Statistical inference and Modeling

SIM Project 2

Inigo Pikabea and Max Tico

Master in Data Science

2024-01-03

Table of Content


```
df <- read.csv("Data/WA_Fn-UseC_-Telco-Customer-Churn.xls")
```

GitHub was used as Version Control System for this project.

The contribution of each member is visible through the following repository: https://github.com/inigopm/SIM-Project-2.git

And the task distribution: https://github.com/inigopm/Projects

Data preparation

As a first step, we imported the training data through the 'read_csv' function.

Then we performed data preparation over the data. It consisted of 4 different steps: Univariate Descriptive Analysis, Data Quality report, Imputation and Profiling.

Univariate Descriptive Analysis

Before performing the descriptive analysis, some variables had to be changed in order to better understand them and also to follow the same characteristics

```
str(df)
```

```
## 'data.frame':    7043 obs. of  21 variables:
##  $ customerID      : chr  "7590-VHVEG" "5575-GNVDE" "3668-QPYBK" "7795
-CFOCW" ...
##  $ gender          : chr  "Female" "Male" "Male" "Male" ...
##  $ SeniorCitizen   : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ Partner         : chr  "Yes" "No" "No" "No" ...
##  $ Dependents      : chr  "No" "No" "No" "No" ...
##  $ tenure          : int  1 34 2 45 2 8 22 10 28 62 ...
##  $ PhoneService    : chr  "No" "Yes" "Yes" "No" ...
##  $ MultipleLines   : chr  "No phone service" "No" "No" "No phone servi
ce" ...
##  $ InternetService : chr  "DSL" "DSL" "DSL" "DSL" ...
##  $ OnlineSecurity  : chr  "No" "Yes" "Yes" "Yes" ...
##  $ OnlineBackup    : chr  "Yes" "No" "Yes" "No" ...
##  $ DeviceProtection: chr  "No" "Yes" "No" "Yes" ...
##  $ TechSupport     : chr  "No" "No" "No" "Yes" ...
##  $ StreamingTV     : chr  "No" "No" "No" "No" ...
##  $ StreamingMovies : chr  "No" "No" "No" "No" ...
##  $ Contract        : chr  "Month-to-month" "One year" "Month-to-month"
"One year" ...
##  $ PaperlessBilling: chr  "Yes" "No" "Yes" "No" ...
##  $ PaymentMethod   : chr  "Electronic check" "Mailed check" "Mailed ch
eck" "Bank transfer (automatic)" ...
##  $ MonthlyCharges  : num  29.9 57 53.9 42.3 70.7 ...
##  $ TotalCharges    : num  29.9 1889.5 108.2 1840.8 151.7 ...
##  $ Churn           : chr  "No" "No" "Yes" "No" ...
```

```
summary(df)
```

```
##   customerID           gender          SeniorCitizen      Partner
##  Length:7043        Length:7043        Min.   :0.0000   Length:7043
##  Class :character   Class :character   1st Qu.:0.0000   Class :charact
er
##  Mode  :character   Mode  :character   Median :0.0000   Mode  :charact
er
##                                        Mean   :0.1621
```

```
##                                          3rd Qu.:0.0000
##                                          Max.   :1.0000
##
##    Dependents            tenure      PhoneService       MultipleLines
##   Length:7043        Min.   : 0.00   Length:7043        Length:7043
##   Class :character   1st Qu.: 9.00   Class :character   Class :characte
r
##   Mode  :character   Median :29.00   Mode  :character   Mode  :characte
r
##                      Mean   :32.37
##                      3rd Qu.:55.00
##                      Max.   :72.00
##
##   InternetService    OnlineSecurity     OnlineBackup       DeviceProtec
tion
##   Length:7043        Length:7043        Length:7043        Length:7043
##   Class :character   Class :character   Class :character   Class :chara
cter
##   Mode  :character   Mode  :character   Mode  :character   Mode  :chara
cter
##
##
##
##
##   TechSupport        StreamingTV        StreamingMovies      Contract
##   Length:7043        Length:7043        Length:7043        Length:7043
##   Class :character   Class :character   Class :character   Class :chara
cter
##   Mode  :character   Mode  :character   Mode  :character   Mode  :chara
cter
##
##
##
##
##   PaperlessBilling   PaymentMethod      MonthlyCharges    TotalCharges
##   Length:7043        Length:7043        Min.   : 18.25   Min.   :  18.8
##   Class :character   Class :character   1st Qu.: 35.50   1st Qu.: 401.4
##   Mode  :character   Mode  :character   Median : 70.35   Median :1397.5
##                                         Mean   : 64.76   Mean   :2283.3
##                                         3rd Qu.: 89.85   3rd Qu.:3794.7
##                                         Max.   :118.75   Max.   :8684.8
##                                                           NA's   :11
##      Churn
##   Length:7043
##   Class :character
##   Mode  :character
##
##
```

```
##
##
```

Some numeric variables variables, corresponding to qualitative concepts, need to be converted to factors:

```
columns_to_factor <- c("gender", "SeniorCitizen", "Partner", "Dependents"
,
                        "PhoneService", "MultipleLines", "InternetService
",
                        "OnlineSecurity", "OnlineBackup", "DeviceProtecti
on",
                        "TechSupport", "StreamingTV", "StreamingMovies",
"Contract",
                        "PaperlessBilling", "PaymentMethod","Churn")

# Loop through each column and convert to factor
for (col in columns_to_factor) {
  df[[col]] <- factor(df[[col]])
}
```

For numeric variables corresponding to real quantitative concepts, we will keep them as numeric but we will create additional factors as a discretization of each one. For this purpose, we created a factor for each numeric variable, consisting of 4 different bins (values).

```
# Monthly charges
# Define the breaks for discretization (bins)
breaks <- seq(0, 120, by = 30)

# Create factor variable using cut()
df$factor_monthlycharges <- cut(df$MonthlyCharges, breaks = breaks, label
s = c("Very_low", "Low", "Medium", "High"), include.lowest = TRUE)

# Total charges
# Define the breaks for discretization (bins)
breaks <- seq(18.8, 8684.8, by = 2000)

# Create factor variable using cut()
df$factor_totalcharges <- cut(df$TotalCharges, breaks = breaks, labels =
c("Very_low", "Low", "Medium", "High"), include.lowest = TRUE, na.rm = TR
UE)
```

After this, we can now perform an Exploratory Data Analysis for each variable. The description and plotting in reference to each variable is shown in the appendix.

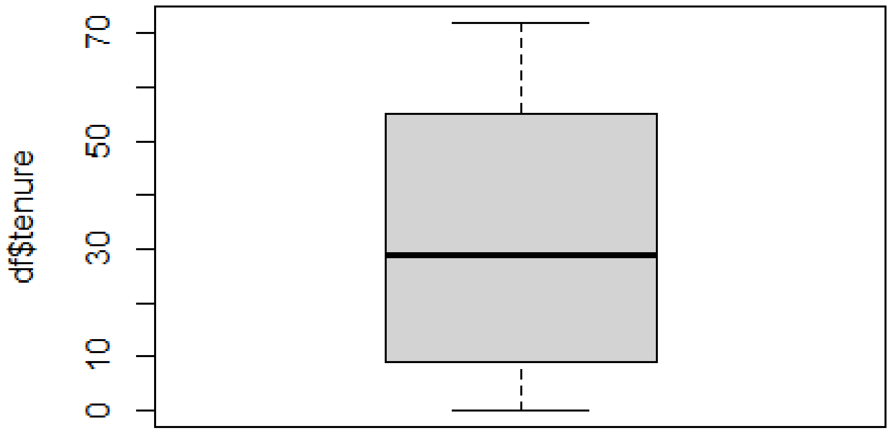
Data Quality Report

Afterwards, for each variable we counted the number of missing values and ranked them according to the sum of missing values. In total, we found 95 missing values, corresponding to variables TotalCharges and factor_totalcharges. We performed a boxplot for every numerical feature in order to identify outliers but no outliers were identified.

After observing the correlations between numerical data, we see that TotalCharges shows a high correlation with both MonthlyCharges (0.6515) and tenure (0.8263). This strong correlation suggests that, in the model creation, TotalCharges might not add independent value in a predictive model if MonthlyCharges and tenure are already included. In building our final model, we'll consider excluding TotalCharges or carefully examine its impact, particularly on metrics like the Akaike Information Criterion (AIC), to avoid potential multicollinearity and ensure model efficiency.

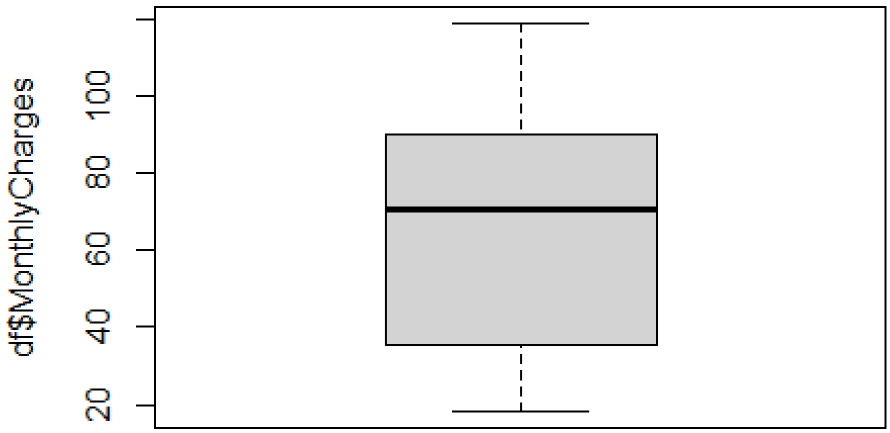
```
missing_counts <- colSums(is.na(df));missing_counts

##            customerID                gender         SeniorCitizen
##                     0                     0                     0
##               Partner            Dependents                tenure
##                     0                     0                     0
##          PhoneService         MultipleLines       InternetService
##                     0                     0                     0
##        OnlineSecurity          OnlineBackup      DeviceProtection
##                     0                     0                     0
##           TechSupport           StreamingTV       StreamingMovies
##                     0                     0                     0
##              Contract      PaperlessBilling         PaymentMethod
##                     0                     0                     0
##        MonthlyCharges          TotalCharges                 Churn
##                     0                    11                     0
## factor_monthlycharges   factor_totalcharges
##                     0                    85

total_missings <- sum(is.na(df));total_missings

## [1] 96

# Correlations

num_corr <- c("tenure", "MonthlyCharges", "TotalCharges")


correlations <- cor(df[, num_corr], use = "complete.obs")
correlations

##                   tenure MonthlyCharges TotalCharges
## tenure         1.0000000      0.2468618    0.8258805
## MonthlyCharges 0.2468618      1.0000000    0.6510648
## TotalCharges   0.8258805      0.6510648    1.0000000
```

```r
# Outlier detection
length(Boxplot(df$tenure, id = list(n = Inf)))
```

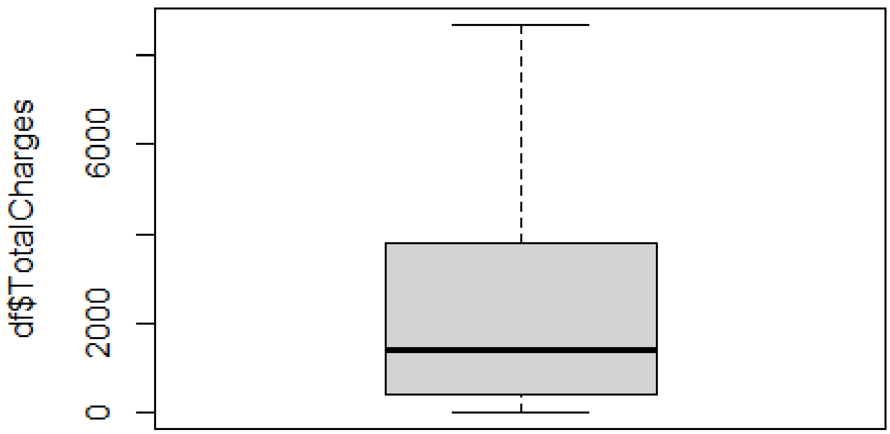



```
## [1] 0
```

```r
length(Boxplot(df$MonthlyCharges, id = list(n = Inf)))
```




```
## [1] 0
length(Boxplot(df$TotalCharges, id = list(n = Inf)))
```



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

We also counted the number of missings per individuals and number of outliers (including multivariant outliers). There are some individuals which have up to two missings.

```
missing_per_individual <- rowSums(is.na(df));missing_per_individual

##    [1] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0
##   [38] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0
##   [75] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0
##  [112] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0
##  [149] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0
##  [186] 0 0 0 0 0 0 0 0 1 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0
…
```

Imputation

As we observed before, there are some variables of our dataset with missing values (TotalCharges & factor_totalcharges). In order to solve this, we imputed those values. As one of this variables is numeric and the other one is categorical, we had to follow different approaches. For the numeric variable, we used imputePCA() function from missMDA library. For the categoric variable we used imputeMCA() function from missMDA library too, which is helpful for imputation of missing values in categorical features.

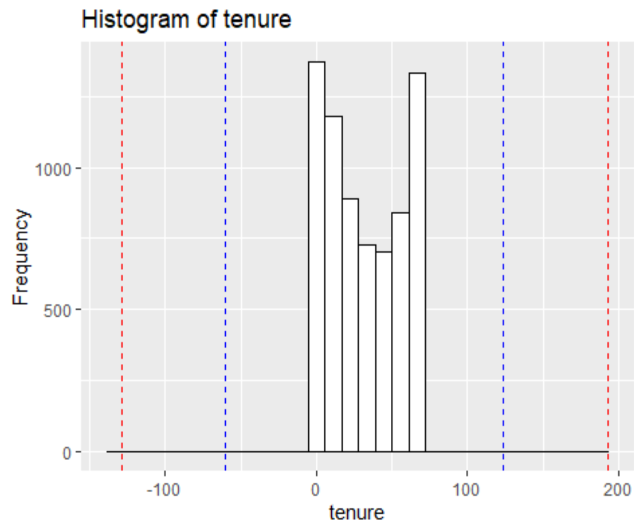
```
# Imputation of numeric variable TotalCharges
res.pca <- imputePCA(df[, c(6,19:20)])
df$TotalCharges <- res.pca$completeObs[, 3]

# Imputation of categorical variable factor_totalcharges
res.mca <- imputeMCA(df[, c(2:5,7:18,21:23)])
df$factor_totalcharges <- res.mca$completeObs[, 19]
```

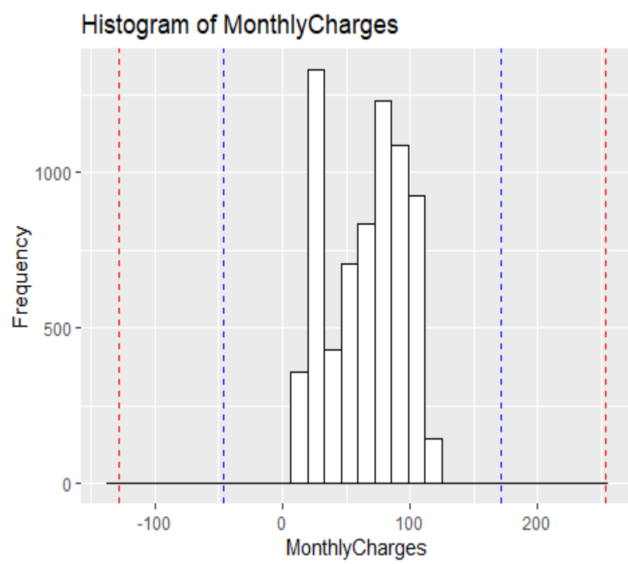
Univariate and Multivariate analysis.

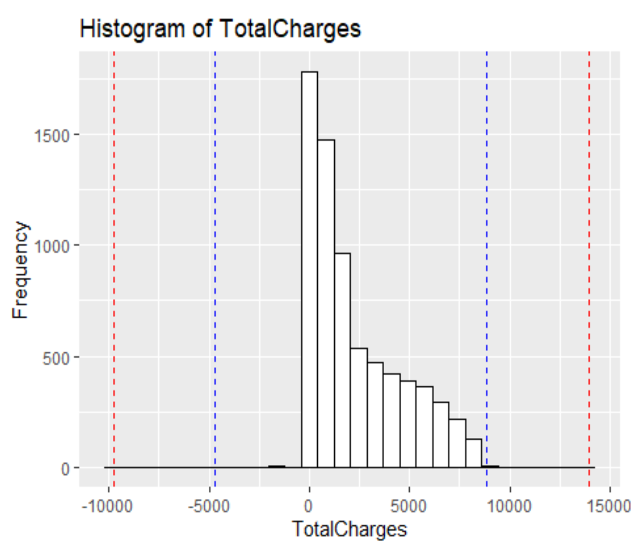
After taking the only 3 numeric values, the analysis showed that there are no univariate outliers. The lack of univariate outliers in the data suggests that the values within each of these variables fall within a reasonable range, without extreme values that could skew the analysis.

We wanted to assess also if multivariate outliers were seen or not. We used the robust Mahalanobis distance and we did not observe any multivariate outlier.

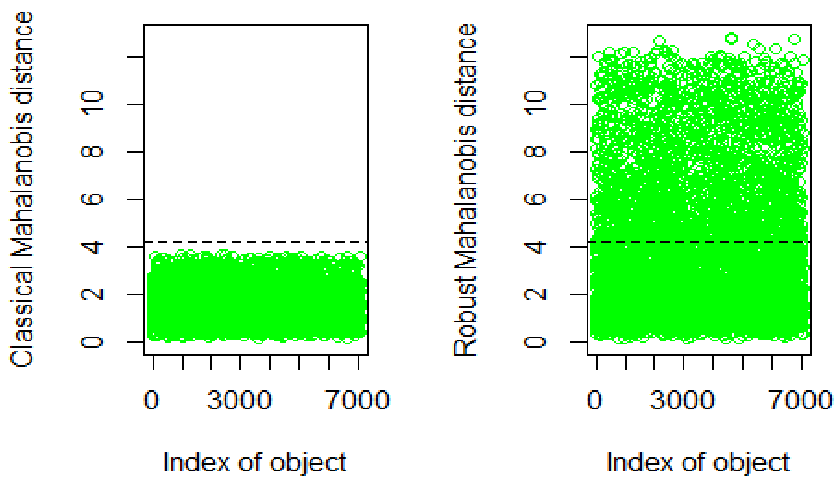
```
num_outliers <- c("tenure", "MonthlyCharges", "TotalCharges")

for(i in 1:length(num_outliers)) {
  columna <- num_outliers[i]

  # Calculate the thresholds
  q1 <- quantile(df[columna],0.25, na.rm = TRUE)
  q3 <- quantile(df[columna],0.75, na.rm = TRUE)
  iqr <- q3 - q1
  mild_l <- q1 - iqr*1.5
  mild_h <- q3 + iqr*1.5
  high_l <- q1 - iqr*3
  high_h <- q3 + iqr*3

  # Create the plot
  p <- ggplot(df, aes(x=!!sym(columna))) +
    geom_histogram(color="black", fill="white", bins=30) +
    geom_vline(aes(xintercept=mild_l), color="blue", linetype="dashed") +
    geom_vline(aes(xintercept=mild_h), color="blue", linetype="dashed") +
    geom_vline(aes(xintercept=high_l), color="red", linetype="dashed") +
    geom_vline(aes(xintercept=high_h), color="red", linetype="dashed") +
    labs(x = columna, y="Frequency", title = paste("Histogram of", column
a))
  # Add the plot to the list
  print(p)
}
```


Histogram of tenure
Frequency
1000
500
0
-100
0
100
200
tenure

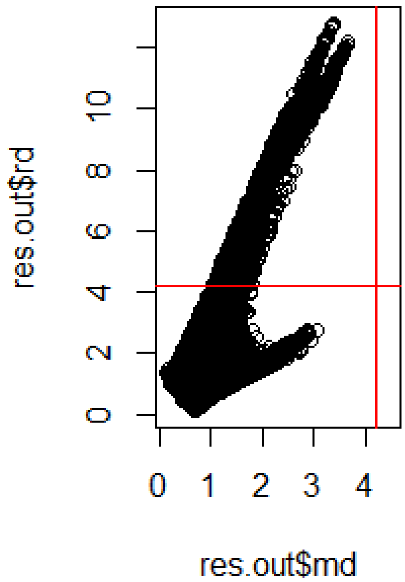

Histogram of MonthlyCharges
Frequency
1000
500
0
-100
0
100
200
MonthlyCharges


Histogram of TotalCharges
Frequency
1500
1000
500
0
-10000
-5000
0
5000
10000
15000
TotalCharges

```
res.out = Moutlier(na.omit(df[, num_outliers]), quantile = 0.9995, col="g
reen")
```




```
outlier_index <- which((res.out$md > res.out$cutoff)&(res.out$rd > res.ou
t$cutoff))
length(outlier_index)

## [1] 0

# par(mfrow=c(1,1))
xlim_range = c(min(res.out$md, na.rm = TRUE), 4.5)
plot( res.out$md, res.out$rd, xlim=xlim_range)
abline(h=res.out$cutoff, col="red")
abline(v=res.out$cutoff, col="red")
```



Profiling

Catdes() is an R function from FactoMineR which is used to describe one factor by categorical variables and/or by qualitative variables. First we analyzed the categorical variables which characterized our binary target variable. In the test.chi2 we can observe that all variables are significant. From those, the ones that exhibited the lowest p.value were Contract, OnlineSecurity, TechSupport, InternetService, PaymentMethod, OnlineBackup & DeviceProtection. We lately focused on the description of each category of our binary variable by the categories of all categorical variables in our dataset. We observed that Two-year contract factor had the lowest p.value among all factors in No category (Churn), meaning that it is a very important factor to describe the No category. The following categories, where corresponding to no Internet Service in different variables. This can be explained as it is the same value repeated in many columns. For the Yes category, we found that Month-to-month factor from Contract variable had lowest p.value, being the category describing 'Yes' the most. For the following categories we observed the same pattern as in 'No' category.

```
catdes.res <- catdes(df,21)
# Assessing which categorical variables characterized the most our target
catdes.res$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
## factor_monthlycharges  3.672933e-73  3
## PaperlessBilling      2.614597e-58  1
## Dependents            3.276083e-43  1
## factor_totalcharges   7.200031e-40  3
## SeniorCitizen         9.477904e-37  1
## Partner               1.519037e-36  1
## MultipleLines         3.464383e-03  2
```

```
# Description of each category by all categorical variables
catdes.res$category
```

```
## $No
##                                         Cla/Mod  Mod/Cla    Global
## Contract=Two year                      97.16814 31.83224 24.066449
## StreamingMovies=No internet service     92.59502 27.30963 21.666903
## StreamingTV=No internet service         92.59502 27.30963 21.666903
## TechSupport=No internet service         92.59502 27.30963 21.666903
## DeviceProtection=No internet service    92.59502 27.30963 21.666903
## OnlineBackup=No internet service        92.59502 27.30963 21.666903
## OnlineSecurity=No internet service      92.59502 27.30963 21.666903
## InternetService=No                      92.59502 27.30963 21.666903
## factor_monthlycharges=Very_low          90.19964 28.81716 23.470112
## PaperlessBilling=No                     83.66992 46.44376 40.778078
## Contract=One year                       88.73048 25.26092 20.914383
## OnlineSecurity=Yes                      85.38881 33.32045 28.666761
## TechSupport=Yes                         84.83366 33.51372 29.021724
## Dependents=Yes                          84.54976 34.48009 29.958824
## Partner=Yes                             80.33510 52.82180 48.303280
## SeniorCitizen=0                         76.39383 87.12795 83.785319
## PaymentMethod=Credit card (automatic)   84.75690 24.93235 21.610109
## InternetService=DSL                     81.04089 37.92037 34.374556
## PaymentMethod=Bank transfer (automatic) 83.29016 24.85504 21.922476
## factor_totalcharges=Medium              84.34442 16.66022 14.510862
## factor_totalcharges=High                86.55738 10.20487  8.661082
## PaymentMethod=Mailed check              80.89330 25.20294 22.887974
## OnlineBackup=Yes                        78.46851 36.83804 34.488144
## DeviceProtection=Yes                    77.49794 36.27754 34.388755
## MultipleLines=No                        74.95575 49.11094 48.132898
## factor_totalcharges=Low                 76.53400 17.83920 17.123385
## MultipleLines=Yes                       71.39010 40.99343 42.183729
## StreamingMovies=Yes                     70.05857 36.99266 38.790288
…
```

Continuation in appendix.

Separation between Train and Test datasets

We will perform separation of the data into train and test.

```
set.seed(123)
# Create an index to randomly split the data
index <- sample(1:nrow(df), nrow(df)*0.8)  # 80% for training, 20% for te
sting
# Create the training set
train_data <- df[index, ]
# Create the testing set
test_data <- df[-index, ]
```

Modeling using numeric variables.

Five logistic regression models were developed to predict customer churn using different combinations and transformations of the numeric variables tenure, MonthlyCharges, and TotalCharges.

Model 1 incorporated all three variables (tenure, MonthlyCharges, TotalCharges). However, high multicollinearity was detected between tenure and TotalCharges, leading to their exclusion due to redundancy. This model yielded an AIC of 5205.915.

Model 2 simplified the approach by using only tenure and MonthlyCharges, resulting in significantly reduced multicollinearity (VIF ~ 1.29 for both variables).

Model 3 explored the effect of transforming MonthlyCharges into a logarithmic scale, hypothesizing a non-linear relationship with churn. This model, however, showed a slight increase in the AIC, suggesting it may not improve the prediction over the simpler Model 2.

Model 4 introduced an interaction term between tenure and MonthlyCharges.

Model 5 took a more complex approach, using polynomial transformations for both tenure and MonthlyCharges. This model achieved the lowest AIC, indicating a better fit. However, the complexity of this model, with higher-order polynomials, might lead to overfitting and interpretability challenges, despite its apparent predictive power.

Model 6 was a combination of model 2 and model 5, but trying to simplify to the maximum our model. We introduced no interaction between our variables, but we added a polynomial of degree two to Tenure. This model obtained an AIC higher than model 5 but lower than all the other models.

Having computed all these models, we decided to take model 6 as our final numerical model. In comparison to model 5, it has a relatively high AIC (>100 of difference), but is simpler. Model 5 consists of the two variables with very high polynomials (7 and 11).

```
mod1 <- glm(Churn ~ tenure + MonthlyCharges + TotalCharges, family="binom
ial", data=train_data)
mod1

##
## Call:  glm(formula = Churn ~ tenure + MonthlyCharges + TotalCharges,
##     family = "binomial", data = train_data)
##
## Coefficients:
##    (Intercept)          tenure  MonthlyCharges    TotalCharges
##     -1.5857884      -0.0634078       0.0297549       0.0001171
##
## Degrees of Freedom: 5633 Total (i.e. Null);  5630 Residual
## Null Deviance:       6566
## Residual Deviance: 5198  AIC: 5206

vif(mod1) # Correlacion muy alta entre Total y tenure. Quitamos Total por
ser ombinacion lineal de tenure.

##         tenure MonthlyCharges   TotalCharges
##      13.192839       2.316826      17.258259

AIC(mod1)

## [1] 5205.915

mod2 <- glm(Churn ~ tenure + MonthlyCharges, family="binomial", data=trai
n_data)
mod2

##
## Call:  glm(formula = Churn ~ tenure + MonthlyCharges, family = "binomi
al",
##     data = train_data)
##
## Coefficients:
##    (Intercept)          tenure  MonthlyCharges
##       -1.74459        -0.05371         0.03200
##
## Degrees of Freedom: 5633 Total (i.e. Null);  5631 Residual
## Null Deviance:       6566
## Residual Deviance: 5201  AIC: 5207

vif(mod2)

##         tenure MonthlyCharges
##       1.287895       1.287895

AIC(mod2)
```

```
## [1] 5207.022

# Using transformations
mod3 <- glm(Churn ~ tenure + log(MonthlyCharges), family="binomial", data
=train_data)
mod3

##
## Call:  glm(formula = Churn ~ tenure + log(MonthlyCharges), family = "b
inomial",
##     data = train_data)
##
## Coefficients:
##         (Intercept)               tenure  log(MonthlyCharges)
##             -6.2412              -0.0501               1.6063
##
## Degrees of Freedom: 5633 Total (i.e. Null);  5631 Residual
## Null Deviance:       6566
## Residual Deviance: 5223  AIC: 5229

vif(mod3)

##              tenure log(MonthlyCharges)
##             1.15893             1.15893

AIC(mod3)

## [1] 5228.922

mod4 <- glm(Churn ~ tenure * MonthlyCharges, family="binomial", data=trai
n_data)
mod4

##
## Call:  glm(formula = Churn ~ tenure * MonthlyCharges, family = "binomi
al",
##     data = train_data)
##
## Coefficients:
##           (Intercept)                 tenure         MonthlyCharges
##            -1.6008300             -0.0625364              0.0299597
## tenure:MonthlyCharges
##             0.0001068
##
## Degrees of Freedom: 5633 Total (i.e. Null);  5630 Residual
## Null Deviance:       6566
## Residual Deviance: 5198  AIC: 5206

vif(mod4)

##                tenure        MonthlyCharges tenure:MonthlyCharges
##             13.305263              2.327414             17.415387
```

```r
AIC(mod4)

## [1] 5206.484

mod5 <- glm(Churn ~ poly(tenure, 7) + poly(MonthlyCharges, 11), family="b
inomial", data=train_data)
mod5

##
## Call:  glm(formula = Churn ~ poly(tenure, 7) + poly(MonthlyCharges,
##     11), family = "binomial", data = train_data)
##
## Coefficients:
##                (Intercept)            poly(tenure, 7)1
##                    -1.4506                   -100.0712
##           poly(tenure, 7)2            poly(tenure, 7)3
##                     3.5032                    -21.9172
##           poly(tenure, 7)4            poly(tenure, 7)5
##                     1.2114                    -10.8841
##           poly(tenure, 7)6            poly(tenure, 7)7
##                    -3.4632                    -11.6118
##  poly(MonthlyCharges, 11)1   poly(MonthlyCharges, 11)2
##                    76.9597                     -3.8481
##  poly(MonthlyCharges, 11)3   poly(MonthlyCharges, 11)4
##                     1.7761                    -17.0677
##  poly(MonthlyCharges, 11)5   poly(MonthlyCharges, 11)6
##                     8.4347                      0.7307
##  poly(MonthlyCharges, 11)7   poly(MonthlyCharges, 11)8
##                    -6.4191                      6.6340
##  poly(MonthlyCharges, 11)9  poly(MonthlyCharges, 11)10
##                     7.3741                      5.1446
## poly(MonthlyCharges, 11)11
##                    -3.1525
##
## Degrees of Freedom: 5633 Total (i.e. Null);  5615 Residual
## Null Deviance:       6566
## Residual Deviance: 5029  AIC: 5067

vif(mod5)

##                              GVIF Df GVIF^(1/(2*Df))
## poly(tenure, 7)          1.687576  7        1.038085
## poly(MonthlyCharges, 11) 1.687576 11        1.024071

AIC(mod5)

## [1] 5067.19

mod6 <- glm(Churn ~ poly(tenure, 2) + MonthlyCharges, family="binomial",
data=train_data)
mod6
```

```
##
## Call:  glm(formula = Churn ~ poly(tenure, 2) + MonthlyCharges, family
= "binomial",
##     data = train_data)
##
## Coefficients:
##     (Intercept)  poly(tenure, 2)1  poly(tenure, 2)2    MonthlyCharges
##        -3.49037         -95.86741          10.09457           0.03246
##
## Degrees of Freedom: 5633 Total (i.e. Null);  5630 Residual
## Null Deviance:       6566
## Residual Deviance: 5188  AIC: 5196
```

```
vif(mod6)
```

```
##                    GVIF Df GVIF^(1/(2*Df))
## poly(tenure, 2) 1.33563  2        1.075032
## MonthlyCharges  1.33563  1        1.155695
```

```
AIC(mod6)
```

```
## [1] 5196.092
```

Residual Analysis

Once our numerical model was chosen, we wanted to validate it. To validate the quality of our model we search for the no-linearity of the variance between the errors of the predictor variables and the class.

Residual plots show pearson residuals in relation with predictor variables and the linear predictor. Within these plots we want to search the absence of clear patterns, which could indicate no-linearities or heteroscedasticity (inconstant variability of the residuals).

In the first two plots (Linear part of poly(tenure, 2) , MonthlyCharges) we do not observe clear/systematic patterns between variables and residuals, suggesting a constant variability of the errors. We do not observe clear signals of no-linearity or heteroscedasticity.

The Linear Predictor plot shows the residuals vs the adjusted values (linear predictor). We expected to observe distributed across the horizontal line in zero. In our plot we observe a tendency of the residuals to deviate from zero, indicating that the model is not capturing well the variability of the residuals.

The Marginal Model Plots show the relationship between each predictor and the response, where the red dotted line represents the predictions of the model and the blue line represents the real data.

Although in poly(tenure,2) the model deviates a little bit from the real data, in general terms we can observe in both plots that the model captures correctly the tendency of the data.

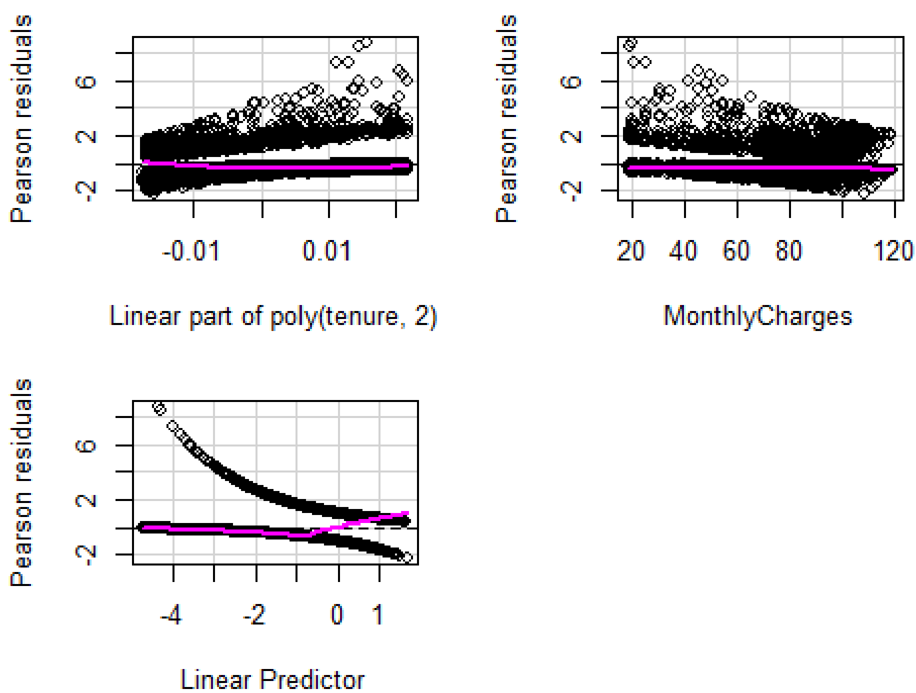
The Effect plots show the relationship between the predictors and the Churn probability. The blue shadow indicates the confidence interval.

In the Tenure effect plot we observe a non-linear relationship, decreasing the probability of churn as Tenure increases. The MonthlyCharges plot indicates that as MonthlyCharges increase, the probability of Churn also increases. This was expected as the customers tend to leave the company of their monthly charges increase / are higher.

Influence Plots visualize the influence of each observation in the model, being the size of the points proportional to Cook's distance (influence measure).

Most of the observations have low leverage values, indicating that there are not high influent data. There are 6 individuals (6119, 269, 4587, 4150, 6425 and 431) with a higher Cook's distance, being the most influent data in the model.

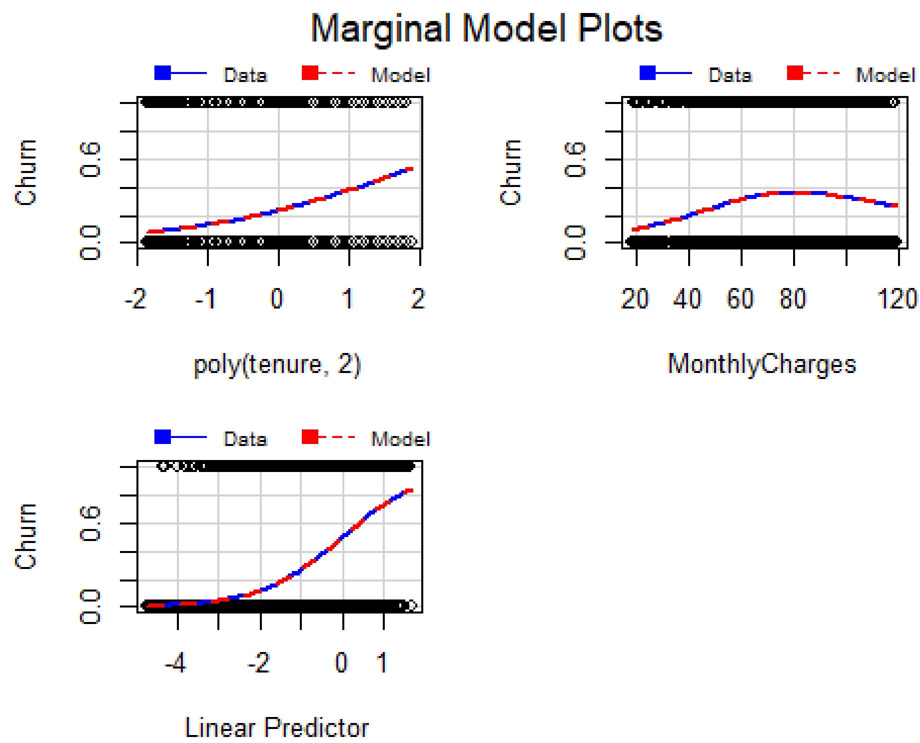
```
par("mar")

## [1] 5.1 4.1 4.1 2.1

par(mar=c(1,1,1,1))
residualPlots(mod6)
```



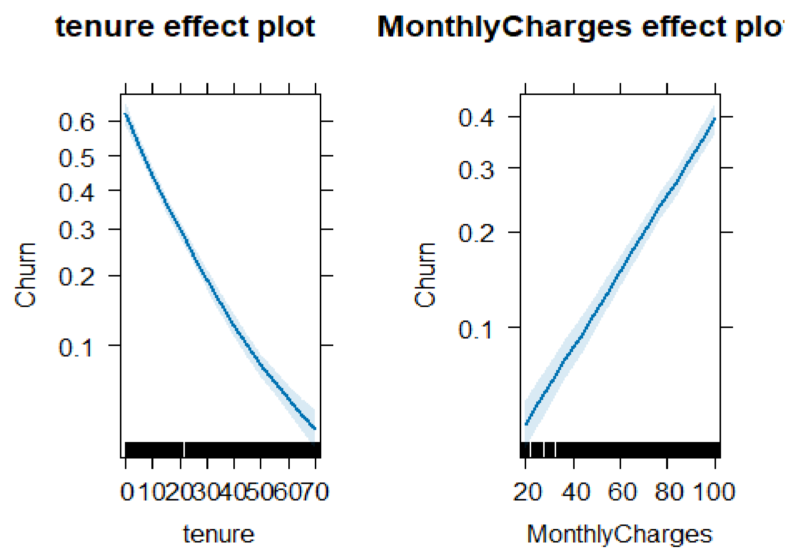
```
##                   Test stat Pr(>|Test stat|)
## poly(tenure, 2)
## MonthlyCharges      0.1865            0.6659
```

```
marginalModelPlots(mod6)
```

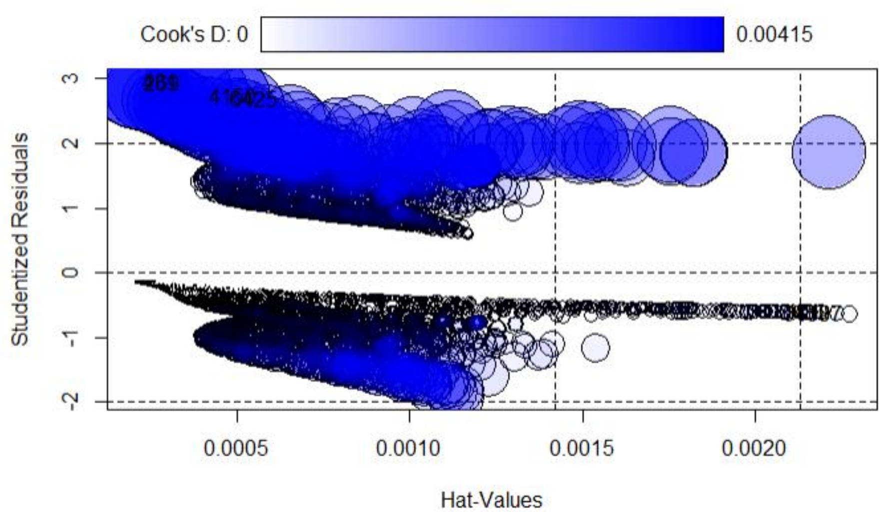
```
## Warning in mmps(...): Splines and/or polynomials replaced by a fitted
linear
## combination
```




```
plot(allEffects(mod6))
```




```
influencePlot(mod6)
```




```
##          StudRes          Hat        CookD
## 6119 -0.6333430 0.0022384799 0.0001246729
## 269   2.9303250 0.0002138174 0.0038319791
## 4587 -0.6384898 0.0022720529 0.0001288309
## 4150  2.7359941 0.0003994393 0.0040849332
## 6425  2.6904483 0.0004610606 0.0041535239
## 431   2.9549735 0.0002097746 0.0040445161
```

Adding Factors to our model

After having created and validated our model with numerical variables, we decided to incorporate all the factor variables of the dataset into our model. As it is a very big model, we wanted to select the most significant variables to Churn. We used two different approaches: Anova and Step. With the results from both approaches, we decided to create two different models (mod8 & mod9). Then we performed a comparison between both of them to see which one of the two had the best AIC. We observed that mod8, corresponding to the model with Anova results, had an AIC of 4988, even higher than the model with all factors (4753). On the other hand mod9, corresponding to step results, had a better AIC, even compared to the original model (4746 vs 4753). To gain validation to mod9, we compared mod8 and mod9 with anova (different from Anova). The anova method shows a p-value when comparing models, which tells us about the improvement of one model compared to the other. In our case, we observed a very significant p-value from mod9, meaning that mod9 significantly improves mod8, so we decided to keep it as our final factor model.

```
mod7 <- glm(Churn ~ poly(tenure, 2) + MonthlyCharges + gender + SeniorCit
izen + Partner
            + Dependents + PhoneService + MultipleLines + InternetService
+ OnlineSecurity
            + OnlineBackup + DeviceProtection + TechSupport + StreamingTV
+ StreamingMovies
            + Contract + PaperlessBilling + PaymentMethod, family="binomi
al", data=train_data)
mod7

##
## Call:  glm(formula = Churn ~ poly(tenure, 2) + MonthlyCharges + gender
+
##     SeniorCitizen + Partner + Dependents + PhoneService + MultipleLine
s +
##     InternetService + OnlineSecurity + OnlineBackup + DeviceProtection
+
##     TechSupport + StreamingTV + StreamingMovies + Contract +
##     PaperlessBilling + PaymentMethod, family = "binomial", data = trai
n_data)
##
## Coefficients:
##                           (Intercept)                          poly(tenure,
2)1
##                             -0.449420                               -48.42
5191
##                     poly(tenure, 2)2                              MonthlyCha
rges
##                             22.786289                                -0.02
6801
##                            genderMale                              SeniorCiti
zen1
##                             -0.040120                                 0.22
6988
##                            PartnerYes                               Dependent
sYes
##                             -0.044109                                -0.06
```

```
3898
##                      PhoneServiceYes          MultipleLinesNo phone ser
vice
##                             0.018492                                   
NA
##                     MultipleLinesYes             InternetServiceFiber o
ptic
##                             0.436108                               1.55
8078
##                    InternetServiceNo     OnlineSecurityNo internet ser
vice
##                            -1.409093                                   
NA
##                    OnlineSecurityYes       OnlineBackupNo internet ser
vice
##                            -0.197354                                   
NA
##                      OnlineBackupYes   DeviceProtectionNo internet ser
vice
##                            -0.006182                                   
NA
##                  DeviceProtectionYes        TechSupportNo internet ser
vice
##                             0.110874                                   
NA
##                       TechSupportYes        StreamingTVNo internet ser
vice
##                            -0.162287                                   
NA
##                       StreamingTVYes    StreamingMoviesNo internet ser
vice
##                             0.556873                                   
NA
##                   StreamingMoviesYes                         ContractOne 
year
##                             0.575622                              -0.72
1985
##                     ContractTwo year                       PaperlessBillin
gYes
##                            -1.903397                               0.33
3092
## PaymentMethodCredit card (automatic)          PaymentMethodElectronic c
heck
##                            -0.153634                               0.27
9782
##              PaymentMethodMailed check 
##                            -0.084553 
## 
## Degrees of Freedom: 5633 Total (i.e. Null);  5610 Residual
```

```
## Null Deviance:       6566
## Residual Deviance: 4706  AIC: 4754

# Removing non-significant variables
Anova(mod7, test="LR")

## Analysis of Deviance Table (Type II tests)
##
## Response: Churn
##                  LR Chisq Df Pr(>Chisq)
## poly(tenure, 2)   210.616  2  < 2.2e-16 ***
## MonthlyCharges      0.578  1    0.44729
## gender              0.308  1    0.57897
## SeniorCitizen       5.782  1    0.01619 *
## Partner             0.261  1    0.60920
## Dependents          0.417  1    0.51859
## PhoneService               0
## MultipleLines       4.902  1    0.02682 *
## InternetService     3.079  1    0.07931 .
## OnlineSecurity      0.984  1    0.32128
## OnlineBackup        0.001  1    0.97482
## DeviceProtection    0.320  1    0.57179
## TechSupport         0.648  1    0.42098
## StreamingTV         2.350  1    0.12527
## StreamingMovies     2.512  1    0.11301
## Contract          108.446  2  < 2.2e-16 ***
## PaperlessBilling   16.060  1  6.138e-05 ***
## PaymentMethod      22.637  3  4.807e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

step(mod7,trace = FALSE)

##
## Call:  glm(formula = Churn ~ poly(tenure, 2) + MonthlyCharges + Senior
Citizen +
##     MultipleLines + InternetService + OnlineSecurity + TechSupport +
##     StreamingTV + StreamingMovies + Contract + PaperlessBilling +
##     PaymentMethod, family = "binomial", data = train_data)
##
## Coefficients:
##                          (Intercept)                        poly(tenure,
2)1
##                             -0.90134                               -49.3
6274
##                    poly(tenure, 2)2                           MonthlyCha
rges
##                             22.81519                                -0.0
1776
##                       SeniorCitizen1         MultipleLinesNo phone ser
```

```
vice
##                          0.23833                                      0.1
6402
##                 MultipleLinesYes            InternetServiceFiber o
ptic
##                          0.38997                                      1.3
3624
##               InternetServiceNo    OnlineSecurityNo internet ser
vice
##                         -1.18649
NA
##               OnlineSecurityYes       TechSupportNo internet ser
vice
##                         -0.24620
NA
##                  TechSupportYes       StreamingTVNo internet ser
vice
##                         -0.20374
NA
##                  StreamingTVYes   StreamingMoviesNo internet ser
vice
##                          0.47110
NA
##              StreamingMoviesYes                        ContractOne
year
##                          0.49470                                     -0.7
2370
##                 ContractTwo year                    PaperlessBillin
gYes
##                         -1.90450                                      0.3
3361
## PaymentMethodCredit card (automatic)         PaymentMethodElectronic c
heck
##                         -0.15245                                      0.2
8077
##           PaymentMethodMailed check
##                         -0.08177
##
## Degrees of Freedom: 5633 Total (i.e. Null);  5615 Residual
## Null Deviance:       6566
## Residual Deviance: 4708  AIC: 4746
```

```r
mod8 <- glm(Churn ~ poly(tenure, 2) + SeniorCitizen + MultipleLines
            + Contract + PaperlessBilling + PaymentMethod, family="binomi
al", data=train_data)
AIC(mod8)
```

```
## [1] 4988.236
```

```r
mod9 <- glm(formula = Churn ~ poly(tenure, 2) + MonthlyCharges + SeniorCi
tizen +
    MultipleLines + InternetService + OnlineSecurity + TechSupport +
    StreamingTV + StreamingMovies + Contract + PaperlessBilling +
    PaymentMethod, family = "binomial", data = train_data)
AIC(mod9)

## [1] 4746.006

anova(mod8, mod9, test="Chisq")

## Analysis of Deviance Table
##
## Model 1: Churn ~ poly(tenure, 2) + SeniorCitizen + MultipleLines + Con
tract +
##     PaperlessBilling + PaymentMethod
## Model 2: Churn ~ poly(tenure, 2) + MonthlyCharges + SeniorCitizen + Mu
ltipleLines +
##     InternetService + OnlineSecurity + TechSupport + StreamingTV +
##     StreamingMovies + Contract + PaperlessBilling + PaymentMethod
##   Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
## 1      5622     4964.2
## 2      5615     4708.0  7   256.23 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Residual Analysis: Factors

We next validated our model with the same approach we used before. For poly(tenure, 2) & MonthlyCharges we do not observe systematic patterns neither heteroscedasticity. This indicates that the transformations and the linear relationship for these variables are correct.

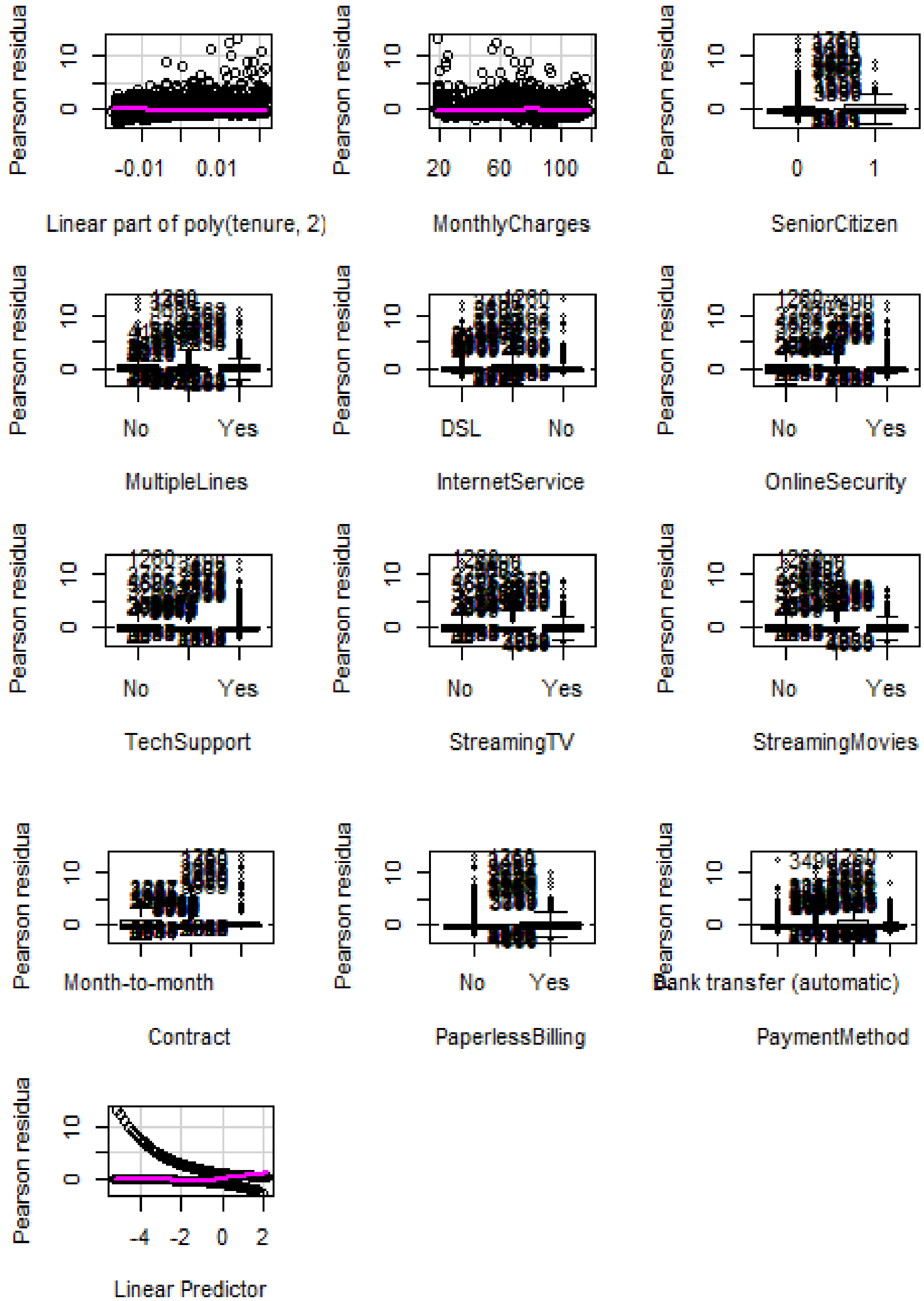
For factor variables we observe that the vast majority of observations are distributed around 0, indicating a uniform performance of the model across the different groups of these variables. Nonetheless, there are some observations which differ from 0, which could suggest they're influent data.

In the Marginal Plots we see that the model follows the tendency of the real data, suggesting that the model adjusts correctly to the variability in these predictors.

In the influence plot, the majority of observations have a low influence in the model. However, there are some individuals containing a relatively high Cook's distance, suggesting that those observations are influent data. These points need a more detailed analysis in order to determine if they represent atypical values.

In comparison with the previous residual analysis, we did observe one individual which keeps being influent in our model, which is 269.

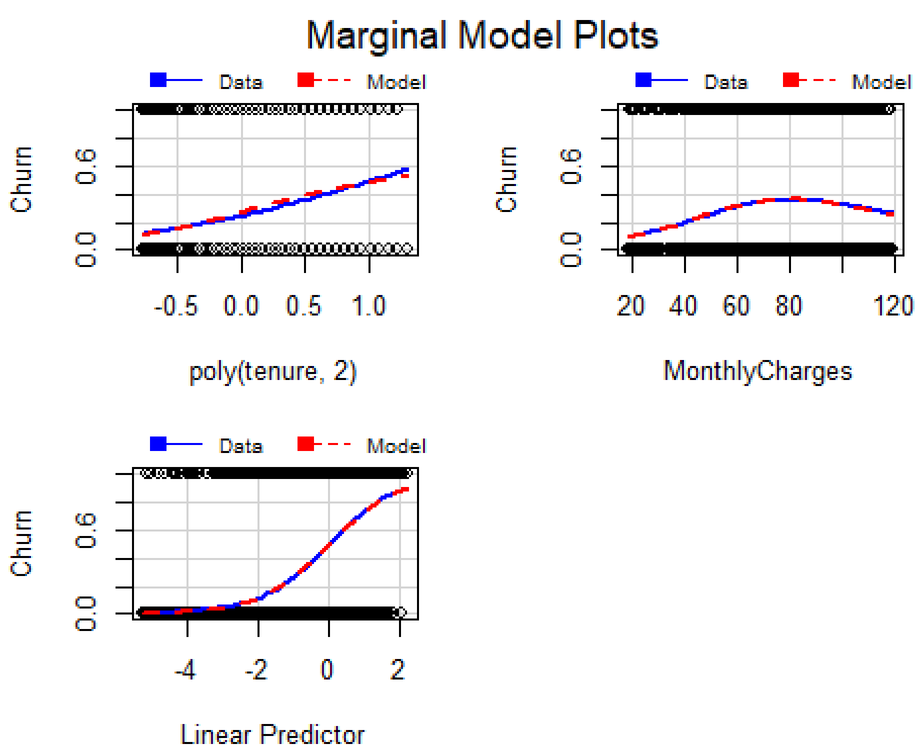
```r
par("mar")

## [1] 5.1 4.1 4.1 2.1

par(mar=c(1,1,1,1))
residualPlots(mod9)
```


Pearson residua
0
10
-0.01
0.01
Linear part of poly(tenure, 2)
20
60
100
MonthlyCharges
0
1
SeniorCitizen
No
Yes
MultipleLines
DSL
No
InternetService
No
Yes
OnlineSecurity
No
Yes
TechSupport
No
Yes
StreamingTV
No
Yes
StreamingMovies
Month-to-month
Contract
No
Yes
PaperlessBilling
Bank transfer (automatic)
PaymentMethod
-4
-2
0
2
Linear Predictor

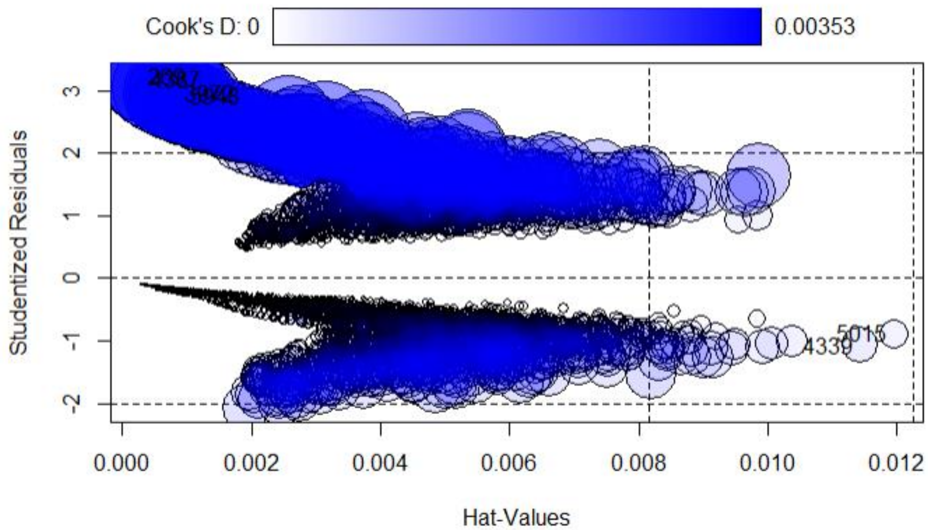
```
##                   Test stat Pr(>|Test stat|)
## poly(tenure, 2)
## MonthlyCharges       1.8327           0.1758
## SeniorCitizen
## MultipleLines
## InternetService
## OnlineSecurity
## TechSupport
## StreamingTV
## StreamingMovies
## Contract
## PaperlessBilling
## PaymentMethod
```

```
marginalModelPlots(mod9)
```

```
## Warning in mmps(...): Splines and/or polynomials replaced by a fitted
linear
## combination

## Warning in mmps(...): Interactions and/or factors skipped
```




```
# plot(allEffects(mod9)) -- da error
influencePlot(mod9)
```




```
##          StudRes          Hat        CookD
## 5015 -0.8934738 0.0119397314 0.0003129484
## 4339 -1.0755797 0.0114287086 0.0004771802
## 269   3.2254433 0.0003091408 0.0028608501
## 4387  3.1847710 0.0003380021 0.0027466249
## 3972  2.9623444 0.0008727474 0.0035341035
## 5948  2.9194994 0.0009529309 0.0034009271
```

Modeling with interactions

In order to build our model with interactions, first we decided to perform an Anova to observe which are the most significant variables in our model and transform those. Here we observed that almost all are significant, but the ones with lower p-value are tenure, InternetService, Contract, PaperlessBilling & PaymentMethod.

As Contract and Tenure were the most significant variables, we decided to create a model with interactions between them. We observed that with this model the AIC value decreased from 4746 (mod9) to 4729 (mod10). To see if mod10 was significantly different from mod9, we performed an anova. Here we obtained a p-value near 0 (6.415e-05) meaning that mod10 was significantly different from mod9.

As creating interactions between the most significant variables exhibited a decrease in the AIC value and significance, we decided to create PaperlessBilling and PaymentMethod (maintaining the interaction before). It showed an AIC value lower than mod9 (4732) but higher than mod10.

Finally we decided to create a model maintaining the first interaction with Tenure and Contract, but creating interactions between the less significant variables, as the approach we performed before didn't performed correctly. In mod12 we introduced an interaction between MonthlyCharges and SeniorCitizen and another one between

OnlineSecurity and Techsupport. We observed a decrease in the AIC value (4724) compared to mod9 and mod10. To see whether this model was significantly different from mod10 we used again anova. Here we observed a significance (p-value of 0.01), meaning that this model is different from the one created before.

```
Anova(mod9, test="LR")

## Analysis of Deviance Table (Type II tests)
##
## Response: Churn
##                    LR Chisq Df Pr(>Chisq)
## poly(tenure, 2)     226.162  2  < 2.2e-16 ***
## MonthlyCharges        2.387  1  0.1223851
## SeniorCitizen         6.611  1  0.0101335 *
## MultipleLines        17.994  2  0.0001238 ***
## InternetService      19.041  1  1.280e-05 ***
## OnlineSecurity        4.989  1  0.0255070 *
## TechSupport           3.206  1  0.0733796 .
## StreamingTV           9.680  1  0.0018624 **
## StreamingMovies      10.947  1  0.0009377 ***
## Contract            110.401  2  < 2.2e-16 ***
## PaperlessBilling     16.168  1  5.795e-05 ***
## PaymentMethod        22.610  3  4.869e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

# Interaction between tenure & Contract
mod10 <- glm(formula = Churn ~ poly(tenure, 2) * Contract + MonthlyCharge
s + SeniorCitizen +
    MultipleLines + InternetService + OnlineSecurity + TechSupport +
    StreamingTV + StreamingMovies + PaperlessBilling +
    PaymentMethod, family = "binomial", data = train_data)
AIC(mod10)

## [1] 4729.532

anova(mod9,mod10,test = "Chisq")

## Analysis of Deviance Table
##
## Model 1: Churn ~ poly(tenure, 2) + MonthlyCharges + SeniorCitizen + Mu
ltipleLines +
##     InternetService + OnlineSecurity + TechSupport + StreamingTV +
##     StreamingMovies + Contract + PaperlessBilling + PaymentMethod
## Model 2: Churn ~ poly(tenure, 2) * Contract + MonthlyCharges + SeniorC
itizen +
##     MultipleLines + InternetService + OnlineSecurity + TechSupport +
##     StreamingTV + StreamingMovies + PaperlessBilling + PaymentMethod
##   Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
## 1      5615     4708.0
```

```
## 2      5611     4683.5  4   24.474 6.415e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# Interaction with PaperlessBilling and PaymentMethod
mod11 <- glm(formula = Churn ~ poly(tenure, 2) * Contract + MonthlyCharge
s + SeniorCitizen +
    MultipleLines + InternetService + OnlineSecurity + TechSupport +
    StreamingTV + StreamingMovies + PaperlessBilling *
    PaymentMethod, family = "binomial", data = train_data)
AIC(mod11)
```

```
## [1] 4732.928
```

```
anova(mod10,mod11,test = "Chisq")
```

```
## Analysis of Deviance Table
##
## Model 1: Churn ~ poly(tenure, 2) * Contract + MonthlyCharges + SeniorC
itizen +
##     MultipleLines + InternetService + OnlineSecurity + TechSupport +
##     StreamingTV + StreamingMovies + PaperlessBilling + PaymentMethod
## Model 2: Churn ~ poly(tenure, 2) * Contract + MonthlyCharges + SeniorC
itizen +
##     MultipleLines + InternetService + OnlineSecurity + TechSupport +
##     StreamingTV + StreamingMovies + PaperlessBilling * PaymentMethod
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1      5611     4683.5
## 2      5608     4680.9  3   2.6035   0.4569
```

```
# Interaction between less significant variables
mod12 <- glm(formula = Churn ~ poly(tenure, 2) * Contract + MonthlyCharge
s * SeniorCitizen +
    MultipleLines + InternetService + OnlineSecurity * TechSupport +
    StreamingTV + StreamingMovies + PaperlessBilling +
    PaymentMethod, family = "binomial", data = train_data)
AIC(mod12)
```

```
## [1] 4724.635
```

```
anova(mod10,mod12,test = "Chisq")
```

```
## Analysis of Deviance Table
##
## Model 1: Churn ~ poly(tenure, 2) * Contract + MonthlyCharges + SeniorC
itizen +
##     MultipleLines + InternetService + OnlineSecurity + TechSupport +
##     StreamingTV + StreamingMovies + PaperlessBilling + PaymentMethod
## Model 2: Churn ~ poly(tenure, 2) * Contract + MonthlyCharges * SeniorC
itizen +
##     MultipleLines + InternetService + OnlineSecurity * TechSupport +
```

```
##     StreamingTV + StreamingMovies + PaperlessBilling + PaymentMethod
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1      5611     4683.5
## 2      5609     4674.6  2    8.897   0.0117 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
anova(mod9,mod12,test = "Chisq")
```

```
## Analysis of Deviance Table
##
## Model 1: Churn ~ poly(tenure, 2) + MonthlyCharges + SeniorCitizen + Mu
ltipleLines +
##     InternetService + OnlineSecurity + TechSupport + StreamingTV +
##     StreamingMovies + Contract + PaperlessBilling + PaymentMethod
## Model 2: Churn ~ poly(tenure, 2) * Contract + MonthlyCharges * SeniorC
itizen +
##     MultipleLines + InternetService + OnlineSecurity * TechSupport +
##     StreamingTV + StreamingMovies + PaperlessBilling + PaymentMethod
##   Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
## 1      5615     4708.0
## 2      5609     4674.6  6   33.372 8.893e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Final Residual Analysis

The residual plots indicate that there is no apparent problematic pattern in the residuals for the linear part of poly(tenure, 2), suggesting that the polynomial transformation of tenure is adequately capturing its relationship with churn. Also for MonthlyCharges, Contract and SeniorCitizen, the residuals are evenly distributed, suggesting that the model fits these variables appropriately without obvious signs of misfit.

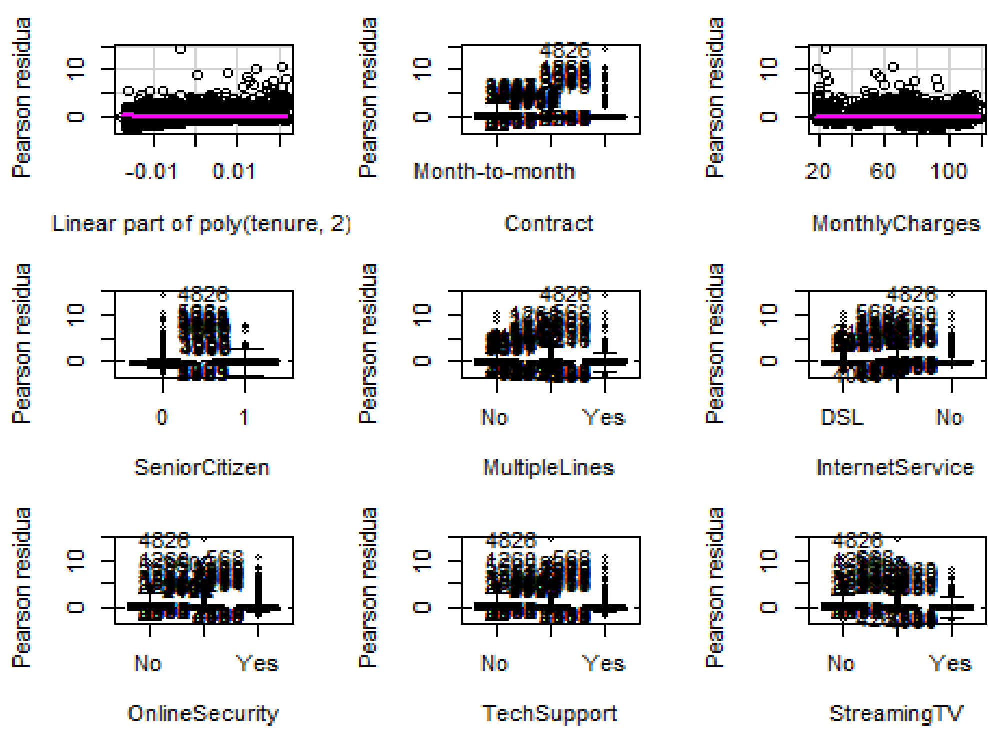
The other categorical variables, such as MultipleLines, InternetService, OnlineSecurity, TechSupport, StreamingTV, and StreamingMovies, also do not show any clear patterns in the residuals, which is generally a good sign. However, there are some outliers in each category that could be potential points of concern. These could be instances of unusual variance not accounted for by the model or could represent unique situations that are not well-represented in the data.

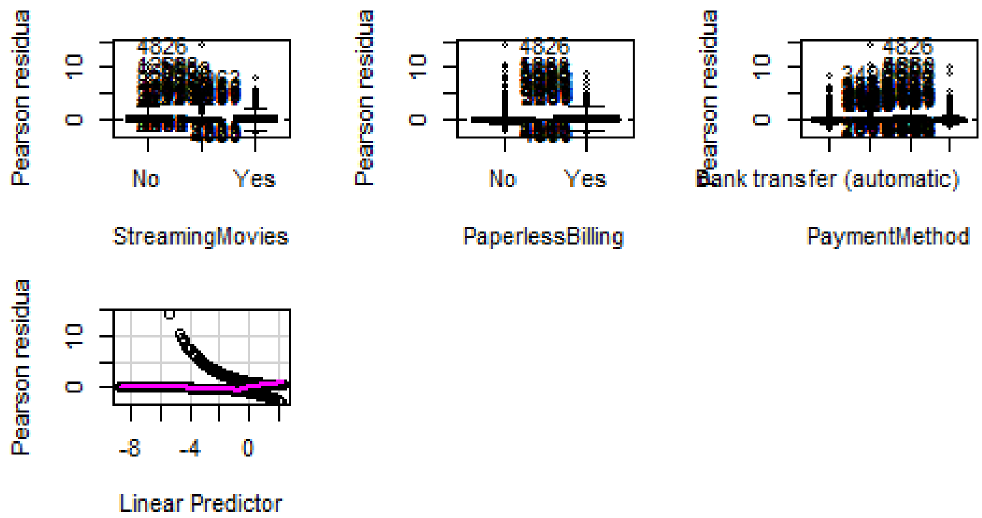
The marginal model plots, which illustrate the relationship between the predictor variables and the probability of churn, show that the model's predictions are in good agreement with the actual data. This suggests that the model is capturing the general trends in the data effectively, particularly for the tenure variable and the linear predictor.

The influence plot reveals several observations with relatively high influence on the model, as indicated by larger Cook's distance values. These points may be outliers or have high leverage and could significantly impact the model's results. Such observations might warrant closer examination and potential exclusion if they are determined to be undue influences.

In concluding the residual analysis, it was considered whether to remove the influential observations to improve model generalization. However, it was decided to retain them, acknowledging that their volume contributes significant information to the dataset. Removing these data points could result in a greater imbalance, detracting from the model's representativeness of the actual customer population. Therefore, to preserve the dataset's integrity and maintain its balance, these variables were kept despite their potential influence.

```
par("mar")

## [1] 5.1 4.1 4.1 2.1

par(mar=c(1,1,1,1))
residualPlots(mod12)
```

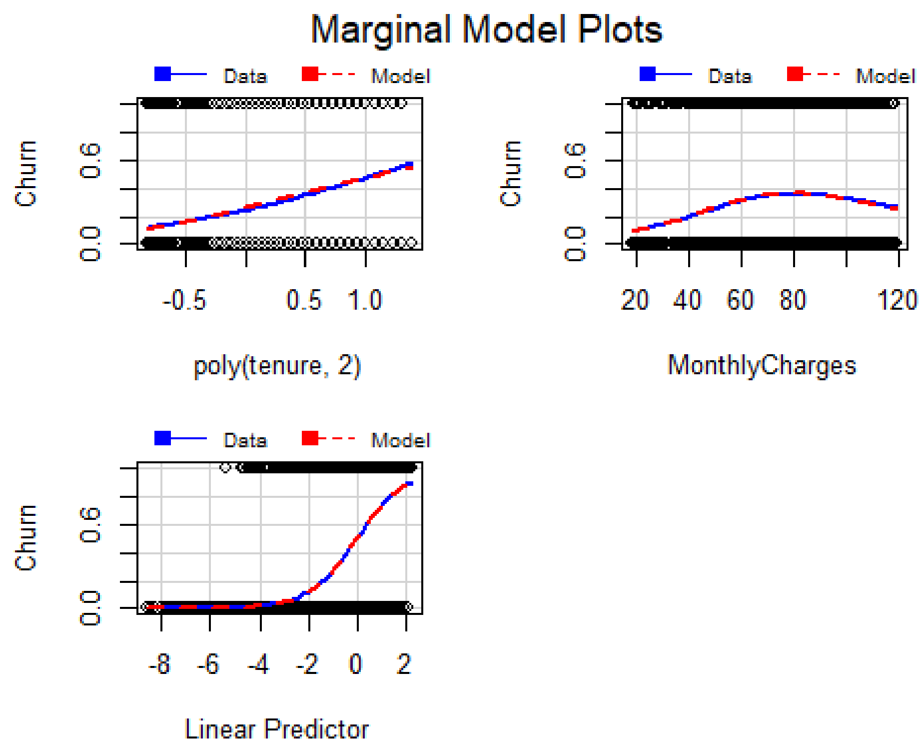




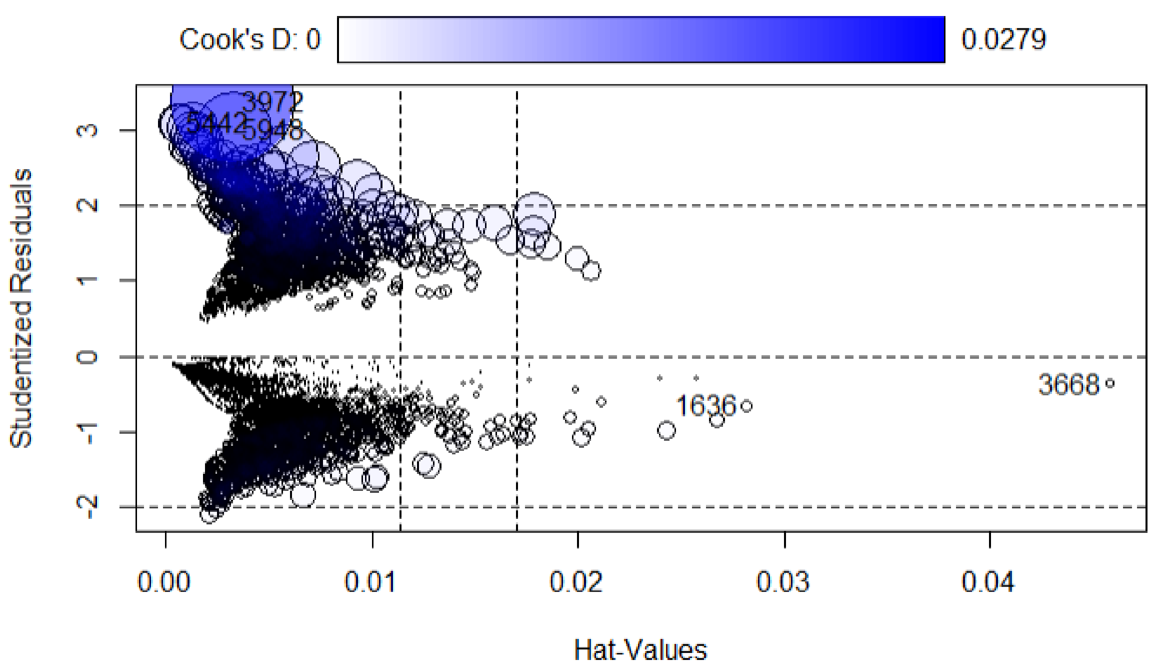

```
##                    Test stat Pr(>|Test stat|)
## poly(tenure, 2)
## Contract
## MonthlyCharges        2.0856           0.1487
## SeniorCitizen
## MultipleLines
## InternetService
## OnlineSecurity
## TechSupport
## StreamingTV
## StreamingMovies
## PaperlessBilling
## PaymentMethod
```

```
marginalModelPlots(mod12)
```

```
## Warning in mmps(...): Splines and/or polynomials replaced by a fitted linear
## combination

## Warning in mmps(...): Interactions and/or factors skipped
```




```r
# plot(allEffects(mod12)) -- da error
influencePlot(mod12)
```




```
##           StudRes          Hat        CookD
## 5442   3.0860806 0.0006317294 0.0028314080
```

```
## 3668 -0.3627866 0.0457483963 0.0001332849
## 1636 -0.6493946 0.0281287837 0.0002747218
## 3972  3.3753143 0.0033021433 0.0278523486
## 5948  3.0089599 0.0033242544 0.0106996871
```

Goodness of fit and Model Interpretation

The results from the mod12 logistic regression model evaluation reveal a strong predictive performance. A key aspect of assessing model quality is its accuracy on unseen data, and with an accuracy of approximately 83.39% on the test set, mod12 demonstrates a robust capacity to predict customer churn. The confusion matrix further underscores this performance, with a substantial number of correct predictions as compared to the incorrect ones. The sensitivity and specificity values indicate a good balance in predicting both the positive and negative classes, although there is a stronger performance in predicting the non-churners (No) over the churners (Yes).

The AUC for the test set, sitting at around 0.8635, suggests that the model has a high ability to discriminate between churners and non-churners. This is further corroborated by the ROC curve. A high AUC value is indicative of a model that provides a good separation between the two classes.

When the model's performance on the training set is compared to the test set, we see a slight dip in both accuracy and AUC. However, this difference is minimal, signaling that the model has not overfitted to the training data and is generalizing well to new, unseen data.

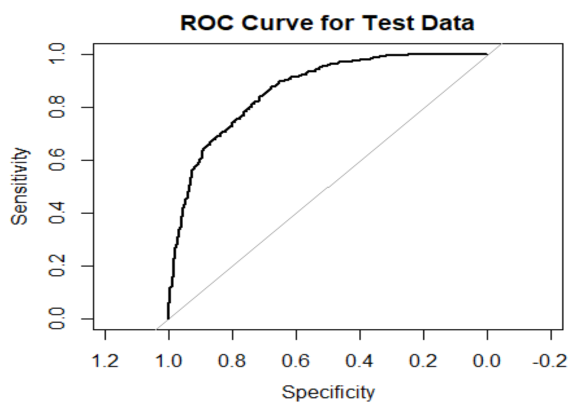
```
# Predict on the test set
test_probabilities <- predict(mod12, newdata = test_data, type = "respons
e")

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

test_predictions <- ifelse(test_probabilities > 0.5, "Yes", "No")

test_data$Churn <- factor(test_data$Churn, levels = c("No", "Yes"))
test_predictions <- factor(test_predictions, levels = c("No", "Yes"))

# Create the confusion matrix
conf_matrix <- confusionMatrix(test_predictions, test_data$Churn)
print(conf_matrix)

## Confusion Matrix and Statistics
##
##            Reference
```

```
## Prediction  No Yes
##        No  978 154
##        Yes  80 197
##
##                Accuracy : 0.8339
##                  95% CI : (0.8134, 0.853)
##     No Information Rate : 0.7509
##     P-Value [Acc > NIR] : 3.389e-14
##
##                   Kappa : 0.5224
##
##  Mcnemar's Test P-Value : 1.823e-06
##
##             Sensitivity : 0.9244
##             Specificity : 0.5613
##          Pos Pred Value : 0.8640
##          Neg Pred Value : 0.7112
##              Prevalence : 0.7509
##          Detection Rate : 0.6941
##    Detection Prevalence : 0.8034
##       Balanced Accuracy : 0.7428
##
##        'Positive' Class : No
##
```

```r
# Calculate accuracy
accuracy <- sum(diag(conf_matrix$table)) / sum(conf_matrix$table)
print(paste("Accuracy on test set:", accuracy))
```

```
## [1] "Accuracy on test set: 0.833924769339957"
```

```r
# ROC curve and AUC
roc_test <- roc(test_data$Churn, test_probabilities)
auc_test <- auc(roc_test)
plot(roc_test, main = "ROC Curve for Test Data")
```




```r
print(paste("AUC for test set:", auc_test))
```

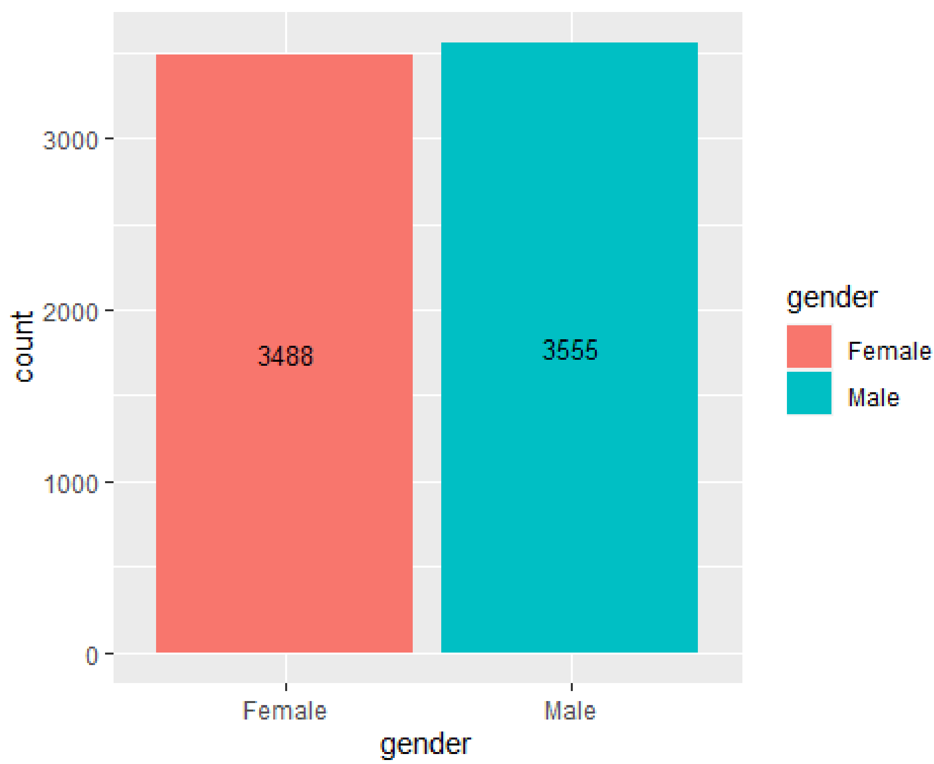
```
## [1] "AUC for test set: 0.863558883880245"

# ------- Same for train set (to see how much it overfitted)
# Calculate predictions and probabilities on the training set
train_probabilities <- predict(mod12, newdata = train_data, type = "respo
nse")

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

train_predictions <- ifelse(train_probabilities > 0.5, "Yes", "No")

# Ensure the predictions are factors with the same levels
train_data$Churn <- factor(train_data$Churn, levels = c("No", "Yes"))
train_predictions <- factor(train_predictions, levels = c("No", "Yes"))

# Create confusion matrix for the training set
train_conf_matrix <- confusionMatrix(train_predictions, train_data$Churn)

# Calculate accuracy for the training set
train_accuracy <- sum(diag(train_conf_matrix$table)) / sum(train_conf_mat
rix$table)

# Calculate ROC and AUC for the training set
train_roc <- roc(train_data$Churn, train_probabilities)
train_auc <- auc(train_roc)

# Output the train accuracy and AUC
print(paste("Accuracy on training set:", train_accuracy))

## [1] "Accuracy on training set: 0.803336883209088"

print(paste("AUC for training set:", train_auc))

## [1] "AUC for training set: 0.848485648729659"

# Check for overfitting
print(paste("Overfitting check - Difference in AUC:", train_auc - auc_tes
t))

## [1] "Overfitting check - Difference in AUC: -0.0150732351505856"
```

Appendix A

variable 1: Gender

Gender is a binary variable composed of two values: Female & Male. We used a barplot to see the numbers of each binary value. We observed 3488 individuals corresponding to Female and 3555 individuals corresponding to Male.

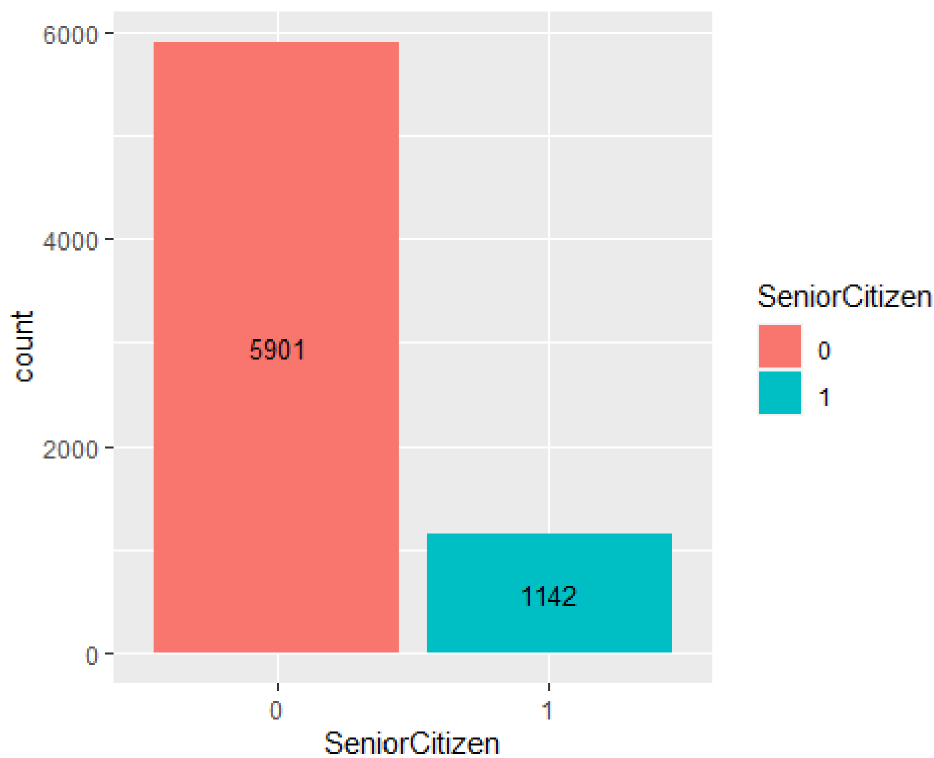
```
summary(df$gender)

## Female   Male
##   3488   3555

ggplot(data=df,aes(gender,fill=gender))+
  geom_bar()+
  stat_count(geom = "text", colour = "black", size = 3.5,
             aes(label = after_stat(count)),position=position_stack(vjus
t=0.5))
```



variable 2: SeniorCitizen

SeniorCitizen is a binary variable composed of two main values: 0 and 1. 0 represents that the customer is not a senior citizen and 1 yes. To visualize the distribution of values in this feature, we used a barplot. Here we observed a more unequal distribution of individuals across these two values: 5901 being not a senior citizen and 1142 being a senior citizen.

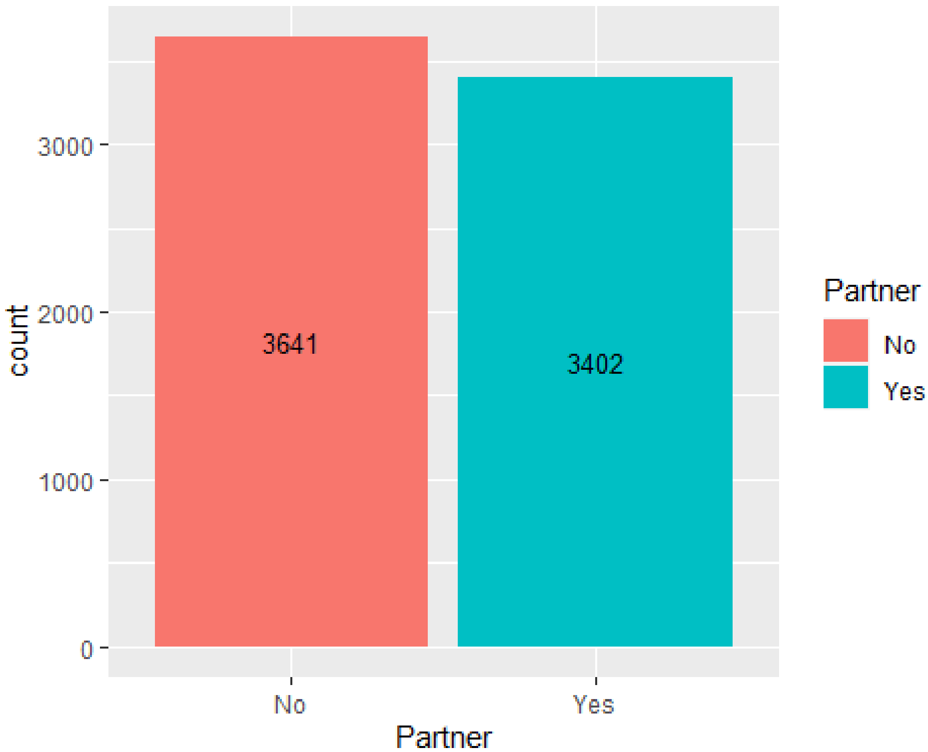
```r
summary(df$SeniorCitizen)

##    0    1 
## 5901 1142

ggplot(data=df,aes(SeniorCitizen,fill=SeniorCitizen))+
  geom_bar()+
  stat_count(geom = "text", colour = "black", size = 3.5,
             aes(label = after_stat(count)),position=position_stack(vjus
t=0.5))
```



variable 3: Partner

Partner is a binary variable with two values: No and Yes (Yes having a partner and No not). We choose a barplot to better understand the distribution of our data across this feature. We did observe a balanced distribution, observing 3641 customer without partner and 3402 with partner.

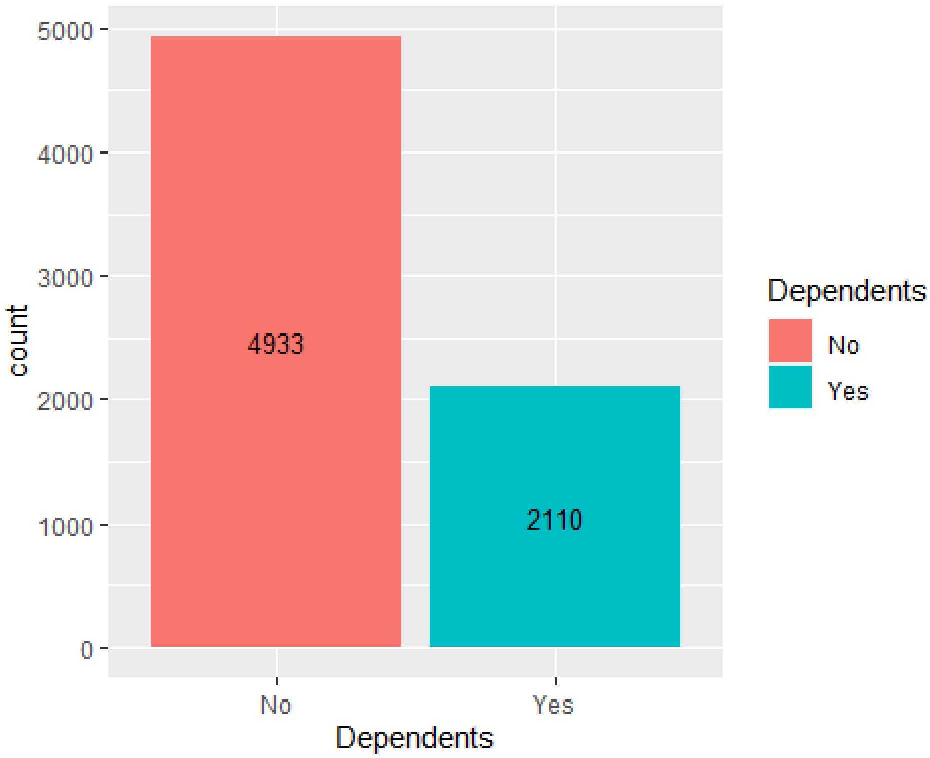
```r
summary(df$Partner)

##   No  Yes 
## 3641 3402

ggplot(data=df,aes(Partner,fill=Partner))+
  geom_bar()+
  stat_count(geom = "text", colour = "black", size = 3.5,
             aes(label = after_stat(count)),position=position_stack(vjus
t=0.5))
```



variable 4: Dependents

Dependents is a binary variable consisting of two main values: Yes (Customer has dependent) and No (Customer has not dependent). Again, we used a bar plot to visualize this variable. We observed 4933 individuals without dependent and 2110 individuals with dependent.

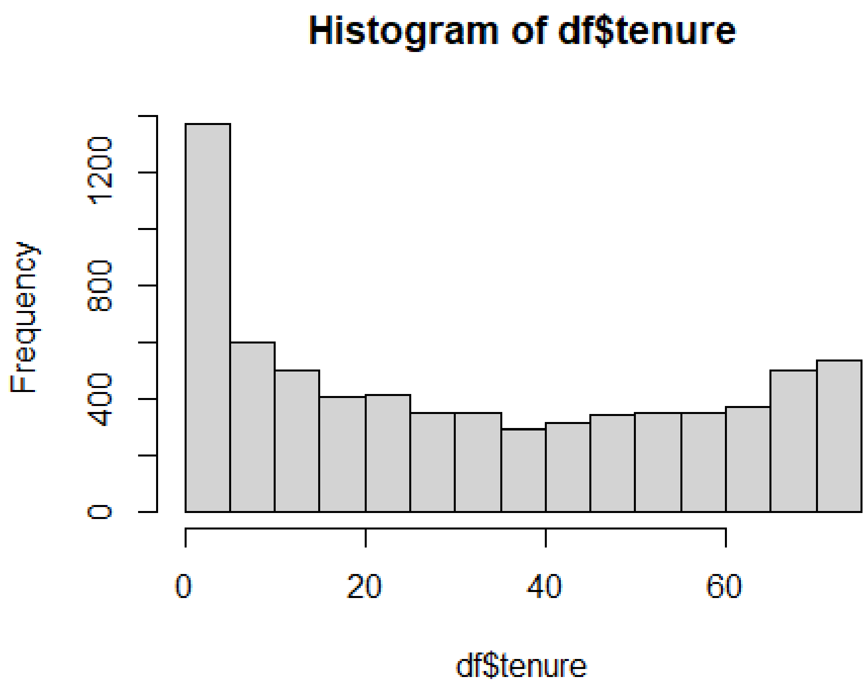
```
summary(df$Dependents)

##   No  Yes
## 4933 2110

ggplot(data=df,aes(Dependents,fill=Dependents))+
  geom_bar()+
  stat_count(geom = "text", colour = "black", size = 3.5,
             aes(label = after_stat(count)),position=position_stack(vjus
t=0.5))
```



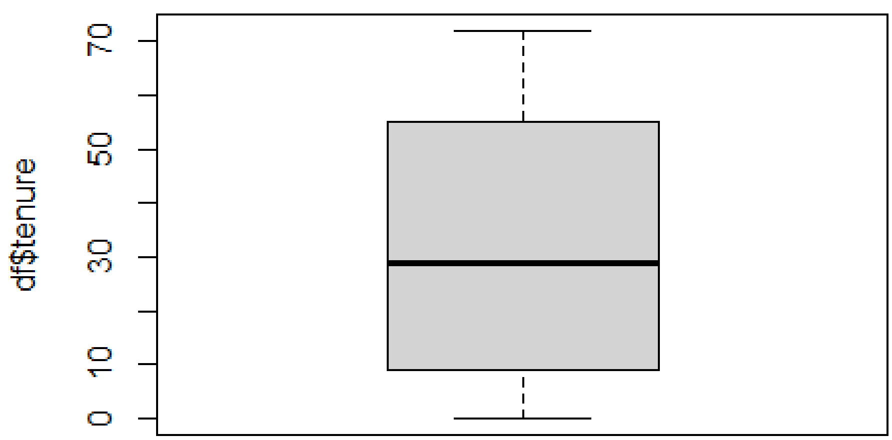
variable 5: tenure

Tenure is a numeric variable. This numeric variable has not NA or missing values. To visualize the distribution of this variable, we used a histogram and a boxplot. We observed that most individuals were comprised between 0 and 5. In the boxplot, we did not observe any outlier in this variable.

```
summary(df$tenure)

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

hist(df$tenure)
```




```
length(Boxplot(df$tenure))
```

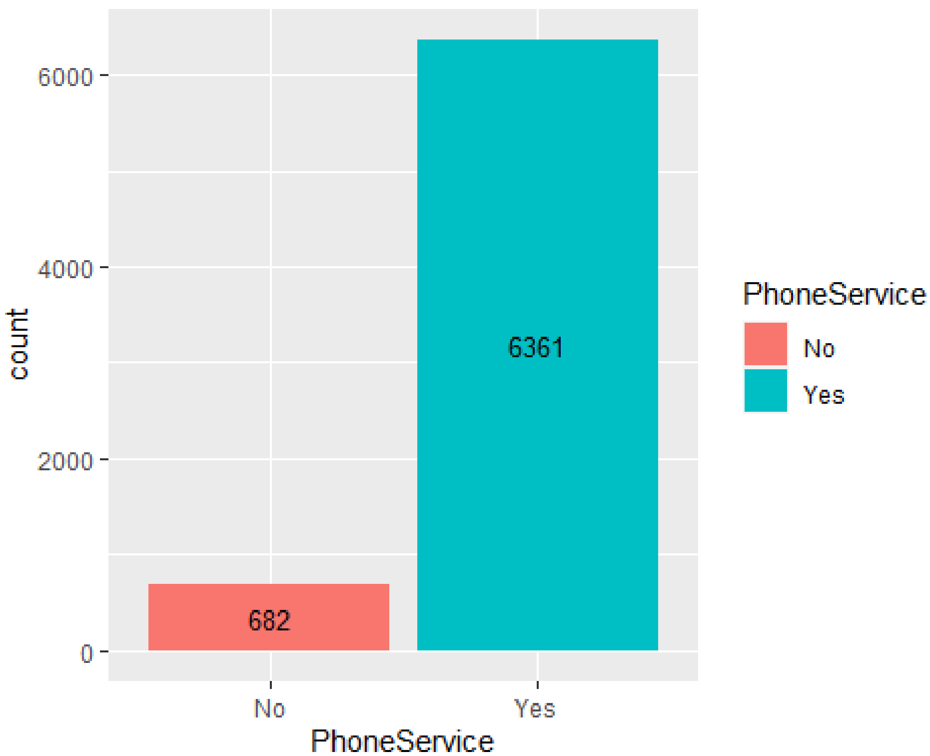



```
## [1] 0
```

variable 6: PhoneService

PhoneService is a binary variable consisting of Yes (Customer with PhoneService) and No (Customer without PhoneService). In the graphic below, we observed an unbalanced distribution of values, observing 682 individuals without Phone service and 6361 individuals with phone service.

```
summary(df$PhoneService)

##   No  Yes
##  682 6361

ggplot(data=df,aes(PhoneService,fill=PhoneService))+
  geom_bar()+
  stat_count(geom = "text", colour = "black", size = 3.5,
             aes(label = after_stat(count)),position=position_stack(vjus
t=0.5))
```



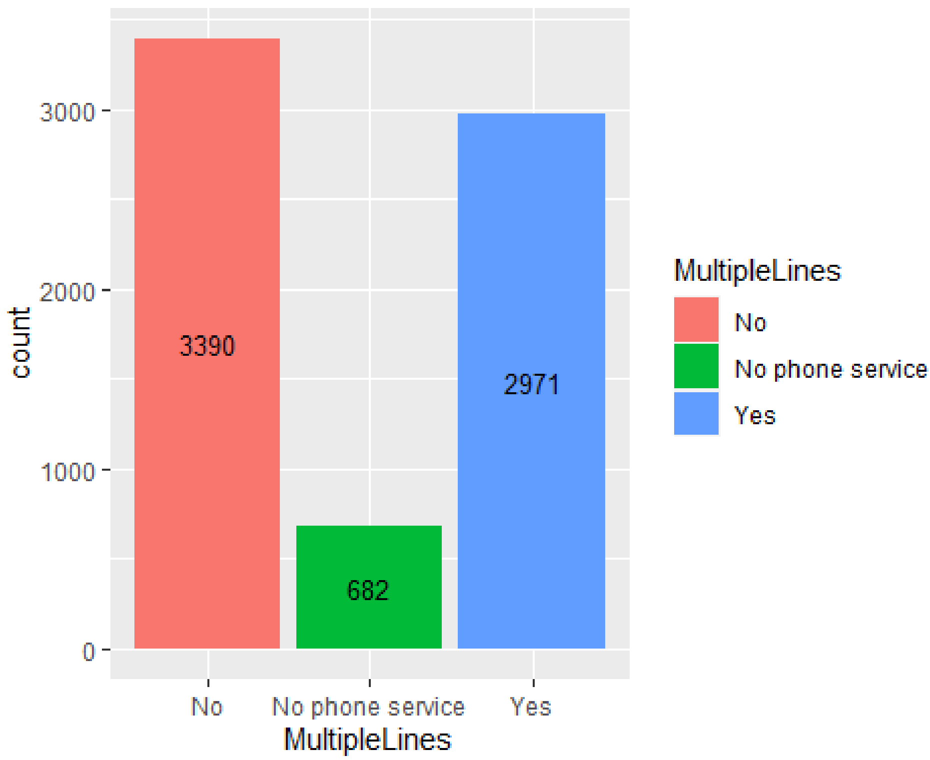
variable 7: MultipleLines

MultipleLines is a factor variable which consists of three values: If customer has Phone Service, how many of them has multiple lines (Yes) or not (No). The third value represents those individual without Phone service. We observed 3390 customers without multiple lines, 2971 with multiple lines and again, consistent with the values from the previous variable (PhoneService), 682 individuals without phone service.

```
summary(df$MultipleLines)
```

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

```
ggplot(data=df,aes(MultipleLines,fill=MultipleLines))+
  geom_bar()+
  stat_count(geom = "text", colour = "black", size = 3.5,
             aes(label = after_stat(count)),position=position_stack(vjus
t=0.5))
```



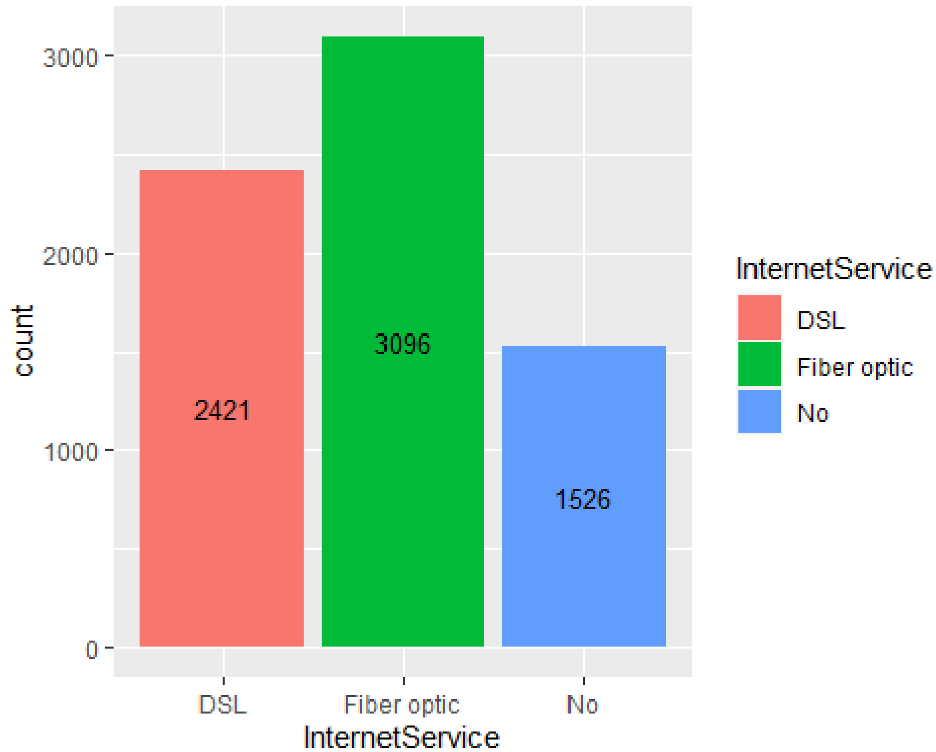
variable 8: InternetService

InternetService is a factor variable consisting of three values: DSL, Fiber Optic and No. In the barplot below, we observed 2421 individuals having DSL, 3096 having Fiber Optic and 1526 without internet service.

```
summary(df$InternetService)
```

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

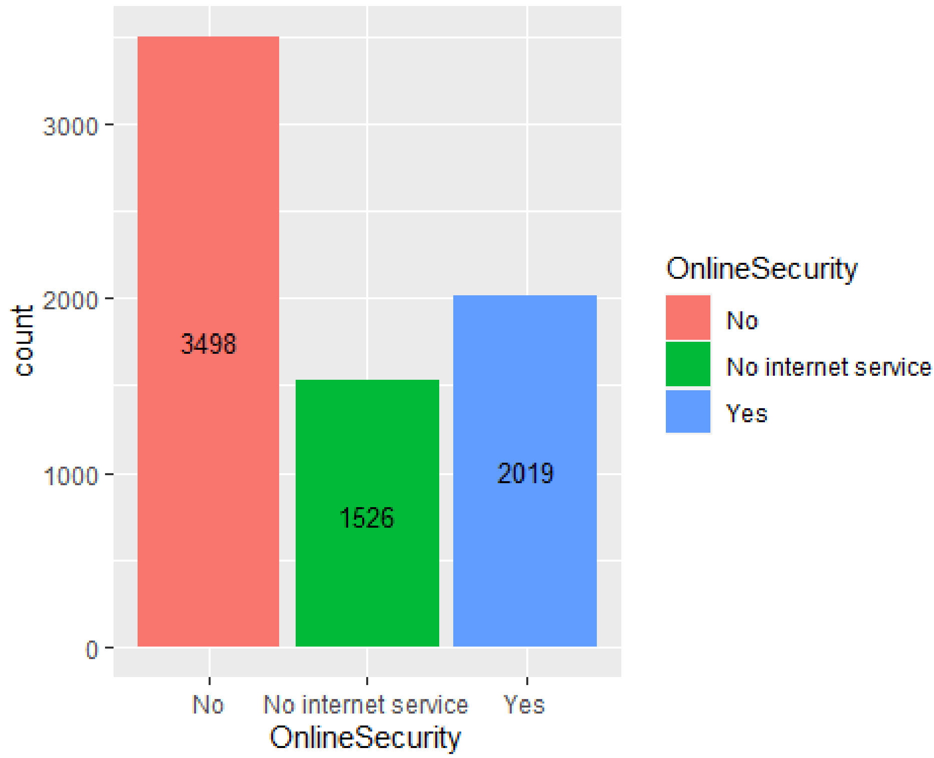
```
ggplot(data=df,aes(InternetService,fill=InternetService))+
  geom_bar()+
  stat_count(geom = "text", colour = "black", size = 3.5,
             aes(label = after_stat(count)),position=position_stack(vjus
t=0.5))
```



variable 9: OnlineSecurity

OnlineSecurity is a factor variable which has three values: From those individuals with internet service, how many of them have online security (Yes) or not (No). The third value corresponds to those customers without internet service. We observed 2019 customers with online security, 3498 without online security, and 1526 individuals without internet service (consistent with previous analysis).

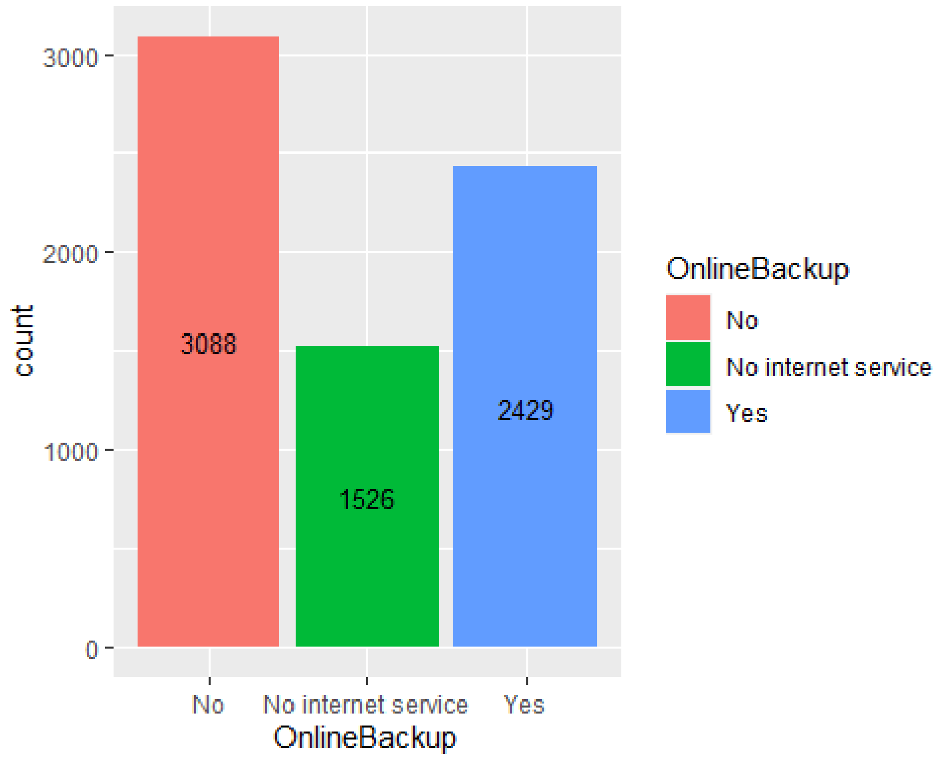
```
summary(df$OnlineSecurity)

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

ggplot(data=df,aes(OnlineSecurity,fill=OnlineSecurity))+
  geom_bar()+
  stat_count(geom = "text", colour = "black", size = 3.5,
            aes(label = after_stat(count)),position=position_stack(vjus
t=0.5))
```



variable 10: OnlineBackup

OnlineBackup is a factor variable which has three values: From those individuals with internet service, how many of them have online backup (Yes) or not (No). The third value corresponds to those customers without internet service. We observed 2429 customers with online backup, 3088 without online backup, and 1526 individuals without internet service (consistent with previous analysis).

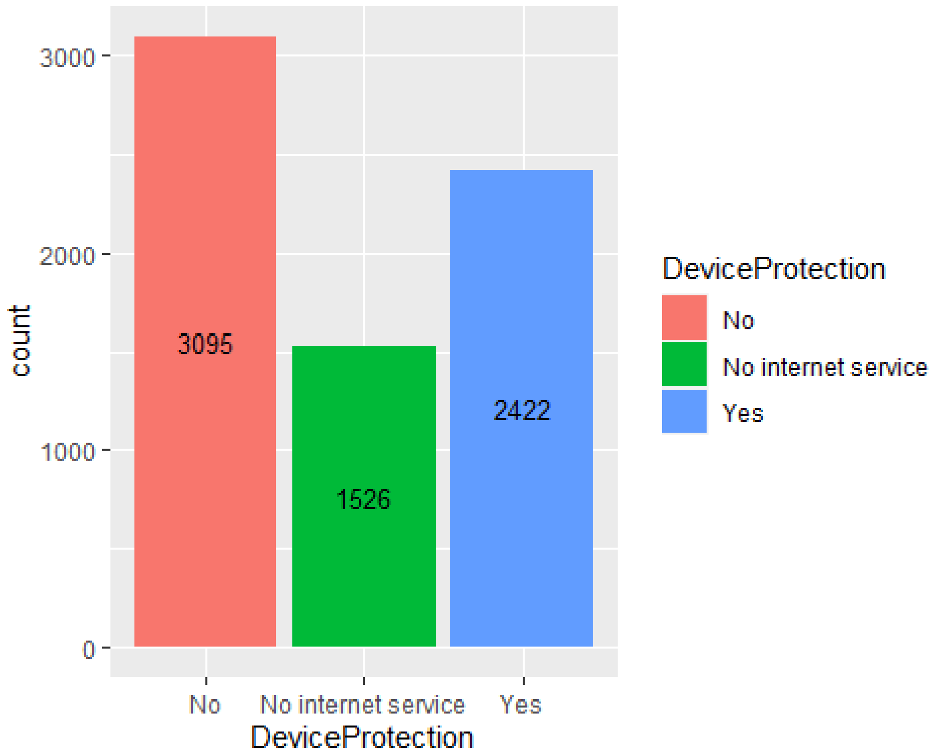
```
summary(df$OnlineBackup)

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

ggplot(data=df,aes(OnlineBackup,fill=OnlineBackup))+
  geom_bar()+
  stat_count(geom = "text", colour = "black", size = 3.5,
              aes(label = after_stat(count)),position=position_stack(vjus
t=0.5))
```



variable 11: DeviceProtection

DeviceProtection is a factor variable which has three values: From those individuals with internet service, how many of them have device protection (Yes) or not (No). The third value corresponds to those customers without internet service. We observed 2422 customers with device protection, 3095 without device protection, and 1526 individuals without internet service (consistent with previous analysis).

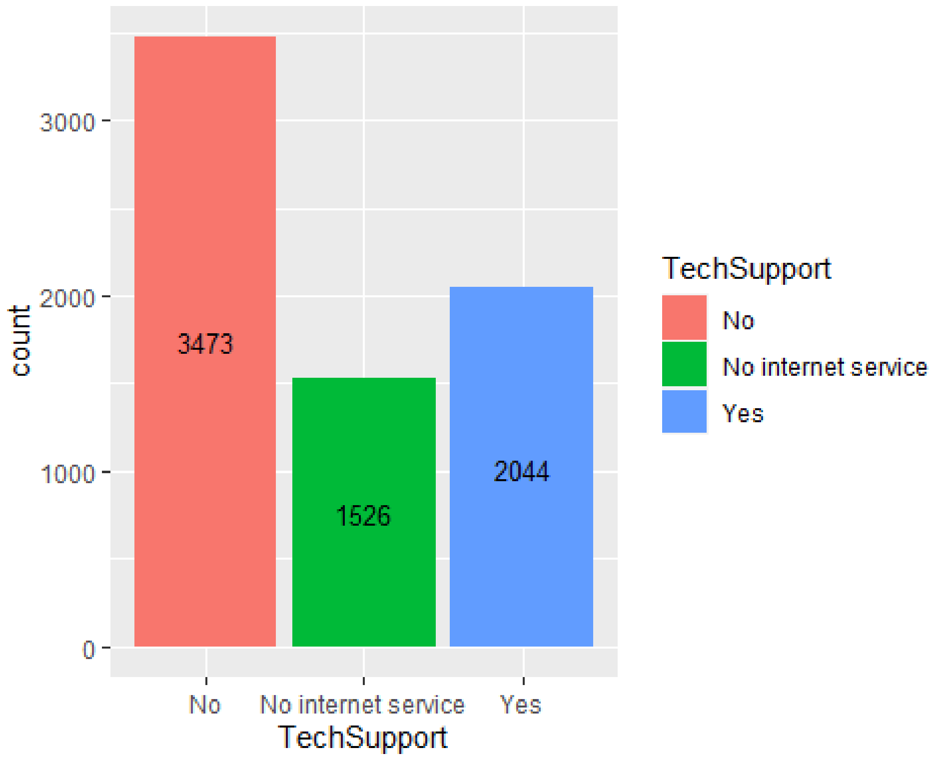
```
summary(df$DeviceProtection)

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

ggplot(data=df,aes(DeviceProtection,fill=DeviceProtection))+
  geom_bar()+
  stat_count(geom = "text", colour = "black", size = 3.5,
             aes(label = after_stat(count)),position=position_stack(vjus
t=0.5))
```



variable 12: TechSupport

TechSupport is a factor variable which has three values: From those individuals with internet service, how many of them have tech support (Yes) or not (No). The third value corresponds to those customers without internet service. We observed 2044 customers with tech support, 3473 without tech support, and 1526 individuals without internet service (consistent with previous analysis).

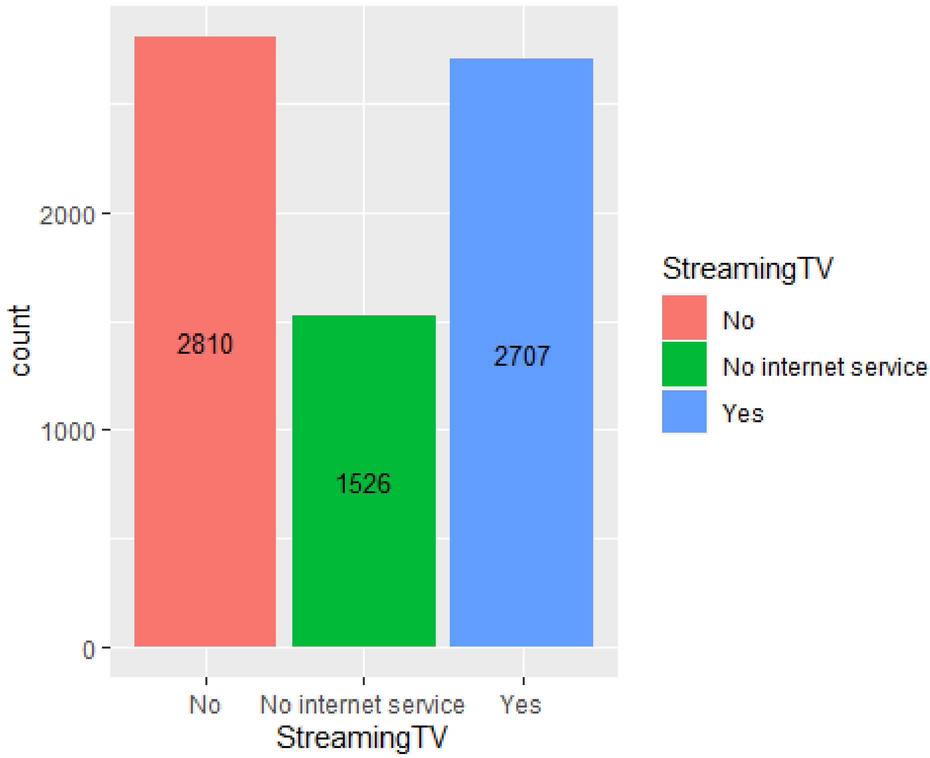
```
summary(df$TechSupport)

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

ggplot(data=df,aes(TechSupport,fill=TechSupport))+
  geom_bar()+
  stat_count(geom = "text", colour = "black", size = 3.5,
            aes(label = after_stat(count)),position=position_stack(vjus
t=0.5))
```



variable 13: StreamingTV

StreamingTV is a factor variable which has three values: From those individuals with internet service, how many of them have StreamingTV (Yes) or not (No). The third value corresponds to those customers without internet service. We observed 2707 customers with StreamingTV, 2810 without StreamingTV, and 1526 individuals without internet service (consistent with previous analysis).

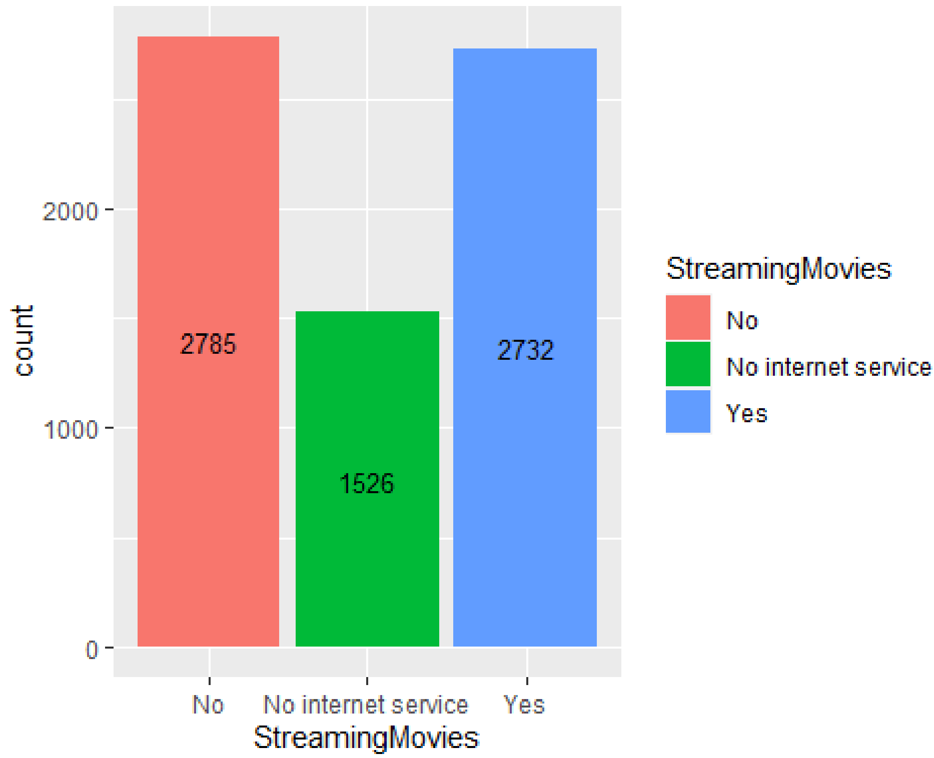
```
summary(df$StreamingTV)

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

ggplot(data=df,aes(StreamingTV,fill=StreamingTV))+
  geom_bar()+
  stat_count(geom = "text", colour = "black", size = 3.5,
              aes(label = after_stat(count)),position=position_stack(vjus
t=0.5))
```



variable 14: StreamingMovies

StreamingMovies is a factor variable which has three values: From those individuals with internet service, how many of them have StreamingMovies (Yes) or not (No). The third value corresponds to those customers without internet service. We observed 2732 customers with StreamingMovies, 2785 without StreamingMovies, and 1526 individuals without internet service (consistent with previous analysis).

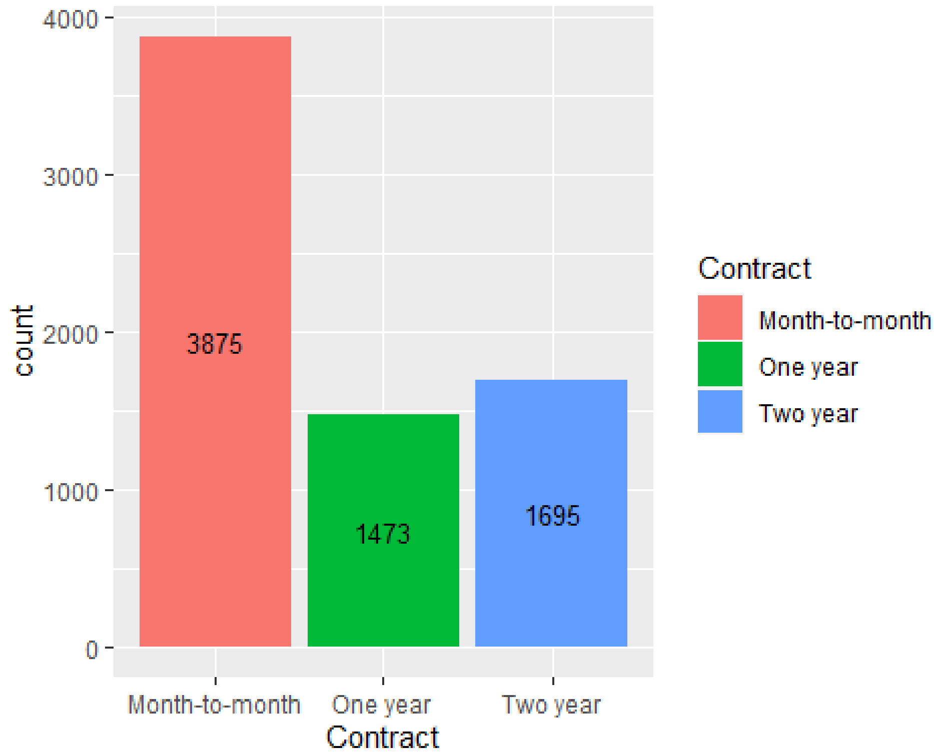
```
summary(df$StreamingMovies)

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

ggplot(data=df,aes(StreamingMovies,fill=StreamingMovies))+
  geom_bar()+
  stat_count(geom = "text", colour = "black", size = 3.5,
             aes(label = after_stat(count)),position=position_stack(vjus
t=0.5))
```



variable 15: Contract

Contract is a factor variable consisting of three values: Month-to-month, One year and Two year. In the plot below, we observed 3875 individuals having a month-to-month contract, 1473 customers with one year contract and the remaining 1695 having a two year contract.

```
summary(df$Contract)

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

ggplot(data=df,aes(Contract,fill=Contract))+
  geom_bar()+
  stat_count(geom = "text", colour = "black", size = 3.5,
             aes(label = after_stat(count)),position=position_stack(vjus
t=0.5))
```



variable 16: PaperlessBilling

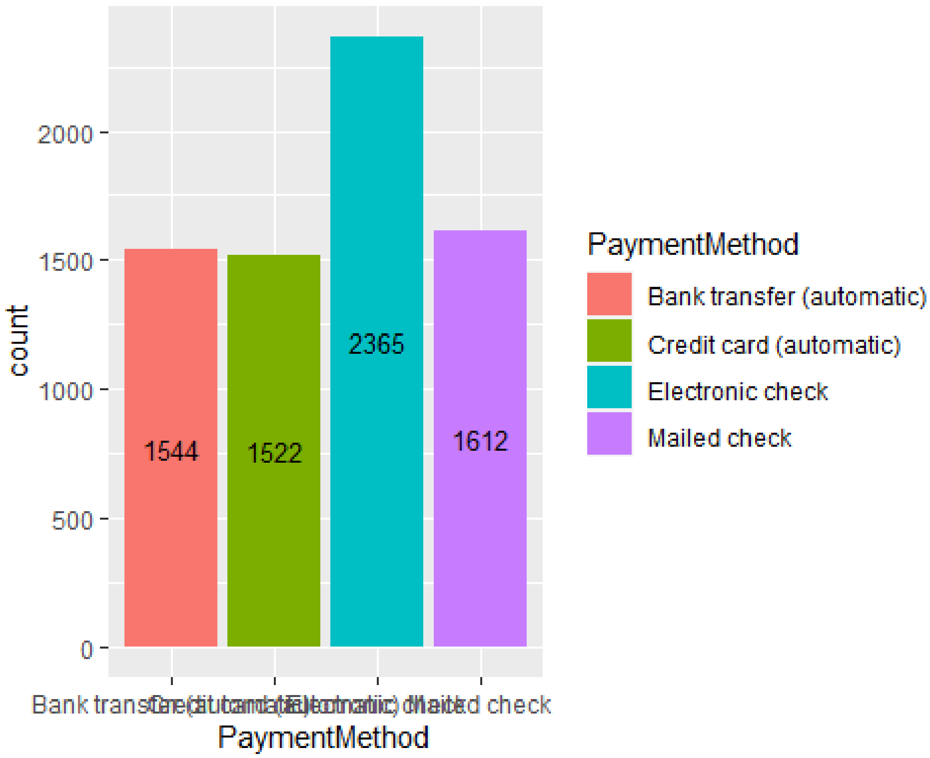
PaperlessBilling is a binary variable. Yes (4171 individuals) represents that the customer has paper billing and No (2872 individuals) the opposite.

```
summary(df$PaperlessBilling)

##   No  Yes
## 2872 4171

ggplot(data=df,aes(PaperlessBilling,fill=PaperlessBilling))+
  geom_bar()+
  stat_count(geom = "text", colour = "black", size = 3.5,
              aes(label = after_stat(count)),position=position_stack(vjus
t=0.5))
```

variable 17: PaymentMethod

PaymentMethod is a factor variable consisting of four main values: Bank transfer (1544 customers), credit card (1522 customers), electronic check (2365 customers) and mailed check (1612 customers).

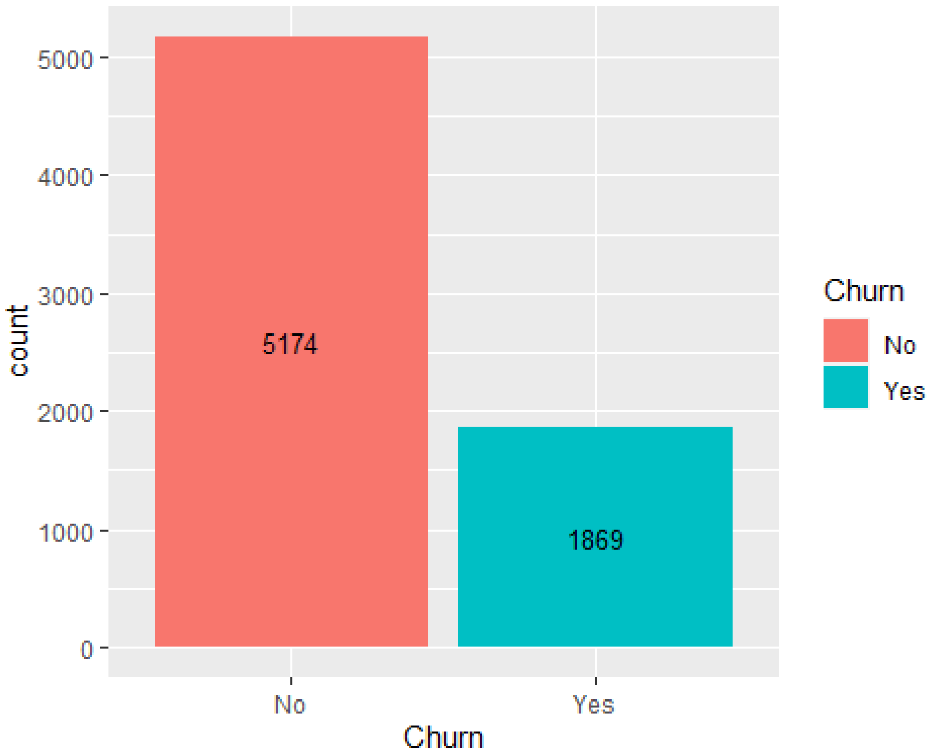
```
summary(df$PaymentMethod)

## Bank transfer (automatic)   Credit card (automatic)          Electroni
c check
##                      1544                      1522
2365
##              Mailed check
##                      1612

ggplot(data=df,aes(PaymentMethod,fill=PaymentMethod))+
  geom_bar()+
  stat_count(geom = "text", colour = "black", size = 3.5,
             aes(label = after_stat(count)),position=position_stack(vjus
t=0.5))
```



variable 18: Churn

Churn is a binary variable representing if customers churned (Yes) or not (No). We found that 1869 individuals churned and 5174 not.

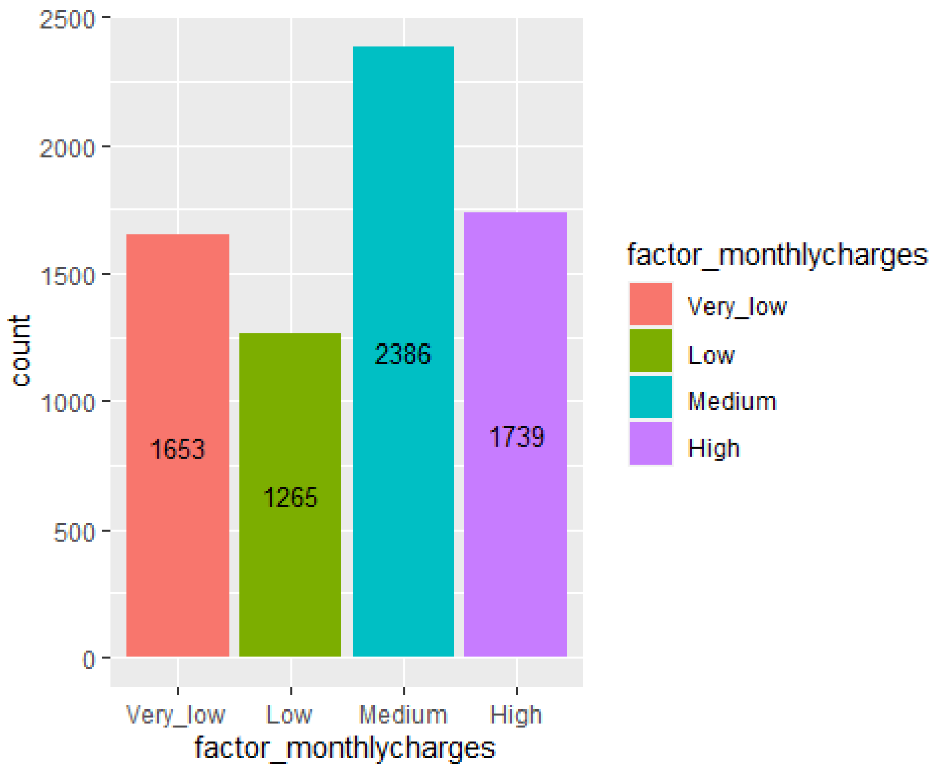
```
summary(df$Churn)

##   No  Yes
## 5174 1869

ggplot(data=df,aes(Churn,fill=Churn))+
  geom_bar()+
  stat_count(geom = "text", colour = "black", size = 3.5,
             aes(label = after_stat(count)),position=position_stack(vjus
t=0.5))
```



variable 19: Monthly Charges (factor)

Monthly charges was originally a numerical variable converted into a categorical variable where each value corresponds to a numeric interval. The values of this feature are_ Very_low, low, medium and high.

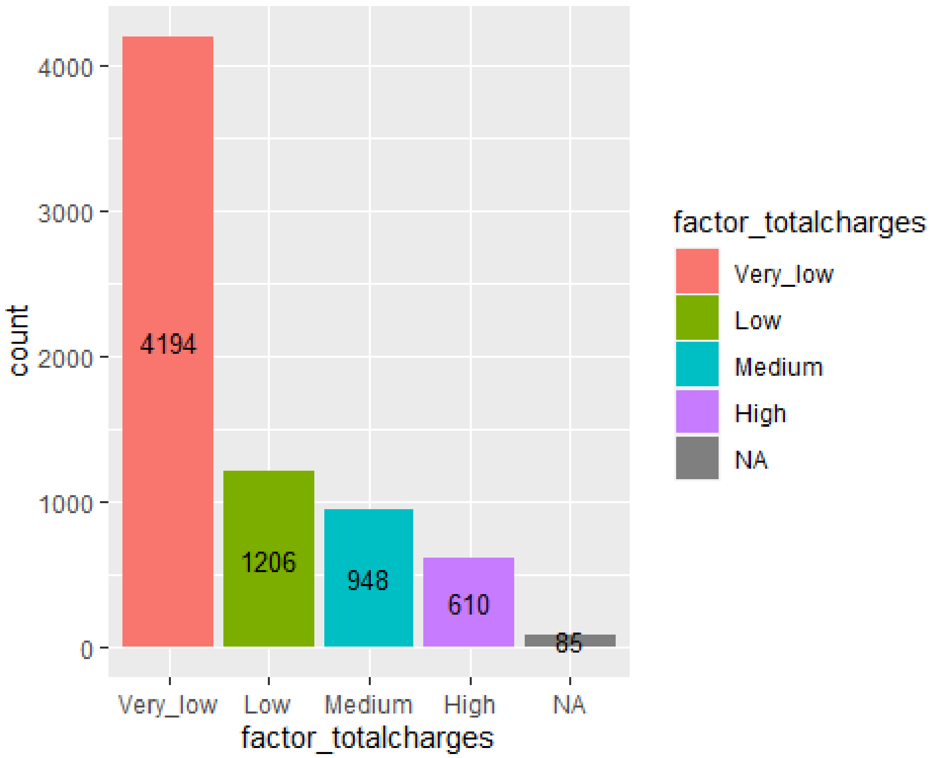
```
summary(df$factor_monthlycharges)

## Very_low      Low   Medium     High
##     1653     1265     2386     1739

ggplot(data=df,aes(factor_monthlycharges,fill=factor_monthlycharges))+
  geom_bar()+
  stat_count(geom = "text", colour = "black", size = 3.5,
             aes(label = after_stat(count)),position=position_stack(vjus
t=0.5))
```



variable 20: Total Charges (factor)

Total charges was originally a numerical variable converted into a categorical variable where each value corresponds to a numeric interval. The values of this feature are_ Very_low, low, medium and high. Here we observed some NA's (85).

```
summary(df$factor_totalcharges)

## Very_low      Low   Medium     High     NA's
##     4194     1206      948      610       85

ggplot(data=df,aes(factor_totalcharges,fill=factor_totalcharges))+
  geom_bar()+
  stat_count(geom = "text", colour = "black", size = 3.5,
              aes(label = after_stat(count)),position=position_stack(vjus
t=0.5))
```


4000
3000
2000
1000
0
count
4194
1206
948
610
85
Very_low
Low
Medium
High
NA
factor_totalcharges
factor_totalcharges
Very_low
Low
Medium
High
NA

Appendix B: Continuation of CatDes Output

```
## StreamingTV=Yes                           69.92981 36.58678 38.435326
## factor_monthlycharges=High                67.22254 22.59374 24.691183
## factor_monthlycharges=Medium              66.09388 30.47932 33.877609
## StreamingTV=No                            66.47687 36.10359 39.897771
## StreamingMovies=No                        66.31957 35.69772 39.542808
## SeniorCitizen=1                           58.31874 12.87205 16.214681
## Partner=No                                67.04202 47.17820 51.696720
## factor_totalcharges=Very_low              68.03805 55.29571 59.704671
## Dependents=No                             68.72086 65.51991 70.041176
## PaperlessBilling=Yes                      66.43491 53.55624 59.221922
## DeviceProtection=No                       60.87237 36.41283 43.944342
## OnlineBackup=No                           60.07124 35.85234 43.844952
## PaymentMethod=Electronic check            54.71459 25.00966 33.579441
## InternetService=Fiber optic               58.10724 34.77000 43.958540
## TechSupport=No                            58.36453 39.17665 49.311373
## OnlineSecurity=No                         58.23328 39.36993 49.666335
## Contract=Month-to-month                   57.29032 42.90684 55.019168
##                                                 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
## factor_monthlycharges=Very_low             5.094110e-80  18.942479
## 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
## factor_totalcharges=Medium                 5.031394e-19   8.911572
## factor_totalcharges=High                   3.707204e-16   8.147762
## 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
## factor_totalcharges=Low                    7.464465e-03   2.675380
## MultipleLines=Yes                          7.843169e-04  -3.358271
## StreamingMovies=Yes                        2.922571e-07  -5.128373
## StreamingTV=Yes                            1.283457e-07  -5.281193
```

```
## factor_monthlycharges=High                  2.192211e-11  -6.692612
## factor_monthlycharges=Medium                3.723184e-23  -9.911155
## 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
## factor_totalcharges=Very_low                3.502145e-37 -12.740926
## 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 55.019168
## OnlineSecurity=No                          41.766724 78.170144 49.666335
## TechSupport=No                             41.635474 77.367576 49.311373
## InternetService=Fiber optic                41.892765 69.395399 43.958540
## PaymentMethod=Electronic check             45.285412 57.303371 33.579441
## OnlineBackup=No                            39.928756 65.971108 43.844952
## DeviceProtection=No                        39.127625 64.794007 43.944342
## PaperlessBilling=Yes                       33.565092 74.906367 59.221922
## Dependents=No                              31.279140 82.557517 70.041176
## factor_totalcharges=Very_low               31.961950 71.910112 59.704671
## Partner=No                                 32.957979 64.205457 51.696720
## SeniorCitizen=1                            41.681261 25.468165 16.214681
## StreamingMovies=No                         33.680431 50.187266 39.542808
## StreamingTV=No                             33.523132 50.401284 39.897771
## factor_monthlycharges=Medium               33.906119 43.285179 33.877609
## factor_monthlycharges=High                 32.777458 30.497592 24.691183
## StreamingTV=Yes                            30.070188 43.552702 38.435326
## StreamingMovies=Yes                        29.941435 43.766720 38.790288
## MultipleLines=Yes                          28.609896 45.478866 42.183729
## factor_totalcharges=Low                    23.466003 15.141787 17.123385
## MultipleLines=No                           25.044248 45.425361 48.132898
## DeviceProtection=Yes                       22.502064 29.159979 34.388755
## OnlineBackup=Yes                           21.531494 27.982879 34.488144
## PaymentMethod=Mailed check                 19.106700 16.479401 22.887974
## factor_totalcharges=High                   13.442623  4.387373  8.661082
## factor_totalcharges=Medium                 15.655577  8.560728 14.510862
## PaymentMethod=Bank transfer (automatic)    16.709845 13.804173 21.922476
## InternetService=DSL                        18.959108 24.558587 34.374556
## PaymentMethod=Credit card (automatic)      15.243101 12.413055 21.610109
## SeniorCitizen=0                            23.606168 74.531835 83.785319
```

```
## Partner=Yes                            19.664903 35.794543 48.303280
## Dependents=Yes                         15.450237 17.442483 29.958824
## TechSupport=Yes                        15.166341 16.586410 29.021724
## OnlineSecurity=Yes                     14.611194 15.783842 28.666761
## Contract=One year                      11.269518  8.881755 20.914383
## PaperlessBilling=No                    16.330084 25.093633 40.778078
## factor_monthlycharges=Very_low          9.800363  8.667737 23.470112
## StreamingMovies=No internet service     7.404980  6.046014 21.666903
## StreamingTV=No internet service         7.404980  6.046014 21.666903
## TechSupport=No internet service         7.404980  6.046014 21.666903
## DeviceProtection=No internet service    7.404980  6.046014 21.666903
## OnlineBackup=No internet service        7.404980  6.046014 21.666903
## OnlineSecurity=No internet service      7.404980  6.046014 21.666903
## InternetService=No                      7.404980  6.046014 21.666903
## Contract=Two year                       2.831858  2.568218 24.066449
##                                              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
## factor_totalcharges=Very_low            3.502145e-37  12.740926
## 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
## factor_monthlycharges=Medium            3.723184e-23   9.911155
## factor_monthlycharges=High              2.192211e-11   6.692612
## StreamingTV=Yes                         1.283457e-07   5.281193
## StreamingMovies=Yes                     2.922571e-07   5.128373
## MultipleLines=Yes                       7.843169e-04   3.358271
## factor_totalcharges=Low                 7.464465e-03  -2.675380
## 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
## factor_totalcharges=High                3.707204e-16  -8.147762
## factor_totalcharges=Medium              5.031394e-19  -8.911572
## 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
## factor_monthlycharges=Very_low              5.094110e-80 -18.942479
## 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
```