Błażej Domagała - WED

# Lista 1

## Zadanie 1

a) Nazwa zbioru: Wine

Obraz zawierający tekst, zrzut ekranu, Czcionka

Opis wygenerowany automatycznie

b) **Krótki tekstowy opis zbioru**

Zbiór danych "Wine" zawiera wyniki analizy chemicznej win wyprodukowanych w określonym regionie we Włoszech przez trzech różnych producentów. Analiza chemiczna dotyczy 13 różnych składników zawartych w winach.

Obraz zawierający zrzut ekranu, tekst, Czcionka

Opis wygenerowany automatycznie

c) Liczba obserwacji w zbiorze: 178

Obraz zawierający tekst, Czcionka, zrzut ekranu, numer

Opis wygenerowany automatycznie

d) Liczba kolumn: 14 (13 atrybutów + 1 kolumna identyfikująca klasę)

Obraz zawierający tekst, zrzut ekranu, Czcionka

Opis wygenerowany automatycznie

e) Zmienna celu:

* Nazwa kolumny z klasą: Class
* Liczba klas: 3 (Klasy 1, 2 i 3)

Obraz zawierający tekst, Czcionka, zrzut ekranu, numer

Opis wygenerowany automatycznie

f) Wykaz i opis cech:

1. **Class**: Zmienna kategoryczna. Klasa wina.
2. **Alcohol**: Zmienna ilościowa. Zawartość alkoholu.
3. **Malic acid**: Zmienna ilościowa. Zawartość kwasu jabłkowego.
4. **Ash**: Zmienna ilościowa. Zawartość popiołu.
5. **Alcalinity of ash**: Zmienna ilościowa. Zasadowość popiołu.
6. **Magnesium**: Zmienna ilościowa. Zawartość magnezu.
7. **Total phenols**: Zmienna ilościowa. Całkowita zawartość fenoli.
8. **Flavanoids**: Zmienna ilościowa. Zawartość flawonoidów.
9. **Nonflavanoid phenols**: Zmienna ilościowa. Zawartość fenoli nieflawonoidowych.
10. **Proanthocyanins**: Zmienna ilościowa. Zawartość proantocyjanidyn.
11. **Color intensity**: Zmienna ilościowa. Intensywność koloru.
12. **Hue**: Zmienna ilościowa. Odcień.
13. **OD280/OD315 of diluted wines**: Zmienna ilościowa. Stosunek absorbancji przy 280 nm do 315 nm w rozcieńczonych winach.
14. **Proline**: Zmienna ilościowa. Zawartość prolina.

Obraz zawierający tekst, zrzut ekranu, Czcionka

Opis wygenerowany automatycznie

## Zadanie 2

1. Zmiana nazw kolumn (pierwszą kolumnę – zmienną celu – proszę nazwać „Class”; nazwy pozostałych kolumn – atrybutów – są podane w pliku „wine.names”)

Obraz zawierający tekst, oprogramowanie, Oprogramowanie multimedialne, zrzut ekranu

Opis wygenerowany automatycznie

1. Polecenie View (fragment print-screena z tabelką)

Obraz zawierający Czcionka, tekst, zrzut ekranu, linia

Opis wygenerowany automatycznie

Obraz zawierający zrzut ekranu, tekst, wyświetlacz, oprogramowanie

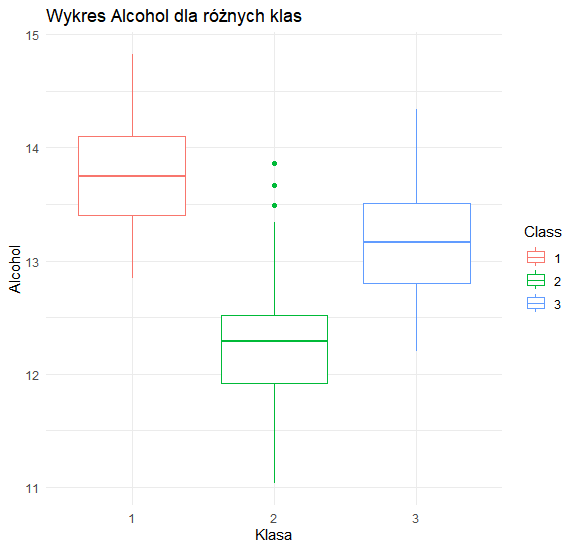
Opis wygenerowany automatycznie

1. Podsumowanie cech (summary)

Obraz zawierający tekst, zrzut ekranu, Czcionka

Opis wygenerowany automatycznie

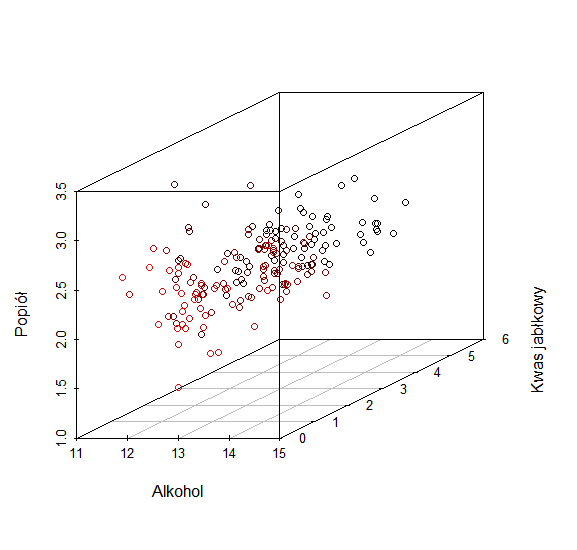
1. Wykres 2D ilustrujący wybraną cechę dla różnych klas



Obraz zawierający tekst, zrzut ekranu

Opis wygenerowany automatycznie

1. Wykres 3D dla trzech wybranych cech (bez klasy)



Obraz zawierający tekst, Czcionka, zrzut ekranu

Opis wygenerowany automatycznie

# Lista 2

1. Dla wybranej cechy wyświetlić wybrane wartości stosując: zakres wartości, sekwencję indeksów (np. co dziesiąty indeks), indeksy ujemne, warunki logiczne.

Kod:

path = "C:\\Users\\petitoff\\Desktop\\repos\\UO\\rok 3\\Wprowadzenie do eksploracji danych\\lista2"

setwd(path) ## ustawienie ścieżki

# Załadowanie danych

wine <- read.csv('wine\\wine.data', header = FALSE)

# Zmień nazwy kolumn

names(wine) <-

c(

'Class',

'Alcohol',

'Malic acid',

'Ash',

'Alcalinity of ash',

'Magnesium',

'Total phenols',

'Flavanoids',

'Nonflavanoid phenols',

'Proanthocyanins',

'Color intensity',

'Hue',

'OD280/OD315 of diluted wines',

'Proline'

)

# a) Dla wybranej cechy wyświetlić wybrane wartości stosując: zakres wartości, sekwencję indeksów (np. co dziesiąty indeks), indeksy ujemne, warunki logiczne.

# Zakres wartości

wine$Alcohol[5:15]

# Sekwencja indeksów

wine$Alcohol[seq(1, nrow(wine), 10)]

# Indeksy ujemne

wine$Alcohol[-(1:5)]

# Warunki logiczne

wine$Alcohol[wine$Alcohol > 14]

Obraz zawierający tekst, zrzut ekranu, Czcionka

Opis wygenerowany automatycznie

1. Wyświetlić wybrane wiersze i kolumny z tabeli.

# b) Wybranie wierszy i kolumn:

selected\_rows\_columns <- wine[1:10, c("Alcohol", "Malic acid")]

print(selected\_rows\_columns)

# indeksy

wine[5:15,]

Obraz zawierający tekst, zrzut ekranu

Opis wygenerowany automatycznie

1. Dodać do tabeli nową kolumnę z wartościami obliczonymi na podstawie innych wybranych kolumn.

# c) Dodanie nowej kolumny:

# Sprawdzanie, czy kolumna istnieje i zawiera dane

if ("Total phenols" %in% names(wine) &&

!all(is.na(wine$`Total phenols`))) {

# Dodanie nowej kolumny

wine$Total.phenols.squared <- wine$`Total phenols` ^ 2

head(wine)

} else {

cat("Column 'Total phenols' does not exist or is empty.")

}

Obraz zawierający tekst, zrzut ekranu, Czcionka

Opis wygenerowany automatycznie

1. Podać wartości podstawowych statystyk dla wybranej kolumny: zakres, średnia, mediana,

# d) Statystyki podstawowe dla wybranej kolumny, np. "Alcohol"

cat("Statystyki dla kolumny 'Alcohol':", "\n")

cat("Zakres: ", min(wine$Alcohol), " - ", max(wine$Alcohol), "\n")

cat("Średnia: ", mean(wine$Alcohol), "\n")

cat("Mediana: ", median(wine$Alcohol), "\n")

cat("Odchylenie standardowe: ", sd(wine$Alcohol), "\n")

cat("Kurtoza: ", moments::kurtosis(wine$Alcohol), "\n")

cat("Skośność: ", moments::skewness(wine$Alcohol), "\n")

cat("Kwantyle: ", quantile(wine$Alcohol, probs = c(0.25, 0.5, 0.75)), "\n")

Obraz zawierający tekst, zrzut ekranu, Czcionka, oprogramowanie

Opis wygenerowany automatycznie

1. Wyznaczyć i zilustrować na wykresie macierz korelacji dla wybranych pięciu zmiennych.

# e) Macierz korelacji dla wybranych pięciu zmiennych i jej wizualizacja

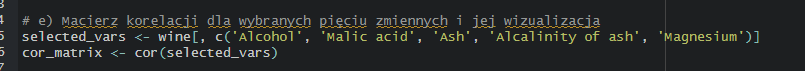
selected\_vars <-

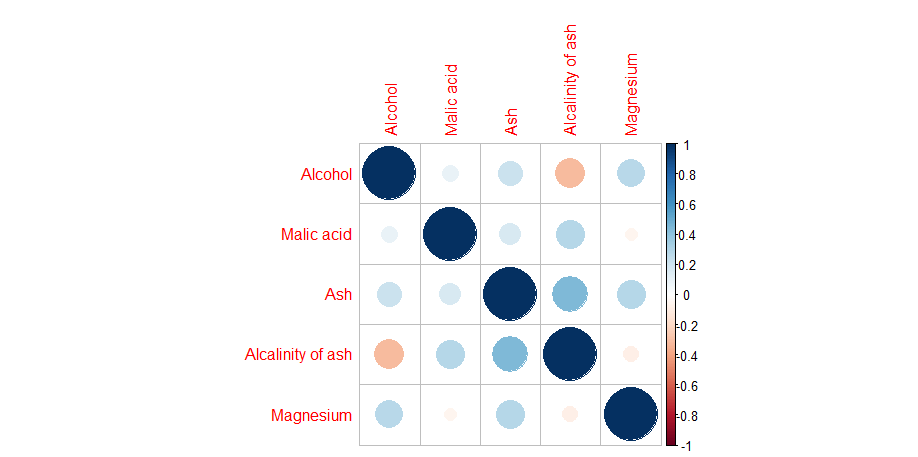
wine[, c('Alcohol', 'Malic acid', 'Ash', 'Alcalinity of ash', 'Magnesium')]

cor\_matrix <- cor(selected\_vars)

library(corrplot)

corrplot(cor\_matrix, method = "circle")





1. Wydrukować histogramy dla trzech różnych zmiennych, przedyskutować wyniki.

# f) Histogramy dla trzech różnych zmiennych

par(mfrow = c(1, 3)) # ustawienie layoutu na 1 wiersz i 3 kolumny

hist(wine$Alcohol,

main = 'Alcohol',

xlab = '',

col = 'skyblue')

hist(

wine$`Malic acid`,

main = 'Malic acid',

xlab = '',

col = 'skyblue'

)

hist(wine$Ash,

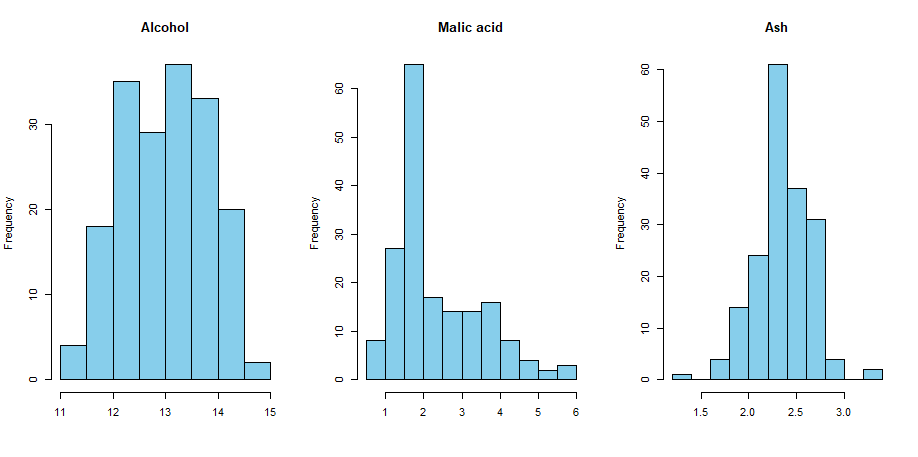
main = 'Ash',

xlab = '',

col = 'skyblue')

Obraz zawierający tekst, zrzut ekranu, Czcionka, oprogramowanie

Opis wygenerowany automatycznie



# Lista 3

## Usuń kolumny z wartościami nominalnymi (identyfikatory, itp.) – jeżeli są, inne niż zmienna celu.

Zakładając, że zmienna celu to V1, a pozostałe kolumny to wartości liczbowe nie ma kolumn z wartościami nominalnymi do usunięcia.

Jeśli jednak byłyby takie kolumny, można by je usunąć za pomocą polecenia subset.

Przykład: Jeśli kolumna V2 była by zmienną nominalną, można by ją usunąć następująco:

wine <- subset(wine, select = -V2) # Usuń kolumnę V2 z ramki danych wine

View(wine) # Wyświetl dane po usunięciu kolumny V2

## b) Zmień nazwy kolumn na nazwy w języku polskim. Nowe nazwy powinny być: krótkie, znaczące, bez polskich znaków i spacji. Wyświetl dane poleceniem View.

library(datasets)

path = "C:\\Users\\petitoff\\Desktop\\repos\\UO\\rok 3\\Wprowadzenie do eksploracji danych\\lista3\\" # używając podwójnych ukośników

setwd(path) ## ustawienie ścieżki

# Załadowanie danych

wine <- read.csv('wine\\wine.data', header = FALSE)

# Oto kod zmieniający nazwy kolumn na nazwy w języku polskim:

nowe\_nazwy <- c(

"Klasa",

'Alkohol',

'Kwas jabłkowy',

'Popiół',

'Alkalność popiołu',

'Magnez',

'Całkowite fenole',

'Flawonoidy',

'Fenole nietrwałe',

'Proantocyjanidy',

'Intensywność koloru',

'Odcień',

'Stężenie odwiedlane win',

'Prolina'

)

Obraz zawierający zrzut ekranu, oprogramowanie, Oprogramowanie multimedialne, tekst

Opis wygenerowany automatycznie

## c) Zmienne o wartościach logicznych (jeżeli są) zapisz jako logiczne (polecenie as.logical).

Obraz zawierający zrzut ekranu, oprogramowanie, Oprogramowanie multimedialne, tekst

Opis wygenerowany automatycznie

wine$Klasa <- as.logical(wine$Klasa)

## d) Upewnij się, że zmienne o wartościach liczbowych są typu liczbowego, a jeżeli nie są, to zapisz je jako numeryczne (as.numeric)

wprowadzone <- c('Alkohol', 'Kwas jabłkowy', 'Popiół', 'Alkalność popiołu', 'Magnez', 'Całkowite fenole', 'Flawonoidy', 'Fenole nietrwałe', 'Proantocyjanidy', 'Intensywność koloru', 'Odcień', 'Stężenie odwiedlane win', 'Prolina')

for (zmienna in wprowadzone) {

if (class(wine[[zmienna]]) != "numeric") {

wine[[zmienna]] <- as.numeric(wine[[zmienna]])

}

}

## e) Zmienną celu zapisz jako mającą wartości nominalne (polecenie as.factor).

wine$Klasa <- as.factor(wine$Klasa)

f) Policz brakujące wartości. Jeżeli są, to dla kolumn o wartościach liczbowych zastąp je wartościami średnimi dla kolumn.

kolumny\_numeryczne <- c('Alkohol', 'Kwas jabłkowy', 'Popiół', 'Alkalność popiołu', 'Magnez', 'Całkowite fenole', 'Flawonoidy', 'Fenole nietrwałe', 'Proantocyjanidy', 'Intensywność koloru', 'Odcień', 'Stężenie odwiedlane win', 'Prolina')

for (kolumna in kolumny\_numeryczne) {

brakujace <- is.na(wine[[kolumna]])

if (sum(brakujace) > 0) {

srednia <- mean(wine[[kolumna]], na.rm = TRUE)

wine[[kolumna]][brakujace] <- srednia

}

}

g) Zapisz przetworzone dane do pliku (razem z nowymi nazwami kolumn). Załącz fragment print-screena zawartości (początku) pliku.

write.csv(wine, file = 'wyniki.csv', row.names = FALSE)

Obraz zawierający tekst, zrzut ekranu, oprogramowanie, komputer

Opis wygenerowany automatycznie

# Lista 4

## Obliczy i narysuje macierz korelacji zmiennych (bez zmiennej celu wyznaczającej klasy)

library(ggplot2)

library(reshape2)

correlation\_matrix <- function(df, threshold) {

# Obliczanie macierzy korelacji

cor\_matrix <- cor(df)

# Rysowanie macierzy korelacji

melted\_cor\_matrix <- melt(cor\_matrix)

plot <- ggplot(data = melted\_cor\_matrix, aes(x=Var1, y=Var2, fill=value)) +

geom\_tile() +

scale\_fill\_gradient2(low = "blue", high = "red", mid = "white",

midpoint = 0, limit = c(-1,1), space = "Lab",

name="Pearson\nCorrelation") +

theme\_minimal() +

theme(axis.text.x = element\_text(angle = 45, vjust = 1,

size = 12, hjust = 1),

axis.text.y = element\_text(size = 12)) +

coord\_fixed()

# Zapisywanie rysunku do pliku

ggsave("correlation\_matrix.png", plot)

# Wypisywanie par zmiennych o korelacji większej niż zadany próg

cor\_pairs <- subset(melted\_cor\_matrix, abs(value) > threshold & Var1 != Var2)

# Usuwanie powtórzeń

cor\_pairs <- cor\_pairs[!duplicated(t(apply(cor\_pairs[,c("Var1","Var2")],1,sort))),]

return(cor\_pairs)

}

# Załadowanie zestawu danych iris

data(iris)

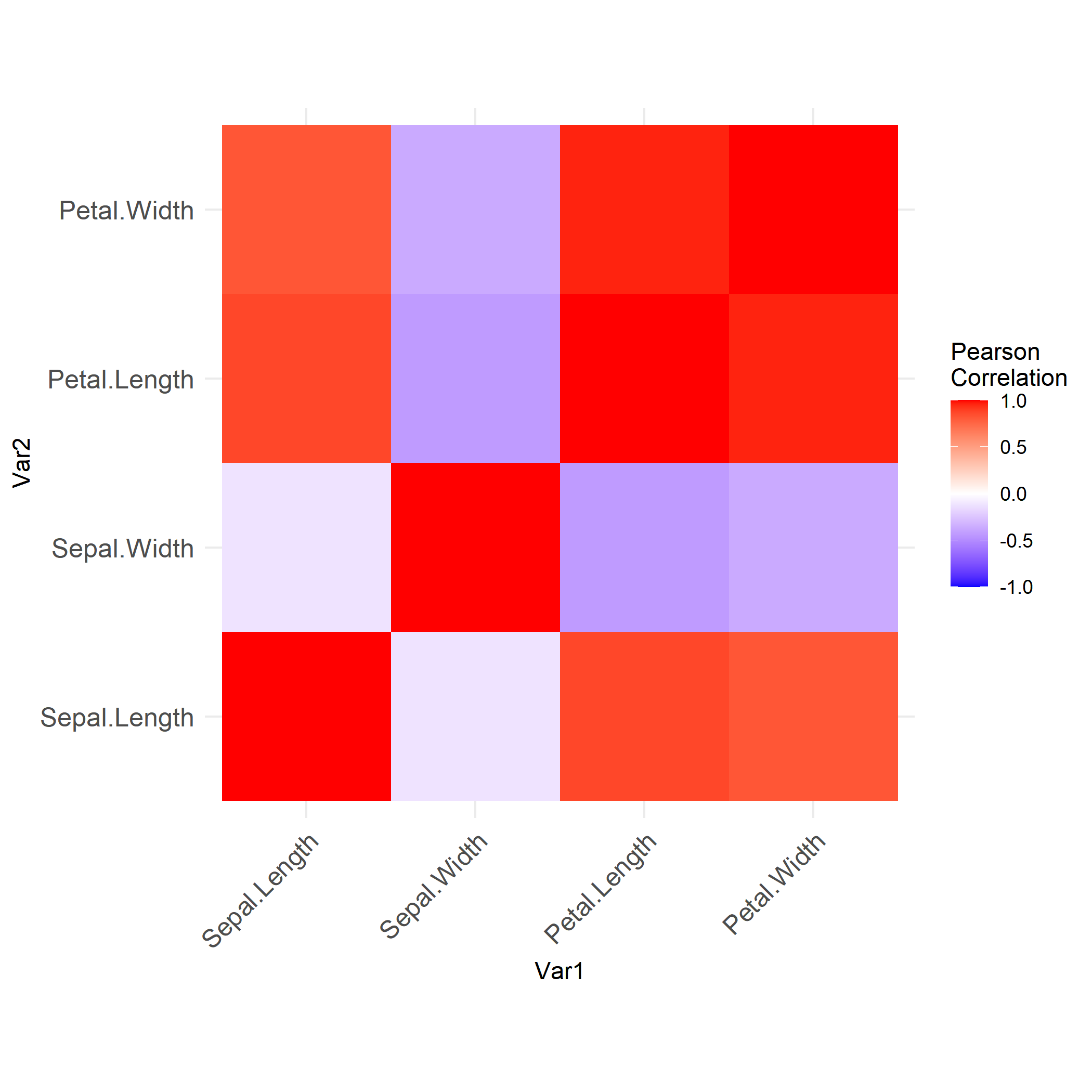
# Usunięcie kolumny Species (bo to jest nasza zmienna celu)

df <- iris[,-5]

# Użycie funkcji na df z progiem 0.5

correlation\_matrix(df, 0.5)

## zapisze rysunek macierzy korelacji do pliku



## c) wypisze pary (nazwy) zmiennych o korelacji większej niż zadany próg oraz odpowiadające im wartości korelacji (wartość progu powinna być argumentem funkcji).

* Proszę uwzględnić ujemne wartości korelacji; czyli przyjmujemy, że np. korelacja równa -0.95 jest powyżej progu 0.9, bo jest to silna korelacja, tylko ujemna (wraz ze wzrostem wartości jednej cechy następuje spadek wartości drugiej cechy).
* Pary proszę wypisać bez powtórek (czyli jeżeli mamy już korelację cechy x z cechą y,to nie wypisujemy korelacji cechy y z x).

Wyniki z konsoli:

Obraz zawierający tekst, Czcionka, zrzut ekranu, numer

Opis wygenerowany automatycznie

# Lista 5

# Dokumentacja Kodu R - Zadanie 2

# Biblioteka datasets

library(datasets)

# Ustawienie ścieżki dostępu do danych

path = "C:\\Users\\petit\\Desktop\\repos\\UO\\rok 3\\Wprowadzenie do eksploracji danych\\lista5\\"

setwd(path)

# -------------- Zadanie 2a: Wczytanie i identyfikacja punktów oddalonych -------------- #

# Wczytanie danych

wine\_data <- read.csv('wine\\wine.data', header = FALSE)

# Funkcja do identyfikacji i zliczania punktów oddalonych przy użyciu IQR

identify\_outliers <- function(data) {

quantiles <- quantile(data, c(.25, .75), na.rm = TRUE)

iqr <- IQR(data, na.rm = TRUE)

lower\_bound <- quantiles[1] - 1.5 \* iqr

upper\_bound <- quantiles[2] + 1.5 \* iqr

outliers <- data[data < lower\_bound | data > upper\_bound]

return(list("dolna\_granica" = lower\_bound, "gorna\_granica" = upper\_bound, "punkty\_oddalone" = outliers))

}

# Część a: Wykrywanie punktów oddalonych dla każdej zmiennej

outliers\_info <- lapply(wine\_data, identify\_outliers)

# Dodanie sztucznego punktu oddalonego, jeśli nie istnieje

for (i in 1:length(outliers\_info)) {

if (length(outliers\_info[[i]]$punkty\_oddalone) == 0) {

wine\_data[i][1] <- outliers\_info[[i]]$gorna\_granica + 1 # Dodanie punktu oddalonego

}

}

# Część b: Wizualizacja punktów oddalonych dla maksymalnie 4 zmiennych

# Wybór pierwszych 4 zmiennych do wizualizacji

selected\_vars <- head(names(wine\_data), 4)

plots <- list()

for (var in selected\_vars) {

plot <- ggplot(wine\_data, aes\_string(x=var)) +

geom\_histogram(binwidth = 1, fill="skyblue", color="black") +

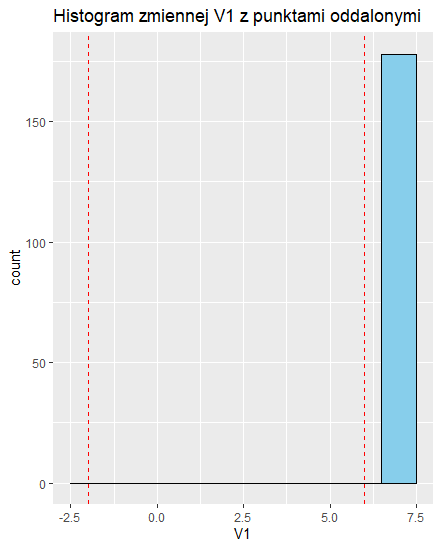
geom\_vline(xintercept = outliers\_info[[var]]$dolna\_granica, color="red", linetype="dashed") +

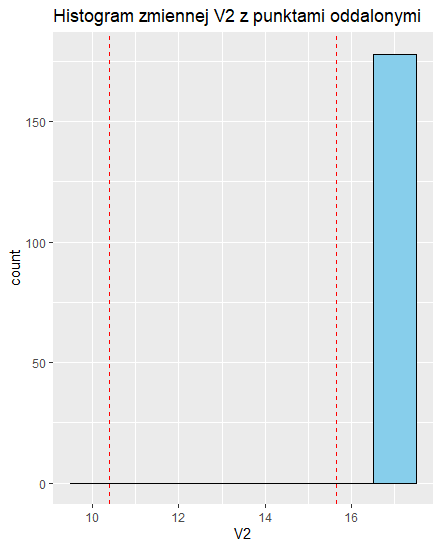
geom\_vline(xintercept = outliers\_info[[var]]$gorna\_granica, color="red", linetype="dashed") +

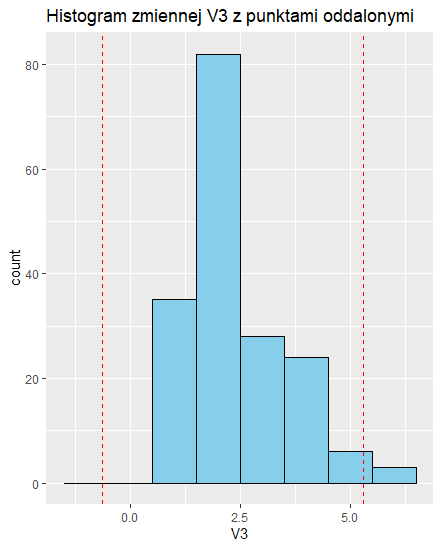
ggtitle(paste("Histogram zmiennej", var, "z punktami oddalonymi"))

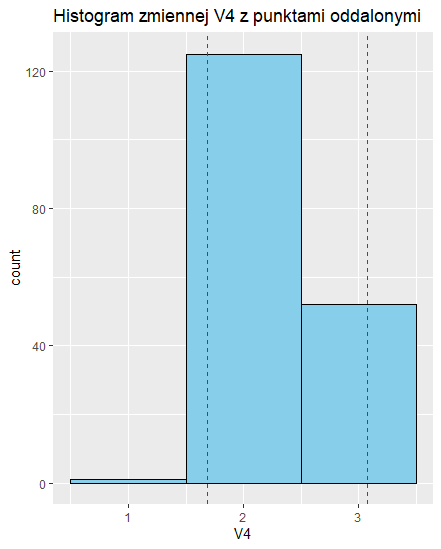
plots[[var]] <- plot

}









# Część c: Usuwanie punktów oddalonych

cleaned\_wine\_data <- wine\_data

for (var in names(wine\_data)) {

cleaned\_wine\_data <- cleaned\_wine\_data[cleaned\_wine\_data[[var]] >= outliers\_info[[var]]$dolna\_granica &

cleaned\_wine\_data[[var]] <= outliers\_info[[var]]$gorna\_granica, ]

}

# Część d: Wizualizacja po usunięciu punktów oddalonych dla tych samych 4 zmiennych

cleaned\_plots <- list()

for (var in selected\_vars) {

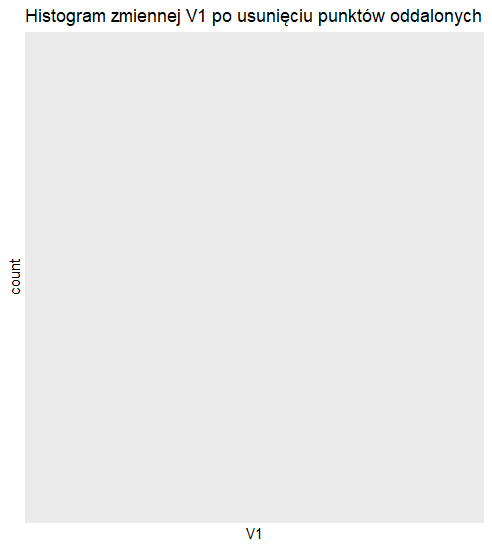
plot <- ggplot(cleaned\_wine\_data, aes\_string(x=var)) +

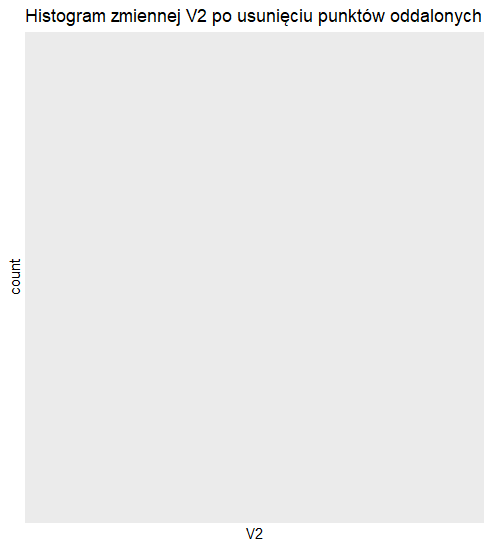
geom\_histogram(binwidth = 1, fill="green", color="black") +

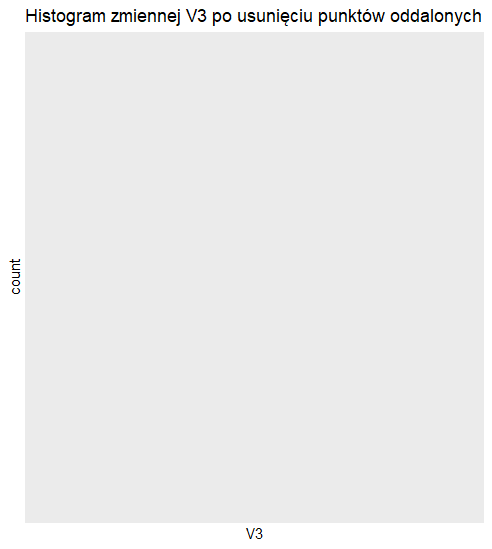
ggtitle(paste("Histogram zmiennej", var, "po usunięciu punktów oddalonych"))

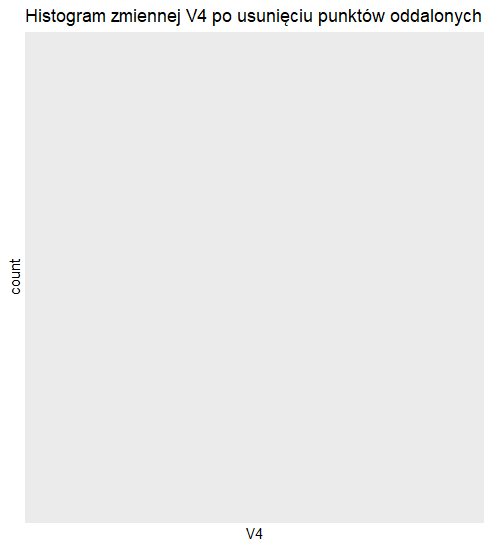
cleaned\_plots[[var]] <- plot

}









# Część e: Zapisanie oczyszczonego zbioru danych

write.csv(cleaned\_wine\_data, "cleaned\_wine\_data.csv", row.names = FALSE)

# Wynik

list("oryginalne\_wykresy" = plots, "oczyszczone\_wykresy" = cleaned\_plots, "plik\_z\_oczyszczonymi\_danymi" = "cleaned\_wine\_data.csv")