

Problemy z Wizualizacją Danych na Wykresach Kołowych

Katarzyna Grzesiuk

1 Wprowadzenie

W poniższym raporcie przedstawione zostaną wyniki ankiety na temat wizualizacji danych na wykresach kołowych.

Główny cel pracy domowej

Tematem pracy domowej jest zbadanie czy pewne ustalone praktyki dotyczące przedstawiania danych faktycznie odpowiadają odbiorcom. W tym raporcie skupimy się na wykresach kołowych i częstych błędach na nich występujących. Pokazane zostanie jak format 3D wykresu wpływa na postrzeganie znajdujących się na nim danych, w jaki sposób źle dobrana skala może oszukać ludzkie oko oraz jak typ wykresu wpływa na czytelność przedstawianych danych.

2 Przeprowadzona ankieta

Obserwacje przedstawione w tym raporcie zostały zebrane na podstawie wyników poniższej ankiety.

https://docs.google.com/forms/d/e/1FAIpQLSeC2qyoNYtKuZ-g6Vb1R1kwz_1cWcTsL0-oDEj02o7zCARQ4A/viewform?fbclid=IwAR2ksxYhhw2kEvzhNTM2aWlFybnw_n3EBFxZH68EWjBoQP3jntXSXGbcEs

Opis ankiety

Przeprowadzona ankieta składała się z pięciu pytań jednokrotnego wyboru. Pierwsze trzy dotyczyły odczytania konkretnych danych z wykresu, a ostatnie dwa pytaly o preferencje ankietowanego odnośnie typu wizualizacji konkretnych danych.

Osoby ankietowane

Ankieta została przeprowadzona na różnorodnej grupie osób. Znajdowały się w niej mężczyźni i kobiety z różnych przedziałów wiekowych. Ankiety wypełniały zarówno osoby, które w trakcie nauki lub kariery zawodowej miały styczność z analizą wykresów jak i zupełnie niezwiązane z tym zagadnieniem.

3 Przeprowadzone eksperymenty i ich wyniki

Pytanie 1

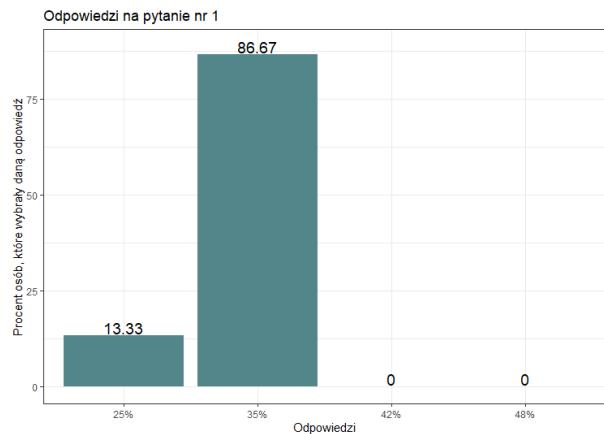
Treść

Pierwszy eksperyment polegał na sprawdzeniu jak ludzkie oko odczytuje liczby z wykresu kołowego gdy wartości na wykresie nie są podpisane. Zadanie polegało na wskazaniu wartości procentowej, która może odpowiadać czerwonemu polu. Do wyboru były cztery różne odpowiedzi: 48%, 25%, 42% and 35%. Ankietowanym został w tym celu przedstawiony poniższy wykres z wymazanymi wartościami.



Wyniki

Ankietowani decydowali między odpowiedziami 25% oraz 35%. Nikt nie zaznaczył odpowiedzi 42% oraz 48%. Wyniki pokazują, że mimo iż każdy trochę inaczej postrzega wartości na wykresie kołowym to wyniki te są zbliżone i na podstawie proporcji jesteśmy w stanie odczytać przybliżoną wartość odpowiadającą danemu wycinkowi. Poniżej przedstawiony jest wykres obrazujący rozkład procentowy odpowiedzi na pytanie pierwsze.



Główny problem

Wykres, który został wykorzystany przy tym pytaniu został wybrany celowo. Przyjrzymy się teraz oryginalnemu wykresowi. Pochodzi on ze strony:

https://www.reddit.com/r/CrappyDesign/comments/fpdv53/a_pie_chart_out_of_178/



Według powyższego wykresu odpowiedzią na nasze pytanie powinno być 48%, jednak nietrudno stwierdzić, że jest to nieintuicyjne, co pokazują nam wyniki ankietowania.

wybrała wartość 35% i nikt nie zdecydował się na 48%. Pokazuje to jak nieczytelny jest ten wykres. Skala przekłamuje wartości, które dodatkowo nie reprezentują niczego sensownego, gdyż jak nietrudno zauważyć sumują się do 178%.

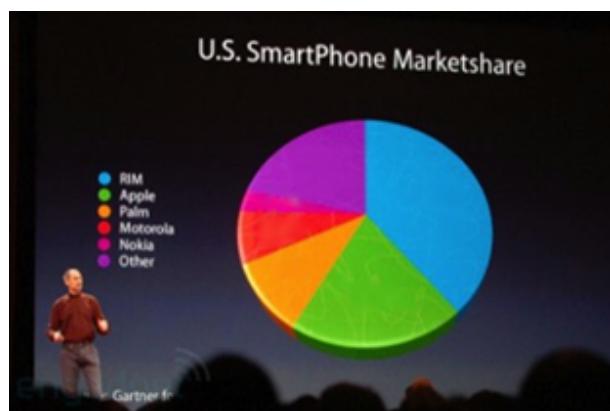
Wnioski

Na podstawie powyższego eksperymentu możemy wywnioskować, że przekłamane wartości zaburzają postrzeganie wykresów kołowych. Wartości przedstawione na orginalnym wykresie nie pokrywają się z wartościami, których byśmy się spodziewali patrząc tylko na podział wykresu na części. Dlatego gdy już decydujemy się na wykres kołowy, ważne jest to żeby zadbać o to by wartości sumowały się do 100%. Dzięki temu podpis i zazanczony wycinek koła będą współgrały i rzetельnie przedstawiały nasze dane. Możemy jednak wywnioskować, że jesteśmy w stanie odczytać przybliżone wartości z wykresu gdy są one przedstawione w poprawnej skali.

Pytanie 2

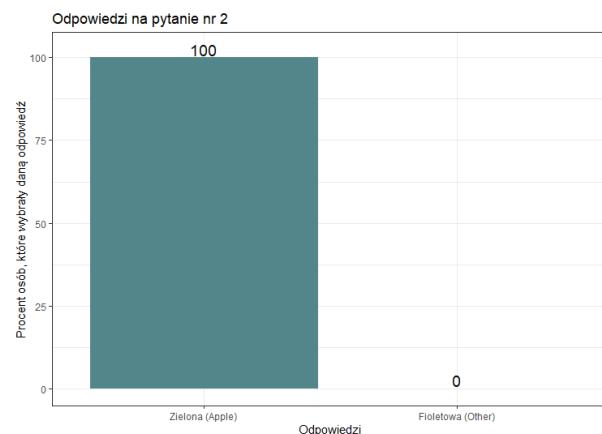
Treść

Drugi eksperyment polegał na sprawdzeniu czy forma 3D wykresu kołowego wywołuje iluzję optyczną czy rzetельnie i proporcjonalnie przedstawia odpowiednie dane. W tym celu przedstawiony został poniższy wykres, ponownie z wymazanymi wartościami. Ankietowani mieli wskazać, który wycinek wykresu jest większy: zielony ("Apple") czy fioletowy ("Other").



Wyniki

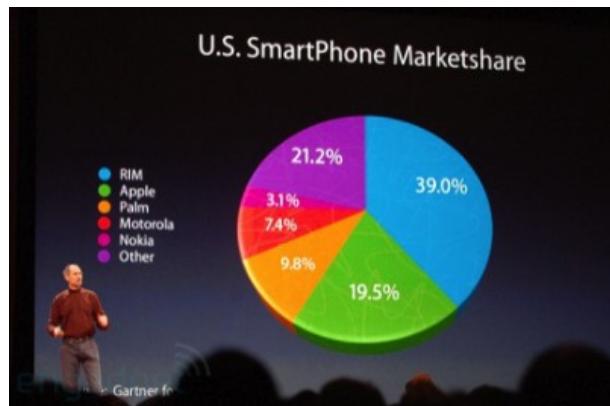
W przypadku pytania drugiego wyniki są jednoznaczne. Wszyscy ankietowani выбрали zieloną część wykresu jako większą. Jest to jasne potwierdzenie, że dla oka ludzkiego zielona wydaje się dominować nad fioletową. Wyniki ankiety zostały przedstawione na wykresie słupkowym.



Główny problem

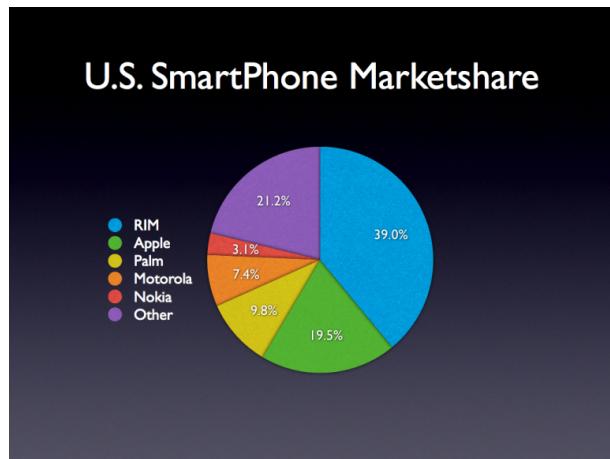
Wykres ten został wybrany aby pokazać w jaki sposób wykresy 3D mogą manipulować danymi poprzez stworzenie iluzji optycznych. Popatrzmy na oryginalny wykres. Pochodzi on ze strony internetowej:

https://2012wc.wordpress.com/2012/03/07/week-07-critical-eye-spotting-the-difference/?fbclid=IwAR0p_VadpTYkkKnyQbJi4Yne0rQ01Khepxs6Tctu1kAkXfz4fax6r5L_ArM



Jak widzimy na oryginalnym wykresie to część fioletowa reprezentuje większą wartość procentową, ale przez format 3D wykresu wydaje się ona mniejsza, co potwierdzają wyniki ankiety, w której wszyscy zaznaczyli część zieloną jako większą. Taki sposób reprezentacji danych prowadzi do manipulacji, gdyż nawet z podpisanymi wartościami na pierwszy plan wysuwa się zielony wycinek reprezentujący firmę Apple, który wydaje się większy niż jest w rzeczywistości. Warto też zobaczyć ten sam wykres w formie 2D, na którym zachowane są odpowiednie proporcje. Na poniższym wykresie wycinki koloru zielonego i fioletowego są podobnej wielkości dzięki czemu dane przedstawione są rzetельnie. Wykres pochodzi ze strony internetowej:

https://2012wc.wordpress.com/2012/03/07/week-07-critical-eye-spotting-the-difference/?fbclid=IwAR0p_VadpTYkkKnyQbJi4Yne0rQ01Khepxs6Tctu1kAkXfz4fax6r5L_ArM



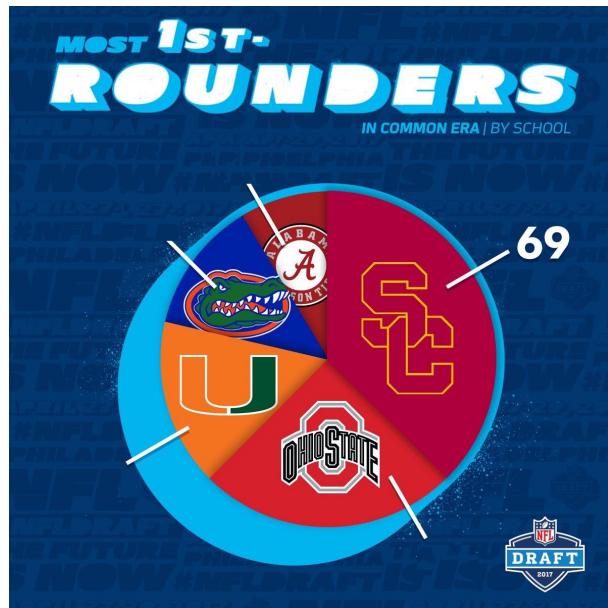
Wnioski

Wykonany eksperyment pokazuje jak wykresy 3D przekłamują dane i sprawiają, że błędnie je odczytujemy, podczas gdy wystarczy użyć wykresu 2D aby pozbyć się tego problemu. Jest to często wykorzystywane w celu manipulowania postrzegania danych przez potencjalnych odbiorców. W ten sposób możemy łatwo uwidoczyć jakąś część wykresu lub sprawić, że jakaś część będzie wdawała się mniej.

Pytanie 3

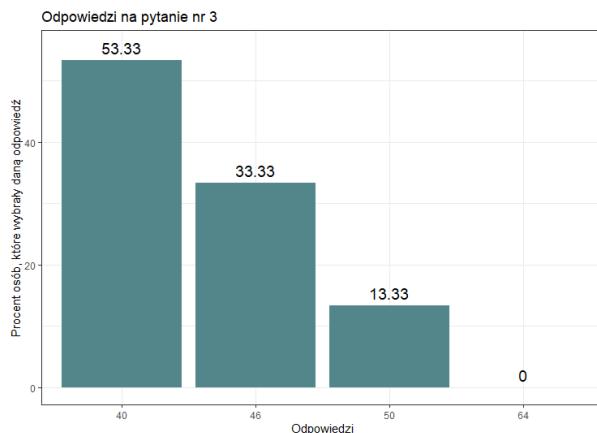
Treść

Przedstawiony został wykres kołowy obrazujący punkty zdobyte przez daną drużynę z nieintuicyjną skalą. Wymazane zostały wszystkie wartości oprócz jednej. Zadaniem ankietowanych było powiedzenie na podstawie pozostawionej wartości ile punktów zdobyła drużyna Ohio State. Do wyboru były 4 odpowiedzi: 46, 64, 50 oraz 40.



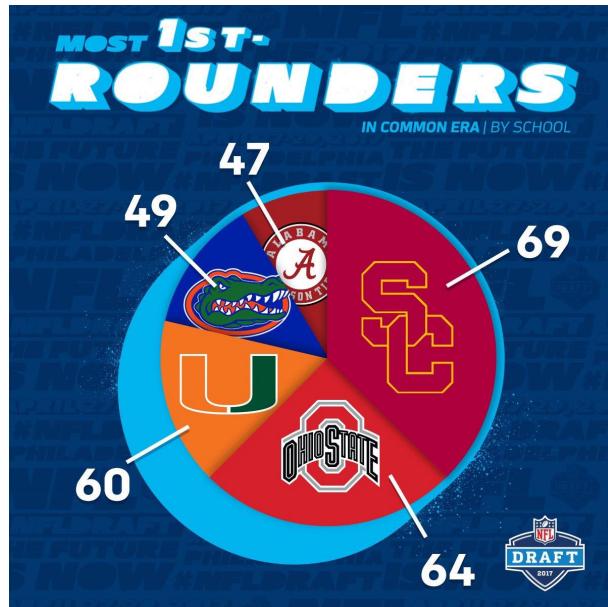
Wyniki

Większość ankietowanych wybrała odpowiedź 40, ale pojawiały się też odpowiedzi 46 oraz 50. Nikt nie zdecydował się na odpowiedź 64. Wyniki ponownie pokazują, że z wykresu kołowego nie da się jednoznacznie odczytać wartości, ale jesteśmy w stanie stwierdzić wartość przybliżoną na podstawie proporcji. Wyniki zostały przedstawione na wykresie słupkowym.



Główny problem

Już patrząc na wykres z jedną wartością możemy zauważyc, że wycinek reprezentujący 69 punktów nie jest nawet połową koła. Teoretycznie w tym wypadku nie przedstawiamy wartości procentowych, ale w takim razie wykres kołowy staje się bezużyteczny, gdyż nie wiemy na podstawie jakiego kryterium przyporządkujemy danej wartości wielkość wycinka. Nie jest to jednak jedyny problem tego wykresu. Spójrzmy na wykres oryginalny. Pochodzi on z poniżej strony internetowej:
<https://imgur.com/wPyNMK5>



Odpowiedzi ankietowanych były rozłożone pomiędzy 46, 40 a 50 punktów. Nikt nie wybrał wartości 64, co nie jest z resztą dziwne, bo już na pierwszy rzut oka widać, że wycinek Ohio State stanowi około $\frac{2}{3}$ wycinka SC. Evidentnie zatem powyższy wykres ma również poważny problem ze skalą, co może wynikać z chęci przedstawienia SC jako najlepszej drużyny. Gdy przyjrzymy się oryginalnemu wykresowi możemy zauważać, że dotyczy to również pozostałych wycinków, gdyż np część o wartości 49 jest znacznie większa od części o wartości 47.

Wnioski

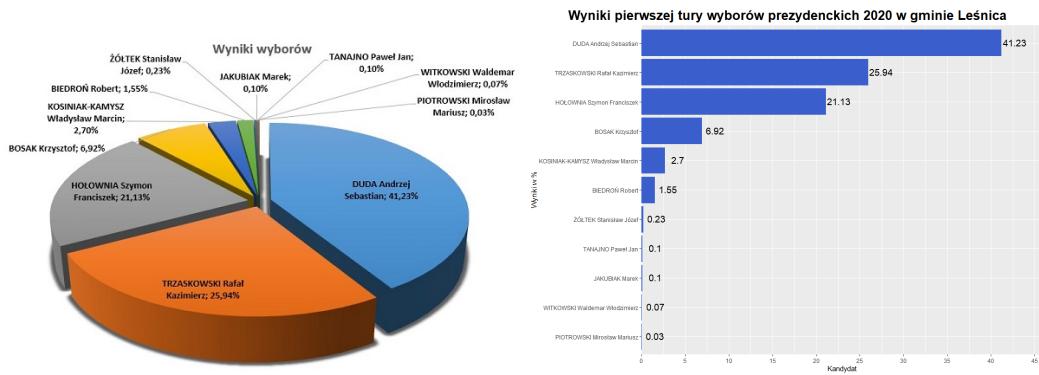
Na rozważanym wykresie nie są przedstawione żadne sensowne proporcje. Wartości nie są procentowe, przez co cała koncepcja przedstawienia danych na wykresie kołowym jest chybiona. Zdecydowanie lepiej sprawdziłby się zwykły wykres słupkowy. Ten przykład pokazuje jak ważne jest zadbanie o dobrą skalę wykresu oraz przemyślenie czy faktycznie dany sposób wizualizacji ma sens w przypadku danych, które chcemy przedstawić.

Pytanie 4

Treść

Dane odnośnie wyników pierwszej tury wyborów prezydenckich zostały przedstawione na dwa sposoby: na wykresie kołowym 3D oraz na wykresie słupkowym. Zadaniem ankietowanego było wskazanie, który wykres jest bardziej czytelny. Wykres kołowy pochodzi ze strony :

<http://lesnica.pl/7200/wyniki-wyborow-prezydenckich-2020-w-gminie-lesnica-i-tura.html?fbclid=IwAR0g1YpxtQhNMdf0G9aBT1HLUG2bbswcZIBM6jUKwicXK-faTMcahXtARX4>

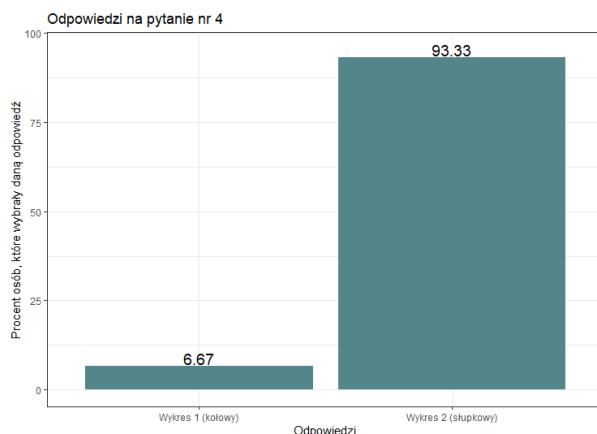


(a) Wykres kołowy

(b) Wykres słupkowy

Wyniki

Tylko nieco ponad 6% osób zdecydowało, że wykres kołowy jest czytelniejszą formą przedstawienia danych. Pozostali ankietowani wybrali wykres słupkowy.



Główny problem

Pytanie to stanowi pewne podsumowanie poprzednich eksperymentów. Sprawdzamy tutaj czy faktycznie w zestawieniu z inną reprezentacją danych wykresy kołowe są mniej czytelne. Porównywane wykresy przedstawiają te same dane. Wykres kołowy został stworzony po wynikach wyborów i przedstawiony na stronie gminy. Wykres słupkowy powstał w oparciu o te same dane przy wykorzystaniu pakietu ggplot. Porównanie to jest istotne, ponieważ nie mamy tu nie tylko wykres kołowy, który już niekoniecznie jest dobrą reprezentacją naszych danych, ale dodatkowo jest on w formie 3D przez co zniekształca otrzymane wyniki.

Wnioski

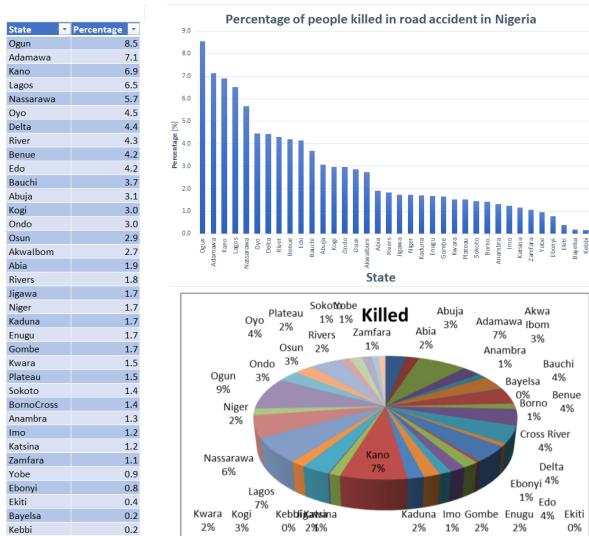
Okazuje się, że faktycznie większość ankietowanych wybrała wykres słupkowy jako bardziej czytelny, co potwierdza, że wykresy kołowe nie zawsze są czytelne. Dodatkowo po raz kolejny mamy do czynienia z formą 3D przez co na pierwszy plan wysuwa się pomarańczowy wycinek koła, co manipuluje postrzeganiem danych. W tym wypadku wykres słupkowy sprawdza się lepiej, ponieważ wyraźnie widać na nim różnice w wielkościach, nazwiska kandydatów oraz otrzymane wyniki.

Pytanie 5

Treść

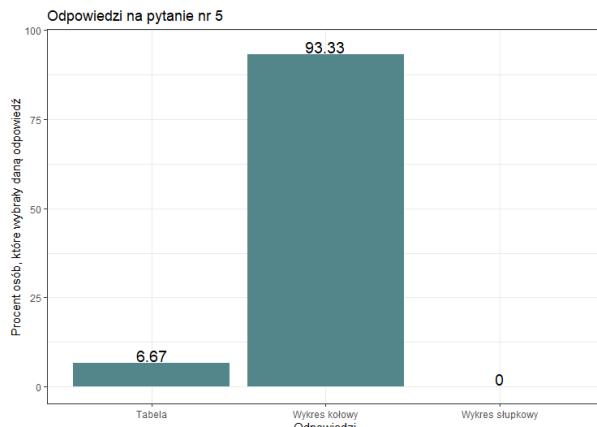
Podobnie jak w poprzednim pytaniu ankietowany ma do porównania różne sposoby wizualizacji tych samych danych. Tym razem do wyboru są trzy formy: tabela, oryginalny wykres kołowy i wykres słupkowy. Wykres kołowy pochodzi z poniższej strony internetowej:

https://www.researchgate.net/figure/Pie-Chart-showing-percentage-of-people-killed-in-road-accident-in-Nigeria_fig6_235325092



Wyniki

Jako najmniej czytelny sposób przedstawienia danych został wybrany wykres kołowy. Tylko około 6% osób zdecydowało się na tabelę, ale nikt nie wybrał wykresu słupkowego jako nieczytelnego.



Główny problem

To pytanie również podsumowuje poprzednie. Przedstawiony wykres pokazuje jaki procent śmiertelnych wypadków drogowych występował w danym regionie Nigerii. Tabela i wykres słupkowy powstały w oparciu o te same dane i reprezentują to samo. Zdecydowana większość ankietowanych wskazała wykres kołowy jako najmniej czytelny z trzech zaproponowanych wizualizacji. Faktycznie ten wykres nie przedstawia wartości dobrze. Potrzeba bardzo dużo skupienia aby porównując rozmiary danych wycinków stwierdzić np. która część wykresu reprezentuje region Plateau. Dodatkowo ponownie pojawia się wykres 3D, przez co wielkości wycinków są zaburzone. Pojawia się tutaj również dziwne wykorzystanie kolorów. Sekwencja kolorystyczna powtarza się ale zastosowany jest gradient, przez co wartości w lewym górnym roku są bardziej wyblakłe, co nie wygląda estetycznie.

Wnioski

Ponownie potwierdza się hipoteza, że wykres kołowy jest mało czytelną formą przedstawienia danych nawet jeśli chcemy przedstawić wartości procentowe. Autor chciał przedstawić na nim dane wraz z podpisami co dało bardzo dziwny rezultat, ze względu na zbyt dużą ilość obserwacji. Użyto formy 3D, wielu kolorów i gradientu podczas gdy wszystkie informacje są właściwie zawarte poza samym obszarem wykresu. Patrzenie na ten wykres przypomina oglądanie prezentacji początkującego użytkownika programu PowerPoint, który chce użyć wszystkich jego dobrodziejstw na raz. Co ciekawe wykres pochodzi z pracy naukowej. Ponownie dowodzi to, że warto przemyśleć w jakiej formie możemy przedstawić nasze dane.

4 Podsumowanie

Przeprowadzone eksperymenty pokazują w jaki sposób błędy na wykresach kołowych mogą manipulować odbiorem przedstawionych na nich danych. Rzeczą, na które warto zwrócić uwagę na takich wykresach z pewnością jest dobranie odpowiedniej skali i sprawdzenie czy przedstawione wartości procentowe sumują się do 100%. Dodatkowo używanie wykresów 3D zazwyczaj nie jest dobrym pomysłem gdy chcemy rzetelnie przedstawić proporcje między danymi, dlatego należy ich unikać. Przede wszystkim warto jednak jest się zastanowić czy na pewno wykres kołowy jest odpowiedni do przedstawienia naszych danych i czy będzie on czytelny. Na podstawie ostatnich dwóch pytań możemy wysnuć wniosek, że zazwyczaj odpowiedź na to pytanie jest negatywna, dlatego dobrą praktyką w przypadku tego typu wykresów jest unikanie ich o ile to możliwe.