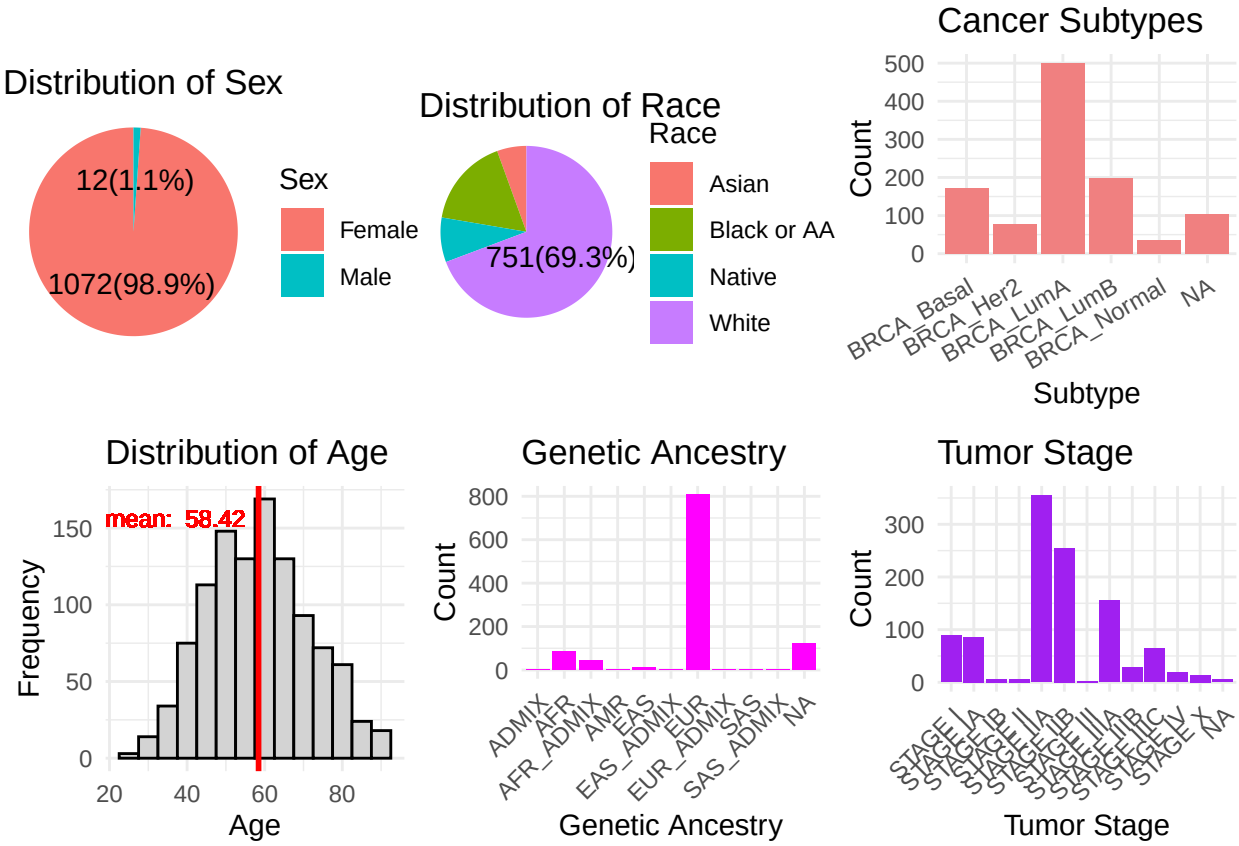
```r
data <- read.table("data_clinical_patient.txt", sep = "\t", header = TRUE)
data.2 <- read.table("brca_tcga_pan_can_atlas_2018_clinical_data.tsv", sep = "\t", header = TRUE)

library(dplyr)

data[data == "" | data == " "] <- NA

data <- data %>%
  mutate(
    RACE = ifelse(is.na(RACE) | RACE == "American Indian or Alaska Native", "Native", RACE),
    RACE = ifelse(is.na(RACE) | RACE == "Black or African American", "Black or AA", RACE))
```

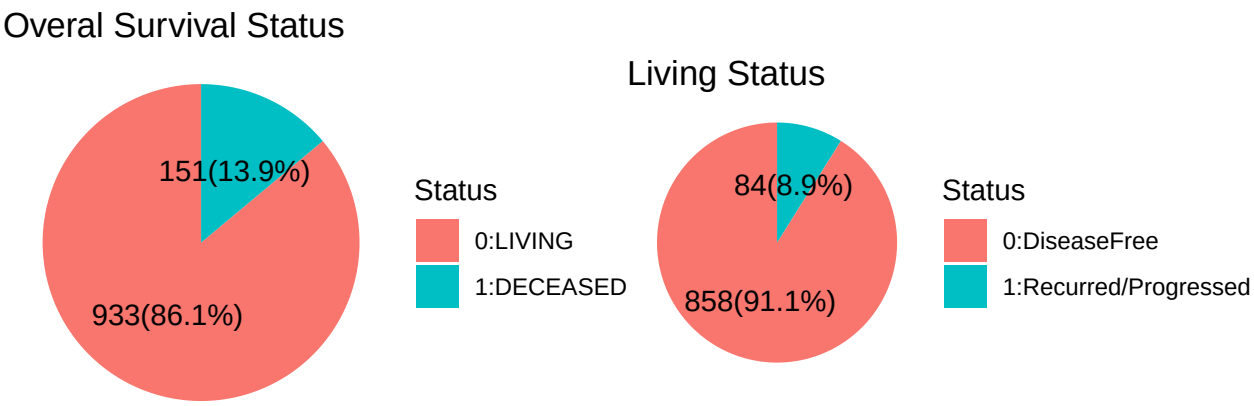
```r
library(ggplot2)
library(gridExtra)

# Plot 1: Pie plot for Sex
data_counts_sex <- data.frame(table(data$SEX))
colnames(data_counts_sex) <- c("SEX", "Count")
data_counts_sex$Percentage <- round((data_counts_sex$Count / sum(data_counts_sex$Count)) * 100, 1)

p1 <- ggplot(data_counts_sex, aes(x = "", y = Count, fill = SEX)) +
  geom_bar(stat = "identity", width = 1) +
  coord_polar(theta = "y") +
  labs(title = "Distribution of Sex", fill = "Sex") +
  theme_void() +  # Remove gridlines and axes
  geom_text(aes(label = paste(Count, "(", Percentage, "%)", sep = "")),
            position = position_stack(vjust = 0.5))

# Plot 2: Pie plot for Race
data_counts_race <- data.frame(table(data$RACE))
colnames(data_counts_race) <- c("RACE", "Count")
data_counts_race$Percentage <- round((data_counts_race$Count / sum(data_counts_race$Count)) * 100, 1)

p2 <- ggplot(data_counts_race, aes(x = "", y = Count, fill = RACE)) +
  geom_bar(stat = "identity", width = 1) +  # Create bars (used for pie chart)
  coord_polar(theta = "y") +  # Convert to pie chart
  labs(title = "Distribution of Race", fill = "Race") +
  theme_void() +  # Remove gridlines and axes
  geom_text(aes(label = ifelse(RACE %in% "White",
                               paste(Count, "(", Percentage, "%)", sep = ""),
                               "")),
            position = position_stack(vjust = 0.5))

# Plot 3: Bar plot for SUBTYPE
p3 <- ggplot(data, aes(x = SUBTYPE)) +
  geom_bar(fill = "lightcoral") +
  labs(title = "Cancer Subtypes", x = "Subtype", y = "Count") +
  theme_minimal()+
  theme(axis.text.x = element_text(angle = 30, hjust = 1))

# Plot 4: Histogram for AGE
p4 <- ggplot(data, aes(x = AGE)) +
  geom_histogram(binwidth = 5, fill = "lightgray", color = "black") +
```

```r
  geom_vline(aes(xintercept = mean(data$AGE)), color = "red", size = 1) +  # Mean line
  geom_text(aes(x = 37, y = 145, label = paste("mean: ", round(mean(data$AGE), 2))),
            color = "red", vjust = -0.5, size = 3) +  # Add text for the mean
  labs(title = "Distribution of Age", x = "Age", y = "Frequency") +
  theme_minimal()

# Plot 5: Barplot for Genetic Ancestry
p5 <- ggplot(data, aes(x = GENETIC_ANCESTRY_LABEL)) +
  geom_bar(fill = "magenta") +
  labs(title = "Genetic Ancestry", x = "Genetic Ancestry", y = "Count") +
  theme_minimal()+
  theme(axis.text.x = element_text(angle = 45, hjust = 1))

# Plot 6: Barplot for Tumor Stage
p6 <- ggplot(data, aes(x = AJCC_PATHOLOGIC_TUMOR_STAGE)) +
  geom_bar(fill = "purple") +
  labs(title = "Tumor Stage", x = "Tumor Stage", y = "Count") +
  theme_minimal()+
  theme(axis.text.x = element_text(angle = 45, hjust = 1))

# Arrange all plots in a 3x3 grid
grid.arrange(p1, p2, p3, p4, p5, p6, nrow = 2, ncol = 3)
```



Survival Analysis

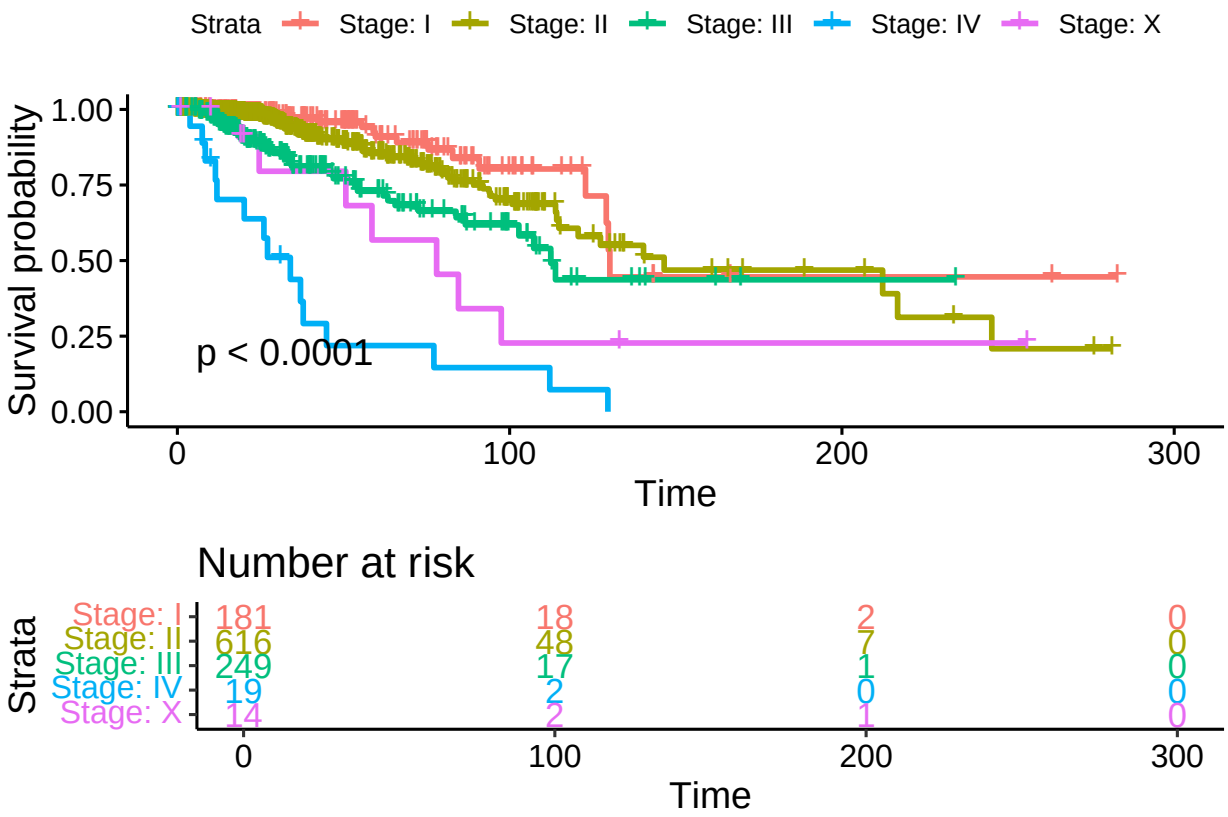
```
library("TCGAbiolinks")
library("survival")
library("survminer")
library("SummarizedExperiment")

data_counts_survival <- data.frame(table(data$OS_STATUS))
colnames(data_counts_survival) <- c("Status", "Count")
data_counts_survival$Percentage <- round((data_counts_survival$Count / sum(data_counts_survival$Count))

p7 <- ggplot(data_counts_survival, aes(x = "", y = Count, fill = Status)) +
  geom_bar(stat = "identity", width = 1) +  # Create bars (used for pie chart)
  coord_polar(theta = "y") +  # Convert to pie chart
  labs(title = "Overal Survival Status", fill = "Status") +
  theme_void() +  # Remove gridlines and axes
  geom_text(aes(label = paste(Count, "(", Percentage, "%)", sep = "")),
            position = position_stack(vjust = 0.5))


library(dbplyr)
data_counts_free <- data.frame(table(data$DFS_STATUS))
colnames(data_counts_free) <- c("Status", "Count")
data_counts_free$Percentage <- round((data_counts_free$Count / sum(data_counts_free$Count)) * 100, 1)

data_counts_free <- data_counts_free %>% filter(Status != "")

p8 <- ggplot(data_counts_free, aes(x = "", y = Count, fill = Status)) +
  geom_bar(stat = "identity", width = 1) +  # Create bars (used for pie chart)
  coord_polar(theta = "y") +  # Convert to pie chart
  labs(title = "Living Status", fill = "Status") +
  theme_void() +  # Remove gridlines and axes
  geom_text(aes(label = paste(Count, "(", Percentage, "%)", sep = "")),
            position = position_stack(vjust = 0.5))

grid.arrange(p7, p8, nrow = 1, ncol = 2)
```




```
clin_df = data[,
                     c("PATIENT_ID",
                       "OS_STATUS",
                       "OS_MONTHS",
                       "DAYS_LAST_FOLLOWUP",
                       "SEX",
                       "AJCC_PATHOLOGIC_TUMOR_STAGE",
                       "SUBTYPE",
                       "RACE",
                       "PRIMARY_LYMPH_NODE_PRESENTATION_ASSESSMENT",
                       "GENETIC_ANCESTRY_LABEL",
                       "RADIATION_THERAPY" ,
                       "PERSON_NEOPLASM_CANCER_STATUS",
                       "ETHNICITY",
                       "HISTORY_NEOADJUVANT_TRTYN",
                       "NEW_TUMOR_EVENT_AFTER_INITIAL_TREATMENT")]

clin_df$deceased = clin_df$OS_STATUS == "1:DECEASED"

clin_df[which(clin_df$AJCC_PATHOLOGIC_TUMOR_STAGE == "N/A"), "AJCC_PATHOLOGIC_TUMOR_STAGE"] = NA
clin_df$AJCC_PATHOLOGIC_TUMOR_STAGE = gsub("[ABC]$", "", clin_df$AJCC_PATHOLOGIC_TUMOR_STAGE)

# Turmor Stage Survival
Surv(clin_df$OS_MONTHS, clin_df$deceased) ~ clin_df$AJCC_PATHOLOGIC_TUMOR_STAGE
```

```
## Surv(clin_df$OS_MONTHS, clin_df$deceased) ~ clin_df$AJCC_PATHOLOGIC_TUMOR_STAGE
```

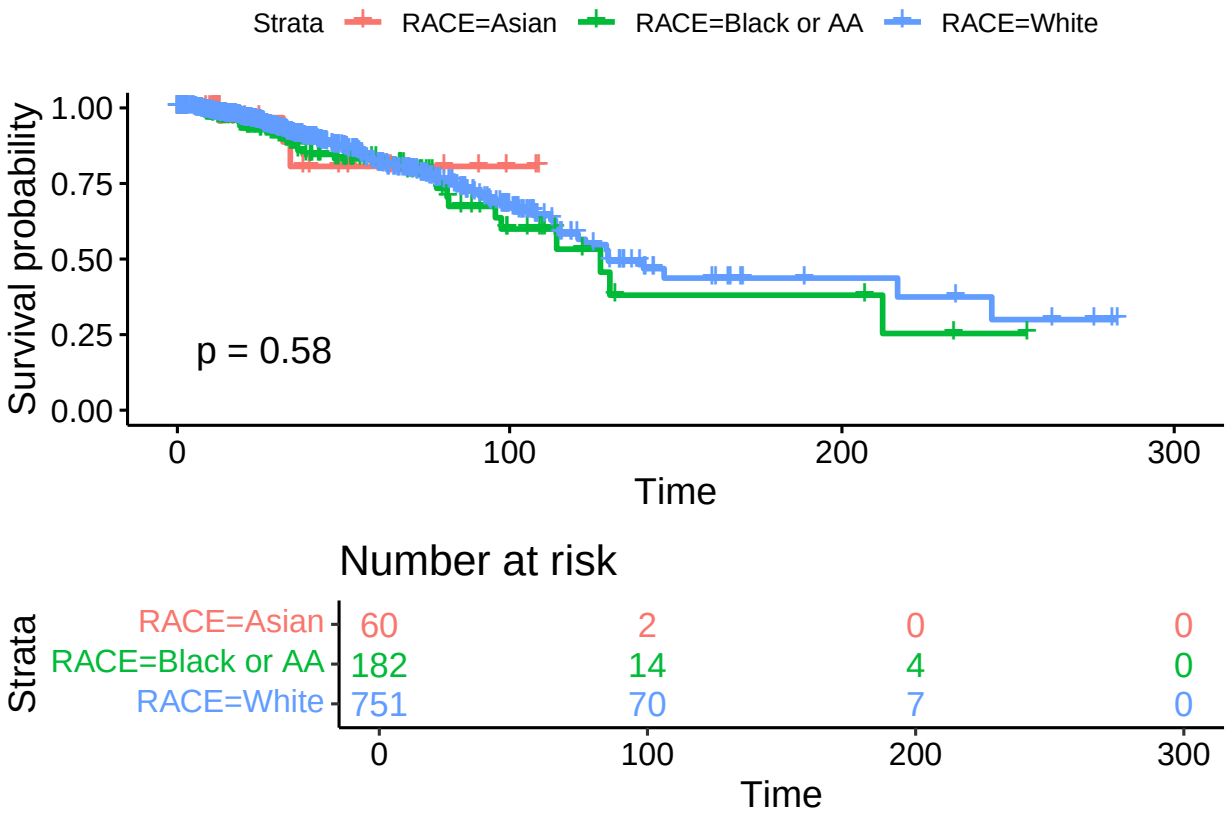
```r
fit = survfit(Surv(OS_MONTHS, deceased) ~ AJCC_PATHOLOGIC_TUMOR_STAGE, data=clin_df)

ggsurvplot(fit, data=clin_df, pval=T, risk.table=T, risk.table.col="strata", risk.table.height=0.35, leg
```




```r
# Race Survival
clin_df[which(clin_df$RACE == "N/A"), "RACE"] = NA
clin_df[which(clin_df$RACE == "Native"), "RACE"] = NA

Surv(clin_df$OS_MONTHS, clin_df$deceased) ~ clin_df$RACE
```

```
## Surv(clin_df$OS_MONTHS, clin_df$deceased) ~ clin_df$RACE
```

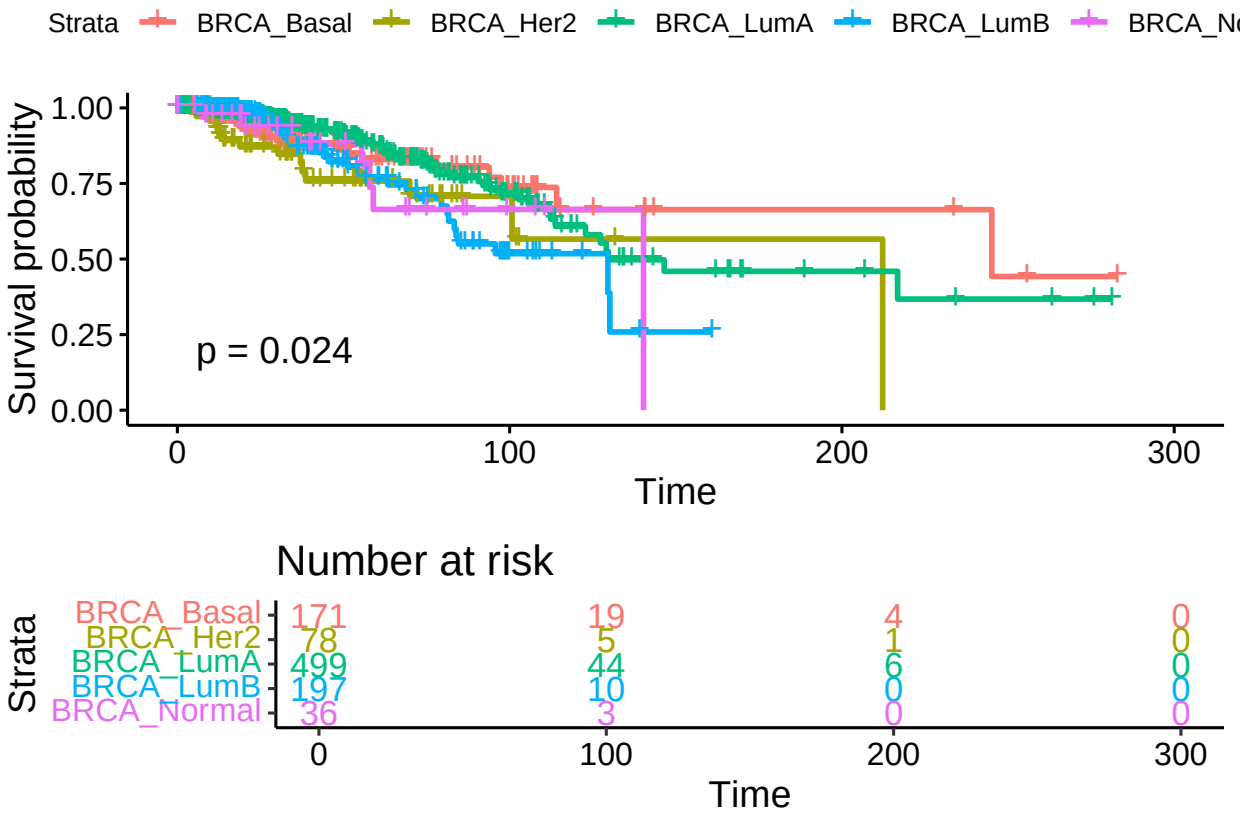
```r
fit_race = survfit(Surv(OS_MONTHS, deceased) ~ RACE, data=clin_df)
ggsurvplot(fit_race, data=clin_df, pval=T, risk.table=T, risk.table.col="strata", risk.table.height=0.3
```




```
# Subtype Survival
clin_df[which(clin_df$SUBTYPE == "N/A"), "SUBTYPE"] = NA
Surv(clin_df$OS_MONTHS, clin_df$deceased) ~ clin_df$SUBTYPE
```

```
## Surv(clin_df$OS_MONTHS, clin_df$deceased) ~ clin_df$SUBTYPE
```

```
fit_subtype = survfit(Surv(OS_MONTHS, deceased) ~ SUBTYPE, data=clin_df)

ggsurvplot(fit_subtype, data=clin_df, pval=T, risk.table=T, risk.table.col="strata", risk.table.height=
```




```
#Lymph Node Survival
clin_df[which(clin_df$PRIMARY_LYMPH_NODE_PRESENTATION_ASSESSMENT == ""), "PRIMARY_LYMPH_NODE_PRESENTATI
table(clin_df$PRIMARY_LYMPH_NODE_PRESENTATION_ASSESSMENT)
```

```
##
##  No Yes
##  33 687
```

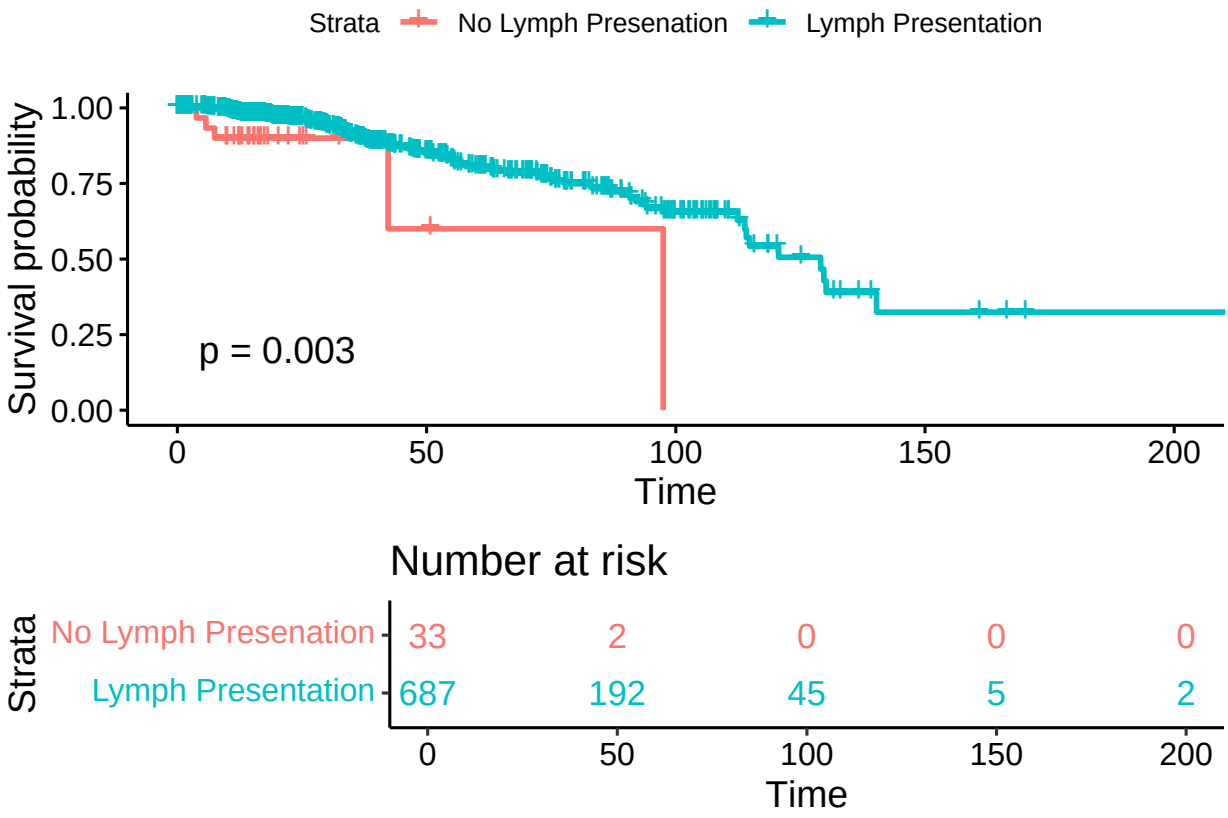
```
Surv(clin_df$OS_MONTHS, clin_df$deceased) ~ clin_df$PRIMARY_LYMPH_NODE_PRESENTATION_ASSESSMENT
```

```
## Surv(clin_df$OS_MONTHS, clin_df$deceased) ~ clin_df$PRIMARY_LYMPH_NODE_PRESENTATION_ASSESSMENT
```

```
fit_lymph = survfit(Surv(OS_MONTHS, deceased) ~ PRIMARY_LYMPH_NODE_PRESENTATION_ASSESSMENT, data=clin_d
fit_lymph$strata
```

```
##  PRIMARY_LYMPH_NODE_PRESENTATION_ASSESSMENT=No
##                                             31
## PRIMARY_LYMPH_NODE_PRESENTATION_ASSESSMENT=Yes
##                                            574
```

```
ggsurvplot(fit_lymph, data=clin_df, pval=T, risk.table=T, risk.table.col="strata", risk.table.height=0.
```



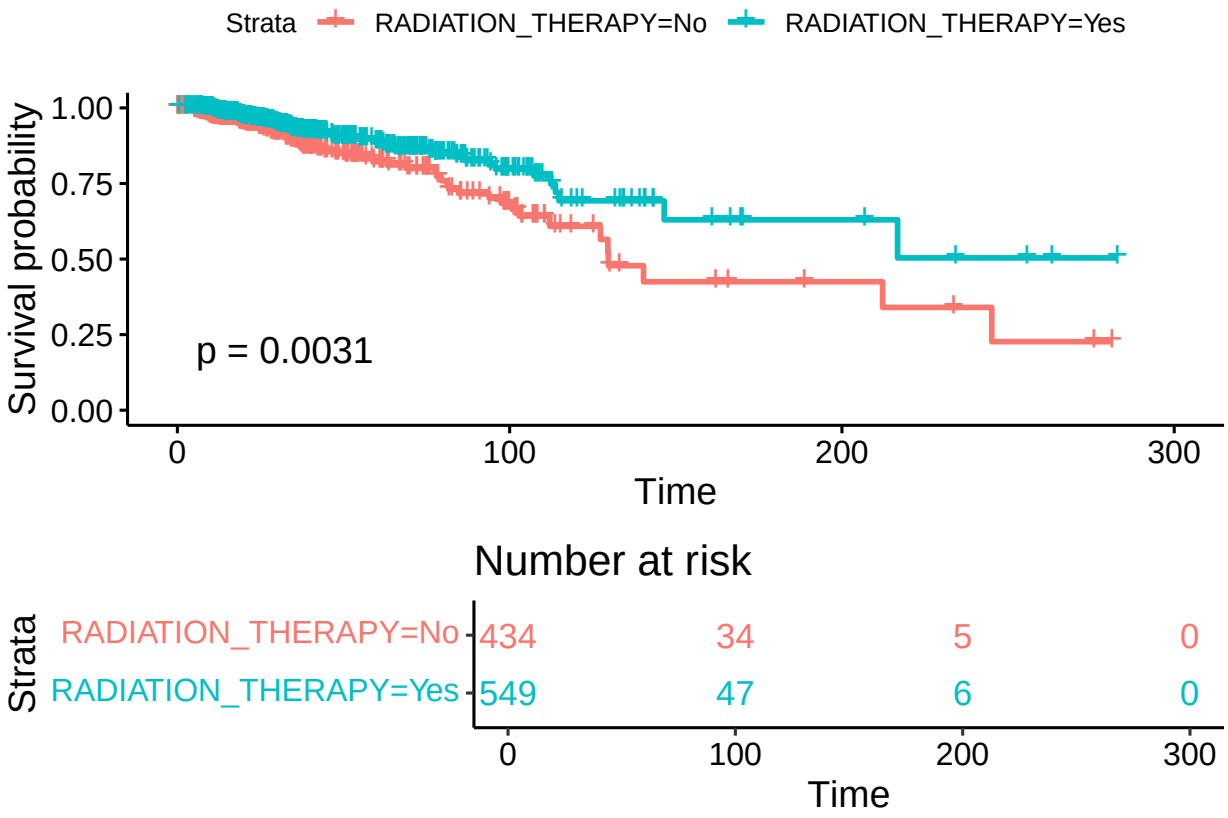

```
# Radiation Surviavl
table(clin_df$RADIATION_THERAPY)
```

```
##
##  No Yes
## 434 549
```

```
Surv(clin_df$OS_MONTHS, clin_df$deceased) ~ clin_df$RADIATION_THERAPY
```

```
## Surv(clin_df$OS_MONTHS, clin_df$deceased) ~ clin_df$RADIATION_THERAPY
```

```
fit_rad = survfit(Surv(OS_MONTHS, deceased) ~ RADIATION_THERAPY, data=clin_df)
ggsurvplot(fit_rad, data=clin_df, pval=T, risk.table=T, risk.table.col="strata", risk.table.height=0.35
```



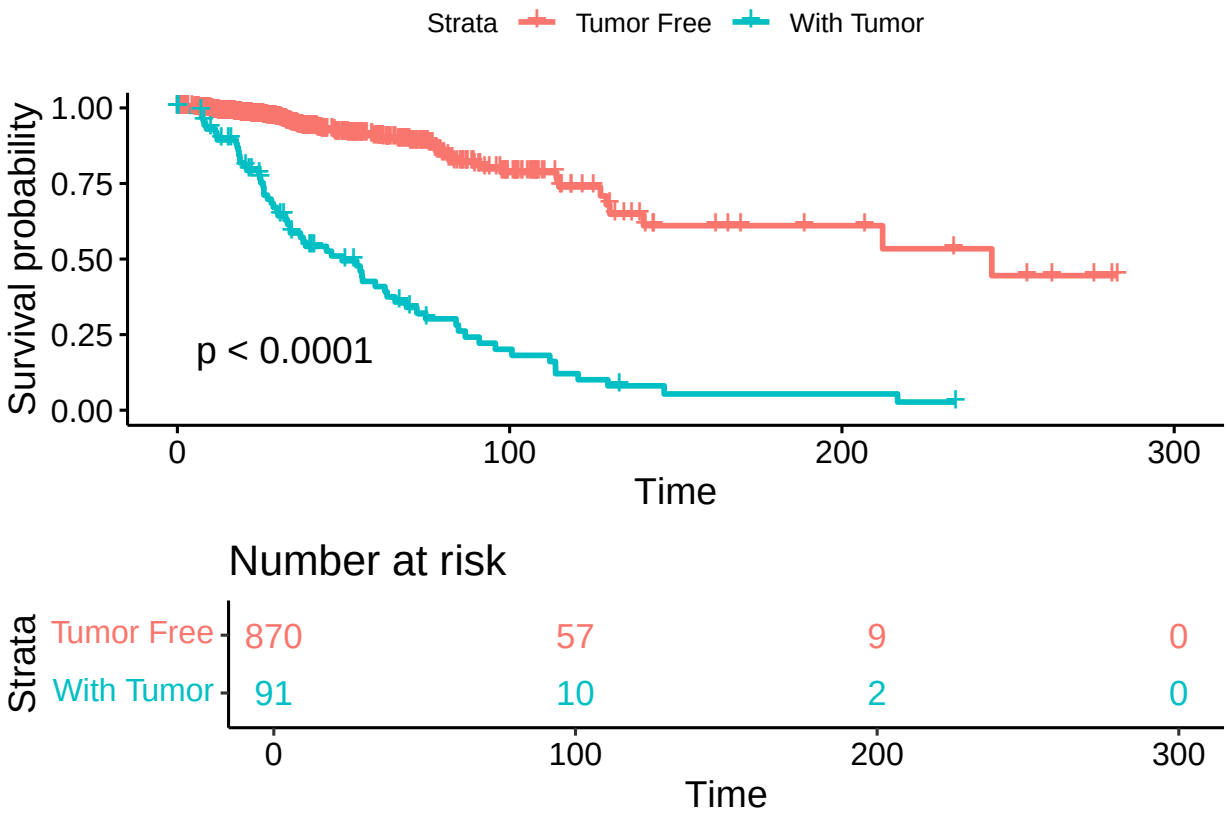

```
# Cancer status
table(data$PERSON_NEOPLASM_CANCER_STATUS)
```

```
##
## Tumor Free With Tumor
##        870         91
```

```
Surv(clin_df$OS_MONTHS, clin_df$deceased) ~ clin_df$PERSON_NEOPLASM_CANCER_STATUS
```

```
## Surv(clin_df$OS_MONTHS, clin_df$deceased) ~ clin_df$PERSON_NEOPLASM_CANCER_STATUS
```

```
fit_status = survfit(Surv(OS_MONTHS, deceased) ~ PERSON_NEOPLASM_CANCER_STATUS, data=clin_df)
ggsurvplot(fit_status, data=clin_df, pval=T, risk.table=T, risk.table.col="strata", risk.table.height=0
```



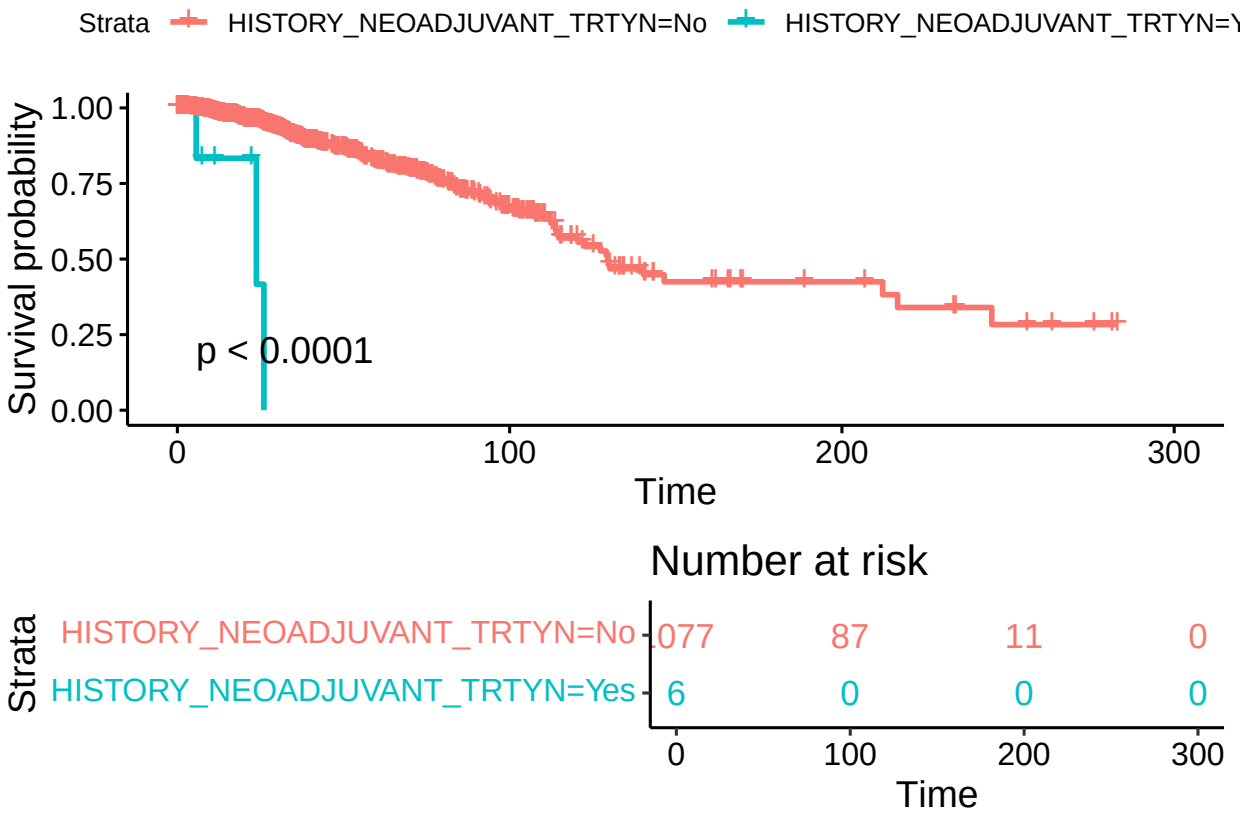

```
# Cancer treat
table(data$HISTORY_NEOADJUVANT_TRTYN)
```

```
##
##   No  Yes
## 1077    6
```

```
Surv(clin_df$OS_MONTHS, clin_df$deceased) ~ data$HISTORY_NEOADJUVANT_TRTYN
```

```
## Surv(clin_df$OS_MONTHS, clin_df$deceased) ~ data$HISTORY_NEOADJUVANT_TRTYN
```

```
fit_treat = survfit(Surv(OS_MONTHS, deceased) ~ HISTORY_NEOADJUVANT_TRTYN, data=clin_df)
ggsurvplot(fit_treat, data=clin_df, pval=T, risk.table=T, risk.table.col="strata", risk.table.height=0.
```



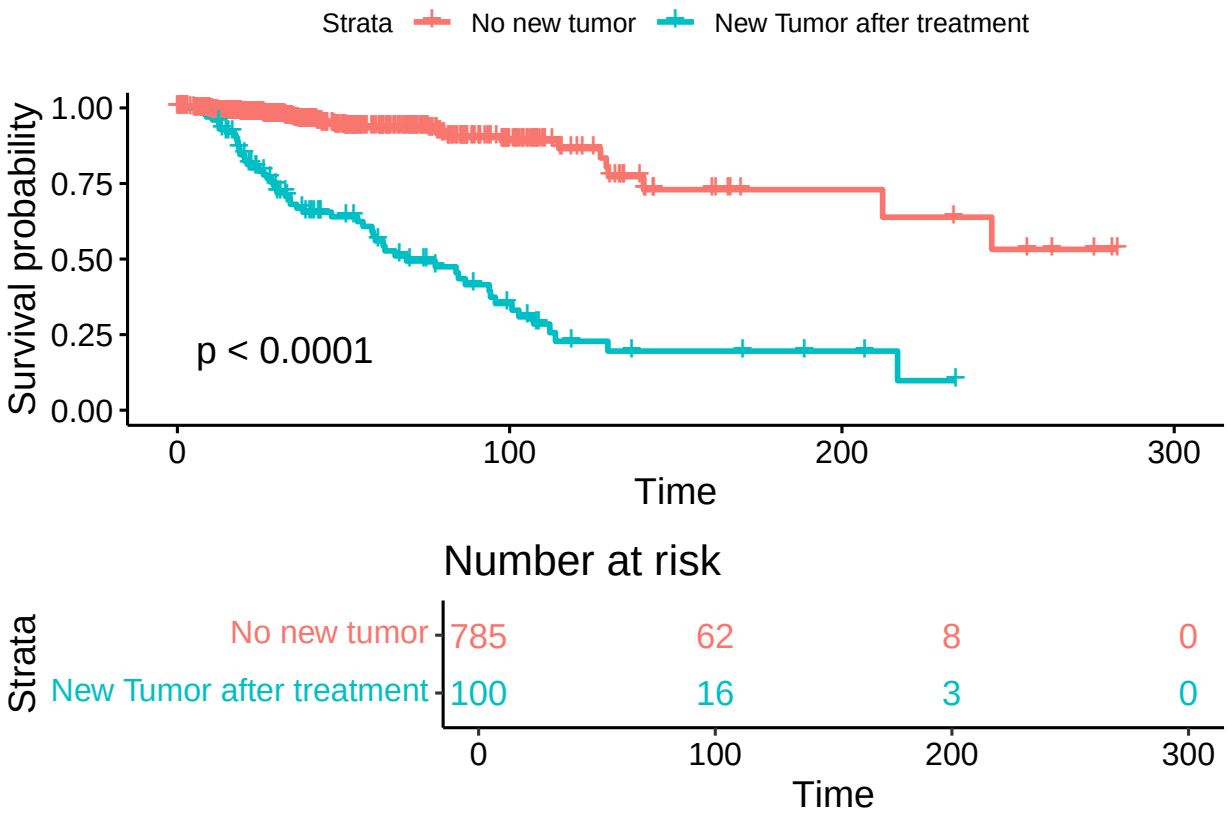

```
#new tumor
table(data$NEW_TUMOR_EVENT_AFTER_INITIAL_TREATMENT)
```

```
##
##  No Yes
## 785 100
```

```
Surv(clin_df$OS_MONTHS, clin_df$deceased) ~ data$NEW_TUMOR_EVENT_AFTER_INITIAL_TREATMENT
```

```
## Surv(clin_df$OS_MONTHS, clin_df$deceased) ~ data$NEW_TUMOR_EVENT_AFTER_INITIAL_TREATMENT
```

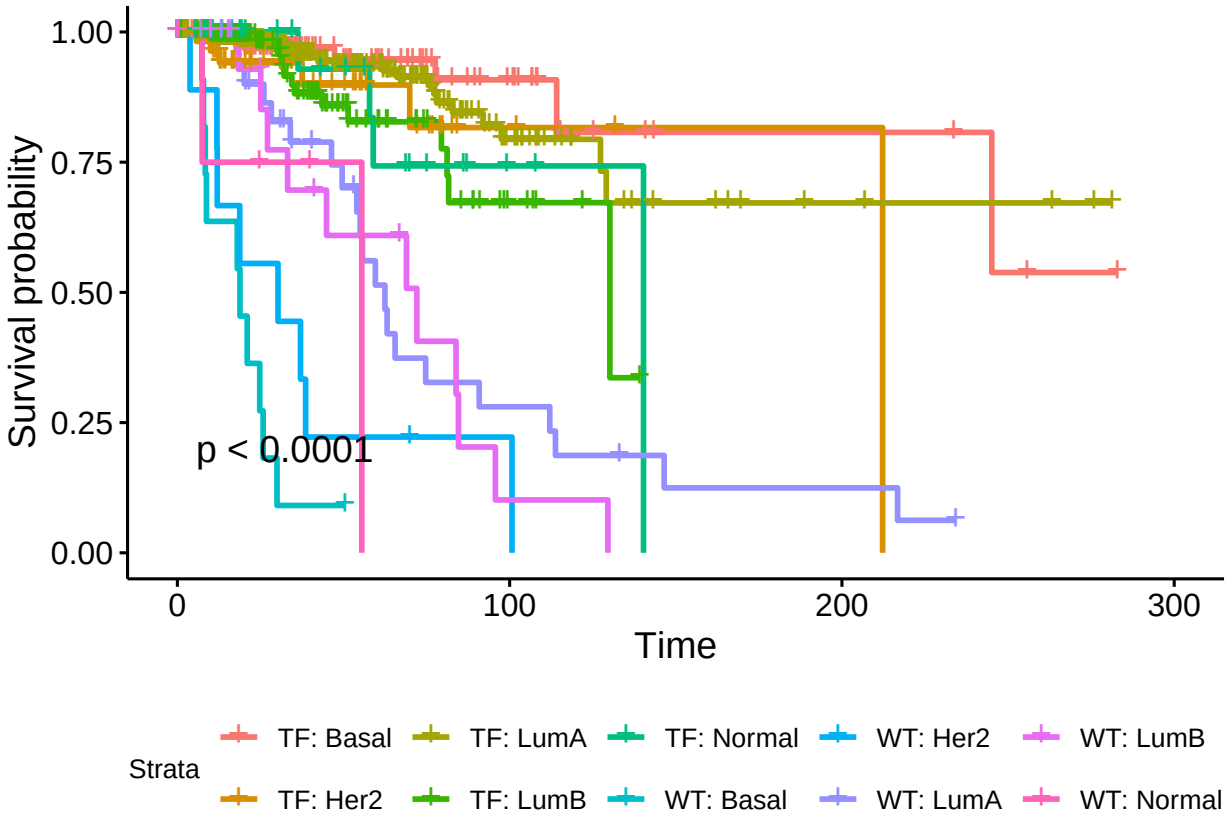
```
fit_new = survfit(Surv(OS_MONTHS, deceased) ~ NEW_TUMOR_EVENT_AFTER_INITIAL_TREATMENT, data=clin_df)
ggsurvplot(fit_new, data=clin_df, pval=T, risk.table=T, risk.table.col="strata", risk.table.height=0.35
```




```
#combinations
Surv(clin_df$OS_MONTHS, clin_df$deceased) ~ clin_df$PERSON_NEOPLASM_CANCER_STATUS + clin_df$SUBTYPE
```

```
## Surv(clin_df$OS_MONTHS, clin_df$deceased) ~ clin_df$PERSON_NEOPLASM_CANCER_STATUS +
##     clin_df$SUBTYPE
```

```
fit_comb = survfit(Surv(OS_MONTHS, deceased) ~ clin_df$PERSON_NEOPLASM_CANCER_STATUS + clin_df$SUBTYPE,
ggsurvplot(fit_comb, data=clin_df, pval=T, legend = "bottom",legend.labs = c("TF: Basal", "TF: Her2", "
```


1.00
0.75
0.50
0.25
0.00
Survival probability
p < 0.0001
0
100
200
300
Time
Strata
TF: Basal
TF: LumA
TF: Normal
WT: Her2
WT: LumB
TF: Her2
TF: LumB
WT: Basal
WT: LumA
WT: Normal