*Obraz zawierający tekst, Czcionka, logo, Grafika

Opis wygenerowany automatycznie*

Uniwersytet Ekonomiczny we Wrocławiu

Przedmiot:

*Statystyczna analiza danych*

Raport z projektu końcowego

*Wioletta Bartos 189484*

*Jakub Szulc 192292*

*Informatyka w biznesie*

*Studia II stopnia, niestacjonarne, rok I*

*Rok akademicki 2023/2024*

*Wrocław, 2024*

**Spis treści**

[1. Regresja liniowa 3](#_Toc168053670)

[2. Regresja logistyczna 11](#_Toc168053671)

[3. Analiza skupień 15](#_Toc168053672)

[4. Analiza głównych składowych. 25](#_Toc168053673)

# 1. Regresja liniowa

Do realizacji regresji liniowej zostały wykorzystane dane z Głównego Urzędu Statystycznego (GUS) przedstawiające liczbę osób bezrobotnych, liczbę osób w wieku produkcyjnym oraz ilość wolnych miejsc pracy w latach 2015-2022. Celem analizy jest zbadanie zależności między liczbą osób bezrobotnych a wybranymi zmiennymi objaśniającymi. Analizowane dane obejmują:

**Zmienna objaśniana:**

**y** - Liczba osób bezrobotnych

**Zmienne objaśniające:**

**x1** - Liczba osób w wieku produkcyjnym

**x2** - Ilość wolnych miejsc pracy

1. Pierwszym etapem było załadowanie biblioteki umożliwiającej zaczytanie plików w formcie *.xlsx*.



1. Następnie wczytano dane do programu R studio.



1. Tworzenie wykresu zależności. Wizualizacja zależności liczby osób bezrobotnych od liczby osób w wieku produkcyjnym.



A graph with numbers and lines

Description automatically generated

1. Tworzenie modelu regresji liniowej. Ustalenie relacji liniowej między liczbą osób bezrobotnych a liczbą osób w wieku produkcyjnym.



1. Dodanie linii regresji do wcześniej utworzonego wykresu punktowego, aby zobrazować dopasowanie modelu do danych. Linia regresji pokazuje przewidywaną liczbę osób bezrobotnych w zależności od liczby osób w wieku produkcyjnym



A graph with numbers and a line

Description automatically generated

1. Ocena jakości dopasowania modelu regresji liniowej.



A computer screen shot of a number

Description automatically generated

Model regresji liniowej sugeruje, że liczba osób w wieku produkcyjnym ma istotny wpływ na liczbę osób bezrobotnych, przy czym każdy dodatkowa osoba w wieku produkcyjnym zwiększa liczbę osób bezrobotnych o około 4 osoby. Model wyjaśnia około 70,58% zmienności w danych, co wskazuje na dobry, choć nie doskonały, poziom dopasowania. Estymatory są istotne statystycznie, co wzmacnia zaufanie do wyników modelu.

1. **Tworzenie wykresu zależności liczby osób bezrobotnych od ilości wolnych miejsc pracy.**



**A graph with numbers and lines

Description automatically generated**

1. Utworzenie modelu regresji liniowej dla liczby osób bezrobotnych i ilości wolnych miejsc pracy.

****

1. Dodanie linii regresji do wcześniej utworzonego wykresu punktowego, aby zobrazować dopasowanie modelu do danych.



**A graph with numbers and lines

Description automatically generated**

1. Ocena jakości dopasowania modelu regresji liniowej.

**A screenshot of a computer error

Description automatically generated**

Model regresji liniowej sugeruje, że liczba wolnych miejsc pracy ma istotny wpływ na liczbę osób bezrobotnych, przy czym każdy dodatkowy wolny etat zmniejsza liczbę osób bezrobotnych o około 7 osób. Model wyjaśnia około 69,56% zmienności w danych, co wskazuje na dobry, choć nie doskonały, poziom dopasowania. Model jest statystycznie istotny, co potwierdza niski p-value zarówno dla wyrazu wolnego, jak i zmiennej liczby wolnych miejsc pracy.

1. **Tworzenie modelu regresji liniowej z jedną zmienną objaśnianą i dwoma zmiennymi objaśniającymi**

****

1. **Posumowanie modelu za pomocą funkcji summary.**

**A screenshot of a computer code

Description automatically generated**

***−3,468e+06 + 2,197e-01\*4,556e+00\****

**INTERPRETACJA ESTYMATORÓW**

**Wyraz wolny: 3,468e+06**

Interpretacja: Gdy liczba osób w wieku produkcyjnym i liczba wolnych miejsc pracy wynoszą 0, przewidywana liczba osób bezrobotnych wynosi -3,468e+06. Wartość ta nie ma praktycznego sensu, ale jest niezbędna do oszacowania linii regresji.

* **Błąd standardowy 1,237e+06**

Interpretacja: Wysoki błąd standardowy sugeruje, że estymator wyrazu wolnego może być mniej precyzyjny.

* **t value: -2,803**
* p-**value: 0,03786**

Interpretacja: Wartość t wynosząca -2,803 oznacza, że oszacowany współczynnik dla wyrazu wolnego jest oddalony o 2,803 standardowych błędów od zera.

Ponieważ p (0,03786) jest mniejsze niż 0,05, co oznacza wyraz wolny jest statycznie istotny na poziome 5%.

**Liczba osób w wieku produkcyjnym: 2,197e-01**

Interpretacja: Z każdym dodatkowym tysiącem osób w wieku produkcyjnym, przewidywana liczba osób bezrobotnych wzrasta średnio o około 220 osób przy założeniu, że wszystkie inne czynniki pozostają stałe.

* **Błąd standardowy 5,065e-02**

Interpretacja: Niski błąd standardowy wskazuje na wysoką precyzję oszacowania wpływu liczby osób w wieku produkcyjnym na liczbę osób bezrobotnych.

* **t value: 4,338**
* **p-value: 0,00745**

Interpretacja: Ponieważ p(0,00745) jest mniejsze niż 0,01, zmienna ta jest statystycznie istotna na poziomie 1%. Oznacza to, że liczba osób w wieku produkcyjnym ma istotny wpływ na liczbę osób bezrobotnych.

**Ilość wolnych miejsc pracy: 4,556e+00**

Interpretacja: Z każdym dodatkowym wolnym miejscem pracy, przewidywana liczba osób bezrobotnych spada średnio o około 5 osób, przy założeniu, że wszystkie inne czynniki pozostają stałe.

* **Błąd standardowy 1,073e+00**

Interpretacja: Niski błąd standardowy wskazuje na wysoką precyzję oszacowania wpływu liczby wolnych miejsc pracy na liczbę osób bezrobotnych

* **t value: -4,245**
* **p-value: 0,00813**

Interpretacja: Ponieważ p(0,00813) jest mniejsze niż 0,01, zmienna ta jest statystycznie istotna na poziomie 1%. Oznacza to, że liczba wolnych miejsc pracy ma istotny wpływ na liczbę osób bezrobotnych.

**STATYSTYKI DOPASOWANIA MODELU**

* **Multiple R-squared: 0,9361**

Interpretacja: Model wyjaśnia 93,61% zmienności w liczbie osób bezrobotnych. Oznacza to, że model bardzo dobrze dopasowuje się do danych.

* **Adjusted R-squared: 0,9105**

Interpretacja: Skorygowany współczynnik determinacji, który uwzględnia liczbę zmiennych w modelu. Wartość 0,9105 wskazuje, że model w 91,05% wyjaśnia zmienność danych po uwzględnieniu liczby zmiennych.

* **F-statistic**: 36,61 na 2 i 5 DF

p-value: 0,001033

Interpretacja: **Wartość F-statistic 36.61** wskazuje, że model regresji wyjaśnia znacznie więcej wariancji zmiennej zależnej niż model bez zmiennych niezależnych. Wysoka wartość F-statistic sugeruje, że przynajmniej jedna ze zmiennych niezależnych ma istotny wpływ na zmienną zależną. Dwa stopnie swobody w liczniku (dla liczby zmiennych niezależnych) i pięć stopni swobody w mianowniku (dla reszt).

**Wartość p(0,001033)** jest znacznie mniejsza niż 0.05, co oznacza, że możemy odrzucić hipotezę zerową. Oznacza to, że model jako całość jest statystycznie istotny, a zmienne niezależne mają istotny wpływ na zmienną zależną.

**RESZTY**

* **Wartości reszt: Min (85072), 1Q (-10940), Median (-58169), 3Q (24222), Max (-76218)**

Interpretacja: Reszty to różnice między wartościami obserwowanymi a przewidywanymi przez model.

* **Najmniejsza wartość reszty (85072)** oznacza, że największa pozytywna różnica między rzeczywistą a przewidywaną wartością zmiennej zależnej wynosi 85072. Wartość ta wskazuje na przypadek, w którym model znacznie zaniżył przewidywaną wartość w porównaniu z rzeczywistą wartością.
* **Pierwszy kwartyl (-10940)** oznacza, że 25% reszt jest mniejszych (bardziej ujemnych) niż -10940. To wskazuje, że dolna ćwiartka reszt jest dość blisko przewidywań modelu, ale z tendencją do ujemnych wartości, co oznacza, że model nieco zawyża przewidywane wartości.
* **Mediana (-58169)** wskazuje, że połowa reszt jest mniejsza (bardziej ujemna) niż -58169, a połowa jest większa (mniej ujemna lub dodatnia). Ujemna mediana sugeruje, że model ma tendencję do zawyżania przewidywanych wartości.
* **Trzeci kwartyl (24222)** oznacza, że 75% reszt jest mniejszych (mniej pozytywnych) niż 24222. To oznacza, że większość reszt jest blisko przewidywanych wartości, ale z tendencją do dodatnich wartości, co oznacza, że model nieco zaniża przewidywane wartości.
* **Największa wartość reszty (-76218)** oznacza, że największa ujemna różnica między rzeczywistą a przewidywaną wartością zmiennej zależnej wynosi -76218. Wartość ta wskazuje na przypadek, w którym model znacznie zawyżył przewidywaną wartość w porównaniu do rzeczywistej wartości.

**PODSUMOWANIE**

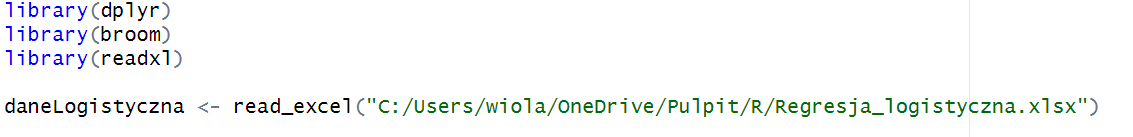
Model regresji liniowej wskazuje, że zarówno liczba osób w wieku produkcyjnym, jak i liczba wolnych miejsc pracy mają istotny wpływ na liczbę osób bezrobotnych. Każdy dodatkowy tysiąc osób w wieku produkcyjnym zwiększa liczbę osób bezrobotnych o około 220 osób, natomiast każdy dodatkowy wolny etat zmniejsza liczbę osób bezrobotnych o około 5 osób. Model wyjaśnia 93,61% zmienności w danych, co wskazuje na bardzo dobry poziom dopasowania. Wszystkie estymatory są istotne statystycznie, co wzmacnia zaufanie do wyników modelu.

# 2. Regresja logistyczna

Celem niniejszej analizy jest zidentyfikowanie czynników wpływających na wysokość opłat za ubezpieczenie zdrowotne. W tym celu wykorzystana zostanie metoda regresji logistycznej, która pozwala na ocenę zależności między zmiennymi niezależnymi a prawdopodobieństwem, że opłaty za ubezpieczenie zdrowotne przekroczą medianę. Dzięki tej metodzie możliwe będzie określenie, które czynniki w największym stopniu determinują wysokość składek ubezpieczeniowych. Analiza opiera się na danych z serwisu Kaggle, dostępnych pod adresem: <https://www.kaggle.com/code/shubhamptrivedi/health-insurance-price-predict-linear-regression>.

**Dane wykorzystane w tej analizie obejmują:**

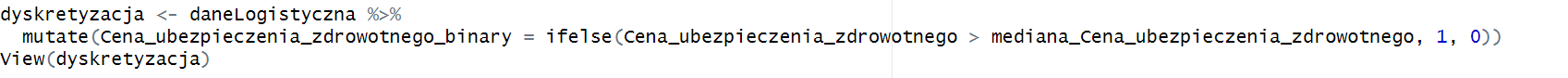
1. **Cena ubezpieczenia zdrowotnego:** Jest to zmienna zależna, która określa wysokość opłat za ubezpieczenie zdrowotne dla każdego ubezpieczonego.
2. **Wiek:** Zmienna niezależna, która wskazuje na wiek ubezpieczonego. Wiek może mieć znaczący wpływ na wysokość składki ubezpieczeniowej ze względu na różne ryzyka zdrowotne związane z wiekiem.
3. **Wskaźnik masy ciała:** Zmienna niezależna, która określa wskaźnik masy ciała ubezpieczonego. BMI jest powszechnie używanym wskaźnikiem, który może wpływać na koszty ubezpieczenia zdrowotnego z uwagi na ryzyka związane z nadwagą i otyłością.
4. Wczytanie bilbliotek i danych.



1. Obliczenie mediany zmiennej ceny ubezpieczenia zdrowotnego.



1. Dyskretyzacja zmiennej ceny ubezpieczenia zdrowotnego. Tworzymy nową zmienną binarną ceny ubezpieczenia zdrowotnego\_binary, która przyjmuje wartość 1, jeśli ceny ubezpieczenia zdrowotnego są większa od mediany, oraz 0 w przeciwnym razie.



1. Przeprowadzenie regresji logistycznej. Przeprowadzamy regresję logistyczną, gdzie zmienną zależną jest ceny ubezpieczenia zdrowotnego \_binary, a zmiennymi objaśniającymi są wiek i wskaźnik masy ciała.



1. Podsumowanie wyników modelu.



A screenshot of a computer code

Description automatically generated

**INTERPRETACJA ESTYMATORÓW**

**Wyraz wolny: -3,891437**

Interpretacja: Logarytm szansy, na to, że zmienna ceny ubezpieczenia zdrowotnego \_binary przyjmie wartość 1 (czyli że opłaty będą powyżej mediany), gdy wszystkie zmienne objaśniające (wiek i BMI) wynoszą 0, jest równy -3,891437.

* **Błąd standardowy: 0,377093**

Interpretacja: Miara niepewności oszacowania parametru wyrazu wolnego. Niska wartość wskazuje na precyzyjne oszacowanie.

* **z value : -10,320**

Interpretacja: Stosunek estymatora do jego błędu standardowego. Wysoka wartość bezwzględna wskazuje na istotność parametru.

* **Pr(>|z|): <2e-16 \*\*\***

Interpretacja: Bardzo niska wartość p sugeruje, że estymator jest istotny statystycznie na poziomie alfa 0,001.

**Wiek: 0,087451**

Interpretacja: Każdy dodatkowy rok życia zwiększa logarytm szansy na to, że ceny ubezpieczenia zdrowotnego będą powyżej mediany, o około 0,087. Wartość ta jest dodatnia, co sugeruje pozytywny związek między wiekiem a wyższymi cenami ubezpieczenia zdrowotnego.

* **Błąd standardowy: 0,005181**

Interpretacja: Niska wartość błędu standardowego wskazuje na precyzyjne oszacowanie tego parametru.

* **z value (Wartość z): 16,879**

Interpretacja: Bardzo wysoka wartość wskazuje na istotność parametru.

* **Pr(>|z|) (Wartość p): <2e-16 \*\*\***

Interpretacja: Bardzo niska wartość p sugeruje, że wiek jest istotnym predyktorem opłat. Możemy z dużą pewnością stwierdzić, że wiek ma istotny wpływ na szanse, że opłaty będą powyżej mediany

**Wskaźnik masy ciała**: **0,015301**

Interpretacja: Każda jednostka wzrostu wskaźnika masy ciała zwiększa logarytm szansy na to, że ceny ubezpieczenia zdrowotnego będą powyżej mediany, o około 0,015. Wartość dodatnia sugeruje pozytywny związek, ale jest mniej wyraźna niż dla wieku.

* **Błąd standardowy: 0,010546**

Interpretacja: Miara niepewności oszacowania tego parametru. Wartość błędu standardowego jest wyższa niż dla wieku, co sugeruje większą niepewność.

* **z value (Wartość z): 1,451**

Interpretacja: : Relatywnie niska wartość z sugeruje, że zmienna wskaźnik masy ciała nie jest istotnym predyktorem w modelu.

**Pr(>|z|) (Wartość p): 0,147**

* Interpretacja: Wartość p jest wyższa niż 0,05, co sugeruje, że wskaźnik masy ciała nie jest istotnym predyktorem ceny ubezpieczenia zdrowotnego na poziomie alfa 0,05.

**Null Deviance 1854.9 na 1337 stopniach swobody.**

Interpretacja: Mierzy, jak dobrze model bez żadnych dopasowuje się do danych. Wysoka wartość sugeruje, że taki model słabo dopasowuje się do danych.

**Residual Deviance: 1468.6 na 1335 stopniach swobody.**

Interpretacja: Mierzy, jak dobrze model z uwzględnionymi predyktorami dopasowuje się do danych. Niższa wartość w porównaniu do null deviance sugeruje, że model z predyktorami lepiej dopasowuje się do danych.

**AIC (Akaike Information Criterion): 1474,6**

Interpretacja: AIC to miara jakości modelu uwzględniająca jego złożoność. Niższa wartość AIC wskazuje na lepszy model. Wartość AIC sama w sobie nie jest interpretowana, ale jest używana do porównywania modeli.

**PODSUMOWANIE**

* Wiek jest istotnym predyktorem, który znacząco wpływa na szanse, że ceny ubezpieczenia zdrowotnego będą powyżej mediany. Każdy dodatkowy rok życia zwiększa te szanse, co potwierdza bardzo niski poziom wartości p.
* Wskaźnik masy ciała nie jest istotnym predyktorem na poziomie istotności alfa 0,05, co sugeruje, że jego wpływ na opłaty nie jest statystycznie znaczący.
* Estymatory mają niskie błędy standardowe, co oznacza, że są one precyzyjne.

# 3. Analiza skupień

Do analizy skupień zostały wykorzystane dane z GUS, które przedstawiają wartości dofinansowań unijnych dla polskich rolników. Dane zostały podzielone na wszystkie województwa Polski, dzięki czemu uzyskano 16 obiektów do analizy. Zmiennymi są: Wsparcie dla nowych użytkowników systemów jakości, Modernizacja gospodarstw rolnych, Inwestycje w gospodarstwach położonych na obszarach Natura 2000, Przetwórstwo i marketing produktów rolnych, Budowa lub modernizacja dróg lokalnych. Dane pochodzą z roku 2021. Celem analizy jest identyfikacja grup województw o podobnych wzorcach finansowania, co umożliwi bardziej efektywne planowanie i alokację środków unijnych oraz wsparcie polityki rolnej.

**1**. Pierwszym krokiem do przeprowadzenia analizy skupień jest zainstalowanie oraz zaczytanie potrzebnych bibliotek.

Obraz zawierający tekst, Czcionka, design

Opis wygenerowany automatycznie

library(readxl):

* Odpowiada za importowanie danych z plików Excel do R. Umożliwia łatwe wczytywanie arkuszy z plików .xls i .xlsx.

library(factoextra):

* Odpowiada za wizualizację wyników analiz statystycznych, w tym analizy głównych składowych (PCA) i analizy skupień. Umożliwia tworzenie atrakcyjnych i informacyjnych wykresów.

library(cluster):

* Odpowiada za wykonywanie różnych metod analizy skupień (klasteryzacji). Zawiera funkcje do takich metod jak k-means, aglomeracyjna analiza skupień oraz analiza siluetowa.

library(NbClust):

* Odpowiada za określanie optymalnej liczby klastrów w analizie skupień. Umożliwia ocenę i porównanie wyników różnych metod klasteryzacji, aby znaleźć najlepszą liczbę klastrów dla danego zestawu danych.

**2.** Następnie zaczytano plik excela do R.



**3.** Przed klasteryzacja należy dokonać okrojenia naszych danych, ponieważ w pierwszej kolumnie znajdują się nazwy województw



**4.** Przed klasteryzacją należy dokonać przeskalowania danych za pomocą funkcji scale(). Funkcja scale() w R przekształca zmienne tak, aby miały średnią równą 0 i odchylenie standardowe równe 1, co zapewnia, że wszystkie zmienne mają porównywalną skalę. Dzięki temu żadna zmienna nie dominuje w analizie skupień, co poprawia jakość klasteryzacji i prowadzi do bardziej wiarygodnych oraz interpretowalnych wyników, ponieważ miary odległości będą bardziej reprezentatywne dla rzeczywistych różnic między danymi.



**5**. Po przeskalowaniu danych można przejść do zbadania odpowiedniej liczby klastrów poprzez przeprowadzenie klasteryzacji. W tym projekcie użyto oraz zbadano dwie metody. Jedna to Metoda łokcia, a druga to Metoda Davies-Bouldin Index.

Metoda Łokcia:



Obraz zawierający tekst, linia, Wykres, zrzut ekranu

Opis wygenerowany automatycznie

Analizując powstały wykres na podstawie naszych danych określono odpowiednią liczbę klastrów. W tym przypadku są to 4 klastry. Metoda łokcia pomaga zidentyfikować punkt, w którym dalsze zwiększanie liczby klastrów przestaje przynosić istotne korzyści w redukcji wewnątrzgrupowej wariancji, sugerując optymalną liczbę klastrów do użycia w analizie skupień.

Metoda Davies-Bouldin Index:

Obraz zawierający tekst, Czcionka, numer, linia

Opis wygenerowany automatycznie

Obraz zawierający tekst, diagram, linia, Wykres

Opis wygenerowany automatycznie

Indeks Daviesa-Bouldina mierzy jakość klasteryzacji poprzez ocenę wewnętrznej spójności klastrów oraz odległości międzyklastrowych. Niższe wartości DBI wskazują na lepiej wyodrębnione i bardziej spójne klastry, co sugeruje lepszą jakość klasteryzacji. W tym przypadku najniższa wartość przypada na 10 klastrów, ale jest to graniczna wartość, zatem przyjęto drugą najniższą, którą jest 6 klastrów.

**6.** W następnym kroku przeprowadzono klasteryzację na podstawie wybranej ilości klastrów dla obu metod – 4 klastry i 6 klastrów. W obu przypadkach użyto tego samego kodu, tylko podmieniono liczby w *centers*.

Analiza 4 klastrów:

Obraz zawierający tekst, Czcionka, zrzut ekranu, linia

Opis wygenerowany automatycznie

Obraz zawierający tekst, zrzut ekranu, Czcionka, numer

Opis wygenerowany automatycznie

Wyniki klasteryzacji metodą k-średnich dla 4 klastrów pokazują, że województwa w klastrze 1 mają bardzo wysokie wartości modernizacji gospodarstw rolnych oraz wysokie wartości inwestycji w gospodarstwach położonych na obszarach Natura 2000. Województwa w klastrze 2 charakteryzują się umiarkowanymi wartościami wsparcia i modernizacji, ale wyższymi wartościami przetwórstwa i marketingu produktów rolnych, podczas gdy województwa w klastrze 3 mają niższe wartości dla wszystkich zmiennych. Województwa w klastrze 4 wyróżniają się najwyższymi wartościami wsparcia dla nowych uczestników systemów jakości oraz wysokimi wartościami we wszystkich pozostałych zmiennych.

Obraz zawierający tekst, Czcionka, zrzut ekranu, linia

Opis wygenerowany automatycznie

Wartości sumy kwadratów wewnątrz klastrów pokazują, że klastry 1 i 4 mają zerową sumę kwadratów, co oznacza, że wszystkie punkty w tych klastrach są identyczne lub zawierają tylko jeden punkt (w tym przypadku posiada tylko jeden punkt). Klastry 2 i 3 mają odpowiednio wartości 10.237733 i 5.774955, wskazując na większe zróżnicowanie wewnątrz tych klastrów. Całkowity procent wyjaśnionej wariancji między klastrami wynosi 78.6%, co oznacza, że większość wariancji w danych jest wyjaśniona przez podział na klastry.

Obraz zawierający tekst, Czcionka, zrzut ekranu, numer

Opis wygenerowany automatycznie

totss (całkowita suma kwadratów):

* Wartość: 75
* Całkowita suma kwadratów odchyleń wszystkich punktów od globalnej średniej. Jest to miara całkowitej wariancji w danych przed klasteryzacją.

withinss (suma kwadratów wewnątrz klastrów):

* Wartości: [0, 10.24, 5.77, 0]
* Suma kwadratów odchyleń punktów w poszczególnych klastrach od ich średnich. Klaster 1 i 4 mają zerową wartość, co sugeruje brak zróżnicowania wewnątrz tych klastrów (mogą zawierać tylko jeden punkt). Klaster 2 i 3 mają odpowiednio 10.24 i 5.77, co wskazuje na większe zróżnicowanie wewnątrz tych klastrów.

tot.withinss (całkowita suma kwadratów wewnątrz klastrów):

* Wartość: 16
* Suma wszystkich withinss. Jest to całkowita suma kwadratów odchyleń wewnątrz wszystkich klastrów, wskazująca na ogólne zróżnicowanie w klastrach po podziale.

betweenss (suma kwadratów między klastrami):

* Wartość: 59
* Suma kwadratów odchyleń między centroidami klastrów a globalną średnią. Wysoka wartość betweenss w porównaniu do totss wskazuje, że znaczna część całkowitej wariancji w danych jest wyjaśniona przez podział na klastry.

size (rozmiary klastrów):

* Wartości: [1, 5, 9, 1]
* Liczba punktów w każdym klastrze. Klastry 1 i 4 zawierają po jednym punkcie, podczas gdy klastry 2 i 3 zawierają odpowiednio 5 i 9 punktów.

iter (liczba iteracji):

* Wartość: 2
* Liczba iteracji potrzebnych do zbieżności algorytmu k-średnich. W tym przypadku algorytm zbiegał się po dwóch iteracjach.

ifault (kod błędu):

* Wartość: 0
* Kod błędu zwracany przez funkcję kmeans. Wartość 0 wskazuje, że algorytm zakończył się sukcesem bez błędów.

**Obraz zawierający tekst, zrzut ekranu, diagram, Wykres

Opis wygenerowany automatycznieWykresy dla 4 klastrów**:

Obraz zawierający tekst, zrzut ekranu, żółty, Wielobarwność

Opis wygenerowany automatycznie

Wykres pierwszy przedstawia wyniki analizy klasteryzacji dla województw z podziałem na 4 klastry, gdzie województwa w poszczególnych klastrach grupują się razem, co oznacza, że mają podobne wartości dofinansowań unijnych. Województwa takie jak Łódzkie, Lubelskie i Wielkopolskie znajdują się w klastrze 1, Podlaskie w klastrze 2, Dolnośląskie, Opolskie i Śląskie w klastrze 3, a Mazowieckie w klastrze 4, co pokazuje ich zróżnicowanie. Wykres drugi, będący mapą cieplną, potwierdza te wyniki, pokazując, że województwa w tych samych klastrach mają podobne wartości, a Mazowieckie wyraźnie różni się od innych województw.

Obraz zawierający tekst, zrzut ekranu, Czcionka, linia

Opis wygenerowany automatycznieTen sam proces przeprowadzono dla 6 klastrów:

Obraz zawierający tekst, zrzut ekranu, Czcionka, numer

Opis wygenerowany automatycznie

Wyniki klasteryzacji dla 6 klastrów pokazują, że województwo w klastrze 1 charakteryzuje się bardzo wysokimi wartościami wsparcia dla nowych uczestników systemów jakości oraz umiarkowanymi wartościami modernizacji gospodarstw rolnych i inwestycji w gospodarstwach położonych na obszarach Natura 2000. Klaster 2, który zawiera trzy województwa, wyróżnia się niskimi wartościami wsparcia dla nowych uczestników systemów jakości, modernizacji gospodarstw rolnych oraz inwestycji w gospodarstwach położonych na obszarach Natura 2000. Klaster 3 ma umiarkowane wartości wsparcia dla nowych uczestników systemów jakości i różne wartości pozostałych zmiennych. Klastry 4, 5 i 6 mają różne kombinacje wartości, z klastrem 6, zawierającym województwo Podlaskie, wyróżniającym się wysokimi wartościami wsparcia dla nowych uczestników systemów jakości oraz umiarkowanymi wartościami pozostałych zmiennych.

Obraz zawierający tekst, Czcionka, zrzut ekranu, linia

Opis wygenerowany automatycznie

Wyniki klasteryzacji pokazują sumy kwadratów odchyleń w obrębie klastrów, co jest miarą wewnętrznej spójności każdego klastra. Konkretne wartości wynoszą: 0.0000 dla klastrów 1 i 6, 1.6575 dla klastra 2, 0.9296 dla klastra 3, 3.6042 dla klastra 4 oraz 1.1645 dla klastra 5. Niskie wartości wskazują na większą jednorodność w obrębie klastra, podczas gdy wyższe wartości sugerują większą zmienność wewnątrz klastra. Wskaźnik między sumą kwadratów odchyleń między klastrami a całkowitą sumą kwadratów odchyleń wynoszący 90.2% oznacza, że 90.2% całkowitej zmienności w danych jest wyjaśniane przez różnice między klastrami, co świadczy o dobrej jakości klasteryzacji.

Obraz zawierający tekst, Czcionka, zrzut ekranu, numer

Opis wygenerowany automatycznie

Wyniki prezentują się następująco, interpretacja wykonana analogicznie jak dla mniejszej liczby klastrów. Jednym zdaniem można stwierdzić, że wyniki wzrosły co znaczy że klastry są lepiej dopasowane.

**Obraz zawierający tekst, zrzut ekranu, diagram, Wykres

Opis wygenerowany automatycznieWykresy dla 6 klastrów**

Obraz zawierający tekst, zrzut ekranu, żółty, Wielobarwność

Opis wygenerowany automatycznie

Wykres pierwszy przedstawia wyniki analizy klasteryzacji dla województw z podziałem na 6 klastrów, wykorzystując dwie główne składowe (Dim1 i Dim2), gdzie województwa w poszczególnych klastrach grupują się razem, co oznacza, że mają podobne wartości dofinansowań unijnych. Klaster 1 obejmuje województwo Mazowieckie, klaster 2 zawiera województwo Podlaskie, klaster 3 Łódzkie, klaster 4 Dolnośląskie, Opolskie i Śląskie, klaster 5 Wielkopolskie i Lubelskie, a klaster 6 Pomorskie i Warmińsko-Mazurskie. Mapa cieplna potwierdza te wyniki, pokazując, że województwa w tych samych klastrach mają podobne wartości, a Mazowieckie i Podlaskie wyraźnie różnią się od innych województw.

**PODSUMOWANIE**

Porównując wyniki klasteryzacji na 4 klastry z wynikami na 6 klastrów, zauważamy, że podział na 6 klastrów prowadzi do bardziej szczegółowej segmentacji województw. W podziale na 4 klastry, niektóre grupy województw były bardziej zbite, natomiast podział na 6 klastrów pozwolił na wyodrębnienie dodatkowych grup województw o specyficznych wartościach dofinansowań. Na przykład, województwo Podlaskie i Mazowieckie, które były oddzielnymi klastrami w podziale na 6 klastrów, zostały przypisane do innych klastrów w podziale na 4 klastry. To sugeruje, że bardziej szczegółowy podział może lepiej odzwierciedlać zróżnicowanie dofinansowań unijnych w poszczególnych województwach.

# 4. Analiza głównych składowych.

Użyte Dane z Głównego Urzędu Statystycznego (GUS) przedstawiają sytuację na rynku pracy w 16 województwach Polski, obejmując liczby osób bezrobotnych, wolnych miejsc pracy, osób w wieku produkcyjnym, absolwentów oraz zlikwidowanych miejsc pracy. Celem analizy głównych składowych (PCA) jest redukcja wymiarowości tych danych, co umożliwia uproszczenie struktury i łatwiejszą interpretację, przy jednoczesnym zachowaniu jak największej ilości informacji. PCA pozwoli zidentyfikować kluczowe czynniki wpływające na sytuację na rynku pracy oraz wykryć wzorce i podobieństwa między województwami, co jest użyteczne dla decydentów politycznych i ekonomicznych w planowaniu działań i alokacji zasobów.

**1**. Pierwszym krokiem do przeprowadzenia analizy głównych składowych jest zainstalowanie oraz zaczytanie potrzebnych bibliotek.

Obraz zawierający tekst, Czcionka, biały, design

Opis wygenerowany automatycznie

library(readxl):

* Importuje dane z plików Excel (.xls i .xlsx) do R.

library(ggplot2):

* Tworzy zaawansowane wykresy i wizualizacje danych w R.

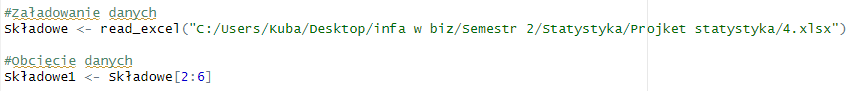
library(psych):

* Przeprowadza analizy psychometryczne i statystyczne, w tym analizę czynnikową i korelacje.

library(clusterSim):

* Służy do klasteryzacji i normalizacji danych oraz oceny jakości klastrów.

**2.** Następnym krokiem było załadowanie danych do R oraz tak jak w poprzedniej analizie obcięcie ich, ponieważ ponownie w pierwszej kolumnie znajdują się wypisane województwa.



**3.** Przeprowadzenie normalizacji zaczytanych danych.

Normalizacja przed analizą głównych składowych (PCA) jest kluczowa, ponieważ dane mogą pochodzić z różnych źródeł i być mierzone w różnych jednostkach, co sprawia, że zmienne mają różne skale. Na przykład, liczba bezrobotnych może być w tysiącach, podczas gdy liczba wolnych miejsc pracy może być w setkach. Normalizacja przekształca wszystkie zmienne do porównywalnej skali, zwykle ze średnią 0 i odchyleniem standardowym 1, co zapobiega dominacji zmiennych o większej skali lub wariancji nad wynikami PCA. Bez normalizacji zmienne o dużej wariancji mogłyby zdominować wyniki, ponieważ PCA identyfikuje główne składowe maksymalizując wariancję. Dzięki normalizacji każda zmienna ma równy wpływ na wyniki analizy, co pozwala na bardziej rzetelne i interpretowalne wyniki.

Obraz zawierający tekst, Czcionka, zrzut ekranu, linia

Opis wygenerowany automatycznie

**4**. Przeprowadzenie analizy głównych składowych (PCA) na zbiorze danych "Składowe1", normalizując i centralizując dane przed analizą, co pozwala na porównywanie zmiennych na tej samej skali. Funkcja prcomp identyfikuje główne składowe, które wyjaśniają największą część wariancji w danych, redukując wymiarowość przy zachowaniu kluczowych informacji. Wyświetlenie wyników za pomocą print(pca) dostarcza informacji o komponentach głównych, ich wariancji i współczynnikach, co ułatwia interpretację i wizualizację danych.

Obraz zawierający tekst, Czcionka, zrzut ekranu, linia

Opis wygenerowany automatycznie

Obraz zawierający tekst, zrzut ekranu, Czcionka, numer

Opis wygenerowany automatycznie

Wyniki analizy głównych składowych (PCA) pokazują, że liczba osób bezrobotnych jest głównie wyjaśniana przez drugą główną składową (PC2) z ładunkiem 0.8785, natomiast ilość wolnych miejsc pracy ma najwyższy ładunek w pierwszej głównej składowej (PC1) z wartością 0.4482. Liczba osób w wieku produkcyjnym jest najbardziej wyjaśniana przez pierwszą główną składową (PC1) z ładunkiem 0.4573, a liczba absolwentów ma najwyższy ładunek również w pierwszej głównej składowej (PC1) z wartością 0.5443. Zlikwidowane miejsca pracy mają najwyższy ładunek w pierwszej głównej składowej (PC1) z wartością 0.4605.

Podsumowanie PCA:

Obraz zawierający tekst, Czcionka, zrzut ekranu, linia

Opis wygenerowany automatycznie

Wyniki analizy głównych składowych (PCA) pokazują, że pierwsza główna składowa (PC1) wyjaśnia 90.8% całkowitej wariancji w danych, druga składowa (PC2) wyjaśnia 5.6%, a trzecia składowa (PC3) 1.85%. Łącznie, pierwsze dwie składowe wyjaśniają 96.4% wariancji, a wszystkie pięć składowych razem wyjaśniają 100% całkowitej wariancji. Oznacza to, że większość informacji zawartej w danych można zredukować do dwóch głównych składowych bez znaczącej utraty informacji.

Obraz zawierający tekst, zrzut ekranu, Czcionka, numer

Opis wygenerowany automatycznie**5.** Analiza wykresów.

Obraz zawierający diagram, linia, tekst, Wykres

Opis wygenerowany automatycznie

Obraz zawierający tekst, zrzut ekranu, diagram, Prostokąt

Opis wygenerowany automatycznie

Obraz zawierający tekst, zrzut ekranu, diagram, Prostokąt

Opis wygenerowany automatycznie

Obraz zawierający tekst, diagram, zrzut ekranu, linia

Opis wygenerowany automatycznie

Pierwszy wykres, zwany "Wykresem osypiska", pokazuje wartości wariancji dla kolejnych głównych składowych. Wykres ten pomaga zidentyfikować liczbę głównych składowych, które warto uwzględnić, zazwyczaj poprzez znalezienie punktu na wykresie, po którym wariancja znacznie spada.

Drugi wykres przedstawia wariancję wyjaśnianą przez każdą główną składową w procentach. Widać, że pierwsza składowa wyjaśnia większość wariancji w danych, co jest zgodne z poprzednim wykresem.

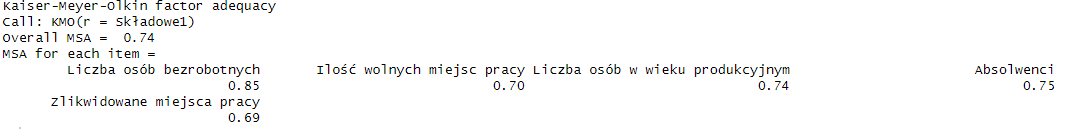
Trzeci wykres to wartości własne głównych składowych, które również wskazują, jak wiele informacji (wariancji) zawiera każda ze składowych.Czwarty wykres, biplot, pokazuje zarówno obiekty (województwa) jak i zmienne w przestrzeni dwóch pierwszych głównych składowych. Umożliwia to wizualizację relacji między obiektami oraz jak zmienne wpływają na te relacje.

**6**. Obliczenie współczynnika KMO.

Współczynnik KMO (Kaiser-Meyer-Olkin) oblicza się w celu oceny adekwatności danych do analizy czynnikowej lub analizy głównych składowych (PCA). Odpowiada on za mierzenie, w jakim stopniu zmienne w zbiorze danych mają wspólne czynniki. Wartości KMO wahają się od 0 do 1, gdzie wartości bliższe 1 wskazują, że zmienne są dobrze skorelowane i odpowiednie do analizy czynnikowej. Współczynnik KMO powyżej 0.6 jest uważany za akceptowalny, powyżej 0.7 za dobry, powyżej 0.8 za bardzo dobry, a powyżej 0.9 za doskonały. Wartości poniżej 0.6 sugerują, że dane mogą nie być odpowiednie do analizy czynnikowej.

Obraz zawierający tekst, Czcionka, typografia

Opis wygenerowany automatycznie



Wyniki współczynnika KMO (Kaiser-Meyer-Olkin) pokazują, że ogólny wskaźnik MSA (Measure of Sampling Adequacy) wynosi 0.74, co oznacza, że dane są odpowiednie do analizy czynnikowej. Szczegółowe wskaźniki MSA dla poszczególnych zmiennych wskazują, że zmienne takie jak liczba osób bezrobotnych (0.85), liczba absolwentów (0.75), liczba osób w wieku produkcyjnym (0.74) oraz ilość wolnych miejsc pracy (0.70) są dobrze skorelowane i odpowiednie do analizy. Jedynie zmienna dotycząca zlikwidowanych miejsc pracy ma niższy wskaźnik MSA (0.69), ale nadal jest akceptowalny.

**PODSUMOWANIE**

Analiza głównych składowych (PCA) na danych z GUS dotyczących rynku pracy w 16 województwach Polski zredukowała wymiarowość danych, zachowując kluczowe informacje. Pierwsza główna składowa (PC1) wyjaśnia 90.8% całkowitej wariancji, a druga (PC2) 5.6%, co oznacza, że łącznie dwie składowe wyjaśniają 96.4% wariancji. Wizualizacje, takie jak wykres osypiska i biplot, pomogły zidentyfikować najważniejsze składowe, a współczynnik KMO 0.74 potwierdził, że dane są odpowiednie do analizy czynnikowej.