Visualisation: Project

Name: Adarsha Mondal | Roll Number: MDS202205

Introduction

A Portuguese banking institution's direct marketing campaigns are covered by the dataset. The marketing campaigns were based on phone conversations. Often, it was necessary to make multiple frequent contacts with the same client in order to determine whether the product (in this case, bank term deposit) would be subscribed ('yes') or not ('no').

Our primary goal is to formulate descriptive statistics and visualization with the graphical tools avilable in R language(mainly, **ggplot2**).

Dataset Information

Our data contains 41188 observations and 21 attributes, where 11 are categorical (within them 4 are binary variables) and 10 are numerical attributes. The table below contains a portion of the analysed data and names of all the attributes, along with their types and values that they contains.

Fig 2: Attribute Description

	Attributes	Kind	Attribute illustration, description	Values of attributes
Bank Client data	age	numeric	age of client	values between 17 and 98
	job	categorical	type of job	management', 'technician', 'entrepreneur', 'blue-collar', 'unknown', 'retired', 'admin.', 'services', 'self-employed', 'unemployed', 'housemaid', 'student'
	marital	categorical	marital status, note: 'divorced' means divorced or widowed	divorced', 'married', 'single', 'unknown'
	education	categorical	degree of education	basic.4y', 'high.school', 'basic.6y', 'basic.9y', 'university.degree', 'illiterate', 'professional.course', 'unknown'
	default	binary	has credit in default?	no', 'yes', 'unknown'
	housing	binary	has housing loan?	no', 'yes', 'unknown'
	loan	binary	has personal loan?	no', 'yes', 'unknown'
Related to last Contact of the current Campaign	contact	categorical	contact communication type	cellular', 'telephone'
	month	categorical	last contact month of year	mar', 'apr','may', 'jun', 'jul', 'aug','sep', 'oct', 'nov', 'dec'
	day_of_week	categorical	day in week	mon', 'tue', 'wed', 'thu', 'fri'
	duration	numeric	last contact duration, in seconds	values b/w 0 and 4918
Other attributes	campaign	numeric	number of contacts performed during this campaign and for this client (included last contact)	values b/w 1 and 56
	pdays	numeric	number of days that passed by after the client was last contacted from a previous campaign, note: 999 means client was not previously contacted	values b/w 0 and 27
	previous	numeric	number of contacts performed before this campaign and for this client	values b/w 0 and 7
	poutcome	categorical	outcome of the previous marketing campaign	failure', 'nonexistent', 'success'
Social and Economic context atributes	emp.var.rate	numeric	employment variation rate --- quarterly indicator	values b/w -3.4 and 1.4
	cons.price.idx	numeric	consumer price index --- monthly indicator	values b/w 92.2 and 94.77
	cons.conf.idx	numeric	consumer confidence index --- monthly indicator	values b/w -50.8 and -26.9
	euribor3m	numeric	euribor 3 month rate --- daily indicator	values b/w 0.634 and 5.045
	nr.employed	numeric	number of employees --- quarterly indicator	values b/w 4964 and 5228
Output variable	subscribed	binary	has the client subscribed a term deposit?	'no', 'yes'

In this case, studying and analyzing socioeconomic attributes (i.e. **emp.var.rate**, **cons.price.idx**, **cons.conf.idx**, **euribor3m**, **nr.employed**) need a extensive knowledge of that specific domain and therefore beyond the purview of our limited descriptive study of this dataset.

Due to this, our exploratory data analysis we would be focusing on the remaining 16 attributes.

Figure 1: Example of the dataset.

age	job	marital	education	default	housing	loan	contact	month	day_of_week	duration	campaign	pdays	previous	poutcome	subscribed
56	housemaid	married	basic.4y	no	no	no	telephone	may	mon	261	1	999	0	nonexistent	no
57	services	married	high.school	unknown	no	no	telephone	may	mon	149	1	999	0	nonexistent	no
37	services	married	high.school	no	yes	no	telephone	may	mon	226	1	999	0	nonexistent	no
40	admin.	married	basic.6y	no	no	no	telephone	may	mon	151	1	999	0	nonexistent	no
56	services	married	high.school	no	no	yes	telephone	may	mon	307	1	999	0	nonexistent	no
45	services	married	basic.9y	unknown	no	no	telephone	may	mon	198	1	999	0	nonexistent	no

Visualization

Age

Summary of **Age** attribute.

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   17.00   32.00   38.00   40.02   47.00   98.00
```

First feature is Age of clients, this is numeric features in range between 17 and 98 years old. We can see Bar plot of this data on *Figure 3* and next to it is the histogram of this values with pointed density and mean value. Histogram is prepared with bandwidth of 5 years. We can see that this plot is right-skewed, but also similar to normal distribution.

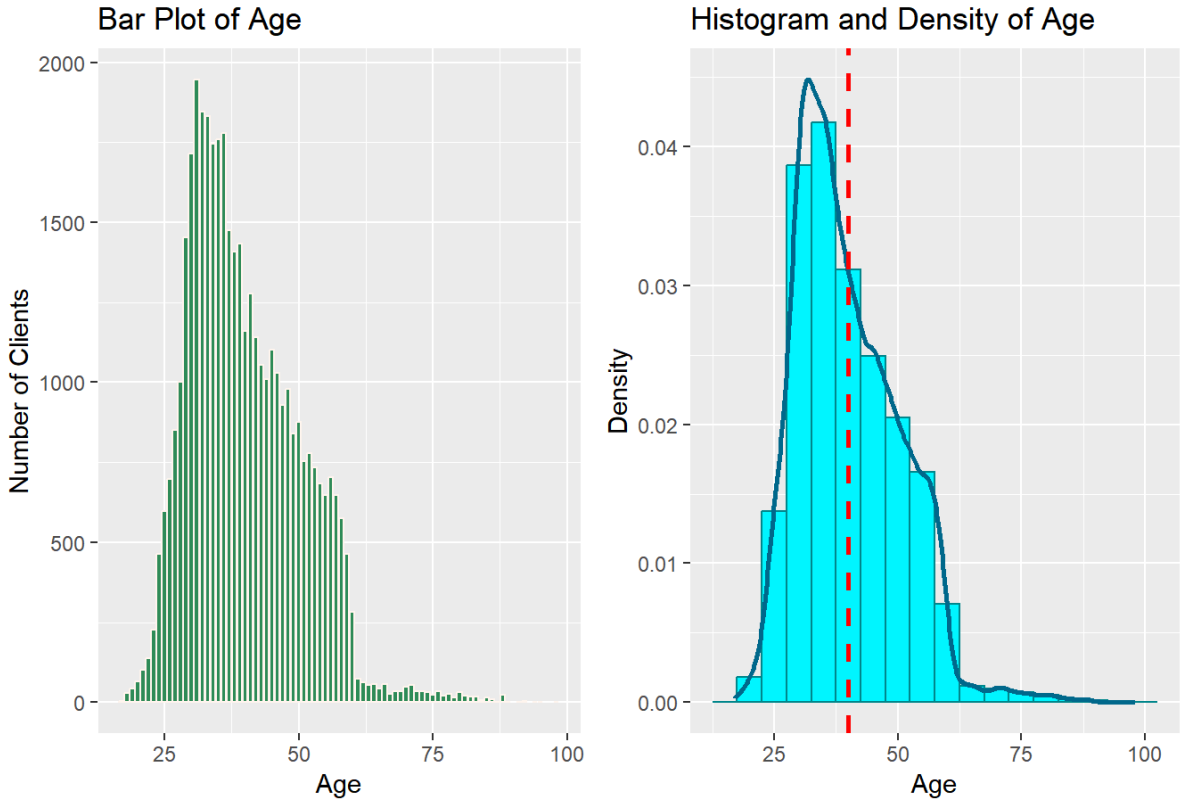


Figure 3

Next, we compare box-plots of Age according to result of campaign. 'yes' means that client subscribed a term deposit, 'no' when didn't do it. Result are presented on *Figure 4*. We see distribution for clients, who subscribed a term deposit is more diffused, but it is because people, who said 'yes' is less.

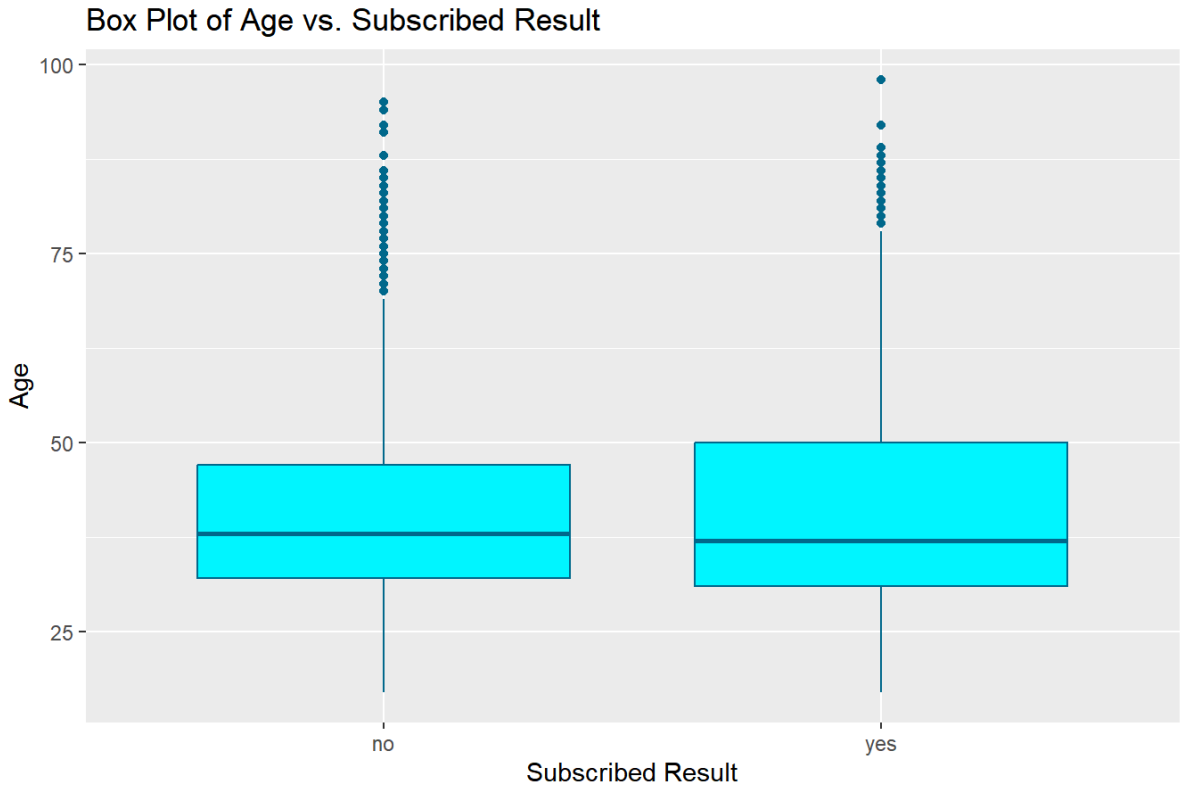
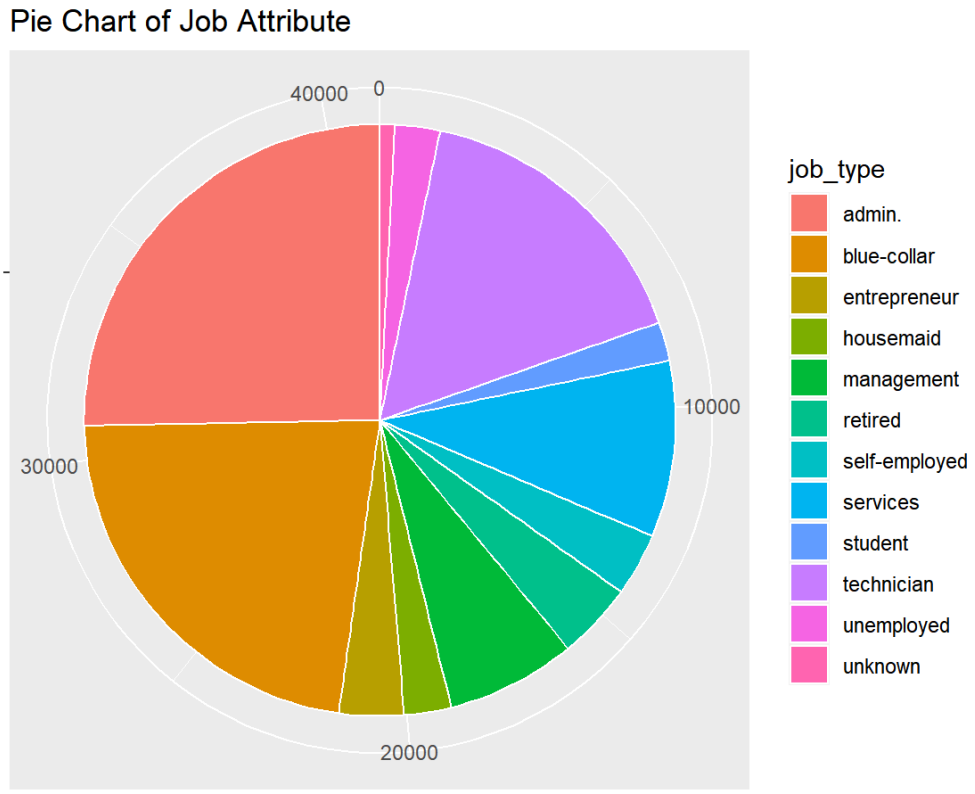


Figure 4: Box-plots of age for different results of campaign

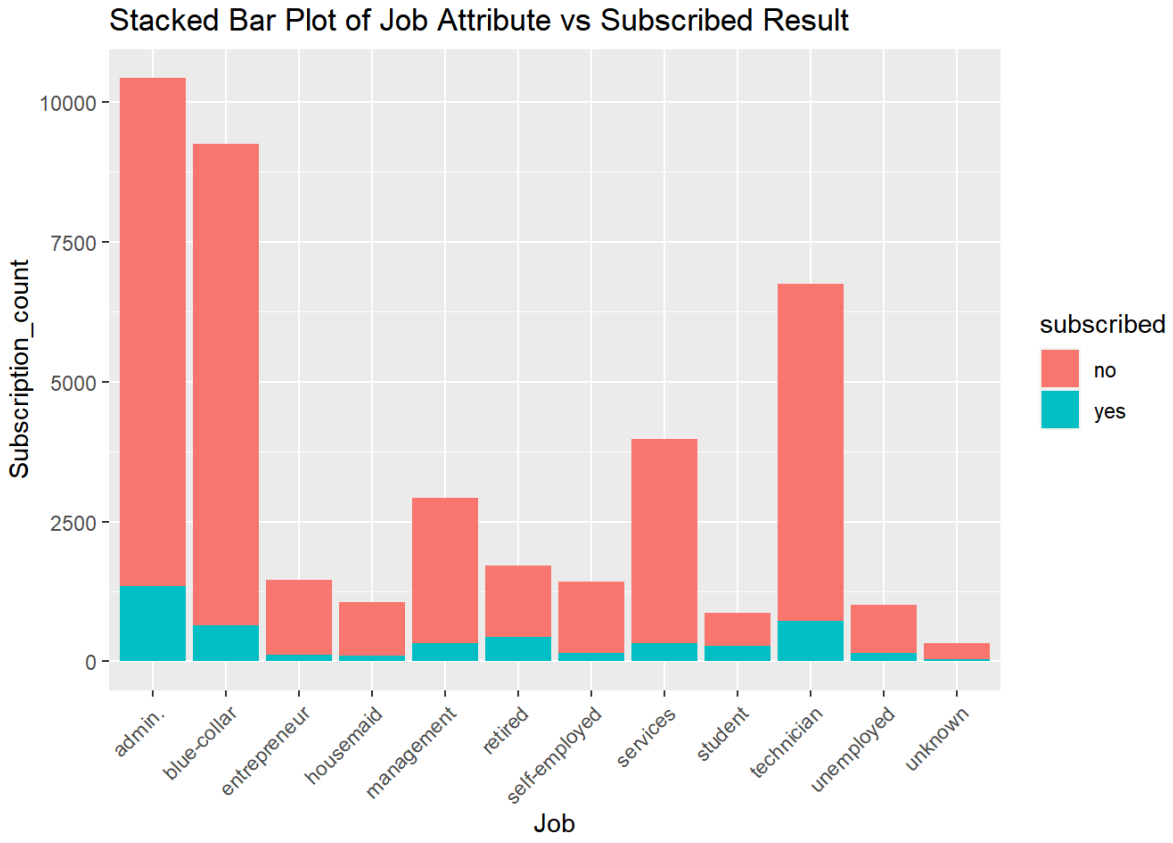
Job

The second attribute is **Job**, it is categorical in type and takes 12 different categories of values i.e. ’management’, ’technician’, ’entrepreneur’, ’blue-collar’, ’unknown’, ’retired’, ’admin.’, ’services’, ’self-employed’, ’unemployed’, ’housemaid’, ’student’. The pie-chart represent the proportions of job categories within the clients.

	job_type	count.Freq	percent
1	admin.	10422	25.303
2	blue-collar	9254	22.468
3	technician	6743	16.371
4	services	3969	9.636
5	management	2924	7.099
6	retired	1720	4.176
7	entrepreneur	1456	3.535
8	self-employed	1421	3.450
9	housemaid	1060	2.574
10	unemployed	1014	2.462
11	student	875	2.124
12	unknown	330	0.801



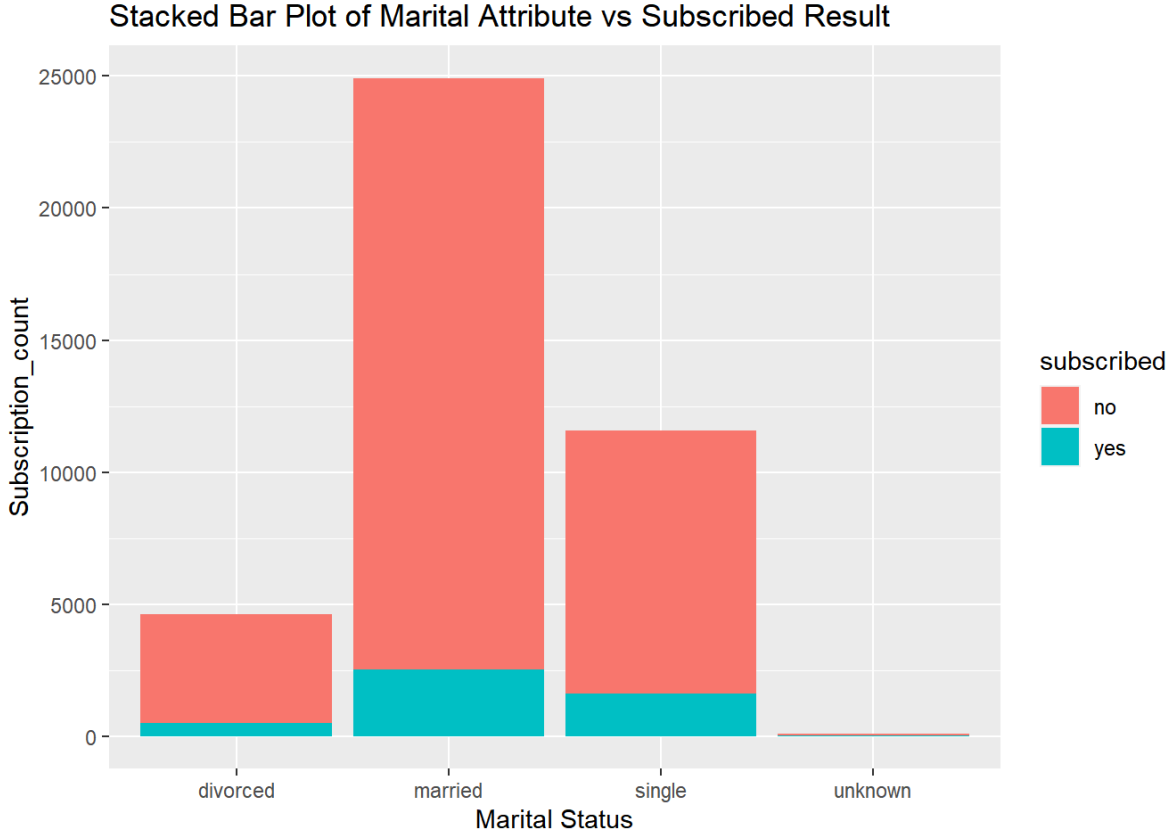
From the bar-plot below, we see that the job type that subscribed to a term deposit. The most of them comes from management, followed by technician, and blue-collar job holders has the least subscription to a term deposit. In later analysis we will treat 'unknown' variables as missing values.



Marital

The third attribute is marital, which is categorical in kind and has range of the following values: 'divorced', 'married', 'single'.

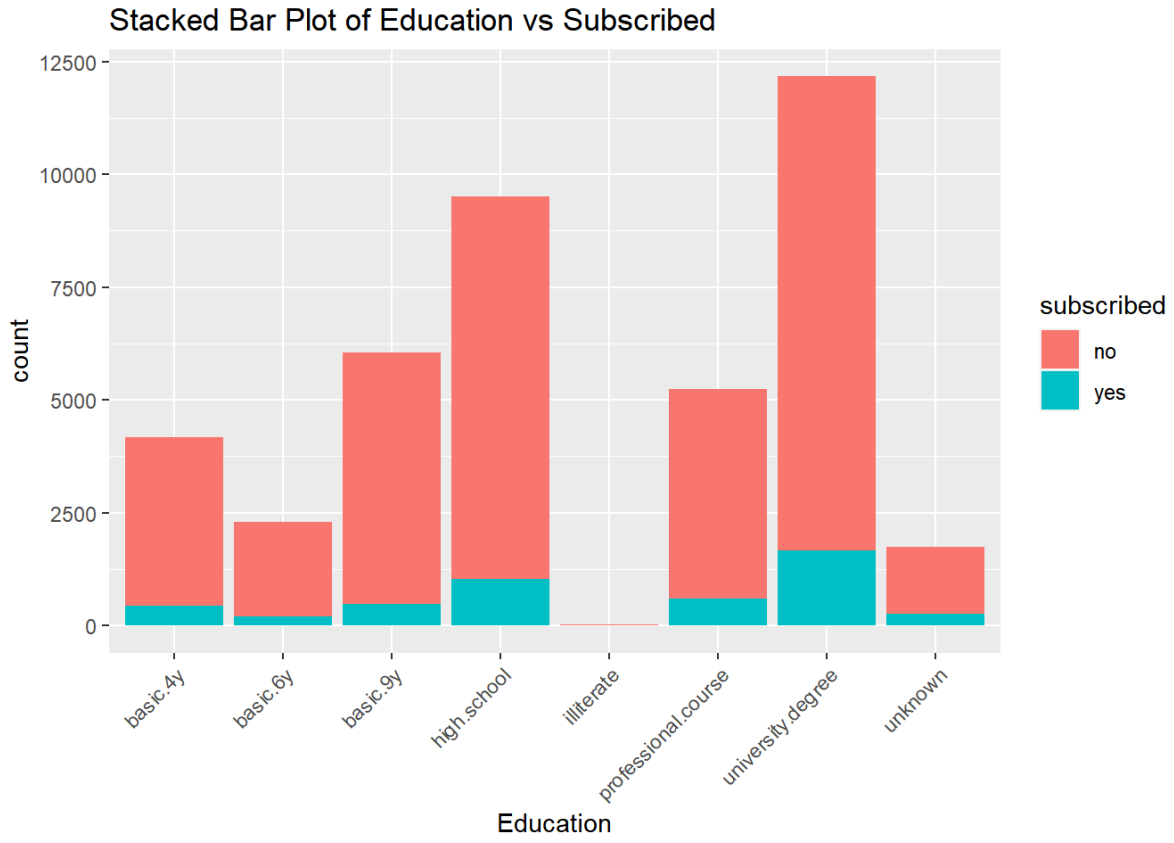
From the bar-plot of marital below, the 'married' has the highest subscription to a term deposit and also has the highest unsubscribed to a term deposit, because marriage people are the biggest group of analysed clients. The 'single' has the second largest subscription and has the second largest unsubscribed to a term deposit with a count of about 9948. In pie charts we can see percentage ratio of individual marital categories for various results of final campaigns.



Education

The fourth attribute in the data set is **Education**, which is categorical in kind and has the following values of attribute: “basic.4y”, “basic.6y”, “basic.9y”, “high.school”, “illiterate”, “professional.course”, “university.degree” and “unknown” . In the following table we see that the highest number of subscribed and unsubscribed to the bank term deposit belong to ‘university.degree’ education category, followed by ‘high.school’ education. The ‘illi terate’ has the least number of subscribed and unsubscribed to the bank term deposit.

	no_count	yes_count	no_%	yes_%
basic.4y	3748	428	9.1	1.039
basic.6y	2104	188	5.108	0.456
basic.9y	5572	473	13.528	1.148
high.school	8484	1031	20.598	2.503
illiterate	14	4	0.034	0.01
professional.course	4648	595	11.285	1.445
university.degree	10498	1670	25.488	4.055
unknown	1480	251	3.593	0.609



Default, housing, loan vs Subscribed

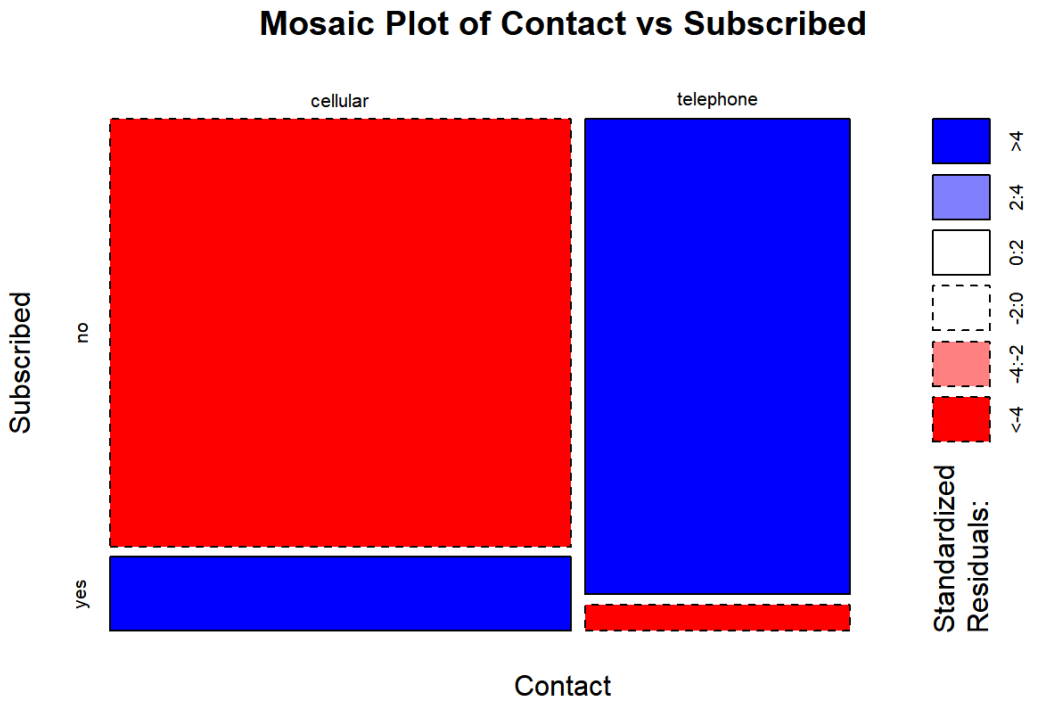
In the tables below we can see the numbers(left table) and percentages(right table) of clients with (“yes”) and without (“no”) any loan default (**default**), housing (**housing**), any sort of loan (**loan**) and subscription to the bank term deposit (**subscribed**).

	no	unknown	yes
default	32588	8597	3
housing	18622	990	21576
loan	33950	990	6248
subscribed	36548	0	4640

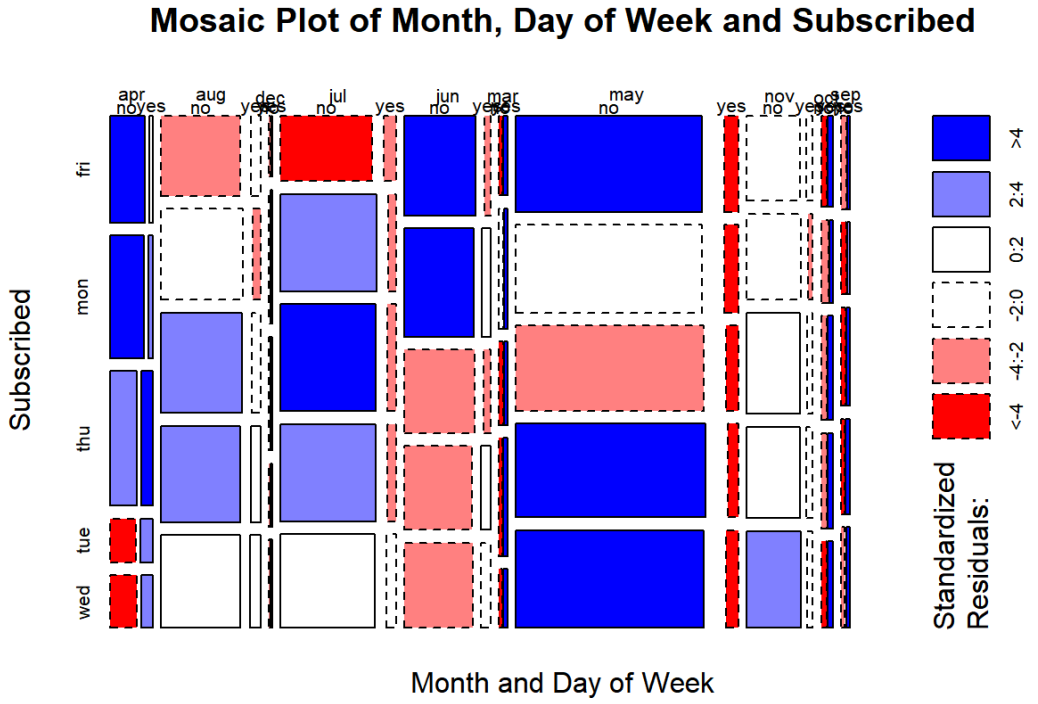
	no_%	unknown_%	yes_%
default	79.12	20.87	0.01
housing	45.21	2.4	52.38
loan	82.43	2.4	15.17
subscribed	88.73	0	11.27

Contact

Here we consider the attribute **contact**, it is categorical in kind and has its values of attributes to be “cellular”, “telephone” and “unknown”. The mosaic plot in figure below shows the means by which the clients were contacted and the number of clients who were contacted have they subscribed or did not subscribe to the bank term deposit.The plot clearly shows that the number of clients that were contacted but did not subscribe was far more than the clients who subscribed to the bank term deposit.



Day & Month



Conclusion