



HKUST
VISLAB

COMP 4462

Data Visualization Tutorial

Leo Yu Ho, Lo
Wenchao Li

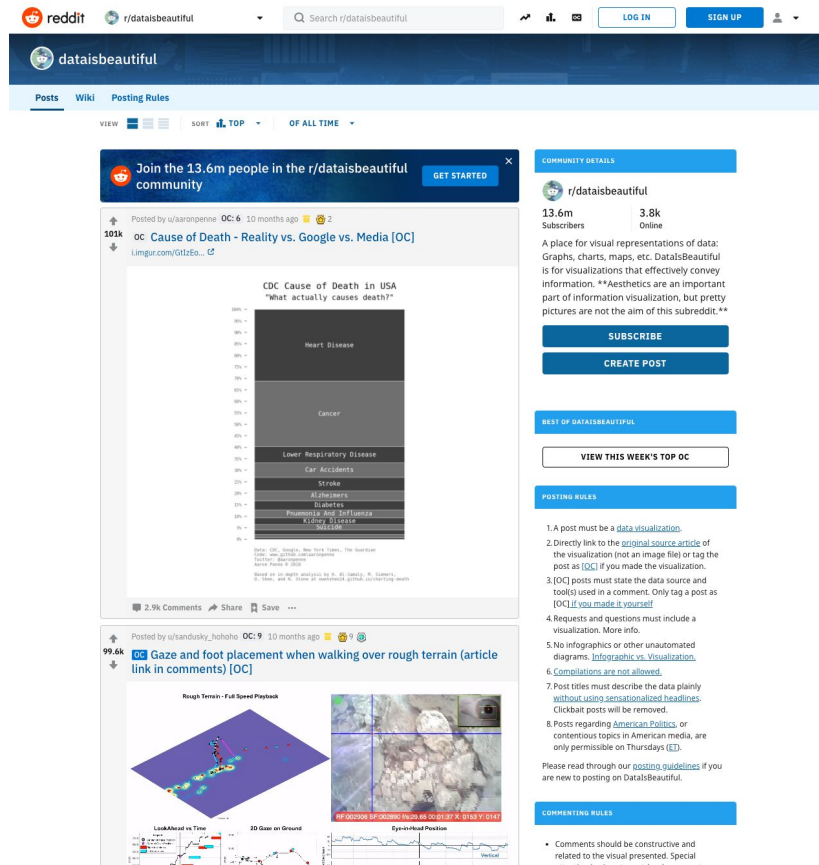
Friday 12 March, 2020
<https://bit.ly/vis-t03>

Course Project & Top-Vis Competition

- Project (35%)
 - Grouping: [Sign up sheet](#) (19 Mar)
 - Phase 1: Proposal presentation (16 Apr)
 - Find a dataset
 - What kind of data? How large is it? Why this dataset?
 - Visualization tasks / data processing / visual encoding
 - Good to have a mock up
 - Phase 2: Project presentation (TBD)
 - Make it real! Coding & demo
 - Share stories in the data
- Top-Vis Competition (10%)
 - ~1.5 mins to present 1 visualization (24 Mar)
 - Video recorded, see instructions: <http://www.yangleni.com/tutorial-record-video.html>
 - Write up a short essay
 - Why you chose this visualization?
 - What data are visualized? How are they encoded? What insights did you find?

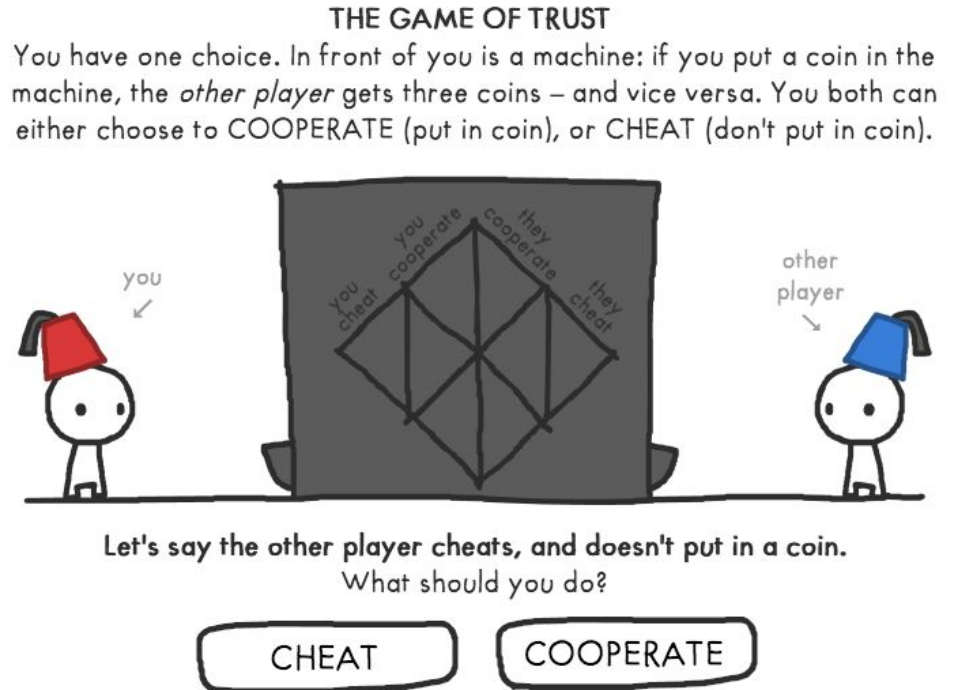
Data is Beautiful

- New visualizations everyday
- Top post of all time
 - Visualization with highest voting of all time
- A lot of remarkable ideas
- Mainstream:
 - Meaning of data > visual effect
 - And some are visually impressive
- Another subreddit: Data is Ugly
 - Lying with charts
 - Deceiving, scam
 - Some are from very authoritative sources
 - Famous news websites
 - Governments
 - Famous companies



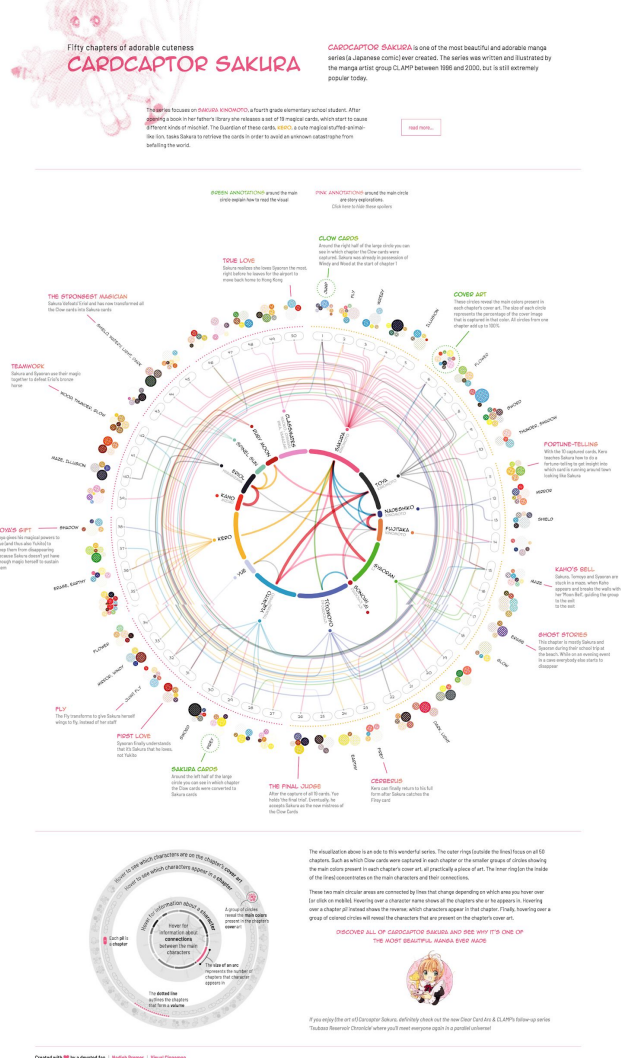
Nick Case

- Narrative visualizations
 - Telling a story with visualizations
- Evolution of Trust
 - Game theory about our society
 - Prisoner dilemma
 - CHEAT?
 - COOPERATE?
 - Interactive
 - Nice graphics and music
 - A sandbox simulator at the end
 - Enjoy!
- More on [Nick Case's webpage](#)



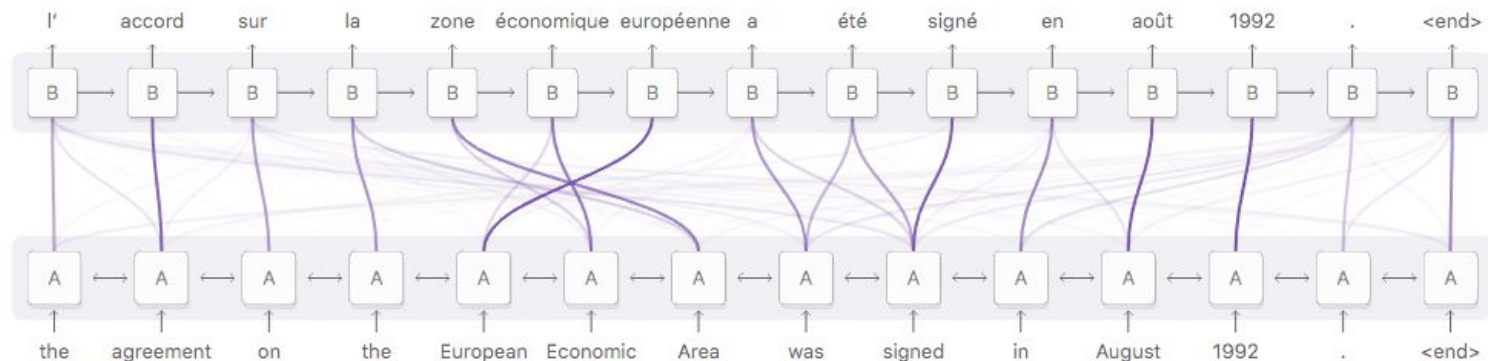
Data Sketches

- Beautiful! Eye pleasing! Fun datasets!
- And they have 24 of them!
- By:
 - [Nadieh Bremer](#)
 - [Susie Lu](#)
- [Cardcaptor Sakura](#)
 - Visualizing 50 chapters of the manga
 - Appeared characters
 - Magic spells
 - Annotations
- Another one on [Dragon Ball Z](#)
- With [explanations](#)!
 - They have journaled the process in details!



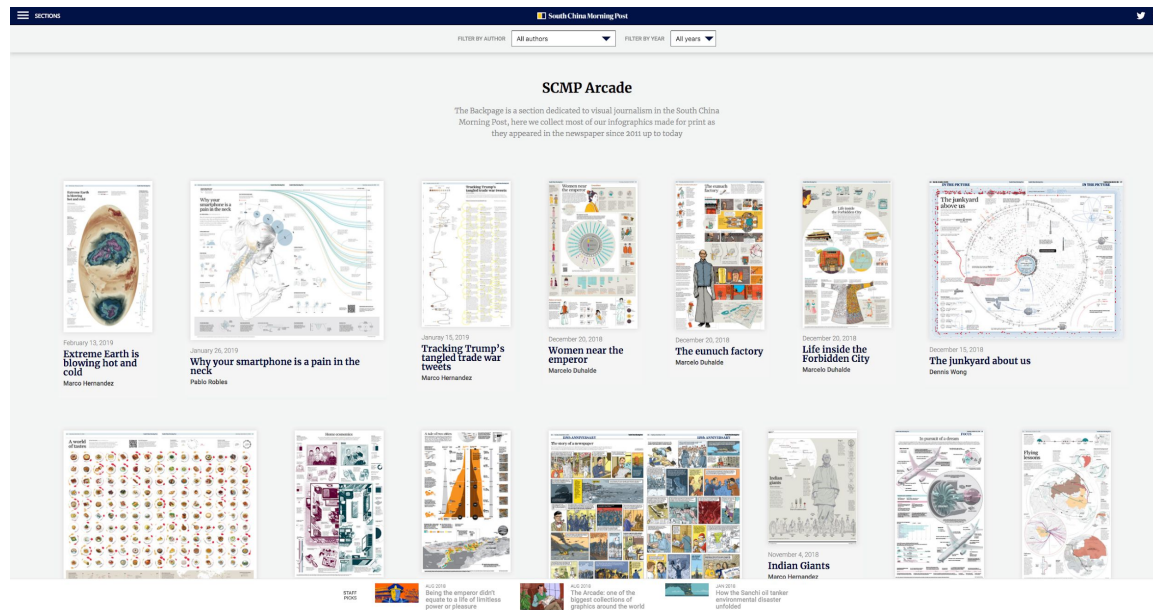
Distill

- Visual Explanation of Machine Learning Algorithms
- Attention and Augmented Recurrent Neural Networks
 - Visualizing a neural translation model
 - Which word in a French sentence \Leftrightarrow which word in English?



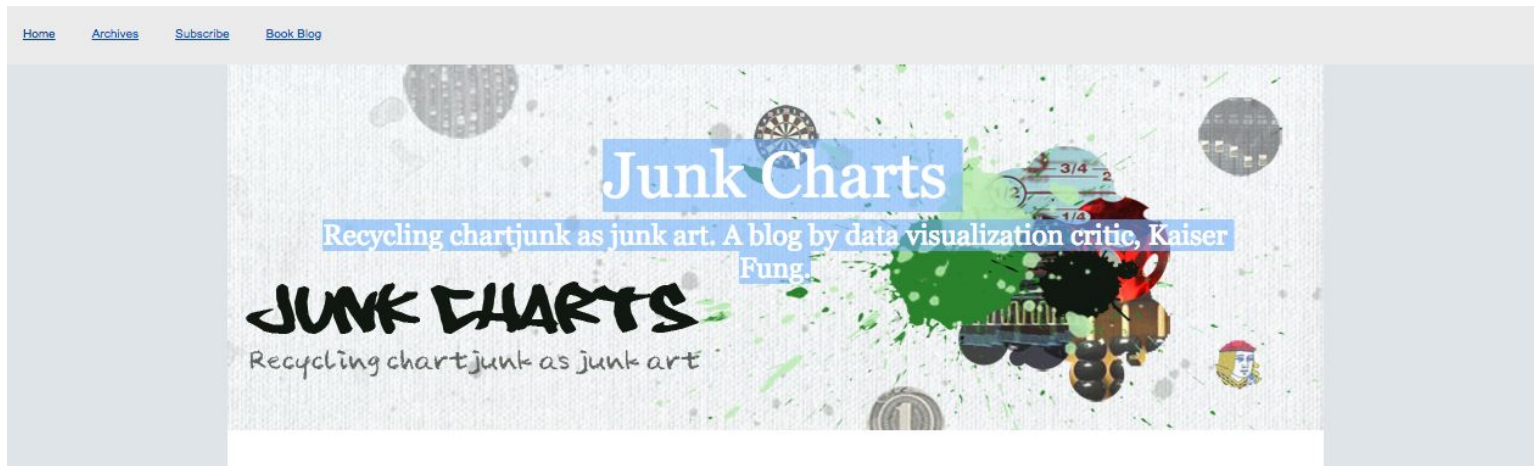
The list of 2019 visualization lists

- 32 lists, each has 10+ visualizations!
- 2019 Review: the year Hong Kong dominated the headlines
- SCMP Print Arcade
 - 262 visualizations from 2011 to 2020



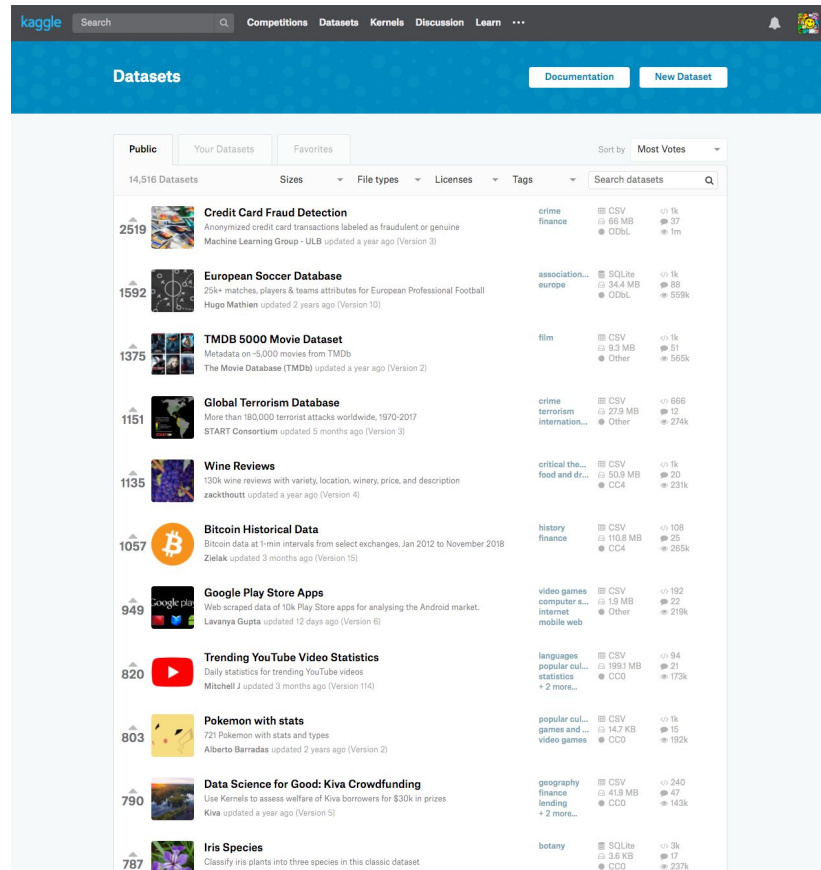
Junk Charts

- A collection of bad visualizations
 - How to lie with visualizations
 - Like [Data is Ugly](#) subreddit
 - With explanations
 - Update frequently



Kaggle Datasets

- No.1 source of datasets
- A lot of datasets
- Data are clean (relatively)
- A lot of kernels (jupyter notebooks)
 - See what the others do with the datasets
- Can seek help very easily
 - Can also raise questions to the authors



The screenshot shows the Kaggle Datasets page. At the top, there's a navigation bar with 'kaggle' logo, a search bar, and links for 'Competitions', 'Datasets', 'Kernels', 'Discussion', and 'Learn'. Below this, a blue header contains the word 'Datasets' and buttons for 'Documentation' and 'New Dataset'. The main content area displays a list of public datasets. The list is sorted by 'Most Votes' and shows 14,516 datasets. Each dataset entry includes a rank, a thumbnail icon, the dataset name, a brief description, the file type, size, and the number of votes. The datasets listed are:

Rank	Dataset Name	Description	File Type	Size	Votes
2519	Credit Card Fraud Detection	Anonymized credit card transactions labeled as fraudulent or genuine Machine Learning Group - ULB updated a year ago (Version 3)	CSV	68 MB	1k
1592	European Soccer Database	25k+ matches, players & teams attributes for European Professional Football Hugo Mathien updated 2 years ago (Version 10)	SQLite	34.4 MB	37
1375	TMDB 5000 Movie Dataset	Metadata on 5,000 movies from TMDB The Movie Database (TMDB) updated a year ago (Version 2)	CSV	9.3 MB	86
1151	Global Terrorism Database	More than 180,000 terrorist attacks worldwide, 1970-2017 START Consortium updated 5 months ago (Version 3)	CSV	27.8 MB	656
1135	Wine Reviews	130k wine reviews with variety, location, winery, price, and description zackthoutt updated a year ago (Version 4)	CSV	50.8 MB	20
1057	Bitcoin Historical Data	Bitcoin data at 1-min intervals from select exchanges, Jan 2012 to November 2018 Zielak updated 3 months ago (Version 15)	CSV	110.8 MB	25
949	Google Play Store Apps	Web scraped data of 10k Play Store apps for analysing the Android market. Lavanya Gupta updated 12 days ago (Version 6)	CSV	1.9 MB	22
820	Trending YouTube Video Statistics	Daily statistics for trending YouTube videos Mitchell J updated 3 months ago (Version 114)	CSV	199.5 MB	21
803	Pokemon with stats	721 Pokemon with stats and types Alberto Barradas updated 2 years ago (Version 2)	CSV	14.7 KB	15
790	Data Science for Good: Kiwa Crowdfunding	Use Kernels to assess welfare of Kiwa borrowers for \$30k in prizes Kiwa updated a year ago (Version 5)	CSV	41.9 KB	47
787	Iris Species	Classify Iris plants into three species in this classic dataset	SQLite	3.6 KB	17

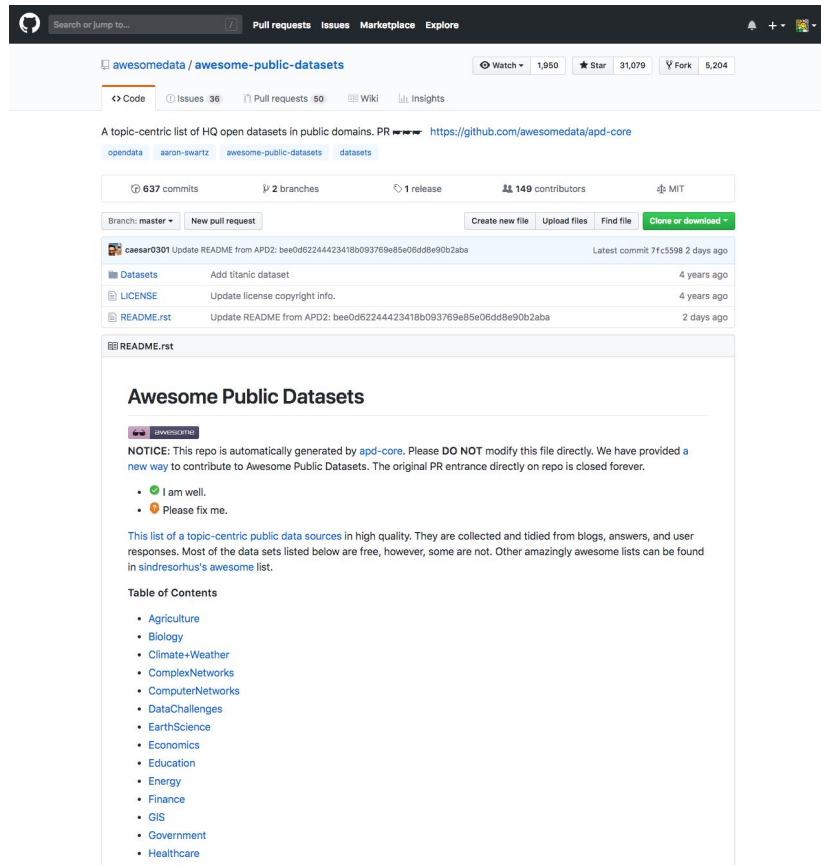
Dataviz Battle on r/dataisbeautiful

- Monthly competition on r/dataisbeautiful
- A lot of submissions for references
- September 2018: Visualize information on all 802 Pokemon
 - Winners are announced in the Dataviz Battle thread of next month
 - For example, October 2018 announced the winners of visualizing Pokemon

The screenshot shows the Reddit interface for the subreddit r/dataisbeautiful. The search bar at the top contains the text "dataviz battle for the month of". Below the search bar, the results are sorted by "NEW" and show a list of posts. The first post is titled "[Battle] DataViz Battle for the month of February 2019: Visualize Physical Harm and Dependence by Drug" and has 21 comments. The second post is titled "[Battle] DataViz Battle for the month of January 2019: Visualize the list of World's Oldest People" and has 75 comments. The third post is titled "[Battle] DataViz Battle for the month of December 2018: Visualize the Freezing and Thawing cycle of Lake Mendota" and has 112 comments. The fourth post is titled "[Battle] DataViz Battle for the month of November 2018: Visualize the List of NASA Astronauts" and has 70 comments. The fifth post is titled "[Battle] DataViz Battle for the month of October 2018: Visualize 859 survey results from r/travel" and has 55 comments. The sixth post is titled "[Battle] DataViz Battle for the month of September 2018: Visualize information on all 802 Pokemon" and has 102 comments. The seventh post is titled "[Battle] DataViz Battle for the month of August 2018: Visualize TSA Claims" and has 94 comments. The eighth post is titled "[Battle] DataViz Battle for the month of July 2018: Make it better: Which Birds prefer Which Seeds" and has 83 comments. The ninth post is titled "[Battle] DataViz Battle for the month of June 2018: Visualize The lives, reigns, and deaths of 68 Roman emperors from 26 BC to 395 AD" and has 99 comments. The tenth post is titled "[Battle] DataViz Battle for the month of May 2018: Visualize 1.6 Million Accidents in England, Scotland, and Wales from 2000-2016" and has 28 comments. The eleventh post is titled "[Lounge] This week is a Bye Week for the DatViz Battles. Use this thread for off-topic discussion, smack talk, and cool suggestions!" and has 12 comments. The twelfth post is titled "[Battle] DataViz Battle for the month of April 2018: Visualize every line from every scene in The Office" and has 81 comments. The thirteenth post is titled "[Battle] DataViz Battle for the month of March 2018: Visualize Over 100,000 Stars" and has 81 comments. On the right side of the page, there is a "COMMUNITY DETAILS" section for r/dataisbeautiful, showing 13.6m subscribers and 3.4k online users. Below this is a "SUBSCRIBE" button and a "CREATE POST" button. At the bottom right, there is a "COMMUNITY OPTIONS" section with links for "About", "Careers", "Press", "Advertise", "Blog", "Help", "The Reddit App", "Reddit Coins", "Reddit Premium", and "Reddit Gifts". At the very bottom right, there is a "Content Policy" and "Privacy Policy" link, and a copyright notice for 2019 Reddit, Inc.

awesome-public-datasets

- A very thorough list
- With active update
- Search Engine subsection
 - Websites that have “search for datasets”
- Data Challenge subsection
 - More Kaggle like websites
- Complementary Collection subsection
 - More dataset lists



Tasks

- Get the whole list of [“Where to find visualizations and datasets” on GitHub](#)
- If you don't have a group yet:
 - a. Find a visualization that interests you
 - b. Post it on [this Trello board](#) with a “how to reach you” message
 - i. [View only](#)
 - ii. [With edit access \(sign up required\)](#)
 - c. Find someone who post visualizations you like
 - d. Form a group if interest matches
 - i. Signup on [Google Docs](#)
- If you already have a group:
 - a. Find some visualizations
 - b. Talk to your group mates
 - c. Find a dataset to work on
 - d. Talk about what interesting insight can be found in the dataset
- Make amazing visualizations!

Next tutorial

Python, Jupyter and
Pandas

- Prepare your Google account beforehand
 - For using [Google Colab](#)
 - Jupyter notebook environment
 - Free!
 - No setup
- Alternatively, you can use jupyter notebook on your computer, but that is cumbersome