

**Name:** Irem Tanrıverdi

**Email:** tanriverdiirem.95@gmail.com

**Country:** Turkey

**College:** Middle East Technical. University

**Specialization:** Data Science

## **1. Problem Description**

ABC Bank wants to sell its term deposit product to customers and before launching the product they want to develop a model which help them in understanding whether a particular customer will buy their product or not (based on customer's past interaction with bank or other Financial Institution).

## 2. Data understanding

First and last 10 variables in the data

	age	job	marital	education	default	balance	housing	loan	contact	day	month	duration	campaign	pdays	previous	poutcome	y
1	58	management	married	tertiary	no	2143	yes	no	unknown	5	may	261	1	-1	0	unknown	no
2	44	technician	single	secondary	no	29	yes	no	unknown	5	may	151	1	-1	0	unknown	no
3	33	entrepreneur	married	secondary	no	2	yes	yes	unknown	5	may	76	1	-1	0	unknown	no
4	47	blue-collar	married	unknown	no	1506	yes	no	unknown	5	may	92	1	-1	0	unknown	no
5	33	unknown	single	unknown	no	1	no	no	unknown	5	may	198	1	-1	0	unknown	no
6	35	management	married	tertiary	no	231	yes	no	unknown	5	may	139	1	-1	0	unknown	no
7	28	management	single	tertiary	no	447	yes	yes	unknown	5	may	217	1	-1	0	unknown	no
8	42	entrepreneur	divorced	tertiary	yes	2	yes	no	unknown	5	may	380	1	-1	0	unknown	no
9	58	retired	married	primary	no	121	yes	no	unknown	5	may	50	1	-1	0	unknown	no
10	43	technician	single	secondary	no	593	yes	no	unknown	5	may	55	1	-1	0	unknown	no
...	...	NA	NA	NA	NA	...	NA	NA	NA	...	NA	...	...	...	...	NA	NA
45202	53	management	married	tertiary	no	583	no	no	cellular	17	nov	226	1	184	4	success	yes
45203	34	admin.	single	secondary	no	557	no	no	cellular	17	nov	224	1	-1	0	unknown	yes
45204	23	student	single	tertiary	no	113	no	no	cellular	17	nov	266	1	-1	0	unknown	yes
45205	73	retired	married	secondary	no	2850	no	no	cellular	17	nov	300	1	40	8	failure	yes
45206	25	technician	single	secondary	no	505	no	yes	cellular	17	nov	386	2	-1	0	unknown	yes
45207	51	technician	married	tertiary	no	825	no	no	cellular	17	nov	977	3	-1	0	unknown	yes
45208	71	retired	divorced	primary	no	1729	no	no	cellular	17	nov	456	2	-1	0	unknown	yes
45209	72	retired	married	secondary	no	5715	no	no	cellular	17	nov	1127	5	184	3	success	yes
45210	57	blue-collar	married	secondary	no	668	no	no	telephone	17	nov	508	4	-1	0	unknown	no
45211	37	entrepreneur	married	secondary	no	2971	no	no	cellular	17	nov	361	2	188	11	other	no

age		job		marital		education	
Min.	:18.00	blue-collar:	9732	divorced:	5207	primary	: 6851
1st Qu.:	33.00	management	:9458	married	:27214	secondary:	23202
Median	:39.00	technician	:7597	single	:12790	tertiary	:13301
Mean	:40.94	admin.	:5171			unknown	: 1857
3rd Qu.:	48.00	services	:4154				
Max.	:95.00	retired	:2264				
		(Other)	:6835				
default		balance		housing		loan	
no	:44396	Min.	: -8019	no	:20081	no	:37967
yes:	815	1st Qu.:	72	yes:	25130	yes:	7244
		Median	: 448				
		Mean	: 1362				
		3rd Qu.:	1428				
		Max.	:102127				
						contact	
						cellular	:29285
						telephone:	2906
						unknown	:13020

day		month		duration		campaign	
Min.	: 1.00	Length:	45211	Min.	: 0.0	Min.	: 1.000
1st Qu.:	8.00	Class	:character	1st Qu.:	103.0	1st Qu.:	1.000
Median	:16.00	Mode	:character	Median	: 180.0	Median	: 2.000
Mean	:15.81			Mean	: 258.2	Mean	: 2.764
3rd Qu.:	21.00			3rd Qu.:	319.0	3rd Qu.:	3.000
Max.	:31.00			Max.	:4918.0	Max.	:63.000

pdays		previous		poutcome		y	
Min.	: -1.0	Min.	: 0.0000	failure:	4901	Length:	45211
1st Qu.:	-1.0	1st Qu.:	0.0000	other	: 1840	Class	:character
Median	: -1.0	Median	: 0.0000	success:	1511	Mode	:character
Mean	: 40.2	Mean	: 0.5803	unknown:	36959		
3rd Qu.:	-1.0	3rd Qu.:	0.0000				
Max.	:871.0	Max.	:275.0000				

```

'data.frame':  45211 obs. of  17 variables:
 $ age      : int  58 44 33 47 33 35 28 42 58 43 ...
 $ job      : Factor w/ 12 levels "admin.", "blue-collar",...: 5 10 3 2 12 5 5 3 6 10 ...
 $ marital  : Factor w/ 3 levels "divorced", "married",...: 2 3 2 2 3 2 3 1 2 3 ...
 $ education: Factor w/ 4 levels "primary", "secondary",...: 3 2 2 4 4 3 3 3 1 2 ...
 $ default  : Factor w/ 2 levels "no", "yes": 1 1 1 1 1 1 1 2 1 1 ...
 $ balance  : int  2143 29 2 1506 1 231 447 2 121 593 ...
 $ housing  : Factor w/ 2 levels "no", "yes": 2 2 2 2 1 2 2 2 2 2 ...
 $ loan     : Factor w/ 2 levels "no", "yes": 1 1 2 1 1 1 2 1 1 1 ...
 $ contact  : Factor w/ 3 levels "cellular", "telephone",...: 3 3 3 3 3 3 3 3 3 3 ...
 $ 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 : Factor w/ 4 levels "failure", "other",...: 4 4 4 4 4 4 4 4 4 4 ...
 $ y        : chr   "no" "no" "no" "no" ...

```

- Bank dataset includes 45211 observations and 17 variables.
- There are 7 numeric variables which are age, balance, day, duration, campaign, pdays, and previous.
- There are 10 categorical variables which are job, marital, education, default, housing, loan, contact, month, poutcome and y.
  1. age (numeric)
  2. job : type of job (categorical: 'admin.', 'blue collar', 'entrepreneur', 'housemaid', 'management', 'retired', 'self-employed', 'services', 'student', 'technician', 'unemployed', 'unknown')
  3. marital : marital status (categorical: 'divorced', 'married', 'single', 'unknown'; note: 'divorced' means divorced or widowed)
  4. education (categorical: 'basic.4y', 'basic.6y', 'basic.9y', 'high.school', 'illiterate', 'professional.course', 'university. Degree', 'unknown')
  5. default: has credit in default? (categorical: 'no', 'yes', 'unknown')
  6. housing: has housing loan? (categorical: 'no', 'yes', 'unknown')
  7. loan: has personal loan? (categorical: 'no', 'yes', 'unknown') related with the last contact of the current campaign:
  8. contact: contact communication type (categorical: 'cellular', 'telephone')
  9. month: last contact month of year (categorical: 'jan', 'feb', 'mar', ..., 'nov', 'dec')
  10. day\_of\_week: last contact day of the week (categorical: 'mon', 'tue', 'wed', 'thu', 'fri')
  11. duration: last contact duration, in seconds (numeric)
  12. campaign: number of contacts performed during this campaign and for this client (numeric, includes last contact)
  13. pdays: number of days that passed by after the client was last contacted from a previous campaign (numeric; 999 means client was not previously contacted)
  14. previous: number of contacts performed before this campaign and for this client (numeric)

15. poutcome: outcome of the previous marketing campaign (categorical: 'failure', 'nonexistent', 'success')

Output variable (desired target):

17. y: has the client subscribed a term deposit? (Binary: 'yes', 'no')

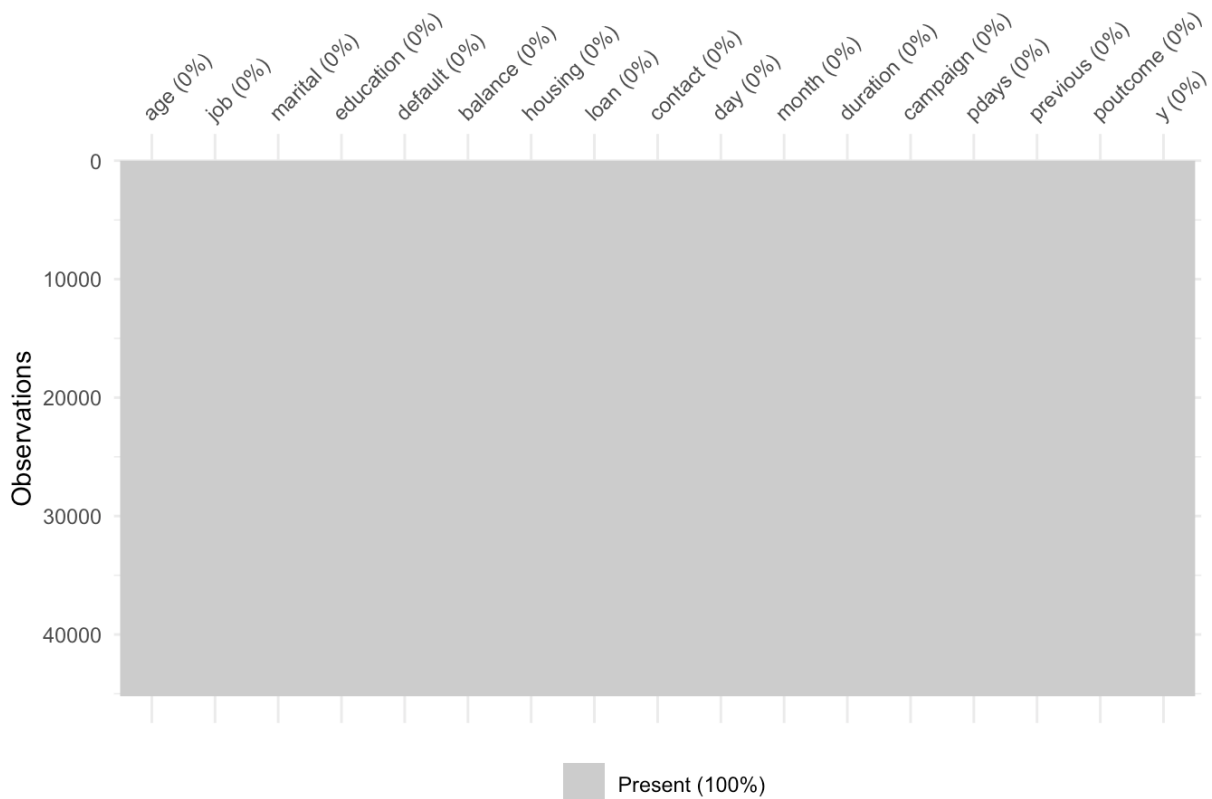
## Exploratory data Analysis

### Is there duplicated rows in the data?

```
# A tibble: 1 x 1
      n
  <int>
1     0
```

- As seen there is not duplicated rows in the data.

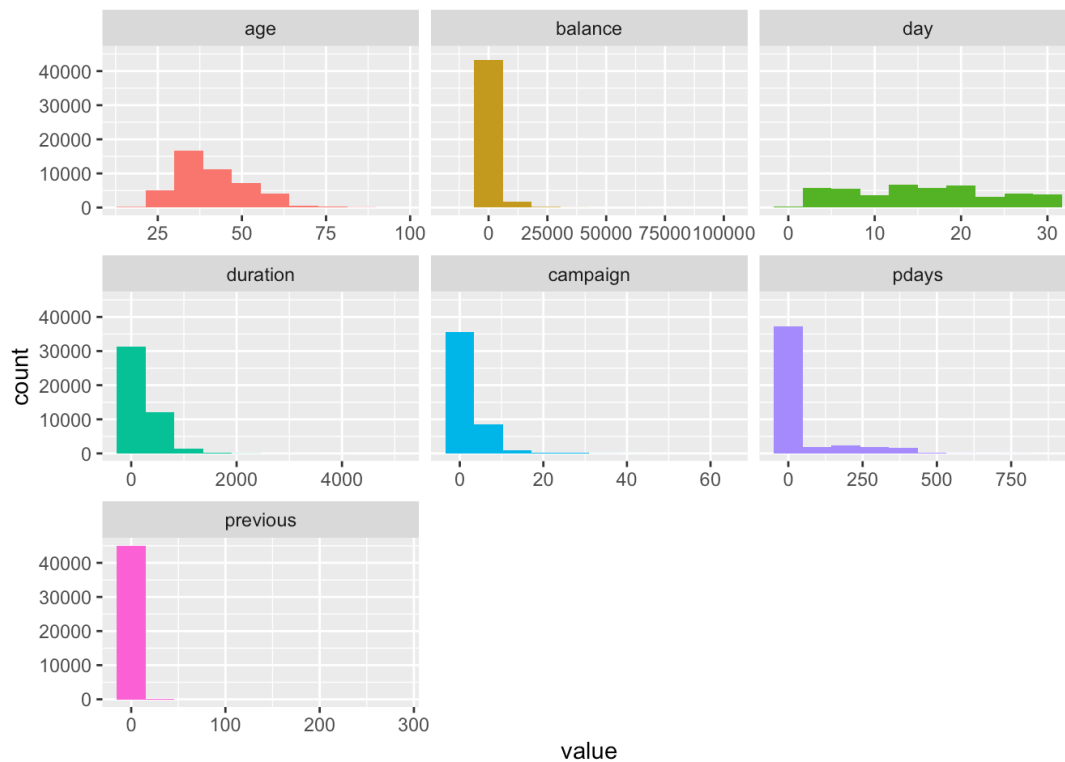
### Is there any missing value in the data?

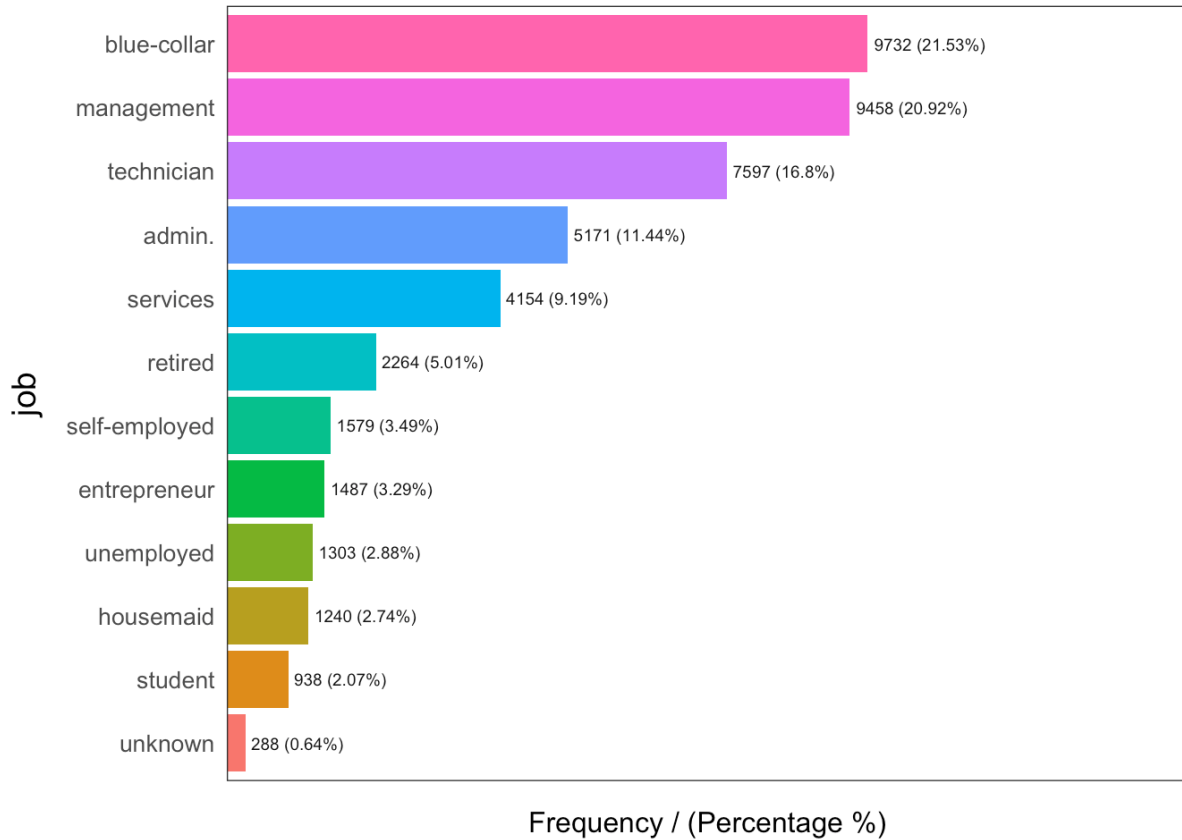


- As seen there is no missing value in the data.

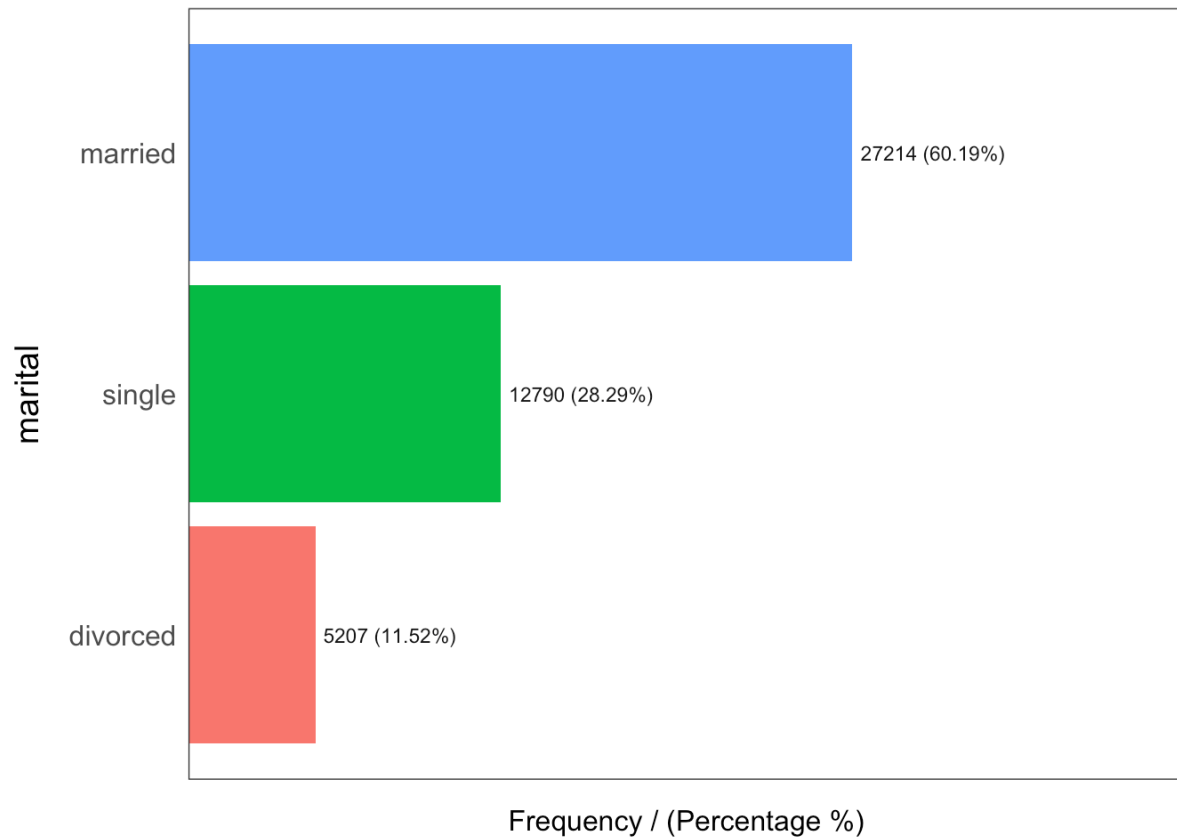
## Frequencies of the categorical variables and distributions of the numeric variables

	variable	mean	std_dev	variation_coef	p_01	p_05	p_25	p_50	p_75
1	age	40.9362102	10.618762	0.2593978	23	27	33	39	48
2	balance	1362.2720577	3044.765829	2.2350644	-627	-172	72	448	1428
3	day	15.8064188	8.322476	0.5265251	2	3	8	16	21
4	duration	258.1630798	257.527812	0.9975393	11	35	103	180	319
5	campaign	2.7638407	3.098021	1.1209115	1	1	1	2	3
6	pdays	40.1978280	100.128746	2.4908994	-1	-1	-1	-1	-1
7	previous	0.5803234	2.303441	3.9692371	0	0	0	0	0
	p_95	p_99	skewness	kurtosis	iqr	range_98	range_80		
1	59	71.0	0.68479520	3.319402	15	[23, 71]	[29, 56]		
2	5768	13164.9	8.36003095	143.735848	1356	[-627, 13164.9]	[0, 3574]		
3	29	31.0	0.09307593	1.940087	13	[2, 31]	[5, 28]		
4	751	1269.0	3.14421378	21.151775	216	[11, 1269]	[58, 548]		
5	8	16.0	4.89848764	42.245178	2	[1, 16]	[1, 5]		
6	317	370.0	2.61562869	9.934296	0	[-1, 370]	[-1, 185]		
7	3	8.9	41.84506609	4509.362118	0	[0, 8.900000000000146]	[0, 2]		



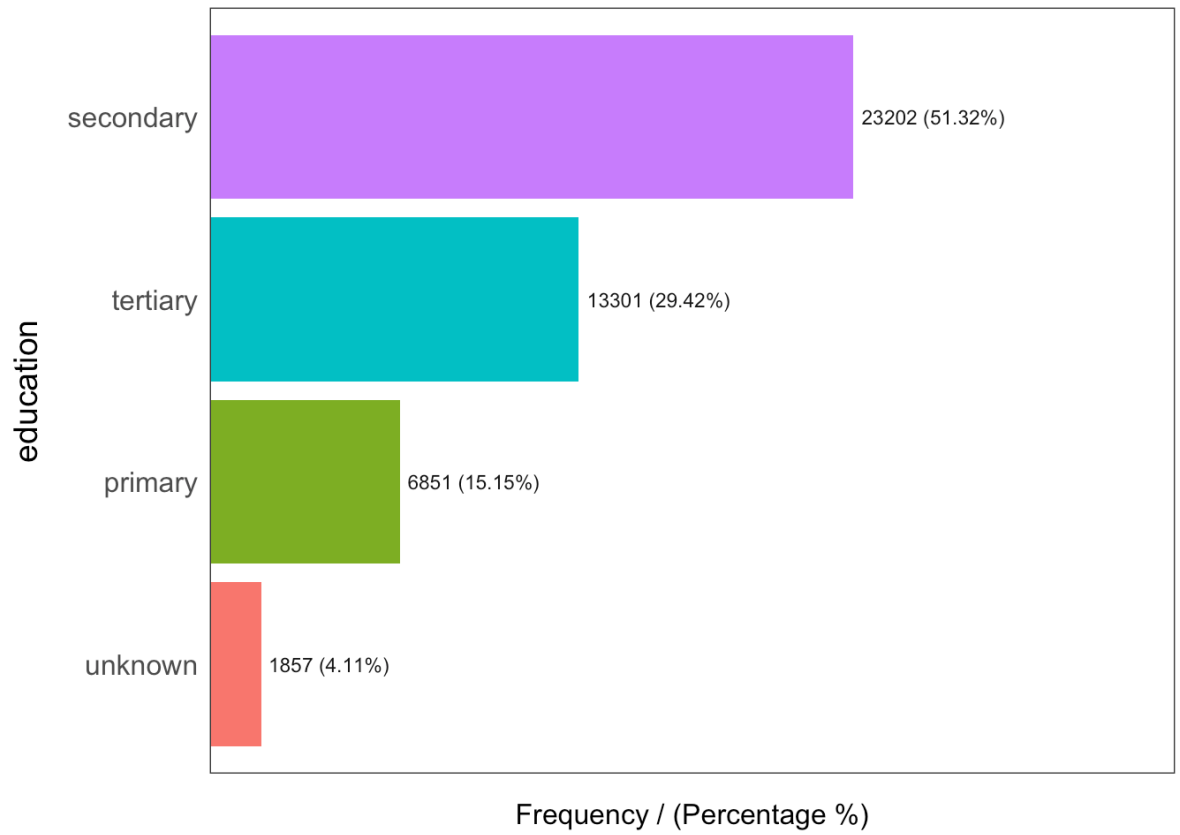


##	job	frequency	percentage	cumulative_perc
## 1	blue-collar	9732	21.53	21.53
## 2	management	9458	20.92	42.45
## 3	technician	7597	16.80	59.25
## 4	admin.	5171	11.44	70.69
## 5	services	4154	9.19	79.88
## 6	retired	2264	5.01	84.89
## 7	self-employed	1579	3.49	88.38
## 8	entrepreneur	1487	3.29	91.67
## 9	unemployed	1303	2.88	94.55
## 10	housemaid	1240	2.74	97.29
## 11	student	938	2.07	99.36
## 12	unknown	288	0.64	100.00

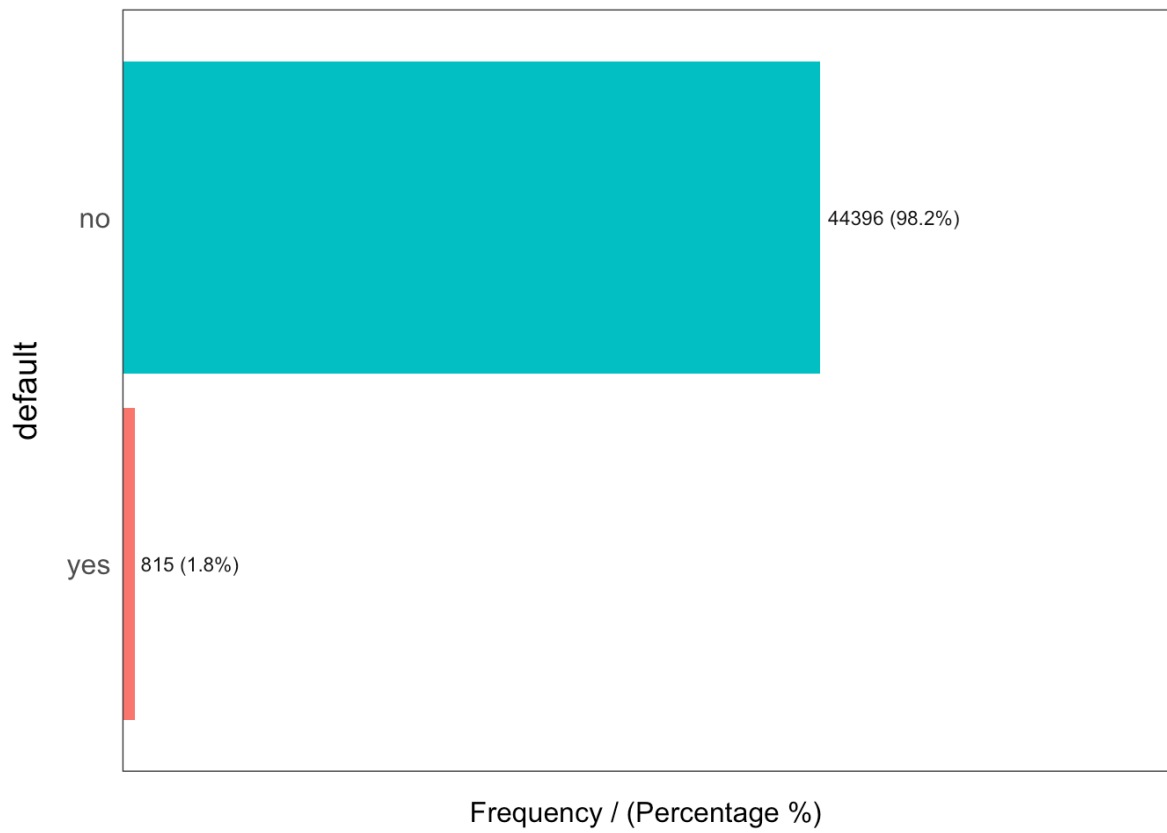


##	marital	frequency	percentage	cumulative_perc
## 1	married	27214	60.19	60.19
## 2	single	12790	28.29	88.48
## 3	divorced	5207	11.52	100.00

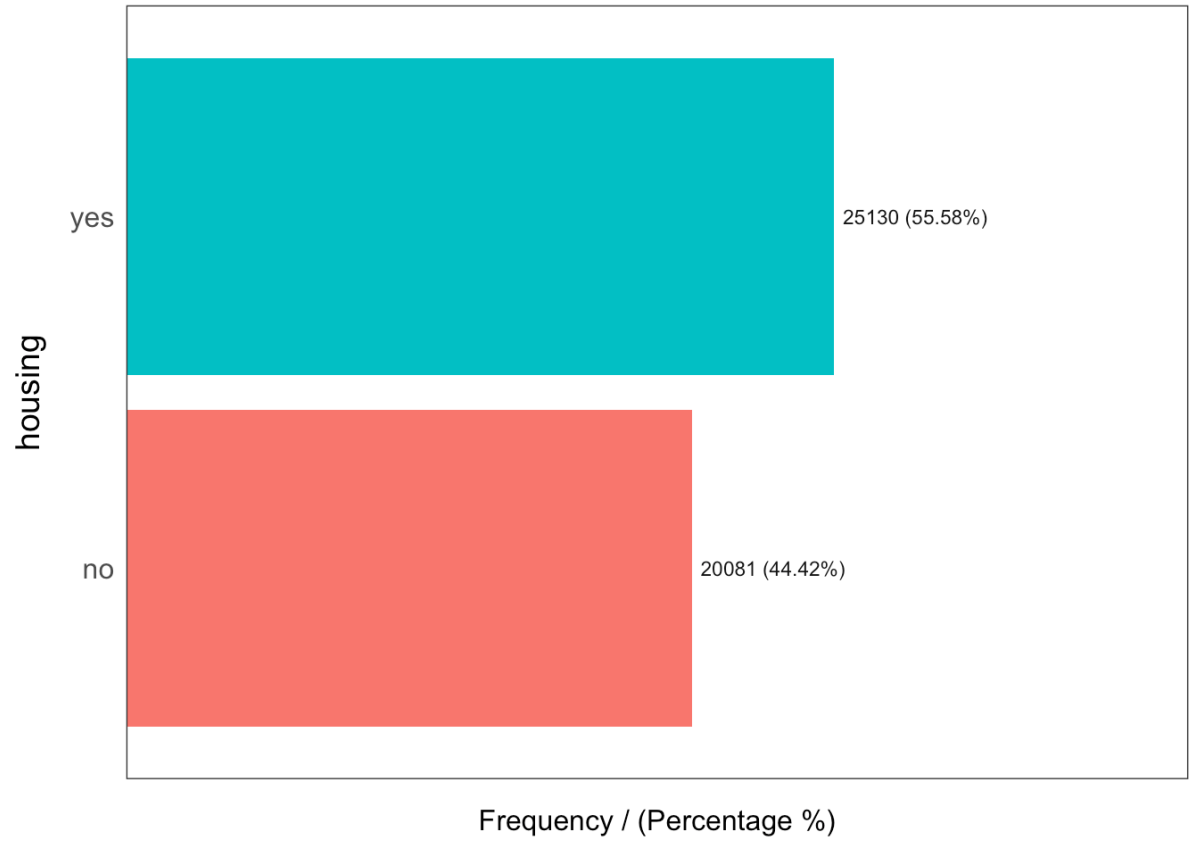




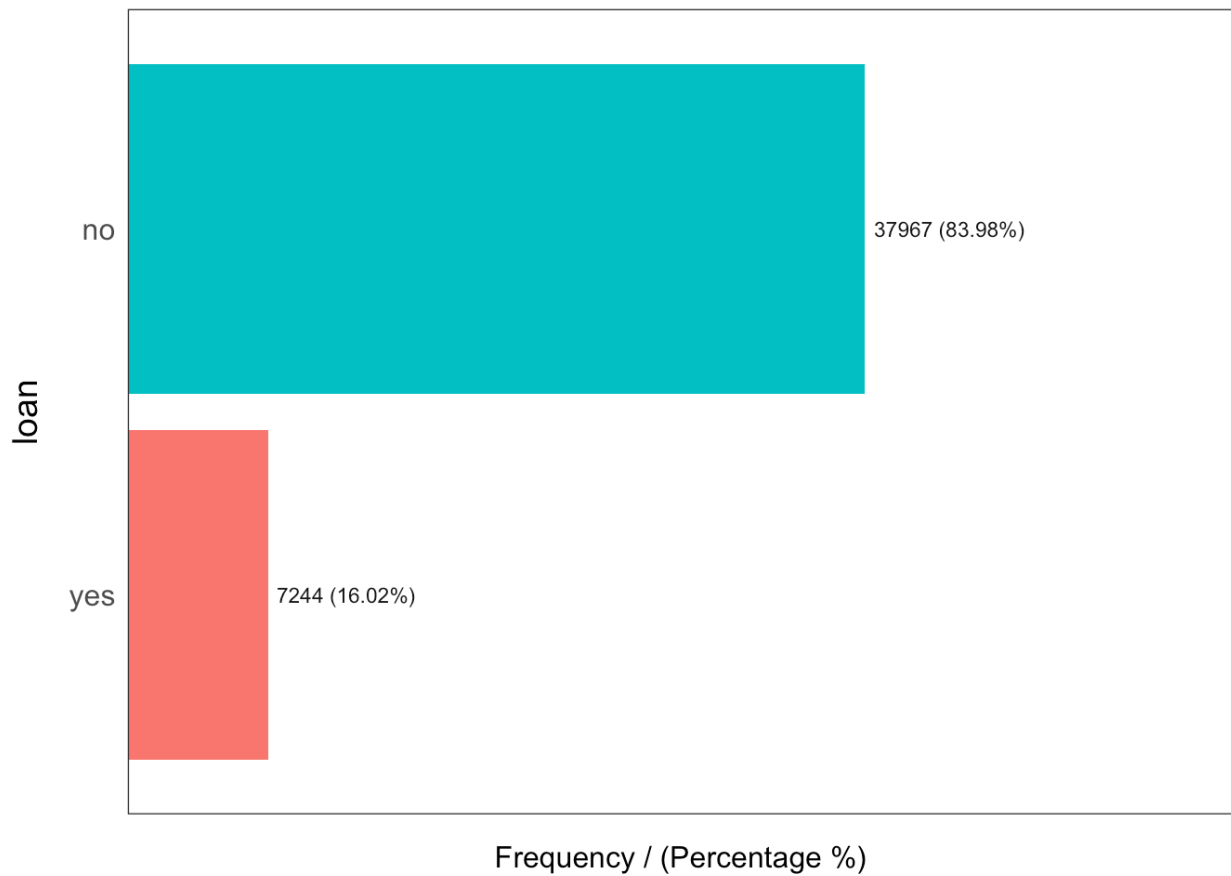
##	education	frequency	percentage	cumulative_perc
## 1	secondary	23202	51.32	51.32
## 2	tertiary	13301	29.42	80.74
## 3	primary	6851	15.15	95.89
## 4	unknown	1857	4.11	100.00



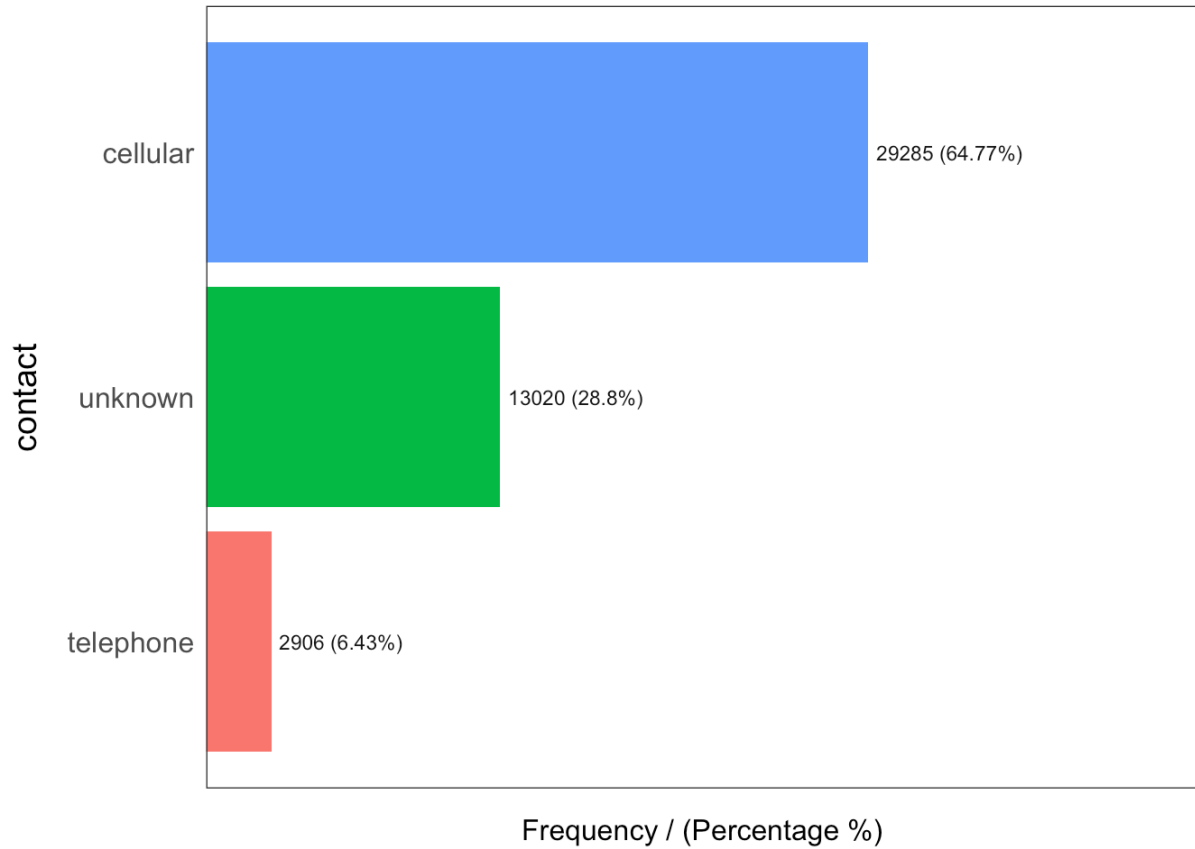
##	default	frequency	percentage	cumulative_perc
## 1	no	44396	98.2	98.2
## 2	yes	815	1.8	100.0



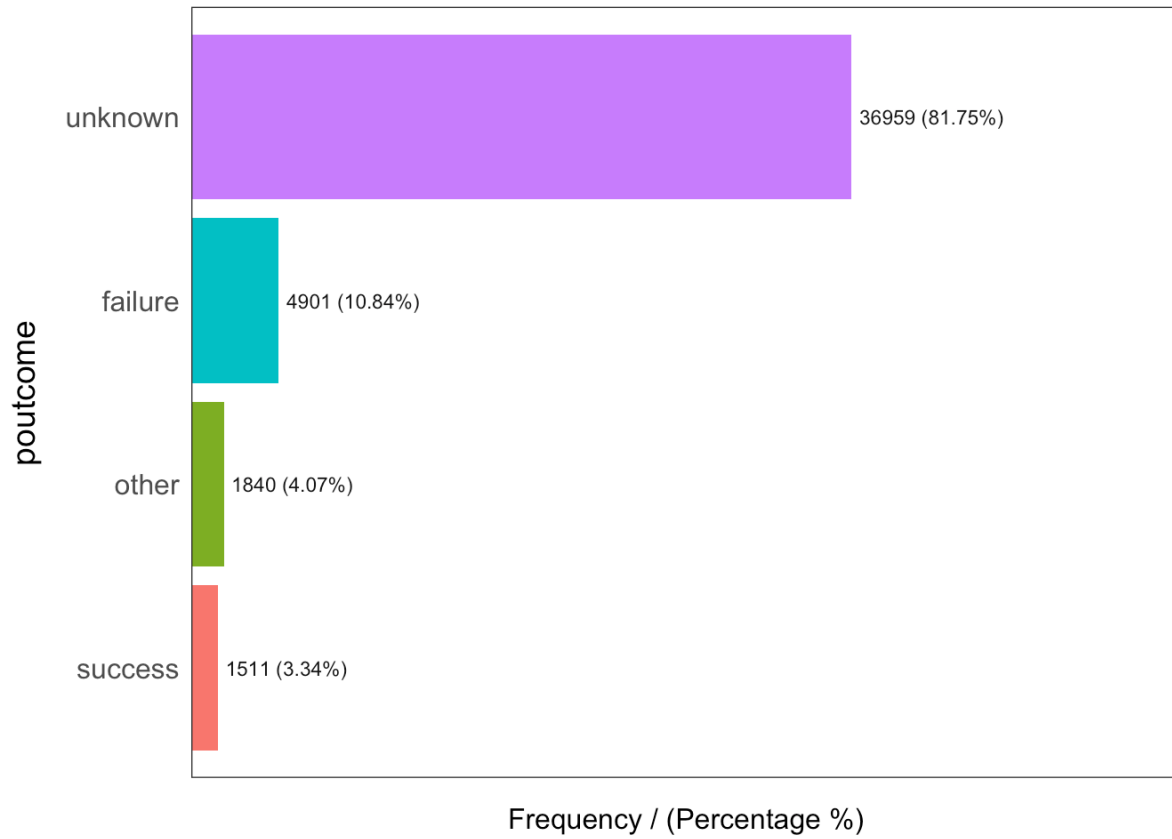
##	housing	frequency	percentage	cumulative_perc
## 1	yes	25130	55.58	55.58
## 2	no	20081	44.42	100.00



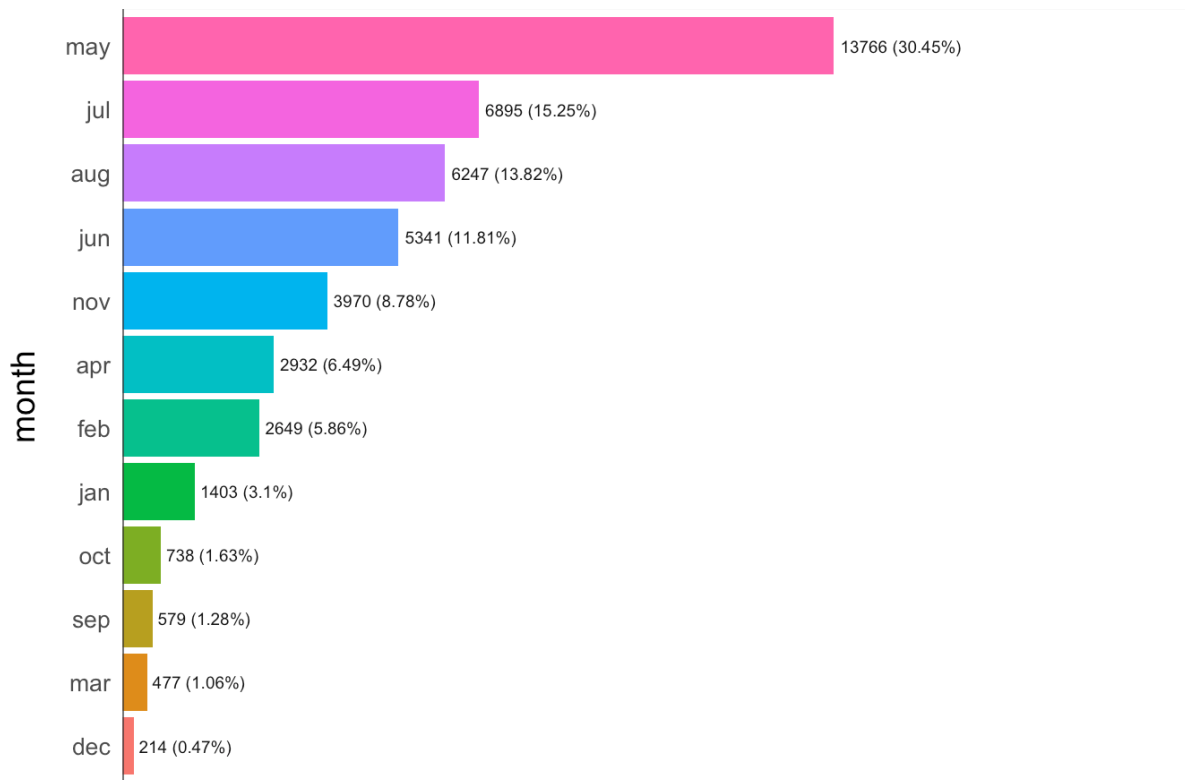
##	loan	frequency	percentage	cumulative_perc
## 1	no	37967	83.98	83.98
## 2	yes	7244	16.02	100.00



##	contact	frequency	percentage	cumulative_perc
## 1	cellular	29285	64.77	64.77
## 2	unknown	13020	28.80	93.57
## 3	telephone	2906	6.43	100.00

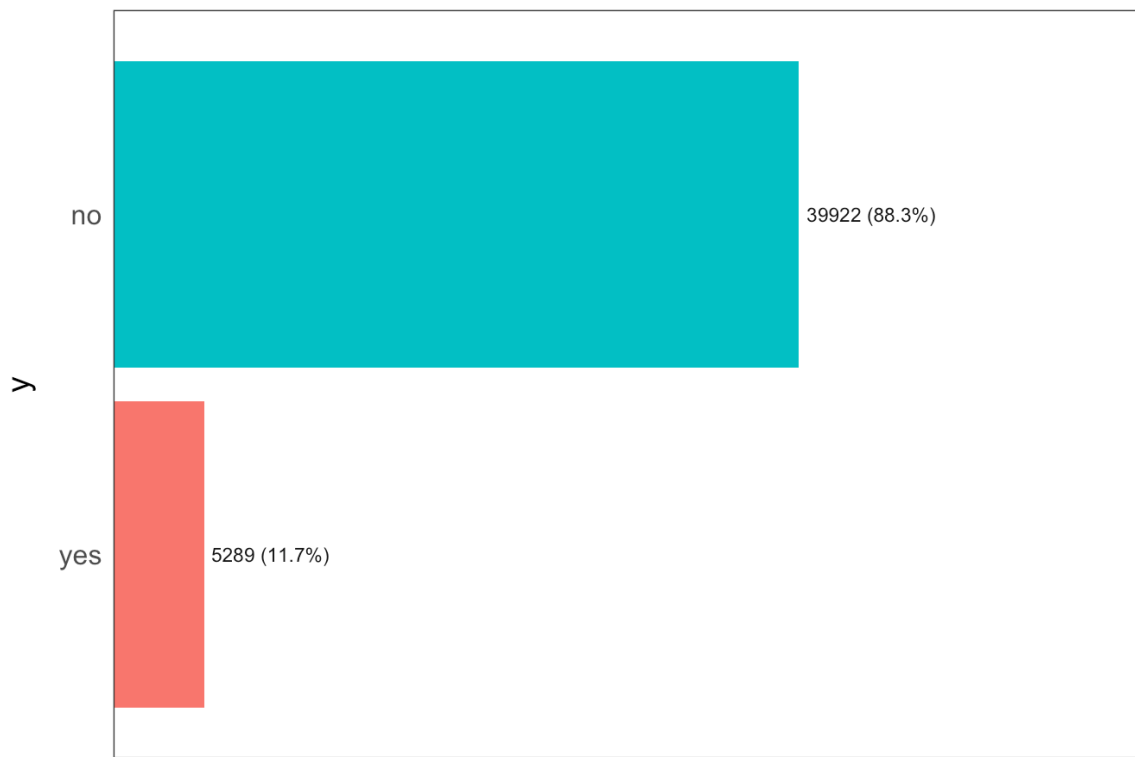


##	poutcome	frequency	percentage	cumulative_perc
## 1	unknown	36959	81.75	81.75
## 2	failure	4901	10.84	92.59
## 3	other	1840	4.07	96.66
## 4	success	1511	3.34	100.00



Frequency / (Percentage %)

##	month	frequency	percentage	cumulative_perc
## 1	may	13766	30.45	30.45
## 2	jul	6895	15.25	45.70
## 3	aug	6247	13.82	59.52
## 4	jun	5341	11.81	71.33
## 5	nov	3970	8.78	80.11
## 6	apr	2932	6.49	86.60
## 7	feb	2649	5.86	92.46
## 8	jan	1403	3.10	95.56
## 9	oct	738	1.63	97.19
## 10	sep	579	1.28	98.47
## 11	mar	477	1.06	99.53
## 12	dec	214	0.47	100.00

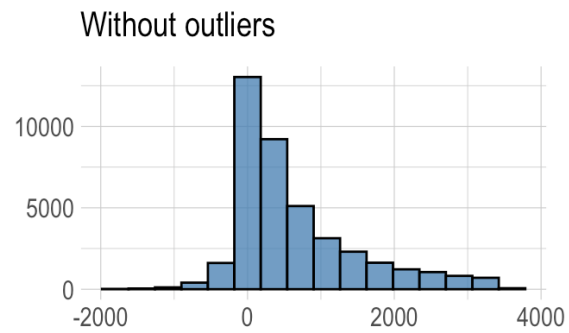
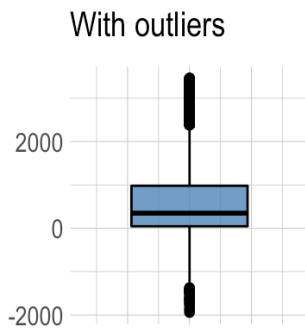
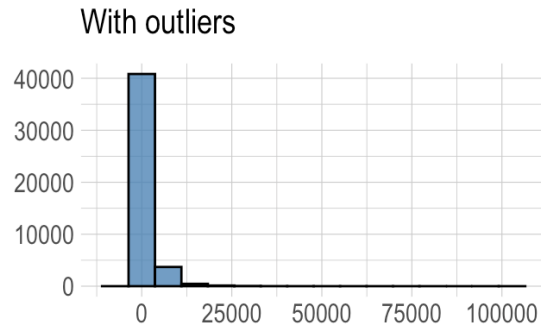
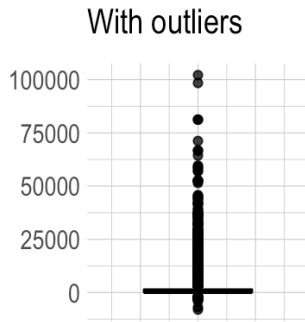


##	y	frequency	percentage	cumulative_perc
## 1	no	39922	88.3	88.3
## 2	yes	5289	11.7	100.0

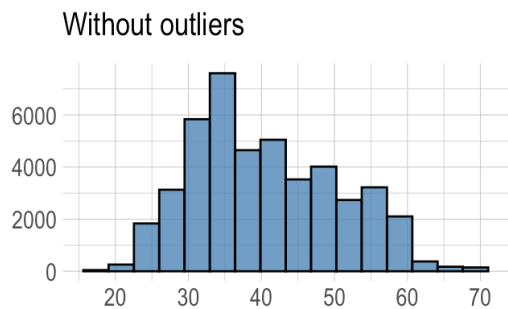
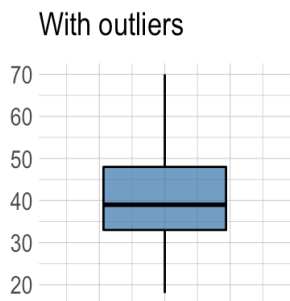
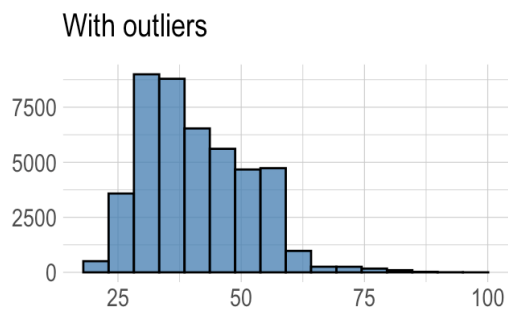
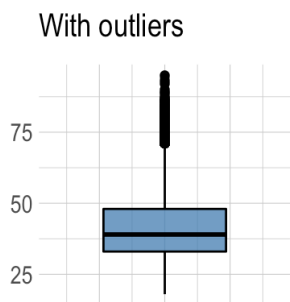
**Do numeric variables have any outlier and what will be shape of the variables exclude the outliers?**



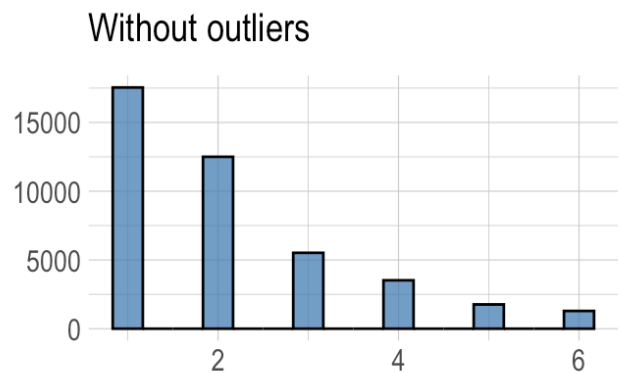
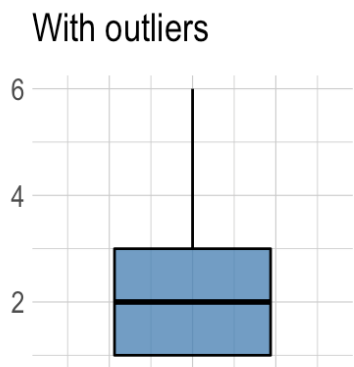
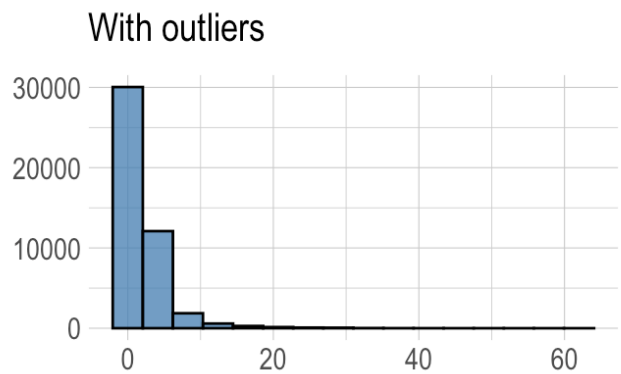
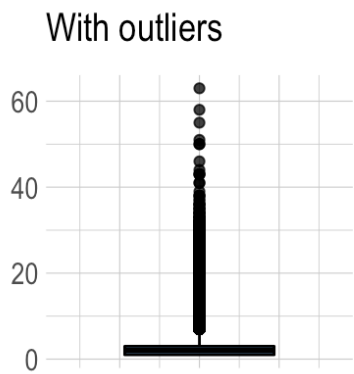
## Outlier Diagnosis Plot (balance)



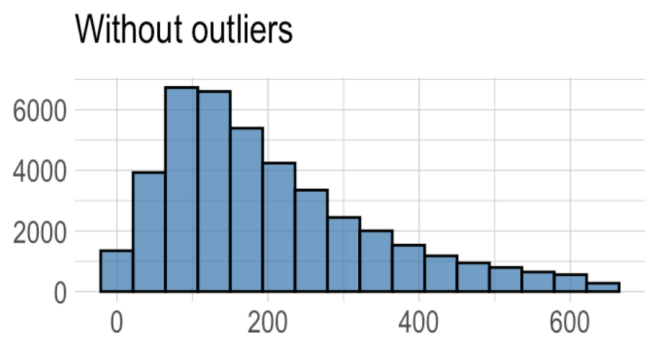
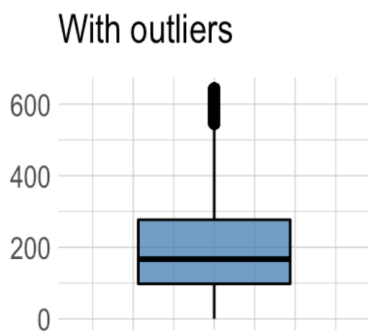
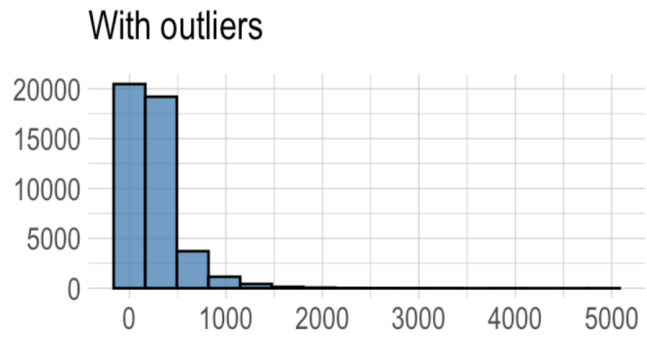
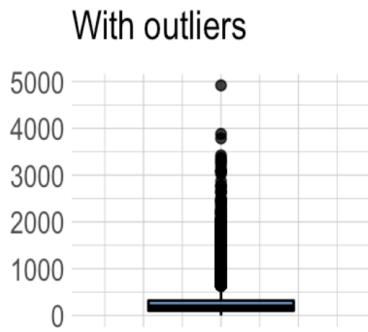
## Outlier Diagnosis Plot (age)



# Outlier Diagnosis Plot (campaign)

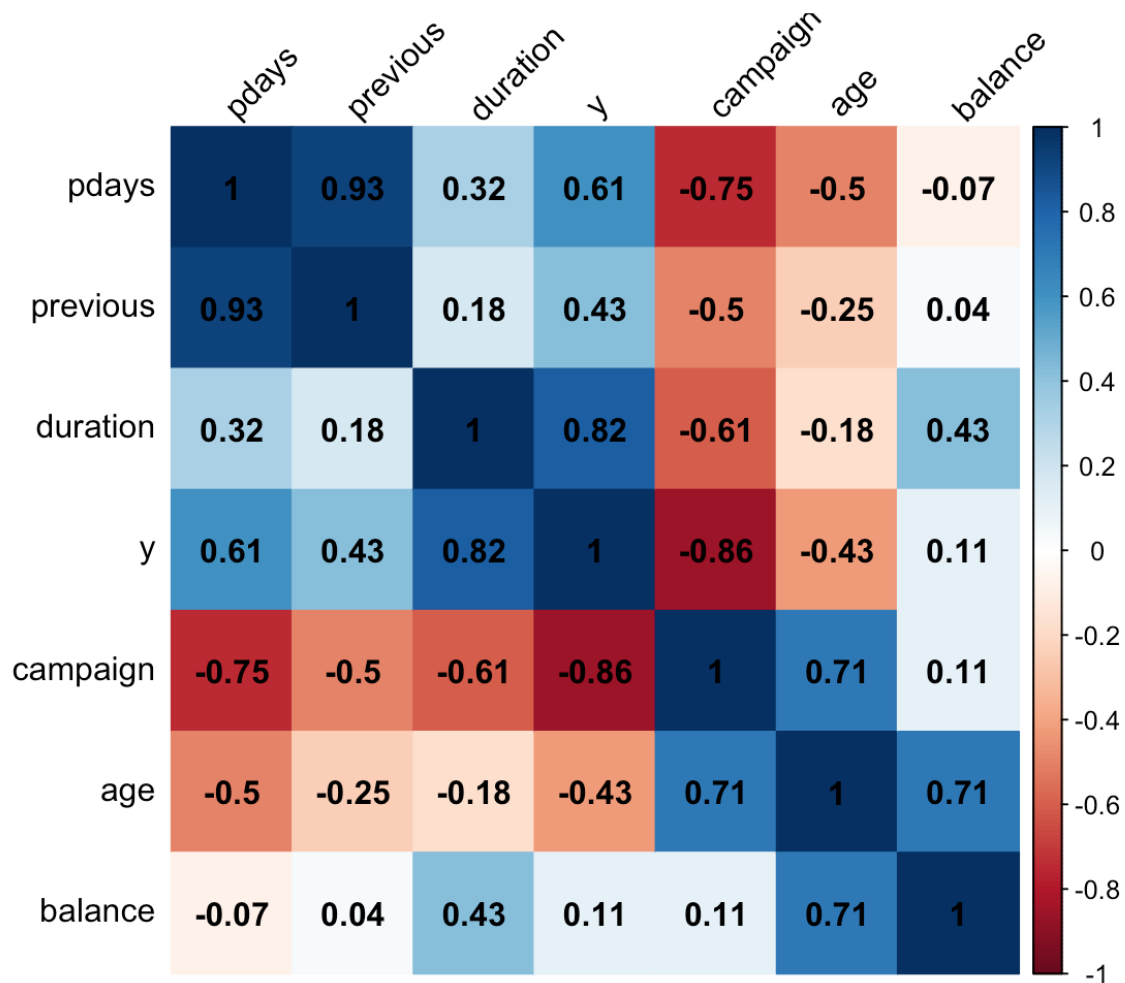


## Outlier Diagnosis Plot (duration)



- As seen all of the 4 numeric variables have outlier.
- Shapes of the variables changed when outliers removed.

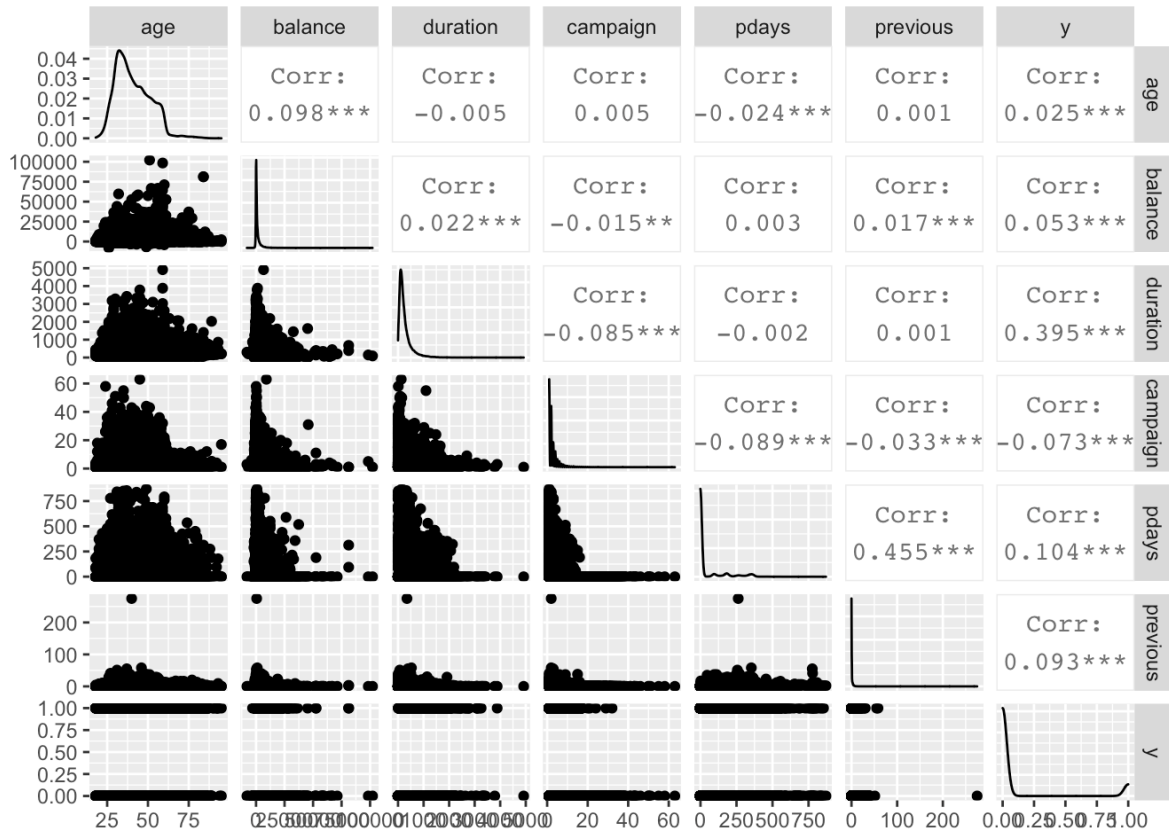
Is there any significant relationship between numeric variables and y, if y taken as numeric (1:no, 0: yes)?



Positive correlations are shown in blue and negative correlations in red color. Color intensity is proportional to the correlation coefficients. Let's look at the correlation matrix to examine which variables have strong relationship with response variable y.

- Between y and duration, there is strong positive relationship.
- Between y and campaign, there is strong negative relationship.

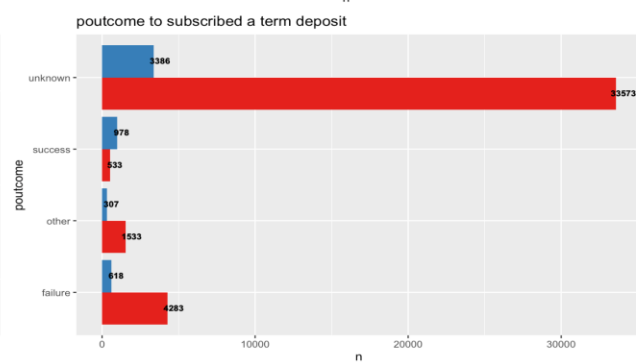
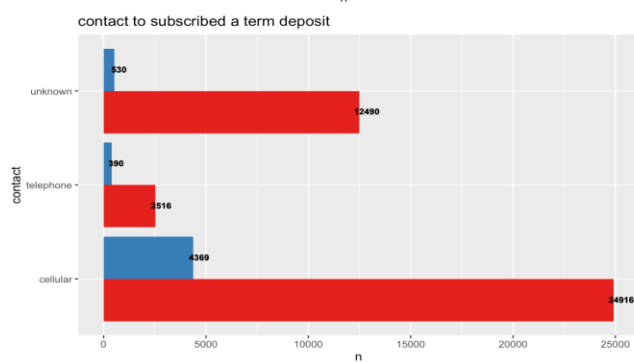
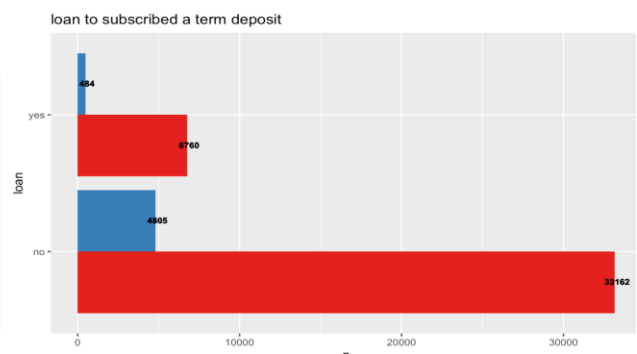
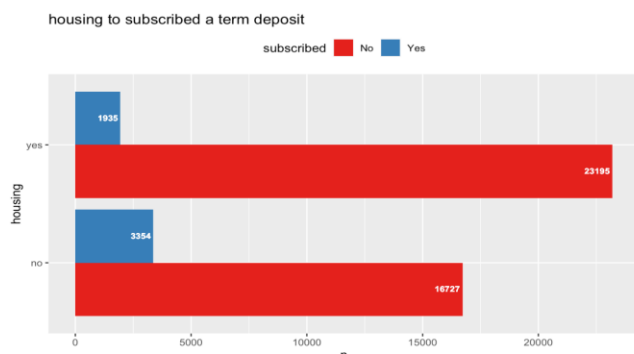
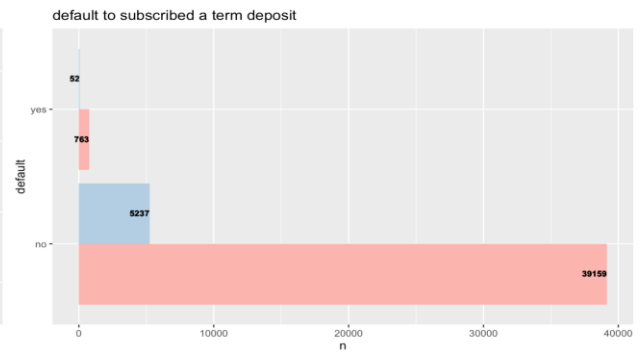
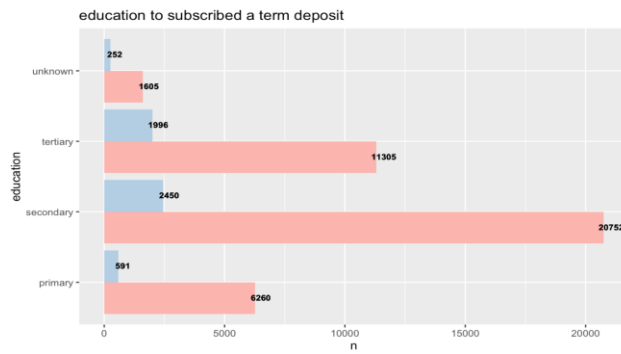
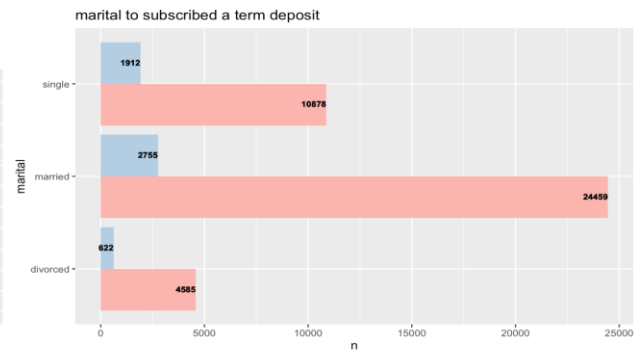
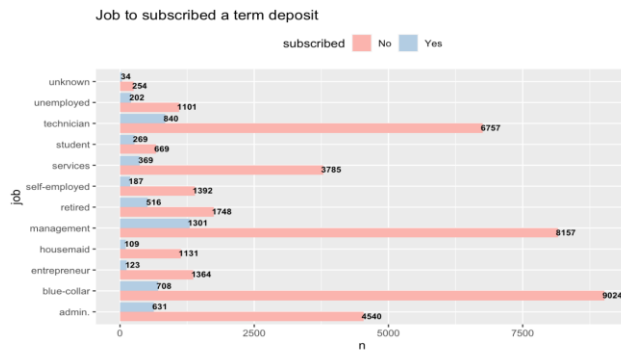
We can also see the relationship between other variables (covariates).



We can see from this plot we can see that if the relationship between variables significant or not.  
 We can see that all the relationship between covariates and y are significant.

**Is there any significant relationship between categorical variables and y?**

$H_0$ : There is not significant relationship between variables (Variables are independent)



	statistic	p.value
job	836.1055	0.000499
marital	196.4959	0.000499
education	238.9235	0.000499
default	22.7235	0.000499
housing	875.6937	0.000499
loan	210.1949	0.000499
contact	1035.714	0.000499
poutcome	4391.507	0.000499

- As seen all p-values are smaller than the significance level of 0.05, so there is significant relationship between categorical variables and y.

### Recommended models for this data set

We see that in EDA part, in response variable, “no” class proportion is 88.3 while “yes” class proportion is 11.7. There is huge difference between two class. Thus, we have imbalance data, and it causes reduction in accuracy of ML algorithms.

### What are the methods to deal with imbalanced data sets?

The methods are widely known as ‘Sampling Methods’. Generally, these methods aim to modify an imbalanced data into balanced distribution using some mechanism. The modification occurs by altering the size of original data set and provide the same proportion of balance.

Below are the methods used to treat imbalanced datasets:

- Undersampling
- Oversampling
- Synthetic Data Generation
- Cost Sensitive Learning

I applied both under sampling and oversampling since you we've lost significant information from the sample when doing undersampling.

After applying one of the methods for imbalanced data; we conduct three ML models which are very popular for classification data.

1. Logistic regression
2. Decision Tree
3. XGBoots