Survival Analysis for Telecommunication Services Subscribers

University of Miami | Mingjun Li

2/22/2020

A telecommunications company is concerned about the number of customers leaving their landline business for cable competitors. Our goal is to help them understand who is leaving and ultimately propose a retention plan to decrease churn and improve revenues.

```
library(tidyr)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(pscl)
```

```
## Warning: package 'pscl' was built under R version 3.6.2
```

```
## Classes and Methods for R developed in the
## Political Science Computational Laboratory
## Department of Political Science
## Stanford University
## Simon Jackman
## hurdle and zeroinfl functions by Achim Zeileis
```

```
library(survival)

#install.packages('survminer')
library(survminer)
```

```
## Warning: package 'survminer' was built under R version 3.6.2
```

```
## Loading required package: ggplot2
```

```
## Loading required package: ggpubr
```

```
## Warning: package 'ggpubr' was built under R version 3.6.2
```

```
## Loading required package: magrittr
```

```
##
## Attaching package: 'magrittr'
```

```
## The following object is masked from 'package:tidyr':
##
##     extract
```

```
#install.packages('rlist')
library(rlist)
```

```
## Warning: package 'rlist' was built under R version 3.6.2
```

```
library(ggpubr)
library(magrittr)
```

```
churn <- read.csv('https://s3.amazonaws.com/douglas2/MAS646/telcoChurn.csv')%>%
  mutate(
    SeniorCitizen = factor(SeniorCitizen,levels=c(1,0)),
    Churn = as.integer((Churn=='Yes')*1)
  )
```

```
names(churn)
```

```
##  [1] "customerID"       "gender"           "SeniorCitizen"
##  [4] "Partner"          "Dependents"       "tenure"
##  [7] "PhoneService"     "MultipleLines"    "InternetService"
## [10] "OnlineSecurity"   "OnlineBackup"     "DeviceProtection"
## [13] "TechSupport"      "StreamingTV"      "StreamingMovies"
## [16] "Contract"         "PaperlessBilling" "PaymentMethod"
## [19] "MonthlyCharges"   "TotalCharges"     "Churn"
```

Which kind of customer is most easy to leave?

##(use customer-based info: gender, age range, and if they have partners and dependents)

##Draw KM curve by each group

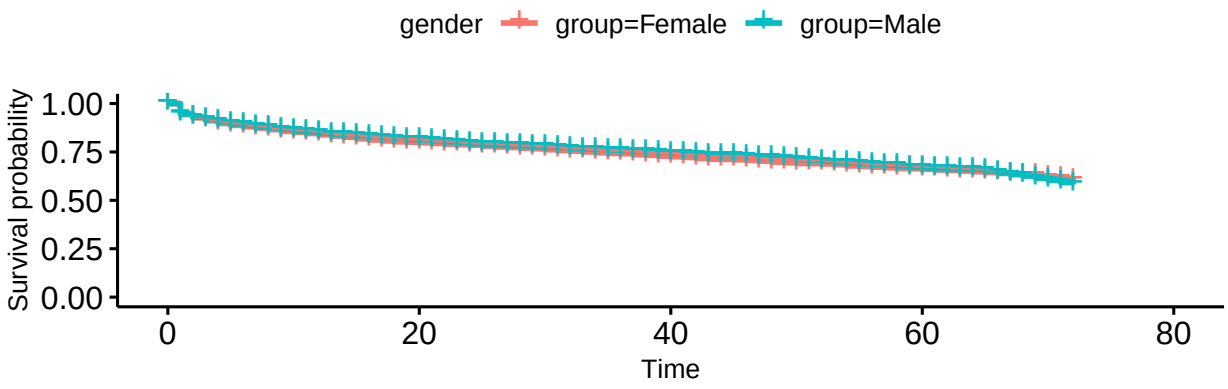
```
#Survminer includes a function surv_fit that acts as a wrapper around survfit. If you use surv_fit inst
KMcurve <- function(d,nameVector,time,churn){
  len=length(nameVector)
  sub=d[nameVector]
  # test if there is difference between groups
  dif = lapply(sub, function(q) survdiff(Surv(tenure,Churn)~q,data=d))
  s<-mapply(
            function(group,colname) {
              ggsurvplot(
                         surv_fit(Surv(d$tenure,d$Churn)~group,data=d),
                          data=d,legend.title = colname,font.x=10,font.y=10
                               )},
            group = sub, colname = names(sub)
              )
  plot_list=list()
  for (i in 1:len){
    plot_list = list.append(plot_list,s[,i]$plot)
  }

  return (list(ggarrange(plotlist=plot_list,ncol=1,nrow=2),dif)) #R function must return a list if mult
}
```

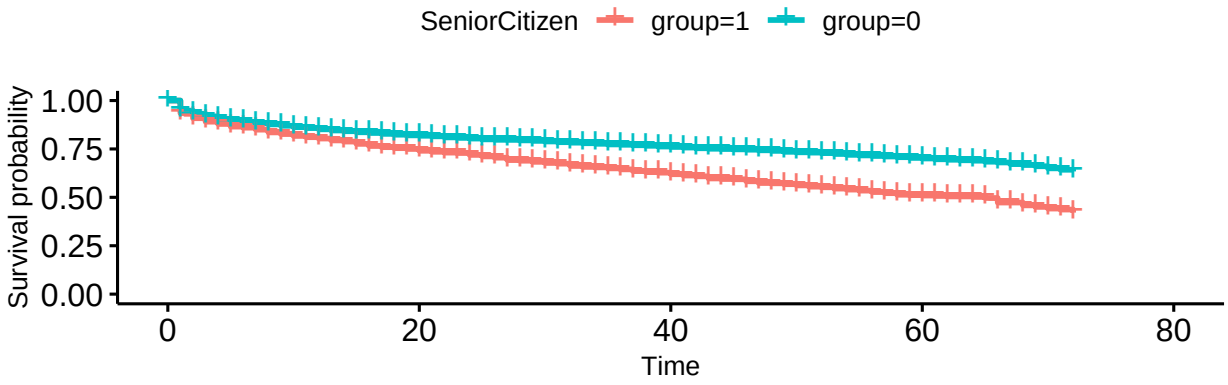
##Gender doesnt make any difference

##Senior Citizen, No partner and No dependents are more likely to leave

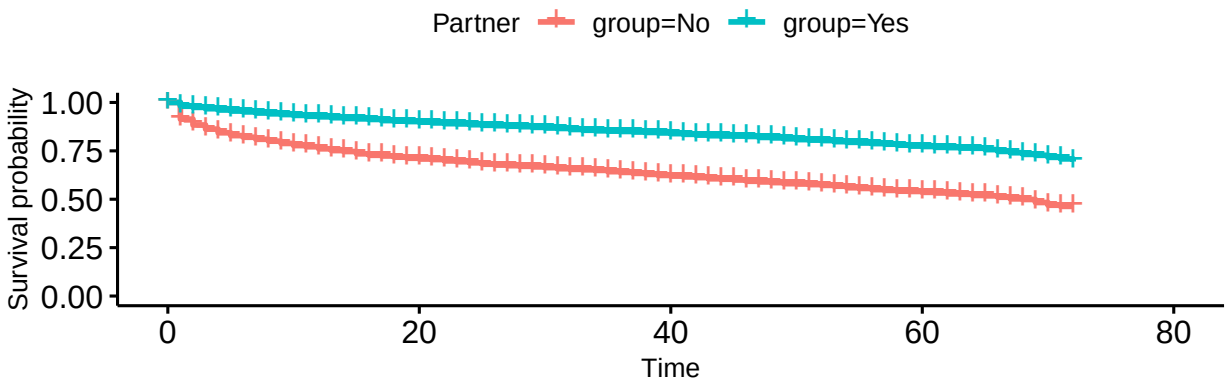
```
cInfo = names(churn)[2:5]
KMcurve(churn,cInfo,'tenure','Churn')
```

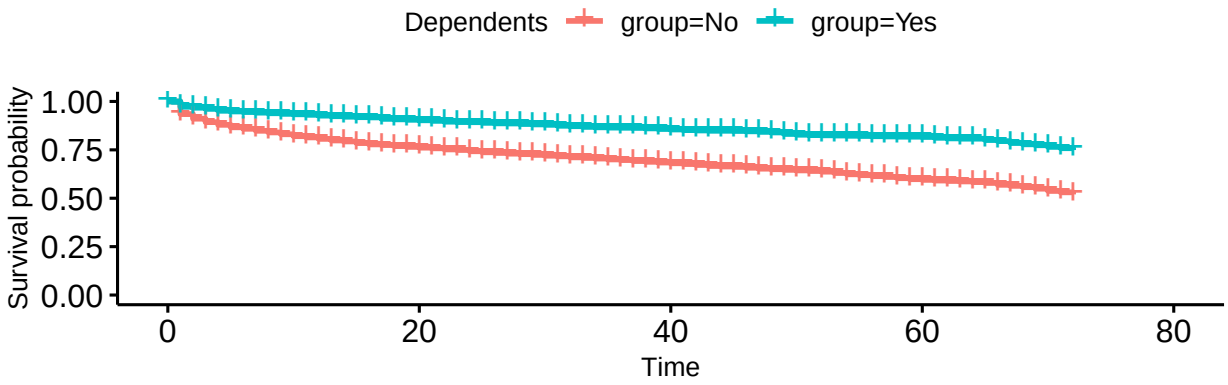
```
## [[1]]
## $`1`
```






```
##
## $`2`
```






```
##
## attr(,"class")
## [1] "list"      "ggarrange"
##
## [[2]]
## [[2]]$gender
## Call:
## survdiff(formula = Surv(tenure, Churn) ~ q, data = d)
##
##             N Observed Expected (O-E)^2/E (O-E)^2/V
## q=Female 3488      939      923     0.261     0.526
## q=Male   3555      930      946     0.255     0.526
##
##  Chisq= 0.5  on 1 degrees of freedom, p= 0.5
##
## [[2]]$SeniorCitizen
## Call:
## survdiff(formula = Surv(tenure, Churn) ~ q, data = d)
##
##        N Observed Expected (O-E)^2/E (O-E)^2/V
## q=1 1142      476      309      89.8       109
## q=0 5901     1393     1560      17.8       109
##
##  Chisq= 110  on 1 degrees of freedom, p= <2e-16
##
## [[2]]$Partner
```

```
## Call:
## survdiff(formula = Surv(tenure, Churn) ~ q, data = d)
##
##          N Observed Expected (O-E)^2/E (O-E)^2/V
## q=No  3641     1200      773       236       424
## q=Yes 3402      669     1096       166       424
##
##  Chisq= 424  on 1 degrees of freedom, p= <2e-16
##
## [[2]]$Dependents
## Call:
## survdiff(formula = Surv(tenure, Churn) ~ q, data = d)
##
##          N Observed Expected (O-E)^2/E (O-E)^2/V
## q=No  4933     1543     1234      77.3       233
## q=Yes 2110      326      635     150.3       233
##
##  Chisq= 233  on 1 degrees of freedom, p= <2e-16
```

##Phone service is not a determinant of leaving, all the other services have influnce.

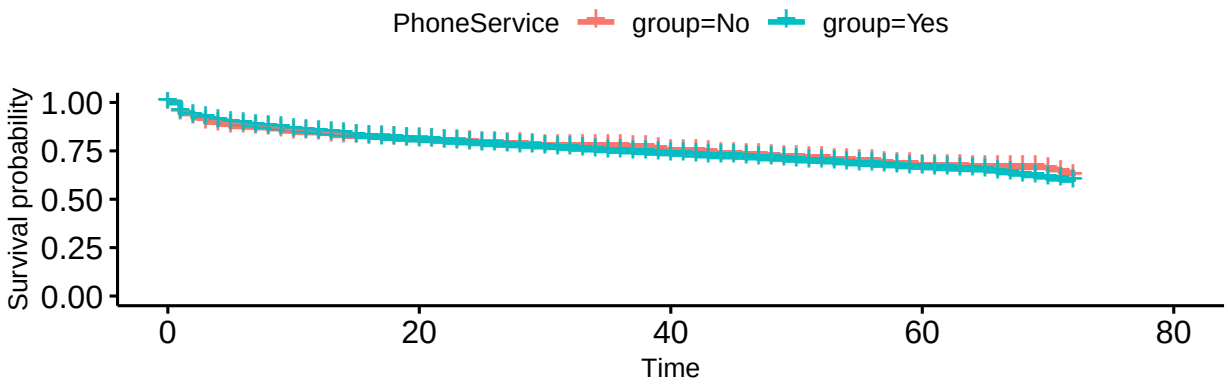
##Specifically, customers who doesn't sign up for internet service are most unlikely to stay.

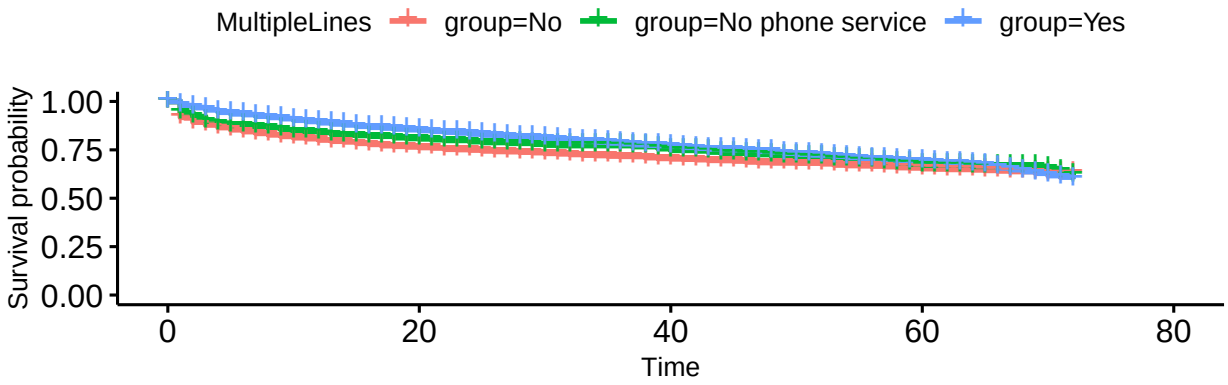
##For internet service, customers who use Fib will more likely to leave than using DSL

##For all the other service, not using service will more likely to leave

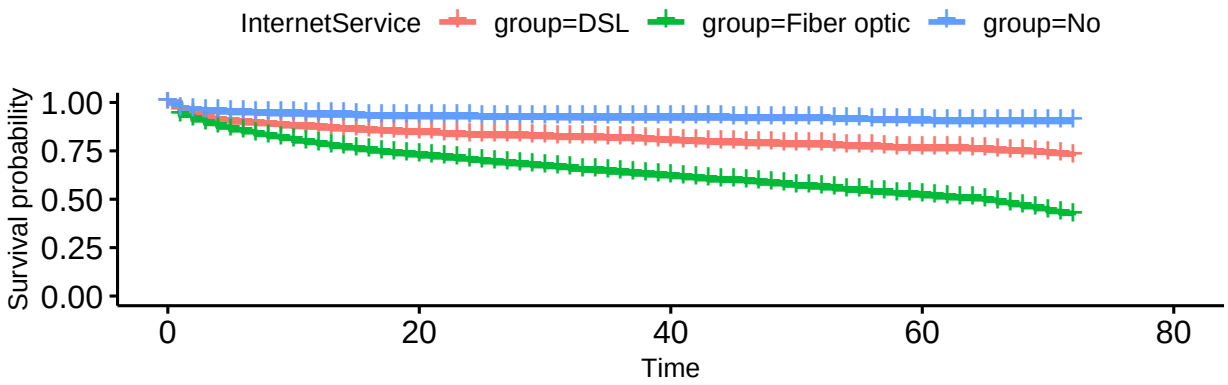
```
service = names(churn)[7:15]
KMcurve(churn,service,'tenure','Churn')
```

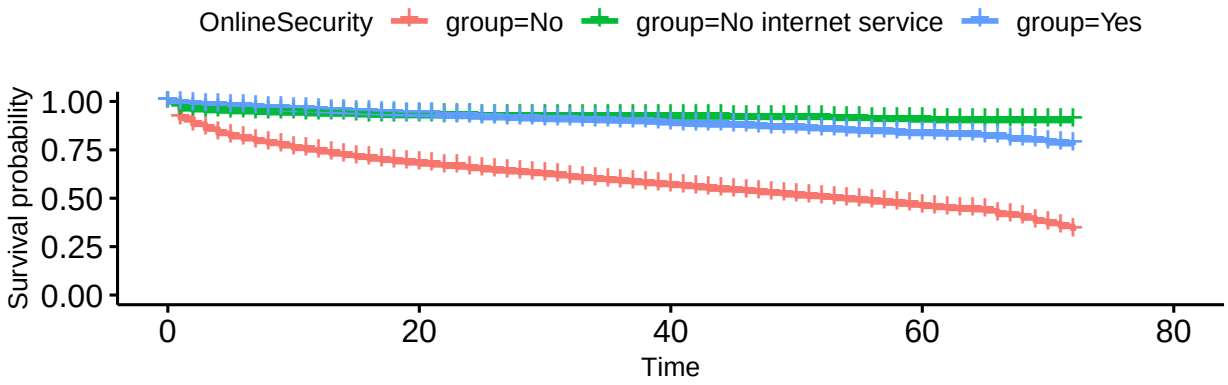
```
## [[1]]
## $`1`
```


PhoneService
group=No
group=Yes
Survival probability
1.00
0.75
0.50
0.25
0.00
0
20
40
60
80
Time

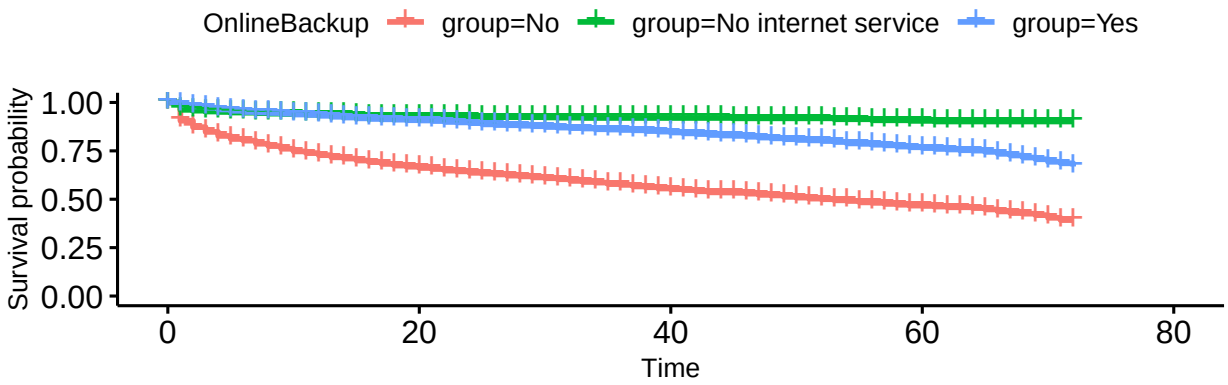

MultipleLines
group=No
group=No phone service
group=Yes
Survival probability
1.00
0.75
0.50
0.25
0.00
0
20
40
60
80
Time

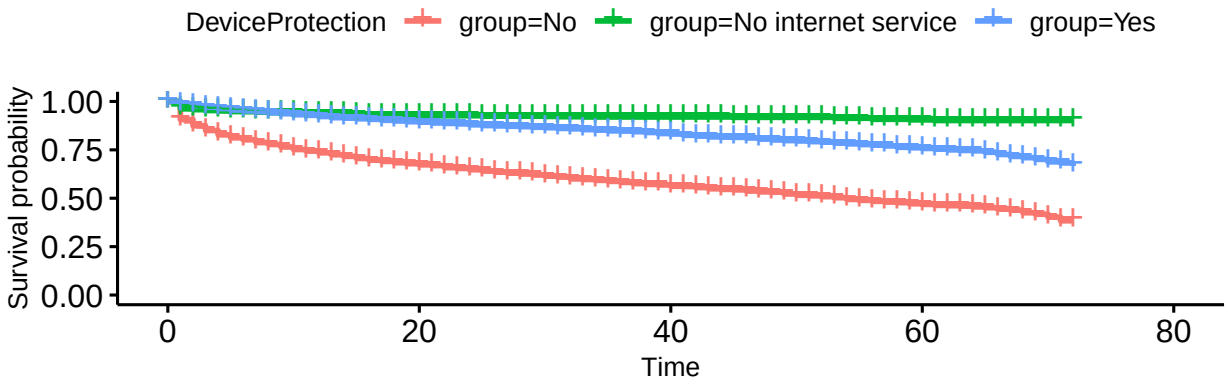
```
##
## $`2`
```


InternetService
group=DSL
group=Fiber optic
group=No
Survival probability
1.00
0.75
0.50
0.25
0.00
0
20
40
60
80
Time

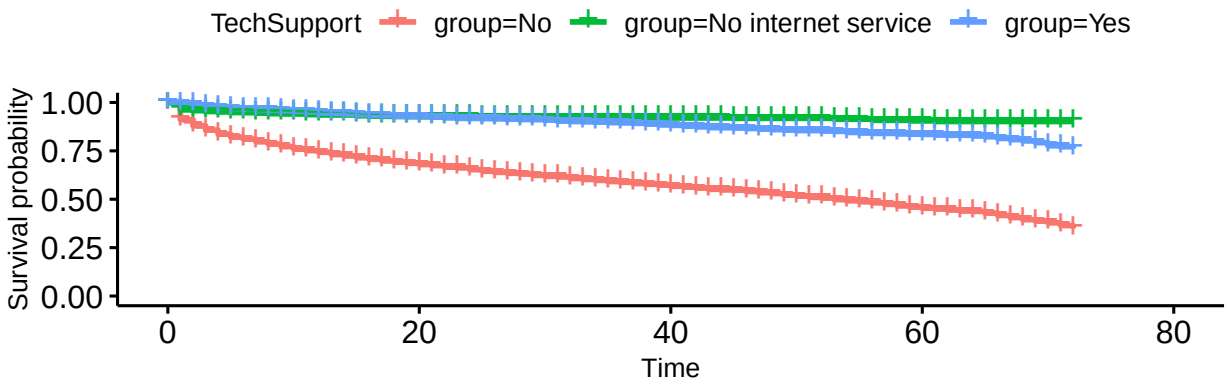

OnlineSecurity
group=No
group=No internet service
group=Yes
Survival probability
1.00
0.75
0.50
0.25
0.00
0
20
40
60
80
Time

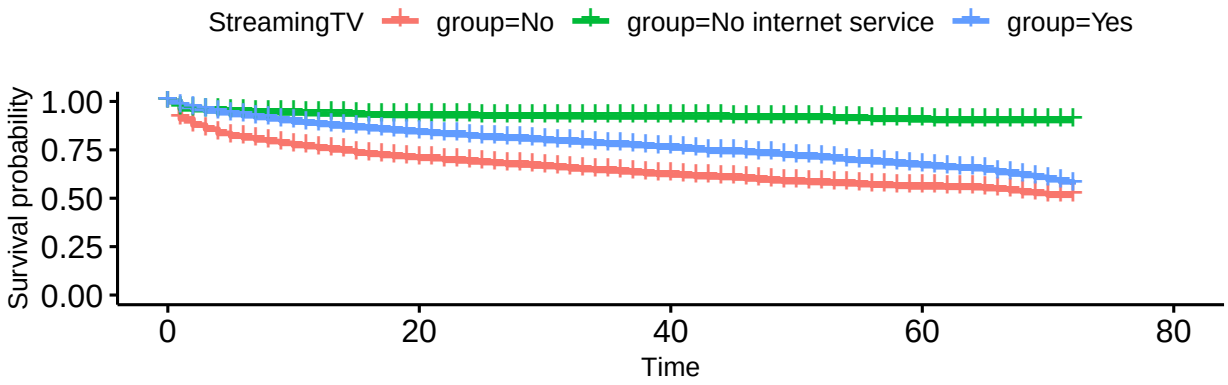
```
##
## $`3`
```


OnlineBackup
group=No
group=No internet service
group=Yes
Survival probability
1.00
0.75
0.50
0.25
0.00
0
20
40
60
80
Time

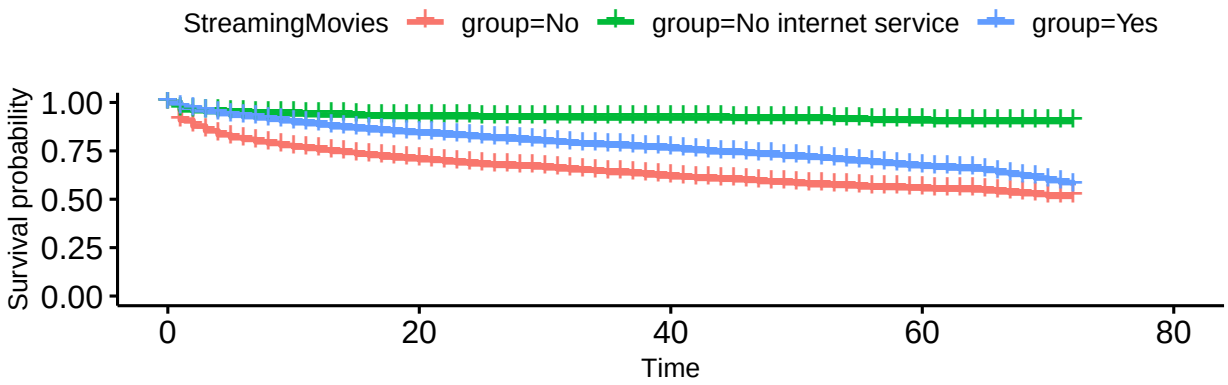

DeviceProtection
group=No
group=No internet service
group=Yes
Survival probability
1.00
0.75
0.50
0.25
0.00
0
20
40
60
80
Time

```
##
## $`4`
```


TechSupport
group=No
group=No internet service
group=Yes
Survival probability
1.00
0.75
0.50
0.25
0.00
0
20
40
60
80
Time


StreamingTV
group=No
group=No internet service
group=Yes
Survival probability
1.00
0.75
0.50
0.25
0.00
0
20
40
60
80
Time

```
##
## $`5`
```




```
##
## attr(,"class")
## [1] "list"      "ggarrange"
##
## [[2]]
## [[2]]$PhoneService
## Call:
## survdiff(formula = Surv(tenure, Churn) ~ q, data = d)
##
##          N Observed Expected (O-E)^2/E (O-E)^2/V
## q=No   682      170      178    0.3828     0.431
## q=Yes 6361     1699     1691    0.0404     0.431
##
##  Chisq= 0.4  on 1 degrees of freedom, p= 0.5
##
## [[2]]$MultipleLines
## Call:
## survdiff(formula = Surv(tenure, Churn) ~ q, data = d)
##
##                       N Observed Expected (O-E)^2/E (O-E)^2/V
## q=No               3390      849      735    17.597    30.276
## q=No phone service  682      170      178     0.383     0.431
## q=Yes              2971      850      955    11.646    24.850
##
##  Chisq= 31  on 2 degrees of freedom, p= 2e-07
##
```

```
## [[2]]$InternetService
## Call:
## survdiff(formula = Surv(tenure, Churn) ~ q, data = d)
##
##                   N Observed Expected (O-E)^2/E (O-E)^2/V
## q=DSL          2421      459      649      55.4      86.4
## q=Fiber optic 3096     1297      832     260.1     477.1
## q=No          1526      113      388     195.4     251.2
##
##  Chisq= 520  on 2 degrees of freedom, p= <2e-16
##
## [[2]]$OnlineSecurity
## Call:
## survdiff(formula = Surv(tenure, Churn) ~ q, data = d)
##
##                            N Observed Expected (O-E)^2/E (O-E)^2/V
## q=No                    3498     1461      794       559      1010
## q=No internet service 1526      113      388       195       251
## q=Yes                   2019      295      686       223       367
##
##  Chisq= 1014  on 2 degrees of freedom, p= <2e-16
##
## [[2]]$OnlineBackup
## Call:
## survdiff(formula = Surv(tenure, Churn) ~ q, data = d)
##
##                            N Observed Expected (O-E)^2/E (O-E)^2/V
## q=No                    3088     1233      664       488       792
## q=No internet service 1526      113      388       195       251
## q=Yes                   2429      523      817       106       197
##
##  Chisq= 821  on 2 degrees of freedom, p= <2e-16
##
## [[2]]$DeviceProtection
## Call:
## survdiff(formula = Surv(tenure, Churn) ~ q, data = d)
##
##                            N Observed Expected (O-E)^2/E (O-E)^2/V
## q=No                    3095     1211      664       450       729
## q=No internet service 1526      113      388       195       251
## q=Yes                   2422      545      816        90       167
##
##  Chisq= 764  on 2 degrees of freedom, p= <2e-16
##
## [[2]]$TechSupport
## Call:
## survdiff(formula = Surv(tenure, Churn) ~ q, data = d)
##
##                            N Observed Expected (O-E)^2/E (O-E)^2/V
## q=No                    3473     1446      788       549       985
## q=No internet service 1526      113      388       195       251
## q=Yes                   2044      310      692       211       349
##
##  Chisq= 990  on 2 degrees of freedom, p= <2e-16
```

```
##
## [[2]]$StreamingTV
## Call:
## survdiff(formula = Surv(tenure, Churn) ~ q, data = d)
##
##                           N Observed Expected (O-E)^2/E (O-E)^2/V
## q=No                   2810      942      624    162.47    252.07
## q=No internet service 1526      113      388    195.36    251.20
## q=Yes                  2707      814      857      2.14      4.09
##
##  Chisq= 368  on 2 degrees of freedom, p= <2e-16
##
## [[2]]$StreamingMovies
## Call:
## survdiff(formula = Surv(tenure, Churn) ~ q, data = d)
##
##                           N Observed Expected (O-E)^2/E (O-E)^2/V
## q=No                   2785      938      614    171.50    264.22
## q=No internet service 1526      113      388    195.36    251.20
## q=Yes                  2732      818      867      2.76      5.33
##
##  Chisq= 378  on 2 degrees of freedom, p= <2e-16
```

##Sining for shorter contract will more likely to leave, monthly contract loses times than yearly contract.

No paperless billing is more likely to leave

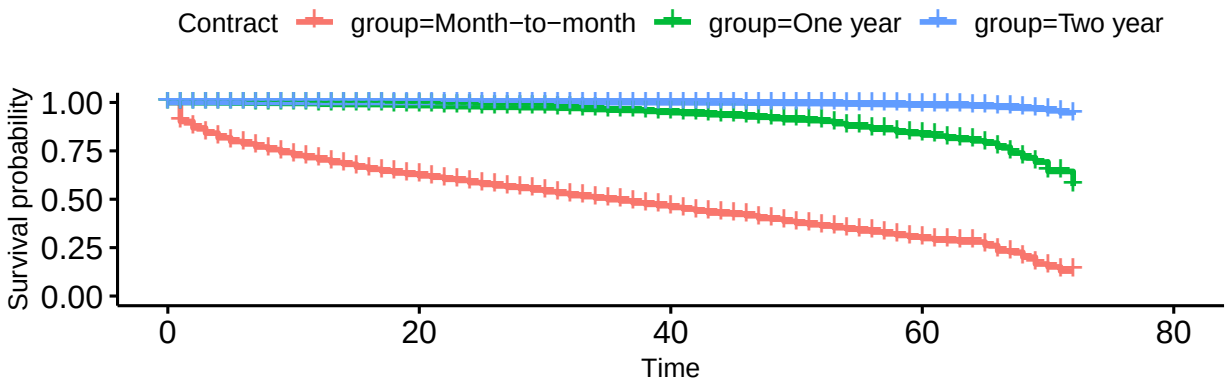
Electronic check is more likely to leave compared with credit card, bank transfer and mailed check

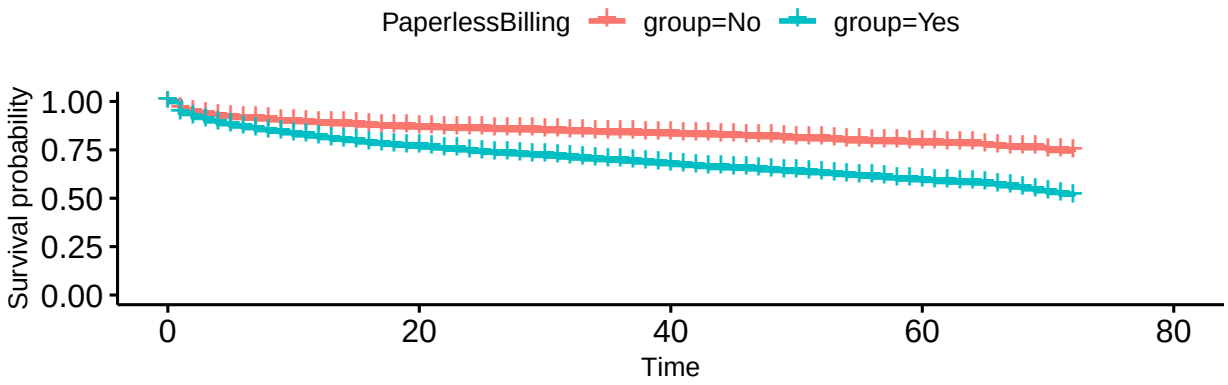
```
unique(churn$PaymentMethod)
```

```
## [1] Electronic check          Mailed check
## [3] Bank transfer (automatic) Credit card (automatic)
## 4 Levels: Bank transfer (automatic) ... Mailed check
```

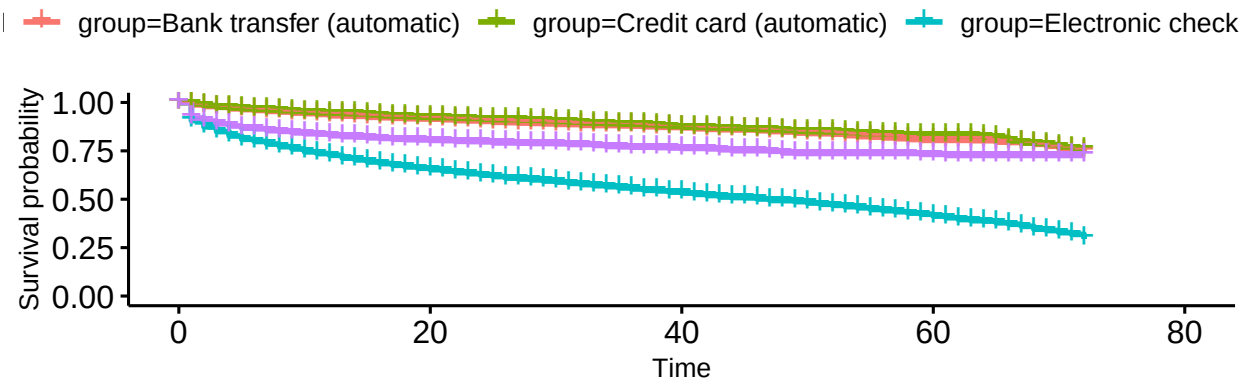
```
#notice that charge is continuous variable
account = names(churn)[16:18]
KMcurve(churn,account,'tenure','Churn')
```

```
## [[1]]
## $`1`
```


Contract
group=Month−to−month
group=One year
group=Two year
Survival probability
1.00
0.75
0.50
0.25
0.00
0
20
40
60
80
Time


PaperlessBilling
group=No
group=Yes
Survival probability
1.00
0.75
0.50
0.25
0.00
0
20
40
60
80
Time

```
##
## $`2`
```




```
##
## attr(,"class")
## [1] "list"      "ggarrange"
##
## [[2]]
## [[2]]$Contract
## Call:
## survdiff(formula = Surv(tenure, Churn) ~ q, data = d)
##
##                     N Observed Expected (O-E)^2/E (O-E)^2/V
## q=Month-to-month 3875     1655      708      1265      2304
## q=One year       1473      166      471       197       270
## q=Two year       1695       48      690       597      1061
##
##  Chisq= 2353  on 2 degrees of freedom, p= <2e-16
##
## [[2]]$PaperlessBilling
## Call:
## survdiff(formula = Surv(tenure, Churn) ~ q, data = d)
##
##          N Observed Expected (O-E)^2/E (O-E)^2/V
## q=No  2872      469      759     110.6       190
## q=Yes 4171     1400     1110      75.6       190
##
##  Chisq= 190  on 1 degrees of freedom, p= <2e-16
##
```

```
## [[2]]$PaymentMethod
## Call:
## survdiff(formula = Surv(tenure, Churn) ~ q, data = d)
##
##                                N Observed Expected (O-E)^2/E (O-E)^2/V
## q=Bank transfer (automatic) 1544      258      512    126.12    178.66
## q=Credit card (automatic)   1522      232      502    145.00    203.63
## q=Electronic check          2365     1071      528    558.93    803.75
## q=Mailed check              1612      308      327      1.14      1.43
##
##  Chisq= 865  on 3 degrees of freedom, p= <2e-16
```

In conclusion, our target customers are younger people with partners and dependents.

In order to retain customer, the overall strategy is to attract more signing up our services. Also, a save and convenient paying environment by credit card and bank transfer will help to retain customers.

```
head(churn)
```

```
##   customerID gender SeniorCitizen Partner Dependents tenure PhoneService
## 1 7590-VHVEG Female             0     Yes         No      1           No
## 2 5575-GNVDE   Male             0      No         No     34          Yes
## 3 3668-QPYBK   Male             0      No         No      2          Yes
## 4 7795-CFOCW   Male             0      No         No     45           No
## 5 9237-HQITU Female             0      No         No      2          Yes
## 6 9305-CDSKC Female             0      No         No      8          Yes
##      MultipleLines InternetService OnlineSecurity OnlineBackup
## 1 No phone service             DSL             No          Yes
## 2               No             DSL            Yes           No
## 3               No             DSL            Yes          Yes
## 4 No phone service             DSL            Yes           No
## 5               No     Fiber optic             No           No
## 6              Yes     Fiber optic             No           No
##   DeviceProtection TechSupport StreamingTV StreamingMovies       Contract
## 1               No          No          No              No Month-to-month
## 2              Yes          No          No              No       One year
## 3               No          No          No              No Month-to-month
## 4              Yes         Yes          No              No       One year
## 5               No          No          No              No Month-to-month
## 6              Yes          No         Yes             Yes Month-to-month
##   PaperlessBilling             PaymentMethod MonthlyCharges TotalCharges
## 1              Yes          Electronic check          29.85        29.85
## 2               No              Mailed check          56.95      1889.50
## 3              Yes              Mailed check          53.85       108.15
## 4               No Bank transfer (automatic)          42.30      1840.75
## 5              Yes          Electronic check          70.70       151.65
## 6              Yes          Electronic check          99.65       820.50
##   Churn
## 1     0
## 2     0
## 3     1
## 4     0
```

```
## 5     1
## 6     1
```

Revenue

tenure is month unit

How much we earn from those without internet service

```
inter <- tapply(churn$MonthlyCharges , churn$InternetService, mean)
inter
```

```
##         DSL Fiber optic          No
##    58.10217    91.50013    21.07919
```

```
km = survfit(Surv(tenure,Churn) ~ InternetService, data=churn)
print(km, print.rmean = T, rmean=60)
```

```
## Call: survfit(formula = Surv(tenure, Churn) ~ InternetService, data = churn)
##
##                                n events *rmean *se(rmean) median 0.95LCL
## InternetService=DSL         2421    459   49.9      0.440     NA      NA
## InternetService=Fiber optic 3096   1297   41.3      0.440     65      60
## InternetService=No          1526    113   55.6      0.402     NA      NA
##                             0.95UCL
## InternetService=DSL              NA
## InternetService=Fiber optic      67
## InternetService=No               NA
##     * restricted mean with upper limit =  60
```

```
c(10.9,10.4,11.4)*inter
```

```
##         DSL Fiber optic          No
##    633.3136    951.6013    240.3028
```

```
c(21.1,19.3,22.6)*inter
```

```
##         DSL Fiber optic          No
##   1225.9558   1765.9525    476.3898
```

```
c(49.9,41.3,55.6)*inter
```

```
##         DSL Fiber optic          No
##    2899.298    3778.955    1172.003
```

Monthly charge more and yearly charge less

```
meanCharge = tapply(churn$MonthlyCharges , churn$Contract,mean)
meanCharge
```

```
## Month-to-month       One year       Two year
##       66.39849       65.04861       60.77041
```

retention rate in specific time range

```
km = survfit(Surv(tenure,Churn) ~ Contract, data=churn)
print(km, print.rmean = T, rmean=30)
```

```
## Call: survfit(formula = Surv(tenure, Churn) ~ Contract, data = churn)
##
##                             n events *rmean *se(rmean) median 0.95LCL
## Contract=Month-to-month 3875   1655   20.9    0.20173     35      32
## Contract=One year       1473    166   29.6    0.06543     NA      72
## Contract=Two year       1695     48   30.0    0.00325     NA      NA
##                         0.95UCL
## Contract=Month-to-month      38
## Contract=One year            NA
## Contract=Two year            NA
##     * restricted mean with upper limit =  30
```