`DataBOOM`: the canon for data science

Databrew

2021-06-09

Contents

Chapter 1

Welcome!

Welcome to `DataBOOM`, a curriculum designed to guide you from your very first line of code towards becoming a professional data scientist.

What this is, and what it isn't

This is not a textbook or a reference manual. It is not exhaustive or comprehensive. It is a *training manual* designed to *empower researchers to do impactful data science.* As such, its tutorials and exercises aim to get you, the researcher, to start writing your own code as quickly as possible and – equally of importance – to *start thinking like a data scientist*, by which we mean tackling ambiguous problems with persistence, independence, and creative problem solving.

Furthermore, this is not a fancy interactive tutorial with bells or whistles. It was purposefully designed to be simple and "analog". You will not be typing your code into this website and getting feedback from a robot, or setting up an account to track your progress, or getting pretty merit badges or points when you complete each module.

Instead, you will be doing your work on your own machine, working with real folders and files, downloading data and moving it around, etc. – all the things you will be doing as a data scientist in the real world.

Who this is for

This curriculum covers everything from the absolute basics of writing code in `R` to machine learning with `tensorflow`. As such, it is designed to be useful to everyone in some way. But the target audience for these tutorials is the student who *wants* to work with data but has *zero* formal training in programming, computer science, or statistics.

This curriculum was originally developed for the **Sewanee Data Institute for Social Good** at Sewanee: The University of the South, TN, USA.

What you will learn

- The **Core theory** unit establishes the conceptual foundations and motivations for this work: what data science is, why it matters, and ethical issues surrounding it: the good, the bad, and the ugly.

The next several units comprise a *core* curriculum for tackling data science problems:

- The **Getting started** unit teaches you how to use `R` (in `RStudio`) to explore and plot data. Here you will add the first and most important tools to your toolbox: working with variables, vectors, dataframes, scripts, and file directories.
- The **Basic `R` workflow** unit teaches you how to bring in your own data and work with it in `R`. You will learn how to format data to simplify analysis and add tools for *data wrangling* (i.e., transforming and re-formatting data to prepare it for plotting and analysis). You will also learn how to conduct basic statistics, from exploratory data analyses (e.g., producing and comparing distributions) to significance testing.
- The **Essential `R` skills** unit equips you with the tools, tricks, and mindset for tackling the most common tasks in data science. This is where you really begin to cut your teeth on real-world data puzzles: figuring out how to use the `R` tools in your toolbag to tackle an ambiguous problem and deliver an excellent data product.

The next several units provide a suite of skills essential to any data science professional:

- The **Interactive dashboards** unit teaches you how to make dashboards and websites for projects using `shiny` in `RStudio`.
- The **Databases** unit teaches you how to access, create, and work with relational databases online using `SQL` and its alternatives.
- The **Documenting your work** unit teaches you to use `R Markdown` to produce beautiful, reproducible data reports. You will also learn about *version control*, using `Git` and `GitHub` to collaborate on shared projects and work on data science teams.
- The **Sharing research** unit teaches you to produce publishable research articles and compelling presentations.

The final unit, **Advanced skills**, introduces you to a variety of advanced data science techniques, from interactive maps to iterative simulations to machine learning, that can help you begin to specialize your skillset.

Who we are

Joe Brew is a data scientist, epidemiologist, and economist. He has worked with the Florida Department of Health (USA), the Chicago Department of Public Health (USA), the Barcelona Public Health Agency (Spain), the Tigray Regional Health Bureau (Ethiopia) and the Manhiça Health Research Center (Mozambique). He is a co-founder of Hyfe and DataBrew. His research focuses on the economics of malaria and its elimination. He earned his BA at Sewanee: The University of the South (2008), an MA at the Institut Catholique de Paris (2009) and an MPH at the Kobenhavns Universitet (2013). He is passionate about international development, infectious disease surveillance, teaching, running, and pizza.

Ben Brew is a data scientist, economist, and health sciences researcher. In addition to co-founding DataBrew, he has spent most of the last few years working with SickKids Hospital in Ontario on machine learning applications for cancer genomics. He earned his BA at Sewanee: The University of the South (2012), and a Master's in Mathetical Models for Economics from the Paris School of Economics (Paris I) (2015). He is passionate about econometrics, applied machine learning, and cycling.

Eric Keen is a data scientist, marine ecologist, and educator. He is the Science Co-director at BCwhales, a research biologist at Marecotel, a data scientist at Hyfe, and a professor of Environmental Studies at Sewanee: the University of the South. He earned his BA at Sewanee (2008) and his PhD at Scripps Institution of Oceanography (2017). His research focuses on the ecology and conservation of whales in developing coastal habitats. He is passionate about whales, conservation, teaching, small-scale farming, running, and bicycles. And pizza.

Part I

Core theory

Chapter 2

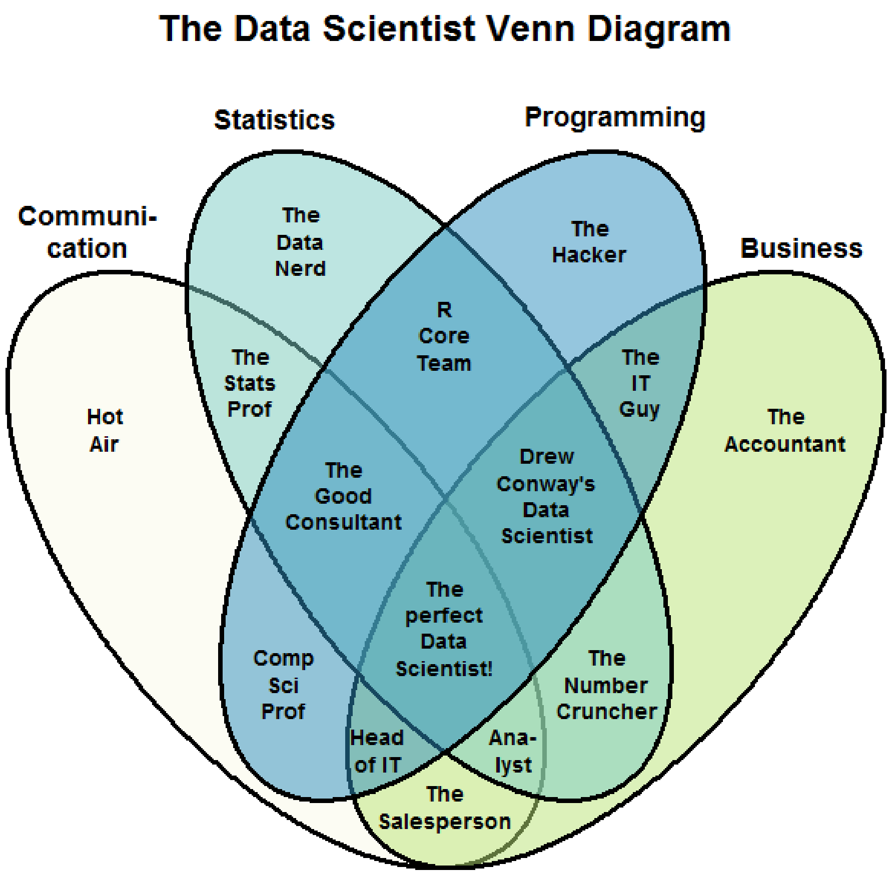
Principles of data science

What is data science?

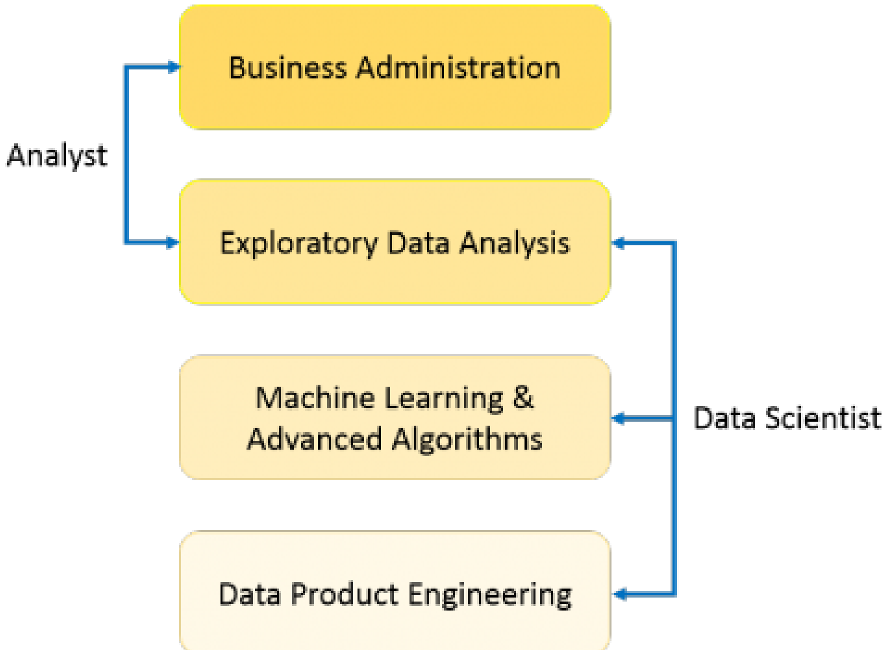
Data science is an interdisciplinary field. Some have argued that it is not a field unto itself, but rather an extension of statistics. In this course, however, we'll take the majority view that data science is its own field: a new field, which combines statistics, mathematics, and computer science.

But we'll go one step outward. Data science is not just the combination of those academic disciplines which form its core; its also something more. Good data science involves domain knowledge (ie, familiarity with the problem being solved), effective communication, an iterative mentality (ie, creating feedback loops for rapid hypothesis testing), a bias to real-world effects rather than theoretical frameworks, and a willingness/desire to work in the real world.

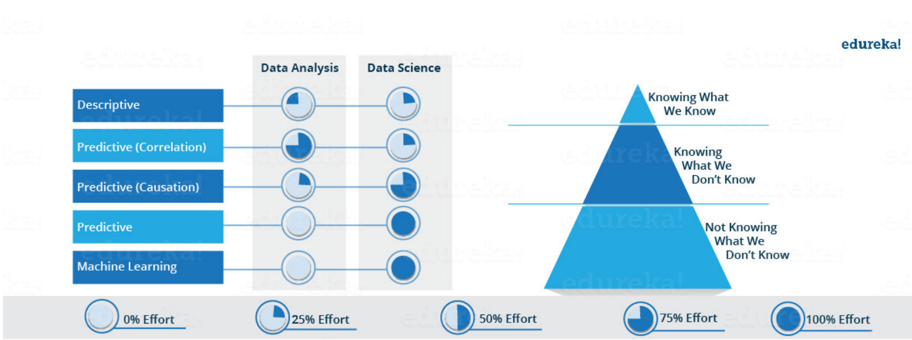
There are a lot of Venn diagrams and figures out there, trying to show what data science is. For example…


The Data Scientist Venn Diagram
Statistics
Programming
Communi-
cation
Business
The Data Nerd
The Hacker
R Core Team
The Stats Prof
The IT Guy
Hot Air
The Accountant
Drew Conway's Data Scientist
The Good Consultant
The perfect Data Scientist!
Comp Sci Prof
The Number Cruncher
Head of IT
Ana-lyst
The Salesperson

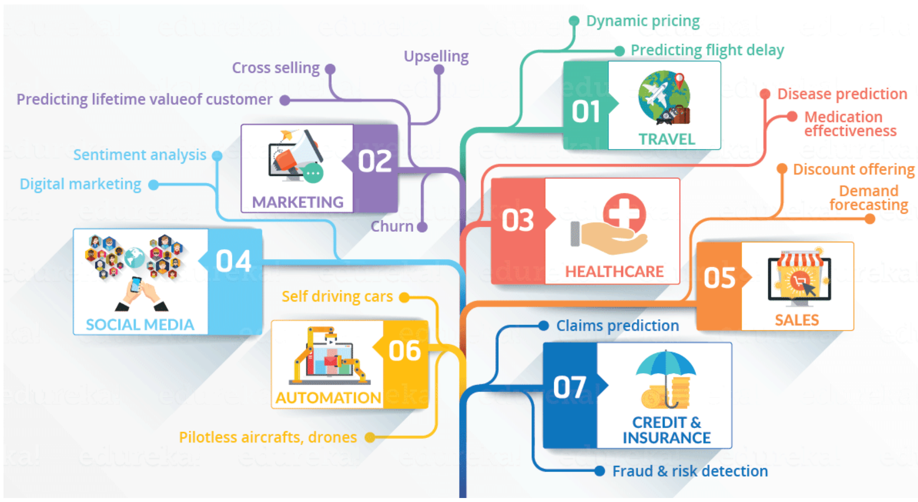
... or ...


Analyst
Business Administration
Exploratory Data Analysis
Machine Learning &
Advanced Algorithms
Data Product Engineering
Data Scientist

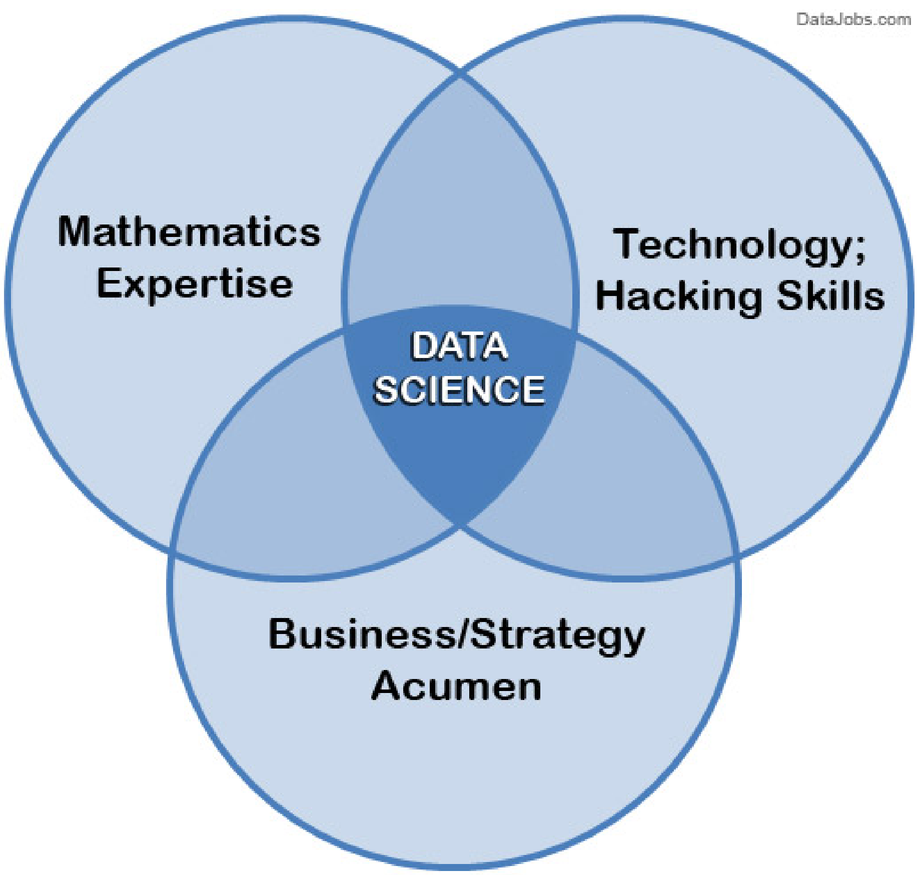
... or ...


edureka!
Data Analysis
Data Science
Descriptive
Predictive (Correlation)
Predictive (Causation)
Predictive
Machine Learning
Knowing What
We Know
Knowing
What We
Don't Know
Not Knowing
What We
Don't Know
0% Effort
25% Effort
50% Effort
75% Effort
100% Effort

... or ...


Dynamic pricing
Predicting flight delay
Cross selling
Upselling
Predicting lifetime valueof customer
01
TRAVEL
Disease prediction
Medication effectiveness
Sentiment analysis
Digital marketing
02
MARKETING
Discount offering
Demand forecasting
Churn
03
HEALTHCARE
04
SOCIAL MEDIA
05
SALES
Self driving cars
Claims prediction
06
AUTOMATION
07
CREDIT & INSURANCE
Pilotless aircrafts, drones
Fraud & risk detection

... or ...


DataJobs.com
Mathematics Expertise
Technology; Hacking Skills
DATA SCIENCE
Business/Strategy Acumen

Our take? These are useful, interesting, and to some degree accurate *but* data science is too new, and too fluid, to be fixed into some static definition. 15 years ago, data science was "dashboards" and "predictive analytics". 10 years ago, data science was "big data" and "data mining". Since then, machine learning and artificial intelligence have taken up an ever-larger chunk of what most people consider data science. And in 5 or 10 years? Who knows.

So, to keep our definition accurate, we'll keep it broad. Data science is simply "doing science with data". And for our purposes, the only difference between our definition and the definition of science itself is not in the word "data" (since nowadays all scientists are, to some degree, "data scientists"), but rather in the word "doing". Data science is about *doing* stuff with data. And that's what this course is going to be about. *DOING.*

Data scientist vs data analysist vs data engineer

There is a lot of fuss out there about what constitutes data science and what doesn't. Our definition is broad enough to encompass lots of things. So, we consider an "analyst" working in business intelligence to be a data scientist; and so too do we think that a data scientist could be an engineer who is processing large amounts of data to extract basic trends. Again, data scientists are those who *do science with data.* That's a lot of people.

What is the data life cycle?

There is a misperception about data science work that it is largely or even exclusively interpretative: that is, a data scientist looks at a big set of data, a light bulb goes off in her head, she has some insight, and then acts on that insight.

The reality is data science is much more than that. And most of data science is a combination of (a) getting data ready for analysis, (b) hypothesis testing, and (c) figuring out what to do with the results of a and b. That is, data science in practice is generally not some artesenal genius staring at a table of numbers until "insight" magically occurs, but rather a lot of work, a lot of structured theories which can be confirmed or falsified, and a lot of imagination applied to the task of implementation. Good, effective data scientists, in practice, are often not those with the most cutting-edge algorithm, but rather those who are able to get the right data and ask the right questions.

In other words, data goes through a whole *lifecycle* of which analysis is just a small part. What is the data lifecycle?

Here's how we conceptualize it:

0. Observation
1. Problem identification and definition
2. Question formation
3. Hypothesis generation
4. Data collection
5. Data processing
6. Model building / hypothesis testing
7. Operationalization
8. Communication / dissemination
9. Action
10. Observation

Does the above look a lot like the sceintific method? That's because it basically is. The main differences are (a) "data processing" (which in reality takes up most of any data scientist's time), (b) the bias towards action, and (c) the iterative / looped nature of the lifecycle.

Data science 'in the wild'

Enough theory - let's talk about what data scientists are actually doing in the real world.

Data scientists are working on a ton of problems. Some examples include:

- Targeted advertising
- Dynamic airline pricing
- Social media feed optimization
- Making video games more fun / addictive
- Identifying and filtering inappropriate content on the internet

- Search autocomplete
- Automating identification of credit card fraud
- Facial recognition
- Voice recognition
- Filtering spam
- Autocorrect
- Virtual assistants
- Preventive maintenance at nuclear facilities
- Identifying tax evaders
- Identifying disease through imagery
- Improving chemotherapy dosage
- Quantifying the likelihood of recedivism to prevent over-incarceration
- Surveilling emergency rooms to predict disease outbreaks
- Improve matchmaking systems (liver transplants, love, etc.)
- Detecting fake news
- Predictive policing

Why R?

This course if largely about learning to *do*, and will largely use R. R is not the only tool in the data scientists' toolbox, but it's a good one, is extremely popular, has a larger community and user base, and can be applied to many fields.

What are some of the other tools and languages used by data scientists? There are plenty (and we won't cover them here). But the main reason we choose R for this course is the fact that it is open-source and free. This matters because…

The reproducibility crisis

There is a crisis in science (and data science): the reproducitibility crisis (also known as the replication crisis). This refers to the fact that many scientific studies have been impossible to reproduce, calling into question the validity of those studies' findings. This is a big deal: if a significant part of science is *wrong* then what do we know, how can we be sure what we know is right, and how can we separate the wheat from the chaff?

Because of this crisis, there has emerged a much needed move to make all science "reproducible". This means making sure that someone else can copy what you did, and get the same results. This is important for identifying scientific fraud, of course, but also for helping us to overcome human bias, mistakes, wishful thinking, etc. Reproducibility is not just a "nice-to-have"; in modern science (and data science), it's a "must".

Good data science must be reproducible. And reproducible science means using tools that others can easily use, and methods that others can easily copy. Programming languages like R and Python are ideal for this.

In this course, we'll focus on *reproducible research*, *literate programming*, *documentation*, and other components of data science (and research in general) which ensure that (a) our methods and findings can be easily sanity-tested by others, and (b) we set ourselves and our projects' up for future collaborations, hand-offs, and expansion.

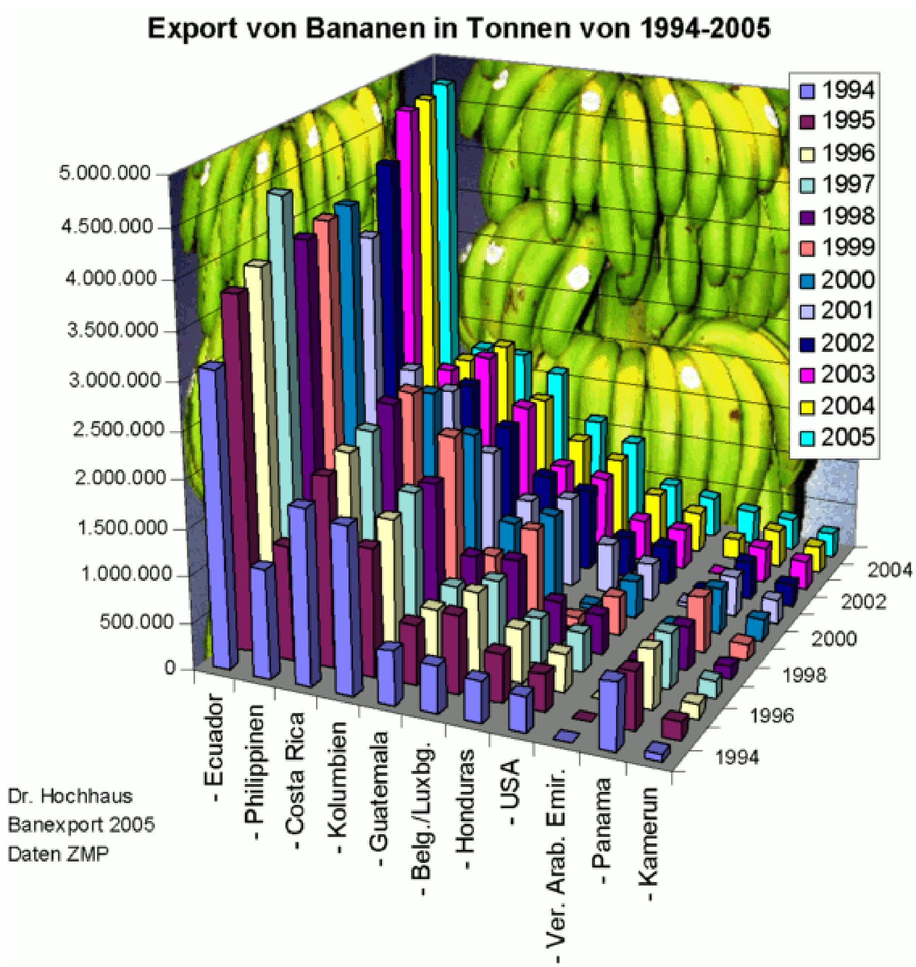
Chapter 3

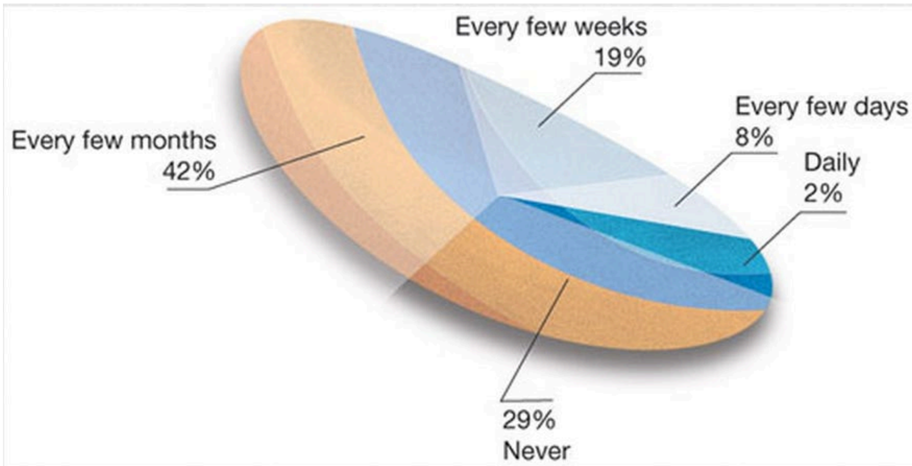
Visualizing data

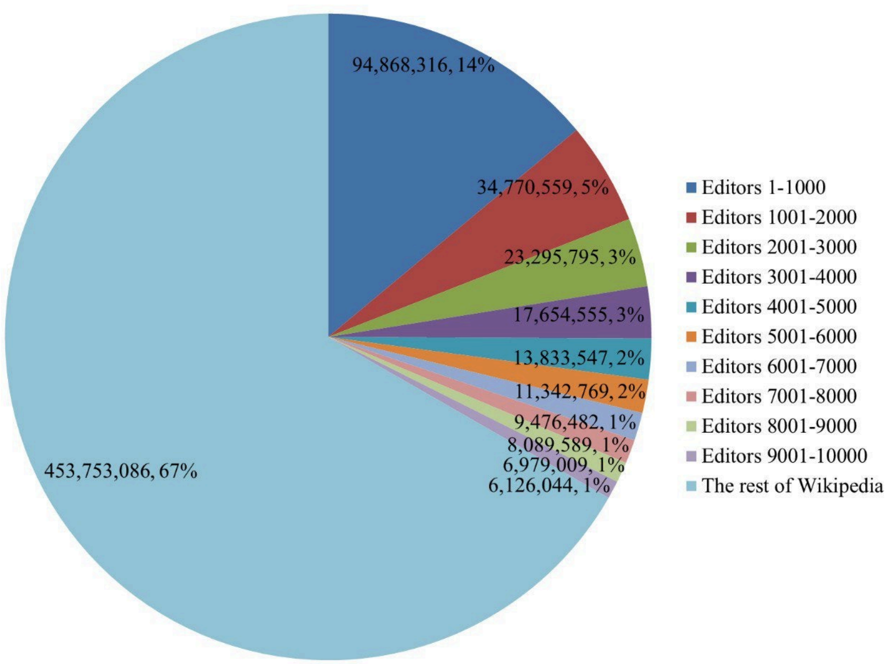
3.1 The importance of visualization

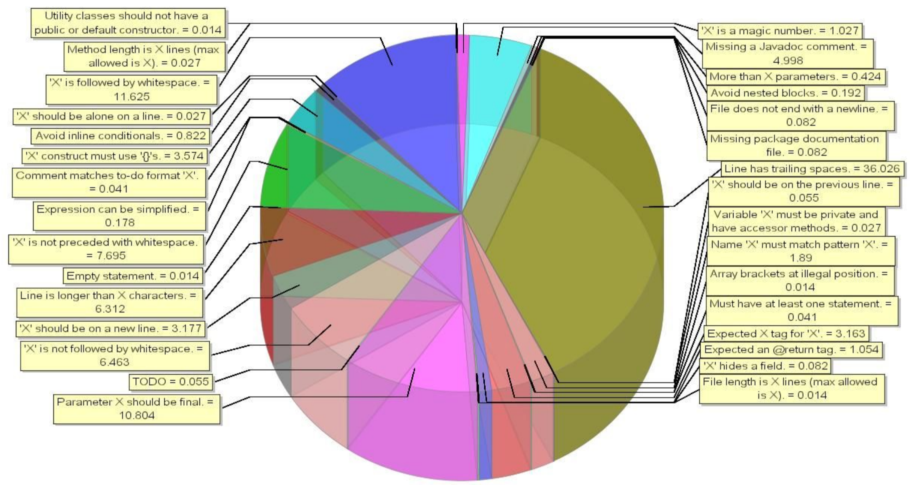
Much of the next few weeks is going to focus on visualization. The reason why is that data visualization is a powerful, quick, and clear tool for communicating data, exploring it, and understanding it. With that in mind, let's get started with some *data viz*.

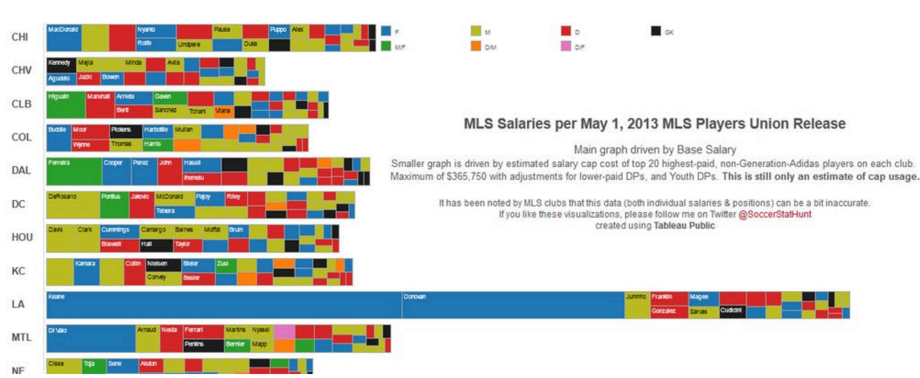
3.2 Bad examples

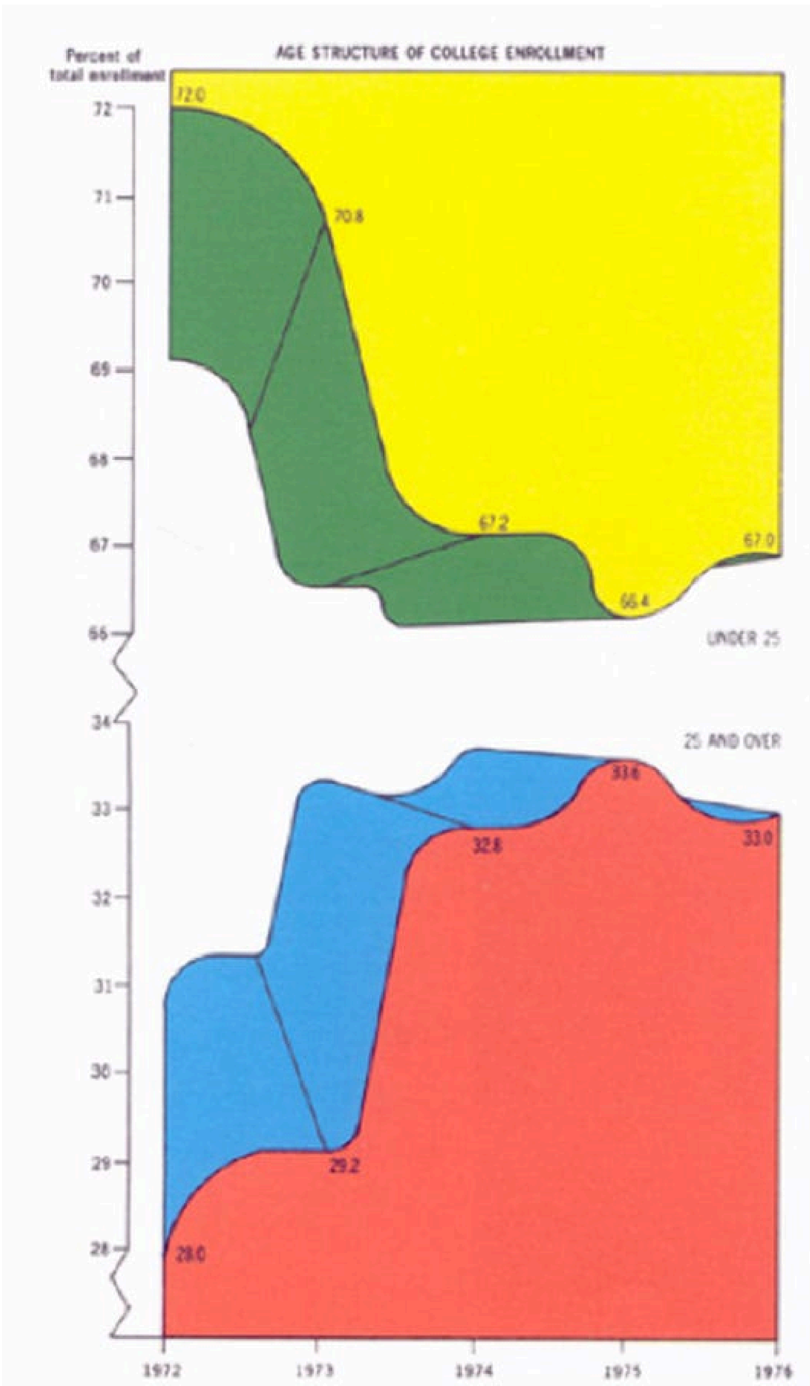


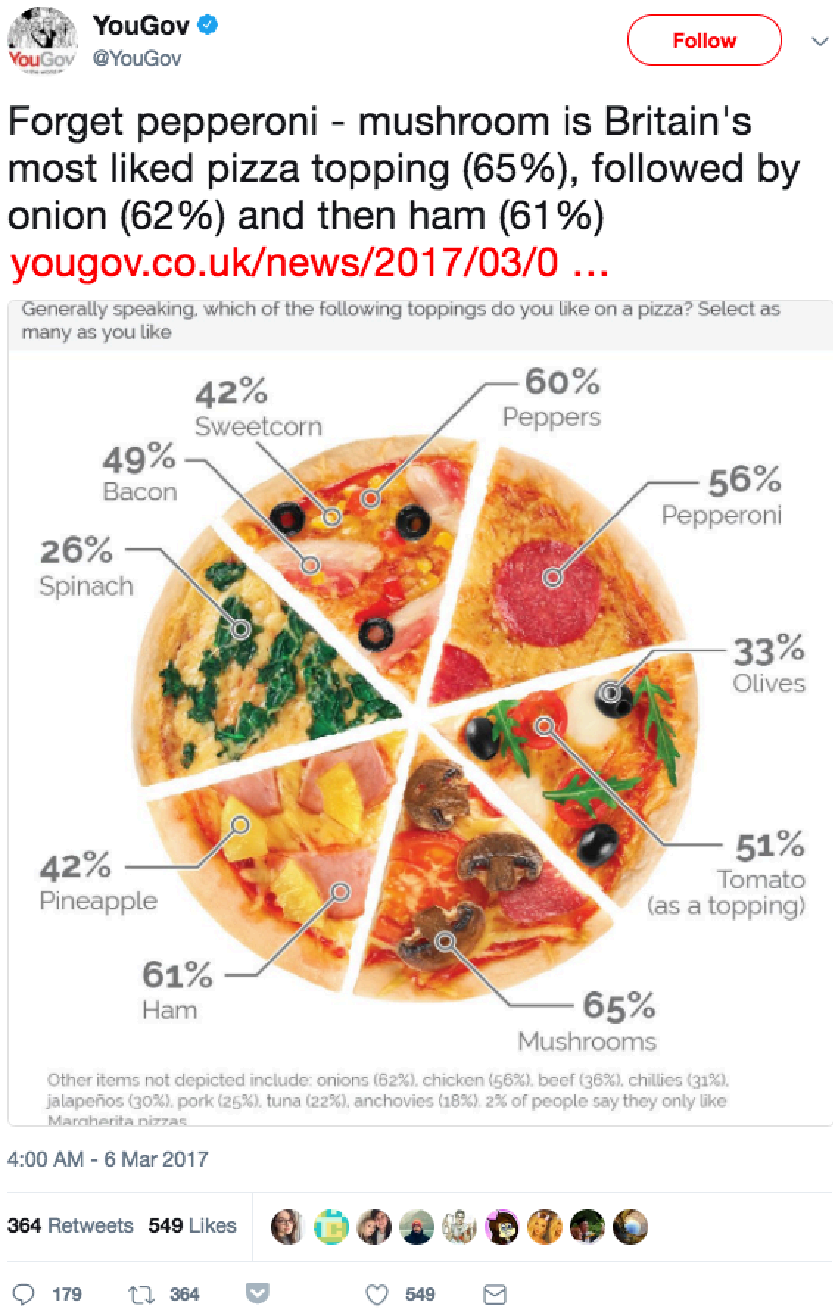

Every few weeks
19%
Every few days
8%
Daily
2%
Every few months
42%
29%
Never


94,868,316, 14%
34,770,559, 5%
23,295,795, 3%
17,654,555, 3%
13,833,547, 2%
11,342,769, 2%
9,476,482, 1%
8,089,589, 1%
6,979,009, 1%
6,126,044, 1%
453,753,086, 67%
Editors 1-1000
Editors 1001-2000
Editors 2001-3000
Editors 3001-4000
Editors 4001-5000
Editors 5001-6000
Editors 6001-7000
Editors 7001-8000
Editors 8001-9000
Editors 9001-10000
The rest of Wikipedia

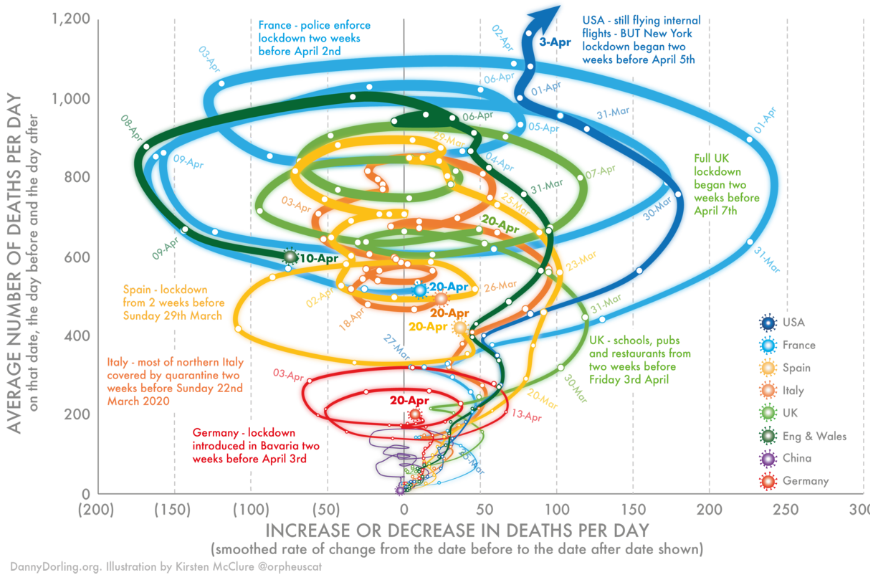

Utility classes should not have a public or default constructor. = 0.014
Method length is X lines (max allowed is X). = 0.027
'X' is followed by whitespace. = 11.625
'X' should be alone on a line. = 0.027
Avoid inline conditionals. = 0.822
'X' construct must use '{}'s. = 3.574
Comment matches to-do format 'X'. = 0.041
Expression can be simplified. = 0.178
'X' is not preceded with whitespace. = 7.695
Empty statement. = 0.014
Line is longer than X characters. = 6.312
'X' should be on a new line. = 3.177
'X' is not followed by whitespace. = 6.463
TODO = 0.055
Parameter X should be final. = 10.804
'X' is a magic number. = 1.027
Missing a Javadoc comment. = 4.998
More than X parameters. = 0.424
Avoid nested blocks. = 0.192
File does not end with a newline. = 0.082
Missing package documentation file. = 0.082
Line has trailing spaces. = 36.026
'X' should be on the previous line. = 0.055
Variable 'X' must be private and have accessor methods. = 0.027
Name 'X' must match pattern 'X'. = 1.89
Array brackets at illegal position. = 0.014
Must have at least one statement. = 0.041
Expected X tag for 'X'. = 3.163
Expected an @return tag. = 1.054
'X' hides a field. = 0.082
File length is X lines (max allowed is X). = 0.014


CHI
CHV
CLB
COL
DAL
DC
HOU
KC
LA
MTL
NE
MLS Salaries per May 1, 2013 MLS Players Union Release
Main graph driven by Base Salary
Smaller graph is driven by estimated salary cap cost of top 20 highest-paid, non-Generation-Adidas players on each club.
Maximum of $365,750 with adjustments for lower-paid DPs, and Youth DPs. This is still only an estimate of cap usage.
It has been noted by MLS clubs that this data (both individual salaries & positions) can be a bit inaccurate.
If you like these visualizations, please follow me on Twitter @SoccerStatHunt
created using Tableau Public

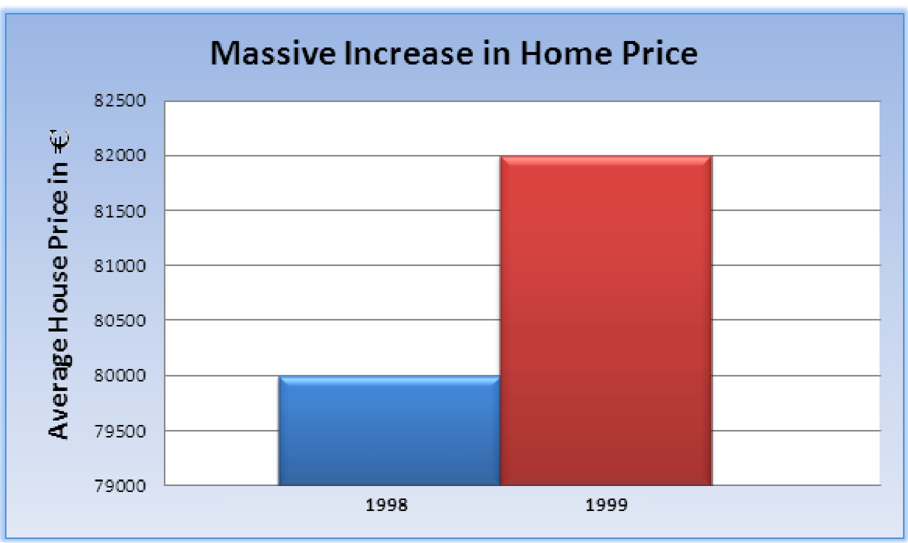

AGE STRUCTURE OF COLLEGE ENROLLMENT
Percent of total enrollment
72
71
70
69
68
67
66
34
33
32
31
30
29
28
72.0
70.8
67.2
66.4
67.0
UNDER 25
25 AND OVER
33.6
32.8
33.0
29.2
28.0
1972
1973
1974
1975
1976



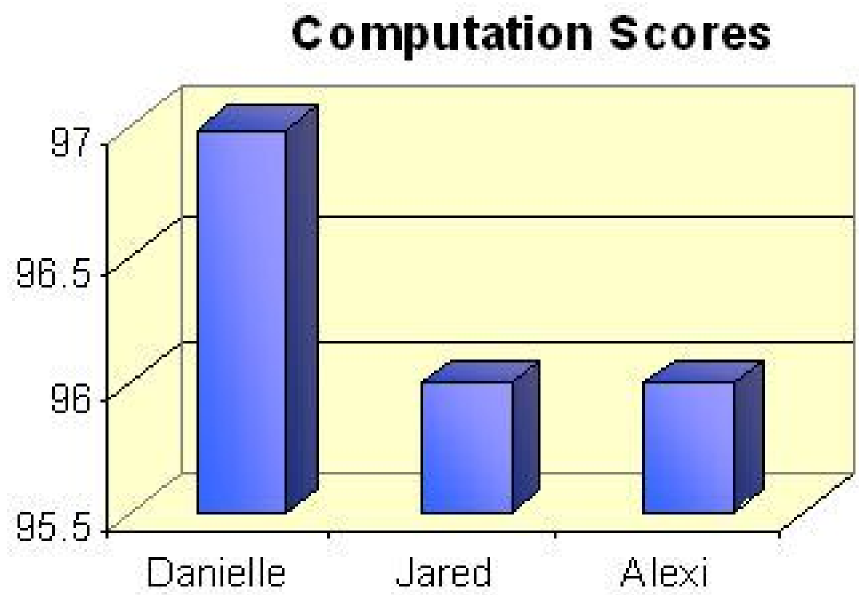
YouGov
@YouGov
Follow
Forget pepperoni - mushroom is Britain's most liked pizza topping (65%), followed by onion (62%) and then ham (61%)
yougov.co.uk/news/2017/03/0 ...
Generally speaking, which of the following toppings do you like on a pizza? Select as many as you like
42% Sweetcorn
60% Peppers
49% Bacon
56% Pepperoni
26% Spinach
33% Olives
42% Pineapple
51% Tomato (as a topping)
61% Ham
65% Mushrooms
Other items not depicted include: onions (62%), chicken (56%), beef (36%), chillies (31%), jalapeños (30%), pork (25%), tuna (22%), anchovies (18%). 2% of people say they only like Margherita pizzas
4:00 AM - 6 Mar 2017
364 Retweets 549 Likes
179
364
549

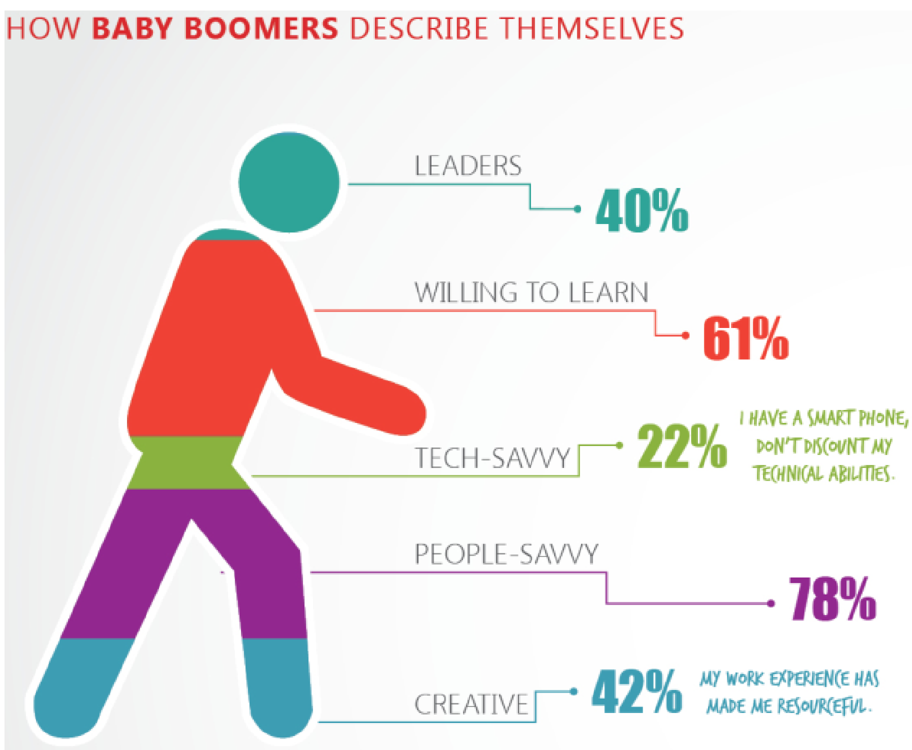


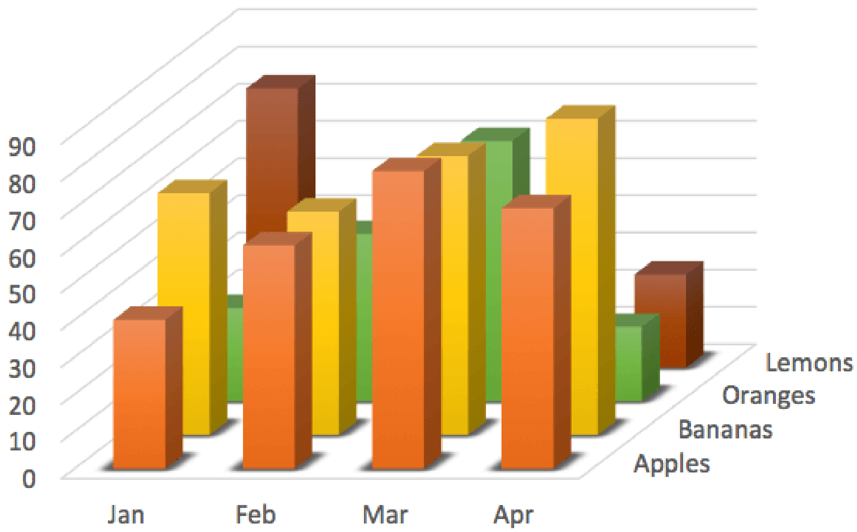
AVERAGE NUMBER OF DEATHS PER DAY
on that date, the day before and the day after
INCREASE OR DECREASE IN DEATHS PER DAY
(smoothed rate of change from the date before to the date after date shown)
France - police enforce lockdown two weeks before April 2nd
USA - still flying internal flights - BUT New York lockdown began two weeks before April 5th
Full UK lockdown began two weeks before April 7th
Spain - lockdown from 2 weeks before Sunday 29th March
Italy - most of northern Italy covered by quarantine two weeks before Sunday 22nd March 2020
UK - schools, pubs and restaurants from two weeks before Friday 3rd April
Germany - lockdown introduced in Bavaria two weeks before April 3rd
USA
France
Spain
Italy
UK
Eng & Wales
China
Germany
DannyDorling.org. Illustration by Kirsten McClure @orpheuscat



Massive Increase in Home Price
Average House Price in €
82500
82000
81500
81000
80500
80000
79500
79000
1998
1999


Computation Scores
97
96.5
96
95.5
Danielle
Jared
Alexi


HOW BABY BOOMERS DESCRIBE THEMSELVES
LEADERS
40%
WILLING TO LEARN
61%
TECH-SAVVY
22%
I HAVE A SMART PHONE, DON'T DISCOUNT MY TECHNICAL ABILITIES.
PEOPLE-SAVVY
78%
CREATIVE
42%
MY WORK EXPERIENCE HAS MADE ME RESOURCEFUL.


90
80
70
60
50
40
30
20
10
0
Jan
Feb
Mar
Apr
Lemons
Oranges
Bananas
Apples