

Customer Churn

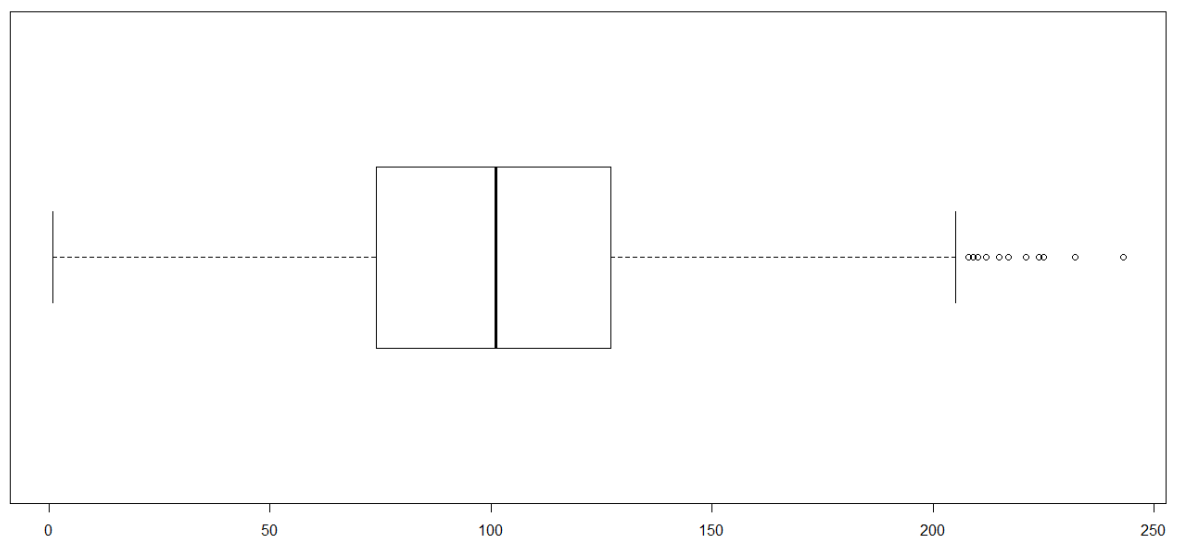
1. EDA (16 Marks)

- How does the data look like, Univariate and bivariate analysis. Plots and charts which illustrate the relationships between variables (4 Marks)
- Look out for outliers and missing values (4 Marks)
- Check for multicollinearity & treat it (4 Marks)
- Summarize the insights you get from EDA (4 Marks)

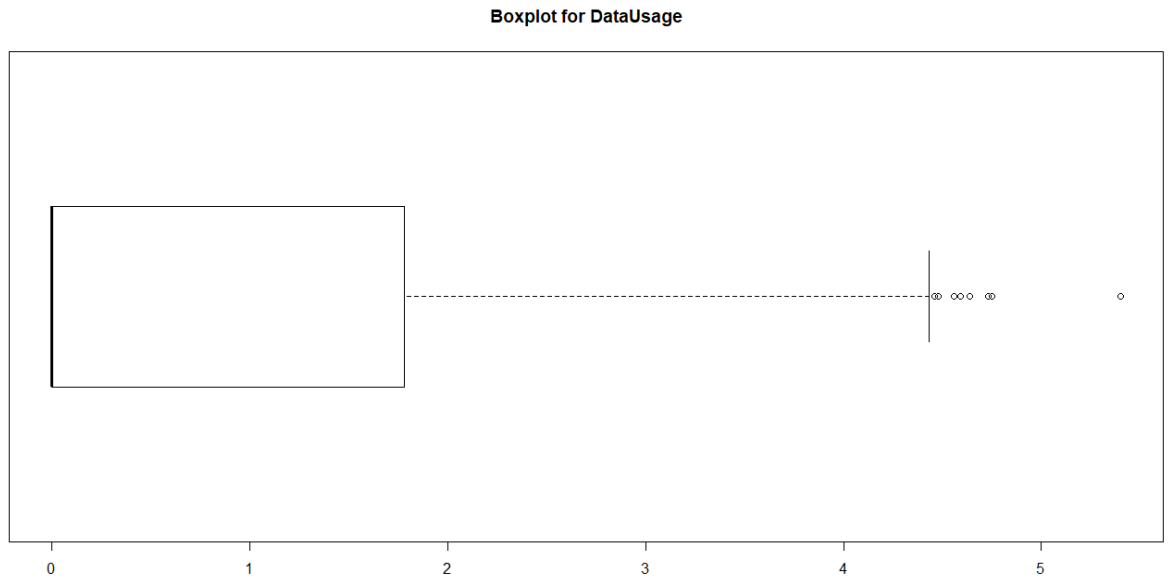
There are 10 variables that will determine the churn in subscribers. The variables are

1. AccountWeeks: This is the number of weeks that the user has an active account. As we can see that the distribution is normal and the average no. of weeks is 101.1 and the standard deviation is 39.8 weeks. There are some outliers in this data but there are no missing values.

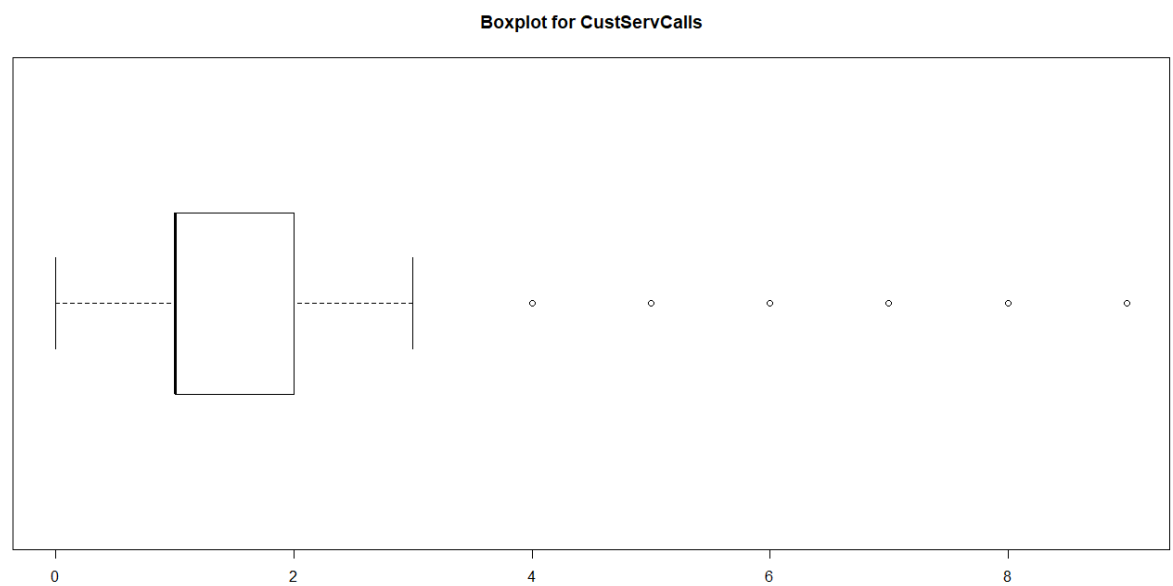
Boxplot for AccountWeeks



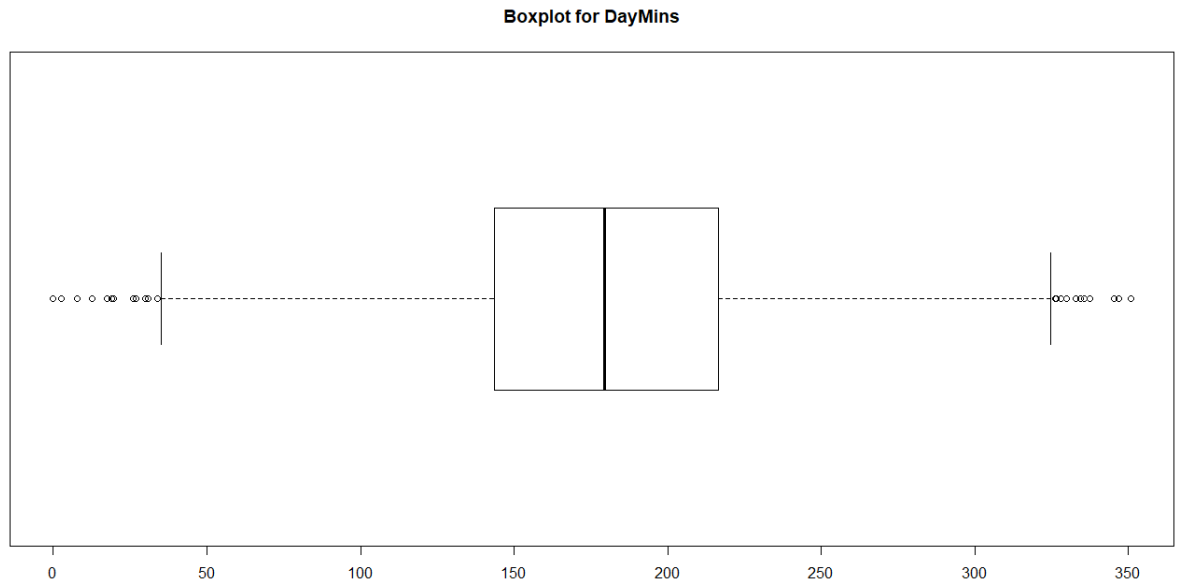
2. ContractRenewal: This is a categorical variable that shows if a user has had a contract renewal or not. The company has seen 90.3% contract renewals. There are no missing values.
3. DataPlan: This variable shows whether the subscriber has subscribed to the data plan. 72.3% subscribers have taken the data plan. There are no missing values here.
4. DataUsage: This variable shows the gigabytes of monthly data usage. On average people use 0.8 GB of data with a standard deviation of 1.3 GB. Since there is a huge percentage of people who have not subscribed to the data plan, this data is skewed. There are no missing values here.



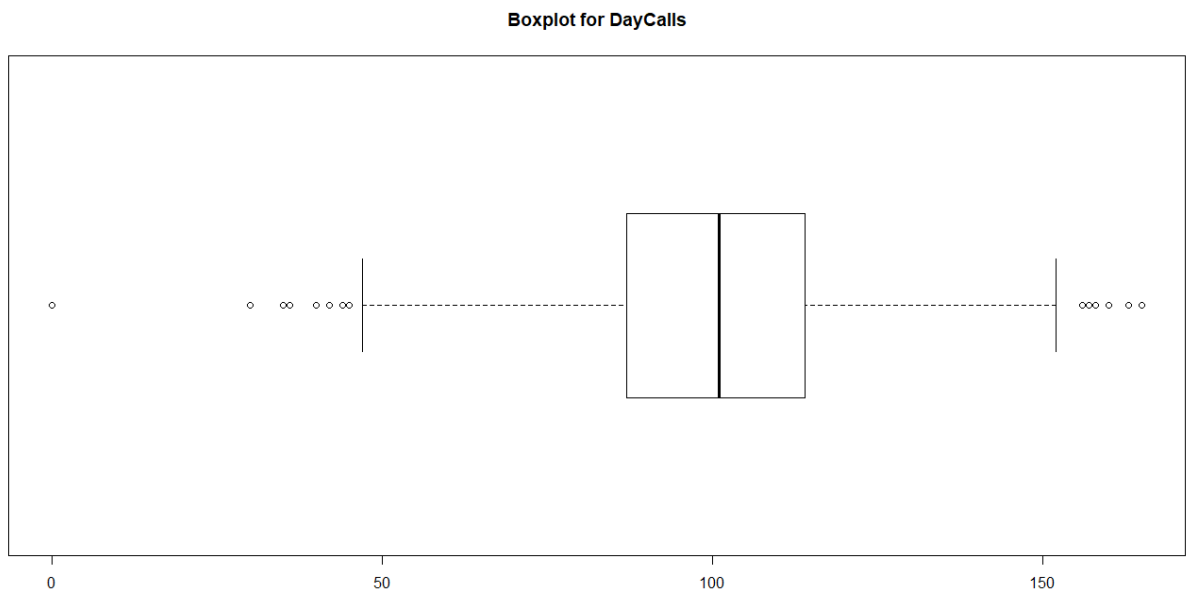
5. CustServeCalls: This is the no. of calls to the customer service by the customer. This is normally distributed with an average of 1.6 calls per customer and a standard deviation of 1.3 calls. There are no missing values.



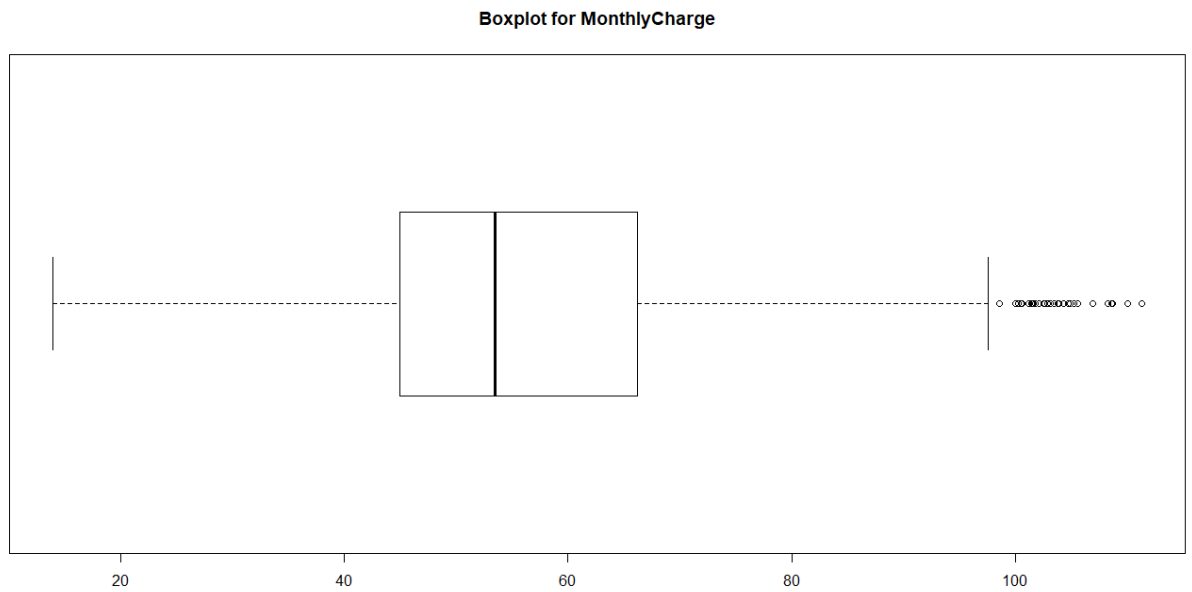
6. DayMins: This is the average number of daytime minutes per month. The average number daytime minutes per month is 179.8 with a standard deviation of 54.5 minutes. This is normally distributed. There are no missing values. There are a few outliers.



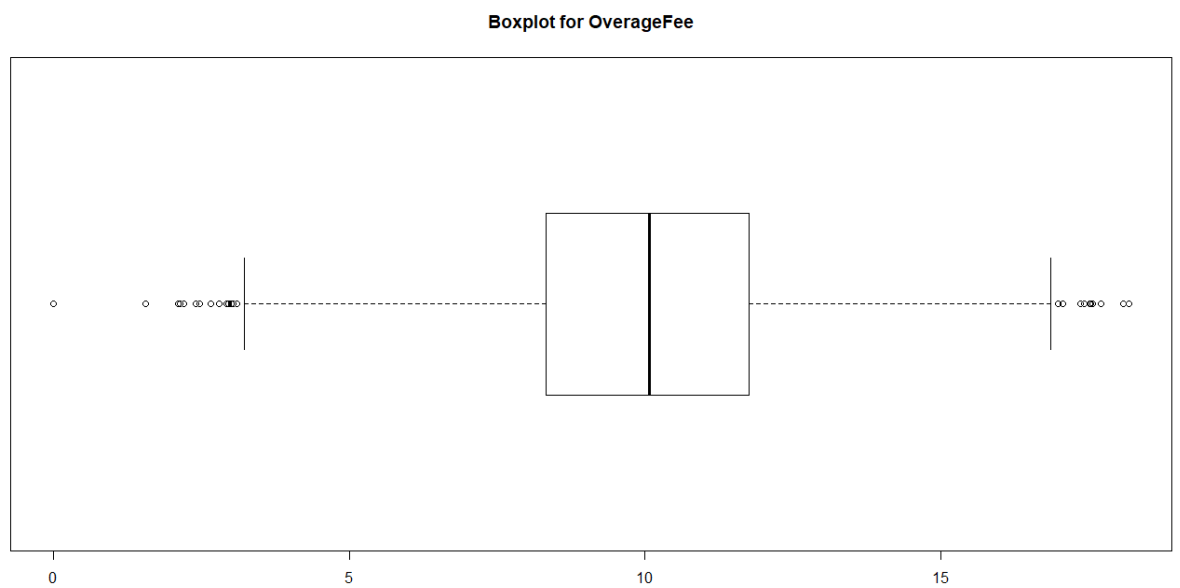
7. **DayCalls:** This is the average number of daytime calls. The average number of daytime calls is 100.4 with a standard deviation of 20.1. This is normally distributed. There are no missing values. There are a few outliers.



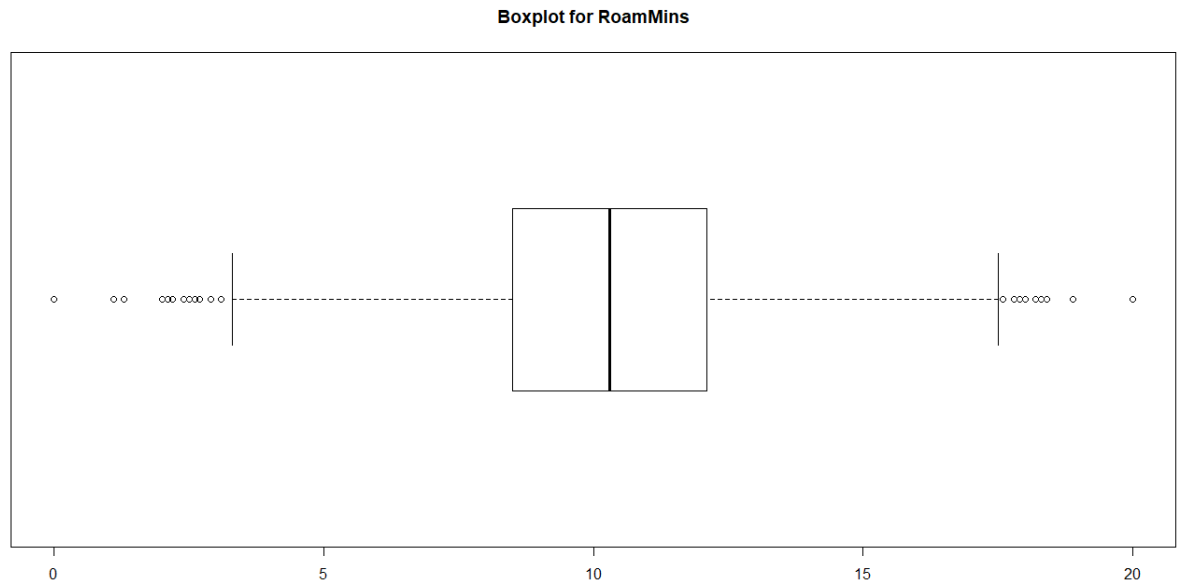
8. **MonthlyCharge:** This is the average monthly bill. This is normally distributed with an average of 56.3 and a standard deviation of 16.4. There are no missing values but there are a few outliers.



9. **OverageFee:** This is the largest overage fee in the last 12 months. This is normally distributed with an average of 10.1 and a standard deviation of 2.5. There are a few outliers but there are no missing values.


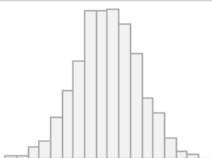
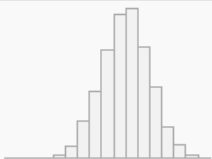
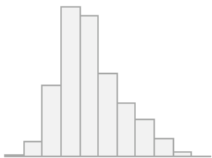
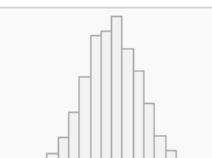
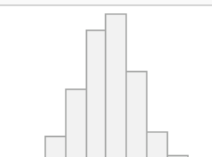


10. **RoamMins:** This is the average number of roaming minutes. This is normally distributed with an average of 10.2 and a standard deviation of 2.8. There are a few outliers but there are no missing values.

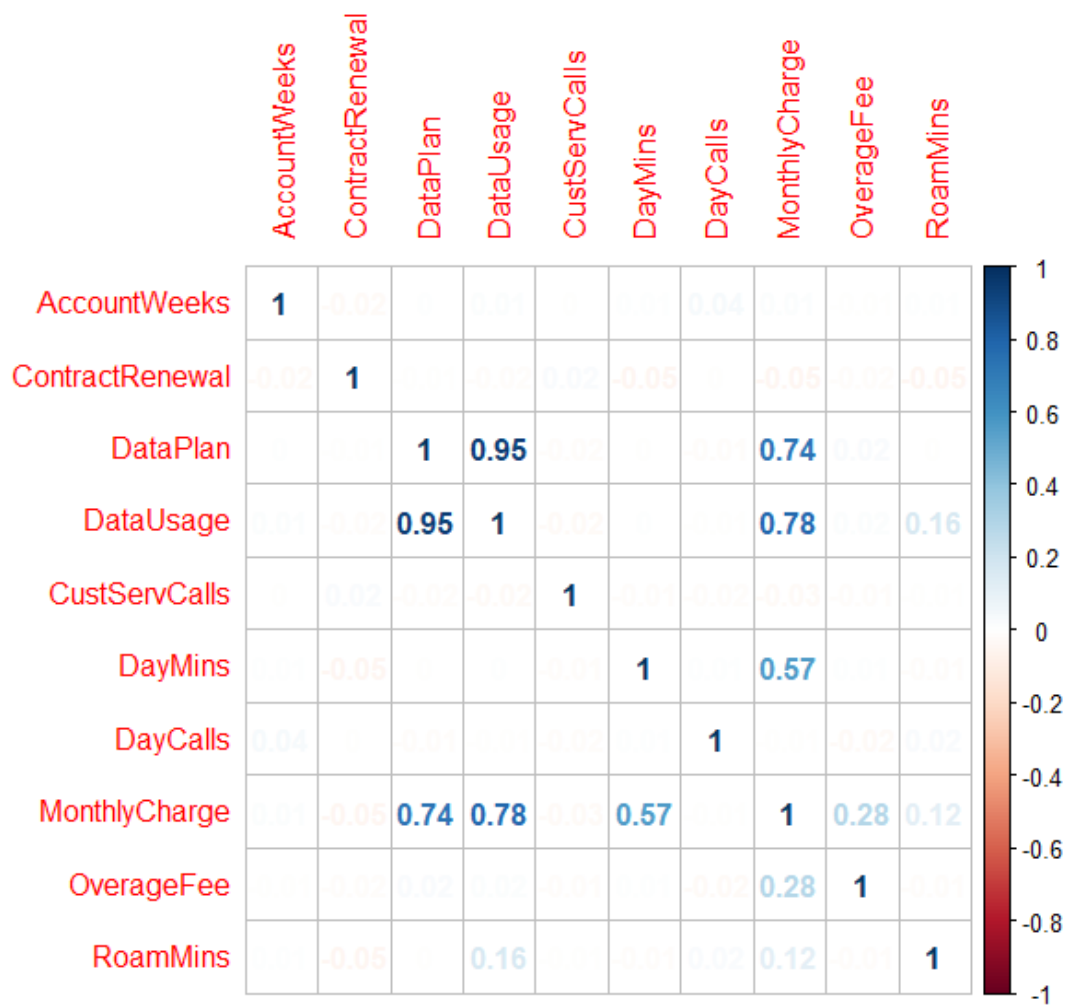


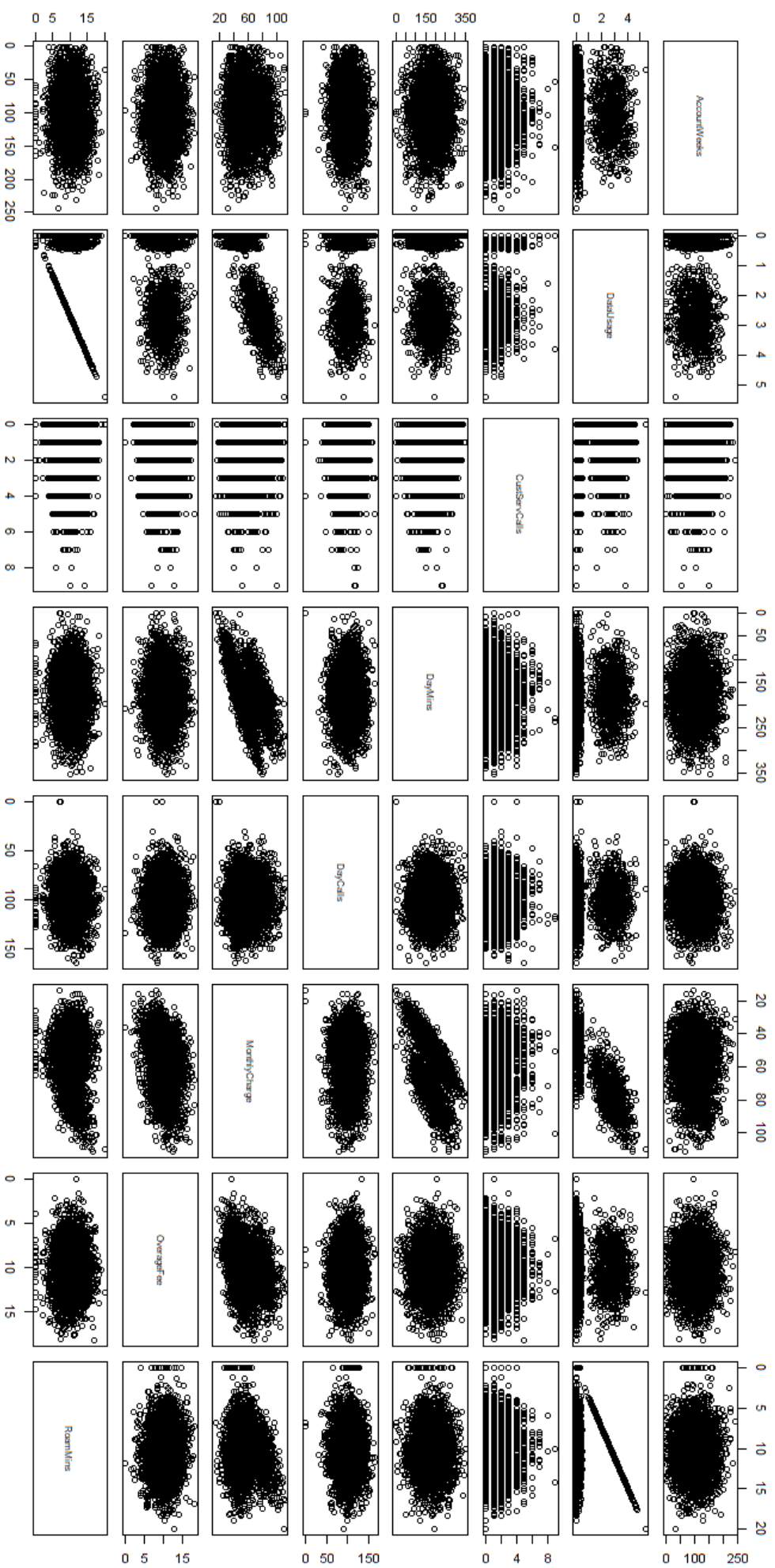
Below is a summary of all variables along with their histograms.

No	Variable	Stats / Values	Freqs (% of Valid)	Graph	Valid	Missing
1	Churn [numeric]	Min : 0 Mean : 0.1 Max : 1	0: 2850 (85.5%) 1: 483 (14.5%)		3333 (100%)	0 (0%)
2	AccountWeeks [numeric]	Mean (sd) : 101.1 (39.8) min < med < max: 1 < 101 < 243 IQR (CV) : 53 (0.4)	212 distinct values		3333 (100%)	0 (0%)
3	ContractRenewal [numeric]	Min : 0 Mean : 0.9 Max : 1	0: 323 (9.7%) 1: 3010 (90.3%)		3333 (100%)	0 (0%)
4	DataPlan [numeric]	Min : 0 Mean : 0.3 Max : 1	0: 2411 (72.3%) 1: 922 (27.7%)		3333 (100%)	0 (0%)
5	DataUsage [numeric]	Mean (sd) : 0.8 (1.3) min < med < max: 0 < 0 < 5.4 IQR (CV) : 1.8 (1.6)	174 distinct values		3333 (100%)	0 (0%)

6	CustServCalls [numeric]	Mean (sd) : 1.6 (1.3) min < med < max: 0 < 1 < 9 IQR (CV) : 1 (0.8)	0: 697 (20.9%) 1: 1181 (35.4%) 2: 759 (22.8%) 3: 429 (12.9%) 4: 166 (5.0%) 5: 66 (2.0%) 6: 22 (0.7%) 7: 9 (0.3%) 8: 2 (0.1%) 9: 2 (0.1%)		3333 (100%)	0 (0%)
7	DayMins [numeric]	Mean (sd) : 179.8 (54.5) min < med < max: 0 < 179.4 < 350.8 IQR (CV) : 72.7 (0.3)	1667 distinct values		3333 (100%)	0 (0%)
8	DayCalls [numeric]	Mean (sd) : 100.4 (20.1) min < med < max: 0 < 101 < 165 IQR (CV) : 27 (0.2)	119 distinct values		3333 (100%)	0 (0%)
9	MonthlyCharge [numeric]	Mean (sd) : 56.3 (16.4) min < med < max: 14 < 53.5 < 111.3 IQR (CV) : 21.2 (0.3)	627 distinct values		3333 (100%)	0 (0%)
10	OverageFee [numeric]	Mean (sd) : 10.1 (2.5) min < med < max: 0 < 10.1 < 18.2 IQR (CV) : 3.4 (0.3)	1024 distinct values		3333 (100%)	0 (0%)
11	RoamMins [numeric]	Mean (sd) : 10.2 (2.8) min < med < max: 0 < 10.3 < 20 IQR (CV) : 3.6 (0.3)	162 distinct values		3333 (100%)	0 (0%)

Below are a few graphs that show the relationship among variables.





As can be seen there is a positive correlation between account weeks and data usage.

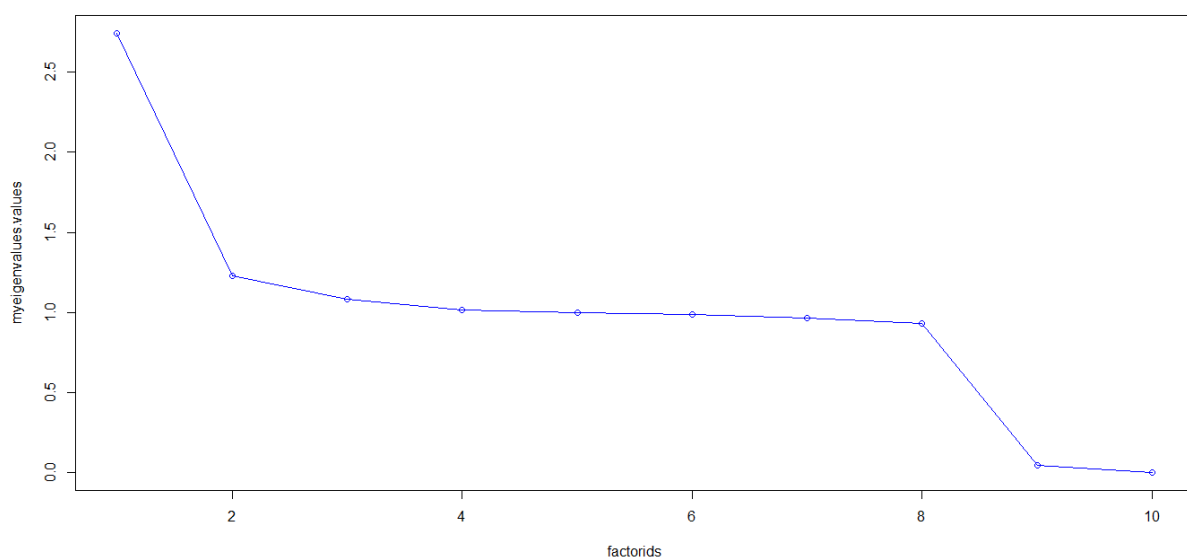
Data usage has a positive correlation with monthly charge since anybody who uses data will have a higher monthly charge. Data usage also has a positive correlation with roaming mins.

Day Minutes has a positive correlation with monthly charge since monthly charge increases with more talk time.

Monthly charge and overage fee are positively related.

This correlation is also shown in a colour coded chart below. Hence this dataset will have to be treated for multicollinearity.

We perform a scree plot to see the possibility of dimensionality reduction and we come across that 8 of the 10 variables have eigenvalues over 1 and hence should be considered as principle factors.



On performing the principal component analysis, we club the highly correlated variables into a single variable to avoid repetition of variables.

	RC1	RC2	RC4	RC3	RC8	RC6	RC5	RC7	h2
Accountweeks	0.01	0	0	0	-0.01	1	0	0.02	1
ContractRenewal	-0.01	-0.03	-0.01	-0.02	1	-0.01	0.01	0	1
DataPlan	0.98	-0.04	-0.02	-0.06	0	0	-0.01	0	0.97
DataUsage	0.99	-0.02	-0.02	0.11	-0.01	0.01	-0.01	0	0.99
CustServCalls	-0.01	-0.01	-0.01	0	0.01	0	1	-0.01	1
DayMins	0.03	1	-0.01	-0.01	-0.02	0	-0.01	0	1
DayCalls	-0.01	0	-0.01	0.01	0	0.02	-0.01	1	1
Monthlycharge	0.79	0.55	0.24	0.08	-0.02	0.01	-0.01	0	1
OverageFee	0.04	0.02	1	-0.01	-0.01	0	-0.01	-0.01	1
RoamMins	0.06	0	-0.01	1	-0.02	0	0	0.01	1

As can be seen here the findings are in line with the preliminary analysis. Here the three variables of MonthlyCharge, DataPlan and DataUsage are strongly correlated and hence can be clubbed into 1 variable. All the other variables are independent of each other.

The variables can be categorised as following:

RC1: DataPlan, MonthlyCharge and DataUsage.

RC2: DayMins

RC3: RoamMins

RC4: OverageFee

RC5: CustServCalls

RC6: AccountWeeks

RC7: DayCalls

RC8: ContractRenewal

On performing the principal component analysis, the variables are also normalised.

2. **Build Models and compare them to get to the best one (39 Marks)**

- Logistic Regression (8 Marks)
- KNN (8 Marks)
- Naive Bayes (8 Marks) (is it applicable here? comment and if it is not applicable, how can you build an NB model in this case?)
- Model Comparison using Model Performance metrics & Interpretation (15 Marks)

For the logistic regression we perform Principal Component Analysis on the data and observe the following results.

	Estimate	Std. Error	z value	Pr(> z)	
(Intercept)	-2.27331	0.06978	-32.579	< 2e-16	***
RC1	-0.35834	0.0626	-5.724	1.04E-08	***
RC2	0.72554	0.05871	12.358	< 2e-16	***
RC3	0.25161	0.05703	4.412	1.03E-05	***
RC4	0.35299	0.05739	6.151	7.70E-10	***
RC5	0.65369	0.05097	12.826	< 2e-16	***
RC6	0.03554	0.05522	0.644	0.52	
RC7	0.07045	0.05508	1.279	0.201	
RC8	-0.60555	0.04253	-14.237	< 2e-16	***

It can be seen that the reduced variables are highly significant. We shall remove the insignificant variables and perform the logistic regression again.

The new model looks like this

	Estimate	Std. Error	z value	Pr(> z)	
(Intercept)	-2.27097	0.06968	-32.593	< 2e-16	***
RC1	-0.35776	0.06254	-5.72	1.06E-08	***
RC2	0.72586	0.05872	12.361	< 2e-16	***
RC3	0.2507	0.05697	4.4	1.08E-05	***
RC4	0.35161	0.0574	6.125	9.06E-10	***
RC5	0.65294	0.05089	12.832	< 2e-16	***
RC8	-0.60543	0.04253	-14.234	< 2e-16	***

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 2758.3 on 3332 degrees of freedom
Residual deviance: 2191.3 on 3326 degrees of freedom
AIC: 2205.3

Number of Fisher Scoring iterations: 5

Model performance

We now split the data into 70:30 ratio for training and testing dataset. We train the model on the bigger dataset and apply it to the smaller dataset and observe the following results. The obtained result is shown below.

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)	
(Intercept)	-2.30971	0.08490	-27.206	< 2e-16	***
RC1	-0.40703	0.07644	-5.325	1.01e-07	***
RC2	0.76226	0.07091	10.750	< 2e-16	***
RC3	0.31792	0.06867	4.629	3.67e-06	***
RC4	0.35731	0.06865	5.205	1.94e-07	***
RC5	0.66117	0.06032	10.961	< 2e-16	***
RC8	-0.59295	0.05159	-11.493	< 2e-16	***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

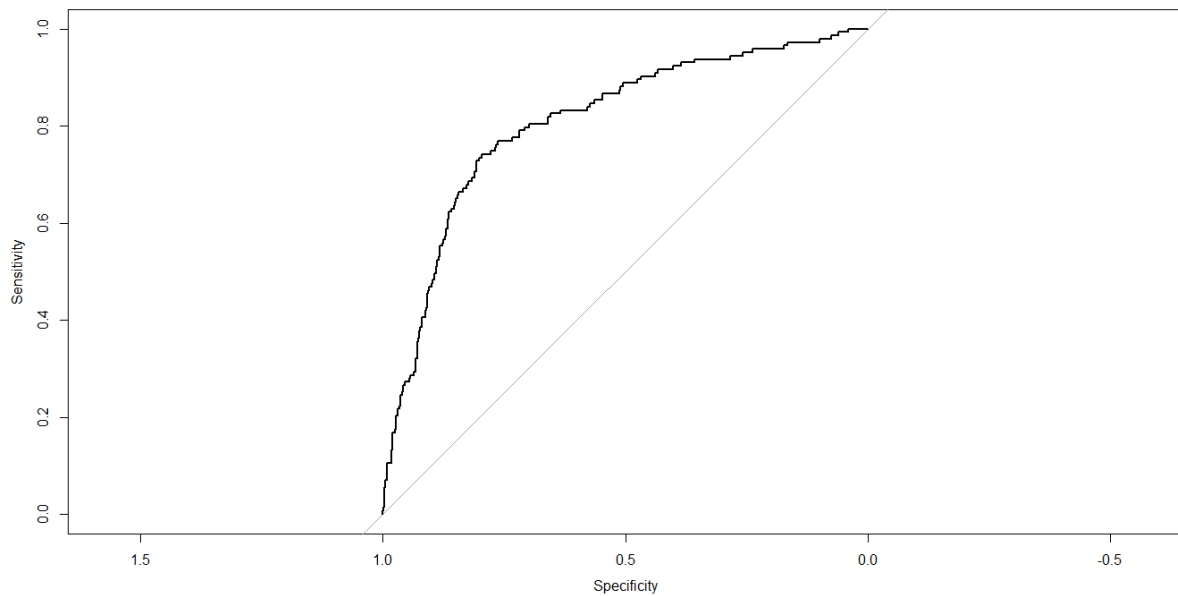
Null deviance: 1946.9 on 2362 degrees of freedom
Residual deviance: 1535.5 on 2356 degrees of freedom
AIC: 1549.5

Number of Fisher Scoring iterations: 6

On applying the model to the testing dataset, we obtain the following confusion matrix.

	obs	
pred	0	1
0	805	115
1	22	28

The model has an accuracy of 85.87%. The ROC curve is as shown below.



The area under the curve is 0.8074.

For the training dataset the results are as follows.

```

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.30971    0.08490 -27.206  < 2e-16 ***
RC1          -0.40703    0.07644  -5.325 1.01e-07 ***
RC2           0.76226    0.07091  10.750  < 2e-16 ***
RC3           0.31792    0.06867   4.629 3.67e-06 ***
RC4           0.35731    0.06865   5.205 1.94e-07 ***
RC5           0.66117    0.06032  10.961  < 2e-16 ***
RC8          -0.59295    0.05159 -11.493  < 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: 1946.9  on 2362  degrees of freedom
Residual deviance: 1535.5  on 2356  degrees of freedom
AIC: 1549.5

Number of Fisher Scoring iterations: 6

```

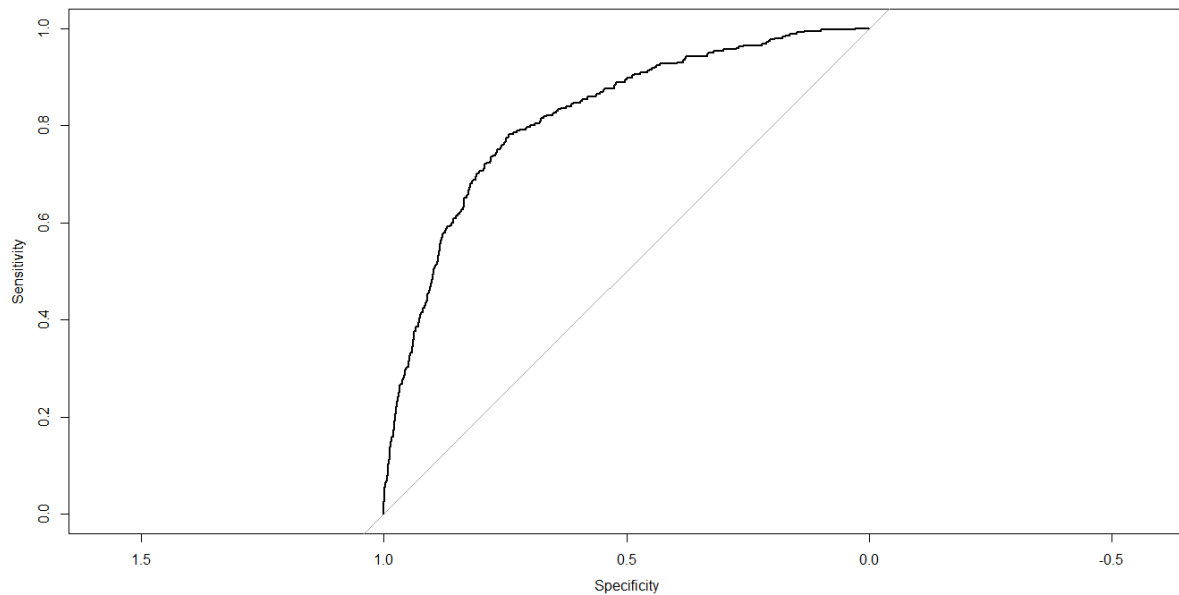
The confusion matrix is as shown below.

```

obs
pred  0  1
  0 1977 274
  1   46   66

```

The accuracy for the training dataset is 86.46%. The ROC curve is as shown below.



The area under the curve is 0.8175.

Since the training dataset and the testing dataset have extremely close values, we can say that the model is highly robust.

k-Nearest Neighbour

For the KNN algorithm first split the data into testing and training algorithms in the ratio of 70:30.

Then the model is trained on the 70% dataset and the findings are applied to the 30% dataset.

Here since the sample size is 3333 which is larger than 400, we can take the value of $k=19$.

We see the following confusion matrix for KNN.

```
y_pred
      0      1
0  826      1
1   61     82
```

The accuracy ratio for this is 93.61%.

The customer loss is 6.88%.

The opportunity lost is 1.2%.

The overall loss percentage is given by $0.95 \times \text{customer lost} + 0.05 \times \text{opportunity lost}$ which is equal to 6.59%.

Naïve-Bayes

For Naïve-Bayes there must be independence among the input variables and therefore it is must to remove multicollinearity and that has been done by applying the Principal Component Analysis. The data is split into two datasets of 70:30 ratio. The bigger dataset is used as the training dataset and is applied to the testing dataset. On applying Naïve-Bayes we obtain the following.

```
y_pred.NB
0 0 1
0 827 0
1 0 143
```

Here we have an accuracy of 100%.

The business lost is 0%.

The opportunity lost is 0%.

The overall loss percentage is given by $0.95 \times \text{business lost} + 0.05 \times \text{opportunity lost}$ which is equal to 0%.

3. Actionable Insights (5 marks)

- Interpretation & Recommendations from the best model

Among all the models, the model with the highest accuracy is Naïve-Bayes, which gives us an accuracy of 100%. The second-best model is the model of k-Nearest Neighbours which gives an accuracy of 93.61%. Logistic regression gives us an accuracy of 85.87%.

Among all the variables observed we have observed that the following variables have significantly contributed to the customer churn. These variables are ContractRenewal, DataPlan, DataUsage, CustServCalls, DayMins, MonthlyCharge, OverageFee and RoamMins. The variables AccountWeeks and DayCalls are not significant and hence were not considered for the logistic regression model.

Hence it can be concluded that in this case the Naïve-Bayes model is the best and can be applied to predict customer churn. Naïve-Bayes model has given us a very high level of accuracy.

The k-Nearest Neighbours also produces a robust model with an accuracy of 93.61%. However it is not as powerful as the Naïve-Bayes.

Logistic Regression has also given us a robust model with an accuracy of 85.87%. However it is not as good as the k-Nearest Neighbour and Naïve-Bayes algorithm.