1. Test Scores
   1. 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 (**Score**) in the course. You will need to import the [Scores.csv](http://content.bellevue.edu/cst/dsc/520/id/resources/scores.csv) dataset that has been provided for you.
      1. Use the appropriate R functions to answer the following questions:
         1. What are the observational units in this study?

Score

* + - 1. Identify the variables mentioned in the narrative paragraph and determine which are categorical and quantitative?

Score – Quantitative

Section –Categorical

* + - 1. 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.sub <- scoresdata[scoresdata$Section == "Regular", ]

sports.sub <- scoresdata[scoresdata$Section == "Sports", ]

* + - 1. 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:
         1. 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.

By only looking at the output of the plot function, I cannot say for certain what section tented to score more points than the other. We can see that the only students who scored 395 points (highest achieved in both sections) where in the Sports section. However, we can also see that the only students who scored 200 points (lowest achieved in both sections) where in the Sports section as well. These plots do not help us understand the likely hood of a student achieving a higher score than the mean of both courses.

Chart, scatter chart

Description automatically generated

Chart, scatter chart

Description automatically generated

* + - * 1. 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.

Every student in one section did not score more points than every student in the other section. Statistical tendency in this context would mean that the students in one class would be more likely to have higher scores than those in the other class (referring to the center of the distribution or mean of the scores). With that said, we ran a Shapiro-Wilk normality test to check the normalcy of each distribution. Both groups came in with a p>0.05, indicating that the distribution is normal. Since we can determine that the distribution is normal, we can test the probability that a student will scorce above or below the mean in each group (mean of the Sports class is 307 and the mean of the Regular class is 337).

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 be influencing the point distributions between the two sections may be the time of the two classes. Since both classes are taught by the same professor each would occur at a different time in the day. Is one at 8AM and the other at 3PM? If there is a large time gap between the two, this may impact some student’s performance. However, this may be less important if the class is taught on-line rather than in person. If both classes are taught on-line, other variables to consider could be age or gender of the students. However, we must be cautious using these variables as well since they may heavily correlate with the class section (e.g. Older Males may tend to choose the sports section over the regular section).

1. We interact with a few datasets in this course, one you are already familiar with, the [2014 American Community Survey](http://content.bellevue.edu/cst/dsc/520/id/resources/acs-14-1yr-s0201.csv) and the second is a [Housing dataset](http://content.bellevue.edu/cst/dsc/520/id/resources/10-week-housing-data/week-6-housing.xlsx), 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:
   1. Use the apply function on a variable in your dataset

Since the apply function cannot be used for one-dimensional data, we must first separate our variable into it’s own dataframe “maxPrice” then use the apply function to find the max sale price.

# Loading the data set week-6-housing

> housing <- read.csv(file = 'R Data/week-6-housing.csv')

# Creating a matrix based on the variable chosen

> dfhousingprice <- data.frame(housing$Sale.Price)

# Using the apply function to find the max housing price

> maxPrice <- apply(dfhousingprice, 2, max)

> maxPrice

**housing.Sale.Price**

**4400000**

* 1. Use the aggregate function on a variable in your dataset

> mean\_pirce <- aggregate(Sale.Price ~ zip5, data = dfhousing, mean)

> mean\_pirce

**zip5 Sale.Price**

**1 98052 649375.4**

**2 98053 672623.7**

**3 98059 645000.0**

**4 98074 951543.8**

* 1. 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

# Splitting out all years built with a sale price above the mean and summing the count of homes falling in that category for that year

yearsabovemean <- subset(count(dfhousing, "year\_built"), dfhousing$Sale.Price > mean(dfhousing$Sale.Price))

yearsabovemean

# Splitting out all years built with a sale price below the mean and summing the count of homes falling in that category for that year

yearsbelowmean <- subset(count(dfhousing, "year\_built"), dfhousing$Sale.Price < mean(dfhousing$Sale.Price))

yearsbelowmean

# Joining the two data frames back together. Giving us a view of the number of homes built and selling above/below the average of all years

yearscombined <- join(yearsabovemean, yearsbelowmean, by = NULL, type = "full", match = "all")

yearscombined

* 1. Check distributions of the data

To check the distributions of the data, we can start by looking at a probability plot of **Sale.Price**. From the below plot, we know that Sale Price is not a normal distribution.

Chart, line chart

Description automatically generated

We see a similar pattern when looking at a probability plot of **square\_feet\_total\_living**.

Chart, line chart

Description automatically generated

Are the two correlated? We can check this visually by plotting both **Sale.Price** and **square\_feet\_total\_living.** We can see here that the two variables are certainly related but there appear to be a few outliers from both variables. However, more analysis is needed before determining whether or not to exclude any strange looking data points.

Chart, scatter chart

Description automatically generated

Another way we can quickly check the distributions of the data is by using ggplot to plot histograms of all numeric variables. As we can see, some of these plots are not so helpful (e.g. lat and ion). However, many do help us see how the data is distributed at a high level (e.g. many more home in the data set where built in the 2000s).

A picture containing diagram

Description automatically generated

* 1. Identify if there are any outliers

After looking at the above distributions and correlation plots. We can identify a few outliers in the **Sale.Price** and **square\_feet\_total\_living plot.** However, maybe the high price and low square footage is actually due to another variable (hosing grade) and that is what is bringing down the price (not the squarefootage). So we also pot out price and **housing grade** (shown in the below scatter plot). Now we see a couple more data points that stand out. Why is the one a the lower right (highlighted in yellow) priced so high but graded so low? Before we identify this an outlier, we need to look into the other variables affecting it’s placement on this chart. Maybe this house has a lot more land than the other homes, bringing up the price.

Chart, scatter chart

Description automatically generated

* 1. Create at least 2 new variables

When looking for a new house, we like to know how much we are paying per square foot. So I will first create a price per square foot variable and add it to the housing data frame.

piceperfoot <- (dfhousing$Sale.Price/dfhousing$square\_feet\_total\_living)

cbind(dfhousing, piceperfoot)

Another important factor to consider when buying a home is how many bathrooms per bedrooms there are. For this we will create a bathperroom variable and add it to the housing data frame.

bathperroom <- ((dfhousing$bath\_full\_count+ bath\_half\_count)/bedrooms)

cbind(dfhousing, piceperfoot)