

Assignment Date	28 September 2022
Student Name	M Hari Haran
Student Roll Number	913119104028
Maximum Marks	2 Marks

## ASSIGNMENT 2

### Question-1:

**Download the dataset**

### Question-2:

**Load the dataset :**

The screenshot shows a Google Colab notebook interface. The left sidebar displays the file explorer with a folder named 'sample\_data' containing a file 'Churn\_Modelling.csv'. The main code area shows the following code:

```
[4] import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns

df=pd.read_csv("Churn_Modelling.csv")

df
```

The output of the code is a preview of the dataset, showing the first five rows of a DataFrame with columns: RowNumber, CustomerId, Surname, CreditScore, Geography, Gender, Age, Tenure, Balance, NumOfProducts, HasCrCard, and IsActiveMember.

RowNumber	CustomerId	Surname	CreditScore	Geography	Gender	Age	Tenure	Balance	NumOfProducts	HasCrCard	IsActiveMember
0	1	Hargrave	619	France	Female	42	2	0.00	1	1	
1	2	Hill	608	Spain	Female	41	1	83807.86	1	0	
2	3	Onio	502	France	Female	42	8	159660.80	3	1	
3	4	Boni	699	France	Female	39	1	0.00	2	0	
4	5	Mitchell	850	Spain	Female	43	2	125510.82	1	1	

The notebook status bar at the bottom indicates that the code was completed at 6:46 AM on 01-10-2022.

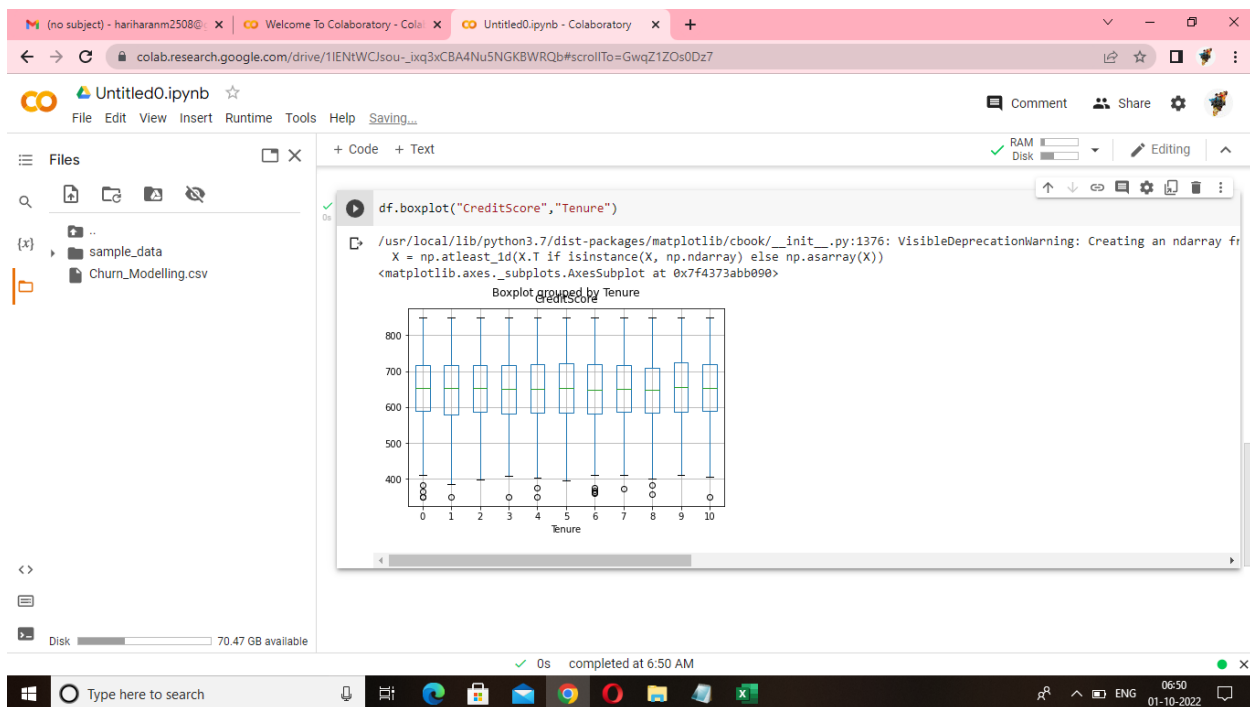
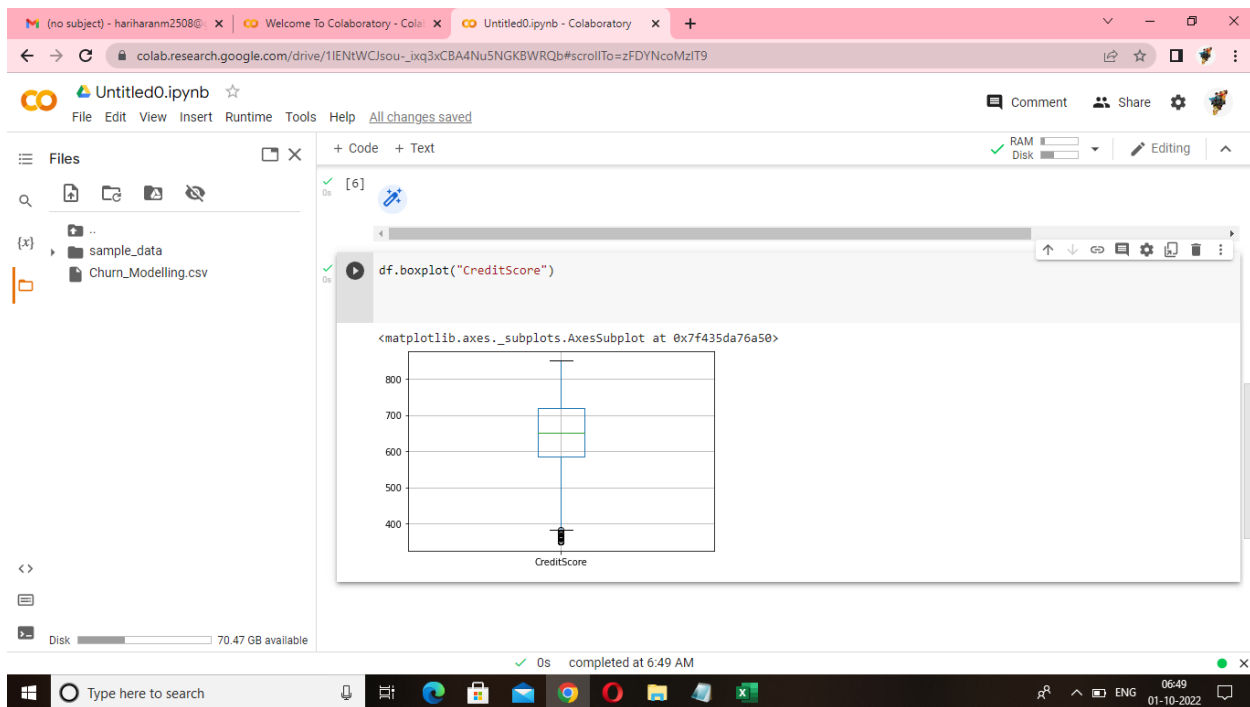
### Question-3:

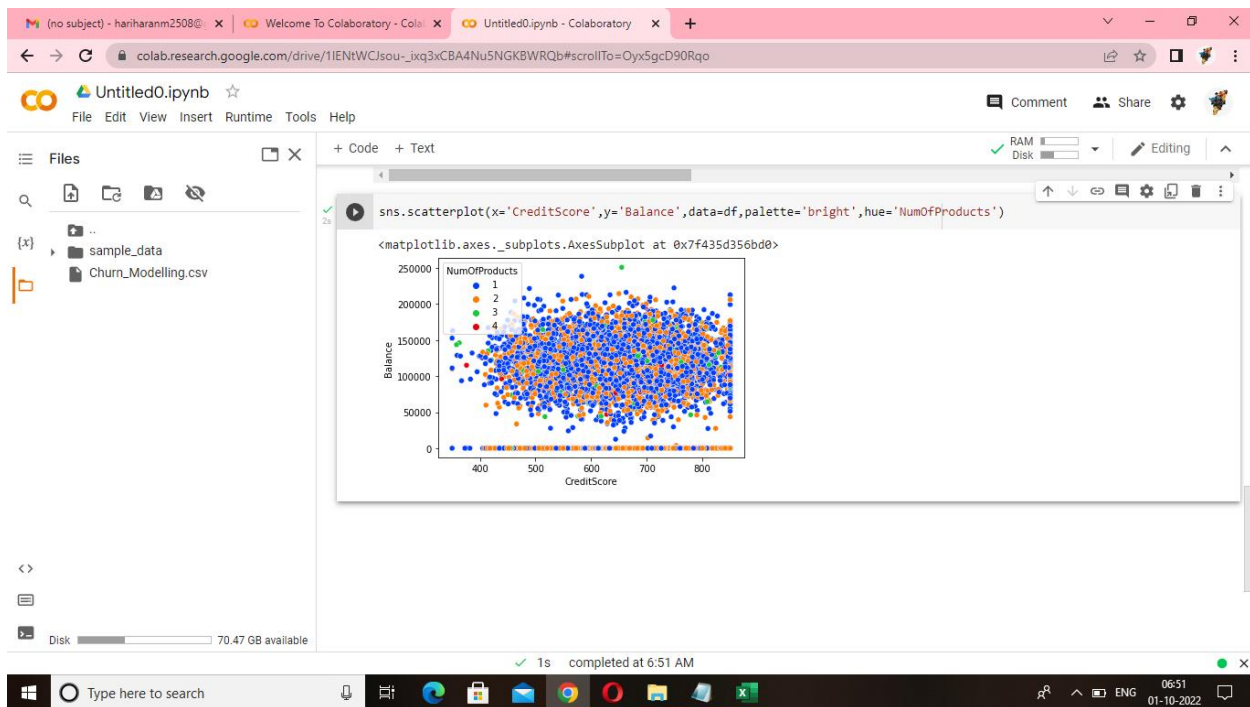
**Perform Below Visualizations.**

**Univariate Analysis**

**Bi – Variate Analysis**

**Mult – Variate Analysis**





#### Question – 4

Perform descriptive statistics on the dataset

Colaboratory interface showing the output of the `df.describe()` command.

```
df.describe()
```

	RowNumber	CustomerId	CreditScore	Age	Tenure	Balance	NumOfProducts	HasCrCard	IsActive
count	10000.00000	1.000000e+04	10000.000000	10000.000000	10000.000000	10000.000000	10000.000000	10000.000000	10000.000000
mean	5000.50000	1.569094e+07	650.528800	38.921800	5.012800	76485.889288	1.530200	0.70550	0.815900
std	2886.89568	7.193619e+04	96.653299	10.487806	2.892174	62397.405202	0.581654	0.45584	0.385690
min	1.00000	1.56570e+07	350.000000	18.000000	0.000000	0.000000	1.000000	0.00000	0.00000
25%	2500.75000	1.562853e+07	584.000000	32.000000	3.000000	0.000000	1.000000	0.00000	0.00000
50%	5000.50000	1.569074e+07	652.000000	37.000000	5.000000	97198.540000	1.000000	1.00000	0.81590
75%	7500.25000	1.575323e+07	718.000000	44.000000	7.000000	127644.240000	2.000000	1.00000	0.81590
max	10000.00000	1.581569e+07	850.000000	92.000000	10.000000	250898.090000	4.000000	1.00000	0.81590

The Colaboratory interface shows the file explorer on the left with 'sample\_data' and 'Churn\_Modelling.csv'. The status bar at the bottom indicates the code was completed at 6:52 AM.

## Question-5

### Handle the Missing Values

The screenshot shows a Google Colab interface with a Jupyter Notebook titled 'Untitled0.ipynb'. The code cell contains the command `df.fillna(5)`, which has been executed. The output is a pandas DataFrame with 10,000 rows and 14 columns. The columns are: RowNumber, CustomerId, Surname, CreditScore, Geography, Gender, Age, Tenure, Balance, NumOfProducts, HasCrCard, and IsActiveMember. The DataFrame shows a subset of rows, including rows 0, 1, 2, 3, 4, and rows 9995 through 9999. The 'Balance' column contains missing values (0.00) for rows 0, 3, 4, 9995, 9996, 9997, and 9998. The 'HasCrCard' column contains missing values (0) for rows 1, 3, 4, 9995, 9996, 9997, and 9998. The 'IsActiveMember' column contains missing values (0) for rows 1, 3, 4, 9995, 9996, 9997, and 9998. The DataFrame is displayed in a table format with a scrollbar on the right.

	RowNumber	CustomerId	Surname	CreditScore	Geography	Gender	Age	Tenure	Balance	NumOfProducts	HasCrCard	IsActiveMember
0	1	15634602	Hargrave	619	France	Female	42	2	0.00	1	1	1
1	2	15647311	Hill	608	Spain	Female	41	1	83807.86	1	0	0
2	3	15619304	Onio	502	France	Female	42	8	159660.80	3	1	1
3	4	15701354	Boni	699	France	Female	39	1	0.00	2	0	0
4	5	15737888	Mitchell	850	Spain	Female	43	2	125510.82	1	1	1
...	...	...	...	...	...	...	...	...	...	...	...	...
9995	9996	15606229	Obijaku	771	France	Male	39	5	0.00	2	1	1
9996	9997	15569892	Johnstone	516	France	Male	35	10	57369.61	1	1	1
9997	9998	15584532	Liu	709	France	Female	36	7	0.00	1	0	0
9998	9999	15682355	Sabbatini	772	Germany	Male	42	3	75075.31	2	1	1
9999	10000	15628319	Walker	792	France	Female	28	4	130142.79	1	1	1

## Question-6:

### Find the outline & replace the outliers

The screenshot shows a Google Colab interface with a Jupyter Notebook titled 'Untitled0.ipynb'. The code cell contains the command `out = df.drop(columns=['Gender', 'Tenure', 'HasCrCard', 'IsActiveMember', 'NumOfProducts', 'Exited']).quantile(q=[0.25, 0.50])`, which has been executed. The output is a pandas DataFrame with 2 rows and 7 columns. The columns are: RowNumber, CustomerId, CreditScore, Age, Balance, EstimatedSalary, and an unnamed column. The DataFrame shows the 0.25 and 0.50 quantiles for the specified columns. The 'Balance' column contains missing values (0.00) for the 0.25 quantile row. The 'EstimatedSalary' column contains missing values (0.00) for the 0.25 quantile row. The DataFrame is displayed in a table format with a scrollbar on the right.

	RowNumber	CustomerId	CreditScore	Age	Balance	EstimatedSalary
0.25	2500.75	15628528.25	584.0	32.0	0.00	51002.110
0.50	5000.50	15690738.00	652.0	37.0	97198.54	100193.915

Colaboratory interface showing a Jupyter Notebook session. The browser tabs include "Welcome To Colaboratory - Colo" and "Untitled0.ipynb - Colaboratory". The URL is [colab.research.google.com/drive/1IENtWCJsou-\\_ixq3xCBA4Nu5NGKBWRQb#scrollTo=tSYh9daU1WLb](https://colab.research.google.com/drive/1IENtWCJsou-_ixq3xCBA4Nu5NGKBWRQb#scrollTo=tSYh9daU1WLb).

The notebook is titled "Untitled0.ipynb" and shows the following code cell:

```
Q1 = out.iloc[0]
Q2=out.iloc[1]
iqr=Q2-Q1
iqr
```

The output of the code cell is a Pandas Series:

RowNumber	2499.750
CustomerId	62209.750
CreditScore	68.000
Age	5.000
Balance	97198.540
EstimatedSalary	49191.805
dtype:	float64

The interface also shows a file explorer on the left with "sample\_data" and "Churn\_Modelling.csv". The status bar indicates "completed at 6:56 AM".

Colaboratory interface showing a Jupyter Notebook session. The browser tabs include "Welcome To Colaboratory - Colo" and "Untitled0.ipynb - Colaboratory". The URL is [colab.research.google.com/drive/1IENtWCJsou-\\_ixq3xCBA4Nu5NGKBWRQb#scrollTo=tVZzKPK71k5\\_](https://colab.research.google.com/drive/1IENtWCJsou-_ixq3xCBA4Nu5NGKBWRQb#scrollTo=tVZzKPK71k5_).

The notebook is titled "Untitled0.ipynb" and shows the following code cell:

```
upper = out.iloc[1]+1.5*iqr
upper
```

The output of the code cell is a Pandas Series:

RowNumber	8.750125e+03
CustomerId	1.578405e+07
CreditScore	7.540000e+02
Age	4.450000e+01
Balance	2.429964e+05
EstimatedSalary	1.739816e+05
dtype:	float64

The interface also shows a file explorer on the left with "sample\_data" and "Churn\_Modelling.csv". The status bar indicates "completed at 6:57 AM".

Colaboratory interface showing a Jupyter Notebook. The browser tabs include "Welcome To Colaboratory - Colo" and "Untitled0.ipynb - Colaboratory". The URL is [colab.research.google.com/drive/1IENtWCJsou-ixq3xCBA4Nu5NGKBWRQb#scrollTo=UEr2eo2J1r4s](https://colab.research.google.com/drive/1IENtWCJsou-ixq3xCBA4Nu5NGKBWRQb#scrollTo=UEr2eo2J1r4s).

The notebook interface shows the "Files" pane on the left with a folder named "sample\_data" containing a file "Churn\_Modelling.csv". The "Code" pane contains the following code:

```
lower = out.iloc[0]-1.5*iqr
lower
```

The output of the code is displayed as a table:

RowNumber	-1.248875e+03
CustomerId	1.553521e+07
CreditScore	4.820000e+02
Age	2.450000e+01
Balance	-1.457978e+05
EstimatedSalary	-2.278560e+04
dtype:	float64

The status bar at the bottom indicates "completed at 6:57 AM".

Colaboratory interface showing a Jupyter Notebook. The browser tabs include "Welcome To Colaboratory - Colo" and "Untitled0.ipynb - Colaboratory". The URL is [colab.research.google.com/drive/1IENtWCJsou-ixq3xCBA4Nu5NGKBWRQb#scrollTo=ZOhCqJ6m10Ca](https://colab.research.google.com/drive/1IENtWCJsou-ixq3xCBA4Nu5NGKBWRQb#scrollTo=ZOhCqJ6m10Ca).

The notebook interface shows the "Files" pane on the left with a folder named "sample\_data" containing a file "Churn\_Modelling.csv". The "Code" pane contains the following code:

```
[18] lower = out.iloc[0]-1.5*iqr
lower

df['CreditScore'] = np.where(df['CreditScore'] > 756, 650.5288, df['CreditScore'])
df['Age'] = np.where(df['Age'] > 62, 38.9218, df['Age'])
```

The output of the code is displayed as a table:

RowNumber	-1.248875e+03
CustomerId	1.553521e+07
CreditScore	4.820000e+02
Age	2.450000e+01
Balance	-1.457978e+05
EstimatedSalary	-2.278560e+04
dtype:	float64

The status bar at the bottom indicates "completed at 6:58 AM".

### Question-7:

Check for Categorical columns and Perform encoding

The screenshot shows a Google Colab environment. The browser tabs include 'Welcome To Colaboratory' and 'Untitled0.ipynb - Colaboratory'. The address bar shows a Google Drive link. The Colab interface has a top bar with 'File', 'Edit', 'View', 'Insert', 'Runtime', 'Tools', and 'Help' menus. On the left, a file explorer shows a folder 'sample\_data' containing 'Churn\_Modelling.csv'. The main area is a code cell with the following Python code:

```
from sklearn.preprocessing import OneHotEncoder
e= OneHotEncoder(sparse=False)
e= e.fit_transform(df)
e

array([[1., 0., 0., ..., 0., 0., 1.],
       [0., 1., 0., ..., 0., 1., 0.],
       [0., 0., 1., ..., 0., 0., 1.],
       ...,
       [0., 0., 0., ..., 0., 0., 1.],
       [0., 0., 0., ..., 0., 0., 1.],
       [0., 0., 0., ..., 0., 1., 0.]])
```

The code cell is executed, showing a green checkmark and '3s' completion time. The bottom status bar indicates 'completed at 6:59 AM'.

## Question-8

Split the data into dependant and independent Variables.

The screenshot shows a Google Colab environment. The browser tabs include 'Welcome To Colaboratory' and 'Untitled0.ipynb - Colaboratory'. The address bar shows a Google Drive link. The Colab interface has a top bar with 'File', 'Edit', 'View', 'Insert', 'Runtime', 'Tools', and 'Help' menus. On the left, a file explorer shows a folder 'sample\_data' containing 'Churn\_Modelling.csv'. The main area is a code cell with the following Python code:

```
x=df.iloc[:, :-1].values
x

array([[1, 15634602, 'Hargrave', ..., 1, 1, 101348.88],
       [2, 15647311, 'Hill', ..., 0, 1, 112542.58],
       [3, 15619304, 'Onio', ..., 1, 0, 113931.57],
       ...,
       [9998, 15584532, 'Liu', ..., 0, 1, 42085.58],
       [9999, 15682355, 'Sabbatini', ..., 1, 0, 92888.52],
       [10000, 15628319, 'Walker', ..., 1, 0, 38190.78]], dtype=object)
```

The code cell is executed, showing a green checkmark and '0s' completion time. The bottom status bar indicates 'completed at 7:00 AM'.

## Question-9

Scale the independent Variables

Colaboratory interface showing a Jupyter Notebook with the following code:

```
from sklearn.preprocessing import StandardScaler
df.head()
```

The output displays the first five rows of the dataset:

	RowNumber	CustomerId	Surname	CreditScore	Geography	Gender	Age	Tenure	Balance	NumOfProducts	HasCrCard	IsActiveMember
0	1	15634602	Hargrave	619.0000	France	Female	42.0	2	0.00	1	1	
1	2	15647311	Hill	608.0000	Spain	Female	41.0	1	83807.86	1	0	
2	3	15619304	Onio	502.0000	France	Female	42.0	8	159660.80	3	1	
3	4	15701354	Boni	699.0000	France	Female	39.0	1	0.00	2	0	
4	5	15737888	Mitchell	650.5288	Spain	Female	43.0	2	125510.82	1	1	

The interface also shows the file explorer on the left with 'sample\_data' and 'Churn\_Modelling.csv' files. The bottom status bar indicates 'completed at 7:02 AM'.

## Question-10

### Split the data into training & testing

Colaboratory interface showing a Jupyter Notebook with the following code:

```
from sklearn.model_selection import train_test_split
x_train,x_test,y_train,y_test = train_test_split(x,y, random_state=0, train_size=.75)

print(x_train.shape)
print(x_test.shape)
print(y_train.shape)
print(y_test.shape)
```

The output displays the shapes of the training and testing sets:

```
(7500, 13)
(2500, 13)
(7500,)
(2500,)
```

The interface also shows the file explorer on the left with 'sample\_data' and 'Churn\_Modelling.csv' files. The bottom status bar indicates 'completed at 7:03 AM'.



