Chapter 2

Displaying and Describing Categorical Data

Homework

p31 5, 6, 7, 8, 9, 10, 21, 22, 23, 25, 31, 32

Your Turn

The Three Rules of Data Analysis

The first three rules of data analysis are simple to remember:

1. **Draw a pichur:** An illustration of the data can show any trends, patterns, or unusual characteristics of the collection that is not obvious in a simple list of the data.

2. **Draw a pichur:** patterns in the data can be seen in a visual representation of the data and you may see things that you would otherwise miss.

3. **Draw a pichur:** a picher is worth a thousand words. Not sure I agree with that, but a picher allows you to more easily explain your words.

Frequency Tables

- We can organize the data by counting the **number** of data values some category of interest.

  - We organize the **counts** into a **frequency table**, which simply records the category names and the total frequency within each category.

Class	Frequency
First	325
Second	285
Third	706
Crew	885

Frequency Tables

- A **relative frequency table** is virtually the same, but gives the percentages or proportion for each category in place of the absolute count in the category.

Class	Count
First	325
Second	285
Third	706
Crew	885

2201

Class	(f/total)100
First	14.77
Second	12.95
Third	32.08
Crew	40.21

$$\frac{325}{2201} = .147660154\ldots$$

- We are going to experience **relative frequencies** often in this course. **Relative frequency** is a proportion, percent, and empirical probability.

Frequency Tables

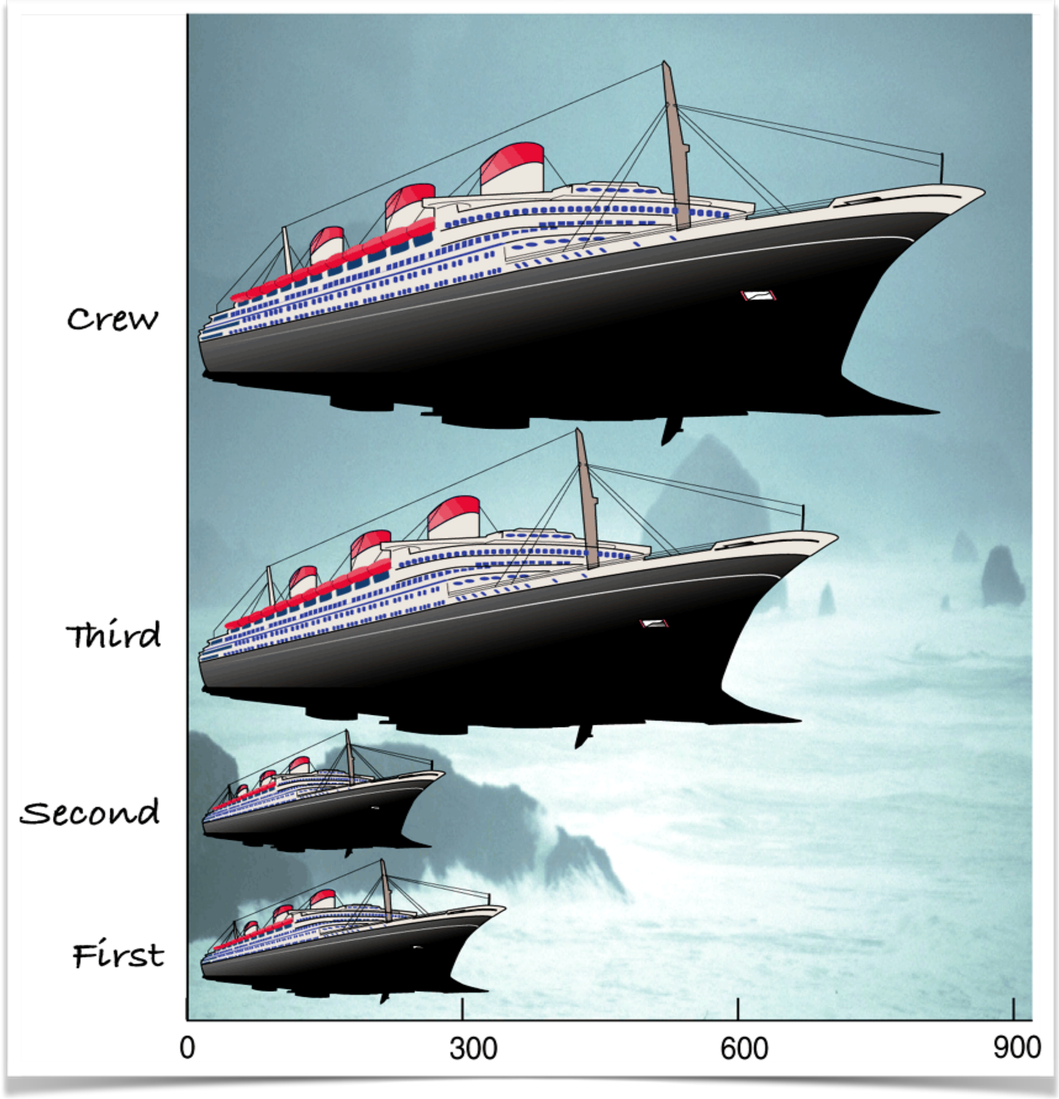
- Both types of tables show how cases are distributed across the categories.

- Frequency tables illustrate the **distribution** of a categorical variable as the table names the possible categories and indicates how frequently each category occurs.

Anything Wrong With This Picture?

- Here is an artistic, clever way to show the number of people on the Titanic:

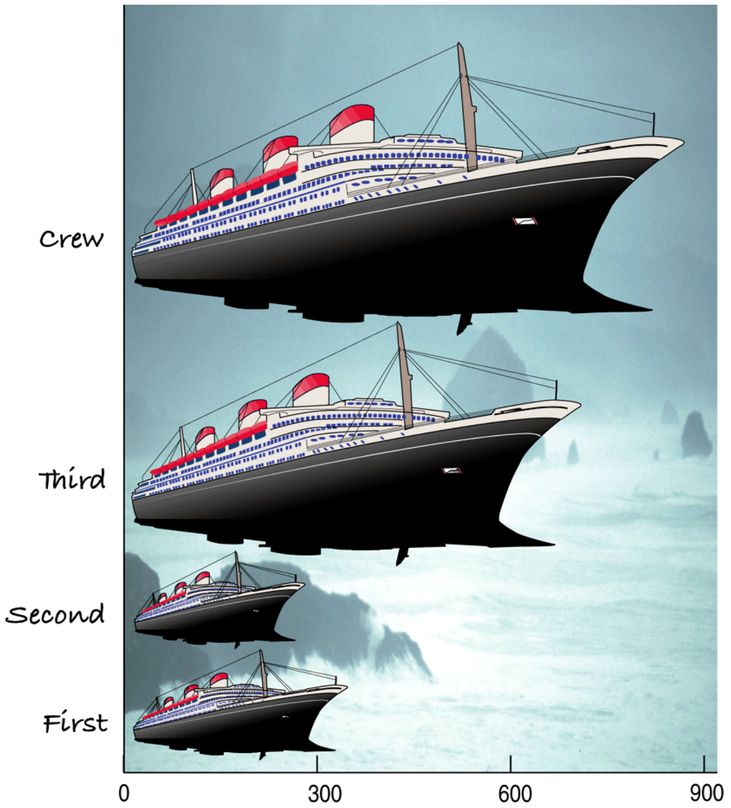
Class	Count
First	325
Second	285
Third	706
Crew	885



- We will not be clever, or artistic in examining data!

The Area Principle

- The ship display makes it look like most of the people on the Titanic were crew members, with a few passengers along for the ride.



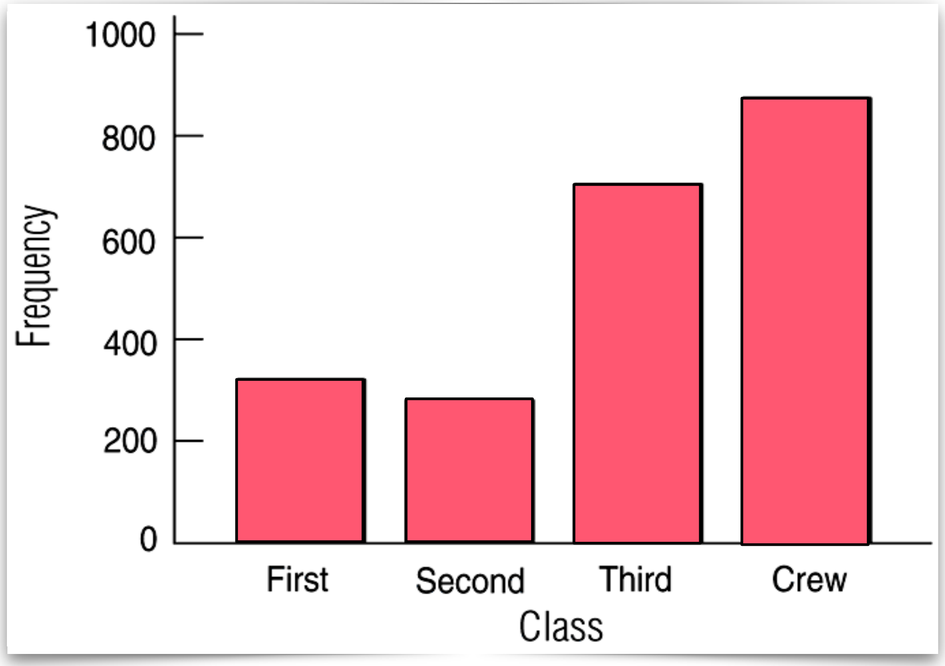
- When looking at each ship, we react to the **area** taken up by the ship, instead of the **length** of the ship. **The length is the only valid characteristic.**
- The ship display violates the **area principle**:
  - The **area** occupied by a part of the graph should match the magnitude of the value it represents.
  - Do not get clever, creative, cute, or fancy .....

Do NOT get fancy

- A **bar chart** displays the distribution of a categorical variable, showing the counts for each category next to each other for easy comparison.

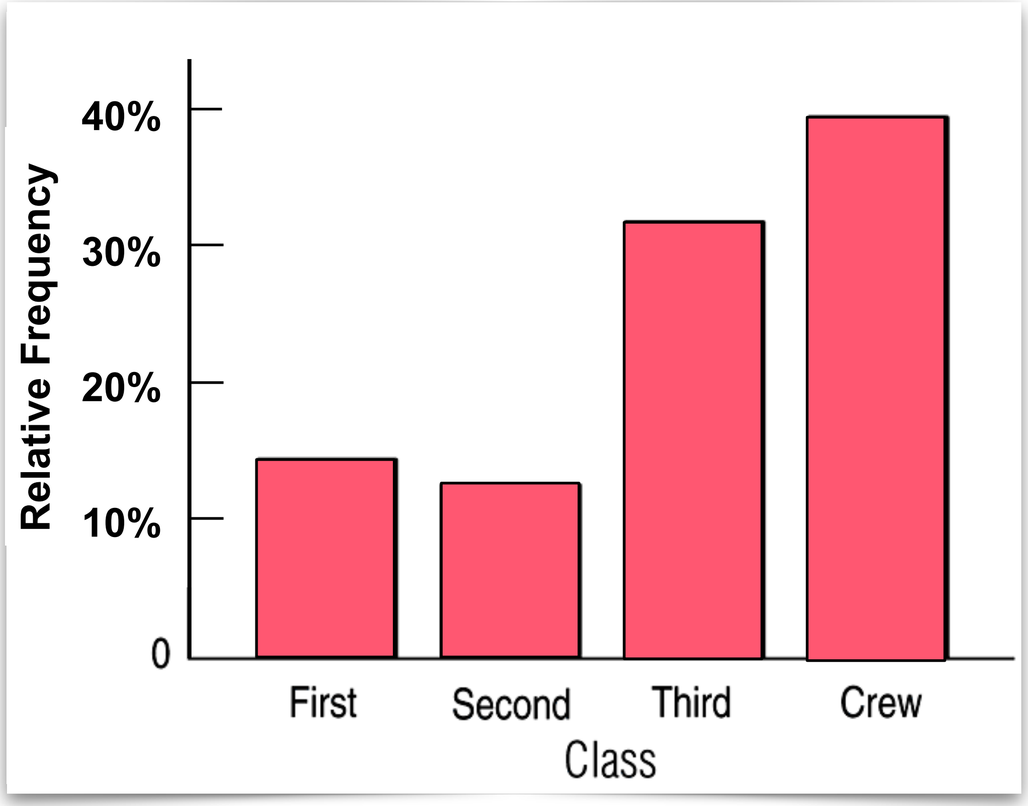
- Thus, a much better display for the ship data is:

- This bar chart stays true to the area principle.



Bar Charts

- A **relative frequency bar chart** displays the relative proportion of counts for each category.



- Simply replace absolute counts with **percentages** or **relative frequencies** for the data:

- A relative frequency bar chart also stays true to the area principle.

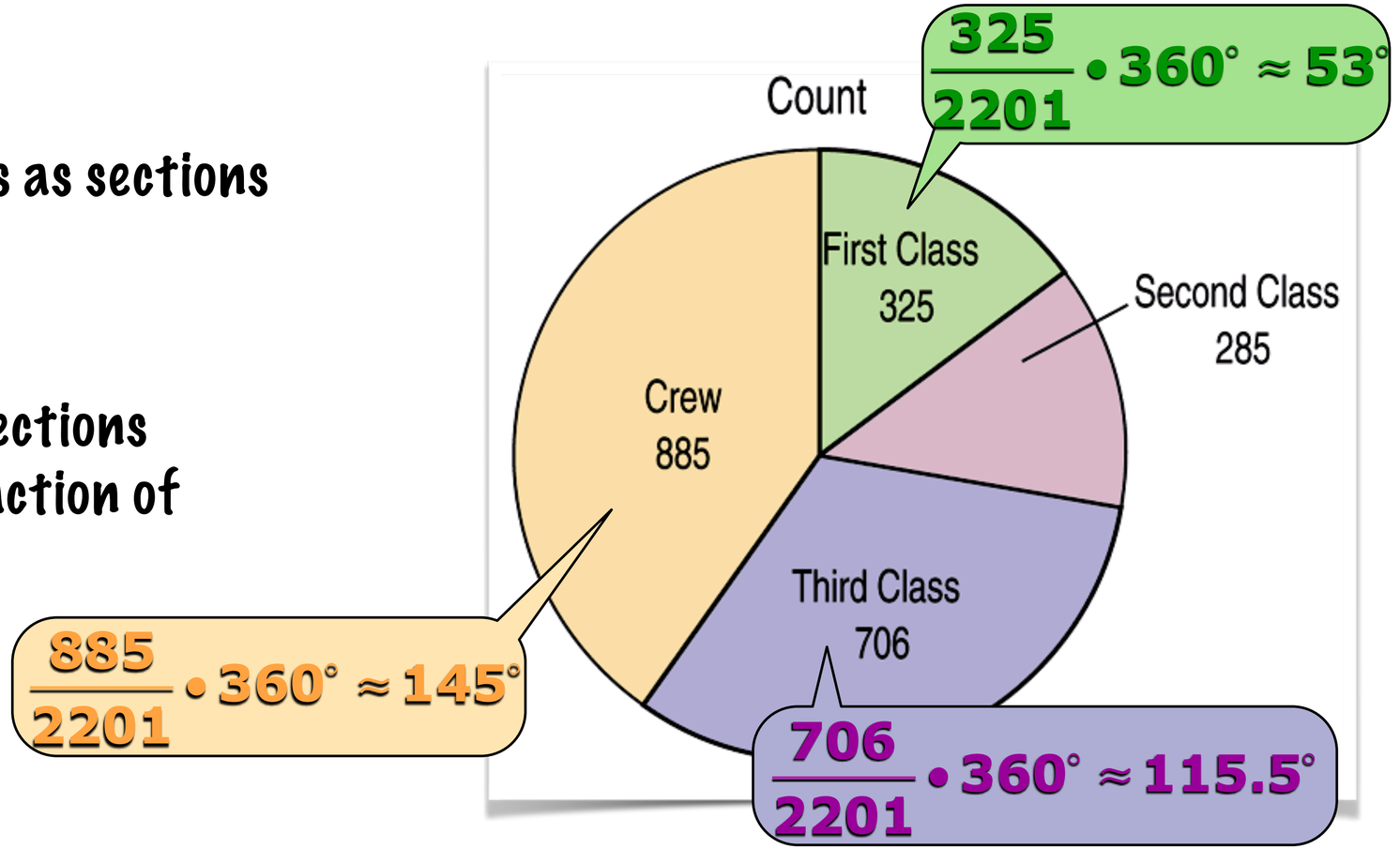
- The only difference between a frequency bar chart and a relative frequency bar chart is the vertical axis label and scaling.

Pie Charts

- If you are interested in illustrating the relative size of **parts of the whole**, a **pie chart** might be your best choice.

- Pie charts show all the categories as sections of a circle.

- Pie charts divide the circle into sections whose size is a **proportional** fraction of the whole for each category.



Contingency Tables (Two Way Tables)

- A **contingency table** (two way table) allows us to look at two categorical variables together.
  - **Important**. We will be referring to contingency table several times later in this course. So remember this topic.

- A contingency table shows how data frequencies are distributed for one variable, **contingent upon** each level of the other variable.

- In our example we can examine the class of ticket **contingent upon** the survival condition:

		Class				
		First	Second	Third	Crew	Total
Survival	Alive	203	118	178	212	711
	Dead	122	167	528	673	1490
	Total	325	285	706	885	2201

Contingency Tables

- The **margins** of the table, both on the right and on the bottom, give totals and the **frequency distributions for each of the variables**.

- The frequency distribution for each variable is called a **marginal distribution** of its respective variable in the contingency table.

		Class				
		First	Second	Third	Crew	Total
Survival	Alive	203	118	178	212	711
	Dead	122	167	528	673	1490
	Total	325	285	706	885	2201

- The marginal distribution of Survival is:

Alive	711
Dead	1490

- The marginal distribution of Class is:

First	Second	Third	Crew
325	285	706	885

Contingency Tables

- Each **cell** of the table gives the count for a **combination** of two conditions.
  - For example, the second cell in the crew column tells us that 673 crew members died when the Titanic sunk.

		Class				
		First	Second	Third	Crew	Total
Survival	Alive	203	118	178	212	711
	Dead	122	167	528	673	1490
	Total	325	285	706	885	2201

- crew **and** died

Conditional Distributions

- A **conditional distribution** shows the distribution of **one variable** for just the individuals who satisfy a **single condition on another variable.**

- The following is the conditional distribution of ticket **Class**, conditional on having **survived:**

Survival	Class: First	Second	Third	Crew	Total
Alive	203	118	178	212	711
Dead	122	167	528	673	1490
Total	325	285	706	885	2201

	Class: First	Second	Third	Crew	Total
Alive	203	118	178	212	711
	28.6%	16.6%	25.0%	29.8%	100%

Conditional Distributions

- The following is the conditional distribution of ticket **Class**, conditional on having perished:

Survival	Class: First	Second	Third	Crew	Total
Alive	203	118	178	212	711
Dead	122	167	528	673	1490
Total	325	285	706	885	2201

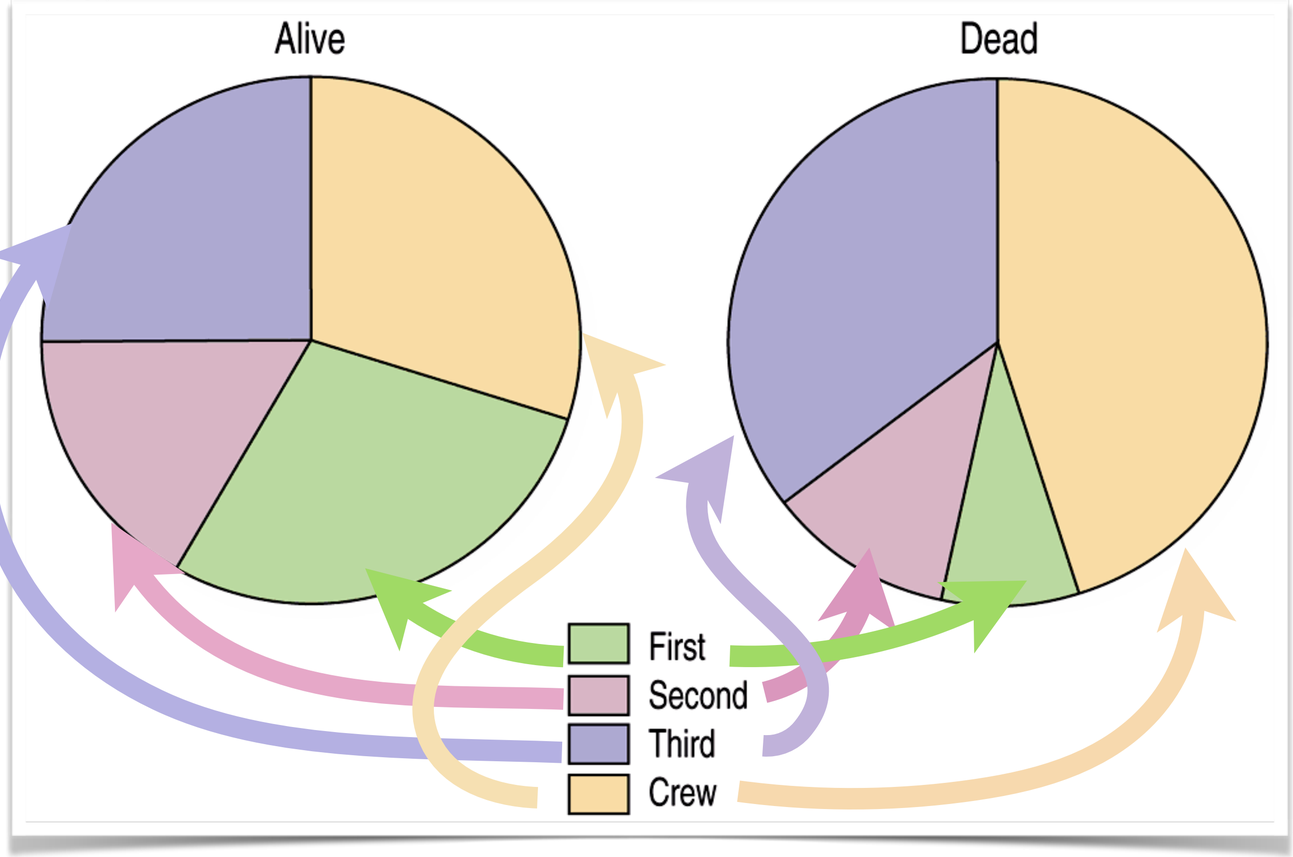
	Class: First	Second	Third	Crew	Total
Dead	122	167	528	673	1490
	8.2%	11.2%	35.4%	45.2%	100%

Conditional Distributions

- The conditional distributions tell us that there is a difference in the distribution of class for those who survived and those who perished.

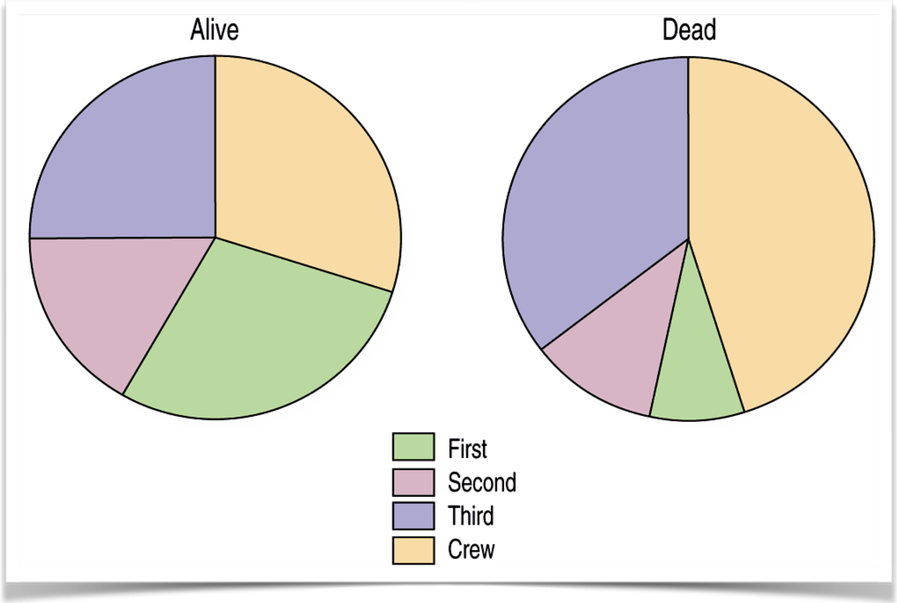
- This is easily seen with pie charts of the two distributions:

- Note the obvious differences in section sizes between the two charts.



Conditional Distributions

- We see that the distribution of Class for the survivors is different from that of the non-survivors.

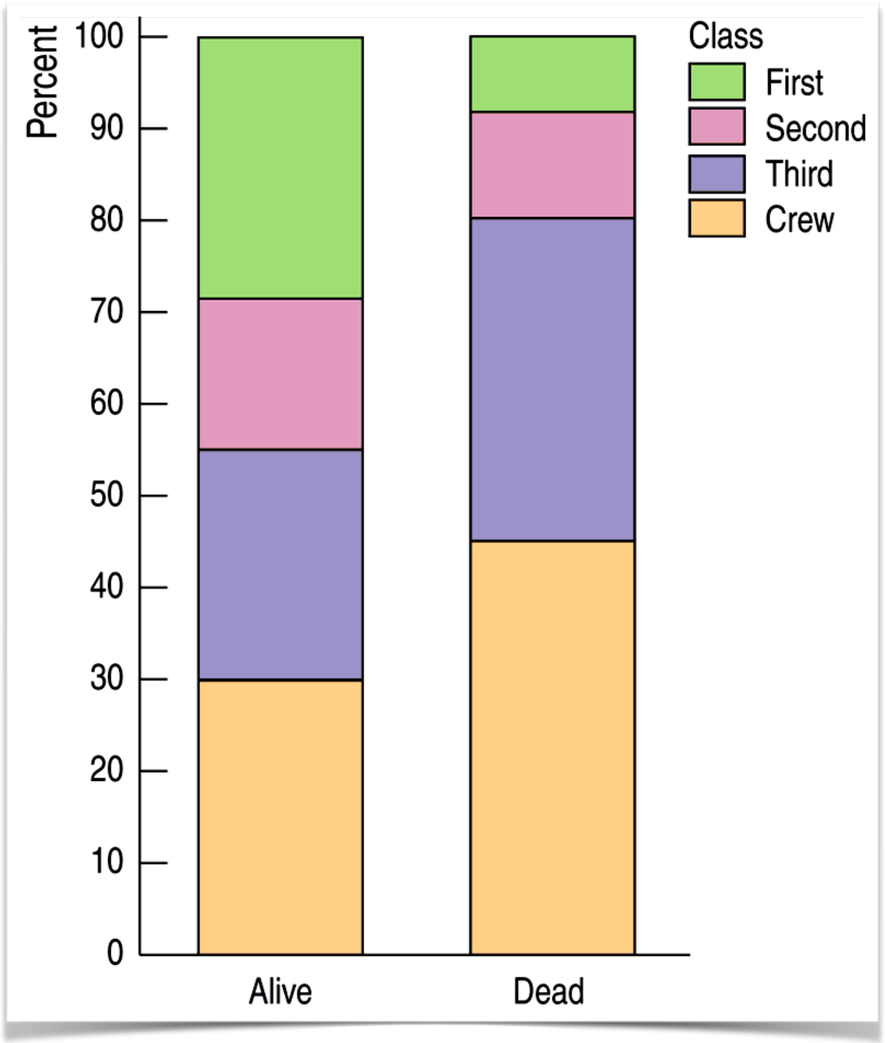


- This leads us to believe that **Class** and **Survival** are associated. That is, they are **not independent**.

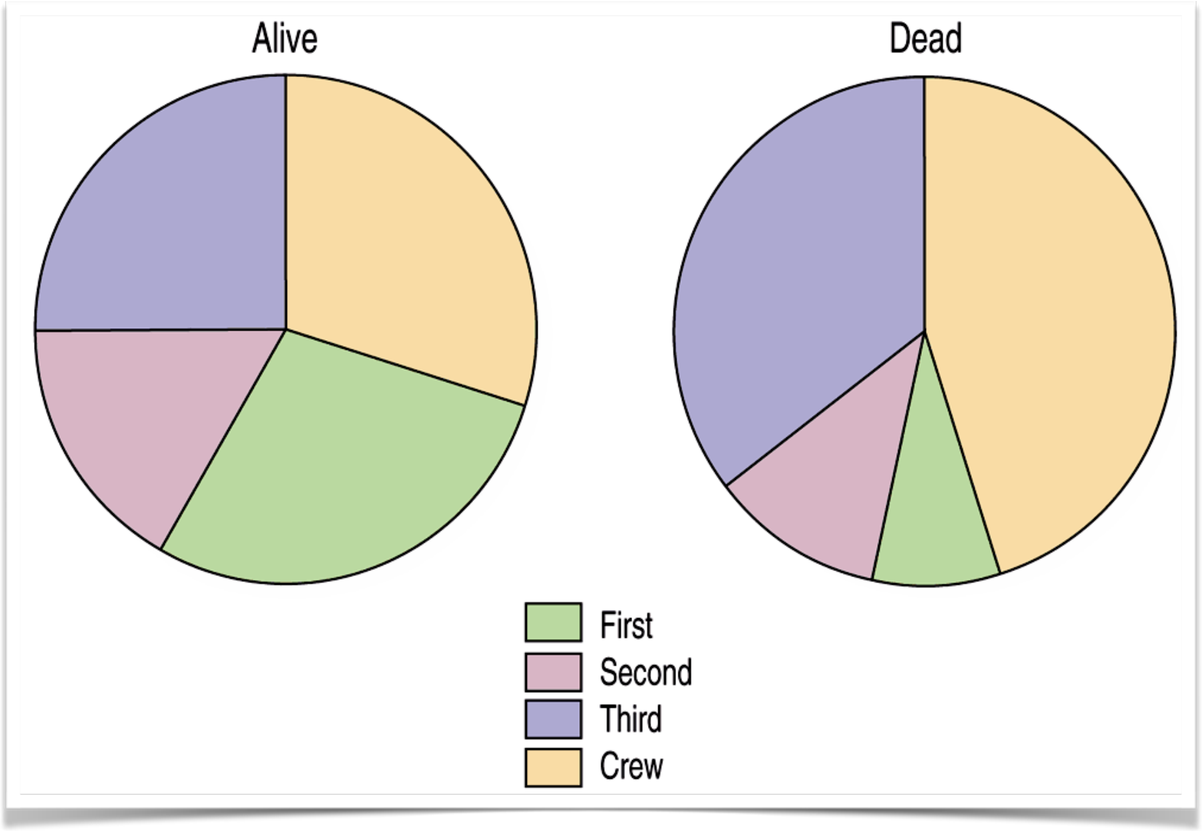
- The variables would be considered **independent** when the **distribution** of categories of one variable in a contingency table is essentially the **same for all categories** of any other variables in the table.

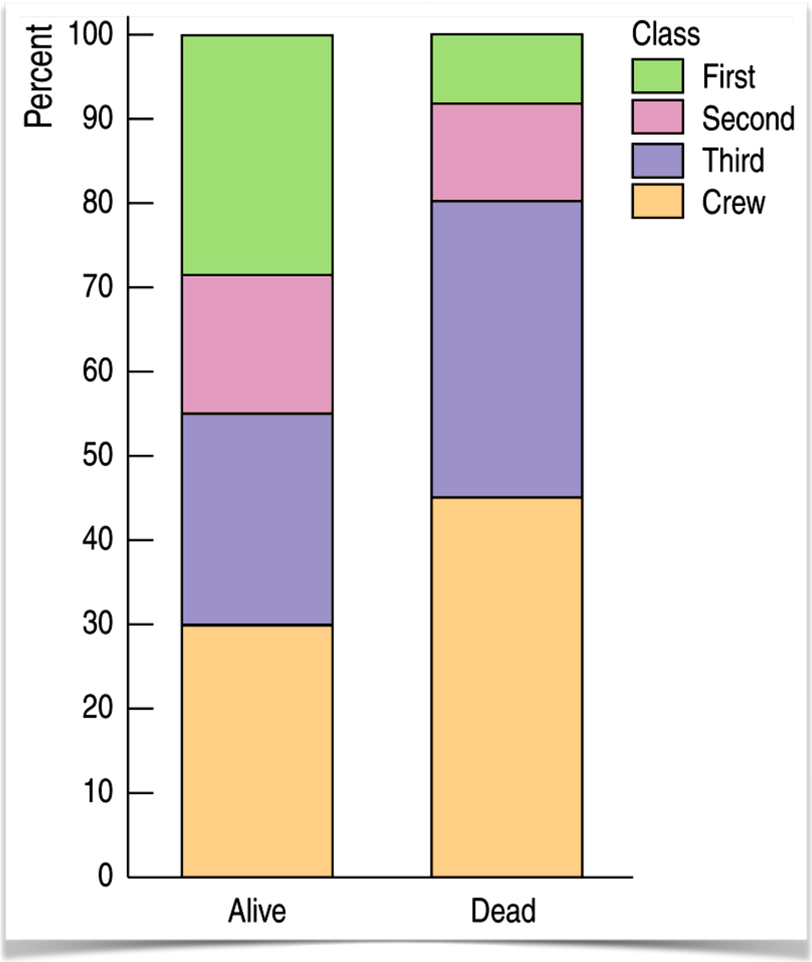
Segmented Bar Charts

- A **segmented bar chart** displays the same information as a pie chart.
- Here is the segmented bar chart for ticket Class by Survival status:
- Each bar is treated as the "whole" and is divided proportionally into segments corresponding to the percentage in each group.



Side-by-Side





Each bar corresponds to a "pie". The information portrayed is the same in each picture but the emphasis is slightly different.

Simpson's Paradox

- Simpson's paradox is a result of averaging done when averages can be misleading.
- Let us peek into the lives of two waiter's in a local eatery, Gyade and JuChi.
- Gyade and JuChi are competing for promotion to night manager. The restaurant manager decides to look at tip count to measure customer satisfaction.

	Gyade	JuChi
Lunch	50 meals - $100	100 meals - $300
Dinner	100 meals - $600	50 meals - $400
Total	150 meals - $700	150 meals - $700

Simpson's Paradox

- Based on the marginal distribution of waiter it appears Gyade and JuChi are equally well regarded by customers.

- Perhaps we should look a little closer.

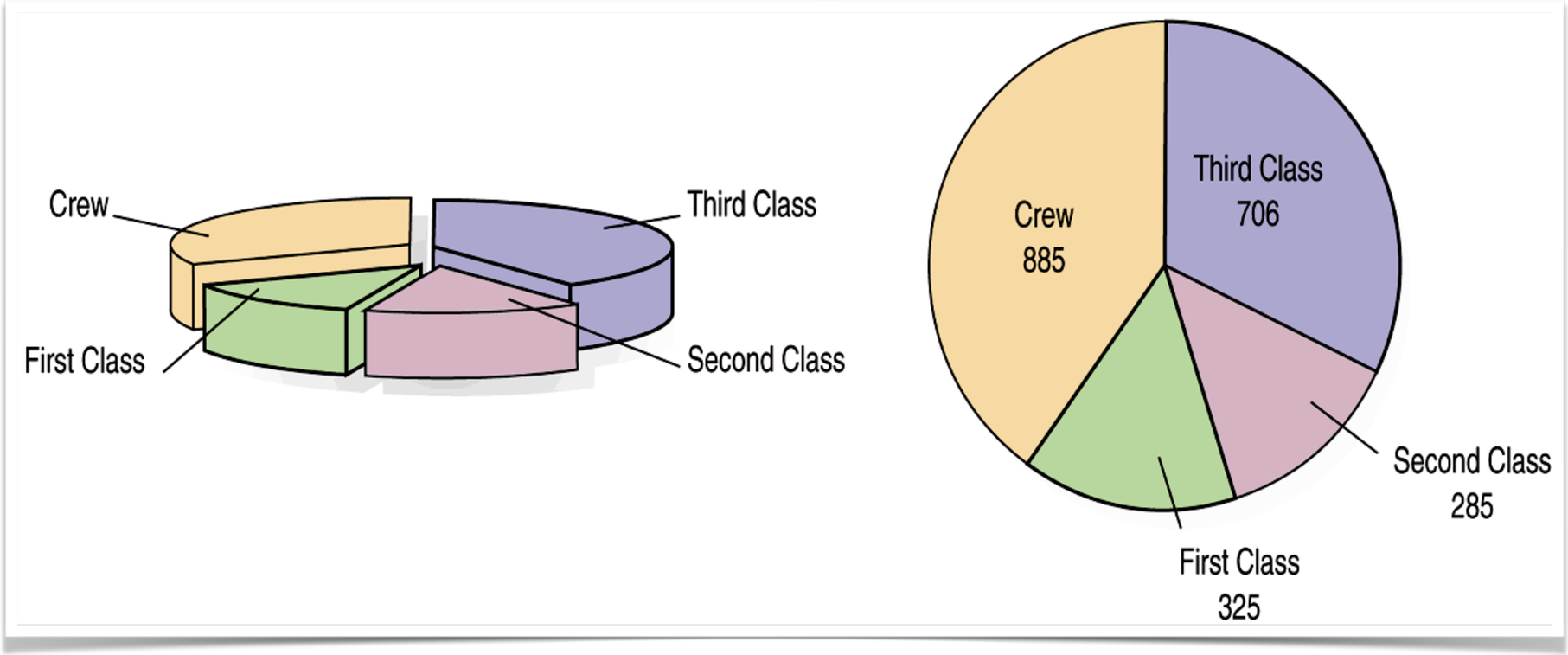
	Gyade	JuChi
Lunch		
Dinner		
Total	150 meals - $700	150 meals - $700

  - Gyade averages $2/lunch and $6/dinner for tips.
  - JuChi averages $3/lunch and $8/dinner in tips.

- Averaging tips across lunch and dinner is unreasonable and is an example of Simpson's Paradox.

What NOT to Do.

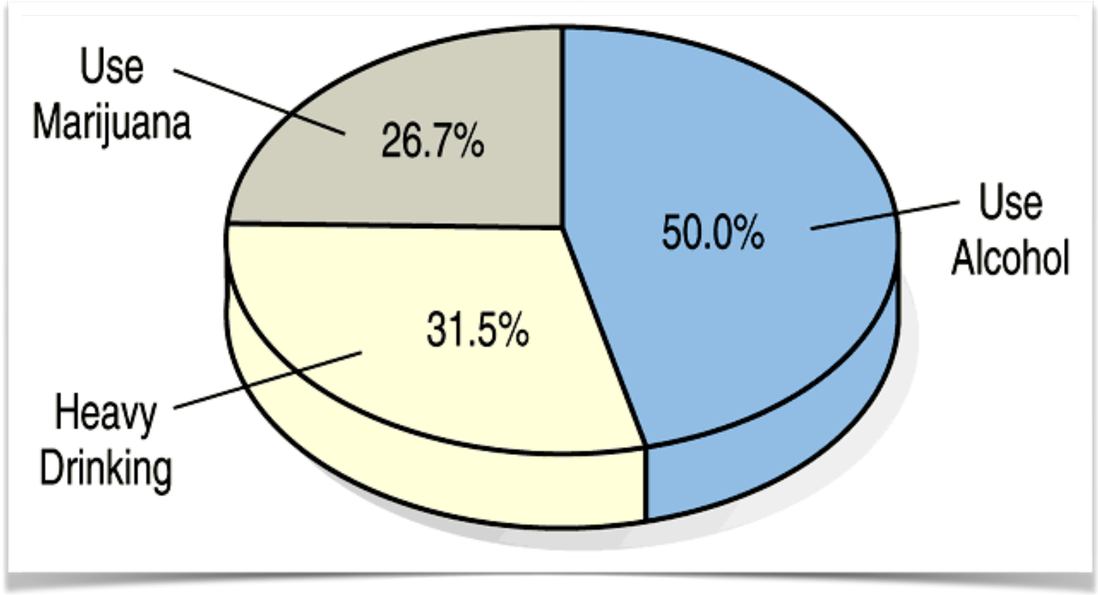
- Do not violate the area principle. In other words, do NOT get cute.



- Some people might like the pie chart on the left better because of the three-dimensional effect. But it is much more difficult to compare fractions of the whole, which is the primary purpose of a pie chart.

What NOT to Do.

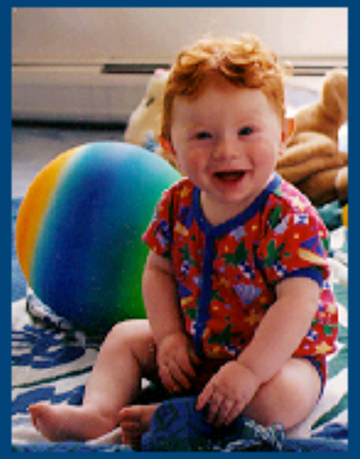
- Make certain your display is honest, and not intended to fool the reader. Your display should show what it purports to show.



- This plot of the percentage of high-school students who engage in specified dangerous behaviors has a few problems.
- List the problems you see.

What NOT to Do.

- Does this make sense?



How common is red hair?

Fair	Red	Medium	Dark	Black
1455	286	2137	1391	116

Total = 5387

What NOT to Do.

- Depends on what you mean by "Fair".

Hair Color

Fair	Red	Medium	Dark	Black
27.01%	5.31%	39.67%	25.88%	21.9%

Total = 5387

Proportion = 1455/5387

.2701 x 100 =27.01%

Caithness County, Scotland

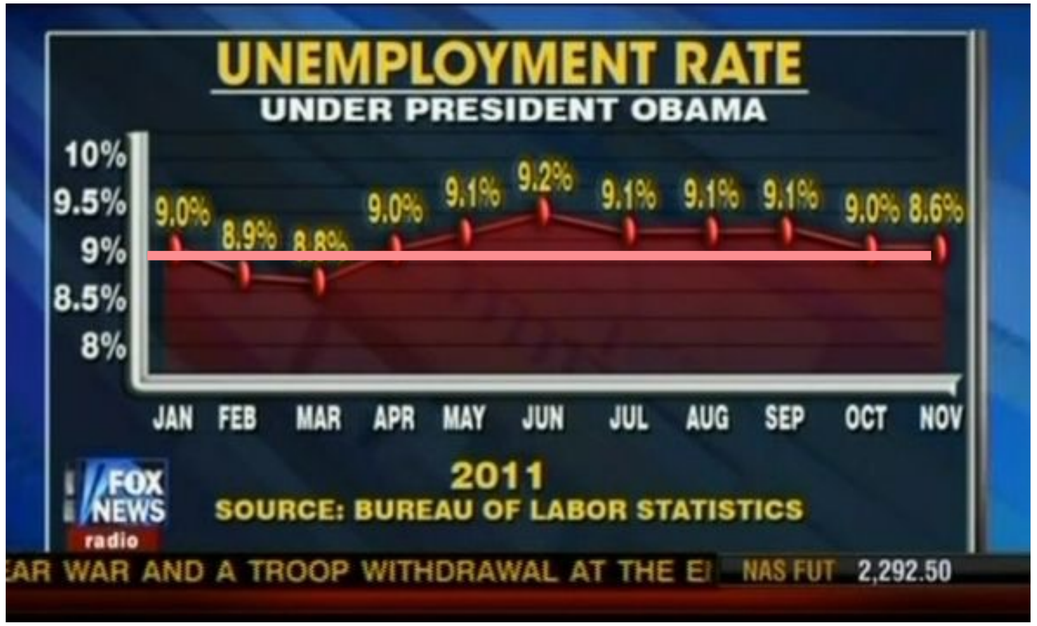
How to Lie With Statistics

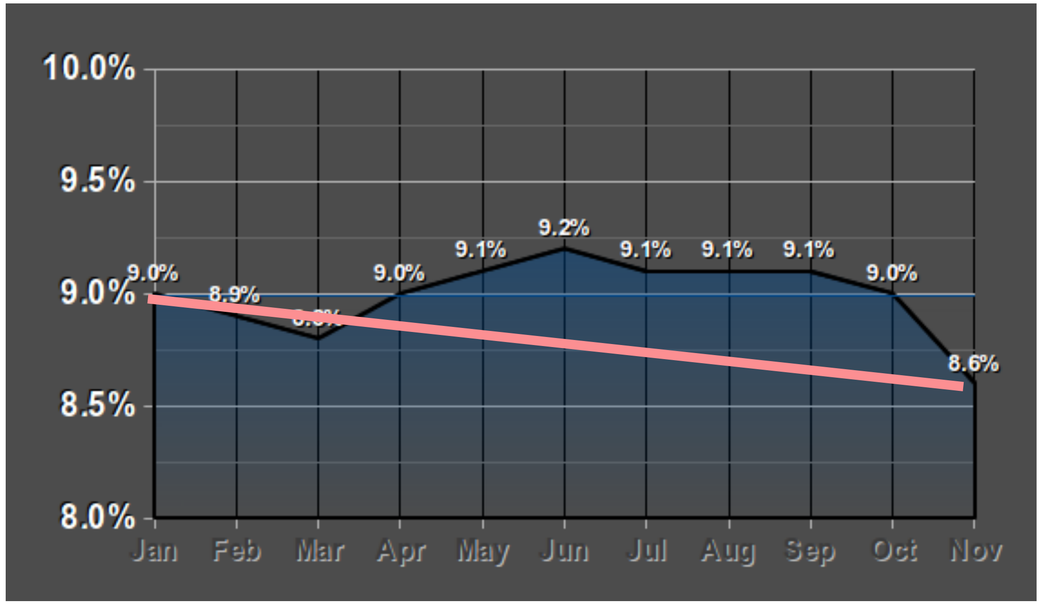
- Statistics do not lie, but there are people who will, intentionally or unintentionally, mislead you. This is especially true of using graphs.

  - View graphs through knowledgeable eyes. Ask how the data was collected, from whom the data was collected, when was the data collected, where did you get your data, and most importantly, why are you being shown the graph.

See anything wrong?

- 9.0% = 8.6% ?
- The maximum differential is only 0.6% (less than 1%).
- Actually indicating a downward trend.
- Note the scale along the vertical axis.

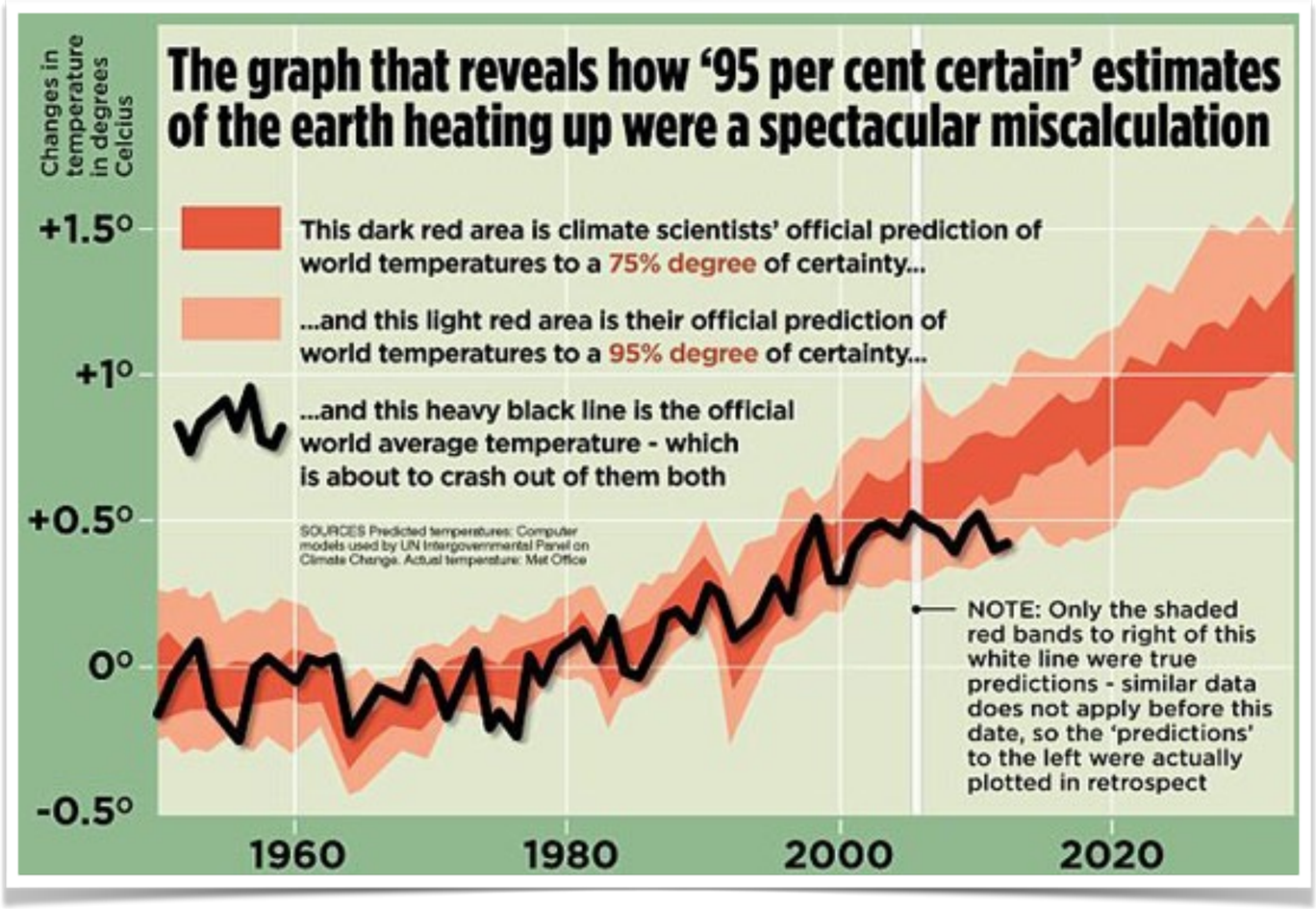




What Global Warming?

- 1st, still inside 95% prediction.

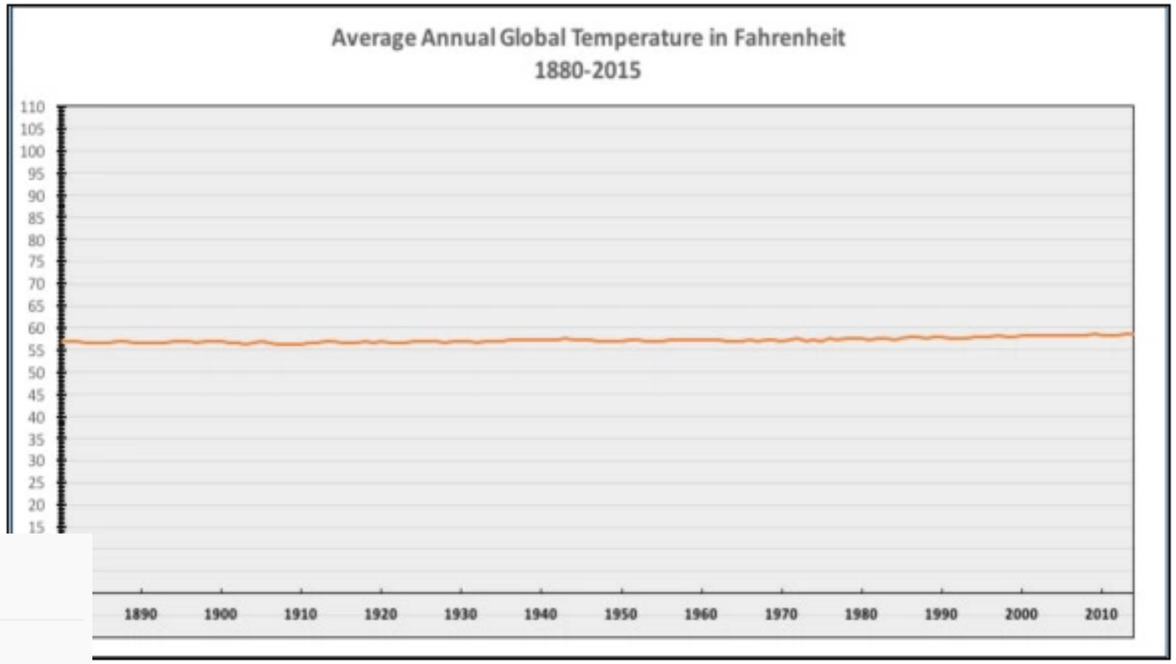
- 2nd, this graph shows air temp? What about ocean temp where most heat resides?

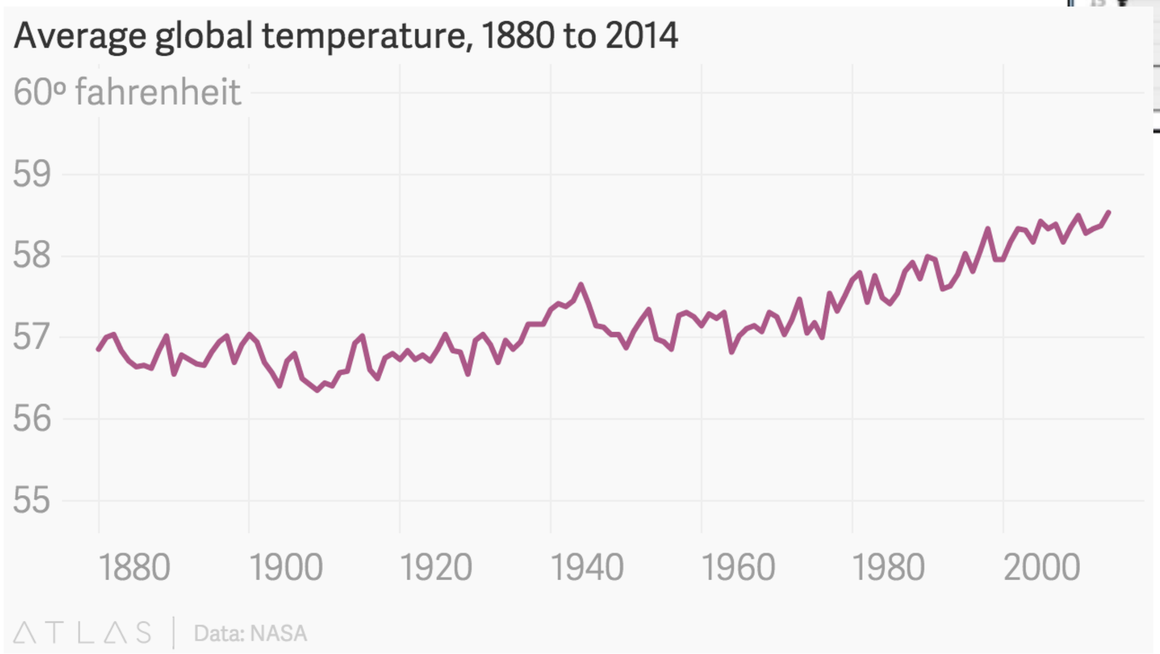


- 3rd, and most importantly, we never trust extrapolation past data. We have no certainty of what comes next.

What Global Warming?

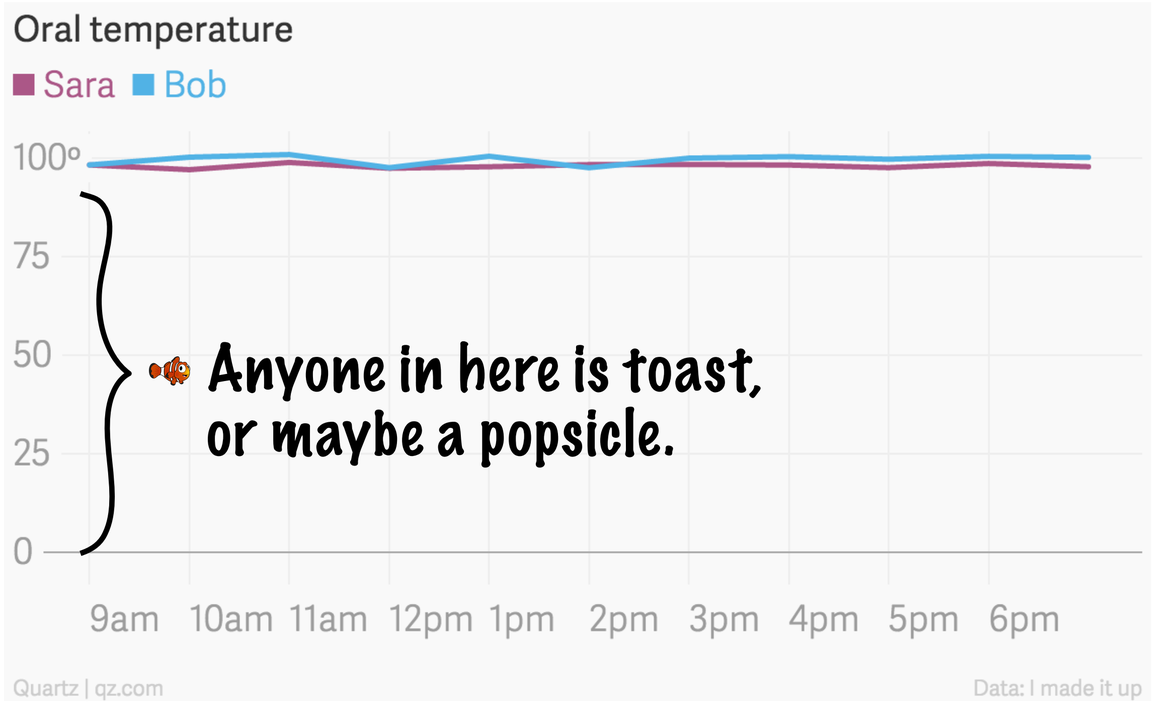
- Sometimes, starting the y axis at zero hides important changes.





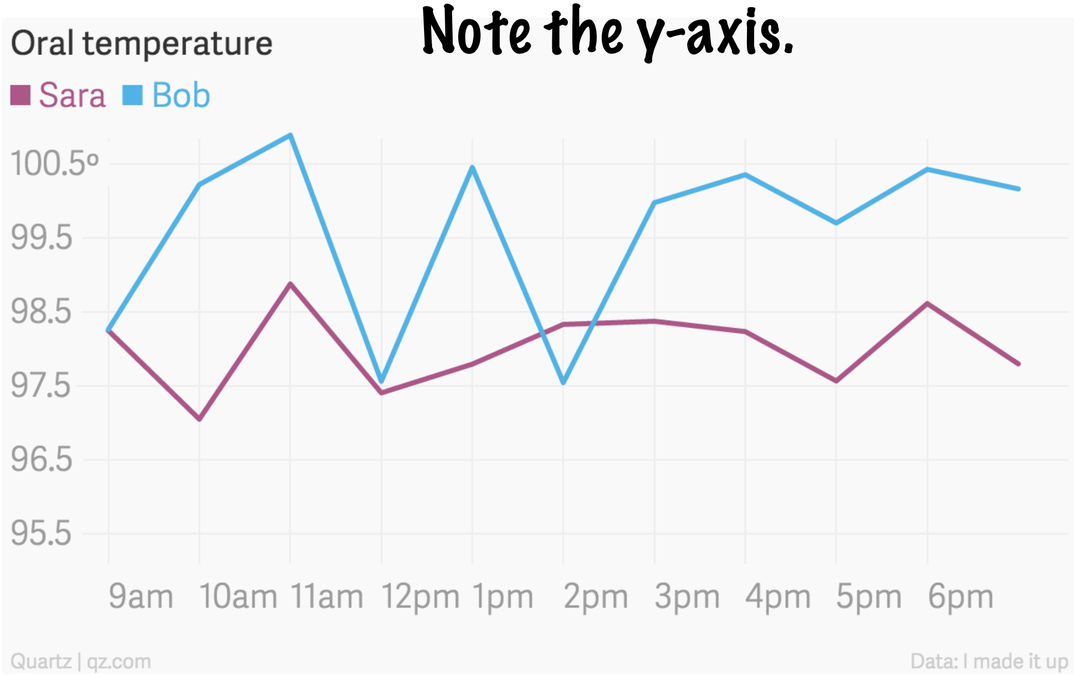
Let Me Illustrate

- Here is an example we can all understand. Who has the fever?



- Anyone in here is toast, or maybe a popsicle.

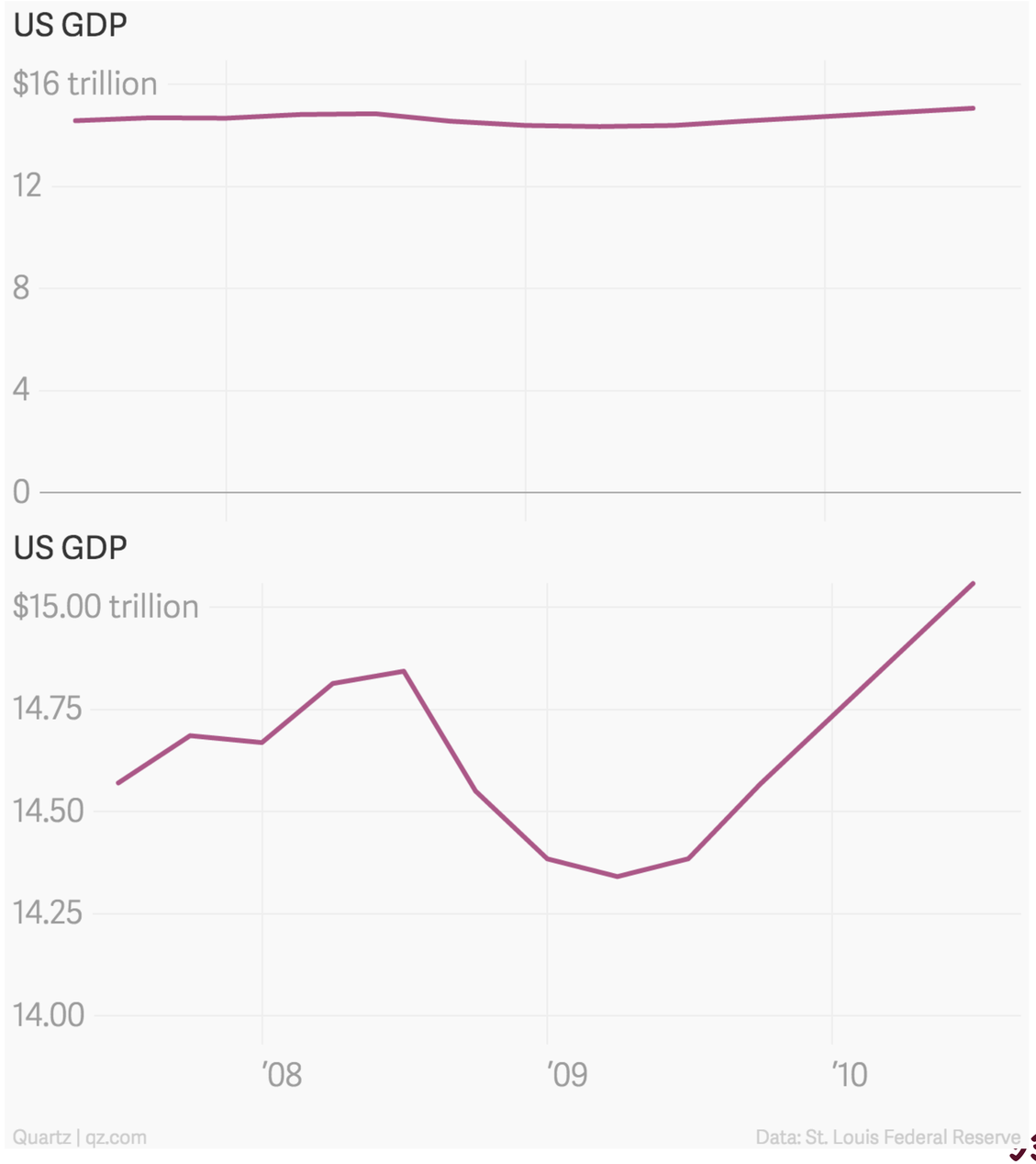
- Aah! Now we can see. Note the y-axis.



What Recovery?

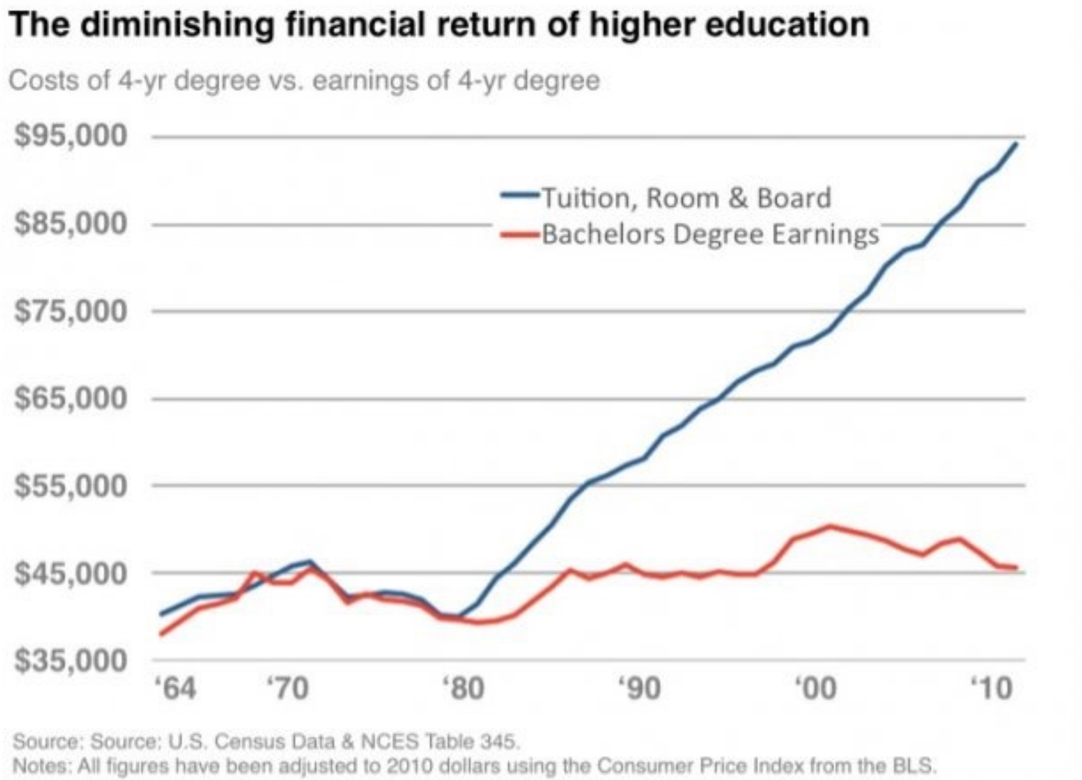
- The economy is a mess. Just look, the growth of the Gross Domestic Product (GDP) is flat.

- Well, you knew this was coming. I guess once Obama took office there was significant growth.



College is a Bad Investment?

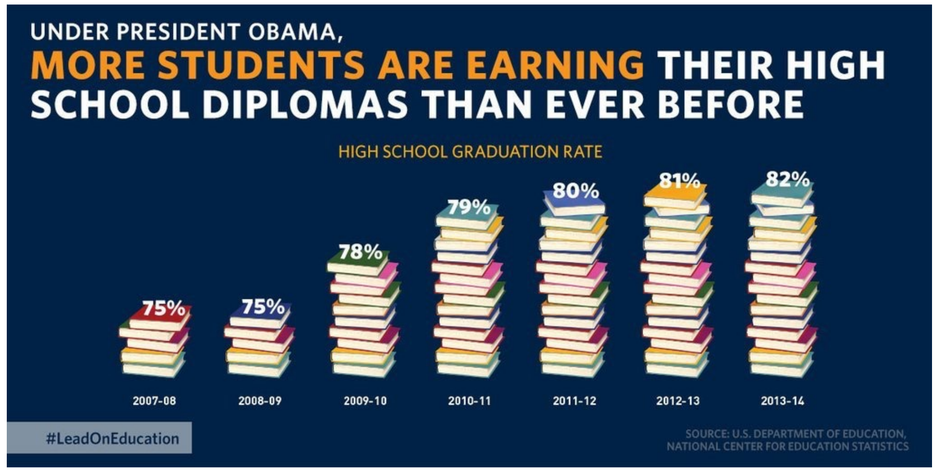
- What about the earnings of those not attending college (Who)? Maybe the differential is worth the investment.

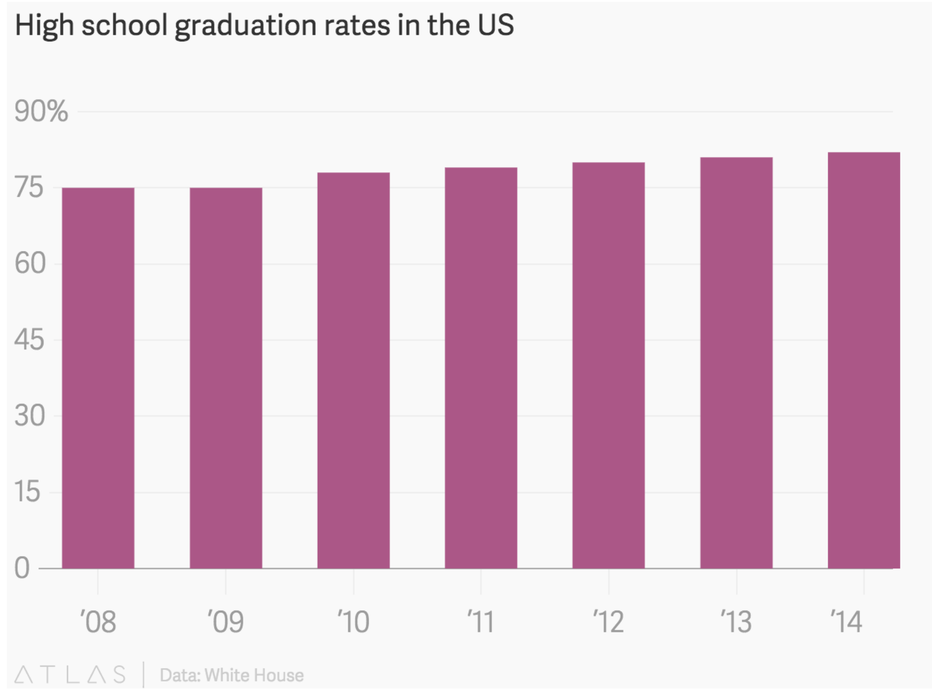


- In this graphic the cost of a 4 year degree is compared directly to the average **1st year** salary. To truly determine the value of that college education, we must compute the expected increase in earnings over not going to college after a **lifetime** of work.

Graduation Rates Are Improving

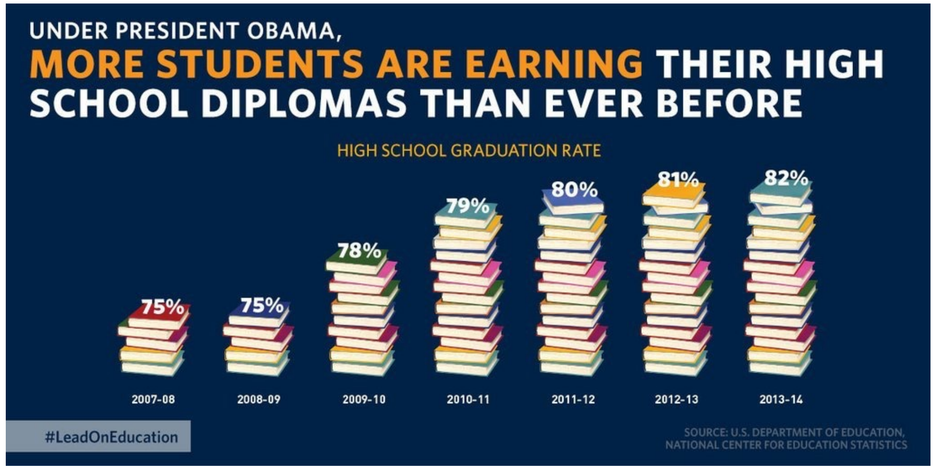
- Here is another close to home. There are several problems with this graphic.
  - **DO NOT** illustrate elements in a graph with pictures (the books here).
  - 5 books equal 75%, thus one book is 15%. 82% should be 5.4667 books.
  - This is how the bar chart should look.
  - But the most serious problem is that the bar chart is not the appropriate graph for this data and what it is intended to show.



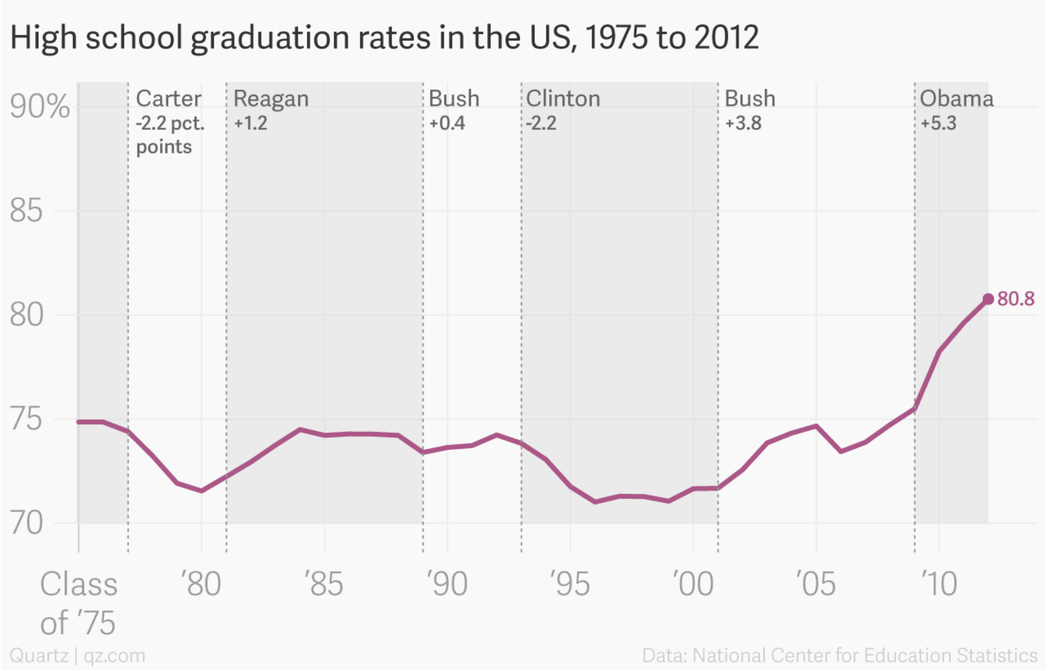


Time Series Graph

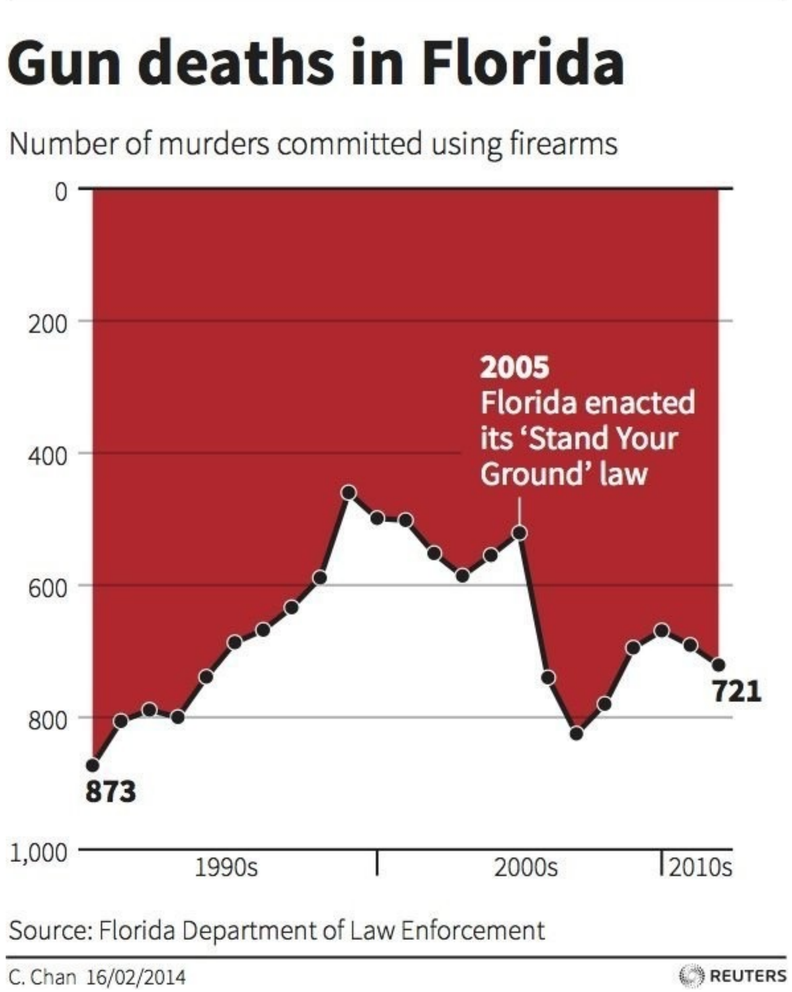
- When the goal is to show changes over time (time series graph) it is preferable to use a frequency polygon (line chart, line graph).



- So yes, graduation rates did increase during Obama's term, but that increase had begun around '96 during the Clinton and Bush years.



Here Is My Personal Favorite!



The only explanation for this graph is an intentional attempt to mislead the reader.