DATA 622 ASSIGNMENT 1

Susanna Wong

2025-02-25

Assignment 1 Prompt

Introduction

This assignment focuses on one of the most important aspects of data science, Exploratory Data Analysis (EDA). Many surveys show that data scientists spend 60-80% of their time on data preparation. EDA allows you to identify data gaps & data imbalances, improve data quality, create better features and gain a deep understanding of your data before doing model training - and that ultimately helps train better models. In machine learning, there is a saying - "better data beats better algorithms" - meaning that it is more productive to spend time improving data quality than improving the code to train the model.

This will be an exploratory exercise, so feel free to show errors and warnings that arise during the analysis. Test the code with both datasets selected and compare the results.

Dataset

A Portuguese bank conducted a marketing campaign (phone calls) to predict if a client will subscribe to a term deposit The records of their efforts are available in the form of a dataset. The objective here is to apply machine learning techniques to analyze the dataset and figure out most effective tactics that will help the bank in next campaign to persuade more customers to subscribe to the bank's term deposit. Download the Bank Marketing Dataset from: https://archive.ics.uci.edu/dataset/222/bank+marketing

Assignment

1. Exploratory Data Analysis

Review the structure and content of the data and answer questions such as:

- Are the features (columns) of your data correlated?
- What is the overall distribution of each variable?
- Are there any outliers present?
- What are the relationships between different variables?
- How are categorical variables distributed?
- Do any patterns or trends emerge in the data?
- What is the central tendency and spread of each variable?
- Are there any missing values and how significant are they?

2. Algorithm Selection

Now you have completed the EDA, what Algorithms would suit the business purpose for the dataset. Answer questions such as:

- Select two or more machine learning algorithms presented so far that could be used to train a model (no need to train models - I am only looking for your recommendations).

- What are the pros and cons of each algorithm you selected?
- Which algorithm would you recommend, and why?
- Are there labels in your data? Did that impact your choice of algorithm?
- How does your choice of algorithm relates to the dataset?
- Would your choice of algorithm change if there were fewer than 1,000 data records, and why?

3. Pre-processing Now you have done an EDA and selected an Algorithm, what pre-processing (if any) would you require for:

- Data Cleaning - improve data quality, address missing data, etc.
- Dimensionality Reduction - remove correlated/redundant data than will slow down training
- Feature Engineering - use of business knowledge to create new features
- Sampling Data - using sampling to resize datasets
- Data Transformation - regularization, normalization, handling categorical variables
- Imbalanced Data - reducing the imbalance between classes

Deliverable

1. Essay (minimum 500 words)

Write a short essay summarizing your findings. Explain your selection of algorithms and how they relate to the data and what you are trying to do. Format: PDF Code This should include your code, as well as the outputs of your code e.g. correlation chart Format:

1. Code

This should be saved in https://rpubs.com. Please provide a link to your code in the submission.

Exploratory Data Analysis

```
library(dplyr)
library("tidyverse")
library(ggplot2)
library(corrplot)
library(forcats)
library(GGally)
```

Data Structure

There are 45,211 observations and 17 columns.

Below is the description of the variables from UCI Machine Learning.

Variable	Description
age	Age of client
job	Type of job the client has *(admin, blue-collar, entrepreneur, housemaid, management, retired, self -employed, services, student, technician, unemployed, unknown)*
martial	Martial status of client *(divorced, married, single, unknown)*
education	Education level of client *(primary, secondary, tertiary, unknown)*
default	Has the client's credit default? *(yes/no)*
balance	average yearly balance of client in euros (numeric)
housing	Does the client have a housing loan? *(yes/no)*

Variable	Description
loan	Does the client have a personal loan? *(yes/no)*
contact	contact communication type of last contact of the current campaign *(telephone, cellular, unknown)*
day	last contact day of the month (numeric)
month	last contact month of year *(jan, feb, mar, ..., dec)*
duration	last contact duration, in seconds
campaign	number of contacts performed during this campaign and for this client (includes last contact)
pdays	number of days that passed by after the client was last contacted from a previous campaign *(-1 means client was not previously contacted)*
previous	number of contacts performed before this campaign and for this client
poutcome	outcome of the previous marketing campaign *(unknown, other, failure, success)*
y	has the client subscribed a term deposit? *yes/no)*

```
raw_data <- read.csv("https://raw.githubusercontent.com/suswong/DATA-622/refs/heads/main/bank-full.csv"

str(data)
```

```
## function (..., list = character(), package = NULL, lib.loc = NULL, verbose = getOption("verbose"),
##     envir = .GlobalEnv, overwrite = TRUE)
```

Data Summary

Below is the summary of the data. It reveals that campaigns reached a broad demographic but was not very effective as indicated by the low percentage of clients that subscribed a term deposit.

Demographic of Client

- The age range is between 18 and 95 with a median of 39.
- Most clients are married and have secondary education.
- The top 3 occupations of the clients are blue-collar workers, management professionals, and technician.

Financial Information of Client

- The clients' balance ranges from -8019 and 102127 euros.
- About 56% of the client has housing loan and very few clients (16%) has personal loan.
- About 1.8% of the clients have credit default.

Contact and Campaign Information

- Most contacts were made via cellular phone but there is a large amount (about 29%) of unknown form of last contact with client (missing data).
- The campaigns were primarily conducted in May (peak), July and August, indicating seasonal campaigns.
- The majority of the clients wer not previously contacted and most previous outcomes were failures, which may indicate ineffective previous campaigns.
- The median number of contacts performed during this campaign is 2 and some clients were contacted up to 63 times.
- Only 5289 out of 45,211 clients (about 11%) subscribed a term deposit, indicating a low success rate.

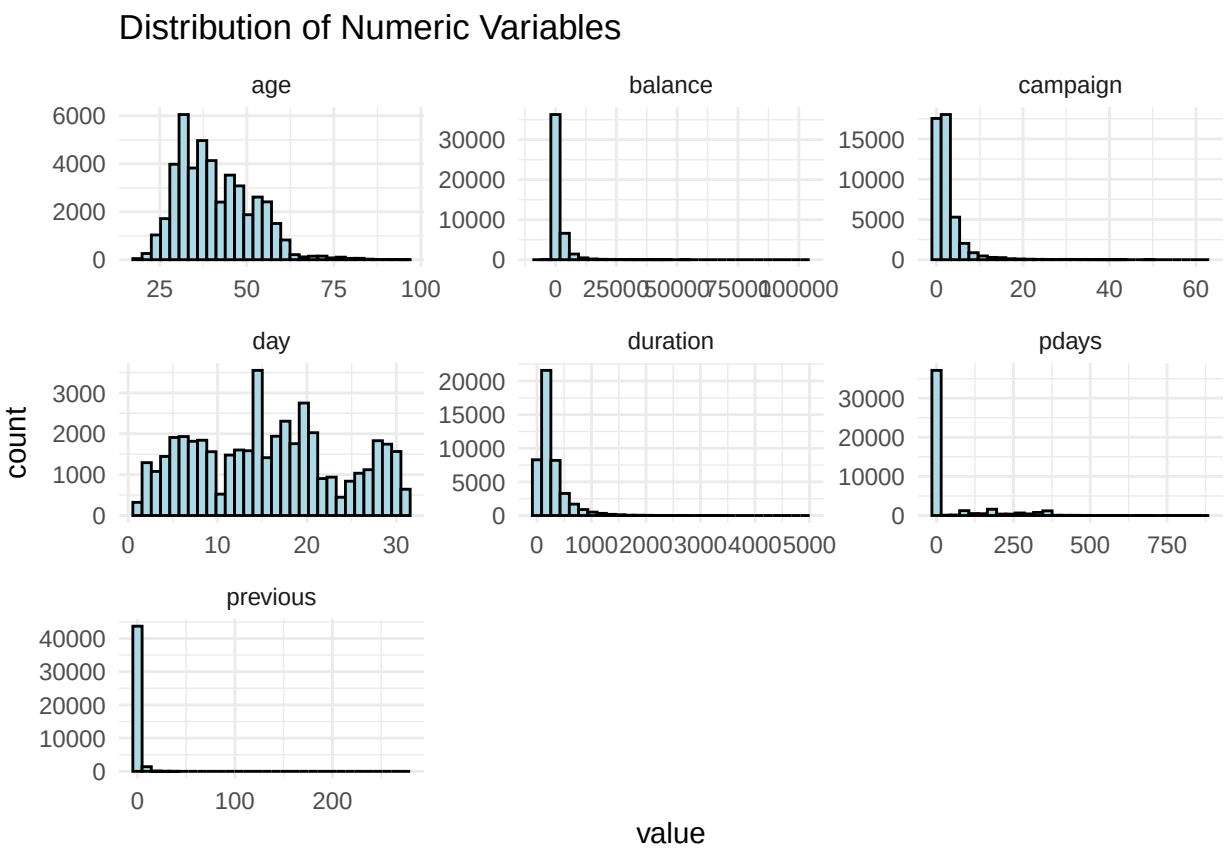
```
factor <- c("job", "marital", "education", "contact", "month", "poutcome", "default", "housing", "loan"

data <- raw_data %>%
  mutate_at(factor, as.factor)

summary(data)
```

```
##       age                 job           marital          education
##  Min.   :18.00   blue-collar:9732   divorced: 5207   primary  : 6851
##  1st Qu.:33.00   management :9458   married :27214   secondary:23202
##  Median :39.00   technician :7597   single  :12790   tertiary :13301
##  Mean   :40.94   admin.     :5171                    unknown  : 1857
##  3rd Qu.:48.00   services   :4154
##  Max.   :95.00   retired    :2264
##                  (Other)    :6835
##  default        balance       housing      loan            contact
##  no :44396   Min.   : -8019   no :20081   no :37967   cellular :29285
##  yes:  815   1st Qu.:    72   yes:25130   yes: 7244   telephone: 2906
##              Median :   448                           unknown  :13020
##              Mean   :  1362
##              3rd Qu.:  1428
##              Max.   :102127
##
##       day            month          duration         campaign
##  Min.   : 1.00   may    :13766   Min.   :   0.0   Min.   : 1.000
##  1st Qu.: 8.00   jul    : 6895   1st Qu.: 103.0   1st Qu.: 1.000
##  Median :16.00   aug    : 6247   Median : 180.0   Median : 2.000
##  Mean   :15.81   jun    : 5341   Mean   : 258.2   Mean   : 2.764
##  3rd Qu.:21.00   nov    : 3970   3rd Qu.: 319.0   3rd Qu.: 3.000
##  Max.   :31.00   apr    : 2932   Max.   :4918.0   Max.   :63.000
##                  (Other): 6060
##      pdays          previous           poutcome       y
##  Min.   : -1.0   Min.   :  0.0000   failure: 4901   no :39922
##  1st Qu.: -1.0   1st Qu.:  0.0000   other  : 1840   yes: 5289
##  Median : -1.0   Median :  0.0000   success: 1511
##  Mean   : 40.2   Mean   :  0.5803   unknown:36959
##  3rd Qu.: -1.0   3rd Qu.:  0.0000
##  Max.   :871.0   Max.   :275.0000
##
```

Distribution of Numerical Variables

All of the numerical variables are right skewed except "day".

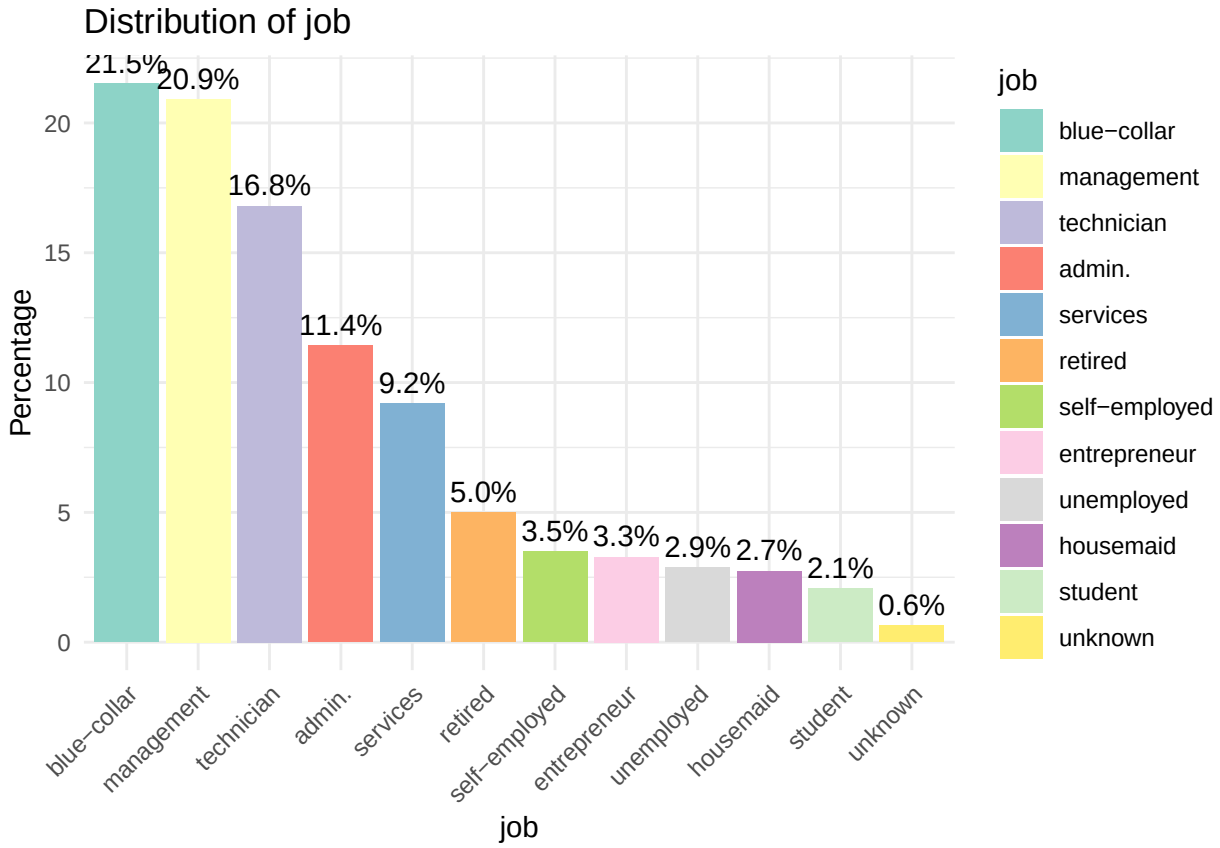
```
data %>%
  select_if(is.numeric) %>%
  gather() %>%
  ggplot(aes(value)) +
  geom_histogram(bins=30, fill="lightblue", color="black") +
  facet_wrap(~key, scales="free") +
  theme_minimal() +
  ggtitle("Distribution of Numeric Variables")
```

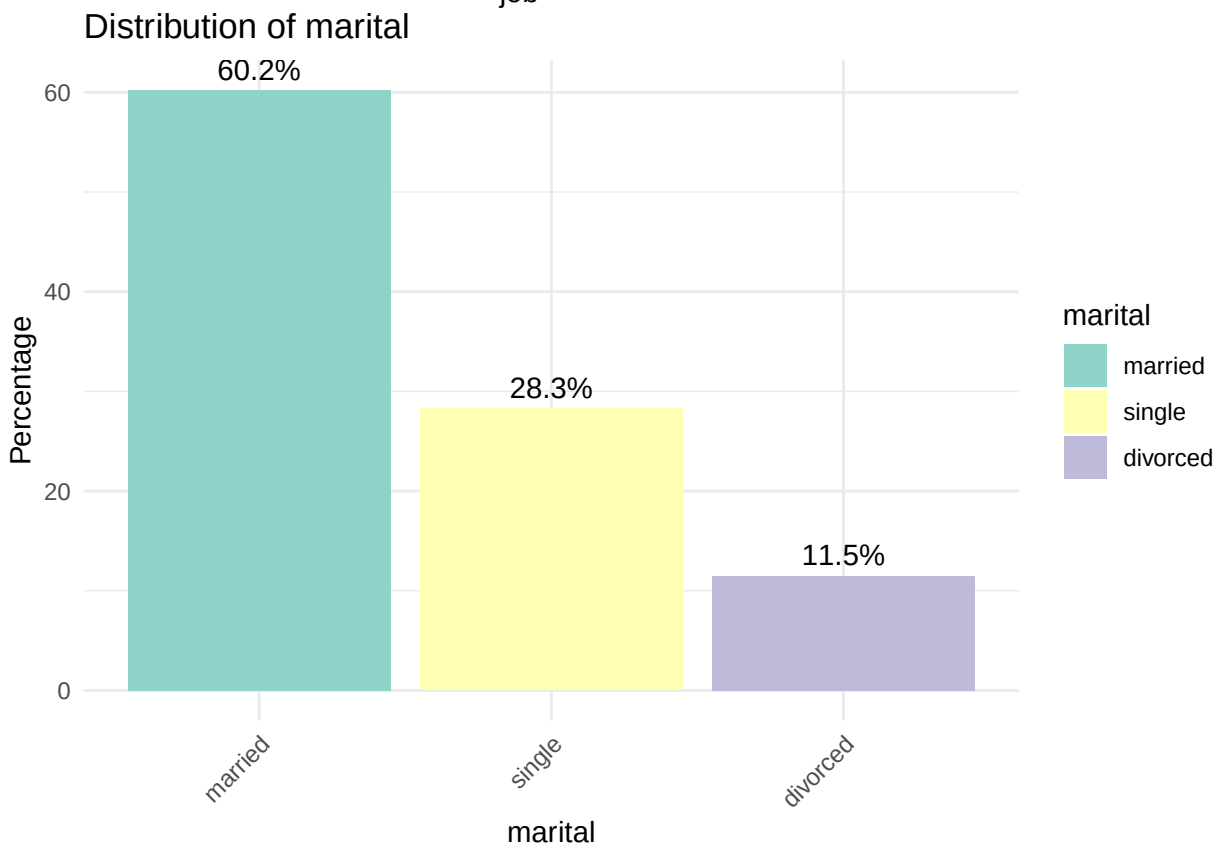


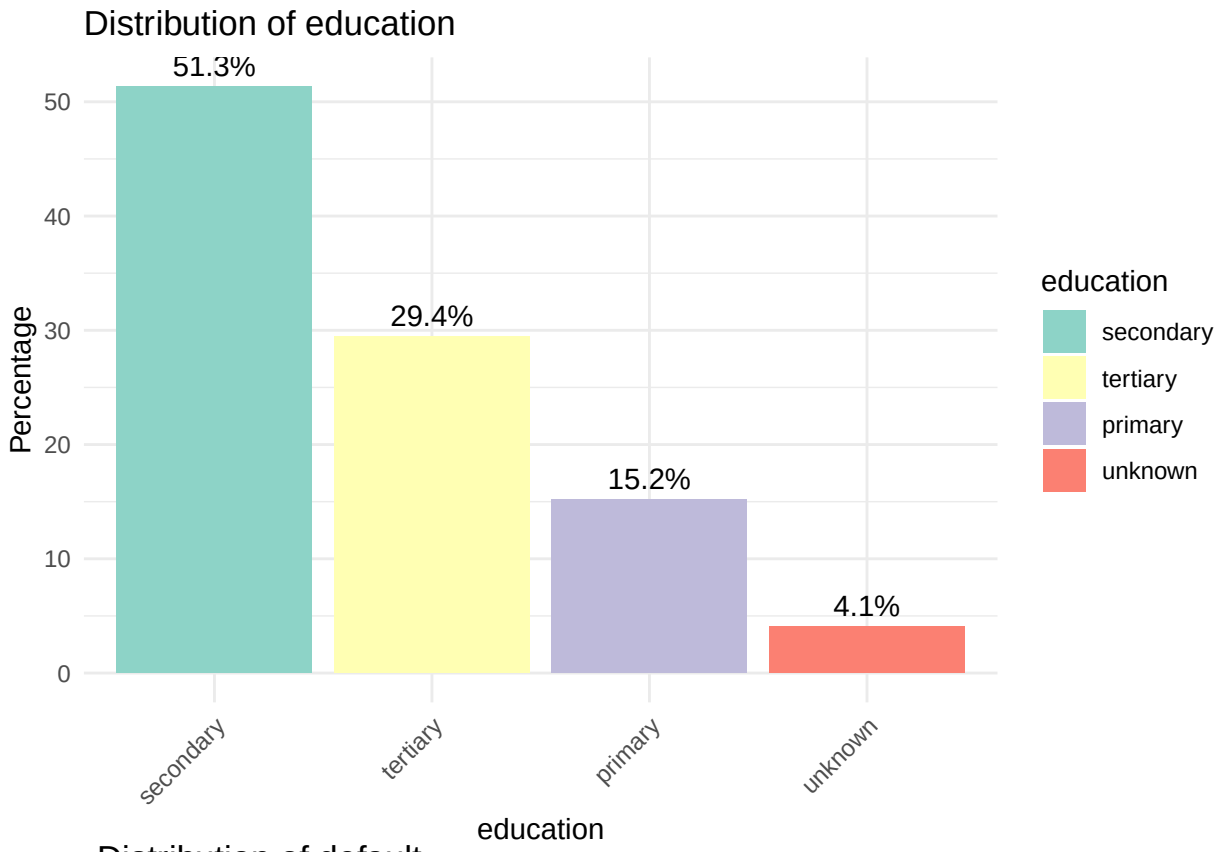
Distribution of Categorical Variables

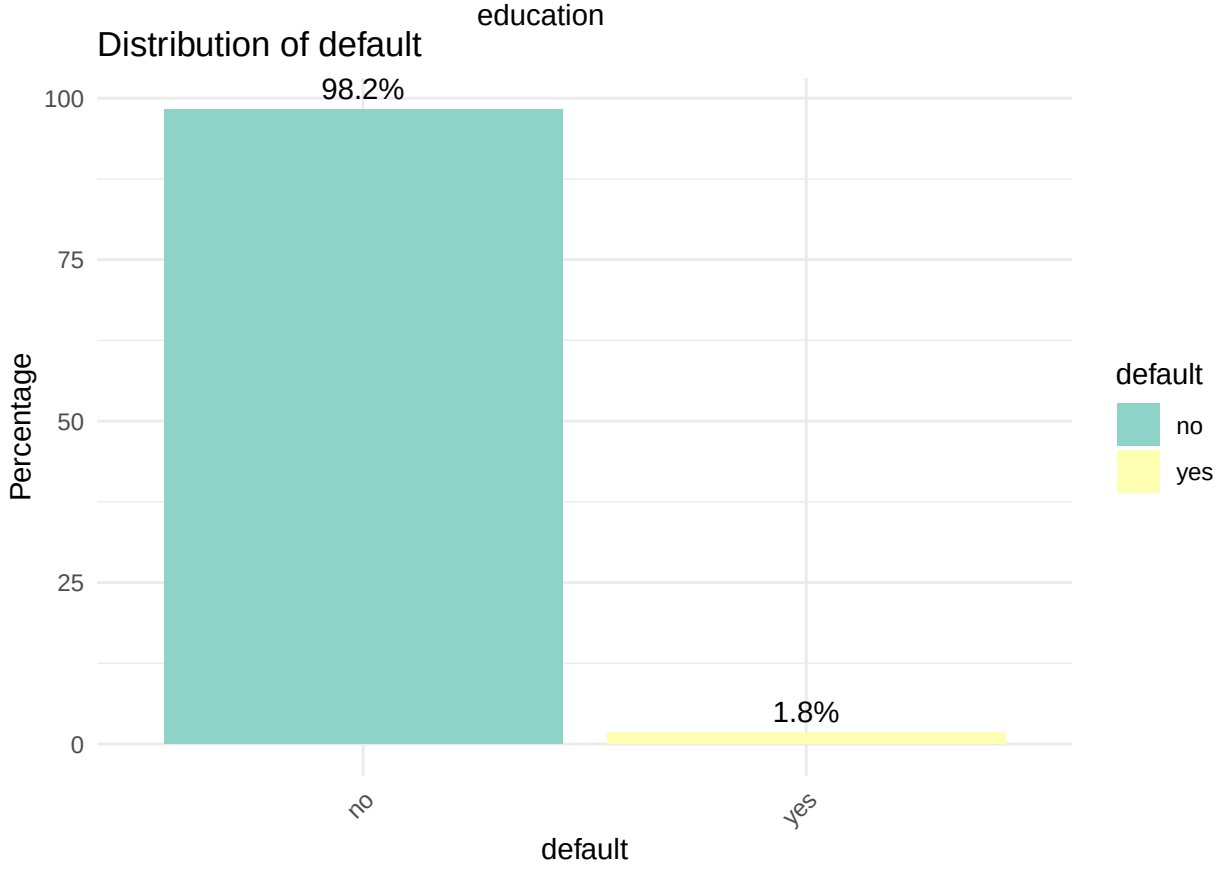
```
categorical_cols <- c("job", "marital", "education", "default", "housing", "loan", "contact", "poutcome

for (col in categorical_cols) {
  data %>%
    count(!!sym(col)) %>%
    mutate(percentage = n / sum(n) * 100) %>%
    mutate(!!col := fct_reorder(!!sym(col), percentage, .desc = TRUE)) %>%
    ggplot(aes_string(x = col, y = "percentage", fill = col)) +
    geom_bar(stat = "identity") +
    geom_text(aes(label = sprintf("%.1f%%", percentage)), vjust = -0.5) +
    theme_minimal() +
    ggtitle(paste("Distribution of", col)) +
    theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
    scale_fill_brewer(palette = "Set3") +
    ylab("Percentage") -> p

  print(p)
}
```

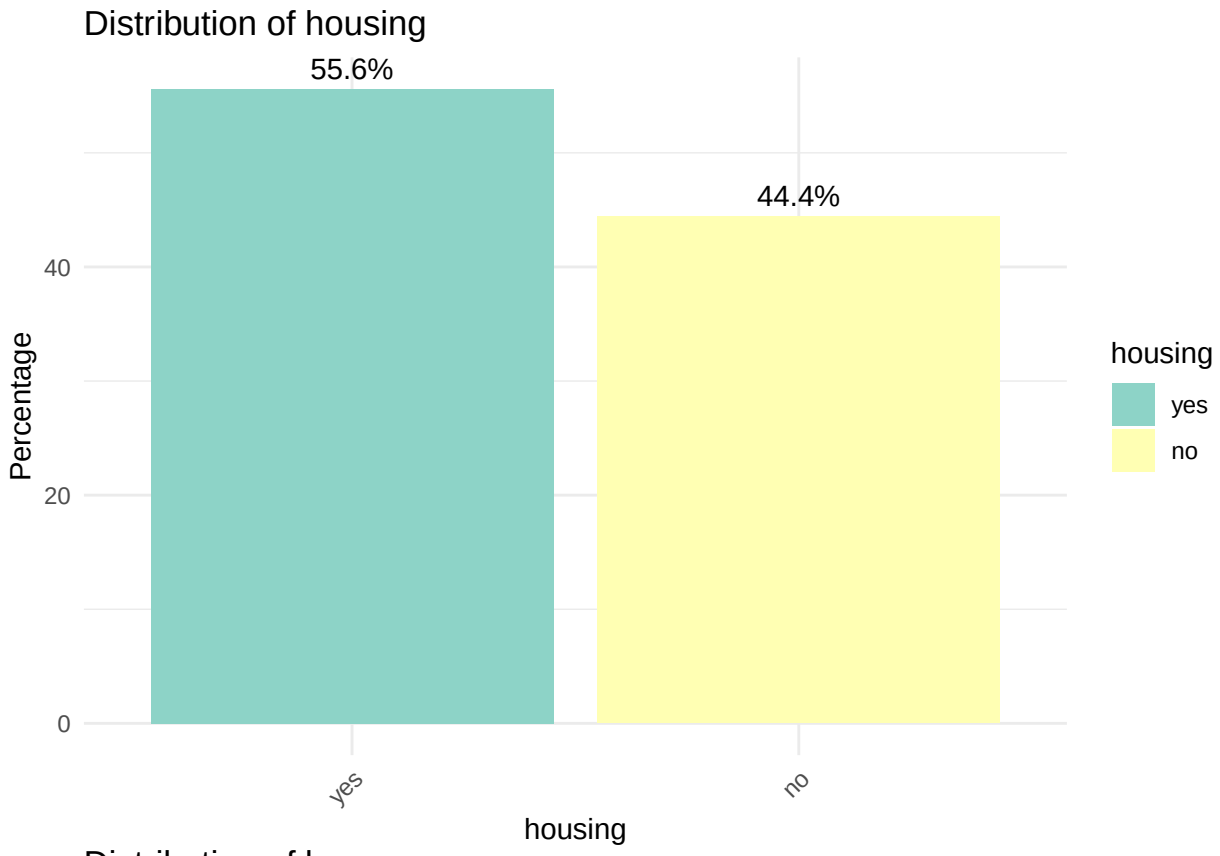
```
## Warning: `aes_string()` was deprecated in ggplot2 3.0.0.
## i Please use tidy evaluation idioms with `aes()`.
## i See also `vignette("ggplot2-in-packages")` for more information.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```

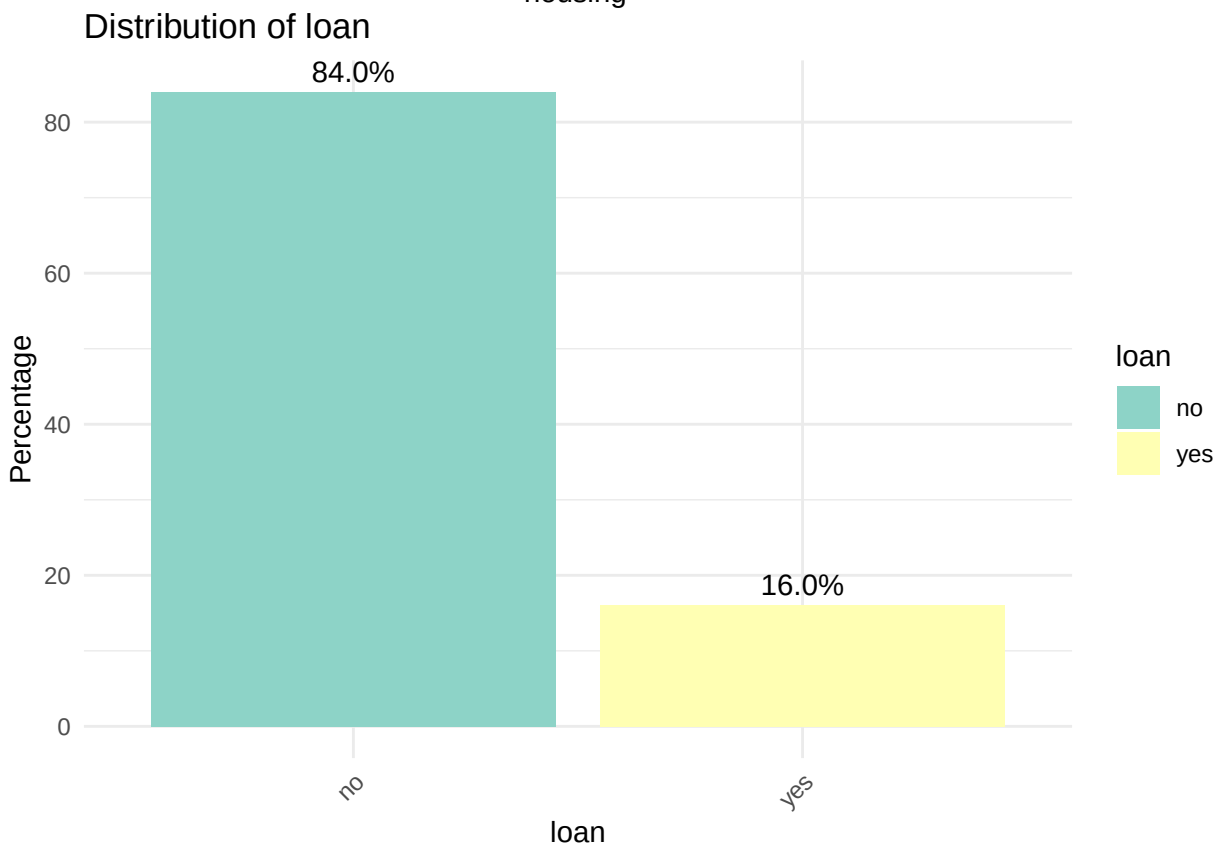

Distribution of job
21.5%
20.9%
16.8%
11.4%
9.2%
5.0%
3.5%
3.3%
2.9%
2.7%
2.1%
0.6%
Percentage
0
5
10
15
20
blue−collar
management
technician
admin.
services
retired
self−employed
entrepreneur
unemployed
housemaid
student
unknown
job
job
blue−collar
management
technician
admin.
services
retired
self−employed
entrepreneur
unemployed
housemaid
student
unknown

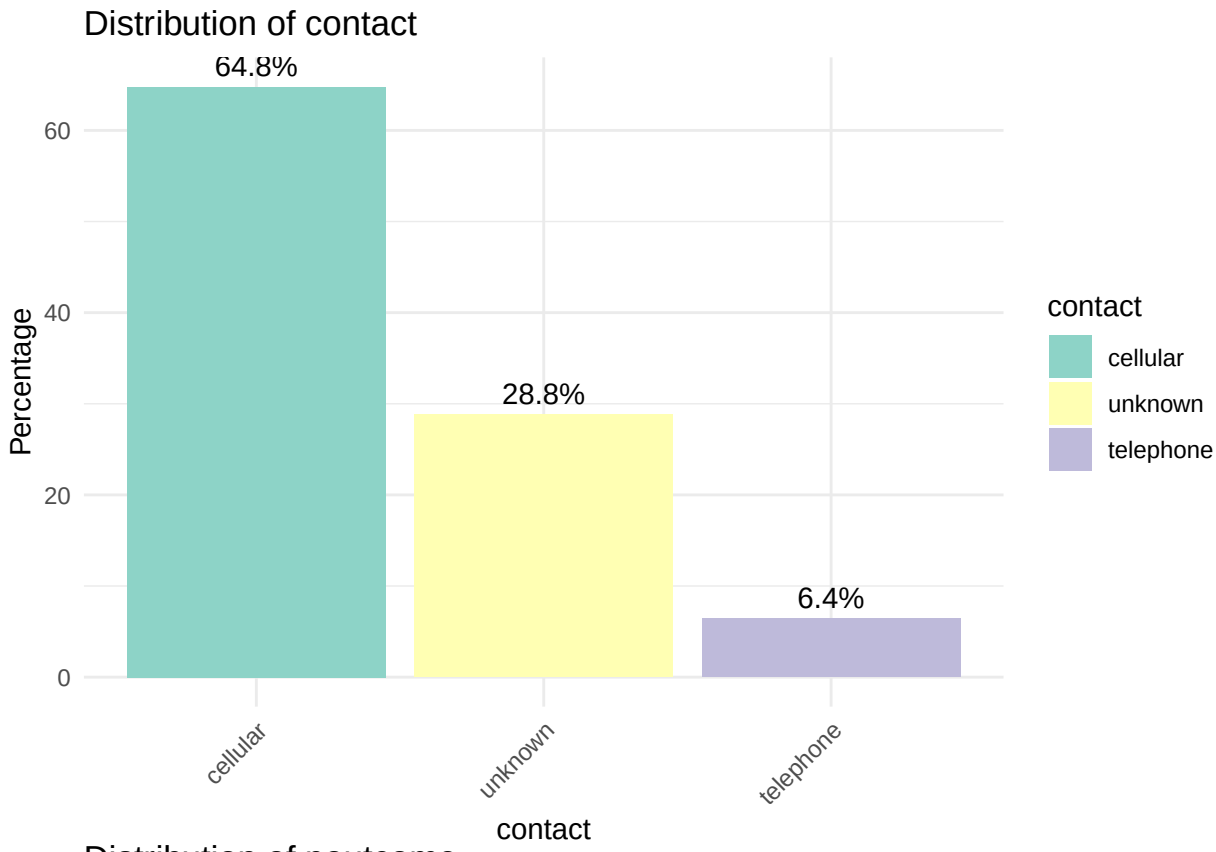

Distribution of marital
60.2%
28.3%
11.5%
Percentage
0
20
40
60
married
single
divorced
marital
marital
married
single
divorced

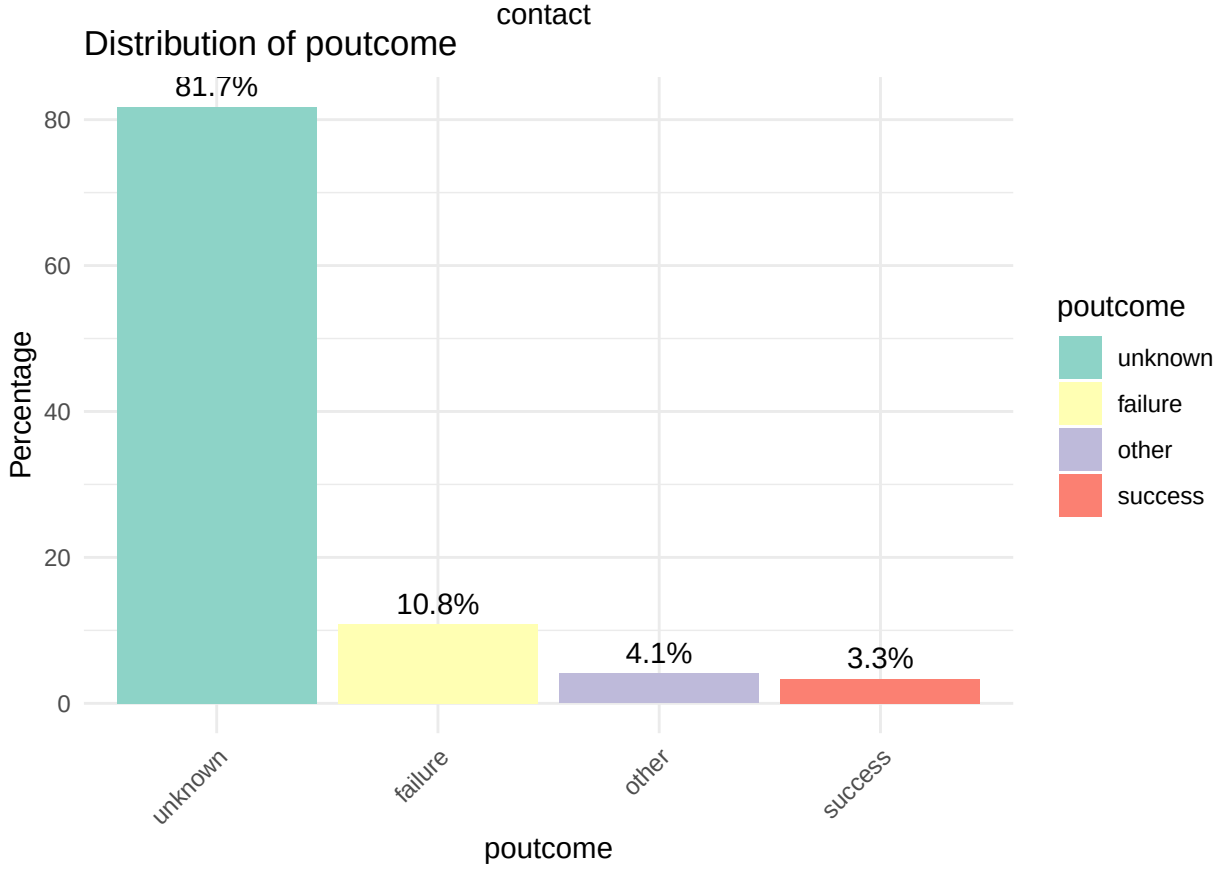

Distribution of education
51.3%
29.4%
15.2%
4.1%
50
40
30
20
10
0
Percentage
secondary
tertiary
primary
unknown
education
education
secondary
tertiary
primary
unknown

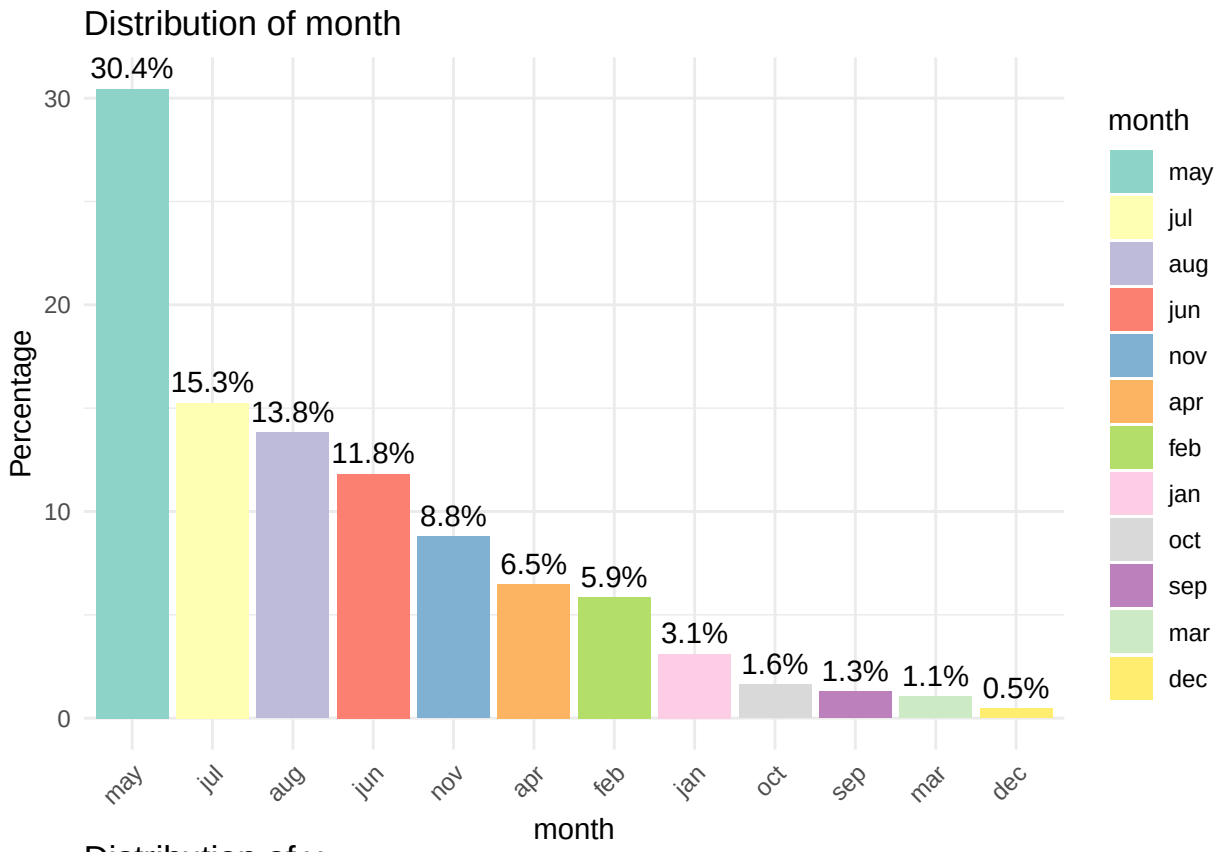

Distribution of default
98.2%
1.8%
100
75
50
25
0
Percentage
no
yes
default
default
no
yes

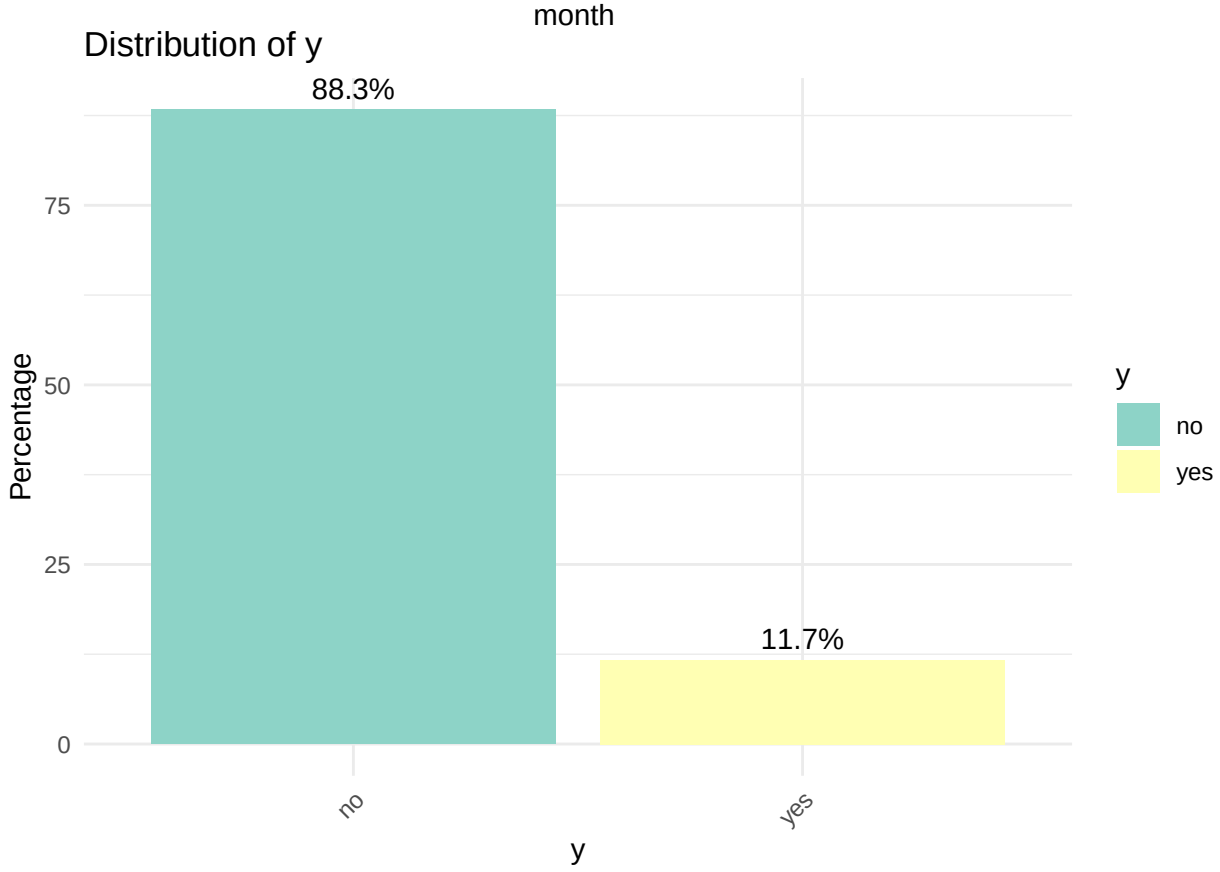

Distribution of housing
55.6%
44.4%
40
20
0
Percentage
yes
no
housing
housing
yes
no


Distribution of loan
84.0%
16.0%
80
60
40
20
0
Percentage
no
yes
loan
loan
no
yes


Distribution of contact
64.8%
28.8%
6.4%
Percentage
0
20
40
60
cellular
unknown
telephone
contact
contact
cellular
unknown
telephone


Distribution of poutcome
81.7%
10.8%
4.1%
3.3%
Percentage
0
20
40
60
80
unknown
failure
other
success
poutcome
poutcome
unknown
failure
other
success


Distribution of month
30.4%
15.3%
13.8%
11.8%
8.8%
6.5%
5.9%
3.1%
1.6%
1.3%
1.1%
0.5%
Percentage
30
20
10
0
may
jul
aug
jun
nov
apr
feb
jan
oct
sep
mar
dec
month
month
may
jul
aug
jun
nov
apr
feb
jan
oct
sep
mar
dec


Distribution of y
88.3%
11.7%
Percentage
75
50
25
0
no
yes
y
y
no
yes

Missing Data

From the code, there are no missing values. However, some client data has unknown values in categorical variables. About 83% of the client data has at least 1 "unknown" value in their data. From the distribution of categorical variables, the percentage of unknown value for: - job is 0.6% - education is 4.1% - contact is 6.4% - poutcome is 81.7%

```
sum(is.na(data))
```

```
## [1] 0
```

```
unknown_rows <- sum(apply(data, 1, function(row) any(row == "unknown")))

total_rows <- nrow(data)
percentage_unknown <- (unknown_rows / total_rows) * 100

percentage_unknown
```

```
## [1] 82.65466
```

There are no duplicated data

```
sum(duplicated(data))
```

```
## [1] 0
```

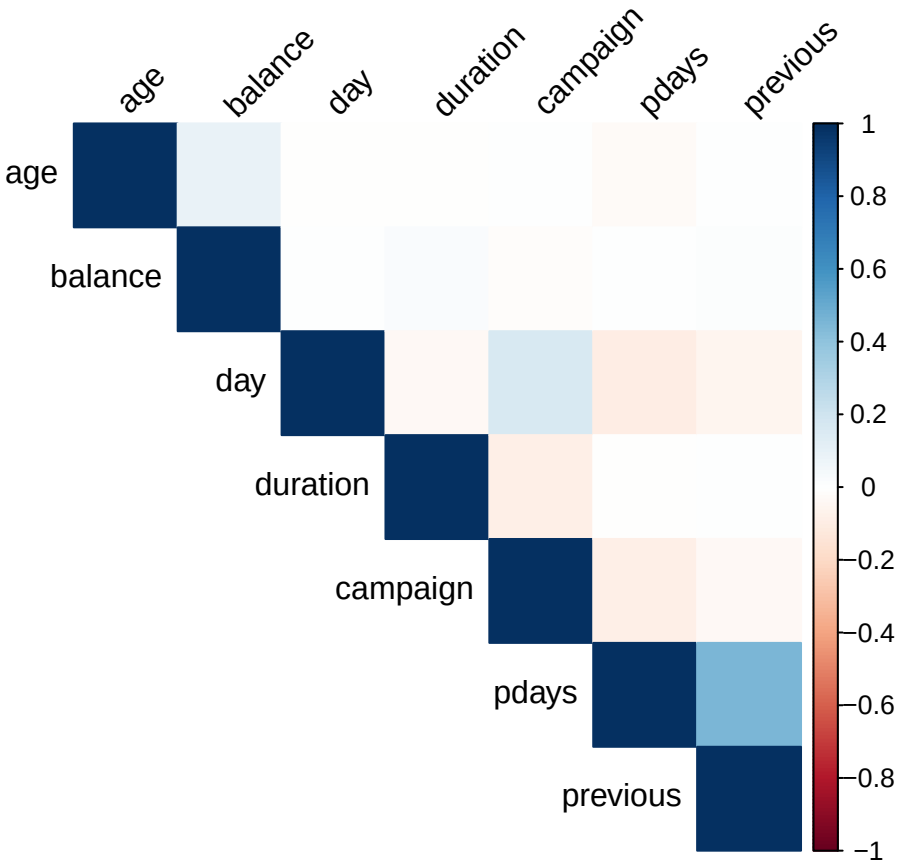
Relationship between variables

Age and balance appear to have a very weak correlation.

pdays and previous have a strong postive correlation.

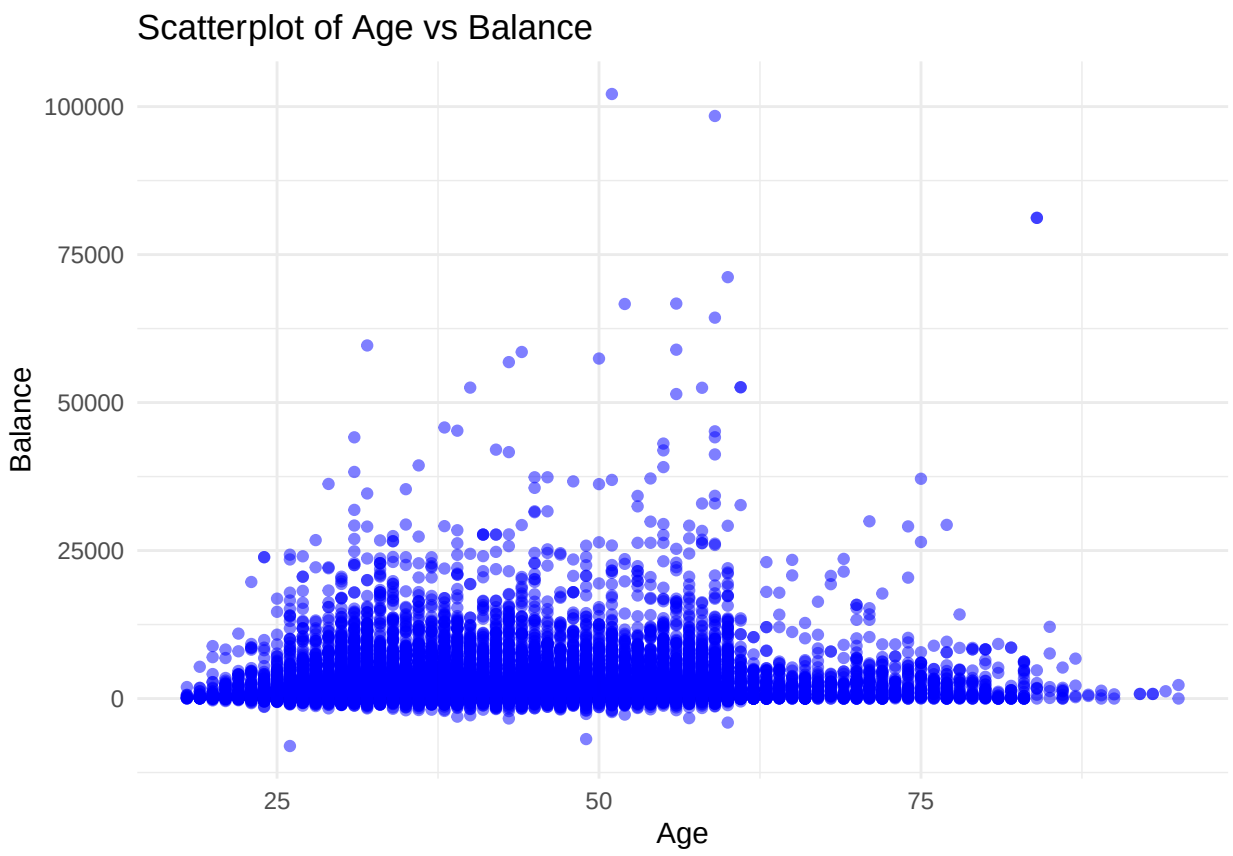
Campaign and duration has a slight negative correlation. This may indicate that the more times a client is contacted, the shorter the call duration becomes.

```
cor_matrix <- cor(data %>% select_if(is.numeric))

corrplot(cor_matrix, method="color", type="upper", tl.col="black", tl.srt=45)
```



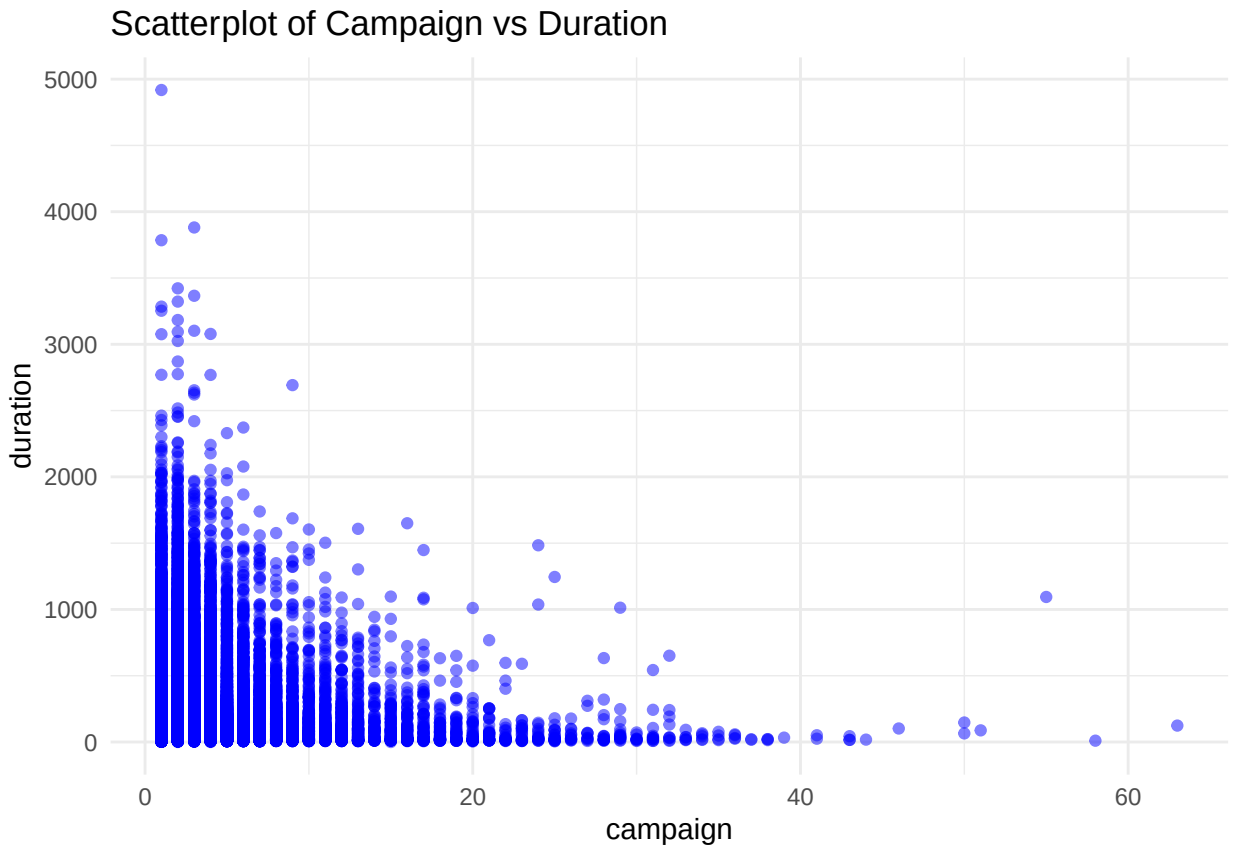
Age and Balance The majority of clients have a balance below 25,000. There are some outliers with significant higher balance. The balance seem spread out across all ages.

```
ggplot(data, aes(x = age, y = balance)) +
  geom_point(alpha = 0.5, color = "blue") +
  labs(title = "Scatterplot of Age vs Balance",
       x = "Age",
       y = "Balance") +
  theme_minimal()
```



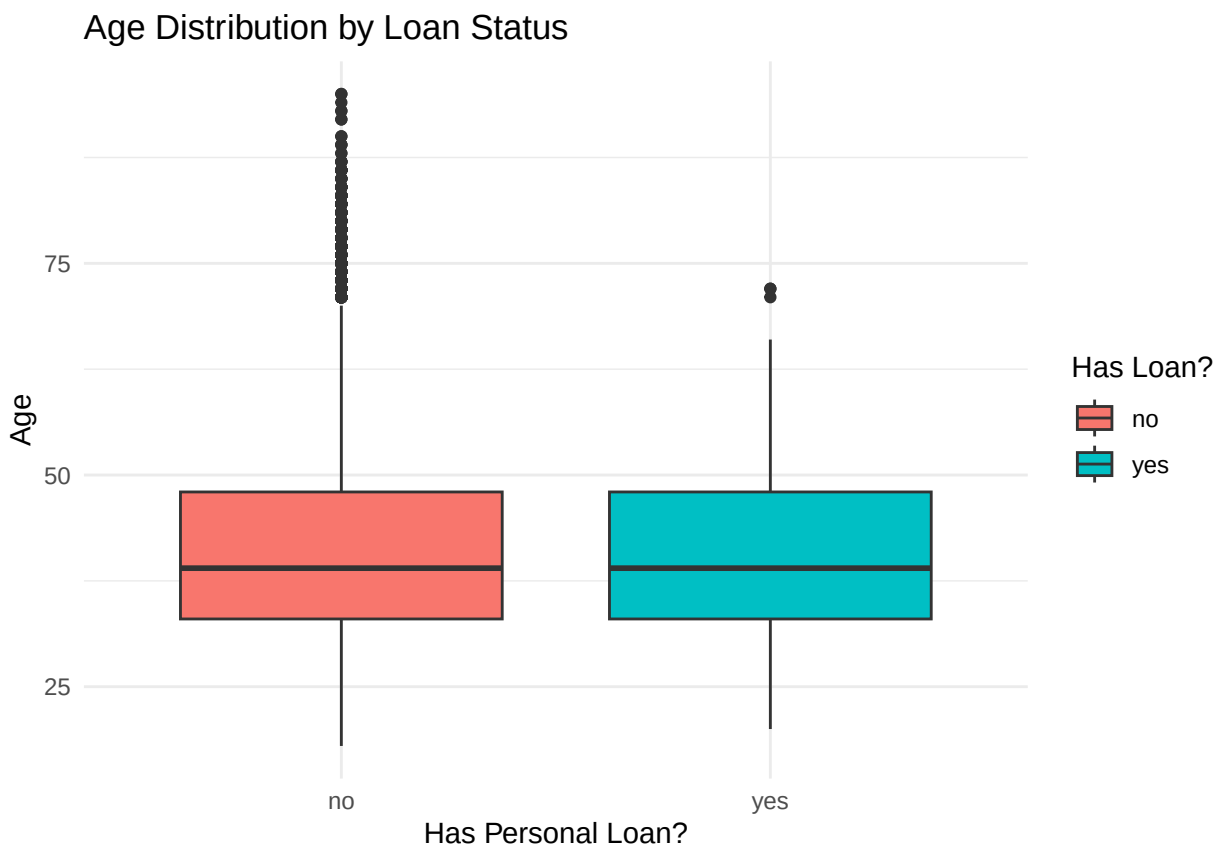
Campaign vs Duration Campaign and duration has a slight negative correlation. This may indicate that the more times a client is contacted, the shorter the call duration becomes.

```
ggplot(data, aes(x = campaign, y = duration)) +
  geom_point(alpha = 0.5, color = "blue") +
  labs(title = "Scatterplot of Campaign vs Duration",
       x = "campaign",
       y = "duration") +
  theme_minimal()
```



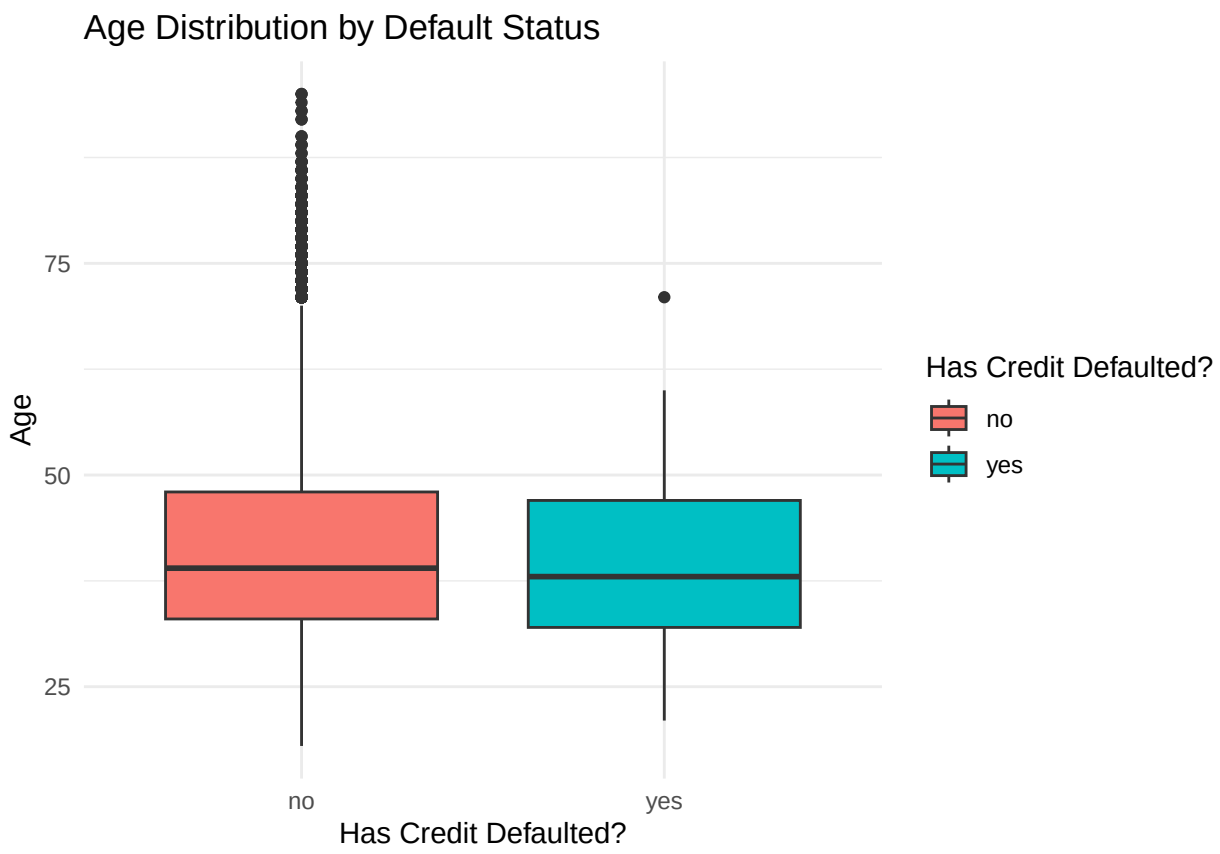
Age vs Loan Clients with no loan has a significant number of high-age outliers.

```
ggplot(data, aes(x = loan, y = age, fill = loan)) +
  geom_boxplot() +
  labs(title = "Age Distribution by Loan Status",
       x = "Has Personal Loan?",
       y = "Age",
       fill = "Has Loan?") +
  theme_minimal()
```



Age vs Default There are significantly more older age outliers in clients who does nothave a credit default. Older individuals are less likely to default as it has a higher median and broader age group.

```
ggplot(data, aes(x = default, y = age, fill = default)) +
  geom_boxplot() +
  labs(title = "Age Distribution by Default Status",
       x = "Has Credit Defaulted?",
       y = "Age",
       fill = "Has Credit Defaulted?") +
  theme_minimal()
```



Subscription v Client Demographics

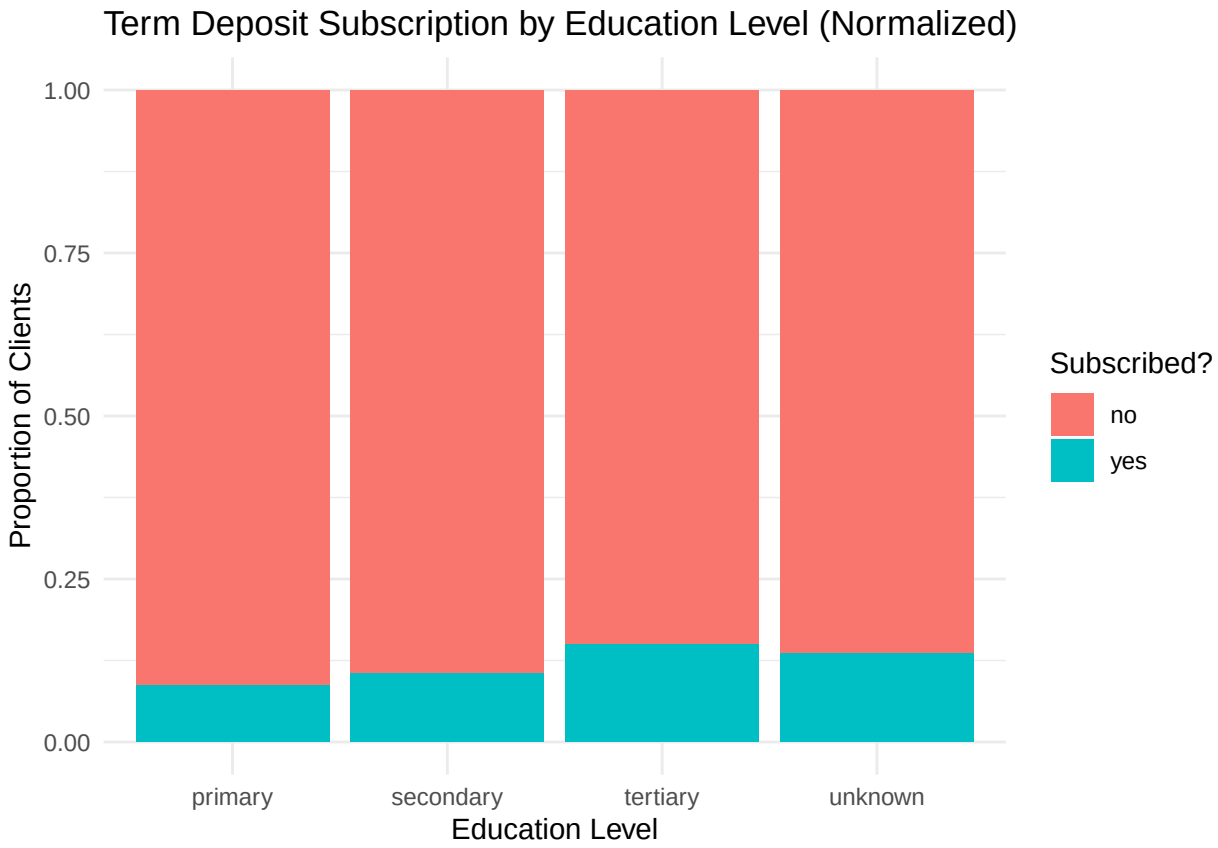
Since the subscription rate ("yes" for term deposit) is 11.7%, there is class imbalance. Proportions normalization the data within each group in necessary to make a fair comparison.

Subscription v Education Level There appears to be slight higher porportion in clients with teritary education level that did subscribe to the term deposit.

```
education_y_counts <- data %>%
  group_by(education, y) %>%
  summarize(n = n()) %>%
  ungroup()
```

```
## `summarise()` has grouped output by 'education'. You can override using the
## `.groups` argument.
```

```
ggplot(education_y_counts, aes(x = education, y = n, fill = y)) +
  geom_bar(stat = "identity", position = "fill") +
  labs(title = "Term Deposit Subscription by Education Level (Normalized)",
       x = "Education Level",
       y = "Proportion of Clients",
       fill = "Subscribed?") +
  theme_minimal()
```

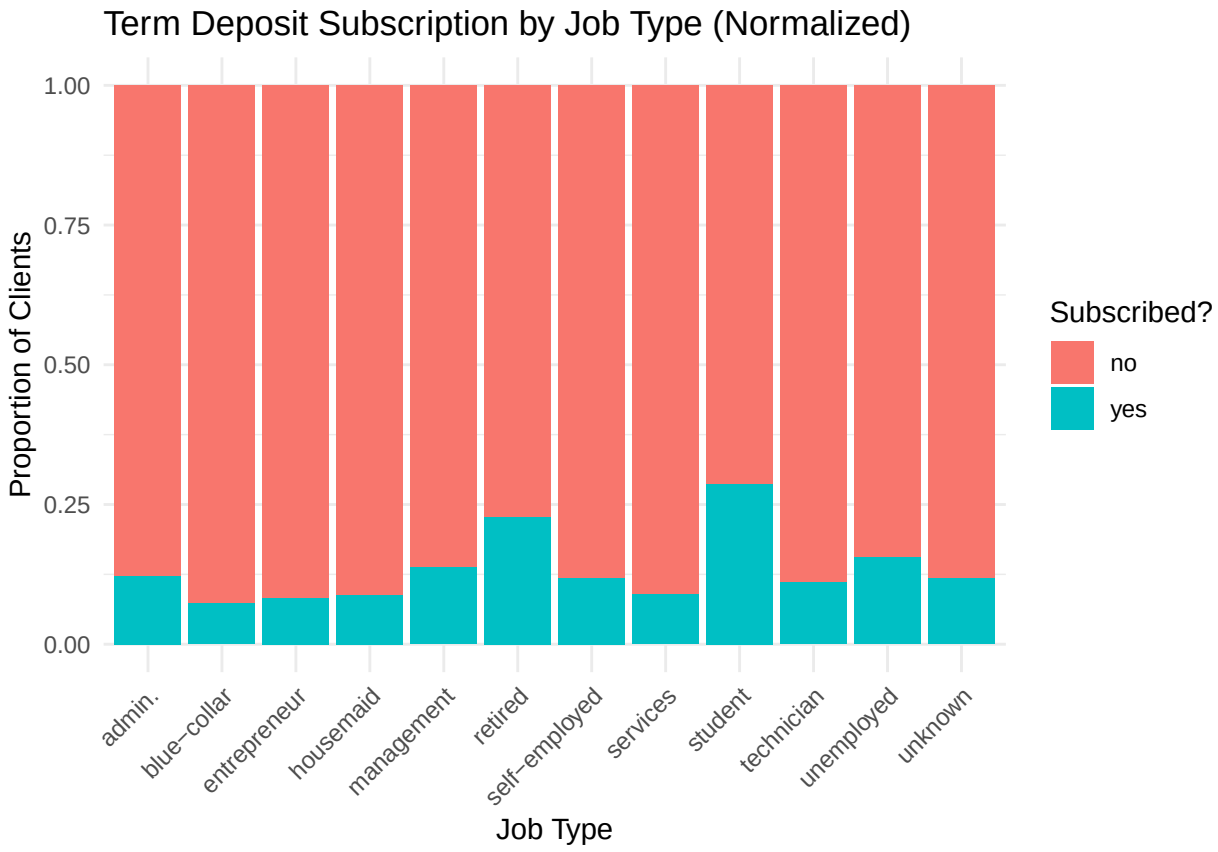


Subscription v Job Type The students, and retired clients has a higher percentage that did subscribe to the term deposit

```
job_y_counts <- data %>%
  group_by(job, y) %>%
  summarize(n = n()) %>%
  ungroup()
```

```
## `summarise()` has grouped output by 'job'. You can override using the `.groups`
## argument.
```

```
ggplot(job_y_counts, aes(x = job, y = n, fill = y)) +
  geom_bar(stat = "identity", position = "fill") +
  labs(title = "Term Deposit Subscription by Job Type (Normalized)",
       x = "Job Type",
       y = "Proportion of Clients",
       fill = "Subscribed?") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```

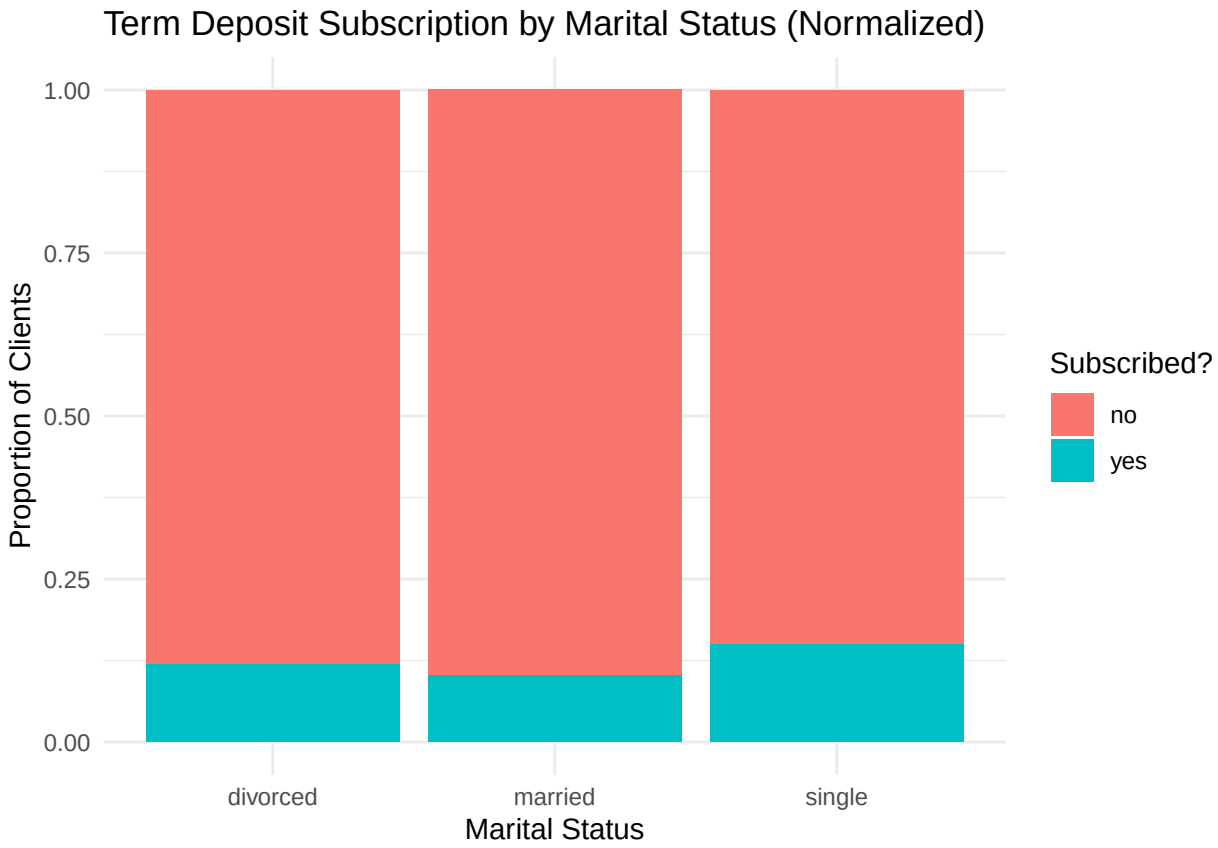


Subscription v Martial Status Clients who are single has a higher percentage that subscribe to the term deposit.

```
marital_y_counts <- data %>%
  group_by(marital, y) %>%
  summarize(n = n()) %>%
  ungroup()
```

```
## `summarise()` has grouped output by 'marital'. You can override using the
## `.groups` argument.
```

```
ggplot(marital_y_counts, aes(x = marital, y = n, fill = y)) +
  geom_bar(stat = "identity", position = "fill") +
  labs(title = "Term Deposit Subscription by Marital Status (Normalized)",
       x = "Marital Status",
       y = "Proportion of Clients",
       fill = "Subscribed?") +
  theme_minimal()
```

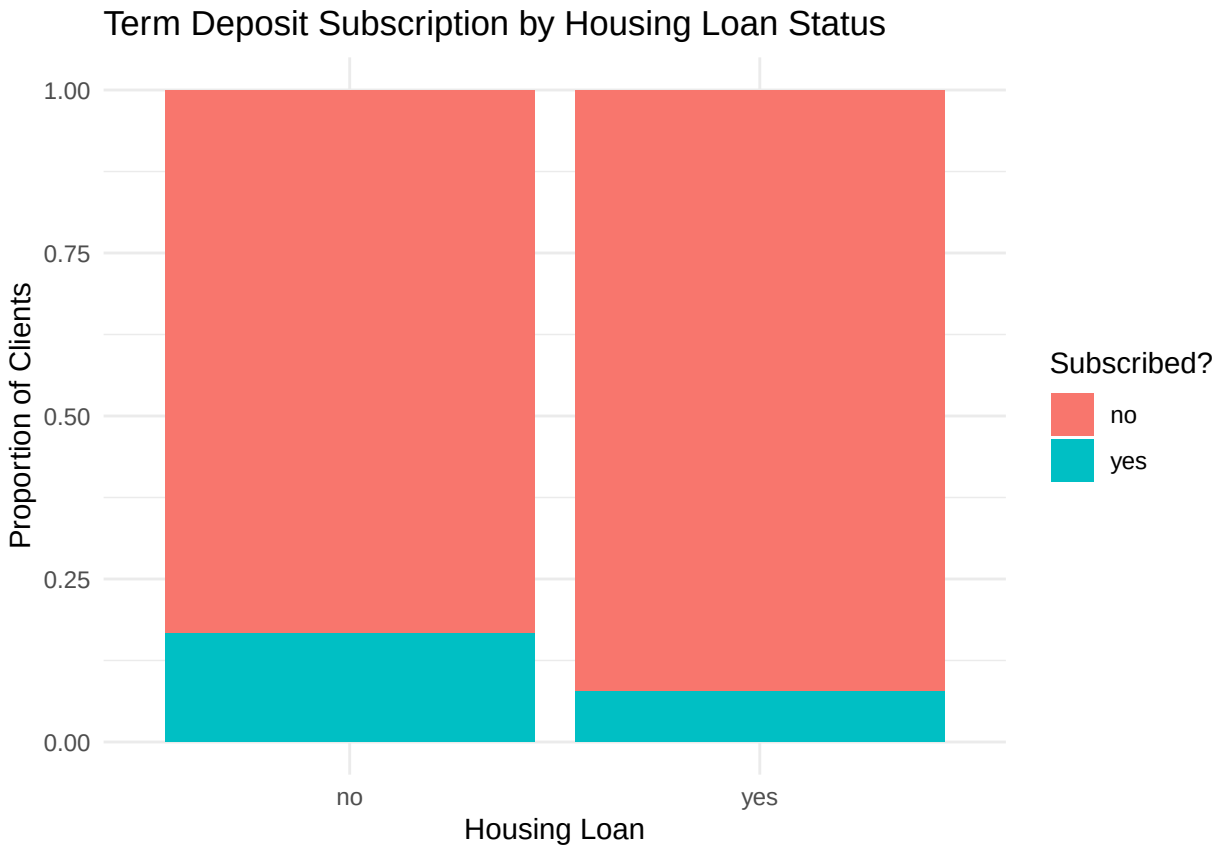


Subscription v Housing Loan Although it is minimal, there appears to be a slighter higher percentage of clients with housing loans that did subscribe to the term deposit.

```
housing_y_counts <- data %>%
  group_by(housing, y) %>%
  summarize(n = n()) %>%
  ungroup()
```

```
## `summarise()` has grouped output by 'housing'. You can override using the
## `.groups` argument.
```

```
ggplot(housing_y_counts, aes(x = housing, y = n, fill = y)) +
  geom_bar(stat = "identity", position = "fill") +
  labs(title = "Term Deposit Subscription by Housing Loan Status",
       x = "Housing Loan",
       y = "Proportion of Clients",
       fill = "Subscribed?") +
  theme_minimal()
```



Subscription v Age There are higher subscription rate among the younger (18-25) and older age group (65+).

```
data <- data %>%
  mutate(age_group = case_when(
    age >= 18 & age <= 25 ~ "18-25",
    age >= 26 & age <= 35 ~ "26-35",
    age >= 36 & age <= 45 ~ "36-45",
    age >= 46 & age <= 55 ~ "46-55",
    age >= 56 & age <= 65 ~ "56-65",
    age >= 66 & age <= 75 ~ "66-75",
    age >= 76 ~ "76+",
    TRUE ~ NA_character_
  ))

table(data$age_group)
```

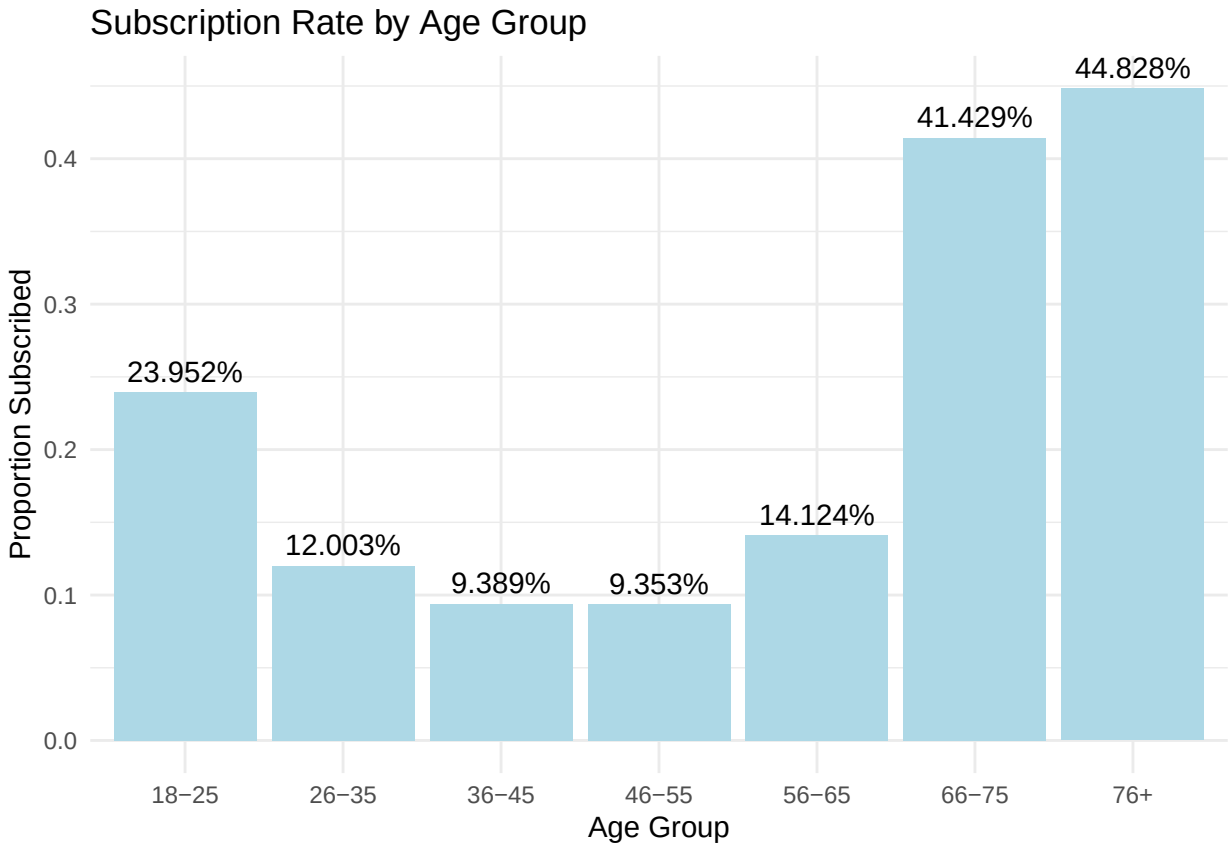
```
##
## 18-25 26-35 36-45 46-55 56-65 66-75   76+
##  1336 15571 13856  9548  4149   490   261
```

```
age_group_summary <- data %>%
  group_by(age_group) %>%
  summarise(
    total = n(),
    subscribed_yes = sum(y == "yes"),
    prop_subscribed = subscribed_yes / total
```

```
  ) %>%
  arrange(age_group)

print(age_group_summary)
```

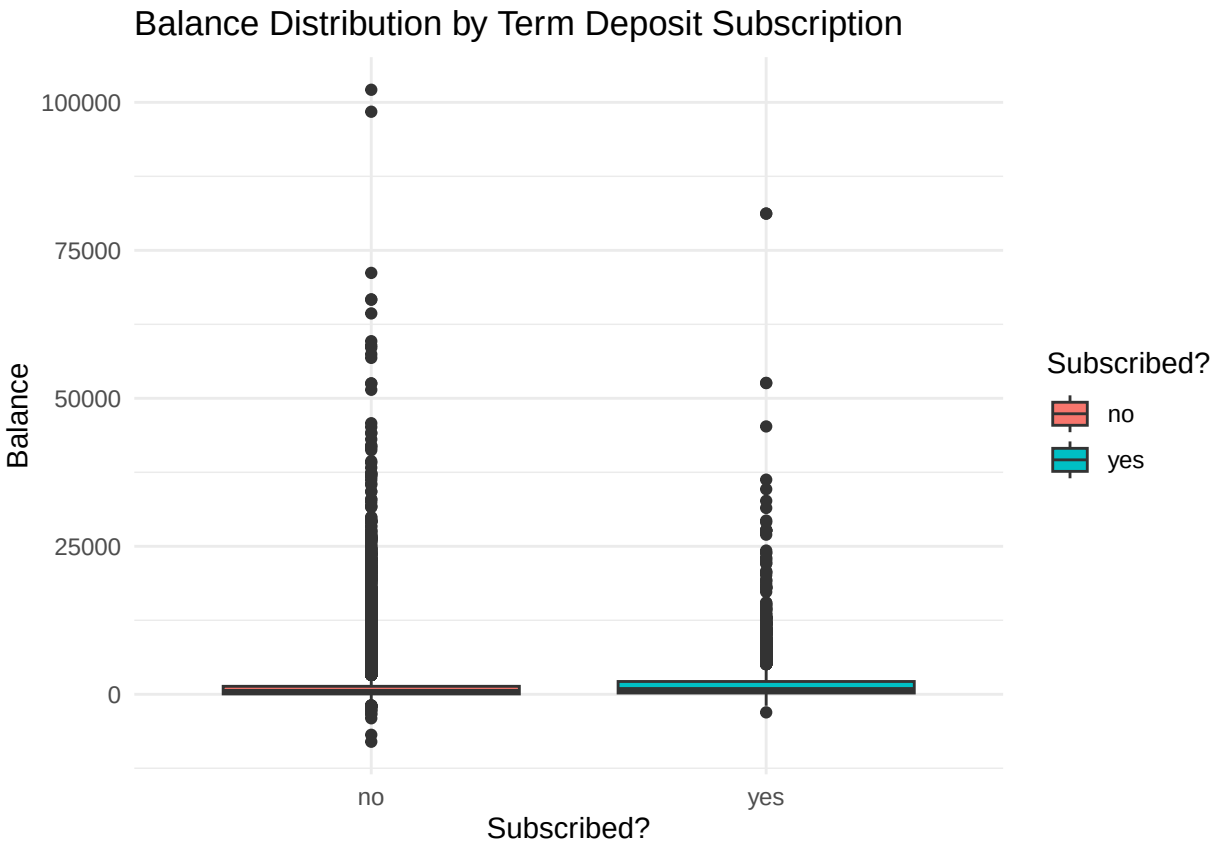
```
## # A tibble: 7 x 4
##   age_group total subscribed_yes prop_subscribed
##   <chr>     <int>          <int>           <dbl>
## 1 18-25      1336            320          0.240
## 2 26-35     15571           1869          0.120
## 3 36-45     13856           1301          0.0939
## 4 46-55      9548            893          0.0935
## 5 56-65      4149            586          0.141
## 6 66-75       490            203          0.414
## 7 76+         261            117          0.448
```

```
ggplot(age_group_summary, aes(x = age_group, y = prop_subscribed)) +
  geom_bar(stat = "identity", fill = "light blue") +
  geom_text(aes(label = scales::percent(prop_subscribed)), vjust = -0.5) +
  labs(
    title = "Subscription Rate by Age Group",
    x = "Age Group",
    y = "Proportion Subscribed"
  ) +
  theme_minimal()
```



Subscription v Balance Extreme outliers are present in both subscribers and non-scribers. However, it is more prominenet in non-subscribers. Both groups present a wide range of balances, but non-subscribers has more high-balance outliers.

```
ggplot(data, aes(x = y, y = balance, fill = y)) +
  geom_boxplot() +
  labs(title = "Balance Distribution by Term Deposit Subscription",
       x = "Subscribed?",
       y = "Balance",
       fill = "Subscribed?") +
  theme_minimal()
```



Subscription v Other variables

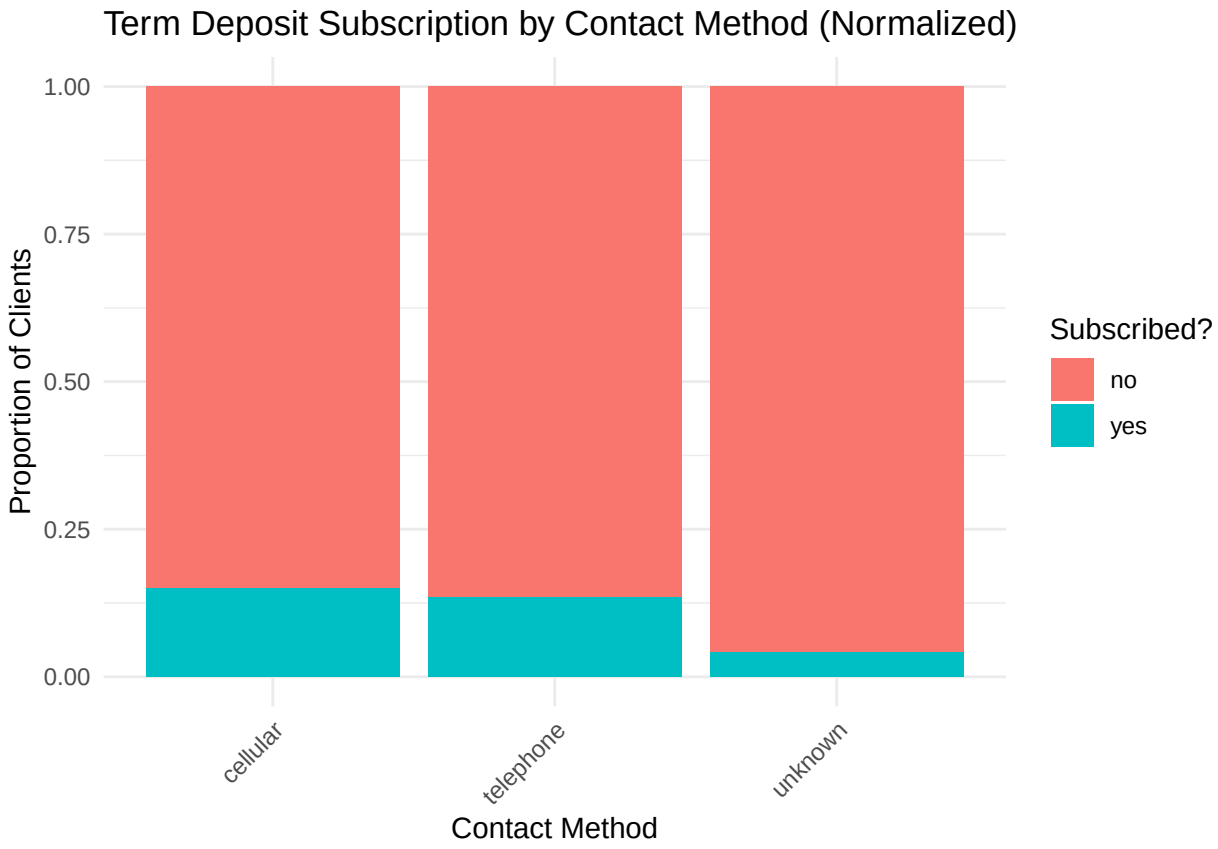
Subscription v Contact Method Celluar has the highest subscription rate, making it the most effective method of marketing campaign.

```
contact_y_counts <- data %>%
  group_by(contact, y) %>%
  summarize(n = n()) %>%
  ungroup()
```

```
## `summarise()` has grouped output by 'contact'. You can override using the
## `.groups` argument.
```

```
ggplot(contact_y_counts, aes(x = contact, y = n, fill = y)) +
  geom_bar(stat = "identity", position = "fill") +
  labs(title = "Term Deposit Subscription by Contact Method (Normalized)",
       x = "Contact Method",
```

```
      y = "Proportion of Clients",
      fill = "Subscribed?") +
  theme_minimal() +
    theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



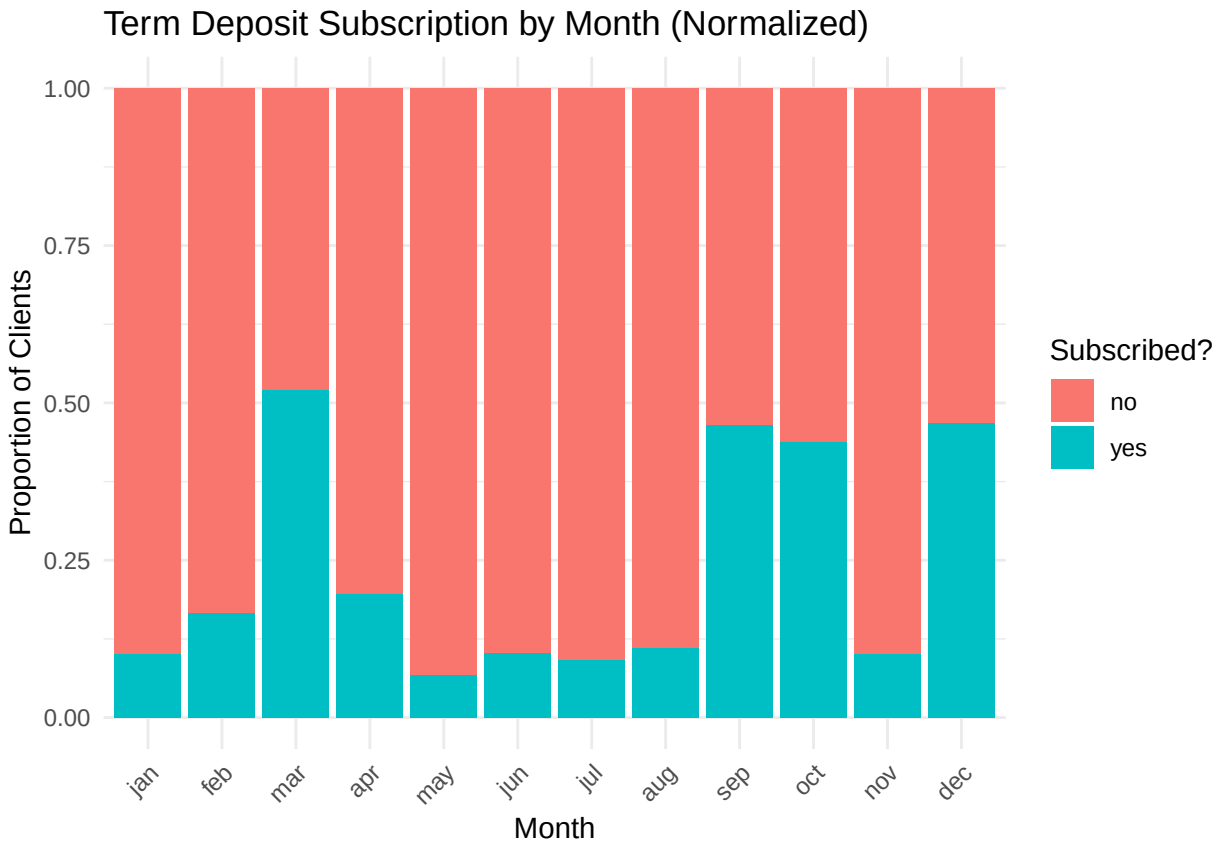
Subscription v Month Subscription rate varies month to month. March, September, and October show higher proportions of term deposit subscriptions. Recall from earlier on the distribution of the months, campaign efforts were aggressive in May at it accounts for about 30% of the campaign, followed by July (15.3%) and Auguest (13.8%). Yet, these months were among the lowest subscription rates, suggesting a not very successful term deposit campaign.

```
month_y_counts <- data %>%
  group_by(month, y) %>%
  summarize(n = n()) %>%
  ungroup()
```

```
## `summarise()` has grouped output by 'month'. You can override using the
## `.groups` argument.
```

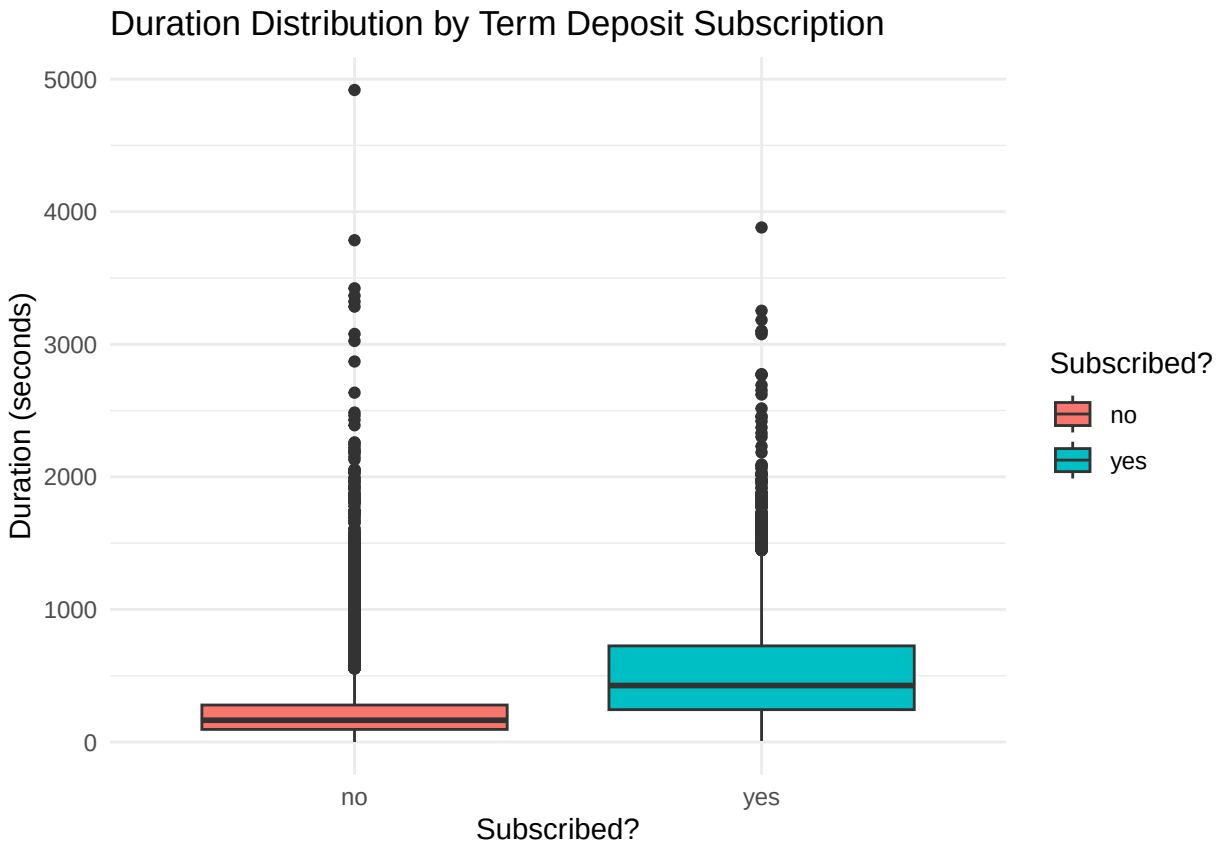
```
# b. Create the normalized stacked bar chart
ggplot(month_y_counts, aes(x = month, y = n, fill = y)) +
  geom_bar(stat = "identity", position = "fill") +
  labs(title = "Term Deposit Subscription by Month (Normalized)",
       x = "Month",
       y = "Proportion of Clients",
       fill = "Subscribed?") +
  theme_minimal() +
```

```
  scale_x_discrete(limits = c("jan", "feb", "mar", "apr", "may", "jun", "jul", "aug", "sep", "oct", "nov
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



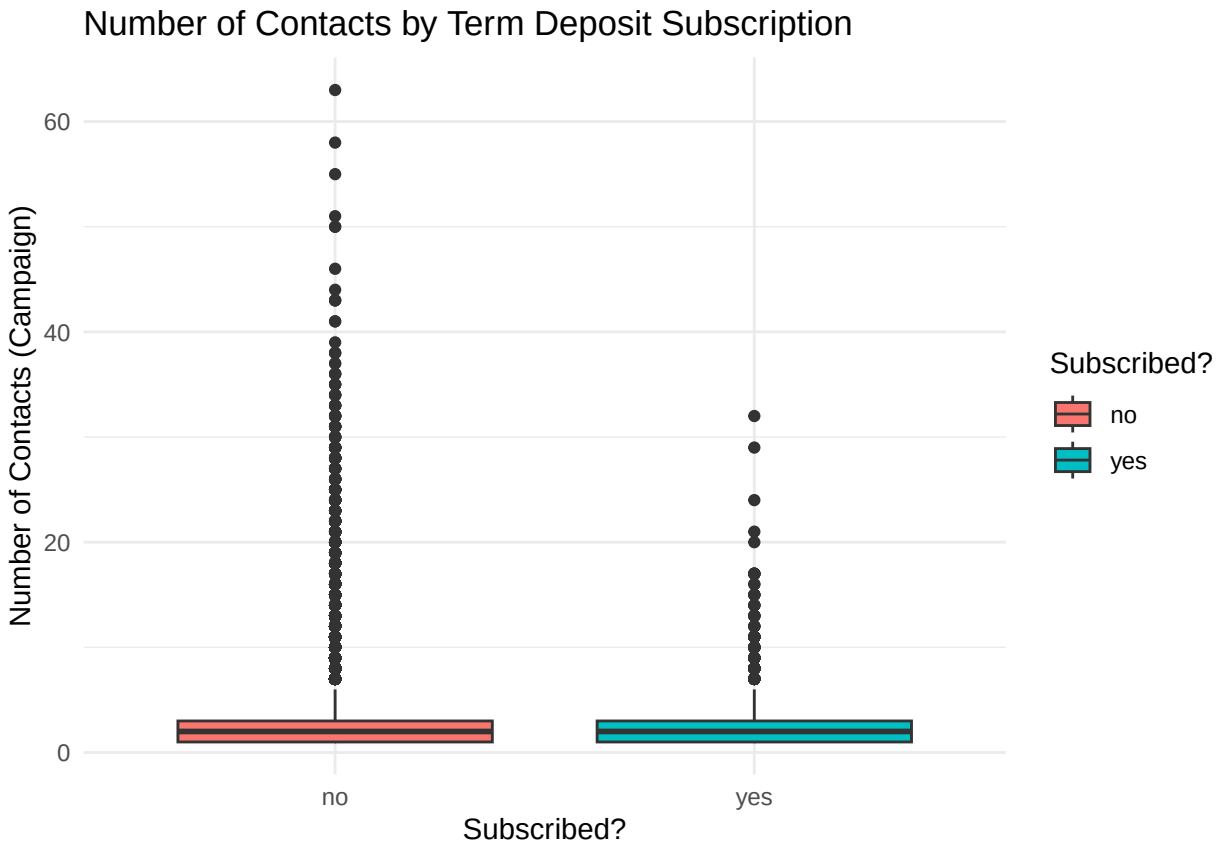
Subscription v Duration Both groups of subscribers and non-subscribers present outliers with very long contact duration. However, it more present in subscribers. The spread of contact duration is greater in subscribers.

```
ggplot(data, aes(x = y, y = duration, fill = y)) +
  geom_boxplot() +
  labs(title = "Duration Distribution by Term Deposit Subscription",
       x = "Subscribed?",
       y = "Duration (seconds)",
       fill = "Subscribed?") +
  theme_minimal()
```



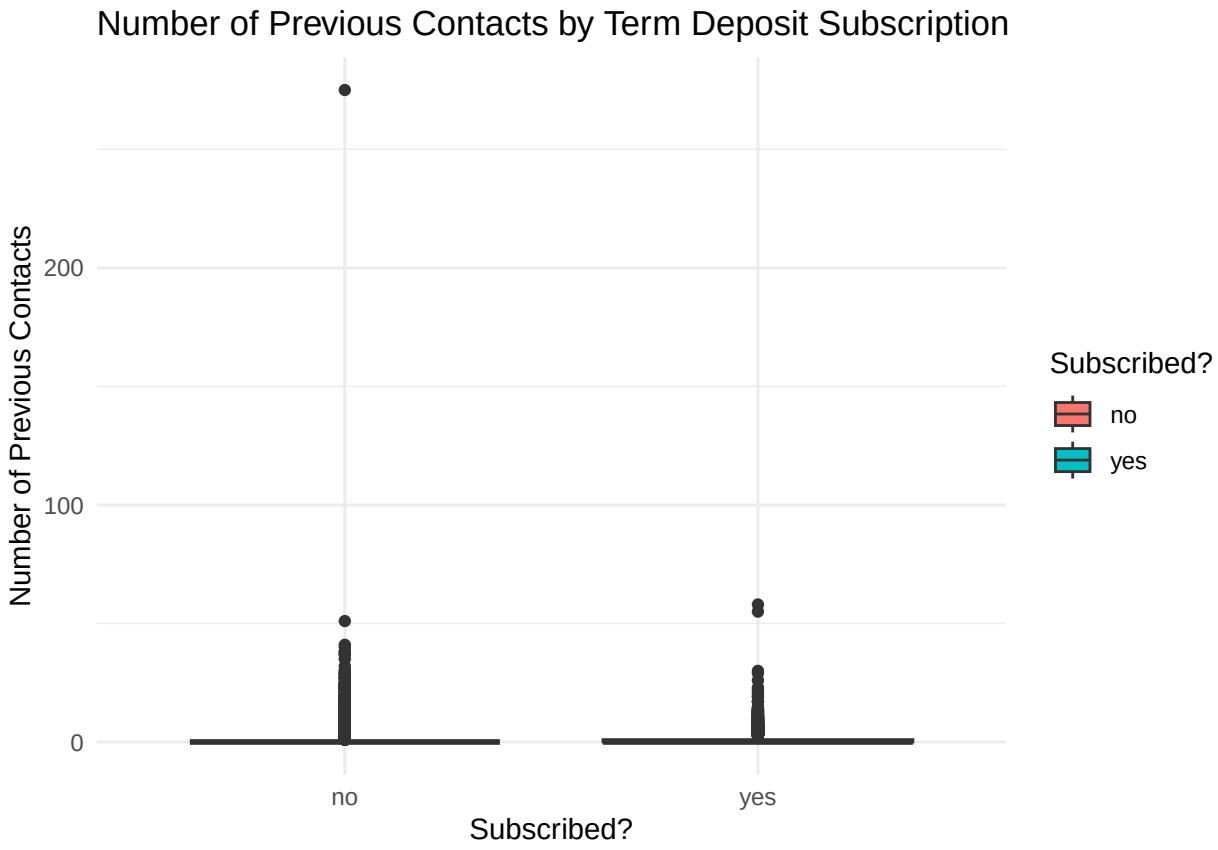
Subscription v Number of contacts Outliers are present in both groups but there are extreme outliers where clients are contacted over 40 times. Repeated contacts may not guarantee more subscribers.

```
ggplot(data, aes(x = y, y = campaign, fill = y)) +
  geom_boxplot() +
  labs(title = "Number of Contacts by Term Deposit Subscription",
       x = "Subscribed?",
       y = "Number of Contacts (Campaign)",
       fill = "Subscribed?") +
  theme_minimal()
```



Subscription v Number of previous contacts Both groups show similar distribution. Most clients regardless if they subscribed or not have very few previous contact. The number of previous contacts may not have much influence on subscription rates.

```
ggplot(data, aes(x = y, y = previous, fill = y)) +
  geom_boxplot() +
  labs(title = "Number of Previous Contacts by Term Deposit Subscription",
       x = "Subscribed?",
       y = "Number of Previous Contacts",
       fill = "Subscribed?") +
  theme_minimal()
```



Algorithm Selection

Since there are labels (term deposits subscriber: yes or no) in our dataset, I would recommend to use supervised learning algorithm such as Random Forest. About 83% of the data contains at 1 "unknown" value. Random Forest can handle missing values without the step of imputation. The dataset contains only about 11% subscribers which shows class imbalance. Random Forest is less sensitive to class imbalance. It uses ensemble learning, which reduces overfitting. However, it is computational slow to use as we have a large dataset. Alternatively, if there were fewer than 1,000 data records, we can use Naive Bases as it is simple and fast, and works well with categorical data. However, it may be sensitive to class imbalance.

Pre-Processing Steps

For optimal model performance, we need to perform pre-processing steps such as data cleaning. We need to address the "unknown value" in the categorical variables. The variable poutcome (previous outcome) has the highest percentage (81.7%) of "unknown" values. I would remove it as most clients regardless if they subscribed or not have very few previous contact. The number of previous contacts may not have much influence on subscription rates.

Due to the right skewness of all numerical variable except day, we would need to perform data transformation such as box cox transformation. Categorical variables will needed to be converted to numerical format using label encoding. Moreover, we should introduce new features such as age groups and contact duration (short, medium, long). Sampling techniques such as SMOTE should be use to address imbalance class.