EDA Bank Marketing

Alexa Victoria Canche Anaya

October 16th 2019

1 Introduction

The data set was taken from the UC Irvine Machine Learning Repository [1], the data is related with direct marketing campaigns (phone calls) of a Portuguese banking institution from May 2008 to November 2010. The classification goal is to predict if the client will subscribe a term deposit (variable y).

1.1 Data Set Features

The data set contains 45,211 rows and 17 columns (variables). Accordingly to the UCI repository we should have 41,188 and 20 variables, but that is not what we had in our csv.

The input variables are:

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')
8. Contact: contact communication type (categorical: 'cellular','telephone')
9. Month: last contact month of year (categorical: 'jan', 'feb', 'mar', ..., 'nov', 'dec')
10. Day: last contact day of the week (categorical: 'mon','tue','wed','thu','fri')

11. Duration: last contact duration, in seconds (numeric). Important note: this attribute highly affects the output target (e.g., if duration=0 then y='no'). Yet, the duration is not known before a call is performed. Also, after the end of the call y is obviously known. Thus, this input should only be included for benchmark purposes and should be discarded if the intention is to have a realistic predictive model.

12. Campaign: number of contacts performed during this campaign and

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 that client

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

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

17. Balance: the balance of the clients

Some inconsistencies with the data were that, e.g. in pdays that instead of the 999 that the repository said it had, we had -1 and in education we didn't have those specifications, we had 'primary', 'secondary', 'tertiary', and 'unknown'. Another one, was that we had and extra variable called 'balance' which in the UCI didn't say we had.

1.2 Data Cleaning

To begin with, we started with checking some basic aspects of our data: the shape of our data as previously said, we were supposedly have a shape of (41188,20) but instead we have (45211,17). We have 7 numerical data and 10 categorical, none duplicated data, and thankfully no missing values as you can see in the graphic of the figure 1.

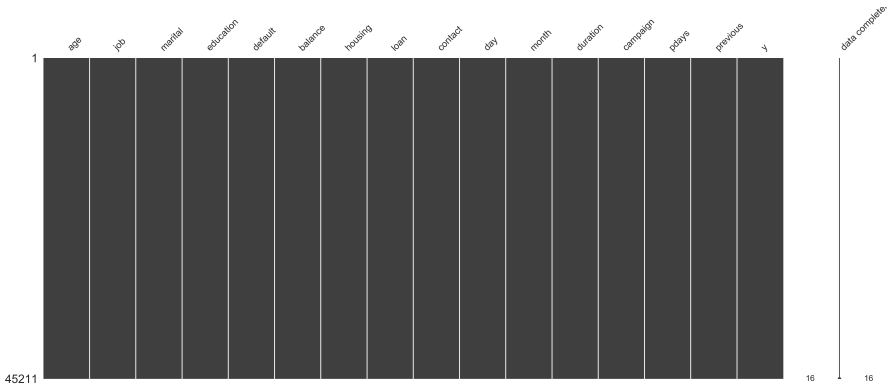


Figure 1: Missing data

1.2.1 Missing data

As we already know in numerical data we have NaN or Null values which describe missing data , but in categorical we have different words, in the case we have 'unknown'. So, what we did is that we searched for 'unknown' values in our categorical variables and delete the ones who had more than *50%* of missing values, try to fill some of the values and to just the others just leave it. After checking our variables only one we had to drop, and two variables we conclude that could help each other to fill those spots. First, what we did was create a contingency table with the variables 'education' and 'job', it was extremely helpful to see for each job what was more likely to have in education, as well as for education what job was more probable to have. Hence, we fill rows that had job with 'unknown' and for example, education 'primary' with the table we could see that the job that repeated the most was 'blue-collar', so we fill it in.

2 Univariate Analysis

2.1 No-Graphical of numerical data

2.1.1 Central Tendency and Dispersion

At first, we did the univariate analysis with the numerical data, we search for the central tendency values. In there we could see for example, that our average client has around 41 years old and the dispersion in the quartiles 33, 39 and 48, you can see more in figures 2, 3 and 4.

	age	balance	day	duration	campaign	pdays	previous
count	45211.00	45211.00	45211.00	45211.00	45211.00	45211.00	45211.00
mean	40.94	1362.27	15.81	258.16	2.76	41.02	0.58
std	10.62	3044.77	8.32	257.53	3.10	99.79	2.30
min	18.00	-8019.00	1.00	0.00	1.00	0.00	0.00
25%	33.00	72.00	8.00	103.00	1.00	0.00	0.00
50%	39.00	448.00	16.00	180.00	2.00	0.00	0.00
75%	48.00	1428.00	21.00	319.00	3.00	0.00	0.00
max	95.00	102127.00	31.00	4918.00	63.00	871.00	275.00

Figure 2: Central Tendency

	age	job	marital	education	default	balance	housing	loan	contact	day	month	duration	campaign	pdays	previous	y
0	32	blue-collar	married	secondary	no	0	yes	no	cellular	20	may	124	1	0	0	no

Figure 3: Central Tendency: Mode

```
age          39.0
balance     448.0
day          16.0
duration    180.0
campaign      2.0
pdays         0.0
previous      0.0
```

Figure 4: Central Tendency: Median

2.1.2 Skeweness and Kurtosis

"It is the degree of distortion from the symmetrical bell curve or the normal distribution. It measures the lack of symmetry in data distribution. It differentiates extreme values in one versus the other tail. A symmetrical distribution will have a skewness of 0."[2], in our data, we had 'age' and 'day' variable were around 0, the rest ('balance', 'campaign', 'pdays' and 'previous') were bigger than zero, meaning they had a positively skewed and the mean and median will be greater than the mode.More in the figure 5. Remember that skeweness is:

- ASF >0, Positive Asymmetry.
- ASF <0, Negative Asymetry.

- ASF = 0, Asymetric.

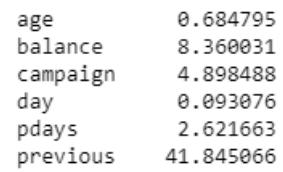


Figure 5: Skeweness

"Kurtosis is all about the tails of the distribution — not the peakedness or flatness. It is used to describe the extreme values in one versus the other tail. It is actually the measure of outliers present in the distribution."[2], since we don't have a very good symmetrical distribution based on our skeweness, the kurtosis won't really say much about our data. More in the figure 6. Remember that kurtosis is:

- Ap >3, Leptokurtic means that data are heavy-tailed or profusion of outliers.
- Ap <3, Platikurtic means the distribution produces fewer and less extreme outliers than does the normal distribution.
- Ap = 3, Mesokurtic means means that if the data follows a normal distribution.

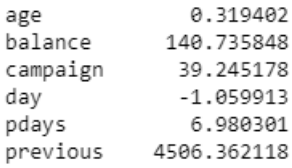


Figure 6: Kurtosis

We changed the value -1 of the pdays columns to 0 so that we can have more correlation with the previous column because both mean the same but they are shown differently. While analyzing the data, and the variables we realized that 'pdays' was described as 'number of days that passed by after the client was last contacted from a previous campaign (numeric; 999 means client was not previously contacted)' and that instead of 999 we had -1 that we changed to 0. Therefore, this means that from all the data the ones who have 0 in the pdays column are new clients and the ones who don't have it are old, what we did was to divide the data in two; new clients and old clients.

2.2 No-Graphical of categorical data

For this section the only thing to do with the categorical data is to count the frequency and their percentage. The categorical variables are job, marital, education, contact and month and binary we have default, housing, loan and our y (the target).

2.3 Graphical of numerical data

The first thing we did here is a histogram of the distribution of all the numerical data, as we saw in the no-graphical analysis 'day' and 'age' are kind of a asymmetrical distribution, and the other ones are passively skewed, which means most of the data, the mean, is on the left side and it's true since you can appreciate it in the graphs (figure 7).

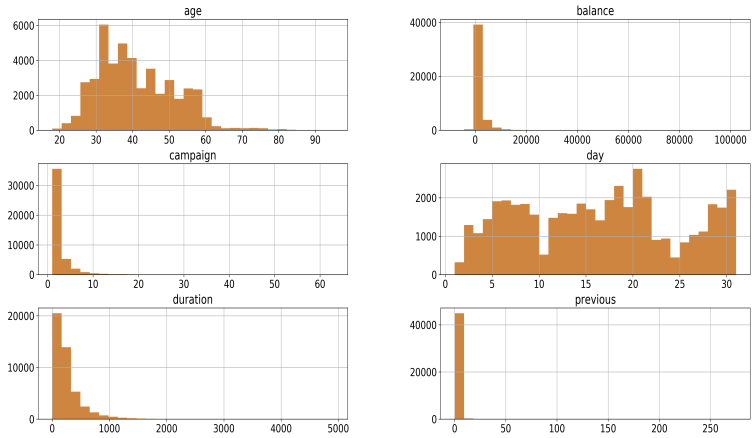


Figure 7: Distribution of numerical data

The 'duration' variable we can see that kind of follows a geometrical normal distribution, we can see it on the black line while the purple one is the real distribution of the data (figure 8).

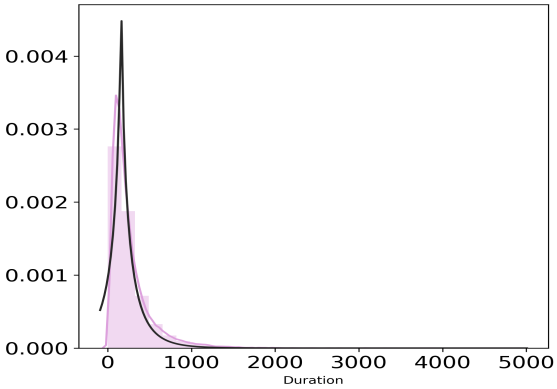
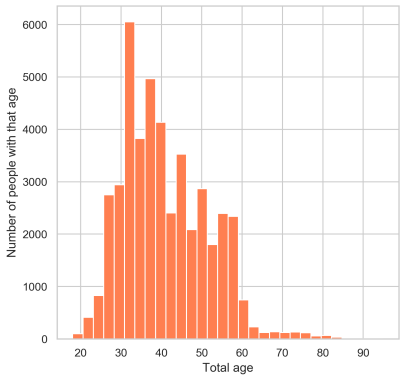
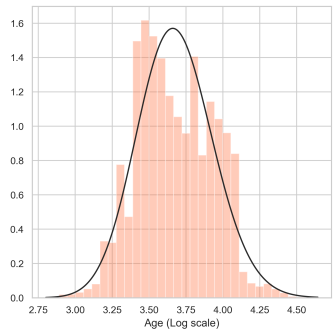


Figure 8: Distribution of duration

The 'age' column was the who had more of normal distribution as we can see in figure 9(a), so what we did was applied a logarithm to the data to change the distribution, and as we can observe in the next figure 9(b) it was good, but it was no the best.



(a) Age distribution



(b) Age distribution (log scale)

Figure 9: Age

Finally in the 'day 'histogram (figure 10) like the 'age' one it kind of looks more like a normal distribution which is good, although it would be good to try to do some transformations.

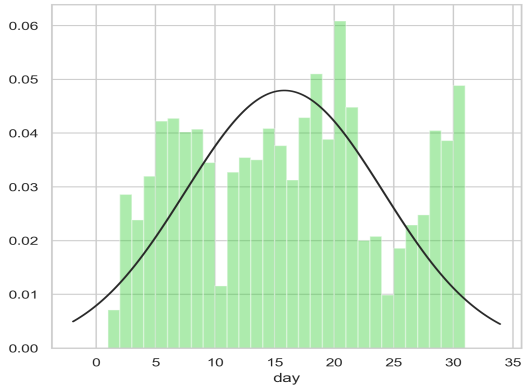


Figure 10: Distribution of day

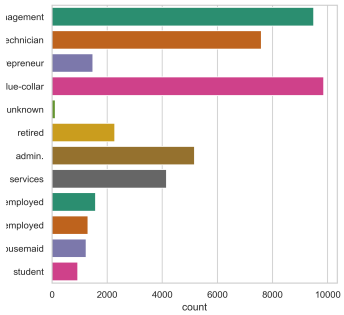
As we said we would split the data in two, we created a boxplot with the days that passed after the client was previously contacted with the data of the old clients, this is different from the figure 7, since here we only have the days of the old clients without the zeros, which makes it clearer for the distribution, in the figure 11, we we that out data is highly skewed on the left (positive), which means most of the data its around 200 days, still we can see some outliers passing 600 days.

Figure 11: Distribution of days for old clients

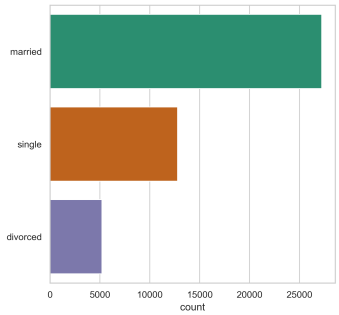
2.4 Graphical of categorical data

Here is better to understand visually the frequency of categorical data, as we can see 10(a) the job frequency, we have around 10,000 clients who are blue-

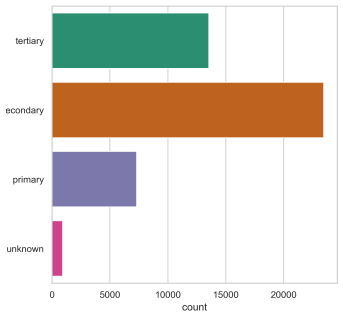
collars and less than 2,000 are students.
In 10(b) the marital status frequency, more than 25,000 clients are married while the minority are divorced.
Figure 10(c), secondary it the education that more than 20,000 have, on the hand the biggest one should be tertiary we have a little less than 15,000 clients.
Figure 10(d) more than *98%* of the people don't have a credit default swap, which is understandable, since is work for big companies and not individuals.
Figure 10(e) more people have a housing loan, but there is just *11.6%* of difference between the people who have and don't have a housing loan, so it's almost equivalent.
Figure 10(f) in the loan variable almost*84%* of the people have a personal loan, while the *16%* doesn't.
Figure 10(g) As it was to be expected the cellphone was the most used way of contact with *64.7%* while the telephone was only *6.43%*.
Figure 10(h) The data set showed that May was the most frequent month, and after checking the success of every month in fact, was still May that had the most success based on the term deposit target with 925 people that subscribed to the term deposit.



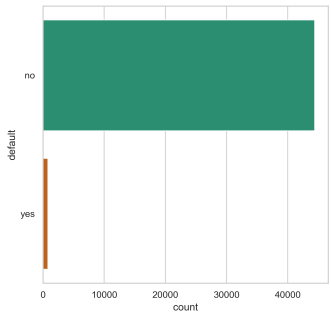
(a) Job



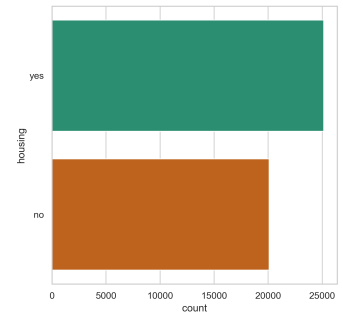
(b) Marital



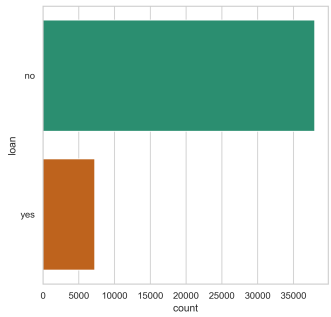
(c) Education



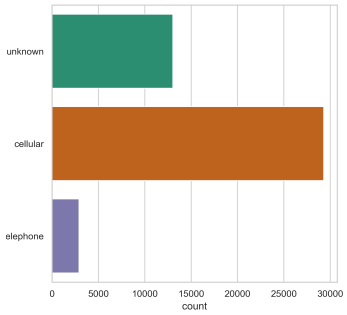
(d) Default



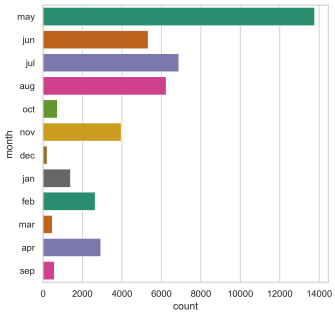
(e) Housing



(f) Loan



(g) Contact



(h) Month

Figure 12: Categorical Frequency

3 Multivariate Analysis

3.1 Correlation, covariance and cross-tabulation

Remember that correlation is a statistical measure that indicates how strongly two variables are related.[3] Where:

- value >0 and value =<1: means that as one variable gets larger the other gets larger: possitive correlation.
- values <0 and values >= -1: means that as one gets larger, the other gets smaller: negative correlation.
- value = 0: means there is no relationship between the variables: zero correlation.

As we can see in the figure 13, we don't have really any strong correlation, the best we have would be between 'previous' and 'pdays' with .45, and since the zeros between the two variables means the same it provoke a better correlation.

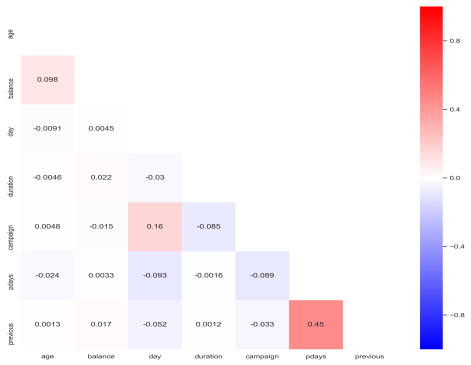


Figure 13: Correlation

As previously said, we dived our data set in two (new and old clients), what we can appreciate in Figure 14 the correlation shown in their distributions, still we don't have a good correlation, in fact it's worst in both of the data sets.

(a) Old clients (b) New clients

Figure 14: Correlation in old and new clients

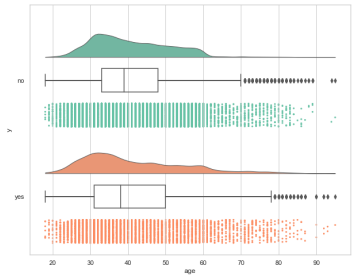
Covariance is a measure of how much two random variables vary together. It's similar to variance, but where variance tells you how a single variable varies, co variance tells you how two variables vary together.[3]

- value >0: y increases and x decreases: large possitive covariance.
- values <0: y decreases and x increases: large negative covariance.
- value = 0: there is no linear tendency: nearly zero covariance.

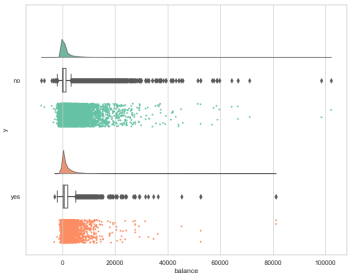
Mostly we have negative values, so we have negative covariance, meaning that 'y' decreases and 'x'increases, but let's remember that almost none of all variables had correlation, therefore they don't have much of a relation, less one that means that one of them increases and the other one decreases.

3.2 Numerical variables and target

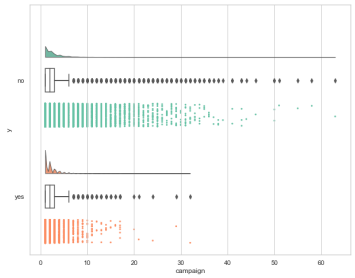
These graphics (figure 15) are very useful. We can see the distribution of the numerical variables divided by the target. In 'age' we can see that the mean of each target is around 38, so we can say that they age is not that big of a factor. And 'day' we can clearly see how the mean for 'yes' of target is exactly 15, while in 'no' is around 17. Balance, campaign, duration and previous as we previously saw we have positive skewed in both targets in all of them, but now we can see no matter the target the data stills the same.



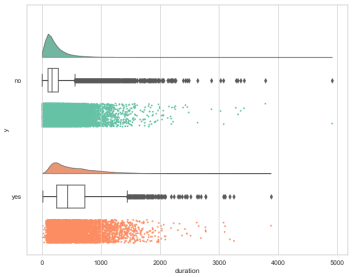
(a) Age



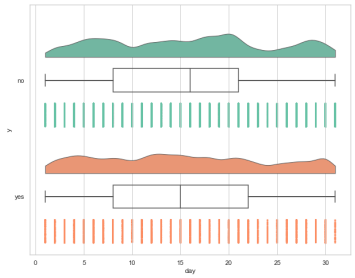
(b) Balance



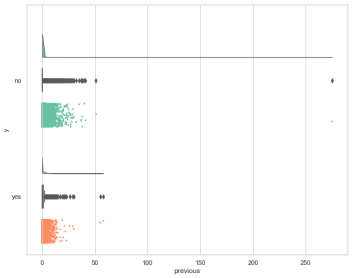
(c) Campaign



(d) Duration



(e) Day

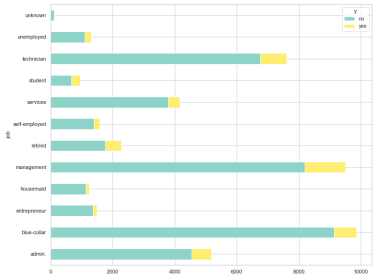


(f) Previous

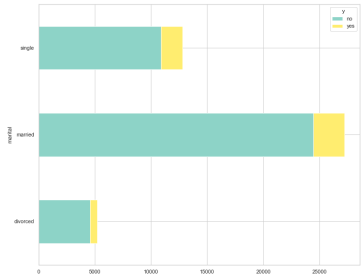
Figure 15: Numerical Values separated by target

3.3 Categorical variables and target

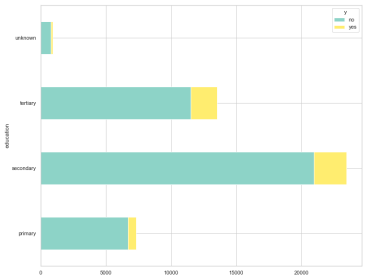
For all the categorical variables most of them said 'no' in the target. Some observations, like in 'job' specifically in 'blue-collar' is the most frequent, it's no the one who had more 'yes' in target it was actually 'management', that happens too in 'housing' even though we have more people with housing loan, 'yes' was more frequent in the ones that don't have a housing loan.



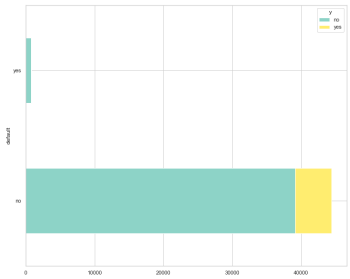
(a) Job



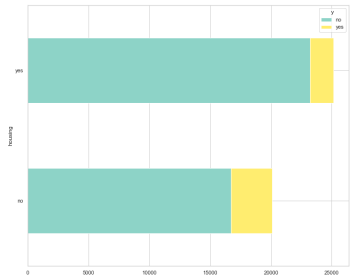
(b) Marital



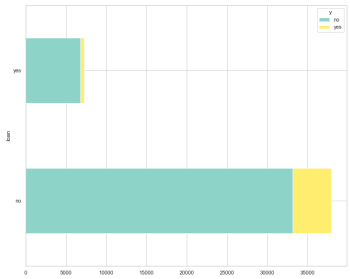
(c) Education



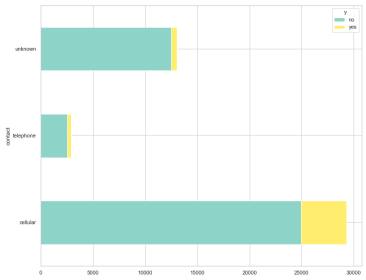
(d) Default



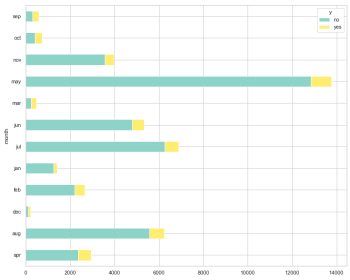
(e) Housing



(f) Loan



(g) Contact



(h) Month

Figure 16: Categorical variable separated by target

3.4 Age and some categorical variables

3.4.1 Age and Job

Here we appreciate and see the mean age of for every job, for example, it was obvious that 'retired' was going to be around 60 years old and students around 26 years old. And as we saw in the mean of 'age' most of them are around 40 years old.

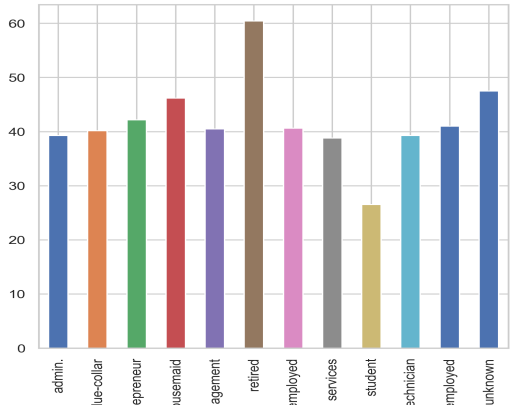


Figure 17: Age and Job

3.4.2 Age and Marital

People who are married, divorce or widowed are around 43 to 45 years old, and single around 33.

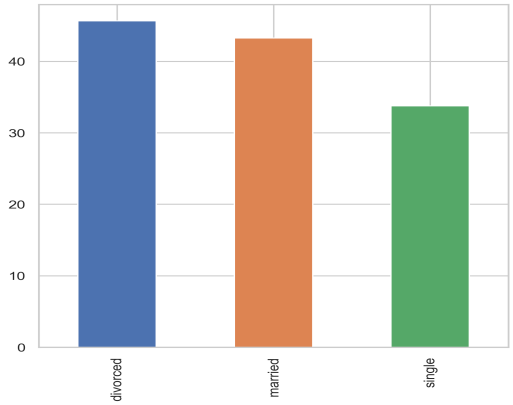


Figure 18: Age and Marital

3.4.3 Age and Education

People who finished school until primary are around 45 years old, tertiary and secondary around 39.

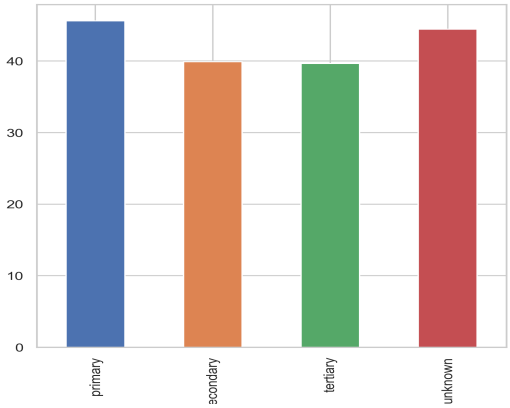
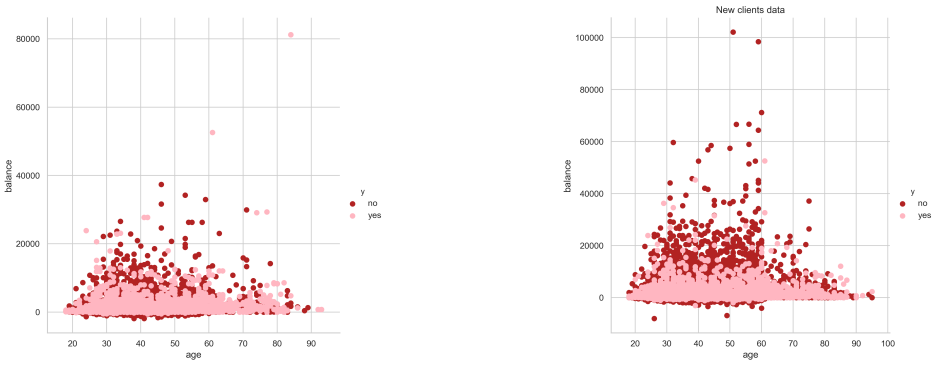


Figure 19: Age and Education

3.5 Age and some numerical variables

Researching we figure it out that balance meant how much money you 'owe' to your credit card and the negatives numbers in here, meant that they have an account with that amount. As you can see, in the old clients data set they data wasn't that well distributed, the maximum amount of the balance didn't go beyond 80,000. On the other hand, in the new clients we have two outliers that, just because of them we see a graph of 100,000 in balance.



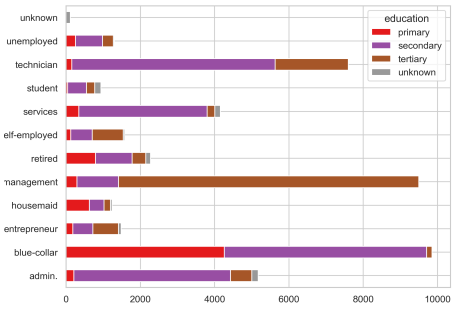
(a) Old clients

(b) New clients

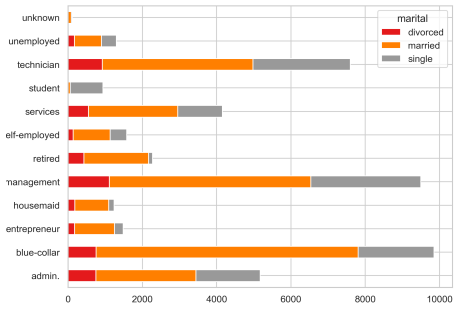
Figure 20: Age of the balance of the clients

3.6 Job and some categorical data

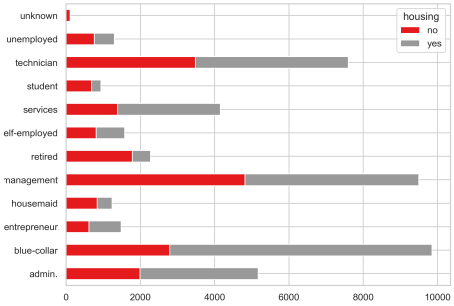
What we tried to do here, was to see the relationships between variables that we thought could mean something. In figure 21, Job-Education we can see we have a large amount of people who work as 'management' or 'blue-collar', but in 'management' of the total 9,497 more than 8,000 studied until tertiary and for 'blue-collar'. Same happens in Job-Marital, in 'blue-collar', or more than 9,000 in total 7,000 are married and also has a loan



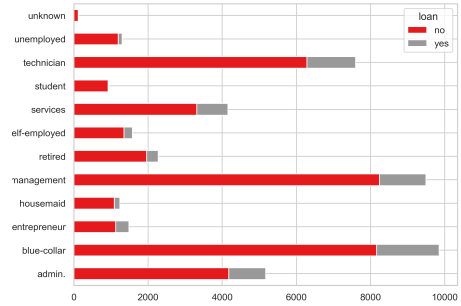
(a) Job-Education



(b) Job-Marital



(c) Job-Housing



(d) Job-Loan

Figure 21: Job and some categorical data

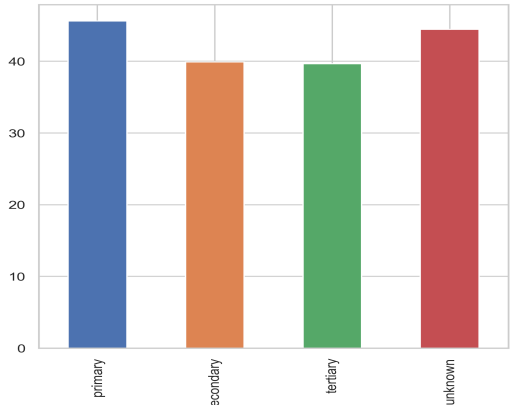


Figure 22: Age and Education

3.7 Job and some numerical data

3.7.1 Job and Balance

We can see that for the 'no' target (the purple) every the balance per job is around 1,000 and the lowest is services with 959. Conversely in the 'yes' target (the green) is a little bit higher, getting to 2,000, in 'self-employed' we can see a peak that goes more than 2,000 in balance.

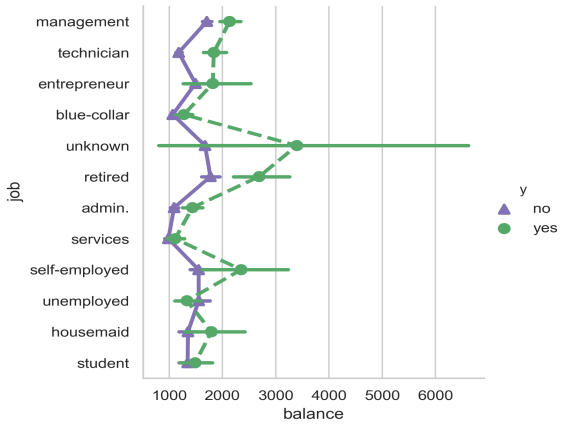


Figure 23: Age-Balance

3.7.2 Job and Duration

The duration, means how much did it took the contact in seconds, here 'management' and 'blue-collar' are almost the same around 2,500 seconds and the lowest were 'students' with a little bit more than 1,000 seconds.

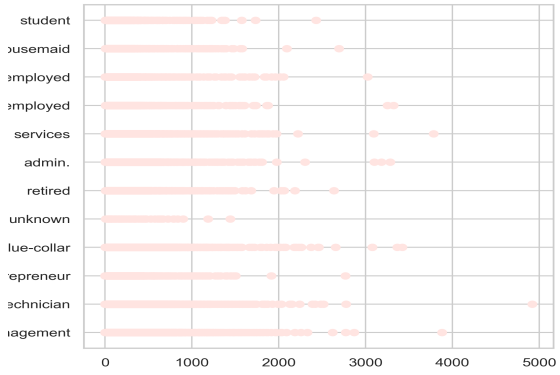


Figure 24: Age-Duration

3.7.3 Month and Campaign

Remembering that 'campaign' meant the number of contacts performed during this campaign, and as we can observe august was the month in which it was most performed contact, and still may was the month who had more success on the target and it only performed around 2.5 contacts.

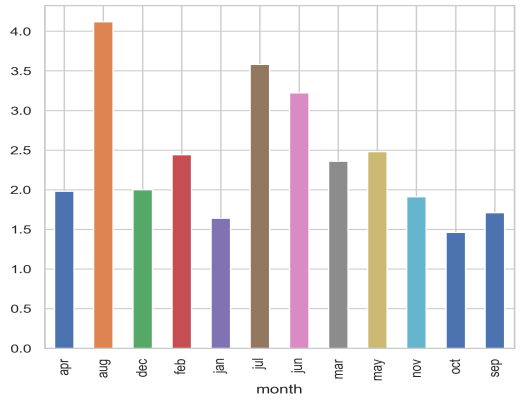


Figure 25: Month-Campaign

3.7.4 Contact and Duration

The cellphone was obviously the most used as we saw previously, and it makes sense that is the one with more duration, however the telephone is almost the same with 228.94 seconds in average and the cellphone had 263.5 in average.

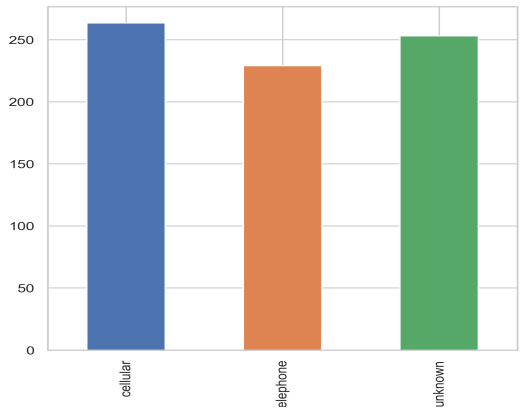
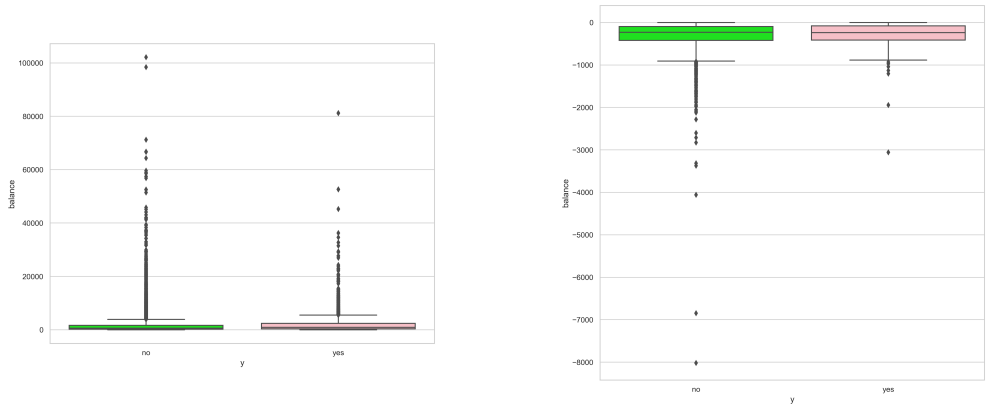


Figure 26: Contact-Duration

3.7.5 Balance

Before we explained that the negative values in balance meant a good thing since mean that the client doesn't have a debt of a credit card, on the contrary the positive values meant the debt that the clients owes to the bank. Here we used box-plots to observe their distribution separately. In the debt part, it looks like is not much, but there are some huge outliers, actually the maximum value here is 102,127 and for the no-debt we have a better distribution, in the 'no' target we can see more outliers than the 'yes', the maximum (means the amount available for a loan depending on the client) is 8,019,



(a) Bad Balance

(b) Good Balance

Figure 27: Balance of the clients

4 Conclusion

To conclude, the percentage of success was calculated and it was a *11.7%*, of all the 45,211 clients only 5,289 people said yes to subscribed to the term deposit which was the target of this marketing campaign. Now, the percentage success for the old clients was *23.07%* and for new clients *9.16%*. With these results we can finish saying that the campaign was not a favorable outcome.

References

[1] Bank marketing data set, 2012.

[2] Skew and kurtosis: 2 important statistics terms you need to know in data science, 2018.

[3] Covariance and correlation, n.d.