B.Sc in Computing

ENTERPRISE DATABASE TECHNOLOGIES CA 1

Daniel Mateus Pires, x00132886

https://github.com/dmateusp/R_CA1

Released: 24th February 2017

TABLE OF CONTENTS

Section 1 - Data Understanding and Data Exploration

1. DATA PRE-PROCESSING

First operation carried out was getting the number of null or empty string values per column (appendix 1.a).

We can see that only very few values are empty (single digit); and some columns have no value missing which indicates that our data is of good quality in terms of completeness.

The next step was to replace null numeric values with the median of the respective columns (appendix 1.b).

Then I created a function to get the mode by gender of a given column and replaced missing values in categorical columns by their respective modes (appendix 1.c).

The mode of PHONE_PLAN is International for both Males and Females, there are also more Males churners than Females churners (plotting the influence of gender could be interesting).

2. DISCRETIZING INCOME

One should be careful with the inclusion / exclusion of lower and upper ranges, the Low-Income category as an example end before 38,000 (37,999 is the last value), this is taken in account in the code.

3. FINDING INFORMATION

For 3.c (appendix 3.c), the get_mode function (created earlier), was used along with the summary function.

See appendix 3c, 3.d, 3.e, 3.f, 3.g for this question

Predictor	**A**	**B**	**C**	**D**	**E**	**F**	**G**
AREA_CODE	Nominal	0	**Mode**: 10040	X	X	X	X
CUST_MOS	Numeric	3/2071	**Min**: 1, **Median**: 11, **Mean**: 16.05, **Max**: 50	Most customers seem to stay during 5 to 15 months	It seems that in the first months, the customer has more chances to Churn, Around 10 months	Skewness: 1.131244 Positively skewed (skewed to the right)	One outlier found in the box plot

					the customer will churn as well (end of one year contract?)		
LONGDIST_FLAG	Nominal	0	**Mode**: 1	X	X	X	X
CALLWAITING_FLAG	Nominal	0	**Mode**: 0	X	X	X	X
NUM_LINES	Numeric	0	**Min**: 1, **Median**: 1, **Mean**: 1.391, **Max**: 3	Most users seem to have 1 number only and only very few have 3 numbers	The number of lines do not seem to bring much insight	Skewness: 2.057503 Strongly positively skewed (skewed to the right)	X
VOICEMAIL_FLAG	Nominal	0	**Mode**: 1	X	X	X	X
MOBILE_PLAN	Nominal	0	**Mode**: 0	X	X	X	X
CONVERGENT_BILLING	Nominal	0	**Mode**: No	X	X	X	X
GENDER	Nominal	0	**Mode**: M	X	Does not bring insight	X	X
INCOME	Ordinal	0	**Mode**: Medium Income	Most users have medium income, twice as many users have high income compared to low income users	Users with low income or high income tend not to churn while medium incomes tend to churn more	X	X
PHONE_PLAN	Ordinal	4/2071	Mode: **International**	Only few users choose the Euro-zone, most of the users opt for the International and	Users having the Euro-Zone or the International phone plan tend to churn while	X	X

				National plans	users with a National or Promo_plan tend to churn less		
EDUCATION	Nominal	8/2071	Mode: **Post Primary**	Post-Primary is the dominant group, the number of High school and Primary school are very low	Primary are churners, Masters tend to churn, PhD tend not to churn	X	X
TOT_MINUTES_USAGE	Numeric	4/2071	**Min**: 0, **Median**: 264, **Mean**: 2036, **Max**: 36237	Clear majority of users use less than 2500 mins	Do not seem to bring insight	Skewness: 1.088757 Positively skewed (skewed to the right)	Graphically, from the boxplot we can see that the data contains a lot of outliers that will need to be cleaned out

4. FINDING OUTLIERS MATHEMATICALLY

I chose TOT_MINUTES_USAGE since its box graph seems to indicate a lot of outliers.

I found 176 outliers using the IQR method while the Z-standardisation method found 69 outliers (appendix 4).

5. SKEWNESS IN TOT_MINUTES_USAGE

My approach was, to first get the skewness value of TOT_MINUTES_USAGE before transformation: 1.088757, (appendix 5.), this positive skewness indicates that the data is skewed on the right (graphically we can see a long right tail). Most of the records will be on the left of the graph.

Z-score standardisation obtained the same skewness so no value was added, my observation is that Z-score uses mean and standard deviation which both are influenced by outliers (which are very present in TOT_MINUTES_USAGE). (appendix 5.a)

Natural log reduced skewness and made it a left-skewness (as opposed to the previous right skewness), it added value (appendix 5.b): -0.7042918

Square root increased the skewness, so it is not appropriate to use it with this data (appendix 5.c): 1.288432

6. RELATIONSHIP BETWEEN VARIABLES AND RESPONSE

a.

To study the relationships, I used the same graphs plotted in appendix 3.e.

My approach was to plot histograms for each variable, colour encoded by the response variable.

Histograms where there are no disproportions between churners and non-churners on at least one of the values or range, might not be of any value for the prediction.

The question mentioned using only numeric variables, but, exploring the data showed more interesting results for overlaid graphs in some ordinal / nominal variables (included in the summary).

Variables that seem to influence churning (from this graphical method):

Income, where from the graph we can infer that Low and High Incomes are more frequent churners than Medium Incomes.

Phone plan, where we see that Euro-zone users are almost only churners, International users also have big churning rates while National and Promo-plan have low churning rates.

Education, where Primary have big churning rates and disproportions can be observed in all other categories besides Post Primary which seems to be balanced (might not help inferring rules).

Area codes where there are clear disproportions on each area.

Variables that seem to have no influence on churning (from this graphical method):

Number of lines and Gender seem to be both almost perfectly balanced, so they might be irrelevant to infer rules.

Variables for which the graph is not explicit:

Customer loyalty and Total minutes usage seem to show balance on some values while some other values show disproportions.

b.

I would expect Income or/and Area code to show up in the classification models as they seem to influence churning rate.

7. CORRELATED VARIABLES

a.

The sum of MINUTES_CURR_MONTH, MINUTES_PREV_MONTH, MINUTES_3MONTHS_AGO correlate with TOT_MINUTES_USAGE, the other variables are not correlated (appendix 7)

b.

The high correlation coefficient confirms the correlation between TOT_MINUTES_USAGE and the other usage metrics (0.9916) confirms the conclusions from the graphical analysis.

The 3 other correlation coefficients confirm that no other pair of numerical variables are correlated.

c.

As demonstrated in 3.c, the attributes that seem to have an important influence on churning are:

- Income
- Phone plan
- Education
- Area code

The attributes that seem to have some influence are:

- Customer loyalty (CUS_MOS)
- Convergent billing

The attributes that seem to be of no value to find churners are:

- Number of lines
- Gender
- Total minutes usage

d.

The variables MINUTES_CURR_MONTH, MINUTES_PREV_MONTH and MINUTES_3MONTHS_AGO should be eliminated because they correlate with Total minutes usage.

However, we showed that Total minutes usage does not seem to bring value in finding churners, so the usage times could be dropped altogether.

Gender and Number of lines could be dropped as well as it seems that they are not bringing any value, the benefit is to keep the training from trying to use meaningless variables and it will also simplify our Decision Trees (less splits).

Section 2 – Data Mining

RESULTS

	ZeroR	PART	JRip	J48
FP	0.479	0.114	0.084 lower than the other algorithms	0.114
FN	0	0.288	0.306 higher than the other algorithms	0.294
Model accuracy	47.8% accuracy	80.25%	80.96%	79.97%
Precision	0.229	0.809	0.822	0.806
True Positive Rate	0.479	0.803	0.810	0.8
False Positive Rate	0	0.205	0.2	0.208
ROC	0.5	0.859	0.846	0.845

ATTRIBUTES KEPT

The attributes kept for the training are: Income, Phone plan, Education, Area code, Customer loyalty (CUS_MOS), Convergent billing

WHAT THE ALGORITHMS DO

ZeroR trains on the proportions of the Response Classes and replicates this proportion on the test set by assigning randomly, that is how we get a result close to a random guessing (50% precision).

PART (appendix 8.a) and JRip (appendix 8.b) both create rules by combining prediction variables.

PART builds a partial C4.5 decision tree in each iteration and makes the “best” leaf into a rule.

J48 (appendix 8.c) creates a decision tree.

ATTRIBUTES USED IN THE PREDICTIONS

PART seem to use AREA_CODE greatly along with EDUCATION and INCOME while JRip rely a lot on AREA_CODE and less on EDUCATION / INCOME.

J48 uses every variable almost equally, besides INCOME which is only used for small part of the decision and CONVERGENT_BILLING that was completely dropped off.

INTERPRETING THE MODELS

The PART inferred rules are to be read in the following manner:

If the phone plan is Euro-Zone and area code is 36785 then that person will churn (39 were classified correctly using this rule),

Else, if income is high and no convergent billing then that person will not churn (510 were classified correctly using this rule, 137 were wrongly classified)

Etc...

The JRip inferred rules are to be read in the following manner:

If income is high and no convergent billing then that person will not churn (510 were classified correctly using this rule, 137 were wrongly classified) -> same rule as in PART

Else, if area code is 10040 then that person will not churn (310 were classified correctly using this rule, 125 were wrongly classified)

Etc…

The J48 tree is to be read in the following manner:

If phone plan is international and education is master then that person will churn (180 were correctly classified using this rule)

If phone plan is euro zone then that person will churn (59 correctly classified using this rule, 9 incorrectly classified)

Etc…

The decisions align with the conclusions drawn by the graphical analysis.

KEY PREDICTORS OF CHURNING

INCOME and AREA_CODE seem to be the two main predictors of churning across all algorithms.

SIGNIFICANT DECISIONS PATHS

From the PART and the JRip output, "Income high and convergent billing no" is a rule that classified 510 records correctly and 137 incorrectly (a very used rule), this rule is significant and the result shows that it is meaningful.

From the J48 tree, the rule "Phone plan international and education post-primary" is a rule that classified 351 records correctly and 71 incorrectly, making it also a significant rule.

OVERALL ASSESSMENT

1. From the conclusions above we can see that some areas a more prone to churning than others, it might be low coverage or low quality of the service in those areas.
   Education is also a factor, that is combined with phone plans or area codes to give multiple rules. The education might have a relation with the age of the customer and therefore its needs and an inadequate offer or phone plan could make these customers churn.

2. The persons with an international phone plan and that have post-primary education or Masters seem to be more likely to churn (J48). The area code 21750 seems to be prone to churning as well.
   These are the clearest rules that we can infer from the algorithms output.

3. The customers that enter in these categories can be monitored more closely, marketing measures can be taken to try and improve customer retention in these areas as well as getting feedback from these customers could be insightful for the company.

APPENDIX WRITTEN REPORT

8.a

```
PART decision list
------------------

PHONE_PLAN = Euro-Zone AND
AREA_CODE = 36785: yes (39.0)

INCOME = High Income AND
CONVERGENT_BILLING = No: no (510.0/137.0)

AREA_CODE = 21750: yes (360.0/72.0)

AREA_CODE = 45987 AND
EDUCATION = Masters: yes (90.0)

AREA_CODE = 10040 AND
INCOME = Low Income: no (271.0/80.0)

PHONE_PLAN = Promo_plan: no (90.0)

EDUCATION = PhD AND
AREA_CODE = 45987: no (90.0)

AREA_CODE = 15563 AND
EDUCATION = Bachelors: no (182.0/72.0)

INCOME = Medium Income AND
AREA_CODE = 45987: yes (90.0)

EDUCATION = Post Primary: yes (88.0)

AREA_CODE = 36785: yes (80.0)

INCOME = High Income AND
PHONE_PLAN = International: no (70.0/20.0)

AREA_CODE = 15563: yes (21.0)

PHONE_PLAN = Euro-Zone AND
CONVERGENT_BILLING = Yes: yes (10.0/3.0)

: no (80.0/38.0)

Number of Rules  :      15
```

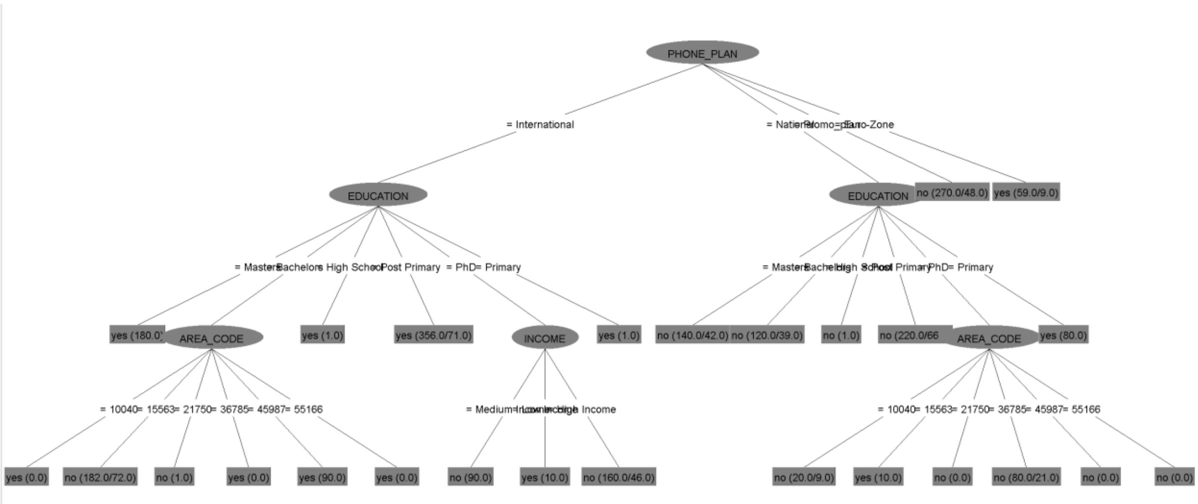
8.b

```
JRIP rules:
===========

(INCOME = High Income) and (CONVERGENT_BILLING = No) => CHURNER=no (510.0/137.0)
(AREA_CODE = 10040) => CHURNER=no (361.0/125.0)
(EDUCATION = PhD) and (AREA_CODE = 45987) => CHURNER=no (90.0/0.0)
(AREA_CODE = 55166) => CHURNER=no (90.0/0.0)
(AREA_CODE = 15563) and (EDUCATION = Bachelors) => CHURNER=no (182.0/72.0)
(INCOME = High Income) and (PHONE_PLAN = International) => CHURNER=no (71.0/21.0)
 => CHURNER=yes (767.0/72.0)

Number of Rules : 7
```

8.c



Data Understanding and Data Exploration

Daniel-Mateus-Pires

PDF config

```
library(knitr)
opts_chunk$set(tidy.opts=list(width.cutoff=60),tidy=TRUE)
```

EuroCom

Dependencies

```
# install.packages('ggplot2')
library(ggplot2)
```

Reading the dataset

```
phones <- read.csv("./eurocomPHONEchurners.csv")
head(phones)
```

```
##   CUST_ID AREA_CODE MINUTES_CURR_MONTH MINUTES_PREV_MONTH
## 1     129     45987                 60                456
## 2     130     15563                  2                  0
## 3     131     10040                  2                  0
## 4     132     21750                678               1222
## 5     133     55166                110                 98
## 6     134     36785                 97                 56
##   MINUTES_3MONTHS_AGO CUST_MOS LONGDIST_FLAG CALLWAITING_FLAG NUM_LINES
## 1                 398       13             0                1         1
## 2                   4        4             0                0         1
## 3                   0        1             0                0         1
## 4                 598       30             1                1         2
## 5                  56       15             1                0         1
## 6                  97        8             0                0         1
##   VOICEMAIL_FLAG MOBILE_PLAN CONVERGENT_BILLING GENDER INCOME
## 1              1           0                Yes      M  88000
## 2              0           0                Yes      M  53000
## 3              1           0                 No      F  29000
## 4              0           1                Yes      M  46000
## 5              1           0                 No      M  98000
## 6              0           1                 No      M 125000
##      PHONE_PLAN   EDUCATION TOT_MINUTES_USAGE CHURNER
## 1 International     Masters               914     yes
## 2 International   Bachelors                 6      no
## 3      National High School                 0      no
## 4 International High School              2498     yes
## 5    Promo_plan High School               264      no
```

```
## 6      National                              250       no
```

1

Data pre-processing

1.a

Getting how many null values, or empty string values there is per column.

```
count_na <- sapply(phones, function(y) sum(length(which(is.na(y) |
    y == ""))))
na_df <- data.frame(count_na)

subset(na_df, na_df$count_na > 0)
```

```
##                      count_na
## MINUTES_3MONTHS_AGO         3
## CUST_MOS                    3
## PHONE_PLAN                  4
## EDUCATION                   8
## TOT_MINUTES_USAGE           4
```

1.b

Replacing na numerics with medians

```
replace_na_with_median <- function(col) {
    median_without_na <- median(col, na.rm = TRUE)
    col[is.na(col)] <- median_without_na
    return(col)
}
```

MINUTES_3MONTHS_AGO

```
phones$MINUTES_3MONTHS_AGO <- replace_na_with_median(phones$MINUTES_3MONTHS_AGO)
```

CUST_MOS

```
phones$CUST_MOS <- replace_na_with_median(phones$CUST_MOS)
```

TOT_MINUTES_USAGE

```
phones$TOT_MINUTES_USAGE <- replace_na_with_median(phones$TOT_MINUTES_USAGE)
```

1.c

Getting the mode for categorical columns PER GENDER

```
get_mode <- function(x) {
    xtable <- table(x)
    idx <- xtable == max(xtable)
    names(xtable)[idx]
}
```

Function to get all modes from a data frame

```
get_modes <- function(x) {
    if (class(x) == "numeric" | class(x) == "integer")
        return("X")
    xtable <- table(x)
    idx <- xtable == max(xtable)
    names(xtable)[idx]
}
```

Displaying modes for males

```
phones_male <- phones[phones$GENDER == "M", ]

modes_male <- data.frame(sapply(phones_male, get_modes))
names(modes_male)[1] <- "MODE_MALE"
modes_male <- subset(modes_male, MODE_MALE != "X")

modes_male
```

```
##                         MODE_MALE
## CONVERGENT_BILLING            Yes
## GENDER                          M
## PHONE_PLAN          International
## EDUCATION            Post Primary
## CHURNER                       yes
```

Displaying modes for females

```
phones_female <- phones[phones$GENDER == "F", ]

modes_female <- data.frame(sapply(phones_female, get_modes))
names(modes_female)[1] <- "MODE_FEMALE"
modes_female <- subset(modes_female, MODE_FEMALE != "X")

modes_female
```

```
##                       MODE_FEMALE
## CONVERGENT_BILLING             No
## GENDER                          F
## PHONE_PLAN          International
## EDUCATION               Bachelors
## CHURNER                        no
```

PHONE_PLAN

```r
phones$PHONE_PLAN[phones$PHONE_PLAN == "" & phones$GENDER ==
    "M"] <- get_mode(phones$PHONE_PLAN[phones$GENDER == "M"])
phones$PHONE_PLAN[phones$PHONE_PLAN == "" & phones$GENDER ==
    "F"] <- get_mode(phones$PHONE_PLAN[phones$GENDER == "F"])
```

EDUCATION

```r
phones$EDUCATION[phones$EDUCATION == "" & phones$GENDER == "M"] <- get_mode(phones$EDUCATION[phones$GEN
    "M"])
phones$EDUCATION[phones$EDUCATION == "" & phones$GENDER == "F"] <- get_mode(phones$EDUCATION[phones$GEN
    "F"])
```

AREA_CODE

```r
phones$AREA_CODE[phones$AREA_CODE == "" & phones$GENDER == "M"] <- get_mode(phones$AREA_CODE[phones$GEN
    "M"])
phones$AREA_CODE[phones$AREA_CODE == "" & phones$GENDER == "F"] <- get_mode(phones$AREA_CODE[phones$GEN
    "F"])
```

LONGDIST_FLAG

```r
phones$LONGDIST_FLAG[phones$LONGDIST_FLAG == "" & phones$GENDER ==
    "M"] <- get_mode(phones$LONGDIST_FLAG[phones$GENDER == "M"])
phones$LONGDIST_FLAG[phones$LONGDIST_FLAG == "" & phones$GENDER ==
    "F"] <- get_mode(phones$LONGDIST_FLAG[phones$GENDER == "F"])
```

CALLWAITING_FLAG

```r
phones$CALLWAITING_FLAG[phones$CALLWAITING_FLAG == "" & phones$GENDER ==
    "M"] <- get_mode(phones$CALLWAITING_FLAG[phones$GENDER ==
    "M"])
phones$CALLWAITING_FLAG[phones$CALLWAITING_FLAG == "" & phones$GENDER ==
    "F"] <- get_mode(phones$CALLWAITING_FLAG[phones$GENDER ==
    "F"])
```

VOICEMAIL_FLAG

```r
phones$VOICEMAIL_FLAG[phones$VOICEMAIL_FLAG == "" & phones$GENDER ==
    "M"] <- get_mode(phones$VOICEMAIL_FLAG[phones$GENDER == "M"])
phones$VOICEMAIL_FLAG[phones$VOICEMAIL_FLAG == "" & phones$GENDER ==
    "F"] <- get_mode(phones$VOICEMAIL_FLAG[phones$GENDER == "F"])
```

MOBILE_PLAN

```r
phones$MOBILE_PLAN[phones$MOBILE_PLAN == "" & phones$GENDER ==
    "M"] <- get_mode(phones$MOBILE_PLAN[phones$GENDER == "M"])
phones$MOBILE_PLAN[phones$MOBILE_PLAN == "" & phones$GENDER ==
    "F"] <- get_mode(phones$MOBILE_PLAN[phones$GENDER == "F"])
```

2

Discretising Income predictor values

```r
head(phones$INCOME)
```

```
## [1]  88000  53000  29000  46000  98000 125000
```

```r
phones$INCOME <- cut(phones$INCOME, breaks = c(0, 37999, 88000,
    max(phones$INCOME)), include.lowest = TRUE, labels = c("Low Income",
    "Medium Income", "High Income"))

head(phones$INCOME)
```

```
## [1] Medium Income Medium Income Low Income    Medium Income High Income
## [6] High Income
## Levels: Low Income Medium Income High Income
```

3.c

```r
get_mode(phones$AREA_CODE)
```

```
## [1] "10040"
```

```r
summary(phones$CUST_MOS)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1.00    6.00   11.00   16.05   26.00   50.00
```

```r
get_mode(phones$LONGDIST_FLAG)
```

```
## [1] "1"
```

```r
get_mode(phones$CALLWAITING_FLAG)
```

```
## [1] "0"
```

```r
summary(phones$NUM_LINES)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   1.000   1.000   1.391   2.000   3.000
```

```r
get_mode(phones$VOICEMAIL_FLAG)
```

```
## [1] "1"
```

```r
get_mode(phones$MOBILE_PLAN)
```

```
## [1] "0"
```

```r
get_mode(phones$CONVERGENT_BILLING)
```

```
## [1] "No"
```

```r
get_mode(phones$GENDER)
```

```
## [1] "M"
```

```r
get_mode(phones$INCOME)
```

```
## [1] "Medium Income"
```

```r
get_mode(phones$PHONE_PLAN)
```

```
## [1] "International"
```

```r
get_mode(phones$EDUCATION)
```

```
## [1] "Post Primary"
```

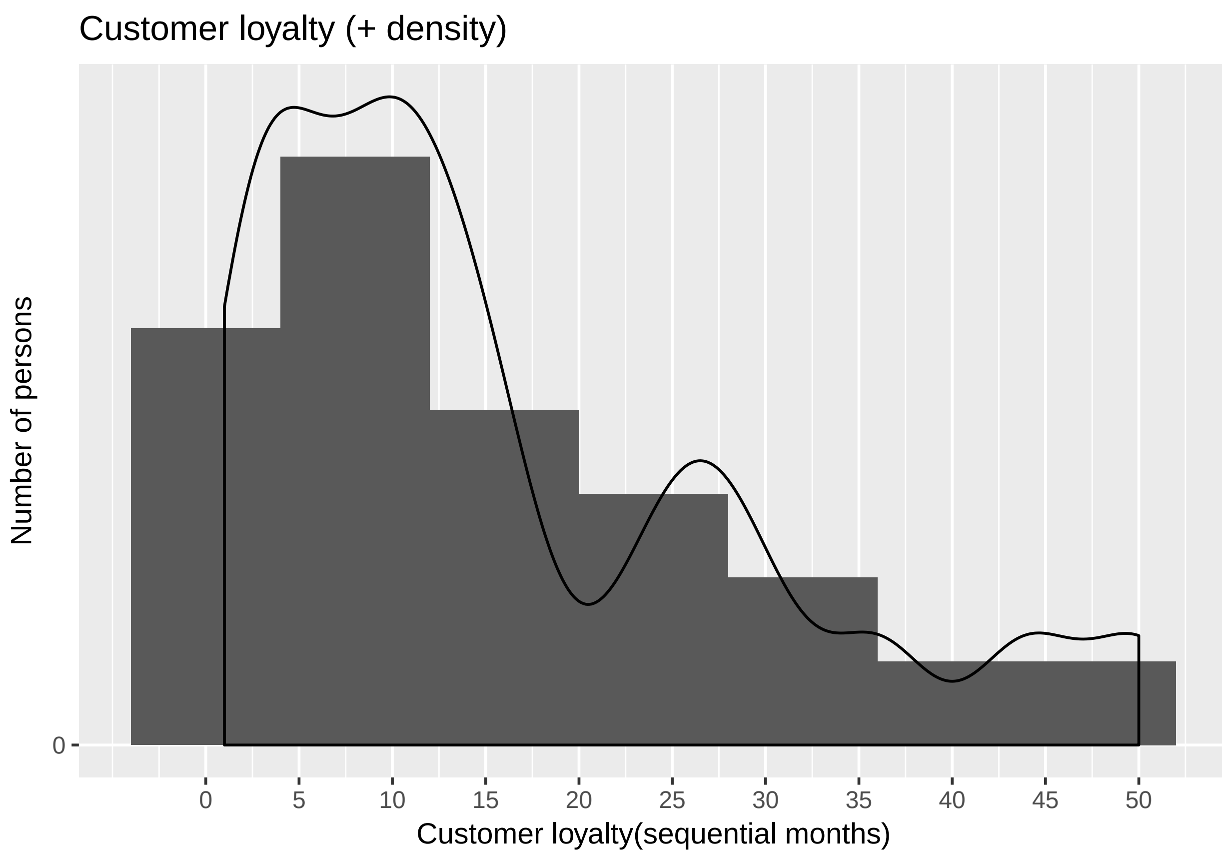
```r
summary(phones$TOT_MINUTES_USAGE)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##       0     116     264    2036    1677   36240
```

3.d

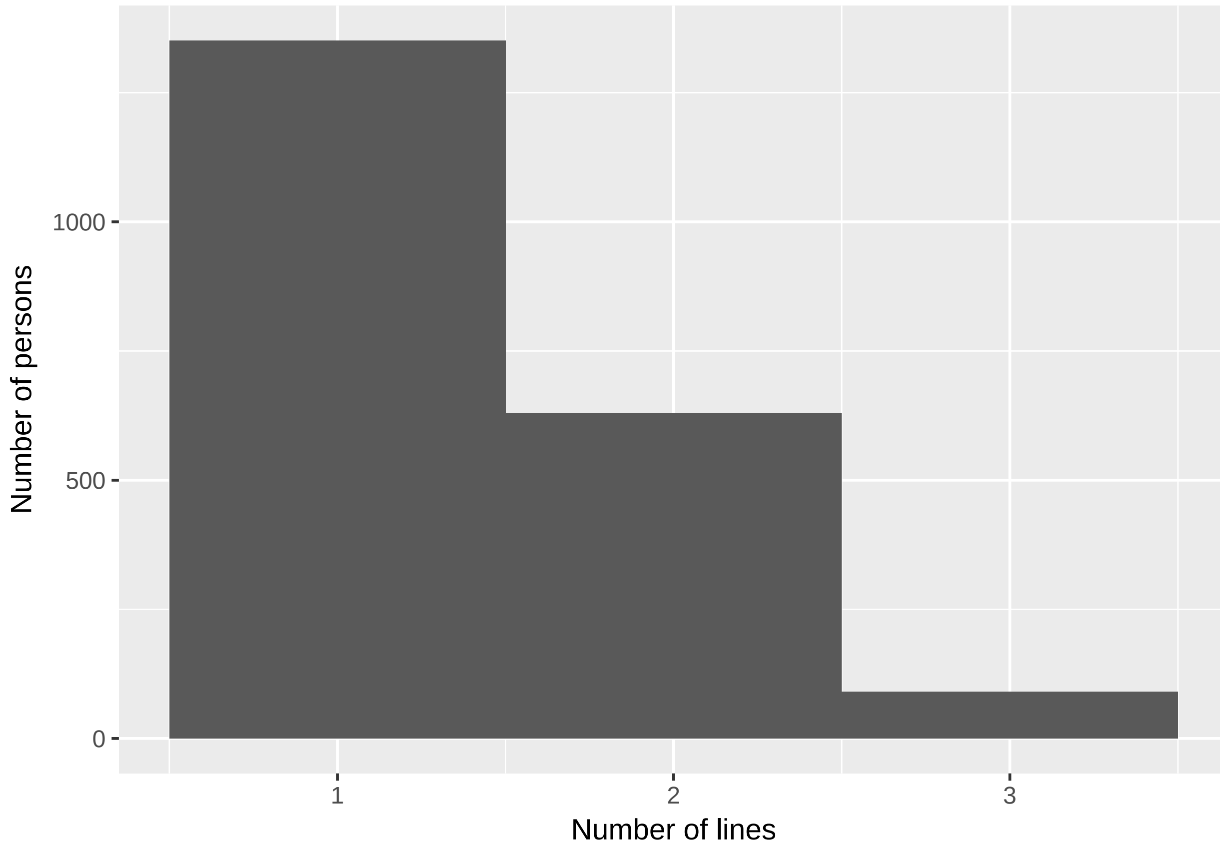
CUST_MOS

```r
ggplot(data = phones, aes(phones$CUST_MOS)) + geom_histogram(binwidth = 8,
    aes(y = ..density..)) + scale_x_continuous(breaks = seq(0,
    60, 5)) + scale_y_continuous(breaks = seq(0, 1000, 50)) +
    labs(x = "Customer loyalty(sequential months)", y = "Number of persons",
        title = "Customer loyalty (+ density)") + geom_density()
```



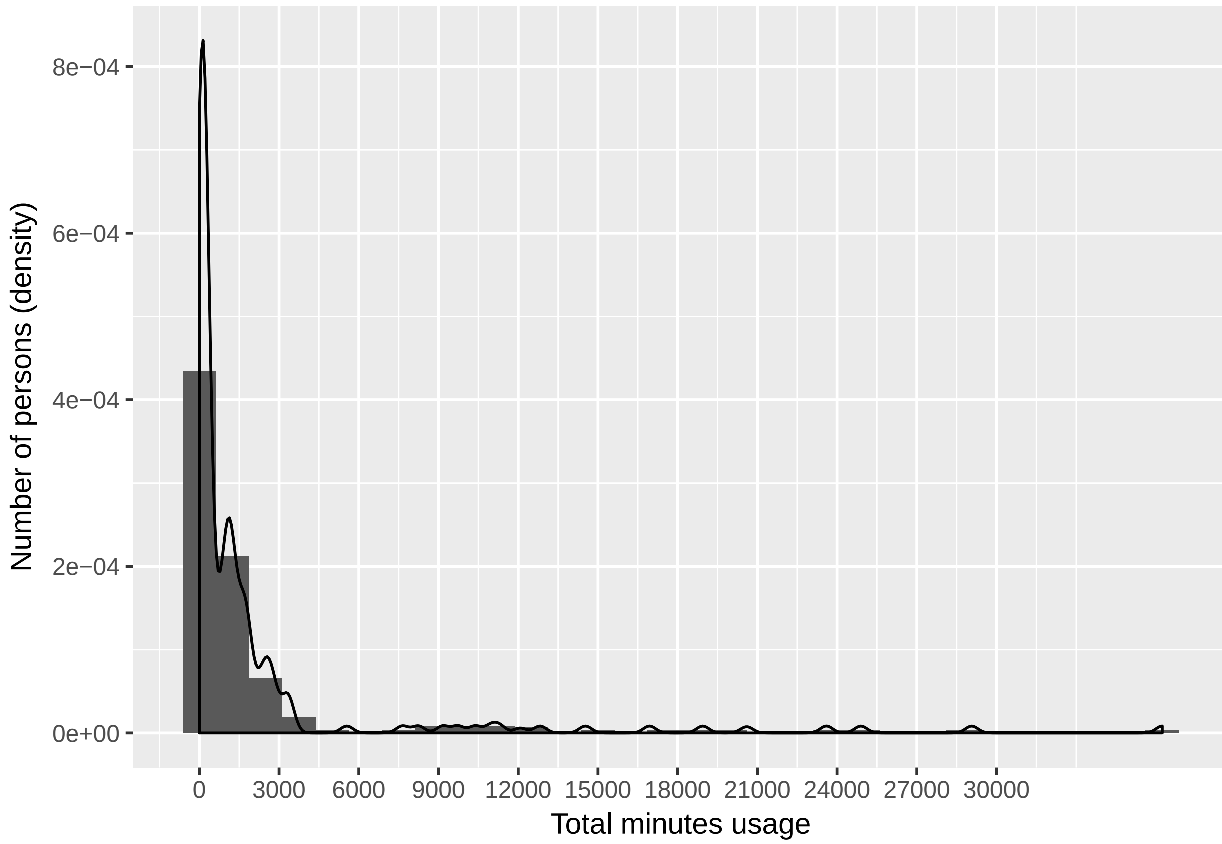
NUM_LINES

```
ggplot(data = phones, aes(phones$NUM_LINES)) + geom_histogram(binwidth = 1) +
    scale_x_continuous(breaks = 0:3) + labs(x = "Number of lines",
    y = "Number of persons")
```



TOT_MINUTES_USAGE

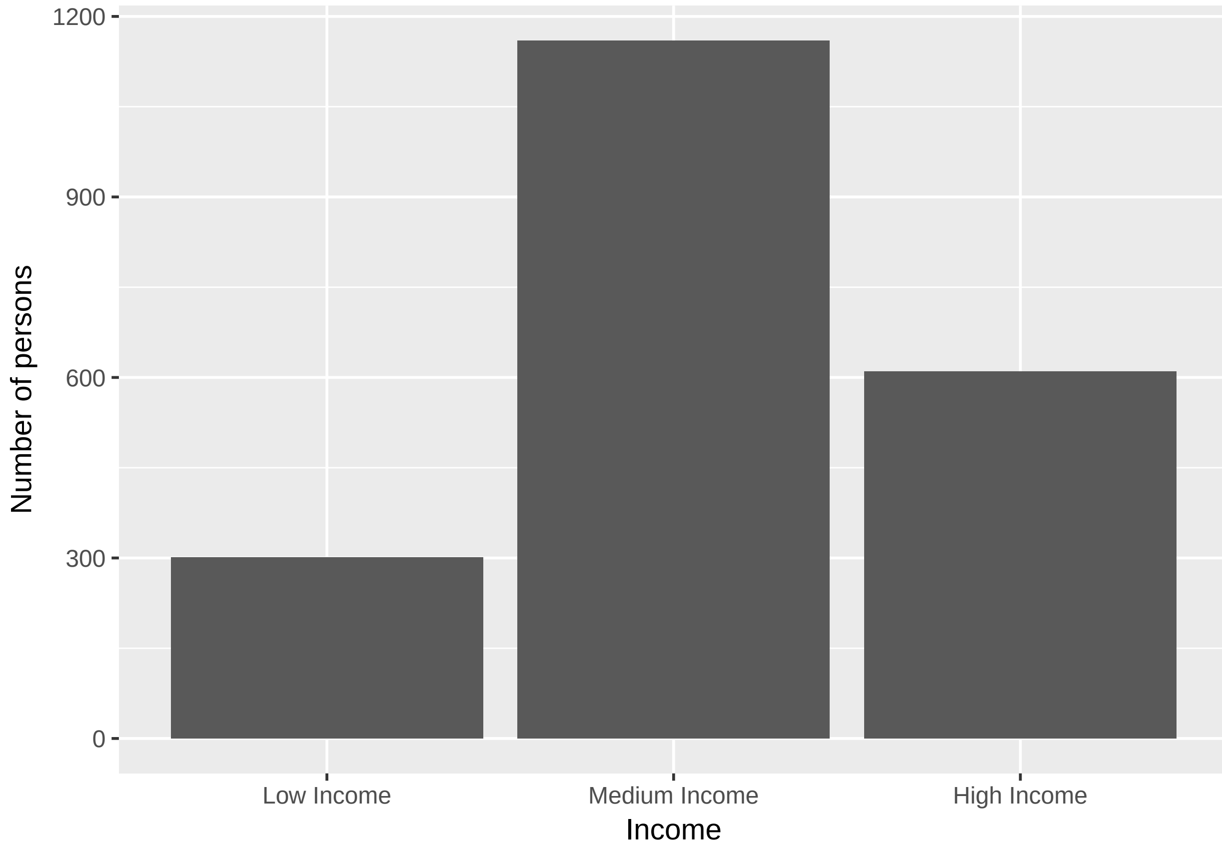
```
ggplot(data = phones, aes(phones$TOT_MINUTES_USAGE)) + geom_histogram(bins = 30,
    aes(y = ..density..)) + scale_x_continuous(breaks = seq(0,
    30000, 3000)) + labs(x = "Total minutes usage", y = "Number of persons (density)") +
    geom_density()
```



INCOME

```
ggplot(data = phones, aes(phones$INCOME)) + geom_histogram(stat = "count") +
    labs(x = "Income", y = "Number of persons")
```

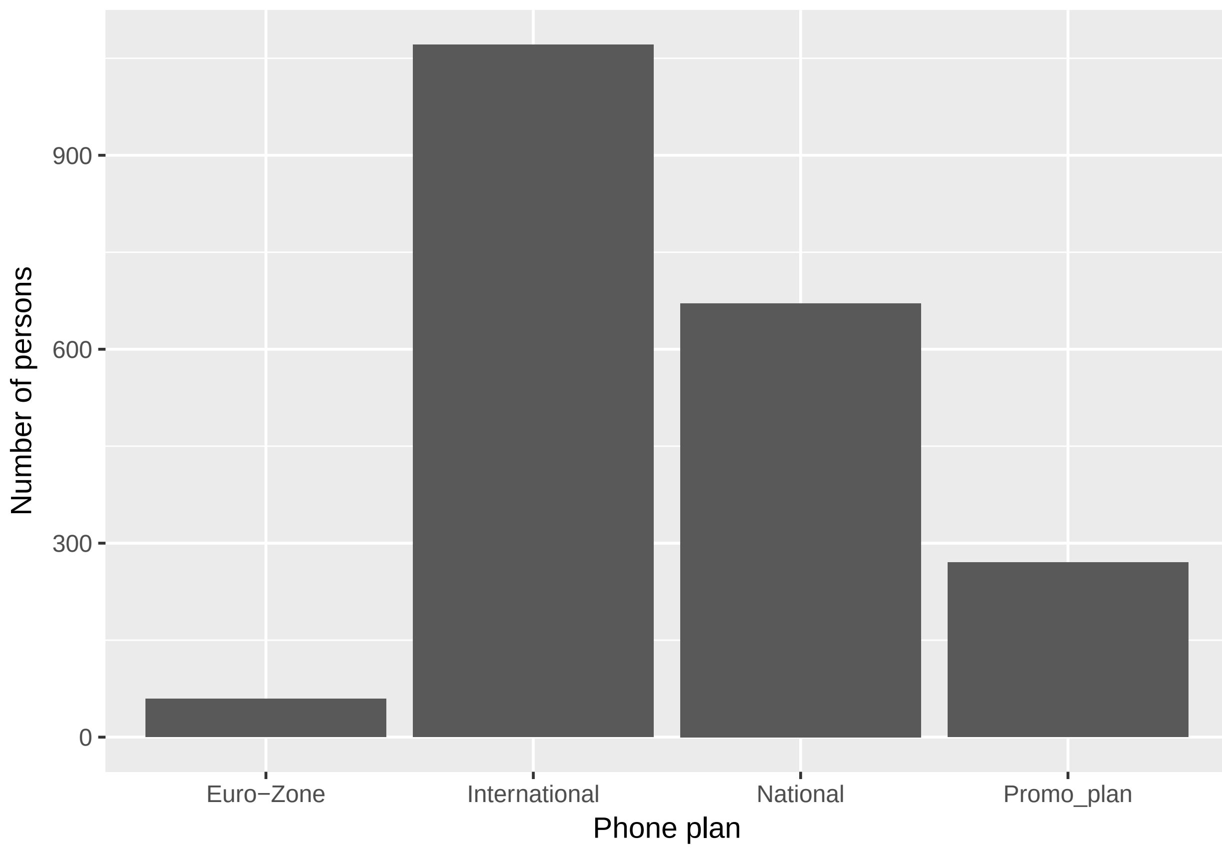
```
## Warning: Ignoring unknown parameters: binwidth, bins, pad
```



PHONE_PLAN

```
ggplot(data = phones, aes(phones$PHONE_PLAN)) + geom_histogram(stat = "count") +
    labs(x = "Phone plan", y = "Number of persons")
```

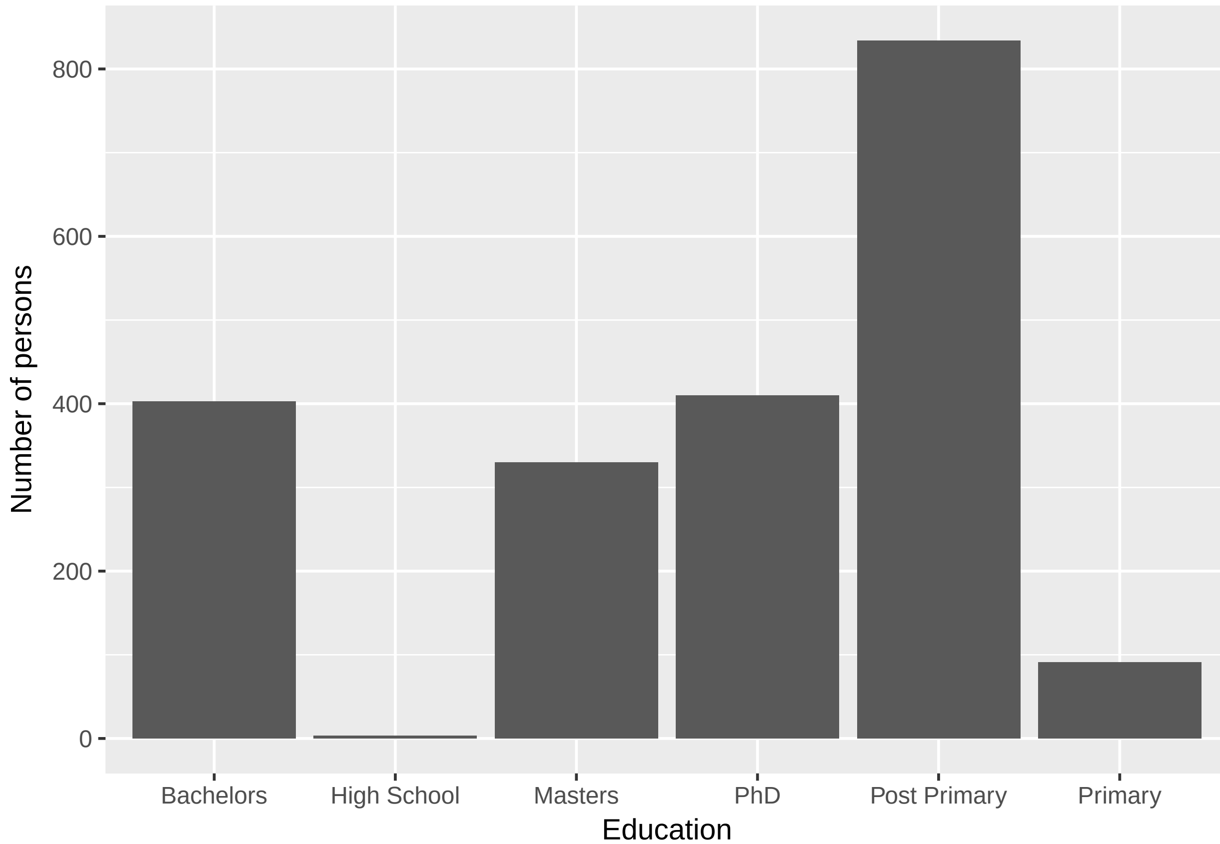
```
## Warning: Ignoring unknown parameters: binwidth, bins, pad
```



EDUCATION

```r
ggplot(data = phones, aes(phones$EDUCATION)) + geom_histogram(stat = "count") +
    labs(x = "Education", y = "Number of persons")
```

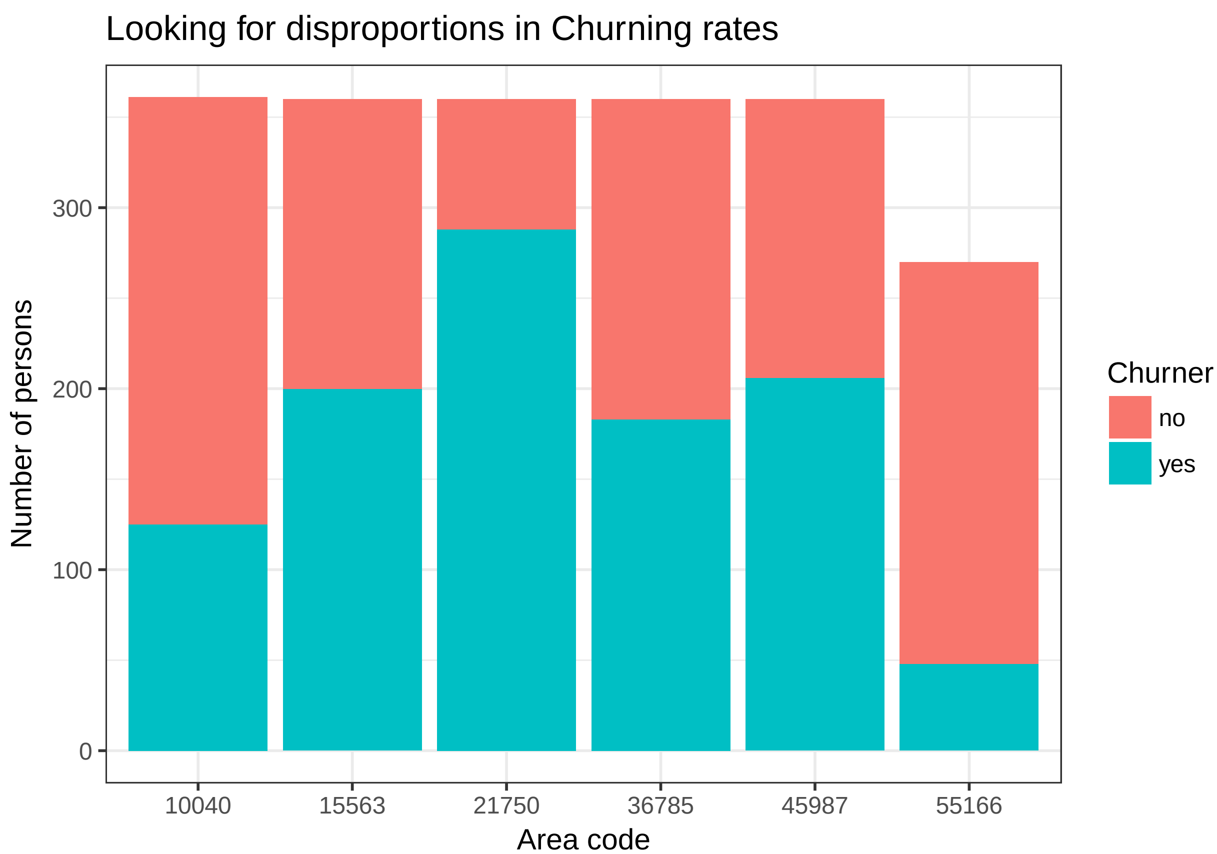
```
## Warning: Ignoring unknown parameters: binwidth, bins, pad
```



3.e ### AREA_CODE

```
ggplot(data = phones, aes(x = phones$AREA_CODE, group = phones$CHURNER,
    fill = phones$CHURNER)) + geom_histogram(stat = "count") +
    theme_bw() + labs(x = "Area code", y = "Number of persons",
    title = "Looking for disproportions in Churning rates", fill = "Churner")
```

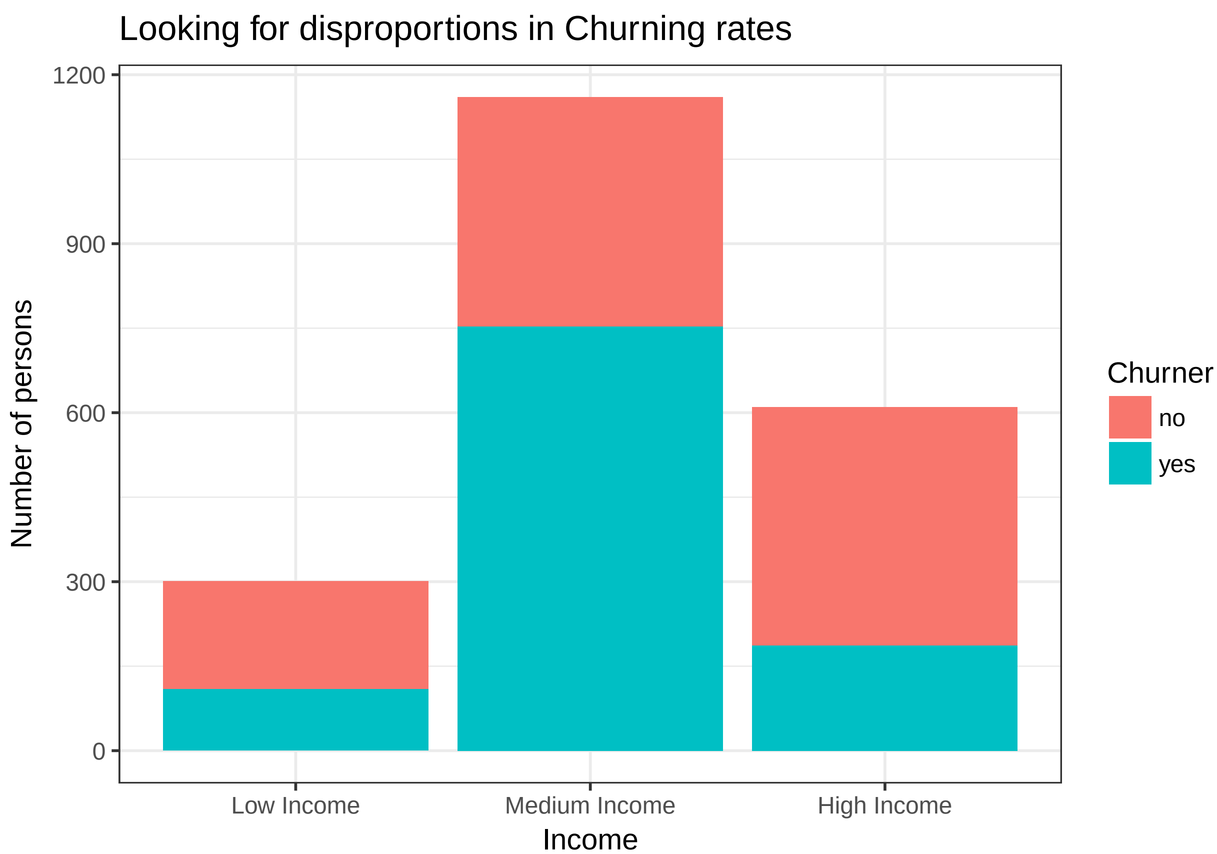
```
## Warning: Ignoring unknown parameters: binwidth, bins, pad
```



INCOME

```
ggplot(data = phones, aes(x = phones$INCOME, group = phones$CHURNER,
    fill = phones$CHURNER)) + geom_histogram(stat = "count") +
    theme_bw() + labs(x = "Income", y = "Number of persons",
    title = "Looking for disproportions in Churning rates", fill = "Churner")
```

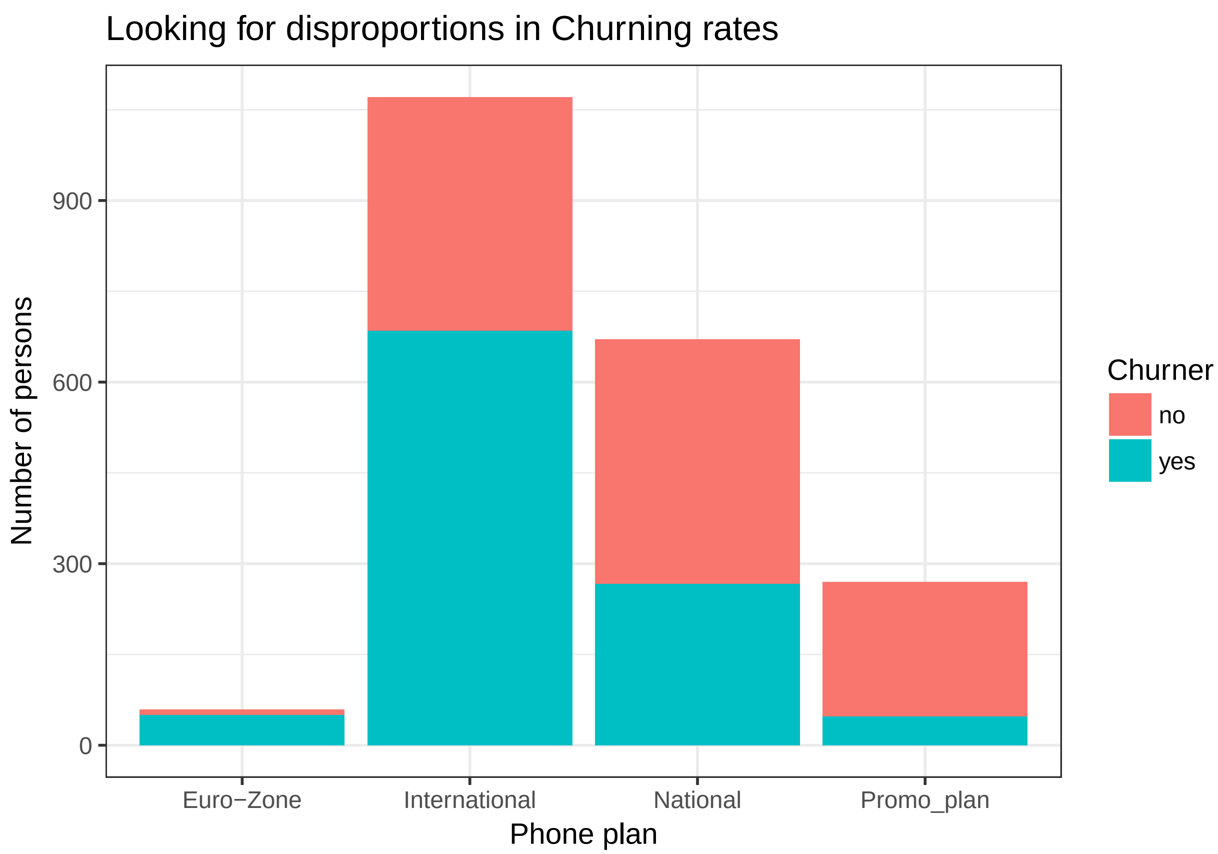
```
## Warning: Ignoring unknown parameters: binwidth, bins, pad
```



PHONE_PLAN

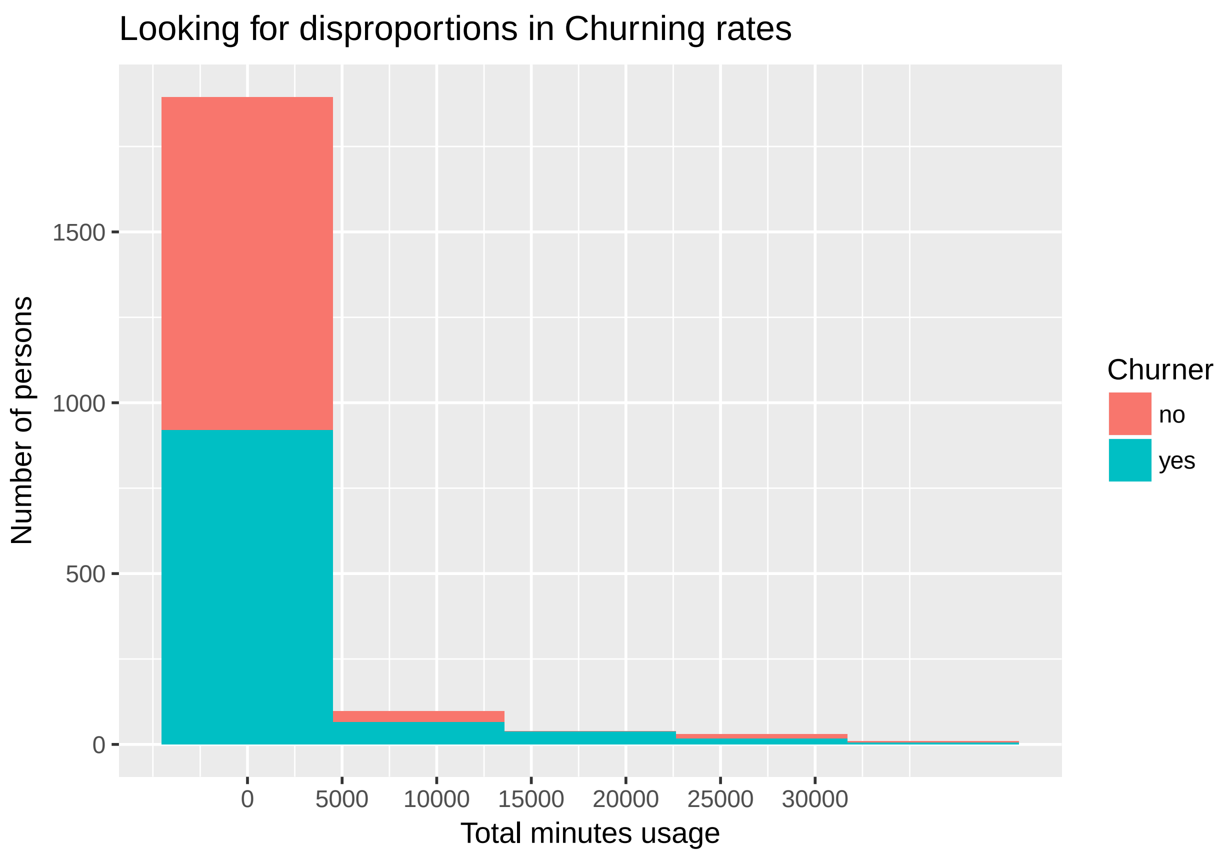
```
ggplot(data = phones, aes(x = phones$PHONE_PLAN, group = phones$CHURNER,
    fill = phones$CHURNER)) + geom_histogram(stat = "count") +
    theme_bw() + labs(x = "Phone plan", y = "Number of persons",
    title = "Looking for disproportions in Churning rates", fill = "Churner")
```

```
## Warning: Ignoring unknown parameters: binwidth, bins, pad
```



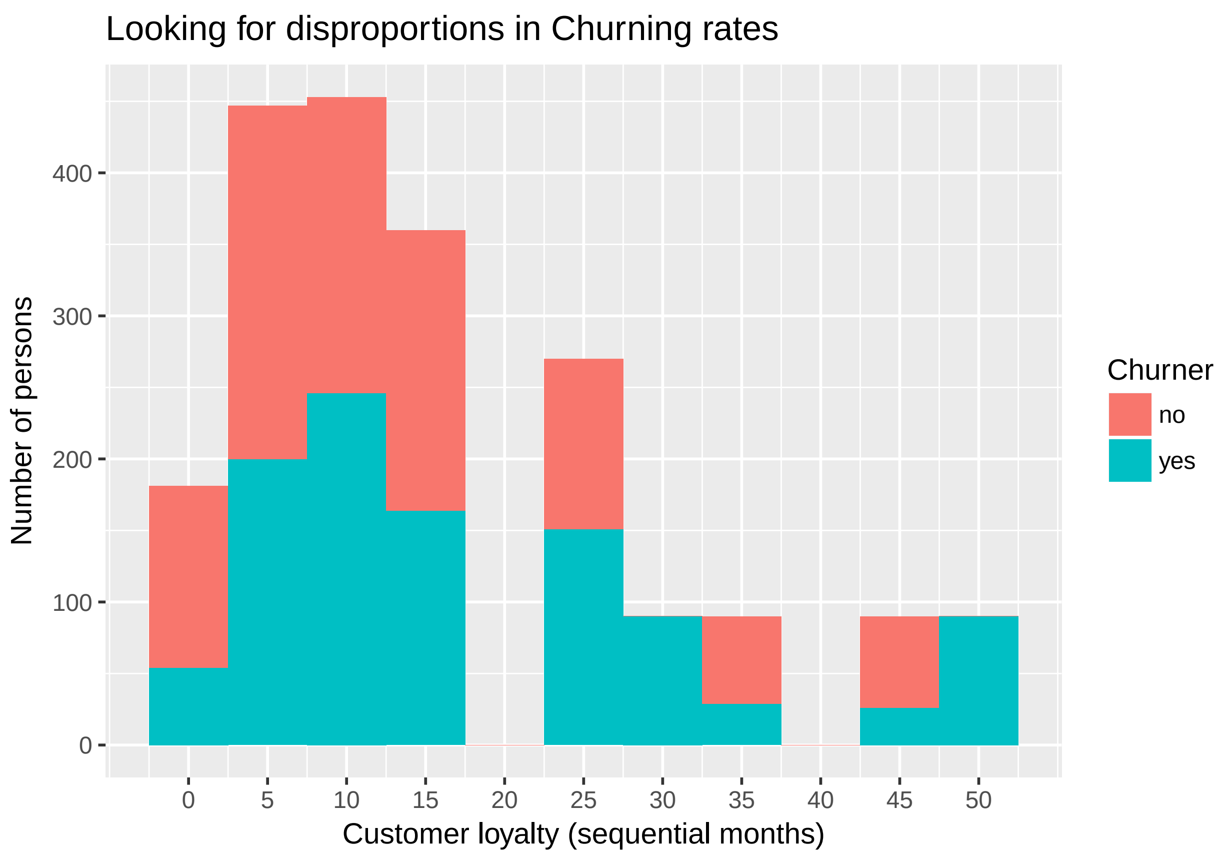
TOT_MINUTES_USAGE

```
ggplot(data = phones, aes(x = phones$TOT_MINUTES_USAGE, group = phones$CHURNER,
    fill = phones$CHURNER)) + geom_histogram(bins = 5) + scale_x_continuous(breaks = seq(0,
    30000, 5000)) + labs(x = "Total minutes usage", y = "Number of persons",
    title = "Looking for disproportions in Churning rates", fill = "Churner")
```



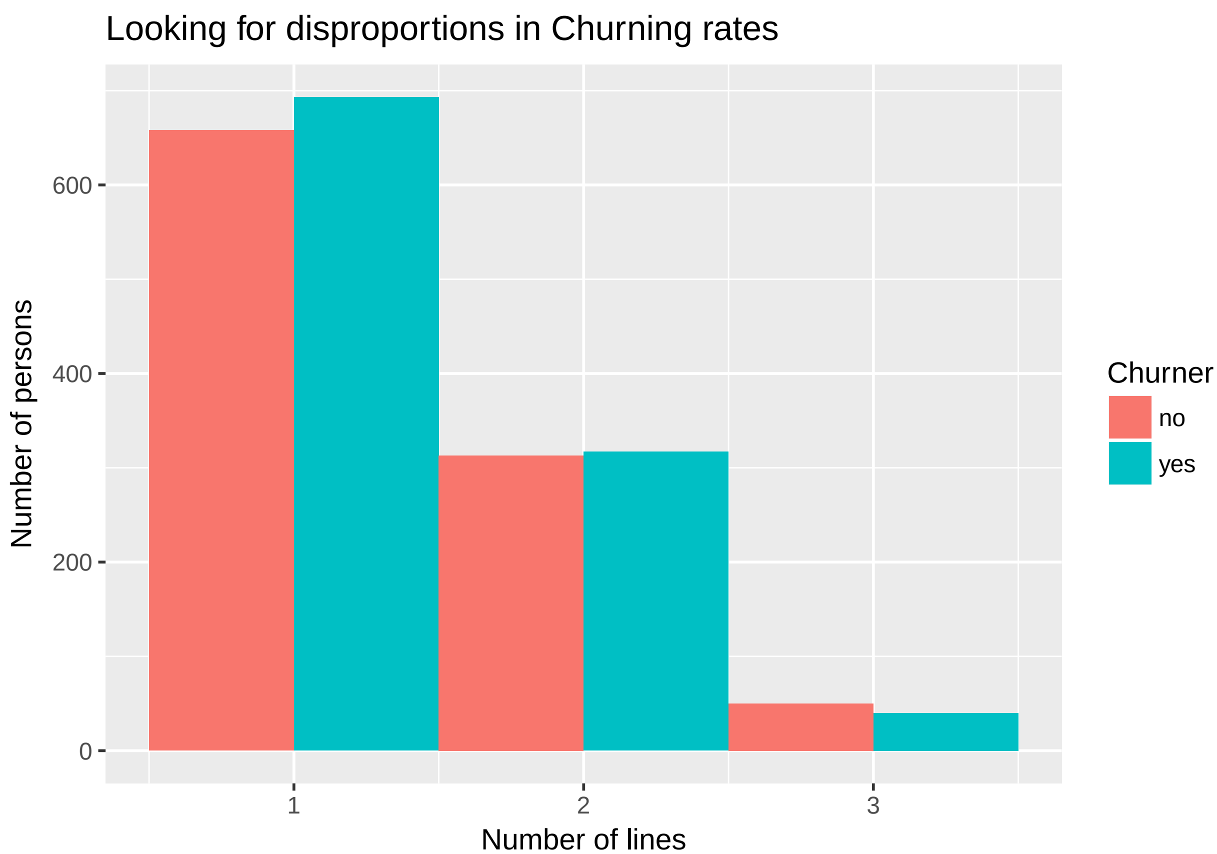
CUST_MOS

```
ggplot(data = phones, aes(x = phones$CUST_MOS, group = phones$CHURNER,
    fill = phones$CHURNER)) + geom_histogram(binwidth = 5) +
    scale_x_continuous(breaks = seq(0, 50, 5)) + labs(x = "Customer loyalty (sequential months)",
    y = "Number of persons", title = "Looking for disproportions in Churning rates",
    fill = "Churner")
```



NUM_LINES

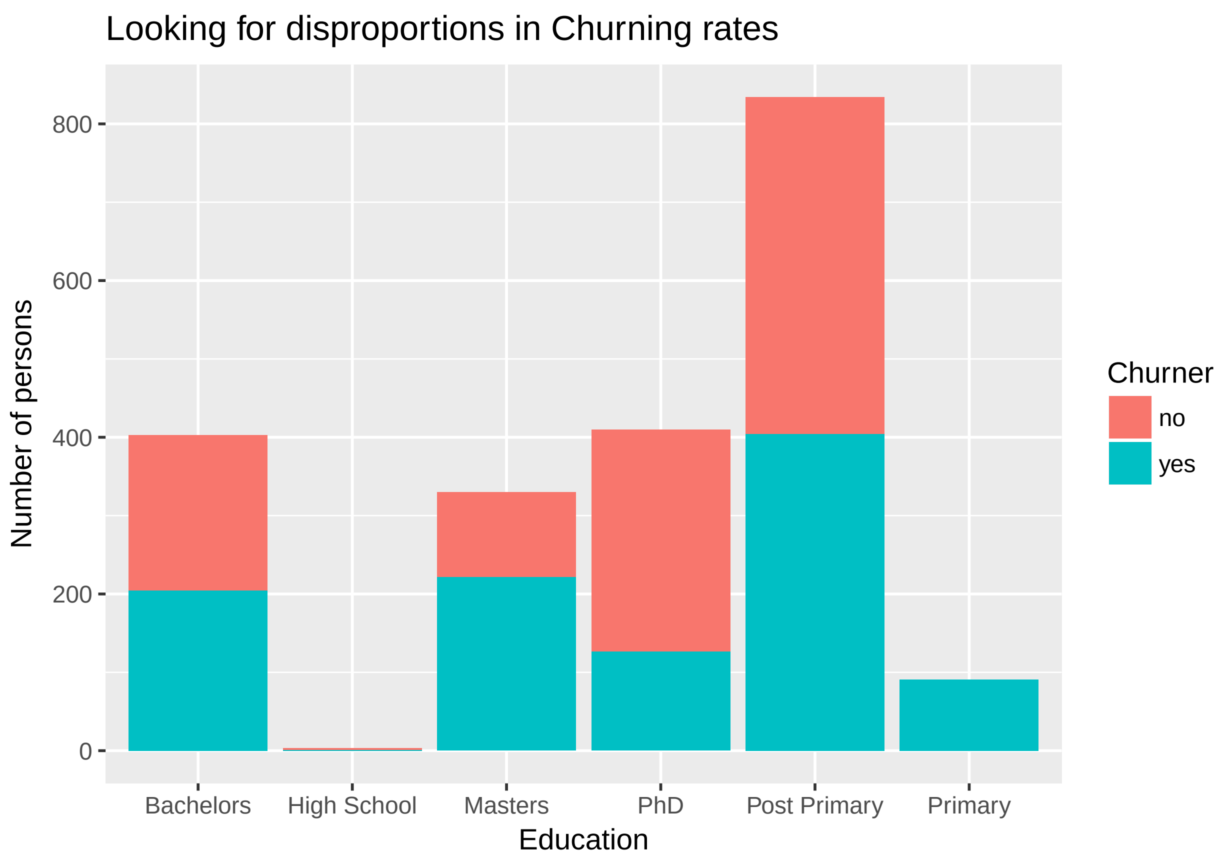
```
ggplot(data = phones, aes(x = phones$NUM_LINES, group = phones$CHURNER,
    fill = phones$CHURNER)) + geom_histogram(binwidth = 1, position = "dodge") +
    scale_x_continuous(breaks = 0:3) + labs(x = "Number of lines",
    y = "Number of persons", title = "Looking for disproportions in Churning rates",
    fill = "Churner")
```



EDUCATION

```
ggplot(data = phones, aes(x = phones$EDUCATION, group = phones$CHURNER,
    fill = phones$CHURNER)) + geom_histogram(stat = "count") +
    labs(x = "Education", y = "Number of persons", title = "Looking for disproportions in Churning rates
        fill = "Churner")
```

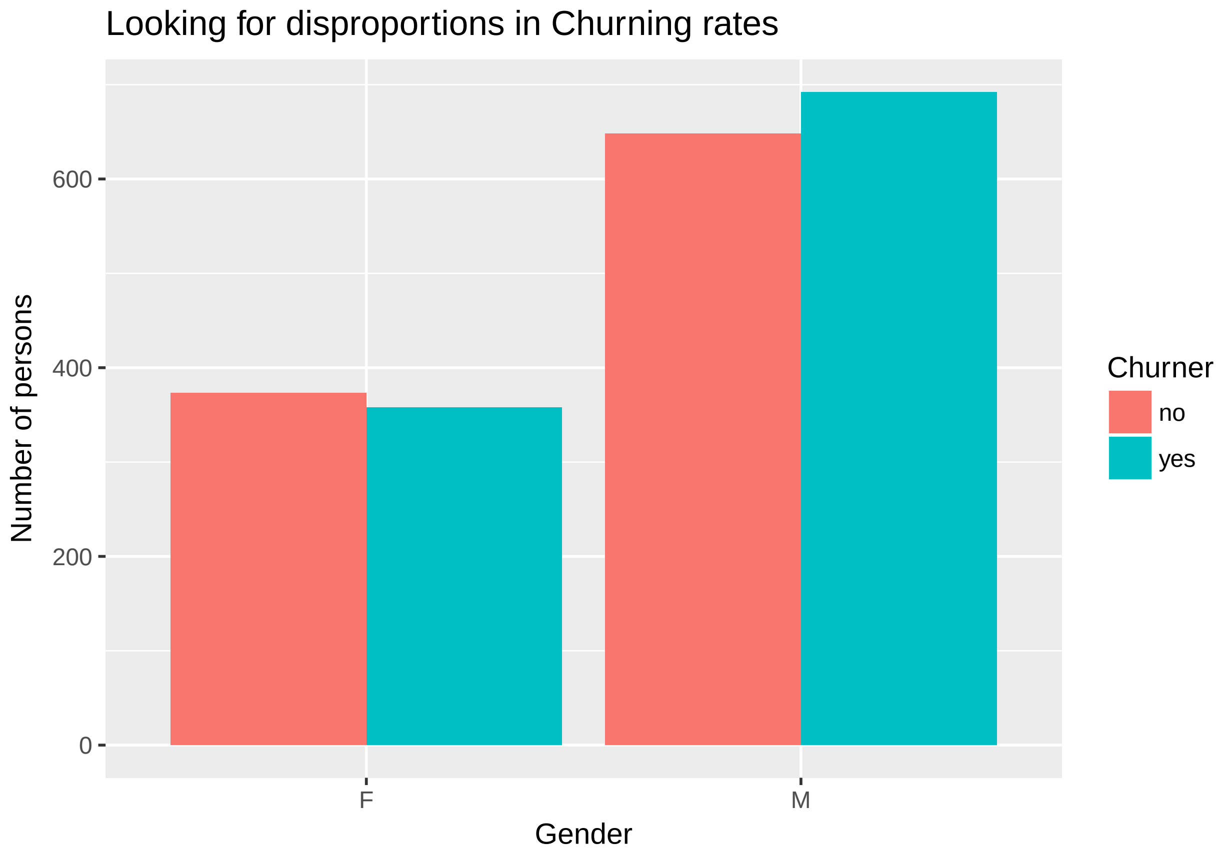
```
## Warning: Ignoring unknown parameters: binwidth, bins, pad
```



GENDER

```
ggplot(data = phones, aes(x = phones$GENDER, group = phones$CHURNER,
    fill = phones$CHURNER)) + geom_histogram(stat = "count",
    position = "dodge") + labs(x = "Gender", y = "Number of persons",
    title = "Looking for disproportions in Churning rates", fill = "Churner")
```

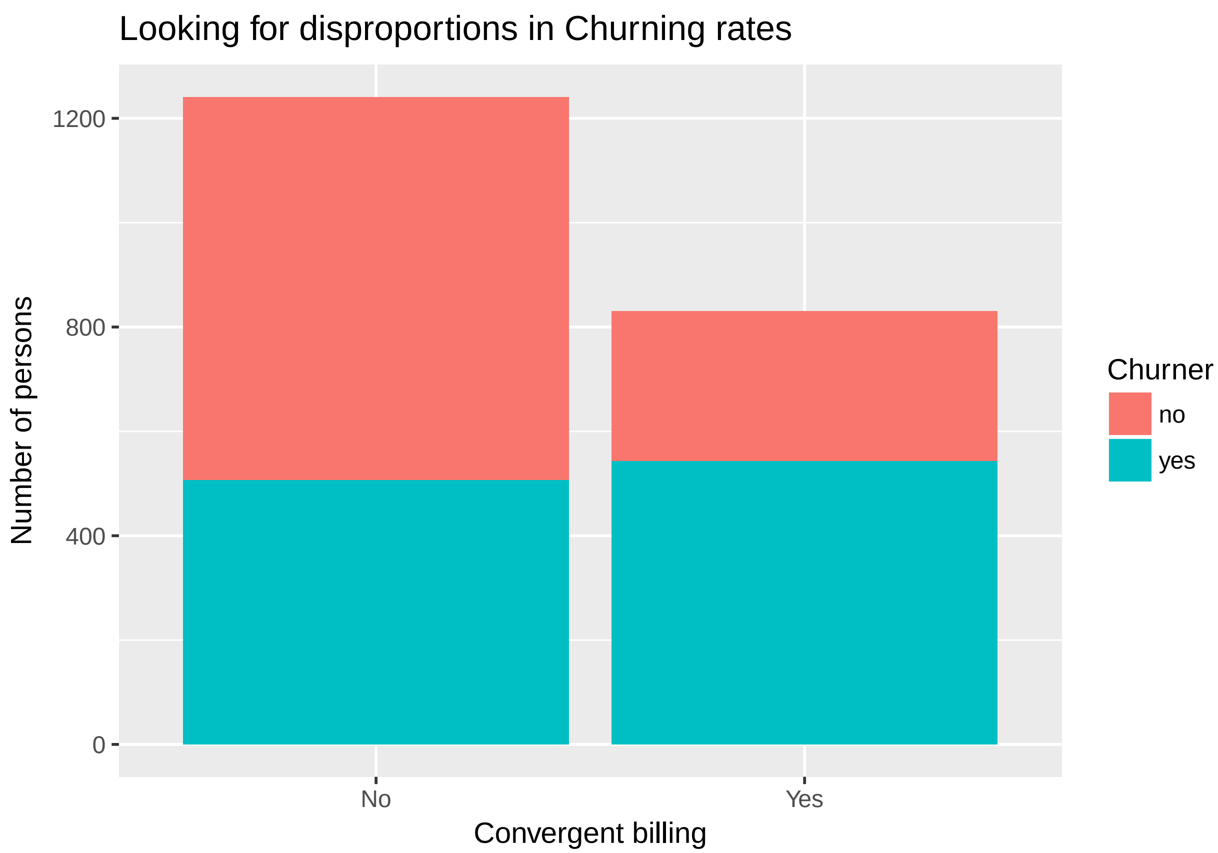
```
## Warning: Ignoring unknown parameters: binwidth, bins, pad
```



CONVERGENT_BILLING

```r
ggplot(data = phones, aes(x = phones$CONVERGENT_BILLING, group = phones$CHURNER,
    fill = phones$CHURNER)) + geom_histogram(stat = "count") +
    labs(x = "Convergent billing", y = "Number of persons", title = "Looking for disproportions in Churn
        fill = "Churner")
```

```
## Warning: Ignoring unknown parameters: binwidth, bins, pad
```



3.f ### CUST_MOS

```r
cust_mos_skew <- (3 * (mean(phones$CUST_MOS) - median(phones$CUST_MOS)))/sd(phones$CUST_MOS)
cust_mos_skew
```

```
## [1] 1.131224
```

NUM_LINES

```r
num_lines_skew <- (3 * (mean(phones$NUM_LINES) - median(phones$NUM_LINES)))/sd(phones$NUM_LINES)
num_lines_skew
```

```
## [1] 2.057503
```

TOT_MINUTES_USAGE

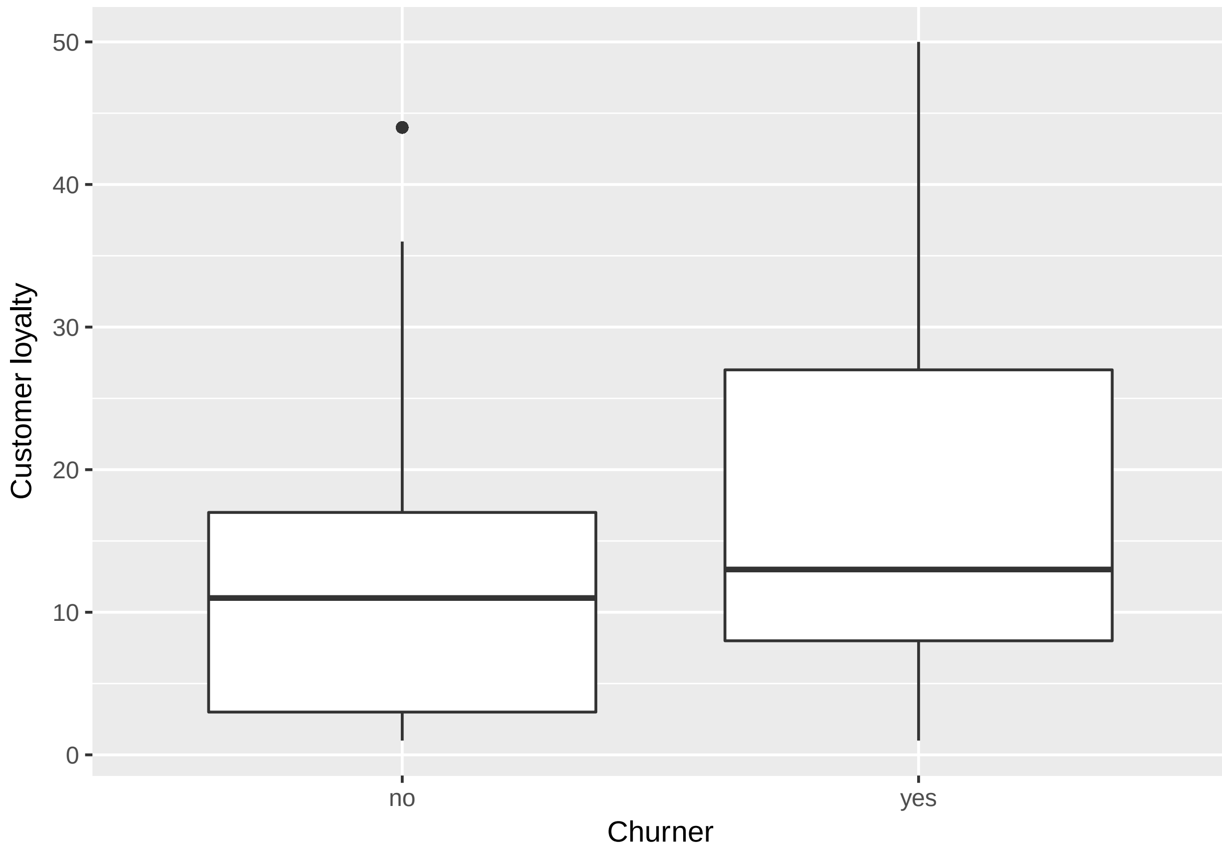
```r
tot_minutes_usage_skew <- (3 * (mean(phones$TOT_MINUTES_USAGE) -
    median(phones$TOT_MINUTES_USAGE)))/sd(phones$TOT_MINUTES_USAGE)
tot_minutes_usage_skew
```

```
## [1] 1.088757
```

3.g

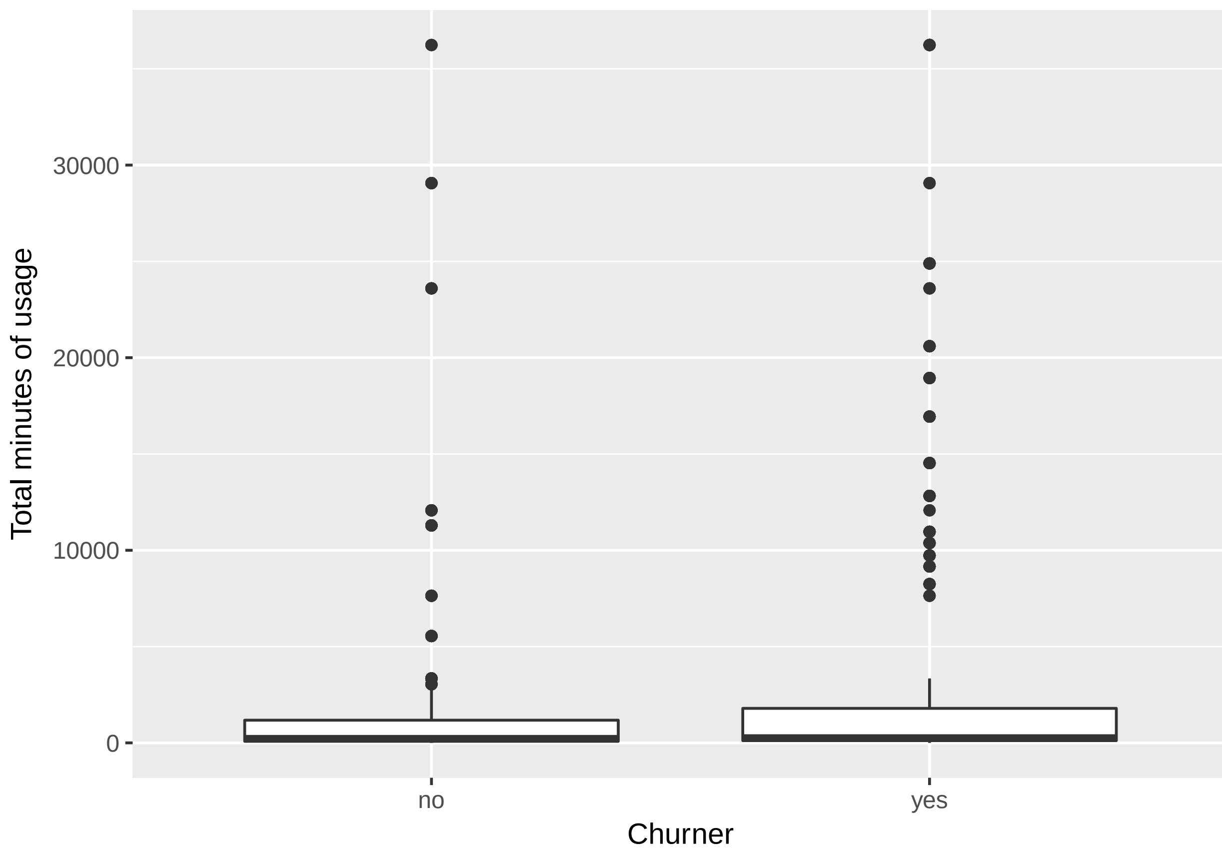
CUST_MOS

```r
ggplot(data = phones, aes(phones$CHURNER, phones$CUST_MOS)) +
    geom_boxplot() + labs(x = "Churner", y = "Customer loyalty")
```



TOT_MINUTES_USAGE

```r
ggplot(data = phones, aes(phones$CHURNER, phones$TOT_MINUTES_USAGE)) +
    geom_boxplot() + labs(x = "Churner", y = "Total minutes of usage")
```



4. Finding outliers mathematically in TOT_MINUTES_USAGE

IQR method

```
summary(phones$TOT_MINUTES_USAGE)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##       0     116     264    2036    1677   36240
```

```
IQR <- 1677 - 116
lower_bound <- 116 - (IQR * 1.5)
upper_bound <- 1677 + (IQR * 1.5)

nrow(phones[phones$TOT_MINUTES_USAGE < lower_bound | phones$TOT_MINUTES_USAGE >
    upper_bound, ])
```

```
## [1] 176
```

Z standardisation method

```
# Z score
z_score_tot_minutes_usage <- scale(phones$TOT_MINUTES_USAGE,
    center = TRUE, scale = TRUE)
# same as (phones$TOT_MINUTES_USAGE -
# mean(phones$TOT_MINUTES_USAGE))/sd(phones$TOT_MINUTES_USAGE)
summary(z_score_tot_minutes_usage)
```

```
##        V1
##  Min.   :-0.41698
##  1st Qu.:-0.39323
```

```
##  Median :-0.36292
##  Mean   : 0.00000
##  3rd Qu.:-0.07354
##  Max.   : 7.00417
```

```r
z_range <- table(z_score_tot_minutes_usage > -3 & z_score_tot_minutes_usage <
    3)
z_range[names(z_range) == FALSE]
```

```
## FALSE
##    69
```

5

```r
tot_mins_before_transfo <- (3 * (mean(phones$TOT_MINUTES_USAGE) -
    median(phones$TOT_MINUTES_USAGE)))/sd(phones$TOT_MINUTES_USAGE)
tot_mins_before_transfo
```

```
## [1] 1.088757
```

5.a

Z-score standardisation see above, we reduced the number of outliers from 176 to 69

```r
tot_mins_z_score <- (3 * (mean(z_score_tot_minutes_usage) - median(z_score_tot_minutes_usage)))/sd(z_sco
tot_mins_z_score
```

```
## [1] 1.088757
```

5.b

Natural log

```r
natural_log_transfo <- log(phones$TOT_MINUTES_USAGE[phones$TOT_MINUTES_USAGE !=
    0])
natural_log_transfo_skewness <- (3 * (mean(natural_log_transfo) -
    median(natural_log_transfo)))/sd(natural_log_transfo)

natural_log_transfo_skewness
```

```
## [1] -0.7042918
```

5.c

Square root

```r
square_root_transfo <- sqrt(phones$TOT_MINUTES_USAGE)
square_root_transfo_skewness <- (3 * (mean(square_root_transfo) -
    median(square_root_transfo)))/sd(square_root_transfo)

square_root_transfo_skewness
```

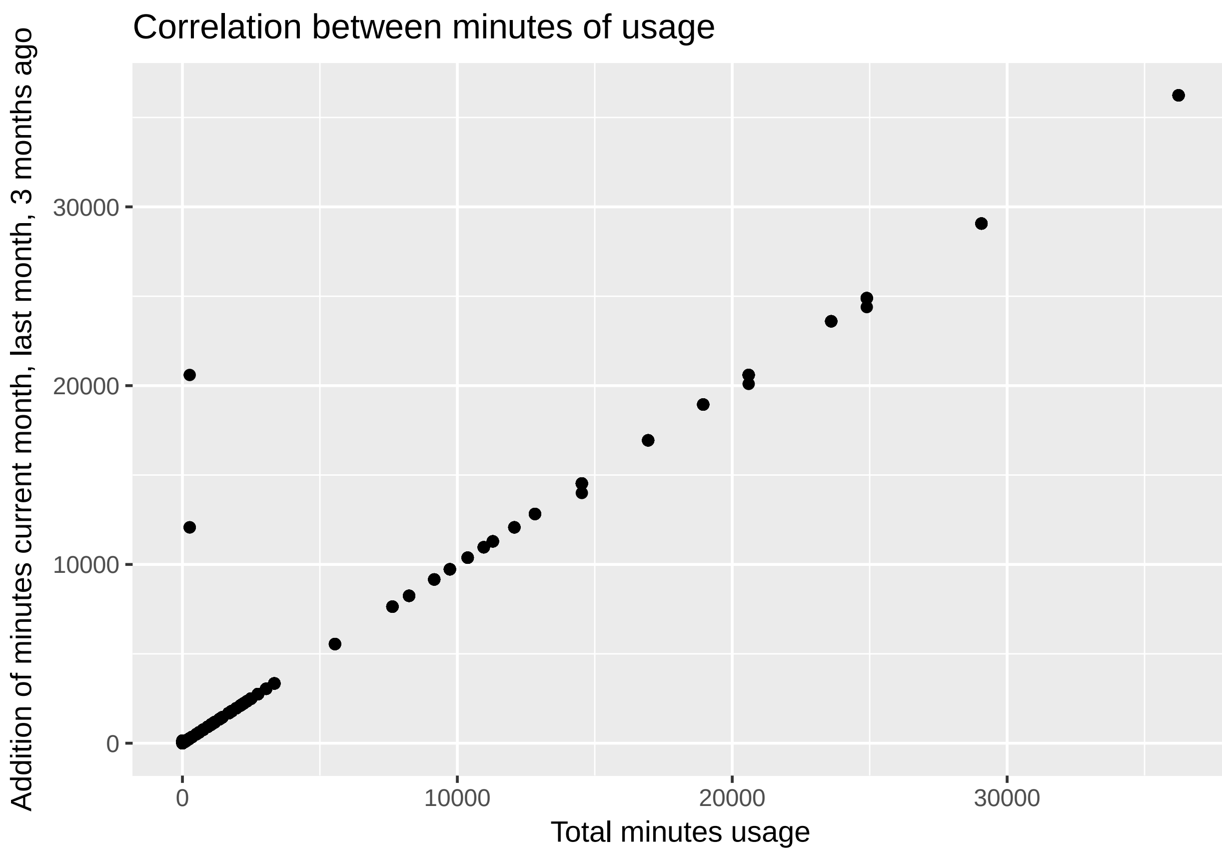
```
## [1] 1.288432
```

7.a.

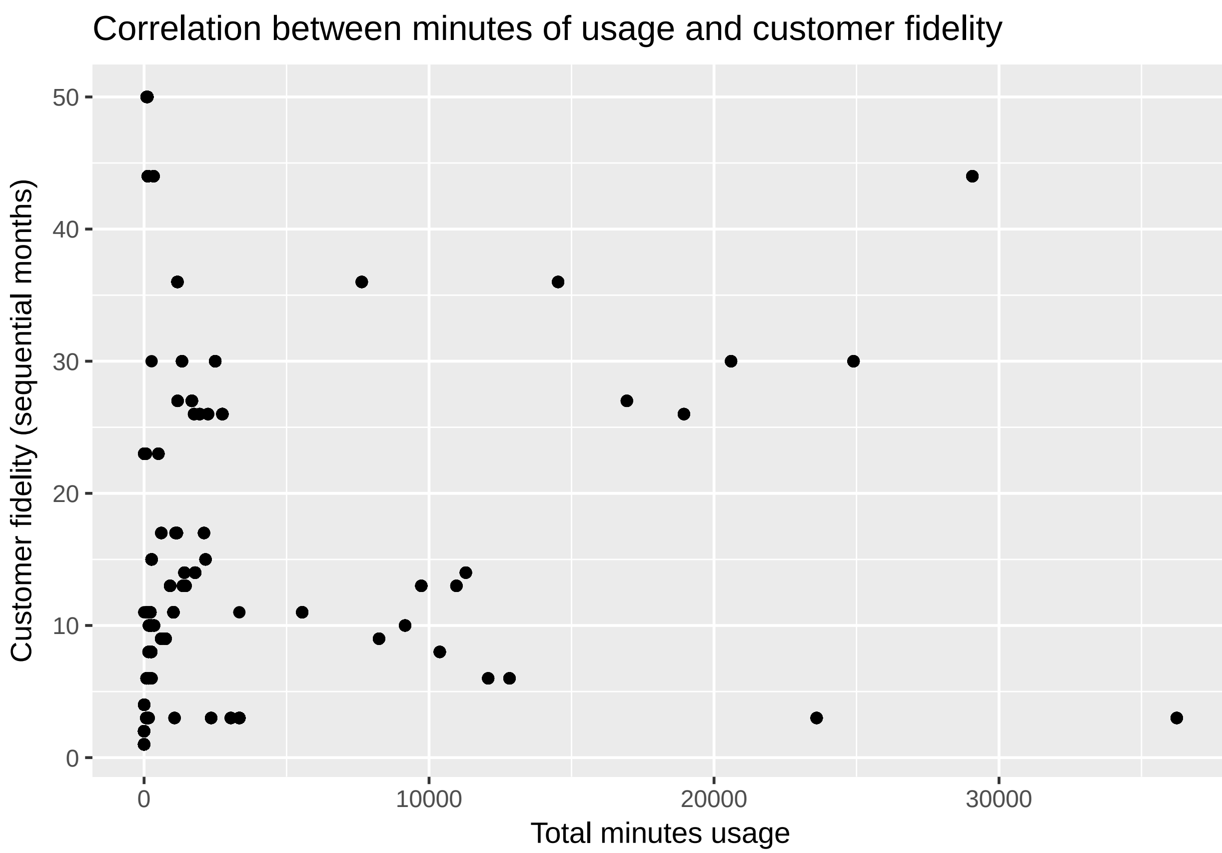
Correlation

Minutes of usage

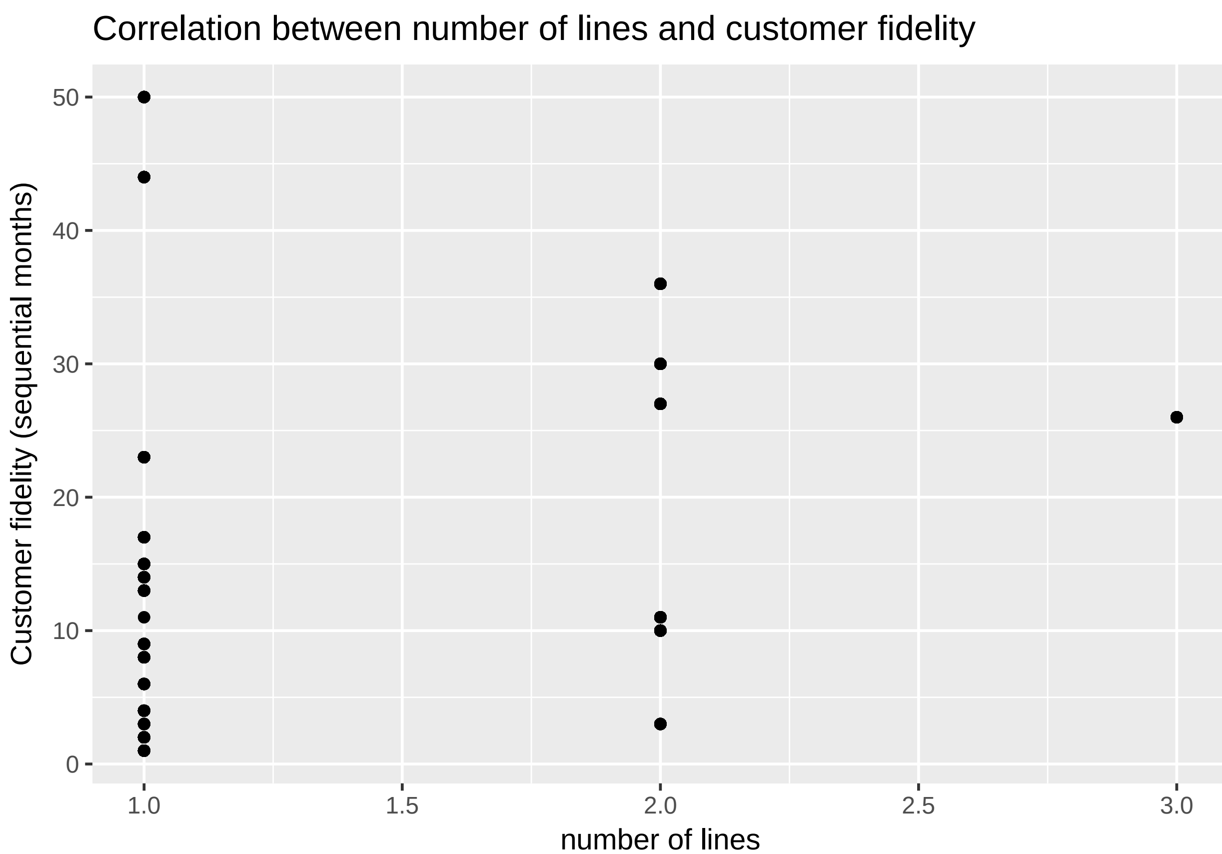
```
ggplot(data = phones, aes(x = phones$TOT_MINUTES_USAGE, y = phones$MINUTES_CURR_MONTH +
    phones$MINUTES_PREV_MONTH + phones$MINUTES_3MONTHS_AGO)) +
    geom_point() + labs(x = "Total minutes usage", y = "Addition of minutes current month, last month, 3
    title = "Correlation between minutes of usage")
```



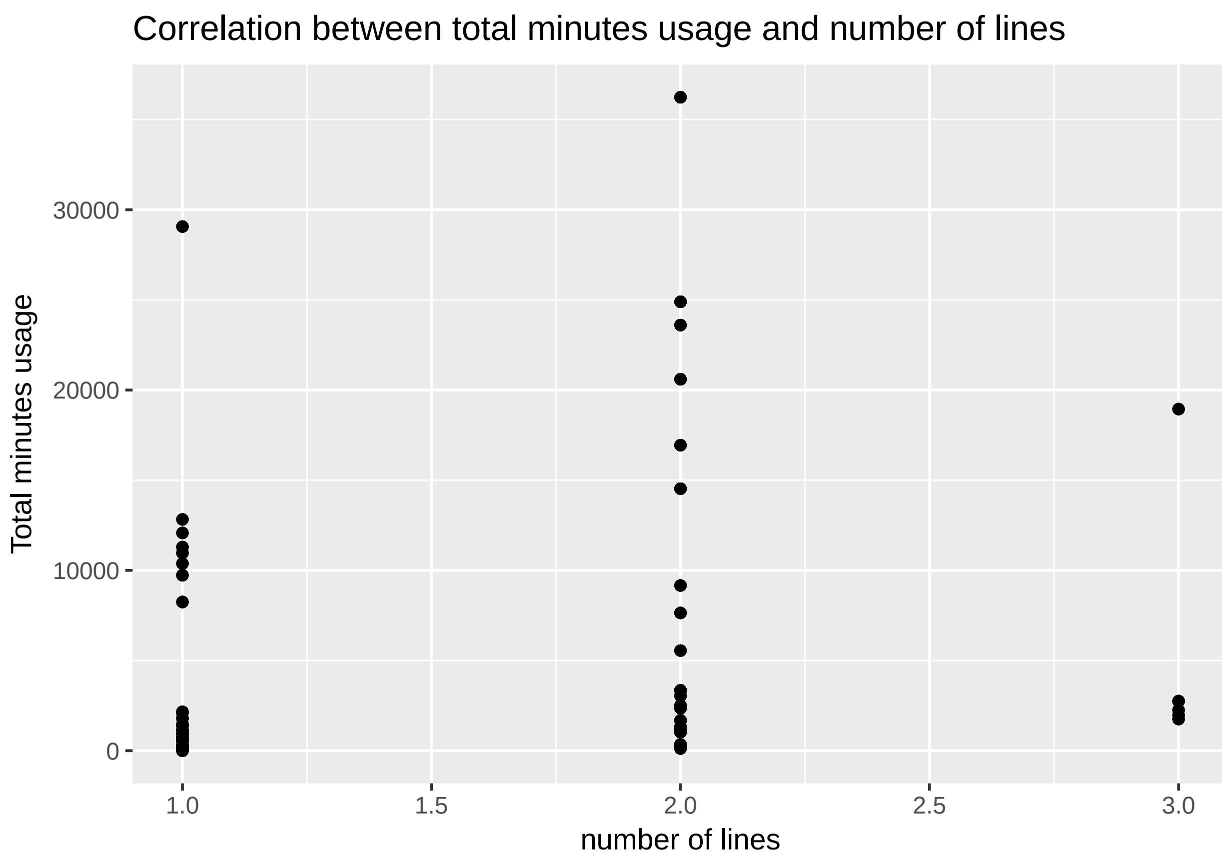

```
ggplot(data = phones, aes(x = phones$TOT_MINUTES_USAGE, y = phones$CUST_MOS)) +
    geom_point() + labs(y = "Customer fidelity (sequential months)",
    x = "Total minutes usage", title = "Correlation between minutes of usage and customer fidelity")
```




```r
ggplot(data = phones, aes(x = phones$NUM_LINES, y = phones$CUST_MOS)) +
    geom_point() + labs(y = "Customer fidelity (sequential months)",
    x = "number of lines", title = "Correlation between number of lines and customer fidelity")
```




```
ggplot(data = phones, aes(x = phones$NUM_LINES, y = phones$TOT_MINUTES_USAGE)) +
    geom_point() + labs(y = "Total minutes usage", x = "number of lines",
    title = "Correlation between total minutes usage and number of lines")
```



7.b

Minutes usage metrics correlation

```
covariance_minutes <- cov(phones$TOT_MINUTES_USAGE, phones$MINUTES_CURR_MONTH +
    phones$MINUTES_PREV_MONTH + phones$MINUTES_3MONTHS_AGO)
covariance_minutes
```

```
## [1] 23778254
```

```
correlation_minutes <- covariance_minutes/(sd(phones$TOT_MINUTES_USAGE) *
    sd(phones$MINUTES_CURR_MONTH + phones$MINUTES_PREV_MONTH +
        phones$MINUTES_3MONTHS_AGO))
correlation_minutes
```

```
## [1] 0.9916396
```

Usage and customer fidelity

```
covariance_minutes_fid <- cov(phones$TOT_MINUTES_USAGE, phones$CUST_MOS)
covariance_minutes_fid
```

```
## [1] 5931.69
```

```
correlation_minutes_fid <- covariance_minutes_fid/(sd(phones$TOT_MINUTES_USAGE) *
    sd(phones$CUST_MOS))
correlation_minutes_fid
```

```
## [1] 0.09075367
```

```r
covariance_lines_fid <- cov(phones$NUM_LINES, phones$CUST_MOS)
covariance_lines_fid
```

```
## [1] 1.550566
```

```r
correlation_lines_fid <- covariance_lines_fid/(sd(phones$NUM_LINES) *
    sd(phones$CUST_MOS))
correlation_lines_fid
```

```
## [1] 0.2031285
```

```r
covariance_lines_minutes <- cov(phones$NUM_LINES, phones$TOT_MINUTES_USAGE)
covariance_lines_minutes
```

```
## [1] 685.4576
```

```r
correlation_lines_minutes <- covariance_lines_minutes/(sd(phones$NUM_LINES) *
    sd(phones$TOT_MINUTES_USAGE))
correlation_lines_minutes
```

```
## [1] 0.2461581
```

Part 2

Preparating Data for learning

```r
keep <- c("INCOME", "PHONE_PLAN", "EDUCATION", "AREA_CODE", "CUS_MOS",
    "CHURNER", "CONVERGENT_BILLING")
phones_learning <- phones[, (names(phones) %in% keep)]
phones_learning$AREA_CODE <- as.factor(phones_learning$AREA_CODE)
head(phones_learning)
```

```
##   AREA_CODE CONVERGENT_BILLING        INCOME    PHONE_PLAN    EDUCATION
## 1     45987                Yes Medium Income International      Masters
## 2     15563                Yes Medium Income International    Bachelors
## 3     10040                 No    Low Income      National  High School
## 4     21750                Yes Medium Income International  High School
## 5     55166                 No   High Income    Promo_plan  High School
## 6     36785                 No   High Income      National Post Primary
##   CHURNER
## 1     yes
## 2      no
## 3      no
## 4     yes
## 5      no
## 6      no
```

Writing the learning data to csv

```r
write.csv(phones_learning, file = "./learning_churners.csv")
```