Introduction

Professor Halterman

Michigan State University

PLS 397 Analyzing and Visualizing Data
Fall 2023

Welcome!

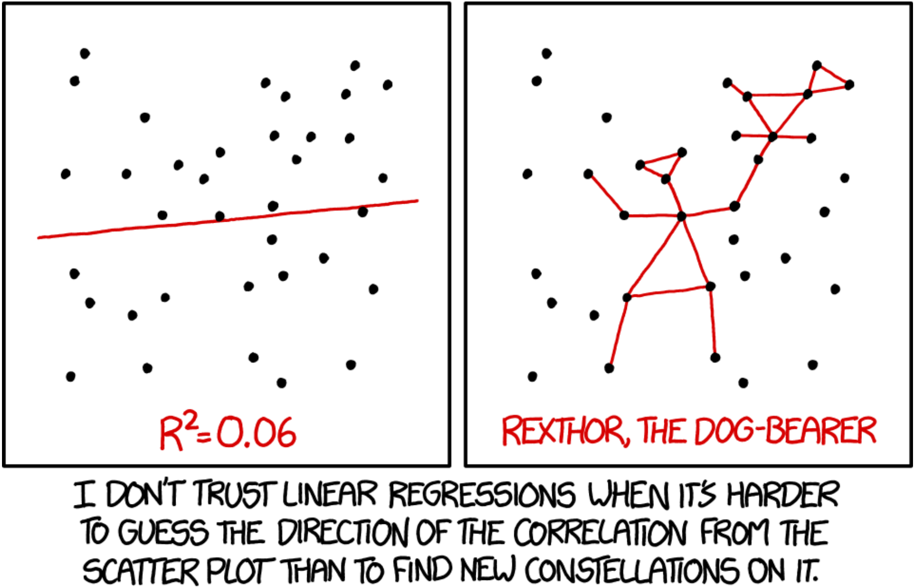
- This class is on analyzing and visualizing data.
- Strong emphasis on data visualization.
- Syllabus is on D2L under “Content"

What We'll Learn

- Why do we visualize data?
- What makes a good (and bad!) data visualization
- How to make accurate and beautiful visualizations using R and ggplot
- How to clean and prepare data for visualization
- How to analyze data and show your results

Why learn to visualize data?

- Makes you a thoughtful consumer of data visualizations
- It's an extremely marketable skill
- You almost always want to start a data analysis with visualization, and you almost always want to communicate your analysis with visualizations.


$R^2 = 0.06$
REXTHOR, THE DOG-BEARER
I DON'T TRUST LINEAR REGRESSIONS WHEN IT'S HARDER TO GUESS THE DIRECTION OF THE CORRELATION FROM THE SCATTER PLOT THAN TO FIND NEW CONSTELLATIONS ON IT.

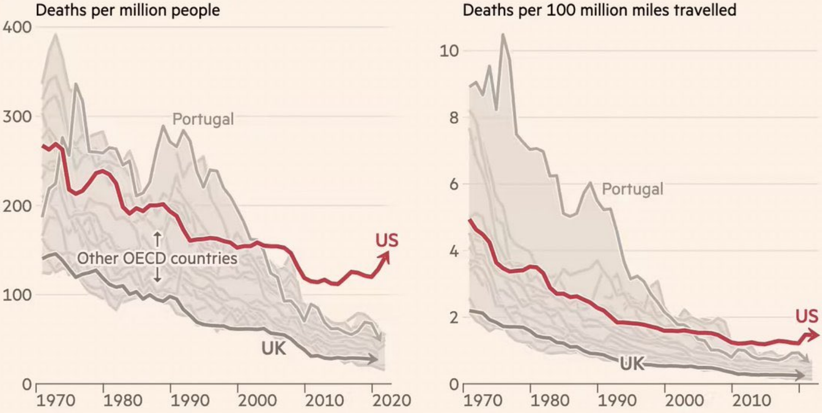
Why make visualizations using code?

Writing code is a pain! Why bother?

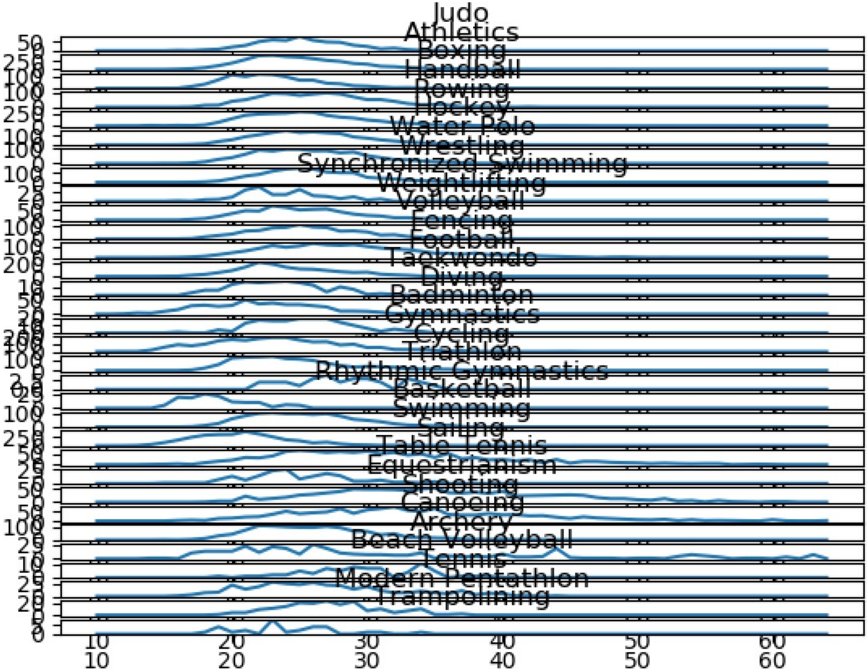
- We often need to analyze, clean, or reformat data first
- Reproducible and accurate
- Easy to customize

The US has much higher road death rates than other developed countries, regardless of how you slice the data

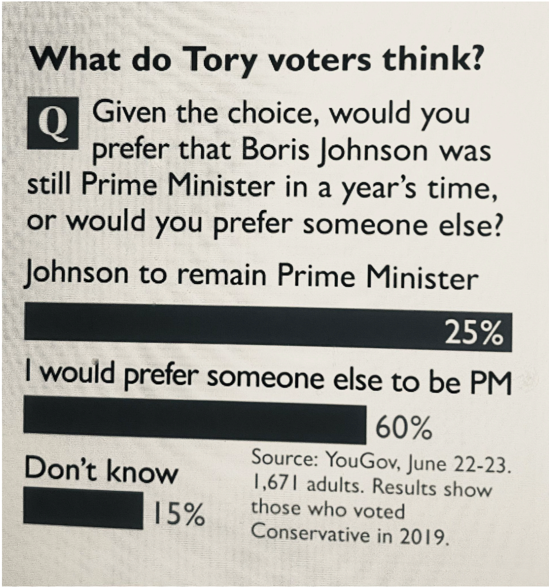
Different measures of road fatality rates, US vs other OECD countries



Sources: FT analysis of OECD; National Highway Traffic Safety Administration (NHTSA)
FT graphic: John Burn-Murdoch / @jburnmurdoch
© FT


Judo
Athletics
Boxing
Handball
Rowing
Hockey
Water Polo
Wrestling
Synchronized Swimming
Weightlifting
Volleyball
Fencing
Football
Taekwondo
Diving
Badminton
Gymnastics
Cycling
Triathlon
Rhythmic Gymnastics
Basketball
Swimming
Sailing
Table Tennis
Equestrianism
Shooting
Canoeing
Archery
Beach Volleyball
Tennis
Modern Pentathlon
Trampolining
50
100
250
200
25
5
0
10
10
20
30
40
50
60

Three kinds of lies...

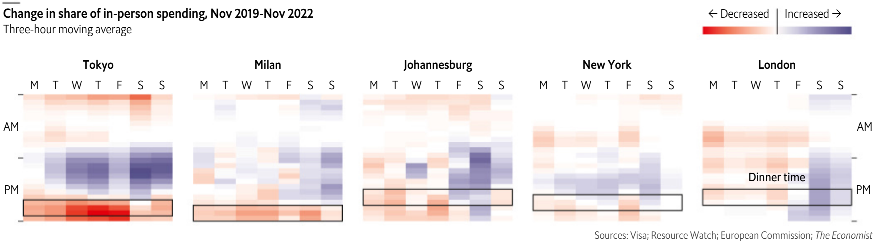


What do Tory voters think?

Q Given the choice, would you prefer that Boris Johnson was still Prime Minister in a year's time, or would you prefer someone else?

Source: YouGov, June 22-23. 1,671 adults. Results show those who voted Conservative in 2019.

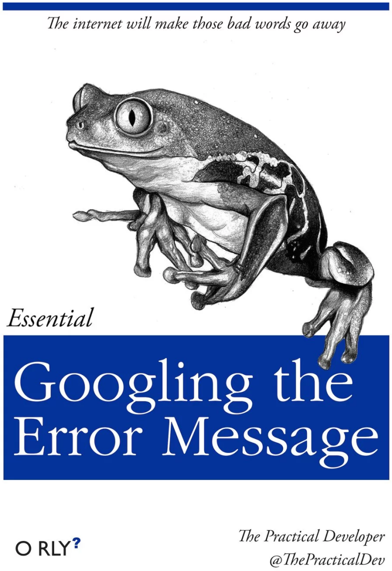
Restaurant Spending (Economist, July 1, 2023)

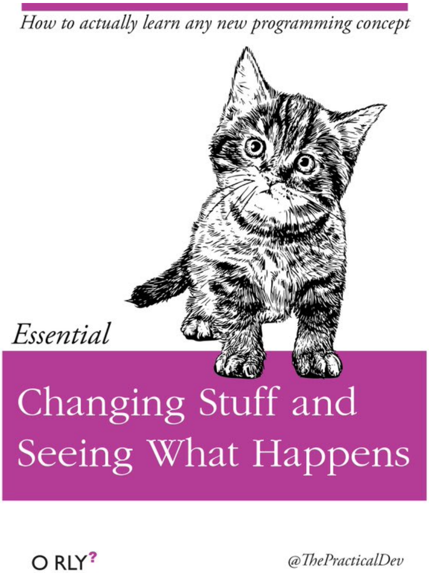


Structure of the class

- This is a hands-on class! Bring your computer.
- Short daily checkins on the reading (graded credit/no credit, with an option of extra credit for excellent responses).
- Short lecture, followed by practical exercises
- Mid-term and final projects–more on this in a bit
- Attendance is crucial
- Willingness to work hard on coding

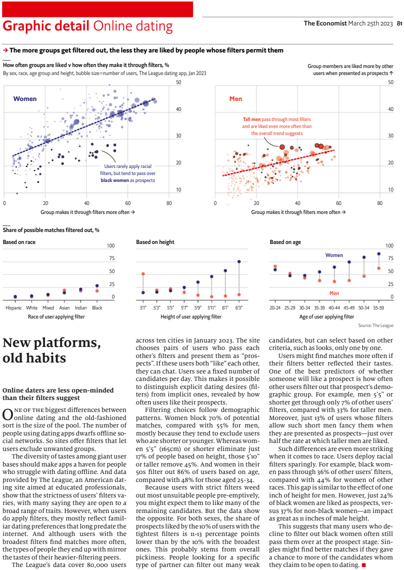
Programming is hard–remember these two weird tricks


The internet will make those bad words go away
Essential
Googling the Error Message
O RLY?
The Practical Developer
@ThePracticalDev


How to actually learn any new programming concept
Essential
Changing Stuff and Seeing What Happens
O RLY?
@ThePracticalDev

Projects

- The mid-term and final projects are **one page** data analyses.
- The mid-term project will use data I provide.
- The final project can use any data you want!



Graphic detail Online dating

The Economist March 25th 2023 81

→ The more groups get filtered out, the less they are liked by people whose filters permit them

How often groups are liked v how often they make it through filters, %

By sex, race, age group and height, bubble size=number of users, The League dating app, Jan 2023

Group members are liked more by other users when presented as prospects ↑
Women
Men
50
40
30
20
10
0
20
40
60
80
Group makes it through filters more often →
Users rarely apply racial filters, but tend to pass over black women as prospects
Tall men pass through most filters and are liked even more often than the overall trend suggests

Share of possible matches filtered out, %

Based on race
Based on height
Based on age
100
75
50
25
0
Women
Men
Hispanic
White
Mixed
Asian
Indian
Black
Race of user applying filter
5'1"
5'3"
5'5"
5'7"
5'9"
5'11"
6'1"
6'3"
Height of user applying filter
20-24
25-29
30-34
35-39
40-44
45-49
50-54
55-59
Age of user applying filter

Source: The League

New platforms, old habits

Online daters are less open-minded than their filters suggest

ONE OF THE biggest differences between online dating and the old-fashioned sort is the size of the pool. The number of people using dating apps dwarfs offline social networks. So sites offer filters that let users exclude unwanted groups.

The diversity of tastes among giant user bases should make apps a haven for people who struggle with dating offline. And data provided by The League, an American dating site aimed at educated professionals, show that the strictness of users' filters varies, with many saying they are open to a broad range of traits. However, when users do apply filters, they mostly reflect familiar dating preferences that long predate the internet. And although users with the broadest filters find matches more often, the types of people they end up with mirror the tastes of their heavier-filtering peers.

The League's data cover 80,000 users across ten cities in January 2023. The site chooses pairs of users who pass each other's filters and present them as "prospects". If these users both "like" each other, they can chat. Users see a fixed number of candidates per day. This makes it possible to distinguish explicit dating desires (filters) from implicit ones, revealed by how often users like their prospects.

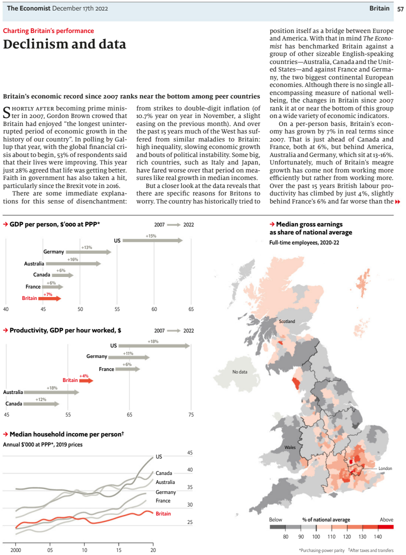
Filtering choices follow demographic patterns. Women block 70% of potential matches, compared with 55% for men, mostly because they tend to exclude users who are shorter or younger. Whereas women 5'5" (165cm) or shorter eliminate just 17% of people based on height, those 5'10" or taller remove 45%. And women in their 50s filter out 86% of users based on age, compared with 48% for those aged 25-34.

Because users with strict filters weed out most unsuitable people pre-emptively, you might expect them to like many of the remaining candidates. But the data show the opposite. For both sexes, the share of prospects liked by the 10% of users with the tightest filters is 11-13 percentage points lower than by the 10% with the broadest ones. This probably stems from overall pickiness. People looking for a specific type of partner can filter out many weak candidates, but can select based on other criteria, such as looks, only one by one.

Users might find matches more often if their filters better reflected their tastes. One of the best predictors of whether someone will like a prospect is how often other users filter out that prospect's demographic group. For example, men 5'5" or shorter get through only 7% of other users' filters, compared with 33% for taller men. Moreover, just 13% of users whose filters allow such short men fancy them when they are presented as prospects—just over half the rate at which taller men are liked.

Such differences are even more striking when it comes to race. Users deploy racial filters sparingly. For example, black women pass through 36% of other users' filters, compared with 44% for women of other races. This gap is similar to the effect of one inch of height for men. However, just 24% of black women are liked as prospects, versus 37% for non-black women—an impact as great as 11 inches of male height.

This suggests that many users who decline to filter out black women often still pass them over at the prospect stage. Singles might find better matches if they gave a chance to more of the candidates whom they claim to be open to dating. ■



The Economist December 17th 2022 Britain 57

Charting Britain's performance

Declinism and data

Britain's economic record since 2007 ranks near the bottom among peer countries

SHORTLY AFTER becoming prime minister in 2007, Gordon Brown crowed that Britain had enjoyed "the longest uninterrupted period of economic growth in the history of our country". In polling by Gallup that year, with the global financial crisis about to begin, 53% of respondents said that their lives were improving. This year just 28% agreed that life was getting better. Faith in government has also taken a hit, particularly since the Brexit vote in 2016.

There are some immediate explanations for this sense of disenchantment: from strikes to double-digit inflation (of 10.7% year on year in November, a slight easing on the previous month). And over the past 15 years much of the West has suffered from similar maladies to Britain: high inequality, slowing economic growth and bouts of political instability. Some big, rich countries, such as Italy and Japan, have fared worse over that period on measures like real growth in median incomes.

But a closer look at the data reveals that there are specific reasons for Britons to worry. The country has historically tried to position itself as a bridge between Europe and America. With that in mind *The Economist* has benchmarked Britain against a group of other sizeable English-speaking countries—Australia, Canada and the United States—and against France and Germany, the two biggest continental European economies. Although there is no single all-encompassing measure of national well-being, the changes in Britain since 2007 rank it at or near the bottom of this group on a wide variety of economic indicators.

On a per-person basis, Britain's economy has grown by 7% in real terms since 2007. That is just ahead of Canada and France, both at 6%, but behind America, Australia and Germany, which sit at 13-16%. Unfortunately, much of Britain's meagre growth has come not from working more efficiently but rather from working more. Over the past 15 years British labour productivity has climbed by just 4%, slightly behind France's 6% and far worse than the ▸▸

→ GDP per person, $'000 at PPP*

→ Productivity, GDP per hour worked, $

→ Median household income per person†

Annual $'000 at PPP*, 2019 prices

→ Median gross earnings as share of national average

Full-time employees, 2020-22

*Purchasing-power parity †After taxes and transfers

Deadlines

- In-class exercises are due the following day at 9am.
- Two “free passes” that you can use to skip in-class assignments, no questions asked.
- The mid-term project is due on **Wednesday, November 1 at 10:20am**.
- The final project is due on **Monday, December 11 at 5pm**.
- Extensions: only in exceptional circumstances. See syllabus.

ChatGPT

ChatGPT, Github Codex, etc, are extremely useful tools.

When are they useful? when they generate something that's difficult to write but easy to verify.

That makes them perfect for helping with data visualization.

Their use is encouraged in this class to help with writing code. However, the text you write should be your own.

First assignment–running an RMarkdown document

On D2L, browse to "Content" and click on Lecture 1. There’s an Rmd file that you should download.

Advice on installing LaTeX:
https://bookdown.org/yihui/rmarkdown-cookbook/install-latex.html

Other notes

No class on Wednesday

I'll be away at a conference. Instead, make sure that you can complete the R refresher assignment (due Monday before class).