Survival Analysis of Breast Cancer Dataset.

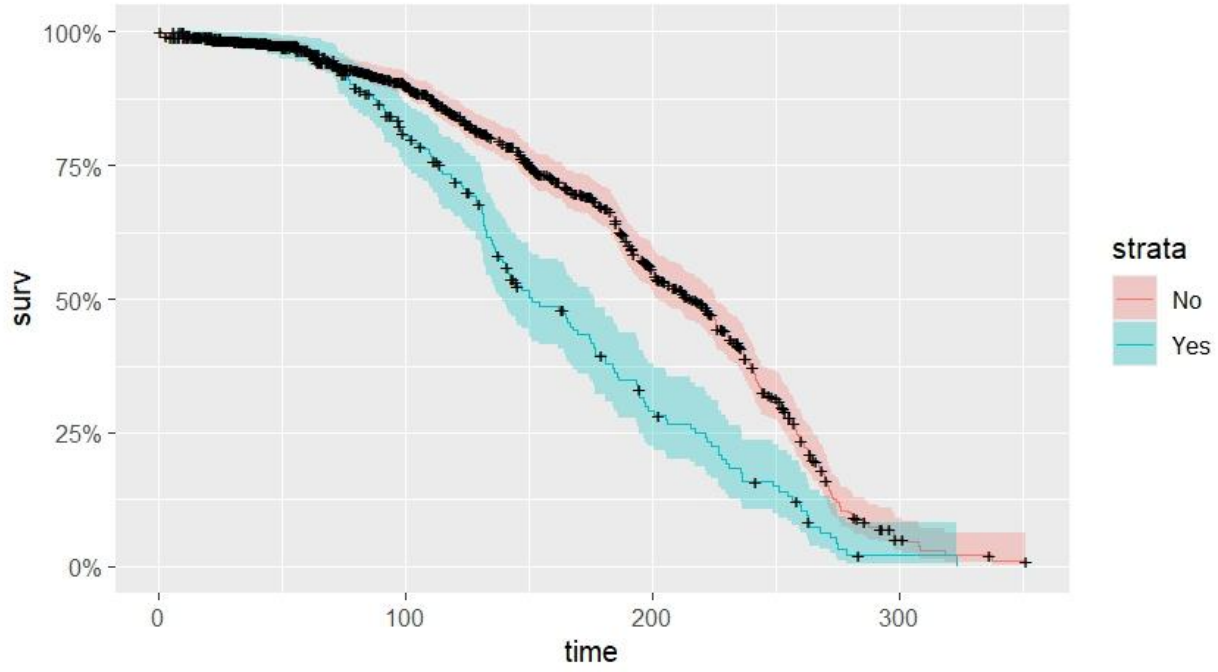
```
#Import the data
data <- read.csv("C:\\Users\\Christianah.O_BROOKS\\Downloads\\Breast Cancer METABRIC.csv", header = TRUE, stringsAsFactors = FALSE)
head(data)
```

```
> head(data)
  Patient.ID   Age Type.of.Breast.Surgery   Cancer.Type
1    MB-0000 75.65              Mastectomy Breast Cancer
2    MB-0002 43.19      Breast Conserving Breast Cancer
3    MB-0005 48.87              Mastectomy Breast Cancer
4    MB-0006 47.68              Mastectomy Breast Cancer
5    MB-0008 76.97              Mastectomy Breast Cancer
6    MB-0010 78.77              Mastectomy Breast Cancer
```

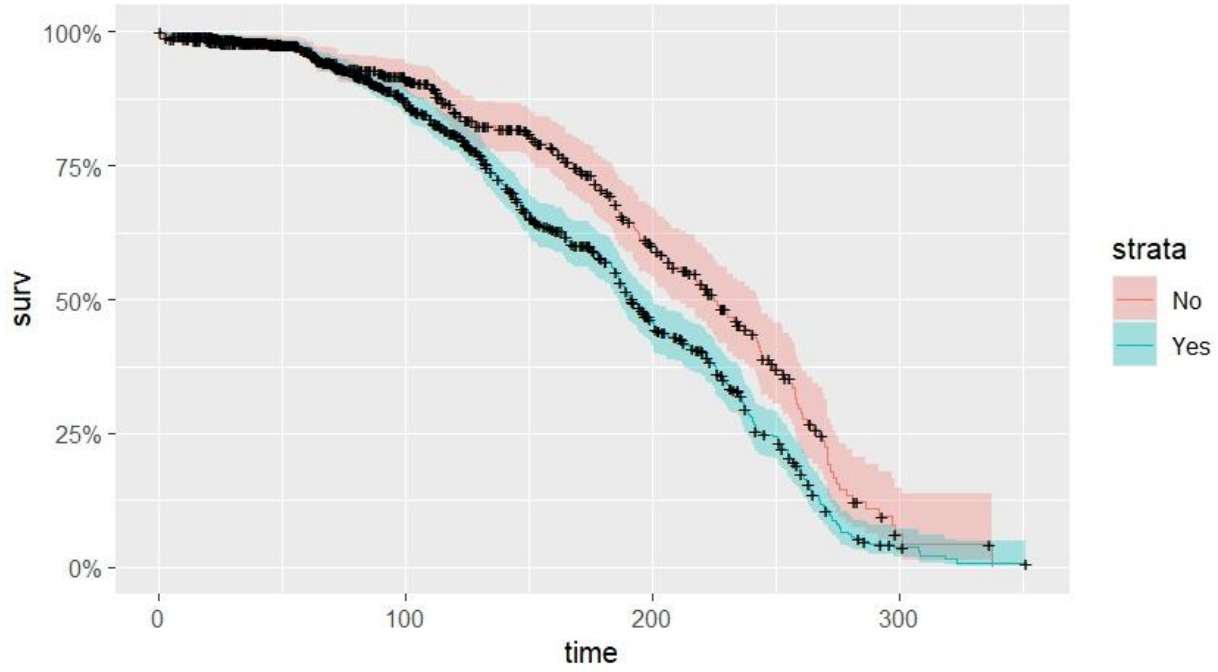
```
#Data Cleaning
Clean_data <- na.omit(data)
Clean_data
View(Clean_data)
cat("Original number of rows:", nrow(data), "\n")
cat("Number of rows after removing missing values:", nrow(Clean_data), "\n")

#Load libraries
library(survival)
library(ranger)
library(ggplot2)
library(dplyr)
library(ggfortify)
library(survminer)
library(tidyr)

#Kaplan-Mier Analysis
Clean_data$PatientStatus <- ifelse(Clean_data$status == "Living", 1, 0)
km <- with(Clean_data, Surv(time, PatientStatus))
head(km, 80)

km_fit <- survfit(Surv(time, PatientStatus) ~ 1, data=Clean_data)
summary(km_fit, times = c(1,30,60,90*(1:10)))
```

```
> km_fit <- survfit(Surv(time, PatientStatus) ~ 1, data=Clean_data)
> summary(km_fit, times = c(1,30,60,90*(1:10)))
Call: survfit(formula = Surv(time, PatientStatus) ~ 1, data = Clean_data)

 time n.risk n.event survival std.err lower 95% CI
    1   1309       0    1.000 0.00000        1.000
   30   1184      20    0.984 0.00347        0.978
   60    989      20    0.966 0.00534        0.955
   90    812      58    0.906 0.00909        0.889
  180    374     211    0.618 0.01778        0.584
  270     47     223    0.141 0.01719        0.111
upper 95% CI
       1.000
       0.991
       0.976
       0.924
       0.654
       0.179
```

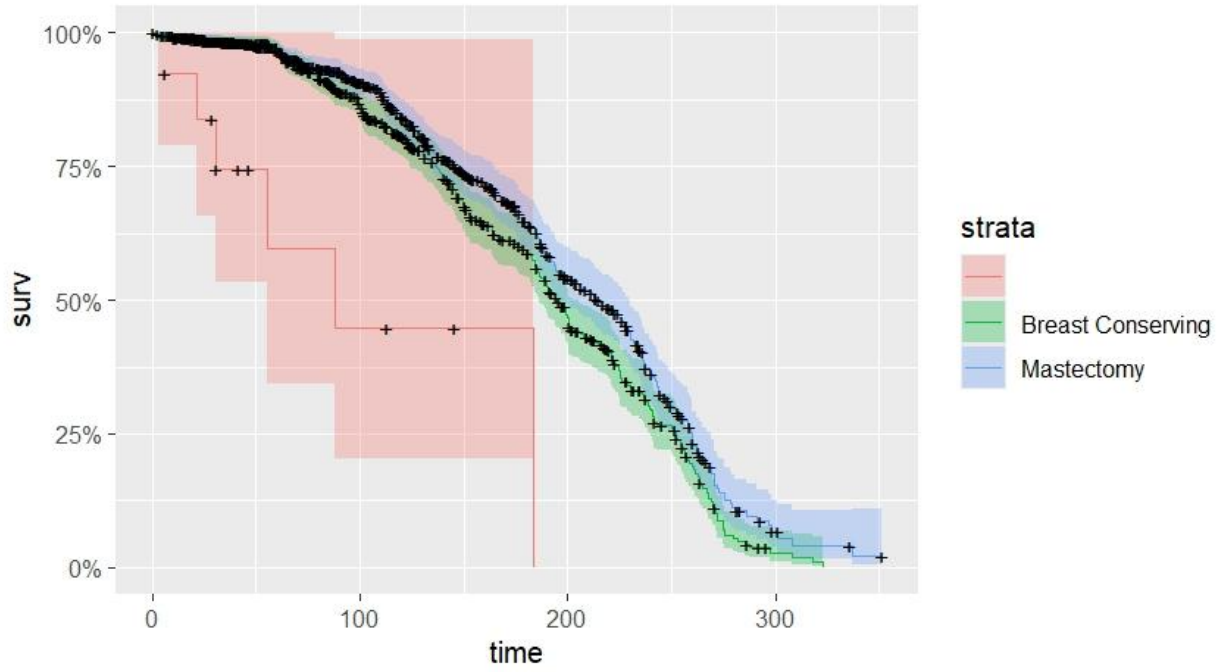
km_trt_fit <- survfit(Surv(time, PatientStatus) ~ Chemotherapy, data=Clean_data)
autoplot(km_trt_fit)



km_trt_fit <- survfit(Surv(time, PatientStatus) ~ Radio.Therapy, data=Clean_data)
autoplot(km_trt_fit)



km_trt_fit <- survfit(Surv(time, PatientStatus) ~ Type.of.Breast.Surgery, data=Clean_data)
autoplot(km_trt_fit)



#Cox Proportional Hazard Model
cox_model <- coxph(Surv(time, PatientStatus) ~ Age + Tumor.Size + ER.Status +
PR.Status + HER2.Status + Chemotherapy + Tumor.Stage, data = Clean_data)

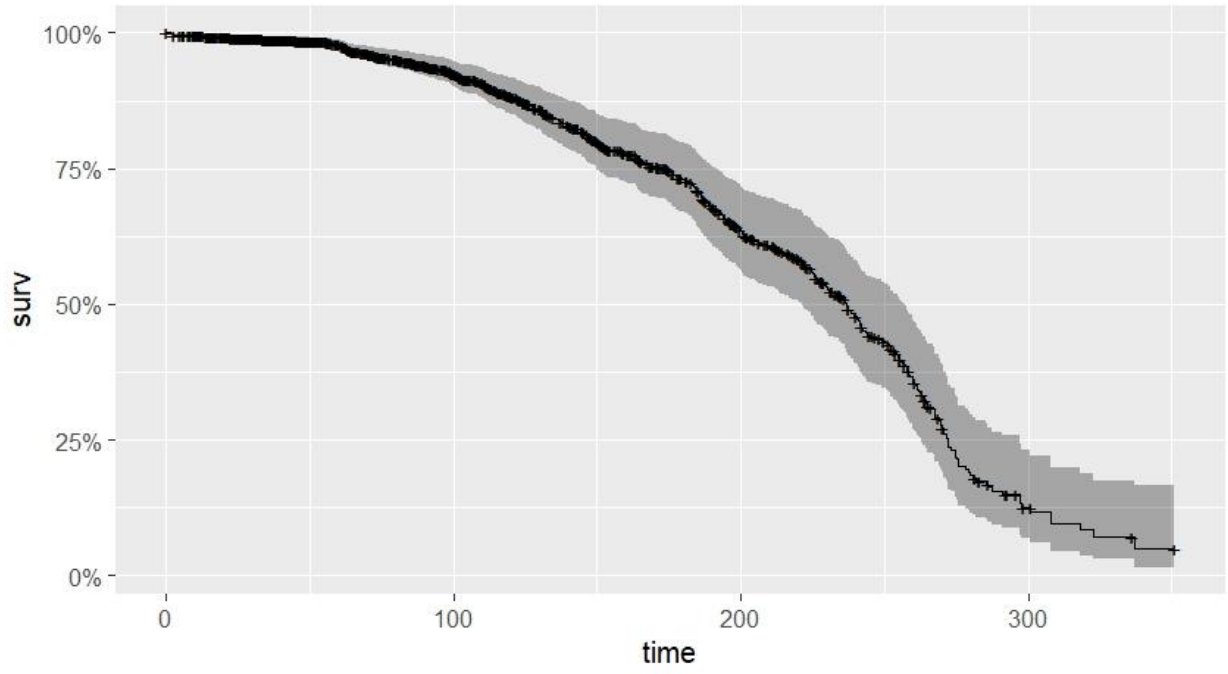
summary(cox_model)

```
                          coef
Age                 -0.0004890
Tumor.Size          -0.0004829
ER.StatusPositive    0.2721494
PR.StatusPositive    0.0473402
HER2.StatusPositive  0.2157262
ChemotherapyYes      0.7666532
Tumor.Stage         -0.0388648
                     exp(coef)
Age                  0.9995112
Tumor.Size           0.9995172
ER.StatusPositive    1.3127832
PR.StatusPositive    1.0484786
HER2.StatusPositive  1.2407626
ChemotherapyYes      2.1525499
Tumor.Stage          0.9618807

                     se(coef)
Age                 0.0041243
Tumor.Size          0.0041351
ER.StatusPositive   0.1440179
PR.StatusPositive   0.1045753
HER2.StatusPositive 0.1486316
ChemotherapyYes     0.1448169
Tumor.Stage         0.0964116
                         z Pr(>|z|)
Age                 -0.119   0.9056
Tumor.Size          -0.117   0.9070
ER.StatusPositive    1.890   0.0588
PR.StatusPositive    0.453   0.6508
HER2.StatusPositive  1.451   0.1467
ChemotherapyYes      5.294  1.2e-07
Tumor.Stage         -0.403   0.6869
```

```
Age
Tumor.Size
ER.StatusPositive    .
PR.StatusPositive
HER2.StatusPositive
ChemotherapyYes     ***
Tumor.Stage
---
Signif. codes:
  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.0
  ‘.’ 0.1 ‘ ’ 1

                    exp(coef)
Age                    0.9995
Tumor.Size             0.9995
ER.StatusPositive      1.3128
PR.StatusPositive      1.0485
HER2.StatusPositive    1.2408
ChemotherapyYes        2.1525
Tumor.Stage            0.9619


                    exp(-coef)
Age                     1.0005
Tumor.Size              1.0005
ER.StatusPositive       0.7617
PR.StatusPositive       0.9538
HER2.StatusPositive     0.8060
ChemotherapyYes         0.4646
Tumor.Stage             1.0396
                    lower .95
Age                    0.9915
Tumor.Size             0.9914
ER.StatusPositive      0.9899
PR.StatusPositive      0.8542
HER2.StatusPositive    0.9272
ChemotherapyYes        1.6206
Tumor.Stage            0.7963
```

```
Age                          1.008
Tumor.Size                   1.008
ER.StatusPositive            1.741
PR.StatusPositive            1.287
HER2.StatusPositive          1.660
ChemotherapyYes              2.859
Tumor.Stage                  1.162

Concordance= 0.558  (se = 0.014 )
Likelihood ratio test= 39.51  on 7 df,   p=2e-06
Wald test            = 43.66  on 7 df,   p=2e-07
Score (logrank) test = 44.64  on 7 df,   p=2e-07
```

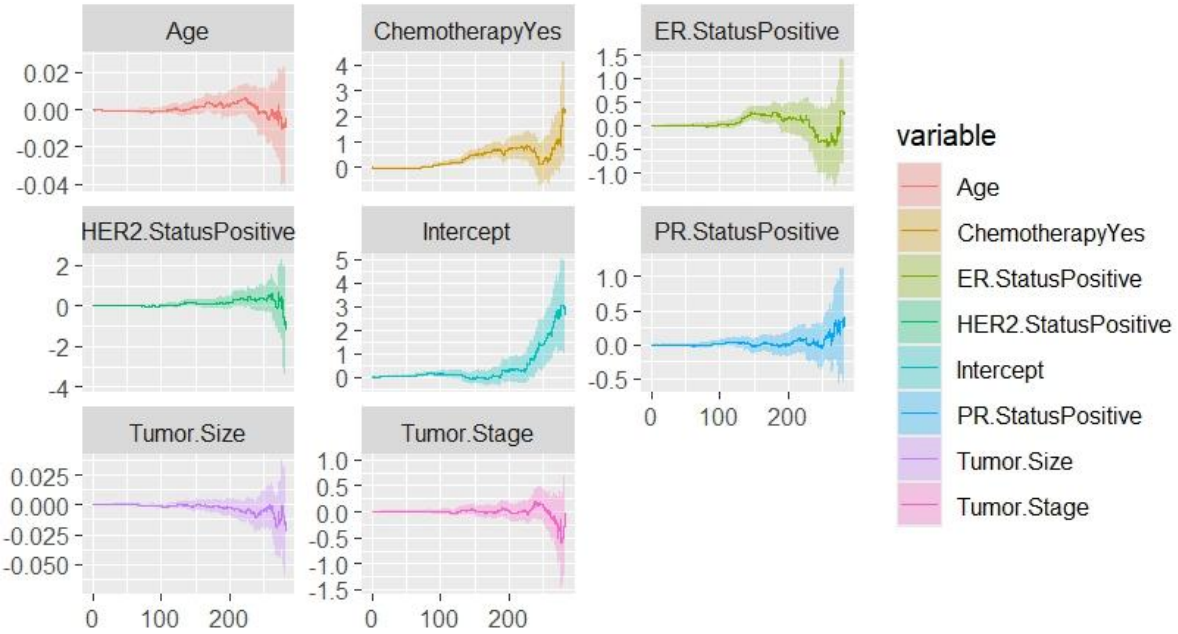
cox_fit <- survfit(cox_model)
autoplot(cox_fit)



#Aalen's additive regression model
aa_fit <-aareg(Surv(time, PatientStatus) ~ Age + Tumor.Size + ER.Status +
PR.Status + HER2.Status + Chemotherapy + Tumor.Stage, data = Clean_data)
summary(aa_fit)

```
                          slope       coef se(coef)
Intercept              3.66e-03  1.37e-03 5.35e-04
Age                    7.80e-06  8.34e-07 7.63e-06
Tumor.Size            -2.76e-05 -7.27e-06 7.86e-06
ER.StatusPositive      2.44e-03  3.98e-04 3.07e-04
PR.StatusPositive      2.64e-04  1.19e-04 2.06e-04
HER2.StatusPositive    1.90e-03  4.48e-04 3.41e-04
ChemotherapyYes        7.90e-03  1.53e-03 3.63e-04
Tumor.Stage            3.19e-04  1.97e-05 1.88e-04
                           z        p
Intercept              2.550 1.06e-02
Age                    0.109 9.13e-01
Tumor.Size            -0.925 3.55e-01
ER.StatusPositive      1.300 1.95e-01
PR.StatusPositive      0.579 5.63e-01
HER2.StatusPositive    1.320 1.88e-01
ChemotherapyYes        4.220 2.43e-05
Tumor.Stage            0.105 9.16e-01

Chisq=28.16 on 7 df, p=0.000205; test weights=aalen
```

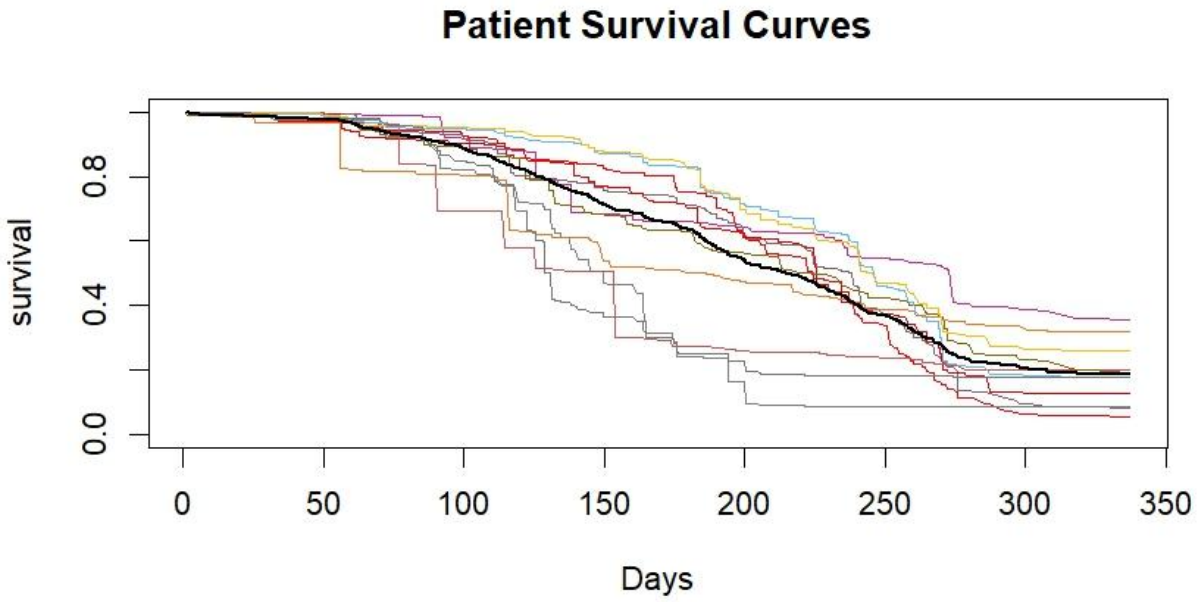
autoplot(aa_fit)




```
#Random forest model
r_fit <- ranger(Surv(time, PatientStatus) ~ Age + Tumor.Size + ER.Status +
          PR.Status + HER2.Status + Chemotherapy + Tumor.Stage,
        data = Clean_data,
        mtry = 4,
        importance = "permutation",
        splitrule = "extratrees",
        verbose = TRUE)
r_fit
```

```
> r_fit
Ranger result

Call:
 ranger(Surv(time, PatientStatus) ~ Age + Tumor.Size + ER.Status +      PR.Statu
s + HER2.Status + Chemotherapy + Tumor.Stage, data = Clean_data,      mtry = 4,
importance = "permutation", splitrule = "extratrees",      verbose = TRUE)

Type:                             Survival
Number of trees:                  500
Sample size:                      1310
Number of independent variables:  7
Mtry:                             4
Target node size:                 3
Variable importance mode:         permutation
Splitrule:                        extratrees
Number of unique death times:     536
Number of random splits:          1
OOB prediction error (1-C):       0.46373
```

```
# Average the survival models
death_times <- r_fit$unique.death.times
surv_prob <- data.frame(r_fit$survival)
avg_prob <- sapply(surv_prob,mean)

# Plot the survival models for each patient
plot(r_fit$unique.death.times,r_fit$survival[1,],
    type = "l",
    ylim = c(0,1),
    col = "red",
    xlab = "Days",
    ylab = "survival",
    main = "Patient Survival Curves")

cols <- colors()
for (n in sample(c(2:dim(Clean_data)[1]), 20)){
  lines(r_fit$unique.death.times, r_fit$survival[n,], type = "l", col = cols[n])
}

lines(death_times, avg_prob, lwd = 2)
legend(500, 0.7, legend = c('Average = black'))
```



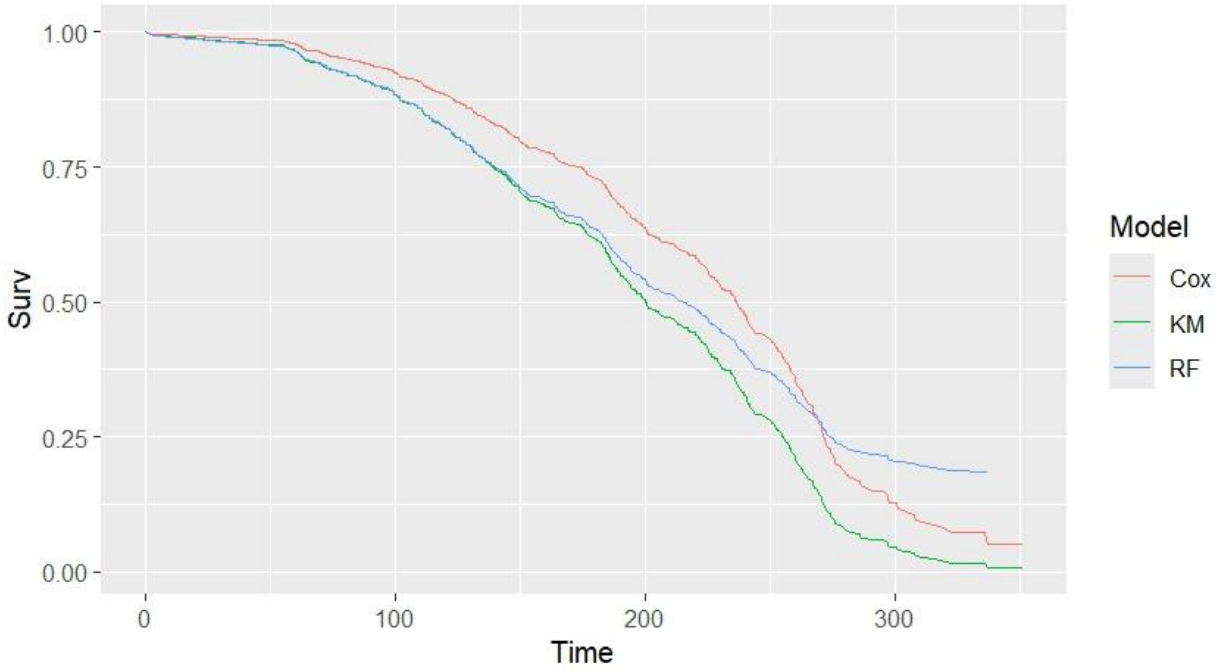

```
vi <- data.frame(sort(round(r_fit$variable.importance, 4), decreasing = TRUE))
names(vi) <- "importance"
head(vi)
```

```
> head(vi)
             importance
Chemotherapy     0.0282
ER.Status        0.0021
Age             -0.0001
PR.Status       -0.0007
Tumor.Size      -0.0017
Tumor.Stage     -0.0031
```

```
cat("Prediction Error = 1 - Harrell's c-index = ", r_fit$prediction.error)
```

```
> cat("Prediction Error = 1 - Harrell's c-index = ", r_fit$prediction.error)
Prediction Error = 1 - Harrell's c-index =  0.46373
```

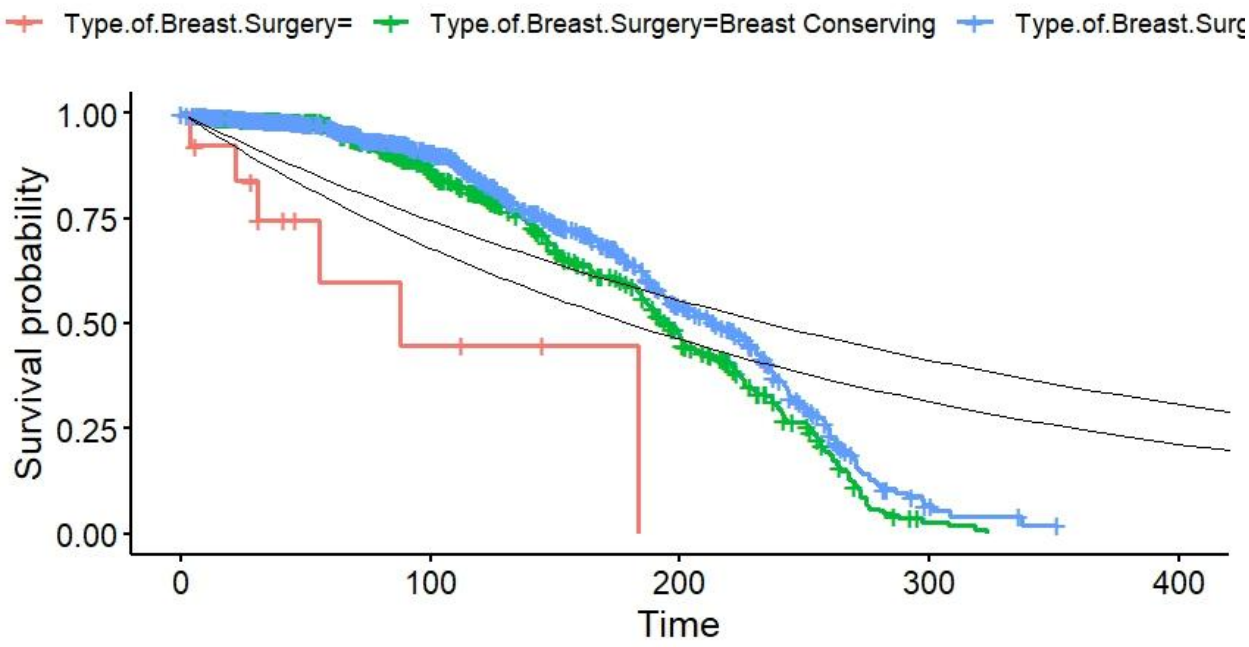
```
#ggplot eyeball comparism of models
kmi <- rep("KM",length(km_fit$time))
km_df <- data.frame(km_fit$time,km_fit$surv,kmi)
names(km_df) <- c("Time","Surv","Model")

coxi <- rep("Cox",length(cox_fit$time))
cox_df <- data.frame(cox_fit$time,cox_fit$surv,coxi)
names(cox_df) <- c("Time","Surv","Model")

rfi <- rep("RF",length(r_fit$unique.death.times))
rf_df <- data.frame(r_fit$unique.death.times,avg_prob,rfi)
names(rf_df) <- c("Time","Surv","Model")


plot_df <- rbind(km_df,cox_df,rf_df)

p <- ggplot(plot_df, aes(x = Time, y = Surv, color = Model))
p + geom_line()
```



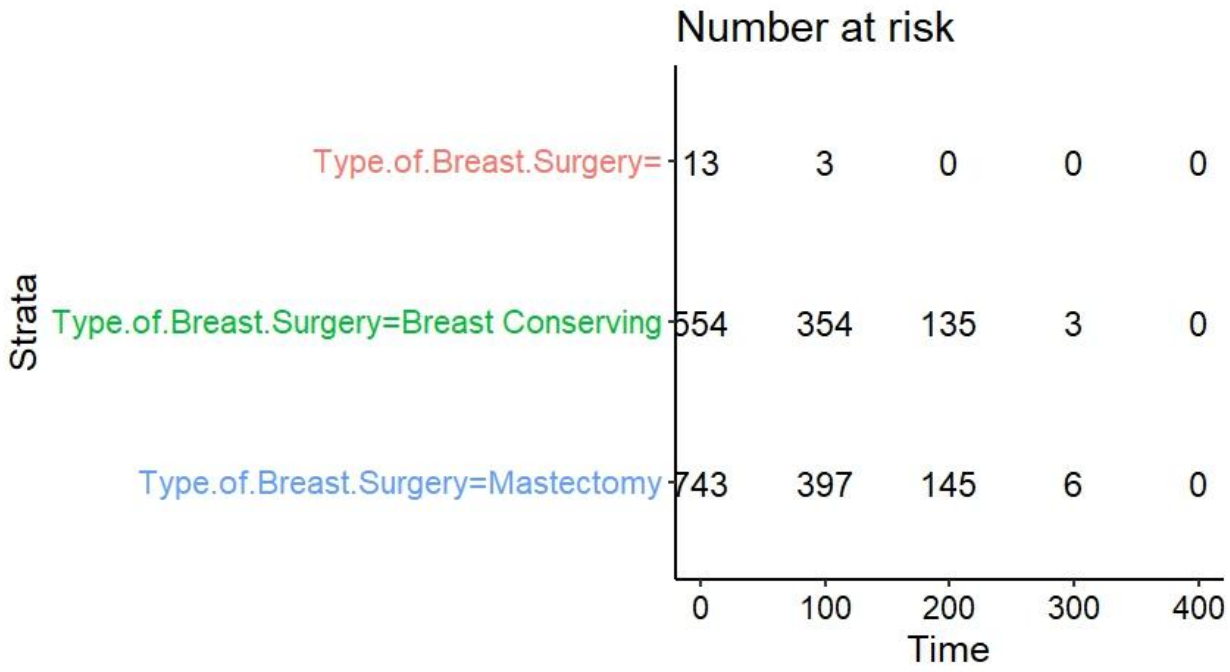
 # Exponential survival model
s <- with(Clean_data,Surv(time,PatientStatus))
fKM <- survfit(s ~ Type.of.Breast.Surgery,data=Clean_data)
sExp <- survreg(s ~ as.factor(Type.of.Breast.Surgery),dist='exp',data=Clean_data)
summary(sExp)

```
> summary(sExp)

Call:
survreg(formula = s ~ as.factor(Type.of.Breast.Surgery), data = Clean_data,
    dist = "exp")
                                                Value Std. Error     z
(Intercept)                                     4.883      0.408 11.96
as.factor(Type.of.Breast.Surgery)Breast Conserving 0.673      0.412  1.63
as.factor(Type.of.Breast.Surgery)Mastectomy        0.944      0.413  2.29
                                                    p
(Intercept)                                     <2e-16
as.factor(Type.of.Breast.Surgery)Breast Conserving  0.103
as.factor(Type.of.Breast.Surgery)Mastectomy         0.022

Scale fixed at 1

Exponential distribution
Loglik(model)= -3785   Loglik(intercept only)= -3791.7
        Chisq= 13.35 on 2 degrees of freedom, p= 0.0013
Number of Newton-Raphson Iterations: 5
n= 1310
```

```
pred.Type.of.Breast.Surgery1 = predict(sExp, newdata=list(Type.of.Breast.Surgery="Breast
Conserving"),type="quantile",p=seq(.01,.99,by=.01))
pred.Type.of.Breast.Surgery2 = predict(sExp,
newdata=list(Type.of.Breast.Surgery="Mastectomy"),type="quantile",p=seq(.01,.99,by=.01))

df = data.frame(y=seq(.99,.01,by=-.01),
Type.of.Breast.Surgery1=pred.Type.of.Breast.Surgery1,
Type.of.Breast.Surgery2=pred.Type.of.Breast.Surgery2)
df_long = gather(df, key= "Type.of.Breast.Surgery", value="time", -y)

p = ggsurvplot(fKM, data = Clean_data, risk.table = T)
p$plot = p$plot + geom_line(data=df_long, aes(x=time, y=y, group=Type.of.Breast.Surgery))
p$plot
```



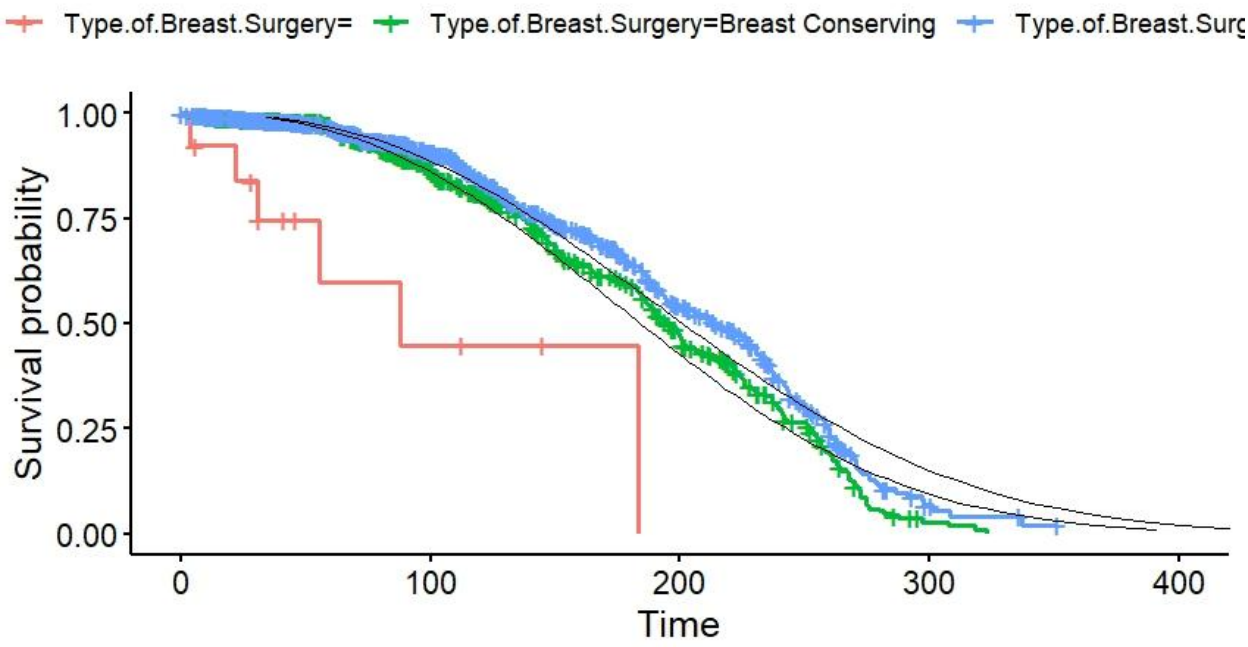

```
p$table
```



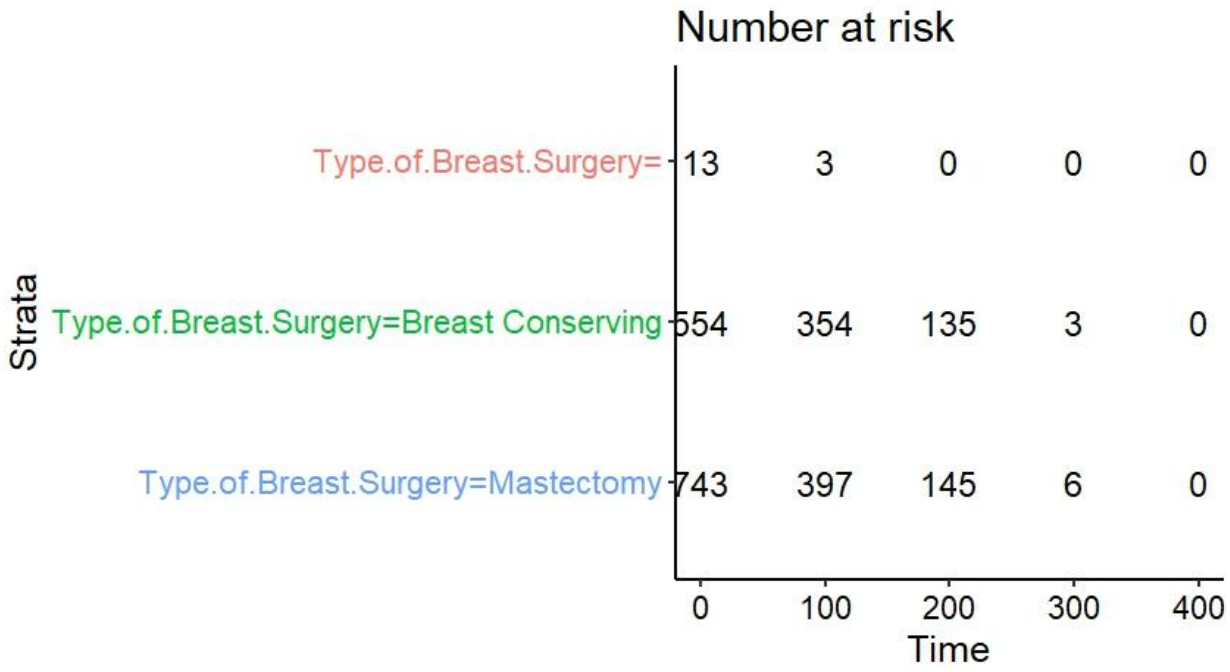
#Weibull survival model
s <- with(Clean_data,Surv(time,PatientStatus))
fKM <- survfit(s ~ Type.of.Breast.Surgery,data=Clean_data)
sWei <- survreg(s ~ as.factor(Type.of.Breast.Surgery), dist = 'weibull', data = Clean_data)
summary(sWei)

```
> summary(sWei)

Call:
survreg(formula = s ~ as.factor(Type.of.Breast.Surgery), data = Clean_data,
    dist = "weibull")
                                                 Value Std. Error      z
(Intercept)                                     4.8016     0.1627  29.52
as.factor(Type.of.Breast.Surgery)Breast Conserving  0.5593     0.1643   3.40
as.factor(Type.of.Breast.Surgery)Mastectomy         0.6477     0.1646   3.93
Log(scale)                                     -0.9206     0.0323 -28.50
                                                    p
(Intercept)                                     < 2e-16
as.factor(Type.of.Breast.Surgery)Breast Conserving 0.00067
as.factor(Type.of.Breast.Surgery)Mastectomy        8.4e-05
Log(scale)                                      < 2e-16

Scale= 0.398

Weibull distribution
Loglik(model)= -3506.6   Loglik(intercept only)= -3514.4
        Chisq= 15.67 on 2 degrees of freedom, p= 4e-04
Number of Newton-Raphson Iterations: 10
n= 1310
```

```
pred.Type.of.Breast.Surgery1 = predict(sWei, newdata=list(Type.of.Breast.Surgery="Breast
Conserving"),type="quantile",p=seq(.01,.99,by=.01))
pred.Type.of.Breast.Surgery2 = predict(sWei,
newdata=list(Type.of.Breast.Surgery="Mastectomy"),type="quantile",p=seq(.01,.99,by=.01))

df = data.frame(y=seq(.99,.01,by=-.01),
Type.of.Breast.Surgery1=pred.Type.of.Breast.Surgery1,
Type.of.Breast.Surgery2=pred.Type.of.Breast.Surgery2)
df_long = gather(df, key= "Type.of.Breast.Surgery", value="time", -y)

p = ggsurvplot(fKM, data = Clean_data, risk.table = T)
p$plot = p$plot + geom_line(data=df_long, aes(x=time, y=y, group=Type.of.Breast.Surgery))
p$plot
```




```
p$table
```


Number at risk
Strata
Type.of.Breast.Surgery=
Type.of.Breast.Surgery=Breast Conserving
Type.of.Breast.Surgery=Mastectomy
13 3 0 0 0
554 354 135 3 0
743 397 145 6 0
0 100 200 300 400
Time