

Raport - Telco

Filip Michewicz 282239
Wiktor Niedźwiedzki 258882

31 marca 2025 Anno Domini

Spis treści

1 Wprowadzenie	4
2 Przygotowanie danych. Podstawowe informacje o danych	5
2.1 Wczytywanie danych	5
2.2 Cechy danych	7
3 Analiza opisowa — wskaźniki sumaryczne	8
3.1 Wskaźniki sumaryczne	8
4 Opis wykonanych analiz	9
5 Analiza opisowa - wykresy	9
5.1 Wskaźniki korelacji	9
5.2 Histogramy i wykresy pudełkowe zmiennych ilościowych	11
5.3 Wykresy słupkowe zmiennych jakościowych	17
5.4 Wykresy rozrzutów	34
6 Analiza opisowa z podziałem na grupy	38
6.1 Histogramy i wykresy pudełkowe	38
6.2 Wykresy słupkowe zmiennych jakościowych	42
6.3 Wykresy rozrzutów	54
7 Dyskusja	59
8 Wnioski	59
9 Rekomendacje	60

Spis wykresów

1	Macierz Korelacji - Klienci Telco	10
2	Histogram dla zmiennej tenure	12
3	Wykres pudełkowy dla zmiennej tenure	13
4	Histogram dla zmiennej MonthlyCharges	14
5	Wykres pudełkowy dla zmiennej MonthlyCharges	15
6	Histogram dla zmiennej TotalCharges	16
7	Wykres pudełkowy dla zmiennej TotalCharges	17
8	Wykres słupkowy zmiennej SeniorCitizen	19
9	Wykres słupkowy zmiennej Partner	20
10	Wykres słupkowy zmiennej Dependents	21
11	Wykres słupkowy zmiennej PhoneService	22
12	Wykres słupkowy zmiennej MultipleLines	23
13	Wykres słupkowy zmiennej InternetService	24
14	Wykres słupkowy zmiennej OnlineSecurity	25
15	Wykres słupkowy zmiennej DeviceProtection	26
16	Wykres słupkowy zmiennej TechSupport	27
17	Wykres słupkowy zmiennej StreamingTV	28
18	Wykres słupkowy zmiennej StreamingMovies	29
19	Wykres słupkowy zmiennej Contract	30
20	Wykres słupkowy zmiennej PaperlessBilling	31
21	Wykres słupkowy zmiennej PaymentMethod	32
22	Wykres słupkowy zmiennej Churn	33
23	Zależność TotalCharges od MonthlyCharges z krzywą wygładzającą	34
24	Zależność TotalCharges od tenure z krzywą wygładzającą	35
25	Zależność MonthlyCharges od tenure z krzywą wygładzającą	36
26	Macierz par zmiennych numerycznych	37
27	Histogram dla zmiennej tenure z podziałem ze względu na churn	39
28	Wykres pudełkowy dla zmiennej tenure z podziałem ze względu na churn	40
29	Histogram dla zmiennej MonthlyCharges z podziałem ze względu na churn	41
30	Wykres pudełkowy dla zmiennej MonthlyCharges z podziałem ze względu na churn	42
31	Wykres pudełkowy dla zmiennej TotalCharges z podziałem ze względu na churn	43
32	Wykres słupkowy zmiennej Dependents z podziałem ze względu na churn	45
33	Wykres słupkowy zmiennej InternetService z podziałem ze względu na churn	46
34	Wykres słupkowy zmiennej OnlineSecurity z podziałem ze względu na churn	47
35	Wykres słupkowy zmiennej OnlineBackup z podziałem ze względu na churn	48

36	Wykres słupkowy zmiennej DeviceProtection z podziałem ze względu na churn	49
37	Wykres słupkowy zmiennej TechSupport z podziałem ze względu na churn	50
38	Wykres słupkowy zmiennej Contract z podziałem ze względu na churn	51
39	Wykres słupkowy zmiennej PaperlessBilling z podziałem ze względu na churn	52
40	Wykres słupkowy zmiennej PaymentMethod z podziałem ze względu na churn	53
41	Wykresy rozrzutów Total Charges ze względu na tenure z podziałem ze względu na zmienną churn	55
42	Wykresy rozrzutów MonthlyCharges ze względu na tenure z podziałem ze względu na zmienną churn	56
43	Macierz par zmiennych numerycznych z podziałem ze względu na churn	58

Spis tabel

1	Opis zmiennych w zbiorze danych	7
2	Tabela z podstawowymi wskaźnikami sumarycznymi dla zmiennych ilościowych	8

```
library(corrplot)
```

```
## Warning: pakiet 'corrplot' został zbudowany w wersji R 4.4.3
```

```
## corrplot 0.95 loaded
```

```
library(ggplot2)
```

```
## Warning: pakiet 'ggplot2' został zbudowany w wersji R 4.4.3
```

```
library(gridExtra)
library(e1071)
```

```
## Warning: pakiet 'e1071' został zbudowany w wersji R 4.4.3
```

```
library(xtable)
```

```
## Warning: pakiet 'xtable' został zbudowany w wersji R 4.4.2
```

```
library(knitr)
```

```
## Warning: pakiet 'knitr' został zbudowany w wersji R 4.4.3
```

```
library(plyr)
```

```
## Warning: pakiet 'plyr' został zbudowany w wersji R 4.4.3
```

```

library(dplyr)

##
## Dołączanie pakietu: 'dplyr'

## Następujące obiekty zostały zakryte z 'package:plyr':
##
##      arrange, count, desc, failwith, id, mutate, rename, summarise,
##      summarize

## Następujący obiekt został zakryty z 'package:gridExtra':
##
##      combine

## Następujące obiekty zostały zakryte z 'package:stats':
##
##      filter, lag

## Następujące obiekty zostały zakryte z 'package:base':
##
##      intersect, setdiff, setequal, union

library(GGally)

## Warning: pakiet 'GGally' został zbudowany w wersji R 4.4.2

## Registered S3 method overwritten by 'GGally':
##   method from
##   +.gg   ggplot2

j <- 0
l <- 1
k <- 0
m <- 0

```

1 Wprowadzenie

W ramach tego zadania przeanalizowano dane dotyczące klientów firmy telekomunikacyjnej Telco, ze szczególnym uwzględnieniem problemu odpływu klientów (churn). Cele dokonanej analizy:

- Zbadanie podstawowych cech zbioru danych, zarówno zmiennych ilościowych, jak i jakościowych.
- Identyfikacja zależności między zmiennymi oraz ocena ich rozkładów.
- Porównanie grup lojalnych klientów (Churn = ‘No’) i tych, którzy odeszli (Churn = ‘Yes’).

Chcemy znaleźć odpowiedzi na pytania:

- Jakie są podstawowe właściwości analizowanych zmiennych?
- Które cechy najlepiej różnicują klientów lojalnych od tych, którzy rezygnują z usług?
- Jakie mogą być główne przyczyny odchodzenia klientów i co firma może zrobić, aby temu przeciwdziałać?

2 Przygotowanie danych. Podstawowe informacje o danych

W tej części dokonano wstępniego opracowania danych, identyfikując wszystkie zmienne oraz korygując ich typy w celu zapewnienia poprawności dalszej analizy. Zamieniono również nazwy będące oryginalnie po angielsku na język polski w celu zwiększenia czytelności oraz ułatwienia interpretacji danych.

2.1 Wczytywanie danych

```
klienci <- read.csv(file="WA_Fn-UseC_-Telco-Customer-Churn.csv",
                     stringsAsFactors=TRUE)

# Niepoprawnie wczytane typy dla: customerID oraz SeniorCitizen

klienci$SeniorCitizen <- as.factor(klienci$SeniorCitizen)
klienci$SeniorCitizen <- revalue(klienci$SeniorCitizen,
                                    c("0" = "Nie", "1" = "Tak"))

klienci$customerID <- as.character(klienci$customerID)

df_table <- data.frame(
  Typ = sapply(klienci, class),
  Opis = c("Unikalny identyfikator klienta",
          "Płeć klienta (Mężczyzna, Kobieta)",
          "Czy klient jest seniorem? (Tak/Nie)",
          "Czy klient ma partnera/partnerkę? (Tak/Nie)",
          "Czy klient ma osoby na utrzymaniu? (Tak/Nie)",
          "Liczba miesięcy, przez które klient korzystał z usług",
          "Czy klient ma usługę telefoniczną? (Tak/Nie)",
          "Czy klient ma wiele linii telefonicznych? (Tak/Nie/Brak usługi telefonicznej)",
          "Typ usługi internetowej (DSL, Światłowód, Brak)",
          "Czy klient ma usługę zabezpieczeń online? (Tak/Nie/Brak internetu)",
          "Czy klient ma usługę kopii zapasowej online? (Tak/Nie/Brak internetu)",
          "Czy klient ma usługę ochrony urządzenia? (Tak/Nie/Brak internetu)",
          "Czy klient ma usługę wsparcia technicznego? (Tak/Nie/Brak internetu)",
          "Czy klient korzysta z usługi streamingu TV? (Tak/Nie/Brak internetu)",
          "Czy klient korzysta z usługi streamingu filmów? (Tak/Nie/Brak internetu)",
          "Rodzaj umowy (Miesięczna, Rocznia, Dwuletnia)",
          "Czy klient korzysta z faktur elektronicznych? (Tak/Nie)",
          "Metoda płatności (np. karta kredytowa, przelew, automatyczne obciążenie)",
          "Miesięczna opłata za usługi",
          "Łączna opłata pobrana od klienta",
          "Czy klient zrezygnował z usług? (Tak/Nie")
  )

# Zamiana nazw angielski na nazwy polskie

klienci$gender <- revalue(klienci$gender,
                           c("Female"="Kobieta", "Male"="Mężczyzna"))
klienci$Partner <- revalue(klienci$Partner,
                           c("Yes"="Tak", "No"="Nie"))
klienci$Dependents <- revalue(klienci$Dependents,
```

```

            c("Yes"="Tak", "No"="Nie"))
klienci$PhoneService <- revalue(klienci$PhoneService,
                                  c("Yes"="Tak", "No"="Nie"))
klienci$MultipleLines <- revalue(klienci$MultipleLines,
                                    c("Yes"="Tak", "No"="Nie",
                                      "No phone service"="Brak usługi telefonicznej"))
klienci$InternetService <- revalue(klienci$InternetService,
                                       c("No"="Brak", "Fiber optic"="Światłowód"))
klienci$OnlineSecurity <- revalue(klienci$OnlineSecurity,
                                     c("No"="Nie", "Yes"="Tak",
                                        "No internet service"="Brak internetu"))
klienci$OnlineBackup <- revalue(klienci$OnlineBackup,
                                 c("No"="Nie", "Yes"="Tak",
                                    "No internet service"="Brak internetu"))
klienci$DeviceProtection <- revalue(klienci$DeviceProtection,
                                       c("No"="Nie", "Yes"="Tak",
                                         "No internet service"="Brak internetu"))
klienci$OnlineSecurity <- revalue(klienci$OnlineSecurity,
                                     c("No"="Nie", "Yes"="Tak",
                                        "No internet service"="Brak internetu"))

## The following 'from' values were not present in 'x': No, Yes, No internet service

klienci$TechSupport <- revalue(klienci$TechSupport,
                                 c("No"="Nie", "Yes"="Tak",
                                    "No internet service"="Brak internetu"))
klienci$StreamingTV <- revalue(klienci$StreamingTV,
                                 c("No"="Nie", "Yes"="Tak",
                                    "No internet service"="Brak internetu"))
klienci$StreamingMovies <- revalue(klienci$StreamingMovies,
                                       c("No"="Nie", "Yes"="Tak",
                                         "No internet service"="Brak internetu"))
klienci$Contract <- revalue(klienci$Contract, c("Month-to-month"="Miesięczna",
                                                 "One year"="Roczną",
                                                 "Two year"="Dwuletnią"))
klienci$PaperlessBilling <- revalue(klienci$PaperlessBilling,
                                       c("Yes"="Tak", "No"="Nie"))
klienci$PaymentMethod <- revalue(klienci$PaymentMethod,
                                   c("Bank transfer (automatic)"=
                                      "Przelew bankowy \n (automatyczny)",
                                      "Credit card (automatic)"=
                                         "Karta kredytowa \n (automatyczna)",
                                      "Electronic check"=
                                         "Czek elektroniczny",
                                      "Mailed check"=
                                         "Czek wysyłany pocztą"))
klienci$Churn <- revalue(klienci$Churn,
                           c("Yes"="Tak", "No"="Nie"))

```

W Tabeli 1. przedstawiono nazwy zmiennych, ich typy oraz opis, który wyjaśnia, co każda z nich reprezentuje.

```
kable(df_table, col.names = c("Zmienna", "Typ", "Opis"),
      caption = "Opis zmiennych w zbiorze danych")
```

Tabela 1: Opis zmiennych w zbiorze danych

Zmienna	Typ	Opis
customerID	character	Unikalny identyfikator klienta
gender	factor	Płeć klienta (Mężczyzna, Kobieta)
SeniorCitizen	factor	Czy klient jest seniorem? (Tak/Nie)
Partner	factor	Czy klient ma partnera/partnerkę? (Tak/Nie)
Dependents	factor	Czy klient ma osoby na utrzymaniu? (Tak/Nie)
tenure	integer	Liczba miesięcy, przez które klient korzysta/korzystał z usług
PhoneService	factor	Czy klient ma usługę telefoniczną? (Tak/Nie)
MultipleLines	factor	Czy klient ma wiele linii telefonicznych? (Tak/Nie/Brak usługi telefonicznej)
InternetService	factor	Typ usługi internetowej (DSL, Światłowód, Brak)
OnlineSecurity	factor	Czy klient ma usługę zabezpieczeń online? (Tak/Nie/Brak internetu)
OnlineBackup	factor	Czy klient ma usługę kopii zapasowej online? (Tak/Nie/Brak internetu)
DeviceProtection	factor	Czy klient ma usługę ochrony urządzenia? (Tak/Nie/Brak internetu)
TechSupport	factor	Czy klient ma usługę wsparcia technicznego? (Tak/Nie/Brak internetu)
StreamingTV	factor	Czy klient korzysta z usługi streamingu TV? (Tak/Nie/Brak internetu)
StreamingMovies	factor	Czy klient korzysta z usługi streamingu filmów? (Tak/Nie/Brak internetu)
Contract	factor	Rodzaj umowy (Miesięczna, Rocznica, Dwuletnia)
PaperlessBilling	factor	Czy klient korzysta z faktur elektronicznych? (Tak/Nie)
PaymentMethod	factor	Metoda płatności (np. karta kredytowa, przelew, automatyczne obciążenie)
MonthlyCharges	numeric	Miesięczna opłata za usługi
TotalCharges	numeric	Łączna opłata pobrana od klienta
Churn	factor	Czy klient zrezygnował z usług? (Tak/Nie)

2.2 Cechy danych

- Zbiór danych zawiera 7043 obserwacje (wiersze) oraz 21 zmiennych (kolumn).
- Zmienne *tenure*, *MonthlyCharges* oraz *TotalCharges* mają typ ilościowy (*numeric*).
- Zmienna *customerID* to identyfikator klienta zapisany jako ciąg znaków.
- Pozostałe zmienne mają typ kategoryczny (*factor*).
- Poza unikalnym identyfikatorem klienta, wszystkie pozostałe zmienne są istotne, ponieważ umożliwiają analizę popularności usług, metod płatności oraz średnich kosztów miesięcznych.

```
klienci$customerID <- NULL
```

- W zbiorze danych występuje 11 brakujących wartości w kolumnie *TotalCharges*, oznaczonych jako *NA*. Braki te można wyjaśnić faktem, że dotyczą nowych klientów, którzy jeszcze nie dokonali płatności.
- Nietypowym rozwiązaniem jest kodowanie zmiennej *SeniorCitizen* za pomocą wartości *0* i *1*, zamiast bardziej czytelnych kategorii, takich jak *Tak* i *Nie*.

3 Analiza opisowa — wskaźniki sumaryczne

3.1 Wskaźniki sumaryczne

Dla danych ilościowych obliczono podstawowe wskaźniki sumaryczne, takie jak minimum, pierwszy i trzeci kwartyl, medianę, maksimum, średnią odchylenie standardowe, skośność i kurtozę.

```
num_cols <- klienci[, sapply(klienci, is.numeric)]  
  
# Obliczenie statystyk opisowych dla każdej kolumny  
num_stats <- apply(num_cols, 2, function(x) {  
  c(  
    Min = min(x, na.rm = TRUE),  
    "1st Qu" = quantile(x, 0.25, na.rm = TRUE),  
    Mediana = median(x, na.rm = TRUE),  
    Średnia = mean(x, na.rm = TRUE),  
    "3rd Qu" = quantile(x, 0.75, na.rm = TRUE),  
    Max = max(x, na.rm = TRUE),  
    "Odch.Stand." = sd(x, na.rm = TRUE),  
    "Skośność" = skewness(x, na.rm = TRUE),  
    "Kurtoza" = kurtosis(x, na.rm = TRUE)  
  )  
})  
  
kable(num_stats, digits=3, caption="Tabela z podstawowymi wskaźnikami sumarycznymi dla zmiennych ilościowych")
```

Tabela 2: Tabela z podstawowymi wskaźnikami sumarycznymi dla zmiennych ilościowych

	tenure	MonthlyCharges	TotalCharges
Min	0.000	18.250	18.800
1st Qu.25%	9.000	35.500	401.450
Mediana	29.000	70.350	1397.475
Średnia	32.371	64.762	2283.300
3rd Qu.75%	55.000	89.850	3794.738
Max	72.000	118.750	8684.800
Odch.Stand.	24.559	30.090	2266.771
Skośność	0.239	-0.220	0.961
Kurtoza	-1.388	-1.258	-0.233

Kurtoza dla wszystkich zmiennych (*tenure*, *MonthlyCharges*, *TotalCharges*) jest ujemna, co oznacza, że ich rozkłady są bardziej płaskie niż rozkład normalny (platykurytyczne). Wskazuje to na większe rozproszenie wartości, czyli brak wyraźnego szczytu i mniej ekstremalnych wartości w porównaniu do typowego rozkładu normalnego.

Skośność informuje o asymetrii rozkładu:

- *tenure* (0.239) – lekko prawostronnie skośny, czyli więcej klientów ma krótszy staż, ale występują też dłuższe okresy.

- *MonthlyCharges* (-0.220) – lekko lewostronnie skośny, co oznacza, że większość klientów płaci stosunkowo niskie kwoty, ale zdarzają się wyższe wartości.
- *TotalCharges* (0.961) – wyraźnie prawostronnie skośny, co sugeruje, że większość klientów zapłaciła stosunkowo niewiele, ale istnieją przypadki z bardzo wysokimi wartościami całkowitych opłat.

4 Opis wykonanych analiz

W analizie przeprowadzono szereg wizualizacji, które miały na celu dokładne zidentyfikowanie kluczowych trendów oraz zależności w analizowanych danych.

Główne elementy analizy obejmowały: - Stworzenie macierzy korelacji, mającej na celu uchwycenie zależności między zmiennymi ilościowymi, takimi jak długość współpracy a łączna wysokość opłat. - Przygotowanie histogramów i wykresów pudełkowych, które umożliwiły szczegółowe zrozumienie rozkładów zmiennych liczbowych oraz ich asymetrii. - Wykonanie wykresów słupkowych dla zmiennych jakościowych, pozwalających na ocenę struktury klientów, np. pod kątem rodzaju umowy czy usług dodatkowych. - Zastosowanie wykresów rozrzutów z nałożonymi krzywymi wygładzającymi, które ukazały zależności między zmiennymi liczbowymi, takimi jak opłaty miesięczne a ryzyko rezygnacji. - Realizację wizualizacji z podziałem na grupy klientów (lojalni vs. rezygnujący), co umożliwiło ocenę różnic w ich zachowaniach.

5 Analiza opisowa - wykresy

W tej części przedstawiono dane w postaci wykresów i dokonano ich podstawowej analizy.

5.1 Wskaźniki korelacji

```
klienci_clean <- na.omit(klienci)

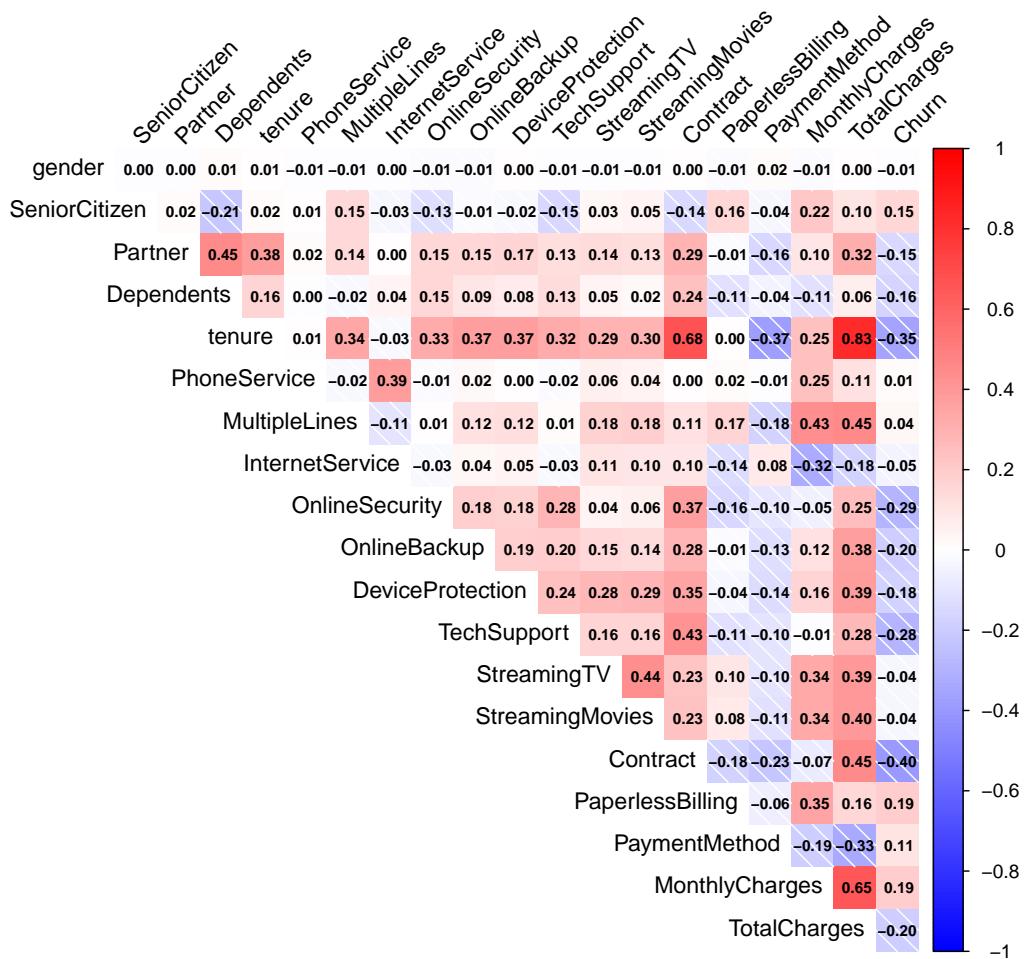
df_numeric <- klienci_clean %>%
  mutate(across(where(is.factor), ~ as.numeric(factor(.)))) %>%
  select(where(is.numeric))

corr_matrix <- cor(df_numeric)

corrplot(corr_matrix, method = "shade", type = "upper",
          col = colorRampPalette(c("blue", "white", "red"))(200),
          tl.col = "black", tl.srt = 45,
          addCoef.col = "black",
          diag = FALSE,
          number.cex=0.7)
```

Na Wykresie 1. przedstawiono macierz korelacji dla wczytanych danych. Widać wyraźne zależności, między innymi:

- Klienci z dłuższym okresem współpracy (*Tenure*) oraz posiadający dłuższe umowy (*Contract*) rzadziej opuszczają firmę, co przekłada się na niższy wskaźnik churn.
- Posiadanie usług, takich jak *OnlineSecurity*, *TechSupport* czy *OnlineBackup*, wiąże się z mniejszym odsetkiem klientów, którzy zdecydowali się odejść.



Wykres 1: Macierz Korelacji - Klienci Telco

- Wyższe miesięczne rachunki (*MonthlyCharges*) oraz cechy, takie jak korzystanie z usług internetowych i e-fakturowanie (*PaperlessBilling*), korelują z wyższym odsetkiem klientów rezygnujących z usług.

Te obserwacje stanowią podstawę do dalszych, bardziej szczegółowych analiz.

5.2 Histogramy i wykresy pudełkowe zmiennych ilościowych

```

wykresy <- list()

for (var in names(num_cols)) {
  # Histogram
  wykres <- ggplot(num_cols, aes_string(x = var)) +
    geom_histogram(bins = 20, fill = "lightblue", color = "black") +
    labs(title = NULL, x = var, y = "Liczność") +
    theme_minimal()

  wykresy <- c(wykresy, list(wykres))

  # Boxplot
  wykres <- ggplot(num_cols, aes_string(y = var)) +
    geom_boxplot(fill = "lightgreen", color = "black") +
    labs(title = NULL, x = NULL, y = var) +
    theme_minimal() +
    coord_flip()

  wykresy <- c(wykresy, list(wykres))
}

## Warning: `aes_string()` was deprecated in ggplot2 3.0.0.
## i Please use tidy evaluation idioms with `aes()`.
## i See also `vignette("ggplot2-in-packages")` for more information.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.

m <- m + 1
print(wykresy[[m]])

```

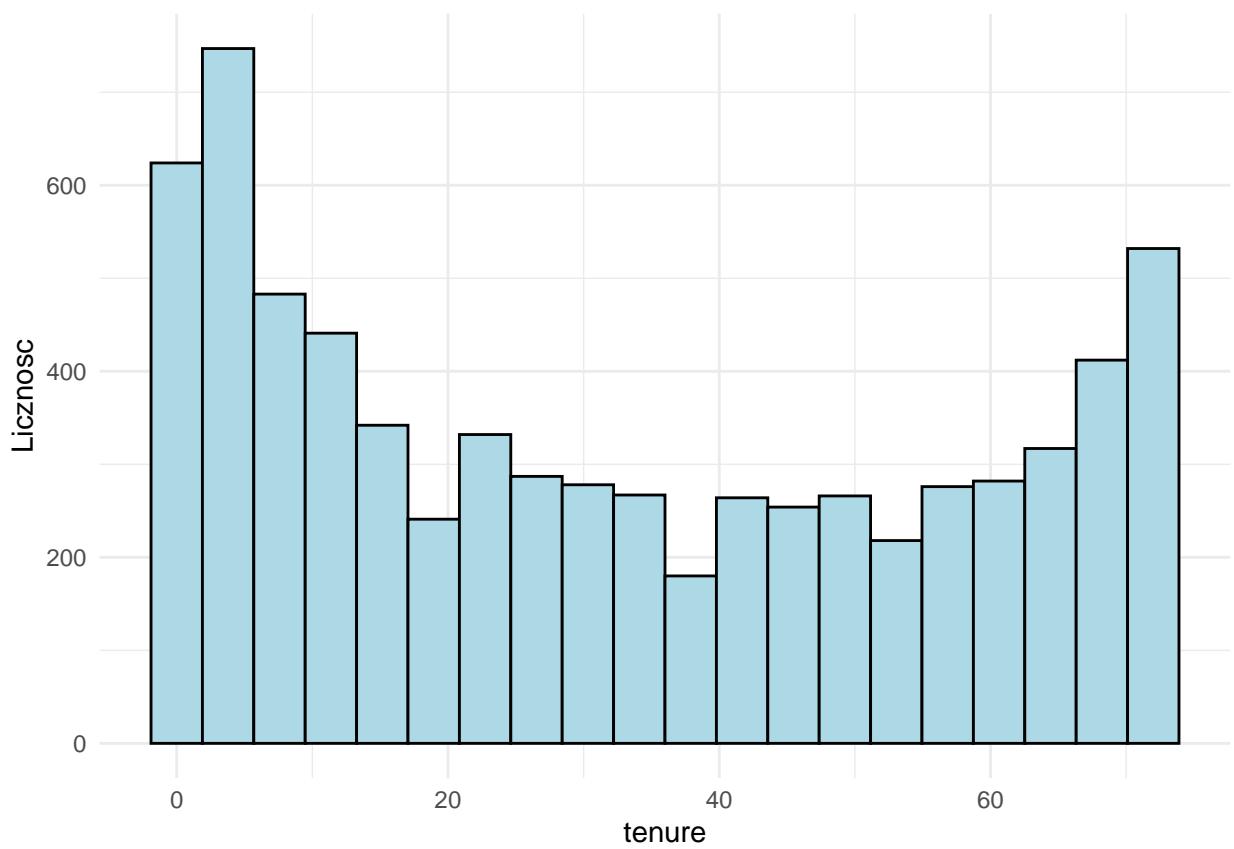
Wykres 2. przedstawia histogram zmiennej *tenure*, który ilustruje długość korzystania przez klientów z usług firmy. Rozkład ten jest dość równomierny, choć zauważalna jest większa liczba klientów, którzy korzystali z usług przez krótki okres czasu. Wskazuje to na wyraźną grupę klientów, którzy niedawno rozpoczęli korzystanie z naszych usług lub zakończyli współpracę w krótkim czasie.

```

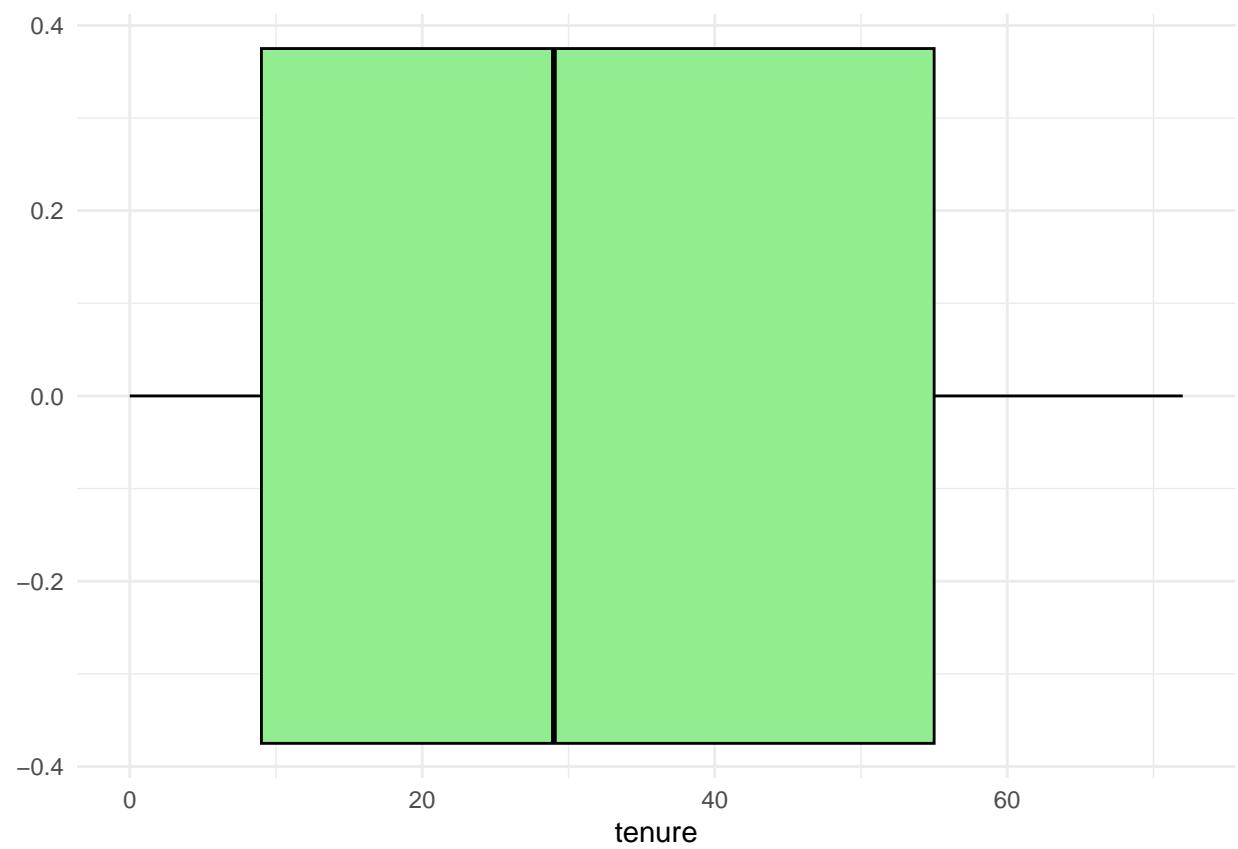
m <- m + 1
print(wykresy[[m]])

```

Wykres 3. przedstawia wykres pudełkowy dla zmiennej *tenure*. Z danych wynika, że wśród naszych klientów znajdują się zarówno osoby nowo pozyskane, jak i ci, którzy współpracują z nami od 6 lat. Środkowa część próby wskazuje, że większość klientów jest z nami od 9 do 55 miesięcy.

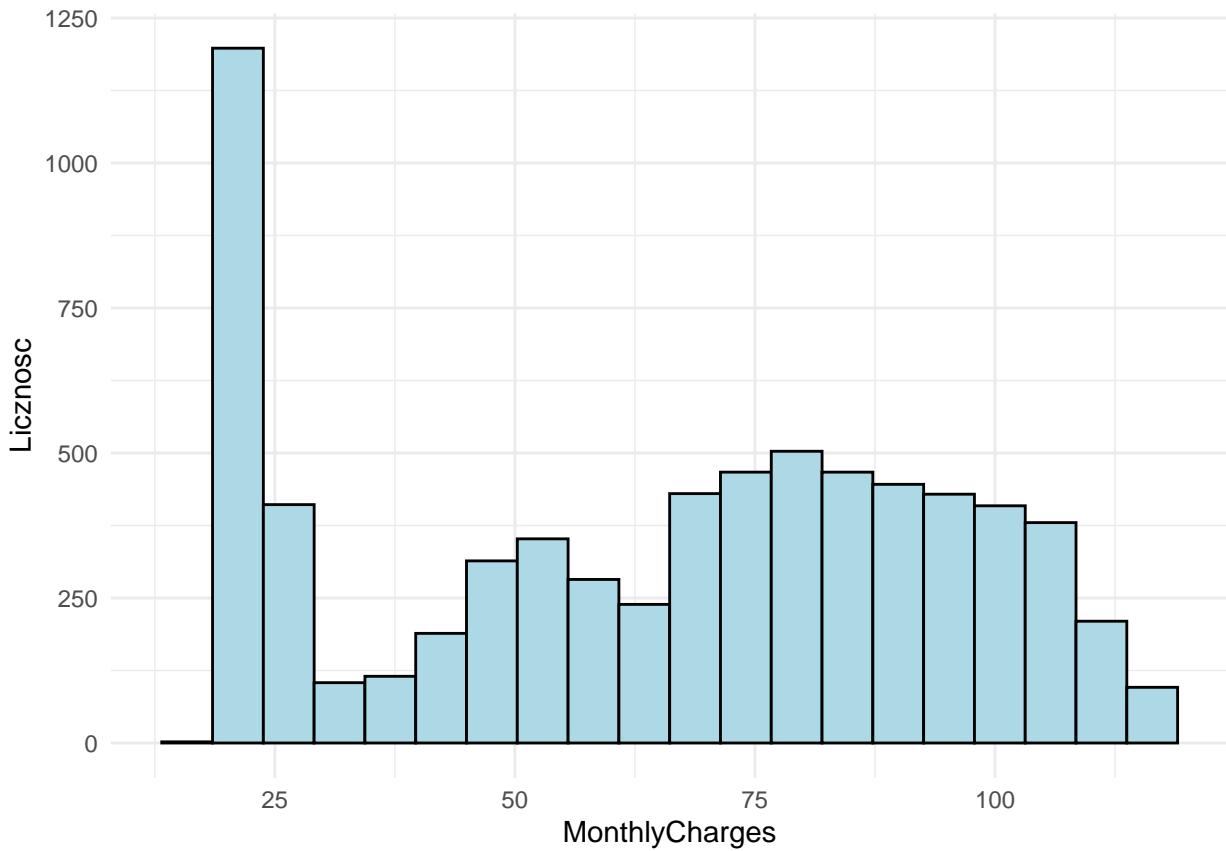


Wykres 2: Histogram dla zmiennej tenure



Wykres 3: Wykres pudełkowy dla zmiennej tenure

```
m <- m + 1
print(wykresy[[m]])
```



Wykres 4: Histogram dla zmiennej *MonthlyCharges*

Wykres 4. przedstawia histogram zmiennej *MonthlyCharges*. Z danych wynika, że wielu klientów korzysta z największych usług, ale wyraźnie wyróżnia się grupa klientów, którzy wydają od około 60 do 120 dolarów miesięcznie.

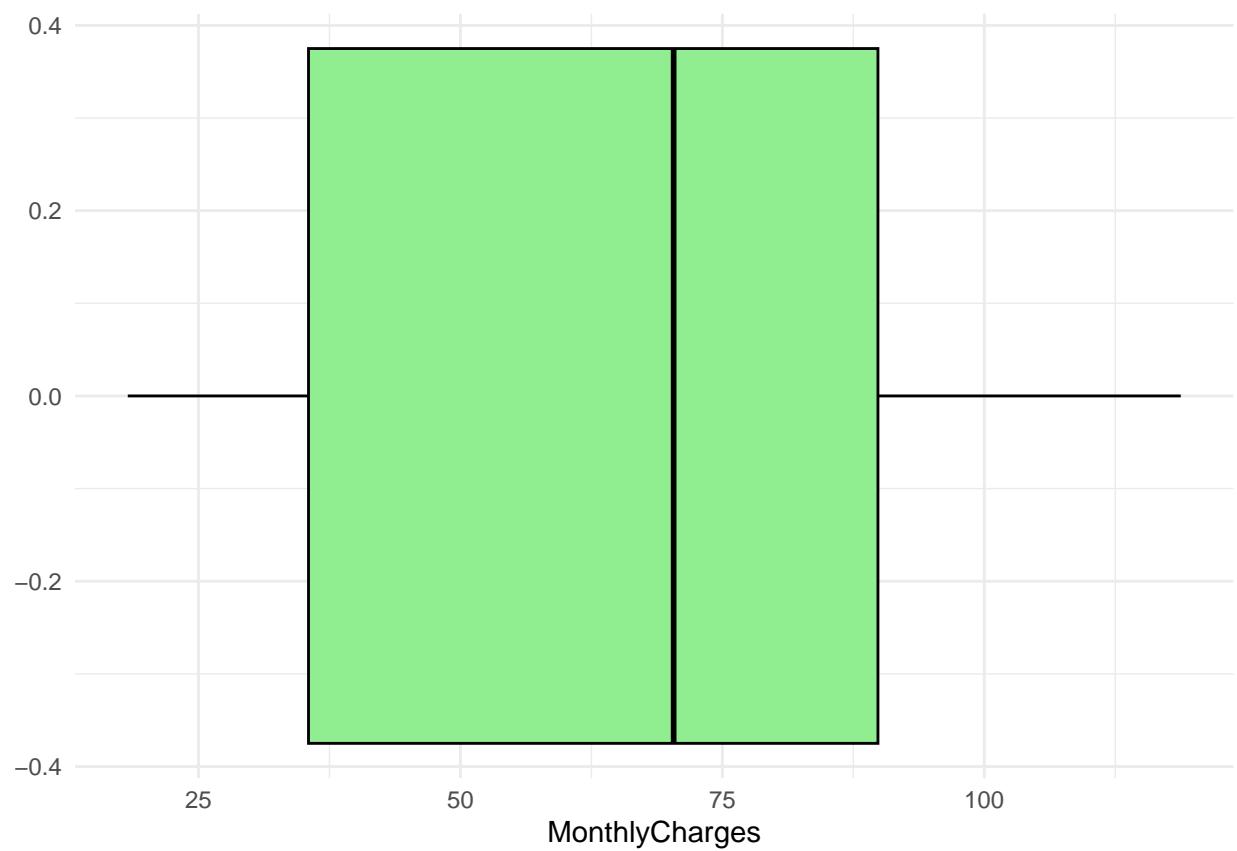
```
m <- m + 1
print(wykresy[[m]])
```

Wykres 5. przedstawia wykres pudelkowy dla zmiennej *MonthlyCharges*. Widać na nim, że większa część klientów ponosi raczej niższe opłaty, jednak mediana jest przesunięta w prawą stronę pudelka, co świadczy o asymetrii prawostronnej (dodatniej) rozkładu.

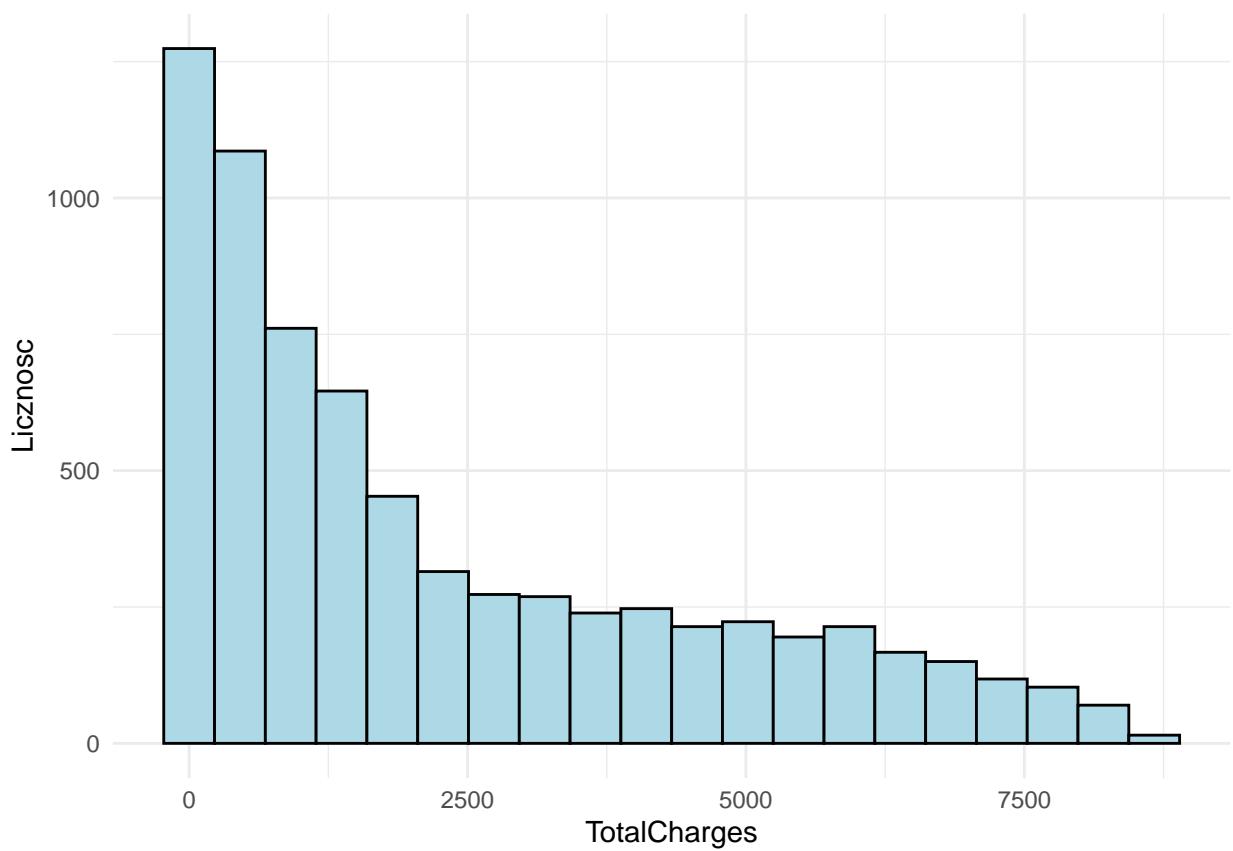
```
m <- m + 1
print(wykresy[[m]])
```

```
## Warning: Removed 11 rows containing non-finite outside the scale range
## ('stat_bin()').
```

Wykres 6. przedstawia histogram dla zmiennej *Total Charges*. Zgodnie z oczekiwaniemi, rozkład jest malejący, jednak warto zwrócić uwagę na tempo tego spadku. Dla niższych wartości (poniżej 2500 dolarów)



Wykres 5: Wykres pudełkowy dla zmiennej MonthlyCharges

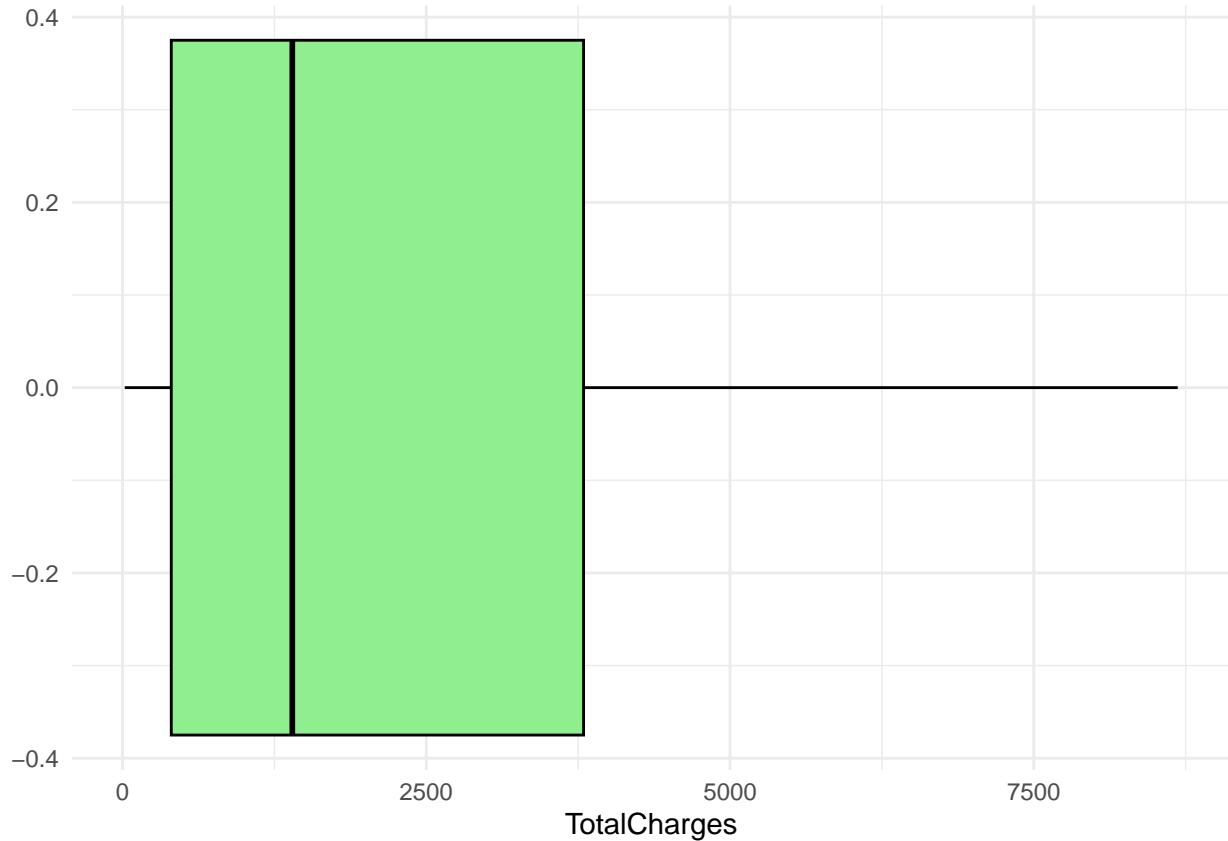


Wykres 6: Histogram dla zmiennej `TotalCharges`

liczność obserwacji zmniejsza się stosunkowo szybko, natomiast w dalszej części rozkładu tempo spadku jest łagodniejsze. Może to sugerować, że większość klientów generuje raczej niższe łączne opłaty, a wyższe wartości występują rzadziej, ale w sposób bardziej rozproszony.

```
m <- m + 1  
print(wykresy[[m]])
```

```
## Warning: Removed 11 rows containing non-finite outside the scale range  
## ('stat_boxplot()').
```



Wykres 7: Wykres pudełkowy dla zmiennej TotalCharges

Wykres 7. przedstawia wykres pudełkowy dla zmiennej *Total Charges*. Pudełko oraz mediana są przesunięte w lewą stronę, co wskazuje, że większość klientów ponosi raczej niższe łączne opłaty. Świadczy to o asymetrii prawostronnej rozkładu, gdzie mniejsze wartości są bardziej skupione, a wyższe rozłożone w dłuższym ogonie. Zaskakujący jest brak obserwacji odstających, co może sugerować, że wartości *Total Charges* są stosunkowo jednolite i nie występują znacząco odbiegające przypadki.

5.3 Wykresy słupkowe zmiennych jakościowych

```
wykresy <- list()  
tytuły <- c()
```

```

# Identyfikatory kolumn jakościowych
zmienne_jakosciowe <- c(1,2,3,4,6,8,15,16,17,20)
zmienne_bez_tel <- c(7)
zmienne_bez_net <- c(9,10,11,12,13,14)

for (i in seq_along(colnames(klienci))) {
  nazwa_zmiennej <- colnames(klienci)[i]

  # Wybór danych z odpowiednim filtrowaniem
  if (i %in% zmienne_jakosciowe) {
    dane <- klienci
  } else if (i %in% zmienne_bez_tel) {
    dane <- klienci %>% filter(.data[[nazwa_zmiennej]] != "Brak usługi telefonicznej")
  } else if (i %in% zmienne_bez_net) {
    dane <- klienci %>% filter(.data[[nazwa_zmiennej]] != "Brak internetu")
  } else {
    next
  }

  t <- paste("Wykres słupkowy zmiennej", nazwa_zmiennej)
  wykresy <- append(wykresy,
    list(ggplot(dane, aes(x=.data[[nazwa_zmiennej]],
      y=after_stat(count)/sum(after_stat(count)),
      fill=.data[[nazwa_zmiennej]])) +
      geom_bar() +
      scale_y_continuous(labels = scales::percent) +
      labs(x="", y="") +
      theme_minimal() +
      theme(legend.position="none",
        axis.text.x = element_text(color = "black")))
  )
}

tytuły <- c(tytuły, t)
}

par(mfrow = c(1, 1))
j <- j+1
#wykresy[[j]]

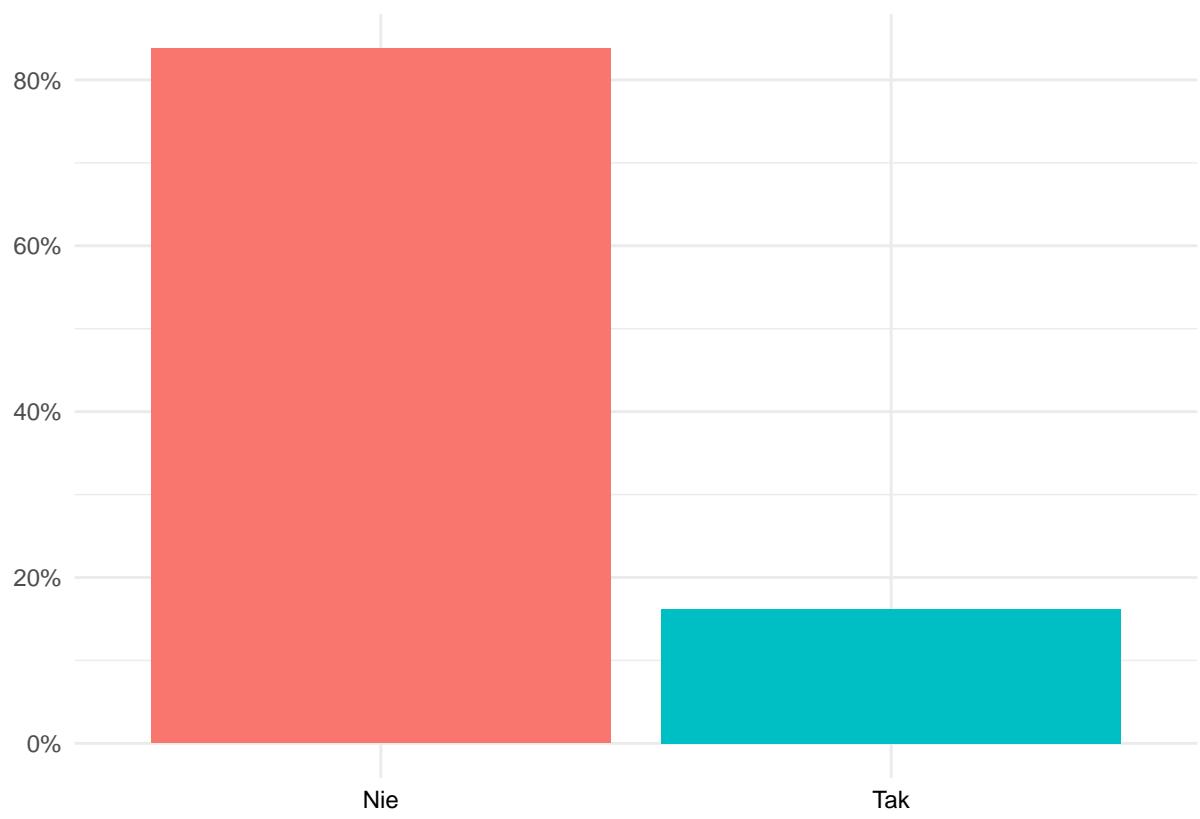
j <- j+1
wykresy[[j]]

```

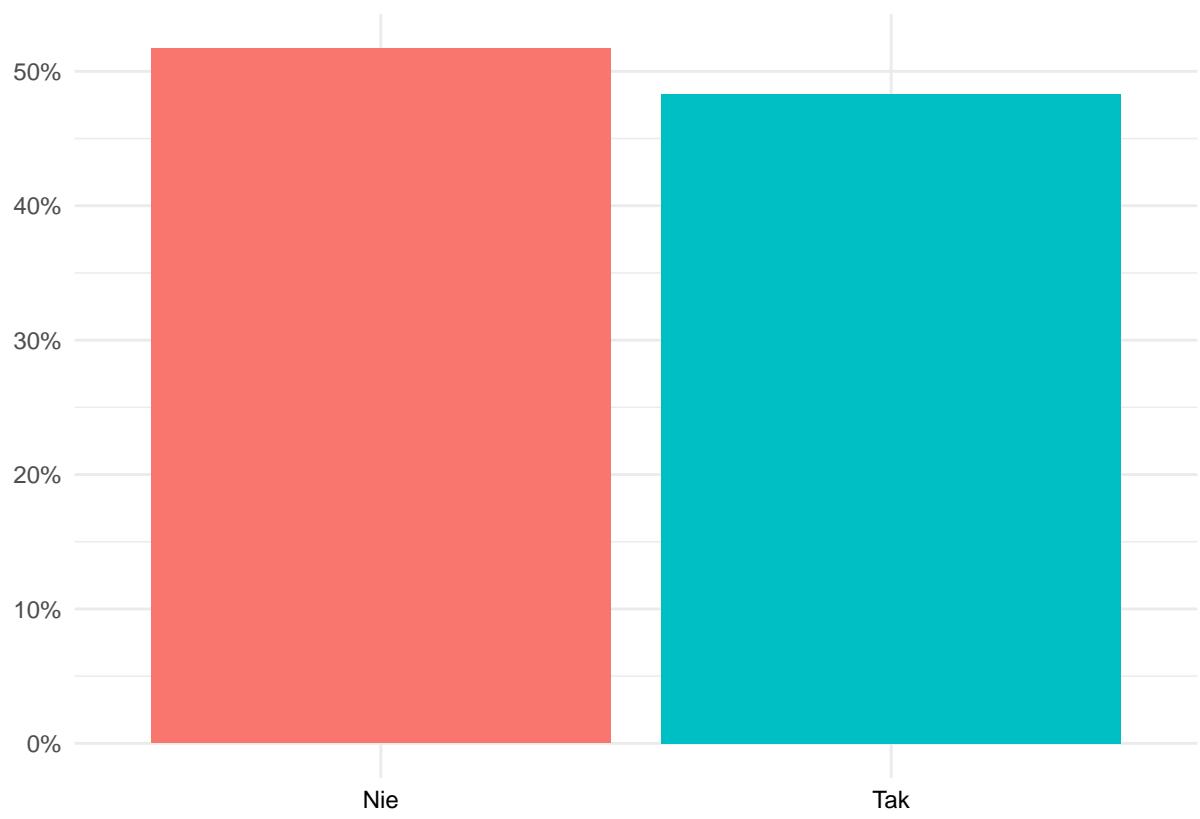
Na wykresie 8. przedstawiono wykres słupkowy dla zmiennej *SeniorCitizen*. Wynika z niego, że większość klientów to osoby, które nie są seniorami. Oznacza to, że w analizowanej grupie przeważają młodsi użytkownicy, co może mieć wpływ na preferencje dotyczące usług oraz wzorce płatności.

```
j <- j+1
wykresy[[j]]
```

Na wykresie 9. przedstawiono wykres słupkowy dla zmiennej *Partner*. Wynika z niego, że rozkład posiadania partnera jest dość równomierny – liczba klientów będących w związku jest zbliżona do liczby osób, które nie mają partnera. Taki wynik sugeruje, że status partnerski nie jest czynnikiem silnie dominującym wśród analizowanych klientów.

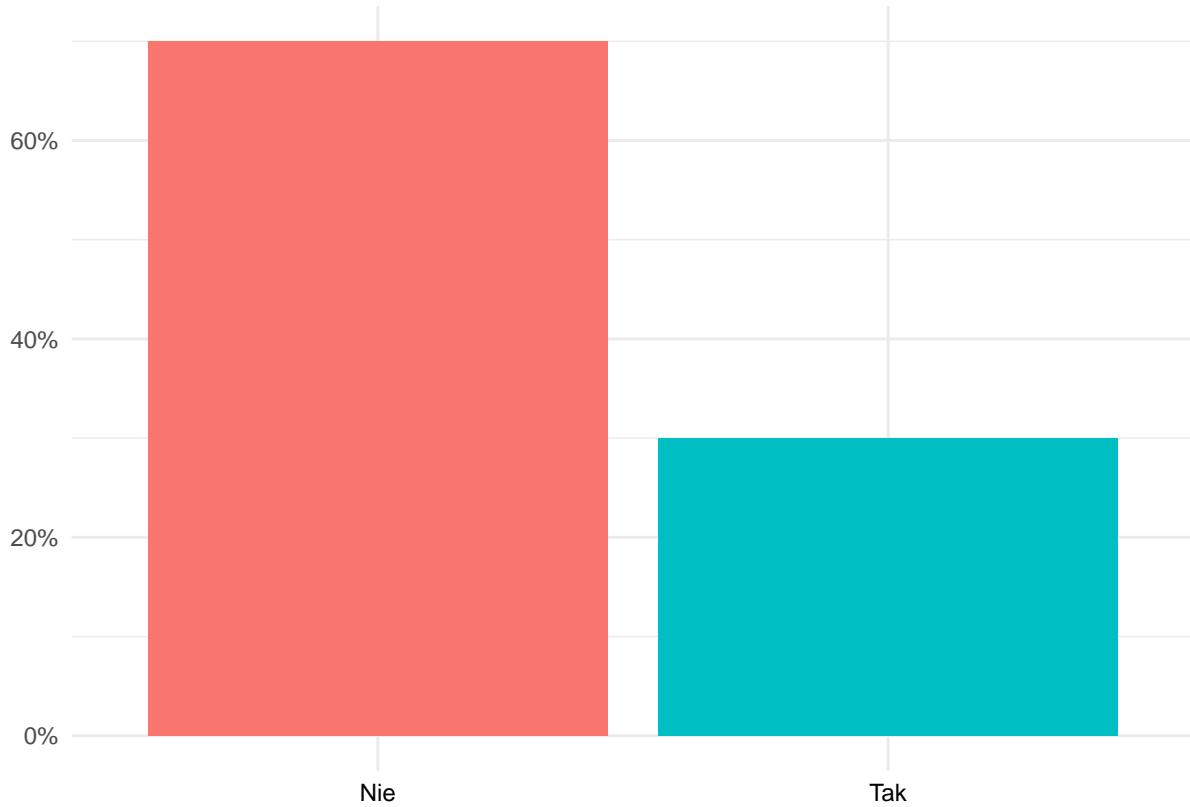


Wykres 8: Wykres słupkowy zmiennej SeniorCitizen



Wykres 9: Wykres słupkowy zmiennej Partner

```
j <- j+1  
wykresy[[j]]
```



Wykres 10: Wykres słupkowy zmiennej Dependents

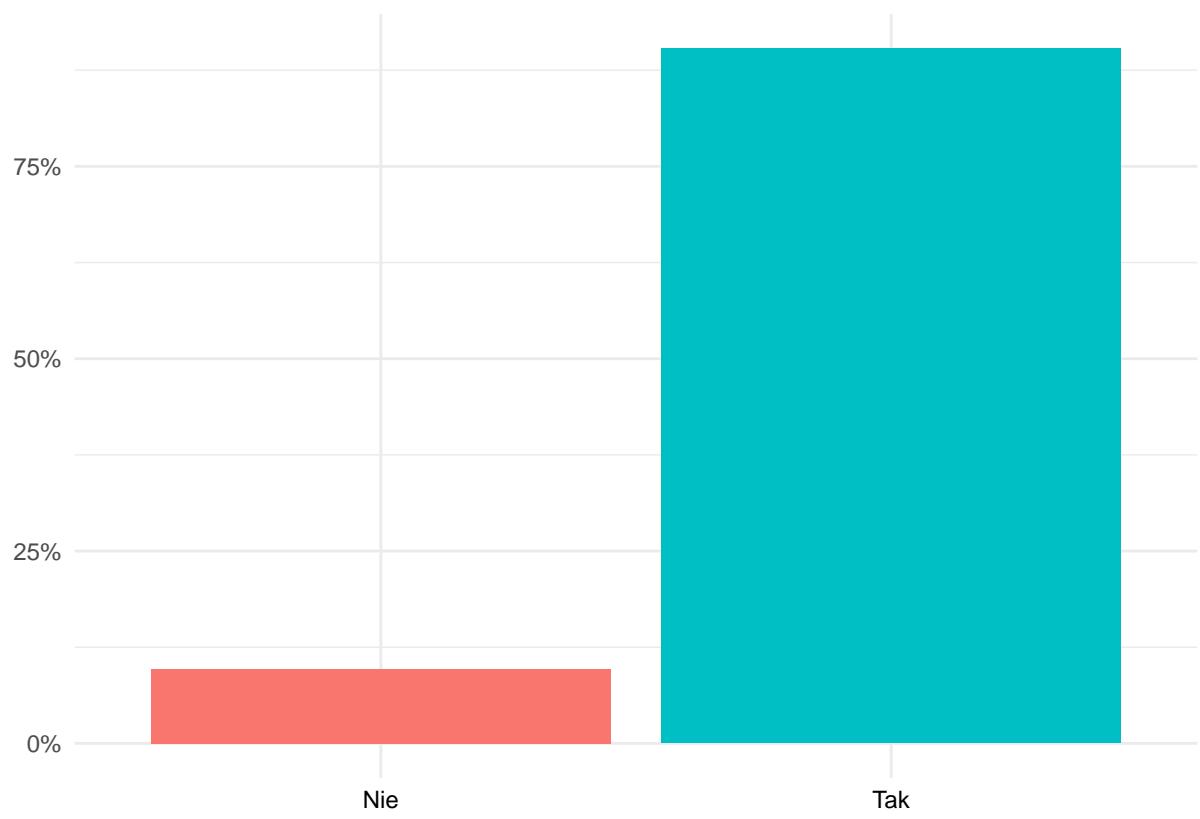
Na wykresie 10. przedstawiono wykres słupkowy dla zmiennej *Dependents*. Wynika z niego, że większość klientów (około 70%) nie ma osób na swoim utrzymaniu. Może to sugerować, że znaczna część użytkowników to osoby samodzielne finansowo, co może wpływać na ich preferencje dotyczące usług – mogą częściej wybierać bardziej elastyczne opcje lub krótsze zobowiązania, niekoniecznie kierując się potrzebami rodzin.

```
j <- j+1  
wykresy[[j]]
```

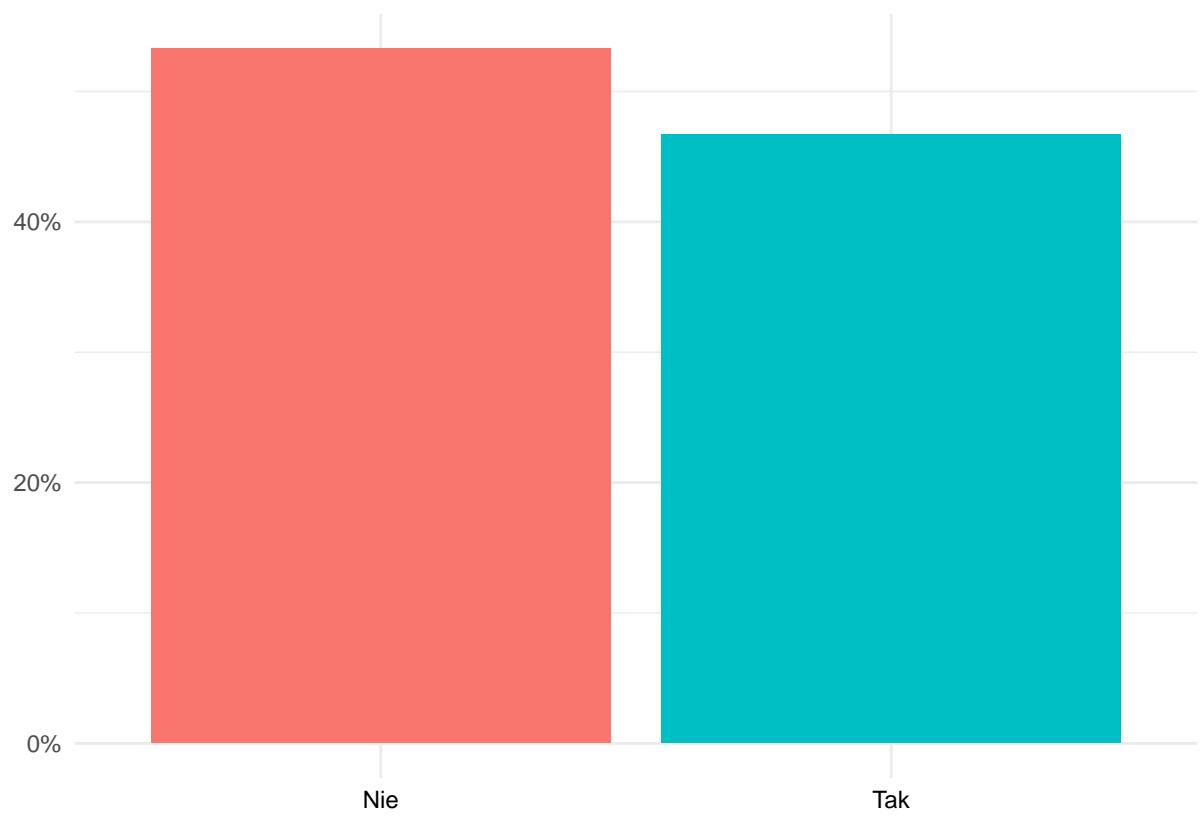
Na wykresie 11. przedstawiono wykres słupkowy dla zmiennej *PhoneService*. Wynika z niego, że zdecydowana większość klientów korzystała z usług telefonicznych, podczas gdy jedynie niewielka grupa (około 10%) nie miała aktywnej usługi telefonicznej. Może to sugerować, że telefonia jest nadal istotnym elementem oferty, choć pewna część klientów może preferować inne formy komunikacji.

```
j <- j+1  
wykresy[[j]]
```

Na wykresie 12. przedstawiono wykres słupkowy dla zmiennej *MultipleLines* wśród klientów, którzy korzystają z usług telefonicznych. Wynika z niego, że wśród tych klientów niewielka większość posiada tylko jedną linię telefoniczną, podczas gdy nieco mniejsza grupa zdecydowała się na korzystanie z wielu linii. Może to



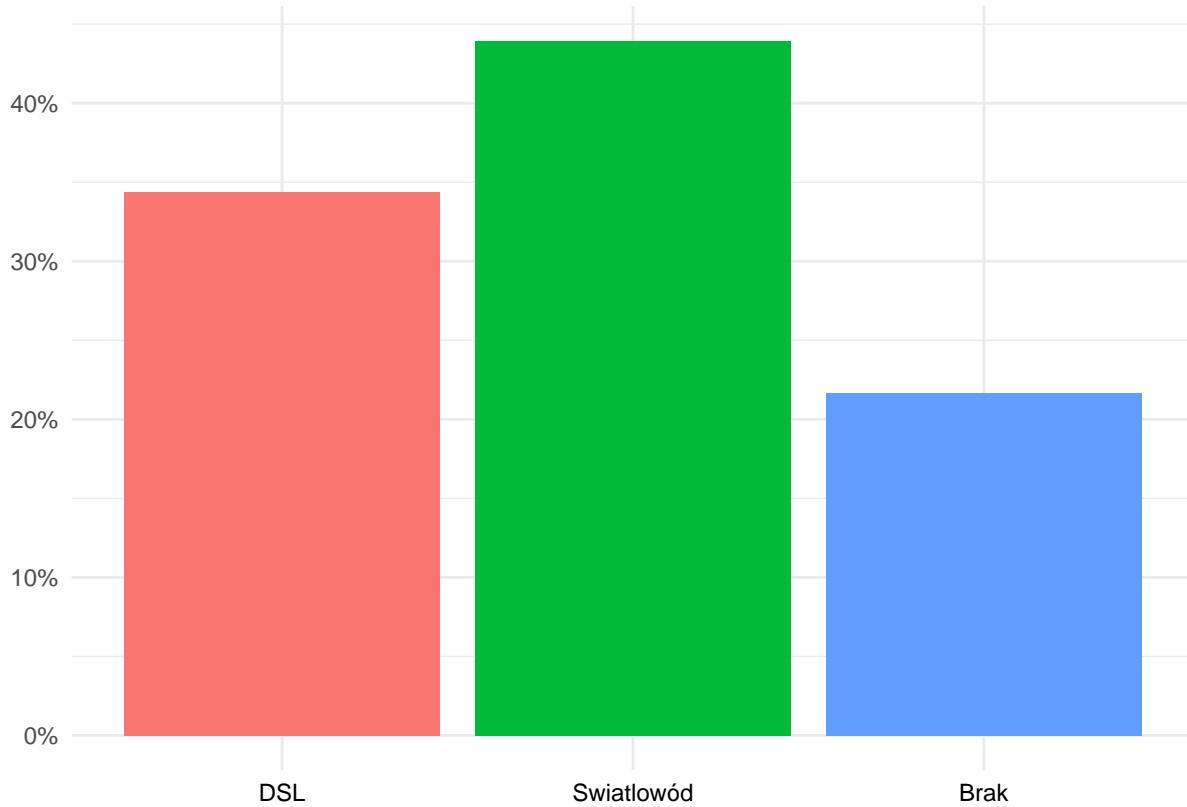
Wykres 11: Wykres słupkowy zmiennej PhoneService



Wykres 12: Wykres słupkowy zmiennej `MultipleLines`

sugerować, że dodatkowe linie telefoniczne nie cieszą się dużą popularnością, co może być efektem indywidualnych potrzeb użytkowników lub specyfiki gospodarstw domowych, które nie wymagają więcej niż jednej linii telefonicznej.

```
j <- j+1  
wykresy[[j]]
```

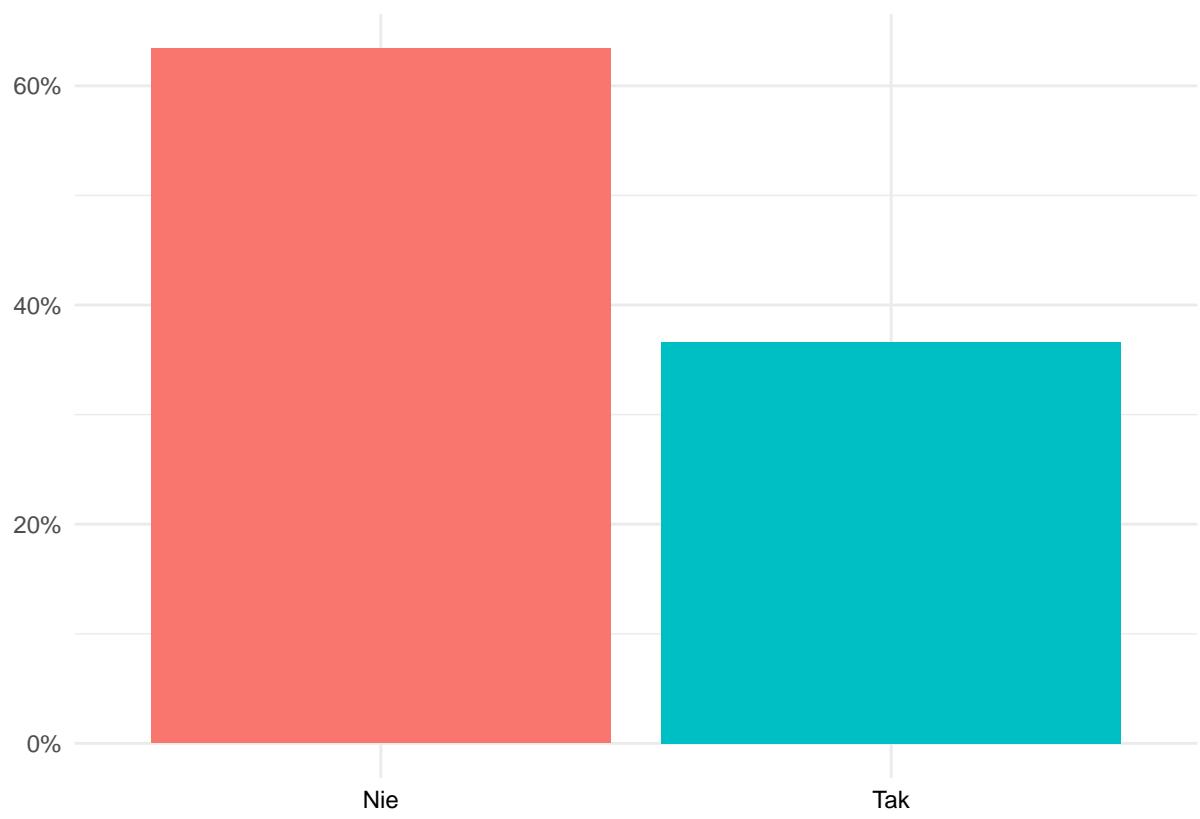


Wykres 13: Wykres słupkowy zmiennej *InternetService*

Na wykresie 13. przedstawiono wykres słupkowy dla zmiennej *InternetService*. Wynika z niego, że około 22% naszych klientów nie korzysta z usług internetowych, natomiast większość (około 43%) korzysta z rozwiązań światłowodowych. Pozostałe 35% korzysta z połączenia DSL. To wskazuje na różnorodność preferencji w zakresie usług internetowych, z dominacją technologii światłowodowych, które mogą oferować szybsze i bardziej stabilne połączenia.

```
j <- j+1  
wykresy[[j]]
```

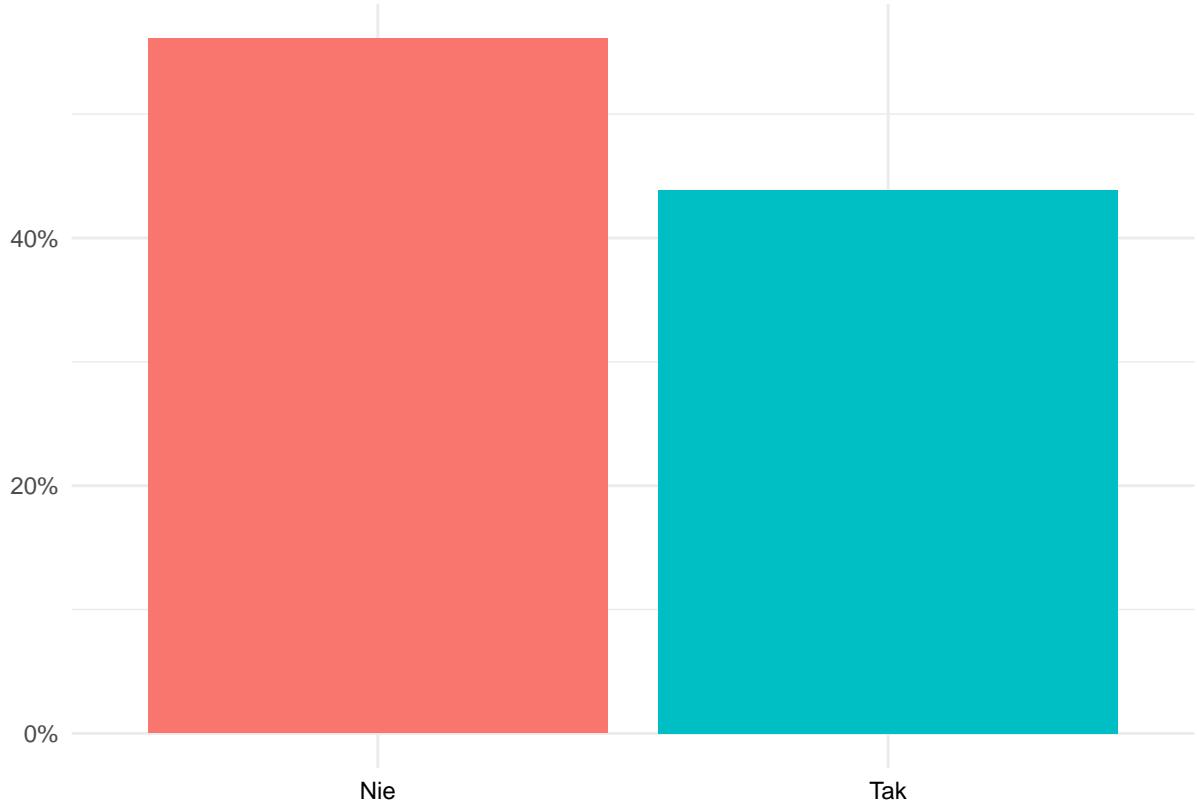
Na wykresie 14. przedstawiono wykres słupkowy dla zmiennej *OnlineSecurity* wśród klientów, którzy korzystają z usług internetowych. Widać, że większość z nich nie korzysta z tej usługi, co może sugerować, że klienci nie zwracają wystarczającej uwagi na bezpieczeństwo w internecie. Może to mieć reperkusje w przyszłości, gdyż brak odpowiedniej ochrony online może prowadzić do większego ryzyka związanych z bezpieczeństwem, co w konsekwencji może skutkować rezygnacją z naszej usługi.



Wykres 14: Wykres słupkowy zmiennej OnlineSecurity

```
j <- j+1  
#wykresy[[j]]
```

```
j <- j+1  
wykresy[[j]]
```

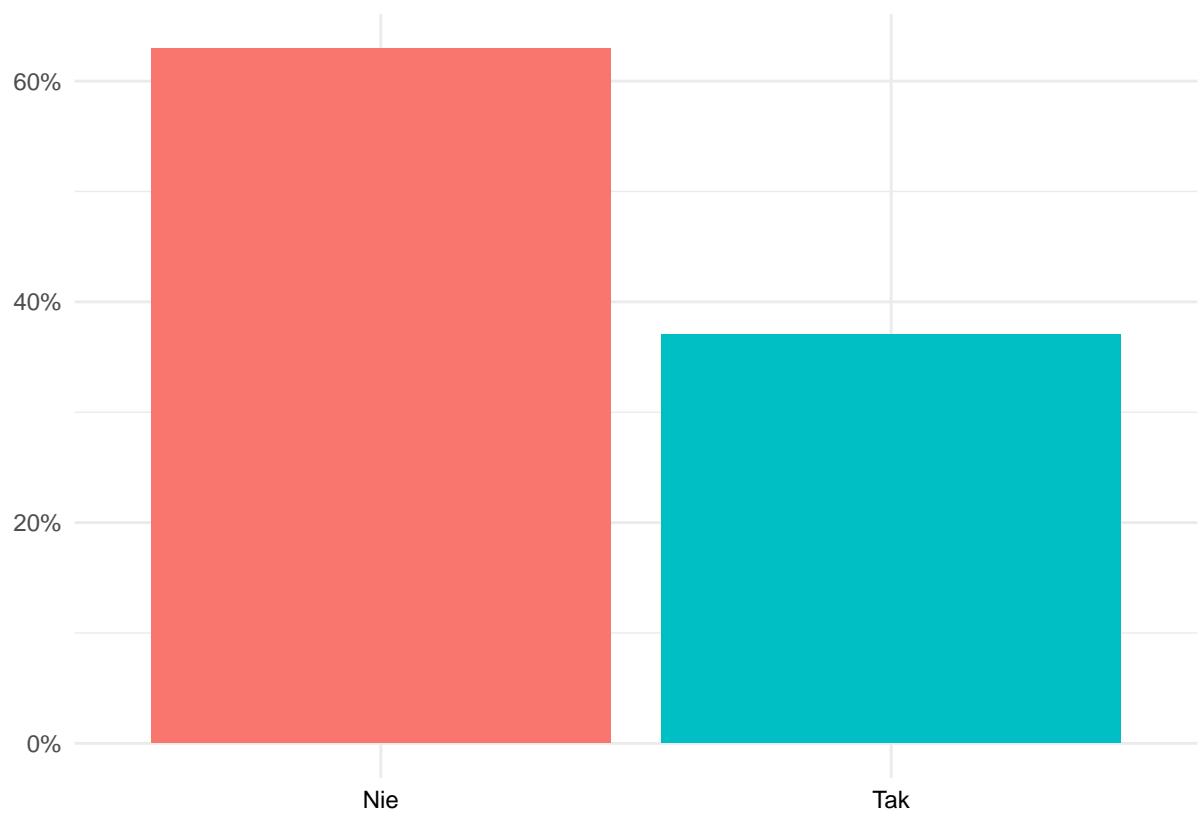


Wykres 15: Wykres słupkowy zmiennej *DeviceProtection*

Na wykresie 15. przedstawiono wykres słupkowy dla zmiennej *DeviceProtection* wśród klientów, którzy korzystają z usług internetowych. Wynika z niego, że podobnie jak w przypadku zmiennej *OnlineSecurity*, większość klientów nie korzysta z tej usługi. Taki brak zainteresowania ochroną urządzeń może prowadzić do problemów związanych z bezpieczeństwem, co w przeszłości może przyczynić się do utraty klientów.

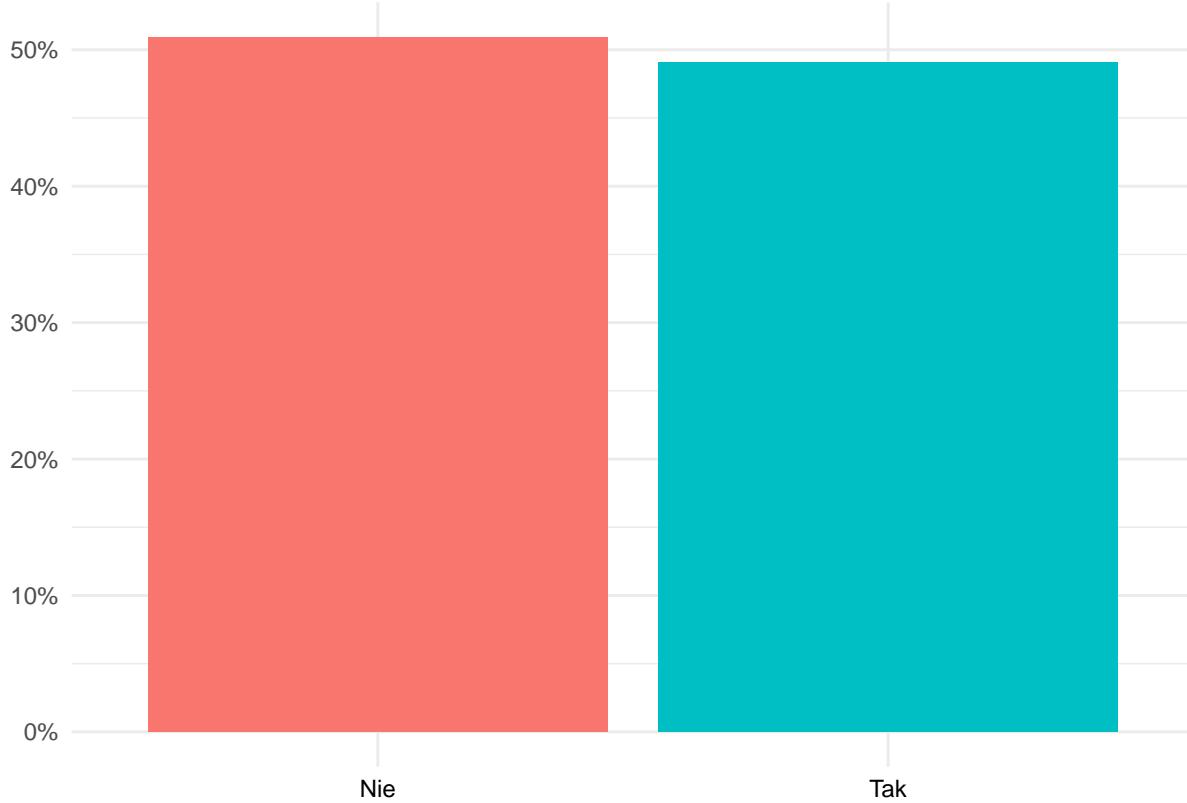
```
j <- j+1  
wykresy[[j]]
```

Na wykresie 16. przedstawiono wykres słupkowy dla zmiennej *TechSupport* wśród klientów, którzy korzystają z usług internetowych. Wynika z niego, że większość klientów nie korzysta z tej usługi. Może to sugerować, że klienci nie czują potrzeby wsparcia technicznego lub są w stanie radzić sobie z problemami samodzielnie. Niemniej jednak, brak korzystania z tej usługi może również wskazywać na potencjalne ryzyko w przyszłości, gdyby klienci napotkali trudności, które wymagałyby pomocy technicznej, co mogłoby prowadzić do ich niezadowolenia lub odejścia od usługi.



Wykres 16: Wykres słupkowy zmiennej TechSupport

```
j <- j+1  
wykresy[[j]]
```



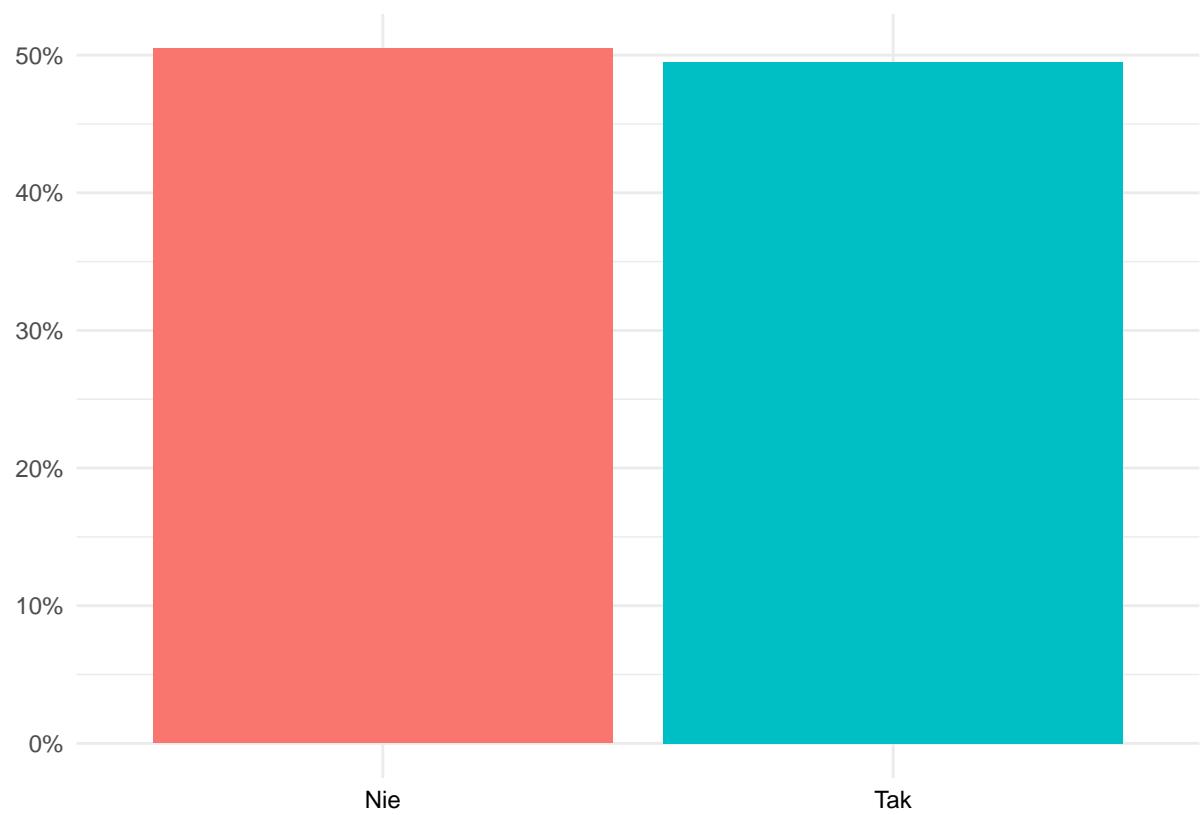
Wykres 17: Wykres słupkowy zmiennej StreamingTV

```
j <- j+1  
wykresy[[j]]
```

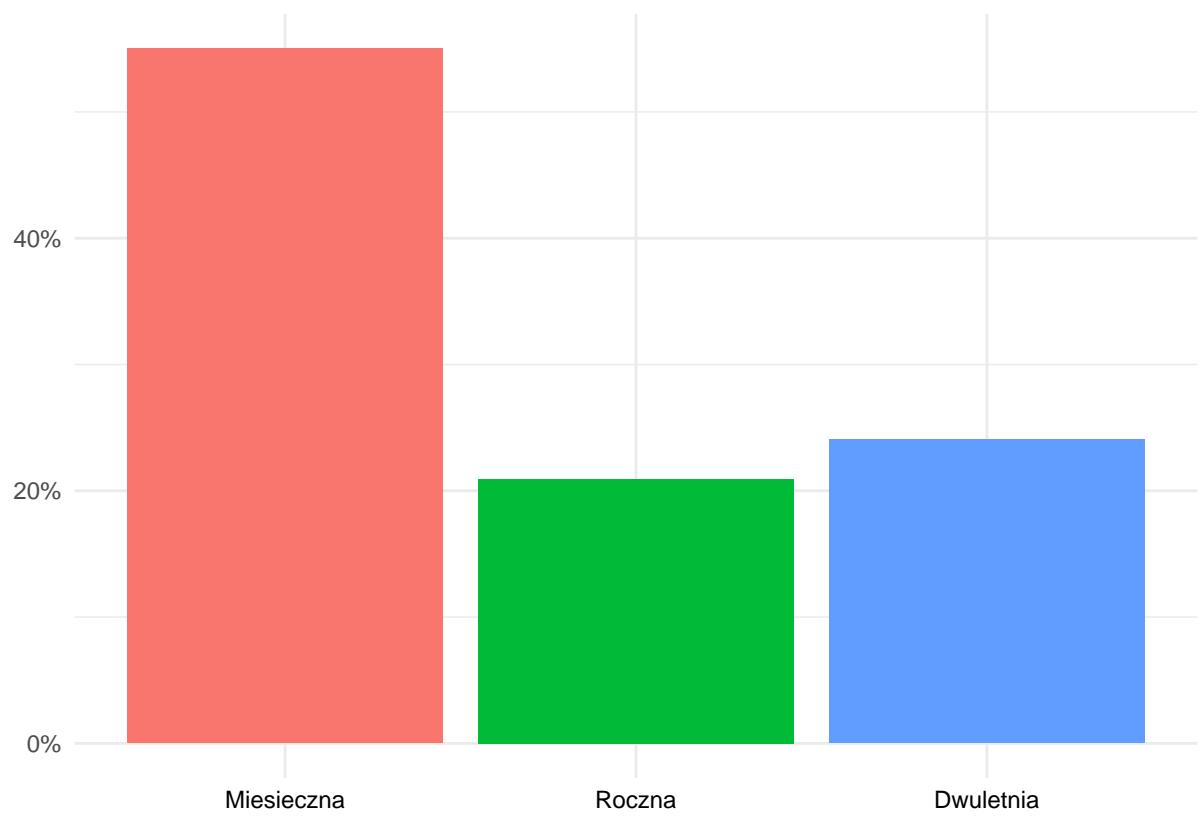
Na wykresach 17. i 18. przedstawiono wykresy słupkowe dla zmiennych *StreamingTV* oraz *StreamingMovies*. Wynika z nich, że równa liczba klientów korzysta z tych udogodnień. Jednakże, jak wskazuje paradoks Simpsona, może to się zmienić w przypadku dogłębniejszej analizy, szczególnie gdy uwzględnimy zmienną *churn*. Może się okazać, że w różnych grupach klientów (np. o różnym statusie “churn”) korzystanie z usług streamingowych ma różne zależności, co może wpływać na rzeczywisty obraz zachowań użytkowników w kontekście rezygnacji z usług.

```
j <- j+1  
wykresy[[j]]
```

Na wykresie 19. przedstawiono wykres słupkowy dla zmiennej *Contract*. Wynika z niego, że większość klientów preferuje kontrakt miesięczny. Opcje roczne i dwuletnie stanowią około 20% każda, przy czym więcej osób wybiera kontrakt dwuletni niż roczny. Taki rozkład może sugerować, że klienci cenią sobie elastyczność miesięcznych umów, co może wynikać z chęci przetestowania usług firmy lub skorzystania z krótkoterminowych korzyści. Jednocześnie część klientów decyduje się na długoterminowe zobowiązania, oczekując w zamian lepszych warunków, takich jak niższe opłaty miesięczne czy dodatkowe korzyści.

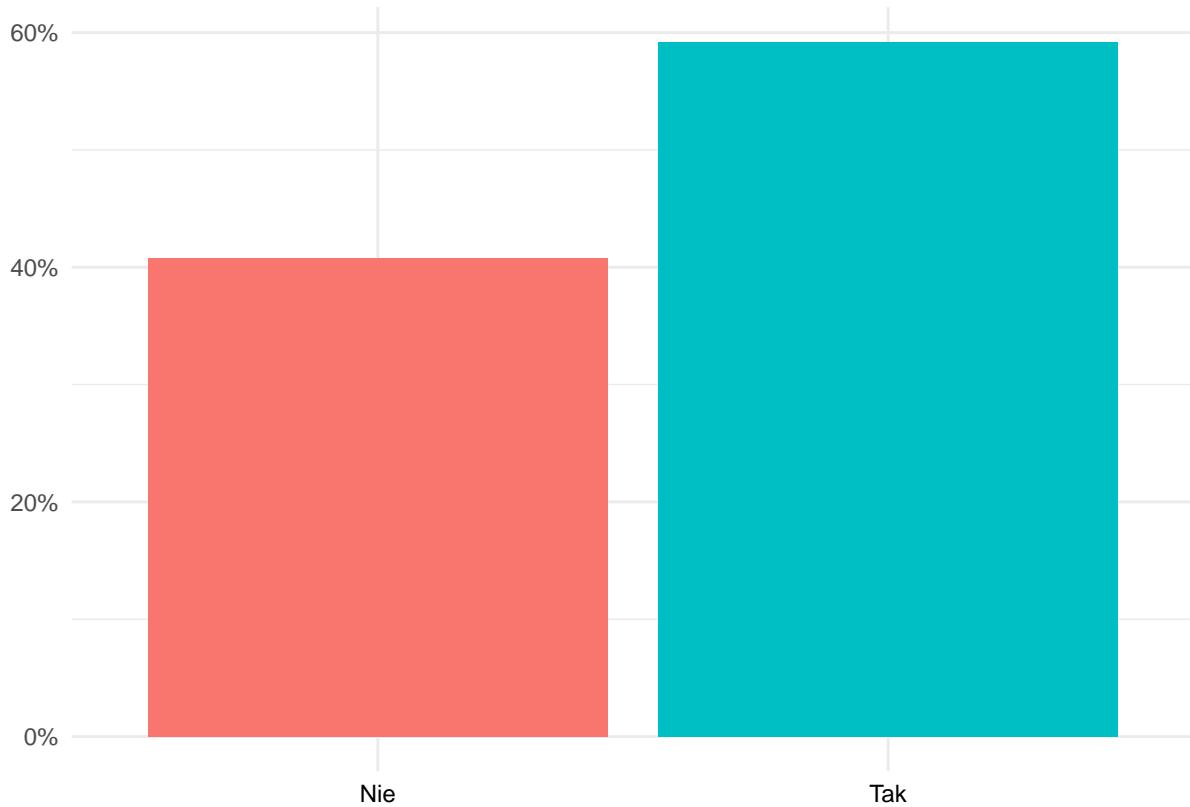


Wykres 18: Wykres słupkowy zmiennej StreamingMovies



Wykres 19: Wykres słupkowy zmiennej Contract

```
j <- j+1  
wykresy[[j]]
```



Wykres 20: Wykres słupkowy zmiennej *PaperlessBilling*

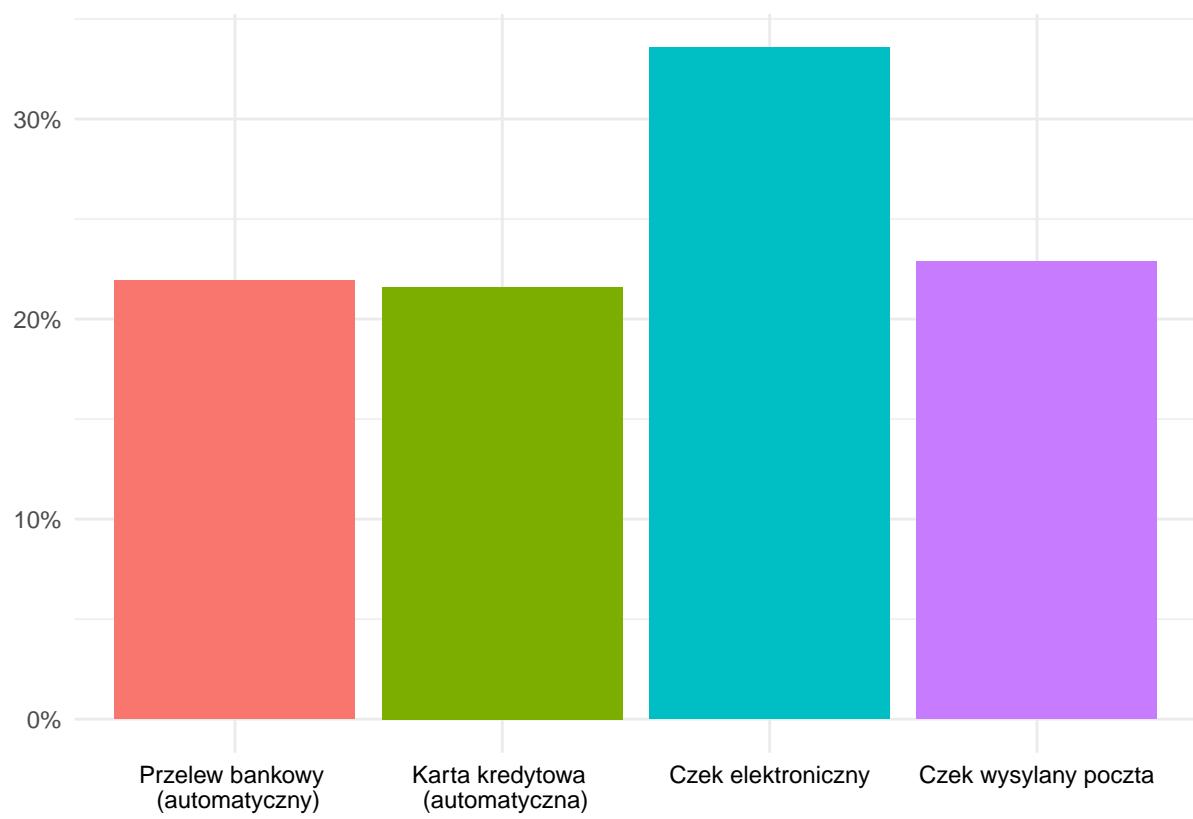
Na wykresie 20. przedstawiono wykres słupkowy dla zmiennej *PaperlessBilling*. Wynika z niego, że większość klientów preferuje rozwiązanie bez papierowych faktur.

```
j <- j+1  
wykresy[[j]]
```

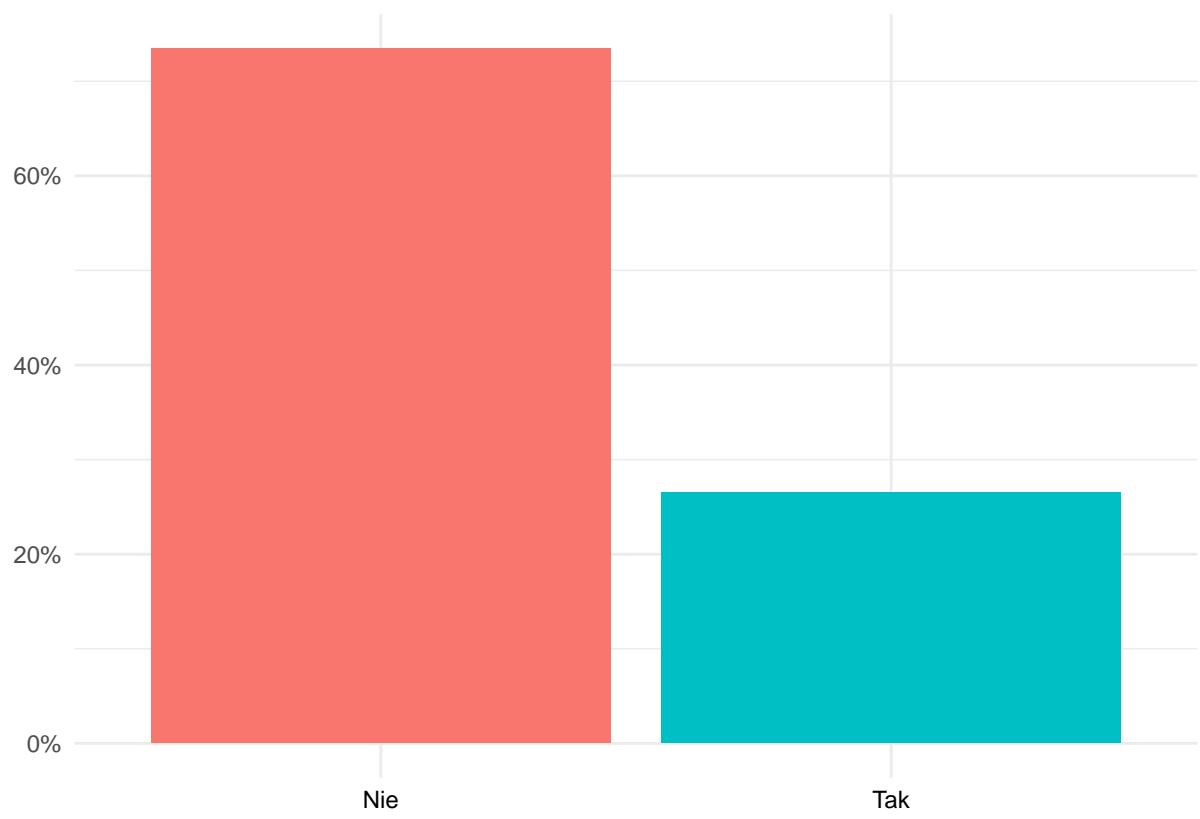
Na wykresie 21. przedstawiono wykres słupkowy dla zmiennej *PaymentMethod*. Widać, że większość osób preferuje płatności za pomocą czeku elektronicznego. Pozostałe metody płatności cieszą się podobną popularnością, co może sugerować, że klienci nie mają wyraźnej preferencji wobec innych opcji płatności, ale cenią sobie różnorodność dostępnych metod.

```
j <- j+1  
wykresy[[j]]
```

Na wykresie 22. przedstawiono wykres słupkowy dla zmiennej *Churn*. Widać, że większość klientów pozostaje lojalna i nie zrezygnowała z umowy, jednak około 25% zdecydowało się na jej rozwiązanie.



Wykres 21: Wykres słupkowy zmiennej PaymentMethod



Wykres 22: Wykres słupkowy zmiennej Churn

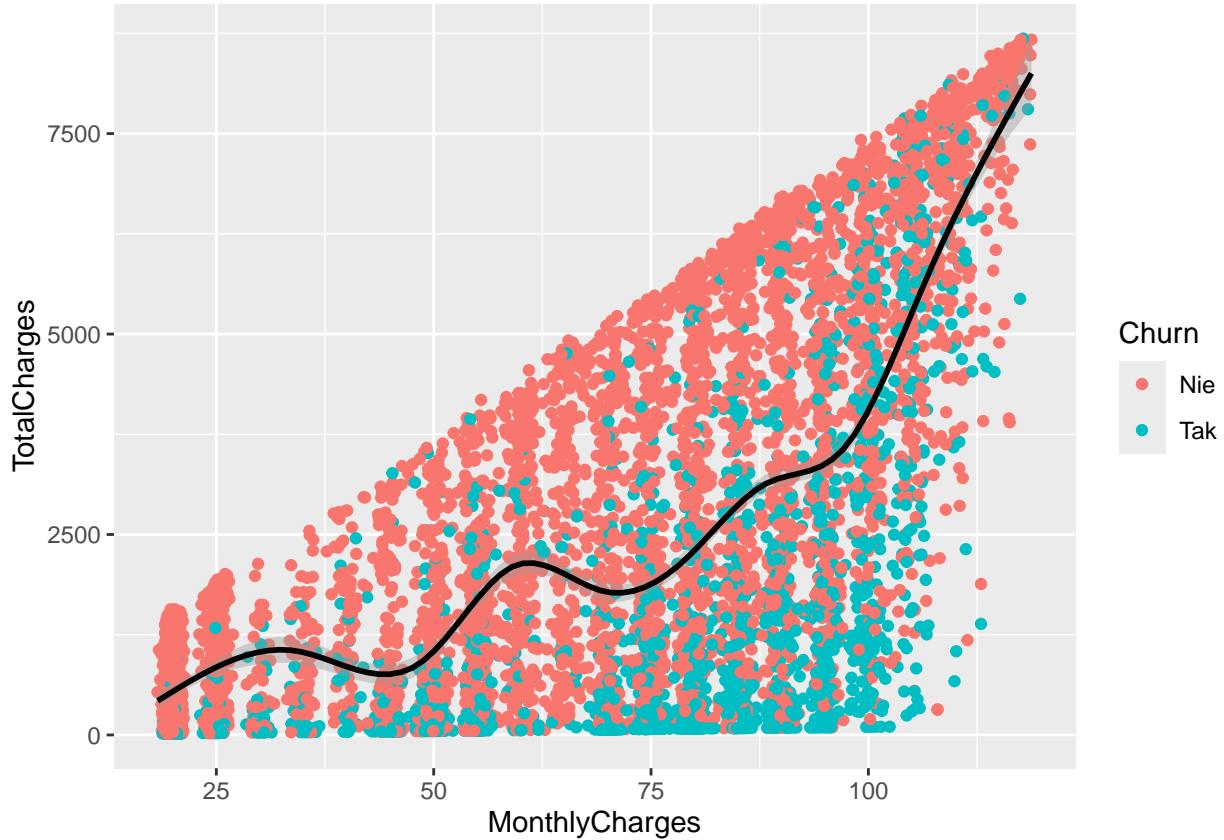
5.4 Wykresy rozrzutów

```
ggplot(klienci, aes(x=MonthlyCharges,
                     y=TotalCharges,
                     color=Churn)) +
  geom_point() +
  stat_smooth(color="black")

## `geom_smooth()` using method = 'gam' and formula = 'y ~ s(x, bs = "cs")'

## Warning: Removed 11 rows containing non-finite outside the scale range
## ('stat_smooth()').

## Warning: Removed 11 rows containing missing values or values outside the scale range
## ('geom_point()').
```



Wykres 23: Zależność TotalCharges od MonthlyCharges z krzywą wygładzającą

Na wykresie 23. przedstawiono zależność *TotalCharges* od *MonthlyCharges* z krzywą wygładzającą, a dane zostały rozdzielone na klientów, którzy odeszli, oraz tych, którzy pozostały. Widać, że istnieje liniowa zależność między maksymalnymi wartościami *TotalCharges* a *MonthlyCharges*, która rośnie proporcjonalnie. Krzywa wygładzająca wskazuje, że dane rozkładają się głównie w średnim zakresie *TotalCharges*, jednak

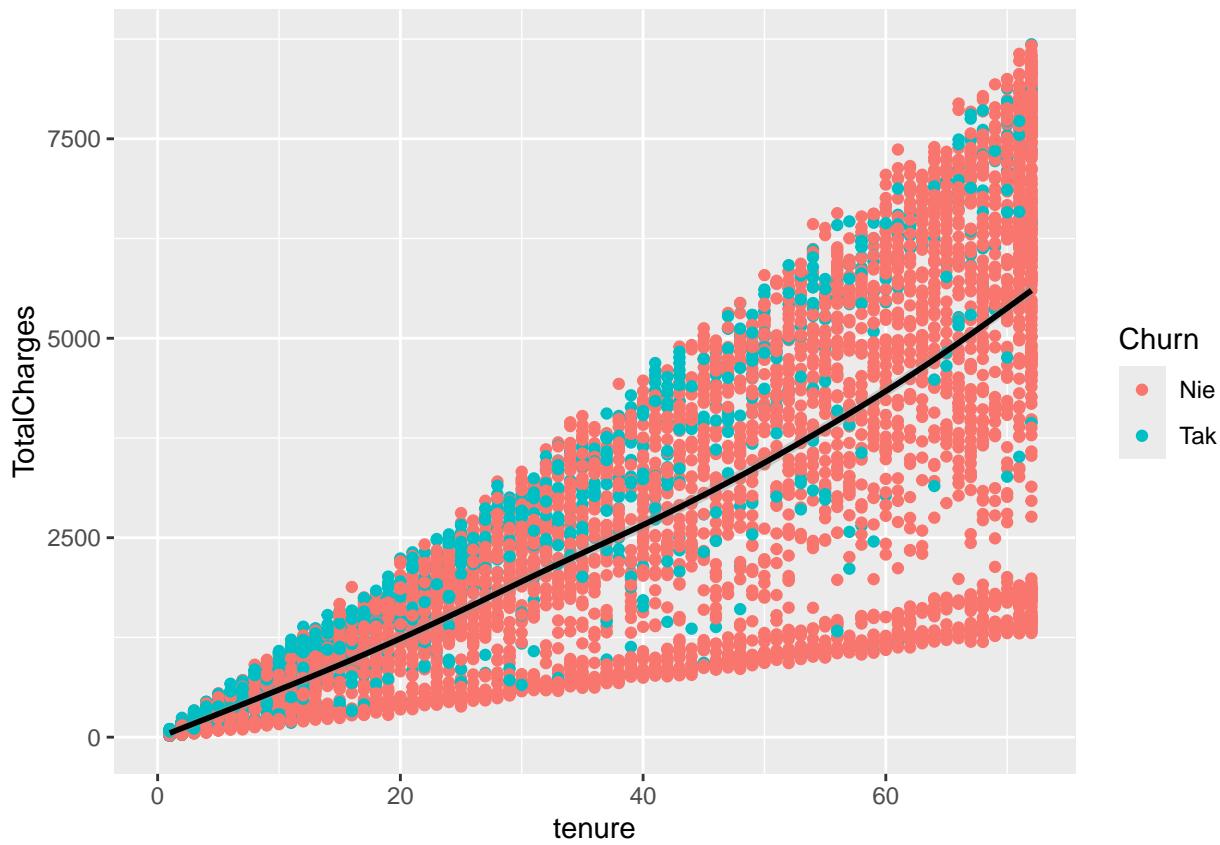
powyżej kwoty 100 dolarów miesięcznie zaczyna się ona zwiększać w sposób liniowy, osiągając wartość maksymalną. Co istotne, nie widać klientów, którzy płacą wysokie miesięczne opłaty, a jednocześnie mają niskie *TotalCharges*. Może to sugerować, że tacy klienci po prostu nie istnieją w zbiorze danych.

```
ggplot(klienci, aes(x=tenure,
                     y=TotalCharges,
                     color=Churn)) +
  geom_point() +
  stat_smooth(color="black")

## `geom_smooth()` using method = 'gam' and formula = 'y ~ s(x, bs = "cs")'

## Warning: Removed 11 rows containing non-finite outside the scale range
## ('stat_smooth()').

## Warning: Removed 11 rows containing missing values or values outside the scale range
## ('geom_point()').
```

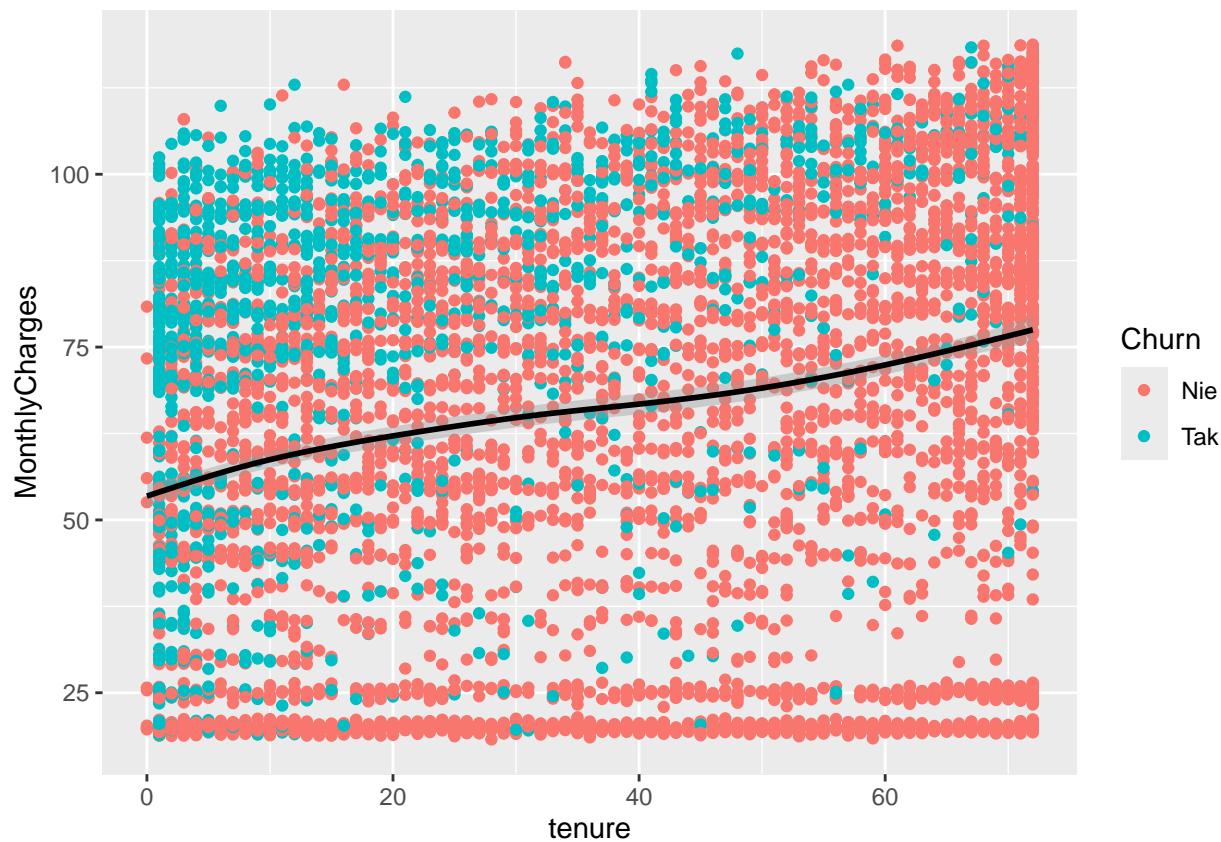


Wykres 24: Zależność *TotalCharges* od *tenure* z krzywą wygładzającą

Na wykresie 24. przedstawiono zależność *TotalCharges* od *tenure* z krzywą wygładzającą, a dane zostały rozdzielone na klientów, którzy odeszli, oraz tych, którzy pozostały. Widać liniową zależność między minimalnymi, maksymalnymi i średnimi wartościami *TotalCharges* a *tenure*, co sugeruje, że wraz z długością trwania umowy klienci generują większe całkowite opłaty. Widać również brak klientów o długiej lojalności,

którzy jednocześnie mają lekko ponad minimalne wydatki całkowite. Może to wskazywać, że lojalni klienci albo korzystają z minimalnych usług, albo wybierają droższe opcje oferowane przez firmę, co może świadczyć o ich skłonności do długoterminowego zaangażowania w usługi, które są bardziej kosztowne, ale dostosowane do ich potrzeb.

```
ggplot(klienci, aes(x=tenure,
                     y=MonthlyCharges,
                     color=Churn)) +
  geom_point() +
  stat_smooth(color="black")  
  
## `geom_smooth()` using method = 'gam' and formula = 'y ~ s(x, bs = "cs")'
```



Wykres 25: Zależność *MonthlyCharges* od *tenure* z krzywą wygładzającą

Na wykresie 25. przedstawiono zależność między wartością *MonthlyCharges* a *tenure*, z nałożoną krzywą wygładzającą, a dane zostały podzielone na klientów, którzy zrezygnowali z usług, oraz tych, którzy pozostały. Widać pełen przekrój klientów, korzystających z naszych usług, obejmujący różne kwoty miesięcznych opłat oraz różnych czas lojalności. Zauważalna jest dodatnia zależność liniowa między średnią a maksymalną wartością *MonthlyCharges*. Klienci korzystający z minimalnych usług są obecni w każdym segmencie. Podobnie jak na wykresie 23., widać rozrzedzenie liczby klientów z długim stażem oraz wydatkami na usługi powyżej minimum. Sugeruje to, że klienci, którzy pozostały z nami na dłużej, albo korzystają z podstawowych usług, albo zdecydowali się na zakup dodatkowych usług.

```

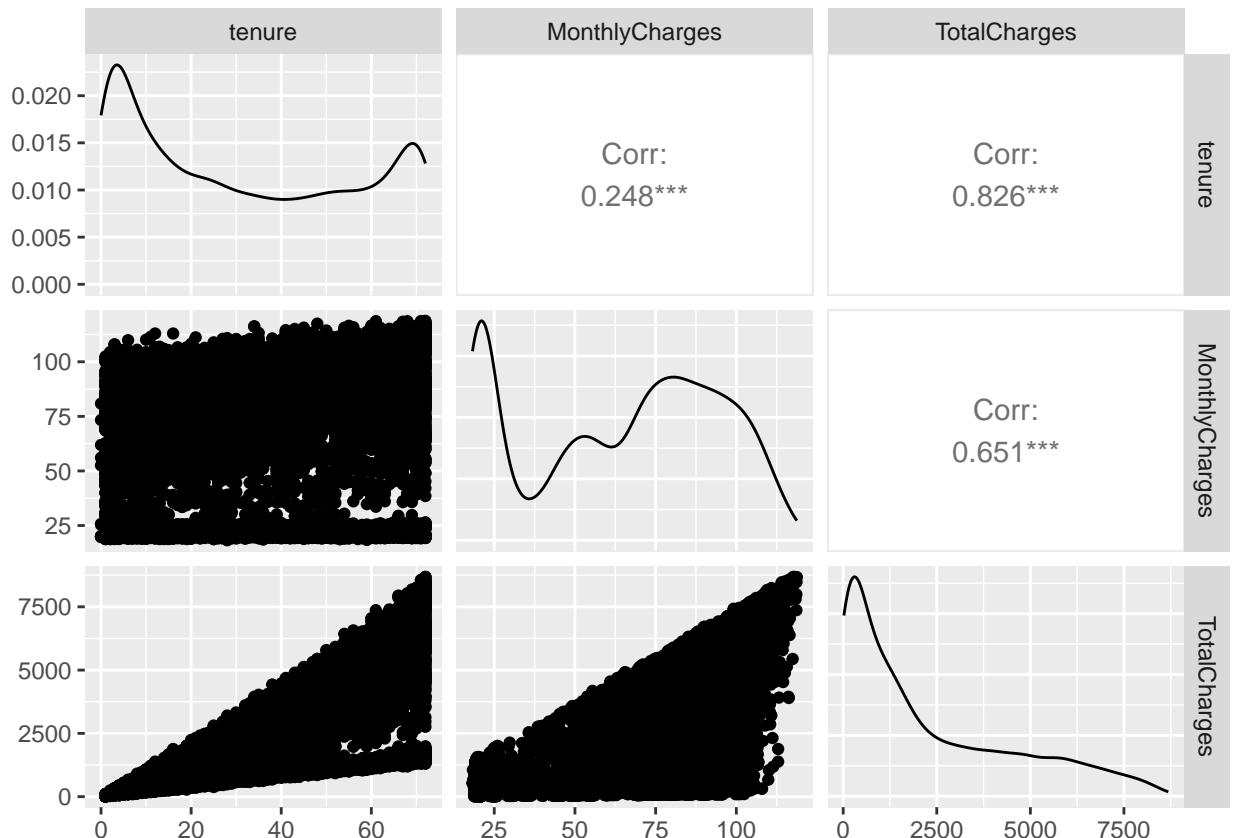
ggpairs(num_cols)

## Warning in ggally_statistic(data = data, mapping = mapping, na.rm = na.rm, :
## Removed 11 rows containing missing values
## Warning in ggally_statistic(data = data, mapping = mapping, na.rm = na.rm, :
## Removed 11 rows containing missing values

## Warning: Removed 11 rows containing missing values or values outside the scale range
## ('geom_point()').
## Removed 11 rows containing missing values or values outside the scale range
## ('geom_point()').

## Warning: Removed 11 rows containing non-finite outside the scale range
## ('stat_density()').

```



Wykres 26: Macierz par zmiennych numerycznych

Na wykresie 26. zaprezentowano macierz par zmiennych numerycznych, w której zidentyfikowano silną korelację liniową (0,826) między zmiennymi *TotalCharges* a *tenure*.

6 Analiza opisowa z podziałem na grupy

```
obecni_klienci <- subset(klienci, subset=(Churn=="Nie"))
dawni_klienci <- subset(klienci, subset=(Churn=="Tak"))
```

6.1 Histogramy i wykresy pudełkowe

```
histogramy <- list()
pudełka <- list()

for (var in names(num_cols)) {
  # Wzór Scotta
  binwidth <- 3.5 * sd(klienci[[var]], na.rm = TRUE) * length(klienci[[var]])^(-1/3)

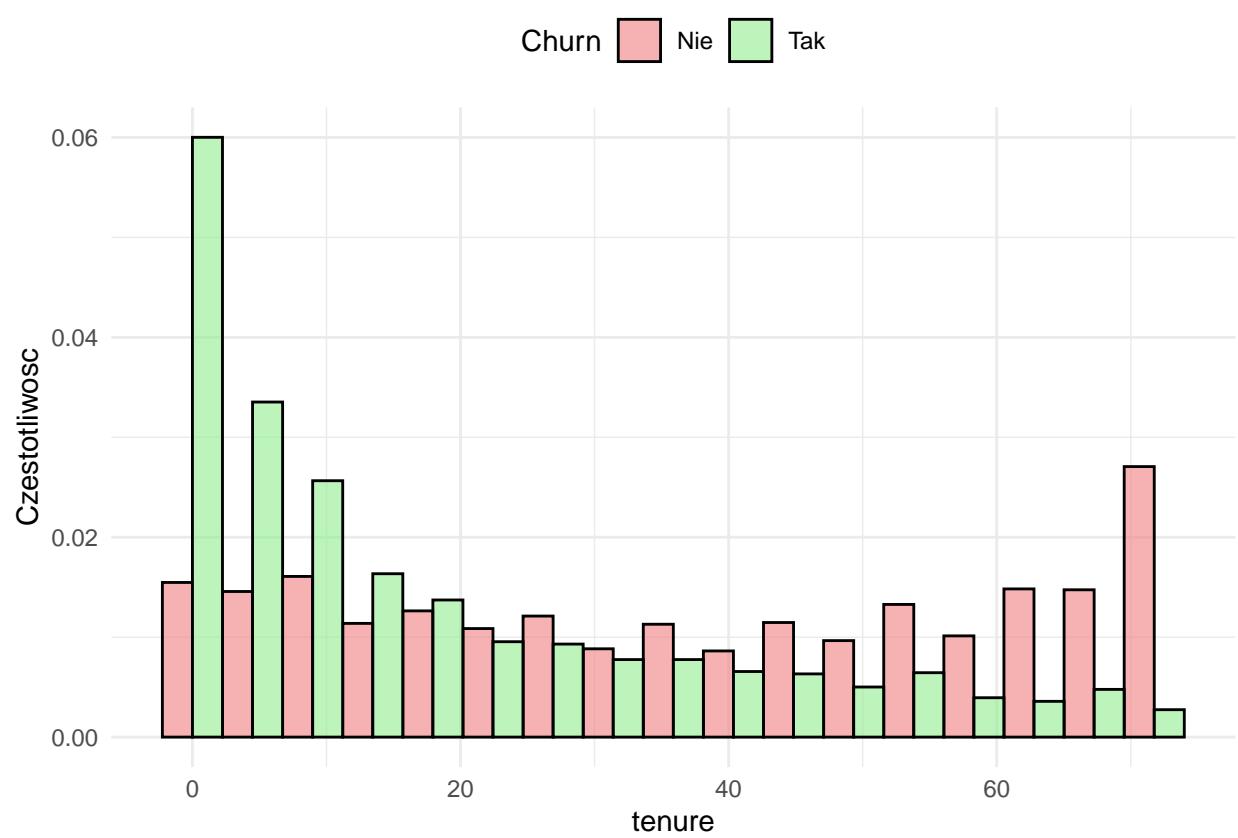
  histogramy <- append(histogramy,
    list(ggplot(klienci, aes(x = .data[[var]],
      fill = Churn)) +
      geom_histogram(position = "dodge",
        binwidth = binwidth,
        color = "black",
        alpha = 0.6,
        aes(y = ..density..)) +
      labs(x = var, y = "Częstotliwość") +
      scale_fill_manual(values = c("Tak" = "lightgreen",
        "Nie" = "lightcoral")) +
      theme_minimal() +
      theme(legend.position = "top")
    ))
}

pudełka <- append(pudełka,
  list(ggplot(klienci, aes(x = Churn, y = .data[[var]],
    fill = Churn)) +
    geom_boxplot() +
    labs(x = "Churn", y = var) +
    scale_fill_manual(values =
      c("Tak" = "lightgreen", "Nie" = "lightcoral")) +
    theme_minimal() +
    theme(legend.position="none")
  ))
}

histogramy[[1]]
```

Warning: The dot-dot notation ('..density..') was deprecated in ggplot2 3.4.0.
i Please use 'after_stat(density)' instead.
This warning is displayed once every 8 hours.
Call 'lifecycle::last_lifecycle_warnings()' to see where this warning was
generated.

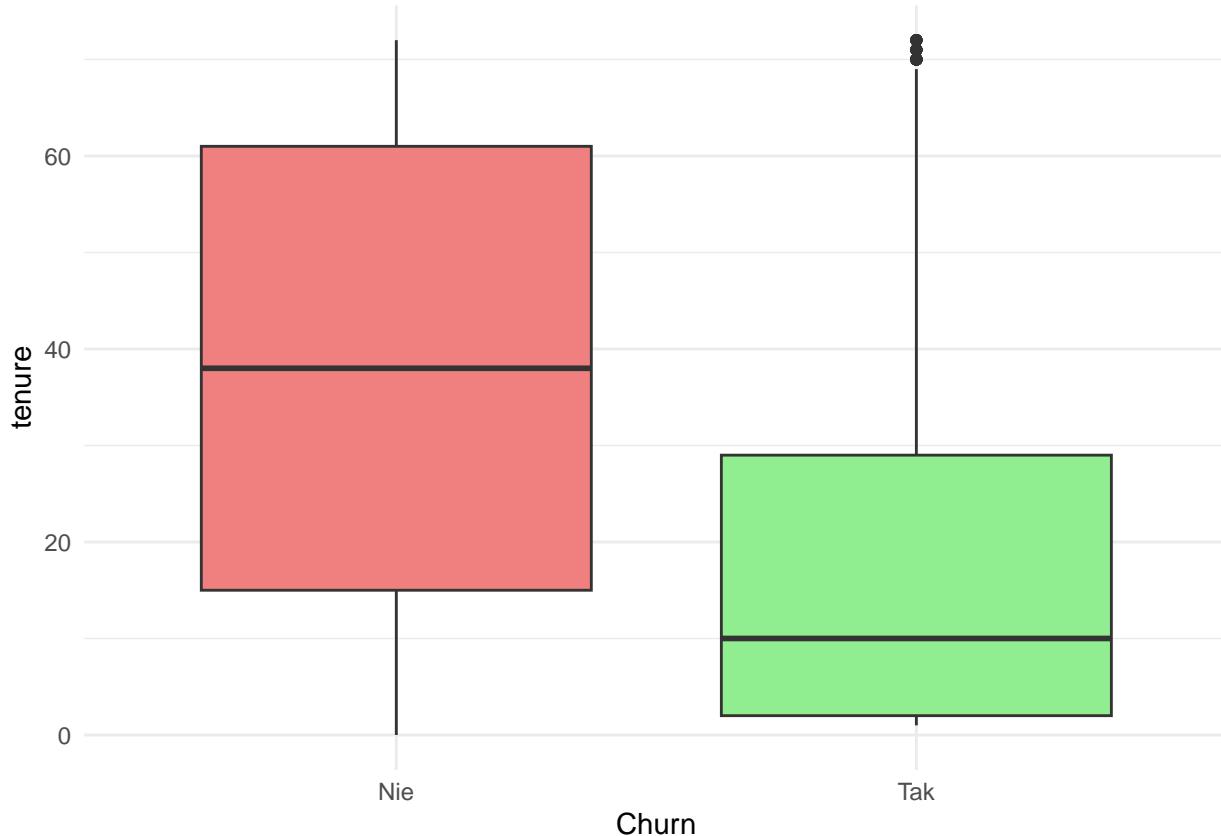
Na wykresie 27. przedstawiono histogram dla zmiennej *tenure*, z podziałem na klientów, którzy odeszli (churn) i tych, którzy pozostali. Podobnie jak w przypadku ogólnego rozkładu, widać równomierny rozkład



Wykres 27: Histogram dla zmiennej tenure z podziałem ze względu na churn

dla osób, które zostały, z wyraźnym szczytem dla klientów o dłuższym stażu. Wypłaszczenie się górką dla niższych wartości jest wynikiem uwzględnienia klientów, którzy odeszli. Ich rozkład przypomina rozkład wykładowiczy, co może sugerować, że ci klienci korzystali z naszych usług przez pierwsze miesiące, jednak nie zostali przekonani, co prowadziło do ich stosunkowo szybkiego odejścia.

pudełka [[1]]

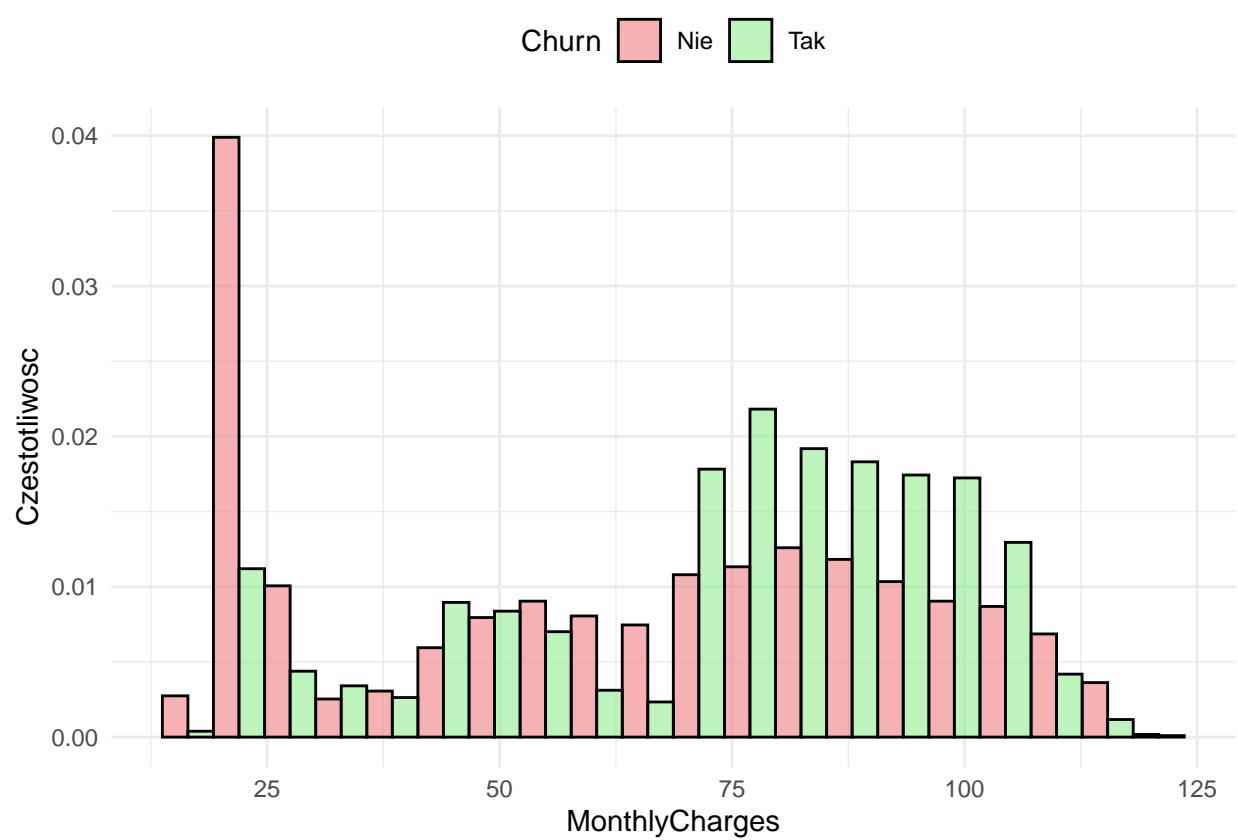


Wykres 28: Wykres pudełkowy dla zmiennej tenure z podziałem ze względu na churn

Na wykresie 28. zaprezentowano wykres pudełkowy dla zmiennej *tenure*, z podziałem na klientów, którzy odeszli (churn) i tych, którzy pozostali. Widać, że klienci, którzy zrezygnowali z usług, korzystali z nich przez krótki okres, podczas gdy ci, którzy pozostali, są lojalni przez dłuższy czas. Zauważalne są również odstające dane w grupie klientów, którzy odeszli, co może wynikać z losowego zakończenia współpracy.

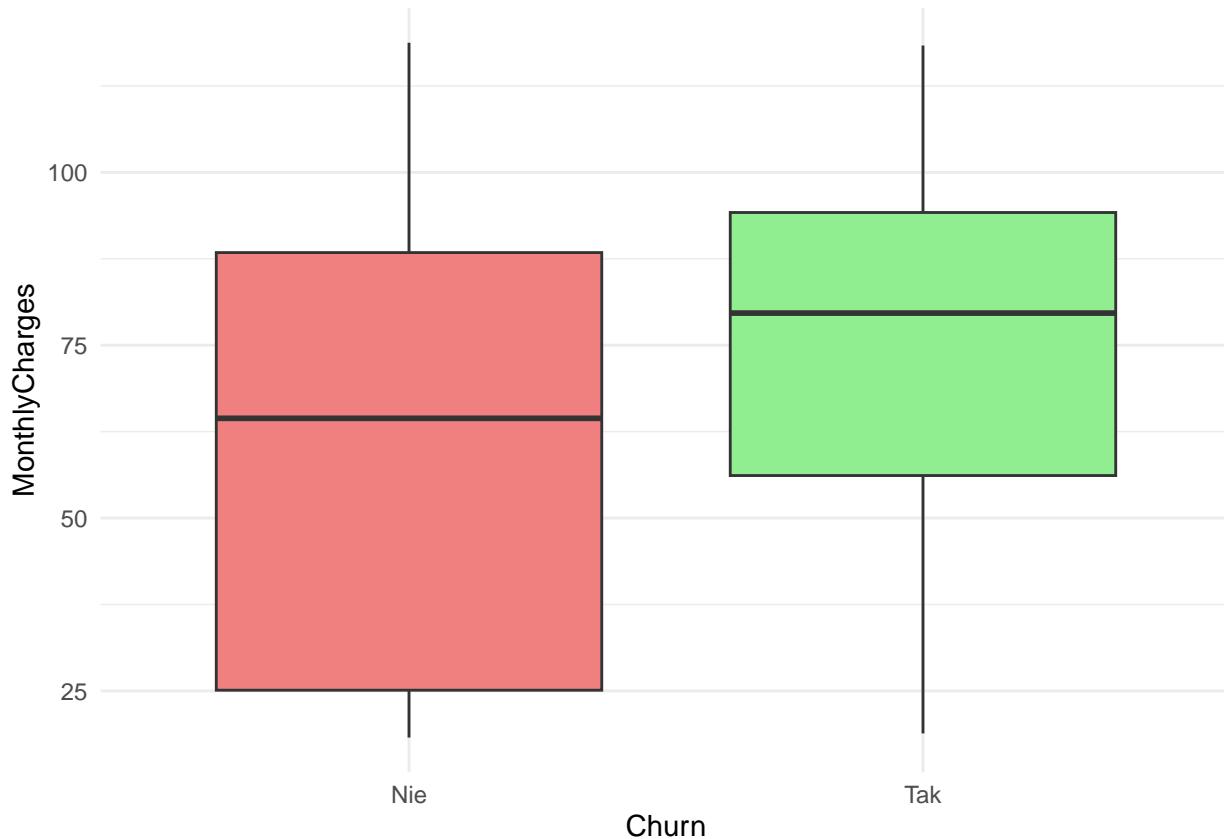
histogramy [[2]]

Na wykresie 29. przedstawiono histogram dla zmiennej *MonthlyCharges*, z podziałem na klientów, którzy odeszli (churn) i tych, którzy pozostali. Widać, że klienci dzielą się na trzy grupy: tych, którzy płacą minimalnie lub nieco ponad, tych, którzy płacą około 50 dolarów, oraz tych, którzy płacą powyżej 60 dolarów. Zauważalny jest również pik na niewielkich kwotach w przypadku klientów, którzy pozostali, co sugeruje, że mamy wielu stałych klientów korzystających z podstawowych usług. Natomiast dla kwot powyżej 60 dolarów widać, że liczba klientów, którzy odeszli, jest większa niż tych, którzy pozostali. Może to sugerować, że droższe usługi oferowane tym klientom nie przypadły im do gustu, lub że potrzebowali ich tylko przez określony czas.



Wykres 29: Histogram dla zmiennej MonthlyCharges z podziałem ze względu na churn

pudełka [[2]]



Wykres 30: Wykres pudełkowy dla zmiennej *MonthlyCharges* z podziałem ze względu na churn

Na wykresie 30. zaprezentowano wykres pudełkowy dla zmiennej *MonthlyCharges*, z podziałem na klientów, którzy odeszli (churn) i tych, którzy pozostali. Widać, że średnia wartość miesięcznych opłat dla klientów, którzy pozostali, mieści się w przedziale od 25 do 88 dolarów. Natomiast dla klientów, którzy odeszli, środkowa część rozkładu pokazuje wyższe kwoty opłat. Sugeruje to, że klienci, którzy pozostali, płacili średnio mniej, podczas gdy ci, którzy odeszli, zazwyczaj korzystali z droższych usług.

#histogramy [[3]]

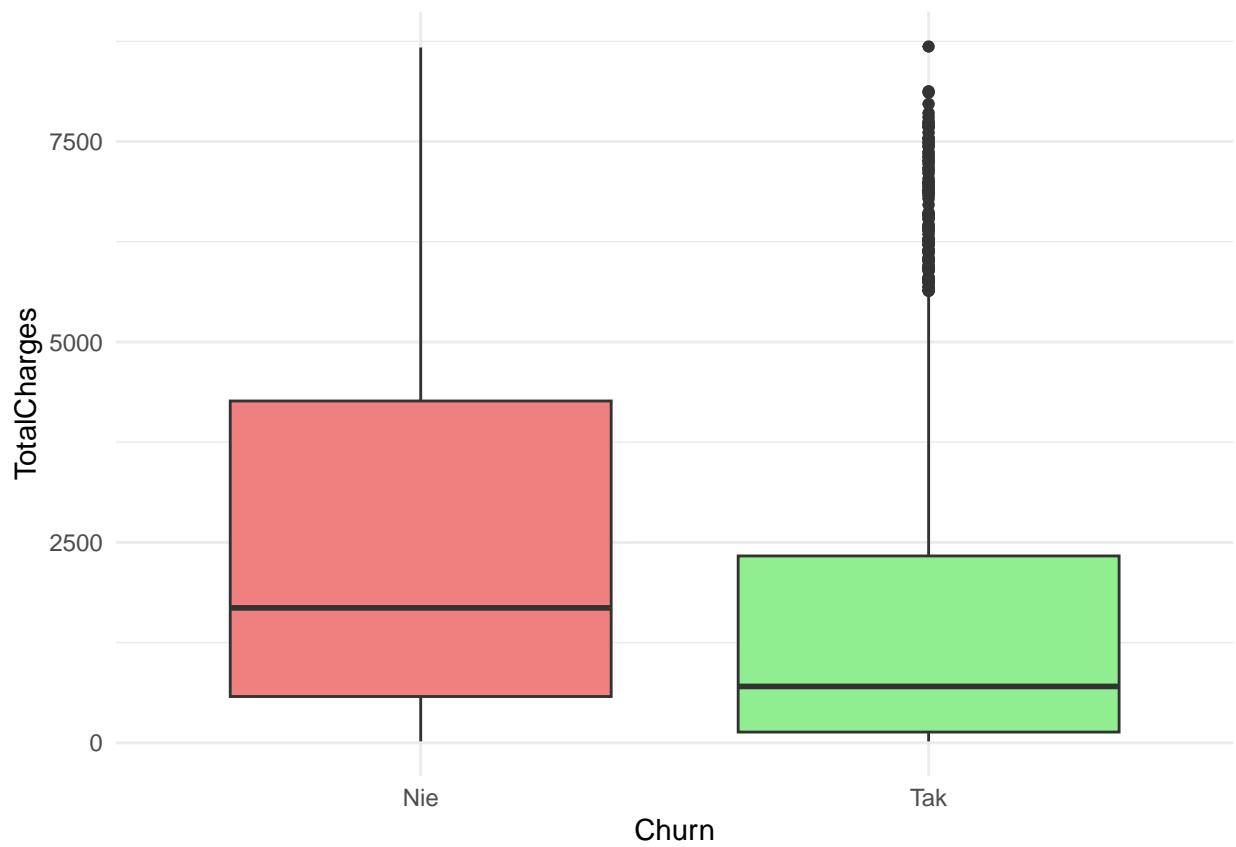
pudełka [[3]]

```
## Warning: Removed 11 rows containing non-finite outside the scale range
## ('stat_boxplot()').
```

Na wykresie 31. zaprezentowano wykres pudełkowy dla zmiennej *TotalCharges*, z podziałem na klientów, którzy odeszli (churn) i tych, którzy pozostali. Widać, że część środkowa klientów, którzy odeszli, zostawiła u nas mniej pieniędzy niż ci, którzy pozostali.

6.2 Wykresy słupkowe zmiennych jakościowych

Ponownie za pomocą for tworzymy wykresy słupkowe zmiennych jakościowych.



Wykres 31: Wykres pudełkowy dla zmiennej TotalCharges z podziałem ze względu na churn

```

wykresy <- list()
tytuły <- c()

for(i in colnames(klienci)){
  if(is.factor(klienci[[i]])){

    obecni_klienci_zmienna <- obecni_klienci[[i]]
    dawni_klienci_zmienna <- dawni_klienci[[i]]

    obecni_klienci_zmienna <- obecni_klienci_zmienna[!obecni_klienci_zmienna %in%
      c("Brak internetu",
        "Brak usługi telefonicznej")]
    dawni_klienci_zmienna <- dawni_klienci_zmienna[!dawni_klienci_zmienna %in%
      c("Brak internetu",
        "Brak usługi telefonicznej")]

    a1 <- table(obecni_klienci_zmienna)
    a2 <- table(dawni_klienci_zmienna)

    a1 <- a1 / sum(a1)
    a2 <- a2 / sum(a2)

    b <- max(a1, a2)

    wykresy <- append(wykresy,
      list(ggplot(data.frame(obecni_klienci_zmienna),
        aes(x = obecni_klienci_zmienna,
          y = after_stat(count) / sum(after_stat(count)),
          label = after_stat(count))) +
        geom_bar(fill = "deepskyblue2") +
        scale_y_continuous(limits = c(0, b),
          labels = scales::percent) +
        labs(x = "", y = "", title = "Obecni klienci") +
        theme(legend.position = "none") + coord_flip() +
        geom_text(stat = "count", aes(label = after_stat(count)),
          size = 5, hjust = 1.1),

        ggplot(data.frame(dawni_klienci_zmienna),
          aes(x = dawni_klienci_zmienna,
            y = after_stat(count) / sum(after_stat(count)))) +
        geom_bar(fill = "tomato") +
        scale_y_continuous(limits = c(0, b),
          labels = scales::percent) +
        labs(x = "", y = "", title = "Dawni klienci") +
        coord_flip() +
        geom_text(stat = "count", aes(label = after_stat(count)),
          size = 5, hjust = 1.1)
      ))
  }
}

t <- paste("Wykres słupkowy zmiennej", i, "z podziałem ze względu na churn")
tytuły <- c(tytuły, t)
}

```

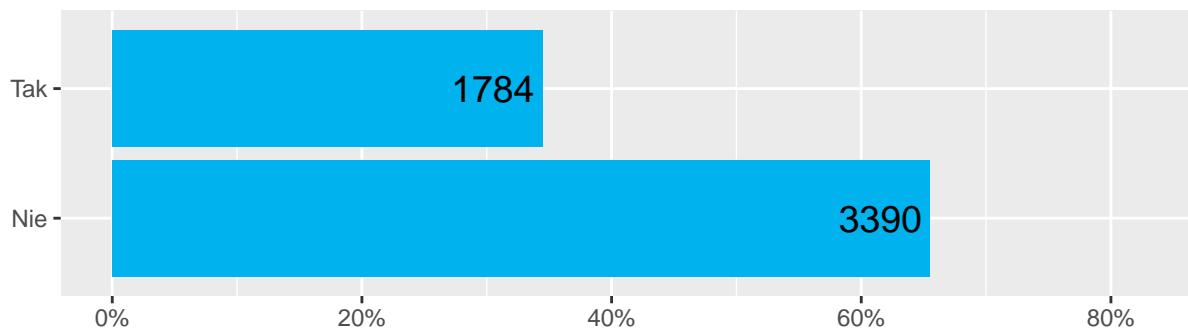
```
#grid.arrange(wykresy[[l]], wykresy[[l+1]], nrow=2)
l <- l+2
k <- k+1
```

```
#grid.arrange(wykresy[[l]], wykresy[[l+1]], nrow=2)
l <- l+2
k <- k+1
```

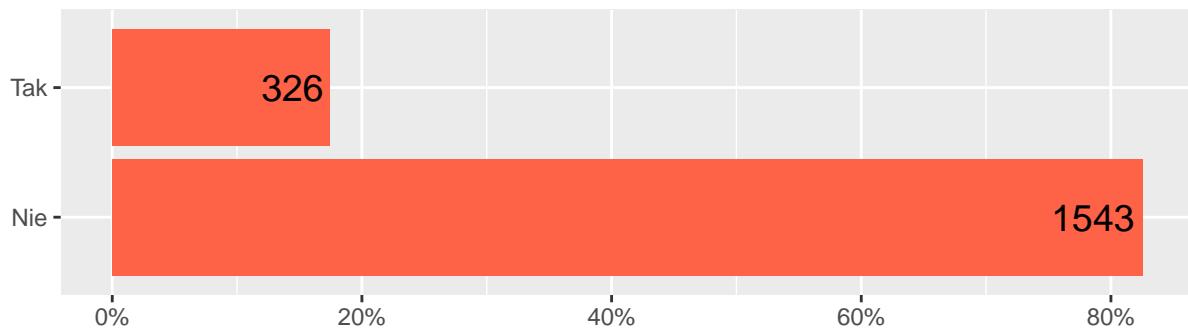
```
#grid.arrange(wykresy[[l]], wykresy[[l+1]], nrow=2)
l <- l+2
k <- k+1
```

```
grid.arrange(wykresy[[1]], wykresy[[1+1]], nrow=2)
```

Obecni klienci



Dawni klienci



Wykres 32: Wykres słupkowy zmiennej Dependents z podziałem ze względu na churn

```
l <- l+2
k <- k+1
```

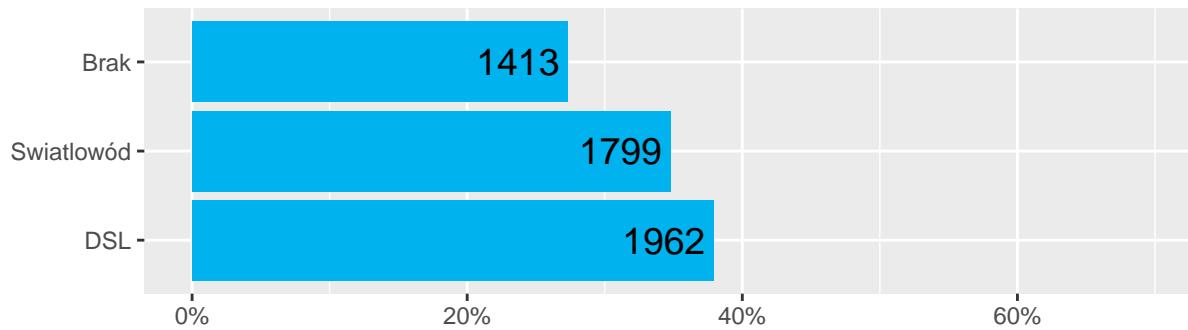
Na wykresie 32. przedstawiono wykres słupkowy dla zmiennej *dependants*, z podziałem na klientów, którzy odeszli (churn) i tych, którzy pozostały. Widać, że klienci, którzy odeszli, nie mają nikogo na utrzymaniu, co sugeruje, że osoby bez osób na utrzymaniu częściej decydują się na rezygnację z usług.

```
#grid.arrange(wykresy[[l]], wykresy[[l+1]], nrow=2)
l <- l+2
k <- k+1
```

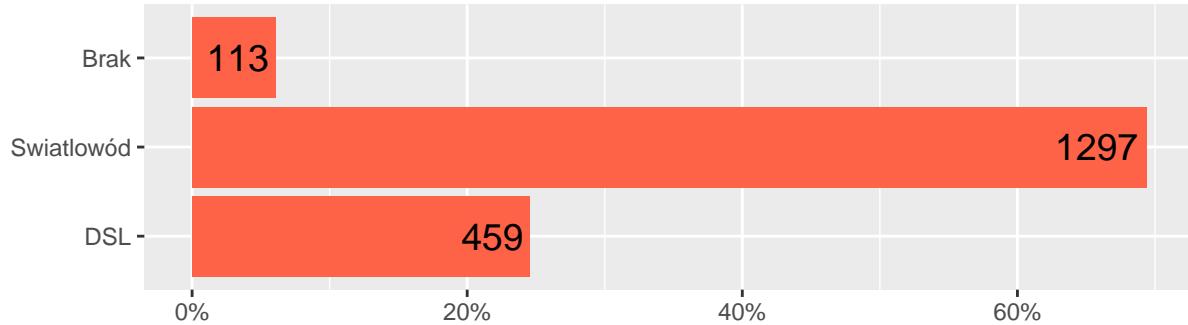
```
#grid.arrange(wykresy[[l]], wykresy[[l+1]], nrow=2)
l <- l+2
k <- k+1
```

```
grid.arrange(wykresy[[1]], wykresy[[1+1]], nrow=2)
```

Obecni klienci



Dawni klienci



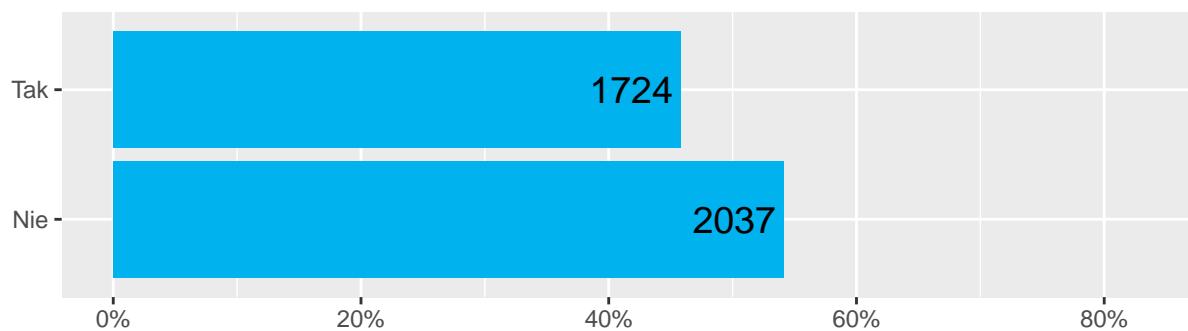
Wykres 33: Wykres słupkowy zmiennej InternetService z podziałem ze względu na churn

```
l <- l+2
k <- k+1
```

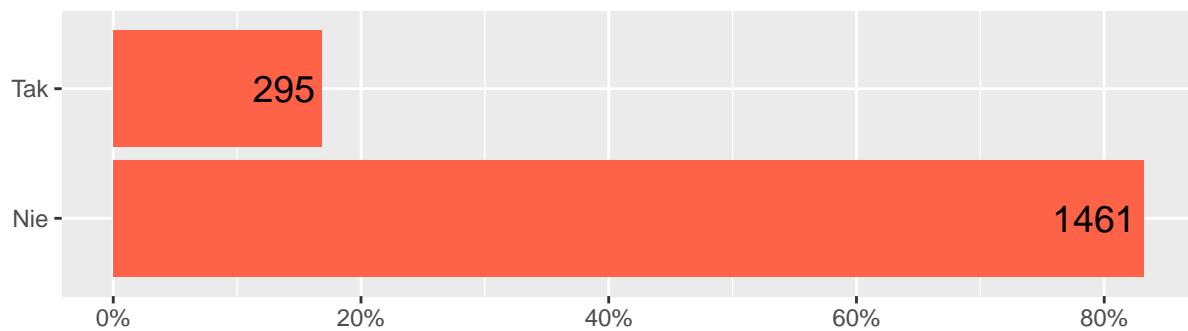
Na wykresie 33. przedstawiono wykres słupkowy dla zmiennej *dependants*, z podziałem na klientów, którzy odeszli (churn) i tych, którzy pozostali. Widać, że klienci, którzy odeszli, w większym stopniu korzystali z internetu i mieli dostęp do lepszych rozwiązań światłowodowych w porównaniu do tych, którzy pozostali.

```
grid.arrange(wykresy[[1]], wykresy[[1+1]], nrow=2)
```

Obecni klienci



Dawni klienci

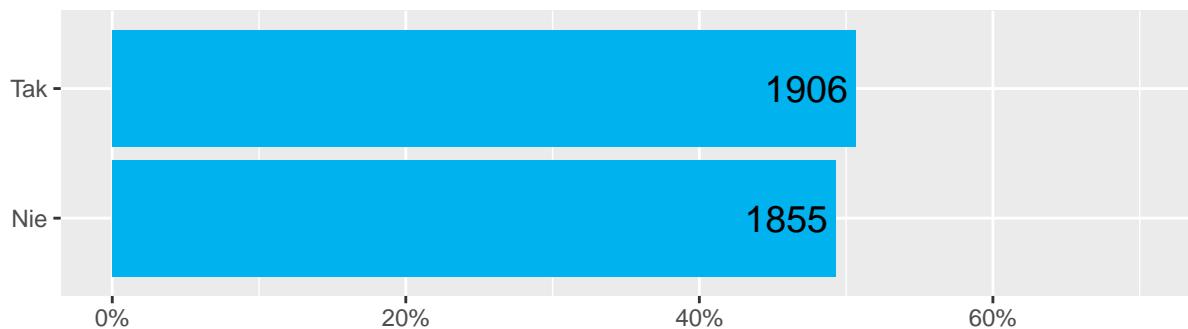


Wykres 34: Wykres słupkowy zmiennej OnlineSecurity z podziałem ze względu na churn

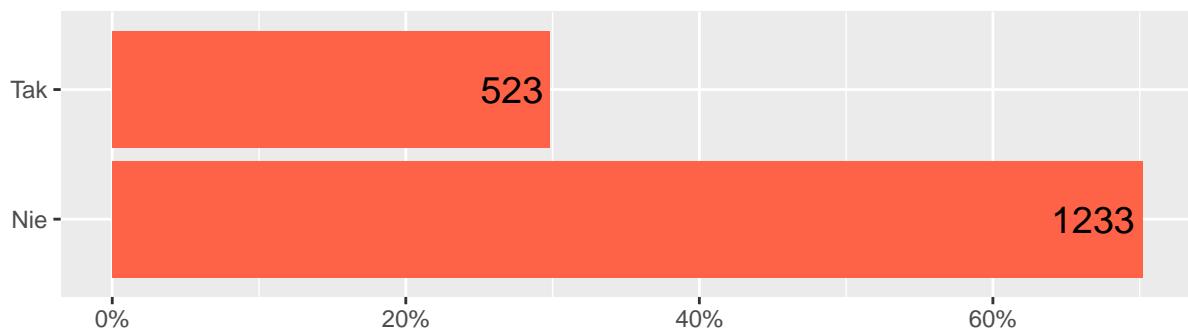
```
l <- l+2  
k <- k+1
```

```
grid.arrange(wykresy[[1]], wykresy[[1+1]], nrow=2)
```

Obecni klienci



Dawni klienci



Wykres 35: Wykres słupkowy zmiennej OnlineBackup z podziałem ze względu na churn

```
l <- l+2  
k <- k+1
```

```
grid.arrange(wykresy[[1]], wykresy[[1+1]], nrow=2)
```

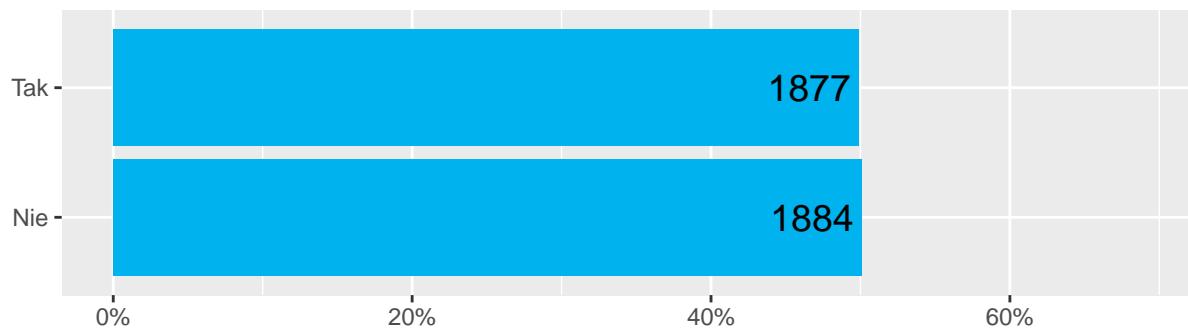
```
l <- l+2  
k <- k+1
```

```
grid.arrange(wykresy[[1]], wykresy[[1+1]], nrow=2)
```

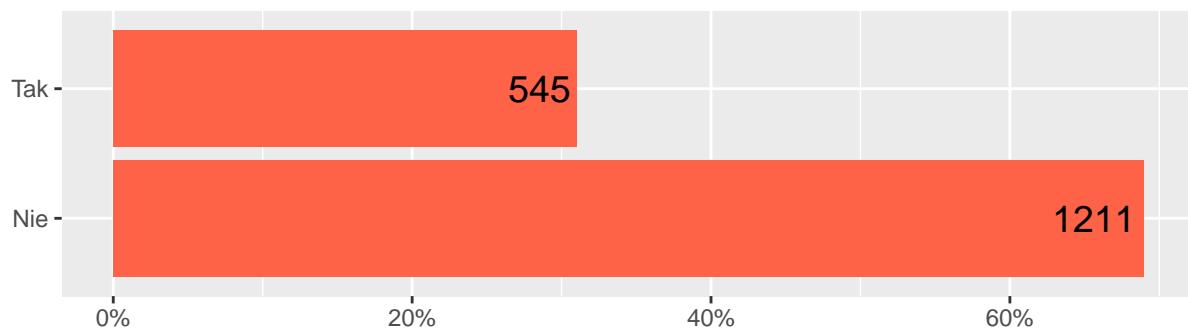
```
l <- l+2  
k <- k+1
```

Na wykresach 34, 35, 36 i 37 przedstawiono wykresy słupkowe dla zmiennych *OnlineSecurity*, *OnlineBackup*, *Device Protection* oraz *TechSupport*, z podziałem na klientów, którzy odeszli (churn) i tych, którzy pozostali.

Obecni klienci

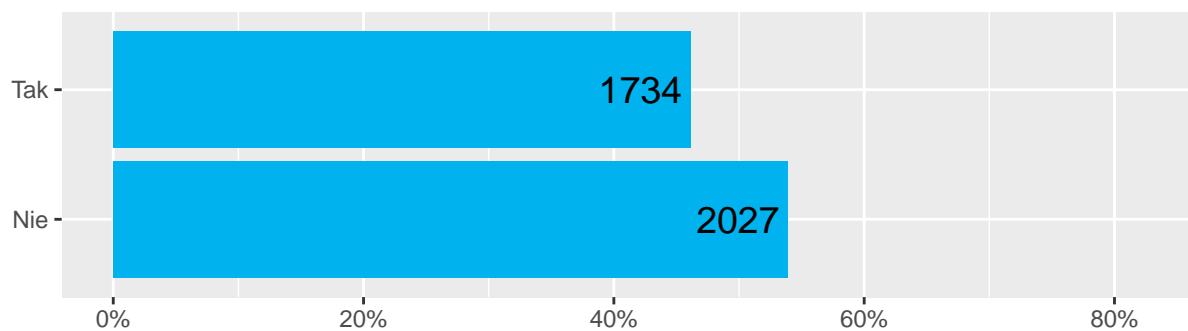


Dawni klienci

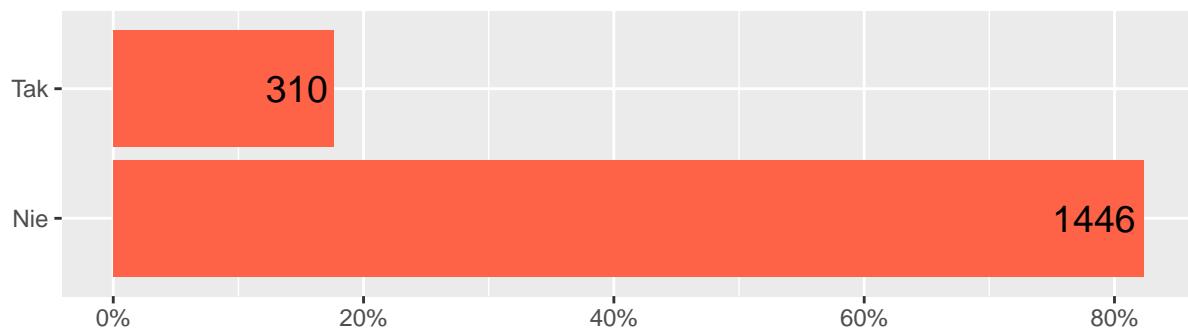


Wykres 36: Wykres słupkowy zmiennej DeviceProtection z podziałem ze względu na churn

Obecni klienci



Dawni klienci



Wykres 37: Wykres słupkowy zmiennej TechSupport z podziałem ze względu na churn

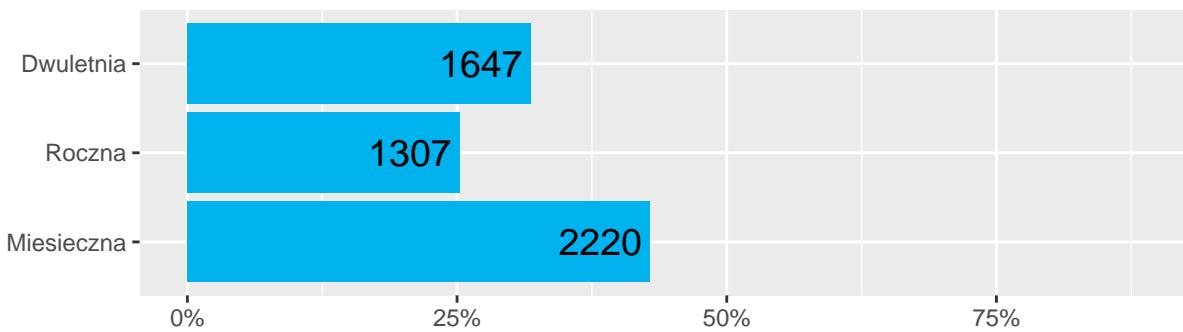
Widać, że w większości klienci, którzy odeszli, nie korzystali z tych rozwiązań, w przeciwieństwie do tych, którzy pozostały. Może to sugerować, że brak skorzystania z tych usług mógł mieć wpływ na decyzję o rezygnacji, wskazując na ich niezadowolenie lub brak postrzeganego zapotrzebowania na dodatkowe usługi zabezpieczające.

```
#grid.arrange(wykresy[[l]], wykresy[[l+1]], nrow=2)
l <- l+2
k <- k+1
```

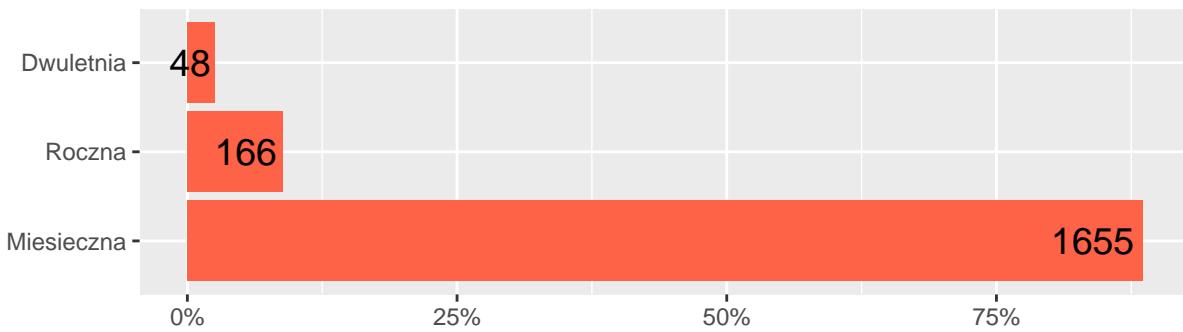
```
#grid.arrange(wykresy[[l]], wykresy[[l+1]], nrow=2)
l <- l+2
k <- k+1
```

```
grid.arrange(wykresy[[1]], wykresy[[1+1]], nrow=2)
```

Obecni klienci



Dawni klienci



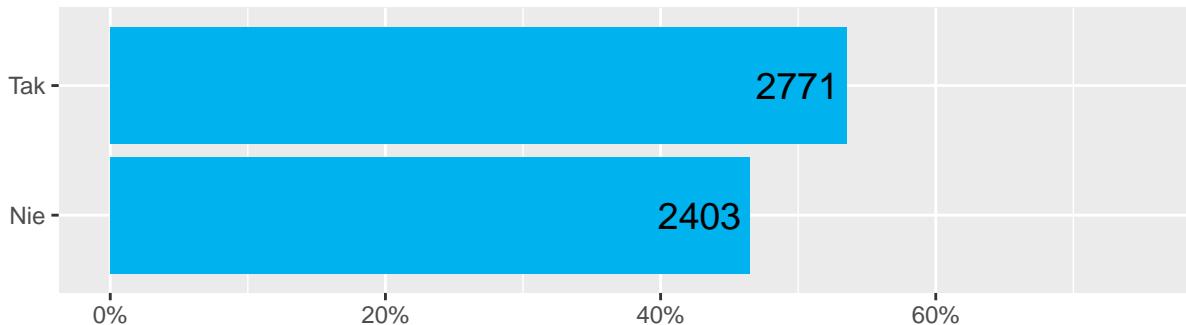
Wykres 38: Wykres słupkowy zmiennej *Contract* z podziałem ze względu na churn

```
l <- l+2
k <- k+1
```

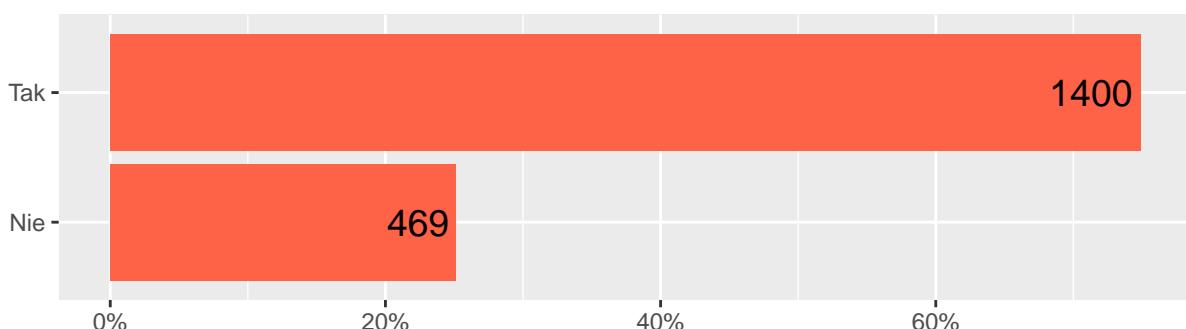
Na wykresie 38. przedstawiono wykres słupkowy dla zmiennej *Contract*, z podziałem na klientów, którzy odeszli (churn) i tych, którzy pozostały. Z analizy wykresu wynika, że klienci, którzy odeszli, w zdecydowanej większości preferowali umowy miesięczne. Natomiast klienci, którzy pozostały, również wybierali umowy miesięczne, ale w znacznie mniejszej skali.

```
grid.arrange(wykresy[[1]], wykresy[[1+1]], nrow=2)
```

Obecni klienci



Dawni klienci



Wykres 39: Wykres słupkowy zmiennej PaperlessBilling z podziałem ze względu na churn

```
l <- l+2  
k <- k+1
```

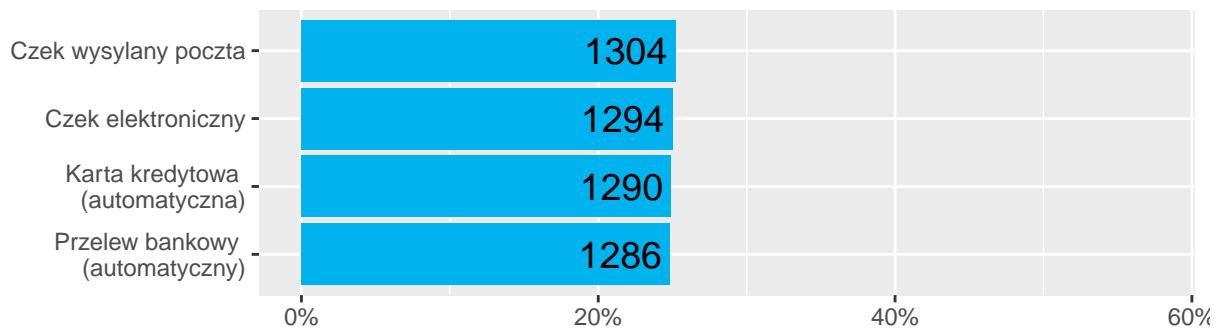
Na wykresie 39. przedstawiono wykres słupkowy dla zmiennej *PaperlessBilling*, z podziałem na klientów, którzy odeszli (churn) i tych, którzy pozostali. Widać, że klienci, którzy odeszli, preferowali biling bezpapierowy. Może to sugerować, że klienci, którzy preferowali tę formę rozliczeń, częściej decydowali się na rezygnację z usług.

```
grid.arrange(wykresy[[1]], wykresy[[1+1]], nrow=2)
```

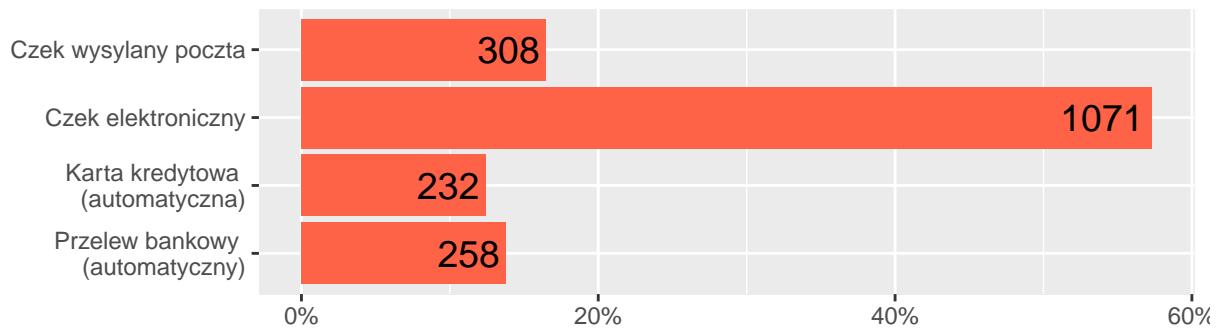
```
l <- l+2  
k <- k+1
```

Na wykresie 40. przedstawiono wykres słupkowy dla zmiennej *PaymentMethod*, z podziałem na klientów, którzy odeszli (churn) i tych, którzy pozostali. Widać, że klienci lojalni korzystają w miarę równo ze wszystkich metod płatności, natomiast w przypadku klientów, którzy odeszli, dominującą metodą płatności jest czek elektroniczny.

Obecni klienci



Dawni klienci



Wykres 40: Wykres słupkowy zmiennej PaymentMethod z podziałem ze względu na churn

6.3 Wykresy rozrzutów

```
aaa1 <- ggplot(obecni_klienci, aes(x=tenure,
                                         y=TotalCharges)) +
  geom_point(color="tomato") + stat_smooth(color="black") +
  theme(legend.position="none") +
  labs(title="Obecni klienci")

aaa2 <- ggplot(dawni_klienci, aes(x=tenure,
                                         y=TotalCharges)) +
  geom_point(color="deepskyblue2") + stat_smooth(color="black") +
  theme(legend.position="none") +
  labs(title="Dawni klienci")

grid.arrange(aaa1, aaa2, ncol=2)

## `geom_smooth()` using method = 'gam' and formula = 'y ~ s(x, bs = "cs")'

## Warning: Removed 11 rows containing non-finite outside the scale range
## ('stat_smooth()').

## Warning: Removed 11 rows containing missing values or values outside the scale range
## ('geom_point()').

## `geom_smooth()` using method = 'gam' and formula = 'y ~ s(x, bs = "cs")'
```

Wykres 41. przedstawia rozkład wartości zmiennej *TotalCharges* w zależności od *tenure*, z podziałem na zmienną *churn*. Widać, że w przypadku byłych klientów, dane koncentrują się wokół wyższych wartości, przy czym ich rozkład wykazuje charakter liniowy. Z kolei dla obecnych klientów, krzywa wygładzająca przechodzi przez środek rozkładu, wskazując na bardziej zróżnicowane zachowania w zakresie *TotalCharges*.

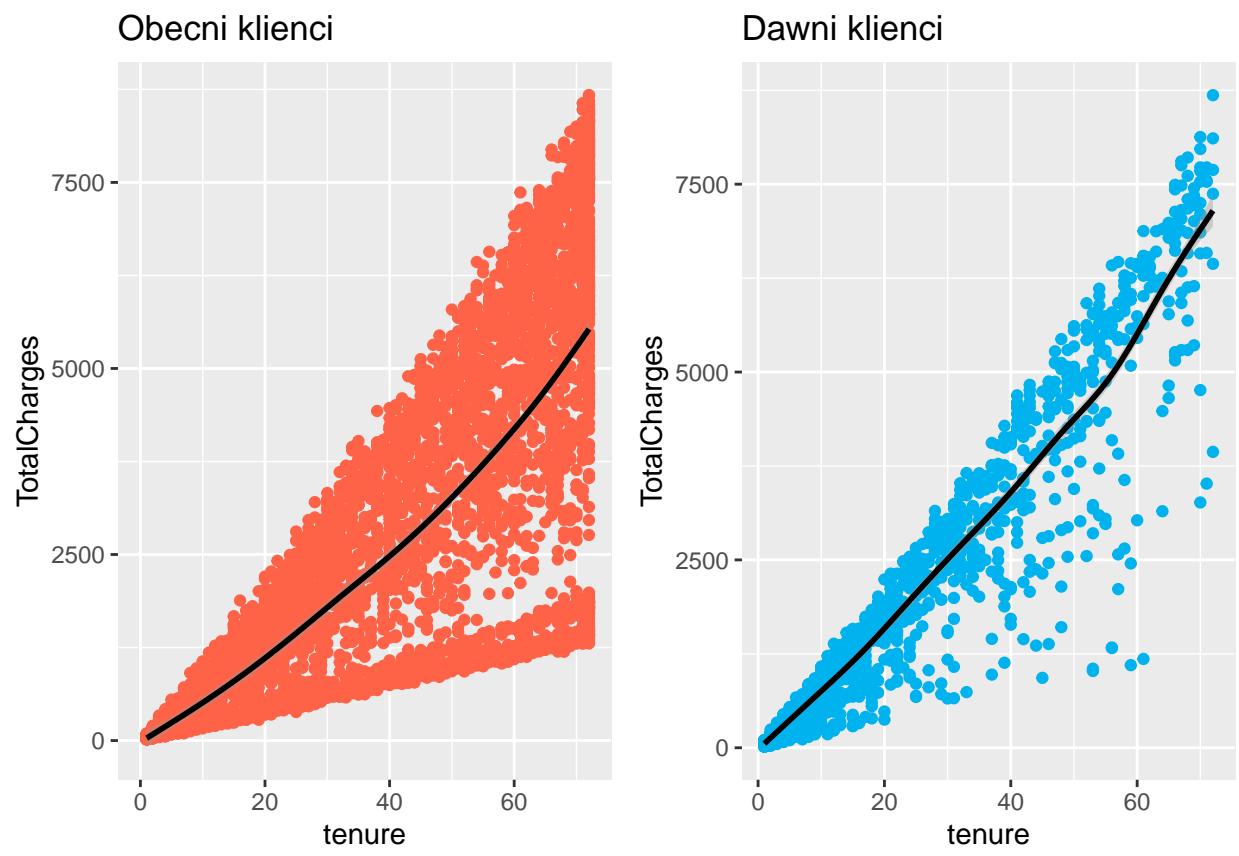
```
maks <- max(obecni_klienci$MonthlyCharges, dawni_klienci$MonthlyCharges)
mini <- min(obecni_klienci$MonthlyCharges, dawni_klienci$MonthlyCharges)

aaa1 <- ggplot(obecni_klienci, aes(x=tenure,
                                         y=MonthlyCharges)) +
  geom_point(color="tomato") + stat_smooth(color="black") +
  theme(legend.position="none") +
  labs(title="Obecni klienci") + ylim(mini, maks)

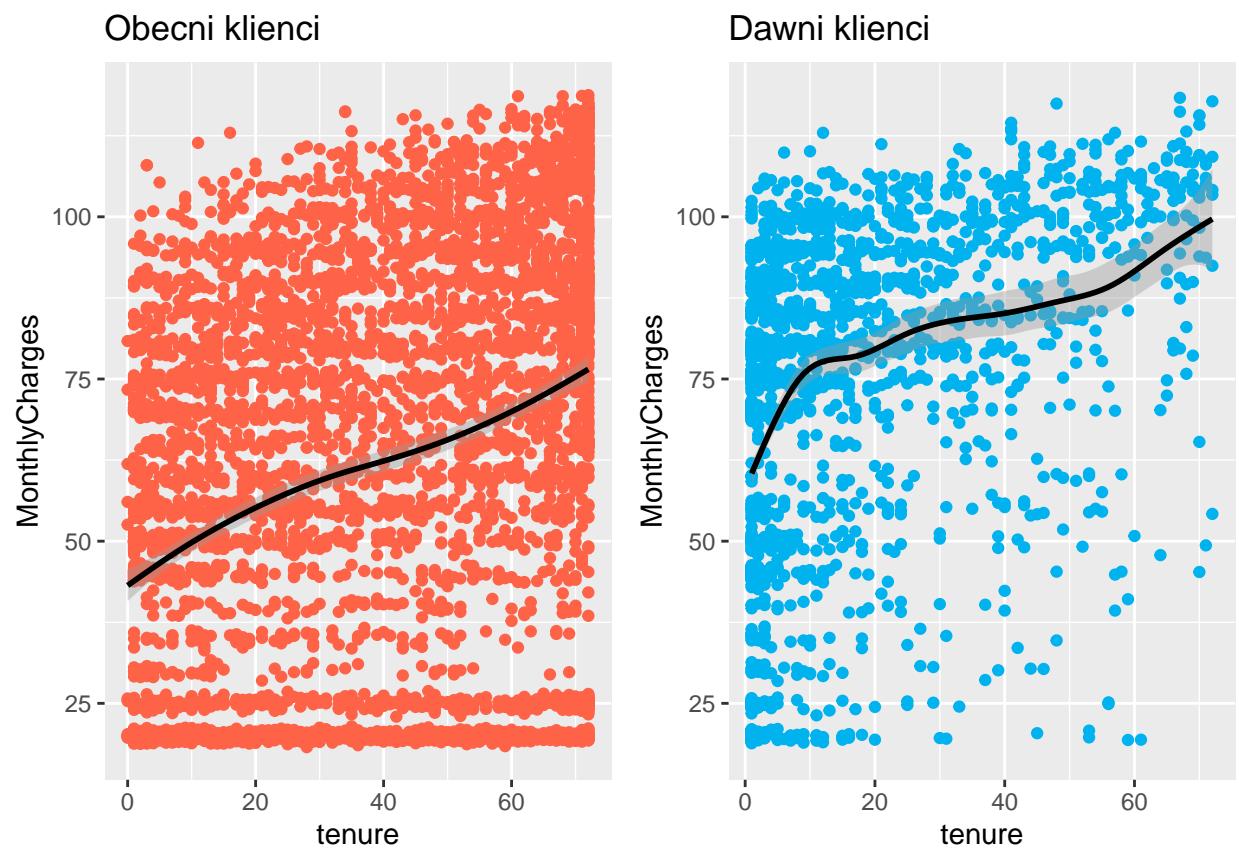
aaa2 <- ggplot(dawni_klienci, aes(x=tenure,
                                         y=MonthlyCharges)) +
  geom_point(color="deepskyblue2") + stat_smooth(color="black") +
  theme(legend.position="none") +
  labs(title="Dawni klienci") + ylim(mini, maks)

grid.arrange(aaa1, aaa2, ncol=2)

## `geom_smooth()` using method = 'gam' and formula = 'y ~ s(x, bs = "cs")'
## `geom_smooth()` using method = 'gam' and formula = 'y ~ s(x, bs = "cs")'
```



Wykres 41: Wykresy rozrzutów Total Charges ze względu na tenure z podziałem ze względu na zmienną churn

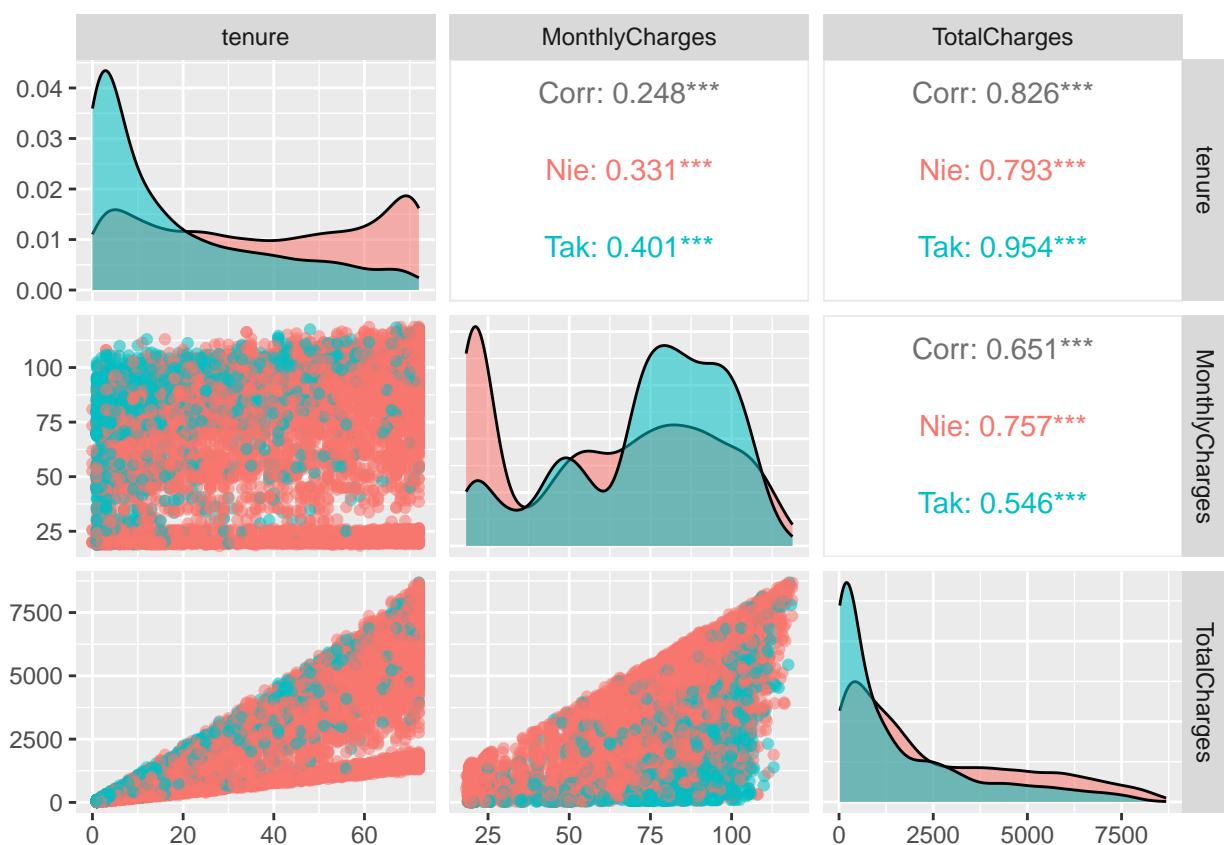


Wykres 42: Wykresy rozrzutów MonthlyCharges ze względu na tenure z podziałem ze względu na zmienną churn

Wykres 42. przedstawia wykresy rozrzutu *MonthlyCharges* w zależności od *tenure* z podziałem na zmienną *churn*. Widać, że obecni klienci stanowią pełny przekrój naszych klientów, obejmując różne poziomy opłat miesięcznych i długości współpracy. Klienci, którzy pozostali, utrzymują się bliżej maksymalnych wartości *MonthlyCharges*, co może sugerować, że korzystali z usług bardziej premium. Widać także, że jest ich zauważalnie więcej w przypadku wysokich *MonthlyCharges* i niskich *tenure*, co sugeruje, że ci klienci początkowo wykupili droższe usługi, ale później z nich zrezygnowali. Z kolei widać również niewielką liczbę klientów, którzy korzystali z tańszych, bardziej podstawowych usług.

```
num_data <- klienci[sapply(klienci, is.numeric)]  
  
num_data$Churn <- klienci$Churn  
  
ggpairs(num_data,  
        columns = 1:(ncol(num_data) - 1),  
        aes(color = factor(num_data$Churn), alpha = 0.5))  
  
## Warning: Use of 'num_data$Churn' is discouraged.  
## i Use 'Churn' instead.  
  
## Warning in ggally_statistic(data = data, mapping = mapping, na.rm = na.rm, :  
## Removed 11 rows containing missing values  
  
## Warning: Use of 'num_data$Churn' is discouraged.  
## i Use 'Churn' instead.  
## Use of 'num_data$Churn' is discouraged.  
## i Use 'Churn' instead.  
  
## Warning in ggally_statistic(data = data, mapping = mapping, na.rm = na.rm, :  
## Removed 11 rows containing missing values  
  
## Warning: Use of 'num_data$Churn' is discouraged.  
## i Use 'Churn' instead.  
  
## Warning: Removed 11 rows containing missing values or values outside the scale range  
## ('geom_point()').  
  
## Warning: Use of 'num_data$Churn' is discouraged.  
## i Use 'Churn' instead.  
  
## Warning: Removed 11 rows containing missing values or values outside the scale range  
## ('geom_point()').  
  
## Warning: Use of 'num_data$Churn' is discouraged.  
## i Use 'Churn' instead.  
  
## Warning: Removed 11 rows containing non-finite outside the scale range  
## ('stat_density()').
```

Na wykresie 43. przedstawiono macierz par zmiennych numerycznych z podziałem ze względu na *churn*. Widać w niej bardzo wysoką (0,954) korelację między *TotalCharges* a *tenure*, co wskazuje, że długość trwania współpracy z klientem (tenure) ma silny wpływ na całkowite opłaty (*TotalCharges*). Klienci, którzy pozostali z firmą przez dłuższy czas, zazwyczaj generują wyższe całkowite opłaty, co sugeruje, że dłuższa lojalność wiąże się z większymi wydatkami na usługi.



Wykres 43: Macierz par zmiennych numerycznych z podziałem ze względu na churn

7 Dyskusja

1. Dłuższy okres współpracy (tenure) i zawieranie długoterminowych umów (roczne, dwuletnie) sprzyjają lojalności klientów.
2. Wyższe miesięczne opłaty (MonthlyCharges) korelują z wyższym ryzykiem rezygnacji.
3. Klienci korzystający z dodatkowych usług (OnlineSecurity, TechSupport, OnlineBackup) wykazują większą lojalność.
4. Usługi telefoniczne są kluczowym elementem oferty – około 90% klientów z nich korzysta.
5. Internet – najczęściej technologia światłowodowa (43%) oraz DSL (35%); około 22% klientów nie korzysta z internetu.
6. Klienci preferują biling elektroniczny (PaperlessBilling) oraz umowy miesięczne ze względu na elastyczność.
7. Klienci odchodzący rezygnują z usług wsparcia IT (wykresy 34–37) przy wysokich opłatach miesięcznych (wykres 29), co może wskazywać na postrzeganie tych usług jako zbędnych lub brak atrakcyjnych pakietów łączących wsparcie z innymi usługami.
8. Około 10% dawnych klientów korzystało z usług przez mniej niż rok, a około 15% przez dwa lata (wykres 27) – wskazuje to na podejście „testowe” w korzystaniu z oferty.
9. Wśród obecnych klientów około 21% korzysta z usług krócej niż rok, a 35% poniżej dwóch lat – warto monitorować tę grupę, by zapobiegać ich odejściu.
10. Ponad 85% dawnych klientów miało umowy miesięczne, co ułatwia rezygnację bez długoterminowych zobowiązań.
11. Duża liczba dawnych klientów korzystała ze światłowodu (wykres 33), co może wskazywać na niezadowolenie z tej usługi.
12. Klienci mają tendencję do dokupowania dodatkowych usług – obecni klienci robią to równomiernie, natomiast dawni intensywnie w pierwszym roku, co może sugerować, że intensywne „testowanie” usług nie przekłada się na długoterminową satysfakcję (wykres 42).
13. Umiarkowana zmienność opłat miesięcznych (wykres 23) oznacza, że wyższe łączne wydatki nie muszą iść w parze z wysokimi miesięcznymi opłatami – mogą występować zarówno wieloletni klienci z niskimi opłatami, jak i nowi wybierający droższe pakiety.
14. Około 20% klientów nie korzysta z internetu, natomiast tylko 10% rezygnuje z usług telefonicznych (wykresy 11 i 13), co sugeruje, że oferta telefoniczna jest bardziej atrakcyjna.

8 Wnioski

- Klienci, którzy odchodzą, zazwyczaj ponoszą wyższe koszty miesięczne, co sugeruje, że dla nich opłaty nie są adekwatne do otrzymywanych usług.
- Dominacja umów miesięcznych wśród klientów odchodzących wskazuje, że wielu z nich korzysta z oferty na zasadzie testowej, nie zobowiązując się długoterminowo.
- Rezygnacja z usług wsparcia IT przy wysokich opłatach sugeruje, że klienci nie widzą wystarczającej wartości w dodatkowych pakietach lub nie są one odpowiednio dopasowane cenowo, przez co są dodatkowo narażeni na większe zagrożenia w internecie, co może ich zniechęcać.
- Brakuje podstawowych danych demograficznych, takich jak wiek klientów, co utrudnia dokładniejszą klasyfikację i identyfikację segmentów, dla których oferta mogłaby być bardziej dopasowana.

Te czynniki łącznie wskazują, że główną przyczyną odchodzenia klientów jest niezadowalająca relacja między kosztem a wartością oferowanych usług, szczególnie w kontekście elastycznych, miesięcznych umów.

9 Rekomendacje

Stworzenie atrakcyjnych pakietów łączonych, które integrują podstawowe usługi ze wsparciem IT w konkurencyjnej cenie.

Wdrożenie programów lojalnościowych premiujących długoterminową współpracę, np. poprzez rabaty czy bonusy.

Personalizacja oferty, w tym przygotowanie specjalnych pakietów startowych dla nowych klientów testujących usługę oraz ekskluzywnych pakietów premium dla klientów lojalnych.

Proaktywna komunikacja z klientami, zwłaszcza tymi, którzy korzystają z usług krócej niż rok lub dwa lata, aby wcześnie zidentyfikować potencjalne niezadowolenie. **Analiza satysfakcji klientów korzystających ze światłowodu**, ponieważ możliwe, że niezadowolenie z tej technologii było przyczyną rezygnacji wielu klientów.

Kontakt z klientami, którzy odeszli przed upływem 24 miesięcy, oraz monitorowanie obecnych klientów w tym przedziale (obecnie 31%), aby zwiększyć ich retencję.

Uatrakcyjnienie usług IT, np. poprzez wprowadzenie okresów próbnych lub oferowanie ich w korzystnych pakietach za niewielką dopłatą.

Identyfikacja klientów o najniższych miesięcznych opłatach i zaproponowanie im dodatkowych usług, które mogłyby zwiększyć ich zaangażowanie i wartość dla firmy.

Nieznajomość szczegółowych informacji o firmie i jej ofercie znacząco utrudnia trafne dopasowanie rekommendacji.