Lista 2

Marta Hałas

2025-04-27

Spis treści

```r
setwd('C:/Users/mhala/Desktop/Eksploracja Danych')
data <- read.csv(file="C:/Users/mhala/Desktop/Eksploracja Danych/uaScoresDataFrame(2).csv",
                 stringsAsFactors = TRUE, header=TRUE)
library(dplyr)
library(VIM)
library(knitr)
library(DataExplorer)
library(ggrepel)
library(factoextra)
library(kableExtra)
library(corrplot)
library(ggplot2)
library(gridExtra)
```

1 Zadanie 2

1.1 a) Dane: City Quality of Life

1.1.1 O czym będzie analiza?

Na podstawie danych z City Quality of Life Dataset, analiza będzie dotyczyć porównania jakości życia w różnych miastach świata. Dane zawierają wiele wskaźników społecznych, ekonomicznych i środowiskowych opisujących warunki życia w miastach z różnych kontynentów.

1.1.2 Opis analizy

- Zastosowana zostanie redukacja wymiarowości za pomocą PCA - wyznaczenie składowych głównych.
- Wizualizacja podobieństw między miastami - czy są np. pogrupowane
- Identyfikacja zmiennych, które mają największy wpływ na rozróżnienie miast

1.2 b) Przygotowanie danych

1.2.1 Zapoznanie się z danymi

Rozmiar danych: liczba wierszy to 266 , liczba kolumn to 21.

Typy poszczególnych cech:

```
typy <- function(d) {
  data.frame(names(d), sapply(d, class))
}

typy(data)
```

```
##                                   names.d. sapply.d..class.
## X                                        X          integer
## UA_Name                            UA_Name           factor
## UA_Country                      UA_Country           factor
## UA_Continent                  UA_Continent           factor
## Housing                           Housing          numeric
## Cost.of.Living             Cost.of.Living          numeric
## Startups                         Startups          numeric
## Venture.Capital           Venture.Capital          numeric
## Travel.Connectivity   Travel.Connectivity          numeric
## Commute                           Commute          numeric
## Business.Freedom         Business.Freedom          numeric
## Safety                             Safety          numeric
## Healthcare                     Healthcare          numeric
## Education                       Education          numeric
## Environmental.Quality Environmental.Quality          numeric
## Economy                           Economy          numeric
## Taxation                         Taxation          numeric
## Internet.Access           Internet.Access          numeric
## Leisure...Culture       Leisure...Culture          numeric
## Tolerance                       Tolerance          numeric
## Outdoors                         Outdoors          numeric
```

UA__Name - zmienna ta została zmieniona na zmienną typu character.

UA__Country - zmienna ta została zmieniona na zmienną typu character.

UA__Continent - zmienna pozostaje zmienną typu factor. Posiada 6 kategorii - levels odpowiadających nazwom kontynentów. Zostaje też usunięta spacja przed każdą z nazw kraju.

```
data$UA_Name<-as.character(data$UA_Name)
data$UA_Country<-as.character(data$UA_Country)
data$UA_Country<-trimws(data$UA_Country)
```

Przydatność w analizie: Zostaje usunięta pierwsza kolumna - nie jest ona potrzebna (kolumna z indeksami - numeracja wierszy od 0)

```
data<-data[,-1]
```

Zmienne ciągłe

```
a<-split_columns(data)$continuous
names(a)
```

```
##  [1] "Housing"               "Cost.of.Living"        "Startups"
##  [4] "Venture.Capital"       "Travel.Connectivity"   "Commute"
##  [7] "Business.Freedom"      "Safety"                "Healthcare"
## [10] "Education"             "Environmental.Quality" "Economy"
## [13] "Taxation"              "Internet.Access"       "Leisure...Culture"
## [16] "Tolerance"             "Outdoors"
```

Zmienne dyskretne

```
b<-split_columns(data)$discrete
names(b)
```

```
## [1] "UA_Name"      "UA_Country"   "UA_Continent"
```

1.2.2 Brakujące obserwacje

Liczba NA - brakujących obserwacji - Liczba braków danych kodowanych za pomocą "NA" wynosi 0. Zakładamy zatem, że brakujące wartości w danych City Quality of Life Dataset są kodowane niestandardowo za pomocą wartości "0".

```
data1<-select_if(data,is.character)
data1[data1==0]<-NA
sum(is.na(data1))
```

```
## [1] 0
```

```
data1<-select_if(data,is.numeric)
data1[data1==0]<-NA
sum(is.na(data1))
```

```
## [1] 140
```

Brakujące dane są tylko dla zmiennych typu numeric.

```
data[data==0]<-NA
sum(is.na(data))
```

```
## [1] 140
```

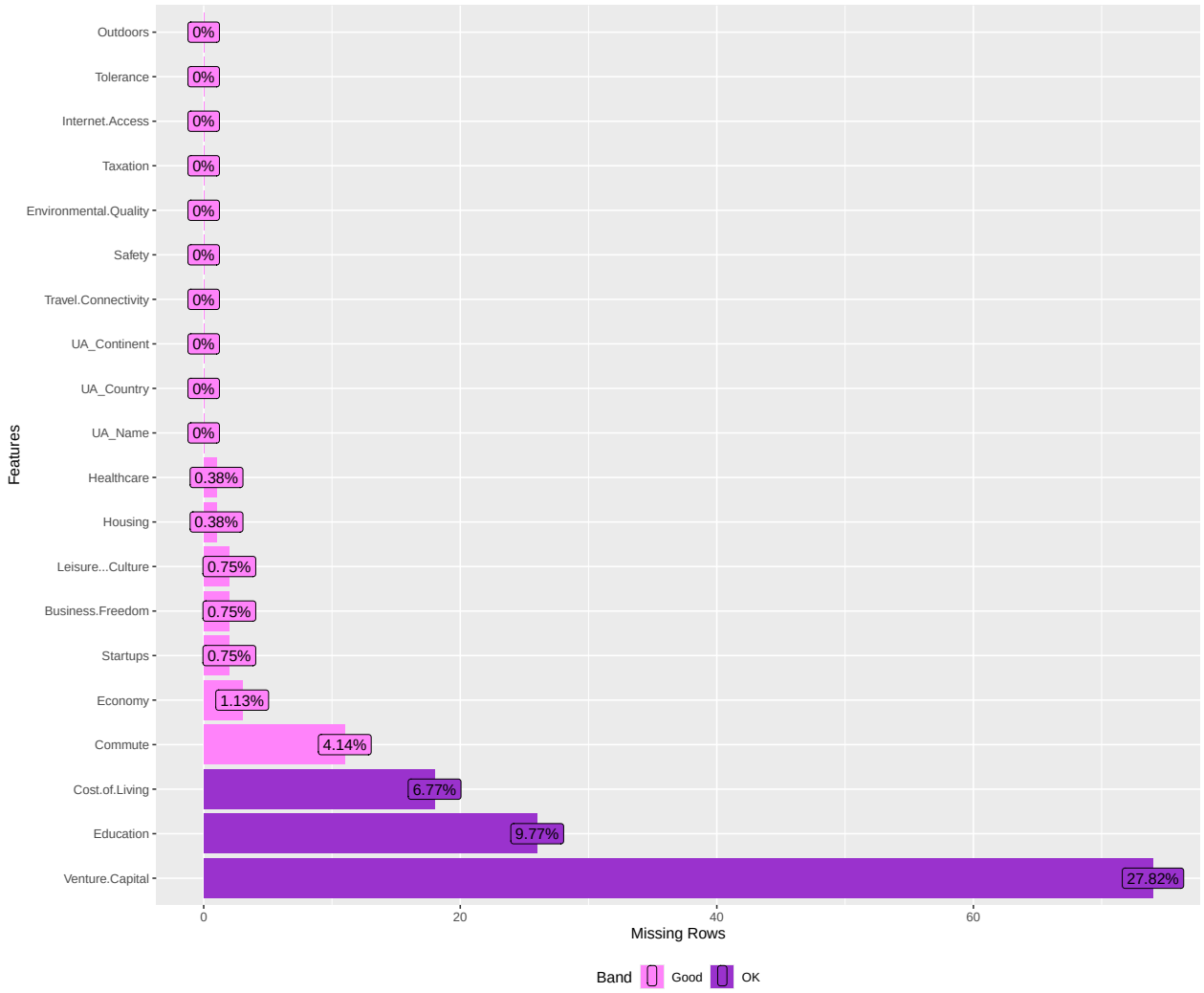
```r
tabela<-unique(which(is.na(data),arr.ind=TRUE)[,2])
tabela%>%
  kable(row.names=FALSE,caption="kolumny,w których występuje wartość NA",col.names="kolumny")
```

Tablica 1: kolumny,w których występuje wartość NA

kolumny
4
5
6
7
9
10
12
13
15
18

```r
p<-plot_missing(data)
```

```r
p + scale_fill_manual(values = c('orchid1','darkorchid3'))
```



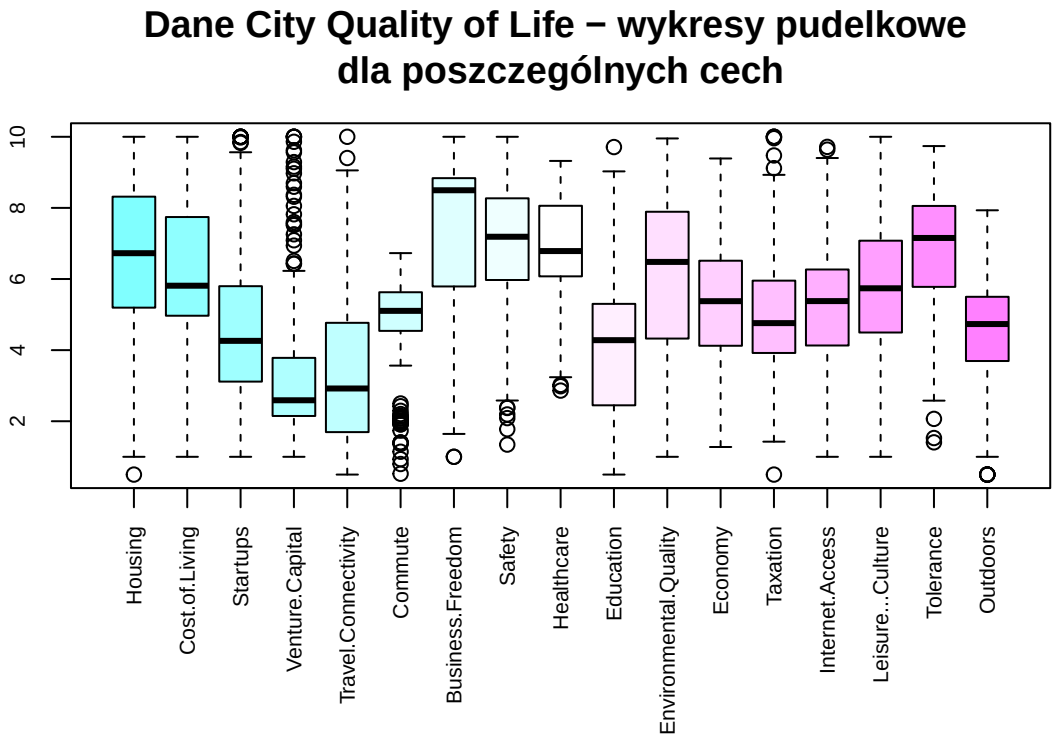
Pominięcie wartości brakujących, to znaczy usunięcie wierszy ich zawierających, powoduje utratę wielu obserwacji. Zatem zastosujemy metodę "KNN" - dla 5 sąsiadów.

```
data<-kNN(data,variable=colnames(data),k=5,imp_var = FALSE)
```

1.2.3 Zmienność (wariancja)

Porównainie zmienności (wariancji)

```
data1<-select_if(data,is.numeric)
par(mar = c(9, 5, 4, 2))
boxplot(data1, las=3, col=cm.colors(17),
        main="Dane City Quality of Life - wykresy pudełkowe \ndla poszczególnych cech",
        cex.axis =0.7)
```

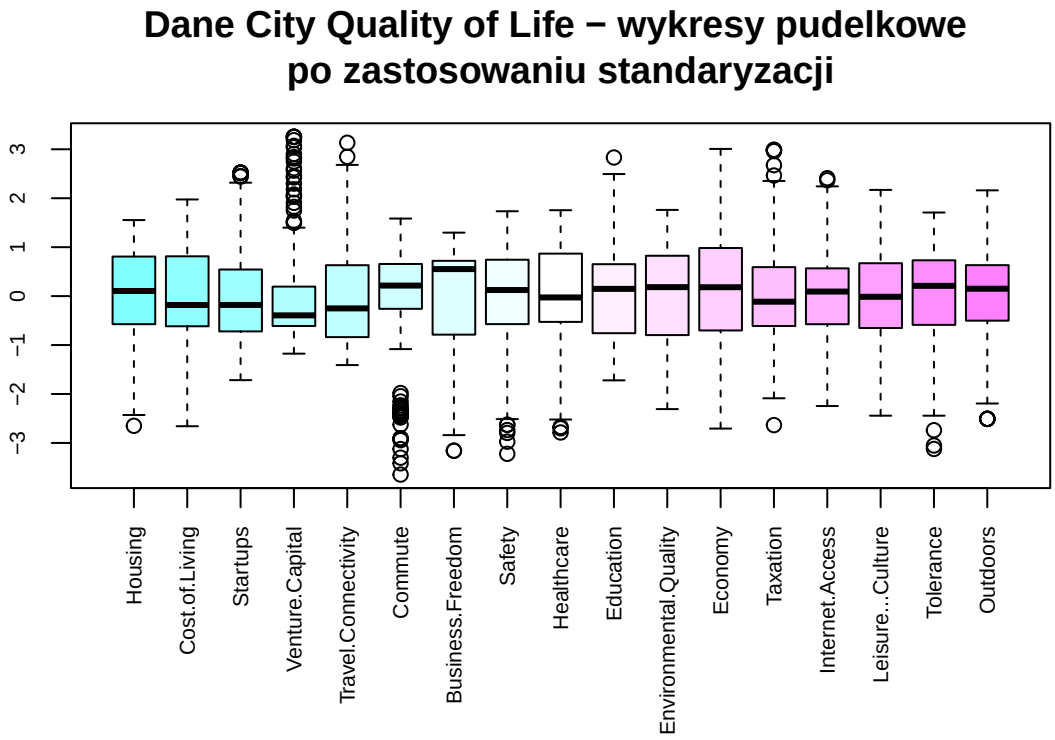


Na powyższym wykresie zauważamy dużą rozbieżność miedzy zakresami zmienności poszcególnych cech. Zakres cechy Commute rózni się znacząco od zakresu zmienności cechy Environmental.Quality, Business.Freedom czy Housing. Venture.Capital (zakres około 2-4), Safety oraz Healthcare (zakres około 6-8) mają również inne zakresy zmienności od siebie samych jak i od pozostałych cech.

Konieczne zastosowanie jest standaryzacji.

Standaryzacja

```
dane.stand <- scale(data1)
par(mar = c(9, 5, 4, 2))
boxplot(dane.stand, las=3,  col=cm.colors(17),
        main="Dane City Quality of Life - wykresy pudełkowe \npo zastosowaniu standaryzacji",
        cex.axis =0.7)
```



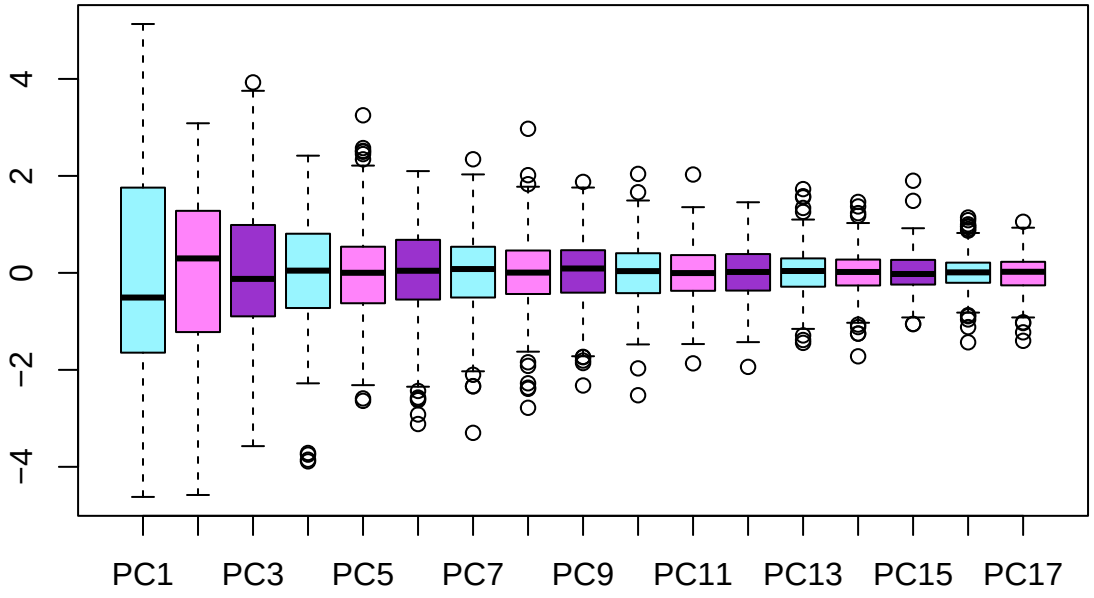
1.3 c) Wyznaczenie składowych głównych

1.3.1 Składowe główne - wizualizacja

Składowe główne

```
data1.pca <- prcomp(data1, scale.=TRUE, center=TRUE, retx=TRUE)

b<-data1.pca$x
boxplot(b, col=c('cadetblue1','orchid1','darkorchid3'))
title("Wykresy pudłkowe dla poszczególnych składowych głównych")
```

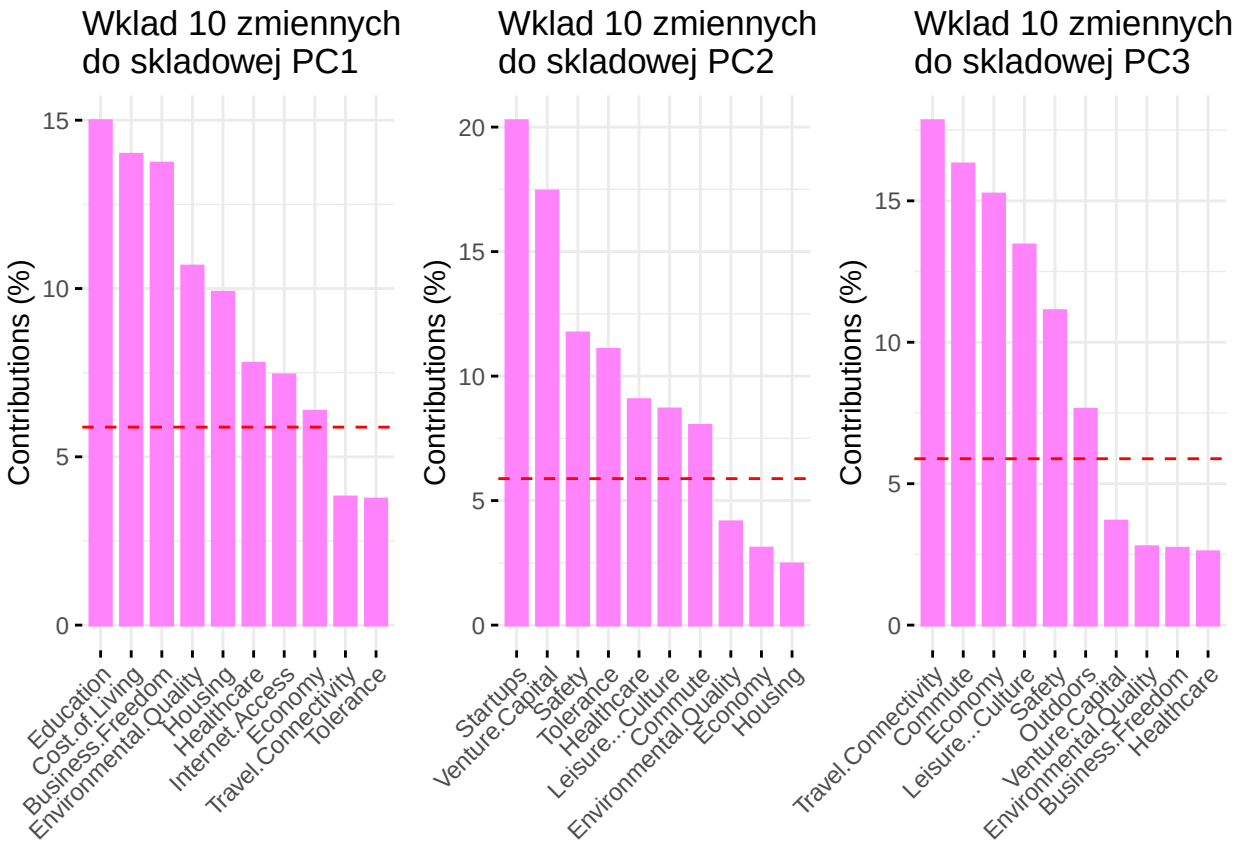


Największy rozrzut posiada składowa PC1 (najszersze "pudełko"), PC2 osiąga drugi największy,a PC3 trzeci największy rorzut. Kolejne składowe główne są reperezentowane przez "pudełka" o coraz mniejszej "szerokości", odpowiadającej malejącej zmienności danych, czyli mają coraz słabszy rozrzut.

1.3.2 Wartości wektorów ładunków, a składowe główne

Wektory ładunków

```
p1<-fviz_contrib(data1.pca, choice="var", axes=1, top=10, color="orchid1", fill="orchid1")+
  labs(title = "Wkład 10 zmiennych \ndo składowej PC1")
p2<-fviz_contrib(data1.pca, choice="var", axes=2,top=10,color="orchid1", fill="orchid1")+
  labs(title = "Wkład 10 zmiennych \ndo składowej PC2")
p3<-fviz_contrib(data1.pca, choice="var", axes=3,top=10,color="orchid1", fill="orchid1")+
  labs(title = "Wkład 10 zmiennych \ndo składowej PC3")
grid.arrange(p1, p2, p3, ncol = 3)
```




```
a<-round(data1.pca$rotation[,1:3],3)

kable(a)
```

	PC1	PC2	PC3
Housing	0.314	0.157	-0.098
Cost.of.Living	0.374	-0.022	-0.109
Startups	-0.158	-0.450	-0.160
Venture.Capital	-0.167	-0.418	-0.192
Travel.Connectivity	-0.195	-0.045	-0.422
Commute	-0.091	0.283	-0.404
Business.Freedom	-0.370	0.081	0.165
Safety	-0.045	0.343	-0.333
Healthcare	-0.279	0.301	-0.161
Education	-0.387	-0.046	-0.058
Environmental.Quality	-0.327	0.204	0.166
Economy	-0.252	-0.176	0.390
Taxation	0.028	0.094	0.030
Internet.Access	-0.273	0.007	0.117
Leisure. . . Culture	-0.070	-0.295	-0.367
Tolerance	-0.194	0.333	-0.041
Outdoors	-0.086	-0.144	-0.276

Najwiekszy wkład (wagę, najwyższe wartości wektorów ładunków) w składową PC1 mają zmienne:

- Education
- Cost.of.Living
- Businees.Freedom

Najwiekszy wkład (wagę, najwyższe wartości wektorów ładunków) w składową PC2 mają zmienne:

- Startups
- Venture.Capital
- Safety

Najwiekszy wkład (wagę, najwyższe wartości wektorów ładunków) w składową PC3 mają zmienne:

- Travel.Connectivity
- Commute
- Economy

Gdy wartości PC1 rosną, poziom edukacji oraz wolność gospodarcza (Business Freedom) maleją, a jednocześnie zauważamy niższe koszty życia (Cost.of.Living dostaje wyższe oceny w przeprowadzonej ankiecie).

Gdy wartości PC1 maleją, poziom edukacji oraz wolność gospodarcza (Business Freedom) wzrasta, a jednocześnie zauważamy wyższe koszty życia (Cost.of.Living dostaje niższe oceny w przeprowadzonej ankiecie).

Gdy wartości PC2 rosną, ilość startapów oraz kapitał inwestycyjny maleją, a jednocześnie zauważamy wzrost bezpieczeństwa.

Gdy wartości PC2 maleją, ilość startapów oraz kapitał inwestycyjny wzrasta, a jednocześnie zauważamy obniżenie poziomu poczucia bezpieczeństwa.

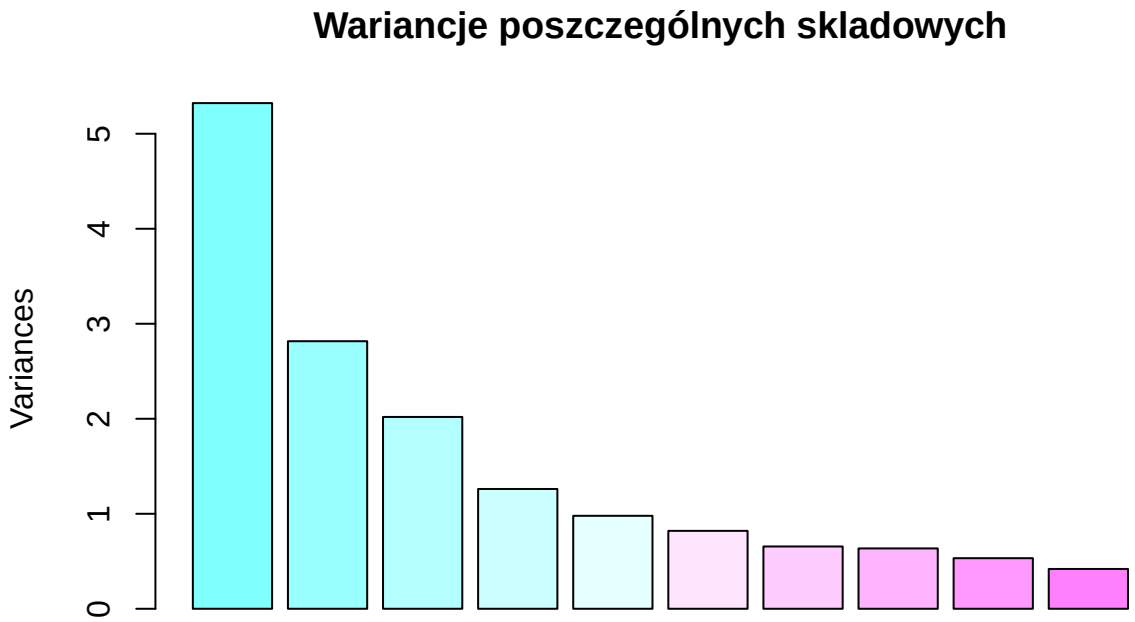
Gdy wartości PC3 rosną, dostępność transporu publicznego oraz czas spędzony na dojazdy (transport) maleją, a jednocześnie zauważamy polepszenie sytuacji gospodarczej.

Gdy wartości PC3 maleją, dostępność transporu publicznego oraz czas spędzony na dojazdy (transport) rosną, a jednocześnie zauważamy pogorszenie sytuacji gospodarzczej.

1.4 d) Zmienność odpowiadająca poszczególnym składowym

1.4.1 Wariancje poszczególnych skałdowych

```
plot(data1.pca, main="Wariancje poszczególnych składowych",col=cm.colors(10))
```

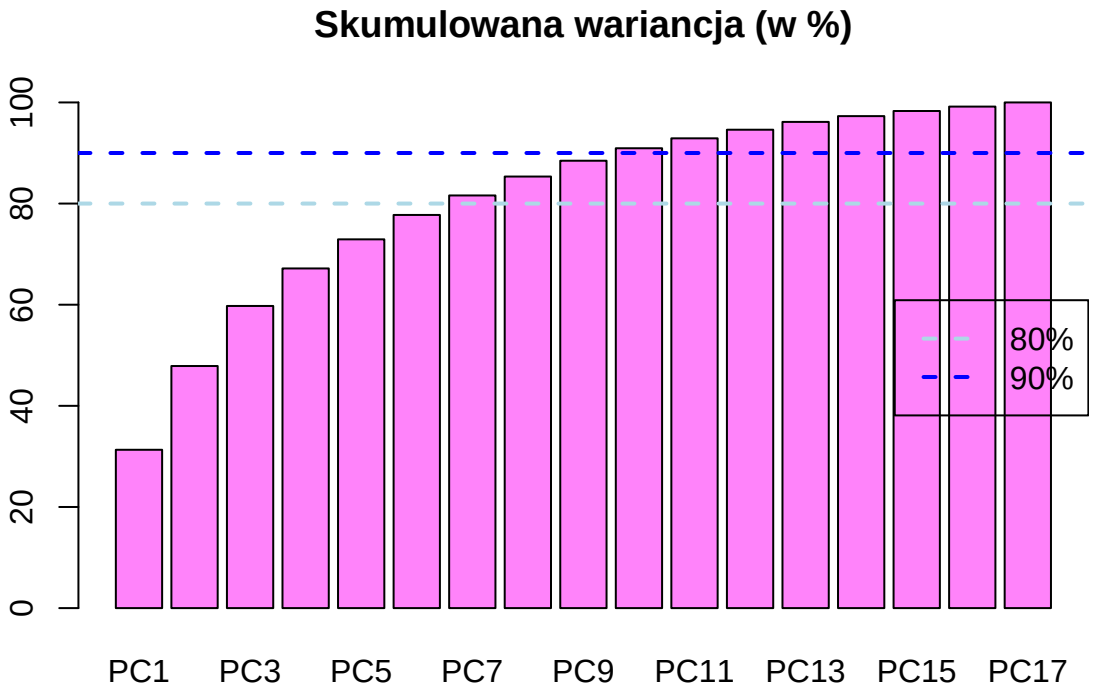


Składowa PC1 posiada największą wariancję, z kolejnymi składowymi wartość wariancji maleje, przechwytują one coraz mniejszą część zmienności danych.

1.4.2 Skumulowana wariancja w procentach

```
variance <- 100*(data1.pca$sdev^2)/sum(data1.pca$sdev^2)
cumulative.variance <- cumsum(variance)
```

```
barplot(cumulative.variance, main="Skumulowana wariancja (w %)",
names.arg=paste0("PC",1:17),col="orchid1")
abline(h=80, col="lightblue", lty=2, lwd=2)
abline(h=90, col="blue", lty=2, lwd=2)
legend("right", legend=c("80%","90%"), lwd=2, lty=2, col=c("lightblue","blue"),)
```



Około 30% całkowitej zmienności danych wyjaśńia składowa głowna PC1

Około 50% całkowitej zmienności danych wyjaśńiają składowe PC1 + PC2

Potrzebujemy:

a) Siedmiu pierwszych składowych głownych (PC1-PC7), aby wyjaśnić 80% całkowitej zmienności danych.

b) Dziesięciu pierwszych składowych głownych (PC1-PC10), aby wyjaśnić 80% całkowitej zmienności danych.

1.5 e) Wizualizacja danych wielowymiarowych

1.5.1 Miasta - powtarzalność nazw

Sprawdzam, czy wszystkie miasta posiadają unikatowe nazwy.

```
length(unique(data$UA_Name))
```

```
## [1] 264
```

```
duplikat<-data$UA_Name[duplicated(data$UA_Name)]
duplikat
```

```
## [1] "Birmingham" "Portland"
```

Dwa miasta mają swoich odpowiedników, dlatego dodam do nich nazwy krajów, aby móc je rozróżnić.

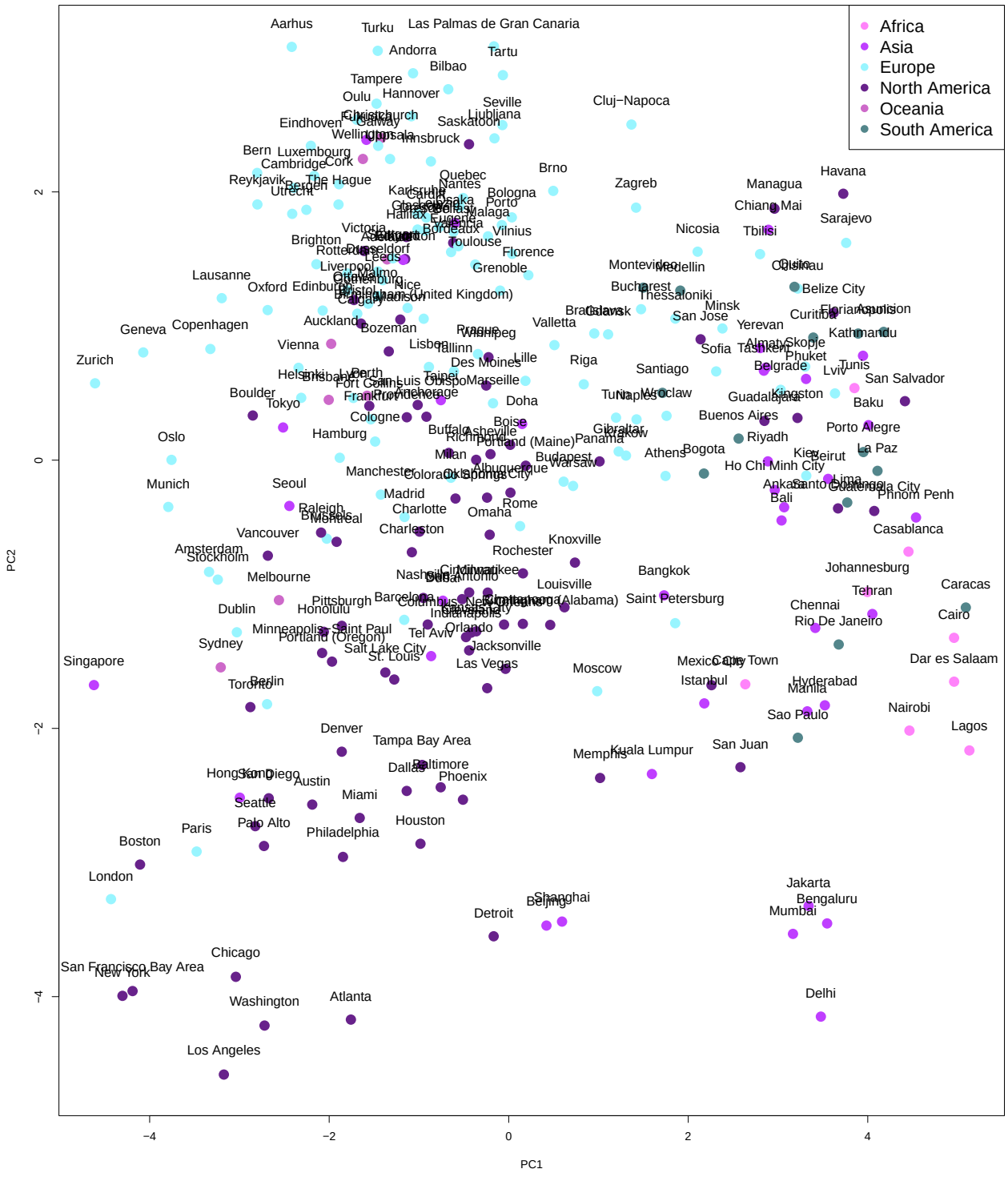
```r
miasta<-data$UA_Name
kontynent<-data$UA_Continent
kraj<-data$UA_Country
data$UA_Name<- ifelse(duplicated(miasta)| duplicated(miasta, fromLast = TRUE),
                              paste0(miasta, " (", kraj, ")"),
                              miasta)
```

1.5.2 Wizualizacja

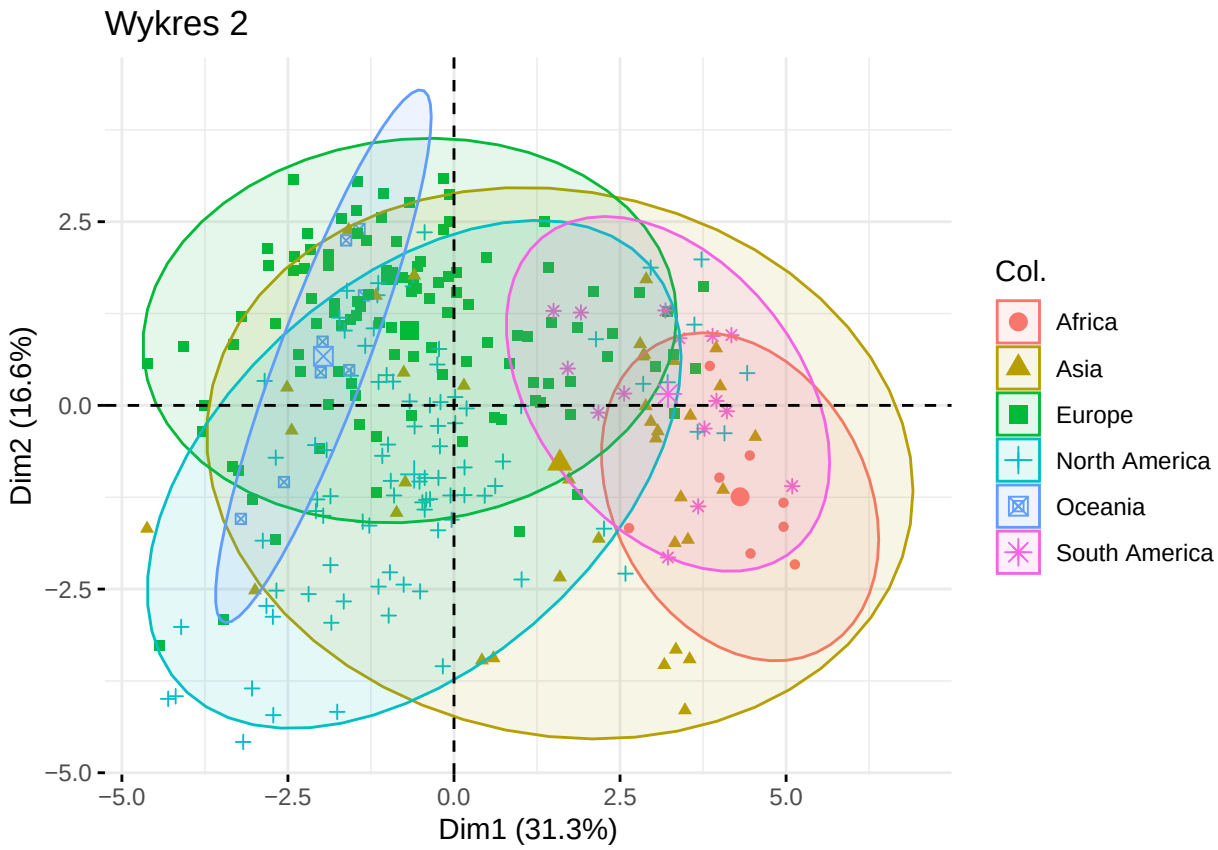
```r
data2<-as.data.frame(data1.pca$x[,1:2])

miasta<-data$UA_Name
kontynent<-data$UA_Continent
kraj<-data$UA_Country
data2$miasto<-miasta
data2$kraj<-kraj
data2$kontynent<-kontynent


kolory <- c("orchid1","darkorchid1","cadetblue1","darkorchid4",'orchid3','cadetblue4')
data2$miasto<-as.factor(data2$miasto)
data2$kontynent<-as.factor(data2$kontynent)
plot(data1.pca$x[,1], data1.pca$x[,2], col=kolory[as.numeric(data2$kontynent)],
     pch=16, xlab="PC1", ylab="PC2", cex=2)
title("Dane City Quality of Life -  wykres rozrzutu 2D - Wykres 1", cex.main=2)
text(data1.pca$x[,1], data1.pca$x[,2]+0.175, labels=data2$miasto, cex=1.2)
legend("topright",
       legend = levels(data2$kontynent),
       col = kolory,
       pch = 16,
       cex = 1.5,
       ncol = 1, text.width = 1.3)
```

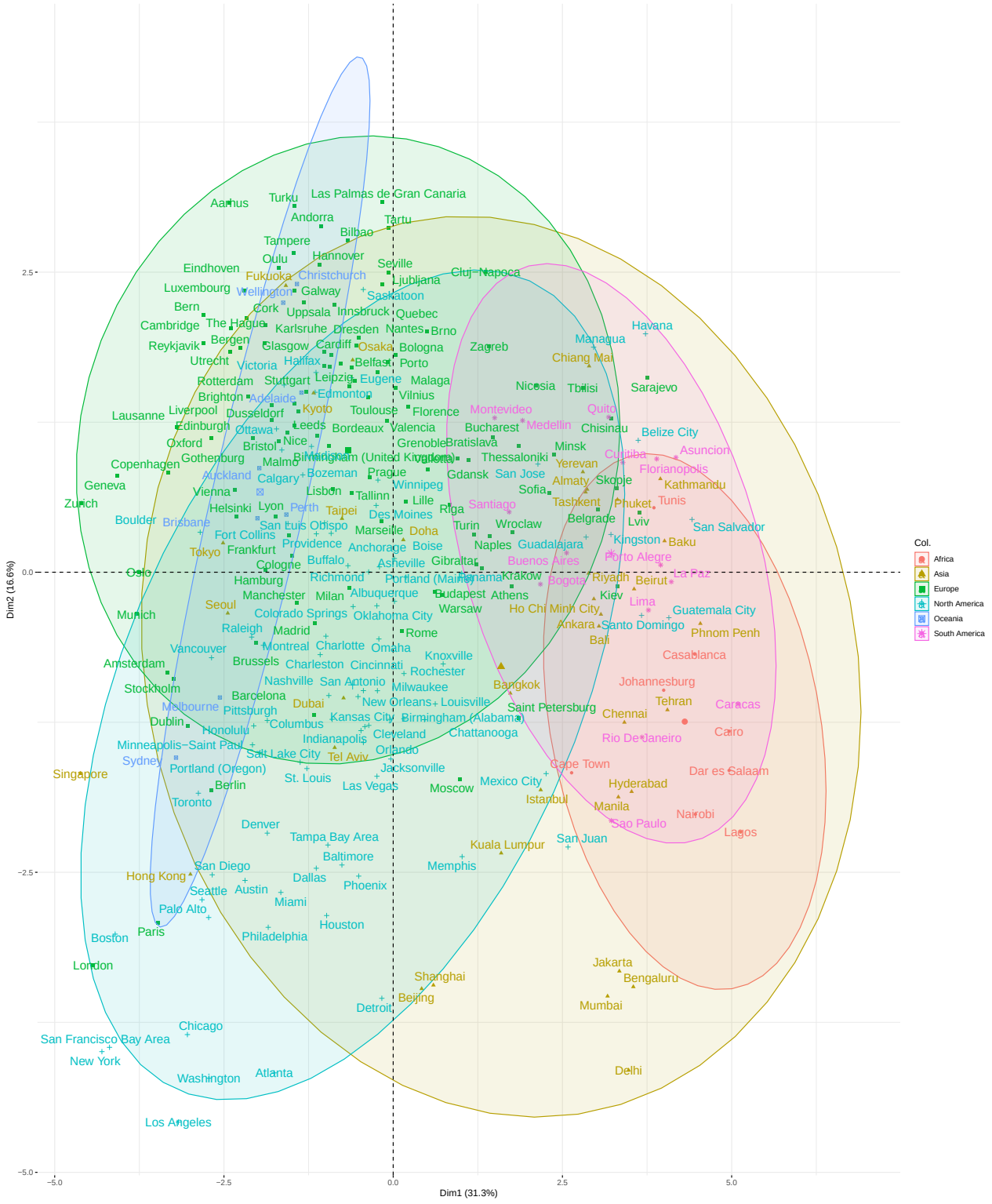



```
fviz_pca_ind(data1.pca, col.ind=data2$kontynent, addEllipses=TRUE, ellipse.level=0.9,
             label="var", repel=TRUE)+
  ggtitle("Wykres 2")
```



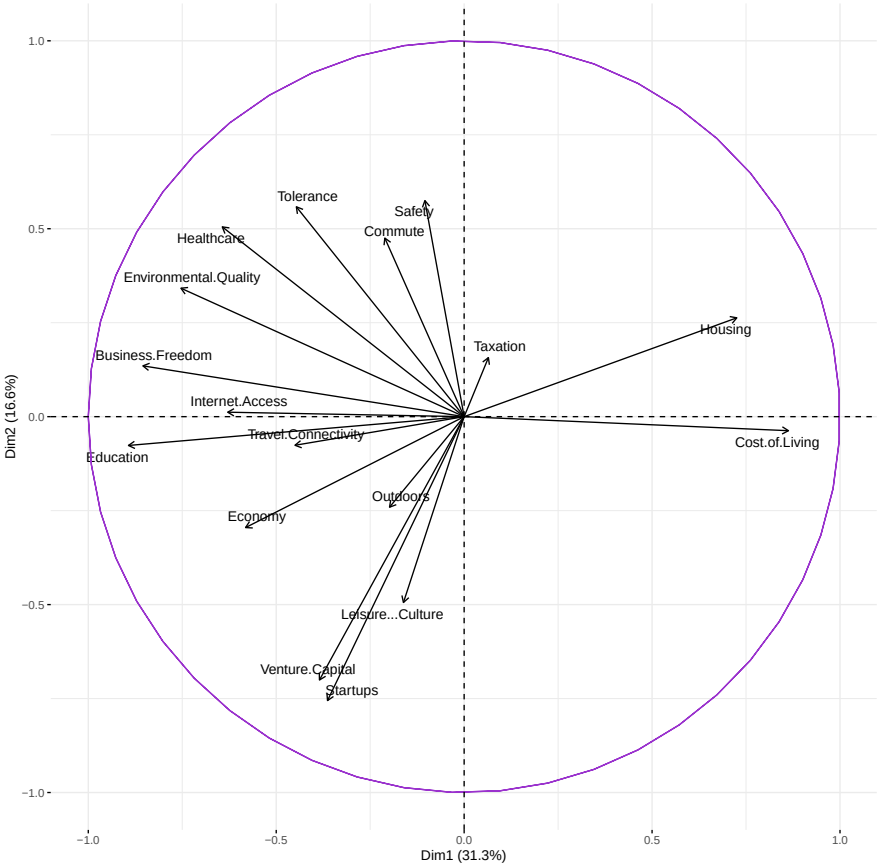

```
fviz_pca_ind(data1.pca, col.ind=data2$kontynent, addEllipses=TRUE, ellipse.level=0.9,
             label="var", repel=TRUE)+
  geom_text_repel(aes(label = data2$miasto, colour = kontynent), max.overlaps = 266,
                  segment.color = NA,size=5, point.padding=1)+
  ggtitle("Wykres 3")+ theme(plot.title = element_text(size = 40))
```

Wykres 3




```
fviz_pca_var(data1.pca,labelsize=4,repel=TRUE,col.var = "black",col.circle = "darkorchid3")+
  ggtitle("Wykres 4")+ theme(plot.title = element_text(size = 30))
```

Wykres 4



1.5.3 Podobieństwa i różnice

- Miasta położone blisko siebie na wykresie 1 mają zbliżone wartości analizowanych cech (duże podobieństwo). Miasta odległe od siebie (duży dystans między punktami na wykresie) charakteryzują się dużymi róznicami, zmiennością pod względem wartości tych cech (słabe podobieństwo).
- Na podstawie wykresu 2 możemy stwierdzić, że obiekty (miasta) układają się w naturlany sposób w grupy (kontynety). Elipsy na wykresie przedstawiają obszary, w których znajduje się 90% obiektów danej grupy. Są oczywiście obiekty (miasta), które "oddstają", wykraczają poza ramy elipsy danej gurpy, ale jest ich nie wiele.
- Najbardziej różniace się miasta od pozostałych (wykres 3):

```
- Delhi
- Los Angeles
- Washington
- Las Palmas de Gran Canaria
- Lagos
- Singapore
```

Lagos posiada największy wskaźnik PC1, natomiast Singapore najmniejszy.

Las Palmas de Gran Canaria posiada największy wskaźnik PC2, natomiast Los Angeles najmniejszy. Również małym wskaźnikiem PC2 charakteryzują się miasta takie jak Washington, Atlanta czy Dheli.

Korzystając z wykresu 4 możemy scharakteryzowac wybrane miasta.

Lagos - Singapore

Koszt życia jest tańszy oraz Housing - lepsze warunki mieszkaniowe, dostępność miekszań jest lepsza w Lagos. Podatki też będą nieco niższe w Lagos niż w Singaporze.

Jeśli natomiast pod uwagę wezmiemy aspekt poziomu edukacji, dostępu do internetu, Ekonomii, Business Freedom (wolności gospodarczej) czy dostępność transportu to o wiele lepiej plasuje się Singapore niż Lagos.

Singapore jest bezpieczniejszy oraz bardziej tolerancyjny niż Lagos.

Startups, Venture.Capital oraz Leisure... Culture są lepiej rozwinięte w Singaporze

Los Angeles - Las Palmas de Gran Canaria

Los Angeles dominuje Las Palmas de Gran Canaria pod względem Leisure... Culture (dostępność do oferty kulturowo-rozrywkowej), Venture Capital (dostępność do inwestycji w przedsiębiorstwa niepubiczne) oraz Startups (dostępność do innowacji, powstaje więcej nowych firm -starupów). Las Palmas de Gran Canaria cechuje się natomiast wysoką tolerancją (Tolerance), wysokim poziomem bezpieczeństwa, oraz wysoką jakością Commute - czas na dojazdy, jakość jazdy w porównaniu do Los Angeles.

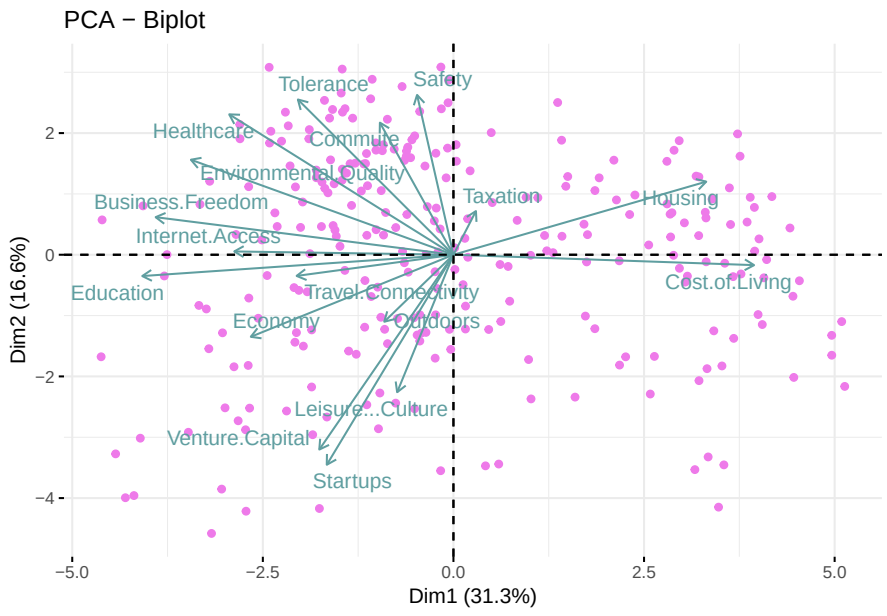
Opieka medyczna, jakość środowiska, podatki (tzn. tańsze) są również lepsze w Las Palmas de Gran Canaria.

Ekonomia jest lepiej rozwinięta w Los Angeles.

1.6 f) Korelacja zmiennych

1.6.1 Biplot

```
variables <- get_pca_var(data1.pca)
fviz_pca_biplot(data1.pca,label="var", repel=TRUE, col.ind="orchid2", col.var = "cadetblue" )
```



Istotna korelacja występuje między zmiennymi:

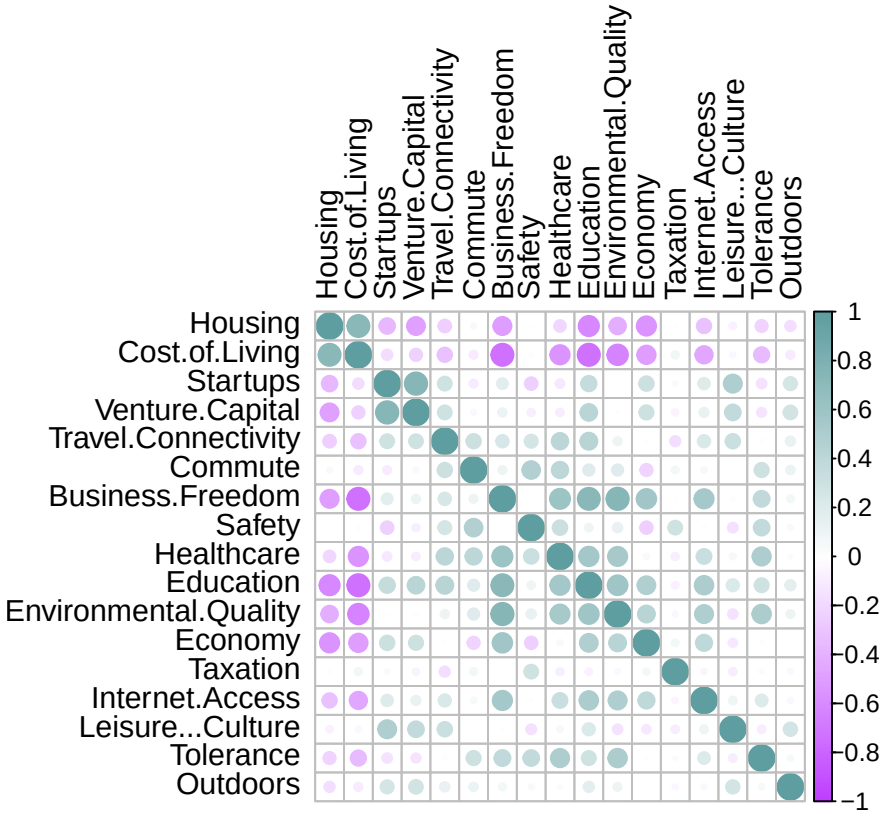
- Housing - Cost.of.Living

- Education - Buisness.Freedom
- Business.Freedom - Environmental.Quality
- Venture.Capital - Startups
- Business.Freedom - Education

Wskazuje na nie podobna długość strzałek (wektorów) na wykresie, mały kąt nachylenia między nimi i ten sam kierunek, zwrot.

1.6.2 Macierz kowariancji

```
correlation.matrix <- cor(data1)
corrplot(correlation.matrix,col = colorRampPalette(c("darkorchid1", "white", "cadetblue"))(200),tl.col=
```

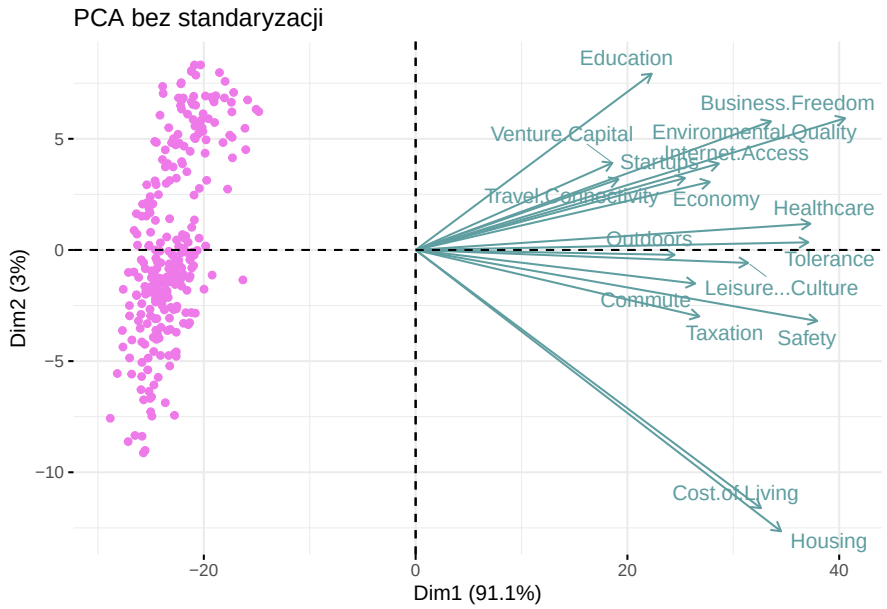


Powyższy wykres macierzy kowariancji potwierdza poprawność wniosków na temat istotnej koleracji między zmiennymi z przeprowadzonej analizy wykresu biplot.

1.7 g) Końcowe wnioski

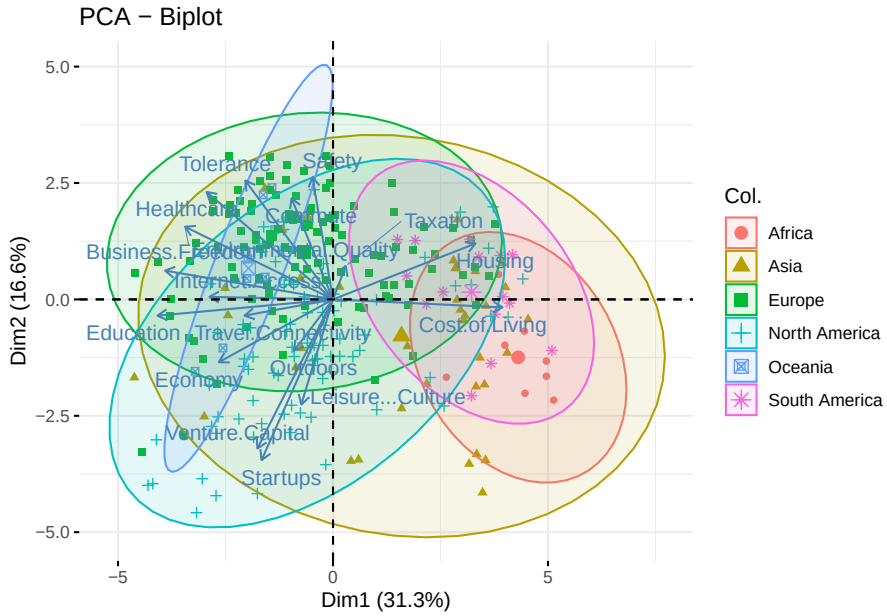
Biplot bez standaryzacji

```
pca <- prcomp(data1, scale. = FALSE, center=FALSE)
fviz_pca_biplot(pca, repel = TRUE, title = "PCA bez standaryzacji",label="var",col.ind="orchid2", col.va
```



- Powyższy wykres przedstawia biplot bez zastosowanej standaryzacji. Stwierdzamy zatem, że niezastosowanie standaryzacji miałoby istotny wpływ na otrzymane wyniki i wnioski. Standaryzacja jest konieczna.
- Do reperzentacji danych użyłam dwóch składowych PC1 i PC2, odpowiadają one za ~50% całej zmienności. Zatem możemy uznać tą reprezentację za zadowalającą.

```
fviz_pca_biplot(data1.pca,label="var", col.ind=data2$kontynent,repel=TRUE,addEllipses=TRUE, elipse.level
```



- Obiekty (miasta) tworzą grupy odpowiadające kontynetom. Istnieją zależności między składowymi PC1, PC2 a grupami kontynentów:
  - Niski koszt życia, słabo rozwinięta edukacja w Afryce i Południowej Ameryce.

- Tolerancja, bezpieczeństwo i opieka medycza były wyżej oceniane w Europie.
- Częstrze powstawanie firm (Startups) i lepszą dostępność do środków inwestycyjnych (Venture.Capital) wyróznia Północną Amerykę.

Z poprzednich punktów wynika również, że miasta, w których dostępność miszkań (Housing) jest wysoka mają niskie koszty życia.

Z kolei te w których wyosko oceniany jest Business.Freedom mają wysoki poziom edukacji i mają niezanieczyszczone środowisko.

Analogicznie rozwój Stratupów w miastach idzie w parze z Venture.Capital.

Ciekawą obserwacją jest, że miasta w Azji są skrajnie zróżnicowane, nie występuje między nimi podobieństwo. Np. Dheli wysoka wartość PC1 (dodatnia), ale ujemna PC2, a Fukuoka przyjmuje wartość ujemną PC1, ale dodatnią PC2.