Szymon Gruda 239661 239661@edu.p.lodz.pl

Zadanie 1.: Problem Set 1

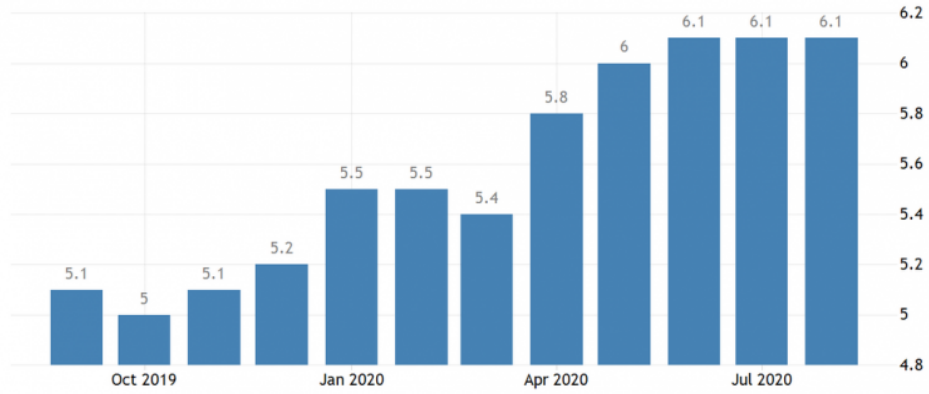
1. Wprowadzenie

Dane wraz ze wykorzystującą je statystyką pozwalają opisywać otaczający ludzi świat i informować co się w nim dzieje. Bardzo ważnym aspektem jest ich prezentacja (wizualizacja). Dużo trudniej jest ludziom zinterpretować długą tabelę z samymi liczbami, lepszą metodą jest zwizualizowanie danych, np. poprzez wykres. Niestety wizualizowanie danych jest podatne na różnego rodzaju manipulacje, które zostaną omówione poniżej.

2. Przykłady manipulacji danymi podczas ich wizualizacji

Rysunek 1. Wykres stopy bezrobocia w Polsce

Rysunek 1 przedstawia wykres stopy bezrobocia w Polsce. Wykres na pierwszy rzut oka wskazuje, że stopa bezrobocia maleje. Niestety tak nie jest, okazuje się że autor wykresu, miesiąc chronologicznie późniejszy umieścił wcześniej. Dodatkowo dziedzina danych to bardzo wąski zbiór trzech miesięcy, przez co wykres nie obrazuje sytuacji w ciągu całego jednego roku, dopuszczalne byłoby przedstawienie danych zebranych w ramach jednego kwartału, ale nie trzech różnych miesięcy, na przestrzeni dwóch lat.



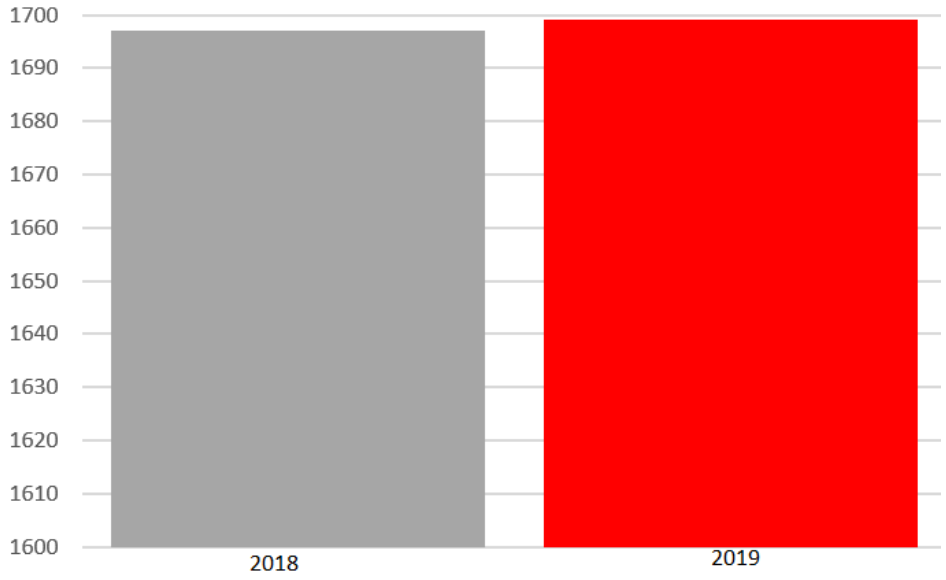
Rysunek 2. Wykres stopy bezrobocia w Polsce zaprezentowany w sposób poprawny

Rysunek 2 również przedstawia wykres stopy bezrobocie w Polsce, jest on natomiast pozbawiony wad wykresu z rysunku 1.

Rysunek 3. Wykres liczby stop cafe w Polsce

Rysunek 3 przedstawia "dynamniczny wzrost liczby stop cafe w Polsce". Infografika nie zawiera dokładnej informacji co prezentuje, ale można się domyślać że liczbę kawiarni na przestrzeni dwóch lat. Liczby zostały przedstawione jako dwa kubki kawki, jeden jest czterokrotnie mniejszy od drugiego.

Czytelnik mógłby wywnioskować na podstawie samej grafiki, że ma doczynienia z dwukrotnym wzrostem liczby, otóż nie, podane obok kubków liczby prezentują wzrost liczby kawiarni o 2.



Rysunek 4. Poprawiony wykres liczby stop cafe w Polsce

Rysunek 4 przedstwia wykres liczby kawiarni na przestrzeni dwóch lat, pozboawiony wad wykresu z rysunku 3.

3. Wnioski

Wykonując zadanie można wywnioskować, że:

— Nieodpowiednią wizualizacją danych można przedstawić dowolną informację, wykorzystując niepasujące do niej dane.
— Bardzo istotnym czynnikiem w wizualizcji danych jest okres czasu, z którego dane będą prezentowane.
— Szczególną uwagę należy zwracać na informacje uzyskiwane z infografik, ponieważ zbyt skupiony na ładnym przedstawianiu danych autor, może zrobić to w sposób nierzetelny.

Wprowadzenie do Data Science i metod uczenia maszynowego **2020/2021**

Prowadzący: mgr inż. Rafał Woźniak Wtorek, 13:15

Jan Karwowski 239671 239671@edu.p.lodz.pl

Zadanie 1.: Problem set 1

1. Wprowadzenie

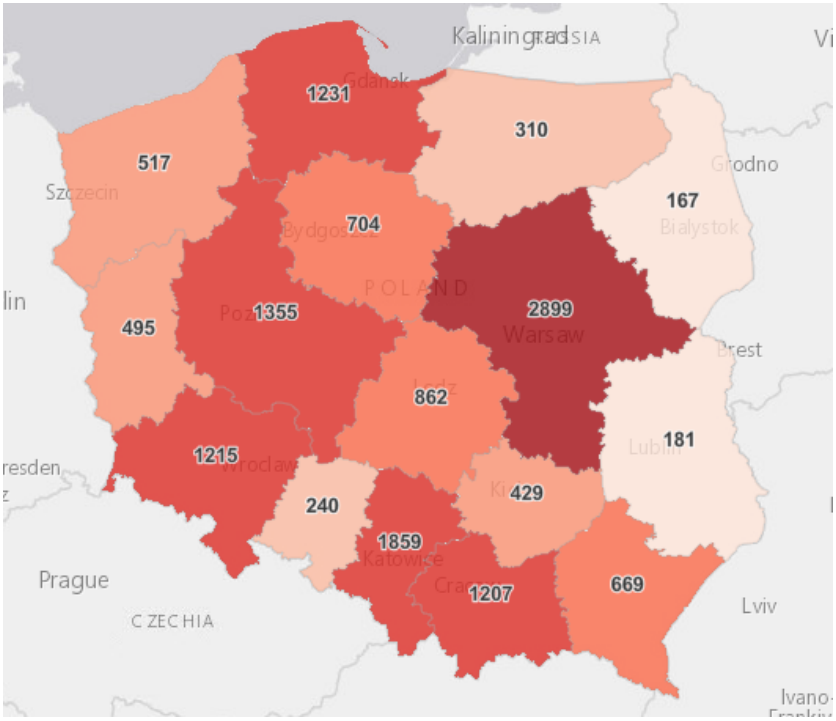
Pośród wielu różnych błędów i manipulacji, których można dokonać z wykorzystaniem statystyki, dość częstym jest mylenie (umyślne lub przypadkowe) wartości *względnych i bezwzględnych*. Najprostszym przykładem tego rodzaju błędu jest przedstawienie na wykresie wartości bezwzględnej pewnej cechy dla np. danych obszarów lub kategorii, podczas gdy z wartości względnych (z uwzględnieniem rozmiaru poszczególnych grup) wynika zupełnie co innego. W ten sposób można zmanipulować społeczeństwo, przykładowo prezentując poparcie wyborców w okręgach wyborczych.

2. Przykład 1: Wartość bezwzględna zarażeń koronawirusem w województwach

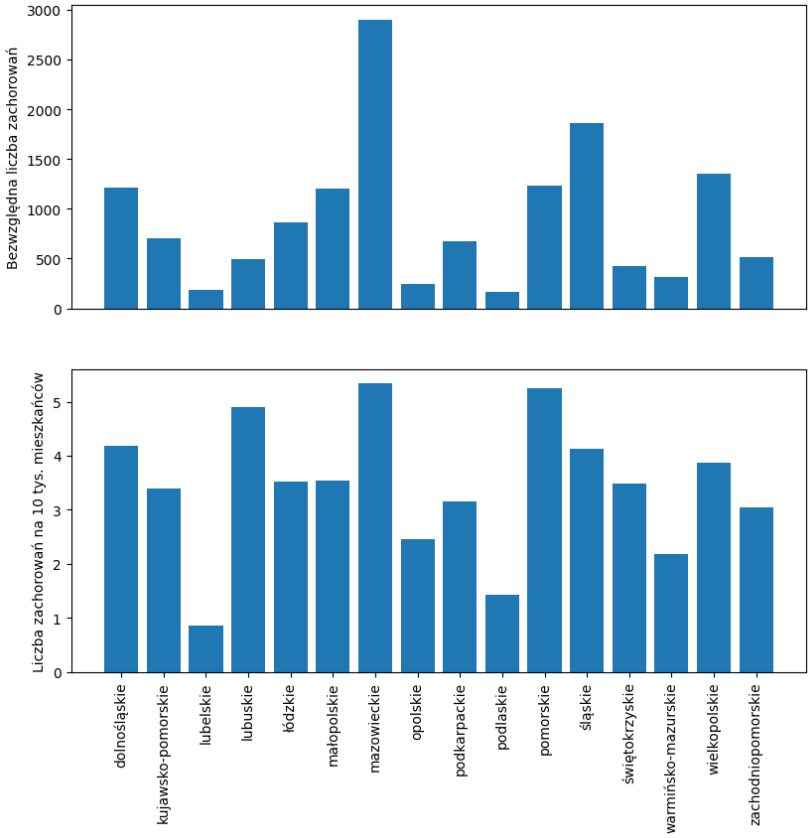
Rysunek 1 (źródło https://www.gov.pl/web/koronawirus/wykaz-zarazen-koronawirusem-sars-cov-2) pokazuje liczbę zarażeń koronawirusem z pojedynczego dnia, dla poszczególnych województw w Polsce. Jak widać z tego rysunku, najgorsza sytuacja jest w województwie mazowieckim, które zdecydowanie wyróżnia się kolorem i wartością. Natomiast dla przykładu sytuacja w województwie lubuskim nie przedstawia się zbyt dramatycznie.

Rysunek 2 prezentuje omawiane dane w nieco inny sposób. Górny wykres przedstawia dokładnie te same wartości co 1, tylko na wykresie słupkowym - *wartość bezwzględną*. Natomiast dolny wykres prezentuje liczbę zakażeń na 10 tys. mieszkańców - *wartość względną*. Jak widać na przykładzie wspomnianych województw mazowieckiego i lubuskiego, poprzednie wnioski nie mają zbyt dużego sensu, kiedy uwzględni się liczbę mieszkańców danego województwa.

Z opisanym tutaj problemem można spotkać się dość często i w zupełnie różnych dziedzinach. Należy więc zawsze uważać, czy mówi się o wartościach względnych czy bezwzględnych i co najważniejsze, które z nich są faktycznie reprezentatywne dla danej dziedziny.

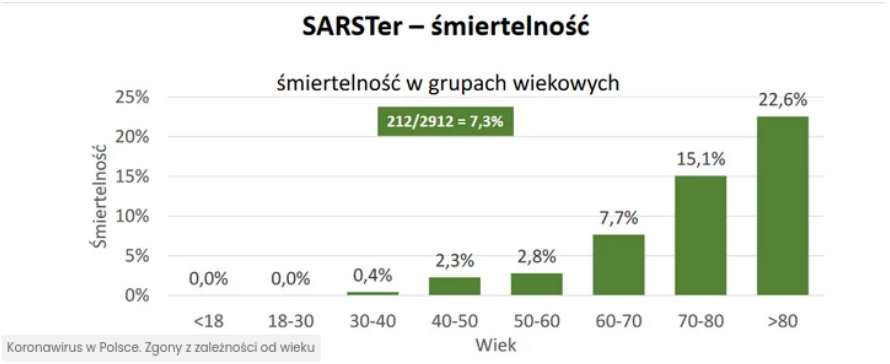


Rysunek 1. Bezwzględna liczba zarażeń koronawirusem w województwach



Rysunek 2. Względna liczba zarażeń koronawirusem w województwach

Wśród zmarłych z powodu czy przy współudziale COVID-19 niemal jedna czwarta (dokładnie 22,6 proc.) miała powyżej 80 lat. 15,1 proc. zmarłych miało od 70 do 80 lat. W przedziale wiekowym 60 - 70 zmarłych było 7,7 proc. W grupie 50+ odsetek zgonów wyniósł 2,8 proc., u pacjentów 40+ było to 2,3 proc.



Rysunek 3. Zgony na koronawirusa w zależności od wieku

3. Przykład 2: Zgony na koronawirusa w zależności do wieku

Drugi przykład jest znacznie bardziej prozaiczny, zawiera natomiast błąd trudny do wychwycenia, na pierwszy rzut oka. Może on być popełniony przypadkowo lub z premedytacją - w tym drugim przypadku byłaby to jednak niezwykle bezczelna manipulacja. Przedstawiony błąd został przedstawiony na rysunku 3 (źródło https://www.medonet.pl/koronawirus/koronawirus-w-polsce,koronawirus-w-polsce–smiertelnosc—co-wiemy-o-ofiarach-,artykul,07236681.html). Na czerwono zakreślono fragment zdania, który zawiera błędną informację. Aby pokazać, że jest ona błędna należy zastanowić się, co przedstawia wykres. Przedstawia on śmiertelność, a więc stosunek chorych, którzy nie przeżyli, do wszystkich chorych z danej kategorii wiekowej. Innymi słowy ze 100 chorych na covid osiemdziesięciolatków około 23 średnio umiera. Natomiast będące częścią artykułu zdanie zawiera zgoła inną informację: ze wszystkich zmarłych z powodu koronawirusa około 23 procent ma powyżej 80 lat. Jak widać jest to więc zupełnie inna wartość: stosunek zmarłych z powodu koronawirusa w wieku powyżej 80 lat do wszystkich zmarłych. Można więc powiedzieć, że ktoś pomylił tutaj znaczenie wartości względnych, dwie zupełnie różne proporcje. Można wysnuć z tego wniosek, że o ile wartości bezwzględne mają znaczenie oczywiste to wartości względne można łatwo zrozumieć źle a nawet opacznie.

Michał Kidawa 239673 239673@edu.p.lodz.pl

Zadanie 1.: Problem Set 1

1. Wprowadzenie

Żyjemy w czasach powszechnej i niezwykle łatwej wymiany informacji. Bardzo często natrafimy na statystyki, które działają na nas opiniotwórczo w ramach określonej dziedziny. Podczas analizy wniosków płynących z danych statystyk warto mieć na uwadze ich źródło oraz czynniki związane z doborem próby. Wpływają one znacznie na uzyskane wyniki i na pierwszy rzut oka są trudne do wykrycia. Zły wybór próby jest niezwykle rozległym tematem. Poniższa analiza problemu obejmuje zagadnienia związane z doborem próby nieodzwierciedlającym rzeczywistego stanu populacji, dobór za małej próby, który prowadzi do uzyskania ekstremalnych wyników oraz wpływ czynnika ludzkiego podczas gromadzenia danych do analizy próby statystycznej.

2. Przykłady błędnego doboru próby do analizy statystycznej

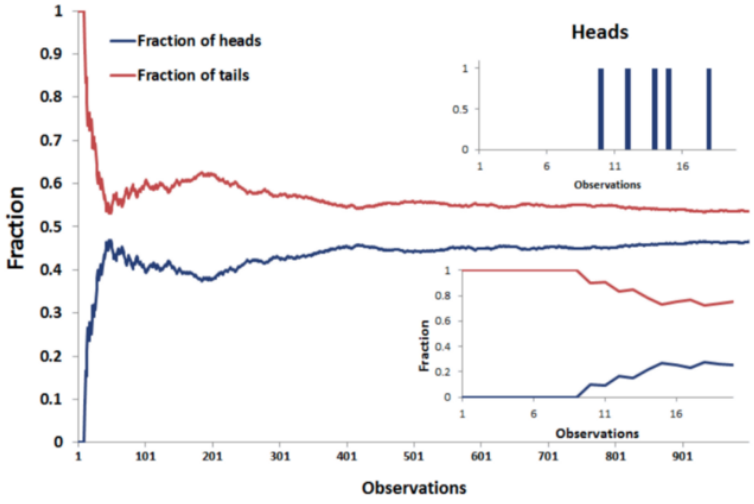
2.1. Próba, które nieodzwierciedla rzeczywistości

Ciekawy przykład doboru próby, która nieodzwierciedlała rzeczywistości można było zaobserwować podczas wyborów prezydenckich w Stanach Zjednoczonych w 1936 roku. Popularny magazyn *Literary Digest* przeprowadził sondę wśród potencjalnych wyborców, która miała wskazać który z kandydatów (Alfred Langdon, czy Franklin D. Roosevelt) zostanie następnym prezydentem Stanów zjednoczonych. Łącznie zebrano dane od około 2.4 miliona osób, co stanowiło gigantyczną liczbę badanych. Były to osoby posiadające telefon i subskrybenci magazynu, które zgodziły się na wzięcie udziału w badaniu. Uczestnicy sondy wskazali Langdona jako następnego prezydenta z przewagą 57% do 43% Roosevelt'a. Wynik okazał się zupełnie odwrotny (62% Roosevelt - 38% Langdon) Może się wydawać, że taka próba była miarodajna. *Literaly Digest* popełniło jednak błąd w doborze osób biorących udział w badaniu. Analizując profil przeciętnego ankietowanego widzimy, że

jest to osoba posiadająca telefon lub subskrypcję magazynu. Można zatem wnioskować, iż była przedstawicielem klasy średniej lub wyższej. W 1936 roku telefon definitywnie był produktem, na który mogła pozwolić sobie znaczna mniejszość społeczeństwa [1]. Tego typu błąd nazywany jest *błędem wyboru*. Oznacza to, że wybór osób, grup lub danych do analizy został przeprowadzony w sposób, który nie zapewnił właściwej randomizacji, powodując w ten sposób, że uzyskana próbka nie jest reprezentatywna dla populacji.

2.2. Małe próby prowadzą do niepoprawnych wyników

Przykłady doboru zbyt małej próby mają miejsce m.in. w branży medycznej i farmakologicznej. Załóżmy, że dokonujemy analizy porównawczą 2 grup badanych, której celem jest określenie częstotliwości występowania powikłań po zastosowaniu terapii. Wykazano, że po określonej terapii w leczeniu że częstotliwość występowania powikłań po zastosowaniu terapii lekiem X wynosi 0.1% co zostało udowodnione w badaniach eksperymentalnych fazy przedklinkinczej. Jeżeli zdefiniujemy grupę pacjentów liczącą 100-200 osób możliwe, że dojdzie do sytuacji w której dane powikłanie może nie wystąpić, co nie znaczy, że dany preparat nie powoduje powikłań. Należy w takim przypadku pamiętać o mocy testu, która jest zależna od wielkości próby użytej w badaniu. Zakładając, że dane zostały wybrane w sposób poprawny większa próba będzie lepiej obrazowała populację. Ponadto w przypadku wybrania mniej licznej próby istnieje szansa, że będą to jednostki wyjątkowe. Problem doboru za małej próby jest ściśle związany z problematyką opisaną w sekcji 2.1. Jeśli liczność próby jest zbyt mała to od razu stanowi niepoprawną reprezentację faktycznej populacji. Załóżmy, że wykonujemy 5 rzutów monetą. Istnieje szansa, że 80% tych rzutów zakończy się "wylosowaniem" tej samej strony. Nie jest to jednak w żaden sposób miarodajna próba i nie zgadza się ona z zasadami prawdopodobieństwa.



Rysunek 1. Wykres prezentujący wyniki rzutu monetą; można zauważyć, że mała liczba prób powoduje ekstremalne wyniki [2]

2.3. Czynnik ludzki przy zbieraniu danych

Istnieją tematy, w przypadku których respondenci mają skłonność do udzielania nieprawdziwych odpowiedzi, aby zaprezentować się w możliwie jak najlepszym świetle. Efekt ten nazwany jest *Efektem społecznych oczekiwań* (ang. social desirability bias). Jest to jeden z najczęściej występujących źródeł błędów, które pływają na wyniki badań ankietowych. Uwydatnia się on szczególnie w przypadku drażliwych społecznie kwestii takich jak religia, polityka czy sprawy osobiste, takie używki lub higiena. Jeśli zadalibyśmy napotkanym na ulicy Polakom pytanie, jak często biorą prysznic uzyskane dane najpewniej byłyby obarczone błędem wynikającej z opisanego powyżej efektu. Ankietowani zawyżaliby celowo liczbę swoich kąpieli ponieważ brak zachowania higieny jest działaniem niepożądanym w społeczeństwie.

3. Rozwiązanie problemów

Aby zapobiegać *błędowi wyboru* należy wykorzystywać metody wybory podgrupy z populacji, które zapewniają jak największą losowość. Jest to możliwe oczywiście do pełnego stopnia, a do tego wykorzystanie odpowiednich metod doboru wiąże się ze zwiększonymi kosztami przeprowadzenia badań. Ponadto warto upewnić się, że wybrana próba jak najlepiej odzwierciedla faktyczny rozkład cech kluczowych względem całej populacji.

Jeśli chodzi o rozmiar próby nie ma jednoznacznego wzoru, który potrafi określić, jaka konkretnie liczność będzie odpowiednia przy danej populacji. Dzięki większej próbie jesteśmy w stanie osiągnąć lepszą moc testu, czyli prawdopodobieństwo niepopełnienia błędu drugiego rzędu.

W celu zwalczania *efektu społecznych oczekiwań* wyróżniono metody, które mają na celu zmniejszenie jego wpływu na wynik badania takie jak użycie testu wymuszonego wyboru, technika losowych odpowiedzi, czy też technika pozornego wariografu [3].

4. Wnioski

Po analizie tematu można dość do wniosku, że:

— przed przyswojeniem i wzięciem za pewnik informacji prezentowanej w postaci statystyk warto poznać metodę zbierania danych
— dobór populacji ma istotny wpływ na wynik prowadzonych badań / eksperymentu
— statystyki mogą być łatwo użyte w celu manipulacji odbiorcą

Literatura

[1] https://www2.math.upenn.edu/ deturck/m170/wk4/lecture/case1.html
[2] https://towardsdatascience.com/lessons-from-how-to-lie-with-statistics-57060c0d2f19
[3] Nederhof AJ. Methods of coping with social desirability bias: A review. „European Journal of Social Psychology". 15 (3), s. 263–80, 1985

Kamil Kowalewski 239676 239676@edu.p.lodz.pl

Zadanie 1.: Problem Set 1

1. Wprowadzenie

Bardzo częstym zjawiskiem jest manipulacja danymi tak, aby zmylić odbiorcę i zmusić go, aby myślał, tak ja chciał autor danego tekstu czy też przekazu. Może być to realizowane w zróżnicowany sposób natomiast poniżej zostanie omówiony problem o nazwie *Korelacja a przyczynowość* (ang. *Correlation vs Causation*). Polega on na powiązaniu dwóch tematyk czy też zjawisk, jakie mają miejsce i są brane pod uwagę w danych badaniach. Można hipotetycznie założyć, że mamy dwie zmienne X oraz Y i wystepuję między nimi korelacja, czyli zależność jednej od drugiej. Zależność ta wyraża się w ten sposób, że przykładowo gdy dwie zmienne X oraz Y wzrastają razem, oraz maleją razem. Celem wyjaśnienia tego można wyróżnić parę scenariuszy, pierwsze z nich jest to, że zjawisko określone jako zmienna X powoduje zjawisko określone przez zmienną Y. Kolejnym przypadkiem jest to, że zjawisko określone przez zmienną Y może powodować zjawisko określone przez zmienną X. Trzecią opcją jest to, że istnieję dodatkowa, trzecia zmienna, która wpływa na wcześniej już wspomniane zmienne X oraz Y. Ostatnią z możliwości jest to, że zmienne X oraz Y są totalnie z sobą niezwiązane a autor artykułu powiązał je, aby osiągnąć swój cel i zmanipulować przesłaniem a co za tym idzie wnioskami, jaki wyciągnie z nich czytelnik.

2. Przykłady błędnie dobranych zjawisk

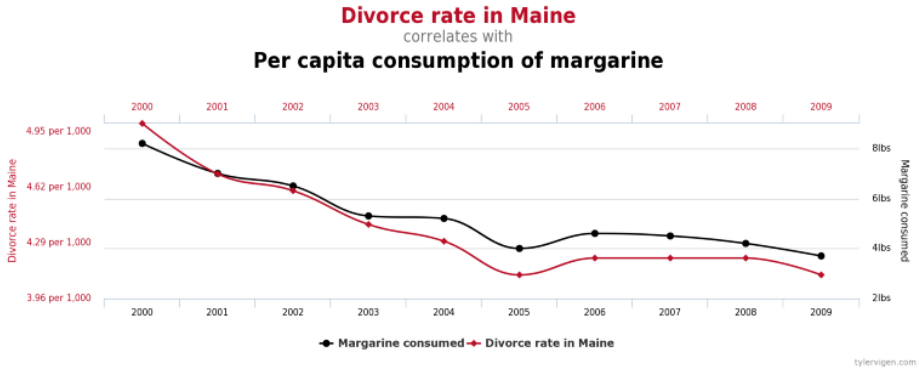
2.1. Wykorzystanie stereotypów i obiegowych opinii

Jednym z ciekawych przykładów obrazujących zjawisko manipulacji jest przedstawiony w artykule [1] sytuacja połączenia palenia papierosów przez uczniów oraz uzyskiwanych przez nich ocen. Autor próbuje narzucić czytelnikowi tezę, że to właśnie przez palenie papierosów uczniowie uzyskują gorsze wyniki w nauce, co jest mocno ugruntowane w społeczeństwie, że osoby, które mają styczność ze środowiskami patologicznym zazwyczaj posiadają

niskie wykształcenie. Autor wykorzystuję tutaj częstą opinię czy też stereotyp. Warto się odnieść tutaj do sekcji 1 gdzie zostało opisane, że zależność może działać w drugą stronę, że właśnie to te złe oceny czy też niepowodzenia w nauce powodują chęć skorzystania ze wcześniej wspomnianej używki celem rozładowania stresu czy też doznania przyjemność z jej przyjmowania. Patrząc na sprawę w wyższego pułapu można też dojść do wniosku, że po prostu osoby z patologicznych rodzin mają skłonność do korzystania z używek w młodym wieku i nie zależy im na edukacji, gdyż po prostu mieli takie wzorce w domu. Jest to przypadek z tą trzecią zmienną przedstawiony w sekcji 1.

2.2. Korelacja zmiennych bez związku

Drugim z przykładów błędnie dobranej korelacji jest zauważenie przez autora korelacji dwóch zmiennych natomiast brak refleksji, że nie są one ze sobą totalnie powiązane. Warto tutaj od razu wspomnieć, że korelacja nie implikuje przyczynowości. To niezmiernie dobrze, że autor zauważył taką pozytywną przemianę w przeciągu tych paru lat przedstawioną na rysunku 1 natomiast powinno zostać to odrzucone jako bezsensowne porównanie.



Rysunek 1. Wykres spożycia margaryny oraz liczby rozwodów w danych latach

2.3. Korelacja tylko na danym przedziale liczbowym

Wartym uwagi jest również fakt, iż wykresy w danych artykułach obejmują pewien ograniczony przedział np. czasu. Zazwyczaj autor dobiera go sobie tak, aby wykazać, to co miał na celu. Niezwykle ważne jest to, że w danym przedziale dwie czy też więcej zmiennych mogą się zachowywać w sposób liniowy natomiast na szerszym przedziale jedna z nich może mieć przebieg logarytmiczny. W takiej sytuacji patrzeć na wykres korelacja może być bardzo słabo zauważalna a wręcz można wyjść z hipotezą, że ona nie występuję. Świetnym przykładem przedstawionym w artykule [2] jest odczucia szczęścia życiowego w stosunku do zarobków rocznych. Przedstawiając tylko początkowy fragment tego wykresu czytelnik może dojść do wniosku, że im większy roczny zarobek tylko wyższy poziom szczęścia, czyli np. Bill Gates byłby w elitarnym gronie osób o najwyższym poziomie szczęścia na naszej planecie. Niestety tak nie jest, ponieważ powyżej pewnej określonej

kwoty ilość pieniędzy przestaja podnosić poziom szczęścia a często nawet go obniża. Stąd jest tak ważna refleksja autora, aby podać dane, które rzetelnie oddają pewną zależność.

3. Rozwiązanie problemów

Rozwiązanie problemów przedstawionych w sekcji 2 z pewnością nie są banalne natomiast poniżej autor tego tekstu chciałbym się odnieść i przedstawić zbiór sugestii, które jego zdaniem mogłyby wyeliminować lub chociaż zmniejszyć liczbę tych błędów.

Odnosząc się do problemu w sekcji 2.1 autorzy tekstów powinni powstrzymać swoje ambicje i ego na rzecz przedstawiania rzetelnych danych i niewprowadzanie w błąd czytelników natomiast sami czytelnicy powinni posiadać wysoki poziom świadomości i podchodzić bardzo krytycznie do przedstawianych im informacji i w miarę możliwości samemu weryfikować informacje, jakie otrzymują, aby nie dać się zmanipulować.

Odnosząc się do problemu w sekcji 2.2 autorzy powinni przed wykorzystaniem danej zależność bardzo dokładnie przemyśleć czy ma ona sens. Prawie każdy inteligentny i myślący odbiorcą z pewnością zauważy bezsensowność porównania, przez co odbiór przygotowane tworu intelektualnego może być negatywny a sama jego ocena może być niezwykle niska. Może to skutkować tym, że opinia o danym autorze będzie niska i z czasem zostanie on wyparty przez innych bardziej rzetelnych twórców.

Odnosząc się do problemu w sekcji 2.3 w szczególności odbiorcy powinni być szczególne uwrażliwieni na tę sytuację, gdyż mogą być łatwy sposób zmanipulowani. Po stronie autorów powinien istnieć pewien kodeks moralny rzetelnego przedstawiania danych. Gdy nawet nie mają możliwość pełnego przedstawienia powinni zasygnalizować o fakcie, że przedstawione dane mogą dawać mylny obraz sytuacji.

4. Wnioski

Podsumowując można powiedzieć, że:

— przedstawiony problem nie jest trywialny i wymaga szczególnej uwagi ze strony autorów oraz czytelników
— dobór zmiennych oraz finalna decyzja o i wykorzystaniu to zagadnienie rozmyte i niejednoznaczne, musi zostać pozostawione wyczuciu osoby przygotowującej dane treści analityczne
— czytelnicy powinni być niezwykle ostrożni i nie powinni wyrabiać sobie opinii na dany temat czytając tylko zasoby z autorów o pewnym zbliżonym stylu myślenia i światopoglądzie

Literatura

[1] https://medium.com/seek-blog/how-to-lie-with-statistics-b671b66399d
[2] https://towardsdatascience.com/lessons-from-how-to-lie-with-statistics-57060c0d2f19