# **Data Visualization**

Jarek Szlichta http://data.science.uoit.ca/

## Visualization

- Looking at data through visualizations can be enlightening and powerful
  - and, sometimes, deceptive...
- Data visualization is a science and an art, and can be very hard to get right;
  - we'll skim the surface
- Basic tools today:
  - Google Spreadsheets
  - Python
  - **R**

## Introductory Video

https://www.youtube.com/watch?v=jbkSRLYSojo



# **Types of Charts**

Bar Chart

### Scatterplots

### Maps



300,000

(T900,000 -

ż



**Top Oil Reserves** 



# **Types of Charts**

### 2D Pie Chart



3d Pie Chart



# **Usefulness of Charts**

### Bar Charts

- Useful when one axis is categories and the other is numeric
- Pie charts
  - Useful when comparing sizes of categories
- Scatterplots
  - Useful when both axes are numeric
- Map
  - Values by geographical region

### Raw Tool

- More advanced/exotic visualization Raw tool
  - <u>http://raw.densitydesign.org/</u>
  - Video:
    - http://raw.densitydesign.org/how-it-works/

0.00	× Cosa resterà di quei ×			
← ⇒ C fi	D app.raw.densitydesign.org			습 🗿 💩 🙆 🖨 🗏
	Choose a Chart			
	Scatter Plot. — A scatter plot, scatterplot, or scattergraph is a type of mathematical diagram using Cartesian coordinates to display values for new variables for a set of data. The data is displayed as a cellection of paints, set having the value of one variable determining the position on the variable data of the value of the other variable data of the other variable data of the other variable data of the value of the other variable data of the other othe	Anviel Diagram (Fineo-Like)	Convex Hull	The two two is enabled of two is the results of two is the results
		Hexagonal Binning	Arallel Coordinates	The time ( The time ( times the time in the finance) Fring time ( The time
		Tastic Fit	Etreargraph	Tasmap
			0	

# **Common Visualization Mistakes**

- Professional data visualizers get excited by the stories they want to tell
  - But sometimes they forget to follow some best practices while doing so
  - It happens to the best of us
- We are going to cover what those best practices are..

### Do not Truncate Axes

- One of the ways a graph can be distorted is by truncating an axis
- This happens when an axis is shortened because one or both of its ends gets cut off

### **Truncating Axis: Allergy Medicines**

- Sometimes a distortion like this is really obvious
- Let's say there are two allergy medicines called Happajoy and Pollaway



#### **Effectiveness of Allergy Medicines**

### **Truncating Axis: Allergy Medicines**

- The bar graph compares how effective these two medicines are at reducing allergy symptoms
- If you quickly glance at the bars, you may think that Happajoy is twice as effective as Pollaway is because its bar is twice as tall
  - But if you examine the graph more closely, you'll see that the y-axis is truncated, starting from 30.2 and going up to only 30.7 percent
  - Happajoy's effectiveness is only 0.2% higher than Pollaway's,
    - which is not as impressive as the results "implied" by the bar graph..

# Happajoy Sales

- Sometimes a truncated axis and the resulting distortion can be more subtle
- the next graph shows the quantity of Happajoy sold from January through April 2014

Happajoy Sales in 2014



## What is the Problem?

- Y-axis is not a problem.
- X-axis has been truncated this time
  - We are seeing sales from less than half the year
  - Truncating a time period like this can give the wrong impression, especially for things that go through cycles
  - The sale of allergy medicine goes through a seasonal cycle since allergy symptoms are typically higher in the spring and lower in the winter

## **Better Way to Show Sales**

- This time the y-axis is proportional and the x-axis covers two full years
  - Sales of Happajoy went down in the winter and up in the spring, but that the rate of sales did not change much from year to year instead of just a few months

Happajoy Sales in 2013-2014



CSCI 2000U Scientific Data Analysis - Jarek Szlichta

# Do not Omit Key Variables

- Context is important, especially when it comes to data
- Carefully consider the ecosystem in which studied variable exists
  - and deliberately seek out other variables that could affect the one you're studying

*C* The first principle is that you must not fool yourself — and you are the easiest person to fool.

- Richard Feynman, 1974 Caltech Graduation Address

### Leading Allergy Medication in Each State



Leading Allergy Medication in Each State

# **Leading Allergy Medication**

- It might seem like Happajoy is the market leader nationally, ahead of Pollaway
- But to get the complete picture you have to pay attention to other variables..

## **Breakdown of Market Shares**

 The margins by which Happajoy leads are significantly less than the margins by which Pollaway leads...
Market Share in Each State



Percent of Allergy Medication Market

CSCI 2000U Scientific Data Analysis - Jarek Szlichta

# Adding up the Numbers

- When you add up the numbers, Pollaway's total sales are more than twice that of Happajoy's
  - The total sales in states where Happajoy is the leader is also significantly less than the total sales in states where Pollaway is the leader..

States	Happajoy	Pollaway	
Wyoming (WY)	299,734	219,037	
North Dakota (ND)	349,814	279,851	
South Dakota (SD)	408,343	341,675	
Montana (MT)	482,400	422,100	
ldaho (ID)	782,040	654,360	
Nebraska (NE)	872,320	798,080	
New Mexico (NM)	1,043,000	834,400	
Nevada (NV)	1,489,860	993,240	
Utah (UT)	1,313,300	1,256,200	
Kansas (KS)	1,414,140	1,183,260	-

# **Do Not Oversimplify**

- The CEO of RediMedico starts off the presentation with the graphic below.
- First you start daydreaming...
- Then, you ask yourself:
  - Compared to what?
  - Compared to when?
  - Compared to whom?



# **Revenue from Top 5 Sellers**

- Thankfully, the CEO of RediMedico agrees and presents the next graphic..
  - which compares the revenues from the five top-selling medicines RediMedico makes:



#### **Revenue from Top 5 Sellers**

# **Number Crunching**

- Increase in revenue between 2013 and 2014 is indeed 18%
- This increase is primarily due to a whopping 225% increase in revenue from a single medicine, Exoalgio
  - Revenue from 3 out of 5 medicines actually dropped
- Using a graphic with a single number and no breakdowns is like writing a news headline without the news story

# Do not Choose the Wrong Form

- Creating a data visualization is a balancing act between form and function
- When choosing a graphic format for your data, you'll have to figure out how to effectively communicate to your audience
  - in an aesthetically pleasing way
- Let's take a look at a couple graphics that use area and shading

### Units Sold by Market Leader in Each State



# **Bubble Graphic**

- The bubble graphic uses area to display the units sold of the top selling allergy medicine in some states.
- Based on bubble size, you can generally tell that more Happajoy was sold in Arizona than in New Mexico
  - But can you tell by how much?
  - Is the Arizona bubble three times bigger than the New Mexico bubble? Four times? It's hard to tell..
  - It is even harder to tell when the bubble sizes are closer together: Did Utah or Kansas sell more Happajoy?

# **Units of Happajoy Sold**

 We run into the same problem with the next graphic, which uses shading to represent Happajoy sales



# **Overall Picture vs Comparisions**

- California is definitely darker than Texas, but how much darker? Two times? Three times? Who knows?
- This is why area and shading are better for giving an overall picture instead of making precise comparisons

### **Percent of Total Revenue**

 If the goal of this graphic is to help readers compare revenues from one year to the next, then something like a bar or pie chart would have been a better choice



### **Do Present Data in Multiple Ways**

- We just covered how different graphic forms are good at doing different things
- So what do you do when you have a lot of data and you want to show different aspects of those data?
- One way to deal with this challenge is to present your data in multiple ways
  - so that taken together your audience gets a more accurate picture of the data as a whole



 For every year, a map that uses shading to represent sales can give people a general sense of how sales have changed across time and location



## **Sales in One State**

 Let's say RediMedico started to advertise heavily in California and New York a few years ago, and the investors are wondering how sales in those states are doing

Happajoy Revenue



## **Multiple States**

### Or you can compare the sales between different states:

**Happajoy Revenue** 



CSCI 2000U Scientific Data Analysis - Jarek Szlichta

### **Do Annotate with Text**

- They say that "a picture is worth a thousand words", but that doesn't mean you should forget about words entirely!
  - You might be wondering what happened between 2009 and 2010?



Happajoy Total Annual Revenue

CSCI 2000U Scientific Data Analysis - Jarek Szlichta

### Headline

### In this case, it would be helpful to add some headline text:

#### Happajoy Total Annual Revenue



# Case Study of an Awesome Infographic

## Case Study

- Let's take a look at all of these pro tips in action by going through a visualization made by real data pros:
  - http://www.nytimes.com/2013/10/16/health/uga nda-fights-stigma-and-poverty-to-take-on-breastcancer.html?\_r=0#g-graphic
- The designers organized the information as a narrative with a step-by-step structure
- This is an interactive graphic, so it's best if you click through the link to get the full experience

### **Case Study**

#### Where is Breast Cancer?

Breast cancer is the world's most common cancer in women, with an estimated 1.4 million new cases diagnosed every year.





- On the first screen, you see a bubble graphic that gives you a general sense of which countries have the most new cases of breast cancer
- After clicking "Begin", you see a scatterplot with proportional axes
  - The scatterplot shows that there is an inverse correlation between breast cancer detection and mortality: as more women are detected with breast cancer, fewer women die from it

# **Pro Tips**

- Notice how the designers use text to write a good headline that grabs the reader's attention ("Where Does Breast Cancer Kill?")
- As you keep clicking "Next", the designers guide you deeper into the scatterplot by highlighting a cluster of countries and providing an annotation that gives you further insight into that cluster
- The designers also present the data in multiple ways and use annotations
  - They use color to add another layer of detail: the development status of various countries
  - if you're curious about the statistics for a specific country, you can mouse over that country's dot to get those numbers

# **Reading List**

### Recommended

- Chart Types
  - https://support.google.com/docs/answer/190718
- Common Visualization Mistakes
  - https://infoactive.co/data-design/ch18.html
- New York Times Article
  - http://www.nytimes.com/2013/10/16/health/ugandafights-stigma-and-poverty-to-take-on-breastcancer.html?\_r=1#g-graphic