P8130_Final_Project_Report

Jiying Wang

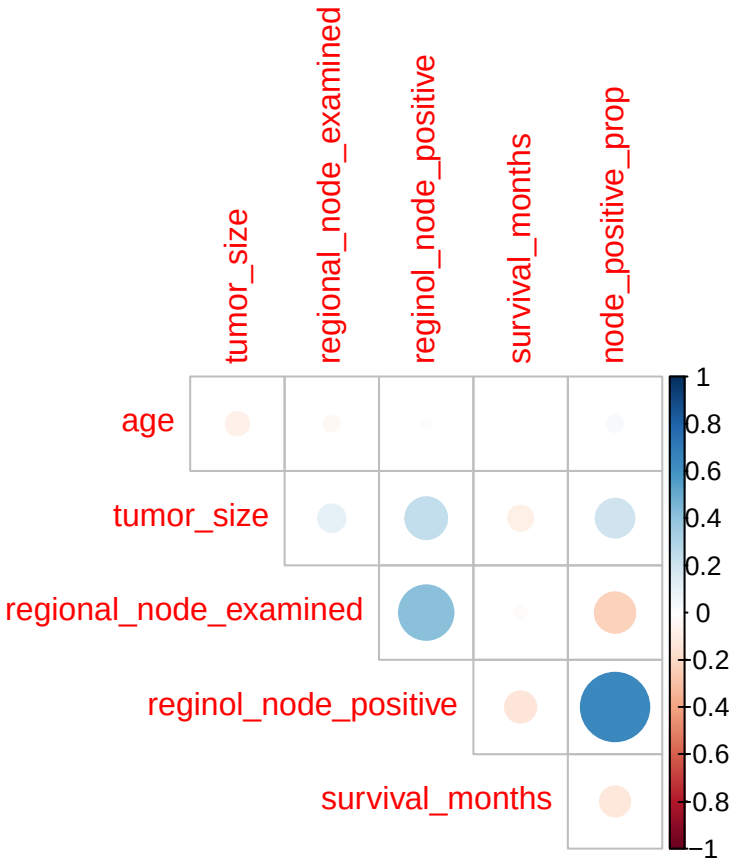
2023-12-15

Introduction

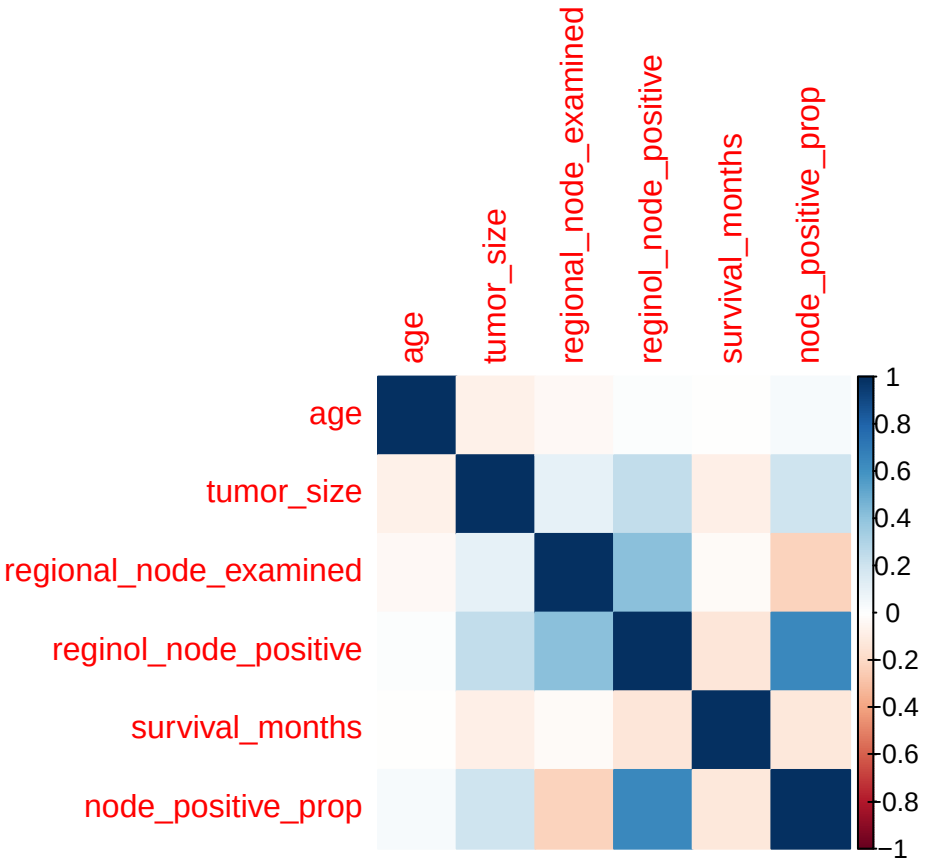
Clean the dataset

EDA

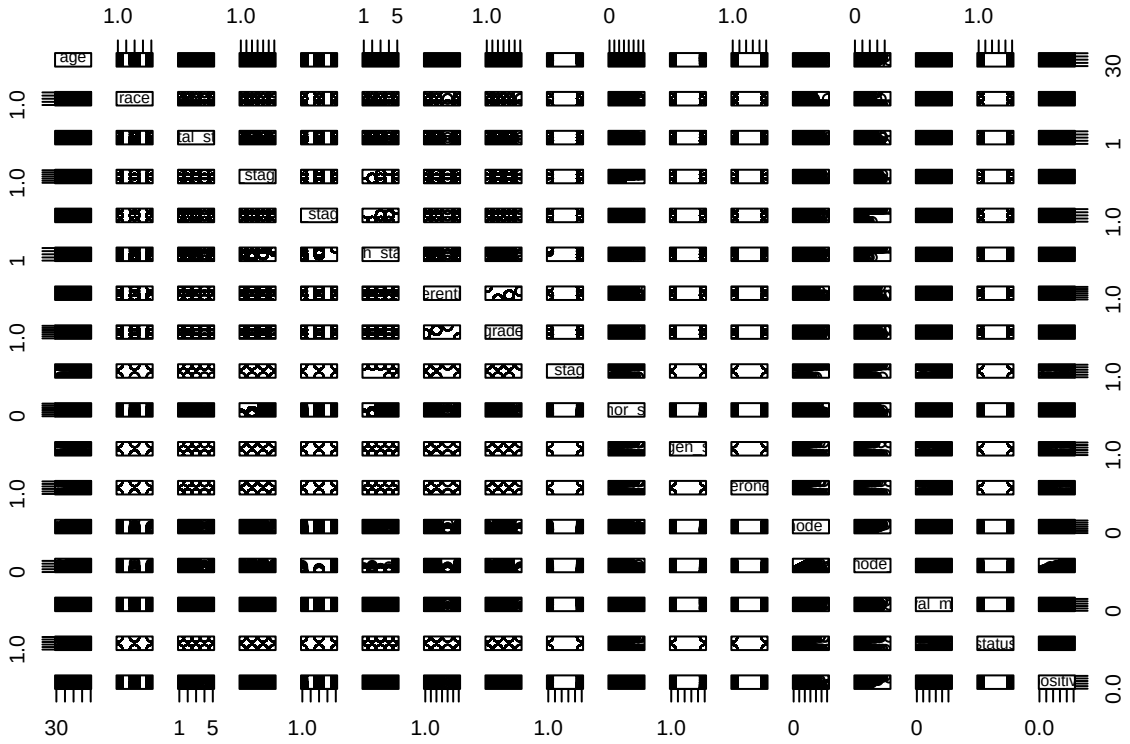
```
# Correlation matrix for all variables
cor_matrix = breastcancer_clean |>
  select_if(is.numeric) |>
  cor()
corrplot::corrplot(cor_matrix, type = "upper", diag = FALSE)
```




```
# Heat map
corrplot::corrplot(cor_matrix, method = "color")
```

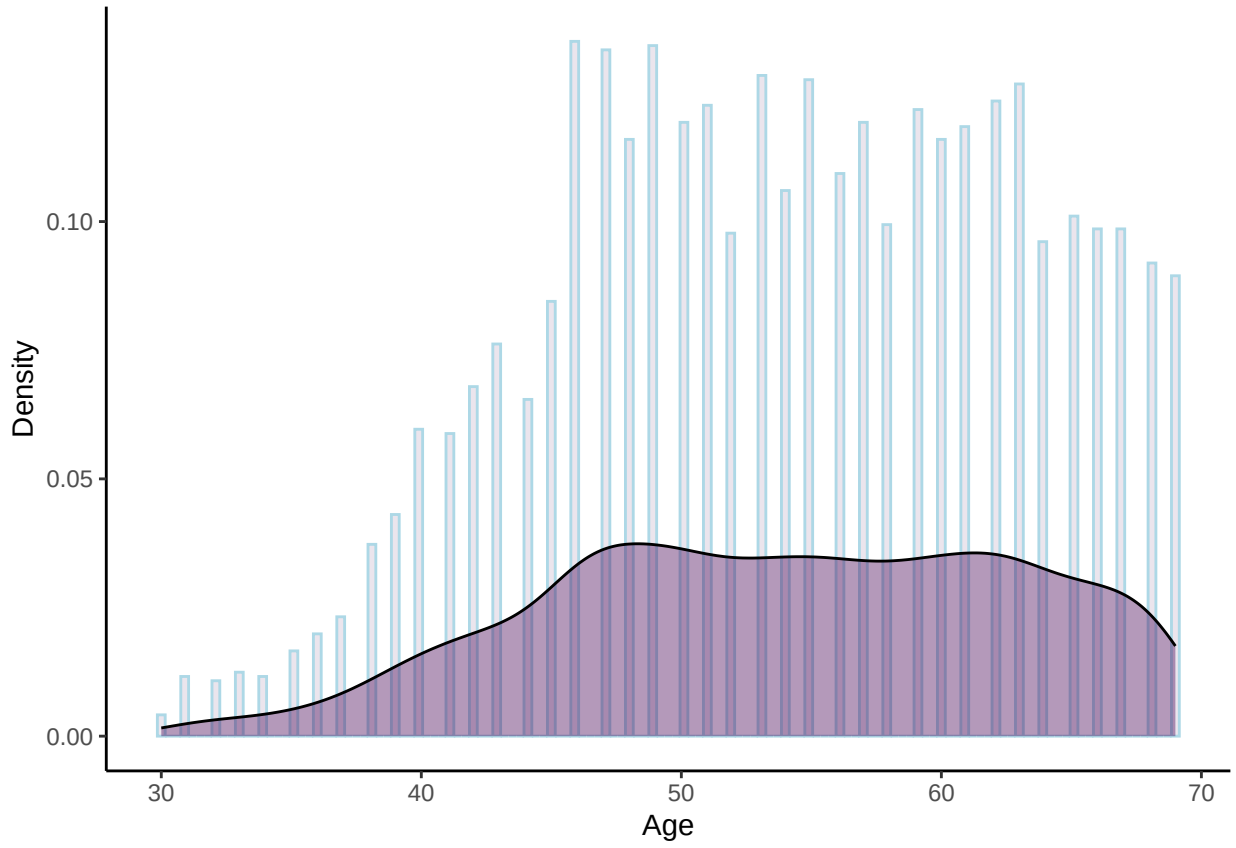



```
# Scatter plot matrix for all variables
pairs(breastcancer_clean)
```

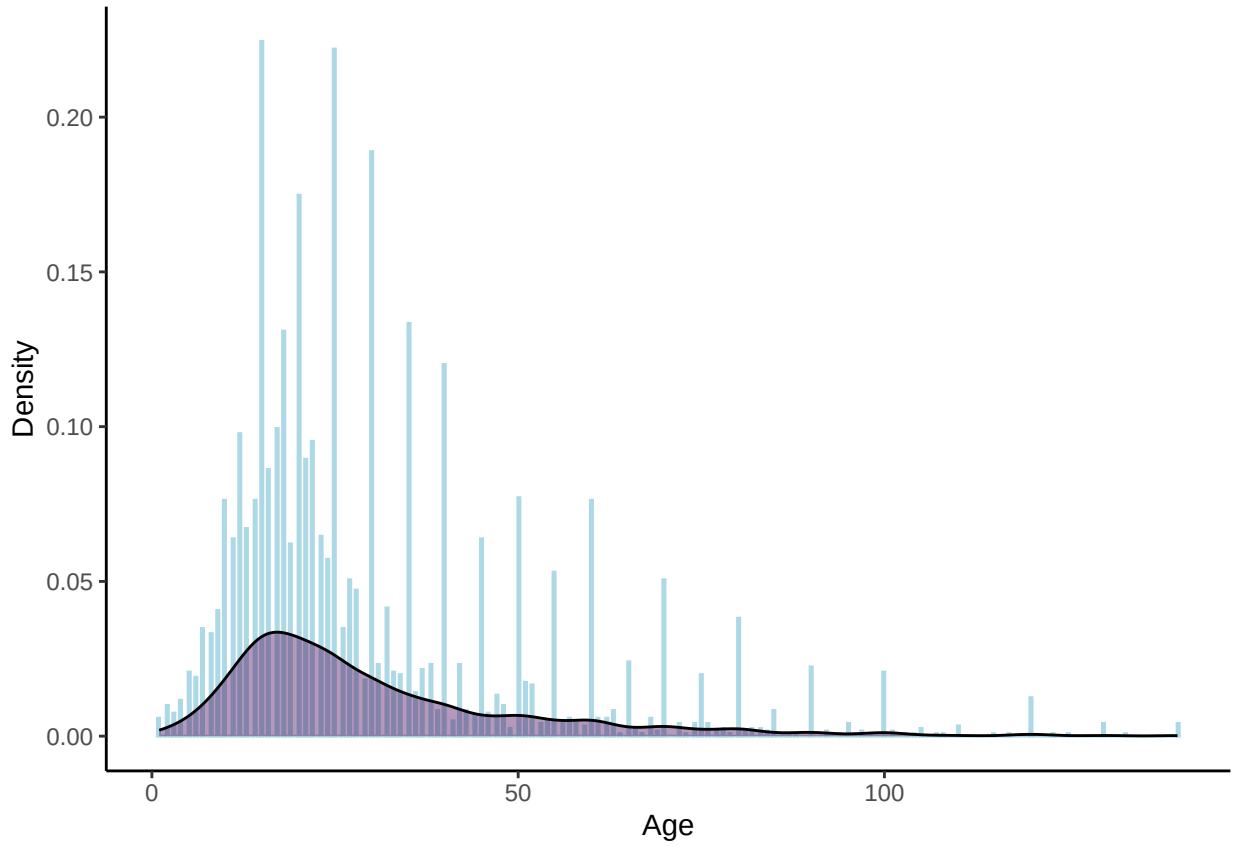



```
# Histograms and density plots
breastcancer_clean |>
  ggplot(aes(x = age, y = ..density.., fill = "purple", alpha = 0.5)) +
  geom_histogram(binwidth = 0.3, colour = "lightblue", alpha = 0.1) +
  geom_density(alpha = 0.4) +
  labs(
    x = "Age",
    y = "Density") +
  scale_fill_viridis_d("") +
  theme_classic() +
  theme(legend.position = "none")
```

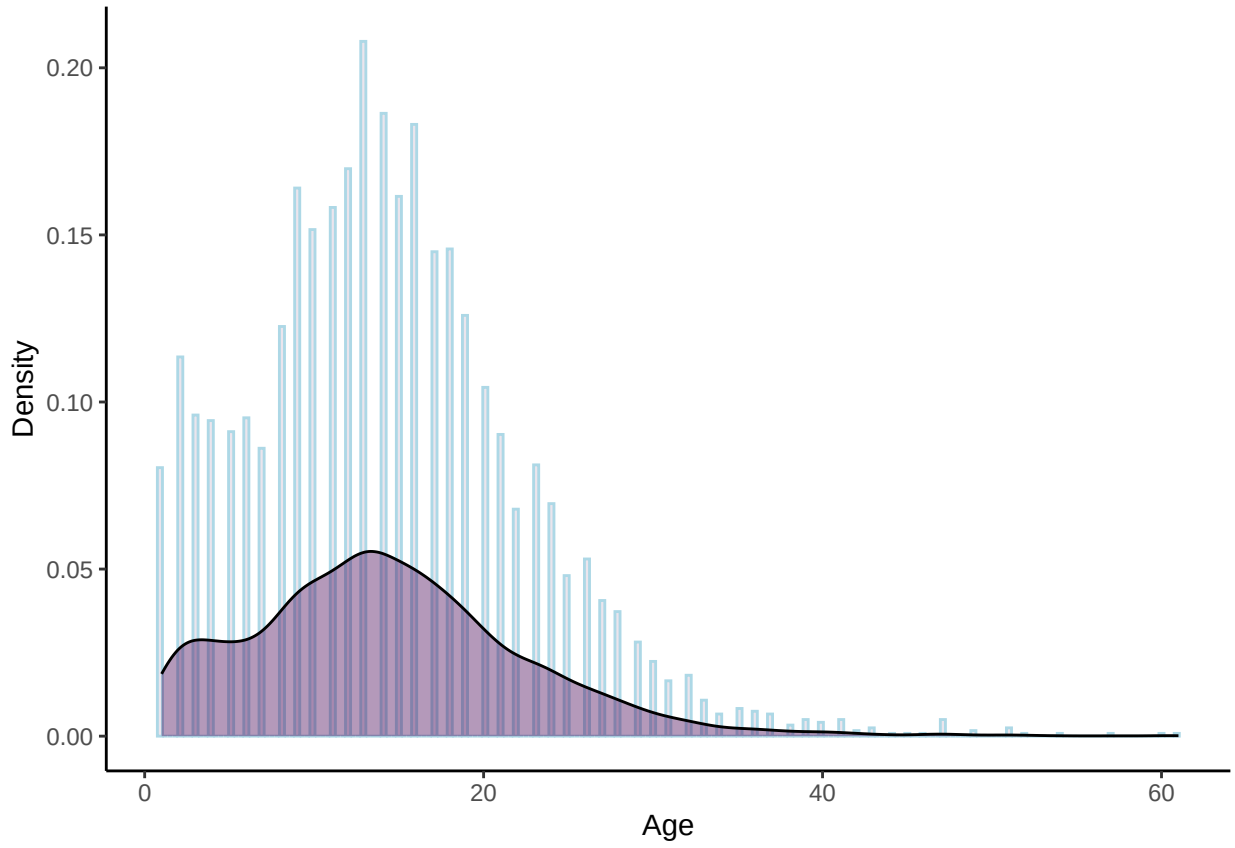
```
## Warning: The dot-dot notation (‘..density..‘) was deprecated in ggplot2 3.4.0.
## i Please use ‘after_stat(density)‘ instead.
## This warning is displayed once every 8 hours.
## Call ‘lifecycle::last_lifecycle_warnings()‘ to see where this warning was
## generated.
```



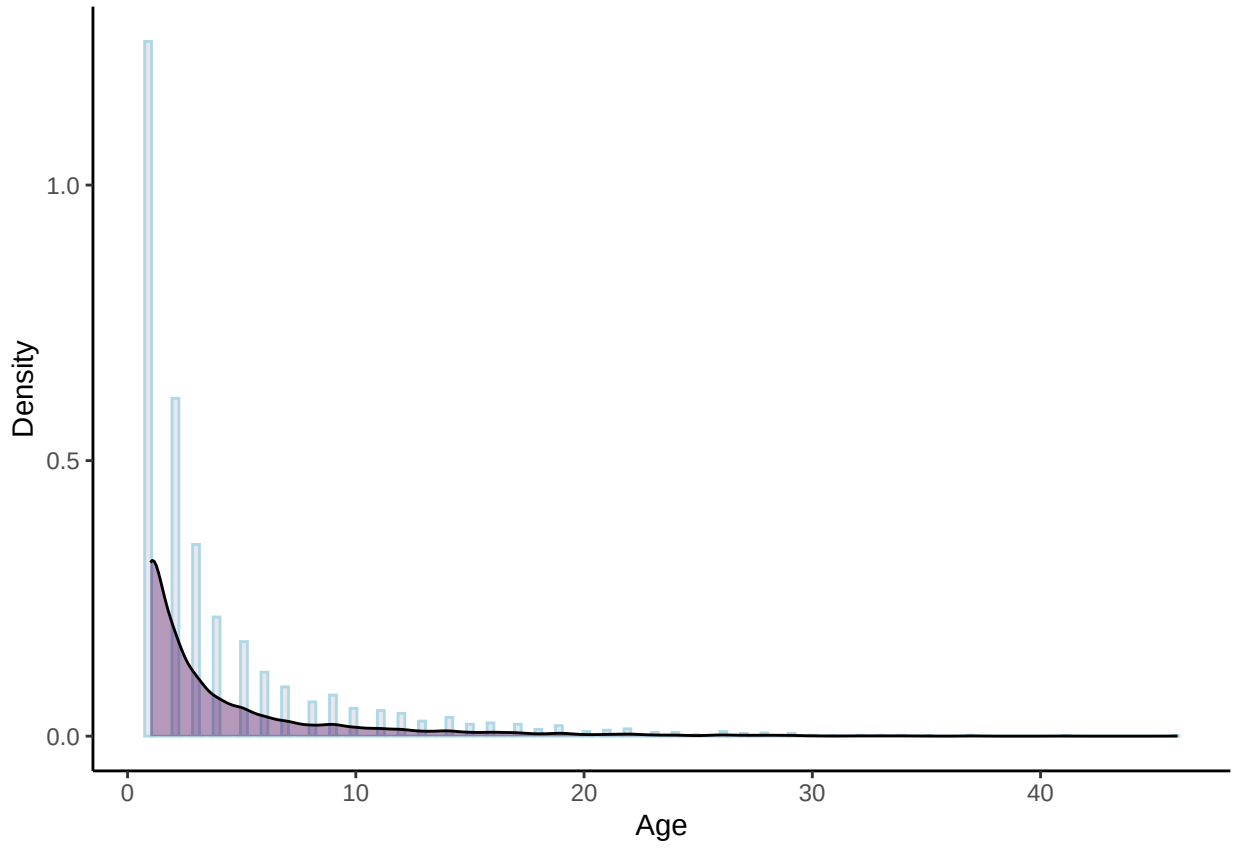

```
breastcancer_clean |>
  ggplot(aes(x = tumor_size, y = ..density.., fill = "purple", alpha = 0.5)) +
  geom_histogram(binwidth = 0.3, colour = "lightblue", alpha = 0.1) +
  geom_density(alpha = 0.4) +
  labs(
    x = "Age",
    y = "Density") +
  scale_fill_viridis_d("") +
  theme_classic() +
  theme(legend.position = "none")
```



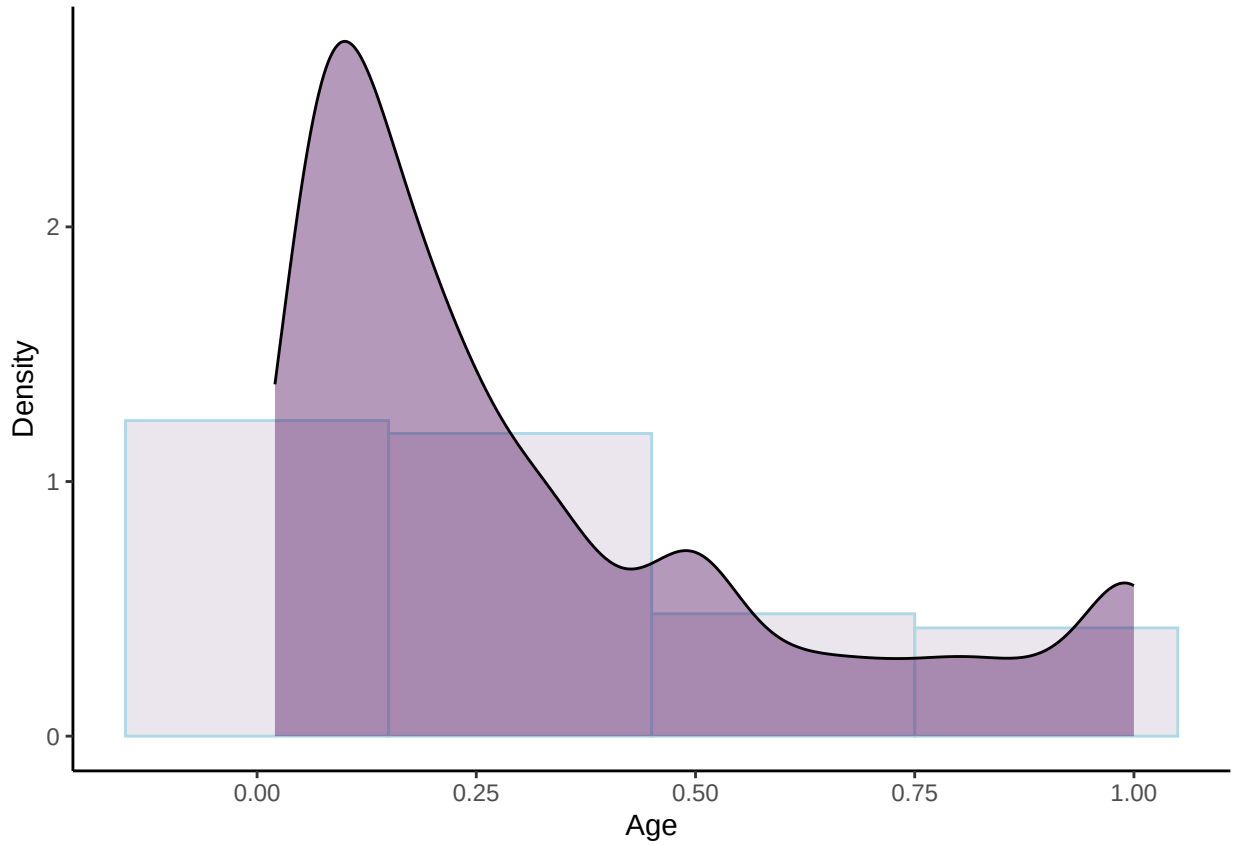

```
breastcancer_clean |>
  ggplot(aes(x = regional_node_examined, y = ..density.., fill = "purple", alpha = 0.5)) +
  geom_histogram(binwidth = 0.3, colour = "lightblue", alpha = 0.1) +
  geom_density(alpha = 0.4) +
  labs(
    x = "Age",
    y = "Density") +
  scale_fill_viridis_d("") +
  theme_classic() +
  theme(legend.position = "none")
```



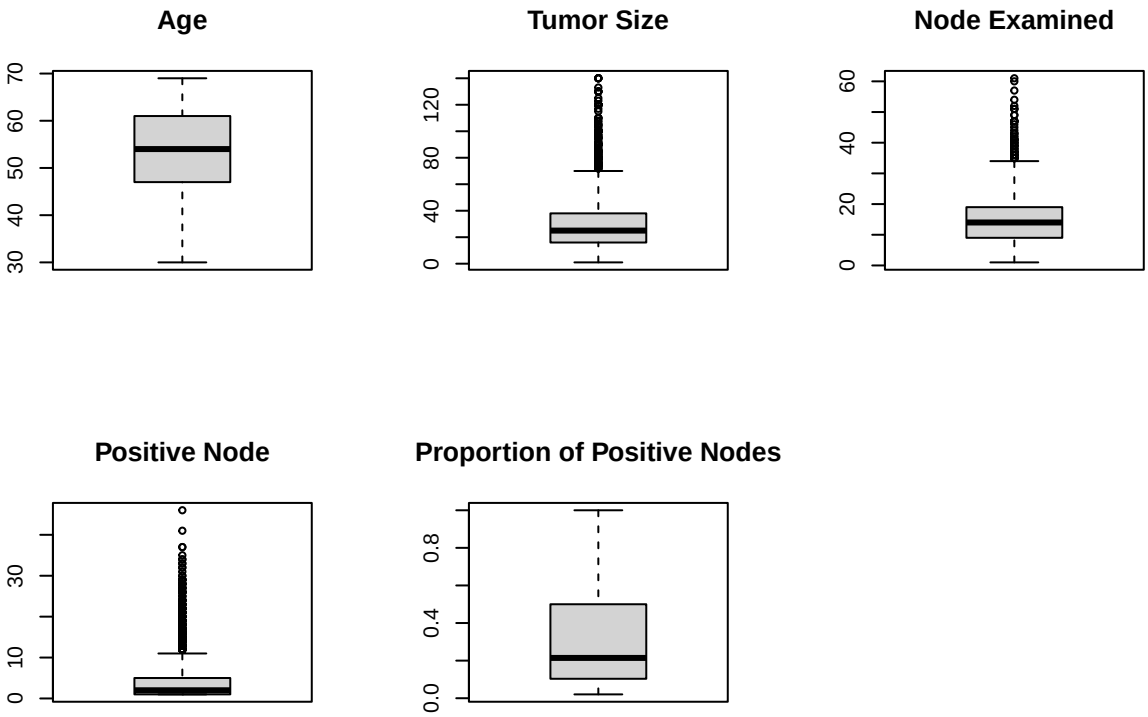

```
breastcancer_clean |>
  ggplot(aes(x = reginol_node_positive, y = ..density.., fill = "purple", alpha = 0.5)) +
  geom_histogram(binwidth = 0.3, colour = "lightblue", alpha = 0.1) +
  geom_density(alpha = 0.4) +
  labs(
    x = "Age",
    y = "Density") +
  scale_fill_viridis_d("") +
  theme_classic() +
  theme(legend.position = "none")
```



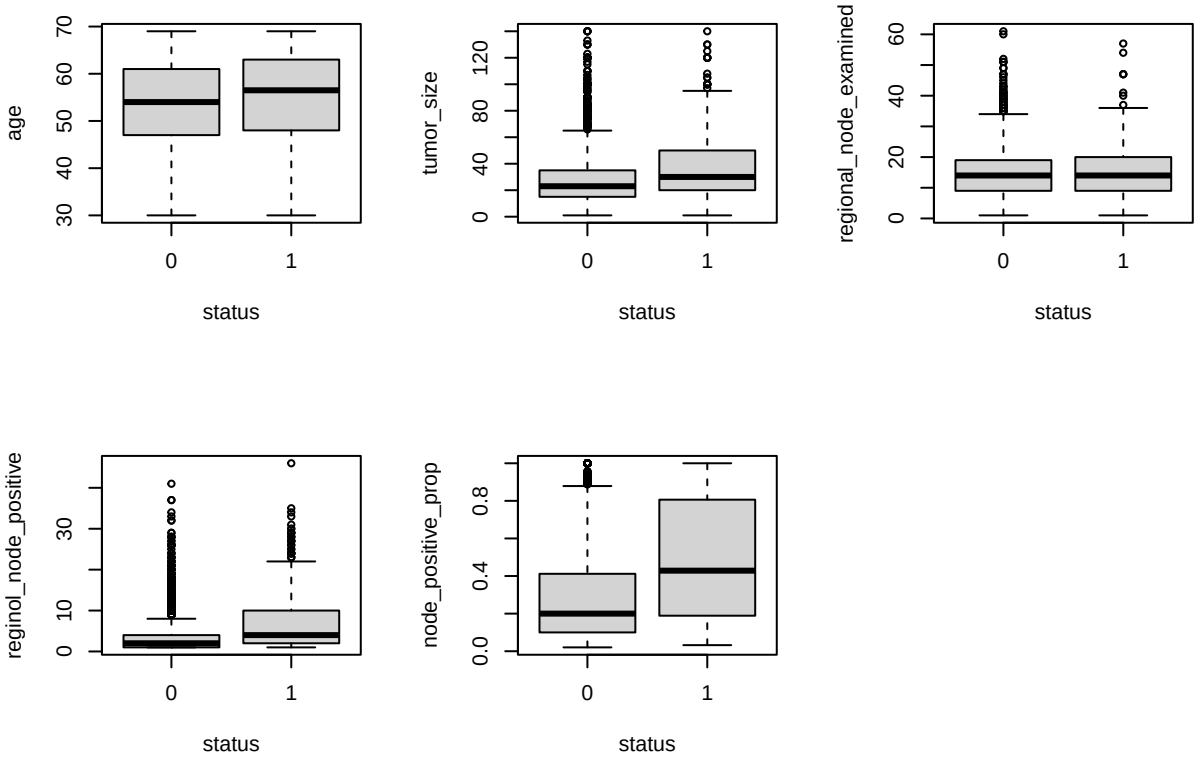

```
breastcancer_clean |>
  ggplot(aes(x = node_positive_prop, y = ..density.., fill = "purple", alpha = 0.5)) +
  geom_histogram(binwidth = 0.3, colour = "lightblue", alpha = 0.1) +
  geom_density(alpha = 0.4) +
  labs(
    x = "Age",
    y = "Density") +
  scale_fill_viridis_d("") +
  theme_classic() +
  theme(legend.position = "none")
```



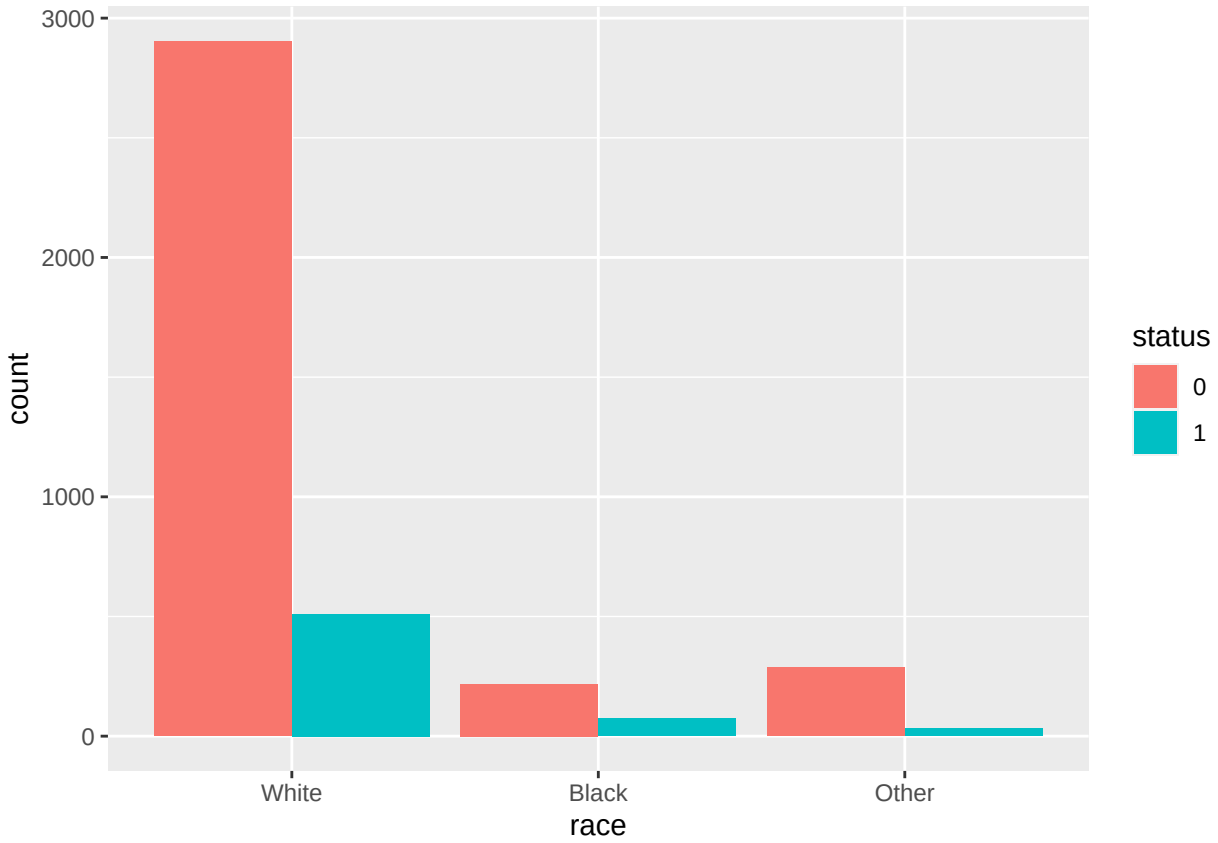

```
# Boxplots
par(mfrow=c(2,3))
boxplot(breastcancer_clean$age, main='Age')
boxplot(breastcancer_clean$tumor_size, main='Tumor Size')
boxplot(breastcancer_clean$regional_node_examined,main='Node Examined' )
boxplot(breastcancer_clean$reginol_node_positive, main='Positive Node')
boxplot(breastcancer_clean$node_positive_prop, main='Proportion of Positive Nodes')
```



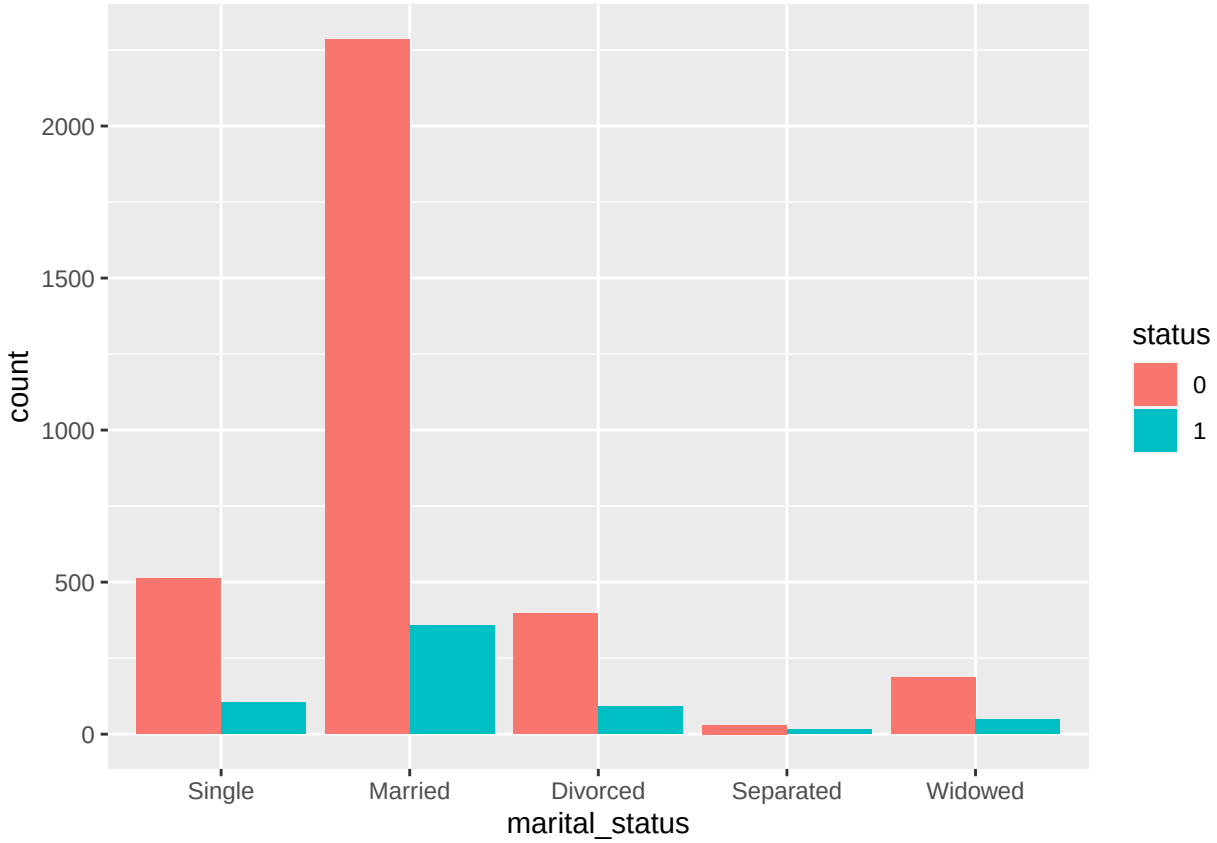

```
par(mfrow = c(2,3))
boxplot(age ~ status, breastcancer_clean)
boxplot(tumor_size ~ status, breastcancer_clean)
boxplot(regional_node_examined ~ status, breastcancer_clean)
boxplot(reginol_node_positive ~ status, breastcancer_clean)
boxplot(node_positive_prop ~ status, breastcancer_clean)
```



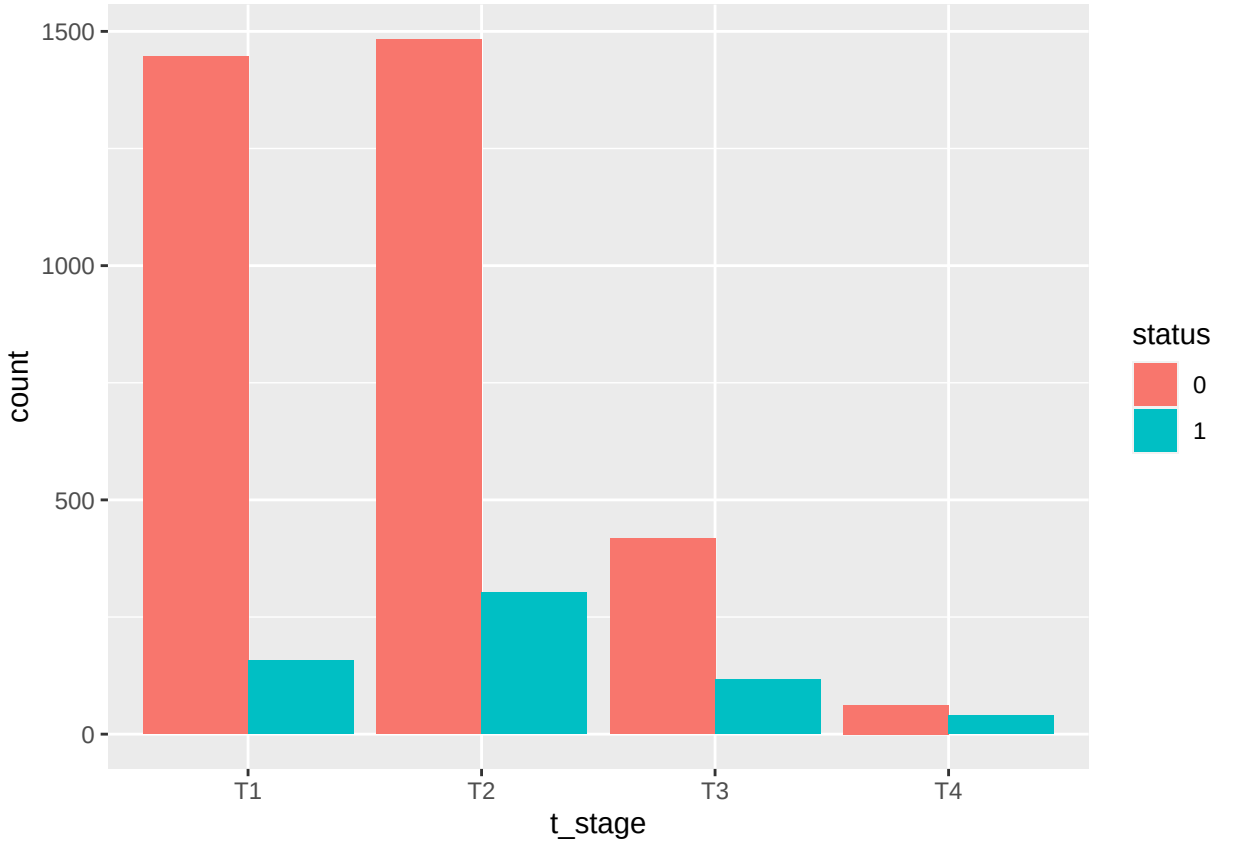

```r
# Barplots for categorical variables
ggplot(breastcancer_clean, aes(x = race, fill = status)) +
  geom_bar(position = "dodge")
```



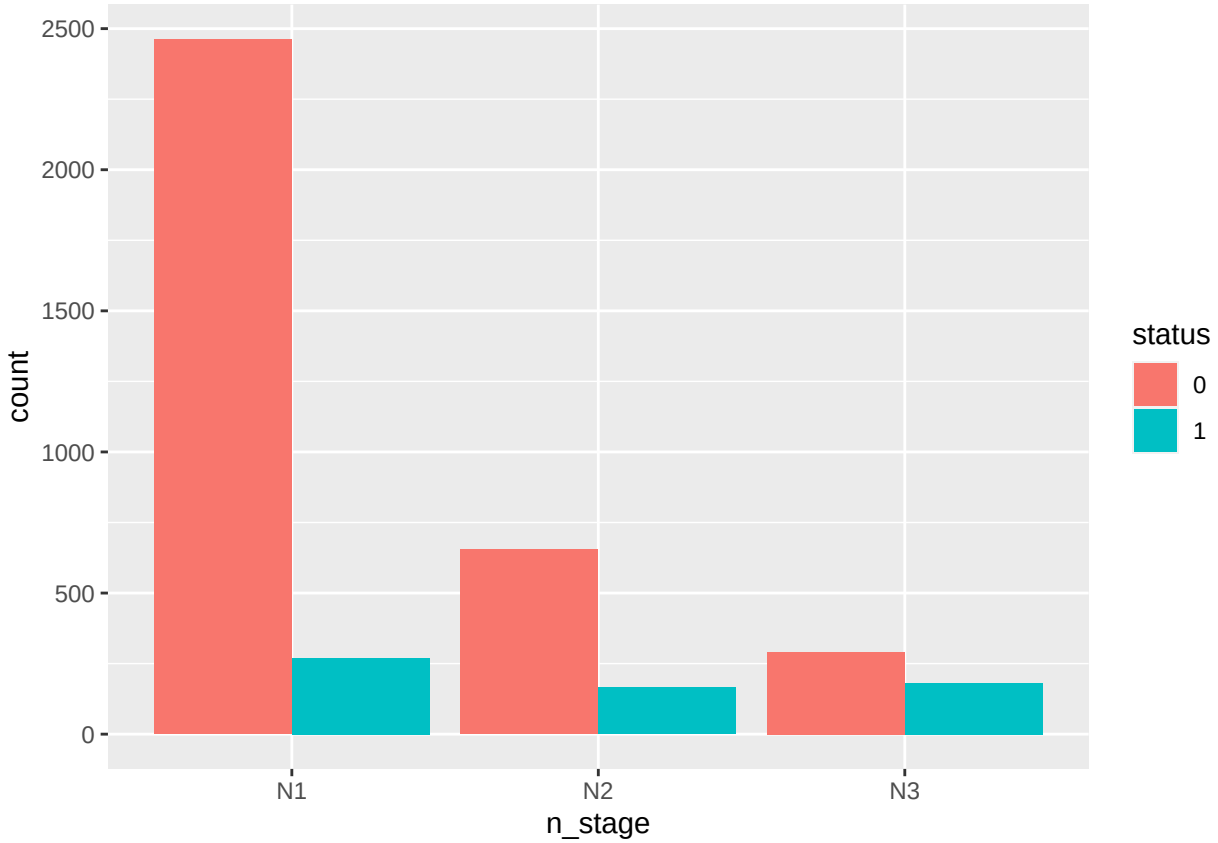

```
ggplot(breastcancer_clean, aes(x = marital_status, fill = status)) +
  geom_bar(position = "dodge")
```



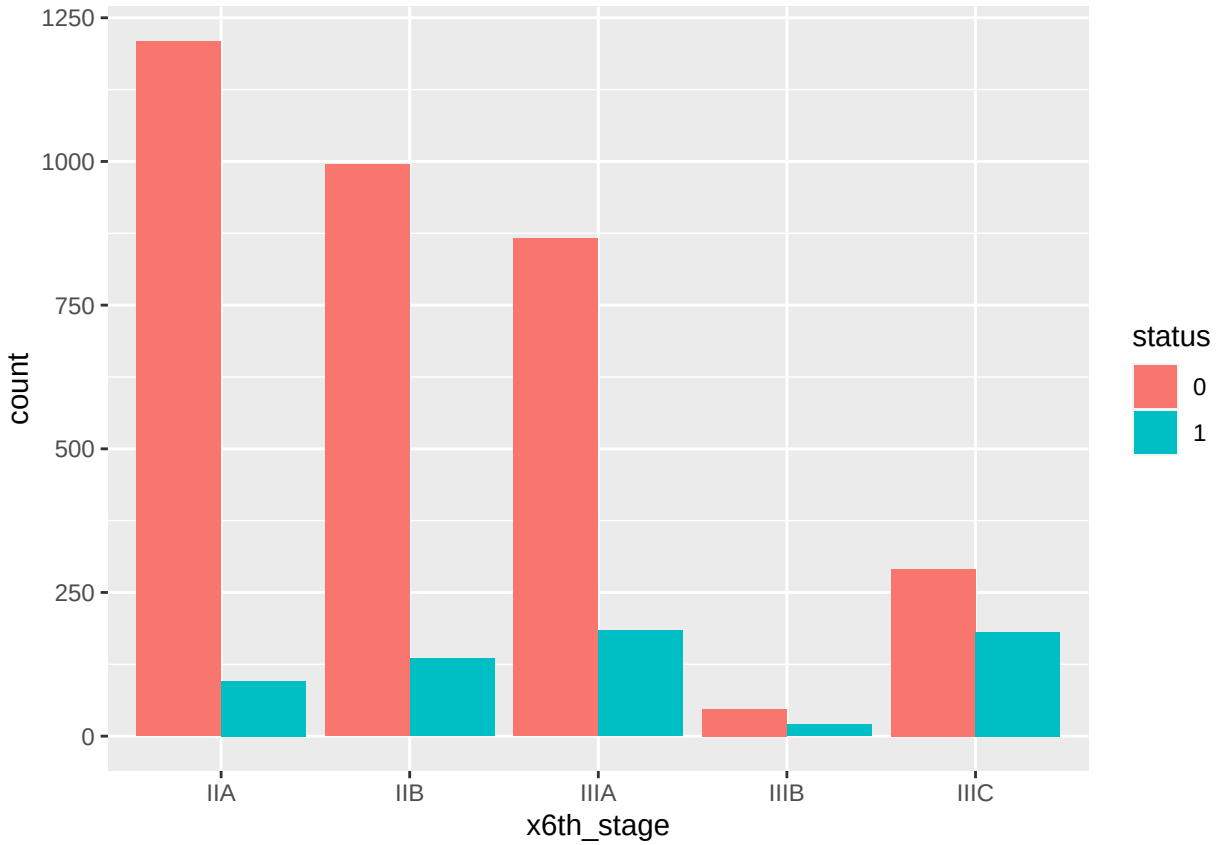

```
ggplot(breastcancer_clean, aes(x = t_stage, fill = status)) +
  geom_bar(position = "dodge")
```



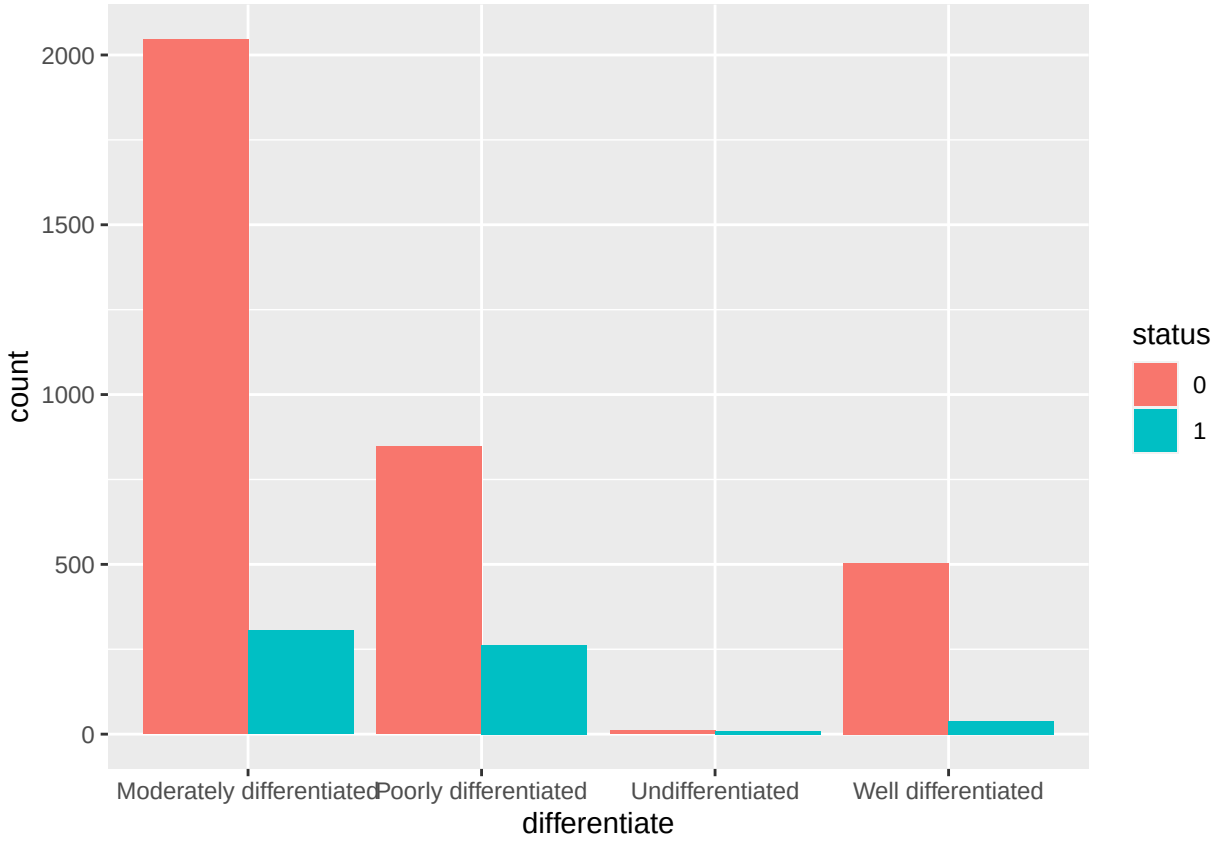

```
ggplot(breastcancer_clean, aes(x = n_stage, fill = status)) +
  geom_bar(position = "dodge")
```



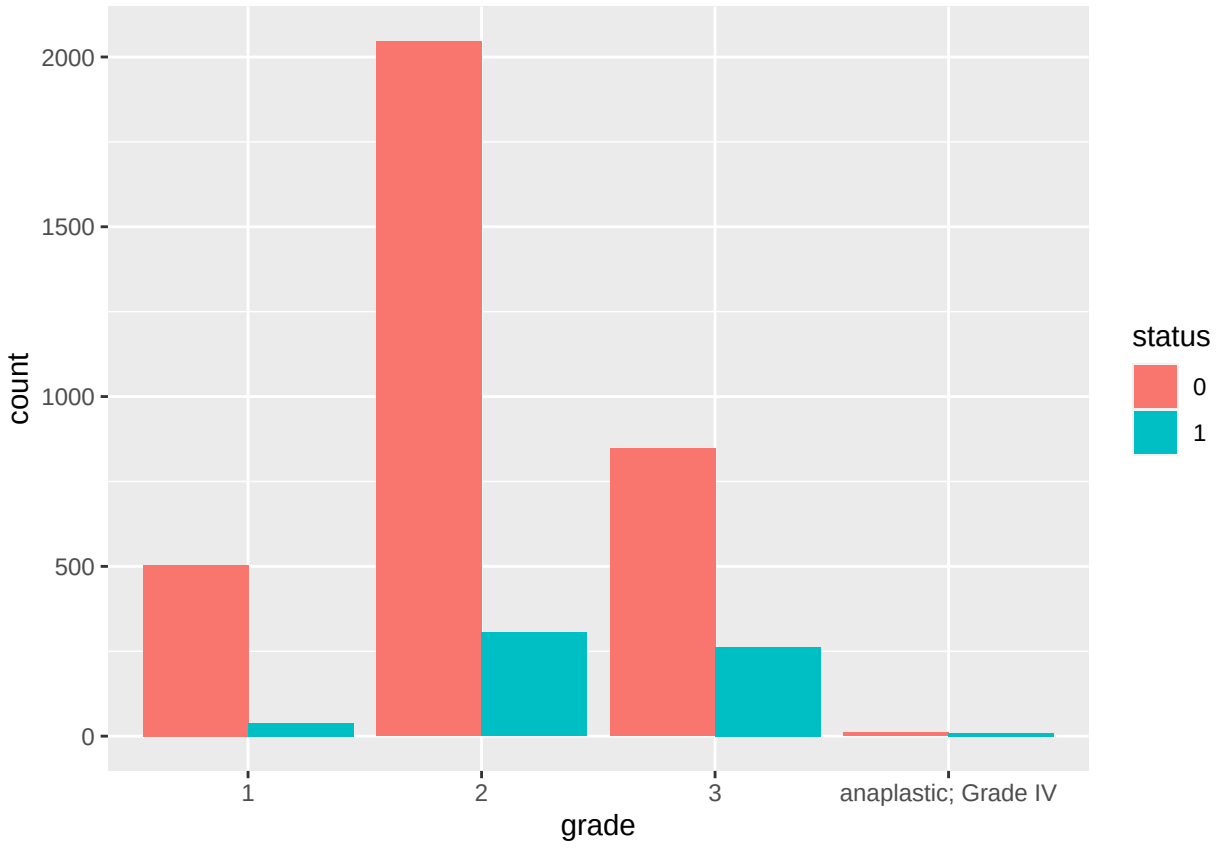

```
ggplot(breastcancer_clean, aes(x = x6th_stage, fill = status)) +
  geom_bar(position = "dodge")
```



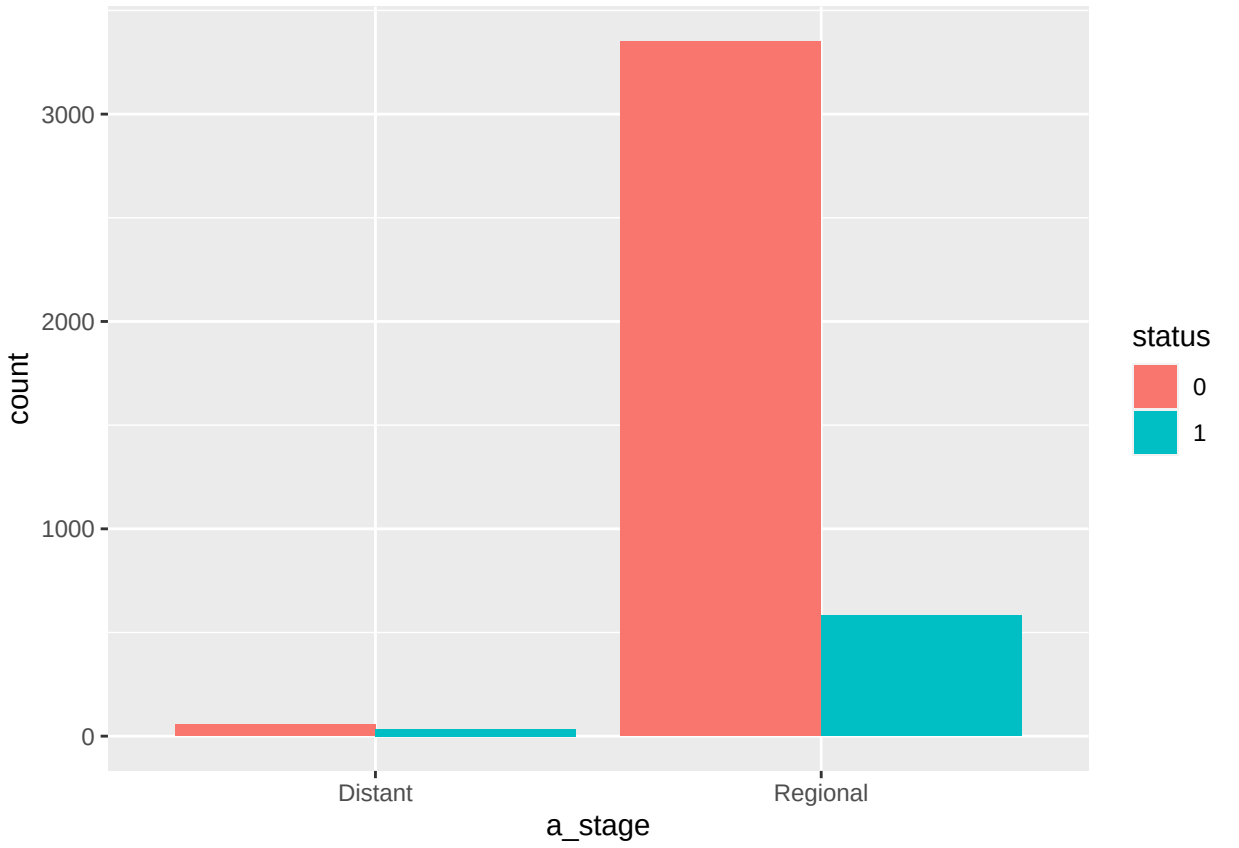

```
ggplot(breastcancer_clean, aes(x = differentiate, fill = status)) +
  geom_bar(position = "dodge")
```



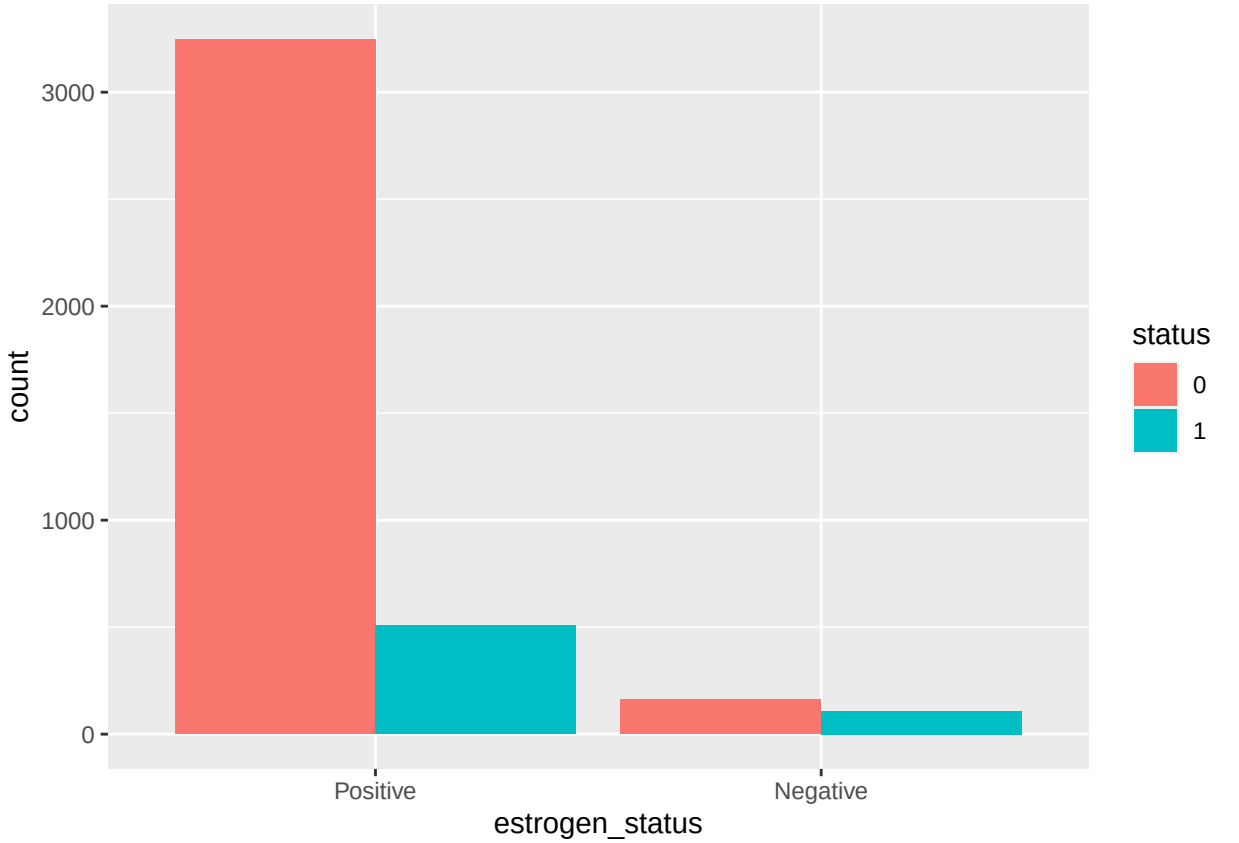

```
ggplot(breastcancer_clean, aes(x = grade, fill = status)) +
  geom_bar(position = "dodge")
```



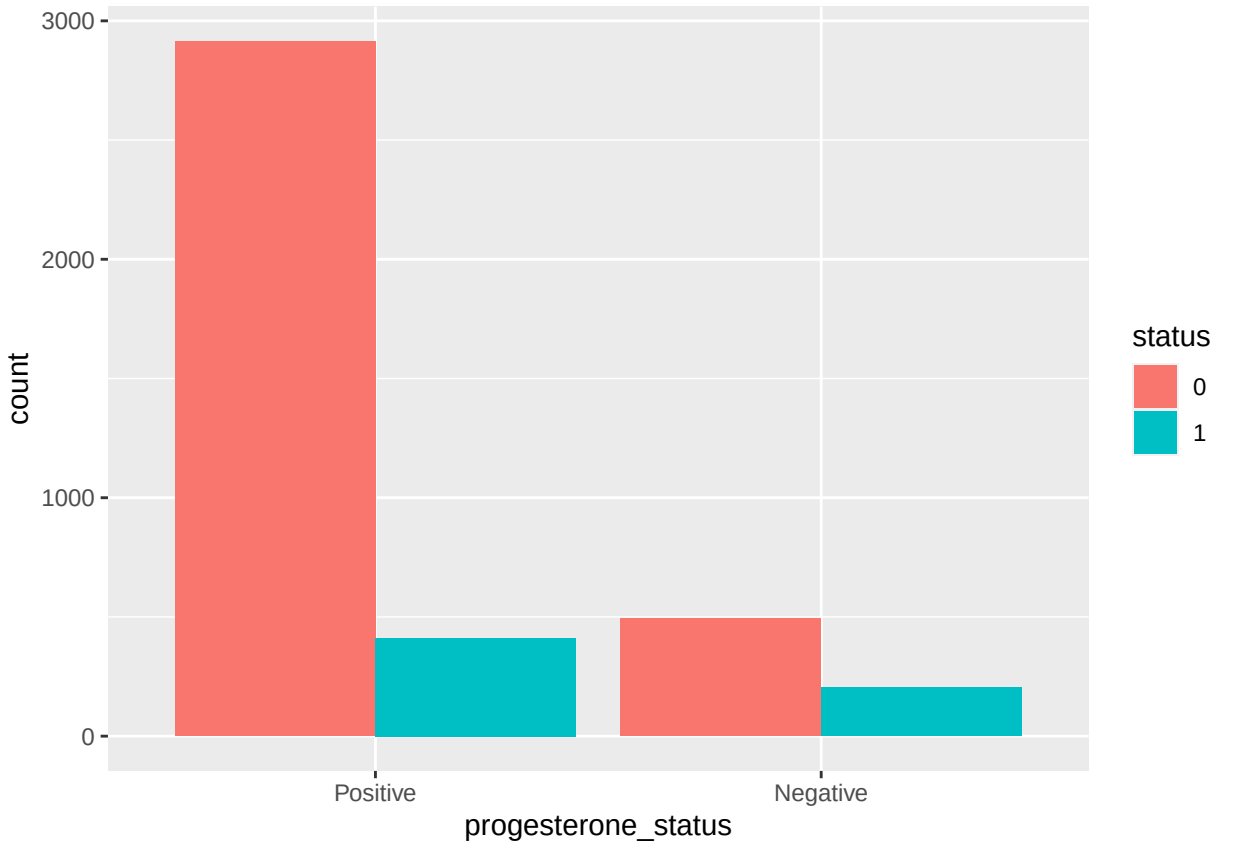

```
ggplot(breastcancer_clean, aes(x = a_stage, fill = status)) +
  geom_bar(position = "dodge")
```



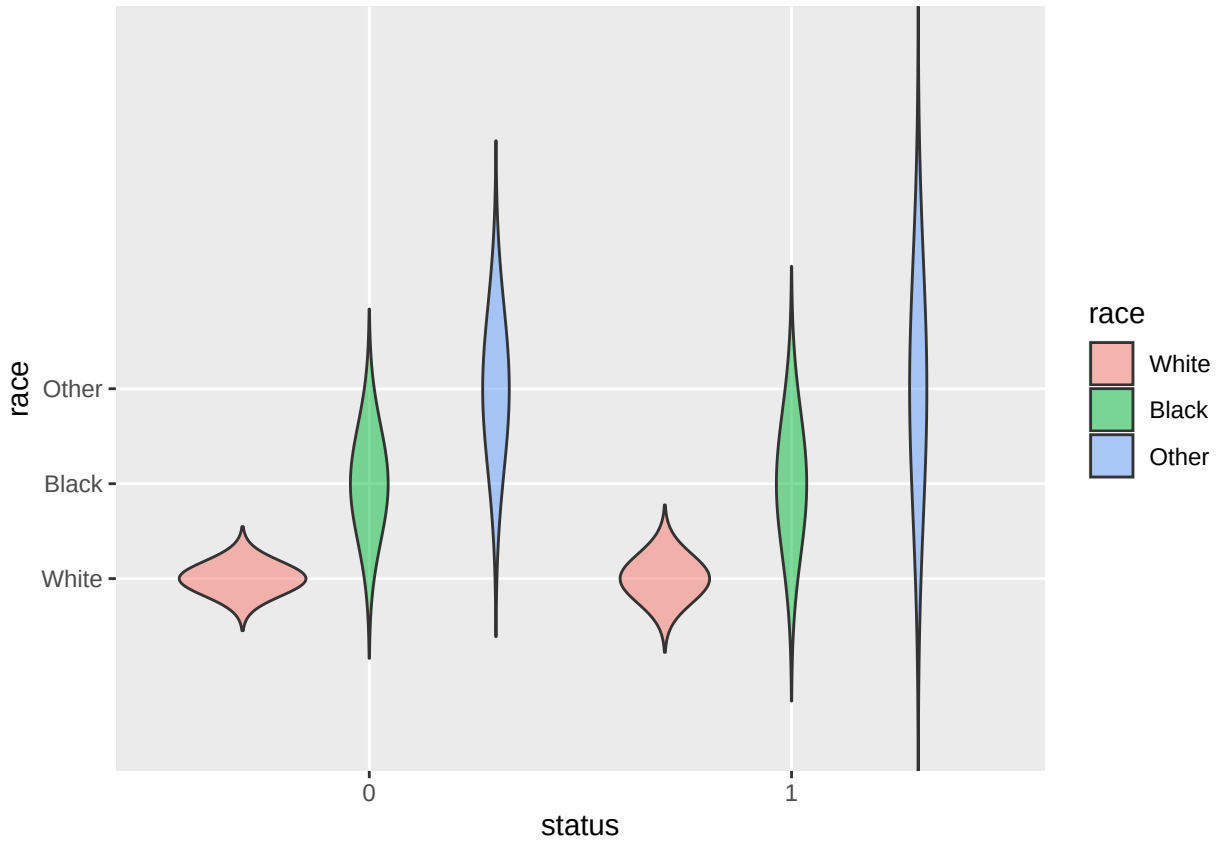

```
ggplot(breastcancer_clean, aes(x = estrogen_status, fill = status)) +
  geom_bar(position = "dodge")
```



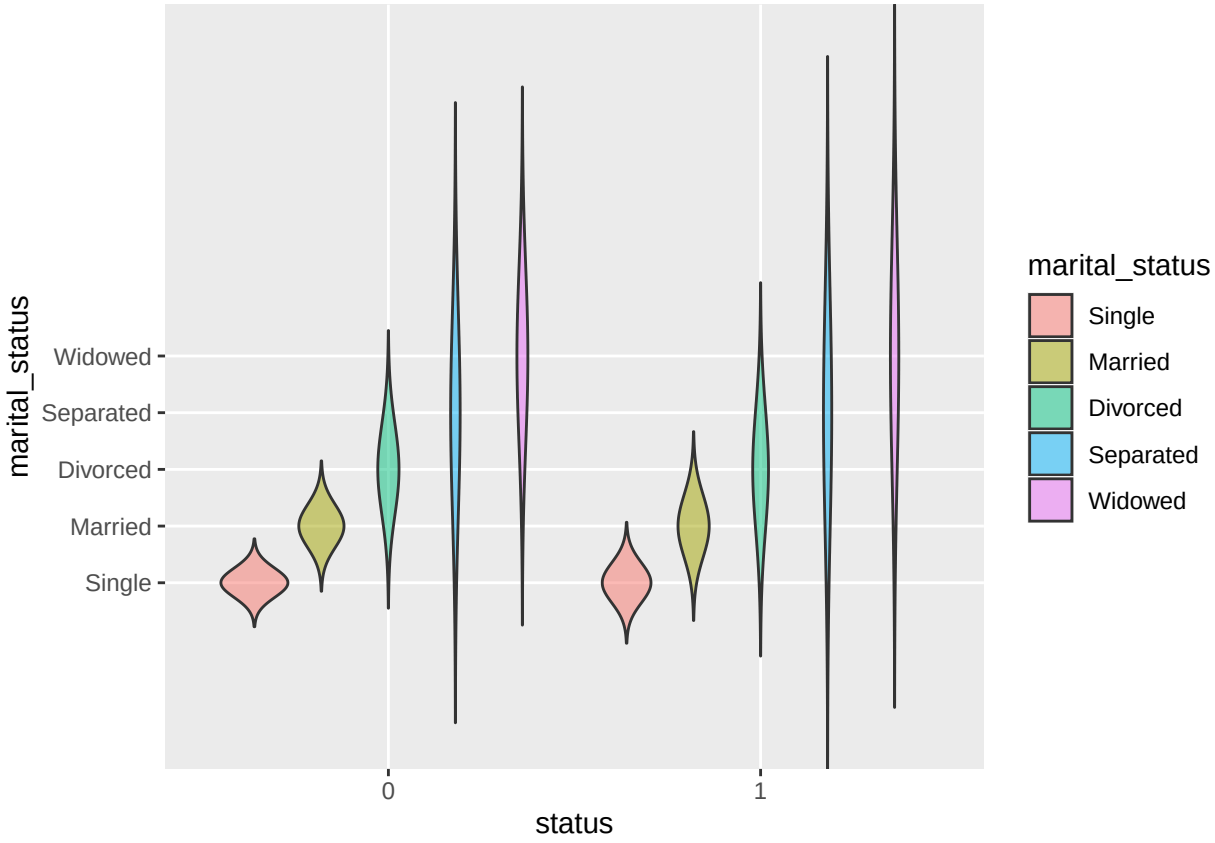

```
ggplot(breastcancer_clean, aes(x = progesterone_status, fill = status)) +
  geom_bar(position = "dodge")
```



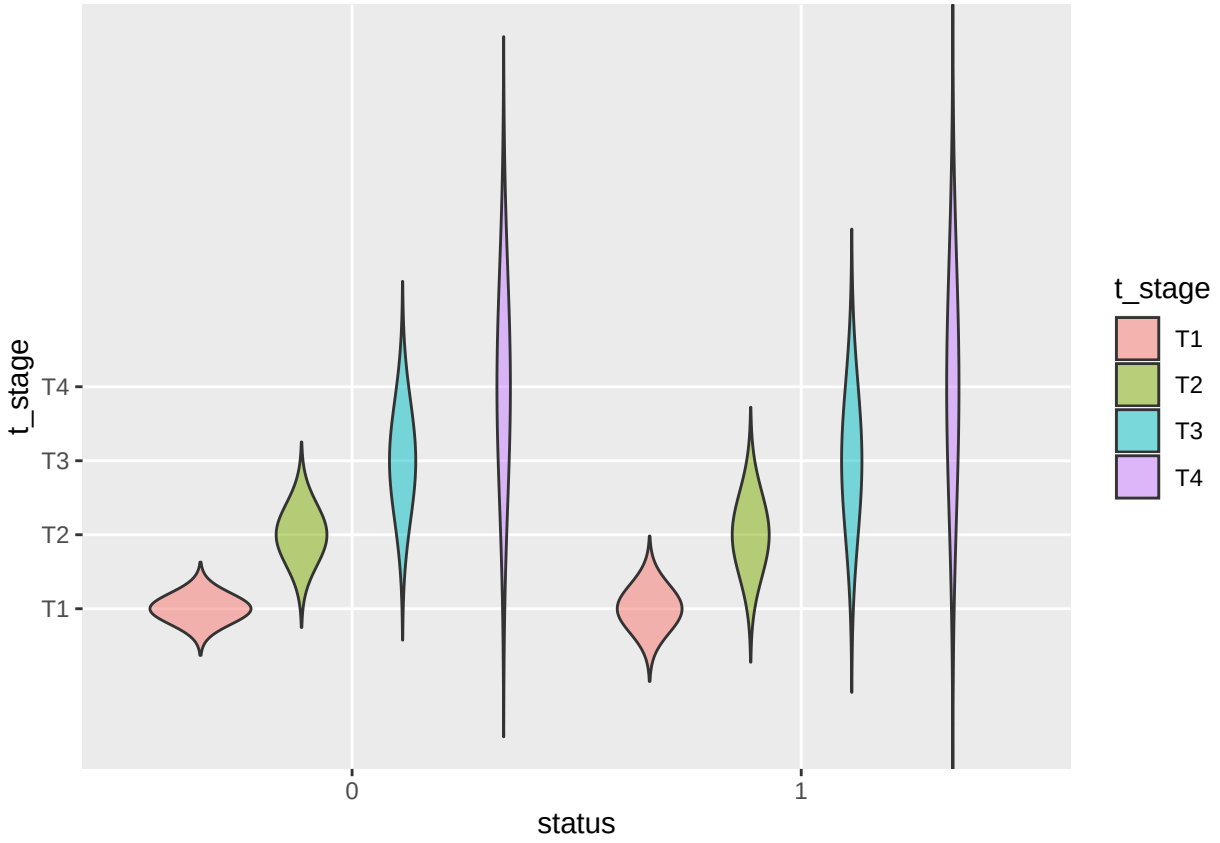

```
# Violin plots
ggplot(breastcancer_clean, aes(x = status, y = race)) +
  geom_violin(aes(fill = race), alpha = .5, trim = FALSE)
```



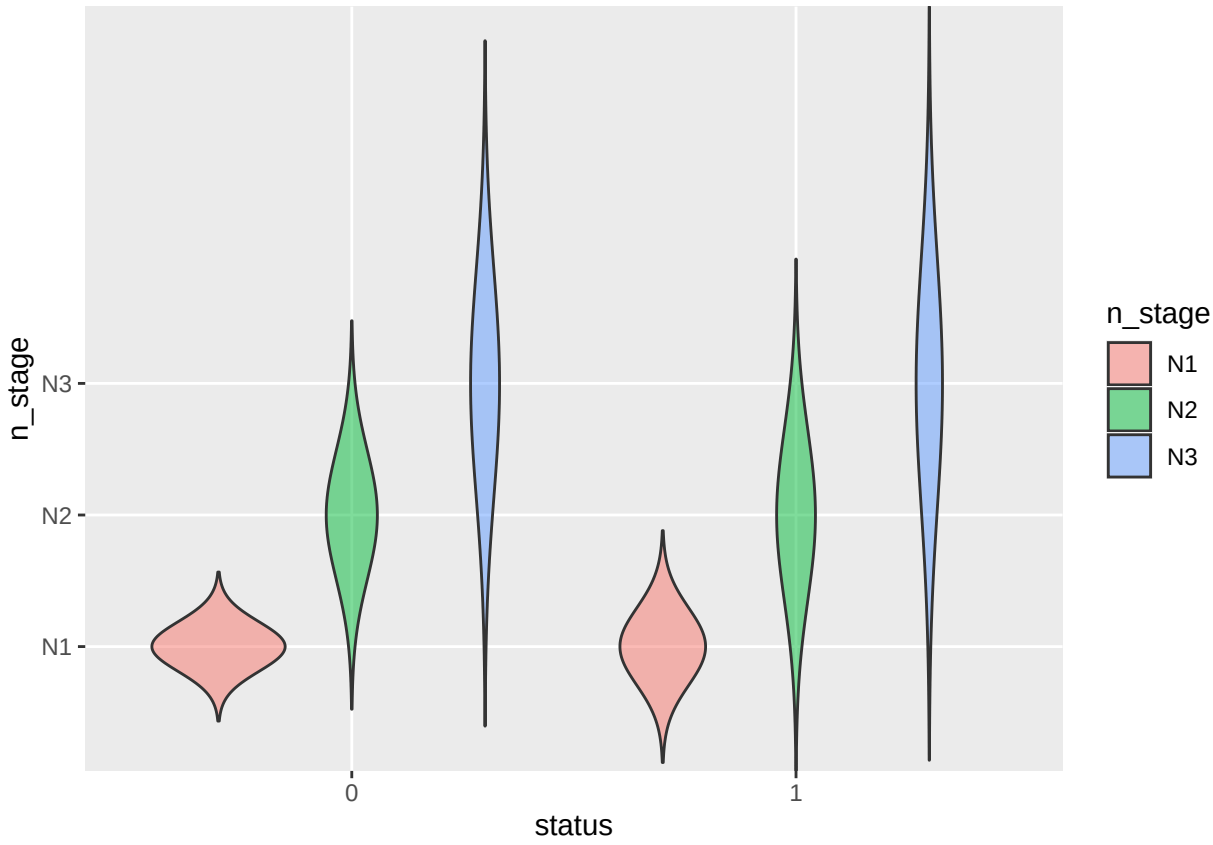

```
ggplot(breastcancer_clean, aes(x = status, y = marital_status)) +
  geom_violin(aes(fill = marital_status), alpha = .5, trim = FALSE)
```



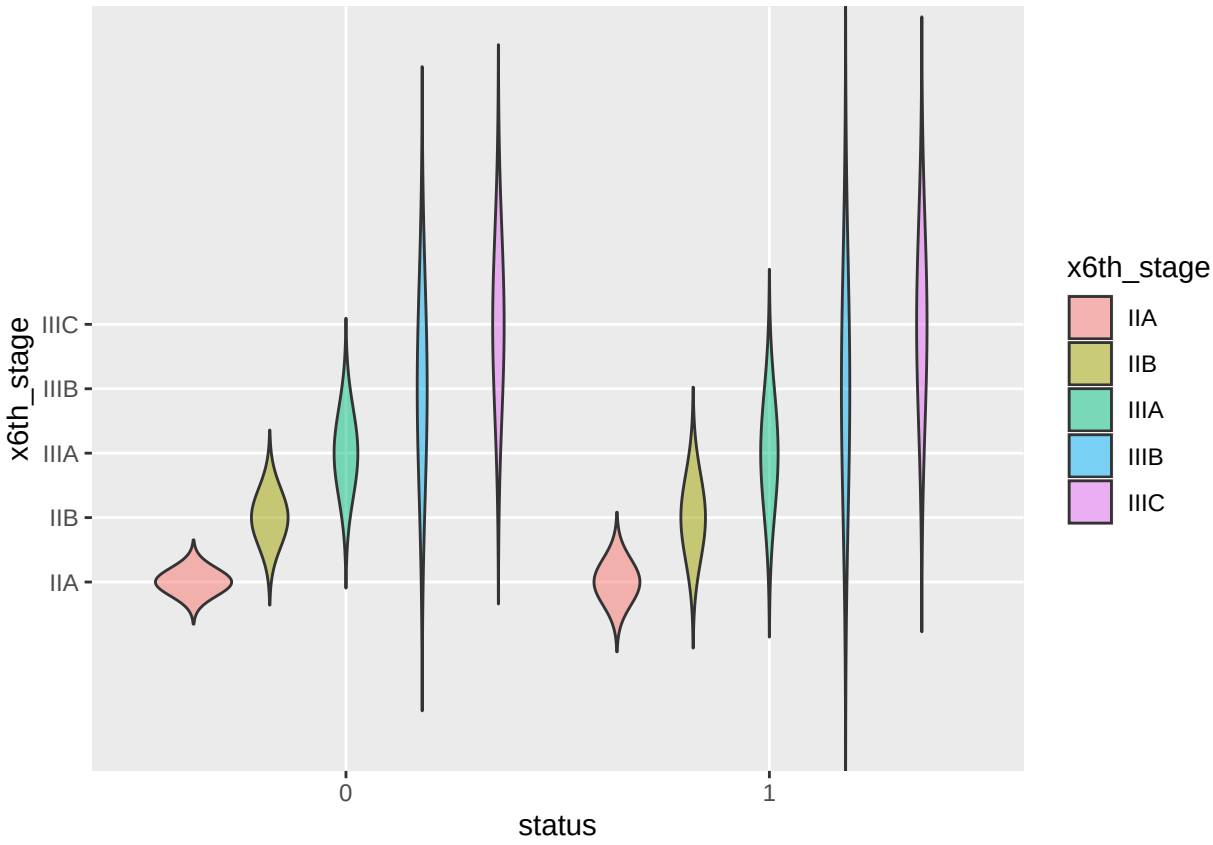

```
ggplot(breastcancer_clean, aes(x = status, y = t_stage)) +
  geom_violin(aes(fill = t_stage), alpha = .5, trim = FALSE)
```



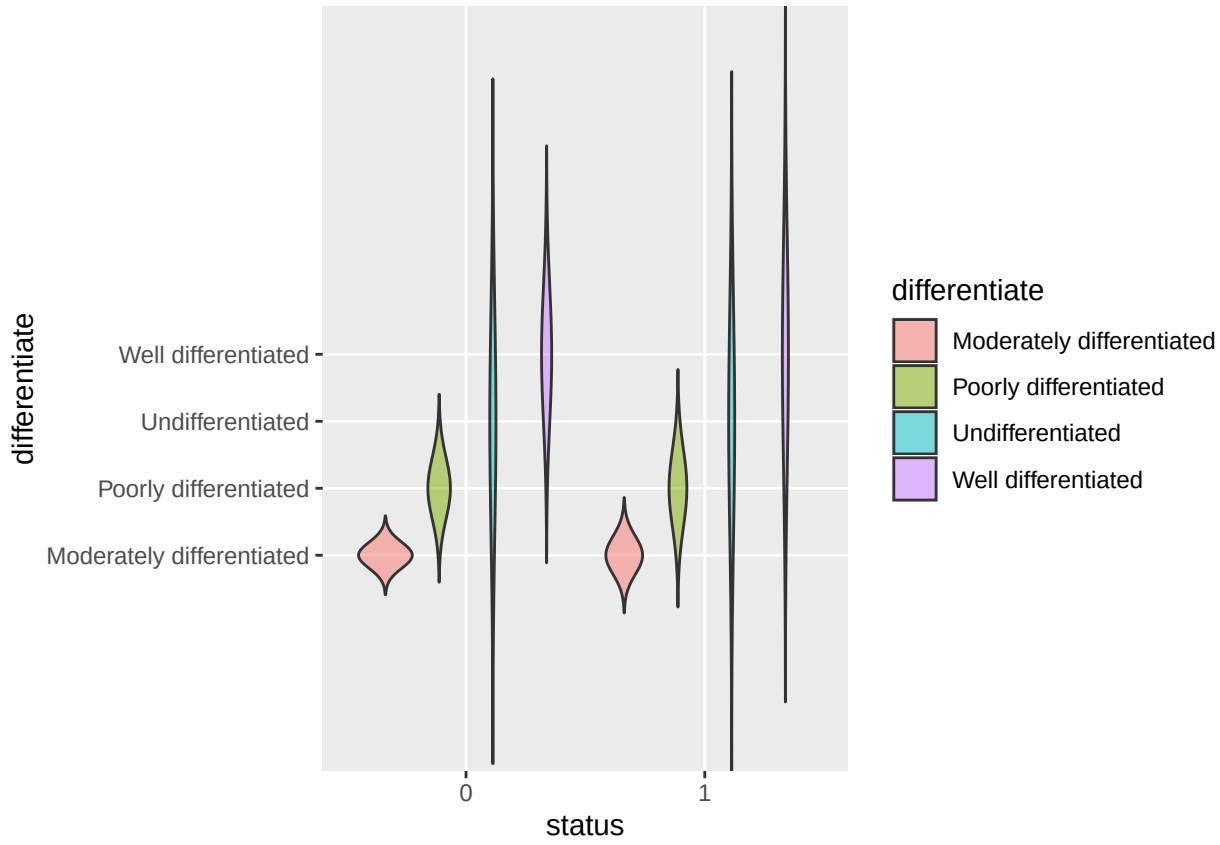

```
ggplot(breastcancer_clean, aes(x = status, y = n_stage)) +
  geom_violin(aes(fill = n_stage), alpha = .5, trim = FALSE)
```



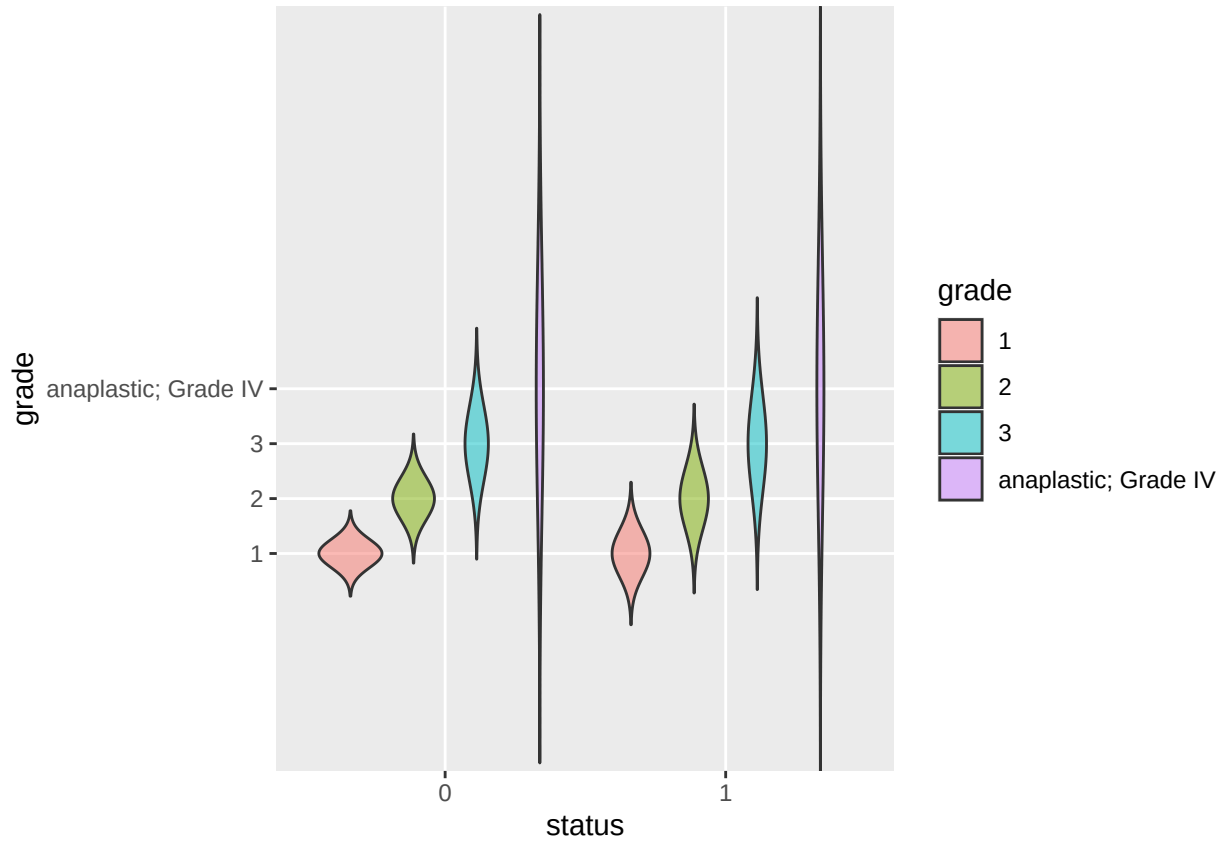

```
ggplot(breastcancer_clean, aes(x = status, y = x6th_stage)) +
  geom_violin(aes(fill = x6th_stage), alpha = .5, trim = FALSE)
```



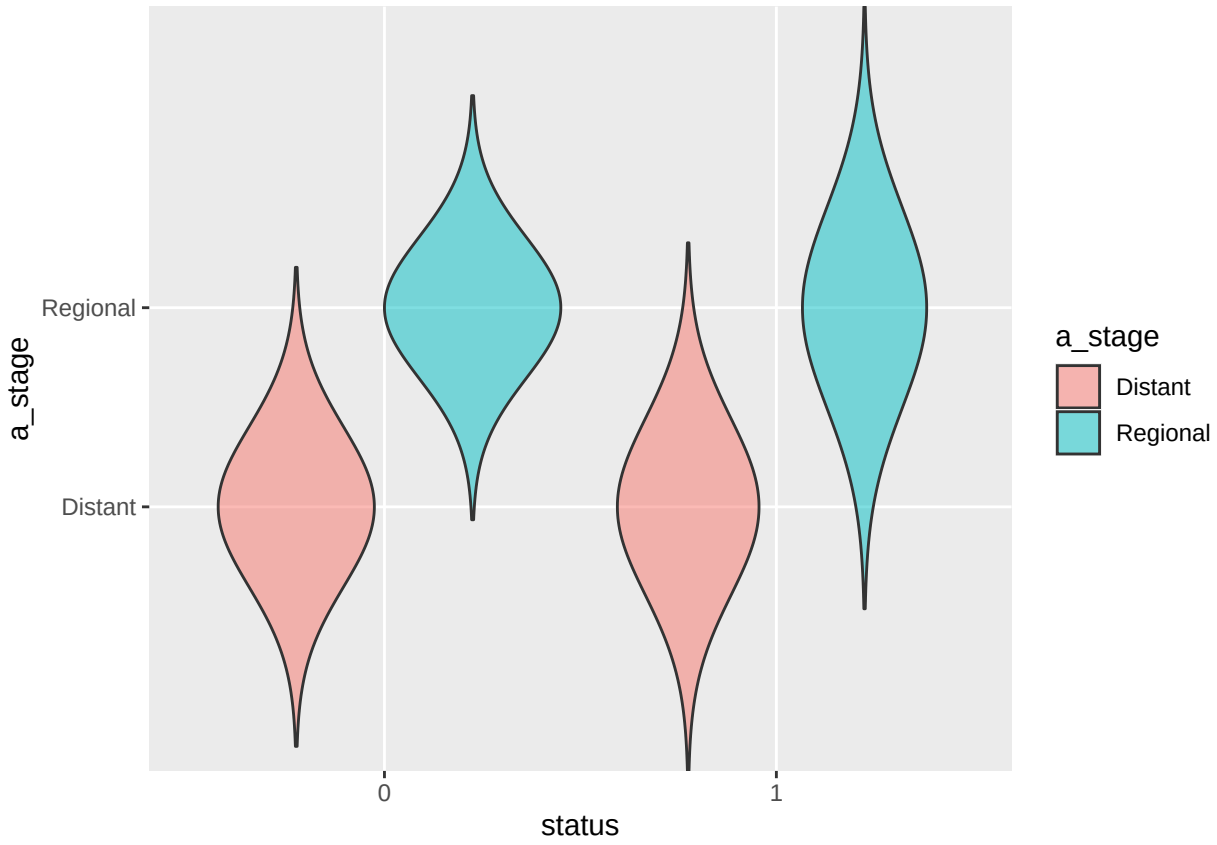

```
ggplot(breastcancer_clean, aes(x = status, y = differentiate)) +
  geom_violin(aes(fill = differentiate), alpha = .5, trim = FALSE)
```



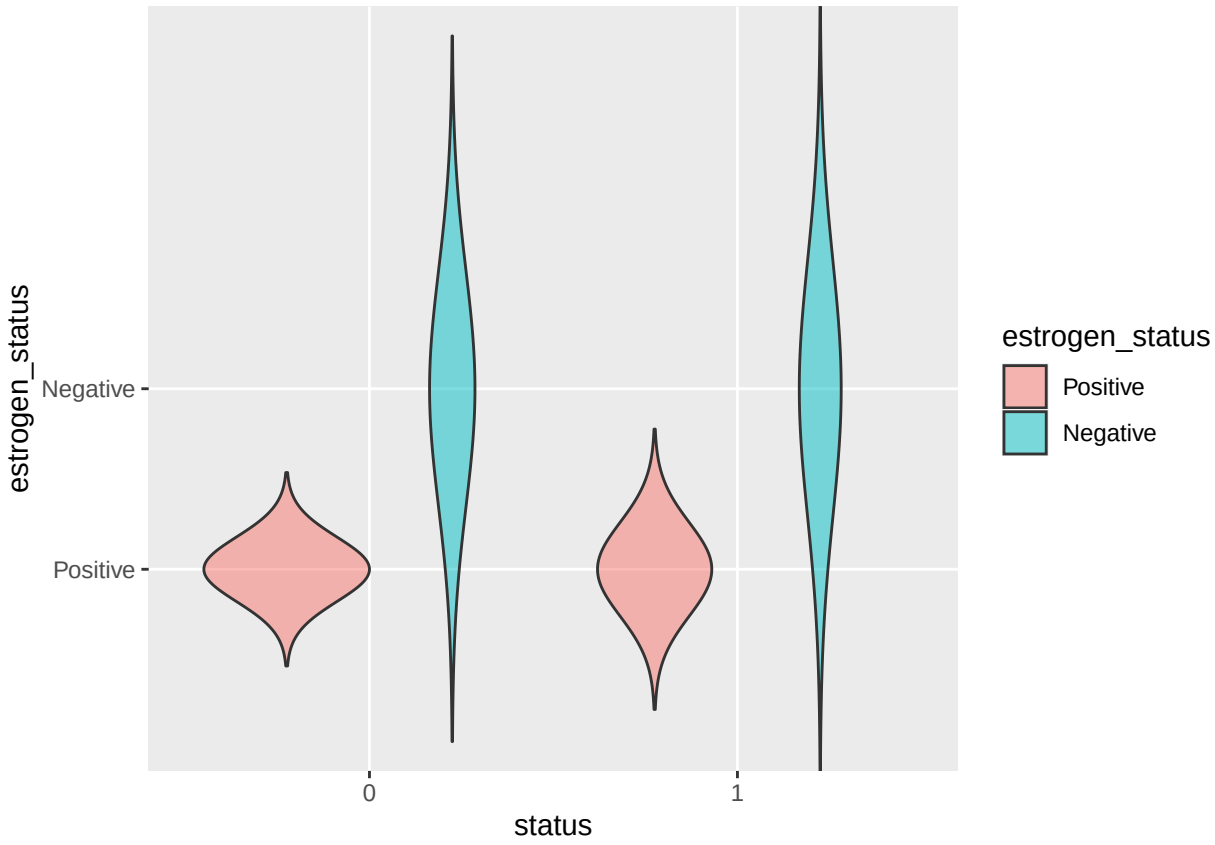

```
ggplot(breastcancer_clean, aes(x = status, y = grade)) +
  geom_violin(aes(fill = grade), alpha = .5, trim = FALSE)
```



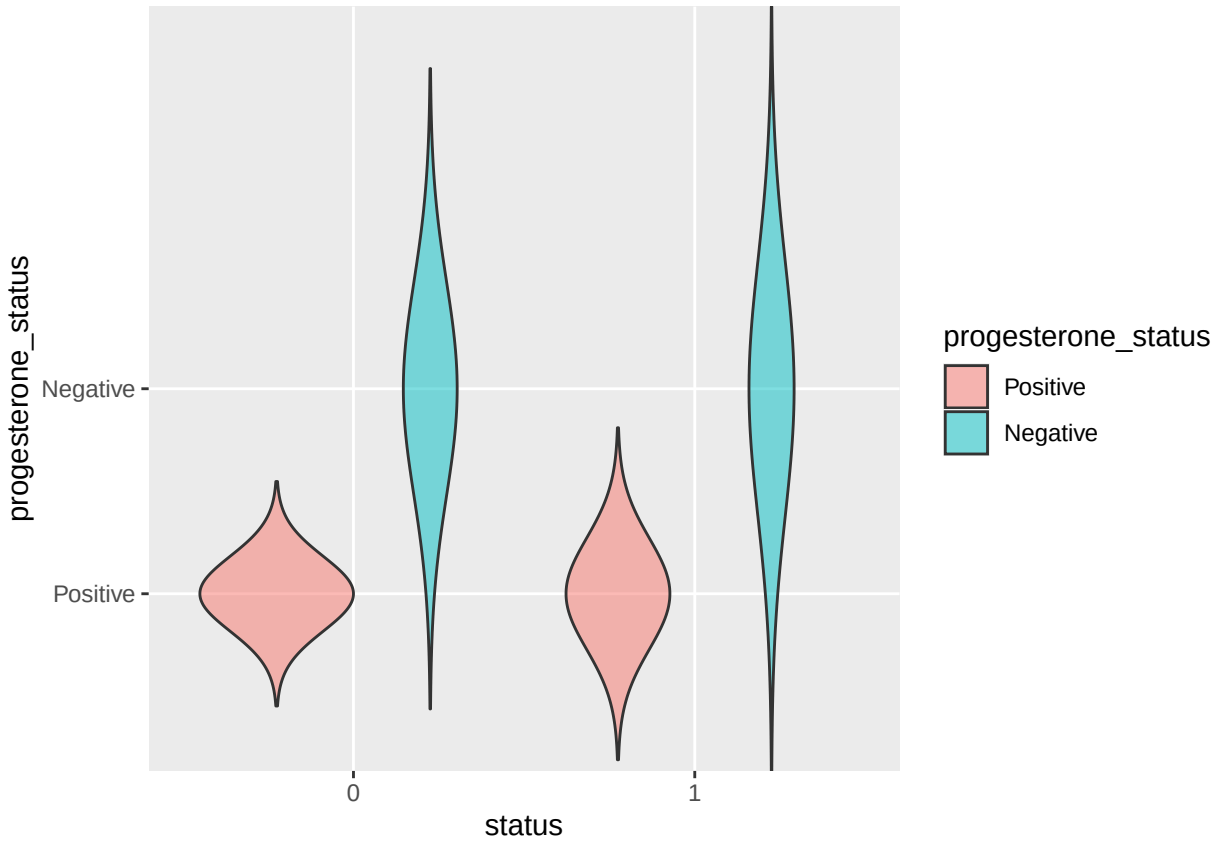

```
ggplot(breastcancer_clean, aes(x = status, y = a_stage)) +
  geom_violin(aes(fill = a_stage), alpha = .5, trim = FALSE)
```




```
ggplot(breastcancer_clean, aes(x = status, y = estrogen_status)) +
  geom_violin(aes(fill = estrogen_status), alpha = .5, trim = FALSE)
```




```
ggplot(breastcancer_clean, aes(x = status, y = progesterone_status)) +
  geom_violin(aes(fill = progesterone_status), alpha = .5, trim = FALSE)
```

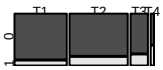



```
# Mosaic Plots
par(mfrow = c(3,4))
mosaicplot(table(breastcancer_clean$race, breastcancer_clean$status), main="Race vs Status", color = TRU
mosaicplot(table(breastcancer_clean$marital_status, breastcancer_clean$status), main="Marital Status vs
mosaicplot(table(breastcancer_clean$t_stage, breastcancer_clean$status), main="T-stage vs Status", colo
mosaicplot(table(breastcancer_clean$n_stage, breastcancer_clean$status), main="N-stage vs Status", colo
mosaicplot(table(breastcancer_clean$x6th_stage, breastcancer_clean$status), main="X6th Stage vs Status"
mosaicplot(table(breastcancer_clean$differentiate, breastcancer_clean$status), main="Differentiate vs S
mosaicplot(table(breastcancer_clean$grade, breastcancer_clean$status), main="Grade vs Status", color = 
mosaicplot(table(breastcancer_clean$a_stage, breastcancer_clean$status), main="A-stage vs Status", colo
mosaicplot(table(breastcancer_clean$estrogen_status, breastcancer_clean$status), main="Estrogen Status 
mosaicplot(table(breastcancer_clean$progesterone_status, breastcancer_clean$status), main="Progesterone
```

Race vs Status

Marital Status vs Statu

T−stage vs Status

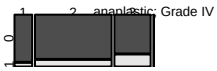


N−stage vs Status

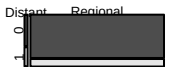
X6th Stage vs Status

Differentiate vs Status

Grade vs Status



A−stage vs Status



Estrogen Status vs Sta

Progesterone Status vs S