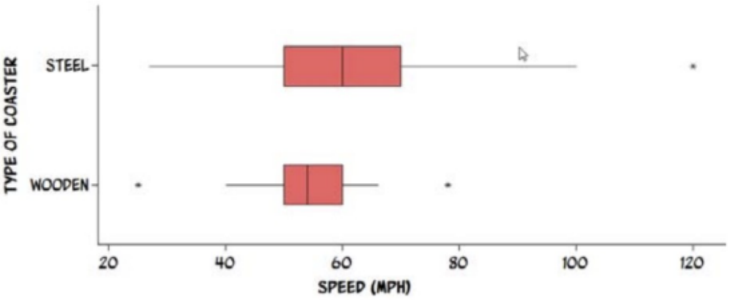
Kyle Thomas

Math 341 Spring 2021, Homework 4

1. The Roller Coaster Database maintains a web site (www.rcdb.com) with data on roller coasters around the world. Some of the data recorded include whether the coaster is made of wood or steel and the maximum speed achieved by the coaster, in miles per hour. The box plots display the distributions of speed by type of coaster for 145 coasters in the United States, as downloaded from the site in November of 2003.



(a) Do these box plots allow you to determine whether there are more wooden or steel roller coasters?

No, this box plot shows the speed of the steel vs wooden coasters.

(B) Do these box plots allow you to say which type has a higher percentage of coasters that go faster than 60 mph? Explain and, if so, answer the question.

Yes, for the steel coasters, 50% of them go over 60 mph. For the wooden coaters, less than 75% of them go over 60 mph.

(c) Do these box plots allow you to say which type has a higher percentage of coasters that go faster than 50 mph? Explain and, if so, answer the question.

No, because for both of them over 25% go faster than 50mph, but looking at the chart more, the steel coasters go much faster than 50 mph, where the wooden ones do not go much faster past 50mph.

(d) Do these box plots allow you to say which type has a higher percentage of coasters that go faster than 45 mph? Explain and, if so, answer the question. Hint: Think twice on this one.

No, they are both greater than 25% that go faster than 45 mph.

(e) Which type of coaster has more "outliers"? Explain how you are deciding.

The wooden coasters have more outliers. The separated dots from the lines represent outliers. The wooden coasters have two, while the steel has one, even though it is a much greater outlier.

(f) Conjecture as to how the mean, median, interquartile range, and standard deviation will change (if at all) if the faster steel coaster (Top Thrill Dragster in Cedar Point

Amusement Park, Sandusky, Ohio) is removed from the data set. Explain your reasoning

The mean and potentially the standard deviation will change the most, because outliers have a heavy sway on both categories. However the median and interquartile range may have a slight change, or no change in all depending on the amount of similar data.

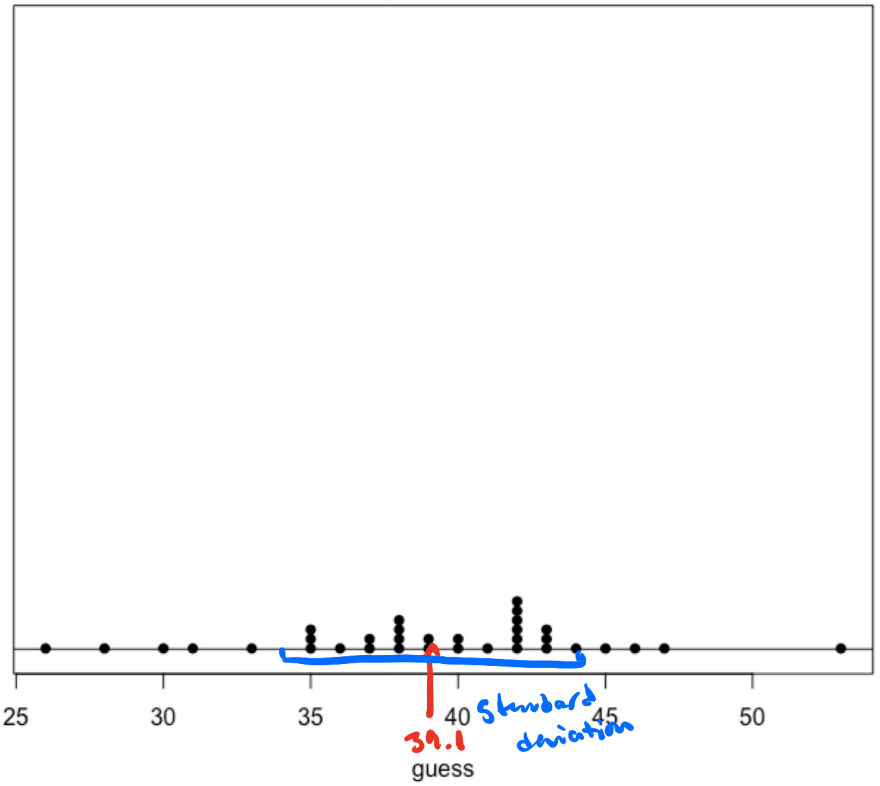
2. A random sample of 10 containers of a particular sun block was selected and their weights were: 10.2, 9.7, 10.1, 10.3, 10.1, 9.8, 9.9, 10.4, 10.3, and 9.8 ounces. Based on the data, would you say the average content of containers of this particular sun block is 10 ounces? Why or why not?

Yes, just looking at the data set, all of the numbers are roughly 10.

3. This problem is about data given in the file ageguesses.csv, which contains the guesses of a volleyball coach's age by the players she currently coaches.

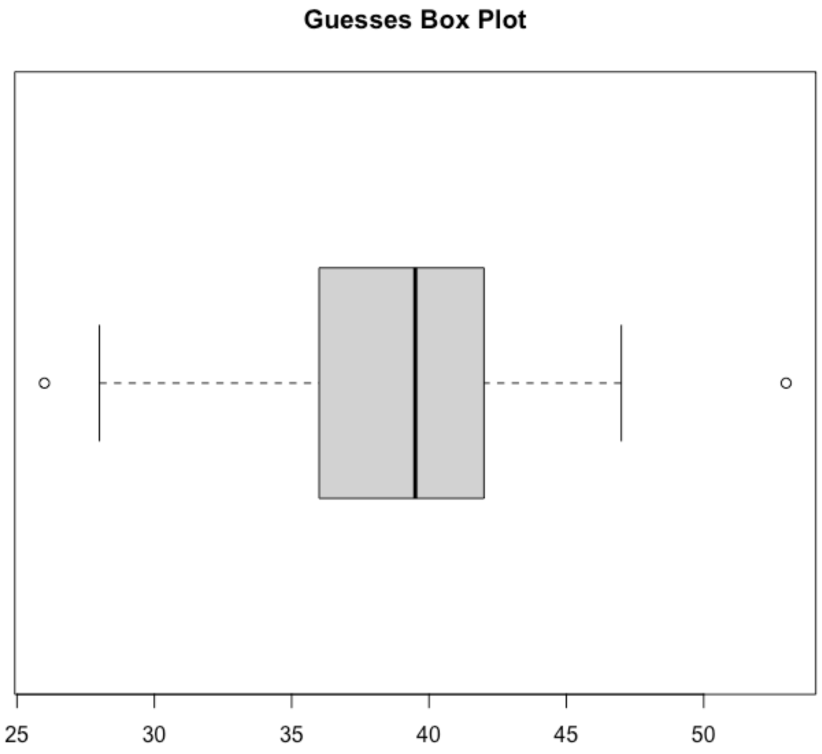
(a) Create a dot plot to illustrate the data.

(b) Find the sample mean *xbar* and sample standard deviation *s* of the data. Label the mean on the dot plot. Also sketch vertical lines indicating the region that is within one sample standard deviation of the sample mean (*xbar* +- *s*)



(c) Create a box plot and list the 5 number summary of this data.

5 number summary: 26 / 36.25 / 39.5 / 42 / 53



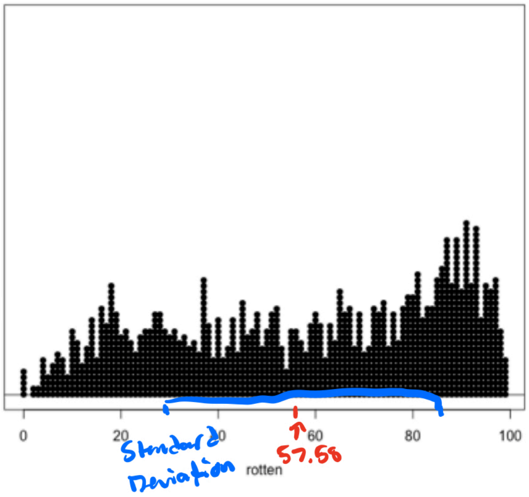
(d) Are there any outliers? If so, list them here and show how to compute them by hand.

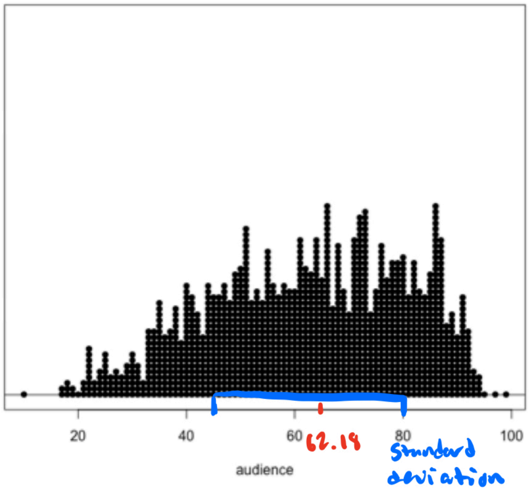
Yes, 26 and 53. To compute them, first you would find the IQR which is Q3-Q1 = 42 - 36.25 = 5.75. The lower outliers will be below Q1 - 1.5 * IQR. The greater outliers will be Q3 + 1.5 * IQR. For this example my lower outliers will be below 27.625, therefore 26 is an outlier. My greater outliers will be greater than 50.625, there for 53 is an outlier.

4.

The data set Hollywood Movies (HollywoodMovies.csv) contains data on movies released in Hollywood between 2012 and 2018. It includes observations of variables such as RottenTomatoes (critics' ratings), AudienceScore (Audience rating via Rotten Tomatoes), Year the movie was released, and Budget, which is the production budget (in millions of dollars).

(a) Create dot plots to compare RottenTomatoes (the critics' ratings) and AudienceScore (they audience ratings via Rotten Tomatoes). How do they compare? Which has the largest mean? Which is more spread out? Compute the sample means and sample standard deviations for each and label them on the dot plots.



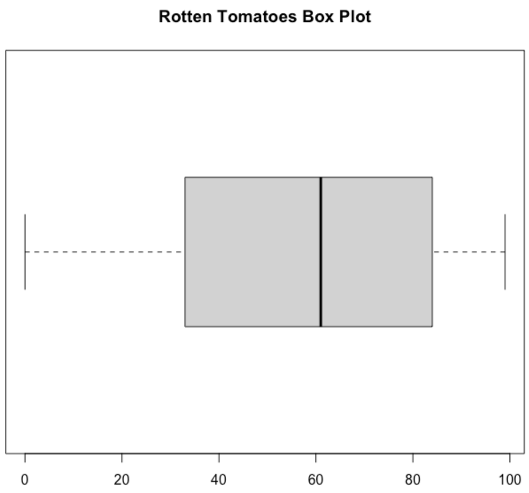


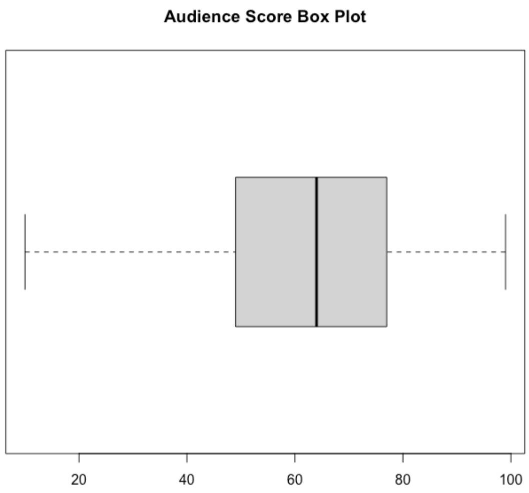
From the dot plots, the Rotten Tomatoes ratings are a lot more spread out from the audience scores. The audience is more likely to give better scores either because they do not watch as many movies or they do not pay as much attention as the Rotten Tomatoes critics. The Rotten Tomatoes critics are more likely to give worse scores, because they have watched a lot more movies and have a much larger sample of movies that they have watched versus the most if not all of the audience.

Rotten Tomatoes:
Mean - 57.58
Standard Deviation - 28.06

Audience Scores:
Mean - 62.18
Standard Deviation - 18.21

(b) Create box plots to compare RottenTomatoes (the critics' ratings) and AudienceScore (they audience ratings via Rotten Tomatoes). How do they compare? Which has the highest median? Which is more spread out? Are there any outliers? Is the explanatory variable categorical or quantitative? Is the response variable categorical or quantitative?





The two box plots appear very similar, however the Rotten Tomatoes has a much lower Q1. The audience score has a slightly higher median at 64, versus Rotten Tomatoes 61. Rotten Tomatoes is more spread out based on their lower scores ranging from 0 to 99 while the audience scores only range form 10 to 99. There are no outliers in the data set. The explanatory variable would be categorical while the response variable would be quantitative.

(c) Do you think the dot plots or box plots explain the data better?

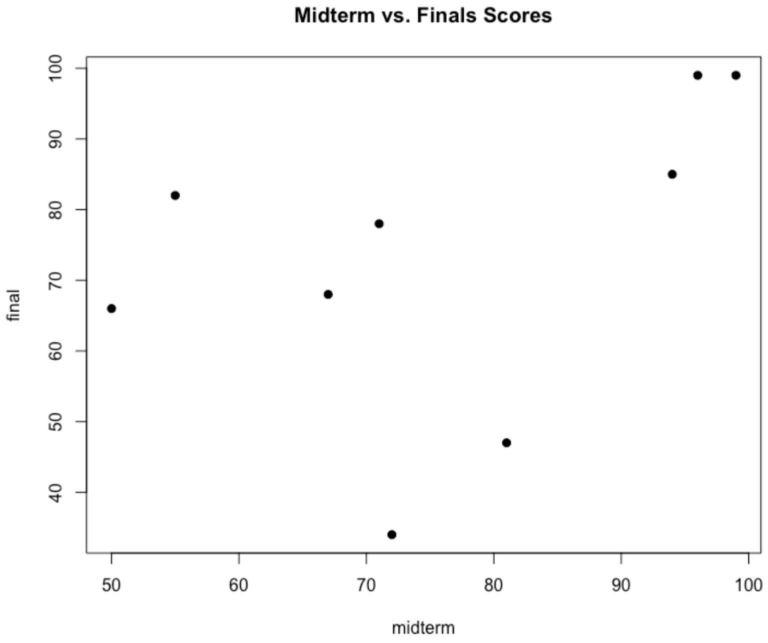
I think the dot plot explains the data better. For me it is a lot easier to see all of the data and how it spreads out on the dot plot more than the box plot.

5. The grades of a class of 9 students on a midterm exam (x) and the final exam (y) are as follows:

x	55	50	71	72	81	94	96	99	67
y	82	66	78	34	47	85	99	99	68

(a) Create a scatterplot to illustrate the data. Which will you use as the explanatory versus response variable?

Kyle Thomas

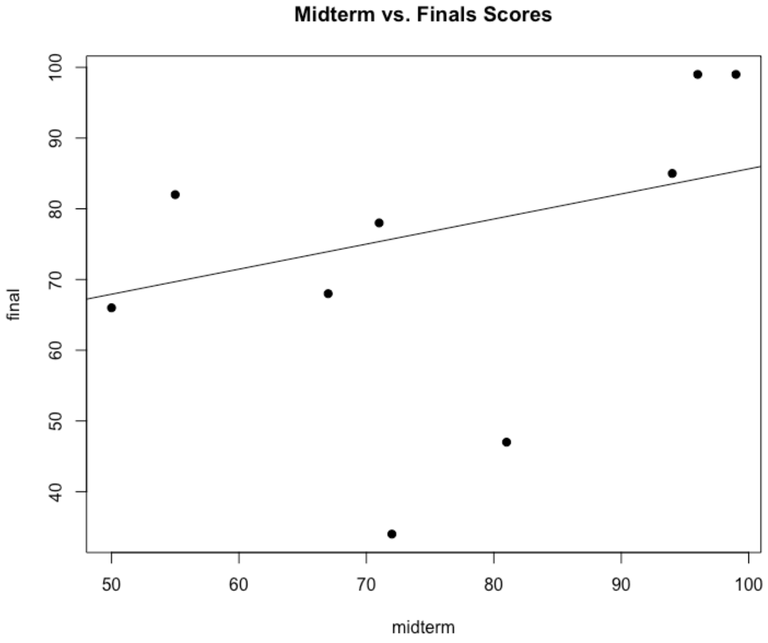


The midterm scores would be the explanatory verses the final scores being the response variables.

(b) First estimate and then compute the correlation coefficient.

I estimated the correlation coefficient would be about .5, after calculating I found the correlation coefficient equal to .4397.

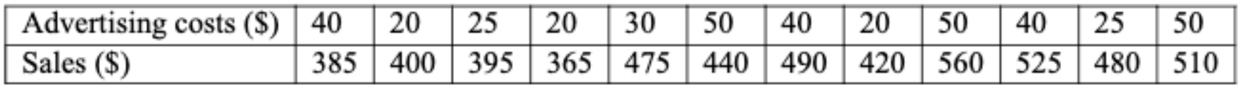
(c) Find and graph the regression line.



(d) Use the line of regression to estimate the final exam grade of a student who got 85 on the midterm exam.

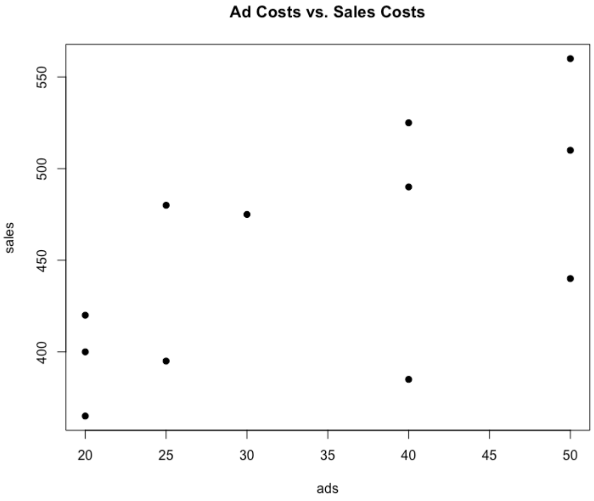
The student who got an 85 on the midterm, would get about an 80 on the final.

6. An entrepreneur sells clothes she makes on Etsy. She conducted a study to determine the relationship between her weekly advertising expenditures and sales. The following is her data:



Advertising costs ($)	40	20	25	20	30	50	40	20	50	40	25	50
Sales ($)	385	400	395	365	475	440	490	420	560	525	480	510

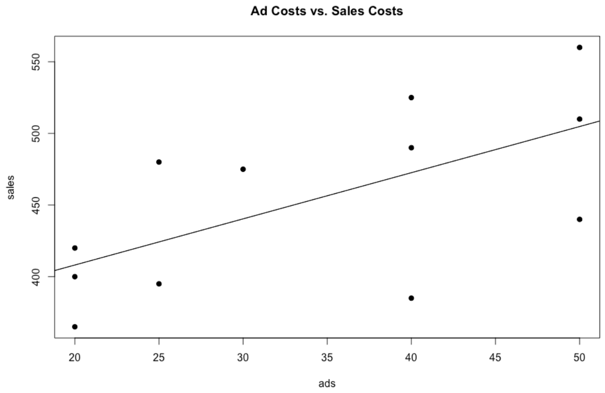
(a) Create a scatterplot to illustrate the data.



(b) First estimate and then compute the correlation coefficient.

I estimate the correlation coefficient to be about .7. After computing, the correlation coefficient is .6348.

(c) Find and graph the regression line.



(d) Use the line of regression to estimate the weekly sales when advertising costs are $35.

When the ad costs are $35, the weekly sales are about $460.