Advanced Statistical Modelling: Logistic Regression

Ricard Meyerhofer & Joel Cantero

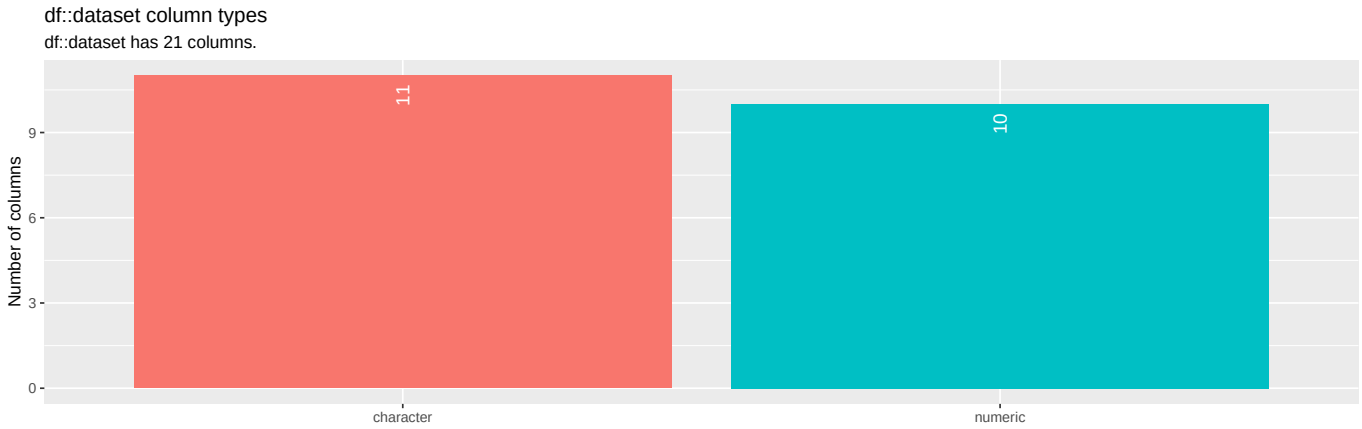
4/11/2019

Exploratory data analysis

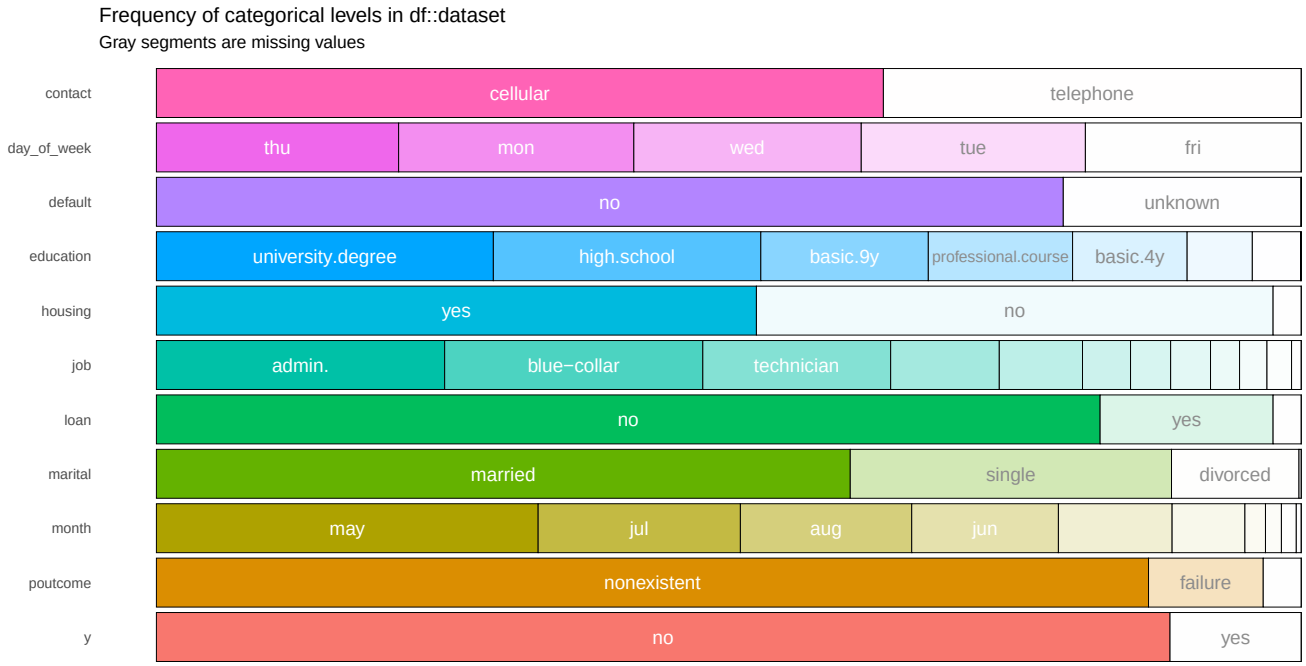
As explained in the problem statement, our dataset is composed by 28645 calls from JYB. JYB has the purpose of reducing the telemarketing costs by decreasing the number of calls to clients not likely to buy the product. This is the list of the available variables:

Variable	Description	Attribute type
id	Customer ID	Client
age	age in years	Client
job	(admin., blue-collar, entrepreneur, housemaid, management, retired, self-employed, services, student, technician, unemployed, unknown)	Client
marital	Marital status (Divorced, married, single, unknown)	Client
education	Education level (basic.4y, basic.6y, basic.9y, high.school, illiterate, professional.course, university.degree, unknown)	Client
default	is he/she a defaulter? (No, yes, unknown)	Client
housing	does he/she has a mortgage? (No, yes, unknown)	Client
loan	does he/she has a personal loan? (No, yes, unknown)	Client
contact	phone type (cellular, telephone)	Call
month	month of the call	Call
day_of_week	day of the call (mon, tue, wed, thu, fri)	Call
campaign	Number of contacts made this campaign for this client (including the current one)	Campaign
pdays	number of days that have passed since the customer was contacted for the last time for a previous campaign (999 means that it was not previously contacted)	Campaign
previous	number of calls made to this client before this campaign	Campaign
poutcome	previous campaign result (failure, nonexistent, success)	Campaign
emp.var.rate	employment variation rate (quarterly)	Indicators
cons.price.idx	Consumer Price Index (monthly)	Indicators
cons.conf.idx	Consumer confidence index (monthly)	Indicators
euribor3m	euribor a 3 mesos (daily)	Indicators
nr.employed	number of employed (quarterly)	Indicators
Y	The customer subscribed the deposit? (yes,no)	Response

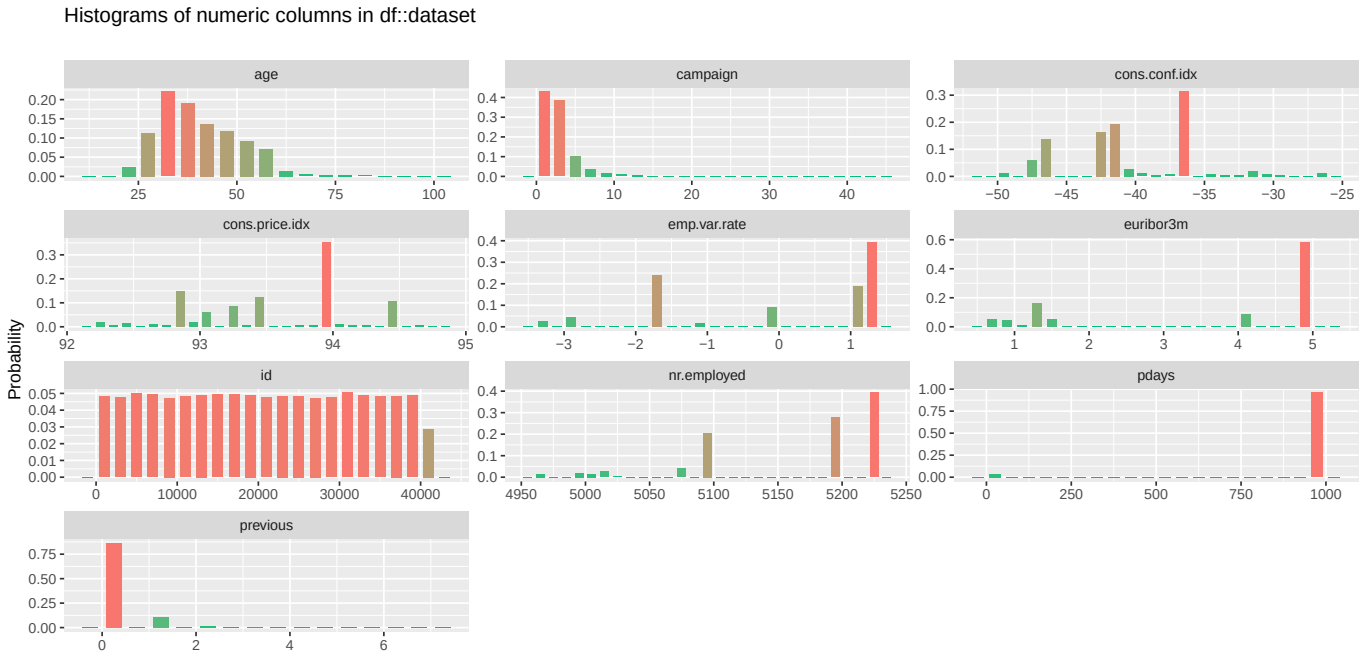
As we can see in the plot, we have 11 factors and 10 numeric values. Furthermore, we have seen that there our dataset is complete which means that it has no missing values. However, this does not imply that there are no outliers.



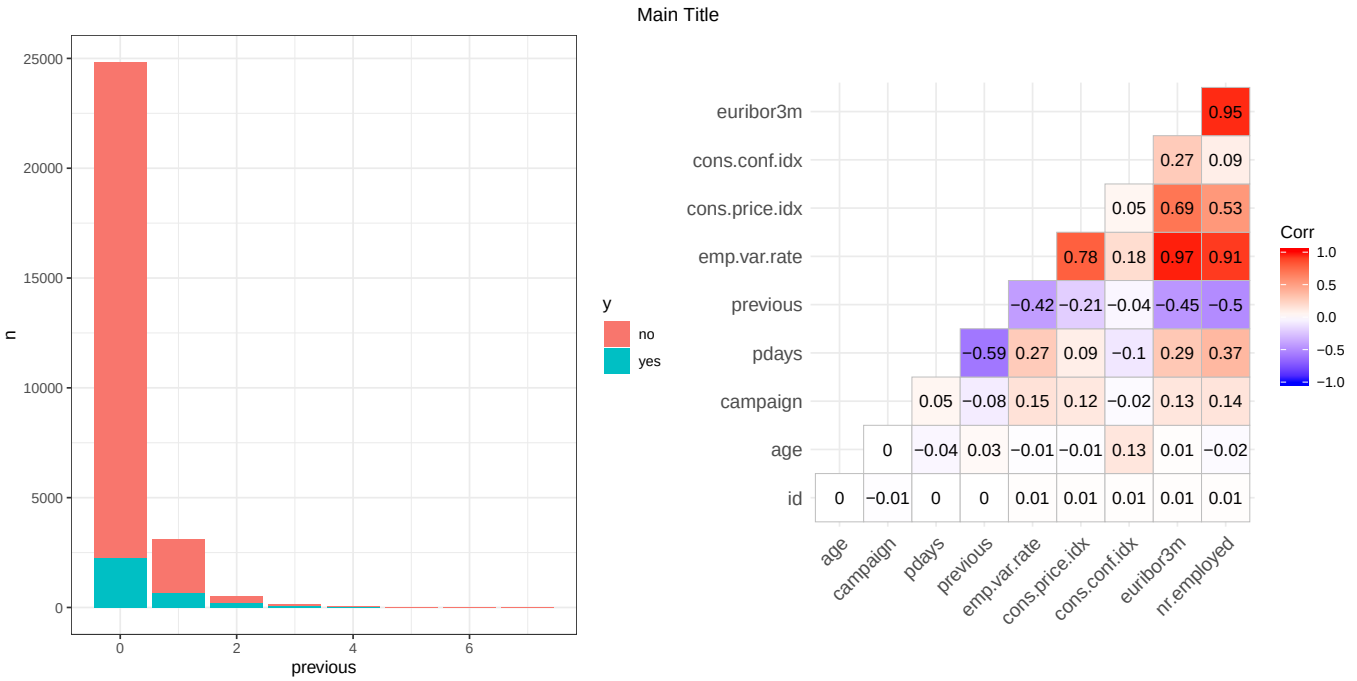
Exploration of the categorical variables where we can see a map with all their possible values and their representation. We can see that there is a clear unbalance in some variables such as *poutcome*, *load*, *contact* or the response variable y. So we might need to resample our dataset.



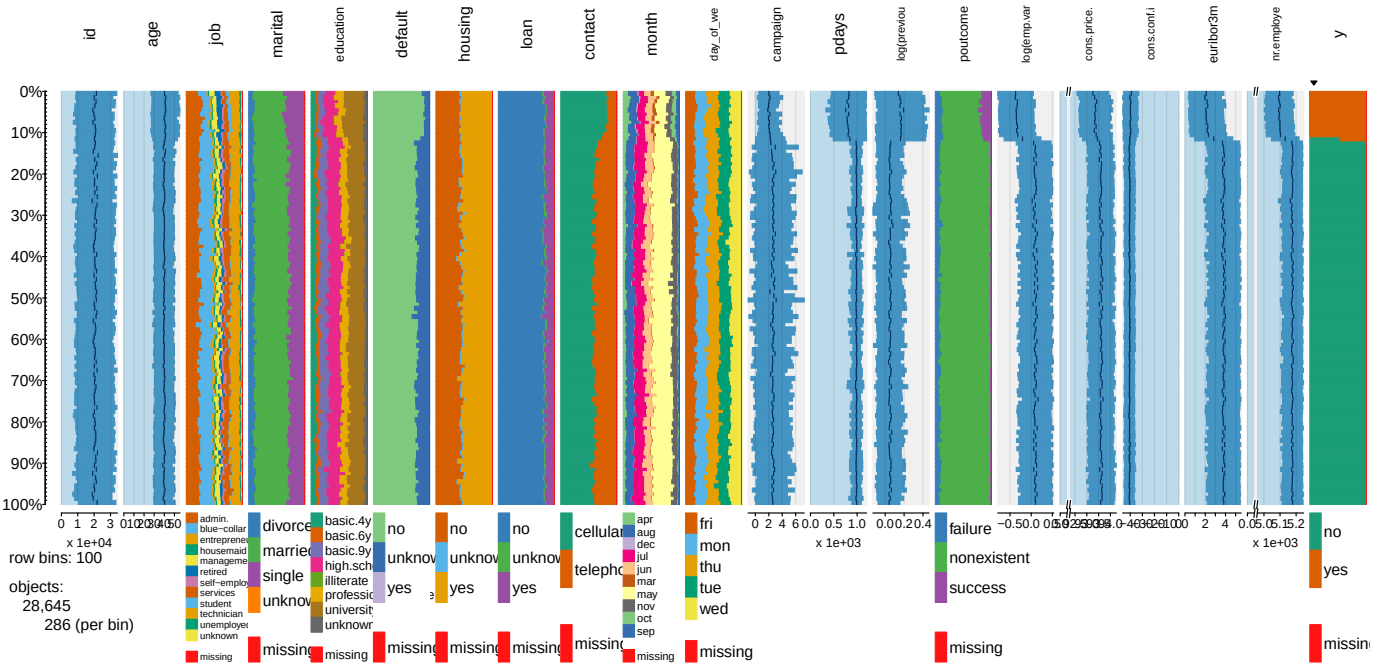
If we now explore of the numerical variables, we see that most of variables are not following a normal. This might be because they are ratios and also are taken most of them in a monthly basis.

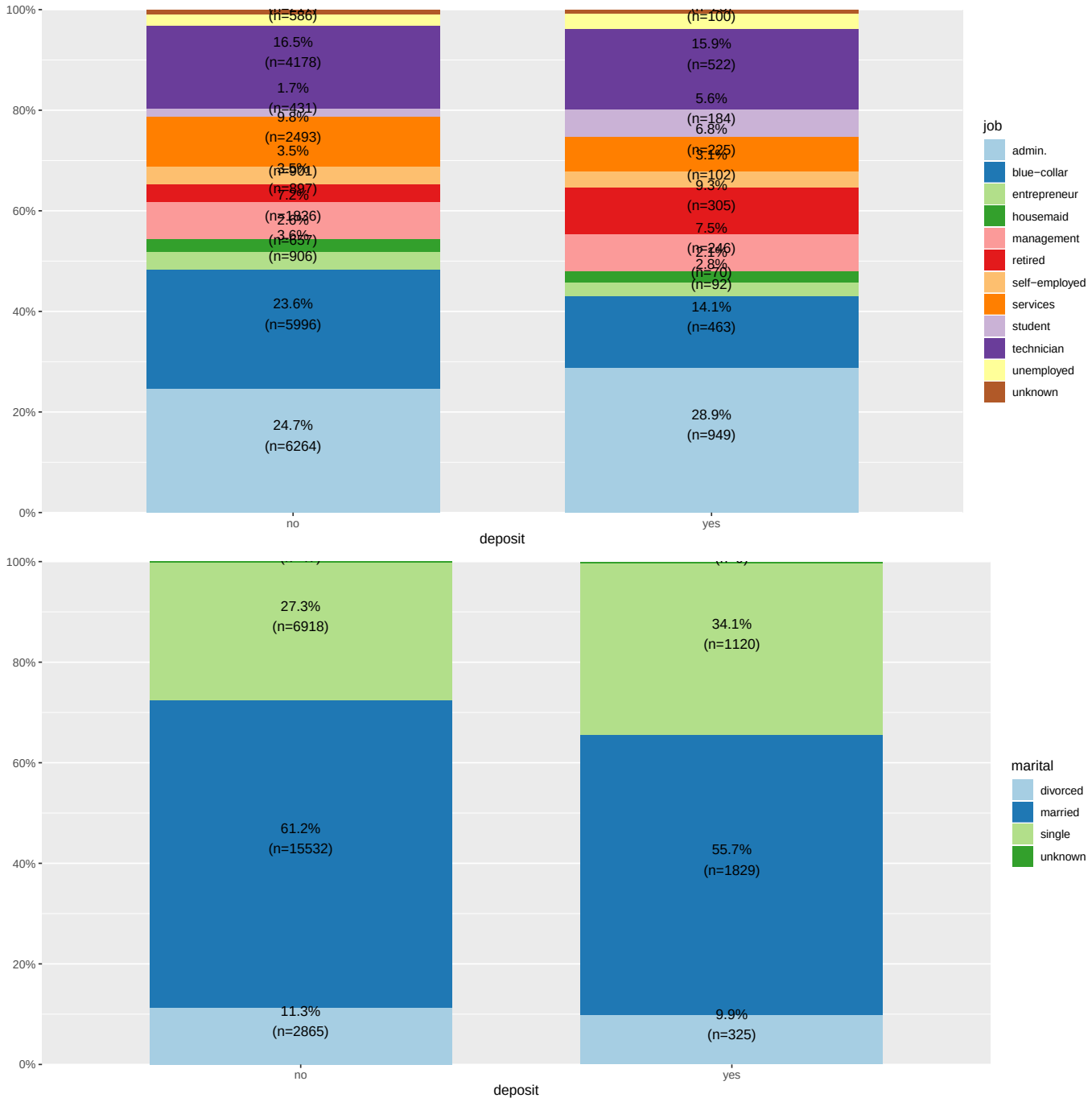


We can also see that the correlation between the variable *previous* and *pdays* is of 0.50 which means that they are moderately correlated. But if we pay attention to number of 0's and 999 and their description, we can see that they are describing the same thing. So we are going to see how these variables are related to the response variable and as we can see below, *pdays* is not necessary. Furthermore, if we take a look at the correlation matrix, we can see that

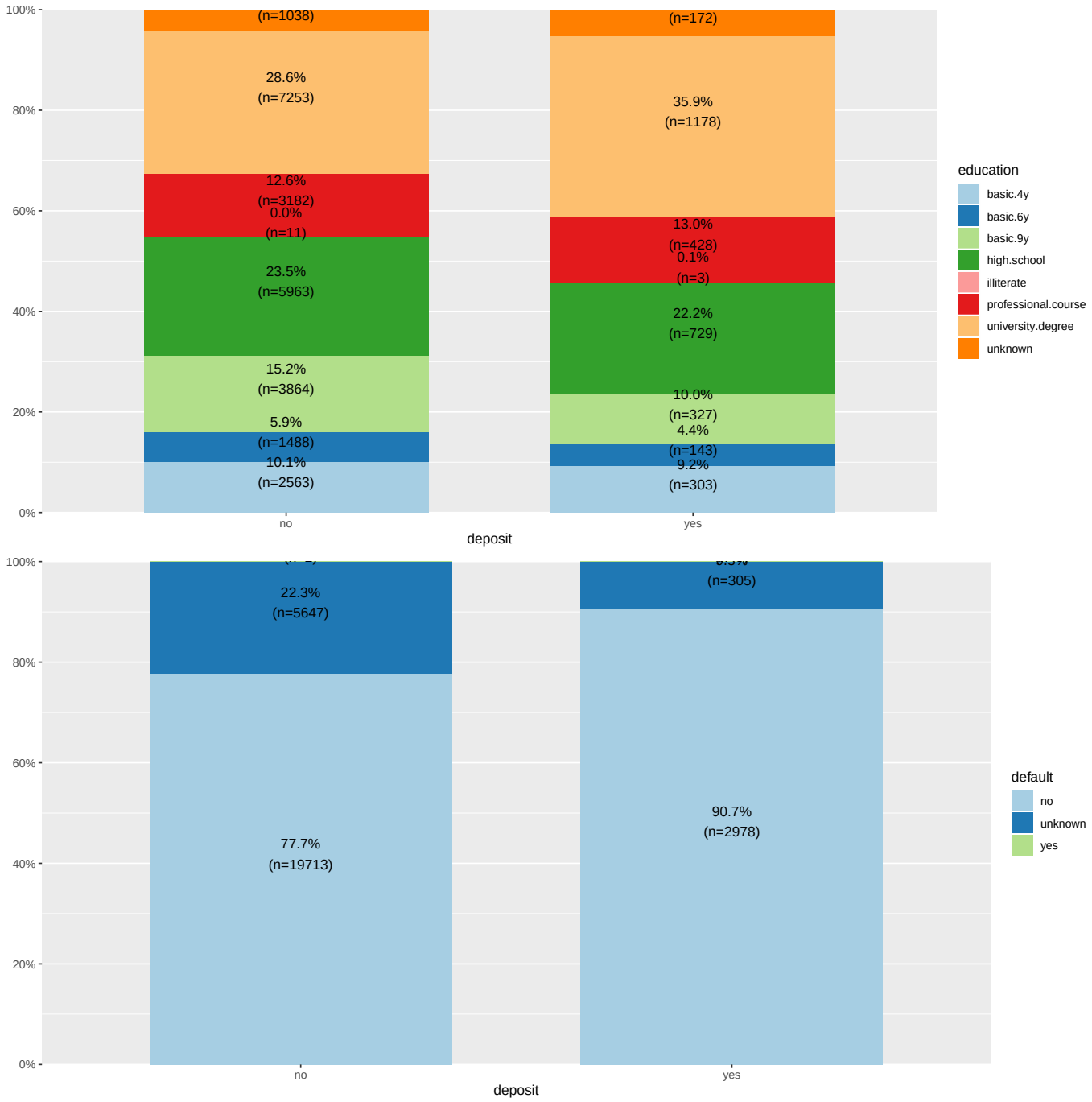


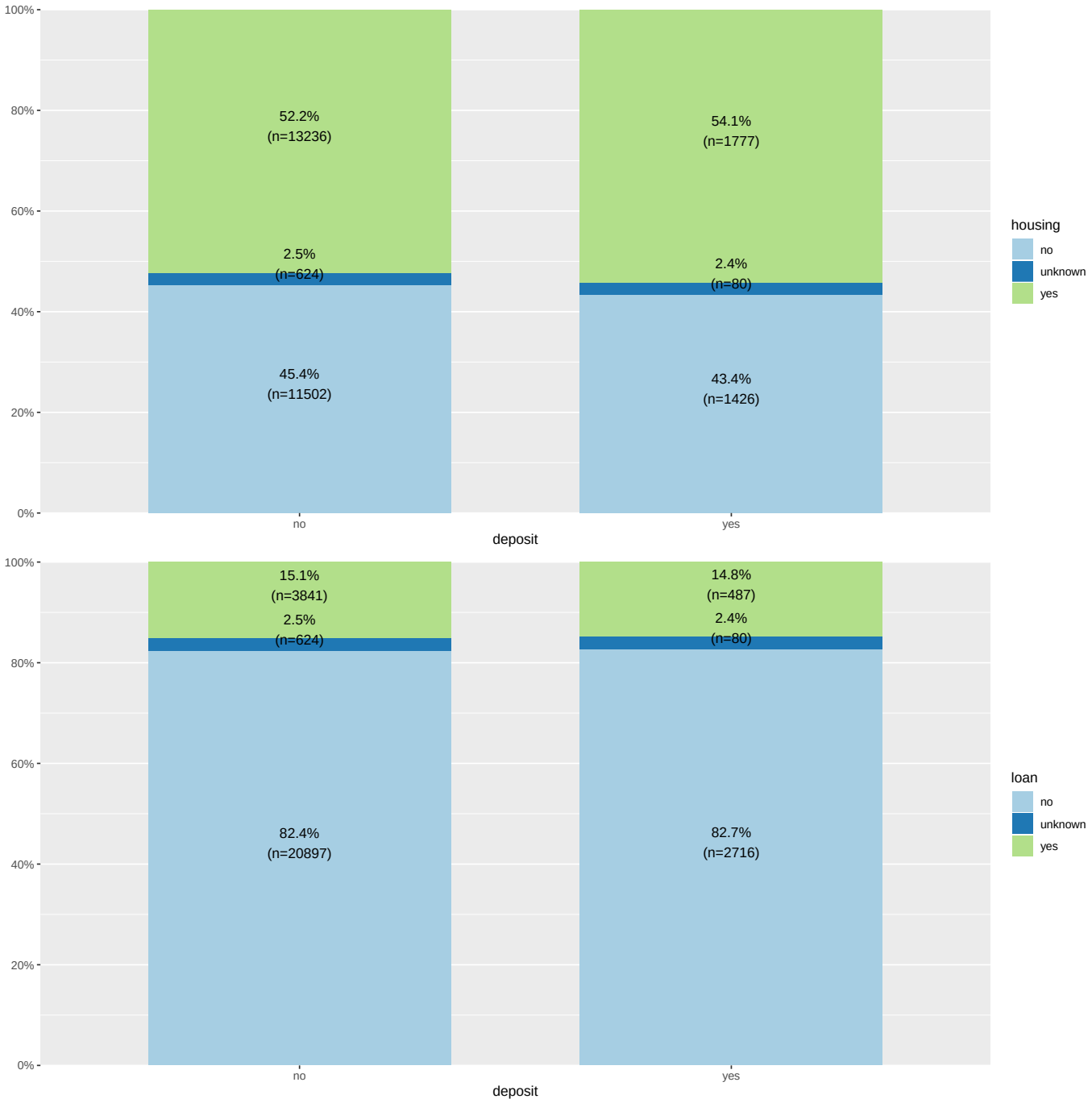
Also we can do a general plot to have a first idea of which variables help us to distinguish between the two groups. As we can see below, it looks like *housing*, *contact* and *potcome* distinguish between the two groups.

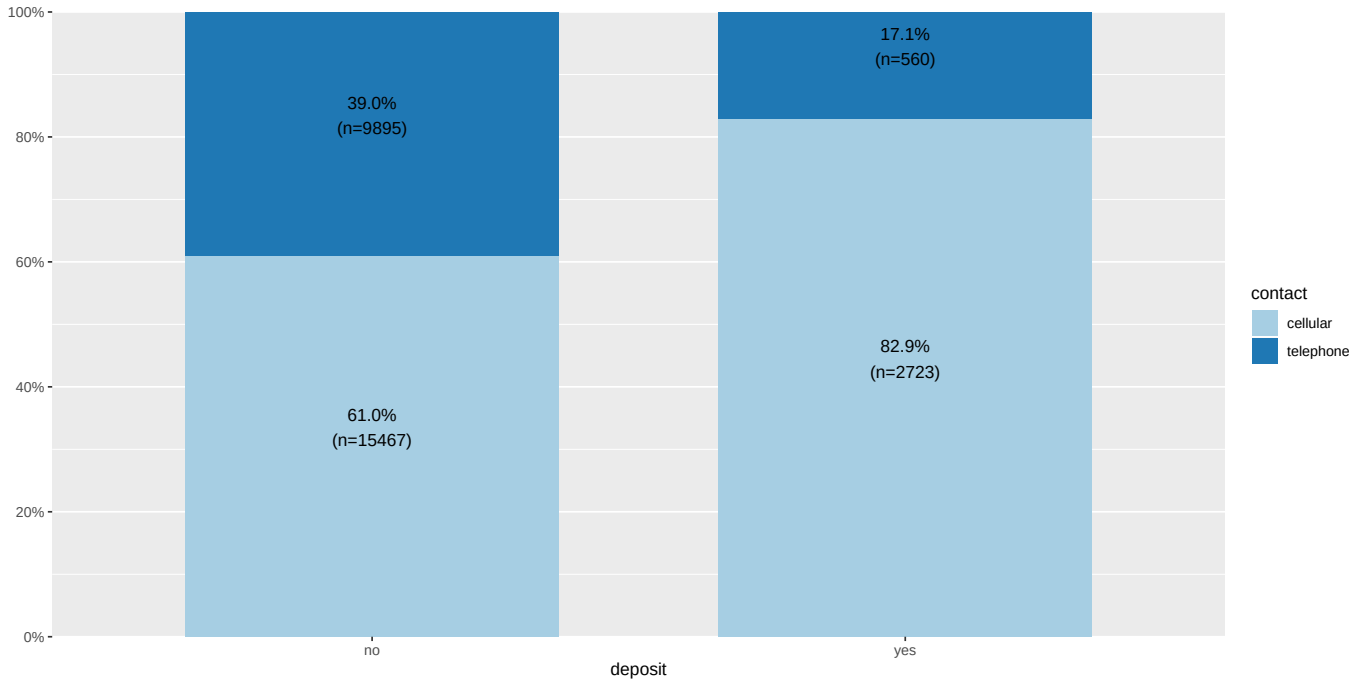


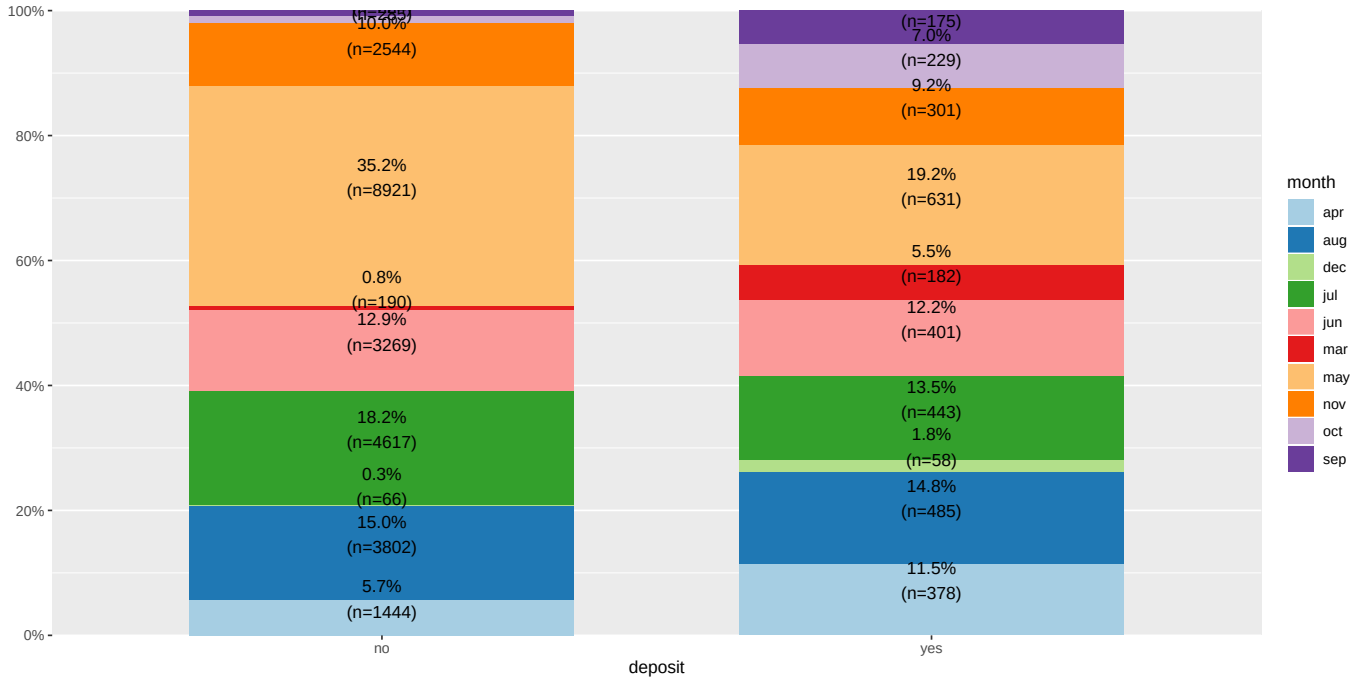

100%
80%
60%
40%
20%
0%
16.5%
(n=4178)
1.7%
(n=431)
9.8%
(n=2493)
3.5%
(n=901)
7.2%
(n=1836)
3.6%
(n=906)
23.6%
(n=5996)
24.7%
(n=6264)
15.9%
(n=522)
5.6%
(n=184)
6.8%
(n=225)
3.1%
(n=102)
9.3%
(n=305)
7.5%
(n=246)
2.8%
(n=92)
14.1%
(n=463)
28.9%
(n=949)
no
yes
deposit
job
admin.
blue−collar
entrepreneur
housemaid
management
retired
self−employed
services
student
technician
unemployed
unknown

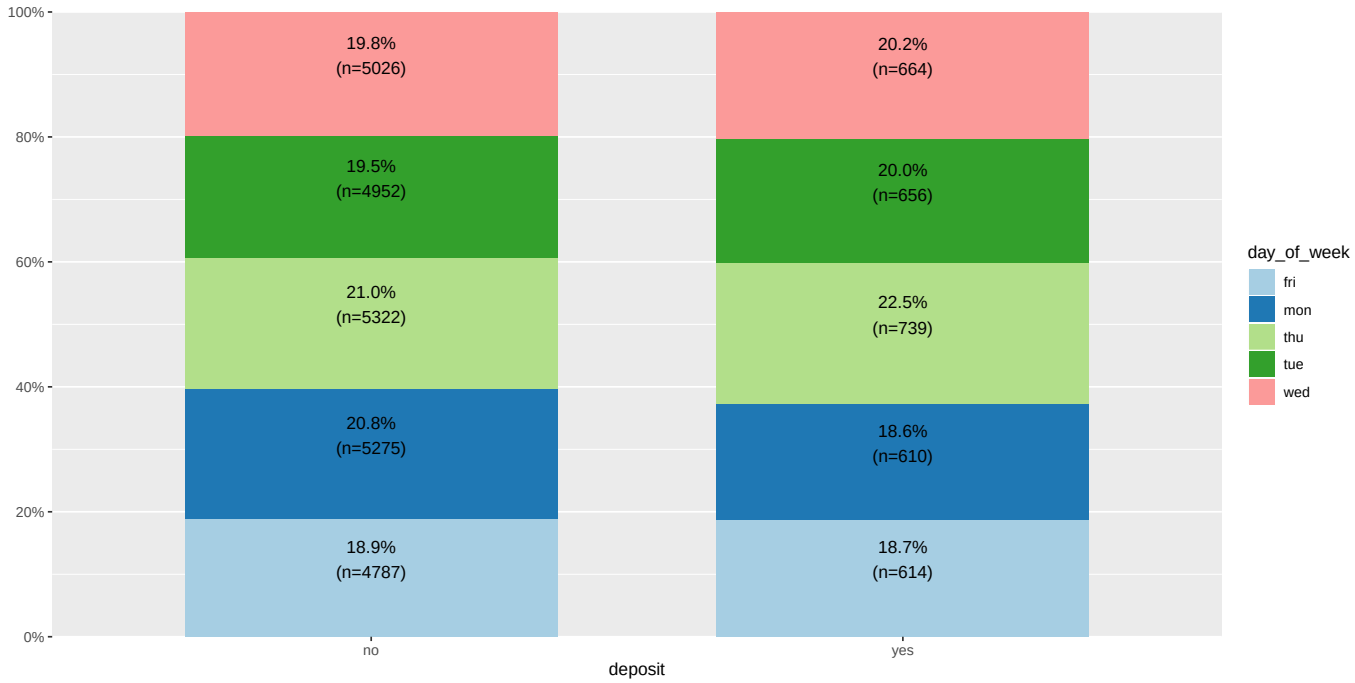
100%
80%
60%
40%
20%
0%
27.3%
(n=6918)
61.2%
(n=15532)
11.3%
(n=2865)
34.1%
(n=1120)
55.7%
(n=1829)
9.9%
(n=325)
no
yes
deposit
marital
divorced
married
single
unknown

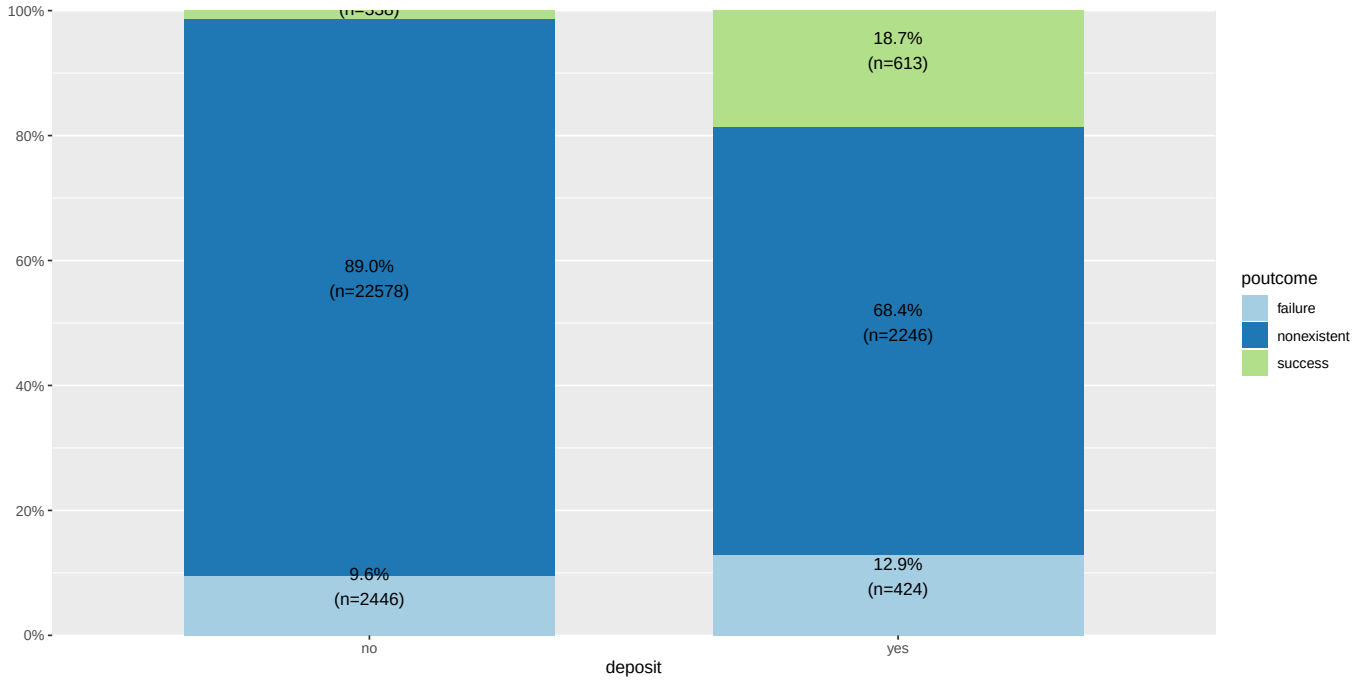

100%
80%
60%
40%
20%
0%
(n=1038)
28.6%
(n=7253)
12.6%
(n=3182)
0.0%
(n=11)
23.5%
(n=5963)
15.2%
(n=3864)
5.9%
(n=1488)
10.1%
(n=2563)
(n=172)
35.9%
(n=1178)
13.0%
(n=428)
0.1%
(n=3)
22.2%
(n=729)
10.0%
(n=327)
4.4%
(n=143)
9.2%
(n=303)
no
yes
deposit
education
basic.4y
basic.6y
basic.9y
high.school
illiterate
professional.course
university.degree
unknown
22.3%
(n=5647)
77.7%
(n=19713)
(n=305)
90.7%
(n=2978)
default
no
unknown
yes

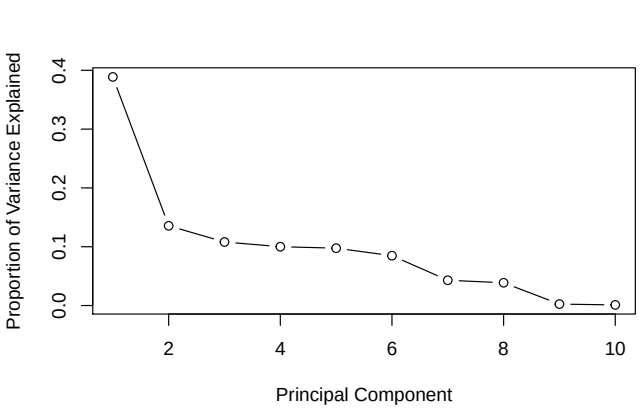

100%
80%
60%
40%
20%
0%
52.2%
(n=13236)
2.5%
(n=624)
45.4%
(n=11502)
54.1%
(n=1777)
2.4%
(n=80)
43.4%
(n=1426)
no
yes
deposit
housing
no
unknown
yes
15.1%
(n=3841)
2.5%
(n=624)
82.4%
(n=20897)
14.8%
(n=487)
2.4%
(n=80)
82.7%
(n=2716)
loan

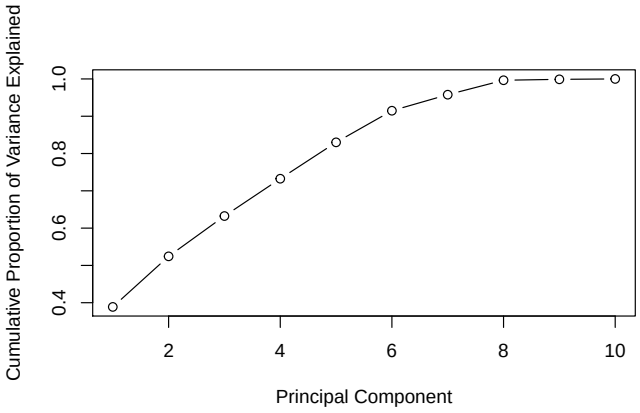

100%
80%
60%
40%
20%
0%
39.0%
(n=9895)
61.0%
(n=15467)
17.1%
(n=560)
82.9%
(n=2723)
no
yes
deposit
contact
cellular
telephone


100%
80%
60%
40%
20%
0%
10.0%
(n=2544)
35.2%
(n=8921)
0.8%
(n=190)
12.9%
(n=3269)
18.2%
(n=4617)
0.3%
(n=66)
15.0%
(n=3802)
5.7%
(n=1444)
(n=175)
7.0%
(n=229)
9.2%
(n=301)
19.2%
(n=631)
5.5%
(n=182)
12.2%
(n=401)
13.5%
(n=443)
1.8%
(n=58)
14.8%
(n=485)
11.5%
(n=378)
no
yes
deposit
month
apr
aug
dec
jul
jun
mar
may
nov
oct
sep

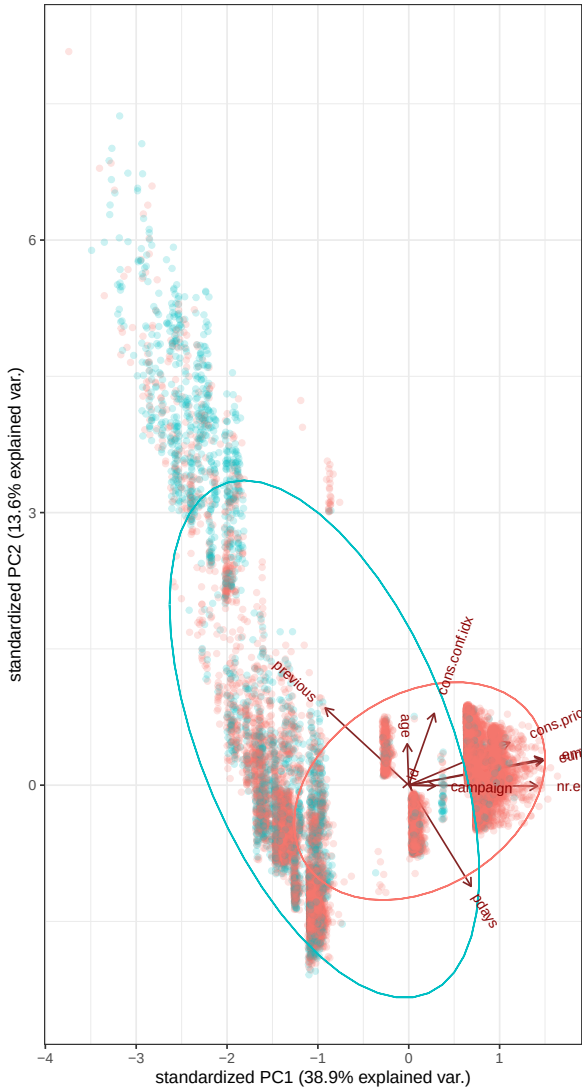

100%
80%
60%
40%
20%
0%
19.8%
(n=5026)
19.5%
(n=4952)
21.0%
(n=5322)
20.8%
(n=5275)
18.9%
(n=4787)
20.2%
(n=664)
20.0%
(n=656)
22.5%
(n=739)
18.6%
(n=610)
18.7%
(n=614)
no
yes
deposit
day_of_week
fri
mon
thu
tue
wed

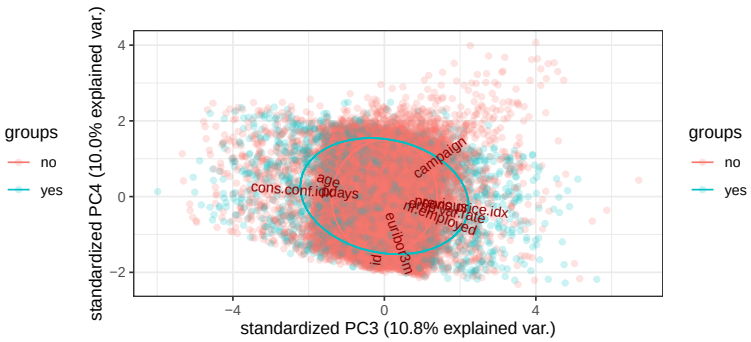

100%
80%
60%
40%
20%
0%
89.0%
(n=22578)
9.6%
(n=2446)
18.7%
(n=613)
68.4%
(n=2246)
12.9%
(n=424)
no
yes
deposit
poutcome
failure
nonexistent
success


Proportion of Variance Explained
0.0
0.1
0.2
0.3
0.4
2
4
6
8
10
Principal Component


Cumulative Proportion of Variance Explained
0.4
0.6
0.8
1.0
2
4
6
8
10
Principal Component

Main Title


standardized PC2 (13.6% explained var.)
standardized PC1 (38.9% explained var.)
previous
cons.conf.idx
age
campaign
pdays
6
3
0
−4
−3
−2
−1
0
1


standardized PC4 (10.0% explained var.)
standardized PC3 (10.8% explained var.)
groups
no
yes
campaign
age
cons.conf.idx
pdays
euribor3m
nr.employed
4
2
0
−2
−4
0
4