Session 1: Do's and Don'ts of Data Visualization

Presenter:

Randy Olson, Ph.D.
DO’S AND DON’TS OF DATA VISUALIZATION

Dr. Randy Olson
rso@randalolson.com
@randal_olson
randalolson.com
ABOUT ME

- AI researcher
  - Postdoc, Computational Genetics Lab @ Penn

- Data tinkerer
  - Data blog: randalolson.com/blog/
  - Community leader: reddit.com/r/DataIsBeautiful

Dr. Randy Olson | Email: rso@randalolson.com | Twitter: @randal_olson | Web: randalolson.com
DATA VIZ TURNS TABLES OF NUMBERS...

<table>
<thead>
<tr>
<th>YEAR</th>
<th>WEEK</th>
<th>ALABAMA</th>
<th>ALASKA</th>
<th>ARIZONA</th>
<th>ARKANSAS</th>
<th>CALIFORNIA</th>
<th>COLORADO</th>
<th>CONNECTICUT</th>
<th>DELAWARE</th>
<th>DISTRICT OF COLUMBIA</th>
<th>FLORIDA</th>
</tr>
</thead>
<tbody>
<tr>
<td>1928</td>
<td>1</td>
<td>3.67</td>
<td>-</td>
<td>1.9</td>
<td>4.11</td>
<td>1.38</td>
<td>8.38</td>
<td>4.5</td>
<td>5.8</td>
<td>0.21</td>
<td></td>
</tr>
<tr>
<td>1928</td>
<td>2</td>
<td>6.25</td>
<td>-</td>
<td>6.4</td>
<td>9.91</td>
<td>1.8</td>
<td>6.02</td>
<td>9</td>
<td>7.3</td>
<td>0.49</td>
<td></td>
</tr>
<tr>
<td>1928</td>
<td>3</td>
<td>7.65</td>
<td>-</td>
<td>4.5</td>
<td>11.15</td>
<td>1.31</td>
<td>2.89</td>
<td>8.81</td>
<td>15.8</td>
<td>0.42</td>
<td></td>
</tr>
<tr>
<td>1928</td>
<td>4</td>
<td>12.58</td>
<td>-</td>
<td>1.9</td>
<td>13.75</td>
<td>1.81</td>
<td>13.71</td>
<td>10.4</td>
<td>1.42</td>
<td>1.49</td>
<td>0.91</td>
</tr>
<tr>
<td>1928</td>
<td>5</td>
<td>8.03</td>
<td>-</td>
<td>0.47</td>
<td>20.79</td>
<td>2.38</td>
<td>5.13</td>
<td>16.8</td>
<td>5.58</td>
<td>4.59</td>
<td>0.49</td>
</tr>
<tr>
<td>1928</td>
<td>6</td>
<td>7.27</td>
<td>-</td>
<td>6.4</td>
<td>26.58</td>
<td>2.79</td>
<td>8.09</td>
<td>17.76</td>
<td>3.45</td>
<td>7.52</td>
<td>1.67</td>
</tr>
<tr>
<td>1928</td>
<td>7</td>
<td>10.0</td>
<td>-</td>
<td>0.95</td>
<td>32.76</td>
<td>2.73</td>
<td>3.94</td>
<td>20.16</td>
<td>4.29</td>
<td>1.27</td>
<td>1.12</td>
</tr>
<tr>
<td>1928</td>
<td>8</td>
<td>13.83</td>
<td>-</td>
<td>1.66</td>
<td>36.44</td>
<td>2.83</td>
<td>4.34</td>
<td>22.7</td>
<td>7.05</td>
<td>3.45</td>
<td>0.77</td>
</tr>
<tr>
<td>1928</td>
<td>9</td>
<td>11.06</td>
<td>-</td>
<td>0.95</td>
<td>33.89</td>
<td>3.84</td>
<td>2.96</td>
<td>22.7</td>
<td>4.01</td>
<td>33.59</td>
<td>2.17</td>
</tr>
<tr>
<td>1928</td>
<td>10</td>
<td>13.98</td>
<td>-</td>
<td>4.03</td>
<td>29.18</td>
<td>5.31</td>
<td>4.04</td>
<td>23.91</td>
<td>4.29</td>
<td>2.23</td>
<td></td>
</tr>
<tr>
<td>1928</td>
<td>11</td>
<td>18.79</td>
<td>-</td>
<td>7.35</td>
<td>20.84</td>
<td>3.5</td>
<td>3.35</td>
<td>25.24</td>
<td>6.44</td>
<td>3.35</td>
<td></td>
</tr>
<tr>
<td>1928</td>
<td>12</td>
<td>22.2</td>
<td>-</td>
<td>6.4</td>
<td>27.4</td>
<td>4.38</td>
<td>2.73</td>
<td>19.09</td>
<td>9.01</td>
<td>38</td>
<td>4.88</td>
</tr>
<tr>
<td>1928</td>
<td>13</td>
<td>21.97</td>
<td>-</td>
<td>7.82</td>
<td>30.1</td>
<td>3.44</td>
<td>4.34</td>
<td>20.1</td>
<td>8.15</td>
<td>4.61</td>
<td></td>
</tr>
<tr>
<td>1928</td>
<td>14</td>
<td>16.78</td>
<td>-</td>
<td>7.82</td>
<td>13.21</td>
<td>2.34</td>
<td>18.05</td>
<td>23.53</td>
<td>6.01</td>
<td>48.85</td>
<td>2.93</td>
</tr>
<tr>
<td>1928</td>
<td>15</td>
<td>15.45</td>
<td>-</td>
<td>8.06</td>
<td>23.06</td>
<td>2.21</td>
<td>8.48</td>
<td>23.4</td>
<td>11.59</td>
<td>5.65</td>
<td></td>
</tr>
<tr>
<td>1928</td>
<td>16</td>
<td>14.89</td>
<td>-</td>
<td>13.27</td>
<td>13.37</td>
<td>2.98</td>
<td>10.06</td>
<td>23.02</td>
<td>6.01</td>
<td>39.67</td>
<td>6.42</td>
</tr>
<tr>
<td>1928</td>
<td>17</td>
<td>16.14</td>
<td>-</td>
<td>1.9</td>
<td>21.28</td>
<td>2.08</td>
<td>9.47</td>
<td>22.45</td>
<td>15.02</td>
<td>35.07</td>
<td>6.56</td>
</tr>
<tr>
<td>1928</td>
<td>18</td>
<td>9.51</td>
<td>-</td>
<td>2.13</td>
<td>24.31</td>
<td>2.25</td>
<td>18.15</td>
<td>24.03</td>
<td>12.88</td>
<td>9.91</td>
<td></td>
</tr>
<tr>
<td>1928</td>
<td>19</td>
<td>13.86</td>
<td>-</td>
<td>28.2</td>
<td>19</td>
<td>2.04</td>
<td>1.97</td>
<td>24.16</td>
<td>18.45</td>
<td>37.79</td>
<td>7.05</td>
</tr>
<tr>
<td>1928</td>
<td>20</td>
<td>14.02</td>
<td>-</td>
<td>1.18</td>
<td>16.57</td>
<td>2.5</td>
<td>13.71</td>
<td>17.69</td>
<td>17.17</td>
<td>48.85</td>
<td>4.88</td>
</tr>
<tr>
<td>1928</td>
<td>21</td>
<td>13.67</td>
<td>-</td>
<td>2.13</td>
<td>9.64</td>
<td>1.67</td>
<td>12.43</td>
<td>22.45</td>
<td>8.58</td>
<td>39.87</td>
<td>9.28</td>
</tr>
<tr>
<td>1928</td>
<td>22</td>
<td>9.92</td>
<td>-</td>
<td>2.13</td>
<td>9.04</td>
<td>1.68</td>
<td>-</td>
<td>22.26</td>
<td>16.31</td>
<td>44.89</td>
<td>12.21</td>
</tr>
<tr>
<td>1928</td>
<td>23</td>
<td>8.3</td>
<td>-</td>
<td>3.32</td>
<td>9.8</td>
<td>1.33</td>
<td>6.61</td>
<td>21.56</td>
<td>4.29</td>
<td>40.08</td>
<td>5.51</td>
</tr>
<tr>
<td>1928</td>
<td>24</td>
<td>5.72</td>
<td>-</td>
<td>48.58</td>
<td>5.52</td>
<td>0.92</td>
<td>8.38</td>
<td>18.33</td>
<td>6.87</td>
<td>25.89</td>
<td>8.23</td>
</tr>
<tr>
<td>1928</td>
<td>25</td>
<td>4.81</td>
<td>-</td>
<td>1.84</td>
<td>1.84</td>
<td>0.71</td>
<td>5.82</td>
<td>18.58</td>
<td>7.73</td>
<td>27.14</td>
<td>6.42</td>
</tr>
<tr>
<td>1928</td>
<td>26</td>
<td>4.02</td>
<td>-</td>
<td>1.42</td>
<td>3.3</td>
<td>0.71</td>
<td>0.89</td>
<td>15.66</td>
<td>15.02</td>
<td>19.83</td>
<td>1.47</td>
</tr>
<tr>
<td>1928</td>
<td>27</td>
<td>1.7</td>
<td>-</td>
<td>0.24</td>
<td>0.52</td>
<td>0.43</td>
<td>4.64</td>
<td>18.64</td>
<td>2.58</td>
<td>17.33</td>
<td>3.28</td>
</tr>
<tr>
<td>1928</td>
<td>28</td>
<td>1.63</td>
<td>-</td>
<td>10.43</td>
<td>0.54</td>
<td>0.37</td>
<td>4.83</td>
<td>12.18</td>
<td>3.43</td>
<td>5.43</td>
<td>2.44</td>
</tr>
<tr>
<td>1928</td>
<td>29</td>
<td>1.29</td>
<td>-</td>
<td>0.95</td>
<td>2.11</td>
<td>0.41</td>
<td>3.55</td>
<td>9.38</td>
<td>-</td>
<td>0.63</td>
<td>-</td>
</tr>
<tr>
<td>1928</td>
<td>30</td>
<td>2.5</td>
<td>-</td>
<td>1.18</td>
<td>0.92</td>
<td>0.36</td>
<td>1.58</td>
<td>6.91</td>
<td>-</td>
<td>0.28</td>
<td>-</td>
</tr>
<tr>
<td>1928</td>
<td>31</td>
<td>0.23</td>
<td>-</td>
<td>1.9</td>
<td>0.27</td>
<td>0.37</td>
<td>0.99</td>
<td>4.25</td>
<td>-</td>
<td>3.97</td>
<td>0.91</td>
</tr>
<tr>
<td>1928</td>
<td>32</td>
<td>0.53</td>
<td>-</td>
<td>0.52</td>
<td>0.22</td>
<td>0.21</td>
<td>1.08</td>
<td>2.47</td>
<td>-</td>
<td>0.84</td>
<td>-</td>
</tr>
<tr>
<td>1928</td>
<td>33</td>
<td>0.27</td>
<td>-</td>
<td>0.71</td>
<td>0.05</td>
<td>0.07</td>
<td>0.59</td>
<td>1.2</td>
<td>1.2</td>
<td>1.46</td>
<td>0.07</td>
</tr>
<tr>
<td>1928</td>
<td>34</td>
<td>1.06</td>
<td>-</td>
<td>0.95</td>
<td>0.27</td>
<td>0.21</td>
<td>0.2</td>
<td>0.7</td>
<td>0.43</td>
<td>0.63</td>
<td>0.14</td>
</tr>
<tr>
<td>1928</td>
<td>35</td>
<td>0.27</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>0.36</td>
<td>-</td>
<td>0.82</td>
<td>-</td>
<td>-</td>
<td>0.42</td>
</tr>
<tr>
<td>1928</td>
<td>36</td>
<td>0.19</td>
<td>-</td>
<td>0.47</td>
<td>0.54</td>
<td>0.22</td>
<td>0.39</td>
<td>0.82</td>
<td>-</td>
<td>-</td>
<td>0.42</td>
</tr>
</tbody>
</table>
... INTO MEANINGFUL INSIGHT

![Heatmap of Polio cases in the United States](image)

- **Alabama**
- **Alaska**
- **Arizona**
- **Arkansas**
- **California**
- **Colorado**
- **Connecticut**
- **Delaware**
- **District of Columbia**
- **Florida**
- **Georgia**
- **Hawaii**
- **Idaho**
- **Illinois**
- **Indiana**
- **Iowa**
- **Kansas**
- **Kentucky**
- **Louisiana**
- **Maine**
- **Maryland**
- **Massachusetts**
- **Michigan**
- **Minnesota**
- **Mississippi**
- **Missouri**
- **Montana**
- **Nebraska**
- **Nevada**
- **New Hampshire**
- **New Jersey**
- **New Mexico**
- **New York**
- **North Carolina**
- **North Dakota**
- **Ohio**
- **Oklahoma**
- **Oregon**
- **Pennsylvania**
- **Rhode Island**
- **South Carolina**
- **South Dakota**
- **Tennessee**
- **Texas**
- **Utah**
- **Vermont**
- **Virginia**
- **Washington**
- **West Virginia**
- **Wisconsin**
- **Wyoming**

Year:
- 1930
- 1935
- 1940
- 1945
- 1950
- 1955
- 1960
- 1965
- 1969

Vaccine introduced:
- 1955

Color scale:
- 0 (Lightest)
- 10
- 20
- 30
- 40
- 50+ (Darkest)
DATA VIZ CAN ALSO MISLEAD

PLANNED PARENTHOOD FEDERATION OF AMERICA:
ABORTIONS UP – LIFE-SAVING PROCEDURES DOWN

2,007,371 in 2010
CANCER SCREENING & PREVENTION SERVICES

289,750 in 2010
ABORTIONS

327,000 in 2013
935,573 in 2013

SOURCE: AMERICANS UNITED FOR LIFE

Dr. Randy Olson | Email: rso@randalolson.com | Twitter: @randal_olson | Web: randalolson.com
LET’S COVER SOME GUIDELINES FOR DATA VISUALIZATION
COLOR IN DATA VISUALIZATION
COLOR SHOULD ADD INFORMATION

Google tutorial searches

- Python
- JavaScript
- R

Google tutorial searches

- Python
- JavaScript
- R

Dr. Randy Olson | Email: rso@randalolson.com | Twitter: @randal_olson | Web: randalolson.com
TYPES OF COLOR SCHEMES

- Qualitative
- Sequential
- Diverging

Dr. Randy Olson | Email: rso@randalolson.com | Twitter: @randal_olson | Web: randalolson.com
CONSIDER COLOR BLINDNESS
CONSIDER COLOR BLINDNESS

ColorOracle.org
A color scheme should:
- Add information
- Encode the data well
- Accommodate color blindness
- Print well – both color and B&W
LIBRARIES WITH GOOD COLOR SCHEMES

- Seaborn (Python)
  - matplotlib’s style library
- ggplot2 (R)
- D3.js (JavaScript)
- … or head to ColorBrewer2.org
COMMON PITFALLS IN DATA VISUALIZATION
UNLABELED CHARTS

- Most common mistake
- Can make your visualization useless
UNLABELED CHARTS

- Most common mistake
- Can make your visualization useless

Dr. Randy Olson | Email: rso@randalolson.com | Twitter: @randal_olson | Web: randalolson.com
USING 3-DIMENSIONAL CHARTS

Google tutorial searches
- Python
- JavaScript
- R
- Other

Google tutorial searches
- Python
- JavaScript
- R
- Other
USING 3-DIMENSIONAL CHARTS

- Python
- JavaScript
- R

Google searches (%)

<table>
<thead>
<tr>
<th>Year</th>
<th>Python</th>
<th>JavaScript</th>
<th>R</th>
</tr>
</thead>
<tbody>
<tr>
<td>2013</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2014</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2015</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Dr. Randy Olson | Email: rso@randalolson.com | Twitter: @randal_olson | Web: randalolson.com
FRIENDS DON’T LET FRIENDS
USE 3-DIMENSIONAL CHARTS
PARTS DON’T SUM TO A MEANINGFUL WHOLE

% of users who use >2 programming languages

- Python 75%
- JavaScript 58%
- R 41%

Dr. Randy Olson | Email: rso@randalolson.com | Twitter: @randal_olson | Web: randalolson.com
PARTS DON’T SUM TO A MEANINGFUL WHOLE

% of users who use >2 programming languages

- R
- JavaScript
- Python

Note: Fake data

Dr. Randy Olson | Email: rso@randalolson.com | Twitter: @randal_olson | Web: randalolson.com
TOO MANY CATEGORIES

Google tutorial searches

Java
Python
PHP
C#
C++
C
Javascript
Ruby
Matlab
Objective-C
Swift
VBA
Perl
lua
Visual Basic

Dr. Randy Olson | Email: rso@randalolson.com | Twitter: @randal_olson | Web: randalolson.com
TOO MANY CATEGORIES

Google tutorial searches

- Java
- PHP
- Python
- Other
BAR CHARTS NOT STARTING AT ZERO

Google tutorial searches for JavaScript

Google tutorial searches for JavaScript

2013  2014  2015

Dr. Randy Olson | Email: rso@randalolson.com | Twitter: @randal_olson | Web: randalolson.com
FAILING TO NORMALIZE DATA

1000s of Python users in...

- Orlando
- NYC
- Chicago

% of Python users in...

- Orlando
- NYC
- Chicago

Note: Fake data

Dr. Randy Olson | Email: rso@randalolson.com | Twitter: @randal_olson | Web: randalolson.com
CHARTJUNK

- Extraneous labels
- Pointless images
- Excessive color
- Fancy animations
CHARTJUNK

- Extraneous labels
- Pointless images
- Excessive color
- Fancy animations
MORE EXAMPLES

- viz.wtf
- Reddit.com/r/DataIsUgly
- Makeover Monday
  - vizwiz.blogspot.com/p/makeover-monday-challenges.html
GOOD PRACTICES IN DATA VISUALIZATION
WHEN APPROPRIATE, SHOW THE DATA
WHEN APPROPRIATE, SHOW THE DATA
CLARIFY CORRELATION != CAUSATION

Number of people who drowned by falling into a pool correlates with Films Nicolas Cage appeared in

Dr. Randy Olson | Email: rso@randalolson.com | Twitter: @randal_olson | Web: randalolson.com
LABEL OBJECTS AND DATA DIRECTLY

Google tutorial searches

- Python
- JavaScript
- R

Dr. Randy Olson | Email: rso@randalolson.com | Twitter: @randal_olson | Web: randalolson.com
Google tutorial searches

- Python
- JavaScript
- R

Dr. Randy Olson | Email: rso@randalolson.com | Twitter: @randal_olson | Web: randalolson.com
LABEL OBJECTS AND DATA DIRECTLY

Google tutorial searches

Python

JavaScript

R

Dr. Randy Olson | Email: rso@randalolson.com | Twitter: @randal_olson | Web: randalolson.com
LABEL OBJECTS AND DATA DIRECTLY

Google tutorial searches

- Python: 10.4%
- JavaScript: 7%
- R: 2.6%

Dr. Randy Olson | Email: rso@randalolson.com | Twitter: @randal_olson | Web: randalolson.com
LABEL OBJECTS AND DATA DIRECTLY

Google tutorial searches

Python 10%

JavaScript 7%

R 3%

Dr. Randy Olson | Email: rso@randalolson.com | Twitter: @randal_olson | Web: randalolson.com
MAKE SURE YOUR VISUALIZATION STANDS BY ITSELF

- Could someone interpret the chart by itself?
- Make sure chart alone can tell your story

Dr. Randy Olson | Email: rso@randalolson.com | Twitter: @randal_olson | Web: randalolson.com
INCLUDE THE UNDERLYING DATA

- When possible
  - Public data is better
- Store in a public repository
  - GitHub.com
  - figshare.com
- Link to the data repository in your visualization
DESCRIBE YOUR METHODS IN DETAIL

- From raw data to the chart
  - What data?
  - Manipulations?
  - Omissions?
- Or just provide the code!
  - GitHub.com
- Link to the methods in your visualization

Use a genetic algorithm to optimize the order to visit the waypoints

Instead of exhaustively looking at every possible solution, genetic algorithms start with a handful of random solutions and continually tinkers with these solutions — always trying something slightly different from the current solutions and keeping the best ones — until they can’t find a better solution any more.

Below, all you need to do is make sure that the file name above matches the file name below (both currently my waypoints--dist-dur.txt) and run the code. The code will read in your route information and use a genetic algorithm to discover an optimized driving route.

```python
import pandas as pd
import numpy as np
waypoint_distances = {}
waypoint_durations = {}
all_waypoints = set()
waypoint_data = pd.read_csv("my-waypoints-dist-dur.txt", sep=\"\t\")
for i, row in waypoint_data.iterrows():
    waypoint_distance[frozenset([row.waypoint1, row.waypoint2])] = row.distance
    waypoint_durations[frozenset([row.waypoint1, row.waypoint2])] = row.duration
all_waypoints.update([row.waypoint1, row.waypoint2])

import random
def compute_fitness(solution):
    """
    This function returns the total distance traveled on the current road trip.
    The genetic algorithm will favor road trips that have shorter total
distances traveled.
    """
    solution_fitness = 0.0
    for index in range(len(solution)):
```

Dr. Randy Olson | Email: rso@randalolson.com | Twitter: @randal_olson | Web: randalolson.com
“IF YOU TORTURE THE DATA LONG ENOUGH, IT WILL CONFESSION TO ANYTHING.”

*How to Lie With Statistics* (1954)
IF IT CAN’T BE REPRODUCED,
THEN IT SHOULDN’T BE TRUSTED
THESE ARE GUIDELINES, NOT RULES
THESE ARE GUIDELINES, NOT RULES

… BUT IT’S HELPFUL TO KNOW “RULES” BEFORE YOU LEARN TO BREAK THEM
Data Visualization Basics with Python
Tips, Techniques, and Best Practices for Effective Chart Visualizations with matplotlib

By Randy Olson
Publisher: O'Reilly Media
Final Release Date: January 2016
Run time: 1 hour 0 minutes

4.8
Read 8 Reviews | Write a Review

In this Data Visualization Basics with Python training course, expert author Randy Olson will teach you how to create effective data visualizations in Python. This course is designed for users that already have some experience with programming in Python.

You will start by learning about the basics of data visualization, including types of charts, common pitfalls and good practices in data visualization, and data sources. Finally, Randy will teach you about matplotlib, including how to use matplotlib in the Jupyter Notebook, matplotlib styles, and subplots and small multiples.

Once you have completed this computer based training course, you will have learned a number of tips, tricks, and best practices for creating effective data visualizations in Python. Working files are included, allowing you to follow along with the author throughout the lessons.

http://shop.oreilly.com/product/0636920046592.do
THANK YOU!