CS174 FA1.2

Darryl Babar, Christian Henry Miguel Caruz, Miguel Soniel & Jan Edgar Tupas BM1

The Bank Marketing dataset from the UCI machine learning repository is a dataset that contains data about the marketing campaign of a Portugese bank. The dataset has categorical and numerical variables with a total of 17 variables. The categorical variables include the following: job; marital; education; contact; poutome. The numerical variables include the following: age; default; balance; housing; loan; duration; campaign; pdays; previous; y.

Variable Name	Role	Type	Demographic	Description
age	Feature	Integer	Age	
job	Feature	Categorical	Occupation	type of job (categorical: 'admin.','blue-collar','entrepreneur','housemaid','mana employed','services','student','technician','unemployed','unknown')
marital	Feature	Categorical	Marital Status	marital status (categorical: 'divorced','married','single','unknown'; note: 'divorce widowed)
education	Feature	Categorical	Education Level	(categorical: 'basic.4y','basic.6y','basic.9y','high.school','illiterate','professional.course','univer
default	Feature	Binary		has credit in default?
balance	Feature	Integer		average yearly balance
housing	Feature	Binary		has housing loan?
loan	Feature	Binary		has personal loan?
contact	Feature	Categorical		contact communication type (categorical: 'cellular','telephone')
day_of_week	Feature	Date		last contact day of the week
month	Feature	Date		last contact month of year (categorical: 'jan', 'feb', 'mar', ..., 'nov', 'dec')
duration	Feature	Integer		last contact duration, in seconds (numeric). Important note: this attribute highly affects the output target (e.g., if duration=0 then y='no'). Yet, the duration is not known before a call is performed. Also, after the end of the call y is obviously known. Thus, this input should only be included for benchmark purposes and should be discarded if the intention is to have a realistic predictive model.
campaign	Feature	Integer		number of contacts performed during this campaign and for this client (numeric, includes last contact)
pdays	Feature	Integer		number of days that passed by after the client was last contacted from a previous campaign (numeric; -1 means client was not previously contacted)
previous	Feature	Integer		number of contacts performed before this campaign and for this client
poutcome	Feature	Categorical		outcome of the previous marketing campaign (categorical: 'failure','nonexistent','success')
y	Target	Binary		has the client subscribed a term deposit?

The variable y is a binary output that shows whether the client subscribed to a term deposit or not. So, we use this as the dependent variable. The remaining variables and their values can be used as the independent variables that can determine the output of the dependent variable y. We can further explore the independent variables to see if it can provide additional information on their relationship with the

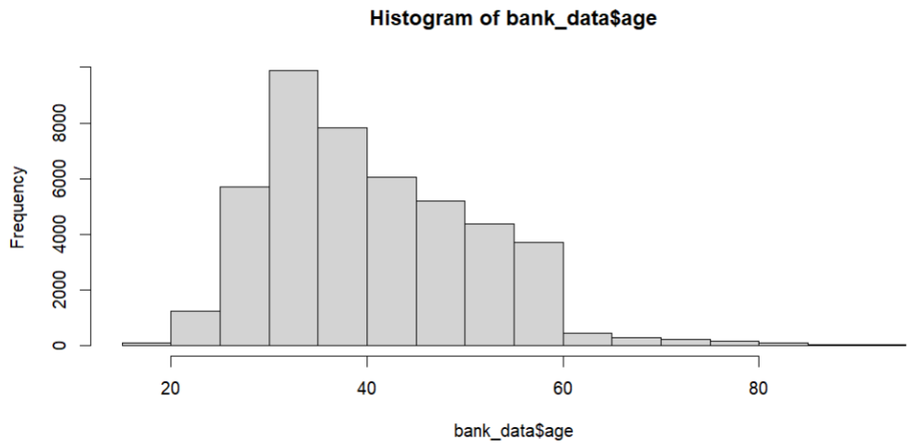
dependent variable. Here, we install the necessary libraries, namely tidyverse and ggplot2, and prepare the dataset by initializing and storing it in bank_data. We show the first six rows of the dataframe.

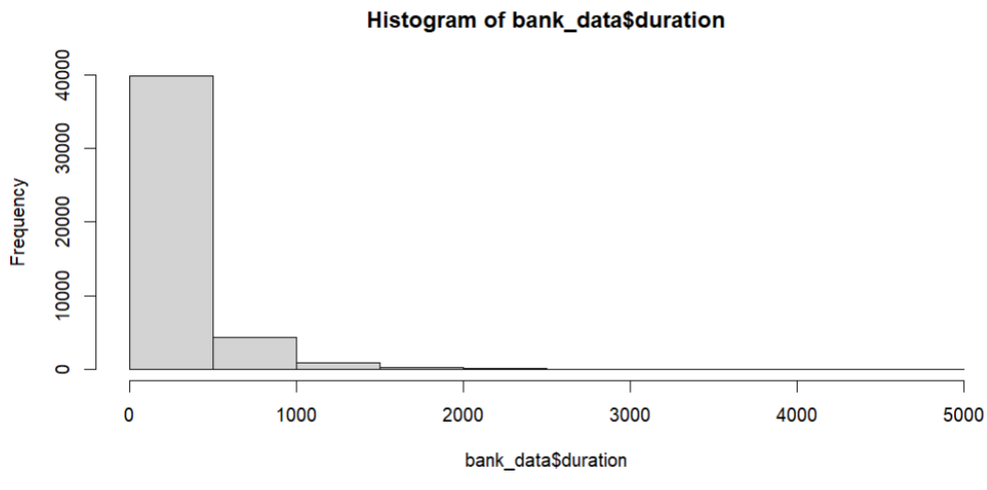
```
install.packages("tidyverse")
library(tidyverse)

install.packages("ggplot2")
library(ggplot2)

bank_data <- read.csv(file = "bank-full.csv", head = TRUE, sep = ";")

head(bank_data)
```

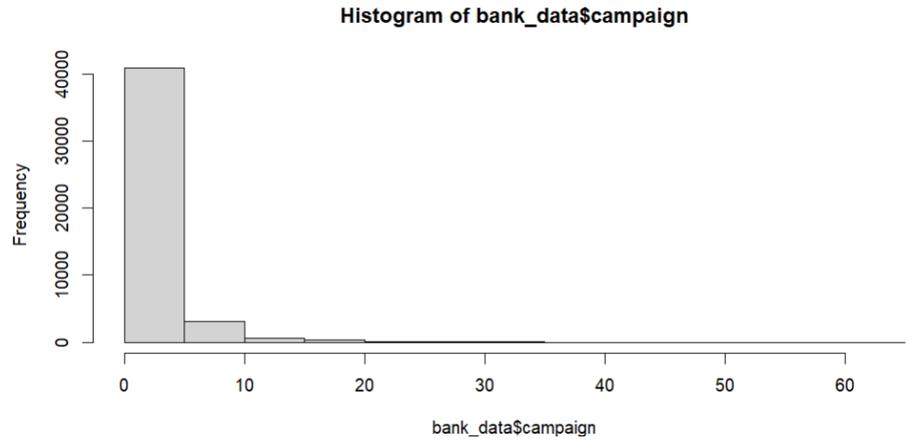
```
  age          job marital education default balance housing loan
1  58   management married  tertiary      no    2143     yes   no
2  44   technician  single secondary      no      29     yes   no
3  33 entrepreneur married secondary      no       2     yes  yes
4  47  blue-collar married   unknown      no    1506     yes   no
5  33      unknown  single   unknown      no       1      no   no
6  35   management married  tertiary      no     231     yes   no
  contact day month duration campaign pdays previous poutcome  y
1 unknown   5   may      261        1    -1        0  unknown no
2 unknown   5   may      151        1    -1        0  unknown no
3 unknown   5   may       76        1    -1        0  unknown no
4 unknown   5   may       92        1    -1        0  unknown no
5 unknown   5   may      198        1    -1        0  unknown no
6 unknown   5   may      139        1    -1        0  unknown no
```

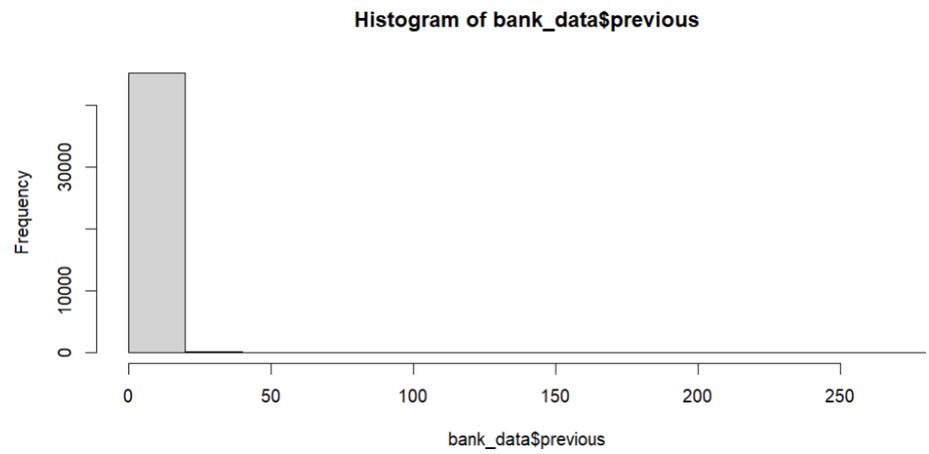
First we will visualize the numerical categories. We will use the histogram to show the numerical categories and their frequency.

```
hist(bank_data$age)
hist(bank_data$duration)
hist(bank_data$campaign)
hist(bank_data$previous)
```









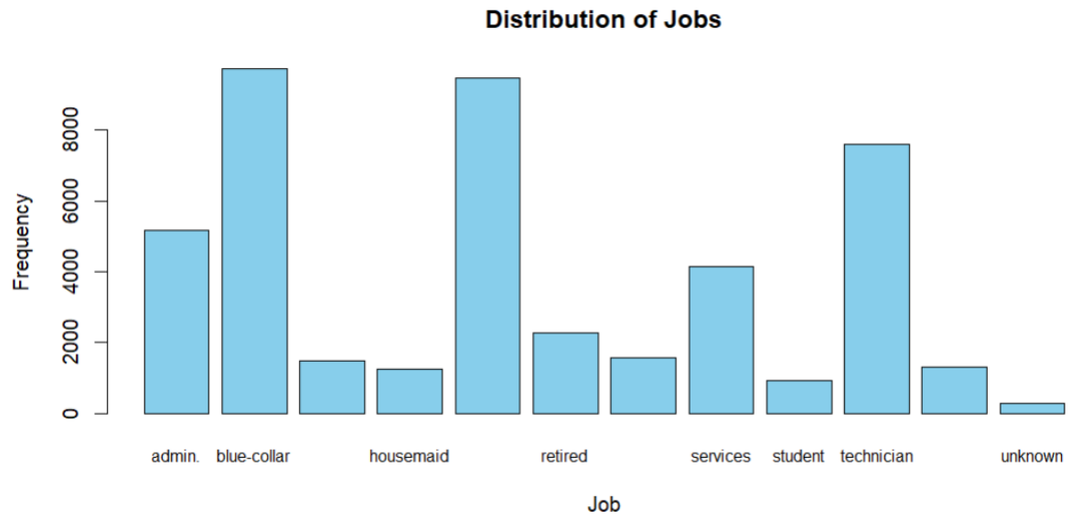
Now we will visualize the categorical variables. Here we use barplots to show their frequency.

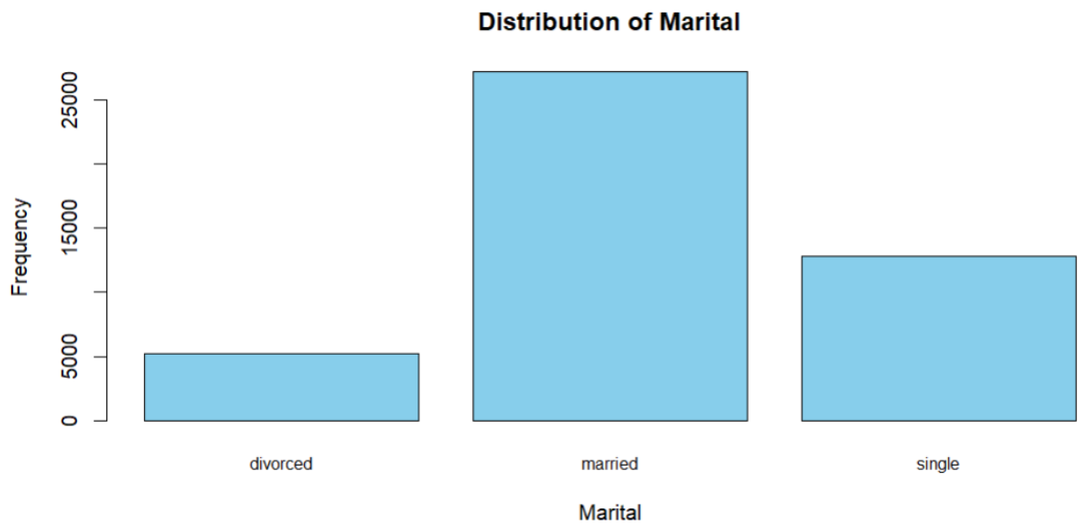
```
job_counts <- table(bank_data$job)
barplot(job_counts,
        main = "Distribution of Jobs",
        xlab = "Job",
        ylab = "Frequency",
        col = "skyblue",
        names.arg = as.character(names(job_counts)),
        cex.names = 0.8)

marital_counts <- table(bank_data$marital)
barplot(marital_counts,
        main = "Distribution of Marital",
        xlab = "Marital",
        ylab = "Frequency",
        col = "skyblue",
        names.arg = as.character(names(marital_counts)),
        cex.names = 0.8)
```

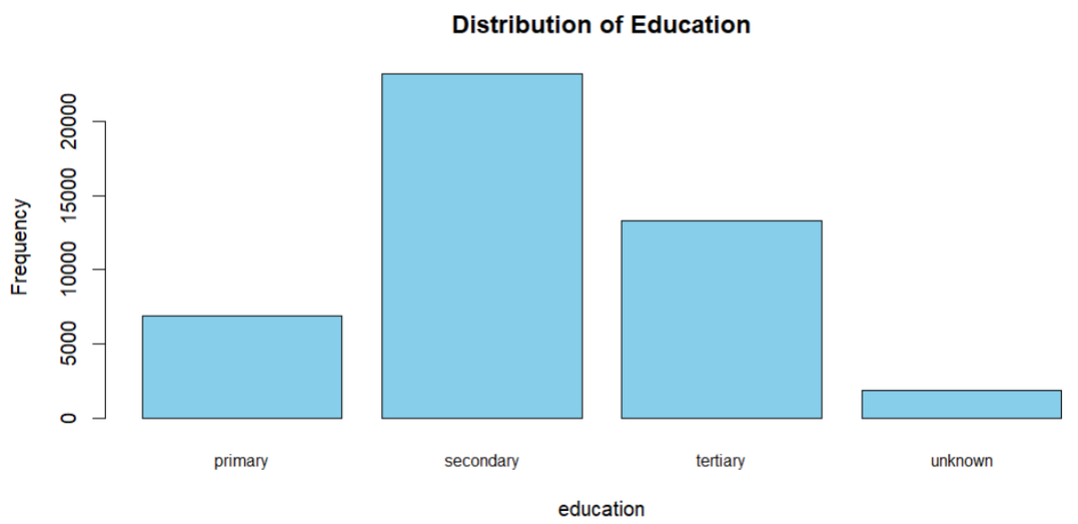
```
education_counts <- table(bank_data$education)
barplot(education_counts,
        main = "Distribution of Education",
        xlab = "education",
        ylab = "Frequency",
        col = "skyblue",
        names.arg = as.character(names(education_counts)),
        cex.names = 0.8)

contact_counts <- table(bank_data$contact)
barplot(marital_counts,
        main = "Distribution of contact",
        xlab = "contact",
        ylab = "Frequency",
        col = "skyblue",
        names.arg = as.character(names(contact_counts)),
        cex.names = 0.8)
```

```
default_counts <- table(bank_data$default)
barplot(default_counts,
        main = "Distribution of contact",
        xlab = "default",
        ylab = "Frequency",
        col = "skyblue",
        names.arg = as.character(names(default_counts)),
        cex.names = 0.8)

housing_counts <- table(bank_data$housing)
barplot(housing_counts,
        main = "Distribution of housing",
        xlab = "housing",
        ylab = "Frequency",
        col = "skyblue",
        names.arg = as.character(names(housing_counts)),
        cex.names = 0.8)
```

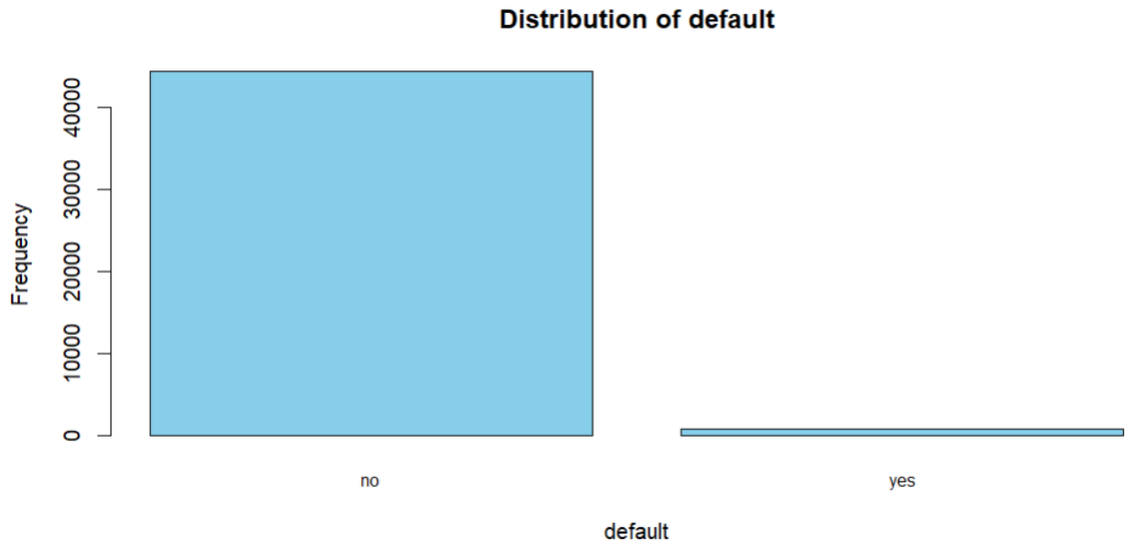
```
loan_counts <- table(bank_data$loan)
barplot(loan_counts,
        main = "Distribution of loan",
        xlab = "loan",
        ylab = "Frequency",
        col = "skyblue",
        names.arg = as.character(names(loan_counts)),
        cex.names = 0.8)

y_counts <- table(bank_data$y)
barplot(y_counts,
        main = "Distribution of loan",
        xlab = "y",
        ylab = "Frequency",
        col = "skyblue",
        names.arg = as.character(names(loan_counts)),
        cex.names = 0.8)
```


Distribution of Jobs
Frequency
0
2000
4000
6000
8000
admin.
blue-collar
housemaid
retired
services
student
technician
unknown
Job

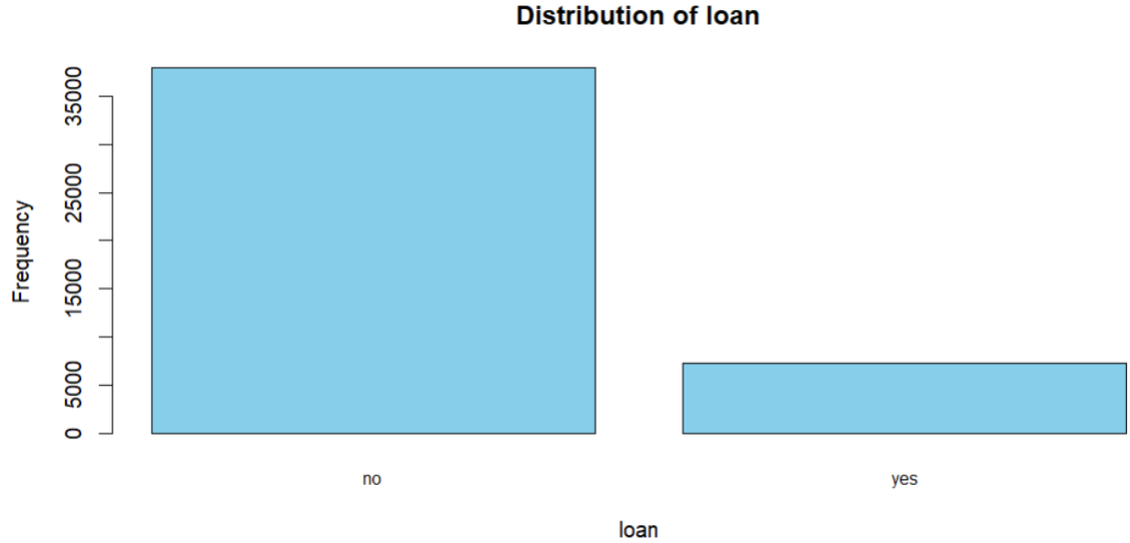

Distribution of Marital
Frequency
0
5000
15000
25000
divorced
married
single
Marital

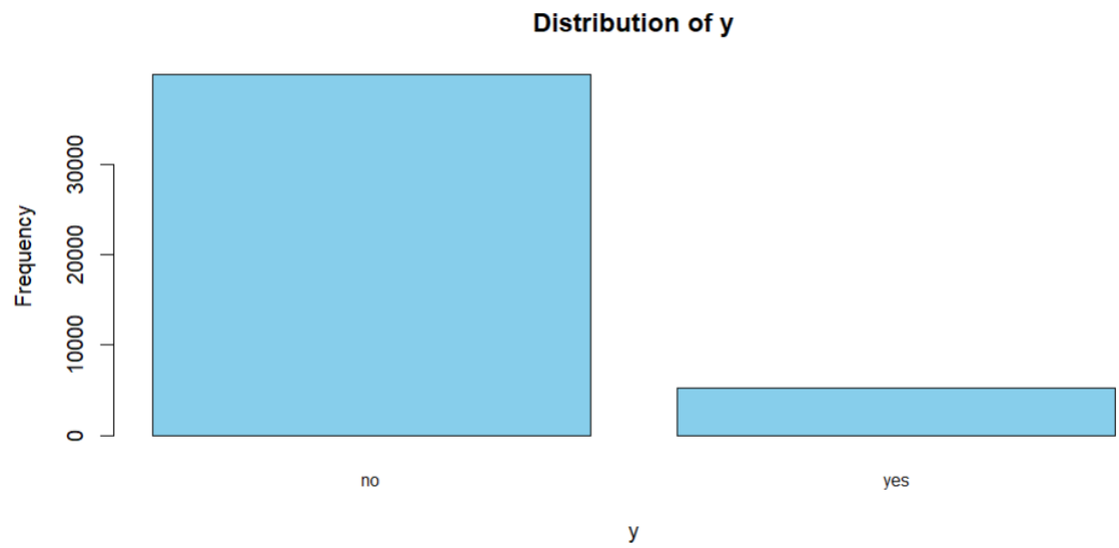

Distribution of Education
Frequency
0
5000
10000
15000
20000
primary
secondary
tertiary
unknown
education


Distribution of contact
Frequency
0
5000
15000
25000
cellular
telephone
unknown
contact


Distribution of default
Frequency
0
10000
20000
30000
40000
no
yes
default


Distribution of housing
Frequency
0
5000
10000
20000
no
yes
housing


Distribution of loan
Frequency
0
5000
15000
25000
35000
no
yes
loan


Distribution of y
Frequency
0
10000
20000
30000
no
yes
y

Based on the graphs, we gain some information. First, the campaigns are highly focused on administators, blue-collars, and technicians in terms of jobs. The clients are mostly married and those married are twice larger than those single. In terms of education, most clients have university education and have no default stay on their cards. Also, most clients have no housing loan and personal loan. In terms of contact, cellular is mostly used. In terms of y, clients that do not subscribe are more than those that do. Therefore, we will use the following independent variables: age, job, marital, education, default, balance, housing, loan, contact, duration, campaign, pdays, previous, and poutcome to predict the dependent variable y. Here we prepare the data for model fitting.

```
selected_columns <- bank_data[, c("age","job","marital","education",
final_data <- data.frame(selected_columns)
head(final_data)

str(final_data)

summary(final_data)
```

```
> head(final_data)
  age          job marital education default balance housing loan
1  58   management married  tertiary      no    2143     yes   no
2  44   technician  single secondary      no      29     yes   no
3  33 entrepreneur married secondary      no       2     yes  yes
4  47  blue-collar married   unknown      no    1506     yes   no
5  33      unknown  single   unknown      no       1      no   no
6  35   management married  tertiary      no     231     yes   no
  contact duration campaign pdays previous poutcome  y
1 unknown      261        1    -1        0  unknown no
2 unknown      151        1    -1        0  unknown no
3 unknown       76        1    -1        0  unknown no
4 unknown       92        1    -1        0  unknown no
5 unknown      198        1    -1        0  unknown no
6 unknown      139        1    -1        0  unknown no

> summary(final_data)
      age            job              marital
 Min.   :18.00   Length:45211       Length:45211
 1st Qu.:33.00   Class :character   Class :character
 Median :39.00   Mode  :character   Mode  :character
 Mean   :40.94
 3rd Qu.:48.00
 Max.   :95.00
  education           default             balance
 Length:45211       Length:45211       Min.   : -8019
 Class :character   Class :character   1st Qu.:    72
 Mode  :character   Mode  :character   Median :   448
                                       Mean   :  1362
                                       3rd Qu.:  1428
                                       Max.   :102127

  housing             loan             contact
Length:45211       Length:45211       Length:45211
Class :character   Class :character   Class :character
Mode  :character   Mode  :character   Mode  :character


   duration        campaign          pdays          previous
Min.   :   0.0   Min.   : 1.000   Min.   : -1.0   Min.   :  0.0000
1st Qu.: 103.0   1st Qu.: 1.000   1st Qu.: -1.0   1st Qu.:  0.0000
Median : 180.0   Median : 2.000   Median : -1.0   Median :  0.0000
Mean   : 258.2   Mean   : 2.764   Mean   : 40.2   Mean   :  0.5803
3rd Qu.: 319.0   3rd Qu.: 3.000   3rd Qu.: -1.0   3rd Qu.:  0.0000
Max.   :4918.0   Max.   :63.000   Max.   :871.0   Max.   :275.0000

   poutcome             y
Length:45211       Length:45211
Class :character   Class :character
Mode  :character   Mode  :character
```

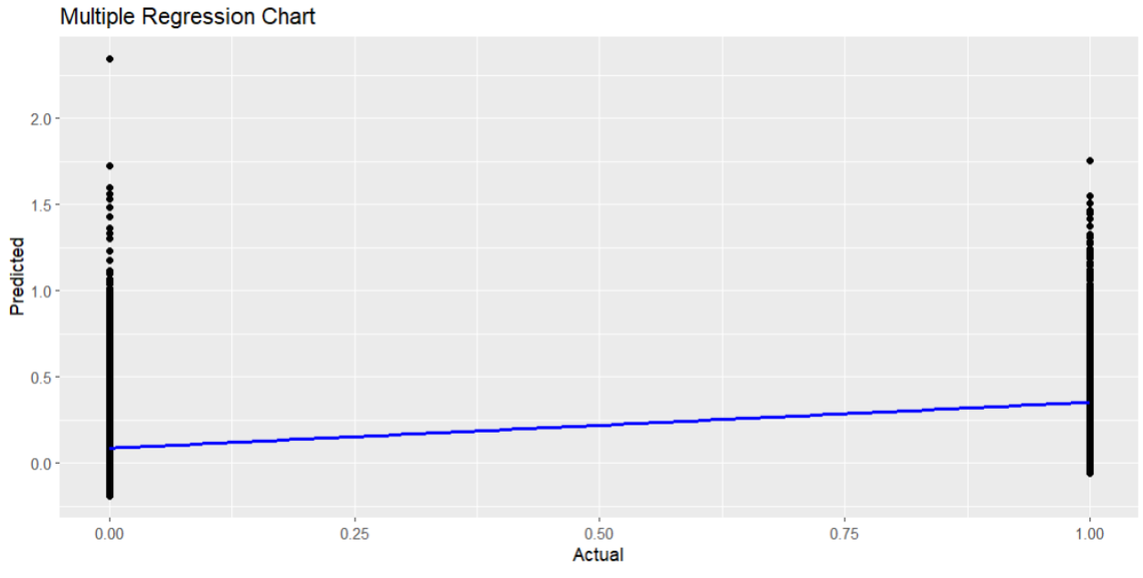
Here, we convert categorical values into factors. We also convert binary yes and no outputs to 1 or 0. We also check for missing data and the results show none.

```
final_data$job <- as.factor(final_data$job)
final_data$marital <- as.factor(final_data$marital)
final_data$education <- as.factor(final_data$education)
final_data$contact <- as.factor(final_data$contact)
final_data$poutcome <- as.factor(final_data$poutcome)
final_data$y <- as.numeric(final_data$y == "yes")
final_data$default <- as.numeric(final_data$default == "yes")
final_data$housing <- as.numeric(final_data$housing == "yes")
final_data$loan <- as.numeric(final_data$loan == "yes")
head(final_data)
```

Now we create a model and fit the data. We are predicting the dependent variable y. Here we will use a regression model. Additionally, since there are multiple independent variables, we will use a multiple regression model.

```
model <- lm(y ~ ., data = final_data)

summary(model)

final_data$predicted_prob <- predict(model, type = "response")

ggplot(final_data, aes(x = y, y = predicted)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE, color = "blue") +
  labs(title = "Multiple Regression Chart",
       x = "Actual",
       y = "Predicted")
```

```
Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)         5.185e-02  1.303e-02   3.981 6.88e-05 ***
age                 2.567e-04  1.588e-04   1.616 0.106144
jobblue-collar     -2.433e-02  4.962e-03  -4.903 9.48e-07 ***
jobentrepreneur    -3.356e-02  8.239e-03  -4.073 4.64e-05 ***
jobhousemaid       -3.689e-02  8.947e-03  -4.124 3.74e-05 ***
jobmanagement      -1.637e-02  5.503e-03  -2.974 0.002940 **
jobretired          4.194e-02  7.701e-03   5.447 5.16e-08 ***
jobself-employed   -2.885e-02  8.068e-03  -3.576 0.000350 ***
jobservices        -1.993e-02  5.723e-03  -3.482 0.000497 ***
jobstudent          8.887e-02  1.013e-02   8.776  < 2e-16 ***
jobtechnician      -2.010e-02  4.981e-03  -4.036 5.45e-05 ***
jobunemployed      -8.569e-03  8.562e-03  -1.001 0.316919
jobunknown         -2.642e-02  1.693e-02  -1.561 0.118598
maritalmarried     -1.099e-02  4.188e-03  -2.625 0.008681 **
maritalsingle       1.424e-02  4.853e-03   2.934 0.003347 **
educationsecondary  6.504e-03  4.160e-03   1.563 0.117953
educationtertiary   2.732e-02  5.204e-03   5.251 1.52e-07 ***
educationunknown    1.441e-02  7.412e-03   1.945 0.051796 .
default            -9.440e-03  9.751e-03  -0.968 0.333015
balance             1.843e-06  4.306e-07   4.281 1.87e-05 ***
housing            -5.613e-02  2.845e-03 -19.724  < 2e-16 ***
loan               -3.268e-02  3.568e-03  -9.158  < 2e-16 ***
contacttelephone   -4.988e-03  5.455e-03  -0.914 0.360516
contactunknown     -5.456e-02  3.144e-03 -17.351  < 2e-16 ***
duration            4.732e-04  5.036e-06  93.963  < 2e-16 ***
campaign           -2.102e-03  4.218e-04  -4.983 6.30e-07 ***
pdays              -2.140e-05  2.726e-05  -0.785 0.432473
previous            1.225e-03  6.649e-04   1.842 0.065523 .
poutcomeother       2.837e-02  7.544e-03   3.761 0.000170 ***
poutcomesuccess     4.445e-01  8.402e-03  52.904  < 2e-16 ***
poutcomeunknown    -2.861e-02  8.145e-03  -3.512 0.000444 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.274 on 45180 degrees of freedom
Multiple R-squared:  0.2738,    Adjusted R-squared:  0.2734
F-statistic: 567.9 on 30 and 45180 DF,  p-value: < 2.2e-16
```



The results show that the residual standard error is 27.4%. Additionally, the multiple r-squared score is 27.38% and the adjusted r-squared is 27.34%. Based on the results, the model struggles to properly predict the dependent variable y with the independent variables given a multiple r-squared score of 27.38%. This means that 27.38% of the variability of the dependent variable y is explained by the model while the remaining precent is unaccounted. We can also see the performance of the model through the regression graph. There are instances and areas in the graph that show that the model can correctly predict the dependent y variable values but it is still outweighed by incorrect predictions. The independent variables contribute to the model's capability in performing the prediction task, but the model performance does not show much efficacy.

REFERENCES

Chee Hua Chew. (2021). *Artificial intelligence, analytics and data science. Volume 1, Core concepts and models*. Cengage Learning Asia Pte Ltd.

freeCodeCamp.org. (2019). R Programming Tutorial - Learn the Basics of Statistical Computing [YouTube Video]. In *YouTube*. https://www.youtube.com/watch?v=_V8eKsto3Ug

Machlis, S. (2015, February 18). *Learn R for beginners with our PDF*. Computerworld. https://www.computerworld.com/article/2884322/learn-r-programming-basics-with-our-pdf.html

Moro,S., Rita,P., and Cortez,P.. (2012). Bank Marketing. UCI Machine Learning Repository. https://doi.org/10.24432/C5K306.

Wickham, H., & Grolemund, G. (2017). *R for data science : import, tidy, transform, visualize, and model data*. O'reilly.