STAT 8426 - Homework 1

Brian Detweiler

February 1, 2017

1. Do you think that data visualization has a significant role in exploratory data analysis? Please give five reasons with explanations in favor of your response.

Answer: Yes. Humans can detect patterns in geometrical shapes and colors much faster than in numerical values. Being able to quickly detect a pattern or anomoly gives the data scientist a direction in which to continue the exploratory data analysis.

Often times, we work with very large data sets. Inspecting this raw data in tabular format is just not practical, and therefore, visualizing it becomes one of the best ways to inspect and comprehend the data as a whole.

Because there are many different ways to display data visually, we have a lot of options. The varying options make data visualization possible for many different scenarios, from spatial data, to time series, to categorical, ordinal, or numerical. There are visualization options for everything.

These options allow us to fit multiple variables into a single visual. In this way, data visualization is a form of compression. We can theoretically compress exabytes of data into a single image visual, while maintaining the core message of the data.

Finally, visualizing data provides an extremely effective method of communication that transcends language and education barriers and helps us get our point across in a clear and concise way.

2. Statistical graphics have been commonly used in media. Recently CNN has published eight graphics about gun violence issues in the following link.

[http://www.cnn.com/2015/12/04/us/gun-violence-graphics/?iid=ob_lockedrail_bottomlarge&iref=obinsite] (http://www.cnn.com/2015/12/04/us/gun-violence-graphics/?iid=ob_lockedrail_bottomlarge&iref=obinsite)

For each of the statistical graphics, provide a brief answer using the following table

i) What is the main message?

ii) How is the data encoded into the graphic?

iii) Do you find any problem of the graphics? Explain clearly.

Answer:

Graphics	Message	Encoding	Problem
1. Countries with Most Guns Per Capita	The US is a clear outlier in gun ownership, far ahead of the next highest ranking country.	Number of guns per 100 people by country.	Since the graphic tops out at 100, it would be easy to mistake this as the percentage of people who own guns, when in fact, the number is higher because some people own multiple guns.
2. Firearm Background Checks by Month since 2012	Gun sales show some correlation with mass shootings which cause gun control debates.	Time series area chart. Y-Axis shows number of background checks, supplemented with mass shooting data.	The base of the chart starts at 1 million. The intent is to show the changes, but this also masks the true number of background checks.
3. US Homicide Rate 1960-2014	The homicide rate is the lowest it's been since the '60s	Time series of homicides per 1000 people with individual data points of interest highlighted by dots	No problems.
4. Handguns kill 20x more people than rifles	Breaking down the type of gun involved in homicides, handguns appeared 20x more than shotguns and rifles.	Bar chart of all homicides broken down by type of gun used.	The time period for this plot (2013) is only listed in the source, and since it is only for one year, this could be cherry picked data (possibly an anomalous year).
5. American deaths caused by terrorism vs. gun violence	Terrorism deaths pale in comparison to gun deaths in the US.	Total number of gun deaths vs. total number of terrorism deaths in the US from 2001-2013 encoded by area of the circle.	Although the graphic is visually appealing, it may actually be underselling the point. We have a hard time perceiving area, so a boring bar chart might convey the data more accurately.
6. Suicides account for 3 in 5 gun deaths in the US	Out of all gun deaths, suicides are the overwhelming explanation.	Donut chart of reason for all gun deaths from 2001-2013.	A donut chart is rarely a good choice for a chart. It is essentially a pie chart with the center hallowed out. A better choice may be a lollipop chart.

Graphics	Message	Encoding	Problem
7. Active Shooter Incidents: Where you're most at risk	Shows that business and schools make up the vast majority of areas prone to mass shootings.	Places where mass shootings took place from 2000 - 2013 encoded as a donut chart.	This suffers from the same issues as the previous chart. This one is harder to read because there are more data points. Again, a lollipop chart or simple bar chart would have been better.
8. How active shooter incidents end.	The incidents usually end with either suicide or law enforcement exchanging gunfire. Almost never by a "good guy with a gun".	160 active shooter incidents between 2000-2013.	None. The bar chart is very effective in displaying the differences in reason, and even lists the total number of incidents covered, so we can perceive proportionality.

3. Provide a list of five people who have a significant contribution in the history of data graphics. Explain why you think they should be in your list of five using the following table:

Answer:

Although many names have contributed to data visualization throughout the years, my list wil

Name	Year	Contribution
Christiaan Huygens	1669	The first plot of a CDF, or continuous distribution function. We use these, along with their counterparts, PDFs (probability distribution functions) regularly in statistics. It is a quick visualization of the probability distribution of a function and allows for easy decision making in statistics.
Leland Wilkinson	1999	The Grammar of Graphics. Taking visualization and breaking it into components that can be described in a grammar is groundbreaking.
Edward Tufte	2001	The Visual Display of Quantitative Information - a beautiful and visually appealing book that is packed with how to (and how not to) do data visualization.
John Tukey	1969	The Box Plot. One of the best plots to show the spread of the data with quantiles and outliers with potentially one or more variables.
Mike Bostock	2011	D3.js - Data Driven Documents is one of the most significant contributions to modern statistics because it allows interactive cross-platform data visualization on the Web.

4. In class we talked about the following graphics. Now fill in the following table.

Answer:

Graphics	Why is it famous?	What influence does it have on modern graphics?
Russian Campaign	It encoded six variables into a visually stunning graphic. The Size of the army over time at two-dimensional locations, the direction they were traveling, and the temperature on the return.	Tufte says this may be the best statistical graphic ever drawn. It provides a benchmark by which to judge other graphics.
Cholera Map	Used in determining the location of the water pump that was responsible for many cholera deaths.	This was groundbreaking in the area of geospatial data visualization, which is now ubiquitous today.
Causes of mortality	Used to prove a point that disease was causing the majority of deaths, not war.	While not the most effective graphic, it demonstrates that graphics can be a powerful tool in communicating and influencing.

5. How do you think the contemporary graphics differ from their earlier counter part? Give at least two specific reasons.

Answer:

Reasons	Earlier graphics	Contemporary graphics
1. Contemporary graphics are moving toward interactivity.	Bound by print, and later slower computers.	Computers are fast, and data can be displayed in a cross-platform browser.
2. They are much easier to generate.	Had to plot by hand, or later, have access to a computer with statistical software (which was not cheap)	Can generate a plot in Excel or Google Sheets with little to no effort, and computers are now cheap.

6. Read the article in the following link and answer the questions below

http://www.jstor.org/stable/3087382?seq=9#page_scan_tab_contents

a) What is the main message of the article?

Answer:

Statisticians often talk about the need to display quantitative information visually, but often fail to do so. There is much value in displaying data visually, as it is easier to comprehend for the reader, but it can also save space if done properly.

b) The author mentioned they learnt two things from their research. What are those two leanings?

Answer:

Number	Learning
1	It takes a lot of work to make nice graphs.
2	Nice graphs are possible, especially when we think hard about why we want to display these numbers

c) The author mentioned that there are problems among statisticians to generate data display. To solve that problem, the paper suggested some steps. What are the steps?

Answer:

1. Identify the key comparisons of interest
2. Display these on small individual plots with comparison lines or axes where appropriate
3. Establish enough control over the graphical display so that small legible plots can be juxtaposed as necessary.

d) The article gives several examples of data tables published in the Journal of American Statistical Association (JASA), which could be displayed using graphics. Pick any one of those examples and explain why a graphical display is better than the table itself.

Answer:

The dataset provided by Mehta, Patel, and Senchaudhuri (2000), is a rather large (for a journal article) dataset with three variables. The authors encoded this as a scatter plot wtih different shapes for RSP = 1 or 0. The plot takes up less than half the space, and it allows the reader to visually notice correlation in the number of days and the dose.

7. Provide a list of graphics/plots/diagrams mentioned in the graphics history time line in the link ; http://www.datavis.ca/milestones/

Answer:

- 134 BC Star chart
- 950 Diagram: planetary movements
- 1280 Diagram: paired comparisons
- 1305 Diagram: knowledge
- 1350 Proto-bar graph
- 1450 Graphs of theoretical relation
- 1603 Pantograph
- 1679 Network diagram on a map
- 1686 Weather Map

- 1686 Bivariate plot
- 1693 Mortality Tables
- 1701 Contour Map
- 1712 Literal line graph
- 1724 Abstract line graph
- 1753 Annotated timeline
- 1758 to 1772 Diagrams of color systems
- 1763 Beta density graph
- 1765 Historical timeline
- 1778 Geological map
- 1782 Thematic map
- 1782 First topographical map
- 1782 Geometric proportional figures
- 1785 Superimposed squares
- 1786 Bar chart, line graph invented
- 1795 Proto-nomogram
- 1800 Automatic time-series graph
- 1801 Large-scale geological map
- 1801 Pie chart
- 1811 Subdivided bar graph
- 1819 Choropleth map
- 1826 Choropleth map
- 1828 Mortality curves
- 1829 Comparative choropleth map
- 1829 Polar-area charts
- 1830 Dot Map
- 1833 Population density
- 1837 Flow Map
- 1843 Contour map of 3D table
- 1846 Logarithmic grid
- 1851 Map with diagrams
- 1857 Coxcombs
- 1861 Modern weather map
- 1863 Semilogarithmic grid
- 1869 Periodic table
- 1869 Stereogram
- 1870 Paris election map
- 1873 Semi-graphic table
- 1874 Age pyramid
- 1874 Semi-graphic scatterplot
- 1874 Two-variable color map
- 1875 Lexis diagram
- 1877 Star plot
- 1879 Stereogram
- 1880 Venn diagram
- 1883 to 1885 Multi-function nomograms
- 1884 Alignment diagrams
- 1884 Pictogram
- 1885 Train schedule graphic
- 1888 Anamorphic maps
- 1889 Social mapping
- 1896 Area rectangles
- 1901 Smoothing time series
- 1904 Butterfly diagram: sunspots

- 1910 Diagrams in textbook
- 1910 Diagrams in textbook
- 1911 Hertzsprung-Russell diagram
- 1914 Pictogram
- 1917 Gantt chart
- 1919 Statistical chartbook
- 1920 Path diagram
- 1925 Control chart
- 1928 Ideograph
- 1928 Nomogram
- 1929 Electroencephalograph
- 1930 Timeline on log scale
- 1933 London Underground map
- 1957 Circular glyphs
- 1966 Triangular glyphs
- 1971 Biplot
- 1971 Star plot
- 1973 Chernoff faces
- 1975 Circular display
- 1975 Scatterplot
- 1977 Cartesian rectangle
- 1978 Linked brushing
- 1979 Geographic correlation diagram
- 1981 Mosaic display
- 1982 Brushing
- 1982 Visibility base map
- 1983 Sieve diagram
- 1985 Grand tour
- 1985 Parallel coordinates plot
- 1987 Interactive linked graphics
- 1988 Interactive grand tours
- 1988 Interactive time-series
- 1989 Interactive maps
- 1990 Multivariate grand tours
- 1990 Textured dot strips
- 1991 Enhanced mosaic display
- 1991 Treemaps
- 2002 Tag cloud, Word cloud
- 2004 Sparkline
- 2009 Chord diagram