be the time frame of the courses as one semester for the Regular section could have had one additional week while the Sports Section had one less week which would cause a difference in scores
```

```
# Assignment: Housing Data Set
# Name: Koby-Hercsky, Theodore
# Date: 2021-04-07

#Use the apply function on a variable in your dataset
apply(Housing_data, 2, median)
#Answer: I used the apply function for my data frame Housing_data to extract each column as a vector that goes to the median that goes by the column.

#Use the aggregate function on a variable in your dataset
aggregate(Housing_data$Sale.Price ~ Housing_data$bedrooms , Housing_data, mean)
#Answer: used the aggregated function to find the mean price by the amount of bedrooms

#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
# Used the ddply function to split the housing data by the zip code and calculating the useless square feet of the lot that can not be lived in for each address.
# Then combined the results back into the data set in order by each zip code
housing_unused_square_feet_by_zip <- ddply(Housing_data, .(zip5), transform,useless_sq_ft = sq_ft_lot - square_feet_total_living)
# As seen in this subset we have 40 observations for the zip code 98052
subset(housing_unused_square_feet_by_zip, zip5 == "98052")

#Check distributions of the data
ggplot(housing_unused_square_feet_by_zip, aes(x = square_feet_total_living, y = Sale.Price, shape = 'color'))
+geom_point(aes(color = zip5))+theme(legend.position = "bottomright")+labs(title = "Legend")+labs(title = "Sale Price by Square Foot")
# The distribution in this plot suggests that most houses on this data set stayed in the square feet of 1,000 to 6,000 in square feet of living space and proceeded in a 90 degree angle.
# Meaning the distribution of this data set shows an upward momentum that states as the square feet increases the sales price goes up

#Identify if there are any outliers
ggplot(housing_unused_square_feet_by_zip, aes(x = square_feet_total_living, y = Sale.Price, shape = 'color'))
+geom_point(aes(color = zip5))+theme(legend.position = "bottomright")+labs(title = "Legend")+labs(title = "Sale Price by Square Foot")
# In this plot we see that we do have outliers that are located to the left and right of our plots as the square feet of total living goes up so should the price.
# Which the outliers are seen to be when the square feet goes up the price stays low or if the square feet is low the price sky rockets.
# These outliers are due to smaller houses being more expencive and in better areas or bigger houses that are in bad areas and need alot of work being why they are so cheap and an outlier.

#Create at least 2 new variables
# added new variable to calculate the price per square foot by using the mutate function
mutate(Housing_data, price_per_square_foot = Sale.Price / square_feet_total_living)
# added new variable to calculate the price per bedroom by using the mutate function
mutate(Housing_data, price_per_bedroom = Sale.Price / bedrooms)
```