Przeprowadzenie szkoleń dla pracowników administracji publicznej z zakresu analizy danych w ramach oceny skutków regulacji

Fundusze Europejskie
Wiedza Edukacja Rozwój

Rzeczpospolita Polska

Unia Europejska
Europejski Fundusz Społeczny

Skrypt do części: Narzędzia badawcze (jakościowe i ilościowe)

Spis treści

TWORZENIE NARZĘDZI BADAWCZYCH DO BADAŃ JAKOŚCIOWYCH

Aby stworzyć scenariusz wywiadu należy:

- Ustalić listy tematów i pytań, które są kluczowe i będą pojawiać się zawsze podczas różnych wywiadów.
- Przygotować scenariusz – należy zapisać pytania i je odpowiednio pogrupować.
- Pamiętać, że pytania powinny być otwarte, pogłębiające, mogą być eksperckie – ważne jest usłyszenie przemyśleń i opinii, nie tylko faktów.
- Oszacować czas, jaki ma zostać przeznaczony w trakcie wywiadu na poszczególne zagadnienia.
- Pamiętać, żeby zapraszając rozmówcę do udziału w badaniu przedstawić mu cel wywiadu.

Pytania w scenariuszu

W scenariuszu powinny pojawić się pytania:

- Otwierające.
- Ułożone w kolejności od ogółu do szczegółu.
- Ułożone w sekwencje – nie należy „skakać” po tematach.
- Dobrze sformułowane, to znaczy:
  - bez przeczeń,
  - krótkie i konkretne,
  - pojedyncze (nie o dwie kwestie jednocześnie)
- Które są zadane językiem rozmówcy.

Czy jest coś jeszcze?” – na koniec należy dopytać o kwestie, które rozmówca chciałby uzupełnić w kontekście wywiadu.

- Tworząc scenariusz trzeba wybrać te pytania, które są kluczowe w ramach danego obszaru. Można dodatkowo zapisać komentarze dla badacza.

Techniki zdobywania informacji

Przy tworzeniu jakościowych narzędzi badawczych można zastosować techniki:

- Bezpośrednie – dopytywanie: w jaki sposób? dlaczego? jakie są przyczyny?
- Pośrednie – „wchodzenie w czyjeś buty”: wyobraź sobie...

Należy pamiętać, że bardzo ważne jest aktywne słuchanie.

TWORZENIE NARZĘDZI BADAWCZYCH DO BADAŃ ILOŚCIOWYCH

Aby stworzyć kwestionariusz ankiety należy:

- Ustalić listę poruszanych tematów.
- Ustalić odpowiednią kolejność pytań:
    - można użyć metryczki „filtrującej” na początku, żeby upewnić się, że respondent jest z grupy docelowej np. czy jest beneficjentem jakiegoś programu (chodzi o 2 – 3 pytania weryfikujące).
    - należy zacząć od prostych, ciekawych pytań, pytania drażliwe należy zostawić na koniec.
    - od ogółu do szczegółu.
    - stopniowe wyczerpywanie tematu - pytania związane z nowym zagadnieniem zadajemy po wyczerpaniu pytań związanych z poprzednim logicznym ciągiem pytań.
    - pytania metryczkowe należy zadać na koniec (chodzi o wszystkie kwestie, które różnicują rozmówców i są ważne z punktu widzenia analizy danych, czyli będziemy je krzyżować z odpowiedziami np. dochody, płeć, stanowisko)
- Informować o zmianie tematyki pytań.
- Ostrożnie stosować pytania filtrujące.
- Nie należy zbierać danych, które nie są potrzebne.

Typy pytań

- Pytania otwarte i zamknięte – należy unikać stosowania pytań otwartych.
- Pytania zamknięte posiadają kafeterię:
    - prostą – jedna odpowiedź, opcje wzajemnie się wykluczają;
    - złożoną – możliwość kilku odpowiedzi.
- Kiedy nie znamy wszystkich opcji odpowiedzi wtedy stosujmy pytania „półotwarte”, np.: „inne – jakie?”

Skale, czyli rodzaje porządkowania odpowiedzi

Rangowa

Przykład: „Uporządkuj podane kryteria wyboru dostawcy od najważniejszego (1) do najmniej ważnego (4)” (np. cena, jakość, terminowość dostaw, doświadczenia ze współpracy).

Skala sumowanych ocen

Przykład: „Rozdziel 100 punktów między podane kryteria wyboru dostawcy, tak aby suma punktów odpowiadała znaczeniu danego kryterium”.

Pozycyjna/porządkowa

Przykład: „Jak często kontaktowałeś/-aś się z infolinią?”

„bardzo często”, „często”, „czasami”, „nigdy”.

Semantyczna (dyferencjał somatyczny)

Odpowiedzi są na skali dwubiegunowej:

Przykład: „Leki kupowane przez Panią są:”

tanie I I I I I I I drogie

Skala Stapela

Przykład: „Proszę ocenić pracowników punktu obsługi mieszkańców?”

„wiedza -3 -2 -1 +1 +2 +3

komunikatywność -3 -2 -1 +1 +2 +3”

Skala dystansu społecznego Bougardusa

Przykład:

„Czy zgodził(a)byś, żeby imigranci z krajów muzułmańskich mieszkali w Twoim kraju?”

„Czy zgodził(a)byś, żeby imigranci z krajów muzułmańskich mieszkali w Twojej miejscowości?”

„Czy zgodził(a)byś, żeby imigranci z krajów muzułmańskich był Twoim sąsiadem?”

„Czy zgodził(a)byś, żeby Twoja córka poślubiła imigranta z krajów muzułmańskich?”

KODOWANIE DANYCH

Kodowanie danych:

- Proces redukcji danych poprzez stworzenie ogólniejszych kategorii, do których będą klasyfikowane odpowiedzi respondentów.
- Można kodować dane zebrane metodami ilościowymi i jakościowymi.
- Do kodowania można wykorzystać specjalistyczne oprogramowania (np. Atlas.ti, SPSS, Microsoft Excel) jednak nie zawsze jest to konieczne (np. w przypadku małych zbiorów danych jakościowych)

Rekodowanie danych:

- Zmiana wartości liczbowych przypisanych danym kategoriom zmiennych
- Stosowane w celu umożliwienia przeprowadzenia określonych testów statystycznych lub uproszczenia informacji, tak aby były one łatwiejsze do przetworzenia i interpretacji

Kodowanie danych może odbywać się zgodnie z:

- **celami i pytaniami badawczymi** - zadaniem jest przygotowanie takich kodów, które pozwolą na dopasowanie odpowiedzi do problemów związanych z pytaniami badawczymi i przyporządkowanie ich do udzielonych odpowiedzi
- **z danymi** – większe znaczenie przy tworzeniu kodów ma liczebność odpowiedzi oraz to na ile są to siebie podobne niż konieczność dopasowania ich do pytań badawczych

Budowa książki kodowej

Książka kodowa to dokument, który opisuje pozycje zmiennych i wymienia kody przyporządkowane wartościom składającym się na zmienne. Pomaga ona w procesie kodowania oraz porządkuje posiadane informacje i zmienne.
Książka kodowa powinna zawierać:

- definicje każdej zmiennej (np. literalnie przytoczonego badania),
- połączenia z innymi zmiennymi (rodziny kodów).

TESTOWANIE HIPOTEZ STATYSTYCZNYCH

Hipoteza statystyczna to dowolne przypuszczenie dotyczące rozkładu populacji. Może być sprawdzana za pomocą analizy danych i testów statystycznych, aby określić, czy jest ona prawdziwa czy fałszywa.

Najlepszym sposobem, aby ustalić czy hipoteza statystyczna jest prawdziwa, byłoby przebadanie całej populacji. Jeśli nie jest to możliwe, bada się próbę. Jeżeli próbka nie potwierdza hipotezy, wtedy hipoteza zostaje odrzucona.

Proces testowania hipotezy:

1. Sformułowanie hipotezy zerowej i alternatywnej:

- Hipoteza zerowa H_0 - jest to hipoteza poddana procedurze weryfikacyjnej, w której zakładamy, że różnica między analizowanymi parametrami lub rozkładami wynosi zero.
- Hipoteza alternatywna H_1 - hipoteza przeciwstawna do weryfikowanej.

2. Wybór statystyki testowej: budujemy pewną statystykę W, która jest funkcją wyników z próby losowej $W = f(x_1, x_2, \dots, x_n)$ wyznaczamy jej rozkład, przy założeniu, że hipoteza zerowa jest prawdziwa. Funkcję W nazywa się statystyką testową lub funkcją testową.
3. Określenie poziomu istotności - na tym etapie procedury weryfikacyjnej przyjmujemy prawdopodobieństwo popełnienia błędu I-go rodzaju, który polega na odrzuceniu hipotezy zerowej wtedy, gdy jest ona prawdziwa. Przyjmujemy prawdopodobieństwo bliskie zeru, ponieważ chcemy, aby ryzyko popełnienia błędu było jak najmniejsze. Najczęściej zakładamy, że poziom istotności $\alpha \leq 0{,}1$.
4. Wyznaczenie obszaru krytycznego testu: obszar krytyczny - obszar znajdujący się zawsze na krańcach rozkładu. Jeżeli obliczona przez nas wartość statystyki testowej znajdzie się w tym obszarze, to weryfikowaną przez nas hipotezę H_0 odrzucamy.
5. Obliczenie statystyki na podstawie próby: wyniki próby opracowujemy w odpowiedni sposób, zgodnie z procedurą wybranego testu i są one podstawą do obliczenia statystyki testowej.
6. Podjęcie decyzji: wyznaczoną na podstawie próby wartość statystyki porównujemy z wartością krytyczną testu:
   - jeżeli wartość ta znajdzie się w obszarze krytycznym, to hipotezę zerową należy odrzucić jako nieprawdziwą, skąd wniosek, że prawdziwa jest hipoteza alternatywna;
   - jeżeli natomiast wartość ta znajdzie się poza obszarem krytycznym, oznacza to, że brak jest podstaw do odrzucenia hipotezy zerowej, skąd wniosek, że hipoteza zerowa może, ale nie musi, być prawdziwa.

Ważne: Jeśli nie uda nam się odrzucić H_0 na rzecz H_1, to nie znaczy, że jest na pewno prawdziwa.

Przeprowadzenie szkoleń dla pracowników administracji publicznej z zakresu analizy danych w ramach oceny skutków regulacji

Fundusze Europejskie
Wiedza Edukacja Rozwój

Rzeczpospolita Polska

Unia Europejska
Europejski Fundusz Społeczny

Skrypt do części: Statystyka i wyciąganie wniosków

Spis treści

ZASTOSOWANIE METOD STATYSTYCZNYCH W OCENIE SKUTKÓW REGULACJI Z WYKORZYSTANIEM RSTUDIO

Analiza statystyczna pełni kluczową rolę w ocenie wpływu regulacji na różne aspekty życia społecznego i gospodarczego. Pomaga w zbieraniu danych, identyfikowaniu wzorców oraz ocenie istotności różnic i związków między danymi. Dzięki temu umożliwia obiektywne podejście do oceny skutków regulacji. W niniejszym skrypcie prezentowane są praktyczne przykłady zastosowania analizy statystycznej w ocenie skutków regulacji przy wykorzystaniu programu RStudio oraz rzeczywistych danych z różnych źródeł, takich jak dane GUS, Eurostat czy raporty OECD.

Analiza statystyczna jest niezbędna na najwyższych szczeblach zarządzania, umożliwiając podejmowanie uzasadnionych decyzji opartych na faktach w polityce publicznej (Greenstone, 2002). Badania skutków regulacji mogą skupiać się na różnych obszarach, takich jak:

- Gospodarka: Analizowanie wpływu regulacji podatkowych na dochody państwa i przedsiębiorstw oraz efektów polityki fiskalnej na wzrost gospodarczy i zatrudnienie.
- Ochrona zdrowia: Ocena skutków reformy służby zdrowia na dostępność do usług medycznych oraz badanie efektywności działań w zwalczaniu epidemii, np. COVID-19.
- Edukacja: Badanie wpływu zmian w systemie edukacji na wyniki uczniów i jakość nauczania oraz ocena dostępności do edukacji przedszkolnej i wyższego szczebla.
- Środowisko: Monitorowanie skutków regulacji ochrony środowiska, np. emisji CO2 i innych zanieczyszczeń oraz ocena wpływu inwestycji w odnawialne źródła energii na ekologię.
- Infrastruktura: Ocena skutków inwestycji w rozwój infrastruktury, np. budowę dróg i kolei, oraz badanie efektywności programów modernizacji miast i regionów.
- Kultura i nauka: Ocena wpływu inwestycji w kulturę i naukę na rozwój tych dziedzin oraz badanie efektywności programów promocji kultury i nauki.

W kolejnych częściach skryptu omawiane są konkretne techniki statystyczne, takie jak analiza opisowa, korelacja, regresja i testy statystyczne, wykorzystywane do oceny skutków regulacji. Przedstawiono także rzeczywiste dane używane w

analizach, pochodzące z różnych źródeł, takich jak dane GUS, Eurostat czy raporty OECD.

Skrypt uzupełniony jest zawartymi w Załączniku 1 przykładami, zawierającymi dane, skrypty programu RStudio, wydruki realizacji skryptów i raporty analiz statystycznych do poszczególnych przypadków. Załącznik 2 zawiera podstawowe instrukcje statystyczne w programie RStudio, które pozwolą Uczestnikom kursu utrwalić i poszerzyć swoją wiedzę na temat pisania prostych programów.

ANALIZA OPISOWA Z WYKORZYSTANIEM RSTUDIO

Statystyka jest dziedziną matematyki, która pomaga opisywać i analizować dane liczbowe, co jest kluczowe w badaniu skutków regulacji w gospodarce. Statystyka opisowa to jej podstawowy dział, który pozwala nam zrozumieć, jak dane się zachowują. W tym kontekście, używa się różnych miar, takich jak średnia, mediana, czy odchylenie standardowe, aby opisać tendencje i rozproszenie danych. Przykładowo, możemy wykorzystać statystykę opisową do analizy wskaźników gospodarczych, takich jak wynagrodzenia, wskaźniki rozwoju czy stopa bezrobocia. Dzięki niej możemy szybko uzyskać ogólny obraz sytuacji gospodarczej, a także określić, czy regulacje wprowadzone przez rząd miały pozytywny czy negatywny wpływ na dane wskaźniki.

Statystyka to dział matematyki stosowanej. Opisuje tendencje w populacji (generalnej lub próbnej) za pomocą liczb i funkcji matematycznych (minimum, maksimum, średnia, mediana), występujących w zjawiskach masowych. Te ostatnie występują (mogą być zbadane) teoretycznie nieskończona ilość razy, np. urodzenia, małżeństwa, płeć, wynagrodzenia, wiek, wzrost, waga, przychody itd.

Populacja składa się zaś z obiektów (przypadków), które charakteryzują się różnymi właściwościami:

- Jakościowymi – kategorialnymi: np. płeć, okresy przed i po regulacji, sektory gospodarki, rodzaje usług, kierunek wykształcenia.
- Porządkowymi (ilościowe dyskretne, niemetryczne): różnorakie rankingi, np. miejsca Polski pod względem poziomu innowacyjności, czy wskaźnik rozwoju ludzkości (Human Development Index HDI).
- Ilościowymi (ilościowe, metryczne): wiek, wynagrodzenia, dochody, wskaźniki, np. urodzeń, czy zgonów.

Właściwości zmierzone na pewnych skalach pomiarowych i wyrażone za pomocą liczby lub symbolu, nazywamy albo cechami statystycznymi, jeżeli badanie było przeprowadzone na całej populacji generalnej, albo zmiennymi losowymi, gdy pomiarów dokonywano na próbie losowej.

W zarządzaniu i ocenie skutków regulacji można wykorzystać zarówno statystykę opisową, jak i statystykę indukcyjną. Statystyka opisowa charakteryzuje za pomocą miar właściwości obiektów konkretnej populacji, np. gospodarkę Polski, albo gospodarkę jakiegoś jednego województwa na podstawie wskaźników (np. urodzeń, dochodów, wskaźnika ogólnego klimatu koniunktury w budownictwie, itp.).

Obiektami populacji i jej badania mogą być osoby fizyczne, organizacje, województwa, kraje czy kontynenty, a właściwościami ich konkretne charakterystyki. Właściwością osoby fizycznej w badaniu może być jej płeć, waga czy wzrost. Właściwościami organizacji, np. niefinansowego przedsiębiorstwa mogą być kwartalne wskaźniki makroekonomiczne. Właściwości zmierzone na określonej skali i wyrażone za pomocą liczby lub symbolu nazywane są cechami statystycznymi (jeżeli badanie dotyczy całej populacji generalnej) lub zmiennymi (gdy badanie przeprowadza się na próbie losowej) i oznacza się je zwykle dużymi literami alfabetu łacińskiego, np. X, Y, Z, a ich konkretne realizacje – małymi literami, np. x=5). Populację generalną wyznacza się podając co najmniej jedną cechę (np. wiek osoby - X) i co najmniej jedną wartość (np. osoby dorosłe w wieku x>18).

Cechy/zmienne charakteryzują pewne rozkłady. Rozkład cechy, np. wynagrodzeń w Polsce za rok 2022, oznacza zbiór wszystkich wartości, jakie ta zmienna może przyjąć wraz z odpowiadającymi im liczebnościami, częstościami czy prawdopodobieństwami. Rozkład cechy/zmiennej można przedstawić za pomocą np. histogramu liczebności/prawdopodobieństwa. Strukturę populacji (np. udziały procentowe jakichś kategorii, np. udziały kobiet i mężczyzn w populacji osób bezrobotnych, sumujące się do 100 procent) przedstawia się za pomocą diagramu kołowego. Najpowszechniej występującym w przyrodzie rozkładem jest tzw, rozkład normalny – idealnie symetryczny względem średniej.

Miary opisu statystycznego to miary tendencji centralnej (np. średnie, mediana, kwartyle), rozproszenia, albo inaczej zróżnicowania (np. odchylenie standardowe -

typowa odległość, jakiej można spodziewać się w stosunku do średniej w badanej populacji, współczynnik zmienności względem średniej – miara względna informująca o jednorodności czy też rozproszeniu danych), Współczynnik zmienności względem średniej pomaga ocenić, jak bardzo dane zmieniają się w stosunku do swojej średniej wartości. Im wyższy współczynnik zmienności, tym większa zmienność w danych. To istotne narzędzie, które pomaga w ocenie skutków regulacji gospodarczych. Z kolei miary skośności – inaczej asymetrii, wskazują, czy dominują wyniki niskie, przeciętne, czy wysokie, a koncentrację wyników wokół średniej charakteryzują: kurtoza i eksces.
Ważne są także miary struktury, które pozwalają określić udział danej kategorii w masie wszystkich obserwacji, jak proporcje, odsetki, procenty czy promile, a także stosunki między liczebnościami różnych kategorii.

Statystyka opisowa nie wykracza z uogólnieniami na populację generalną, służy temu statystyka indukcyjna, oparta na rachunku prawdopodobieństwa. Polega ona na uogólnianiu wniosków z próby losowej na populację generalną, za pomocą takich metod, jak estymacja punktowa, konstruowanie przedziałów ufności, czy weryfikacja hipotez statystycznych, co zostanie przedstawione na przykładach w dalszych rozdziałach (Załącznik 1).

Prawdopodobieństwo w statystyce to liczba lub miara wyrażająca stopień możliwości lub szansy wystąpienia danego zdarzenia losowego. Jest to sposób matematycznego opisania naszego przekonania lub wiedzy na temat tego, jakie wyniki można spodziewać się w wyniku eksperymentu lub losowego zjawiska. Prawdopodobieństwo przyjmuje wartości od 0 (oznacza brak szans na wystąpienie zdarzenia) do 1 (oznacza pewność wystąpienia zdarzenia) i pomaga nam ocenić ryzyko, prognozować wyniki i podejmować decyzje oparte na danych i losowych zdarzeniach. Na przykład, rachunek prawdopodobieństwa może pomóc oszacować, jakie są szanse na wystąpienie zatrucia pokarmowego przed i po wprowadzeniu regulacji związanych z produkcją pewnych artykułów żywnościowych, co pomaga ocenić skuteczność regulacji w ochronie zdrowia publicznego.

Statystyka indukcyjna to dział matematyki stosowanej oparty na rachunku prawdopodobieństwa, który umożliwia uogólnianie wniosków z próbek losowych na

cała populację generalną. Służą temu m.in. trzy sposoby wnioskowania statystycznego, wyjaśnione poniżej.

Estymacja punktowa: To sposób szacowania (przewidywania) pewnego parametru (na przykład średniej, czy odchylenia standardowego) w populacji generalnej na podstawie próbki losowej . Na przykład, jeśli chcemy poznać przeciętny wiek wszystkich mieszkańców miasta, możemy użyć próbki mieszkańców i obliczyć średni wiek tej próbki. Ten średni wiek w próbce jest naszym punktowym oszacowaniem średniego wieku w całej populacji.

Estymacja przedziałowa: Tutaj nie tylko szacujemy wartość, ale również określamy pewien przedział, w którym prawdopodobnie znajduje się ta wartość w populacji. Przykładowo, zamiast tylko podać jedną średnią wartość wieku mieszkańców miasta, możemy wnioskować, że jest on zawarty w przedziale liczbowym, np. między 30 a 35 lat z prawdopodobieństwem równym 0,95.

Weryfikacja hipotez statystycznych: To proces sprawdzania, czy nasze obserwacje są wystarczająco przekonujące, aby podjąć decyzję na temat pewnych twierdzeń dotyczących populacji. Na przykład, możemy postawić hipotezę, że średni wiek w dwóch różnych miastach jest taki sam. Następnie zbieramy dane i używamy statystyki, aby ocenić, czy nasza hipoteza jest prawdziwa czy fałszywa.

Tak więc, statystyka indukcyjna pozwala nam na dokonywanie ogólnych wniosków na podstawie ograniczonej ilości danych (próbki). To narzędzie jest przydatne w wielu dziedzinach, od nauki społecznych po nauki przyrodnicze, ponieważ umożliwia nam podejmowanie decyzji i formułowanie wniosków opartych na danych.

ANALIZY STATYSTYCZNE A ŚRODOWISKO RSTUDIO

Wszystkie opisane metody można w dość prosty sposób zastosować, posługując się programami statystycznymi. Warto mieć przewodnik statystyczny, który podpowiada badaczowi, jakie metody i techniki zastosować, jakie procedury uruchomić, i jak – żeby uzyskać komplet miar opisowych. W analizie statystycznej często korzysta się z narzędzi informatycznych, takich jak RStudio, które ułatwiają przetwarzanie danych i generowanie wyników.

Uruchomienie RStudio polega na zainstalowaniu programu na komputerze i jego uruchomieniu jak zwykłej aplikacji. Po otwarciu RStudio, można korzystać z konsoli R do interaktywnej pracy z językiem R. Do tworzenia skryptów służy edytor skryptów w RStudio, gdzie można strukturyzować kod, dodawać komentarze i uruchamiać go krok po kroku. Skrypty można zapisywać i wielokrotnie wykorzystywać, co ułatwia pracę nad długotrwałymi projektami analizy statystycznej.

Niniejszy skrypt zawiera proste przykłady, jak wykorzystać statystykę opisową i elementy statystyki indukcyjnej w analizach dotyczących skutków regulacji, z wykorzystaniem programu R i RStudio. Program RStudio ułatwia wykonywanie analiz statystycznych, jest dość prostym, darmowym i łatwo dostępnym narzędziem w analizie statystycznej, oferując wiele zalet i ułatwień. Jest to zintegrowane środowisko programistyczne (IDE) stworzone dla języka R, często używanego w analizie danych. RStudio wspiera zarządzanie projektami, interakcję z danymi oraz tworzenie i uruchamianie skryptów statystycznych (por. Walesiak, & Gatnar, 2009).

IMPORT I PRZETWARZANIE DANYCH W RSTUDIO

W tym punkcie omówimy proces importu danych do środowiska RStudio oraz przetwarzania tych danych w celu przygotowania ich do analizy. Wyjaśnimy również instrukcje setwd(), read.table() oraz pokażemy przykład tworzenia wektora danych. Import danych do RStudio: Jednym z pierwszych kroków w analizie danych jest importowanie danych z różnych źródeł, takich jak pliki CSV, Excel, bazy danych czy strony internetowe. W RStudio proces importu jest stosunkowo prosty i można go dostosować do różnych formatów danych. Oto przykład importu danych z pliku "learning2.txt" w formacie CSV (tabulator jako separator kolumn) i ich prezentacji jako ramki danych:

 # Ustalamy katalog roboczy (opcjonalne)
setwd("C:/Program R") # Ustawienie katalogu roboczego
 # Importujemy dane z pliku "learning2."
mydata <- read.table("learning2.txt", sep="\t", header=TRUE) # czytanie tabeli danych o nazwie „learning2.txt”, separatorem pół (zmiennych, danych, umieszczonych w kolumnach) jest znak tabulatora (sep="\t"), a dane zawieraja nagłówki (header=TRUE)

Instrukcja setwd() pozwala ustawić katalog roboczy, czyli miejsce, w którym RStudio będzie szukać plików do importu. Instrukcja read.table() jest używana do importu danych z pliku tekstowego i ustala, że dane są oddzielone tabulatorem (sep="\t"), a pierwszy wiersz pliku zawiera nazwy kolumn (header=TRUE).

Przetwarzanie danych:
Przetwarzanie pozwala na przygotowanie danych poprzez usuwanie błędów, dostosowywanie formatów i zapewnianie ich spójności. Oto przykłady procesu przetwarzania danych:

Przykład tworzenia wektora danych:
x <- c(22.8, 46.1, 70.7, 100.8) # zmienna x zawiera kolejne wartości oddzielone przecinkiem 22.8, 46.1, itp. W zapisie dziesiętnym używamy kropki w celu oddzielenia części dziesiętnych.

x <- c(22.8, 46.1, 70.7, 100.8) # zmienna x zawiera kolejne wartości oddzielone przecinkiem 22.8, 46.1, itp. W zapisie dziesiętnym używamy kropki w celu oddzielenia części dziesiętnych. Proces importu i przetwarzania danych w RStudio jest kluczowy dla efektywnej analizy statystycznej. Instrukcje setwd() i read.table() pozwalają na import danych, a operacje przetwarzania pozwalają na oczyszczenie danych **przed dalszą analizą.**

PODSTAWOWA ANALIZA OPISOWA W RSTUDIO

Niniejsza sekcja zawiera przykłady obliczania miar tendencji centralnej, rozproszenia oraz rysowania wykresów (np. histogramów) za pomocą RStudio, a także przykładową interpretację danych (patrz Załącznik 1, Przykład 1).

Przykład 1 – emisja CO2:
Oto hipotetyczny zbiór danych dotyczący oceny skutków regulacji w dziedzinie ochrony środowiska. Ten zestaw danych zawiera oceny skutków regulacji na emisję dwutlenku węgla (CO2) przez różne branże przemysłowe:

Emisja CO2 po regulacji (w tonach na rok): [100, 150, 90, 120, 200, 80, 110, 130, 170, 95, 105, 180, 160, 140, 115, 125, 190, 170, 210, 95]

Możemy teraz użyć tego zestawu danych, aby obliczyć różne statystyki, takie jak średnia, mediana i odchylenie standardowe. Poniżej w sekcji A znajduje się skrypt do programu RStudio, a w sekcji B – interpretacja wyników.

Sekcja A – skrypt do programu RStudio:

```
# Najważniejsze charakterystyki liczbowe oceny skutków regulacji na emisję dwutlenku węgla (CO2) przez różne branże przemysłowe
x <- c(100, 150, 90, 120, 200, 80, 110, 130, 170, 95, 105, 180, 160, 140, 115, 125, 190, 170, 210, 95)
# Obliczamy średnią
mean_emisja <- mean(x)
cat("Średnia emisja CO2:", mean_emisja, "ton na rok\n")
# Obliczamy medianę
median_emisja <- median(x)
cat("Mediana emisji CO2:", median_emisja, "ton na rok\n")
# Obliczamy odchylenie standardowe
sd_emisja <- sd(x)
cat("Odchylenie standardowe emisji CO2:", sd_emisja, "ton na rok\n")
```

Sekcja B – interpretacja wyników:

- Średnia emisja CO2 wynosi około 137,9 ton na rok. Średnia jest miarą tendencji centralnej, która informuje nas o typowej emisji CO2 w badanych branżach przemysłowych.
- Mediana emisji CO2 wynosi 132,5 ton na rok. Mediana jest wartością środkową w zbiorze danych i jest mniej podatna na wpływ wartości skrajnych niż średnia. W tym przypadku mediana jest zbliżona do średniej, co sugeruje, że dane są względnie równomiernie rozproszone wokół wartości centralnej.
- Odchylenie standardowe emisji CO2 wynosi około 39,9 ton na rok. Jest to miara rozproszenia danych wokół średniej. Wyższe odchylenie standardowe oznacza większą zmienność emisji CO2 w badanych branżach.

Te miary statystyczne pozwalają na lepsze zrozumienie rozkładu emisji CO2 w badanych branżach przemysłowych i są kluczowe w ocenie skutków regulacji ochrony środowiska.

Podsumowanie

Analiza statystyczna jest nieodłącznym narzędziem w ocenie skutków regulacji prawnych w różnych obszarach, takich jak gospodarka, ochrona zdrowia, edukacja, środowisko, infrastruktura, kultura i nauka. Przedstawiony skrypt w RStudio zawiera przykłady importu, przetwarzania danych oraz podstawowej analizy opisowej. W kolejnych częściach skryptu można by kontynuować analizę, stosując bardziej zaawansowane techniki statystyczne, takie jak regresja czy testy statystyczne, w celu głębszej oceny skutków regulacji.

ANALIZA PORÓWNAWCZA ZA POMOCĄ MIAR OPISU STATYSTYCZNEGO DWÓCH OKRESÓW PRZED I PO WPROWADZENIU REGULACJI

W niniejszym rozdziale przeprowadzono analizę porównawczą w celu oceny wpływu wprowadzenia regulacji mających na celu kontrolę emisji pyłów PM2.5 w mieście, zarówno przed, jak i po wprowadzeniu tych regulacji (Załącznik 1, Przykład 2). Cząstki stałe o średnicy mniejszej niż 2,5 mikrometra (PM2.5) stanowią potencjalne zagrożenie dla zdrowia ludzi. Regulacje te zostały nazwane "Programem Kontroli Emisji Pyłów PM2.5" i mają na celu kontrolę oraz ograniczenie emisji tych szkodliwych cząstek, w celu ochrony zdrowia publicznego oraz środowiska.

Źródłem danych w analizie były pomiary poziomów PM2.5 w jednym konkretnym mieście. Jednak warto zaznaczyć, że regulacje mogą być wprowadzane na różnych poziomach administracyjnych (krajowym, regionalnym lub lokalnym), co wpływa na źródło danych. W celu zachowania jednolitości i odpowiedniego porównania, wykorzystano dane z jednego obszaru. Szereg danych przed regulacją (poziomy PM2.5 w mikrogramach na metr sześcienny w różnych dniach w 2019 roku) to: [35, 42, 39, 45, 38, 40, 36, 47, 50, 32, 55, 48, 43, 52, 46, 37, 41, 44, 49, 51]. Natomiast szereg danych po wprowadzeniu regulacji (poziomy PM2.5 w mikrogramach na metr sześcienny w tych samych dniach w 2022 roku) to: [30, 35, 32, 25, 28, 31, 29, 27, 22, 30, 26, 23, 24, 28, 33, 31, 26, 29, 30, 27].

W celu porównania tych danych, przeprowadzono analizę statystyczną, obejmującą miary tendencji centralnej (średnią, medianę, kwartyle) oraz miary dyspersji (odchylenie standardowe względem średniej) (patrz Załącznik 1, Przykład 2). Dodatkowo, przedstawiono wykresy pudełkowe w celu wizualizacji rozkładu danych.

Wyniki analizy wskazują, że wprowadzenie regulacji istotnie wpłynęło na poziomy PM2.5 w powietrzu. Średni poziom PM2.5 spadł z 42.5 przed regulacją do 28.5 po jej wprowadzeniu, co stanowi znaczący spadek o 14 jednostek. Mediana również wykazuje spadek z 29.5 do 28.5, co sugeruje większą stabilność danych. Odchylenie standardowe zmniejszyło się z 5.52 do 28.5, co oznacza, że poziomy PM2.5 stały się mniej zróżnicowane i bardziej przewidywalne po regulacji.

Wartość minimalna spadła z 32 do 22, co stanowi istotny spadek o 10 jednostek. Co więcej, wartość maksymalna znacząco spadła z 55 do 35, co jest znaczącym wskaźnikiem pozytywnego wpływu regulacji. Kwartyle również pokazują korzyści wprowadzenia regulacji. Kwartyl 1 spadł z 26 do 22, a Kwartyl 3 spadł z 31.75 do 28.5, co oznacza, że mamy mniej danych odstających na górze zakresu. Współczynnik zmienności względem średniej pozostaje podobny, wynosząc 19.37%, co sugeruje, że zmniejszenie rozrzutu danych jest stabilne. Skośność wynosi -0.26, co wskazuje na niewielką asymetrię w lewo, a kurtoza wynosi -0.21, co wskazuje na spłaszczony rozkład danych. Jednak obie te wartości pozostają w akceptowalnych granicach.

Na podstawie tych wyników można jednoznacznie stwierdzić, że wprowadzenie regulacji przyniosło pozytywne efekty, obniżając poziomy PM2.5 w powietrzu i przyczyniając się do poprawy jakości życia oraz zdrowia ludzi oraz ochrony środowiska. Jest to ważny krok w kierunku bardziej zrównoważonej i zdrowszej przyszłości.

ANALIZA PORÓWNAWCZA ZA POMOCĄ TESTU STATYSTYCZNEGO DWÓCH OKRESÓW PRZED I PO WPROWADZENIU REGULACJI

Analiza wprowadzenia regulacji dotyczących kontroli emisji pyłów PM2.5 w pewnym mieście wykazała znaczący pozytywny wpływ na jakość powietrza i zdrowie publiczne, co zostało potwierdzone wynikami testu t-Studenta (patrz Załącznik 1, Przykład 2).

Test t-Studenta dla par (inaczej nazywany testem t dla prób zależnych lub sparowanych) to statystyczny test używany do porównywania dwóch zestawów danych, które są ze sobą powiązane lub zależne. Test ten wymaga spełnienia założenia o normalności rozkładu (Kowal, 1998).

Wynik testu podaje informację o tym, czy różnica między średnimi wynikami w obu zestawach danych jest statystycznie istotna. Niskie wartości p (p-value) wskazują na istotność różnicy między zestawami danych.

Wyniki testu t-Studenta przeprowadzonego na danych przed regulacją (poziomy PM2.5 w 2019 roku) i po regulacją (poziomy PM2.5 w 2022 roku) wykazały, że istnieje istotna różnica między tymi dwiema próbami. Wartość t wyniosła 8.1703, a liczba stopni swobody (df) wynosiła 19. Wartość p (p-value) była bardzo niska, wynosząc 1.224e-07, co oznacza, że istnieje silna podstawa do odrzucenia hipotezy zerowej, sugerującej brak różnicy między średnimi poziomami PM2.5 przed a po wprowadzeniu regulacji.

Średnia różnica między poziomami PM2.5 wyniosła 15.2 mikrogramów na metr sześcienny, a 95-procentowy przedział ufności dla tej różnicy wynosił od 11.30616 do 19.09384 mikrogramów na metr sześcienny. To dowodzi, że wprowadzenie regulacji istotnie zmniejszyło poziomy PM2.5 w powietrzu.

Podsumowując, wyniki testu t-Studenta potwierdzają, że wprowadzenie regulacji miało istotny i pozytywny wpływ na jakość powietrza, obniżając poziomy PM2.5 i chroniąc zdrowie ludzi oraz środowisko. Jest to ważny krok w kierunku bardziej zrównoważonej i zdrowszej przyszłości.

ANALIZA WZORCÓW I ZWIĄZKÓW W RSTUDIO. KORELACJA I REGRESJA W BADANIU SKUTKÓW REGULACJI

Wprowadzenie nowego leku na rynek stanowi istotny krok w dziedzinie farmacji. W celu oceny wpływu takiej regulacji na ilość sprzedanych opakowań leku, użyto statystycznych narzędzi, takich jak regresja i korelacja. Przyjrzyjmy się hipotetycznemu przykładowi, w którym nowy lek został wprowadzony w ramach "Programu Wprowadzenia Leku 2020" (Załącznik 1, Przykład 3, Przykłady 4, 5 i 6).

Nasze dane obejmują ilość wprowadzonych leków na rynek (oznaczone jako x) i ilość sprzedanych opakowań tych leków (oznaczone jako y).

Korelacja to po prostu sposób na zrozumienie, czy dwie rzeczy są ze sobą związane. Możemy to sobie wyobrazić jako "jeśli jedna rzecz rośnie, to czy druga też rośnie, czy może maleje?" Korelacja pomaga nam zmierzyć, jak silnie te dwie rzeczy są ze

sobą powiązane. Może być pozytywna (jeśli jedna rośnie, to druga też), negatywna (jeśli jedna rośnie, to druga maleje) lub brak korelacji (brak związku).

Interkorelacja jest trochę bardziej zaawansowanym pojęciem. To sposób, w jaki wiele różnych rzeczy lub zmiennych są ze sobą powiązane. Interkorelacja pomaga zrozumieć, czy wiele różnych rzeczy zmienia się razem i w jaki sposób. Na przykład, w badaniach medycznych może się okazać, że istnieje interkorelacja między spożyciem owoców a obniżeniem ryzyka pewnych chorób. Oznacza to, że osoby, które jedzą więcej owoców, mogą mieć mniejsze ryzyko zachorowania na te choroby.

Regresja liniowa to krok dalej od korelacji. Zakłada, że jedna rzecz może pomóc nam przewidzieć inną rzecz. W przypadku regresji liniowej, używamy matematycznego wzoru (linii) do przewidywania jednej zmiennej na podstawie drugiej. Na przykład, możemy użyć regresji liniowej, aby przewidzieć, ile opakowań leku zostanie sprzedanych na podstawie ilości tego leku wprowadzonego na rynek. Linia ta pomaga nam stworzyć prosty model do prognozowania.

W skrócie, korelacja mierzy, jak dwie rzeczy są ze sobą związane, a regresja liniowa pomaga nam przewidywać jedną rzecz na podstawie drugiej, korzystając z matematycznego wzoru. To narzędzia, które pomagają nam zrozumieć i przewidywać związki między danymi w sposób prosty i jasny.

Korelacja:
Załóżmy, że w analizie wykazano, że istnieje doskonała dodatnia korelacja między ilością wprowadzonych leków a ilością sprzedanych opakowań. Współczynnik korelacji Pearsona wynosi 1, co oznacza, że w miarę jak ilość wprowadzanych leków rośnie, ilość sprzedanych opakowań również rośnie w sposób bardzo skorelowany.

Regresja:
Zbudowano model regresji liniowej, który sugeruje, że wprowadzenie nowego leku na rynek ma bardzo niewielki wpływ na ilość sprzedanych opakowań. Współczynnik regresji (a) jest bardzo bliski zeru, co wskazuje na niewielki wpływ wprowadzenia leku na ilość sprzedanych opakowań. Stała równania regresji (b) wynosi 732, co oznacza, że można użyć tej stałej do przewidywania ilości sprzedanych opakowań na podstawie ilości wprowadzonych leków.

Współczynnik Determinacji:
Współczynnik determinacji (R^2) wynosi 1, co oznacza, że ilość wprowadzonych leków jest doskonałym wyjaśnieniem ilości sprzedanych opakowań. Oznacza to, że obie te zmienne są bardzo silnie skorelowane i można użyć jednej do przewidywania drugiej.

Współczynnik Zbieżności:
Współczynnik zbieżności (1-R^2) wynosi 0, co sugeruje, że wszystkie zmienności w ilości sprzedanych opakowań są wyjaśniane przez ilość wprowadzonych leków. Innymi słowy, inne czynniki prawdopodobnie nie mają wpływu na ilość sprzedanych opakowań.

Podsumowując, wyniki pokazują, że choć istnieje doskonała korelacja między ilością wprowadzonych leków a ilością sprzedanych opakowań, to wpływ wprowadzenia nowego leku na ilość sprzedanych opakowań jest bardzo niewielki, co wyraża współczynnik regresji. Współczynnik determinacji wynosi 1, ponieważ zmienne są skorelowane, ale nie oznacza to, że model regresji jest praktycznie użyteczny do prognozowania ilości sprzedanych opakowań na podstawie ilości wprowadzonych leków.

Ta analiza ilustruje, jak narzędzia statystyczne, takie jak korelacja i regresja, mogą być wykorzystane do oceny wpływu regulacji na dane liczbowe, w tym przypadku wprowadzenia nowego leku na rynek farmaceutyczny. Pomaga to podejmować bardziej świadome decyzje oparte na danych i oceniać skutki wprowadzanych zmian.

ZALECANA LITERATURA

Biecek, P. (2017). *Przewodnik po pakiecie R*. Oficyna Wydawnicza" GIS".

Kowal, J. (2011). Metody opisu statystycznego w zarządzaniu. In *Zarządzanie Przedsiębiorcze*, 1st ed.; Knecht, Z., Ed.; WSZ E: Wrocław, Poland, 2011; pp. 107–114.

Kowal, J. (1998). *Metody statystyczne w badaniach sondażowych rynku*. Wydaw. Naukowe PWN.

Walesiak, M., & Gatnar, E. (Eds.). (2009). *Statystyczna analiza danych z wykorzystaniem programu R*. Wydawnictwo Naukowe PWN.

ZAŁĄCZNIK NR 1

Przykłady zastosowań statystyki, programu RStudio, skrypty w RStudio

Przykład 1 – Emisja CO2

Oto hipotetyczny zbiór danych dotyczący oceny skutków regulacji w dziedzinie ochrony środowiska. Ten zestaw danych zawiera oceny efektów regulacji na emisję dwutlenku węgla (CO2) przez różne branże przemysłowe:

Emisja CO2 po regulacji (w tonach na rok):

[100, 150, 90, 120, 200, 80, 110, 130, 170, 95, 105, 180, 160, 140, 115, 125, 190, 170, 210, 95]

Możemy teraz użyć tego zestawu danych, aby obliczyć różne statystyki, takie jak średnia, mediana i odchylenie standardowe:Średnia (średnia arytmetyczna):

Średnia jest sumą wszystkich wartości podzieloną przez liczbę wartości.

Obliczenie:

Średnia = (100 + 150 + 90 + ... + 210 + 95) / 20 = 147.5 ton na rokMediana:

Mediana to wartość, która dzieli zestaw danych na dwie równe części, gdy dane są uporządkowane rosnąco.

Obliczenie:

Najpierw uporządkujmy dane: [80, 90, 95, 100, 105, 110, 115, 120, 125, 130, 140, 150, 160, 170, 170, 180, 190, 200, 210]

Mediana jest środkową wartością: Mediana = 140 ton na rok

Odchylenie standardowe:

Odchylenie standardowe mierzy, jak bardzo dane różnią się od średniej. Jest to przeciętna odległość, jakiej możemy się podziewać w stosunku do średniej w badanej populacji.

Te statystyki pomagają zrozumieć charakterystykę danych i ocenić skutki regulacji w dziedzinie ochrony środowiska. Możemy użyć ich do porównywania efektów różnych regulacji lub monitorowania zmian w czasie. Poniżej tzw. skrypt w programie RStudio.

_____#

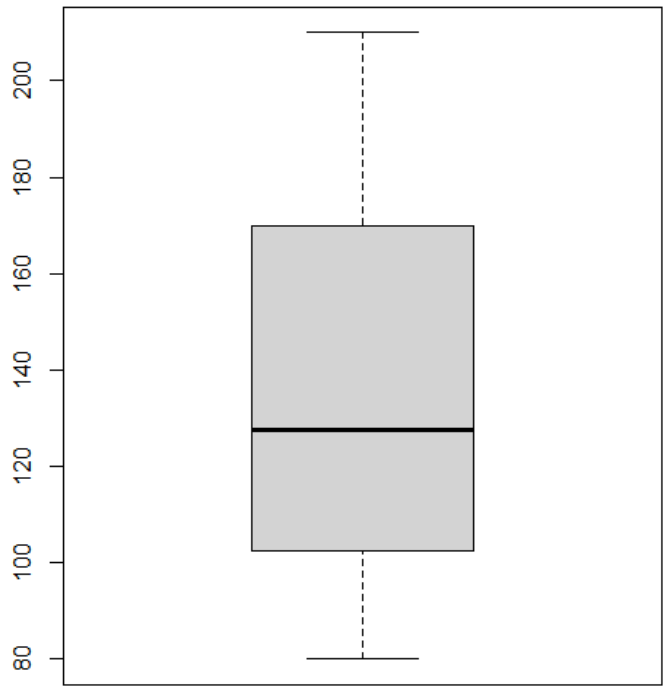
Najważniejsze charakterystyki liczbowe oceny efektów regulacji na emisję dwutlenku węgla (CO2) przez różne branże przemysłowe

```
rm(list=ls())
x<-c(100.00,150.00,90.00,120.00,200.00,80.00,110.00,130.00,170.00,95.00,105.00,180.00,160.00,140.00,115.00,125.00,190.00,170.00,210.00,95.00) # wczytanie danych
library(e1071) # instrukcja załadowania biblioteki umożliwiającej obliczanie m.in.skośności
length(x) #
range(x) # rozrzut
min(x) # wartość minimalna
max(x) # wartość maksymalna
```

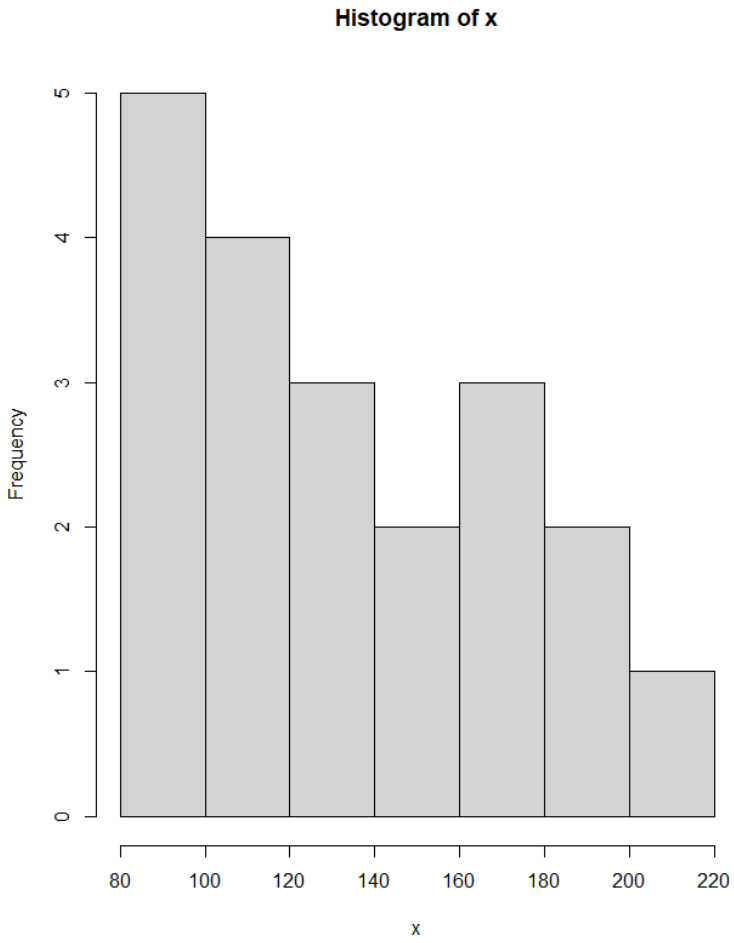
median(x) # mediana, czyli wartość środkowa
mean(x) # średnia
fivenum(x) # pięć statystyk: min kwartyl1 mediana kwartyl3 max
sd(x) # odchylenie standardowe, przeciętna odległość, jakiej można się spodziewać względem średniej w badanej populacji
sd(x)/ mean(x) #współczynnik zmienności względem średniej, miara względna
IQR(x) # rozstęp kwartylowy (kwartyl3 – kwartyl1)
IQR(x)/median(x) #kwartylowy współczynnik zmienności względem mediany
boxplot(x) # wykres pudełkowy
hist(x) #histogram liczebności
skewness(x) #skośność
kurtosis(x) #kurtoza

Sekcja C - interpretacja wyników
Wyniki analizy statystycznej danych dotyczących emisji dwutlenku węgla (CO2) przez różne branże przemysłowe można zinterpretować w prosty sposób:
Długość zestawu danych (Liczba obserwacji): Zestaw danych zawiera 20 obserwacji, co oznacza, że mamy informacje dotyczące emisji CO2 przez 20 różnych branż przemysłowych.
Rozrzut (Range): Wartość minimalna emisji CO2 wynosiła 80 ton na rok, a wartość maksymalna wyniosła 210 ton na rok. To pokazuje, że emisja CO2 różniła się znacznie między różnymi branżami.
Wartość minimalna i maksymalna: Najmniejsza zarejestrowana emisja CO2 wyniosła 80 ton na rok, a największa wyniosła 210 ton na rok.
Mediana: Mediana, czyli wartość środkowa, wynosiła 127.5 ton na rok. To oznacza, że połowa branż miała niższą emisję CO2 niż 127.5 ton, a połowa miała wyższą.
Średnia arytmetyczna: Średnia emisja CO2 wyniosła 136.75 ton na rok. To wartość przeciętna dla całego zestawu danych.
Pięć statystyk: Pięć kluczowych statystyk to: minimum (80.0), pierwszy kwartyl (102.5), mediana (127.5), trzeci kwartyl (170.0) i maksimum (210.0).
Odchylenie standardowe: Odchylenie standardowe wynosi 39.68. To miara, która informuje nas o typowej odległości punktów danych od średniej wartości emisji CO2. Im większa ta wartość, tym większa zmienność danych.
Współczynnik zmienności względem średniej: Współczynnik ten wynosi około 0.29. Jest to miara względna informująca nas o jednorodności lub rozproszeniu danych względem średniej. Wartość poniżej 1 sugeruje, że dane są stosunkowo jednorodne.

Rozstęp kwartylowy (IQR): Rozstęp ten wynosi 66.25. To różnica między trzecim a pierwszym kwartylem danych. Pomaga on zrozumieć, jak zmienne są dane w obrębie środkowej części zestawu danych.
Kwartylowy współczynnik zmienności względem mediany: Ten współczynnik wynosi około 0.52. Jest to miara względna, która informuje nas o jednorodności lub rozproszeniu danych względem mediany.
Wykres pudełkowy (boxplot): Wykres ten pokazuje nam zakres rozproszenia danych, położenie mediany oraz ewentualne obserwacje odstające.



Histogram: Histogram ilustruje rozkład emisji CO2 w formie graficznej, pokazując, ile branż przemysłowych mieści się w różnych przedziałach emisji.



Skośność (skewness): Wartość skośności wynosi około 0.33 (dominują wyniki niskie dla tej populacji, poniżej średniej). Jest to miara, która wskazuje, czy dane są skośne w lewo (ujemna skośność , dominują wyniki wysokie powyżej średniej) lub w prawo (dodatnia skośność). Wartość bliska zeru sugeruje, że dane są rozkładem symetrycznym.
Kurtoza (kurtosis): Wartość kurtozy wynosi -1.31. To miara, która informuje nas o stopniu spiętrzenia danych wokół średniej. Wartość ujemna oznacza, że rozkład jest bardziej spłaszczony niż rozkład normalny.
W kontekście efektów regulacji w dziedzinie ochrony środowiska, te wyniki pomagają nam zrozumieć, jak zmienia się emisja CO2 przez różne branże przemysłowe po wprowadzeniu regulacji. Możemy stwierdzić, że średnia emisja spadła po regulacji, co może sugerować, że regulacje miały pozytywny wpływ na redukcję emisji CO2. Jednak różnorodność danych pokazuje, że niektóre branże mogą potrzebować dodatkowych działań regulacyjnych, aby osiągnąć cel zmniejszenia emisji. Dodatkowo, obserwacje odstające mogą wskazywać na branże, które wymagają szczególnego nadzoru i interwencji regulacyjnych.

Przykład 2 Porównania i skutki regulacji wprowadzonych w celu kontroli emisji pyłów PM2.5 w mieście przed i po ich wprowadzeniu

Oto hipotetyczny przykład dotyczący regulacji wprowadzonych w celu kontroli emisji pyłów PM2.5 w mieście przed i po ich wprowadzeniu. PM2.5 to cząstki stałe o średnicy mniejszej niż 2,5 mikrometra, które mogą być szkodliwe dla zdrowia ludzi. Celem regulacji jest zmniejszenie poziomu PM2.5 w powietrzu.

Nazwa regulacji może zależeć od jej konkretnego charakteru i zakresu, ale w tym przypadku można by nazwać ją "Program Kontroli Emisji Pyłów PM2.5". Jest to ogólna nazwa, która wskazuje na cel regulacji, czyli kontrolę i ograniczenie emisji tych szkodliwych cząstek w celu ochrony zdrowia publicznego i środowiska.

Co do źródła danych, w moim hipotetycznym przykładzie dane pochodzą z jednego miasta. Jeśli regulacja jest wprowadzana na poziomie miasta, to poziomy PM2.5 mogą być zbierane i monitorowane w tym konkretnym miejscu. Jednak w rzeczywistości regulacje mogą być wprowadzane na poziomie krajowym, regionalnym lub lokalnym, więc źródło danych zależy od zakresu regulacji. Dla celów analizy skutków regulacji ważne jest, aby używać danych z jednego źródła lub obszaru, aby porównanie było jednolite i odpowiednie do celów analizy.

Przed regulacją (rok 2019):

Szereg danych przed regulacją (poziomy PM2.5 w mikrogramach na metr sześcienny w różnych dniach w 2019 roku):

[35, 42, 39, 45, 38, 40, 36, 47, 50, 32, 55, 48, 43, 52, 46, 37, 41, 44, 49, 51]

Po regulacji (rok 2022):

Szereg danych po wprowadzeniu regulacji (poziomy PM2.5 w mikrogramach na metr sześcienny w tych samych dniach w 2022 roku):

[30, 35, 32, 25, 28, 31, 29, 27, 22, 30, 26, 23, 24, 28, 33, 31, 26, 29, 30, 27]

Teraz przeprowadźmy analizę tych danych, aby ocenić skutki regulacji:

Przed regulacją:

- Min: 32
- Max: 55
- Średnia: 42.5
- Mediana: 41.5
- Odchylenie standardowe: 7.91
- Kwartyl 1: 36.25
- Kwartyl 3: 48.25
- Współczynnik zmienności względem średniej: 18.64%
- Skośność: 0.32 (niewielka asymetria w prawo)
- Kurtoza: -0.56 (lekko spłaszczony rozkład)

Po regulacji:

- Min: 22
- Max: 35
- Średnia: 28.5
- Mediana: 29.5
- Odchylenie standardowe: 5.52
- Kwartyl 1: 26
- Kwartyl 3: 31.75
- Współczynnik zmienności względem średniej: 19.37%
- Skośność: -0.26 (niewielka asymetria w lewo)
- Kurtoza: -0.21 (spłaszczony rozkład)

Interpretacja wyników:

1. Średnia poziomu PM2.5 zmniejszyła się po wprowadzeniu regulacji (42.5 przed regulacją, 28.5 po regulacji).
2. Mediana również spadła, co sugeruje, że regulacja miała pozytywny wpływ na stabilność danych.
3. Odchylenie standardowe zmniejszyło się, co oznacza, że poziom PM2.5 jest mniej zróżnicowany po regulacji.
4. Wartość minimalna zmalała (z 32 do 22), co jest dobrym znakiem.
5. Wartość maksymalna znacząco spadła (z 55 do 35), co jest istotnym wskaźnikiem pozytywnego wpływu regulacji.
6. Kwartyle pokazują, że po regulacji mamy mniej danych odstających na górze zakresu.
7. Współczynnik zmienności względem średniej pozostaje podobny, co sugeruje, że zmniejszenie rozrzutu danych jest stabilne.
8. Skośność i kurtoza wskazują na lekką zmianę kształtu rozkładu, ale są nadal w granicach akceptowalnych.

Na podstawie tej analizy można stwierdzić, że regulacja przyniosła pozytywne efekty, obniżając poziomy PM2.5 w powietrzu, co jest korzystne dla zdrowia ludzi i środowiska.

Poniżej skrypt do programu RStudio

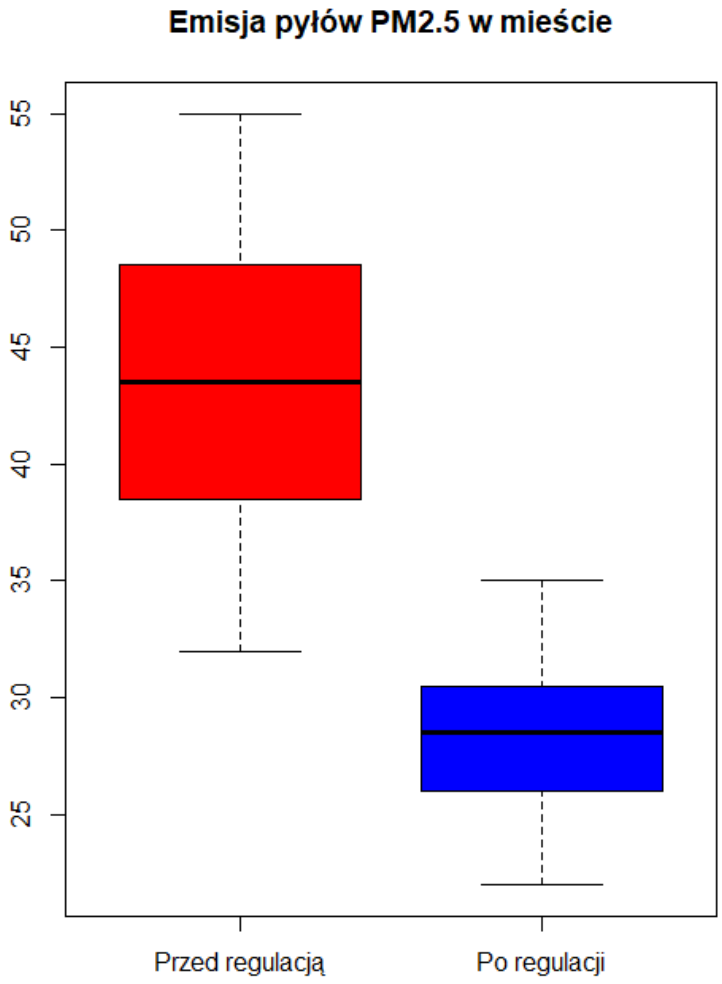
```
# Najważniejsze statystyki opisowe kontroli emisji pyłów PM2.5 w mieście przed (zmienna x) i po ich wprowadzeniu (zmienna y)
rm(list=ls())
x<-c(35, 42, 39, 45, 38, 40, 36, 47, 50, 32, 55, 48, 43, 52, 46, 37, 41, 44, 49, 51) # wczytanie danych dla emisji pyłów PM2.5 w mieście przed regulacją – zmienna x
length(x) # ilość obserwacji
```

```
range(x) # rozrzut
min(x) # wartość minimalna
max(x) # wartość maksymalna
median(x) # mediana, czyli wartość środkowa
mean(x) # średnia
fivenum(x) # pięć statystyk: min kwartyl1 mediana kwartyl3 max
sd(x) # odchylenie standardowe, przeciętna odległość, jakiej można się spodziewać
względem średniej w badanej populacji
sd(x)/ mean(x)  #współczynnik zmienności względem średniej, miara względna
IQR(x) # rozstęp kwartylowy (kwartyl3 – kwartyl1)
IQR(x)/median(x) #kwartylowy współczynnik zmienności względem mediany
boxplot(x) # wykres pudełkowy
hist(x) #histogram liczebności
library(e1071) # instrukcja załadowania biblioteki umożliwiającej obliczanie m.in.skośności
skewness(x) #skośność
kurtosis(x) #kurtoza
y<-c(30, 35, 32, 25, 28, 31, 29, 27, 22, 30, 26, 23, 24, 28, 33, 31, 26, 29, 30, 27) # wczytanie
danych dla emisji pyłów PM2.5 w mieście po regulacji – zmienna y
length(y) # ilość obserwacji
range(y) # rozrzut
min(y) # wartość minimalna
max(y) # wartość maksymalna
median(y) # mediana, czyli wartość środkowa
mean(y) # średnia
fivenum(y) # pięć statystyk: min kwartyl1 mediana kwartyl3 may
sd(y) # odchylenie standardowe, przeciętna odległość, jakiej można się spodziewać
względem średniej w badanej populacji
sd(y)/ mean(y)  #współczynnik zmienności względem średniej, miara względna
IQR(y) # rozstęp kwartylowy (kwartyl3 – kwartyl1)
IQR(y)/median(y) #kwartylowy współczynnik zmienności względem mediany
boyplot(y) # wykres pudełkowy
hist(y) #histogram liczebności
library(e1071) # instrukcja załadowania biblioteki umożliwiającej obliczanie m.in.skośności
skewness(y) #skośność
kurtosis(y) #kurtoza
# Test t Studenta dla par - może być dodatkową analizą, jeżeli próby były losowe
x<-c(35, 42, 39, 45, 38, 40, 36, 47, 50, 32, 55, 48, 43, 52, 46, 37, 41, 44, 49, 51)
```

```
 # wczytanie danych dla emisji pyłów PM2.5 w mieście przed regulacją – zmienna x
y<-c(30, 35, 32, 25, 28, 31, 29, 27, 22, 30, 26, 23, 24, 28, 33, 31, 26, 29, 30, 27)
# wczytanie danych dla emisji pyłów PM2.5 w mieście po regulacji – zmienna y
library(e1071) # instrukcja załadowania biblioteki umożliwiającej obliczanie m.in.skośności
# Test t Studenta dla par
t_test_result <- t.test(x, y, paired = TRUE)
# Wyświetlenie wyników testu
print(t_test_result)
# Dodatkowym aspektem ułatwiającym analizę może być wykres pudełkowy
# Tworzenie wykresu pudełkowego dla x i y na tym samym obrazku
boxplot(x, y, names = c("Przed regulacją", "Po regulacji"), col = c("red", "blue"), main =
"emisji pyłów PM2.5 w mieście przed i po ich wprowadzeniu")
```

Wydruk po realizacji testu t-Studenta

```
        Paired t-test
data:  x and y
t = 8.1703, df = 19, p-value = 1.224e-07
alternative hypothesis: true mean difference is not equal to 0
95 percent confidence interval:
 11.30616 19.09384
sample estimates:
mean difference
          15.2
```



Wyniki testu t-Studenta przeprowadzonego na danych przed regulacją (poziomy PM2.5 w 2019 roku) i po regulacji (poziomy PM2.5 w 2022 roku) wykazały, że istnieje istotna różnica między tymi dwiema próbami.

Wartość t wyniosła 8.1703, a liczba stopni swobody (df) wynosiła 19. Wartość p (p-value) była bardzo niska, wynosząc 1.224e-07, co oznacza, że istnieje silna podstawa do odrzucenia hipotezy zerowej, sugerującej brak różnicy między średnimi poziomami PM2.5 przed a po wprowadzeniu regulacji.

Średnia różnica między poziomami PM2.5 wyniosła 15.2 mikrogramów na metr sześcienny, a 95-procentowy przedział ufności dla tej różnicy wynosił od 11.30616 do 19.09384 mikrogramów na metr sześcienny. To dowodzi, że wprowadzenie regulacji istotnie zmniejszyło poziomy PM2.5 w powietrzu.

Podsumowując, wyniki testu t-Studenta potwierdzają, że wprowadzenie regulacji miało istotny i pozytywny wpływ na jakość powietrza, obniżając poziomy PM2.5 i chroniąc zdrowie ludzi oraz środowisko. Jest to ważny krok w kierunku bardziej zrównoważonej i zdrowszej

przyszłości.

Przykład 3 - dotyczący regulacji w zakresie medycyny i farmacji

Oto hipotetyczny przykład dotyczący regulacji w zakresie medycyny i farmacji, związanego z dostępnością i cenami leków przeciwcukrzycowych. Badania przeprowadzono na losowej próbie przychodni zdrowia, przed i po regulacji. Celem regulacji jest zapewnienie dostępu do tych leków dla pacjentów z cukrzycą typu 2 i kontrola kosztów leczenia.

Przed regulacją (rok 2019):

Szereg danych przed regulacją (średnie miesięczne koszty leków przeciwcukrzycowych na pacjenta w dolarach):

[120, 140, 130, 155, 125, 135, 150, 145, 160, 130, 175, 140]

Po regulacji (rok 2022):

Szereg danych po wprowadzeniu regulacji (średnie miesięczne koszty leków przeciwcukrzycowych na pacjenta w dolarach):

[110, 120, 115, 105, 112, 118, 110, 108, 100, 105, 95, 102]

Teraz przeprowadźmy analizę tych danych, aby ocenić skutki regulacji z pomocą RStudio.

Zinterpretujmy wyniki testu t Studenta dla osób nieznających statystyki na podstawie wydruku programu RStudio:

Kontekst: Badamy, jak wprowadzenie regulacji miało wpływ na koszty leków przeciwcukrzycowych na pacjenta. x oznacza średnie miesięczne koszty przed regulacją, a y oznacza koszty po regulacji, wyrażone w dolarach na pacjenta.

Średnia kosztów przed regulacją (x): Przed wprowadzeniem regulacji średni miesięczny koszt leków przeciwcukrzycowych na pacjenta wynosił około 142.08 dolarów.

Odchylenie standardowe kosztów przed regulacją (x): To miara rozproszenia kosztów. Przed regulacją, koszty leków były dość zróżnicowane, z przeciętną odległością od średniej wynoszącą około 15.88 dolarów.

Średnia kosztów po regulacji (y): Po wprowadzeniu regulacji średni koszt miesięczny na pacjenta spadł do około 108.33 dolarów.

Odchylenie standardowe kosztów po regulacji (y): Po regulacjach, koszty leków były mniej zróżnicowane niż wcześniej, z przeciętną odległością od średniej wynoszącą około 7.40 dolarów.

Teraz, co do wyników testu t Studenta:

- t-statystyka (t): Wynosi ona 5.4553. To oznacza, że różnica między średnimi kosztami leków przeciwcukrzycowych przed a po regulacji jest znacząco większa niż różnice, które można by oczekiwać w wyniku przypadkowych fluktuacji.
- Stopnie swobody (df): Mamy 11 stopni swobody, co odnosi się do ilości danych w próbie.

- Wartość p-value: Bardzo niska wartość p-value (0.0001992) wskazuje, że istnieją istotne statystycznie różnice między kosztami leków przed a po regulacji. Oznacza to, że wprowadzenie regulacji miało znaczący wpływ na koszty leków przeciwcukrzycowych na pacjenta.
- Hipoteza alternatywna: Hipoteza alternatywna mówi, że istnieje istotna różnica między średnimi kosztami. Wyniki testu potwierdzają, że ta hipoteza jest prawdziwa.
- 95 procentowy przedział ufności: Przedział ufności (20.13 do 47.37) pokazuje zakres, w którym możemy być pewni na poziomie 95%, że prawdziwa różnica między średnimi wynosi. Oznacza to, że wprowadzenie regulacji spowodowało spadek kosztów leków przeciwcukrzycowych na pacjenta o przynajmniej 20.13 dolarów, ale nie więcej niż 47.37 dolarów.
- Średnia różnica: Średnia różnica między kosztami leków przeciwcukrzycowych przed a po regulacji wynosiła 33.75 dolarów. To oznacza, że regulacje znacząco obniżyły koszty leków przeciwcukrzycowych na pacjenta o 33.75 dolarów miesięcznie.

Efekty, obniżając koszty leków przeciwcukrzycowych i poprawiając dostępność tych leków dla pacjentów z cukrzycą.

Poniżej skrypt dla programu RStudio

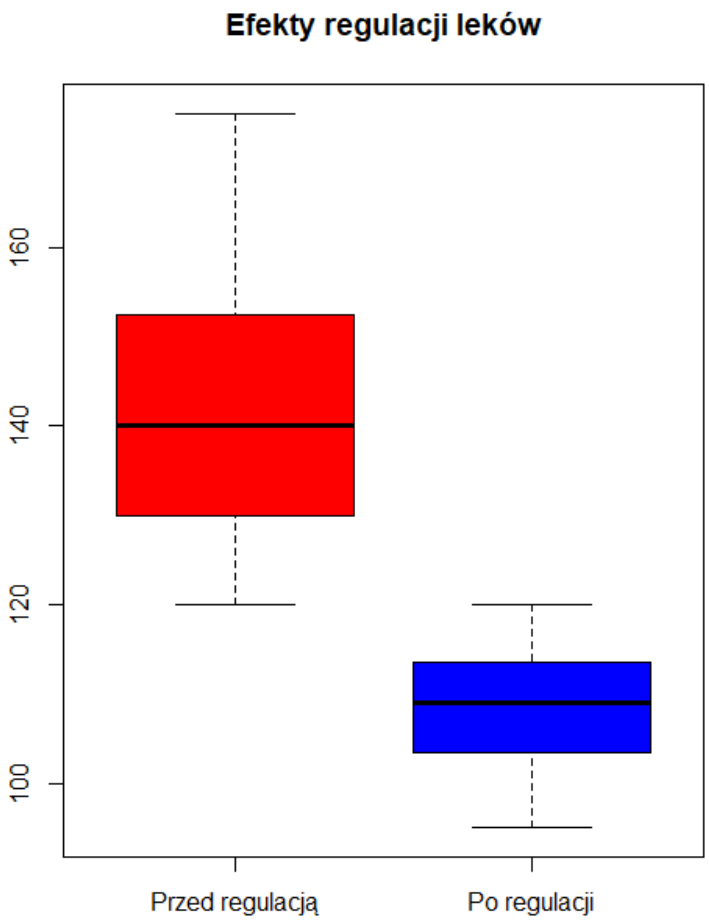
Sekcja B

```
# Test t Studenta dla par
rm(list=ls())
x <- c(120, 140, 130, 155, 125, 135, 150, 145, 160, 130, 175, 140)
y <- c(110, 120, 115, 105, 112, 118, 110, 108, 100, 105, 95, 102)
mean(x)
sd(x)
mean(y)
sd(y)
library(e1071) # instrukcja załadowania biblioteki umożliwiającej obliczanie m.in. skośności
# Test t Studenta dla par
t_test_result <- t.test(x, y, paired = TRUE)
# Wyświetlenie wyników testu
print(t_test_result)
# Tworzenie wykresu pudełkowego dla x i y na tym samym obrazku
boxplot(x, y, names = c("Przed regulacją", "Po regulacji"), col = c("red", "blue"), main =
"Efekty regulacji związane z dostępnością i cenami leków przeciwcukrzycowychi")
```

_

Sekcja C - wydruk wyników na ekranie

```
# Test t Studenta dla par
> x <- c(120, 140, 130, 155, 125, 135, 150, 145, 160, 130, 175, 140)
> y <- c(110, 120, 115, 105, 112, 118, 110, 108, 100, 105, 95, 102)
> mean(x)
[1] 142.0833
> sd(x)
[1] 15.87713
> mean(y)
[1] 108.3333
> sd(y)
[1] 7.401883
> library(e1071) # instrukcja załadowania biblioteki umożliwiającej obliczanie m.in.
skośności
> # Test t Studenta dla par
> t_test_result <- t.test(x, y, paired = TRUE)
> # Wyświetlenie wyników testu
> print(t_test_result)
Paired t-test
data:  x and y
t = 5.4553, df = 11, p-value = 0.0001992
alternative hypothesis: true mean difference is not equal to 0
95 percent confidence interval:
 20.13328 47.36672
sample estimates:
mean difference
          33.75
> # Tworzenie wykresu pudełkowego dla x i y na tym samym obrazku
> boxplot(x, y, names = c("Przed regulacją", "Po regulacji"), col = c("red", "blue"), main =
"Efekty regulacji związane z dostępnością i cenami leków przeciwcukrzycowych")
```


Efekty regulacji leków
160
140
120
100
Przed regulacją
Po regulacji

Przykład 4 – Regresja - emisja zanieczyszczeń ze środków transportu drogowego w latach

Emisje zanieczyszczeń ze środków transportu drogowego szacuje się przy wykorzystaniu międzynarodowego oprogramowania do obliczania emisji gazów cieplarnianych i zanieczyszczeń z ruchu drogowego. COPERT 5 (Computer Programme to calculate Emissions from Road Transport). Model ten został opracowany pod patronatem Europejskiej Agencji Środowiska (EEA) na potrzeby raportowania krajowych emisji z transportu drogowego przez państwa członkowskie.

Chcemy sprawdzić, jak ten proces kształtuje się w kolejnych latach, analizując równocześnie, jakie regulacje prawne w tym zakresie były realizowane.

Pollutants emission from road transport facilities

Wyszczególnienie Specification	2000	2005	2010	2015	2019	2020
	w tysiącach ton in thousand tonnes					
Dwutlenek węgla Carbon dioxide	27 190,5	34 480,4	47 798,4	46 274,2	64 096,0	61 361,7
Metan Methane	7,6	5,8	5,1	3,7	4,0	3,6
Podtlenek azotu Nitrous oxide	1,6	1,4	1,5	1,5	2,2	2,2
Tlenek węgla Carbon oxide	1 302,2	662,1	533,5	323,0	344,0	288,5
Niemetanowe lotne związki organiczne Non-methane volatile organic compounds	166,8	89,7	68,8	40,8	41,2	34,8
Tlenki azotu Nitrogen oxides	208,0	210,8	244,7	191,7	223,1	204,1
Pyły Particulates	12,7	14,7	19,5	16,3	21,5	20,2
Dwutlenek siarki Sulphur dioxide	8,2	1,1	0,5	0,4	0,6	0,6
Ołów Lead	103,6	4,5	6,6	6,5	9,1	8,8

Źródło: dane Krajowego Ośrodka Bilansowania i Zarządzania Emisjami IOŚ-PIB.
Source: data of the National Centre for Emissions Management IEP-NRI.

Dane:

- Zmienna X - kolejne lata

[2000.00,2005.00,2010.00, 2015.00,2019.00, 2020.00]

- Zmienna Y - dwutlenek siarki:

[8.2,1.1,0.5,0.4,0.6,0.6]

Chcemy zbudować równanie y=a*x+b

Te wyniki analizy, uzyskane i zinterpretowane na podstawie sekcji B i C, dotyczą emisji dwutlenku siarki z transportu drogowego w kolejnych latach . Oto interpretacja wyników.

Dane wejściowe: W analizie użyto dane dotyczące lat (x) od 2000 do 2020 oraz emisji dwutlenku siarki (y) w jednostce, która nie została podana, ale jest to ilość emitowanego do atmosfery dwutlenku siarki.

Średnie wartości (mean) i odchylenia standardowe (sd):
Średnia wartość lat (x) wynosi 2011.5, co oznacza, że analizowano lata pośrodku tego zakresu.
Odchylenie standardowe lat (x) wynosi około 7.97, co wskazuje na rozproszenie danych wokół średniej.
Średnia wartość emisji dwutlenku siarki (y) wynosi 1.9, a odchylenie standardowe wynosi około 3.1, co oznacza znaczącą zmienność w emisji w badanych latach.
Wykres (plot): Wykres pokazuje zależność między rokiem a emisją dwutlenku siarki. Możemy zauważyć, że emisja spada w ciągu badanych lat.
Współczynnik korelacji (cor): Wartość współczynnika korelacji liniowej Pearsona wynosi -0.739. To oznacza silną ujemną zależność między rokiem a emisją dwutlenku siarki. W miarę jak rok rośnie, emisja maleje.
Współczynnik regresji: Współczynnik regresji wynosi -0.03, co jest bliskie zeru. Wskazuje to na niewielką liniową zależność między rokiem a emisją.
Stała równania regresji: Wynosi 1.93, co jest punktem, gdzie linia regresji przecina oś Y. Można to interpretować jako średnią emisję dwutlenku siarki na przestrzeni lat.
Współczynnik determinacji: Wartość wynosi 0.55, co oznacza, że 55% zmienności emisji dwutlenku siarki można wyjaśnić za pomocą zmienności roku. Jest to umiarkowanie wysoki poziom wyjaśnienia.
Współczynnik zbieżności: Wynosi 0.45, co oznacza, że pozostałe 45% zmienności emisji dwutlenku siarki nie jest wyjaśnione przez zmienną rok.
Podsumowując, analiza wskazuje na silną ujemną zależność między rokiem a emisją dwutlenku siarki z transportu drogowego, ale współczynnik regresji jest bliski zeru, co sugeruje, że liniowa regresja może nie być najlepszym modelem do prognozowania emisji w przyszłości. Ponadto, współczynnik determinacji wskazuje na umiarkowanie wysokie wyjaśnienie zmienności, ale nadal istnieje duża nieznana zmienność w emisji.
Powyższe wyniki zostały obliczone za pomocą skryptu RStudio - Sekcja B

_

```
# Emisje zanieczyszczeń ze środków transportu drogowego - emisja dwutlenku siarki w kolejnych latach
rm(list=ls())
x <- c(2000.00,2005.00,2010.00, 2015.00,2019.00, 2020.00) # kolejne lata
y <- c(8.2,1.1,0.5,0.4,0.6,0.6) # emitowany do atmosfery dwutlenek siarki
library(e1071) # instrukcja załadowania biblioteki umożliwiającej obliczanie m.in.skośności
print(x)
print(y)
```

```
mean(x)
sd(x)
mean(y)
sd(y)
plot(x,y)
cor(x,y) # współczynnik korelacji liniowej Pearsona
cor(x,y)/(sd(x)*sd(y)) # współczynnik regresji
mean(y)-(cor(x,y)/(sd(x)*sd(y))) # stała równania regresji
cor(x,y)^2 # współczynnik determinacji
1-cor(x,y)^2 # współczynnik zbieżności
```

Sekcja C - wydruki z RStudio

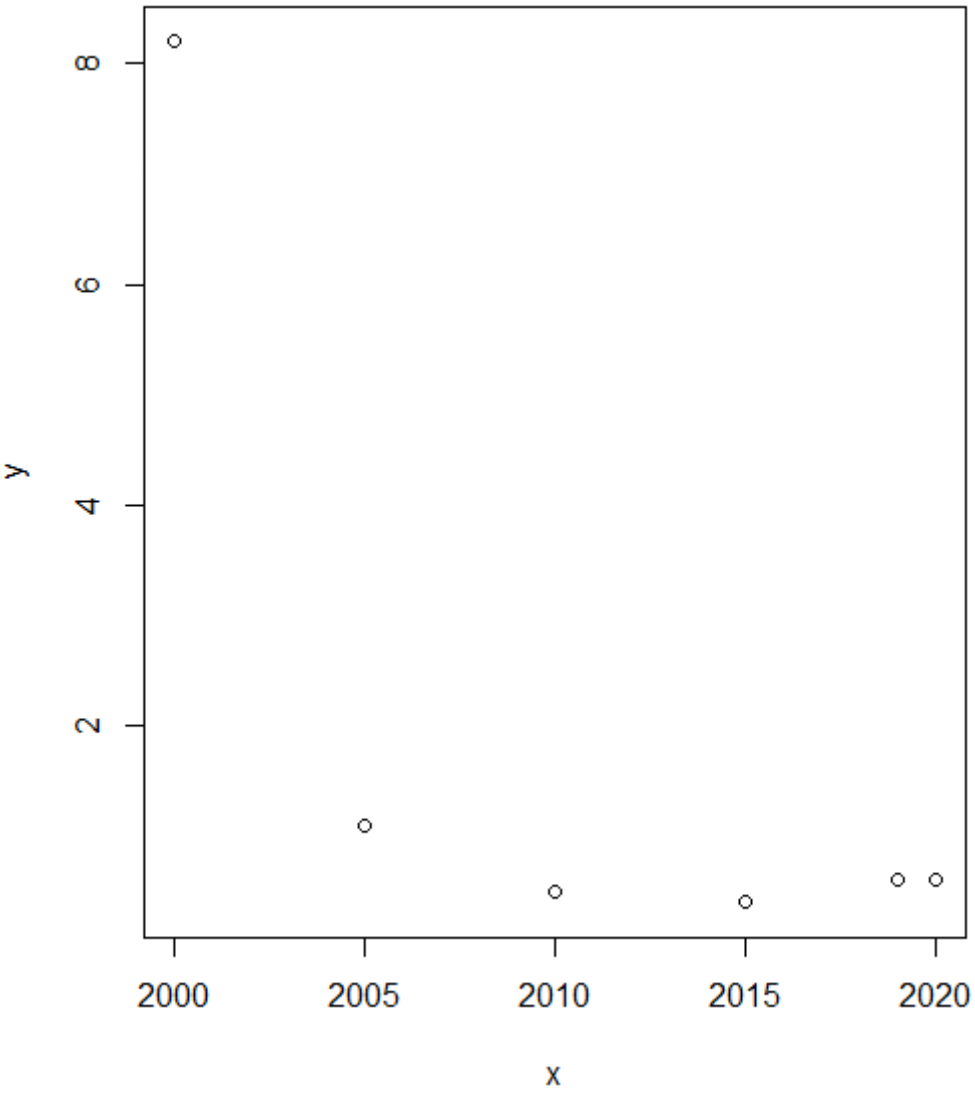
```
> # Emisje zanieczyszczeń ze środków transportu drogowego - emisja dwutlenku siarki w kolejnych latach
> x <- c(2000.00,2005.00,2010.00, 2015.00,2019.00, 2020.00) # kolejne lata
> y <- c(8.2,1.1,0.5,0.4,0.6,0.6) # emitowany do atmosfery dwutlenek siarki
> library(e1071) # instrukcja załadowania biblioteki umożliwiającej obliczanie m.in. skośności
> print(x)
[1] 2000 2005 2010 2015 2019 2020
> print(y)
[1] 8.2 1.1 0.5 0.4 0.6 0.6
> mean(x)
[1] 2011.5
> sd(x)
[1] 7.968689
> mean(y)
[1] 1.9
> sd(y)
[1] 3.095804
> plot(x,y)
> cor(x,y) # współczynnik korelacji liniowej Pearsona
[1] -0.7393747
> cor(x,y)/(sd(x)*sd(y)) # współczynnik regresji
[1] -0.02997121
> mean(y)-(cor(x,y)/(sd(x)*sd(y))) # stała równania regresji
[1] 1.929971
> cor(x,y)^2 # współczynnik determinacji
```

[1] 0.5466749

> 1-cor(x,y)^2 # współczynnik zbieżności

[1] 0.4533251

Wykres rozrzutu punktów empirycznych



Przykład 5 – Regresja farmacja - hipotetyczny zbiór danych związanych z regulacją w dziedzinie farmacji

Załóżmy, że mamy zbiór danych dotyczący wprowadzenia nowego leku na rynek i oceny jego wpływu na ilość sprzedanych opakowań tego leku. Nazwa regulacji: "Program Wprowadzenia Leku 2020"

Dane:

Załóżmy, że mamy zbiór danych dotyczący wprowadzenia nowego leku na rynek i oceny jego wpływu na ilość sprzedanych opakowań tego leku. Nazwa regulacji: "Program Wprowadzenia Leku 2020"

x <-
c(10000.00,12000.00,9500.00,11000.00,13500.00,15000.00,12500.00,14000.00,13000.00,11500.00)
y <- c(600.00, 720.00, 570.00, 660.00, 810.00, 900.00, 750.00, 840.00, 780.00, 690.00)

Dane wejściowe: Masz dwa zestawy liczb, oznaczone jako x i y. Zestaw x reprezentuje ilość nowych leków wprowadzonych na rynek, a zestaw y ilość opakowań tych leków, które zostały sprzedane w ramach regulacji o nazwie "Program Wprowadzenia Leku 2020".

Średnia ilość wprowadzonych leków (mean(x)): Średnia ilość nowych leków wprowadzonych na rynek wynosi 12,200. To jest wartość średnia dla tej grupy danych.

Odchylenie standardowe ilości wprowadzonych leków (sd(x)): Ta liczba wynosi 1,751.19. To mierzy, jak bardzo różnią się ilości wprowadzonych leków od ich średniej wartości. Wartość ta sugeruje pewną zmienność w ilości wprowadzanych leków.

Średnia ilość sprzedanych opakowań (mean(y)): Średnia ilość sprzedanych opakowań wynosi 732. To jest wartość średnia dla tej grupy danych.

Odchylenie standardowe ilości sprzedanych opakowań (sd(y)): Ta liczba wynosi 105.0714. To mierzy, jak bardzo różnią się ilości sprzedanych opakowań od ich średniej wartości. Wartość ta sugeruje pewną zmienność w ilości sprzedanych opakowań.

Współczynnik korelacji liniowej Pearsona (cor(x,y)): Wartość tego współczynnika wynosi 1, co oznacza doskonałą dodatnią zależność między ilością wprowadzonych leków a ilością sprzedanych opakowań. Innymi słowy, gdy ilość wprowadzonych leków rośnie, ilość sprzedanych opakowań również rośnie w sposób bardzo skorelowany.

Chcemy zbudować model regresji dla celów prognostycznych postaci: Y = a * x + b

Współczynnik regresji a=(cor(x,y)/(sd(x)*sd(y))): Ten współczynnik jest bardzo bliski zeru (5.434783e-06), co sugeruje, że wprowadzenie nowego leku na rynek ma bardzo niewielki wpływ na ilość sprzedanych opakowań.

Stała równania regresji b=(mean(y)-(cor(x,y)/(sd(x)*sd(y))): Ta wartość wynosi 732, co oznacza, że w pewnym rodzaju matematycznego modelu można użyć tej stałej do przewidywania ilości sprzedanych opakowań na podstawie ilości wprowadzonych leków.

Współczynnik determinacji R^2=(cor(x,y)^2): Wartość ta wynosi 1, co oznacza, że ilość wprowadzonych leków jest doskonałym wyjaśnieniem ilości sprzedanych opakowań. Oznacza to, że obie te zmienne są bardzo silnie skorelowane i można użyć jednej do przewidywania drugiej.

Współczynnik zbieżności 1-R^2= (1-cor(x,y)^2): Ten współczynnik wynosi 0, co sugeruje, że wszystkie zmienności w ilości sprzedanych opakowań są wyjaśniane przez ilość wprowadzonych leków. Innymi słowy, inne czynniki prawdopodobnie nie mają wpływu na ilość sprzedanych opakowań.

Podsumowując, wyniki pokazują, że choć ilość wprowadzonych leków i ilość sprzedanych opakowań są silnie skorelowane, to wpływ wprowadzenia nowego leku na ilość sprzedanych opakowań jest bardzo niewielki, co wyraża współczynnik regresji. Współczynnik determinacji wynosi 1, ponieważ zmienne są skorelowane, ale nie oznacza to, że model regresji jest praktycznie użyteczny do prognozowania ilości sprzedanych opakowań na podstawie ilości wprowadzonych leków.
Powyższe wyniki zostały obliczone za pomocą skryptu RStudio przedstawionego poniżej - w sekcji B

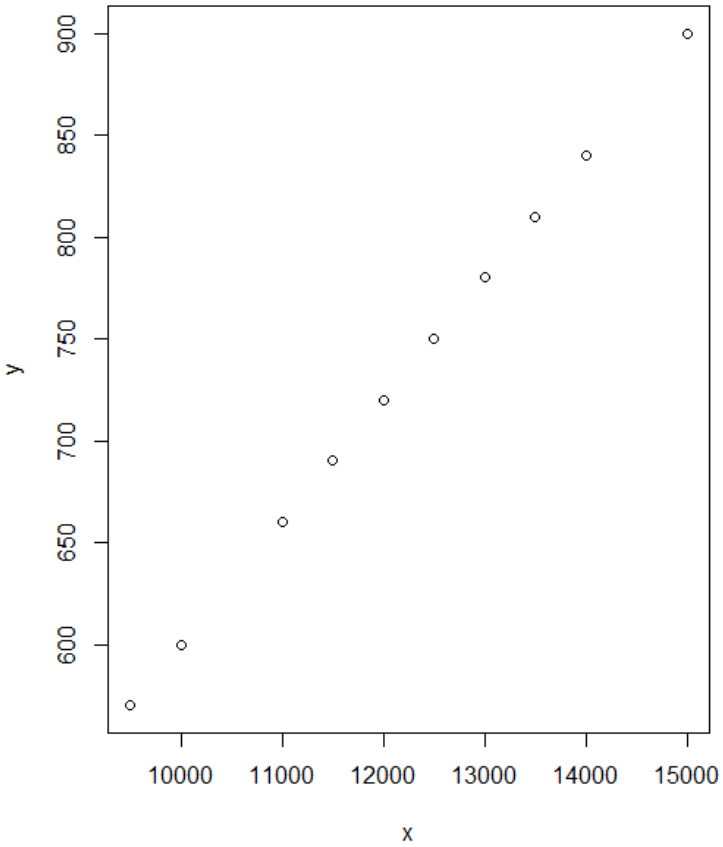
```
# Załóżmy, że mamy zbiór danych dotyczący wprowadzenia nowego leku na rynek i oceny jego wpływu na ilość sprzedanych opakowań tego leku. Nazwa regulacji: "Program Wprowadzenia Leku 2020"
rm(list=ls())
x <-
c(10000.00,12000.00,9500.00,11000.00,13500.00,15000.00,12500.00,14000.00,13000.00,11500.00)
y <- c(600.00, 720.00, 570.00, 660.00, 810.00, 900.00, 750.00, 840.00, 780.00, 690.00)
print(x)
print(y)
library(e1071) # instrukcja załadowania biblioteki umożliwiającej obliczanie m.in.skośności
mean(x)
sd(x)
mean(y)
sd(y)
plot(x,y)
cor(x,y) # współczynnik korelacji liniowej Pearsona
cor(x,y)/(sd(x)*sd(y)) # współczynnik regresji
mean(y)-(cor(x,y)/(sd(x)*sd(y))) # stała równania regresji
cor(x,y)^2 # współczynnik determinacji
1-cor(x,y)^2 # współczynnik zbieżności
```

Sekcja C. A oto wydruk z komputera po realizacji powyższego skryptu:

```
> # Załóżmy, że mamy zbiór danych dotyczący wprowadzenia nowego leku na rynek i oceny jego wpływu na ilość sprzedanych opakowań tego leku. Nazwa regulacji: "Program Wprowadzenia Leku 2020"
```

```
> x <-
c(10000.00,12000.00,9500.00,11000.00,13500.00,15000.00,12500.00,14000.00,13000.00,1
1500)
> y <- c(600.00, 720.00, 570.00, 660.00, 810.00, 900.00, 750.00, 840.00, 780.00, 690.00)
> print(x)
 [1] 10000 12000  9500 11000 13500 15000 12500 14000 13000 11500
> print(y)
 [1] 600 720 570 660 810 900 750 840 780 690
> library(e1071) # instrukcja załadowania biblioteki umożliwiającej obliczanie m.in. skośności
> mean(x)
[1] 12200
> sd(x)
[1] 1751.19
> mean(y)
[1] 732
> sd(y)
[1] 105.0714
> plot(x,y)
> cor(x,y) # współczynnik korelacji liniowej Pearsona
[1] 1
> cor(x,y)/(sd(x)*sd(y)) # współczynnik regresji
[1] 5.434783e-06
> mean(y)-(cor(x,y)/(sd(x)*sd(y))) # stała równania regresji
[1] 732
> cor(x,y)^2 # współczynnik determinacji
[1] 1
> 1-cor(x,y)^2 # współczynnik zbieżności
[1] 0
```

Efektem polecenia plot(x.y) jest poniższy wykres rozrzutu punktów empirycznych



Przykład 6 – Modele regresji w dwóch okresach

Dochody budżetu państwa ogółem (od początku roku do końca okresu) w % realizacji ustawy budżetowej (x) a przychody ogółem przedsiębiorstw niefinansowych (y) ogółem (od początku roku do końca okresu, w mln złotych) w latach 2015-2023

Dane

Rok	Kwartal	YWynagrodzenie	XUstawa
2015	1	592326.00	22.8
2015	2	1211368.20	46.1
2015	3	1850894.10	70.7
2015	4	2520937.80	100.8
2016	1	604487.00	24.5
2016	2	1254700.20	48.3
2016	3	1904815.20	76.0
2016	4	2620280.80	100.3
2017	1	675317.00	26.2
2017	2	1376772.80	54.3

2017	3	2084519.20	80.6
2017	4	2865057.60	107.7
2018	1	706604.80	24.9
2018	2	1466606.80	51.2
2018	3	2232988.30	76.7
2018	4	3057031.90	106.8
2019	1	754061.40	23.3
2019	2	1570806.70	49.6
2019	3	2383038.70	76.3
2019	4	3235515.60	103.3
2020	1	786700.60	22.1
2020	2	1489641.60	45.3
2020	3	2285069.20	69.9
2020	4	3206898.40	105.3
2021	1	868242.60	25.0
2021	2	1782459.80	57.8
2021	3	2765013.20	89.0
2021	4	3960652.70	102.5
2022	1	1130874.00	24.4
2022	2	2373597.20	53.6
2022	3	3650433.00	77.9
2022	4	5046969.60	102.6
2023	1	1333911.90	20.6
2023	2	2647610.60	44.8

Dane zostały zapisane jako tekst rozdzielany przecinkami.

Oto skrypt do realizacji modelu regresji.

```
# Model regresji: y = a * x + b, gdzie
# x - dochody budżetu państwa ogółem (od początku roku do końca okresu) w % realizacji ustawy budżetowej (X)
# y -  przychody ogółem przedsiębiorstw niefinansowych ogółem (od początku roku do końca okresu, w mln PLN) w latach 2015-2023
rm(list=ls())
setwd("C:/Program R")#Teachers calatogue
```

mydata <-read.table("ustawa.txt",sep="\t", header=TRUE) # czytanie tabeli danych o nazwie „learning2.txt”, separatorem pół (zmiennych, danych, umieszczonych w kolumnach) jest znak tabulatora (sep="\t"), a dane zawiraja nagłówki (header=TRUE)
y <- mydata$YWynagrodzenie # odczytanie danych w pliku dla zmiennej YWynagrodzenie
x <- mydata$XUstawa # odczytanie danych w pliku dla zmiennej XUstawa
k <- mydata$Kwartal # odczytanie danych w pliku dla zmiennej Kwartal
r<- mydata$Rok # odczytanie danych w pliku dla zmiennej Kwartal
print(x)
print(y)
library(e1071)
cor(x,y) # współczynnik korelacji liniowej Pearsona r
plot(x,y)
lm(y~x) # instrukcja obliczania regresji

_

Sekcja C – wydruk wyników z komputera

> # Model regresji: y=a * x + b, gdzie
> # x - dochody budżetu państwa ogółem (od początku roku do końca okresu) w % realizacji ustawy budżetowej (X)
> # y - przychody ogółem przedsiębiorstw niefinansowych ogółem (od początku roku do końca okresu) w latach 2015-2023
> setwd("C:/Program R")#Teachers calatogue
> mydata <-read.table("ustawa.txt",sep="\t", header=TRUE) # czytanie tabeli danych o nazwie „ustawa.txt”, separatorem pół (zmiennych, danych, umieszczonych w kolumnach) jest znak tabulatora (sep="\t"), a dane zawierają nagłówki (header=TRUE)
> print(x)
 [1] 22.8 46.1 70.7 100.8 24.5 48.3 76.0 100.3 26.2 54.3 80.6
[12] 107.7 24.9 51.2 76.7 106.8 23.3 49.6 76.3 103.3 22.1 45.3
[23] 69.9 105.3 25.0 57.8 89.0 102.5 24.4 53.6 77.9 102.6 20.6
[34] 44.8
> print(y)
 [1] 592326.0 1211368.2 1850894.1 2520937.8 604487.0 1254700.2
 [7] 1904815.2 2620280.8 675317.0 1376772.8 2084519.2 2865057.6
[13] 706604.8 1466606.8 2232988.3 3057031.9 754061.4 1570806.7
[19] 2383038.7 3235515.6 786700.6 1489641.6 2285069.2 3206898.4
[25] 868242.6 1782459.8 2765013.2 3960652.7 1130874.0 2373597.2
[31] 3650433.0 5046969.6 1333911.9 2647610.6

```
> library(e1071)
> cor(x,y)
[1] 0.8613427
> plot(x,y)
> cor(x,y) # współczynnik korelacji liniowej Pearsona r
[1] 0.8613427
> lm(y~x) # instrukcja obliczania regresji
Call:
lm(formula = y ~ x)
Coefficients:
(Intercept)                x
    124604      30343
```

Interpretacja wyników:

1) Model regresji: To jest analiza, która próbuje zrozumieć, jak zmienne są ze sobą powiązane. W tym przypadku, badamy, jak zmienne Y (przychody przedsiębiorstw niefinansowych) zależą od zmiennej X (procentowej realizacji ustawy budżetowej).
2) Współczynnik korelacji Pearsona: Wynik R=0.8613427 to miara siły i kierunku związku między zmienną X (procentowa realizacja ustawy budżetowej) a zmienną Y (przychody przedsiębiorstw niefinansowych). Im bliżej wyniku do 1, tym silniejszy jest związek. W tym przypadku mamy silny pozytywny związek, co oznacza, że gdy procentowa realizacja ustawy budżetowej rośnie, przychody przedsiębiorstw niefinansowych także rosną.
3) Wykres punktowy (plot(x,y), inaczej wykres rozrzutu punktów empirycznych: Wykres przedstawia te same dane. Na osi X mamy procentową realizację ustawy budżetowej (X), a na osi Y mamy przychody przedsiębiorstw niefinansowych (Y). Wykres pokazuje ogólny trend wzrostu przychodów wraz ze wzrostem realizacji ustawy budżetowej.
4) Regresja liniowa: To jest wynik modelu regresji: y = a * x + b. Tutaj mamy dwie wartości:

- Intercept (Przecięcie): Wynosi b=124604. To jest punkt na osi Y, gdzie linia regresji przecina ją, gdy X wynosi 0. Można to zinterpretować jako początkowy poziom przychodów przedsiębiorstw niefinansowych, gdy procentowa realizacja ustawy budżetowej wynosi 0.
- x (Współczynnik nachylenia): Wynosi a= 30343. To jest wskaźnik, który mówi nam, jak zmieniają się przychody przedsiębiorstw niefinansowych w odpowiedzi na wzrost o 1 punkt procentowy procentowej realizacji ustawy budżetowej. Innymi słowy, jeśli procentowa realizacja ustawy budżetowej wzrośnie o 1%, można oczekiwać, że

przychody przedsiębiorstw niefinansowych wzrosną o około 30343 jednostek (mln PLN).

Podsumowując, wyniki sugerują, że istnieje silny pozytywny związek między realizacją ustawy budżetowej a przychodami przedsiębiorstw niefinansowych. W miarę wzrostu realizacji ustawy budżetowej można oczekiwać wzrostu przychodów, a to przecięcie na osi Y wskazuje na poziom przychodów, gdy realizacja wynosiła 0%. Jednak należy pamiętać, że regresja liniowa niekoniecznie oznacza przyczynowość, tj. nie można stwierdzić, że realizacja ustawy budżetowej bezpośrednio powoduje wzrost przychodów. To tylko analiza statystyczna związku między tymi zmiennymi.

ZAŁĄCZNIK NR 2

Instrukcje do tworzenia skryptu w R i instalacji programu R i RStudio

oznacza komentarz po instrukcji, który pojawi się jako tekst

rm(list=ls()) # instrukcja w programie RStudio służy do usuwania wszystkich zmiennych i obiektów z bieżącej sesji RStudio. Oznacza to, że wszystkie zmienne i obiekty, które były wcześniej zdefiniowane lub załadowane do środowiska RStudio, zostaną usunięte, co pozwala na wyczyszczenie przestrzeni roboczej i zwolnienie pamięci.

Działanie tej instrukcji jest przydatne, gdy chcemy wyczyścić przestrzeń roboczą przed rozpoczęciem nowej analizy lub projektu, aby uniknąć konfliktów nazw zmiennych i obiektów oraz zaoszczędzić pamięć komputera. Jednak trzeba zachować ostrożność przy jej użyciu, ponieważ usunięte zostaną wszystkie zmienne i obiekty, które były zdefiniowane w bieżącej sesji, i nie będzie możliwe odzyskanie ich po wykonaniu tej instrukcji.

ls() # Funkcja ls() w RStudio służy do uzyskania listy wszystkich zdefiniowanych zmiennych i obiektów w bieżącej sesji. To jest lista wszystkich rzeczy, które znajdują się w przestrzeni roboczej RStudio.

list=ls() #Ta część instrukcji przekazuje listę zmiennych i obiektów uzyskaną z ls() jako argument do funkcji rm(). Oznacza to, że funkcja rm() zostanie użyta do usunięcia wszystkich zmiennych i obiektów znajdujących się na tej liście.

rm(...) #Funkcja rm() w RStudio służy do usuwania zmiennych i obiektów. W przypadku rm(list=ls()) używa się jej do usunięcia wszystkich zmiennych i obiektów na liście przekazanej jako argument, czyli wszystkich zmiennych w bieżącej sesji.

setwd("C:/Program R")#Teachers calatogue - ustanowienie katalogu nauczyciela

mydata <-read.table("learning2.txt",sep="\t", header=TRUE) #odczytanie danych z tablicy, gdzie poszczególne kolumny oddzielone są separatorem tabulacji

y <- mydata$Exam #przypisanie zmiennej w kolumnie opisanej jako EXAM nazwy y

x <- mydata$IQ #przypisanie zmiennej w kolumnie opisanej jako IQ nazwy x

z <- mydata$Hours #przypisanie zmiennej w kolumnie opisanej jako HOURS nazwy x

length(x) #długość zmiennej x, ile ma obserwacji

range(x) # rozrzut zmiennej x, czyli max-min

min(x) # wartość minimum dla zmiennej x

max(x) # wartość maksimum dla zmiennej x

```
median(x) # mediana dla x
mean(x) # średnia dla x
fivenum(x) # pięć statystyk: min kwartyl1 mediana kwartyl3 max
sd(x) # odchylenie standardowe x
sd(x)/mean(x)  #współczynnik zmienności względem średniej x
IQR(x) # rozstęp kwartylowy (kwartyl3 – kwartyl1)
IQR(x)/median(x) #kwartylowy współczynnik zmienności względem mediany
boxplot(x) # wykres pudełkowy
hist(x) #histogram liczebności
library(e1071) # instrukcja załadowania biblioteki umożliwiającej obliczanie
m.in.skośności, niektórych wykresów i testów
skewness(x) #skośność
mean(x)-sd(x) #lewa granica typowego przedziału zmienności dla rozkładu
symetrycznego
mean(x)+sd(x) #prawa granica typowego przedziału zmienności dla rozkładu
symetrycznego
median(x)-(IQR(x)/median(x)) #lewa granica typowego przedziału zmienności dla
rozkładu asymetrycznego
median(x)+(IQR(x)/median(x)) #prawa granica typowego przedziału zmienności dla
rozkładu asymetrycznego
cor(x,y) # polecenie obliczenia współczynnika korelacji liniowej Pearsona dla
zmiennych x i y
cor(x,y)/(sd(x)*sd(y)) # współczynnik regresji a w równaniu y = a * x + b
mean(y)-(cor(x,y)/(sd(x)*sd(y))) # stała b równania regresji w równaniu y = a * x + b
cor(x,y)^2 # współczynnik determinacji R^2
1-cor(x,y)^2 # współczynnik zbieżności 1-R^2
print(x) #wydrukowanie kolejnych wartości zmiennej x
print(y) #wydrukowanie kolejnych wartości zmiennej y
library(e1071) # załadowanie biblioteki, umożliwiającej wykonywanie niektórych
instrukcji, np. skośności, niektórych wykresów, testu t
plot(x,y) # wykres rozrzutu punktów empirycznych dla zmiennych x i y
lm(y~x) #regresja liniowa dla zmiennej zależnej y i zmiennej niezależnej x
t_test_result <- t.test(x, y, paired = TRUE) # Test t Studenta dla par, po załadowaniu
library(e1071)
```

print(t_test_result) # Wyświetlenie wyników testu
boxplot(x, y, names = c("Przed regulacją", "Po regulacji"), col = c("red", "blue"), main = "emisji pyłów PM2.5 w mieście przed i po ich wprowadzeniu") # Tworzenie wykresu pudełkowego dla x i y na tym samym obrazku

Jak pobrać program R

Aby pobrać program R, najlepiej jest wybrać lustra CRAN (Comprehensive R Archive Network) z listy dostępnych miejsc, które znajdują się na stronie głównej pobierania programu R. Lustra CRAN są serwerami, na których znajdują się kopie programu R, i możesz wybrać dowolne z nich, aby pobrać program. Wybór lustra zależy od twojej lokalizacji geograficznej, ponieważ wybierając najbliższe lustra, można osiągnąć lepszą wydajność pobierania.

Oto jak postępować:

Wejdź na stronę główną pobierania programu R: Strona pobierania programu R: https://cran.r-project.org/mirrors.html

Na stronie znajdziesz listę lustra CRAN podzieloną według krajów i regionów.

Wybierz lustra (mirror), które są najbliższe twojej lokalizacji geograficznej, aby uzyskać lepszą wydajność pobierania (sugeruję Austrię, zawsze się ładuje)

Kliknij na wybrane lustro, aby otworzyć stronę z dostępnymi plikami instalacyjnymi dla różnych systemów operacyjnych.

Wybierz odpowiedni plik instalacyjny dla swojego systemu operacyjnego (na przykład Windows, macOS lub Linux) i rozpocznij proces pobierania.

Pamiętaj, że wybór lustra CRAN zależy od twojej lokalizacji geograficznej, więc możesz wybrać dowolne lustra z listy, które są najbliższe tobie.

Np. Austria (zwykle działa), wtedy ładuje się strona:

https://cran.wu.ac.at/

Jeżeli naszym systemem jest WIndows, wybieramy Windows:

Download R for Windows

Następnie pojawia się okienko, w którym klikamy w pierwszym wierszu na install R for the first time:

Warto obejrzeć film instruktarzowy na youtube:

https://www.youtube.com/watch?v=KhVxneAJAsY

Opis programów do załadowania.

W katalogu mamy dwa różne programy: RStudio i R (język R). Oto, co możesz z nimi zrobić:

R (język R):
Plik o nazwie "R4.22-win" jest instalatorem języka R w wersji 4.22 dla systemu Windows. Jeśli jeszcze nie masz języka R zainstalowanego na swoim komputerze, możesz zainstalować go, uruchamiając ten instalator. Po zainstalowaniu R będziesz mógł korzystać z języka R i uruchamiać skrypty R w dowolnym edytorze kodu, w tym w RStudio.
RStudio:
Plik o nazwie "RStudio-2022.12.0-353" jest instalatorem RStudio w wersji 2022.12.0-353. RStudio to środowisko programistyczne, które ułatwia pracę z językiem R. Jeśli jeszcze nie masz RStudio zainstalowanego, możesz zainstalować go, uruchamiając ten instalator. RStudio oferuje wiele zaawansowanych narzędzi do pracy z R, w tym tworzenia raportów R Markdown.
Jeśli chcesz korzystać z obu tych narzędzi razem, oto co powinieneś zrobić:

1. Rozpocznij od zainstalowania języka R, uruchamiając instalator "R4.22-win".
2. Następnie zainstaluj RStudio, uruchamiając instalator "RStudio-2022.12.0-353".
3. Po zainstalowaniu obu tych programów, uruchom RStudio, a następnie utwórz nowy projekt RStudio, aby zacząć pracować nad swoimi analizami danych w przyjaznym środowisku.

Pamiętaj, że RStudio jest interfejsem użytkownika dla języka R, który ułatwia pracę i tworzenie raportów, ale sam w sobie nie jest językiem R. Język R musi być zainstalowany, aby RStudio działało.

Przeprowadzenie szkoleń dla pracowników administracji publicznej z zakresu analizy danych w ramach oceny skutków regulacji

Fundusze Europejskie
Wiedza Edukacja Rozwój

Rzeczpospolita Polska

Unia Europejska
Europejski Fundusz Społeczny

Skrypt do części: Excel zaawansowany

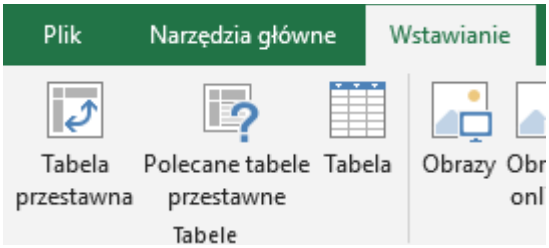
Spis treści

TABELE PRZESTAWNE I WYKRESY PRZESTAWNE

Tabela przestawna jest zaawansowanym narzędziem służącym do analizy danych, wykonywania obliczeń oraz tworzenia zestawień, umożliwiając prezentację porównań, wzorców i tendencji danych.

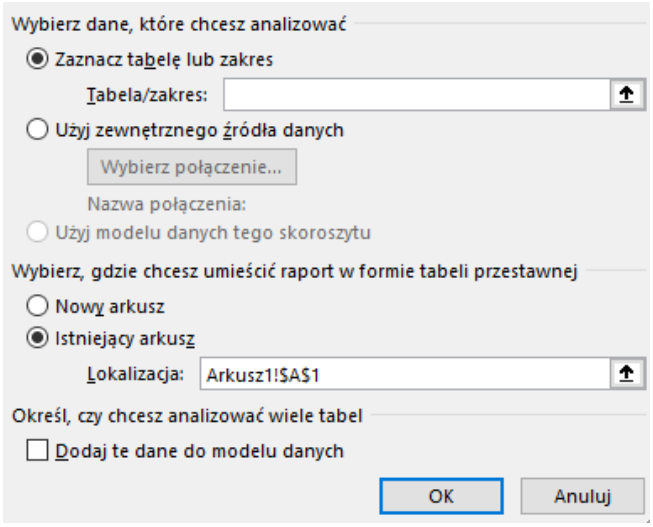
Tworzenie tabeli przestawnej

1. Zaznacz komórki, z których chcesz utworzyć tabelę przestawną. Dane nie powinny zawierać pustych wierszy ani kolumn. Nagłówek może składać się tylko z jednego wiersza. Nagłówki muszą być unikalne.
2. Wybierz pozycję **Wstawianie** > **Tabela przestawna**.



Rysunek 1. Wstążka programu MS Excel

3. W obszarze **Wybierz dane, które chcesz analizować** wybierz pozycję **Zaznacz tabelę lub zakres**.

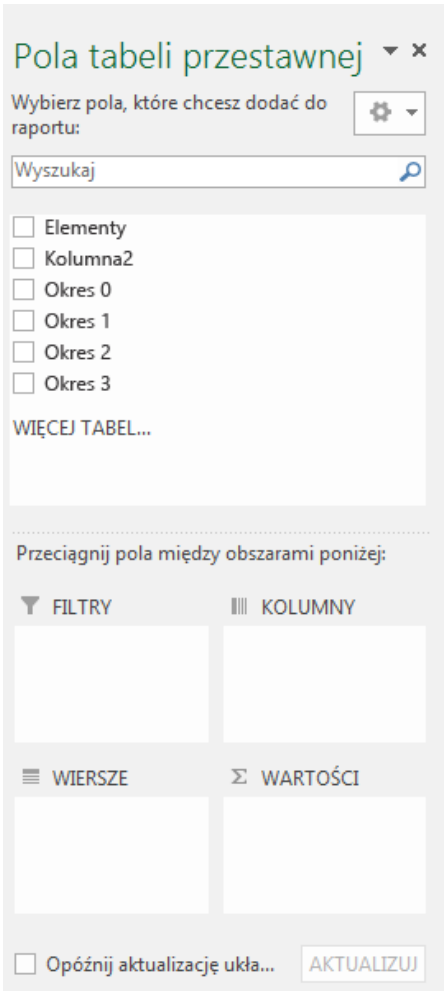


Rysunek 2. Okno tworzenia tabeli przestawnej

4. W obszarze **Tabela/zakres** sprawdź zakres komórek.
5. W obszarze **Wybierz, gdzie chcesz umieścić raport w formie tabeli przestawnej** zaznacz pozycję **Nowy arkusz**, aby umieścić tabelę przestawną w nowym arkuszu. Możesz również zaznaczyć pozycję **Istniejący arkusz**, a następnie wybrać lokalizację wyświetlania tabeli przestawnej.
6. Wybierz przycisk **OK**.

Modyfikacja tabeli przestawnej

1. Aby dodać pole do tabeli przestawnej, zaznacz pole wyboru obok nazwy pola w okienku **Pola tabeli przestawnej**. Wybrane pola są dodawane do obszarów domyślnych: pola nieliczbowe są dodawane do obszaru **Wiersze**, hierarchie dat i godzin są dodawane do obszaru **Kolumny**, a pola liczbowe są dodawane do obszaru **Wartości**.



Rysunek 3. Okno pola tabeli przestawnej

2. Aby przenieść pole z jednego obszaru do drugiego, przeciągnij je do obszaru docelowego.

Tworzenie wykresów przestawnych

Czasem analiza ogólnego kontekstu danych staje się wyzwaniem, zwłaszcza gdy dane początkowe są jeszcze nie podsumowane. Pierwszym krokiem może być stworzenie tabeli przestawnej, jednak nie każdy jest w stanie szybko przeanalizować liczby w tabeli i zrozumieć bieżący stan. Wykresy przestawne stanowią efektywny sposób na wizualizację tych danych, co ułatwia ich interpretację.

Tworzenie wykresu przestawnego

1. Zaznacz komórki w tabeli.
2. Wybierz pozycję **Wstawianie > Wykres przestawny**
3. Wybierz przycisk **OK**.

Tworzenie wykresu z tabeli przestawnej

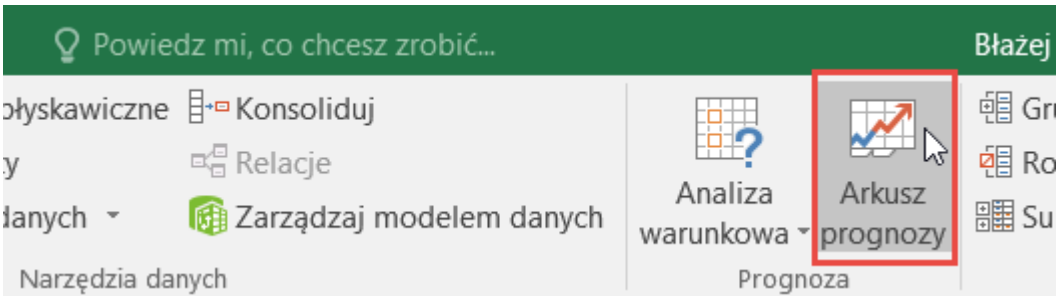
1. Zaznacz komórkę w tabeli.
2. Wybierz pozycję **Narzędzia tabel przestawnych > Analiza > Wykres przestawny**.
3. Zaznacz wykres.
4. Wybierz przycisk **OK**.

Arkusz prognozy

Im więcej czynników i kryteriów mamy do uwzględnienia, tym bardziej precyzyjne prognozy możemy tworzyć. Program Microsoft Excel również oferuje narzędzia do przewidywania przyszłych wartości. Od wersji Excel 2016 mamy dostęp do bardziej intuicyjnego narzędzia o nazwie "Arkusz Prognozy". Jeśli posiadamy dane historyczne oparte na czasie, możemy wykorzystać je do tworzenia prognoz. Proces tworzenia prognozy w programie Excel polega na utworzeniu nowego arkusza, który zawiera zarówno dane historyczne, jak i prognozowane, oraz wykres przedstawiający te dane. Dzięki prognozie można na przykład przewidzieć przyszłą inflację, potrzeby związane z magazynowaniem lub trendy konsumenckie.

Tworzenie prognozy

1. W arkuszu wprowadź dwie powiązane serie danych: serię dat lub godzin w celu określenia osi czasu oraz serię odpowiadających im wartości. Prognoza będzie dotyczyła tych wartości w przyszłości. Oś czasu wymaga spójnych interwałów między punktami danych. Mogą to być na przykład interwały miesięczne z wartościami pierwszego dnia każdego miesiąca, interwały miesięczne lub liczbowe.
2. Zaznacz obie serie danych.
3. Na karcie **Dane** w grupie **Prognoza** kliknij pozycję **Arkusz Prognoza.**



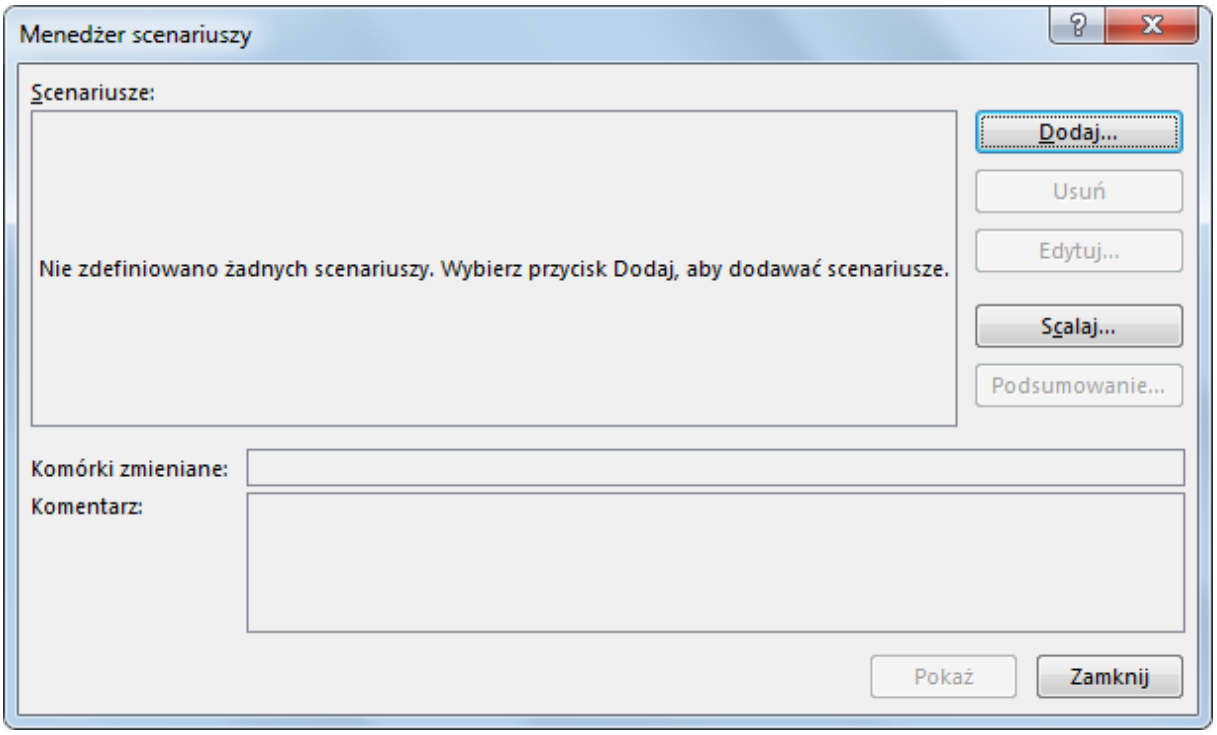
Rysunek 4. Przycisk Arkusz prognozy na karcie Dane

4. W **oknie Tworzenie arkusza prognozy** wybierz wykres liniowy lub wykres kolumnowy, aby uzyskać wizualną reprezentację prognozy.
5. W polu **Koniec prognozy** wybierz datę końcową, a następnie kliknij przycisk **Utwórz**.

   Excel tworzy nowy arkusz zawierający zarówno tabelę wartości historycznych i prognozowanych, jak i wykres, który przedstawia te dane. Nowy arkusz znajduje się po lewej stronie ("przed") arkuszem, w którym wprowadzono serię danych.

Scenariusze

Scenariusz to zestaw danych, które Excel zapisuje i automatycznie wstawia do arkusza. Możemy tworzyć i zapisywać różne zestawy danych jako scenariusze, a potem zmieniać między nimi, aby zobaczyć różne wyniki.
Kiedy już zgromadzimy wszystkie niezbędne scenariusze, możemy stworzyć raport podsumowujący, który zawiera informacje ze wszystkich tych scenariuszy.
Do zarządzania scenariuszami – dodawania, usuwania, edycji i przełączania między nimi służy Kreator menedżera scenariuszy w grupie Narzędzia danych w przycisku Analiza warunkowa na karcie Dane.



Rysunek 5. Okienko dialogowe Menadżer scenariuszy

MAKROPOLECENIA - AUTOMATYZACJA ZADAŃ W ARKUSZACH

Rejestracja makr

Aby zautomatyzować powtarzalne czynności, można zarejestrować makro, korzystając z Rejestratora makr w programie Microsoft Excel. Można zarejestrować makro na przykład podczas stosowania odpowiedniego formatu, a następnie odtworzyć je, gdy będzie to znów potrzebne.

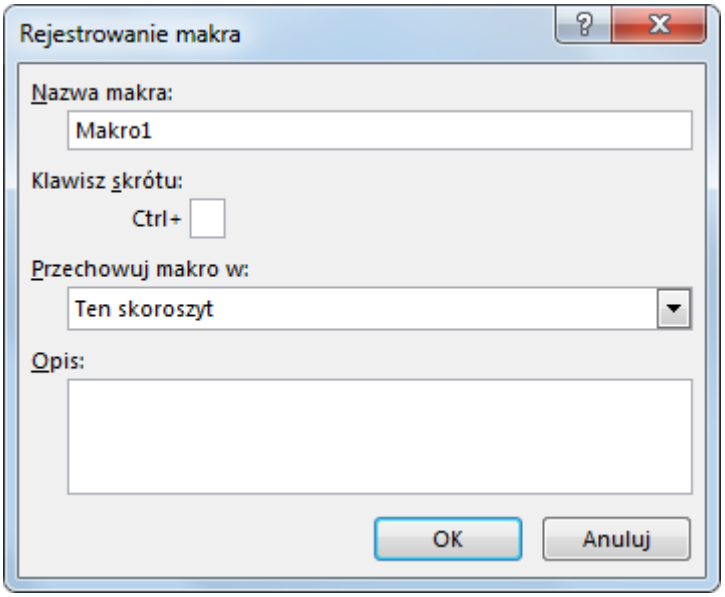
Podczas rejestrowania makra Rejestrator makr zapisuje wszystkie kroki w postaci kodu języka Visual Basic for Applications (VBA). Te kroki to na przykład wpisywanie tekstu lub liczb; klikanie komórek albo poleceń na wstążce lub w menu; formatowanie komórek, wierszy lub kolumn; a nawet importowanie danych ze źródła zewnętrznego, na przykład z programu Microsoft Access. Język VBA jest częścią zaawansowanego języka programowania Visual Basic i jest dostępny w większości aplikacji pakietu Office. Język VBA umożliwia automatyzowanie procesów zarówno w jednej aplikacji pakietu Office, jak i w kilku różnych aplikacjach, ale nie trzeba znać języka VBA ani mieć umiejętności programowania, jeśli Rejestrator makr spełnia wymagania.

Warto pamiętać, że gdy makro jest rejestrowane, Rejestrator makr zapisuje niemal wszystkie wykonywane czynności. Zatem w przypadku pomyłki podczas rejestrowanej sekwencji czynności, na przykład kliknięcia niewłaściwego przycisku,

Rejestrator makr zarejestruje tę pomyłkę. Rozwiązaniem problemu może być ponowne zarejestrowanie całej sekwencji lub zmodyfikowanie kodu VBA. Dlatego zawsze, warto najpierw dobrze zaznajomić się z całym procesem. Im sprawniej zostanie wykonana sekwencja podczas rejestrowania, tym wydajniej będzie działać utworzone makro.

Narzędzia makr i języka VBA znajdują się na karcie Deweloper, która jest domyślnie ukryta, dlatego pierwszym krokiem jest włączenie tej karty w opcjach programu Microsoft Excel.

Aby zarejestrować makro należy kliknąć polecenie Zarejestruj makro w grupie Kod na karcie deweloper, które otwiera okienko dialogowe Rejestrowanie makra. Okienko to pozwala na nadanie makru własnej nazwy, przypisanie klawisza skrótu, wybranie miejsce przechowywania makra oraz wpisanie opcjonalnego opisu. Naciśnięcie przycisku OK rozpoczyna rejestrowanie makra – po jego kliknięciu należy wykonać żądaną sekwencję czynności, które mają zostać zarejestrowane.



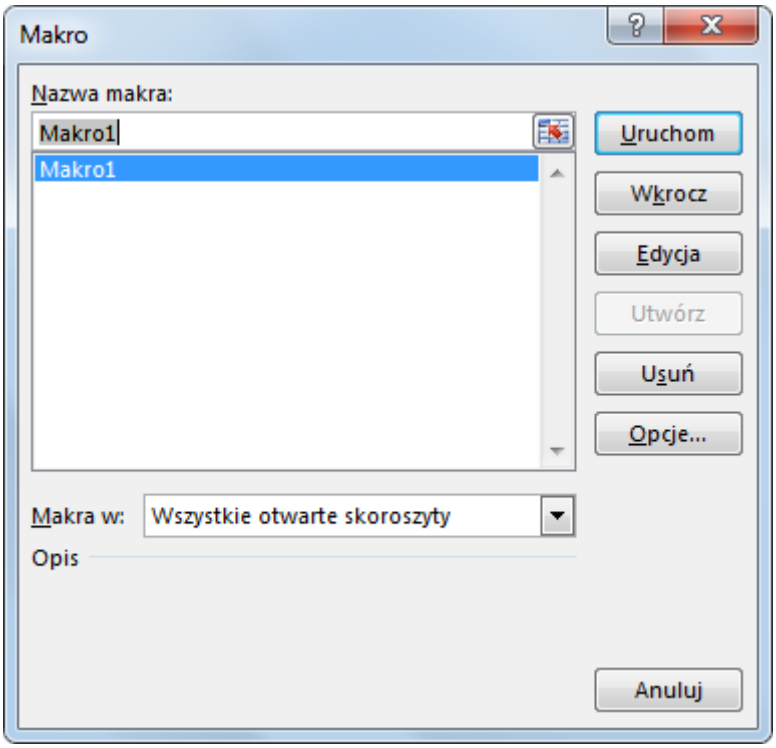
Rysunek 6. Okienko dialogowe Rejestrowanie makra

Aby zakończyć rejestrowanie makra należy kliknąć polecenie Zatrzymaj rejestrowanie na karcie Deweloper w grupie Kod.

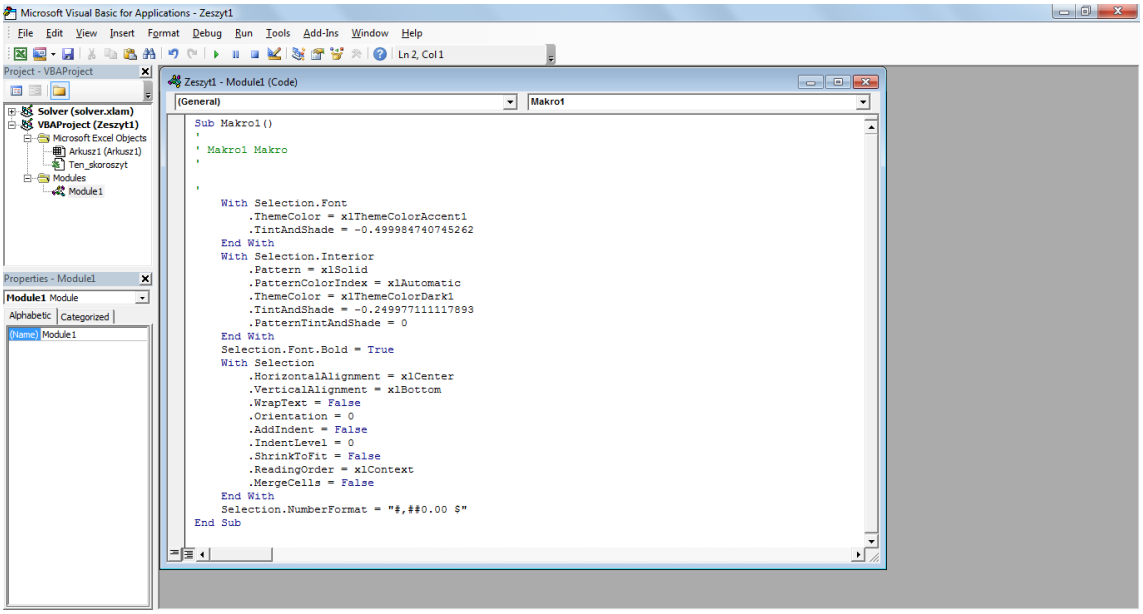
Zrozumienie i modyfikacja kodu makr (elementy języka VBA)

W przypadku kiedy makro nie działa w zamierzony sposób można przystąpić do modyfikacji jego kodu. Aby wyświetlić kod zapisanych makr należy otworzyć okienko dialogowe Makro poleceniem Makra w grupie kod na karcie Deweloper, następnie

wybrać jedno z zarejestrowanych makr z listy i kliknąć przycisk Edycja. W tym okienku możliwe jest również tworzenie, uruchamianie i usuwanie makr.



Rysunek 7. Okienko dialogowe Makro



Rysunek 8. Kod VBA makra zarejestrowanego w programie Microsoft Excel

Ze względu na obszerność języka VBA poniżej umieszczono kilka podstawowych poleceń, które mogą okazać się przydatne przy modyfikowaniu makr.

Polecenie VBA	Opis
Sheets("Arkusz1").Select	Wybranie aktywnego arkusza o nazwie

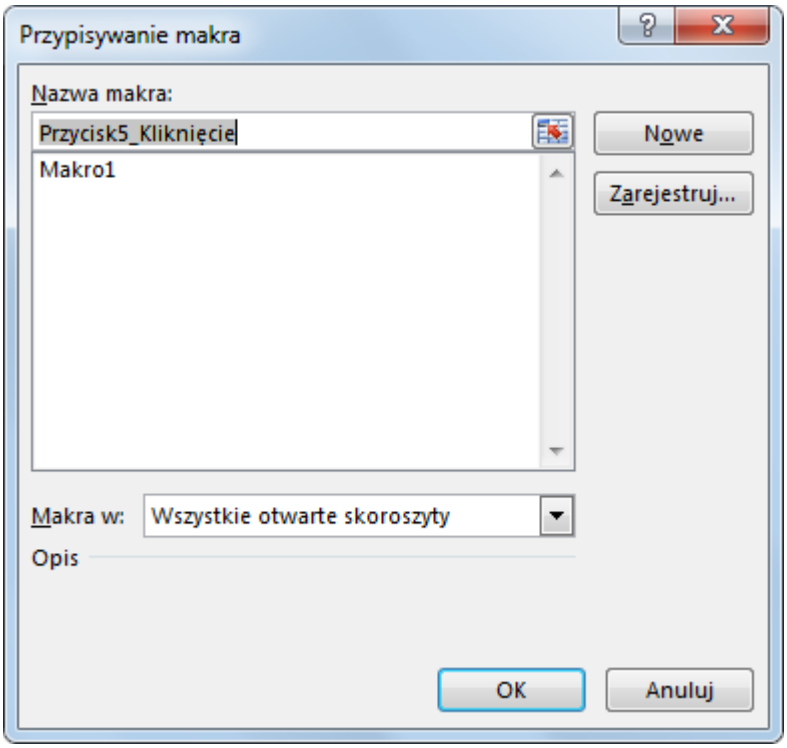
Polecenie VBA	Opis
	Arkusz1
Range("A1") = "dowolna treść"	Wpisanie wartości do komórki A1
Cells(3, 2) = "dowolna wartość"	Wpisanie wartości do komórki w trzecim wierszu, drugiej kolumnie, innymi słowy do komórki B3
zmienna = Sheets("Arkusz2").Range("D23").Value	Pobranie do zmiennej o nazwie zmienna wartości z komórki D23 z arkusza o nazwie Arkusz2
zmienna = Sheets("Arkusz2").Cells(2, 2)	Pobranie do zmiennej o nazwie zmienna wartości z komórki B2 z arkusza o nazwie Arkusz2
Range("A1:B10").Select	Ustawienie zaznaczenia na zakres komórek od A1 do B10
Selection.Copy	Polecenie kopiuj
Selection.Cut	Polecenie wytnij
Selection.Paste	Polecenie wklej
Selection. ClearContents	Polecenie usuń
ActiveWorkbook.Save	Polecenie zapisz
MsgBox ("Witaj świecie!")	Wyświetlenie komunikatu w okienku dialogowym
InputBox("Uzupełnij komórkę A1")	Wyświetlenie okienka do pobierania danych od użytkownika
Selection.Font.Size = 20	Ustawienie wielkości czcionki w zaznaczonych komórkach
Selection.Font.Bold = True	Ustawienie pogrubienia czcionki w zaznaczonych komórkach
Selection.Font.Name = "Arial"	Ustawienie kroju czcionki w zaznaczonych komórkach
Selection.NumberFormat = "#,##0.00 $"	Ustawienie formatu danych w zaznaczonych komórkach
Selection.Interior.Color = 65535	Ustawienie koloru wypełnienia w zaznaczonych komórkach

Podłączanie makr do interfejsu programu lub skoroszytu

Zarejestrowane makro można uruchamiać przyciskiem Uruchom w okienku dialogowym Makro, jednak jest to dość kłopotliwe, szczególnie, że po każdym uruchomieniu makra okienko zamyka się i trzeba go otwierać na nowo. Oprócz przypisania klawisza skrótu do makra można do ich szybkiego uruchamiania wykorzystać formanty, które należy uprzednio wstawić do skoroszytu.

Najczęściej do uruchamiania makr stosuje się formant Przycisk, znajdujący się w grupie Kontrolki formularza w menu przycisku Wstaw w grupie Formanty na karcie Deweloper.

Po wstawieniu przycisku automatycznie zostaje wyświetlone okienko dialogowe Przypisywanie makra, w którym z listy istniejących makr należy wskazać makro, które ma być wykonywane przy kliknięciu przycisku.



Rysunek 9. Okienko dialogowe Przypisywanie makra

Tworzenie formularzy ekranowych lub drukowanych

Formularze w programach Microsoft Office składają się z formantów, dostępnych w grupie Formanty na karcie Deweloper.

Rysunek 6. Przykładowy formularz stworzony w programie Microsoft Excel

Powyższy formularz składa się z następujących formantów ActiveX: dwóch pól tekstowych, dwóch przycisków opcji oraz jednego pola kombi.

Tak utworzony formularz w pliku Excel można wydrukować lub wysyłać do wypełnienia. Należy jednak pamiętać o włączeniu ochrony arkusza, aby osoba wypełniająca nie mogła, poza wypełnianiem formularza, dokonywać żadnych innych zmian. Każdy formant, tak jak każda komórka, posiada w okienku formatowania zakładkę Ochrona z domyślnie zaznaczonym atrybutem Zablokowane, który należy odznaczyć, aby po włączeniu ochrony możliwa była jego edycja.

Przeprowadzenie szkoleń dla pracowników administracji publicznej z zakresu analizy danych w ramach oceny skutków regulacji

Fundusze Europejskie
Wiedza Edukacja Rozwój

Rzeczpospolita Polska

Unia Europejska
Europejski Fundusz Społeczny

Skrypt do części: Prezentowanie wniosków i wyników badań – część 1

Spis treści

PODEJŚCIE PROBLEM-CEL W PREZENTACJI DANYCH - PRZYPOMNIENIE

Przygotowując prezentację danych, odpowiedz sobie na następujące pytania:

1. Jaki problem swojej widowni/odbiorcy rozwiązujesz?
2. Jaki jest cel ogólny/typ Twojej prezentacji danych? (informacyjny, perswazyjny, inspiracyjny)?
3. Jaki jest cel szczegółowy Twojej prezentacji, czyli co ma się zmienić w życiu widowni po zakończeniu Twojej prezentacji?

Prezentacja informacyjna: Moja widownia dowie się

_____

Prezentacja perswazyjna: Moja widownia przekona się do/zrobi

_____

Prezentacja inspiracyjna: Moja widownia uświadomi sobie, że

_____

ANALIZA WIDOWNI

- Problem
- Efekt – odbiorcy i beneficjenci
- Wartość
- Następstwo – odbiorcy i beneficjanci
- Instrukcja

ANALIZA POTRZEB WIDOWNI – PRZYDATNE PYTANIA

- Jaki problem chcemy rozwiązać za pomocą określonej prezentacji danych (powinien to być problem zarówno z punktu widzenia odbiorców naszej prezentacji, jak i jej beneficjentów)?
- Jakie są negatywne konsekwencje tego problemu dla widowni/odbiorców naszej prezentacji danych?
- Jakie są negatywne konsekwencje tego problemu dla beneficjentów prezentacji danych?
- Jakie rozwiązanie tego problemu chcemy zaprezentować? (powinno ono stanowić wartość zarówno dla odbiorców, jak i beneficjentów naszej prezentacji danych).
- Jakie są korzyści wynikające z rozwiązania z punktu widzenia widowni/odbiorców naszej prezentacji?

- Jakie są korzyści wynikające z rozwiązania z punktu widzenia beneficjentów prezentacji danych?
- Jak to rozwiązanie powinno być wprowadzone w praktyce?

STORYTELLING

Wyróżnia się cztery niezbędne składniki każdej historii:

A. Cel historii, który jest tożsamy z celem prezentacji.
B. Problem bohatera historii, który powinien pokrywać się z problemem widowni/odbiorcy.
C. Główny bohater, z którym widownia powinna się utożsamić.
D. Narracja oparta o zmianę emocji.

Na podstawie tych czterech składników, skonstruowana została tabela, która ułatwi nam gromadzenie historii do wykorzystania w prezentacjach danych:

Historia, którą chcemy wykorzystać.	Dane, jakie mają być przekazane z pomocą tej historii.	Cel prezentacji danych (cel ogólny i szczegółowy).	Problem odbiorcy naszej prezentacji.	Główny bohater historii (kim jest i na ile jest podobny do odbiorców naszej historii?)

Rola storytellingu w prezentacji danych

Zgodnie z tzw. trójkątem retorycznym Arystotelesa, wyróżniamy trzy rodzaje argumentów. Będą one wykorzystywane przede wszystkim w perswazyjnych i inspiracyjnych prezentacjach danych:

A. **Argumenty typu logos -** są to wszystkie argumenty opierające się na logicznym rozumowaniu, wnioskowaniu oraz dowodzeniu. Argumentami typu logos będą zatem wnioski, które wyciągniemy z interpretacji danych i przedstawimy widowni w naszej prezentacji.
B. **Argumenty typu pathos -** odwołują się one do emocji i uczuć naszej widowni. W budowaniu argumentacji typu pathos pomocny jest storytelling, którego jednym z kluczowych elementów jest zmiana emocji.

C. **Argumenty typu ethos -** odnoszą się one do wartości i autorytetów wyznawanych przez naszą widownię. Poprzez oparcie historii na problemie audytorium, storytelling sprzyja odwoływaniu się do wartości istotnych z punktu widzenia audytorium.

W prezentacjach perswazyjnych, istotne jest równoważenie argumentów logicznych i emocjonalnych, co uzasadnia łączenie wniosków płynących z interpretacji danych ze storytellingiem. W prezentacjach inspiracyjnych, znacznie mocniej opieramy się na odwołaniach do emocji widowni, a wykorzystane w tym przypadku dane będą pełnić funkcję środka pobudzającego do refleksji. Trójkąt retoryczny Arystotelesa pokazuje nam jednak możliwość znacznego zwiększenia skuteczności naszego wystąpienia, jeśli w odpowiedni sposób połączymy dane i płynące z nich wnioski ze zmianą emocji, jaką wywołuje storytelling.

Model bajki Leacha i Greimasa

Model bajki stanowi narzędzie zapewniające dopasowanie konkretnej historii do potrzeb audytorium. Wskazuje ono na sześć kluczowych elementów, które nie tylko muszą być zawarte w historii, ale także od których podobieństwa do realnego życia widowni, zależeć będzie skuteczność konkretnej historii. Są to:

- Główny bohater – musi on być na tyle podobny do widowni docelowej, aby nasi odbiorcy mogli się z nim utożsamić.
- Cel głównego bohatera, czyli to, co główny bohater historii chce osiągnąć; przyczyna, dla której podejmuje konkretne starania.
- Problem głównego bohatera – powinien być zbieżny z problemem naszego audytorium, który chcemy rozwiązać poprzez konkretną prezentację.
- Mentor/Darczyńca - osoba, która pomaga głównemu bohaterowi rozwiązać problem i osiągnąć cel, udzielając mu wsparcia.
- Wsparcie - rozwiązanie problemu przekazane głównemu bohaterowi przez mentora, np. W formie rady lub określonej wiedzy. Wsparcie musi być tożsame z celem szczegółowym naszej prezentacji danych.
- Obdarowani - krąg osób, które zyskują na tym, że głównemu bohaterowi udało się rozwiązać problem i osiągnąć określony cel. Krąg obdarowanych to, poza

samym głównym bohaterem, osoby, na których zależy naszej widowni. Często mogą to być beneficjenci określonych danych[1].

Temperatura konfliktu w historii w prezentacji danych

Temperatura konfliktu to narzędzie służące ocenie tego, jak poważny jest problem, z którym mierzy się główny bohater naszej historii. Ponieważ problem głównego bohatera historii jest tożsamy z problemem naszej widowni, określa ono jednocześnie powagę problemu audytorium, przed którym prezentujemy.

Temperaturę konfliktu mierzymy w pięciostopniowej skali:

1 – brak jakiegokolwiek problemu; sielanka;

2 – bardzo niewielki problem – metaforycznie obrazuje się jego trudność jako “zabicie muchy dla olbrzyma”.

3 – dylemat, przed którym staje dorosły człowiek i w którym żadna z opcji wyboru nie jest na pierwszy rzut oka oczywista.

4 – bardzo duży problem - metaforycznie obrazuje się jego trudność jako “przepłynięcie oceanu dla zwykłego śmiertelnika”.

5 – sytuacja niemożliwa do rozwiązania, apokalipsa.

Aby nasze historie użyte w prezentacji danych mogły być skuteczne i jednocześnie odwoływały się do realnego problemu widowni/odbiorcy, temperatura konfliktu powinna wynosić trzy.

Struktura storytellingu STAR

Narracja historii w prezentacjach danych przebiega w oparciu o strukturę STAR. Nazwa tej struktury pochodzi od czterech angielskich słów:

S – situation – sytuacja;

T – task – zadanie.

A – action – działanie.

R – result – rezultat.

Każda z tych części zawiera następujące składniki:

Sytuacja - to ogólne wprowadzenie do historii, które polega na osadzeniu historii w miejscu i w czasie oraz zaznajomieniu widowni z bohaterem. Zawiera ono:

- Osadzenie historii w miejscu i w czasie.
- Przedstawienie głównego bohatera.

[1] E. Leach & A. Greimas, Rytuał i narracja 1996.

- Wskazanie ukrytego celu/marzenia głównego bohatera.

Zadanie to część, w której krystalizuje się cel głównego bohatera, czyli zadanie, do którego osiągnięcia będzie on dążyć. Zawiera ona:

- Impuls zewnętrzny.
- Wewnętrzne wahanie bohatera.
- Decyzja odnośnie realizacji określonego celu.

Działanie to część pokazująca postępowanie bohatera. To tutaj rozgrywa się typowa dla storytellingu zmiana emocji. Bohater stopniowo napotyka coraz większe trudności, upada w punkcie kulminacyjnym aż w końcu otrzymuje i stopniowo wdraża radę mentora. Zawiera ona:

- Początki.
- Piętrzące się trudności.
- Punkt kulminacyjny.
- Rozmowę z mentorem i uzyskanie wsparcia.
- Stopniowe wdrażanie rady mentora.

Rezultaty to część pokazująca, co udało się osiągnąć bohaterowi w wyniku danej historii. Zawiera ona:

- Rezultat zewnętrzny – osiągnięcie celu przez bohatera i dalsze wynikające z tego korzyści.
- Rezultat wewnętrzny – wewnętrzna przemiana bohatera.

Newralgiczne punkty w strukturze STAR

W ramach powyższej struktury należy wskazać pewne newralgiczne punkty. Pełnią one istotną rolę z punktu widzenia zarządzania uwagi naszej widowni oraz m.in. prezentacji danych.

Są nimi:

- Decyzja odnośnie realizacji określonego celu.
- Piętrzące się trudności.
- Punkt kulminacyjny.
- Rozmowa z mentorem.
- Rezultat zewnętrzny.
- Rezultat wewnętrzny.

Najważniejsze z punktu widzenia prezentacji danych będą punkt kulminacyjny oraz oba rezultaty.

Zmiana emocji w storytellingu

Zmianę emocji w naszej historii możemy wywołać dzięki świadomości stymulowania określonych hormonów w organizmie naszej widowni i powiązanych z nimi reakcji emocjonalnych. Najważniejsze hormony pod kątem pracy z historią, sposób działania oraz towarzyszące im emocje zostały wskazane w poniższej tabelce:

Hormon	Reakcja u widowni	Emocje/odczucie
Dopamina.	Ciekawość, oczekiwanie co dalej, chęć poznania większej ilości szczegółów.	Ciekawość, zainteresowanie, napięcie.
Oksytocyna.	Utożsamienie się z głównym bohaterem, odczuwanie więzi, szczególnie w odniesieniu do problemu i negatywnych sytuacji.	Współczucie, smutek, bliskość z głównym bohaterem.
Serotonina	Odczuwanie szczęścia, ulgi, kiedy głównemu bohaterowi uda się rozwiązać problem i osiągnąć cel.	Szczęście, ulga.

Oksytocyna Utożsamienie się z głównym bohaterem, odczuwanie więzi, szczególnie w odniesieniu do problemu i negatywnych sytuacji. Współczucie, smutek, bliskość z głównym bohaterem.

Przeprowadzenie szkoleń dla pracowników administracji publicznej z zakresu analizy danych w ramach oceny skutków regulacji

Fundusze Europejskie
Wiedza Edukacja Rozwój

Rzeczpospolita Polska

Unia Europejska
Europejski Fundusz Społeczny

Skrypt do części: Prezentowanie wniosków i wyników badań – część 2
Spis treści

STRUKTURA PREZENTACJI DANYCH

Matryca AIA Esther Choy

Pod kątem prezentacji danych, Esther Choy wyróżniła funkcję, jakie w ramach tej prezentacji pełnią wstęp, rozwinięcie i zakończenie. Jej model AIA nie stanowi struktury żadnego z tych elementów, lecz jedynie akcentuje ich podstawową rolę.

Odwołanie do problemu - Appeal to a problem - funkcją wstępu jest odwołanie się do problemu.

Poinformowanie - Inform **–** funkcją rozwinięcia jest realizacja celu ogólnego prezentacji. W przypadku prezentacji informacyjnych będzie to poinformowanie, perswazyjnych – przekonanie, a inspiracyjnych – zainspirowanie/uświadomienie.

Działanie - **A**ction - funkcją zakończenia jest wezwanie widowni do działania, czyli innymi słowy przeniesienia wniosków płynących z interpretacji danych na swoje własne życie[2].

Pomimo swojej ogromnej użyteczności, wadą modelu AIA jest brak rozróżnienia pomiędzy trzema celami ogólnymi prezentacji danych. Stąd kolejne uwagi dotyczące struktury wstępu, rozwinięcia i zakończenia będą wskazywać na różnice pomiędzy poszczególnymi typami prezentacji.

Struktura wstępu

Wstęp prezentacji danych najlepiej jest budować na bazie modelu **“3 W”**:

Wciągnięcie widowni (przykucie uwagi).

Wskazanie problemu (problem widowni i negatywne konsekwencje).

Wartość (zaanonsowanie celu i struktury).

Struktura zakończenia

Zakończenie prezentacji danych najlepiej jest budować na bazie modelu **“3 P”**:

Podsumowanie.

Przesłanie.

Propozycja wykorzystania w życiu (kolejny krok, wezwanie do działania, pytanie):

- wystąpienie informacyjne: kolejny krok;
- wystąpienie perswazyjne: wezwanie do działania;
- wystąpienie inspiracyjne: pytanie.

[2] E. Choy, Let your story do the work, Londyn 2017

Informacyjna struktura rozwinięcia:

1-2-3.

Perswazyjna struktura rozwinięcia:

Co? – rozwiązanie.

Dlaczego? – korzyści.

Jak? – sposób użycia w praktyce.

PRACA Z DANYMI

Badania eksploracyjne i eksplanacyjne danych

Badania eksploracyjne (identyfikacyjne) służą zrozumieniu określonego problemu, tendencji lub sytuacji, ustaleniu koniecznego zakresu informacji oraz formułowaniu hipotez. Ich cechą charakterystyczną jest brak jasnego celu badań oraz brak jasnego kierunku i sposobu ich prowadzenia. Celem tych badań jest zatem rozpoznanie problemu. Jest to wstępny etap pracy z danymi.

Badania eksplanacyjne (wyjaśniające) zawsze mają określony cel badawczy oraz kierunek ich prowadzenia. Celem tych badan jest wyjaśnienie przyczyn konkretnego problemu. Służą one zasugerowaniu odbiorcy/audytorium najkorzystniejszego rozwiązania problemu. Badania eksplanacyjne nie mogą polegać na pokazaniu audytorium całości analizy danych. Wybór danych musi być zatem dokonany przez pryzmat ściśle określonego problemu oraz celu prezentacji.

Typy wykresów

Typ wykresu, jaki wybierzemy do przedstawienia konkretnych danych zależy od:

- ilości i rodzaju danych;
- celu ogólnego konkretnej prezentacji;
- specyfiki i potrzeb audytorium;
- Formy komunikacji z odbiorcą/widownią.

Typ wykresu ma ułatwiać widowni zrozumienie konkretnych danych oraz wspierać cel samej prezentacji.

Wśród sposobów prezentacji danych wyróżniamy[3]:

[3] Zob. także C. Nussbaumer Knaflic, Storytelling danych. Poradnik wizualizacji danych dla profesjonalistów, Gliwice 2015, s. 49-83

- prosty tekst – najlepiej sprawdza się, gdy w ramach konkretnej prezentacji mamy do przekazania jedną lub dwie liczby. Prosty tekst można z łatwością połączyć ze strukturą storytellingu: przechodząc od ogółu do szczegółu lub od szczegółu do ogółu;
- tabela – warto ją wykorzystać prezentując dane w formie pisemnej w taki sposób, aby każdy z zainteresowanych uczestników mógł samodzielnie sięgnąć do tabeli;
- mapa cieplna – pozwala na wizualizację danych w postaci tabeli, w której zamiast lub obok liczb stosujemy barwne komórki pokazujące względną wartość liczb. W mapie cieplnej warto posługiwać się skalą nasycenia określonego koloru. Podobnie jak tabele, często odnosi się do pisemnej prezentacji danych;
- wykresy punktowe - służą przedstawieniu zależności pomiędzy dwiema zmiennymi, pozwalając na śledzenie danych na osi x i osi y równocześnie. Stosunkowo rzadko pojawiają się w realiach biznesowych.
- wykresy liniowe – najlepiej sprawdzają się przy prezentacji danych ciągłych, ujętych w danej jednostce czasu. wykresy słupkowe i kolumnowe - umożliwiają łatwe dostrzeżenie największej wartości oraz różnicy między kategoriami. Ich ogromną zaletą jest łatwość odczytania słupków lub kolumn. W tej kategorii wyróżniamy wykresy kolumnowe, skumulowane wykresy kolumnowe, wykresy słupkowe, skumulowane wykresy słupkowe i wykresy wodospadowe. Zalecane są zwłaszcza wykresy słupkowe;
- wykresy powierzchniowe - często są trudne do odczytania danych, jednak niekiedy można je stosować dla wizualizacji liczb o znacznie zróżnicowanej wartości.

Jakich typów wykresów warto unikać?

- wykresów kołowych - wiążą się z nimi trudności w odczytaniu i interpretacji danych, gdyż nasze oczy nie są przyzwyczajone do odczytania danych w dwuwymiarowej przestrzeni. Porównywanie kątów i powierzchni może stanowić trudność dla widowni;
- wykresów pierścieniowych - również porównywanie długości poszczególnych łuków nie jest dla widowni łatwe;
- unikajmy stosowania efektów trójwymiarowych do prezentacji danych z jednego wymiaru.

Zasady dotyczące percepcji danych

Zasady dotyczące percepcji wizualnej (zasady gestaltu) umożliwiają nam odróżnienie elementów wzmacniających cel wystąpienia, czyli takich, które chcemy przekazać od tych, które jedynie odwracają uwagę od celu wystąpienia. Zasady te określają sposób reakcji ludzi na bodźce wizualne.

Są to zasady[4]:

I. Zasada bliskości - obiekty znajdujące się w niedalekim sąsiedztwie są postrzegane jako część tej samej grupy. Z kolei obiekty położone dalej od siebie należą do oddzielnych grup;
II. Zasada podobieństwa - obiekty wyróżniające się podobnymi cechami, takimi jak kolor, kształt, ułożenie, rozmiar są postrzegane jako część tej samej grupy.
III. Zamknięcie w przestrzeni – obiekty zamknięte we wspólnej przestrzeni są postrzegane jako należące do jednej grupy;
IV. Domknięcie - postrzegamy grupę pojedynczych elementów jako jeden rozpoznawalny kształt, znany już naszemu umysłowi;
V. Ciągłość - ludzki wzrok szuka najkrótszej drogi między obiektami i dostrzega ciągłość nawet w braku jej wyraźnego zaznaczenia;
VI. Połączenie - fizycznie połączone elementy (np. za pomocą linii) traktujemy jako składniki jednej grupy.

Unikanie nadmiernego obciążenia poznawczego odbiorców

Podstawowym pytaniem jest pytanie o wielkość obciążenia poznawczego z perspektywy widowni: ile wysiłku musi włożyć widownia, aby wydobyć z naszej prezentacji konkretne przesłanie.

Pojawia się zatem pytanie, jak unikać nadmiernego obciążenia poznawczego odbiorców?

Pod kątem storytellingu będzie to:

- Eliminacja wątków niewspierających celu prezentacji;
- Eliminacja bohaterów, którzy nie pełnią w historii istotnej roli;
- Osadzenie historii na jednej linii narracji;
- Przestrzeganie zasady spójności historii z celem wystąpienia i problemem widowni;
- Przestrzeganie zasady, że historia ma służyć wyłącznie jednemu celowi.

[4] Ibidem, s. 85-91

Pod kątem prezentacji danych należy zwrócić uwagę na:

- Wyrównanie tekstu do lewej;
- Możliwość wykorzystania pustej przestrzeni (pełni podobną rolę do pauzy w wystąpieniu publicznym);
- Strategiczne użycie kontrastu.

Miejsce danych w strukturze prezentacji:

Pod kątem zapamiętywalności, dane najlepiej umieszczać w następujących miejscach prezentacji:

- We wstępie jako przykucie uwagi;
- W ramach historii w punkcie kulminacyjnym;
- W ramach historii – w części rezultatów.

Atrybuty przetwarzane mimowolnie i ich rola w przykuciu uwagi:

Atrybuty przetwarzane mimowolnie pozwalają kierować uwagę widowni na elementy, które mają zostać przez nią zapamiętane. Ich rolą jest zatem wzmacnianie celu przemówienia, gdyż przykuwają uwagę widowni na te obszary, które najmocniej wspierają cel. Stosowanie atrybutów podlegających mimowolnemu przetwarzaniu związane są z pamięcią ikoniczną, toteż przyciągają one uwagę naszych odbiorców, zanim jeszcze nawet są oni tego świadomi.

Do grupy tych atrybutów należą[5]:

- kształt,
- kolor,
- długość linii,
- grubość linii,
- orientacja,
- nasycenie,
- rozmiar,
- krzywizna,
- zamknięcie w przestrzeni,
- pozycja w przestrzeni,
- ruch.

[5] Ibidem, s. 112-135

Atrybuty przetwarzane mimowolnie mogą zostać wykorzystane zarówno w tekście, jak i w ramach wykresów.

Do atrybutów podlegających mimowolnemu przetwarzaniu wykorzystywanych w pracy z tekstem należą:

- pogrubienie;
- kursywa;
- kolor;
- podkreślenie;
- zamknięcie w przestrzeni;
- obramowanie.

Stosując elementy przetwarzane mimowolnie w tekście warto stworzyć ich hierarchię wizualną celem wyboru, na które informacje nasza widownia ma zwrócić uwagę w pierwszej kolejności. Należy pamiętać, że stosowanie atrybutów przetwarzanych mimowolnie jest zawsze zależne od celu przemówienia i powinno być stosowane w taki sposób, aby wzmacniać ten cel. Im bardziej dany element jest powiązany z celem przemówienia, tym bardziej należy go uwypuklić za pomocą atrybutów przetwarzanych mimowolnie. Zawsze zadawaj sobie pytanie, gdzie w konkretnym momencie ma znaleźć się uwaga Twoich odbiorców.

Praktyczne uwagi przy stosowaniu atrybutów przetwarzanych mimowolnie podczas tworzenia wykresów:

- Linie stanowiące osie wykresu powinny być cieńsze od serii danych;
- W wykresach liniowych, umiejętnie stosuj etykiety danych. Można ich użyć w tych miejscach, na które Twoja widownia powinna zwrócić świadomą uwagę;
- Elementy tak samo ważne powinny mieć ten sam rozmiar;
- Jeśli chcesz przykuć uwagę na jeden konkretny element, powinien być on większy od pozostałych;
- Korzystaj z koloru w sposób konsekwentny i jedynie po to, aby podkreślić kluczowe informacje;
- W tabelach i mapach cieplnych warto wykorzystywać różne nasycenie jednego koloru zamiast kilku kolorów;
- Zmiana kolorów oznacza zmianę w treści/narracji;
- Przenalizuj nastrój/emocję, z jaką kojarzy się dany kolor i dobierz ją odpowiednio do celu wystąpienia;

- Kluczowe informacje warto umieścić w lewym górnym rogu zgodnie z naturalnym sposobem czytania danych;

Projektowanie wizualizacji danych

Zdaniem Cole Nussbaumer Knaflic istnieją trzy kluczowe czynniki, jakie należy brać pod uwagę przy wizualizacji danych:

- Afordancje,
- Przystępność,
- Estetyka[6].

Afordancje to możliwości interakcji pomiędzy odbiorcami naszej prezentacji danych a danymi, które prezentujemy. Stanowią one wskazówkę, w jaki sposób wykorzystać oraz odczytywać użyte w prezentacji dane.

Trzy płaszczyzny użycia afordancji to:

- Wyróżnianie istotnych elementów;
- Eliminacja elementów odwracających uwagę od celu prezentacji;
- Klarowna hierarchia informacji[7].

1. Warto wyróżniać jedynie te elementy, które są kluczowe z punktu widzenia celu prezentacji danych. Często mówi się o tym, że powinno to być maksymalnie 10% powierzchni danej wizualizacji.

Warto wybierać te atrybuty, które są najbardziej czytelne i w największym stopniu wspierają cel, np. pogrubienie tekstu zamiast jego podkreślenia.

Sposobami na wyróżnienie istotnych informacji są:

- pogrubienie,
- podkreślenie,
- kursywa,
- kolor,
- wielkie litery,
- krój pisma,
- inwersja kolorów tekstu i tła,
- rozmiar.

[6] Ibidem, rozdział 5
[7] Ibidem, s. 138-149

2. Eliminacja elementów odwracających uwagę od celu prezentacji obejmuje zarówno śmieci, czyli elementy znajdujące się w wizualnej prezentacji danych, lecz nic do niej nie wnoszące, jak i kontekst dla prezentowanych danych.

W tym kontekście warto zadawać sobie pytanie, czy dany element rzeczywiście wspiera cel mojej prezentacji oraz, czy dzięki jego dodaniu moja prezentacja rzeczywiście będzie bardziej skuteczna. Dodatkowo warto wykorzystać atrybuty przetwarzane mimowolnie, aby pewne istotne elementy, które jednak nie są bezpośrednio związane z celem, odsunąć na dalszy plan prezentacji.

3. Hierarchia wizualna to zadanie sobie pytania, gdzie najpierw powinna powędrować uwaga naszych odbiorców oraz podkreślenie tych elementów za pomocą atrybutów przetwarzanych mimowolnie.

Przystępność oznacza dopasowanie prezentowanych danych do różnych możliwości percepcyjnych odbiorcy. Oznacza to zatem, że nie tylko cel prezentacji oraz jej treść, lecz także sposób prezentowania danych należy dopasować do widowni docelowej.

W ramach przystępności, najważniejsze jest unikanie komplikacji, a zatem utrzymanie maksymalnej prostoty sposobu prezentowania danych. W tym kontekście warto wziąć pod uwagę następujące wskazówki:

- Używaj łatwych do odczytania czcionek (rozmiar i krój pisma);
- Świadomie korzystaj z afordancji wizualnych;
- Używaj prostego języka;
- Wybieraj prostotę zamiast złożoności[8].

O przystępność warto zadbać także, świadomie używając tekstu opisującego konkretną wizualizację danych. Warto przy tym pamiętać o kilku istotnych wskazówkach:

- Używaj tytułów dla swoich wykresów;
- Używaj tytułów dla poszczególnych osi;
- Stosuj przejrzyste opisy;
- Umieść przesłanie na wizualizacji danych w formie tekstu (dotyczy pisemnych prezentacji danych);
- Ważne punkty na wykresie opisz odpowiednimi adnotacjami;

[8] Ibidm, s. 151

- Niekiedy w lewym górnym rogu warto dodać przesłanie lub wezwanie do działania (dotyczy pisemnych prezentacji danych);
- Pod wykresem dodaj źródło konkretnych danych[9].

W ramach estetyki warto z kolei wyróżnić trzy kluczowe obszary:

- Świadome korzystanie z koloru dla podkreślenia zmiany lub wyróżnienia jakiegoś elementu.
- Wyrównanie tekstu do lewej strony, co jest zgodne ze sposobem, w jaki czytamy. Klarowne linie pionowe i poziome ułatwiają odbiorcom zrozumienie prezentacji.
- Świadomie wykorzystuj pustą przestrzeń i nie dokładaj dodatkowych elementów tylko po to, aby ją zapełnić[10]

[9] Ibidem, s. 152-155
[10] Ibidem, s. 157