Data 622 Machine Learning and Big Data_HW1

Jiaxin Zheng

2025-09-28

PartI

1.1 Assignment 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.

1.2 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

1.3 Bank Direct Marketing Data Set Description

There is a total number of 45,211 client recors in this data set, The data consists of 17 varibales, Input variables:

1 - age (numeric)

2 - job : type of job (categorical: "admin.","unknown","unemployed","management","housemaid","entrepreneur","student", "blue-collar","self-employed","retired","technician","services")

3 - marital : marital status (categorical: "married","divorced","single"; note: "divorced" means divorced or widowed)

4 - education (categorical: "unknown","secondary","primary","tertiary")

5 - default: has credit in default? (binary: "yes","no")

6 - balance: average yearly balance, in euros (numeric)

7 - housing: has housing loan? (binary: "yes","no")

8 - loan: has personal loan? (binary: "yes","no")

related with the last contact of the current campaign:

9 - contact: contact communication type (categorical: “unknown”,“telephone”,“cellular”) 10 - day: last contact day of the month (numeric)

11 - month: last contact month of year (categorical: “jan”, “feb”, “mar”, . . . , “nov”, “dec”)

12 - duration: last contact duration, in seconds (numeric) # Other attributes:

13 - campaign: number of contacts performed during this campaign and for this client (numeric, includes last contact)

14 - pdays: number of days that passed by after the client was last contacted from a previous campaign (numeric, -1 means client was not previously contacted)

15 - previous: number of contacts performed before this campaign and for this client (numeric)

16 - poutcome: outcome of the previous marketing campaign (categorical: “unknown”,“other”,“failure”,“success”)

Output variable (desired target): 17 - y - has the client subscribed a term deposit? (binary: “yes”,“no”)

Part II: EDA

There are 17 variables includes “y” and 45,211 observations in the dataset. “y” is the target variable which represent whether a client subscribes to a term deposit or not.

```
# Load Libraries
library(tidyverse)
library(corrplot)
library(dplyr)
library(ggplot2)
library(knitr)
library(skimr)
library(readr)
library(forcats)
library(scales)
```

```
# Read data file
df <- read.csv("https://raw.githubusercontent.com/Jennyjjxxzz/HW1/refs/heads/main/bank-full.csv", sep =

head (df)
```

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

```r
summary(df)
```

```
##       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
##
##
##
```

```r
str(df)
```

```
## '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" ...
```

2.1 Missing Data

There appear to be no missing values in this data set, and there is no need to use any methods to impute the missing values.

```
colSums(is.na(df))
```

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

```
# Organize the dataset as Numeric and Categorical

cat_cols  <- c("job","marital","education","default","housing","loan",
               "contact","month","poutcome","y")
num_cols  <- c("age","balance","day","duration","campaign","pdays","previous")

df <- df %>%
  mutate(across(all_of(cat_cols), as.factor),
         across(all_of(num_cols), as.numeric))
```

2.2 Numeric Feature Distribution

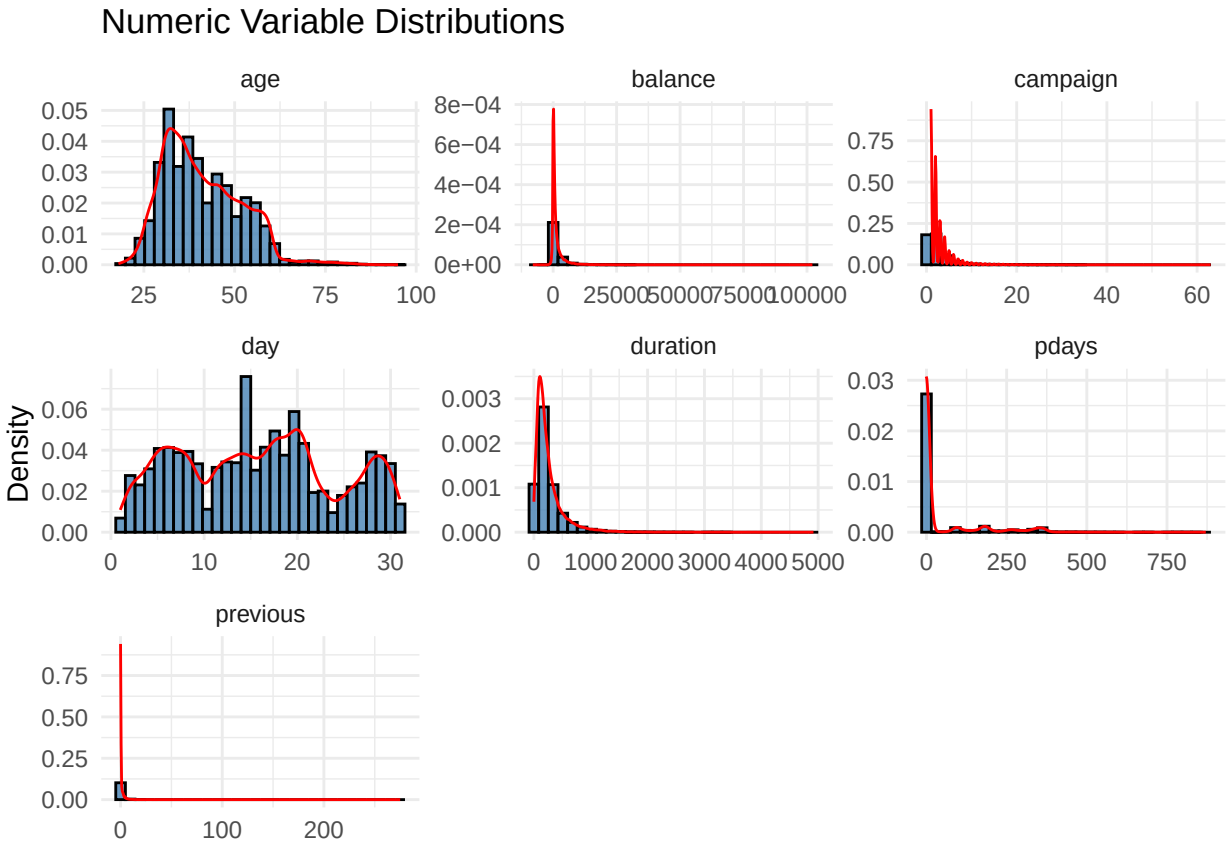
All the numeric variables are right skewed, expect"day". "Day" appears to approximate distribution.

```
# Distribution graphic for Numeric data
numeric_plot <- df %>%
  pivot_longer(all_of(num_cols), names_to = "variable", values_to = "value") %>%
  ggplot(aes(value)) +
  geom_histogram(aes(y = ..density..), alpha = 0.8, bins = 30, fill = "steelblue", color="black") +
  geom_density(color = "red", size = 0.5) +
  facet_wrap(~ variable, scales = "free") +
  labs(title = "Numeric Variable Distributions", x = NULL, y = "Density")+
  theme_minimal()
```

```
## Warning: Using ‘size‘ aesthetic for lines was deprecated in ggplot2 3.4.0.
## i Please use ‘linewidth‘ instead.
## This warning is displayed once every 8 hours.
## Call ‘lifecycle::last_lifecycle_warnings()‘ to see where this warning was
## generated.
```

```
numeric_plot
```

```
## 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.
```

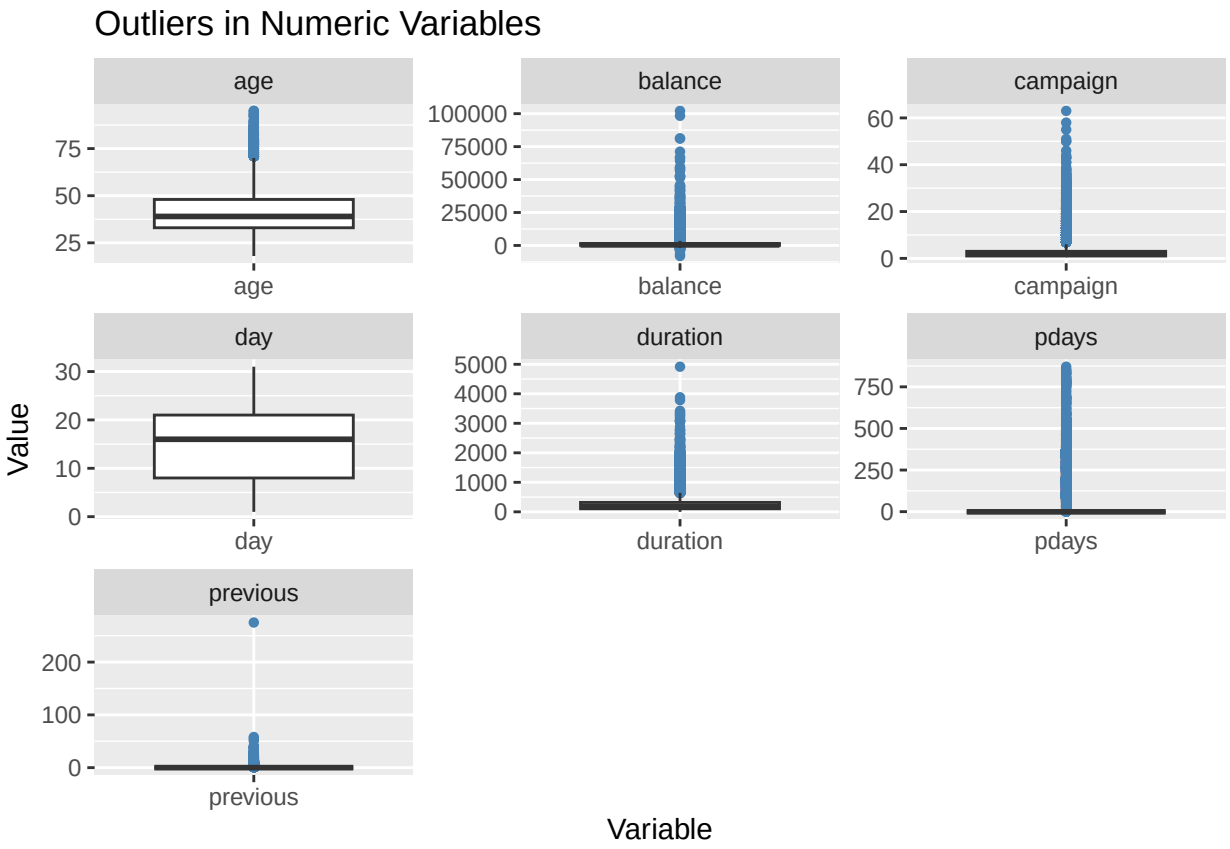


2.3 Are there any outliers

Outlier analysis via boxplot below: Almost all the feature appears to have outliers, with exception of “day”.

Addition: I Use IQR function to calculate how many outlier in each feature via 1.5 * IQR age 487, balance 4729, campaign 3064, duration 3235, pdays 8257, previous 8257.

```
# Boxplots for analysis the outlier in numeric data
numeric_boxplot <- df %>%
  pivot_longer(cols = all_of(num_cols), names_to = "variable", values_to = "value") %>%
  ggplot(aes(x = variable, y = value)) +
  geom_boxplot(outlier.colour = "steelblue", outlier.shape = 16) +
  facet_wrap(~ variable, scales = "free") +
  labs(title = "Outliers in Numeric Variables", x = "Variable", y = "Value")


numeric_boxplot
```




```
# Use IQR function to calculate the outlyiers
count_outliers <- function(x) {
  q1 <- quantile(x, 0.25, na.rm = TRUE)
  q3 <- quantile(x, 0.75, na.rm = TRUE)
  iqr <- q3 - q1 # Interquartile range

  lower <- q1 - 1.5 * iqr
  upper <- q3 + 1.5 * iqr
  sum(x < lower | x > upper, na.rm = TRUE)
}

# Apply to all numeric columns
outlier_counts <- sapply(df[num_cols], count_outliers)
outlier_counts
```

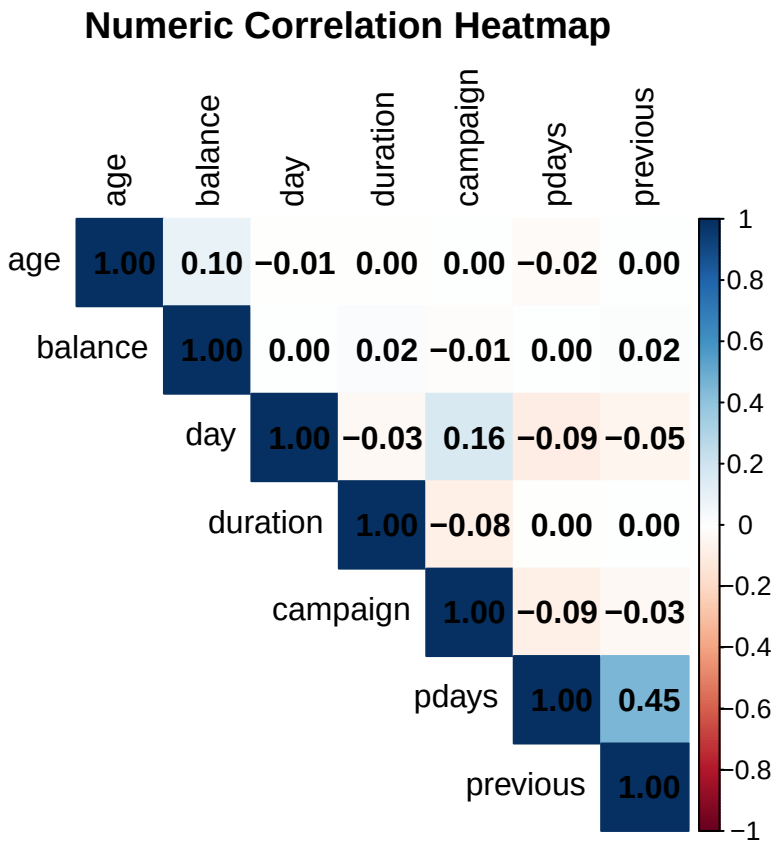
```
##       age   balance       day  duration  campaign     pdays  previous
##       487      4729         0      3235      3064      8257      8257
```

2.4 Correlation of Numerical Variables

Below is the correlation heatmap for numerical variables. Overall, numeric variable doesn't not have strong relation in this dataset.

MILD RELATIONSHIPS: - Pdays and Previous – Correlation is 0.46 This might means there is relationship between before and after campaign. Who were contacted before campaign are more likely to be contacted again.

```r
# Correlation among numeric features

numeric_corr <- df %>% select(all_of(num_cols))
corr_m   <- cor(numeric_corr, use = "pairwise.complete.obs")
corrplot::corrplot(corr_m, method = "color", type = "upper",
                   tl.col = "black", addCoef.col = NA,
                   title = "Numeric Correlation Heatmap", mar = c(0,0,2,0))
```

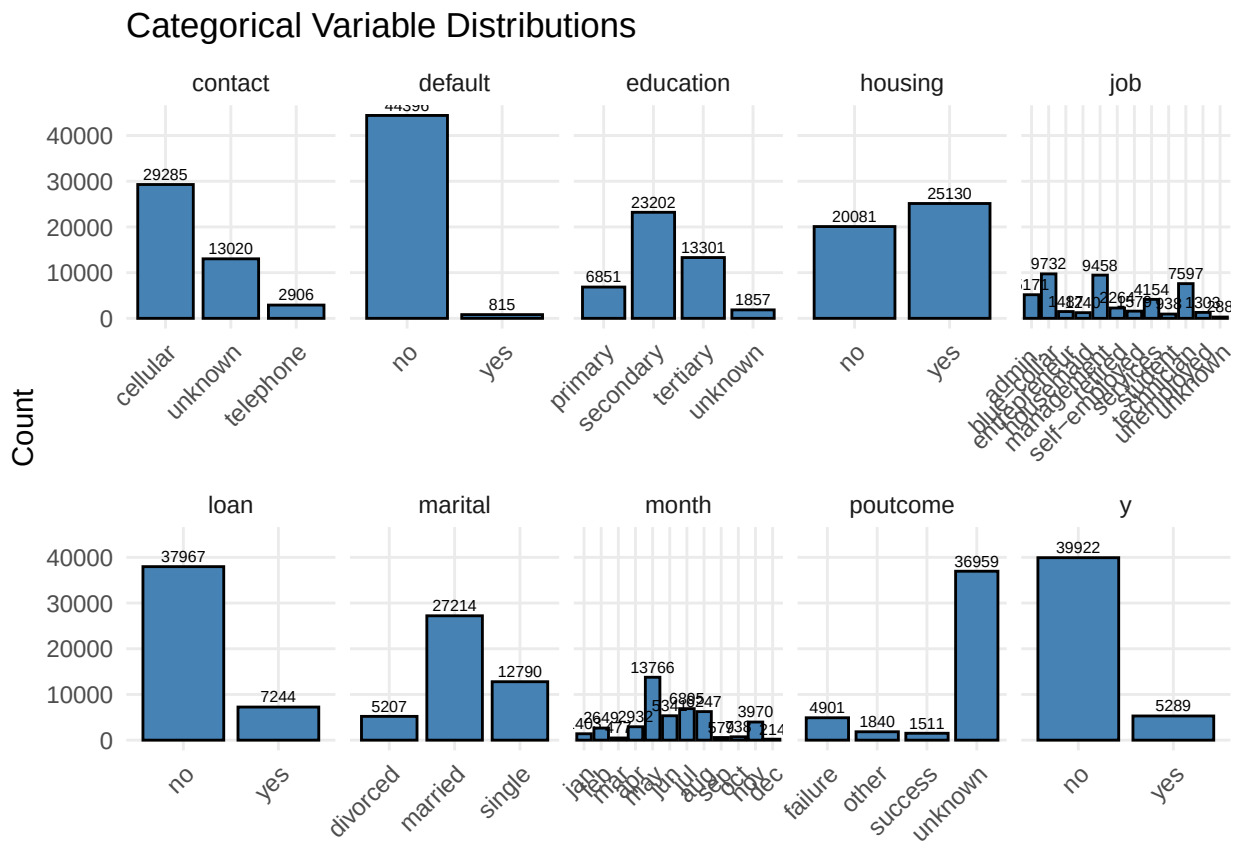


2.5 Categirical Feature Distribution

```r
# choose your categorical columns
cat_cols <- c("job","marital","education","default","housing","loan",
              "contact","poutcome","month","y")

# Make "month" in calendar order
if ("month" %in% names(df)) {
  df <- df %>% mutate(month = factor(month,
    levels = c("jan","feb","mar","apr","may","jun","jul","aug","sep","oct","nov","dec")))
}

# build summary table
cat_long <- df %>%
  pivot_longer(all_of(cat_cols), names_to = "variable", values_to = "level") %>%
```

```
  group_by(variable, level) %>%
  summarise(n = n(), .groups = "drop_last") %>%
  mutate(pct = n / sum(n)) %>%
  group_by(variable) %>%
  mutate(level = fct_reorder(level, pct, .desc = TRUE)) %>%
  ungroup()
```

```
cat_plot <-
  ggplot(cat_long, aes(x = level, y = n)) +
  geom_col(fill = "steelblue", color = "black", width = 0.85) +
  geom_text(aes(label = n), vjust = -0.4, size = 2) +
  facet_wrap(~ variable, scales = "free_x") +
  labs(title = "Categorical Variable Distributions", x = NULL, y = "Count") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1),
        panel.grid.minor = element_blank()) +
    facet_wrap(~variable, scales = "free_x", ncol = 5)

cat_plot
```

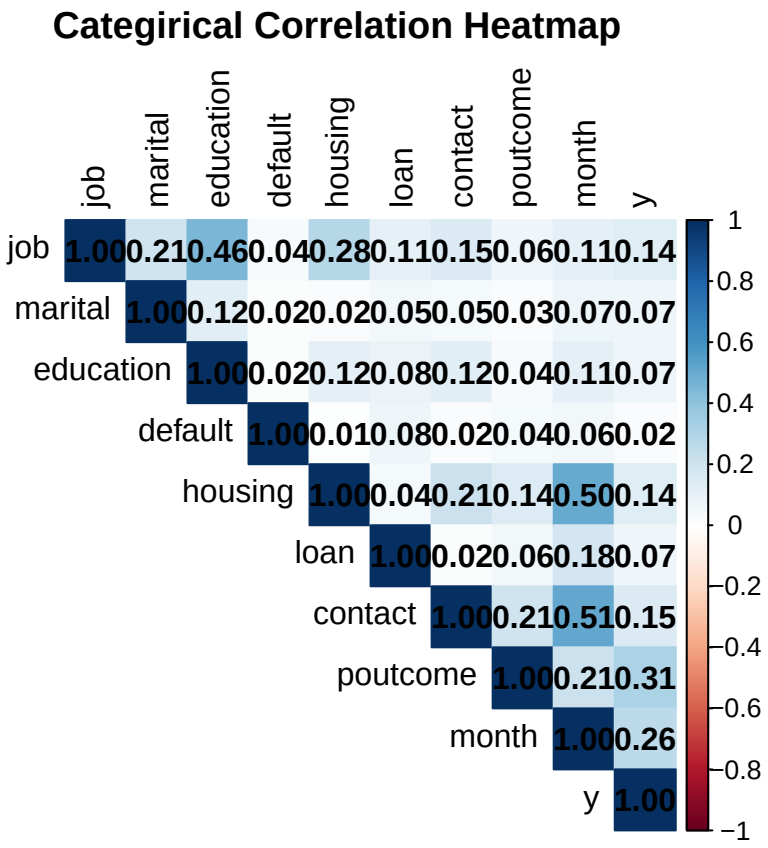


2.6 Categirical Correlation Heatmap

Below is the correlation heatmap for categorical variables. Overall, categorical variable also doesn't not have strong relation in this dataset.

MILD RELATIONSHIPS: - Contact and Month – Correlation of 0.51 - Housing and Month – Correlation of 0.50 - Job and Education – Correlation of 0.46

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

cramers_v <- function(x, y) {
  tbl <- table(x, y)
  if (min(dim(tbl)) < 2) return(NA_real_)
  chi2 <- suppressWarnings(chisq.test(tbl, correct = FALSE)$statistic)
  n <- sum(tbl); r <- nrow(tbl); k <- ncol(tbl)
  phi2 <- chi2 / n
  V <- sqrt(phi2 / min(k - 1, r - 1))
  as.numeric(V)
}

# Build symmetric matrix
Vmat <- matrix(NA_real_, nrow = length(cat_cols), ncol = length(cat_cols),
               dimnames = list(cat_cols, cat_cols))

for (i in seq_along(cat_cols)) {
  for (j in seq_along(cat_cols)) {
    Vmat[i, j] <- if (i == j) 1 else cramers_v(df[[cat_cols[i]]], df[[cat_cols[j]]])
  }
}

# Categirical Correlation Heatmap
corrplot(Vmat,
         method = "color", type = "upper",
         tl.col = "black", addCoef.col = NA,
         title = "Categirical Correlation Heatmap",
         mar = c(0,0,2,0))
```

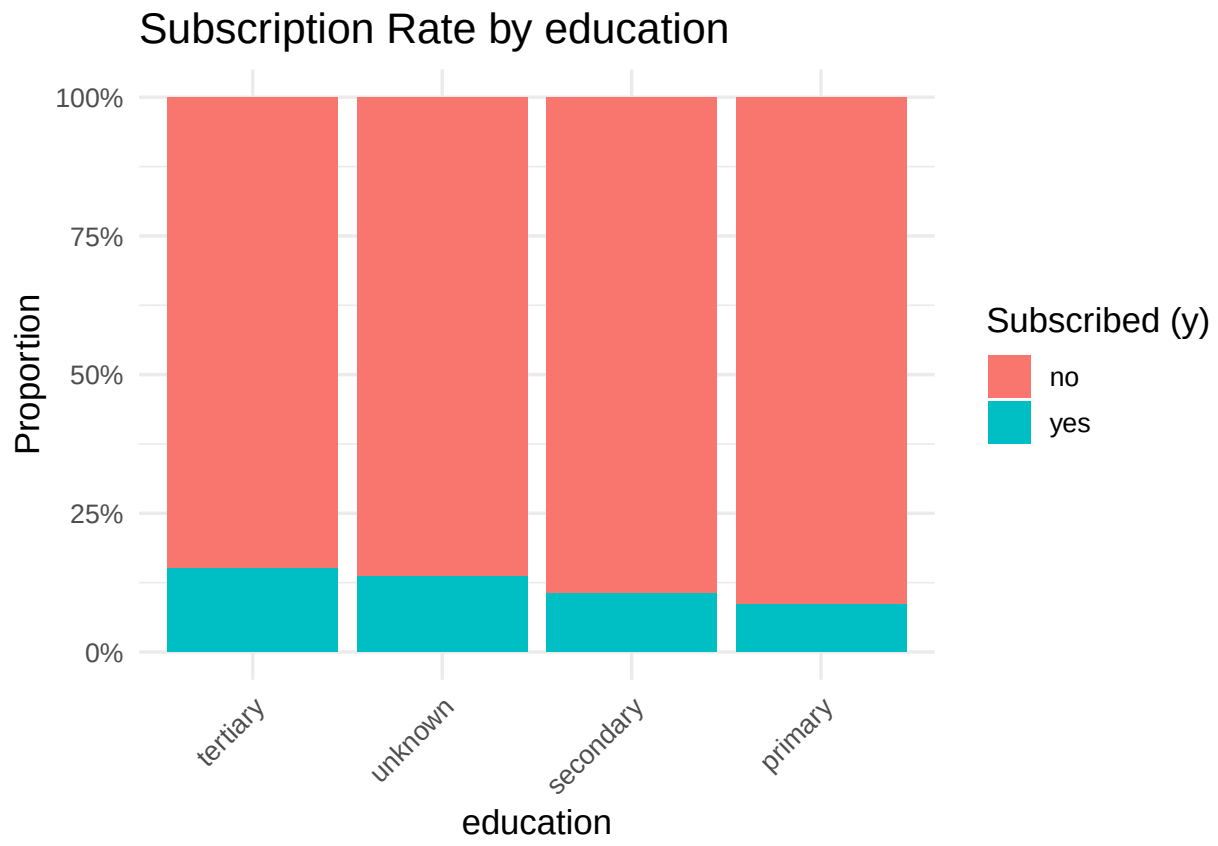


2.7 Subscrption Vs. Variables

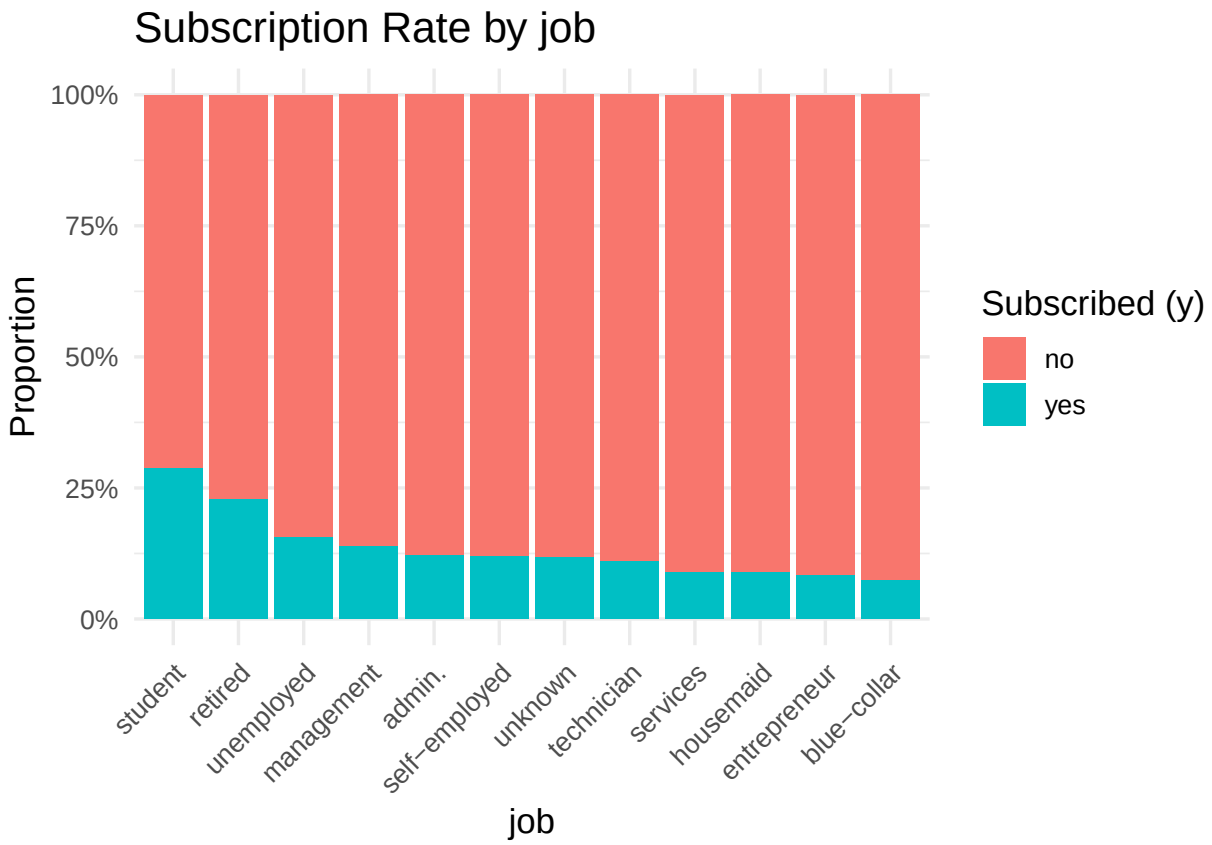
1. Subscription vs. Education: Clients with teritary education level have highest proportion.
2. Subscription vs. Job: Students has highest percentage did subscribe to the term deposit.
3. Subscription vs. Marital: Clients who are single has highest percentage did subscribe to the term deposit, and fellow by clients who are divorced.
4. Subscription vs. Housing: Higher percentage of clients with housing loans did not subscribe.
5. Subscription vs. Age: Age group 18-25 and age group 65+ have higher subscription percentage.

```
plot_y_by_cat <- function(catvar) {
  stopifnot(catvar %in% names(df))
  agg <- df %>%
    count(!!sym(catvar), y, name = "n") %>%
    group_by(!!sym(catvar)) %>%
    mutate(pct = n / sum(n)) %>%
    ungroup()

  # order levels by success rate
  order_tbl <- agg %>%
    filter(y == "yes") %>%
    arrange(desc(pct)) %>%
    pull(!!sym(catvar)) %>%
    as.character()

  ggplot(
```

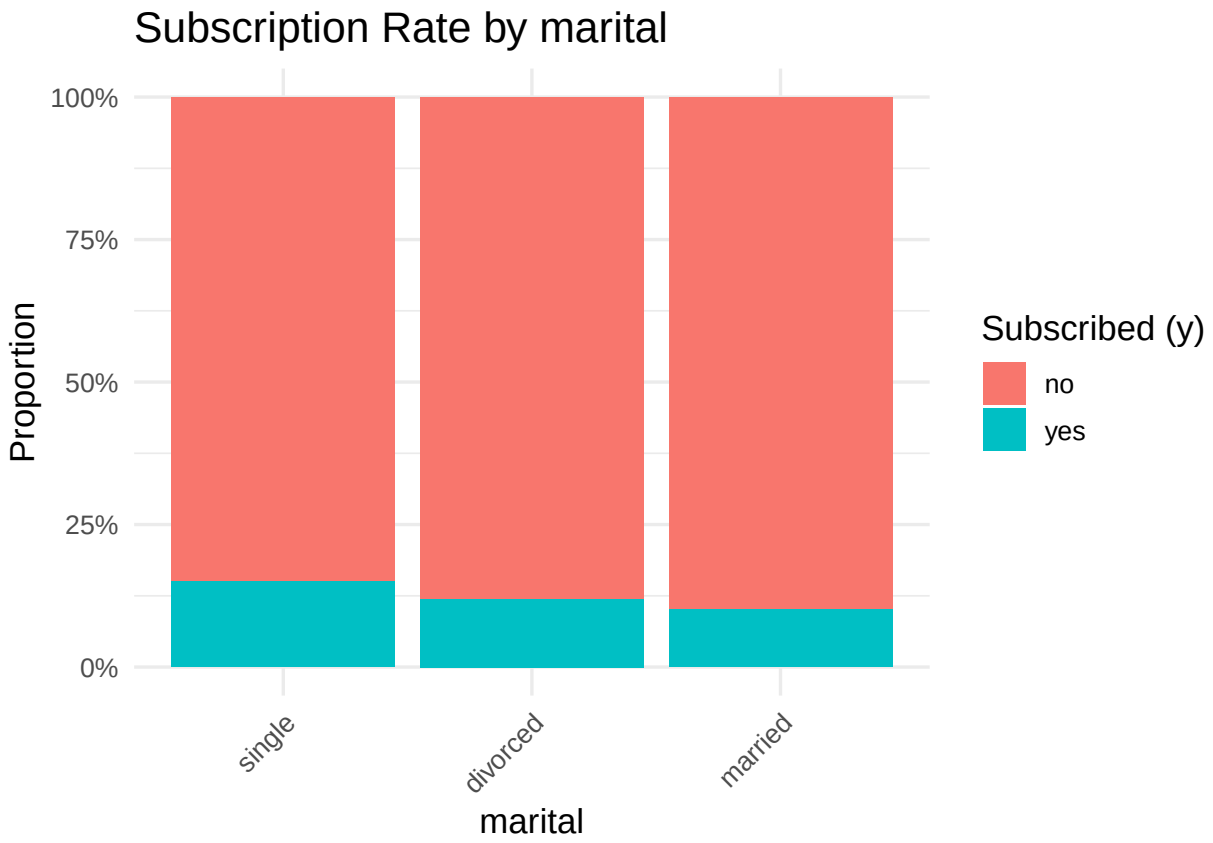
```
    agg %>% mutate(!!catvar := factor(!!sym(catvar), levels = order_tbl)),
    aes(x = !!sym(catvar), y = pct, fill = y)
  ) +
    geom_col(position = "fill") +
    scale_y_continuous(labels = percent) +
    labs(
      title = paste("Subscription Rate by", catvar),
      x = catvar, y = "Proportion", fill = "Subscribed (y)"
    ) +
    theme_minimal(base_size = 13) +
    theme(axis.text.x = element_text(angle = 45, hjust = 1))
}

# Plot all
plot_y_by_cat("education")
```




```
plot_y_by_cat("job")
```




```
plot_y_by_cat("marital")
```




```
plot_y_by_cat("housing")
```


```r
# Subscription Rate by Age Group
data <- df %>%
  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_
  ),
  age_group = factor(age_group, levels = c("18-25","26-35","36-45","46-55","56-65","66-75","76+"))
  )

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

ggplot(age_group_summary, aes(x = age_group, y = prop_subscribed)) +
```

```
  geom_bar(stat = "identity", fill = "lightblue") +
  geom_text(aes(label = scales::percent(prop_subscribed, accuracy = 0.1)),
            vjust = -0.5, size = 3.5) +
  labs(
    title = "Subscription Rate by Age Group",
    x = "Age Group",
    y = "Proportion Subscribed"
  ) +
  theme_minimal(base_size = 13)
```