Assigment 2

Alícia Chimeno Sarabia and Bruna

2023-12-02

Libraries

Load dataset

Data context

This dataset contains information about customers. Demographic data,

Data exploration

```r
dim(df)
```

```
## [1] 7043   21
```

```r
names(df)
```

```
##  [1] "customerID"       "gender"           "SeniorCitizen"    "Partner"
##  [5] "Dependents"       "tenure"           "PhoneService"     "MultipleLines"
##  [9] "InternetService"  "OnlineSecurity"   "OnlineBackup"     "DeviceProtection"
## [13] "TechSupport"      "StreamingTV"      "StreamingMovies"  "Contract"
## [17] "PaperlessBilling" "PaymentMethod"    "MonthlyCharges"   "TotalCharges"
## [21] "Churn"
```

```r
#str(df)
#summary(df)
```

We only have NA values in *TotalCharges*.

```r
summary(is.na(df))
```

```
##  customerID        gender         SeniorCitizen    Partner
##  Mode :logical   Mode :logical   Mode :logical   Mode :logical
##  FALSE:7043      FALSE:7043      FALSE:7043      FALSE:7043
##
```

```
##   Dependents        tenure        PhoneService    MultipleLines
##   Mode :logical   Mode :logical   Mode :logical   Mode :logical
##   FALSE:7043      FALSE:7043      FALSE:7043      FALSE:7043
##
##   InternetService OnlineSecurity  OnlineBackup    DeviceProtection
##   Mode :logical   Mode :logical   Mode :logical   Mode :logical
##   FALSE:7043      FALSE:7043      FALSE:7043      FALSE:7043
##
##   TechSupport     StreamingTV     StreamingMovies  Contract
##   Mode :logical   Mode :logical   Mode :logical   Mode :logical
##   FALSE:7043      FALSE:7043      FALSE:7043      FALSE:7043
##
##   PaperlessBilling PaymentMethod   MonthlyCharges  TotalCharges
##   Mode :logical    Mode :logical   Mode :logical   Mode :logical
##   FALSE:7043       FALSE:7043      FALSE:7043      FALSE:7032
##                                                    TRUE :11
##     Churn
##   Mode :logical
##   FALSE:7043
##
```

Variable Description

In total, we have 21 variables related to demographic, services, and accountant data. One is the ID, three are numerical variables, and 17 are categorical (? binary) variables. We will conduct a descriptive analysis and a data quality report for each variable, considering aspects such as the number of missing values, errors, and the distribution or balance of the variable…

1. customerID

Demographic data

2. gender Is a binary variable (female/male).

```r
sum(is.na(df$gender))
```

```
## [1] 0
```

```r
table(df$gender)
```

```
##
## Female   Male
##   3488   3555
```

3. SeniorCitizen It is a binary variable. Levels: 1(=yes)/0(=no).

```r
sum(is.na(df$SeniorCitizen))
```

```
## [1] 0
```

```r
table(df$SeniorCitizen)
```

```
##
##    0    1
## 5901 1142
```

4. Partner It is a binary variable. Levels: Yes/No.

```r
sum(is.na(df$Partner))
```

```
## [1] 0
```

```r
table(df$Partner)
```

```
##
##   No  Yes
## 3641 3402
```

5. Dependents It is a binary variable. Levels: Yes/No.

```r
sum(is.na(df$Dependents))
```

```
## [1] 0
```

```r
table(df$Dependents)
```

```
##
##   No  Yes
## 4933 2110
```

```r
#plots
par(mfrow = c(2, 2))
barplot(table(df$gender), main = "Distribution of gender",xlab = "Gender",col = "skyblue")
barplot(table(df$SeniorCitizen), main = "Distribution of SeniorCitizen",xlab = "SeniorCitizen",col = "sl
barplot(table(df$Partner), main = "Distribution of Partner",xlab = "Partner",col = "skyblue")
barplot(table(df$Dependents), main = "Distribution of Dependents",xlab = "Dependents",col = "skyblue")
```

Services of the costumer data

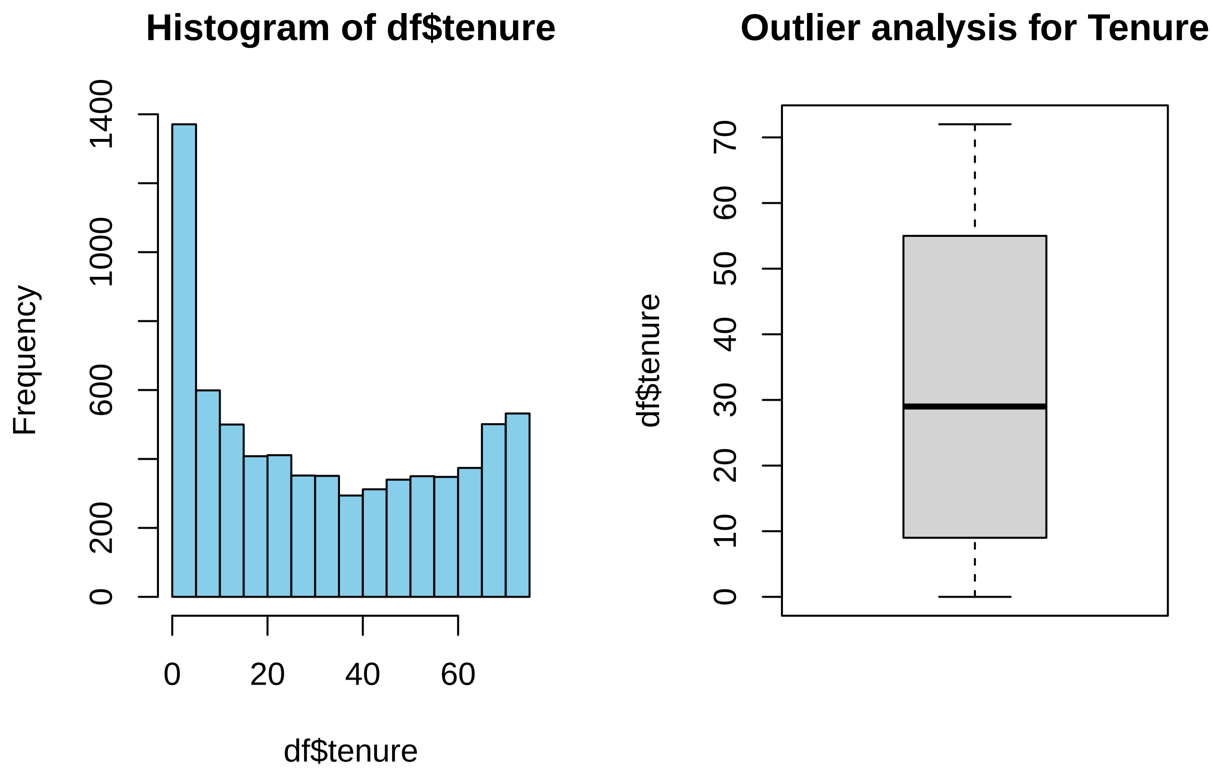
Services that each customer has signed up for:

6. tenure It is a numerical variable that indicates the duration, in months, that the customer has stayed with the company. We shall explore the statistics of the variable and look for the *outliers*

```
summary(df$tenure)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00    9.00   29.00   32.37   55.00   72.00
```

```
par(mfrow = c(1, 2))
hist(df$tenure,breaks=20, col="skyblue")
Boxplot(df$tenure, main="Outlier analysis for Tenure")
```




```
par(mfrow = c(1, 1))
sm_t <- summary(df$tenure)
iqr_t <- sm_t["3rd Qu."] - sm_t["1st Qu."]
# Mild Outliers
mild_ub_t <- sm_t["3rd Qu."] + 1.5 * iqr_t
mild_lb_t <- sm_t["1st Qu."] - 1.5 * iqr_t
length(which(df$tenure > mild_ub_t | df$tenure < mild_lb_t)) # number of mild outliers
```

```
## [1] 0
```

```
# Severe Outliers
severe_ub_t <- sm_t["3rd Qu."] + 3 * iqr_t
severe_lb_t <- sm_t["1st Qu."] - 3 * iqr_t
length(which(df$tenure > severe_ub_t | df$tenure < severe_lb_t)) # number of severe outliers
```

```
## [1] 0
```

There are *no mild nor severe outliers* in Tenure.

7. PhoneService It is a binary variable. Levels: Yes/No.

```
sum(is.na(df$PhoneService))
```

```
## [1] 0
```

```r
table(df$PhoneService)
```

```
##
##   No  Yes
##  682 6361
```

8. MultipleLines Categorical variable with 3 levels, No/No phone service/Yes.

```r
sum(is.na(df$MultipleLines))
```

```
## [1] 0
```

```r
table(df$MultipleLines)
```

```
##
##               No No phone service              Yes
##             3390              682             2971
```

Check inconsistencies: - Cannot happen that a costumer has not phoneservice and multiplelines.

```r
subset(df, MultipleLines == "Yes" & PhoneService == "No")
```

```
##  [1] customerID       gender           SeniorCitizen    Partner
##  [5] Dependents       tenure           PhoneService     MultipleLines
##  [9] InternetService  OnlineSecurity   OnlineBackup     DeviceProtection
## [13] TechSupport      StreamingTV      StreamingMovies  Contract
## [17] PaperlessBilling PaymentMethod    MonthlyCharges   TotalCharges
## [21] Churn
## <0 rows> (or 0-length row.names)
```

9. InternetService Categorical variable with 3 levels: DSL/Fiber optic/No.

```r
table(df$InternetService)
```

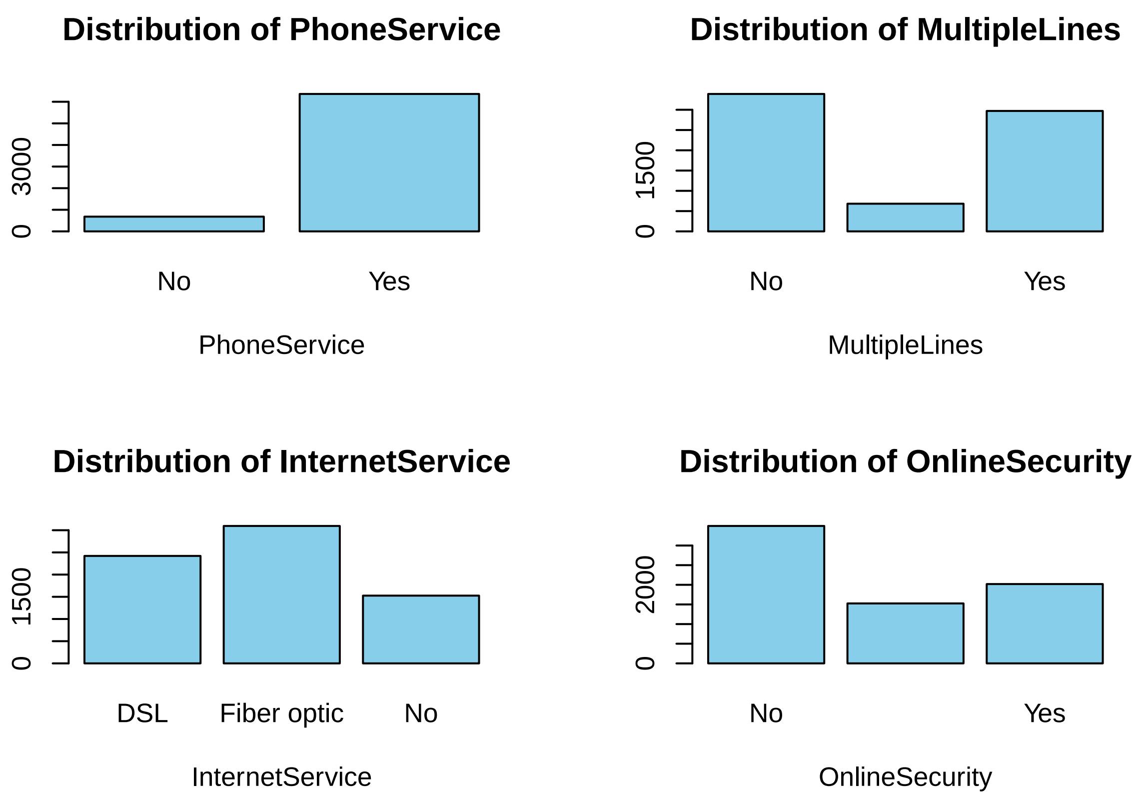
```
##
##         DSL Fiber optic          No
##        2421        3096        1526
```

10. OnlineSecurity Categorical variable with 3 levels: No/No internet service/Yes

```r
table(df$OnlineSecurity)
```

```
##
##                  No No internet service                 Yes
##                3498                1526                2019
```

```r
#plots:
par(mfrow = c(2, 2))
barplot(table(df$PhoneService), main = "Distribution of PhoneService",xlab = "PhoneService",col = "skybl
barplot(table(df$MultipleLines), main = "Distribution of MultipleLines",xlab = "MultipleLines",col = "sl
barplot(table(df$InternetService), main = "Distribution of InternetService",xlab = "InternetService",col
barplot(table(df$OnlineSecurity), main = "Distribution of OnlineSecurity",xlab = "OnlineSecurity",col =
```



Check consistency

```r
sum(df$InternetService == "No")
```

```
## [1] 1526
```

```r
sum(df$OnlineSecurity == "No internet service")
```

```
## [1] 1526
```

```r
nrow(subset(df, InternetService == "No" & OnlineSecurity == "No internet service"))
```

```
## [1] 1526
```

11. OnlineBackup Categorical variable with 3 levels: No/No internet service/Yes

```r
table(df$OnlineBackup)
```

```
##
##                  No No internet service                 Yes
##                3088                1526                2429
```

```r
# Check concistency
sum(df$OnlineBackup == "No internet service") #1526
```

```
## [1] 1526
```

```r
sum(df$OnlineSecurity == "No internet service") #1526
```

```
## [1] 1526
```

12. DeviceProtection Categorical variable with 3 levels: No/No internet service/Yes

```r
table(df$DeviceProtection)
```

```
##
##                  No No internet service                 Yes
##                3095                1526                2422
```

```r
# Check concistency
sum(df$OnlineSecurity == "No internet service") #1526
```

```
## [1] 1526
```

```r
sum(df$DeviceProtection == "No internet service") #1526
```

```
## [1] 1526
```

13. TechSupport Categorical variable with 3 levels: No/No internet service/Yes

```r
table(df$TechSupport)
```

```
##
##                  No No internet service                 Yes
##                3473                1526                2044
```

```r
#Check consistency
sum(df$DeviceProtection == "No internet service") #1526
```

```
## [1] 1526
```

```r
sum(df$TechSupport == "No internet service") #1526
```

```
## [1] 1526
```

14. StreamingTV Categorical variable with 3 levels: No/No internet service/Yes

```r
table(df$StreamingTV)
```

```
##
##                  No No internet service                 Yes
##                2810                1526                2707
```

```r
#Check consistency
sum(df$TechSupport == "No internet service") #1526
```

```
## [1] 1526
```

```r
sum(df$StreamingTV == "No internet service") #1526
```

```
## [1] 1526
```

15. StreamingMovies Categorical variable with 3 levels: No/No internet service/Yes

```r
table(df$StreamingMovies)
```

```
##
##                  No No internet service                 Yes
##                2785                1526                2732
```

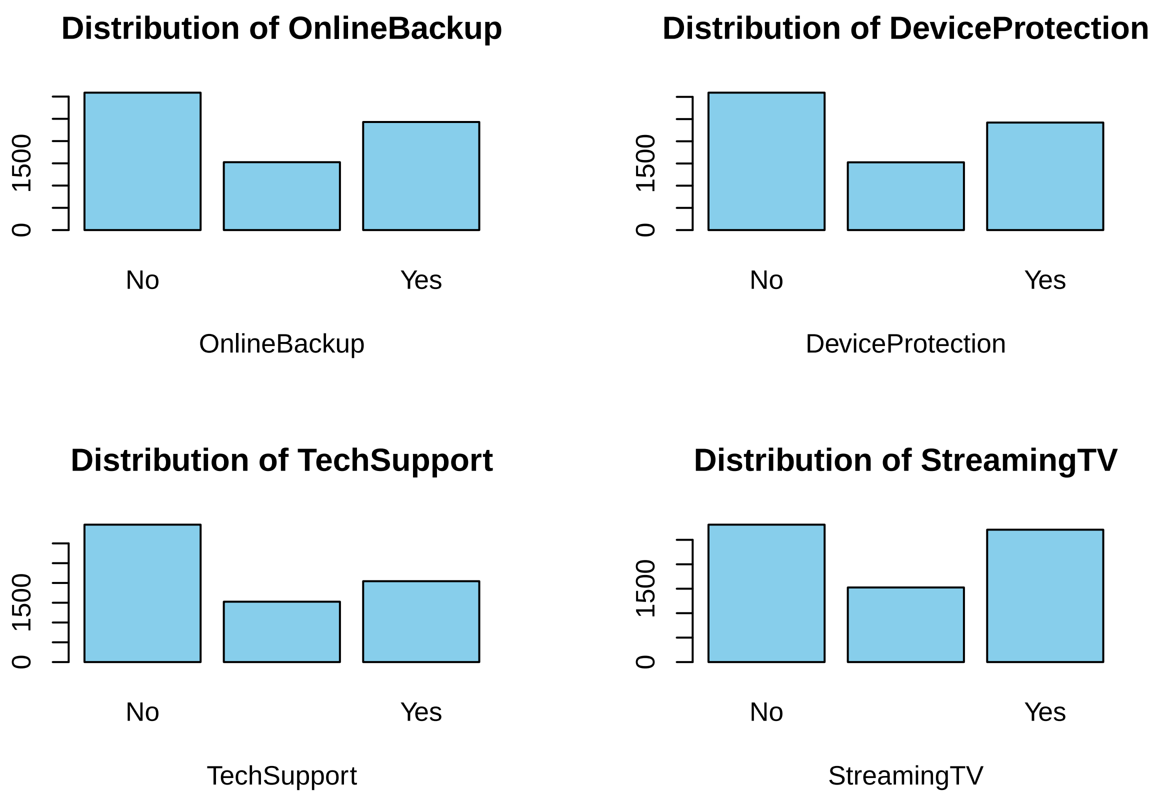
```r
#Check consistency
sum(df$StreamingTV == "No internet service") #1526
```

```
## [1] 1526
```

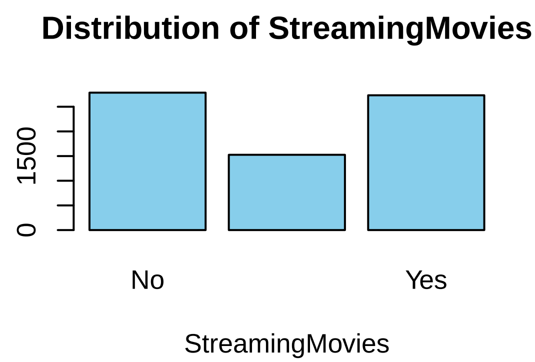
```r
sum(df$StreamingMovies == "No internet service") #1526
```

```
## [1] 1526
```

```r
#plots:
par(mfrow = c(2, 2))
barplot(table(df$OnlineBackup), main = "Distribution of OnlineBackup",xlab = "OnlineBackup",col = "skybl
barplot(table(df$DeviceProtection), main = "Distribution of DeviceProtection",xlab = "DeviceProtection"
barplot(table(df$TechSupport), main = "Distribution of TechSupport",xlab = "TechSupport",col = "skyblue
barplot(table(df$StreamingTV), main = "Distribution of StreamingTV",xlab = "StreamingTV",col = "skyblue
```




```
barplot(table(df$StreamingMovies), main = "Distribution of StreamingMovies",xlab = "StreamingMovies",col
```



Customer account data

16. Contract Categorical variable with 3 levels: Month-to-month/One year/Two year

```r
table(df$Contract)
```

```
##
## Month-to-month       One year       Two year
##           3875           1473           1695
```

17. PaperlessBilling It is a binary variable. Levels: No/Yes

```r
table(df$PaperlessBilling)
```

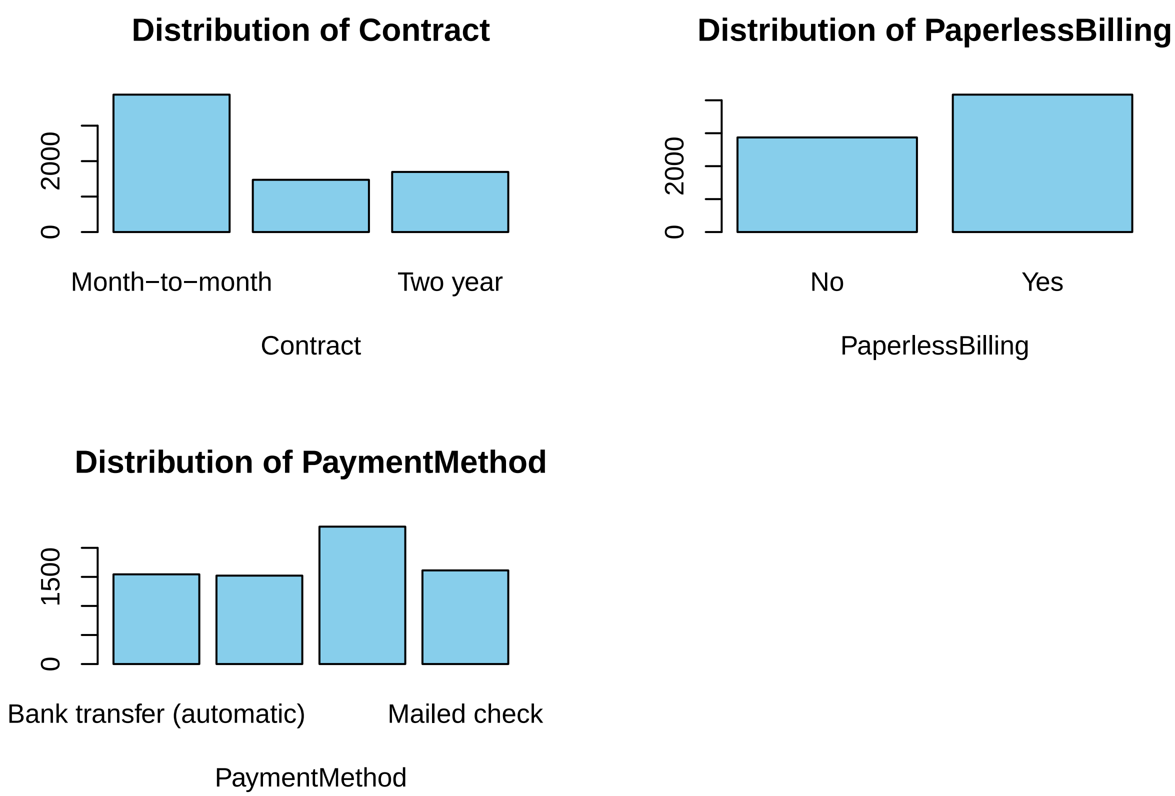
```
##
##   No  Yes
## 2872 4171
```

18. PaymentMethod Categorical variable with 4 levels: Bank transfer (automatic)/Credit card (automatic)/Electronic check/Mailed check

```r
table(df$PaymentMethod)
```

```
##
## Bank transfer (automatic)   Credit card (automatic)          Electronic check
##                      1544                      1522                      2365
##              Mailed check
##                      1612
```

```r
#plots
par(mfrow = c(2, 2))
barplot(table(df$Contract), main = "Distribution of Contract",xlab = "Contract",col = "skyblue")
barplot(table(df$PaperlessBilling), main = "Distribution of PaperlessBilling",xlab = "PaperlessBilling"
barplot(table(df$PaymentMethod), main = "Distribution of PaymentMethod",xlab = "PaymentMethod",col = "sl
```

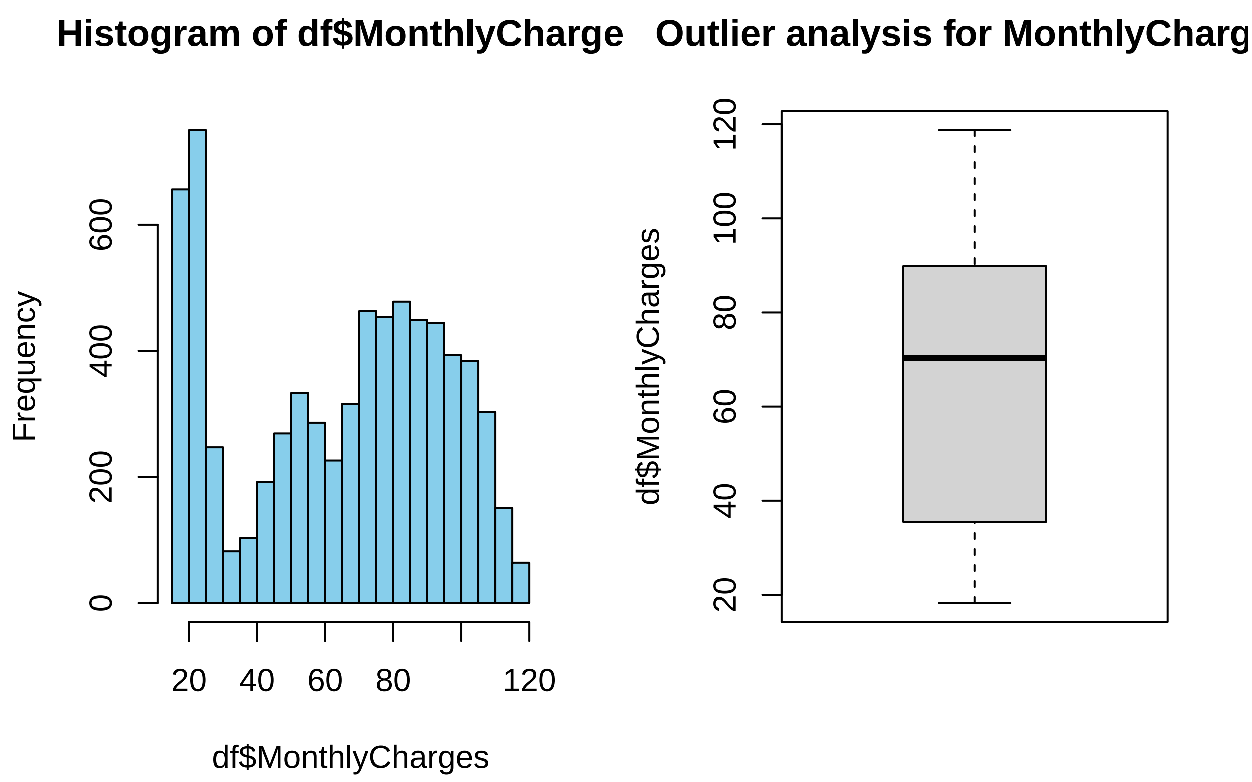


19. MonthlyCharges It is a numerical variable.

```r
summary(df$MonthlyCharges)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   18.25   35.50   70.35   64.76   89.85  118.75
```

```
par(mfrow = c(1,2 ))
hist(df$MonthlyCharges,breaks=20,col="skyblue")
Boxplot(df$MonthlyCharges, main="Outlier analysis for MonthlyCharges")
```



Let's look for *outliers*.

```
sm <- summary(df$MonthlyCharges)
iqr <- sm["3rd Qu."] - sm["1st Qu."]
# Mild Outliers
mild_ub <- sm["3rd Qu."] + 1.5 * iqr
mild_lb <- sm["1st Qu."] - 1.5 * iqr
length(which(df$MonthlyCharges > mild_ub | df$MonthlyCharges < mild_lb)) # number of mild outliers
```

```
## [1] 0
```

```
# Severe Outliers
severe_ub <- sm["3rd Qu."] + 3 * iqr
severe_lb <- sm["1st Qu."] - 3 * iqr
length(which(df$MonthlyCharges > severe_ub | df$MonthlyCharges < severe_lb)) # number of severe outlier
```

```
## [1] 0
```

There are no mild nor severe outliers in MonthlyCharges.

20. TotalCharges (numeric) It is a numerical variable.

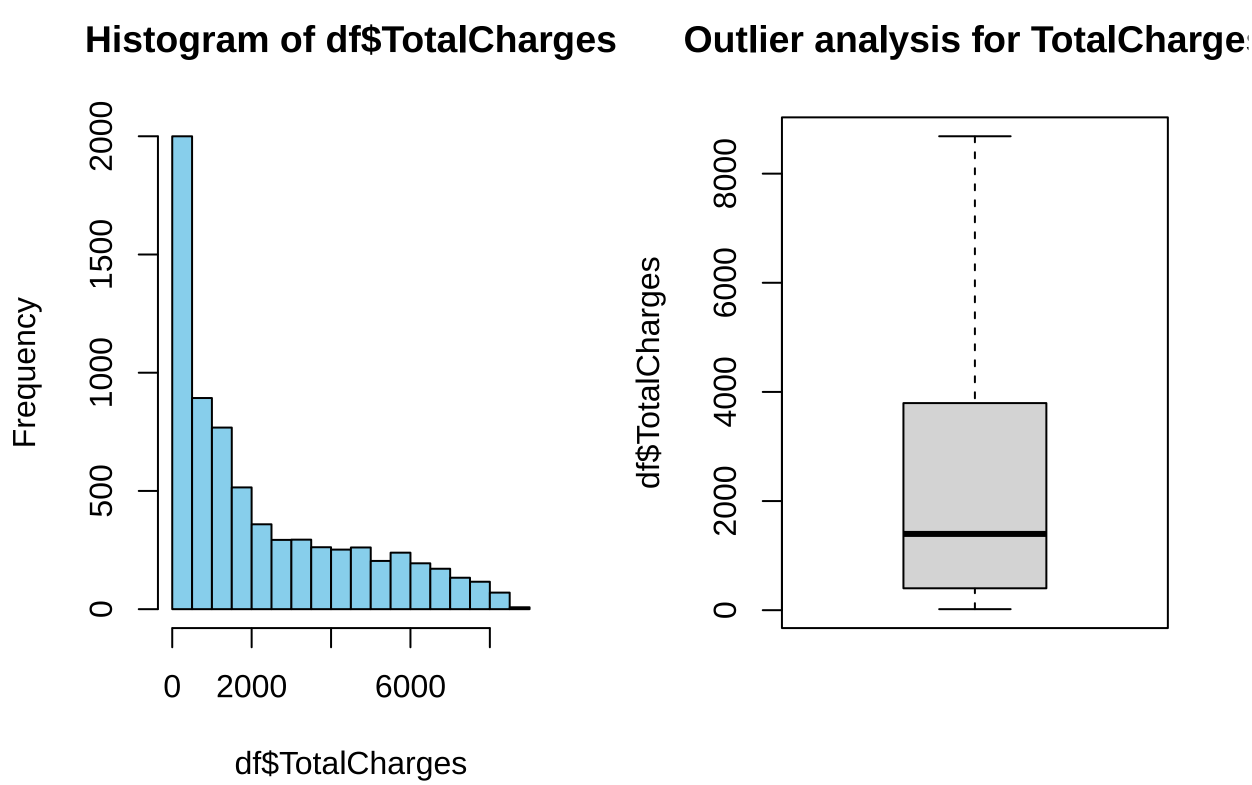
```
summary(df$TotalCharges)
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.     NA's 
##     18.8    401.4   1397.5   2283.3   3794.7   8684.8       11
```

```
sum(is.na(df$TotalCharges))
```

```
## [1] 11
```

```
par(mfrow = c(1, 2))
hist(df$TotalCharges,breaks=20,col="skyblue")
Boxplot(df$TotalCharges, main="Outlier analysis for TotalCharges")
```



Let's look for *outliers*.

```
sm <- summary(df$TotalCharges)
iqr <- sm["3rd Qu."] - sm["1st Qu."]
# Mild Outliers
mild_ub <- sm["3rd Qu."] + 1.5 * iqr
mild_lb <- sm["1st Qu."] - 1.5 * iqr
length(which(df$TotalCharges > mild_ub | df$TotalCharges < mild_lb)) # number of mild outliers
```

```
## [1] 0
```

```r
# Severe Outliers
severe_ub <- sm["3rd Qu."] + 3 * iqr
severe_lb <- sm["1st Qu."] - 3 * iqr
length(which(df$TotalCharges > severe_ub | df$TotalCharges < severe_lb)) # number of severe outliers
```

```
## [1] 0
```

There are no mild nor severe outliers.

Target:

21. Churn It is the target variable. It is binary, describes whether the customer churned or not (Yes or No).

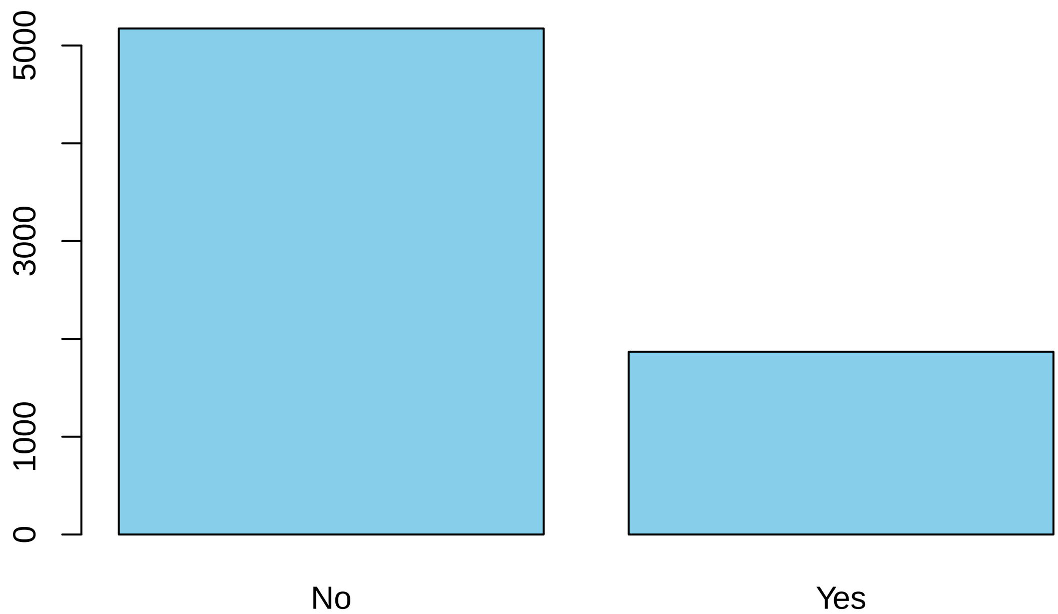
```r
table(df$Churn)
```

```
##
##   No  Yes
## 5174 1869
```

```r
prop.table(table(df$Churn))
```

```
##
##        No       Yes
## 0.7346301 0.2653699
```

```r
barplot(table(df$Churn), col="skyblue")
```



Data preprocessing

Recode variables into correct type

We shall reconvert the type of certain variables that are encoded with wrong type. First, we convert the character variables (except the ID) into factors.

```
char_cols <- which(sapply(df, is.character))
df[, char_cols[-1]]<- lapply(df[, char_cols[-1]], as.factor)
```

Also, we convert the numerical variable SeniorCitizen into a factor.

```
df$SeniorCitizen<- factor(df$SeniorCitizen)
```

Data imputation

```
summary(is.na(df))
```

```
##  customerID          gender          SeniorCitizen     Partner
##  Mode :logical    Mode :logical    Mode :logical    Mode :logical
##  FALSE:7043       FALSE:7043       FALSE:7043       FALSE:7043
```

```
##
##   Dependents        tenure        PhoneService    MultipleLines
##   Mode :logical   Mode :logical   Mode :logical   Mode :logical
##   FALSE:7043      FALSE:7043      FALSE:7043      FALSE:7043
##
##   InternetService OnlineSecurity  OnlineBackup    DeviceProtection
##   Mode :logical   Mode :logical   Mode :logical   Mode :logical
##   FALSE:7043      FALSE:7043      FALSE:7043      FALSE:7043
##
##   TechSupport     StreamingTV     StreamingMovies  Contract
##   Mode :logical   Mode :logical   Mode :logical   Mode :logical
##   FALSE:7043      FALSE:7043      FALSE:7043      FALSE:7043
##
##   PaperlessBilling PaymentMethod   MonthlyCharges  TotalCharges
##   Mode :logical    Mode :logical   Mode :logical   Mode :logical
##   FALSE:7043       FALSE:7043      FALSE:7043      FALSE:7032
##                                                    TRUE :11
##     Churn
##   Mode :logical
##   FALSE:7043
##
```

Only the variable TotalCharges has NA's.

The missing data corresponds to the individuals that have not payed yet the charges of the current month, we can guess that are new clients of the company.

Duplicate values: no

```r
dim(df)
```

```
## [1] 7043   21
```

```r
length(unique(df$customerID))
```

```
## [1] 7043
```

These NA exist because the costumer hasn't payed yet that month (tenure is 0). We convert these NA to 0.

```r
ll <- which(is.na(df$TotalCharges))
df[ll,"TotalCharges"] <- 0
summary(is.na(df$TotalCharges))
```

```
##    Mode   FALSE
## logical    7043
```

Correlation between categorical

The categorical variables MultipleLines and PhoneService are 100% correlated. We might have multicollinearity between these two variables.

```r
contingency_table<-table(df$MultipleLines,df$PhoneService)
sqrt(chisq.test(contingency_table)$statistic / (sum(contingency_table) * (min(dim(contingency_table)) -
```

```
## X-squared
##         1
```

Profiling

```r
res.cat=catdes(df, 21)
res.cat$test.chi2
```

```
##                       p.value df
## Contract         5.863038e-258  2
## OnlineSecurity   2.661150e-185  2
## TechSupport      1.443084e-180  2
## InternetService  9.571788e-160  2
## PaymentMethod    3.682355e-140  3
## OnlineBackup     2.079759e-131  2
## DeviceProtection 5.505219e-122  2
## StreamingMovies   2.667757e-82  2
## StreamingTV       5.528994e-82  2
## PaperlessBilling  2.614597e-58  1
## Dependents        3.276083e-43  1
## SeniorCitizen     9.477904e-37  1
## Partner           1.519037e-36  1
## MultipleLines     3.464383e-03  2
```

```r
head(res.cat$category)
```

```
## $No
##                                          Cla/Mod  Mod/Cla   Global
## Contract=Two year                       97.16814 31.83224 24.06645
## StreamingMovies=No internet service     92.59502 27.30963 21.66690
## StreamingTV=No internet service         92.59502 27.30963 21.66690
## TechSupport=No internet service         92.59502 27.30963 21.66690
## DeviceProtection=No internet service    92.59502 27.30963 21.66690
## OnlineBackup=No internet service        92.59502 27.30963 21.66690
## OnlineSecurity=No internet service      92.59502 27.30963 21.66690
## InternetService=No                      92.59502 27.30963 21.66690
## PaperlessBilling=No                     83.66992 46.44376 40.77808
## Contract=One year                       88.73048 25.26092 20.91438
## OnlineSecurity=Yes                      85.38881 33.32045 28.66676
## TechSupport=Yes                         84.83366 33.51372 29.02172
## Dependents=Yes                          84.54976 34.48009 29.95882
## Partner=Yes                             80.33510 52.82180 48.30328
## SeniorCitizen=0                         76.39383 87.12795 83.78532
## PaymentMethod=Credit card (automatic)   84.75690 24.93235 21.61011
## InternetService=DSL                     81.04089 37.92037 34.37456
## PaymentMethod=Bank transfer (automatic) 83.29016 24.85504 21.92248
## PaymentMethod=Mailed check              80.89330 25.20294 22.88797
```

```
## OnlineBackup=Yes                       78.46851 36.83804 34.48814
## DeviceProtection=Yes                   77.49794 36.27754 34.38875
## MultipleLines=No                       74.95575 49.11094 48.13290
## MultipleLines=Yes                      71.39010 40.99343 42.18373
## StreamingMovies=Yes                    70.05857 36.99266 38.79029
## StreamingTV=Yes                        69.92981 36.58678 38.43533
## StreamingTV=No                         66.47687 36.10359 39.89777
## StreamingMovies=No                     66.31957 35.69772 39.54281
## SeniorCitizen=1                        58.31874 12.87205 16.21468
## Partner=No                             67.04202 47.17820 51.69672
## Dependents=No                          68.72086 65.51991 70.04118
## PaperlessBilling=Yes                   66.43491 53.55624 59.22192
## DeviceProtection=No                    60.87237 36.41283 43.94434
## OnlineBackup=No                        60.07124 35.85234 43.84495
## PaymentMethod=Electronic check         54.71459 25.00966 33.57944
## InternetService=Fiber optic            58.10724 34.77000 43.95854
## TechSupport=No                         58.36453 39.17665 49.31137
## OnlineSecurity=No                      58.23328 39.36993 49.66634
## Contract=Month-to-month                57.29032 42.90684 55.01917
##                                             p.value    v.test
## Contract=Two year                      3.588830e-187  29.178937
## StreamingMovies=No internet service     6.584621e-98  20.999812
## StreamingTV=No internet service         6.584621e-98  20.999812
## TechSupport=No internet service         6.584621e-98  20.999812
## DeviceProtection=No internet service    6.584621e-98  20.999812
## OnlineBackup=No internet service        6.584621e-98  20.999812
## OnlineSecurity=No internet service      6.584621e-98  20.999812
## InternetService=No                      6.584621e-98  20.999812
## PaperlessBilling=No                     1.072745e-60  16.435085
## Contract=One year                       3.593041e-57  15.935502
## OnlineSecurity=Yes                      1.606459e-50  14.947938
## TechSupport=Yes                         1.323174e-46  14.334963
## Dependents=Yes                          3.572324e-46  14.265846
## Partner=Yes                             6.170871e-37  12.696658
## SeniorCitizen=0                         3.024931e-34  12.202212
## PaymentMethod=Credit card (automatic)   6.408166e-32  11.758206
## InternetService=DSL                     2.545367e-26  10.614727
## PaymentMethod=Bank transfer (automatic)  1.180908e-24  10.250207
## PaymentMethod=Mailed check              3.226893e-15   7.881803
## OnlineBackup=Yes                        3.021982e-12   6.976698
## DeviceProtection=Yes                    2.173366e-08   5.597602
## MultipleLines=No                        6.262488e-03   2.733712
## MultipleLines=Yes                       7.843169e-04  -3.358271
## StreamingMovies=Yes                     2.922571e-07  -5.128373
## StreamingTV=Yes                         1.283457e-07  -5.281193
## StreamingTV=No                          6.049871e-27 -10.748094
## StreamingMovies=No                      1.092934e-27 -10.904833
## SeniorCitizen=1                         3.024931e-34 -12.202212
## Partner=No                              6.170871e-37 -12.696658
## Dependents=No                           3.572324e-46 -14.265846
## PaperlessBilling=Yes                    1.072745e-60 -16.435085
## DeviceProtection=No                     1.116896e-99 -21.192627
## OnlineBackup=No                        3.366400e-112 -22.509287
## PaymentMethod=Electronic check         1.790860e-136 -24.864755
```

```
## InternetService=Fiber optic            2.289126e-148 -25.941138
## TechSupport=No                         1.899538e-183 -28.883947
## OnlineSecurity=No                      6.171504e-190 -29.396034
## Contract=Month-to-month                3.620915e-283 -35.959308
##
## $Yes
##                                          Cla/Mod   Mod/Cla   Global
## Contract=Month-to-month                42.709677 88.550027 55.01917
## OnlineSecurity=No                      41.766724 78.170144 49.66634
## TechSupport=No                         41.635474 77.367576 49.31137
## InternetService=Fiber optic            41.892765 69.395399 43.95854
## PaymentMethod=Electronic check         45.285412 57.303371 33.57944
## OnlineBackup=No                        39.928756 65.971108 43.84495
## DeviceProtection=No                    39.127625 64.794007 43.94434
## PaperlessBilling=Yes                   33.565092 74.906367 59.22192
## Dependents=No                          31.279140 82.557517 70.04118
## Partner=No                             32.957979 64.205457 51.69672
## SeniorCitizen=1                        41.681261 25.468165 16.21468
## StreamingMovies=No                     33.680431 50.187266 39.54281
## StreamingTV=No                         33.523132 50.401284 39.89777
## StreamingTV=Yes                        30.070188 43.552702 38.43533
## StreamingMovies=Yes                    29.941435 43.766720 38.79029
## MultipleLines=Yes                      28.609896 45.478866 42.18373
## MultipleLines=No                       25.044248 45.425361 48.13290
## DeviceProtection=Yes                   22.502064 29.159979 34.38875
## OnlineBackup=Yes                       21.531494 27.982879 34.48814
## PaymentMethod=Mailed check             19.106700 16.479401 22.88797
## PaymentMethod=Bank transfer (automatic) 16.709845 13.804173 21.92248
## InternetService=DSL                    18.959108 24.558587 34.37456
## PaymentMethod=Credit card (automatic)  15.243101 12.413055 21.61011
## SeniorCitizen=0                        23.606168 74.531835 83.78532
## Partner=Yes                            19.664903 35.794543 48.30328
## Dependents=Yes                         15.450237 17.442483 29.95882
## TechSupport=Yes                        15.166341 16.586410 29.02172
## OnlineSecurity=Yes                     14.611194 15.783842 28.66676
## Contract=One year                      11.269518  8.881755 20.91438
## PaperlessBilling=No                    16.330084 25.093633 40.77808
## StreamingMovies=No internet service     7.404980  6.046014 21.66690
## StreamingTV=No internet service         7.404980  6.046014 21.66690
## TechSupport=No internet service         7.404980  6.046014 21.66690
## DeviceProtection=No internet service    7.404980  6.046014 21.66690
## OnlineBackup=No internet service        7.404980  6.046014 21.66690
## OnlineSecurity=No internet service      7.404980  6.046014 21.66690
## InternetService=No                      7.404980  6.046014 21.66690
## Contract=Two year                       2.831858  2.568218 24.06645
##                                              p.value     v.test
## Contract=Month-to-month                3.620915e-283  35.959308
## OnlineSecurity=No                      6.171504e-190  29.396034
## TechSupport=No                         1.899538e-183  28.883947
## InternetService=Fiber optic            2.289126e-148  25.941138
## PaymentMethod=Electronic check         1.790860e-136  24.864755
## OnlineBackup=No                        3.366400e-112  22.509287
## DeviceProtection=No                     1.116896e-99  21.192627
## PaperlessBilling=Yes                    1.072745e-60  16.435085
```

```
## Dependents=No                              3.572324e-46  14.265846
## Partner=No                                 6.170871e-37  12.696658
## SeniorCitizen=1                            3.024931e-34  12.202212
## StreamingMovies=No                         1.092934e-27  10.904833
## StreamingTV=No                             6.049871e-27  10.748094
## StreamingTV=Yes                            1.283457e-07   5.281193
## StreamingMovies=Yes                        2.922571e-07   5.128373
## MultipleLines=Yes                          7.843169e-04   3.358271
## MultipleLines=No                           6.262488e-03  -2.733712
## DeviceProtection=Yes                       2.173366e-08  -5.597602
## OnlineBackup=Yes                           3.021982e-12  -6.976698
## PaymentMethod=Mailed check                 3.226893e-15  -7.881803
## PaymentMethod=Bank transfer (automatic)   1.180908e-24 -10.250207
## InternetService=DSL                        2.545367e-26 -10.614727
## PaymentMethod=Credit card (automatic)     6.408166e-32 -11.758206
## SeniorCitizen=0                            3.024931e-34 -12.202212
## Partner=Yes                                6.170871e-37 -12.696658
## Dependents=Yes                             3.572324e-46 -14.265846
## TechSupport=Yes                            1.323174e-46 -14.334963
## OnlineSecurity=Yes                         1.606459e-50 -14.947938
## Contract=One year                          3.593041e-57 -15.935502
## PaperlessBilling=No                        1.072745e-60 -16.435085
## StreamingMovies=No internet service        6.584621e-98 -20.999812
## StreamingTV=No internet service            6.584621e-98 -20.999812
## TechSupport=No internet service            6.584621e-98 -20.999812
## DeviceProtection=No internet service       6.584621e-98 -20.999812
## OnlineBackup=No internet service           6.584621e-98 -20.999812
## OnlineSecurity=No internet service         6.584621e-98 -20.999812
## InternetService=No                         6.584621e-98 -20.999812
## Contract=Two year                         3.588830e-187 -29.178937
```

```
res.cat$quanti.var
```

```
##                      Eta2       P-value
## tenure         0.12406504 7.999058e-205
## TotalCharges   0.03933251  2.127212e-63
## MonthlyCharges 0.03738671  2.706646e-60
```

```
res.cat$quanti
```

```
## $No
##                   v.test Mean in category Overall mean sd in category
## tenure          29.55784         37.56997     32.37115       24.11145
## TotalCharges    16.64270       2549.91144   2279.73430     2329.72904
## MonthlyCharges -16.22582         61.26512     64.76169       31.08964
##                Overall sd       p.value
## tenure           24.55774 5.207314e-192
## TotalCharges   2266.63354  3.418341e-62
## MonthlyCharges   30.08791  3.312724e-59
##
## $Yes
##                   v.test Mean in category Overall mean sd in category
## MonthlyCharges  16.22582         74.44133     64.76169       24.65945
```

```
## TotalCharges   -16.64270       1531.79609   2279.73430      1890.31709
## tenure         -29.55784         17.97913     32.37115        19.52590
##                Overall sd       p.value
## MonthlyCharges   30.08791  3.312724e-59
## TotalCharges   2266.63354  3.418341e-62
## tenure           24.55774 5.207314e-192
```

Regarding to the results of the test Chi^2 all correlations with the variables are significant since the $p-value$ is less than 0,05. Since the response variable is binary, we have different results for each answer and also for all outcomes of the categorical parameters.

For example, we can analyse in detail the variable "Contract". For the customers that haven't churned, the correlation between the ones that have a contract of two year is directly proportional and it's the highest relation. However, we can see that is the costumer has churned the ones that have a two-year contract have an strong negative correlation. The ones that have a month-to-month contract are the opposite of the previous answer; they have the highest positive correlation with the costumers that have churned and the negative with the ones that haven't.

Besides the latter variable, we can observe the parameter that have a higher positive correlation with the costumers that churn is the parameter "OnlineSecurity" and "TechSupport" when the answer is "No". The parameters that have a negative relation with the costumers that churn are when they haven't hired an Internet Service. We can see that all parameters that have an answer that is "No internet service" have also a negative relation with the response variable "Yes". We can deduce that they might have multicollinearity with the parameter Internet Service, but we will check it later.

The parameters that have a higher positive relation with the costumers that don't churn are the ones that have a negative relation when the response variable is "Yes", that we have analysed before. In the same vein, we can observe that the parameters that have a negative relation with the costumers that churn are "OnlineSecurity" and "TechSupport" when the answer is "No", the same parameters that have a positive relation when the costumers churn. We can see that the target answer "Yes" and "No" have an approximate opposite correlations with the explanatory variables.

Modelling

Data transformations:

Recall that the following variables:

- OnlineSecurity
- OnlineBackup
- DeviceProtection
- TechSupport
- StreamingTV
- StreamingMovies

are categorical variables with 3 levels: No/No internet service/Yes.

We observe that they contain "No internet service" as a response. We have a variable called *InternetService* that is a categorical variable with 3 levels: DSL/Fiber optic/No. Whenever *InternetService*="No" implies -> var="No internet service". Therefore we decided to transform the level "No internet service" into "No" in the 6 variables above since this variable will specify.

```r
df$OnlineSecurity[df$OnlineSecurity=="No internet service"] <- "No"
df$OnlineBackup[df$OnlineBackup=="No internet service"] <- "No"
df$DeviceProtection[df$DeviceProtection=="No internet service"] <- "No"
df$TechSupport[df$TechSupport=="No internet service"] <- "No"
df$StreamingTV[df$StreamingTV=="No internet service"] <- "No"
df$StreamingMovies[df$StreamingMovies=="No internet service"] <- "No"
```

We saw that *MultipleLines* is 100% related with *PhoneService*. The reason is similar as the previous parameters: one answer of *MultipleLines* is "No phone service". We set this answer to "No" since we don't lose the information because it is contained inside the parameter *PhoneService*.

```r
df$MultipleLines[df$MultipleLines=="No phone service"] <- "No"
summary(df)
```

```
##   customerID           gender     SeniorCitizen Partner    Dependents
##  Length:7043        Female:3488   0:5901        No :3641   No :4933
##  Class :character   Male  :3555   1:1142        Yes:3402   Yes:2110
##  Mode  :character
##
##
##
##      tenure      PhoneService          MultipleLines     InternetService
##  Min.   : 0.00   No : 682     No              :4072   DSL        :2421
##  1st Qu.: 9.00   Yes:6361     No phone service:   0   Fiber optic:3096
##  Median :29.00                Yes             :2971   No         :1526
##  Mean   :32.37
##  3rd Qu.:55.00
##  Max.   :72.00
##             OnlineSecurity              OnlineBackup
##  No                 :5024   No                 :4614
##  No internet service:   0   No internet service:   0
##  Yes                :2019   Yes                :2429
##
##
##
##            DeviceProtection               TechSupport
##  No                 :4621    No                 :4999
##  No internet service:   0    No internet service:   0
##  Yes                :2422    Yes                :2044
##
##
##
##              StreamingTV              StreamingMovies           Contract
##  No                 :4336   No                 :4311   Month-to-month:3875
##  No internet service:   0   No internet service:   0   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.   :   0.0   No :5174
##  1st Qu.: 398.6   Yes:1869
##  Median :1394.5
##  Mean   :2279.7
##  3rd Qu.:3786.6
##  Max.   :8684.8
```

Modelling:

```
set.seed(1234)
m <- floor(0.7*nrow(df))
train_d <- sample(seq_len(nrow(df)),size = m)

train <- df[train_d,]
test <- df[-train_d,]
```

Target variable is Churn.

Numerical Variables

We start the modelling by the null model.

```
mod0 <- glm(Churn ~ 1, data=train, family=binomial)
mod0$deviance
```

```
## [1] 5694.218
```

We continue by adding the numerical variables and assessing the model.

```
which(sapply(df, is.numeric))
```

```
##         tenure MonthlyCharges   TotalCharges
##              6             19             20
```

We start by *tenure*

```
mod1 <- glm(Churn ~ tenure, data=train, family=binomial)
mod1$deviance;AIC(mod0,mod1) #summary(mod1)
```

```
## [1] 5040.677
```

```
##      df      AIC
## mod0  1 5696.218
## mod1  2 5044.677
```

```r
anova( mod0, mod1,  test="Chisq")
```

```
## Analysis of Deviance Table
##
## Model 1: Churn ~ 1
## Model 2: Churn ~ tenure
##   Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
## 1      4929     5694.2
## 2      4928     5040.7  1   653.54 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Add MonthlyCharges

```r
mod2 <- glm(Churn ~ tenure + MonthlyCharges, data=train, family=binomial)
mod2$deviance
```

```
## [1] 4467.45
```

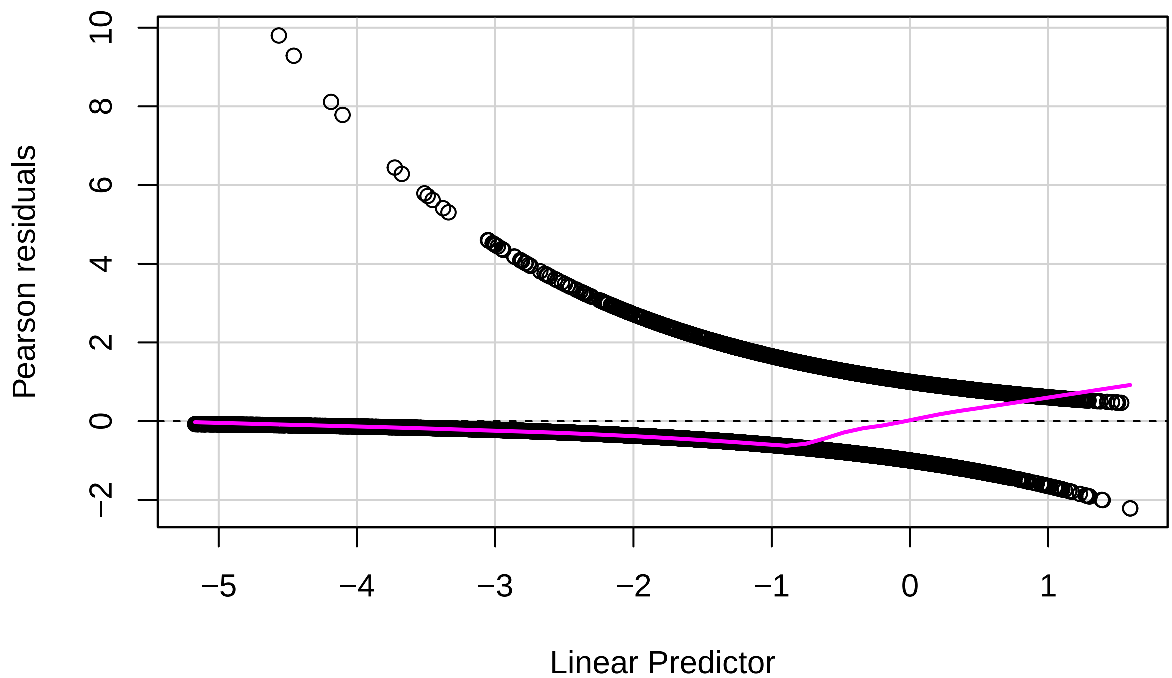
```r
AIC(mod2) #4473.45
```

```
## [1] 4473.45
```

```r
anova( mod1, mod2,  test="Chisq")
```

```
## Analysis of Deviance Table
##
## Model 1: Churn ~ tenure
## Model 2: Churn ~ tenure + MonthlyCharges
##   Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
## 1      4928     5040.7
## 2      4927     4467.5  1   573.23 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
residualPlot(mod2)
```



Add TotalCharges

```r
mod3 <- glm(Churn ~ tenure + MonthlyCharges + TotalCharges, data=train, family=binomial)
mod3$deviance
```

```
## [1] 4460.555
```

```r
anova( mod2, mod3,  test="Chisq") #significant
```

```
## Analysis of Deviance Table
##
## Model 1: Churn ~ tenure + MonthlyCharges
## Model 2: Churn ~ tenure + MonthlyCharges + TotalCharges
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1      4927     4467.5
## 2      4926     4460.6  1   6.8951 0.008643 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

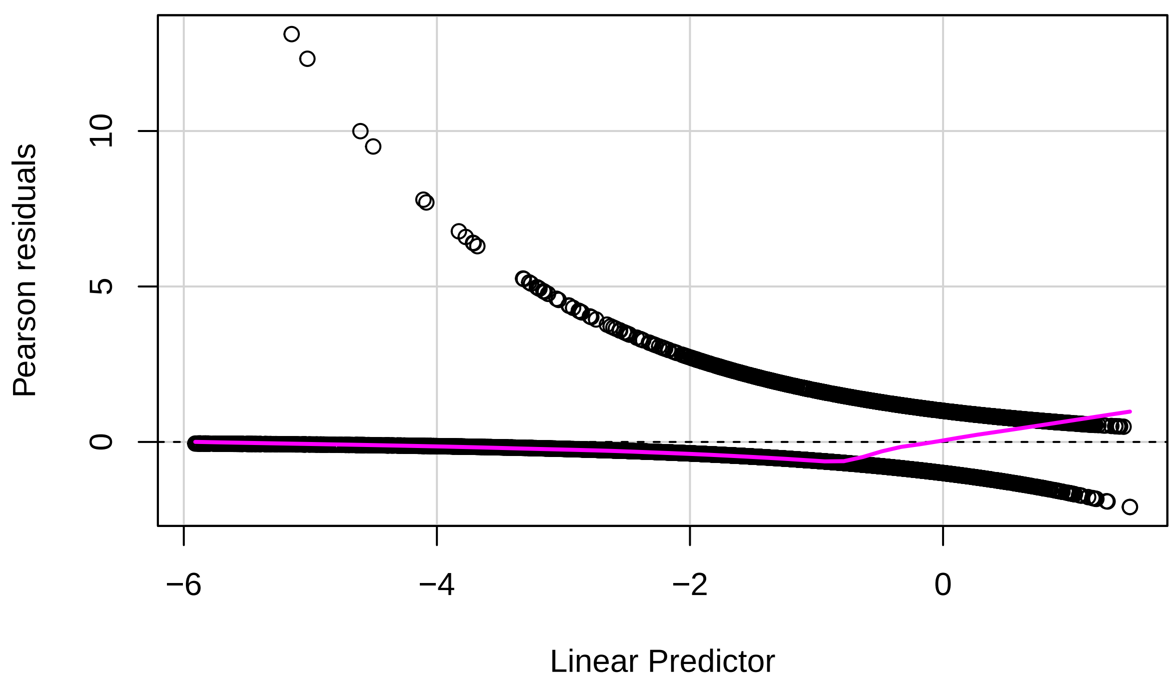
```r
AIC(mod3) #4468.55
```

```
## [1] 4468.555
```

```r
vif(mod3)
```

```
##         tenure MonthlyCharges   TotalCharges
##      14.730657       2.271293      18.869079
```

```r
residualPlot(mod3)
```



It is significant enough but we can also see that *TotalCHarges* has a high VIF, so it has high multicollinearity. We decide to not include it in the model.

Inlfuential data

```r
infl <- influence.measures(mod3)

sum(residuals(mod3,'deviance')^2)
```

```
## [1] 4460.555
```

```r
sum(residuals(mod3,'pearson')^2)
```

```
## [1] 5196.056
```

```r
influential_indices <- which(infl$is.inf == TRUE)
length(influential_indices)
```

```
## [1] 209
```

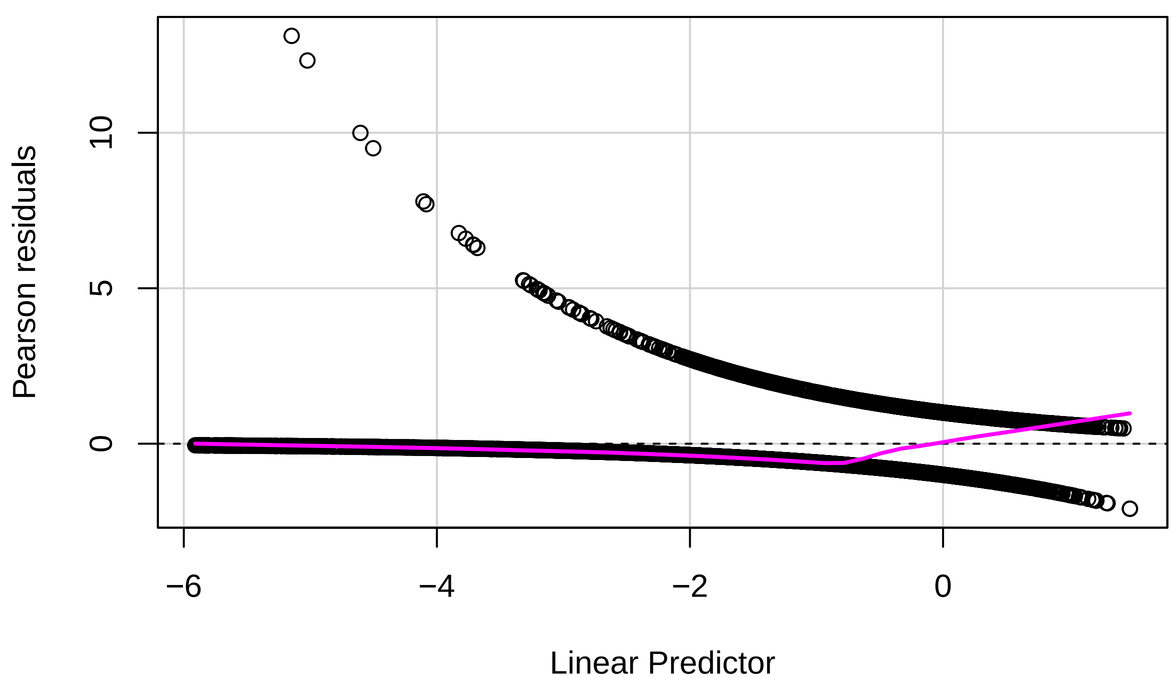
```r
length(train$customerID)
```

```
## [1] 4930
```

We have 209 influential points out of 4930.

Residuals

```r
residualPlot(mod3)
```



The residuals need to be nearer to the 0.

Categorical Variables

Now, we shall add the categorical variables. The order of addition is significant, therefore we start by adding the most correlated variables with the target.

Contract InternetService StreamingMovies StreamingTV TechSupport DeviceProtection OnlineBackup OnlineSecurity PaperlessBilling Dependents MultipleLines SeniorCitizen Partner PaymentMethod PhoneService

Contract

We start with *Contract* variable.

```
mod4 <- glm(Churn ~ tenure + MonthlyCharges + Contract, data=train, family=binomial)
AIC(mod4) #4302.2 better
```

```
## [1] 4302.234
```

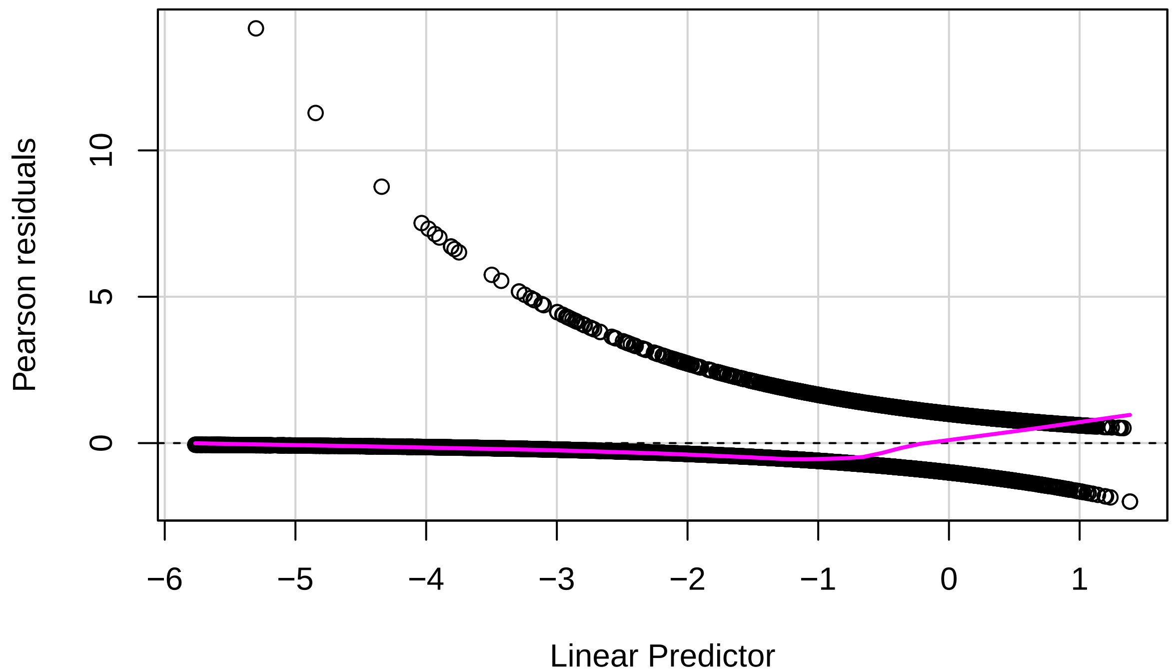
```
anova( mod3, mod4,  test="Chisq") #significant
```

```
## Analysis of Deviance Table
##
## Model 1: Churn ~ tenure + MonthlyCharges + TotalCharges
## Model 2: Churn ~ tenure + MonthlyCharges + Contract
##   Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
## 1      4926     4460.6
## 2      4925     4292.2  1   168.32 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
vif(mod4)
```

```
##                    GVIF Df GVIF^(1/(2*Df))
## tenure         1.707900  1        1.306867
## MonthlyCharges 1.300967  1        1.140599
## Contract       1.361428  2        1.080186
```

```
residualPlot(mod4)
```



We add the parameter because it improves the model.

InternetService

```r
mod5 <- glm(Churn ~ tenure + MonthlyCharges + Contract + InternetService, data=train, family=binomial)
AIC(mod5) #4254.1 better
```

```
## [1] 4254.114
```

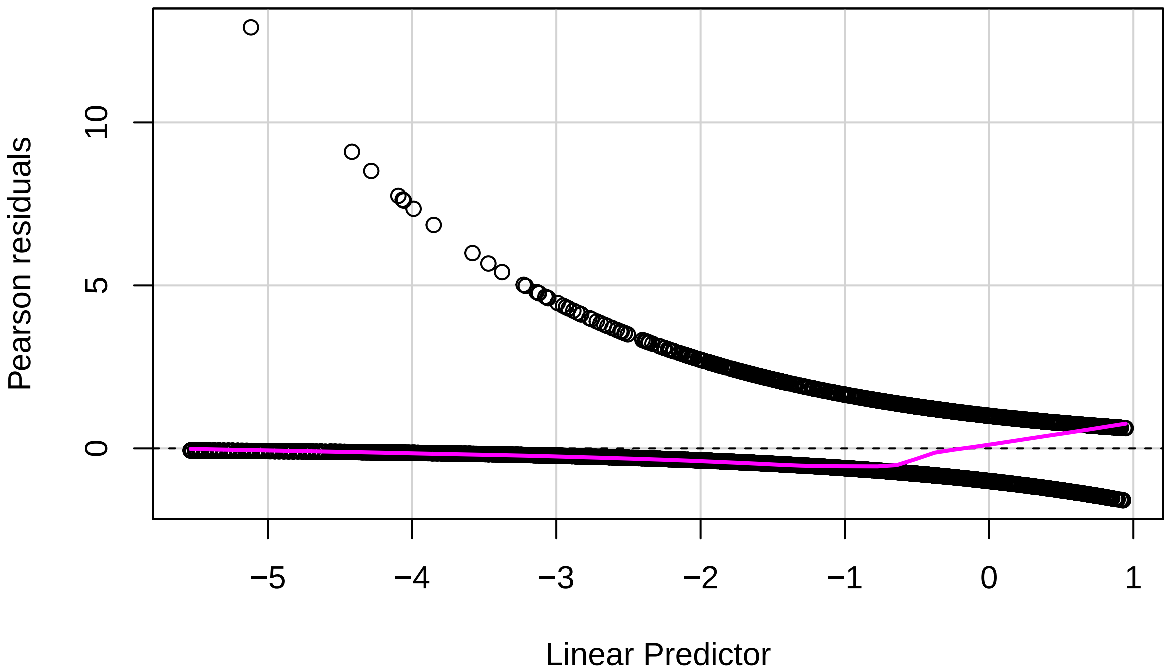
```r
anova( mod4, mod5,  test="Chisq") #significant
```

```
## Analysis of Deviance Table
##
## Model 1: Churn ~ tenure + MonthlyCharges + Contract
## Model 2: Churn ~ tenure + MonthlyCharges + Contract + InternetService
##   Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
## 1      4925     4292.2
## 2      4923     4240.1  2    52.12 4.811e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
vif(mod5)
```

```
##                      GVIF Df GVIF^(1/(2*Df))
## tenure           1.738643  1        1.318576
## MonthlyCharges   6.009378  1        2.451403
## Contract         1.450931  2        1.097518
## InternetService  5.338238  2        1.520021
```

```r
residualPlot(mod5)
```



StreamingMovies

```r
mod6 <- glm(Churn ~ tenure + MonthlyCharges + Contract + InternetService + StreamingMovies, data=train,
AIC(mod6) #4238.6 better
```

```
## [1] 4238.552
```

```r
anova( mod5, mod6,  test="Chisq") #significant
```

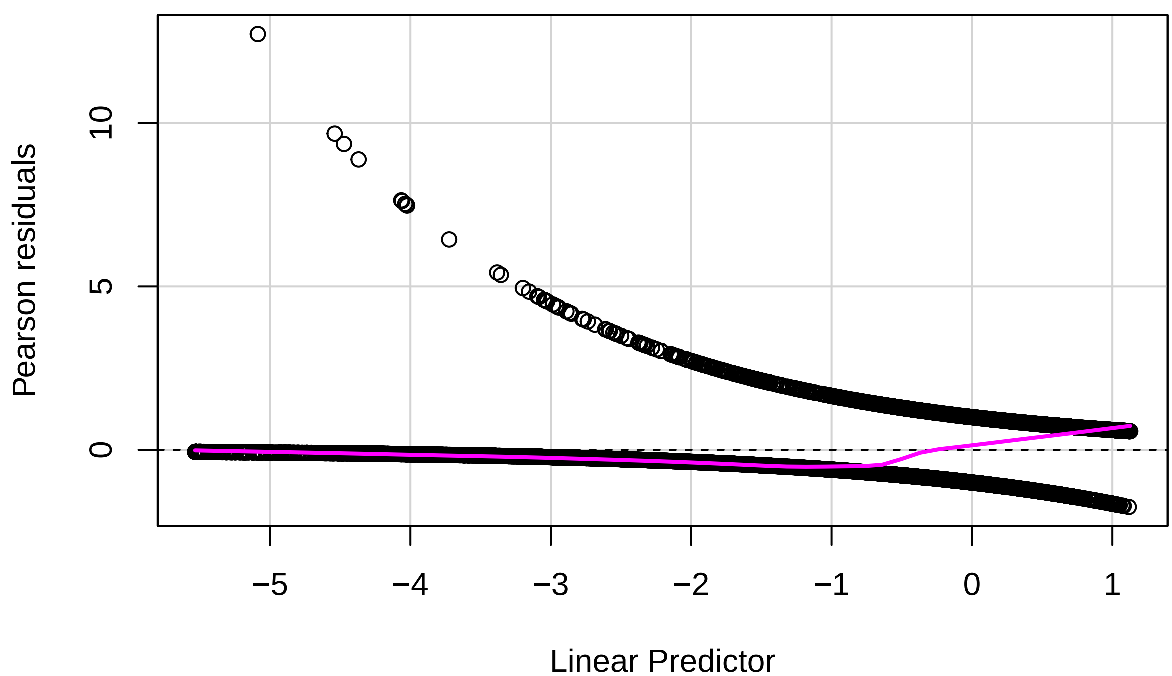
```
## Analysis of Deviance Table
##
## Model 1: Churn ~ tenure + MonthlyCharges + Contract + InternetService
## Model 2: Churn ~ tenure + MonthlyCharges + Contract + InternetService +
##     StreamingMovies
```

```
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1      4923     4240.1
## 2      4922     4222.6  1   17.563 2.78e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
vif(mod6)
```

```
##                     GVIF Df GVIF^(1/(2*Df))
## tenure          1.734387  1        1.316961
## MonthlyCharges  9.114445  1        3.019014
## Contract        1.447519  2        1.096872
## InternetService 6.680296  2        1.607677
## StreamingMovies 1.878425  1        1.370556
```

```r
residualPlot(mod6)
```



The model has improved but the VIF is becoming higher.

StreamingTV

```r
mod7 <- glm(Churn ~ tenure + MonthlyCharges + Contract + InternetService + StreamingMovies + StreamingT
AIC(mod7) #4213.5 better
```

```
## [1] 4213.55
```

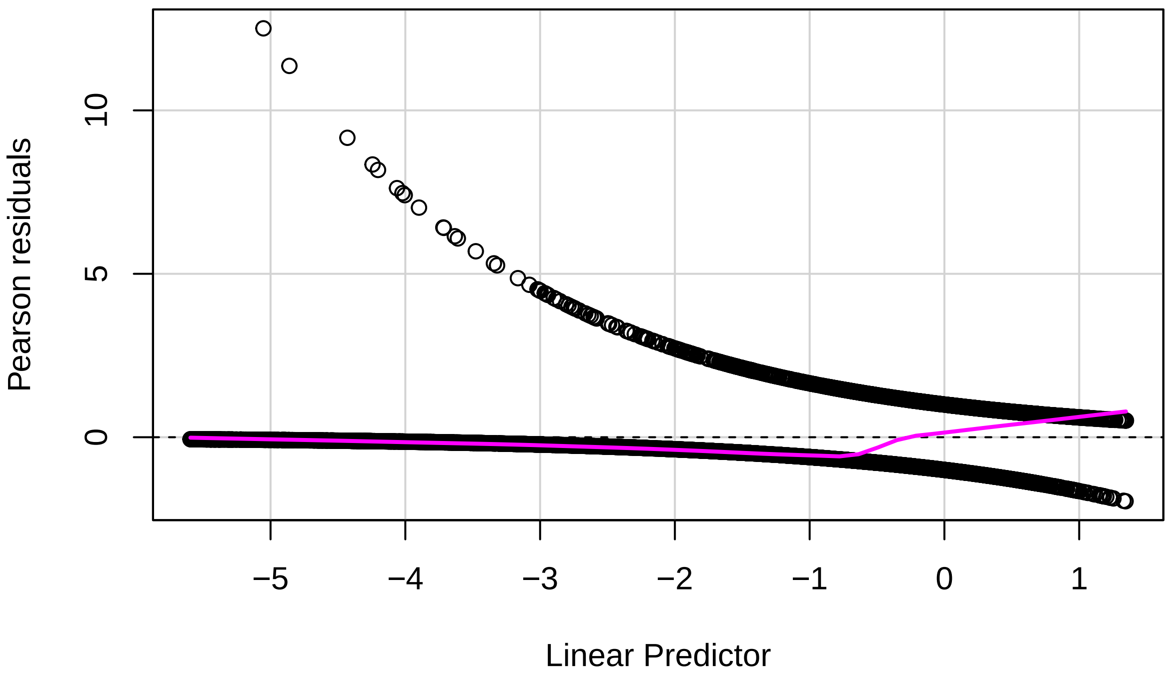
```r
anova( mod6, mod7,  test="Chisq") #significant
```

```
## Analysis of Deviance Table
##
## Model 1: Churn ~ tenure + MonthlyCharges + Contract + InternetService +
##     StreamingMovies
## Model 2: Churn ~ tenure + MonthlyCharges + Contract + InternetService +
##     StreamingMovies + StreamingTV
##   Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
## 1      4922     4222.6
## 2      4921     4195.5  1   27.002 2.033e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
vif(mod7)
```

```
##                      GVIF Df GVIF^(1/(2*Df))
## tenure           1.732269  1        1.316157
## MonthlyCharges  12.166459  1        3.488045
## Contract         1.443988  2        1.096203
## InternetService  7.954251  2        1.679383
## StreamingMovies  1.860165  1        1.363878
## StreamingTV      1.906895  1        1.380904
```

```r
residualPlot(mod7)
```



MonthlyCharges has a high VIF. We'll may need to add transformations or maybe discard this parameter. For now, we'll keep the parameters that we have been adding.

TechSupport

```
mod8 <- glm(Churn ~ tenure + MonthlyCharges + Contract + InternetService + StreamingMovies + StreamingT
#summary(mod8) #4208.3 better
AIC(mod8)
```

```
## [1] 4208.273
```

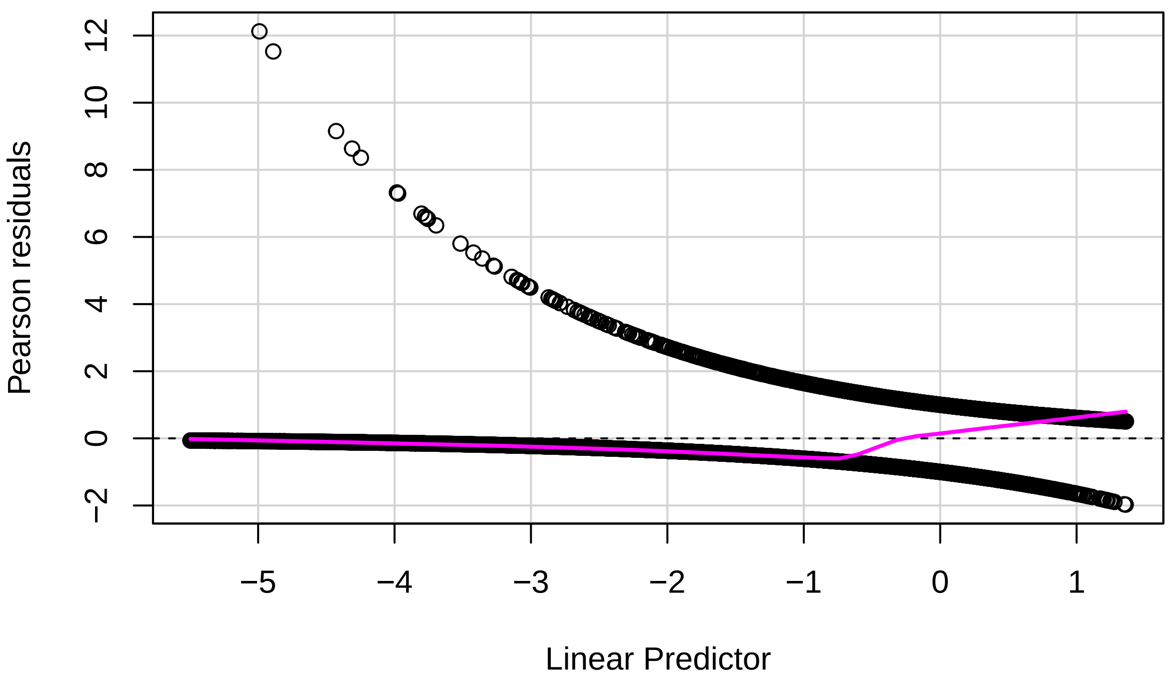
```
anova( mod7, mod8,  test="Chisq") #significant
```

```
## Analysis of Deviance Table
##
## Model 1: Churn ~ tenure + MonthlyCharges + Contract + InternetService +
##     StreamingMovies + StreamingTV
## Model 2: Churn ~ tenure + MonthlyCharges + Contract + InternetService +
##     StreamingMovies + StreamingTV + TechSupport
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1      4921     4195.5
## 2      4920     4188.3  1   7.2764 0.006987 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
vif(mod8)
```

```
##                       GVIF Df GVIF^(1/(2*Df))
## tenure            1.732344  1        1.316185
## MonthlyCharges   13.838376  1        3.719997
## Contract          1.475851  2        1.102201
## InternetService   9.342986  2        1.748322
## StreamingMovies   1.893830  1        1.376165
## StreamingTV       1.943568  1        1.394119
## TechSupport       1.294163  1        1.137613
```

```r
residualPlot(mod8)
```



Including *TechSupport* improves the model.

DeviceProtection

```r
mod9 <- glm(Churn ~ tenure + MonthlyCharges + Contract + InternetService + StreamingMovies + StreamingT
summary(mod9) #4209.3 worse
```

```
##
## Call:
## glm(formula = Churn ~ tenure + MonthlyCharges + Contract + InternetService +
```

```
##     StreamingMovies + StreamingTV + TechSupport + DeviceProtection,
##     family = binomial, data = train)
##
## Coefficients:
##                              Estimate Std. Error z value Pr(>|z|)
## (Intercept)                   0.20725    0.24332   0.852 0.394345
## tenure                       -0.03217    0.00250 -12.868  < 2e-16 ***
## MonthlyCharges               -0.01417    0.00558  -2.539 0.011129 *
## ContractOne year             -0.84846    0.12453  -6.813 9.54e-12 ***
## ContractTwo year             -1.71130    0.21068  -8.123 4.55e-16 ***
## InternetServiceFiber optic    1.49636    0.20259   7.386 1.51e-13 ***
## InternetServiceNo            -1.33473    0.19328  -6.906 5.00e-12 ***
## StreamingMoviesYes            0.41040    0.10661   3.850 0.000118 ***
## StreamingTVYes                0.51843    0.10817   4.793 1.64e-06 ***
## TechSupportYes               -0.27817    0.10447  -2.663 0.007751 **
## DeviceProtectionYes           0.09141    0.09477   0.965 0.334789
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 5694.2  on 4929  degrees of freedom
## Residual deviance: 4187.3  on 4919  degrees of freedom
## AIC: 4209.3
##
## Number of Fisher Scoring iterations: 6
```

```r
AIC(mod9)
```

```
## [1] 4209.343
```

```r
anova( mod8, mod9,  test="Chisq") #not significant
```

```
## Analysis of Deviance Table
##
## Model 1: Churn ~ tenure + MonthlyCharges + Contract + InternetService +
##     StreamingMovies + StreamingTV + TechSupport
## Model 2: Churn ~ tenure + MonthlyCharges + Contract + InternetService +
##     StreamingMovies + StreamingTV + TechSupport + DeviceProtection
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1      4920     4188.3
## 2      4919     4187.3  1  0.93092   0.3346
```

We don't add the parameter to the model. It does not improve it.

OnlineBackup

```r
mod10 <- glm(Churn ~ tenure + MonthlyCharges + Contract + InternetService + StreamingMovies + Streaming
AIC(mod10) #4209.6 worse
```

```
## [1] 4209.632
```

```r
anova( mod8, mod10,  test="Chisq") #not significant
```

```
## Analysis of Deviance Table
##
## Model 1: Churn ~ tenure + MonthlyCharges + Contract + InternetService +
##     StreamingMovies + StreamingTV + TechSupport
## Model 2: Churn ~ tenure + MonthlyCharges + Contract + InternetService +
##     StreamingMovies + StreamingTV + TechSupport + OnlineBackup
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1      4920     4188.3
## 2      4919     4187.6  1  0.64158   0.4231
```

We don't add the parameter to the model. It does not improve it.

OnlineSecurity

```r
mod11 <- glm(Churn ~ tenure + MonthlyCharges + Contract + InternetService + StreamingMovies + Streaming
AIC(mod11) #4199 better
```

```
## [1] 4198.953
```

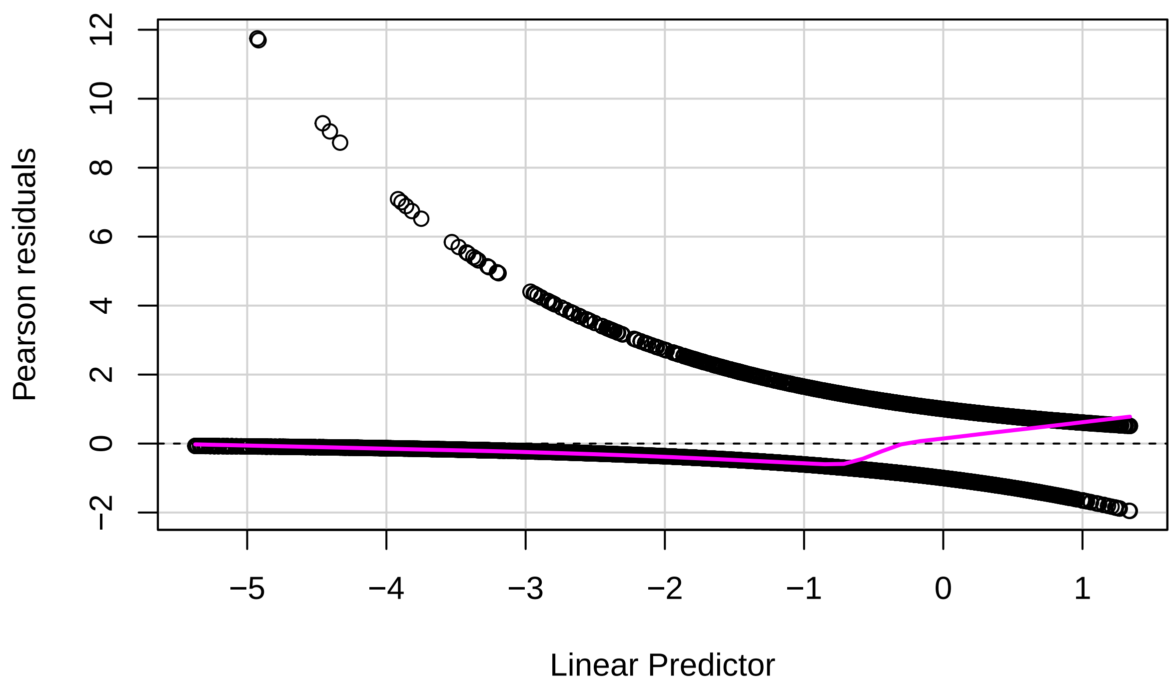
```r
anova( mod8, mod11,  test="Chisq") #significant
```

```
## Analysis of Deviance Table
##
## Model 1: Churn ~ tenure + MonthlyCharges + Contract + InternetService +
##     StreamingMovies + StreamingTV + TechSupport
## Model 2: Churn ~ tenure + MonthlyCharges + Contract + InternetService +
##     StreamingMovies + StreamingTV + TechSupport + OnlineSecurity
##   Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
## 1      4920     4188.3
## 2      4919     4177.0  1   11.321 0.0007665 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
vif(mod11)
```

```
##                      GVIF Df GVIF^(1/(2*Df))
## tenure           1.744624  1        1.320842
## MonthlyCharges  15.487373  1        3.935400
## Contract         1.492903  2        1.105371
## InternetService 10.866851  2        1.815624
## StreamingMovies  1.971177  1        1.403986
## StreamingTV      2.028530  1        1.424265
## TechSupport      1.296059  1        1.138446
## OnlineSecurity   1.242751  1        1.114787
```

```r
residualPlot(mod11)
```



We keep the parameter

PaperlessBilling

```r
mod12 <- glm(Churn ~ tenure + MonthlyCharges + Contract + InternetService + StreamingMovies + Streaming
summary(mod12) #4184.5 better
```

```
##
## Call:
## glm(formula = Churn ~ tenure + MonthlyCharges + Contract + InternetService +
##     StreamingMovies + StreamingTV + TechSupport + OnlineSecurity +
##     PaperlessBilling, family = binomial, data = train)
##
## Coefficients:
##                              Estimate Std. Error z value Pr(>|z|)
## (Intercept)                 -0.206715   0.251517  -0.822 0.411150
## tenure                      -0.031980   0.002512 -12.730  < 2e-16 ***
## MonthlyCharges              -0.006893   0.005737  -1.202 0.229554
## ContractOne year            -0.774511   0.125366  -6.178 6.49e-10 ***
## ContractTwo year            -1.575801   0.211901  -7.436 1.03e-13 ***
## InternetServiceFiber optic   1.162390   0.211629   5.493 3.96e-08 ***
```

```
## InternetServiceNo          -1.216241   0.195326  -6.227 4.76e-10 ***
## StreamingMoviesYes          0.328093   0.109142   3.006 0.002646 **
## StreamingTVYes              0.412453   0.111023   3.715 0.000203 ***
## TechSupportYes             -0.293252   0.105072  -2.791 0.005255 **
## OnlineSecurityYes          -0.325252   0.105781  -3.075 0.002107 **
## PaperlessBillingYes         0.354796   0.087670   4.047 5.19e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 5694.2  on 4929  degrees of freedom
## Residual deviance: 4160.5  on 4918  degrees of freedom
## AIC: 4184.5
##
## Number of Fisher Scoring iterations: 6
```

```r
AIC(mod12)
```

```
## [1] 4184.475
```

```r
anova( mod11, mod12,  test="Chisq") #significant
```

```
## Analysis of Deviance Table
##
## Model 1: Churn ~ tenure + MonthlyCharges + Contract + InternetService +
##     StreamingMovies + StreamingTV + TechSupport + OnlineSecurity
## Model 2: Churn ~ tenure + MonthlyCharges + Contract + InternetService +
##     StreamingMovies + StreamingTV + TechSupport + OnlineSecurity +
##     PaperlessBilling
##   Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
## 1      4919     4177.0
## 2      4918     4160.5  1   16.478 4.923e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
vif(mod12)
```

```
##                       GVIF Df GVIF^(1/(2*Df))
## tenure            1.760119  1        1.326695
## MonthlyCharges   15.519259  1        3.939449
## Contract          1.507661  2        1.108092
## InternetService  10.973792  2        1.820075
## StreamingMovies   1.970408  1        1.403712
## StreamingTV       2.035605  1        1.426746
## TechSupport       1.298079  1        1.139333
## OnlineSecurity    1.247294  1        1.116823
## PaperlessBilling  1.111928  1        1.054480
```

We keep the parameter

Dependents

```r
mod13 <- glm(Churn ~ tenure + MonthlyCharges + Contract + InternetService + StreamingMovies + Streaming
summary(mod13) #4177.2 better
```

```
##
## Call:
## glm(formula = Churn ~ tenure + MonthlyCharges + Contract + InternetService +
##     StreamingMovies + StreamingTV + TechSupport + OnlineSecurity +
##     PaperlessBilling + Dependents, family = binomial, data = train)
##
## Coefficients:
##                              Estimate Std. Error z value Pr(>|z|)
## (Intercept)                 -0.160331   0.252462  -0.635  0.52538
## tenure                      -0.031654   0.002520 -12.559  < 2e-16 ***
## MonthlyCharges              -0.006595   0.005749  -1.147  0.25137
## ContractOne year            -0.746604   0.125870  -5.932 3.00e-09 ***
## ContractTwo year            -1.536143   0.212595  -7.226 4.99e-13 ***
## InternetServiceFiber optic   1.133942   0.212173   5.344 9.07e-08 ***
## InternetServiceNo           -1.193933   0.195766  -6.099 1.07e-09 ***
## StreamingMoviesYes           0.317729   0.109348   2.906  0.00366 **
## StreamingTVYes               0.412210   0.111213   3.706  0.00021 ***
## TechSupportYes              -0.287327   0.105193  -2.731  0.00631 **
## OnlineSecurityYes           -0.317077   0.105920  -2.994  0.00276 **
## PaperlessBillingYes          0.351625   0.087803   4.005 6.21e-05 ***
## DependentsYes               -0.291003   0.096298  -3.022  0.00251 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 5694.2  on 4929  degrees of freedom
## Residual deviance: 4151.2  on 4917  degrees of freedom
## AIC: 4177.2
##
## Number of Fisher Scoring iterations: 6
```

```r
AIC(mod13)
```

```
## [1] 4177.206
```

```r
anova( mod12, mod13,  test="Chisq") #significant
```

```
## Analysis of Deviance Table
##
## Model 1: Churn ~ tenure + MonthlyCharges + Contract + InternetService +
##     StreamingMovies + StreamingTV + TechSupport + OnlineSecurity +
##     PaperlessBilling
## Model 2: Churn ~ tenure + MonthlyCharges + Contract + InternetService +
##     StreamingMovies + StreamingTV + TechSupport + OnlineSecurity +
##     PaperlessBilling + Dependents
```

```
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1      4918     4160.5
## 2      4917     4151.2  1   9.2692  0.00233 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
vif(mod13)
```

```
##                       GVIF Df GVIF^(1/(2*Df))
## tenure            1.773404  1        1.331692
## MonthlyCharges   15.562560  1        3.944941
## Contract          1.522708  2        1.110847
## InternetService  10.992492  2        1.820849
## StreamingMovies   1.973305  1        1.404744
## StreamingTV       2.037770  1        1.427505
## TechSupport       1.299374  1        1.139901
## OnlineSecurity    1.247956  1        1.117120
## PaperlessBilling  1.112626  1        1.054811
## Dependents        1.027601  1        1.013706
```

We keep the parameter

MultipleLines

```r
mod14 <- glm(Churn ~ tenure + MonthlyCharges + Contract + InternetService + StreamingMovies + Streaming
AIC(mod14) #4162.2 better
```

```
## [1] 4162.18
```

```r
anova( mod13, mod14,  test="Chisq") #significant
```

```
## Analysis of Deviance Table
##
## Model 1: Churn ~ tenure + MonthlyCharges + Contract + InternetService +
##     StreamingMovies + StreamingTV + TechSupport + OnlineSecurity +
##     PaperlessBilling + Dependents
## Model 2: Churn ~ tenure + MonthlyCharges + Contract + InternetService +
##     StreamingMovies + StreamingTV + TechSupport + OnlineSecurity +
##     PaperlessBilling + Dependents + MultipleLines
##   Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
## 1      4917     4151.2
## 2      4916     4134.2  1   17.026 3.688e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
vif(mod14)
```

```
##                       GVIF Df GVIF^(1/(2*Df))
## tenure            1.860860  1        1.364133
```

```
## MonthlyCharges   19.785122  1        4.448047
## Contract          1.529039  2        1.112000
## InternetService  12.562934  2        1.882664
## StreamingMovies   2.104685  1        1.450753
## StreamingTV       2.150829  1        1.466570
## TechSupport       1.346109  1        1.160219
## OnlineSecurity    1.283323  1        1.132838
## PaperlessBilling  1.113149  1        1.055059
## Dependents        1.028391  1        1.014096
## MultipleLines     1.749163  1        1.322559
```

We keep the parameter

SeniorCitizen

```r
mod15 <- glm(Churn ~ tenure + MonthlyCharges + Contract + InternetService + StreamingMovies + Streaming
AIC(mod15) #4155.7 better
```

```
## [1] 4155.702
```

```r
anova( mod14, mod15,  test="Chisq") #significant
```

```
## Analysis of Deviance Table
##
## Model 1: Churn ~ tenure + MonthlyCharges + Contract + InternetService +
##     StreamingMovies + StreamingTV + TechSupport + OnlineSecurity +
##     PaperlessBilling + Dependents + MultipleLines
## Model 2: Churn ~ tenure + MonthlyCharges + Contract + InternetService +
##     StreamingMovies + StreamingTV + TechSupport + OnlineSecurity +
##     PaperlessBilling + Dependents + MultipleLines + SeniorCitizen
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1      4916     4134.2
## 2      4915     4125.7  1   8.4782 0.003594 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
vif(mod15)
```

```
##                       GVIF Df GVIF^(1/(2*Df))
## tenure            1.889241  1        1.374497
## MonthlyCharges   19.790331  1        4.448632
## Contract          1.536772  2        1.113403
## InternetService  12.635139  2        1.885363
## StreamingMovies   2.104216  1        1.450592
## StreamingTV       2.148543  1        1.465791
## TechSupport       1.353673  1        1.163474
## OnlineSecurity    1.286526  1        1.134251
## PaperlessBilling  1.114284  1        1.055597
## Dependents        1.056349  1        1.027789
## MultipleLines     1.752169  1        1.323695
## SeniorCitizen     1.113813  1        1.055374
```

We keep the parameter

Partner

```
mod16 <- glm(Churn ~ tenure + MonthlyCharges + Contract + InternetService + StreamingMovies + Streaming
AIC(mod16) #4157.7 worse
```

```
## [1] 4157.677
```

```
anova( mod15, mod16,  test="Chisq") #not significant
```

```
## Analysis of Deviance Table
##
## Model 1: Churn ~ tenure + MonthlyCharges + Contract + InternetService +
##     StreamingMovies + StreamingTV + TechSupport + OnlineSecurity +
##     PaperlessBilling + Dependents + MultipleLines + SeniorCitizen
## Model 2: Churn ~ tenure + MonthlyCharges + Contract + InternetService +
##     StreamingMovies + StreamingTV + TechSupport + OnlineSecurity +
##     PaperlessBilling + Dependents + MultipleLines + SeniorCitizen +
##     Partner
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1      4915     4125.7
## 2      4914     4125.7  1 0.024971   0.8744
```

We don't keep the parameter

PaymentMethod

```
mod17 <- glm(Churn ~ tenure + MonthlyCharges + Contract + InternetService + StreamingMovies + Streaming
AIC(mod17) #4139.4 better
```

```
## [1] 4139.434
```

```
anova( mod15, mod17,  test="Chisq") #significant
```

```
## Analysis of Deviance Table
##
## Model 1: Churn ~ tenure + MonthlyCharges + Contract + InternetService +
##     StreamingMovies + StreamingTV + TechSupport + OnlineSecurity +
##     PaperlessBilling + Dependents + MultipleLines + SeniorCitizen
## Model 2: Churn ~ tenure + MonthlyCharges + Contract + InternetService +
##     StreamingMovies + StreamingTV + TechSupport + OnlineSecurity +
##     PaperlessBilling + Dependents + MultipleLines + SeniorCitizen +
##     PaymentMethod
##   Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
## 1      4915     4125.7
## 2      4912     4103.4  3   22.269 5.735e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
vif(mod17)
```

```
##                      GVIF Df GVIF^(1/(2*Df))
## tenure            1.963626  1        1.401295
## MonthlyCharges   19.895259  1        4.460410
## Contract          1.543913  2        1.114694
## InternetService  13.046889  2        1.900539
## StreamingMovies   2.110866  1        1.452882
## StreamingTV       2.164001  1        1.471054
## TechSupport       1.357356  1        1.165056
## OnlineSecurity    1.291867  1        1.136603
## PaperlessBilling  1.120742  1        1.058651
## Dependents        1.057502  1        1.028349
## MultipleLines     1.753352  1        1.324142
## SeniorCitizen     1.116591  1        1.056689
## PaymentMethod     1.332467  3        1.049001
```

PhoneService

```r
mod18 <- glm(Churn ~ tenure + MonthlyCharges + Contract + InternetService + StreamingMovies + Streaming
AIC(mod18)#4139.4 it does not change anything
```

```
## [1] 4139.379
```

```r
anova( mod17, mod18,  test="Chisq") #not significant
```

```
## Analysis of Deviance Table
##
## Model 1: Churn ~ tenure + MonthlyCharges + Contract + InternetService +
##     StreamingMovies + StreamingTV + TechSupport + OnlineSecurity +
##     PaperlessBilling + Dependents + MultipleLines + SeniorCitizen +
##     PaymentMethod
## Model 2: Churn ~ tenure + MonthlyCharges + Contract + InternetService +
##     StreamingMovies + StreamingTV + TechSupport + OnlineSecurity +
##     PaperlessBilling + Dependents + MultipleLines + SeniorCitizen +
##     PaymentMethod + PhoneService
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1      4912     4103.4
## 2      4911     4101.4  1    2.055   0.1517
```

We don't include the parameter

Inlfuential data

We check the influential data after including the categorical variables

```r
infl_2 <- influence.measures(mod17)

sum(residuals(mod17,'deviance')^2)
```

```
## [1] 4103.434
```

```r
sum(residuals(mod17,'pearson')^2)
```

```
## [1] 4919.679
```

```r
influential_indices_2 <- which(infl_2$is.inf == TRUE)
length(influential_indices_2)
```

```
## [1] 98
```

```r
length(train$customerID)
```

```
## [1] 4930
```

The influential data has reduced until 98 tuples.

Interactions

We need to search for interactions. Possible interactions:

- Dependents and Multiple Lines

```r
mod19 <- glm(Churn ~ tenure + MonthlyCharges + Contract + InternetService + StreamingMovies + Streaming
 #4140.4 worse
AIC(mod19)
```

```
## [1] 4140.355
```

```r
anova( mod17, mod19,  test="Chisq") #not significant
```

```
## Analysis of Deviance Table
##
## Model 1: Churn ~ tenure + MonthlyCharges + Contract + InternetService +
##     StreamingMovies + StreamingTV + TechSupport + OnlineSecurity +
##     PaperlessBilling + Dependents + MultipleLines + SeniorCitizen +
##     PaymentMethod
## Model 2: Churn ~ tenure + MonthlyCharges + Contract + InternetService +
##     StreamingMovies + StreamingTV + TechSupport + OnlineSecurity +
##     PaperlessBilling + Dependents * MultipleLines + SeniorCitizen +
##     PaymentMethod
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1      4912     4103.4
## 2      4911     4102.4  1   1.0787    0.299
```

We don't include the interaction since it is not significative

- MonthlyCharges and InternetService

```r
mod20 <- glm(Churn ~ tenure + InternetService * MonthlyCharges + Contract + StreamingMovies + Streaming
AIC(mod20) #4133.7 better
```

```
## [1] 4133.664
```

```r
anova( mod17, mod20,  test="Chisq") #significant
```

```
## Analysis of Deviance Table
##
## Model 1: Churn ~ tenure + MonthlyCharges + Contract + InternetService +
##     StreamingMovies + StreamingTV + TechSupport + OnlineSecurity +
##     PaperlessBilling + Dependents + MultipleLines + SeniorCitizen +
##     PaymentMethod
## Model 2: Churn ~ tenure + InternetService * MonthlyCharges + Contract +
##     StreamingMovies + StreamingTV + TechSupport + OnlineSecurity +
##     PaperlessBilling + Dependents + MultipleLines + SeniorCitizen +
##     PaymentMethod
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1      4912     4103.4
## 2      4910     4093.7  2   9.7694 0.007561 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
vif(mod20)
```

```
## there are higher-order terms (interactions) in this model
## consider setting type = 'predictor'; see ?vif
```

```
##                                        GVIF Df GVIF^(1/(2*Df))
## tenure                             2.079881  1        1.442179
## InternetService                 9738.807709  2        9.934052
## MonthlyCharges                    21.386127  1        4.624514
## Contract                           1.550405  2        1.115864
## StreamingMovies                    2.374759  1        1.541025
## StreamingTV                        2.416906  1        1.554640
## TechSupport                        1.374225  1        1.172273
## OnlineSecurity                     1.300790  1        1.140522
## PaperlessBilling                   1.124965  1        1.060644
## Dependents                         1.056690  1        1.027954
## MultipleLines                      1.897486  1        1.377493
## SeniorCitizen                      1.115802  1        1.056315
## PaymentMethod                      1.346214  3        1.050797
## InternetService:MonthlyCharges 11466.767397  2       10.348091
```

We improved the model but multicolinearity worse ??? ???DUBTE pk apareixen les variables com si estiguessin tambe per separat i no nomes com una interaccio? A l'anterior model ens surt nomes la interaccio.

- SeniorCitizen and PaymentMethod

```r
mod21 <- glm(Churn ~ tenure + InternetService + MonthlyCharges + Contract + StreamingMovies + Streaming
AIC(mod21) #4133 better and also better than mod20
```

```
## [1] 4133.038
```

```r
anova( mod17, mod21,  test="Chisq") #significant
```

```
## Analysis of Deviance Table
##
## Model 1: Churn ~ tenure + MonthlyCharges + Contract + InternetService +
##     StreamingMovies + StreamingTV + TechSupport + OnlineSecurity +
##     PaperlessBilling + Dependents + MultipleLines + SeniorCitizen +
##     PaymentMethod
## Model 2: Churn ~ tenure + InternetService + MonthlyCharges + Contract +
##     StreamingMovies + StreamingTV + TechSupport + OnlineSecurity +
##     PaperlessBilling + Dependents + MultipleLines + SeniorCitizen *
##     PaymentMethod
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1      4912     4103.4
## 2      4909     4091.0  3   12.396 0.006144 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
anova( mod20, mod21,  test="Chisq") #not significant
```

```
## Analysis of Deviance Table
##
## Model 1: Churn ~ tenure + InternetService * MonthlyCharges + Contract +
##     StreamingMovies + StreamingTV + TechSupport + OnlineSecurity +
##     PaperlessBilling + Dependents + MultipleLines + SeniorCitizen +
##     PaymentMethod
## Model 2: Churn ~ tenure + InternetService + MonthlyCharges + Contract +
##     StreamingMovies + StreamingTV + TechSupport + OnlineSecurity +
##     PaperlessBilling + Dependents + MultipleLines + SeniorCitizen *
##     PaymentMethod
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1      4910     4093.7
## 2      4909     4091.0  1   2.6261   0.1051
```

```r
vif(mod21) #better multicollinearity
```

```
## there are higher-order terms (interactions) in this model
## consider setting type = 'predictor'; see ?vif
```

```
##                                  GVIF Df GVIF^(1/(2*Df))
## tenure                       1.973899  1        1.404955
## InternetService             13.127210  2        1.903457
## MonthlyCharges              19.972402  1        4.469049
## Contract                     1.548154  2        1.115459
## StreamingMovies              2.114568  1        1.454155
```

```
## StreamingTV                   2.168544  1        1.472598
## TechSupport                   1.359278  1        1.165881
## OnlineSecurity                1.292280  1        1.136785
## PaperlessBilling              1.120630  1        1.058598
## Dependents                    1.058287  1        1.028731
## MultipleLines                 1.759302  1        1.326387
## SeniorCitizen                 6.564344  1        2.562098
## PaymentMethod                 2.413718  3        1.158193
## SeniorCitizen:PaymentMethod 10.225907  3        1.473274
```

```
mod22 <- glm(Churn ~ tenure + InternetService * MonthlyCharges + Contract + StreamingMovies + Streaming
AIC(mod22) #4126.8 better
```

```
## [1] 4126.835
```

```
anova( mod21, mod22,  test="Chisq") #significant
```

```
## Analysis of Deviance Table
##
## Model 1: Churn ~ tenure + InternetService + MonthlyCharges + Contract +
##     StreamingMovies + StreamingTV + TechSupport + OnlineSecurity +
##     PaperlessBilling + Dependents + MultipleLines + SeniorCitizen *
##     PaymentMethod
## Model 2: Churn ~ tenure + InternetService * MonthlyCharges + Contract +
##     StreamingMovies + StreamingTV + TechSupport + OnlineSecurity +
##     PaperlessBilling + Dependents + MultipleLines + SeniorCitizen *
##     PaymentMethod
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1      4909     4091.0
## 2      4907     4080.8  2   10.203 0.006088 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
anova( mod20, mod22,  test="Chisq") #significant
```

```
## Analysis of Deviance Table
##
## Model 1: Churn ~ tenure + InternetService * MonthlyCharges + Contract +
##     StreamingMovies + StreamingTV + TechSupport + OnlineSecurity +
##     PaperlessBilling + Dependents + MultipleLines + SeniorCitizen +
##     PaymentMethod
## Model 2: Churn ~ tenure + InternetService * MonthlyCharges + Contract +
##     StreamingMovies + StreamingTV + TechSupport + OnlineSecurity +
##     PaperlessBilling + Dependents + MultipleLines + SeniorCitizen *
##     PaymentMethod
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1      4910     4093.7
## 2      4907     4080.8  3   12.829 0.005021 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
vif(mod22)
```

```
## there are higher-order terms (interactions) in this model
## consider setting type = 'predictor'; see ?vif
```

```
##                                     GVIF Df GVIF^(1/(2*Df))
## tenure                          2.092433  1        1.446525
## InternetService              9747.368394  2        9.936235
## MonthlyCharges                 21.496711  1        4.636455
## Contract                        1.554570  2        1.116613
## StreamingMovies                 2.379677  1        1.542620
## StreamingTV                     2.420865  1        1.555913
## TechSupport                     1.375906  1        1.172990
## OnlineSecurity                  1.300799  1        1.140526
## PaperlessBilling                1.124887  1        1.060607
## Dependents                      1.057390  1        1.028295
## MultipleLines                   1.905667  1        1.380459
## SeniorCitizen                   6.580622  1        2.565272
## PaymentMethod                   2.445976  3        1.160759
## InternetService:MonthlyCharges 11487.448457 2      10.352754
## SeniorCitizen:PaymentMethod    10.277317  3        1.474506
```

Having both interactions improves the model but VIF gets worse. The best model is with SeniorCitizen and PaymentMethod interaction (mod21)

###Second Order variable

```r
mod23 <- glm(Churn ~ tenure + I(tenure^2) + InternetService + MonthlyCharges + Contract + StreamingMovi
AIC(mod23) #4088.4 better
```

```
## [1] 4088.366
```

```r
anova( mod21, mod23,  test="Chisq") #significant
```

```
## Analysis of Deviance Table
##
## Model 1: Churn ~ tenure + InternetService + MonthlyCharges + Contract +
##     StreamingMovies + StreamingTV + TechSupport + OnlineSecurity +
##     PaperlessBilling + Dependents + MultipleLines + SeniorCitizen *
##     PaymentMethod
## Model 2: Churn ~ tenure + I(tenure^2) + InternetService + MonthlyCharges +
##     Contract + StreamingMovies + StreamingTV + TechSupport +
##     OnlineSecurity + PaperlessBilling + Dependents + MultipleLines +
##     SeniorCitizen * PaymentMethod
##   Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
## 1      4909     4091.0
## 2      4908     4044.4  1   46.672 8.392e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
vif(mod23)
```

```
## there are higher-order terms (interactions) in this model
## consider setting type = 'predictor'; see ?vif

##                                  GVIF Df GVIF^(1/(2*Df))
## tenure                      15.110913  1        3.887276
## I(tenure^2)                 14.413478  1        3.796509
## InternetService             13.143356  2        1.904042
## MonthlyCharges              20.658589  1        4.545172
## Contract                     1.830861  2        1.163225
## StreamingMovies              2.155609  1        1.468199
## StreamingTV                  2.220993  1        1.490300
## TechSupport                  1.373947  1        1.172155
## OnlineSecurity               1.306102  1        1.142848
## PaperlessBilling             1.124076  1        1.060225
## Dependents                   1.060211  1        1.029666
## MultipleLines                1.824384  1        1.350697
## SeniorCitizen                6.421969  1        2.534160
## PaymentMethod                2.503172  3        1.165239
## SeniorCitizen:PaymentMethod 10.118072  3        1.470674
```

```r
mod23.1 <- glm(Churn ~ tenure + I(tenure^2) + InternetService + Contract + StreamingMovies + StreamingTV
AIC(mod23.1) #4093.9 worse
```

```
## [1] 4093.873
```

```r
anova( mod23, mod23.1,  test="Chisq") #significant
```

```
## Analysis of Deviance Table
##
## Model 1: Churn ~ tenure + I(tenure^2) + InternetService + MonthlyCharges +
##     Contract + StreamingMovies + StreamingTV + TechSupport +
##     OnlineSecurity + PaperlessBilling + Dependents + MultipleLines +
##     SeniorCitizen * PaymentMethod
## Model 2: Churn ~ tenure + I(tenure^2) + InternetService + Contract + StreamingMovies +
##     StreamingTV + TechSupport + OnlineSecurity + PaperlessBilling +
##     Dependents + MultipleLines + SeniorCitizen * PaymentMethod
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1      4908     4044.4
## 2      4909     4051.9 -1  -7.5068 0.006147 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
vif(mod23.1) #better vif
```

```
## there are higher-order terms (interactions) in this model
## consider setting type = 'predictor'; see ?vif

##                                  GVIF Df GVIF^(1/(2*Df))
```

```
## tenure                       15.094283  1        3.885136
## I(tenure^2)                  14.395726  1        3.794170
## InternetService               1.753349  2        1.150713
## Contract                      1.832458  2        1.163479
## StreamingMovies               1.439408  1        1.199753
## StreamingTV                   1.476549  1        1.215133
## TechSupport                   1.176693  1        1.084755
## OnlineSecurity                1.145979  1        1.070504
## PaperlessBilling              1.123469  1        1.059938
## Dependents                    1.059050  1        1.029102
## MultipleLines                 1.406194  1        1.185831
## SeniorCitizen                 6.416355  1        2.533052
## PaymentMethod                 2.500773  3        1.165053
## SeniorCitizen:PaymentMethod  10.110887  3        1.470499
```

Removing *MonthlyCharges* from the model is getting a bit worse the AIC but the change is significant and it improves the VIF.

For improving the multicollinearity we add log in *tenure*

```r
mod23.4 <- glm(Churn ~ log(tenure + 0.01) + I(tenure^2) + InternetService + Contract + StreamingMovies +
AIC(mod23.4) #4059.53
```

```
## [1] 4059.531
```

```r
vif(mod23.4)
```

```
## there are higher-order terms (interactions) in this model
## consider setting type = 'predictor'; see ?vif
```

```
##                                  GVIF Df GVIF^(1/(2*Df))
## log(tenure + 0.01)           2.500964  1        1.581444
## I(tenure^2)                  2.794150  1        1.671571
## InternetService              1.770563  2        1.153527
## Contract                     1.731667  2        1.147139
## StreamingMovies              1.429558  1        1.195641
## StreamingTV                  1.458661  1        1.207750
## TechSupport                  1.172948  1        1.083027
## OnlineSecurity               1.140765  1        1.068066
## PaperlessBilling             1.125341  1        1.060821
## Dependents                   1.057858  1        1.028522
## MultipleLines                1.385364  1        1.177015
## SeniorCitizen                6.404190  1        2.530650
## PaymentMethod                2.532835  3        1.167529
## SeniorCitizen:PaymentMethod 10.154436  3        1.471553
```

We keep this last model.

Inlfuential data

We check the influential data after including the interactions and the second order variables.

```r
infl_3 <- influence.measures(mod23.4)

sum(residuals(mod23.4,'deviance')^2)
```

```
## [1] 4017.531
```

```r
sum(residuals(mod23.4,'pearson')^2)
```

```
## [1] 4952.141
```

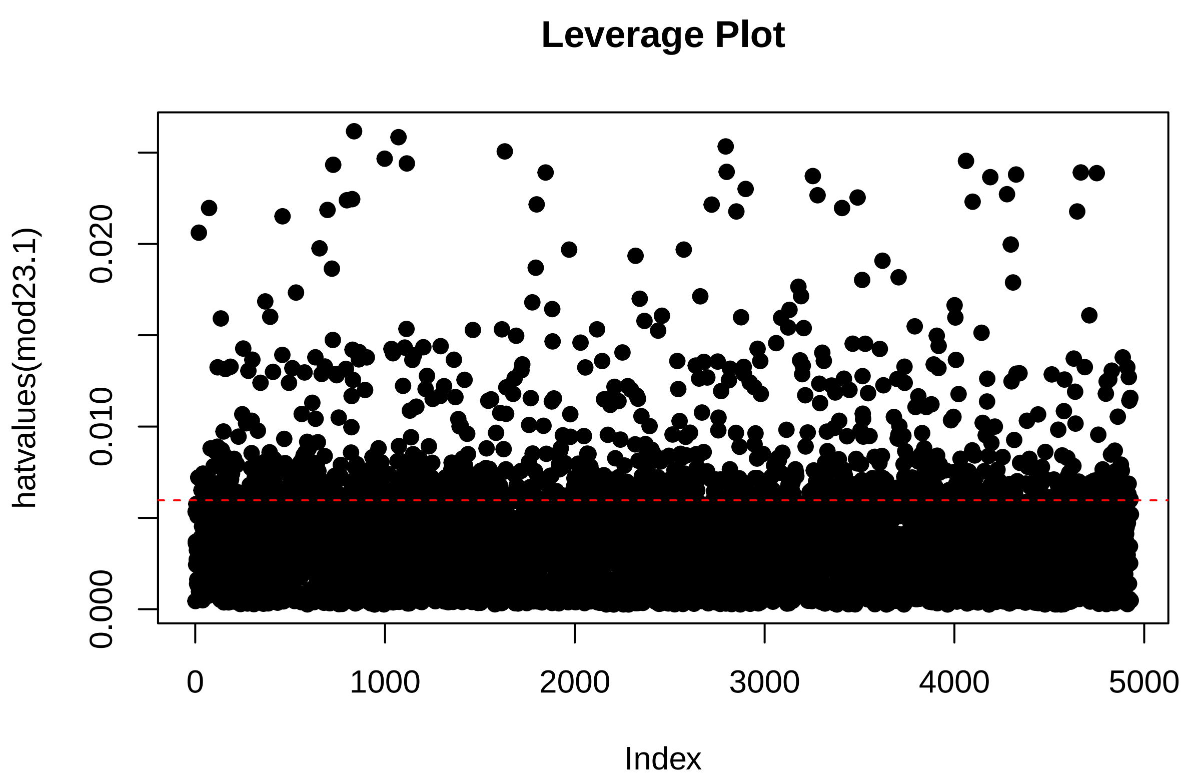
```r
influential_indices_3 <- which(infl_3$is.inf == TRUE)
length(influential_indices_3)
```

```
## [1] 399
```

```r
length(train$customerID)
```

```
## [1] 4930
```

```r
#Leverage values
plot(hatvalues(mod23.1), pch = 19, main = "Leverage Plot")
abline(h = 2 * ncol(model.matrix(mod23.1))/length(df$customerID), col = "red", lty = 2)
```



We have more influential data than before, 399 tuples.

Predictions

```
#selecting the parameters that we have in the model

test_data <- test[c(3,5,6,8,9,10,13,14,15,16,17,18)]

pred_prob <- predict(mod23.4, newdata = test_data, type="response")

churn_pred<- ifelse(pred_prob>0.5,"Yes","No")

table(churn_pred)
```

```
## churn_pred
##   No  Yes
## 1677  436
```

```
table(test$Churn)
```

```
##
##   No  Yes
## 1547  566
```

```
#Confusion table

tt <- table(churn_pred, test$Churn);tt
```

```
##
## churn_pred   No  Yes
##        No  1409  268
##        Yes  138  298
```

```
100*sum(diag(tt))/sum(tt) #80.79
```

```
## [1] 80.78561
```

The accuracy of our model is good, it is 80.79.

```
roc_curve <- roc(test$Churn, pred_prob)
```

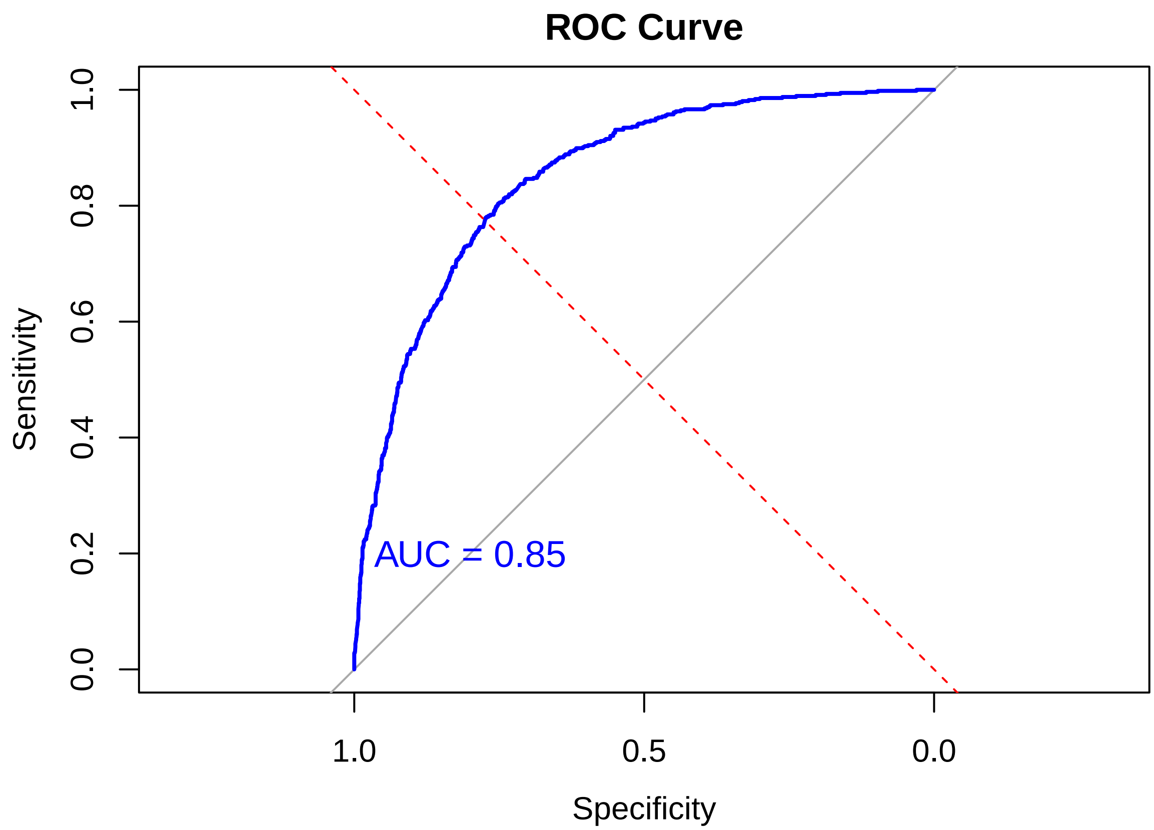
```
## Setting levels: control = No, case = Yes

## Setting direction: controls < cases
```

```
# Plot the ROC curve
plot(roc_curve, main = "ROC Curve", col = "blue", lwd = 2)

# Add diagonal reference line for comparison
abline(a = 0, b = 1, lty = 2, col = "red")

# Add AUC (Area Under the Curve) value to the plot
text(0.8, 0.2, paste("AUC =", round(auc(roc_curve), 2)), col = "blue", cex = 1.2)
```



Our Area Under the Curve for ROC curve is 0.85 so it is high

Our final model is

```
coef(mod23.4)
```

```
##                            (Intercept)
##                           5.747068e-02
##                      log(tenure + 0.01)
##                          -5.438358e-01
##                            I(tenure^2)
##                          -4.494426e-05
##            InternetServiceFiber optic
##                           7.544949e-01
##                      InternetServiceNo
##                          -9.744106e-01
##                       ContractOne year
##                          -7.534039e-01
##                       ContractTwo year
##                          -1.895286e+00
##                     StreamingMoviesYes
##                           2.624637e-01
##                         StreamingTVYes
##                           3.305712e-01
##                         TechSupportYes
##                          -2.174029e-01
```

```
##                                  OnlineSecurityYes
##                                      -2.801188e-01
##                                PaperlessBillingYes
##                                       3.294340e-01
##                                      DependentsYes
##                                      -2.300625e-01
##                                   MultipleLinesYes
##                                       3.244615e-01
##                                     SeniorCitizen1
##                                      -1.540301e-01
##                 PaymentMethodCredit card (automatic)
##                                      -2.543356e-01
##                        PaymentMethodElectronic check
##                                       2.736901e-01
##                            PaymentMethodMailed check
##                                      -2.447431e-01
## SeniorCitizen1:PaymentMethodCredit card (automatic)
##                                       8.653999e-01
##        SeniorCitizen1:PaymentMethodElectronic check
##                                       2.843971e-01
##            SeniorCitizen1:PaymentMethodMailed check
##                                       1.101151e+00
```

$Y = -0.58 - 0.08 tenure + 0.0007 tenure^2 + 0.75 InternetServiceFiberoptic - 0.92 InternetServiceNo - 0.72 ContractOneyear - 1$

Això ho podem posar a l'annex i deixem els comentaris al report

Univariate

```r
names(train)
```

```
##  [1] "customerID"       "gender"           "SeniorCitizen"    "Partner"
##  [5] "Dependents"       "tenure"           "PhoneService"     "MultipleLines"
##  [9] "InternetService"  "OnlineSecurity"   "OnlineBackup"     "DeviceProtection"
## [13] "TechSupport"      "StreamingTV"      "StreamingMovies"  "Contract"
## [17] "PaperlessBilling" "PaymentMethod"    "MonthlyCharges"   "TotalCharges"
## [21] "Churn"
```

```r
mod <- glm(Churn ~ gender, data=train, family=binomial)
summary(mod)
```

```
##
## Call:
## glm(formula = Churn ~ gender, family = binomial, data = train)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.00637    0.04542 -22.158   <2e-16 ***
## genderMale  -0.03499    0.06460  -0.542    0.588
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 5694.2  on 4929  degrees of freedom
## Residual deviance: 5693.9  on 4928  degrees of freedom
## AIC: 5697.9
##
## Number of Fisher Scoring iterations: 4
```

```
mod2 <- glm(Churn ~ SeniorCitizen, data=train, family=binomial)
summary(mod2)
```

```
##
## Call:
## glm(formula = Churn ~ SeniorCitizen, family = binomial, data = train)
##
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)    -1.19026    0.03682  -32.33   <2e-16 ***
## SeniorCitizen1  0.88226    0.08027   10.99   <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: 5694.2  on 4929  degrees of freedom
## Residual deviance: 5577.9  on 4928  degrees of freedom
## AIC: 5581.9
##
## Number of Fisher Scoring iterations: 4
```

```
mod3 <- glm(Churn ~ Partner, data=train, family=binomial)
summary(mod3)
```

```
##
## Call:
## glm(formula = Churn ~ Partner, family = binomial, data = train)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.70909    0.04215  -16.82   <2e-16 ***
## PartnerYes  -0.71326    0.06676  -10.68   <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: 5694.2  on 4929  degrees of freedom
## Residual deviance: 5576.5  on 4928  degrees of freedom
## AIC: 5580.5
##
## Number of Fisher Scoring iterations: 4
```

```r
mod4 <- glm(Churn ~ Dependents, data=train, family=binomial)
summary(mod4)
```

```
##
## Call:
## glm(formula = Churn ~ Dependents, family = binomial, data = train)
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)   -0.78158    0.03662  -21.34   <2e-16 ***
## DependentsYes -0.97564    0.08228  -11.86   <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: 5694.2  on 4929  degrees of freedom
## Residual deviance: 5534.9  on 4928  degrees of freedom
## AIC: 5538.9
##
## Number of Fisher Scoring iterations: 4
```

```r
mod5 <- glm(Churn ~ tenure, data=train, family=binomial)
summary(mod5)
```

```
##
## Call:
## glm(formula = Churn ~ tenure, family = binomial, data = train)
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.010348   0.050517   0.205    0.838
## tenure      -0.038339   0.001679 -22.837   <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: 5694.2  on 4929  degrees of freedom
## Residual deviance: 5040.7  on 4928  degrees of freedom
## AIC: 5044.7
##
## Number of Fisher Scoring iterations: 4
```

```r
mod6 <- glm(Churn ~ PhoneService, data=train, family=binomial)
summary(mod6)
```

```
##
## Call:
## glm(formula = Churn ~ PhoneService, family = binomial, data = train)
##
## Coefficients:
```

```
##                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)      -1.1415     0.1076 -10.611   <2e-16 ***
## PhoneServiceYes   0.1299     0.1128   1.151     0.25
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 5694.2  on 4929  degrees of freedom
## Residual deviance: 5692.9  on 4928  degrees of freedom
## AIC: 5696.9
##
## Number of Fisher Scoring iterations: 4
```

```r
mod7 <- glm(Churn ~ MultipleLines, data=train, family=binomial)
summary(mod7)
```

```
##
## Call:
## glm(formula = Churn ~ MultipleLines, family = binomial, data = train)
##
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)      -1.12350    0.04348 -25.841  < 2e-16 ***
## MultipleLinesYes  0.23006    0.06505   3.537 0.000405 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 5694.2  on 4929  degrees of freedom
## Residual deviance: 5681.7  on 4928  degrees of freedom
## AIC: 5685.7
##
## Number of Fisher Scoring iterations: 4
```

```r
mod8 <- glm(Churn ~ InternetService, data=train, family=binomial)
summary(mod8)
```

```
##
## Call:
## glm(formula = Churn ~ InternetService, family = binomial, data = train)
##
## Coefficients:
##                             Estimate Std. Error z value Pr(>|z|)
## (Intercept)                 -1.47098    0.06258 -23.506   <2e-16 ***
## InternetServiceFiber optic  1.13842    0.07611  14.957   <2e-16 ***
## InternetServiceNo           -1.11658    0.13582  -8.221   <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: 5694.2  on 4929  degrees of freedom
## Residual deviance: 5132.9  on 4927  degrees of freedom
## AIC: 5138.9
##
## Number of Fisher Scoring iterations: 5
```

```
mod9 <- glm(Churn ~ OnlineSecurity, data=train, family=binomial)
summary(mod9)
```

```
##
## Call:
## glm(formula = Churn ~ OnlineSecurity, family = binomial, data = train)
##
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)       -0.79719    0.03633  -21.94   <2e-16 ***
## OnlineSecurityYes -0.96472    0.08405  -11.48   <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: 5694.2  on 4929  degrees of freedom
## Residual deviance: 5544.3  on 4928  degrees of freedom
## AIC: 5548.3
##
## Number of Fisher Scoring iterations: 4
```

```
mod10 <- glm(Churn ~ OnlineBackup, data=train, family=binomial)
summary(mod10)
```

```
##
## Call:
## glm(formula = Churn ~ OnlineBackup, family = binomial, data = train)
##
## Coefficients:
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)     -0.91109    0.03891 -23.414  < 2e-16 ***
## OnlineBackupYes -0.34507    0.07016  -4.919 8.72e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 5694.2  on 4929  degrees of freedom
## Residual deviance: 5669.4  on 4928  degrees of freedom
## AIC: 5673.4
##
## Number of Fisher Scoring iterations: 4
```

```
mod11 <- glm(Churn ~ DeviceProtection, data=train, family=binomial)
summary(mod11)
```

```
##
## Call:
## glm(formula = Churn ~ DeviceProtection, family = binomial, data = train)
##
## Coefficients:
##                     Estimate Std. Error z value Pr(>|z|)
## (Intercept)         -0.93239    0.03909 -23.852  < 2e-16 ***
## DeviceProtectionYes -0.27669    0.06963  -3.973 7.09e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 5694.2  on 4929  degrees of freedom
## Residual deviance: 5678.1  on 4928  degrees of freedom
## AIC: 5682.1
##
## Number of Fisher Scoring iterations: 4
```

```r
mod12 <- glm(Churn ~ TechSupport, data=train, family=binomial)
summary(mod12)
```

```
##
## Call:
## glm(formula = Churn ~ TechSupport, family = binomial, data = train)
##
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)    -0.80594    0.03674  -21.94   <2e-16 ***
## TechSupportYes -0.86397    0.08058  -10.72   <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: 5694.2  on 4929  degrees of freedom
## Residual deviance: 5566.6  on 4928  degrees of freedom
## AIC: 5570.6
##
## Number of Fisher Scoring iterations: 4
```

```r
mod13 <- glm(Churn ~ StreamingTV, data=train, family=binomial)
summary(mod13)
```

```
##
## Call:
## glm(formula = Churn ~ StreamingTV, family = binomial, data = train)
##
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)    -1.14795    0.04263 -26.931  < 2e-16 ***
## StreamingTVYes  0.30561    0.06551   4.665 3.09e-06 ***
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 5694.2  on 4929  degrees of freedom
## Residual deviance: 5672.6  on 4928  degrees of freedom
## AIC: 5676.6
##
## Number of Fisher Scoring iterations: 4
```

```
mod14 <- glm(Churn ~ StreamingMovies, data=train, family=binomial)
summary(mod14)
```

```
##
## Call:
## glm(formula = Churn ~ StreamingMovies, family = binomial, data = train)
##
## Coefficients:
##                    Estimate Std. Error z value Pr(>|z|)
## (Intercept)        -1.12512    0.04254 -26.449  < 2e-16 ***
## StreamingMoviesYes  0.24849    0.06550   3.794 0.000148 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 5694.2  on 4929  degrees of freedom
## Residual deviance: 5679.9  on 4928  degrees of freedom
## AIC: 5683.9
##
## Number of Fisher Scoring iterations: 4
```

```
mod15 <- glm(Churn ~ Contract, data=train, family=binomial)
summary(mod15)
```

```
##
## Call:
## glm(formula = Churn ~ Contract, family = binomial, data = train)
##
## Coefficients:
##                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)      -0.30975    0.03876  -7.992 1.33e-15 ***
## ContractOne year -1.73958    0.10521 -16.535  < 2e-16 ***
## ContractTwo year -3.29329    0.18611 -17.695  < 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: 5694.2  on 4929  degrees of freedom
## Residual deviance: 4736.2  on 4927  degrees of freedom
## AIC: 4742.2
##
## Number of Fisher Scoring iterations: 6
```

```r
mod16 <- glm(Churn ~ PaperlessBilling, data=train, family=binomial)
summary(mod16)
```

```
##
## Call:
## glm(formula = Churn ~ PaperlessBilling, family = binomial, data = train)
##
## Coefficients:
##                     Estimate Std. Error z value Pr(>|z|)
## (Intercept)         -1.62562    0.06013  -27.04   <2e-16 ***
## PaperlessBillingYes  0.93196    0.07182   12.98   <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: 5694.2  on 4929  degrees of freedom
## Residual deviance: 5512.4  on 4928  degrees of freedom
## AIC: 5516.4
##
## Number of Fisher Scoring iterations: 4
```

```r
mod17 <- glm(Churn ~ PaymentMethod, data=train, family=binomial)
summary(mod17)
```

```
##
## Call:
## glm(formula = Churn ~ PaymentMethod, family = binomial, data = train)
##
## Coefficients:
##                                      Estimate Std. Error z value Pr(>|z|)
## (Intercept)                          -1.59686    0.08266 -19.319   <2e-16 ***
## PaymentMethodCredit card (automatic) -0.15101    0.11847  -1.275    0.202
## PaymentMethodElectronic check         1.40923    0.09627  14.638   <2e-16 ***
## PaymentMethodMailed check             0.13813    0.11233   1.230    0.219
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 5694.2  on 4929  degrees of freedom
## Residual deviance: 5246.3  on 4926  degrees of freedom
## AIC: 5254.3
##
## Number of Fisher Scoring iterations: 4
```

```r
mod18 <- glm(Churn ~ MonthlyCharges, data=train, family=binomial)
summary(mod18)
```

```
##
## Call:
## glm(formula = Churn ~ MonthlyCharges, family = binomial, data = train)
```

```
##
## Coefficients:
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)    -2.120267   0.090047  -23.55   <2e-16 ***
## MonthlyCharges  0.016008   0.001166   13.73   <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: 5694.2  on 4929  degrees of freedom
## Residual deviance: 5491.4  on 4928  degrees of freedom
## AIC: 5495.4
##
## Number of Fisher Scoring iterations: 4
```

```
mod19 <- glm(Churn ~ TotalCharges, data=train, family=binomial)
summary(mod19)
```

```
##
## Call:
## glm(formula = Churn ~ TotalCharges, family = binomial, data = train)
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -5.713e-01  4.451e-02  -12.84   <2e-16 ***
## TotalCharges -2.257e-04  1.726e-05  -13.07   <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: 5694.2  on 4929  degrees of freedom
## Residual deviance: 5494.9  on 4928  degrees of freedom
## AIC: 5498.9
##
## Number of Fisher Scoring iterations: 4
```

```
AIC(mod)
```

```
## [1] 5697.925
```