Group Name	"Data Pirates"	Data Report Intake Link	Repository Link	
Name	Email	Country	University/Company	Specialisation
Reema Al-Otaibi	otaibi_reema@outlook.com	Saudi Arabia	AOU	Data Science
Tarek Mohamed	tarek.mohamed.abdullah@gmail.com	Egypt	Fayoum University	Data science

Bank Marketing Campaign

Understanding the Problem:

- ABC Bank wants to sell it's 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).
- Bank wants to use ML model to shortlist customer whose chances of buying the product is more so that their marketing channel (tele marketing, SMS/email marketing etc) can focus only to those customers whose chances of buying the product is more.

Understanding the DataSet:

- There are 2 dataset files: bank and bank-additional.
- Bank-additional with all examples (41188) and 20 inputs, ordered by date (from May 2008 to November 2010).

Try of Data:

1. **Input variables:**
    - bank client data:

1 - age (*numeric*)

2- job (*categorical*: ‘admin.’ , ‘blue-collar’ , ‘entrepreneur’ , ‘housemaid’ , ‘management’ , ‘retired’ , ‘services’ , ‘student’ , ‘technician’ , ‘unemployed’ , ‘unknown’)

3 - marital (*categorical*: ‘divorced' , ‘married' , ‘single' , ‘unknown')

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: (*categorical*: ‘cellular’ , 'telephone')

9 - month: (*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*).
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.

- other attributes:

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’)

- social and economic context attributes

16 - emp.var.rate: employment variation rate - quarterly indicator (*numeric*)

17 - cons.price.idx: consumer price index - monthly indicator (*numeric*)

18 - cons.conf.idx: consumer confidence index - monthly indicator (*numeric*)

19 - euribor3m: euribor 3 month rate - daily indicator (*numeric*)

20 - nr.employed: number of employees - quarterly indicator (*numeric*)

2. **Output variable (desired target):**
   21 - y - has the client subscribed a term deposit? (*binary*: ‘yes' , ‘no')

Problems in the Data:

- It’s not clear from the beginning that the data has missing values.

Out[155]:

	age	job	marital	education	default	housing	loan	contact	month	day_of_week	duration	campaign	pdays	previous	poutcome
0	56	housemaid	married	basic.4y	no	no	no	telephone	may	mon	261	1	999	0	nonexistent
1	57	services	married	high.school	unknown	no	no	telephone	may	mon	149	1	999	0	nonexistent
2	37	services	married	high.school	no	yes	no	telephone	may	mon	226	1	999	0	nonexistent
3	40	admin.	married	basic.6y	no	no	no	telephone	may	mon	151	1	999	0	nonexistent
4	56	services	married	high.school	no	no	yes	telephone	may	mon	307	1	999	0	nonexistent
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
41183	73	retired	married	professional.course	no	yes	no	cellular	nov	fri	334	1	999	0	nonexistent
41184	46	blue-collar	married	professional.course	no	no	no	cellular	nov	fri	383	1	999	0	nonexistent
41185	56	retired	married	university.degree	no	yes	no	cellular	nov	fri	189	2	999	0	nonexistent
41186	44	technician	married	professional.course	no	no	no	cellular	nov	fri	442	1	999	0	nonexistent
41187	74	retired	married	professional.course	no	yes	no	cellular	nov	fri	239	3	999	1	failure

41188 rows × 21 columns

```
In [156]: bank_df.isna().sum()
Out[156]: age             0
          job             0
          marital         0
          education       0
          default         0
          housing         0
          loan            0
          contact         0
          month           0
          day_of_week     0
          duration        0
          campaign        0
          pdays           0
          previous        0
          poutcome        0
          emp.var.rate    0
          cons.price.idx  0
```

- Some values are not written neatly or in a proper way.

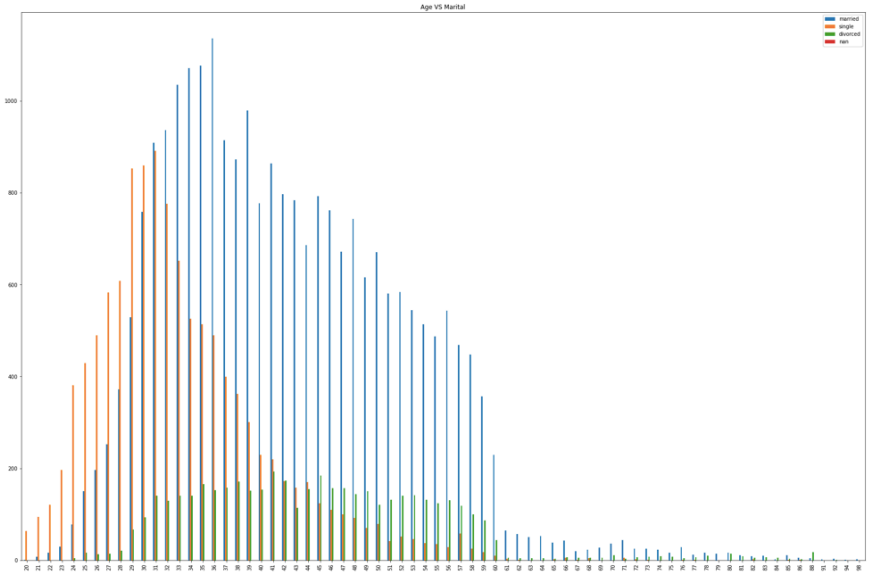
Out[178]:

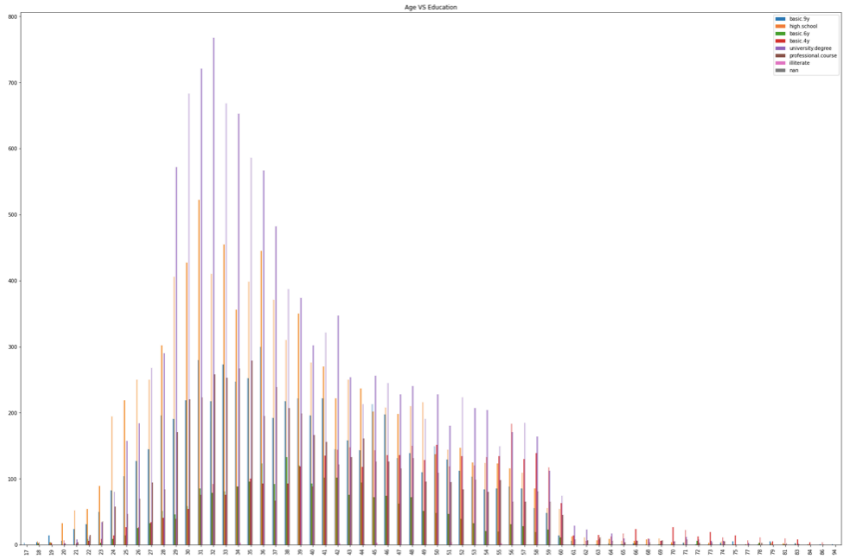
	age	job	marital	education	default	housing	loan	contact	month	day_of_week	duration	campaign	pdays	previous	poutcome	em
0	56	housemaid	married	basic.4y	no	no	no	telephone	may	mon	261	1	999	0	nonexistent	
1	57	services	married	high.school	NaN	no	no	telephone	may	mon	149	1	999	0	nonexistent	
2	37	services	married	high.school	no	yes	no	telephone	may	mon	226	1	999	0	nonexistent	
3	40	admin.	married	basic.6y	no	no	no	telephone	may	mon	151	1	999	0	nonexistent	
4	56	services	married	high.school	no	no	yes	telephone	may	mon	307	1	999	0	nonexistent	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	
1183	73	retired	married	professional.course	no	yes	no	cellular	nov	fri	334	1	999	0	nonexistent	
1184	46	blue-collar	married	professional.course	no	no	no	cellular	nov	fri	383	1	999	0	nonexistent	
1185	56	retired	married	university.degree	no	yes	no	cellular	nov	fri	189	2	999	0	nonexistent	
1186	44	technician	married	professional.course	no	no	no	cellular	nov	fri	442	1	999	0	nonexistent	
1187	74	retired	married	professional.course	no	yes	no	cellular	nov	fri	239	3	999	1	failure	

cons.conf.idx	euribor3m	nr.employed	y
-36.4	4.857	5191.0	no
-36.4	4.857	5191.0	no
-36.4	4.857	5191.0	no
-36.4	4.857	5191.0	no
-36.4	4.857	5191.0	no
...	...	...	...
-50.8	1.028	4963.6	yes
-50.8	1.028	4963.6	no
-50.8	1.028	4963.6	no
-50.8	1.028	4963.6	yes
-50.8	1.028	4963.6	no

Admin. In jobs, basic.4y - university.degree in education, ‘y’ the name of the last column in the data set.

Most of the data concentration is found in ages less than 61.





Approaches:

- Replaced yes values with 1 and no values with 0.
- Renamed ‘y’ column to ‘deposited?’
- Will clean the values in the data set.
- Decide whether data for age > 60 are outliers or not and drop then based on that.