Data preparation

“Give me six hours to chop down a tree and I will spend the first four sharpening the axe.”

— Abraham Lincoln (?)

Know Your Questions

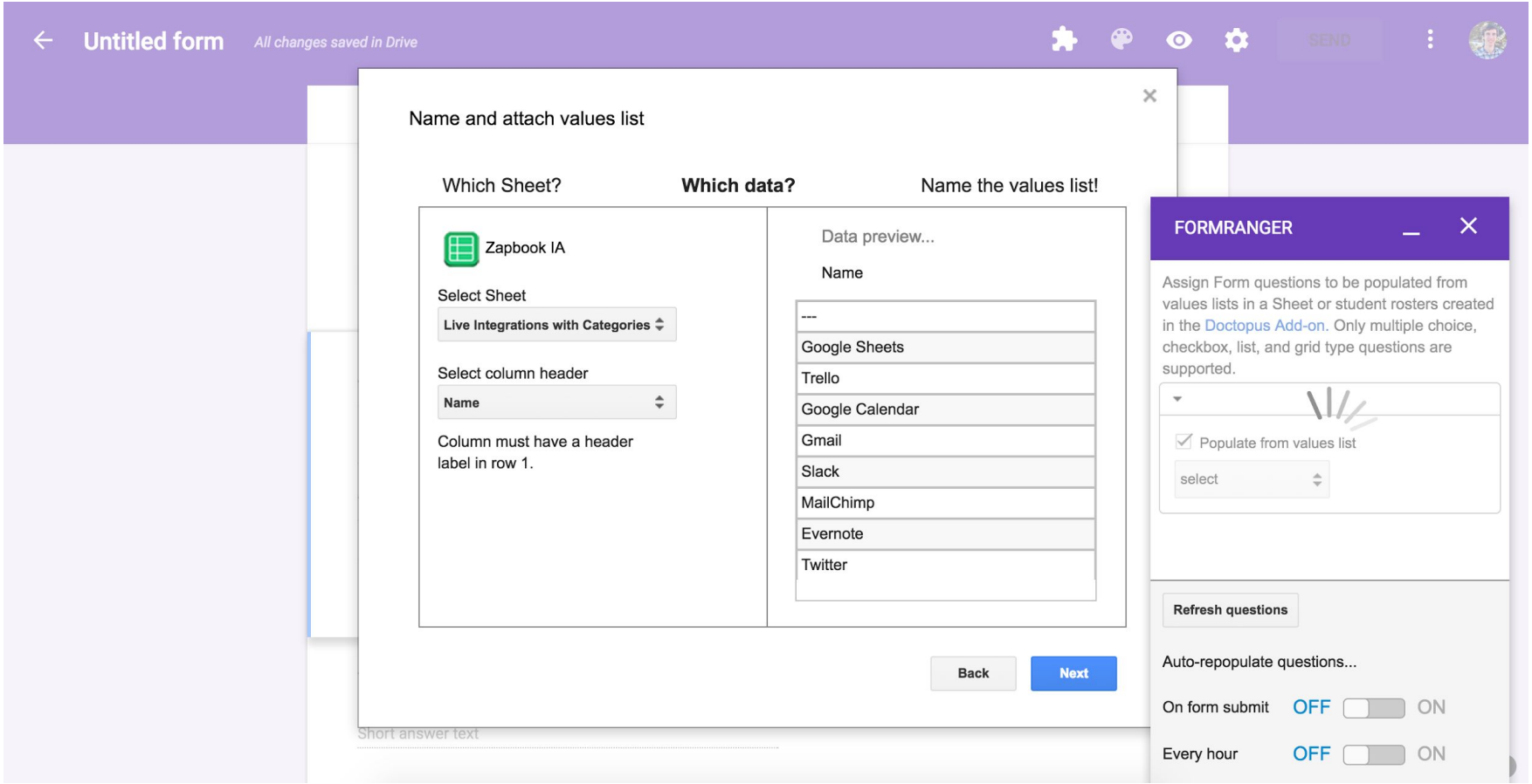
Gathering your data

https://vimeo.com/27800118

Manual

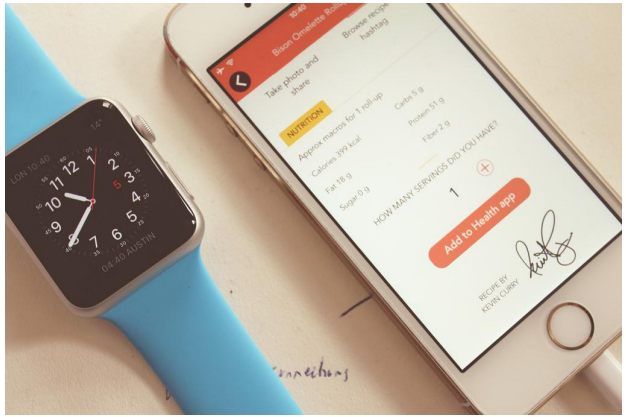
https://vimeo.com/27800118

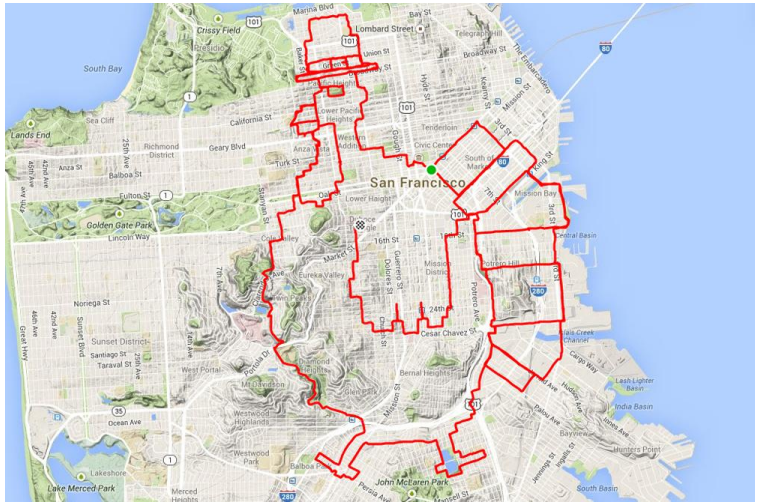
Self Generated or User Generated



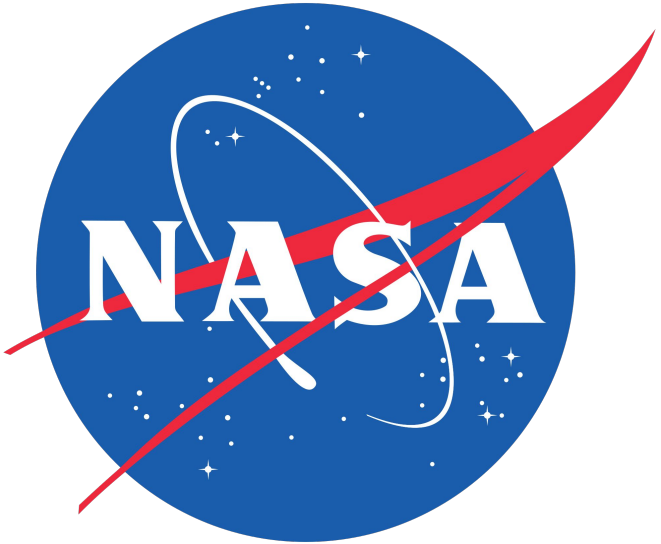
Self Generated or User Generated

Fit Bit, Nike+, iPhone Health App, Strava, Biometrics etc.





Existing Data Sources



Tidy data / Don’t mix different type of data in the same column

Columns: attribute (or variable)

Summit Name	Country	Lat.	Elevation (ft.)	Prominence (ft).	First ascent
Mt. Everest	Nepal/Tibet	27º59' N	29028	29028	1953
Nanga Parbat	Pakistan	35º14' N	26657	15118	1953
Namcha Barwa	China	29º38' N	25531	13471	1992
Jengish Chokusu	Kyrgyzstan/China	42º02' N	24406	13609	1956
Aconcagua	Argentina	32º39' S	22841	22841	1897
Chimborazo	Ecuador	1º28' S	20561	13523	1880
Mt. McKinley (Denali)	US	63º06' N	20320	20138	1913
Mt. Logan	Canada	60º34' N	19550	17224	1925
Cristobal Colon	Colombia	10º50' N	18701	18074	1939
Gora Elbrus	Russia	43º21' N	18510	15554	1874
Pico de Orizaba (Citlaltepetl)	Mexico	19º02' N	18491	16148	1848
Damavand	Iran	35º57' N	18405	15311	905
Bogda Shan	China	43º48' N	17864	13523	1981
Vinson Massif	Antarctica	78º31' S	16050	16050	1966
Puncak Jaya	Indonesia	4º04' S	16023	16023	1962
Mont Blanc	France/Italy	45º50' N	15777	15406	1786
Klyuchevskaya Volcano	Russia	56º03' N	15584	15252	1788
Mauna Kea	US	19º49' N	13796	13796	1823
Kinabalu	Malaysia	6º04' N	13435	13435	1851

Rows: observations

Tidy data is easier to work with

For most charting tools (Excel, Illustrator, etc.), omit commas, symbols, leters for quantitative data:

$2,745	#	2745
+76%	#	76
65mph	#	65
212°F	#	212
March 1, 2017	#	20170301
73°41'W	#	-73.41

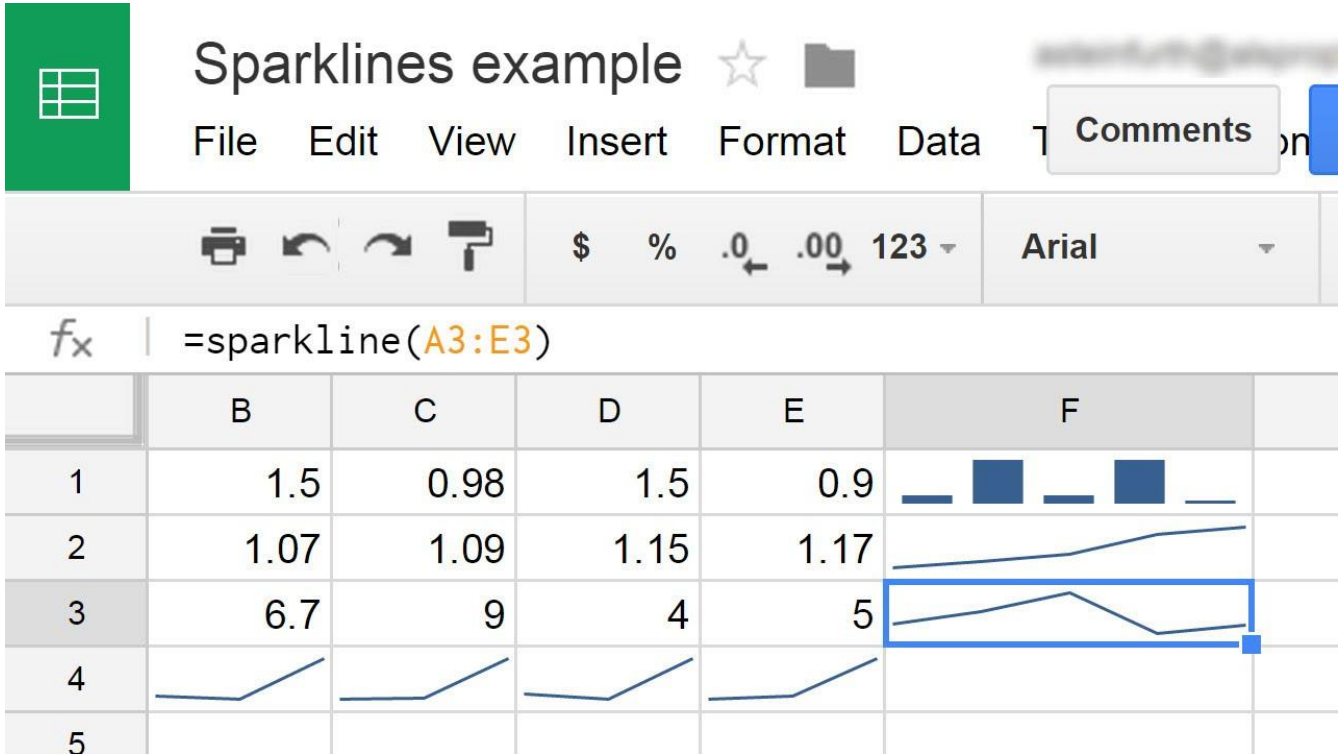
CSV files / Comma-separated values

```
Rank,Summit Name,Country,Lat.,Long.,Elevation (ft.),Elevation (m),Prominence (ft).,First ascent
1,Mt. Everest,Nepal/Tibet,27.59,86.55,29028,8848,29028,1953
2,Nanga Parbat,Pakistan,35.14,74.36,26657,8125,15118,1953
3,Namcha Barwa,China,29.38,95.03,25531,7782,13471,1992
4,Jengish Chokusu,Kyrgyzstan/China,42.02,80.07,24406,7439,13609,1956
5,Aconcagua,Argentina,-32.39,-70.01,22841,6962,22841,1897
6,Chimborazo,Ecuador,-1.28,-78.49,20561,6267,13523,1880
7,Mt. McKinley (Denali),US,63.06,-151.03,20320,6194,20138,1913
8,Mt. Logan,Canada,60.34,-140.24,19550,5959,17224,1925
9,Kilimanjaro,Tanzania,-3.04,37.21,19340,5895,19308,1889
10,Cristobal Colon,Colombia,10.50,-73.41,18701,5700,18074,1939
11,Gora Elbrus,Russia,43.21,42.26,18510,5642,15554,1874
12,Pico de Orizaba (Citlaltepetl),Mexico,19.02,-97.16,18491,5636,16148,1848
13,Damavand,Iran,35.57,52.07,18405,5610,15311,905
14,Bogda Shan,China,43.48,88.20,17864,5445,13523,1981
15,Vinson Massif,Antarctica,-78.31,-85.36,16050,4892,16050,1966
16,Puncak Jaya,Indonesia,-4.04,137.11,16023,4884,16023,1962
17,Mont Blanc,France/Italy,45.50,6.52,15777,4809,15406,1786
18,Klyuchevskaya Volcano,Russia,56.03,160.39,15584,4750,15252,1788
19,Mauna Kea,US,19.49,-155.28,13796,4205,13796,1823
20,Kinabalu,Malaysia,6.04,116.34,13435,4095,13435,1851
```

Data analysis

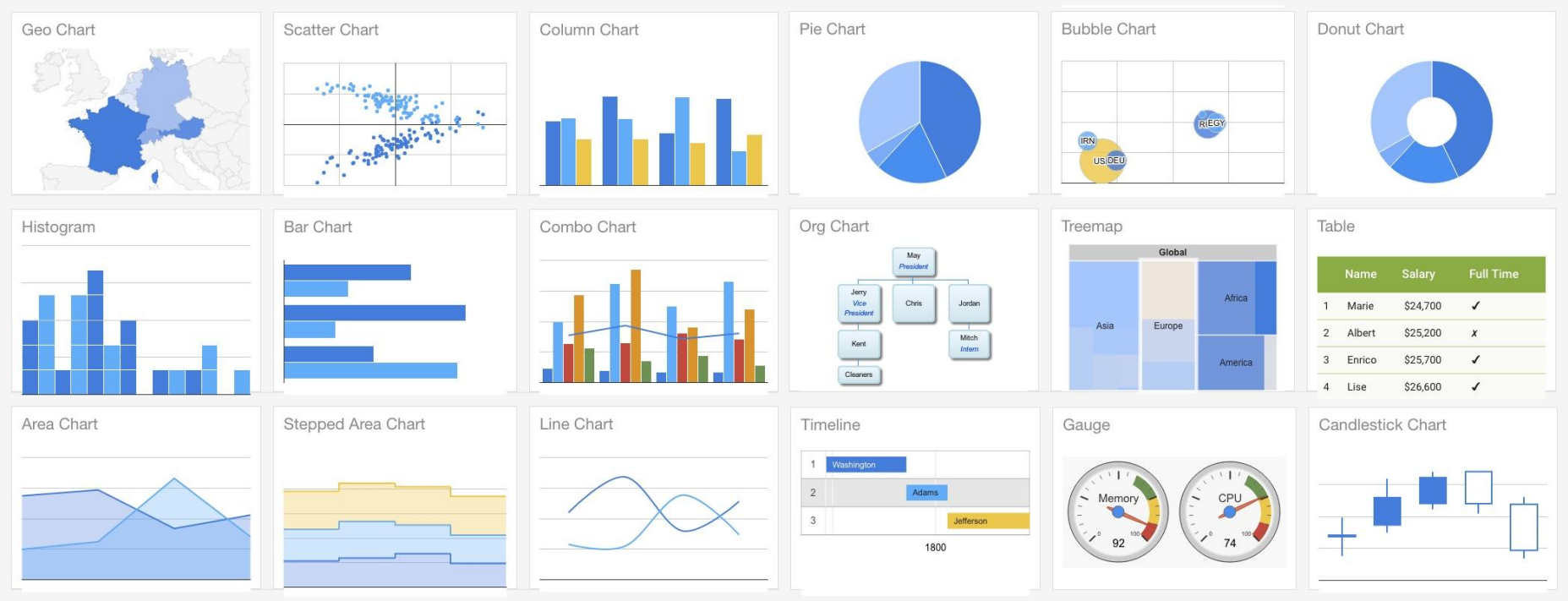
Suggested tool

Sparklines - Google Sheets (EXAMPLE)



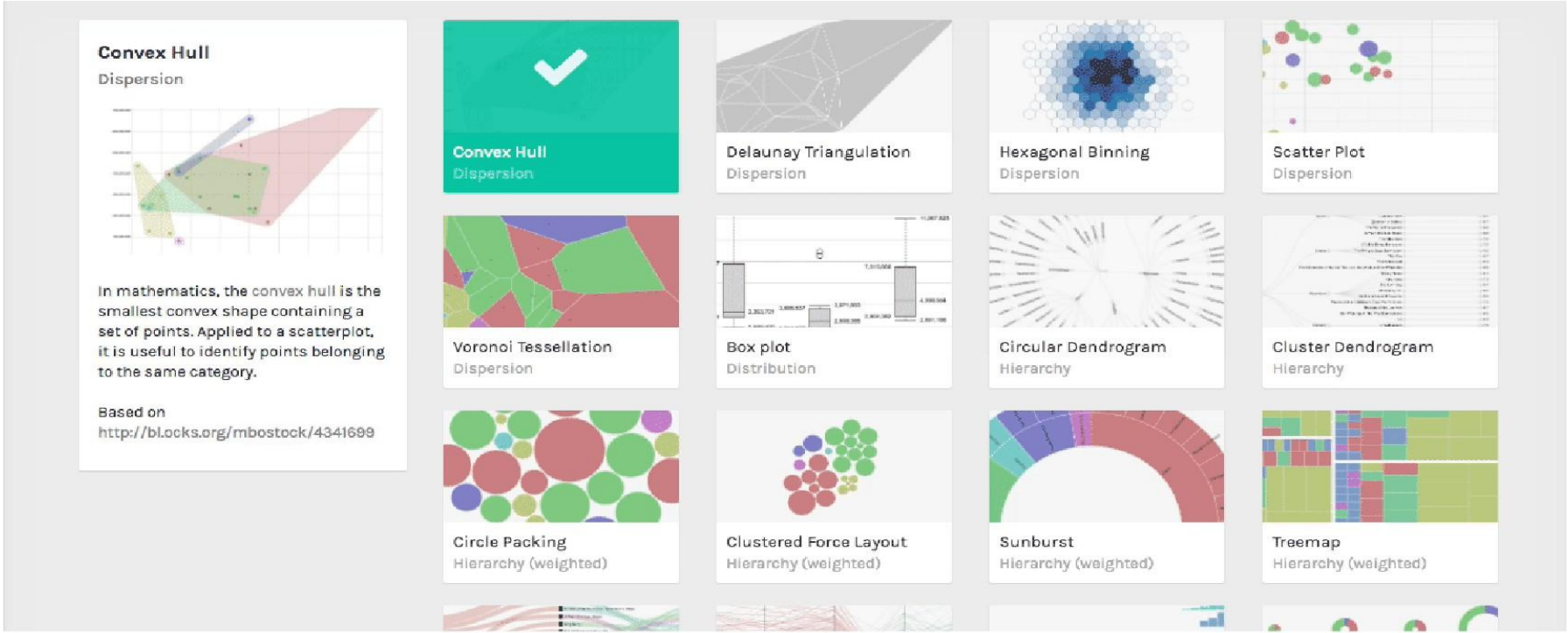
Suggested tool

Google Charts - requires some coding knowledge



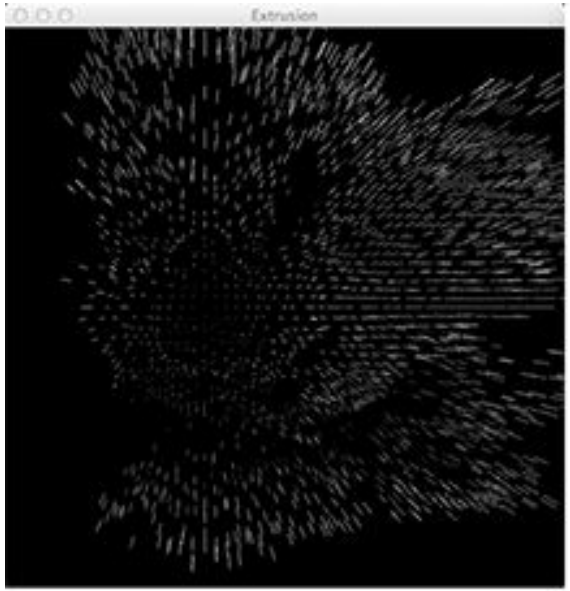
Suggested tool

[RAWGraphs](#)

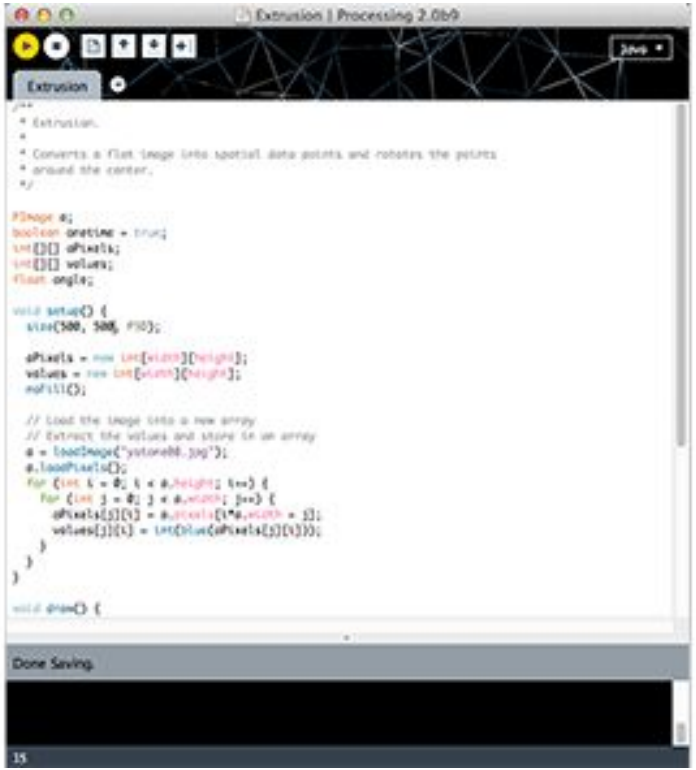


Suggested tool

Processing



Display Window



Toolbar

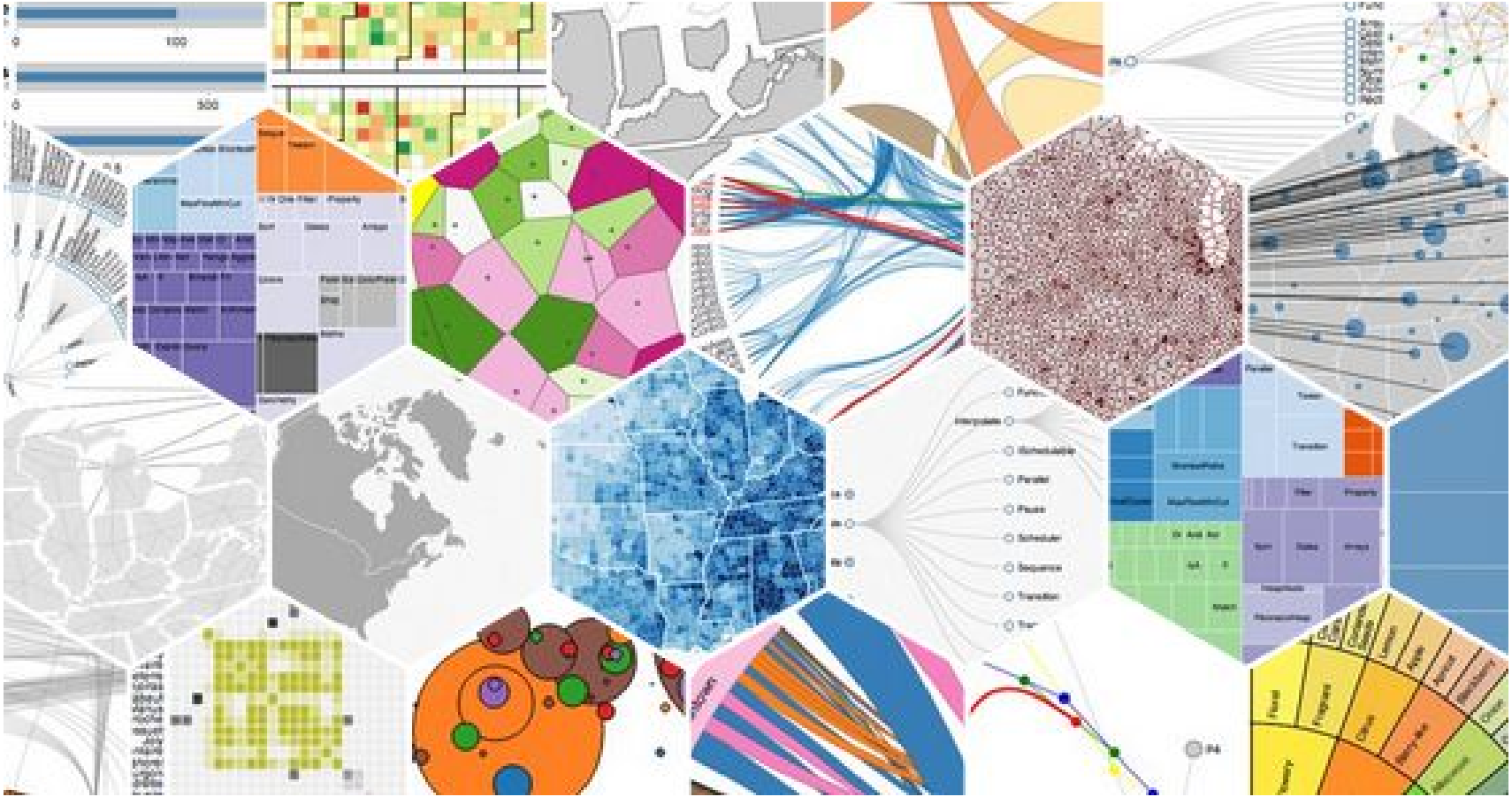
Tabs

Text Editor

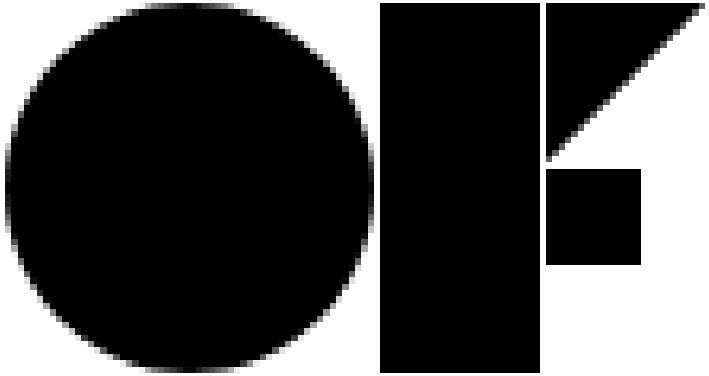
Message Area

Text Area

Advanced: D3 d3js.org



Advanced: C++ Libraries



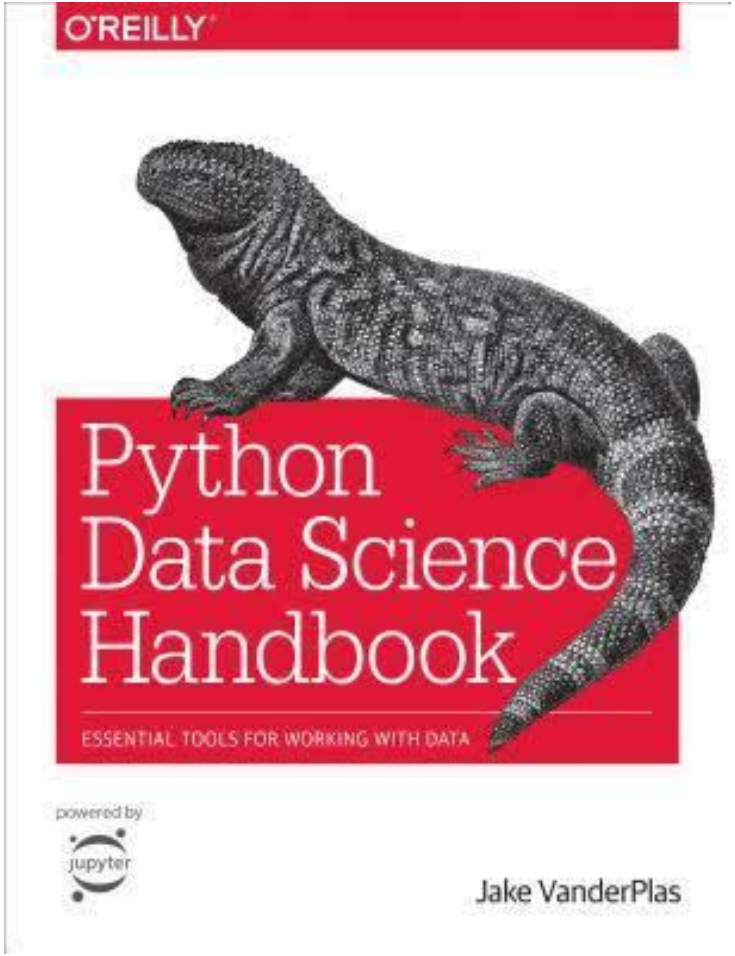
Advanced: Python

Python resources

Python Data Science Handbook:

Essential Tools for Working with Data

[Python Tutorial]



Advanved: R - Statistical Computing

RStudio

	ID	Hometown	State	Country	Latitude	Longitude	Miles_to_SVA	Age	temp	temp_devi
4	Orange dog	Chicago	IL	USA	41.87811	-87.62980	712.40	38.00000	82	
3	Rose cat	Pennsylvania	PA	USA	41.20332	-77.19452	169.94	32.00000	78	
17	Pacific blue turtle	Paramus	NJ	USA	40.94454	-74.07542	13.22	25.00000	70	
6	Sea foam panda	Huntington	NY	USA	40.86815	-73.42568	31.49	22.00000	70	
14	Turquoise bear	New Hyde Park	NY	USA	40.73510	-73.68791	16.51	24.00000	70	
7	Blue wolf	Summit	NJ	USA	40.71464	-74.36461	19.07	31.00000	70	
1	Pink rabbit	Holmdel	NJ	USA	40.34511	-74.18403	29.05	22.00000	72	
11	Deep blue elephant	Cherry Hill	NJ	USA	39.92681	-75.02463	78.37	30.00000	80	
9	Blue-green dog	Dayton	OH	USA	39.75895	-84.19161	541.93	44.00000	70	
10	Navy blue dog	Seoul	Seoul	South Korea	37.56654	126.97797	6868.00	23.00000	75	
12	Jade vulture	Anyang	Kyeonggi	South Korea	37.39425	126.95682	6880.45	24.00000	75	
2	Blue alpaca	Los Angeles	CA	USA	34.05223	-118.24368	2446.75	27.00000	85	
8	Dark slate cat	Nanjing	Jiangsu	China	32.06025	118.79688	7343.15	23.00000	78	

Showing 1 to 13 of 20 entries

Console ~/Desktop/Visualizing Information class/survey/R-sketches/

```
Removed 2 rows containing missing values (geom_point).
> ggplot(students, aes(x=Age, y=temp, size=followers_total, color=spicy)) + geom_point() + scale_size_area(ma
x_size=10) + geom_text(aes(label=ID), size=1, color="black") + scale_color_gradient2(low="blue", high="red")
Warning messages:
1: Removed 2 rows containing missing values (geom_point).
2: Removed 2 rows containing missing values (geom_text).
> ggplot(students, aes(x=Age, y=temp, size=followers_total, color=spicy)) + geom_point() + scale_size_area(ma
x_size=10) + geom_text(aes(label=ID), size=1, color="black") + scale_color_gradient2(low="blue", high="red",
midpoint=5)
Warning messages:
1: Removed 2 rows containing missing values (geom_point).
2: Removed 2 rows containing missing values (geom_text).
> ggplot(students, aes(x=Age, y=temp, size=followers_total, color=spicy)) + geom_point() + scale_size_area(ma
x_size=10) + geom_text(aes(label=ID), size=1, color="black") + scale_color_gradient2(low="blue", high="red",
mid="purple", midpoint=5)
Warning messages:
1: Removed 2 rows containing missing values (geom_point).
2: Removed 2 rows containing missing values (geom_text).
> ggplot(students, aes(x=Age, y=temp, size=followers_total, color=spicy)) + geom_point() + scale_size_area(ma
x_size=10) + geom_text(aes(label=ID), size=2, color="black", vjust=-2) + scale_color_gradient2(low="blue", hi
gh="red", mid="purple", midpoint=5)
Warning messages:
1: Removed 2 rows containing missing values (geom_point).
2: Removed 2 rows containing missing values (geom_text).
>
```

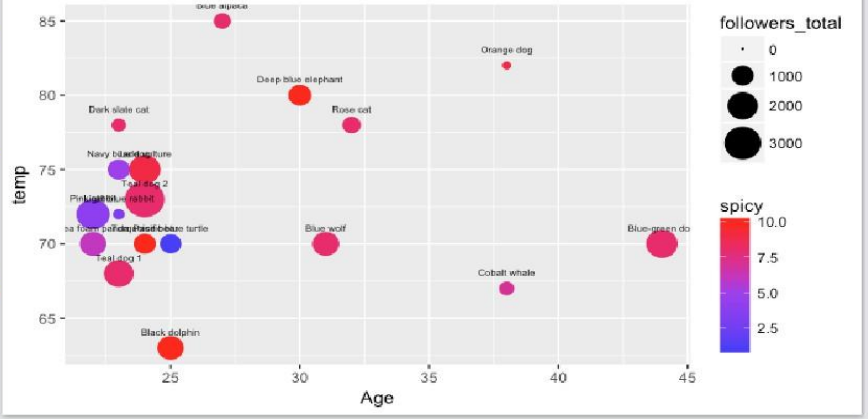
Environment | History

To Console | To Source

```
library("RColorBrewer", lib.loc="/Library/Frameworks/R.framework/Versions/3.4/Resources/li...
followers <- read.csv("followers_num_tall.csv")
View(followers)
treemap(id=followers$ID, area=followers$total, group=followers$ID, color=followers$type, l...
map.market(id=followers$ID, area=followers$total, group=followers$ID, color=followers$type...
treemap(id=followers$ID, area=followers$total, group=followers$ID, color=followers$type_co...
treemap(id=followers$ID, area=followers$amt, group=followers$ID,
    color=followers$type_color, lab = c(TRUE, TRUE), main="social media followers",
    pal="Oranges", linecol="white", textcol="white")
map.market(id=followers$ID, area=followers$amt, group=followers$ID, color=followers$type_c...
map.market(id=followers$ID, area=followers$percent, group=followers$ID, color=followers$ty...
map.market(id=followers$ID, area=followers$percent, group=followers$type, color=followers$...
map.market(id=followers$type, area=followers$percent, group=followers$ID, color=followers$...
map.market(id=followers$type, area=followers$amt, group=followers$ID, color=followers$type...
amt
map.market(id=followers$type, area=followers$amt, group=followers$ID, color=followers$type...
map.market(id=followers$amt, area=followers$amt, group=followers$ID, color=followers$type_...
treemap(id=followers$amt, area=followers$amt, group=followers$ID, color=followers$type_col...
map.market(id=followers$amt, area=followers$amt, group=followers$ID, color=followers$type_...
map.market(id=followers$amt, area=followers$amt, group=followers$ID, color=followers$type_...
```

Files | Plots | Packages | Help | Viewer

Zoom | Export | Publish

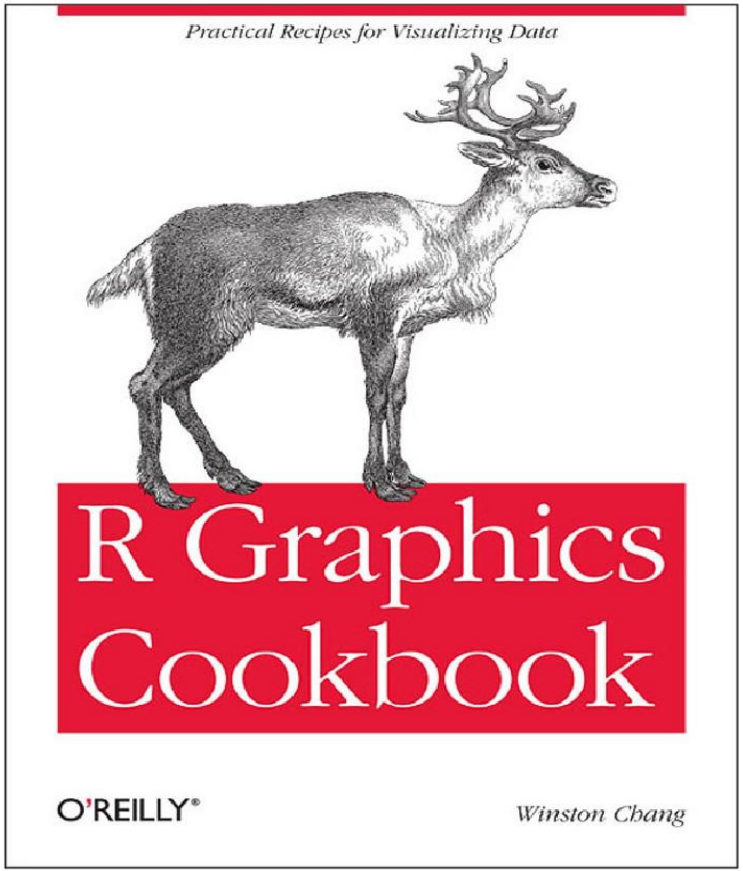


R resources

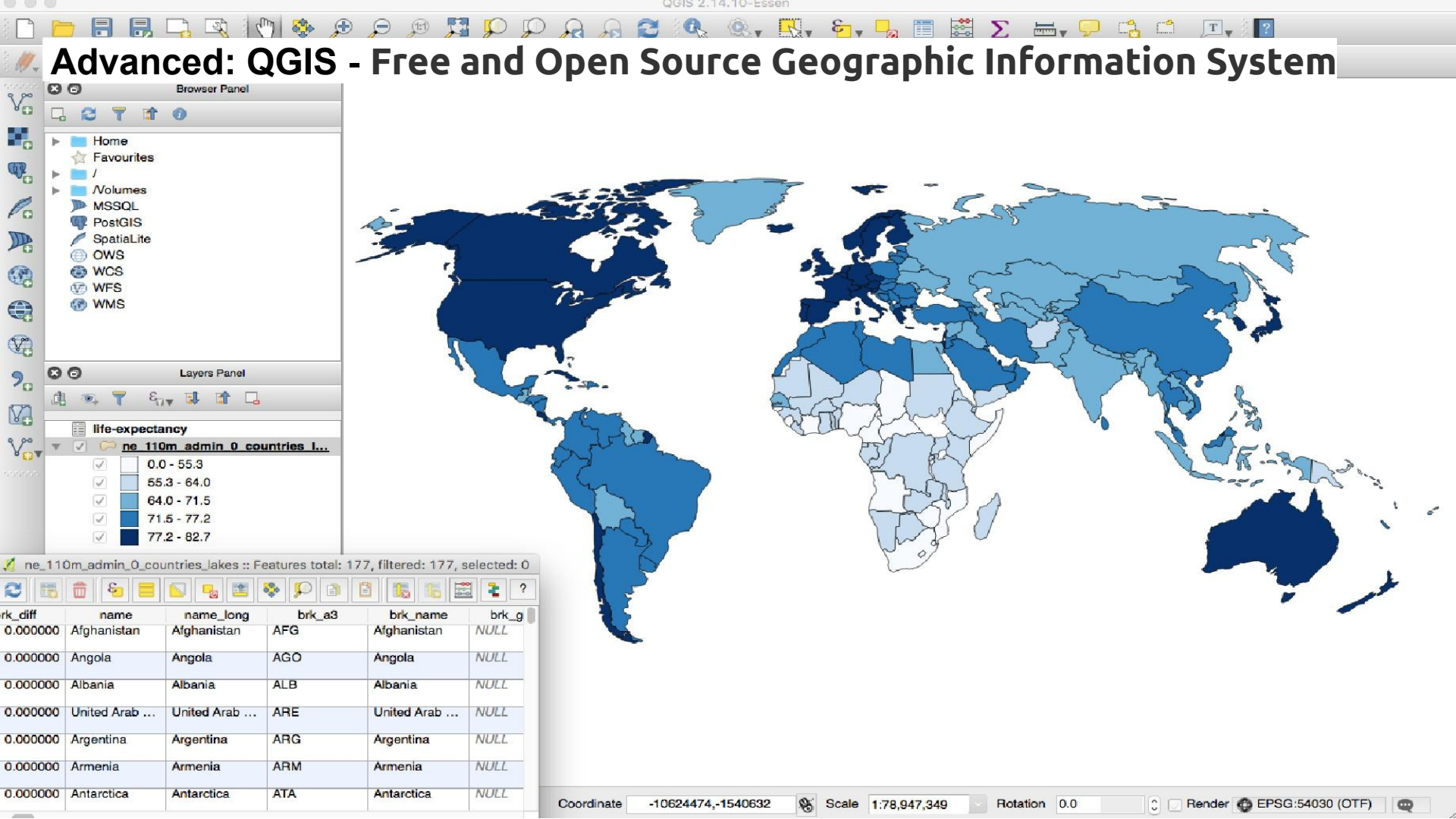
First download R
Then download R Studio

R Graphics Cookbook

[R Tutorial]



Advanced: QGIS - Free and Open Source Geographic Information System



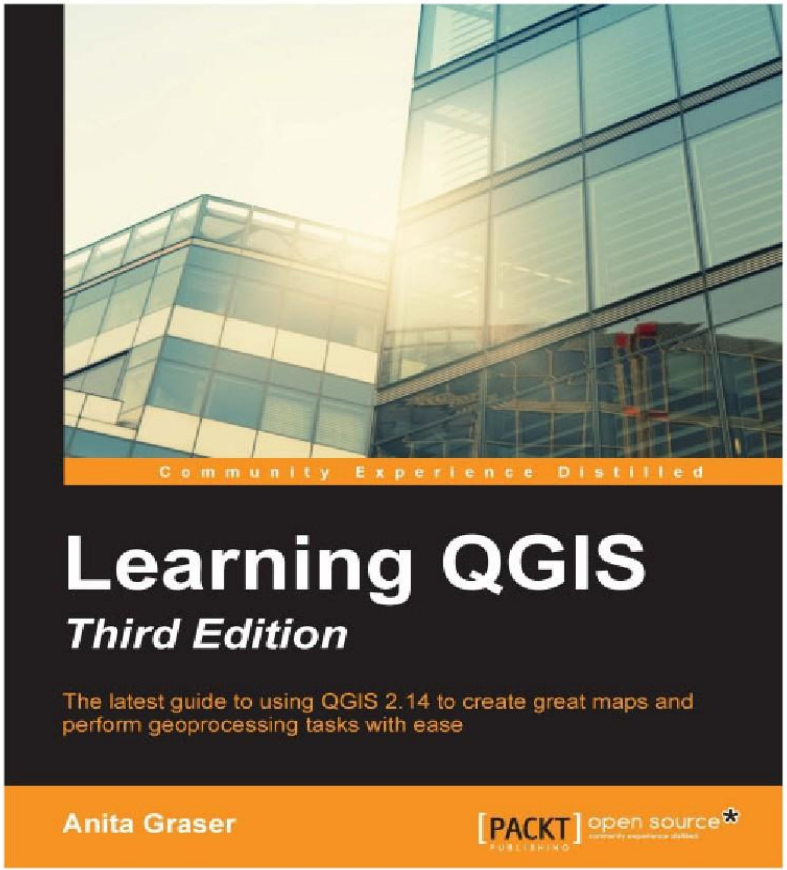
QGIS resources

QGIS download

Tutorials
Learning QGIS book

OpenStreetMap
OpenStreetMap data extracts

Natural Earth
Trimble Data Marketplace



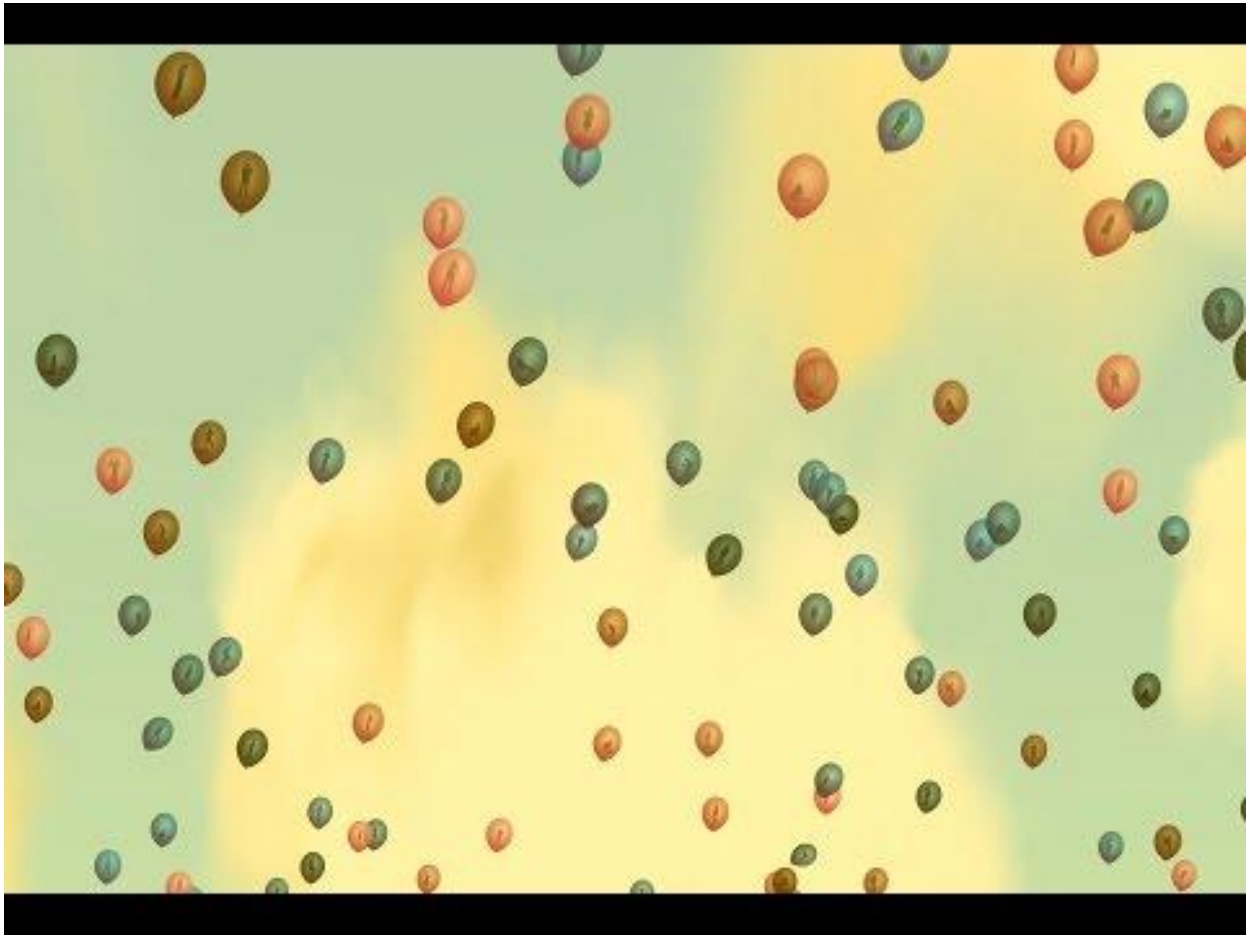
Data Art vs. Data Visualization

Thousands of Exhausted Things - The Office of Creative Research

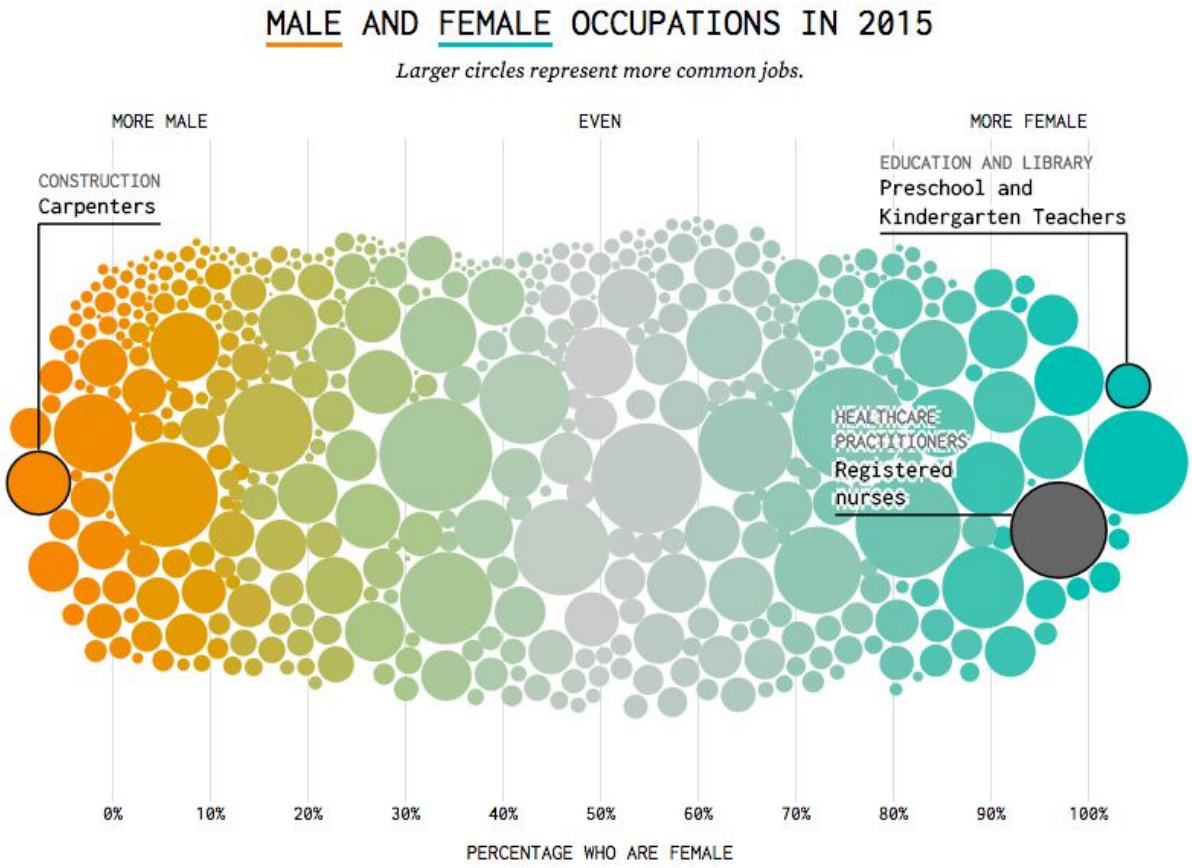
LINK

Serendipity - Kyle McDonald

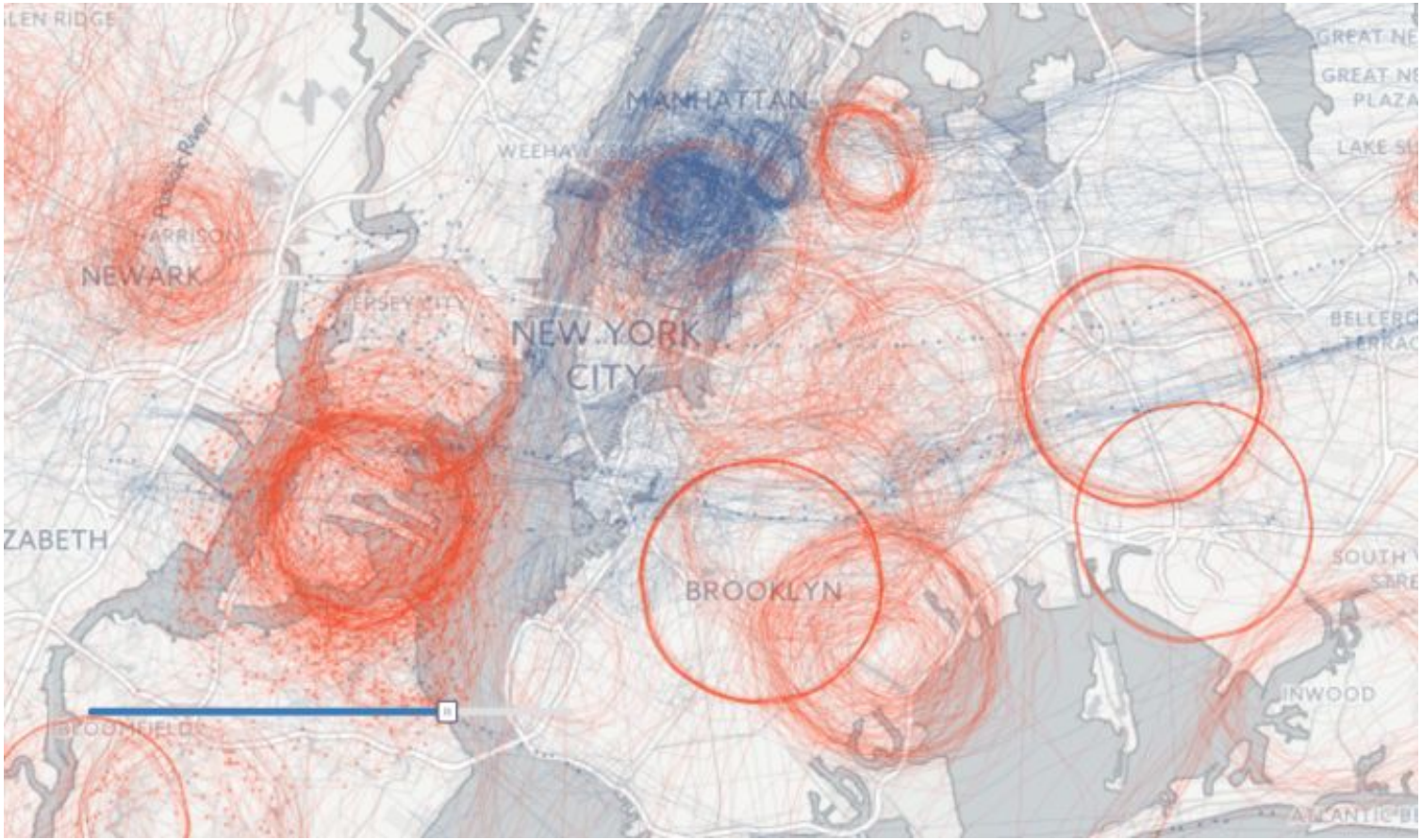
I Want You to Want Me - Jonathan Harris



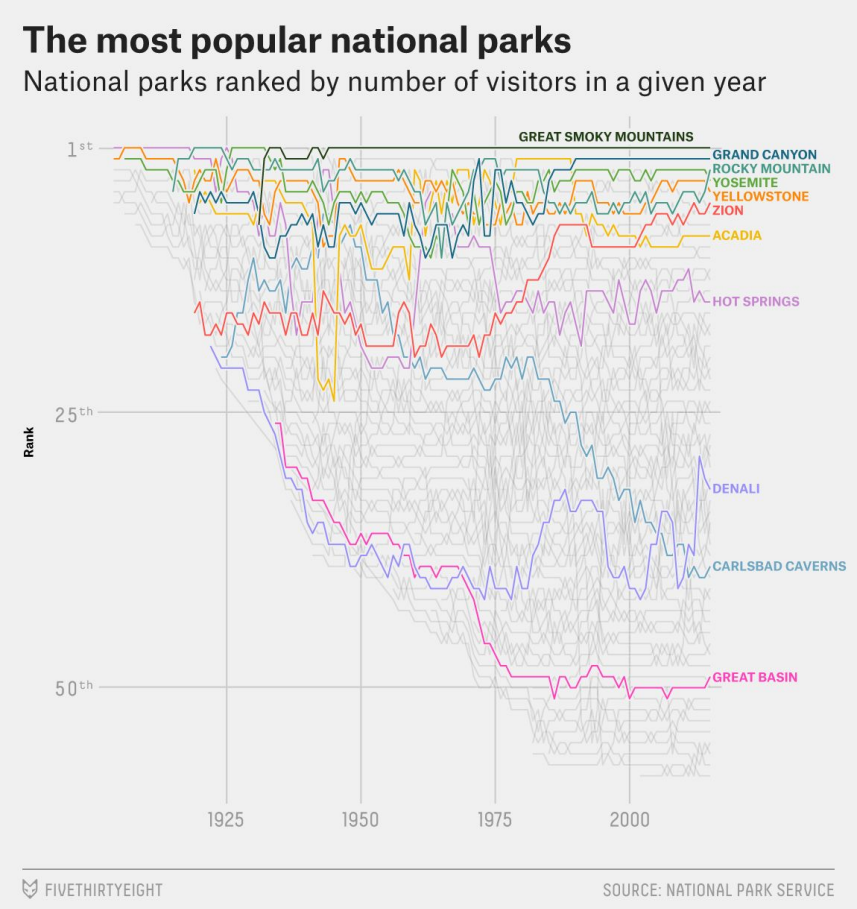
Occupations - Nathan Yau



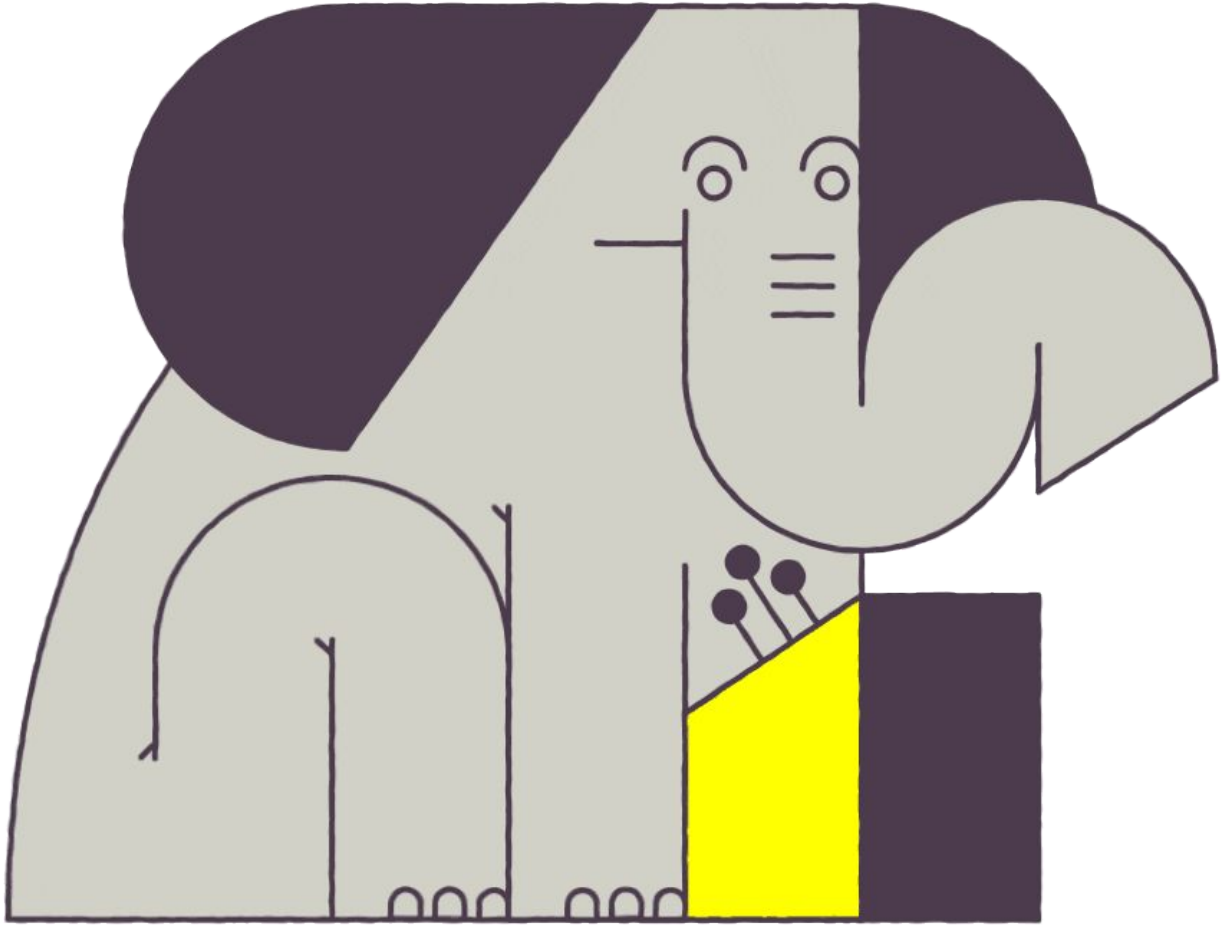
Spies in the Skies - Buzzfeed



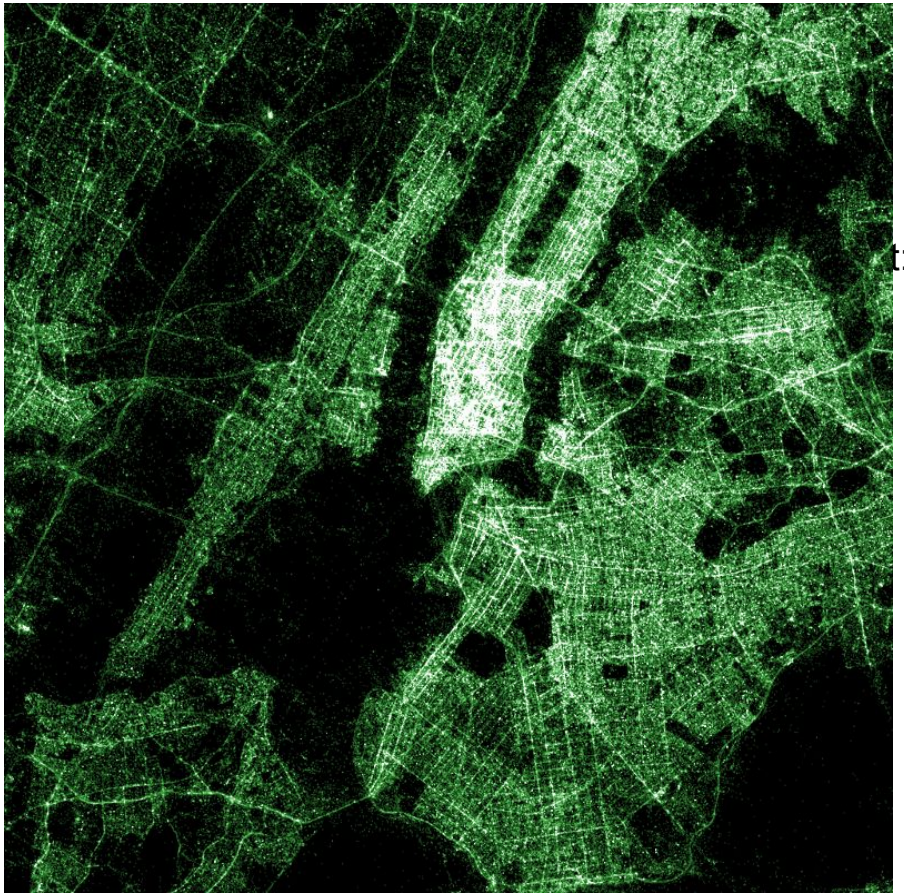
National Parks-FiveThirtyEight

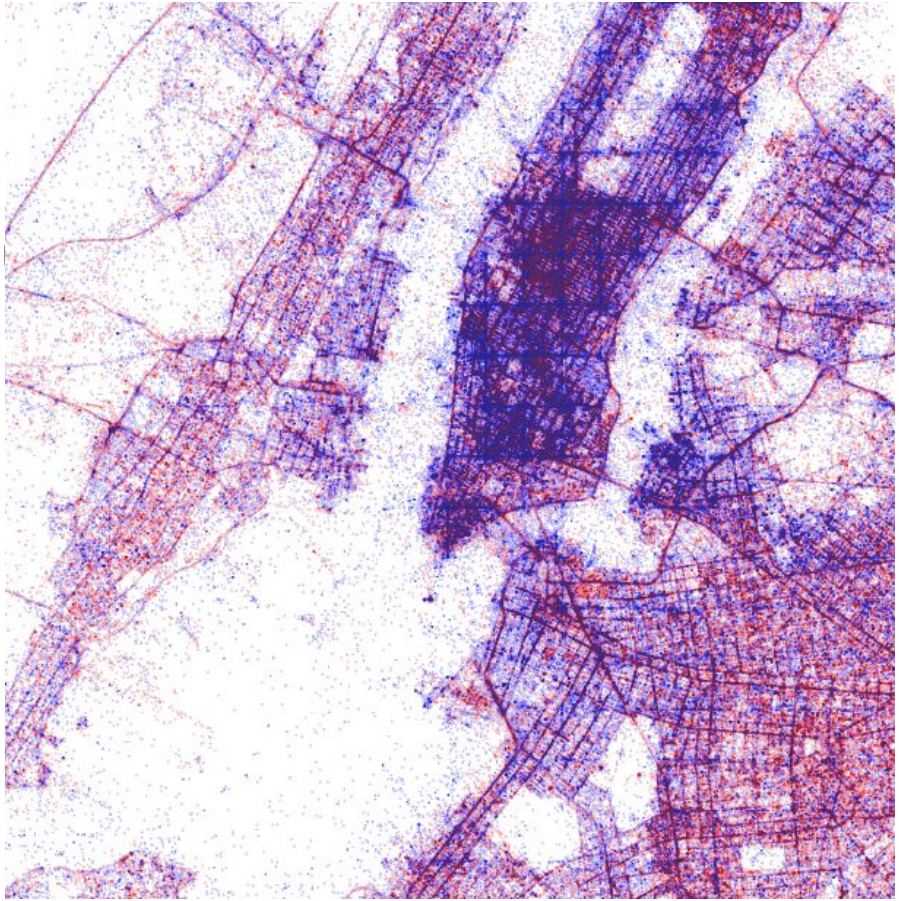


Republican Stump Speech-FiveThirtyEight



Shazam and One Billion Recognitions-Umar Hansa





Assignment

Take your assigned data set and formulate the following:

1. **Before looking at the specifics of the data, develop a set of questions you want answered with this data set. You may find that the data doesn't quite provide enough of what you want to concern, in that case, modify your questions or gather extra data.**
2. **Your data should be clean, but in the case that it's not, tidy it according to your needs. Document this process.**
3. **Organize your data according to all that you know about organizational methods. Be prepared to defend this as they relate to the questions you want answered and the message to be relayed.**
4. **Use an easier charting software (Sheets with sparklines, google charts, Processing, or RawGraphs (or if you feel inclined, one of the advanced methods or languages) to visualize your data.**

We will present in class next Monday. BTW- your work can be more data art-but still must present all of the above.