# Capstone Project 1

**(Module - Basic Data Science)**

Perform exploratory data analysis on the given dataset using Python, Numpy, Pandas, Matplotlib, Seaborn.

**Bank Dataset**

*This dataset pertains to the direct marketing campaigns of a banking institution in Portugal. The campaigns primarily involved making phone calls to potential clients. Multiple contacts were often made with the same client to determine whether they would subscribe to a bank term deposit or not. The outcome was recorded as either 'yes' or 'no'.*

**Attribute/ Variable Information:**

age (numeric): The age of the client.

job (categorical): The type of job the client has, including categories such as 'admin.', 'bluecollar', 'entrepreneur', 'housemaid', 'management', 'retired', 'selfemployed', 'services', 'student', 'technician', 'unemployed', and 'unknown'.

marital (categorical): The marital status of the client, with categories including 'divorced', 'married', 'single', and 'unknown'. Note that 'divorced' also includes those who are widowed.

education (categorical): The level of education of the client, with categories such as 'basic.4y', 'basic.6y', 'basic.9y', 'high.school', 'illiterate', 'professional.course', 'university.degree', and 'unknown'.

default: Indicates whether the client has a credit in default, with categories 'no', 'yes', and 'unknown'.

balance: The average yearly balance of the client's bank account in euros (numeric).

housing: Indicates whether the client has a housing loan, with categories 'no', 'yes', and 'unknown'.

loan: Indicates whether the client has a personal loan, with categories 'no', 'yes', and 'unknown'.

contact: The type of contact communication used, with categories 'cellular' and 'telephone'.

day: The last contact day of the month (numeric 1-31).

month: The last contact month of the year, with categories 'jan', 'feb', 'mar', ..., 'nov', 'dec'.

duration: The duration of the last contact in seconds (numeric). Note that this attribute strongly affects the output target, but it's only known after the call is made and should be discarded for realistic predictive models.

campaign: The number of contacts performed during this campaign for the client (numeric, includes the last contact).

pdays: The number of days that passed after the client was last contacted from a previous campaign (numeric). A value of 999 indicates that the client was not previously contacted.

previous: The number of contacts performed before this campaign for the client (numeric).

poutcome: The outcome of the previous marketing campaign, with categories 'failure', 'nonexistent', and 'success'.

target: Indicates whether the client has subscribed to a term deposit, with binary values "yes" or "no".

The entries which are unknown are blank and may be considered as NaN values.

**Add Python comments to explain your code.**

**For each of the five sections given below, mention conclusions/observations drawn by you after each section.**

1. Section 1 (10 marks)
    a. Analyse data types of features and verify they hold data same as that of their datatype. Update if required.
    b. Check for Nan values in data and remove them using appropriate method, if any.
2. Section 2 (20 marks)
    a. Check for duplicates, view duplicated rows, and remove them, if any.
    b. Check for outliers using boxplot and statistical method, and remove them, if any.
    c. For categorical features, draw countplot. Write your observations.
    d. For numeric features, draw a histogram. Compute and about skewness of variables and apply transformation function, if needed.
3. Section 3 (20 marks)
    a. Come up with scatter plot with hue parameter. Write your observations from the plot.
    b. Compute correlation among independent features and demonstrate it using heatmap.
    c. Apply any scaling method to at least two independent features.
    d. Convert categorical features into numeric ones using appropriate encoding techniques.
4. Section 4: Compute correlation of each independent feature with dependent variable ‘target’. Select seven most important independent features from the set. (5 marks)

5. Section 5: Consider **'target'** column is prediction variable. Apply feature selection method to the dataset using SelectKBest() to reduce the dataset size to **7** features (5 marks).

**Note –**

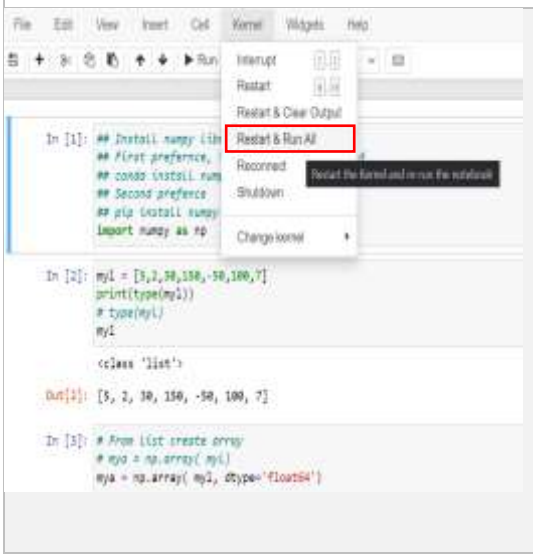
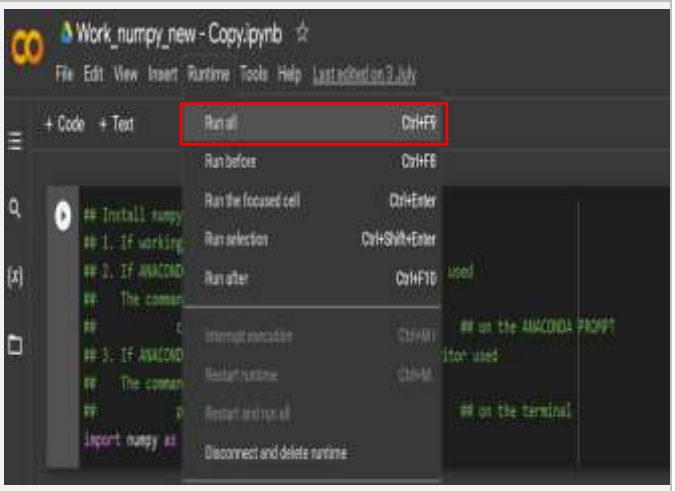
1. The comments and conclusions/observations carry marks for each section.

2. Students should submit their capstone project work in the form of .ipynb code file through mail to Dr. Hetal Gandhi (hetalg@regenesys.net).

Before submission verify that all cells are executed in the notebook from top to bottom-

**STEP A.** Address the solution to questions section-wise. Mention section and sub-section number before addressing it.

**STEP B.** Execute all cells all at-a-time at the end.

| **JUPYTER NOTEBOOK** | **COLAB** |
|---|---|
|  |  |

**STEP C.** Verify there is no error for any cell. If there is any error or any irrelevant output, address it and start from STEP-1.

**STEP D.** Download the file, once all is okay.

**STEP E.** Attach it to the mail, after renaming as per the instructions given.

3. Please rename the filename in the format as given below:

Name_of_student-Group_19-Capstone_project_1.ipynb.

(Add your full name in place of Name_of_student)

**4. The subject of the email should be same as the file name.**

5. Duration of the project work is set to one month from the date of project work announcement. Last date of submission will be **11th September 2023.**

## ADDITIONAL TIPS

If during implementation, you are facing any problem with reading your dataset in .csv format or uploading file in proper format:

I) IF YOU ARE USING **JUPYTER NOTEBOOK EDITOR**,

a. Please refer to the video recordings of **Basic Data Science pandas and matplotlib hands-on sessions wherein file access process is explained.** To know the path of where your python is installed, we use **pwd** command. At that path place your dataset file downloaded from LMS.

b. Then your file becomes accessible with these commands

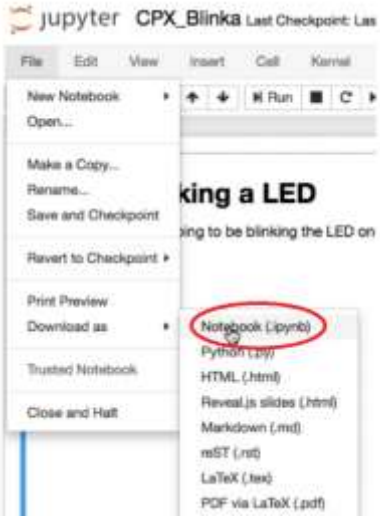
```
import pandas as pd
df = pd.read_csv('your_file_name.csv')

## If .csv (dataset) file is separated by ';' and not ','
df = pd.read_csv('your_file_name.csv', sep=';')
```

c. Work with it as per the problem statement.

d. To save your file properly in .ipynb format and download it.

Refer to this link- https://learn.adafruit.com/circuitpython-with-jupyter-notebooks/sharing-jupyter-notebooks, the details available will help you.




It will get downloaded in your download's folder/ default set for all downloads. Then share that file as attachment in the mail. The file format need to be .ipynb.

II) IF YOU ARE USING **COLAB as editor for python**, refer to this link to use .csv file by following all steps demonstrated.

To read a .csv file in Python with COLAB