Spis treści

Rozdział 1

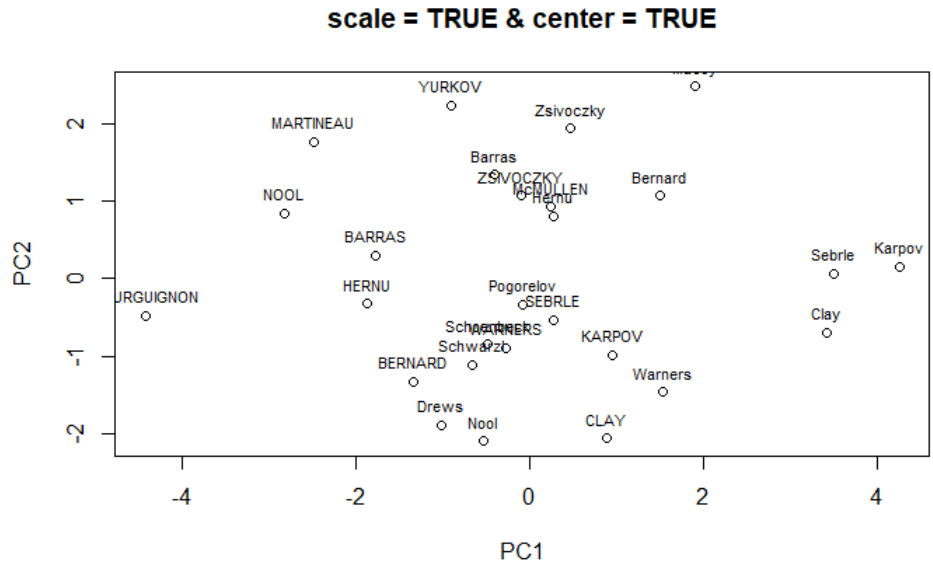
Wstęp

W ramach sprawozdania z drugiego laboratorium przetestowano działanie funkcji *prcomp*() oraz *princomp*(), co zostało przedstawione w rozdziale 2. Następnie, w rozdziale 3, przeprowadzono redukcję wymiarowości metodą PCA. Redukcja wymiarowości została przeprowadzona na danych [1]. Dane z zadania pierwszego zostały wybrane ponownie w celu sprawdzenia czy redukcja wymiarowości pozwoli na wizualizację niewykrytych we wcześniejszej analizie zależności. W rozdziale 4 podsumowano analizę.

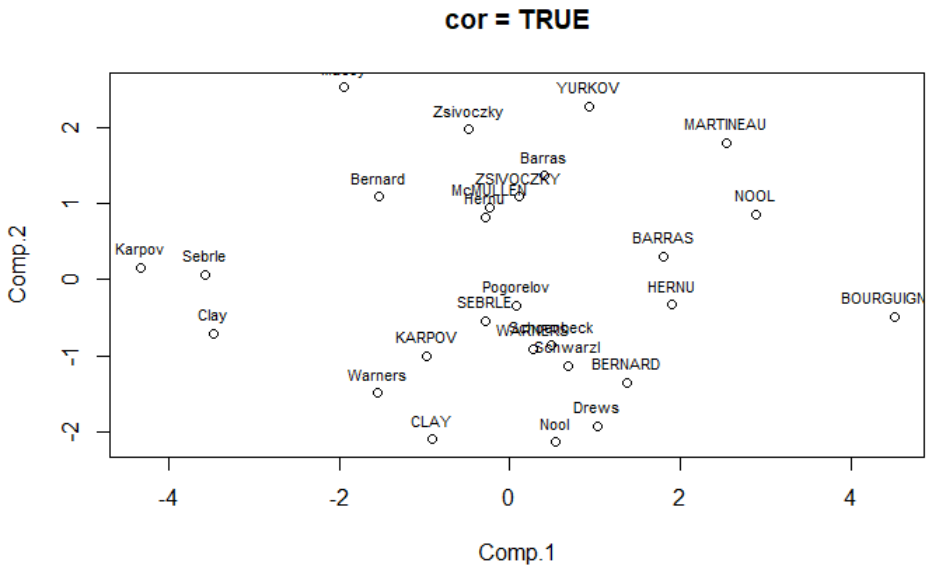
Rozdział 2

Testy funkcji *prcomp*() oraz *princomp*()

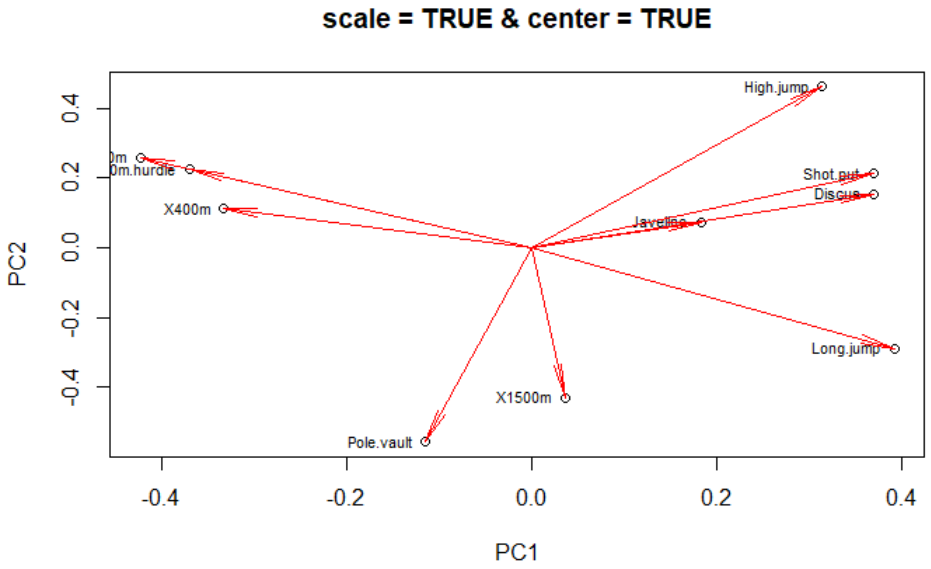
Pierwszy etap testowania polegał na przetestowaniu funkcji *prcomp*() z parametrami *scale* i *center* mającymi wartość *TRUE* oraz funkcji *princomp*() z parametrem *cor* mającym wartość *TRUE*. Wykresy obserwacji dla funkcji *prcomp*() oraz *princomp*() zostały przedstawione kolejno na 2.1 i 2.2. Jedyna różnica polega na fakcie, że wykres dla funkcji *princomp*() jest lustrzanym odbiciem wykresu dla funkcji *prcomp*(). Wykresy zmiennych dla tych funkcji także stanowią swoje lustrzane odbicie, co zostało pokazane na 2.3 i 2.4.



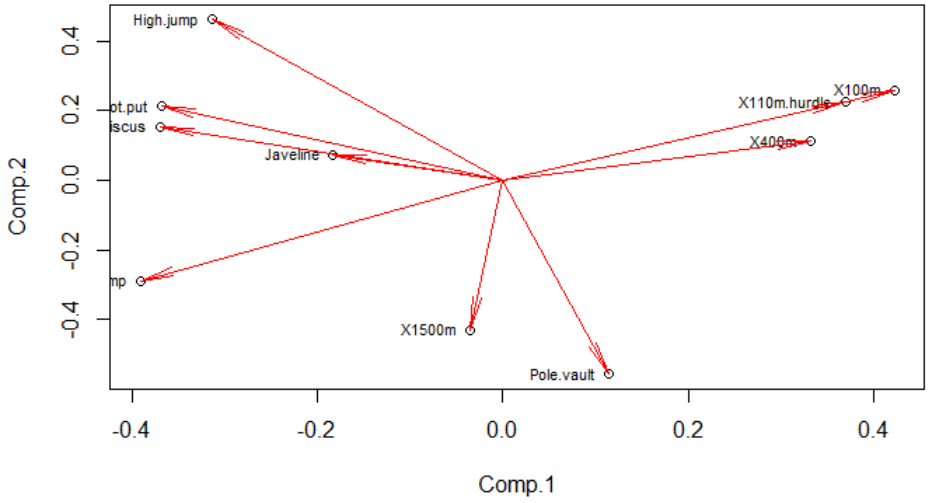
Rysunek 2.1: Wykres obserwacji dla funkcji *prcomp*().



Rysunek 2.2: Wykres obserwacji dla funkcji *princomp*().

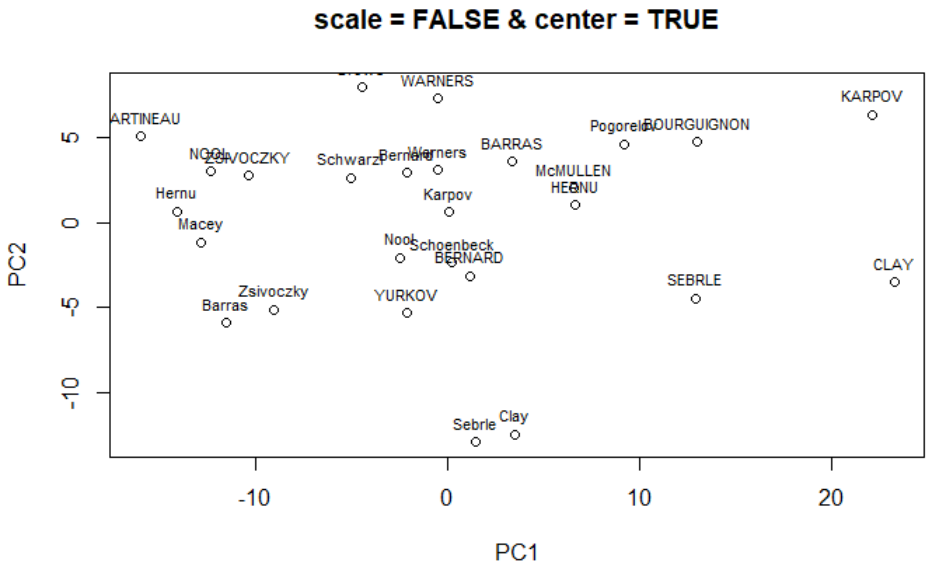


Rysunek 2.3: Wykres zmiennych dla funkcji *prcomp*().

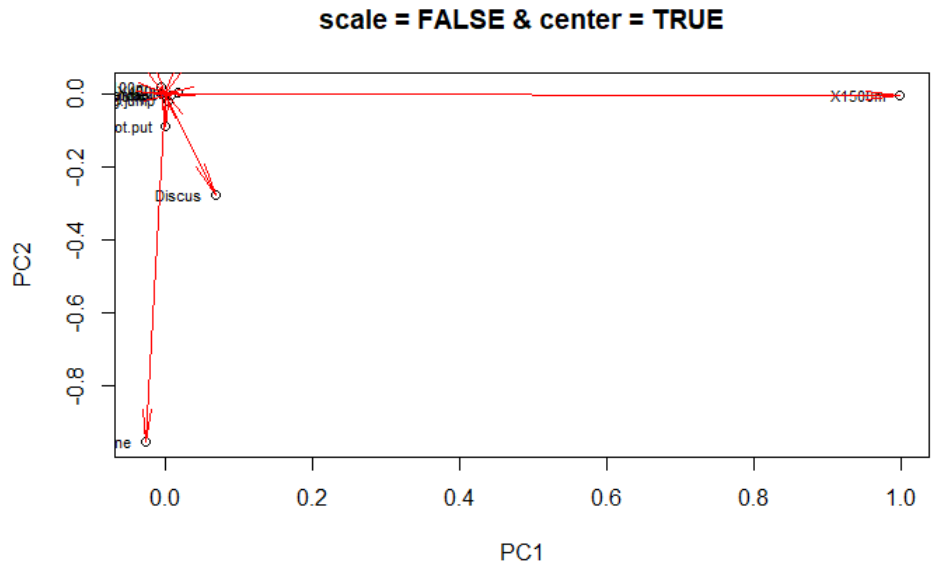


Rysunek 2.4: Wykres zmiennych dla funkcji *princomp*().

Następnie w funkcji *prcomp*() zmieniono wartość parametru *scale* na *FALSE*. Wykres obserwacji zmienił się znacząco, co zostało pokazane na wykresie 2.5. Jednak największa zmiana dotyczyła wykresu zmiennych, co zostało przedstawione na wykresie 2.6 - wykres stał się nieczytelny.

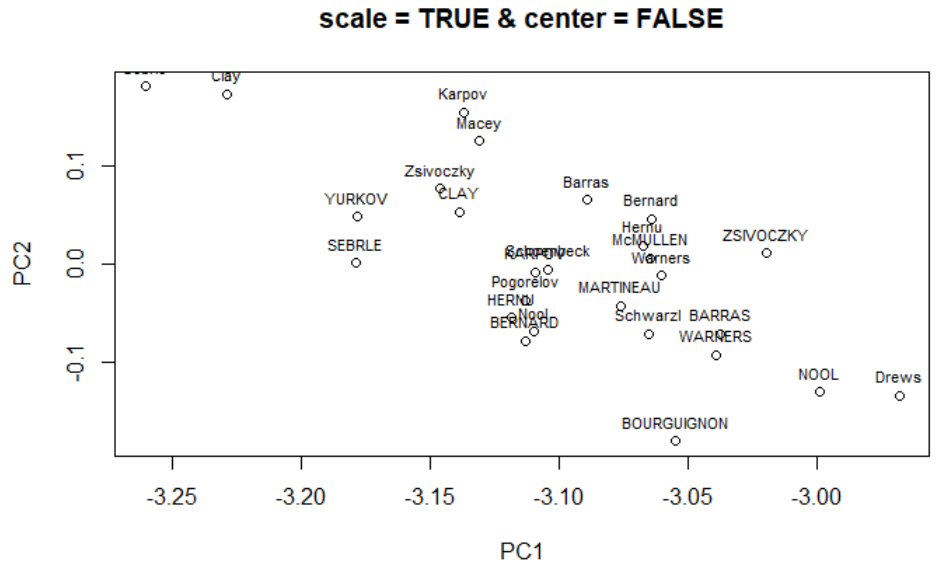


Rysunek 2.5: Wykres obserwacji dla funkcji *prcomp*() z parametrem *scale* o wartości *FALSE*.

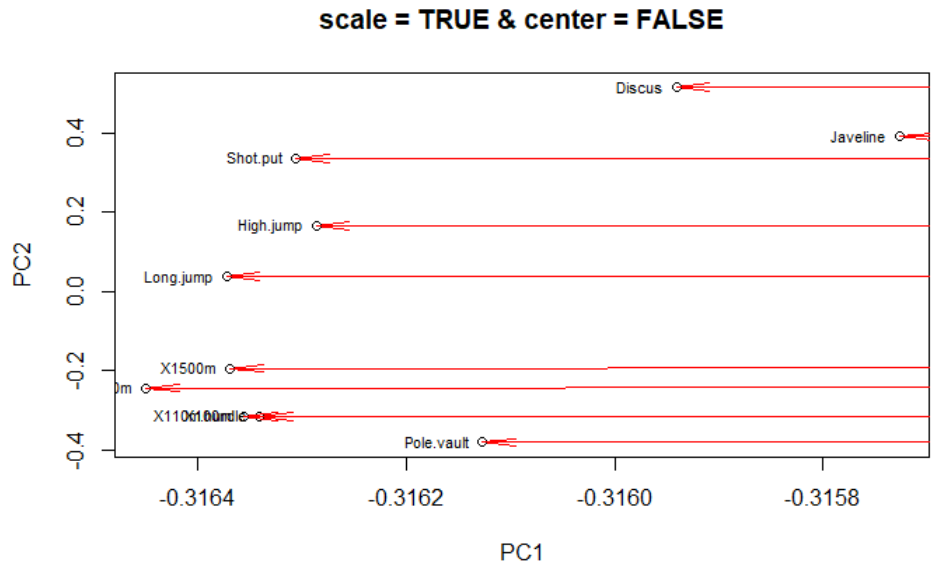


Rysunek 2.6: Wykres zmiennych dla funkcji *prcomp*() z parametrem *scale* o wartości *FALSE*.

Następnie w funkcji *prcomp*() zmieniono wartość parametru *center* na *FALSE* (wartość parametru *scale* była ustawiona na *TRUE*). Wykres obserwacji zmienił się znacząco, co zostało pokazane na wykresie 2.7. Podobnie jak we wcześniejszym przykładzie największa zmiana dotyczyła wykresu zmiennych, co zostało przedstawione na wykresie 2.8. Z wykresu da się odczytać dane, jednak sam wykres jest niezbyt pomocny.

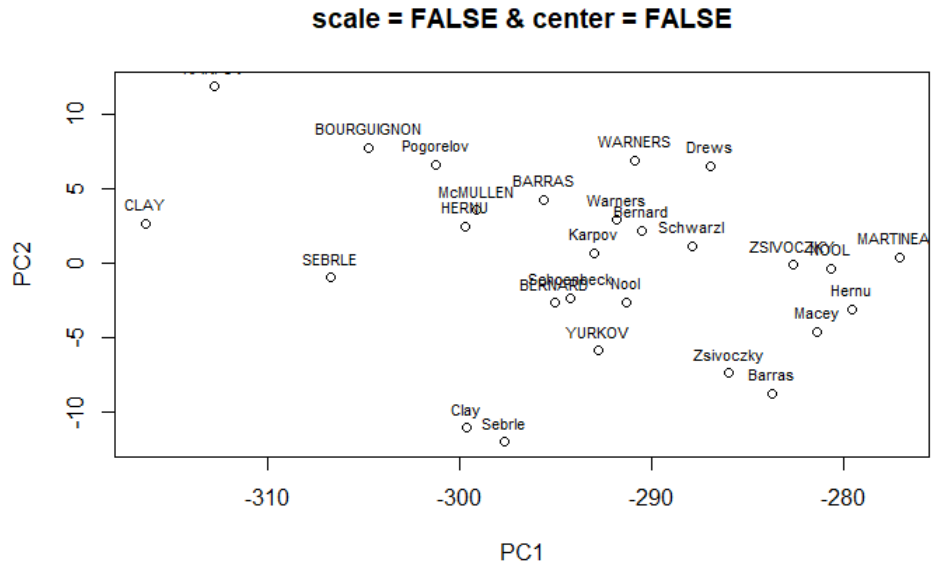


Rysunek 2.7: Wykres obserwacji dla funkcji *prcomp*() z parametrem *center* o wartości *FALSE*.

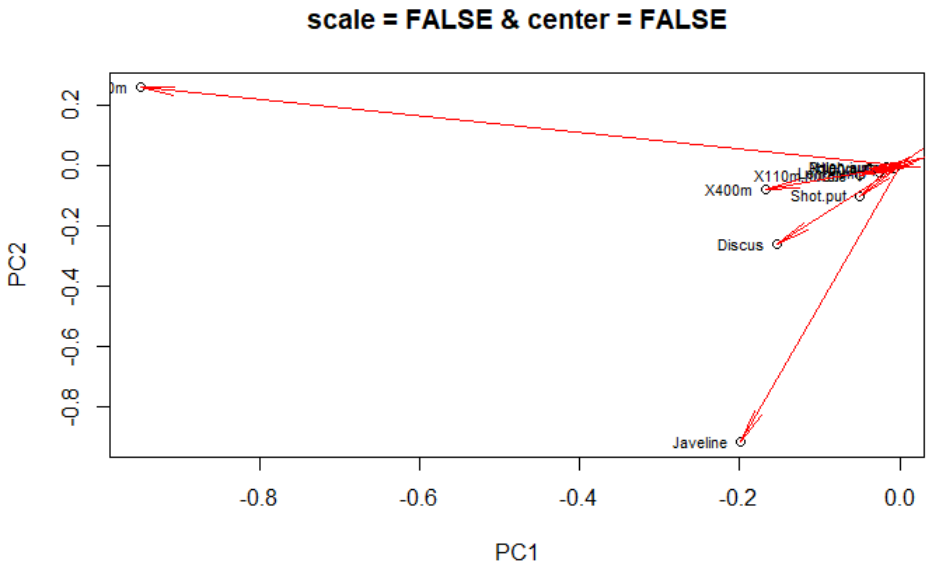


Rysunek 2.8: Wykres zmiennych dla funkcji *prcomp*() z parametrem *center* o wartości *FALSE*.

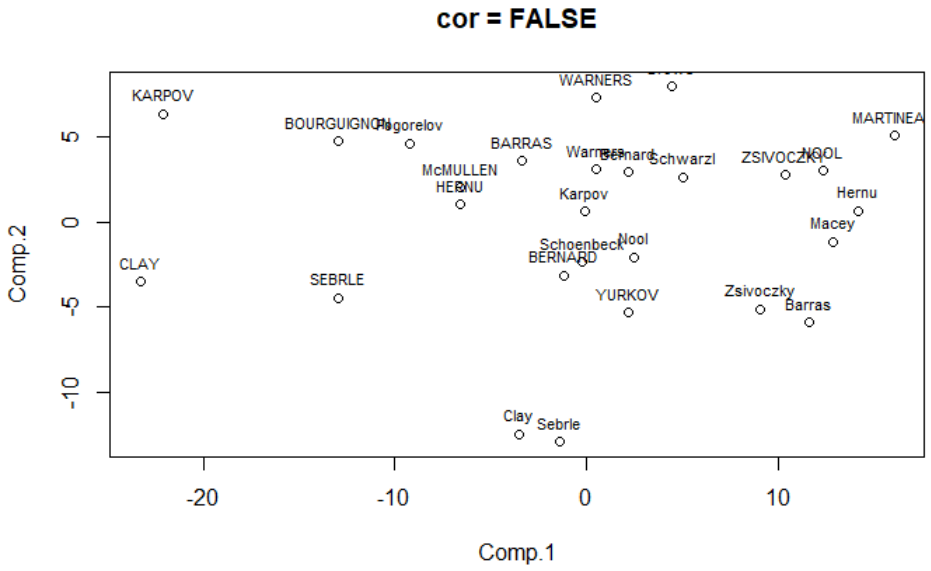
Na ostatnim etapie w funkcji *prcomp*() zmieniono wartości obydwu parametrów na *FALSE*. Wykres obserwacji został pokazany na wykresie 2.9, natomiast wykres zmiennych na wykresie 2.10. W przypadku wykresu obserwacji, obserwacje zmieniły położenie na wykresie. Natomiast w przypadku wykresu zmiennych - wykres stał się nieczytelny. W przypadku funkcji *princomp*() z parametrem *cor* mającym wartość *FALSE* wykresy obserwacji zostały pokazane kolejno na wykresach 2.11 i 2.12. Można zauważyć, że wykresy funkcji *princomp*() z parametrem *cor* równym *FALSE* odpowiadają wykresom funkcji *prcomp*() z parametrem *scale* równym *FALSE* i parametrem *center* równym *TRUE*.



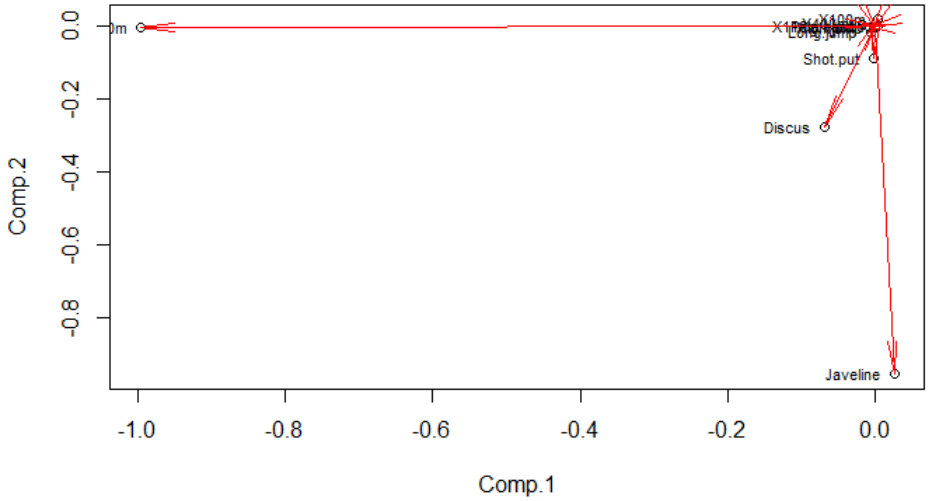
Rysunek 2.9: Wykres obserwacji dla funkcji *prcomp*() z obydwoma parametrami o wartości *FALSE*.



Rysunek 2.10: Wykres zmiennych dla funkcji *prcomp*() z obydwoma parametrami o wartości *FALSE*.



Rysunek 2.11: Wykres obserwacji dla funkcji *princomp*() z parametrem *cor* o wartości *FALSE*.



Rysunek 2.12: Wykres zmiennych dla funkcji *princomp*() z parametrem *cor* o wartości *FALSE*.

Rozdział 3

Redukcja wymiarowości metodą PCA

3.1 Ocena danych wejściowych

Ze względu na fakt, że wybrane dane zostały już we wcześniejsze analizie poddane procesowi oceny, w tym protokole zostaną tylko wspomniane podstawowe informacje na temat danych. Osoby biorące udział w badaniu były osobami młodymi - średnia wieku wynosi 30 lat, a najstarszy uczestnik miał mniej niż 50 lat. Na podstawie polecenia *table* wiadomo także, że w badaniu wzięło udział 40 kobiet i 35 mężczyzn, a także, że występują dwie grupy chorych oraz jedna grupa kontrolna. W danych występują także trzy brakujące wartości. Na podstawie nazw kolumn można było wywnioskować, że dane pochodzą z badania krwi. I tak:

- hsCRP - białko C-reaktywne, bierze udział w odpowiedzi immunologicznej, wzrasta w stanach zapalnych, infekcjach oraz w wyniku zawału mięśnia sercowego [3],
- ERY - erytrocyty, odpowiadają za przenoszenie tlenu,
- PLT - płytki krwi, odgrywają istotną rolę w procesach krzepnięcia krwi,
- HGB - hemoglobina, białko, odpowiada za przenoszenie tlenu,
- HCT - hematokryt, stosunek objętości krwinek do objętości krwi [2],
- MCHC - średnie stężenie hemoglobiny w erytrocytach [4],
- MON - monocyty, odpowiadają za fagocytozę,
- LEU - leukocyty, chronią organizm przed zakażeniami i nowotworami.

3.2 Redukcja wymiarowości

W analizie z laboratorium pierwszego niektóre dane zostały pominięte, analiza PCA wymaga, aby wszystkie dane były numeryczne, ale nie stanowiło to większego problemu. Dane typu *character* takie jak grupa i płeć został przekształcone do danych numerycznych w następujący sposób:

- grupa *CHOR1* została zastąpiona wartością 1,
- grupa *CHOR2* została zastąpiona wartością 2,
- grupa *KONTROLNA* została zastąpiona wartością 0,
- płeć *m* została zastąpiona wartością 0,

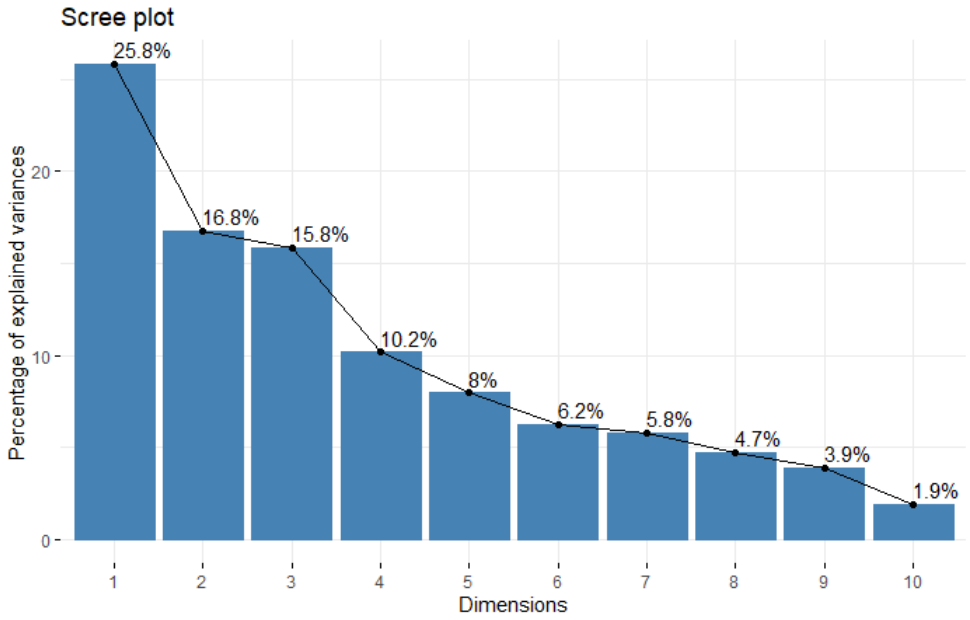
- płeć k została zastąpiona wartością 1.

Następnie zostały użyta funkcje z pakietu *factoextra* - *get_eigenvalue*(). Wyniki działania funkcji został przedstawiony w tabeli 3.1. Na podstawie kolumny *cumulative.variance.percent* można stwierdzić, że cztery pierwsze wartości własne odpowiadają aż za 68.59% zmienności danych.

	eigenvalue	**variance percent**	**cumulative variance percent**
Dim.1	2.83969014	25.8153649	25.81536
Dim.2	1.84571828	16.7792571	42.59462
Dim.3	1.74051258	15.8228417	58.41746
Dim.4	1.11921921	10.1747201	68.59218
Dim.5	0.87598456	7.9634960	76.55568
Dim.6	0.68701327	6.2455752	82.80125
Dim.7	0.63790364	5.7991240	88.60038
Dim.8	0.51748918	4.7044471	93.30483
Dim.9	0.43005540	3.9095945	97.21442
Dim.10	0.20710393	1.8827630	99.09718
Dim.11	0.09930981	0.9028164	100.00000

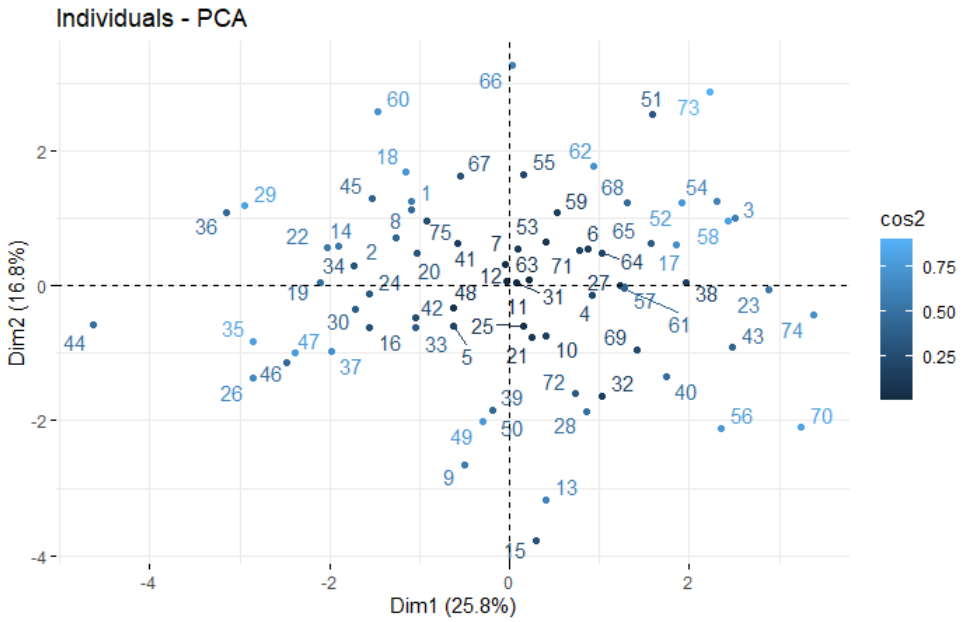
Tablica 3.1: Wyniki funkcji *get_eigenvalue*().

Następny etap polegał na wizualizacji zmienności danych na wykresie osuwiskowym, co zostało przedstawione na wykresie 3.1. Wykres osuwiskowy przedstawia to samo, co tabela 3.1 jednak w bardziej przystępnej, graficznej formie.



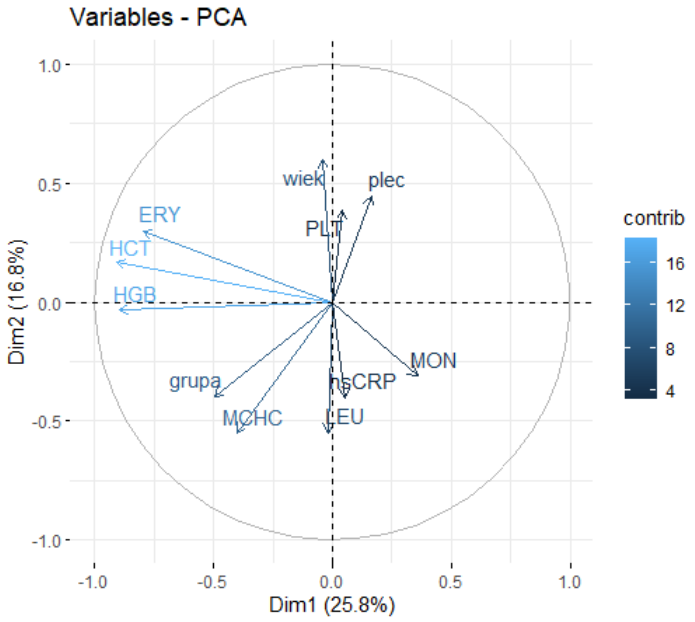
Rysunek 3.1: Wykres osuwiskowy zmienności danych.

Następnie na wykresie 3.2 przedstawiono wykres dla obserwacji. Niestety nie jest on zbyt czytelny i nie pozwala na wyciągnięcie konkretnych wniosków.

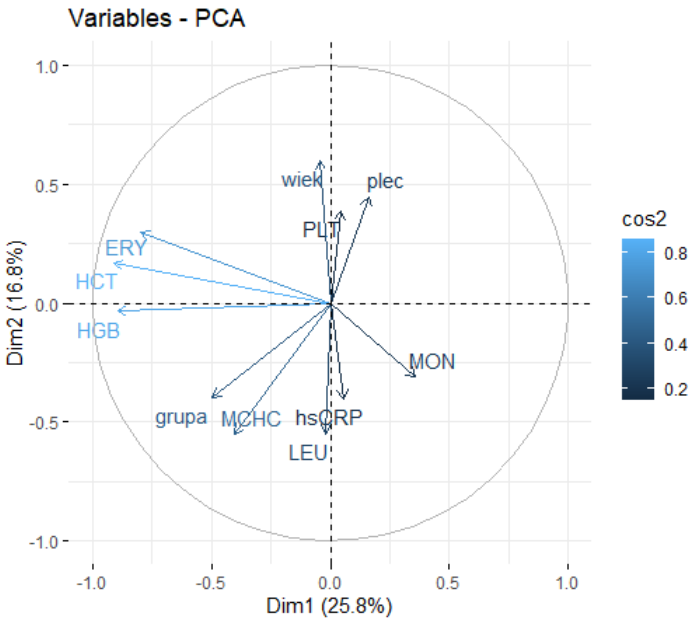


Rysunek 3.2: Wykres dla obserwacji.

Następnie na wykresie 3.3 przedstawiono wykres udziału poszczególnych zmiennych, natomiast na wykresie 3.4 przedstawiono wykres jakości reprezentacji zmiennych. Z wykresów wynika, że zmienne *HGB*, *HCT* oraz *ERY* są silnie dodatnio skorelowane, podobnie jak *grupa* i *MCHC*. Natomiast kolejno *MON* i *płeć* są ujemnie skorelowane do powyższych grup zmiennych.

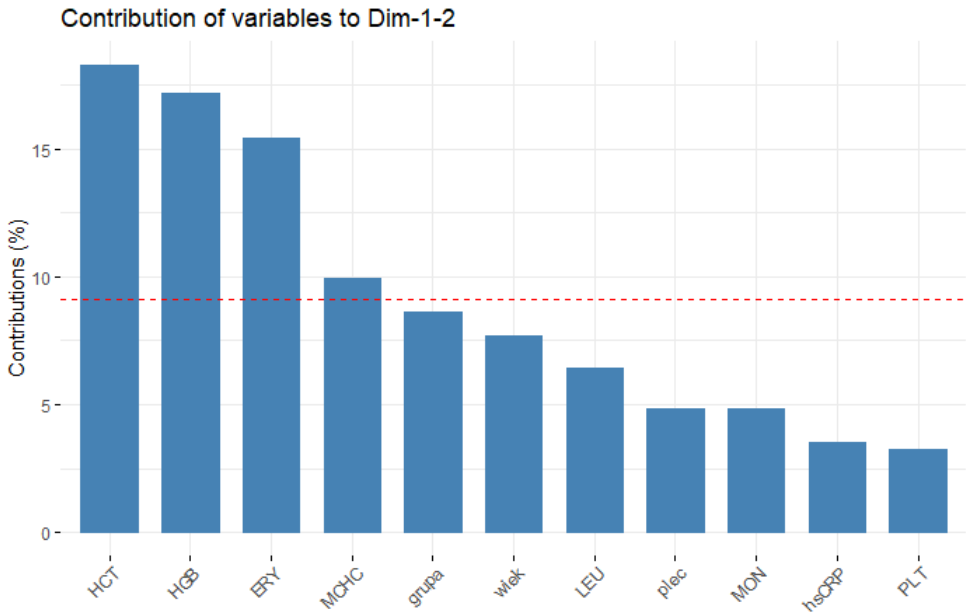


Rysunek 3.3: Wykres udziału poszczególnych zmiennych.

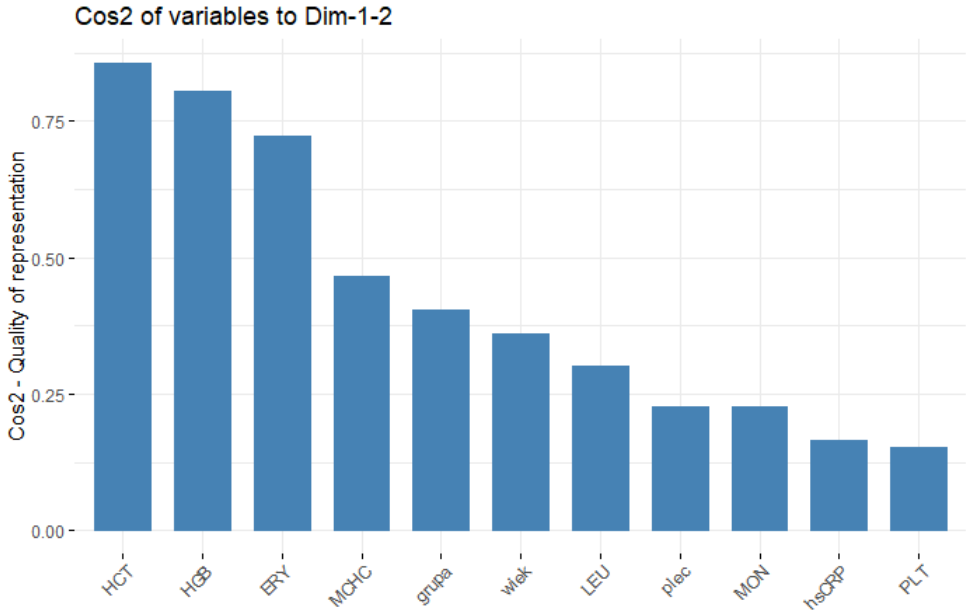


Rysunek 3.4: Wykres jakości reprezentacji zmiennych.

Następnie na wykresach 3.5 i 3.6 przedstawiono te same dane, tylko w innej formie. Z wykresów wynika, że zmienne *HCT*, *HGB*, *ERY* są najlepiej reprezentowane w składowej głównej. Także te same zmienne mają największy udział w składowej głównej.

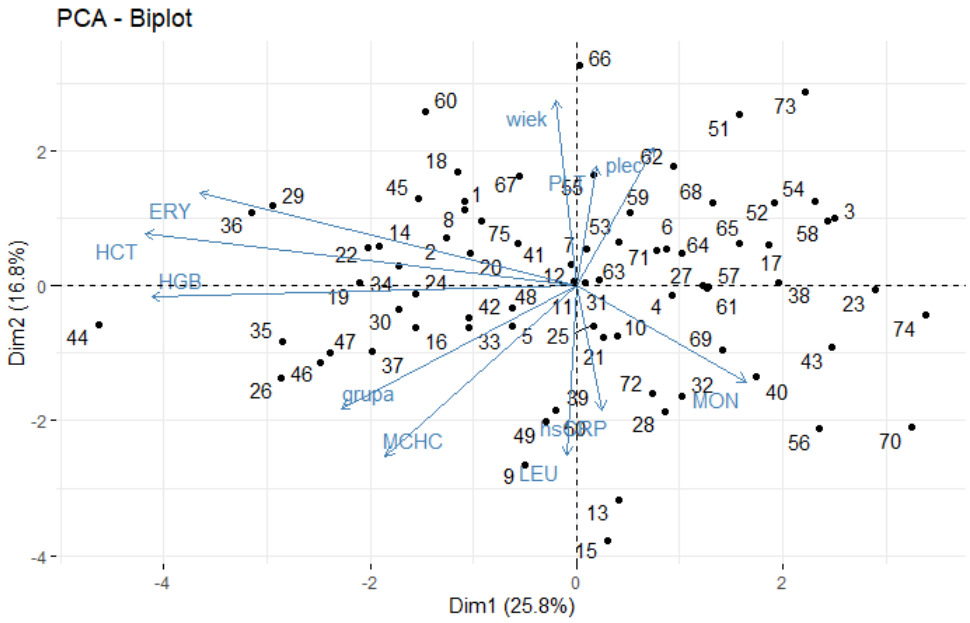


Rysunek 3.5: Wykres udziału poszczególnych zmiennych.



Rysunek 3.6: Wykres jakości reprezentacji zmiennych.

Na ostatnim wykresie przedstawiono wykres zarówno obserwacji, jak i zmiennych. Niestety nie jest on zbyt czytelny i nie przekazuje zbyt dużo wartościowych informacji, chociaż wskazuje, które zmienne mają wpływ, na które obserwacje.



Rysunek 3.7: Wykres obserwacji i zmiennych.

Rozdział 4

Podsumowanie

Najciekawszym wnioskiem płynącym z analizy jest fakt, że wyniki powyższej analizy różnią się od wyników analizy z laboratorium pierwszego. W laboratorium pierwszym została wykryta korelacja zmiennych *HCT* oraz *MCHC*, natomiast powyższa analiza wskazała na brak korelacji między tymi zmiennymi, natomiast wskazała na korelację między zmiennymi *MCHC* oraz *grupa*. Niestety ze względu na brak wiedzy dziedzinowej nie można stwierdzić, która analiza jest błędna.

Literatura

[1] Dane poddane analizie. [on-line] www.cs.put.poznan.pl/kgutowska/PSwBB/dane/przykladoweDane-Projekt.csv.

[2] Hematokryt. [on-line] https://lekarzebezkolejki.pl/blog/hematokryt-hct-normy-co-oznacza-niski-i-wysoki-poziom-hct/w-603.

[3] Białko hscrp. [on-line] https://www.labtestsonline.pl/test/hs-crp.

[4] Mchc. [on-line] https://www.medonet.pl/zdrowie/pytania-do-lekarzy,o-czym-swiadczy-mchc-ponizej-normy-,porada,43426672.html.