5231 pj V1

```
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(survival)
library(survminer)
```

```
Loading required package: ggplot2

Loading required package: ggpubr


Attaching package: 'survminer'

The following object is masked from 'package:survival':

    myeloma
```

```
library(ggplot2)
library(corrplot)
```

```
corrplot 0.95 loaded
```

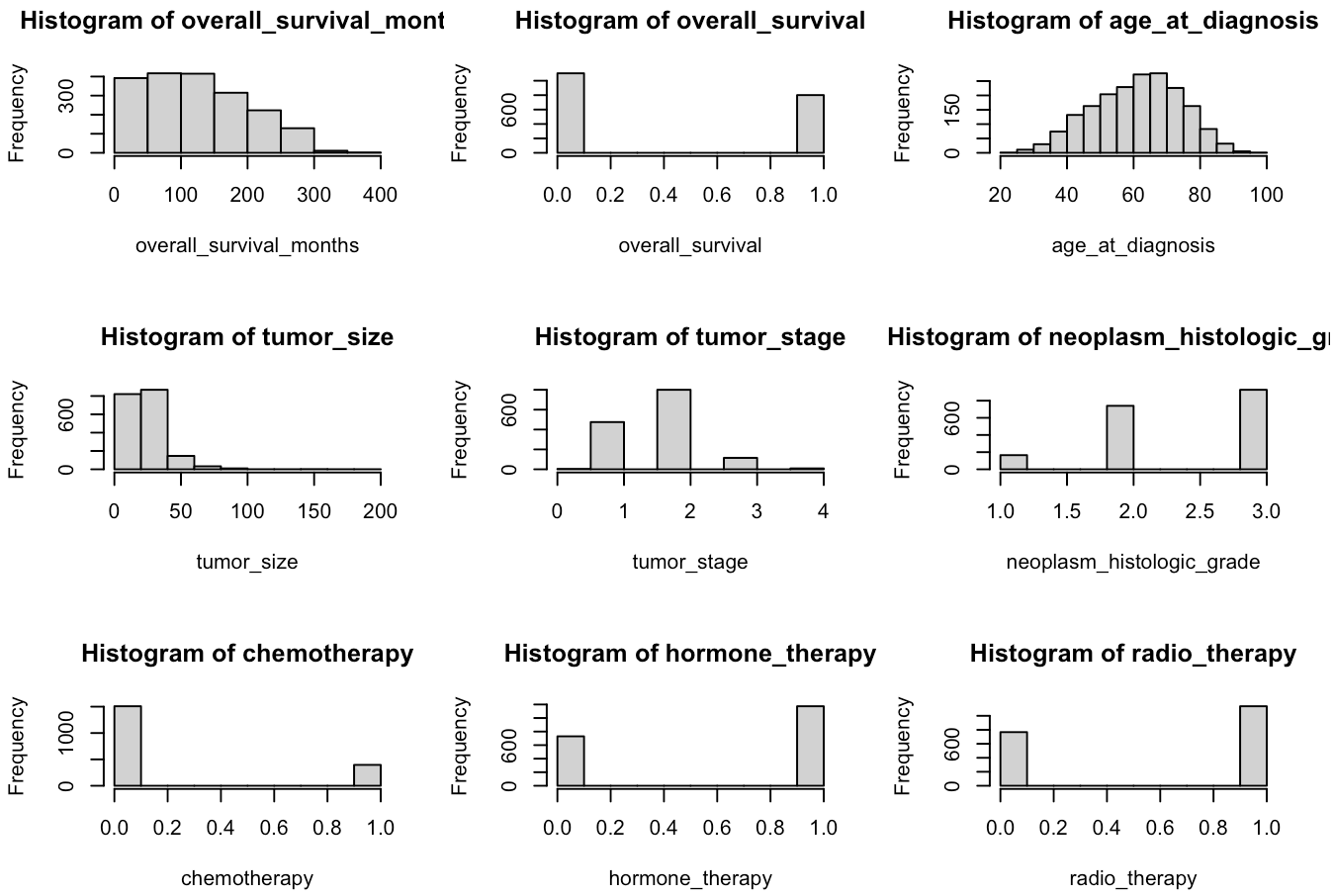
```
library(glmnet)
```

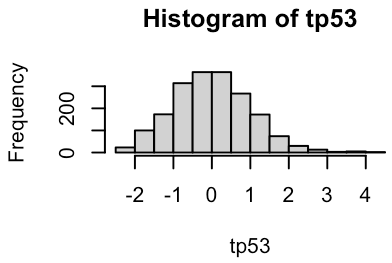
```
Loading required package: Matrix

Loaded glmnet 4.1-8
```

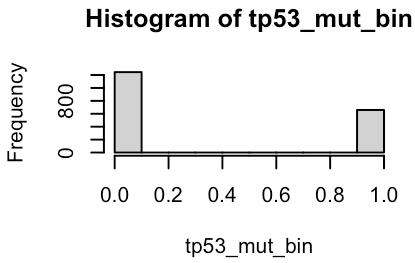
```
df <- read.csv('~/Downloads/METABRIC_RNA_Mutation.csv')
```

```
#data selection: we choose clinical data and the gene we interested:tp53
#tp53: This gene is consider to have a tumor suppressor effect
df$tp53_mut_bin <- ifelse(df$tp53_mut=='0', 0, 1)
df_model<-df %>% select(overall_survival_months,overall_survival,age_at_diagnosis,tumor_s
```

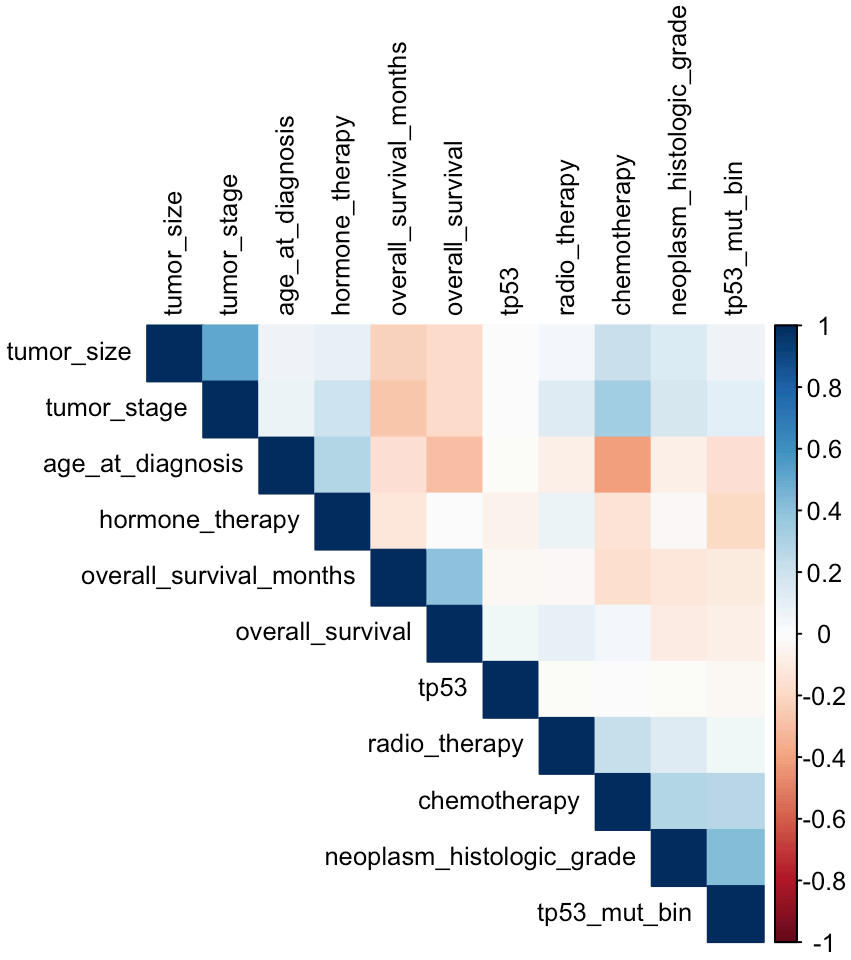
```r
#感觉hist不适合0，1的数据，你们有啥想法么
numeric_df <- df_model %>% select(where(is.numeric))
numeric_vars <- names(numeric_df)

par(mfrow = c(3, 3))

for (col in numeric_vars) {
  hist(df_model[[col]],
       main = paste("Histogram of", col),
       xlab = col)
}
```



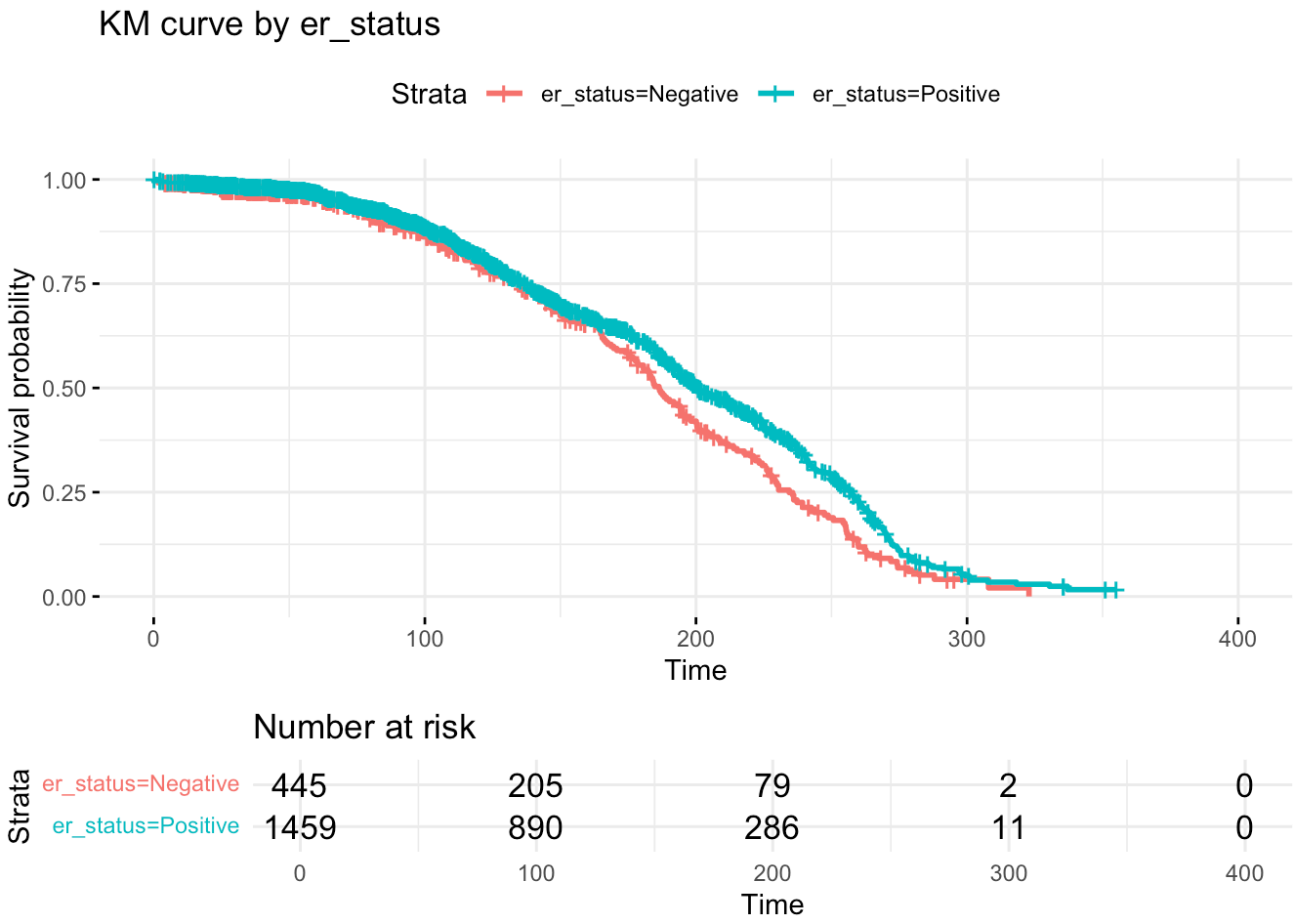


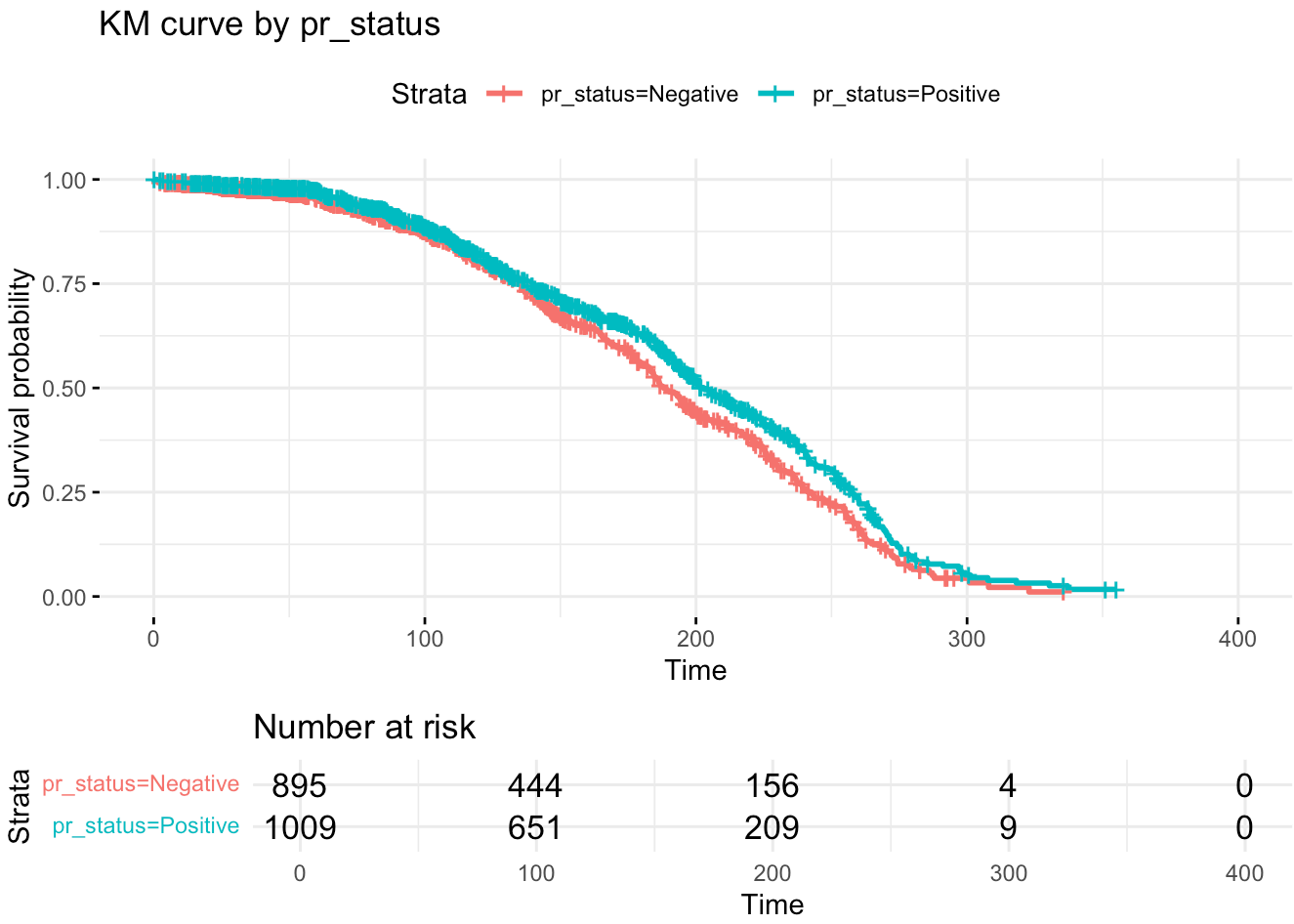


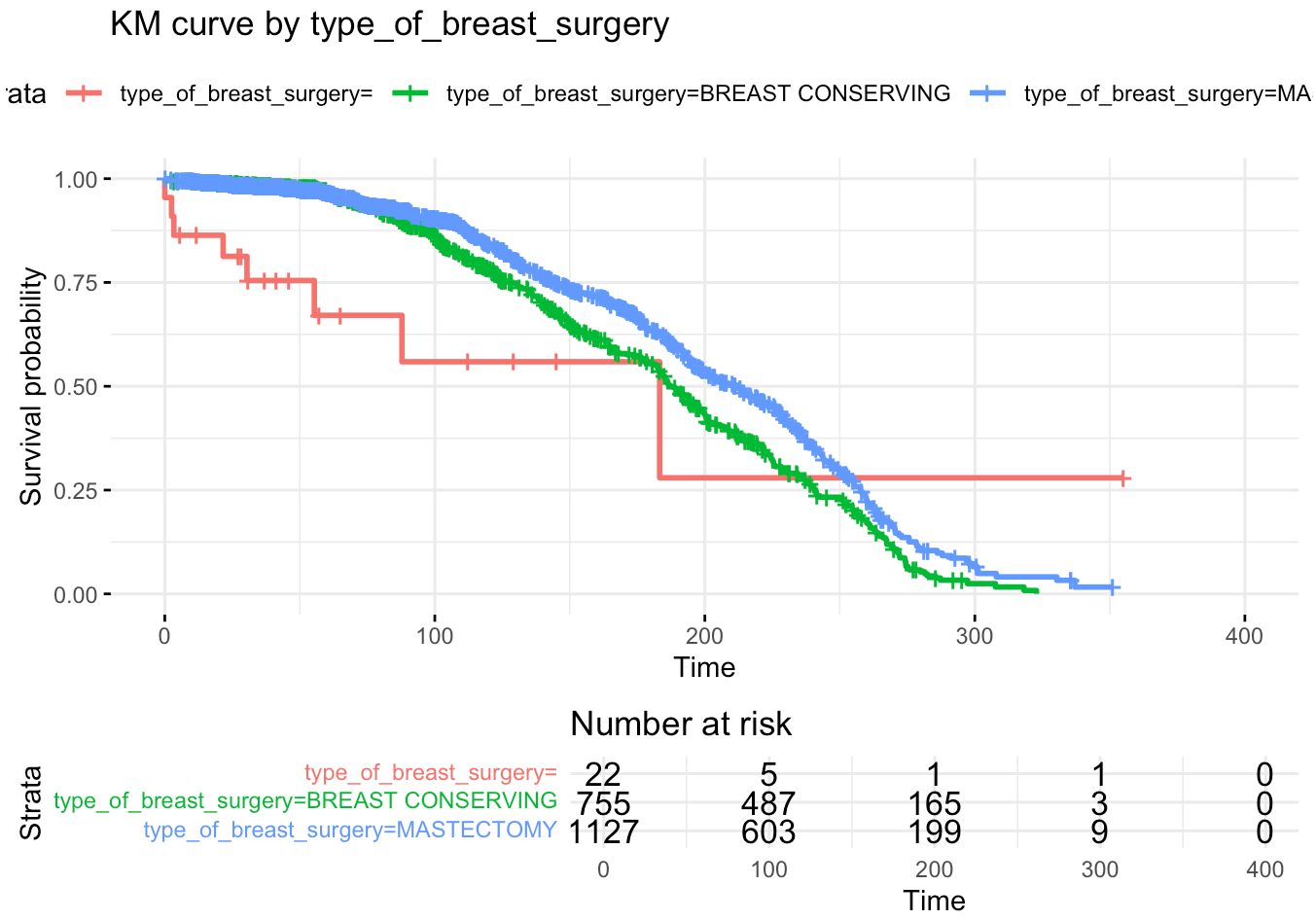

```r
cor_mat <- cor(numeric_df, use = "complete.obs")
corrplot(cor_mat, method = "color", type = "upper",
         tl.cex = 0.8, tl.col = "black", order = "hclust")
```

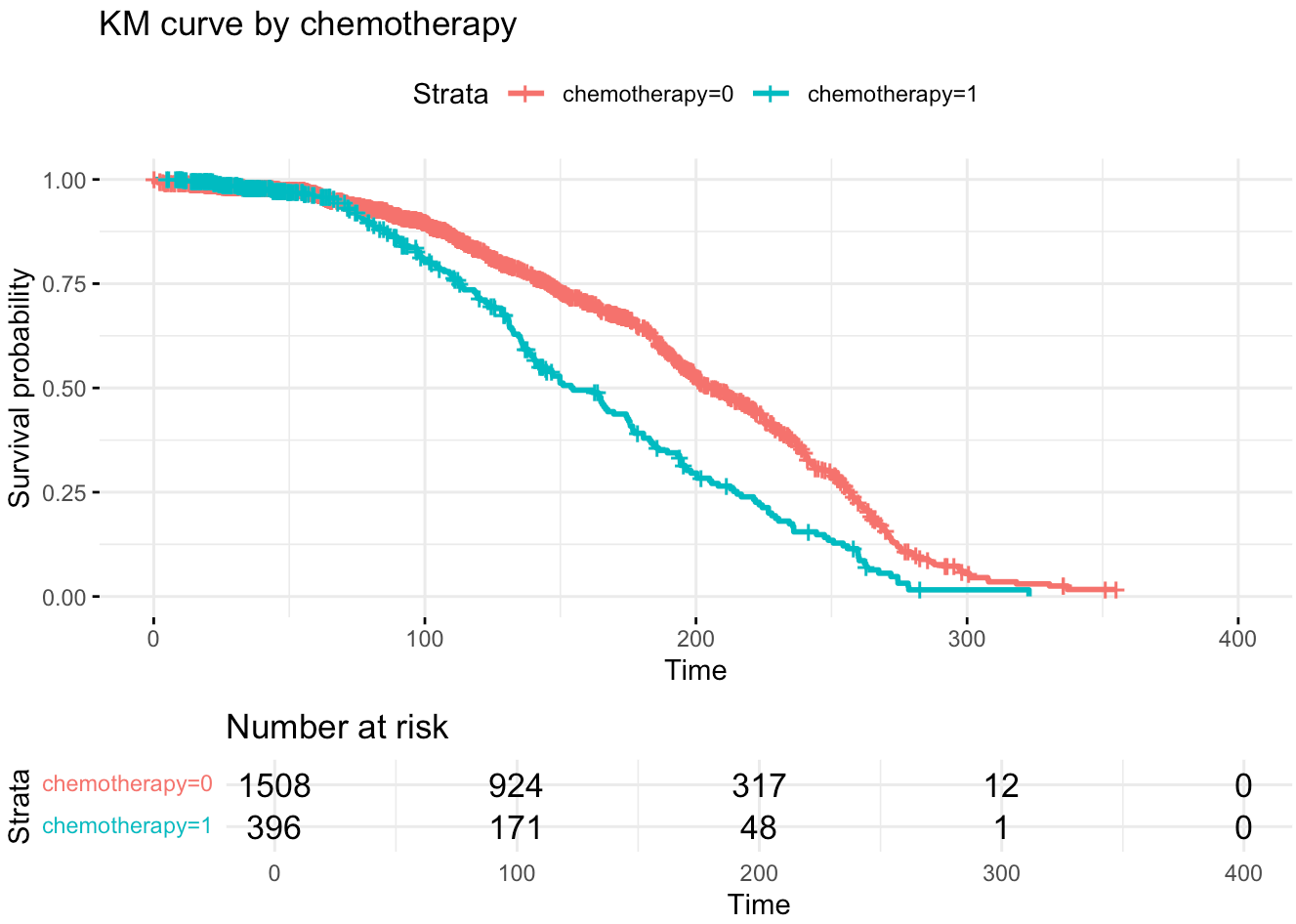


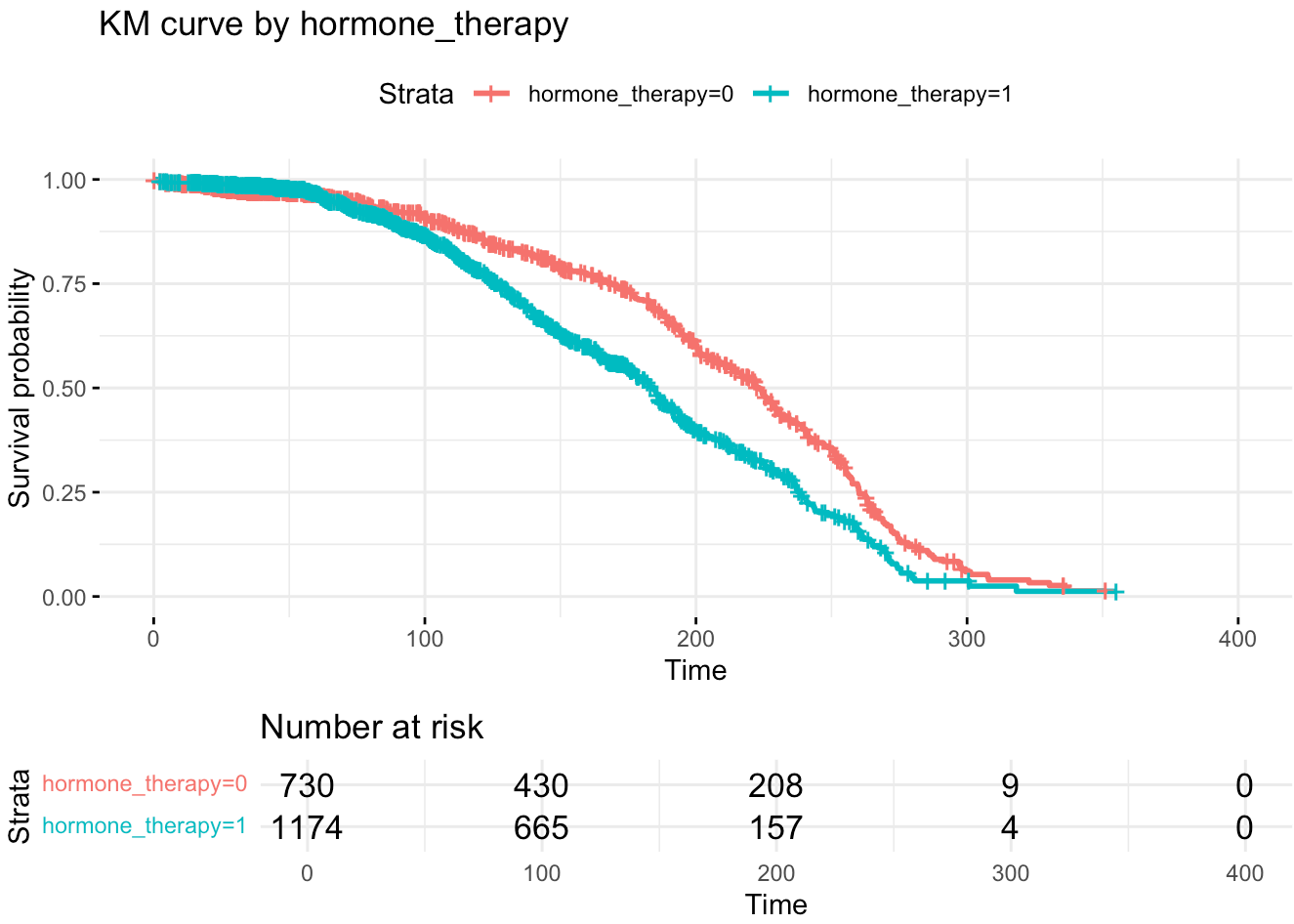

```r
#KM plot
#looking for group feature,I assume the unique length less than 3 is also group not numer
group_vars <- names(df_model)[sapply(df_model, function(x) {
  is.factor(x) || is.character(x) || (is.numeric(x) && length(unique(x)) <= 3)
})]
group_vars <- group_vars[group_vars != "overall_survival"]
for (var in group_vars) {
  formula <- as.formula(paste("Surv(overall_survival_months, overall_survival) ~", var))
  model <- survfit(formula, data = df_model)

  model$call <- list(formula = formula)

  print(ggsurvplot(model, data = df_model,
           risk.table = TRUE,
           title = paste("KM curve by", var),
           ggtheme = theme_minimal()))
}
```

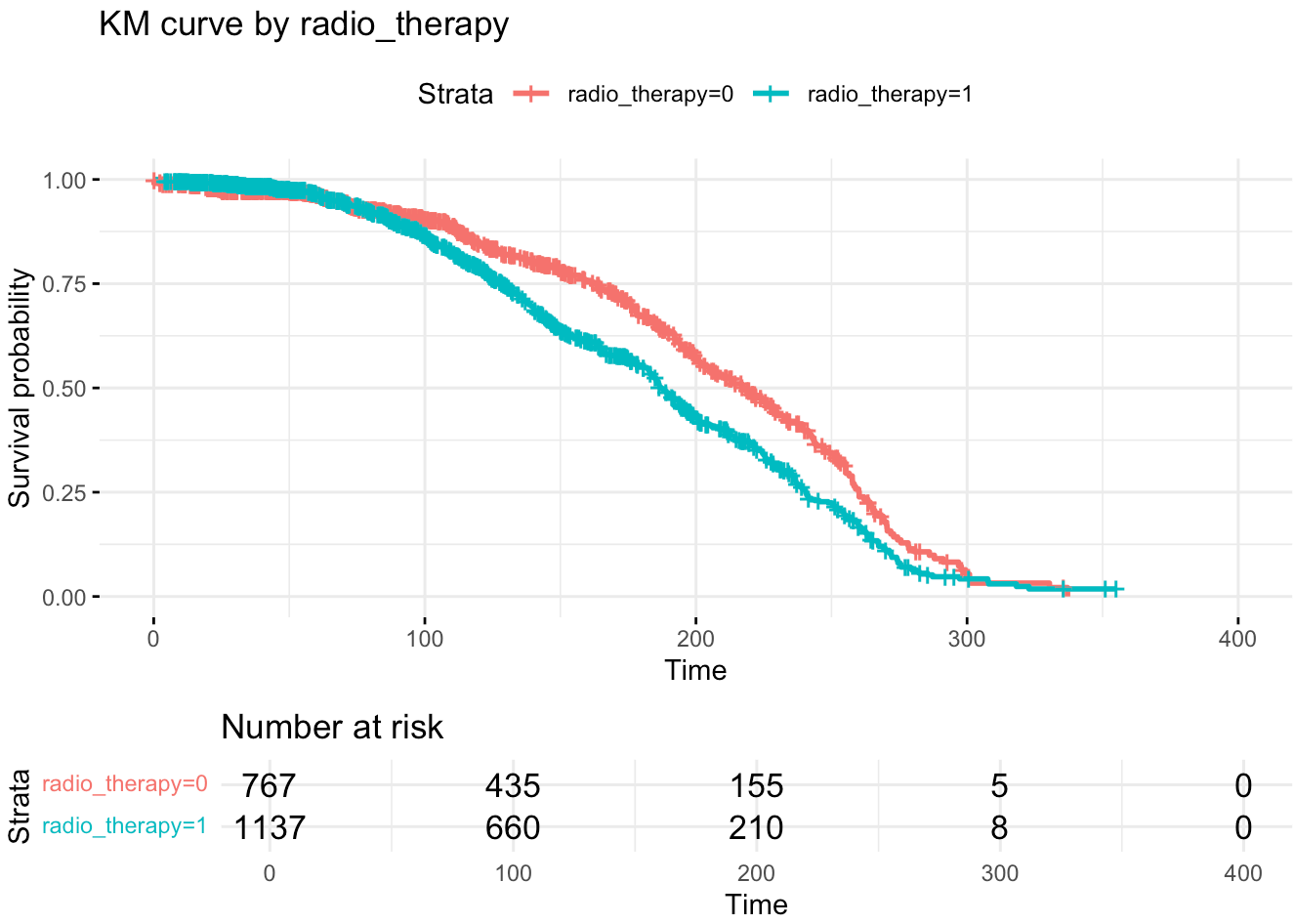

KM curve by er_status
Strata
er_status=Negative
er_status=Positive
1.00
0.75
0.50
0.25
0.00
Survival probability
0
100
200
300
400
Time
Number at risk
Strata
er_status=Negative
er_status=Positive
445
205
79
2
0
1459
890
286
11
0
0
100
200
300
400
Time

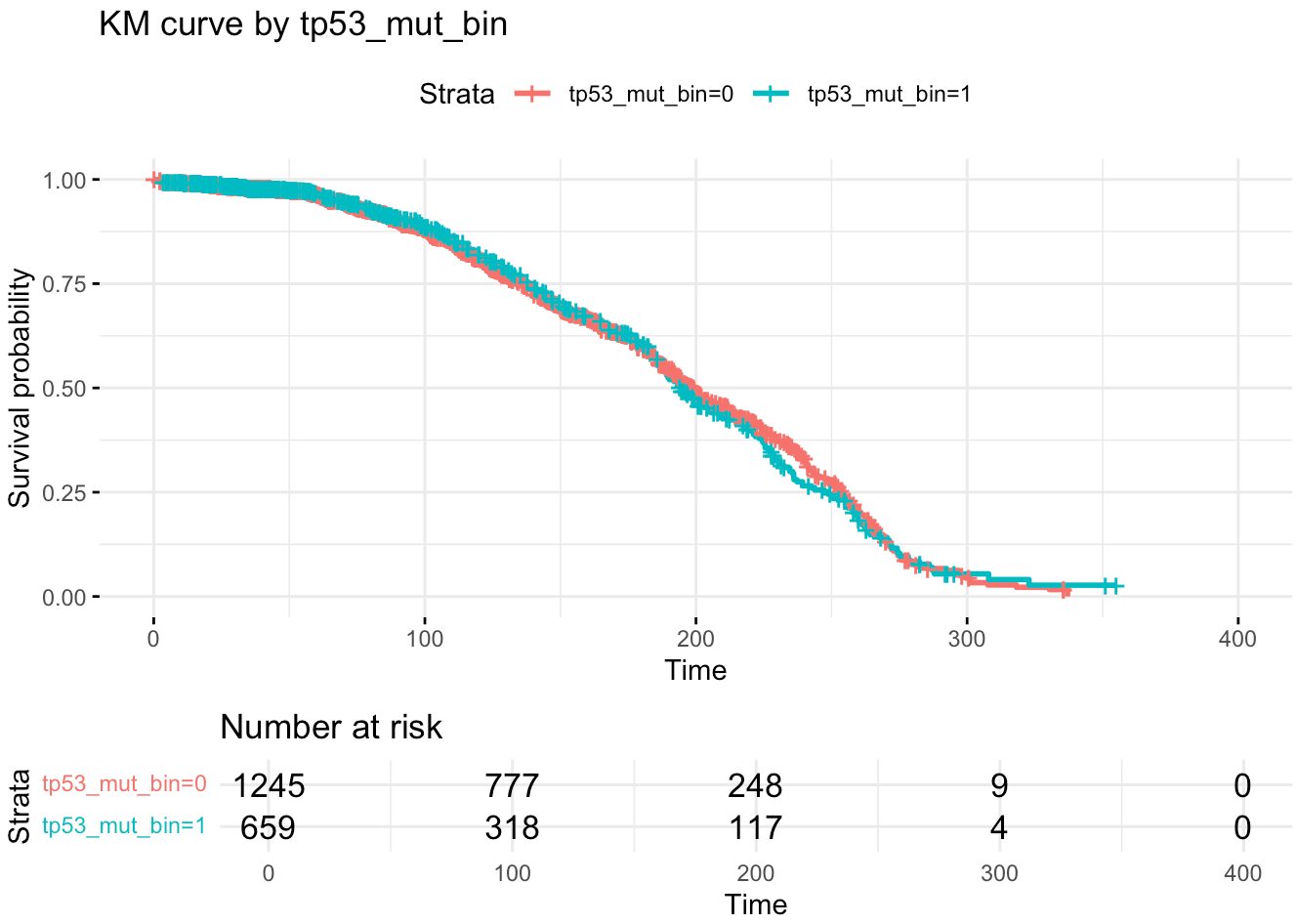

KM curve by pr_status
Strata
pr_status=Negative
pr_status=Positive
Survival probability
1.00
0.75
0.50
0.25
0.00
0
100
200
300
400
Time
Number at risk
Strata
pr_status=Negative
pr_status=Positive
895
444
156
4
0
1009
651
209
9
0
0
100
200
300
400
Time


KM curve by type_of_breast_surgery
ata
type_of_breast_surgery=
type_of_breast_surgery=BREAST CONSERVING
type_of_breast_surgery=MA
Survival probability
1.00
0.75
0.50
0.25
0.00
0
100
200
300
400
Time
Number at risk
Strata
type_of_breast_surgery=
type_of_breast_surgery=BREAST CONSERVING
type_of_breast_surgery=MASTECTOMY
22 5 1 1 0
755 487 165 3 0
1127 603 199 9 0
0
100
200
300
400
Time

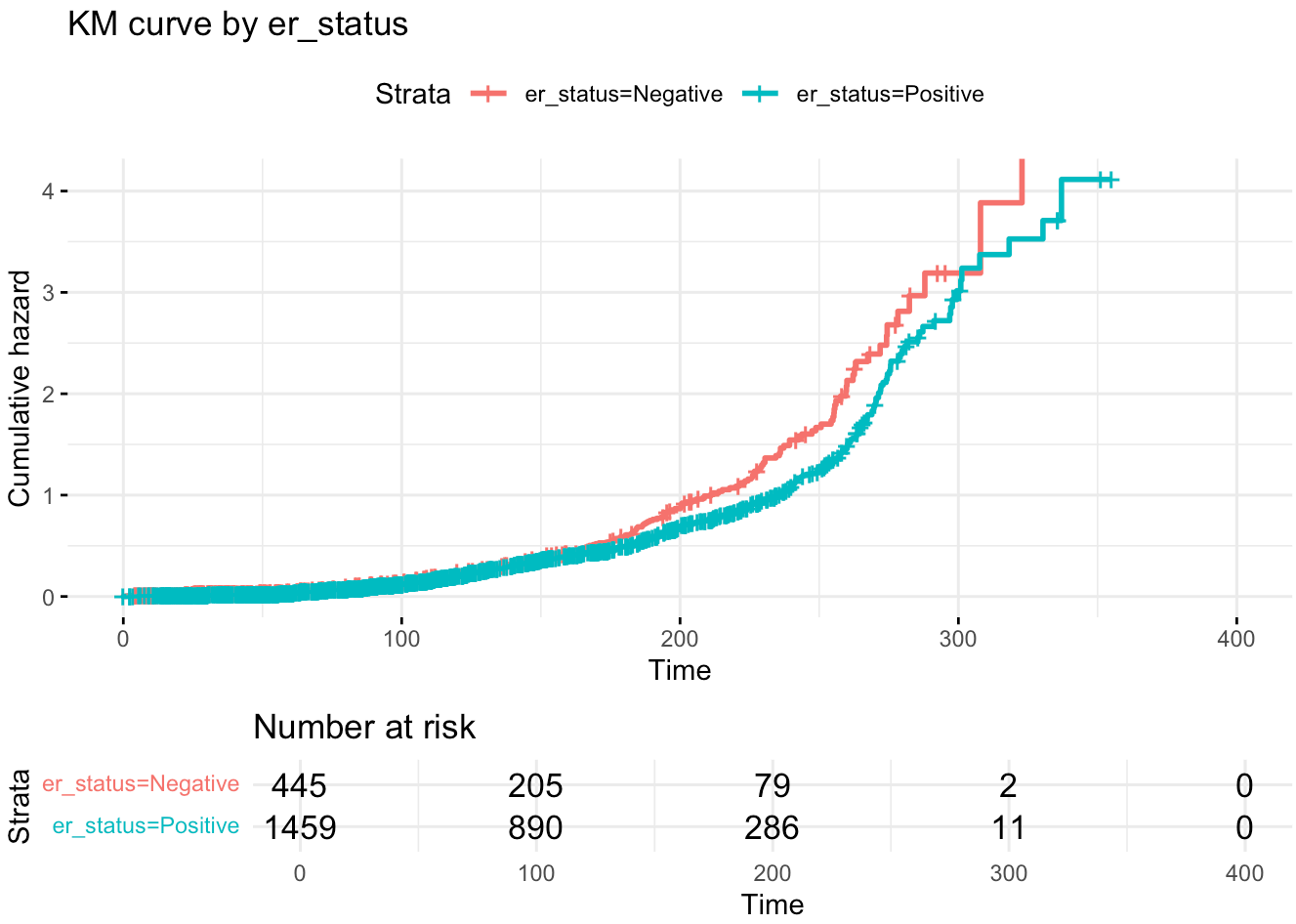

KM curve by chemotherapy
Strata
chemotherapy=0
chemotherapy=1
Survival probability
1.00
0.75
0.50
0.25
0.00
0
100
200
300
400
Time
Number at risk
Strata
chemotherapy=0
chemotherapy=1
1508
924
317
12
0
396
171
48
1
0
0
100
200
300
400
Time

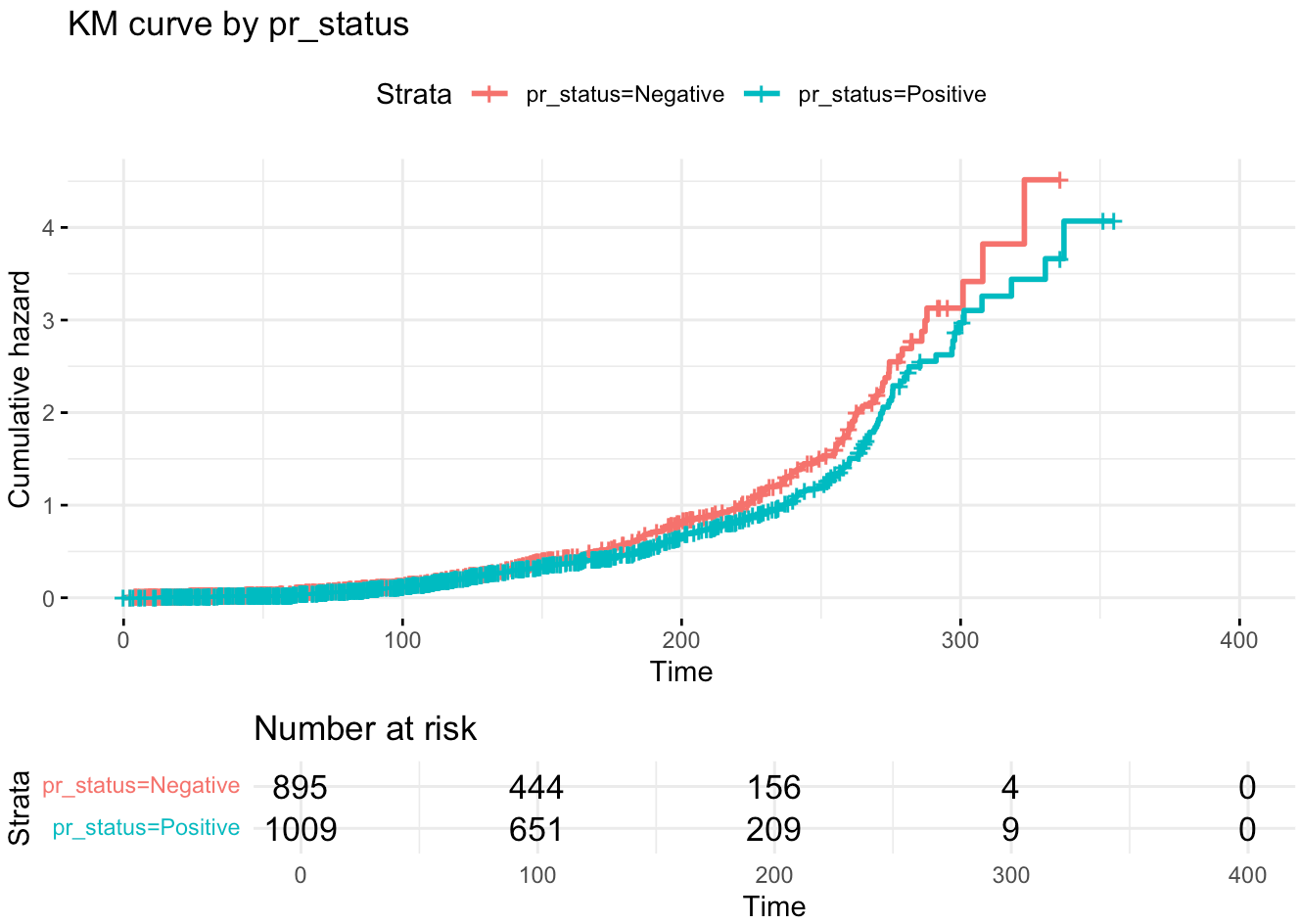

KM curve by hormone_therapy
Strata
hormone_therapy=0
hormone_therapy=1
Survival probability
1.00
0.75
0.50
0.25
0.00
0
100
200
300
400
Time
Number at risk
Strata
hormone_therapy=0
hormone_therapy=1
730
430
208
9
0
1174
665
157
4
0
0
100
200
300
400
Time

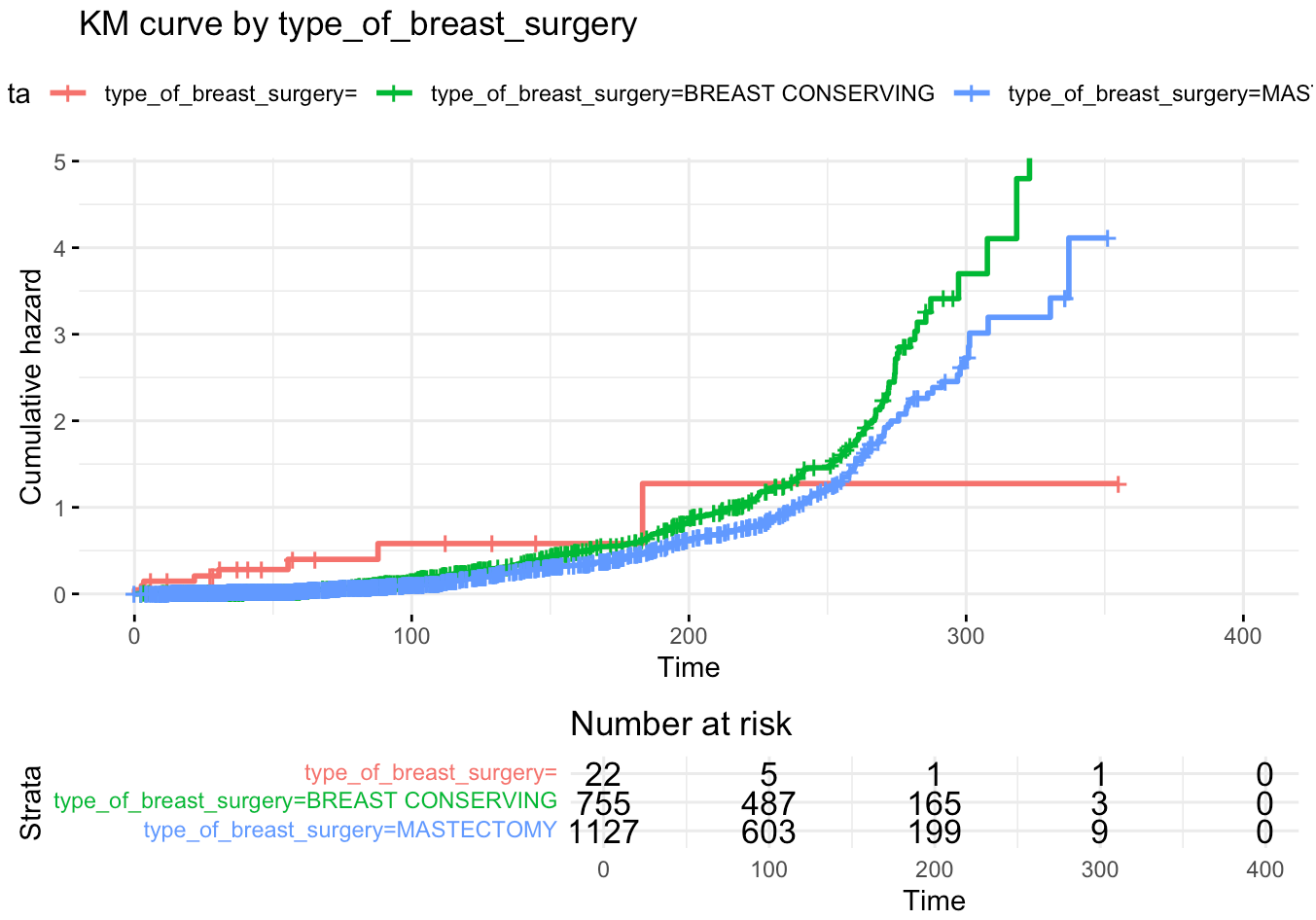

KM curve by radio_therapy
Strata
radio_therapy=0
radio_therapy=1
Survival probability
1.00
0.75
0.50
0.25
0.00
0
100
200
300
400
Time
Number at risk
Strata
radio_therapy=0
radio_therapy=1
767
435
155
5
0
1137
660
210
8
0
0
100
200
300
400
Time

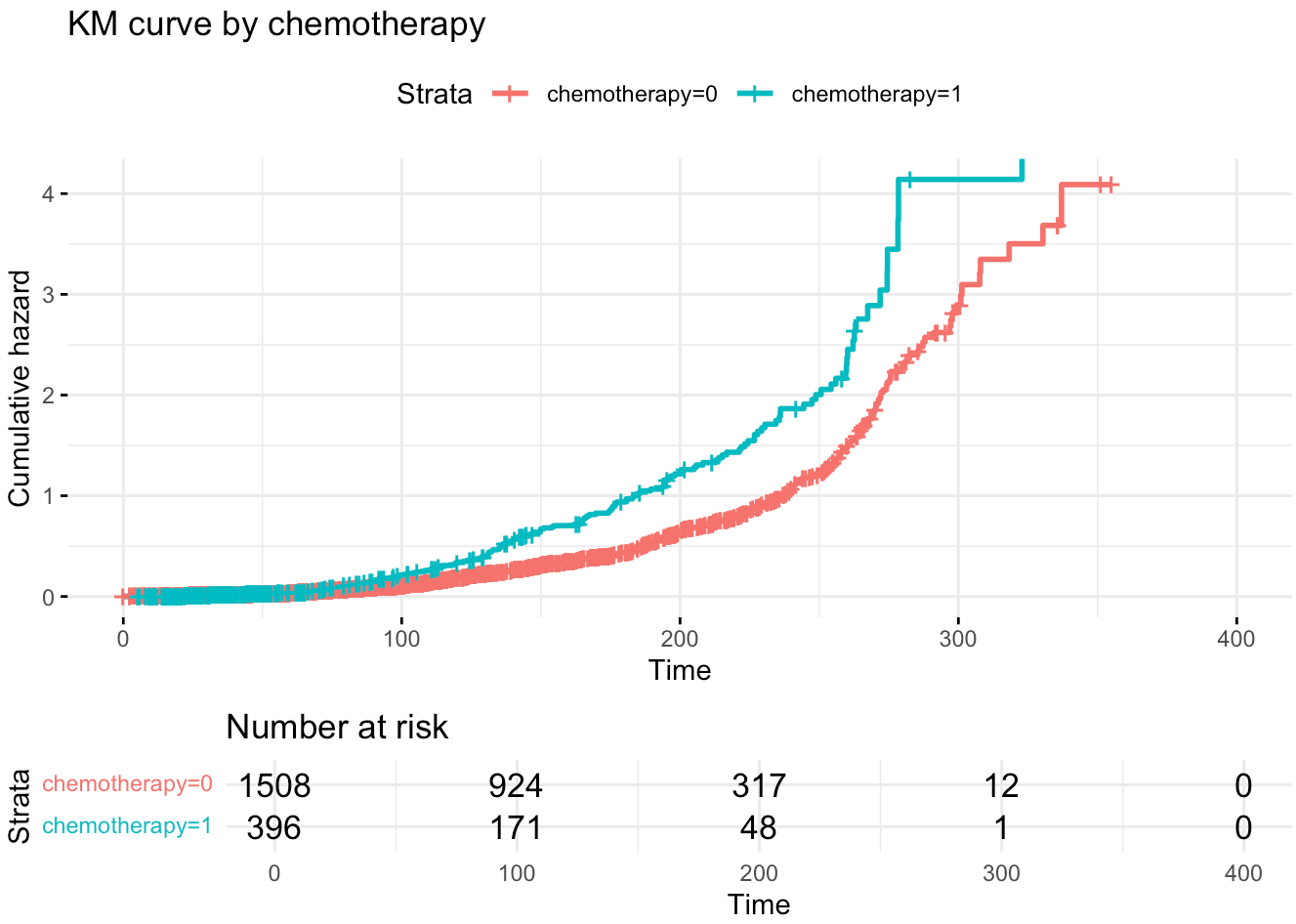


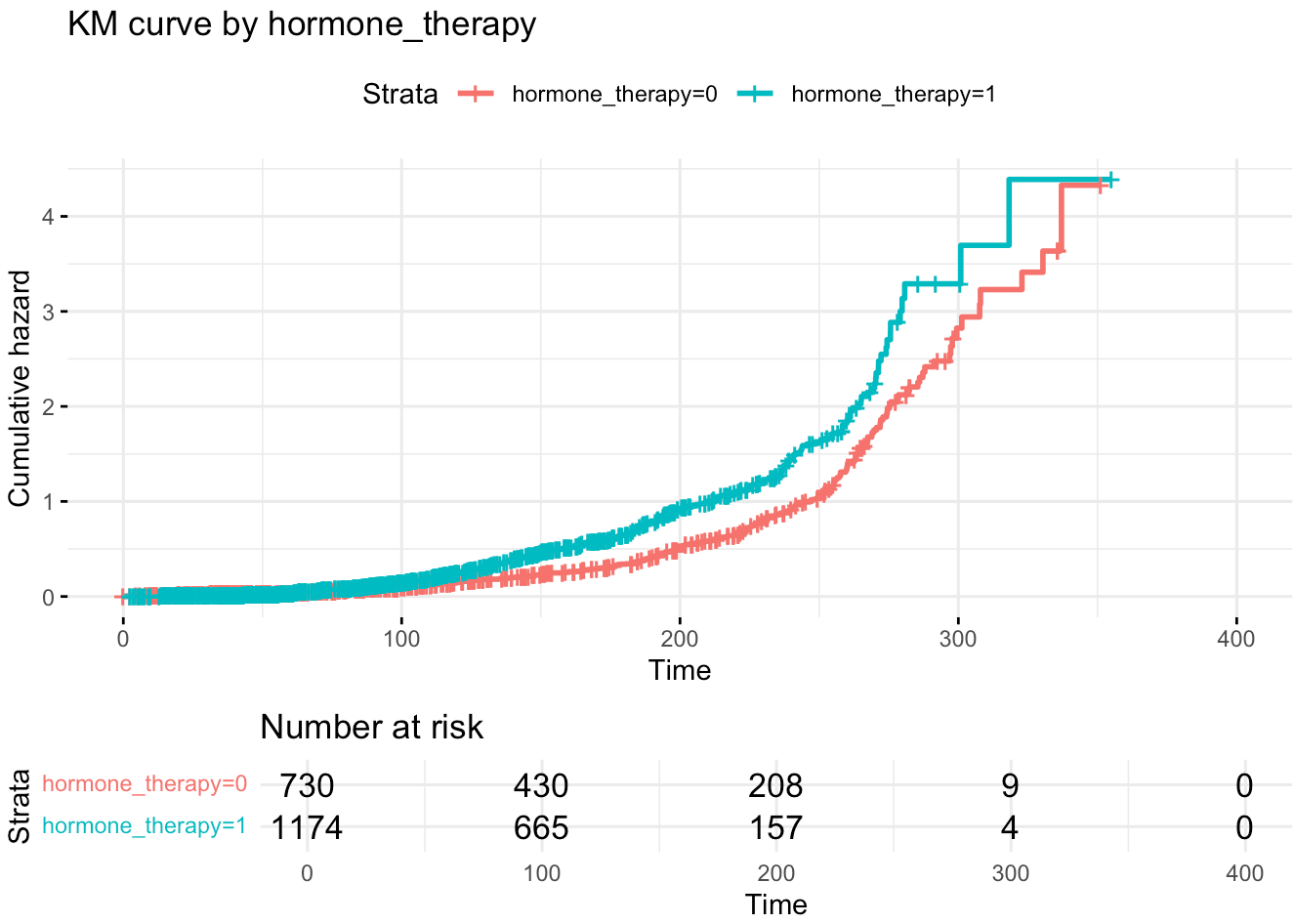

```r
#Cumulative Hazard Plot
for (var in group_vars) {
  formula <- as.formula(paste("Surv(overall_survival_months, overall_survival) ~", var))
  model <- survfit(formula, data = df_model)

  model$call <- list(formula = formula)

  print(ggsurvplot(model,fun ="cumhaz", data = df_model,
           risk.table = TRUE,
           title = paste("KM curve by", var),
           ggtheme = theme_minimal()))
}
```

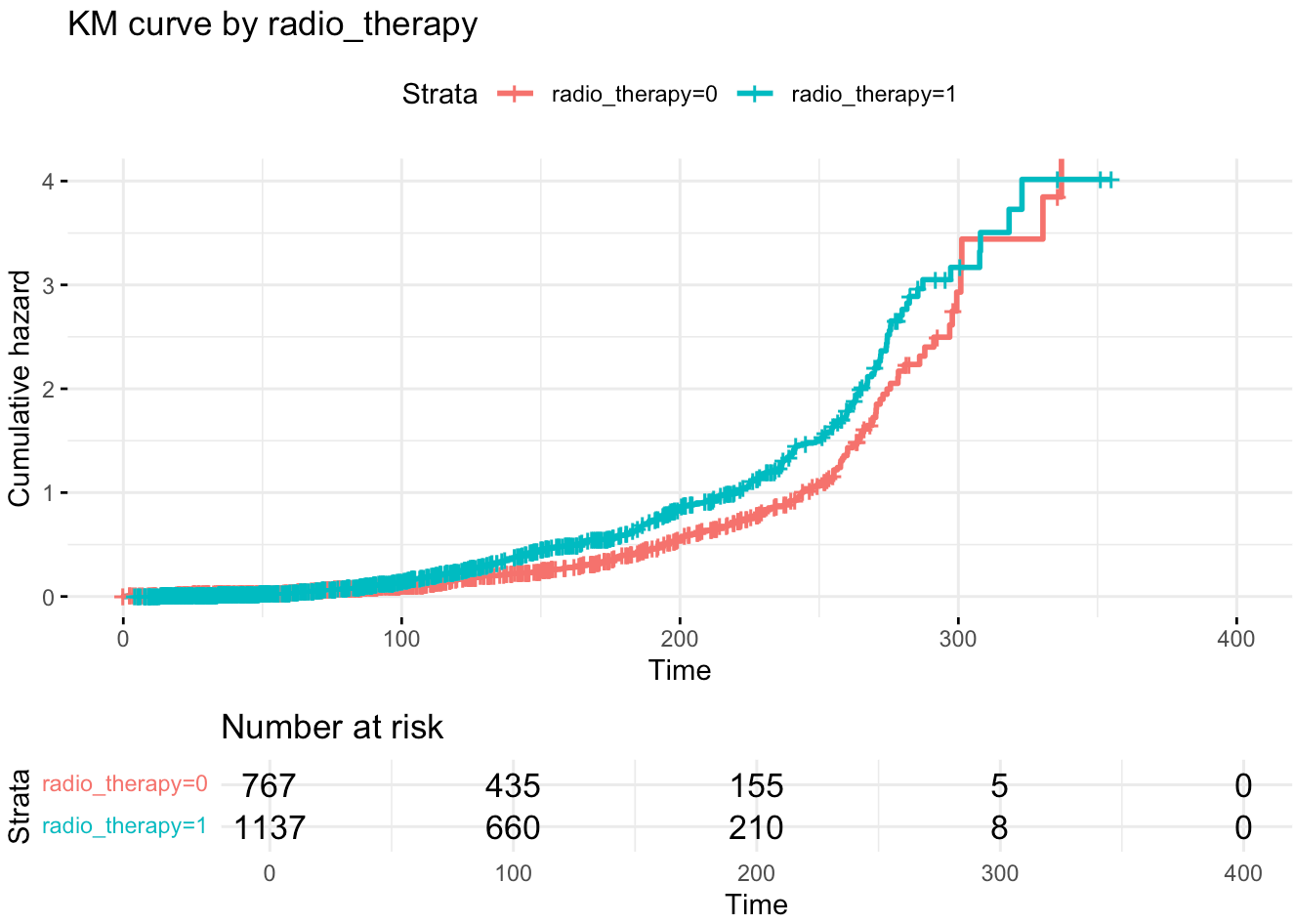

KM curve by er_status
Strata
er_status=Negative
er_status=Positive
Cumulative hazard
0
1
2
3
4
0
100
200
300
400
Time
Number at risk
Strata
er_status=Negative
er_status=Positive
445
205
79
2
0
1459
890
286
11
0
0
100
200
300
400
Time

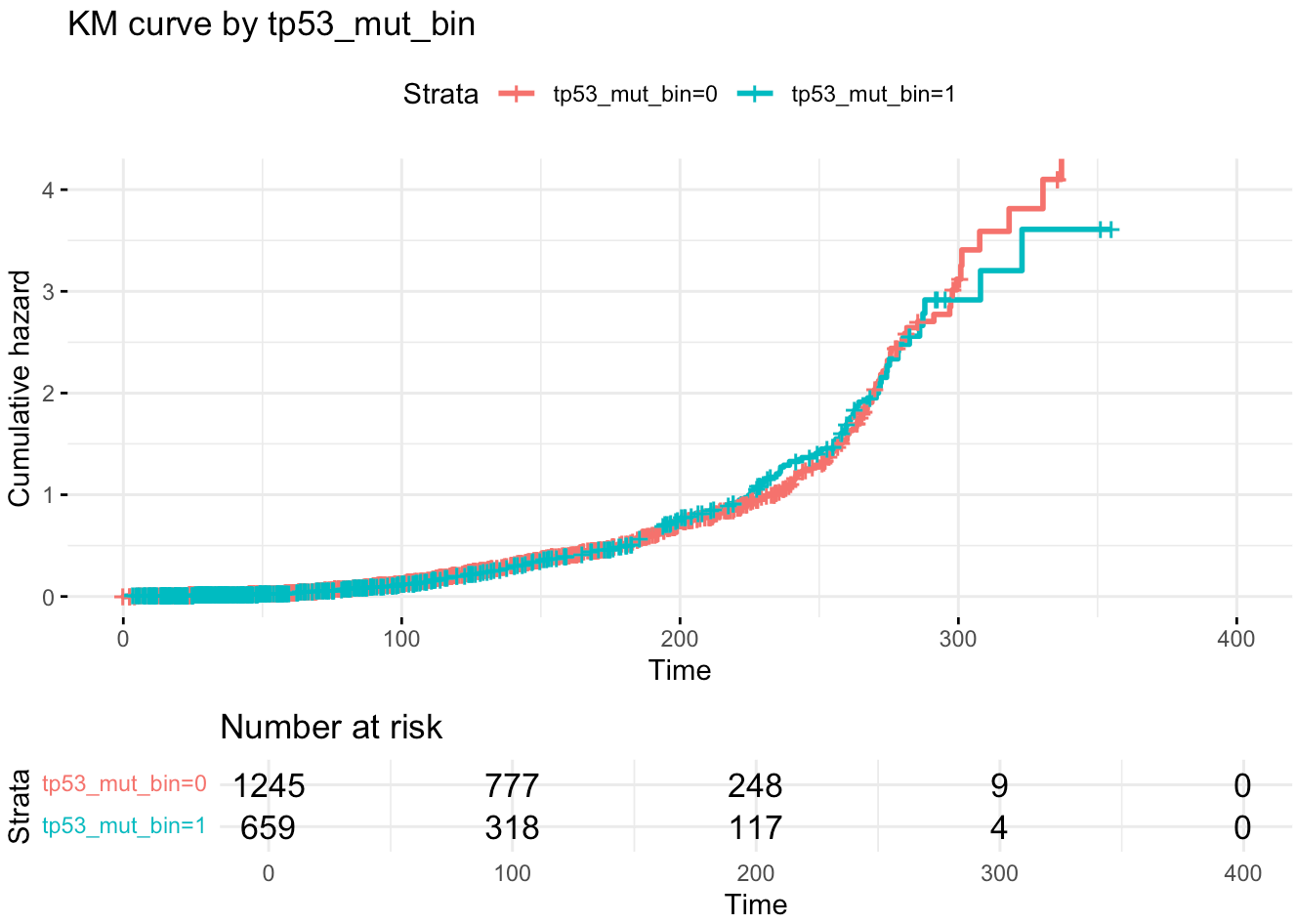

KM curve by pr_status
Strata
pr_status=Negative
pr_status=Positive
Cumulative hazard
0
1
2
3
4
0
100
200
300
400
Time
Number at risk
Strata
pr_status=Negative
pr_status=Positive
895
444
156
4
0
1009
651
209
9
0
0
100
200
300
400
Time


KM curve by type_of_breast_surgery
ta
type_of_breast_surgery=
type_of_breast_surgery=BREAST CONSERVING
type_of_breast_surgery=MAS
Cumulative hazard
0
1
2
3
4
5
0
100
200
300
400
Time
Number at risk
Strata
type_of_breast_surgery= 22 5 1 1 0
type_of_breast_surgery=BREAST CONSERVING 755 487 165 3 0
type_of_breast_surgery=MASTECTOMY 1127 603 199 9 0
0
100
200
300
400
Time


KM curve by chemotherapy
Strata
chemotherapy=0
chemotherapy=1
Cumulative hazard
0
1
2
3
4
0
100
200
300
400
Time
Number at risk
Strata
chemotherapy=0
chemotherapy=1
1508
924
317
12
0
396
171
48
1
0
0
100
200
300
400
Time


KM curve by hormone_therapy
Strata
hormone_therapy=0
hormone_therapy=1
Cumulative hazard
0
1
2
3
4
0
100
200
300
400
Time
Number at risk
Strata
hormone_therapy=0
hormone_therapy=1
730
430
208
9
0
1174
665
157
4
0
0
100
200
300
400
Time


KM curve by radio_therapy
Strata
radio_therapy=0
radio_therapy=1
Cumulative hazard
0
1
2
3
4
0
100
200
300
400
Time
Number at risk
Strata
radio_therapy=0
radio_therapy=1
767
435
155
5
0
1137
660
210
8
0
0
100
200
300
400
Time




```r
get_cox <- function(var, df){
  formula <- as.formula(paste("Surv(overall_survival_months, overall_survival) ~", var))
  model <- coxph(formula, data = df_model)
  var1 <- rownames(summary(model)$coefficients)
  p_value <- summary(model)$coefficients[,"Pr(>|z|)"]
  hr <- summary(model)$coefficients[, "exp(coef)"]
  return(list(variable = var1,
              p_value = p_value,
              harzed_ratio = hr))
}

cox_var <- c(
  "age_at_diagnosis", "tumor_size", "tumor_stage", "neoplasm_histologic_grade",
  "er_status", "pr_status", "type_of_breast_surgery", "chemotherapy",
  "hormone_therapy", "radio_therapy", "tp53", "tp53_mut_bin"
)
cox_result <- lapply(cox_var, get_cox, df = df_model)
cox_result_df <- do.call(rbind, lapply(cox_result, as.data.frame))
```

```r
cox_result_df
```

	variable	p_value	harzed_ratio
1	age_at_diagnosis	6.226937e-04	0.9897605
2	tumor_size	9.513677e-01	0.9998237
3	tumor_stage	1.205711e-01	1.1141142
4	neoplasm_histologic_grade	7.374502e-01	0.9819642
5	er_statusPositive	4.315518e-03	0.7903588
6	pr_statusPositive	1.090340e-02	0.8345292
type_of_breast_surgeryBREAST CONSERVING	type_of_breast_surgeryBREAST CONSERVING	6.705815e-01	0.8553684
type_of_breast_surgeryMASTECTOMY	type_of_breast_surgeryMASTECTOMY	2.284567e-01	0.6431166
11	chemotherapy	1.656809e-11	1.7640230
12	hormone_therapy	4.336906e-09	1.5438562
13	radio_therapy	1.519481e-05	1.3816083
14	tp53	2.091148e-05	1.1611382
15	tp53_mut_bin	8.329323e-01	1.0161538

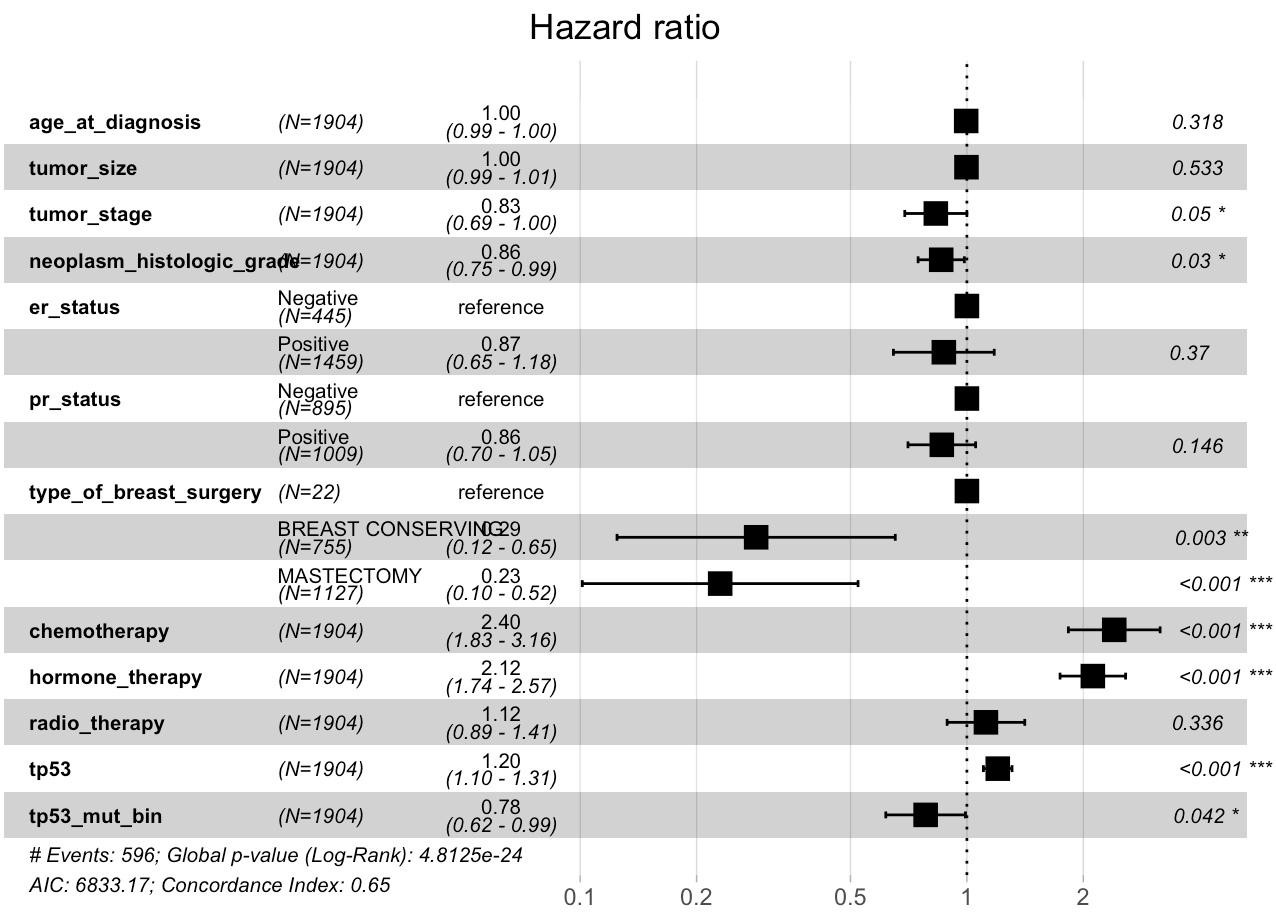
```
cox_fit_total <- coxph(Surv(overall_survival_months,overall_survival) ~ ., data = df_mode
summary(cox_fit_total)
```

```
Call:
coxph(formula = Surv(overall_survival_months, overall_survival) ~
    ., data = df_model)

  n= 1354, number of events= 596
   (550 observations deleted due to missingness)

                                          coef exp(coef)  se(coef)      z
age_at_diagnosis                      -0.004186  0.995822  0.004188 -1.000
tumor_size                            -0.002483  0.997520  0.003980 -0.624
tumor_stage                           -0.185139  0.830989  0.094337 -1.963
neoplasm_histologic_grade             -0.152800  0.858302  0.070276 -2.174
er_statusPositive                     -0.137270  0.871735  0.153140 -0.896
pr_statusPositive                     -0.149445  0.861186  0.102767 -1.454
type_of_breast_surgeryBREAST CONSERVING -1.254714  0.285157  0.422680 -2.968
type_of_breast_surgeryMASTECTOMY      -1.468909  0.230177  0.418861 -3.507
chemotherapy                           0.877482  2.404836  0.139354  6.297
hormone_therapy                        0.749782  2.116538  0.099544  7.532
```

```
radio_therapy                             0.113396  1.120076  0.117912  0.962
tp53                                      0.183670  1.201620  0.043447  4.227
tp53_mut_bin                             -0.245785  0.782090  0.121011 -2.031
                                         Pr(>|z|)
age_at_diagnosis                         0.317528
tumor_size                               0.532611
tumor_stage                              0.049701 *
neoplasm_histologic_grade                0.029685 *
er_statusPositive                        0.370056
pr_statusPositive                        0.145887
type_of_breast_surgeryBREAST CONSERVING  0.002993 **
type_of_breast_surgeryMASTECTOMY         0.000453 ***
chemotherapy                             3.04e-10 ***
hormone_therapy                          4.99e-14 ***
radio_therapy                            0.336198
tp53                                     2.36e-05 ***
tp53_mut_bin                             0.042245 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

                                         exp(coef) exp(-coef) lower .95
age_at_diagnosis                            0.9958     1.0042    0.9877
tumor_size                                  0.9975     1.0025    0.9898
tumor_stage                                 0.8310     1.2034    0.6907
neoplasm_histologic_grade                   0.8583     1.1651    0.7479
er_statusPositive                           0.8717     1.1471    0.6457
pr_statusPositive                           0.8612     1.1612    0.7041
type_of_breast_surgeryBREAST CONSERVING     0.2852     3.5068    0.1245
type_of_breast_surgeryMASTECTOMY            0.2302     4.3445    0.1013
chemotherapy                                2.4048     0.4158    1.8301
hormone_therapy                             2.1165     0.4725    1.7414
radio_therapy                               1.1201     0.8928    0.8890
tp53                                        1.2016     0.8322    1.1035
tp53_mut_bin                                0.7821     1.2786    0.6170
                                         upper .95
age_at_diagnosis                            1.0040
tumor_size                                  1.0053
tumor_stage                                 0.9998
neoplasm_histologic_grade                   0.9851
er_statusPositive                           1.1769
pr_statusPositive                           1.0533
type_of_breast_surgeryBREAST CONSERVING     0.6529
type_of_breast_surgeryMASTECTOMY            0.5231
chemotherapy                                3.1601
hormone_therapy                             2.5725
radio_therapy                               1.4113
tp53                                        1.3084
tp53_mut_bin                                0.9914

Concordance= 0.652  (se = 0.013 )
Likelihood ratio test= 143.2  on 13 df,   p=<2e-16
```

```
Wald test            = 150.9  on 13 df,   p=<2e-16
Score (logrank) test = 156.5  on 13 df,   p=<2e-16
```

```
ggforest(cox_fit_total, data = df_model)
```




```
df_lasso <- df_model[df_model$overall_survival_months > 0, ]
df_lasso <- na.omit(df_lasso)
X <- model.matrix(Surv(overall_survival_months,overall_survival) ~ ., data = df_lasso)[,
y <- Surv(df_lasso$overall_survival_months,df_lasso$overall_survival)
cox_lasso <- cv.glmnet(X, y, family = "cox", alpha = 1)
selected_vars <- rownames(coef(cox_lasso))[
  as.vector(coef(cox_lasso, s = "lambda.min")) != 0
]
selected_vars
```

```
[1] "age_at_diagnosis"
[2] "tumor_size"
[3] "tumor_stage"
[4] "neoplasm_histologic_grade"
[5] "er_statusPositive"
[6] "pr_statusPositive"
[7] "type_of_breast_surgeryBREAST CONSERVING"
[8] "type_of_breast_surgeryMASTECTOMY"
```

```
 [9] "chemotherapy"
[10] "hormone_therapy"
[11] "radio_therapy"
[12] "tp53"
[13] "tp53_mut_bin"
```