Obraz zawierający tekst, urządzenie, wskaźnik

Opis wygenerowany automatycznie

Analiza zamożności wybranych krajów UE

Mateusz Nowak, Damian Okoń, Robert Zamiar

Analityka Biznesowa, 2021/2022, Semestr IV

# Spis treści

[Spis treści 2](#_Toc106439729)

[Wstęp 4](#_Toc106439730)

[1 Metody analizy 4](#_Toc106439731)

[1.1 Opis zbioru danych 4](#_Toc106439732)

[1.1.1 Zmienność 5](#_Toc106439733)

[1.1.2 Korelacje i ich istotność 6](#_Toc106439734)

[1.1.3 Wstępna analiza zmiennych kategorycznych 9](#_Toc106439735)

[1.1.4 Dodatkowa analiza 10](#_Toc106439736)

[1.2 Opis metod analizy 11](#_Toc106439737)

[2 Wyniki 11](#_Toc106439738)

[2.1 Odpowiedź na pierwsze pytanie badawcze. (Do jakiej minimalnej ilości wymiarów można zredukować predyktory, jednocześnie zachowując wysoki procent wyjaśnianej wariancji tych predyktorów? Co opisują te wymiary?) 11](#_Toc106439739)

[2.1.1 Przygotowania do analizy 11](#_Toc106439740)

[2.1.2 Właściwa analiza 12](#_Toc106439741)

[2.1.3 Podsumowanie 15](#_Toc106439742)

[2.2 Odpowiedź na drugie pytanie badawcze (W jakie skupiska można połączyć kraje, przy użyciu dwóch różnych metod? Czy skupiska są takie same?) 15](#_Toc106439743)

[2.2.1 Grupowanie hierarchiczne 15](#_Toc106439744)

[2.2.2 Grupowanie k-means 16](#_Toc106439745)

[2.2.3 Porównanie grupowań 18](#_Toc106439746)

[2.2.4 Podsumowanie 19](#_Toc106439747)

[2.3 Odpowiedź na trzecie pytanie badawcze (Jak wygląda model regresji liniowej wykonany na wszystkich predyktorach? Czy model przeszedł pozytywnie weryfikacje (testy autokorelacji, heteroskedastyczności i normalności reszt)? Jakie jest dopasowanie modelu i jego interpretacja?) 19](#_Toc106439748)

[2.3.1 Wstępna analiza zmiennej zależnej – korelacje 20](#_Toc106439749)

[2.3.2 Regresja – budowa modelu 20](#_Toc106439750)

[2.3.3 Weryfikacja modelu 21](#_Toc106439751)

[2.3.4 Podsumowanie 22](#_Toc106439752)

[2.4 Odpowiedź na czwarte pytanie badawcze (Czy korzystając z nowych zmiennych powstałych przy pomocy przeprowadzonej analizy PCA, uproszczony model regresji liniowej ma dużo gorsze dopasowanie? Czy ten model nadaje się do predykcji?) 22](#_Toc106439753)

[2.4.1 Analiza modelu regresji liniowej z jednym wymiarem 23](#_Toc106439754)

[2.4.2 Analiza modelu regresji liniowej z dwoma wymiarami 23](#_Toc106439755)

[2.4.3 Statystyczne porównanie modeli 24](#_Toc106439756)

[2.4.4 Podsumowanie 24](#_Toc106439757)

[2.5 Odpowiedź na piąte pytanie badawcze (Jak modele klasyfikacyjne radzą sobie z rozpoznawaniem krajów pod względem posiadanej waluty na podstawie danych makroekonomicznych?) 24](#_Toc106439758)

[2.5.1 Podstawowe modele klasyfikacyjne 24](#_Toc106439759)

[2.5.2 Modele usprawniające klasyfikację 26](#_Toc106439760)

[2.5.3 Podsumowanie 28](#_Toc106439761)

[3 Wnioski 28](#_Toc106439762)

[Załącznik 1: Wkład członków zespołu projektowego 29](#_Toc106439763)

[Załącznik 2: Poprawki pierwotnych wersji analiz 29](#_Toc106439764)

[Załącznik 3: Zbiór danych 29](#_Toc106439765)

[Spis tabel i rysunków 30](#_Toc106439766)

# Wstęp

Głównym celem projektu jest zbadanie wpływu różnych zmiennych opisujących zjawiska ekonomiczne oraz społeczne, na ogólny poziom zamożności danych krajów, a także zbudowanie dobrego modelu predykcyjnego, pozwalającego na przewidywanie wartości PKB per capita kraju, po podaniu odpowiednich danych wejściowych. W związku z tym, w trakcie projektu wykonano wiele analiz, wykorzystując metody uczenia nienadzorowanego i nadzorowanego, które pozwoliły odpowiedzieć na następujące pytania badawcze:

1. Do jakiej minimalnej ilości wymiarów można zredukować predyktory, jednocześnie zachowując wysoki procent wyjaśnianej wariancji tych predyktorów? Co opisują te wymiary?
2. W jakie skupiska można połączyć kraje, przy użyciu dwóch różnych metod? Czy skupiska są takie same?
3. Jak wygląda model regresji liniowej wykonany na wszystkich predyktorach? Czy model przeszedł pozytywnie weryfikacje (testy autokorelacji, heteroskedastyczności i normalności reszt)? Jakie jest dopasowanie modelu i jego interpretacja?
4. Czy korzystając z nowych zmiennych powstałych przy pomocy przeprowadzonej analizy PCA, uproszczony model regresji liniowej ma dużo gorsze dopasowanie? Czy ten model nadaje się do predykcji?
5. Jak modele klasyfikacyjne radzą sobie z rozpoznawaniem krajów pod względem posiadanej waluty na podstawie danych makroekonomicznych?

# Metody analizy

## Opis zbioru danych

Etykiety z nazwami krajów (nazwy w języku angielskim), a także zmienne numeryczne zaczerpnięto z serwisu Eurostat, natomiast zmienne kategoryczne zostały utworzone przez zespół na podstawie źródeł z serwisów geopolitycznych. Zmienne numeryczne podane są w ujęciu procentowym na 1000 mieszkańców, bądź per capita. Zastosowane formuły przy łączeniu zbiorów można podejrzeć w arkuszu dane. Jedyną wartością brakującą w zbiorze była ilość lekarzy dla Czech. Brak ten w finalnym zbiorze uzupełniono obliczając średnią wszystkich wartości wskaźnika liczba lekarzy na 1000 mieszkańców.

Tabela 1. Zmienne wykorzystane w projekcie

|  |  |  |
| --- | --- | --- |
| zmienna | wyjaśnienie | rodzaj zmiennej |
| country | kraj | etykieta |
| GDP\_pc | Produkt krajowy brutto per capita w euro | numeryczna |
| high\_tech\_trade\_pc | Handel wysokimi technologiami per capita w euro | numeryczna |
| r&d\_gdp\_pct | Wydatki na badania i rozwój jako procent PKB | numeryczna |
| r&d\_bud\_pct | Wydatki na badania i rozwój jako procent budżetu | numeryczna |
| use\_cloud\_pct | Procent ludzi korzystających z chmury | numeryczna |
| weeknd\_work\_pct | Procent ludzi pracujących w weekendy (15-64) | numeryczna |
| emp\_deadline\_pct | Procent ludzi pracujących intensywnie i z deadlinem (15-64) | numeryczna |
| working\_pop\_pct | Procent populacji w wieku produkcyjnym (15-64) | numeryczna |
| phycisians\_per\_1000 | lekarze na 1000 mieszkańców | numeryczna |
| sea\_access | Czy kraj posiada dostęp do morza | binarna (1 – tak, 0 – nie) |
| joined\_EU | dekada dołączenia do Unii Europejskiej | zmienna kategoryczna (Founding member (1957), 1970s, 1980s, 1990s, 2000s, 2010s) |
| is\_euro\_currency | Czy walutą jest euro? | binarna (1 – tak, 0 – nie) |
| nuclear\_electricity | Czy kraj posiada elektorwnie atomowe? | binarna (1 – tak, 0 – nie) |

### Zmienność

Przygotowane przez nas dane są przekrojowe – pochodzą z roku 2015, więc nie są szeregiem czasowym. Z powodu przyszłego wykorzystania danych do analizy regresji, dla zmiennych numerycznych obliczono współczynnik zmienności oraz odchylenie standardowe. Dla każdej etykiety zmiennych kategorycznych obliczono liczebność. Tabela numer 2 przedstawia wyniki obliczeń.

Tabela 2. Zmienność i odchylenie standardowe zmiennych numerycznych

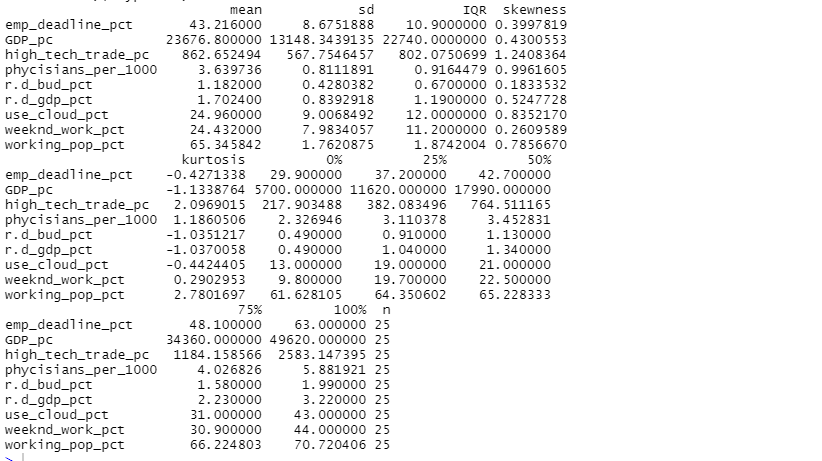
|  |  |  |  |
| --- | --- | --- | --- |
| zmienna | średnia | Odchylenie standardowe | Współczynnik zmienności |
| GDP\_pc | 23677 | 13148,34 | 55,53% |
| high\_tech\_trade\_pc | 863 | 567,75 | 65,81% |
| r&d\_gdp\_pct | 2 | 0,84 | 49,30% |
| r&d\_bud\_pct | 1 | 0,43 | 36,21% |
| use\_cloud\_pct | 25 | 9,00 | 36,09% |
| weeknd\_work\_pct | 24 | 7,98 | 32,68% |
| emp\_deadline\_pct | 43 | 8,68 | 20,07% |
| working\_pop\_pct | 65 | 1,76 | 2,70% |
| phycisians\_per\_1000 | 4 | 0,81 | 22,29% |

Możemy zauważyć, że największą zmiennością charakteryzują się zmienne związane z handlem wysokimi technologiami, a także PKB per capita. Zmienne te posiadają dosyć wysokie odchylenie standardowe z powodu zróżnicowania tych wartości w poszczególnych krajach Unii Europejskiej.

Tabela 3. Charakterystyka zmiennych kategorycznych

|  |  |  |
| --- | --- | --- |
| zmienna | etykieta | liczebność |
| sea\_access | 0  1 | 4  21 |
| nuclear\_electricity | 0  1 | 11  14 |
| is\_euro\_currency | 0  1 | 9  16 |
| joined\_EU | Founding member  1970s  1980s  1990s  2000s  2010s | 5  3  3  3  10  1 |

W przypadku zmiennych kategorycznych zdecydowana większość państw posiada dostęp do morza. Czternaście państw posiada elektrownie jądrowe, a szesnaście państw używa waluty euro. Porównując liczebności państw ze względu na dekadę dołączenia do Unii Europejskiej, zdecydowana większość państw dołączyła w latach 2000 – 2010. Ponadto przy wykorzystaniu dodatku Rcmdr do programu Rstudio obliczono statystyki jak skośność, kurtoza oraz kwantyle zmiennych numerycznych.

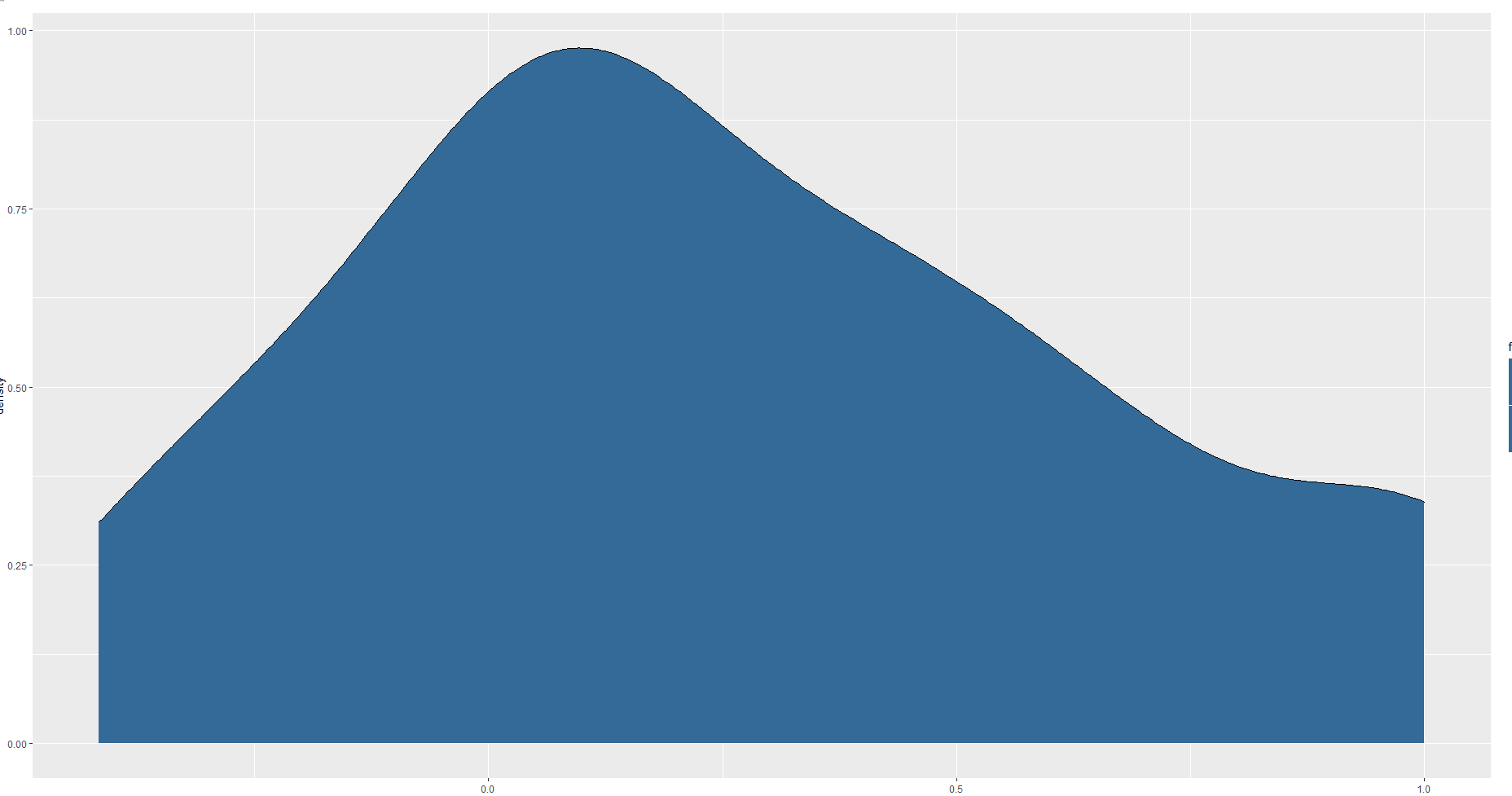


Rysunek 1. Skośność, kurtoza i kwantyle

Ze statystyk opisowych, widocznych na rysunku numer 1, możemy dowiedzieć się, że zmienne posiadają asymetryczny rozkład, z czego większość posiada prawostronny. Z kolei wyniki kurtozy informują nas, że w większości badanych przez nas zmiennych rozkład jest mniej wysmukły niż rozkład normalny (k<0) – istnieje mniej dodatnich wartości niż w przypadku rozkładu normalnego.

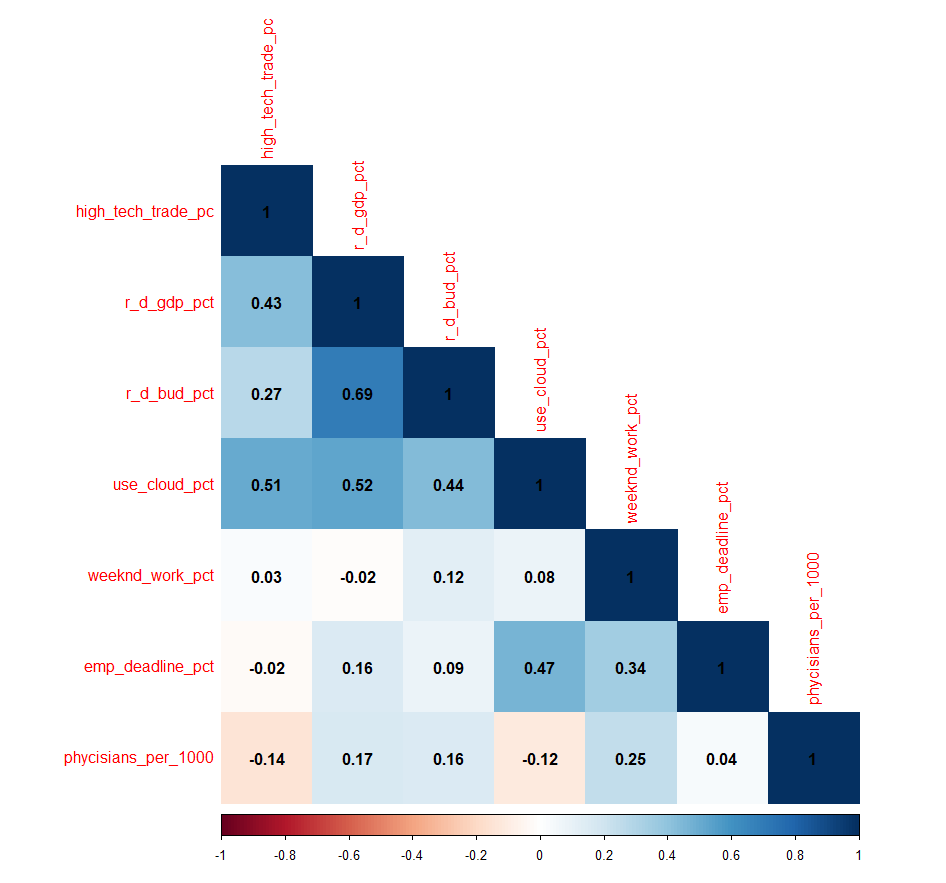
### Korelacje i ich istotność

Na początku – w celu lepszej wizualizacji macierzy korelacyjnej, na jej podstawie wykonano wykres gęstości korelacji między zmiennymi.



Rysunek 2. Wykres gęstości

Dzięki wykonanej wizualizacji możemy dowiedzieć się, że wartości korelacji posiadają rozkład raczej symetryczny, a ich najwięcej wartości statystyki zawiera się w przedziale [0;0.3]. Następnie utworzono korelogram.



Rysunek 3. Korelogram

Najbardziej skorelowane były: zmienna GDP\_pc oraz zmienne odpowiadające m.in. handlem wysokimi technologiami oraz użyciem technologii „cloud” czy wydatkami na badania rozwojowe. Przyjęte przez nas założenia zgadzały się z wizualizacją na wykresie gęstości – korelacje w większości przypadków są słabe. Założenia potwierdza macierz wartości p-value wykonana w programie Rcmdr. Można z niej odczytać, że najwięcej istotnych korelacji zachodzi między PKB per capita, a ww. zmiennymi.

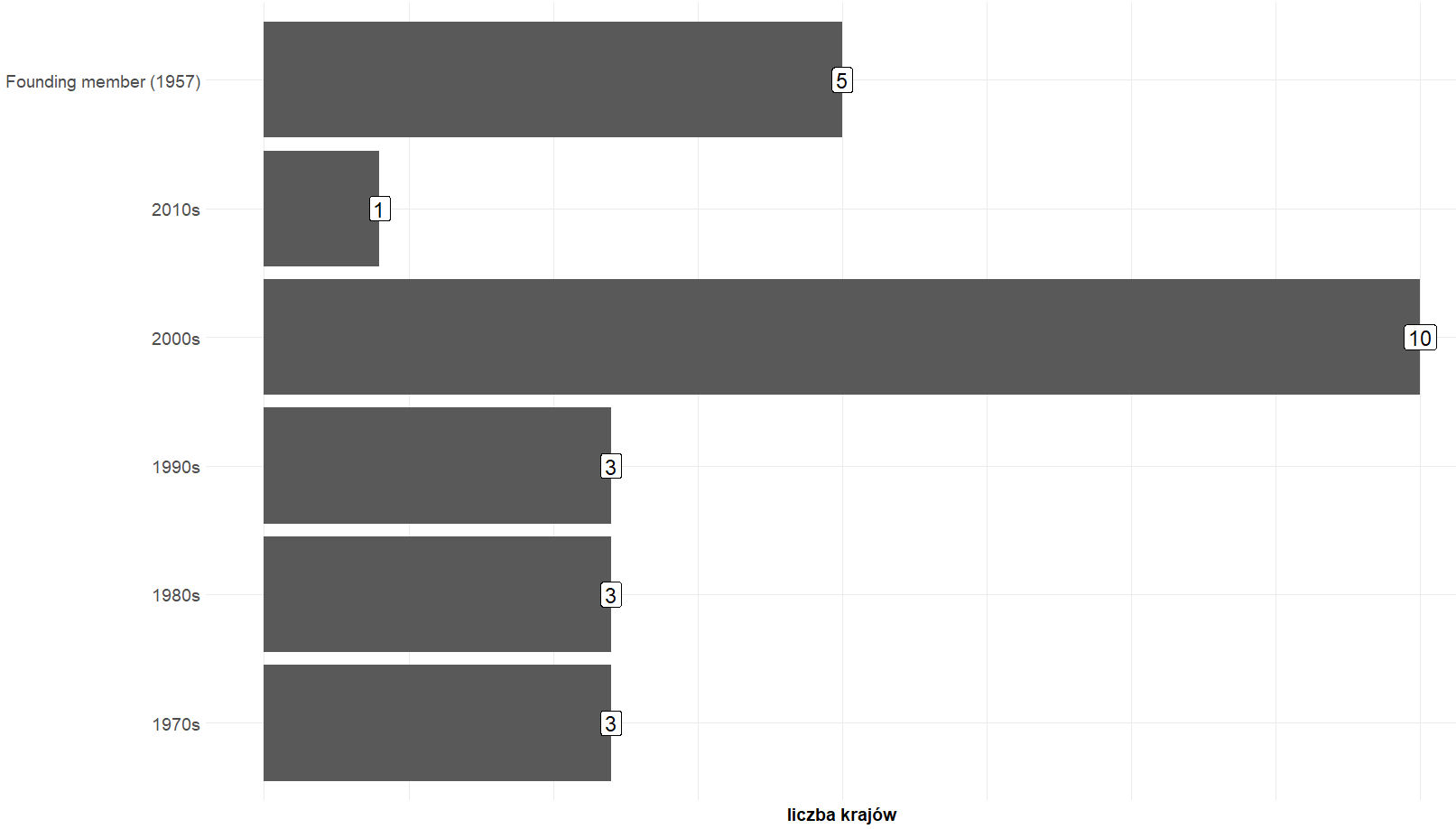
Obraz zawierający stół

Opis wygenerowany automatycznie

Rysunek 4. Macierz p-value

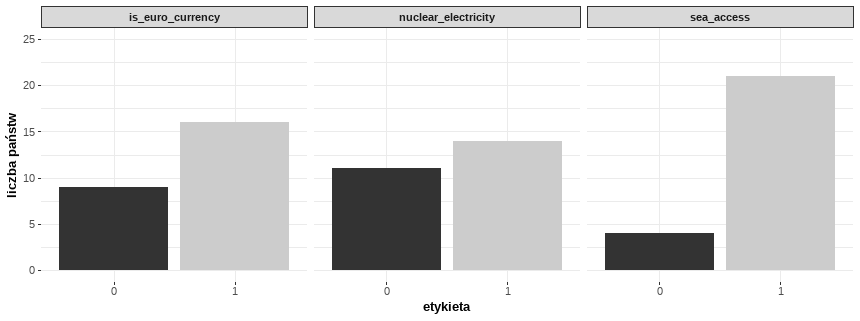
### Wstępna analiza zmiennych kategorycznych

Zmienna „joined\_EU” jest zmienną utworzoną, w celu grupowania krajów. Zmienna ta została utworzona do przyszłej analizy skupień lub klasyfikacji. Niemożliwe dla niej jest wykonanie wcześniejszych analiz. Można natomiast utworzyć wykres słupkowy pokazujący rozkład zmiennej – zliczanie rekordów.



Rysunek 5. Lata dołączenia krajów do UE - wykres słupkowy

Zgodnie z rysunkiem numer 5, większość krajów obecnej Unii Europejskiej, to kraje, które dołączyły do Wspólnoty w pierwszej dekadzie XX wieku. Z kolei zmienne sea\_access, is\_euro\_currency, nuclear\_electricity są binarne. Na rysunku numer 6 zestawiono liczebności etykiet tych zmiennych. Z wykresu można wywnioskować, że liczba państw posiadających elektrownie atomowe jest zbliżona do liczby państw ich nie posiadających. Pozostałe dwie zmienne mają zróżnicowane liczebności.



Rysunek 6. Porównanie zmiennych binarnych

### Dodatkowa analiza

W ramach dodatkowej analizy, postanowiono przyjrzeć się bliżej zmiennej nuclear\_electricity, której liczebności grup są podobne. Zbadano, czy państwa posiadające elektrownie jądrową różnią się od państw, które nie posiadają elektrowni atomowych ze względu na produkt krajowy brutto per capita w euro.

Rozważania rozpoczęto od wykonania testu Shapiro-Wilka, w celu zbadania normalności rozkładów w każdej grupie. W teście przyjęto poziom istotności 0,05. Wyniki obliczeń widoczne są na rysunku numer 7.

Obraz zawierający stół

Opis wygenerowany automatycznie

Rysunek 7. Test Shapiro-Wilka - wyniki

Wartość p-value poniżej 0,05 dla jednej grupy świadczy o tym, że dane nie są normalnie dystrybuowane. Wykonano więc test Manna–Whitneya. Postawiono hipotezę zerową mówiącą o tym, że nie ma różnicy pomiędzy grupami, oraz hipotezę alternatywną, mówiącą że są różnice pomiędzy grupami.

Obraz zawierający tekst, niebo, mapa, jasne

Opis wygenerowany automatycznie

Rysunek 8. Test Manna-Whitneya – wyniki

Zgodnie z wynikami widocznymi na rysunku numer 8, p-value na poziomie 0,68 mówi, że nie ma podstaw do odrzucenia hipotezy zerowej. Grupy nie różnią się między sobą ze względu na produkt krajowy brutto per capita w euro.

Należy opisać zbiór danych (zawierający wszystkie zmienne, które weszły w analizy) oraz jak ten zbiór powstał: skąd ściągnięto dane, jakim transformacjom został poddany (co zrobiono z brakami danych, wartościami odstającymi, czy zastosowano wskaźnik). Tabele i rysunki należy formatować wg poniższych wzorów.

## Opis metod analizy

Po wykonaniu podstawowych statystyk opisowych na danych, zaczęto przeprowadzać pierwsze analizy przy wykorzystaniu metod uczenia nienadzorowanego, chcąc poznać zależności pomiędzy poszczególnymi zmiennymi. Aby odpowiedzieć na pierwsze pytanie badawcze wykonano analizę PCA, dzięki której udało się zminimalizować liczbę predyktorów, przy zachowaniu zadowalającego poziomu wyjaśnianej zmienności. Następnie wykonano analizę skupień przy użyciu dwóch metod (grupowania hierarchicznego i k-średnich), celem uzyskania odpowiedzi na drugie pytanie badawcze. Kolejny etap rozważań dotyczył zagadnień uczenia nadzorowanego. Do wyjaśnienia zjawisk, o których traktowały pytania badawcze numer 3 i 4 wykorzystano metody regresji liniowej. Odpowiednio najpierw przy użyciu wszystkich predyktorów, a potem korzystając z nowych wymiarów, uzyskanych dzięki wykonanej wcześniej analizie PCA. Na koniec, chcąc odpowiedzieć na ostatnie pytanie badawcze, skorzystano z metod klasyfikacji.

# Wyniki

## Odpowiedź na pierwsze pytanie badawcze. (Do jakiej minimalnej ilości wymiarów można zredukować predyktory, jednocześnie zachowując wysoki procent wyjaśnianej wariancji tych predyktorów? Co opisują te wymiary?)

### Przygotowania do analizy

Jako pierwszą wykonano analizę składowych głównych. Najpierw jednak przeprowadzono testy, sprawdzające czy przeprowadzenie PCA jest w ogóle możliwe. Jako pierwszy zrobiono test sferyczności Bartletta. P-value tego testu wyszło bliskie zeru, co pozwala na przeprowadzenie PCA.

Obraz zawierający tekst

Opis wygenerowany automatycznie

Rysunek 9. Test sferyczności Bartletta

Następnie sprawdzono kryterium KMO, którego wynik ogólny (Overall MSA) większy od 0,5 również dopuszcza wykonanie PCA.

Obraz zawierający stół

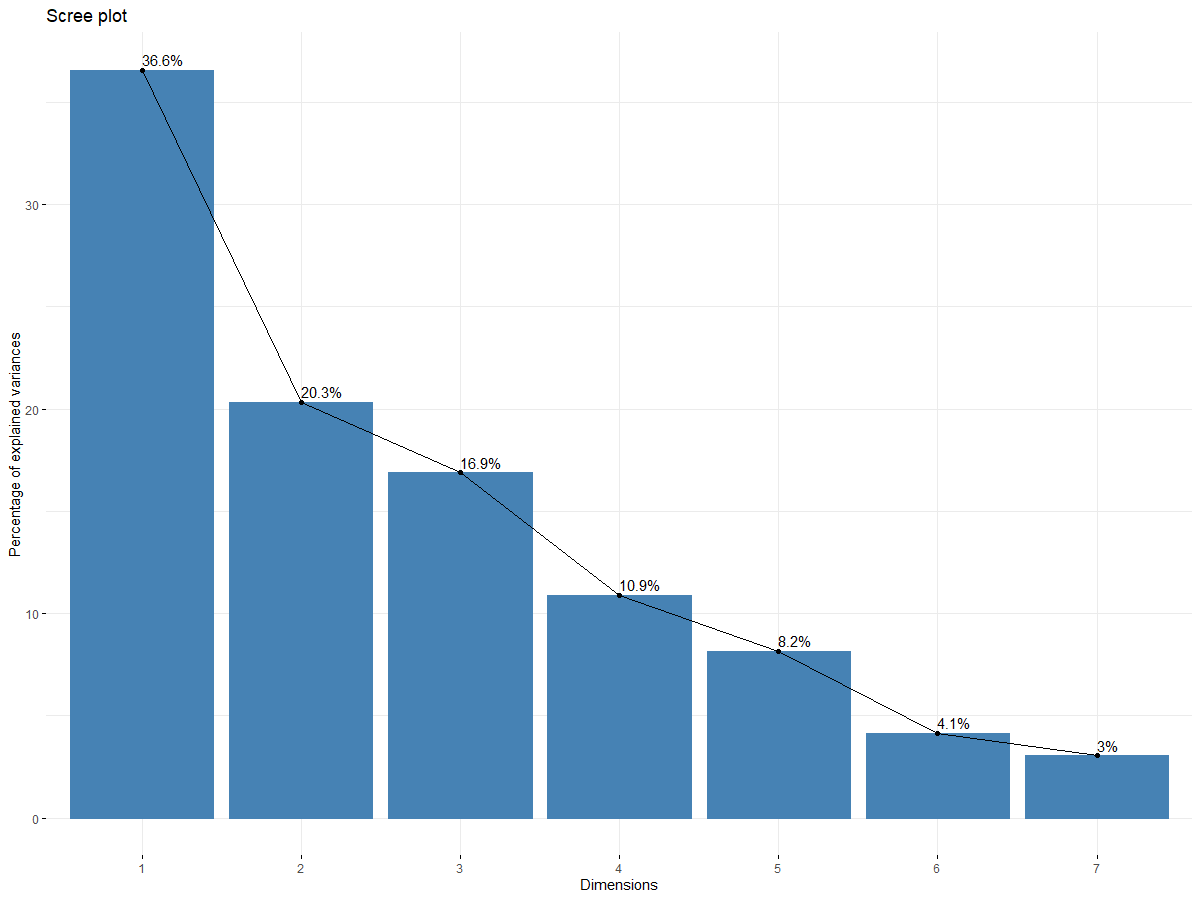
Opis wygenerowany automatycznie

Rysunek 10. Kryterium KMO

Na sam koniec przeprowadzono standaryzację zmiennych

### Właściwa analiza

Samą analizę PCA wykonano przy użyciu biblioteki FactoMiner w języku R. Najpierw przeprowadzono PCA dla liczby wymiarów równej liczbie predyktorów liczbowych, czyli dla 7 oraz stworzono wykres osypiska za pomocą funkcji fviz\_screeplot.



Rysunek 11. Wykres osypiska

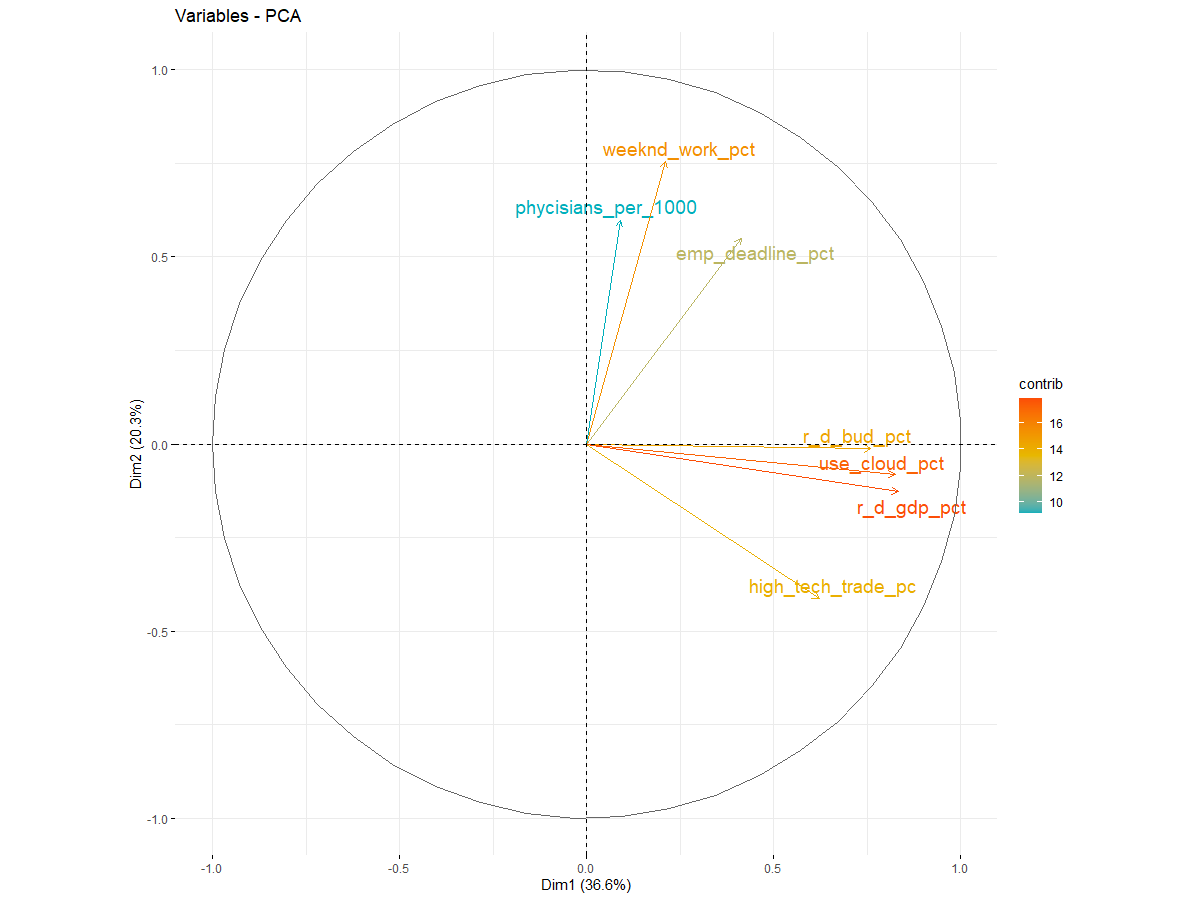
Spłaszczenie wykresu osypiska następuje po 2 wymiarach, które wyjaśniają 57% zmienności.

Obraz zawierający tekst

Opis wygenerowany automatycznie

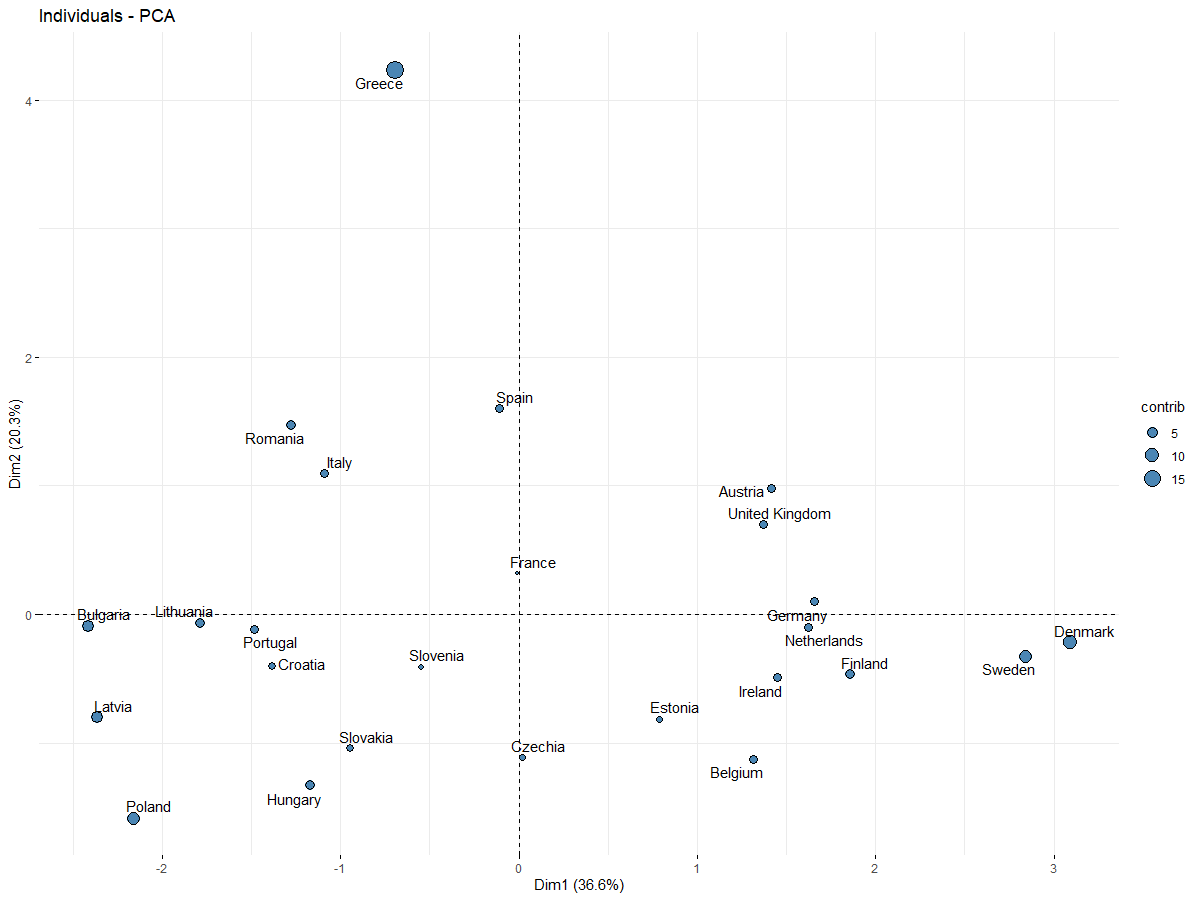
Rysunek 12. Badanie wymiarów

Pierwszy wymiar wyjaśnia handel wysokimi technologiami per capita w 62%, wydatki na badania i rozwój jako procent PKB i procent budżetu kolejno w 83% i 76% oraz procent ludzi korzystających z chmury w 82%. Drugi wymiar z kolei, pokrywa się z procentem ludzi pracujących w weekendy w 75% oraz z pracownikami mającymi deadline w 55%. Ostatnia zmienna, czyli liczba lekarzy na tysiąc mieszkańców w jest wyjaśniana przez drugi wymiar w 60%. Następnie wykonano PCA dla 2 wymiarów. Wyniki lepiej widoczne są na wykresie ładunków czynnikowych (rysunek numer 13).



Rysunek 13. Analiza PCA - wykres ładunków czynnikowych

Zgodnie z wcześniejszymi obserwacjami cztery zmienne układają się zgodnie z wymiarem pierwszym (oś X), trzy zmienne natomiast układają się wzdłuż wymiaru drugiego (oś Y). Można zatem przyjąć, że wymiar pierwszy opisuje nakłady pieniężne na technologię oraz stopień rozwoju technologicznego danego państwa. Drugi wymiar natomiast opisuje kraje bardziej pod kątem społecznym. Ostatnim etapem wykonanej analizy PCA jest naniesienie danych na wykres (rysunek numer 14).



Rysunek 14. Wykres obserwacji

Zgodnie z wykresem przedstawionym na Rysunku 14, kraje można podzielić na 4 grupy zgodnie z kwadrantami, w których się znajdują. Pierwszy kwadrant zawiera Austrię, Wielką Brytanię oraz Niemcy. Są to państwa wysoko rozwinięte oraz dość silnie inwestujące w nowe technologie. Charakteryzują się również wysoką wartością przynajmniej jednego czynnika opisującego wymiar drugi. W przypadku Austrii jest to liczba lekarzy na tysiąc mieszkańców, z kolei w przypadku Wielkiej Brytanii jest to liczba pracowników pracujących w ścisłych terminach. Mimo tego, że Niemcy należą do tego kwadrantu, w żadnej z jego składowych nie odstają mocno od średniej. W kolejnym kwadrancie znalazły się kraje skandynawskie, kraje byłego Beneluksu, Irlandia, Czechy oraz Estonia. Podobnie jak państwa z poprzedniego kwadrantu, kraje te również charakteryzują wyższym nakładem inwestycyjnym na nowe technologie niż pozostałe państwa. Wyróżnia je natomiast niższy wynik w wymiarze drugim. Przeciętnie mniej mieszkańców tych krajów pracuje w weekendy oraz mniejszy ich odsetek pracuje z deadline’ami. Ponadto mogą mieć oni dostęp do mniejszej ilości lekarzy w przeliczeniu na tysiąc mieszkańców. Następny kwadrant stanowią głównie kraje postsowieckie, w tym Polska. Państwa te przeznaczają znacznie mniej środków na rozwój nowoczesnych technologii od swoich poprzedników. Są jednak stosunkowo podobne do krajów z kwadrantu drugiego pod względem wymiaru społecznego. Polska w tej grupie jest państwem najbardziej odstającym pod względem niskiego wyniku w wymiarze drugim. Na taki rezultat wpływ ma najniższa w zestawieniu liczba lekarzy na tysiąc mieszkańców. Do ostatniego kwadrantu należą Rumunia oraz kraje śródziemnomorskie, takie jak Grecja, Włochy, Hiszpania i Francja. Państwa te również przeznaczają mniejszą część budżetu i PKB na rozwój. Wyróżnia je natomiast przeciętnie większa liczba lekarzy na tysiąc mieszkańców, większy odsetek obywateli pracujących w weekendy oraz z deadline’ami. Na tle tych krajów wyraźnie wyróżnia się Grecja. Państwo to przoduje w liczbie pracowników weekendowych, liczbie lekarzy na tysiąc mieszkańców, a także jest drugie w ilości pracowników, pracujących w ścisłych terminach.

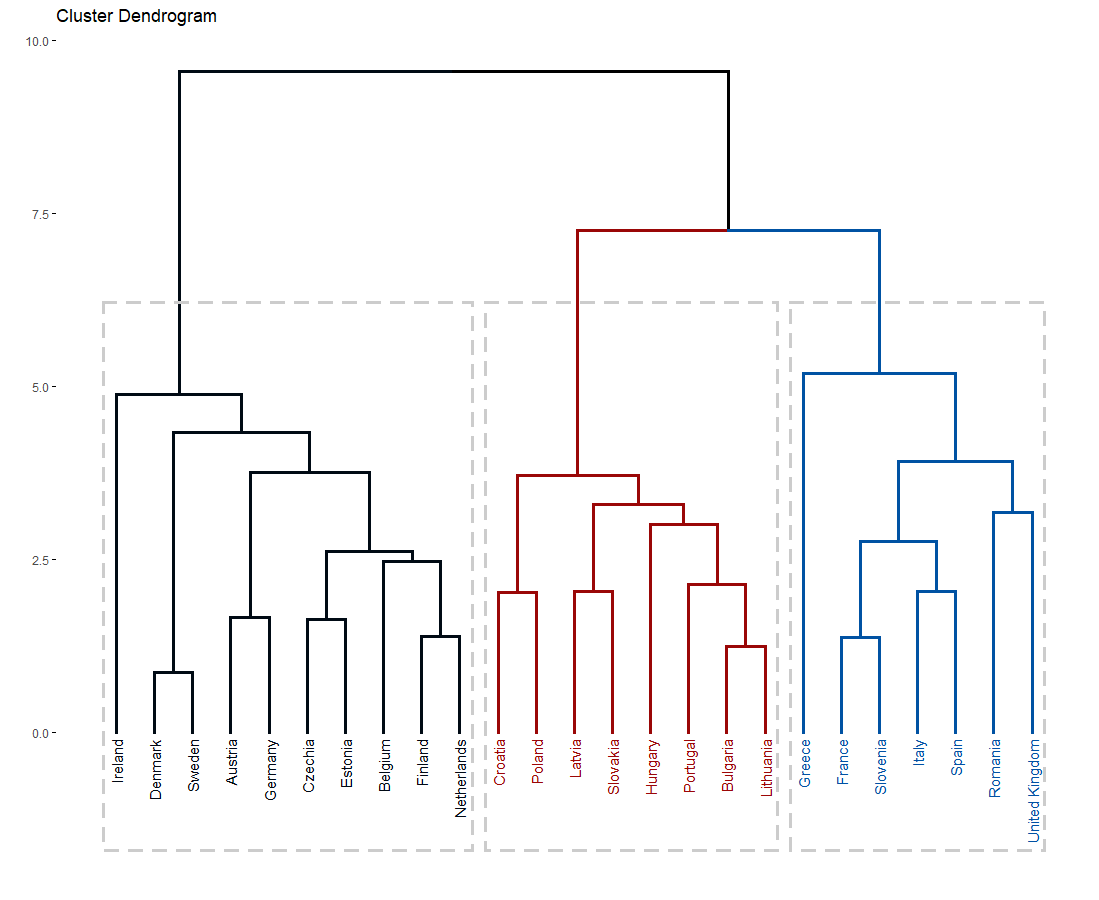
### Podsumowanie

Maksymalnie można zredukować liczbę predyktorów do dwóch, jednocześnie zachowując 57% zmienności. Pierwszy wymiar opisuje poziom nakładów pieniężnych na technologie oraz stopień rozwoju technologicznego danego państwa. Drugi wymiar z kolei opisuje aspekt społeczny tych krajów, głównie pod kątem sposobu pracy ich obywateli, a także dostępu do wykwalifikowanej opieki medycznej.

## Odpowiedź na drugie pytanie badawcze (W jakie skupiska można połączyć kraje, przy użyciu dwóch różnych metod? Czy skupiska są takie same?)

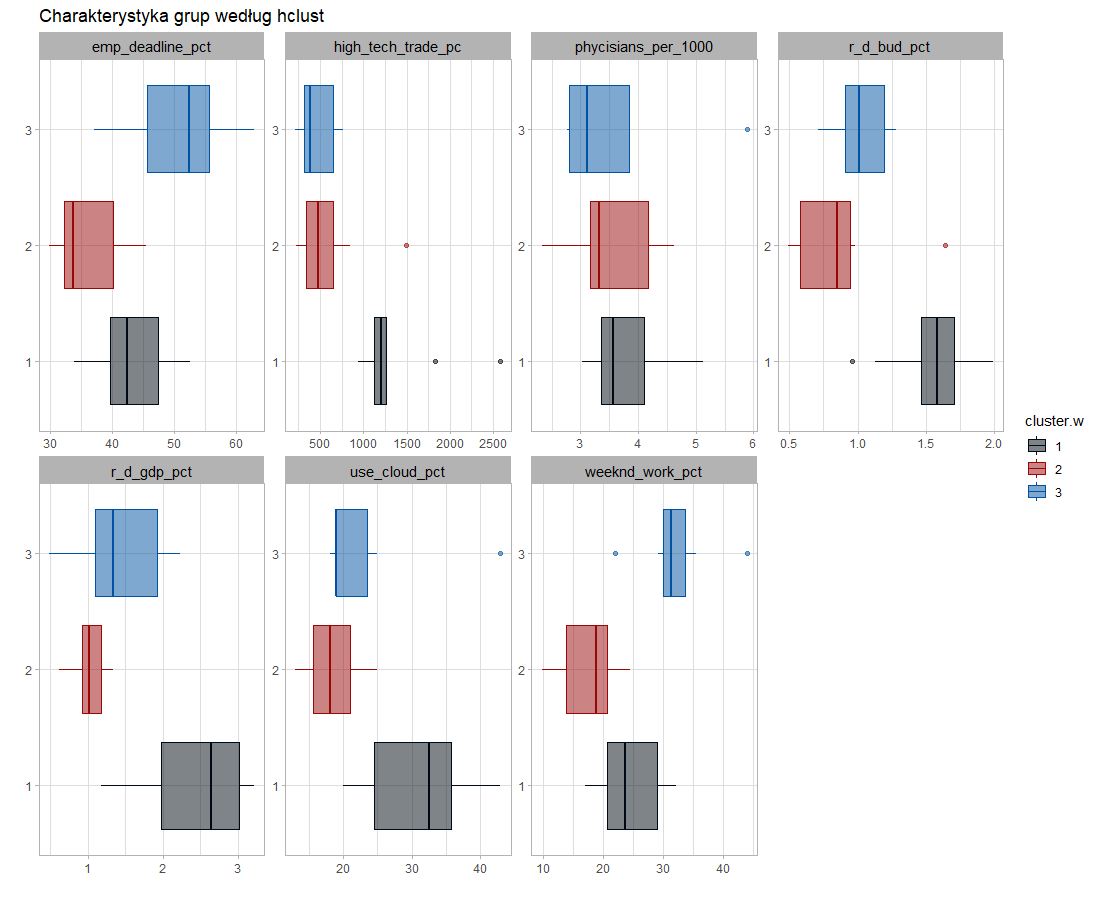
### Grupowanie hierarchiczne

W celu wykonania grupowania niezbędna jest macierz dystansu. Do jej utworzenia zastosowano odległość euklidesową przy pomocy języka R. Potem skorzystano z grupowania hclust, przy użyciu metody „ward.D2”. Wynik grupowania można przedstawić w postaci dendrogramu.



Rysunek 15. Grupowanie hierarchiczne – dendrogram

Zgodnie z rysunkiem numer 15 pogrupowanie danych na trzy części wydaje się być najodpowiedniejsze. Można stwierdzić, że grupy całkiem dobrze pokrywają się z podziałem wynikającym z analizy PCA. Prawie wszystkie kraje grupy pierwszej to państwa, należące do pierwszych dwóch kwadrantów analizy składowych głównych. Jedynym wyjątkiem jest Wielka Brytania, która trafiła do grupy z krajami z kwadrantu czwartego. Do grupy drugiej z kolei trafiły głównie kraje postkomunistyczne, znajdujące się w trzecim kwadrancie analizy PCA. Do ostatniej grupy należą państwa czwartego kwadrantu, a także wcześniej wymieniona Wielka Brytania oraz Słowenia, która w analizie składowych głównych znajdowała się w kwadrancie trzecim.

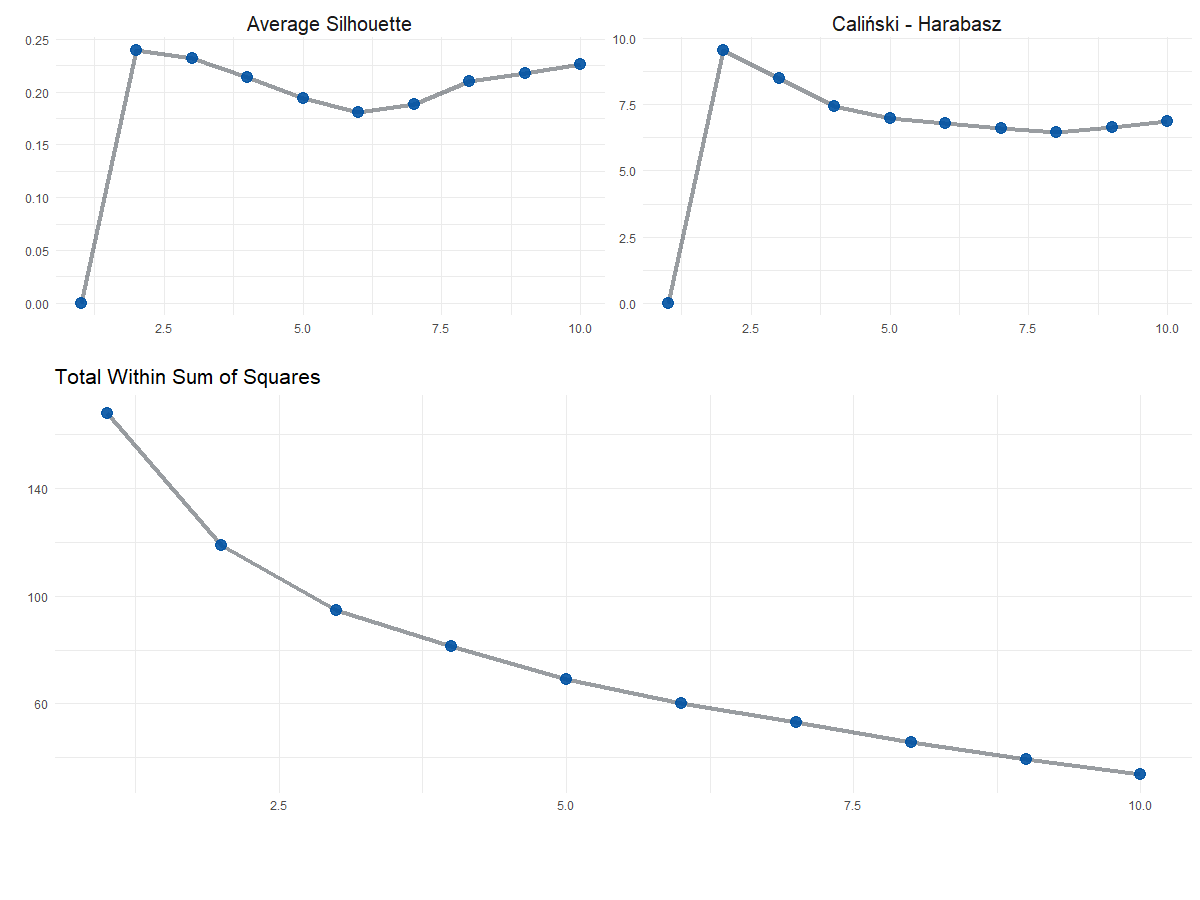


Rysunek 16. Charakterystyka grup - grupowanie hierarchiczne

Przechodząc do charakterystyki grup, od razu można zobaczyć, że pierwsza grupa to kraje najlepiej rozwinięte oraz przeznaczające najwięcej środków na rozwój i nowe technologie. Charakteryzują się najwyższą wartością mediany handlu wysokimi technologiami oraz wydatkami na rozwój. Nawet bez wstępnego sprawdzenia członków danych grup można wysnuć wniosek, że to głównie kraje Europy zachodniej i północnej. Druga grupa to z kolei państwa przeznaczające najmniej funduszy na nowe technologie. W kategoriach, gdzie pierwsza grupa przodowała, druga grupa notuje najczęściej najsłabszy wynik. Można podejrzewać, że głównie są to państwa byłego bloku wschodniego. Ostatnia grupa charakteryzuje się wysokim odsetkiem ludności pracującej w ścisłych terminach oraz populacji pracującej w weekendy. Jest to najprawdopodobniej grupa, w skład której wchodzą kraje śródziemnomorskie.

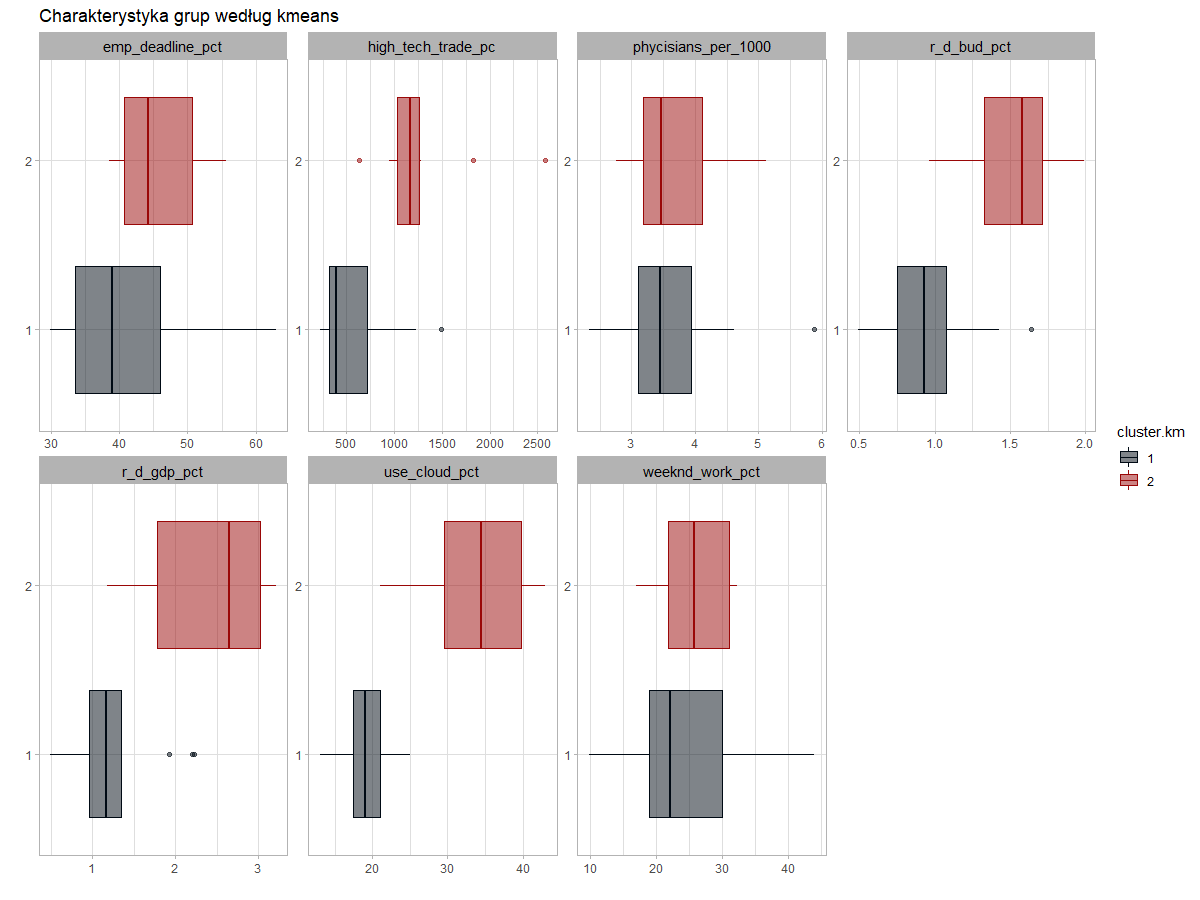
### Grupowanie k-means

Następnie przeprowadzono grupowanie metodą k-means (k-średnich) również przy pomocy języka R. Chcąc ustalić liczbę grupowań, jaką należało wykonać, kierowano się wykresem osypiska, a także kryteriami Calińskiego-Harabasza oraz Average Silhouette.



Rysunek 17. Wykres osypiska, kryteria Average Silhouette oraz Calińskiego-Harabasza – metoda k-means

Wykres osypiska wskazuje na dwie lub trzy grupy, natomiast pozostałe kryteria sugerują grupowanie na dwa. Przeprowadzono zatem analizę skupień na dwóch grupach metodą k-means, a następnie skonfrontowano wyniki z metodą grupowania hierarchicznego.

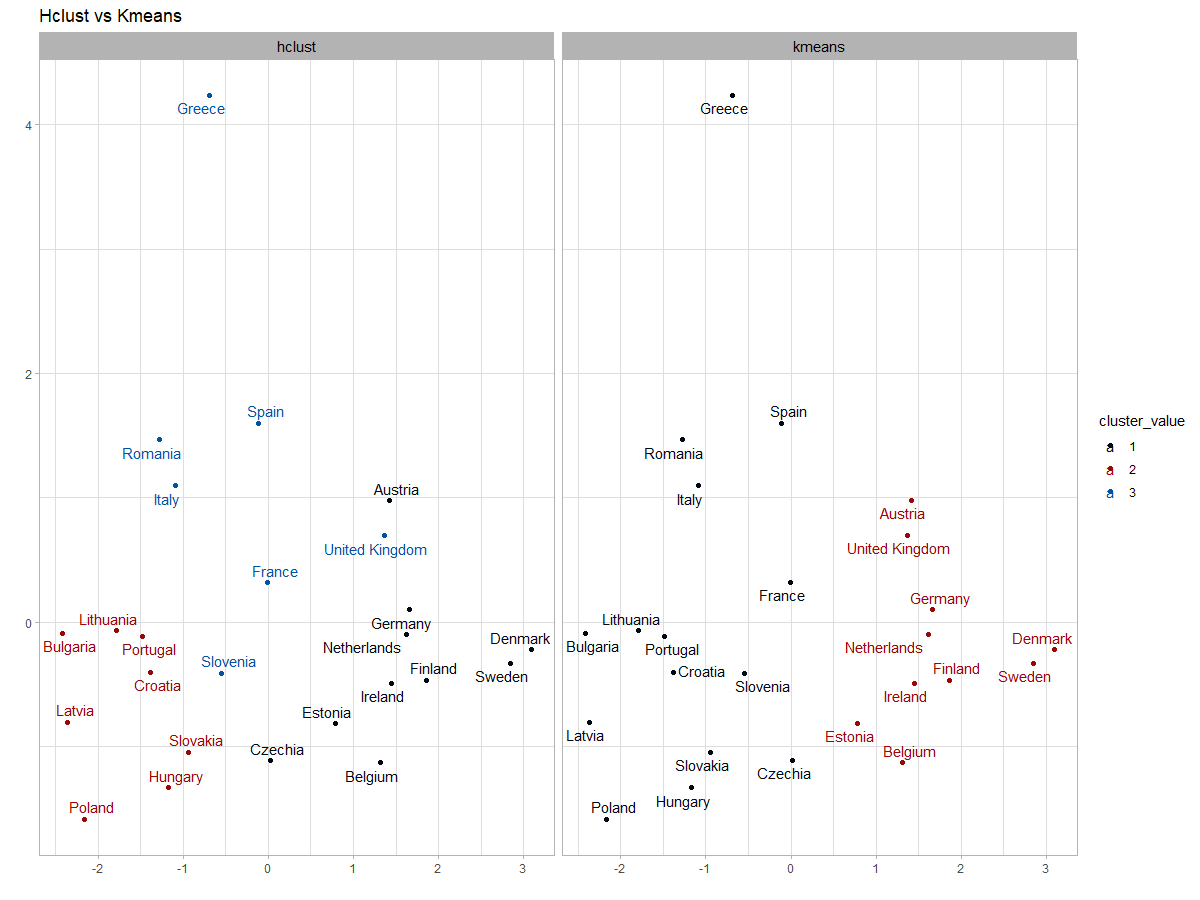


Rysunek 18. Charakterystyka grup - metoda k-means

Charakterystyka grup powstałych w wyniku grupowania k-means jest bardzo prosta. Członkowie grupy drugiej charakteryzują się przeciętnie wyższymi wartościami w każdej kategorii. Można zatem domniemać, że druga grupa to szeroko pojęta Europa Zachodnia, a grupa pierwsza to były blok komunistyczny.

### Porównanie grupowań

Ogólny podział krajów na grupy według metody k-średnich i grupowania hierarchicznego wygląda następująco (rysunek 19).



Rysunek 19. Podział krajów - metoda grupowania hierarchicznego i k-średnich

Zgodnie z wcześniejszymi przypuszczeniami do grupy pierwszej w przypadku grupowania k-means w znacznej większości należą państwa postsowieckie. Dodatkowo w tej grupie znajdują się państwa śródziemnomorskie takie jak Grecja, Hiszpania czy Włochy. Graniczną wartością, decydującą o przynależności poszczególnego kraju wydaje się być zero na osi X (wartość średnia zestandaryzowanego wymiaru pierwszego w analizie PCA).

Grupowanie hierarchiczne z kolei bardziej odpowiada poszczególnym kwadrantom analizy PCA, co dokładniej opisano w części poświęconej grupowaniu hierarchicznemu. Ponadto podział ten zdaje się lepiej odzwierciedlać geograficzne położenie państw względem siebie.

### Podsumowanie

Wykonując analizę skupień metodą grupowania hierarchicznego otrzymano trzy grupy, natomiast przy pomocy metody k-średnich dwie. Pierwsza metoda pozwoliła na uzyskanie grup w dużej mierze zbieżnych z kwadrantami wcześniej wykonanej analizy PCA, druga z kolei dzieli kraje starego kontynentu na państwa bogatego zachodu i północy oraz biedniejszego wschodu i południa.

## Odpowiedź na trzecie pytanie badawcze (Jak wygląda model regresji liniowej wykonany na wszystkich predyktorach? Czy model przeszedł pozytywnie weryfikacje (testy autokorelacji, heteroskedastyczności i normalności reszt)? Jakie jest dopasowanie modelu i jego interpretacja?)

### Wstępna analiza zmiennej zależnej – korelacje

Na podstawie rysunku numer 4 widać, że zmienna GDP\_pc jest najbardziej skorelowana z zmienną „handel wysokimi technologiami per capita”, wydatkami budżetowymi na badania i rozwój oraz procentem osób korzystających z technologii chmurowych. Z kolei zgodnie z rysunkiem numer 5, test na istotność korelacji wykazał, że istotnie ze skorelowane ze zmienną objaśnianą są między innymi zmienne opisujące wydatki na rozwój (r&d), procent osób korzystających z usług chmurowych, handel wysokimi technologiami i procent ludzi pracujących. Chcąc jednak bardziej przebadać zależności między zmiennymi, do pierwszego modelu regresji liniowej postanowiono użyć wszystkich predyktorów dostępnych w zbiorze.

### Regresja – budowa modelu

Przy pomocy dodatku RCommander do programu Rstudio stworzono model regresji liniowej.

Obraz zawierający tekst, paragon

Opis wygenerowany automatycznie

Rysunek 20. Model regresji liniowej - wszystkie predyktory liczbowe

Predyktorami istotnie wpływającymi na zmienną objaśnianą GDP\_pc w utworzonym modelu są: handel wysokimi technologiami, procent PKB przeznaczony na badania rozwojowe oraz procent ludzi pracujących w weekendy – czyli w głównej mierze czynniki gospodarcze. Dopasowanie modelu do danych rzeczywistych wynosi 88%, natomiast duża różnica (prawie 7 punktów procentowych) między współczynnikiem determinacji, a skorygowanym współczynnikiem determinacji, wynika z faktu, że do modelu dołożono niepotrzebne predyktory.

Po wstępnej weryfikacji pierwszego modelu regresji liniowej, do kolejnego modelu wybrano 3 istotne zmienne z modelu 1.

Obraz zawierający tekst, paragon

Opis wygenerowany automatycznie

Rysunek 21. Model regresji liniowej - wybrane predyktory

W tym modelu wszystkie zmienne łącznie z wyrazem wolnym okazały się istotne, zmalał współczynnik determinacji, co jest efektem zmniejszenie liczby predyktorów. Zmalała również różnica między R2, a skorygowanym R2. Użyto zatem zmiennych istotnie wpływających na dopasowanie modelu.

### Weryfikacja modelu

**Autokorelacja reszt**

Obraz zawierający tekst

Opis wygenerowany automatycznie

Rysunek 22. Test Durbina-Watson

Pierwszym etapem weryfikacji modelu było sprawdzenie, czy w modelu występuje autokorelacja reszt. Hipotezą zerową w wykonywanym teście jest brak występowania autokorelacji reszt, natomiast alternatywną – występowanie autokorelacji. W naszym przypadku wynik p-value jest wyższy niż 0.05. Oznacza to, że nie ma podstaw do odrzucenia hipotezy zerowej. Autokorelacja reszt nie występuje.

Zjawisko autokorelacji w danych makro lub mikroekonomicznych może być spowodowane niewłaściwym doborem danych (nie uwzględniając np. czynników inflacyjnych) lub kryzysami ekonomicznymi, które mają odzwierciedlenie w danych statystycznych.

**Heteroskedastyczność**

Obraz zawierający tekst

Opis wygenerowany automatycznie

Rysunek 23. Test Breuscha-Pagana

Celem sprawdzenia heteroskedastyczności przeprowadzono test Breuscha-Pagana. Hipoteza H0 tego testu mówi, że heteroskedastyczność reszt nie występuje (wariancja reszt jest stała). H1z kolei stwierdza, że heteroskedastyczność jest obecna. Zgodnie z wartością p-value wykonanego testu (rysunek 7) przyjmujemy hipotezę H0. Heteroskedastyczność reszt nie występuje. Gdyby była ona obecna, model miałby błędne wnioskowanie statystyczne.

**Normalność składnika losowego**

Obraz zawierający tekst

Opis wygenerowany automatycznie

Rysunek 24. Test Doornika-Hansena

Normalność składnika losowego sprawdzono przy pomocy testu Doornika-Hansena przy wykorzystaniu oprogramowania Gretl. Hipoteza H0 mowi, że rozkład składnika losowego jest normalny. Ponieważ p-value jest wyższe od poziomu istnotności alfa=0.05, przyjmujemy hipotezę zerową. Model został właściwie zweryfikowany.

### Podsumowanie

Model przedstawia się równaniem:

**y = -12317 + 9,33\*high\_tech\_trade\_pc + 9004,64\*r.d\_gdp\_pct + 516,2\*weekend\_work\_pct**

Równanie modelu oznacza, że prognozowana wartość zmiennej objaśnianej jaką jest GDP\_PC zwiększy się o kolejno: 9.33 euro, jeżeli zmienna związana z handlem wysokimi technologiami zwiększy się o jednostkę. Wzrośnie również o 9000,64 euro, jeżeli zmienna związana z procentem wydatków na rozwój państwa zwiększy się o jeden punkt procentowy. Z kolei wzrost zmiennej związanej z pracą w weekendy o jedną jednostkę zwiększy prognozowaną wartość PKP per capita o 516.2 euro.

Model wyjaśnia rzeczywistość w 79%, co jest wysoką wartością. Wartość Pozwala to na wykorzystanie modelu w celach predykcyjnych, przykładowo do przewidywań wartości zmiennej objaśnianej innych krajów dodanych do zbioru (technika wykorzystywana w uczeniu maszynowym).

Model również pozytywnie przeszedł wszystkie weryfikacyjne testy statystyczne, co dokładniej opisano w poprzednim paragrafie.

## Odpowiedź na czwarte pytanie badawcze (Czy korzystając z nowych zmiennych powstałych przy pomocy przeprowadzonej analizy PCA, uproszczony model regresji liniowej ma dużo gorsze dopasowanie? Czy ten model nadaje się do predykcji?)

W wyniku redukcji wymiarów otrzymano dwie składowe – PC2 opisującą kraje pod kątem społecznym, oraz PC1, w skład której weszły predyktory odpowiedzialne za nakłady pieniężne na technologię oraz szeroko rozumiany postęp technologiczny. Dwie zmienne z pierwszego wymiaru (high\_tech\_trade\_pc i r.d\_gdp\_pct) oraz jedna zmienna z drugiego wymiaru – weekend\_work\_pct były istotnymi predyktorami w stworzonych przez nas modelach regresji liniowej.

### Analiza modelu regresji liniowej z jednym wymiarem

Do wstępnej analizy użyto wymiaru, w skład którego wchodzą zmienne związane z sprawami rozwojowymi państw.

Obraz zawierający tekst, paragon, zrzut ekranu

Opis wygenerowany automatycznie

Rysunek 25. Model regresji liniowej z jednym wymiarem

Pierwszy wymiar wyjaśnia zmienną objaśnianą (PKB per capita) w 79%. Zarówno predyktor, jak i wyraz wolny okazały się istotnie wpływać na zmienną celu. Uzyskany model jest istotny – p-value  
wynosi 0.

### Analiza modelu regresji liniowej z dwoma wymiarami

Obraz zawierający tekst, paragon, zrzut ekranu

Opis wygenerowany automatycznie

Rysunek 26. Model regresji liniowej z dwoma wymiarami

W modelu regresji liniowej z dwoma wymiarami predyktorem nieistotnie wpływającym na zmienną objaśnianą jest wymiar opisujący kraje pod względem polityki społecznej. W porównaniu do modelu z poprzedniego punktu należy zauważyć, że różnica pomiędzy R2, a skorygowanym R2 powiększyła się. Do modelu dodano zatem nieistotny predyktor, jakim jest PC2. Jeżeli chodzi o te miary, warto zauważyć, że współczynnik determinacji z poprzedniego punktu jest praktycznie taki sam jak w modelu z dwoma wymiarami.

### Statystyczne porównanie modeli

Obraz zawierający tekst

Opis wygenerowany automatycznie

Rysunek 27. Test ANOVA

Test wykonano przy hipotezie zerowej mówiącej, że oba modele są równe przydatne w celach predykcyjnych i H1 mówiącej o tym, że lepszy wydaje się większy model – z dwoma wymiarami. Otrzymana statystyka testowa jest większa niż poziom istotności, co oznacza, że nie ma podstaw do odrzucenia hipotezy zerowej. Do predykcji nadają się dwa wykonane przez nas modele. Lepszym rozwiązaniem jednak, zgodnym z prawem ekonometrii byłoby wykorzystanie pierwszego modelu.

### Podsumowanie

Uzyskany model regresji liniowej z użyciem pierwszego wymiaru przedstawia się następująco:  
**y = 23677 + 11664 \* PC1** i charakteryzuje się współczynnikiem determinacji na poziomie 79%, identycznym do modelu regresji wykonanego wcześniej na samych predyktorach. Dzięki przekształceniu zmiennych przy pomocy PCA i rozłożeniu ich na wymiary otrzymujemy proste równanie, w którym zamiast 5 predyktorów mamy jeden. Taki uproszczony model może równie dobrze posłużyć do przewidywania wartości PKB per capita, korzystając z danych związanych z rozwojem ekonomiczno-technologicznym danego państwa.

## Odpowiedź na piąte pytanie badawcze (Jak modele klasyfikacyjne radzą sobie z rozpoznawaniem krajów pod względem posiadanej waluty na podstawie danych makroekonomicznych?)

Do budowy modeli jako zmienną celu wykorzystano predyktor binarny – is\_euro\_currency. Zmienna przyjmuje wartości 0 – dla krajów nie będących w systemie walutowym euro oraz 1 – dla krajów posiadających walutę euro jako walutę główną.

### Podstawowe modele klasyfikacyjne

Wstępną klasyfikację wykonano na podstawie modelu k-najbliższych sąsiadów. Model ten wykorzystuje bliskość sklasyfikowanych już rekordów na podstawie zadanych predyktorów – odległości w przestrzeni przez nie wyznaczonej.

Obraz zawierający stół

Opis wygenerowany automatycznie

Rysunek 28. Model k-najbliższych sąsiadów

Model wykazuje dokładność na poziomie 88%, najtrafniej przewiduje przypadki pozytywny (sensitivity), natomiast w przypadku przypadków negatywnych model pomylił się trzykrotnie (specitifity).

Jednym z podstawowych modeli klasyfikacji jest również model “Naiwnego Bayesa”. Do predykcji zmiennej celu wykorzystuje on prawdopodobieństwo warunkowe.

**Obraz zawierający tekst, paragon, zrzut ekranu

Opis wygenerowany automatycznie**

Rysunek 29. Model Naive-Bayes

Model ten wypadł gorzej niż poprzedni. Jednym z powodów gorszej predykcji może być fakt, że wartości predyktorów użytych do klasyfikacji nie są jednorodne. *Naive Bayes* jest używany zazwyczaj do predykcji danych tekstowych i kategorycznych, gorzej sprawdza się przy danych niejednorodnych jak na przykład dane makroekonomiczne. Z kolei drzewa klasyfikacyjne i regresyjne są formą drzew decyzyjnych, wykorzystujące predyktory do jak najlepszego podziału obserwacji ze względu na zmienną celu.

Obraz zawierający stół

Opis wygenerowany automatycznie

Rysunek 30. Model - drzewo klasyfikacyjne

Predykcje modelu są gorsze niż w poprzednich przypadkach. Ma to związek z podziałem drzewa według zmiennej GDP per capita, która różni zdecydowanie w poszczególnych rekordach. Na łącznie sześć nietrafnych predykcji mogą więc mieć wpływ obserwacje odstające. Model przewidział nietrafnie 4 przypadki negatywne (zakwalifikował je jako pozytywne), a dwa pozytywne zakwalifikował jako negatywne.

### Modele usprawniające klasyfikację

Modelami usprawniającymi klasyfikację są modele lasu losowego oraz boosting. Las losowy bazuje na baggingu – losuje dane ze zwracaniem tworząc drzewa decyzyjne. Oprócz próbki danych las losowy za każdym razem losuje inny dobór predyktorów. Jego praca się kończy, gdy utworzy zadaną liczbę drzew, a stworzone przez algorytm modele wskażą większościa głosów najlepsze rozwiązanie. Boosting z kolei to uczenie modelu na próbce danych. Działa on iteracyjnie, za każdym razem “poprawia” wcześniejsze błędy – przywiązuje do nich większe wagi.

Obraz zawierający stół

Opis wygenerowany automatycznie

Rysunek 31. Wyniku boostingu

Obraz zawierający tekst

Opis wygenerowany automatycznie

Rysunek 32.Wyniki lasu losowego

Z powodu małej próbki danych oba algorytmy przedstawiają te same wyniki – wyszły jednak gorzej niż k-najbliższych sąsiadów. Może być to spowodowane tzw. “przetrenowaniem” modelu.

Obraz zawierający stół

Opis wygenerowany automatycznie

Rysunek 33. Porównanie modeli k-ns i boostingu

Modele te mylą się w przypadku Czech, czego powodem może być wyższe od średniej PKB per capita. K-najbliższych sąsiadów myli się w przypadku Węgier i Rumunii, co także jest spowodowane bliskością niektórych predyktorów w przestrzeni, kraje te zostały niepoprawnie sklasyfikowane jako należące do strefy walutowej euro. Boosting z kolei pomylił się w przypadku krajów posiadających euro – Litwy i Królestwa Niderlandów, które zostały sklasyfikowane niepoprawnie jako kraje nieposiadające europejskiej waluty jako środka płatniczego.

### Podsumowanie

Wykonane analizy w większości przypadków odzwierciedlają rzeczywistość, wykorzystane modele w większości przypadków trafnie klasyfikują kraje ze względu na przynależność do strefy euro, jednakże w niektórych przypadkach należałoby użyć innego doboru predyktorów.

# Wnioski

Odpowiadając na pierwsze pytanie badawcze ustalono, że liczbę predyktorów można zredukować do 2, przy zachowaniu 57% zmienności. Dzięki takiemu zabiegowi udało się uprościć interpretacje w późniejszych analizach tj. regresja.

Z kolei przeprowadzając analizę skupień udało się zaobserwować podobieństwa między poszczególnymi krajami, co potwierdziło zdroworozsądkowe przypuszczenia oparte na ogólnej wiedzy historycznej i ekonomicznej na temat tych państw. Klasteryzacja pozwoliła zidentyfikować regiony w UE o najsilniejszej gospodarce, jak i te o najsłabszej

Odpowiadając następnie na pytania badawcze numer 3 i 4 zdołano zbudować dwa dobre modele predykcyjne, dzięki którym możliwe jest trafne przewidywanie poziomu zamożności państwa, podając wyłącznie kilka parametrów. Dzięki wykonanej regresji wiadome jest, które składowe najistotniej wpływają na PKB per capita danego kraju.

Na koniec wykonując klasyfikacje (ze zmienną celu charakteryzującą walutę, obowiązującą w tym państwie) udało się zbudować dobre modele, które mogą posłużyć do ustalenia z prawie 90% dokładnością, czy w danym kraju w Europie obowiązującą walutą jest euro.

Wnioski należy sformułować w kontekście pytań badawczych, w postaci zwartego tekstu bez tabel i rysunków (np. odpowiadając na pierwsze pytanie badawcze ustalono, że …). W tym miejscu należy też wskazać wszystkie ograniczenia przeprowadzonych badań.

# Załącznik 1: Wkład członków zespołu projektowego

Efekt 1: Praca wspólna

Efekt 2: Mateusz Nowak (czyszczenie i łączenie danych, wstępny opis danych)

Efekt 3: Robert Zamiar (obliczenia i sprawozdanie), Damian Okoń (poprawki sprawozdania i dodatkowy test statystyczny)

Efekt 4: Praca wspólna (obliczenia PCA), Damian Okoń (wykresy), Mateusz Nowak (sprawozdanie)

Efekt 5: Praca wspólna (obliczenia analiza skupień), Damian Okoń (wykresy), Mateusz Nowak (sprawozdanie)

Efekt 6: Robert Zamiar (obliczenia, sprawozdanie), Mateusz Nowak (sprawozdanie)

Efekt 7: Robert Zamiar (obliczenia, sprawozdanie), Mateusz Nowak (sprawozdanie)

Efekt 8 (dodatkowa analiza) : Robert Zamiar (obliczenia, sprawozdanie), Mateusz Nowak (sprawozdanie)

Efekt 8 (Całość) : Mateusz Nowak

# Załącznik 2: Poprawki pierwotnych wersji analiz

Efekty 1, 2, 6, 7 – Brak uwag.

Efekt 3: Poprawa tabelek, współczynnik zmienności (pierwotnie policzony był dla wszystkich zmiennych), korelacje, zmiana nazwy histogram na wykres słupkowy, przeprowadzenie dodatkowego testu Manna Whitneya.

Efekt 4: Zmiana nazwy biplot na wykres ładunków czynnikowych.

Efekt 5: Zmiana nazwy metoda euklidesowa na odległość euklidesowa.

# Załącznik 3: Zbiór danych

Data.xlsx – Oryginalny zbiór danych.

data\_pca\_clustering.csv – Dane użyte do PCA i analizy skupień.

data\_regr\_PCA.csv – Dane użyte do regresji, przy wykorzystaniu wymiarów PCA.

# Spis tabel i rysunków

[Rysunek 1. Skośność, kurtoza i kwantyle 6](#_Toc106439767)

[Rysunek 2. Wykres gęstości 7](#_Toc106439768)

[Rysunek 3. Korelogram 8](#_Toc106439769)

[Rysunek 4. Macierz p-value 8](file:///D:\Dokumenty\Studia\Analityka\IV_semestr\Zarządzanie%20projektem%20analiza%20danych%20gospodarczych\Sprawko_calosc.docx#_Toc106439770)

[Rysunek 5. Lata dołączenia krajów do UE - wykres słupkowy 9](#_Toc106439771)

[Rysunek 6. Porównanie zmiennych binarnych 9](#_Toc106439772)

[Rysunek 7. Test Shapiro-Wilka - wyniki 10](#_Toc106439773)

[Rysunek 8. Test Manna-Whitneya – wyniki 10](#_Toc106439774)

[Rysunek 9. Test sferyczności Bartletta 11](#_Toc106439775)

[Rysunek 10. Kryterium KMO 11](#_Toc106439776)

[Rysunek 11. Wykres osypiska 12](#_Toc106439777)

[Rysunek 12. Badanie wymiarów 12](#_Toc106439778)

[Rysunek 13. Analiza PCA - wykres ładunków czynnikowych 13](#_Toc106439779)

[Rysunek 14. Wykres obserwacji 14](#_Toc106439780)

[Rysunek 15. Grupowanie hierarchiczne – dendrogram 15](#_Toc106439781)

[Rysunek 16. Charakterystyka grup - grupowanie hierarchiczne 16](#_Toc106439782)

[Rysunek 17. Wykres osypiska, kryteria Average Silhouette oraz Calińskiego-Harabasza – metoda k-means 17](#_Toc106439783)

[Rysunek 18. Charakterystyka grup - metoda k-means 18](#_Toc106439784)

[Rysunek 19. Podział krajów - metoda grupowania hierarchicznego i k-średnich 19](#_Toc106439785)

[Rysunek 20. Model regresji liniowej - wszystkie predyktory liczbowe 20](#_Toc106439786)

[Rysunek 21. Model regresji liniowej - wybrane predyktory 20](#_Toc106439787)

[Rysunek 22. Test Durbina-Watson 21](#_Toc106439788)

[Rysunek 23. Test Breuscha-Pagana 21](#_Toc106439789)

[Rysunek 24. Test Doornika-Hansena 22](#_Toc106439790)

[Rysunek 25. Model regresji liniowej z jednym wymiarem 23](#_Toc106439791)

[Rysunek 26. Model regresji liniowej z dwoma wymiarami 23](#_Toc106439792)

[Rysunek 27. Test ANOVA 24](#_Toc106439793)

[Rysunek 28. Model k-najbliższych sąsiadów 25](#_Toc106439794)

[Rysunek 29. Model Naive-Bayes 25](#_Toc106439795)

[Rysunek 30. Model - drzewo klasyfikacyjne 26](#_Toc106439796)

[Rysunek 31. Wyniku boostingu 27](#_Toc106439797)

[Rysunek 32.Wyniki lasu losowego 27](#_Toc106439798)

[Rysunek 33. Porównanie modeli k-ns i boostingu 28](#_Toc106439799)

[Tabela 1. Zmienne wykorzystane w projekcie 4](#_Toc106439800)

[Tabela 2. Zmienność i odchylenie standardowe zmiennych numerycznych 5](#_Toc106439801)

[Tabela 3. Charakterystyka zmiennych kategorycznych 5](#_Toc106439802)