Bank Customers

Karolina Popiołek

2024-08-08

1. Dane

1.1. Załączenie danych i wyświetlenie podglądu kilku pierwszych wierszy

```
Churn.Modeling <- read.csv("C:/Users/Karolina/Documents/Churn Modeling.csv")
attach(Churn.Modeling)
length(Churn.Modeling$Surname)
```

```
## [1] 10000
```

```
head(Churn.Modeling)
```

```
##   RowNumber CustomerId  Surname CreditScore Geography Gender Age Tenure
## 1         1   15634602 Hargrave         619    France Female  42      2
## 2         2   15647311     Hill         608     Spain Female  41      1
## 3         3   15619304     Onio         502    France Female  42      8
## 4         4   15701354     Boni         699    France Female  39      1
## 5         5   15737888 Mitchell         850     Spain Female  43      2
## 6         6   15574012      Chu         645     Spain   Male  44      8
##     Balance NumOfProducts HasCrCard IsActiveMember EstimatedSalary Exited
## 1      0.00             1         1              1       101348.88      1
## 2  83807.86             1         0              1       112542.58      0
## 3 159660.80             3         1              0       113931.57      1
## 4      0.00             2         0              0        93826.63      0
## 5 125510.82             1         1              1        79084.10      0
## 6 113755.78             2         1              0       149756.71      1
```

1.2. Sprawdzenie struktury danych

Występuje łącznie 14 zmiennych (9 - integer, 3 - character, 2 - numeric).

```
str(Churn.Modeling)
```

```
## 'data.frame':    10000 obs. of  14 variables:
##  $ RowNumber      : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ CustomerId     : int  15634602 15647311 15619304 15701354 15737888 15574012 15592531 15656148 157
92365 15592389 ...
##  $ Surname        : chr  "Hargrave" "Hill" "Onio" "Boni" ...
##  $ CreditScore    : int  619 608 502 699 850 645 822 376 501 684 ...
##  $ Geography      : chr  "France" "Spain" "France" "France" ...
##  $ Gender         : chr  "Female" "Female" "Female" "Female" ...
##  $ Age            : int  42 41 42 39 43 44 50 29 44 27 ...
##  $ Tenure         : int  2 1 8 1 2 8 7 4 4 2 ...
##  $ Balance        : num  0 83808 159661 0 125511 ...
##  $ NumOfProducts  : int  1 1 3 2 1 2 2 4 2 1 ...
##  $ HasCrCard      : int  1 0 1 0 1 1 1 1 0 1 ...
##  $ IsActiveMember : int  1 1 0 0 1 0 1 0 1 1 ...
##  $ EstimatedSalary: num  101349 112543 113932 93827 79084 ...
##  $ Exited         : int  1 0 1 0 0 1 0 1 0 0 ...
```

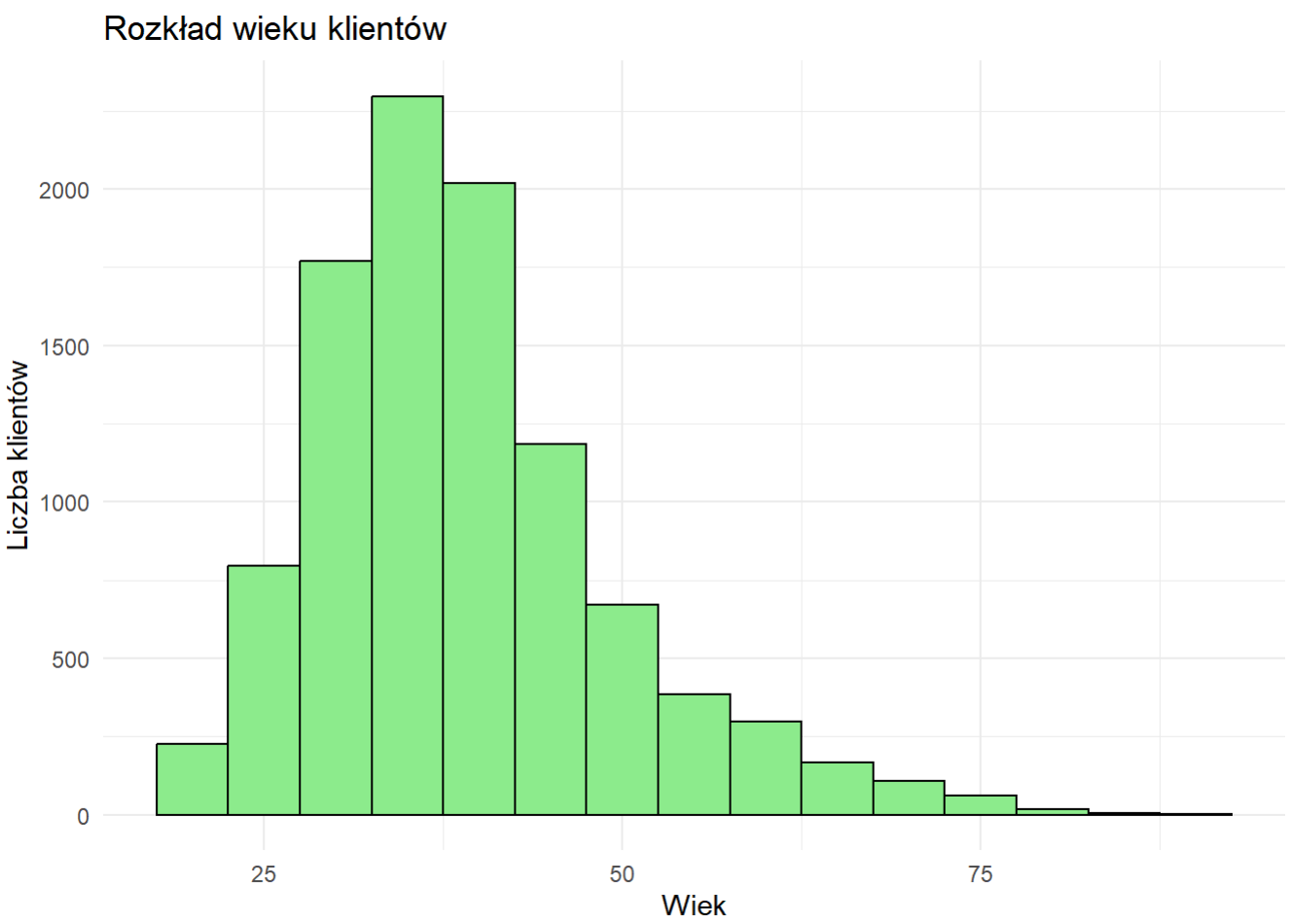
1.3. Konwertowanie zmiennych kategorycznych

```
Churn.Modeling$Geography <- as.factor(Churn.Modeling$Geography)
Churn.Modeling$Gender <- as.factor(Churn.Modeling$Gender)
Churn.Modeling$HasCrCard <- as.factor(Churn.Modeling$HasCrCard)
Churn.Modeling$IsActiveMember <- as.factor(Churn.Modeling$IsActiveMember)
Churn.Modeling$Exited <- as.factor(Churn.Modeling$Exited)
```

2. Rozkład danych

2.1. Rozkład danych ze względu na wiek

```
ggplot(Churn.Modeling, aes(x = Age)) +
  geom_histogram(binwidth = 5, fill = "lightgreen", color = "black") +
  theme_minimal() +
  labs(title = "Rozkład wieku klientów", x = "Wiek", y = "Liczba klientów")
```

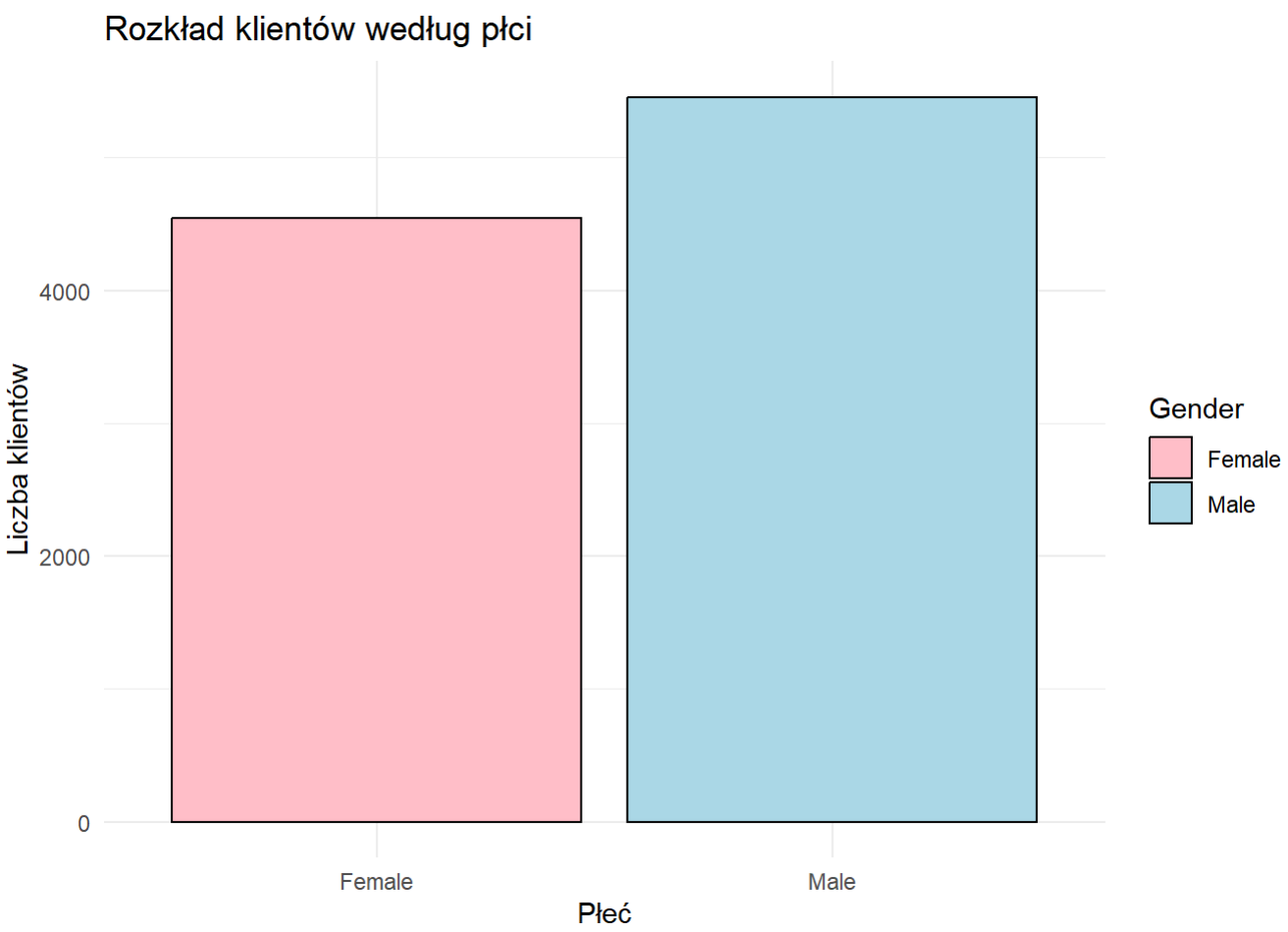


2.2. Rozkład danych ze względu na płeć

```
#Liczba klientów wg płci
gender_distribution <- table(Churn.Modeling$Gender)
print(gender_distribution)
```

```
##
## Female   Male
##   4543   5457
```

```
#Wizualizacja
ggplot(Churn.Modeling, aes(x = Gender, fill = Gender)) +
  geom_bar(color = "black") +
  scale_fill_manual(values = c("Female" = "pink", "Male" = "lightblue")) +
  theme_minimal() +
  labs(title = "Rozkład klientów według płci", x = "Płeć", y = "Liczba klientów")
```



2.3. Rozkład danych ze względu na kraj

```
geography_Churn.Modeling <- Churn.Modeling %>%
  group_by(Geography) %>%
  summarise(Count = n())

# Sprawdzam wyniki
print(geography_Churn.Modeling)
```

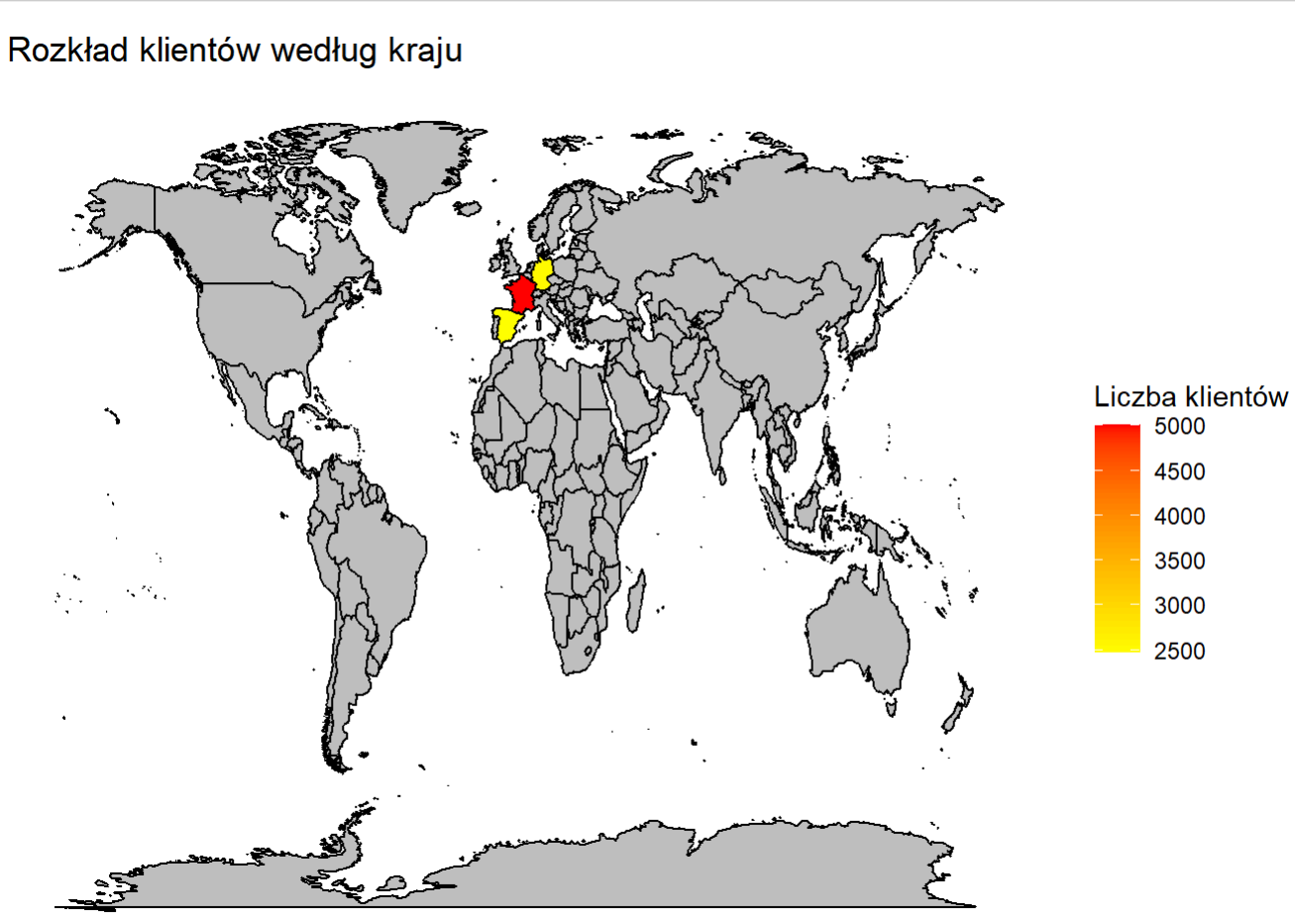
```
## # A tibble: 3 × 2
##   Geography Count
##   <fct>     <int>
## 1 France     5014
## 2 Germany    2509
## 3 Spain      2477
```

```
# Wczytuję mapę
world_map <- map_data("world")

# Dopasowanie nazw krajów z bazy do nazw na mapie
geography_Churn.Modeling <- geography_Churn.Modeling %>%
  mutate(Country = recode(Geography,
                          "Spain" = "Spain",
                          "Germany" = "Germany",
                          "France" = "France"))

# Połączenie danych z mapą
map_Churn.Modeling <- left_join(world_map, geography_Churn.Modeling, by = c("region" = "Country"))

# Wizualizacja
ggplot(map_Churn.Modeling, aes(x = long, y = lat, group = group, fill = Count)) +
  geom_polygon(color = "black") +
  scale_fill_gradient(low = "yellow", high = "red", na.value = "grey") +
  theme_minimal() +
  labs(title = "Rozkład klientów według kraju", fill = "Liczba klientów") +
  theme(axis.text = element_blank(),
        axis.title = element_blank(),
        panel.grid = element_blank())
```

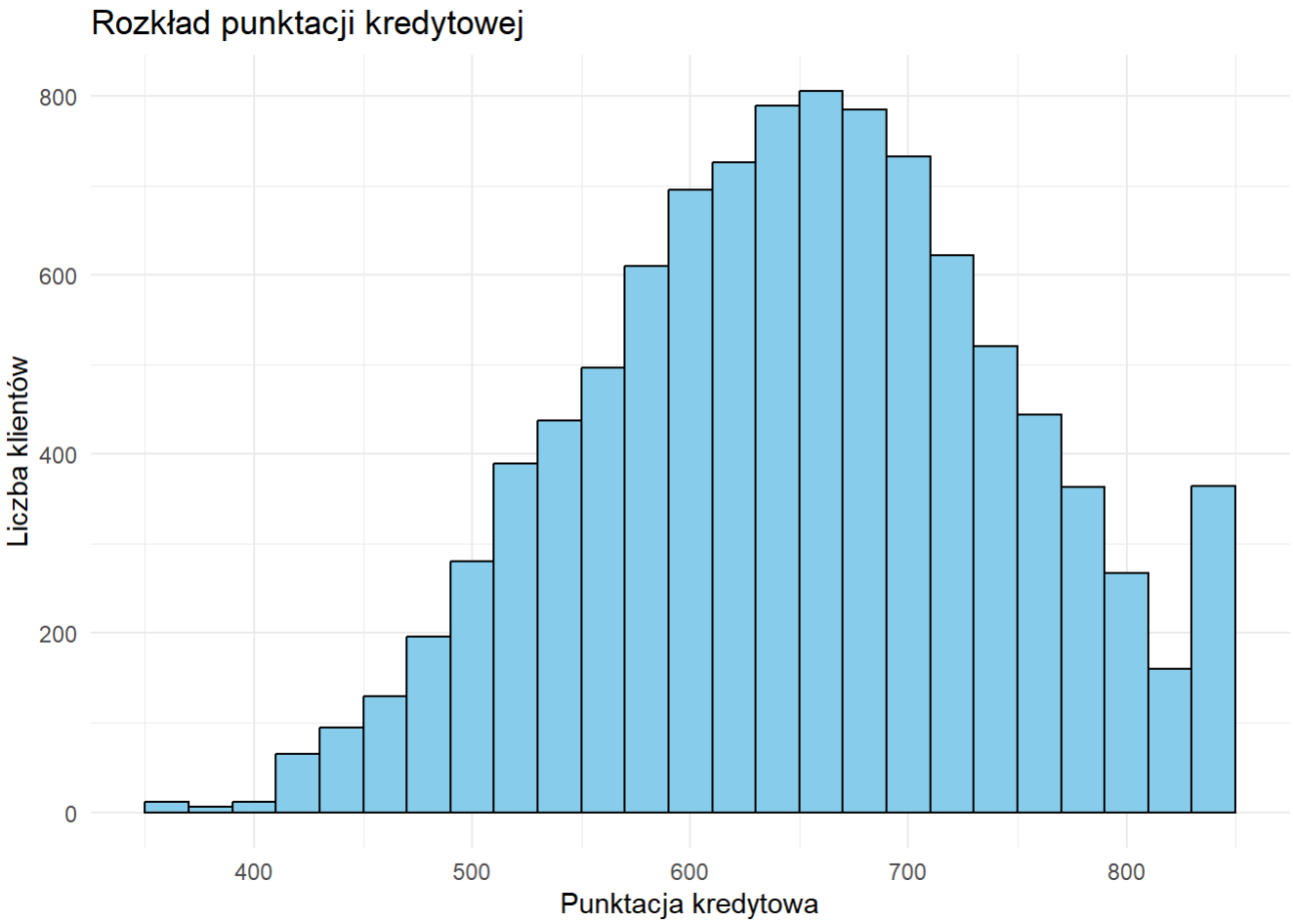


2.4. Rozkład danych ze względu na Credit Score

```
# Podstawowe statystyki dla punktacji kredytowej
creditscore_summary <- summary(Churn.Modeling$CreditScore)
print(creditscore_summary)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   350.0   584.0   652.0   650.5   718.0   850.0
```

```
# Histogram punktacji kredytowej
ggplot(Churn.Modeling, aes(x = CreditScore)) +
  geom_histogram(binwidth = 20, fill = "skyblue", color = "black") +
  theme_minimal() +
  labs(title = "Rozkład punktacji kredytowej", x = "Punktacja kredytowa", y = "Liczba klientów")
```



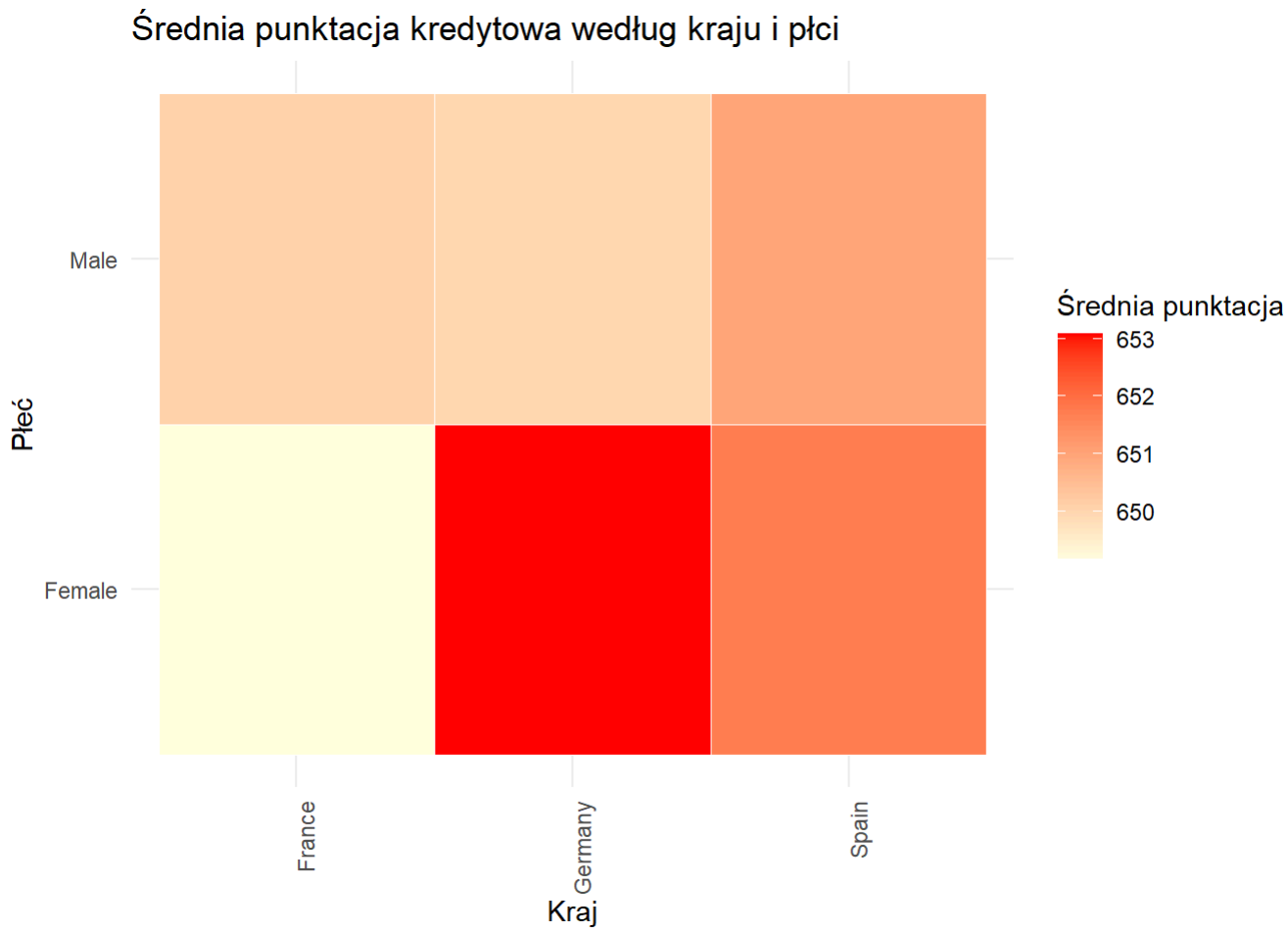
2.4.1. Czy Credit Score zależy od płci i kraju?

Analizę przeprowadzono za pomocą Heatmapy. Im kolor jest ciemniejszy, tym wyższe wartości średniej punktacji kredytowej reprezentuje.

```
heatmap_data <- Churn.Modeling %>%
  group_by(Geography, Gender) %>%
  summarise(Mean_CreditScore = mean(CreditScore, na.rm = TRUE))
```

```
## `summarise()` has grouped output by 'Geography'. You can override using the
## `.groups` argument.
```

```
# Heatmapa punktacji kredytowej według kraju i płci
ggplot(heatmap_data, aes(x = Geography, y = Gender, fill = Mean_CreditScore)) +
  geom_tile(color = "white") +
  scale_fill_gradient(low = "lightyellow", high = "red") +
  theme_minimal() +
  labs(title = "Średnia punktacja kredytowa według kraju i płci", x = "Kraj", y = "Płeć", fill = "Średn
ia punktacja") +
  theme(axis.text.x = element_text(angle = 90, hjust = 1))
```

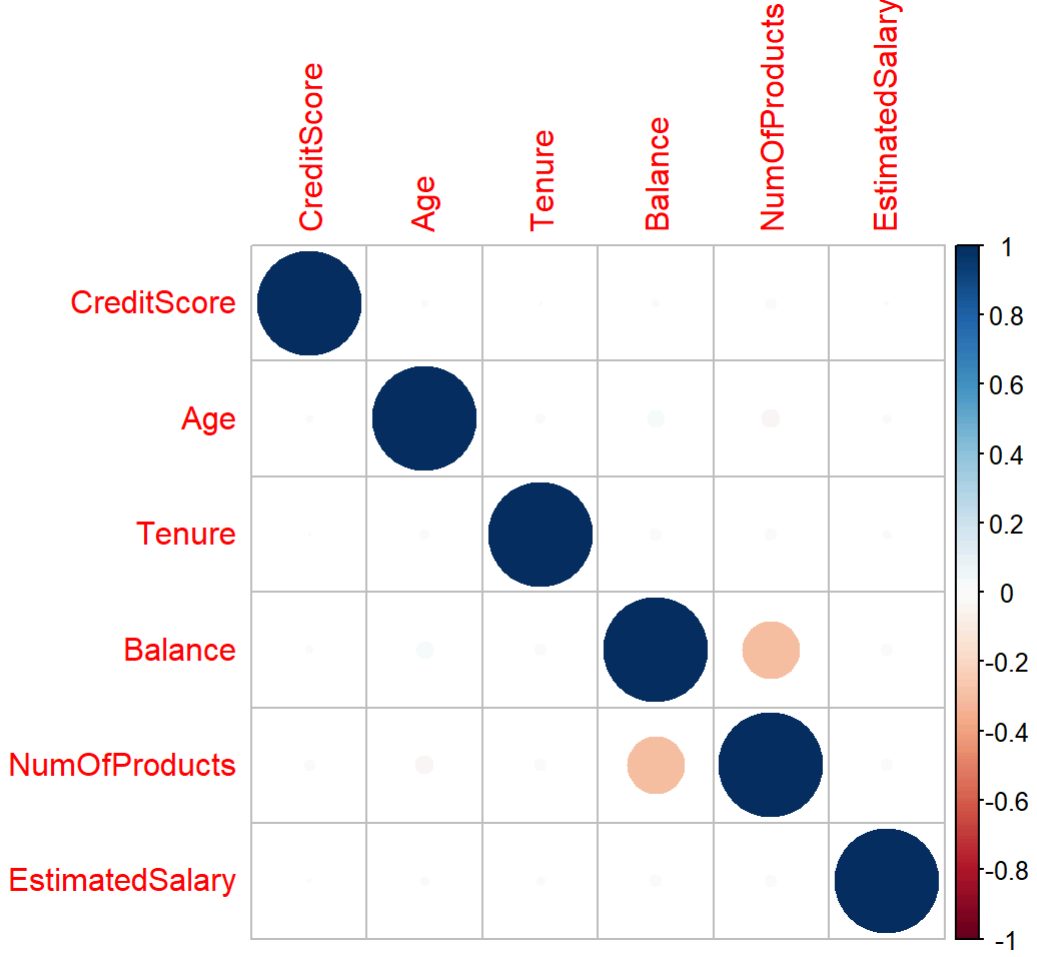


Wnioski: Generalnie różnice w średnich punktacjach kredytowej są stosunkowo niewielkie. Może to oznaczać, że systemy oceniania kredytowego w analizowanych krajach są do siebie zbliżone.

Najwyższą średnią punktację kredytową wyróżniają się kobiety w Niemczech, natomiast najniższą kobiety we Francji. W Hiszpani punktacje obu płci są do siebie bardzo podobne.

3. Analiza korelacji

```
# Korelacja między zmiennymi numerycznymi
corr_matrix <- cor(Churn.Modeling %>% select(CreditScore, Age, Tenure, Balance, NumOfProducts, Estimate
dSalary))
corrplot(corr_matrix, method = "circle")
```




```
# Korelacja ze zmienną "Exited"
correlation <- cor(Churn.Modeling %>% select(-Exited) %>% select_if(is.numeric), as.numeric(as.character(Churn.Modeling$Exited)))
print(correlation)
```

```
##                        [,1]
## RowNumber       -0.016571371
## CustomerId      -0.006247987
## CreditScore     -0.027093540
## Age              0.285323038
## Tenure          -0.014000612
## Balance          0.118532769
## NumOfProducts   -0.047819865
## EstimatedSalary  0.012096861
```

Wnioski:

1. Występuje słaba korelacja pomiędzy liczbą produktów a saldem konta. Zależność ta jest ujemna, a to oznacza, że gdy wzrasta liczba produktów, saldo konta maleje i odwrotnie.
2. Większość zmiennych ma bardzo słabą korelację ze zmienną “Exited”, a zatem zmienne takie jak: punkt kredytowy, długość zatrudnienia, liczba produktów i oszacowane wynagrodzenie nie mają silnego wpływu na decyzję o opuszczeniu banku.
3. Starsi klienci mają nieco wyższą tendencję do opuszczania banku, na co wskazuje umiarkowana dodatnia korelacja.
4. Klienci posiadający wyższe saldo konta mogą mieć nieco wyższą tendencję do opuszczania banku, co sugeruje dodatnia, choć słaba korekacja.

4. Klienci, którzy opuścili bank

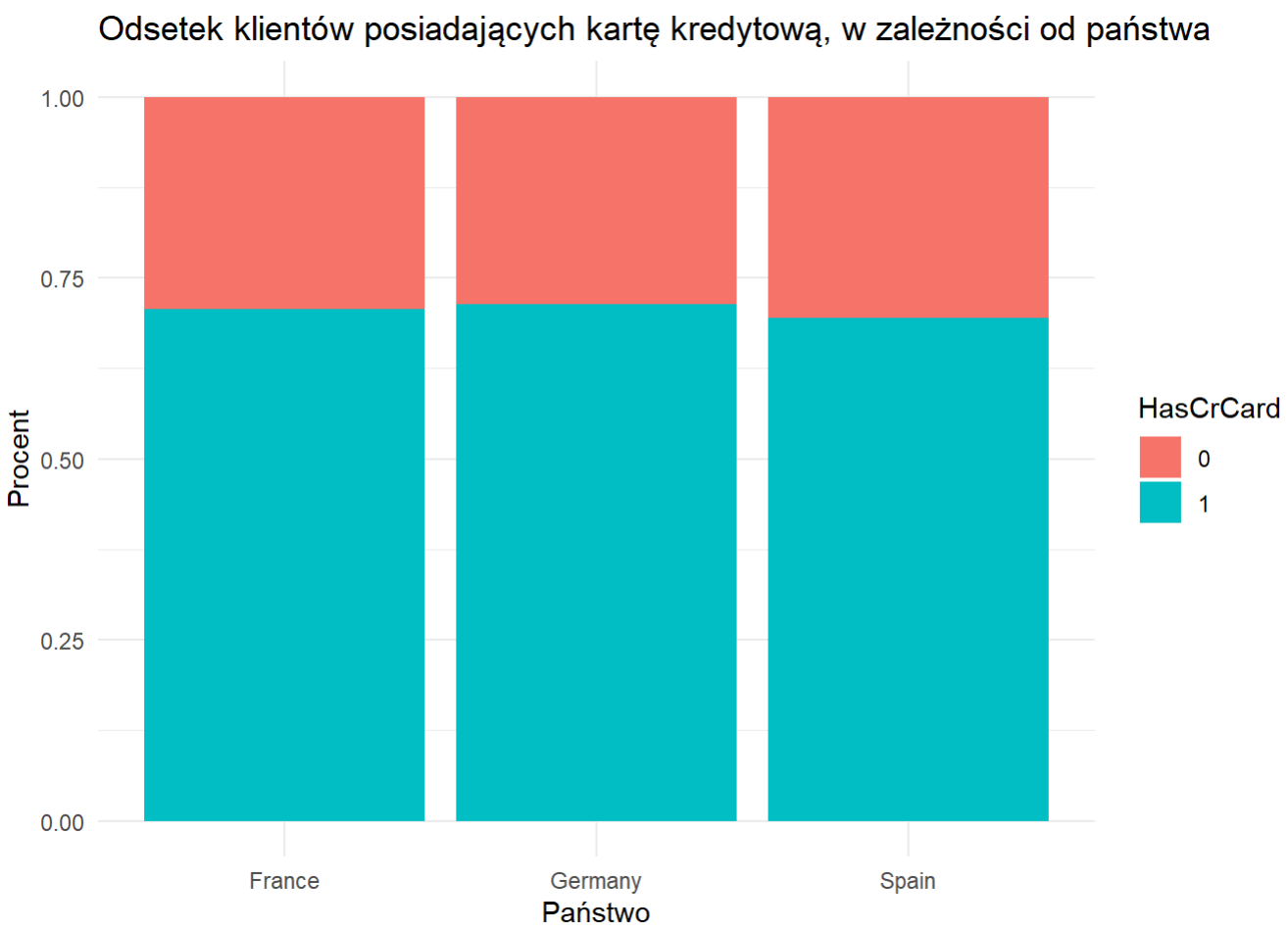
```
# Klienci, którzy odeszli, w zależności od państwa
ggplot(Churn.Modeling, aes(x = Geography, fill = Exited)) +
  geom_bar(position = "fill") +
  theme_minimal() +
  labs(title = "Odsetek klientów, którzy odeszli, w zależności od państwa", x = "Państwo", y = "Procen
t")
```


```
# Klienci, którzy odeszli, w zależności od płci
ggplot(Churn.Modeling, aes(x = Gender, fill = Exited)) +
  geom_bar(position = "fill") +
  theme_minimal() +
  labs(title = "Odsetek klientów, którzy odeszli, w zależności od płci", x = "Płeć", y = "Procent")
```

Wniosek: Wśród klientów, którzy opuścili bank największy odsetek stanowili klienci z Niemiec. Biorąc pod uwagę płeć, wśród klientów opuszczających bank przeważały kobiety.

5. Klienci posiadający kartę kredytową w zależności od państwa

```
# Klienci, którzy odeszli, w zależności od państwa
ggplot(Churn.Modeling, aes(x = Geography, fill = HasCrCard)) +
  geom_bar(position = "fill") +
  theme_minimal() +
  labs(title = "Odsetek klientów posiadających kartę kredytową, w zależności od państwa", x = "Państw
o", y = "Procent")
```



Wniosek: Zdecydowana większość klientów (około 75% z nich) posiada kartę kredytową.