

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)
# Area.Code powinnien być zmiennej jako ciow
```

- poznajmy rozmiar naszych danych:

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

— są więc 21 zmienne i 3333 obserwacji;

- sprawdźmy ich typ:

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

Tabela 1: Hello

Zmienna ‘Churn.’ mówi o tym, czy dany klient zrezygnował z oferty.

- sprawdźmy czy pojawiają się wartości brakujące:

```
sum(sapply(df, function(x) sum(is.na(x))))  
## [1] 0
```

- usuńmy dane pełniące rolę indentyfikatora:

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

3 Wybór zmiennych

Teraz podzielimy zmmienne ze względu na ich typ oraz wykonamy kilka wykresów, które pomogą w zauważeniu pewnych zależności i wyborze najistotniejszych pod względem naszej analizy atrybutów.

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

```

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

```

```

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

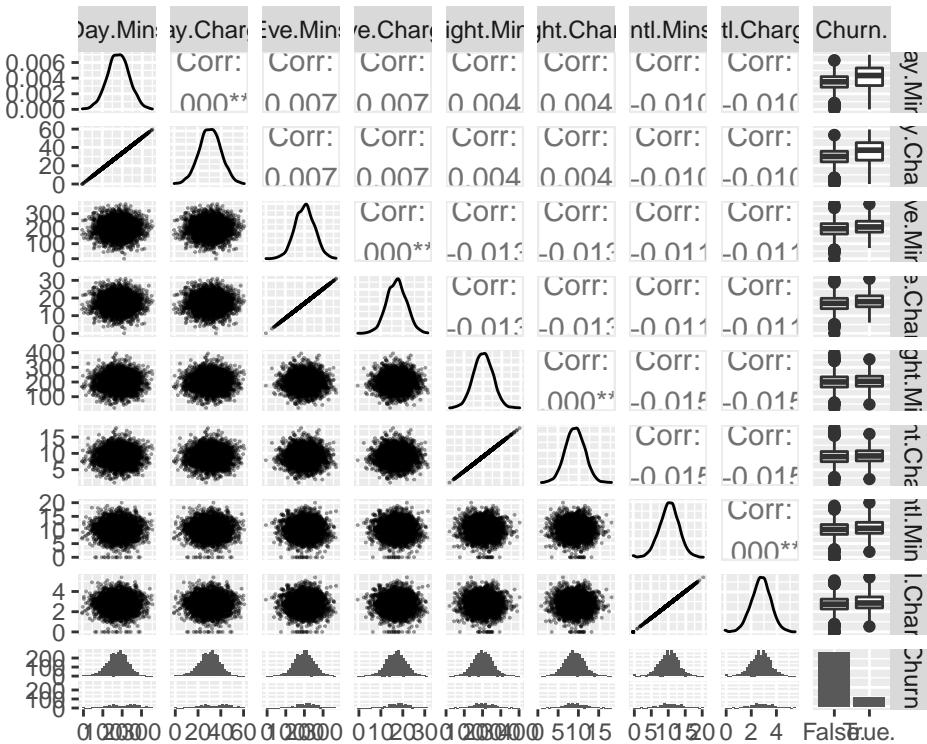
```

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.

```

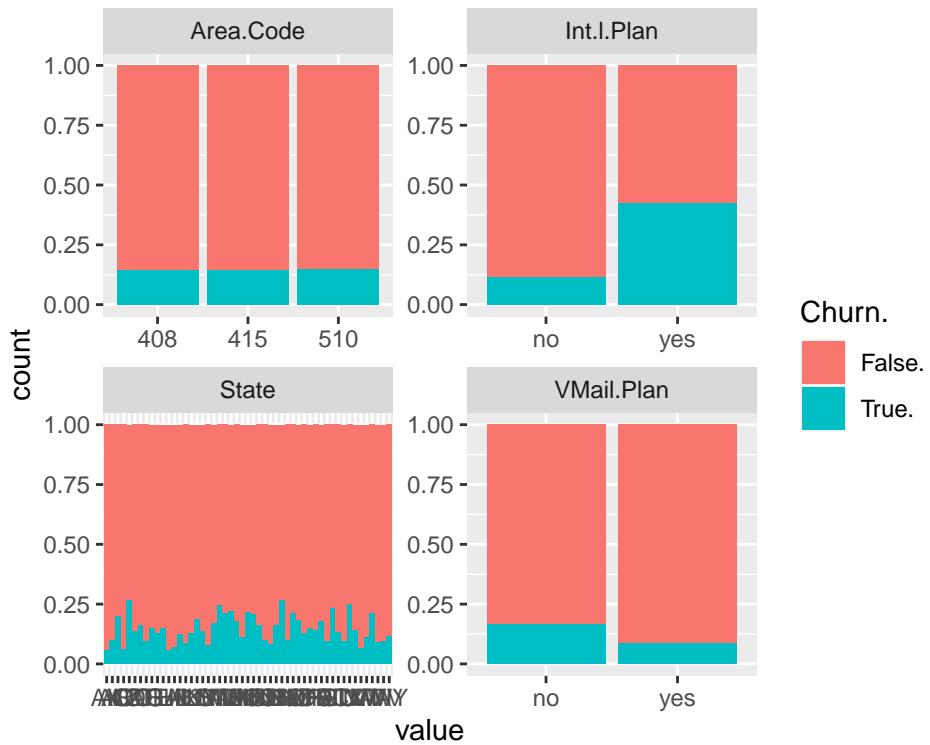
numerics <- subset(numerics, select=-c(Day.Charge, Eve.Charge, Night.Charge, Intl.Charge))

```

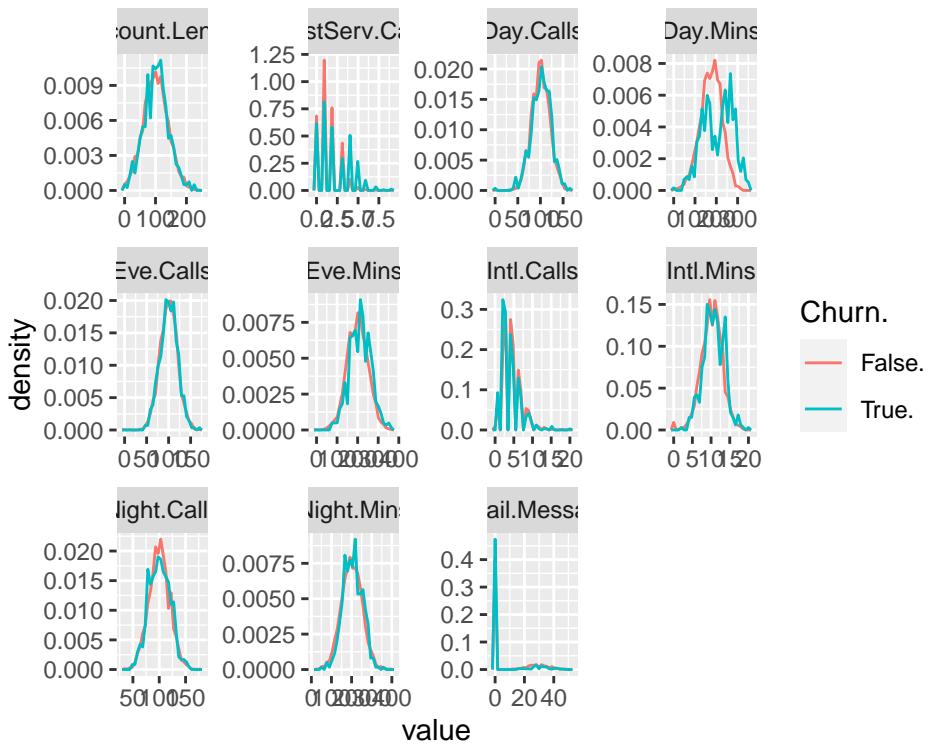
```

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

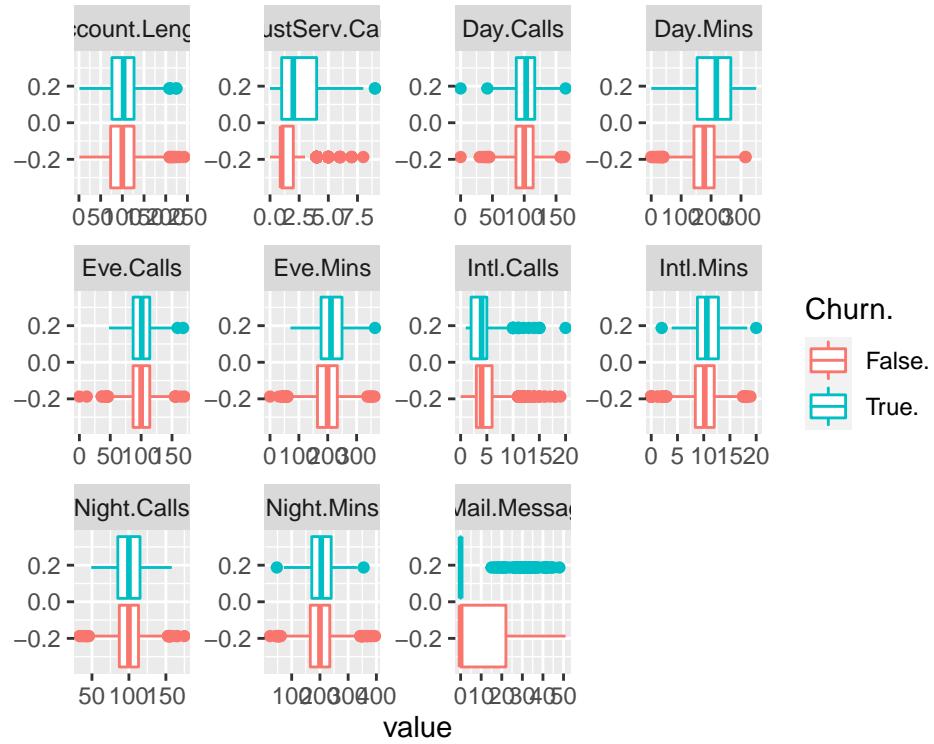
```



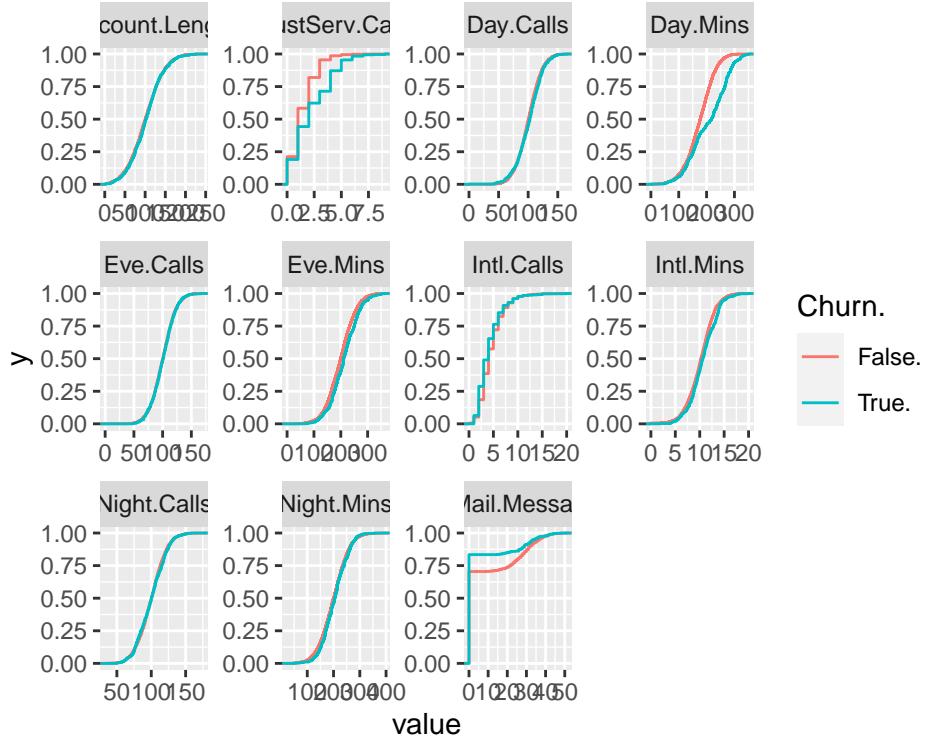
```
ggplot(gather(numerics, "key", "value", -Churn.), aes(x=value, color=Churn.)) +
  geom_freqpoly(aes(y=..density..), position="identity") +
  facet_wrap(~key, scales='free')
```



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



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



```

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 2: Wyniki testu Kolmogorova-Smirnowa

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

Po przeanalizowaniu wykresów, decydujemy się na dalszą analizę następujących zmiennych:

- 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, Churn.))
```

Wyznaczmy dla nich wskaźniki sumaryczne.

```
# w<U+0142>asna funkcja zwracaj<U+0105>ca wska<U+017A>niki 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), kurtosis(x), skewness(x))
  names(statistics) <- c("Srednia", "Q1", "Mediana", "Q3", "IQR", "Min", "Max",
                          "Wariancja", "Odchylenie standardowe", "Kurtoza", "Skosnosc")
  return(statistics)
}
```

Tabela 3: Wskaźniki sumaryczne dla wybranych zmiennych

	Srednia	Q1	Mediana	Q3	IQR	Min	Max	Wariancja	Odchylenie standardowe	Kurtoza	Skosnosc
CustServ.Calls	1.562856	1.0	1.0	2.0	1.0	0	9.0	1.730517	1.315491	4.726519	1.0908683
Day.Mins	179.775097	143.7	179.4	216.4	72.7	0	350.8	2966.696486	54.467389	2.978290	-0.0290640
Eve.Mins	200.980348	166.6	201.4	235.3	68.7	0	363.7	2571.894016	50.713844	3.023792	-0.0238667

5 Etap III

6 Etap IV