

List 1

Mikołaj Langner, Marcin Kostrzewa

31.3.2021

1 Wstęp

W niniejszym sprawozdaniu zajmować się będziemy danymi sieci telefonii komórkowej ze względu na rezygnacje 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
```

— 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 identyfikatora:

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

3 Wybór zmiennych

Teraz podzielimy zmmienne ze względu na typ oraz wykonamy kilka wykresów, które pomogą w zauważeniu pewnych zależności i wyborze najistotniejszych zmiennych.

```
library(ggplot2)  
library(ggmosaic)  
library(GGally)  
library(tidyr)  
library(dplyr)  
library(EnvStats)  
library(DescTools)
```

```

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.)

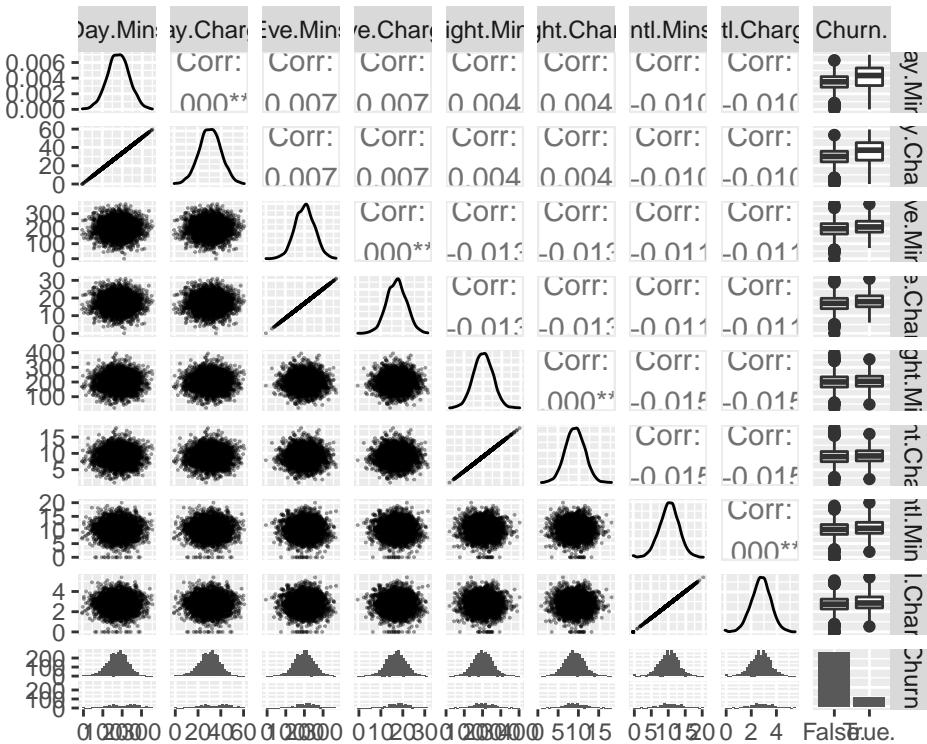
```

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.

```

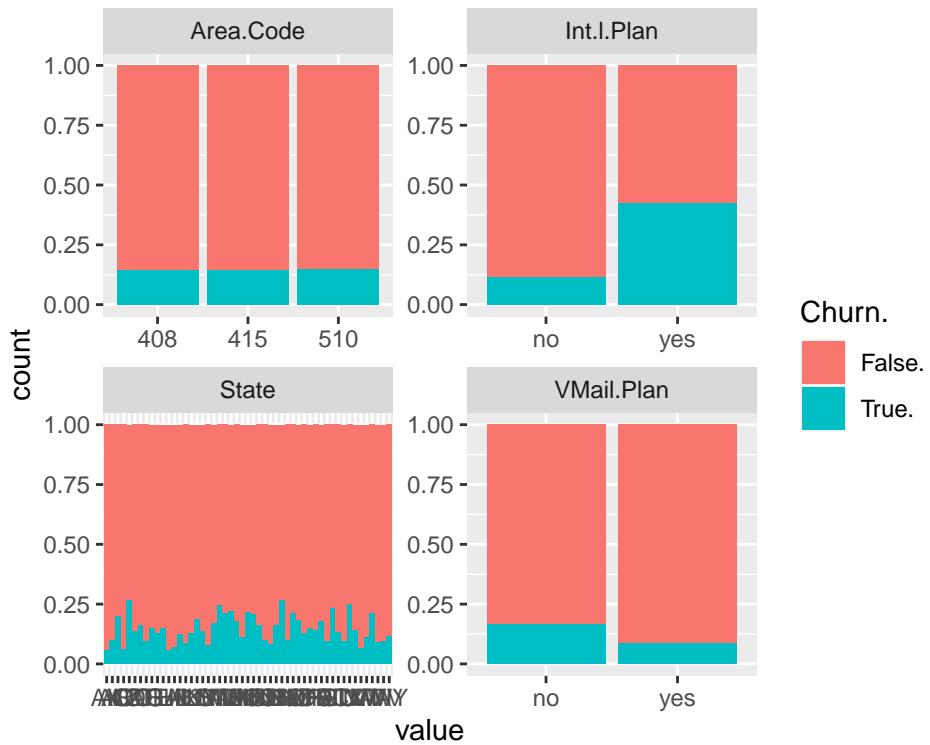
numerics <- subset(numerics, select=-c(Day.Charge, Eve.Charge, Night.Charge, Intl.Charge))

```

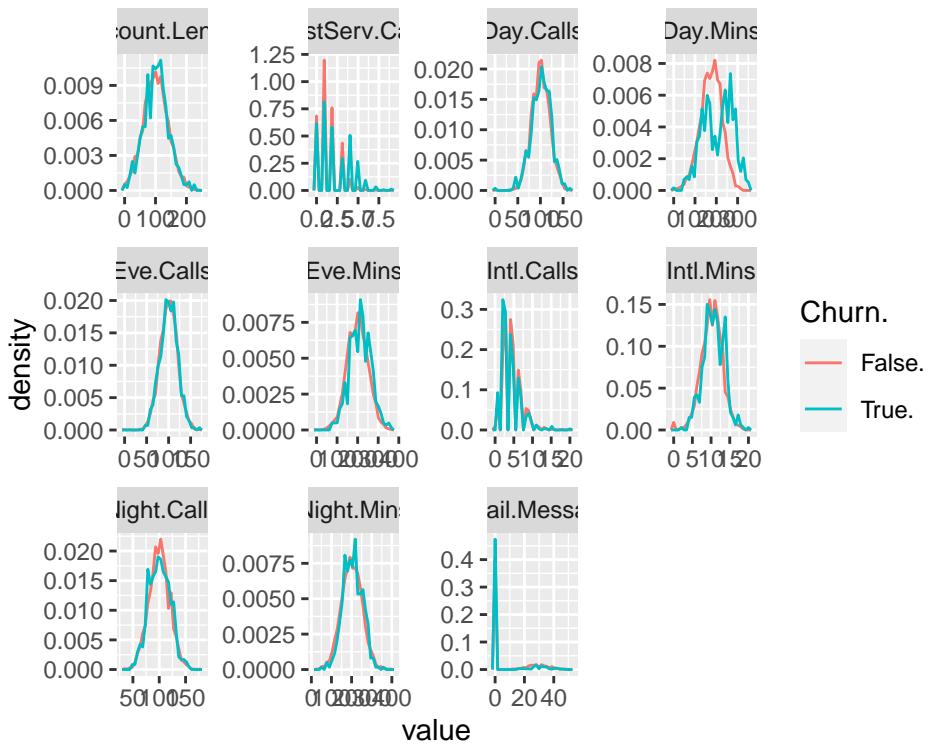
```

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

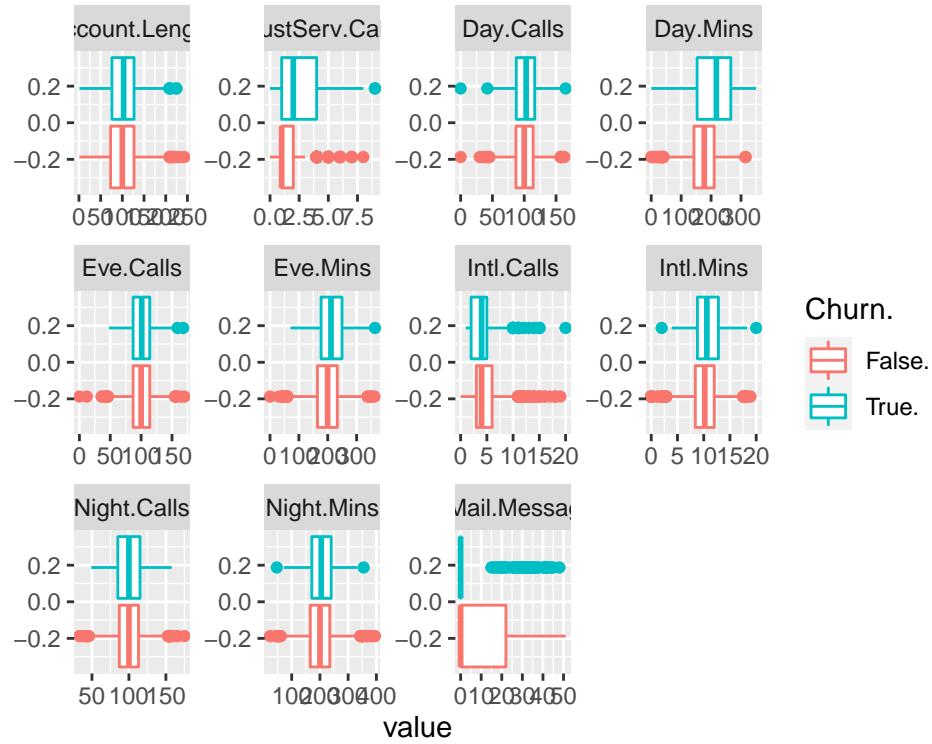
```



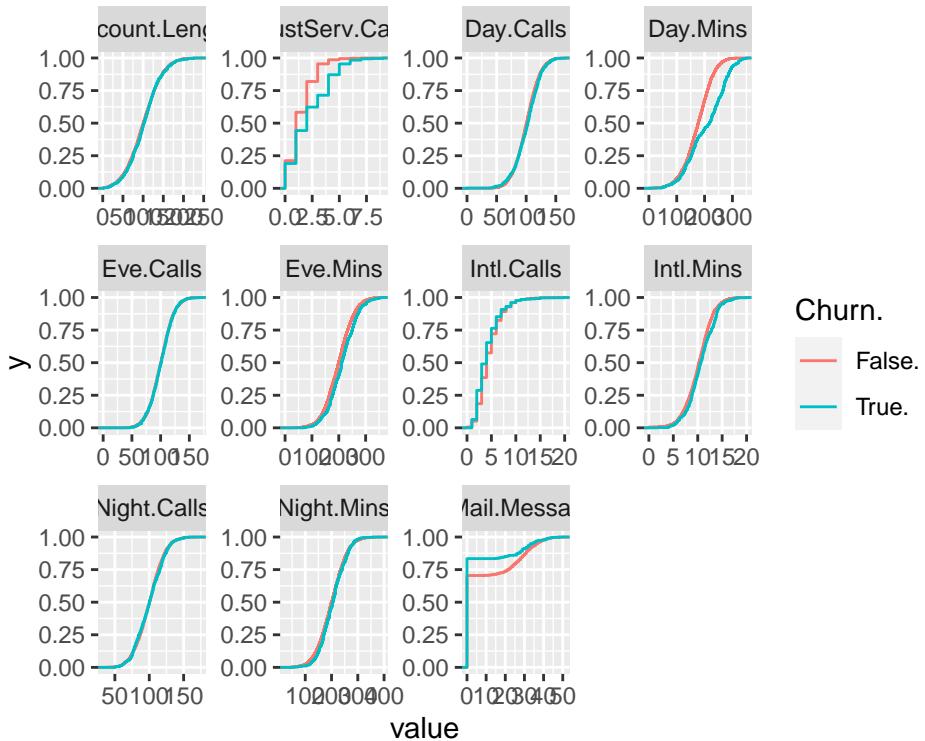
```
ggplot(gather(numerics, "key", "value", -Churn.), aes(x=value, color=Churn.)) +
  geom_freqpoly(aes(y=..density..), position="identity") +
  facet_wrap(~key, scales='free')
```



```
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')
```



```

## Warning in ks.test(yes[[feature]], no[[feature]]): p-value will be approximate in the presence
of ties

## [[1]]
## [[1]][[1]]
## [1] "Account.Length"
##
## [[1]]$statistic
##           D
## 0.03894301
##
## [[1]]$p.value
## [1] 0.5581609
##
##
## [[2]]
## [[2]][[1]]
## [1] "VMail.Message"
##
## [[2]]$statistic
##           D
## 0.1298071
##
## [[2]]$p.value
## [1] 1.804775e-06
##
##
## [[3]]
## [[3]][[1]]
## [1] "Day.Mins"
##
## [[3]]$statistic
##           D
## 0.3172082
##
## [[3]]$p.value
## [1] 0
##
##
## [[4]]
## [[4]][[1]]
## [1] "Day.Calls"
##
## [[4]]$statistic
##           D
## 0.05563256
##
## [[4]]$p.value
## [1] 0.1550801
##
##
## [[5]]
## [[5]][[1]]
## [1] "Eve.Mins"

```

```

## 
## [[5]]$statistic
##           D
## 0.1166198
## 
## [[5]]$p.value
## [1] 2.643629e-05
## 
## 
## [[6]]
## [[6]][[1]]
## [1] "Eve.Calls"
## 
## [[6]]$statistic
##           D
## 0.01922851
## 
## [[6]]$p.value
## [1] 0.9980118
## 
## 
## [[7]]
## [[7]][[1]]
## [1] "Night.Mins"
## 
## [[7]]$statistic
##           D
## 0.05513784
## 
## [[7]]$p.value
## [1] 0.1622501
## 
## 
## [[8]]
## [[8]][[1]]
## [1] "Night.Calls"
## 
## [[8]]$statistic
##           D
## 0.04013512
## 
## [[8]]$p.value
## [1] 0.5189055
## 
## 
## [[9]]
## [[9]][[1]]
## [1] "Intl.Mins"
## 
## [[9]]$statistic
##           D
## 0.1007606
## 
## [[9]]$p.value

```

```

## [1] 0.0004559583
##
##
## [[10]]
## [[10]][[1]]
## [1] "Intl.Calls"
##
## [[10]]$statistic
##           D
## 0.1054201
##
## [[10]]$p.value
## [1] 0.000206202
##
##
## [[11]]
## [[11]][[1]]
## [1] "CustServ.Calls"
##
## [[11]]$statistic
##           D
## 0.2404511
##
## [[11]]$p.value
## [1] 0

```

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 Etap II

5 Etap III

6 Etap IV