CAPSTONE PROJECT

FINAL REPORT

June 2021

TELECOM CHURN ANALYSIS

Submitted By,

Ajuna P John

ACKNOWLEDGEMENT

I would like to express my sincere gratitude to my mentor Ms. Keya Choudhury Ganguli for her valuable guidance and creative suggestions offered during this project.

I would also like to sincerely thank Mr Faisal for his constant encouragement and the Great Lakes management for giving an opportunity to work on a relevant problem in Telecom Domain.

I would like to extend my gratitude to my family who has provided immense help and support during various phases of the project.

Contents

1. Introduction

1.1 Defining problem statement

The senior management in a telecom provider organization is worried about the rising customer attrition levels. Additionally, a recent independent survey has suggested that the industry will face increasing churn rates and decreasing ARPU (average revenue per unit). The effort to retain customers so far has been very reactive. Only when the customer calls to close their account is when the company acts. That has not proved to be a great strategy so far. The management team is keen to take more proactive measures on this front. You as a data scientist are tasked to derive insights, predict the potential behaviour of customers, and then recommend steps to reduce churn.

1.2 Need of the study/project

Customer churn is a major problem and one of the most important concerns for large companies with highly competitive services. Due to the direct effect on the revenues of the companies, especially in the telecom field, companies are seeking to develop means to predict potential customer to churn. Therefore, finding factors that increase customer churn is important to take necessary actions to reduce this churn. On the other hand, predicting the customers who are likely to leave the company will represent potentially large additional revenue source if it is done in the early phase.

1.3 Understanding business/social opportunity

The telecommunications sector has become one of the main industries in developed countries. The technical progress and the increasing number of operators raised the level of competition. Companies are working hard to survive in this competitive market depending on multiple strategies. To apply the strategy, which is most profitable, less costly, and much easier, companies must decrease the potential of customer's churn, known as "the customer movement from one provider to another".

When it comes to the telecommunications segment, there is great room for opportunities. The wealth and the amount of customer data that carriers collect can contribute a lot to shift from a reactive to a proactive position. The emergence of sophisticated artificial intelligence and data analytics techniques further help leverage this rich data to address churn in a much more effective manner.

2. Data Report

2.1 Understanding how data was collected in terms of time, frequency, and methodology

Dataset "Telecom_Sampled.csv" that contains 26518 observations, and 81 variables was provided for study. Generating hypothesis is key to unlock any analytics project. We first list down our understanding to derive insights through various approach and then proceed from there.

2.2 Visual inspection of data (rows, columns, descriptive details)

- There are 26518 observations and 81 variables in the dataset
- Number of rows in the dataset: 26518
- Number of Columns in the dataset: 81
- Target variable is "churn" with values "0" and "1" as "Non churn" and "churn" respectively. All the other variables are independent variables.

The snapshot of the data is provided below:

	mou_Mean	totmrc_Mean	rev_Range	mou_Range	change_mou	drop_blk_Mean	drop_vce_Range	owylis_vce_Range
1	190.25	63.9400	26.00	43	-11.25	4.6666667	8	
2	443.00	39.9900	5.10	199	-78.00	4.3333333	5	
3	400.50	44.9900	13.88	172	-67.50	2.0000000	1	
4	53.50	34.6675	18.56	78	12.50	1.3333333	2	
5	37.00	21.0425	35.79	74	-33.00	5.6666667	1	
6	1181.50	44.9900	62.30	381	-67.50	5.6666667	2	
7	543.50	69.9900	16.94	462	326.50	13.0000000	3	
8	1463.25	84.9900	97.03	1280	-824.25	28.3333333	19	
9	93.50	65.4750	15.96	23	9.50	1.0000000	1	
10	1225.50	35.9900	4.20	107	10.50	10.0000000	5	
11	1196.50	39.9900	63.65	82	57.50	10.3333333	2	
12	610.00	49.9900	131.46	694	-315.00	13.6666667	11	
13	1354.00	84.9900	37.51	506	-173.00	12.0000000	2	
14	787.50	44.6150	303.01	975	107.50	12.0000000	6	
15	233.00	39.9900	35.71	309	-130.00	11.3333333	13	

2.3 Understanding of attributes (variable info, renaming if required)

Class of the dataset is "data.frame". There are 47 continuous and 34 categorical variables.

Continuous Variable	Description
mou_Mean	Mean number of monthly minutes of use
totmrc_Mean	Mean total monthly recurring charge
rev_Range	Range of revenue (charge amount)
mou_Range	Range of number of minutes of use
change_mou	Percentage change in monthly minutes of use vs previous three month average
drop_blk_Mean	Mean number of dropped or blocked calls
drop_vce_Range	Range of number of dropped (failed) voice calls
owylis_vce_Range	Range of number of outbound wireless to wireless voice calls
mou_opkv_Range	Range of unrounded minutes of use of off-peak voice calls
months	Total number of months in service
totcalls	Total number of calls over the life of the customer
eqpdays	Number of days (age) of current equipment
custcare_Mean	Mean number of customer care calls
callwait_Mean	Mean number of call waiting calls
iwylis_vce_Mean	Mean number of inbound wireless to wireless voice calls
callwait_Range	Range of number of call waiting calls
ccrndmou_Range	Range of rounded minutes of use of customer care calls
adjqty	Billing adjusted total number of calls over the life of the customer
ovrrev_Mean	Mean overage revenue
rev_Mean	Mean monthly revenue (charge amount)
ovrmou_Mean	Mean overage minutes of use
comp_vce_Mean	Mean number of completed voice calls
plcd_vce_Mean	Mean number of attempted voice calls placed
avg3mou	Average monthly minutes of use over the previous three months
avgmou	Average monthly minutes of use over the life of the customer
avg3qty	Average monthly number of calls over the previous three months
avgqty	Average monthly number of calls over the life of the customer
avg6mou	Average monthly minutes of use over the previous six months
avg6qty	Average monthly number of calls over the previous six months
opk_dat_Mean	Mean number of off-peak data calls
retdays	Number of days since last retention call
roam_Mean	Mean number of roaming calls
recv_sms_Mean	Mean number of received SMS calls

blck_dat_Mean	Mean number of blocked (failed) data calls
mou_pead_Mean	Mean unrounded minutes of use of peak data calls
da_Mean	Mean number of directory assisted calls
da_Range	Range of number of directory assisted calls
datovr_Mean	Mean revenue of data overage
datovr_Range	Range of revenue of data overage
drop_dat_Mean	Mean number of dropped (failed) data calls
drop_vce_Mean	Mean number of dropped (failed) voice calls
adjmou	Billing adjusted total minutes of use over the life of the customer
totrev	Total revenue
adjrev	Billing adjusted total revenue over the life of the customer
avgrev	Average monthly revenue over the life of the customer
comp_dat_Mean	Mean number of completed data calls
plcd_dat_Mean	Mean number of attempted data calls placed

Categorical Variable	**Description**
income	Estimated income
crclscod	Credit class code
asl_flag	Account spending limit
prizm_social_one	Social group letter only
area	Geographic area
refurb_new	Handset: refurbished or new
hnd_webcap	Handset web capability
marital	Marital status
ethnic	Ethnicity roll-up code
age1	Age of first household member
age2	Age of second household member
models	Number of models issued
hnd_price	Current handset price
actvsubs	Number of active subscribers in household
uniqsubs	Number of unique subscribers in the household
forgntvl	Foreign travel dummy variable
dwlltype	Dwelling unit type
dwllsize	Dwelling size
mailordr	Mail order buyer
occu1	Occupation of first household member

mtrcycle	Motorcycle indicator
numbcars	Known number of vehicles
truck	Truck indicator
wrkwoman	Working woman in household
churn	Instance of churn between 31-60 days after observation date
solflag	Infobase no phone solicitation flag
proptype	Property type detail
mailresp	Mail responder
cartype	Dominant vehicle lifestyle
car_buy	New or used car buyer
children	Children present in household
csa	Communications local service area
div_type	Division type code
Customer_ID	Unique tournament specific customer ID for scoring purposes

3. *Exploratory data analysis*

3.1 Univariate analysis (distribution and spread for every continuous attribute, distribution of data in categories for categorical ones)

Summary of the dataset

The 6 numbers (minimum value, 1st quantile, Median, Mean, 3rd Quantile and Maximum value) that help to describe the centre, spread and shape of the data are determined. Visually checked and found that there are missing value and outliers in almost all the variables.

***Plots for various continuous variables provided in Annexure 1**

Boxplot of continuous variables are plotted (plots are given above) and found that outliers are present in most of the variables. Outliers in the variables are treated by flooring and capping the observations. After finding out the subset of outliers greater than 75% quantile and below 25% quantile, they are capped with maximum and minimum values.

Boxplot, histogram, and density diagram of the treated values are taken for analysis. Boxplot can also be used for finding correlation between numerical and categorical variables.

Mean number of monthly minutes of use (mou_Mean), Range of number of minutes of use (mou_Range), Mean number of dropped or blocked calls (drop_blk_Mean), Range of number of dropped (failed) voice calls (drop_vce_Range), Range of number of outbound wireless to wireless voice calls (owylis_vce_Range), Range of unrounded minutes of use of off-peak voice calls (mou_opkv_Range), Total number of calls over the life of the customer (totcalls), Billing adjusted total number of calls over the life of the customer (adjqty), Mean number of completed voice calls (comp_vce_Mean), Mean number of attempted voice calls placed (plcd_vce_Mean), Average monthly minutes of use over the previous three months (avg3mou), Average monthly minutes of use over the life of the customer (avgmou), Average monthly number of calls over the previous three months (avg3qty), Average monthly number of calls over the life of the customer (avgqty),Average monthly minutes of use over the previous six months (avg6mou), Average monthly number of calls over the previous six months (avg6qty), Billing adjusted total minutes of use over the life of the customer (adjmou),Total revenue (totrev), Mean number of dropped (failed) voice calls (drop_vce_Mean) values are slightly left skewed and hence median value is in lower side. Mean total monthly recurring charge (totmrc_Mean) data is also not uniform.

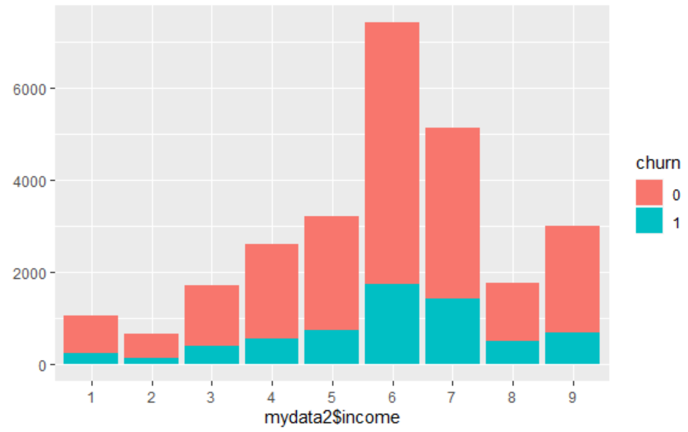
3.1 Bivariate analysis (relationship between different variables, correlations)

We have analysed the relationship between independent categorical variables with Target variable “churn”.

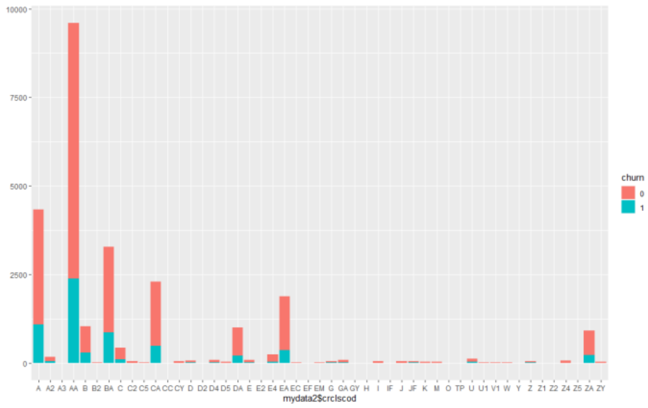
From Chi square, Fisher test and visual analysis, it is found that the following variables are insignificant

Foreign travel dummy variable (forgntvl), Motorcycle indicator (mtrcycle), Truck indicator (truck), Range of unrounded minutes of use of off-peak voice calls (mou_opkv_Range), Total number of calls over the life of the customer (totcalls), Billing adjusted total number of calls over the life of the customer (adjqty), Mean monthly revenue (charge amount) (rev_Mean), Mean overage minutes of use (ovrmou_Mean), Mean number of completed voice calls (comp_vce_Mean), Average monthly minutes of use over the life of the customer (avgmou), Average monthly number of calls over the previous six months (avg6qty), Mean number of off-peak data calls (opk_dat_Mean), Mean number of roaming calls (roam_Mean), Mean number of blocked (failed) data calls (blck_dat_Mean), Mean unrounded minutes of use of peak data calls (mou_pead_Mean), Mean revenue of data overage (datovr_Mean), Range of revenue of data overage (datovr_Range), Mean number of dropped (failed) data calls (drop_dat_Mean), Billing adjusted total minutes of use over the life of the customer (adjmou), Total revenue (totrev), Billing adjusted total revenue over the life of the customer (adjrev), Average monthly revenue over the life of the customer (avgrev), Mean number of completed data calls (comp_dat_Mean), Customer ID and Mean number of attempted data calls placed (plcd_dat_Mean).

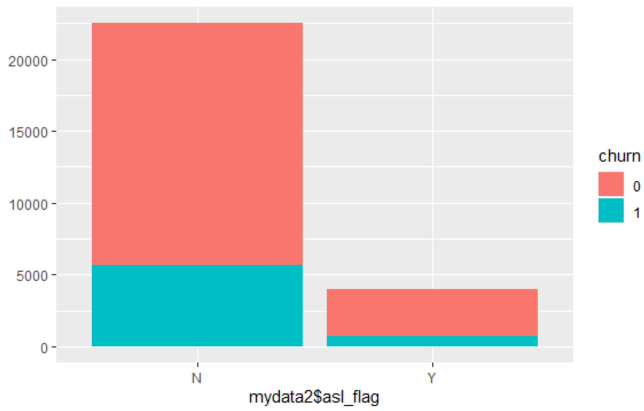
1. Income

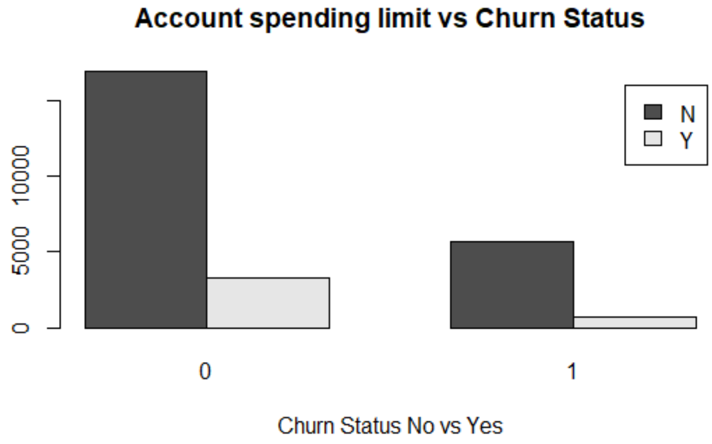


2. Credit class code



3. Account spending limit





Probability of variable proportions:

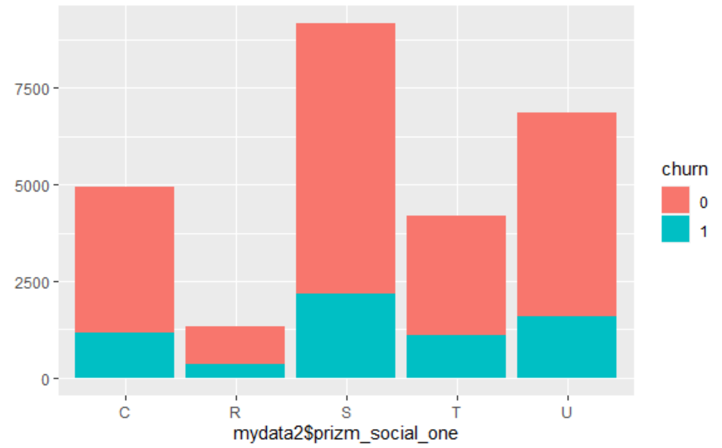
prop.table on as_flag and churn

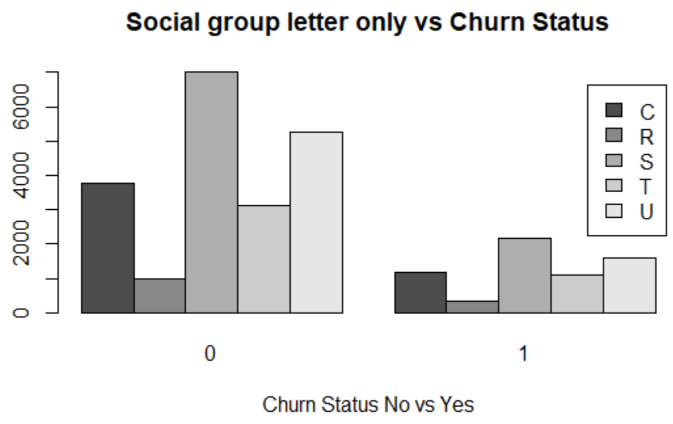
	0	1
N	74.97670	25.02330
Y	81.79764	18.20236

Table on as_flag and churn

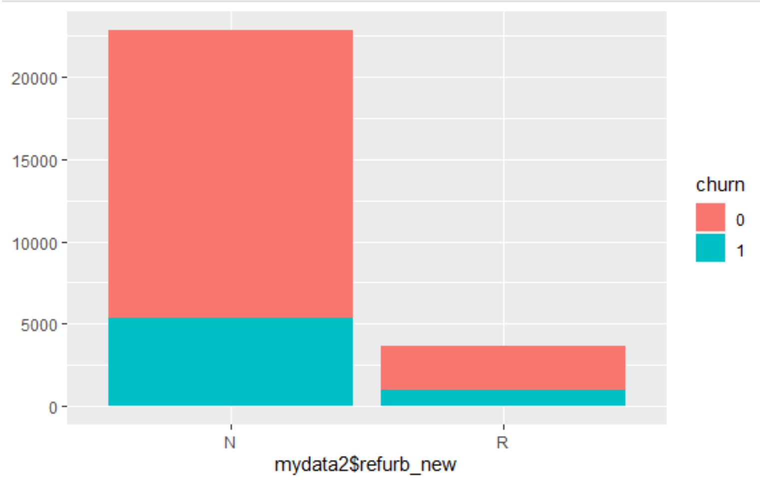
	N	Y
0	16896	3258
1	5639	725

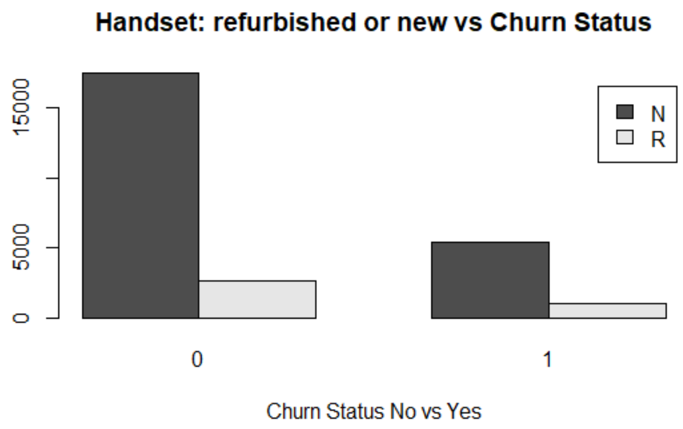
4. Social group letter





5. Handset: refurbished or new





Probability of variable proportions:

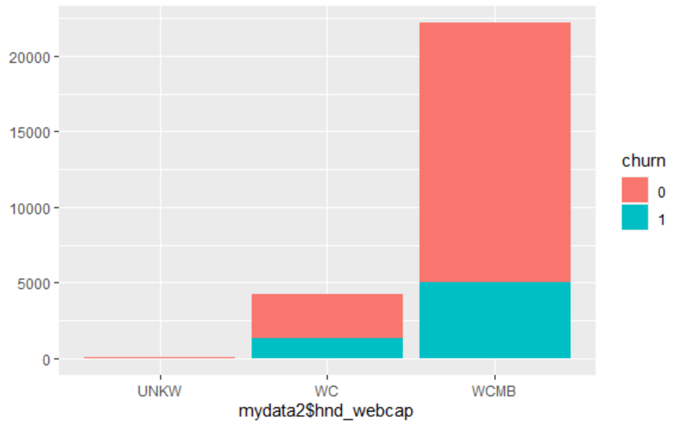
prop.table on refurb_new and churn

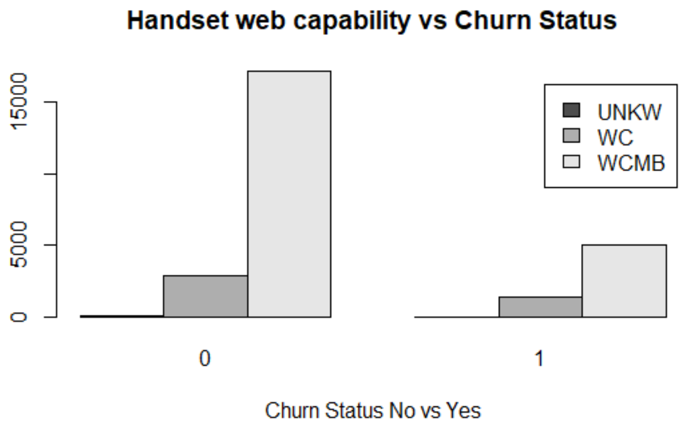
	0	1
N	76.52976	23.47024
R	72.69494	27.30506

Table on refurb_new and churn

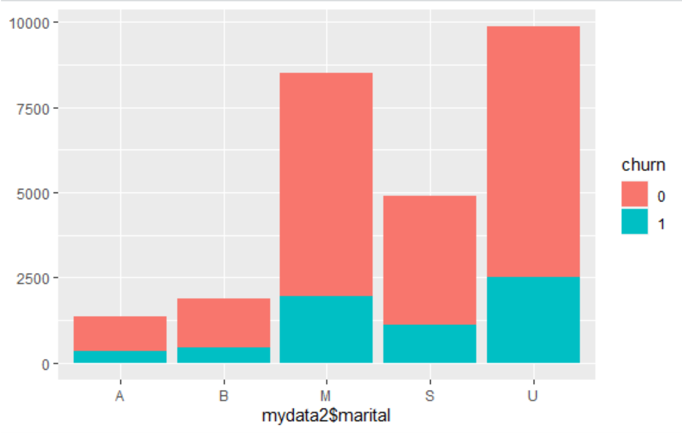
	N	R
0	17497	2657
1	5366	998

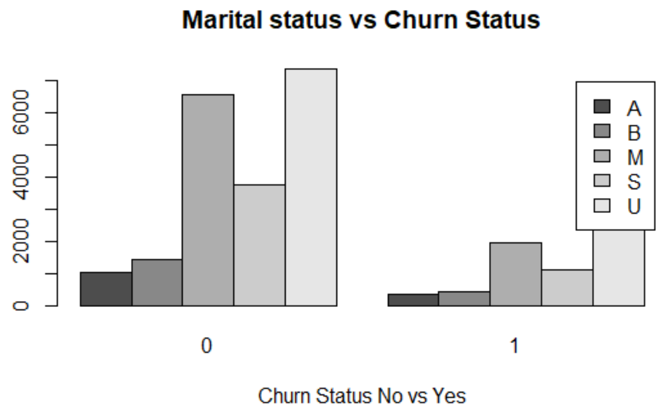
6. Handset web capability



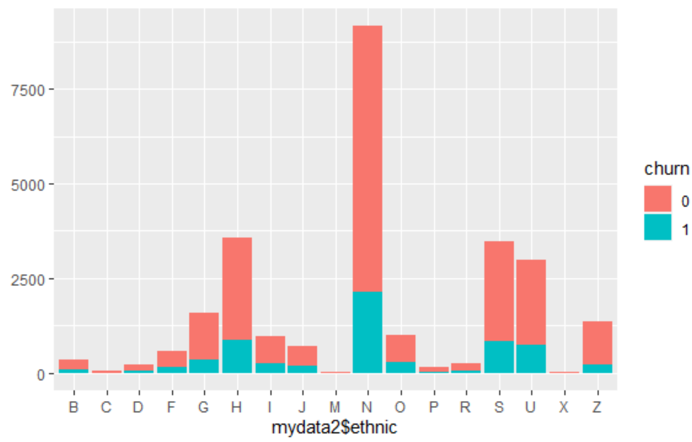


7. Marital status

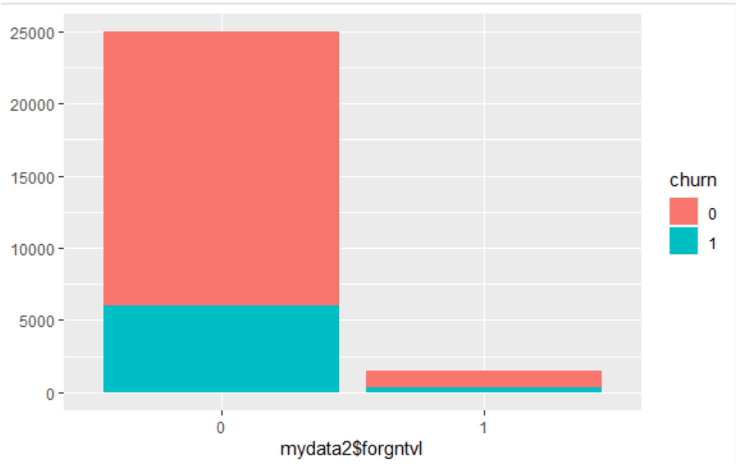


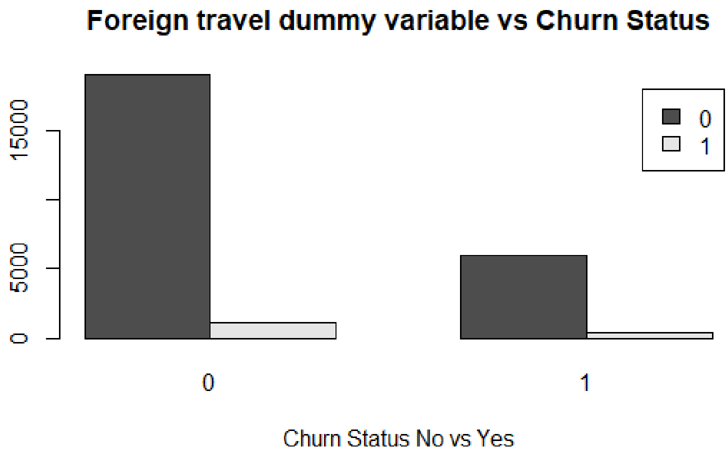


8. Ethnicity roll-up code



9. Foreign travel dummy variable





Probability of variable proportions:

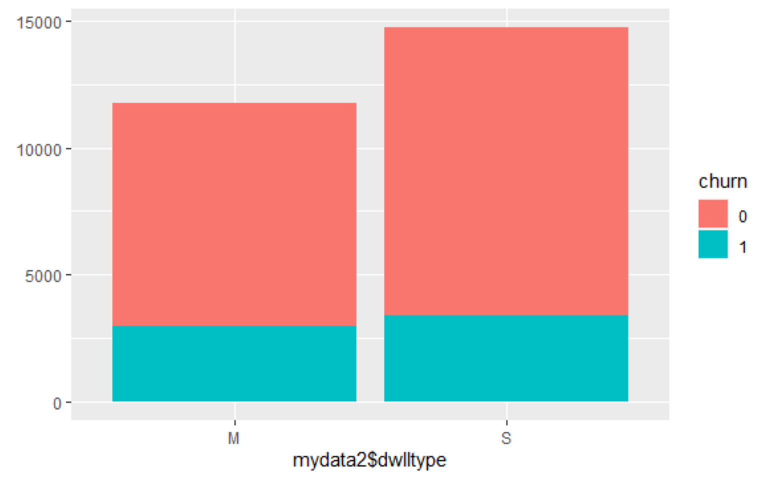
prop.table on forgntvl and churn

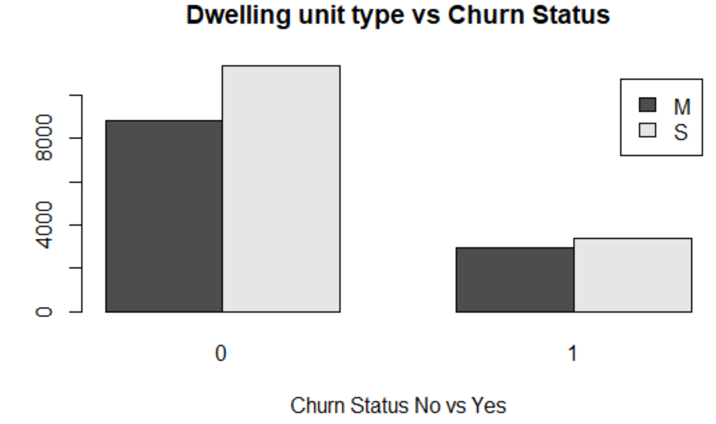
	0	1
0	75.97200	24.02800
1	76.48221	23.51779

Table on forgntvl and churn

	0	1
0	18993	1161
1	6007	357

10. Dwelling unit type





Probability of variable proportions:

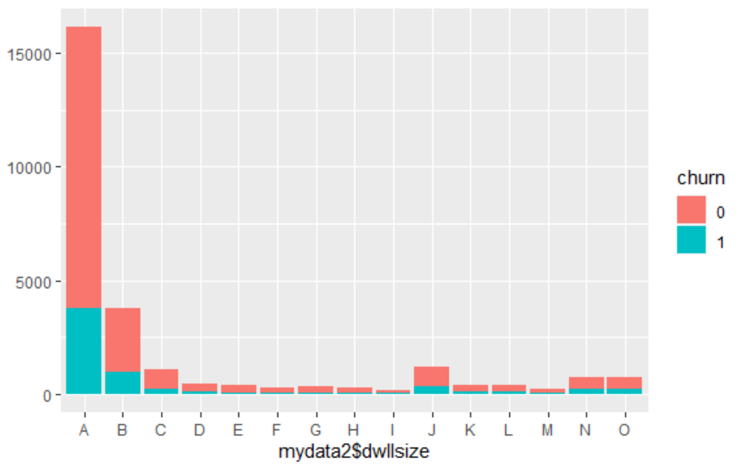
prop.table on dwlltype and churn

	0	1
M	74.79606	25.20394
S	76.96271	23.03729

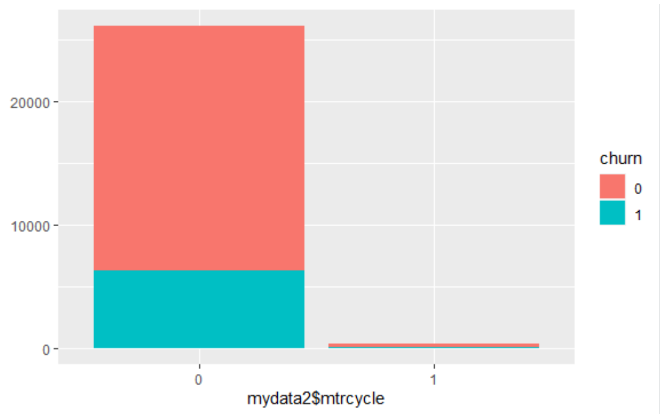
Table on dwlltype and churn

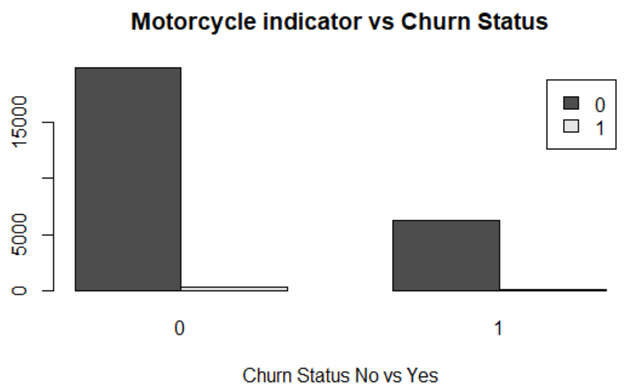
	M	S
0	8802	11352
1	2966	3398

11. Dwelling size



12. Motorcycle indicator





Probability of variable proportions:

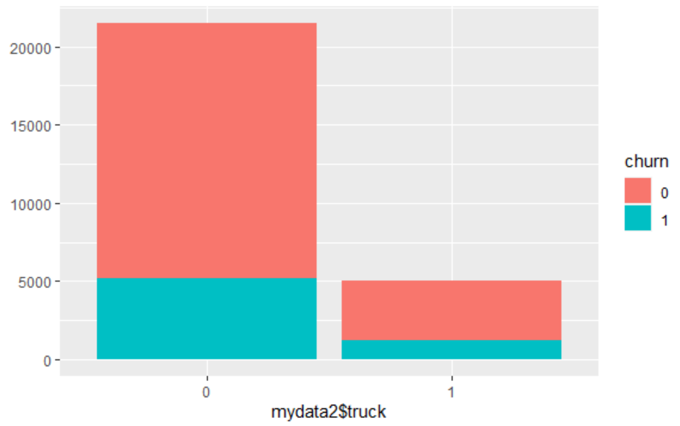
prop.table on mtrcycle and churn

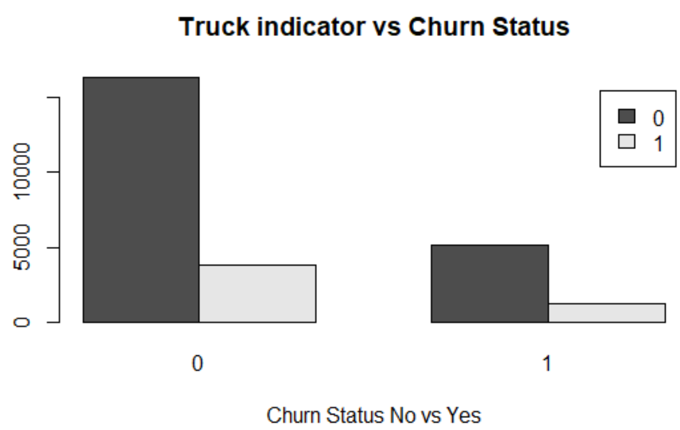
	0	1
0	76.0263	23.9737
1	74.1573	25.8427

Table on mtrcycle and churn

	0	1
0	19890	264
1	6272	92

13. Truck indicator





Probability of variable proportions:

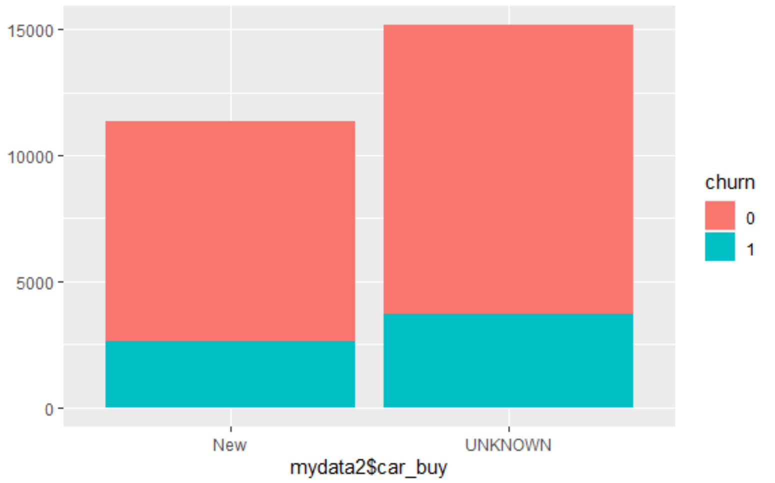
prop.table on truck and churn

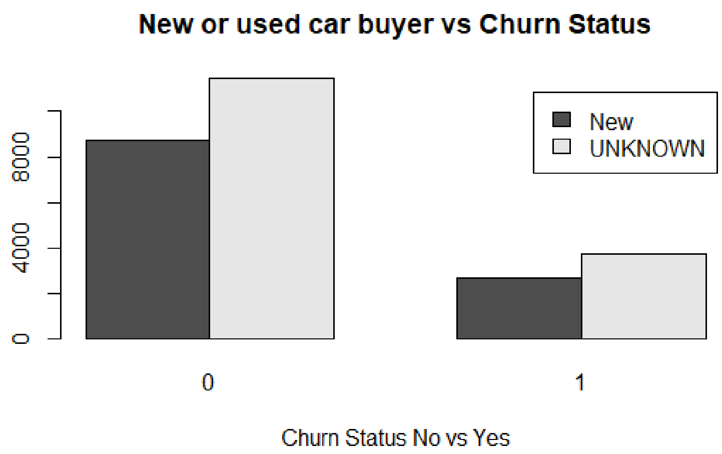
	0	1
0	75.92395	24.07605
1	76.33313	23.66687

Table on truck and churn

	0	1
0	16332	3822
1	5179	1185

14. New or used car buyer





Probability of variable proportions:

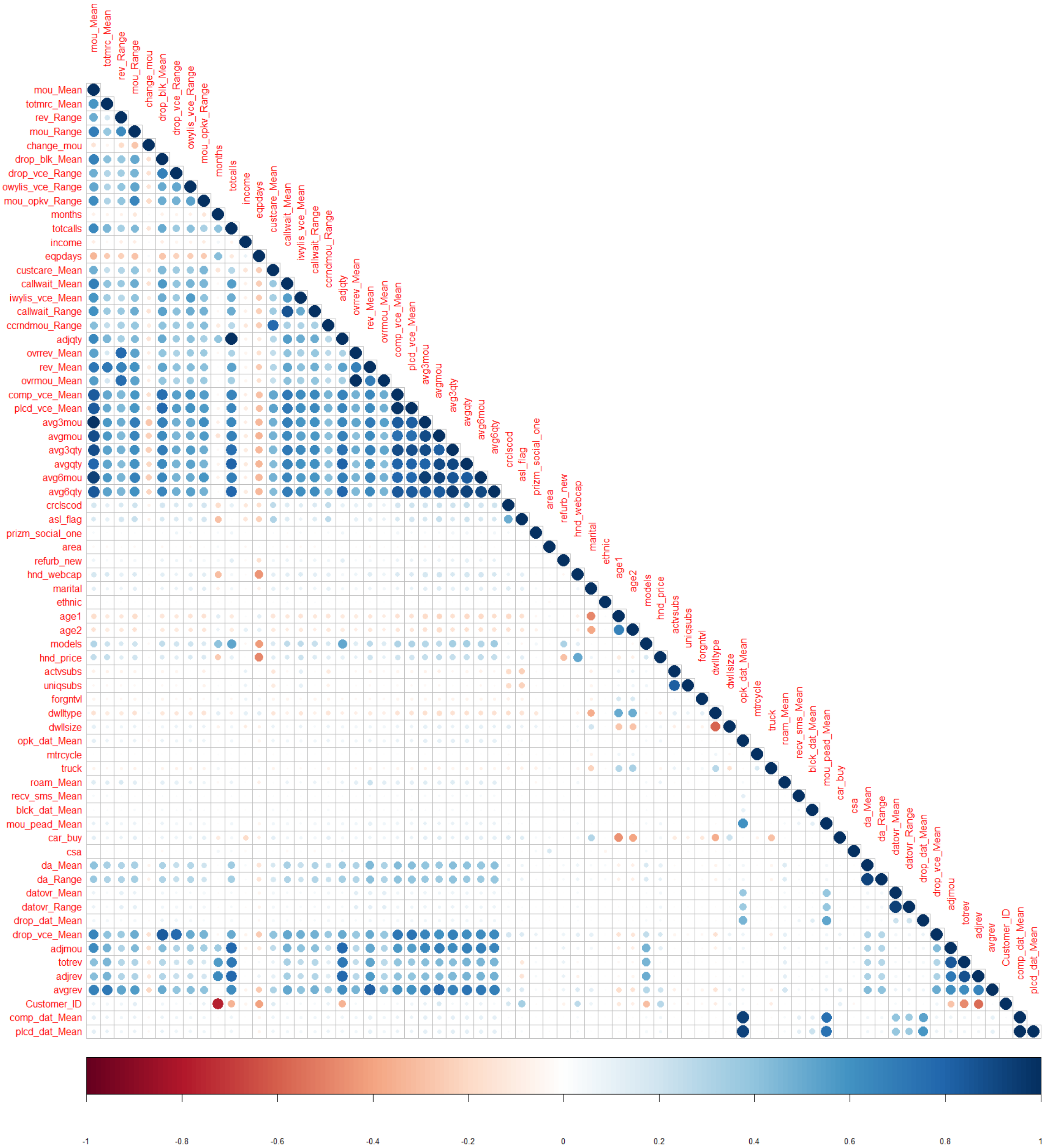
prop.table on car_buy and churn

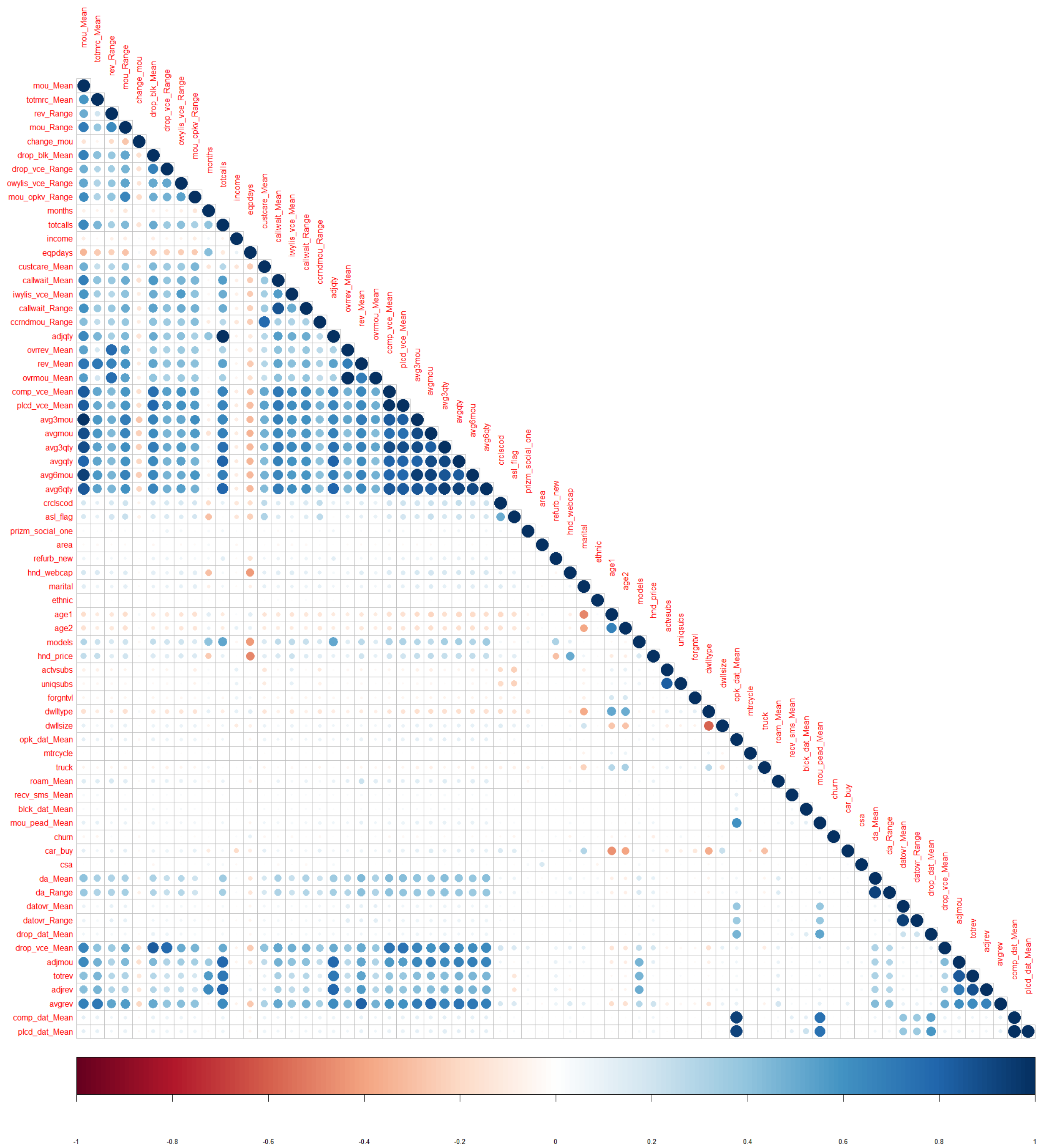
	0	1
New	76.66960	23.33040
UNKNOWN	75.50105	24.49895

Table on car_buy and churn

	New	UNKNOWN
0	8702	11452
1	2648	3716

1. Correlation between the variables





[The accurate picture can be obtained from the R Code]

- Billing adjusted total number of calls and Total number of calls over the life of the customer, Mean overage minutes of use and Mean overage revenue, Mean number of attempted voice calls placed and Mean number of completed voice calls are highly correlated
- Mean number of attempted data calls placed, Mean number of completed data calls and Mean number of off-peak data calls are highly correlated
- Mean number of monthly minutes of use is highly correlated to Mean number of completed voice calls, Mean number of attempted voice calls placed, Average monthly minutes of use over the previous three months, Average monthly minutes of use over the life of the customer, Average monthly number of calls over the previous three months, Average monthly minutes of use over the previous six months and Average monthly number of calls over the previous six months
- Average monthly number of calls over the previous three months is highly correlated to Mean number of monthly minutes of use, Mean number of completed voice calls, Mean number of attempted voice calls placed and Average monthly minutes of use over the previous three months
- Average monthly number of calls over the previous six months is highly correlated to Mean number of monthly minutes of use, Total number of calls over the life of the customer, Billing adjusted total number of calls over the life of the customer, Mean number of completed voice calls, Mean number of attempted voice calls placed, Average monthly minutes of use over the previous three months, Average monthly minutes of use over the life of the customer, Average monthly number of calls over the previous three months, Average monthly number of calls over the life of the customer and Average monthly minutes of use over the previous six months

3.2 Removal of unwanted variables

Found missing values in 42 variables. Variables with missing values more than 40% are discarded/ removed from analysis. 11 out of 81 variables are removed.

mailordr, occu1, numbcars, retdays, wrkwoman, solflag, proptype, mailresp, cartype, children and div_type are the variables that are discarded.

Customer ID is a unique code given to each customer and does not seem to be relevant for analysis, hence dropping it as well.

Now the working dataset has 26518 observations and 69 variables. 1 Dependant variable and 68 independent variables.

3.3 Missing Value treatment

The analysis obtained suggest that many variables still have missing values in them. Out of the 69 variables, 31 variables having missing values are also having vital other information, we may replace the missing values with “median value” or “KNN value” to factor them instead of discarding them

Missing values in the variables `avg6mou`, `avg6qty`, `mou_Mean`, `totmrc_Mean`, `rev_Range`, `mou_Range`, `change_mou`, `ovrrev_Mean`, `rev_Mean`, `ovrmou_Mean` which are continuous with less than 10% of data values missing is replaced with the median value of the column values.

Missing values in the variables "prizm_social_one", "income", "dwlltype", "dwllsize", "hnd_webcap", "marital", "ethnic", "age1", "age2", "hnd_price", "forgntvl", "mtrcycle", "truck", "roam_Mean", "car_buy", "csa", "da_Mean", "da_Range", ”datovr_Mean", "datovr_Range", "area" which are categorical with less than 40% of data values missing are replaced with KNN of the column values.

3.4 Outlier treatment

Visually checked for outliers and found that there are outliers in almost all the variables. Boxplot of continuous variables are also plotted (plots are given above) and found that outliers are present in most of the variables. Outliers in the variables are treated by flooring and capping the observations. After finding out the subset of outliers greater than 75% quantile and below 25% quantile, they are capped with maximum and minimum values. There are some variables which has values equal to zero or close to zero and hence those variables are not transformed.

3.5 Variable transformation

Conversion to factors:

The variables (forgntv1, mtrcycle, truck, churn) are converted to factors

The income levels are converted into ordered factors

eqpdays (“Number of days (age) of current equipment”) has negative values. We have fixed them with corresponding absolute values

Factor Analysis

1. Bartlett test

   Ho: All dimensions are unrelated
   Ha: All dimensions are related

   Rejecting the null hypothesis (p value< .05) indicate that a factor analysis may be useful with the data.

Chisq	p.value	df
6024.339	3.700552e-321	2346

   If p-value is much lesser than 0.05, we reject null hypothesis, indicating that there may be statistically significant interrelationship between variables in our dataset and it is an ideal case for dimension reduction. Pvalue is the calculated probability of making a Type 1 error.

   Here Ho: the correlation matrix is same as the identity matrix

   Ha: Correlation matrix is different from identity matrix

2. **KMO test for checking sampling adequacy**

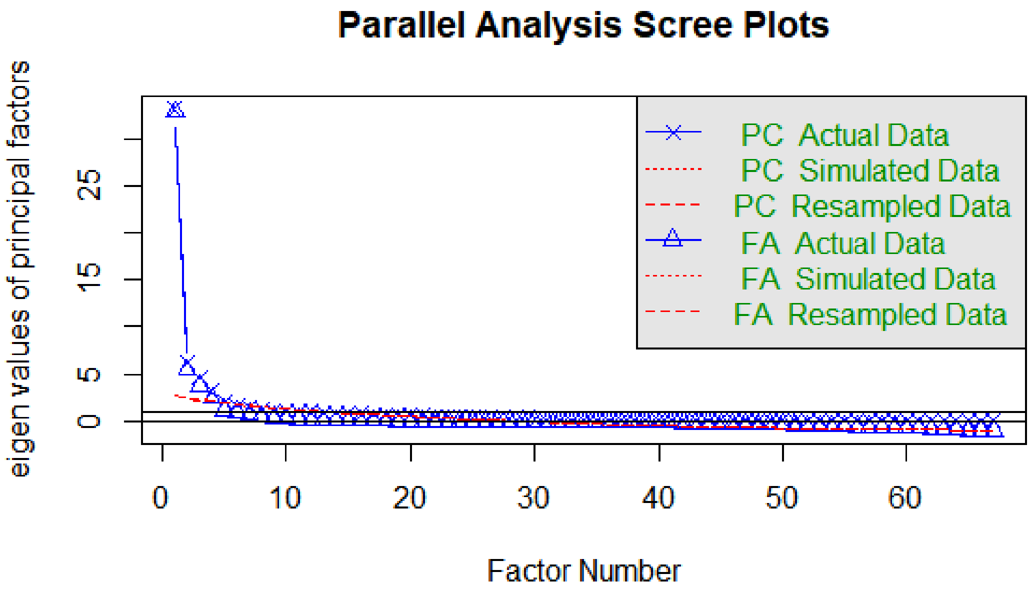
Kaiser-Meyer-Olkin factor adequacy

We have obtained MSA for each item and overall MSA. Here the output Overall MSA is 0.89 which indicate the sampling is adequate for factor analysis. Hence, we need not collect more samples.

MSA value for blck_dat_Mean and drop_dat_Mean is less than .5, hence would not be useful for factor analysis.

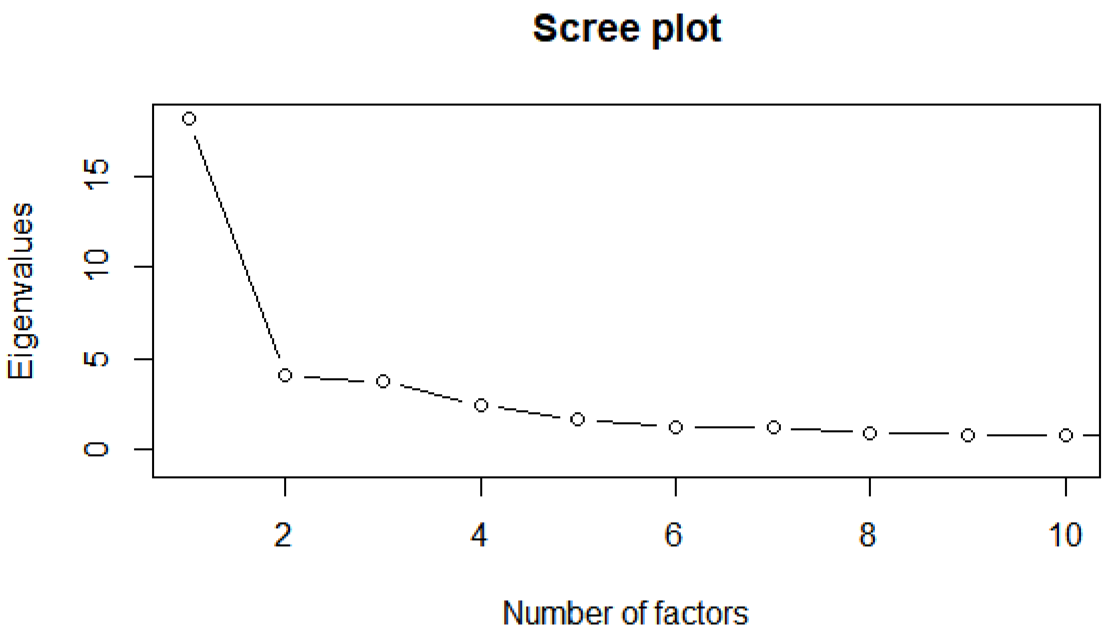
3. Parallel Analysis

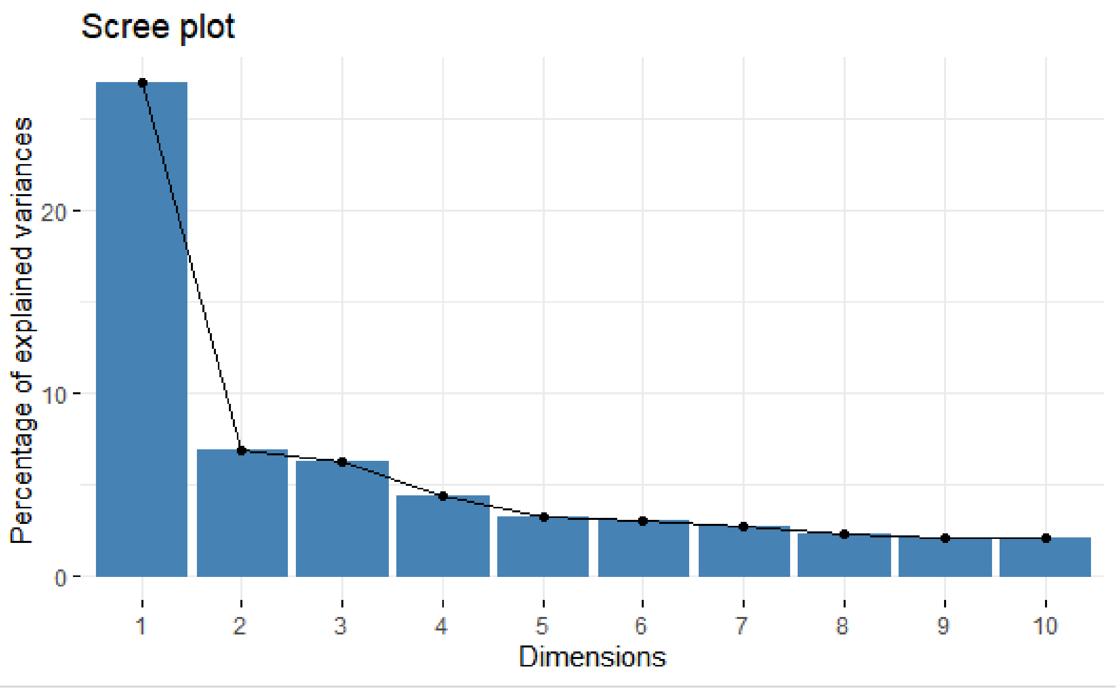
According to parallel analysis, the number of factors suggested is 4



4. Extract eigen values

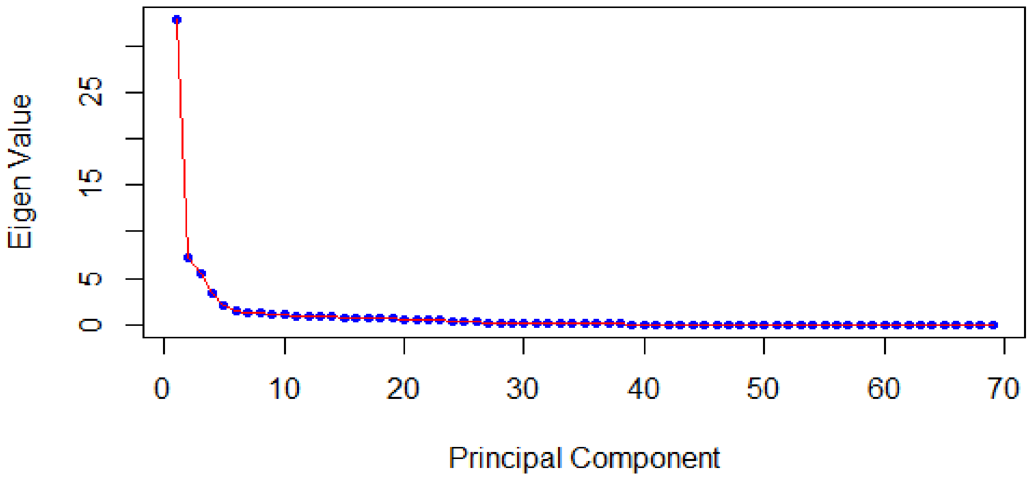
Eigen value is the basis for selecting number of factors. Eigen values for all variables are extracted.





Scree plot suggests that 3 is the desirable number of factors for analysis of the given dataset

Kaisers



We shall check what happens if we do factor analysis.

We received output that test of the hypothesis that 3 factors are sufficient. Proportion of variance explained by each factor (from Proportion Var 0.26 0.06 0.05) is:

Factor1: 25%

Factor2: 6%

Factor3: 5%

Variable “Models” is crossloading in factor 1 and 3.

There is illogical loading, hence we need to do rotation.

Varimax rotation is recommended as it is Orthogonal rotation. The loadings which are higher became more higher and which are lower became much lower.

Standardized loadings (pattern matrix) based upon correlation matrix are obtained. Test of the hypothesis that 3 components are sufficient.

The root mean square of the residuals (RMSR) is 0.06
with the empirical chi square 2123332 with prob < 0

Fit based upon off diagonal values = 0.95

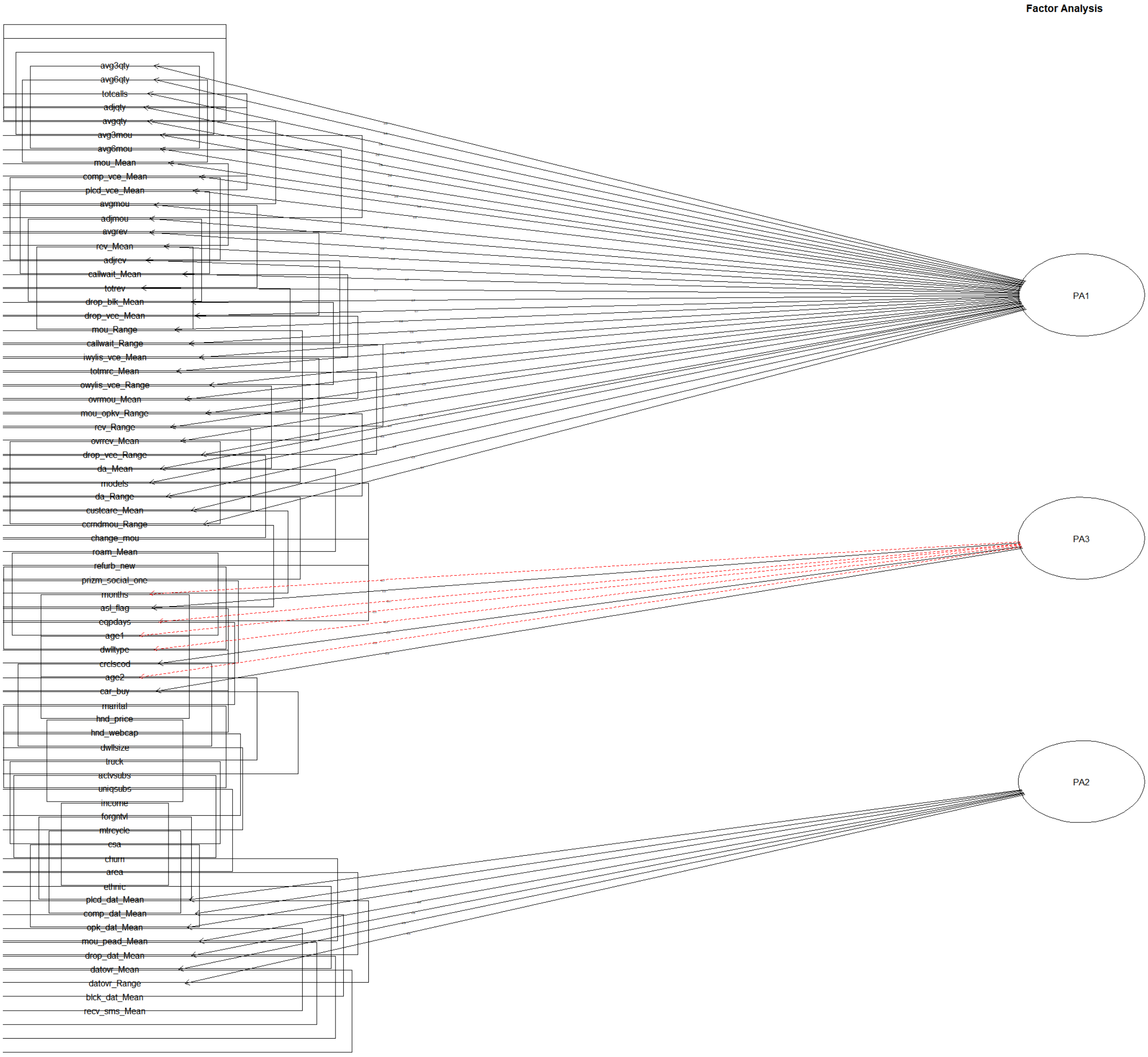
While checking cumulative variance, all the 3 factors contribute 38% of the explanation

Proportion of variance explained by each factor (from Proportion Var 0.26 0.06 0.05) is:

Factor1: 25%

Factor2: 6%

Factor3: 7%



[The accurate picture can be obtained from the R Code]

Proportion of variance explained by each factors (from Proportion Var: 0.25 0.07 0.06) is:

Factor1 explains 25% of variance and can be categorised as “Minutes of Usage and Revenue”:

Mean number of monthly minutes of use (mou_Mean), Mean total monthly recurring charge (totmrc_Mean), Range of revenue (charge amount) (rev_Range), Range of number of minutes of use (mou_Range), Mean number of dropped or blocked calls (drop_blk_Mean), Range of number of dropped (failed) voice calls (drop_vce_Range), Range of number of outbound wireless to wireless voice calls (owylis_vce_Range), Range of unrounded minutes of use of off-peak voice calls (mou_opkv_Range), Total number of calls over the life of the customer (totcalls), Mean number of call waiting calls (callwait_Mean), Mean number of inbound wireless to wireless voice calls (iwylis_vce_Mean), Range of number of call waiting calls (callwait_Range), Billing adjusted total number of calls over the life of the customer (adjqty), Mean overage revenue (ovrrev_Mean), Mean monthly revenue (charge amount) (rev_Mean), Mean overage minutes of use (ovrmou_Mean), Mean number of completed voice calls (comp_vce_Mean), Mean number of attempted voice calls placed (plcd_vce_Mean), Average monthly minutes of use over the previous three months (avg3mou), Average monthly minutes of use over the life of the customer (avgmou), Average monthly number of calls over the previous three months (avg3qty), Average monthly number of calls over the life of the customer (avgqty), Average monthly minutes of use over the previous six months (avg6mou), Average monthly number of calls over the previous six months (avg6qty), Mean number of dropped (failed) voice calls (drop_vce_Mean), Billing adjusted total minutes of use over the life of the customer (adjmou), Total revenue (totrev), Billing adjusted total revenue over the life of the customer (adjrev), Mean number of customer care calls (custcare_Mean), Range of rounded minutes of use of customer care calls (ccrndmou_Range), Number of models issued (models), Mean number of directory assisted calls (da_Mean), Range of number of directory assisted calls (da_Range) and Average monthly revenue over the life of the customer (avgrev)

Factor3 explains 6% of variance categorised as “Customer Characteristics”:

Total number of months in service (months), Number of days (age) of current equipment (eqpdays), Credit class code (crclscod), Account spending limit (asl_flag), Age of first household member (age1), Age of second household member (age2), New or used car buyer (car_buy) and Dwelling unit type (dwlltype)

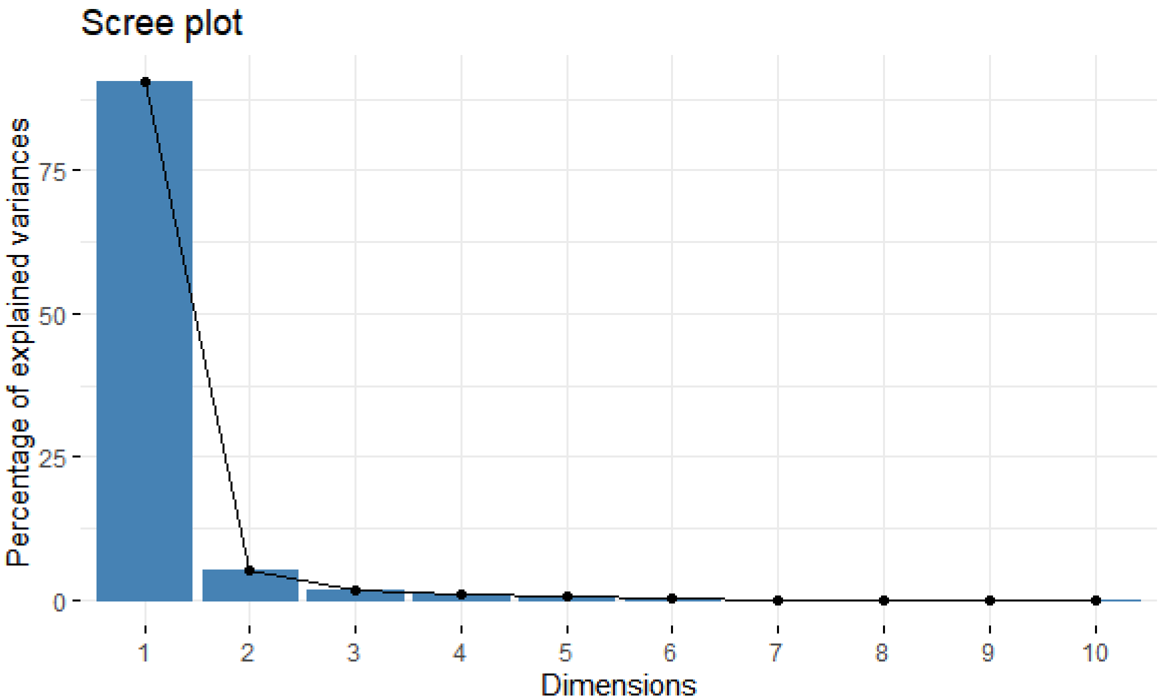
Factor2 explains 7% of variance categorised as “Customer Data usage”:

Mean number of off-peak data calls (opk_dat_Mean), Mean unrounded minutes of use of peak data calls (mou_pead_Mean), Mean revenue of data overage (datovr_Mean), Mean number of dropped (failed) data calls (drop_dat_Mean), Mean number of completed data calls (comp_dat_Mean), Mean number of attempted data calls placed (plcd_dat_Mean) and Range of revenue of data overage (datovr_Range)

PCA

Principal does Principal Components factor Analysis i.e principal component extraction with assumption that factors are correlated.

While checking cumulative variance, first 3 principal components explain 97% of the variance. PC1 contributes almost all the variation in the data.




```
> head(eig.mydata)
          eig   variance cumvariance
1 23994.6739 90.4844781    90.48448
2  1386.6528  5.2291001    95.71358
3   450.0719  1.6972317    97.41081
4   260.1777  0.9811360    98.39195
5   143.4531  0.5409651    98.93291
6   126.5329  0.4771585    99.41007
```

Standard deviation which calculates how much variation in the original data each principal component accounts for.

4. Business insights from EDA

4.1 Is the data unbalanced? If so, what can be done? Please explain in the context of the business

Balance of the target variable: 24% customers have churned and 76% has not churned.

4.2 Business insights

- Most of the variables (continuous/categorical) follow normal distribution (analysis from histogram and density plots).
- Churn does not seem to be highly correlated with any of the variables.
- Factors such as minutes of Usage, revenue, customer characteristics and data usage impact customer churn to a great extend
- Variables that are related to cost and billing, network service quality, and usage of services are highly significant and correlated with customer churn.

5. Model building and interpretation

5.1 Build various models

Linear model

Linear model helps to represent the dependant variable as a linear combination of independent variables. Simple linear regression works well when the dependant variable is normally distributed.

Summary of the model is derived. After checking on the pvalue of each variable, rev_Range, drop_vce_Range, owylis_vce_Range, totcalls, custcare_Mean, callwait_Mean, callwait_Range, ccrndmou_Range, adjqty, ovrrev_Mean, avg3mou, avg3qty, avgqty, avg6mou, avg6qty , crclscod, prizm_social_one, area, marital, age2, forgntvl, dwlltype, opk_dat_Mean, mtrcycle, truck, roam_Mean, recv_sms_Mean, blck_dat_Mean, mou_pead_Mean, car_buy, csa, da_Mean, da_Range, datovr_Mean, datovr_Range, drop_dat_Mean, adjmou, adjrev, avgrev, plcd_dat_Mean and comp_dat_Mean seems to be less significant according to the linear model.

In the model, Residual Standard error is .419 means if I forecast my churn, I would predict with an error of 42%.

Multiple R squared 0.03982 implies that in this model, only 3.98% of the variation in churn is explained by all other factors taken together. Adjusted R squared signifies that in real life situation, all the factors together explain only 3.74% of variation and not 3.99%. There is no significant improvement in performance of the model even after dropping few variables.

P value gives you the actual risk or level of significance by which null hypothesis is rejected. As p value here is < 2.2e-16 which is less than the significant level of .05%, we accept H1 – the mean of at least one group is different. We can conclude that there is significant difference between the groups.

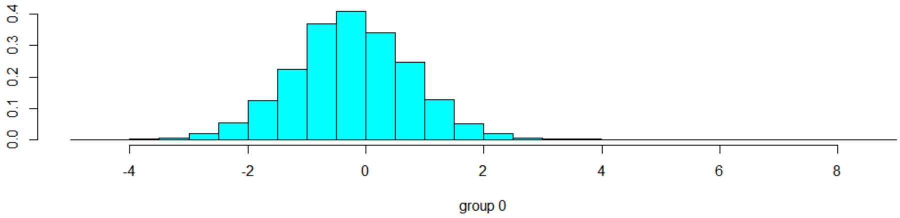
Splitting the dataset into train and test dataset

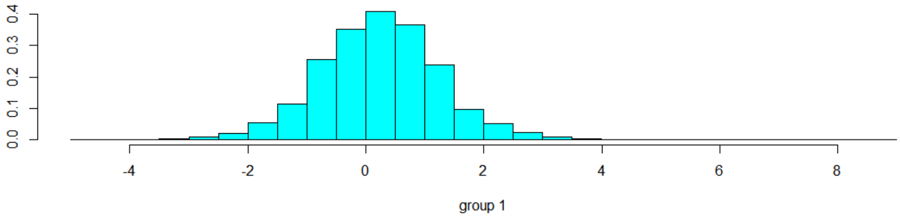
Before building the models, we split the given dataset into train and test datasets in 70:30 split.
Dimension of test data: 7955 observations and 70 variables
Dimension of train data: 18563 observations and 70 variables.
There is 24% default probability and 76% non-default in train and test dataset

Linear Discriminant Analysis (LDA)

Performed Linear Discriminant Analysis on the dataset

From prior probabilities obtained, There is 24% default probability and 76% non-default.





Confusion matrix

	Actual	
Predicted	0	1
0	5996	1875
1	50	34

The model could predict well in train and test dataset. The model gave 76.07% accuracy in train dataset and 75.8% accuracy in test dataset. Area under the curve: 0.6273

Factors that are related to customer characteristics such as Handset: refurbished or new (refurb_new), Number of unique subscribers in the household (uniqsubs), Motorcycle indicator (mtrcycle), Truck indicator (truck), Number of models issued (models), Estimated income (income), Mean number of call waiting calls (callwait_Mean) is found to be significant in influencing customer churn according to this model.

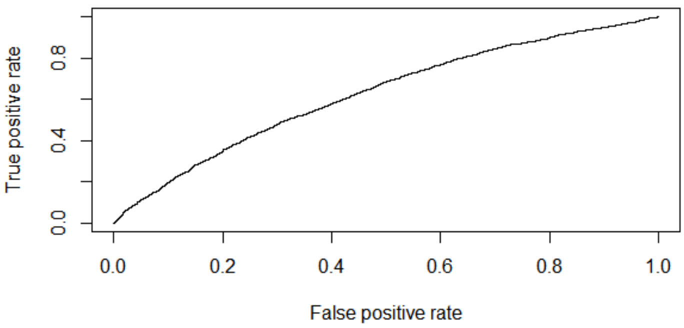
Logistic Regression model

After performing logistic regression analysis on the train dataset, we could find few variables as insignificant and hence dropped those variables and performed analysis on the dataset again.

The model predicted well on train and test dataset with accuracy of 75.98% in train dataset and 75.86% accuracy in test dataset.

Confusion Matrix:

	FALSE	TRUE
0	6011	35
1	1885	24



Area under the curve is 0.6254837. This means, If I build a model on my training dataset & then look at a new set of data, & pick from it random customers who cancelled and not cancelled the service, then 62.6% of the time, the churned customers will have higher predicted churn and the non-churn customers will have low predicted churn.

Factors categorized as “Customer Characteristics” and “Usage minutes” such as Number of days (age) of current equipment (eqpdays), Mean number of monthly minutes of use (mou_mean), Average monthly minutes of use over the life of the customer (avgmou), Account spending limit (asl_flag), Mean number of dropped or blocked calls (drop_blk_Mean), Age of first household member (age1), Number of unique subscribers in the household (uniqsubs), Range of number of minutes of use (mou_Range), Total number of months in service (months), Mean overage minutes of use (ovrmou_Mean) and Number of active subscribers in household (actvsubs) is found to be top 10 important variables according to logistic regression model.

Step AIC Regression method

This method provides an output with all the variables that are significant and gave the best model with AIC: 19794.

According to this method, it is found that generalized linear model performs best in predicting churn. Generalized linear model provides an accuracy of 59.73% on train data and 59.21% on test data.

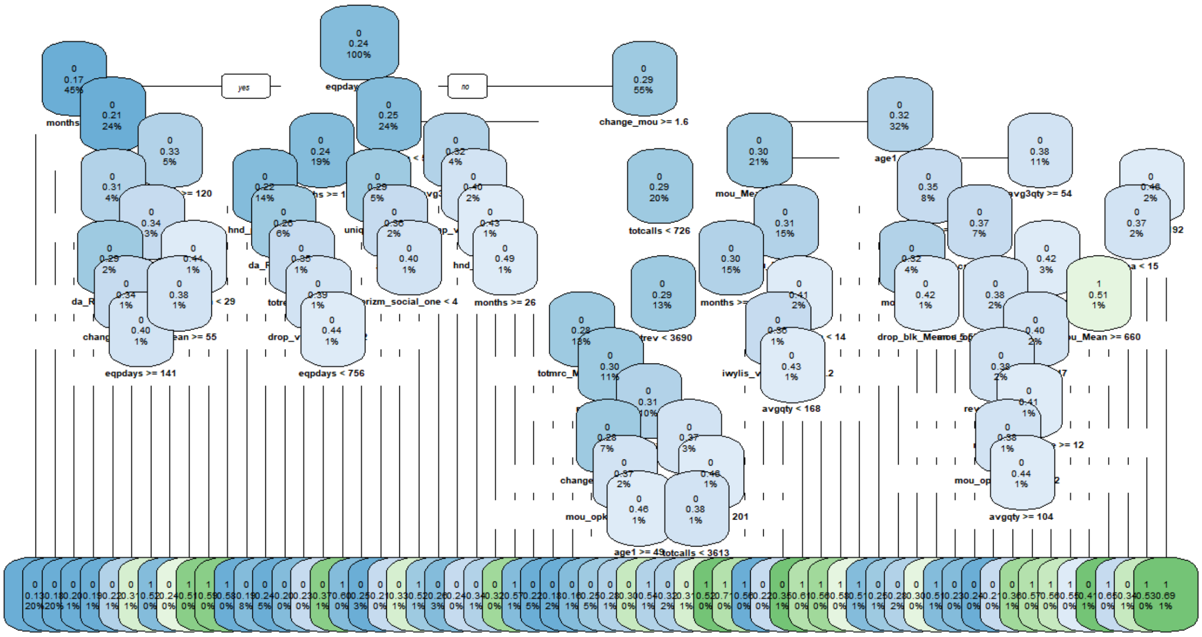
Confusion Matrix:

	pred 0	1
0	3570	2476
1	769	1140

From this method, factors categorized as “Customer Characteristics” and “Usage minutes” such as Mean number of monthly minutes of use (mou_mean), Mean number of dropped or blocked calls (drop_blk_Mean), Number of days (age) of current equipment (eqpdays), Average monthly minutes of use over the life of the customer (avgmou), Account spending limit (asl_flag), Age of first household member (age1), Number of unique subscribers in the household (uniqsubs), Range of number of minutes of use (mou_Range), Total number of months in service (months), Mean overage minutes of use (ovrmou_Mean) and Number of active subscribers in household (actvsubs) is found to be top 10 important variables.

CART Model

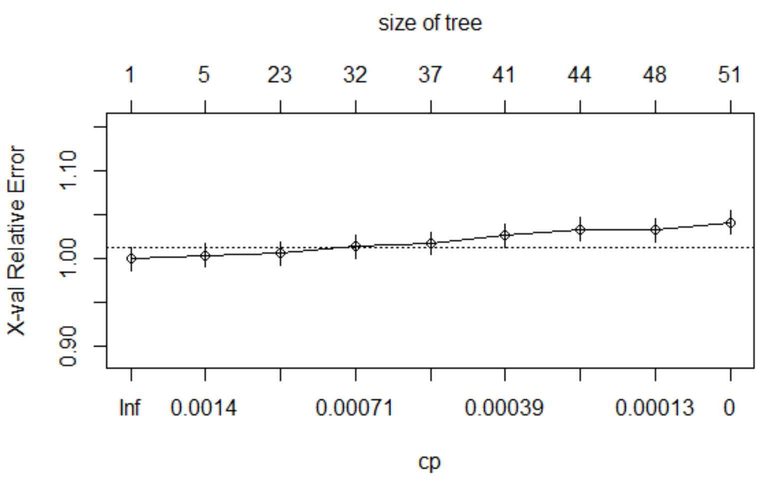
The decision tree method is a powerful and popular predictive machine learning technique that is used for both classification and regression. So, it is also known as Classification and Regression Trees (CART).



[The accurate picture can be obtained from the R Code]

Variables used in tree construction:

age1, area, asl_flag, avg3qty, avg6mou, avg6qty, avgmou, avgqty, change_mou, comp_vce_Mean, csa, da_Range, drop_blk_Mean, drop_vce_Mean, eqpdays, hnd_price, hnd_webcap, iwylis_vce_Mean, months, mou_Mean, mou_opkv_Range, mou_Range, ovrmou_Mean, ovrrev_Mean, prizm_social_one, rev_Mean, totcalls, totmrc_Mean, totrev and uniqsubs



The model predicted well in train and test datasets with an accuracy of 77.22% on train data and 73.68% on test data

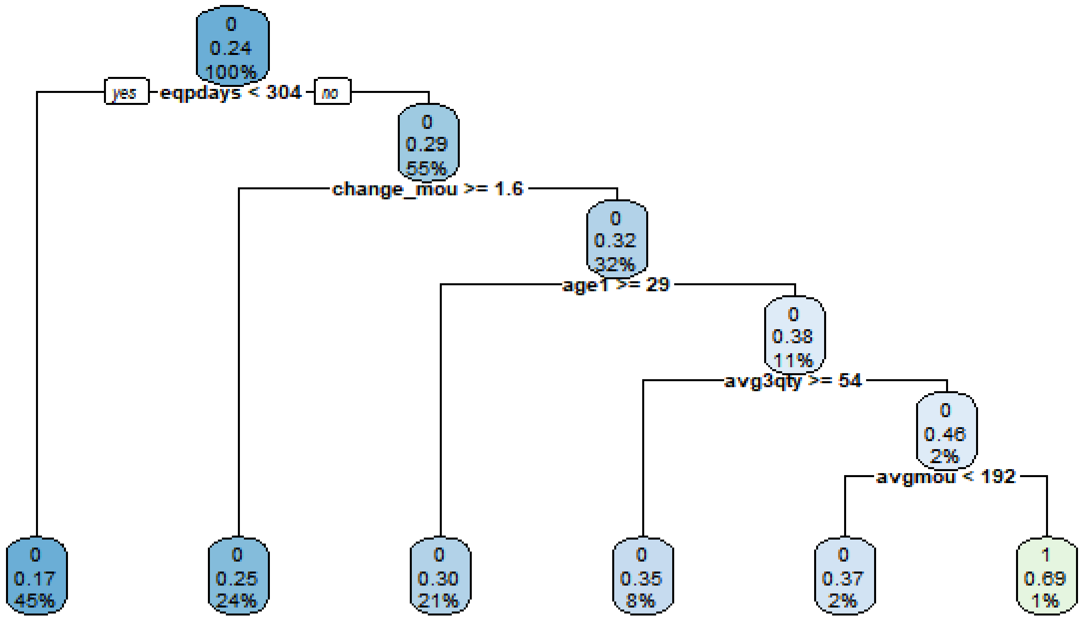
Confusion Matrix:

```
  predCT_test
      0    1
0 5577  469
1 1625  284
```

We will have to prune the data considering .002 as the pruned parameter from the rpart plot

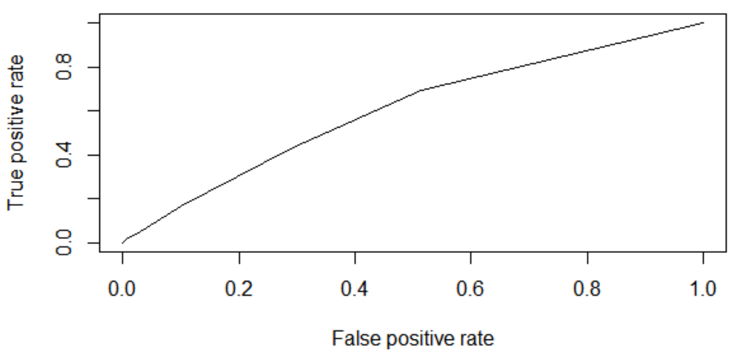
For Pruned Data:

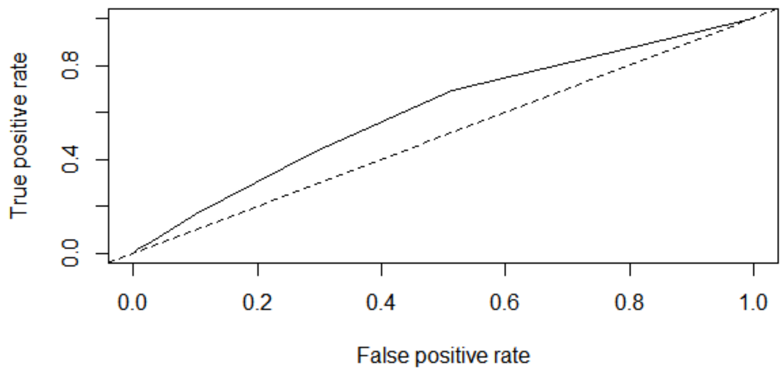
Variables used in tree construction:
[1] age1 avg3qty avgmou change_mou eqpdays



[The accurate picture can be obtained from the R Code]

ROC for pruned tree





Plotting AUC:

Area under the curve is around 0.6025479

CART Model is around to 75.98% accurate in predicting churn on test data

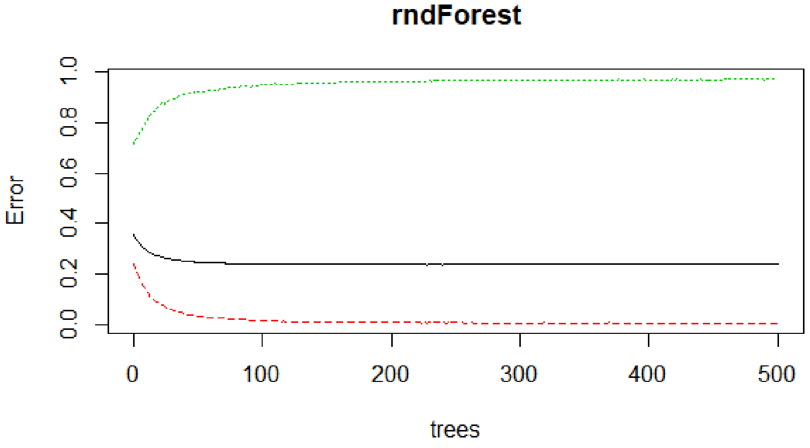
Factors categorized under “Minutes of usage” and “Equipment characteristics” such as Billing adjusted total minutes of use over the life of the customer (Adjmou), Age of first household member (age1), Average monthly minutes of use over the previous three months (avg3mou), Average monthly number of calls over the previous three months (avg3qty), Average monthly minutes of use over the previous six months (avg6mou), Average monthly number of calls over the previous six months (avg6qty), Average monthly minutes of use over the life of the customer (avgmou), Average monthly number of calls over the life of the customer (avgqty), Percentage change in monthly minutes of use vs previous three month average (change_mou), Mean number of completed voice calls (comp_vce_Mean), Number of days (age) of current equipment (eqpdays), Current handset price (hnd_price), Handset web capability (hnd_webcap), Total number of months in service (months), Mean number of monthly minutes of use (mou_Mean), Range of number of minutes of use (mou_Range) and Mean number of attempted voice calls placed (plcd_vce_Mean) affect customer churn as per CART model

Random Forest Model

The Random Forest is a classification algorithm that contains several decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset.
It uses bagging and feature randomness when building each individual tree to try to create an uncorrelated forest of trees whose prediction is more accurate than that of any individual tree.

Performed Random Forest analysis on train dataset. Here Type 1 error is extremely high.

```
Confusion matrix:
      0   1 class.error
0 13999 109 0.007726113
1  4325 130 0.970819304
```



The model predicted well on train data with an accuracy of 97.02%

```
       0     1
 0 14108     0
 1   553  3902
```

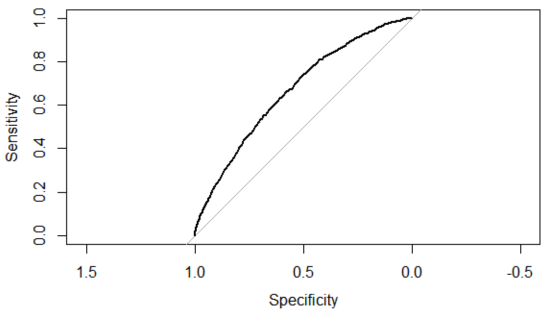
The model predicted on test data has an accuracy of 76.38%

```
     0    1
0 6012   34
1 1845   64
```

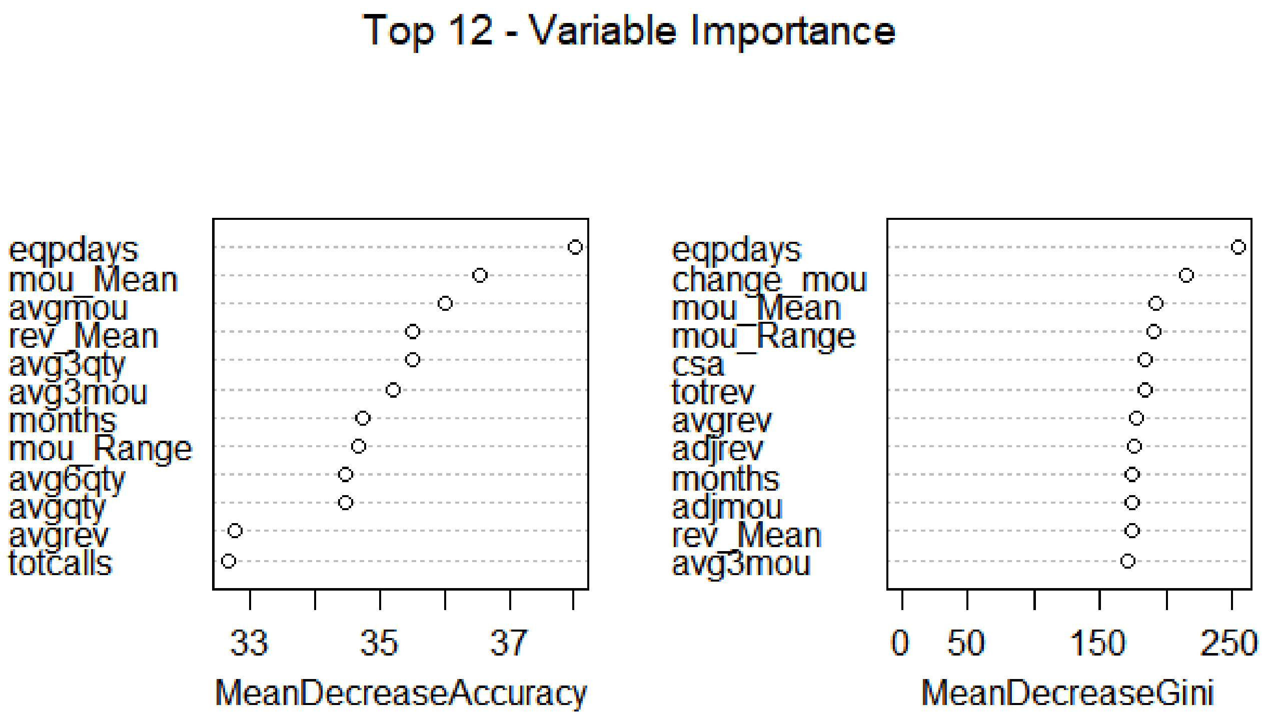
We tune the random forest model and find that OOB Error is least at mtry=8 and hence performed random forest again on the dataset.
OOB error has improved from 23.89% to 23.79%. Classification error has also decreased in the new model.

```
Confusion matrix:
      0   1 class.error
0 14007 101 0.007159059
1  4315 140 0.968574635
```

The tuned model predicted customer churn on test dataset with an accuracy of 76.34%.



Area under the curve: 0.6686



Age of current equipment/ service, Minutes of usage, monthly revenue are few important factors affecting customer churn as per the Random Forest model.

6. Observations & Recommendations

Comparison of various models

Model	AUC	Accuracy of prediction on test dataset
GLM		59.21%
LDA	0.6273	75.80%
Logistic Regression Model	0.6255	75.86%
CART	0.6025	75.98%
Random Forest	0.6686	76.34%

Considering model accuracy, AUC value, number of false negative predictions and various other factors, Random Forest model is found to be the best model to predict customer churn.

Observation:

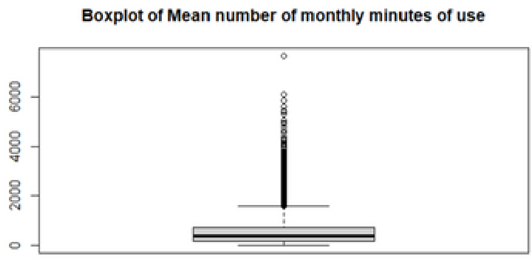
1. Minutes of usage and Revenue influence churn
2. Customer's and customer's equipment characteristics influence churn
3. Cost and billing impact customer behaviour
4. Network and service quality highly influence customer churn
5. The intercept is significant. This constitutes to the effects of levels of categorical variables that were removed in the model

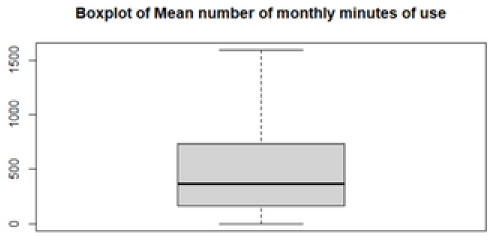
Recommendation:

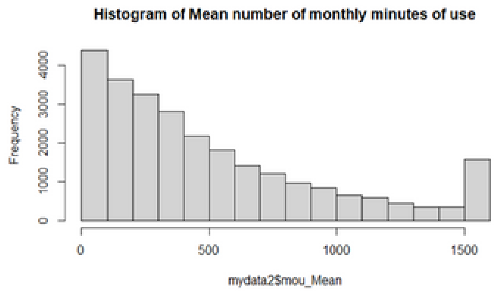
1. Variable 'Minutes of usage' is highly significant variable. If the minutes of usage falls, there is high probability that the customer will churn. Hence, need to offer more flexible and customised plans to encourage increase in mean time of usage
2. Rate of plan needs to be revised proactively to increase revenue and profit of the organisation
3. Need to work on the network quality as dropped and blocked calls are highly significant and influence in customer churn. Any improvement in the network quality can increase customer satisfaction and hence will be able to retain them
4. Targeted marketing and customised plans based on customer characteristics such as age is necessary to reduce churn.

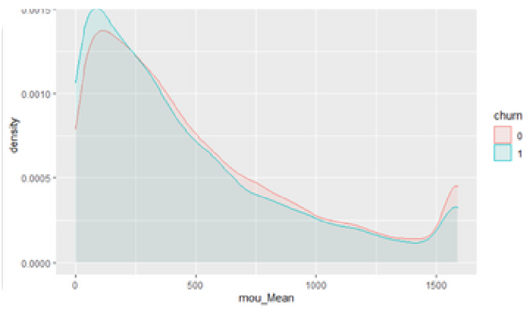
ANNEXURE 1 - DIAGRAMS FOR CONTINOUS VARIABLES

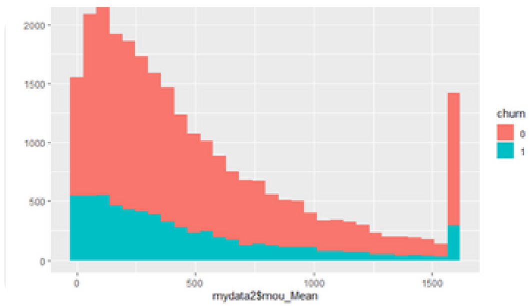
AN1 - Anlaysis of Mean number of monthly minutes of use



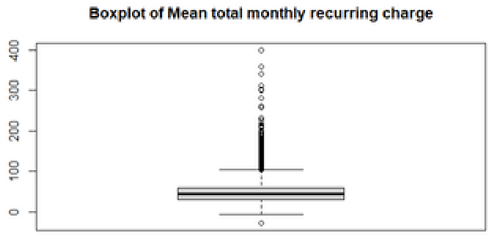


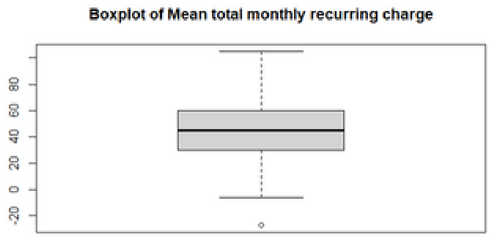


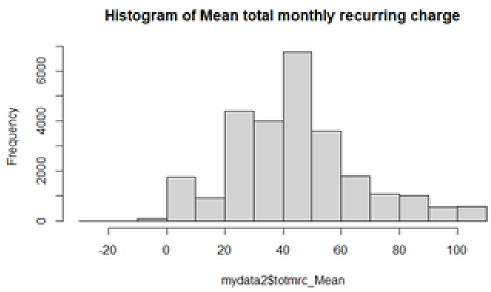


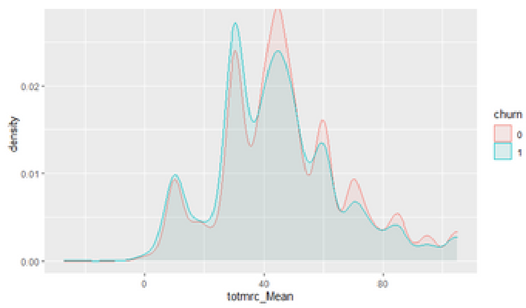


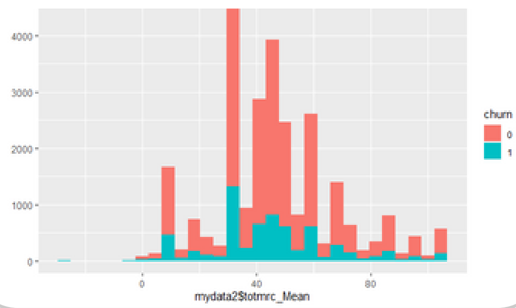
AN2 - Anlaysis of Mean total monthly recurring charge



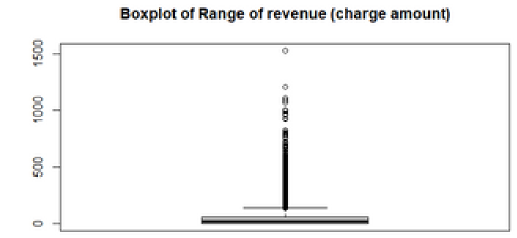


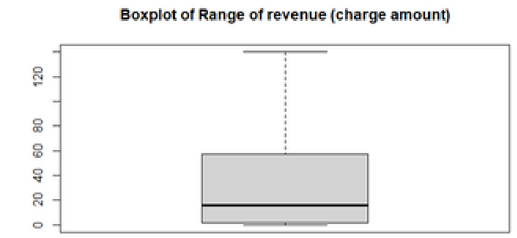


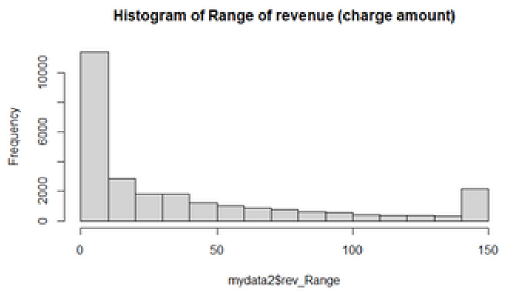


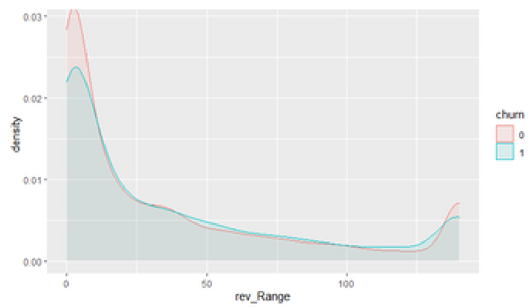


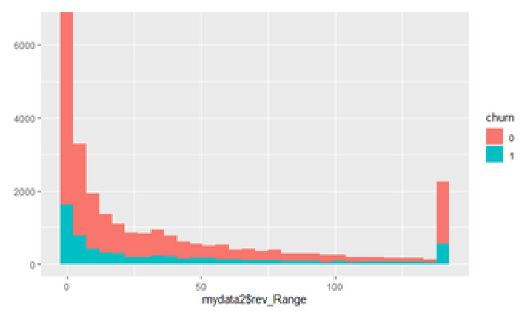
AN3 - Range of revenue (charge amount)



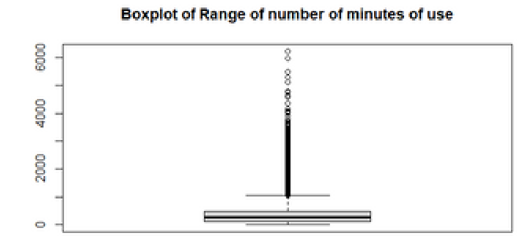


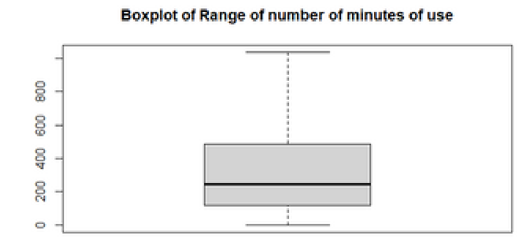


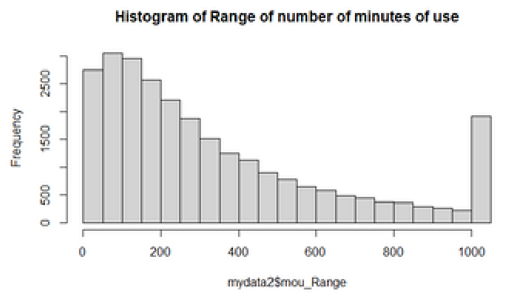


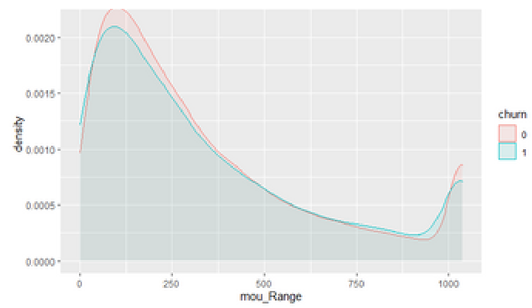


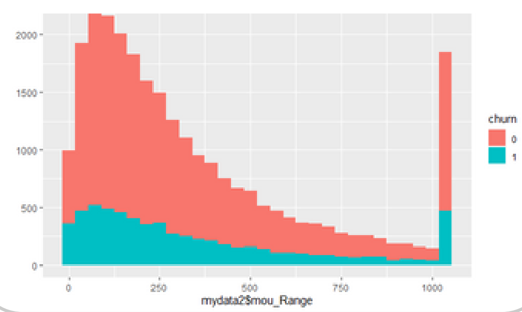
AN4 - Range of number of minutes of use



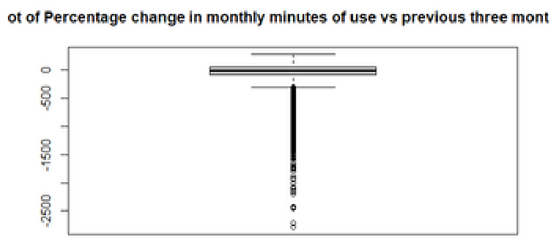


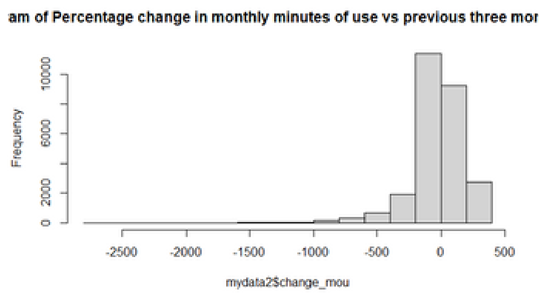


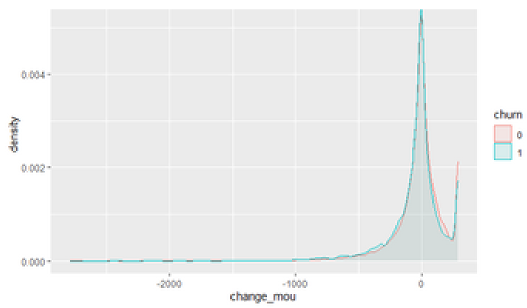


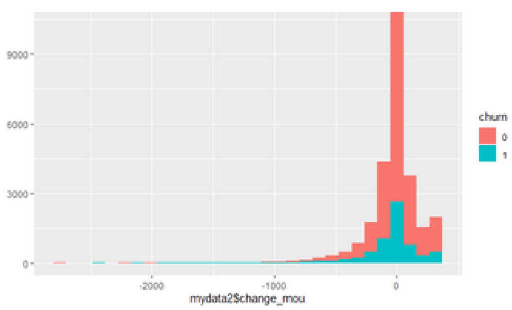


AN5 - Percentage change in monthly minutes of use vs previous 3 months

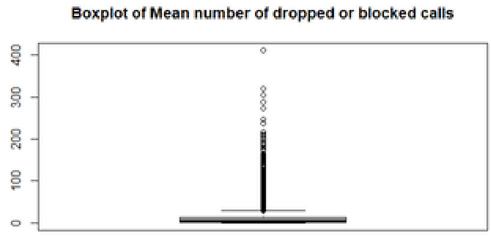


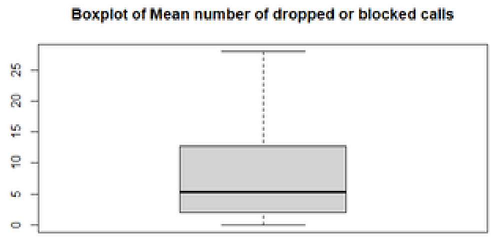


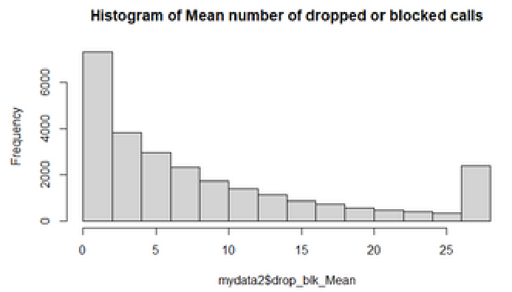


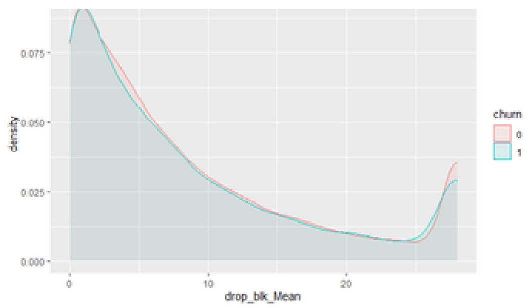


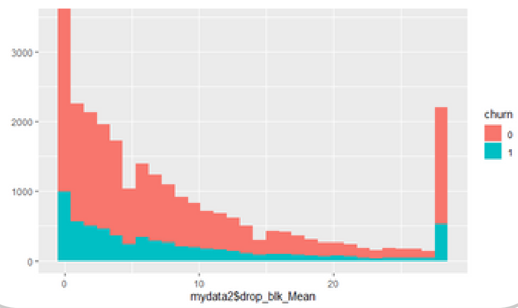
AN6 - Anlaysis of Mean number of dropped or blocked calls



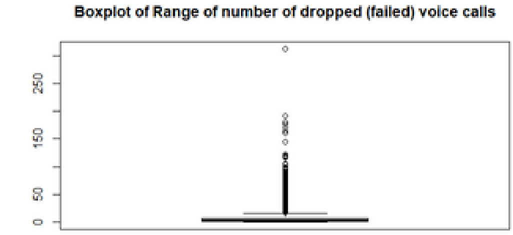


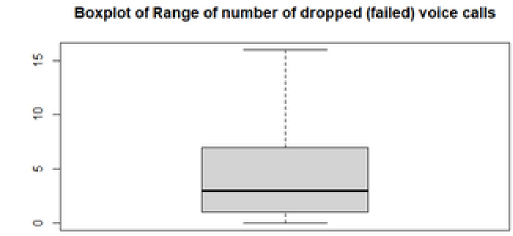


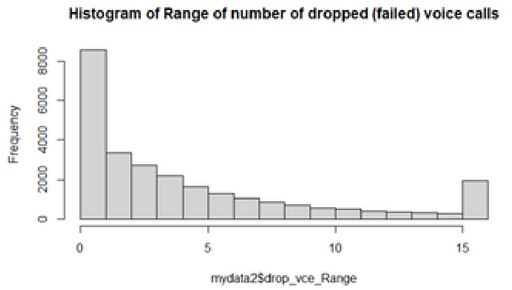


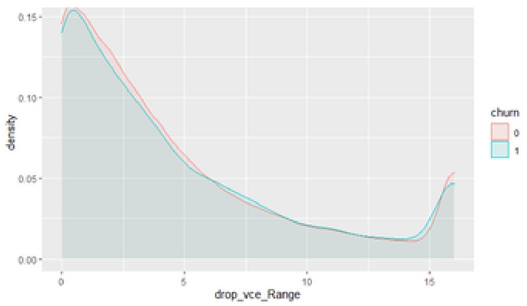


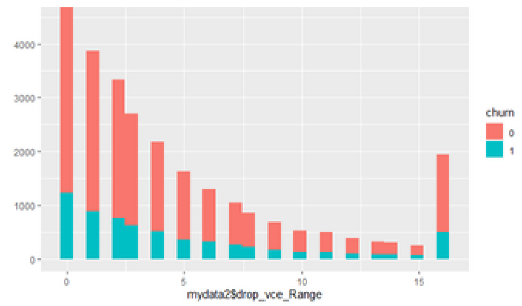
AN7 - Range of number of dropped (failed) voice calls



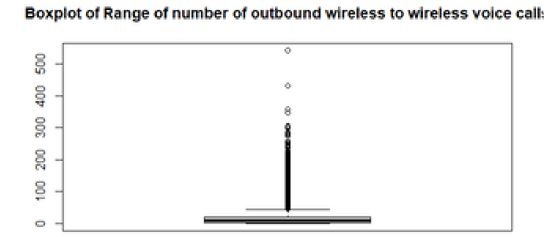


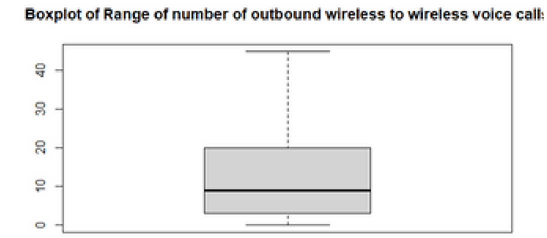


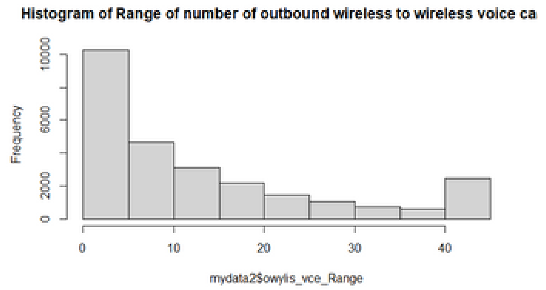


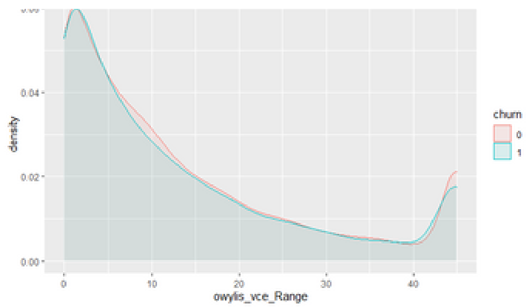


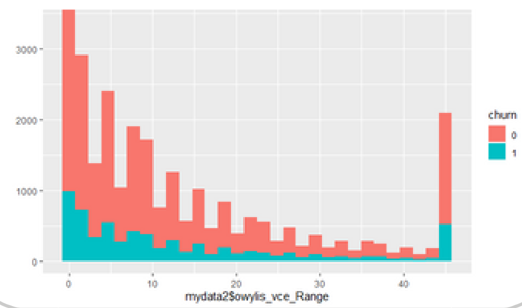
AN8 - Anlaysis of Range of number of outbound wireless to wireless calls



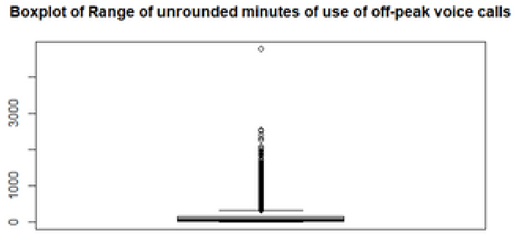


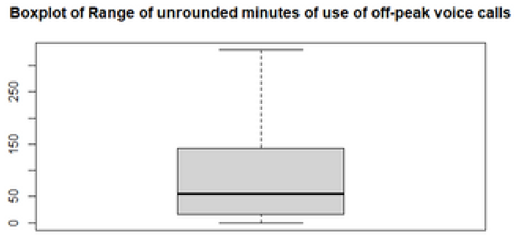


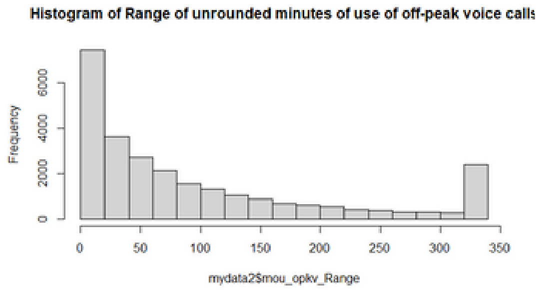


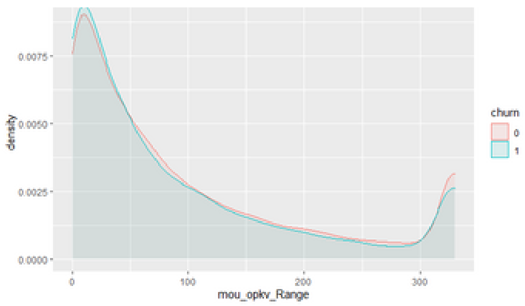


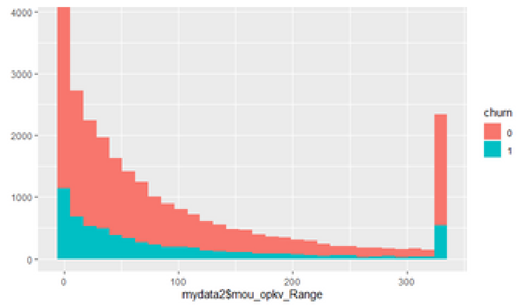
AN9 - Range of unrounded minutes of use off-peak voice calls



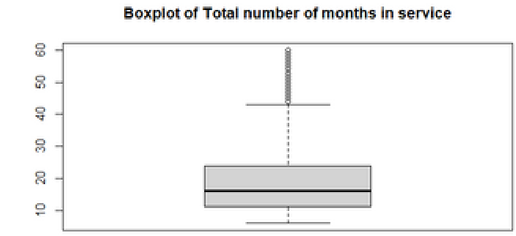


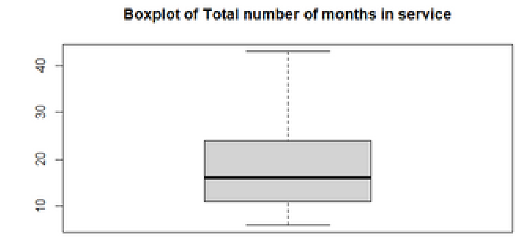


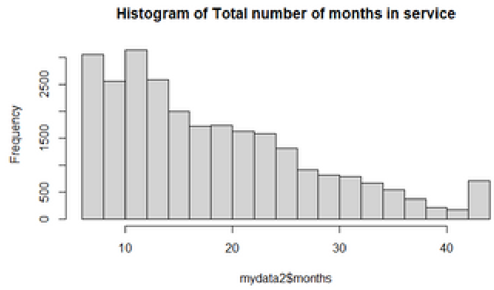


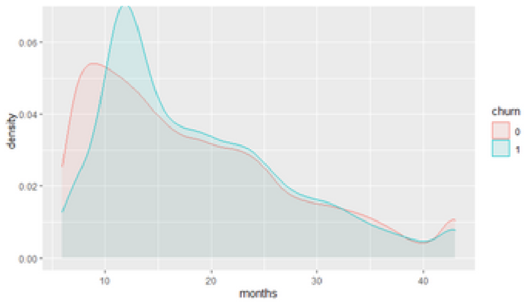


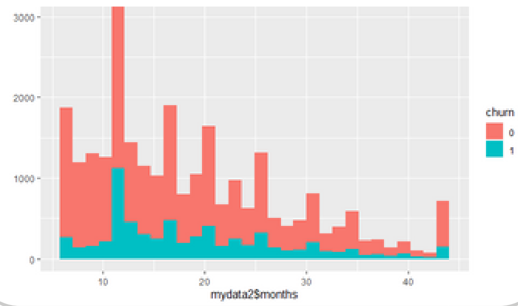
AN10 - Anlaysis of Total number of months in service



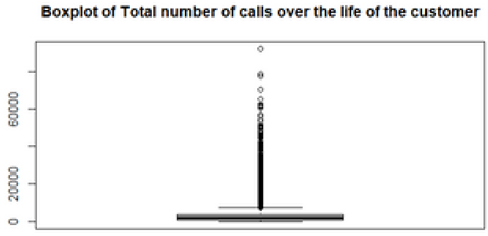


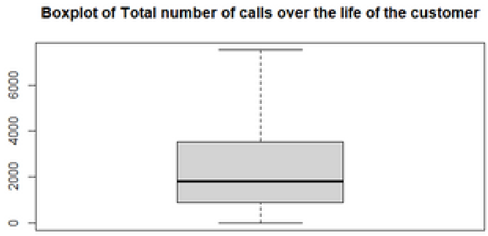


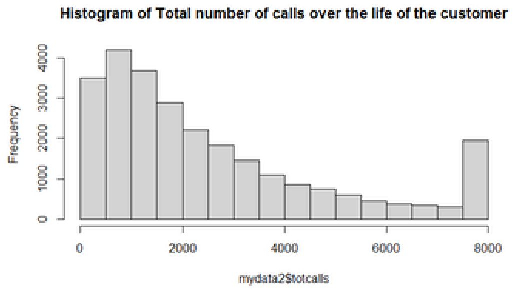


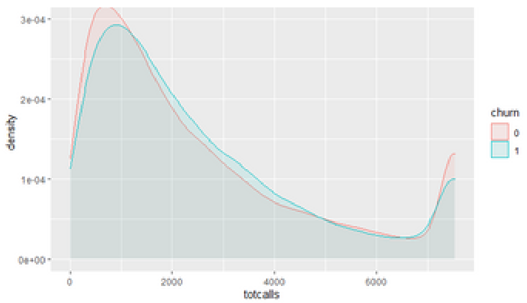


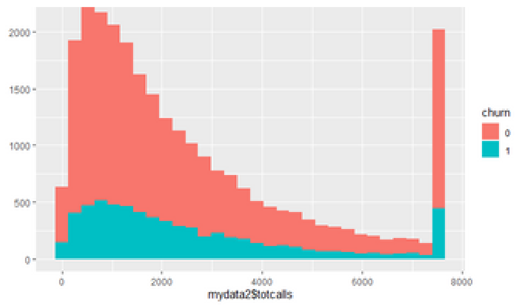
AN11 - Range of unrounded minutes of use off-peak voice calls



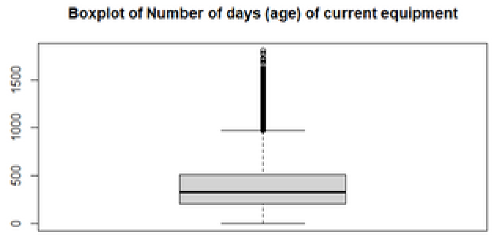


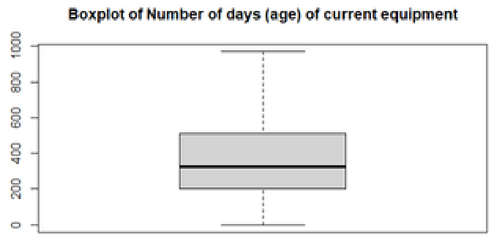


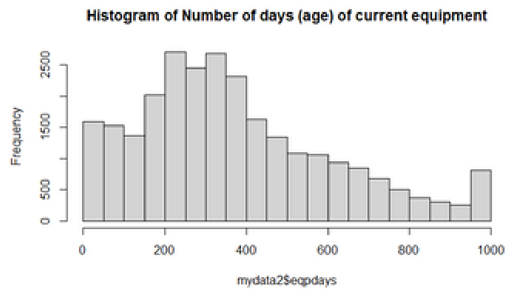


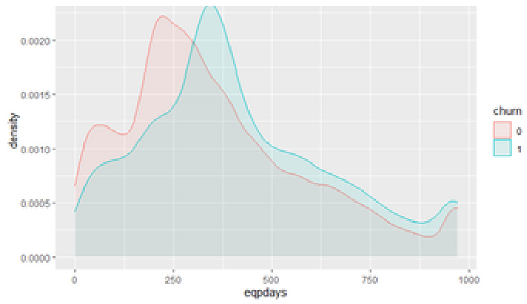


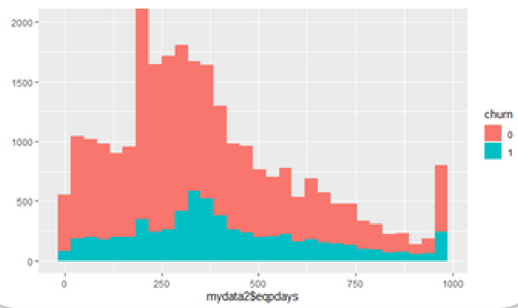
AN12 - Anlaysis of Number of days of current equipment



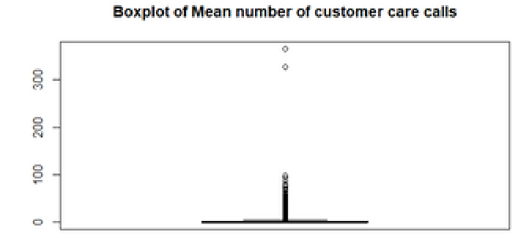


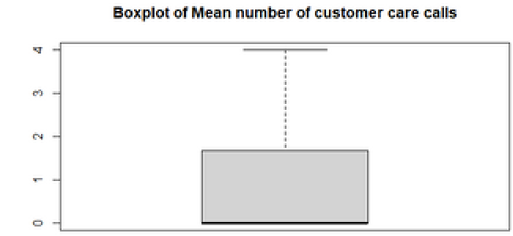


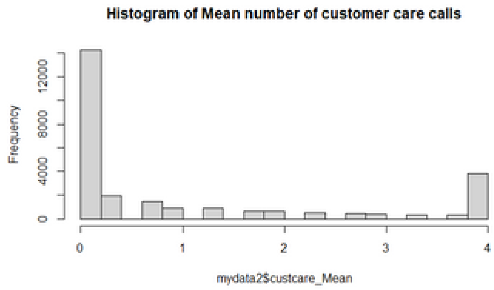


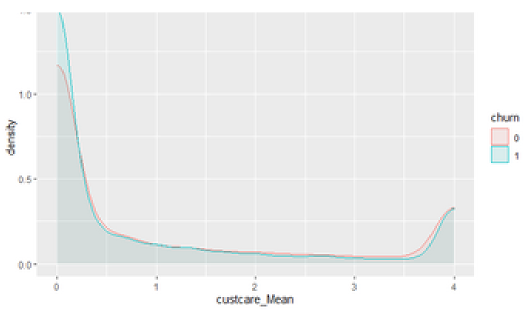


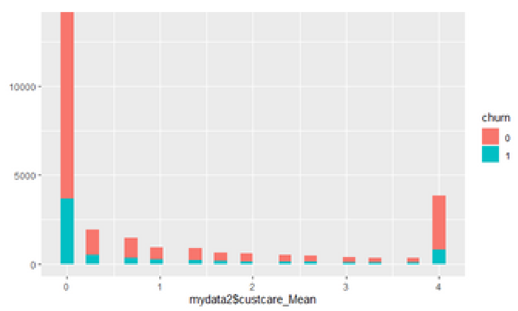
AN13 - Range of Mean number of customer care calls



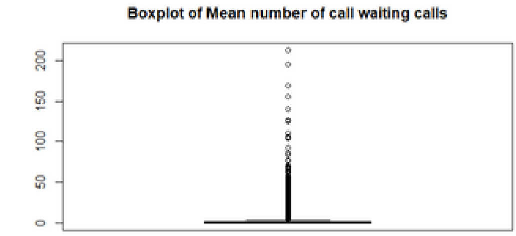


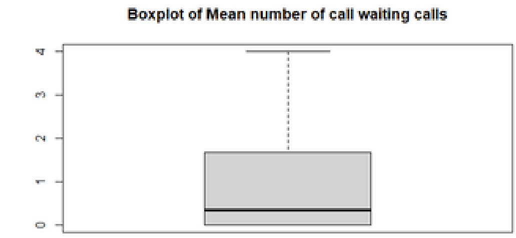


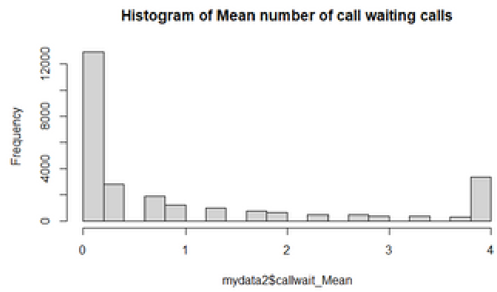


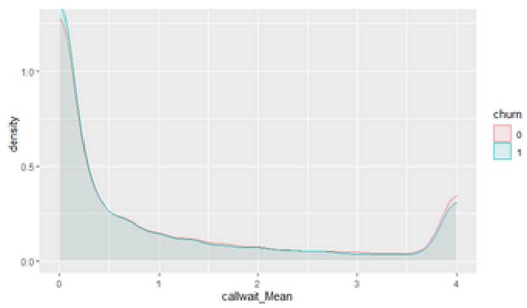


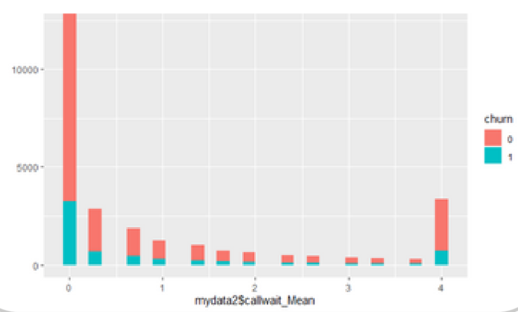
AN14 - Anlaysis of Mean number of call waiting calls



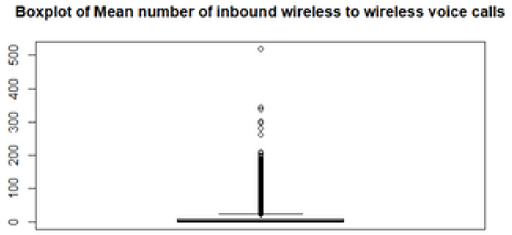


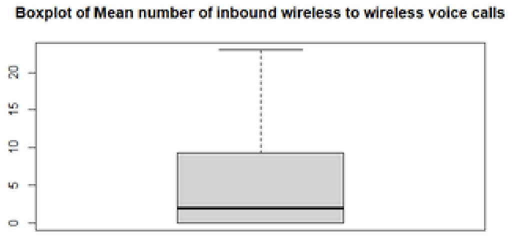


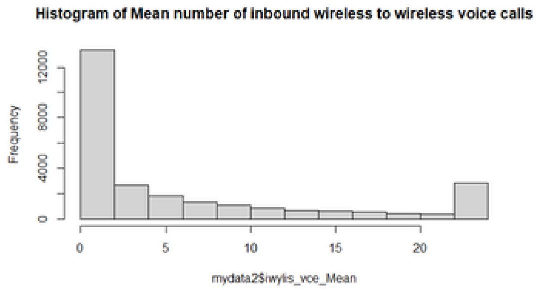


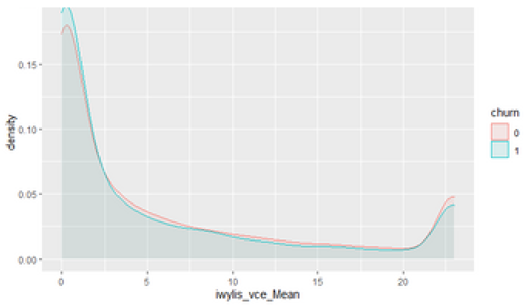


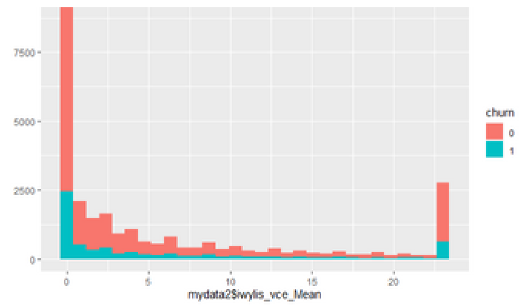
AN15 - Range of inbound wireless to wireless voice calls



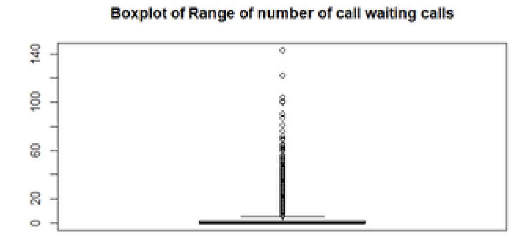


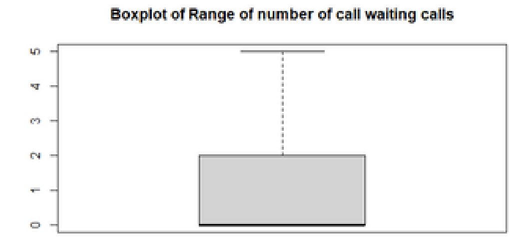


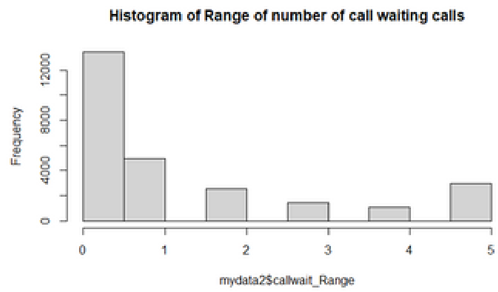


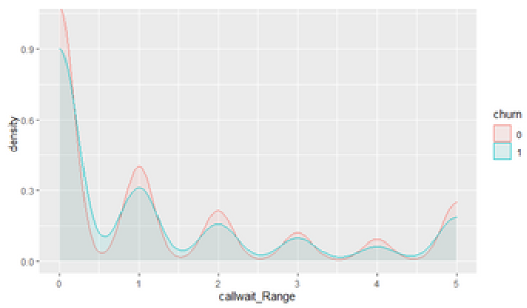


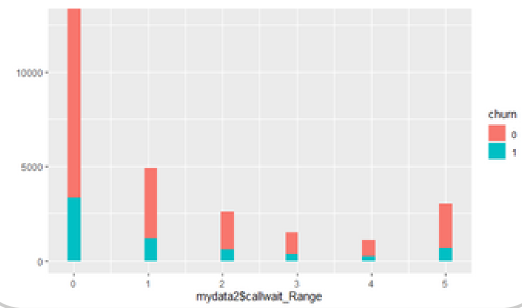
AN16 - Anlaysis of Range of number of call waiting calls



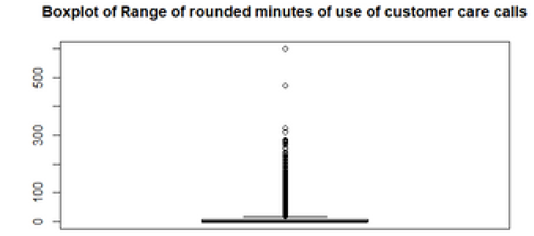


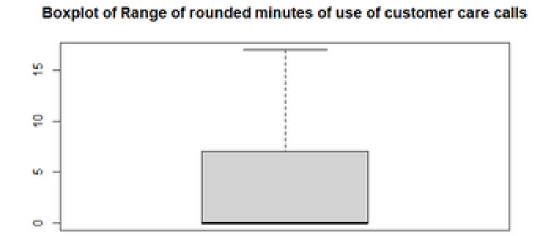


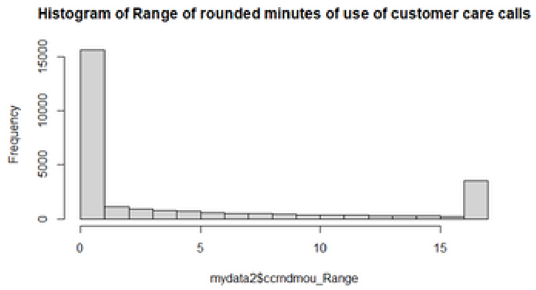


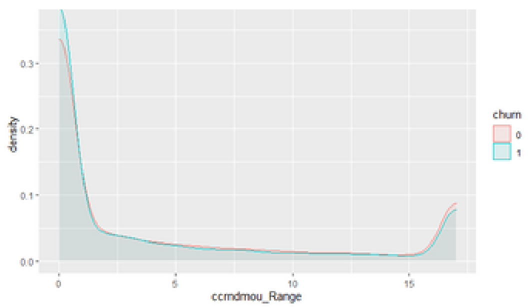


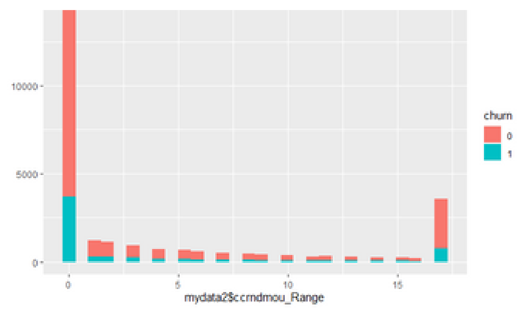
AN17 - Range of unrounded minutes of use off-peak voice calls



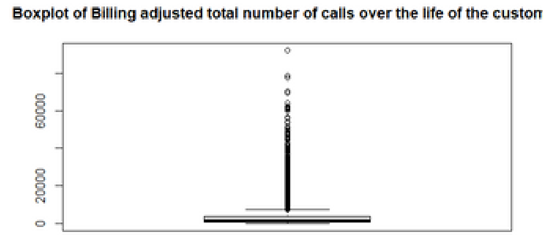


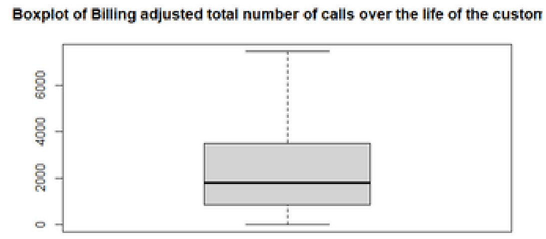


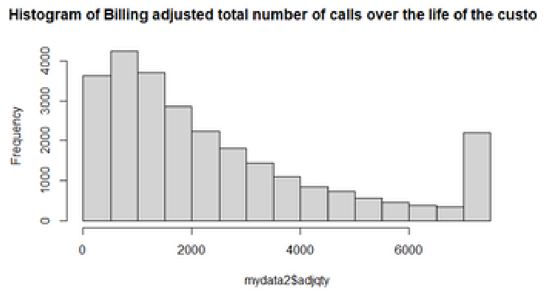


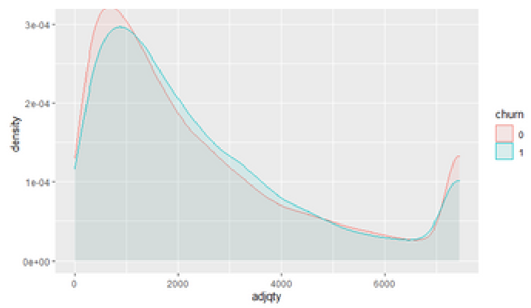


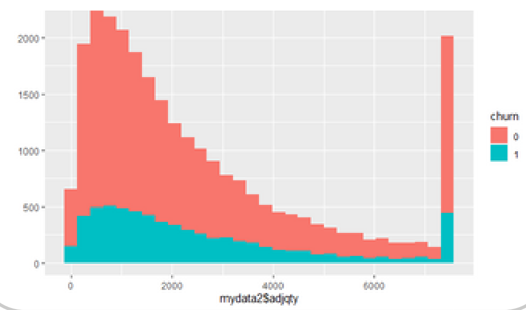
AN18 - Anlaysis of Total number of calls over the life of the customer



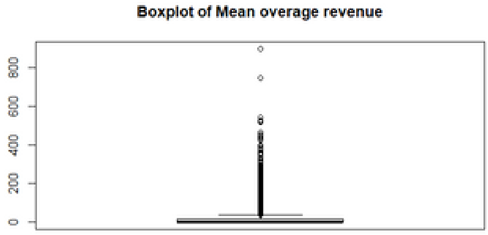


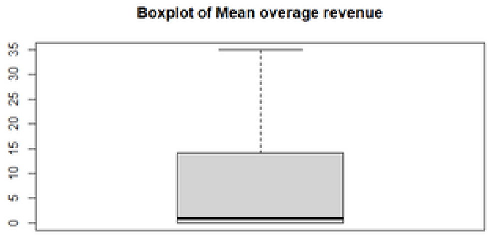


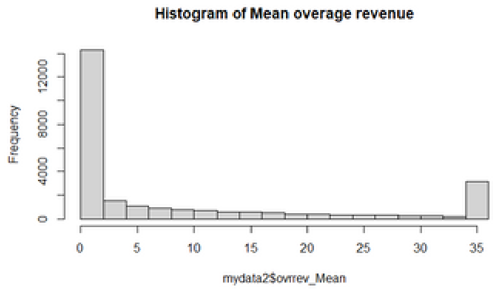


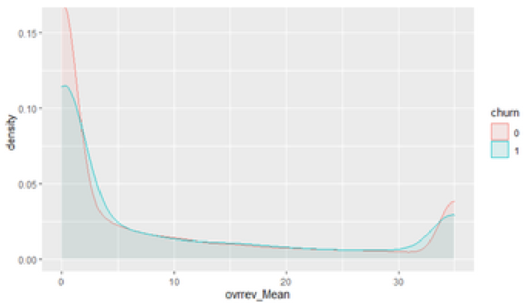


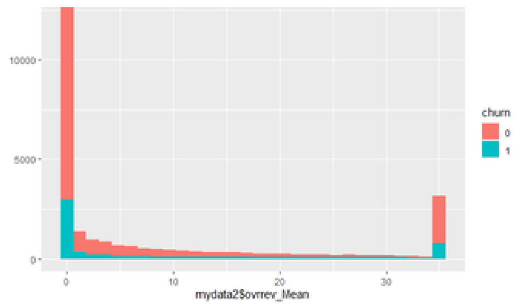
AN19 - Range of mean overage revenue



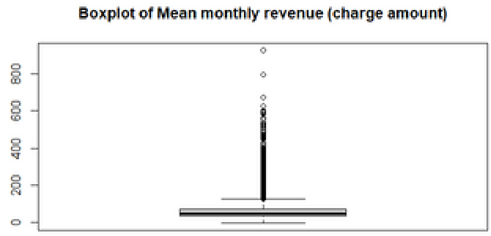


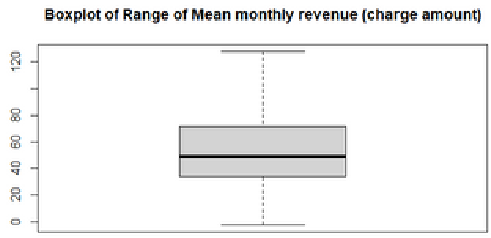


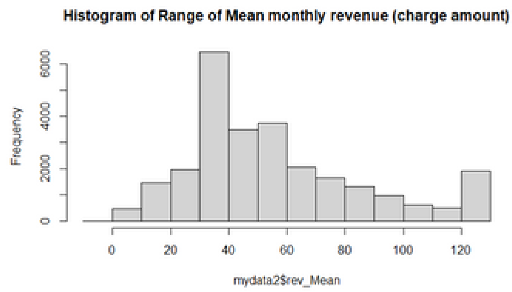


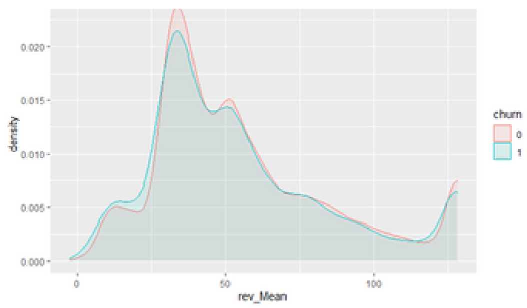


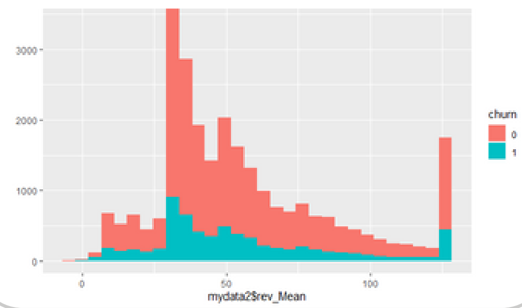
AN20 - Anlaysis of Mean monthly revenue (charge amount)



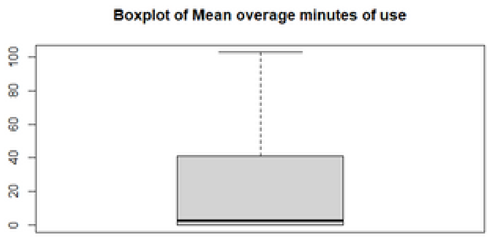


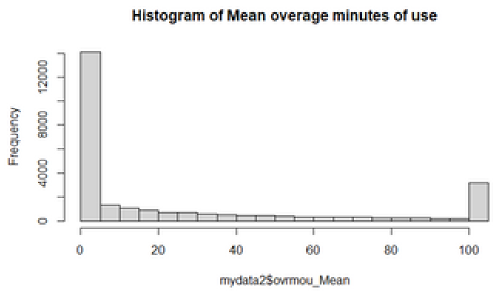


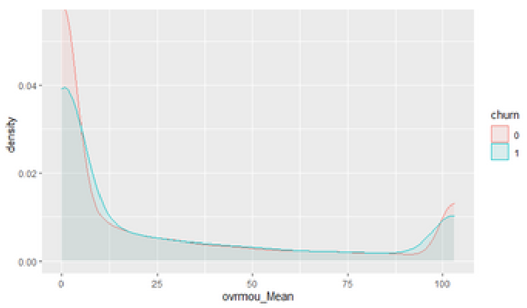


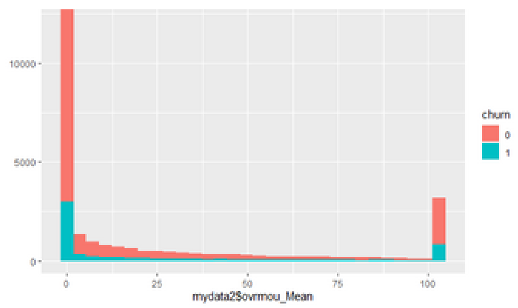


AN21 - Range of mean overage minutes of use

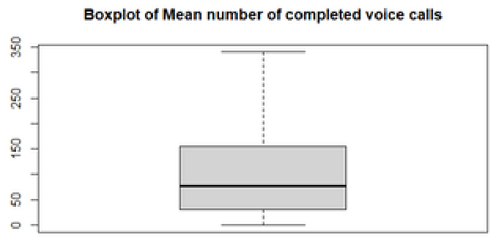


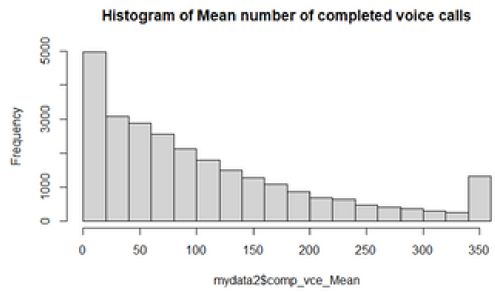


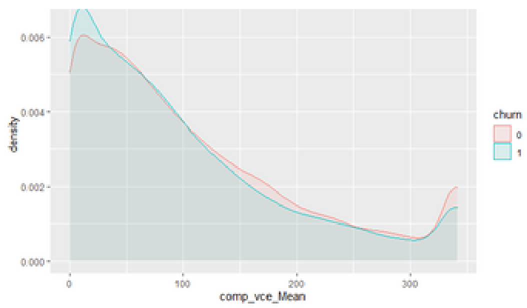


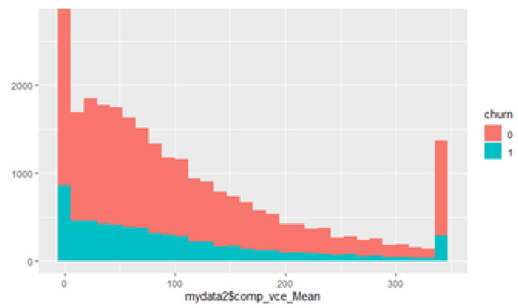


AN22 - Anlaysis of mean number of completed voice calls

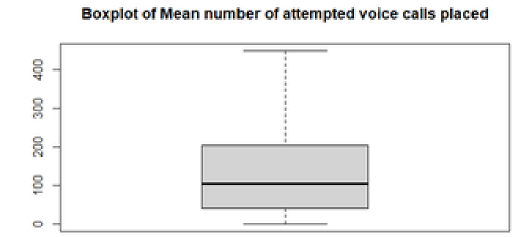


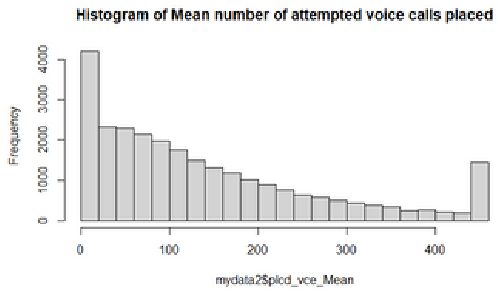


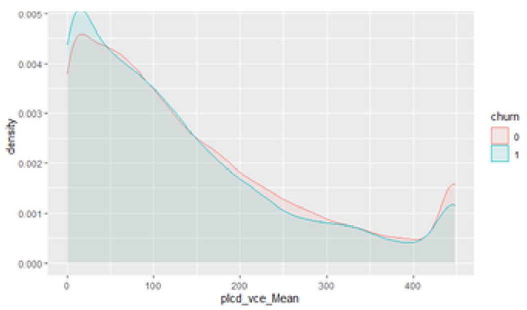


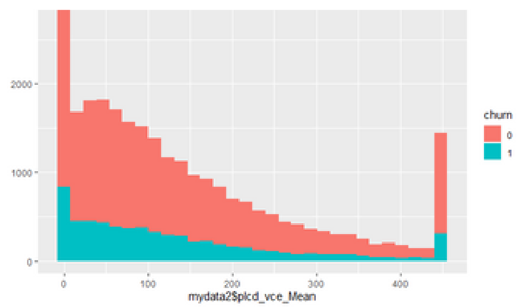


AN23 - Mean numner of attempted voice calls placed

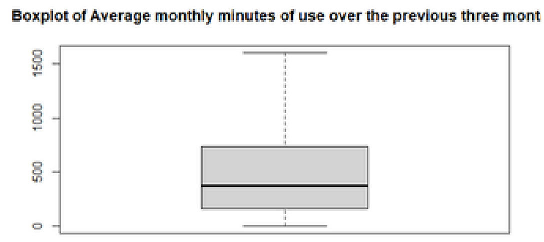


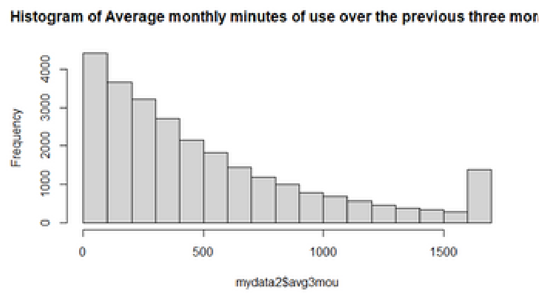


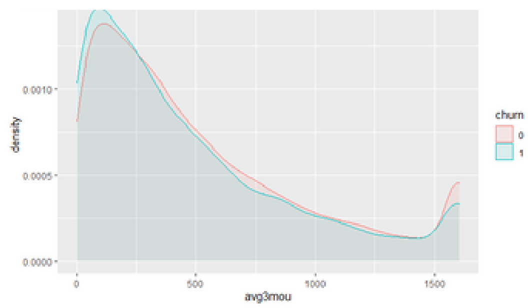


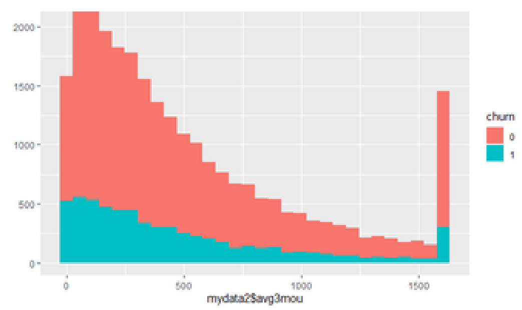


AN24 - Average monthly minutes of use over previous three months

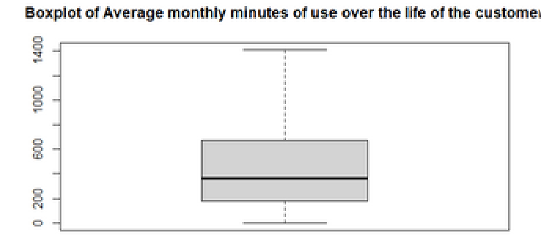


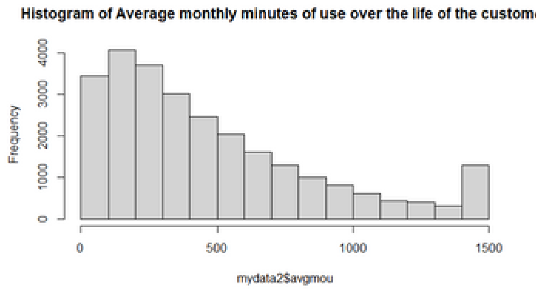


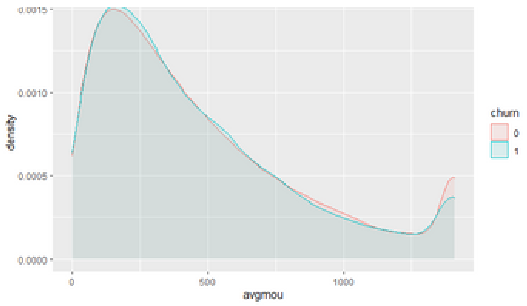


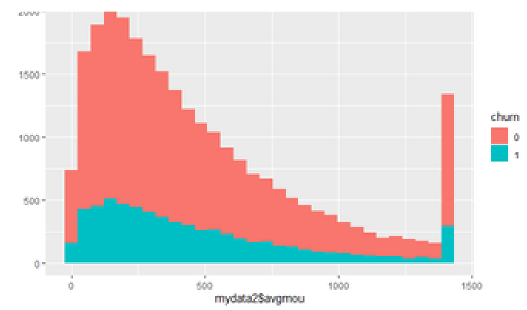


AN25 - Average monthly minutes of use over the life of the customer

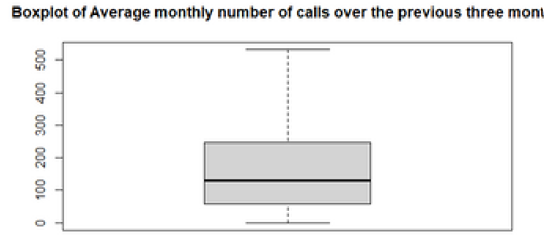


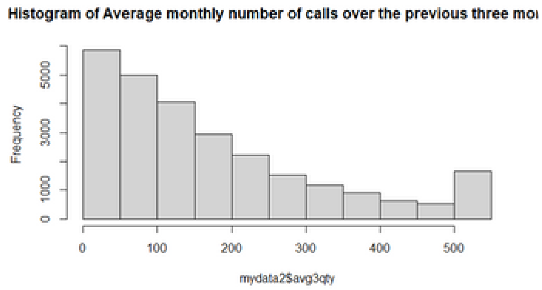


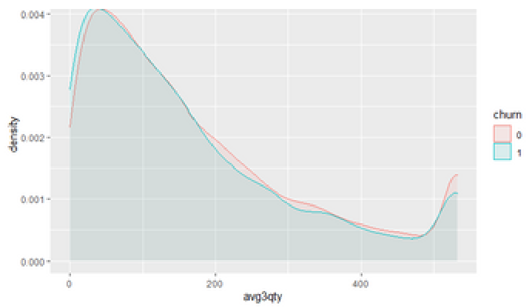


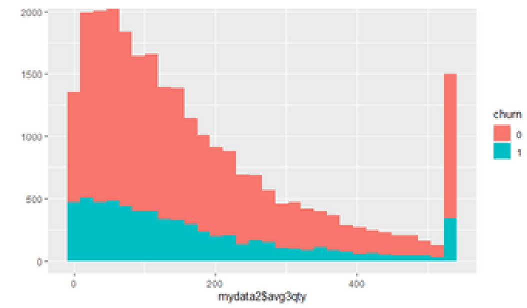


AN26 - Anlaysis of monthly number of calls over the previous three months

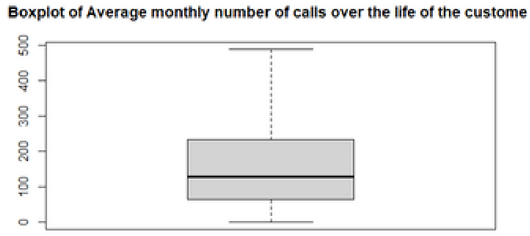


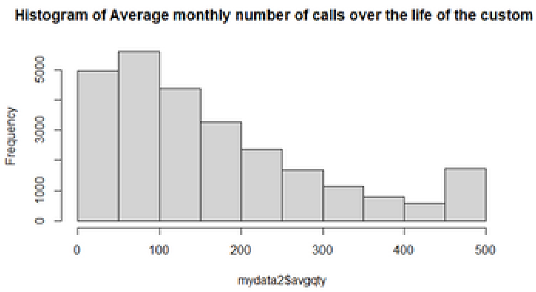


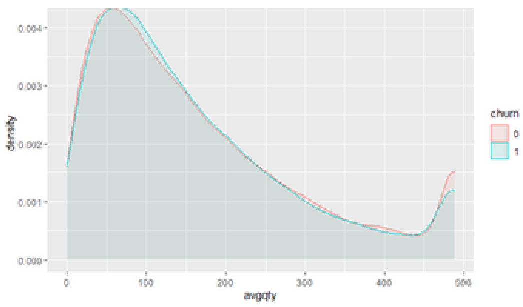


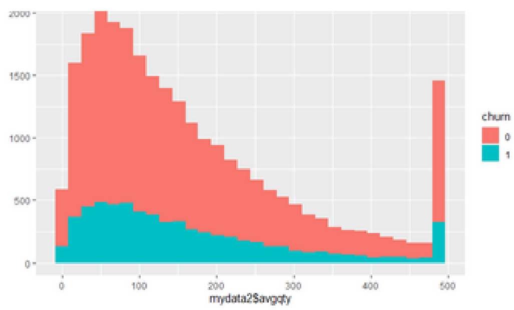


AN27 - Analysis of Average monthly number of calls over the life

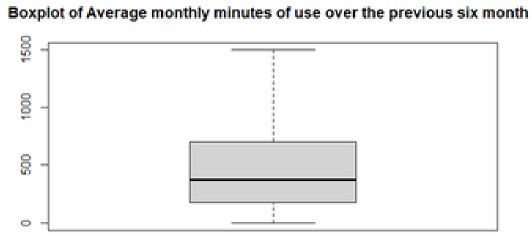


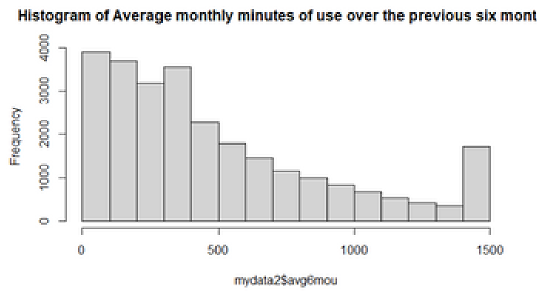


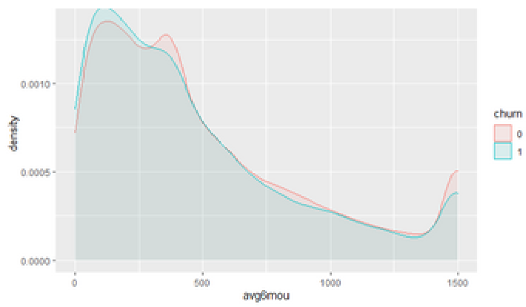


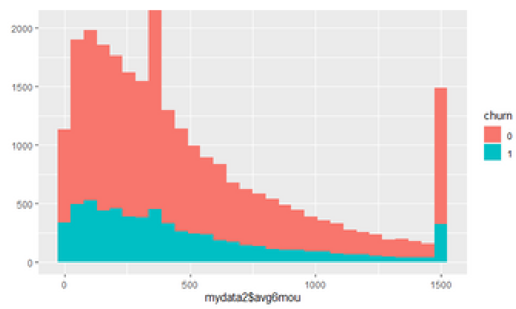


AN28 - Anlaysis of Average monthly minutes of use over six months

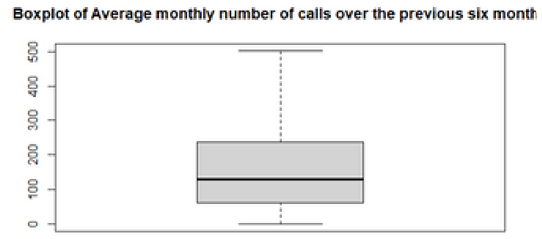


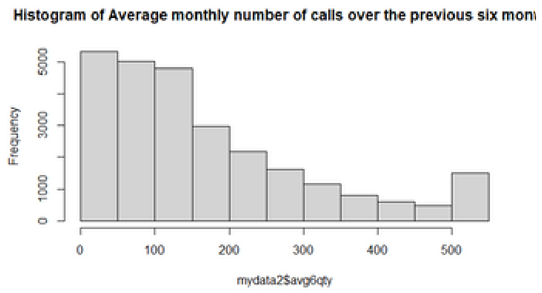


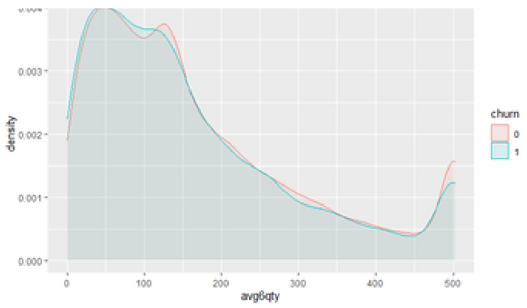


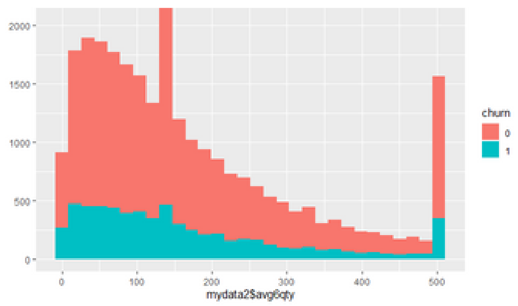


AN29 - Average monthly number of calls over six months

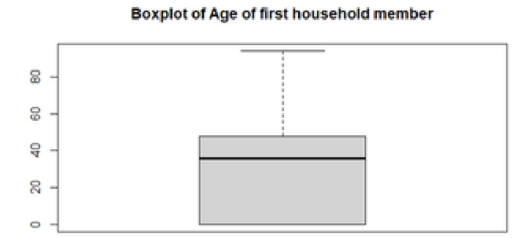


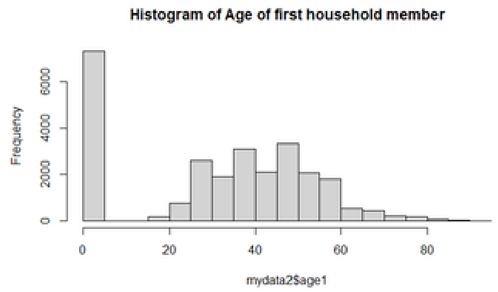


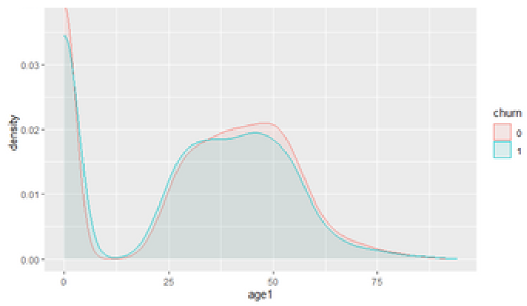


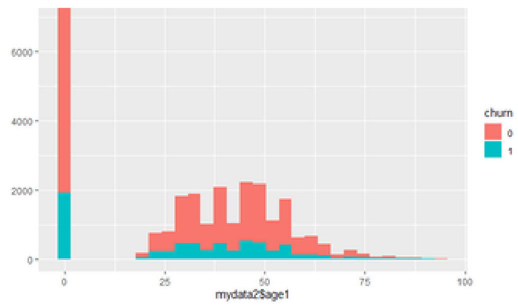


AN30 - Age of first household member

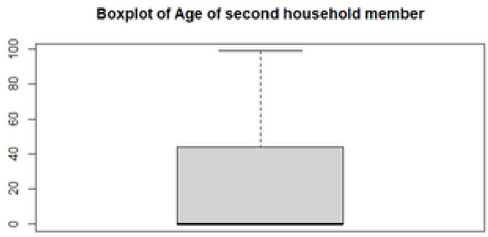


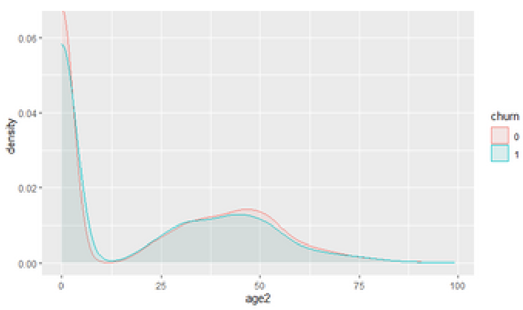


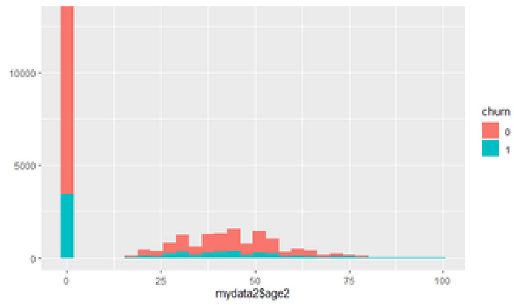




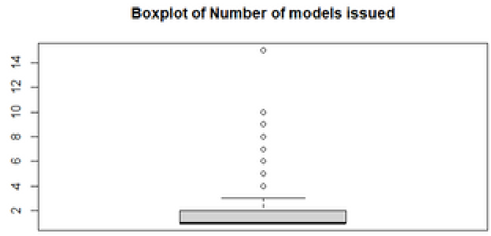
AN31 - Age of second household member

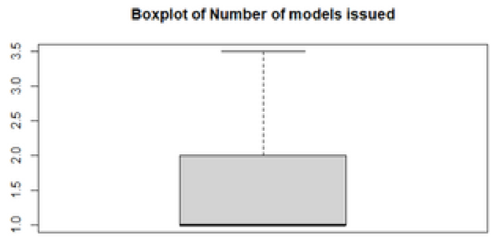


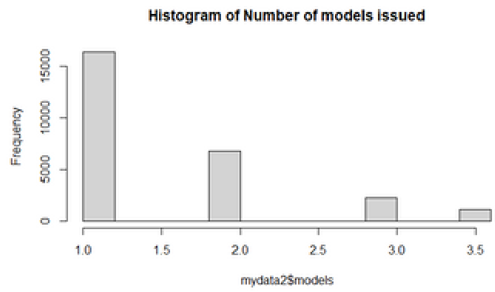


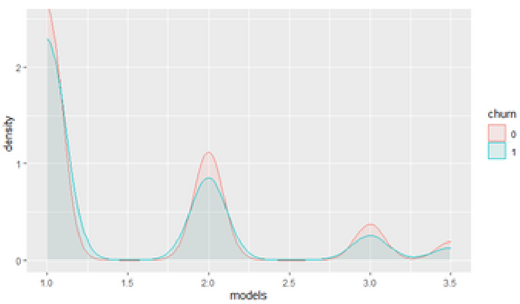


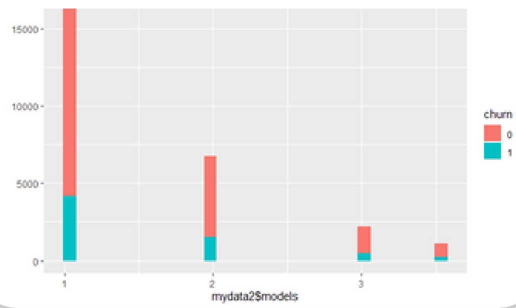
AN32 - Anlaysis of Number of models issued



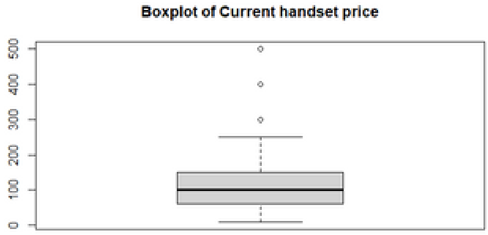


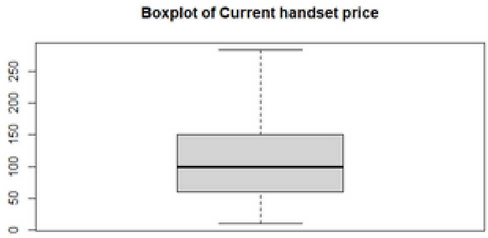


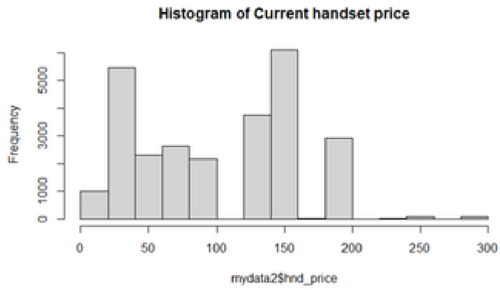


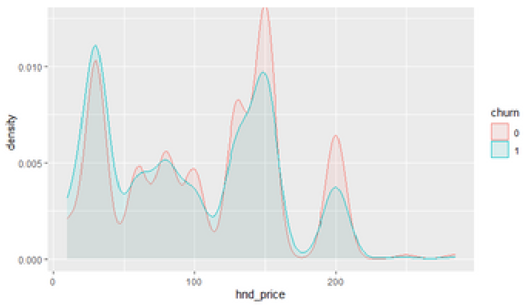


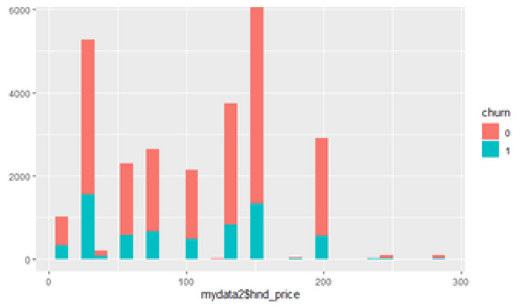
AN33 - Analysis of current handset price



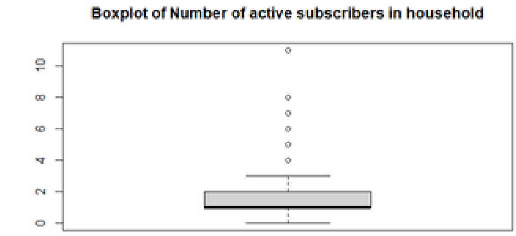


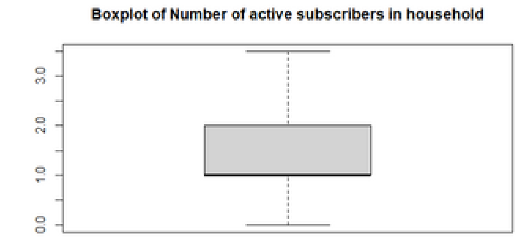


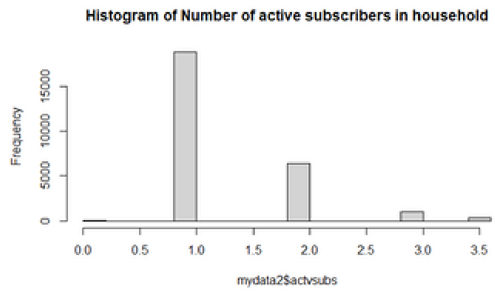


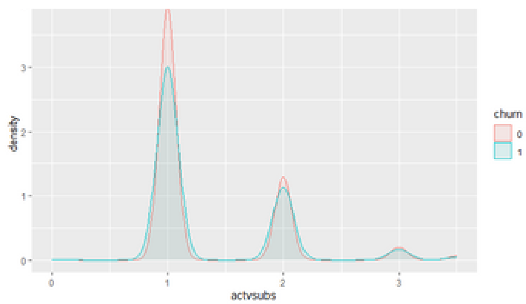


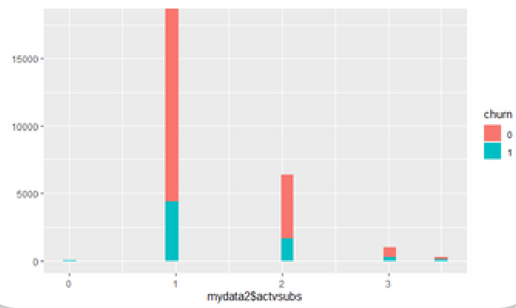
AN34 - Anlaysis of active subscribers in household



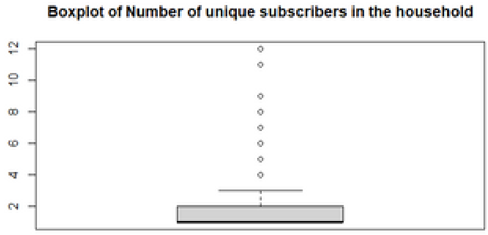


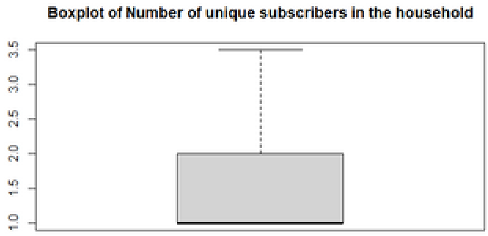


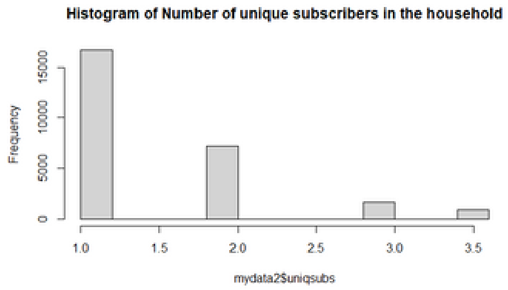


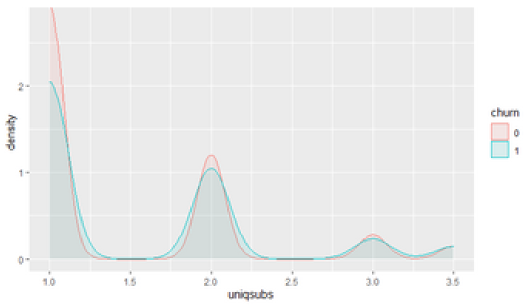


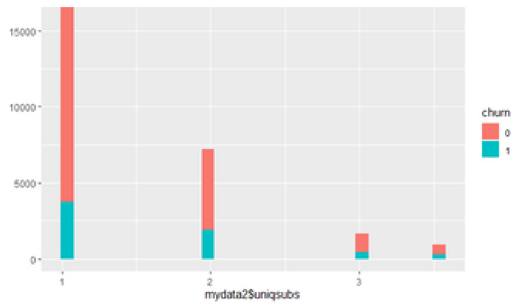
AN35 - Analysis of unique users in the household



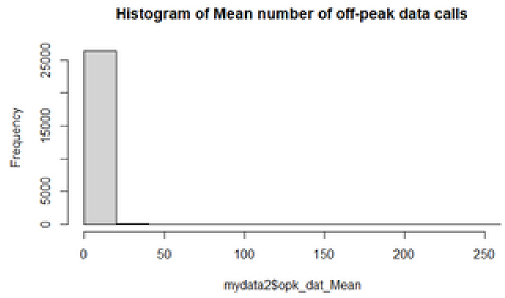


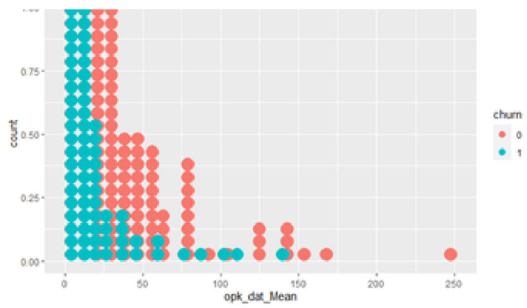


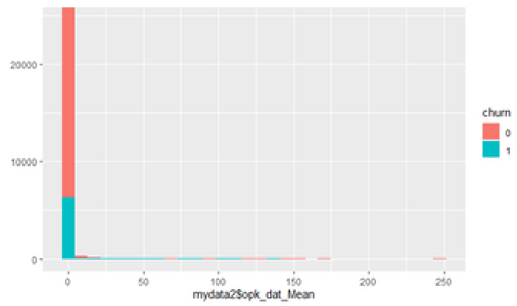




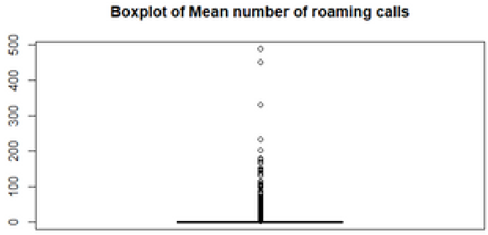
AN36 - Anlaysis of mean number of off-peak data calls

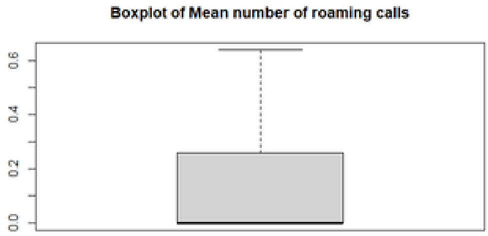


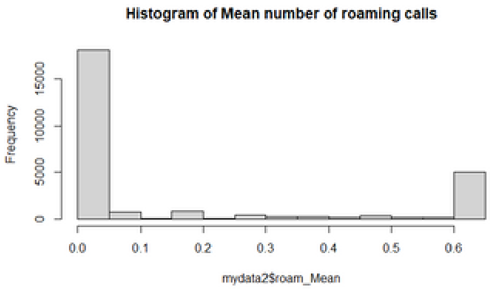


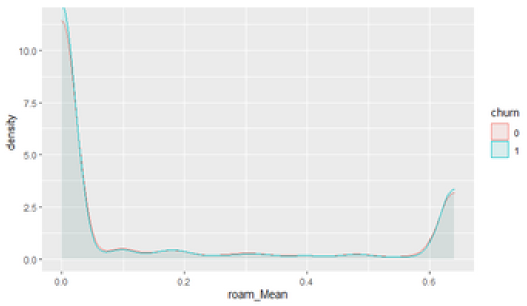


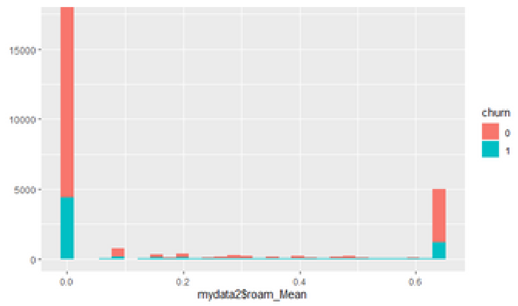
AN37 - Analysis of mean number of roaming calls



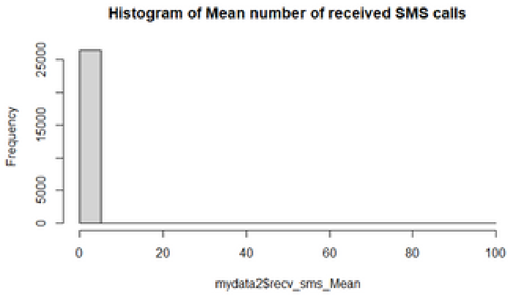


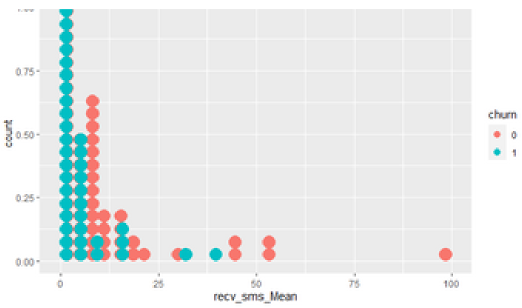


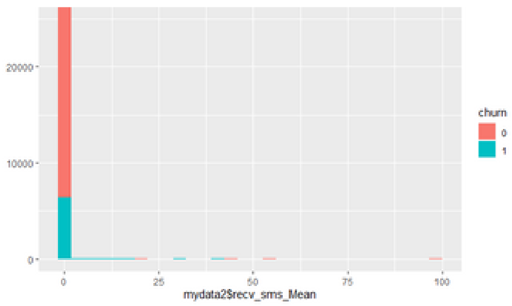




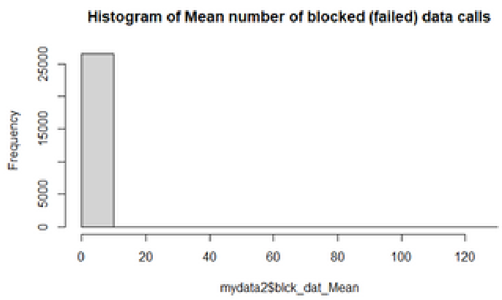
AN38 - Anlaysis of mean number of received SMS calls

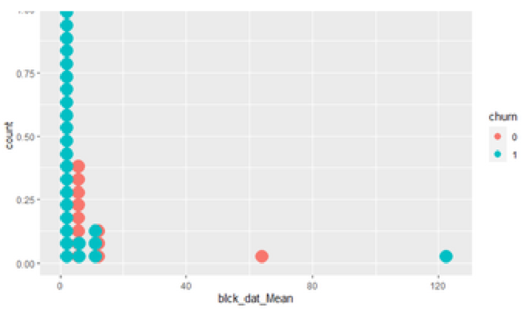


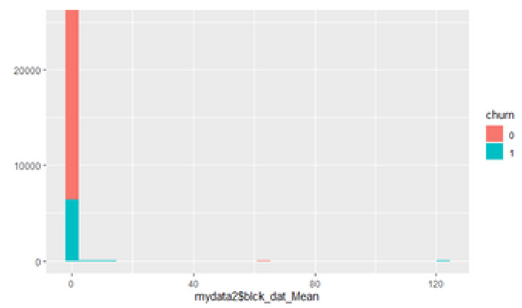




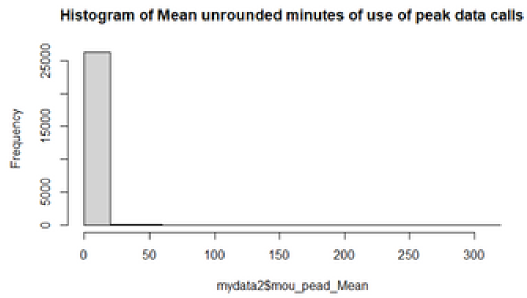
AN39 - Analysis of mean number of blocked (failed) data calls

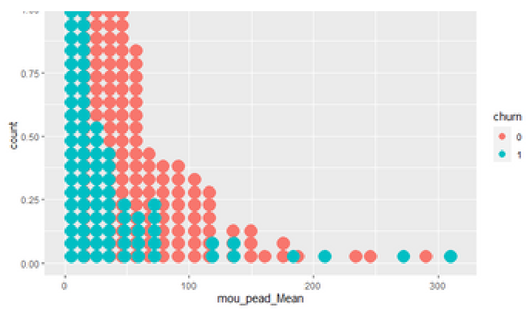


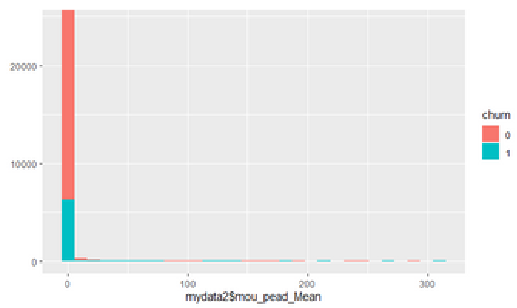




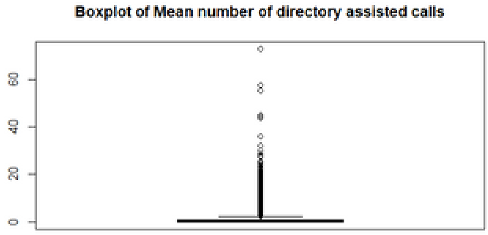
AN40 - Anlaysis of mean minutes of use of peak data calls

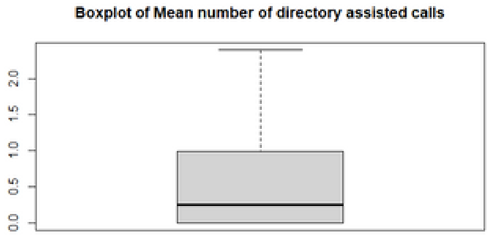


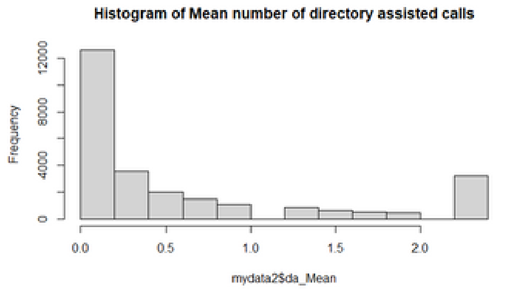


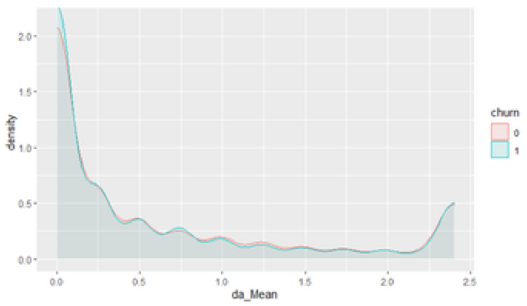


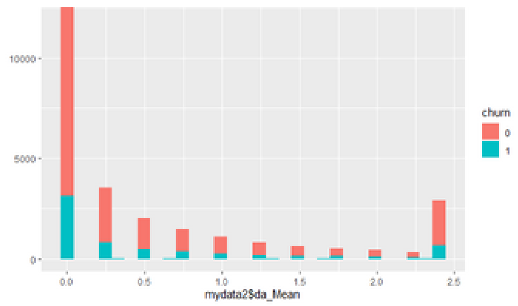
AN41 - Analysis of mean number of directory assisted calls



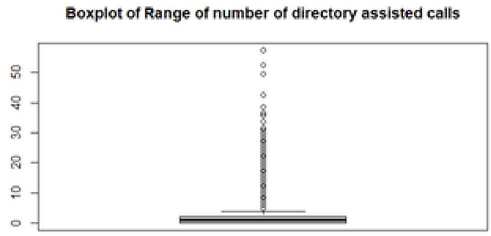


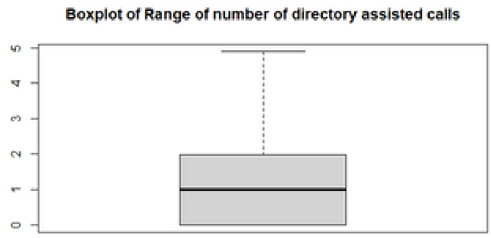


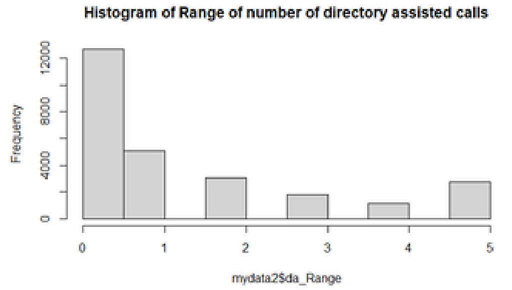


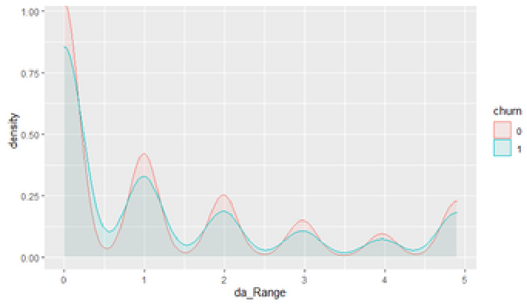


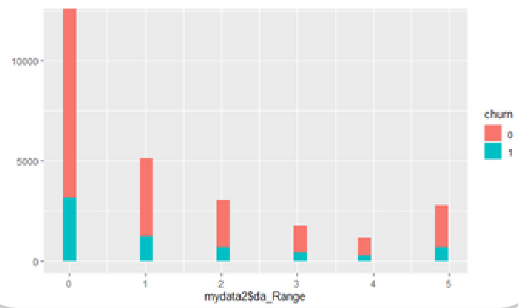
AN42 - Anlaysis of number of directory assisted calls



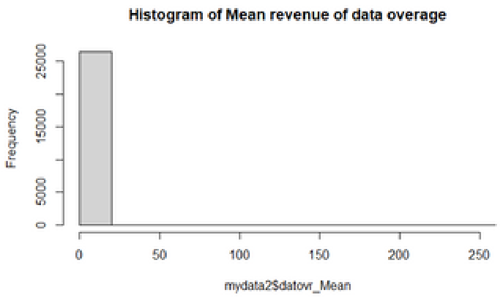


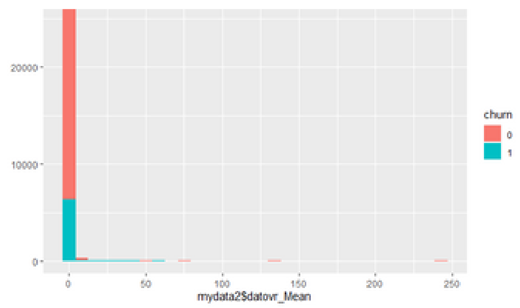


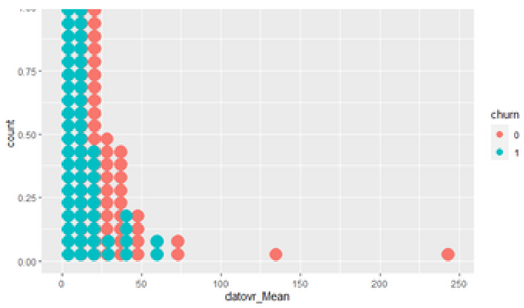




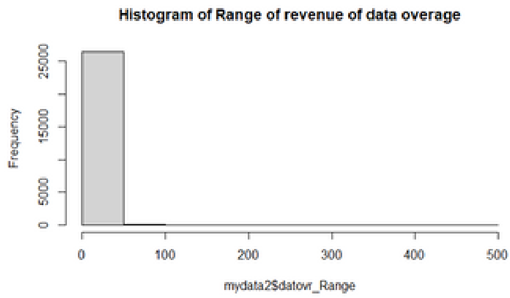
AN43 - Analysis of mean revenue of data overage

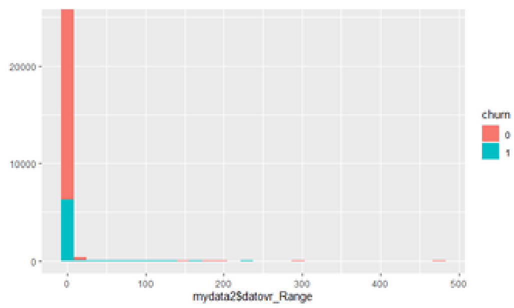


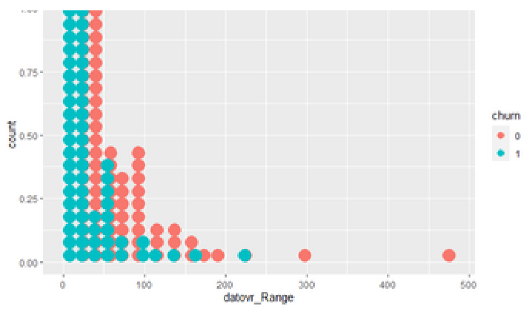




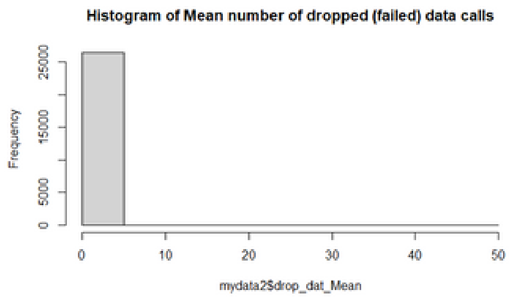
AN44 - Anlaysis of range of revenue of data overage

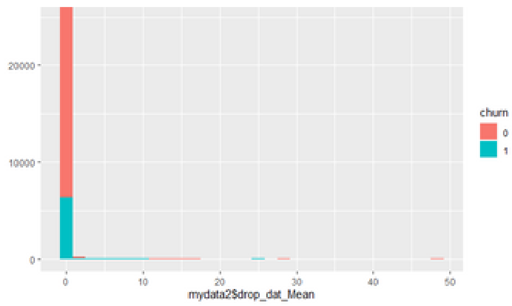


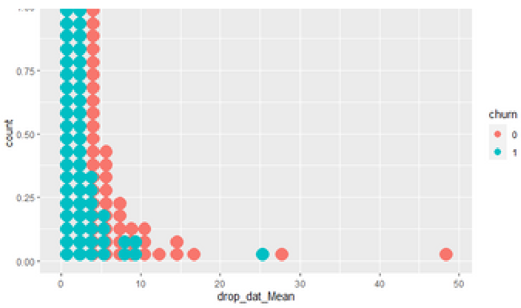




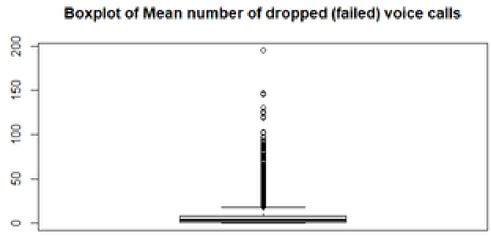
AN45 - Analysis of dropped (failed) data calls

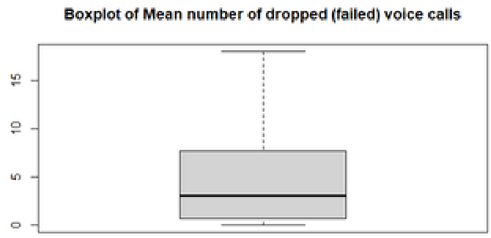


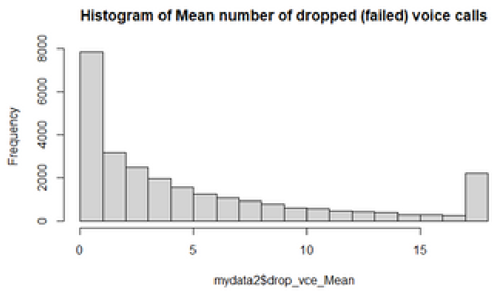


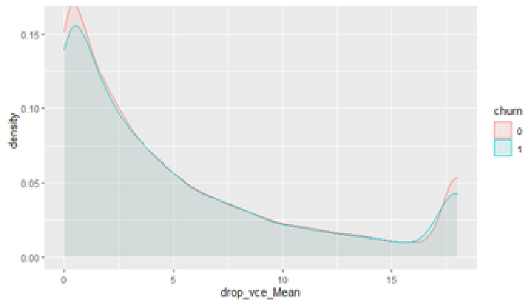


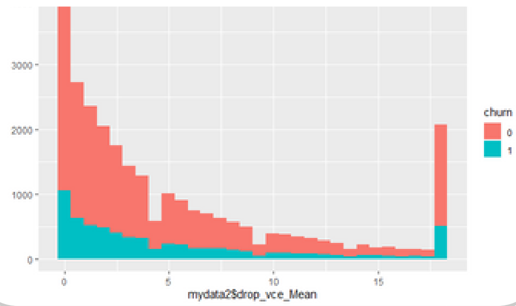
AN46 - Anlaysis of mean number of dropped (failed) voice calls



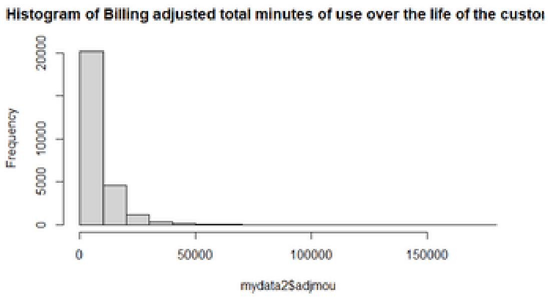


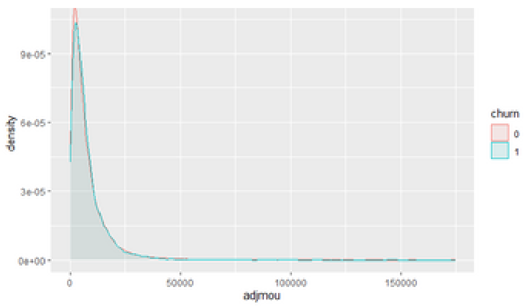


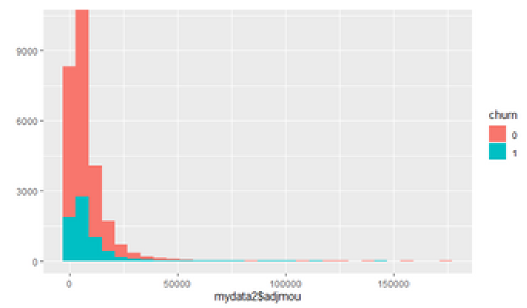


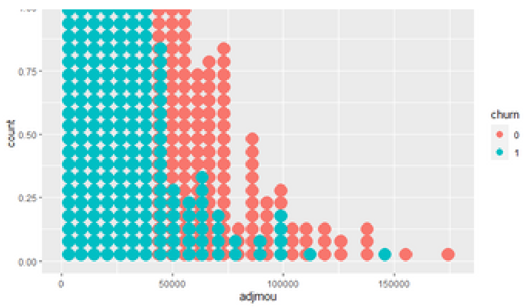


AN47 - Analysis of billing adjusted total minutes of use over life

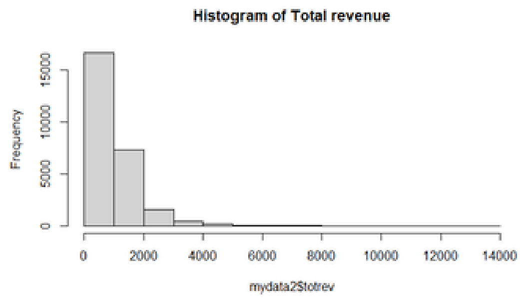


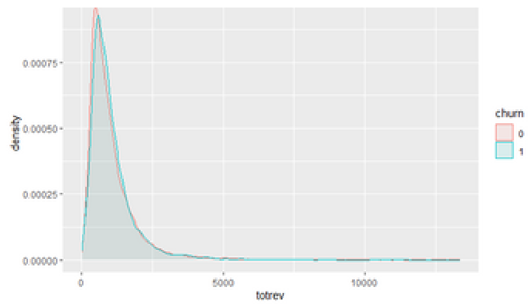


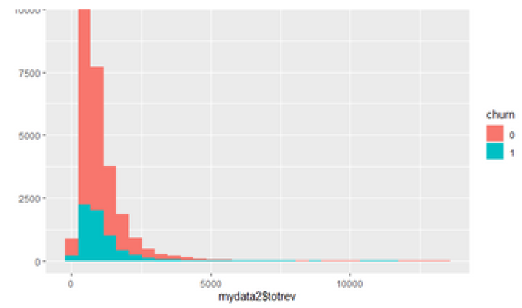


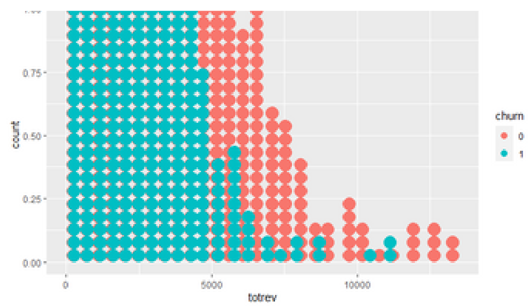


AN48 - Anlaysis of Total Revenue

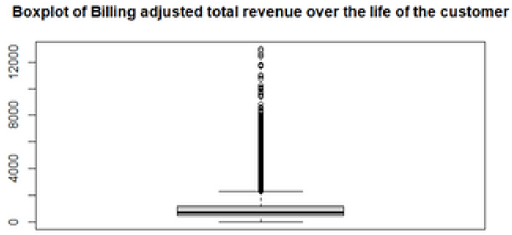


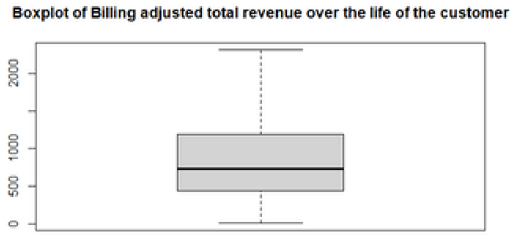


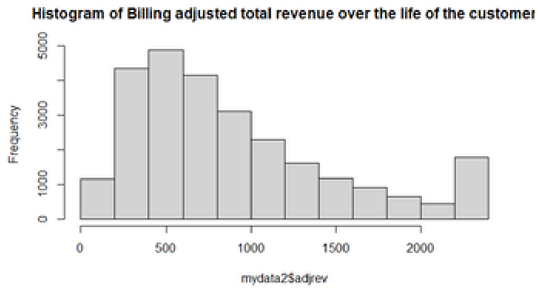


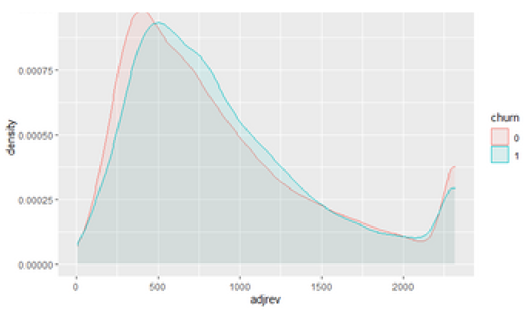


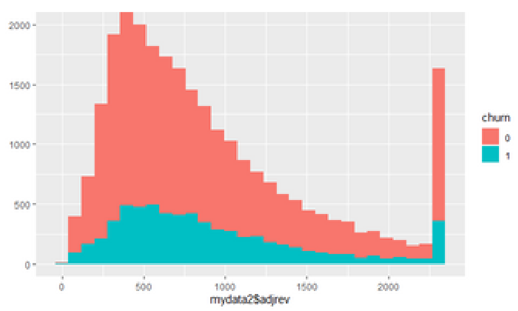
AN49 - Analysis of adjusted total revenue over life



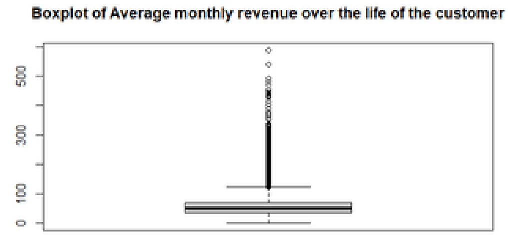


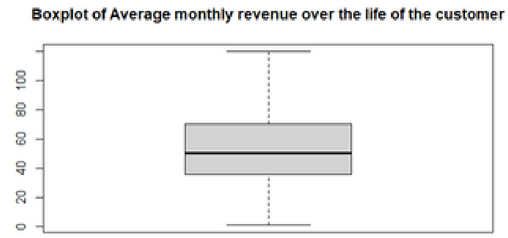


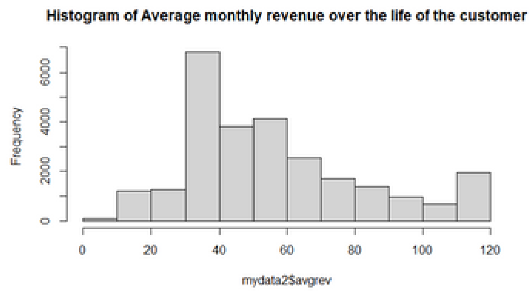


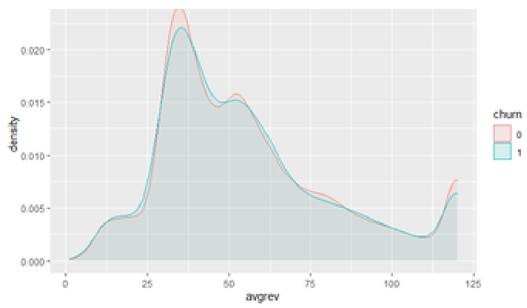


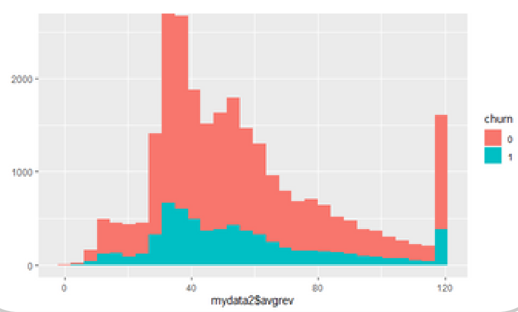
AN50 - Anlaysis of Average monthly revenue over the life



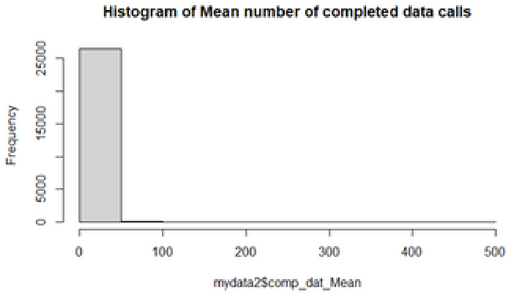


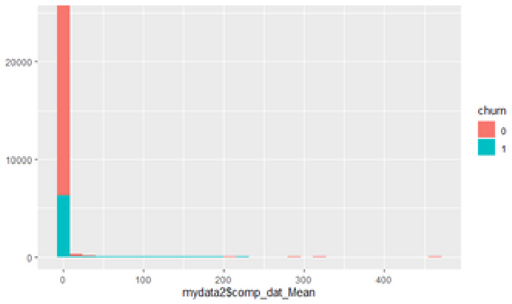


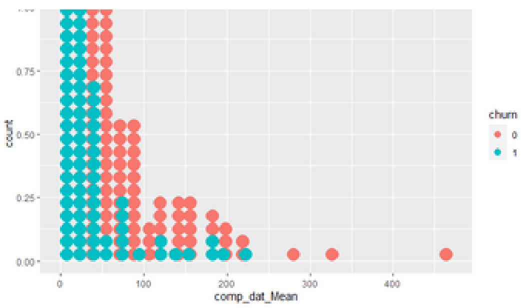




AN51 - Analysis of mean number of completed data calls







AN52 - Anlaysis of mean number of attempted data calls placed