What is data?

GEOG 30323

August 25, 2015

Introductions

My data journey

Course goals

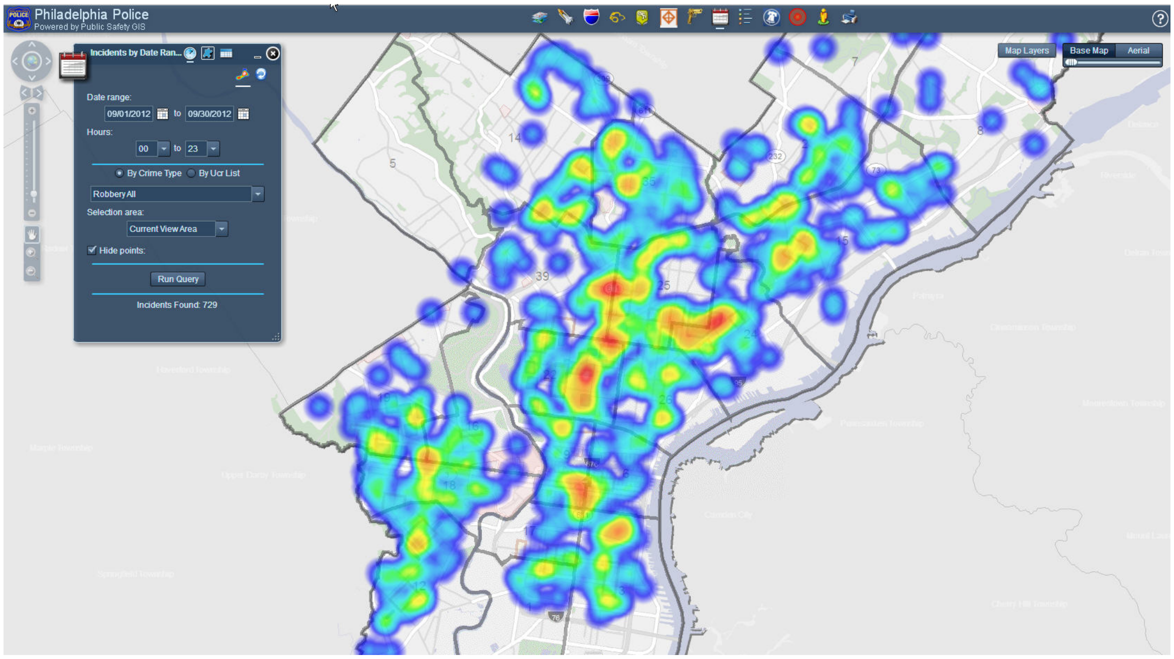
Syllabus

What is data?

Brainstorm: how might you use data if you were…

- …a librarian?
- …a police officer?
- …a small shoe store owner?
- …a general manager of a professional basketball franchise?

Policing



Date range: 09/01/2012 to 09/30/2012

Hours: 00 to 23

● By Crime Type ○ By Ucr List

Robbery All

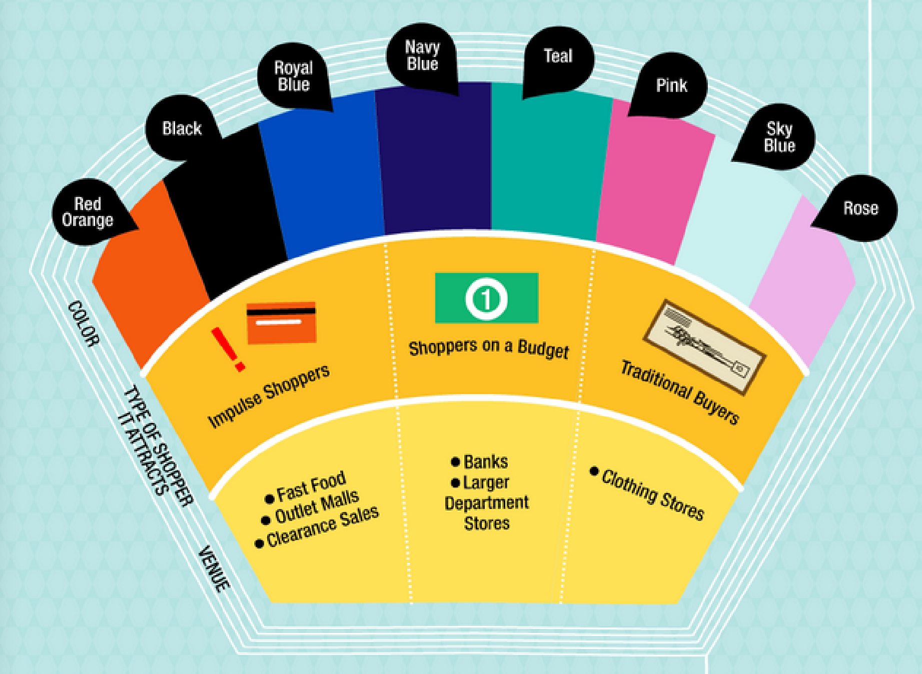
Selection area: Current View Area

☑ Hide points:

Incidents Found: 729

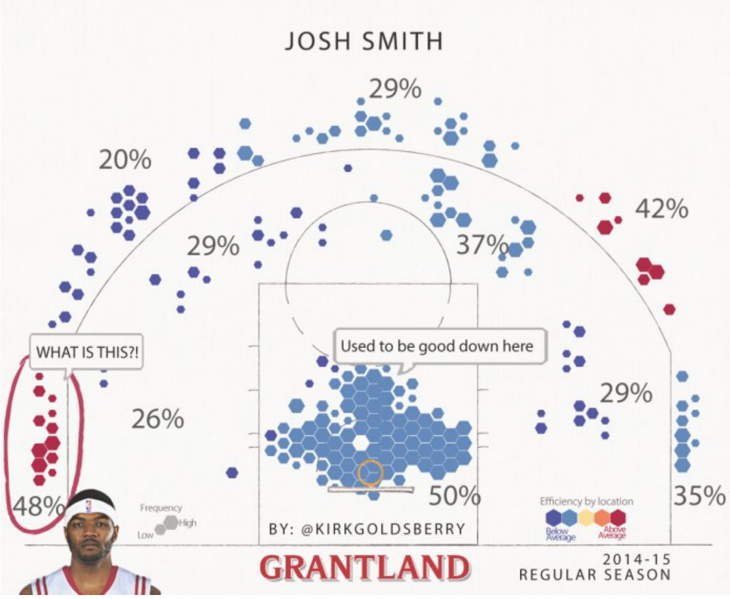
Source: technical.ly (https://technical.ly/philly/2012/11/20/philadelphia-police-gis-new-system-crime-tracking/)

Retail



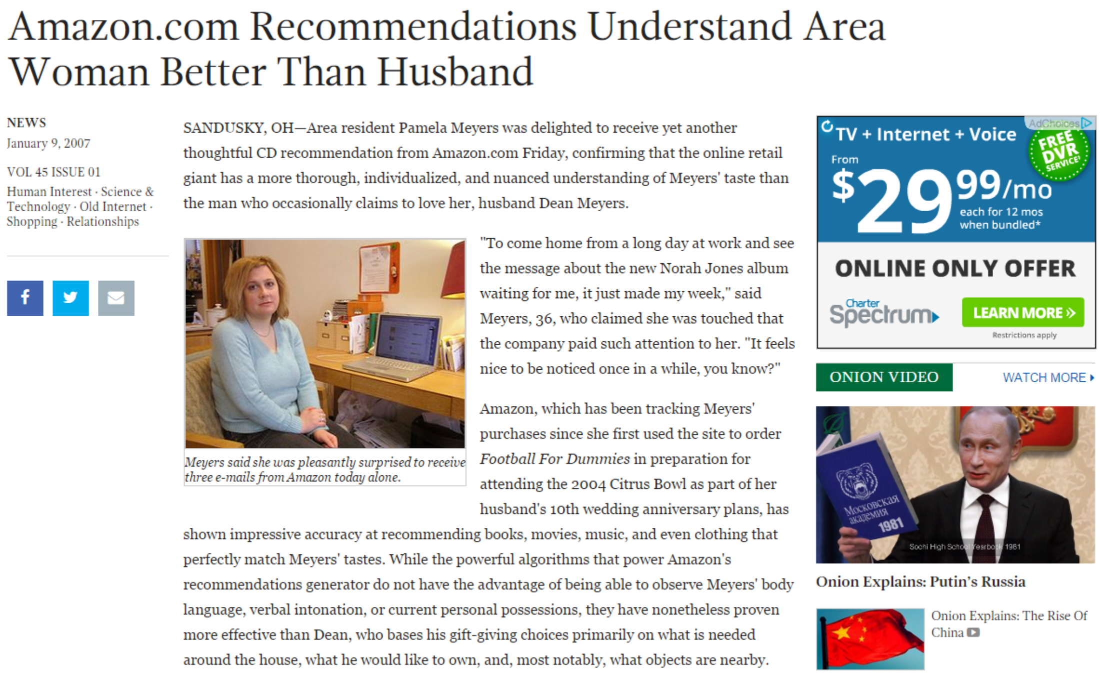
Source: Humayun Khan/Shopify (http://www.shopify.com/blog/14254569-why-all-sale-signs-are-red-the-science-of-color-in-retail)

Sports



Source: Kirk Goldsberry/*Grantland* (http://grantland.com/the-triangle/nba-least-efficient-shooters-josh-smith-michael-carter-williams-trey-burke-nerlens-noel-dion-waiters/)

Data science



Amazon.com Recommendations Understand Area Woman Better Than Husband

NEWS
January 9, 2007

VOL 45 ISSUE 01

Human Interest · Science & Technology · Old Internet · Shopping · Relationships

SANDUSKY, OH—Area resident Pamela Meyers was delighted to receive yet another thoughtful CD recommendation from Amazon.com Friday, confirming that the online retail giant has a more thorough, individualized, and nuanced understanding of Meyers' taste than the man who occasionally claims to love her, husband Dean Meyers.

Meyers said she was pleasantly surprised to receive three e-mails from Amazon today alone.

"To come home from a long day at work and see the message about the new Norah Jones album waiting for me, it just made my week," said Meyers, 36, who claimed she was touched that the company paid such attention to her. "It feels nice to be noticed once in a while, you know?"

Amazon, which has been tracking Meyers' purchases since she first used the site to order *Football For Dummies* in preparation for attending the 2004 Citrus Bowl as part of her husband's 10th wedding anniversary plans, has shown impressive accuracy at recommending books, movies, music, and even clothing that perfectly match Meyers' tastes. While the powerful algorithms that power Amazon's recommendations generator do not have the advantage of being able to observe Meyers' body language, verbal intonation, or current personal possessions, they have nonetheless proven more effective than Dean, who bases his gift-giving choices primarily on what is needed around the house, what he would like to own, and, most notably, what objects are nearby.

ONION VIDEO WATCH MORE ▸

Sochi High School Yearbook 1981

Onion Explains: Putin's Russia

Onion Explains: The Rise Of China

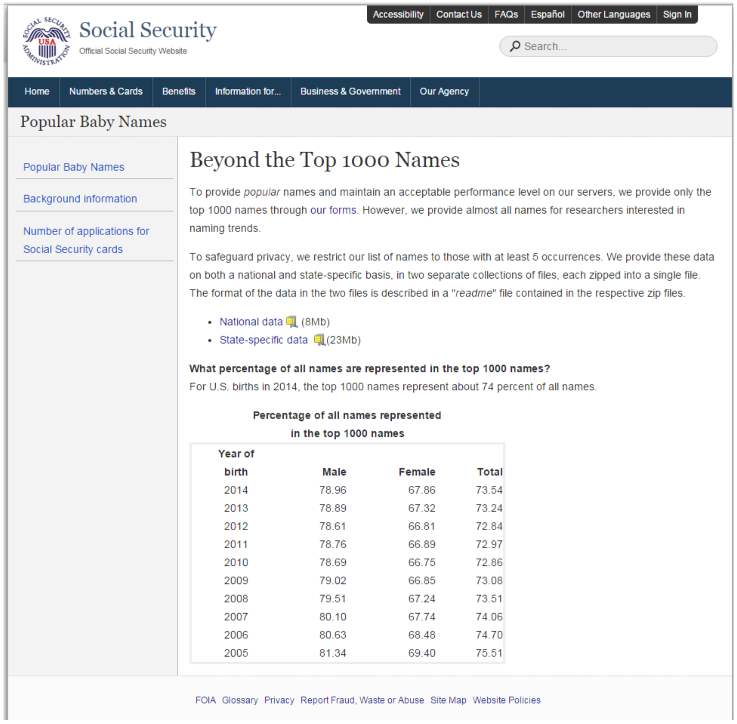
Source: *The Onion* (http://www.theonion.com/article/amazoncom-recommendations-understand-area-woman-be-2121)

What you will learn in this course

This course covers the process of **exploratory data analysis**, which includes how to:

- Find data
- Load data
- Clean data
- Summarize data
- Visualize data
- Present data
- Example question: what are the most popular female baby names in Texas, and how has this changed over time?

Finding data



Beyond the Top 1000 Names

To provide *popular* names and maintain an acceptable performance level on our servers, we provide only the top 1000 names through our forms. However, we provide almost all names for researchers interested in naming trends.

To safeguard privacy, we restrict our list of names to those with at least 5 occurrences. We provide these data on both a national and state-specific basis, in two separate collections of files, each zipped into a single file. The format of the data in the two files is described in a "*readme*" file contained in the respective zip files.

- National data (8Mb)
- State-specific data (23Mb)

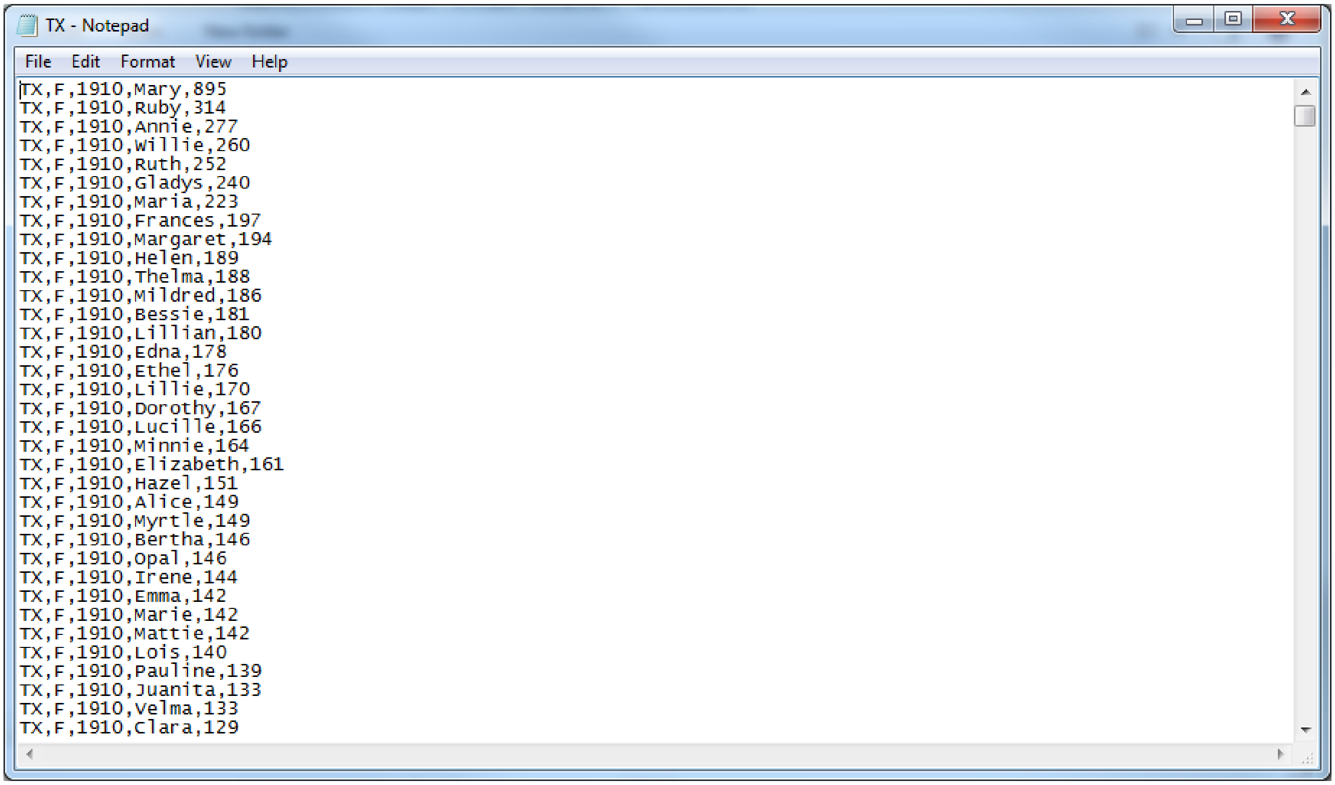
What percentage of all names are represented in the top 1000 names?
For U.S. births in 2014, the top 1000 names represent about 74 percent of all names.

Percentage of all names represented in the top 1000 names

Year of birth	Male	Female	Total
2014	78.96	67.86	73.54
2013	78.89	67.32	73.24
2012	78.61	66.81	72.84
2011	78.76	66.89	72.97
2010	78.69	66.75	72.86
2009	79.02	66.85	73.08
2008	79.51	67.24	73.51
2007	80.10	67.74	74.06
2006	80.63	68.48	74.70
2005	81.34	69.40	75.51

Source: Social Security Administration (http://www.ssa.gov/oact/babynames/limits.html)

Loading and cleaning data

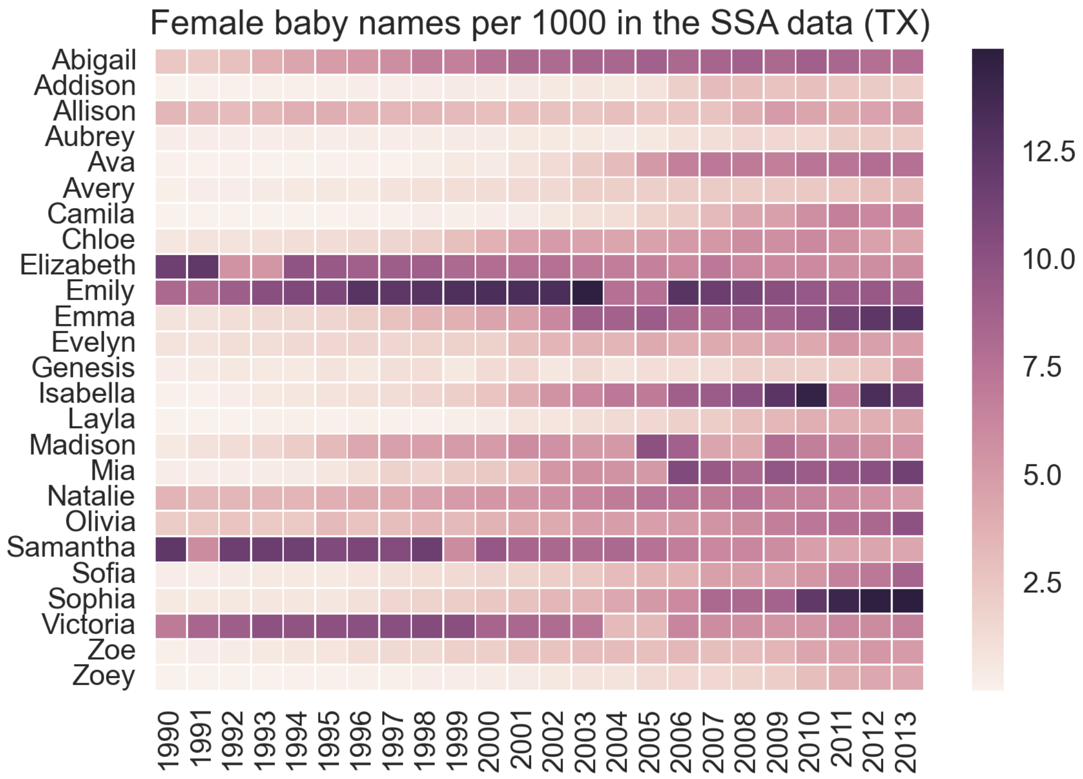


```
TX,F,1910,Mary,895
TX,F,1910,Ruby,314
TX,F,1910,Annie,277
TX,F,1910,Willie,260
TX,F,1910,Ruth,252
TX,F,1910,Gladys,240
TX,F,1910,Maria,223
TX,F,1910,Frances,197
TX,F,1910,Margaret,194
TX,F,1910,Helen,189
TX,F,1910,Thelma,188
TX,F,1910,Mildred,186
TX,F,1910,Bessie,181
TX,F,1910,Lillian,180
TX,F,1910,Edna,178
TX,F,1910,Ethel,176
TX,F,1910,Lillie,170
TX,F,1910,Dorothy,167
TX,F,1910,Lucille,166
TX,F,1910,Minnie,164
TX,F,1910,Elizabeth,161
TX,F,1910,Hazel,151
TX,F,1910,Alice,149
TX,F,1910,Myrtle,149
TX,F,1910,Bertha,146
TX,F,1910,Opal,146
TX,F,1910,Irene,144
TX,F,1910,Emma,142
TX,F,1910,Marie,142
TX,F,1910,Mattie,142
TX,F,1910,Lois,140
TX,F,1910,Pauline,139
TX,F,1910,Juanita,133
TX,F,1910,Velma,133
TX,F,1910,Clara,129
```

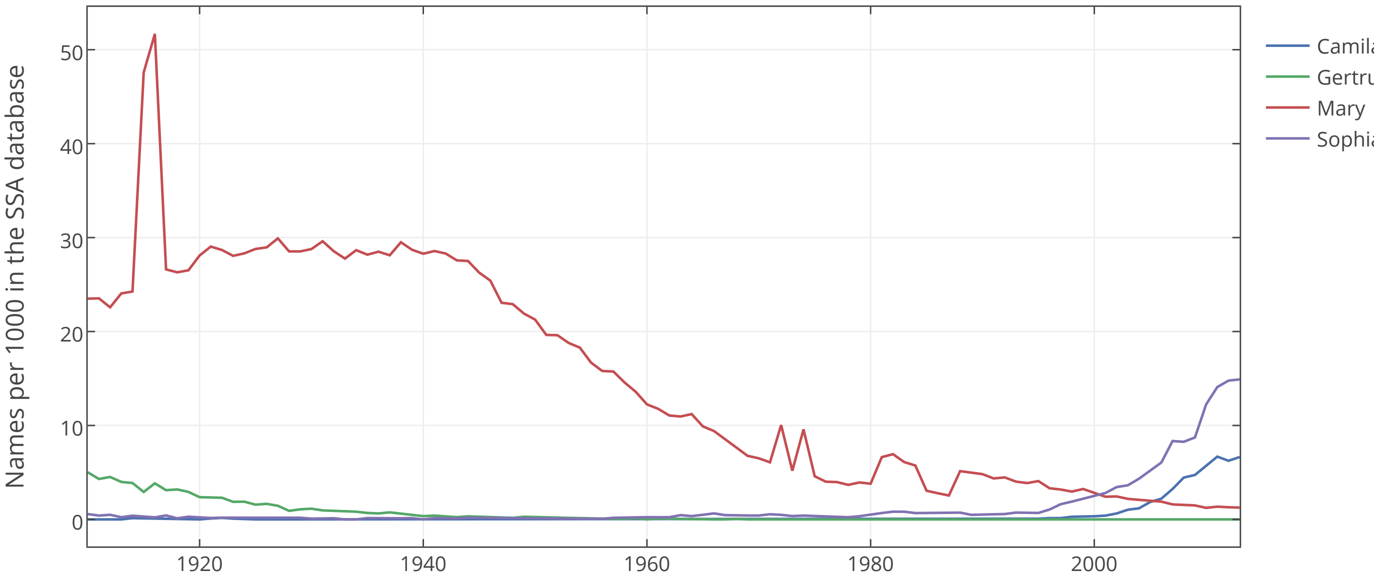
Summarizing data

	state	gender	year	name	count	per1000	rank
177374	TX	F	2013	Sophia	2381	14.905098	1
177375	TX	F	2013	Emma	2043	12.789213	2
177376	TX	F	2013	Isabella	1941	12.150691	3
177377	TX	F	2013	Mia	1830	11.455829	4
177378	TX	F	2013	Olivia	1598	10.003506	5
177379	TX	F	2013	Emily	1449	9.070763	6
177380	TX	F	2013	Sofia	1380	8.638822	7
177381	TX	F	2013	Abigail	1250	7.825020	8
177382	TX	F	2013	Ava	1227	7.681040	9
177383	TX	F	2013	Victoria	1067	6.679437	10

Visualizing data



Presenting data



Edit ch

Tools of the trade

- Your main tool: the Python programming language
- Guest appearances from: CartoDB, Tableau

- More to come on Python next week!

Assignment 1: Data journalism

- **Data journalism**: movement in journalism toward interactive, data-driven content
- Popular publications: The Upshot (*New York Times*) (http://www.nytimes.com/upshot/?_r=0), FiveThirtyEight (http://fivethirtyeight.com/), *The Guardian* (http://www.theguardian.com/data), ProPublica (https://www.propublica.org/), many more

Your task:

- Find an example of data journalism from the past two years in a recognized publication
- Write a response that 1) summarizes the article, 2) discusses its data sources, and 3) discusses its methods of analysis and visualization.
- At least one paragraph, no longer than one page