Data622_Assignment_1

Natalie Kalukeerthie and Anna Moy

2025-09-19

```
library(tidyverse)
library(ggplot2)
library(corrplot)
library(dplyr)

#import full bank dataset

# Import the CSV file from github
bank <- read_csv2("https://raw.githubusercontent.com/nk014914/Data622/refs/heads/main/bank-full.csv", s
```

First, we'll look at the structure of the data.

```
## spc_tbl_ [45,211 x 17] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ age      : num [1:45211] 58 44 33 47 33 35 28 42 58 43 ...
##  $ job      : chr [1:45211] "management" "technician" "entrepreneur" "blue-collar" ...
##  $ marital  : chr [1:45211] "married" "single" "married" "married" ...
##  $ education: chr [1:45211] "tertiary" "secondary" "secondary" "unknown" ...
##  $ default  : chr [1:45211] "no" "no" "no" "no" ...
##  $ balance  : num [1:45211] 2143 29 2 1506 1 ...
##  $ housing  : chr [1:45211] "yes" "yes" "yes" "yes" ...
##  $ loan     : chr [1:45211] "no" "no" "yes" "no" ...
##  $ contact  : chr [1:45211] "unknown" "unknown" "unknown" "unknown" ...
##  $ day      : num [1:45211] 5 5 5 5 5 5 5 5 5 5 ...
##  $ month    : chr [1:45211] "may" "may" "may" "may" ...
##  $ duration : num [1:45211] 261 151 76 92 198 139 217 380 50 55 ...
##  $ campaign : num [1:45211] 1 1 1 1 1 1 1 1 1 1 ...
##  $ pdays    : num [1:45211] -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 ...
##  $ previous : num [1:45211] 0 0 0 0 0 0 0 0 0 0 ...
##  $ poutcome : chr [1:45211] "unknown" "unknown" "unknown" "unknown" ...
##  $ y        : chr [1:45211] "no" "no" "no" "no" ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   age = col_double(),
##   ..   job = col_character(),
##   ..   marital = col_character(),
##   ..   education = col_character(),
##   ..   default = col_character(),
##   ..   balance = col_double(),
##   ..   housing = col_character(),
##   ..   loan = col_character(),
##   ..   contact = col_character(),
##   ..   day = col_double(),
##   ..   month = col_character(),
##   ..   duration = col_double(),
##   ..   campaign = col_double(),
##   ..   pdays = col_double(),
##   ..   previous = col_double(),
##   ..   poutcome = col_character(),
##   ..   y = col_character()
##   .. )
##  - attr(*, "problems")=<externalptr>
```

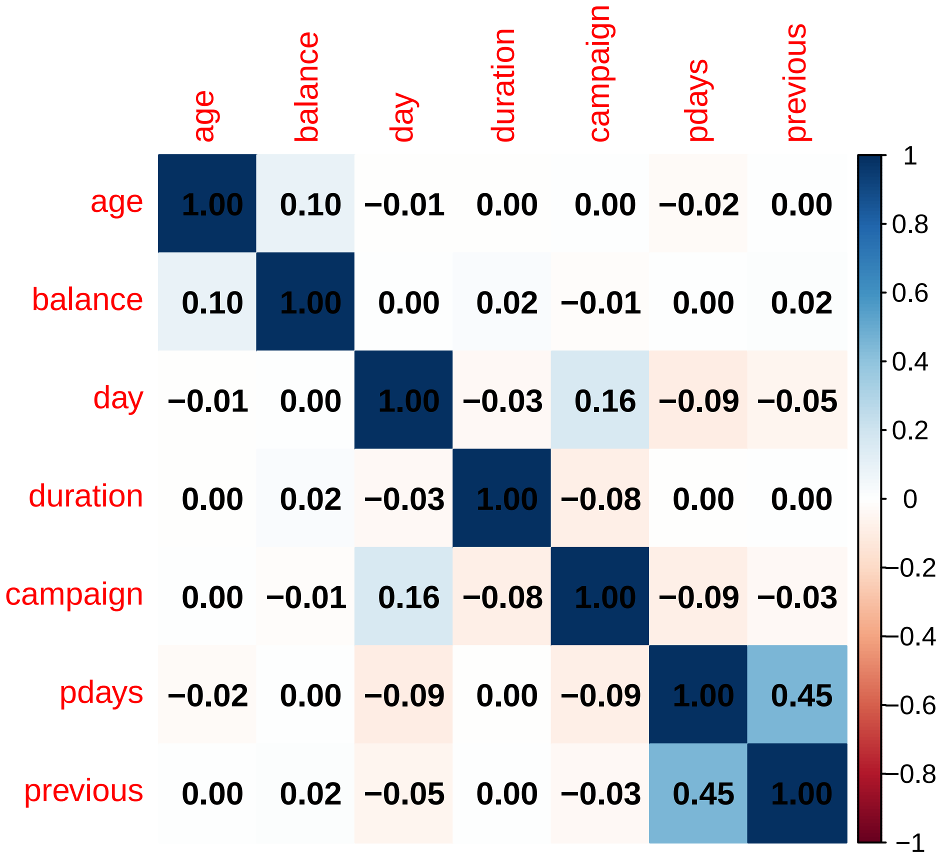
```
##       age            job              marital           education
##  Min.   :18.00   Length:45211       Length:45211       Length:45211
##  1st Qu.:33.00   Class :character   Class :character   Class :character
##  Median :39.00   Mode  :character   Mode  :character   Mode  :character
##  Mean   :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
##
##
##
```

Our chosen dataset,‘Bank-Full’, has 45,211 rows and 17 columns. It contains both numerical variables (e.g., age, balance, duration) and categorical variables (e.g., job, marital status, education). Our target variable is y, which indicates whether a client subscribed to a term deposit. Having a large dataset allows us to run models more effectively.

The summary provides central tendency (mean, median) and spread (quartiles, min, max) for numeric variables, and frequency counts for categorical ones. This helps us quickly spot imbalances or unusual values.

1. Are the features (columns) of your data correlated?

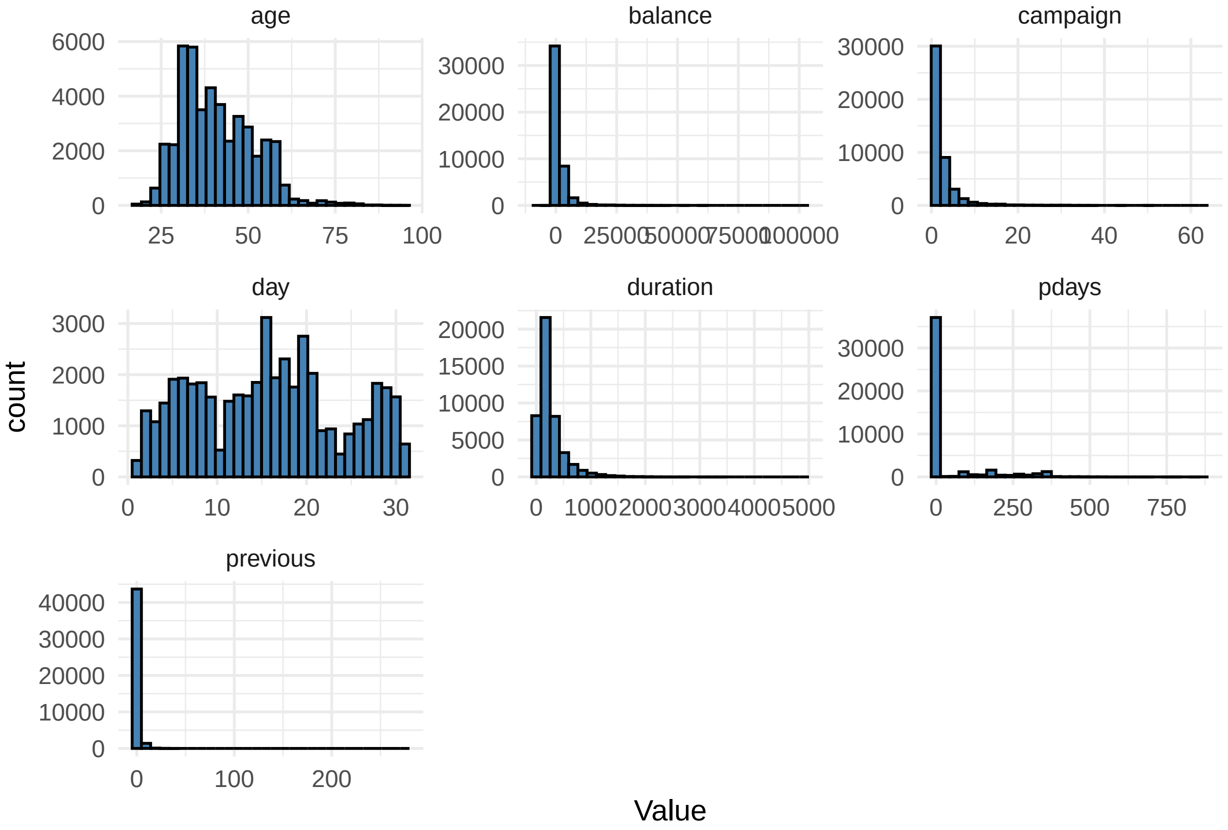
```
numeric_vars <- bank %>% select_if(is.numeric)
corr_matrix <- cor(numeric_vars)
corrplot(corr_matrix, method = "color", addCoef.col = "black")
```



The correlation matrix shows that most numeric features are weakly correlated with each other. For example, duration and campaign have a very low correlation. This indicates that predictors are largely independent, which is good for most modeling approaches. With some indicating zero it means there is no correlation with other features in the data.

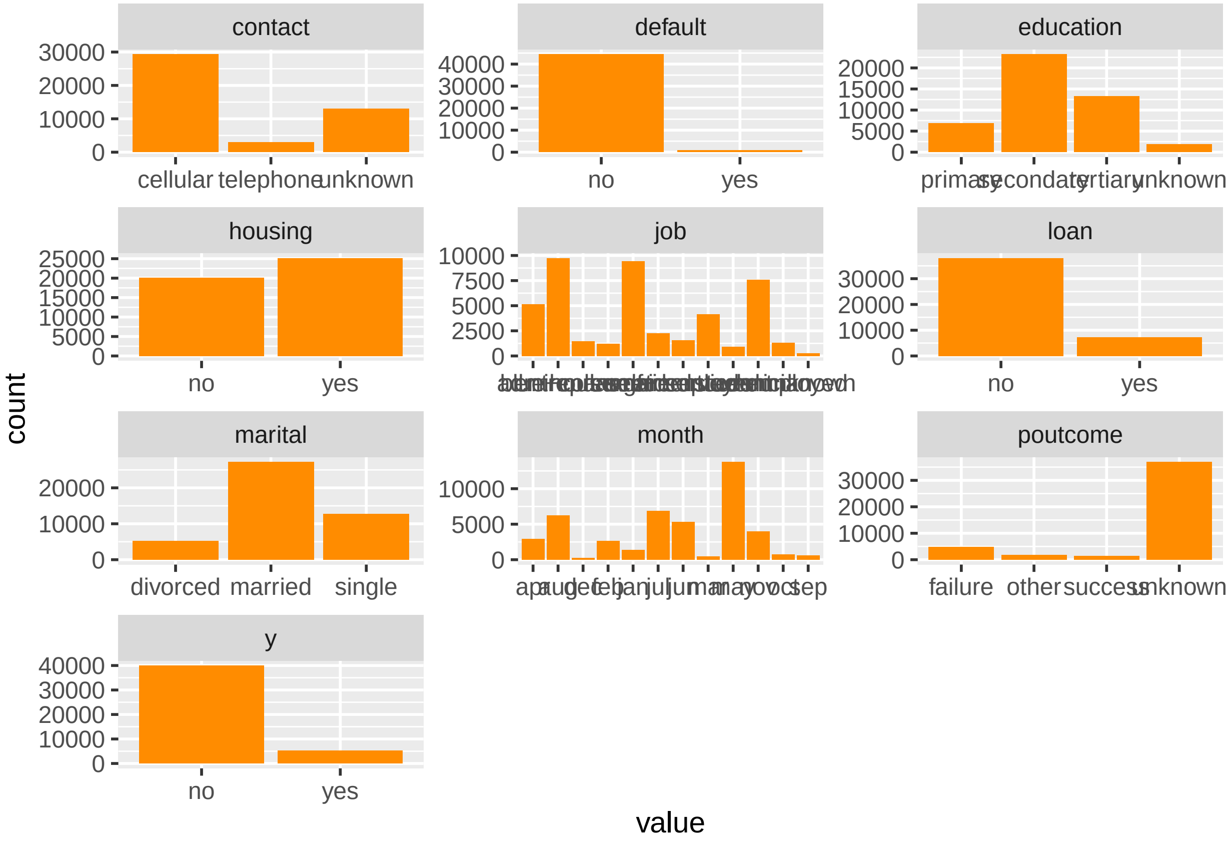
2. What is the overall distribution of each variable?

```
#numerical
numeric_vars %>%
  pivot_longer(cols = everything(), names_to = "Variable", values_to = "Value") %>%
  ggplot(aes(x = Value)) +
  facet_wrap(~Variable, scales = "free") +
  geom_histogram(bins = 30, fill = "steelblue", color = "black") +
  theme_minimal()
```



```
#categorical
categorical_data <- select(bank, where(is.character))

categorical_data %>%
  pivot_longer(cols = everything()) %>%
  ggplot(aes(x = value)) +
  facet_wrap(~name, scales = "free", ncol = 3) +
  geom_bar(fill = "darkorange")
```

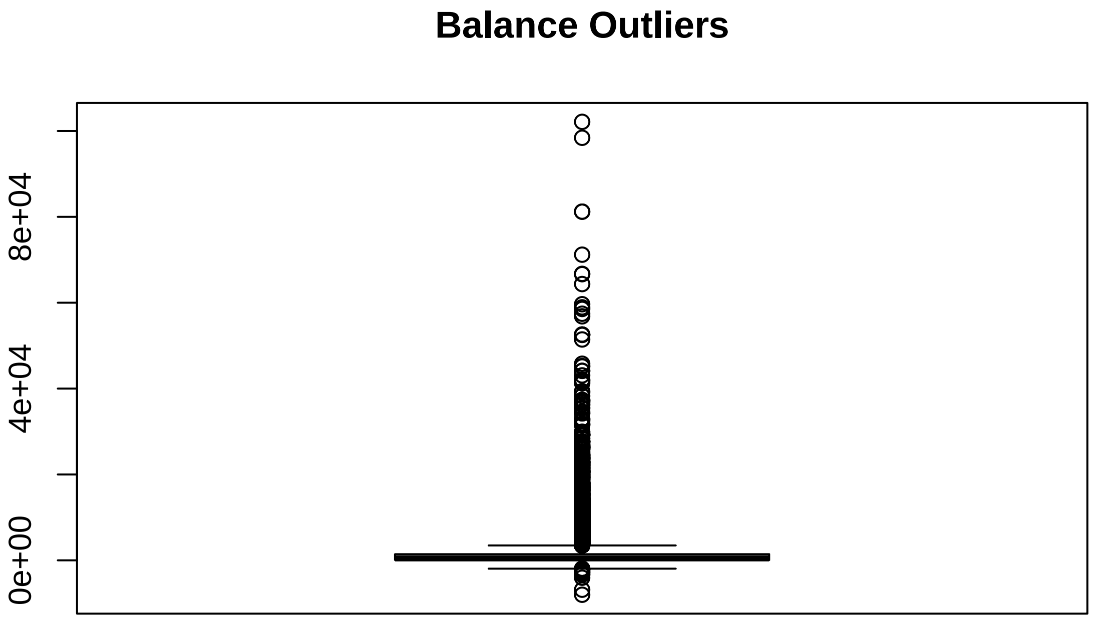


Numerical data: We see that age is right-skewed, with most clients being 25–60. Balance is highly skewed with extreme outliers, along with campaign and duration also being skewed.

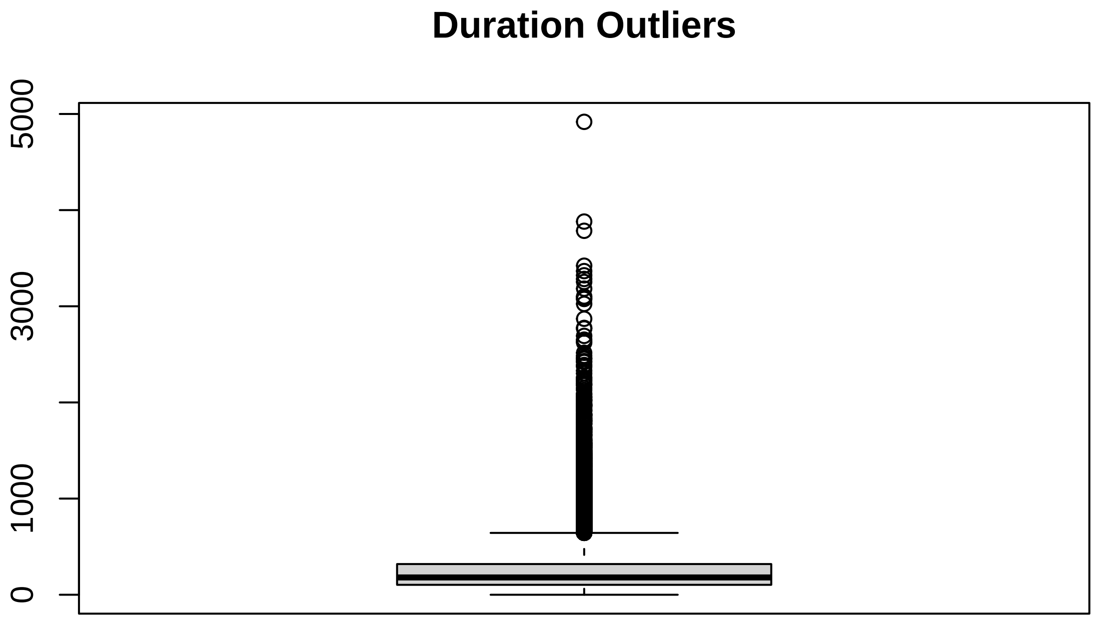
Categorical data: There is imbalance of data as the subscribe a term deposit has higher numbers in no than yes. The distribution of the features are varying in distributions whcih some higher than others.

3. Are there any outliers present?

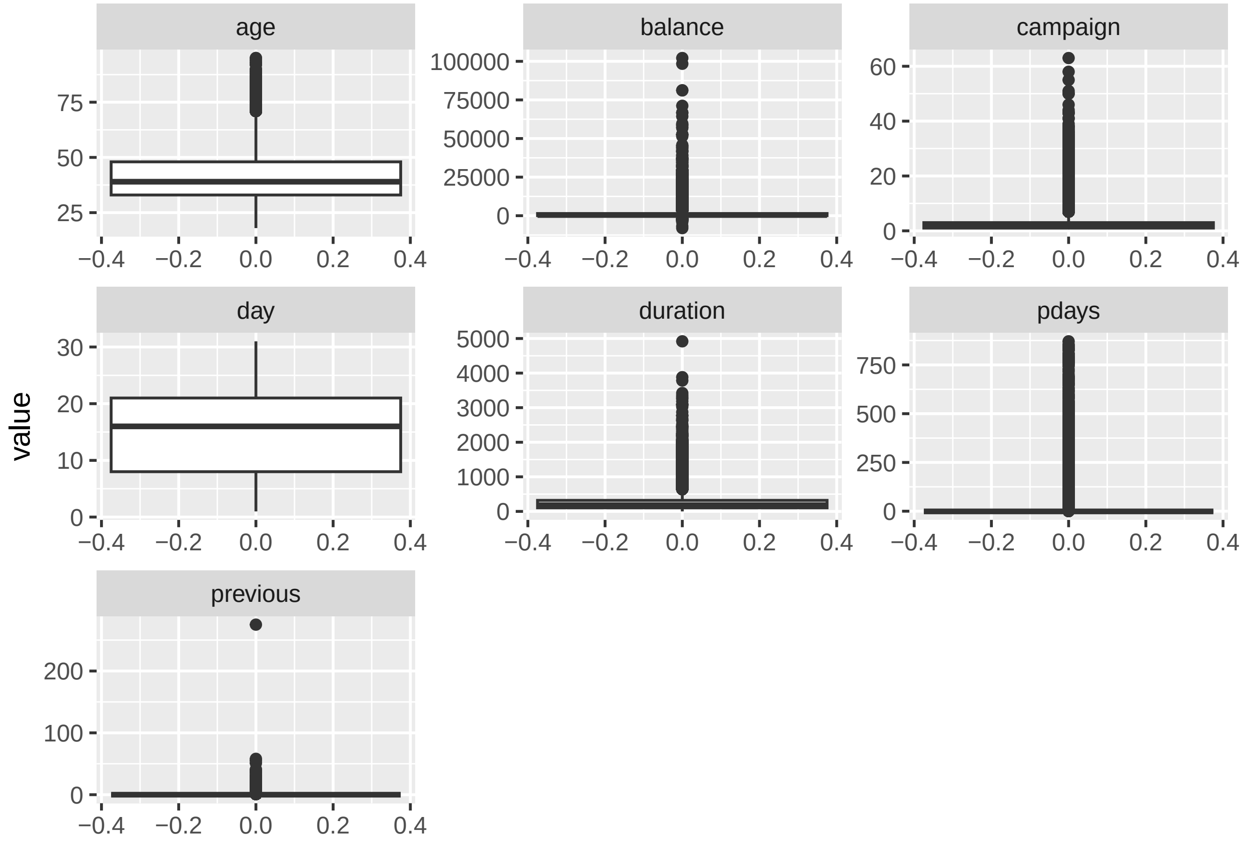
```
boxplot(bank$balance, main = "Balance Outliers")
```



```r
boxplot(bank$duration, main = "Duration Outliers")
```



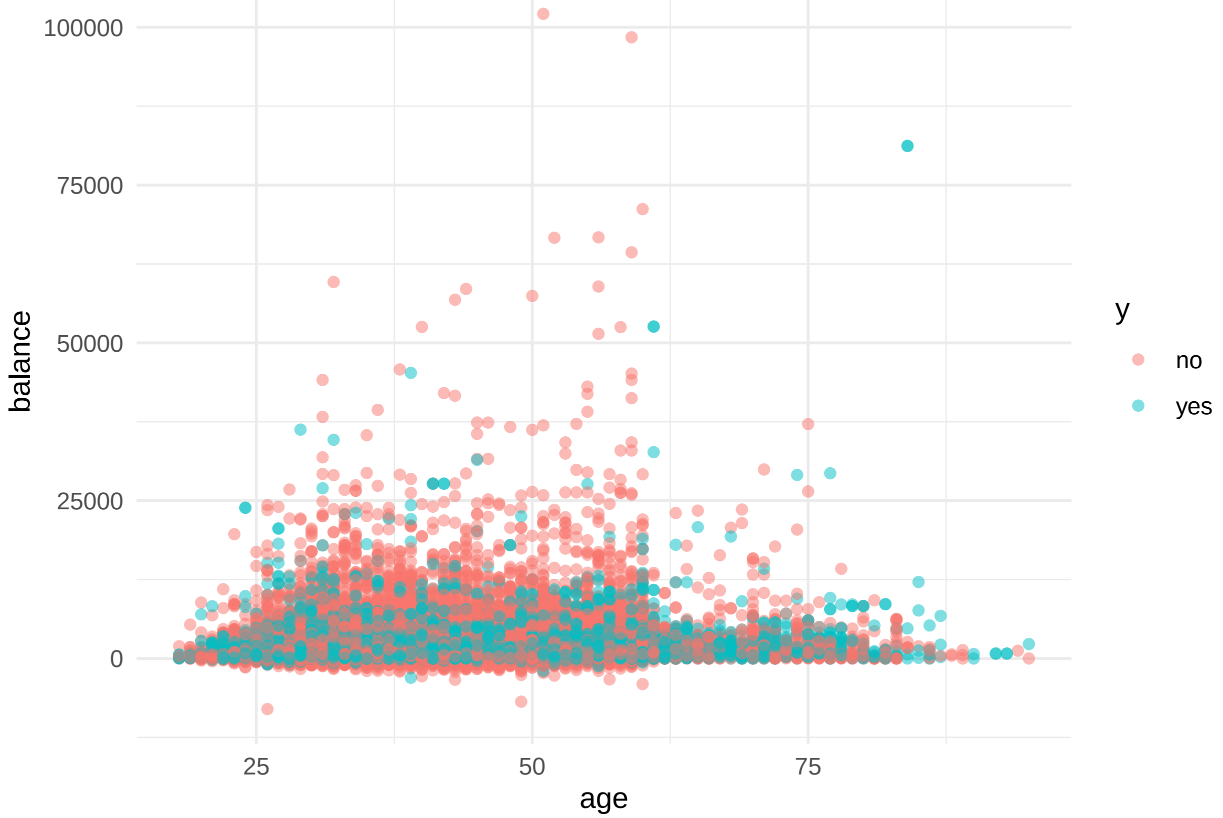
```
numeric_vars %>%
  pivot_longer(cols = everything()) %>%
  ggplot(aes(y = value)) +
  facet_wrap(~name, scales = "free", ncol = 3) +
  geom_boxplot()
```



There does seem to be some outliers present. Balance and duration show extreme outliers that may affect model training.

4. What are the relationships between different variables?

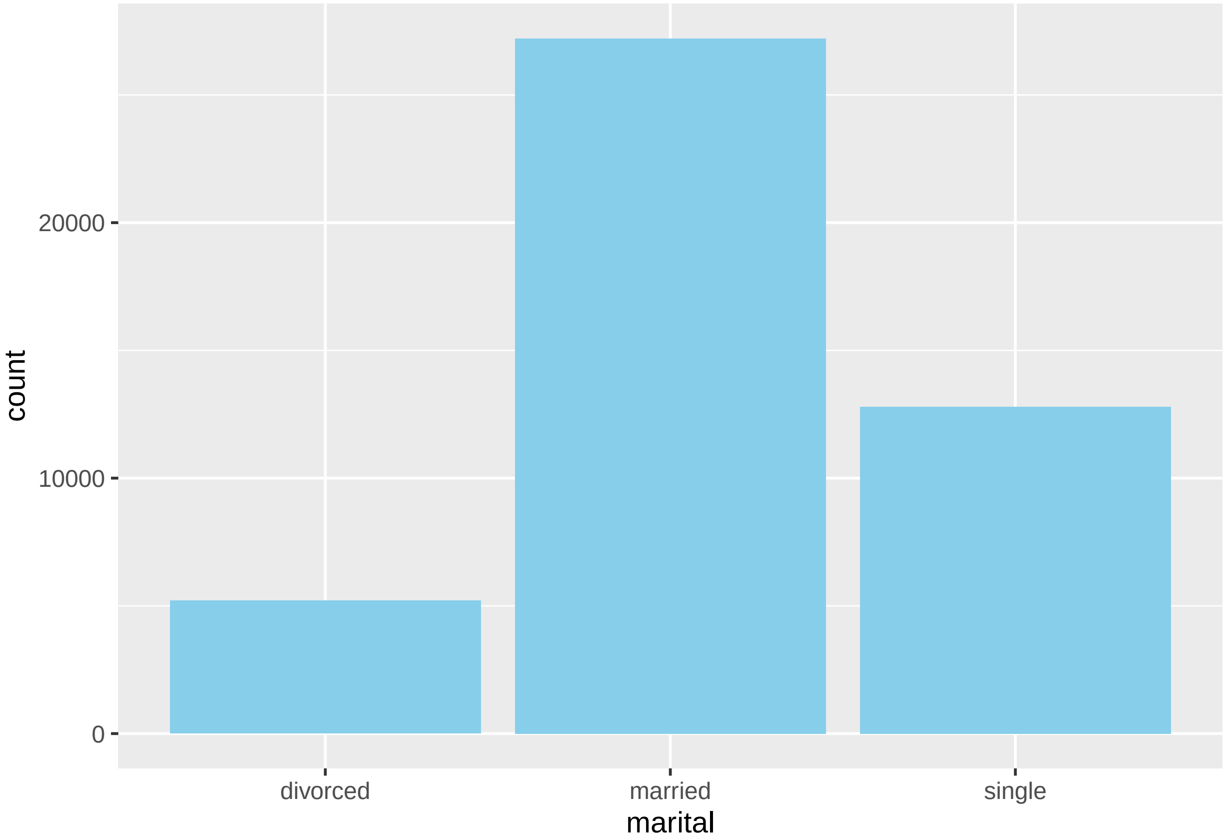
```
ggplot(bank, aes(x = age, y = balance, color = y)) +
  geom_point(alpha = 0.5) +
  theme_minimal()
```



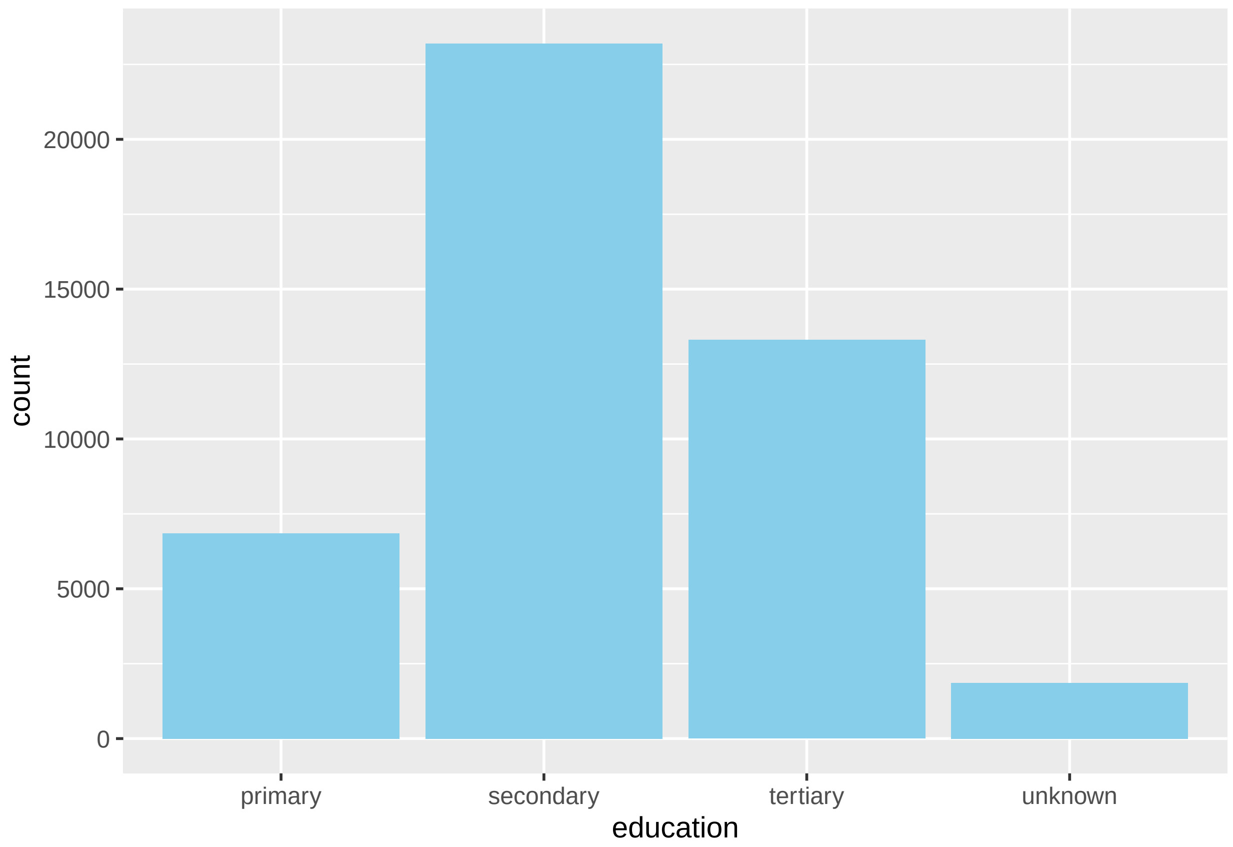
The scatterplot of age vs. balance shows that subscription decisions are not strictly determined by financial status. Younger clients (20s–30s) often have lower balances and mixed responses, while middle-aged clients (30s–50s) with moderate balances appear somewhat more likely to subscribe, possibly due to financial stability and savings goals. Older clients (60+) frequently have higher balances but show fewer "yes" outcomes, suggesting other priorities. Overall, the relationship is noisy (no single trend is distinct) indicating that age and balance alone cannot reliably predict subscription outcomes.

5. How are categorical variables distributed?

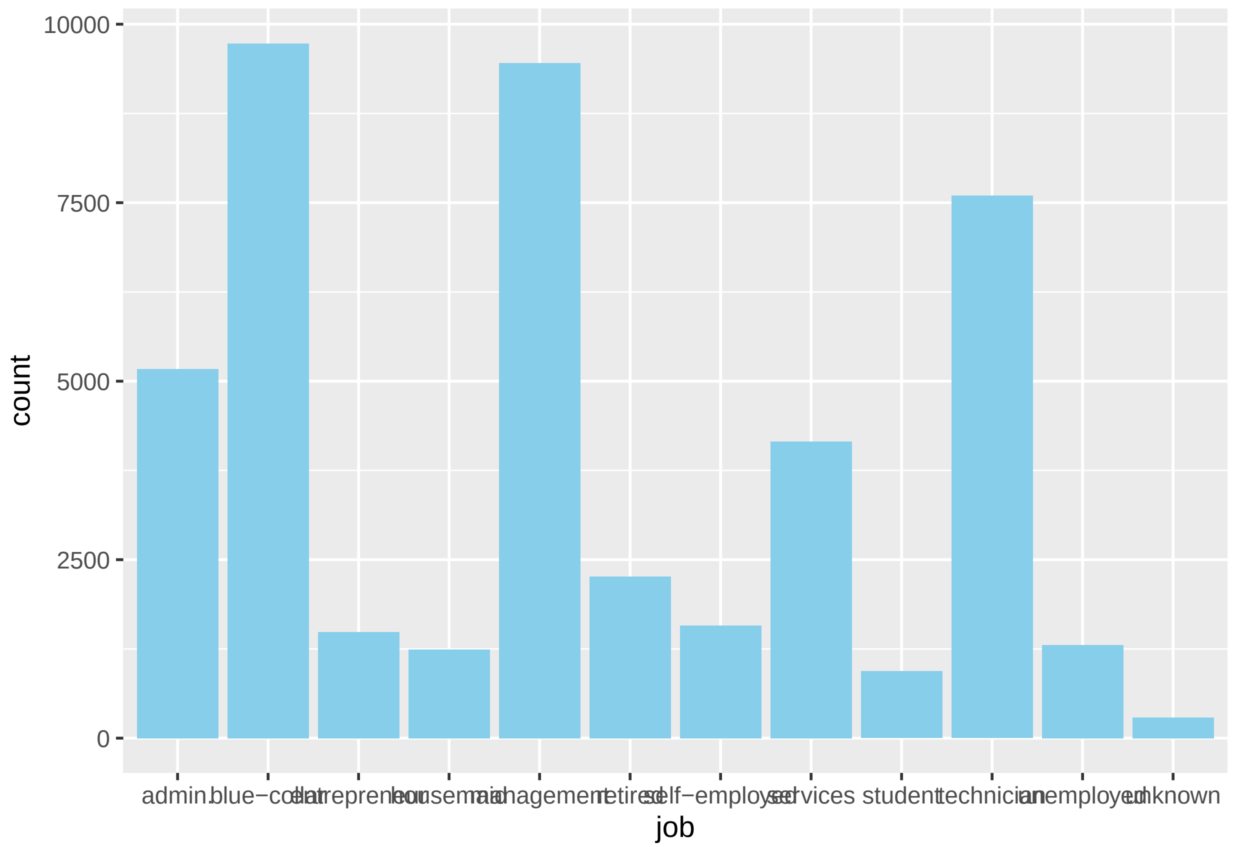
```
ggplot(bank, aes(x = marital)) + geom_bar(fill = "skyblue")
```



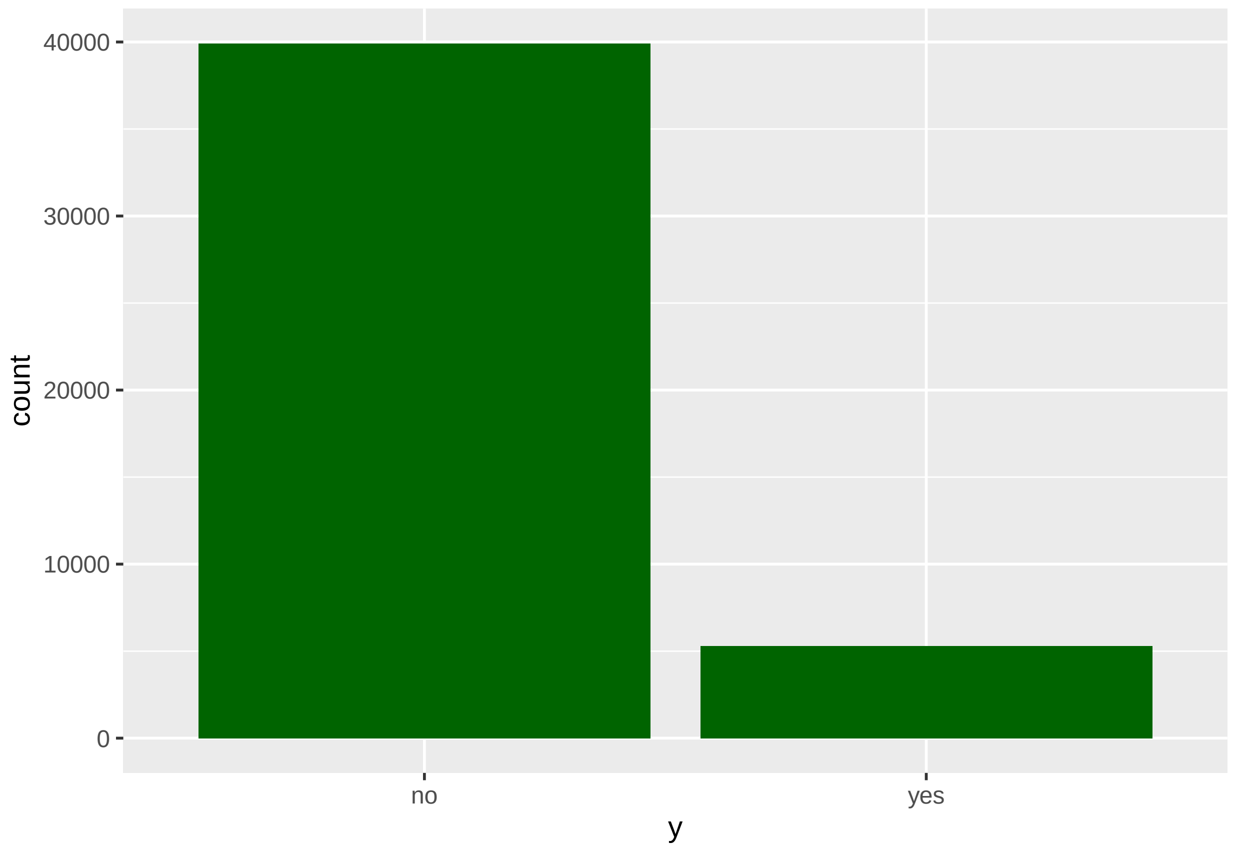
```
ggplot(bank, aes(x = education)) + geom_bar(fill = "skyblue")
```



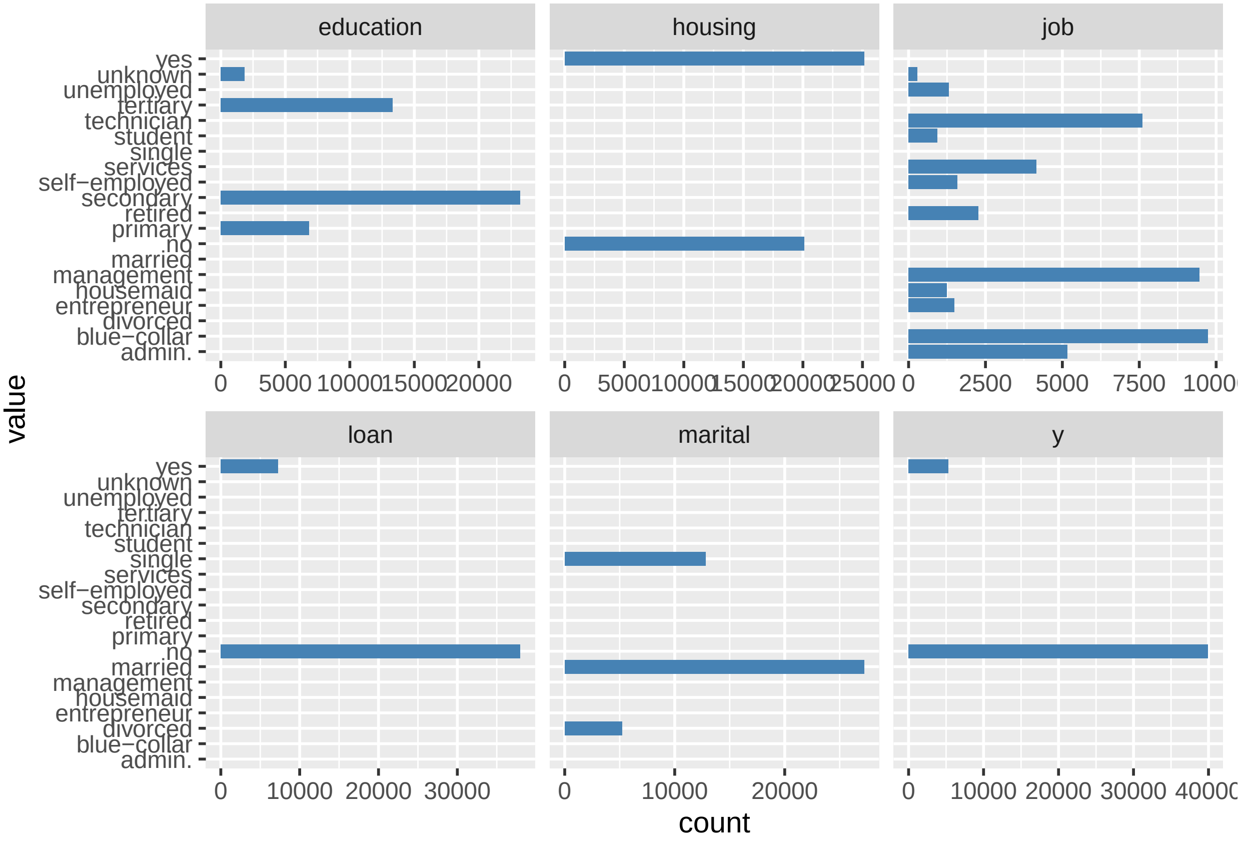
```
ggplot(bank, aes(x = job)) + geom_bar(fill = "skyblue")
```



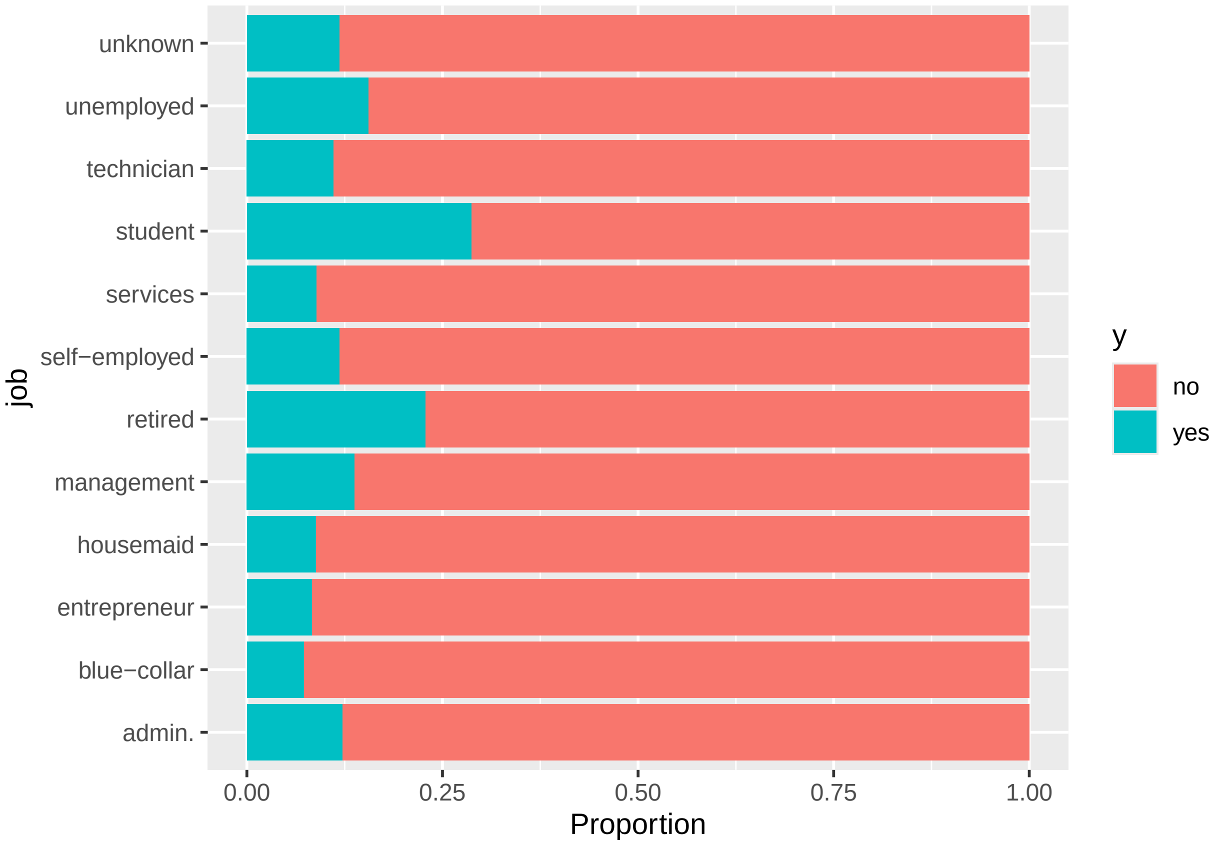
```r
ggplot(bank, aes(x = y)) + geom_bar(fill = "darkgreen")
```



```
bank %>%
  select(job, marital, education, housing, loan, y) %>%
  pivot_longer(cols = everything()) %>%
  ggplot(aes(x = value)) +
  geom_bar(fill = "steelblue") +
  facet_wrap(~name, scales = "free_x") +
  coord_flip()
```



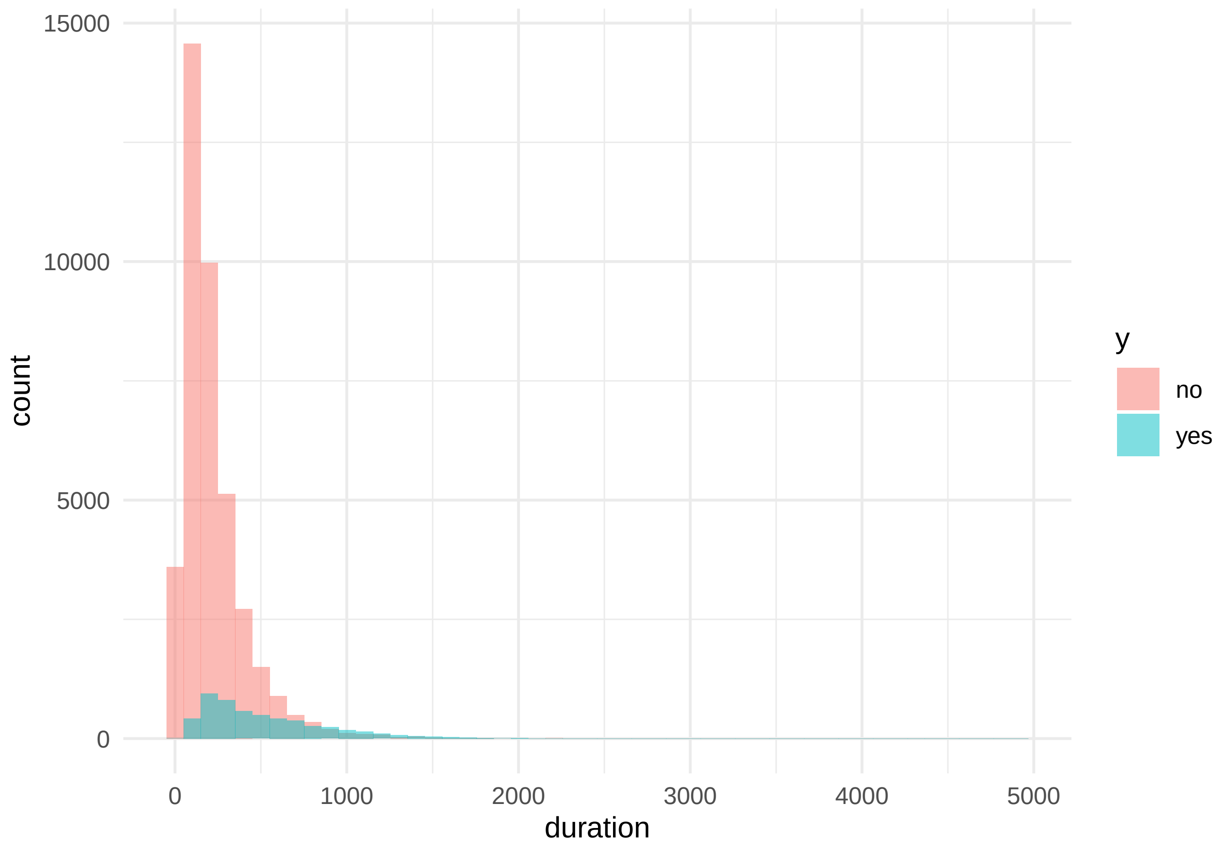
```r
ggplot(bank, aes(x = job, fill = y)) +
  geom_bar(position = "fill") +
  ylab("Proportion") +
  coord_flip()
```



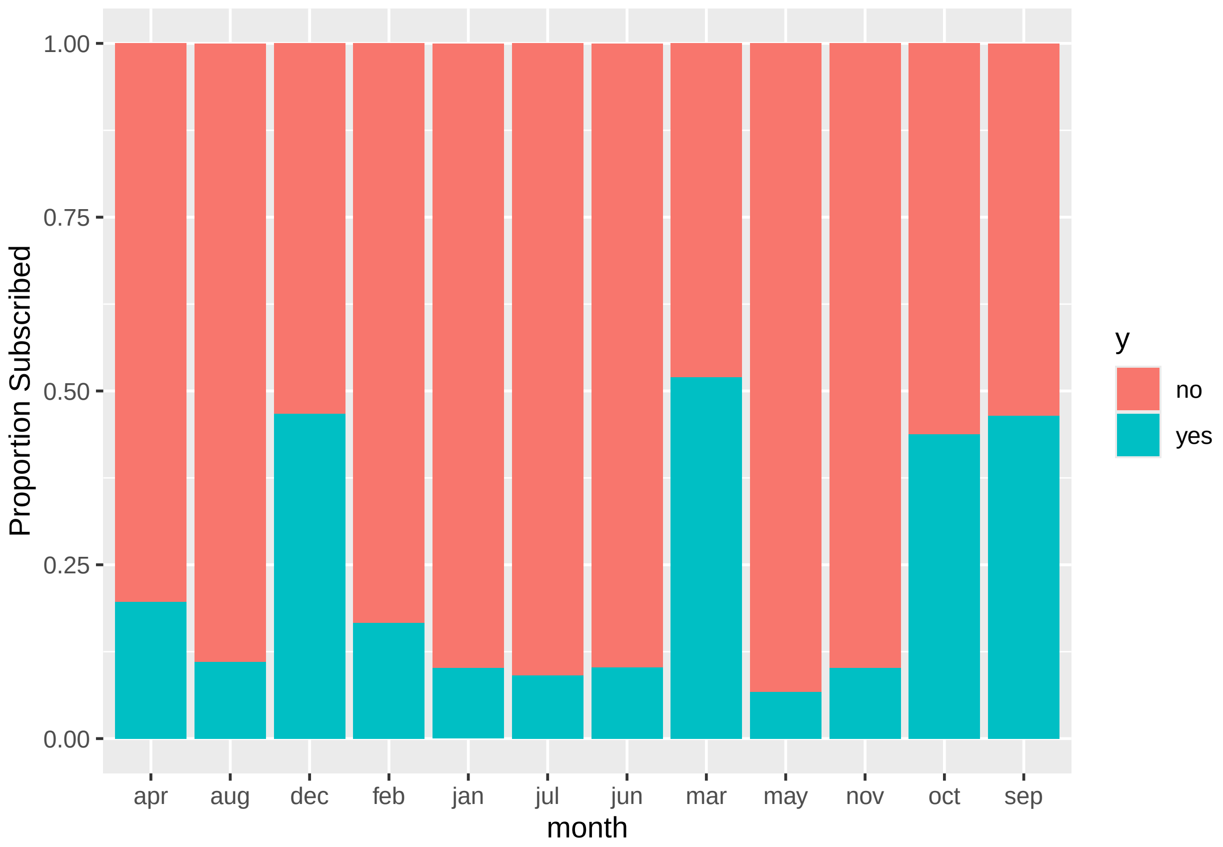
Most clients are married and have secondary education. Blue-collar, management, and technician are the most common jobs. The target variable y shows a heavy imbalance, with many more “no” responses than “yes.”

6. Do any patterns or trends emerge in the data?

```
ggplot(bank, aes(x = duration, fill = y)) +
  geom_histogram(bins = 50, position = "identity", alpha = 0.5) +
  theme_minimal()
```



```r
ggplot(bank, aes(x = month, fill = y)) +
  geom_bar(position = "fill") +
  ylab("Proportion Subscribed")
```



Yes, we're able to see a few patterns from the data, namely:

- Call duration is strongly predictive (longer calls are much more likely to end in a "yes.")
- Middle-aged clients with stable jobs and balances appear somewhat more responsive.
- Campaign outcome is imbalanced, meaning the bank struggled to achieve a high conversion rate overall.
- There were more yes in the month of december, march, october and september than other months.

7. What is the central tendency and spread of each variable?

```
summary(numeric_vars)
```

```
##       age           balance            day           duration
##  Min.   :18.00   Min.   : -8019   Min.   : 1.00   Min.   :   0.0
##  1st Qu.:33.00   1st Qu.:    72   1st Qu.: 8.00   1st Qu.: 103.0
##  Median :39.00   Median :   448   Median :16.00   Median : 180.0
##  Mean   :40.94   Mean   :  1362   Mean   :15.81   Mean   : 258.2
##  3rd Qu.:48.00   3rd Qu.:  1428   3rd Qu.:21.00   3rd Qu.: 319.0
##  Max.   :95.00   Max.   :102127   Max.   :31.00   Max.   :4918.0
##     campaign         pdays          previous
##  Min.   : 1.000   Min.   : -1.0   Min.   :  0.0000
##  1st Qu.: 1.000   1st Qu.: -1.0   1st Qu.:  0.0000
##  Median : 2.000   Median : -1.0   Median :  0.0000
##  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
```

The central tendency and spread of the numeric variables provide insights into the typical client profile as well as the variability in the data.

- `age`: mean about 40 years, range 18–95.

The age variable has a mean of approximately 40 years, with clients ranging from 18 to 95 years old, indicating that the dataset includes a wide age spectrum, though most clients fall within the 25–60 range.

- `balance`: median near 448 but very high max (>80,000) showing extreme outliers.

The balance variable is heavily skewed: while the median balance is around 448, some clients have extremely high balances exceeding 80,000, revealing the presence of significant outliers that could influence model performance.

- `duration`: median ~180s, but some calls last over an hour.

Similarly, duration, which measures the length of the last call in seconds, has a median value around 180 seconds, but a few calls exceed an hour, again highlighting extreme cases.

- `campaign`: median 2 contacts, max 63 (also extreme outliers).

The campaign variable, representing the number of contacts performed during this campaign for a client, has a median of 2 and a maximum of 63, showing that some clients were contacted repeatedly. Overall, these statistics indicate that most clients cluster around typical values (middle-aged, moderate balances, short calls, and few campaign contacts), but the presence of extreme outliers and wide spreads suggests that the dataset has heavy tails and variability that should be considered during modeling.

8. Are there any missing values and how significant are they?

```
summary(numeric_vars)
```

```
##       age           balance            day           duration     
##  Min.   :18.00   Min.   : -8019   Min.   : 1.00   Min.   :   0.0  
##  1st Qu.:33.00   1st Qu.:    72   1st Qu.: 8.00   1st Qu.: 103.0  
##  Median :39.00   Median :   448   Median :16.00   Median : 180.0  
##  Mean   :40.94   Mean   :  1362   Mean   :15.81   Mean   : 258.2  
##  3rd Qu.:48.00   3rd Qu.:  1428   3rd Qu.:21.00   3rd Qu.: 319.0  
##  Max.   :95.00   Max.   :102127   Max.   :31.00   Max.   :4918.0  
##     campaign          pdays           previous       
##  Min.   : 1.000   Min.   : -1.0   Min.   :  0.0000  
##  1st Qu.: 1.000   1st Qu.: -1.0   1st Qu.:  0.0000  
##  Median : 2.000   Median : -1.0   Median :  0.0000  
##  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  
```

The dataset contains no missing values, so no imputation is required.