Obraz zawierający tekst, urządzenie, wskaźnik

Opis wygenerowany automatycznieObraz zawierający tekst, logo, Czcionka, Grafika

Opis wygenerowany automatycznie

,,Co wpływa na bogactwo społeczeństwa?”

Gabriela Kuczyńska, Martyna Stalmach, Klaudia Grabowska

Analityka biznesowa, 2022/2023, semestr IV

**Spis treści**

[1. Wstęp do analizy 3](#_Toc138796955)

[2. Metody badawcze: 3](#_Toc138796956)

[3. Wyniki 4](#_Toc138796957)

[3.1. Analiza rozkładu badanych zmiennych 4](#_Toc138796958)

[3.2. Analiza PCA 9](#_Toc138796959)

[3.2.1. Test sferyczności Bartletta 9](#_Toc138796960)

[3.2.2. Kryterium KMO (Kaiser-Meyer-Olkin factor adequacy) 9](#_Toc138796961)

[3.2.3. Wizualizacje analizy 10](#_Toc138796962)

[3.2.4. **Odpowiedź na pytanie badawcze ,**,Czy i jak można zredukować wymiary badanych danych?”**:** 12](#_Toc138796963)

[3.3. Analiza skupień (grupowanie) 13](#_Toc138796964)

[3.3.1. Wyświetlenie Dendrogramu 13](#_Toc138796965)

[3.3.2. Metoda WSS 14](#_Toc138796966)

[3.3.3. Metoda Silhouette 14](#_Toc138796967)

[3.3.4. Metoda gap\_stat 15](#_Toc138796968)

[3.3.5. Graficzne przedstawienie grup 16](#_Toc138796969)

[3.3.6. **Odpowiedź na pytanie badawcze „**Ile i jakie grupy utworzą badane dane?”: 16](#_Toc138796970)

[3.4. Sprawdzenie korelacji – które zmienne są istotnie skorelowane ze zmienną PKB na osobę 17](#_Toc138796971)

[3.4.1. Macierz korelacji 17](#_Toc138796972)

[3.4.2. Sprawdzenie istotności współczynników korelacji 18](#_Toc138796973)

[3.4.3. Budowa modelu regresji 19](#_Toc138796974)

[3.4.4. Weryfikacja modelu 21](#_Toc138796975)

[3.4.5. **Odpowiedź na pytanie badawcze** ,,Jak wygląda model regresji badanych danych?”: 21](#_Toc138796976)

[3.5. Przygotowanie danych do analizy wymiarów PCA jako predyktorów 22](#_Toc138796977)

[3.5.1. Jeden wymiar jako predyktor 23](#_Toc138796978)

[3.5.2. Dwa wymiary jako predyktory 23](#_Toc138796979)

[3.5.3. Dwa wymiary jako predyktory 24](#_Toc138796980)

[3.5.4. **Odpowiedź na pytanie badawcze** „Czy wymiary utworzone na etapie analizy PCA są dobrymi predyktorami zmiennej PKB?”**:** 24](#_Toc138796981)

[Załącznik 1: Wkład członków zespołu projektowego 24](#_Toc138796982)

[Załącznik 2: Poprawki pierwotnych wersji analiz 24](#_Toc138796983)

# Wstęp do analizy

W projekcie dokonano analizy wpływu zmiennych na PKB per capita. PKB per capita to jeden z najważniejszych wskaźników gospodarczych, który odzwierciedla wartość produkcji krajowej brutto podzielonej przez liczbę mieszkańców kraju. Jest to wskaźnik, który pokazuje poziom życia i potencjał konsumpcyjny mieszkańców kraju. Celem projektu było przeprowadzenie analizy wpływu różnych zmiennych na poziom PKB per capita, takich jak produkcja energii, wydajność pracowników, zarobki roczne, czy np. wskaźnik konsumpcji. Niektóre wyniki były szczególnie zaskakujące. Na pierwszy rzut oka nie wydawałoby się, że dany czynnik mógłby mieć istotny wpływ na zmienną zależną. W ramach projektu przeprowadzono analizy statystyczne, które umożliwiły zidentyfikowanie zmiennych, które najbardziej wpływają na poziom PKB per capita oraz pomogły określić relacje między nimi. Analiza została przeprowadzana przy użyciu różnych narzędzi i metod statystycznych, takich jak analiza korelacji, PCA, grupowanie, regresja. W projekcie odpowiedziano na pytania badawcze takie jak:

**1. ,,Czy i jak można zredukować wymiary badanych danych?”**

**2. ,,Ile i jakie grupy utworzą badane dane?”**

**3. ,,Jak wygląda model regresji badanych danych?”**

**4. ,,Czy wymiary utworzone na etapie analizy PCA są dobrymi predyktorami zmiennej PKB?”**

# Metody badawcze:

Aby odpowiedzieć na pytanie badawcze czy i jak zredukować wymiary badanych danych użyto metody analizy głównych składowych. Zbadano korelację, wykonano test sferyczności Barletta oraz test KMO, aby upewnić się, że analiza jest możliwa i jej wyniki są miarodajne. Wynikiem analizy były dwa wymiary, wyjaśniające ponad 80% zmienności.

Przeprowadzono analizę skupień w celu odpowiedzi na pytanie dotyczące liczby i rodzaju grup w  badanych danych. Do tego celu wykorzystano funkcję Clusterboot. Na podstawie wyników tej analizy ustalono optymalną liczbę grup oraz metodę. Następnie, w celu potwierdzenia wyników, sporządzono wykresy metodą WSS (Within-Cluster Sum of Squares), Silhouette oraz Gap\_stat. Po przeprowadzeniu tych analiz, podjęto decyzję o wyborze dwóch grup jako optymalnej liczby. Algorytm Clara pomógł zwizualizować grupy w sposób bardziej czytelny. Ostatecznie, wybrano metodę Warda i wykonano dendrogram, który wyraźnie przedstawiał zróżnicowanie między grupami.

Odpowiadając na pytanie jak wygląda model regresji najpierw zbadano korelacje, następnie sprawdzono istotność współczynników korelacji. Kolejnym krokiem było zbudowanie i weryfikacja modelu regresji. W tym celu posłużono się funkcją regresji liniowej oraz wykresem na bazie współczynnika determinacji, aby wybrać najbardziej dopasowany model.

Na pytanie czy wymiary utworzone na etapie analizy PCA są dobrymi predyktorami zmiennej PKB znaleziono odpowiedź posługując się połączeniem analizy PCA i regresji. Metoda polega na użyciu wymiarów z PCA jako predyktorów i porównaniu czy lepiej wyjaśniają model razem czy może pierwszy wymiar sam wyjaśnia więcej zmienności. Badane zjawisko badano za pomocą polecenia Anovy.

# Wyniki

## Analiza rozkładu badanych zmiennych

*Tabela 1. Podstawowe statystyki* przedstawia podstawowe statystyki, takie jak średnia, odchylenie standardowe, rozstęp, skośność, kurtozę, oraz wszystkie kwartyle.

*Tabela 1. Podstawowe statystyki*

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **mean** | **sd** | **IQR** | **cv** | **skewness** | **kurtosis** | **0%** | **25%** | **50%** | **75%** | **100%** |
| PKB na osobę | 30717,10 | 18938,84 | 22820,00 | 0,62 | 1,10 | 0,87 | 6630,00 | 15685 | 25500 | 38505 | 83590 |
| Budżet na badania i rozwój | 1,31 | 0,54 | 0,67 | 0,42 | 0,76 | 0,98 | 0,53 | 0,92 | 1,33 | 1,59 | 32905 |
| Internet use | 87,23 | 8,42 | 14,42 | 0,10 | -0,43 | -0,69 | 67,95 | 81,01 | 87,75 | 95,43 | 99,03 |
| Produkcja energii | 7,66 | 3,78 | 4,39 | 0,49 | 1,23 | 2,07 | 32,00 | 32599,00 | 7,31 | 9,28 | 19,40 |
| Wskaźnik konsumpcji | 105,65 | 40,30 | 47,50 | 0,38 | 1,82 | 4,77 | 53,00 | 76,00 | 97,00 | 123,50 | 251 |
| Wydajność pracowników | 98,60 | 28,52 | 36,45 | 0,29 | 1,13 | 2,16 | 49,00 | 77,10 | 98,50 | 113,55 | 188,20 |
| Zarobki roczne | 21547,36 | 7797,94 | 12166,97 | 0,36 | 0,36 | -0,54 | 10640,64 | 15051,68 | 21985,37 | 27218,65 | 40872,86 |
| Zwolnienia ze szpitala | 2291,16 | 984,10 | 1322,74 | 0,43 | 0,42 | -0,44 | 691,83 | 1627,95 | 2061,64 | 2950,69 | 4509,10 |

Poniższa *Tabela 2. Korelacje między zmiennymi* przedstawia korelacje między predyktorami oraz zmienną zależną. Korelacja między nimi jest mierzona za pomocą współczynnika korelacji Pearsona, który przyjmuje wartości między -1 a 1. Współczynnik korelacji Pearsona 0 oznacza brak korelacji, a wartość -1 lub 1 oznacza idealną korelację negatywną lub pozytywną odpowiednio. Można zauważyć, że korelacje zmiennej zależnej z predyktorami są silne. Osiągają wartości ponad 0,50, co daje nam duże prawdopodobieństwo, że predyktory mają znaczący wpływ na zmienną zależną. Korelacje między predyktorami również są wysokie, jedynie zwolnienia ze szpitala korelują negatywnie z pozostałymi predyktorami, zostały one przedstawione w *Tabela 2. Korelacje między zmiennymi*.

*Tabela 2. Korelacje między zmiennymi*

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  | Budżet na badania i rozwój | Internet use | PKB na osobę | Produkcja energii | Wskaźnik konsumpcji | Wydajność pracowników | Zarobki roczne | Zwolnienia ze szpitala |
| PKB na osobę |  | 0,581 | 0,751 | 1,000 | 0,673 | 0,943 | 0,882 | 0,872 | -0,472 |
| Budżet na badania i rozwój |  | 1,000 | 0,659 | 0,581 | 0,220 | 0,437 | 0,378 | 0,734 | -0,179 |
| Internet use |  | 0,659 | 1,000 | 0,751 | 0,402 | 0,671 | 0,651 | 0,796 | -0,397 |
| Produkcja energii |  | 0,220 | 0,402 | 0,673 | 1,000 | 0,623 | 0,748 | 0,521 | -0,432 |
| Wskaźnik konsumpcji |  | 0,437 | 0,671 | 0,943 | 0,623 | 1,000 | 0,907 | 0,784 | -0,387 |
| Wydajność pracowników |  | 0,378 | 0,651 | 0,882 | 0,748 | 0,907 | 1,000 | 0,759 | -0,474 |
| Zarobki roczne |  | 0,734 | 0,796 | 0,872 | 0,521 | 0,784 | 0,759 | 1,000 | -0,442 |
| Zwolnienia ze szpitala |  | -0,179 | -0,397 | -0,472 | -0,432 | -0,387 | -0,474 | -0,442 | 1,000 |

Two-sided p-value to wartość prawdopodobieństwa, która mówi o tym, jakie jest prawdopodobieństwo, że różnice między dwiema grupami lub zmiennymi jest wynikiem przypadku, a nie rzeczywistej różnicy między nimi. Wynik ten wyraża się w skali od 0 do 1, gdzie wartość p-value bliska 0 oznacza, że istnieje małe prawdopodobieństwo, że różnica między zmiennymi jest wynikiem przypadku, a wartość p-value bliska 1 sugeruje, że nie ma statystycznie istotnej różnicy między zmiennymi. Poziom istotności jaki zakładamy to 0,05, więc bierzemy pod uwagę tylko p-value mniejsze od 0,05.

*Tabela 3. Dwustronne p-value*

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Budżet na badania i rozwój | Internet use | PKB na osobę | Produkcja energii | Wskaźnik konsumpcji | Wydajność pracowników | Zarobki roczne | Zwolnienia ze szpitala |
| Budżet na badania i rozwój |  | <.0001 | 0.0006 | 0.2336 | 0.0141 | 0.0363 | <.0001 | 0.3348 |
| Internet use | <.0001 |  | <.0001 | 0.0251 | <.0001 | <.0001 | <.0001 | 0.0272 |
| PKB na osobę | 0.0006 | <.0001 |  | <.0001 | <.0001 | <.0001 | <.0001 | 0.0073 |
| Produkcja energii | 0.2336 | 0.0251 | <.0001 |  | 0.0002 | <.0001 | 0.0027 | 0.0151 |
| Wskaźnik konsumpcji | 0.0141 | <.0001 | <.0001 | 0.0002 |  | <.0001 | <.0001 | 0.0315 |
| Wydajność pracowników | 0.0363 | <.0001 | <.0001 | <.0001 | <.0001 |  | <.0001 | 0.0071 |
| Zarobki roczne | <.0001 | <.0001 | <.0001 | 0.0027 | <.0001 | <.0001 |  | 0.0128 |
| Zwolnienia ze szpitala | 0.3348 | 0.0272 | 0.0073 | 0.0151 | 0.0315 | 0.0071 | 0.0128 |  |

Jak można zauważyć dla powyższej tabeli przedstawiającej wartości p-value – *Tabela 3. Dwustronne p-value*, w większości przypadków p-value jest < 0,05.

Poniższy wykres przedstawia korelacje w postaci graficznej (*Rysunek 1. Wykres przedstawiający podstawowe korelacje między predyktorami z naszego zbioru.*). Najbardziej skorelowane predyktory ze zmienną zależną to: wskaźnik konsumpcji, wydajność pracowników, zarobki roczne.

*Obraz zawierający wykres

Opis wygenerowany automatycznieRysunek 1. Wykres przedstawiający podstawowe korelacje między predyktorami z naszego zbioru*

Z poniższych boxplotów można odczytać informacje na temat zmiennych, które pozwolą na szybką analizę rozkładu wartości próbki, wykrycie wartości skrajnych i skośności rozkładu.

*Rysunek 2. Wykresy pudełkowe wybranych zmiennych*

Obraz zawierający wykres

Opis wygenerowany automatycznieObraz zawierający wykres

Opis wygenerowany automatycznie

Obraz zawierający wykres

Opis wygenerowany automatycznieObraz zawierający wykres

Opis wygenerowany automatycznieObraz zawierający wykres

Opis wygenerowany automatycznieObraz zawierający wykres

Opis wygenerowany automatycznie

Mediana w powyższym rozkładzie wynosi delikatnie mniej niż 100 euro. Rozkład jest prawostronnie skośny. Wartość odstająca – ok. 250 euro.

Mediana w powyższym rozkładzie wynosi delikatnie ponad 2 000/ 100 000 osób.

Mediana w powyższym rozkładzie wynosi delikatnie ponad 2 000 euro.

Mediana w powyższym rozkładzie wynosi ok.1,3%. Rozkład jest prawostronnie skośny. Wartość odstająca – delikatnie mniej niż 3%.

Mediana w powyższym rozkładzie wynosi ok. 88%.

Mediana w powyższym rozkładzie wynosi delikatnie ponad 2 000 euro. Rozkład jest prawostronnie skośny. Wartość odstająca – ponad 80 000 euro rocznie.

Obraz zawierający wykres

Opis wygenerowany automatycznieObraz zawierający wykres

Opis wygenerowany automatycznie

Mediana w powyższym rozkładzie wynosi ok. 7 euro/kg. Wartość odstająca – delikatnie mniej niż 20 euro/kg. Rozkład jest prawostronnie skośny.

Mediana w powyższym rozkładzie wynosi ok. 100% UE-27. Wartość odstająca – ok. 185% UE-27. Rozkład jest prawostronnie skośny.

## Analiza PCA

PCA wykonuje się tylko wtedy, gdy miedzy zmiennymi występują korelacje.Korelacje są wyraźne, wiec PCA jest możliwa do wykonania.

### Test sferyczności Bartletta

|  |  |  |
| --- | --- | --- |
| chisq | p-value | df |
| 161.1922 | 1.279058e-23 | 21 |

P-value bardzo małe równe 1.279058e-23 mniejsze niż 0.05, czyli zakładany poziom istotności.

### Kryterium KMO (Kaiser-Meyer-Olkin factor adequacy)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | | | | |  |
| Overall MSA = 0.82 |  | |  | |  |
| MSA for each item | | | | | |
| Zwolnienia ze szpitala | | Użycie internetu | | Budżet na rozwój | |
| 0.83 | | 0.92 | | 0.73 | |
| Produkcja energii | | Wydajność pracowników | | Zarobki roczne | |
| 0.83 | | 0.76 | | 0.83 | |
| Wskaźnik konsumpcji |  | |  | |  |
| 0.81 |  | |  | |  |

Powyższe mocne korelacje oznaczają dużą szansę na powodzenie PCA. KMO > 0.5, więc PCA jest jak najbardziej dopuszczalna.

### Wizualizacje analizy

*Wykres 1. Zawartości informacji w składowych.*

Obraz zawierający wykres

Opis wygenerowany automatycznie

Wybrano 2 składowe, ponieważ łącznie wyjaśniają ponad 78% zmienności – można to łatwo odczytać z   *Wykres 1. Zawartości informacji w składowych.* – dwie pierwsze składowe wyjaśniają ponad 78 % (63,1% + 15,5% = 78,6%).

*Wykres 2. Wykres osypiska*

Obraz zawierający wykres

Opis wygenerowany automatycznie

Widać, iż tylko dwie składowe spełniają kryterium Keisera.

Obraz zawierający wykres

Opis wygenerowany automatycznie

*Wykres 3. Wizualizacja procentu wyjaśnianej wariancji dla różnej liczby składowych.*

Dwie składowe wyjaśniają około 78% wariancji.

PC1 = zarobki i ogólne bogactwo kraju.

PC2 = kapitał na rozwój i stan służby zdrowia

*Wykres 4. Wykres krajów według składowych PCA*

Obraz zawierający wykres

Opis wygenerowany automatycznie

### **Odpowiedź na pytanie badawcze ,**,Czy i jak można zredukować wymiary badanych danych?”**:**

Badane dane spełniły warunki do wykonania PCA. Wynikiem tej analizy uzyskano dwa wymiary. Pierwszy wymiar wyjaśniający 63% zmienności i drugi wymiar wyjaśniający 15,5%. Kraje biedniejsze, mniej rozwinięte usytuowały się po lewej stronie wykresu. Kraje bogatsze, lepiej rozwinięte znajdują się po prawej. Zwolnienia ze szpitala wliczają w sobie również śmierci, dlatego odwrotnie korelują z PKB.

## Analiza skupień (grupowanie)

### Wyświetlenie Dendrogramu

Wybrano grupowanie hierarchiczne po przeprowadzeniu badania funkcją clusterboot. Stabilność tej metody oscyluje między 92-94%. Badany zbiór danych jest dość mały, więc ta metoda wydaje się najbardziej odpowiednia.

*Wykres 5. Dendrogram Państw – dodatkowa analiza*

Obraz zawierający tekst, diagram, linia, Wykres

Opis wygenerowany automatycznie

Najbliżej siebie są Czechy i Estonia. Odległość na wykresie tych dwóch zmiennych od siebie jest najmniejsza. Odczytuje się to na podstawie dendrogramu i najniżej położonej kreski łączącej pary. Dendrogram wyraźnie wyznacza dwie grupy. Odczytuje się go metodą "bottom up", czyli od dołu. Dzieli się zbiór na dwie grupy i odczytuje, która grupa ma przewagę dla poszczególnych zmiennych.

### Metoda WSS

*Wykres 6. Dendrogram Państw – dodatkowa analiza*

Obraz zawierający tekst, linia, diagram, Wykres

Opis wygenerowany automatycznie

### Metoda Silhouette

*Wykres 7. Silhouette (dodatkowa analiza)*

Obraz zawierający tekst, linia, diagram, Wykres

Opis wygenerowany automatycznie

### Metoda gap\_stat

*Wykres 8. Gap\_stat (dodatkowa analiza)*

Obraz zawierający tekst, diagram, linia, Wykres

Opis wygenerowany automatycznie

Dla dwóch pierwszych wykresów najbardziej odpowiedni wydaje się podział na dwie grupy. Dla wykresu nr 4, jednej grupy. Biorąc pod uwagę przedstawione wizualizacje dokonuje się podziału na dwie grupy.

### Graficzne przedstawienie grup

*Wykres 9. Wykres według grupowania (dodatkowy)*

Obraz zawierający tekst, diagram, Wykres, numer

Opis wygenerowany automatycznie

### **Odpowiedź na pytanie badawcze „**Ile i jakie grupy utworzą badane dane?”:

Optymalne badania ilości grup oraz dendrogram pomogły podjąć decyzję o wyborze 2 grup. Badane grupy można scharakteryzować następująco:

* W 1 grupie przeważają państwa wysoko rozwinięte, z dużą ilością użytkowników internetowych, dużym budżetem na rozwój.
* W 2 grupie znajdują się kraje, które cechuje duża liczba zwolnień ze szpitala pacjentów z oddziałów kardiologicznych (w tym śmierci), z niskim wskaźnikiem konsumpcji i mało wydajnymi pracownikami.

Odstające wartości przyjęły Irlandia, Luxemburg, Szwajcaria i Bułgaria. Zauważalne skupiska można dostrzec wokół punktów (-1,5; 0) i (1,5; 1). Ewidentnym wnioskiem z przeprowadzonej analizy jest bliskość siebie krajów byłego bloku radzieckiego i ich opóźnienie względem krajów zachodnich i Skandynawii.

## Sprawdzenie korelacji – które zmienne są istotnie skorelowane ze zmienną PKB na osobę

### Macierz korelacji

*Wykres 10. Wykres rozrzutu*

Obraz zawierający tekst, linia, zrzut ekranu, diagram

Opis wygenerowany automatycznie

Poniższa *Tabela 5. Korelacje między zmiennymi* przedstawia korelacje między predyktorami oraz zmienną zależną. Korelacja między nimi jest mierzona za pomocą współczynnika korelacji Pearsona, który przyjmuje wartości między -1 a 1. Współczynnik korelacji Pearsona 0 oznacza brak korelacji, a wartość -1 lub 1 oznacza idealną korelację negatywną lub pozytywną odpowiednio. Można zauważyć, że korelacje zmiennej zależnej z predyktorami są silne. Osiągają wartości ponad 0,50, co daje duże prawdopodobieństwo, że predyktory mają znaczący wpływ na zmienną zależną. Korelacje między predyktorami również są wysokie, jedynie zwolnienia ze szpitala korelują negatywnie z pozostałymi predyktorami, zostały one przedstawione w *Tabela 5. Korelacje między zmiennymi*.

*Tabela 5. Korelacje między zmiennymi*

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  | Budżet na badania i rozwój | Internet use | PKB na osobę | Produkcja energii | Wskaźnik konsumpcji | Wydajność pracowników | Zarobki roczne | Zwolnienia ze szpitala |
| PKB na osobę |  | 0,581 | 0,751 | 1,000 | 0,673 | 0,943 | 0,882 | 0,872 | -0,472 |
| Budżet na badania i rozwój |  | 1,000 | 0,659 | 0,581 | 0,220 | 0,437 | 0,378 | 0,734 | -0,179 |
| Internet use |  | 0,659 | 1,000 | 0,751 | 0,402 | 0,671 | 0,651 | 0,796 | -0,397 |
| Produkcja energii |  | 0,220 | 0,402 | 0,673 | 1,000 | 0,623 | 0,748 | 0,521 | -0,432 |
| Wskaźnik konsumpcji |  | 0,437 | 0,671 | 0,943 | 0,623 | 1,000 | 0,907 | 0,784 | -0,387 |
| Wydajność pracowników |  | 0,378 | 0,651 | 0,882 | 0,748 | 0,907 | 1,000 | 0,759 | -0,474 |
| Zarobki roczne |  | 0,734 | 0,796 | 0,872 | 0,521 | 0,784 | 0,759 | 1,000 | -0,442 |
| Zwolnienia ze szpitala |  | -0,179 | -0,397 | -0,472 | -0,432 | -0,387 | -0,474 | -0,442 | 1,000 |

### Sprawdzenie istotności współczynników korelacji

Two-sided p-value to wartość prawdopodobieństwa, która mówi o tym, jakie jest prawdopodobieństwo, że różnice między dwiema grupami lub zmiennymi jest wynikiem przypadku, a nie rzeczywistej różnicy między nimi. Wynik ten wyraża się w skali od 0 do 1, gdzie wartość p-value bliska 0 oznacza, że istnieje małe prawdopodobieństwo, że różnica między zmiennymi jest wynikiem przypadku, a wartość p-value bliska 1 sugeruje, że nie ma statystycznie istotnej różnicy między zmiennymi. Poziom istotności jaki zakładamy to 0,05, więc bierze się pod uwagę tylko p-value mniejsze od 0,05.

*Tabela 6. Dwustronne p-value*

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Budżet na badania i rozwój | Internet use | PKB na osobę | Produkcja energii | Wskaźnik konsumpcji | Wydajność pracowników | Zarobki roczne | Zwolnienia ze szpitala |
| Budżet na badania i rozwój |  | <.0001 | 0.0006 | 0.2336 | 0.0141 | 0.0363 | <.0001 | 0.3348 |
| Internet use | <.0001 |  | <.0001 | 0.0251 | <.0001 | <.0001 | <.0001 | 0.0272 |
| PKB na osobę | 0.0006 | <.0001 |  | <.0001 | <.0001 | <.0001 | <.0001 | 0.0073 |
| Produkcja energii | 0.2336 | 0.0251 | <.0001 |  | 0.0002 | <.0001 | 0.0027 | 0.0151 |
| Wskaźnik konsumpcji | 0.0141 | <.0001 | <.0001 | 0.0002 |  | <.0001 | <.0001 | 0.0315 |
| Wydajność pracowników | 0.0363 | <.0001 | <.0001 | <.0001 | <.0001 |  | <.0001 | 0.0071 |
| Zarobki roczne | <.0001 | <.0001 | <.0001 | 0.0027 | <.0001 | <.0001 |  | 0.0128 |
| Zwolnienia ze szpitala | 0.3348 | 0.0272 | 0.0073 | 0.0151 | 0.0315 | 0.0071 | 0.0128 |  |

Jak można zauważyć na powyższej tabeli *Tabela 6. Dwustronne p-value*. - Wszystkie zmienne są istotnie (p<0,05) skorelowane ze zmienną objaśniającą. A co za tym idzie: do modelu jako wstępne zmienne objaśniające bierzemy zmienne: Budżet na badania i rozwój, Internet use, Produkcja energii, Wskaźnik konsumpcji, Wydajność pracowników, Zarobki roczne, Zwolnienia ze szpitala.

### Budowa modelu regresji

*Tabela 7. Model regresji 1*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Estimate | Std. Error | T value | Pr(>|t|) |
| (Intercept) | -24789.2162 | 13098.2674 | -1.893 | 0.0711 |
| Budżet na badania i rozwój | 3754.6242 | 2640.4132 | 1.422 | 0.1685 |
| Internet use | 125.1901 | 175.5932 | 0.713 | 0.4830 |
| Produkcja energii | 743.1344 | 349.3446 | 2.127 | 0.0443\* |
| Wskaźnik konsumpcji | 326.8950 | 54.7137 | 5.975 | 0.00000431 \*\*\* |
| Wydajność pracowników | -64.7916 | 89.1568 | -0.727 | 0.4747 |
| Zarobki roczne | 0.4109 | 0.2761 | -1.488 | 0.1503 |
| Zwolnienia ze szpitala | -1.3204 | 1.0312 | -1.281 | 0.2131 |

|  |
| --- |
| Residual standard error: 4625 on 23 degrees of freedom |
| Multipe R-squared: 0.9543, Adjusted R-squaredL 0.9404 |
| F-statistic: 68.58 on 7 and 23 DF, p-value: 6.714e-14 |

Wartość współczynnika determinacji 0.9543 oznacza, że 95.43% zmienności w zmiennej zależnej jest wyjaśniane przez zastosowany model regresji. Wysoka wartość sugeruje, że model bardzo dobrze dopasowuje się do danych. Wartość p-value na poziomie 0.00 sugeruje, że model regresji jest statystycznie istotny. Nie wszystkie zmienne są jednak istotne. Zmiennymi istotnymi są: Produkcja energii, Wskaźnik konsumpcji więc te zmienne bierzemy. Poza tym bierzemy jeszcze zmienną – Zarobki roczne, gdyż jego p-value jest stosunkowo nie duże na tle pozostałych zmiennych.

Obraz zawierający tekst, diagram, Prostokąt, linia

Opis wygenerowany automatycznie

*Tabela 8. Model regresji 2*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Estimate | Std. Error | T value | Pr(>|t|) |
| (Intercept) | -21473.4797 | 2631.4551 | -8.160 | 0.00000000916\*\*\* |
| Produkcja energii | 608.4336 | 291.8271 | 2.085 | 0.046664\* |
| Wskaźnik konsumpcji | 285.0313 | 37.7064 | 7.559 | 0.00000003937\*\*\* |
| Zarobki roczne | 0.8084 | 0.1785 | 4.528 | 0.000108\*\*\* |

|  |
| --- |
| Residual standard error: 4721 on 27 degrees of freedom |
| Multipe R-squared: 0.9441, Adjusted R-squaredL 0.9379 |
| F-statistic: 151.9 on 3 and 27 DF, p-value: 2.2e-16 |

W tym przypadku wartość współczynnika determinacji 0.9441 oznacza, że 94.41% zmienności w zmiennej zależnej jest wyjaśniane przez zastosowany model regresji. Wysoka wartość sugeruje, że model bardzo dobrze dopasowuje się do danych. Wartość p-value na poziomie 0.00 sugeruje, że model regresji jest statystycznie istotny. Zmiennymi objaśniającymi pozostają zatem: Produkcja energii, Wskaźnik konsumpcji oraz Zarobki roczne.

Przyjmuje się postać modelu:

### Weryfikacja modelu

|  |  |  |  |
| --- | --- | --- | --- |
| Shapiro-Wilk normality test | Data:  residuals.RegModel.2 | W = 0.85403 | P-value = 0.000618 |
| Pearson chi-square normality test | Data:  residuals.RegModel.2 | P = 6.4194 | P-value = 0.2675 |

Wniosek: Test Shapiro-Wilka wykazał, iż należy odrzucić hipotezę o normalności reszt. Z kolei test Pearsona pozwolił przyjąć hipotezę o normalności reszt.

### **Odpowiedź na pytanie badawcze** ,,Jak wygląda model regresji badanych danych?”:

Tak kształtuje się ostateczny model:

Wszystkie zmienne składające się na model są wskaźnikami; jedynie zarobki roczne są zmienną numeryczną(PPS).

Na PKB na osobę ma wpływ: Produkcja energii, Wskaźnik konsumpcji oraz zarobki roczne – zmienne te wyjaśniają zmienną PKB na osobę w 95,4%. W pozostałych 4,6% wpływają na nią inne czynniki (100%-R^2). Wzrost produkcji energii o 1 jednostkę zwiększa PKB na osobę 608.4336 jednostek. Wzrost Wskaźnika konsumpcji o 1 jednostkę powoduje wzrost PKB na osobę o 285.0313 jednostek. Wzrost Zarobków rocznych o 1 jednostkę PPS powoduje wzrost PKB na osobę o 0.8084 jednostek. Weryfikacja założeń modelu wykazała, że rozkład normalny reszt modelu jest wątpliwy, co stanowi ograniczenie zbudowanego modelu.

## Przygotowanie danych do analizy wymiarów PCA jako predyktorów

|  |  |  |
| --- | --- | --- |
| PC1 | PC2 | PKB |
| 0.45087740 | -0.02533994 | 36110 |
| -1.99773379 | 0.10484738 | 6630 |
| -0.47542992 | 0.75198449 | 18460 |
| 1.10227925 | -0.37684856 | 48970 |
| 0.57488351 | 1.56116149 | 35950 |
| -0.54451790 | 1.28975462 | 15410 |
| 1.80526374 | -3.01578544 | 59560 |
| -0.77346761 | 0.09911590 | 17780 |
| 0.17963923 | -0.42684440 | 25180 |
| 0.31555281 | -0.50479459 | 33250 |
| -0.71345657 | 0.25572405 | 12710 |
| -0.17823568 | -1.08661419 | 27230 |
| -0.06439735 | -0.94399050 | 25500 |
| -1.11656382 | -0.17465838 | 12540 |
| -1.11924365 | 0.47494161 | 14060 |
| 1.91724832 | -0.74613572 | 83590 |
| -1.22862795 | -0.15115828 | 13310 |
| -0.32548584 | -0.68661401 | 22890 |
| 0.84673078 | 0.48683559 | 41980 |
| 0.39782325 | 0.52494130 | 38090 |
| -0.83845955 | 0.06585457 | 13070 |
| -0.69747467 | -1.33127543 | 18670 |
| -1.33501849 | -0.74040608 | 9300 |
| -0.56813363 | -0.36734830 | 20770 |
| -1.00578037 | -0.05868041 | 15960 |
| 0.28742950 | 0.95201911 | 37150 |
| 0.63735994 | 0.34429875 | 44180 |
| 0.74242906 | 1.66702859 | 38920 |
| 1.52460131 | -0.13848490 | 70150 |
| 1.57291755 | 2.31949999 | 61950 |
| 0.62699114 | -0.12302831 | 32910 |

### Jeden wymiar jako predyktor

|  |  |
| --- | --- |
| Współczynniki | P- value |
| Wyraz wolny | 2e-16 |
| PC1 | 3,532e-16 |

|  |  |  |  |
| --- | --- | --- | --- |
| Współczynnik R^2 | 0,9022 | Dopasowany wsp. R^2 | 0,8988 |

|  |  |  |  |
| --- | --- | --- | --- |
| Statystyka F | 267,5 | P-value | 3,3532e-16 |

Wymiar 1 PC1 wyjaśnia 90% zmiennej celu PKB.

Parametry modelu są istotne obie wartości p -value są mniejsze od 0,05.

### Dwa wymiary jako predyktory

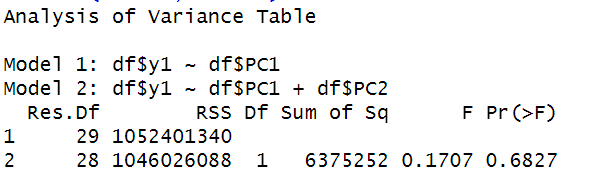
|  |  |
| --- | --- |
| Współczynniki | P- value |
| Wyraz wolny | 2e-16 |
| PC1 | 1,06e-15 |
| PC2 | 0,683 |

|  |  |  |  |
| --- | --- | --- | --- |
| Współczynnik R^2 | 0,9028 | Dopasowany wsp. R^2 | 0,8958 |

|  |  |  |  |
| --- | --- | --- | --- |
| Statystyka F | 130 | P-value | 6,73e-15 |

Uzyskano istotny model F- statistic, wyjaśnia on 90% zmienności zmiennej PKB. PC2 za to okazało się nieistotne.

### Dwa wymiary jako predyktory



Po porównaniu modeli stosując polecenie Anovy statystyka F Pr(>F) wynosi 0.6827 i jest większa

od przyjętego poziomu istotności alfa=0.05. Można zatem przyjąć hipotezę H0. Oznacza to, że oba modele są równie dobre do przewidywania wyniku. Należy jednak wybrać model mniejszy

### **Odpowiedź na pytanie badawcze** „Czy wymiary utworzone na etapie analizy PCA są dobrymi predyktorami zmiennej PKB?”**:**

Tak, wymiary utworzone na etapie analizy PCA są dobrymi predyktorami zmiennej PKB. Należy jednak wybrać tylko jeden z nich PC1.

## Załącznik 1: Wkład członków zespołu projektowego

Projekt został przeprowadzony w pełni opierając się na współpracy. Każdy członek zespołu brał czynny udział we wszystkich etapach projektu. Od samego początku podjęłyśmy decyzję o równoczesnym zaangażowaniu wszystkich członków zespołu. Wszystkie miałyśmy możliwość wniesienia swojego wkładu, przedstawienia pomysłów i podejmowania decyzji w grupie. Praca była podzielona w sposób elastyczny, aby umożliwić nam efektywne reagowanie na zmieniające się wymagania i wyzwania projektowe. Regularnie się komunikowałyśmy i dzieliłyśmy się zadaniami, aby każda z nas mogła przyczynić się do różnych aspektów projektu. Ważne było również, że utrzymywałyśmy otwarty dialog i regularnie organizowałyśmy spotkania, w których omawialiśmy postęp. Podsumowując, nasz projekt opierał się w całości na współpracy. Każda z nas przyczyniła się do stworzenia ostatecznej wersji naszego projektu.

## Załącznik 2: Poprawki pierwotnych wersji analiz

Poprawione zostały efekty: 3,4,6 oraz 7.

**W efekcie 3** poprawiłyśmy nazewnictwo z określeń y oraz x na kolejno zmienna zależna i predyktor. Każdemu wykresowi pudełkowemu został również przypisany komentarz zgodnie ze wskazaniem. Dodane zostało także podsumowanie korelacji między zmiennymi a także komentarz interpretacyjny do *Rysunek 1. Wykres przedstawiający* podstawowe *korelacje między predyktorami z naszego zbioru.* Przy interpretacji Two-sided p-value uzwgledniłyśmy również odwołanie do poziomu istotności alfa.

Jeżeli chodzi o poprawki **w efekcie 4** przede wszystkim poprawiony przez nas został Test sferyczności Barletta oraz Kryterium KMO. Wyniki zostały ujęte w tabelach według sugestii. Poprawiony został także wniosek do testu sferyczności Barletta. Zostało także uwzględnione w tekście odniesienie do *Wykres 2. Wykres osypiska*. Usunięty został również PCA-Biplot. We wnioskach z PCA znalazła się również odpowiedź na pytanie badawcze, oraz poprawione odwołanie – zamiast do zmiennych odwołałyśmy się do utworzonych wymiarów.

**W efekcie 6** poprawiłyśmy wnioski przy modelu regresji. Wyjaśniłyśmy dlaczego zmienna Zarobki roczne weszła do modelu mimo że jej p-value > 0.05. Odpowiedź na pytanie badawcze również została wyraźnie zaznaczona. **W efekcie 7** poprawiłyśmy interpretację oraz odpowiedź na pytanie badawcze według sugestii.