Data Report

2025-11-03

Introduction

One of the first step when looking at a new data set is to investigate the quality of the data. This pdf presents my own script with function calls to quickly access whether the values in the data set are plausible or if actions are required.

Packages required are:

```r
# For the script
library(ggplot2)
library(tidyverse)

# For the pdf
library(knitr)
```

The user is only required to load the data as a dataframe with the correct data types. The script only handles the numerical, factor and boolean data types. Data types such as strings are not handled and should be investigated seperately. The result in the pdf is automatically produced.

```r
data <- read.csv2("bank.csv", stringsAsFactors=TRUE)
# The original data does not include boolean or character variables.
# These are added to include functionalities in the script.
data$boolean <- c(rep(TRUE, 2740), rep(FALSE, 1370), rep(NA, 411))
data$char <- c("hej")
```

Tables

Table 1: **Unproccessed** variables

variable	type
char	character

Table 2: Summary of **numeric** variables

	age	balance	day	duration	campaign	pdays	previous
Min	19	-3313	1	4	1	-1	0
Max	87	71188	31	3025	50	871	25
Median	39	444	16	185	2	-1	0
Mean	41.17	1422.658	15.915	263.961	2.794	39.767	0.543
Number of NA	0	0	0	0	0	0	0
Percentage NA	0	0	0	0	0	0	0

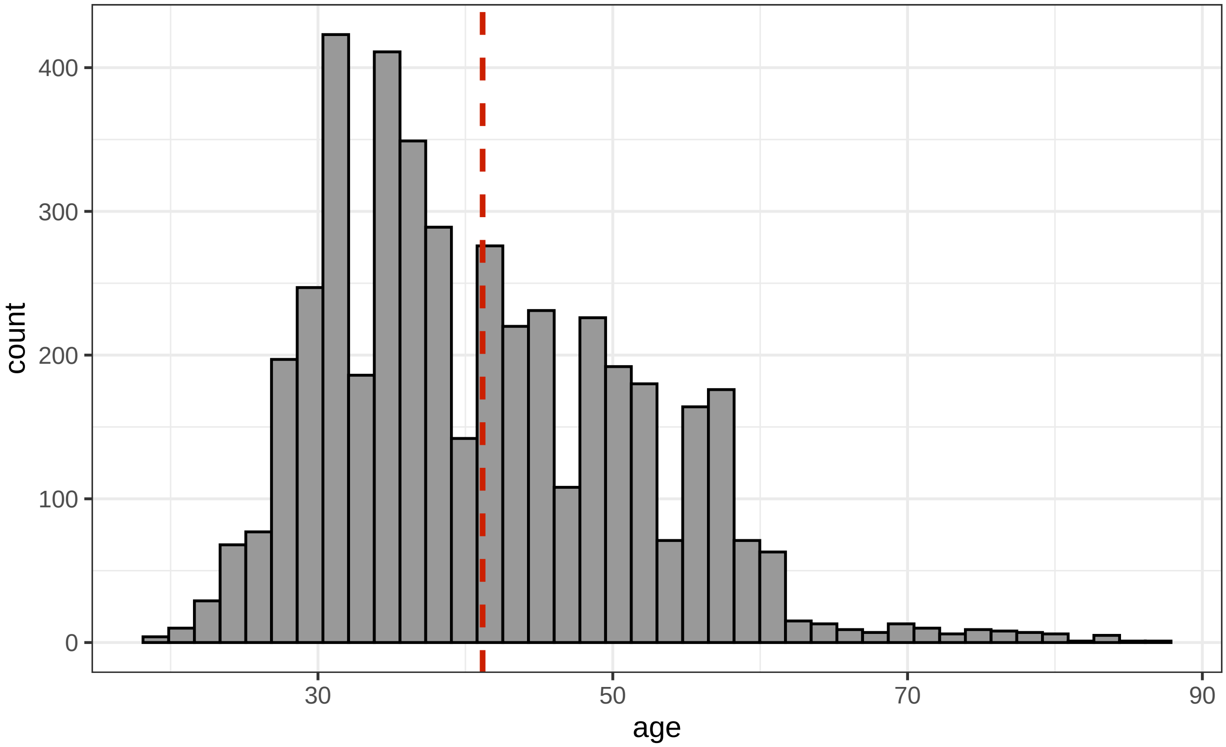
Table 3: Summary of **factor** variables

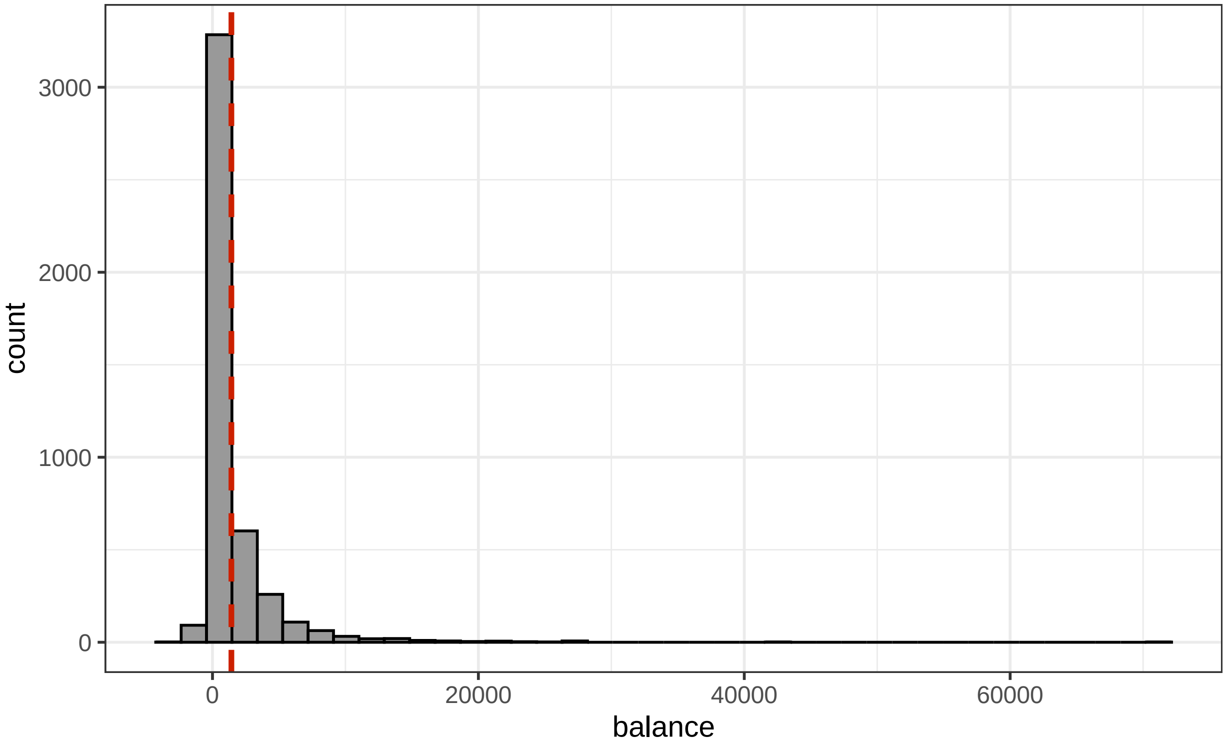
	x
job	12
marital	3
education	4
default	2
housing	2
loan	2
contact	3
month	12
poutcome	4
y	2

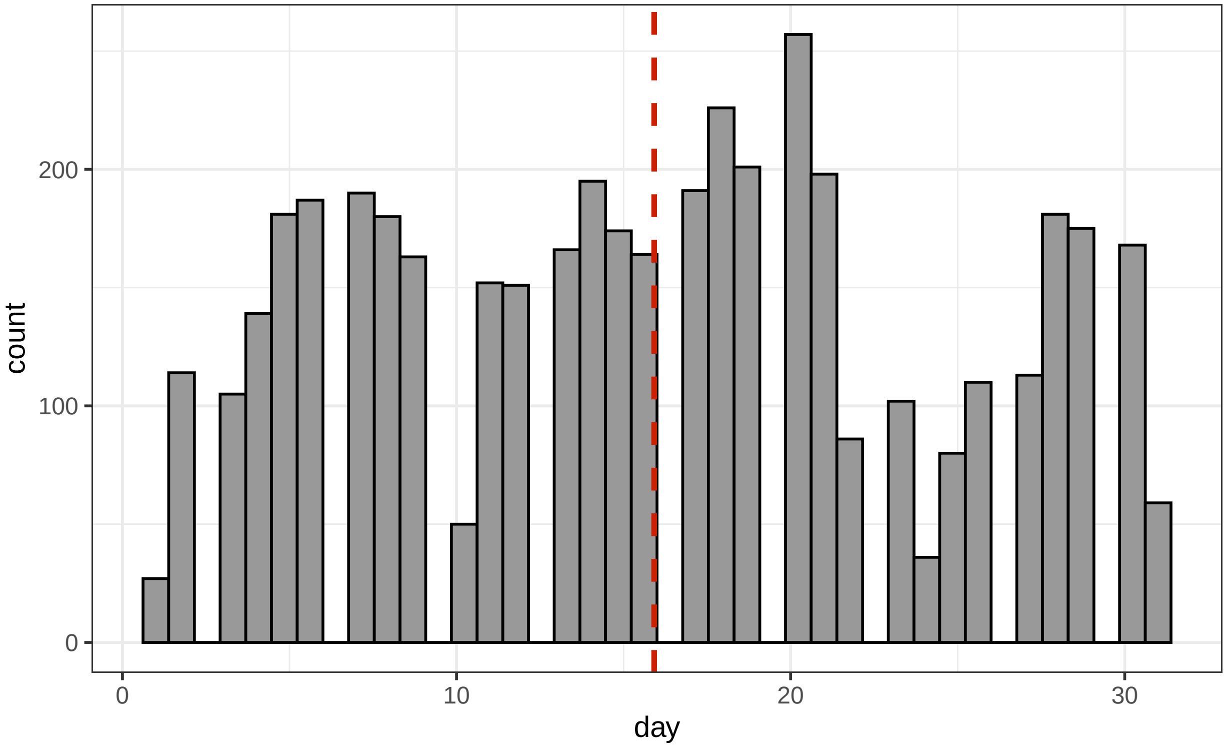
Table 4: Summary of **boolean** variables

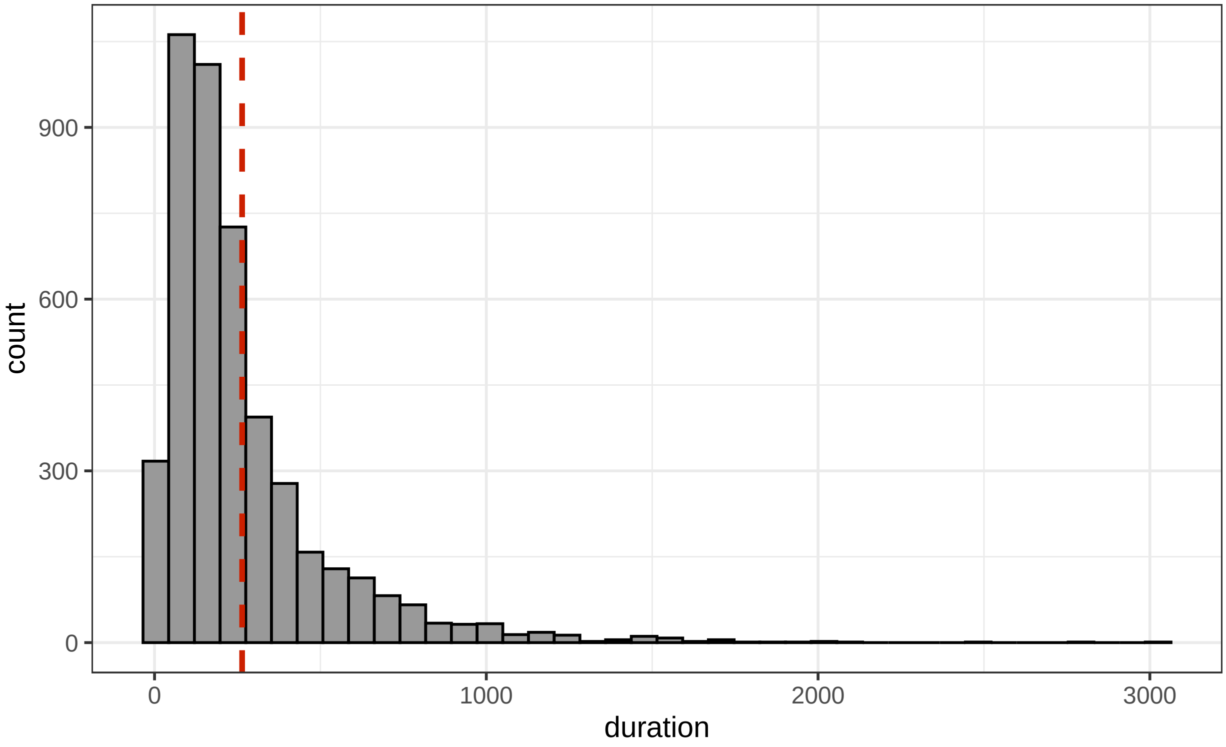
	boolean
True	2740
False	1370
True(%)	0.6666667
NA(#)	411
NA(%)	0.09090909

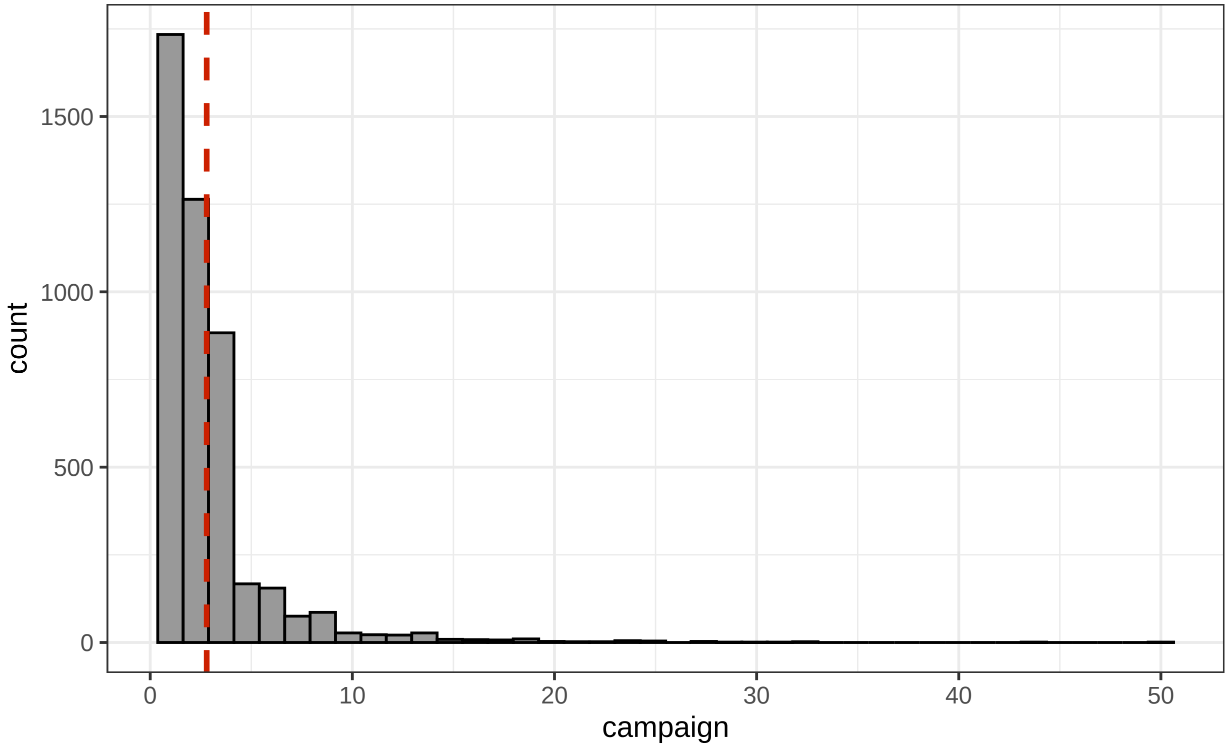
Figures for numerical variables

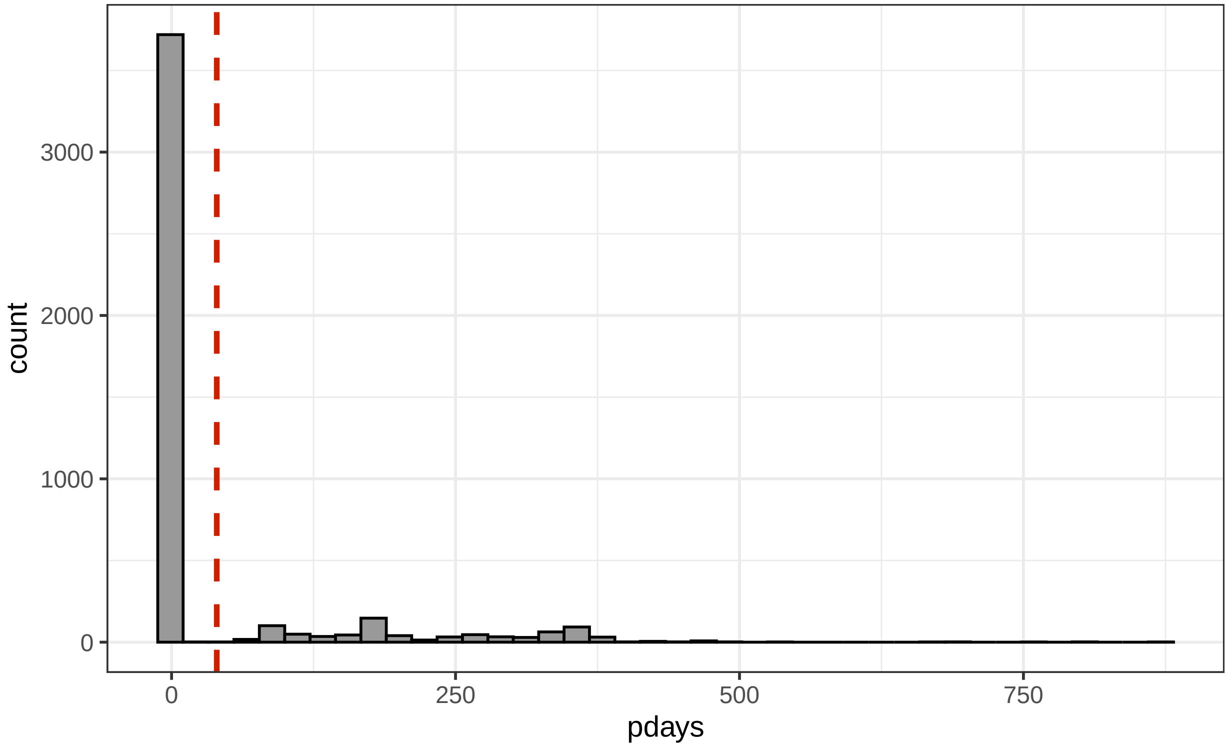


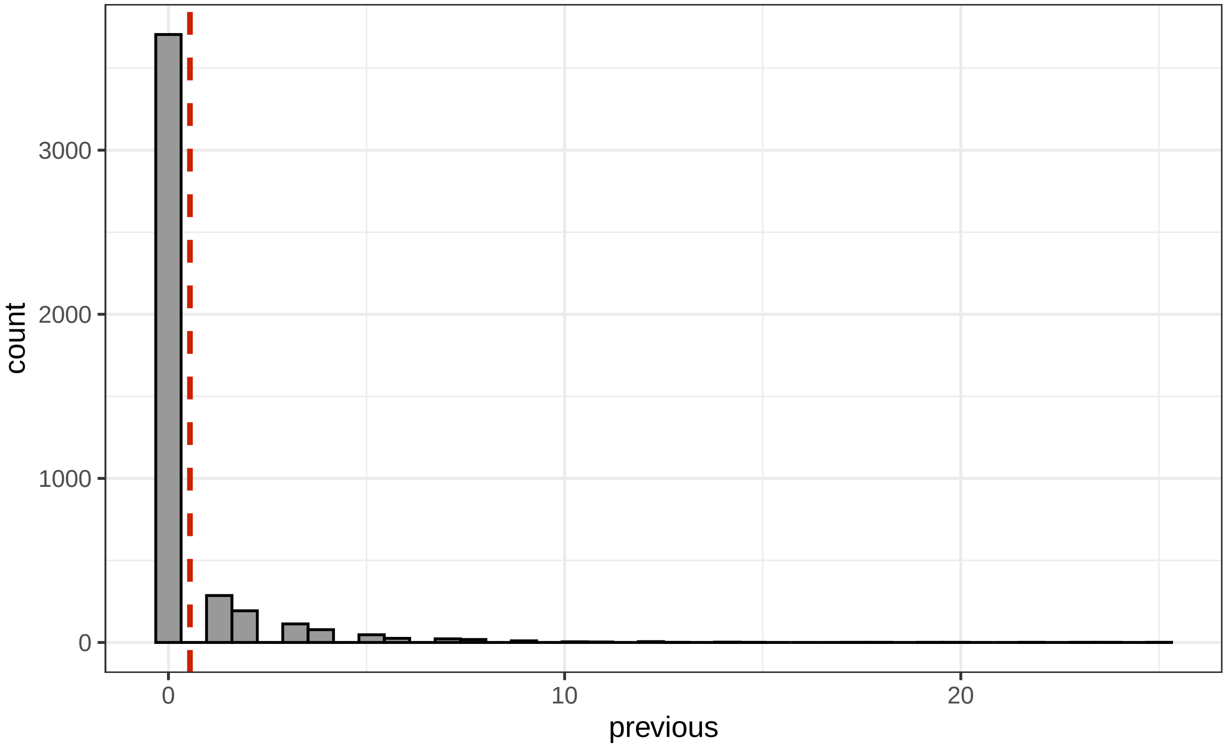



count
200
100
0
0
10
20
30
day


count
900
600
300
0
0
1000
2000
3000
duration

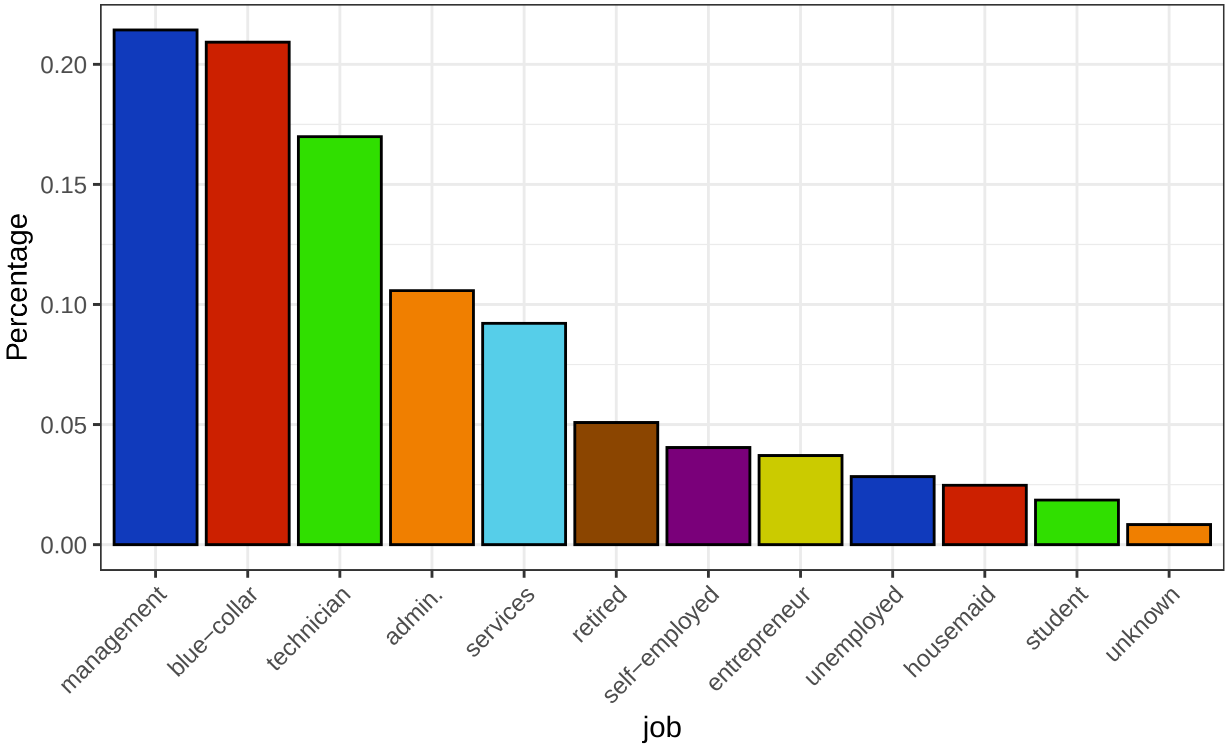

count
1500
1000
500
0
0
10
20
30
40
50
campaign

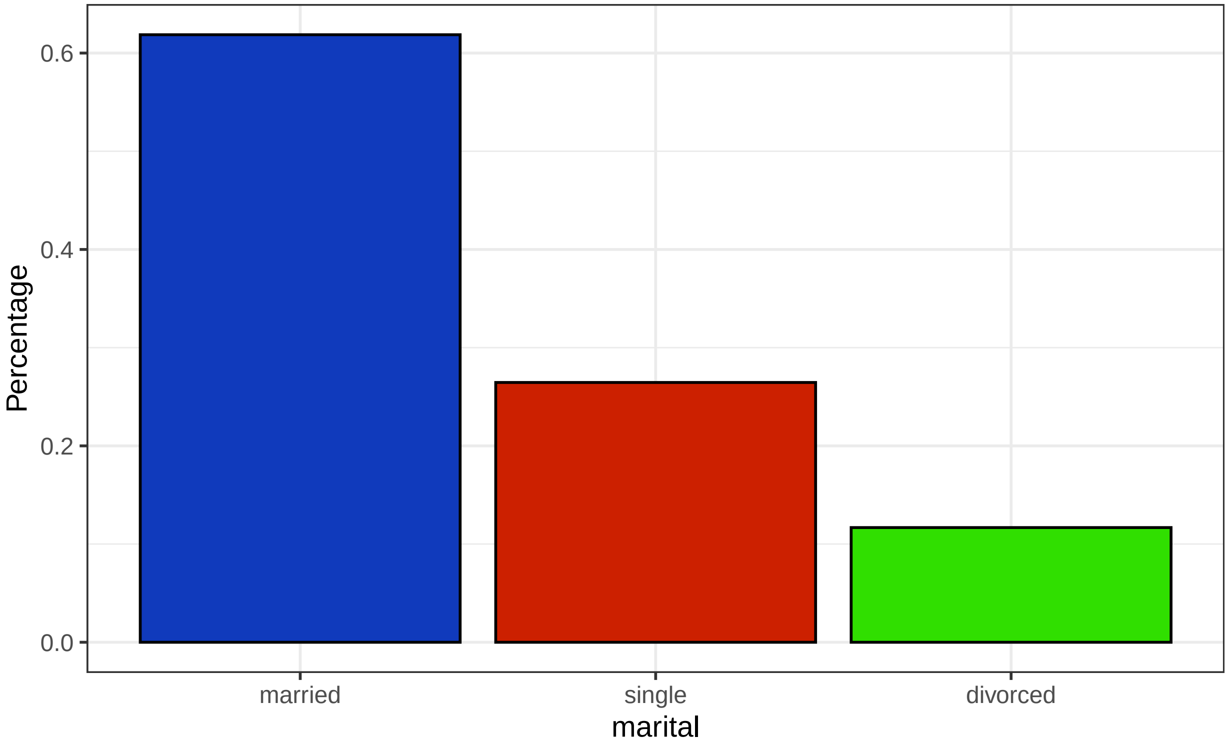

count
3000
2000
1000
0
0
250
500
750
pdays

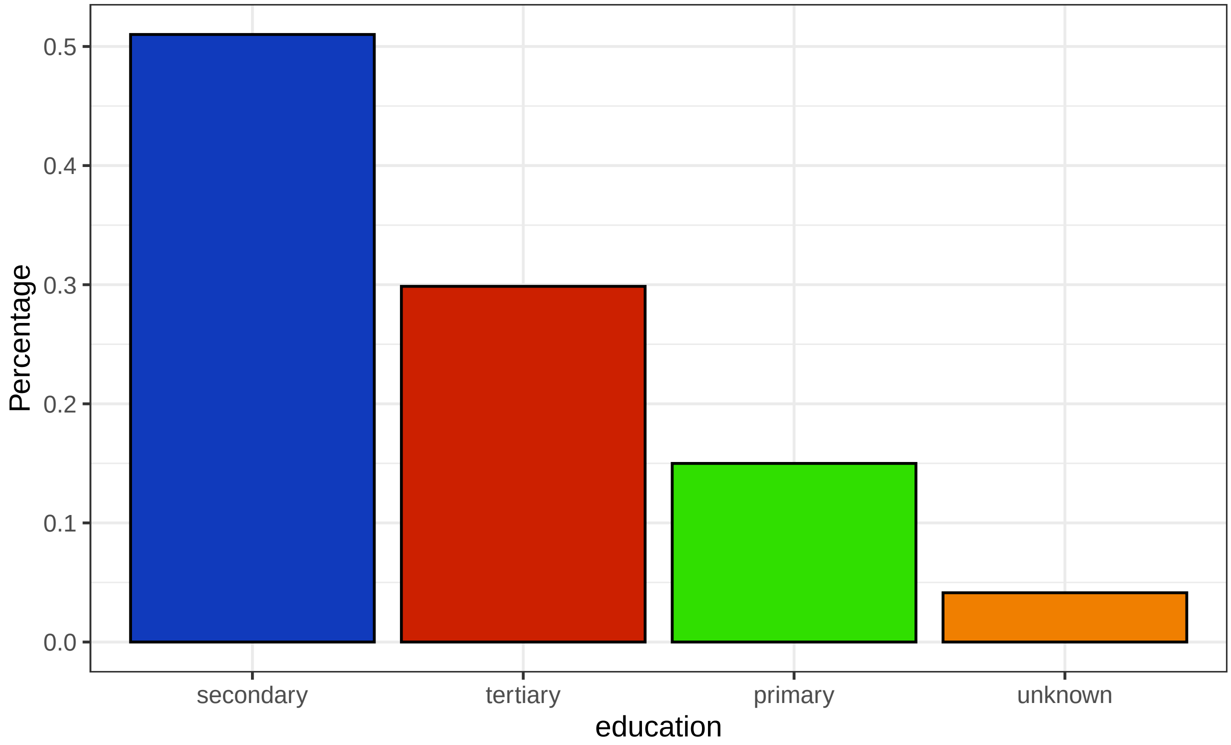


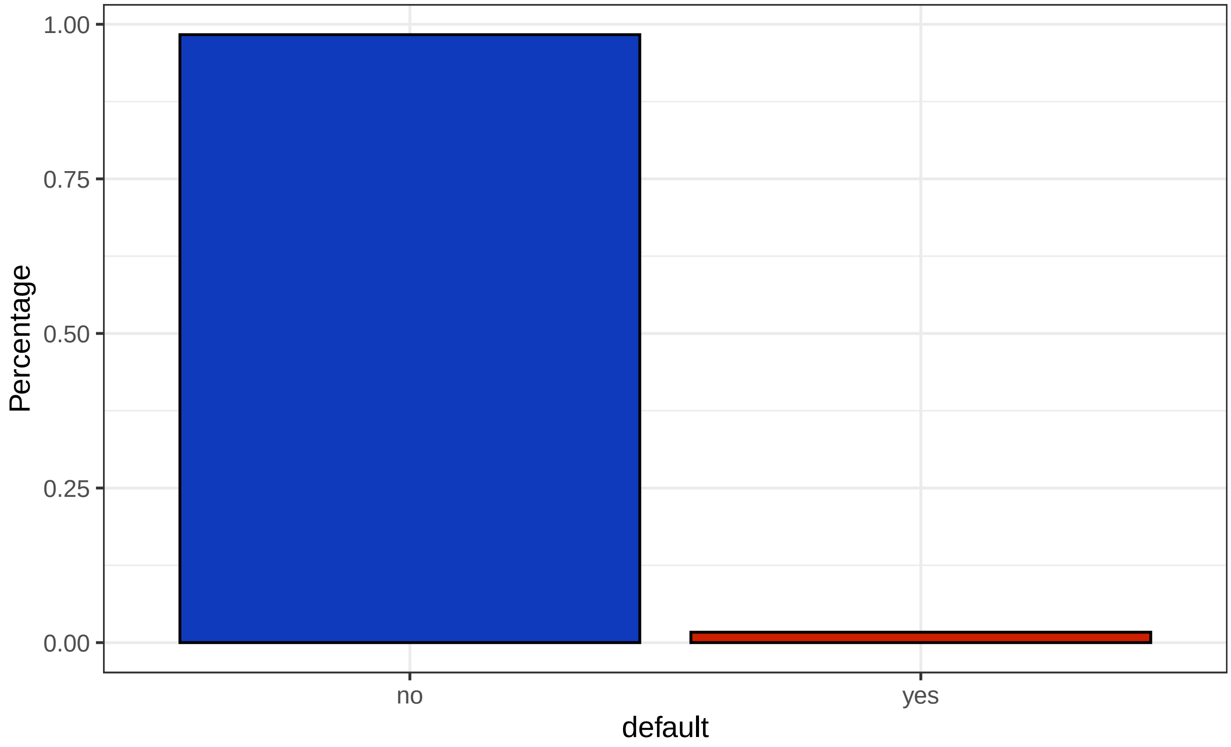
count
3000
2000
1000
0
0
10
20
previous

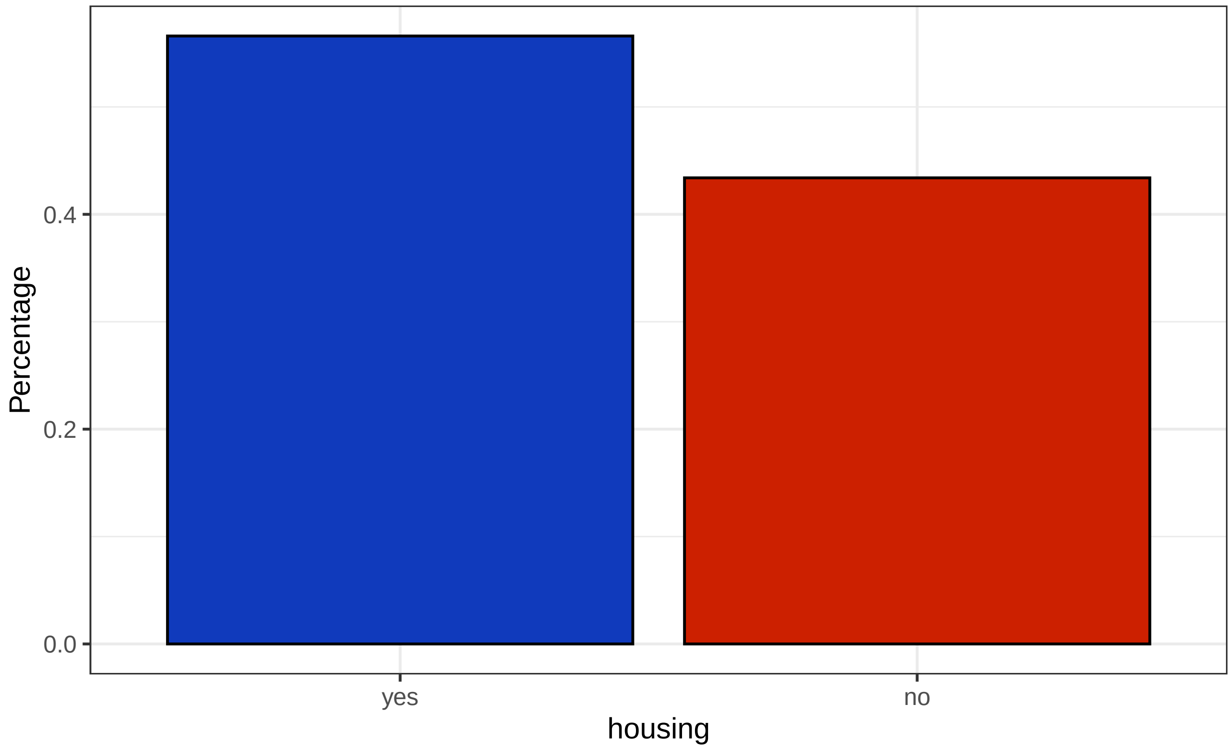
Figures for factor variables

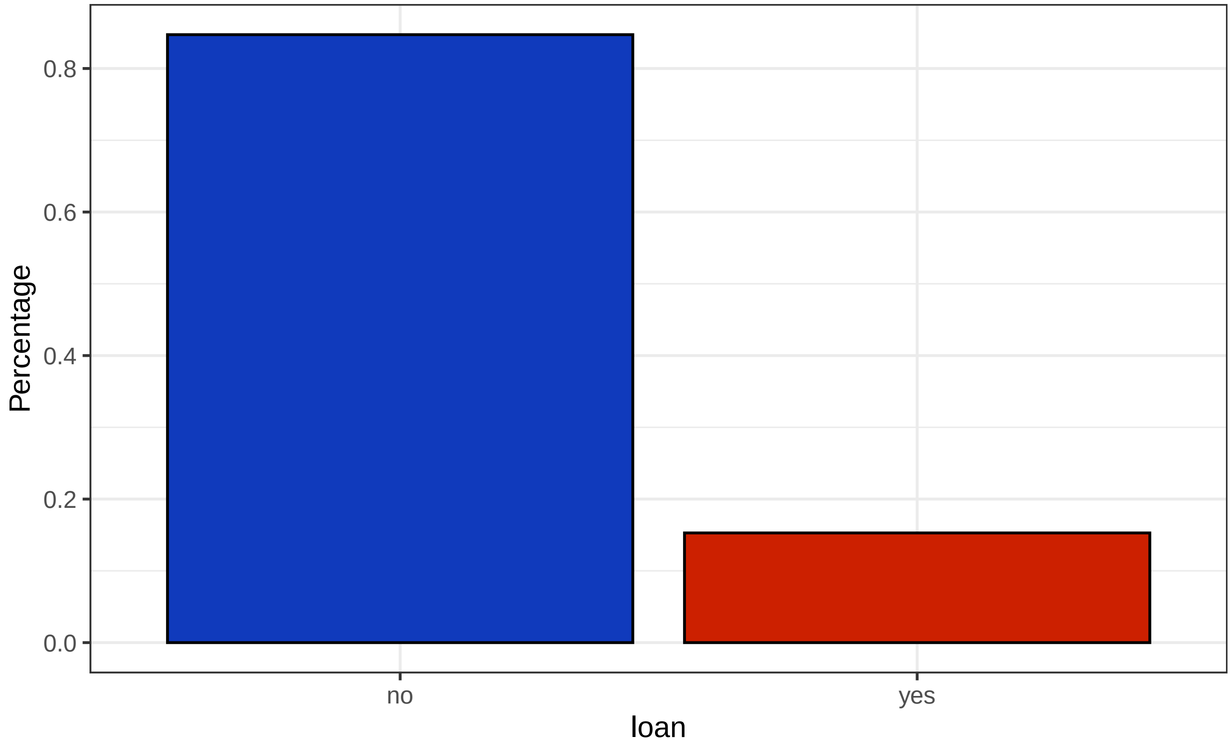


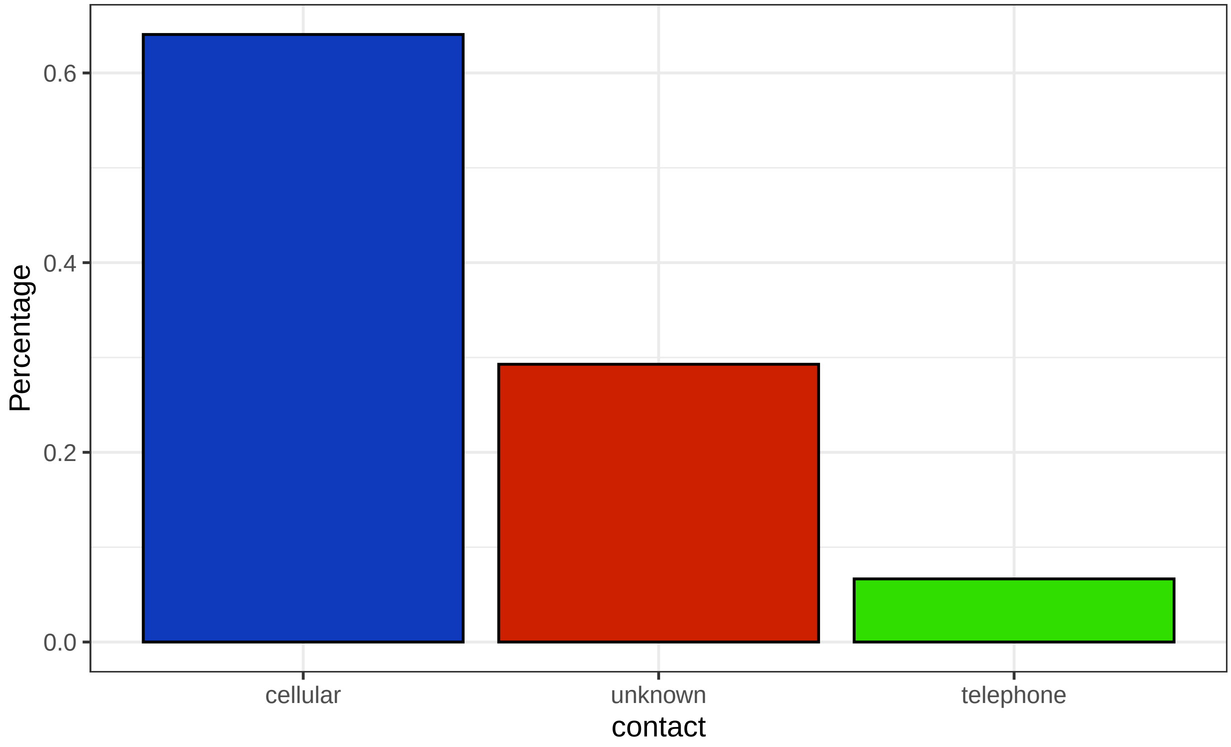


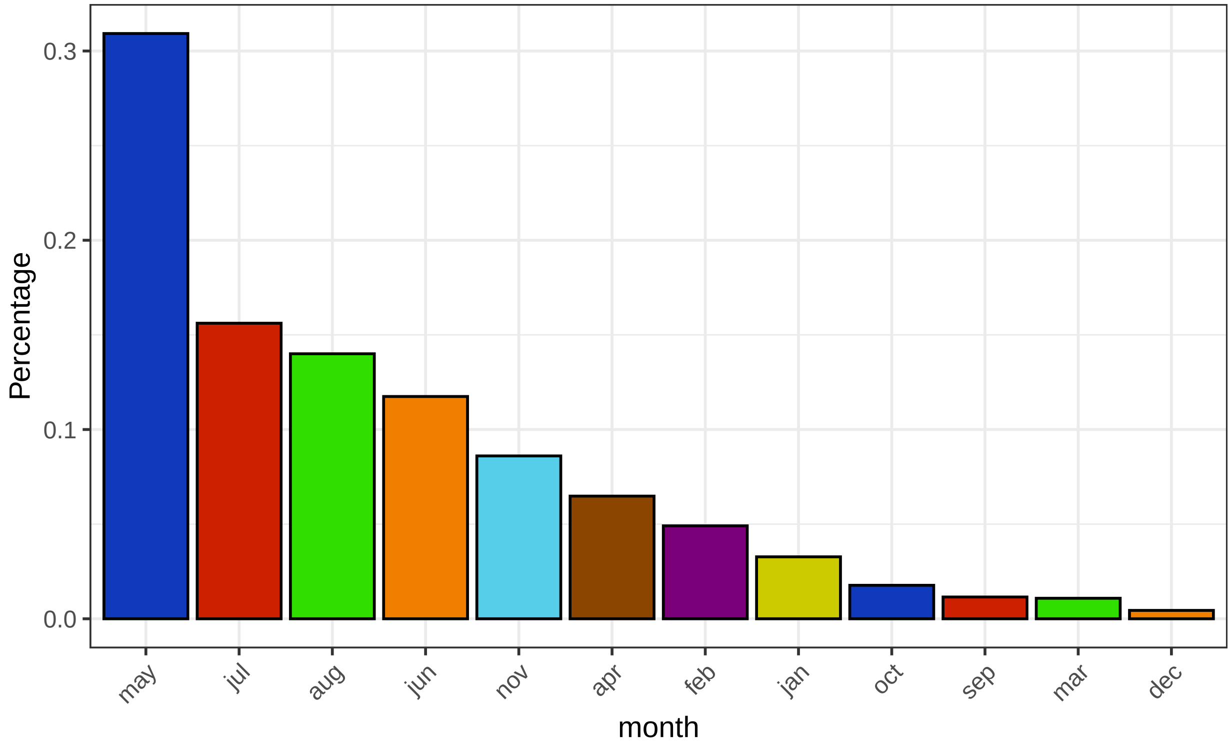

Percentage
0.5
0.4
0.3
0.2
0.1
0.0
secondary
tertiary
primary
unknown
education

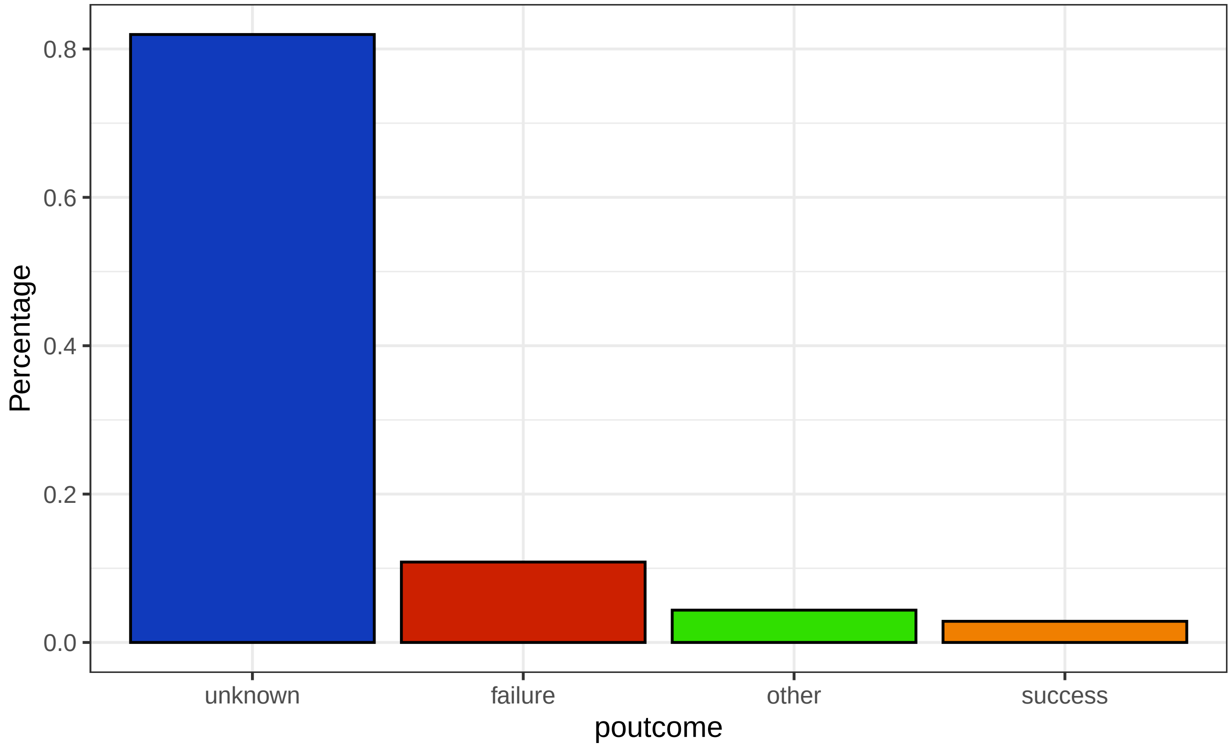

Percentage
1.00
0.75
0.50
0.25
0.00
no
yes
default

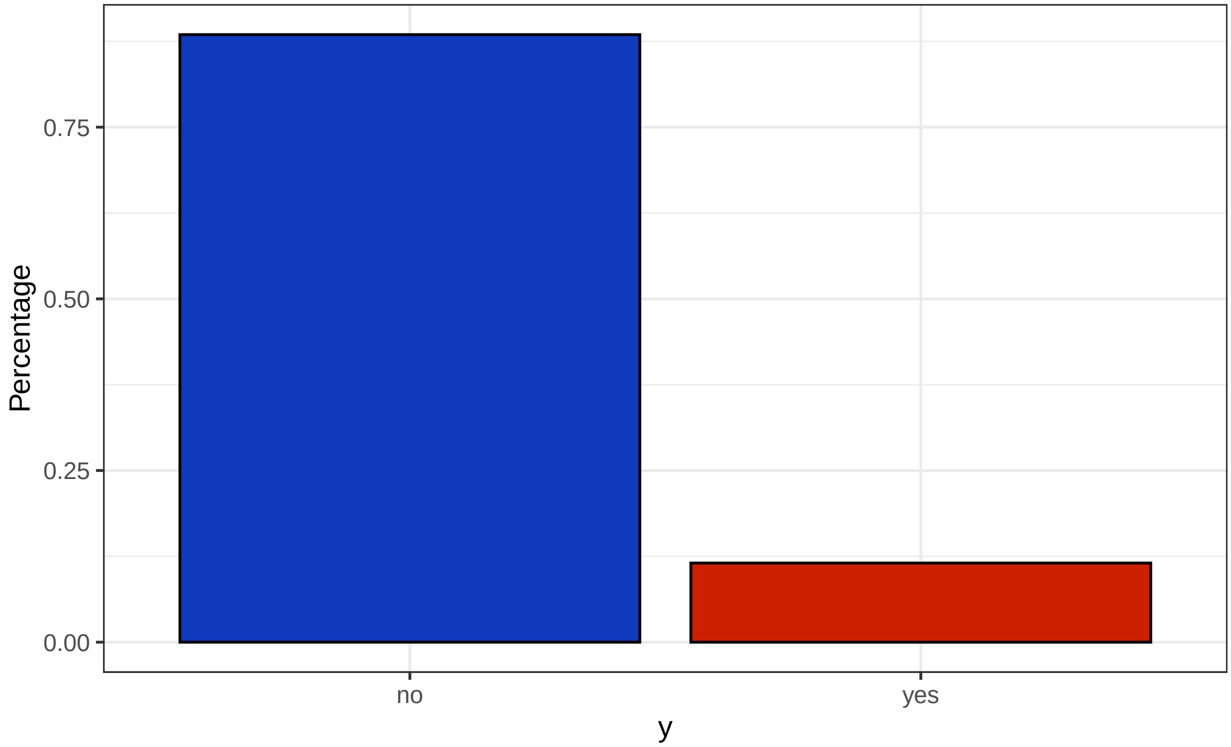

Percentage
0.4
0.2
0.0
yes
no
housing


Percentage
0.8
0.6
0.4
0.2
0.0
no
yes
loan


Percentage
0.6
0.4
0.2
0.0
cellular
unknown
telephone
contact


Percentage
0.3
0.2
0.1
0.0
may
jul
aug
jun
nov
apr
feb
jan
oct
sep
mar
dec
month


0.8
0.6
0.4
0.2
0.0
Percentage
unknown
failure
other
success
poutcome


0.75
0.50
0.25
0.00
Percentage
no
yes
y

Time series

The script can also be used to compare time series. The data variables are the monthly unemployment rates for men and women separately or both combined from 2001 to 2025.

```
library(readxl)
data <- read_excel("arbetsloshet.xlsx")
colnames(data)
```

```
## [1] "År"      "Månad"   "Båda"    "Män"     "Kvinnor"
```

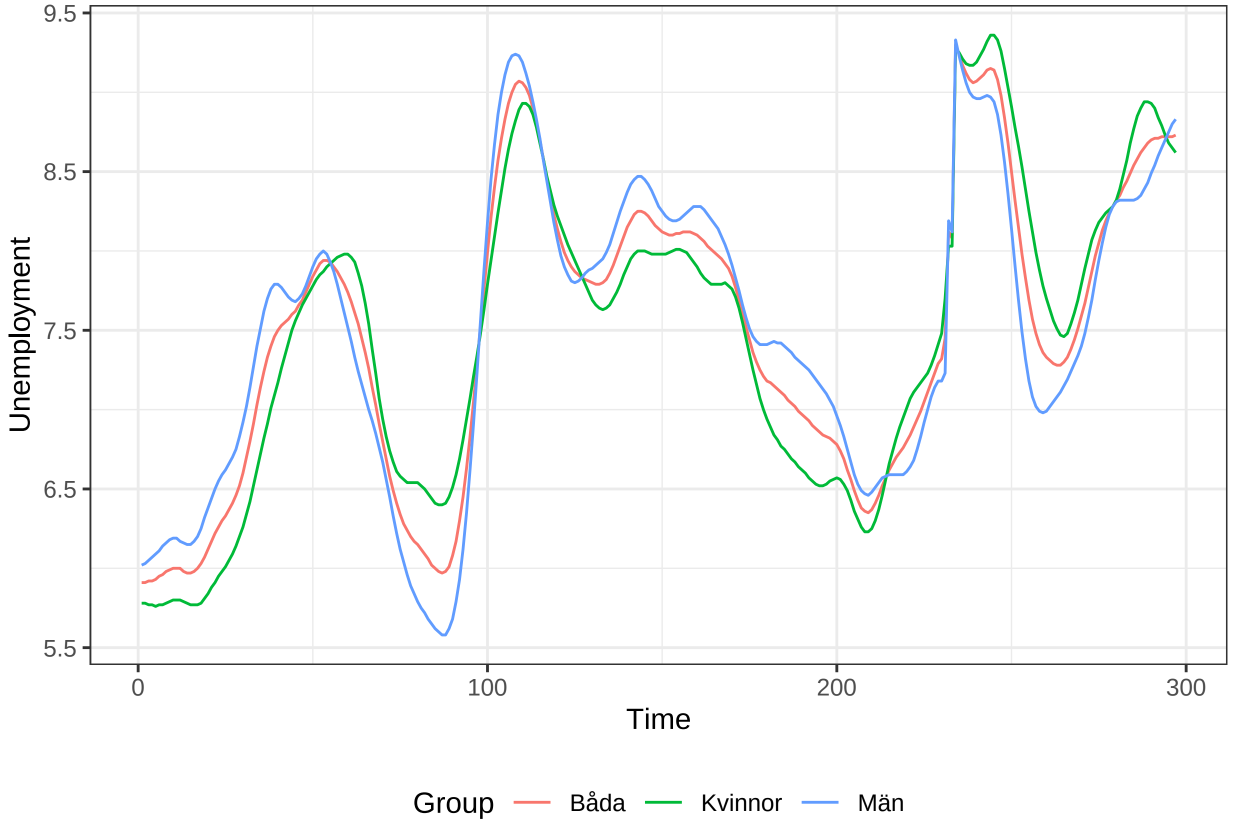
Since the time data are separated in two variables a quick temporary solution is to create a simple time variable.

The user needs to specify which variables that are needed to be plotted.

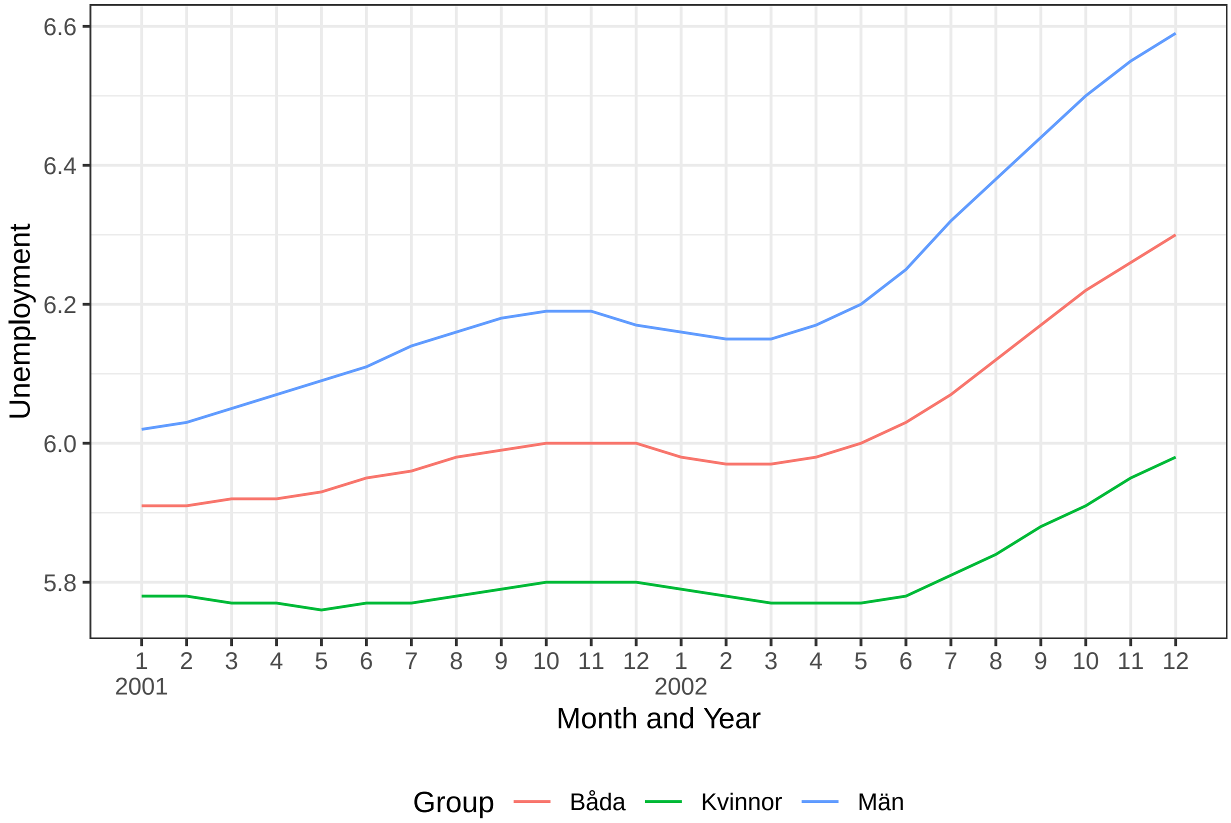
```
plot_data <- long_data(data, cols = c("Båda", "Män", "Kvinnor"))
```

The plot can be modified in many ways, such as if the time points are visualised as dots in the plot, the size of the dots, user defined color palette for the lines, legend position and so on.

```
plot <- plot_line(plot_data,
                  x_val="tid",
                  y_val="value",
                  group_by="group",
                  points=FALSE,
                  legend_pos="bottom")
plot <- plot + labs(x ="Time", y="Unemployment", color="Group")
plot
```



The x-axis could also be modified to present the original month and year time points.


Unemployment
Month and Year
Group
Båda
Kvinnor
Män