Enterprise Database Technology CA1

SHAJUN DOMINIC
X00109340

Table of Contents

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.

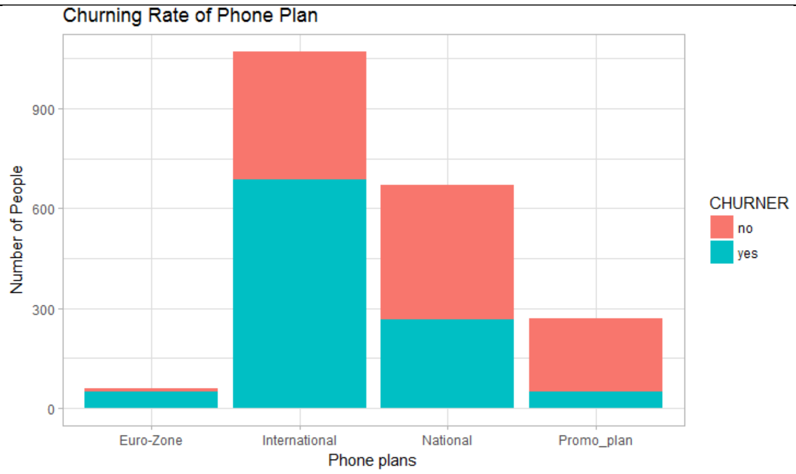
c) A mode function was created in order to get mode for some predictor variable and summary function was also used to get details for each predictor variable.

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 us	No useful information

								e less than 30 00 minutes	
CUST_ID	Ordinal	2070	1	1035.1	246	1035	597.8092		

	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		



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.

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

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

CUST_MOS <= 44 AND
AREA_CODE <= 45987 AND
CONVERGENT_BILLING = Yes AND
AREA_CODE > 15563: yes (410.0)

CUST_MOS <= 44 AND
AREA_CODE > 45987: no (270.0/48.0)

CUST_MOS <= 44 AND
IncomeRange = High Income AND
TOT_MINUTES_USAGE <= 3344 AND
EDUCATION = PhD AND
TOT_MINUTES_USAGE <= 1792 AND
MOBILE_PLAN <= 0: no (160.0/44.0)

CUST_MOS <= 44 AND
CUST_MOS <= 3 AND
AREA_CODE <= 21750 AND
LONGDIST_FLAG <= 0: no (181.0/54.0)

CUST_MOS > 44: yes (90.0)

CUST_MOS <= 3 AND
AREA_CODE > 21750: no (89.0)

AREA_CODE > 36785 AND
EDUCATION = Bachelors: yes (90.0)
```

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

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

(INCOME >= 75000) and (INCOME <= 75000) => CHURNER=no (180.0/0.0)
(INCOME >= 91000) and (TOT_MINUTES_USAGE <= 2748) => CHURNER=no (560.0/150.0)
(INCOME <= 39000) and (CUST_MOS <= 6) => CHURNER=no (271.0/82.0)
(AREA_CODE <= 15563) and (CUST_MOS <= 11) => CHURNER=no (270.0/115.0)
(AREA_CODE <= 10040) => CHURNER=no (90.0/26.0)
(INCOME >= 91000) and (CONVERGENT_BILLING = No) => CHURNER=no (20.0/7.0)
 => CHURNER=yes (680.0/10.0)

Number of Rules : 7
```

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

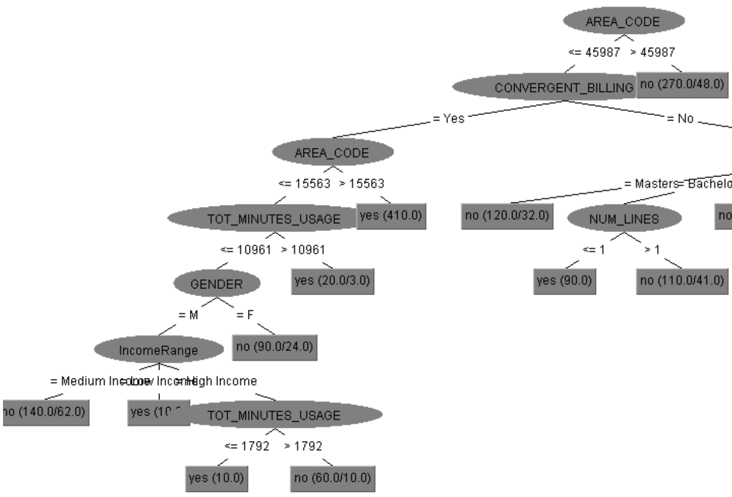


Figure 3 J48 Tree

```
J48 pruned tree
------------------

AREA_CODE <= 45987
|   CONVERGENT_BILLING = Yes
|   |   AREA_CODE <= 15563
|   |   |   TOT_MINUTES_USAGE <= 10961
|   |   |   |   GENDER = M
|   |   |   |   |   IncomeRange = Medium Income: no (140.0/62.0)
|   |   |   |   |   IncomeRange = Low Income: yes (10.0)
|   |   |   |   |   IncomeRange = High Income
|   |   |   |   |   |   TOT_MINUTES_USAGE <= 1792: yes (10.0)
|   |   |   |   |   |   TOT_MINUTES_USAGE > 1792: no (60.0/10.0)
|   |   |   |   GENDER = F: no (90.0/24.0)
|   |   |   TOT_MINUTES_USAGE > 10961: yes (20.0/3.0)
|   |   AREA_CODE > 15563: yes (410.0)
|   CONVERGENT_BILLING = No
|   |   EDUCATION = Masters: no (120.0/32.0)
|   |   EDUCATION = Bachelors
|   |   |   NUM_LINES <= 1: yes (90.0)
|   |   |   NUM_LINES > 1: no (110.0/41.0)
|   |   EDUCATION = High School: no (1.0)
|   |   EDUCATION = Post Primary
|   |   |   CUST_MOS <= 15: no (260.0/81.0)
|   |   |   CUST_MOS > 15
|   |   |   |   AREA_CODE <= 21750: yes (159.0/10.0)
|   |   |   |   AREA_CODE > 21750: no (50.0/13.0)
|   |   EDUCATION = PhD: no (270.0/52.0)
|   |   EDUCATION = Primary: yes (1.0)
AREA_CODE > 45987: no (270.0/48.0)

Number of Leaves  :     17

Size of the tree :      28
```

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 = TR
UE)
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)] = medi
anMinutesUsage

//MODE - PHONE_PLAN
modePhone <- (table(phonecsv$PHONE_PLAN)) == max(table(phonecsv$PHON
E_PLAN))
modePhone <- names(table(phonecsv$PHONE_PLAN))[modePhone]
phonecsv$PHONE_PLAN[phonecsv$PHONE_PLAN == ""] <- modePhone

//MODE - EDUCATION
modeEducation <- (table(phonecsv$EDUCATION)) == max(table(phonecsv$E
DUCATION))
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 Inco
me", "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(st
at="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_contin
uous(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="cou
nt") + 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 Loy
alty (Months)" ,y="Number of persons") + labs(title="Histogram For C
ustomer 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="CHU
RNER")+ theme_light()
```

GENDER

```
ggplot(data=phonecsv ,aes(x=phonecsv$GENDER, group=phonecsv$CHURNER,
fill=phonecsv$CHURNER)) + geom_histogram(stat="count")  + labs(x="Ge
nder" ,y="Number of People", title="Churning Rates", fill="CHURNER")
+ theme_light()
```

G)

```
ggplot(data=phonecsv ,aes(phonecsv$CHURNER, phonecsv$CUST_MOS)) + ge
om_boxplot() + labs(x="Churner" ,y="Customer Duration (Months)")

ggplot(data=phonecsv ,aes(phonecsv$CHURNER, phonecsv$TOT_MINUTES_USA
GE)) + 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_MIN
UTES_USAGE)) / sd(phonecsv$TOT_MINUTES_USAGE)
```

Q5 - Skewness

```
totminutesusage_skew <-(3*(mean(phonecsv$TOT_MINUTES_USAGE)-median(p
honecsv$TOT_MINUTES_USAGE)))/sd(phonecsv$TOT_MINUTES_USAGE)
```

Z-Score

```
zscore_totminsusage_skew <-(3*(mean(z_data)-median(z_data)))/sd(z_da
ta)
```

Natural Log

```
natural_log <- log(phonecsv$TOT_MINUTES_USAGE[phonecsv$TOT_MINUTES_U
SAGE != 0])
natural_log_skew <-(3*(mean(natural_log)-median(natural_log)))/sd(na
tural_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$TOT_MINUTES_USAGE, y=phonecsv$MINUTE
S_CURR_MONTH)) + labs(x="Total Minutes Usage", y="Minutes Current Mo
nth" ,title="Total Minutes Vs Minutes Current Month")  + geom_point(
)

ggplot(phonecsv ,aes(x=phonecsv$NUM_LINES, y=phonecsv$TOT_MINUTES_US
AGE)) + labs(x="Number Of Lines", y="Total Minutes Usage" ,title="Nu
m 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_M
OS)) + labs(x="Total Minutes Usage", y="Months" ,title="Total Minute
s 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.