Telecom User Churn Prediction

• Dataset: https://www.kaggle.com/blastchar/telco-customer-churn

```
> setwd("~/Downloads")

> #DATA PREPROCESSING
> data <- read.csv("Telco.csv")

> View(data)
```

	customerID	gender	SeniorCitizen	Partner	Dependents	tenure	PhoneService	MultipleLines	InternetService	OnlineSecurity	OnlineBackup	DeviceProtection
1	7590-VHVEG	Female	0	Yes	No	1	No	No phone service	DSL	No	Yes	No
2	5575-GNVDE	Male	0	No	No	34	Yes	No	DSL	Yes	No	Yes
3	3668-QPYBK	Male	0	No	No	2	Yes	No	DSL	Yes	Yes	No
4	7795-CFOCW	Male	0	No	No	45	No	No phone service	DSL	Yes	No	Yes
5	9237-HQITU	Female	0	No	No	2	Yes	No	Fiber optic	No	No	No
6	9305-CDSKC	Female	0	No	No	8	Yes	Yes	Fiber optic	No	No	Yes
7	1452-KIOVK	Male	0	No	Yes	22	Yes	Yes	Fiber optic	No	Yes	No
8	6713-OKOMC	Female	0	No	No	10	No	No phone service	DSL	Yes	No	No
9	7892-POOKP	Female	0	Yes	No	28	Yes	Yes	Fiber optic	No	No	Yes
10	6388-TABGU	Male	0	No	Yes	62	Yes	No	DSL	Yes	Yes	No

TechSupport	StreamingTV	StreamingMovies	Contract	PaperlessBilling	PaymentMethod	MonthlyCharges	TotalCharges	Churn
No	No	No	Month-to-month	Yes	Electronic check	29.85	29.85	No
No	No	No	One year	No	Mailed check	56.95	1889.50	No
No	No	No	Month-to-month	Yes	Mailed check	53.85	108.15	Yes
Yes	No	No	One year	No	Bank transfer (automatic)	42.30	1840.75	No
No	No	No	Month-to-month	Yes	Electronic check	70.70	151.65	Yes
No	Yes	Yes	Month-to-month	Yes	Electronic check	99.65	820.50	Yes
No	Yes	No	Month-to-month	Yes	Credit card (automatic)	89.10	1949.40	No
No	No	No	Month-to-month	No	Mailed check	29.75	301.90	No
Yes	Yes	Yes	Month-to-month	Yes	Electronic check	104.80	3046.05	Yes
No	No	No	One year	No	Bank transfer (automatic)	56.15	3487.95	No

```
> str(data)
'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 service" ...
 $ 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 check" "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(data)
  customerID          gender          SeniorCitizen      Partner           Dependents           tenure      PhoneService       MultipleLines
 Length:7043        Length:7043        Min.   :0.0000   Length:7043        Length:7043        Min.   : 0.00   Length:7043        Length:7043
 Class :character   Class :character   1st Qu.:0.0000   Class :character   Class :character   1st Qu.: 9.00   Class :character   Class :character
 Mode  :character   Mode  :character   Median :0.0000   Mode  :character   Mode  :character   Median :29.00   Mode  :character   Mode  :character
                                       Mean   :0.1621                                         Mean   :32.37
                                       3rd Qu.:0.0000                                         3rd Qu.:55.00
                                       Max.   :1.0000                                         Max.   :72.00

 InternetService    OnlineSecurity     OnlineBackup       DeviceProtection   TechSupport        StreamingTV        StreamingMovies      Contract
 Length:7043        Length:7043        Length:7043        Length:7043        Length:7043        Length:7043        Length:7043        Length:7043
 Class :character   Class :character   Class :character   Class :character   Class :character   Class :character   Class :character   Class :character
 Mode  :character   Mode  :character   Mode  :character   Mode  :character   Mode  :character   Mode  :character   Mode  :character   Mode  :character



 PaperlessBilling   PaymentMethod      MonthlyCharges    TotalCharges       Churn
 Length:7043        Length:7043        Min.   : 18.25   Min.   :  18.8   Length:7043
 Class :character   Class :character   1st Qu.: 35.50   1st Qu.: 401.4   Class :character
 Mode  :character   Mode  :character   Median : 70.35   Median :1397.5   Mode  :character
                                       Mean   : 64.76   Mean   :2283.3
                                       3rd Qu.: 89.85   3rd Qu.:3794.7
                                       Max.   :118.75   Max.   :8684.8
                                                        NA's   :11
```

```
> sapply(data, function(x) sum(is.na(x)))
      customerID           gender    SeniorCitizen          Partner       Dependents           tenure     PhoneService    MultipleLines  InternetService
               0                0                0                0                0                0                0                0                0
  OnlineSecurity     OnlineBackup DeviceProtection      TechSupport      StreamingTV  StreamingMovies         Contract PaperlessBilling    PaymentMethod
               0                0                0                0                0                0                0                0                0
  MonthlyCharges     TotalCharges            Churn
               0               11                0

> data$TotalCharges=ifelse(is.na(data$TotalCharges),ave(data$TotalCharges,FUN=function(x) mean(x, na.rm=TRUE)),data$TotalCharges)

> sapply(data, function(x) sum(is.na(x)))
      customerID           gender    SeniorCitizen          Partner       Dependents           tenure     PhoneService    MultipleLines  InternetService
               0                0                0                0                0                0                0                0                0
  OnlineSecurity     OnlineBackup DeviceProtection      TechSupport      StreamingTV  StreamingMovies         Contract PaperlessBilling    PaymentMethod
               0                0                0                0                0                0                0                0                0
  MonthlyCharges     TotalCharges            Churn
               0                0                0
```

```
> unique(data[c("gender")])
  gender
1 Female
2   Male

> table(data$gender)

Female   Male
  3488   3555

> unique(data[c("SeniorCitizen")])
   SeniorCitizen
1              0
21             1

> table(data$SeniorCitizen)

   0    1
5901 1142

> unique(data[c("Partner")])
  Partner
1     Yes
2      No

> table(data$Partner)

  No  Yes
3641 3402
```

```
> unique(data[c("Dependents")])
  Dependents
1         No
7        Yes

> table(data$Dependents)

  No  Yes
4933 2110

> unique(data[c("MultipleLines")])
     MultipleLines
1 No phone service
2               No
6              Yes

> table(data$MultipleLines)

             No No phone service              Yes
           3390              682             2971

> unique(data[c("PhoneService")])
  PhoneService
1           No
2          Yes

> table(data$PhoneService)

  No  Yes
 682 6361

> unique(data[c("InternetService")])
   InternetService
1              DSL
5      Fiber optic
12              No

> table(data$InternetService)

        DSL Fiber optic          No
       2421        3096        1526
```

```
> unique(data[c("OnlineSecurity")])
       OnlineSecurity
1                  No
2                 Yes
12 No internet service

> table(data$OnlineSecurity)

                  No No internet service                 Yes
                3498                1526                2019

> unique(data[c("OnlineBackup")])
         OnlineBackup
1                 Yes
2                  No
12 No internet service

> table(data$OnlineBackup)

                  No No internet service                 Yes
                3088                1526                2429

> unique(data[c("DeviceProtection")])
      DeviceProtection
1                  No
2                 Yes
12 No internet service

> table(data$DeviceProtection)

                  No No internet service                 Yes
                3095                1526                2422

> unique(data[c("TechSupport")])
          TechSupport
1                  No
4                 Yes
12 No internet service

> table(data$TechSupport)

                  No No internet service                 Yes
                3473                1526                2044
```

```
> unique(data[c("StreamingTV")])
           StreamingTV
1                   No
6                  Yes
12 No internet service

> table(data$StreamingTV)

                  No No internet service                 Yes
                2810                1526                2707

> unique(data[c("StreamingMovies")])
       StreamingMovies
1                   No
6                  Yes
12 No internet service

> table(data$StreamingMovies)

                  No No internet service                 Yes
                2785                1526                2732
```

```
> unique(data[c("Contract")])
         Contract
1  Month-to-month
2        One year
12       Two year

> table(data$Contract)

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

> unique(data[c("PaperlessBilling")])
  PaperlessBilling
1              Yes
2               No

> table(data$PaperlessBilling)

  No  Yes
2872 4171


> unique(data[c("PaymentMethod")])
             PaymentMethod
1         Electronic check
2             Mailed check
4 Bank transfer (automatic)
7   Credit card (automatic)

> table(data$PaymentMethod)

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

> unique(data[c("Churn")])
  Churn
1    No
3   Yes

> table(data$Churn)

  No  Yes
5174 1869
```

```
> data$Partner = factor(data$Partner, levels = c('Yes', 'No'), labels = c(1, 0))
> data$Dependents = factor(data$Dependents, levels = c('Yes', 'No'), labels = c(1, 0))
> data$PhoneService = factor(data$PhoneService, levels = c('Yes', 'No'), labels = c(1, 0))
> data$OnlineSecurity = factor(data$OnlineSecurity, levels = c('Yes', 'No'), labels = c(1, 0))
> data$OnlineBackup = factor(data$OnlineBackup, levels = c('Yes', 'No'), labels = c(1, 0))
> data$DeviceProtection = factor(data$DeviceProtection, levels = c('Yes', 'No'), labels = c(1, 0))
> data$TechSupport = factor(data$TechSupport, levels = c('Yes', 'No'), labels = c(1, 0))
> data$StreamingTV = factor(data$StreamingTV, levels = c('Yes', 'No'), labels = c(1, 0))
> data$StreamingMovies = factor(data$StreamingMovies, levels = c('Yes', 'No'), labels = c(1, 0))
> data$PaperlessBilling = factor(data$PaperlessBilling, levels = c('Yes', 'No'), labels = c(1, 0))
> data$Churn = factor(data$Churn, levels = c('Yes', 'No'), labels = c(1, 0))
> data$Partner <- as.integer(as.character(data$Partner))
> data$Dependents <- as.integer(as.character(data$Dependents))
> data$PhoneService <- as.integer(as.character(data$PhoneService))
> data$OnlineSecurity <- as.integer(as.character(data$OnlineSecurity))
> data$OnlineBackup <- as.integer(as.character(data$OnlineBackup))
> data$DeviceProtection <- as.integer(as.character(data$DeviceProtection))
> data$TechSupport <- as.integer(as.character(data$TechSupport))
> data$StreamingTV <- as.integer(as.character(data$StreamingTV))
> data$StreamingMovies <- as.integer(as.character(data$StreamingMovies))
> data$PaperlessBilling <- as.integer(as.character(data$PaperlessBilling))
> data$Churn <- as.integer(as.character(data$Churn))
```

	customerID	gender	SeniorCitizen	Partner	Dependents	tenure	PhoneService	MultipleLines	InternetService	OnlineSecurity	OnlineBackup	DeviceProtection
1	7590-VHVEG	Female	0	1	0	1	0	No phone service	DSL	0	1	0
2	5575-GNVDE	Male	0	0	0	34	1	No	DSL	1	0	1
3	3668-QPYBK	Male	0	0	0	2	1	No	DSL	1	1	0
4	7795-CFOCW	Male	0	0	0	45	0	No phone service	DSL	1	0	1
5	9237-HQITU	Female	0	0	0	2	1	No	Fiber optic	0	0	0
6	9305-CDSKC	Female	0	0	0	8	1	Yes	Fiber optic	0	0	1
7	1452-KIOVK	Male	0	0	1	22	1	Yes	Fiber optic	0	1	0
8	6713-OKOMC	Female	0	0	0	10	0	No phone service	DSL	1	0	0
9	7892-POOKP	Female	0	1	0	28	1	Yes	Fiber optic	0	0	1

TechSupport	StreamingTV	StreamingMovies	Contract	PaperlessBilling	PaymentMethod	MonthlyCharges	TotalCharges	Churn
0	0	0	Month-to-month	1	Electronic check	29.85	29.85	0
0	0	0	One year	0	Mailed check	56.95	1889.50	0
0	0	0	Month-to-month	1	Mailed check	53.85	108.15	1
1	0	0	One year	0	Bank transfer (automatic)	42.30	1840.75	0
0	0	0	Month-to-month	1	Electronic check	70.70	151.65	1
0	1	1	Month-to-month	1	Electronic check	99.65	820.50	1
0	1	0	Month-to-month	1	Credit card (automatic)	89.10	1949.40	0
0	0	0	Month-to-month	0	Mailed check	29.75	301.90	0
1	1	1	Month-to-month	1	Electronic check	104.80	3046.05	1

```
> str(data)
'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         : int  1 0 0 0 0 0 0 0 1 0 ...
 $ Dependents      : int  0 0 0 0 0 0 1 0 0 1 ...
 $ tenure          : int  1 34 2 45 2 8 22 10 28 62 ...
 $ PhoneService    : int  0 1 1 0 1 1 1 0 1 1 ...
 $ MultipleLines   : chr  "No phone service" "No" "No" "No phone service" ...
 $ InternetService : chr  "DSL" "DSL" "DSL" "DSL" ...
 $ OnlineSecurity  : int  0 1 1 1 0 0 0 1 0 1 ...
 $ OnlineBackup    : int  1 0 1 0 0 0 1 0 0 1 ...
 $ DeviceProtection: int  0 1 0 1 0 1 0 0 1 0 ...
 $ TechSupport     : int  0 0 0 1 0 0 0 0 1 0 ...
 $ StreamingTV     : int  0 0 0 0 0 1 1 0 1 0 ...
 $ StreamingMovies : int  0 0 0 0 0 1 0 0 1 0 ...
 $ Contract        : chr  "Month-to-month" "One year" "Month-to-month" "One year" ...
 $ PaperlessBilling: int  1 0 1 0 1 1 1 0 1 0 ...
 $ PaymentMethod   : chr  "Electronic check" "Mailed check" "Mailed check" "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           : int  0 0 1 0 1 1 0 0 1 0 ...
```

```
> data$Partner = factor(data$Partner, levels = c('Yes', 'No'), labels = c(1, 0))
> data$Dependents = factor(data$Dependents, levels = c('Yes', 'No'), labels = c(1, 0))
> data$PhoneService = factor(data$PhoneService, levels = c('Yes', 'No'), labels = c(1, 0))
> data$OnlineSecurity = factor(data$OnlineSecurity, levels = c('Yes', 'No'), labels = c(1, 0))
> data$OnlineBackup = factor(data$OnlineBackup, levels = c('Yes', 'No'), labels = c(1, 0))
> data$DeviceProtection = factor(data$DeviceProtection, levels = c('Yes', 'No'), labels = c(1, 0))
> data$TechSupport = factor(data$TechSupport, levels = c('Yes', 'No'), labels = c(1, 0))
> data$StreamingTV = factor(data$StreamingTV, levels = c('Yes', 'No'), labels = c(1, 0))
> data$StreamingMovies = factor(data$StreamingMovies, levels = c('Yes', 'No'), labels = c(1, 0))
> data$PaperlessBilling = factor(data$PaperlessBilling, levels = c('Yes', 'No'), labels = c(1, 0))
> data$Churn = factor(data$Churn, levels = c('Yes', 'No'), labels = c(1, 0))
> data$Partner <- as.integer(as.character(data$Partner))
> data$Dependents <- as.integer(as.character(data$Dependents))
> data$PhoneService <- as.integer(as.character(data$PhoneService))
> data$OnlineSecurity <- as.integer(as.character(data$OnlineSecurity))
> data$OnlineBackup <- as.integer(as.character(data$OnlineBackup))
> data$DeviceProtection <- as.integer(as.character(data$DeviceProtection))
> data$TechSupport <- as.integer(as.character(data$TechSupport))
> data$StreamingTV <- as.integer(as.character(data$StreamingTV))
> data$StreamingMovies <- as.integer(as.character(data$StreamingMovies))
> data$PaperlessBilling <- as.integer(as.character(data$PaperlessBilling))
> data$Churn <- as.integer(as.character(data$Churn))
```

	customerID	gender	SeniorCitizen	Partner	Dependents	tenure	PhoneService	MultipleLines	InternetService	OnlineSecurity	OnlineBackup	DeviceProtection
1	7590-VHVEG	Female	0	1	0	1	0	No phone service	DSL	0	1	0
2	5575-GNVDE	Male	0	0	0	34	1	No	DSL	1	0	1
3	3668-QPYBK	Male	0	0	0	2	1	No	DSL	1	1	0
4	7795-CFOCW	Male	0	0	0	45	0	No phone service	DSL	1	0	1
5	9237-HQITU	Female	0	0	0	2	1	No	Fiber optic	0	0	0
6	9305-CDSKC	Female	0	0	0	8	1	Yes	Fiber optic	0	0	1
7	1452-KIOVK	Male	0	0	1	22	1	Yes	Fiber optic	0	1	0
8	6713-OKOMC	Female	0	0	0	10	0	No phone service	DSL	1	0	0
9	7892-POOKP	Female	0	1	0	28	1	Yes	Fiber optic	0	0	1

TechSupport	StreamingTV	StreamingMovies	Contract	PaperlessBilling	PaymentMethod	MonthlyCharges	TotalCharges	Churn
0	0	0	Month-to-month	1	Electronic check	29.85	29.85	0
0	0	0	One year	0	Mailed check	56.95	1889.50	0
0	0	0	Month-to-month	1	Mailed check	53.85	108.15	1
1	0	0	One year	0	Bank transfer (automatic)	42.30	1840.75	0
0	0	0	Month-to-month	1	Electronic check	70.70	151.65	1
0	1	1	Month-to-month	1	Electronic check	99.65	820.50	1
0	1	0	Month-to-month	1	Credit card (automatic)	89.10	1949.40	0
0	0	0	Month-to-month	0	Mailed check	29.75	301.90	0
1	1	1	Month-to-month	1	Electronic check	104.80	3046.05	1

```
> sapply(data, function(x) sum(is.na(x)))
      customerID           gender    SeniorCitizen          Partner       Dependents           tenure     PhoneService    MultipleLines  InternetService
               0                0                0                0                0                0                0                0                0
  OnlineSecurity     OnlineBackup DeviceProtection      TechSupport      StreamingTV  StreamingMovies         Contract PaperlessBilling    PaymentMethod
            1526             1526             1526             1526             1526             1526                0                0                0
  MonthlyCharges     TotalCharges            Churn
               0                0                0
```

	customerID	gender	SeniorCitizen	Partner	Dependents	tenure	PhoneService	MultipleLines	InternetService	OnlineSecurity	OnlineBackup	DeviceProtection
11	9763-GRSKD	Male	0	1	1	13	1	No	DSL	1	0	
12	7469-LKBCI	Male	0	0	0	16	1	No	No	*NA*	*NA*	*NA*
13	8091-TTVAX	Male	0	1	0	58	1	Yes	Fiber optic	0	0	
14	0280-XJGEX	Male	0	0	0	49	1	Yes	Fiber optic	0	1	
15	5129-JLPIS	Male	0	0	0	25	1	No	Fiber optic	1	0	
16	3655-SNQYZ	Female	0	1	1	69	1	Yes	Fiber optic	1	1	
17	8191-XWSZG	Female	0	0	0	52	1	No	No	*NA*	*NA*	*NA*
18	9959-WOFKT	Male	0	0	1	71	1	Yes	Fiber optic	1	0	

```
> data$OnlineSecurity=ifelse(is.na(data$OnlineSecurity),ave(data$OnlineSecurity,FUN=function(x) mean(x, na.rm=TRUE)),data$OnlineSecurity)

> data$OnlineBackup=ifelse(is.na(data$OnlineBackup),ave(data$OnlineBackup,FUN=function(x) mean(x, na.rm=TRUE)),data$OnlineBackup)

> data$DeviceProtection=ifelse(is.na(data$DeviceProtection),ave(data$DeviceProtection,FUN=function(x) mean(x, na.rm=TRUE)),data$DeviceProtection)

> data$TechSupport=ifelse(is.na(data$TechSupport),ave(data$TechSupport,FUN=function(x) mean(x, na.rm=TRUE)),data$TechSupport)

> data$StreamingTV=ifelse(is.na(data$StreamingTV),ave(data$StreamingTV,FUN=function(x) mean(x, na.rm=TRUE)),data$StreamingTV)

> data$StreamingMovies=ifelse(is.na(data$StreamingMovies),ave(data$StreamingMovies,FUN=function(x) mean(x, na.rm=TRUE)),data$StreamingMovies)

> sapply(data, function(x) sum(is.na(x)))
      customerID           gender    SeniorCitizen          Partner       Dependents           tenure     PhoneService    MultipleLines  InternetService
               0                0                0                0                0                0                0                0                0
  OnlineSecurity     OnlineBackup DeviceProtection      TechSupport      StreamingTV  StreamingMovies         Contract PaperlessBilling    PaymentMethod
               0                0                0                0                0                0                0                0                0
  MonthlyCharges     TotalCharges            Churn
               0                0                0
```

	customerID	gender	SeniorCitizen	Partner	Dependents	tenure	PhoneService	MultipleLines	InternetService	OnlineSecurity	OnlineBackup	DeviceProtection
11	9763-GRSKD	Male	0	1	1	13	1	No	DSL	1.0000000	0.0000000	0.00000
12	7469-LKBCI	Male	0	0	0	16	1	No	No	0.3659598	0.4402755	0.43900
13	8091-TTVAX	Male	0	1	0	58	1	Yes	Fiber optic	0.0000000	0.0000000	1.00000
14	0280-XJGEX	Male	0	0	0	49	1	Yes	Fiber optic	0.0000000	1.0000000	1.00000
15	5129-JLPIS	Male	0	0	0	25	1	No	Fiber optic	1.0000000	0.0000000	1.00000
16	3655-SNQYZ	Female	0	1	1	69	1	Yes	Fiber optic	1.0000000	1.0000000	1.00000
17	8191-XWSZG	Female	0	0	0	52	1	No	No	0.3659598	0.4402755	0.43900
18	9959-WOFKT	Male	0	0	1	71	1	Yes	Fiber optic	1.0000000	0.0000000	1.00000

	customerID	gender	SeniorCitizen	Partner	Dependents	tenure	PhoneService	MultipleLines	InternetService	OnlineSecurity	OnlineBackup	DeviceProtection
1	7590-VHVEG	Female	0	1	0	1	0	No phone service	DSL	0.0000000	1.0000000	0.00000
2	5575-GNVDE	Male	0	0	0	34	1	No	DSL	1.0000000	0.0000000	1.00000
3	3668-QPYBK	Male	0	0	0	2	1	No	DSL	1.0000000	1.0000000	0.00000
4	7795-CFOCW	Male	0	0	0	45	0	No phone service	DSL	1.0000000	0.0000000	1.00000
5	9237-HQITU	Female	0	0	0	2	1	No	Fiber optic	0.0000000	0.0000000	0.00000

	customerID	gender	SeniorCitizen	Partner	Dependents	tenure	PhoneService	MultipleLines	InternetService	OnlineSecurity	OnlineBackup	DeviceProtection
1	7590-VHVEG	Female	0	1	0	1	0	No phone service	DSL	0.0000000	1.0000000	0.00000
2	5575-GNVDE	Male	0	0	0	34	1	No	DSL	1.0000000	0.0000000	1.00000
3	3668-QPYBK	Male	0	0	0	2	1	No	DSL	1.0000000	1.0000000	0.00000
4	7795-CFOCW	Male	0	0	0	45	0	No phone service	DSL	1.0000000	0.0000000	1.00000
5	9237-HQITU	Female	0	0	0	2	1	No	Fiber optic	0.0000000	0.0000000	0.00000

```
> data$MultipleLines <- as.factor(mapvalues(data$MultipleLines,from=c("No phone service"),to=c("No")))
```

	customerID	gender	SeniorCitizen	Partner	Dependents	tenure	PhoneService	MultipleLines	InternetService	OnlineSecurity	OnlineBackup	DeviceProtection
1	7590-VHVEG	Female	0	1	0	1	0	No	DSL	0.0000000	1.0000000	0.00000
2	5575-GNVDE	Male	0	0	0	34	1	No	DSL	1.0000000	0.0000000	1.00000
3	3668-QPYBK	Male	0	0	0	2	1	No	DSL	1.0000000	1.0000000	0.00000
4	7795-CFOCW	Male	0	0	0	45	0	No	DSL	1.0000000	0.0000000	1.00000
5	9237-HQITU	Female	0	0	0	2	1	No	Fiber optic	0.0000000	0.0000000	0.00000

```
> data$customerID <- NULL
```

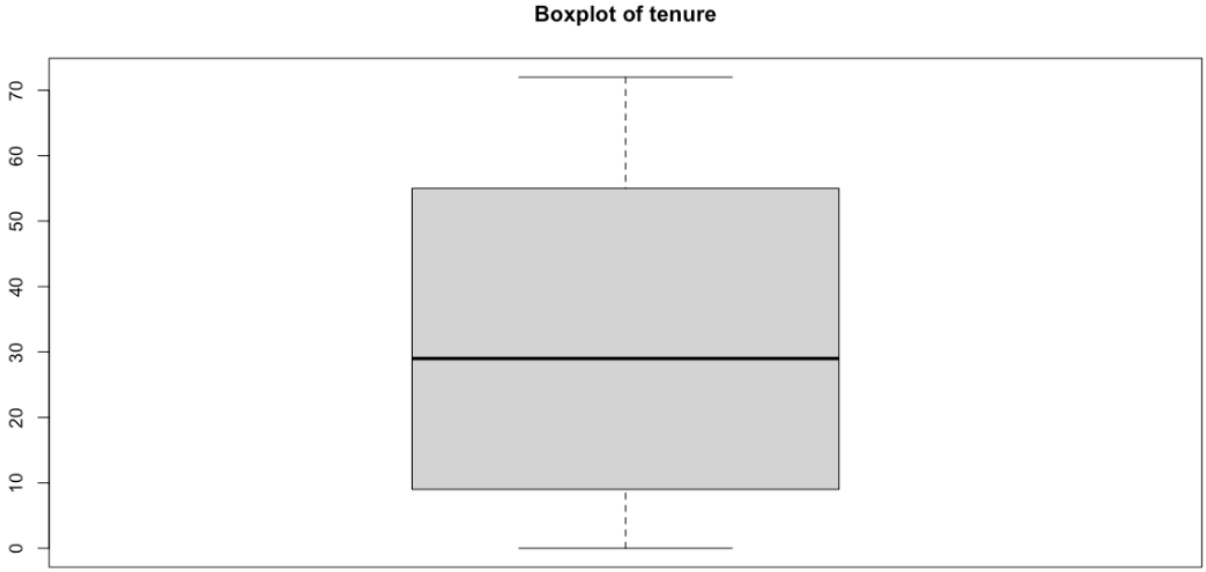
	gender	SeniorCitizen	Partner	Dependents	tenure	PhoneService	MultipleLines	InternetService	OnlineSecurity	OnlineBackup	DeviceProtection	TechSupport
1	Female	0	1	0	1	0	No	DSL	0.0000000	1.0000000	0.0000000	0.0000
2	Male	0	0	0	34	1	No	DSL	1.0000000	0.0000000	1.0000000	0.0000
3	Male	0	0	0	2	1	No	DSL	1.0000000	1.0000000	0.0000000	0.0000
4	Male	0	0	0	45	0	No	DSL	1.0000000	0.0000000	1.0000000	1.0000
5	Female	0	0	0	2	1	No	Fiber optic	0.0000000	0.0000000	0.0000000	0.0000

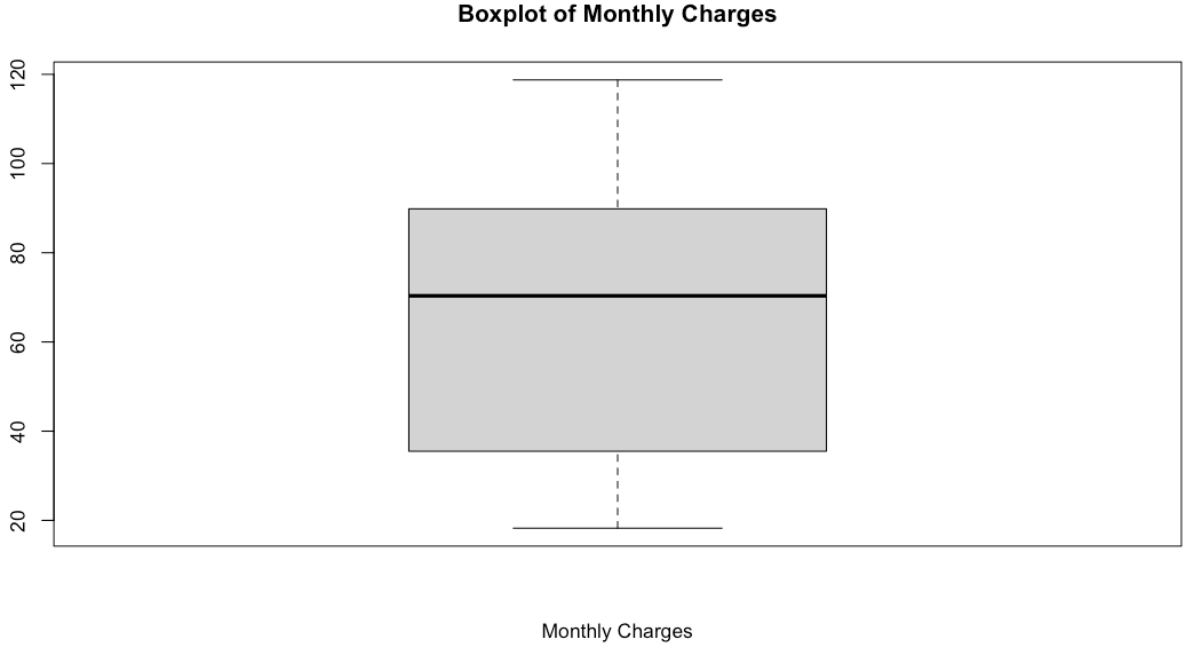
```
> min(data$tenure)
[1] 0

> max(data$tenure)
[1] 72
```

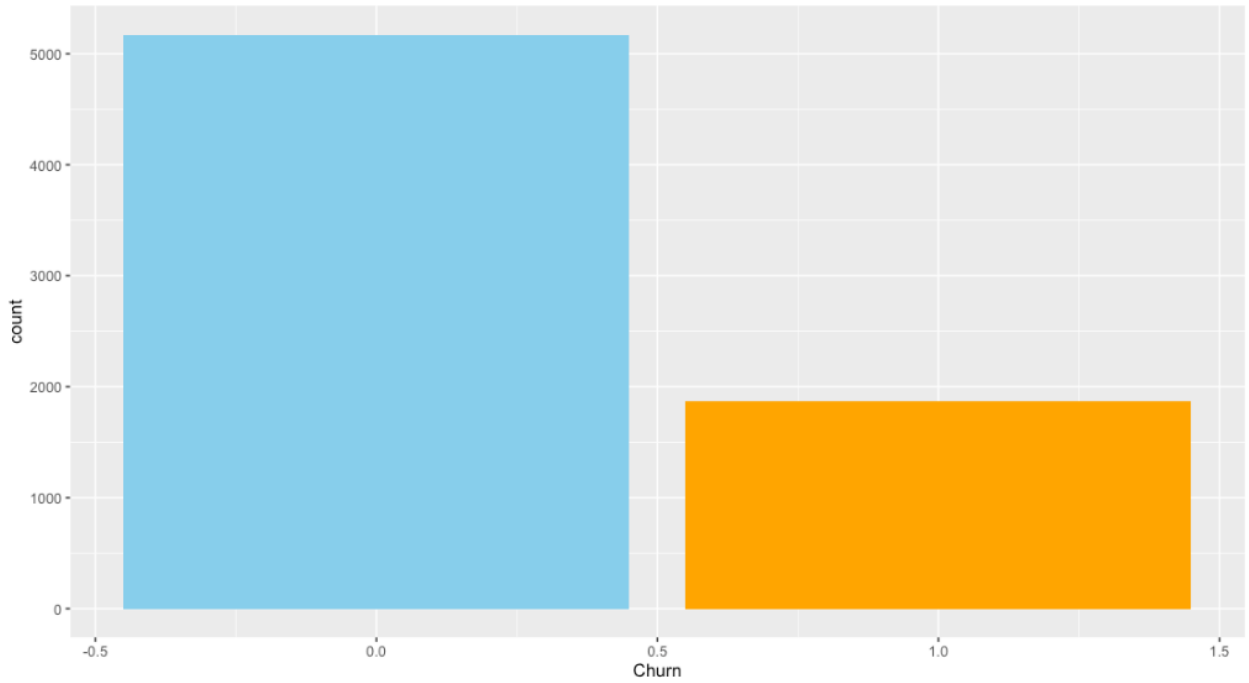
```
> data$MultipleLines = factor(data$MultipleLines, levels = c('Yes', 'No'), labels = c(1, 0))

> data$MultipleLines <- as.integer(as.character(data$MultipleLines))
```

	gender	SeniorCitizen	Partner	Dependents	tenure	PhoneService	MultipleLines	InternetService	OnlineSecurity	OnlineBackup	DeviceProtection	TechSupport
1	Female	0	1	0	1	0	0	DSL	0.0000000	1.0000000	0.0000000	0.0000
2	Male	0	0	0	34	1	0	DSL	1.0000000	0.0000000	1.0000000	0.0000
3	Male	0	0	0	2	1	0	DSL	1.0000000	1.0000000	0.0000000	0.0000
4	Male	0	0	0	45	0	0	DSL	1.0000000	0.0000000	1.0000000	1.0000
5	Female	0	0	0	2	1	0	Fiber optic	0.0000000	0.0000000	0.0000000	0.0000

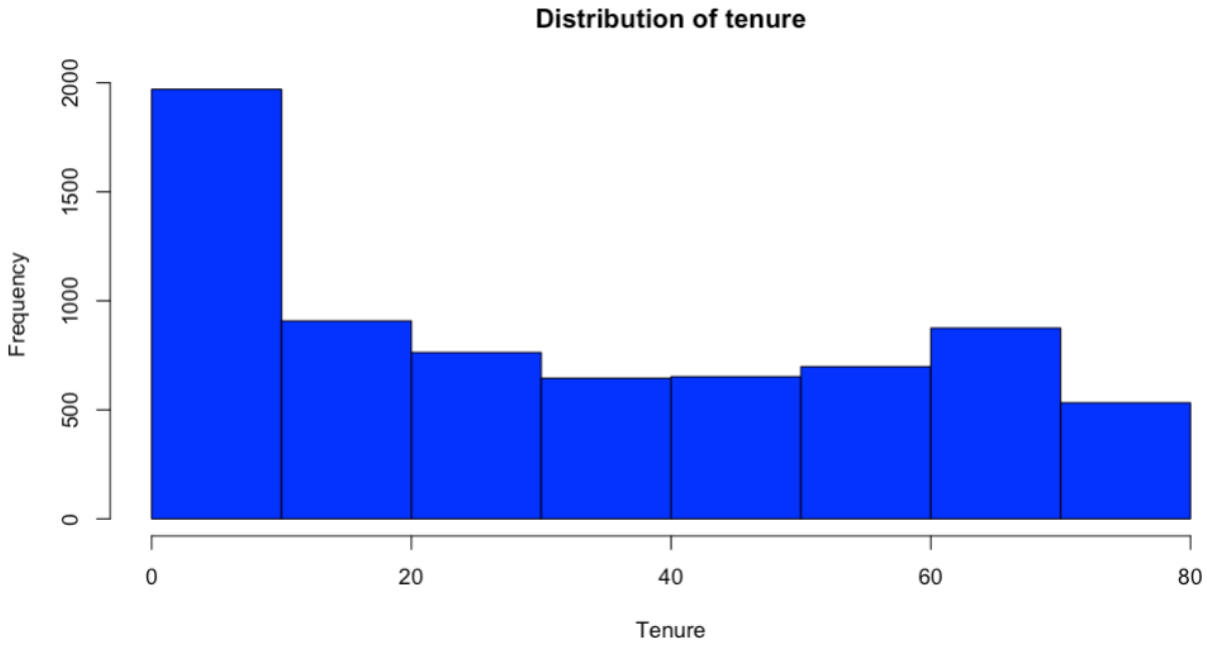
```
> boxplot(data$tenure,main="Boxplot of tenure",xlab="Tenure")

> boxplot(data$MonthlyCharges,main="Boxplot of Monthly Charges",xlab="Monthly Charges")

> ggplot(data, aes(x = Churn))+ geom_histogram(stat = "count", fill = c("sky blue", "orange"))
```

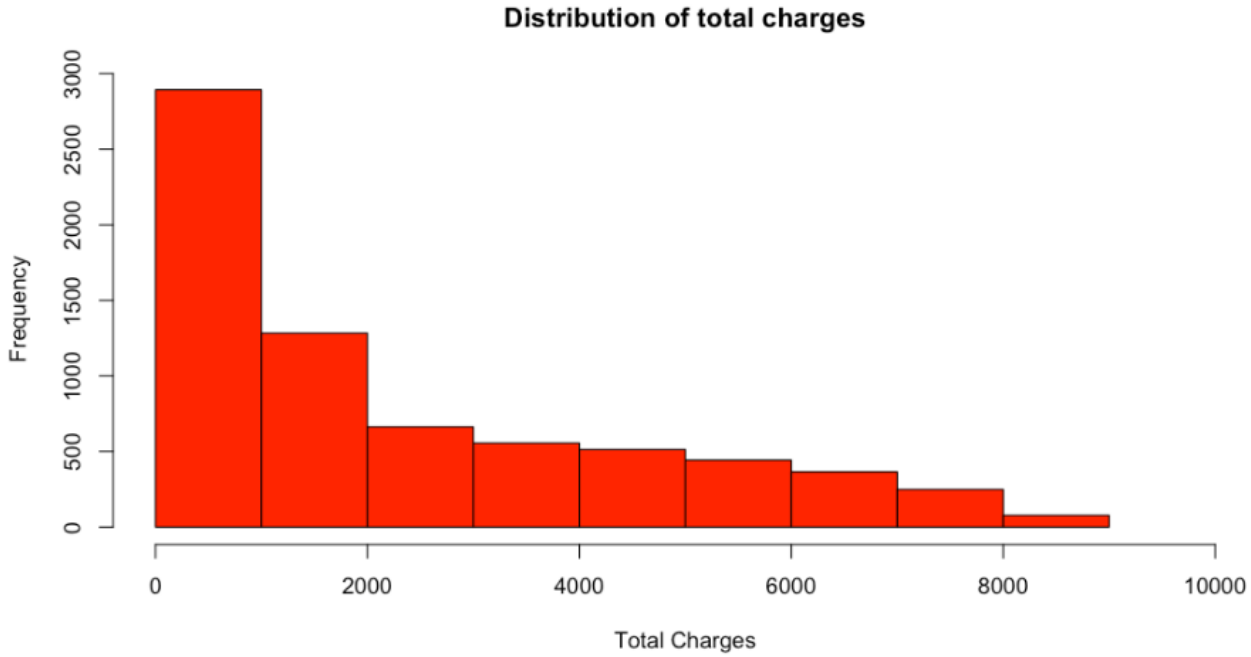

Boxplot of tenure
0
10
20
30
40
50
60
70
Tenure

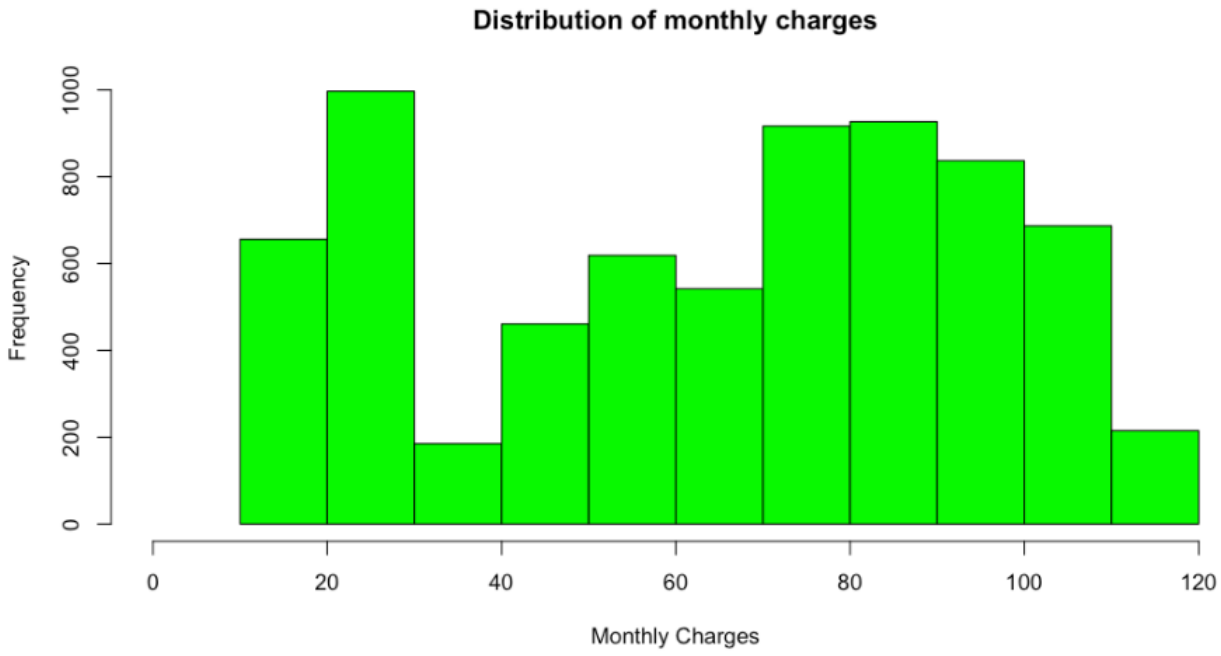

Boxplot of Monthly Charges
20
40
60
80
100
120
Monthly Charges

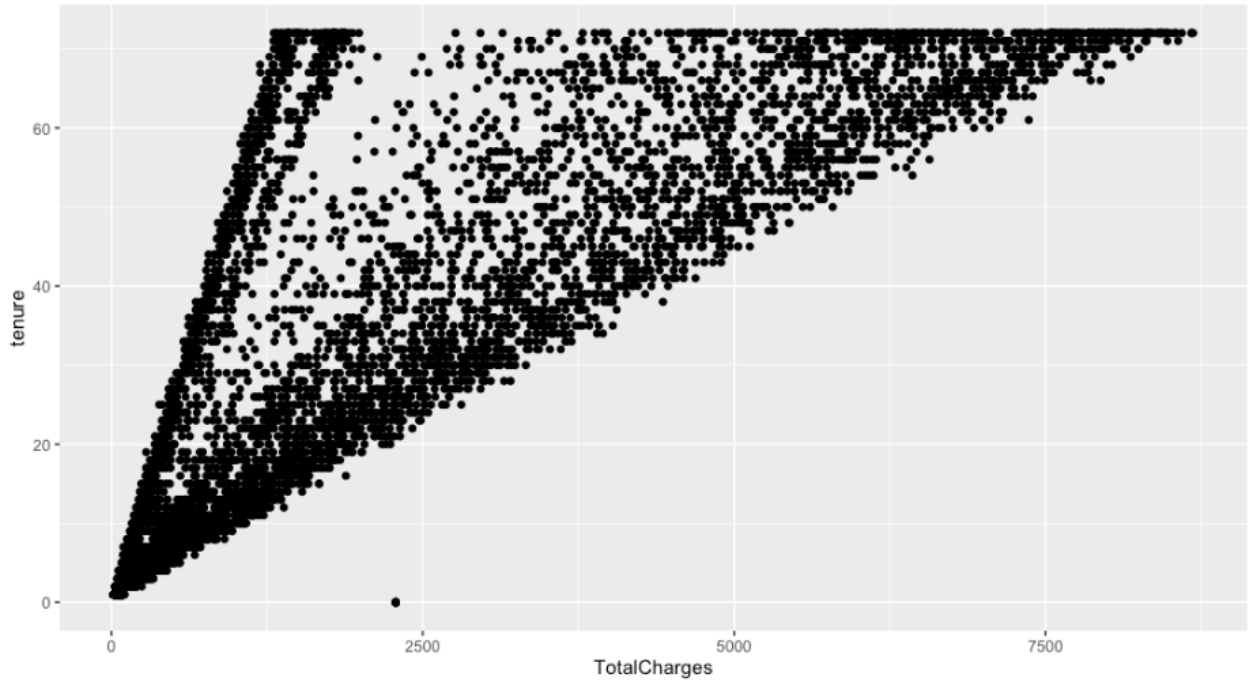



```
> hist(data$tenure,main="Distribution of tenure",xlab="Tenure",xlim=c(0,80),col="blue",breaks=8)

> hist(data$TotalCharges,main="Distribution of total charges",xlab="Total Charges",xlim=c(0,10000),col="red",breaks=9)

> hist(data$MonthlyCharges,main="Distribution of monthly charges",xlab="Monthly Charges",xlim=c(0,120),col="green",breaks=11)

> ggplot(data=data, mapping=aes(x=TotalCharges, y=tenure)) + geom_point()
```

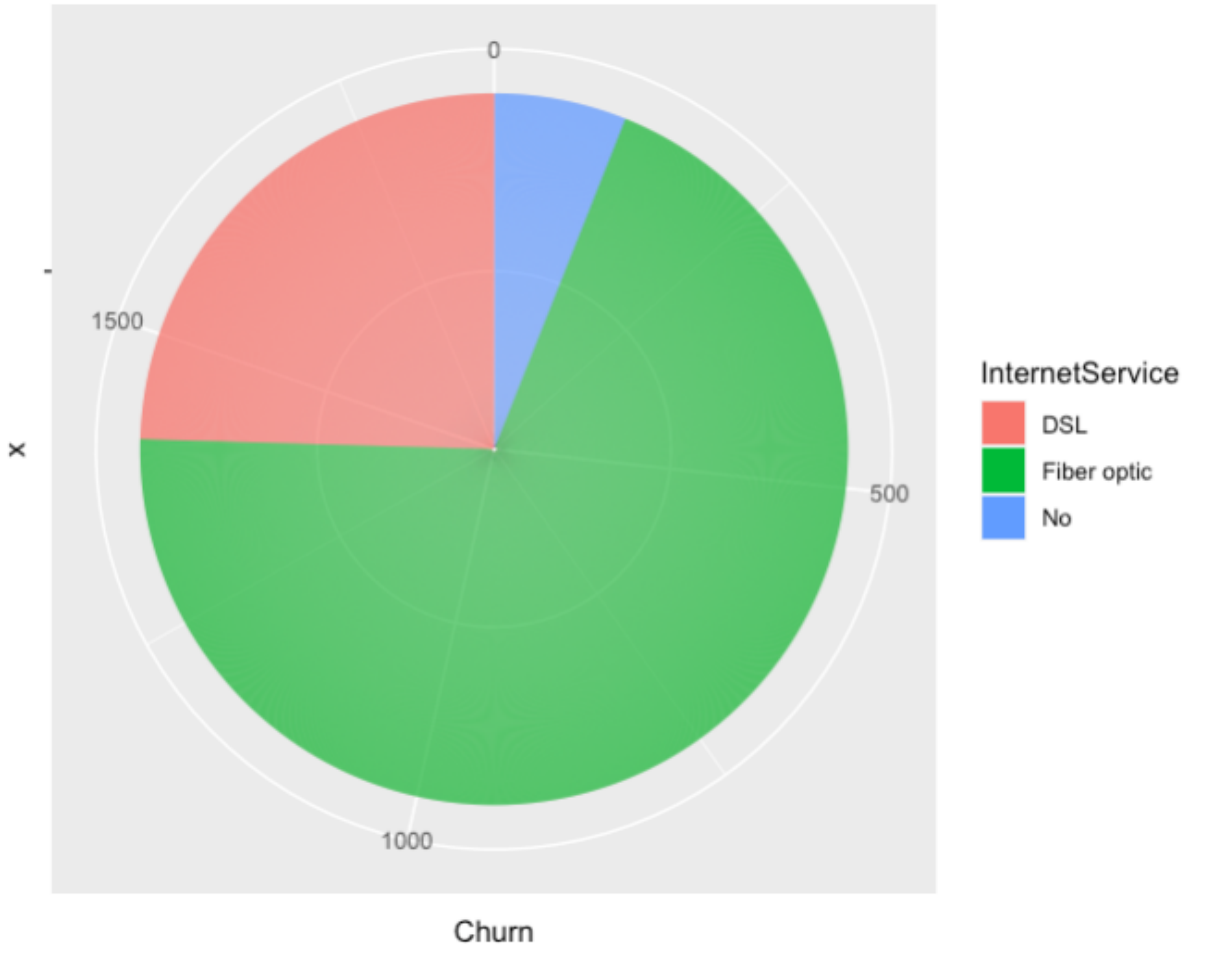


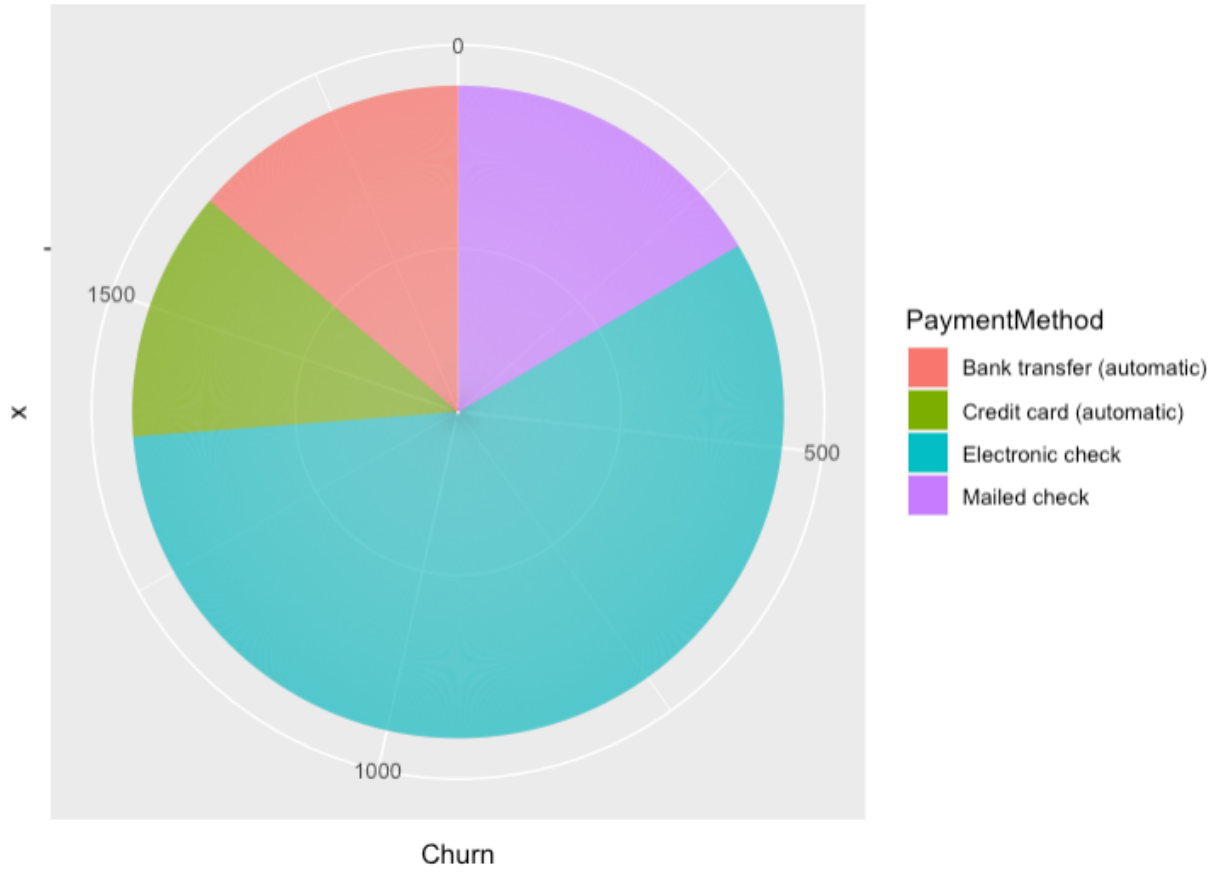

Distribution of total charges
Frequency
0
500
1000
1500
2000
2500
3000
0
2000
4000
6000
8000
10000
Total Charges

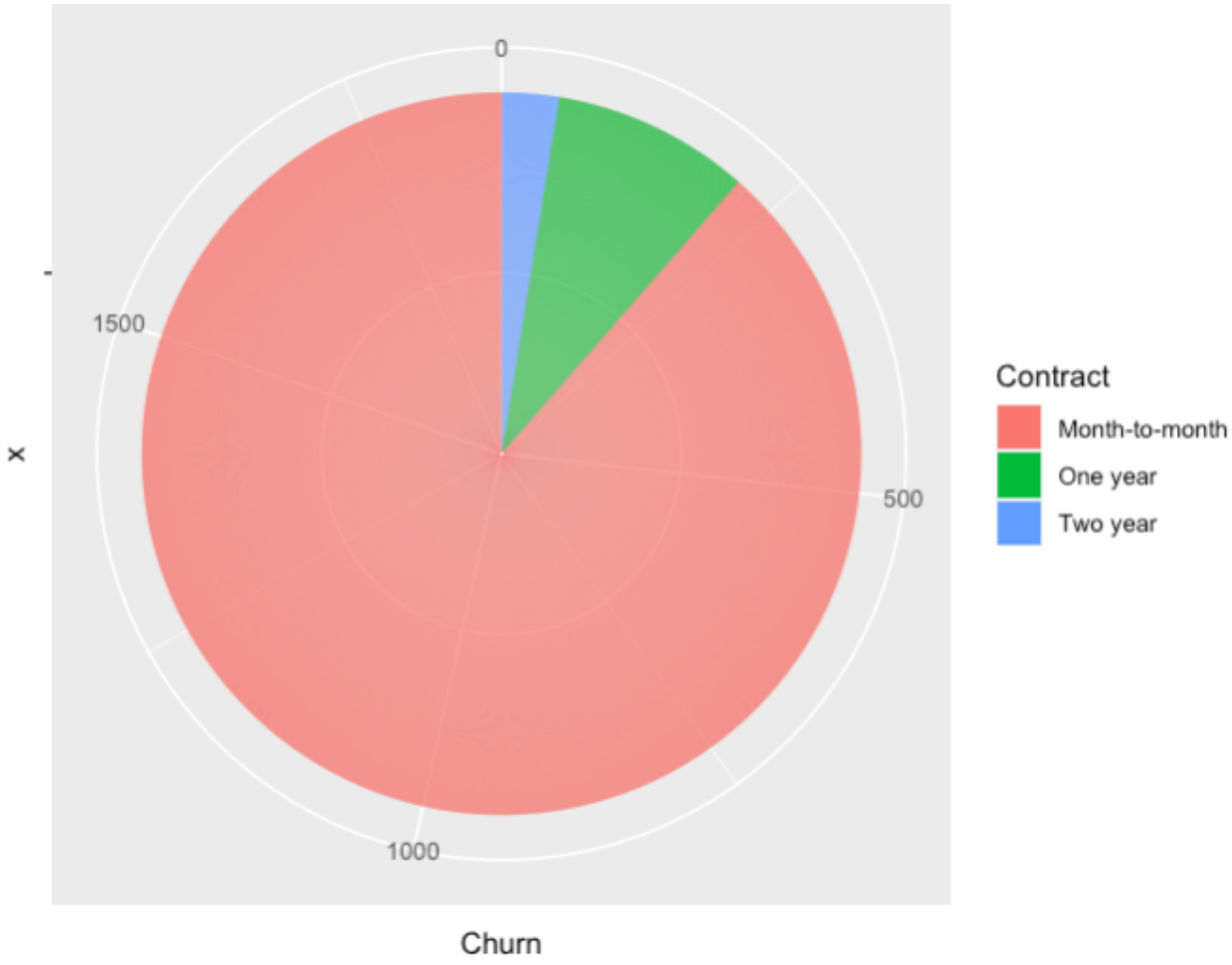

Distribution of monthly charges
Frequency
0
200
400
600
800
1000
0
20
40
60
80
100
120
Monthly Charges


tenure
0
20
40
60
0
2500
5000
7500
TotalCharges

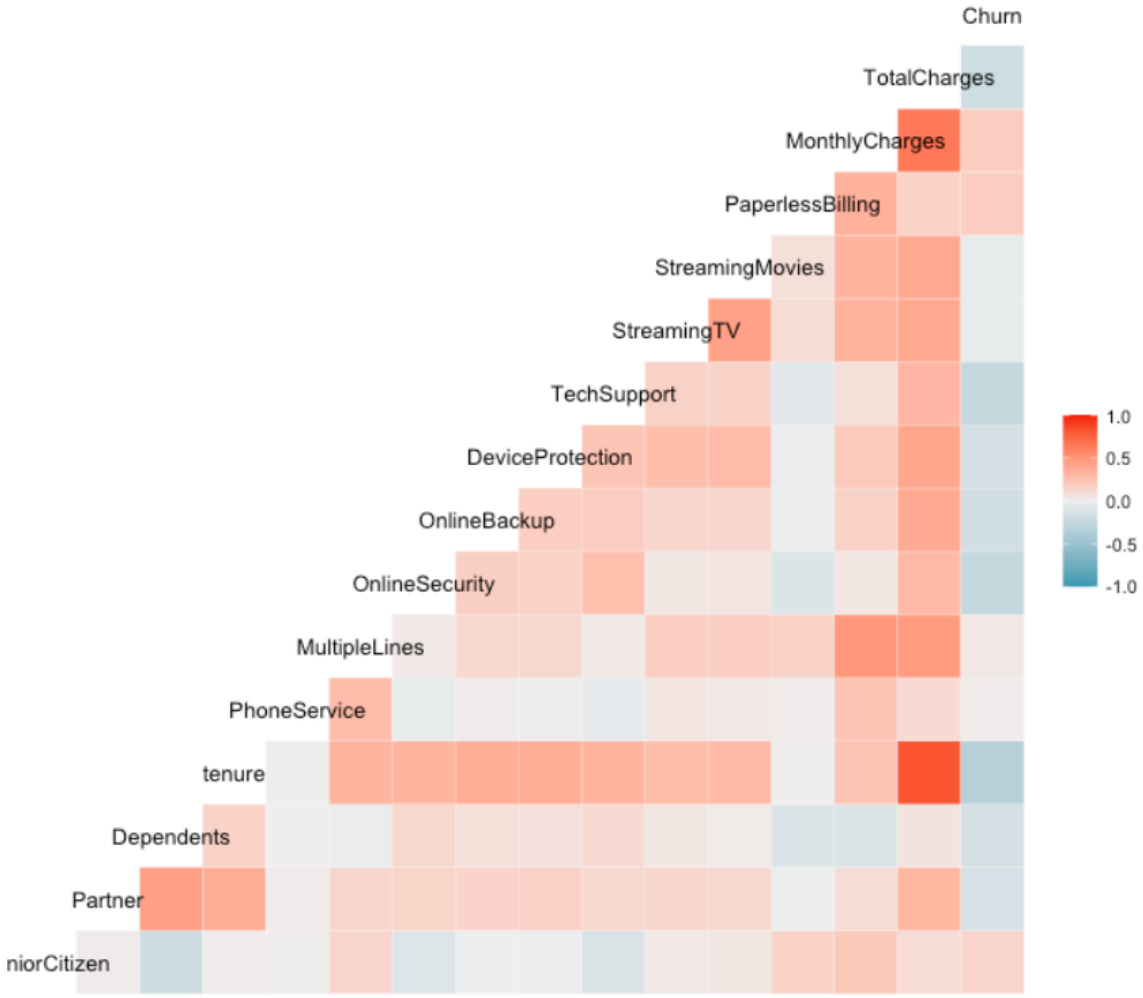
```
> ggplot(data, aes(x="", y=Churn, fill=InternetService))+geom_bar(stat = "identity")+coord_polar("y")
> ggplot(data, aes(x="", y=Churn, fill=PaymentMethod))+geom_bar(stat = "identity")+coord_polar("y")
> ggplot(data, aes(x="", y=Churn, fill=Contract))+geom_bar(stat = "identity")+coord_polar("y")
```




0
500
1000
1500
x
Churn
PaymentMethod
Bank transfer (automatic)
Credit card (automatic)
Electronic check
Mailed check




```
> #Correlation
> ggcorr(data)
```




```
> cor(data$tenure,data$TotalCharges)
[1] 0.8247573
> #Regression Analysis
>
> #logistic regression
>
> newdata <- select(data,SeniorCitizen,Partner,Dependents,tenure,PhoneService,MultipleLines,Online .... [TRUNCATED]

> View(newdata)

> install.packages("caTools")
trying URL 'https://cran.rstudio.com/bin/macosx/contrib/4.0/caTools_1.18.2.tgz'
Content type 'application/x-gzip' length 243209 bytes (237 KB)
==================================================
downloaded 237 KB


The downloaded binary packages are in
        /var/folders/lc/5v5k3fr16n335lxrv9pms7b00000gn/T//RtmpG5ji9S/downloaded_packages

> library(caTools)
```

	SeniorCitizen	Partner	Dependents	tenure	PhoneService	MultipleLines	OnlineSecurity	OnlineBackup	DeviceProtection	TechSupport
1	0	1	0	1	0	0	0.0000000	1.0000000	0.0000000	0.000000
2	0	0	0	34	1	0	1.0000000	0.0000000	1.0000000	0.000000
3	0	0	0	2	1	0	1.0000000	1.0000000	0.0000000	0.000000
4	0	0	0	45	0	0	1.0000000	0.0000000	1.0000000	1.000000
5	0	0	0	2	1	0	0.0000000	0.0000000	0.0000000	0.000000

StreamingTV	StreamingMovies	PaperlessBilling	MonthlyCharges	TotalCharges	Churn
0.0000000	0.0000000	1	29.85	29.85	0
0.0000000	0.0000000	0	56.95	1889.50	0
0.0000000	0.0000000	1	53.85	108.15	1
0.0000000	0.0000000	0	42.30	1840.75	0
0.0000000	0.0000000	1	70.70	151.65	1

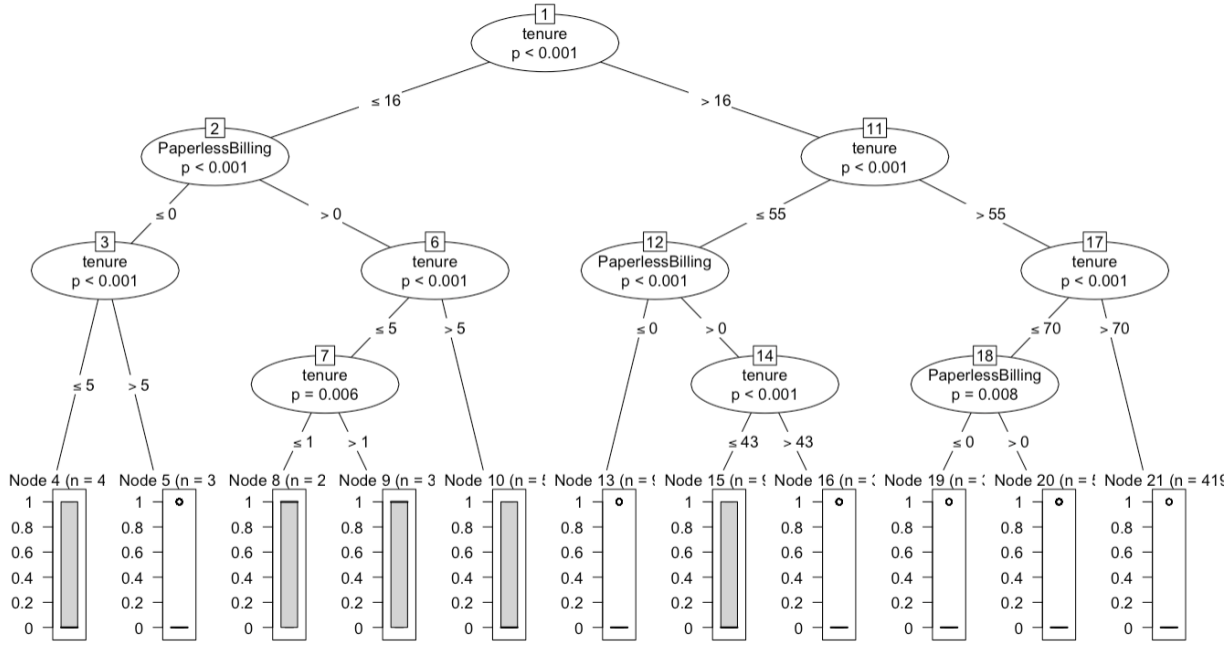
```
> split <- sample.split(data,SplitRatio=0.8)

> split
 [1] FALSE  TRUE  TRUE FALSE  TRUE  TRUE  TRUE  TRUE  TRUE FALSE FALSE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
[20]  TRUE

> training <- subset(data,split=="TRUE")

> testing <- subset(data,split=="FALSE")

> model <- glm(Churn~.,training,family="binomial")

> summary(model)

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

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.8937  -0.6816  -0.2828   0.7462   3.4497

Coefficients:
                                    Estimate Std. Error z value Pr(>|z|)
(Intercept)                         7.262e-01  9.171e-01   0.792 0.428465
genderMale                         -1.513e-02  7.222e-02  -0.210 0.834045
SeniorCitizen                       2.114e-01  9.414e-02   2.246 0.024703 *
Partner                             2.384e-02  8.672e-02   0.275 0.783359
Dependents                         -1.150e-01  1.008e-01  -1.142 0.253548
tenure                             -5.708e-02  6.829e-03  -8.358  < 2e-16 ***
PhoneService                       -9.383e-02  7.279e-01  -0.129 0.897440
MultipleLines                       3.412e-01  1.974e-01   1.729 0.083856 .
InternetServiceFiber optic          1.382e+00  8.948e-01   1.545 0.122338
InternetServiceNo                  -1.807e+00  1.537e+00  -1.175 0.239981
OnlineSecurity                     -3.374e-01  2.003e-01  -1.684 0.092104 .
OnlineBackup                       -8.958e-02  1.973e-01  -0.454 0.649795
DeviceProtection                    1.567e-01  1.972e-01   0.795 0.426871
TechSupport                        -1.821e-01  2.025e-01  -0.899 0.368569
StreamingTV                         4.003e-01  3.668e-01   1.091 0.275114
StreamingMovies                     4.868e-01  3.659e-01   1.330 0.183365
ContractOne year                   -6.125e-01  1.180e-01  -5.191 2.10e-07 ***
ContractTwo year                   -1.489e+00  2.005e-01  -7.422 1.15e-13 ***
PaperlessBilling                    3.581e-01  8.324e-02   4.302 1.70e-05 ***
PaymentMethodCredit card (automatic) -5.509e-02  1.274e-01  -0.432 0.665405
PaymentMethodElectronic check       3.344e-01  1.056e-01   3.168 0.001537 **
PaymentMethodMailed check          -3.516e-03  1.292e-01  -0.027 0.978290
MonthlyCharges                     -2.582e-02  3.564e-02  -0.724 0.468769
TotalCharges                        2.956e-04  7.774e-05   3.803 0.000143 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 6549.8  on 5633  degrees of freedom
Residual deviance: 4691.6  on 5610  degrees of freedom
AIC: 4739.6

Number of Fisher Scoring iterations: 6
```

```
> anova(model, test="Chisq")
Analysis of Deviance Table

Model: binomial, link: logit

Response: Churn

Terms added sequentially (first to last)
```

	Df	Deviance	Resid. Df	Resid. Dev	Pr(>Chi)	
NULL			5634	6530.2		
gender	1	0.04	5633	6530.2	0.8440764	
SeniorCitizen	1	136.44	5632	6393.8	< 2.2e-16	***
Partner	1	136.53	5631	6257.2	< 2.2e-16	***
Dependents	1	28.39	5630	6228.8	9.903e-08	***
tenure	1	679.67	5629	5549.2	< 2.2e-16	***
PhoneService	1	4.28	5628	5544.9	0.0386756	*
MultipleLines	1	146.65	5627	5398.2	< 2.2e-16	***
InternetService	2	524.03	5625	4874.2	< 2.2e-16	***
OnlineSecurity	1	25.79	5624	4848.4	3.803e-07	***
OnlineBackup	1	3.82	5623	4844.6	0.0505915	.
DeviceProtection	1	0.19	5622	4844.4	0.6633365	
TechSupport	1	28.46	5621	4815.9	9.560e-08	***
StreamingTV	1	22.23	5620	4793.7	2.415e-06	***
StreamingMovies	1	12.93	5619	4780.8	0.0003229	***
Contract	2	86.40	5617	4694.4	< 2.2e-16	***
PaperlessBilling	1	19.15	5616	4675.2	1.206e-05	***
PaymentMethod	3	22.66	5613	4652.6	4.762e-05	***
MonthlyCharges	1	1.63	5612	4650.9	0.2015916	
TotalCharges	1	18.90	5611	4632.0	1.380e-05	***

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

```
> testing$Churn <- as.character(testing$Churn)

> testing$Churn[testing$Churn=="No"] <- "0"

> testing$Churn[testing$Churn=="Yes"] <- "1"

> fitted.results <- predict(model,newdata=testing,type='response')

> fitted.results <- ifelse(fitted.results > 0.5,1,0)

> misClasificError <- mean(fitted.results != testing$Churn)

> print(paste('Logistic Regression Accuracy',1-misClasificError))
[1] "Logistic Regression Accuracy 0.803267045454545"

> print("Confusion Matrix for Logistic Regression")
[1] "Confusion Matrix for Logistic Regression"

> table(testing$Churn, fitted.results > 0.5)

    FALSE TRUE
  0   912  114
  1   163  219


> install.packages("party")
trying URL 'https://cran.rstudio.com/bin/macosx/contrib/4.0/party_1.3-7.tgz'
Content type 'application/x-gzip' length 961557 bytes (939 KB)
==================================================
downloaded 939 KB


The downloaded binary packages are in
        /var/folders/lc/5v5k3fr16n335lxrv9pms7b00000gn/T//Rtmpr60poK/downloaded_packages

> library(party)


> #Decision Tree

> tree <- ctree(Churn~tenure+PaperlessBilling, training)

> plot(tree)
```


1
tenure
p < 0.001
≤ 16
> 16
2
PaperlessBilling
p < 0.001
11
tenure
p < 0.001
≤ 0
> 0
≤ 55
> 55
3
tenure
p < 0.001
6
tenure
p < 0.001
12
PaperlessBilling
p < 0.001
17
tenure
p < 0.001
≤ 5
> 5
≤ 5
> 5
≤ 0
> 0
≤ 70
> 70
7
tenure
p = 0.006
14
tenure
p < 0.001
18
PaperlessBilling
p = 0.008
≤ 1
> 1
≤ 43
> 43
≤ 0
> 0
Node 4 (n = 4
Node 5 (n = 3
Node 8 (n = 2
Node 9 (n = 3
Node 10 (n = 5
Node 13 (n = 5
Node 15 (n = 5
Node 16 (n = 3
Node 19 (n = 3
Node 20 (n = 5
Node 21 (n = 419
1
0.8
0.6
0.4
0.2
0