DSCI 445 Project

Group 6

2025-10-24

```
data <- read.csv("bank-full[1].csv", sep = ";")
str(data)
```

```
## 'data.frame':    45211 obs. of  17 variables:
##  $ age      : int  58 44 33 47 33 35 28 42 58 43 ...
##  $ job      : chr  "management" "technician" "entrepreneur" "blue-collar" ...
##  $ marital  : chr  "married" "single" "married" "married" ...
##  $ education: chr  "tertiary" "secondary" "secondary" "unknown" ...
##  $ default  : chr  "no" "no" "no" "no" ...
##  $ balance  : int  2143 29 2 1506 1 231 447 2 121 593 ...
##  $ housing  : chr  "yes" "yes" "yes" "yes" ...
##  $ loan     : chr  "no" "no" "yes" "no" ...
##  $ contact  : chr  "unknown" "unknown" "unknown" "unknown" ...
##  $ day      : int  5 5 5 5 5 5 5 5 5 5 ...
##  $ month    : chr  "may" "may" "may" "may" ...
##  $ duration : int  261 151 76 92 198 139 217 380 50 55 ...
##  $ campaign : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ pdays    : int  -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 ...
##  $ previous : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ poutcome : chr  "unknown" "unknown" "unknown" "unknown" ...
##  $ y        : chr  "no" "no" "no" "no" ...
```

```
summary(data)
```

```
##       age            job              marital           education
##  Min.   :18.00   Length:45211       Length:45211       Length:45211
##  1st Qu.:33.00   Class :character   Class :character   Class :character
##  Median :39.00   Mode  :character   Mode  :character   Mode  :character
##  Mean   :40.94
##  3rd Qu.:48.00
##  Max.   :95.00
##    default             balance         housing              loan
##  Length:45211       Min.   : -8019   Length:45211       Length:45211
##  Class :character   1st Qu.:    72   Class :character   Class :character
##  Mode  :character   Median :   448   Mode  :character   Mode  :character
##                     Mean   :  1362
##                     3rd Qu.:  1428
##                     Max.   :102127
##    contact              day           month              duration
##  Length:45211       Min.   : 1.00   Length:45211       Min.   :   0.0
##  Class :character   1st Qu.: 8.00   Class :character   1st Qu.: 103.0
##  Mode  :character   Median :16.00   Mode  :character   Median : 180.0
##                     Mean   :15.81                      Mean   : 258.2
```

```
##                     3rd Qu.:21.00                       3rd Qu.: 319.0
##                     Max.   :31.00                       Max.   :4918.0
##     campaign          pdays          previous          poutcome
##  Min.   : 1.000   Min.   : -1.0   Min.   :  0.0000   Length:45211
##  1st Qu.: 1.000   1st Qu.: -1.0   1st Qu.:  0.0000   Class :character
##  Median : 2.000   Median : -1.0   Median :  0.0000   Mode  :character
##  Mean   : 2.764   Mean   : 40.2   Mean   :  0.5803
##  3rd Qu.: 3.000   3rd Qu.: -1.0   3rd Qu.:  0.0000
##  Max.   :63.000   Max.   :871.0   Max.   :275.0000
##       y
##  Length:45211
##  Class :character
##  Mode  :character
##
##
##
```

```
data$y <- as.factor(data$y)
data$job <- as.factor(data$job)
data$marital <- as.factor(data$marital)
data$education <- as.factor(data$education)
data$contact <- as.factor(data$contact)
data$month <- as.factor(data$month)
data$poutcome <- as.factor(data$poutcome)
data$housing <- as.factor(data$housing)
data$loan <- as.factor(data$loan)
```

```
numeric_vars <- c("age", "balance", "duration", "campaign", "pdays", "previous")
summary(data[numeric_vars])
```

```
##       age           balance          duration         campaign
##  Min.   :18.00   Min.   : -8019   Min.   :   0.0   Min.   : 1.000
##  1st Qu.:33.00   1st Qu.:    72   1st Qu.: 103.0   1st Qu.: 1.000
##  Median :39.00   Median :   448   Median : 180.0   Median : 2.000
##  Mean   :40.94   Mean   :  1362   Mean   : 258.2   Mean   : 2.764
##  3rd Qu.:48.00   3rd Qu.:  1428   3rd Qu.: 319.0   3rd Qu.: 3.000
##  Max.   :95.00   Max.   :102127   Max.   :4918.0   Max.   :63.000
##      pdays          previous
##  Min.   : -1.0   Min.   :  0.0000
##  1st Qu.: -1.0   1st Qu.:  0.0000
##  Median : -1.0   Median :  0.0000
##  Mean   : 40.2   Mean   :  0.5803
##  3rd Qu.: -1.0   3rd Qu.:  0.0000
##  Max.   :871.0   Max.   :275.0000
```

```
colSums(data == "unknown")
```

```
##       age       job   marital education   default   balance   housing      loan
##         0       288         0      1857         0         0         0         0
##   contact       day     month  duration  campaign     pdays  previous  poutcome
##     13020         0         0         0         0         0         0     36959
##         y
##         0
```

```
sum(data$pdays == -1)
```

```
## [1] 36954
```

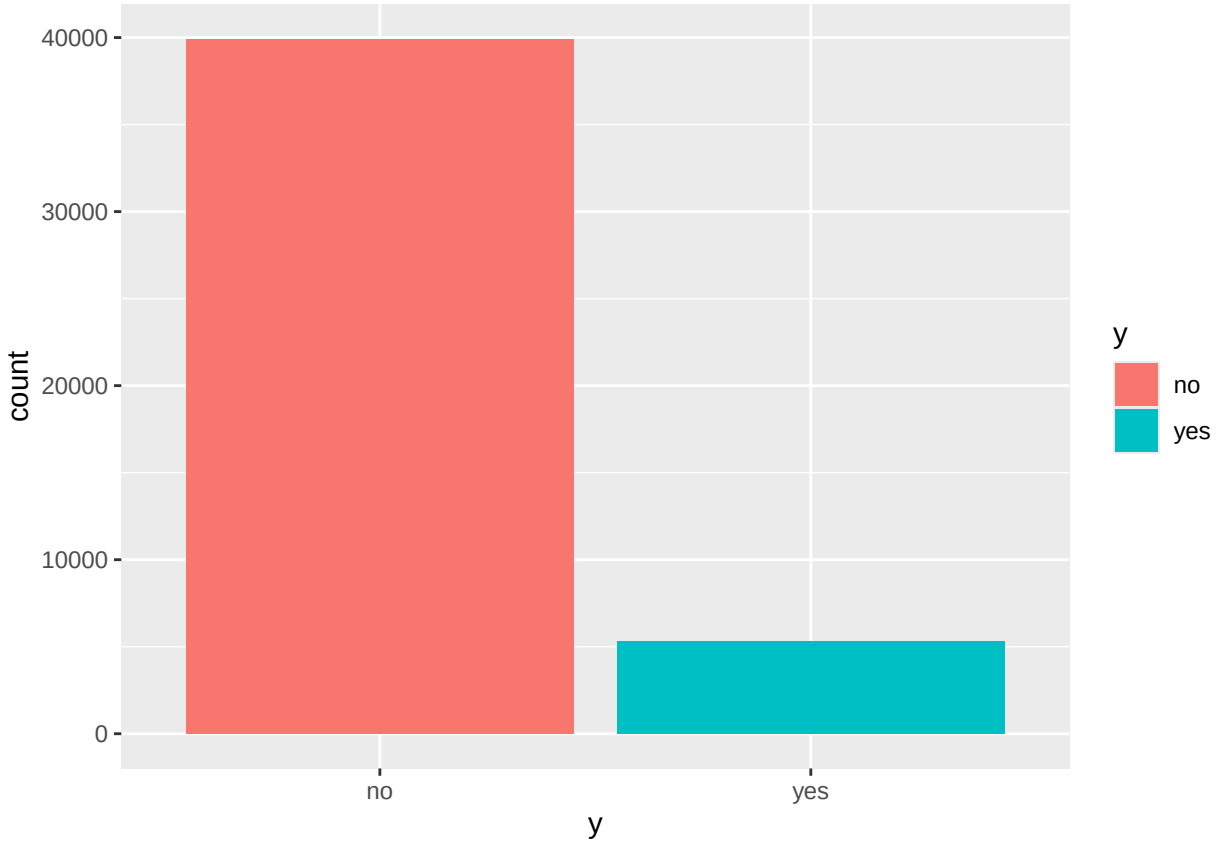
```r
library(ggplot2)

table(data$y)
```

```
##
##    no   yes
## 39922  5289
```

```r
prop.table(table(data$y))
```

```
##
##        no       yes
## 0.8830152 0.1169848
```

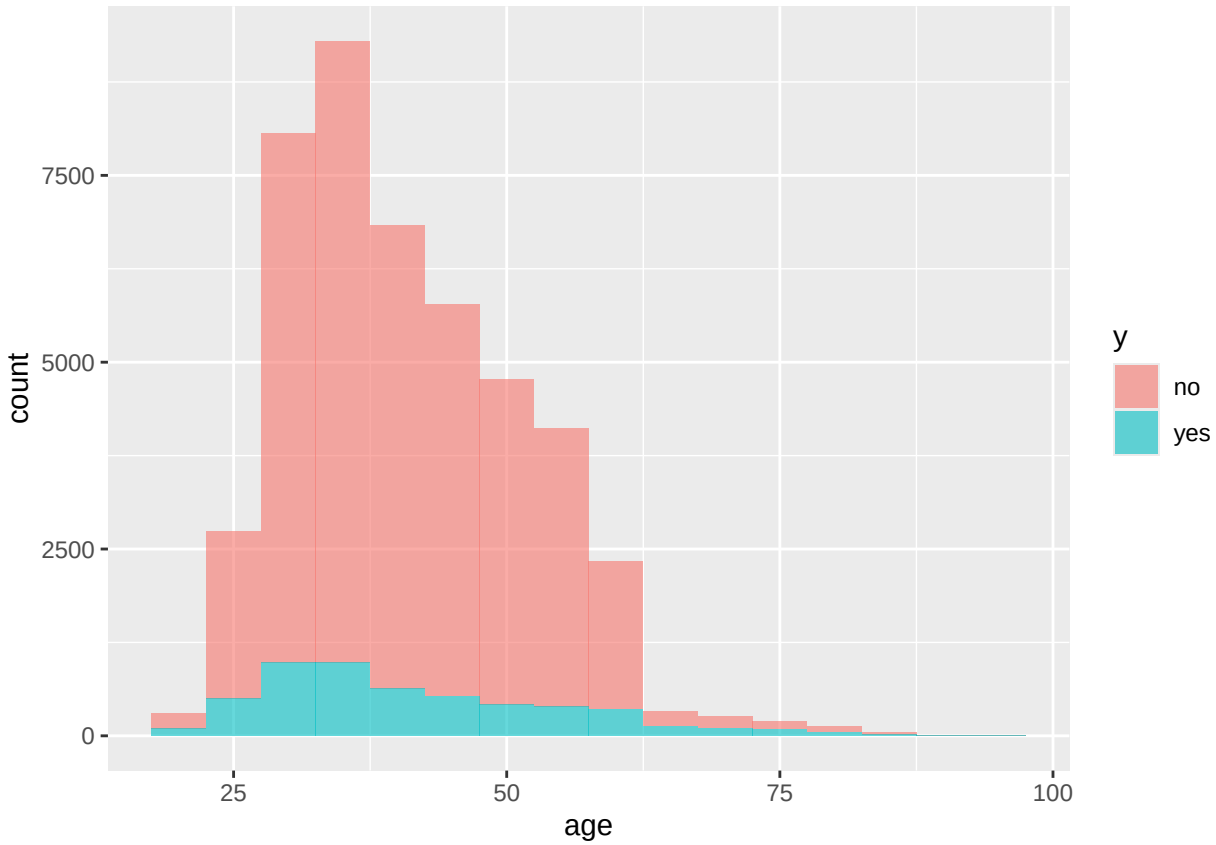
```r
ggplot(data, aes(x = y, fill = y)) + geom_bar()
```




```r
success_rate <- mean(data$y == "yes") * 100
cat("Success rate:", round(success_rate, 2), "%\n")
```

```
## Success rate: 11.7 %
```

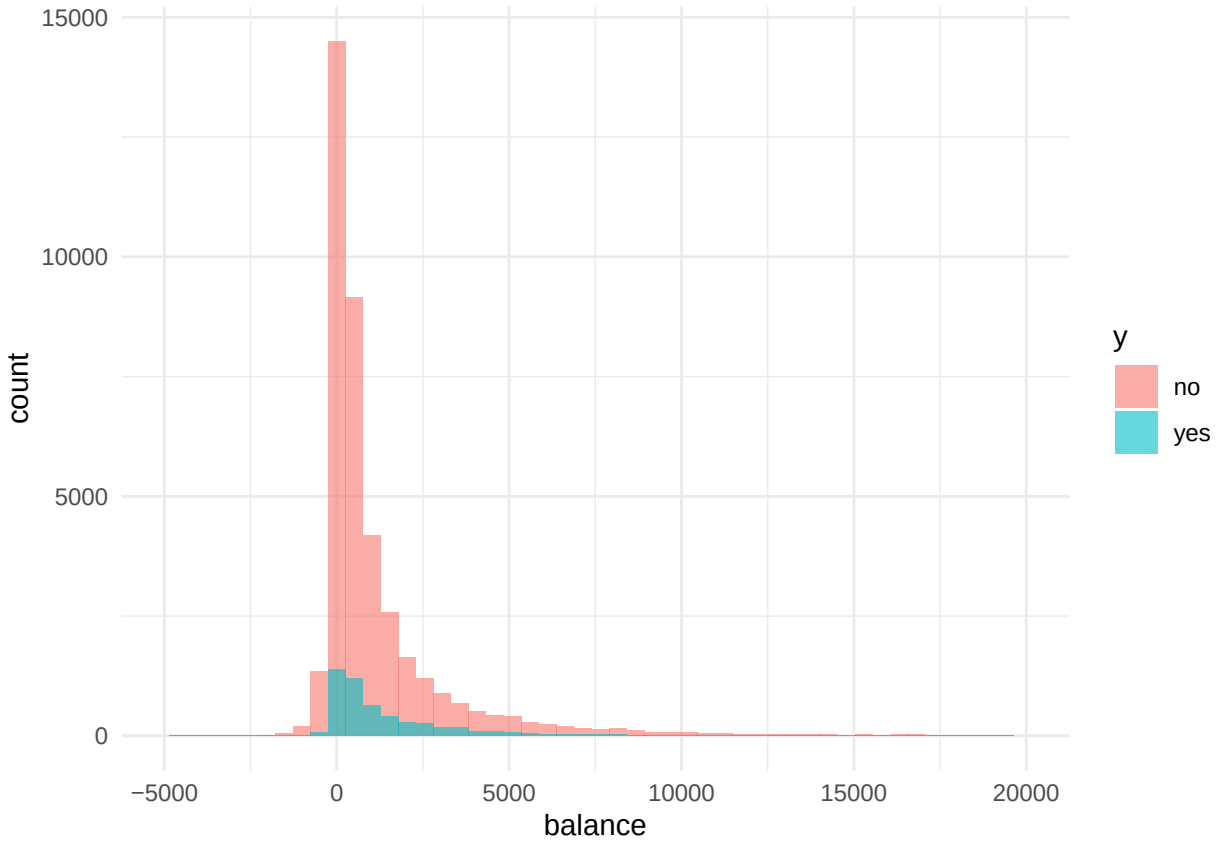
```r
ggplot(data, aes(x = age, fill = y)) + geom_histogram(binwidth = 5, alpha = 0.6)
```



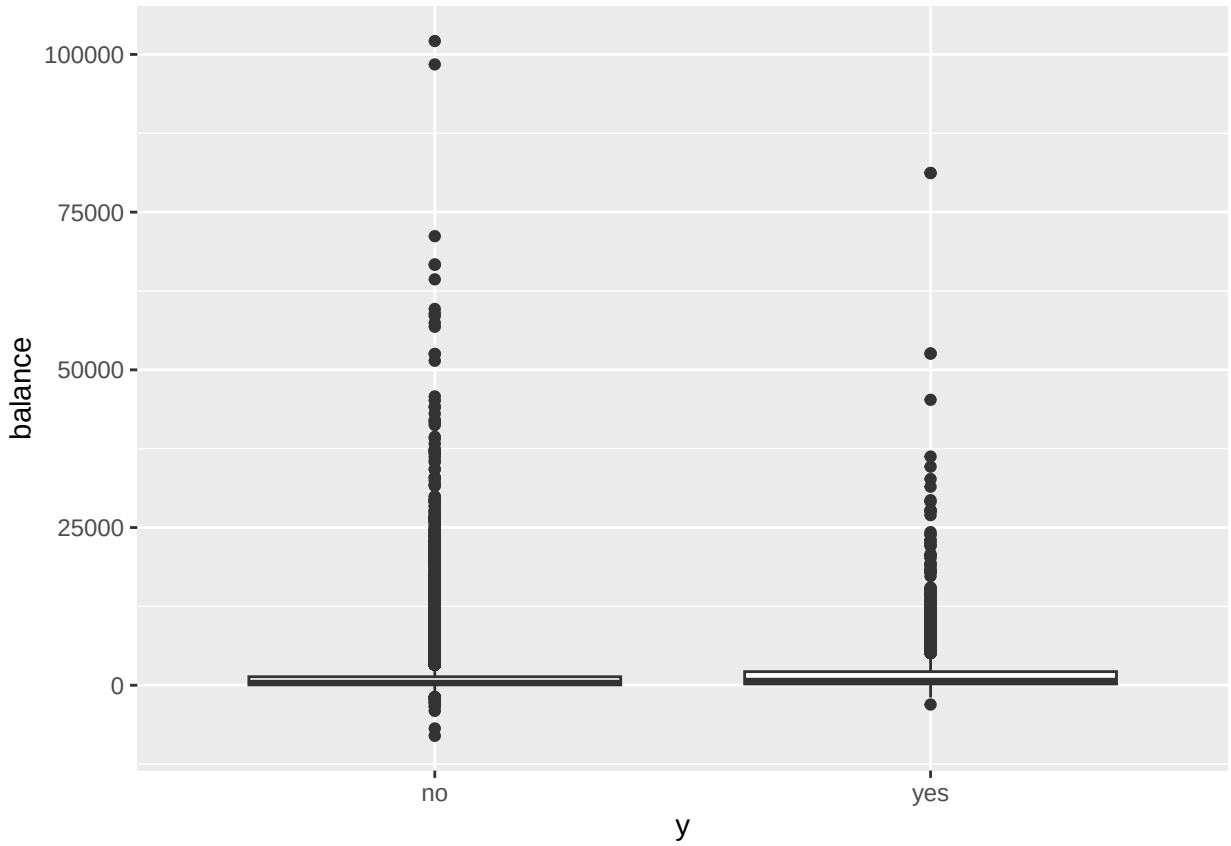

```
ggplot(data, aes(x = balance, fill = y)) +
  geom_histogram(bins = 50, alpha = 0.6, position = "identity") +
  scale_x_continuous(limits = c(-5000, 20000)) +
  theme_minimal()
```

```
## Warning: Removed 195 rows containing non-finite outside the scale range
## (`stat_bin()`).
```

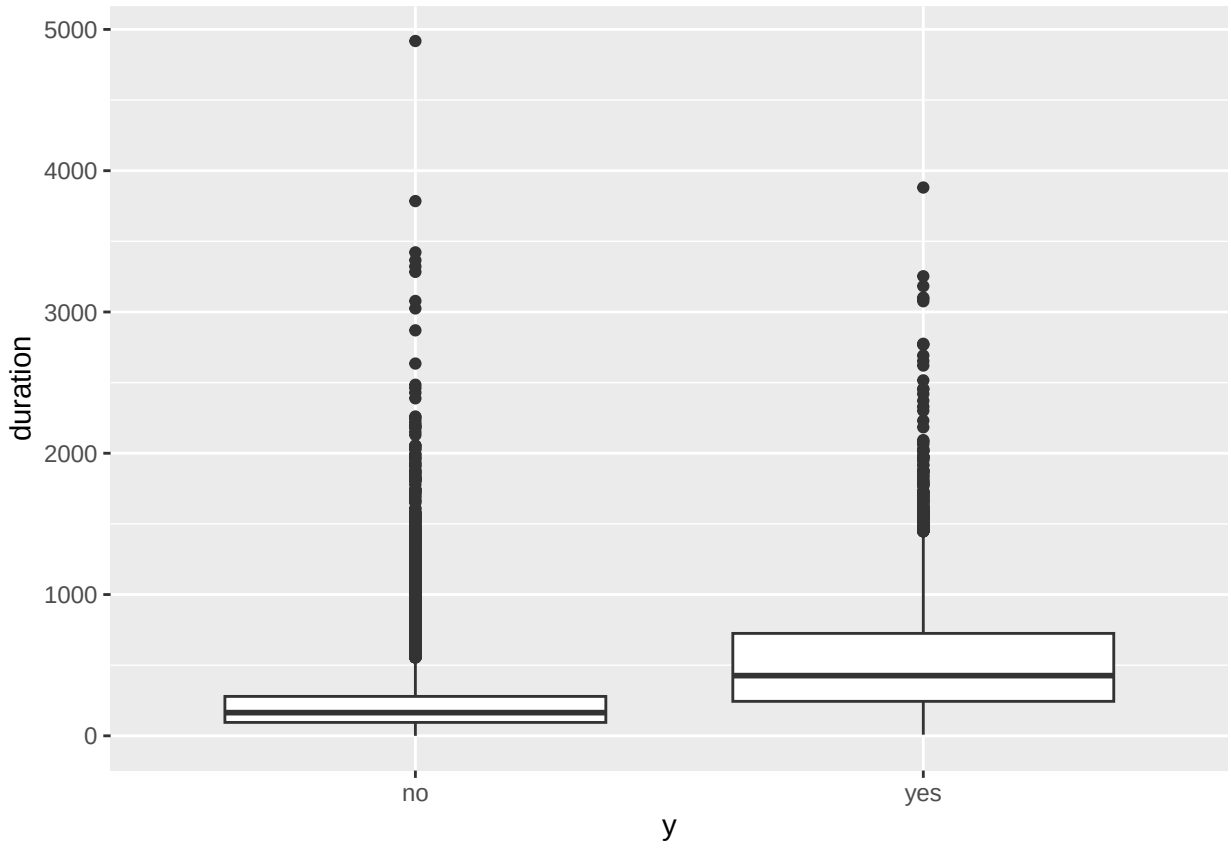
```
## Warning: Removed 4 rows containing missing values or values outside the scale range
## (`geom_bar()`).
```




```
ggplot(data, aes(y = balance, x = y)) + geom_boxplot()
```



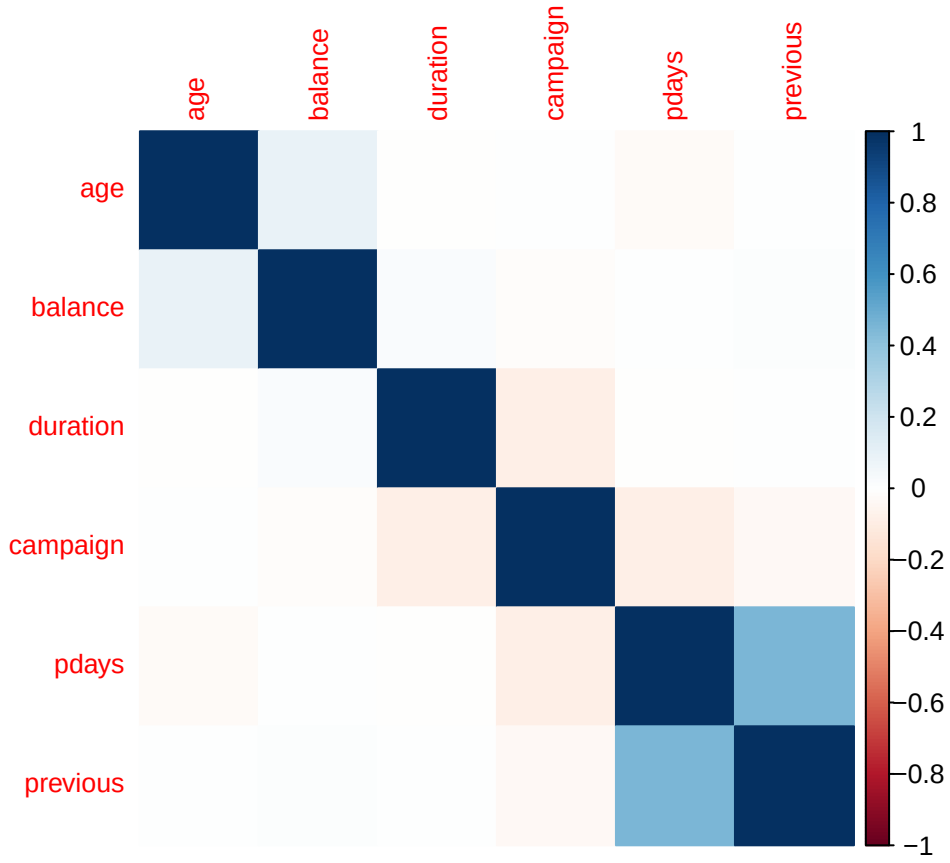

```
ggplot(data, aes(x = y, y = duration)) + geom_boxplot()
```



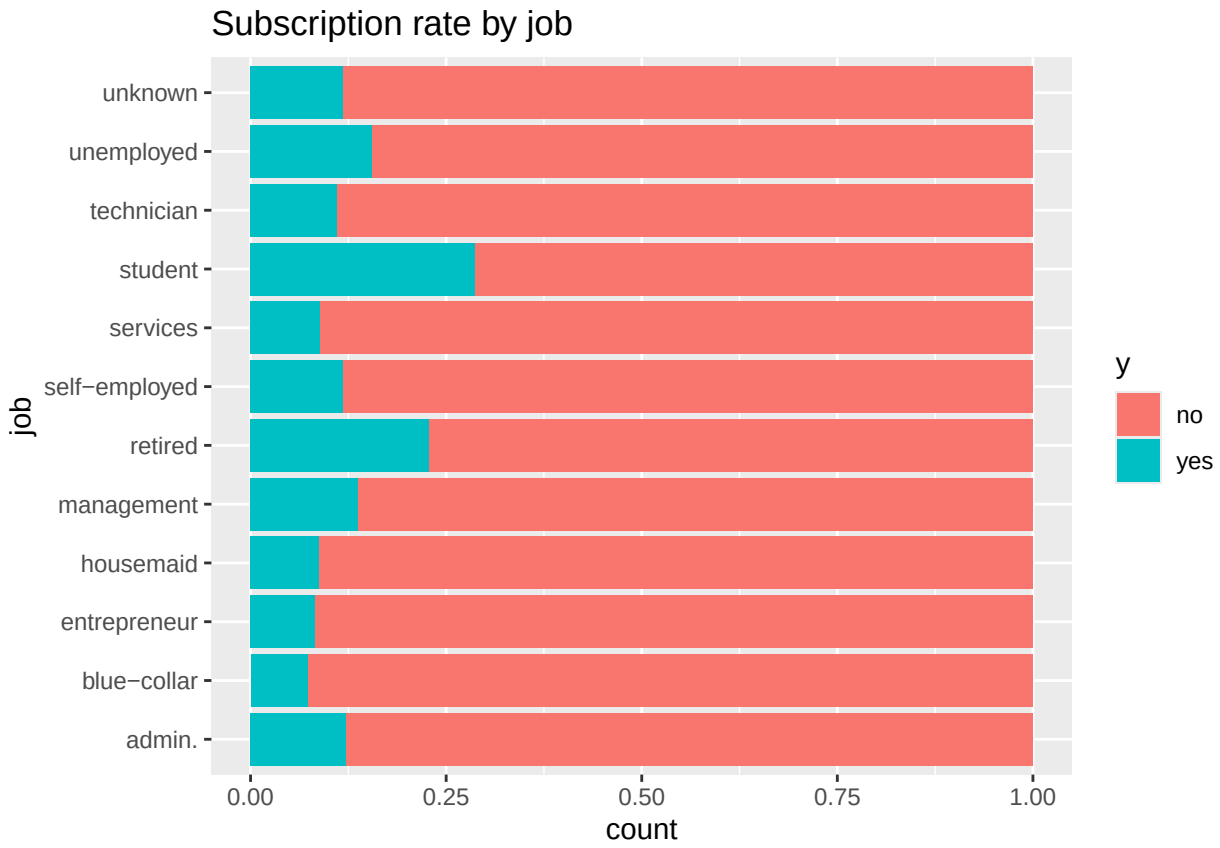

```
library(corrplot)
```

```
## corrplot 0.95 loaded
```

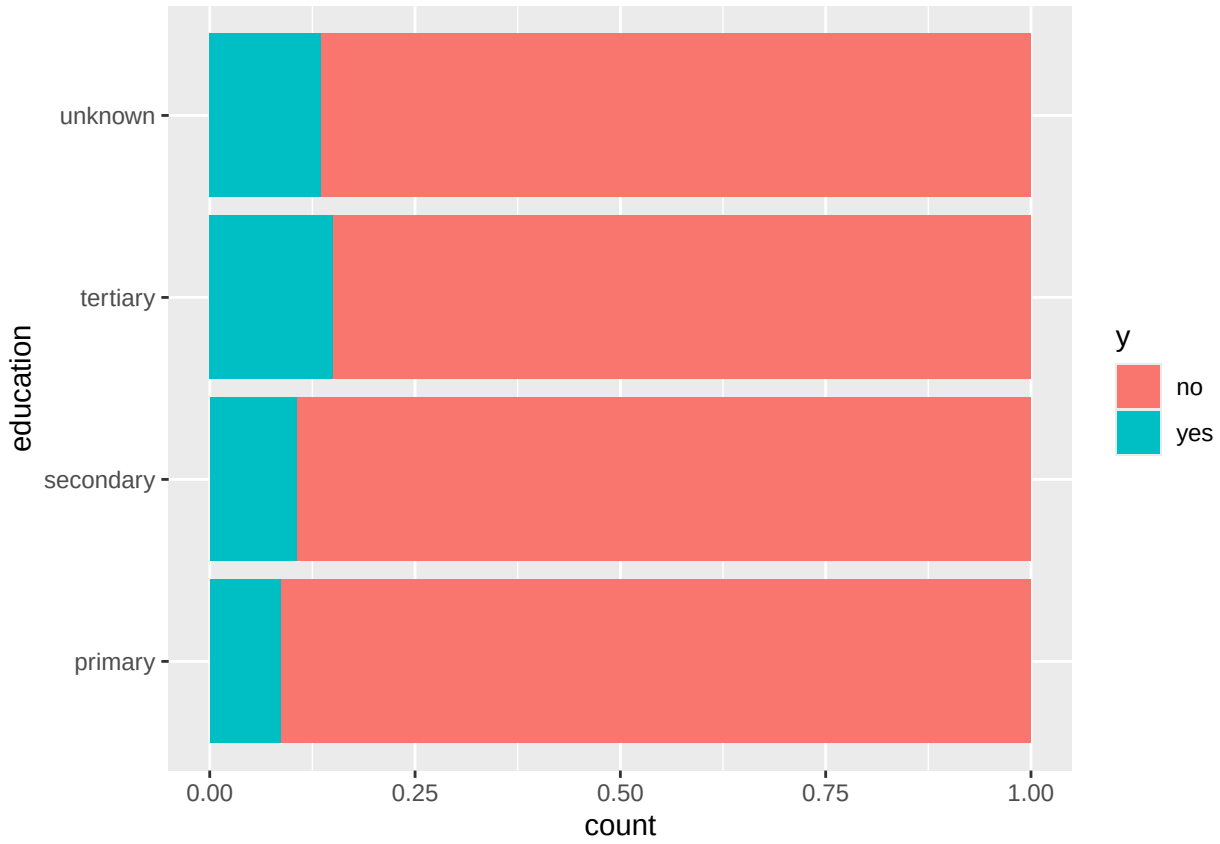
```
nums <- data[numeric_vars]
corrplot(cor(nums), method = "color", tl.cex = 0.8)
```



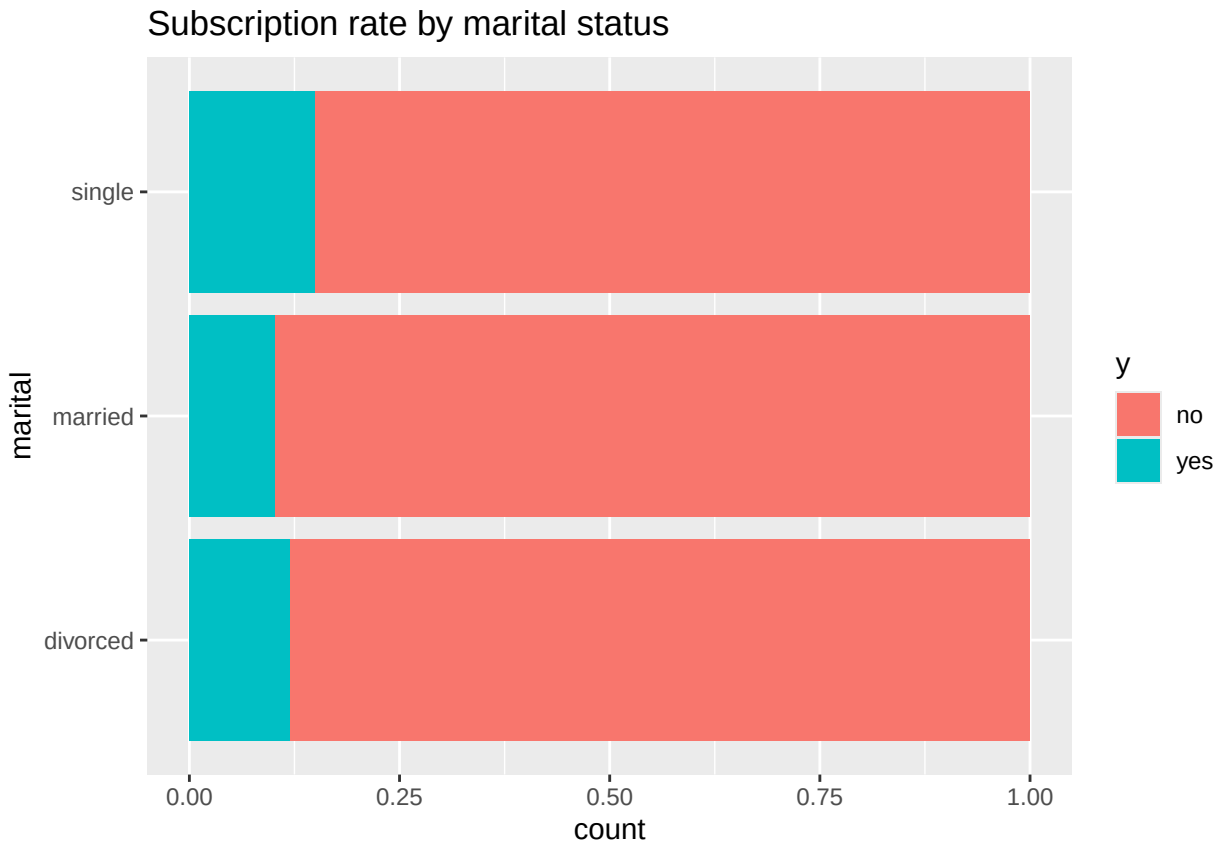

```
ggplot(data, aes(x = job, fill = y)) +
  geom_bar(position = "fill") +
  coord_flip() +
  labs(title = "Subscription rate by job")
```



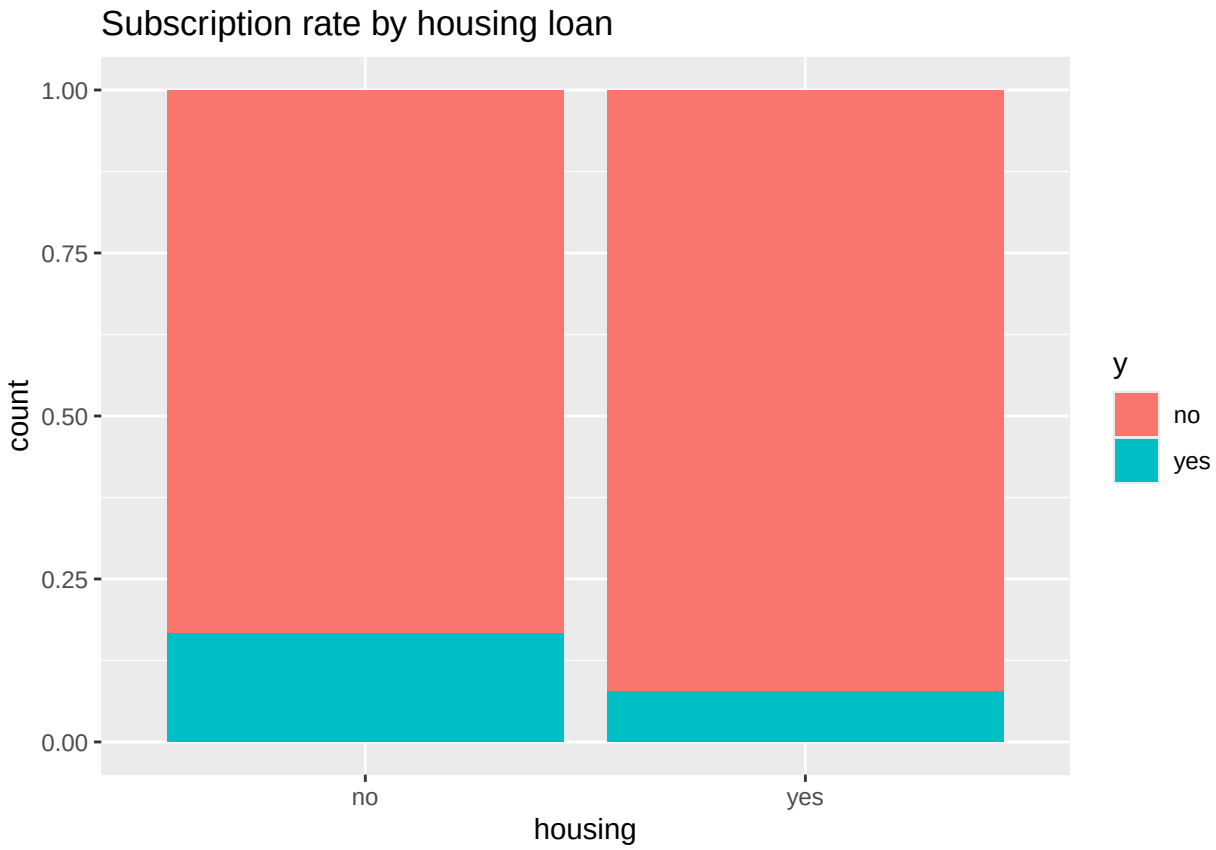

```
ggplot(data, aes(x = education, fill = y)) +
  geom_bar(position = "fill") +
  coord_flip()
```



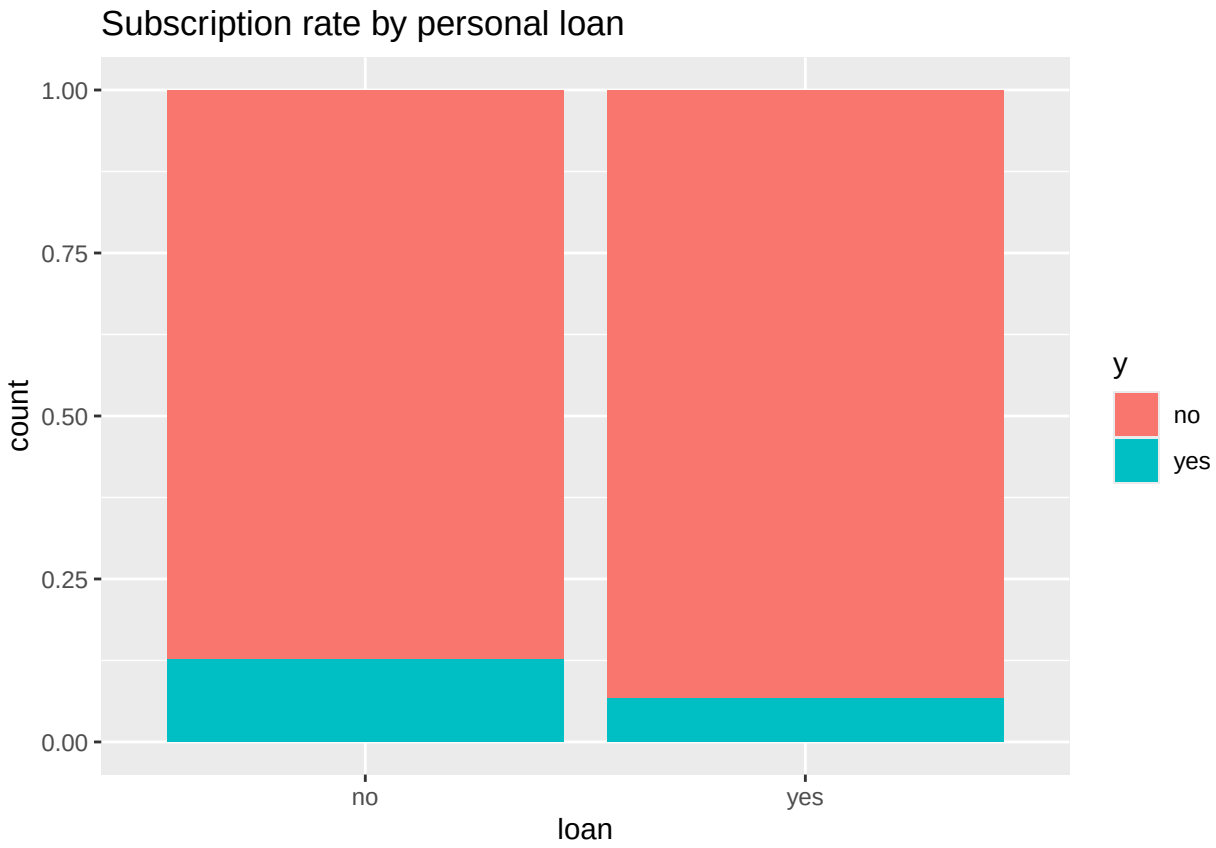

```r
ggplot(data, aes(x = marital, fill = y)) +
  geom_bar(position = "fill") +
  coord_flip() +
  labs(title = "Subscription rate by marital status")
```



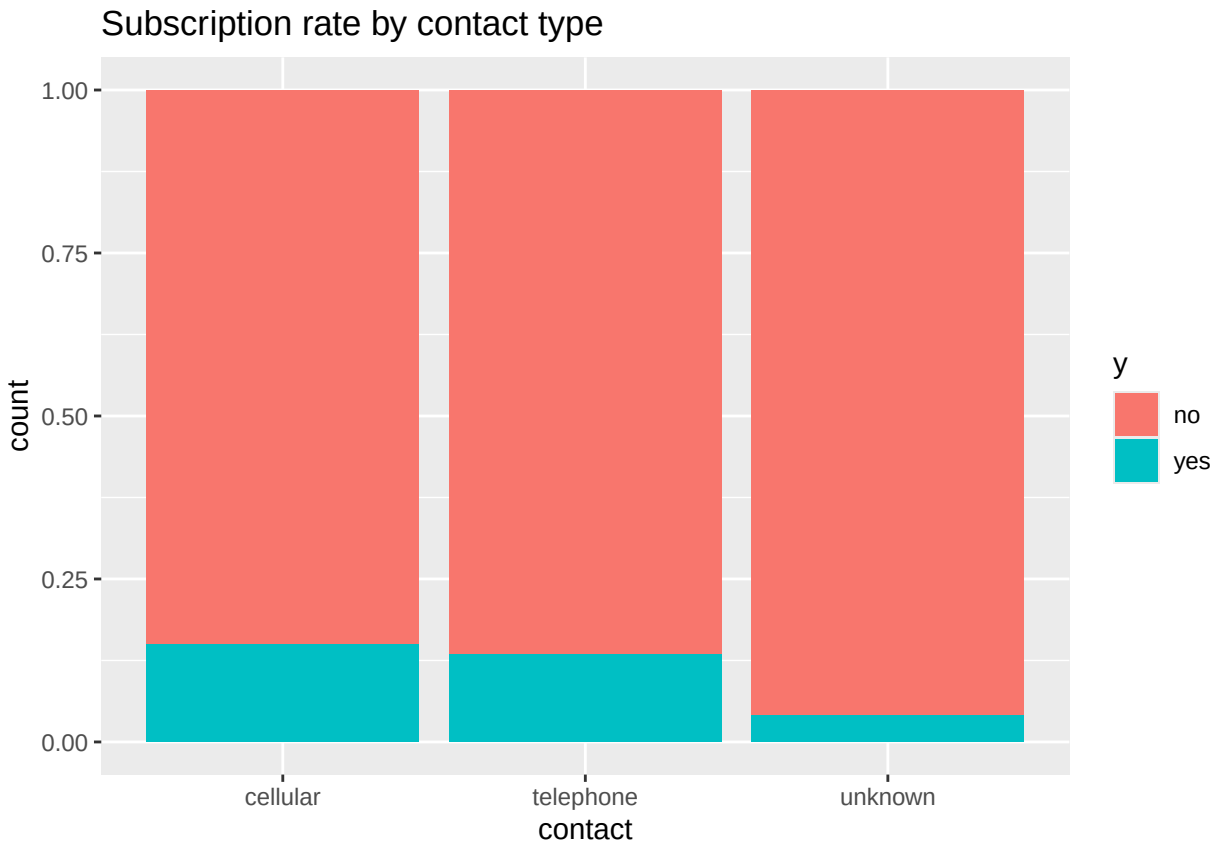

```r
ggplot(data, aes(x = housing, fill = y)) +
  geom_bar(position = "fill") +
  labs(title = "Subscription rate by housing loan")
```



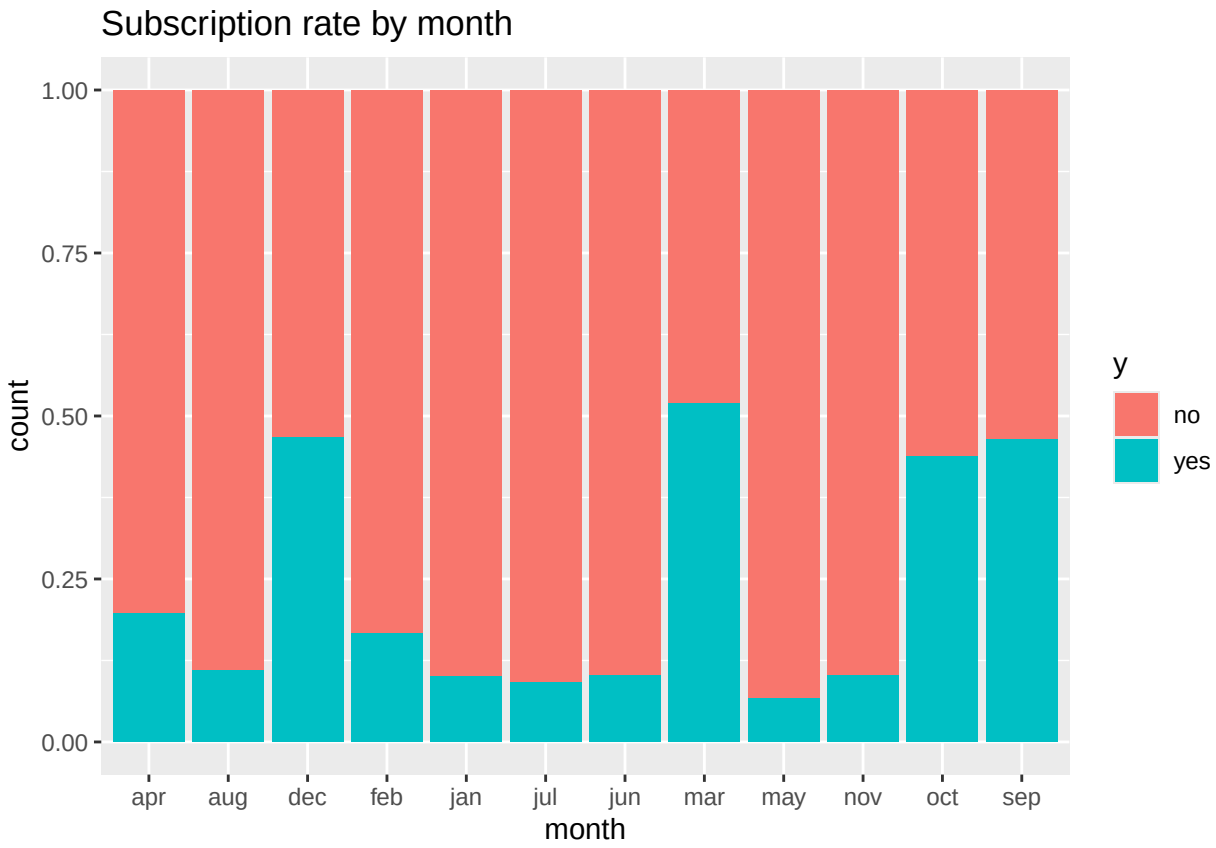

```
ggplot(data, aes(x = loan, fill = y)) +
  geom_bar(position = "fill") +
  labs(title = "Subscription rate by personal loan")
```



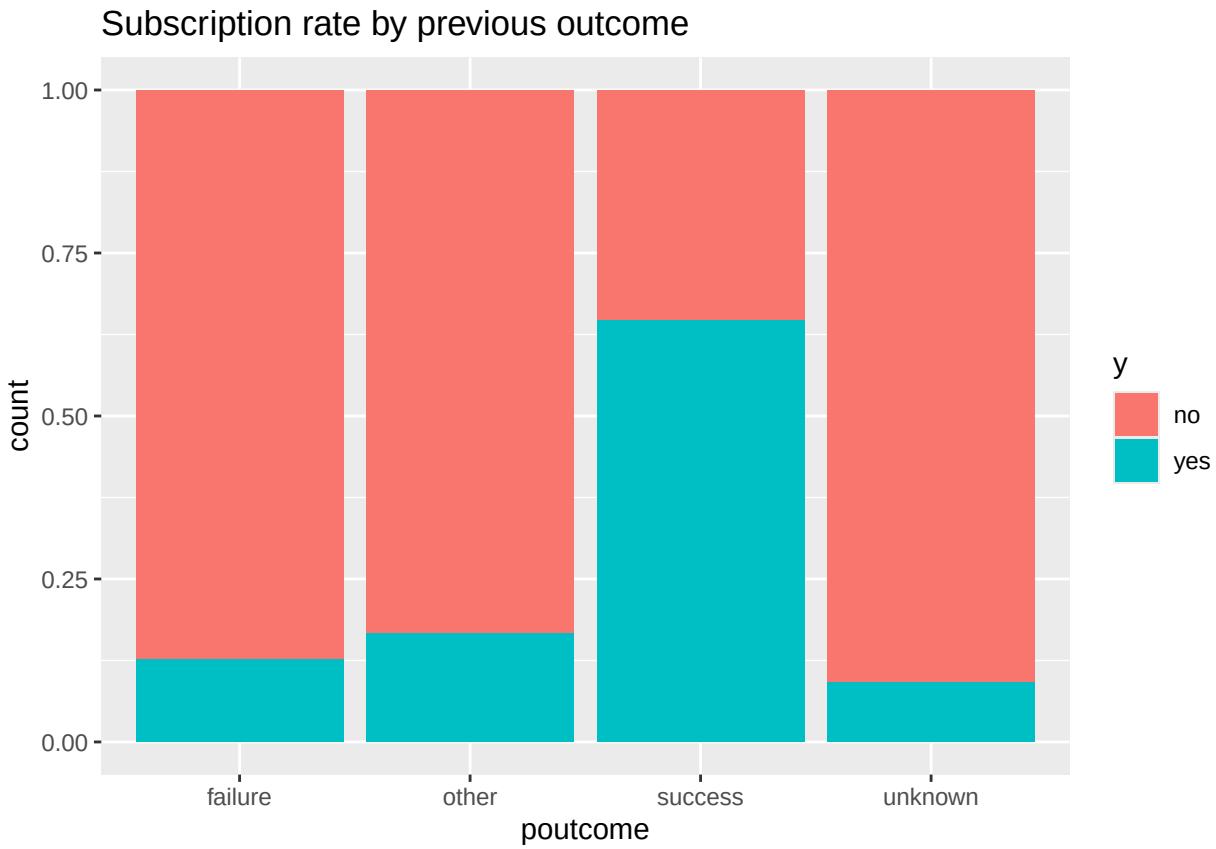

```
ggplot(data, aes(x = contact, fill = y)) +
  geom_bar(position = "fill") +
  labs(title = "Subscription rate by contact type")
```




```
ggplot(data, aes(x = month, fill = y)) +
  geom_bar(position = "fill") +
  labs(title = "Subscription rate by month")
```




```
ggplot(data, aes(x = poutcome, fill = y)) +
  geom_bar(position = "fill") +
  labs(title = "Subscription rate by previous outcome")
```



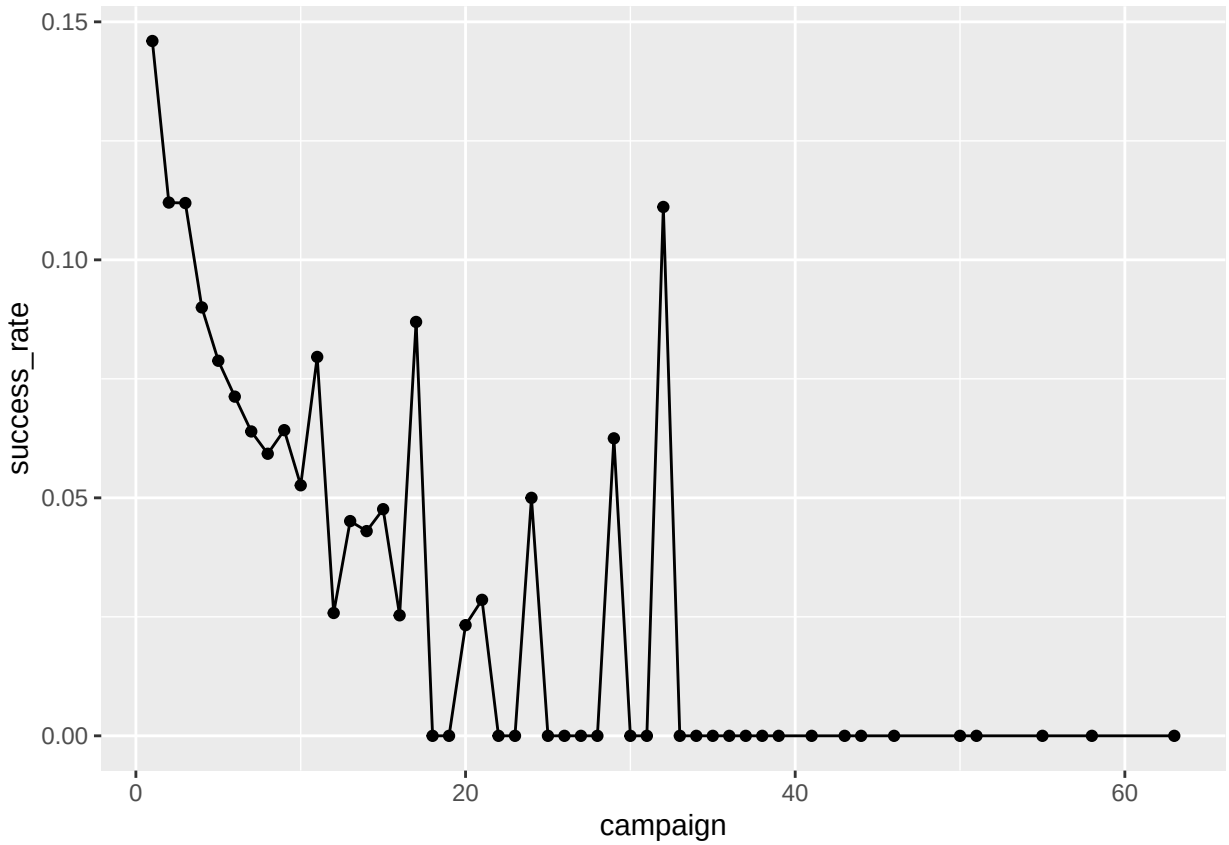

```
library(dplyr)
```

```
##
## 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
```

```
data |>
  group_by(campaign) |>
  summarise(success_rate = mean(y == "yes")) |>
  ggplot(aes(campaign, success_rate)) + geom_line() + geom_point()
```


0.15
0.10
0.05
0.00
success_rate
0
20
40
60
campaign