Bank Marketing (Campaign)

Group Name: Data Science Group (Bushra 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

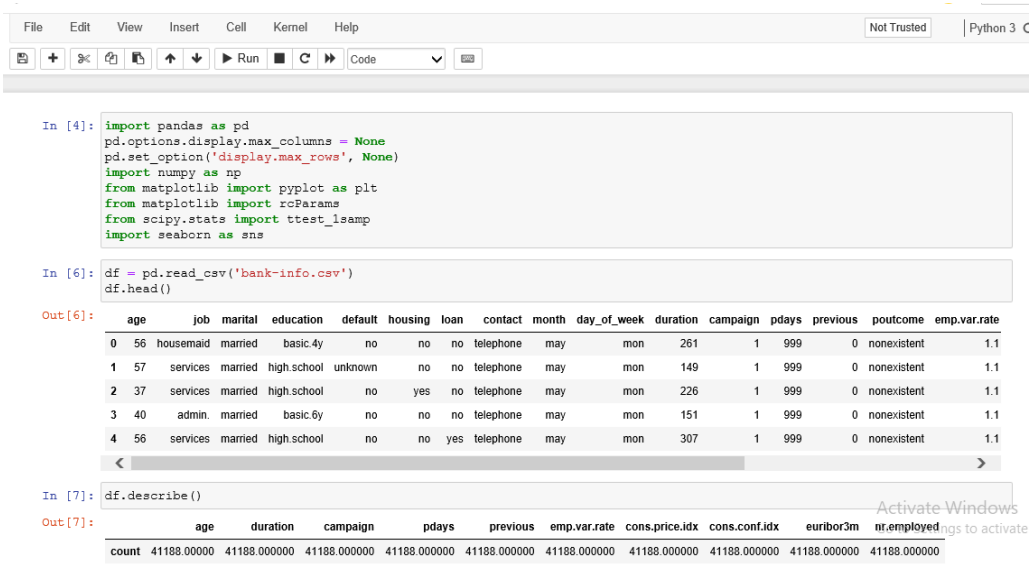
Specialization: Data Science

Exploratory Data Analysis Steps:

1. **Basics of Attributes in Dataset**
   - ✓ Determine the numeric and categorical attributes in the dataset
   - ✓ Basic Statistical Descriptions of Data - Mean, Median, Mode & Midrange
   - ✓ Dispersion of Data: Range, Quartiles, Variance, Standard Deviation, and Interquartile Range -
2. **Handling Missing Values in Dataset**
   - ✓ Handling missing values in the dataset
   - ✓ Handling missing values using Pandas & NumPy | Python Programming
3. **Outlier Detection**
   - ✓ Detecting and Filtering Outliers from Data
4. **Distribution of Datasets**
   - ✓ Handling a Skewed Dataset using Square Root Function

To perform EDA, we perform following steps:

1- Import libraries and show data description.



```
In [4]: import pandas as pd
pd.options.display.max_columns = None
pd.set_option('display.max_rows', None)
import numpy as np
from matplotlib import pyplot as plt
from matplotlib import rcParams
from scipy.stats import ttest_1samp
import seaborn as sns
```

```
In [6]: df = pd.read_csv('bank-info.csv')
df.head()
```

Out[6]:

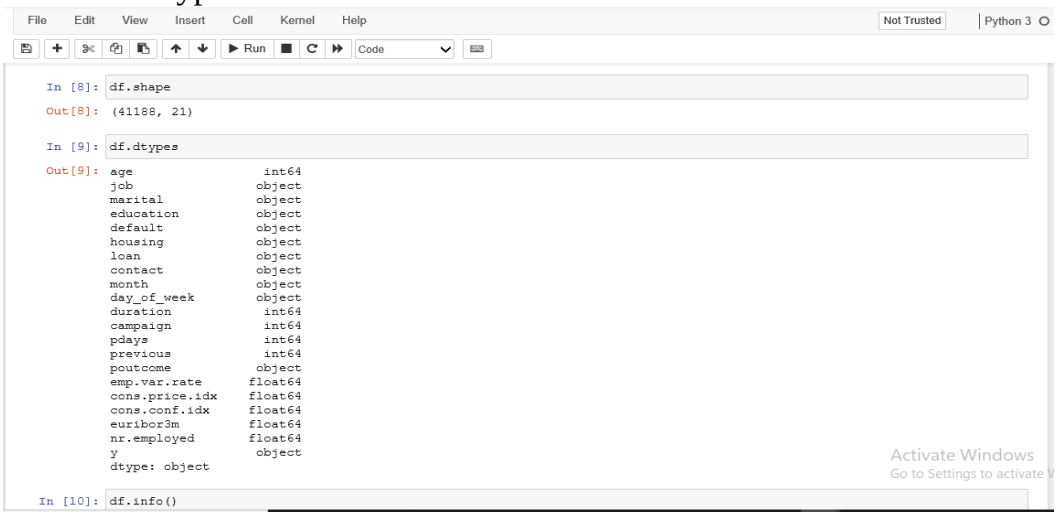
	age	job	marital	education	default	housing	loan	contact	month	day_of_week	duration	campaign	pdays	previous	poutcome	emp.var.rate
0	56	housemaid	married	basic.4y	no	no	no	telephone	may	mon	261	1	999	0	nonexistent	1.1
1	57	services	married	high.school	unknown	no	no	telephone	may	mon	149	1	999	0	nonexistent	1.1
2	37	services	married	high.school	no	yes	no	telephone	may	mon	226	1	999	0	nonexistent	1.1
3	40	admin.	married	basic.6y	no	no	no	telephone	may	mon	151	1	999	0	nonexistent	1.1
4	56	services	married	high.school	no	no	yes	telephone	may	mon	307	1	999	0	nonexistent	1.1

```
In [7]: df.describe()
```

Out[7]:

	age	duration	campaign	pdays	previous	emp.var.rate	cons.price.idx	cons.conf.idx	euribor3m	nr.employed
count	41188.00000	41188.000000	41188.000000	41188.000000	41188.000000	41188.000000	41188.000000	41188.000000	41188.000000	41188.000000

2- Check data types



```
In [8]: df.shape
Out[8]: (41188, 21)

In [9]: df.dtypes
Out[9]: age               int64
        job              object
        marital          object
        education        object
        default          object
        housing          object
        loan             object
        contact          object
        month            object
        day_of_week      object
        duration          int64
        campaign          int64
        pdays             int64
        previous          int64
        poutcome         object
        emp.var.rate    float64
        cons.price.idx  float64
        cons.conf.idx   float64
        euribor3m       float64
        nr.employed     float64
        y                object
        dtype: object

In [10]: df.info()
```


3- Check data information(null or not null)

jupyter EDA_Bank Last Checkpoint: Last Sunday at 6:15 PM (autosaved)

Logout

File Edit View Insert Cell Kernel Help

Not Trusted | Python 3

Run Code

```
In [10]: df.info()

<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
```


4- Print sum of null values

jupyter EDA_Bank Last Checkpoint: Last Sunday at 6:15 PM (autosaved)

Logout

File Edit View Insert Cell Kernel Help

Not Trusted | Python 3

Run Code

```
False

In [12]: print(df.isnull().sum())

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
cons.conf.idx     0
euribor3m         0
nr.employed       0
y                 0
dtype: int64
```

5- show data set relationship

jupyter EDA_Bank Last Checkpoint: Last Sunday at 6:15 PM (autosaved)

Logout

File Edit View Insert Cell Kernel Help

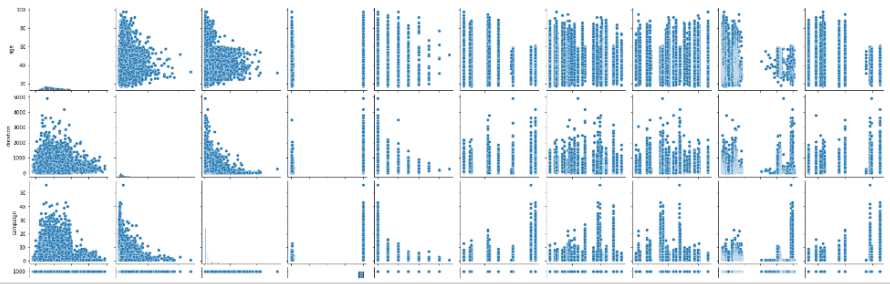
Not Trusted | Python 3

Run Code

```
nr.employed       0
y                 0
dtype: int64

In [13]: sns.pairplot(df)

Out[13]: <seaborn.axisgrid.PairGrid at 0x1f38002b1c0>
```



```
In [17]: sns.scatterplot(x='job', y='education', data=df, hue='education')
```

6- Draw heat map