PID: ____

Last Name, First Name: __________________________________

Section: ____

Approximate time spent to complete this assignment: ______ hour(s)

Homework 1
Math 11, UCSD, Winter 2018
Due on Tuesday, 16th January

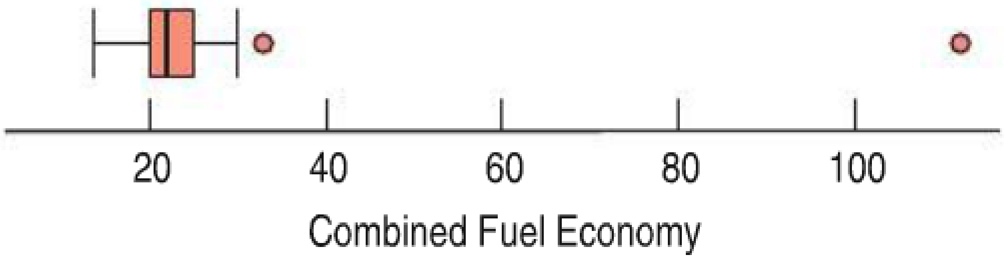
Readings: Chapter 1, Chapter 2 (2.1 and the section on Simpson's paradox on page 31), Chapter 3 (you may skip the section on stem and leaf displays and dotplots), Chapter 4 (sections 4.1-4.3), Chapter 5 (sections 5.1-5.2), and Chapter 6 (but skip Kendall's Tau, Spearman's Rho, and Straightening Scatterplots)

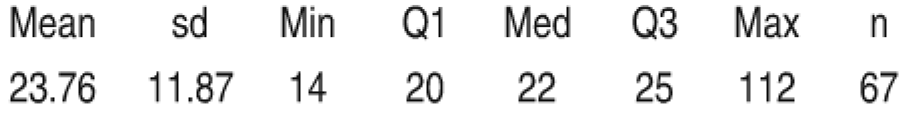
Exercise 1

After entering test scores from her statistics class of 15 students, the instructor calculated the mean and median scores. Upon checking, she discovered that she had entered the top score as 89 when it should have been 98. When she corrects the score, how will the mean and median be affected?

Exercise 2

The boxplot shows the fuel economy ratings for 67 model year 2012 subcompact cars. Some summary statistics are also provided. The extreme outlier is the **Mitsubishi i-MiEV**, an electric car whose electricity usage is equivalent to 112 miles per gallon.



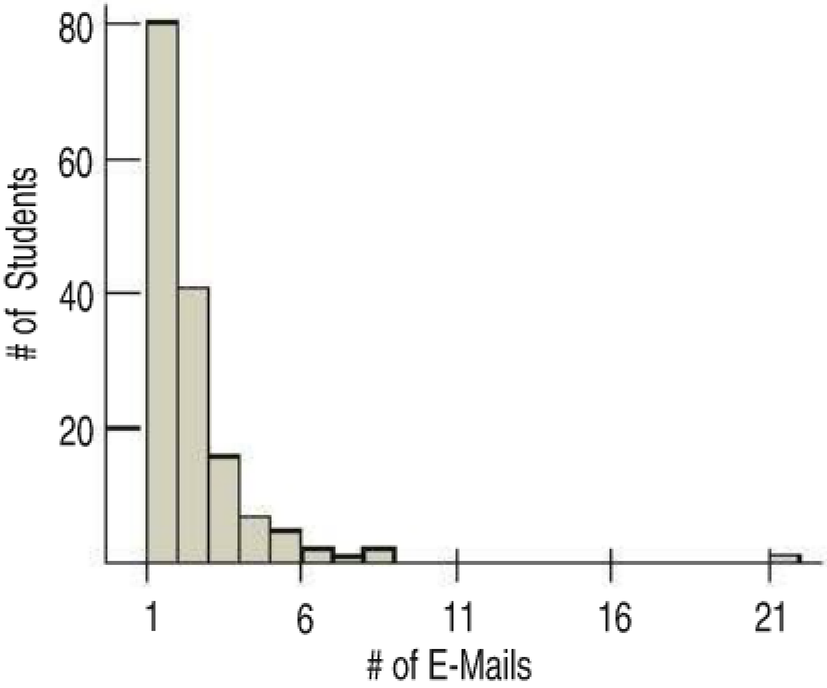


Mean	sd	Min	Q1	Med	Q3	Max	n
23.76	11.87	14	20	22	25	112	67

If the electric car is removed from the dataset, how will the standard deviation be affected?

Exercise 3

A university teacher saved every email received from students in a large Introductory Statistics class during the entire term. he then counted, for each student who had sent him at least one email, how many emails each student had sent.



(a) From the histogram, would you expect the mean of the median to be larger? Explain.

(b) Write a few sentences describing this distribution (shape, center, spread, unusual features).

(c) Which summary statistics would you choose to summarize the center and spread in these data? Why?

Exercise 4

Here are the annual numbers of deaths from tornadoes in the United States from 1998 through 2013 (Source: NOAA):
130, 94, 40, 40, 555, 54, 35, 38, 67, 81, 125, 21, 45, 553, 70, 54
Find these statistics:

(a) mean

(b) median and quartiles

(c) range and IQR

Exercise 5

During contract negotiations, a company seeks to change the number of sick days employees may take, saying that the annual "average" is 7 days of absence per employee. The union negotiators counter that the "average" employee misses only 3 days of work each year. Explain how both sides might be correct, identifying the measure of center you think each side is using and why the difference might exist.

Exercise 6

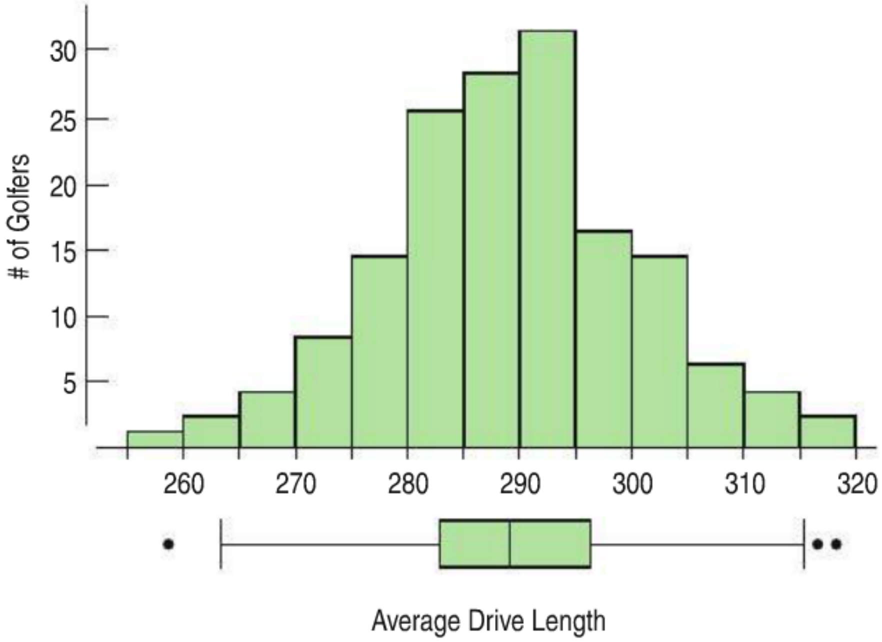
In each part examine the two given sets of numbers. Without doing any calculations, decide which set has the larger standard deviation and why. Then check by finding the standard deviations.

(a) Set I: 4,7,7,7,10, Set II: 4,6,7,8,10

(b) Set I: 100,140,150,160,200, Set II: 10, 50, 60, 70, 110

Exercise 7

The display shows the average drive distance (in yards) for 155 professional golfers during a week on the men's PGA tour in 2013.



(a) Describe this distribution.

(b) Approximately what proportion of professional male golfers drive, on average, 280 years or less?

(c) Estimate the mean by examining the histogram.

(d) Do you expect the mean to be smaller than, approximately equal to, or larger than the median? Why?

Exercise 8

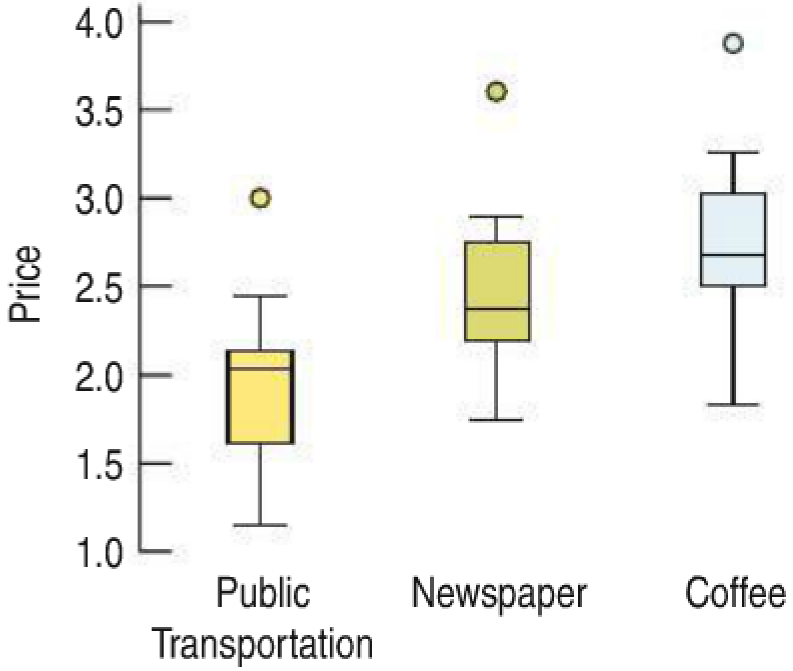
A meteorologist preparing a talk on global warming compiled a list of weekly low temperatures (in degree Fahrenheit) he observed at his southern Florida home last year. The coldest temperature for any week was 36F, but he inadvertently recorded the Celsius value of 2 degrees. Assuming that he correctly listed all the other temperatures, explain how this error will affect the mean, median, range, IQR, and standard deviation.

Exercise 9

A online jewelry company compiled statistics of the zip codes of 500 of its customers. The summary statistics it got were as follows:
Count: 500, Mean: 64,970.0, Standard Deviation: 23,523.0, Median: 64,871, IQR: 44,183, Q1: 46,050, Q3: 90,233
What can these statistics tell you about the company's sales?

Exercise 10

To help travelers know what to expect, researchers collected the prices of commodities in 16 cities throughout the world. Here are boxplots comparing the prices of a ride on public transportation, a newspaper, and a cup of coffee in 16 cities (prices in US dollars):



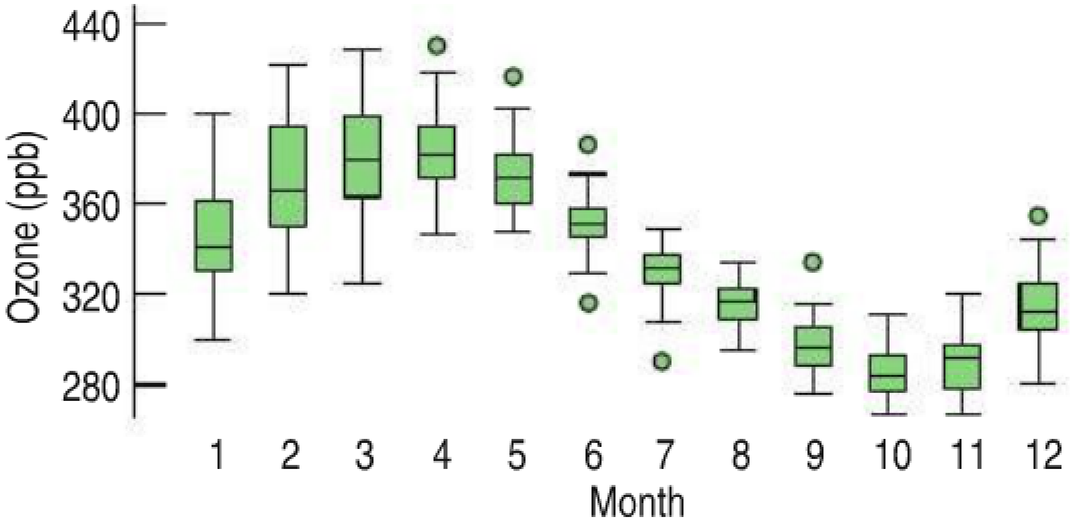
(a) On average, which commodity is the most expensive?

(b) Is a newspaper always more expensive than a ride on public transportation? Explain.

(c) Does the presence of outlier affect your conclusions in part a or b?

Exercise 11

Ozone levels (in parts per billion) were recorded at the sites in New Jersey monthly between 1926 and 1971. Here are boxplots of the data for each month (over the 46 years), lined up in order (January=1, etc.):



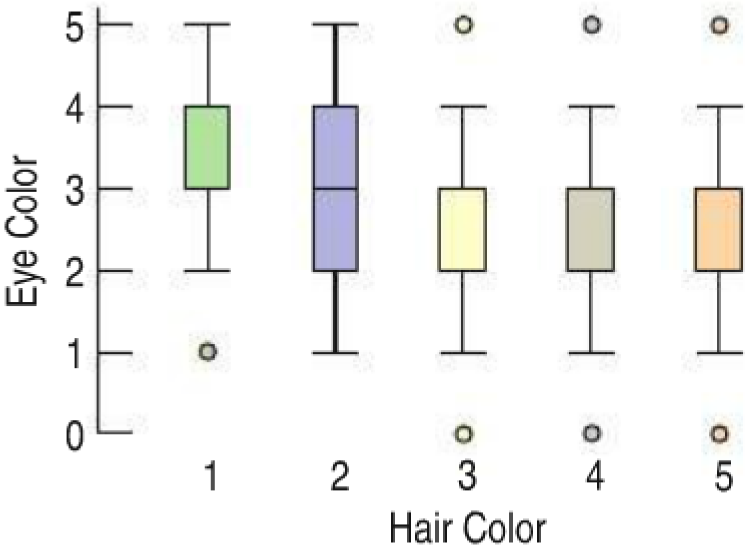
(a) In what month was the highest recorded ozone level?

(b) Which month has the largest IQR?

(c) Which month has the smallest range?

Exercise 12

A survey of 1021 school-age children was conducted by randomly selecting children from several large urban elementary schools. Two of the questions concerned eye and hair color. In the survey the following codes were used:
Hair colors: 1-Blond, 2-Brown, 3-Black, 4-Red, 5-Other
Eye colors: 1-Blue, 2-Green, 3-Brown, 4-Grey, 5-Other
The statistics students analyzing the data were asked to study the relationship between eye and hair color. They produced this plot:



Is their graph appropriate? If so, summarize the findings. If not, explain why not.

Exercise 13

A specialty foods company sells “gourmet hams” by mail order. The hams vary in size from 4.25 to 7.45 pounds, with a mean weight of 6.1 pounds, and standard deviation of 0.7 pounds. The quartiles and median weights are 5.65, 6.2, and 6.6 pounds.

(a) Find the range and IQR of the weights.

(b) Do you think the distribution of weights is symmetric or skewed? If skewed, which way? Why?

(c) If these weights were expressed in ounces (1 pound = 16 ounces), what would the mean, standard deviation, quartiles and IQR be?

(d) When the company ships the hams, the box and packing material add 30 ounces. What are the mean, standard deviation, quartiles, median, IQR and range of weights of the boxes (in ounces)?

(e) One customer made a special order of a 10-pound ham. Which of the summary statistics of part (d) might not change if that data value were added to the distribution?

Exercise 14

A high school senior uses the internet to get information on February temperatures in the town where he'll be going to college. He finds a website listing some statistics, but they are in degrees Celsius. The conversion formula is $°F = 9/5°C + 32$. Determine the Fahrenheit equivalents of the statistics below: Maximum temperature: $10°C$, Range: $32°C$, Mean: $2°C$, Standard Deviation: $9°C$, Median: $1°C$, IQR: $17°C$

Exercise 15

In Chapter 4, we looked at three outliers in terms of average wind speed per month in the Hopkins Forest. Each was associated with an unusually strong storm, but which was the most remarkable for its month?

Here are summary statistics for the months in question:

	February	**June**	**August**
Mean	2.324	0.857	0.63
SD	1.577	0.795	0.597
Outlier	6.73	3.93	2.53

(a) What are their z-scores?

(b) Which was the most extraordinary event?

Exercise 16

If we assume that the conditions for correlation are met, which of the following are true? If false, explain briefly.

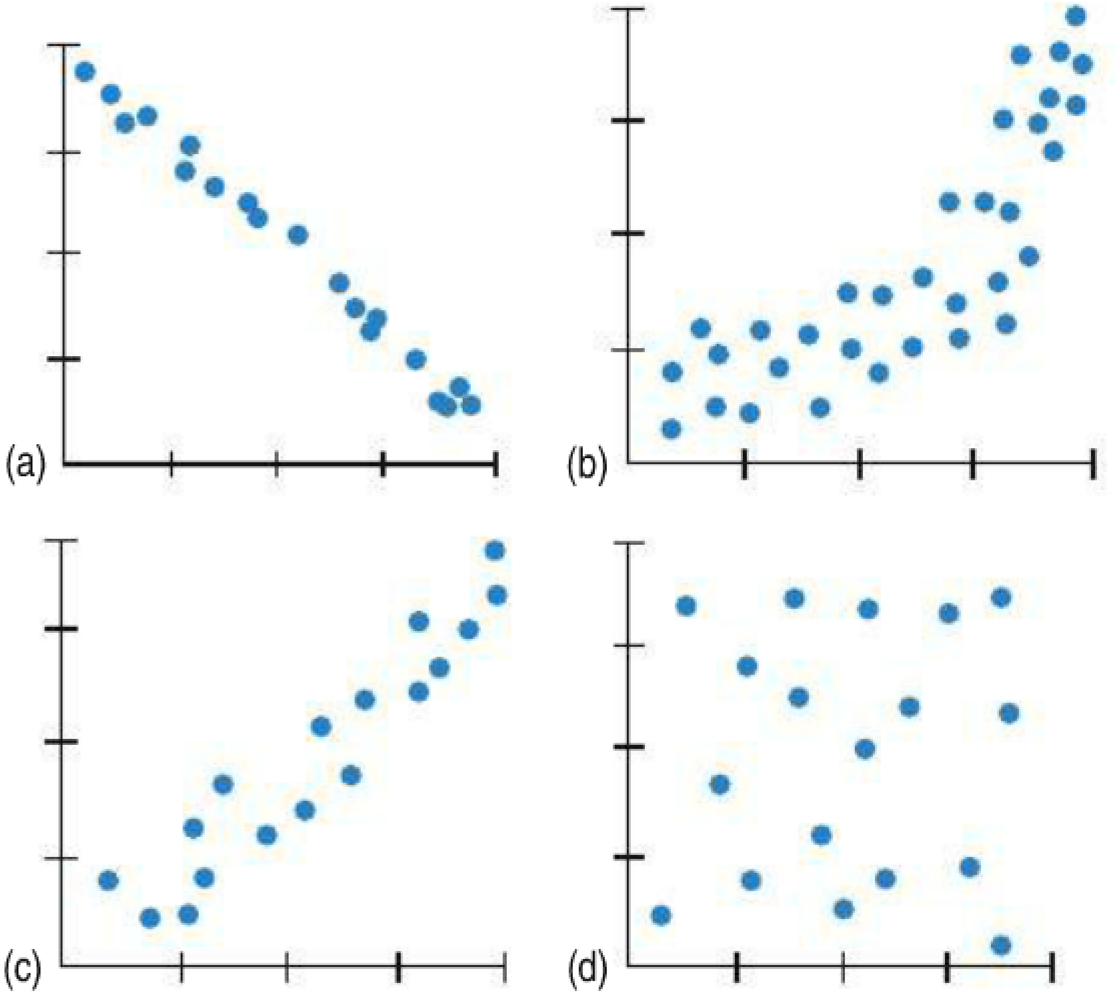
(a) A correlation of 0.02 indicates a strong positive association.

(b) Standardizing the variables will make the correlation 0.

(c) Adding an outlier can dramatically change the correlation.

Exercise 17

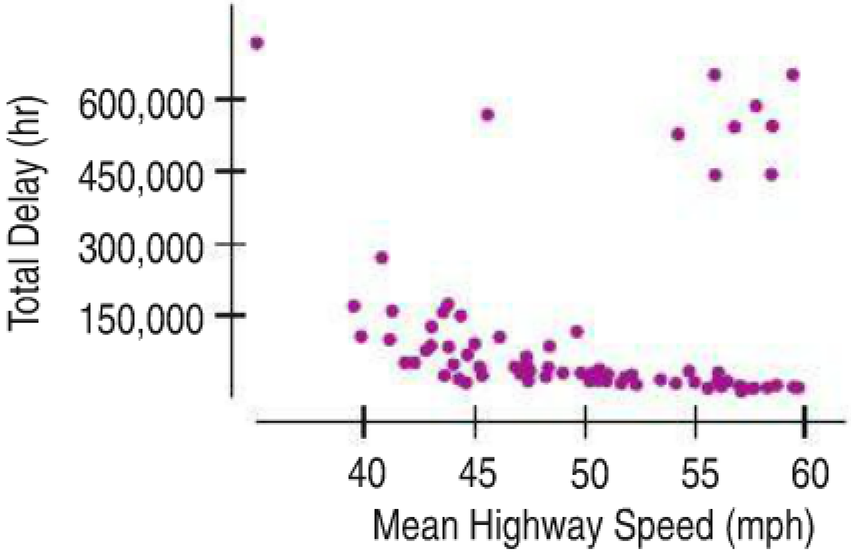
Here are several scatterplots. The calculated correlations are $-0.977, -0.021, 0.736$, and 0.951.



Which is which?

Exercise 18

A study of traffic delays in 68 US cities found the following relationship between Total Delays (in total hours lost) and Mean Highway Speeds:



Is it appropriate to summarize the strength of association with a correlation? Explain.

Exercise 19

A researcher investigating the association between two variables collected some data and was surprised when they calculated the correlation. They had expected to find a fairly strong association, yet the correlation was near 0. Discouraged, they didn't bother making a scatterplot. Explain how the scatterplot might still reveal the strong association that was anticipated.

Exercise 20

Students in an economics class were analyzing some data and wrote the following conclusions. Explain the mistakes that they made.

(a) “There was a very strong correlation of 1.22 between Life Expectancy and GDP”.

(b) “The correlation between Literacy Rate and GDP was 0.83. This shows that countries wanting to increase their standard of living should invest heavily in education.”

Exercise 21

A survey of the world’s nations in 2014 shows a strong positive correlation between percentage of the country using smart phones and life expectancy.

(a) Does this mean that smart phones are good for your health?

(b) What might explain the strong correlation?

Exercise 22

The correlation between Fuel Efficiency (as measured in miles per gallon) and Price of 150 cars at a large dealership is $r = -0.34$. Explain whether or not each of these possible conclusions is justified:

(a) The more you pay, the lower the fuel efficiency of your car will be.

(b) The form of the relationship between Fuel Efficiency and Price is moderately linear.

(c) There are several outliers that explain the low correlation.

(d) If we measure Fuel Efficiency in kilometers per liter instead of miles per gallon, the correlation will increase.

Exercise 23

A polling organization is checking its database to see if the data sources it used sampled the same ZIP codes. The variable Datasource is set to $1, 2$ or 3 depending on which of its three sources were used. The organization determines that the correlation between Datasource and the 5-digit ZIP code is -0.0229. It concludes that the correlation is low enough to state that there is no dependency between ZIP Code and Source of Data. Comment.