# Raport Lista 2

Eksploracja danych

Dominik Kowalczyk i Matylda Mordal

2025-05-01

## Spis treści



# 1 Analiza składowych głównych (Principal Component Analysis (PCA))

## 1.1 Przygotowanie danych i ich opis

Dane pochodzą z pliku CSV i zawierają informacje o jakości życia w różnych miastach świata, które posłużą do analizy danych (z pliku "uaScoresDataFrame.csv") dostępnego pod linkiem: https://www.kaggle.com/datasets/orhankaramancode/city-quality-of-life-dataset.

```
#Wczytanie zbioru danych
danePCA <- read.csv(file="uaScoresDataFrame.csv", stringsAsFactors = TRUE)
```

Zweryfikujmy i spawdźmy z jakimi danymi mamy doczynenia

Tabela 1: Opis danych PCA

| Indeks | Nazwa zmiennej | Typ zmiennej | Opis zmiennej |
|---|---|---|---|
| 1 | X | integer | Indeks porządkowy |
| 2 | UA_Name | factor | Nazwa obszaru miejskiego |
| 3 | UA_Country | factor | Kraj obszaru miejskiego |
| 4 | UA_Continent | factor | Kontynent obszaru miejskiego |
| 5 | Housing | numeric | Wskaźnik standardu zamieszkania |

| | | | |
|---|---|---|---|
| 6 | Cost.of.Living | numeric | Wskaźnik kosztów życia |
| 7 | Startups | numeric | Wskaźnik liczby startupów |
| 8 | Venture.Capital | numeric | Wskaźnik kapitału wysokiego ryzyka |
| 9 | Travel.Connectivity | numeric | Wskaźnik łączności podróżniczej |
| 10 | Commute | numeric | Wskaźnik dojazdów do pracy |
| 11 | Business.Freedom | numeric | Wskaźnik swobody działalności gospodarczej |
| 12 | Safety | numeric | Wskaźnik bezpieczeństwa |
| 13 | Healthcare | numeric | Wskaźnik opieki zdrowotnej |
| 14 | Education | numeric | Wskaźnik edukacji |
| 15 | Environmental.Quality | numeric | Wskaźnik jakości środowiska |
| 16 | Economy | numeric | Wskaźnik ekonomii |
| 17 | Taxation | numeric | Wskaźnik opodatkowania |
| 18 | Internet.Access | numeric | Wskaźnik dostępu do internetu |
| 19 | Leisure...Culture | numeric | Wskaźnik rekreacji i kultury |
| 20 | Tolerance | numeric | Wskaźnik tolerancji |
| 21 | Outdoors | numeric | Wskaźnik aktywności na świeżym powietrzu |

W zbiorze danych nie występują brakujące wartości (NA), Struktura i podstawowe statystyki zmiennych zostały sprawdzone za pomocą str() i summary(). Żadne nieprawidłowości w typach danych ani w rozkładzie zmiennych nie zostały zidentyfikowane. Kolumna `X` zawiera wyłącznie numerację wierszy i nie wnosi istotnych informacji, zatem nie należy brać jej pod uwagę podczas naszej analizy.

Metodę PCA można zastosować tylko na zmiennych rodzaju ilościowego, zatem wyodrębmy je (pamiętając o usunięciu kolumny `X`):

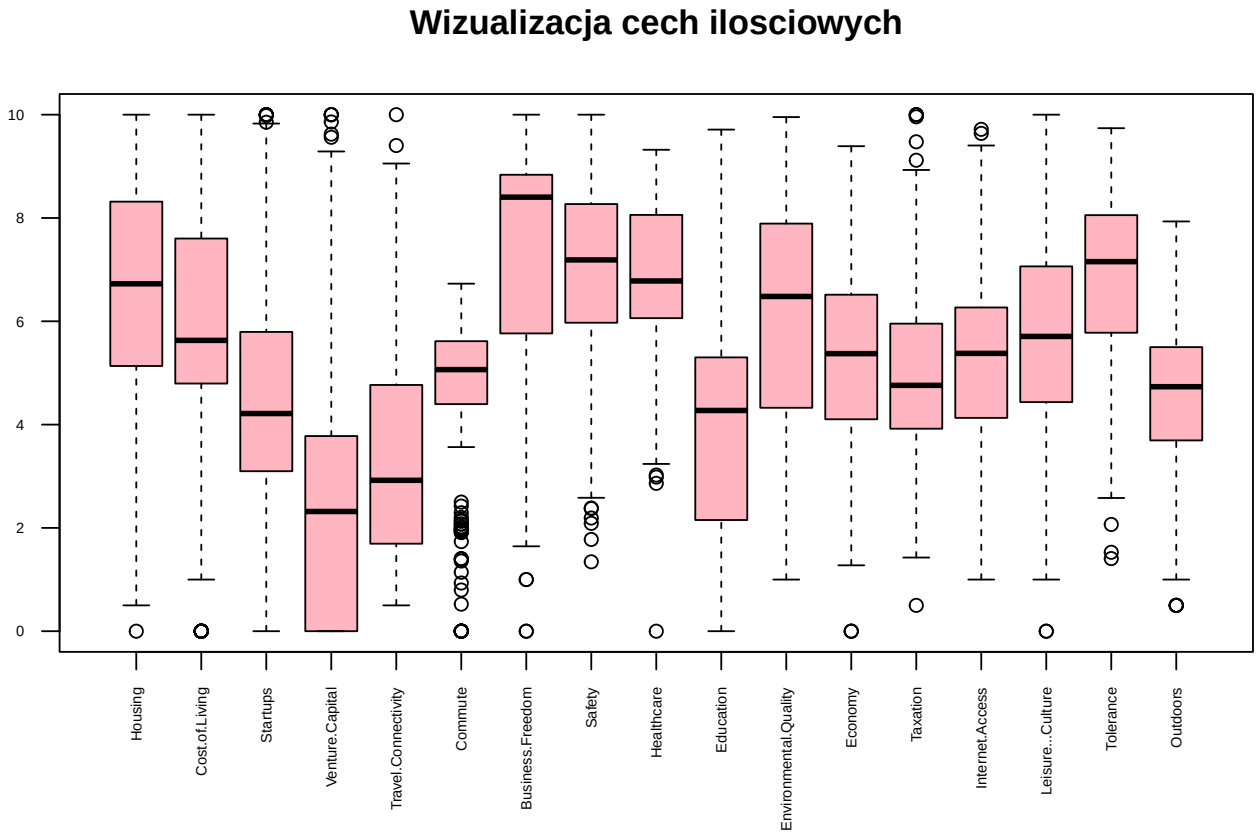
```
#Wyodębnienie zmiennych ilościowych
daneIlosciowe <- danePCA %>%
  select(where(is.numeric)) %>%
  select(-X)
```

Dodatkowo, aby móc efektownie zastosować metodę PCA musimy stwierdzić czy potrzebna jest standaryzacja naszych danych ilościowych. W tym celu wyliczymy ich wariację oraz porównamy je wizualnie.

Tabela 2: Wariancje cech ilościowych

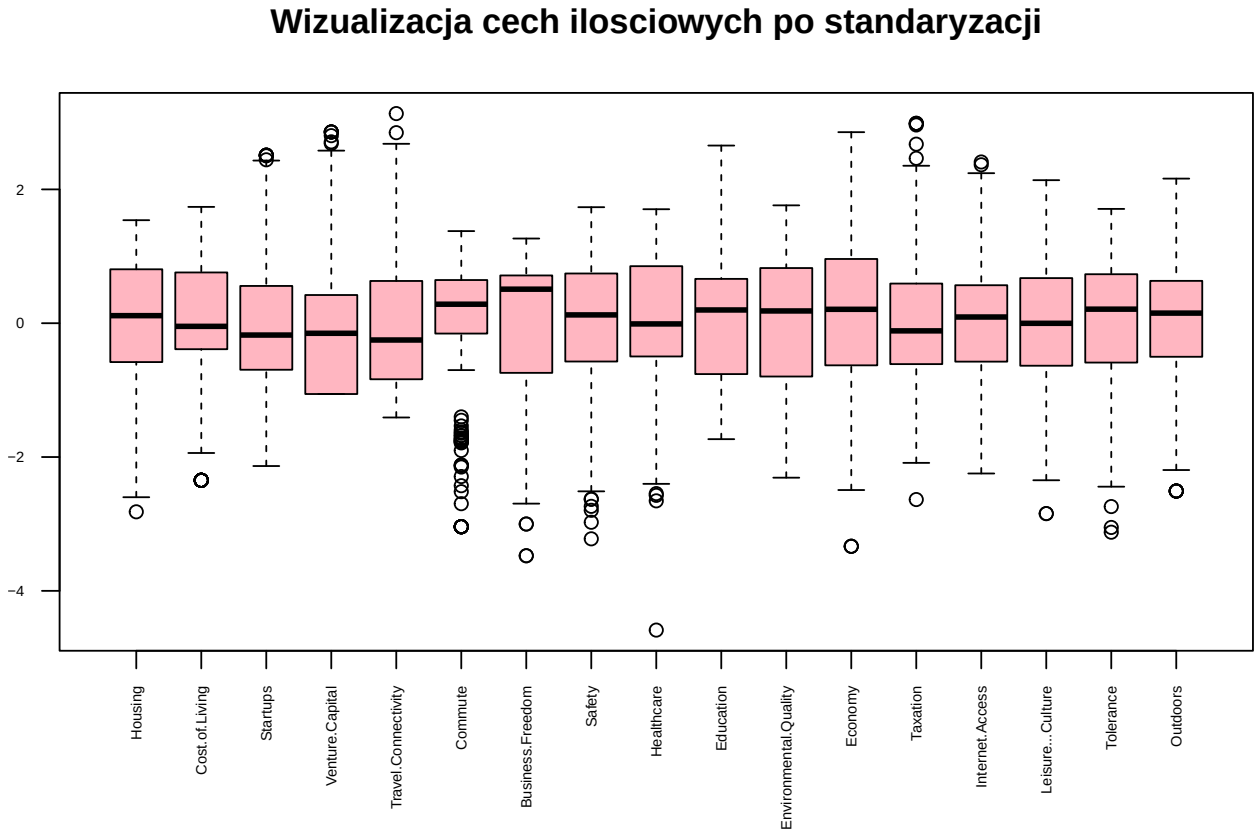
| Cecha | Wariancja |
|---|---|
| Housing | 5.265 |
| Cost.of.Living | 5.988 |
| Startups | 4.635 |
| Venture.Capital | 6.520 |
| Travel.Connectivity | 4.375 |

| | |
|---|---|
| Commute | 2.320 |
| Business.Freedom | 4.450 |
| Safety | 3.051 |
| Healthcare | 2.196 |
| Education | 4.897 |
| Environmental.Quality | 4.840 |
| Economy | 2.302 |
| Taxation | 2.855 |
| Internet.Access | 3.505 |
| Leisure...Culture | 4.027 |
| Tolerance | 2.974 |
| Outdoors | 2.534 |



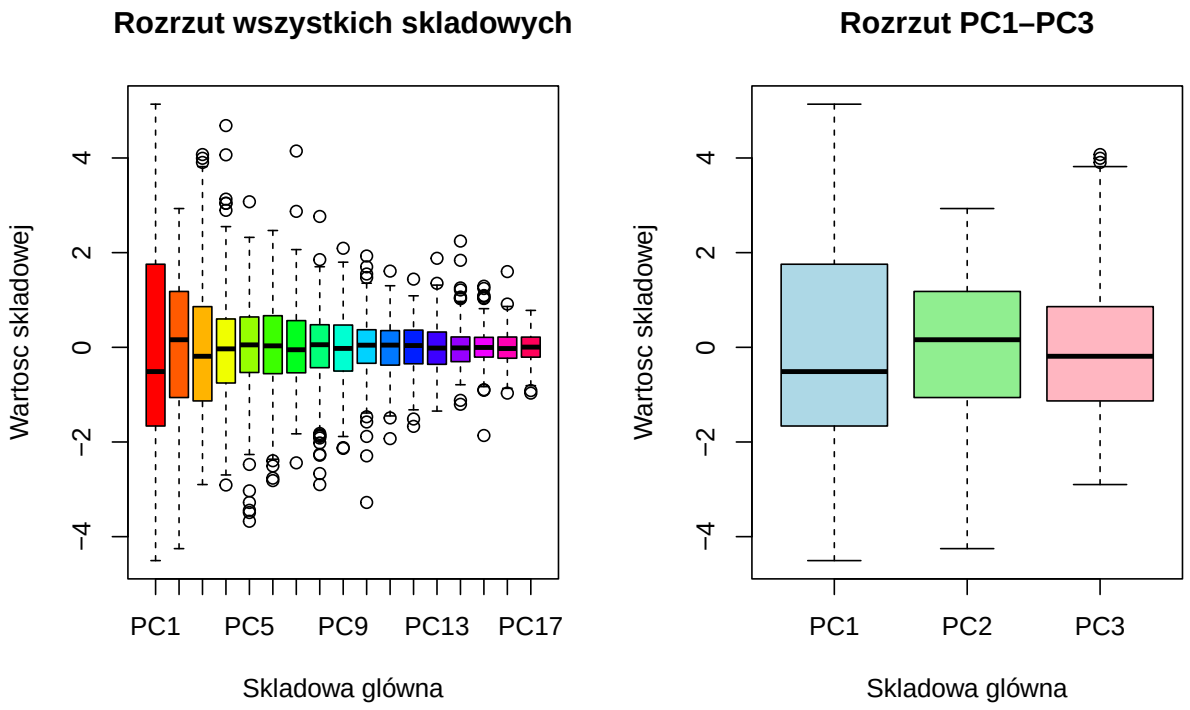

Rysunek 1: Wizualizacja cech ilościowych

Jak widać na Rysunku 1 oraz Tabeli 2 rozrzut (wariancja) między cechami ilościowymi jest chaotyczny, znacząco zróżnicowany - niektóre cechy mają duży zakres, a inne są mocno skupione wokół środka. Stąd, można stwierdzić, że należy przeprowadzić standaryzację danych.




Rysunek 2: Wizualizacja cech ilościowych po standaryzacji

Mając ustandaryzowane dane (jak na Rysunku 2) możemy przejść do wyzanczania, wyliczania oraz porównywania składowych głównych.




Rysunek 3: Rozrzut składowych głównych

Składowe główne są uporządkowane względem wariancji - PC1 ma charakteryzuje się największym rozrzutem, a PC17 najmniejszym, co jest wyraźnie ukazane na Rysunkach 3.

Przyjrzyjmy się głębiej interesującymi nas składowymi głównymi: PC1, PC2, PC3:

Tabela 3: Największe obciążenia zmiennych na PC1

| Zmienna | Loading |
|---|---|
| Education | -0.403 |
| Business.Freedom | -0.377 |
| Environmental.Quality | -0.326 |
| Housing | 0.308 |
| Healthcare | -0.280 |
| Internet.Access | -0.276 |

Tabela 4: Największe obciążenia zmiennych na PC2

| Zmienna | Loading |
|---|---|
| Startups | -0.483 |
| Venture.Capital | -0.427 |
| Leisure...Culture | -0.365 |
| Tolerance | 0.355 |

| | |
|---|---|
| Safety | 0.287 |
| Environmental.Quality | 0.253 |

Tabela 5: Największe obciążenia zmiennych na PC3

| Zmienna | Loading |
|---|---|
| Commute | -0.506 |
| Travel.Connectivity | -0.340 |
| Safety | -0.333 |
| Cost.of.Living | -0.331 |
| Housing | -0.314 |
| Economy | 0.309 |

**PC1**

Pierwsza składowa główna - PC1 (Tabela 3) wydaje się kontrastować miasta o dobrej sytuacji mieszkaniowej z tymi, które charakteryzują się wyższym poziomem rozwoju społeczno-ekonomicznego w innych obszarach. Wysokie wartości na PC1 wskazują na miasta, gdzie dostępność i warunki mieszkaniowe są relatywnie lepsze, ale może to iść w parze z niższymi wskaźnikami w edukacji, swobodzie gospodarczej, jakości środowiska, opiece zdrowotnej i dostępie do internetu. Z kolei niskie wartości PC1 sugerują miasta z gorszą sytuacją mieszkaniową, ale potencjalnie silniejszymi wynikami w wymienionych aspektach rozwoju.

**PC2**

Druga składowa główna - PC2 (Tabela 4 ogólnie odróżnia miasta o wyższym poziomie kapitału społecznego i jakości życia od tych z silniejszym ekosystemem innowacji i kultury. Wysokie wartości PC2 wskazują na miasta z większą tolerancją, bezpieczeństwem i lepszą jakością środowiska, ale potencjalnie z mniejszą aktywnością startupową i uboższą ofertą czasu wolnego o charakterze komercyjnym. Natomiast niskie wartości PC2 sugerują miasta z dynamicznym środowiskiem startupów, bogatą ofertą kulturalną i rozrywkową, ale mogą borykać się z niższym poziomem tolerancji, bezpieczeństwa i gorszą jakością środowiska.

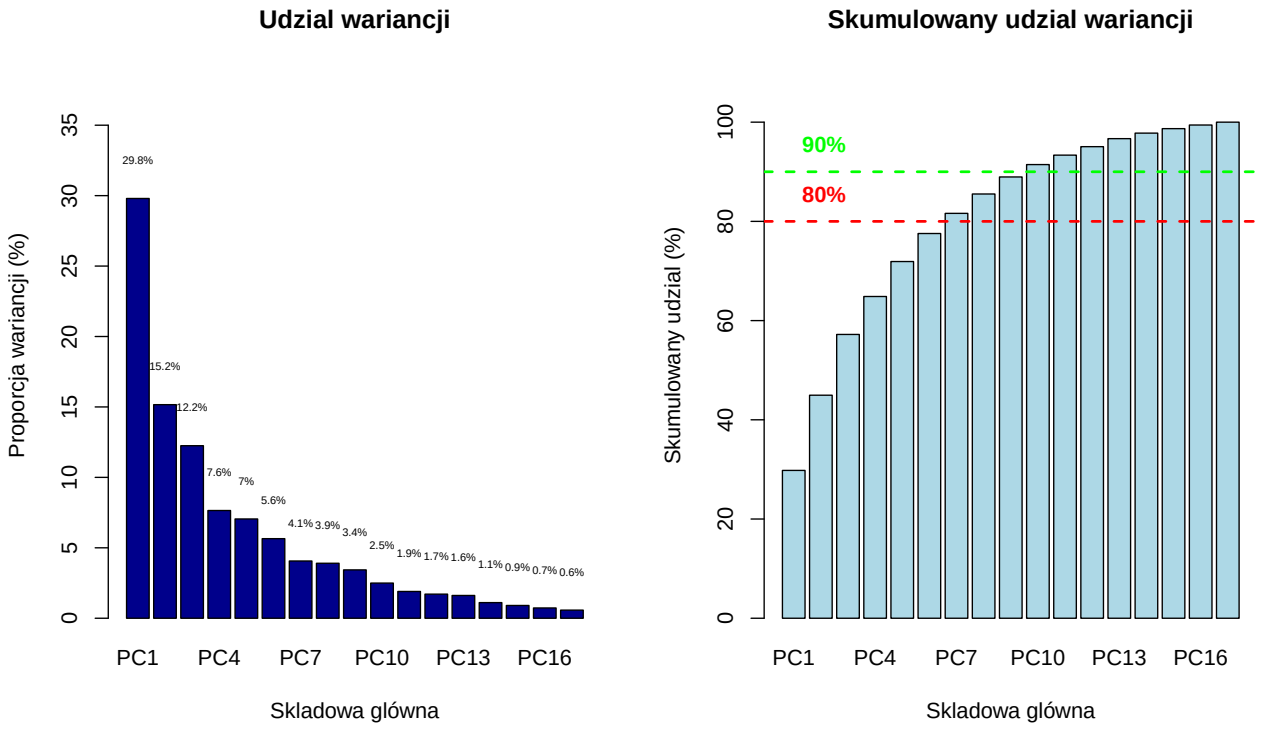
**PC3** Trzecia składowa główna - PC3 (Tabela5 zdaje się przeciwstawiać miasta z silną kondycją ekonomiczną tym, które oferują lepszą jakość życia pod względem codziennych doświadczeń i kosztów. Wysokie wartości na PC3 wskazują na miasta z dynamiczną gospodarką, ale mogą wiązać się z dłuższymi dojazdami, gorszą komunikacją zewnętrzną, niższym poziomem bezpieczeństwa, wyższymi kosztami życia i trudniejszą sytuacją mieszkaniową. Z kolei niskie wartości PC3 sugerują miasta o potencjalnie słabszej gospodarce, lecz z krótszymi dojazdami, lepszą łącznością, wyższym bezpieczeństwem, niższymi kosztami utrzymania i lepszym rynkiem mieszkaniowym.

## 1.2 Zmienność odpowiadająca poszczególnym składowym

Zbadajmy teraz, jaki procent wyjaśnionej wariancji (zmienności) odpowiada poszczególnym składowym.

Tabela 6: Udział wariancji wyjaśnionej przez składniki główne

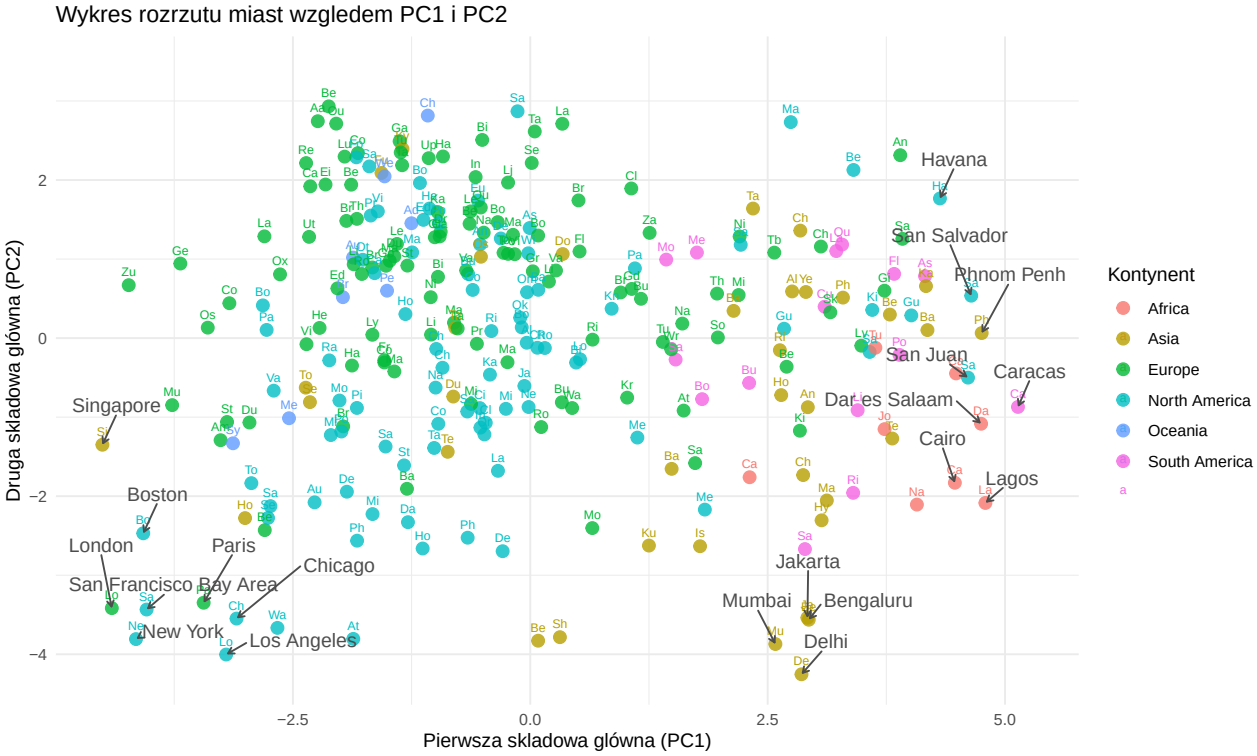
| Skladowa | Udzial | Skumulowany |
|---|---|---|
| PC1 | 29.80 | 29.80 |
| PC2 | 15.16 | 44.96 |
| PC3 | 12.25 | 57.21 |
| PC4 | 7.65 | 64.86 |
| PC5 | 7.05 | 71.90 |
| PC6 | 5.65 | 77.55 |
| PC7 | 4.06 | 81.62 |
| PC8 | 3.90 | 85.52 |
| PC9 | 3.43 | 88.95 |
| PC10 | 2.50 | 91.45 |
| PC11 | 1.90 | 93.35 |
| PC12 | 1.71 | 95.06 |
| PC13 | 1.62 | 96.68 |
| PC14 | 1.11 | 97.79 |
| PC15 | 0.91 | 98.69 |
| PC16 | 0.73 | 99.42 |
| PC17 | 0.58 | 100.00 |




Rysunek 4: Udział jednostkowy i skumulowany wariancji dla składowych głównych (%)

Jak możemy wywnioskować z Rysunków 4 oraz Tabeli 6 większość zmienności danych można wyjaśnić za pomocą niewielkiej liczby składowych. Już wcześniej przebadane 3 pierwsze składowe stanowią ponad połowę (bo 57,21%) całkowitej wariancji.

Liczba składowych potrzebna do wyjaśnienia 80% zmienności to 7. Liczba składowych potrzebna do wyjaśnienia 90% zmienności to 10.




Rysunek 5: Rozrzut miast względem skladowych

Rysunek 5 przedstawia rozrzut miast w przestrzeni dwóch pierwszych (o największej zmienności) składowych głównych PC1 (oś X) i PC2 (oś Y), reprezentujących dwa główne kierunki zmienności w analizowanych danych.

Miasta zaznaczone strzałkami to miasta najabardziej oddalone od pozostałej większości miast.

Havana (Kuba): Położona w prawym górnym kwadrancie (wysokie PC1, wysokie PC2), sugeruje dobrą sytuację mieszkaniową, wysoki poziom tolerancji i bezpieczeństwa oraz dobrą jakość środowiska. Jednakże, może to iść w parze z niższym poziomem edukacji, mniejszą swobodą gospodarczą, słabszą opieką zdrowotną, gorszym dostępem do internetu, mniejszą aktywnością startupową i uboższą ofertą czasu wolnego. Sytuacja polityczno-gospodarcza Kuby może tłumaczyć ograniczenia w gospodarce i technologii, przy jednoczesnych silnych więziach społecznych i unikalnej kulturze. Jakość środowiska i bezpieczeństwo mogą być relatywnie wysokie.

Caracas (Wenezuela): W prawym środkowym obszarze (wysokie PC1, lekko ujemne PC2), wskazuje na dobrą sytuację mieszkaniową (z zastrzeżeniami co do kryzysu), niższy poziom

tolerancji, bezpieczeństwa i jakości środowiska, ale potencjalnie wyższą aktywność startupową i bogatszą ofertę czasu wolnego. Interpretacja wymaga ostrożności ze względu na trudną sytuację w kraju.

Mumbai i Bengaluru (Indie): W dolnej prawej części (wysokie PC1, niskie PC2), sugerują dobrą sytuację mieszkaniową, niski poziom tolerancji, bezpieczeństwa i jakości środowiska, ale wysoką aktywność startupową, dostępność kapitału i bogatą ofertę czasu wolnego, szczególnie w technologicznym Bengaluru. Mumbai jako centrum finansowe również wykazuje te cechy. Podobnie Delhi (Indie) w tym samym obszarze, charakteryzuje się dobrą sytuacją mieszkaniową, niższym poziomem tolerancji i środowiska, lecz dynamicznym środowiskiem startupowym i kulturalnym.

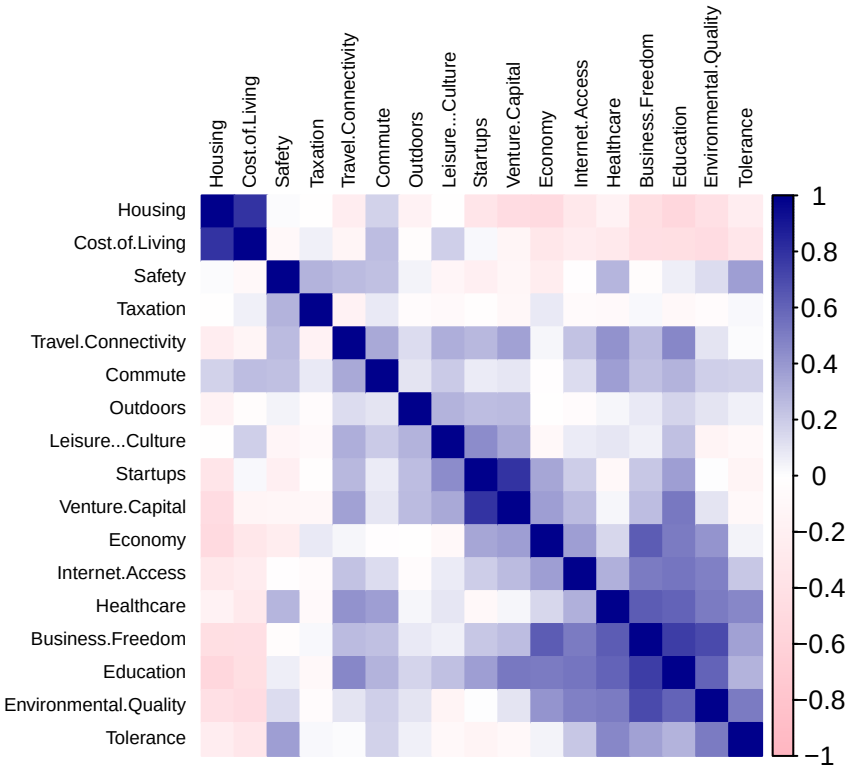
Singapur: Położony nisko i lekko na lewo (niskie PC1, niskie PC2), sugeruje relatywnie gorszą sytuację mieszkaniową, ale wyższy poziom edukacji, swobody gospodarczej, opieki zdrowotnej i dostępu do internetu. Niska wartość PC2 wskazuje na niższy poziom tolerancji i bezpieczeństwa oraz jakość środowiska, ale wysoką aktywność startupową i bogatą ofertę kulturalną. Wysokie koszty życia i nieruchomości w Singapurze kontrastują z doskonałą edukacją i gospodarką opartą na innowacjach.

Paryż (Francja): W lewym dolnym kwadrancie (niskie PC1, niskie PC2), wskazuje na gorszą sytuację mieszkaniową (choć lepszą niż Singapur), wysoki poziom edukacji, swobody gospodarczej i opieki zdrowotnej. Podobnie jak Singapur, niska wartość PC2 sugeruje niższy poziom tolerancji i środowiska, ale wysoką aktywność startupową i bogatą kulturę. Wysokie ceny nieruchomości w Paryżu idą w parze z silną gospodarką i bogatym życiem kulturalnym.

Los Angeles (USA): W lewym dolnym rogu (niskie PC1, bardzo niskie PC2), sugeruje znacząco gorszą sytuację mieszkaniową, ale wysoki poziom edukacji, swobody gospodarczej i opieki zdrowotnej. Bardzo niska wartość PC2 wskazuje na niski poziom tolerancji i środowiska, ale bardzo wysoką aktywność startupową i ofertę kulturalną. Wysokie koszty nieruchomości i problemy społeczne w Los Angeles kontrastują z dynamiczną gospodarką i przemysłem rozrywkowym.

## 1.3 Analiza korelacji

Zacznijmy od przeanalizowania macierzy korelacji, aby później porównać wnioski z dwuwykresem.




Rysunek 6: Macierz korelacji zmiennych

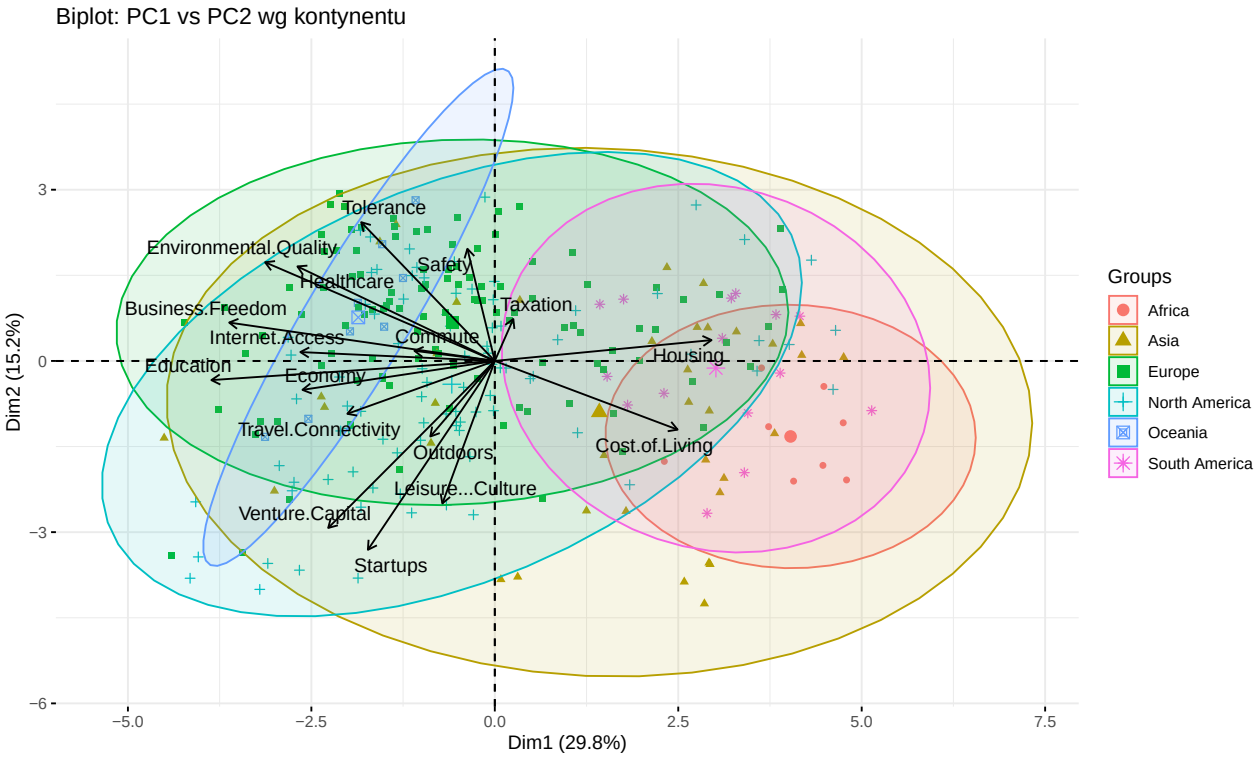
Analiza korelacji między cechami miast przedstawiona na wykresie ujawnia kilka istotnych zależności. Wysoki standard mieszkaniowy wiąże się z wyższymi kosztami życia, ale jednocześnie z niższym poziomem edukacji i gorszą kondycją ekonomiczną. Z kolei wysokie koszty życia idą w parze z niższą jakością środowiska.

Środowisko sprzyjające powstawaniu startupów jest silnie powiązane z dostępnością kapitału inwestycyjnego i w mniejszym stopniu z bogatą ofertą kulturalno-rozrywkową. Kapitał wysokiego ryzyka wykazuje również lekką pozytywną korelację z poziomem wykształcenia mieszkańców.

Komunikacja miejska koreluje słabo dodatnio z dostępem do opieki zdrowotnej i poziomem edukacji, podobnie jak dojazdy do pracy i opieka zdrowotna. Wolność prowadzenia działalności gospodarczej silnie koreluje z lepszą opieką zdrowotną, wyższym poziomem edukacji, lepszą jakością środowiska i dobrą kondycją ekonomiczną, natomiast słabiej i negatywnie z cenami mieszkań. Bezpieczeństwo wydaje się być powiązane z tolerancją.

Opieka zdrowotna jest dodatnio skorelowana z poziomem edukacji i jakością środowiska. Interesującym jest fakt, że wyższy poziom wykształcenia wiąże się z niższą jakością środowiska, słabszą gospodarką i mniejszym dostępem do internetu. Kondycja ekonomiczna nie wykazuje znaczących korelacji poza wspomnianymi. Opodatkowanie nie jest silnie powiązane z innymi zmiennymi, choć zauważalna jest lekka korelacja z bezpieczeństwem. Dostęp do internetu, oprócz negatywnej korelacji z edukacją, pozytywnie koreluje z wolnością gospodarczą i w mniejszym stopniu z jakością środowiska. Oferta kulturalno-rozrywkowa jest słabo

powiązana jedynie ze zmienną dotyczącą startupów. Tolerancja silnie koreluje z jakością środowiska i opieką zdrowotną. Aktywności na świeżym powietrzu wykazują słabe powiązania ze startupami, ofertą kulturalno-rozrywkową i kapitałem venture.




Rysunek 7: PC1 vs PC2 wg kontynentu

Z Rysunku 7 możemy wywnioskować, że miasta z poszczególnych kontynentów tworzą wyraźne grupy, co potwierdza istnienie podobieństw oraz różnic wewnątrz regionów, kontynentów.

Dwuwykres głównych składowych (PC1 i PC2), wyjaśniający łącznie 45% wariancji, prezentuje rozkład miast oraz wpływ poszczególnych zmiennych. Długie, zbliżone wektory zmiennych takich jak edukacja, wolność gospodarcza, jakość środowiska i opieka zdrowotna wskazują na ich silną, pozytywną korelację i sugerują wspólny czynnik rozwoju społeczno-ekonomicznego, który znacząco kształtuje PC1. Koszty życia znajdują się po przeciwnej stronie, co implikuje ich negatywny związek z tymi wskaźnikami.

Startup-y, kapitał wysokiego ryzyka oraz kultura tworzą odrębną grupę, wykazując silną korelację między sobą, ale słabszą z innymi zmiennymi społecznymi, co może odzwierciedlać niezależny wymiar przedsiębiorczości i aktywności miejskiej. Bezpieczeństwo jest powiązane z Tolerancją i Opeką Zdrowotną. Prostopadłe ustawienie wektorów, np. opodatkowania względem edukacji i startup-ów, sugeruje brak silnej korelacji. Długość wektorów wskazuje, że edukacja i wolność gospodarcza są lepiej reprezentowane w przestrzeni PC1 i PC2 niż opodatkowanie czy aktywności na świeżym powietrzu.

Analiza rozmieszczenia kontynentów ujawnia, że miasta Europy, Oceanii i Ameryki Północnej charakteryzują się wyższymi wartościami zmiennych społeczno-gospodarczych. Miasta Afryki

i część Azji są bliżej zmiennych związanych z mieszkaniami i kosztami życia. Miasta Ameryki Południowej grupują się w obszarze wysokich kosztów życia, cen mieszkań i opodatkowania, będąc jednocześnie oddalone od zmiennych takich jak edukacja, jakość środowiska czy wolność gospodarcza, co sugeruje odmienne priorytety lub wyzwania rozwojowe w tych regionach. Ogólnie, położenie kontynentów wskazuje na wyższy poziom rozwoju społecznego i infrastrukturalnego w Europie, Ameryce Północnej i Oceanii, w przeciwieństwie do części miast azjatyckich, południowoamerykańskich i afrykańskich, które mogą skupiać się na innych aspektach rozwoju.

Zarówno dwuwykres, jak i macierz korelacji dostarczają komplementarnych wniosków. Dwuwykres oferuje intuicyjną wizualizację struktury zależności, podczas gdy analiza korelacji zapewnia precyzyjne, liczbowe oceny powiązań między zmiennymi, tworząc razem pełniejszy obraz analizowanych danych.

## 1.4 Wnioski końcowe

Przeprowadzona w dokumencie analiza Principal Component Analysis (PCA) dostarczyła szeregu istotnych wniosków dotyczących struktury danych opisujących jakość życia w różnych miastach świata. Analiza ta pozwoliła na zredukowanie złożoności danych poprzez identyfikację głównych składowych, które w efektywny sposób wyjaśniają większość zaobserwowanej zmienności.

Jednym z kluczowych rezultatów analizy jest ujawnienie złożonych relacji między różnymi wskaźnikami jakości życia. Na przykład, zaobserwowano, że wysoki standard mieszkaniowy jest często powiązany z wyższymi kosztami życia, co jest zgodne z intuicją ekonomiczną. Jednakże, co ciekawe, może on również korelować z niższym poziomem edukacji i gorszą kondycją ekonomiczną, co sugeruje potencjalne nierówności społeczne. Z kolei środowisko sprzyjające rozwojowi startupów wykazuje silne powiązanie z dostępnością kapitału wysokiego ryzyka, co podkreśla kluczową rolę inwestycji w innowacyjność miejską.

Analiza PCA umożliwiła również identyfikację charakterystycznych profilów miast z różnych kontynentów. Miasta europejskie, zlokalizowane w Oceanii i Ameryce Północnej, generalnie charakteryzują się wyższymi wartościami wskaźników społeczno-gospodarczych, co wskazuje na ich relatywnie wysoki poziom rozwoju. Natomiast miasta afrykańskie i część azjatyckich często wykazują większy nacisk na wskaźniki związane z mieszkaniami i kosztami życia, co może odzwierciedlać odmienne priorytety rozwojowe lub wyzwania, z którymi się te regiony mierzą. Miasta Ameryki Południowej wyróżniają się skupieniem na wysokich kosztach życia, cenach mieszkań i opodatkowaniu, przy jednoczesnym oddaleniu od zmiennych takich jak edukacja czy jakość środowiska, co sugeruje specyficzne uwarunkowania ekonomiczne i społeczne tego regionu.

Kluczowym aspektem metodologicznym analizy PCA była konieczność standaryzacji danych. Zróżnicowany rozrzut i wariancja między analizowanymi cechami ilościowymi wymusiły zastosowanie tej procedury, aby zapewnić, że każda zmienna wnosi równy wkład do analizy. Standaryzacja umożliwiła obiektywne porównanie zmiennych o różnych skalach i jednostkach, co jest fundamentalne dla poprawnego działania PCA i interpretacji uzyskanych wyników.