Enterprise Database Technology CA1

Shajun Dominic

X00109340

Table of Contents

[Section 1 2](#_Toc478341629)

[1) Data Processing 2](#_Toc478341630)

[2) Discretizing Income 2](#_Toc478341631)

[3) Information Assessment 2](#_Toc478341632)

[4) Outliers 5](#_Toc478341633)

[5) Skewness 5](#_Toc478341634)

[6) Histogram and overlay of numeric variable 5](#_Toc478341635)

[7) Correlated Variables 6](#_Toc478341636)

[Section 2 8](#_Toc478341637)

[PART 8](#_Toc478341638)

[JRip 9](#_Toc478341639)

[J48 10](#_Toc478341640)

[Overall Assessment 12](#_Toc478341641)

[Appendix 13](#_Toc478341642)

[Section 1 13](#_Toc478341643)

[Q1 – Data Processing 13](#_Toc478341644)

[Q2 – Discretizing Income 13](#_Toc478341645)

[Q3 – Information Assessment 14](#_Toc478341646)

[Q4 - Outliers 17](#_Toc478341647)

[Q5 - Skewness 17](#_Toc478341648)

[Q7 – Correlated Variables 18](#_Toc478341649)

[Section 2 19](#_Toc478341650)

# Section 1

## 1) Data Processing

First of all, I got a count list of values from the data frame that had null values or where there was empty data. This showed me how many columns had null values. From this we can see that there was only few empty values, so our data is of good quality.

Then I replaced the null values in CUS\_MOS, MINUTES\_3MONTHS\_AGO and TOT\_MINUTES\_USAG with the median values for each of the columns specified above.

For missing categorical values, they were replaced by the mode value by gender for that predictor variable.

The mode value for PHONE\_PLAN is "International".

## 2) Discretizing Income

I added a new column into my data frame in order for me to compare if the discretization had worked out according to my calculation. The lower range income value ends at 37,999 since in the document it said values less than 38000 should be called less income.

## 3) Information Assessment

a) The attribute type for each predictor variable has been found and has been listed out in the table below.

b) There was no duplicated data. But multiple CUST\_ID which meant that the customer left and came back. Overall there was no duplicated data.

c) A mode function was created in order to get mode for some predictor variable and summary function was also used to get details for each predictor variable.

For parts a, c, d, e, f, g please look at the table below.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **ATTRIBUTE** | **MAX** | **MIN** | **MEAN** | **MODE** | **MEDIAN** | **Standard Deviation** | **PART D** | **PART E** |
| **AREA\_CODE** | Nominal |  |  |  | 10040 |  | 15893.23 |  |  |
| **MINUTES\_CURR\_MONTH** | Numeric | 14000 | 1.0 | 747 | 2 | 105.0 | 2017.135 |  |  |
| **MINUTES\_PREV\_MONTH** | Numeric | 16754 | 0.0 | 863.9 | 0 | 98.0 | 2468.028 |  |  |
| **MINUTES\_3MONTHS\_AGO** | Numeric | 12456 | 0.0 | 452.5 | 0 | 97.0 | 1183.944 |  |  |
| **CUST\_MOS** | Numeric | 50 | 1 | 16.05 | 11 | 11 | 13.38527 | Lots of customers stay between 5 and 15 months | After 10 months, the customers tend to churn faster |
| **LONGDIST\_FLAG** | Nominal |  |  |  | 1 |  |  |  |  |
| **CALLWAITING\_FLAG** | Nominal | 1 | 0 | 0.4346 | 0 | 0 | 0.4958205 |  |  |
| **NUM\_LINES** | Numeric | 3 | 1 | 1.391 | 1 | 1 | 0.5702769 | Majority of users tend to have only one connection | N/A |
| **VOICEMAIL\_FLAG** | Nominal |  |  |  | 1 |  |  |  |  |
| **MOBILE\_PLAN** | Nominal | 1 | 0 | 0.3477 | No | 0 | 0.4763418 |  |  |
| **CONVERGENT\_BILLING** | Nominal |  |  |  | No |  |  |  |  |
| **GENDER** | Nominal |  |  |  | M |  |  |  | Equal Numbers of people churn and not churn |
| **INCOME** | Numeric | 320000 | 17000 | 85784 | 75000, 80000 | 75000 | 66740.6 | Majority of people have medium income, which is followed by high income and hen low income | People with medium income tends to churn more when compared with low and high income |
| **PHONE\_PLAN** | Nominal |  |  |  | International |  |  |  | People with International and National phone plans tend to churn more |
| **EDUCATION** | Nominal |  |  |  | Post Primary |  |  |  | People in post primary and PhD tend to have the highest churning rate while the primary has no churning |
| **TOT\_MINUTES\_USAGE** | Numeric | 36237 | 0 | 2036 | 0 | 264 | 4883.004 | Most of the users tend to use less than 3000 minutes | No useful information |
| **CUST\_ID** | Ordinal | 2070 | 1 | 1035.1 | 246 | 1035 | 597.8092 |  |  |

|  |
| --- |
| Creating graphic visualisation allowed me to distinguish between the users or people in our context. The bar charts allowed me to see that lots of customers tend to stay between 5 and 15 months on a particular plan. I also found out that people on international and national plans tend to churn more when compared with the rest of the plans.  It also allowed me to see that an equal number in both gender churn and not churn. Also, I was able to see that majority of people who had low and high income churn more than the people with medium income. |

|  |  |  |
| --- | --- | --- |
|  | **PART F** | **PART G** |
| **AREA\_CODE** |  |  |
| **MINUTES\_CURR\_MONTH** |  |  |
| **MINUTES\_PREV\_MONTH** |  |  |
| **MINUTES\_3MONTHS\_AGO** |  |  |
| **CUST\_MOS** | Skewness: 1.061088  Positively skewed to the right | Few outliers have been found |
| **LONGDIST\_FLAG** |  |  |
| **CALLWAITING\_FLAG** |  |  |
| **NUM\_LINES** | Skewness: 1.136485  Positively skewed to the right |  |
| **VOICEMAIL\_FLAG** |  |  |
| **MOBILE\_PLAN** |  |  |
| **CONVERGENT\_BILLING** |  |  |
| **GENDER** |  |  |
| **INCOME** |  |  |
| **PHONE\_PLAN** |  |  |
| **EDUCATION** |  |  |
| **TOT\_MINUTES\_USAGE** | Skewness: 1.087828  Positively skewed to the right | An outlier has been found |
| **CUST\_ID** |  |  |

## 4) Outliers

IQR:

A huge number of outliers has been found in TOT\_MINUTES\_USAGE. So, I decided to work with that predictor variable. After getting the upper bound and lower bound using IQR method I found 176 outliers.

Z Score:

The Z-Score value was calculated but was unable to find any outliers. (Code in appendix)

## 5) Skewness

http://growingknowing.com/GKStatsBookSkewness.php

When we created the graph for TOT\_MINUTES\_USAGE we also calculated the skewness. We know that the skewness value is 1.087282 and that the graph is positively skewed to the right.

Z-Score Standardisation:

Skewness on the z-score value returned the same value as the skewness on TOT\_MINUTES\_USAGE.

Natural Log:

The skewness of TOT\_MINUTES\_USAGE is -0.53953. It indicates that the TOT\_MINUTES\_USAGE distribution is skewed towards the left.

Square Root:

Skewness on the square root was greater than the skewness on TOT\_MINUTES\_UAGE. Square root skewness returned 1.286773. Therefore, this information is not useful on TOT\_MINUTES\_USAGE.

## 6) Histogram and overlay of numeric variable

Since I had already used histograms for creating graphs in part 3 I'll be using those graphs in this analysis section. I'll be mainly focusing on the churning effects from the histograms because it allows us to see how the churning affects are between different categories of people.

a)

People in different phone plans have different churning rates. The majority of people in International and national plans have the highest churn rates. While the Euro zone and promo plan have the least churn rates and least number of people.

For income, we can see from the graph that people who have medium income tend to churn less when compared with low and high income. Also, we can see that lots of people tend to leave between 10 on 15 months. For example, this could be due to their contract being finished after 12 months’ contract. But still many people stay longer than 15 months.

Variable that have no impact on churning rates is gender. Gender seems to be a situation where 50% stay and the other 50% leave for both male and female.

b)

The variables that I expect to make a significant appearance in any data mining classification model is income and phone plan. As we can see from income graph that the majority of medium income sty while the others leave. Phone plan also shows us how people in different plans tend to leave more than the other plans.

## 7) Correlated Variables

a)

TOT\_MINUTES\_USAGE with MINUTES\_CURR\_MONTH looks to be correlated from the graph. The scatter plot look to have a positively linear scatter plot.

b)

Using the method cor() to find the correlation coefficient value, I found out that it returned 0.8844 which is close to 1. This indicates that the variables are positively linearly related and the scatter plot falls almost along a straight line with positive slope. Using cov() covariance method I found out that it also returned a positive value which also indicates that the scatter plot is positively linearly related. The other 3 correlation a covariance checking yielded no useful information.

c)

|  |  |
| --- | --- |
| Attributes that influence churning rate | Attributes that have no influence in churning rate |
| AREA\_CODE | GENDER |
| EDUCATION | NUM\_LINES |
| INCOME |  |
| PHONE\_PLAN |  |
| CONVERGENT\_BILLING |  |

d)

The variables that can be eliminated from the dataset are MINUTES\_3MONTHS\_AGO, MINUTES\_CURR\_MONTH and MINUTES\_PREV\_MONTH. These variables should be eliminated because there is another variable TOT\_MINUTES\_USAGE which contains the total of all these values.

Also from the above table we can see that GENDER and NUM\_LINES have no influence in the churning rate of customers. The decision tree will be less is one of the benefits.

# Section 2

Explanation of classifiers:

**PART** - Builds a partial C4.5 decision tree in each iteration and makes the "best" leaf into a rule

**JRip** - Creates Rules by combining prediction variables

**J48** - Creates a decision tree

|  |  |
| --- | --- |
|  | PART |
| **Parameter** | CHURNER |
| **How does it decide if it is a churner or not?** | Groups multiple attributes and compares and tests them to see if they are correlated. |
| **Does it make sense to you?** | Yes, because by looking rules in the decision list we can see the list of rules that tell us why the users have left the telecom. |
| **Key predictors of churn** | INCOME, CUST\_MOS, AREA\_CODE |
| **Significant rules/ decision tree path** | CUST\_MOS > 44: yes (90.0) |
| **Proportion of false positives** | 0.0965 |
| **Proportion of false negatives** | 0.2314 |
| **Overall error rate and overall model accuracy** | The accuracy of correctly classified instances is: 81.8%  The overall error state is: 18.2% |
| **Precision** | 0.903 |
| **Sensitivity(Recall) - True Positive Rate** | 0.694 |
| **Specificity - False Positive Rate** | 0.9319 |
| **ROC** | 0.863 |
| **FP** | 25 |
| **FN** | 103 |
| **TN** | 342 |
| **TP** | 234 |

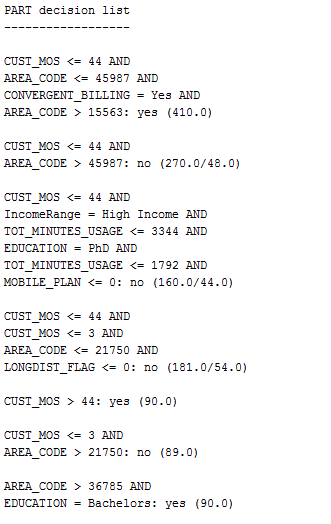


Figure 1 PART DECISION LIST

|  |  |
| --- | --- |
|  | JRip |
| **Parameter** | CHURNER |
| **How does it decide if it is a churner or not?** | It predicts the mean for numeric class and mode for nominal class. |
| **Does it make sense to you?** | Yes, because the 16 rules allow us to see for what values the customer is likely to leave the telecom. |
| **Key predictors of churn** | INCOME, AREA\_CODE, |
| **Significant rules/ decision tree path** | (INCOME >= 75000)  and (LONGDIST\_FLAG  >=1) => CHURNER =no |
| **Proportion of false positives** | 0.0382 |
| **Proportion of false negatives** | 0.2367 |
| **Overall error rate and overall model accuracy** | The accuracy of correctly classified instances is: 82.9%  The overall error state is: 17.0% |
| **Precision** | 0.962 |
| **Sensitivity(Recall)-True Positive Rate** | 0.6706 |
| **Specificity - False Positive Rate** | 0.9755 |
| **ROC** | 0.867 |
| **FP** | 9 |
| **FN** | 111 |
| **TN** | 358 |
| **TP** | 226 |

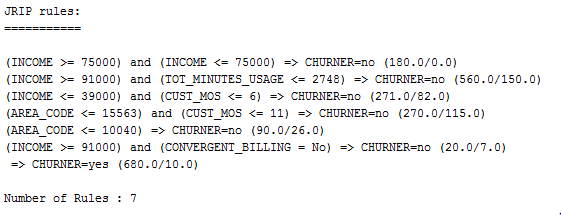


Figure 2 JRIP RULES

|  |  |
| --- | --- |
|  | J48 |
| **Parameter** | CHURNER |
| **How does it decide if it is a churner or not?** | It creates a decision tree based on the file read in and creates a tree path of what variables makes customers churn |
| **Does it make sense to you?** | Yes, because we can follow the tree path and see for what specific values does the customer churn from the telecom |
| **Key predictors of churn** | INCOME, AREA\_CODE, TOT\_MINUTES\_USAGE |
| **Significant rules/ decision tree path** | INCOME = HIGH INCOME & TOT\_MINUTES\_USAFE <= 1792: yes |
| **Proportion of false positives** | 0.0797 |
| **Proportion of false negatives** | 0.1217 |
| **Overall error rate and overall model accuracy** | The accuracy of correctly classified instances is: 82.1%  The overall error state is: 17.9% |
| **Precision** | 0.920 |
| **Sensitivity(Recall) - True Positive Rate** | 0.6855 |
| **Specificity - False Positive Rate** | 0.9745 |
| ROC | 0.857 |
| FP | 20 |
| FN | 106 |
| TN | 765 |
| TP | 231 |

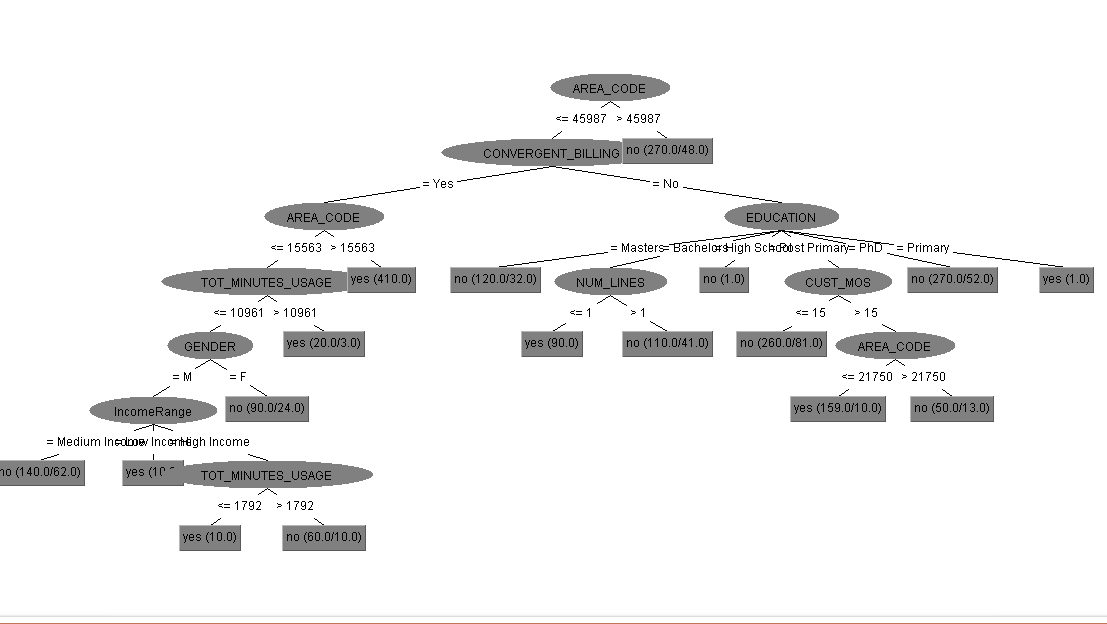
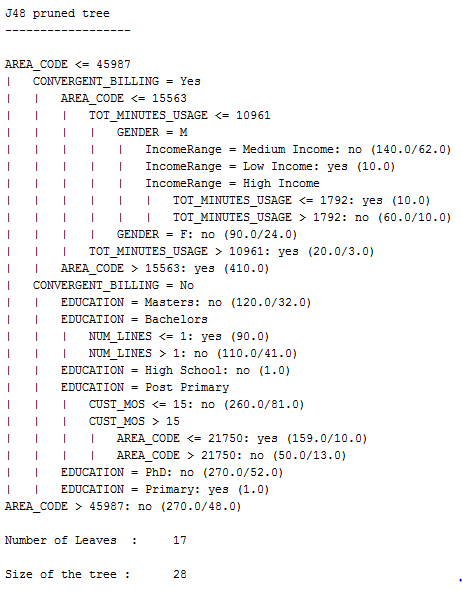


Figure J48 Tree



Using other predictor variables as the parameters yielded no outcome to show if the customer will churn or not.

## Overall Assessment

1) From this data mining, we get an information that certain areas are more prone to churning than others. Income is also a factor for churning because people with low, medium or high income might not need the plan they are currently on. It could be due to less call duration and the area they are in could have low service or coverage.

2)

People whose education is post primary and have been staying with a telecom for more than 15 months and is in an area code below 21750 churn more according to the J48 pruned tree.

People whose gender is male and education is Bachelors churn more according to the PART classifier.

3)

More customer feedback would be a great way in order for a company to know the improvements that they should make. More offers would attract more customers and also allows the current customers to switch to different mobile plans that suit their needs. The customers that are in the categories specified above should be monitored closely.

# Appendix

## Section 1

### Q1 – Data Processing

//Read in the csv file

phonecsv <- read.csv(file = "C:/EDT CA/phone.csv")

//LIST of all columns with null values

list <- data.frame(lapply(phonecsv, function(DATA) sum(length(which(is.na(DATA) | DATA == "")))))

//MEDIAN - MINUTES\_3\_MONTHS\_AGO

median(phonecsv$MINUTES\_3MONTHS\_AGO, na.rm = TRUE)

median\_3monthsago <- median(phonecsv$MINUTES\_3MONTHS\_AGO, na.rm = TRUE)

is.na(phonecsv$MINUTES\_3MONTHS\_AGO) = median\_3monthsago

//MEDIAN - CUST\_MOS

median(phonecsv$CUST\_MOS, na.rm = TRUE)

median\_custmos <- median(phonecsv$CUST\_MOS, na.rm = TRUE)

phonecsv$MINUTES\_3MONTHS\_AGO[is.na(phonecsv$MINUTES\_3MONTHS\_AGO)] = median3MonthsAgo

//MEDIAN - TOT\_MINUTES\_USAGE

median\_totminutesusage <- median(phonecsv$TOT\_MINUTES\_USAGE, na.rm = TRUE)

phonecsv$TOT\_MINUTES\_USAGE[is.na(phonecsv$TOT\_MINUTES\_USAGE)] = medianMinutesUsage

//MODE - PHONE\_PLAN

modePhone <- (table(phonecsv$PHONE\_PLAN)) == max(table(phonecsv$PHONE\_PLAN))

modePhone <- names(table(phonecsv$PHONE\_PLAN))[modePhone]

phonecsv$PHONE\_PLAN[phonecsv$PHONE\_PLAN == ""] <- modePhone

//MODE - EDUCATION

modeEducation <- (table(phonecsv$EDUCATION)) == max(table(phonecsv$EDUCATION))

modeEducation <- names(table(phonecsv$EDUCATION))[modeEducation]

phonecsv$EDUCATION[phonecsv$EDUCATION == ""] <- modeEducation

### Q2 – Discretizing Income

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

### Q3 – Information Assessment

B)

anyDuplicated(phonecsv)

[1] 0

> anyDuplicated(phonecsv$CUST\_ID)

[1] 152

c)

//Get Mode function

getMode <- function(MODE){

modevalues <- table(MODE) == max(table(MODE))

return(names(table(MODE))[modevalues])

}

summary(phonecsv)

getMode(phonecsv$AREA\_CODE)

getMode(phonecsv$MINUTES\_CURR\_MONTH)

getMode(phonecsv$MINUTES\_PREV\_MONTH)

getMode(phonecsv$MINUTES\_3MONTHS\_AGO)

getMode(phonecsv$CUST\_MOS)

getMode(phonecsv$LONGDIST\_FLAG)

getMode(phonecsv$CALLWAITING\_FLAG)

getMode(phonecsv$NUM\_LINES)

getMode(phonecsv$VOICEMAIL\_FLAG)

getMode(phonecsv$MOBILE\_PLAN)

getMode(phonecsv$CONVERGENT\_BILLING)

getMode(phonecsv$GENDER)

getMode(phonecsv$INCOME)

getMode(phonecsv$PHONEPLAN)

getMode(phonecsv$EDUCATION)

getMode(phonecsv$TOT\_MINUTES\_USAGE)

getMode(phonecsv$CUST\_ID)

sd(phonecsv$AREA\_CODE)

sd(phonecsv$MINUTES\_CURR\_MONTH)

sd(phonecsv$MINUTES\_PREV\_MONTH)

sd(phonecsv$MINUTES\_3MONTHS\_AGO)

sd(phonecsv$CUST\_MOS)

sd(phonecsv$LONGDIST\_FLAG)

sd(phonecsv$CALLWAITING\_FLAG)

sd(phonecsv$NUM\_LINES)

sd(phonecsv$VOICEMAIL\_FLAG)

sd(phonecsv$MOBILE\_PLAN)

sd(phonecsv$CONVERGENT\_BILLING)

sd(phonecsv$GENDER)

sd(phonecsv$INCOME)

sd(phonecsv$PHONEPLAN)

sd(phonecsv$EDUCATION)

sd(phonecsv$TOT\_MINUTES\_USAGE)

sd(phonecsv$CUST\_ID)

d)

**INCOME**

p <- ggplot(phonecsv ,aes(phonecsv$IncomeRange)) + geom\_histogram(stat="count", col="red", fill="green") + labs(x="Income" ,y="Number of People") + theme\_light() + labs(title="Histogram for Income")

**TOTAL\_MINUTESUSAGE**

ggplot(data=phonecsv,aes(phonecsv$TOT\_MINUTES\_USAGE))+scale\_x\_continuous(breaks = seq(0, 35000, 3000)) + geom\_histogram(bins = 30) + labs(x="Minutes Used", y="No of People")

**PHONE\_PLAN**

ggplot(phonecsv,aes(phonecsv$PHONE\_PLAN)) + geom\_histogram(stat="count") + labs(x="Phone plan" ,y="Number of People")

**NUM\_LINES**

ggplot(phonecsv ,aes(phonecsv$NUM\_LINES)) + geom\_histogram(stat="count") + labs(x="Number of Connections" ,y="Number of People")

**CUST\_MOS**

ggplot(phonecsv ,aes(phonecsv$CUST\_MOS)) + geom\_histogram(binwidth=8) + scale\_x\_continuous(breaks = seq(0,60,5)) + labs(x="Customer Loyalty (Months)" ,y="Number of persons") + labs(title="Histogram For Customer Loyalty")

**MINUTES PREVIOUS MONTH**

ggplot(phonecsv ,aes(phonecsv$MINUTES\_PREV\_MONTH)) + geom\_histogram(fill="blue") + labs(x="Number of Minutes" ,y="Number of People") + labs(title="Histogram For Prev Month Minutes") + scale\_x\_continuous(breaks = seq(0,20000,2000))

**Minutes Curr Month**

ggplot(phonecsv ,aes(phonecsv$MINUTES\_CURR\_MONTH)) + geom\_histogram(fill="orange") + labs(x="Number of Minutes" ,y="Number of People") + labs(title="Histogram For Current Month Minutes") + scale\_x\_continuous(breaks = seq(0,20000,2000))

F)

skewness(phonecsv$CUST\_MOS)

skewness(phonecsv$NUM\_LINES)

skewness(phonecsv$TOT\_MINUTES\_USAGE)

E)

**INCOME**

ggplot(data=phonecsv ,aes(x=phonecsv$INCOME, group=phonecsv$CHURNER, fill=phonecsv$CHURNER)) + geom\_histogram(stat="count") + labs(x="Income" ,y="Number of People", title="Churning Rates Based on Income", fill="CHURNER") + theme\_light()

**PHONE\_PLAN**

ggplot(data=phonecsv ,aes(x=phonecsv$PHONE\_PLAN, group=phonecsv$CHURNER, fill=phonecsv$CHURNER)) + geom\_histogram(stat="count") + labs(x="Phone plans" ,y="Number of People", title="Churning Rate of Phone Plan", fill="CHURNER") + theme\_light()

**TOT\_MINUTES\_USAGE**

ggplot(phonecsv ,aes(x=phonecsv$TOT\_MINUTES\_USAGE, group=phonecsv$CHURNER, fill=phonecsv$CHURNER)) + geom\_histogram(bins = 7) + labs(x="Total Minutes Used" ,y="Number of People", title="Churning Rate", fill="CHURNER")+ theme\_light() + scale\_x\_continuous(breaks = seq(0,40000,5000))

**CUST\_MOS**

ggplot(phonecsv ,aes(x=phonecsv$CUST\_MOS, group=phonecsv$CHURNER, fill=phonecsv$CHURNER)) + geom\_histogram(binwidth=5) + scale\_y\_continuous(breaks = seq(0,500,50)) + scale\_x\_continuous(breaks = seq(0,50,5))+ labs(x="Months(CUST\_MOS)" ,y="Number of People", title="Churning Rate Per Month", fill="CHURNER")+ theme\_light()

**EDUCATION**

ggplot(data=phonecsv ,aes(x=phonecsv$EDUCATION, group=phonecsv$CHURNER, fill=phonecsv$CHURNER)) + geom\_histogram(stat="count") + labs(x="Education" ,y="Number of People", title="Churning Rate", fill="CHURNER")+ theme\_light()

**GENDER**

ggplot(data=phonecsv ,aes(x=phonecsv$GENDER, group=phonecsv$CHURNER, fill=phonecsv$CHURNER)) + geom\_histogram(stat="count") + labs(x="Gender" ,y="Number of People", title="Churning Rates", fill="CHURNER") + theme\_light()

G)

ggplot(data=phonecsv ,aes(phonecsv$CHURNER, phonecsv$CUST\_MOS)) + geom\_boxplot() + labs(x="Churner" ,y="Customer Duration (Months)")

ggplot(data=phonecsv ,aes(phonecsv$CHURNER, phonecsv$TOT\_MINUTES\_USAGE)) + geom\_boxplot() + labs(x="Churner" ,y="ToTal Minutes")

ggplot(data=phonecsv ,aes(phonecsv$CHURNER, phonecsv$PHONE\_PLAN)) + geom\_boxplot() + labs(x="Churner" ,y="Plans")

### Q4 - Outliers

**IQR**

|  |
| --- |
| totalminsusage\_IQR <- IQR(phonecsv$TOT\_MINUTES\_USAGE)  lowerbound <- 116 - (totalminsusage\_IQR \* 1.5)  upperbound <- 1677 + (totalminsusage\_IQR \* 1.5)  nrow(phonecsv[phonecsv$TOT\_MINUTES\_USAGE < lowerbound |  phonecsv$TOT\_MINUTES\_USAGE > upperbound,])  **ZSCORE**  z\_data <- (phonecsv$TOT\_MINUTES\_USAGE - mean(phonecsv$TOT\_MINUTES\_USAGE)) / sd(phonecsv$TOT\_MINUTES\_USAGE) |
|  |
| |  | | --- | |  | |

### Q5 - Skewness

totminutesusage\_skew <-(3\*(mean(phonecsv$TOT\_MINUTES\_USAGE)-median(phonecsv$TOT\_MINUTES\_USAGE)))/sd(phonecsv$TOT\_MINUTES\_USAGE)

#### Z-Score

zscore\_totminsusage\_skew <-(3\*(mean(z\_data)-median(z\_data)))/sd(z\_data)

#### Natural Log

natural\_log <- log(phonecsv$TOT\_MINUTES\_USAGE[phonecsv$TOT\_MINUTES\_USAGE != 0])

natural\_log\_skew <-(3\*(mean(natural\_log)-median(natural\_log)))/sd(natural\_log)

#### Square Root

sqaure\_root <- sqrt(phonecsv$TOT\_MINUTES\_USAGE)

square\_root\_skew <- (3\*(mean(sqaure\_root) -median(sqaure\_root)))/sd(sqaure\_root)

summary(square\_root\_skew)

square\_root\_skew

### Q7 – Correlated Variables

a)

ggplot(phonecsv ,aes(x=phonecsv$TOT\_MINUTES\_USAGE, y=phonecsv$MINUTES\_CURR\_MONTH)) + labs(x="Total Minutes Usage", y="Minutes Current Month" ,title="Total Minutes Vs Minutes Current Month") + geom\_point()

ggplot(phonecsv ,aes(x=phonecsv$NUM\_LINES, y=phonecsv$TOT\_MINUTES\_USAGE)) + labs(x="Number Of Lines", y="Total Minutes Usage" ,title="Num Lines Vs Total Minutes") + geom\_point()

ggplot(phonecsv ,aes(x=phonecsv$TOT\_MINUTES\_USAGE, y=phonecsv$INCOME)) + labs(x="Total Minutes Usage", y="Income" ,title="Total Minutes Vs Income") + geom\_point()

ggplot(phonecsv ,aes(x=phonecsv$TOT\_MINUTES\_USAGE, y=phonecsv$CUST\_MOS)) + labs(x="Total Minutes Usage", y="Months" ,title="Total Minutes Vs Customer Months") + geom\_point()

b)

cor(phonecsv$TOT\_MINUTES\_USAGE, phonecsv$MINUTES\_CURR\_MONTH)

cov(phonecsv$TOT\_MINUTES\_USAGE, phonecsv$MINUTES\_CURR\_MONTH)

cor(phonecsv$NUM\_LINES, phonecsv$TOT\_MINUTES\_USAGE)

cov(phonecsv$NUM\_LINES, phonecsv$TOT\_MINUTES\_USAGE)

cor(phonecsv$TOT\_MINUTES\_USAGE, phonecsv$INCOME)

cov(phonecsv$TOT\_MINUTES\_USAGE, phonecsv$INCOME)

cor(phonecsv$TOT\_MINUTES\_USAGE, phonecsv$CUST\_MOS)

cov(phonecsv$TOT\_MINUTES\_USAGE, phonecsv$CUST\_MOS)

## Section 2

phonecsv <- phonecsv[-(1)]

phonecsv <- phonecsv[-(2)]

phonecsv <- phonecsv[-(2)]

phonecsv <- phonecsv[-(2)]

|  |
| --- |
| write.csv(phonecsv, file = "c:/EDT CA 1/weka.csv") |
|  |
| |  | | --- | |  | |

weka.classifiers.rules.PART = Builds a partial C4.5 decision tree in each iteration and makes the "best" leaf into a rule.