Techniki Wizualizacji Danych
Raport z pracy domowej nr 3

Jakub Piwko

19 listopada 2021

Wprowadzenie

Niniejszy raport został opracowany w ramach przedmiotu Techniki Wizualizacji Danych na kierunku Inżynieria i Analiza Danych w semestrze zimowym 2021-2022. Jego celem jest zaprezentowanie wyników eksperymentu, badającego na grupie osób występowanie pozytywnego i negatywnego wpływu niektórych praktyk w wizualizacji danych na odczyt wykresów.

Eksperyment

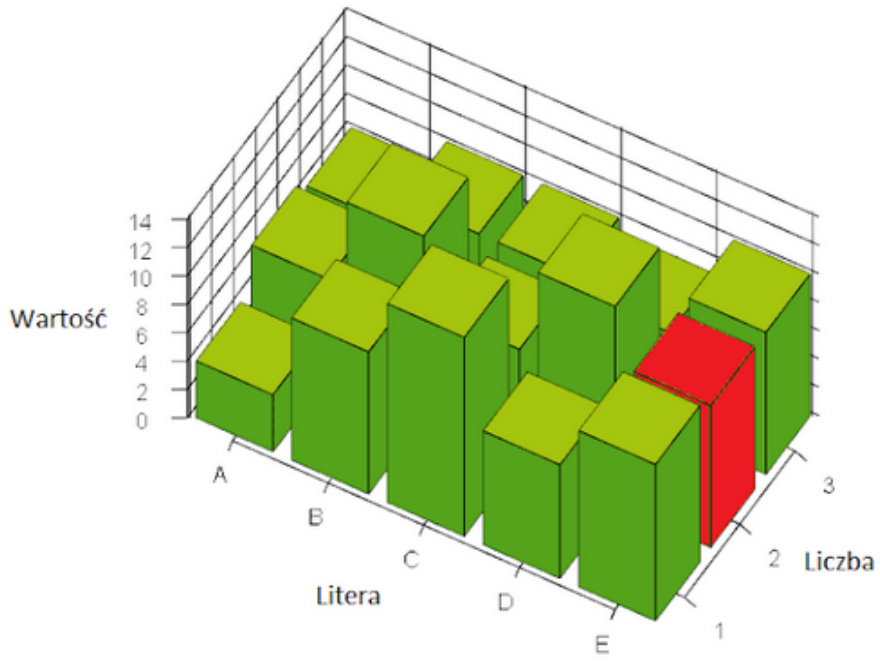
Opis

Celem mojego eksperymentu było sprawdzenie, czy problematyczne wykresy naprawdę wpływają na odczytywanie danych wśród społeczeństwa. Sporządziłem 5 różnych wykresów na bazie nieprawdziwych danych. Zależało mi na różnorodności rodzajów, ale też na tym, aby były to typy wizualizacji, które spotykamy na co dzień w mediach. Do każdego z nich sporządziłem jedno pytanie, które wymagało poprawnego odczytania danych, a czasem też drobnej operacji na odczytanych wartościach. Dwa spośród wykresów są próbą kontrolną. Zostały one sporządzone z zachowaniem najważniejszych zasad prezentacji danych, a odpowiedzi na pytania z nimi związane mają potwierdzić wiarygodność wyników. Pozostałe wizualizacje są specjalnie skonstruowane tak, aby uwypuklać ich wady.

Ankieta

Ankieta składała się z 6 różnych pytań. Pytania od 1 do 5 miały za zadanie sprawdzić trudność odczytywania danych z rożnego rodzaju wykresów. W pytaniu 6. ankietowani mieli wskazać, które wykresy ich zdaniem były najłatwiejsze i najtrudniejsze do odczytania. Ankieta w postaci formularza Google została rozesłana drogą elektroniczną. Ankietowani mieli dowolną ilość czasu na odpowiedź. Udało mi się uzyskać odpowiedzi od 25 osób. Poniżej znajdują się pytania, wykresy i odpowiedzi, z którymi musieli mierzyć się badani, wraz ze wskazaniem poprawnej odpowiedzi i problemami mogącymi utrudnić odczyt.

1. Jaka wartość odpowiada kolumnie E2 oznaczonej czerwonym kolorem?



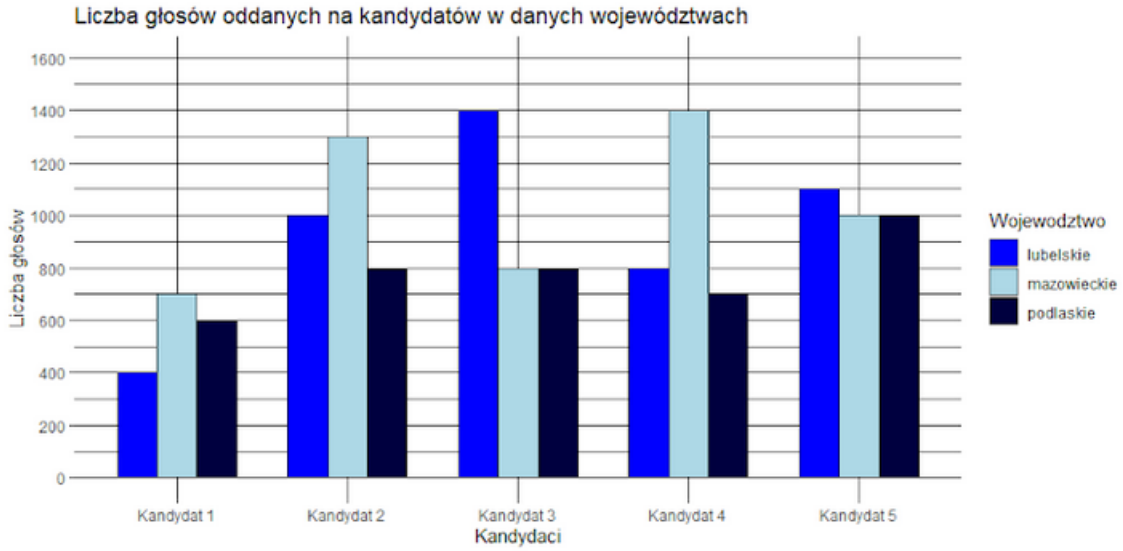
- A. 8
- B. 8.5
- C. 9
- D. 10
- E. 11

Rysunek 1: Pytanie 1

Prawidłowa odpowiedź: D

Problemy z wykresem: Podstawową wadą wizualizacji jest to, że wykorzystuje kolumnowy wykres 3D. Podanie poprawnej wartości przy tak niekorzystnym kącie i krzykliwych barwach jest kwestią szacowania, a nie faktycznego odczytania wysokości słupka z osi. Dodatkowo wykresowi brakuje sensownych podpisów. Dane nic nie mówią odbiorcy, przez co nie wie, z jakim problemem ma do czynienia.

2. Ile głosów oddano na kandydata nr 5 w województwie mazowieckim?



- A. 1200
- B. 1100
- C. 1000
- D. 900
- E. 800

Rysunek 2: Pytanie 2

Prawidłowa odpowiedź: C

Problemy z wykresem: Wykres i pytanie są analogiczne do tych z pytania 1, jednak są pozbawione wcześniejszych problemów. Wizualizacja została poprawiona w 2D, aby wartości były łatwiejsze do odczytania. Same dane zostały nazwane, przez co wiadomo na jaki temat jest wizualizacja. Aby ukryć odzwierciedlenie wykresu pierwszego, dodatkowo wartości na osi y zostały przeskalowane razy 100.

3. Uszereguj gatunki według ich procentowego udziału we wszystkich zwierzętach rosnąco w oparciu o przedstawiony wykres kołowy: *

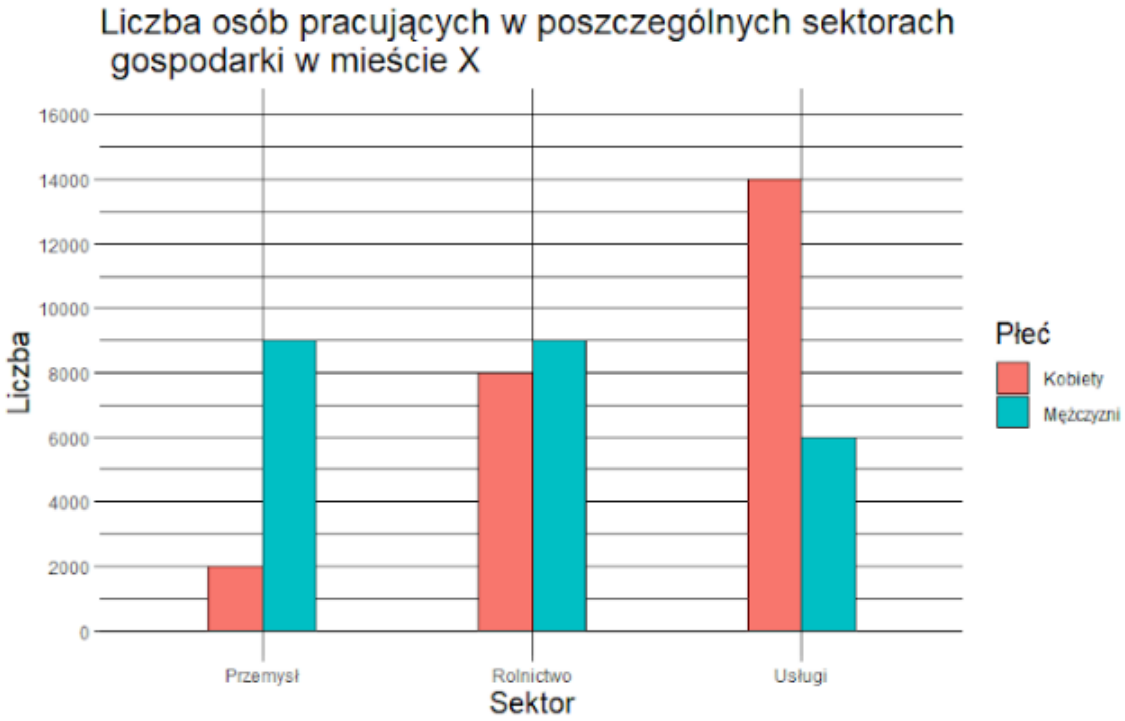
- A. Papuga, Chomik, Rybki, Świnka, Pies, Kot
- B. Rybki, Papuga, Chomik, Świnka, Kot, Pies
- C. Papuga, Rybki, Chomik, Świnka, Kot, Pies
- D. Papuga, Rybki, Chomik, Świnka, Pies, Kot
- E. Rybki, Papuga, Chomik, Świnka, Pies, Kot

Rysunek 3: Pytanie 3

Poprawna odpowiedź: C

Problemy z wykresem: Ta wizualizacja danych za pomocą wykresu kołowego jest szczególnie niekorzystna. Wartości są bardzo zbliżone, przez co różnice między kątami wycinków są trudne do zauważenia. Brak tu uporządkowania. Można powiedzieć, że dane są porozrzucane po wykresie. Dodatkowym utrudnieniem jest dobór kolorów i brak konturów. Najbardziej to widać, między wartościami kota i rybek, gdzie niebieski i turkusowy zlewają się w jeden wycinek.

4. W którym sektorze gospodarki pracuje najwięcej osób? *

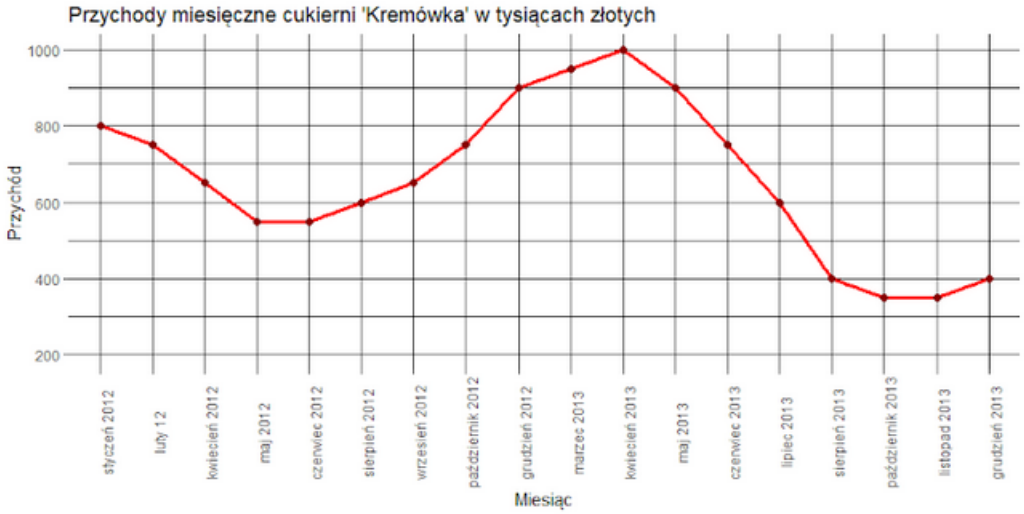


- ◯ W każdym sektorze pracuje taka sama liczba osób
- ◯ W przemyśle
- ◯ W rolnictwie
- ◯ W usługach
- ◯ W przemyśle i rolnictwie

Rysunek 4: Pytanie 4

Prawidłowa odpowiedź: D
Problemy z wykresem: Wykres jest próbą kontrolną. Jego odczytanie nie powinno stanowić problemu.

5. Jak poprawnie opisać zależność między przychodami miesięcznymi w sierpniu 2012, a przychodami 9 miesięcy później? *



- A. Przychody miesięczne zwiększyły się o ok. 50%
- B. Przychodzy miesięczne zwiększyły się o ok. 100%
- C. Przychody miesięczne zmniejszyły się o ok. 50%
- D. Przychody miesięczne zmniejszyły się o ok. 33%
- E. Przychody miesięczne się nie zmieniły

Rysunek 5: Pytanie 5

Prawidłowa odpowiedź: A

Problemy z wykresem: W tej wizualizacji problemem są wartości na osiach. Oś **X** reprezentująca czas jest wybrakowana, gdyż brakuje na niej niektórych miesięcy. Nie jest to dobrze odzwierciedlone na skali, nie ma też żadnego ostrzeżenia o przeskokach. Skala **Y** zaczyna się od wartości 200, zamiast od 0. Wszystko to powoduje, że odczytanie potrzebnych wartości jest podchwytliwe i wymaga czujności.

6. Wykres z którego pytania twoim zdaniem był najłatwiejszy, a z którego najtrudniejszy do odczytania? *

	Z pytania 1	Z pytania 2	Z pytania 3	Z pytania 4	Z pytania 5
Najłatwiejszy	☐	☐	☐	☐	☐
Najtrudniejszy	☐	☐	☐	☐	☐

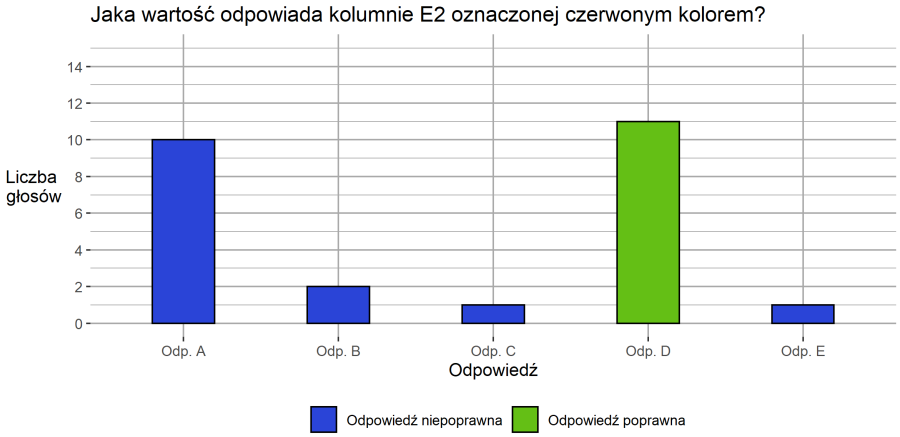
Rysunek 6: Pytanie 6

Wyniki

Prezentacja i analiza wyników

Jaka wartość odpowiada kolumnie E2 oznaczonej kolorem czerwonym?

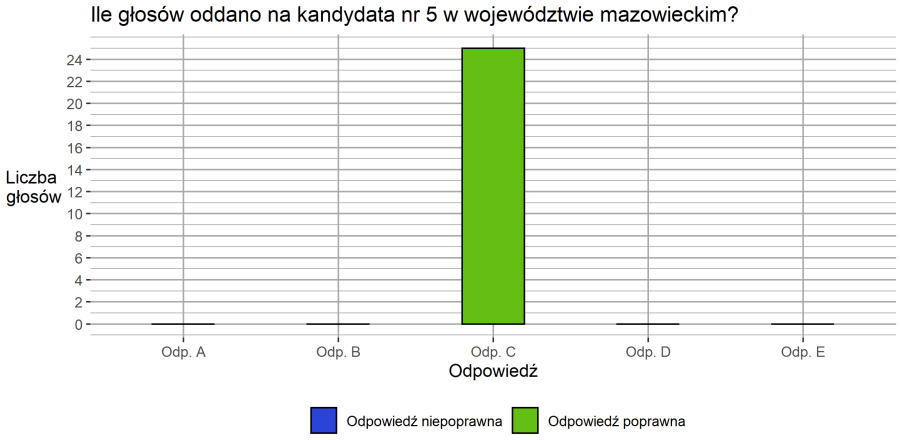
Odpowiedź	A. 8	B. 8,5	C. 9	D.10	E.11
Liczba głosów	10	2	1	11	1
Procent	40%	8%	4%	44%	4%



Poprawną odpowiedź wybrało zaledwie 44% ankietowanych. Zaskakująco dużo, bo aż 10, osób wskazało odpowiedź A. Iluzja wykresu 3D okazała się bardzo myląca, gdyż słupek czerwony ma dokładnie taką samą wartość jak sąsiedni oznaczony E3, ale kolor i perspektywa sugerowały ankietowanym, że słupek E2 jest mniejszy. Mimo, że wśród odpowiedzi były dwie pośrednie wartości między 8 a 10, które wybrały w sumie 3 osoby, to wybór częściej padał na najmniejszą, najbardziej odbiegającą od poprawnej.

2. Ile głosów oddano na kandydata nr 5 w województwie mazowieckim?

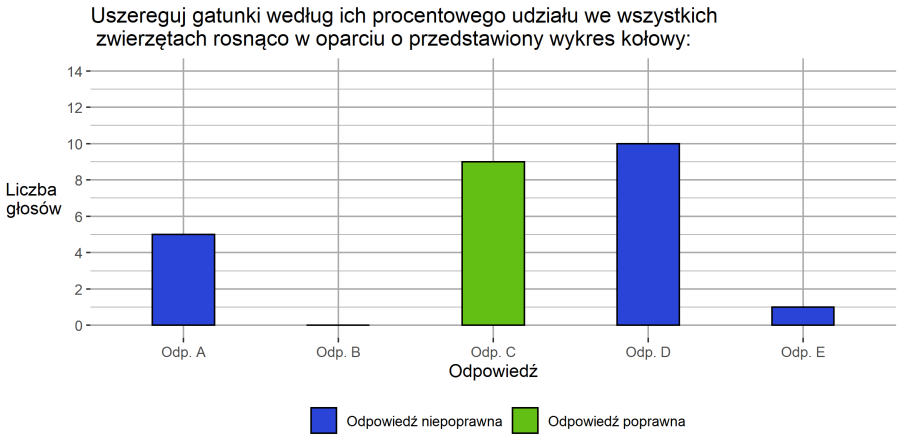
Odpowiedź	A. 1200	B. 1100	C. 1000	D.900	E. 800
Liczba głosów	0	0	25	0	0
Procent	0%	0%	100%	0%	0%



Na to pytanie wszyscy badani odpowiedzieli poprawnie. Przypomnieć warto, że wykres w tym poleceniu był odzwierciedleniem wykresu z pytania 1 w wersji 2D z nazwanymi i przeskalowanymi danymi. Okazuje się, że analogiczne wartości są odczytywane zupełnie inaczej w zależności od tego, czy użyto wykresu 2D czy 3D. Zastosowanie trójwymiarowości na wykresie zaburza poprawny odczyt danych. Perspektywa, niekorzystne kolory, przykryta przez słupki skala sprawiają, że wizualizacja jest nieczytelna, co mija się z jej celem.

3. Uszereguj gatunki według ich procentowego udziału we wszystkich zwierzętach rosnąco w oparciu o przedstawiony wykres kołowy:

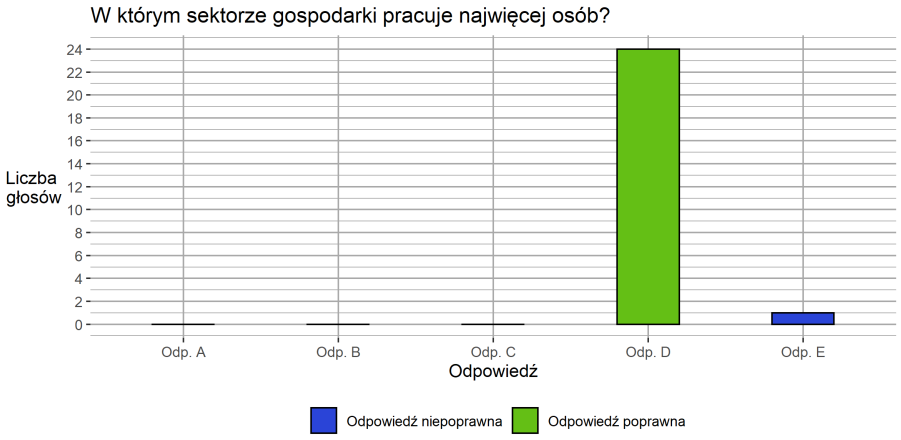
Odpowiedź	A. Papuga, Chomik, Rybki, Świnka, Pies, Kot	B. Rybki, Papuga, Chomik, Świnka, Kot, Pies	C. Papuga, Rybki, Chomik, Świnka, Kot, Pies	D. Papuga, Rybki, Chomik, Świnka, Pies, Kot	E. Rybki, Papuga, Chomik, Świnka, Pies, Kot
Liczba głosów	5	0	9	10	1
Procent	20%	0%	36%	40%	4%



Odpowiedź poprawna nie jest w tym przypadku najczęściej wybierana. Otrzymała 9 głosów, podczas gdy sąsiednia odpowiedź otrzymała o 1 więcej. Aż 5 osób wybrało kolejność A. Można uzasadnić dlaczego akurat tak rozłożyły się wybory badanych. Odpowiedź A różni się od odpowiedzi C jedynie pod względem ustawienia rybek i chomika. Na wykresie te wartości są oddzielone dużym polem odpowiadającym kotowi, który uniemożliwia poprawną ocenę kątów tych dwóch gatunków. Z kolei poprawna odpowiedź i odpowiedź D mają zamienioną kolejność kota i psa. Podobnie, błąd oceny może wynikać z nieuporządkowania, ale przede wszystkim wynika z kolorów pól kota i rybek. Podobieństwo barw sprawia, że wartość dla kota zlewa się z rybkami i wydaje się znacznie większa od pola odpowiadającemu psu, a tak nie jest. W sumie aż 64% odpowiedzi zawierało niewłaściwe ułożenie kota i psa w szeregu. Może to wskazywać, ze ocena kolejności ostatnich dwóch gatunków stanowiła największy problem. Wyniki tego polecenia potwierdzają, że wykres kołowy nie jest sprzymierzeńcem w prezentowaniu danych, a na pewno nie wykres nieuporządkowany i o tak niekorzystnym doborze kolorów.

4. W którym sektorze gospodarki pracuje najwięcej osób?

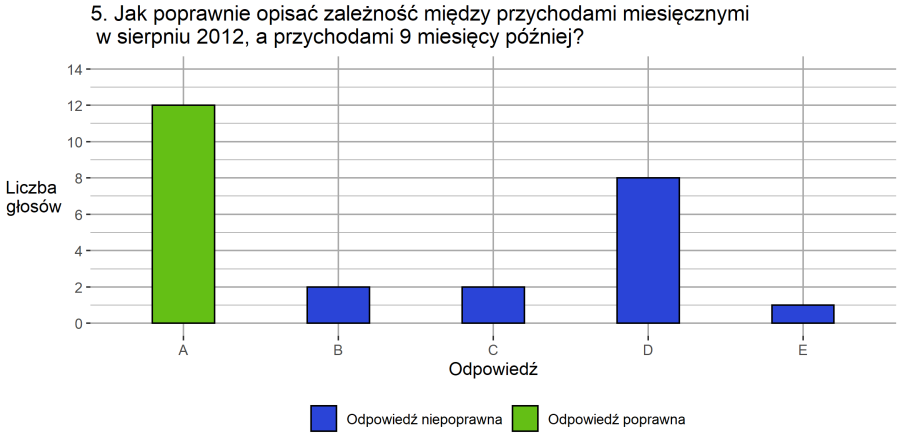
Odpowiedź	A. W każdym sektorze pracuje taka sama liczba osób	B. W przemyśle	C. W rolnictwie	D. W usługach	E. W przemyśle i rolnictwie
Liczba głosów	0	0	0	24	1
Procent	0%	0%	0%	96%	4%



Spośród 25 osób, 24 wybrały odpowiedź poprawną. Można uznać, że to bardzo dobry wynik. Pomijając jeden głos na odpowiedź E, który może wynikać na przykład z błędu rachunkowego sumowania, badani nie mieli problemu z odczytem danych. Dobór skali, koloru czy podział na płeć nie miały złego wpływu na wybór poprawnej odpowiedzi. Wizualizacja wykonana z zachowaniem dobrych praktyk okazała się łatwa do odczytania.

5. Jak poprawnie opisać zależność między przychodami miesięcznymi w sierpniu 2012, a przychodami 9 miesięcy później?

Odpowiedź	A. Przychody miesięczne zwiększyły się o ok. 50%	B. Przychody miesięczne zwiększyły się o ok. 100%	C. Przychody miesięczne zmniejszyły się o ok. 50%	D. Przychody miesięczne zmniejszyły się o ok. 33%	E. Przychody miesięczne się nie zmieniły
Liczba głosów	12	2	2	8	1
Procent	48%	8%	8%	32%	4%



Na to pytanie poprawnej odpowiedzi udzieliło 48% ankietowanych. Można było się spodziewać lepszego rezultatu, zważywszy na to, że ta wizualizacja wymagała od odbiorcy jedynie większego skupienia przy odczytywaniu danych, przez wspomniane wcześniej wybrakowane daty i rozpoczynającą się od wartości 200 oś **Y**. Patrząc na to, że aż 8 osób wskazało odpowiedź D, można powiedzieć, że bardziej problematyczna

była oś **X**. Aby odnaleźć przychód 9 miesięcy później, należało odnaleźć przychód w maju 2013 roku. Badani intuicyjnie odliczyli po prostu 9 podziałek na skali i odczytali niewłaściwą wartość. To pokazuje jak ważne są poprawne skale na wykresach. Niedbałość w zbieraniu i rozmieszczaniu danych przynosi niechciane problemy.

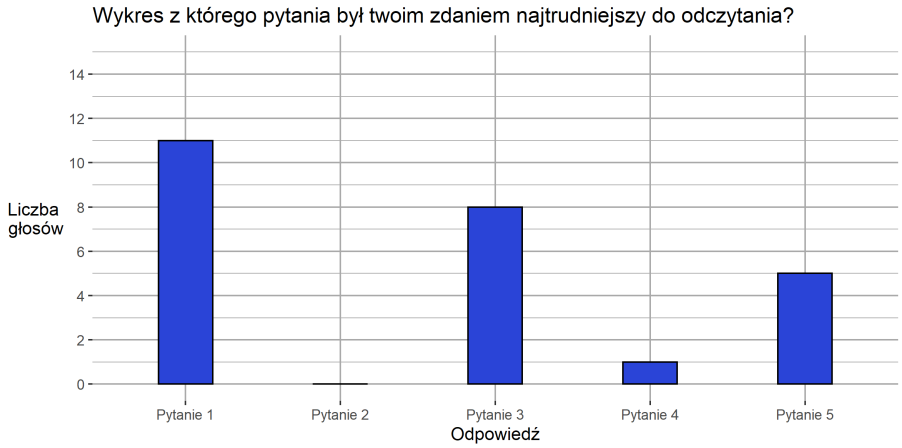
6. Wykres z którego pytania twoim zdaniem był najłatwiejszy do odczytania?

Odpowiedź	Z pytania 1	Z pytania 2	Z pytania 3	Z pytania 4	Z pytania 5
Liczba głosów	1	14	3	7	0
Procent	4%	56%	12%	28%	0%

Najwięcej osób wskazało wykres 2 jako najprostszy w odczytaniu. Dużo postawiło także na 4. W sumie na te dwie wizualizacje zagłosowało 84% ankietowanych. Wiedząc, jak dobrze w tych dwóch pytaniach poradzili sobie badani, takie wartości nie są zaskakujące, a można oczekiwać nawet wyższych. Liczba 3 głosów przy wykresie kołowym może trochę niepokoić. Ciekawe jest to, że 2 spośród osób, które udzieliły takiej odpowiedzi, odpowiedziały dobrze na pytanie 3. Być może jest to kwestia sympatii lub przyzwyczajenia do takiego typu wizualizacji, bo nie da się ukryć, kołowa reprezentacja danych jest bardzo popularna w mediach.

6. Wykres z którego pytania twoim zdaniem był najtrudniejszy do odczytania?

Odpowiedź	Z pytania 1	Z pytania 2	Z pytania 3	Z pytania 4	Z pytania 5
Liczba głosów	11	0	8	1	5
Procent	44%	0%	32%	4%	20%



Spośród 25 aż 24 wybrało wykres 1, 3 lub 5. Nie pozostawia to złudzeń, że wykresy wskazane jako najtrudniejsze do odczytania, to właśnie te, które powinny sprawić i sprawiły problemy przy udzieleniu poprawnej odpowiedzi. Duża liczba wskazań wykresu kołowego rozwiewa też wątpliwości z poprzedniego pytania. Interesujący jest fakt, że głosów na wykres 3D jest najwięcej. Odczyt wartości z tej wizualizacji był naprawdę trudny, ale nie spodziewałem się, że badani wskażą akurat ten wykres jako ten najbardziej kłopotliwy. Może to wynika z pewnego porównania z korzystniejszym wykresem 2, a może rzeczywiście trójwymiarowe słupki są największą zmorą wizualizacji danych.

Wnioski

Uważam, że przedstawiony eksperyment i jego wyniki w dużej mierze potwierdzają, że niekorzystne wizualizacje w istocie negatywnie wpływają na odczyt danych. Podsumowując, w pytaniach zawierających wykresy nakierowane na błąd ze strony badanego, średnio tylko 43% udzielonych odpowiedzi było poprawnych. Dodatkowo 96% ankietowanych wskazało jeden z tych wykresów jako najtrudniejszy do odczytania. Jest to druzgocąca statystyka, szczególnie w zestawieniu z dwoma pozostałymi wykresami, przy których poprawne odpowiedzi są udzielane przez prawie 100% pytanych, a 86% spośród nich wskazuje je jako najłatwiejsze do odczytu. Tworzenie wykresów ma na celu wizualizowanie danych w takich sposób, aby przedstawić je w sposób jak najprostszy, najwygodniejszy i najszybszy do odczytania przez ludzkie oko. Można sobie więc zadać pytanie: Jaki jest sens tworzenia wizualizacji, które nie spełniają tych założeń? Żaden. Jednak nadal w różnych mediach napotykamy takie przykłady. Mimo, że do dyspozycji jest mnóstwo przyjaznych typów wykresów i narzędzi do prezentowania różnych danych, to wizualizacje kołowe, przesadna kolorystyka, ubogie nazewnictwo, niedokładne skale, trójwymiarowe grafiki są nadużywane. Potwierdzenie działania dobrych praktyk na wykresach skłania też do wniosku, że o ile nie chcemy, aby wykres stanowił narzędzie do dezinformacji, to zawsze warto dbać o czytelność wizualizacji. Nie kosztuje to więcej czasu, nie wymaga też większego nakładu pracy, a gwarantuje, że język wykresów będzie zrozumiały dla każdego.