Breifly Report

This IBM Sample Dataset has information about Telco customers and if they left the company within the last month (churn). Each row represents a unique costumer, while the columns contain information about customer's services, account and demographic data.

The intention is to predict customers with greater potential to leave the company.

1. Data

Code:

```
telco <- read.csv("~/Shiqi.RDATA/telco/Telco-Customer-Churn.csv")
head(telco)
summary(telco)
```

Output:

```
> summary(telco)
      customerID      gender     SeniorCitizen    Partner    Dependents     tenure      PhoneService          MultipleLines
 0002-ORFBO:   1   Female:3488   Min.   :0.0000   No :3641   No :4933   Min.   : 0.00   No : 682     No            :3390
 0003-MKNFE:   1   Male  :3555   1st Qu.:0.0000   Yes:3402   Yes:2110   1st Qu.: 9.00   Yes:6361     No phone service: 682
 0004-TLHLJ:   1                 Median :0.0000                         Median :29.00                Yes           :2971
 0011-IGKFF:   1                 Mean   :0.1621                         Mean   :32.37
 0013-EXCHZ:   1                 3rd Qu.:0.0000                         3rd Qu.:55.00
 0013-MHZWF:   1                 Max.   :1.0000                         Max.   :72.00
 (Other)   :7037
    InternetService            OnlineSecurity             OnlineBackup             DeviceProtection             TechSupport
 DSL        :2421   No                 :3498   No                 :3088   No                 :3095   No                 :3473
 Fiber optic:3096   No internet service:1526   No internet service:1526   No internet service:1526   No internet service:1526
 No         :1526   Yes                :2019   Yes                :2429   Yes                :2422   Yes                :2044


             StreamingTV            StreamingMovies          Contract    PaperlessBilling                 PaymentMethod
 No                 :2810   No                 :2785   Month-to-month:3875   No :2872         Bank transfer (automatic):1544
 No internet service:1526   No internet service:1526   One year      :1473   Yes:4171         Credit card (automatic)  :1522
 Yes                :2707   Yes                :2732   Two year      :1695                    Electronic check         :2365
                                                                             Mailed check             :1612


 MonthlyCharges    TotalCharges    Churn
 Min.   : 18.25   Min.   :  18.8   No :5174
 1st Qu.: 35.50   1st Qu.: 401.4   Yes:1869
 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
```

- 7043 observations with 21 variables.
- There are only 11 missing data in the TotalCharges field. And we can replace it with MonthlyCharges*tenure(The time customers have stayed).
- There are three continuous variables and they are Tenure, MonthlyCharges and TotalCharges. SeniorCitizen is in 'int' form, that can be changed to categorical.

Code:

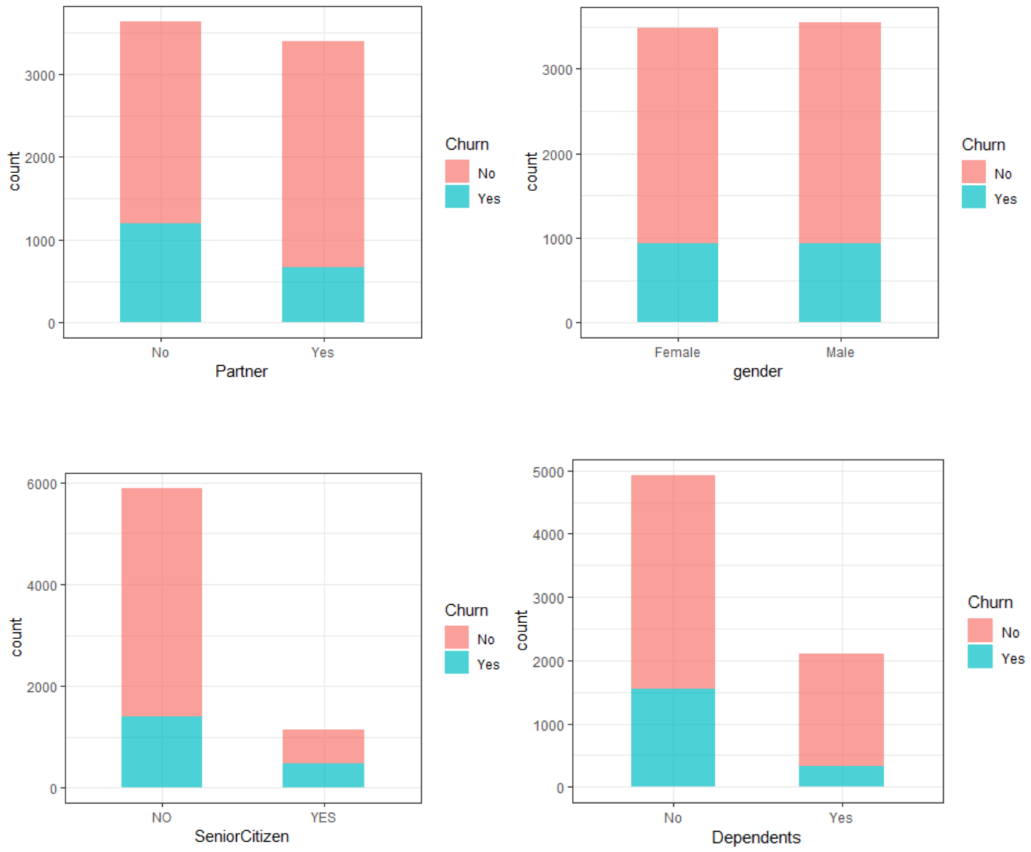
```
telco$TotalCharges <- ifelse(is.na(telco$TotalCharges==TRUE),
                             telco$MonthlyCharges*telco$tenure,
                             telco$TotalCharges)

telco$SeniorCitizen <- as.factor(ifelse(telco$SeniorCitizen==1, 'YES', 'NO'))
```

2. Exploratory Data Analysis (EDA)

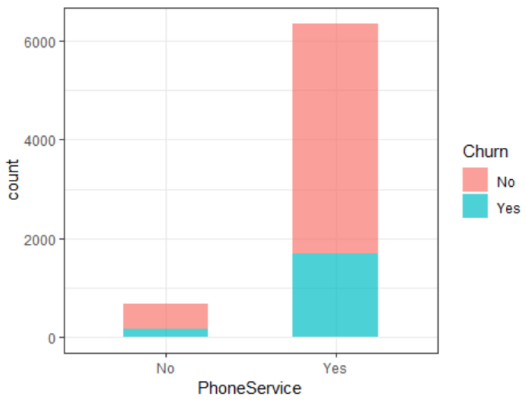
Code:

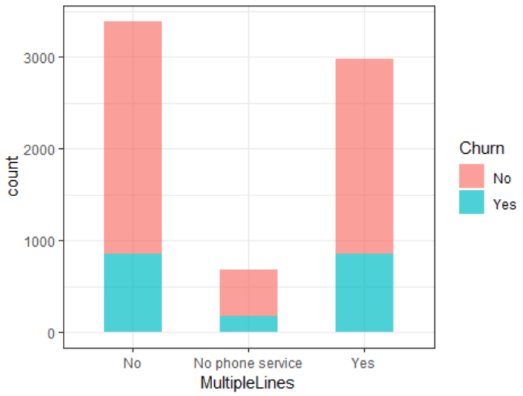
```
ggplot(telco)+geom_bar(aes(x=gender,fill=Churn),stat="count",alpha = 0.7,width=0.5)+theme_bw()
ggplot(telco)+geom_bar(aes(x=SeniorCitizen,fill=Churn),stat="count",alpha = 0.7,width=0.5)+theme_bw()
ggplot(telco)+geom_bar(aes(x=Partner,fill=Churn),stat="count",alpha = 0.7,width=0.5)+theme_bw()
ggplot(telco)+geom_bar(aes(x=Dependents,fill=Churn),stat="count",alpha = 0.7,width=0.5)+theme_bw()
ggplot(telco)+geom_bar(aes(x=PhoneService,fill=Churn),stat="count",alpha = 0.7,width=0.5)+theme_bw()
ggplot(telco)+geom_bar(aes(x=MultipleLines,fill=Churn),stat="count",alpha = 0.7,width=0.5)+theme_bw()
ggplot(telco)+geom_bar(aes(x=InternetService,fill=Churn),stat="count",alpha = 0.7,width=0.5)+theme_bw()
ggplot(telco)+geom_bar(aes(x=OnlineSecurity,fill=Churn),stat="count",alpha = 0.7,width=0.5)+theme_bw()
ggplot(telco)+geom_bar(aes(x=OnlineBackup,fill=Churn),stat="count",alpha = 0.7,width=0.5)+theme_bw()
ggplot(telco)+geom_bar(aes(x=DeviceProtection,fill=Churn),stat="count",alpha = 0.7,width=0.5)+theme_bw()
ggplot(telco)+geom_bar(aes(x=TechSupport,fill=Churn),stat="count",alpha = 0.7,width=0.5)+theme_bw()
ggplot(telco)+geom_bar(aes(x=StreamingTV,fill=Churn),stat="count",alpha = 0.7,width=0.5)+theme_bw()
ggplot(telco)+geom_bar(aes(x=StreamingMovies,fill=Churn),stat="count",alpha = 0.7,width=0.5)+theme_bw()
ggplot(telco)+geom_bar(aes(x=Contract,fill=Churn),stat="count",alpha = 0.7,width=0.5)+theme_bw()
ggplot(telco)+geom_bar(aes(x=PaperlessBilling,fill=Churn),stat="count",alpha = 0.7,width=0.5)+theme_bw()
ggplot(telco)+geom_bar(aes(x=PaymentMethod,fill=Churn),stat="count",alpha = 0.7,width=0.5)+theme_bw()
```

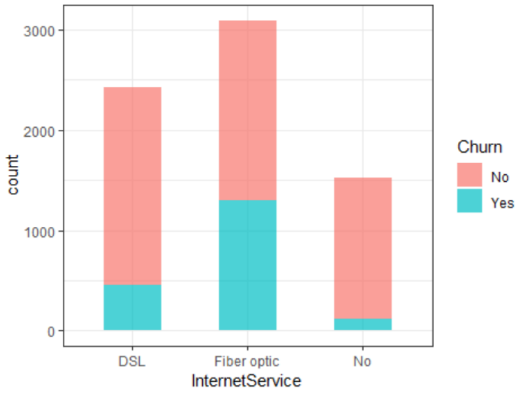
Output:

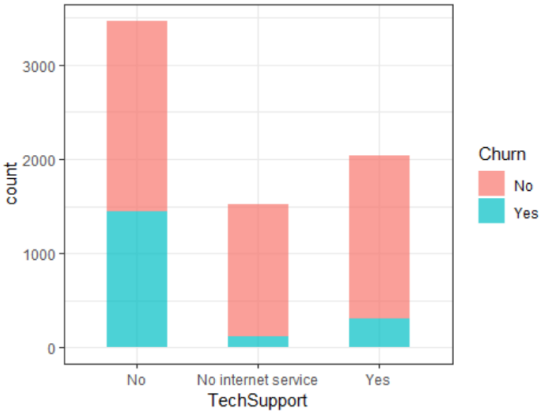


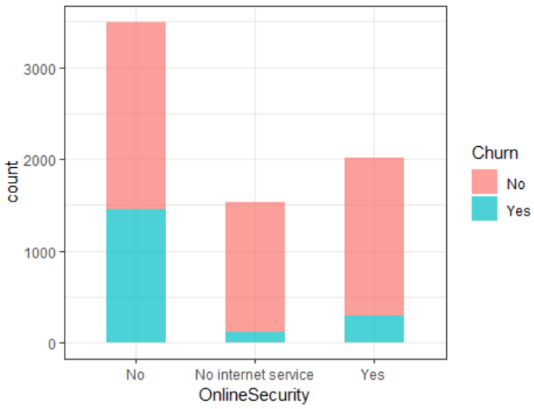
- The churn percent is almost equal in case of Male and Females
- The percent of churn is higher in case of senior citizens
- Customers with Partners and Dependents have lower churn rate as compared to those who don't have partners & Dependents.

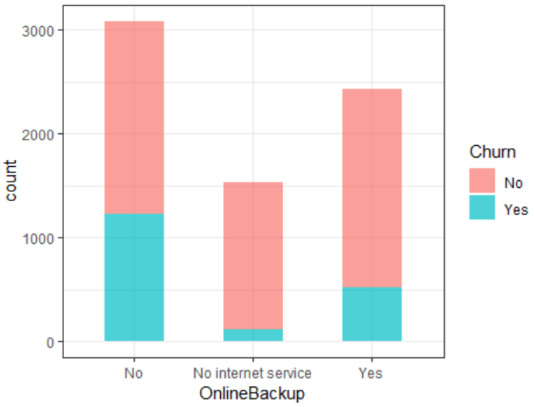




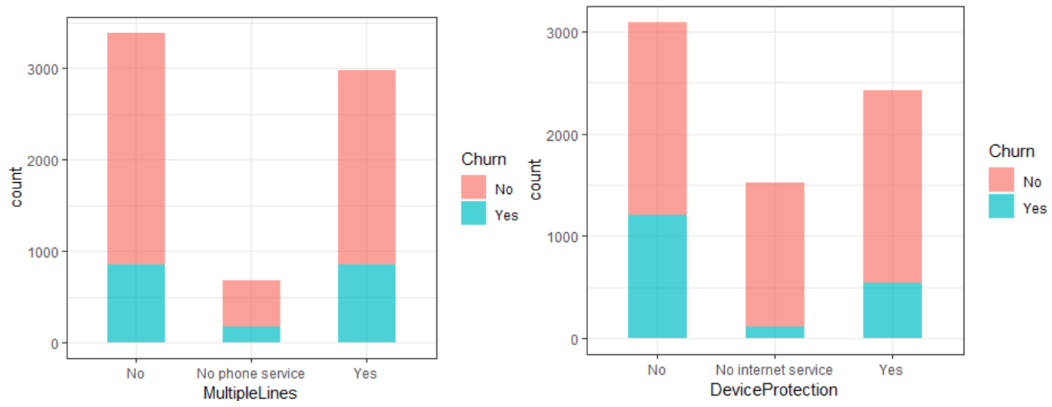


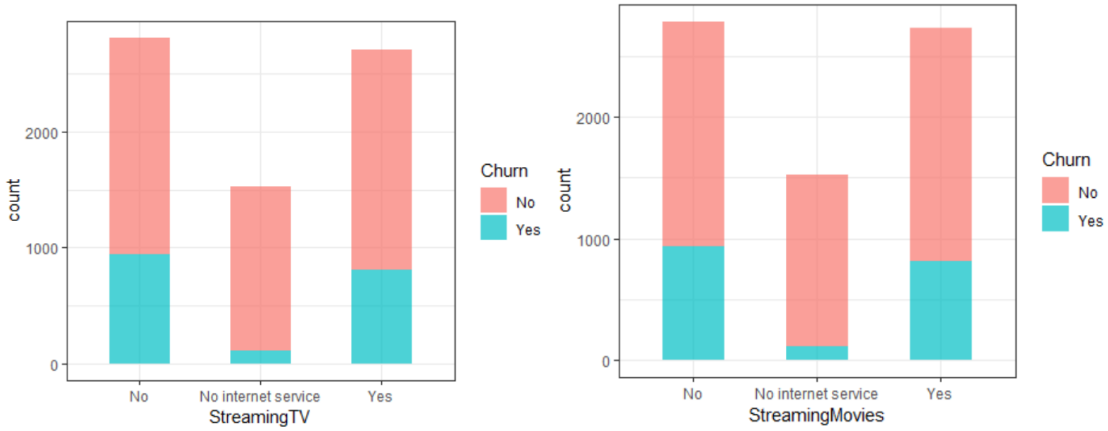


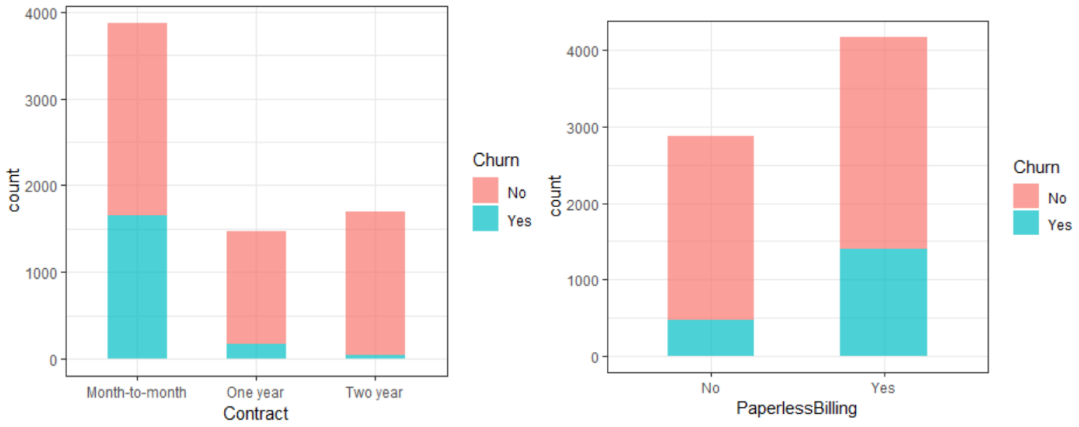


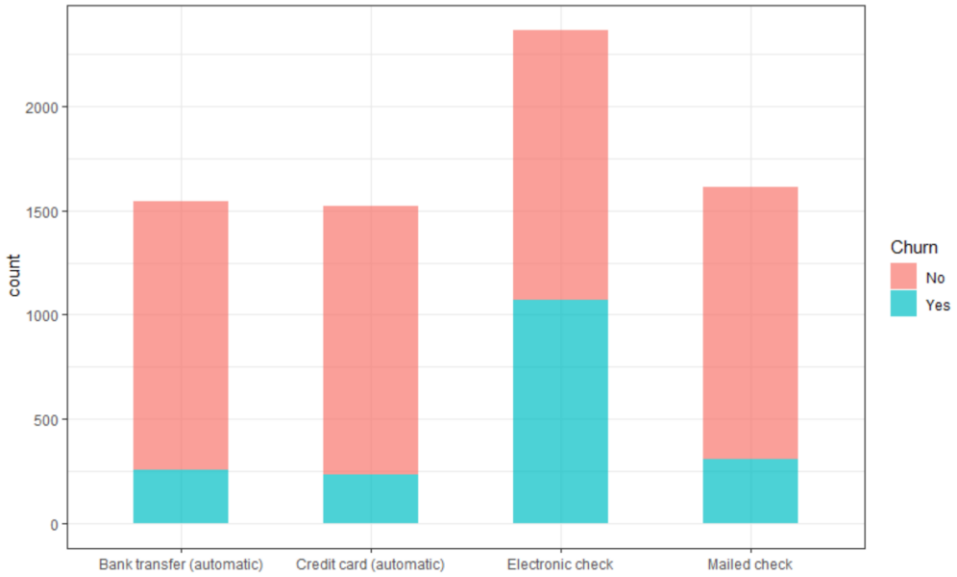


- Churn rate is much higher in case of Fiber Optic InternetServices.
- Customers who do not have services like No OnlineSecurity , OnlineBackup andTechSupport have left the platform in the past month.


count
3000
2000
1000
0
No
No phone service
Yes
MultipleLines
Churn
No
Yes
count
3000
2000
1000
0
No
No internet service
Yes
DeviceProtection
Churn
No
Yes


count
2000
1000
0
No
No internet service
Yes
StreamingTV
Churn
No
Yes
count
2000
1000
0
No
No internet service
Yes
StreamingMovies
Churn
No
Yes


count
4000
3000
2000
1000
0
Month-to-month
One year
Two year
Contract
Churn
No
Yes
count
4000
3000
2000
1000
0
No
Yes
PaperlessBilling
Churn
No
Yes

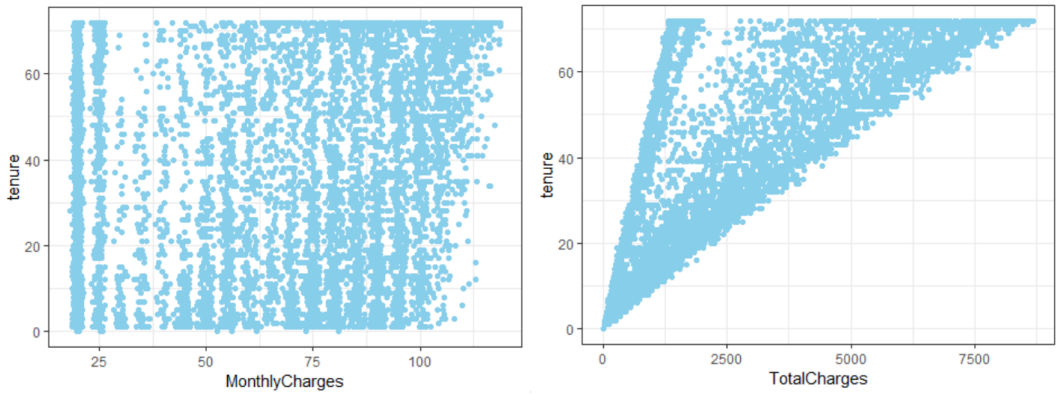


- A larger percent of Customers with monthly subscription have left when compared to Customers with one or two year contract.
- Churn percent is higher in case of cutsomers having paperless billing option.
- Customers who have ElectronicCheck PaymentMethod tend to leave the platform more when compared to other options.

Code:

```
ggplot(telco)+geom_point(aes(x=MonthlyCharges,y=tenure), colour="skyblue")+theme_bw()
ggplot(telco)+geom_point(aes(x=TotalCharges,y=tenure),colour="skyblue")+theme_bw()
```

Output:



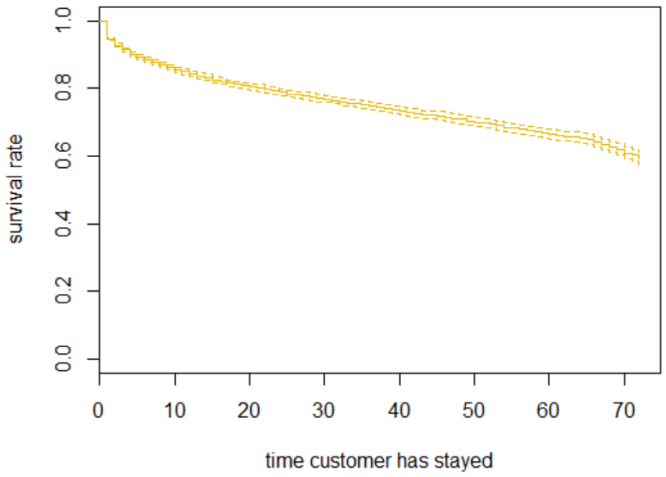
- We observe that there is no relation to the monthly payments. What I think about is that many clients remain for a long time without hiring new services; in contrast, some already come with more expensive plans.

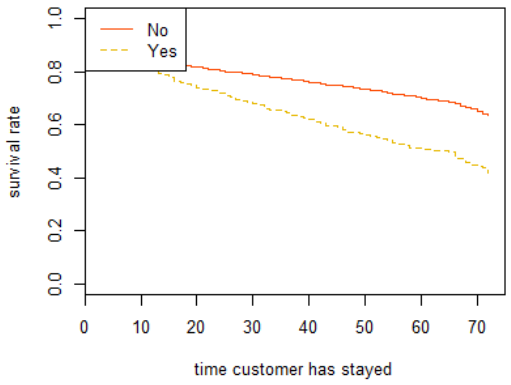
3. Survival Curve

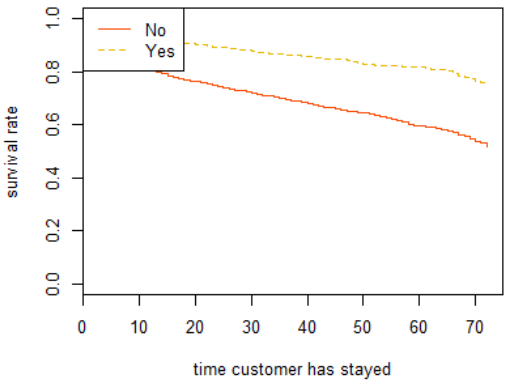
Code:

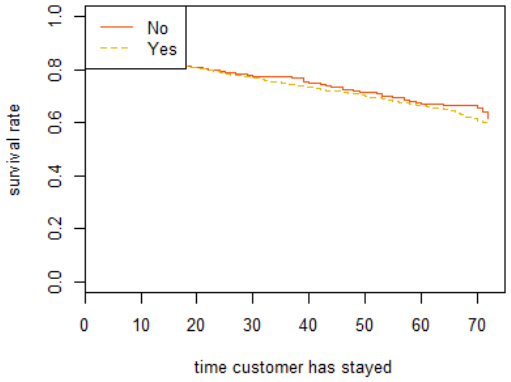
```
telco$Churn <- ifelse(telco$Churn=='Yes',1,0)

plot(survfit(Surv(telco$tenure,telco$Churn)~telco$SeniorCitizen),
     col=c("#FC4E07", "#E7B800"),lty=c(1,2),
     xlab="time customer has stayed", ylab="survival rate")
legend("topleft",c("No","Yes"),col=c("#FC4E07", "#E7B800"),lty=c(1,2))
plot(survfit(Surv(telco$tenure,telco$Churn)~telco$Dependents),
     col=c("#FC4E07", "#E7B800"),lty=c(1,2),
     xlab="time customer has stayed",ylab="survival rate")
legend("topleft",c("No","Yes"),col=c("#FC4E07", "#E7B800"),lty=c(1,2))
plot(survfit(Surv(telco$tenure,telco$Churn)~telco$PhoneService),
     col=c("#FC4E07", "#E7B800"),lty=c(1,2),
     xlab="time customer has stayed",ylab="survival rate")
legend("topleft",c("No","Yes"),col=c("#FC4E07", "#E7B800"),lty=c(1,2))
plot(survfit(Surv(telco$tenure,telco$Churn)~telco$InternetService),
     col=c("#FC4E07", "#E7B800","grey"),lty=c(1,2),
     xlab="time customer has stayed", ylab="survival rate")
legend("topleft",c("DSL","Fiber optic","No" ),
         col=c("#FC4E07", "#E7B800","grey"),lty=c(1,2))
```

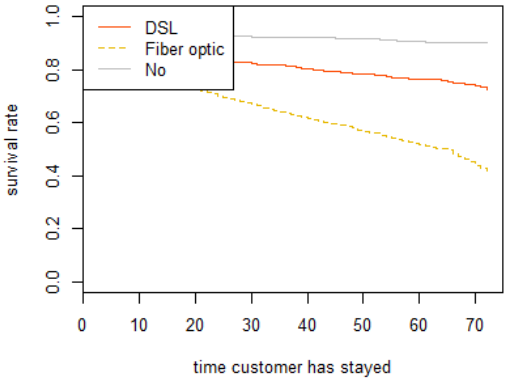
Output:







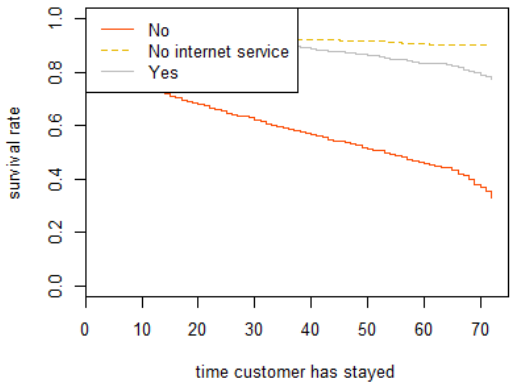


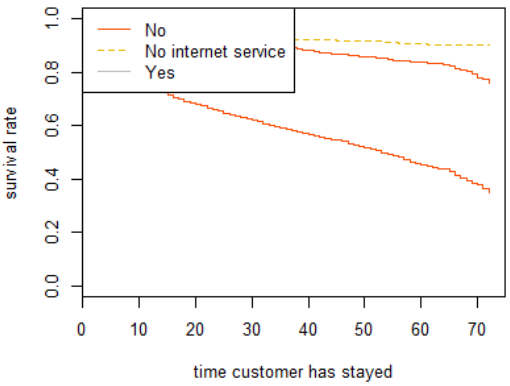


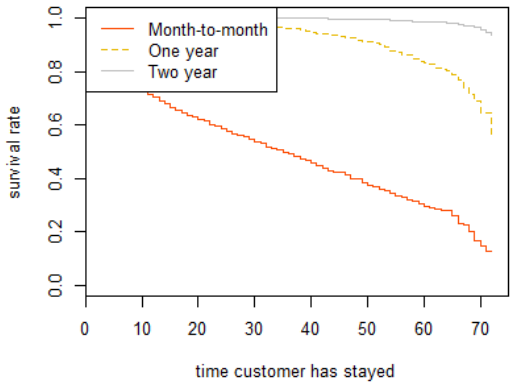
Code:

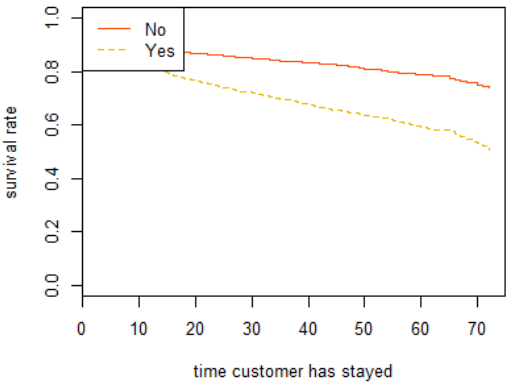
```
plot(survfit(Surv(telco$tenure,telco$Churn)~telco$OnlineSecurity),
     col=c("#FC4E07", "#E7B800","grey"),lty=c(1,2),
     xlab="time customer has stayed",ylab="survival rate")
legend("topleft",c("No","No internet service","Yes"    ),
       col=c("#FC4E07", "#E7B800","grey"),lty=c(1,2))
plot(survfit(Surv(telco$tenure,telco$Churn)~telco$TechSupport),
     col=c("#FC4E07", "#E7B800","#FC4E07"),lty=c(1,2),
     xlab="time customer has stayed",ylab="survival rate")
legend("topleft",c("No","No internet service","Yes"    ),
       col=c("#FC4E07", "#E7B800","grey"),lty=c(1,2))
plot(survfit(Surv(telco$tenure,telco$Churn)~telco$Contract),
     col=c("#FC4E07", "#E7B800","grey"),lty=c(1,2),
     xlab="time customer has stayed",ylab="survival rate")
legend("topleft",c("Month-to-month" ,"One year","Two year"     ),
       col=c("#FC4E07", "#E7B800","grey"),lty=c(1,2))

plot(survfit(Surv(telco$tenure,telco$Churn)~telco$PaymentMethod),
     col=c("#FC4E07", "#E7B800","grey", "skyblue"),lty=c(1,2),
     xlab="time customer has stayed",ylab="survival rate")
legend("topleft",c("Bank transfer (automatic)" ,"Credit card (automatic)",
                   "Electronic check","Mailed check"),
       col=c("#FC4E07", "#E7B800","grey", "#skyblue"),lty=c(1,2))
```

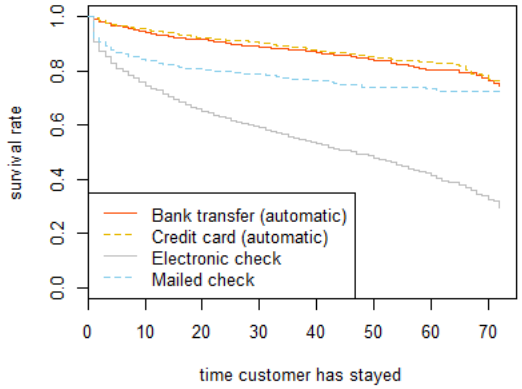
Output:











- We can see the output of the survival curve is pretty much like the conclusion we get from the descriptive analytical (EDA).
- The volume of older people leaving the company is much higher than the volume of non-elderly
- There is no behavior difference between women and men but there is great difference between people with partners e without it

- There is a huge churn tendency in Fiber Optic Services. That might show a great insatisfcation with this service.
- The company should explore some ways to making customers use internet services since it makes the customers stay longer.

4. Supervised Machine Leaning

4.1 Data Transformation

Code:

```
#replce "No internet service/No phone service" with "No"
telco_sub <- telco[,c(-1,-6, -19,-20,-21)]
telco_sub <- data.frame(lapply(telco_sub, function(x) {
  gsub("No internet service", "No", x)}))
telco_sub <- data.frame(lapply(telco_sub, function(x) {
  gsub("No phone service", "No", x)}))
telco_final<-cbind(telco[,c(1,6, 19,20,21)],telco_sub)
```

4.2 Creating Training and Test Sets

Code:

```
n <- nrow(telco_final)
ntrain <- round(n*0.7)
set.seed(116)
tindex <- sample(n,ntrain)

train <- telco_final[tindex,]
test <- telco_final[-tindex,]
```

4.3 Random Forest

Code:

```
#use random forest to predice if the customers will stay or
rf <- randomForest(telco_sub,y=telco$Churn, ntree=500, mtry=2, importance=TRUE)
#use trained model rf, predict test set response values
prediction <- predict(rf, newdata=test, type="class")
table(prediction, test$Churn)
misclassification_error_rate <- sum(test$Churn != prediction) / nrow(test)*100
importance(rf)
```

Output:

```
> table(prediction, test$Churn)

prediction    0    1
         0 1455  261
         1  110  287
> misclassification_error_rate
[1] 17.65263
```

- The error rate is still high and we should keep improving our model

4.4 Cox Proportion Risk Model

Code:

```
telco_final$Churn <- ifelse(telco_final$Churn==0,FALSE,TRUE)
fit=coxph(Surv(tenure,Churn)~SeniorCitizen+Dependents+
            PhoneService+InternetService+OnlineSecurity+
            TechSupport+Contract+PaperlessBilling+PaymentMethod+
            TotalCharges+MonthlyCharges,data=telco_final)
summary(fit)
```

Output:

```
> summary(fit)
Call:
coxph(formula = Surv(tenure, Churn) ~ SeniorCitizen + Dependents +
    PhoneService + InternetService + OnlineSecurity + TechSupport +
    Contract + PaperlessBilling + PaymentMethod + TotalCharges +
    MonthlyCharges, data = telco_final)

  n= 7043, number of events= 1869

                                        coef  exp(coef)   se(coef)       z Pr(>|z|)
SeniorCitizenYES                    1.261e-02  1.013e+00  5.616e-02   0.225  0.82235
DependentsYes                      -1.834e-01  8.324e-01  6.399e-02  -2.867  0.00415 **
PhoneServiceYes                     3.551e-01  1.426e+00  1.133e-01   3.136  0.00171 **
InternetServiceFiber optic          4.427e-01  1.557e+00  1.058e-01   4.184 2.87e-05 ***
InternetServiceNo                  -1.720e+00  1.791e-01  1.714e-01 -10.037  < 2e-16 ***
OnlineSecurityYes                  -3.365e-01  7.142e-01  6.802e-02  -4.948 7.50e-07 ***
TechSupportYes                     -1.961e-01  8.219e-01  6.792e-02  -2.887  0.00388 **
ContractOne year                   -1.268e+00  2.813e-01  1.007e-01 -12.594  < 2e-16 ***
ContractTwo year                   -3.713e+00  2.441e-02  2.022e-01 -18.365  < 2e-16 ***
PaperlessBillingYes                 1.544e-01  1.167e+00  5.644e-02   2.735  0.00623 **
PaymentMethodCredit card (automatic) -9.818e-03  9.902e-01  9.069e-02  -0.108  0.91379
PaymentMethodElectronic check       3.937e-01  1.482e+00  7.270e-02   5.414 6.15e-08 ***
PaymentMethodMailed check           5.232e-01  1.687e+00  8.674e-02   6.031 1.63e-09 ***
TotalCharges                       -1.624e-03  9.984e-01  4.041e-05 -40.185  < 2e-16 ***
MonthlyCharges                      3.688e-02  1.038e+00  2.734e-03  13.490  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1


                                    exp(coef) exp(-coef) lower .95 upper .95
SeniorCitizenYES                      1.01269     0.9875   0.90713   1.13053
DependentsYes                         0.83240     1.2013   0.73428   0.94363
PhoneServiceYes                       1.42638     0.7011   1.14244   1.78089
InternetServiceFiber optic            1.55685     0.6423   1.26528   1.91560
InternetServiceNo                     0.17908     5.5840   0.12799   0.25056
OnlineSecurityYes                     0.71424     1.4001   0.62510   0.81609
TechSupportYes                        0.82192     1.2167   0.71948   0.93895
ContractOne year                      0.28133     3.5546   0.23094   0.34271
ContractTwo year                      0.02441    40.9726   0.01642   0.03627
PaperlessBillingYes                   1.16694     0.8569   1.04473   1.30345
PaymentMethodCredit card (automatic)  0.99023     1.0099   0.82897   1.18286
PaymentMethodElectronic check         1.48238     0.6746   1.28550   1.70941
PaymentMethodMailed check             1.68740     0.5926   1.42358   2.00011
TotalCharges                          0.99838     1.0016   0.99830   0.99846
MonthlyCharges                        1.03757     0.9638   1.03202   1.04314

Concordance= 0.928  (se = 0.002 )
Rsquare= 0.571   (max possible= 0.988 )
Likelihood ratio test= 5961  on 15 df,   p=<2e-16
Wald test            = 2572  on 15 df,   p=<2e-16
Score (logrank) test = 4619  on 15 df,   p=<2e-16
```

- The coefficients represent that when the other independent variables remain unchanged, the value of X_i increases by one unit, the risk becomes exp (β_1) times as much as the original value.
- Overall, the model has a higher significance. However, there are some inconsistencies with the EDA stage, which need further study.