

Advanced Statistical Modelling: Logistic Regression

Ricard Meyerhofer & Joel Cantero

4/11/2019

Exploratory data analysis

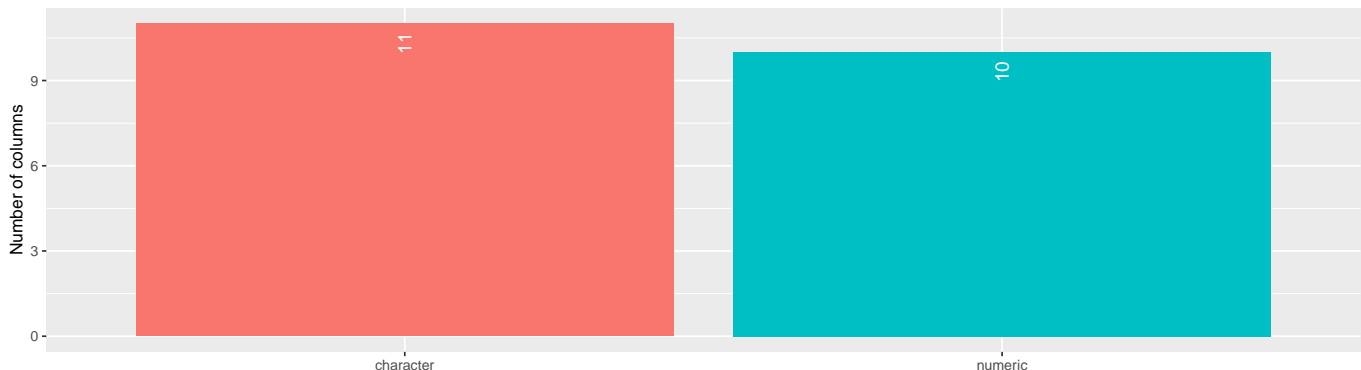
As explained in the problem statement, our dataset is composed by 28645 calls from JYB. JYB has the purpose of reducing the telemarketing costs by decreasing the number of calls to clients not likely to buy the product. This is the list of the available variables:

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 meses (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.

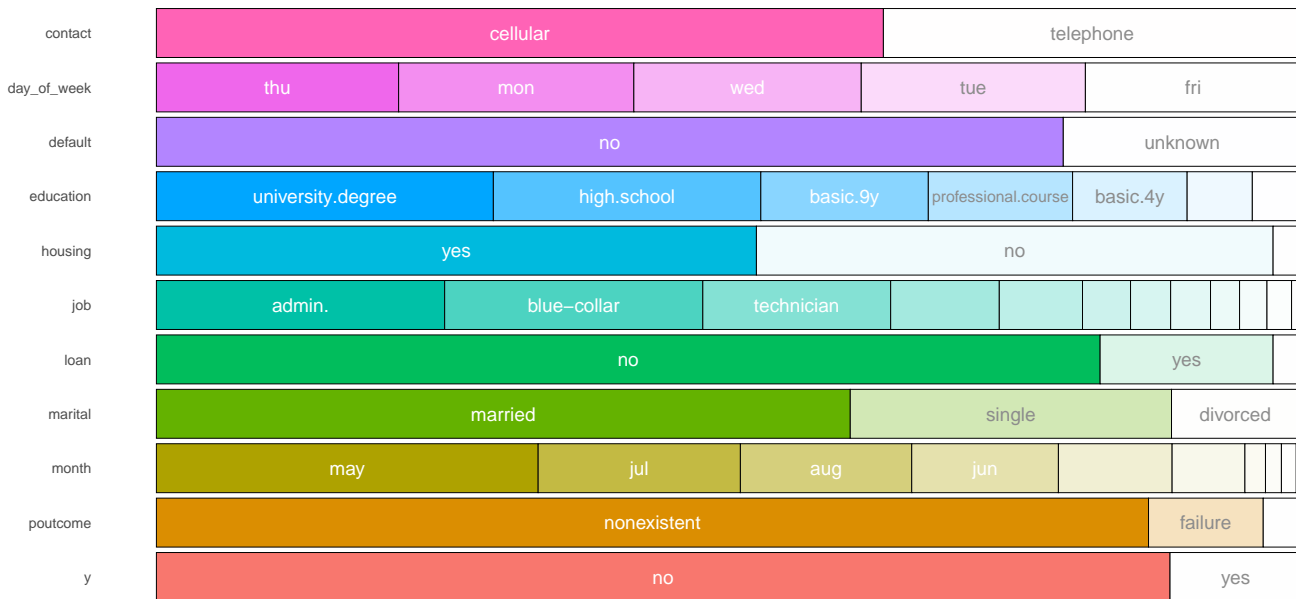
df::dataset column types

df::dataset has 21 columns.



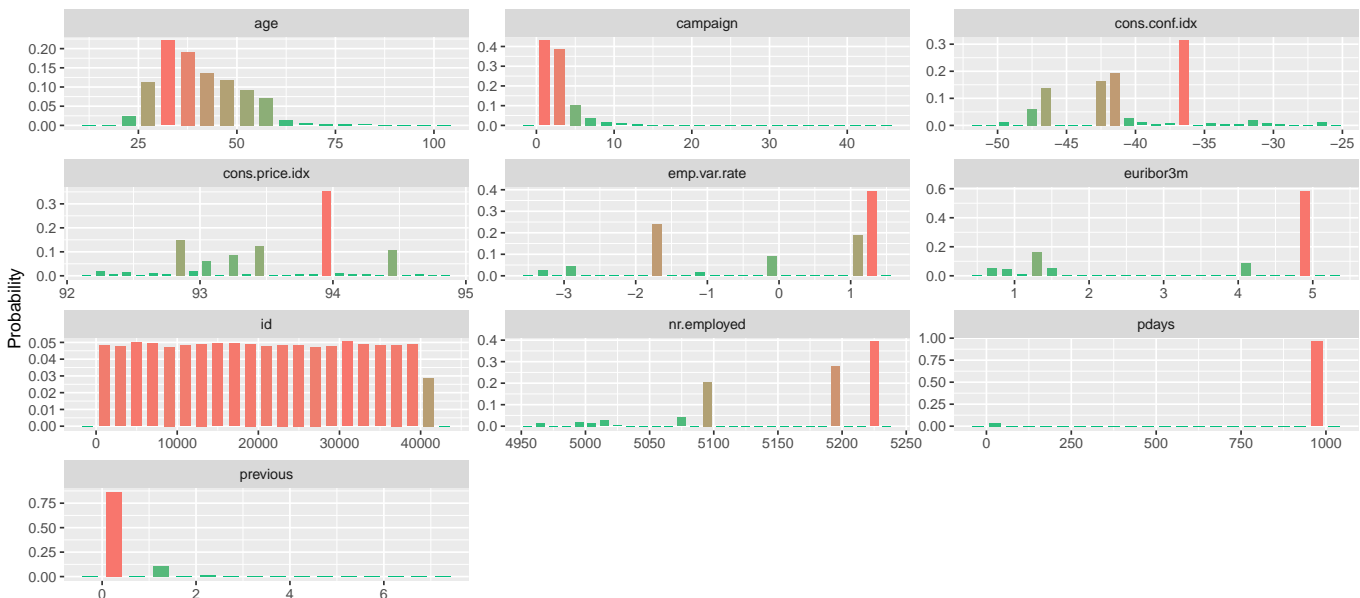
Exploration of the categorical variables where we can see a map with all their possible values and their representation. We can see that there is a clear unbalance in some variables such as *poutcome*, *load*, *contact* or the response variable *y*. So we might need to resample our dataset.

Frequency of categorical levels in df::dataset
Gray segments are missing values

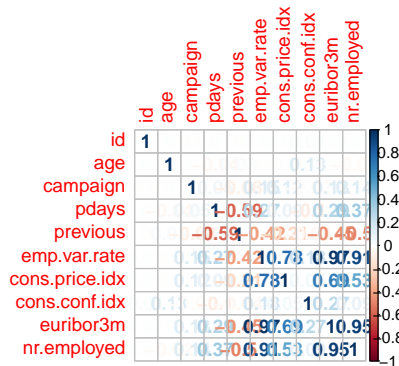
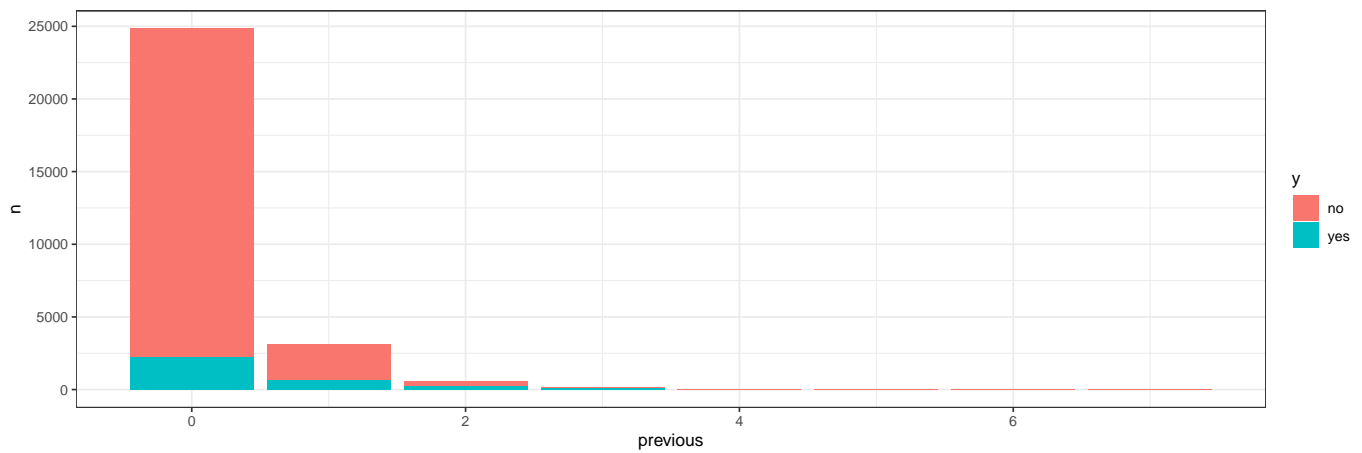


If we now explore of the numerical variables, we see that most of variables are not following a normal. This might be because they are ratios and also are taken most of them in a monthly basis.

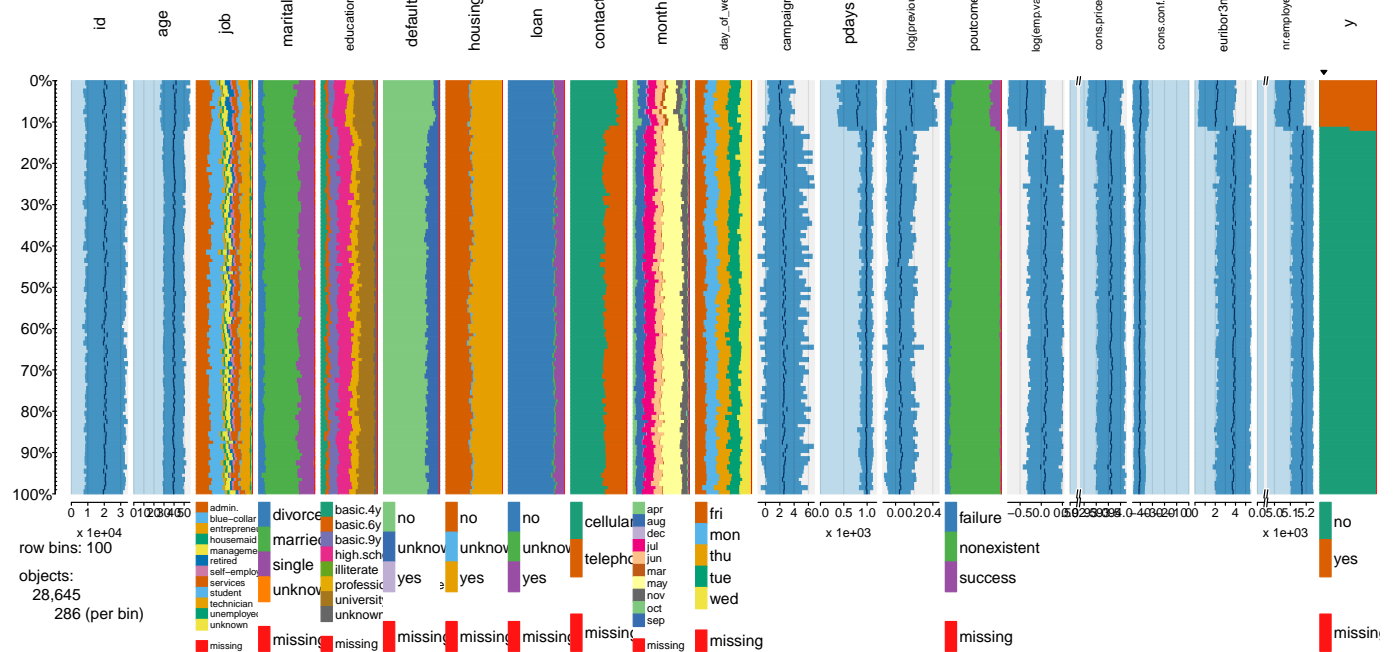
Histograms of numeric columns in df::dataset

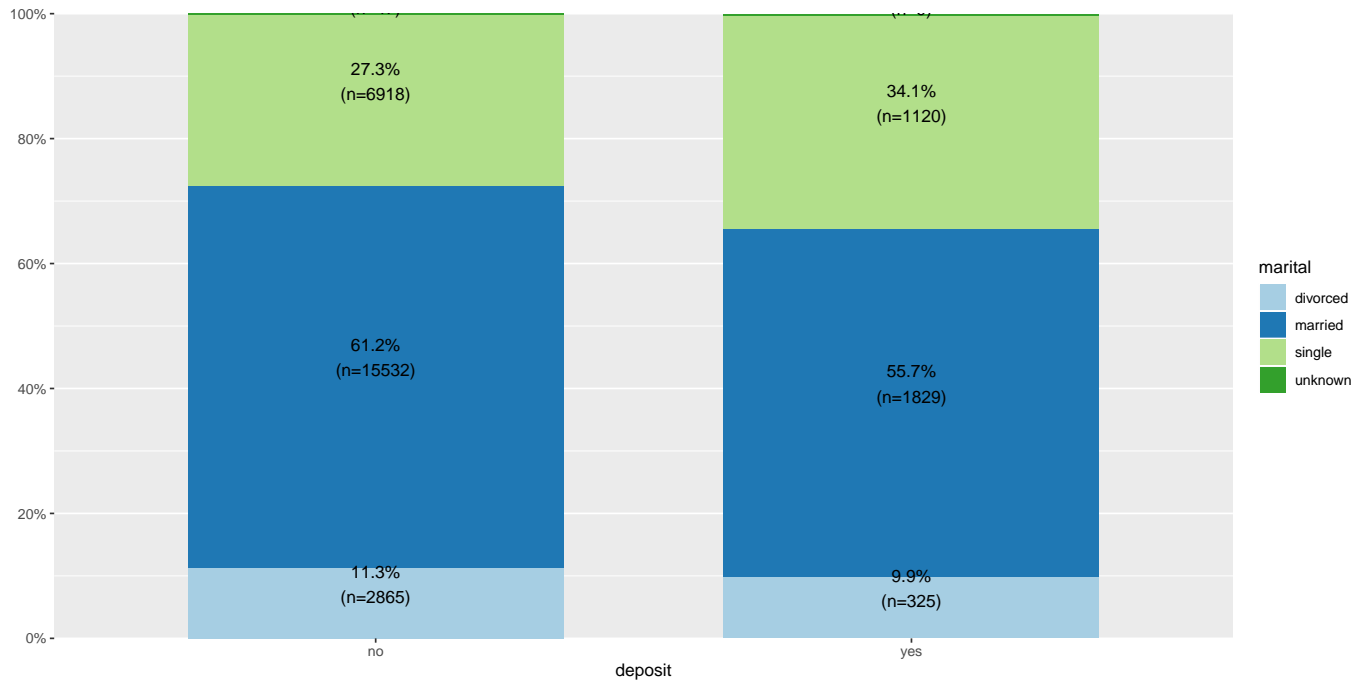
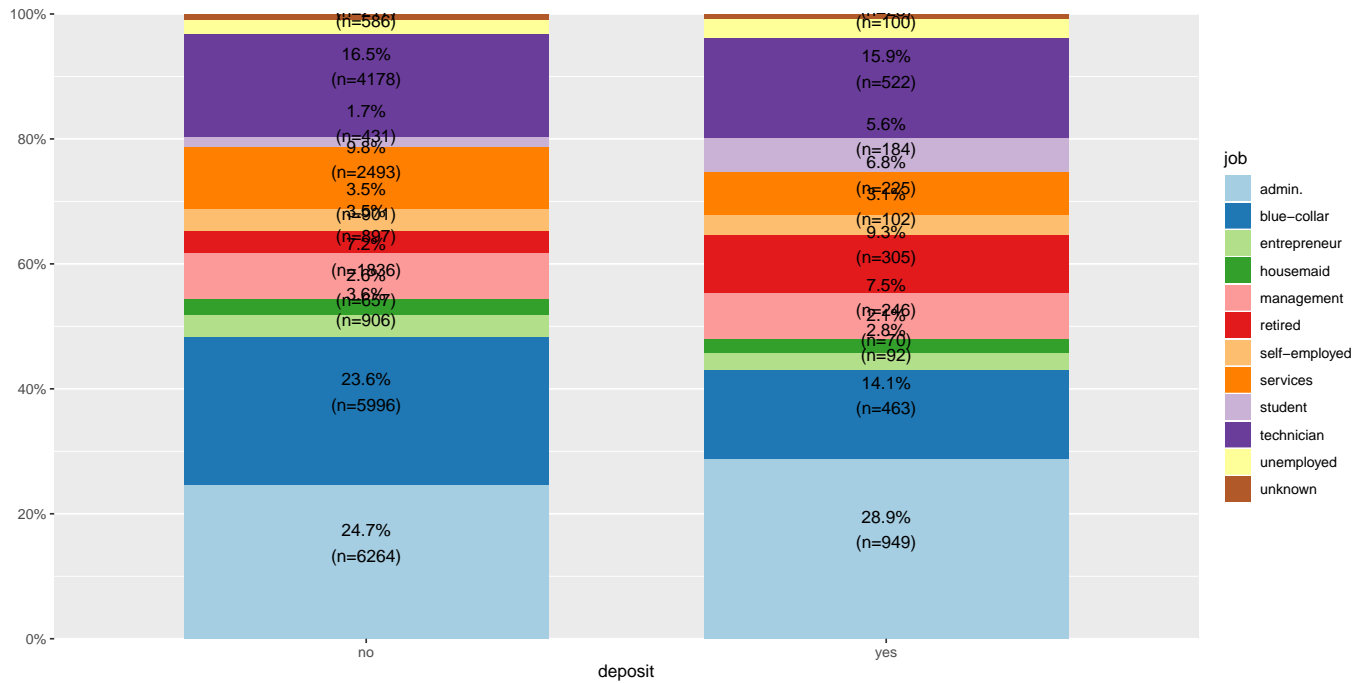


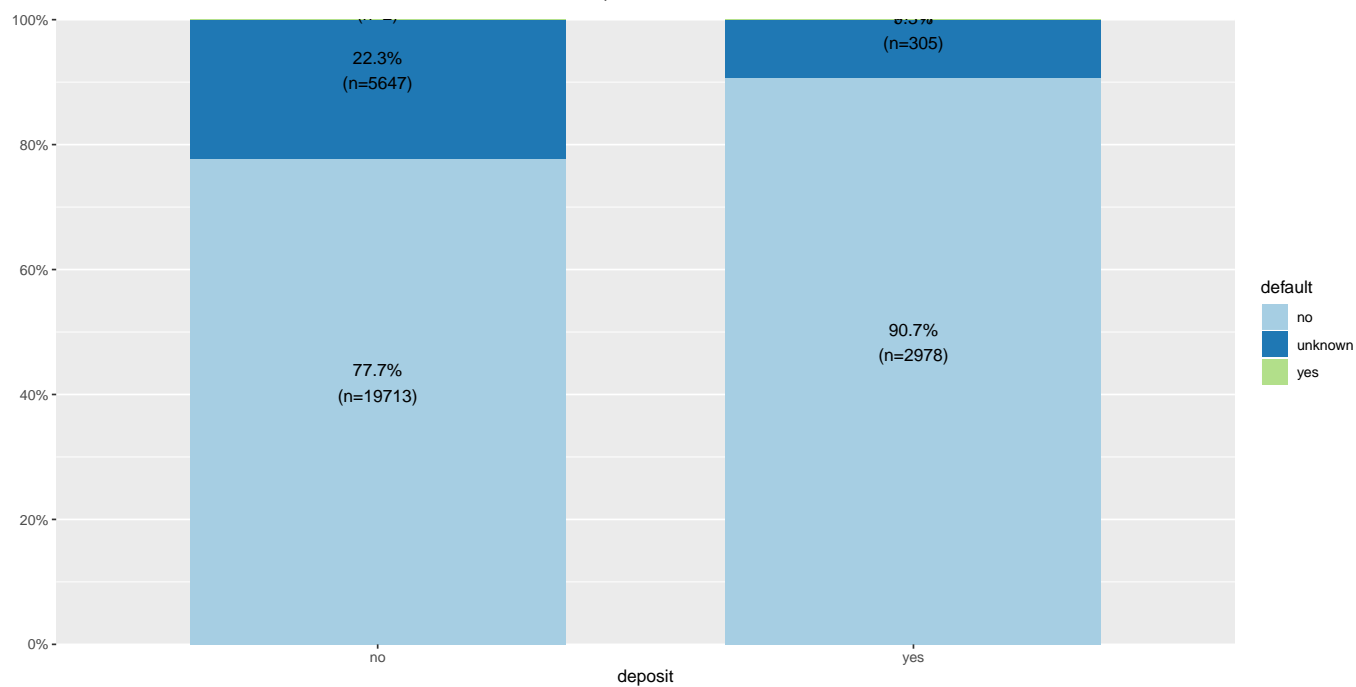
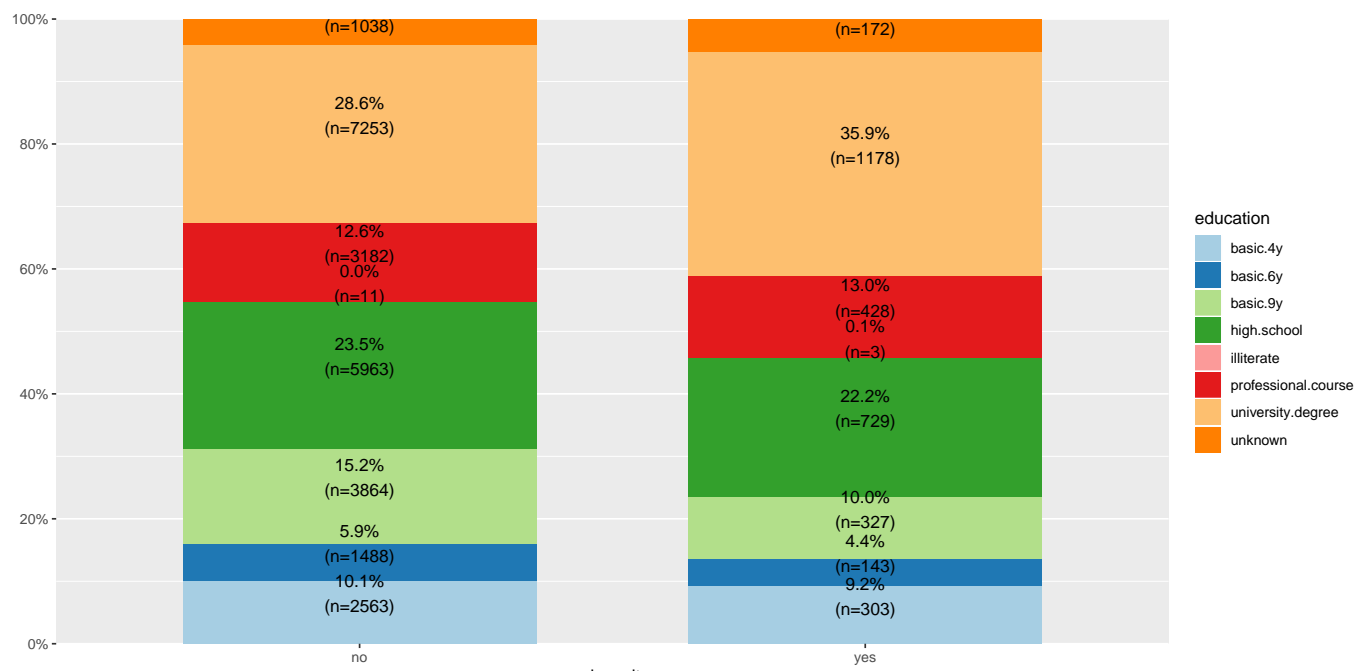
We can also see that the correlation between the variable *previous* and *pdays* is of 0.50 which means that they are moderately correlated. But if we pay attention to number of 0's and 999 and their description, we can see that they are describing the same thing. So we are going to see how these variables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

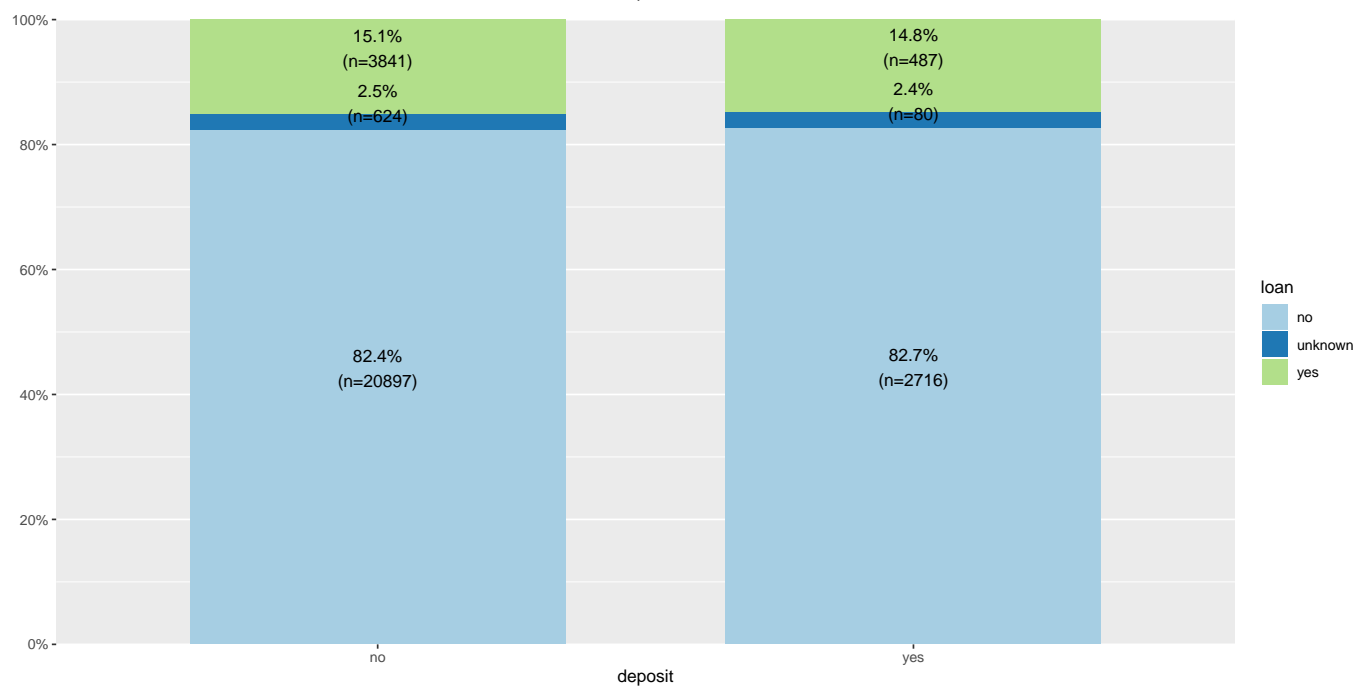
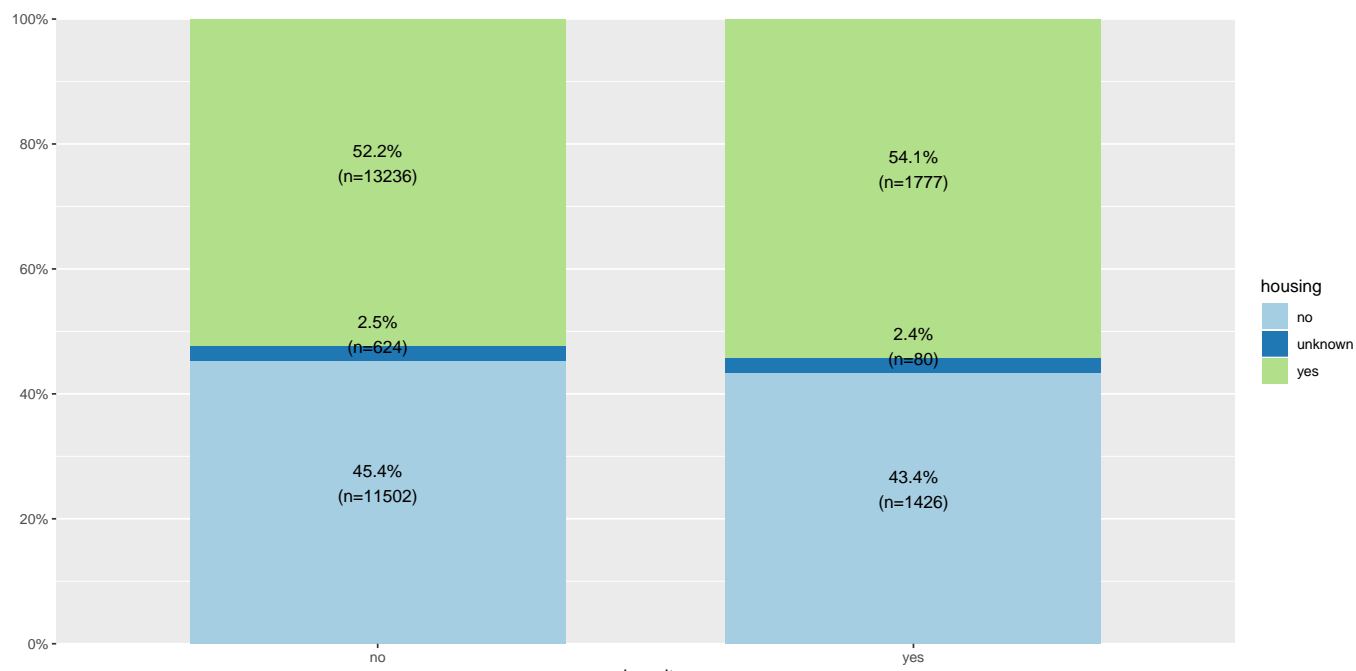


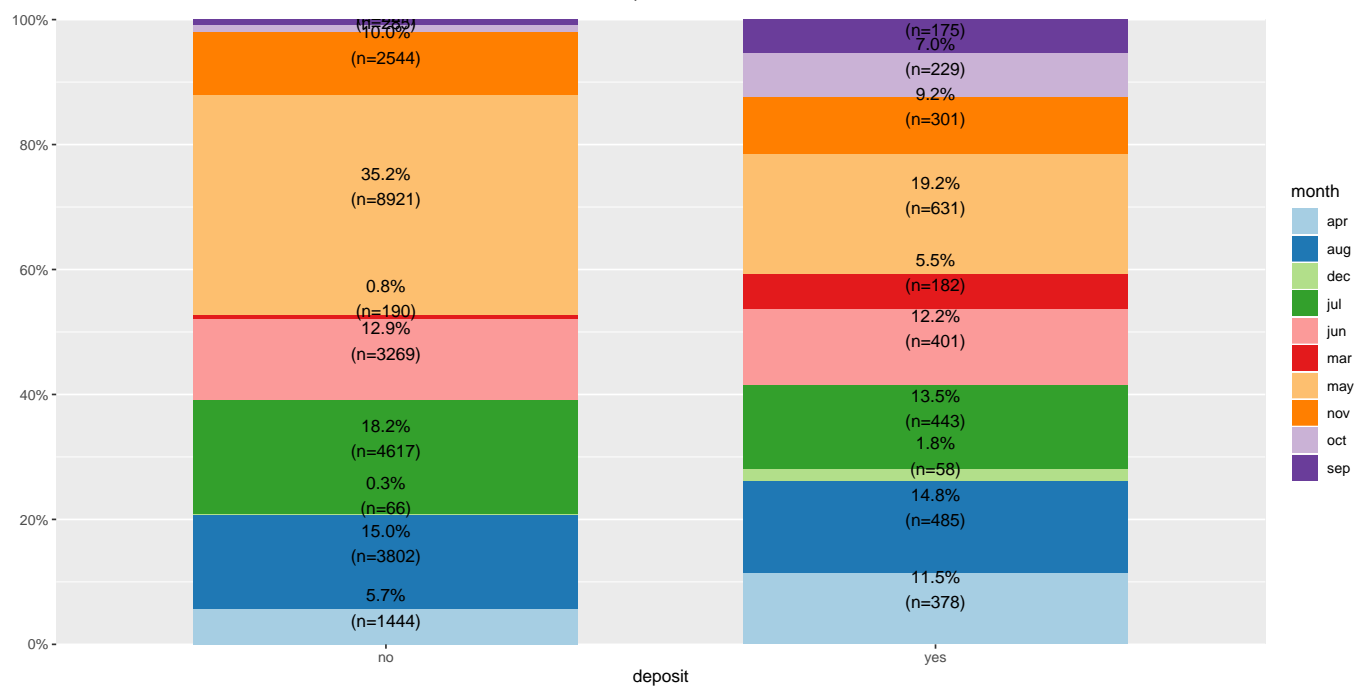
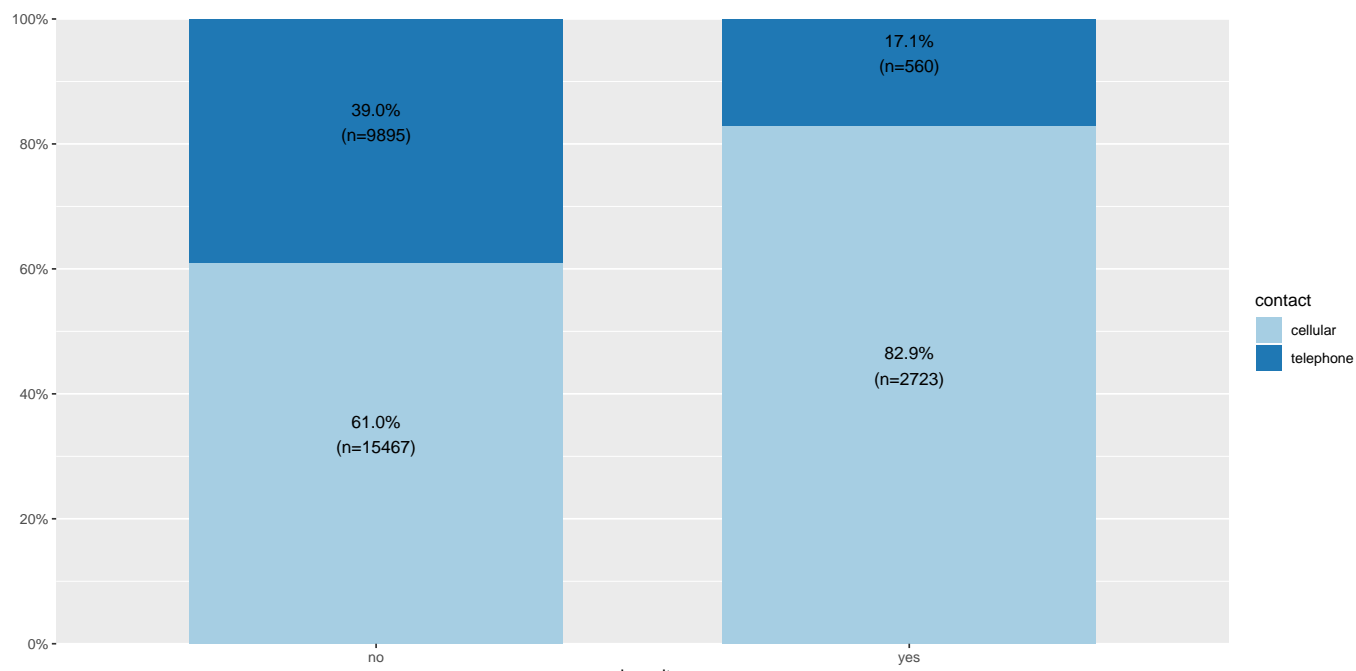
Also we can do a general plot to have a first idea of which variables help us to distinguish between the two groups. As we can see below, it looks like *housing*, *contact* and *potcome* distinguish between the two groups.

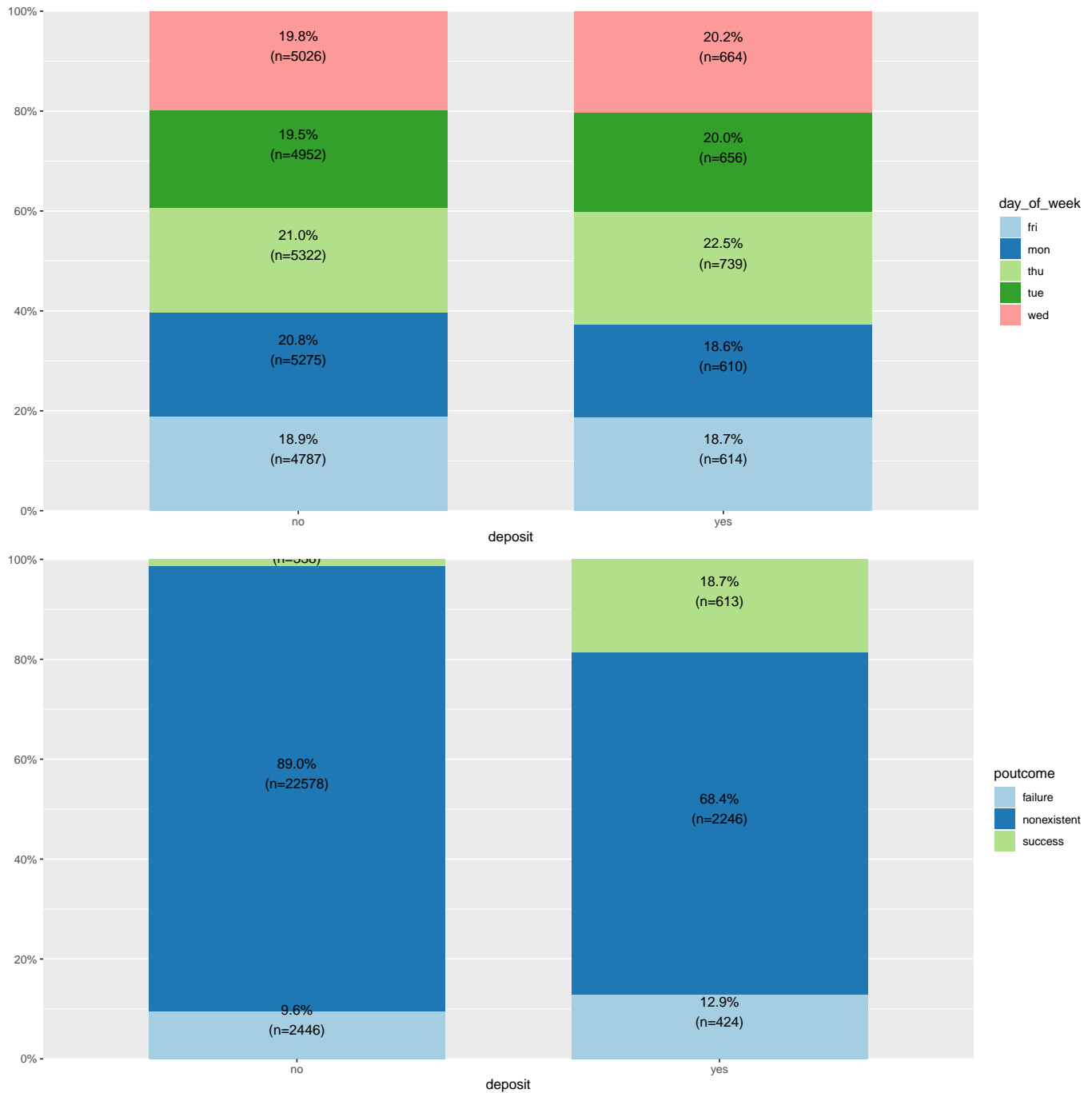












```
## Standard deviations (1, ..., p=10):
## [1] 1.9716147 1.1644768 1.0395572 1.0002886 0.9875492 0.9207305 0.6557383
## [8] 0.6222420 0.1573153 0.1027072
##
## Rotation (n x k) = (10 x 10):
##           PC1      PC2      PC3      PC4
## id      0.004506029  0.021154154 -0.124687451 -0.95127099
## age     -0.005616099  0.255664435 -0.638150224  0.20467279
## campaign 0.097175573 -0.004735733  0.291777420  0.21728167
## pdays   0.227640482 -0.624931366 -0.266334073  0.02798564
## previous -0.304987637  0.477274000  0.285864152 -0.03501187
## emp.var.rate 0.488883765 0.161364701 0.092138090 -0.02069430
## cons.price.idx 0.369176132 0.263569587 0.275626551 -0.03837764
## cons.conf.idx 0.093037098 0.443915834 -0.504008914  0.04172509
```


## euribor3m	0.490591266	0.152432006	0.004528449	-0.01384333
## nr.employed	0.470141863	-0.005639947	0.030433559	-0.01176826
##	PC5	PC6	PC7	PC8
## id	0.28102826	-0.005800433	0.0044427073	0.00613086
## age	0.37762520	-0.584546998	0.0304187623	0.01579130
## campaign	0.87048398	0.316617553	0.0008463865	-0.01077610
## pdays	0.03836421	0.019926332	-0.2346932432	-0.65495808
## previous	-0.01177674	-0.131843142	0.1640744053	-0.74264287
## emp.var.rate	-0.05225892	-0.069998822	0.0647781843	-0.04896583
## cons.price.idx	-0.02638851	-0.258379430	-0.7353312139	0.01701850
## cons.conf.idx	-0.09765905	0.682957811	-0.1776045028	-0.10557465
## euribor3m	-0.06728135	0.015892279	0.2126984088	-0.05925605
## nr.employed	-0.03972467	-0.043601770	0.5434710076	-0.04199807
##	PC9	PC10		
## id	0.0007902179	-0.0005189965		
## age	0.0021766183	0.0014533275		
## campaign	0.0006834990	-0.0090969866		
## pdays	0.0009043996	0.0016957086		
## previous	-0.0164190922	0.0042647868		
## emp.var.rate	0.7937135179	0.2860604343		
## cons.price.idx	-0.3135366270	0.1001703270		
## cons.conf.idx	-0.0702127884	0.1221842838		
## euribor3m	-0.0612042026	-0.8237689946		
## nr.employed	-0.5126027810	0.4631466045		

