



# MINI GROUP ASSIGNMENT

TOPIC – FINANCE

ANALYSIS ON BANK MARKETING

GROUP – 4



# INDEX

1. Team Members
2. Introduction to the Problem Statement
3. Aim
4. Technologies Used & Skills Used
5. Data Description
6. Data Collection and Cleaning
7. Problem Solving Steps
8. What could have been done better?
9. Takeaways and Conclusions
10. Future Steps



# TEAM MEMBERS

Hithashree J.

Aditya Aryan

Shrawani Hemant Deshmukh

Adarsh Kumar



# INTRODUCTION TO THE PROBLEM STATEMENT

The problem is that the Bank Marketing campaigns of a Portuguese banking institution need to identify the factors that cause the customers to tend to take the subscription, as well as Bank Marketing campaigns of a Portuguese banking institution need to identify the reasons behind the customer which make them not take the subscription.



# AIM

- ✓ To Determine/Analyse factors for the subscription and non-subscription using the ITP and NPV techniques.
- ✓ To generate a meaningful insight from the data which is provided using data science concepts.
- ✓ Analyzing a huge amount of data using modern tools. Collecting and arranging it in such a way that it is more sophisticated and straightforward.
- ✓ Predict the customer behavior and attract the customers in future according to their need. Based on this, they can use different marketing strategies to attract more customers.



### **TECNOLOGIES USED**

- ✓ Python
- ✓ Jupyter Notebook

### **SKILLS USED**

- ✓ Programming skills
- ✓ Visualization skills

# DATASET DESCRIPTION

- **Bank Marketing:** The data is related to direct marketing campaigns of a Portuguese banking institution. The marketing campaigns were based on phone calls. Often, more than one contact with the same client was required to assess if the product (bank term deposit) would be (or not) subscribed.
- In the dataset, there are 17 columns and 4521 rows.
- There are no null values and duplicate values or any missing values.
- There are 10 categorical attributes and 7 numerical attributes in the data set.
- Minimum age of the customers is 19 and maximum is 87. Mean age is 41.
- Mean balance is 1422.

### These are the attributes in the dataset:

- 1 - age: (numeric)
- 2 - job: type of job (categorical)
- 3 - marital: marital status (categorical)
- 4 - education (categorical)
- 5 - default: has credit in default? (binary)
- 6 - balance: average yearly balance, in Euros (numeric)
- 7 - housing: has a housing loan? (binary)
- 8 - loan: has personal loan? (binary)
- 9 - contact: contact communication type (categorical)
- 10 - day: last contact day of the month (numeric)
- 11 - month: last contact month of year (categorical)
- 12 - duration: last contact duration, in seconds (numeric)
- 13 - campaign: number of contacts performed during this campaign and for this client (numeric)
- 14 - P-days: number of days that passed by after the client was last contacted from a previous campaign (numeric, -1 means client was not previously contacted)
- 15 - previous: number of contacts performed before this campaign and for this client (numeric)
- 16 - poutcome: outcome of the previous marketing campaign (categorical)
- 17 - y - has the client subscribed to a term deposit? (binary)



# DATA COLLECTION AND CLEANING

- Data is collected from the 'bank' CSV file.
- As there were double quotes (""") all over the data in the CSV file, it could not be read directly into a dataframe in the desired manner. For this, we first read the file and removed all the double quotes from the text. We wrote the text back into the file and then we could import it.

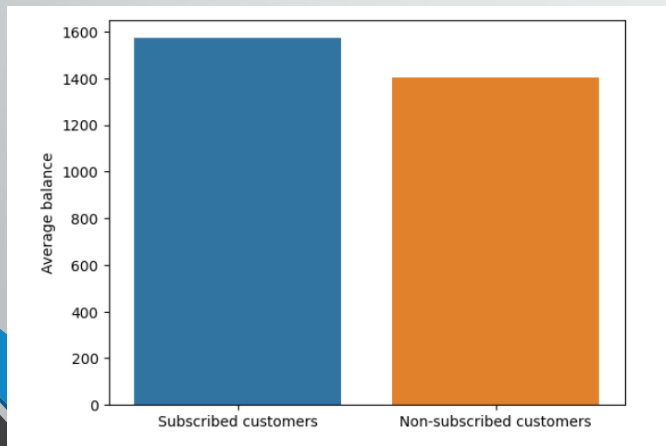
```
import pandas as pd
t=open('bank.csv','r').read().replace('"','') # opening the csv file, reading it and making
# modifications in the text
open('bank.csv','w').write(t) # writing the modified text into the csv file
df=pd.read_csv('bank.csv',sep=';') # reading the csv file using ';' as separator into a dataframe
df.head()
```

	age	job	marital	education	default	balance	housing	loan	contact	day	month	duration	campaign	pdays	pi
0	30	unemployed	married	primary	no	1787	no	no	cellular	19	oct	79	1	-1	
1	33	services	married	secondary	no	4789	yes	yes	cellular	11	may	220	1	339	
2	35	management	single	tertiary	no	1350	yes	no	cellular	16	apr	185	1	330	
3	30	management	married	tertiary	no	1476	yes	yes	unknown	3	jun	199	4	-1	
4	59	blue-collar	married	secondary	no	0	yes	no	unknown	5	may	226	1	-1	

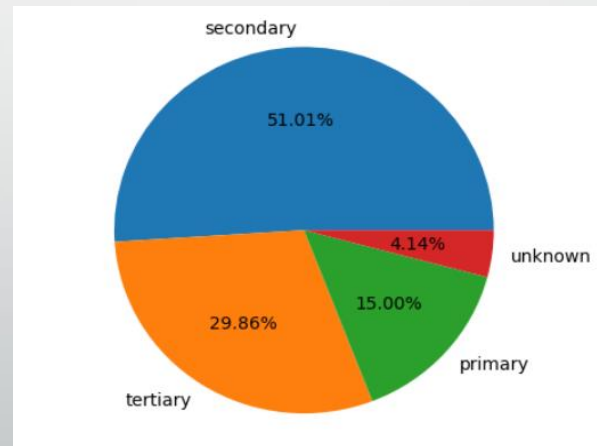
Data cleaning was not required as it had no wrongly identified datatypes, null values or missing values.

# PROBLEM SOLVING STEPS

- For finding out the average balance of the customers, we first used mean function on the balance attribute where 'y' is yes and no, i.e. the customer has taken the subscription or not. Then we plotted a bar plot using subscribed and non-subscribed customers on x-axis and mean balance on y-axis. We found out that the average balance of subscribed customers is 1571.96 Euros and of non-subscribed customers is 1403.21 Euros.



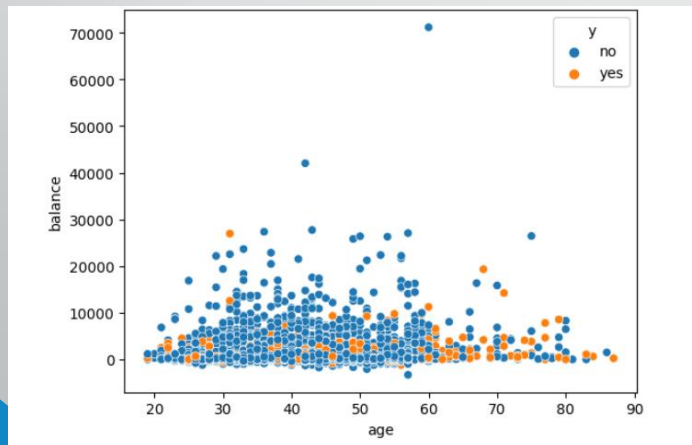
- We plotted a pie plot to understand the education background for the customers. We understood that 51.01% have secondary education, 29.86% have tertiary education, 15% have primary education and that of 4.14% customers is unknown.



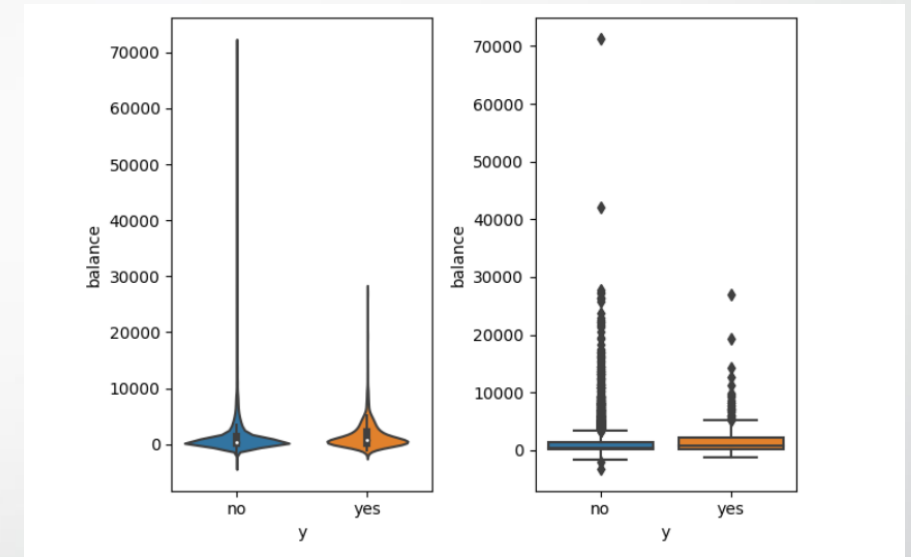
- We introduced a new column called season to understand how the customer engagement is there in each season. We applied 4 seasons according to the month such as winter, spring, summer and autumn. We analyzed that winter season has less customer engagement.

```
def season(m): # function which returns the s
    if m in ['dec', 'jan', 'feb']:
        return 'winter'
    if m in ['mar', 'apr', 'may']:
        return 'spring'
    if m in ['jun', 'jul', 'aug', 'sep']:
        return 'summer'
    return 'autumn'
df['season'] = df.month.apply(season) # using '
df.head()
```

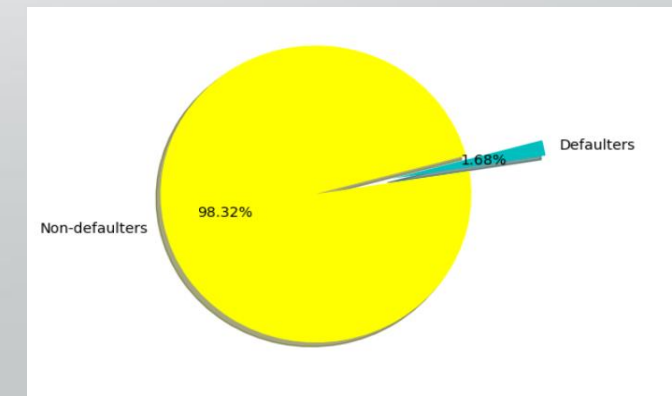
- It has been seen that more the number of pdays, more are the subscribers. But here in this case the customers who have -1 in pdays are the people who are not contacted before.
- After replacing -1 in pdays to NaN value we get the actual no. of customers who have recently been contacted and it is seen that subscription is more for recently contacted clients.
- The number of people who took subscription having balance less than zero are less.
- We found out the maximum balance for each job and we analyzed that retired customers have the maximum balance followed by the customers who are entrepreneurs. Unemployed customers and students have least balance.
- It has been seen that till the middle age around 60, balance is more but the no. of subscribed customers is less. Whereas above age 60 we can see there are a greater number of subscribers.



- We see that the people having maximum balance have less subscribers than the ones with medium balance. We can see that violin plot shows full distribution of the data and box plot shows the outliers of the data.



We analyzed that the people who have credit in account by default are less than the ones who don't. 98.32% customers do not have credit by default in their account.



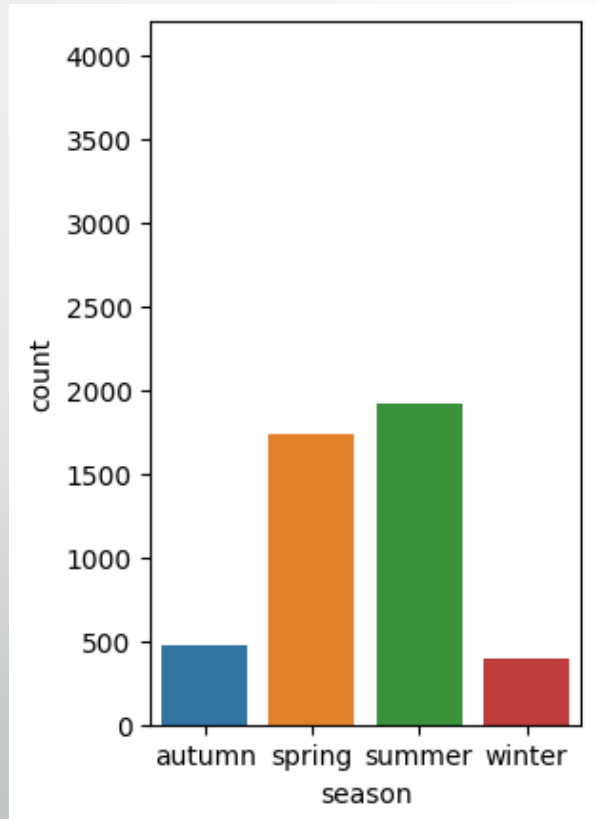
# WHAT COULD HAVE BEEN DONE BETTER?

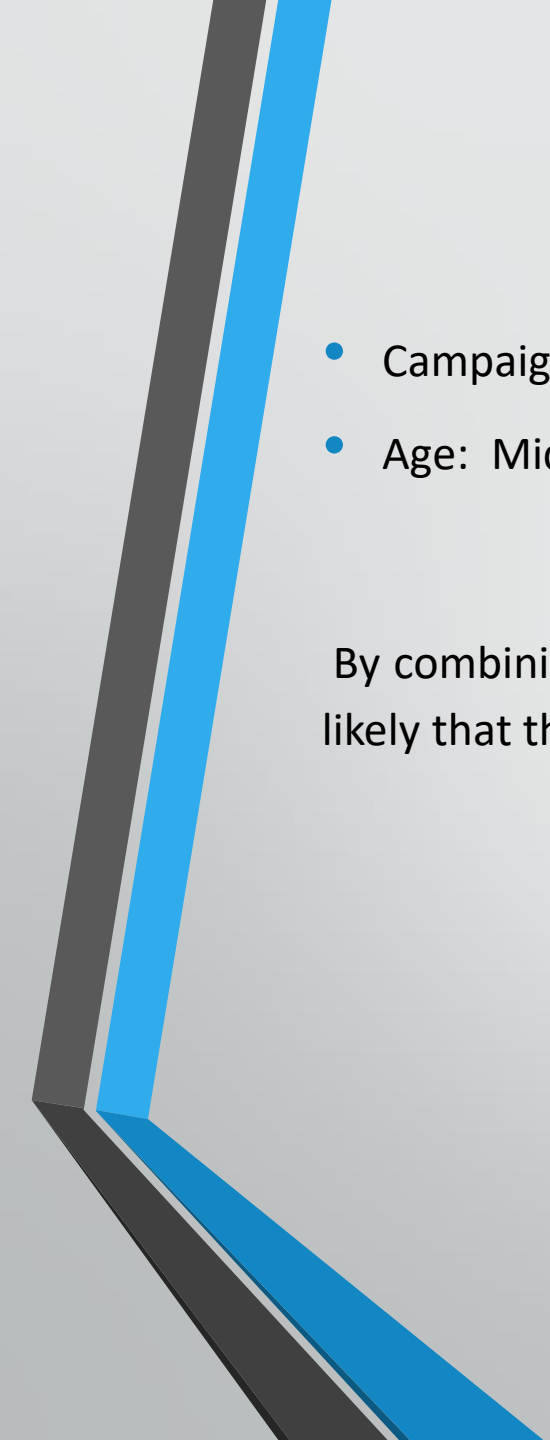
The 'day' column in the data could be converted into 'object' datatype from 'int64' as it could be considered a categorical column.

loan	contact	day	month	duration
no	cellular	19	oct	79
yes	cellular	11	may	220
no	cellular	16	apr	185
yes	unknown	3	jun	199
no	unknown	5	may	226

# TAKEAWAY AND CONCLUSIONS

- Seasonality: Least contacted months were of winter; they should have targeted more during winter.



- 
- Campaign calls should not be done more than thrice in order to save time and effort.
  - Age: Middle aged people should be targeted as they are the least ones to be subscribing.

By combining all these strategies and simplifying the market audience the next campaign should address, it is likely that the next marketing campaign of the bank will be more effective than the current one.

# FUTURE STEPS

- The data can be analyzed to find out how the number of customers with the subscription varies with the number of customers with an ongoing housing or personal loan so that more emphasis on customers in a particular category can be put accordingly.
- Using the 'duration' column, we can find out whether time spent on call with each customer affects the likeliness of the customer taking the term deposit.
- Similarly, it can be found out whether customers who are contacted a greater number of times are more or less likely to take the subscription.
- We can focus on customers who had taken the subscription in the past according to the data in the 'poutcome' column.
- Machine learning can be applied to make predictions from the data.



THANK YOU