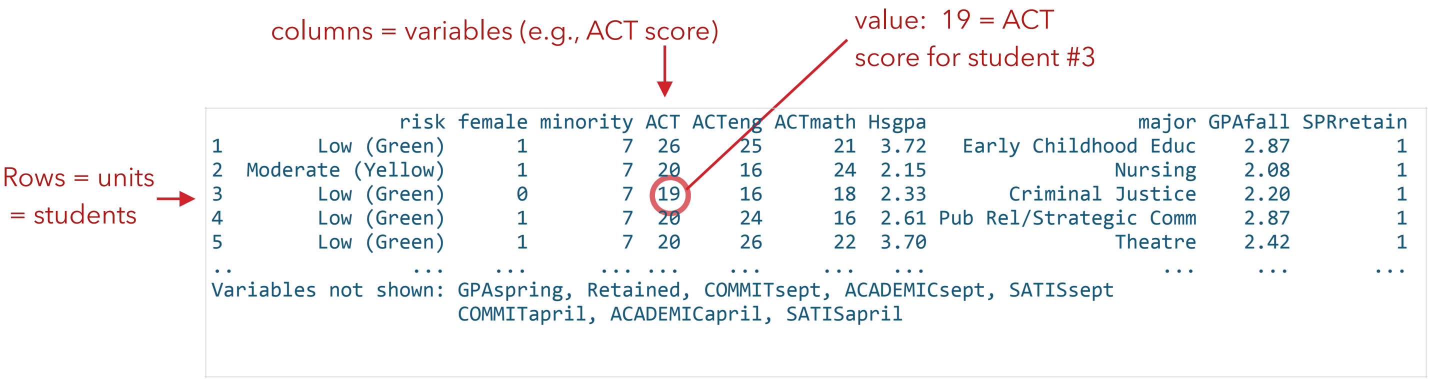
Activity #12: Summarizing Data

Data are the raw material of statistics*. If we're going to summarize a data set, it will help if our data are **tidy****.

In tidy data:
- Tables store data in rows and columns
- Rows = units (a.k.a. individuals, cases, subjects, or observations)
- Columns = variables (a.k.a. measurements)
- Each observation has a value for each variable (unless some data is missing)

In this activity, we'll summarize some data about a 200 randomly selected first-year students from St. Ambrose in 2013. Below, I've printed values for the first 5 units and 10 variables. The actual data set contains 200 rows and 18 columns.



columns = variables (e.g., ACT score)

value: 19 = ACT score for student #3

Rows = units = students

```
                  risk female minority ACT ACTeng ACTmath Hsgpa                 major GPAfall SPRretain
1          Low (Green)      1        7  26     25      21  3.72  Early Childhood Educ    2.87         1
2   Moderate (Yellow)      1        7  20     16      24  2.15               Nursing    2.08         1
3          Low (Green)      0        7  19     16      18  2.33      Criminal Justice    2.20         1
4          Low (Green)      1        7  20     24      16  2.61 Pub Rel/Strategic Comm    2.87         1
5          Low (Green)      1        7  20     26      22  3.70               Theatre    2.42         1
..                 ...    ...      ... ...    ...     ...   ...                   ...     ...       ...
Variables not shown: GPAspring, Retained, COMMITsept, ACADEMICsept, SATISsept
                     COMMITapril, ACADEMICapril, SATISapril
```

The 18 variables in this dataset include 6 **factor** variables:

variable	values
risk	Risk of not returning in 2014 (green = low risk, yellow, red = high risk)
female	0 = male; 1 = female
minority	0 = Caucasian; 1 = racial minority
major	Declared major
SPRretain	0 = student did not return in Spring; 1 = student did return in Spring
Retained	0 = student did not return in 2014; 1 = student did return in 2014

and 12 quantitative (continuous) variables:

variable	values (minimum – maximum)
ACT	ACT composite score (17 – 34)
ACTeng	ACT English score (12 – 36)
ACTmath	ACT Math score (15 – 34)
HSgpa	High school GPA (1.96 – 4.00)
GPAfall	GPA in first Fall semester (0.86 – 3.78)
GPAspring	GPA in first Spring semester (0.21 – 4.00)
COMMITsept	Student's commitment, on 09/2013, to return to St. Ambrose in 2014 (1 – 7 scale)
COMMITapril	Student's commitment, on 04/2014, to return to St. Ambrose in 2014 (1 – 7 scale)
ACADEMICsept	Self-reported academic skills on 09/2013 (1 – 7 scale)
ACADEMICapril	Self-reported academic skills on 04/2014 (1 – 7 scale)
SATISsept	Satisfaction with St. Ambrose, as of 09/2013 (1 – 7 scale)
SATISapril	Satisfaction with St. Ambrose, as of 04/2014 (1 – 7 scale)

This data comes from the MAP-Works surveys you may have taken when you were freshmen.

* Pruim, R. (2011). Foundations and Applications of Statistics: An Introduction Using R, p. 1.
** Wickham, H. (2014). Tidy Data http://vita.had.co.nz/papers/tidy-data.html

1. Looking at all 3,600 values (200 units x 18 variables) won't give us a good feel for the data. Instead, we can create some tabular, graphical, and numeric summaries of the data.

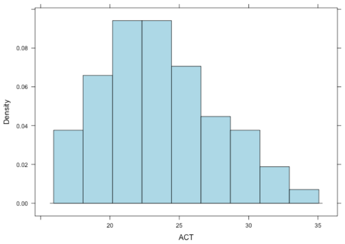
   Let's first suppose we're interested in understanding the distribution of a variable in this dataset. For **factor** (categorical) variables, we may want to construct a table. Factor variables, which may have numeric or string values, are used to place units into categories.

   Here's a table for the **risk** variable:

```
High (Red)    Low (Green)    Moderate (Yellow)       Total
         8            145                   47         200
```

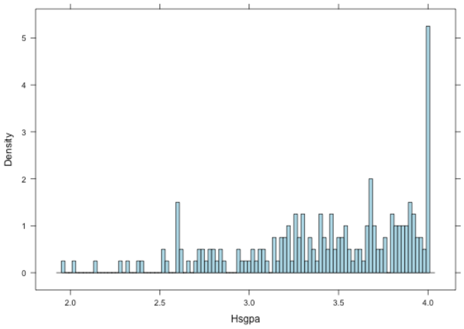
   We can construct tables for quantitative variables, too, as long as they don't have too many possible values:

```
ACT scores: 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34
 frequency:  5 11 11 17 23 17 16 24 13 17 11  8  8  8  5  3  2  1
```

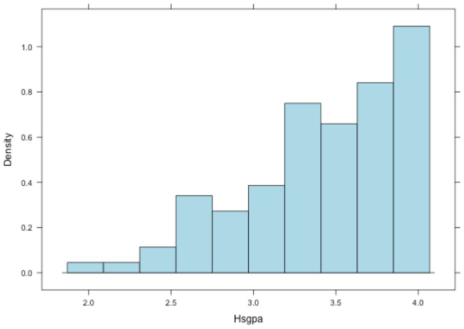


   A tabular view of quantitative data can be converted to a visual display like the histogram displayed to the right. With histograms, we can choose various bin widths (or number of bars).

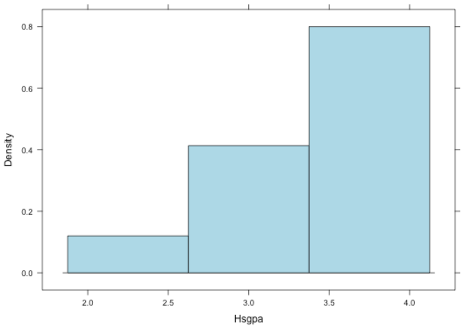
   Which of the following histograms do you think best displays the distribution of high school GPAs? Why?



`binwidth = 0.02`

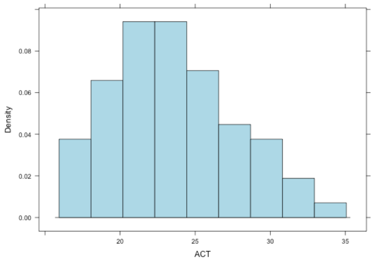


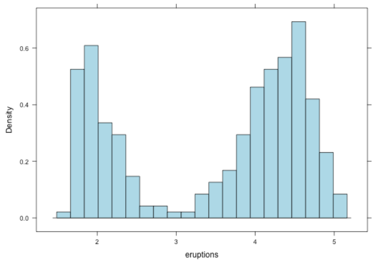
`binwidth = 0.22`

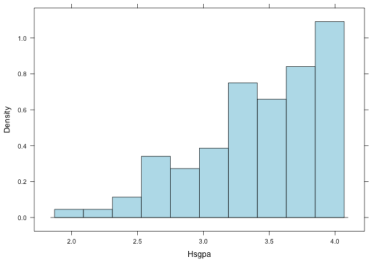


`binwidth = 0.75`

2. Histograms tell us about the shape of our data. Describe the shape of each of these distributions:



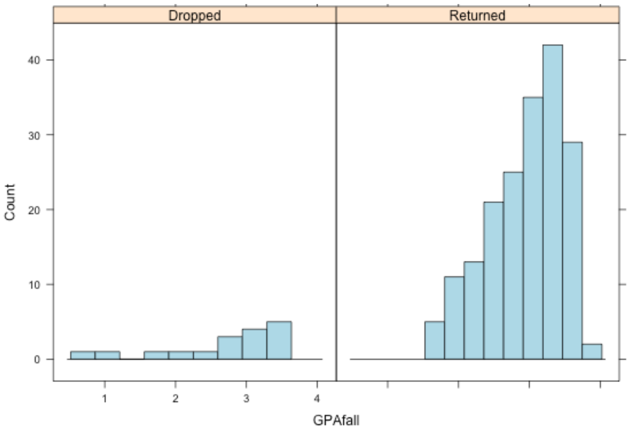




Variable:	ACT scores	geyser eruption times	HS GPA
Skew:	____________	____________	____________
Mode(s):	____________	____________	____________

3. We can create side-by-side histograms to compare values on a variable across categories of a factor.

   To the right is a histogram displaying Fall semester GPAs for students who did or did not return to St. Ambrose the next year. What can you interpret from this? Is Fall GPA a good predictor of student retention?



4. One disadvantage of histograms is that the data values are lost (by combining them into bins). If you have a relatively small dataset and want to see the distribution of all the values, you can create a stemplot:

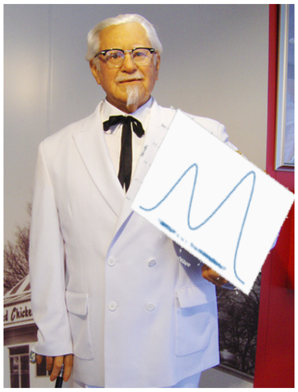
```
Stemplot for Fall semester GPA  ( the decimal point is 1 digit(s) to the left of the | )

   8 | 6
  10 |
  12 | 1
  14 | 6777
  16 | 49
  18 | 23
  20 | 00027778898
  22 | 00123339133369
  24 | 00022799900344778
  26 | 02357777791337
  28 | 000022223777788023333588
  30 | 0000000066666778882333444557889
  32 | 00002445555556779023335567788888
  34 | 0000000002334577777777000000003679
  36 | 1133345713358
```

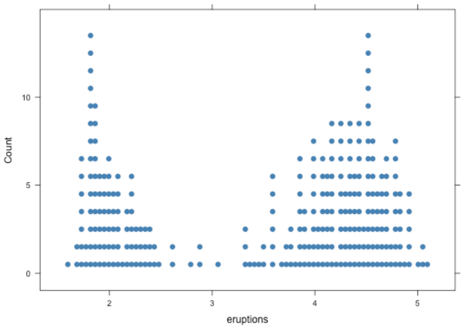
How would you describe the shape of this distribution?

5. Another disadvantage of histograms is their appearance depends on the width of the bins. To deal with this, we can graph a **kernel density plot**. A kernel density plot provides a smoothed distribution of a variable can be easier to describe or understand.

   You can read more about kernel density estimation at: http://en.wikipedia.org/wiki/Kernel_density_estimation. For the geyser eruption data, the idea is this:

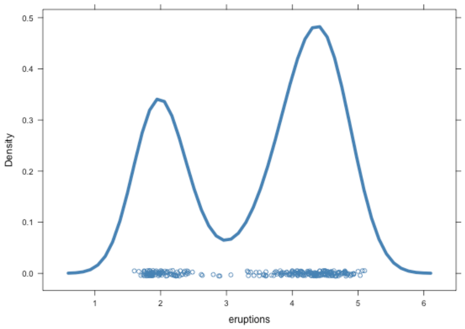


Named after the famous Colonel Density



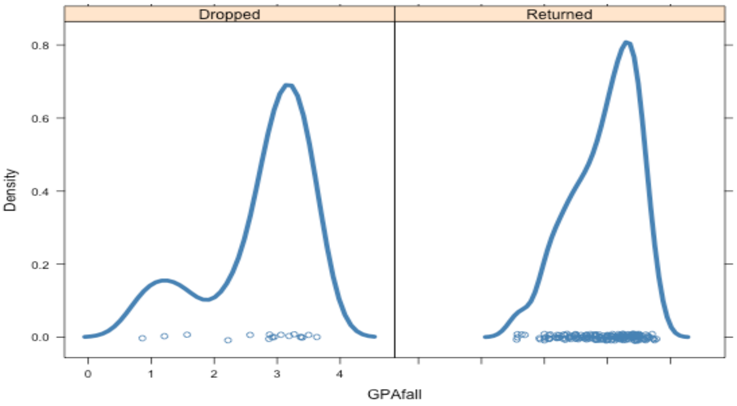
1. Create a dotplot of the data showing all observations in the dataset.

2. Replace each dot with a handful of sand. When you drop the sand, it begins to pile up into a smooth distribution.



6. With density plots, we can compare distributions in a single plot. Here, once again, are the Fall GPAs for students who did and did not return to St. Ambrose the next year.

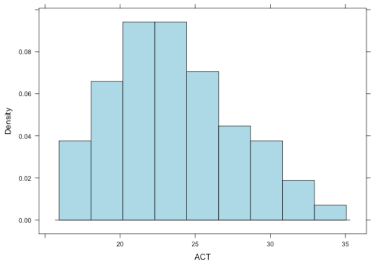
   If you had to select a number or numbers that **best summarize** the distribution of Fall GPAs for each group, what number(s) would you choose? Why?

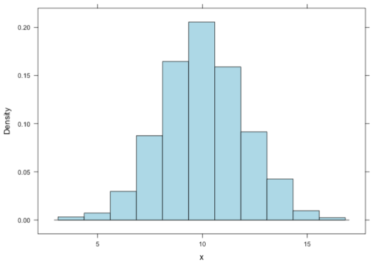


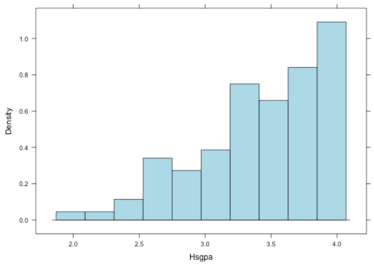
7. Visual displays are important, but we may want a more precise numerical summary of our data. For example, we might want to know the center of a distribution.

   Two measures of central tendency that you're already familiar with are the mean and median. In the assignment associated with this activity, you'll derive definitions for both the mean and median. For now, let's quickly review what they represent.

   Interpret the mean and median for one of the distributions. How can we use the mean and median to determine the skewness of a distribution?







Variable:	ACT scores	(simulated variable)	HS GPA
Mean:	23.72	10.03	3.4315
Median:	23.50	10.04	3.5000
20% Trimmed mean	23.42	10.02	3.5083

Interpretation of mean: ______________________________

Interpretation of median: ______________________________

How to determine skew from mean and median: ______________________________

What's a trimmed mean? ______________________________

8. For the geyser eruption data, we could calculate the mean and median eruption times: $\bar{X} = 3.4878$, $\text{Median} = 4$. Eruption times are measured in minutes. What would happen to the mean and median if we changed to seconds?

Mean (in minutes) = 3.4878 Mean (in seconds) = ______________________________

Median (in minutes) = 4 Median (in seconds) = ______________________________

Explain the following: Given $\bar{X} = \hat{\mu}_x = \left(\frac{1}{n}\right)\sum_{i=1}^{n} x_i$

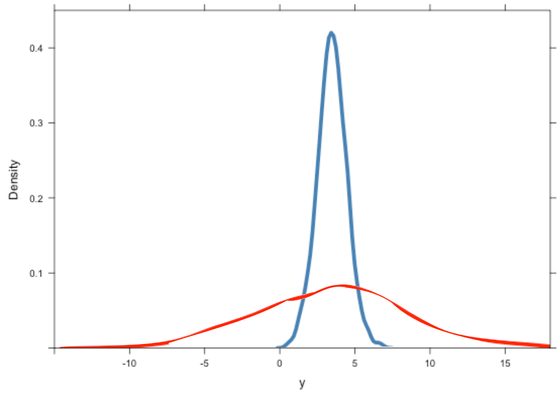
$$\hat{\mu}_{ax+b} = \left(\frac{1}{n}\right)\sum_{i=1}^{n} ax_i + b$$

$$= \left(\frac{1}{n}\right)\left(ax_1 + b + ax_2 + b + \ldots + ax_n + b +\right)$$

$$= \left(\frac{1}{n}\right)\left[a\left(x_1 + x_2 + \ldots + x_n\right) + nb\right]$$

$$= \left(\frac{1}{n}\right)\left[a\sum_{i=1}^{n} x_i + nb\right] = a\frac{\sum_{i=1}^{n} x_i}{n} + b = a\bar{X} + b$$

9. The center of a distribution doesn't tell the whole story.



To the right, you can see two variables that have nearly identical means and medians. While they don't differ in terms of their centers, they certainly differ in terms of their **dispersion**. The red distribution is more spread out than the blue distribution.

Almost all the data in the blue distribution is close to 3.5, while a much larger proportion of data in the red distribution is far away from 3.5.

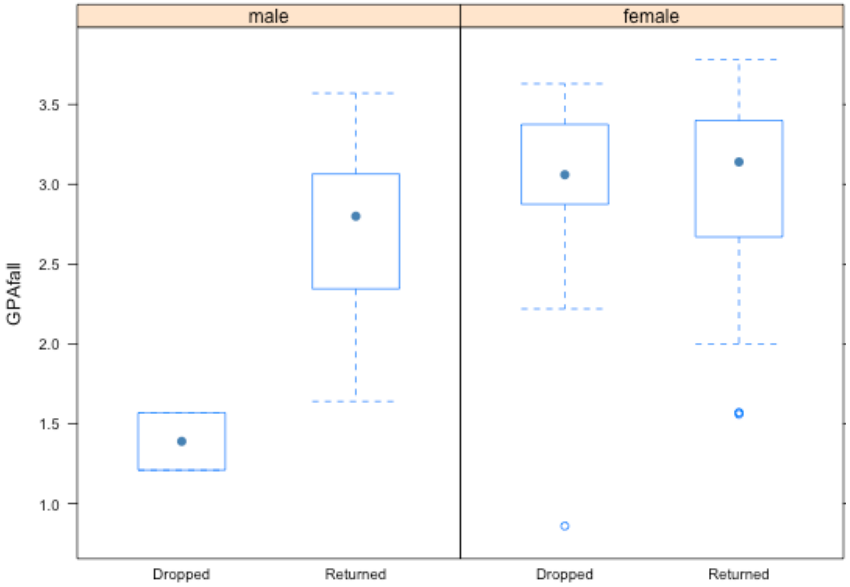
To more precisely state this fact, we can use **quantiles**.

A **p-quantile** of a distribution is a number **q** such that the **proportion of the distribution that is less than q** is **p**.

0.2-quantile = 20th percentile = value that divides the distribution into 20% below and 80% above.
0.5-quantile = 50th percentile = median = value that divides the distribution into 50% below and 50% above.

Suppose we have a small dataset with the values: **1, 2, 3, 4, 5**. Identify the 25th and 50th percentiles of the data.

10. Boxplots visualize some important quantiles of a distribution. Interpret and Identify the key features of the following box plot of Fall semester GPAs:



.

IQR = ____

11. Visualizations (histograms and density plots) show us the shape of a distribution. Measures such as the mean, median, and percentiles tell us about the location of a distribution. The IQR, much like the range, can inform us about the dispersion of a distribution. One limitation of the IQR and range is that they're only based on two values.

    Another important way to describe the dispersion of a distribution would be to calculate the deviations (distances) from each value to the mean. Write the formula for a deviation:

    If we calculated deviations for our Fall semester GPAs, we'd have 200 different deviations. To get a single measure of the variability in GPAs, we could find the sum of those deviations. Identify 2 limitations with the sum of these deviations and propose a calculation that will address those limitations:

    Limitation #1: ____

    Solution #1: ____

    Limitation #2: ____

    Solution #2: ____

12. Two different formulas for the standard deviation of a variable are displayed below.

$$\mathrm{SD}\left[x\right]=\sqrt{\mathrm{var}\left[x\right]}=\sqrt{\sigma_x^2}=\sigma_x=\sqrt{\frac{\sum_{i=1}^{n}\left(x_i-\mu\right)^2}{n}}$$

$$\text{unbiased estimate of }\mathrm{SD}\left[x\right]=\sqrt{\hat{\sigma}_x^2}=\sqrt{s_x^2}=s_x=\sqrt{\frac{\sum_{i=1}^{n}\left(x_i-\bar{X}\right)^2}{n-1}}$$

We'll learn what an unbiased estimate is (and why we need n-1 in the denominator) in the next activity.
For Fall semester GPAs, the (unbiased estimate of the) standard deviation was calculated to be **0.5537**. Interpret.

13. The standard deviation of geyser eruption times was calculated to be **1.1414 minutes**. Suppose we converted all the times to seconds. What would happen to the standard deviation?

SD (in minutes) = 1.1414 SD (in seconds) = ______________________________

Suppose we then had to add 3 seconds to each value (because of a faulty stopwatch). What would happen to the standard deviation?

SD(x) = 1.1414 SD (60x + 3) = ______________________________

Explain the following:

Given $\mathrm{var[x]}=\sigma_x^2=\frac{\sum_{i=1}^{n}\left(x_i-\mu\right)^2}{n}$

$$\mathrm{var[ax+b]}=\sigma_{ax+b}^2=\frac{\sum_{i=1}^{n}\left(\left(ax_i+b\right)-\left(a\mu+b\right)\right)^2}{n}=\frac{\left[\left(ax_1+b\right)-\left(a\mu+b\right)\right]^2+\ldots+\left[\left(ax_n+b\right)-\left(a\mu+b\right)\right]^2}{n}$$

$$=\frac{\left[\left(ax_1-a\mu+b\right)\right]^2+\ldots+\left[\left(ax_n-a\mu\right)\right]^2}{n}=\frac{\left(a^2x_1^2-2a^2\mu x_1+a^2\mu^2\right)+\ldots+\left(a^2x_n^2-2a^2\mu x_n+a^2\mu^2\right)}{n}$$

$$=\frac{a^2\left(x_1^2-2\mu x_1+\mu^2\right)+\ldots+a^2\left(x_n^2-2\mu x_n+\mu^2\right)}{n}=\frac{\sum_{i=1}^{n}a^2\left(x_i-\mu\right)^2}{n}=\frac{a^2\sum_{i=1}^{n}\left(x_i-\mu\right)^2}{n}=a^2\sigma_x^2$$

14. Suppose we add one outlier to our geyser eruption dataset. The longest eruption (out of all 272 observations in the dataset) is 5.1 minutes. Suppose we change that largest value to 51.0 minutes. What impact, if any, would that have on the following:

Mean = 3.4878. The outlier would cause the **mean** to: **increase** **decrease** **no change** **unknown**

Trimmed mean = 3.6165. The outlier would cause this to: **increase** **decrease** **no change** **unknown**

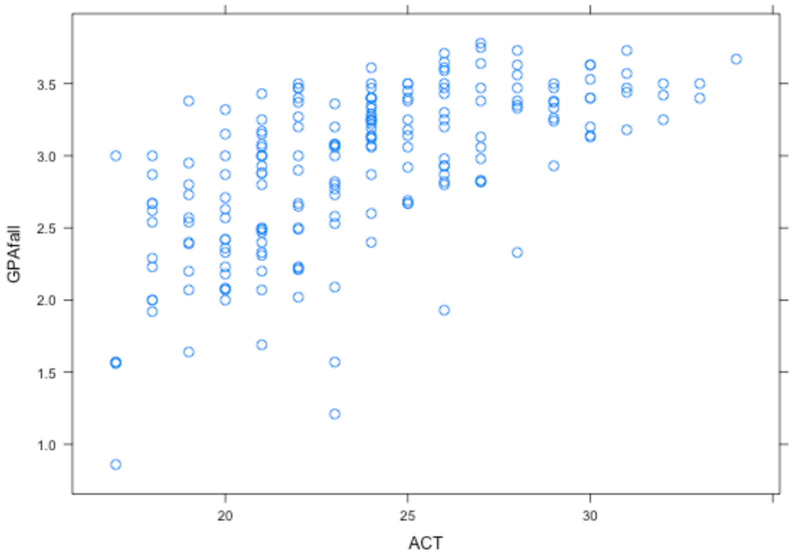
Median = 4. The outlier would cause the **median** to: **increase** **decrease** **no change** **unknown**

Std. Dev. = 1.1414. The outlier would cause the **std. dev.** to: **increase** **decrease** **no change** **unknown**

Original IQR = 2.2915. The outlier would cause the **IQR** to: **increase** **decrease** **no change** **unknown**

15. So far, we've primarily been examining a single variable at a time. We're often interested in the relationship between two or more variables, as we saw with the boxplots comparing GPAs by gender and retention status.

Scatterplots are a good way to examine the relationship between two quantitative variables. Here's a scatterplot showing the relationship between ACT scores and Fall semester GPAs:



How might we summarize this relationship? How might we summarize the strength of this relationship?

16. Between 1967–1977, there was a moratorium on the death penalty in the U.S. One reason for this was the belief that racial discrimination influenced death penalty sentencing. When the racial bias argument was tested in court*, supporters of the death penalty produced the following contingency table..

	Defendant		
Penalty	Black	White	**Total**
Death	17	19	**36**
Not	149	141	**290**
Total	**166**	**160**	**326**

If we randomly select a defendant from this table, calculate the following:

P(death | white) = ________________________________

P(death | black) = ________________________________

Relative rate = ____________________________________

Interpretation: ____

Let's add another variable to our tables – the race of the victims.

For <u>white</u> victims:

	Defendant		
Penalty	Black	White	**Total**
Death	11	19	**30**
Not	52	132	**184**
Total	**63**	**151**	**214**

For <u>black</u> victims:

	Defendant		
Penalty	Black	White	**Total**
Death	6	0	**6**
Not	97	9	**106**
Total	**103**	**9**	**112**

Calculate the following for each table:

P(death | white) = __________________________ __________________________

P(death | black) = __________________________ __________________________

* Radelet, M. (1981). Racial characteristics and imposition of the death penalty. American Sociological Review, 46:918-927.