

Lista 1

Mikołaj Langner, Marcin Kostrzewa

31.3.2021

1 Wstęp

W niniejszym sprawozdaniu zajmować się będziemy danymi dotyczącymi klientów pewnej sieci telefonii komórkowej. Naszym zadaniem będzie odkrycie zależności między zmiennymi, które określą przyczyny rezygnacji klientów z oferty (churn analysis).

2 Wczytanie i identyfikacja danych

Wczytajmy dane z pliku i przeprowadźmy ich wstępna analizę i obróbkę:

```
df <- read.csv('churn.txt', stringsAsFactors = TRUE)
df$Area.Code = as.factor(df$Area.Code)
```

- poznajmy rozmiar naszych danych:

```
dim(df)
## [1] 3333 21
```

— jest 21 zmiennych i 3333 obserwacji;

- sprawdźmy ich typy:

Tabela 1: Typy zmiennych

	State	Account.Length	Area.Code	Phone	Int.l.Plan	VMail.Plan	VMail.Message	Day.Mins	Day.Calls	Day.Charge	Eve.Mins
Typ zmiennej	factor	integer	factor	factor	factor	integer	numeric	integer	numeric	numeric	numeric

	Eve.Calls	Eve.Charge	Night.Mins	Night.Calls	Night.Charge	Intl.Mins	Intl.Calls	Intl.Charge	CustServ.Calls	Churn.
Typ zmiennej	integer	numeric	numeric	integer	numeric	numeric	integer	numeric	integer	factor

- sprawdźmy czy pojawiają się wartości brakujące:

```
sum(is.na(df))
## [1] 0
```

— nie ma więc brakujących obserwacji;

- usuńmy dane pełniące rolę identyfikatora (numer telefonu):

```
df <- subset(df, select=-Phone)
```

3 Analiza zmiennych, szkanie zależności

3.1 Wskaźniki sumaryczne

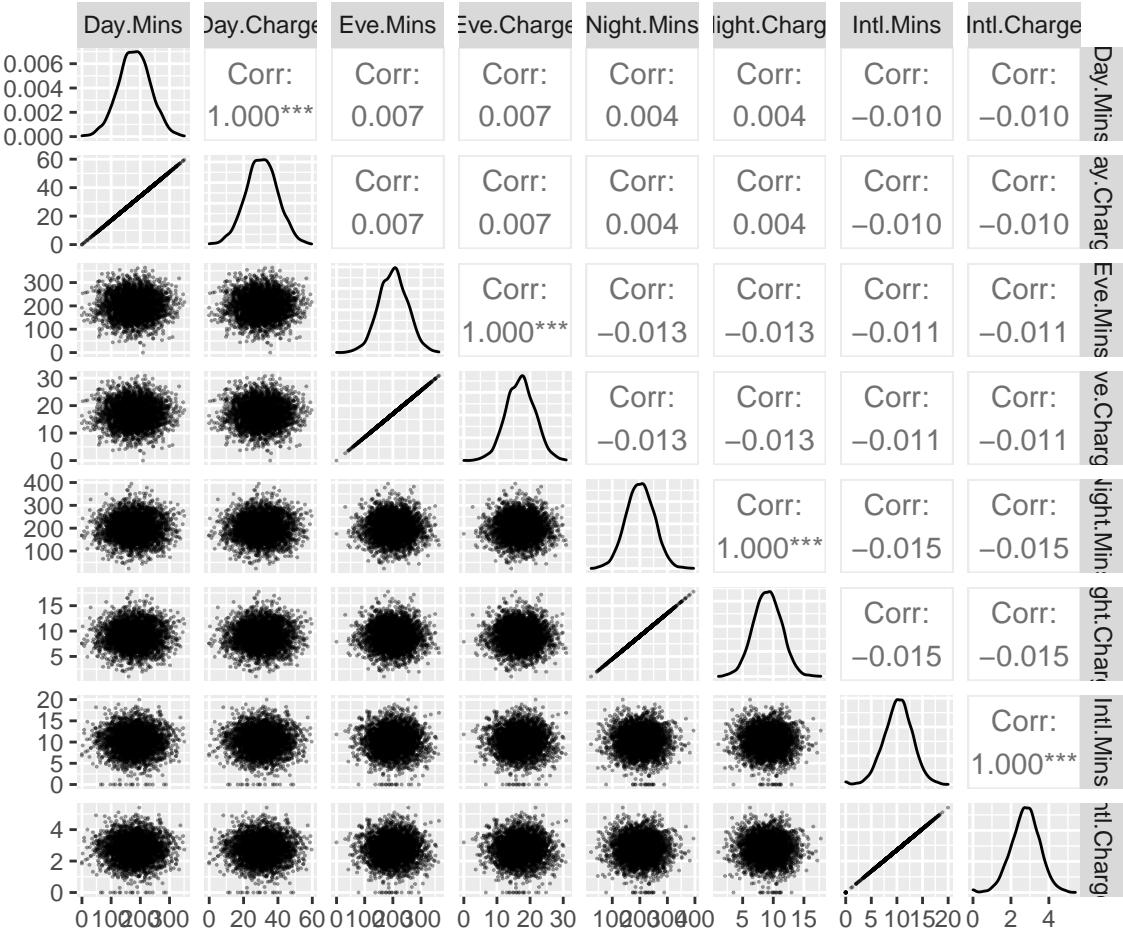
```
# wczytanie potrzebnych bibliotek
library(ggplot2)
library(ggmosaic)
library(GGally)
library(tidyr)
library(dplyr)
library(EnvStats)
library(DescTools)
library(moments)
```

Teraz podzielimy zmienne ze względu na ich typ (jakościowe — **factors**, ilościowe — **numerics**) oraz wykonamy kilka wykresów, które pomogą w zauważeniu pewnych zależności i wyborze najistotniejszych pod względem naszej analizy atrybutów.

```
factors <- subset(df, select=sapply(df, is.factor))
numerics <- subset(df, select=sapply(df, function(x) !is.factor(x)))
```

Sprawdźmy zależności pomiędzy zmiennymi ciągłymi.

```
continuous <- subset(numerics, select=sapply(numerics, function(x) !is.integer(x)))
ggpairs(continuous,
        lower=list(continuous=wrap("points", alpha=.4, size=.01)))
```



Możemy zauważyc, że zmienne z przyrostkami ‘.Mins’ oraz ‘.Charge’ są ze sobą idealnie skorelowane. Odrzućmy zatem od razu dane z przyrostkiem ‘.Charge’ dla ułatwienia dalszej analizy. Nie ma natomiast korelacji pomiędzy pozostałymi atrybutami.

```
numerics <- subset(numerics, select=-c(Day.Charge, Eve.Charge, Night.Charge, Intl.Charge))
```

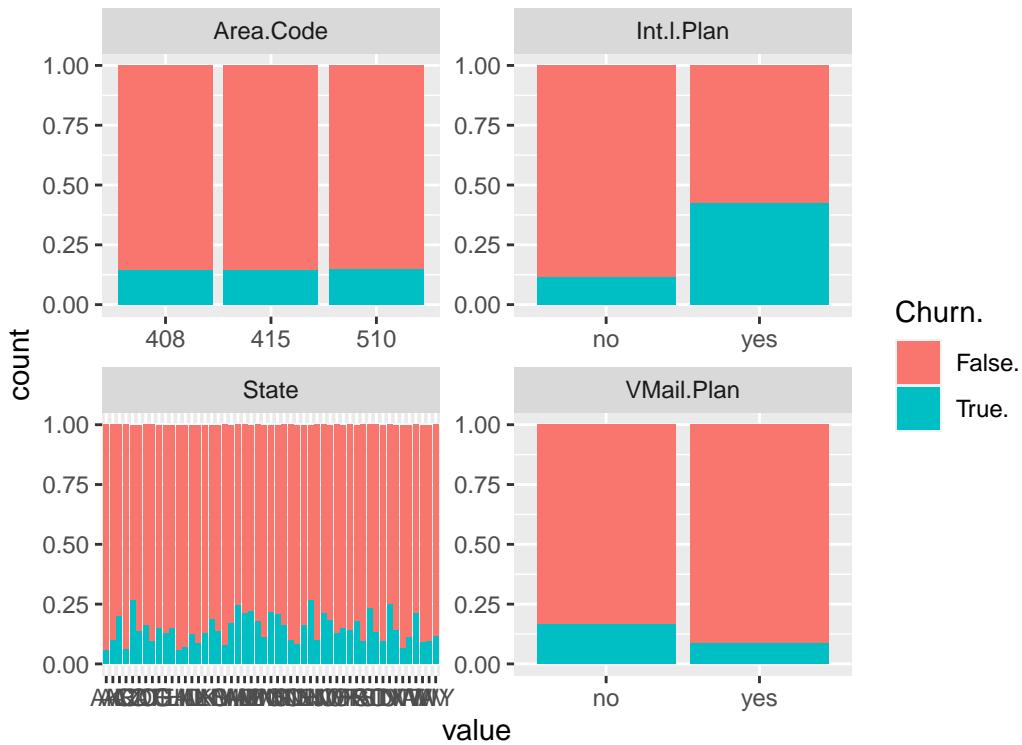
Wykonamy teraz wykresy zmiennych ilościowych, dzieląc klientów na dwie grupy:

- tych, którzy zrezygnowali — Churn. == TRUE ,
- tych, którzy pozostali lojalni — Churn. == FALSE.

```
numerics <- data.frame(numerics, Churn. = df$Churn.)
```

Poniżej wykresy słupkowe dla danych jakościowych.

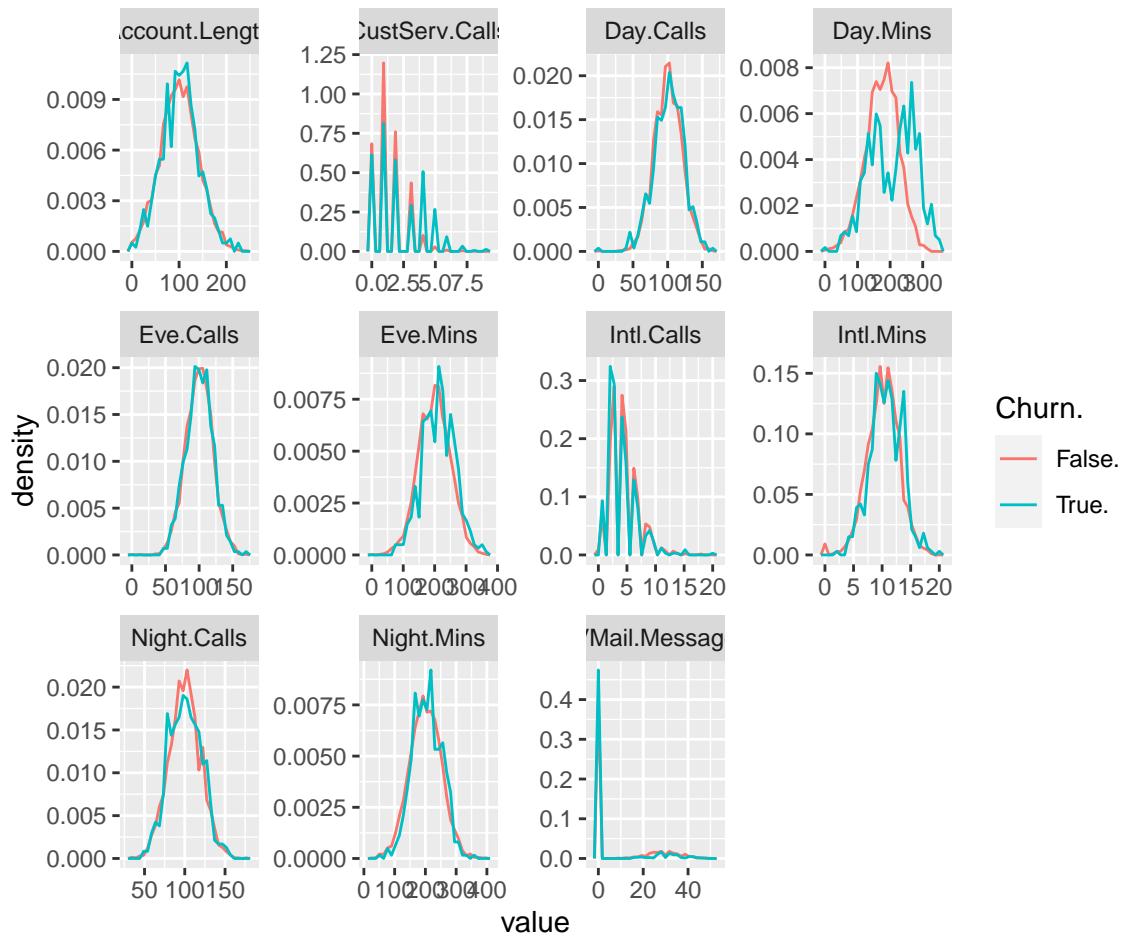
```
ggplot(gather(factors, "key", "value", -Churn.), aes(value, fill=Churn.)) +
  geom_bar(position="fill") +
  facet_wrap(~key, scales='free')
```



Możemy zauważać, że osoby, które posiadały plan międzynarodowy, jak i te, które nie posiadały planu skrzyki głosowej częściej rezygnowały z usług. Zmienne `Area.Code` i `State` nie pokazują żadnych istotnych różnic pomiędzy tymi dwoma grupami.

Poniżej znajdują się wykresy empirycznych gęstości.

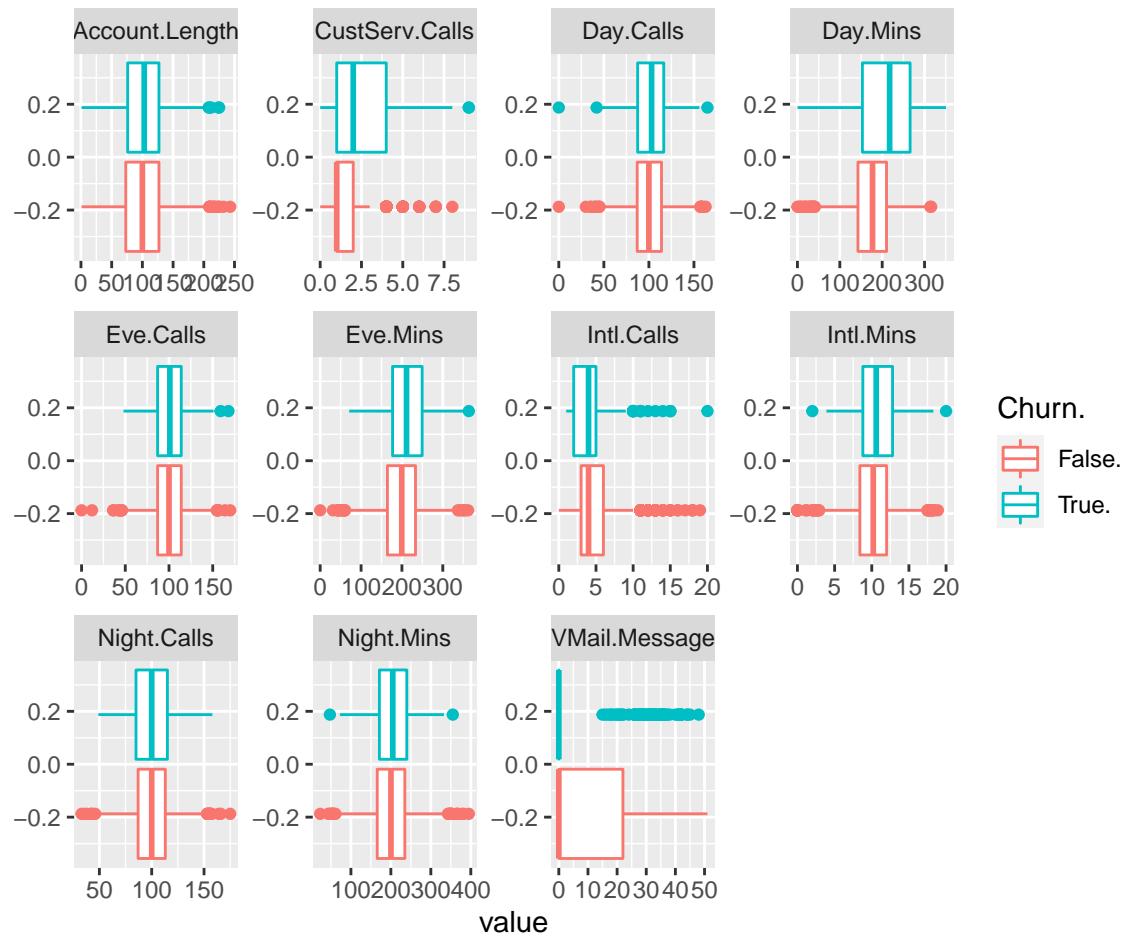
```
ggplot(gather(numerics, "key", "value", -Churn.), aes(x=value, color=Churn.)) +
  geom_freqpoly(aes(y=..density..), position="identity") +
  facet_wrap(~key, scales='free')
```



Widoczne gołym okiem różnice są zauważalne w przypadku zmiennych: `Day.Mins`, `Customer.Service.Calls`, `Eve.Calls`.

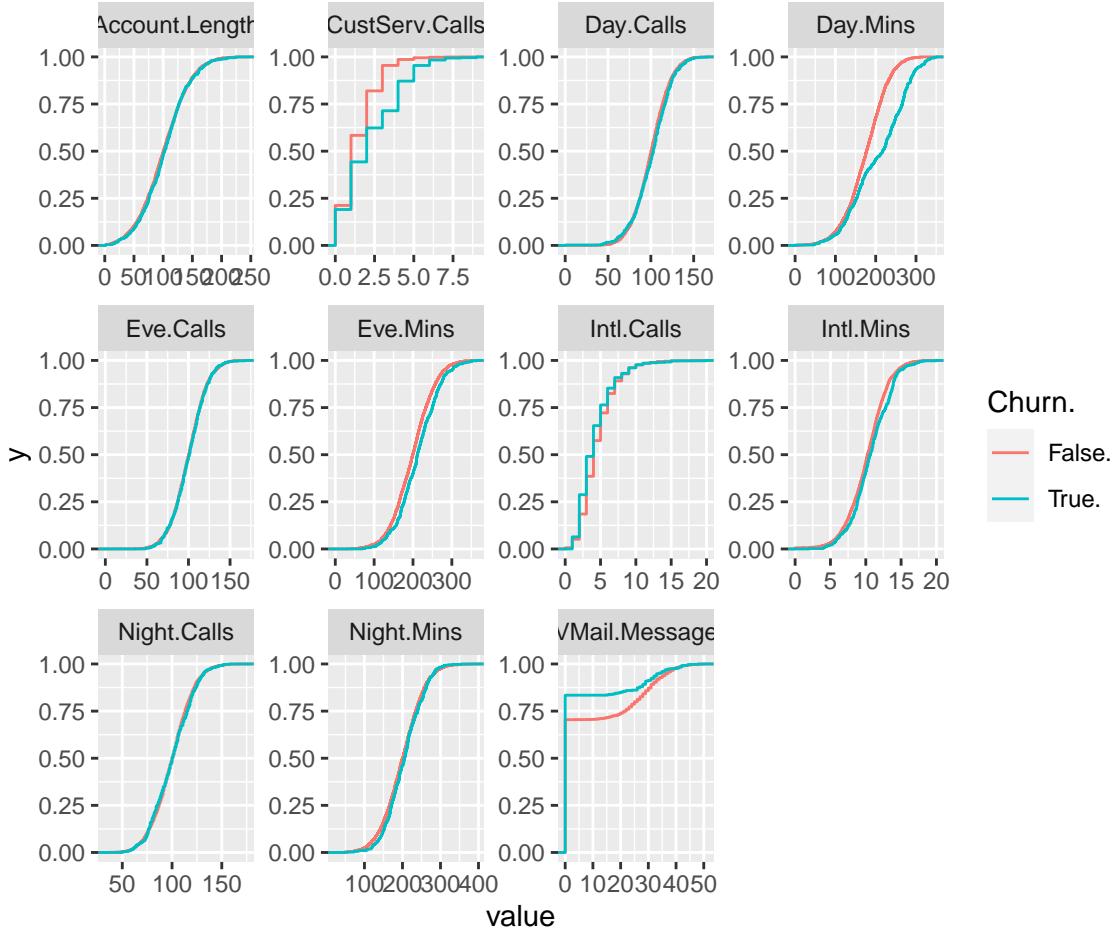
Poniżej generujemy wykresy pudełkowe.

```
ggplot(gather(numerics, "key", "value", -Churn.), aes(value, color=Churn.)) +
  geom_boxplot(aes(x=value)) +
  facet_wrap(~key, scales='free')
```



Ponownie, duże różnice uwidaczniają się dla zmiennych: Day.Mins, Customer.Service.Calls, Eve.Calls. Stworzymy również wykresy empirycznych dystrybuant.

```
ggplot(gather(numerics, "key", "value", -Churn.), aes(value, color=Churn.)) +
  stat_ecdf() +
  facet_wrap(~key, scales='free')
```



Coś tam, coś tam ...

By wykryć, dla których zmiennych następuje najważniejsza różnica pomiędzy klientami lojalnymi, a tymi którzy zrezygnowali z usług, posłużymy się również testem Kołmogorova-Smirnova. Poniżej znajduje się funkcja, która wyznacza wyniki tego testu dla naszych zmiennych.

```
churn.kstest <- function(feature) {
  yes <- subset(numerics, subset=Churn=="True.")
  no <- subset(numerics, subset=Churn=="False.")
  return(c(ks.test(yes[[feature]], no[[feature]])[c("statistic", "p.value")]))
}
```

Tabela poniżej zbiera wyniki przeprowadzonych testów statystycznych.

Tabela 2: Wyniki testu Kolmogorova-Smirnova

Zmienna	Account.Length	VMail.Message	Day.Mins	Day.Calls	Eve.Mins	Eve.Calls	Night.Mins	Night.Calls	Intl.Mins	Intl.Calls	CustServ.Calls
statistic	0.0389430	0.1298071	0.3172082	0.0556326	0.1166198	0.0192285	0.0551378	0.0401351	0.1007606	0.1054201	0.2404511
pvalue	0.5581609	0.0000018	0.0000000	0.1550801	0.0000264	0.9980118	0.1622501	0.5189055	0.0004560	0.0002062	0.0000000

Możemy zauważyć, że testy na największe różnice (duża wartość zmiennej **statistic**, mała zmiennej **pvalue**) wskazują w przypadku zmiennych: CustServ.Calls, Day.Mins, Eve.Mins,

Po dogłębnym przeanalizowaniu wykresów i wyników testów Kołmogorova-Smirnova, zauważmy, że istotne dla naszej analizy to następujące zmienne:

- ilościowych:

- CustServ.Calls,
- Day.Mins,
- Eve.Mins;
- jakościowych
 - Int.l.Plan,
 - VMail.Plan,
 - Churn.

4 Analiza wybranych zmiennych

Skupmy się jedynie na wybranych zmiennych:

```
important <- subset(df, select=c(CustServ.Calls, Day.Mins, Eve.Mins, Int.l.Plan,
                                 VMail.Plan, VMail.Message, Churn.))
```

Wyznaczmy dla nich wskaźniki sumaryczne.

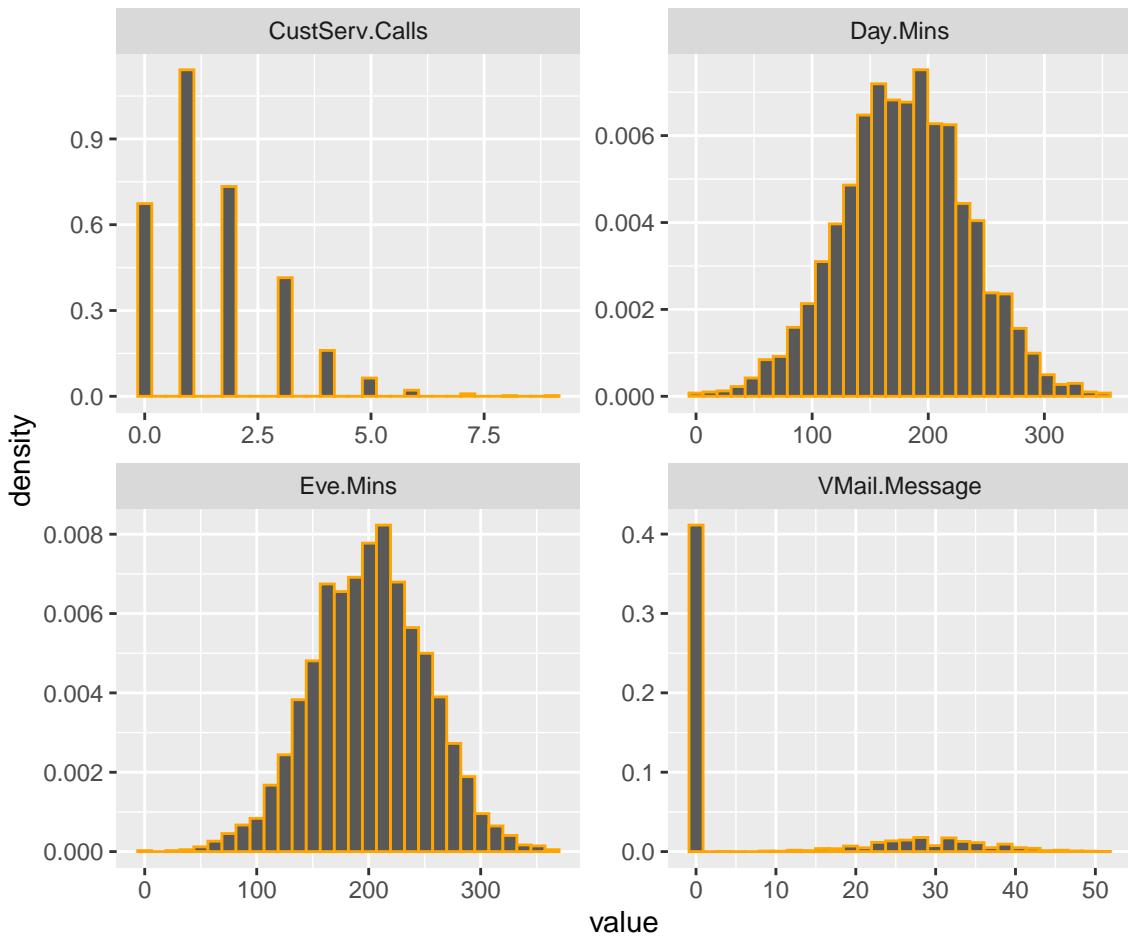
```
my_summary <- function(x) {
  statistics <- c(mean(x), quantile(x, 0.25), median(x), quantile(x, 0.75),
                 IQR(x), min(x), max(x), var(x), sd(x), sd(x) / mean(x),
                 kurtosis(x), skewness(x))
  names(statistics) <- c("Srednia", "Q1", "Mediana", "Q3", "IQR", "Min", "Max",
                         "Wariancja", "Odchylenie standardowe", "Wspolczynnik_zmienlosci",
                         "Kurtoza", "Skosnosc")
  return(statistics)
}
```

Tabela 3: Wskazniki sumaryczne dla wybranych zmiennych

	Srednia	Q1	Mediana	Q3	IQR	Min	Max	Wariancja	Odchylenie standardowe	Wspolczynnik_zmienlosci	Kurtoza	Skosnosc
CustServ.Calls	1.562856	1.0	1.0	2.0	1.0	0	9.0	1.730517	1.315491	0.8417223	4.726519	1.0908683
Day.Mins	179.775097	143.7	179.4	216.4	72.7	0	350.8	2966.696486	54.467389	0.3029752	2.978290	-0.0290640
Eve.Mins	200.980348	166.6	201.4	235.3	68.7	0	363.7	2571.894016	50.713844	0.2523324	3.023792	-0.0238667
VMail.Message	8.099010	0.0	0.0	20.0	20.0	0	51.0	187.371347	13.688365	1.6901282	2.947148	1.2642543

Przedstawimy również wartości tych zmiennych na histogramach.

```
subset = subset(important, select=-c(Churn., Int.l.Plan, VMail.Plan))
ggplot(gather(subset, 'key', 'value'), aes(x=value)) +
  geom_histogram(aes(y=..density..), position="identity", color="orange") +
  facet_wrap(~key, scales='free')
```



Day.Mins i Eve.Mins mają rozkład symetryczny, natomiast pozostałe CustServ.Calls i VMail.Message mają rozkład prawostronnie skośny. Dodatkowo, te dwie zmienne charakteryzują się one dużą zmiennością.

Teraz przyjrzyjmy się bliżej wybranym zmiennym jakościowym.

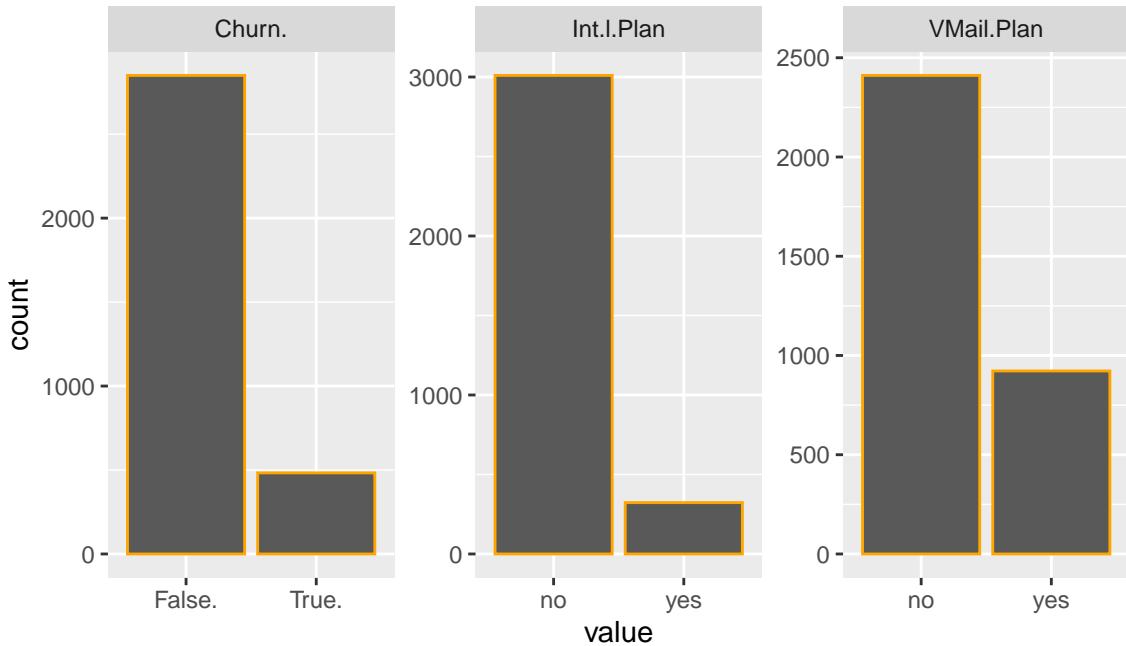
Churn	Count
False.	2850
True.	483

Int.l.Plan	Count
no	3010
yes	323

VMail.Plan	Count
no	2411
yes	922

Stworzymy dla tych zmiennych wykresy słupkowe.

```
ggplot(gather(important, "key", "value", -c(CustServ.Calls, Day.Mins, Eve.Mins, VMail.Message)), aes(value)) +
  geom_bar(position="dodge", color='orange') +
  facet_wrap(~key, scales='free')
```



Łatwo stwierdzić, że większość klientów była lojalna ($\approx 86\%$), nie miała wykupionego planu międzynarodowego ($\approx 90\%$) oraz nie miało dostępu do planu poczty głosowej ($\approx 72\%$).

4.1 Analiza wybranych zmiennych z podziałem na grupy

Poniższe tabele zawierają informacje o wartościach wskaźników sumarycznych dla zmiennych ilościowych, tym razem uwzględniają one podział klientów na grupy.

Tabela 4: Day.Mins

	Srednia	Q1	Median	Q3	IQR	Min	Max	Wariancja	Odchylenie standardowe	Wspolczynnik zmienności	Kurtoza	Skosnosc
False.	175.18	142.83	177.2	210.30	67.47	0	315.6	2518.2	50.18	0.29	2.99	-0.23
True.	206.91	153.25	217.6	265.95	112.70	0	350.8	4760.7	69.00	0.33	2.19	-0.20

Tabela 5: Eve.Mins

	Srednia	Q1	Median	Q3	IQR	Min	Max	Wariancja	Odchylenie standardowe	Wspolczynnik zmienności	Kurtoza	Skosnosc
False.	199.04	164.5	199.6	233.20	68.70	0.0	361.8	2529.30	50.29	0.25	3.03	-0.04
True.	212.41	177.1	211.3	249.45	72.35	70.9	363.7	2675.88	51.73	0.24	2.90	0.03

Tabela 6: CustServ.Calls

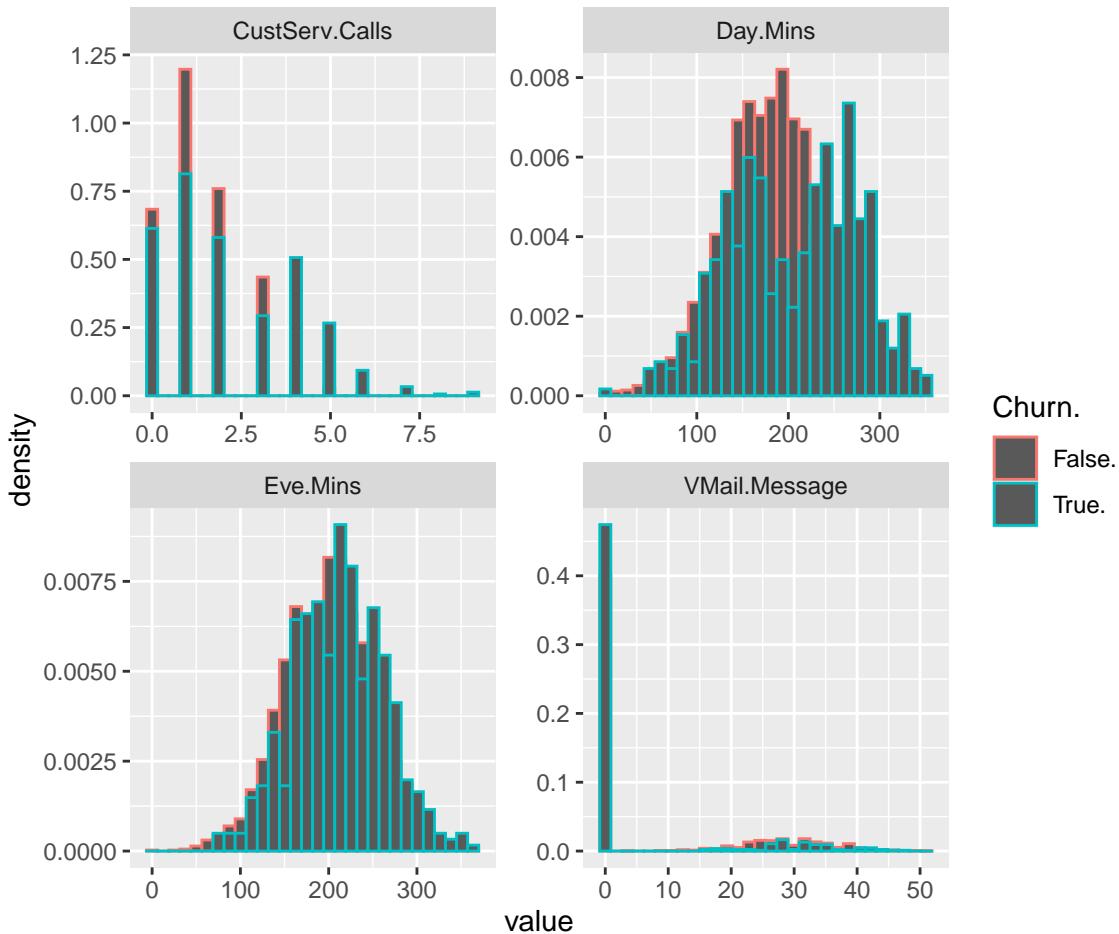
	Srednia	Q1	Median	Q3	IQR	Min	Max	Wariancja	Odchylenie standardowe	Wspolczynnik zmienności	Kurtoza	Skosnosc
False.	1.45	1	1	2	1	0	8	1.35	1.16	0.80	4.21	0.89
True.	2.23	1	2	4	3	0	9	3.43	1.85	0.83	2.89	0.70

Tabela 7: VMail.Message

	Srednia	Q1	Median	Q3	IQR	Min	Max	Wariancja	Odchylenie standardowe	Wspolczynnik zmienności	Kurtoza	Skosnosc
False.	8.60	0	0	22	22	0	51	193.58	13.91	1.62	2.71	1.17
True.	5.12	0	0	0	0	0	48	140.66	11.86	2.32	5.52	2.03

Podobnie jak powyżej, przedstawimy wartości zmiennych, korzystając z histogramów. Tym razem uwzględniamy podział na grupy.

```
subset = subset(important, select=-c(Int.l.Plan, VMail.Plan))
ggplot(gather(subset, 'key', 'value', -Churn.), aes(x=value, color=Churn.)) +
  geom_histogram(aes(y=..density..), position="identity") +
  facet_wrap(~key, scales='free')
```



Zauważamy, że

Przyjrzymy się teraz zmiennym jakościowym po podziale na grupy.

Tabela 8: Int.l.Plan

	False.	True.
no	0.89	0.11
yes	0.58	0.42

Wykonamy także wykresy rozrzutu, przedstawiające zależności między zmiennymi ilościowymi.

```
subset = subset(important, select=-c(Int.l.Plan, VMail.Plan))
subset %>% ggpairs(., mapping = ggplot2::aes(color=Churn.),
columns=1:4,
```

Tabela 9: VMail.Plan

	False.	True.
no	0.83	0.17
yes	0.91	0.09

```
lower=list(continuous=wrap("points", alpha=.4, size=.01)),
upper=list(continuous="blank"),
diag=list(continuous="blank"))
```



5 Podsumowanie

Najważniejsze wnioski z naszej analizy to:

- jakiś wniosek