Statistical Thinking for Forensic Practitioners

Quiz on Part 4: Types of Data

1 Misleading visualizations

Data visualizations can be used to summarize characteristics of a dataset. However, if we're not careful, the message we try to convey using these visualizations may be misinterpreted. To learn how to construct a "good" visualization, it can be helpful to examine examples of "bad" or misleading visualizations. We will explore some data visualization faux pas here.

1. Figure 1 shows a bar chart from USA Today. The lesson to be abstracted from this visualization is to always check the axes of a graph.

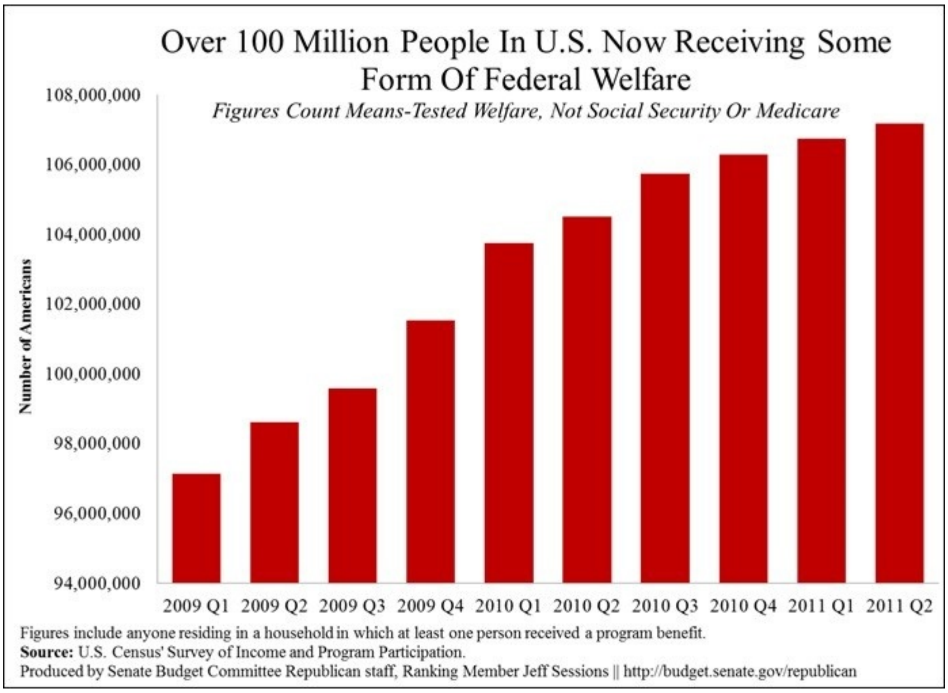


Figure 1: Source

(a) What do you believe is the author's intended message?
The author seems to be suggesting that the number of people in the U.S. on some form of federal welfare has increased considerably between 2009 Q1 and 2011 Q2.

(b) Calculate the actual percentage change in the number of Americans on federal welfare between 2009 Q1 and 2011 Q2 using 97 million and 107 million, respectively, as the true bar heights. How does this affect the intended message of the author you identified above?
$100\% \cdot \left(\frac{107-97}{97}\right) = 10.3\%$

2. Figure 2 shows a bar chart comparing the number of ”People on Welfare” against the number of ”People with a full time job” in the US. We again observe the same axis problem identified in question 1 above. However, it’s also important to be cognizant of the data sources used to construct a visualization.

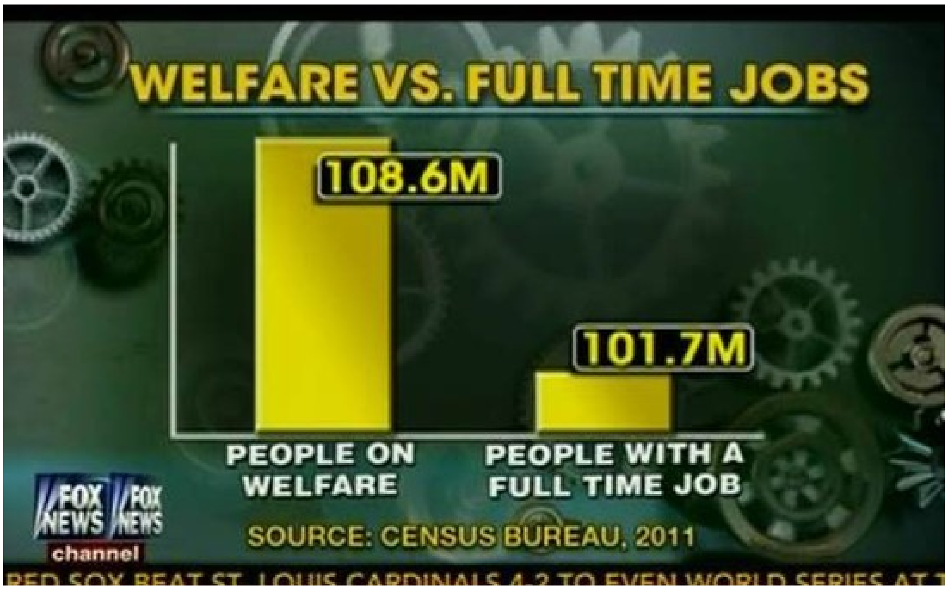


Figure 2: Source

(a) What do you believe is the author’s intended message?

The author seems to be suggesting that the number of people in the U.S. on welfare is much higher than people with a full time job.

(b) The author of this plot retrieved the data from the US Census Bureau in 2011. Consider the following excerpt describing the source of data (Source):

> *Fox’s 108.6 million figure for the number of “people on welfare” comes from a Census Bureau’s account of participation in means-tested programs, which include “anyone residing in a household in which one or more people received benefits” in the fourth quarter of 2011, thus including individuals who did not themselves receive government benefits. On the other hand, the “people with a full time job” figure Fox used included only individuals who worked, not individuals residing in a household where at least one person works.*

How does this affect the intended message of the author you identified above?

There are likely considerably more people who live in the same household as someone with a full time job than 101.7 million. These two bars are not comparable.

3. Consider the following data of the number of graduates from a community college between 1999 and 2003 (Source):

Year	1999	2000	2001	2002	2003
# of Graduates	140	180	200	210	160

Figure 3 graphically displays this number of graduates dataset. These plots demonstrate how visualizations can be used to exaggerate patterns.

(a) Identify the difference in how these two charts were constructed.

The two charts differ in their vertical axis scales.

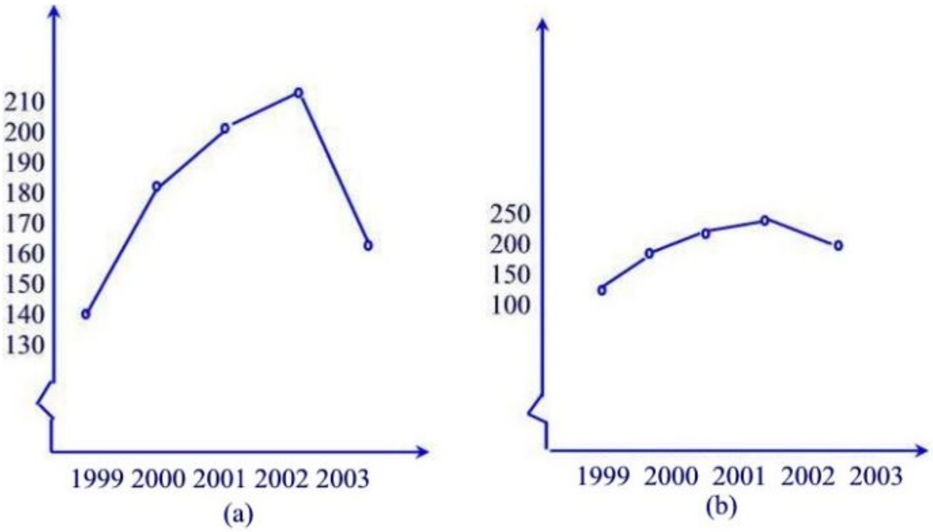
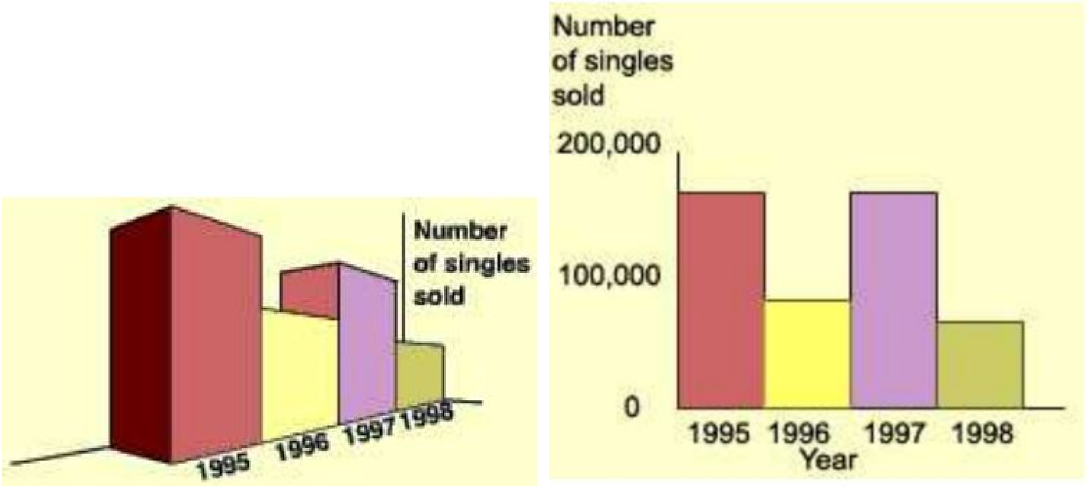


Figure 3: Source

(b) How might the difference identified in part (a) change the conclusions drawn from these two plots?

The first chart seems to indicate a sharp decrease in the number of graduates from 2002 to 2003. This decrease does not seem nearly as drastic in the second plot.

4. Figure 4 shows the number of single mattresses sold by a mattress company in the years 1995 through 1998 as a 3D and 2D box plot. For the reason(s) you will identify below, it is always recommended to use a 2D box plot over a 3D one.



(a) 3D bar chart of Number of singles beds sold by a mattress company

(b) Same data as in Figure 4a shown as a 2D bar chart

Figure 4: Source

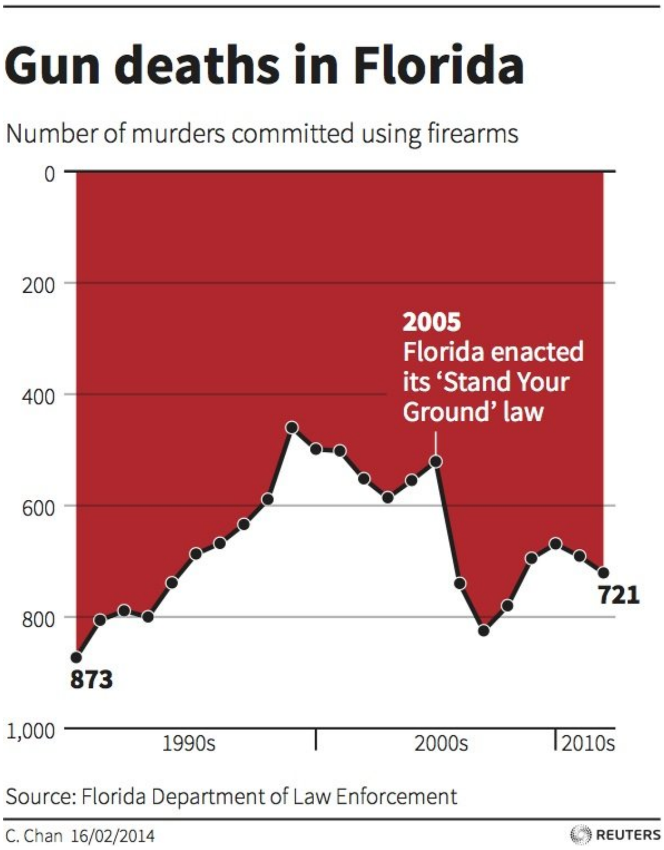
(a) Compared to Figure 4b, what might be misunderstood about the number of singles sold in 1995 and 1997 in Figure 4a?

It appears as if more single mattresses were sold in 1995 than in 1997 in the 3D bar chart. This is not the case, as evidenced by the bars' similar heights in the 2D bar chart.

(b) List one or more features of Figure 4b that it make it easier to compare the number of singles sold between years compared to Figure 4a.

The 2D bar chart actually has a labeled vertical axis. Additionally, as mentioned previously, there isn't a danger of misinterpreting height based on perspective differences in the 2D bar chart.

5. Figure 5 shows a line chart of the number of gun deaths in Florida over time (Source). This example affirms the lesson to always check the axes.



Gun deaths in Florida

Number of murders committed using firearms

2005
Florida enacted its 'Stand Your Ground' law

Source: Florida Department of Law Enforcement

C. Chan 16/02/2014 REUTERS

Figure 5: Source

(a) What do you believe is the author's intended message?

The author seems to be suggesting that the number of gun deaths decreased sharply after the Stand Your Ground law was enacted.

(b) Considering the orientation of the vertical axis, how do the data affect the intended message of the author you identified above?

This sharp change actually corresponds to a sharp increase in gun deaths after Stand Your Ground. Thus, the intended message seems to be directly contradicted by the actual data.

2 Descriptive statistics

Graphs provide a succinct, *visual* summary of characteristics of a dataset. Descriptive statistics provide a *numeric* summary of characteristics of a dataset. Both are useful tools for learning about patterns in data. We will explore the relationship between graphs and descriptive statistics here. Click on the following link to access a Descriptive Statistics applet.

Once you open the applet, you should see on the left a table of values under a “Sample data:” header. These are the data that we will explore, so make sure not to click “Clear” or “Random Sample.” If you do, then reload the page. We will practice identifying different numerical summaries; namely, Mean, Median, Standard Deviation, and Inter-Quartile Range. You should see these 4 statistics listed next to the dataset with “Guess” and “Actual” checkboxes. Do not click the “Actual” checkboxes until instructed. Next to these 4 statistics should be a dotplot graphically showing the 34 datapoints. Dotplots convey similar information as histograms, but histograms are more often used. Click on the “Histogram” option under “Select display type:” to change the plot to a histogram. Do not click on the “Boxplot” option until instructed.

1. Upon changing the plot to a histogram, a “Number of bins:” bar should have appeared. Slide the bar to the left and observe what happens to the plot. Although histograms are meant to summarize continuous quantitative data, why might we avoid setting the number of bins to be very small (e.g., 3, 2, or 1)?

   Too few bins means that we are excessively summarizing our data. Considering the extreme of only one bin yields a histogram in which we lose location information of the individual data points.

2. Now move “Number of bins:” bar to the right and observe what happens to the plot. Why might we avoid setting the number of bins to be large (e.g., 50+ in this case)?

   At the other extreme, we may not be summarizing our data enough if we use a large number of bins. Setting the bin number above 50 yields many bars of height 1, which is not particularly useful if our goal is to understand the overall behavior of the data.

3. Set the number of bins to 5. Based on this visualization of the dataset, describe the shape of the distribution including the modality (uniform, unimodal, bimodal, multimodal) and the symmetry (symmetric, left-skewed, right-skewed). (Hint: see slides 17 and 18 of the Part 4 lecture slides)

   The distribution is unimodal and left skewed.

4. Based on your answer to question 3, is the average of these data less than or greater than the median? Explain. (Hint: think about which of the two, mean or median, is more sensitive to the shape of a distribution)

   The mean tends to be “pulled” towards extreme values or skewness more than the median. Since the data are skewed left, we expect the mean to be “to the left of” or less than the median.

5. Return the number of bins to 20 and click the “Guess” checkbox next to “Mean:”. A red line should appear on the plot. You can click and drag this red line to move it. Drag the red line to where you believe the average is. After you’ve made your guess, click the “Guess” checkbox next to “Median:” and drag the blue line to where you believe the median is. You may record those values here.

   Answers will vary.

6. Now verify whether your guesses were correct by clicking the “Actual” checkboxes next to “Mean:” and “Median:”. Does anything surprise you about where the actual mean and median are? (Note: your answer will be subjective, so there isn’t really a “correct” answer)

   Answers will vary. One may overestimate how separate the mean and median are (the person writing this assignment did, at least).

7. Unclick the checkboxes next to “Mean:” and “Median:” to remove this information from the plot. We will now consider measures of spread. Namely, the standard deviation and the IQR. Considering again your answer to question 3, do you believe the standard deviation will be greater than or less than the IQR. (Hint: consider slides 22/23 of the Part 4 lecture slides think about which measure of spread, the standard deviation or the IQR, would be more sensitive to a handful of data deviating far from the center of the data)

   Similar to the mean, the standard deviation is sensitive to (and grows with) large deviations from the “center” (i.e., mean) of the data. The IQR, being a measure of spread of the middle 50% of the data (75th percentile minus 25th percentile), is less affected by extreme values on the periphery.

8. Click the "Guess" checkbox next to "Std dev:" and a rectangle representing one standard deviation away from the mean will appear on the plot. Move the sides of the rectangle to where you believe the standard deviation is. After you've made your guess, click the "Guess" checkbox next to "IQR:" and drag each of the the blue rectangle's legs to where you believe Q1 and Q3 are. You may record the standard deviation and IQR values here.

   Answers will vary.

9. Now verify whether your guesses were correct by clicking the "Actual" checkboxes next to "Std dev:" and "IQR:". Does anything surprise you about the actual standard deviation and IQR? (Note: again, your answer will be subjective, so there isn't a "correct" answer)

   Answers will vary.

10. Unclick the checkboxes next to "Std dev:" and "IQR:" to remove this information from the plot. Click the "Boxplot" checkbox under the "Select display type" header. A boxplot will appear on the plot. Does the shape of this boxplot affect the conclusions made in question 3? Explain.

    No. The boxplot displays typical left-skewed behavior. The modality of a dataset cannot be ascertained using a boxplot.