Description

Telecom Customer Churn Prediction

Customer Churn is a burning problem for Telecom companies. In this report, we simulate one such case of customer churn where we work on a data of post-paid customers with a contract. The data has information about the customer usage behaviour, contract details and the payment details. The data also indicates which were the customers who cancelled their service. Based on this past data, we shall build a model which can predict whether a customer will cancel their service in the future or not.

DataSet

Variables	Description	Type
Churn	1 if customer cancelled service, 0 if not	Categorical
AccountWeeks	number of weeks customer has had active account	Continuous
ContractRenewal	1 if customer recently renewed contract, 0 if not	Categorical
DataPlan	1 if customer has data plan, 0 if not	Categorical
DataUsage	gigabytes of monthly data usage	Continuous
CustServCalls	number of calls into customer service	Continuous
DayMins	average daytime minutes per month	Continuous
DayCalls	average number of daytime calls	Continuous
MonthlyCharge	average monthly bill	Continuous
OverageFee	largest overage fee in last 12 months	Continuous
RoamMins	average number of roaming minutes	Continuous

The dataset has 3333 observations with 11 variables. Churn is considered as the Dependent variable and all other attributes as Independent variables.

Assumption

- The data has one dependent variable and other response variables

Hypothesis Formulation

The assignment aim is to identify the predictor variables which are significant for customer Churn.

Null Hypothesis (Ho): No predictor can predict Churn

Alternate Hypothesis (Ha): At least one of the predictors can predict customer Churn

Importing libraries

library(grid)

library(gridExtra)

library(lattice)

library(ModelMetrics)

library(corrplot)

library(ineq)

library(ROCR)

library(caret)

library(tidyverse)

library(readxl)

library(dplyr)

library(rpart)

library(ggplot2)

library(rpart.plot)

library(dplyr)

library(ggplot2)

library(VIF)

library(lmtest)

library(car)

library(e1071)

library(class)

library(MASS)

Analysis of Dataset

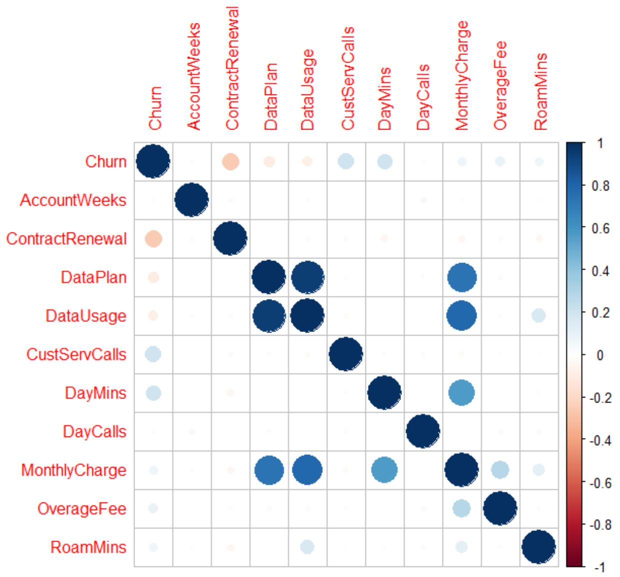
1. Check for missing values

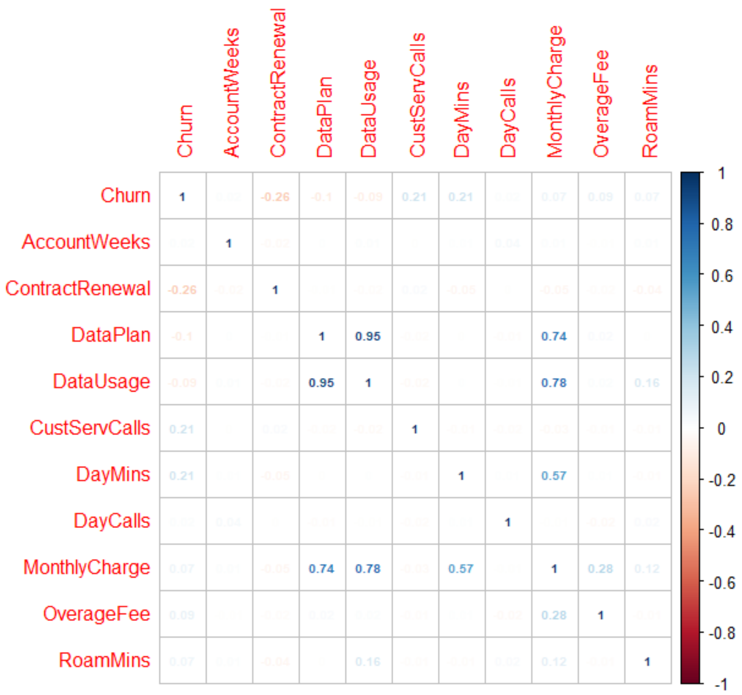
Here we do not have null/ missing values

2. Balance of the target variable

14.5% customers have churned and 85.5% has not churned

3. Correlation among all variables





Data Usage and Data Plan are highly corelated. Monthly Charge is also highly correlated with Data Usage, Data Plan and Day Mins. Churn does not seem to be highly corelated with any of the variables. Churn has maximum correlation with Contract Renewal, Customer Service Calls and Day Mins. Contract Renewal, Data Plan and Data usage is negatively correlated to Churn

4. **Convert binary variables into factors**

Here 3 variables - Churn, ContractRenewal and DataPlan are converted into factors

5. **Summary of the dataset**

Churn	**AccountWeeks**	**ContractRenewal**	**DataPlan**	**DataUsage**	**CustServCalls**
0:2850	Min. : 1.0	0: 323	0:2411	Min. :0.0000	Min. :0.000
1: 483	1st Qu.: 74.0	1:3010	1: 922	1st Qu.:0.0000	1st Qu.:1.000
	Median :101.0			Median :0.0000	Median :1.000
	Mean :101.1			Mean :0.8165	Mean :1.563
	3rd Qu.:127.0			3rd Qu.:1.7800	3rd Qu.:2.000
	Max. :243.0			Max. :5.4000	Max. :9.000

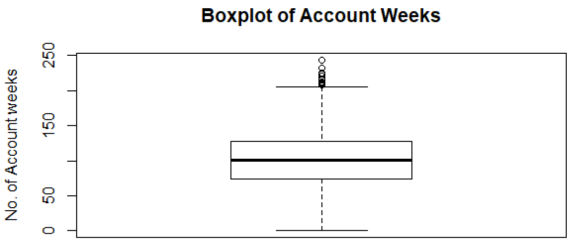
DayMins	**DayCalls**	**MonthlyCharge**	**OverageFee**	**RoamMins**
Min. : 0.0	Min. : 0.0	Min. : 14.00	Min. : 0.00	Min. : 0.00
1st Qu.:143.7	1st Qu.: 87.0	1st Qu.: 45.00	1st Qu.: 8.33	1st Qu.: 8.50
Median :179.4	Median :101.0	Median : 53.50	Median :10.07	Median :10.30
Mean :179.8	Mean :100.4	Mean : 56.31	Mean :10.05	Mean :10.24
3rd Qu.:216.4	3rd Qu.:114.0	3rd Qu.: 66.20	3rd Qu.:11.77	3rd Qu.:12.10
Max. :350.8	Max. :165.0	Max. :111.30	Max. :18.19	Max. :20.00

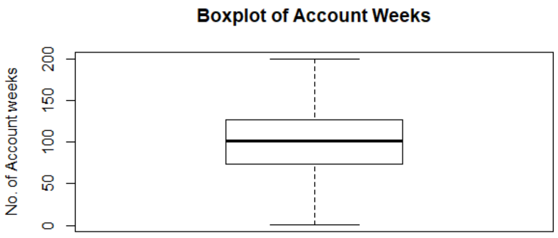
Exploratory Data Analysis

1.Continuous Variables

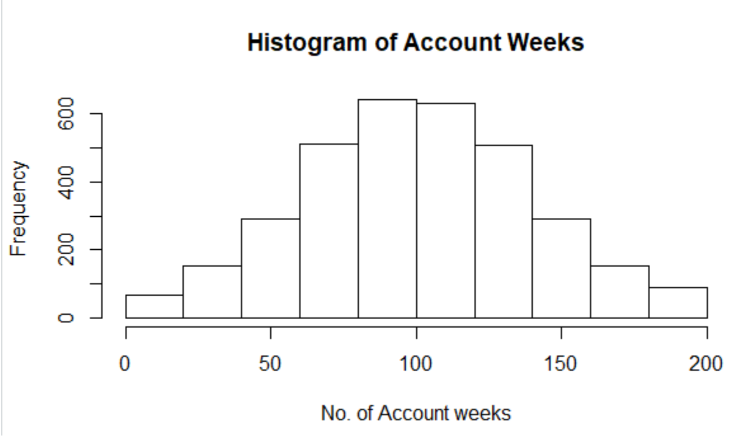
- ### AccountWeeks

Boxplot

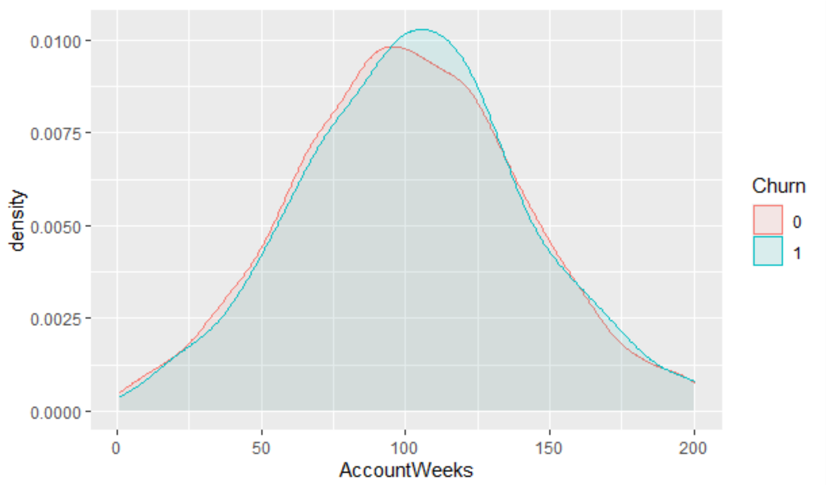




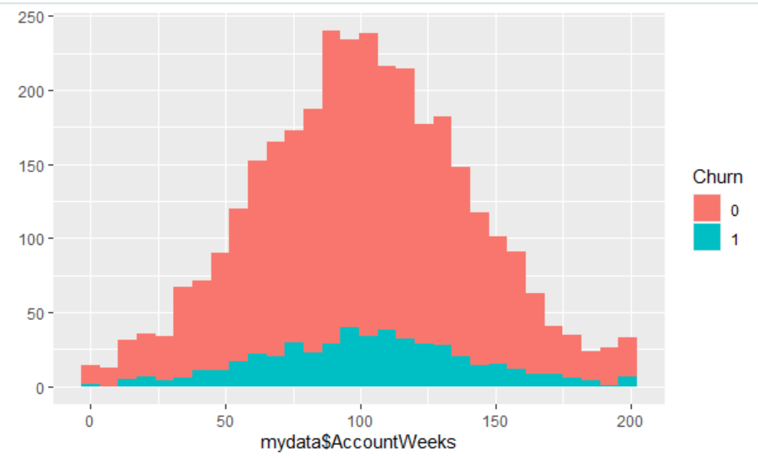
Histogram



Density Plot

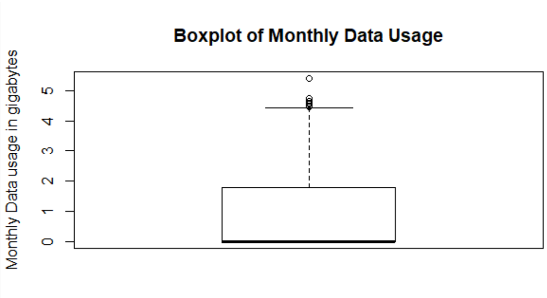


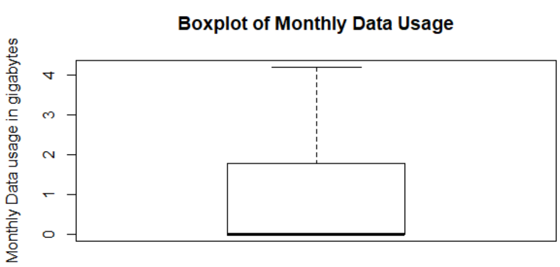
Qplot

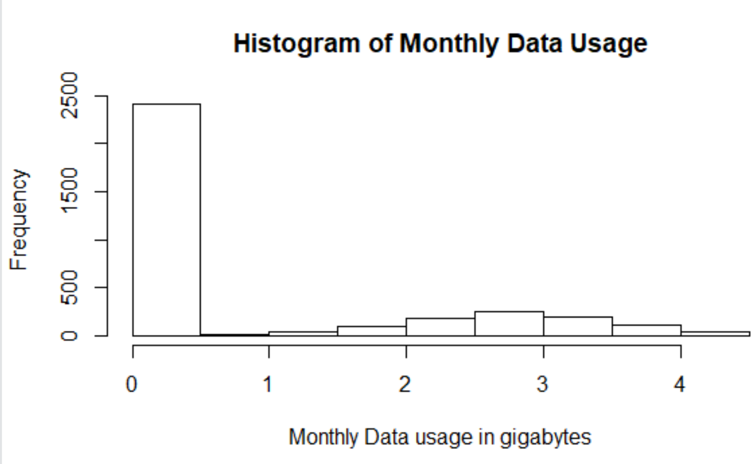


- ## Data Usage

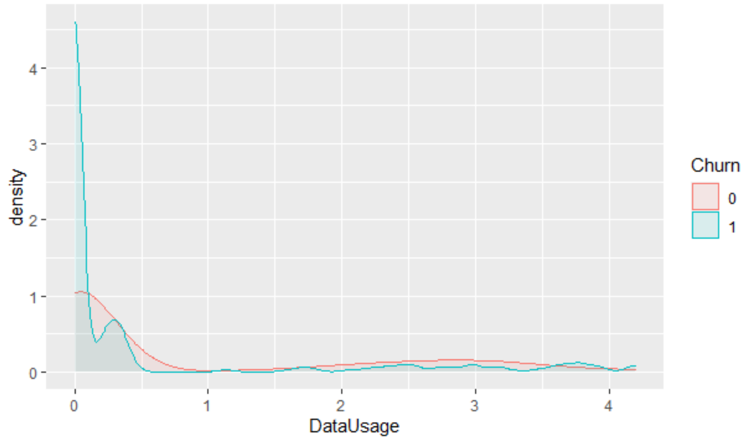
Boxplot



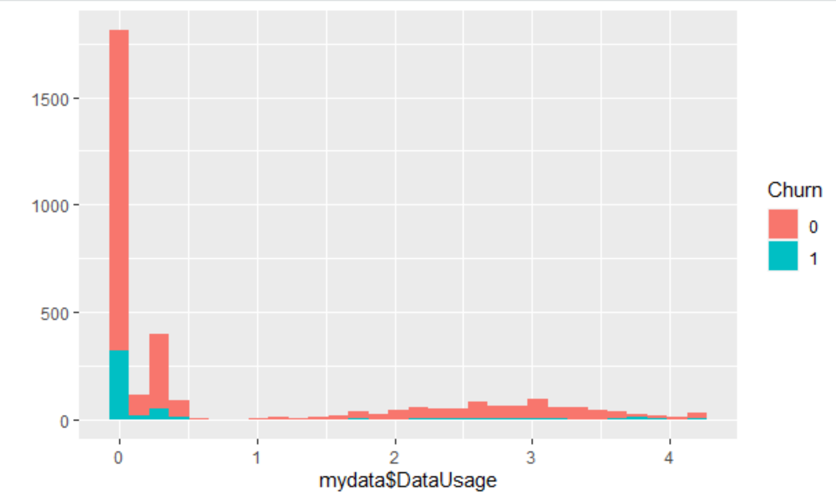




Density Plot

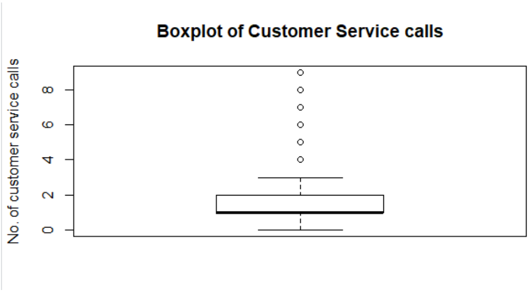


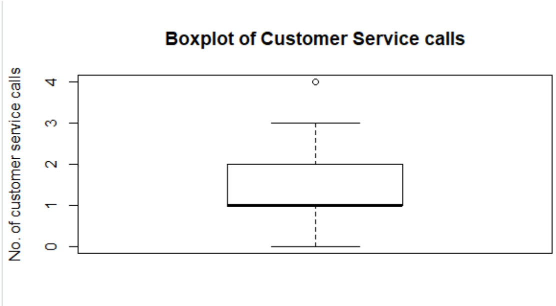
Qplot

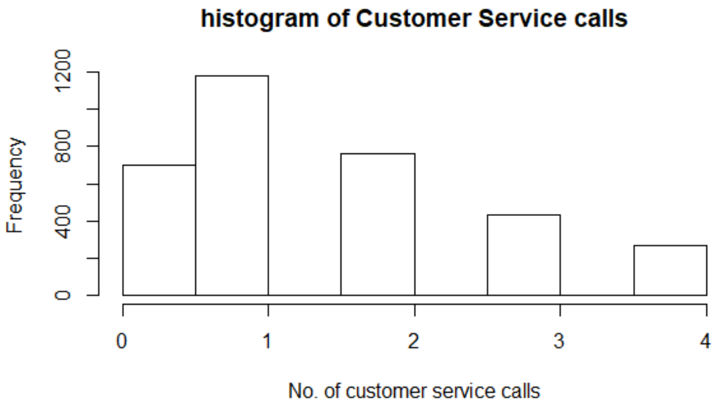


- **CustServCalls**

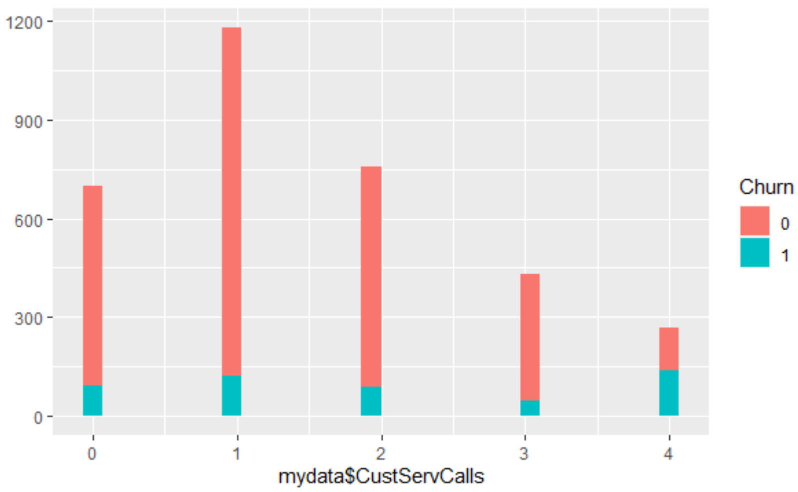
Boxplot



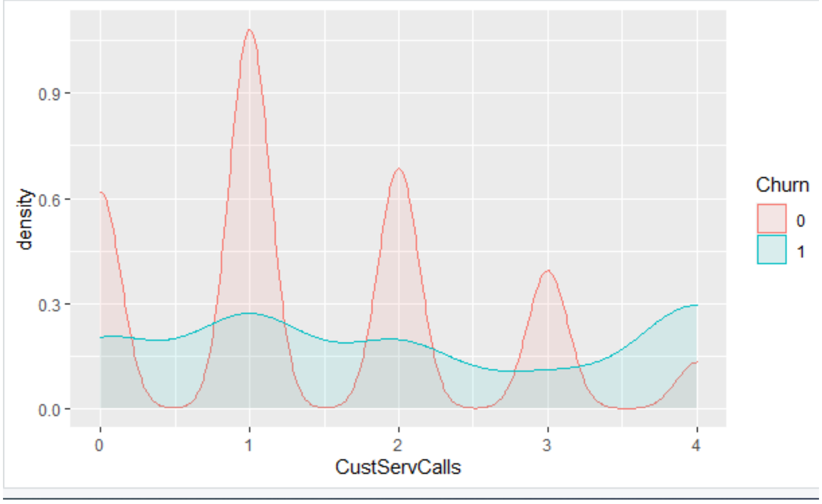




Qplot



Density Plot



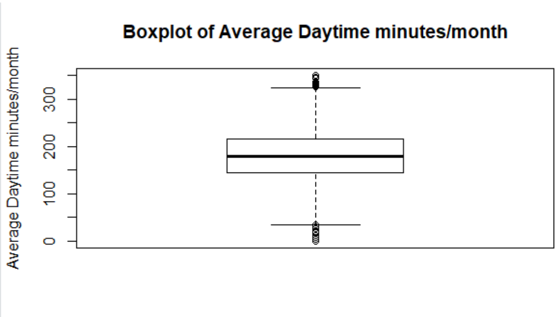
Though this is a numeric variable, from the above figure, we understand that it has specific levels. So, we will explore its influence on Churn using proportion table

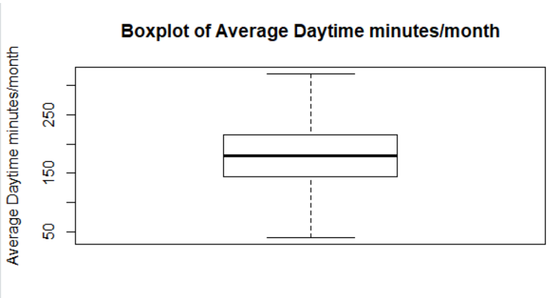
prop.table(table(mydata$CustServCalls,mydata$Churn),1)*100

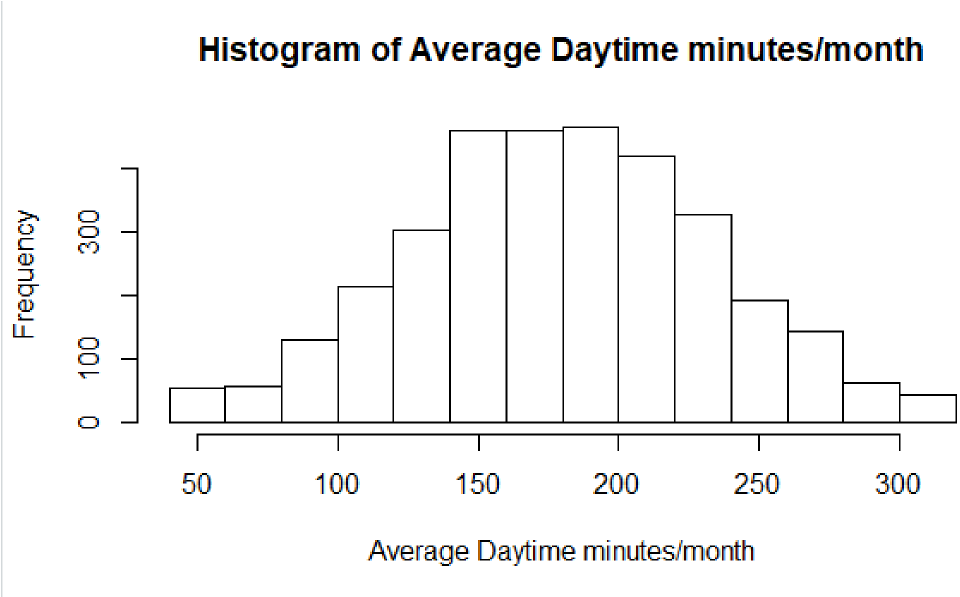
	0	1
0	86.80057	13.19943
1	89.66977	10.33023
2	88.53755	11.46245
3	89.74359	10.25641
4	48.31461	51.68539

- ## DayMins

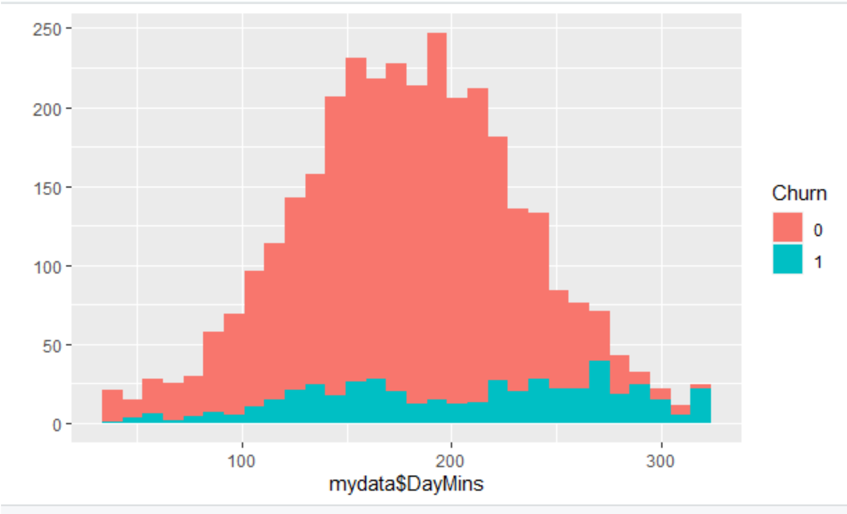
Boxplot



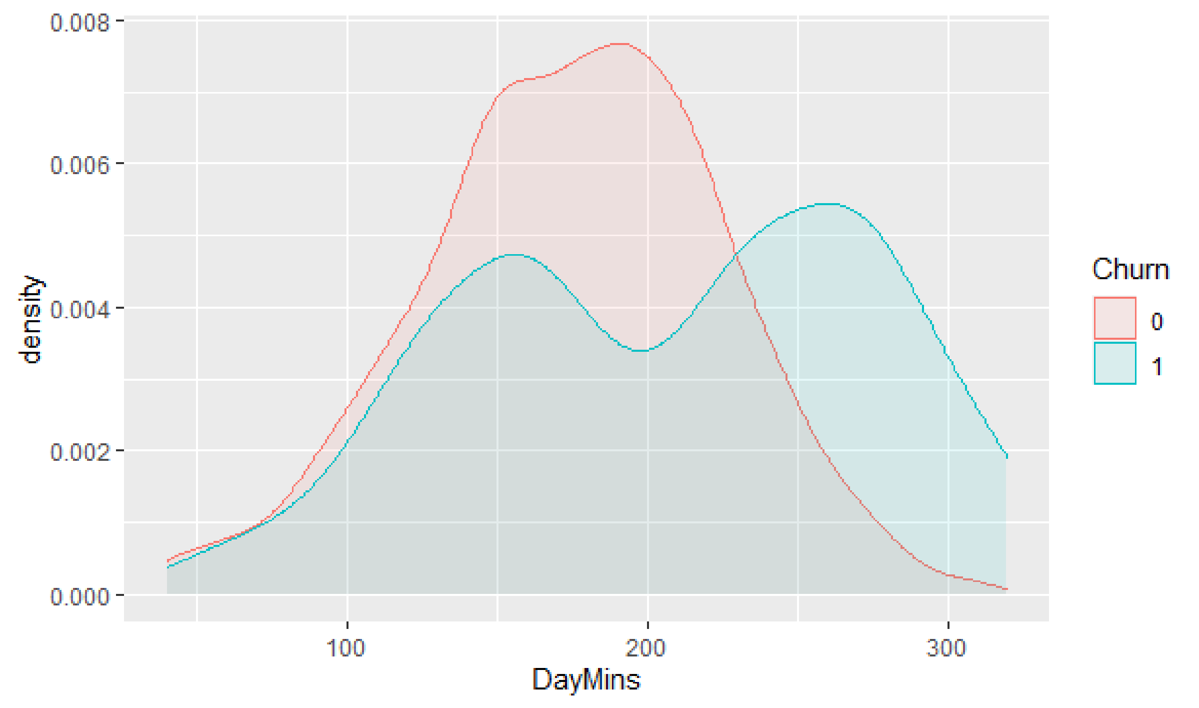



Histogram of Average Daytime minutes/month
Frequency
0
100
300
50
100
150
200
250
300
Average Daytime minutes/month

Qplot

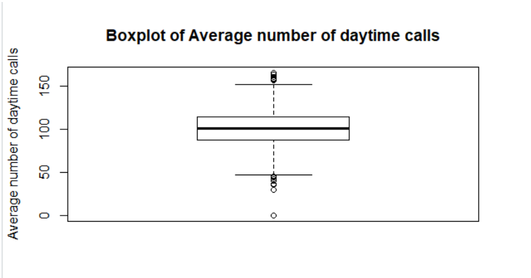

250
200
150
100
50
0
100
200
300
mydata$DayMins
Churn
0
1

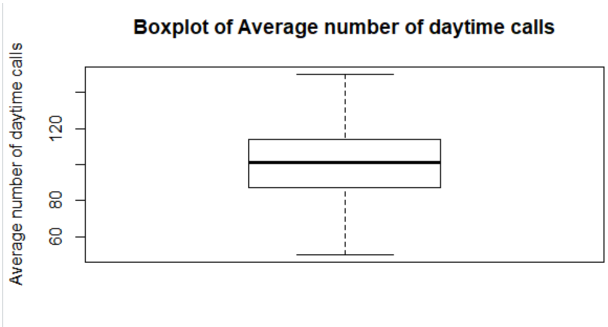
Density Plot

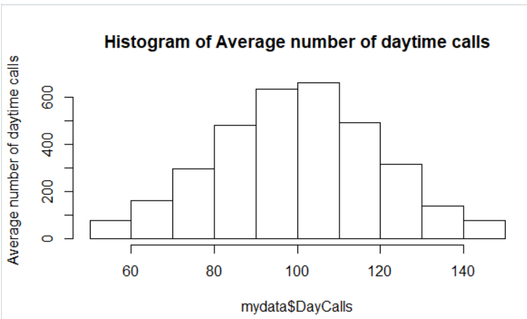


- **DayCalls**

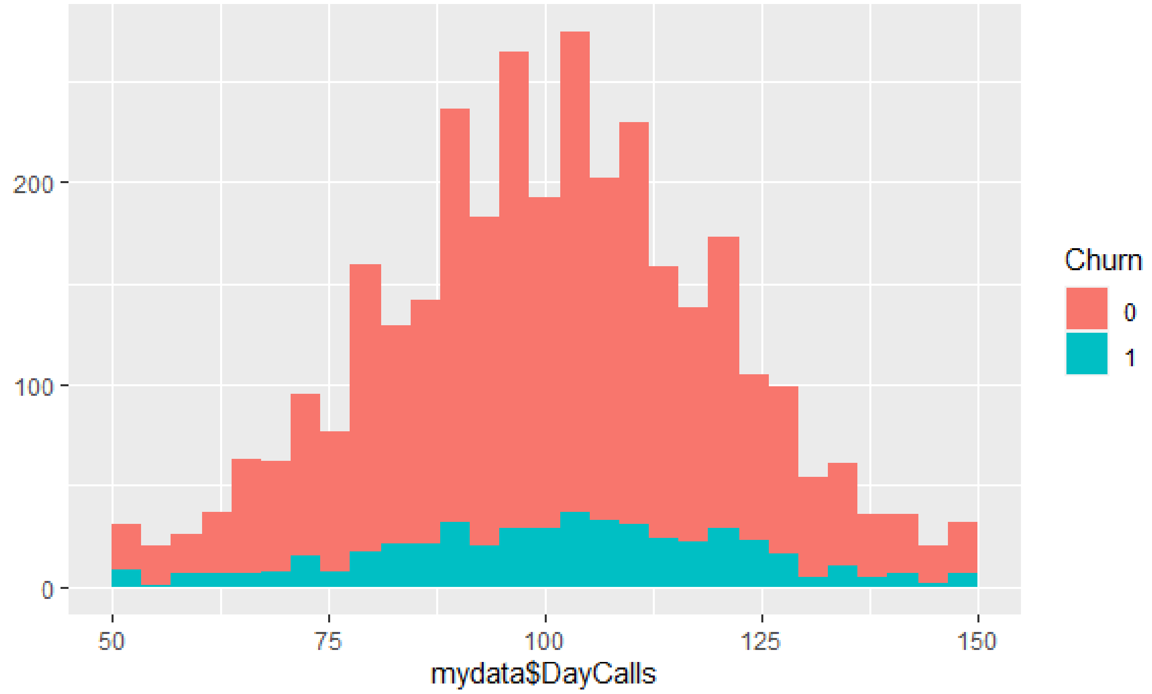
Boxplot



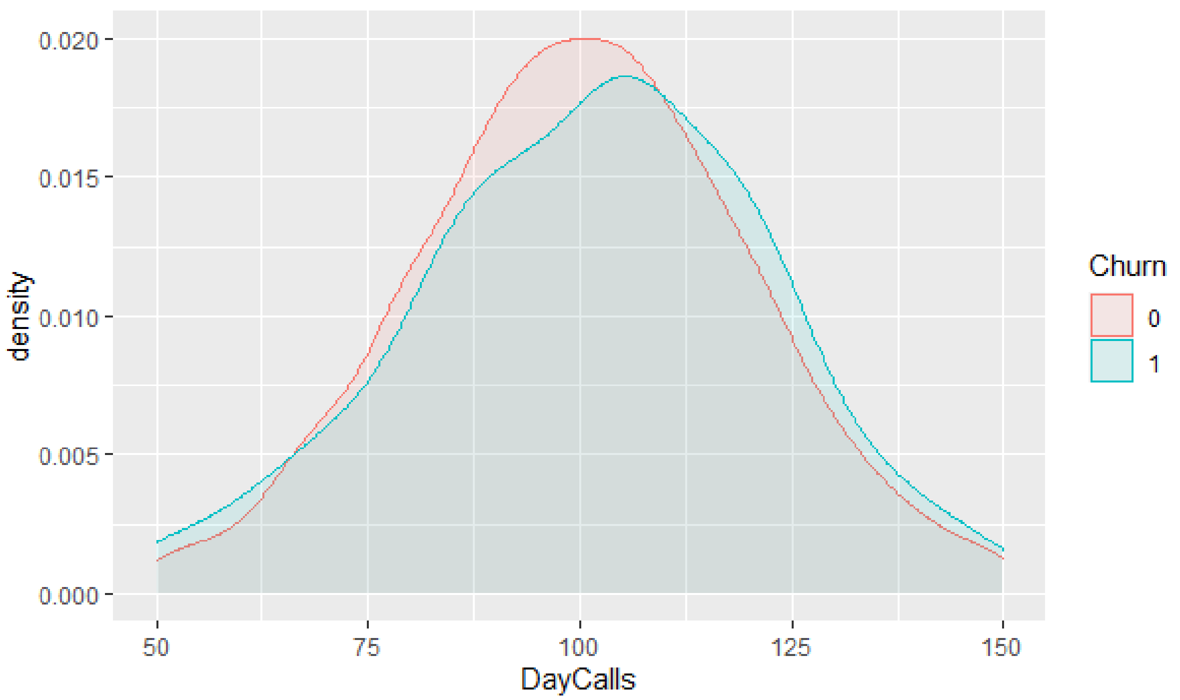




Qplot

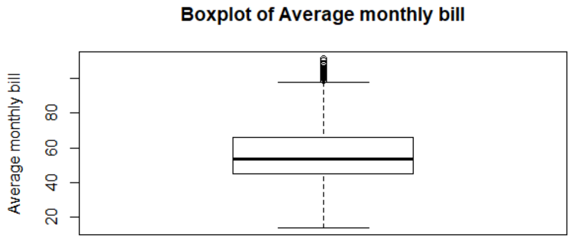


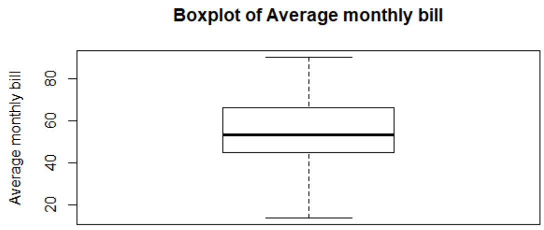
Density Plot

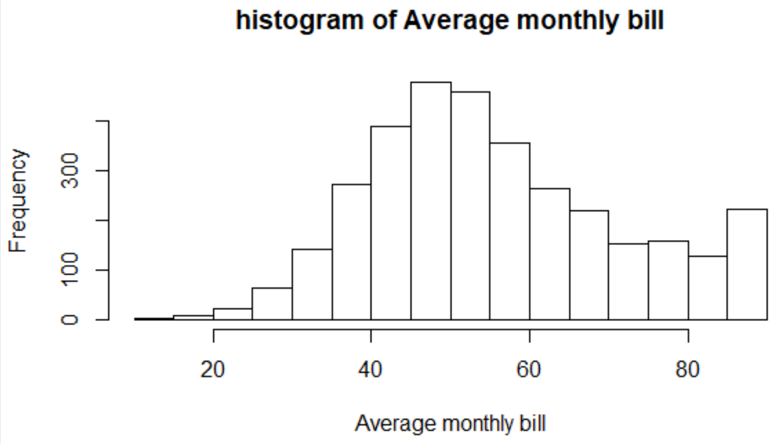


- **MonthlyCharge**

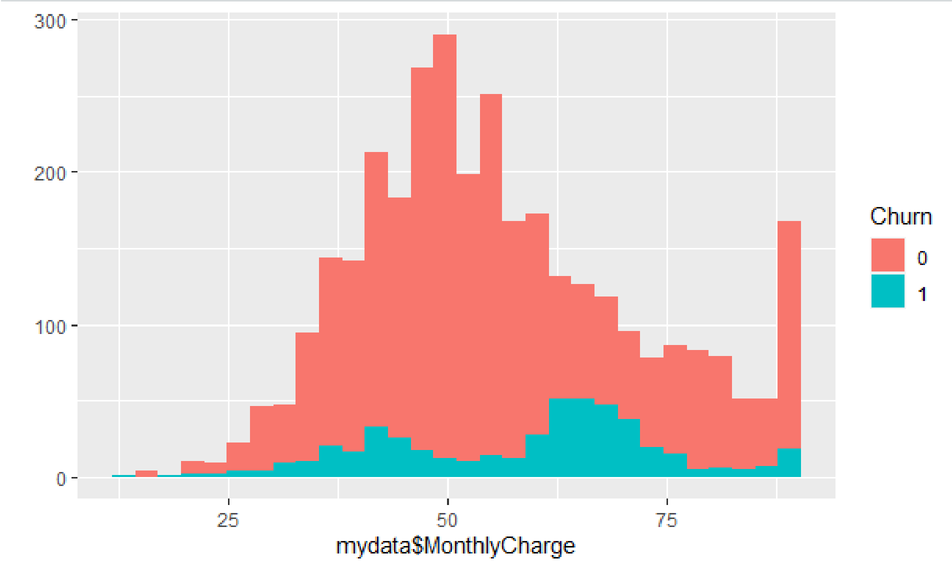
Boxplot



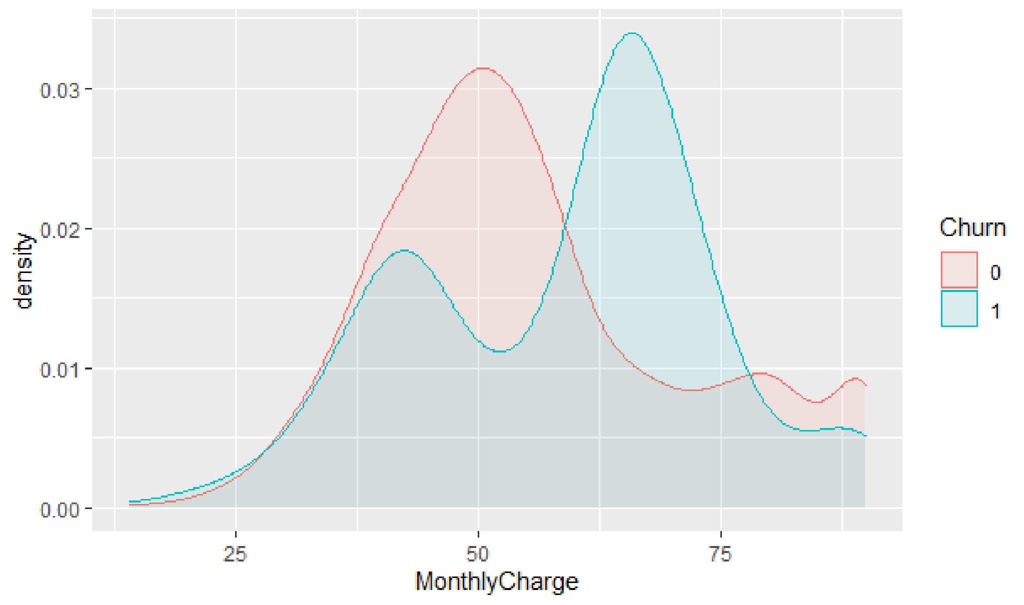




Qplot

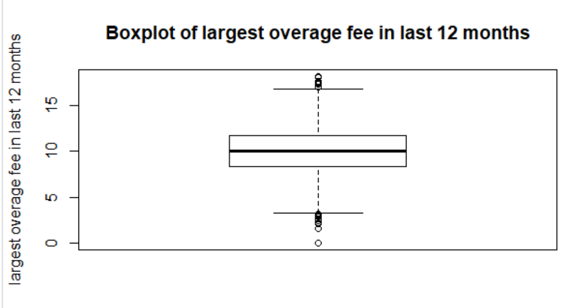


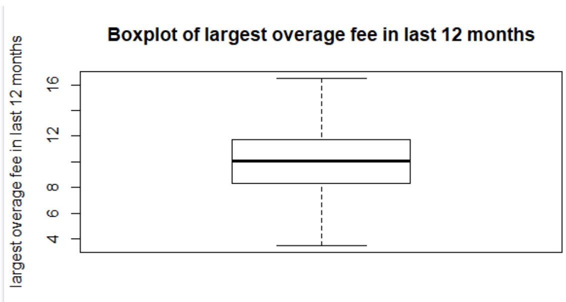
Density Plot

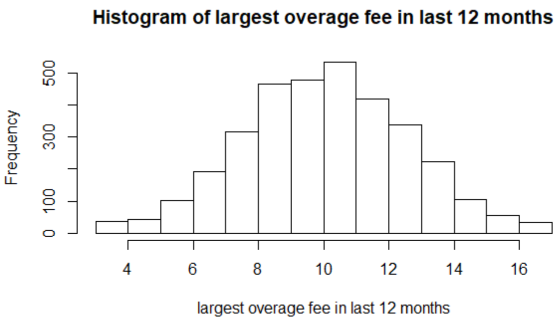


- ## OverageFee

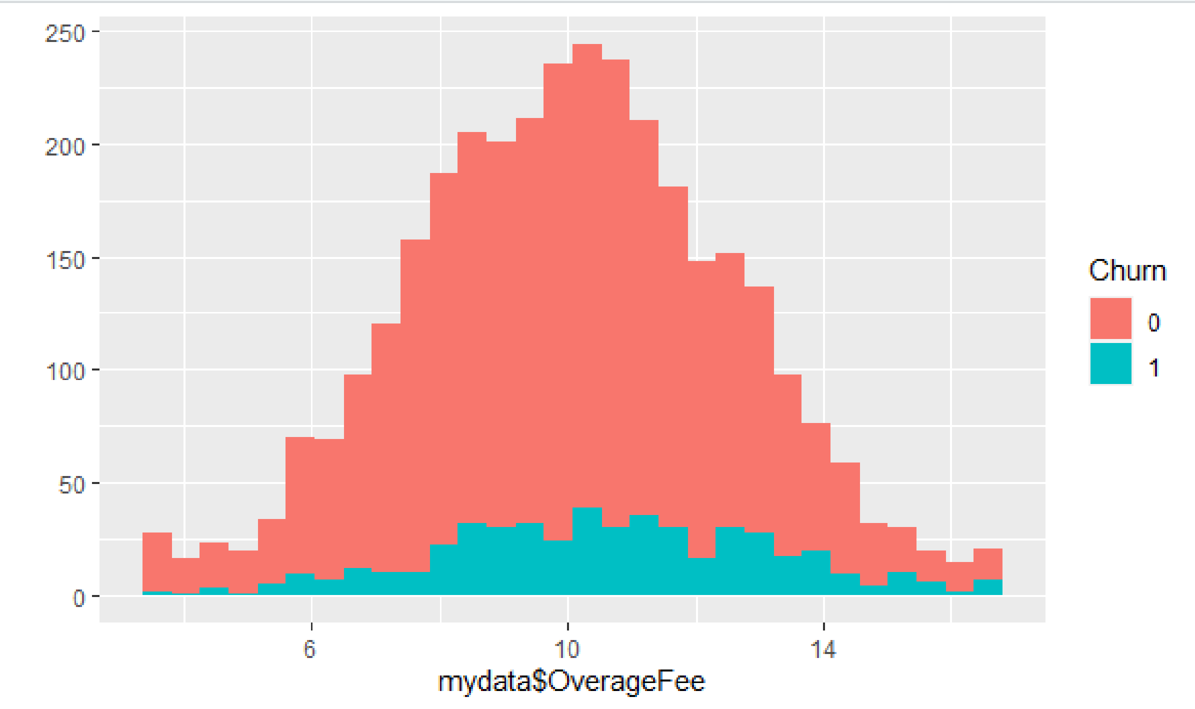
Boxplot



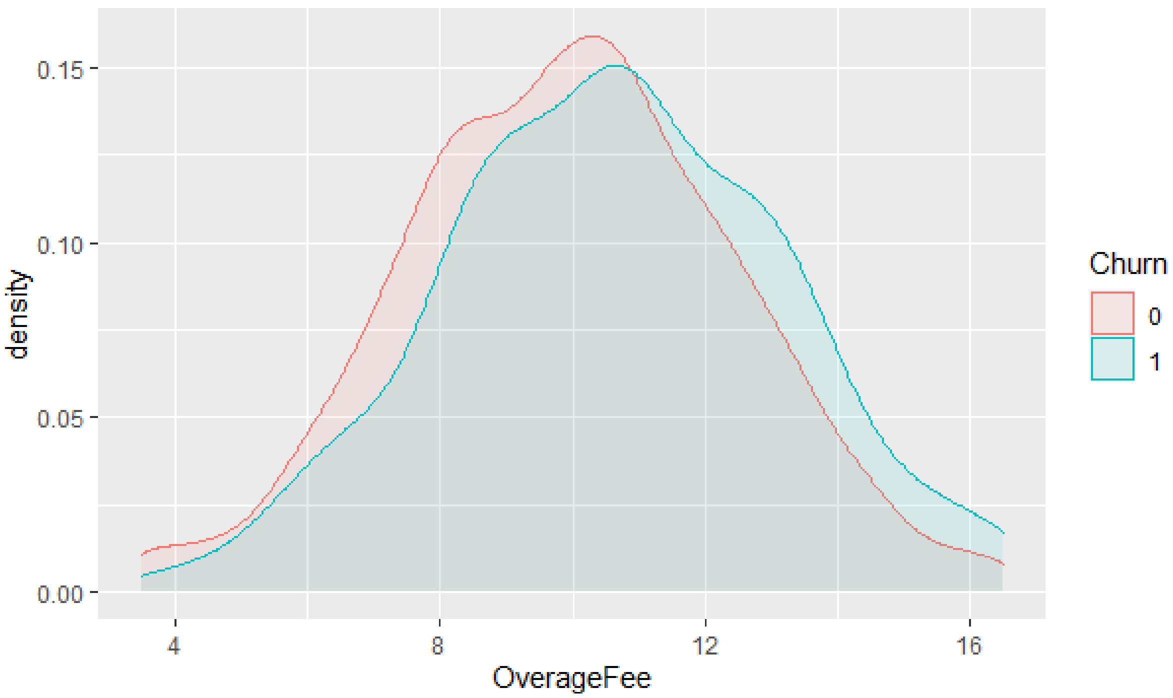




Qplot

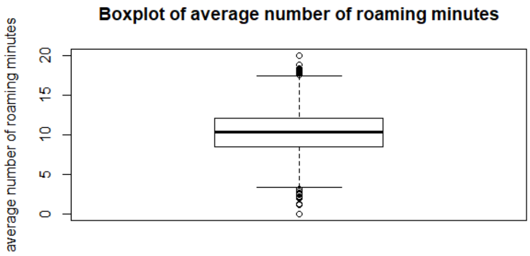


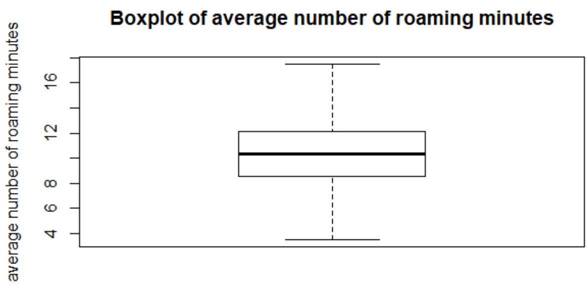
Density Plot

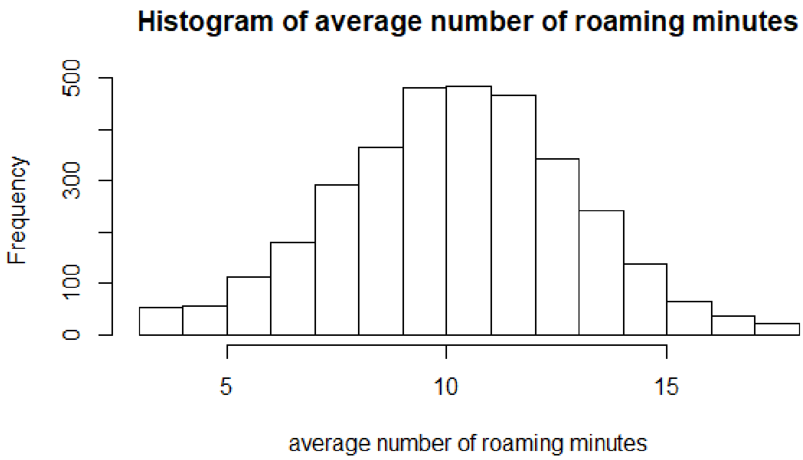


- **RoamMins**

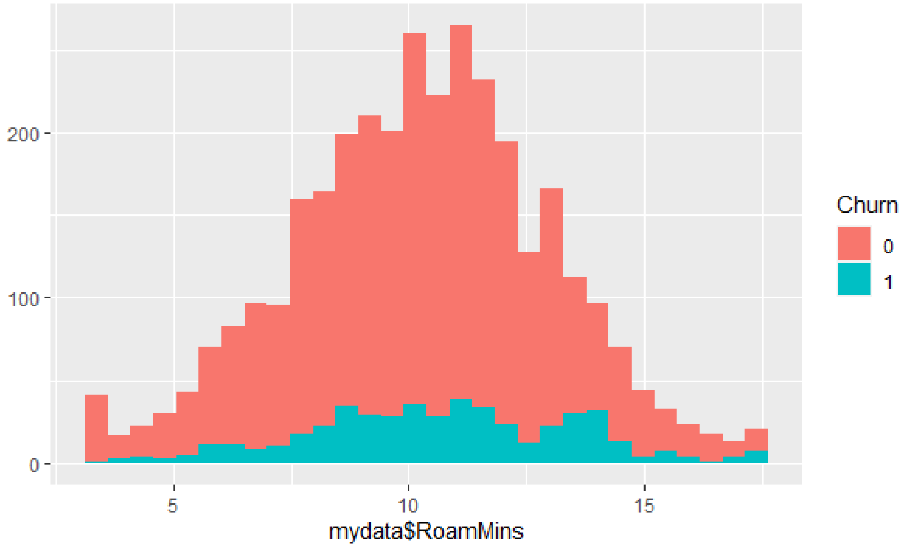
Boxplot



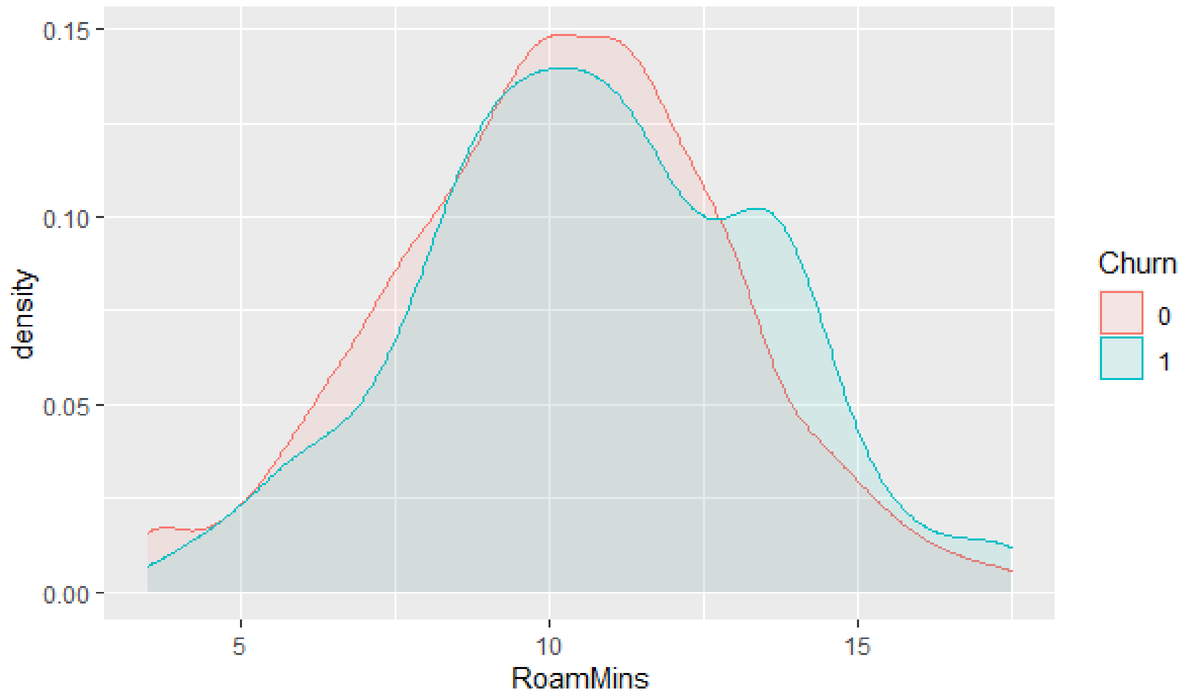




Qplot

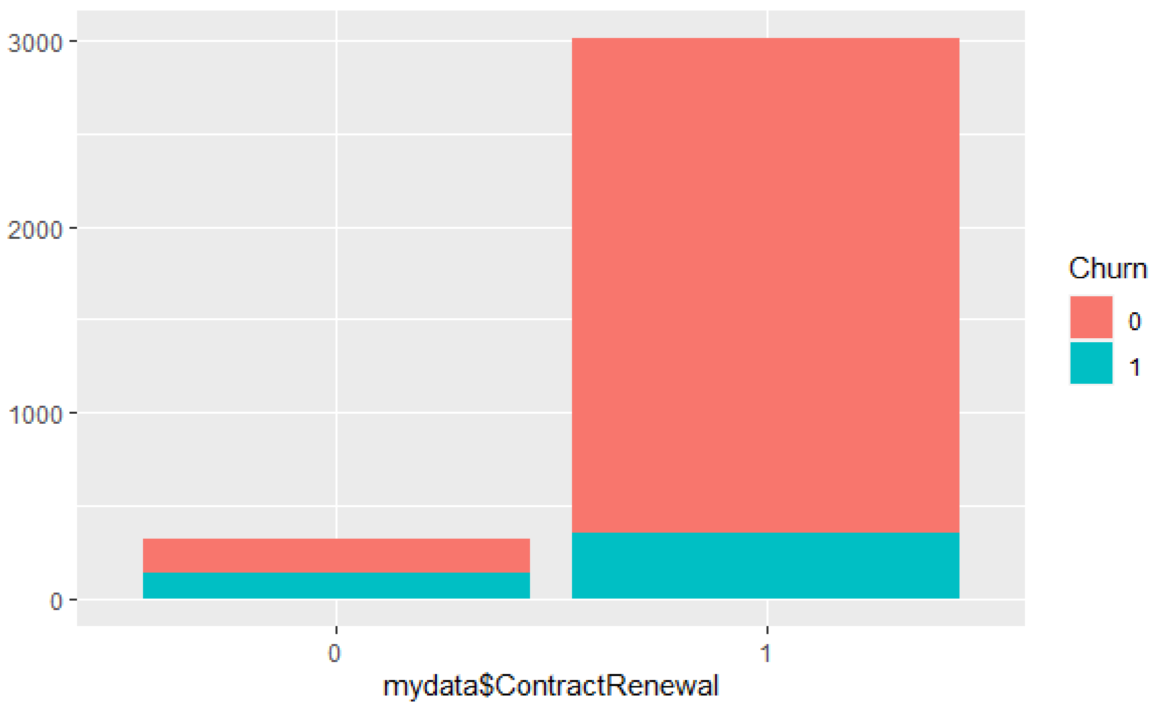


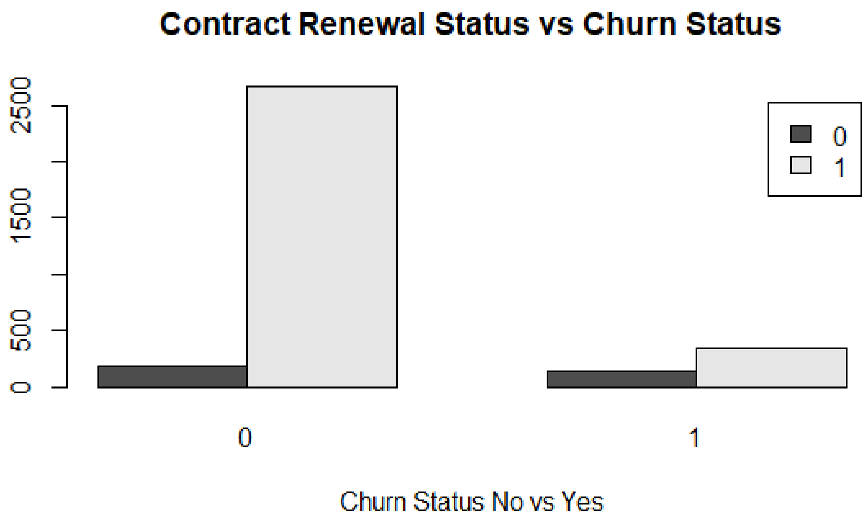
Density Plot



2. Categorical Variables

- **ContractRenewal**





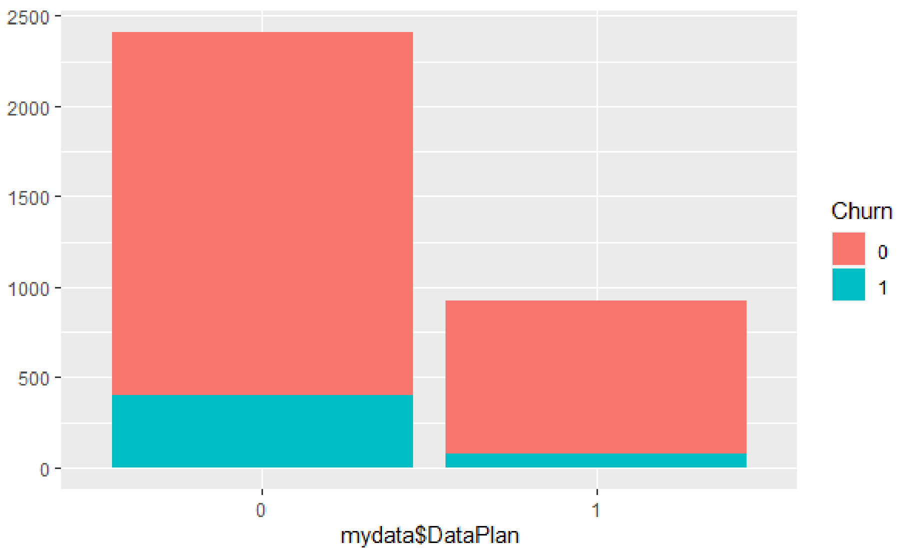
prop.table(table(mydata$ContractRenewal,mydata$Churn),1)*100

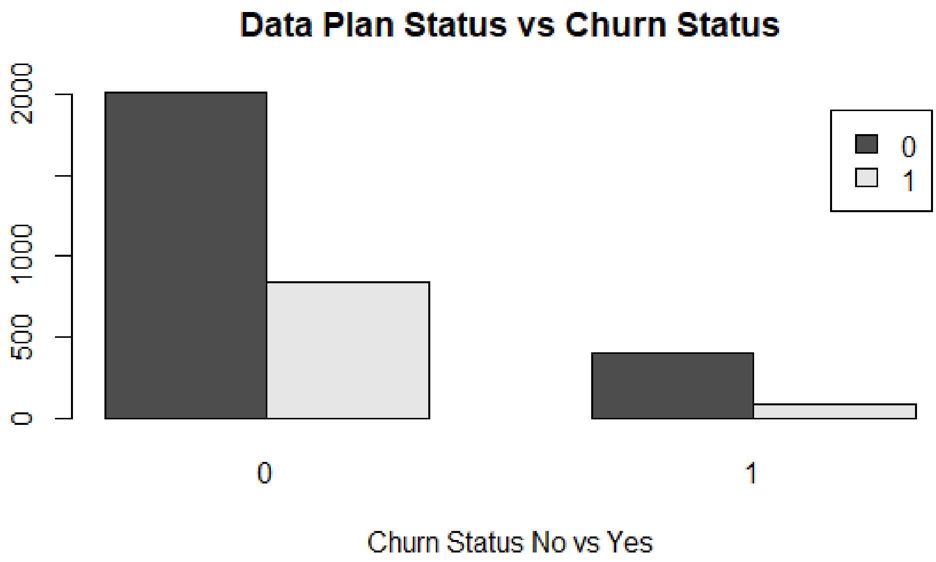
	0	1
0	57.58514	42.41486
1	88.50498	11.49502

table(mydata$Churn, mydata$ContractRenewal)

	0	1
0	186	2664
1	137	346

- ## DataPlan





prop.table(table(mydata$DataPlan,mydata$Churn),1)*100

	0	1
0	83.28494	16.71506
1	91.32321	8.67679

table(mydata$Churn, mydata$DataPlan)

	0	1
0	2008	842
1	403	80

The probability of an account churning is higher if the account has not subscribed to a data plan.

- ## Churn

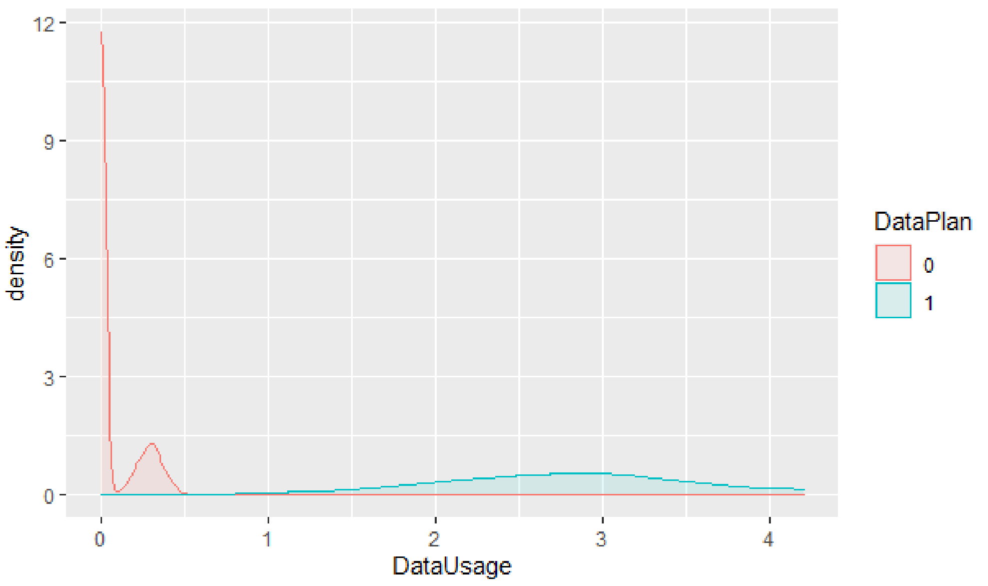
prop.table(table(mydata$Churn))*100

0	1
85.50855	14.49145

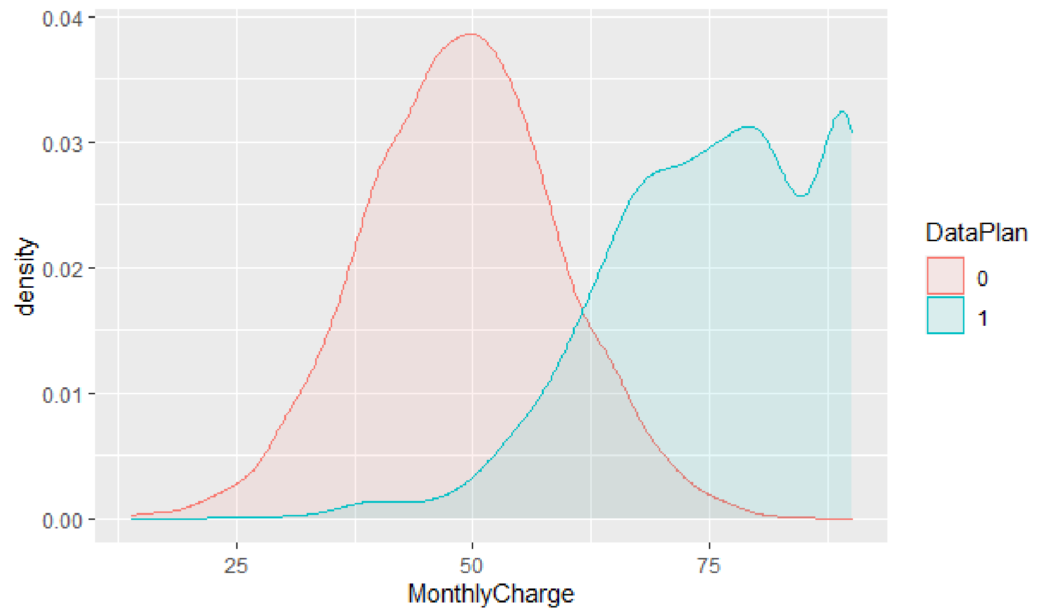
In the given dataset,customers who has canceled service vs not canceled service is 14.49% and 85.51% respectively

2. Relation between variables

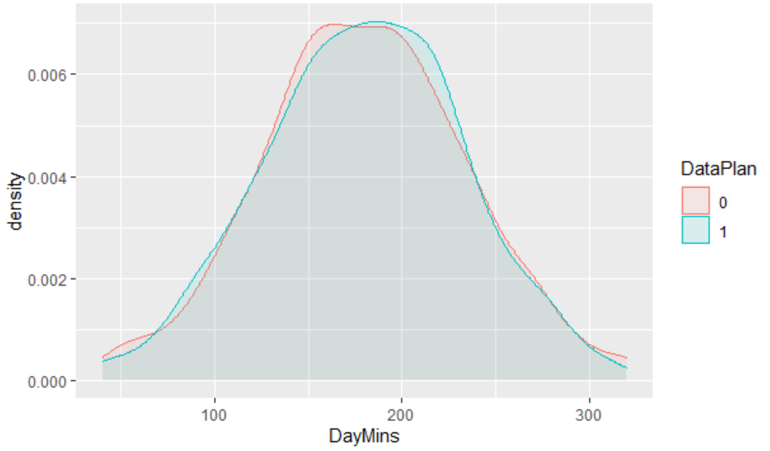
Data Usage & Data Plan



Monthly Charge & Data Plan



Day Mins & Data Plan

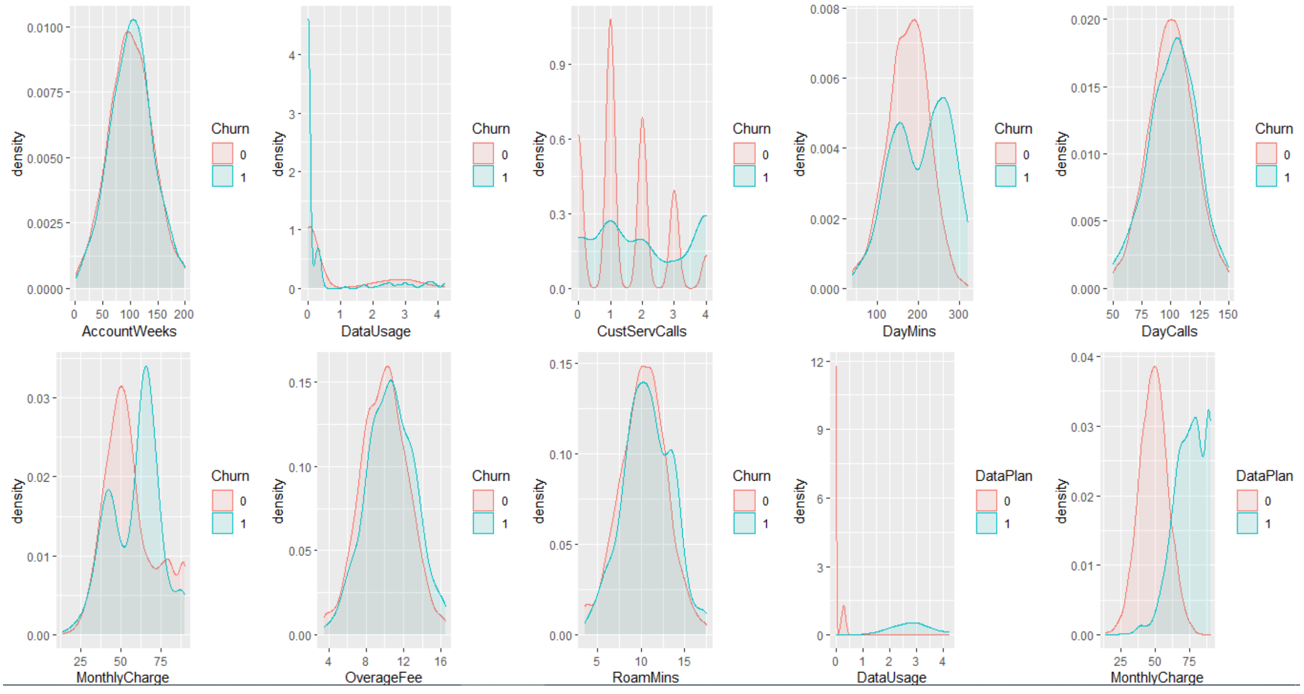


Contract Renewal & Data Plan

table(mydata$ContractRenewal, mydata$DataPlan)

	0	1
0	231	92
1	2180	830

Density plots of all variables:



Data Slicing

Splitting the dataset into train and test dataset

dim(testdata)

[1] 999 11

> dim(traindata)

[1] 2334   11

85.48% of train data has not cancelled and 14.52% has cancelled service

85.59% of train data has not cancelled and 14.41% has cancelled service

table(traindata$Churn)

```
   0    1 
1995  339
```

Logistic Regression

Model 1

Including all the variables:

Call: glm(formula = traindata$Churn ~ ., family = "binomial", data = traindata)

```
Coefficients:
    (Intercept)      AccountWeeks  ContractRenewal1        DataPlan1
     -6.196e+00        -6.692e-05        -1.865e+00       -1.954e+00
      DataUsage      CustServCalls          DayMins         DayCalls
     -1.008e+00         5.302e-01        -1.227e-02        2.771e-03
  MonthlyCharge        OverageFee          RoamMins
      1.472e-01        -1.094e-01         9.546e-02
```

Degrees of Freedom: 2333 Total (i.e. Null); 2323 Residual

Null Deviance: 1934

Residual Deviance: 1569 AIC: 1591

glm(formula = traindata$Churn ~ ., family = "binomial", data = traindata)

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.8694  -0.5253  -0.3666  -0.2191   2.9595
Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)      -6.196e+00  6.774e-01  -9.148  < 2e-16 ***
AccountWeeks     -6.692e-05  1.654e-03  -0.040 0.967737
ContractRenewal1 -1.865e+00  1.715e-01 -10.879  < 2e-16 ***
DataPlan1        -1.954e+00  6.789e-01  -2.879 0.003992 **
DataUsage        -1.008e+00  4.629e-01  -2.177 0.029498 *
CustServCalls     5.302e-01  5.375e-02   9.865  < 2e-16 ***
DayMins          -1.227e-02  7.886e-03  -1.556 0.119610
DayCalls          2.771e-03  3.277e-03   0.846 0.397798
```

```
MonthlyCharge    1.472e-01  4.721e-02   3.117 0.001827 **
OverageFee      -1.094e-01  8.268e-02  -1.323 0.185827
RoamMins         9.546e-02  2.743e-02   3.480 0.000502 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Dispersion parameter for binomial family taken to be 1)
    Null deviance: 1934.3  on 2333  degrees of freedom
Residual deviance: 1568.8  on 2323  degrees of freedom
```

AIC: 1590.8

Number of Fisher Scoring iterations: 5

AccountWeeks, DayMins, DayCalls and OverageFee seems to be less significant

There could be multicollinearity

<u>Odds Ratio of all variables</u>

```
exp(coef(logmodel)) #Odds ratio
    (Intercept)     AccountWeeks ContractRenewal1       DataPlan1
    0.002037072      0.999933086      0.154853412      0.141643006
      DataUsage    CustServCalls          DayMins        DayCalls
    0.365063605      1.699330017      0.987800899      1.002775132
  MonthlyCharge       OverageFee         RoamMins
    1.158536641      0.896385904      1.100165887
```

<u>Probability</u>

```
exp(coef(logmodel))/(1+exp(coef(logmodel))) #Probability
    (Intercept)     AccountWeeks ContractRenewal1       DataPlan1
    0.002032931      0.499983271      0.134089236      0.124069438
      DataUsage    CustServCalls          DayMins        DayCalls
    0.267433403      0.629537702      0.496931508      0.500692822
  MonthlyCharge       OverageFee         RoamMins
    0.536723176      0.472681168      0.523847137
```

We shall create a null model for comparison with the created model. A null model does not have independent variable coefficients

Call: glm(formula = traindata$Churn ~ 1, family = "binomial", data = traindata)

Coefficients:

(Intercept)

    -1.772

Degrees of Freedom: 2333 Total (i.e. Null); 2333 Residual

Null Deviance: 1934

Residual Deviance: 1934   AIC: 1936

Call:

glm(formula = traindata$Churn ~ 1, family = "binomial", data = traindata)

Deviance Residuals:

```
    Min       1Q   Median       3Q      Max  
-0.5603  -0.5603  -0.5603  -0.5603   1.9644  
```

Coefficients:

```
            Estimate Std. Error z value Pr(>|z|)    
(Intercept) -1.77240    0.05875  -30.17   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

(Dispersion parameter for binomial family taken to be 1)

```
    Null deviance: 1934.3  on 2333  degrees of freedom
Residual deviance: 1934.3  on 2333  degrees of freedom
AIC: 1936.3
```

Number of Fisher Scoring iterations: 4

Likelihood ratio test

Model 1: traindata$Churn ~ AccountWeeks + ContractRenewal + DataPlan +

DataUsage + CustServCalls + DayMins + DayCalls + MonthlyCharge +

OverageFee + RoamMins

Model 2: traindata$Churn ~ 1

	#Df	LogLik	Df	Chisq	Pr(>Chisq)	
1	11	-784.39				
2	1	-967.14	-10	365.5	< 2.2e-16	***

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

p value is 2.2e-16. Null hypothesis is rejected. Hence the model is a valid one.

<u>Check for multicollinearity</u>

Heteroscedasticity test

VIF values of the variables :

AccountWeeks	ContractRenewal	DataPlan	DataUsage	CustServCalls
1.006819	1.055450	17.158205	73.345912	1.067725

DayMins	DayCalls	MonthlyCharge	OverageFee	RoamMins
39.739575	1.005319	112.488165	9.264825	1.213842

VIF values of DataPlan, DataUsage, DayMins, MonthlyCharge, OverageFee are too high (>5)

Hence there is multicollinearity

Data Usage and Data Plan are highly correlated. Monthly Charge is also highly correlated with Data Usage, Data Plan and Day Mins.

The multicolliniearity has caused the inflated VIF values for correlated variables, making the model unreliable.

Model 2:

We will create a model after dropping DataUsage and Monthly Charge

```
Call:  glm(formula = traindata$Churn ~ ., family = "binomial", data = traindata[,
    -c(5, 9)])

Coefficients:
     (Intercept)      AccountWeeks  ContractRenewal1        DataPlan1
      -5.8266021         0.0001276        -1.8544423       -0.8432731
   CustServCalls           DayMins          DayCalls       OverageFee
       0.5187575         0.0117894         0.0024574        0.1313893
        RoamMins
       0.1003900
Degrees of Freedom: 2333 Total (i.e. Null);  2325 Residual
Null Deviance:      1934
Residual Deviance: 1581         AIC: 1599

glm(formula = traindata$Churn ~ ., family = "binomial", data = traindata[,
    -c(5, 9)])

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.8573  -0.5273  -0.3674  -0.2249   2.9067

Coefficients:
                   Estimate Std. Error z value Pr(>|z|)
(Intercept)      -5.8266021  0.6501531  -8.962  < 2e-16 ***
AccountWeeks      0.0001276  0.0016461   0.078    0.938
ContractRenewal1 -1.8544423  0.1704254 -10.881  < 2e-16 ***
DataPlan1        -0.8432731  0.1661374  -5.076 3.86e-07 ***
CustServCalls     0.5187575  0.0533146   9.730  < 2e-16 ***
DayMins           0.0117894  0.0012405   9.504  < 2e-16 ***
DayCalls          0.0024574  0.0032711   0.751    0.453
OverageFee        0.1313893  0.0271639   4.837 1.32e-06 ***
```

```
RoamMins          0.1003900  0.0248231   4.044 5.25e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Dispersion parameter for binomial family taken to be 1)
    Null deviance: 1934.3  on 2333  degrees of freedom
Residual deviance: 1581.4  on 2325  degrees of freedom
```

AIC: 1599.4

Number of Fisher Scoring iterations: 5

Now AccountWeeks and DayCalls seem to be less significant

Odds Ratio of variable in new model:

```
exp(coef(logmodel1)) #Odds ratio
    (Intercept)    AccountWeeks ContractRenewal1       DataPlan1
    0.002948077     1.000127631      0.156540222     0.430299819
   CustServCalls        DayMins         DayCalls      OverageFee
    1.679939003     1.011859129      1.002460402     1.140411677
       RoamMins
    1.105601987
```

Probability in new model :

exp(coef(logmodel1))/(1+exp(coef(logmodel1))) #Probability

(Intercept)	AccountWeeks	ContractRenewal1	DataPlan1
0.002939412	0.500031906	0.135352164	0.300845888
CustServCalls	DayMins	DayCalls	OverageFee
0.626857179	0.502947306	0.500614345	0.532800157
RoamMins			
0.525076436			

VIF values for the new model:

AccountWeeks	ContractRenewal	DataPlan	CustServCalls	DayMins
1.004141	1.051596	1.017276	1.058836	1.026393
DayCalls	OverageFee	RoamMins		
1.004381	1.019200	1.013826		

The values are less than 5, hence there is no multicollinearity

Print likelihood of the new model

(Intercept)	AccountWeeks	ContractRenewal1	DataPlan1
0.002948077	1.000127631	0.156540222	0.430299819
CustServCalls	DayMins	DayCalls	OverageFee
1.679939003	1.011859129	1.002460402	1.140411677
RoamMins			
1.105601987			

If there is 1 unit change in CustServCalls, there is 1.679939003 units change in the odds of Churn being '1'

#Probability=1.679939003/1+1.679939003 = 0.6268572

#If there is 1 unit increase in CustServCalls, probability of customer canceling the service increases by 62.69%

We shall predict on test data:

table(testdata$Churn,(predictTest>0.16))

Confusion Matrix:

	FALSE	TRUE
0	665	190
1	35	109

table(testdata$Churn,(predictTest>0.5))

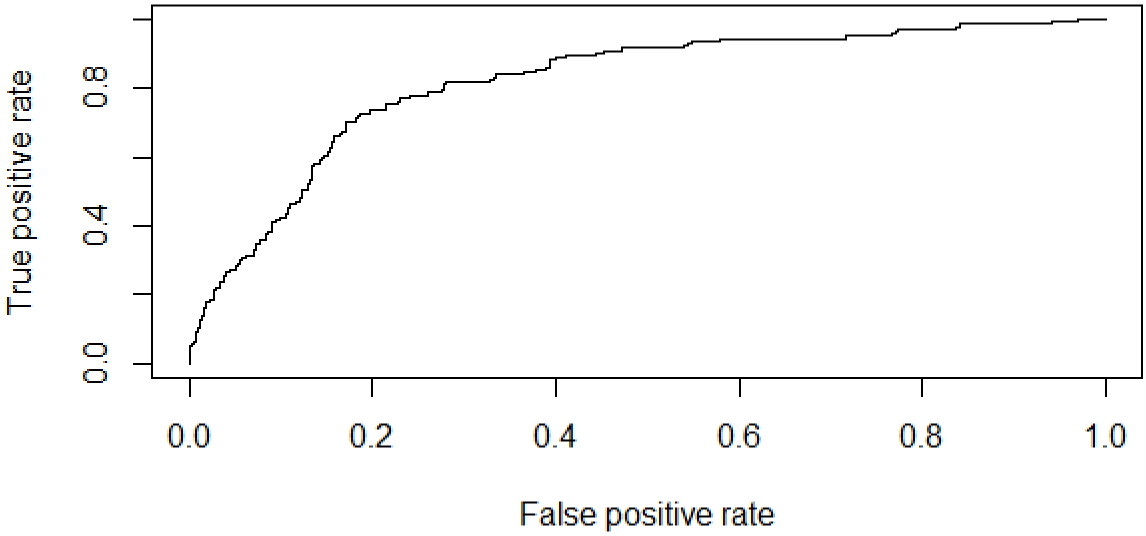
Confusion Matrix:

	FALSE	TRUE
0	840	15
1	120	24

Accuracy of the model is 86.48649%.

This predicts well on test data

Test set AUC: 0.8189327



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 82% of the time, the churned customers will have higher predicted churn and the non-churn customers will have low predicted churn.

Model 3

We will use a stepwise variable reduction function using VIF values. The function works like this:

- It uses the full set of explanatory variables.
- It calculates VIF for each variable,
- It removes the variable with the single highest value,
- It then recalculates all VIF values with the new set of variables,

It removes the variable with the next highest value, and so on, until all values are below the threshold.

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.8852  -0.5313  -0.3680  -0.2266   2.9254

Coefficients:
                 Estimate Std. Error z value Pr(>|z|)
(Intercept)      -5.55093    0.51916 -10.692  < 2e-16 ***
ContractRenewal1 -1.85669    0.17036 -10.899  < 2e-16 ***
DataPlan1        -0.84055    0.16602  -5.063 4.13e-07 ***
CustServCalls     0.51838    0.05325   9.735  < 2e-16 ***
DayMins           0.01178    0.00124   9.503  < 2e-16 ***
OverageFee        0.13017    0.02712   4.800 1.59e-06 ***
RoamMins          0.10042    0.02480   4.049 5.15e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1934.3  on 2333  degrees of freedom
Residual deviance: 1581.9  on 2327  degrees of freedom
```

AIC: 1595.9

```
Number of Fisher Scoring iterations: 5
```

Model 4

Model tuning and building model using balanced data using caret function

Generalized Linear Model

2334 samples

10 predictor

2 classes: '0', '1'

No pre-processing

Resampling: Cross-Validated (5 fold, repeated 3 times)

Summary of sample sizes: 1867, 1867, 1867, 1867, 1868, 1867, ...

Addtional sampling using up-sampling

Resampling results:

Accuracy	Kappa
0.7523568	0.3221023

variable importance

	Overall
CustServCalls	100.0000
ContractRenewal1	85.4596
MonthlyCharge	21.6999
DataUsage	18.4421
RoamMins	17.6624
DataPlan1	10.2646
DayMins	7.5541
OverageFee	6.5668
AccountWeeks	0.2107
DayCalls	0.0000

We shall predict on the test data

Confusion Matrix:

	0	1
0	643	212
1	30	114

Accuracy of 75.77578

Specificity and Sensitivity also shows that it is a good model

K Nearest Neighbour Algorithm

Model1

Removing correlated variables at k=7 gives better model performance:

Confusion Matrix:

	0	1
0	837	18
1	128	16

Overall Accuracy of 85.49%

Model2

Confusion Matrix:

	0	1
0	843	12
1	112	32

Overall Acuracy of 87.59%

Model3

Using Caret function

k-Nearest Neighbors

2334 samples

10 predictor

2 classes: '0', '1'

No pre-processing

Resampling: Cross-Validated (5 fold, repeated 3 times)

Summary of sample sizes: 1867, 1867, 1867, 1867, 1868, 1868, ...

Addtional sampling using up-sampling

Resampling results across tuning parameters:

k	Accuracy	Kappa
5	0.5825456	0.09842483
7	0.5582674	0.09337804
9	0.5534159	0.09000191

Accuracy was used to select the optimal model using the largest value.

The final value used for the model was k = 5.

Model4

After normalising continuous variables

k-Nearest Neighbors

2334 samples

10 predictor

2 classes: '0', '1'

No pre-processing

Resampling: Cross-Validated (3 fold)

Summary of sample sizes: 1556, 1556, 1556

Resampling results across tuning parameters:

k	Accuracy	Kappa
5	0.8924593	0.4702685
7	0.9027421	0.5105868
9	0.9010283	0.4892428
11	0.9005998	0.4778552
13	0.8950300	0.4318318
15	0.8971722	0.4283900
17	0.8907455	0.3876009
19	0.8856041	0.3337831
21	0.8834619	0.3136189
23	0.8787489	0.2759238

Accuracy was used to select the optimal model using the largest value.

The final value used for the model was k = 7.

Accuracy was used to select the optimal model using the largest value.

We shall now predict on test data

Confusion Matrix and Statistics:

```
          Reference
Prediction   0   1
         0 841  89
         1  14  55

               Accuracy : 0.8969
                 95% CI : (0.8764, 0.9151)
    No Information Rate : 0.8559
    P-Value [Acc > NIR] : 7.230e-05

                  Kappa : 0.4666

 Mcnemar's Test P-Value : 3.067e-13
            Sensitivity : 0.38194
            Specificity : 0.98363
         Pos Pred Value : 0.79710
         Neg Pred Value : 0.90430
             Prevalence : 0.14414
         Detection Rate : 0.05506
   Detection Prevalence : 0.06907
      Balanced Accuracy : 0.68279

       'Positive' Class : 1
```

NAÏVE BAYES

A-priori probabilities:

```
Y
        0         1
0.8547558 0.1452442

Conditional probabilities:
   AccountWeeks
Y        [,1]     [,2]
  0 100.9248 39.81273
  1 102.4189 39.56451

   ContractRenewal
Y            0          1
  0 0.06666667 0.93333333
  1 0.26843658 0.73156342

   DataPlan
Y           0         1
  0 0.7082707 0.2917293
  1 0.8259587 0.1740413

   DataUsage
Y         [,1]     [,2]
  0 0.8536341 1.281012
  1 0.5861652 1.202507

   CustServCalls
Y        [,1]     [,2]
  0 1.425063 1.101551
  1 2.050147 1.511443
```

```
  DayMins
Y       [,1]     [,2]
  0 175.7853 50.34229
  1 206.1029 68.66804

  DayCalls
Y       [,1]     [,2]
  0 100.2581 19.57531
  1 100.7876 21.29996

  MonthlyCharge
Y       [,1]     [,2]
  0 55.58782 15.93325
  1 59.27109 15.74741

  OverageFee
Y        [,1]     [,2]
  0  9.960682 2.489379
  1 10.612183 2.432404

  RoamMins
Y       [,1]     [,2]
  0 10.17840 2.662703
  1 10.80029 2.748679
```

Output gives prior probabilities

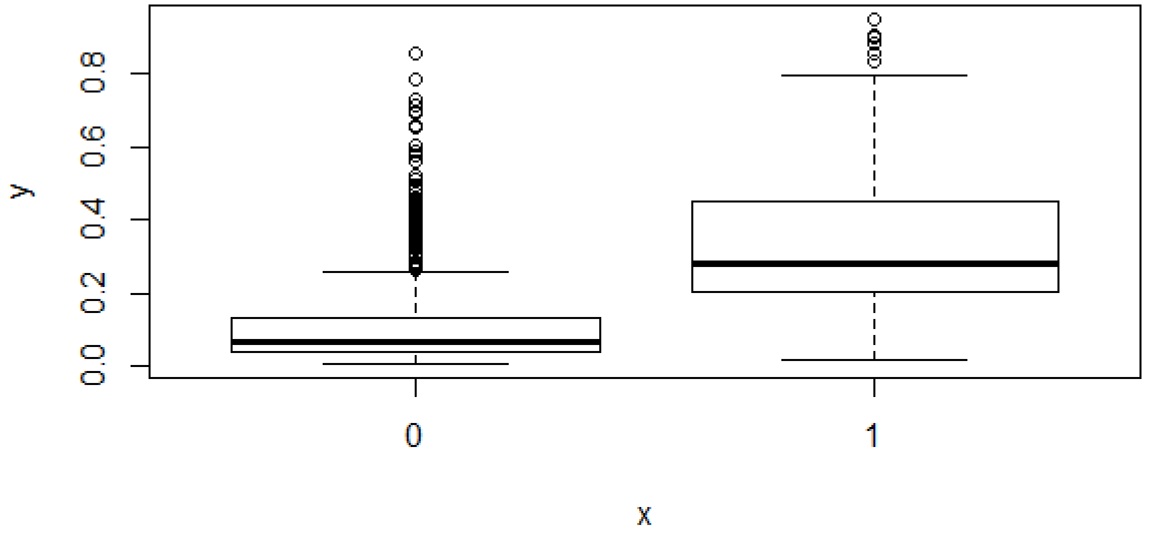
Churned customers have 102.4189 number of active account weeks with std deviation of 39.56451,

#0.5861652 gigabytes of monthly data usage with std dev of 1.202507, 2.050147 calls made to customer service with std dev of 1.511443,

#206.1029 average daytime mins/month with std dev of 68.66804,100.7876 average daytime calls with std dev of 21.29996, 59.27109 of monthly charge with std dev of 15.74741,

#10.612183 of largest overage fee in last 12 months with std dev of 2.432404, 10.80029 average roaming mins with std dev of 2.748679

We shall now predict on test data:



Confusion matrix

	0	1
0	840	15
1	119	25

Accuracy is 86.58%

Specificity and sensitivity shows that this is a good model

	Accuracy	**Sensitivity**	**Specificity**
Logistic Regression	86.2	34.7	94.9
K Nearest Neighbors	91.9	46.6	99.6
Naive Bayes	87.6	24.3	98.2

Accuracy and Sensitivity are relatively higher for **K Nearest Neighbors**