

O'REILLY®

Podstawy wizualizacji danych

Zasady tworzenia
atrakcyjnych wykresów



Helion 

Claus O. Wilke

Tytuł oryginału: Fundamentals of Data Visualization: A Primer on Making Informative and Compelling Figures

Tłumaczenie: Leszek Sagalara

ISBN: 978-83-283-6126-3

© 2020 Helion SA

Authorized Polish translation of the English edition of Fundamentals of Data Visualization ISBN 9781492031086 © 2019 Claus O. Wilke

This translation is published and sold by permission of O'Reilly Media, Inc., which owns or controls all rights to publish and sell the same.

The O'Reilly logo is a registered trademark of O'Reilly Media, Inc. Fundamentals of Data Visualization, the cover image, and related trade dress are trademarks of O'Reilly Media, Inc.

All rights reserved. No part of this book may be reproduced or transmitted in any form or by any means, electronic or mechanical, including photocopying, recording or by any information storage retrieval system, without permission from the Publisher.

Wszelkie prawa zastrzeżone. Nieautoryzowane rozpowszechnianie całości lub fragmentu niniejszej publikacji w jakiegokolwiek postaci jest zabronione. Wykonywanie kopii metodą kserograficzną, fotograficzną, a także kopiowanie książki na nośniku filmowym, magnetycznym lub innym powoduje naruszenie praw autorskich niniejszej publikacji.

Wszystkie znaki występujące w tekście są zastrzeżonymi znakami firmowymi bądź towarowymi ich właścicieli.

Autor oraz Helion SA dołożyli wszelkich starań, by zawarte w tej książce informacje były kompletne i rzetelne. Nie biorą jednak żadnej odpowiedzialności ani za ich wykorzystanie, ani za związane z tym ewentualne naruszenie praw patentowych lub autorskich. Autor oraz Helion SA nie ponoszą również żadnej odpowiedzialności za ewentualne szkody wynikłe z wykorzystania informacji zawartych w książce.

Helion SA

ul. Kościuszki 1c, 44-100 Gliwice

tel. 32 231 22 19, 32 230 98 63

e-mail: helion@helion.pl

WWW: <http://helion.pl> (księgarnia internetowa, katalog książek)

Drogi Czytelniku!

Jeżeli chcesz ocenić tę książkę, zajrzyj pod adres

<http://helion.pl/user/opinie/powida>

Możesz tam wpisać swoje uwagi, spostrzeżenia, recenzję.

Printed in Poland.

- Kup książkę
- Poleć książkę
- Oceń książkę

- Księgarnia internetowa
- [Lubię to!](#) » [Nasza społeczność](#)

Spis treści

Wstęp	9
1. Wprowadzenie	15
Brzydkie, złe i błędne wykresy	16
<hr/>	
Część I. Od danych do wizualizacji	19
2. Wizualizacja danych — estetyczne odwzorowywanie danych	21
Estetyka a typy danych	21
Estetyczne odwzorowywanie wartości danych przy użyciu skal	24
3. Układy współrzędnych i osie	27
Układ współrzędnych kartezjańskich	27
Osie nieliniowe	30
Krzywoliniowe układy współrzędnych	34
4. Skale kolorów	37
Kolor jako narzędzie do odróżniania	37
Kolory jako reprezentacja wartości danych	39
Kolor jako narzędzie do wyróżniania	41
5. Spis wizualizacji	45
Wielkości	45
Rozkłady	46
Proporcje	47
Relacje $x - y$	48
Dane geoprzestrzenne	49
Niepewność	49

6. Wizualizacja wielkości	51
Wykresy słupkowe	51
Słupki zgrupowane i stosowe	55
Wykresy kropkowe i mapy cieplne	58
7. Wizualizacja rozkładów: histogramy i wykresy gęstości	63
Wizualizacja pojedynczego rozkładu	63
Jednoczesna wizualizacja wielu rozkładów	68
8. Wizualizacja rozkładów: dystrybuanty empiryczne i wykresy Q-Q	73
Dystrybuanty empiryczne	73
Rozkłady o wysokim współczynniku skośności	76
Wykresy kwantylowo-kwantylowe	79
9. Wizualizacja wielu rozkładów jednocześnie	83
Wizualizacja rozkładów wzdłuż osi pionowej	83
Wizualizacja rozkładów wzdłuż osi poziomej	88
10. Wizualizacja proporcji	93
Przypadek dla diagramów kołowych	93
Przypadek dla słupków sąsiadujących	96
Przypadek dla słupków stosowych i gęstości stosowych	98
Oddzielne wizualizacje proporcji jako części składowych całości	99
11. Wizualizacja zagnieżdżonych proporcji	103
Błędna prezentacja zagnieżdżonych proporcji	103
Wykresy mozaikowe i mapy drzewa	104
Zagnieżdżone diagramy kołowe	108
Zestawy równoległe	110
12. Wizualizacja powiązań między zmiennymi ilościowymi	113
Wykresy punktowe	113
Korelogramy	116
Redukcja wymiarów	119
Dane sparowane	122
13. Wizualizacja szeregów czasowych i innych funkcji zmiennej niezależnej	125
Pojedyncze szeregi czasowe	125
Wiele szeregów czasowych i krzywe dawka-odpowiedź	128
Szeregi czasowe dwóch lub więcej zmiennych zależnych	130

14. Wizualizacja trendów	135
Wygładzanie	135
Pokazywanie trendów za pomocą zdefiniowanej postaci funkcyjnej	140
Usuwanie trendu i dekompozycja szeregu czasowego	143
15. Wizualizacja danych geoprzestrzennych	149
Odwzorowania	149
Warstwy	155
Kartogramy	158
Anamorfozy	161
16. Wizualizacja niepewności	165
Ramowanie prawdopodobieństw jako częstotliwości	165
Wizualizacja niepewności estymacji punktowych	169
Wizualizacja niepewności dopasowań krzywych	179
Wykresy hipotetycznych wyników	181

Część II. Zasady projektowania wykresów **185**

17. Zasada proporcjonalnego atramentu	187
Wizualizacje wzdłuż osi liniowych	187
Wizualizacje wzdłuż osi logarytmicznych	192
Bezpośrednie wizualizacje obszarów	194
18. Postępowanie z nakładającymi się punktami	197
Częściowa przezroczystość i wibrowanie	197
Histogramy 2D	200
Izolinie	203
19. Częste niebezpieczeństwa powodowane używaniem kolorów	209
Kodowanie zbyt wielu lub nieistotnych informacji	209
Używanie niemonotonicznych skal kolorów do kodowania wartości danych	212
Niedostosowanie do osób z zaburzeniami rozpoznawania barw	214
20. Kodowanie nadmiarowe	219
Projektowanie wykresów z kodowaniem nadmiarowym	219
Projektowanie wykresów bez legendy	224
21. Wykresy wielopanelowe	229
Małe wielokrotności	229
Wykresy zestawione	234

22. Tytuły, podpisy i tabele	239
Tytuły i podpisy wykresu	239
Tytuły osi i legend	240
Tabele	244
23. Zachowanie równowagi między danymi i kontekstem	247
Zapewnienie odpowiedniej ilości kontekstu	247
Siatka tła	252
Dane połączone w pary	256
Podsumowanie	258
24. Używaj większych etykiet osi	259
25. Unikaj rysowania konturów	263
26. Nie idź w 3D	271
Unikaj nieuzasadnionej trójwymiarowości	271
Unikaj trójwymiarowych skali położenia	273
Właściwe zastosowanie wizualizacji 3D	278
<hr/>	
Część III. Inne zagadnienia	281
27. Najpopularniejsze formaty plików graficznych	283
Bitmapa a grafika wektorowa	283
Bezstratna i stratna kompresja grafiki bitmapowej	285
Konwersja pomiędzy formatami obrazów	287
28. Wybór odpowiedniego oprogramowania wizualizacyjnego	289
Odtwarzalność i powtarzalność	289
Analiza danych kontra prezentacja danych	291
Oddzielenie zawartości od projektu	293
29. Opowiadanie historii i wysuwanie wniosków	297
Czym jest historia?	297
Twórz wykresy dla generałów	300
Przechodź stopniowo do bardziej złożonych wykresów	304
Spraw, aby Twoje wykresy zapadały w pamięć	304
Zachowaj spójność, ale nie powtarzaj się	307
<hr/>	
Dodatki	311
Bibliografia	313
Źródła	317

Wizualizacja wielkości

W wielu scenariuszach interesuje nas wielkość jakiegoś zbioru liczb. Przykładowo możemy wizualizować łączną wielkość sprzedaży różnych marek samochodów, łączną liczbę osób mieszkających w różnych miastach lub wiek olimpijczyków uprawiających różne sporty. We wszystkich tych przypadkach mamy zestaw kategorii (np. marki samochodów, miasta czy dyscypliny sportowe) oraz wartość ilościową dla każdej kategorii. Określam te przypadki jako wizualizacje wielkości, ponieważ w tych wizualizacjach główny nacisk będzie położony na wielkość wartości ilościowych. Standardową wizualizacją w tym scenariuszu jest wykres słupkowy, który posiada kilka wariantów, w tym słupki proste, zgrupowane i ułożone w stosy. Alternatywą dla wykresu słupkowego jest wykres kropkowy i mapa cieplna.

Wykresy słupkowe

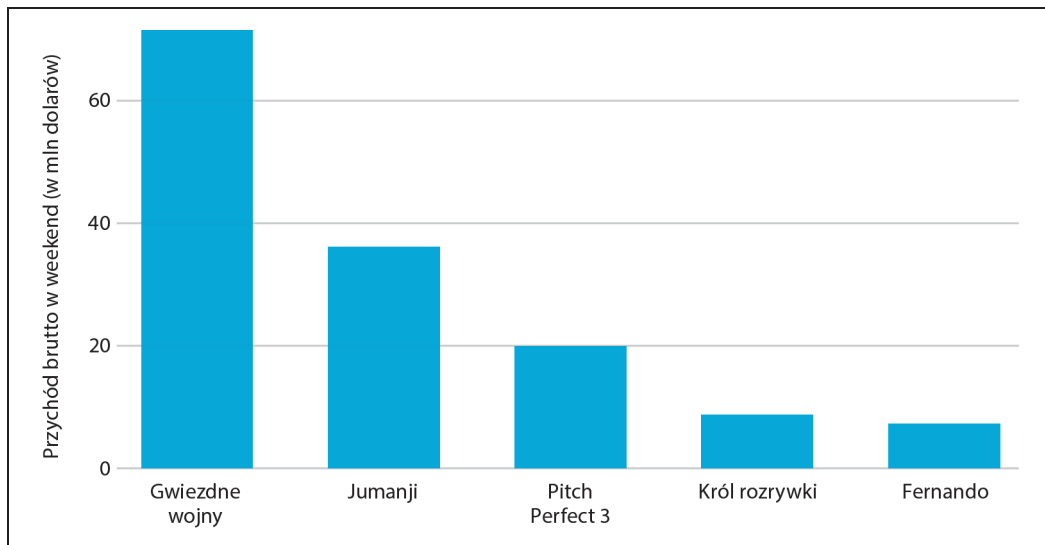
Aby uatrakcyjnić koncepcję wykresu słupkowego, rozważmy całkowitą sprzedaż biletów na najpopularniejsze filmy z danego weekendu. W tabeli 6.1 przedstawiono pięć filmów najlepiej sprzedających się podczas weekendu poprzedzającego Boże Narodzenie w 2017 roku. *Gwiezdne wojny: Ostatni Jedi* był zdecydowanie najpopularniejszym filmem tego weekendu, wyprzedził filmy znajdujące się na czwartym i piątym miejscu, *Król rozrywki* i *Fernando*, prawie dziesięciokrotnie.

Tabela 6.1. Najbardziej kasowe filmy podczas weekendu 22 – 24 grudnia 2017 r.

Źródło danych: Box Office Mojo (<http://www.boxofficemojo.com>). Wykorzystano za ich zgodą

Pozycja	Tytuł	Przychód brutto w weekend (w dolarach)
1	<i>Gwiezdne wojny: Ostatni Jedi</i>	71 565 498
2	<i>Jumanji: Przygoda w dżungli</i>	36 169 328
3	<i>Pitch Perfect 3</i>	19 928 525
4	<i>Król rozrywki</i>	8 805 843
5	<i>Fernando</i>	7 316 746

Ten rodzaj danych jest powszechnie wizualizowany za pomocą pionowych słupków. Dla każdego filmu tworzymy słupek, który zaczyna się od zera i rozciąga aż do wartości pieniężnej określającej przychód brutto ze sprzedaży biletów na dany film w czasie weekendu (rysunek 6.1). Wizualizacja ta nazywana jest **wykresem słupkowym** lub **diagramem słupkowym**.



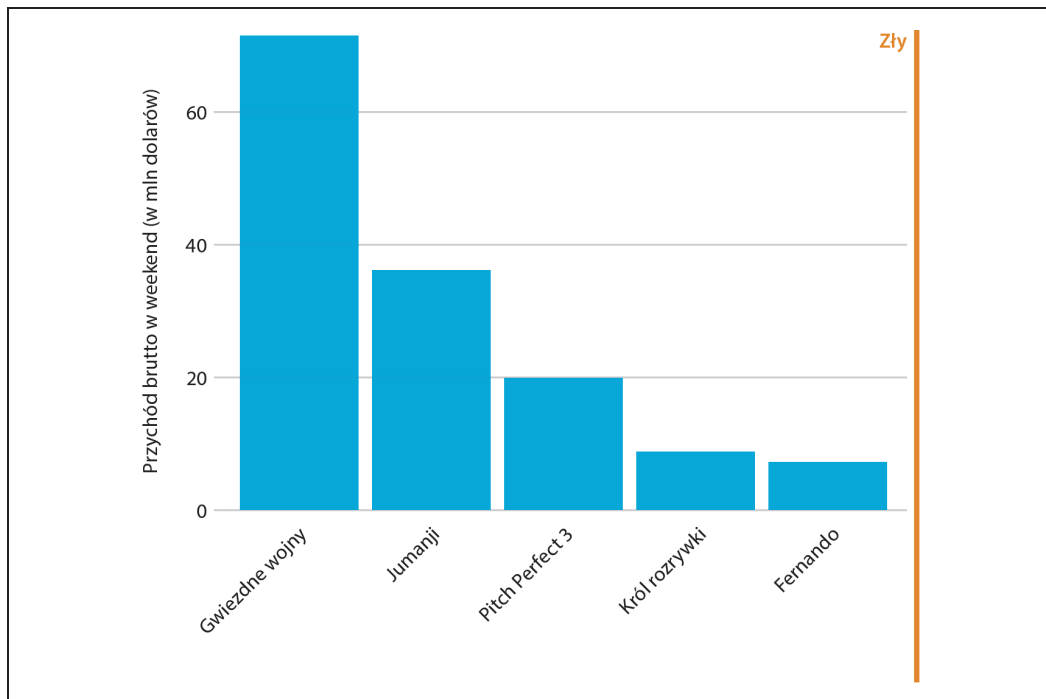
Rysunek 6.1. Najbardziej kasowe filmy podczas weekendu 22 – 24 grudnia 2017 r. przedstawione jako wykres słupkowy. Źródło danych: Box Office Mojo (<http://www.boxofficemojo.com>). Wykorzystano za ich zgodą

Jednym z problemów, z którymi często spotykamy się w przypadku pionowych słupków, jest to, że etykiety identyfikujące każdy słupek zajmują dużo przestrzeni poziomej. W rzeczywistości rysunek 6.1 musiałem wykonać dość szeroko i rozsunąć słupki, żeby zmieścić pod nimi tytuły filmów. Aby zaoszczędzić miejsce w poziomie, mogliśmy umieścić słupki bliżej siebie i obrócić etykiety (rysunek 6.2). Nie jestem jednak wielkim zwolennikiem obracanych etykiet. Powstałe w ten sposób wykresy uważam za niezręczne i trudne do odczytania. A z mojego doświadczenia wynika, że gdy etykiety są zbyt długie, by można je było umieścić poziomo, nie wyglądają też dobrze obrócone.

Lepszym rozwiązaniem dla długich etykiet jest zazwyczaj zamiana osi x i y , tak aby słupki ułożone były poziomo (rysunek 6.3). Po zamianie osi otrzymujemy zwartą formę, w której wszystkie elementy wizualne, w tym całość tekstu, są zorientowane poziomo. W rezultacie rysunek jest znacznie łatwiejszy do odczytania niż rysunek 6.2 czy nawet rysunek 6.1.

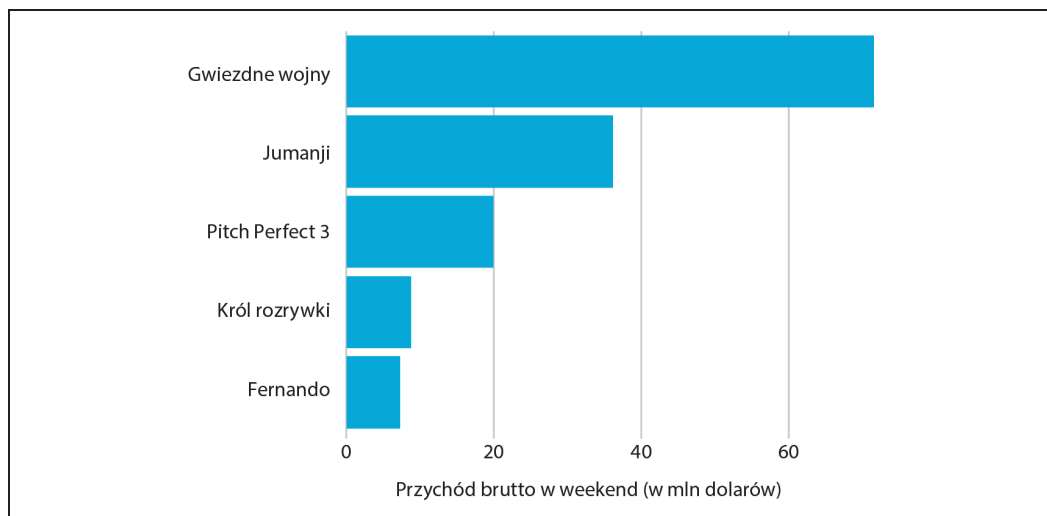
Niezależnie od tego, czy ustawimy słupki pionowo, czy poziomo, musimy zwrócić uwagę na kolejność ich ułożenia. Często widzę wykresy słupkowe, na których słupki są ułożone arbitralnie lub według jakiegoś kryterium, które nie ma znaczenia w kontekście wykresu. Niektóre programy do tworzenia wykresów domyślnie układają słupki w kolejności alfabetycznej etykiet; możliwe są też inne, podobnie arbitralne ustawienia (rysunek 6.4). Ogólnie rzecz biorąc, tego rodzaju wykresy są bardziej mylące i mniej intuicyjne niż wykresy, w których słupki są ułożone zgodnie z ich wielkością.

Powinniśmy zmieniać rozmieszczenie słupków tylko wtedy, gdy nie ma naturalnego uporządkowania reprezentowanych przez nie kategorii. Ilekroć istnieje naturalna kolejność (tj. gdy nasza zmienna kategoria jest czynnikiem uporządkowanym), powinniśmy zachować tę kolejność w wizualizacji. Dla przykładu rysunek 6.5 przedstawia medianę rocznych dochodów w USA według grup wiekowych. W tym przypadku słupki powinny być ułożone w kolejności rosnącego wieku. Sortowanie według wysokości słupków, przy jednoczesnym wymieszaniu grup wiekowych, nie ma sensu (rysunek 6.6).



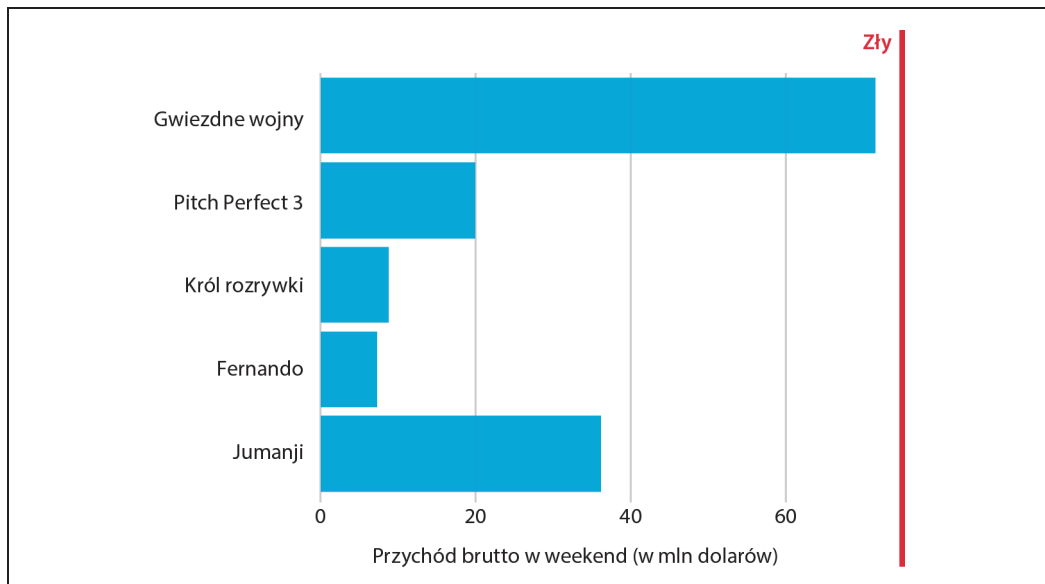
Rysunek 6.2. Najbardziej dochodowe filmy podczas weekendu 22 – 24 grudnia 2017 r. przedstawione jako wykres słupkowy z obróconymi etykietami wskaźników osi. Obrócone etykiety wskaźników osi są zazwyczaj trudne do odczytania i wymagają niewygodnego wykorzystania miejsca pod wykresem. Z tych powodów generalnie uważam, że wykresy z obróconymi etykietami wskaźników są brzydkie.

Źródło danych: Box Office Mojo (<http://www.boxofficemojo.com>). Wykorzystano za ich zgodą



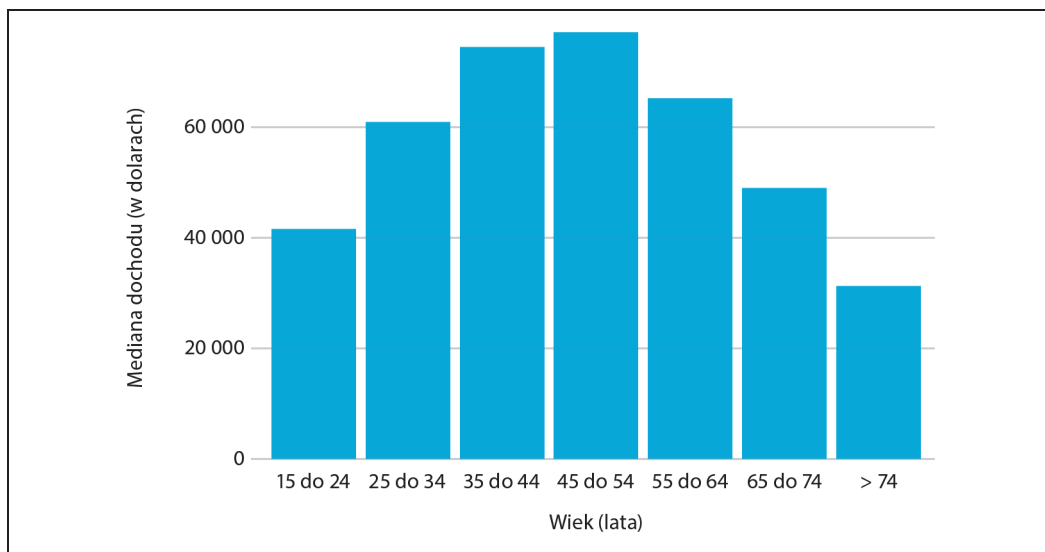
Rysunek 6.3. Najbardziej dochodowe filmy podczas weekendu 22 – 24 grudnia 2017 r. przedstawione w formie poziomego wykresu słupkowego. Źródło danych: Box Office Mojo (<http://www.boxofficemojo.com>).

Wykorzystano za ich zgodą

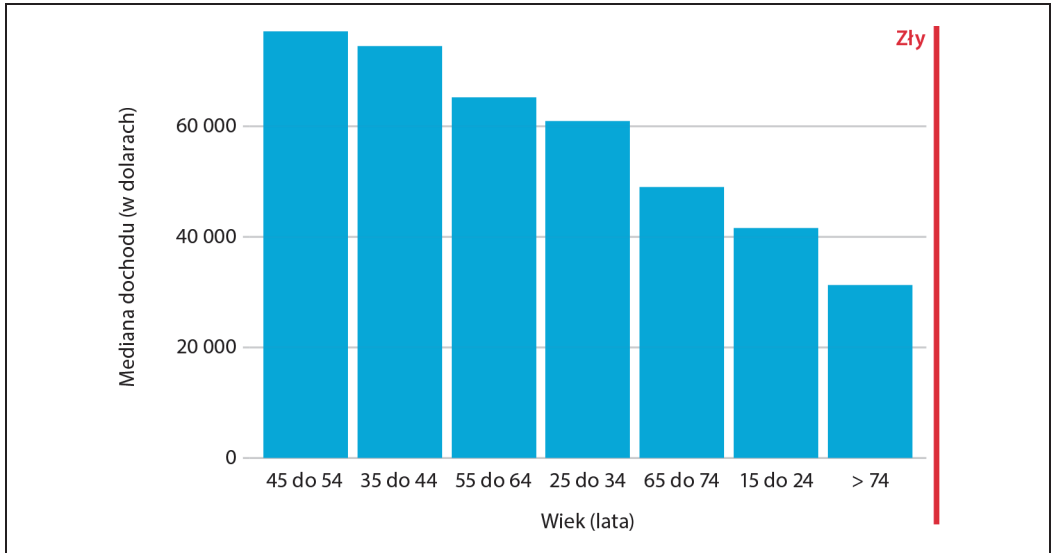


Rysunek 6.4. Najbardziej dochodowe filmy podczas weekendu 22 – 24 grudnia 2017 r. przedstawione w formie poziomego wykresu słupkowego. Tutaj słupki zostały umieszczone w kolejności malejącej pod względem długości tytułów filmów. Taki układ słupków jest arbitralny, nie służy sensownemu celowi i sprawia, że powstały wykres jest znacznie mniej intuicyjny niż rysunek 6.3.

Źródło danych: Box Office Mojo (<http://www.boxofficemojo.com>). Wykorzystano za ich zgodą



Rysunek 6.5. Mediana dochodu gospodarstw domowych w USA w 2016 r. w stosunku do grup wiekowych. Grupa wiekowa 45 – 54 ma najwyższy średni dochód. Źródło danych: US Census Bureau



Rysunek 6.6. Mediana dochodu gospodarstw domowych w USA w 2016 r. w stosunku do grup wiekowych, posortowana według dochodów. Choć ta kolejność słupków wygląda atrakcyjnie wizualnie, kolejność grup wiekowych jest teraz myląca. Źródło danych: US Census Bureau

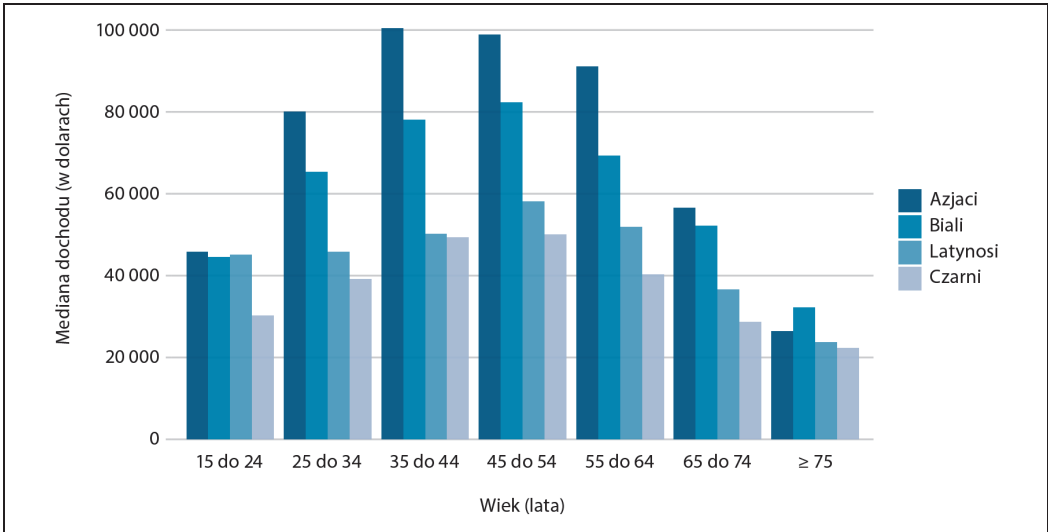


Zwracaj uwagę na kolejność słupków. Jeśli reprezentują kategorie nieuporządkowane, należy je uszeregować według rosnących lub malejących wartości danych.

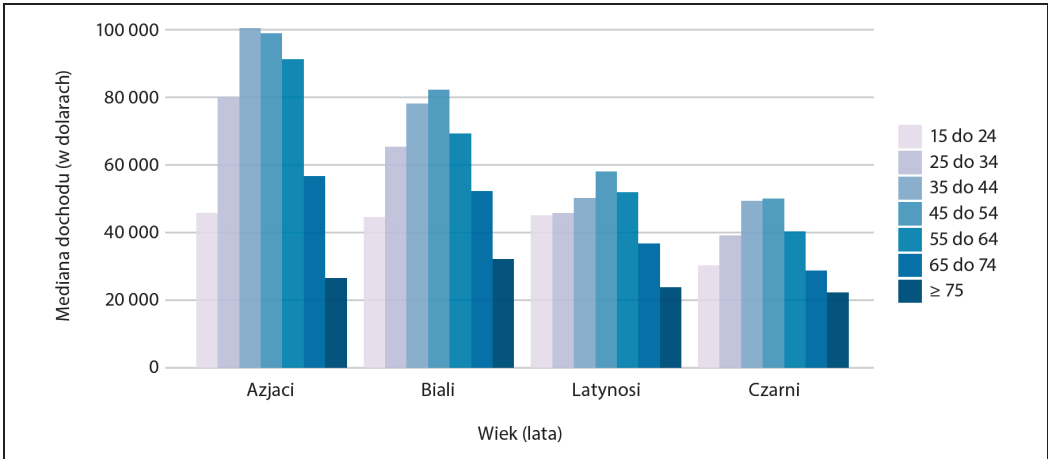
Słupki zgrupowane i stosowe

Wszystkie przykłady z poprzedniego podrozdziału pokazywały, w jaki sposób wielkość ilościowa różniła się w odniesieniu do jednej zmiennej kategorialnej. Często jednak jesteśmy zainteresowani jednocześnie dwiema zmiennymi kategorialnymi. Dla przykładu US Census Bureau podaje medianę poziomów dochodów w rozbiciu na wiek i rasę. Możemy zwizualizować ten zbiór danych za pomocą **zgrupowanego wykresu słupkowego** (rysunek 6.7). W zgrupowanym wykresie słupkowym tworzymy grupę słupków na każdej pozycji wzdłuż osi x , określonej przez jedną zmienną kategorialną, a następnie rysujemy słupki w obrębie każdej grupy zgodnie z inną zmienną kategorialną.

Zgrupowane wykresy słupkowe pokazują wiele informacji jednocześnie i mogą być dezorientujące. W rzeczywistości, mimo że na rysunku 6.7 nie oznaczyłem wykresu jako zły lub brzydki, trudno mi go odczytać. Szczególnie trudno porównać średnie dochody w różnych grupach wiekowych dla danej grupy rasowej. Tak więc wykres ten jest odpowiedni tylko wtedy, gdy interesują nas przede wszystkim różnice w poziomie dochodów między grupami rasowymi, oddzielnie dla poszczególnych grup wiekowych. Jeśli bardziej zależy nam na ogólnym wzorcu poziomu dochodów w poszczególnych grupach rasowych, lepiej przedstawić rasę wzdłuż osi x , a wiek jako osobne słupki w obrębie każdej grupy rasowej (rysunek 6.8).

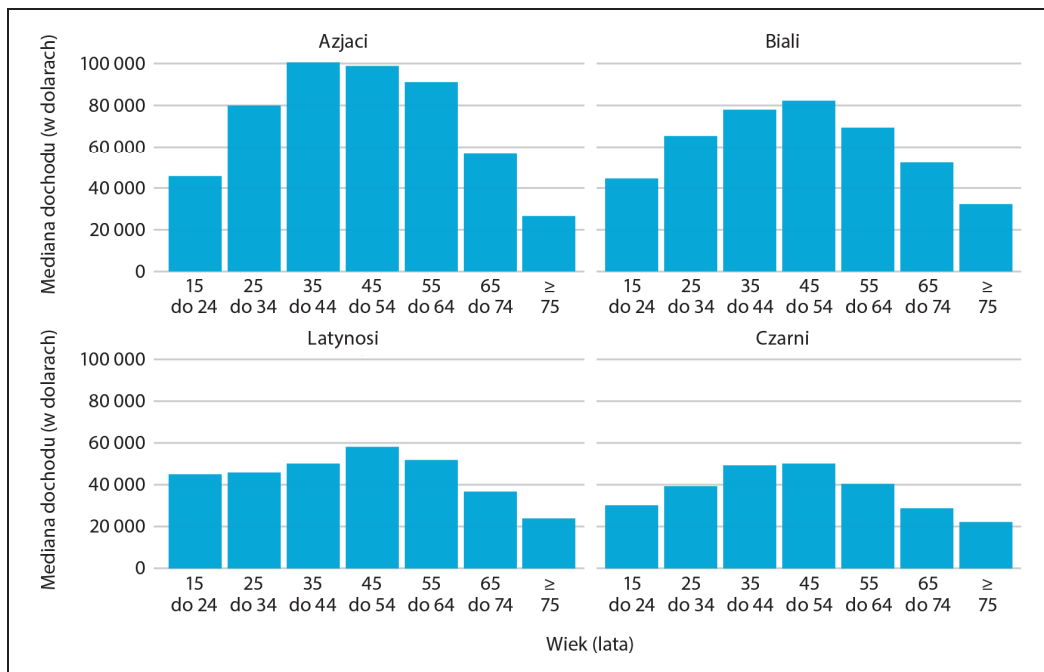


Rysunek 6.7. Mediana dochodu gospodarstw domowych w USA w 2016 r. w stosunku do grup wiekowych i ras. Grupy wiekowe przedstawione są wzdłuż osi x, a dla każdej grupy wiekowej istnieją cztery słupki odpowiadające medianie dochodu odpowiednio Azjatów, Białych, Latynosów i Czarnych. Źródło danych: US Census Bureau



Rysunek 6.8. Mediana dochodu gospodarstw domowych w USA w 2016 r. w stosunku do grup wiekowych i ras. W przeciwieństwie do rysunku 6.7, teraz rasa jest przedstawiona wzdłuż osi x, a dla każdej rasy pokazujemy siedem słupków odpowiadających siedmiu grupom wiekowym. Źródło danych: US Census Bureau

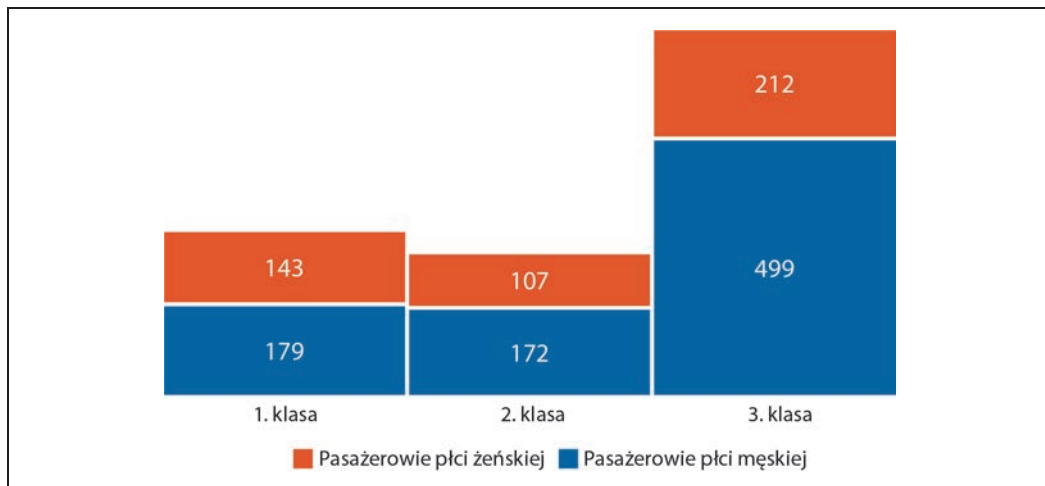
Rysunki 6.7 i 6.8 kodują jedną zmienną kategoryjną na podstawie położenia wzdłuż osi x i drugą na podstawie koloru słupka. W obu przypadkach kodowanie na podstawie położenia jest łatwe do odczytania, natomiast kodowanie na podstawie koloru słupka wymaga większego wysiłku umysłowego, ponieważ musimy mentalnie dopasować kolory słupków do kolorów w legendzie. Możemy uniknąć tego dodatkowego wysiłku umysłowego, przedstawiając cztery oddzielne wykresy słupkowe zamiast jednego zgrupowanego wykresu słupkowego (rysunek 6.9). Wybór jednej z tych opcji jest ostatecznie kwestią gustu. Prawdopodobnie wybrałbym rysunek 6.9, ponieważ eliminuje on konieczność stosowania różnych kolorów słupków.



Rysunek 6.9. Mediana dochodu gospodarstw domowych w USA w 2016 r. w stosunku do grup wiekowych i ras. Zamiast wyświetlać te dane jako zgrupowany wykres słupkowy, jak na rysunkach 6.7 i 6.8, teraz pokazujemy je jako cztery oddzielne zwykłe wykresy słupkowe. To rozwiązanie ma tę zaletę, że nie musimy kodować żadnej ze zmiennych kategorialnych za pomocą koloru paska. Źródło danych: US Census Bureau

Zamiast rysować grupy słupków obok siebie, czasami lepiej ułożyć je jeden na drugim. Układanie w stos jest przydatne, gdy suma wartości reprezentowanych przez poszczególne słupki w stosie jest sama w sobie znaczącą wartością. Tak więc, chociaż nie miałyby sensu układanie w stos wartości mediany dochodu na rysunku 6.7 (suma dwóch wartości mediany dochodu nie jest wartością znaczącą), sensowne może być układanie w stos weekendowego przychodu brutto na rysunku 6.1 (suma wartości weekendowego przychodu brutto dwóch filmów jest całkowitą wartością przychodu brutto dla dwóch filmów łącznie). Układanie w stosy jest również właściwe, gdy poszczególne słupki reprezentują liczby. Przykładowo w zbiorze danych ludzi możemy albo liczyć mężczyzn i kobiety oddzielnie, albo policzyć ich razem. Jeśli ustawimy słupek reprezentujący liczbę kobiet na słupku reprezentującym liczbę mężczyzn, to łączna wysokość słupków reprezentuje całkowitą liczbę osób bez względu na płeć.

Zademonstruję tę zasadę, korzystając ze zbioru danych na temat pasażerów transatlantyka *Titanic*, który zatonął 15 kwietnia 1912 roku. Na pokładzie było około 1300 pasażerów, nie licząc załogi. Pasażerowie podróżowali w jednej z trzech klas (pierwsza, druga lub trzecia), a na statku było prawie dwa razy więcej pasażerów płci męskiej niż kobiet. Aby zobrazować podział pasażerów według klas i płci, możemy narysować osobne słupki dla każdej klasy i płci oraz ustawiać słupki reprezentujące kobiety na słupkach reprezentujących mężczyzn, oddzielnie dla każdej klasy (rysunek 6.10). Połączone słupki reprezentują łączną liczbę pasażerów w każdej klasie.



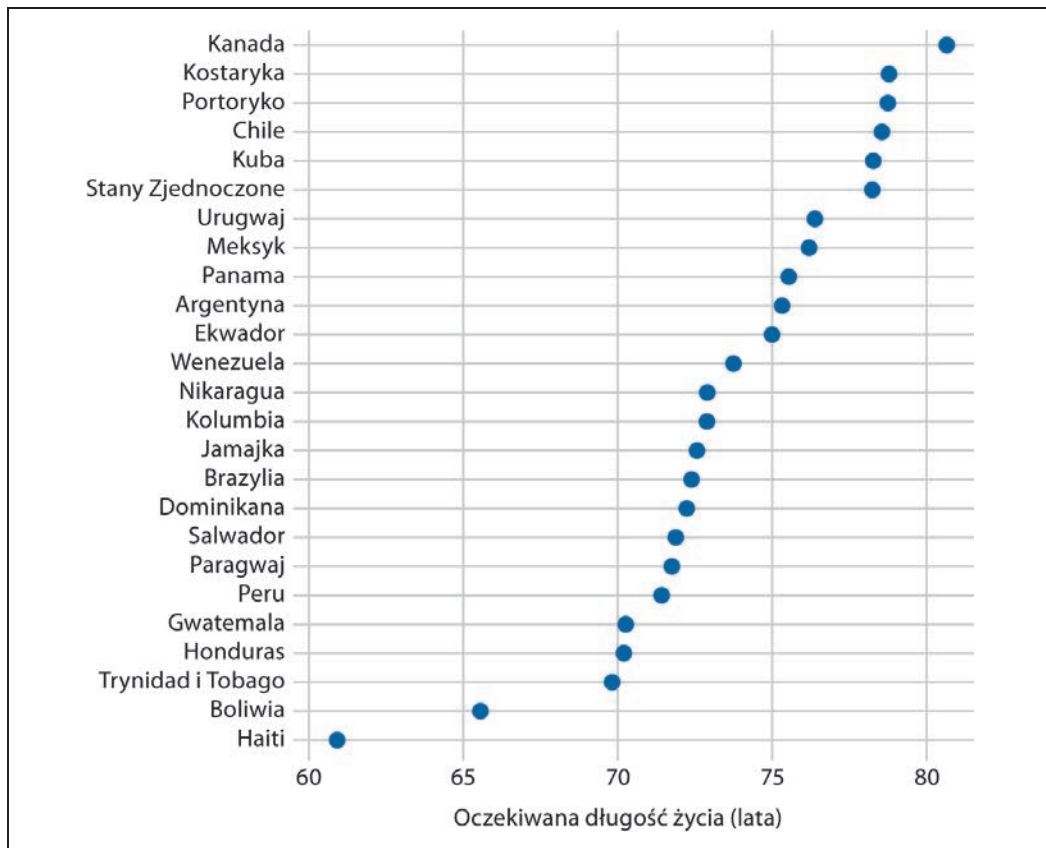
Rysunek 6.10. Liczba pasażerów płci żeńskiej i męskiej na Titanicu podróżujących w 1., 2. i 3. klasie.
 Źródło danych: Encyklopedia Titanica

Rysunek 6.10 różni się od poprzednio przedstawionych wykresów słupkowych, ponieważ nie ma wyraźnej osi y . Zamiast tego pokazałem rzeczywiste wartości liczbowe, które reprezentuje każdy pasek. Gdy wykres ma pokazywać tylko niewielką liczbę różnych wartości, sensowne jest dodanie rzeczywistych liczb do wykresu. To znacznie zwiększa ilość informacji przenoszonych przez wykres bez wprowadzania dodatkowego wizualnego szumu i eliminuje potrzebę użycia wyraźnej osi y .

Wykresy kropkowe i mapy cieplne

Słupki nie są jedyną metodą wizualizacji wielkości. Jednym z ważnych ograniczeń słupków jest to, że muszą one zaczynać się od zera, tak że długość słupka jest proporcjonalna do wyświetlanej wielkości. W niektórych zbiorach danych może to być niepraktyczne lub nawet przesłaniać kluczowe cechy. W tym przypadku możemy wskazać wielkości, umieszczając kropki w odpowiednich miejscach wzdłuż osi x lub y .

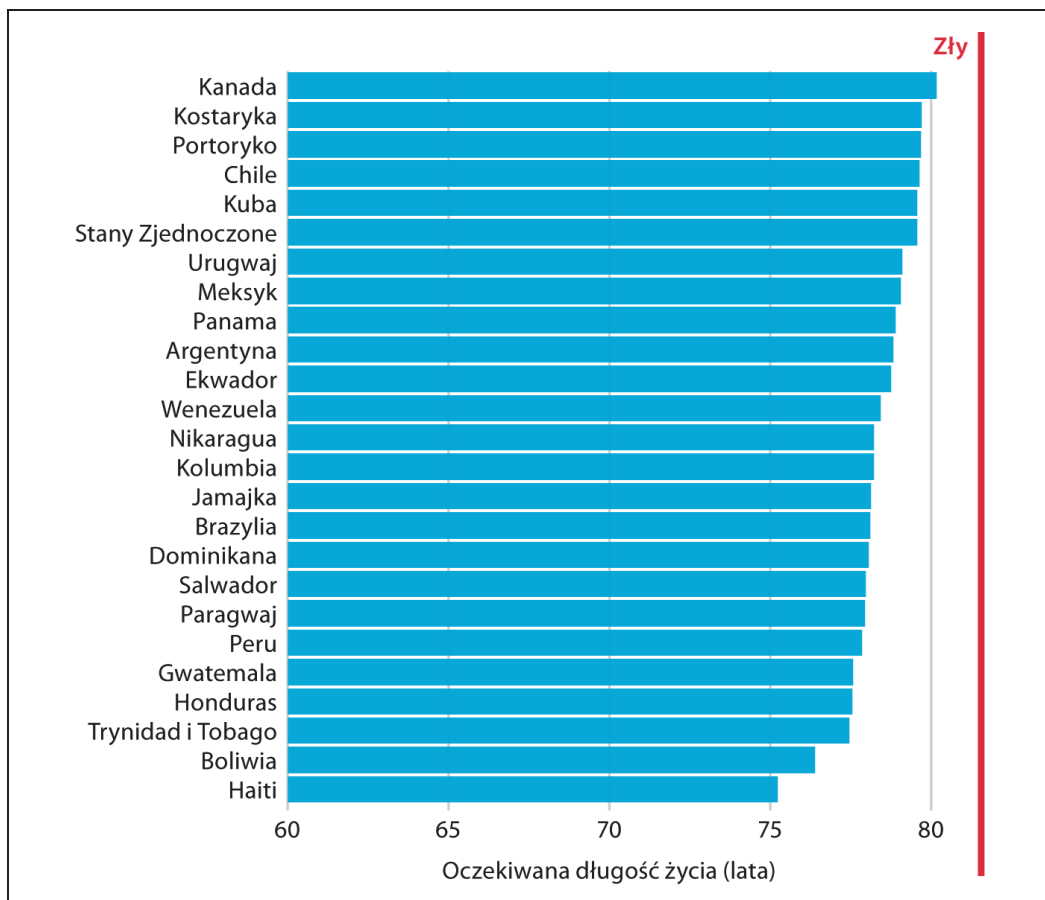
Rysunek 6.11 przedstawia to podejście wizualizacyjne dla zbioru danych dotyczących oczekiwanej długości życia w 25 krajach Ameryki Północnej i Południowej. Obywatele tych krajów mają oczekiwaną długość życia między 60 a 81 lat, a każda indywidualna wartość oczekiwanej długości życia jest oznaczona niebieską kropką w odpowiedniej lokalizacji wzdłuż osi x . Ograniczając zakres osi do przedziału od 60 do 81 lat, wykres podkreśla kluczowe cechy tego zbioru danych: Kanada ma najwyższą średnią długość życia spośród wszystkich wymienionych krajów, a Boliwia i Haiti mają znacznie niższą średnią długość życia niż wszystkie inne kraje. Gdybyśmy użyli słupków zamiast kropek (rysunek 6.12), uzyskalibyśmy znacznie mniej przekonujący wykres. Ponieważ na tym rysunku słupki są bardzo długie, przy niewielkich różnicach w ich długości wzrok przyciąga raczej środek słupków niż ich punkty końcowe, a wykres nie przekazuje swojego przesłania.



Rysunek 6.11. Oczekiwane długości życia w krajach Ameryki Północnej i Południowej na rok 2007.
Źródło danych: Gapminder

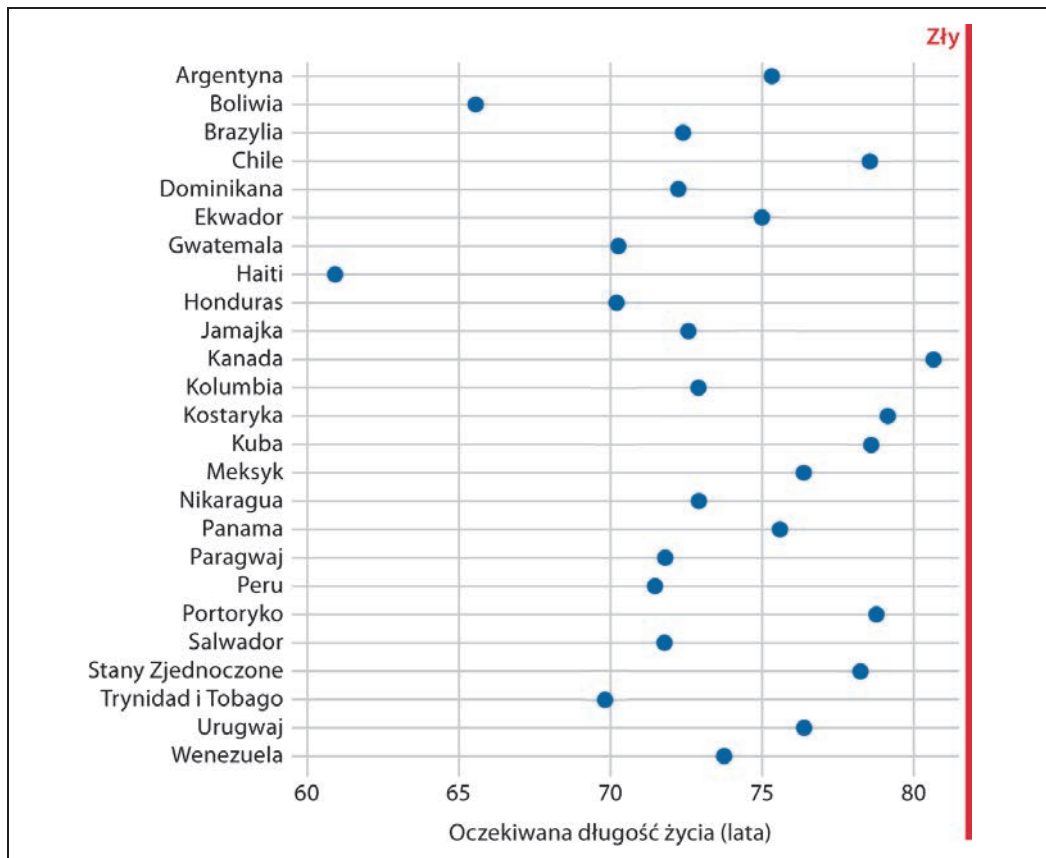
Niezależnie od tego, czy używamy słupków, czy kropek, musimy jednak zwracać uwagę na kolejność wartości danych. Na rysunkach 6.11 i 6.12 kraje uporządkowane są w porządku malejącym według oczekiwanej długości życia. Gdybyśmy zamiast tego uporządkowali je alfabetycznie, skończyłoby się to nieuporządkowaną chmurą punktów, która jest myląca i nie przekazuje wyraźnego komunikatu (rysunek 6.13).

Wszystkie dotychczasowe przykłady przedstawiały wielkości według lokalizacji wzdłuż skali pozycji albo przy użyciu punktu końcowego słupka, albo z wykorzystaniem położenia kropki. W bardzo dużych zbiorach danych żadna z tych opcji może nie być odpowiednia, ponieważ wynikowy wykres stanie się zbyt zagęszczony. Już na rysunku 6.7 widzieliśmy, że tylko siedem grup czterech wartości danych może spowodować, że wykres będzie skomplikowany i niełatwy do odczytania. Gdybyśmy mieli 20 grup po 20 wartości danych, tego rodzaju wykres prawdopodobnie byłby dość mylący.



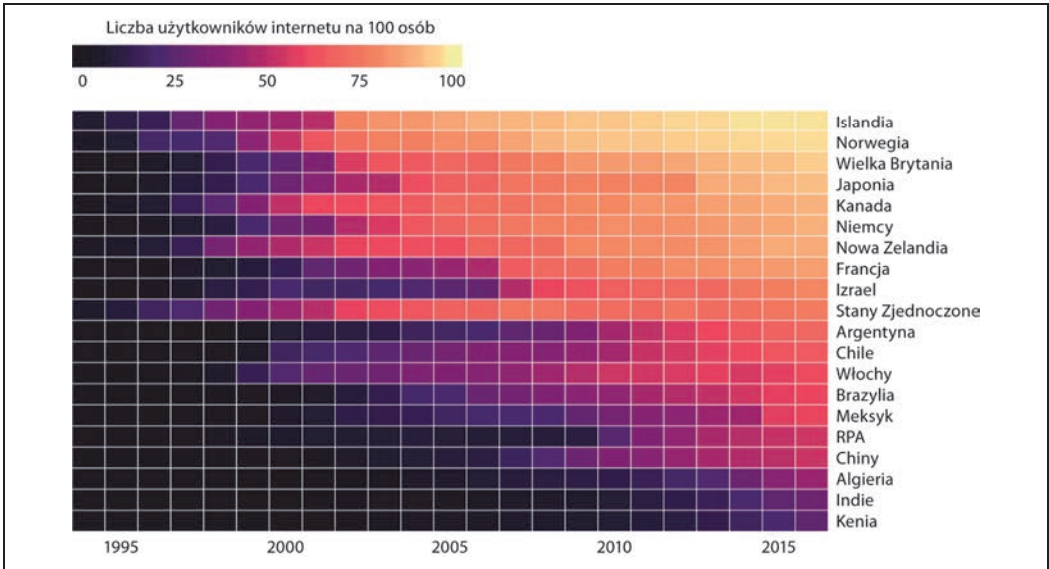
Rysunek 6.12. Oczekiwane długości życia w krajach Ameryki Północnej i Południowej na rok 2007 przedstawione w formie słupków. Taka wizualizacja jest nieodpowiednia dla tego zestawu danych. Słupki są zbyt długie i odwracają uwagę od kluczowej cechy danych — różnic w oczekiwanej długości życia w różnych krajach. Źródło danych: Gapminder

Zamiast odwzorowywać wartości danych na położenia za pomocą słupków lub kropek, możemy odwzorować wartości danych na kolory. Taki wykres nazywany jest **mapą cieplną** (ang. *heatmap*). Rysunek 6.14 wykorzystuje to podejście, aby pokazać odsetek użytkowników internetu w 20 krajach na przestrzeni 23 lat, od 1994 do 2016 roku. Choć ta wizualizacja utrudnia określenie dokładnych wartości prezentowanych danych (np. jaki jest dokładny odsetek internautów w Stanach Zjednoczonych w 2015 r.), doskonale sprawdza się w podkreślaniu bardziej ogólnych trendów. Widzimy, w których krajach korzystanie z internetu rozpoczęło się wcześniej, a w których nie, oraz które kraje mają dużą dostępność internetu w ostatnim roku ujętym w zbiorze danych (2016).

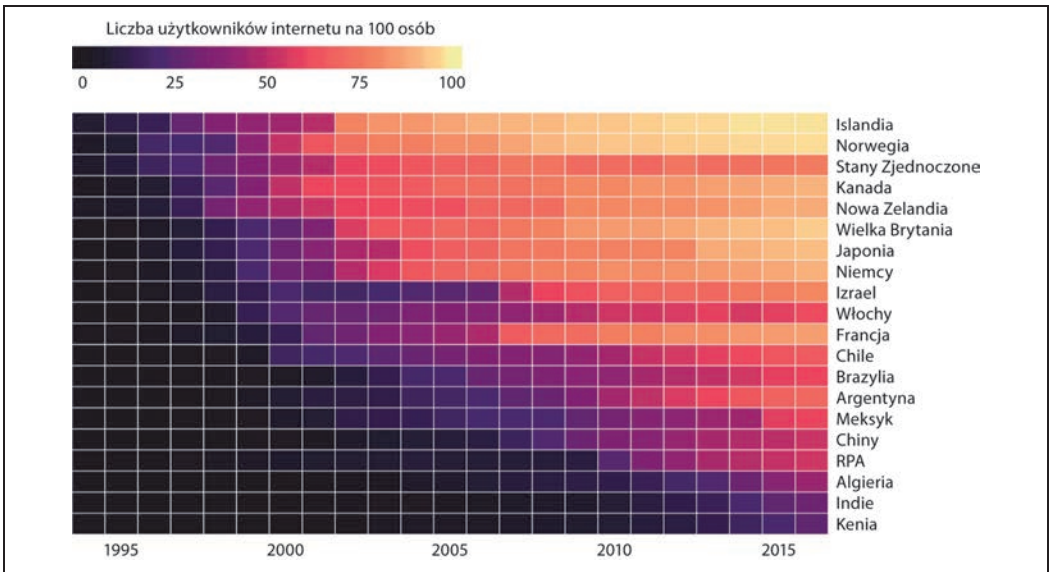


Rysunek 6.13. Oczekiwane długości życia w krajach Ameryki Północnej i Południowej na rok 2007. Tutaj kraje są uporządkowane alfabetycznie, co powoduje, że kropki tworzą nieuporządkowaną chmurę punktów. Utrudnia to odczytanie wykresu, w związku z czym zasługuje na określenie go jako „zły”. Źródło danych: Gapminder

Podobnie jak w przypadku wszystkich innych metod wizualizacji omówionych w tym rozdziale, przy tworzeniu map cieplnych należy zwrócić uwagę na kolejność wartości kategoryzowanych danych. Na rysunku 6.14 kraje są uporządkowane według odsetka internautów w roku 2016. Takie uporządkowanie plasuje Wielką Brytanię, Japonię, Kanadę i Niemcy przed Stanami Zjednoczonymi, ponieważ wszystkie te kraje miały większą dostępność do internetu w 2016 roku niż Stany Zjednoczone, mimo że w Stanach Zjednoczonych korzystanie z internetu było wcześniej bardzo intensywne. Opcjonalnie moglibyśmy uporządkować kraje według tego, jak wcześnie zaczęły odnotowywać znaczące wykorzystanie internetu. Na rysunku 6.15 kraje są uporządkowane według roku, w którym wykorzystanie internetu po raz pierwszy przekroczyło 20%. Na tym wykresie Stany Zjednoczone plasują się na trzeciej pozycji od góry i wyróżniają relatywnie niskim wykorzystaniem internetu w 2016 roku w porównaniu z tym, jak wcześnie zaczęto z niego korzystać. Podobny wzorec można zaobserwować we Włoszech. Z kolei w Izraelu i Francji korzystanie z internetu rozpoczęło się stosunkowo późno, ale szybko zyskało na popularności.



Rysunek 6.14. Rozpowszechnienie dostępu do internetu w czasie, dla wybranych krajów. Kolor reprezentuje odsetek internautów w danym kraju i roku. Kraje zostały uporządkowane pod względem odsetka internautów w 2016 roku. Źródło danych: Bank Światowy



Rysunek 6.15. Rozpowszechnienie dostępu do internetu w czasie, dla wybranych krajów. Kraje zostały uporządkowane według roku, w którym wykorzystanie internetu po raz pierwszy przekroczyło 20%. Źródło danych: Bank Światowy

Rysunki zarówno 6.14, jak i 6.15 są poprawnymi reprezentacjami danych. Wybór jednego z nich zależy od historii, którą chcemy przekazać. Jeśli nasza historia dotyczy korzystania z internetu w 2016 roku, rysunek 6.14 będzie prawdopodobnie lepszym wyborem. Jeśli jednak nasza historia dotyczy tego, w jakim stopniu wczesne lub późniejsze rozpowszechnienie internetu wiąże się z jego obecnym wykorzystaniem, wówczas preferowany jest rysunek 6.15.

PROGRAM PARTNERSKI

— GRUPY HELION —

1. ZAREJESTRUJ SIĘ
2. PREZENTUJ KSIĄŻKI
3. ZBIERAJ PROWIZJĘ

Zmień swoją stronę WWW w działający bankomat!

Dowiedz się więcej i dołącz już dzisiaj!

<http://program-partnerski.helion.pl>

GRUPA
Helion 

Jak wygląda wykres, który decyduje o sukcesie prezentacji?

Aby skutecznie przekazać wiedzę płynącą z zebranych danych, szczególnie w przypadku nauk przyrodniczych, ekonomicznych i społecznych, warto sięgnąć po narzędzia do wizualizacji. To proste z pozoru zadanie wymaga często sięgnięcia po coraz większe i coraz bardziej złożone zbiory danych, a dostępne narzędzia do wizualizacji zapewniają oszałamiający wybór rozwiązań i opcji, w których łatwo się pogubić. Co więcej, podczas przygotowywania prezentacji należy wziąć pod uwagę szczególne właściwości ludzkiego umysłu w zakresie postrzegania i przyswajania informacji. Efektywna wizualizacja danych jest bardzo istotną sprawą: nierzadko od jakości przekazywanych w ten sposób informacji zależy trafność podejmowanych decyzji.

Ten poradnik pomoże Ci uniknąć często spotykanych problemów z wizualizacją danych. Zawiera wskazówki, dzięki którym szybko zaczniesz tworzyć świetnie wyglądające i bogate w treść wykresy. Nauczysz się bezbłędnego dobierania najlepszego w danej sytuacji sposobu wizualizacji, poznasz reguły stosowania kolorów, wielkości i rodzaju fontu, zachowywania odpowiednich proporcji poszczególnych części wykresu. Przekonasz się, że dobrymi decyzjami co do jego estetyki możesz zapewnić mu przejrzystość i elegancję. Znajdziesz tu również obszerny katalog wizualizacji, co ułatwi zapoznanie się z powszechnie używanymi metodami wizualizowania danych. Ważną częścią książki jest zestaw przykładów dobrze i źle przygotowanych wykresów z wyczerpującymi wyjaśnieniami.

W książce między innymi:

- kolor jako cenne narzędzie wyróżniania danych
- dostarczanie kluczowych informacji na wiele sposobów
- dobór technik wizualizacji do różnych rodzajów danych
- reguły pomocne w projektowaniu estetycznych wykresów
- technika stosowania wykresów w większym dokumencie

Dr Claus O. Wilke jest profesorem biologii integracyjnej na Uniwersytecie Teksasńskim w Austin. Jest autorem lub współautorem ponad 170 publikacji naukowych obejmujących zagadnienia z zakresu biologii obliczeniowej, modelowania matematycznego, bioinformatyki, biologii ewolucyjnej, biochemii białek, wirusologii i statystyki. Napisał również kilka popularnych pakietów języka R wykorzystywanych do wizualizacji danych, takich jak cowplot i ggirdges. Jest współautorem pakietu ggplot2.

Helion
helion.pl
HELION SA
ul. Kosciuszki 1c
44-100 Gliwice
tel.: 32 230 98 63
helion@helion.pl

Sprawdź nasze szkolenia!
SZKOLENIA
AKADEMIA IT & BUSINESS
HELIONSZKOLENIA.PL

KOD KORZYŚCI
Sięgnij po więcej ▶
ISBN 978-83-283-6126-3
9 788328 361263
Cena: 67,00 zł