

```
##  
## Attaching package: 'dplyr'  
## The following objects are masked from 'package:stats':  
##  
##     filter, lag  
## The following objects are masked from 'package:base':  
##  
##     intersect, setdiff, setequal, union
```

# Raport 1 - Analiza firmy Telekomunikacyjnej

Romana Żmuda

249706

Adrian Kit

249746

5 kwietnia 2020

## Spis treści

<b>1 Krótki opis problemu, czyli co analizujemy i dlaczego?</b>	<b>2</b>
<b>2 Opis eksperymentów</b>	<b>2</b>
<b>3 Etap 1 - Przygotowanie danych</b>	<b>3</b>
3.1 A . . . . .	3
3.2 B . . . . .	4
<b>4 Etap 2 - Analiza opisowa</b>	<b>5</b>
4.1 A - Podstawowe wskaźniki sumaryczne . . . . .	5
4.1.1 Cechy jakościowe . . . . .	5
4.1.2 Cechy ilościowe . . . . .	7
4.2 B . . . . .	16
<b>5 ETAP 3</b>	<b>17</b>
5.0.1 Jakościowe . . . . .	17
5.0.2 Ilościowe . . . . .	20
<b>6 ETAP 4</b>	<b>24</b>

## 1 Krótki opis problemu, czyli co analizujemy i dlaczego?

- Co analizujemy? Badamy zachowanie klientów sieci telekomunikacyjnej w celu przewidzenia przyszłych zachowań klientów na podstawie 3333 klientów oraz 21 cechach charakteryzujących. Między innymi mamy:
  - Skąd pochodzi?
  - Jak długo korzysta/ł z planu?
  - Ile dany klient dzwonił w określonej porze dnia?
  - Czy przystąpił do planu poczty głosowej lub planu międzynarodowego?
  - Ile razy dzwonił na infolini?

- Czy zrezygnował z oferty firmy?
- Nasz problem badawczy? Badanie wpływu oferty firmy na pozostanie klientów oraz wy ciąganie wniosków z analizy danych klientów, którzy odeszli w celu utrzymania większej ilości klientów. Tym samym interesuje Nas fakt, jakie zmiany musi wprowadzić firma, aby utrzymać jak najwięcej klientów oraz sprawdzić opłacalność Planów zaproponowanych dla klientów, które z nich realnie przyczynią się do utrzymania klienta. Jak również bardziej spersonalizować usługi do potrzeb klientów.

## 2 Opis eksperymentów

W naszym raporcie korzystaliśmy z histogramów, boxplotów i wielu innych graficznych reprezentacji danych uwzględniając poszczególne typy zmiennych. Więcej o samym użyciu, zakresach i rozkładach będziemy pisać w poszczególnych podsekcjach, aby zachować przejrzystość raportu.

## 3 Etap 1 - Przygotowanie danych

### 3.1 A

Wczytujemy dane do przestrzeni R :

```
dane <- read.csv(file="churn.txt")
attach(dane)
```

Sprawdzamy, czy dane zostały dobrze rozpoznane:

```
Jakosciowe <- which(sapply(dane, is.factor)) #Cechy jakościowe
Jakosciowe

##      State       Phone Int.l.Plan VMail.Plan      Churn.
##        1          4           5           6          21

Ilosciowe <- which(sapply(dane, is.numeric)) #Cechy ilościowe
Ilosciowe

## Account.Length       Area.Code   VMail.Message      Day.Mins      Day.Calls
##        2                  3             7                 8                   9
## Day.Charge       Eve.Mins     Eve.Calls     Eve.Charge    Night.Mins
##        10                 11            12                 13                  14
## Night.Calls   Night.Charge   Intl.Mins   Intl.Calls   Intl.Charge
##        15                 16            17                 18                  19
## CustServ.Calls
##        20

sapply(dane, class)
```

```

##           State Account.Length      Area.Code       Phone   Int.l.Plan
## "factor"    "integer"          "integer"    "factor"    "factor"
## VMail.Plan  VMail.Message     Day.Mins     Day.Calls  Day.Charge
## "factor"    "integer"         "numeric"    "integer"    "numeric"
## Eve.Mins    Eve.Calls        Eve.Charge   Night.Mins  Night.Calls
## "numeric"   "integer"         "numeric"    "numeric"    "integer"
## Night.Charge Intl.Mins      Intl.Calls  Intl.Charge CustServ.Calls
## "numeric"   "numeric"        "integer"    "numeric"    "integer"
##             Churn.          "factor"
## "factor"

```

Porównując otrzymane dane z oficjalnym opisem zmiennych z pliku *churn.txt*:

Nazwa	State	Area code	Int.l.Plan	VMail.plan	Chunk
Typ	Jakościowe	Jakościowe	Jakościowe	Jakościowe	Jakościowe
Nazwa	Account.Length	Phone	VMail.Message	Day.Mins	
Typ	Jakościowe	0	Jakościowe	Jakościowe	

- Zauważamy, że R źle zinterpretował funkcje *Phone*, u nas pełni rolę identyfikatora, jednak jego liczbowy zapis program R utożsamia z typem ilościowym, ciągłym.
- Zmienna *Area.Code* również źle została rozpoznana, gdyby była cechą ilościową, to liczenie średniej od liczb porządkujących dostarczałoby informacji do niczego nieprzydatnych. Zmienna ta występuje, jako ciąg liczbowy, więc znowu R funkcjonalnie przydzielił ją do cechy Ilościowej.

## 3.2 B

- Rozmiar danych:

```

ncol(dane) # ilość kolumn
## [1] 21

nrow(dane) # ilość przypadek
## [1] 3333

```

- Typy poszczególnych cech:

Nazwa	Porządkowe	Nominalne	Ciągłe	Dyskretne
State		X		
Account.Length			X	
Area.Code	X			
Int.l.Plan	X			
VMail.Plan	X			
VMail.Message				X
Day.Mins			X	
Day.Calls				X
Day.Charge		X		
Eve.Mins			X	
Eve.Calls				X
Eve.Charge			X	
Night.Mins			X	
Night.Calls				X
Night.Charge			X	
Intl.Mins			X	
Intl.Calls				X
Intl.Charge			X	
Intl.Calls				X
CustServ.Calls				X
Churn.	X			

- Zmienna *Phone* jest zmienną niepotrzebną w dalszej analizie, gdyż pełni ona rolę identyfikatora danego klienta.

```
dane <- select(dane, -Phone) #usunięcie niepotrzebnej zmiennej Phone
```

- Nie stwierdziliśmy żadnych braków w danych:

```
ilebrakujacych<-sum(is.na(dane))

ilebrakujacych #nie brakuje żadnej wartości

## [1] 0
```

- Nie zauważliśmy żadnych nietypowych wartości

## 4 Etap 2 - Analiza opisowa

### 4.1 A - Podstawowe wskaźniki sumaryczne

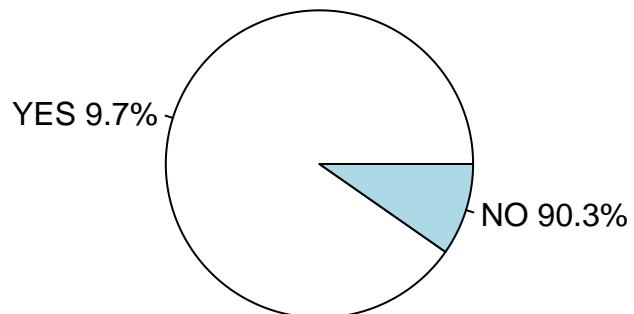
#### 4.1.1 Cechy jakościowe

- Zmienna: *Int.l.Plan*

Zmienna informuje o przystąpieniu do planu międzynarodowego

```
## Int.1.Plan  
##   no   yes  
## 3010   323
```

## Rozkład planu międzynarodowego

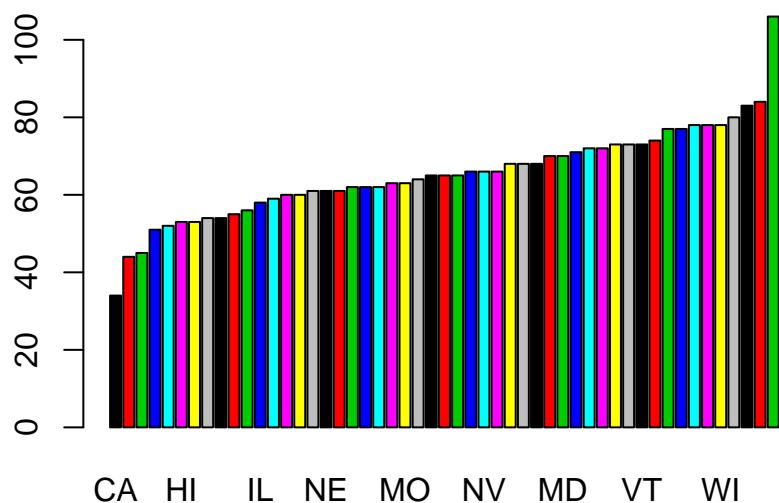


90

- Zmienna: State

```
liczebnosci_State <- table(dane$State)  
k<-sort(table(dane$State))  
barplot(k, col=1:length(levels(State)))  
title("Wykres słupkowy (barplot) dla zmiennej State")
```

## Wykres słupkowy (barplot) dla zmiennej State



```
mediana_State<-median(k)
srednia_State<-mean(k)
range_State<-range(k)
mediana_State #mediana ilości osób przypadających na dany Stan

## [1] 65

srednia_State # średnia ilości osób przypadających na dany Stan

## [1] 65.35294

range_State #zakres wartości zmiennej State

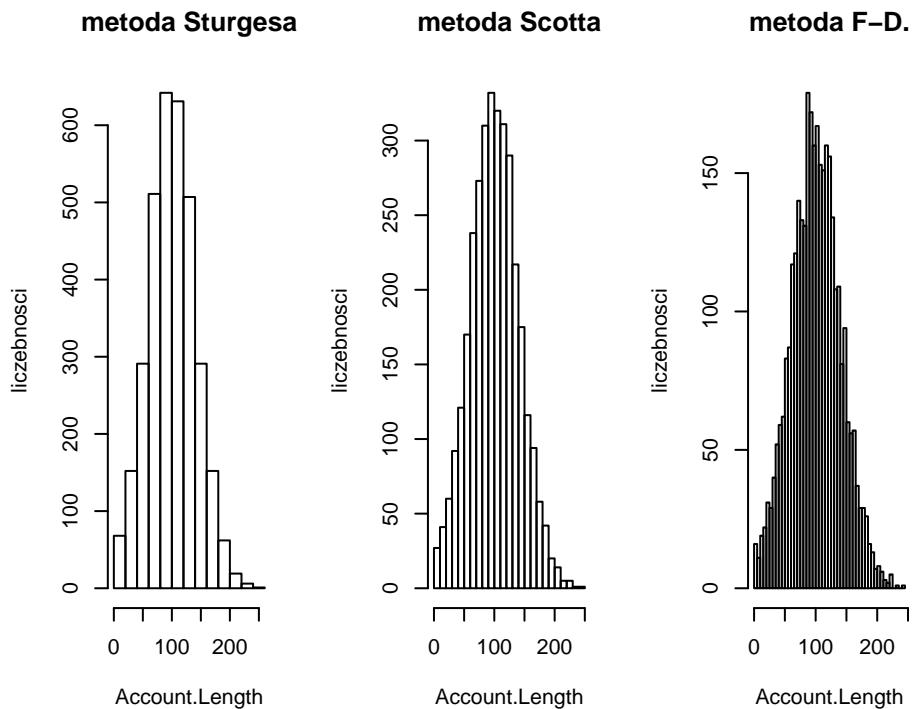
## [1] 34 106
```

Wnioski: Najwięcej osób jest ze stanu WV (Virginia Zachodnia), a najmniej w CA (California), co obrazują dwie mocno skrajne wartości.

### 4.1.2 Cechy ilościowe

- Zmienna: *Account.Length*

Zważywszy na duży rozrzut danych (zakres od 1 do 243) stosujemy 3 metody wyboru optymalnej szerokości przedziału w histogramie.



Poszczególne ilości klas dla różnych metod:

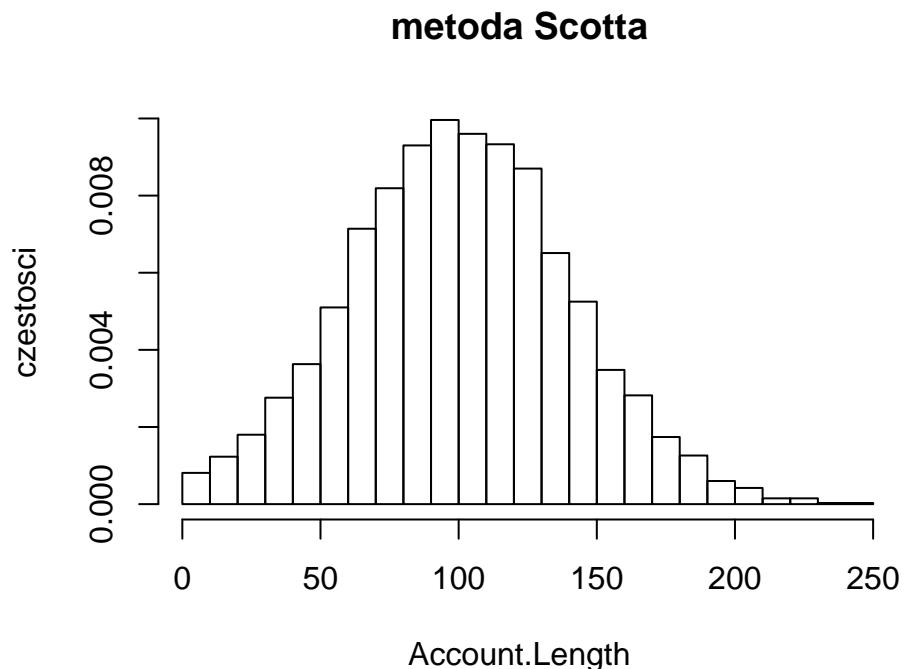
```
nclass.Sturges(Account.Length) #Sturges
## [1] 13

nclass.scott(Account.Length) #Scott
## [1] 26

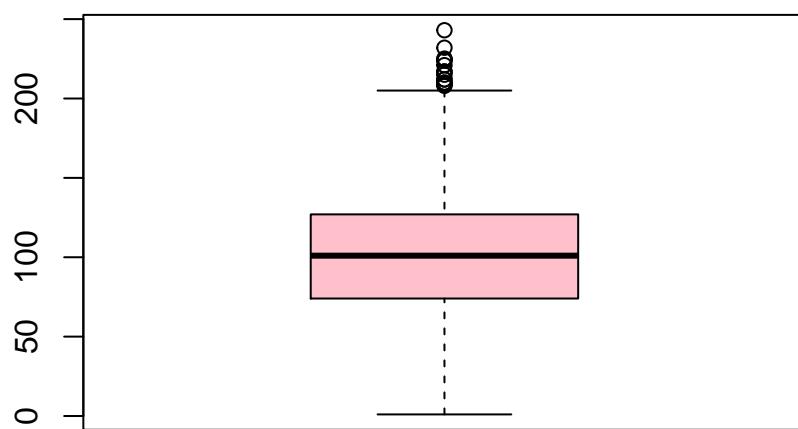
nclass.FD(Account.Length) #Freedman-Diaconis
## [1] 35
```

Naszym zdaniem, metoda Scotta jest najbardziej optymalna, ponieważ zachowuje przejrzystość przy dużej liczbie klas.

Wykres częstości metodą Scotta oraz wskaźniki sumaryczne:



### Wykres pudelkowy (boxplot) – zmienna Account.Length



```
Summary_Account.Length<-my.summary(Account.Length)
Summary_Account.Length #wskaźniki sumaryczne
```

```
##           min         Q1      median        mean         Q3        max       var
## 1.000000  74.000000 101.000000 101.06481 127.00000 243.00000 1585.80012
```

```

##           sd      IQR
##    39.82211  53.00000

getmode <- function(v) {
  uniqv <- unique(v)
  uniqv[which.max(tabulate(match(v, uniqv)))]
}
getmode(dane$Account.Length) #dominanta

## [1] 105

skewness_Account.Length<-skewness(Account.Length)
skewness_Account.Length #skośność

## [1] 0.09656281

splaszczenie_Account.Length<-kurtosis(Account.Length)
splaszczenie_Account.Length # spłaszczenie

## [1] 2.890526

```

Wnioski: Histogram jest jednomodalny, prawostronnie skośny, wyostrzony (rozkład leptokuryczny)

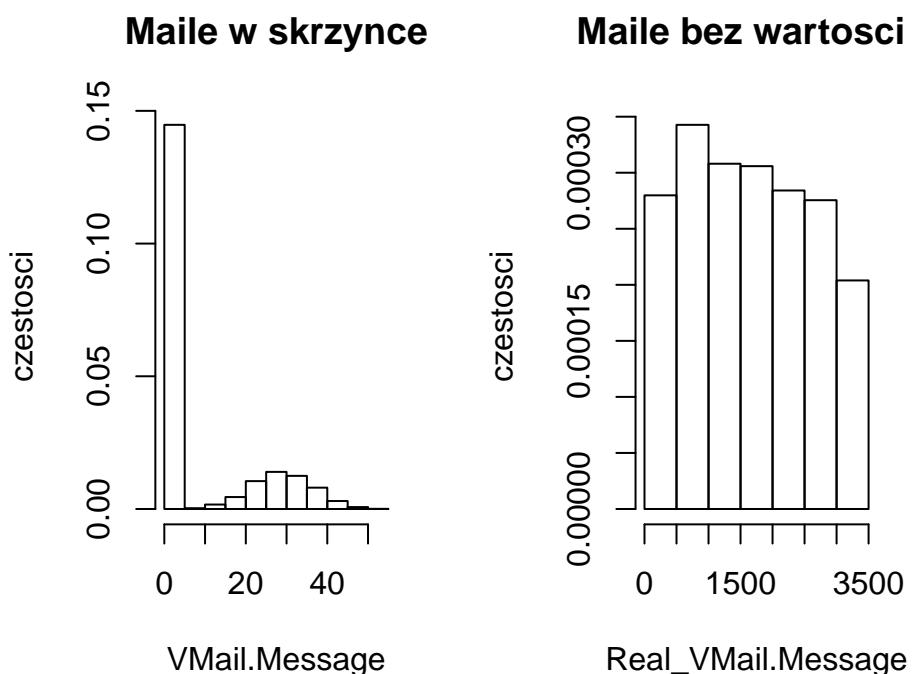
- Zmienna: *VMail.Message*

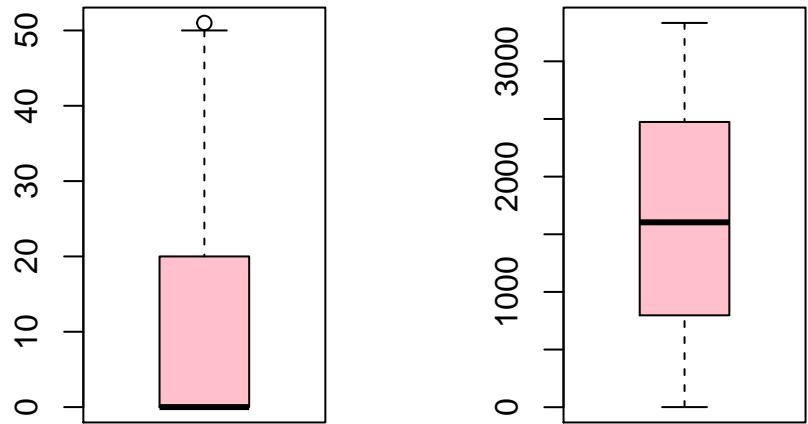
Zanim zajmiemy się analizą tej zmienny rozważmy ile osób realnie korzysta z oferty Poczty głosowej

```
## VMail.Plan
##      no      yes
## 72.33723 27.66277
```

Widzimy, że z oferty poczty głosowej korzysta mniej niż jedna trzecia osób, ponieważ wartość zero występuje bardzo często( reprezentacja osób, które nie biorą udziału w ofercie) usuwamy ją, aby zobaczyć, jak realnie klienci wykorzystują możliwości planu poczty głosowej.

Histogram zmiennej: Ilości wiadomości w poczcie głosowej





```

Summary_VMail.Message<-my.summary(VMail.Message)
Summary_VMail.Message

##      min       Q1     median      mean       Q3      max      var      sd
## 0.00000 0.00000 0.00000 8.09901 20.00000 51.00000 187.37135 13.68837

#wskaźniki sumaryczne dla zmiennej VMail.Message

Summary_Real_VMail.Message<-my.summary(Real_VMail.Message)
Summary_Real_VMail.Message

##      min       Q1     median      mean       Q3      max
## 1.0000 797.5000 1603.5000 1645.0694 2473.5000 3333.0000
##      var      sd      IQR
## 922125.3676 960.2736 1676.0000

#wskaźniki sumaryczne dla zmiennej VMail.Message bez wartości "0"

zakres <- range(dane$VMail.Message)
zakres #zakres dla normalnej zmiennej ilości wiadomości w poczcie

## [1] 0 51

```

```

zakres <- range(dane_Real_VMail$Message$VMail$Message)
zakres # zakres zmiennej bez wartości 0

## [1] 4 51

skewness_VMail$Message<-skewness(VMail$Message)
skewness_VMail$Message #skośność zmiennej VMail$Message

## [1] 1.264254

skewness_Real_VMail$Message<-skewness(Real_VMail$Message)
skewness_Real_VMail$Message #skośność zmiennej bez wartości 0

## [1] 0.06354856

splaszczanie_VMail$Message<-kurtosis(VMail$Message)
splaszczanie_VMail$Message #spłaszczenie zmiennej VMail$Message

## [1] 2.947148

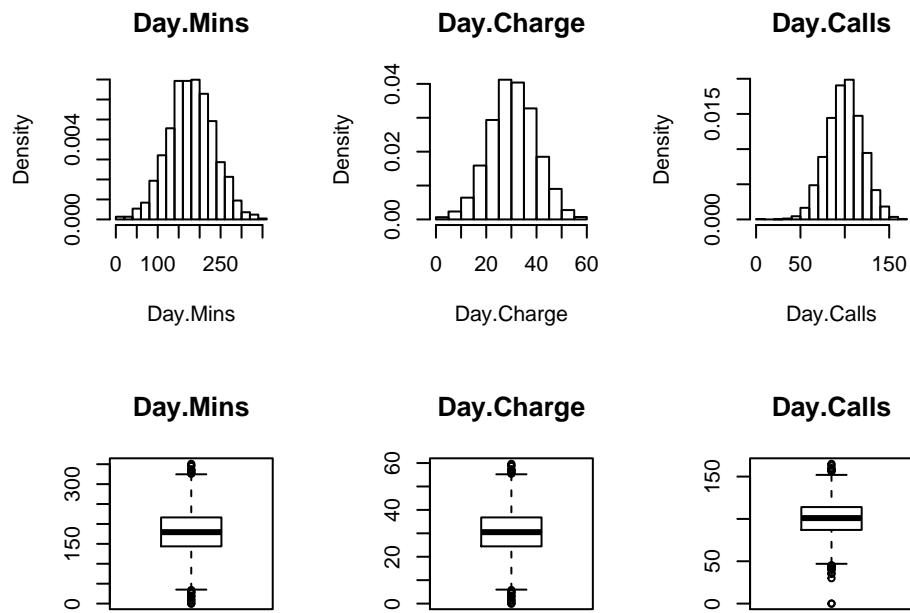
splaszczanie_Real_VMail$Message<-kurtosis(Real_VMail$Message)
splaszczanie_Real_VMail$Message #spłaszczenie zmiennej bez wartości 0

## [1] 1.801158

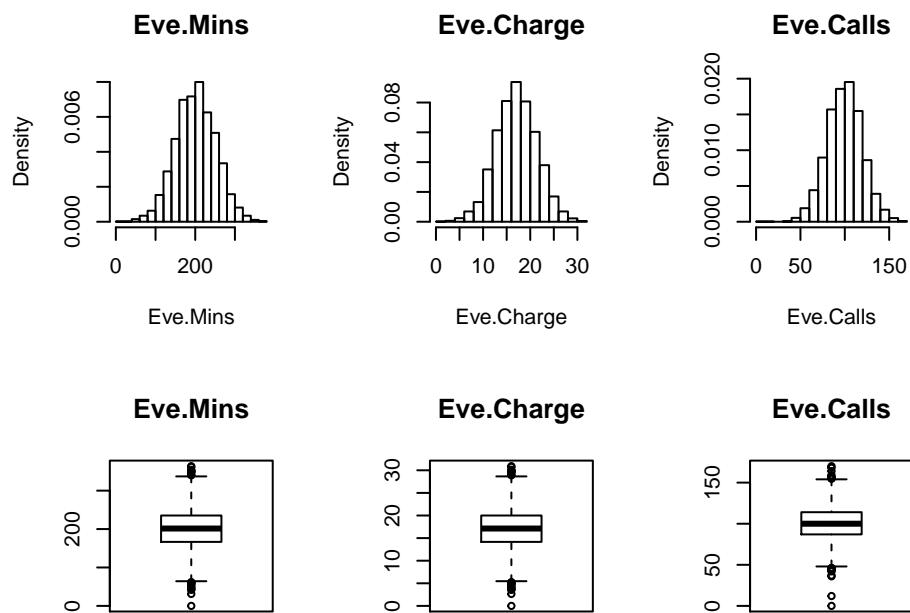
```

Wnioski: Widzimy, że wartość "0" mocno zaburza możliwości obserwacji. Klienci w znacznym stopniu korzystają z funkcji poczty głosowej.

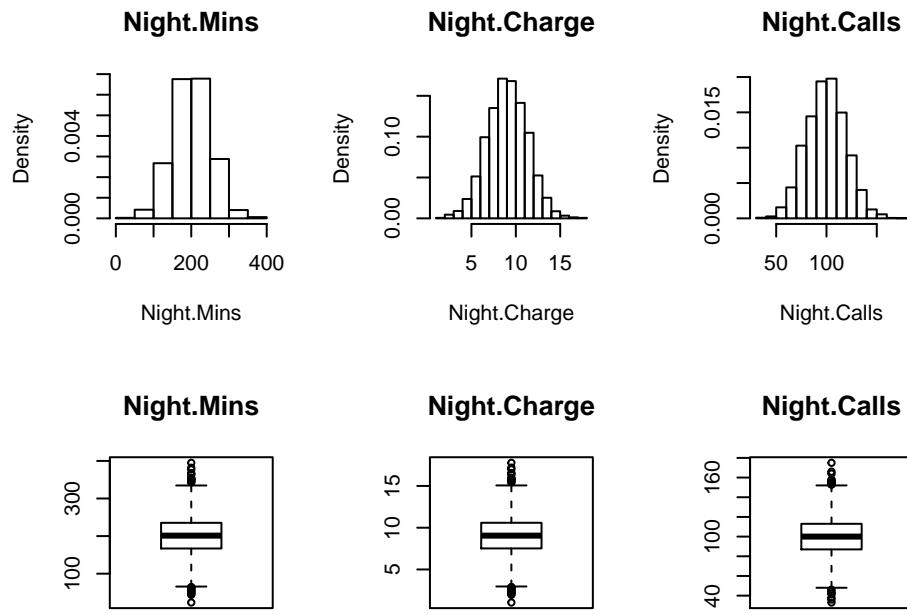
- Zmienne: minut, liczby połączeń, opłat za okres: do południa, po południu, w nocy oraz połączeń międzynarodowych
- Analiza minut, połączeń oraz opłat za poranek



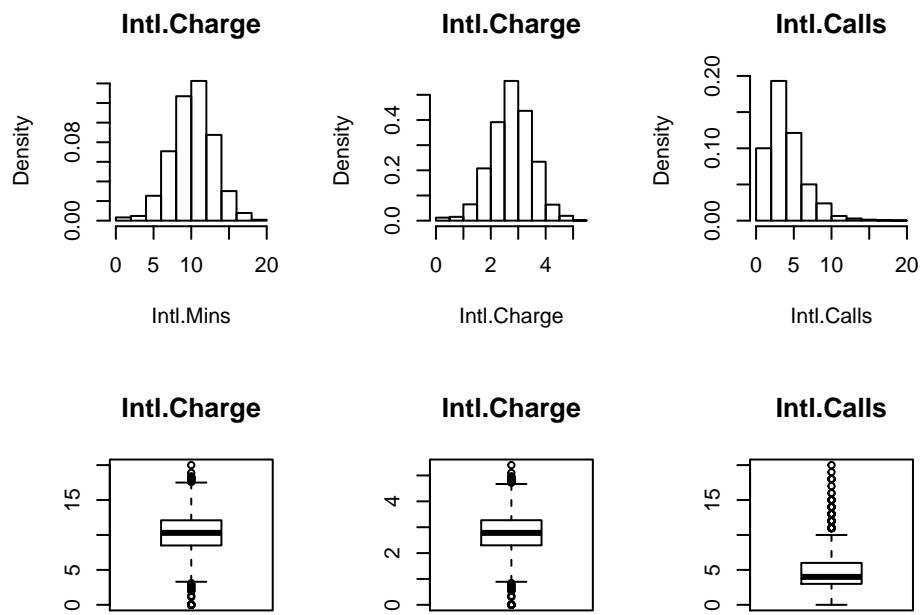
Analiza minut, połączeń oraz opłat za popołudnie



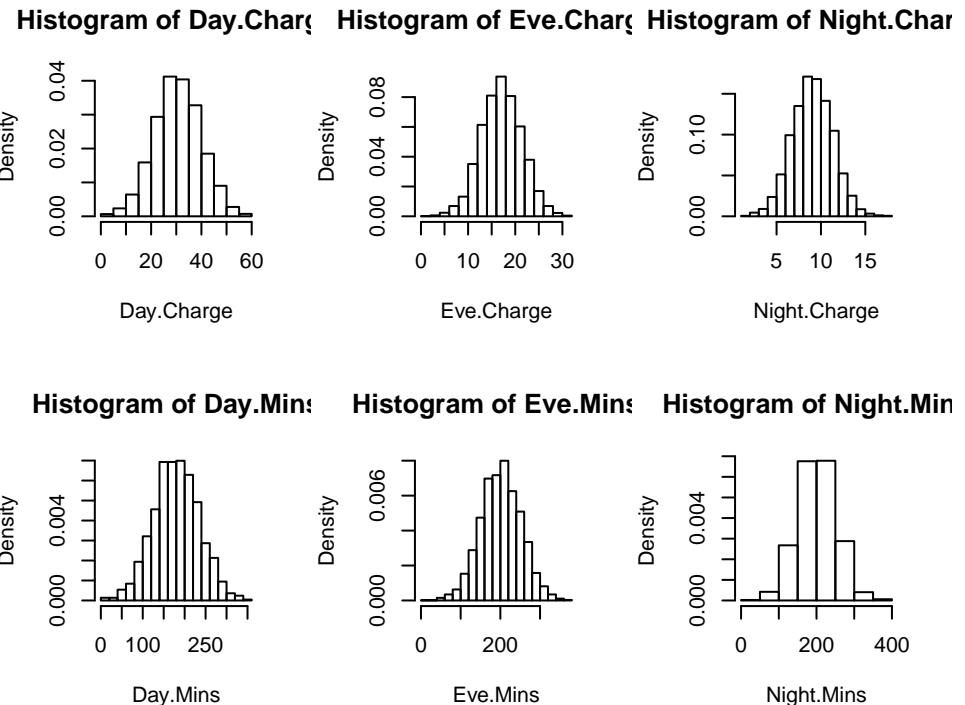
Analiza minut, połączeń oraz opłat za wieczór



Analiza minut, połączeń oraz opłat za połączenia międzynarodowe



Porównanie różnych pór w ciągu dnia z ilością wykonywanych połączeń wraz z długością połączeń



Zauważamy, że każdy z tych histogramów ma podobny kształt oraz na podobnym poziomie prawdopodobienstwo występowania.

```
skewness_Day.Charge<-skewness(Day.Charge)
skewness_Eve.Charge<-skewness(Eve.Charge)
skewness_Night.Charge<-skewness(Night.Charge)
c_skewness_calls <-c(skewness_Day.Charge,skewness_Eve.Charge,skewness_Night.Charge)
c_skewness_calls #skośność opłat dla różnych pór dnia

## [1] -0.029070178 -0.023847250  0.008882237
```

Wszystkie połączenia mają rozkład bardzo zbliżony do rozkładu normalnego  
Wskaźniki sumaryczne dla długości rozmów w różnych porach dnia:

```
Summary_Day.Mins<-my.summary(Day.Mins)
Summary_Eve.Mins<-my.summary(Eve.Mins)
Summary_Night.Mins<-my.summary(Night.Mins)
c_summary_calls_mins<-cbind(Summary_Day.Mins,Summary_Eve.Mins,Summary_Night.Mins)
c_summary_calls_mins

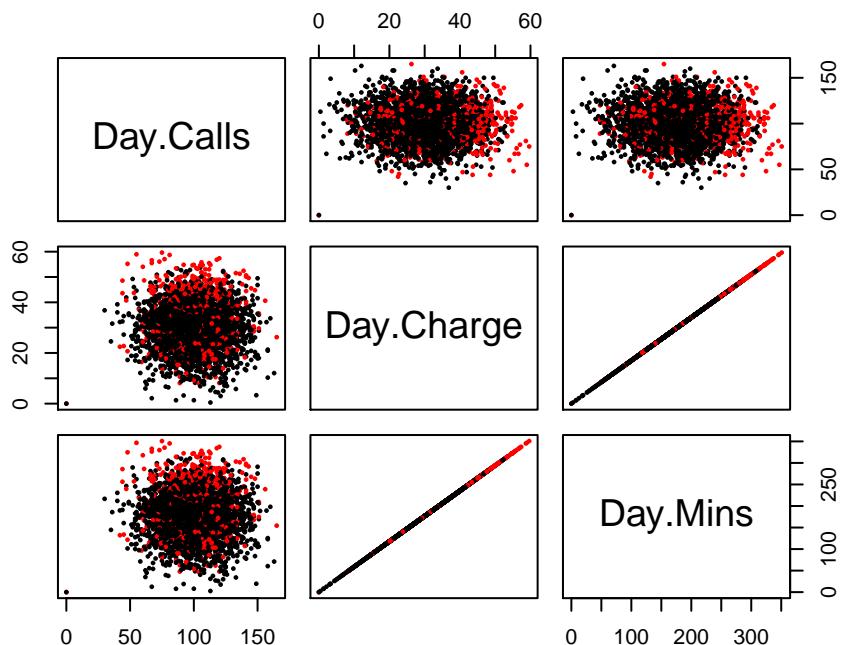
##           Summary_Day.Mins Summary_Eve.Mins Summary_Night.Mins
## min          0.00000      0.00000     23.20000
## Q1         143.70000    166.60000    167.00000
## median     179.40000    201.40000    201.20000
## mean        179.77510    200.98035    200.87204
## Q3         216.40000    235.30000    235.30000
## max        350.80000    363.70000    395.00000
```

## var	2966.69649	2571.89402	2557.71400
## sd	54.46739	50.71384	50.57385
## IQR	72.70000	68.70000	68.30000

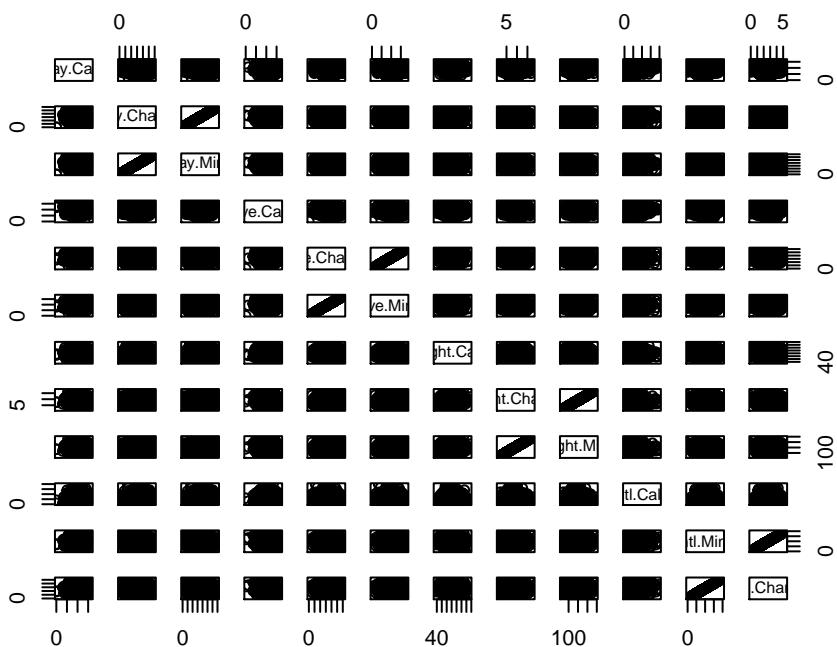
## 4.2 B

Szukanie korelacji między zmiennymi W tym podpunkcie przeanalizujemy dane pod kątem potencjalnych korelacji cech. W tym celu skorzystamy z wykresów rozrzutu.

Wykres rozrzutu między połączeniami w ciągu dnia, a minutami



Korelacja wielu zmiennych



Widać, że jedyne korelacje występują między długością trwania połączeń a opłatami, niezależnie od pory dnia.

## 5 ETAP 3

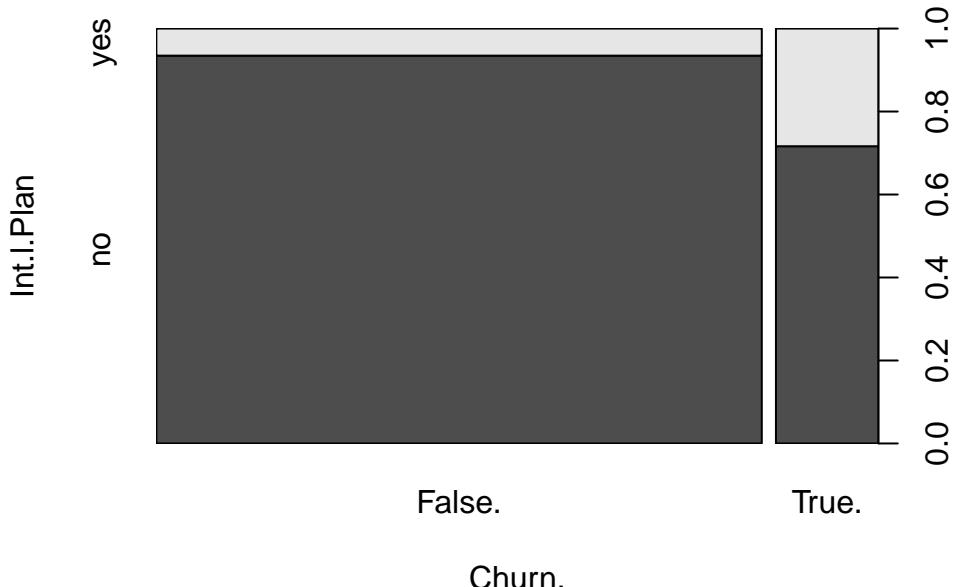
W tym etapie spróbujemy znaleźć przyczyny rezygnacji klientów z usług firmy poprzez analizę poszczególnych cech z podziałem klientów na lojalnych i tych, którzy zakończyli współpracę.

### 5.0.1 Jakościowe

- Zmienna: *Int.l.Plan* Plan międzynarodowy

```
dane.lojalni <- subset(dane, Churn=="False.") #zostali
dane.odeszli <- subset(dane, Churn=="True.") #odeszli
plot(Int.l.Plan~Churn., main="Rezygnacje klientów a plan międzynarodowy")
```

## Rezygnacje klientów a plan międzynarodowy

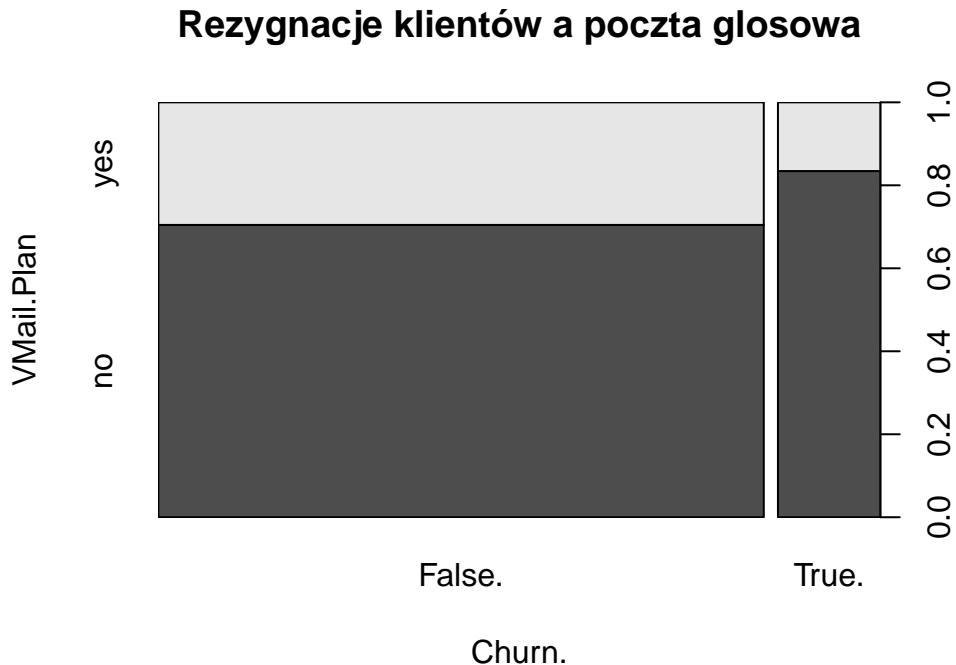


Można zauważać tu znaczne zróżnicowanie. Klienci, którzy używali tego pakietu częściej rezygnowali z usług firmy.

OPCJONALNIE: Spośród klientów, którzy odeszli, 30% z nich miało plan międzynarodowy, natomiast 95% klientów lojalnych nie posiada tego pakietu.

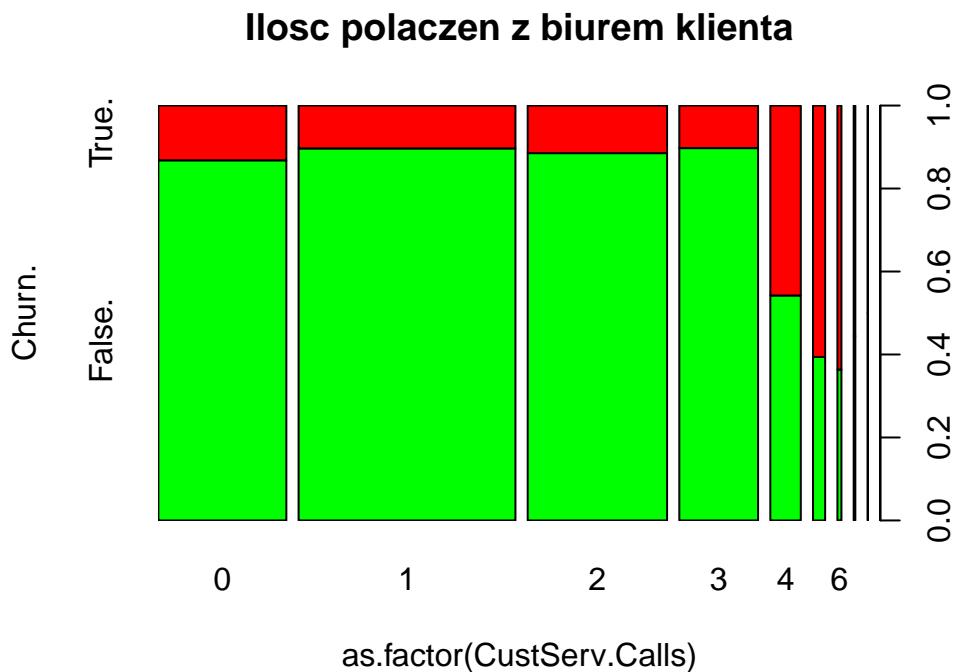
- Zmienna: *VoiceMail Plan* poczty głosowej

```
plot(VMail.Plan~Churn., main="Rezygnacje klientów a poczta głosowa")
```



Plan poczty głosowej nie miał większego wpływu na rezygnowanie z usług firmy.

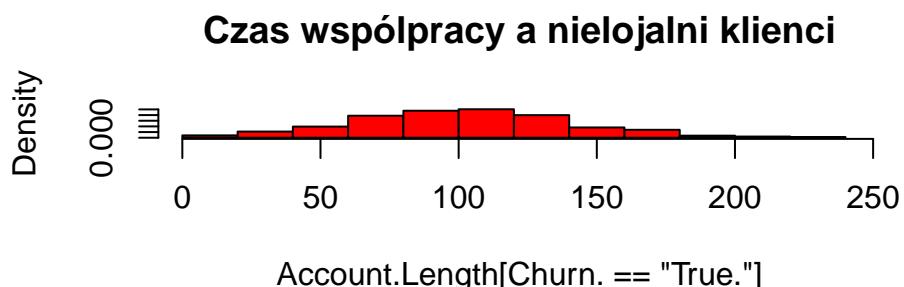
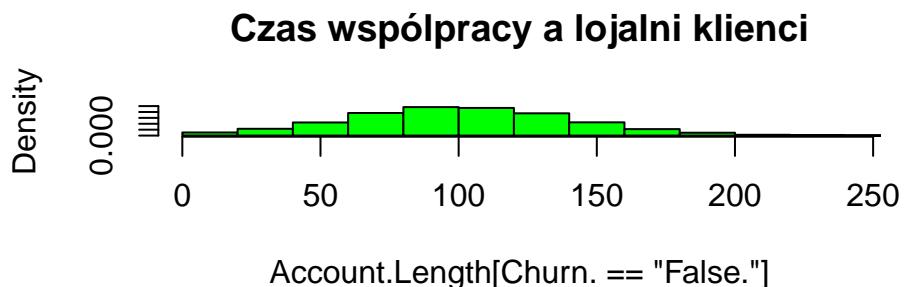
- Zmienna: *Cust.ser.calls*



Widać, że klienci, którzy zdecydowali się odejść częściej dzwoniли na infolinię. Dla 4 i więcej połączeń mamy jest to ponad 50%. Może to sugerować niezadowolenie z jakości świadczonej usługi.

### 5.0.2 Ilościowe

- Zmienna: *Account length*

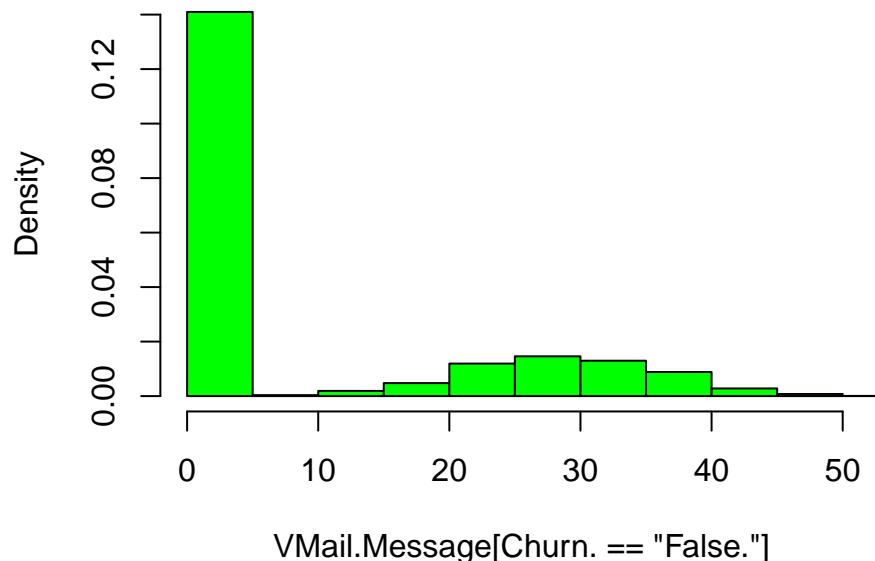


```
## Churn.: False.
##      min       Q1     median       mean       Q3      max   var
## 1.00000  73.00000 100.00000 100.79368 127.00000 243.00000 1590.60186
##      sd       IQR
## 39.88235  54.00000
## -----
## Churn.: True.
##      min       Q1     median       mean       Q3      max   var
## 1.00000  76.00000 103.00000 102.66460 127.00000 225.00000 1557.70885
##      sd       IQR
## 39.46782  51.00000
```

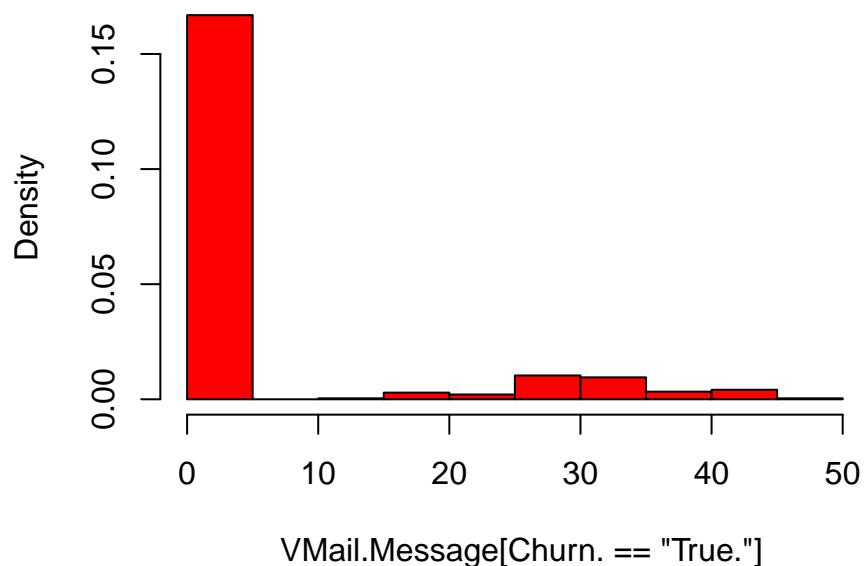
Wykresy są niemal identyczne, co powinno oznaczać, że długość współpracy z firmą nie miała wpływu na dalsze korzystanie z jej usług. Może to świadczyć o niewystarczającym nagradzaniu lojalności klientów.

- Zmienna: *Vmail Message*

### Ilosc wiadomosci glosowych a lojalni klienci



### Ilosc wiadomosci glosowych a nielojalni klienci



```
## Churn.: False.
##      min       Q1     median      mean       Q3      max    var
## 0.000000 0.000000 0.000000 8.604561 22.000000 51.000000 193.575058
##      sd       IQR
## 13.913125 22.000000
## -----
## Churn.: True.
```

```

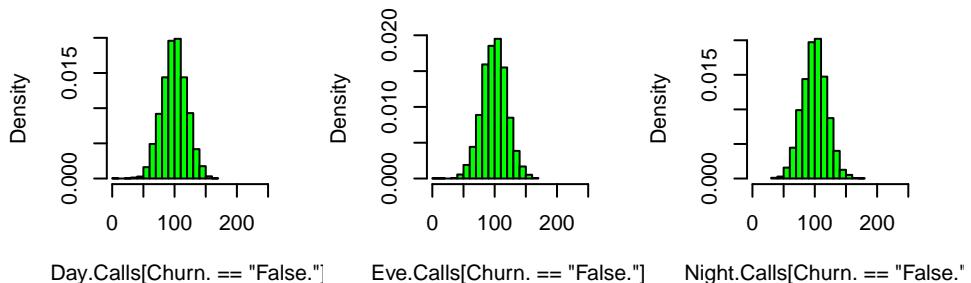
##      min       Q1     median      mean       Q3      max   var
## 0.000000 0.000000 0.000000 5.115942 0.000000 48.000000 140.662878
##      sd      IQR
## 11.860138 0.000000

```

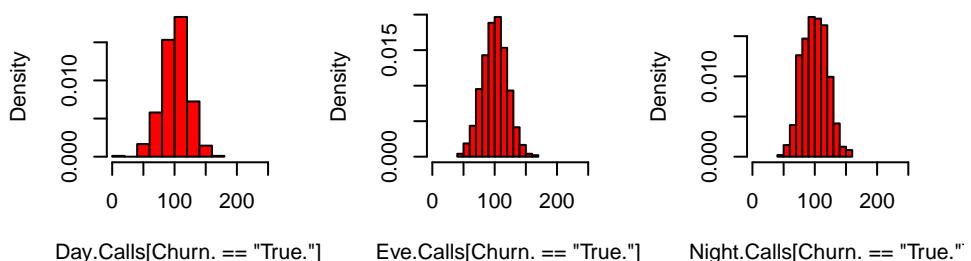
Poczta głosowa jest rzadko wybierana w obu grupach, więc nie powinna być przyczyną rezygnacji klienta z usług firmy.

- Zmienna: *Analiza rozmów w różnych porach dobowych*

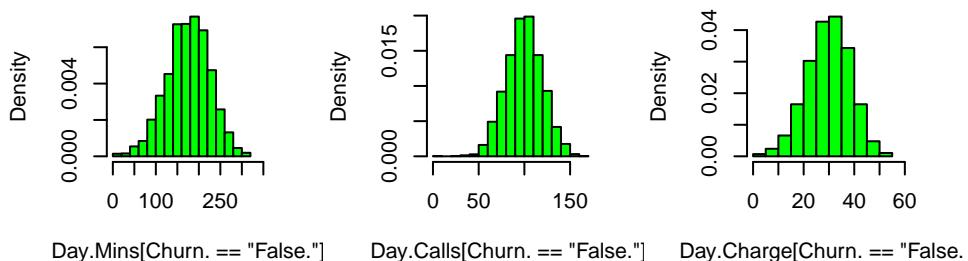
**ram of Day.Calls[Churn. == "False."]** **ram of Eve.Calls[Churn. == "False."]** **ram of Night.Calls[Churn. == "False."]**



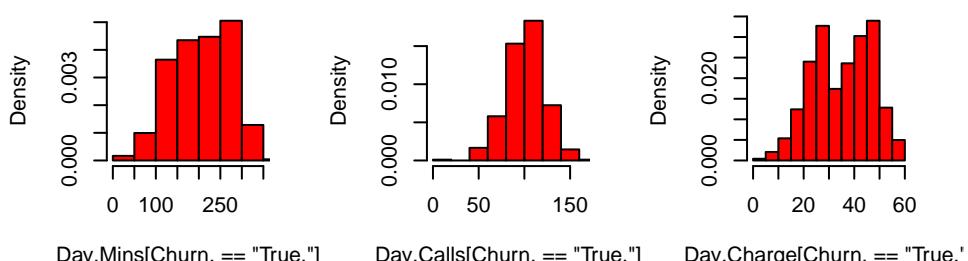
**ram of Day.Calls[Churn. == "True."]** **ram of Eve.Calls[Churn. == "True."]** **ram of Night.Calls[Churn. == "True."]**



ram of Day.Mins[Churn. == "False."], ram of Day.Calls[Churn. == "False."], ram of Day.Charge[Churn. == "False."]

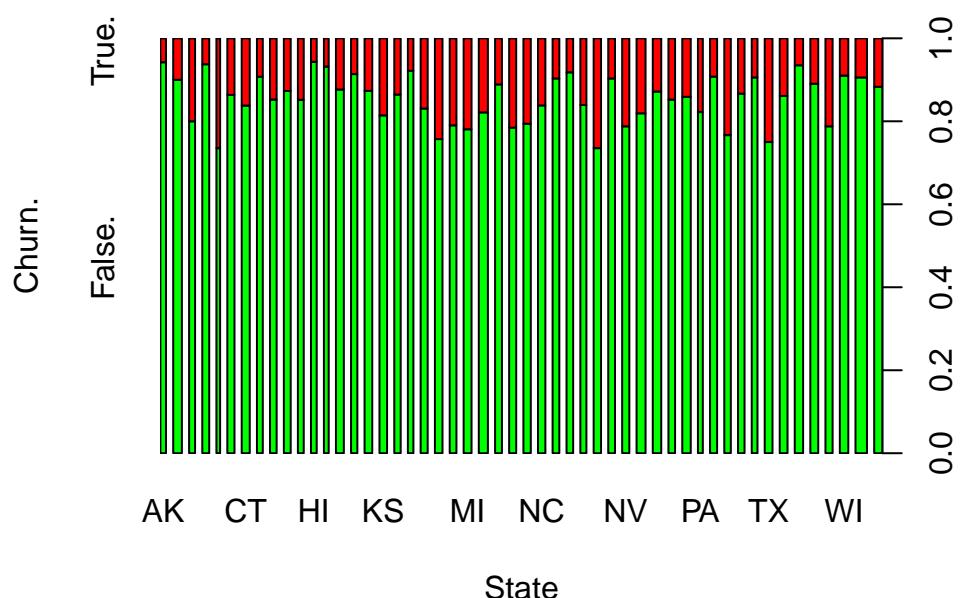


ram of Day.Mins[Churn. == "True."], ram of Day.Calls[Churn. == "True."], ram of Day.Charge[Churn. == "True."]



Tylko dzień, widać że lojalni klienci dużo więcej rozmawiają, w pełni korzystają z oferty, dlatego chcą ją kontynuować. Opłacalność idzie w parze z pełnym wykorzystaniem.

- Zmienna: *State*



Procent osób które odeszły różni się w zależności od stanu.

## 6 ETAP 4

Co wiemy o klientach sieci?

- Klienci korzystali głównie z połączeń
- Pora dnia nie ma wpływu na długość i ilość połączeń
- Mała popularność usługi poczty głosowej i planu międzynarodowego
- Spopularyzowanie usług poczty głosowej i planu międzynarodowego
- Klienci rezygnowali z usług firmy bez względu na długość trwania współpracy, być może lojalność powinna być lepiej nagradzana.
- Zbyt wielu klientów korzystających z planu międzynarodowego zrezygnowało z usług firmy. Powinny zostać wprowadzone niezbędne korekty bazujące na ofertach konkurencji.
- Niezadowolenie z usług obsługi klienta. Ponad 50% klientów, która wykonała 4 i więcej połączeń zdecydowała się zakończyć współpracę
- Klienci, którzy zrezygnowali z usług naszej firmy prowadzili dłuższe rozmowy od lojalnych klientów. Na tej podstawie można przypuszczać, że konkurencja oferuje lepsze ceny za dłuższe połączenia, warto więc poprawić cenę w naszej firmie