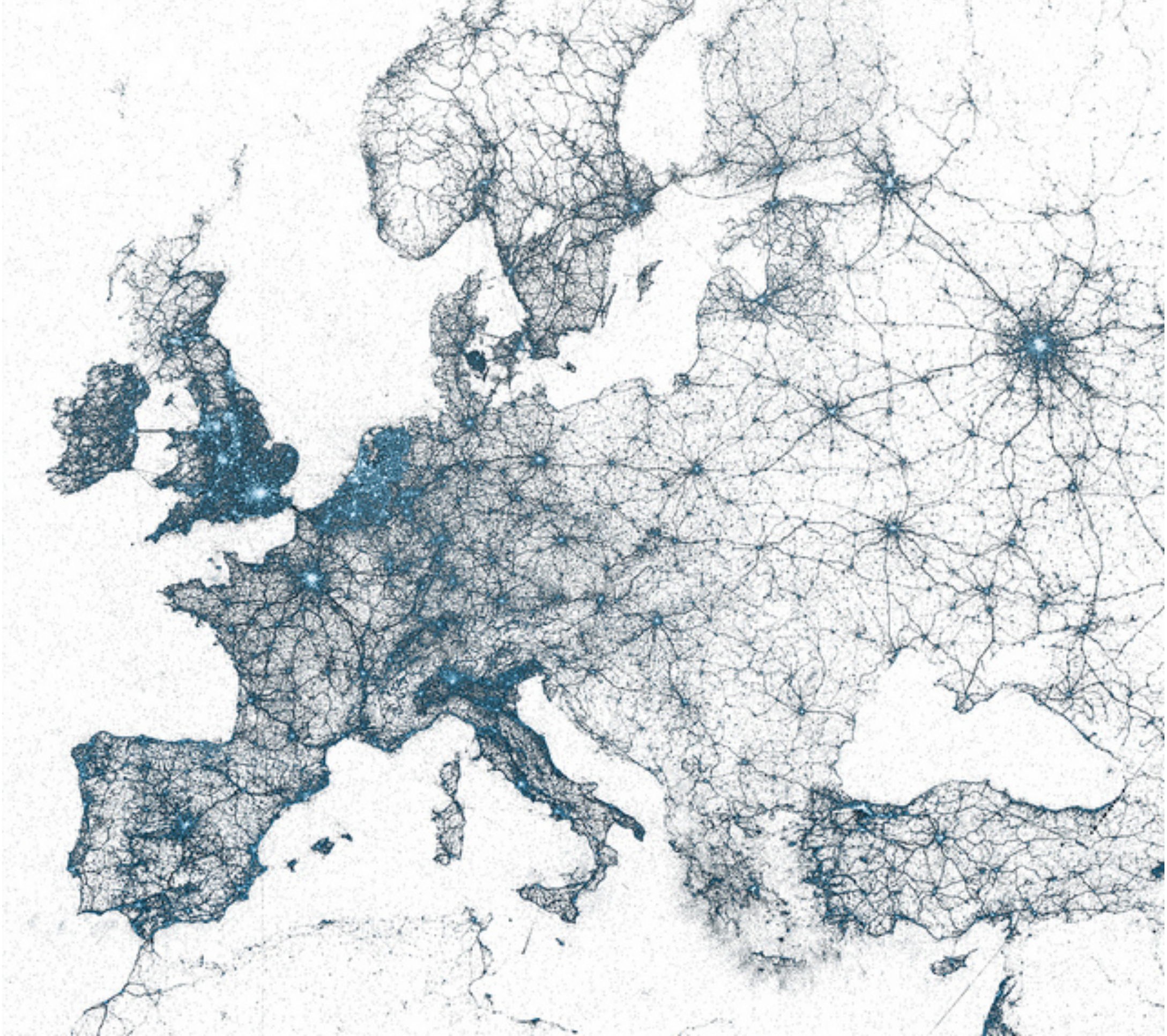
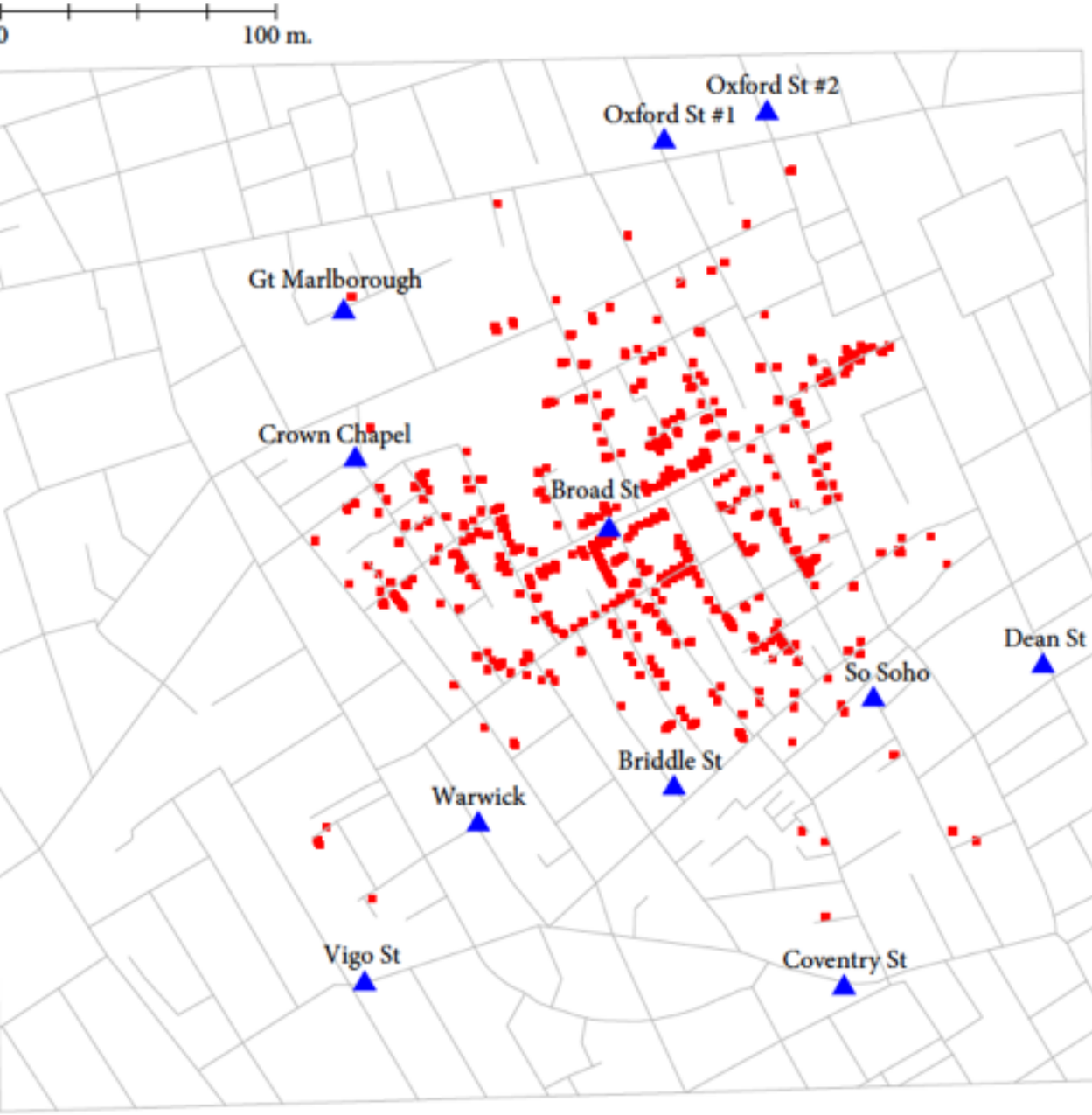
Wizualizacja danych.
Co robić, czego nie robić, dlaczego warto poznać pakiet ggplot2 dla R

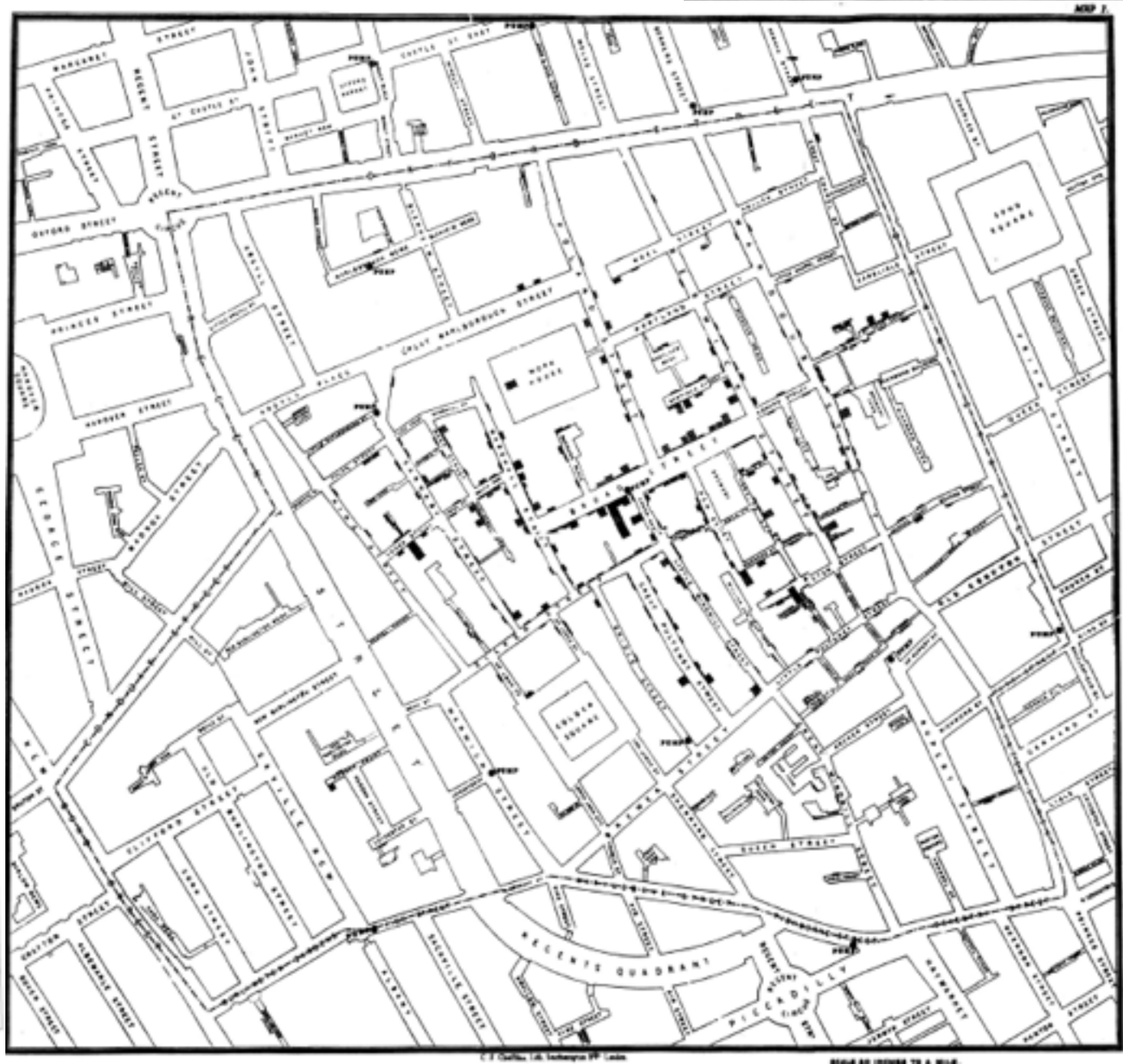
Przemysław Biecek
Uniwersytet Warszawski
Politechnika Warszawska

Czy wizualizacja danych jest przydatna?

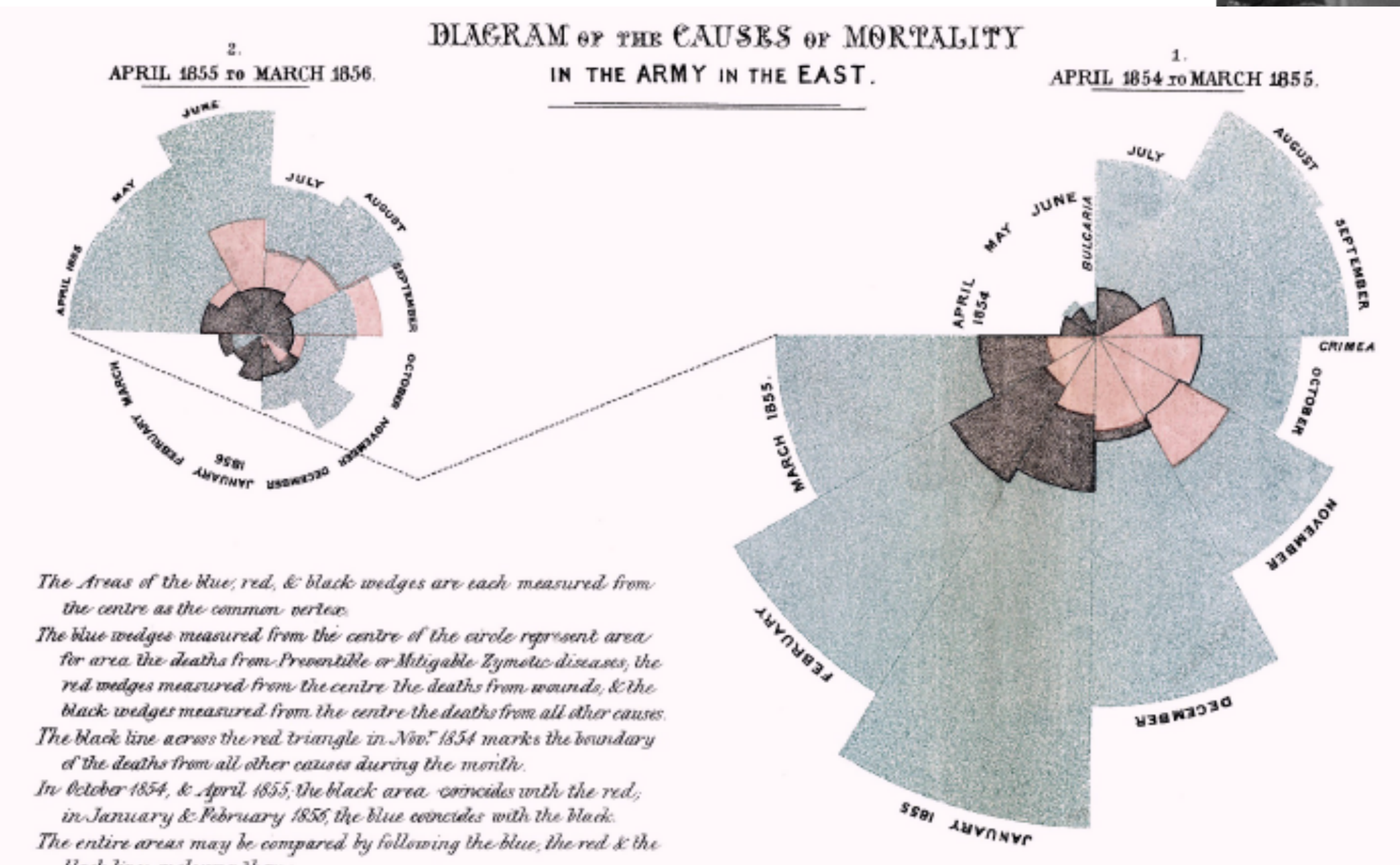


Kim jest John Snow?





Kim jest Florence Nightingale?



The Areas of the blue, red, & black wedges are each measured from the centre as the common vertex.

The blue wedges measured from the centre of the circle represent area for area the deaths from Preventible or Mitigable Zymotic diseases, the red wedges measured from the centre the deaths from wounds, & the black wedges measured from the centre the deaths from all other causes.

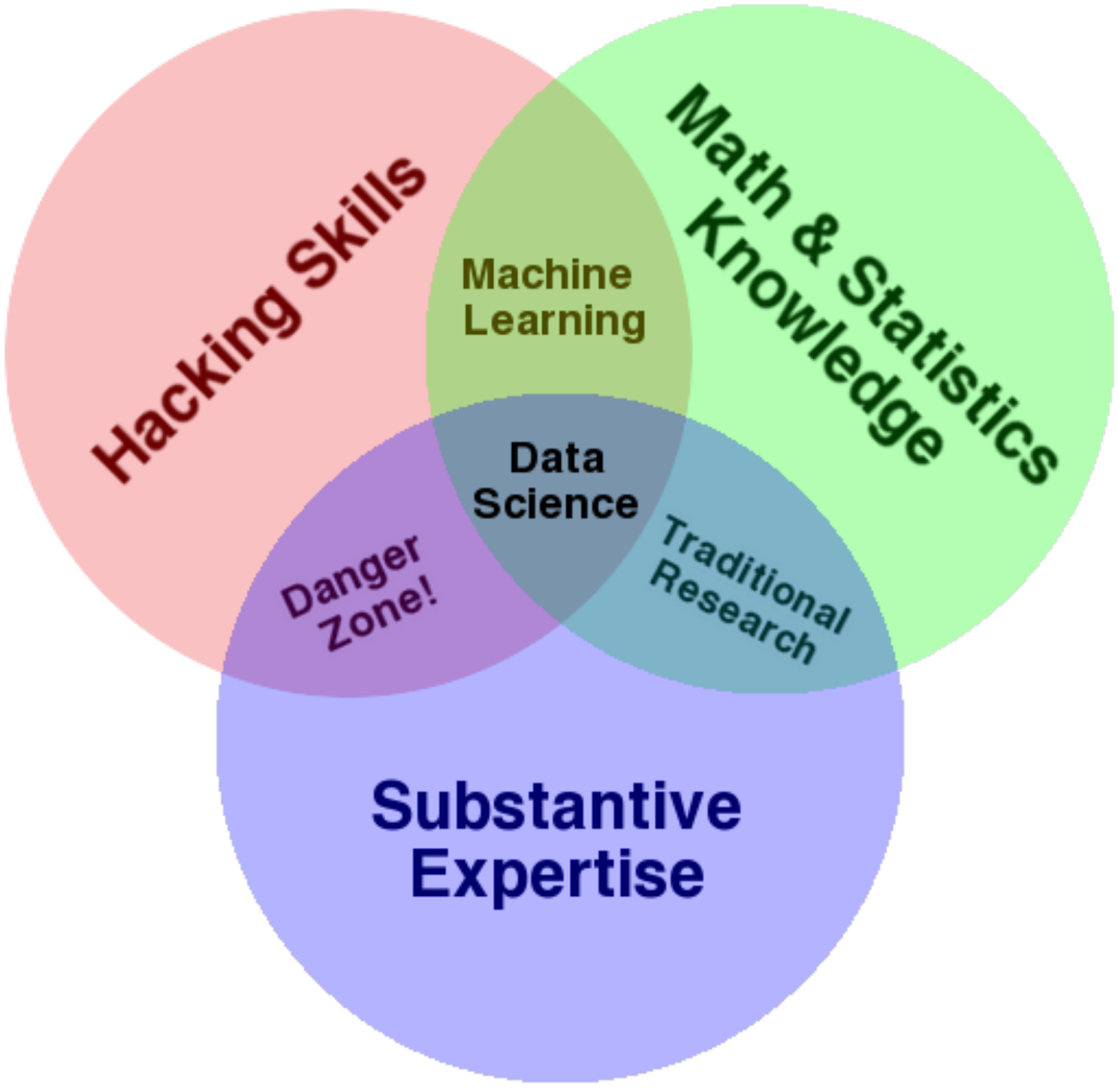
The black line across the red triangle in Nov.r 1854 marks the boundary of the deaths from all other causes during the month.

In October 1854, & April 1855, the black area coincides with the red; in January & February 1856, the blue coincides with the black.

The entire areas may be compared by following the blue, the red & the black lines enclosing them.

Ale czy to jest Data Science?

Zagadka 1:
Gdzie jest wizualizacja danych?



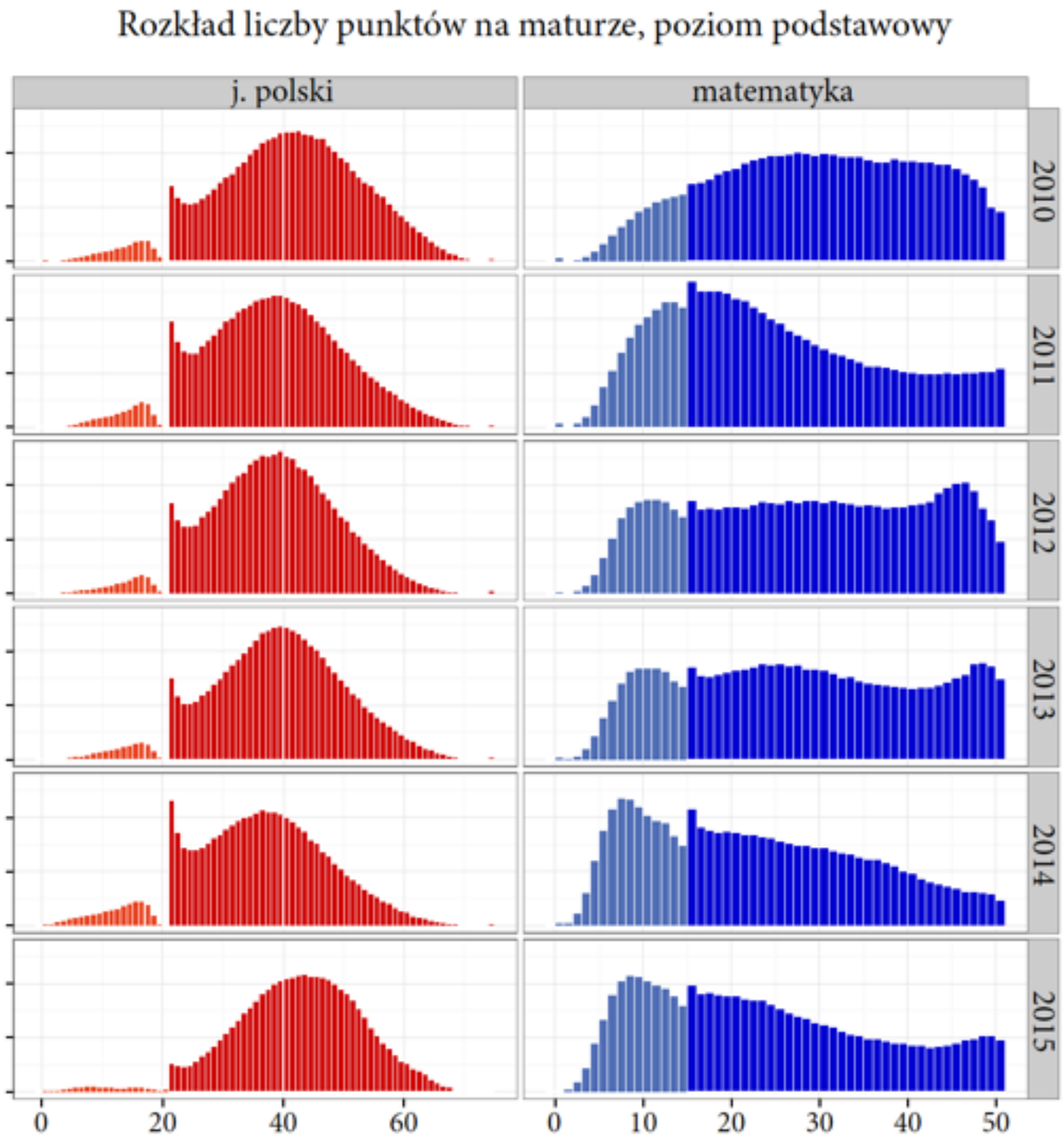
Zagadka 2:
Tabela? Wykres? Słownie?

Liczba ludności Polski	W tym kobiet	W tym mężczyzn
38 511 800	51,6%	48,4%

W wyniku przeprowadzenia Narodowego Spisu Powszechnego w roku 2011 ustalono, że w Polsce mieszka 38 511 800 osób, z czego 48,4% to mężczyźni, a 51,6% to kobiety.

Zagadka 2:
Tabela? Wykres? Słownie?

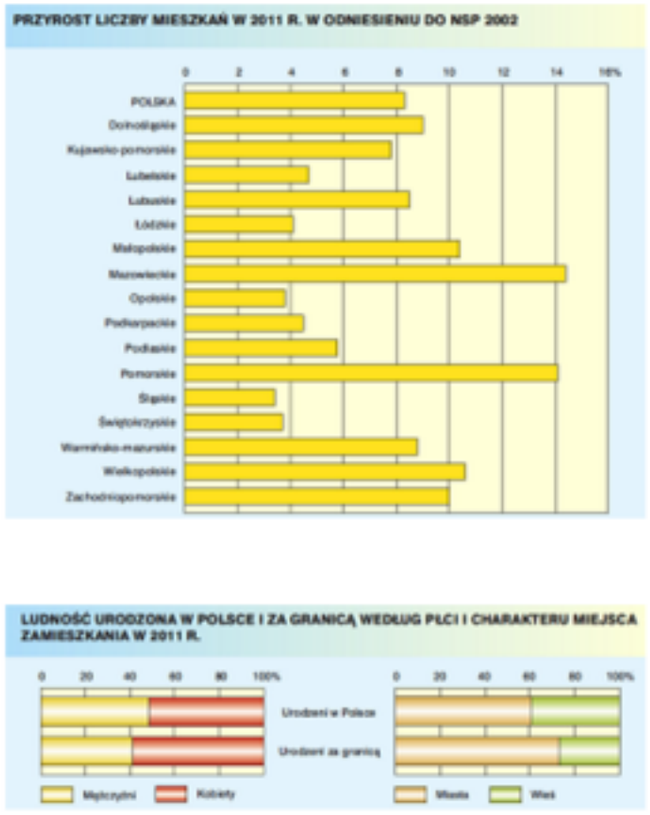
punkty	przedmiot	2010	2011	2012	2013	2014	2015
...	...	...	...	...	...	...	...
6	j. polski	0,09	0,09	0,09	0,09	0,25	0,16
7	j. polski	0,12	0,14	0,11	0,12	0,28	0,16
8	j. polski	0,16	0,18	0,12	0,14	0,34	0,19
9	j. polski	0,19	0,22	0,14	0,19	0,36	0,19
10	j. polski	0,23	0,27	0,18	0,21	0,40	0,17
11	j. polski	0,25	0,29	0,20	0,25	0,45	0,16
12	j. polski	0,28	0,31	0,23	0,28	0,47	0,15
13	j. polski	0,34	0,36	0,27	0,31	0,50	0,13
14	j. polski	0,37	0,41	0,32	0,37	0,61	0,13
15	j. polski	0,42	0,47	0,37	0,41	0,68	0,16
16	j. polski	0,49	0,57	0,45	0,45	0,73	0,16
17	j. polski	0,54	0,67	0,50	0,50	0,74	0,17
18	j. polski	0,54	0,62	0,46	0,44	0,63	0,14
19	j. polski	0,34	0,34	0,26	0,27	0,31	0,10
20	j. polski	0,13	0,09	0,09	0,09	0,07	0,06
21	j. polski	0,02	0,01	0,01	0,01	0,01	0,10
22	j. polski	1,90	2,72	2,43	2,28	3,76	0,90
23	j. polski	1,60	2,20	1,96	1,78	2,80	0,82
24	j. polski	1,46	1,95	1,80	1,56	2,36	0,81
25	j. polski	1,44	1,91	1,80	1,59	2,28	0,85
...	...	...	...	...	...	...	...

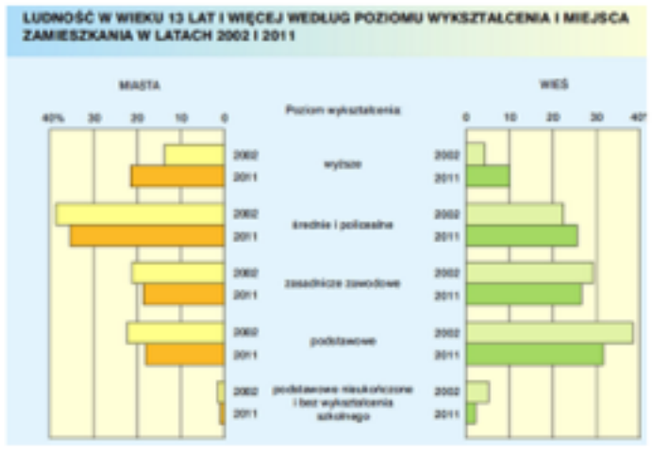


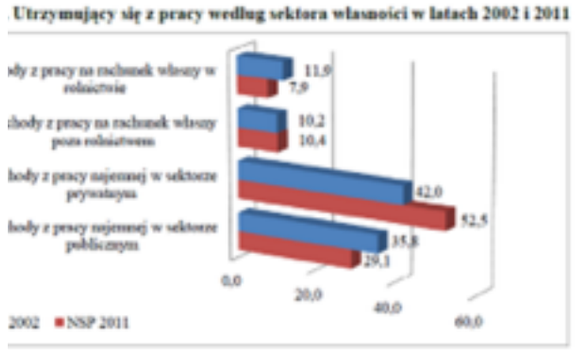
Czy wizualizacja danych jest prosta?

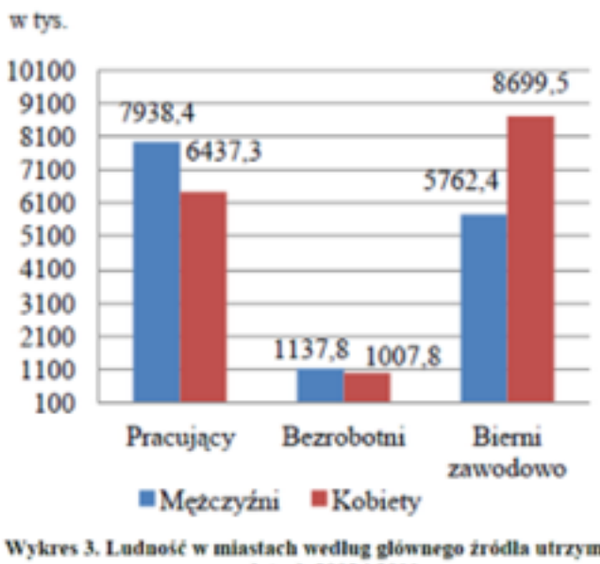
Raporty GUS

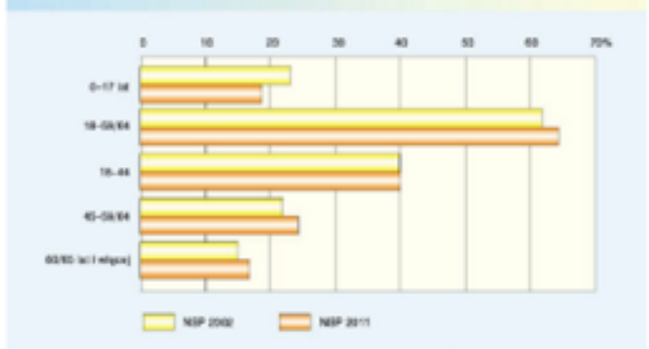
Dlaczego należy uważać z Excelem?



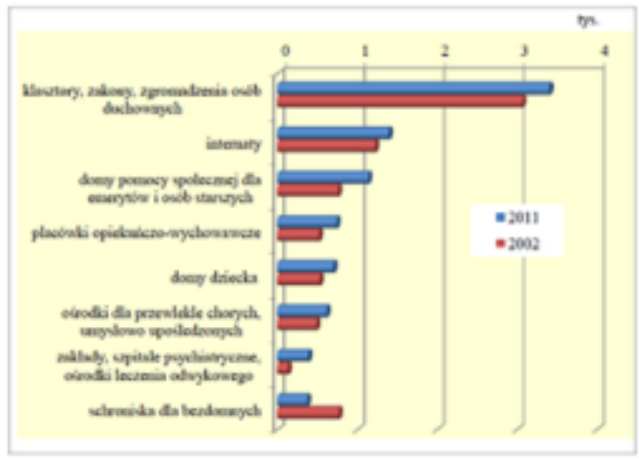




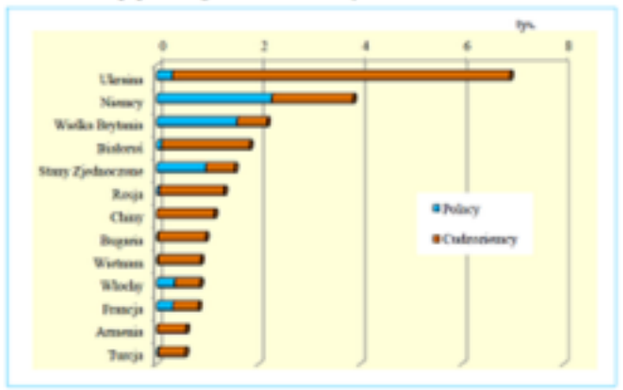


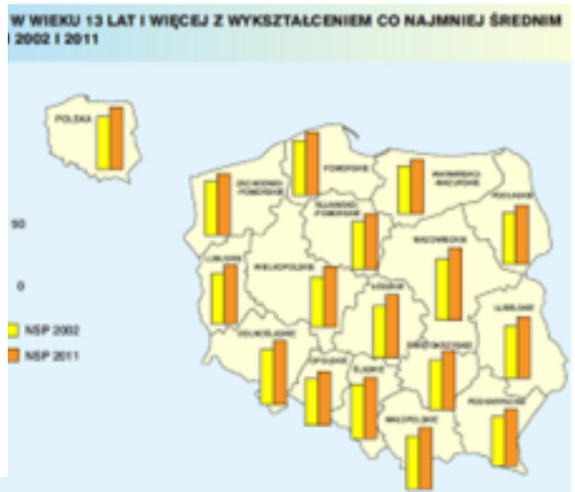


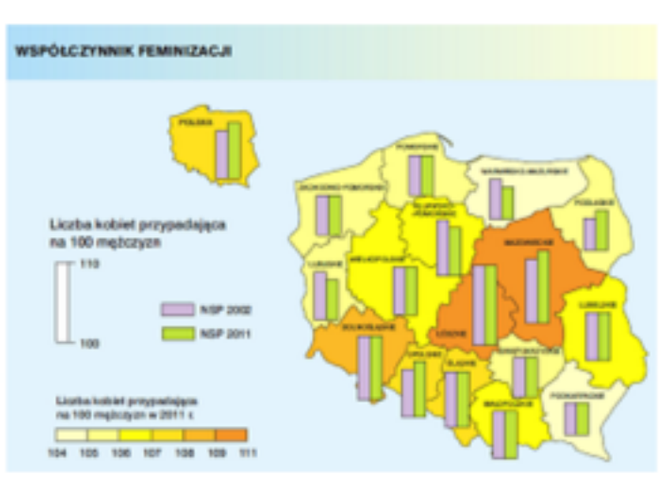
Wykres 11. Obiekty zbiorowego zakwaterowania według wybranych typów w latach 2002 i 2011



Wykres 12. Imigranci przebywający w Polsce czasowo powyżej 3 miesięcy według kraju poprzedniego zamieszkania i obywatelstwa w 2011 r.

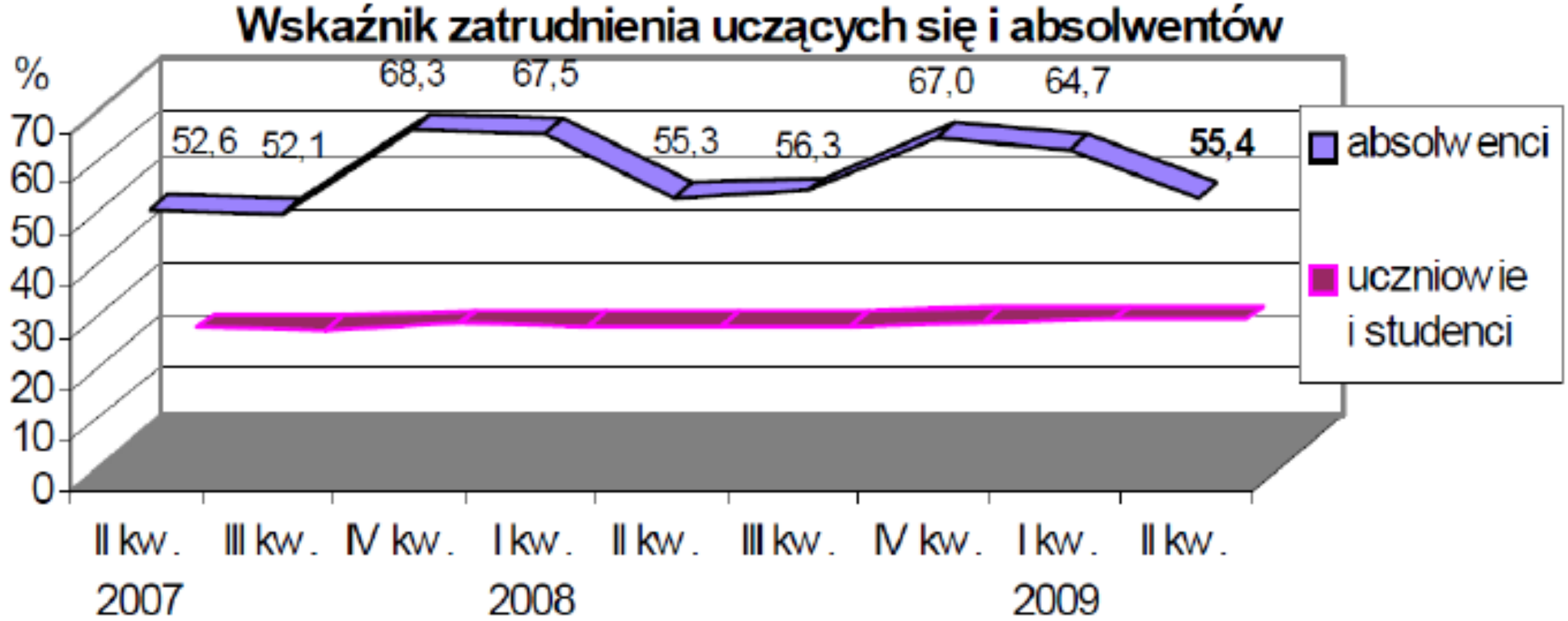




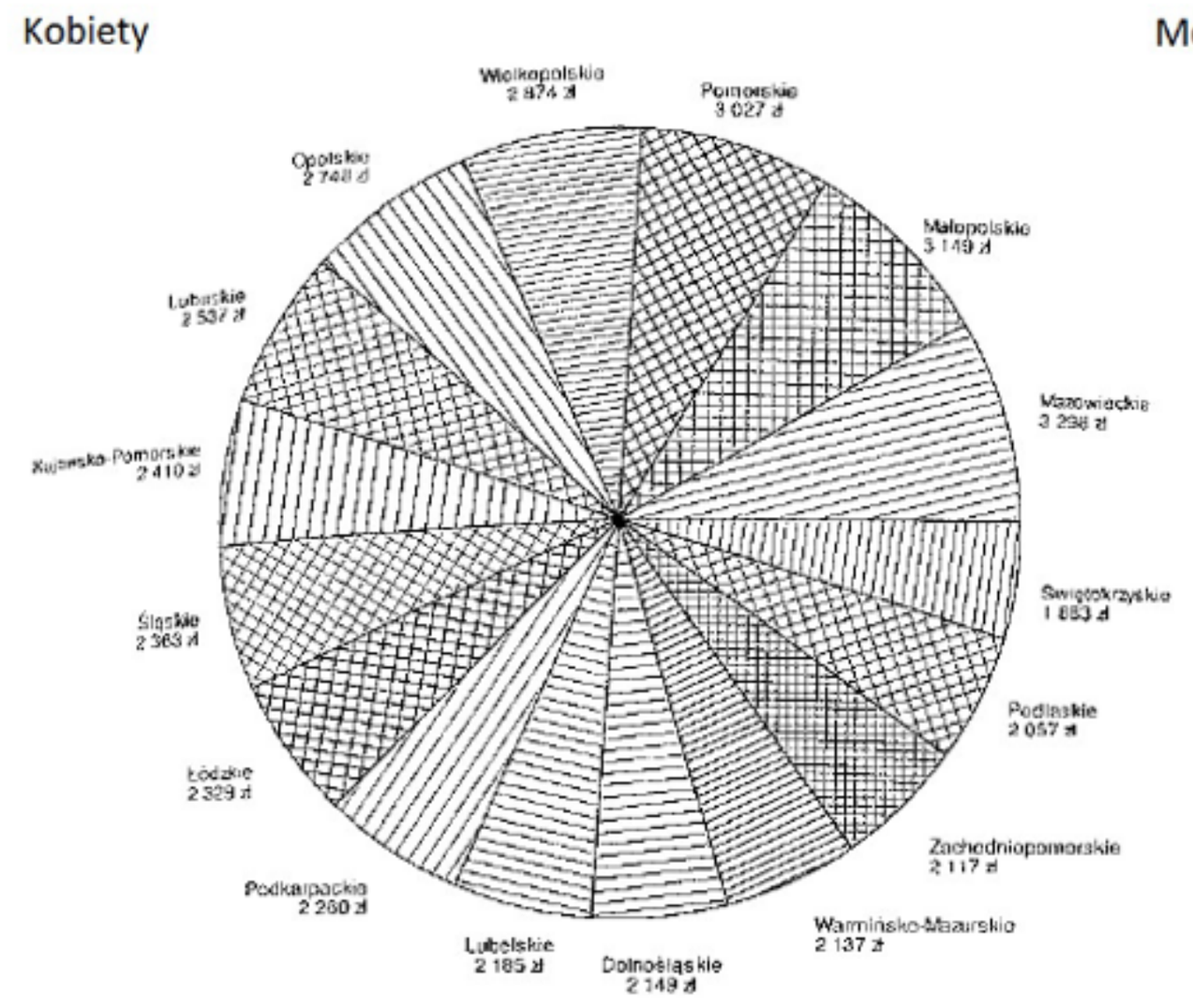


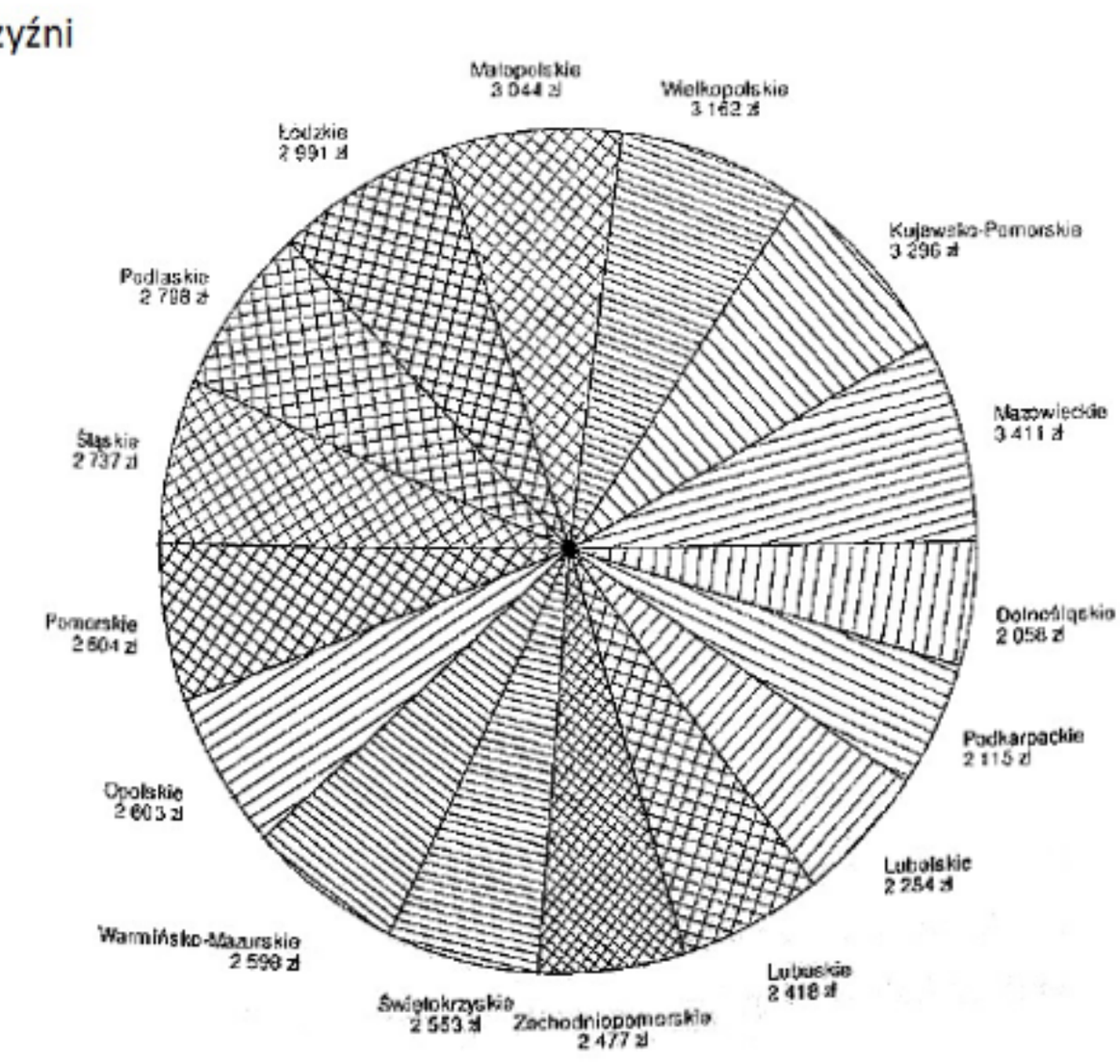
Raporty GUS

Dlaczego należy uważać z Excelem?

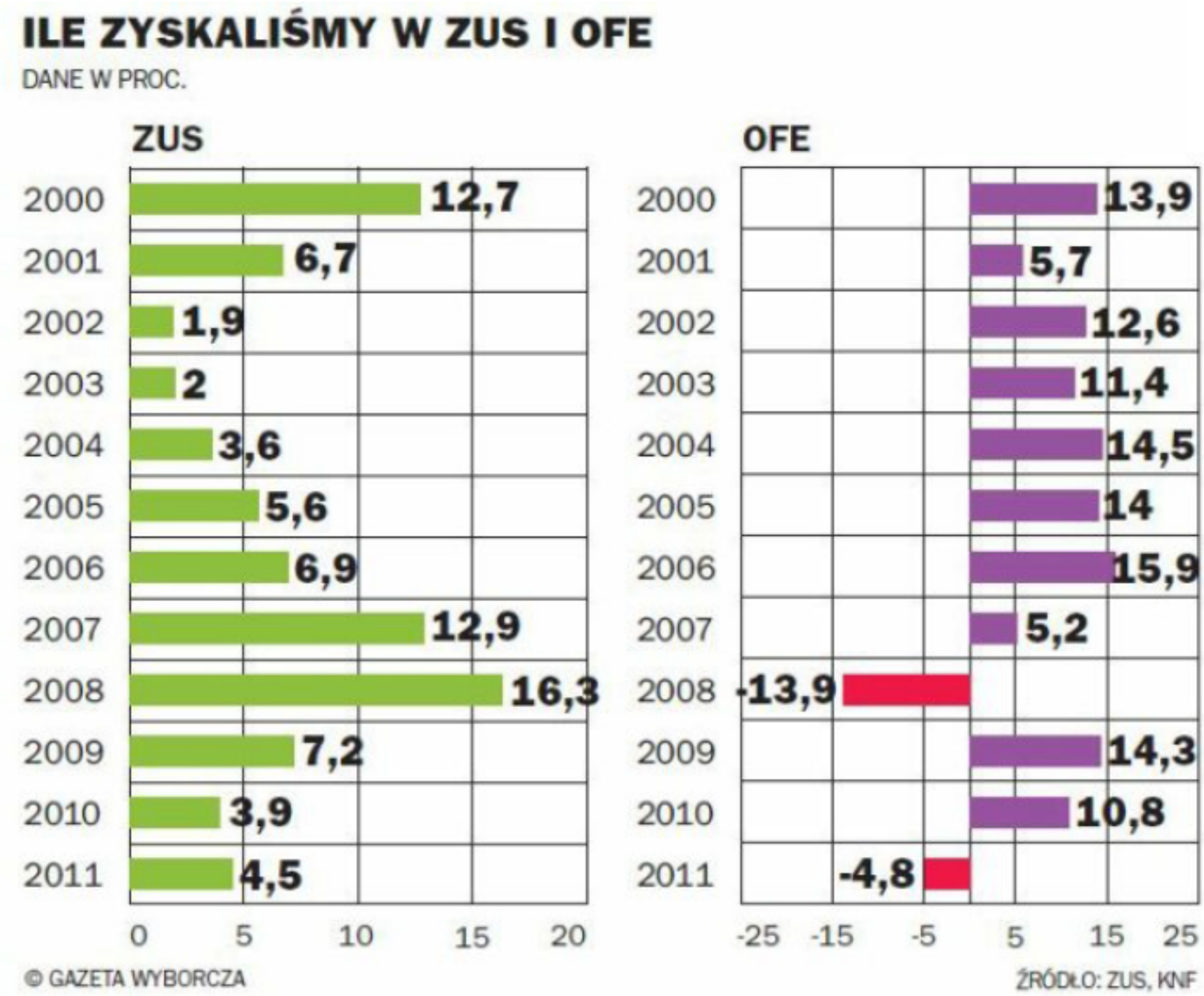


Średni dochód kobiet jest inny niż mężczyzn…





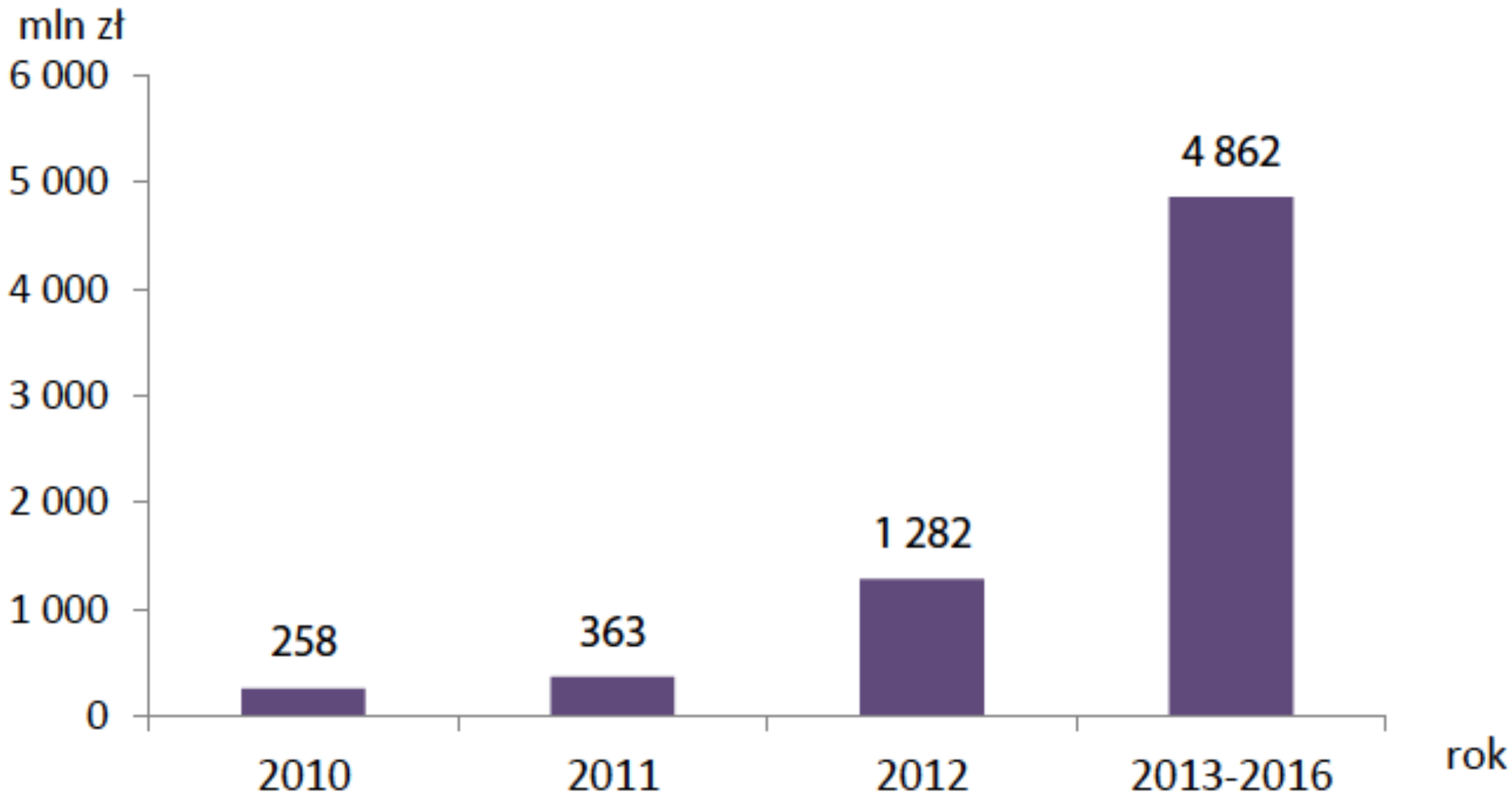
Z ZUS można zyskać znacznie więcej! Czyżby?



Znaczny wzrost kosztów użytkowania! Czyżby?

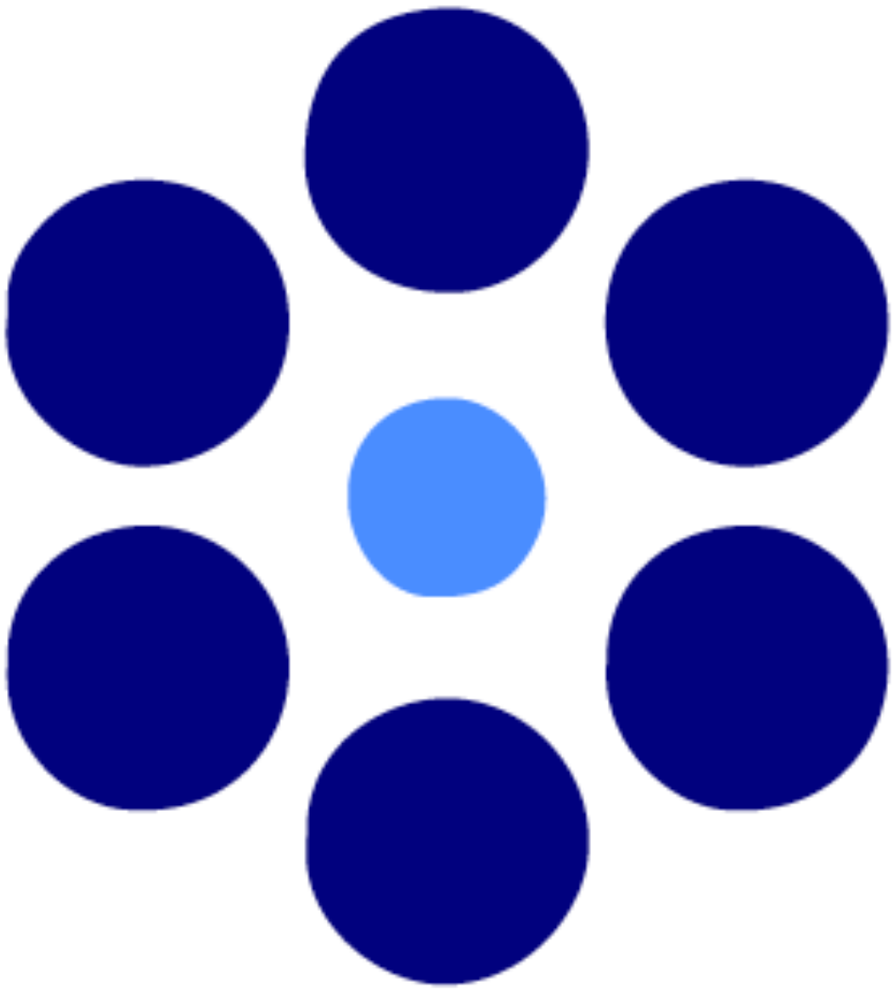
Inwestujemy coraz więcej w B+R! Czyżby?

I.3 Wydatki deklarowane przez przedsiębiorców na B+R w programach NCBiR w latach 2010-2016.

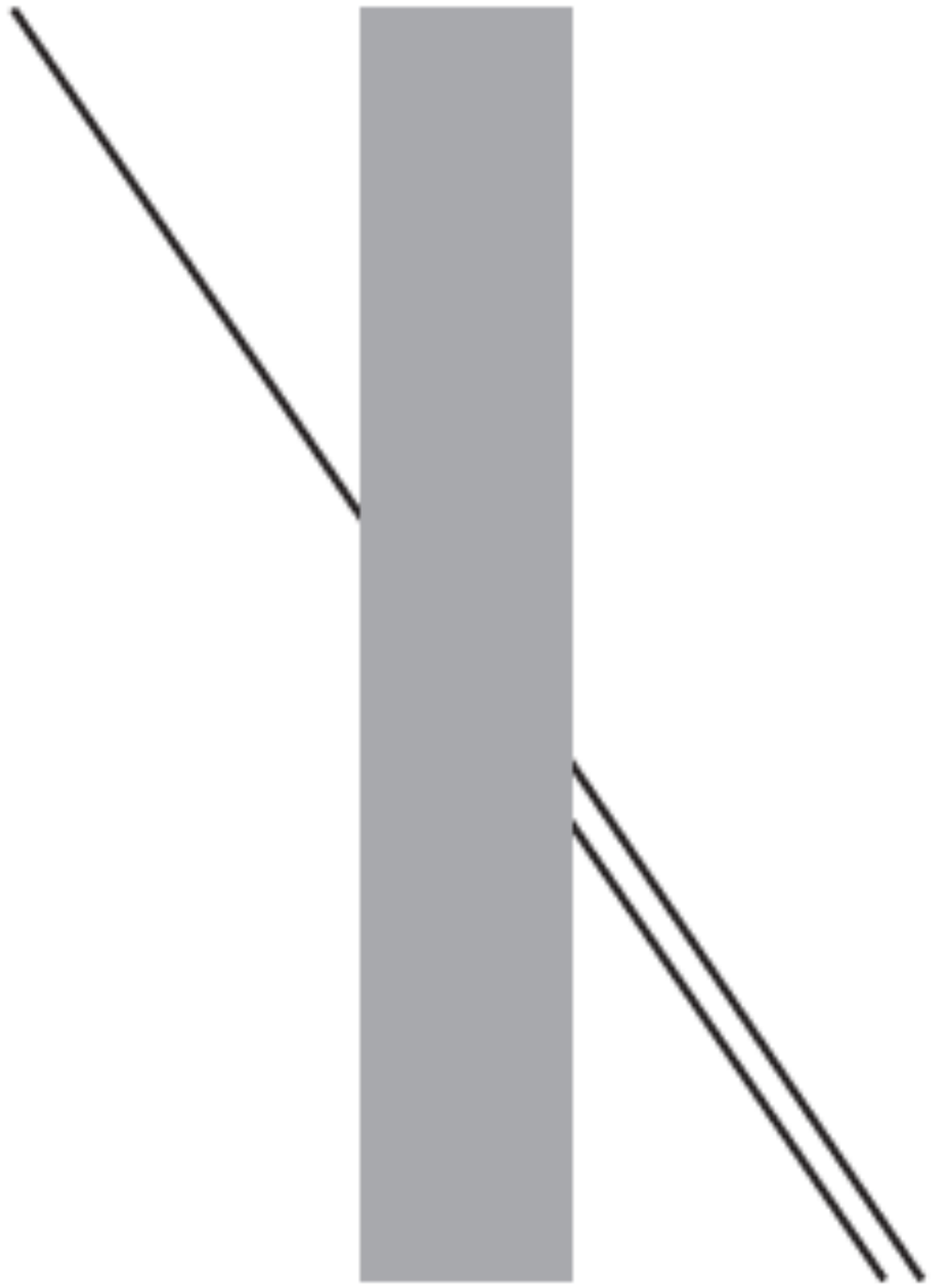


źródło: Narodowe Centrum Badań i Rozwoju

Które z jasnobiebieskich kół jest większe?

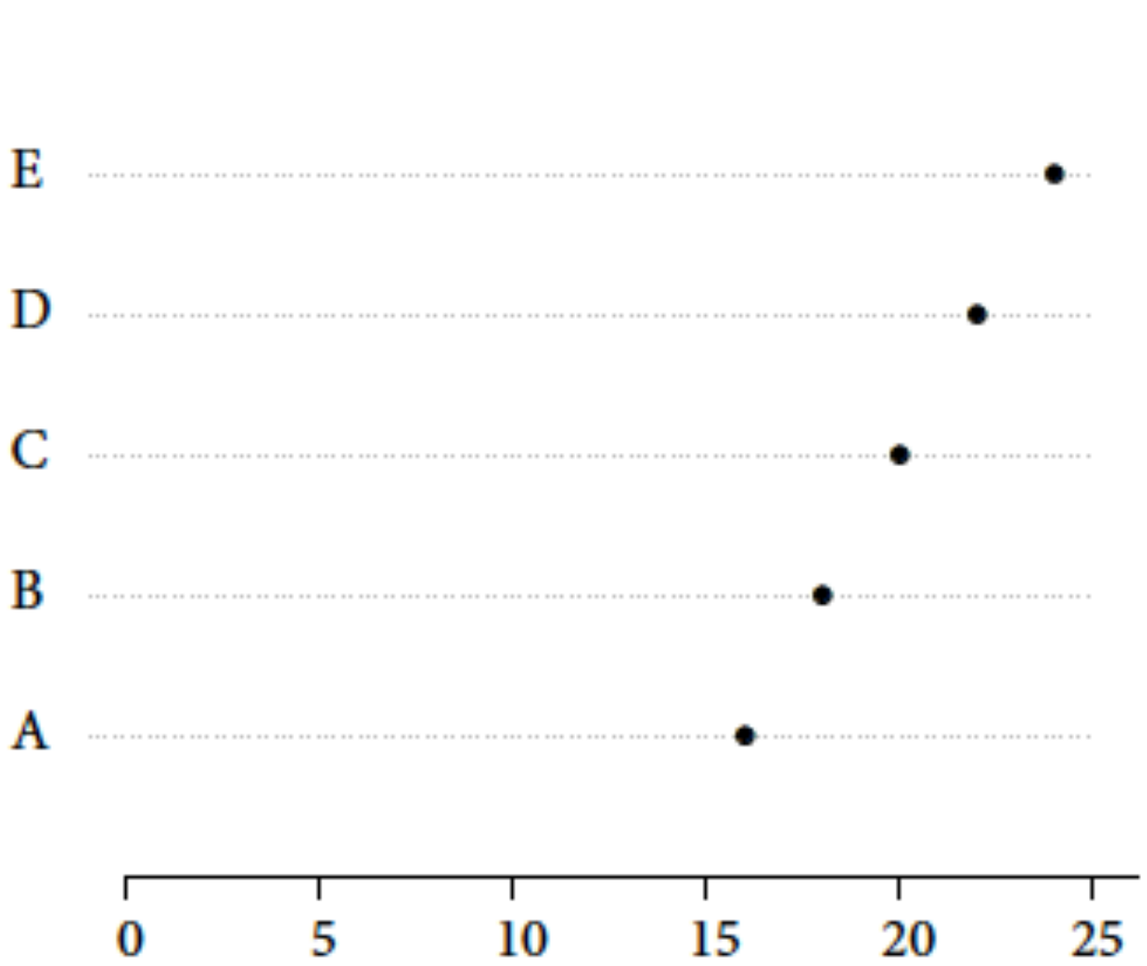


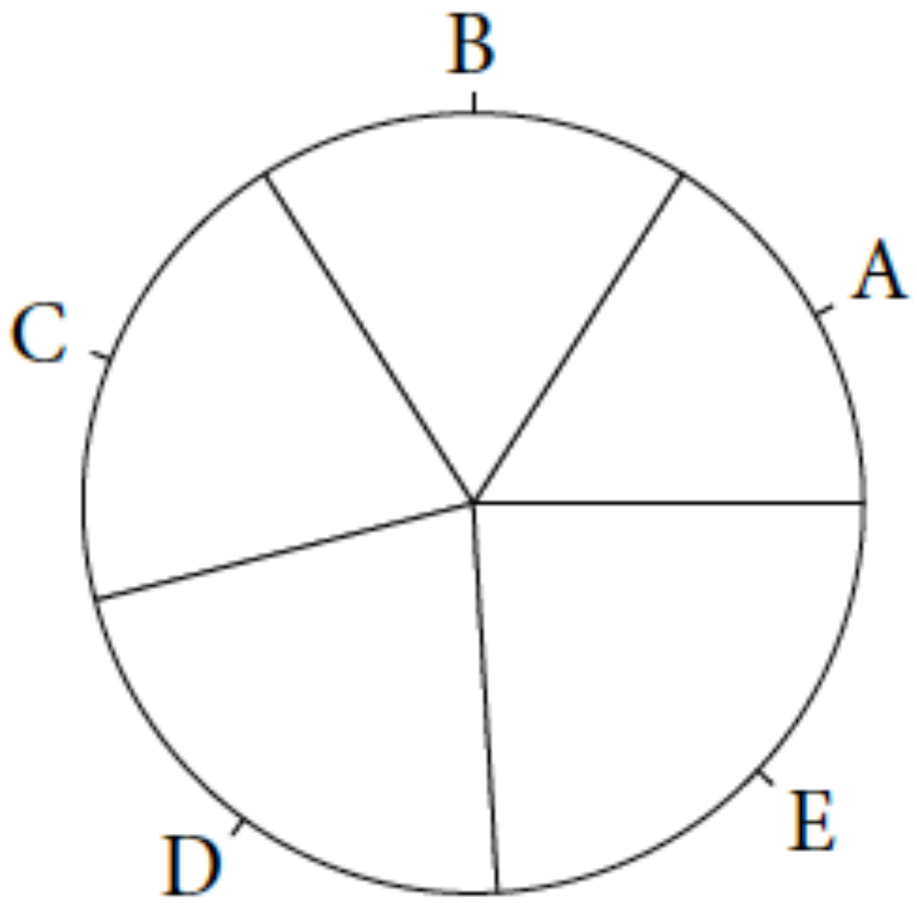
Która z prawych linii jest przedłużeniem lewej?

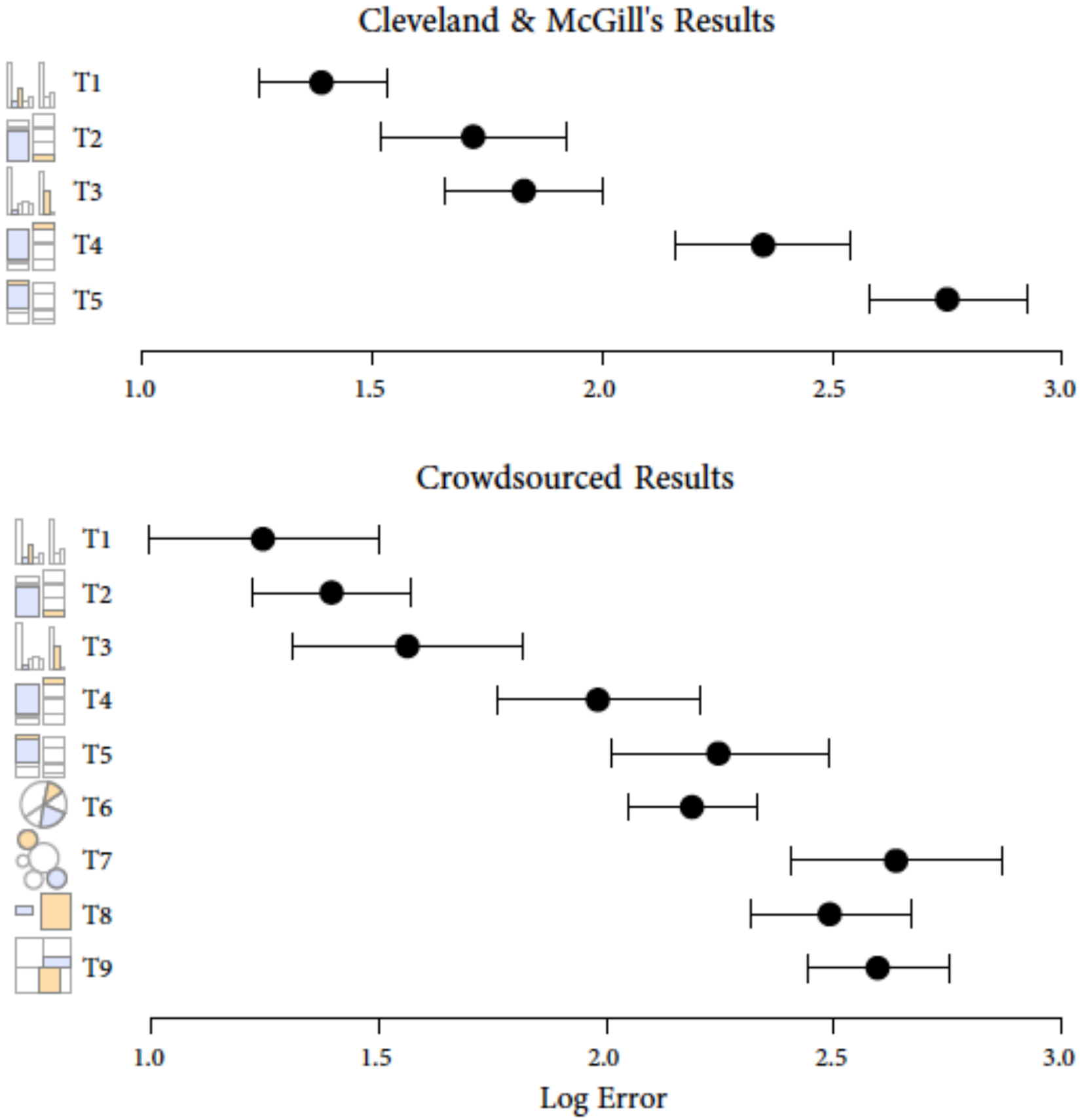


A więc jak żyć?

Zagadka 3:
Pola, kąty, pozycje czy długości?



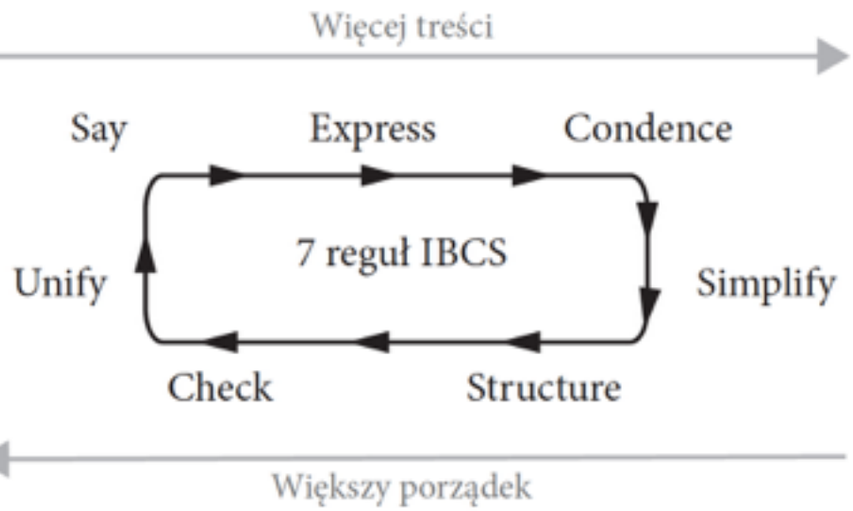


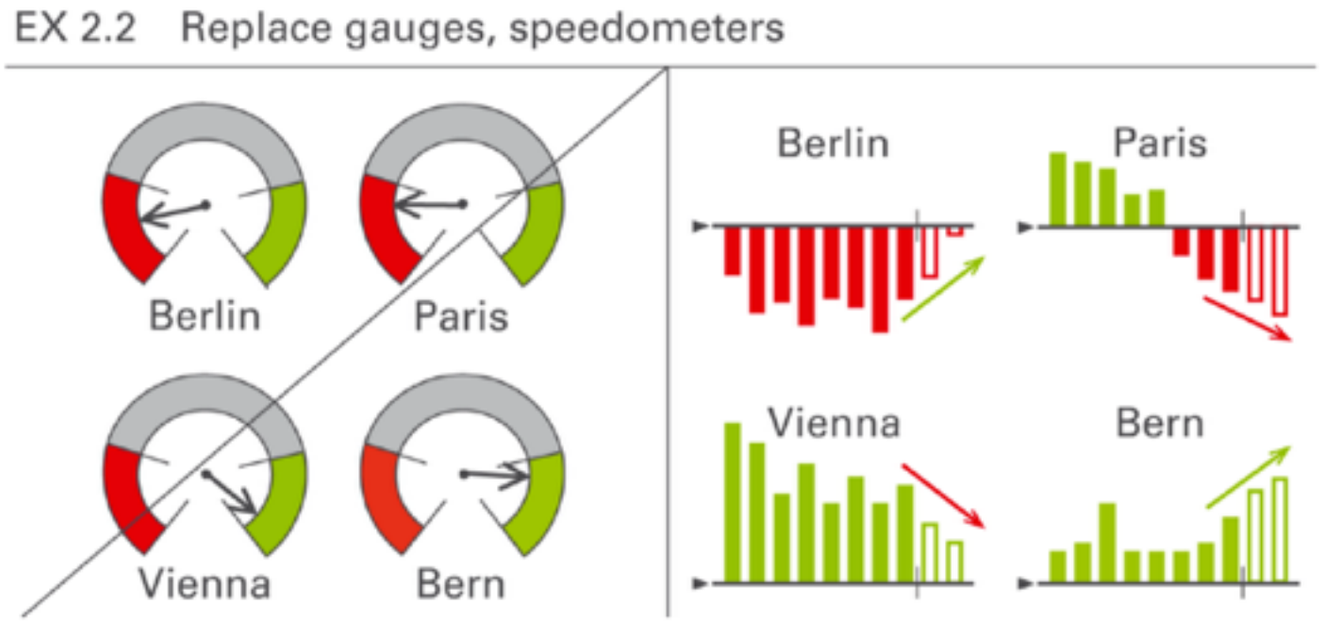


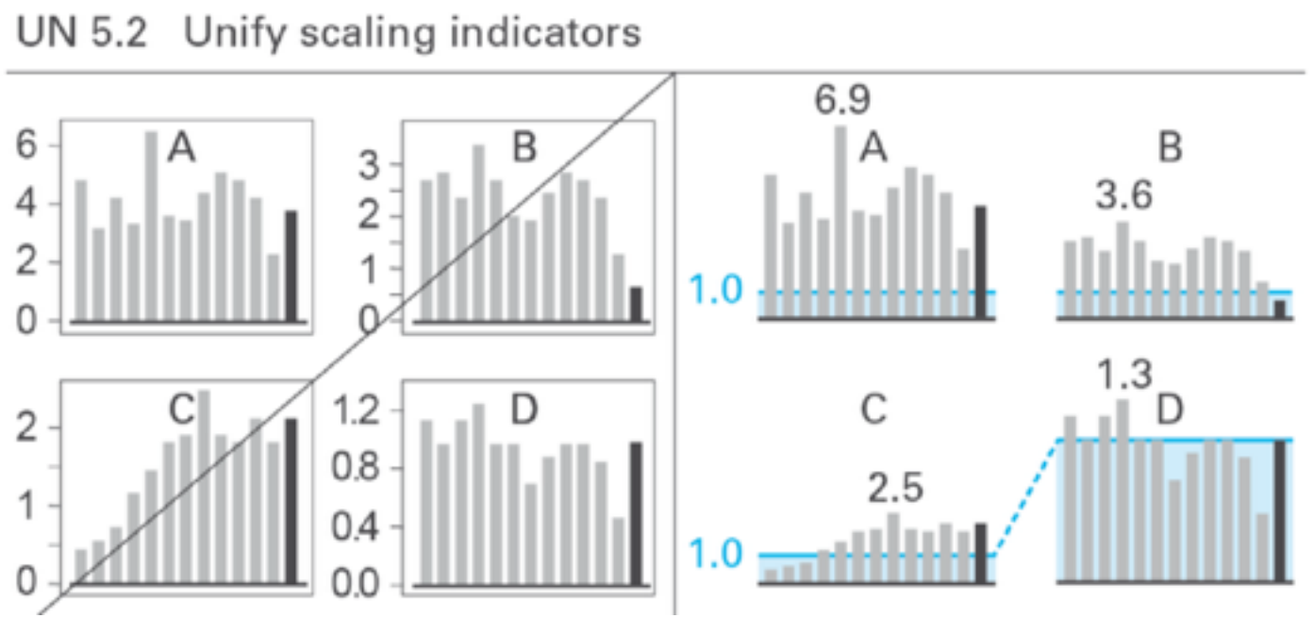
Jeffrey Heer and Michael Bostock. Crowdsourcing graphical perception: using mechanical turk to assess visualization design

The International Business Communication Standards
http://www.ibcs-a.org/

Say	Przekazuj konkretny, istotny komunikat.
Unify	Stosuj jednolite oznaczenia i wyróżnienia.
Condence	Zwiększaj ilość informacji, dbaj o treściwość.
Check	Zapewnij wizualność spójność wykresów i tabel.
Express	Wybierz odpowiednią formę (opis, tabelę, wykres) dla prezentowanych danych.
Simplify	Unikaj bałaganu i zbędnych ozdobników.
Structure	Zadbaj o czytelną strukturę.





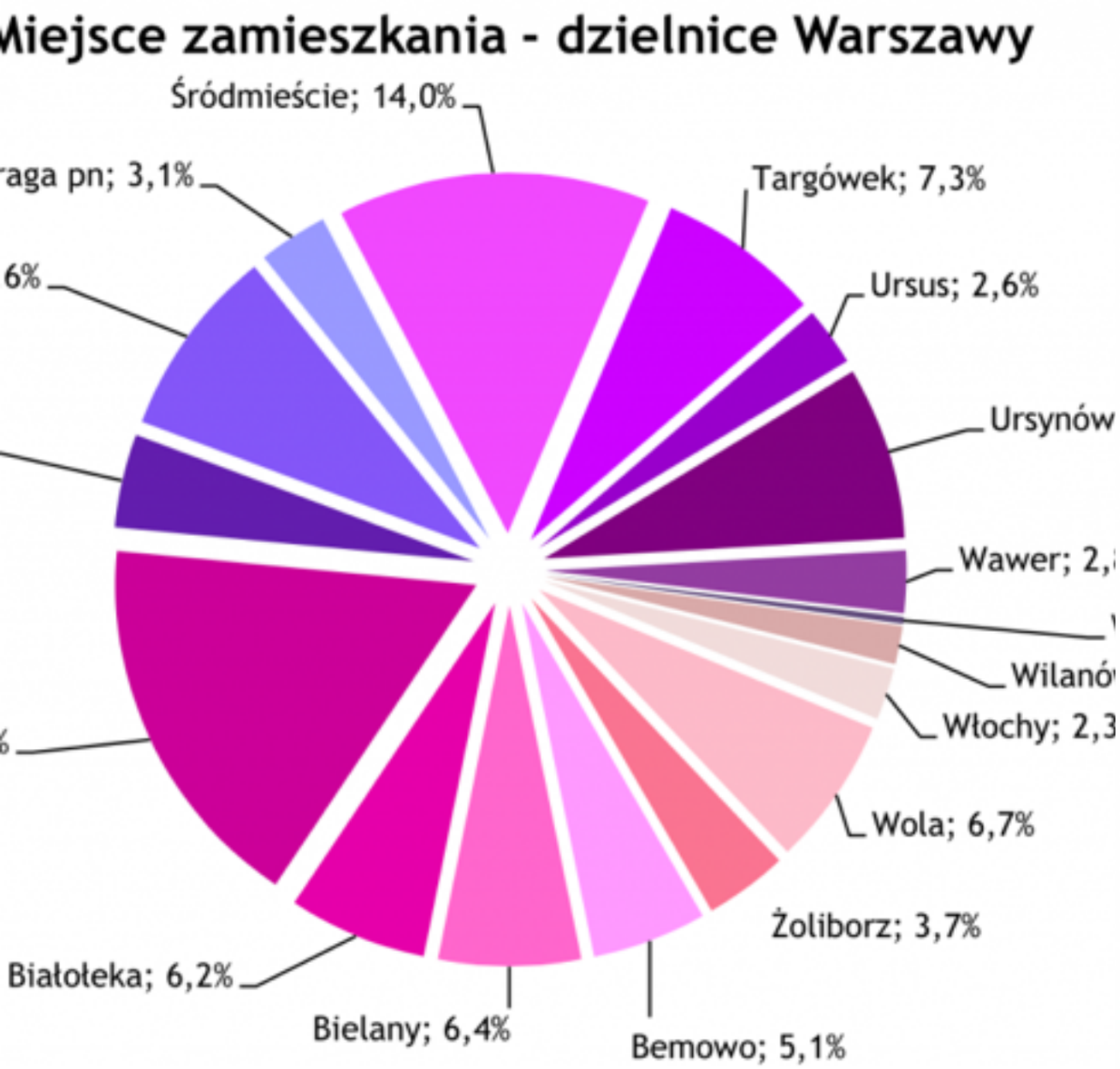


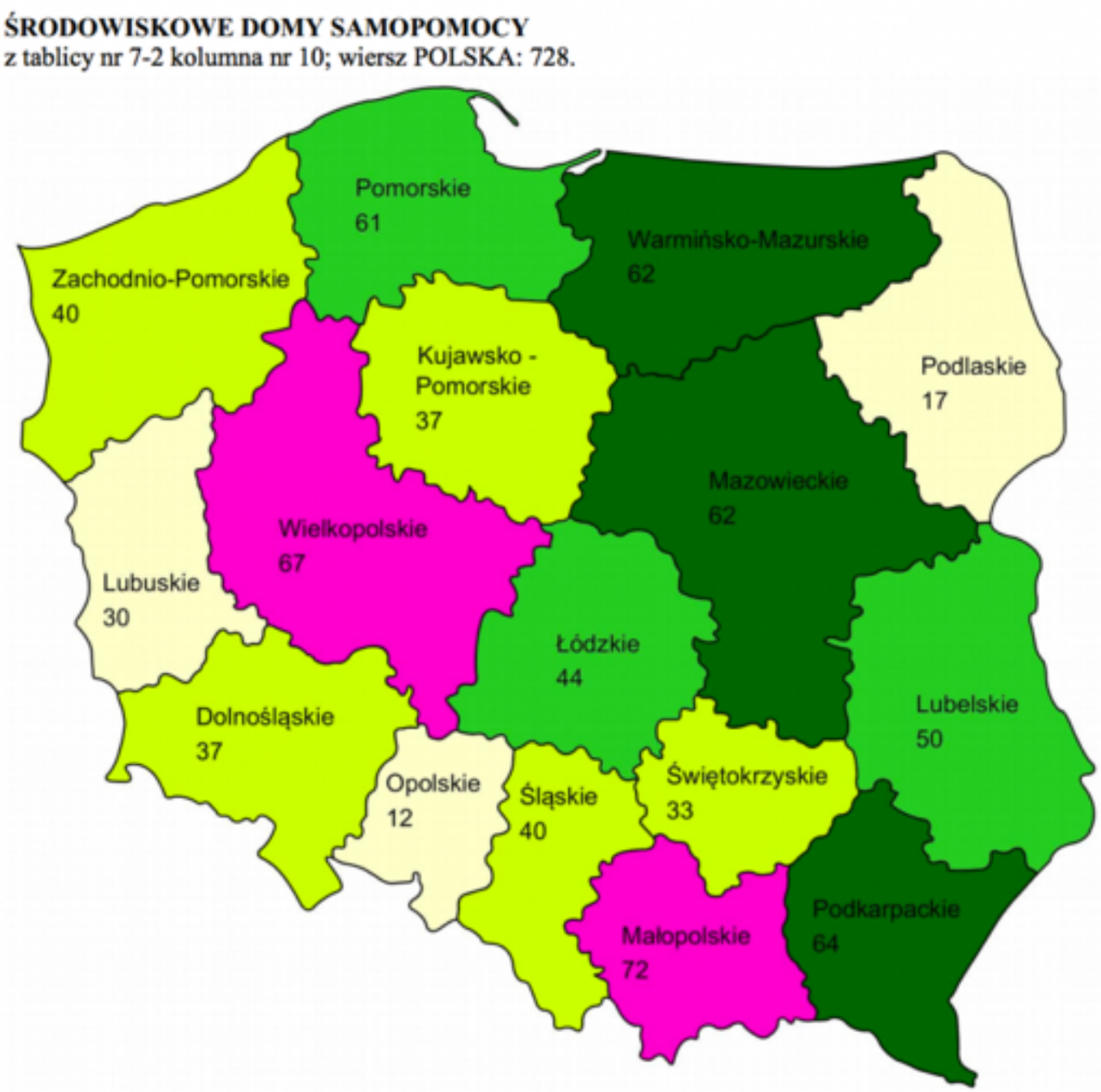
iHealth,
Punkty, łamana czy słupki?

Ok, w takim razie używamy słupków.

A jak dobierać kolory?

Nie ma różu bez ognia

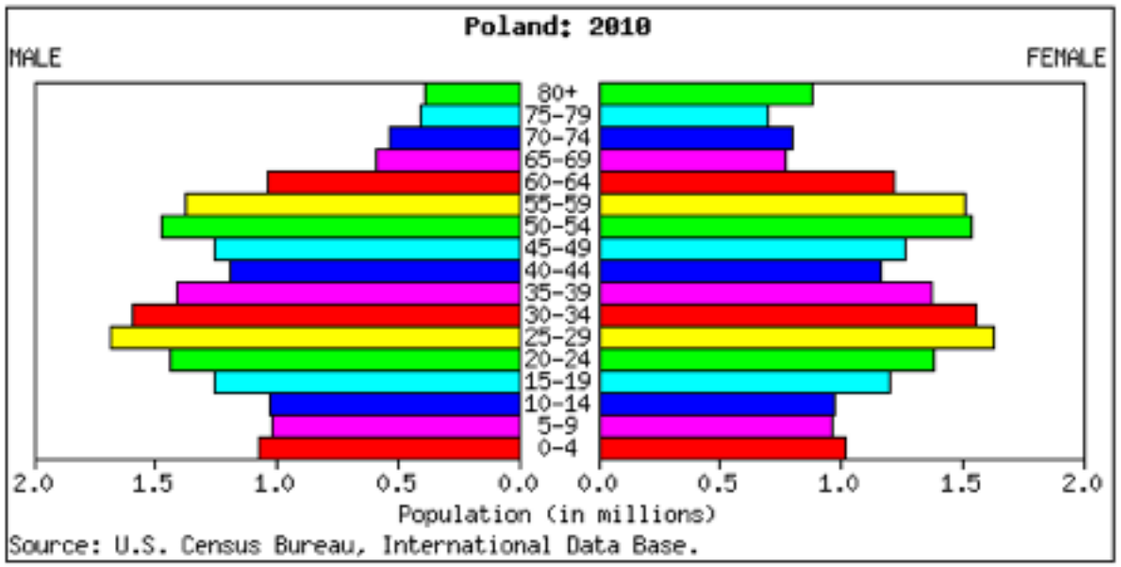




Nie ma różu bez ognia

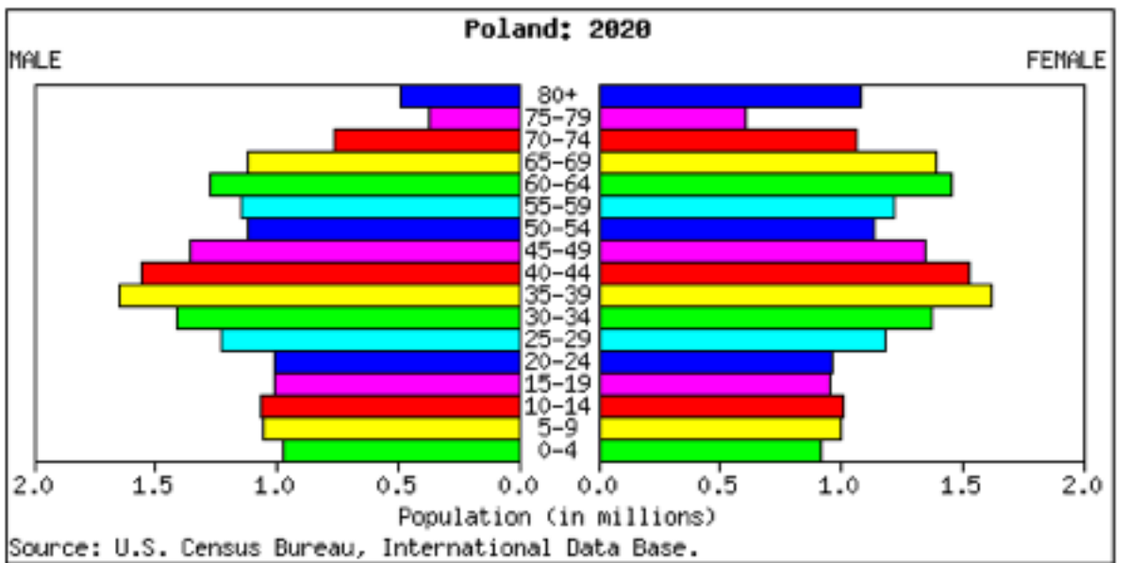
Poland Population Pyramid for 2010

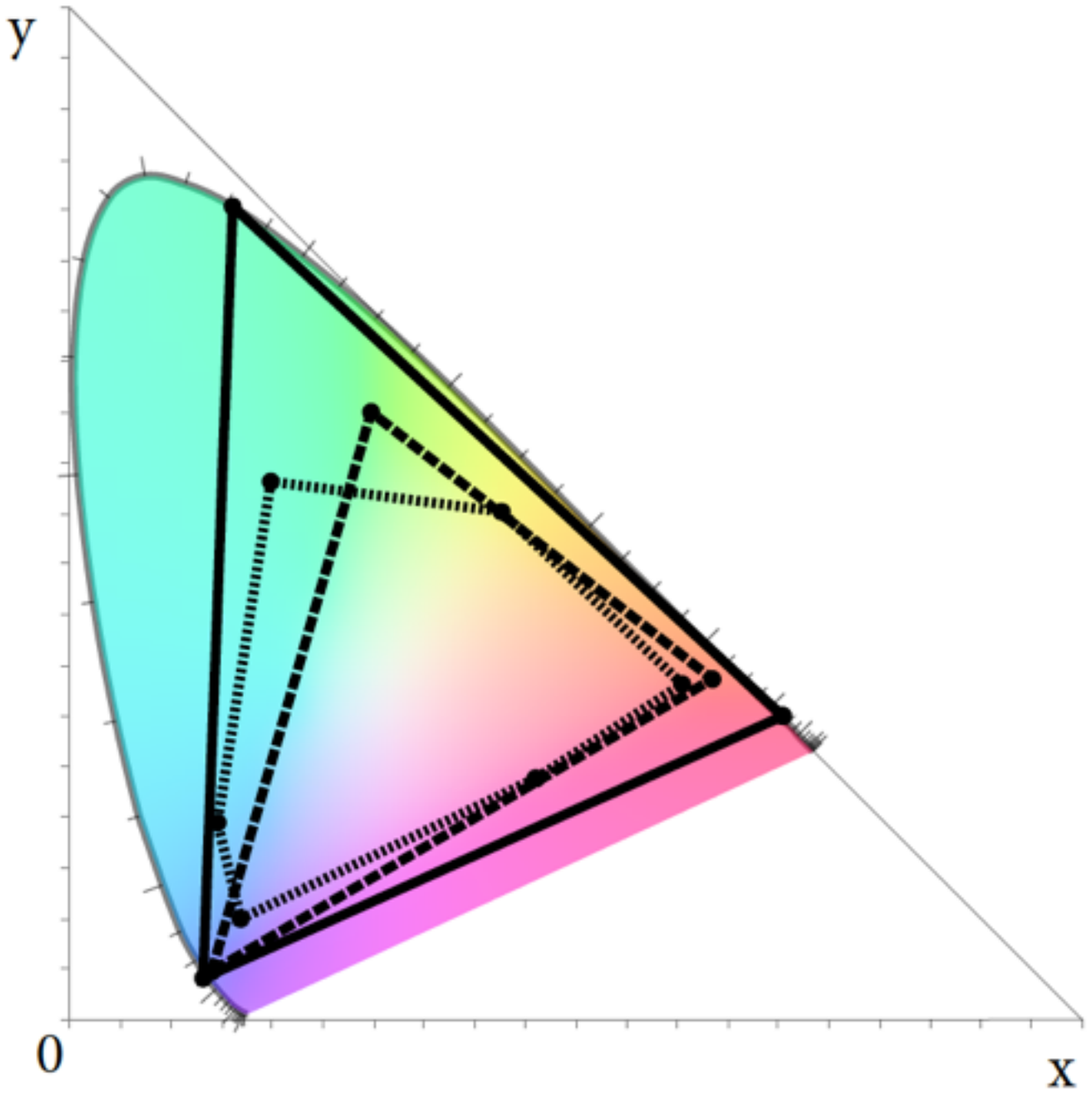
Age and sex distribution for the year 2010:

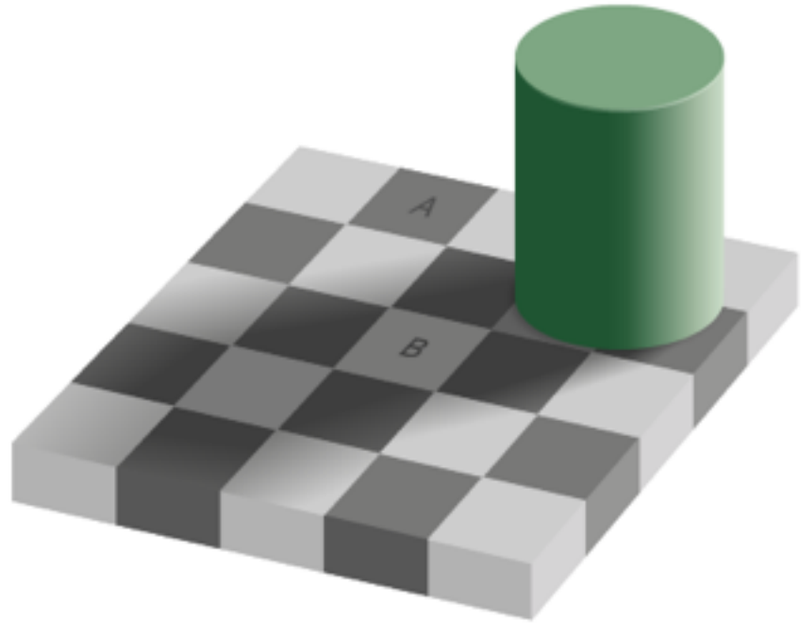


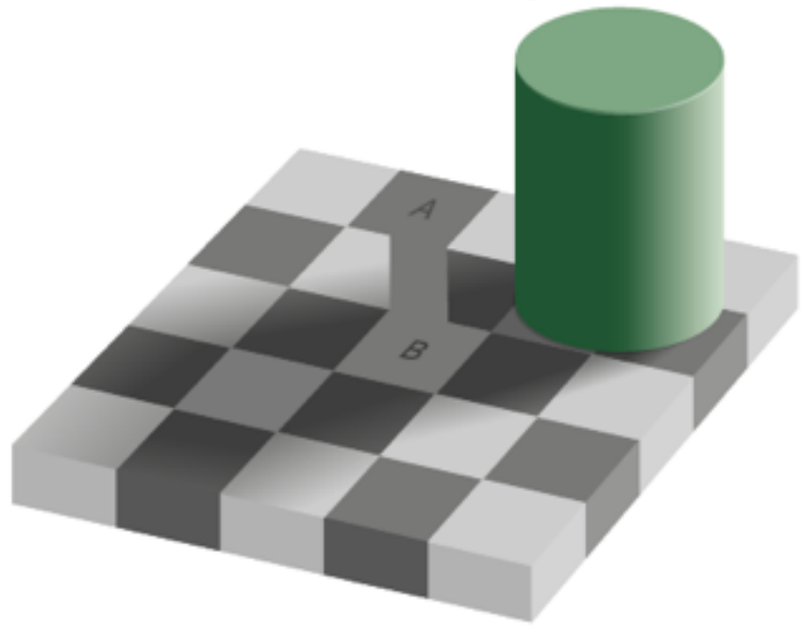
Poland Population Pyramid for 2020

Predicted age and sex distribution for the year 2020:

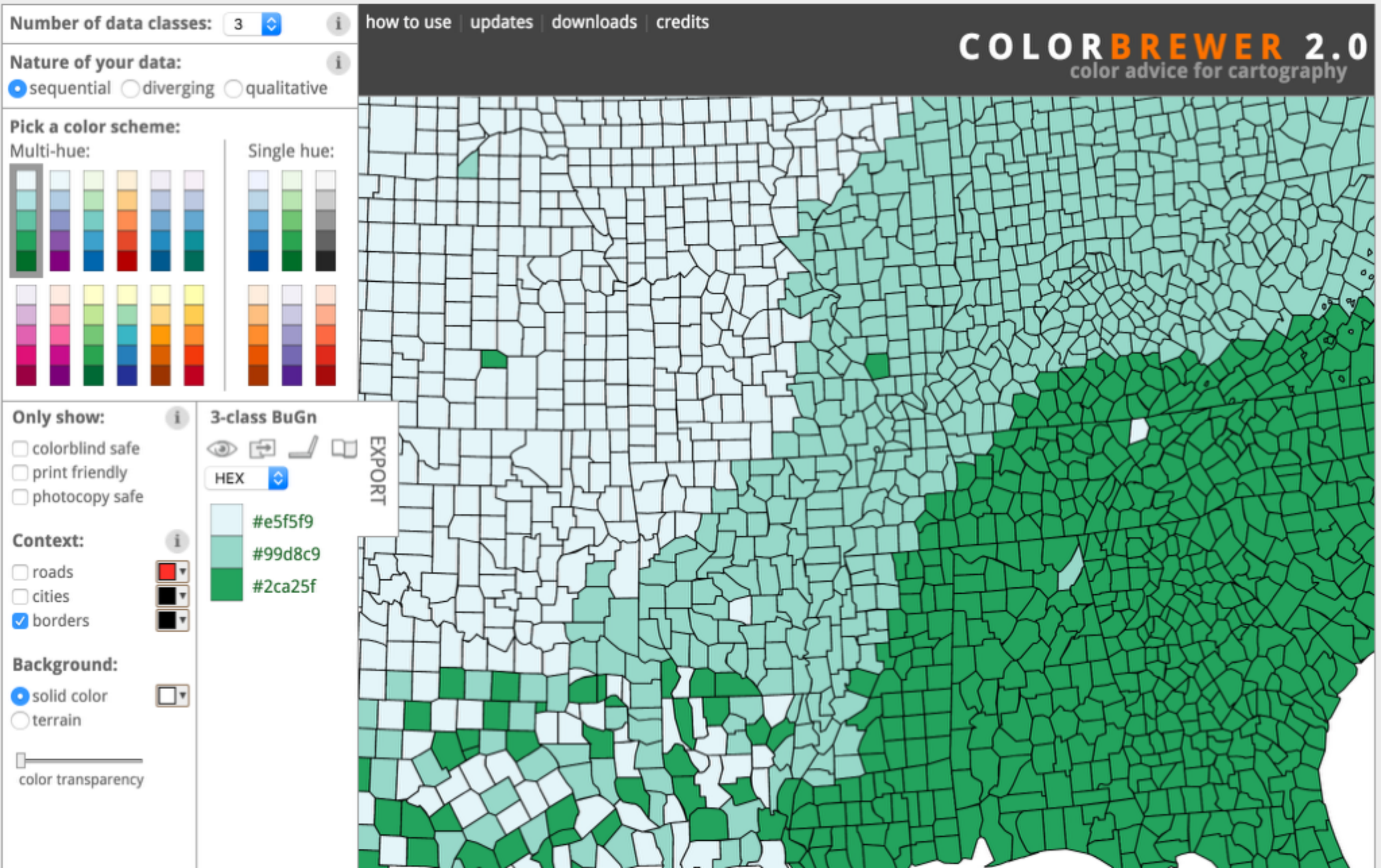



y
0
x

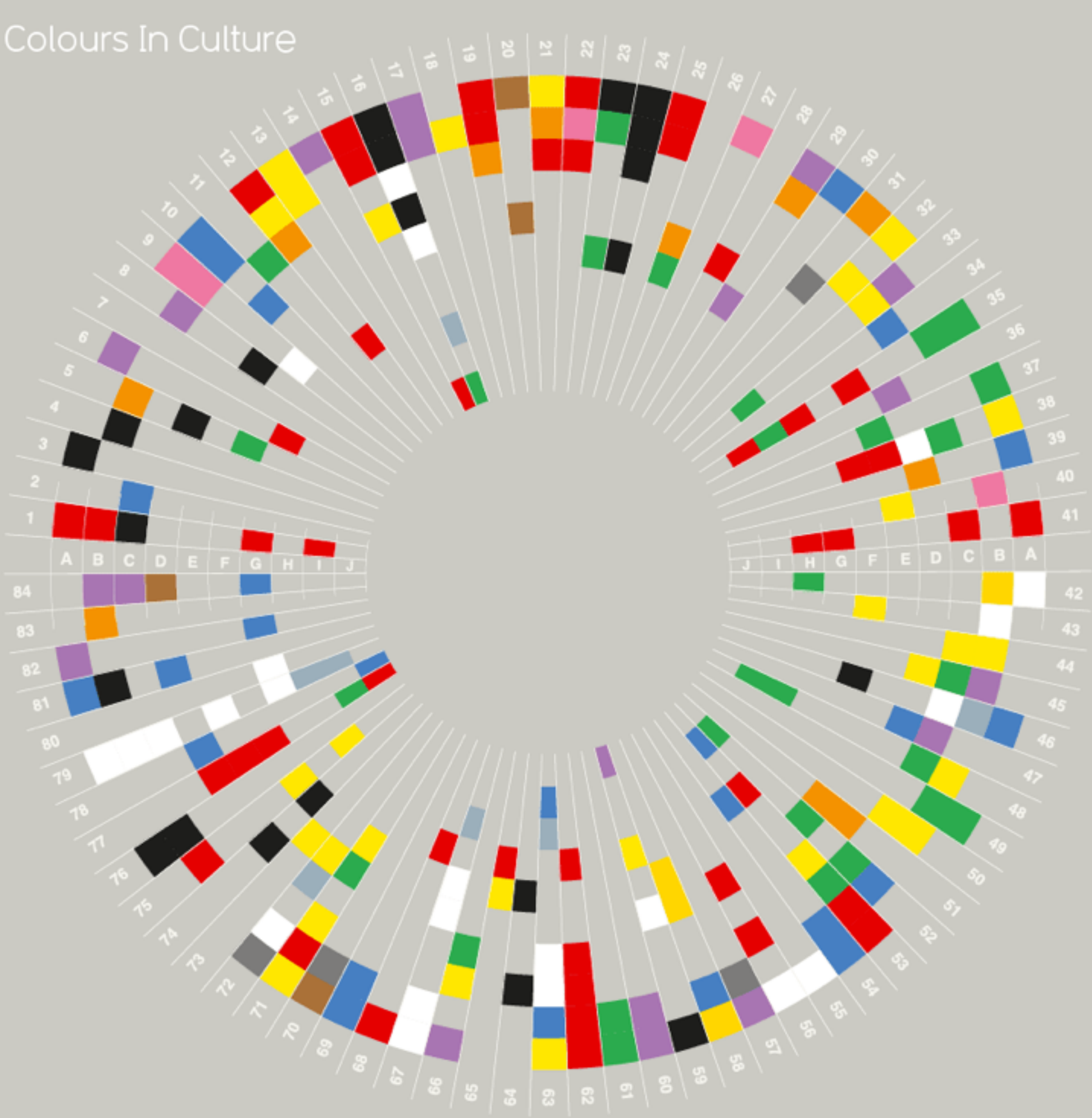

A
B


A
B

Kim jest Cynthia Brewer?



Colours In Culture



A Western / American
B Japanese
C Hindu
D Native American
E Chinese
F Asian
G Eastern European
H Muslim
I African
J South American

1 Anger
2 Art / Creativity
3 Authority
4 Bad Luck
5 Balance
6 Beauty
7 Calm
8 Celebration
9 Children
10 Cold
11 Compassion
12 Courage
13 Cowardice
14 Cruelty
15 Danger
16 Death
17 Decadence
18 Deceit

19 Des
20 Ear
21 Ene
22 Ero
23 Ete
24 Evi
25 Exc
26 Fan
27 Fer
28 Fer
29 Fla
30 Fre
31 Fri
32 Fur
33 Go
34 Go
35 Go
36 Gra

37 Growth
38 Happiness
39 Healing
40 Healthy
41 Heat
42 Heaven
43 Holiness
44 Illness
45 Insight
46 Intelligence
47 Intuition
48 Religion
49 Jealousy
50 Joy
51 Learning
52 Life
53 Love
54 Loyalty

55 Luxury
56 Marriage
57 Modesty
58 Money
59 Mourning
60 Mystery
61 Nature
62 Passion
63 Peace
64 Penance
65 Power
66 Personal power
67 Purity
68 Radicalism
69 Rational
70 Reliable
71 Repels Evil
72 Respect

73 Ro
74 Sel
75 Str
76 Sty
77 Su
78 Tro
79 Tru
80 Tru
81 Un
82 Virt
83 Wa
84 Wis

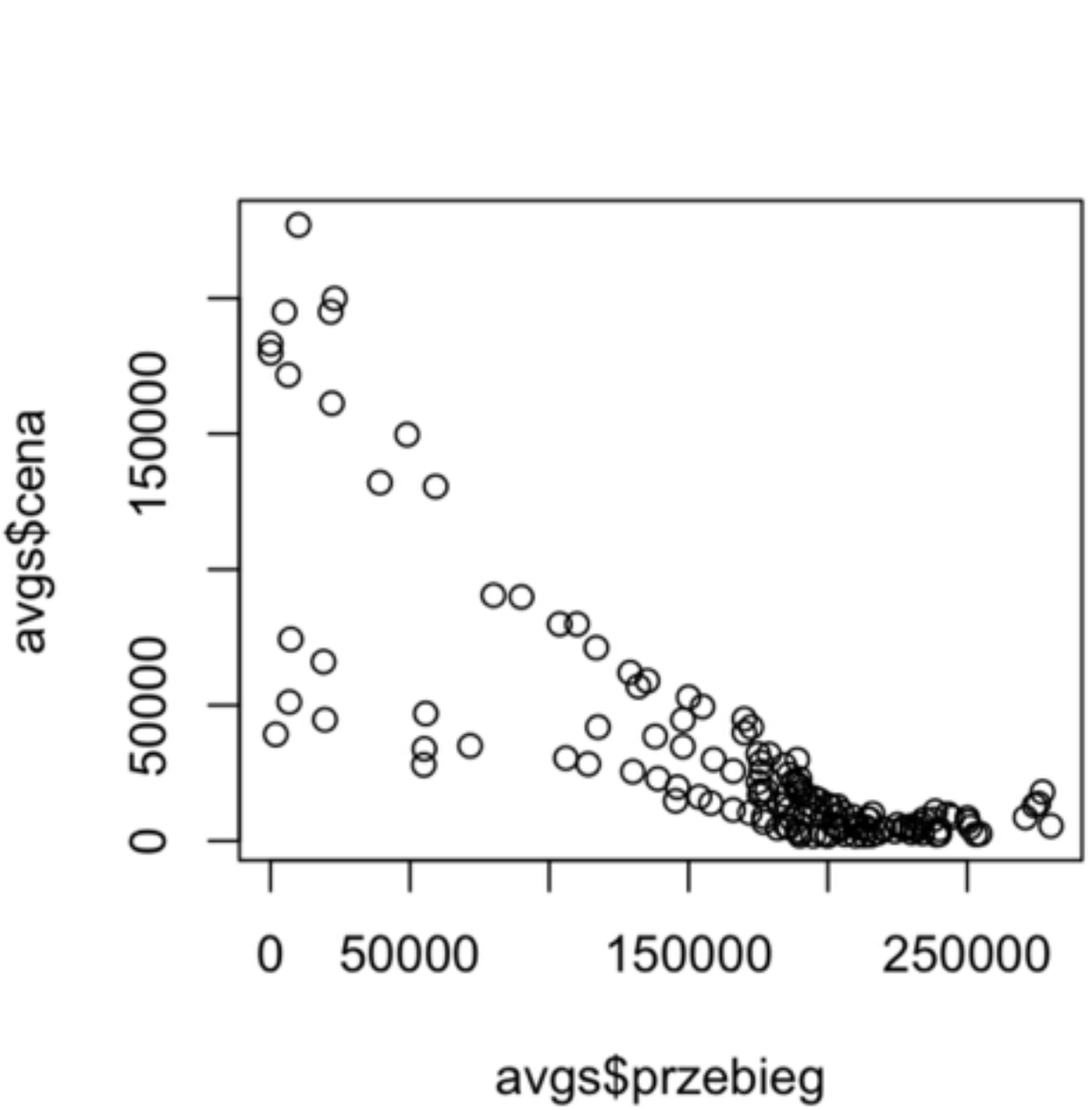
Yellow
Grey
Gold
Silver

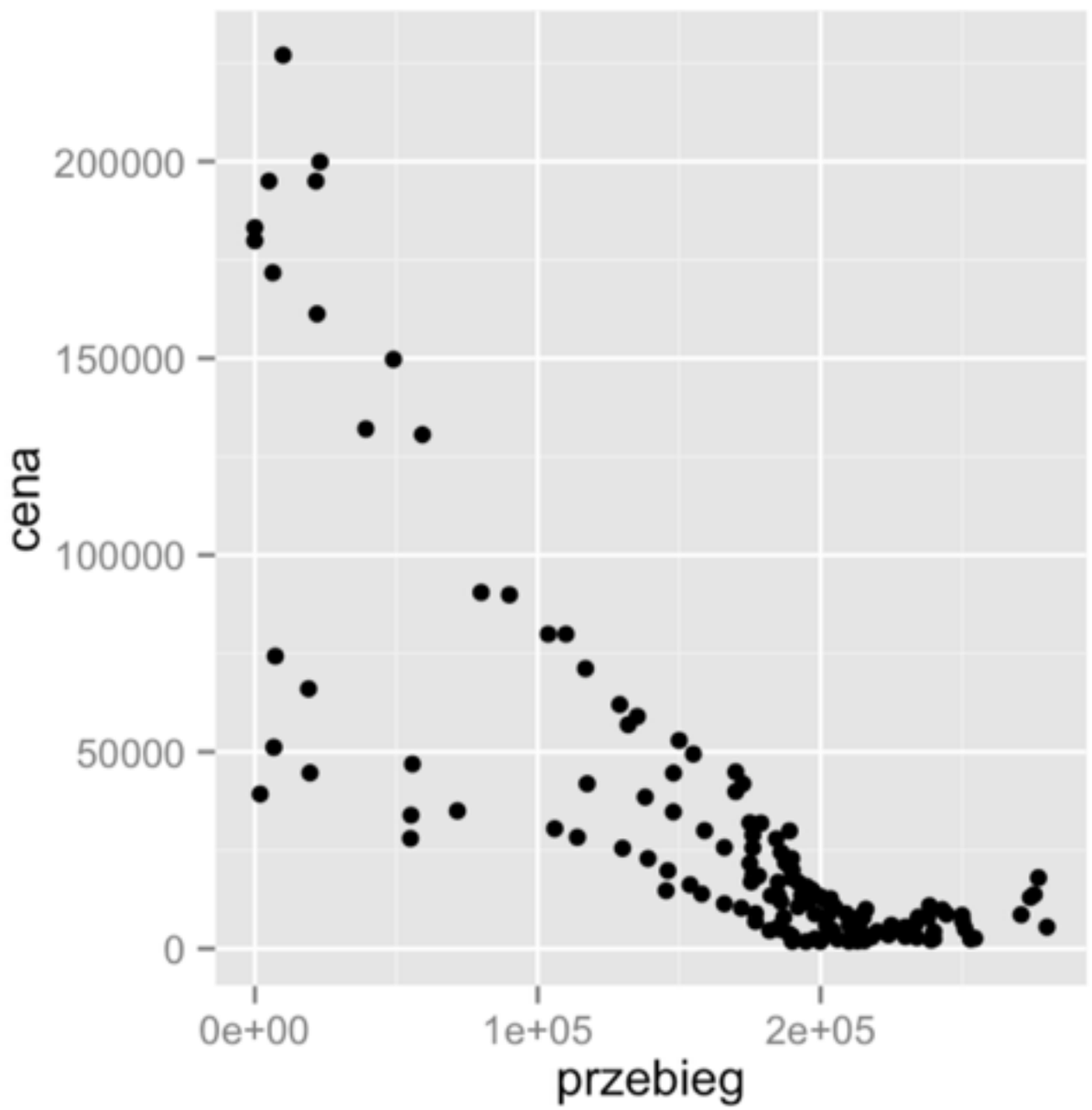
source: Pantone, ColorMatters & w

Czas na R

Trzy podejścia do tworzenia statycznej grafiki

```
plot(avgs$przebieg, avgs$cena)           # base graphics
ggplot(avgs, aes(x=przebieg, y=cena)) + # ggplot graphics
    geom_point()
```



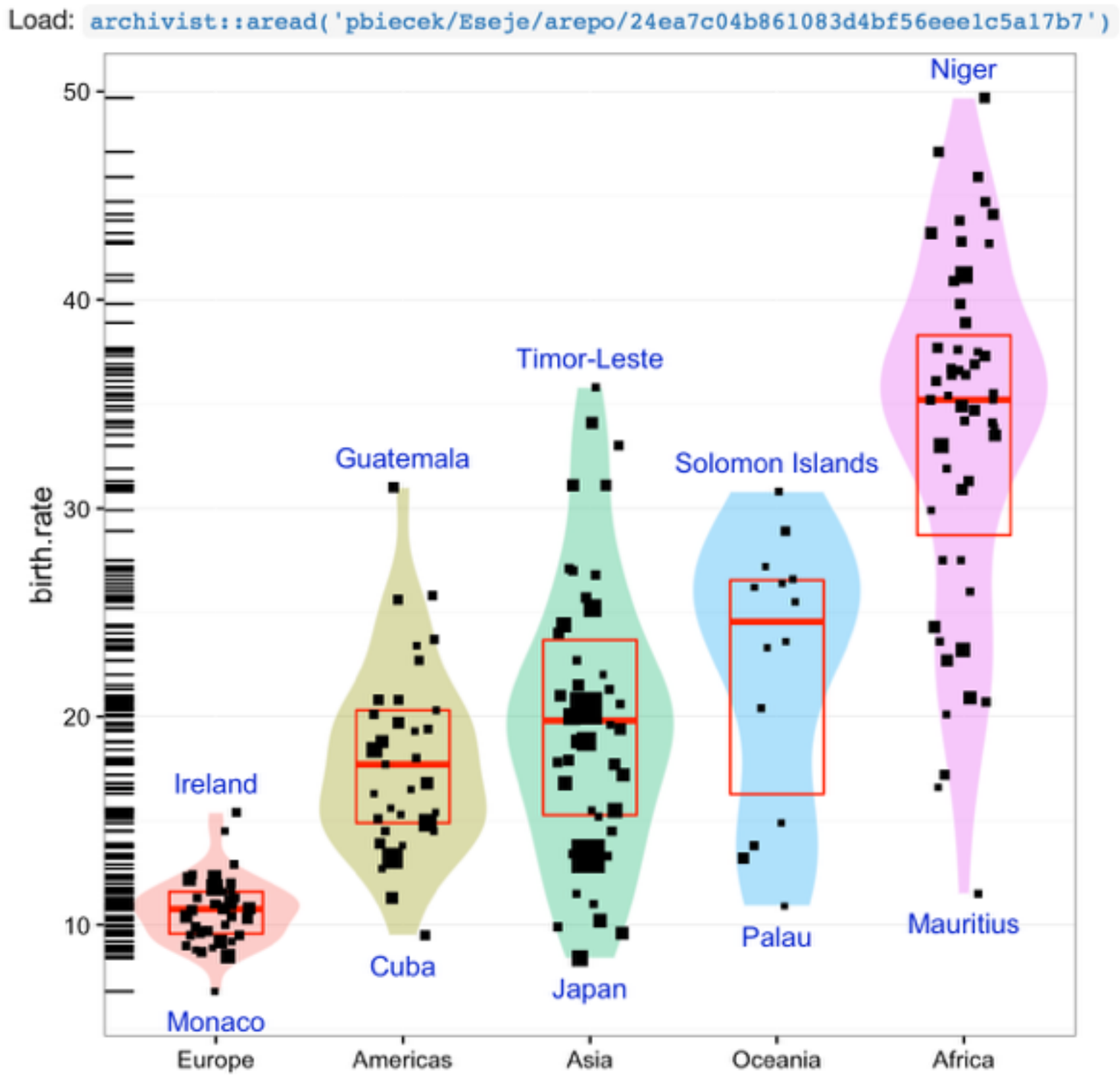


A więc dlaczego ggplot2?

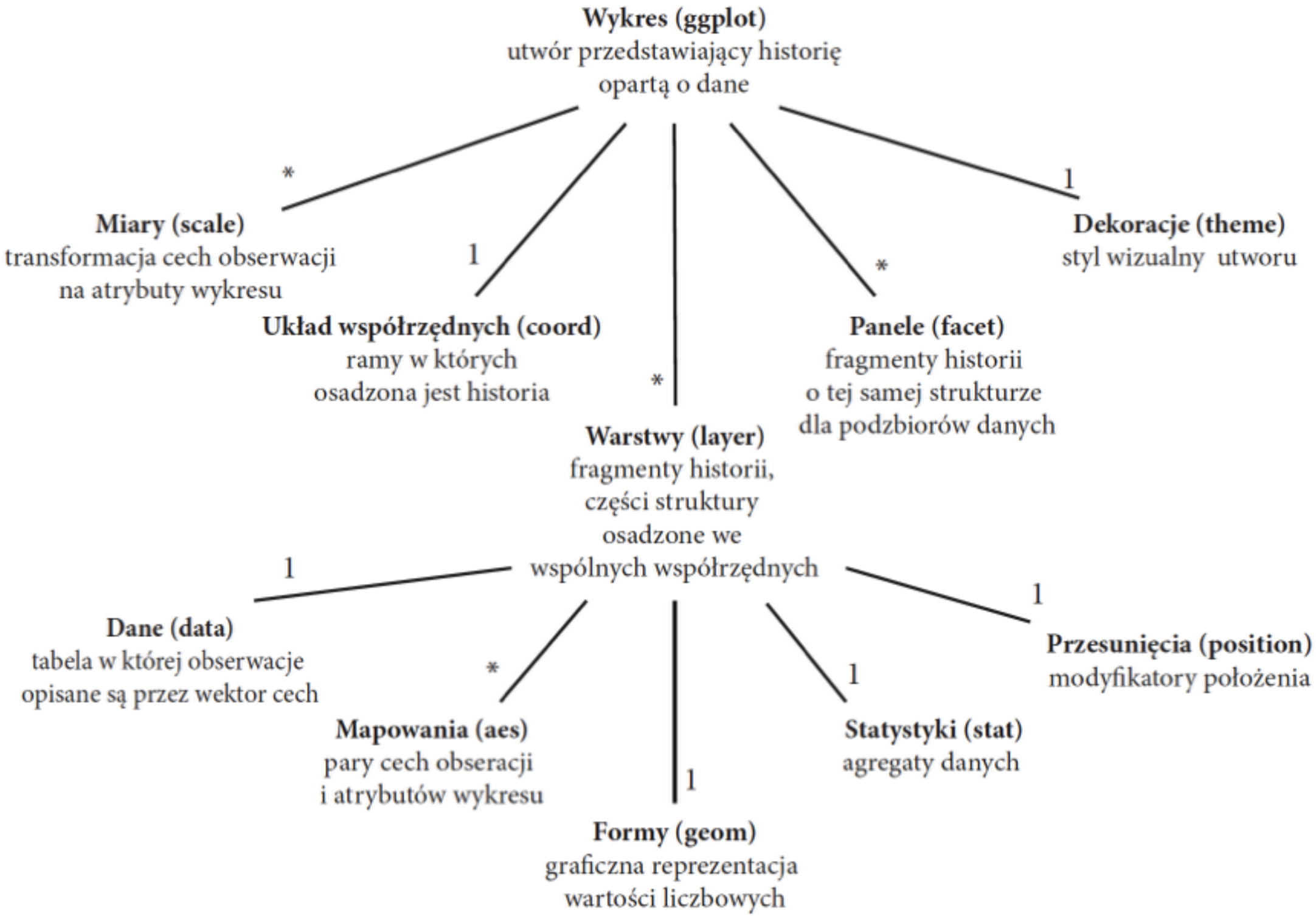
Pakiety:
ggplot2
archivist
ggThemeAssistant

Showcase:

https://rawgit.com/pbiecek/Eseje/master/GamatykaGrafiki.html



http://bit.ly/1Zeu9s6



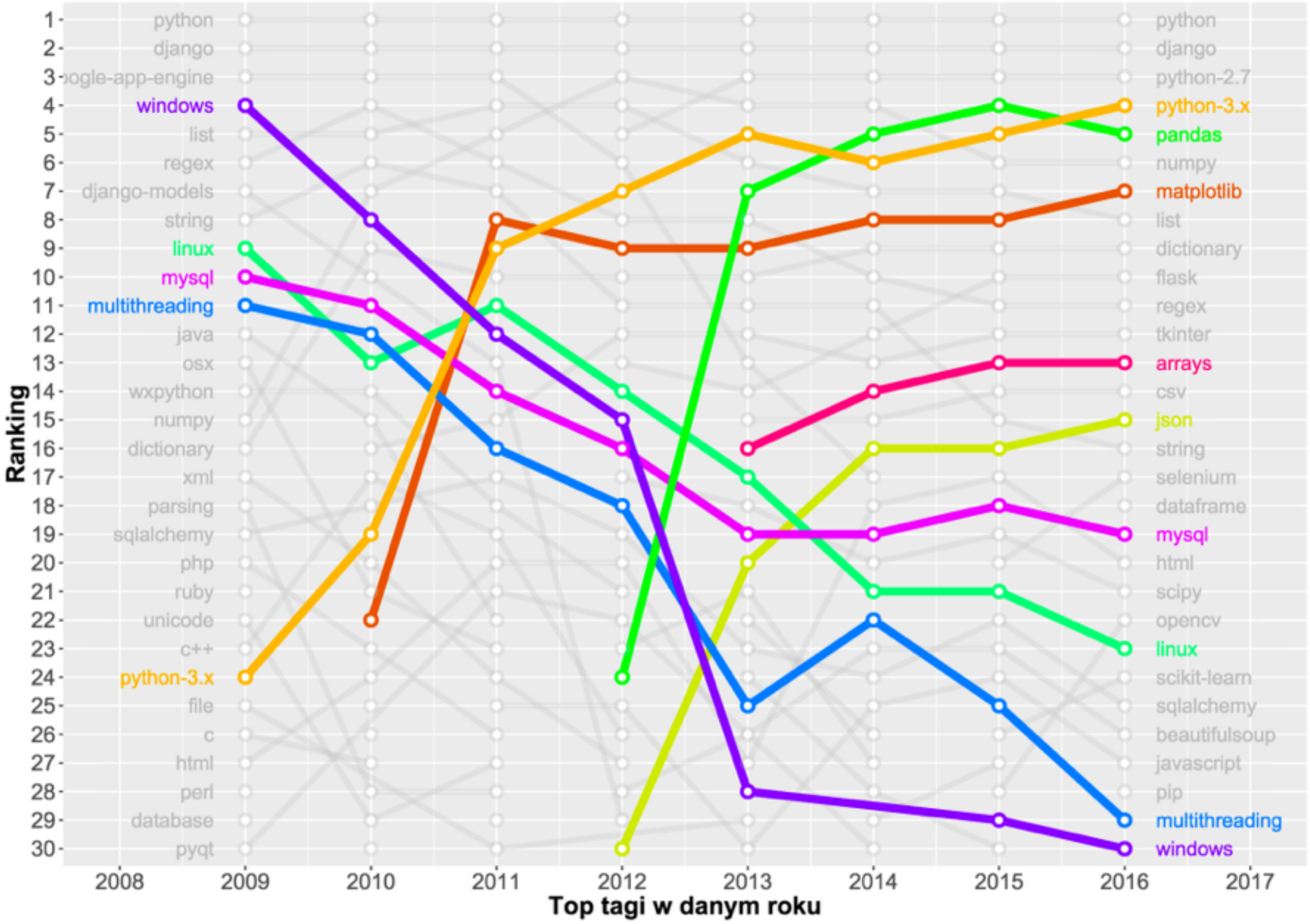
datahero.tech

Analiza rynku pracy dla osób pracujących z danymi:

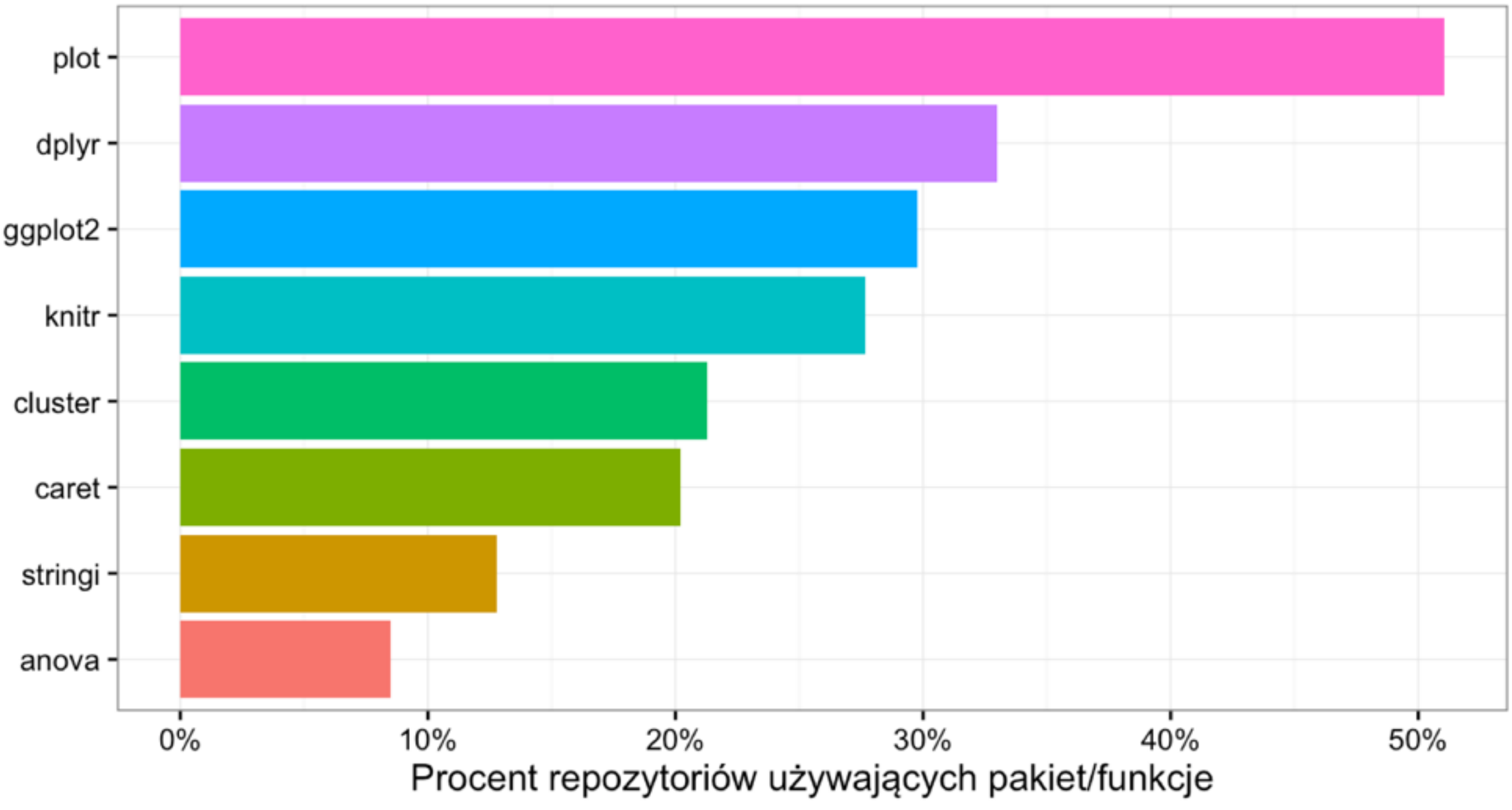
- Olga Mierzwa @ DataHero - analiza StackOverflow

- Konrad Więcko, Krzysztof Słomczyński @ MI2 - zbieranie i analiza pracuj.pl

StackOverflow:



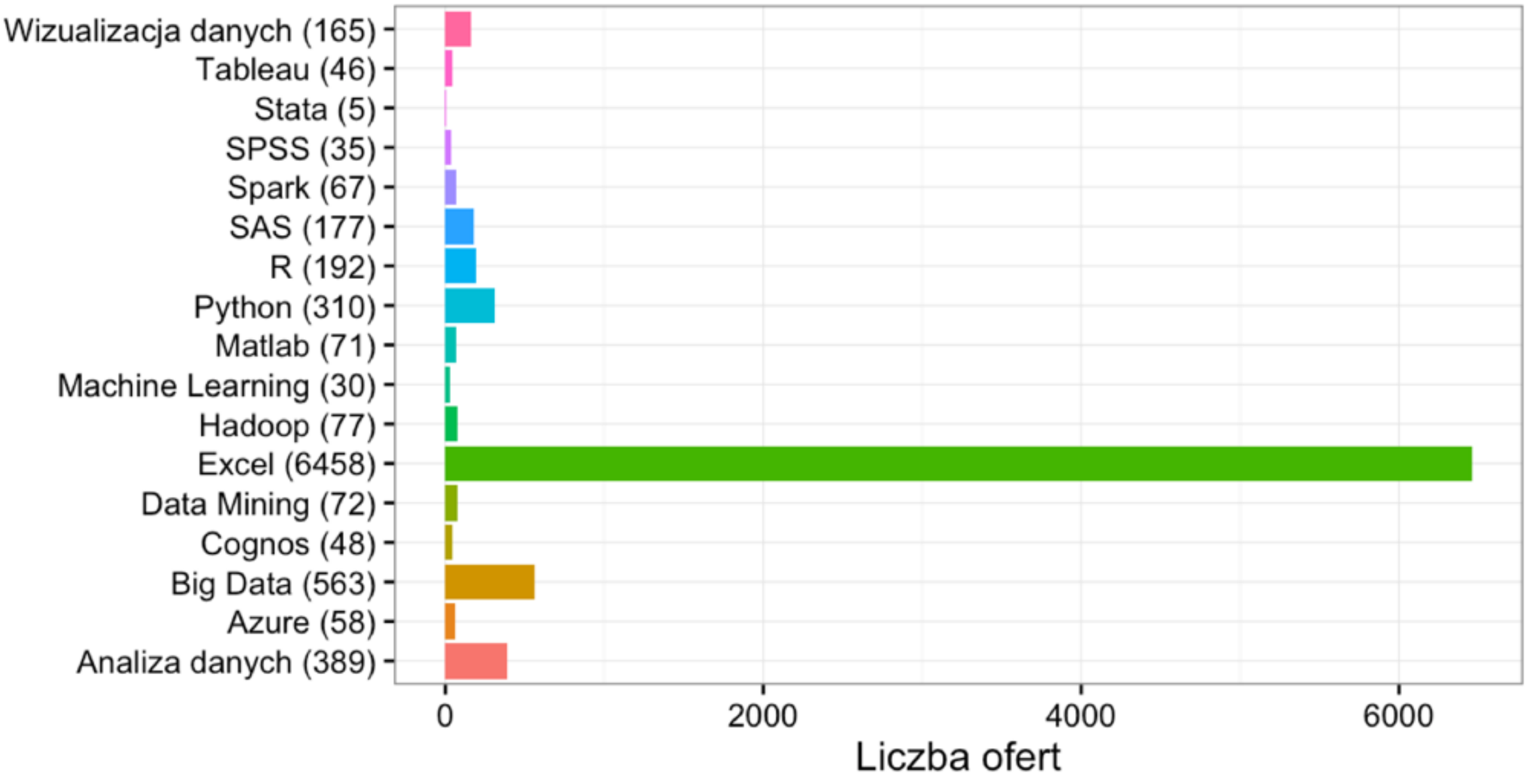
GitHub:
Skanujemy repozytoria R i sprawdzamy kto i jak często korzysta
z określonych funkcji/funkcjonalności.



Analiza danych z pracuj.pl na 1000 ofert dla analityków:

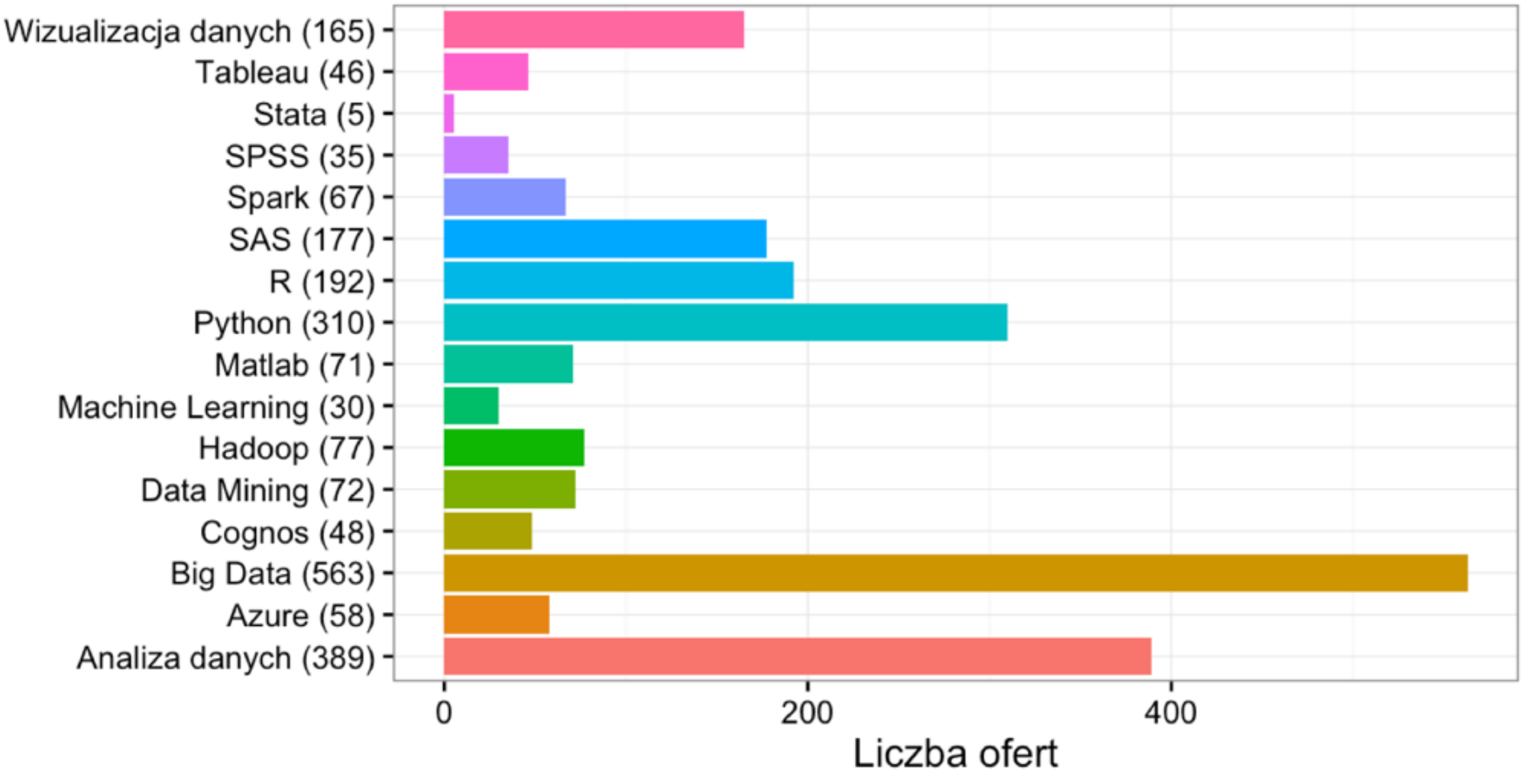
- 326 jest z Warszawy,
- 112 z Krakowa,
- 100 z Wrocławia,
- 62 z Poznania,
- około 35 z Gdańska i Katowic.

Analiza danych z pracuj.pl
Excel jest bardzo popularny w ofertach,
choć w wielu przypadkach nie dotyczy głębszych analiz



Po usunięciu Excela:

- nieznacznie więcej ofert dla R niż dla SAS,
- Matlab, SPSS, Stata mniej popularne
- dużo ofert dla Pythona, w większości dla programistów nie analityków
- Machine Learning, Big Data, takich ofert jest juz całkiem sporo



BECOME DATA HERO

Looking for a job in data world? Subscribe offers newsletter.

Email Address

Subscribe

HIRE DATA HERO

Having troubles finding one?

- Access and hire within the community of data enthusiasts
- Reach out to candidates based on their skills that match your search
- Use big data analytics power and hire from datahero.tech database

Add offer

JOBS BOARD

MACHINE LEARNING ENGINEER

THINKAPPS

San Francisco, USA

machine-learning

Added 15 days ago, 60 opens

SENIOR DATA SCIENTIST

CODEWISE

Kraków, Poland

algorithms spark statistics logistic-regression machine-learning

Added 23 days ago, 275 opens

BIOINFORMATICS, COMPUTATIONAL GENOMICS

CENTRE OF NEW TECHNOLOGIES, UNIVERSITY OF WARSAW

Warsaw, Poland

bayesian-networks bigdata database matlab

Added 26 days ago, 115 opens

DATA SCIENTIST

CODILIME SP. Z O.O.

Warszawa

machine-learning scikit r nlp deep-learning

Added 34 days ago, 124 opens

Więcej informacji:

Odkrywać! Ujawniać! Objaśniać!

http://biecek.pl/Eseje

Information is Beautiful

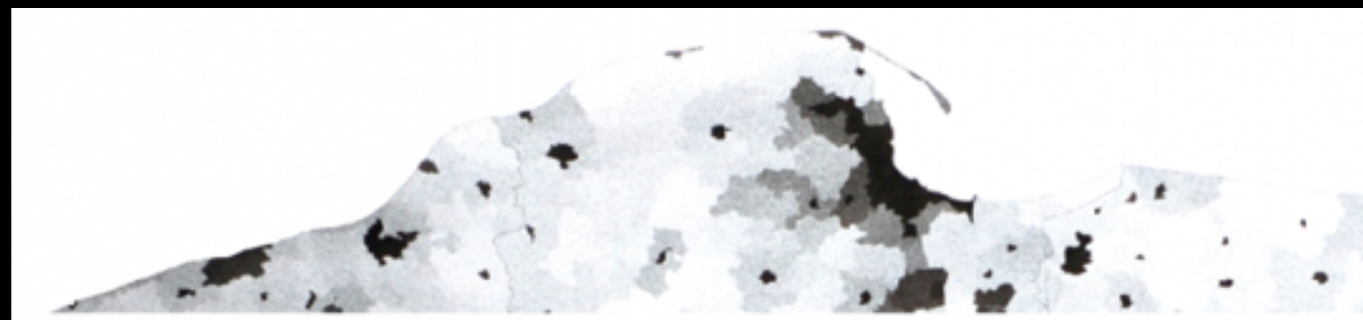
http://www.informationisbeautiful.net/

FlowingData

http://flowingdata.com/

Charts and Things

http://chartsnthings.tumblr.com/

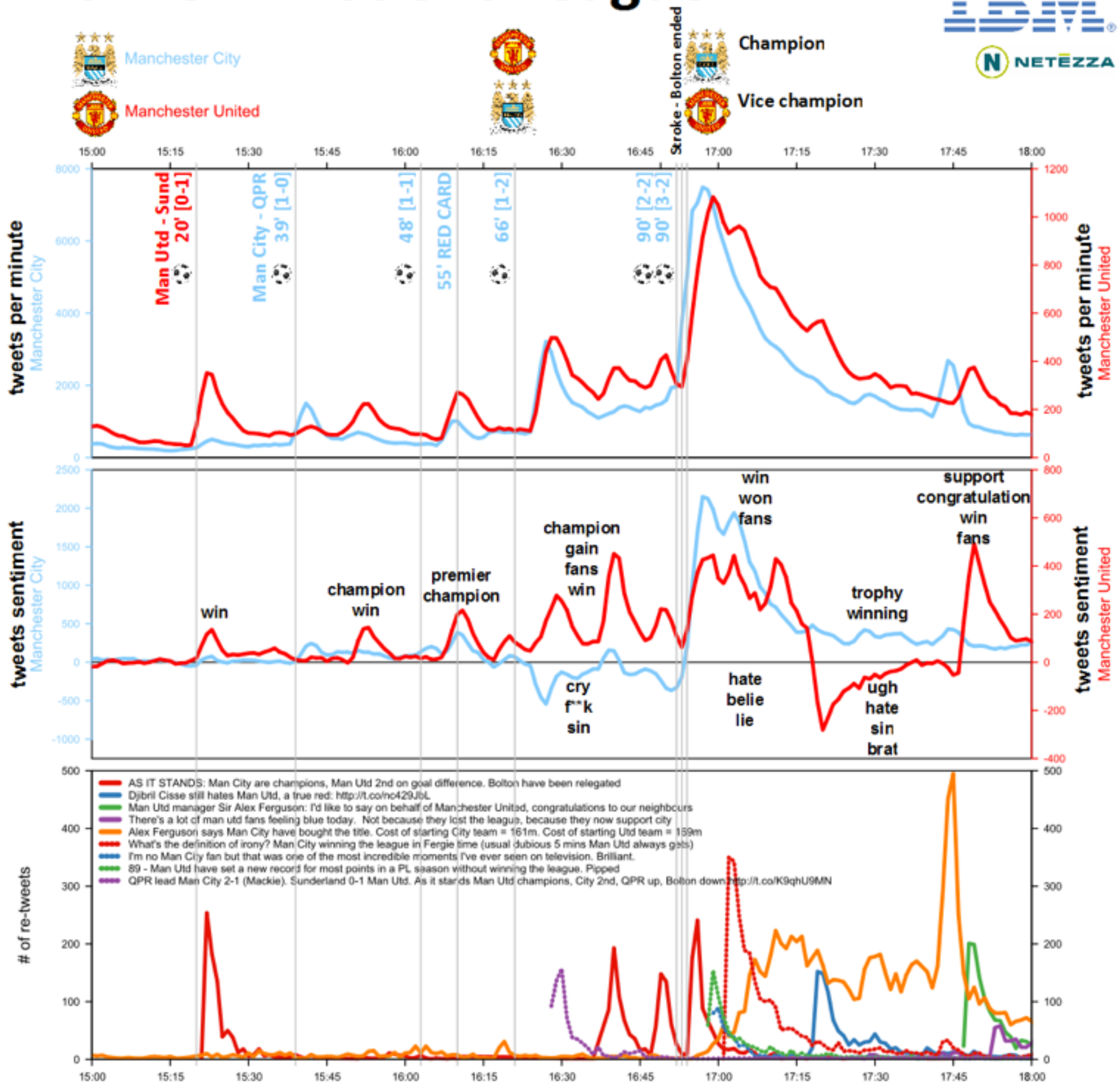


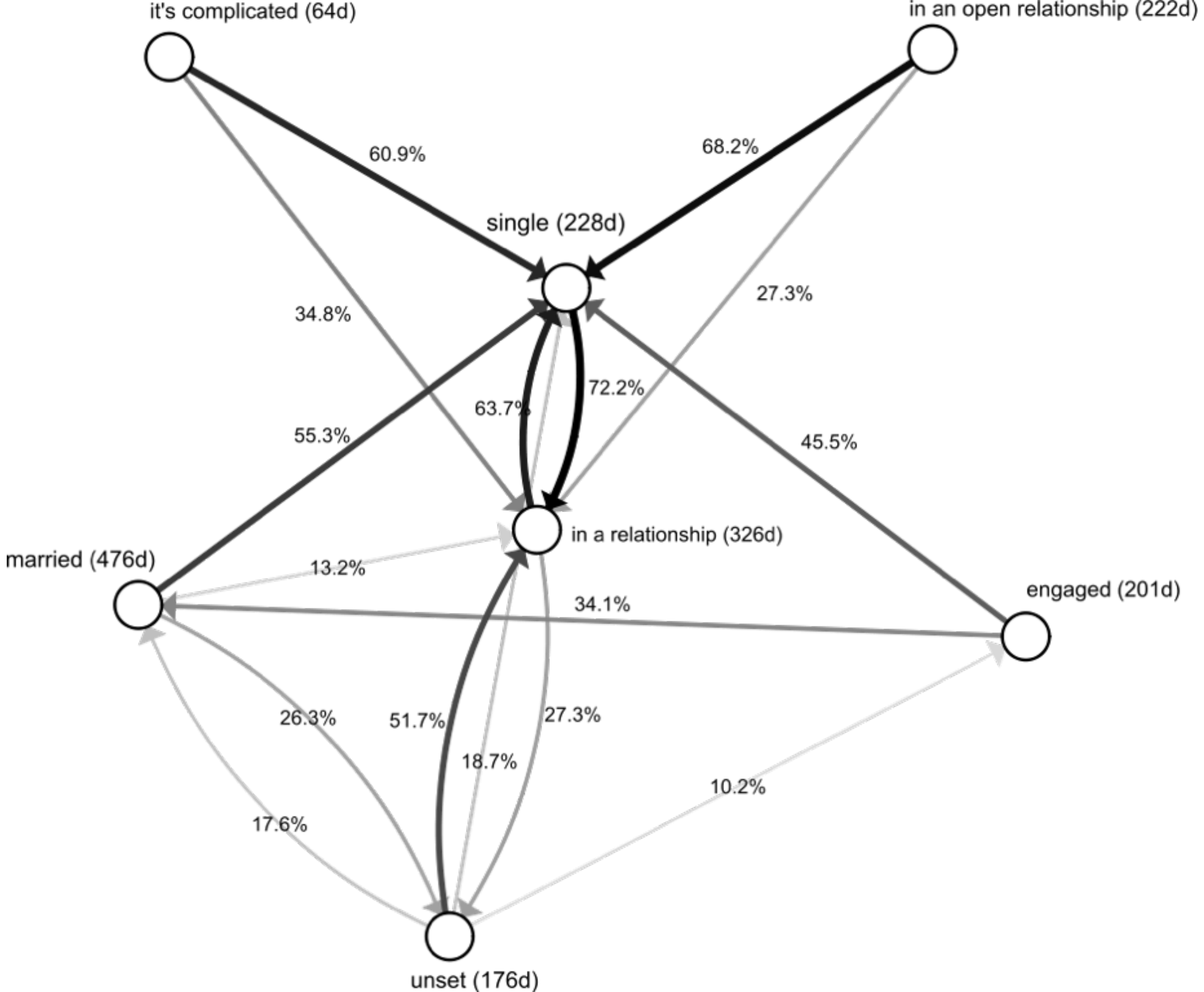
Przemysław Biecek

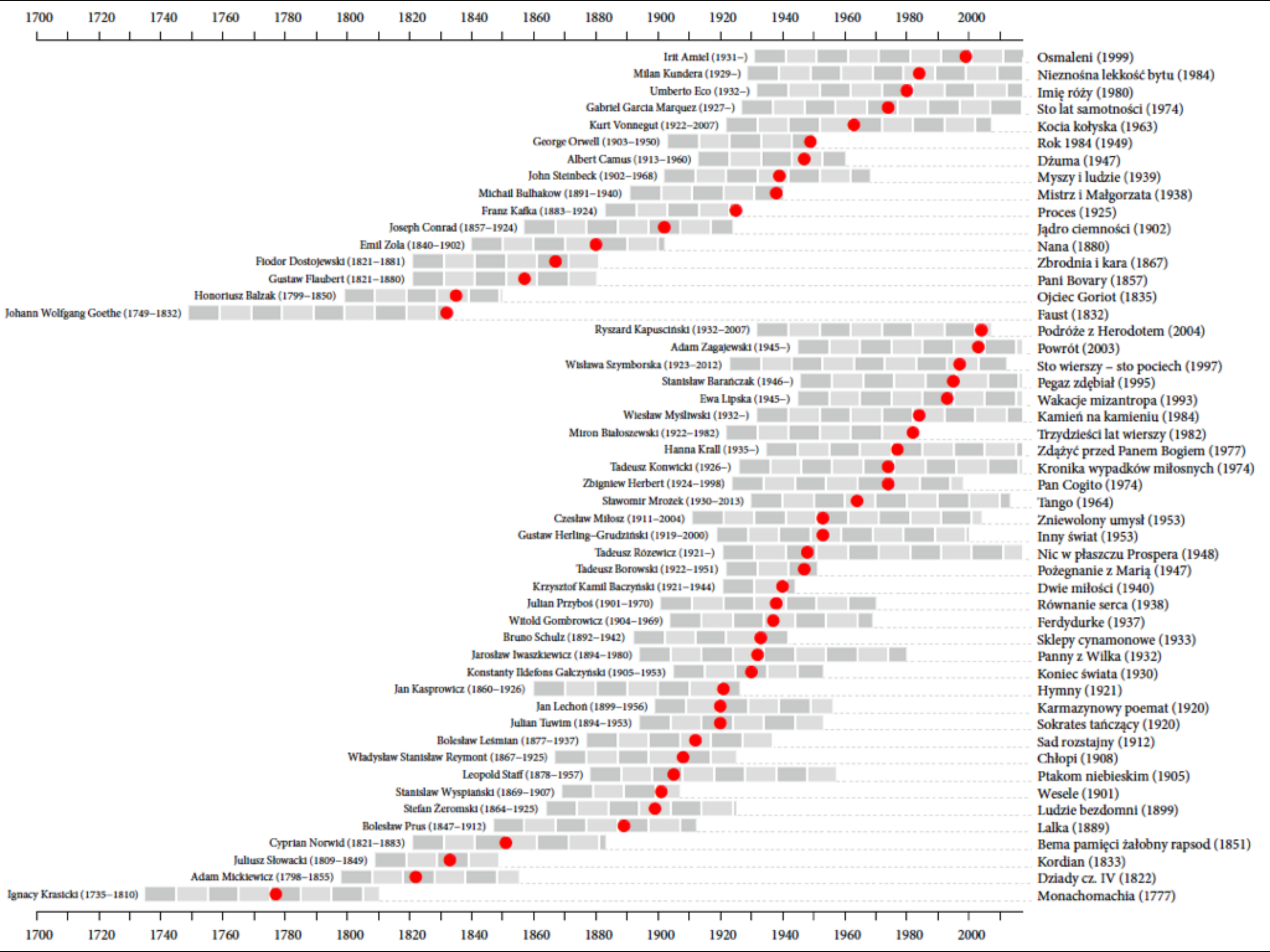
Odkrywać! Ujawniać! Objaśniać!

Zbiór esejów o sztuce prezentowania danych

Premier Emotions League




it's complicated (64d)
in an open relationship (222d)
60.9%
68.2%
single (228d)
34.8%
27.3%
63.7%
72.2%
55.3%
45.5%
in a relationship (326d)
married (476d)
13.2%
engaged (201d)
34.1%
26.3%
51.7%
27.3%
18.7%
10.2%
17.6%
unset (176d)


1700 1720 1740 1760 1780 1800 1820 1840 1860 1880 1900 1920 1940 1960 1980 2000
Irit Amiel (1931–)
Osmaleni (1999)
Milan Kundera (1929–)
Nieznośna lekkość bytu (1984)
Umberto Eco (1932–)
Imię róży (1980)
Gabriel Garcia Marquez (1927–)
Sto lat samotności (1974)
Kurt Vonnegut (1922–2007)
Kocia kołyska (1963)
George Orwell (1903–1950)
Rok 1984 (1949)
Albert Camus (1913–1960)
Dżuma (1947)
John Steinbeck (1902–1968)
Myszy i ludzie (1939)
Michail Bułhakow (1891–1940)
Mistrz i Małgorzata (1938)
Franz Kafka (1883–1924)
Proces (1925)
Joseph Conrad (1857–1924)
Jądro ciemności (1902)
Emil Zola (1840–1902)
Nana (1880)
Fiodor Dostojewski (1821–1881)
Zbrodnia i kara (1867)
Gustaw Flaubert (1821–1880)
Pani Bovary (1857)
Honoriusz Balzak (1799–1850)
Ojciec Goriot (1835)
Johann Wolfgang Goethe (1749–1832)
Faust (1832)
Ryszard Kapuściński (1932–2007)
Podróże z Herodotem (2004)
Adam Zagajewski (1945–)
Powrót (2003)
Wisława Szymborska (1923–2012)
Sto wierszy – sto pociech (1997)
Stanisław Barańczak (1946–)
Pegaz zdębiał (1995)
Ewa Lipska (1945–)
Wakacje mizantropa (1993)
Wiesław Myśliwski (1932–)
Kamień na kamieniu (1984)
Miron Białoszewski (1922–1982)
Trzydzieści lat wierszy (1982)
Hanna Krall (1935–)
Zdążyć przed Panem Bogiem (1977)
Tadeusz Konwicki (1926–)
Kronika wypadków miłosnych (1974)
Zbigniew Herbert (1924–1998)
Pan Cogito (1974)
Sławomir Mrożek (1930–2013)
Tango (1964)
Czesław Miłosz (1911–2004)
Zniewolony umysł (1953)
Gustaw Herling–Grudziński (1919–2000)
Inny świat (1953)
Tadeusz Różewicz (1921–)
Nic w płaszczu Prospera (1948)
Tadeusz Borowski (1922–1951)
Pożegnanie z Marią (1947)
Krzysztof Kamil Baczyński (1921–1944)
Dwie miłości (1940)
Julian Przyboś (1901–1970)
Równanie serca (1938)
Witold Gombrowicz (1904–1969)
Ferdydurke (1937)
Bruno Schulz (1892–1942)
Sklepy cynamonowe (1933)
Jarosław Iwaszkiewicz (1894–1980)
Panny z Wilka (1932)
Konstanty Ildefons Gałczyński (1905–1953)
Koniec świata (1930)
Jan Kasprowicz (1860–1926)
Hymny (1921)
Jan Lechoń (1899–1956)
Karmazynowy poemat (1920)
Julian Tuwim (1894–1953)
Sokrates tańczący (1920)
Bolesław Leśmian (1877–1937)
Sad rozstajny (1912)
Władysław Stanisław Reymont (1867–1925)
Chłopi (1908)
Leopold Staff (1878–1957)
Ptakom niebieskim (1905)
Stanisław Wyspiański (1869–1907)
Wesele (1901)
Stefan Żeromski (1864–1925)
Ludzie bezdomni (1899)
Bolesław Prus (1847–1912)
Lalka (1889)
Cyprian Norwid (1821–1883)
Bema pamięci żałobny rapsod (1851)
Juliusz Słowacki (1809–1849)
Kordian (1833)
Adam Mickiewicz (1798–1855)
Dziady cz. IV (1822)
Ignacy Krasicki (1735–1810)
Monachomachia (1777)
1700 1720 1740 1760 1780 1800 1820 1840 1860 1880 1900 1920 1940 1960 1980 2000

Pytania?