Group Name: Data Science Group (Boshra and Omer) - Data Glacier

Name:

1- Boshra Abdualrahman Eisa

2- Omer Salih Dawood Omer

Email: bush.eisa9@gmail.com , omercomail@gmail.com

Country: Saudi Arabia

College/Company: Prince Sattam Bin Abdulaziz University

Specialization: Data Science

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.

Data understanding

The data is related with direct marketing campaigns of a Portuguese banking institution. The classification goal is to predict if the client will subscribe a term deposit.

Data downloaded from: https://archive.ics.uci.edu/ml/datasets/Bank+Marketing

It consists of four tables:

1) bank-additional-full.csv with all examples (41188) and 20 inputs, ordered by date.

2) bank-additional.csv with 10% of the examples (4119), randomly selected from 1), and 20 inputs.

3) bank-full.csv with all examples and 17 inputs, ordered by date (older version of this dataset with less inputs).

4) bank.csv with 10% of the examples and 17 inputs, randomly selected from 3 (older version of this dataset with less inputs).

Data Types:

Column	Data Type
Age, duration, campaign, pdays, emp.var.rate, cons.price.idx, cons.conf.idx, euribor3m, nr.employed	Numeric
Job, Marital, education, default, housing, loan, contact, month, day_of_week, poutcome	Categorical
y	Binary

What are the problems in the data (number of NA values, outliers, skewed etc.)?

1. Files need some processing by removing Semi-colons and convert file from single row to multiple.
2. Identifying any NA values:

```
Import pandas as pd

Import numpy as np

Import matplotlib.pyplot as plt

# Reading the data

df = pd.read_csv("bank-additional-full.csv")

print(df.shape)

print(df.info)
```

Output:

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 41188 entries, 0 to 41187
Data columns (total 21 columns):
 #    Column           Non-Null Count   Dtype
---   ------           --------------   -----
 0    age              41188 non-null   int64
 1    job              41188 non-null   object
 2    marital          41188 non-null   object
 3    education        41188 non-null   object
 4    default          41188 non-null   object
 5    housing          41188 non-null   object
 6    loan             41188 non-null   object
 7    contact          41188 non-null   object
 8    month            41188 non-null   object
 9    day_of_week      41188 non-null   object
 10   duration         41188 non-null   int64
 11   campaign         41188 non-null   int64
 12   pdays            41188 non-null   int64
 13   previous         41188 non-null   int64
 14   poutcome         41188 non-null   object
 15   emp.var.rate     41188 non-null   float64
 16   cons.price.idx   41188 non-null   float64
 17   cons.conf.idx    41188 non-null   float64
 18   euribor3m        41188 non-null   float64
 19   nr.employed      41188 non-null   float64
 20   y                41188 non-null   object
dtypes: float64(5), int64(5), object(11)
memory usage: 6.6+ MB
None
```

Applying this step to all files, we found no null values.

3. Check for outliers, by using the following code

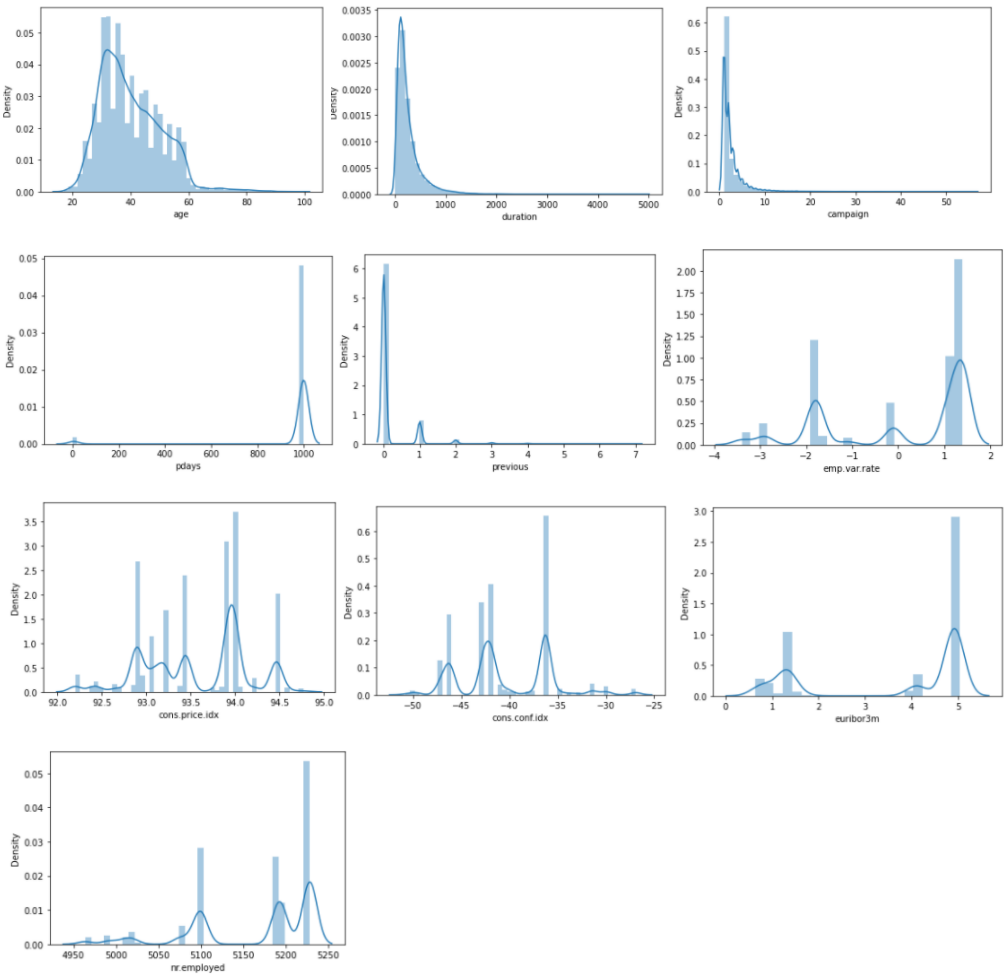
```
import pandas as pd

import numpy as np

import matplotlib.pyplot as plt

# Reading the data

df = pd.read_csv("bank-additional.csv")

df.describe()

Q1 = df.quantile(0.25)

Q3 = df.quantile(0.75)

IQR = Q3 - Q1

print(IQR)

print(df < (Q1 - 1.5 * IQR)) |(df > (Q3 + 1.5 * IQR))
```

After running above code, we found no outliners.

4. Checking for skewness, using the following code

```
for col in data:
        print(col)
        print(skew(data[col]))
        plt.figure()
        sns.distplot(data[col])
        plt.show()
```

Some of the columns were skewed, either positively or negatively.



What approaches you are trying to apply on your data set to overcome problems like NA value, outlier etc. and why?

We firstly started by removing all the semi-colons from the file to convert it into a multiple cell file. This was done so we can perform the analysis effectively and efficiently.

Secondly, since the data did not have any null values nor outliers, no approaches were necessary.

Lastly, we had to overcome the skewness of some of the columns by applying the square root of each column.