

# Wstęp do Eksploracji Danych

Politechnika Warszawska  
Anna Kozak

# Anna Kozak



anna.kozak@pw.edu.pl

MS Teams



@kozaka93

# Hubert Ruczyński



hubert.ruczynski.stud@pw.edu.pl

MS Teams



@HubertR21

# Bartłomiej Sobieski



bartlomiej.sobieski.stud@pw.edu.pl

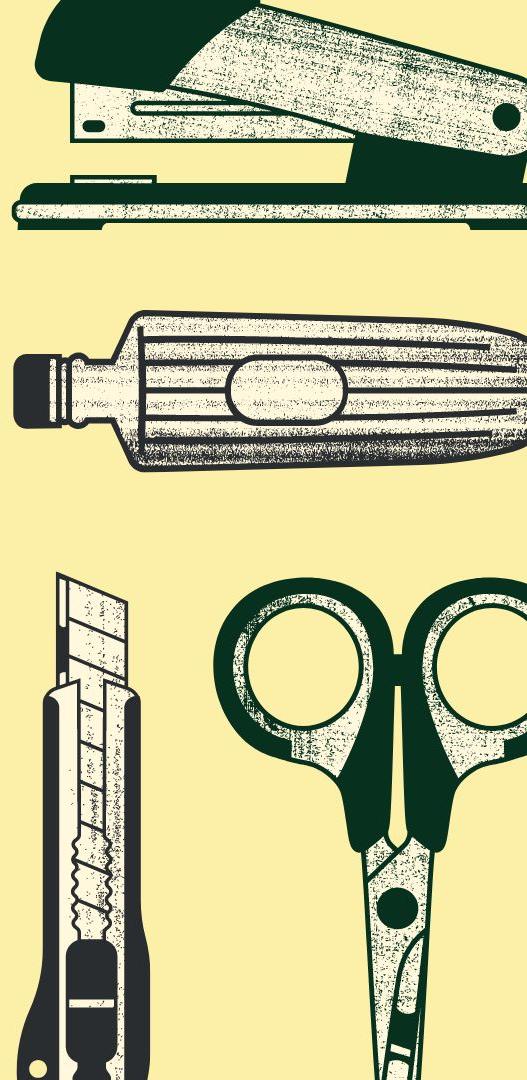
MS Teams



@sobieskibj

# Strona przedmiotu

[https://github.com/MI2-Education/  
2023L-ExploratoryDataAnalysis](https://github.com/MI2-Education/2023L-ExploratoryDataAnalysis)



# Wykład

Na wykładzie będą przedstawione zarówno teoretyczne aspekty pracy z danymi, jak i praktyczne.

15 wykładów =  $13 \times (\text{wykład/projekt}) + 2 \times \text{prezentacje projektów}$

# Projekty

- 2 projekty w ciągu semestru
  - zespoły 3 osobowe, różne podczas 1 i 2 projektu
  - projekt trwa 7-8 tygodni
  - 24p (P1) i 20p (P2) za projekt
- \*(w tym do 5p za pracę na zajęciach projektowych)

# Laboratorium

- praca w R i Python
- powtórzenie operacji na danych (R: dplyr, tidyr; Python: pandas)
- wstęp do narzędzi pozwalających na estetyczne prezentowanie danych
- różne sposoby oceny zmiennych, danych, wizualizacji
- 6 x praca domowa (4 x 5p + 2 x 10p)
- 3 x wejściówka (3 x 2p)

# Ocena końcowa

Suma punktów z prac domowych i projektów:

$$4 \times 5 + 2 \times 10 + 24 + 20 + 3 \times 2 = 90$$

$$(PD) + (PD) + (P1) + (P2) + (W) = (O)$$

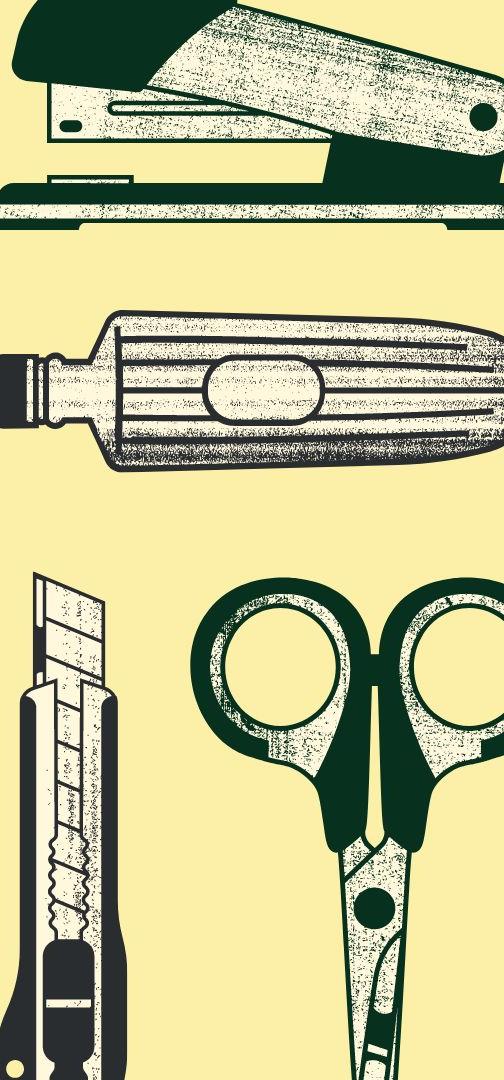
Aby zaliczyć kurs należy uzyskać ponad 45 punktów,  
w tym co najmniej 50% punktów z każdego z projektów.

Zajęcia laboratoryjne są obowiązkowe, w ciągu semestru dopuszczalne  
są co najwyżej dwie nieusprawiedliwione nieobecności.

Oceny będą wystawiane zgodnie z tabelą:

Ocena	3	3.5	4	4.5	5 <sub>∞</sub>
Punkty	(45, 54]	(54, 63]	(63, 72]	(72, 81]	(81, )

# Pytania?



# Eksploracja danych

# Dane

Mogą być generowane przez:

- ?

# Dane

Mogą być generowane przez:

- banki,
- ubezpieczenia,
- portale społecznościowe,
- firmy telekomunikacyjne,
- szpitale,
- dane eksperymentalne,
- tekst,
- mapy,
- sklepy internetowe,
- ...

# Eksploracja danych - czym jest?

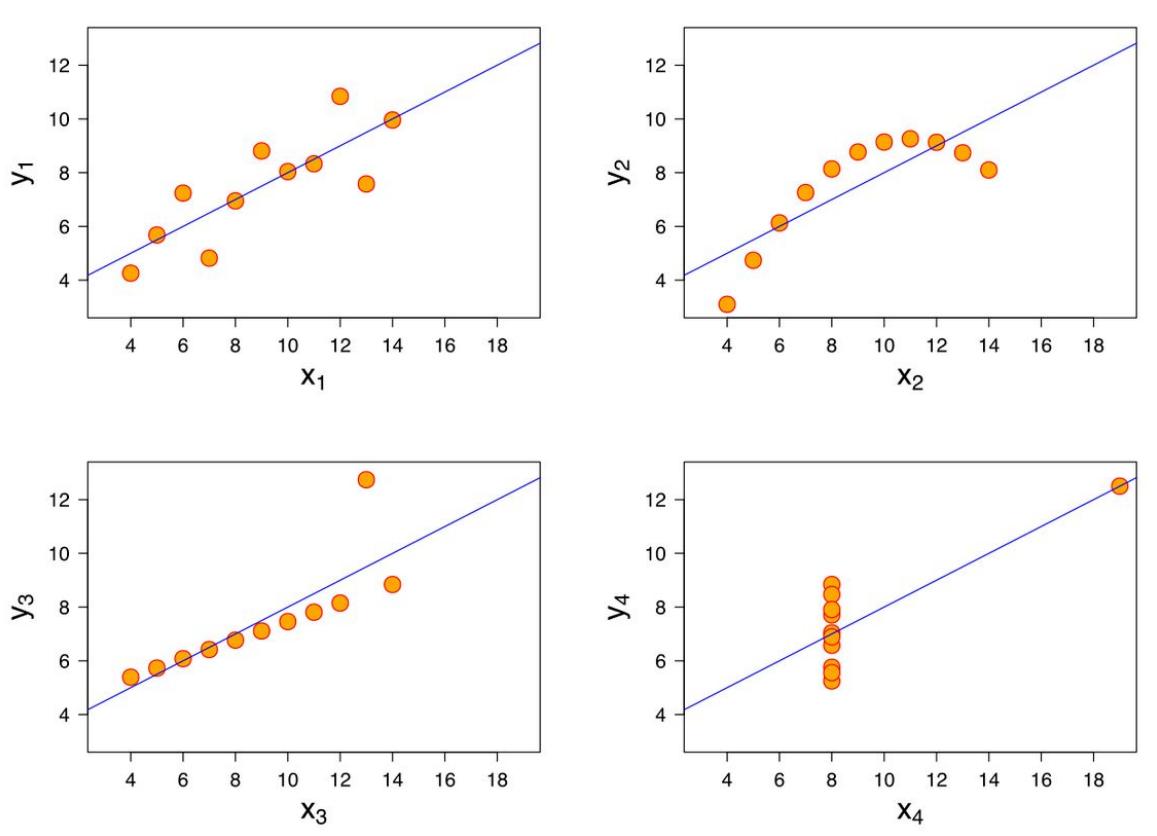
“proces odkrywania nietrywialnych, dotychczas nieznanych, potencjalnie użytecznych reguł, zależności, trendów”

Cel: analiza danych w celu lepszego ich zrozumienia

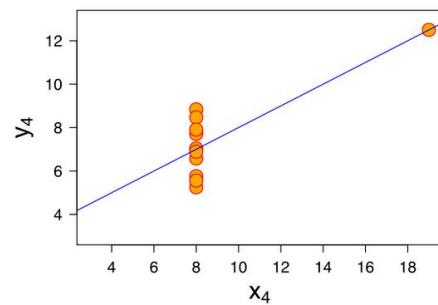
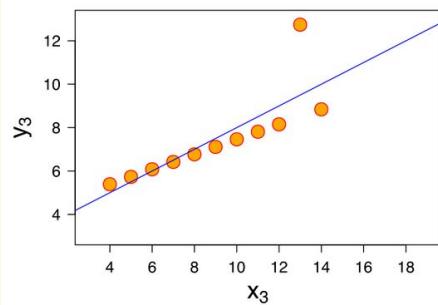
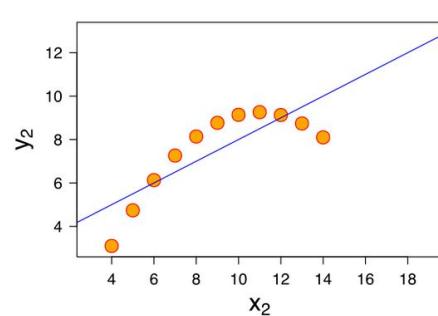
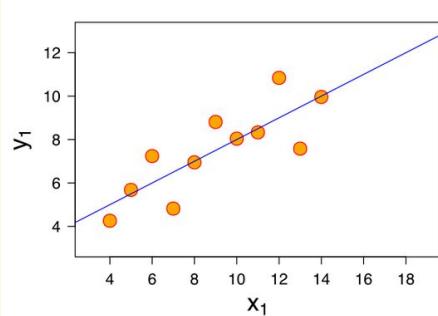
# Eksploracja danych - czym jest?

Na eksplorację danych składa się wiele dyscyplin, między innymi:

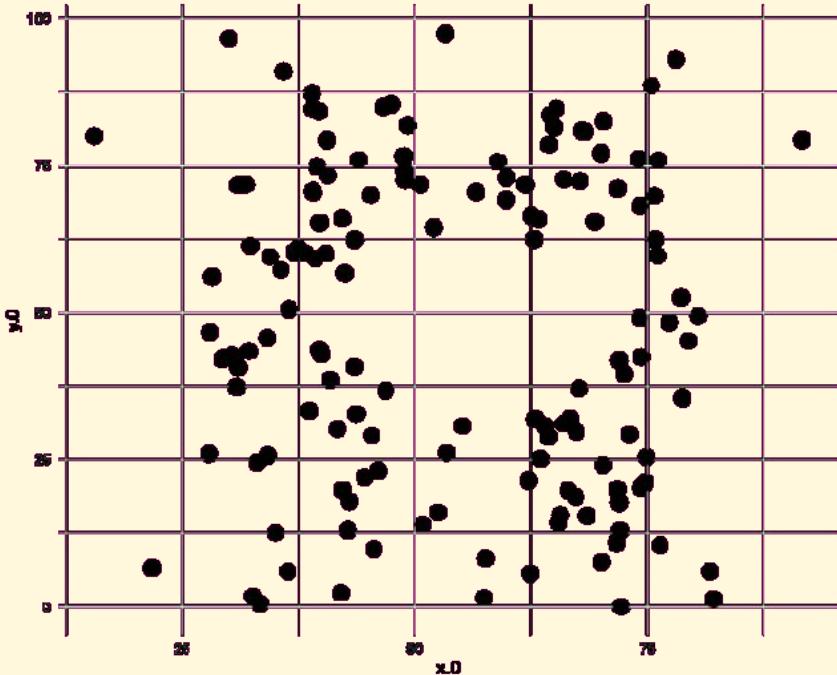
- bazy danych
- statystyka
- uczenie maszynowe
- wizualizacja danych
- wyszukiwanie informacji



## Kwartet Anscombe'a



Cecha	Wartość
Średnia arytmetyczna zmiennej $x$	9
Wariancja zmiennej $x$	11
Średnia arytmetyczna zmiennej $y$	7.50 (identyczna do dwóch cyfr po przecinku)
Wariancja zmiennej $y$	4.122 lub 4.127 (identyczna do trzech cyfr po przecinku)
Współczynnik korelacji pomiędzy zmiennymi	0.816 (identyczny do trzech cyfr po przecinku)



## The Datasaurus Dozen

13 zestawów danych ma te same statystyki zbiorcze (średnia x/y, odchylenie standardowe x/y i korelacja Pearsona) z dokładnością do dwóch miejsc po przecinku, a jednocześnie drastycznie różni się wyglądem.

# Jak rozpoznać rodzaj zmiennej?

*“dane liczbowe to nie tylko liczby”*

# Typy danych

Zmienne jakościowe (nazywane również *wyliczeniowymi, czynnikowymi* lub *kategorycznymi*), to zmienne przyjmujące określona liczbę wartości (najczęściej nie liczbowych). Zmienne te można dalej podzielić na:

- *binarne* (nazywane również dwumianowymi, dychotomicznymi) np. płeć (poziomy: kobieta/mężczyzna),
- *nominalne* (nazywane również zmiennymi jakościowymi nieuporządkowanymi) np. marka samochodu,
- *uporządkowane*, np. wykształcenie (poziomy: podstawowe/średnie/wyższe), ocena z przedmiotu.

# Typy danych

Zmienne ilościowe, z których można dodatkowo wyróżnić:

- *zliczenia* (liczba wystąpień pewnego zjawiska, opisywana liczbą całkowitą), np. liczba lat nauki, liczba wypadków,
- *ilorazowe*, czyli zmienne mierzone w skali, w której można dzielić wartości (ilorazy mają sens). Np. długość w metrach (coś jest 2 razy dłuższe, 10 razy krótsze itp.),
- *przedziałowe* (nazywane też interwałowymi), mierzone w skali, w której można odejmować wartości (wyznaczać długość przedziału).

# Struktura zbioru danych

ID	PŁEĆ	ZAWÓD	WZROST	DATA URODZENIA
ID_23	K	INFORMATYK	158	1978-03-12
ID_45	K	PRAWNIK	178	1989-05-29
ID_46	M	MATEMATYK	183	1991-01-19
ID_89	M	INFORMATYK	167	1982-02-20
ID_101	K	LEKARZ	163	1973-02-23

# Narzędzia do wizualizacji danych

- programistyczne (R, Python, JavaScript)
- programy graficzne (Inkscape)
- programy dedykowane do wizualizacji danych (Tableau, Power BI)

# PROJEKTY

# Cel zajęć projektowych

- wykorzystanie i utrwalenie zdobytej wiedzy z wykładu oraz laboratoriów
- praktyczna praca z danymi
- ćwiczenie sposobu prezentacji wyników

# Zasady

- 2 projekty w ciągu semestru
- zespoły 3 osobowe, różne podczas 1 i 2 projektu
- projekt trwa 7-8 tygodni
- 24p (P1) i 20p (P2) za projekt (w tym do 5p za pracę na zajęciach projektowych)

# **PROJEKT 1**

Zadanie: Przygotowanie plakatu na zadany temat

Rezultat: Plakat w formacie A2 wydrukowany + sesja plakatowa podczas wykładu

Zajęcia:

- wspólnie dyskusje
- prezentacje kolejnych etapów

# Ocena

Za projekt można otrzymać od 0 do 24 punktów, z czego:

- 5p (1 x 1p, 2 x 2p) uzyskuje się za przedstawienie postępu prac w danym tygodniu
- 5p uzyskuje się za przygotowanie estetycznych wykresów (dwa lub więcej)
- 5p uzyskuje się, jeżeli przygotowane wykresy mają wszystkie niezbędne elementy do poprawnego odczytania danych (tytuł, podtytuł, adnotacje na osiach, legenda, jednostki, opis jak czytać wykres)
- 5p uzyskuje się za estetykę i pomysłowość aranżacji wykresów i opisów w jedną całość
- 4p uzyskuje się za prezentację projektu

## Za tydzień

- podział na zespoły 3 osobowe
- "burza mózgów"

**Temat projektu to...**

**MUZYKA**

# **Jakie prace zostały wykonane w poprzednich latach?**

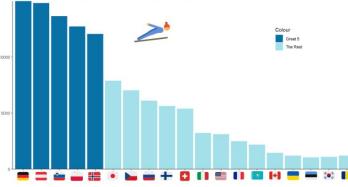
2022/2023 Sport

2021/2022 Plakaty, które zmieniają spojrzenie na klimat i środowisko

2020/2021, 2018/2019 Filmy, seriale, książki, audiobooki, gry

# See how fast they fly

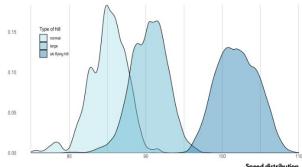
Number of jumps per country (trainings excluded)



## The Great 5

Ski jumping is still rather a niche sport - there is a huge difference between ski jumping popularity between Germany, Austria, Slovenia, Poland and Norway and the rest of the world. Because of that popularity, those countries have won 20 consecutive Nations Cups, and are therefore called "The Great 5". To get acquainted with this weird sport, let's have a look at one of the key aspects of a great, long ski jump: the take-off speed.

Distribution of take-off speed by type of hill



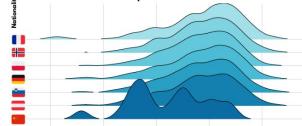
## The bigger the hill, the faster you go!

First thing we need to refer to when measuring take-off speeds, is of course the size of the hill. As it turns out, ski flying gives competitors as much as 10-20 km/h faster take-off than most commonly used, large hill. Needless to say, this extra ordinary velocity results in extra ordinary distance - and possibly even world records!

## Speed by country

Due to different techniques and proficiency of their technical teams jumpers from different countries achieve different take-off speeds. On the chart we represented The Great 5 nations, China and France as they have interesting distributions.

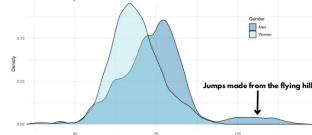
Distribution of take-off speed for countries



## Who goes faster: men or women?

As we see on the chart below: men athletes more often have high take-off speeds in comparison to women. This effect is mostly caused by the type of hills they compete in - as for now, there are no women's competitions on flying hills.

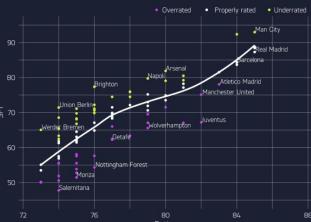
Take-off speed: men vs women



# FIFA VS REALITY

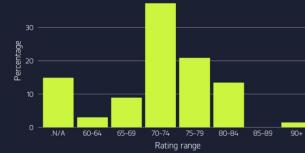
With the release of every single FIFA people argue about the ratings of players and call the creators biased. But is FIFA really that bad in terms of ratings and statistics?

## FIFA TEAM RATINGS VS SPI

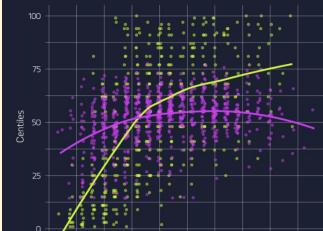


The first graph shows the dependence of the teams' overall rating in FIFA 23 on the soccer power index as of the start of the 22/23 season. SPI determines how good a team is based on match statistics from the last few months. The trend is that the higher the SPI, the higher the rating. However, at the extreme values of the rating, the trend line is a bit more vertical, which indicates that FIFA tries not to create extremely strong and extremely weak teams. It can be also observed that the greatest discrepancies between the ratings and the SPI are in case of the weaker teams.

## CURRENT RATINGS OF PERSPECTIVE PLAYERS



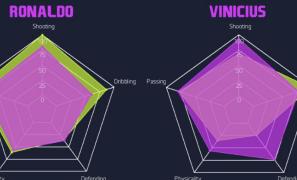
## DEPENDENCE OF FORWARDS STATISTICS ON AGE



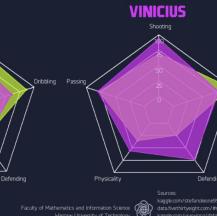
Players in FIFA are rated also on factors unrelated to their skills such as age of a player. The graph on the left compares the strikers' statistics in reality and in FIFA. According to the trend lines of these statistics, older players, despite a significant decrease in technical skills and physical ability, do not lose their statistics in FIFA. Similarly, young players whose form has exploded receive inadequate and underestimated statistics.

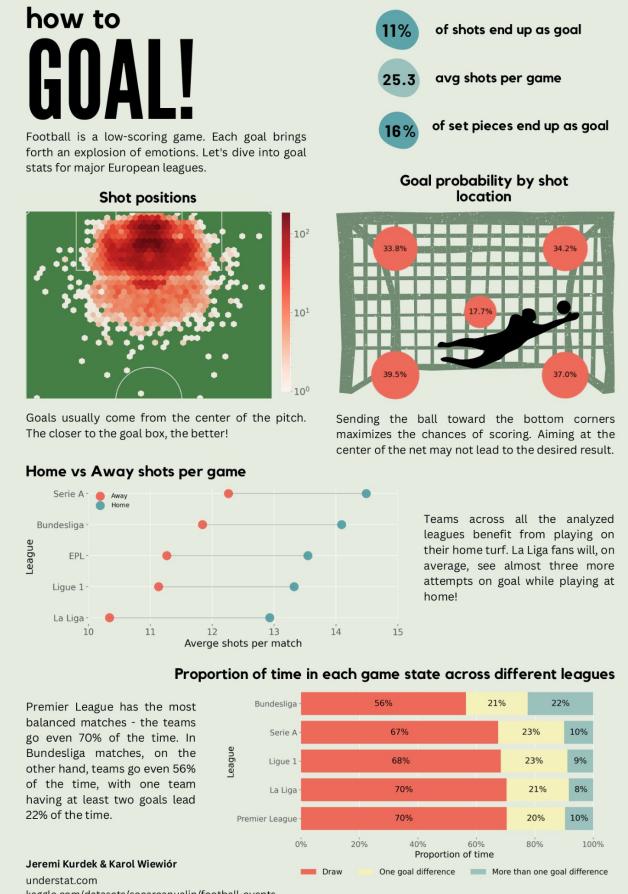
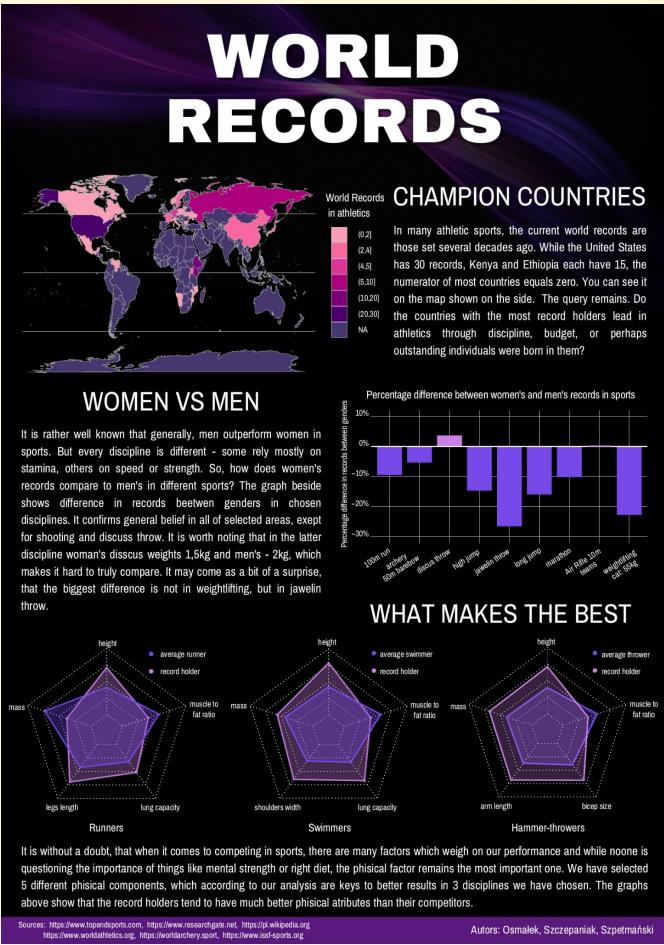
Great examples are Vinicius Jr (born in 2000) and Cristiano Ronaldo (born in 1985). Below you will find a comparison of their statistics in percentiles against other footballers playing in the same position.

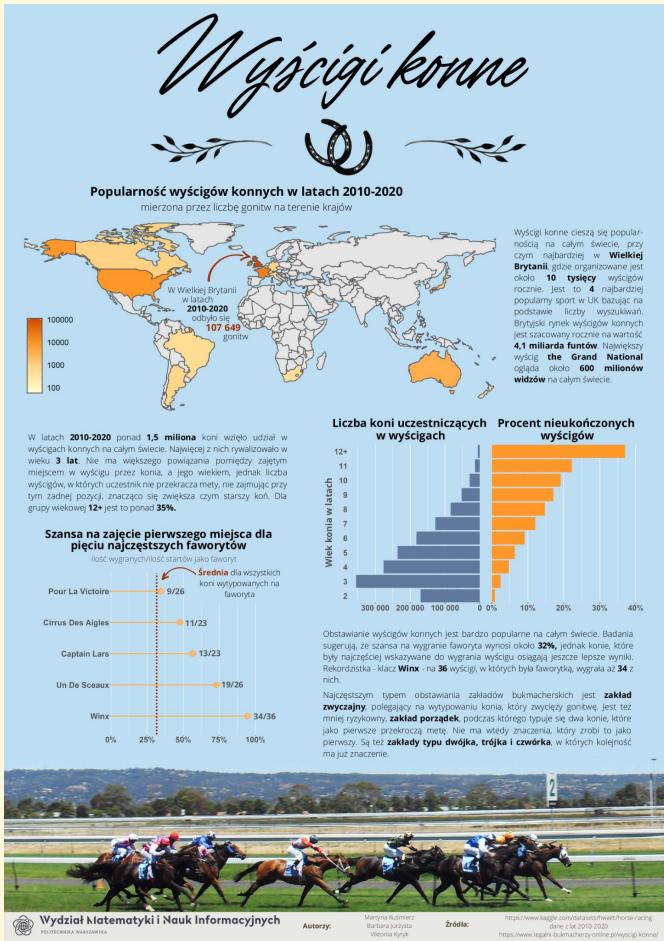
### RONALDO

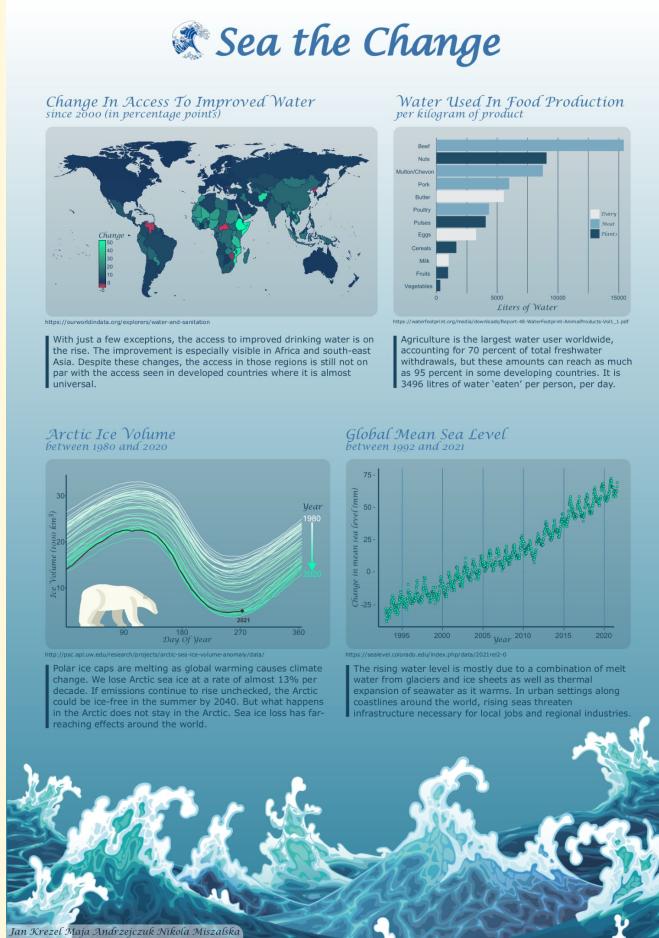
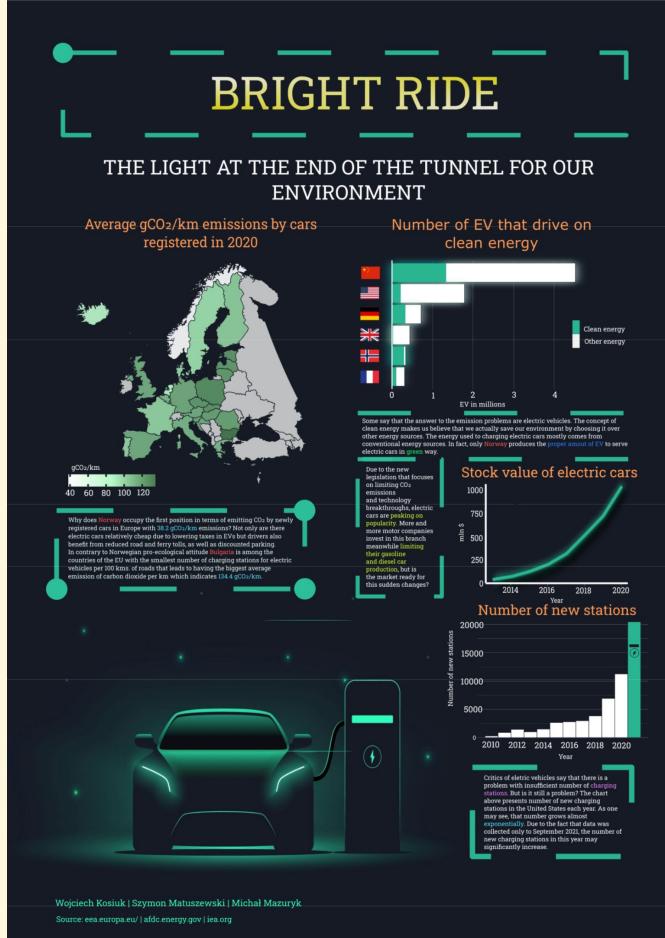


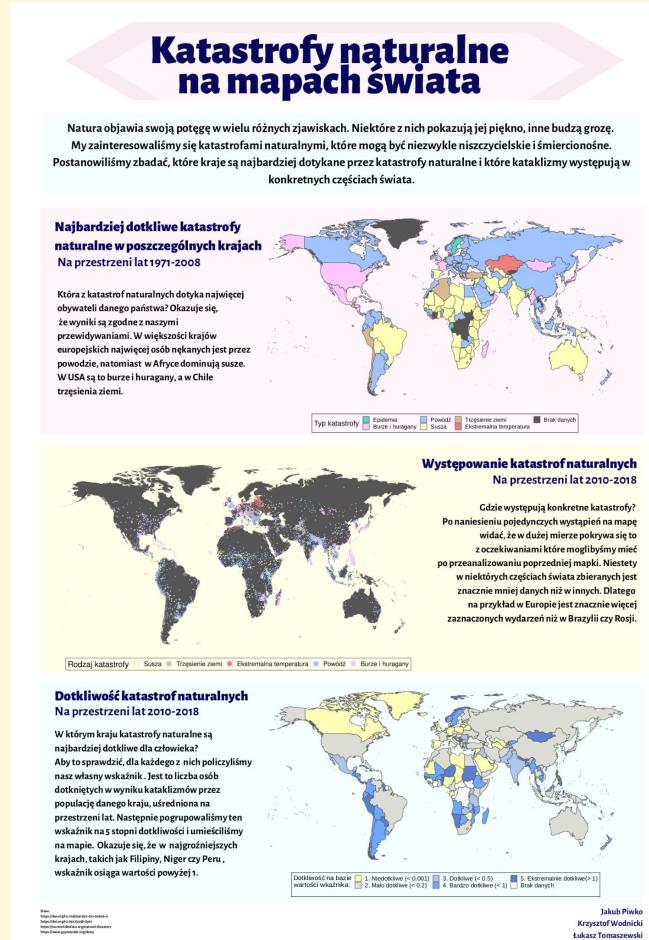
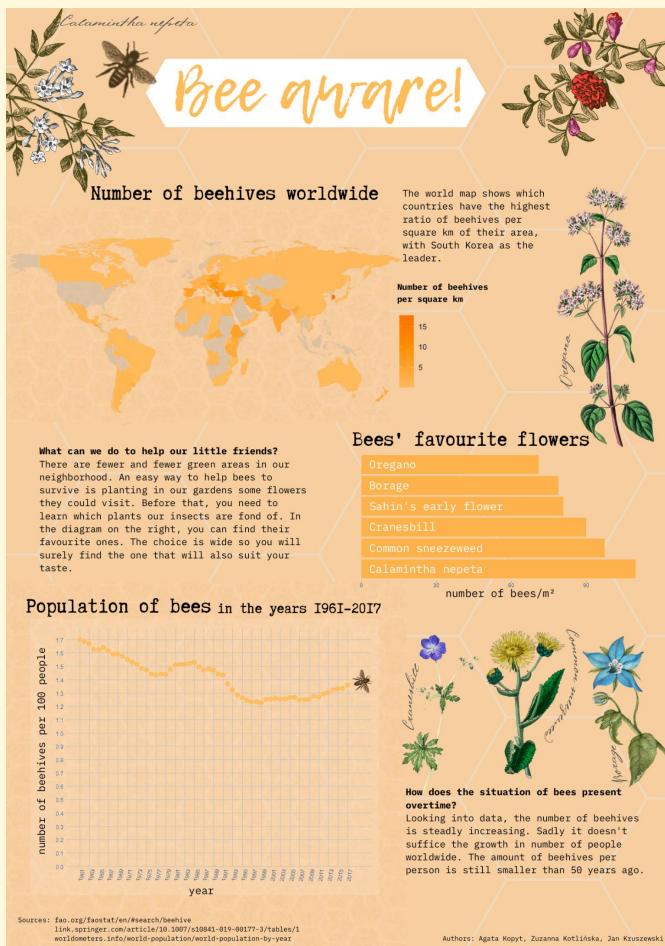
### VINICIUS







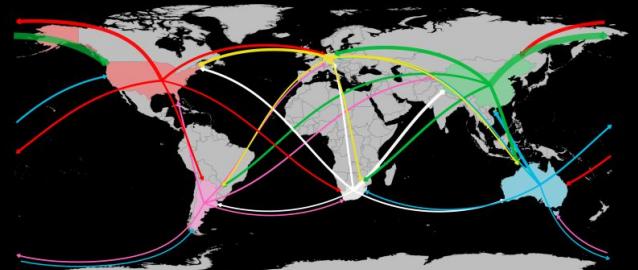




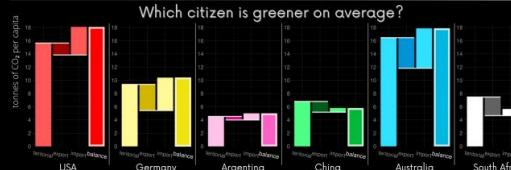
<https://medium.com/responsibleml/posters-that-change-the-perspective-on-climate-and-the-environment-c3682c0f6c39>

# The hidden emissions

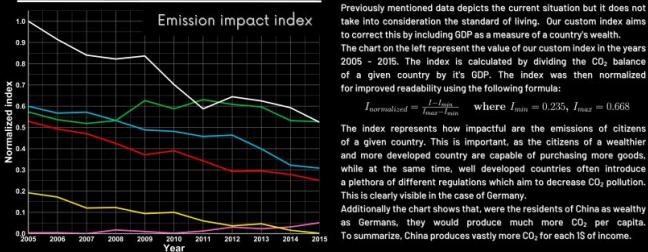
CO<sub>2</sub> in production and trade



Arrows coming out of a country have the same color as the country and they represent export from it to all others. The width of the arrow represents the amount of CO<sub>2</sub> exported in 2015. From this chart we can clearly see that some countries, like China, export much more than they import; while other countries, like Australia, do the opposite. Those exports and imports are usually omitted when calculating countries' emissions. This leads to significant changes in total emissions.



The other factor that usually is not taken into account is the number of people living in the country. Although the map would imply that China is the biggest polluter in the world, it does not necessarily reflect the whole truth. The chart above presents carbon dioxide values per capita in 2015. From the per capita perspective there seems to be a discrepancy between what general public reckons and what the numbers say. As one can observe, it would appear that USA CO<sub>2</sub> demand surpasses the Chinese by a whooping 10 tonnes disparity.



Faculty of Mathematics and Information Science  
Patryk Rakus Kacper Trębacz Małwina Wojewoda Data source: idata.org stats.oecd.org

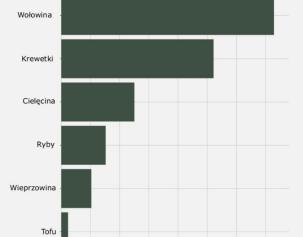
# Produkcja żywności a ekologia

Emisja CO<sub>2</sub> per capita w roku 2013  
w wyniku produkcji żywności



Badania sugerują, że 20% dwutlenku węgla, który wytwarzamy, pochodzi z produkcji żywności. Marnowanie jedzenia również przyczynia się do transmisji nadmiarowego dwutlenku węgla do atmosfery. Najbardziej odpowiedzialne za ten stan rzeczy są kraje rozwinięte.

Emisja CO<sub>2</sub> w wyniku produkcji żywności w przeliczeniu na 1000 kcal

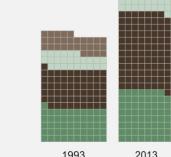


Spożycie wybranych produktów w 2013 roku  
W kilogramach per capita



David Pludowski  
Antoni Zajko  
Grzegorz Kiersnowski  
źródła: FAOSTAT, Kaggle, OWID

CO<sub>2</sub>[t]



Produkcja i marnowanie żywności  
W latach 1993 i 2013

Typy żywności

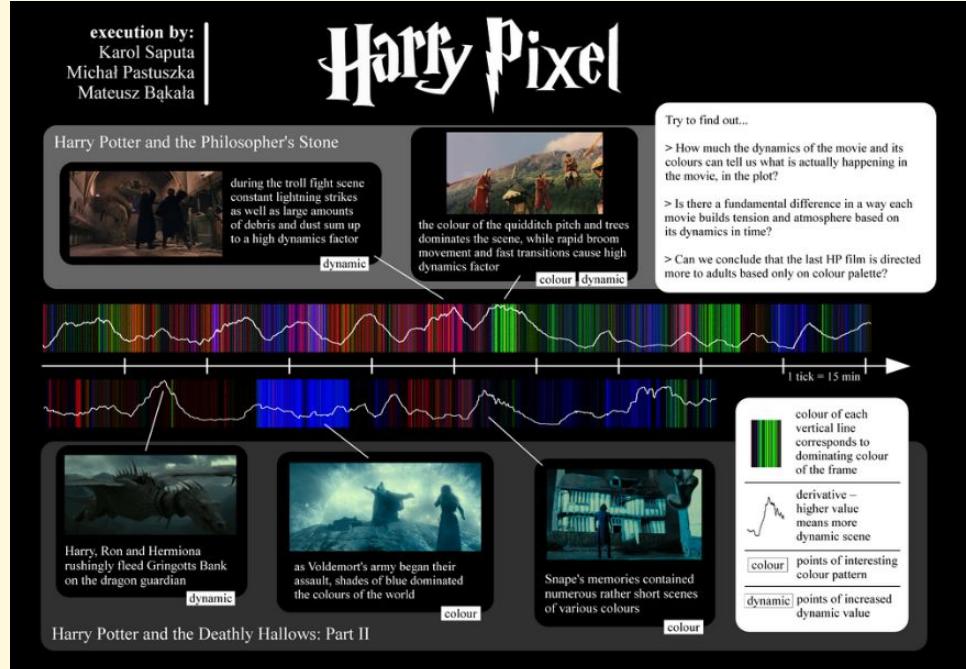
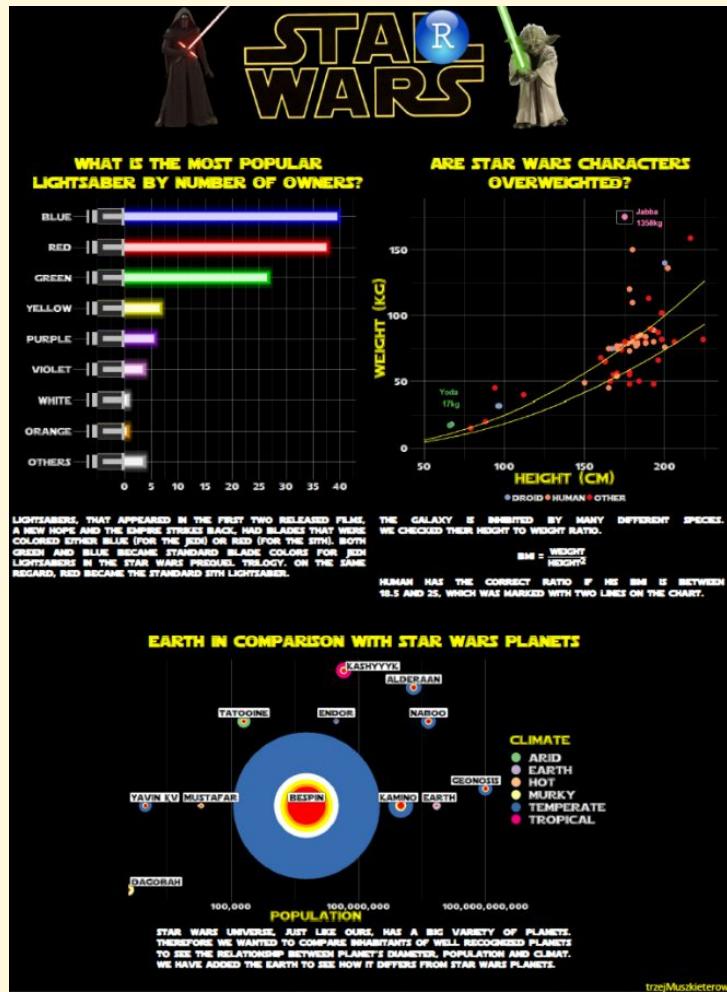
- Mięso i nabiał zmarnowane
- Warzywa i owoce zmarnowane
- Mięso i nabiał
- Warzywa i owoce

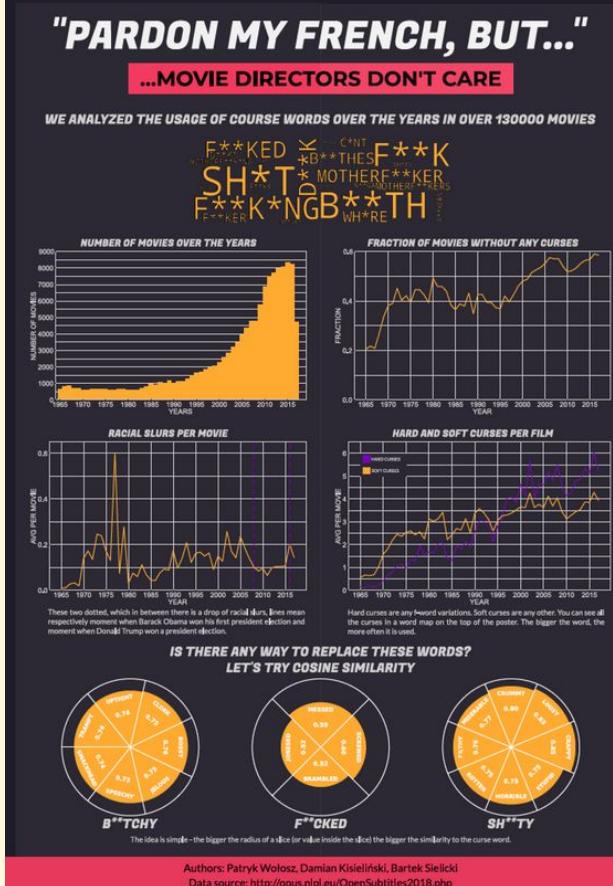
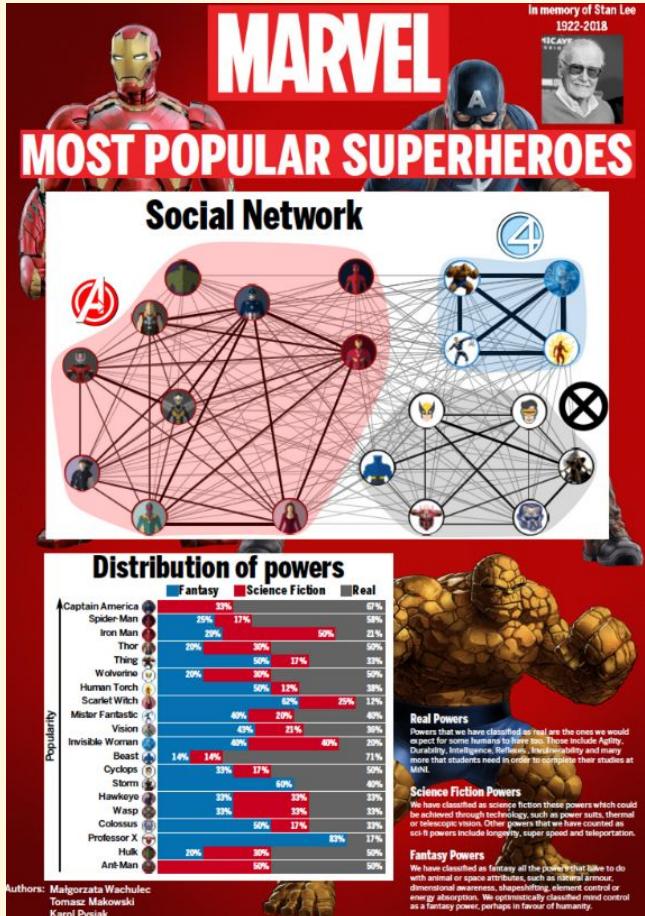
~10min ton

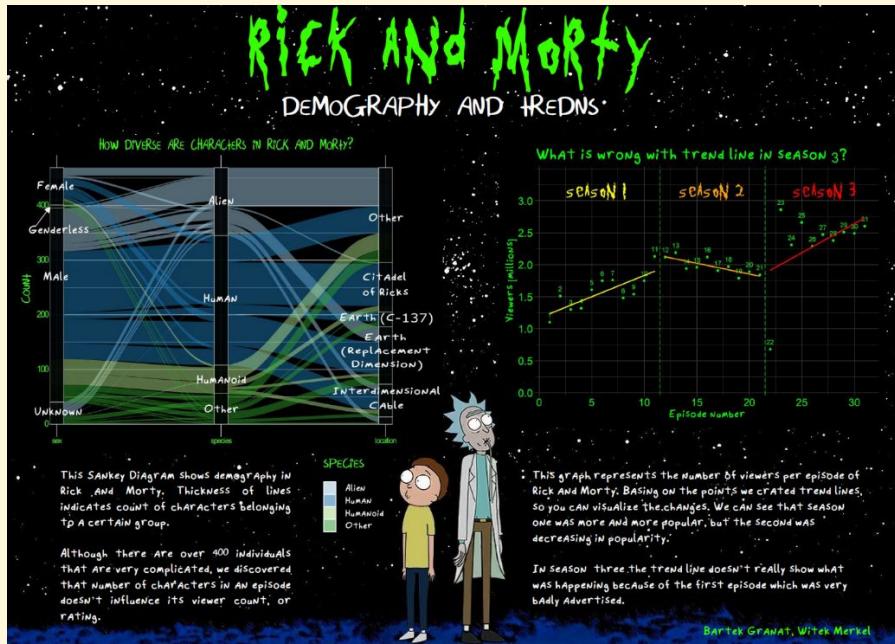
1993

2013

<https://medium.com/responsibleml/posters-that-change-the-perspective-on-climate-and-the-environment-c3682c0f6c39>



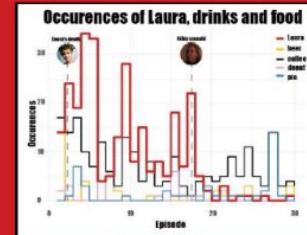




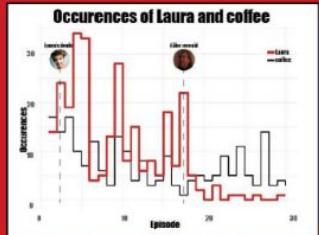
# Did coffee help to solve the murder case?

The main theme of series called Twin Peaks is murder of Laura Palmer. Investigation is lead by FBI agent Dale Cooper, who is a huge fan of coffee and cherry pie. Actually everyone who lives in Twin Peaks is a huge fan of coffee and cherry pie. Let's find out if their favourite food and drinks helped them with finding the murderer.

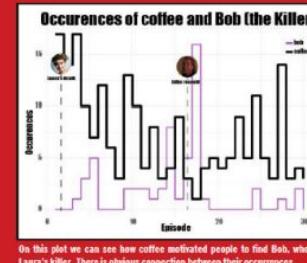
The most important thing on charts below are 2 marked moments: Laura's death and killer revelation. Apart from that, we can observe occurrences of specific words in subtitles per episode. There were 30 episodes in total in Twin Peaks series.



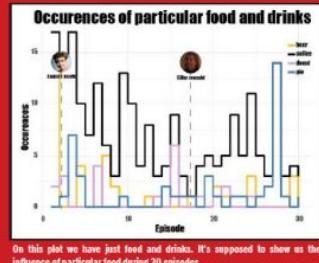
On this plot we can see all most popular food and drinks that were consumed during the story. We can observe the fact that coffee is the most consumed drink during the whole series. Which is pretty obvious - it's Twin Peaks favorite!



On this plot we can observe mentions of Laura and coffee. We can see that at the beginning of investigation, when everybody is excited about it, there are much more occurrences of Laura name and more coffee consumption.



On this plot we can see how coffee motivated people to find Bob, who was Laura's killer. There is obvious connection between their occurrences.



On this plot we have just food and drinks. It's supposed to show us the influence of particular food during 30 episodes.

