2nd project

2023-12-17

Data preparation

Firstly, we must convert characters to factors and take a look to the summary of our data.

```
df[sapply(df, is.character)] <- lapply(df[sapply(df, is.character)], as.factor)
df$SeniorCitizen = factor(df$SeniorCitizen)
summary(df)
```

```
##       customerID      gender     SeniorCitizen Partner    Dependents
##  0002-ORFBO:   1   Female:3488   0:5901        No :3641   No :4933
##  0003-MKNFE:   1   Male  :3555   1:1142        Yes:3402   Yes:2110
##  0004-TLHLJ:   1
##  0011-IGKFF:   1
##  0013-EXCHZ:   1
##  0013-MHZWF:   1
##  (Other)   :7037
##      tenure      PhoneService          MultipleLines     InternetService
##  Min.   : 0.00   No : 682     No              :3390   DSL        :2421
##  1st Qu.: 9.00   Yes:6361     No phone service: 682   Fiber optic:3096
##  Median :29.00                Yes             :2971   No         :1526
##  Mean   :32.37
##  3rd Qu.:55.00
##  Max.   :72.00
##
##              OnlineSecurity              OnlineBackup
##  No                 :3498   No                 :3088
##  No internet service:1526   No internet service:1526
##  Yes                :2019   Yes                :2429
##
##             DeviceProtection              TechSupport
##  No                 :3095    No                 :3473
##  No internet service:1526    No internet service:1526
##  Yes                :2422    Yes                :2044
##
##               StreamingTV              StreamingMovies          Contract
##  No                 :2810   No                 :2785   Month-to-month:3875
##  No internet service:1526   No internet service:1526   One year      :1473
##  Yes                :2707   Yes                :2732   Two year      :1695
##
##  PaperlessBilling                   PaymentMethod  MonthlyCharges
##  No :2872         Bank transfer (automatic):1544   Min.   : 18.25
##  Yes:4171         Credit card (automatic)  :1522   1st Qu.: 35.50
##                   Electronic check         :2365   Median : 70.35
##                   Mailed check             :1612   Mean   : 64.76
##                                                    3rd Qu.: 89.85
##                                                    Max.   :118.75
##
##   TotalCharges    Churn
##  Min.   :  18.8   No :5174
##  1st Qu.: 401.4   Yes:1869
##  Median :1397.5
##  Mean   :2283.3
##  3rd Qu.:3794.7
##  Max.   :8684.8
##  NA's   :11
```

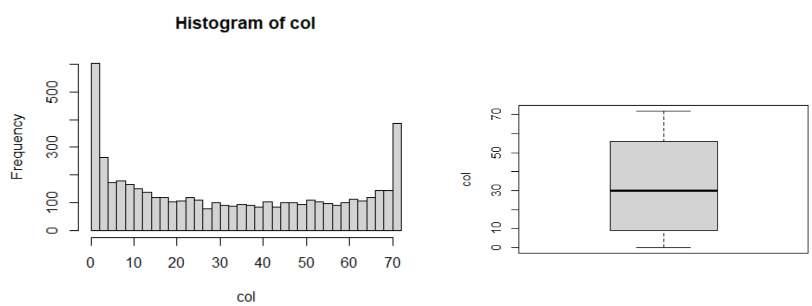
Next step is to split data to train and test, where 30% will be the test set.

```r
sample <- sample(c(TRUE, FALSE), nrow(df), replace=TRUE, prob=c(0.7,0.3))
train  <- df[sample, ]
test   <- df[!sample, ]
```

Exploration of variables and outliers

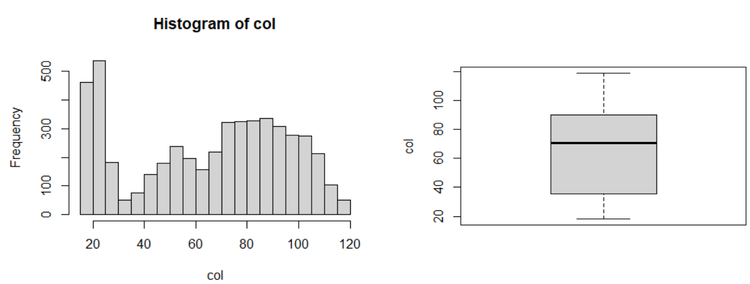
1. **tenure**

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##    0.00    9.00   30.00   32.69   56.00   72.00
```



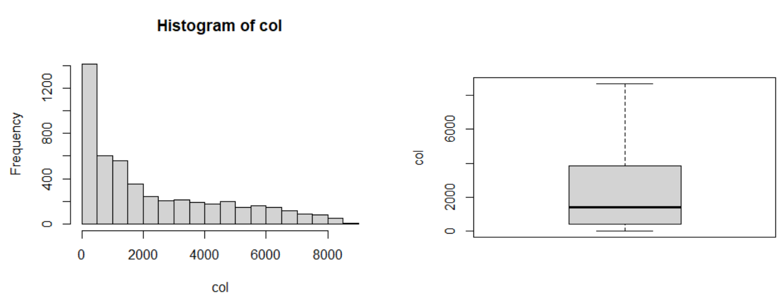
2. **MonthlyCharges**

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   18.25   35.25   70.40   64.74   89.85  118.75
```

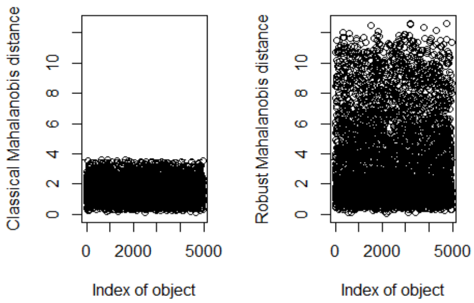


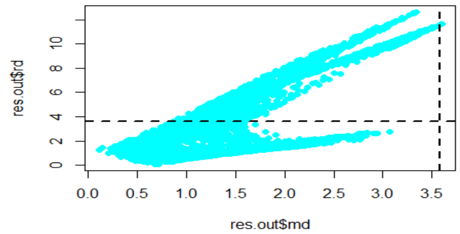
3. **TotalCharges**

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's 
##    18.8   399.7  1404.5  2302.8  3864.8  8684.8       7
```



It is observed that our 3 numerical variables do not have univariate outliers. So, now we check for multivariate outliers and remove them.





There are 3 observations that are after the cutoff point. Let's observe these rows. It appears that the outliers have distinct characteristics, and some of the features that differ between the outliers include:

All of them are Males, have partners and no dependents. They also have a Phone service, while the other variables differ, with some of them with 3 categories, having one person in each, and other categories having 2 of them and tenure, MonthlyCharges and TotalCharges vary across the outliers. Lastly, 2 of them did not Churn. So there is pattern in these 3 outliers. We remove them from the dataset.

```
##        customerID    gender   SeniorCitizen Partner Dependents     tenure
##   3211-ILJTT:1    Female:0    0:2           No :0    No :3      Min.   :17.0
##   5787-KXGIY:1    Male  :3    1:1           Yes:3    Yes:0      1st Qu.:33.0
##   9681-KYGYB:1                                                  Median :49.0
##   0002-ORFBO:0                                                  Mean   :46.0
##   0003-MKNFE:0                                                  3rd Qu.:60.5
##   0004-TLHLJ:0                                                  Max.   :72.0
##   (Other)   :0
##   PhoneService          MultipleLines    InternetService
##   No :0         No              :2     DSL        :0
##   Yes:3         No phone service:0     Fiber optic:2
##                 Yes             :1     No         :1
##

##             OnlineSecurity              OnlineBackup
##   No                 :1      No                 :1
##   No internet service:1      No internet service:1
##   Yes                :1      Yes                :1
##
##            DeviceProtection              TechSupport              StreamingTV
##   No                 :1       No                 :2    No                 :2
##   No internet service:1       No internet service:1    No internet service:1
##   Yes                :1       Yes                :0    Yes                :0
##
##            StreamingMovies              Contract PaperlessBilling
##   No                 :2      Month-to-month:2    No :1
##   No internet service:1      One year      :0    Yes:2
##   Yes                :0      Two year      :1
##
##                   PaymentMethod MonthlyCharges   TotalCharges   Churn
##   Bank transfer (automatic):1     Min.   :19.30   Min.   :1214   No :2
##   Credit card (automatic)  :1     1st Qu.:44.85   1st Qu.:1259   Yes:1
##   Electronic check         :1     Median :70.40   Median :1305
##   Mailed check             :0     Mean   :59.30   Mean   :2226
##                                   3rd Qu.:79.30   3rd Qu.:2732
##                                   Max.   :88.20   Max.   :4159
```

Missing Values

What is the percentage of missing values?

```
missing_mean = mean(is.na(train)) ;missing_mean*100

## [1] 0.006708258
```

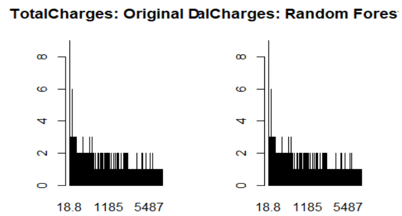
Observe the rows that have missing values. It appears that the rows with missing values have NaNs in the "TotalCharges" column. The "TotalCharges" column has 7 missing values in these rows. Additionally, there are specific characteristics common to these rows, such as having "PhoneService" with "Yes," "Contract" being "Two year," and "PaymentMethod" being "Mailed check, "PaperlessBilling" being no, "Partners" and "Dependents" being yes and tenure being 0."

```
##       customerID    gender   SeniorCitizen Partner Dependents     tenure
##  2520-SGTTA:1    Female:3   0:7           No :1   No :0      Min.   :0
##  2775-SEFEE:1    Male  :4   1:0           Yes:6   Yes:7      1st Qu.:0
##  3213-VVOLG:1                                                Median :0
##  4075-WKNIU:1                                                Mean   :0
##  4367-NUYAO:1                                                3rd Qu.:0
##  5709-LVOEQ:1                                                Max.   :0
##  (Other)   :1
##  PhoneService          MultipleLines    InternetService
##  No :0        No              :3     DSL        :3
##  Yes:7        No phone service:0     Fiber optic:0
##               Yes             :4     No         :4
##
##             OnlineSecurity              OnlineBackup
##  No                 :1      No                 :0
##  No internet service:4      No internet service:4
##  Yes                :2      Yes                :3
##
##            DeviceProtection              TechSupport              StreamingTV
##  No                 :1       No                 :1    No                 :1
##  No internet service:4       No internet service:4    No internet service:4
##  Yes                :2       Yes                :2    Yes                :2
##
##            StreamingMovies           Contract PaperlessBilling
##  No                 :2      Month-to-month:0   No :6
##  No internet service:4      One year      :0   Yes:1
##  Yes                :1      Two year      :7
##
##                    PaymentMethod MonthlyCharges   TotalCharges Churn
##  Bank transfer (automatic):1     Min.   :19.85   Min.   : NA   No :7
##  Credit card (automatic)  :0     1st Qu.:22.68   1st Qu.: NA   Yes:0
##  Electronic check         :0     Median :25.75   Median : NA
##  Mailed check             :6     Mean   :43.86   Mean   :NaN
##                                  3rd Qu.:67.62   3rd Qu.: NA
##                                  Max.   :80.85   Max.   : NA
##                                                  NA's   :7
```

Impute missing values with random Forest

```
set.seed(17)
rf_imp <- missForest(train[,c(2,3,4,5,7,8,9,10,11,12,13,14,15,16,17,18,21,6,19,20)],
variablewise=T, verbose=T)
```

The evaluation of the imputation by plots is same before and after imputation, but we also have to check the summary of the datasets before and after. We can see that the mean was 2302.9 before imputation and it became 2300.8 after. The statistics did not alter, thus we proceed with the analysis.



Evaluate imputation by statistics

```
## [1] "TotalCharges"

## [1] "Original:"

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##    18.8   399.5  1406.0  2302.9  3864.1  8684.8       7

## [1] "Random Forest:"

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    18.8   396.5  1402.7  2300.8  3857.0  8684.8
```

Oversampling

Our data are imbalanced, thus we oversample the train data, so Churn yes and no are almost equally distributed, using the ROSE library.

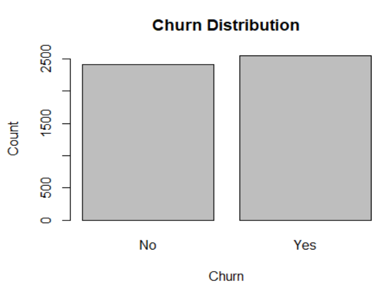
```
train_rose <- ovun.sample(Churn~., data = train, N = nrow(train), p = 0.5)$data
```

Original Data:

```
Churn
No :3659
Yes:1307
```

After oversampling:

```
Churn
No :2417  (48.67%)
Yes:2549  (51.32%)
```



Target Churn profiling

We use catdes to find relationships of Churn variable with the other variables of the dataset and with their respective categories. This output is in the ANNEX.

```
res.cat <-catdes(train, num.var=20)
res.cat$test.chi
```

```
res.cat$quanti

res.cat$quanti.var

res.cat$category
```

Insights:

Contract: Customers with month-to-month contracts are significantly more likely to churn compared to those with one-year or two-year contracts. This suggests that longer contract terms contribute to higher customer retention.

OnlineSecurity and TechSupport: The absence of online security and tech support is strongly associated with higher churn. Customers who do not have these services are more likely to churn, highlighting the importance of providing robust online security and technical support features.

InternetService: Fiber optic internet service is associated with higher churn compared to DSL. Companies may need to investigate and address issues related to fiber optic service quality or explore incentives for customers to stay with fiber optic plans.

PaymentMethod: Customers using electronic checks as their payment method are more likely to churn compared to other payment methods. Encouraging alternative payment methods might be beneficial in reducing churn.

Tenure: Shorter tenure is associated with higher churn. Customers who have been with the company for a shorter period are more likely to churn. Encouraging customer loyalty programs or providing incentives for longer-term relationships could be effective.

Dependents and PaperlessBilling: The presence of dependents and paperless billing is associated with lower churn. Customers with dependents and those who opt for paperless billing are less likely to churn. Companies may consider promoting paperless billing options and targeting family-oriented services.

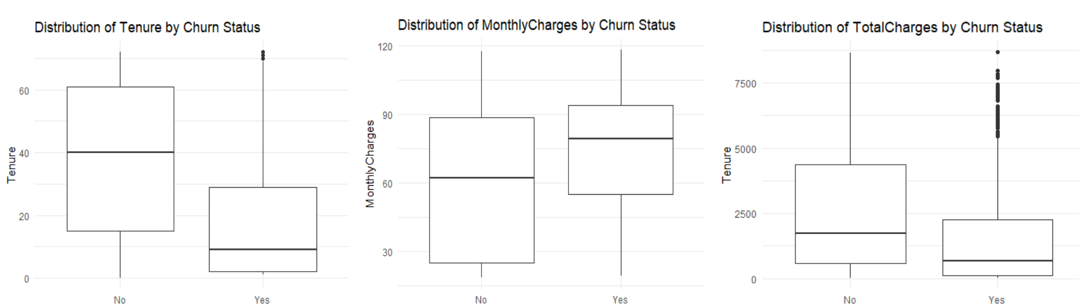
Partner: Customers without a partner are more likely to churn compared to those with a partner. Promoting family or partner plans and providing targeted offers could be strategies to improve retention.

SeniorCitizen: Senior citizens are less likely to churn compared to non-senior citizens. This suggests that services catering to senior citizens may contribute to higher customer loyalty.

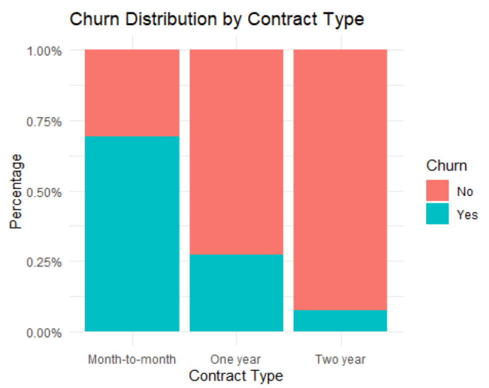
MultipleLines: Having multiple lines (phone service) is associated with higher churn. Companies may need to explore reasons behind this association, such as service quality or pricing issues related to multiple lines.

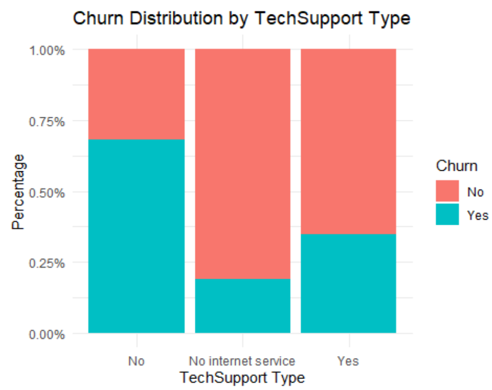
MonthlyCharges and TotalCharges: For customers who do not churn, lower monthly charges, higher total charges, and longer tenure are observed. This implies that customers who are more cost-sensitive and have a longer history with the company are less likely to churn.

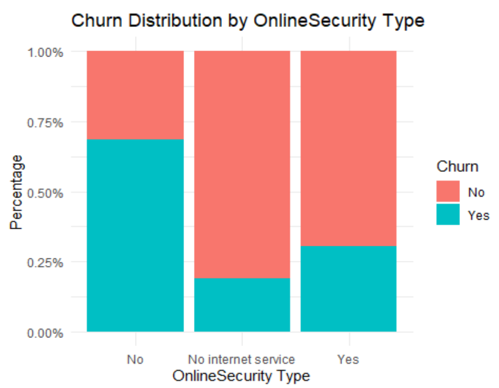
Plot Churn by numeric variables

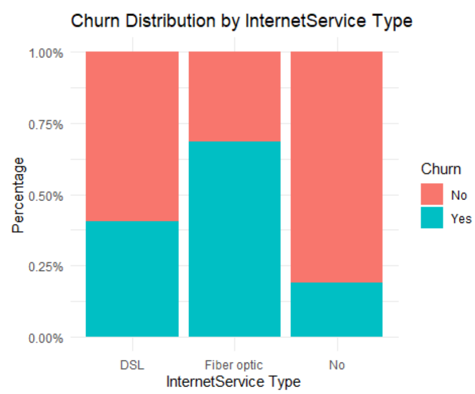


Plot Churn by factors









Future Selection

Find importance of each variable, by using a glm model and checking the coefficients of the variables in order to obtain variable importance.

```
set.seed(123)
model <- glm(Churn ~ ., data = train, family = "binomial")
summary_result = summary(model)
variable_importance <- coef(model)

## [1] "Variable Importance:"

##                      (Intercept)                       genderMale
##                     3.3824915778                     0.0017140610
##                   SeniorCitizen1                       PartnerYes
##                     0.3538701494                    -0.0750944367
##                    DependentsYes                           tenure
##                    -0.1359025215                    -0.0517717300
##                  PhoneServiceYes    MultipleLinesNo phone service
##                     0.8873273054                               NA
##                 MultipleLinesYes       InternetServiceFiber optic
##                     0.6978161096                     2.8187324382
##                InternetServiceNo   OnlineSecurityNo internet service
##                    -2.8044995466                               NA
##                OnlineSecurityYes     OnlineBackupNo internet service
##                    -0.2001390877                               NA
##                  OnlineBackupYes DeviceProtectionNo internet service
##                     0.2995366198                               NA
##              DeviceProtectionYes      TechSupportNo internet service
##                     0.2368652780                               NA
```

```
##                         TechSupportYes       StreamingTVNo internet service
##                           0.2117756661                                   NA
##                         StreamingTVYes   StreamingMoviesNo internet service
##                           0.8600564606                                   NA
##                     StreamingMoviesYes                      ContractOne year
##                           1.1576584331                          -0.5847971593
##                       ContractTwo year                   PaperlessBillingYes
##                          -1.3946331334                           0.2728656120
## PaymentMethodCredit card (automatic)        PaymentMethodElectronic check
##                          -0.1219166281                           0.4004829216
##              PaymentMethodMailed check                       MonthlyCharges
##                          -0.1335196928                          -0.0826874979
##                           TotalCharges
##                           0.0002541057
```

After this model we store the features that are significant (p_value <0.05)

```
##                                    Estimate   Std. Error   z value     Pr(>|z|)
## (Intercept)                    3.3824915778 9.353906e-01  3.616127 2.990434e-04
## SeniorCitizen1                 0.3538701494 1.000596e-01  3.536594 4.053215e-04
## tenure                        -0.0517717300 5.780290e-03 -8.956598 3.348506e-19
## MultipleLinesYes               0.6978161096 2.030362e-01  3.436905 5.884025e-04
## InternetServiceFiber optic     2.8187324382 9.088388e-01  3.101466 1.925651e-03
## InternetServiceNo             -2.8044995466 9.151037e-01 -3.064679 2.179034e-03
## StreamingTVYes                 0.8600564606 3.671887e-01  2.342274 1.916665e-02
## StreamingMoviesYes             1.1576584331 3.722352e-01  3.110019 1.870753e-03
## ContractOne year              -0.5847971593 1.125780e-01 -5.194596 2.051645e-07
## ContractTwo year              -1.3946331334 1.697650e-01 -8.215082 2.120201e-16
## PaperlessBillingYes            0.2728656120 8.125386e-02  3.358186 7.845572e-04
## PaymentMethodElectronic check  0.4004829216 1.054777e-01  3.796850 1.465464e-04
## MonthlyCharges                -0.0826874979 3.610324e-02 -2.290307 2.200353e-02
## TotalCharges                   0.0002541057 6.861021e-05  3.703614 2.125496e-04
```

Keep only the features that are significant (also in test set)

```r
test <- test[, c('SeniorCitizen', 'Partner', 'tenure', 'MultipleLines',
'InternetService', 'OnlineBackup', 'StreamingTV', 'StreamingMovies', 'Contract',
'PaperlessBilling', 'PaymentMethod', 'MonthlyCharges', 'TotalCharges', 'Churn')]
```

Transformations

We have to asses if our numeric variables need to be transformed. Firstly, we check skewness and kurtosis

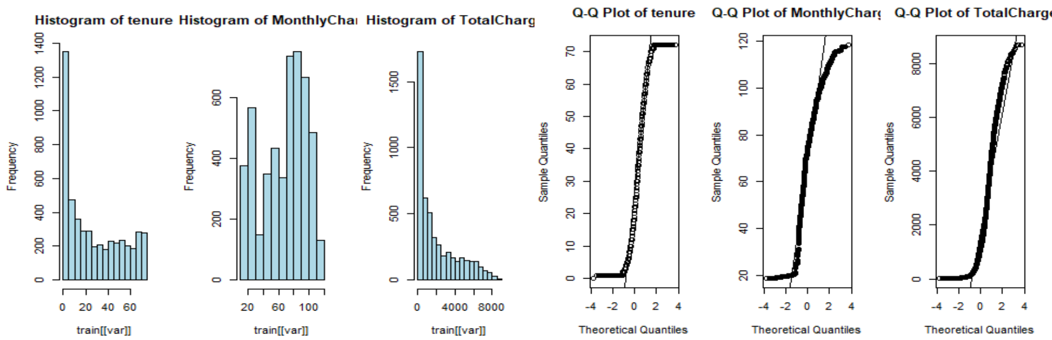
Tenure: The positively skewed distribution with a skewness of 0.50 suggests that most customers have shorter tenures. We could check different transformations because some might be beneficial.

MonthlyCharges: The slightly negatively skewed distribution with a skewness of -0.37 indicates a slight tail to the right. We might choose to transform this variable to achieve a more symmetric distribution.

TotalCharges: The positively skewed distribution with a skewness of 1.11 suggests a concentration of customers with lower total charges. The kurtosis of 3.12 indicates heavy tails. Transforming this variable (e.g., logarithmic transformation) might help mitigate the impact of extreme values and achieve a more symmetric distribution. By checking also the q-q plots it seems that we could try transformations mainly to MonthlyCharges and TotalCharges

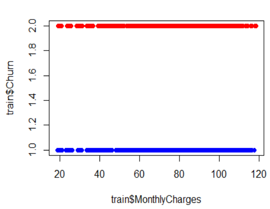
```r
skewness(train[, numeric_vars])

##         tenure MonthlyCharges   TotalCharges
##      0.5034220     -0.3689311      1.1078426

kurtosis(train[, numeric_vars])
```

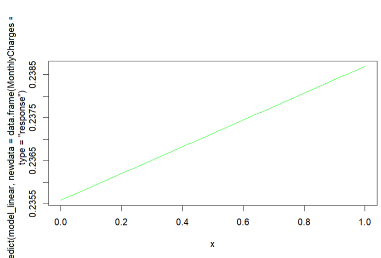
```
##          tenure MonthlyCharges   TotalCharges
##        1.814840       1.893477       3.119993
```

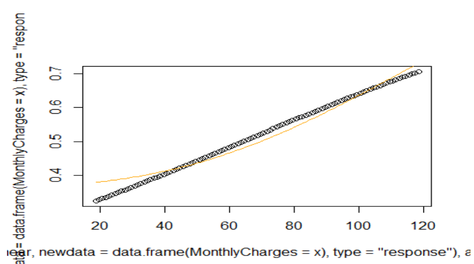


However, we will continue the assessment of possible evaluations. We are plotting the relationships of these numeric variables with Churn variable. Then we fit a glm model in order to plot a fitting logistic regression curve. If the curve is not capturing the underlying pattern well, it might suggest that a linear relationship between the numeric variable and the log-odds of churn is not appropriate. In such cases, we might consider transformations. Though as it can be seen in the plots this is not the case in any of the three variables. We also fit a quadratic model and plot it, to check linearity. Thus, we will not transform them, but in the next section that we are going to fit a linear model, different transformations of the variables will be tested (poly, log etc.) in order to check which one performs better.

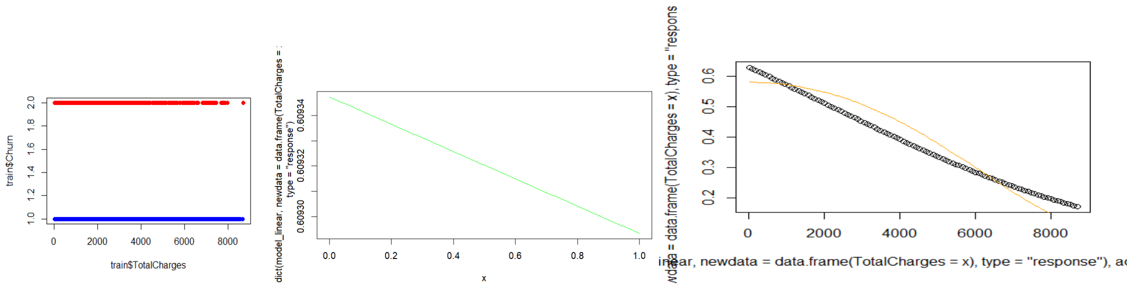
```r
# Plot the relationship
plot(train$MonthlyCharges, train$Churn, pch = 16, col = ifelse(train$Churn == "Yes",
"red", "blue"))

# Fit logistic regression without transformation
model_linear <- glm(Churn ~ MonthlyCharges, data = train, family = "binomial")

# Plot the logistic regression curve
plot(curve(predict(model_linear, newdata = data.frame(MonthlyCharges = x), type =
"response"), add = TRUE, col = "green"))

# Fit logistic regression with a quadratic transformation
train$MonthlyCharges_squared <- train$MonthlyCharges^2
model_quadratic <- glm(Churn ~ MonthlyCharges_squared, data = train, family =
"binomial")

# Plot the logistic regression curve with quadratic transformation
curve(predict(model_quadratic, newdata = data.frame(MonthlyCharges_squared = x^2), type
= "response"), add = TRUE, col = "orange")
```



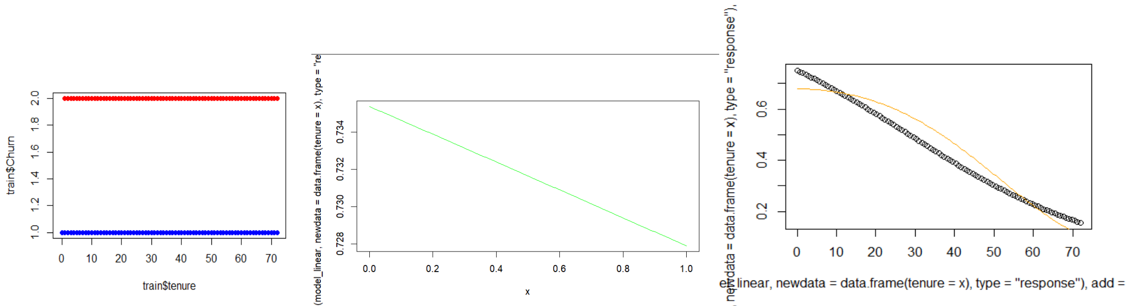




```
TotalCharges
```




```
tenure
```



Modeling using numeric variables

There is a function to evaluate training and testing in models that it is called multiple times

```r
evaluation <- function(model,train){
  # Predict probabilities on the training set
  train$predicted_prob <- predict(model, type = "response")

  # ROC curve and AUC for the training set
  roc_train <- roc(train$Churn, train$predicted_prob)
  auc_train <- auc(roc_train)

  # Confusion matrix for the training set
  threshold_train <- 0.5  # Adjust threshold as needed
  train$predicted_class <- ifelse(train$predicted_prob >= threshold_train, "Yes", "No")
  conf_matrix_train <- table(Actual = train$Churn, Predicted = train$predicted_class)

  # Calculate recall and F1-score for the training set
  recall_train <- conf_matrix_train["Yes", "Yes"] / sum(conf_matrix_train["Yes", ])
  precision_train <- conf_matrix_train["Yes", "Yes"] / sum(conf_matrix_train[, "Yes"])
  f1_score_train <- 2 * (precision_train * recall_train) / (precision_train +
recall_train)

  # Output results for the training set
  cat("Training Set Metrics:\n")
  cat("AUC:", auc_train, "\n")

  cat("ROC:", roc_train, "\n "))
  cat("Confusion Matrix:\n", conf_matrix_train, "\n")
  cat("Recall:", recall_train, "\n")
  cat("F1-score:", f1_score_train, "\n\n")


  # Predict probabilities on the test set
  test$predicted_prob <- predict(model, newdata = test, type = "response")

  # Confusion matrix for the test set
```

```r
  threshold_test <- 0.5  # Adjust threshold as needed
  test$predicted_class <- ifelse(test$predicted_prob >= threshold_test, "Yes", "No")
  conf_matrix_test <- table(Actual = test$Churn, Predicted = test$predicted_class)

  # Calculate recall for the test set
  recall_test <- conf_matrix_test["Yes", "Yes"] / sum(conf_matrix_test["Yes", ])
  precision_test <- conf_matrix_test["Yes", "Yes"] / sum(conf_matrix_test[, "Yes"])
  f1_score_test <- 2 * (precision_test * recall_test) / (precision_test + recall_test)

  # Output results for the test set
  cat("Test Set Metrics:\n")
  cat("Confusion Matrix:\n", conf_matrix_test, "\n")
  cat("Recall:", recall_test, "\n")
  cat("F1-score:", f1_score_test, "\n")
}
```

There were test around 20 different models only for numerical variables. We tried logarithmic, polynomial, quadratic, cubic transformations along with interaction terms or only by keeping the main effects, or 1 and 2 out of the 3 variables. All these models were compared with confusion matrices and ROC for training after using the function above. We keep at last the best one, but for space and reading complexity reasons we will not show all the metrics, but just the best one. The metrics of six models will be put in the annex. The F1-Score in the test set is around 61%, which is not that great, but we remind that the oversampling did not take place in the test set, thus the test set is imbalanced, and this has influence in the results. However, recall in test set is around 75% and the metrics for training are very good. Consequently, further modeling along with factor has to take place, so the results would be better and overfitting would might be further avoided.

```r
set.seed(123)
#model <- glm(Churn ~ MonthlyCharges + TotalCharges + tenure, data = train, family = 
"binomial")  # 13
#model <- glm(Churn ~ MonthlyCharges + I(MonthlyCharges^2) + TotalCharges + 
I(TotalCharges^2) + tenure, data = train, family = "binomial")  #15
#model <- glm(Churn ~ I(MonthlyCharges^2) + I(TotalCharges^2) + tenure, data = train, 
family = "binomial") # 9
#model <- glm(Churn ~ I(MonthlyCharges^2) + TotalCharges + tenure, data = train, family 
= "binomial") # 4
#model <- glm(Churn ~ I(MonthlyCharges^2) + I(TotalCharges^2) + I(tenure^2), data = 
train, family = "binomial") # 18
#model <- glm(Churn ~ MonthlyCharges * TotalCharges + tenure, data = train, family = 
"binomial") # 12
#model <- glm(Churn ~ MonthlyCharges + TotalCharges*tenure, data = train, family = 
"binomial") # 17
#model <- glm(Churn ~ MonthlyCharges*tenure + TotalCharges, data = train, family = 
"binomial") # 10

#TAKE ONE OUT MODELS

#model <- glm(Churn ~ MonthlyCharges + TotalCharges, data = train, family = "binomial") 
#21
#model <- glm(Churn ~ tenure + TotalCharges, data = train, family = "binomial")  #22
#model <- glm(Churn ~ MonthlyCharges + tenure, data = train, family = "binomial")   #16

#LOG
#small_const <- 1e-5
#model <- glm(Churn ~ log(MonthlyCharges + small_const) + log(TotalCharges + 
small_const) + log(tenure + small_const), data = train, family = "binomial")   #3

#poly

#model <- glm(Churn ~ poly(MonthlyCharges,2) + poly(TotalCharges,2) + poly(tenure,2), 
data = train, family = "binomial")  # 6
#model <- glm(Churn ~ poly(MonthlyCharges,3) + poly(TotalCharges,3) + poly(tenure,3), 
```

```r
data = train, family = "binomial") # 7
#model <- glm(Churn ~ poly(MonthlyCharges,3) + poly(TotalCharges,3) + poly(tenure,2),
data = train, family = "binomial")   #8
#model <- glm(Churn ~ poly(MonthlyCharges,3) + poly(TotalCharges,2) + poly(tenure,3),
data = train, family = "binomial")  # 2
#model <- glm(Churn ~ poly(MonthlyCharges,2) + poly(TotalCharges,3) + poly(tenure,3),
data = train, family = "binomial")  # 4
model <- glm(Churn ~ poly(MonthlyCharges,2) + poly(TotalCharges,2) + poly(tenure,3),
data = train, family = "binomial")  # 1 BEST


#I^3
#model <- glm(Churn ~ I(MonthlyCharges^3) + I(TotalCharges^3) + I(tenure^3), data =
train, family = "binomial") #20
#model <- glm(Churn ~ I(MonthlyCharges^3) + I(TotalCharges^3) + I(tenure^2), data =
train, family = "binomial") #11
#model <- glm(Churn ~ I(MonthlyCharges^3) + I(TotalCharges^2) + I(tenure^2), data =
train, family = "binomial") #14
#model <- glm(Churn ~ I(MonthlyCharges^2) + I(TotalCharges^3) + I(tenure^2), data =
train, family = "binomial")  #19

evaluation(model,train)
```

```
Training Set Metrics:

AUC: 0.8135801

Confusion Matrix:

 1853 654 667 1761

Recall: 0.7291925

F1-score: 0.7272352


Test Set Metrics:

AUC: 0.8207639

Confusion Matrix:

 1114 146 408 434

Recall: 0.7482759

F1-score: 0.6104079
```

ANNEX

RESULTS OF CATDES FOR CHURN PROFILING

```
res.cat <-catdes(train, num.var=20)
res.cat$test.chi

##                       p.value df
## Contract         2.291710e-278  2
## OnlineSecurity   4.650500e-195  2
## TechSupport      2.233583e-174  2
## InternetService  1.064039e-156  2
## PaymentMethod    1.183206e-143  3
## DeviceProtection 2.225330e-131  2
## OnlineBackup     1.040239e-130  2
## StreamingTV       4.627008e-91  2
## StreamingMovies   3.048941e-90  2
## Dependents        1.347363e-53  1
## PaperlessBilling  1.671381e-53  1
## Partner           2.371358e-43  1
## SeniorCitizen     4.951196e-36  1
## MultipleLines     2.717989e-04  2

res.cat$quanti

## $No
##                   v.test Mean in category Overall mean sd in category
## tenure          29.53497         38.01655     27.60693       24.35900
## TotalCharges    17.41002       2553.16447   2005.94729     2314.42627
## MonthlyCharges -15.85219         60.67956     67.33133       30.94244
##                Overall sd       p.value
## tenure           24.18307 1.024264e-191
## TotalCharges   2156.61992  6.926568e-68
## MonthlyCharges   28.79128  1.357604e-56
##
## $Yes
##                   v.test Mean in category Overall mean sd in category
## MonthlyCharges  15.85219         73.63864     67.33133       25.00717
## TotalCharges   -17.41002       1487.06775   2005.94729     1851.70508
## tenure         -29.53497         17.73637     27.60693       19.40480
##                Overall sd       p.value
## MonthlyCharges   28.79128  1.357604e-56
## TotalCharges   2156.61992  6.926568e-68
## tenure           24.18307 1.024264e-191

res.cat$quanti.var

##                      Eta2       P-value
## tenure         0.17569276 1.460165e-210
## TotalCharges   0.06104907  5.747761e-70
## MonthlyCharges 0.05061269  5.185864e-58

res.cat$category

## $No
##                                       Cla/Mod  Mod/Cla   Global
## Contract=Two year                    92.63521 33.30575 17.49899
## StreamingMovies=No internet service  80.83028 27.38933 16.49215
## StreamingTV=No internet service      80.83028 27.38933 16.49215
## TechSupport=No internet service      80.83028 27.38933 16.49215
## DeviceProtection=No internet service 80.83028 27.38933 16.49215
```

```
## OnlineBackup=No internet service        80.83028 27.38933 16.49215
## OnlineSecurity=No internet service      80.83028 27.38933 16.49215
## InternetService=No                      80.83028 27.38933 16.49215
## OnlineSecurity=Yes                      69.51736 33.96773 23.78172
## Dependents=Yes                          67.10526 35.87091 26.01692
## PaperlessBilling=No                     62.90060 47.91063 37.07209
## Contract=One year                       72.62905 25.03103 16.77406
## TechSupport=Yes                         65.17375 34.91932 26.07733
## Partner=Yes                             59.29010 55.97849 45.95248
## PaymentMethod=Credit card (automatic)   67.74194 25.19652 18.10310
## SeniorCitizen=0                         53.00274 88.00165 80.80950
## PaymentMethod=Bank transfer (automatic) 64.80086 24.90691 18.70721
## InternetService=DSL                     59.52824 39.67729 32.44060
## PaymentMethod=Mailed check              60.73858 25.85850 20.72090
## DeviceProtection=Yes                    56.11602 36.82251 31.93717
## OnlineBackup=Yes                        55.64369 37.73273 33.00443
## MultipleLines=No                        51.47373 49.85519 47.14056
## MultipleLines=Yes                       45.39819 39.38767 42.22714
## StreamingTV=Yes                         44.91569 36.36740 39.40797
## StreamingMovies=Yes                     44.30693 37.02938 40.67660
## StreamingMovies=No                      40.43253 35.58130 42.83125
## StreamingTV=No                          40.00000 36.24328 44.09988
## SeniorCitizen=1                         30.43022 11.99835 19.19050
## Partner=No                              39.64232 44.02151 54.04752
## PaperlessBilling=Yes                    40.28800 52.08937 62.92791
## Dependents=No                           42.18835 64.12909 73.98308
## OnlineBackup=No                         33.61244 34.87795 50.50342
## DeviceProtection=No                     33.77587 35.78817 51.57068
## InternetService=Fiber optic             31.38801 32.93339 51.06726
## PaymentMethod=Electronic check          27.54860 24.03806 42.46879
## TechSupport=No                          31.94250 37.69135 57.43053
## OnlineSecurity=No                       31.49022 38.64295 59.72614
## Contract=Month-to-month                 30.85172 41.66322 65.72694
##                                               p.value    v.test
## Contract=Two year                       3.302328e-204  30.492241
## StreamingMovies=No internet service      2.802028e-95  20.710215
## StreamingTV=No internet service          2.802028e-95  20.710215
## TechSupport=No internet service          2.802028e-95  20.710215
## DeviceProtection=No internet service     2.802028e-95  20.710215
## OnlineBackup=No internet service         2.802028e-95  20.710215
## OnlineSecurity=No internet service       2.802028e-95  20.710215
## InternetService=No                       2.802028e-95  20.710215
## OnlineSecurity=Yes                       1.326697e-61  16.561312
## Dependents=Yes                           2.942761e-54  15.510565
## PaperlessBilling=No                      6.876683e-54  15.455972
## Contract=One year                        3.017374e-53  15.360394
## TechSupport=Yes                          7.841082e-44  13.884716
## Partner=Yes                              1.432931e-43  13.841447
## PaymentMethod=Credit card (automatic)    4.005106e-37  12.730453
## SeniorCitizen=0                          7.958380e-37  12.676729
## PaymentMethod=Bank transfer (automatic)  6.137566e-28  10.957190
## InternetService=DSL                      2.322336e-26  10.623288
## PaymentMethod=Mailed check               2.950281e-18   8.713348
## DeviceProtection=Yes                     6.391286e-13   7.191864
## OnlineBackup=Yes                         5.166961e-12   6.900912
## MultipleLines=No                         1.908308e-04   3.730855
## MultipleLines=Yes                        7.982639e-05  -3.944921
## StreamingTV=Yes                          1.950639e-05  -4.270467
## StreamingMovies=Yes                      3.461636e-07  -5.096405
## StreamingMovies=No                       7.350194e-24 -10.071955
## StreamingTV=No                           1.430021e-27 -10.880357
```

```
## SeniorCitizen=1                           7.958380e-37 -12.676729
## Partner=No                                1.432931e-43 -13.841447
## PaperlessBilling=Yes                      6.876683e-54 -15.455972
## Dependents=No                             2.942761e-54 -15.510565
## OnlineBackup=No                          1.529950e-103 -21.607409
## DeviceProtection=No                      1.058880e-105 -21.835878
## InternetService=Fiber optic              1.367645e-139 -25.151301
## PaymentMethod=Electronic check           2.889584e-148 -25.932170
## TechSupport=No                           1.615015e-169 -27.752734
## OnlineSecurity=No                        1.895168e-197 -29.978078
## Contract=Month-to-month                  1.490170e-281 -35.855863
##
## $Yes
##                                            Cla/Mod   Mod/Cla   Global
## Contract=Month-to-month                  69.148284 88.544527 65.72694
## OnlineSecurity=No                        68.509777 79.717536 59.72614
## TechSupport=No                           68.057504 76.147509 57.43053
## PaymentMethod=Electronic check           72.451399 59.945077 42.46879
## InternetService=Fiber optic              68.611987 68.262064 51.06726
## DeviceProtection=No                      66.224131 66.535896 51.57068
## OnlineBackup=No                          66.387560 65.319733 50.50342
## Dependents=No                            57.811649 83.326795 73.98308
## PaperlessBilling=Yes                     59.712000 73.205179 62.92791
## Partner=No                               60.357675 63.554335 54.04752
## SeniorCitizen=1                          69.569780 26.010200 19.19050
## StreamingTV=No                           60.000000 51.549627 44.09988
## StreamingMovies=No                       59.567466 49.705767 42.83125
## StreamingMovies=Yes                      55.693069 44.134955 40.67660
## StreamingTV=Yes                          55.084313 42.291095 39.40797
## MultipleLines=Yes                        54.601812 44.919576 42.22714
## MultipleLines=No                         48.526271 44.566497 47.14056
## OnlineBackup=Yes                         44.356315 28.520989 33.00443
## DeviceProtection=Yes                     43.883985 27.304825 31.93717
## PaymentMethod=Mailed check               39.261419 15.849353 20.72090
## InternetService=DSL                      40.471757 25.578658 32.44060
## PaymentMethod=Bank transfer (automatic) 35.199139 12.828560 18.70721
## SeniorCitizen=0                          46.997259 73.989800 80.80950
## PaymentMethod=Credit card (automatic)   32.258065 11.377011 18.10310
## Partner=Yes                              40.709904 36.445665 45.95248
## TechSupport=Yes                          34.826255 17.693213 26.07733
## Contract=One year                        27.370948  8.944684 16.77406
## PaperlessBilling=No                      37.099402 26.794821 37.07209
## Dependents=Yes                           32.894737 16.673205 26.01692
## OnlineSecurity=Yes                       30.482642 14.123186 23.78172
## StreamingMovies=No internet service      19.169719  6.159278 16.49215
## StreamingTV=No internet service          19.169719  6.159278 16.49215
## TechSupport=No internet service          19.169719  6.159278 16.49215
## DeviceProtection=No internet service     19.169719  6.159278 16.49215
## OnlineBackup=No internet service         19.169719  6.159278 16.49215
## OnlineSecurity=No internet service       19.169719  6.159278 16.49215
## InternetService=No                       19.169719  6.159278 16.49215
## Contract=Two year                         7.364787  2.510789 17.49899
##                                                p.value     v.test
## Contract=Month-to-month                  1.490170e-281  35.855863
## OnlineSecurity=No                        1.895168e-197  29.978078
## TechSupport=No                           1.615015e-169  27.752734
## PaymentMethod=Electronic check           2.889584e-148  25.932170
## InternetService=Fiber optic              1.367645e-139  25.151301
## DeviceProtection=No                      1.058880e-105  21.835878
## OnlineBackup=No                          1.529950e-103  21.607409
## Dependents=No                             2.942761e-54  15.510565
```

```
## PaperlessBilling=Yes                      6.876683e-54  15.455972
## Partner=No                                1.432931e-43  13.841447
## SeniorCitizen=1                           7.958380e-37  12.676729
## StreamingTV=No                            1.430021e-27  10.880357
## StreamingMovies=No                        7.350194e-24  10.071955
## StreamingMovies=Yes                       3.461636e-07   5.096405
## StreamingTV=Yes                           1.950639e-05   4.270467
## MultipleLines=Yes                         7.982639e-05   3.944921
## MultipleLines=No                          1.908308e-04  -3.730855
## OnlineBackup=Yes                          5.166961e-12  -6.900912
## DeviceProtection=Yes                      6.391286e-13  -7.191864
## PaymentMethod=Mailed check                2.950281e-18  -8.713348
## InternetService=DSL                       2.322336e-26 -10.623288
## PaymentMethod=Bank transfer (automatic)  6.137566e-28 -10.957190
## SeniorCitizen=0                           7.958380e-37 -12.676729
## PaymentMethod=Credit card (automatic)    4.005106e-37 -12.730453
## Partner=Yes                               1.432931e-43 -13.841447
## TechSupport=Yes                           7.841082e-44 -13.884716
## Contract=One year                         3.017374e-53 -15.360394
## PaperlessBilling=No                       6.876683e-54 -15.455972
## Dependents=Yes                            2.942761e-54 -15.510565
## OnlineSecurity=Yes                        1.326697e-61 -16.561312
## StreamingMovies=No internet service       2.802028e-95 -20.710215
## StreamingTV=No internet service           2.802028e-95 -20.710215
## TechSupport=No internet service           2.802028e-95 -20.710215
## DeviceProtection=No internet service      2.802028e-95 -20.710215
## OnlineBackup=No internet service          2.802028e-95 -20.710215
## OnlineSecurity=No internet service        2.802028e-95 -20.710215
## InternetService=No                        2.802028e-95 -20.710215
## Contract=Two year                        3.302328e-204 -30.492241
```

model <- glm(Churn ~ poly(MonthlyCharges,2) + poly(TotalCharges,2) + poly(tenure,2), data = train, family = "binomial")

Training Set Metrics:

AUC: 0.8104709

Confusion Matrix:

1887 675 633 1740

Recall: 0.7204969

F1-score: 0.726817

Test Set Metrics:

AUC: 0.8150953

Confusion Matrix:

1127 164 395 416

Recall: 0.7172414

F1-score: 0.5981308

```
model <- glm(Churn ~ poly(MonthlyCharges,3) + poly(TotalCharges,3) + poly(tenure,3), data = train, family = "binomial")
```

Training Set Metrics:

AUC: 0.8140574

Confusion Matrix:

1869 694 651 1721

Recall: 0.7126294

F1-score: 0.7190307

Test Set Metrics:

AUC: 0.8214968

Confusion Matrix:

1127 156 395 424

Recall: 0.7310345

F1-score: 0.6061472

```
model <- glm(Churn ~ poly(MonthlyCharges,3) + poly(TotalCharges,3) + poly(tenure,2), data = train, family = "binomial")
```

Training Set Metrics:

AUC: 0.8124606

Confusion Matrix:

1863 687 657 1728

Recall: 0.715528

F1-score: 0.72

Test Set Metrics:

AUC: 0.8186149

Confusion Matrix:

1121 160 401 420

Recall: 0.7241379

F1-score: 0.5995717

```
model <- glm(Churn ~ poly(MonthlyCharges,2) + poly(TotalCharges,3) + poly(tenure,3), data = train, family = "binomial")
```

Training Set Metrics:

AUC: 0.8139946

Confusion Matrix:

1866 684 654 1731

Recall: 0.7167702

F1-score: 0.72125

Test Set Metrics:

AUC: 0.8215546

Confusion Matrix:

1124 155 398 425

Recall: 0.7327586

F1-score: 0.6058446

```
model <- glm(Churn ~ poly(MonthlyCharges,2) + poly(TotalCharges,2) + poly(tenure,3), data = train, family = "binomial")
```

Training Set Metrics:

AUC: 0.8135801

Confusion Matrix:

1853 654 667 1761

Recall: 0.7291925

F1-score: 0.7272352

Test Set Metrics:

AUC: 0.8207639

Confusion Matrix:

1114 146 408 434

Recall: 0.7482759

F1-score: 0.6104079

```
model <- glm(Churn ~ poly(MonthlyCharges,3) + poly(TotalCharges,2) + poly(tenure,3), data = train, family = "binomial")
```

Training Set Metrics:

AUC: 0.8134064

Confusion Matrix:

1857 676 663 1739

Recall: 0.7200828

F1-score: 0.7220262

Test Set Metrics:

AUC: 0.8206993

Confusion Matrix:

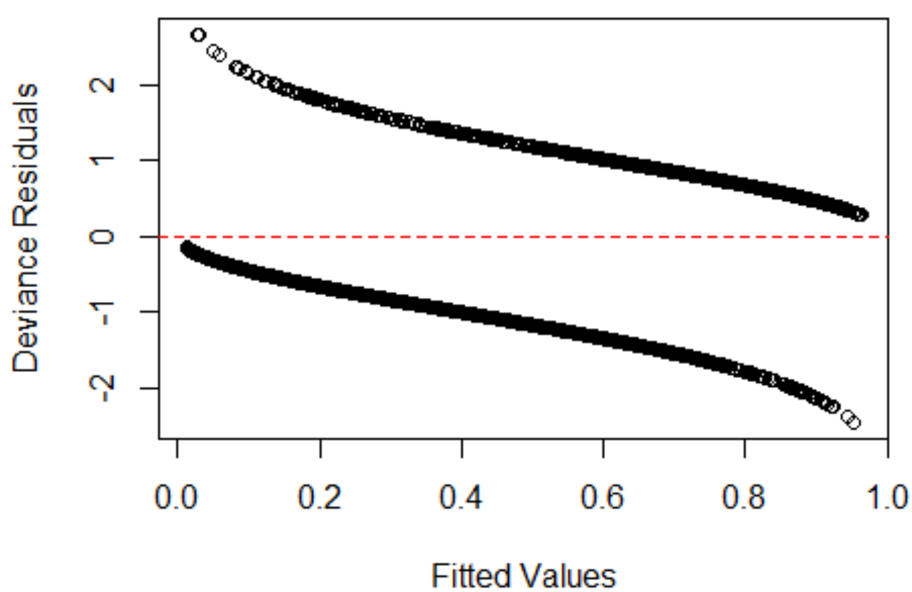
1132 155 390 425

Recall: 0.7327586

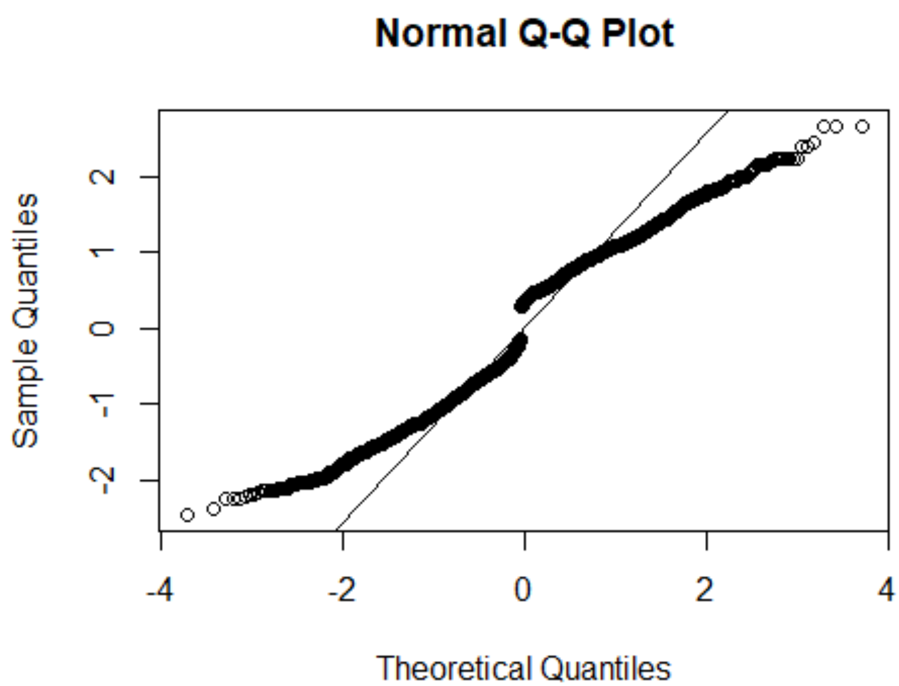
F1-score: 0.609319

Residual Analysis of Numeric Model

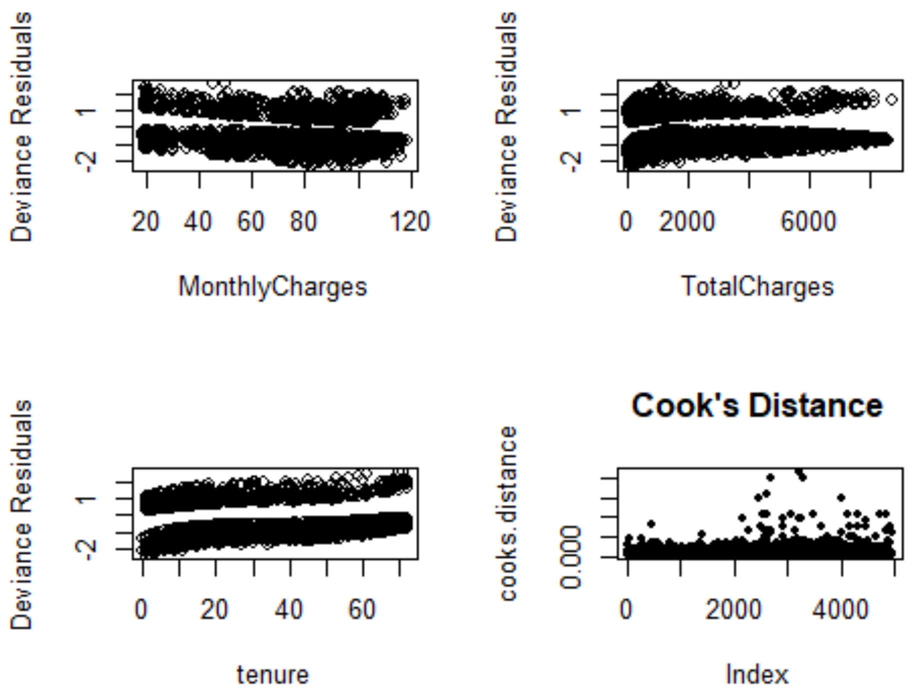
```
residuals <- residuals(model, type = "deviance")
#For checking heteroscedasticity
plot(fitted(model), residuals, ylab = "Deviance Residuals", xlab =
"Fitted Values")
abline(h = 0, col = "red", lty = 2)
```



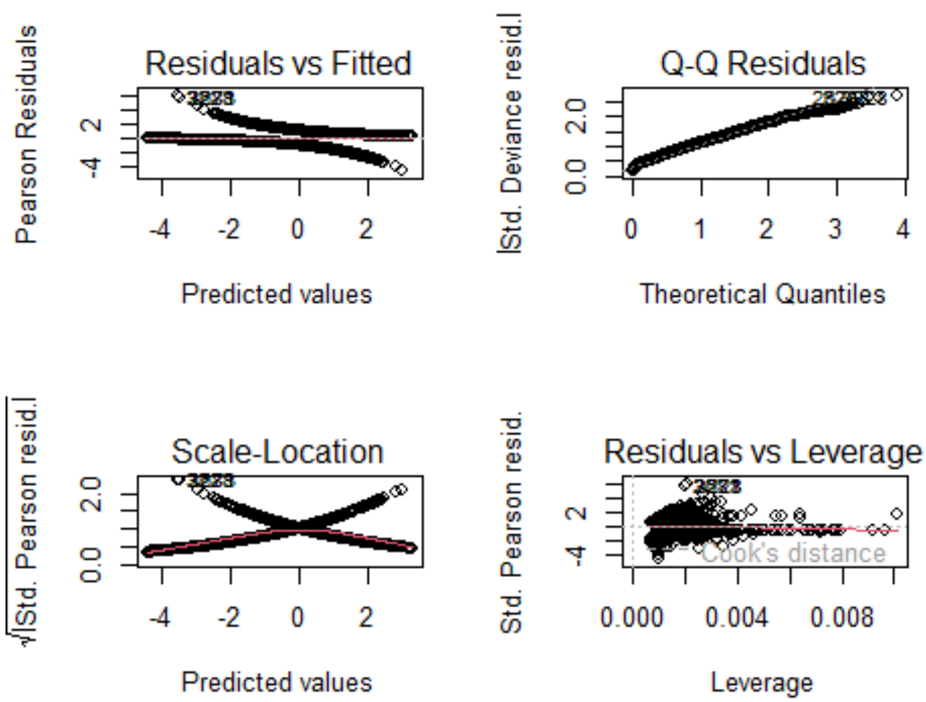
```
qqnorm(residuals)
qqline(residuals)
```



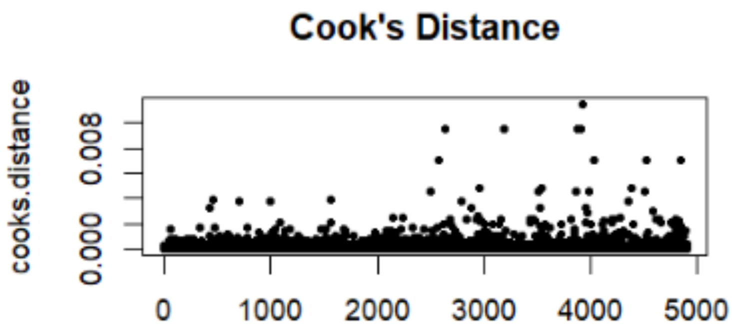

```
par(mfrow = c(2, 2))
plot(train$MonthlyCharges, residuals, ylab = "Deviance Residuals", xlab
= "MonthlyCharges")
plot(train$TotalCharges, residuals, ylab = "Deviance Residuals", xlab =
"TotalCharges")
plot(train$tenure, residuals, ylab = "Deviance Residuals", xlab =
"tenure")
cooks.distance <- cooks.distance(model)
plot(cooks.distance, pch = 20, main = "Cook's Distance")
```


Deviance Residuals
1
-2
20 40 60 80 120
MonthlyCharges
Deviance Residuals
1
-2
0 2000 6000
TotalCharges
Deviance Residuals
1
-2
0 20 40 60
tenure
Cook's Distance
cooks.distance
0.000
0 2000 4000
Index

```r
plot(model)
```


Residuals vs Fitted
Pearson Residuals
2
-4
-4 -2 0 2
Predicted values
Q-Q Residuals
|Std. Deviance resid.|
2.0
0.0
0 1 2 3 4
Theoretical Quantiles
Scale-Location
√|Std. Pearson resid.|
2.0
0.0
-4 -2 0 2
Predicted values
Residuals vs Leverage
Std. Pearson resid.
2
-4
Cook's distance
0.000 0.004 0.008
Leverage

Most of the points have a Cook's distance under .001, so to remove outliers from the model we remove data points with a Cook's Distance greater than .001.




```
cooksd <- cooks.distance(model)
outliers <- which(cooksd > 0.001)
train_clean <- train[-outliers, ]
```

```
# Model before removing outliers (training set):
# Confusion Matrix: 1739 540 700 2004  / AUC: 0.8220867  / F1-score:
0.7637195

# Check model after removing outliers
model_cleaned <- glm(Churn ~ poly(MonthlyCharges,2) +
poly(TotalCharges,2) + poly(tenure,3), data = train_clean, family =
"binomial")
evaluation(model_cleaned,train)



Setting levels: control = No, case = Yes

Setting direction: controls < cases

Training Set Metrics:

AUC: 0.8212205

[1] "Confusion Matrix:"

      Predicted

Actual    No  Yes

   No  1755  736

   Yes  543 1941

Recall: 0.781401
```

```
F1-score: 0.7521798


Setting levels: control = No, case = Yes

Setting direction: controls < cases

Test Set Metrics:

AUC: 0.8201131

[1] "Confusion Matrix:"

      Predicted

Actual   No  Yes

   No  1064  439

   Yes  127  423

Recall: 0.7690909

F1-score: 0.5991501
```

The purpose of this residual analysis for the numeric model is to provide insights into its performance and identified potential areas for improvement. We created diagnostic plots, including fitted vs. residuals and Q-Q plots. Most helpful was was the plot of Cook's distance, which facilitated identifying a threshold for outliers, which we set at 0.001. The subsequent removal of data points exceeding this threshold aimed to improve the model's reliability by reducing the impact of influential observations on our model.

Adding factors to the numeric model

```r
library(caret)

# Function designed to test which combination of factors leads to the
highest F-Score.
find_best_model <- function(data) {
    formula_template <- as.formula("Churn ~ poly(MonthlyCharges,2) +
poly(TotalCharges,2) + poly(tenure,3)")

  # List of factors to consider
  factors <- c("SeniorCitizen", "Partner",
"MultipleLines","InternetService",
               "OnlineBackup", "StreamingTV", "StreamingMovies",
"Contract",
               "PaperlessBilling", "PaymentMethod")

  # Initialize variables to store the best model and F-score
```

```r
  best_model <- NULL
  best_f_score <- 0

  # Iterate through all possible combinations of factors
  for (i in 0:(2^length(factors) - 1)) {
    # Create a binary vector indicating which factors to include
    factor_subset <- as.logical(intToBits(i))[1:length(factors)]

    # Extract selected factors
    selected_factors <- factors[factor_subset]

    # Construct the formula
    formula <- reformulate(c("poly(MonthlyCharges,2)",
"poly(TotalCharges,2)", "poly(tenure,3)", selected_factors), response =
"Churn")

    # Fit the logistic regression model
    model <- glm(formula, data = data, family = "binomial")

    # Make predictions on the training set
    predictions <- predict(model, newdata = data, type = "response")

    # Make binary predictions
    binary_predictions <- ifelse(predictions > 0.5, 1, 0)

    confusion_matrix <- table(data$Churn, binary_predictions)

    precision <- confusion_matrix[2, 2] / sum(confusion_matrix[, 2])
    recall <- confusion_matrix[2, 2] / sum(confusion_matrix[2, ])
    f_score <- 2 * (precision * recall) / (precision + recall)

    # Update the best model if the F-score is higher
    if (f_score > best_f_score) {
      best_model <- model
      best_f_score <- f_score
    }
  }

  # Return the best model and its F-score
  return(list(model = best_model, f_score = best_f_score))
}

result <- find_best_model(train)

best_model <- result$model
best_f_score <- result$f_score


print(best_model)
```

```
##
## Call:  glm(formula = formula, family = "binomial", data = data)
##
## Coefficients:
##                         (Intercept)
poly(MonthlyCharges, 2)1
##                            -0.51322                               -
5.09857
##             poly(MonthlyCharges, 2)2
poly(TotalCharges, 2)1
##                             6.82500                               -
32.27251
##               poly(TotalCharges, 2)2
poly(tenure, 3)1
##                            11.87311                               -
24.49113
##                     poly(tenure, 3)2
poly(tenure, 3)3
##                             7.79648                               -
15.78657
##                       SeniorCitizen1
PartnerYes
##                             0.32461                               -
0.19053
##          InternetServiceFiber optic
InternetServiceNo
##                             1.13563                               -
0.90240
##       OnlineBackupNo internet service
OnlineBackupYes
##                                  NA                               -
0.17730
##        StreamingTVNo internet service
StreamingTVYes
##                                  NA
0.42654
##   StreamingMoviesNo internet service
StreamingMoviesYes
##                                  NA
0.54736
##                     ContractOne year
ContractTwo year
##                            -0.89791                               -
2.00175
##                  PaperlessBillingYes  PaymentMethodCredit card
(automatic)
##                             0.31065                               -
0.18911
##        PaymentMethodElectronic check             PaymentMethodMailed
check
```

```
##                              0.10748                              -
0.04075
##
## Degrees of Freedom: 4942 Total (i.e. Null);  4922 Residual
## Null Deviance:       6850
## Residual Deviance: 4757  AIC: 4799
```

```
print(paste("Best F-score:", best_f_score))
```

```
## [1] "Best F-score: 0.788732394366197"
```

```
#Best model according to F-Score

f_model <- glm(Churn ~ poly(MonthlyCharges, 2) + poly(TotalCharges, 2)
+ poly(tenure, 3) + SeniorCitizen + InternetService + StreamingTV +
StreamingMovies + Contract, data = train, family = "binomial")

# f_model <- glm(Churn ~ poly(MonthlyCharges, 2) + poly(TotalCharges,
2) + poly(tenure, 3) + SeniorCitizen + InternetService + OnlineBackup +
StreamingTV + StreamingMovies + Contract + PaperlessBilling +
PaymentMethod, data = train, family = "binomial")

evaluation(f_model, train)
```

```
## Warning in predict.lm(object, newdata, se.fit, scale = 1, type = if
(type == :
## prediction from rank-deficient fit; attr(*, "non-estim") has
doubtful cases
```

```
## Training Set Metrics:
## AUC: 0.8540055
## [1] "Confusion Matrix:"
##       Predicted
## Actual   No  Yes
##    No  1784  637
##    Yes  478 2053
## Recall: 0.8111418
## F1-score: 0.7864394
```

```
## Warning in predict.lm(object, newdata, se.fit, scale = 1, type = if
(type == :
## prediction from rank-deficient fit; attr(*, "non-estim") has
doubtful cases
```

```
## Setting levels: control = No, case = Yes
## Setting direction: controls < cases
```

```
## Test Set Metrics:
## AUC: 0.8375479
## [1] "Confusion Matrix:"
##       Predicted
```

```
## Actual   No  Yes
##    No  1119  424
##    Yes  107  434
## Recall: 0.8022181
## F1-score: 0.6204432
```

In this section we extended the numeric model by adding a select group of factors. A function was created to search for a combination of factors to maximize the F-score. The selected factors were added to the model, and an evaluation was done, which showed a slight improvement in the model for both the training and the test data set.

ANOVA test on variables

```
# Perform ANOVA on the categorical variables to check for significance
anova_results <- anova(f_model, test = "Chisq")

# Display the ANOVA results
print(anova_results)
```

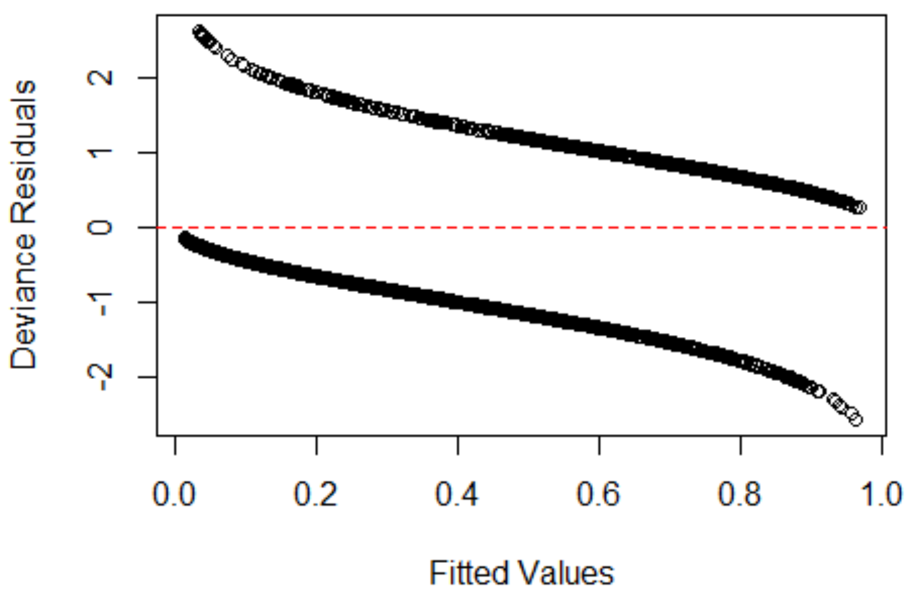
```
## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
## Response: Churn
##
## Terms added sequentially (first to last)
##
##
##                           Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
## NULL                                       4942     6850.0
## poly(MonthlyCharges, 2)  2   330.18      4940     6519.9 < 2.2e-16
***
## poly(TotalCharges, 2)    2  1175.99      4938     5343.9 < 2.2e-16
***
## poly(tenure, 3)          3   170.94      4935     5172.9 < 2.2e-16
***
## SeniorCitizen            1    53.98      4934     5119.0 2.022e-13
***
## InternetService          2    88.13      4932     5030.8 < 2.2e-16
***
## StreamingTV              1    22.12      4931     5008.7 2.562e-06
***
## StreamingMovies          1    36.07      4930     4972.6 1.899e-09
***
## Contract                 2   184.18      4928     4788.5 < 2.2e-16
***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# All variables are significant predictors
```

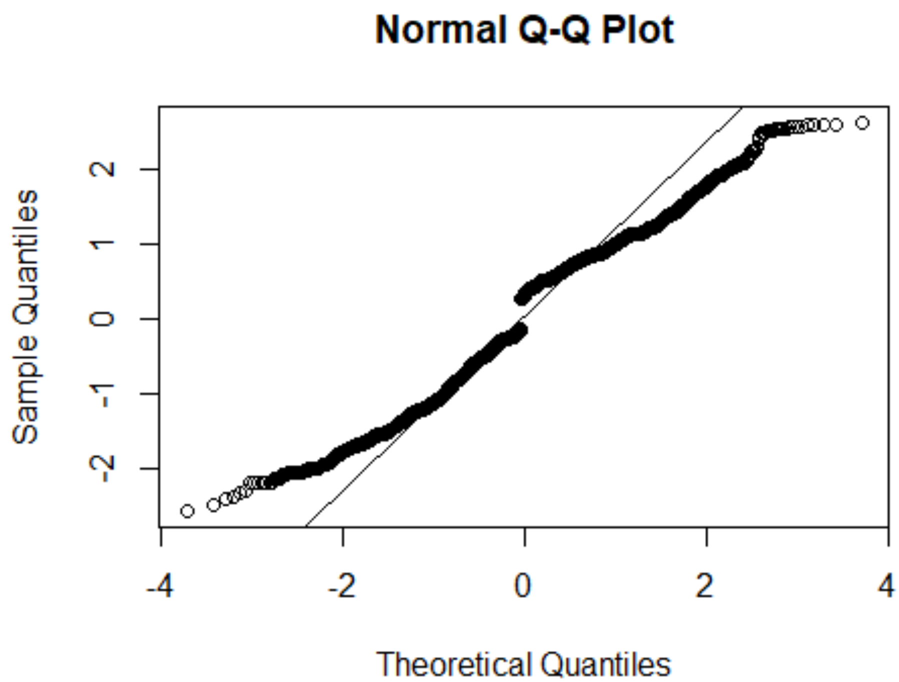
We then performed an ANOVA test to check the significance of these factors. Fortunately, all factors were found to be significant, thus none had to be removed.

Residual Analysis of Numeric + Factor Model

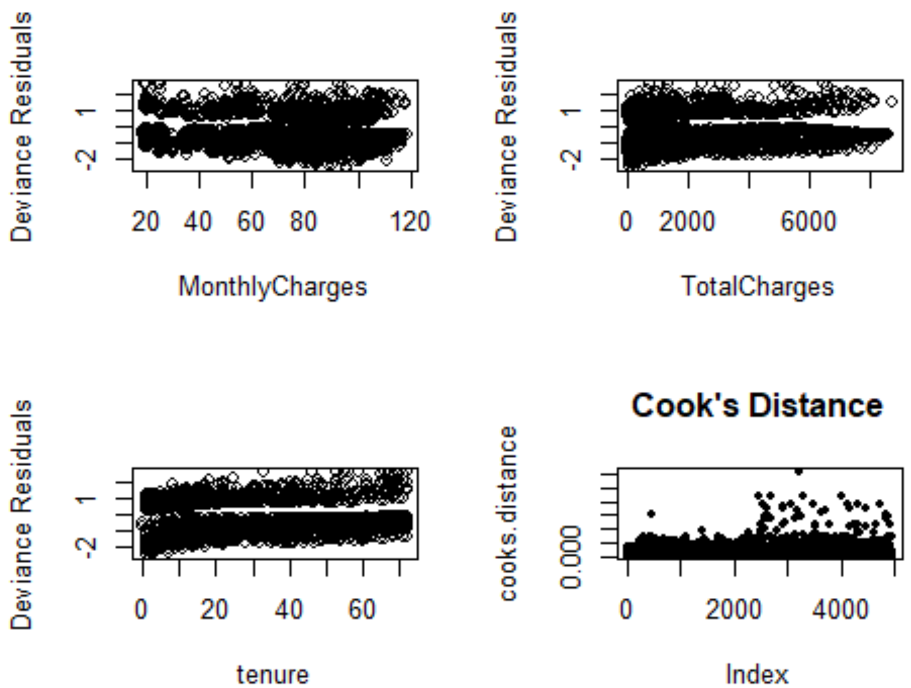
```
residuals2 <- residuals(f_model, type = "deviance")
#For checking heteroscedasticity
plot(fitted(f_model), residuals2, ylab = "Deviance Residuals", xlab =
"Fitted Values")
abline(h = 0, col = "red", lty = 2)
```



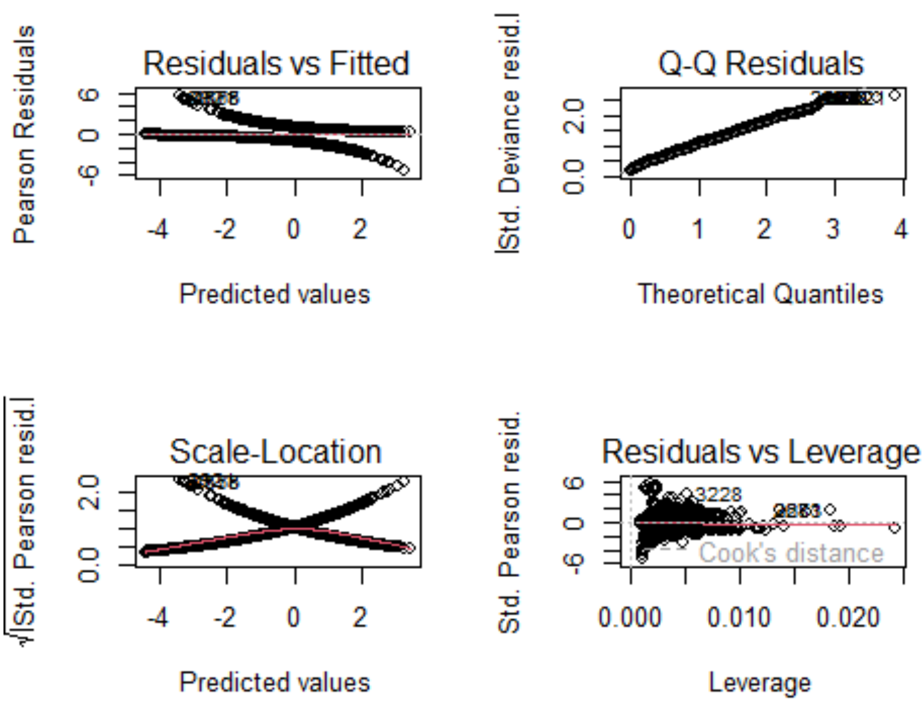

```
qqnorm(residuals2)
qqline(residuals2)
```




```r
par(mfrow = c(2, 2))
plot(train$MonthlyCharges, residuals2, ylab = "Deviance Residuals",
xlab = "MonthlyCharges")
plot(train$TotalCharges, residuals2, ylab = "Deviance Residuals", xlab
= "TotalCharges")
plot(train$tenure, residuals2, ylab = "Deviance Residuals", xlab =
"tenure")
cooks.distance <- cooks.distance(f_model)
plot(cooks.distance, pch = 20, main = "Cook's Distance")
```




```
plot(f_model)
```




```
# Now the normal Q-Q Plot does not follow such a linear pattern
```

This second residual analysis also aims to provide insights into its performance and identified potential areas for improvement. We created diagnostic plots, including fitted vs. residuals and Q-Q plots. Most helpful was again the plot of Cook's distance, which facilitated identifying a threshold for outliers, which we set at 0.002.

Remove data points with a Cook's Distance greater than .002

```
cooksd2 <- cooks.distance(f_model)
outliers2 <- which(cooksd2 > 0.002)
train_clean2 <- train[-outliers2, ]


# Create model without  outliers
# f_model_cleaned <- glm(Churn ~ poly(MonthlyCharges, 2) +
poly(TotalCharges, 2) + poly(tenure, 3) + SeniorCitizen +
InternetService + OnlineBackup + StreamingTV + StreamingMovies +
Contract + PaperlessBilling + PaymentMethod, data = train_clean2,
family = "binomial")

f_model_cleaned <- glm(Churn ~ poly(MonthlyCharges, 2) +
poly(TotalCharges, 2) + poly(tenure, 3) + SeniorCitizen +
InternetService + StreamingTV + StreamingMovies + Contract, data =
train_clean2, family = "binomial")


#Test it on the train data
evaluation(f_model_cleaned, train)
```

```
## Training Set Metrics:
## AUC: 0.8531192
## [1] "Confusion Matrix:"
##       Predicted
## Actual   No  Yes
##    No  1790  631
##    Yes  488 2043
## Recall: 0.8071908
## F1-score: 0.7850144

## Warning in predict.lm(object, newdata, se.fit, scale = 1, type = if
(type == :
## prediction from rank-deficient fit; attr(*, "non-estim") has
doubtful cases

## Setting levels: control = No, case = Yes
## Setting direction: controls < cases

## Test Set Metrics:
## AUC: 0.8373466
## [1] "Confusion Matrix:"
```

```
##        Predicted
## Actual   No  Yes
##    No  1126  417
##    Yes  107  434
## Recall: 0.8022181
## F1-score: 0.6235632
```

To improve model reliability, data points with a Cook’s Distance greater than 0.002 were removed from the dataset. The refined mode was created from this data set which excluded those outliers and was then evaluated again. The resulting model was only ever so slightly improved.