Practical No.: 02

2) Data Wrangling II

Create an "Academic performance" dataset of students and perform the following operations usingPython.

1. **Scan all variables for missing values and inconsistencies. If there are missing values and/or inconsistencies, use any of the suitable techniques to deal with them.**
2. **Scan all numeric variables for outliers. If there are outliers, use any of the suitable techniquesto deal with them.**
3. **Apply data transformations on at least one of the variables. The purpose of this transformation should be one of the following reasons: to change the scale for better understanding of the variable, to convert a non-linear relation into a linear one, or to decrease the skewness and convert the distribution into a normal distribution.**

Reason and document your approach properly.

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
%matplotlib inline
```

```
data=pd.read_csv("tecdiv.csv")
```

```
data
```

3]:

	Timestamp	Email Address	Name	Email	Roll no	PRN No.	Mobile No.	First year: Sem 1	First year: Sem 2	Se
0	1/17/2022 12:45:09	sejal.zambare19@pccoepune.org	Sejal Zambare	sejal.zambare19@gmail.com	TECOC359	72026841K	8208217782	8.40	8.60	
1	1/17/2022 12:45:44	rushikesh.thorat19@pccoepune.org	Rushikesh Vilas Thorat	rushikesh.thorat19@pccoepune.org	TECOC347	72026776F	9021261925	8.14	8.14	
2	1/17/2022 12:46:10	atharv.sontakke19@pccoepune.org	Atharv Sontakke	atharv123sontakke@gmail.com	TECOC340	72026742M	9009804629	6.61	6.61	
3	1/17/2022 12:46:21	amisha.sherekar19@pccoepune.org	Amisha Sunil Sherekar	amisha.sherekar19@pccoepune.org	TECOC328	72026696D	8698227548	7.20	7.30	
4	1/17/2022 12:46:31	saurabh.sawardekar19@pccoepune.org	Saurabh Raju Sawardekar	saurabh.sawardekar19@pccoepune.org	TECOC326	72026682D	7774072850	7.05	7.45	

```
print("The last five rows are as follows: ")
data.tail()
```

```
The last five rows are as follows:
```

:

	Timestamp	Email Address	Name
59	1/20/2022 9:24:40	pratik.meshram20@pccoepune.org	Pratik Amrut Meshram
60	1/20/2022 9:36:14	prasad.zore19@pccoepune.org	Prasad Zore
61	1/20/2022 9:42:34	sudhir.varu19@pccoepune.org	SUDHIR VARU
			Bhagyashree

```
data.describe()
```

[6]:

	Mobile No.	First year: Sem 1	First year: Sem 2	Second year: Sem 1	Second year: Sem 2
count	6.400000e+01	64.000000	64.000000	64.000000	64.000000
mean	8.623097e+09	8.834219	9.095469	9.292031	9.377187
std	9.132070e+08	11.187839	11.171986	0.528523	0.495185
min	7.028870e+09	0.000000	0.000000	6.900000	7.200000
25%	7.766559e+09	7.237500	7.655000	9.050000	9.140000
50%	8.805720e+09	8.260000	8.400000	9.445000	9.450000
75%	9.335094e+09	8.802500	9.115000	9.645000	9.725000
max	9.975810e+09	95.000000	95.000000	9.910000	9.950000

```
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 64 entries, 0 to 63
Data columns (total 11 columns):
 #   Column                 Non-Null Count  Dtype
---  ------                 --------------  -----
 0   Timestamp              64 non-null     object
 1   Email Address          64 non-null     object
 2   Name                   64 non-null     object
 3   Email                  64 non-null     object
 4   Roll no                64 non-null     object
 5   PRN No.                64 non-null     object
 6   Mobile No.             64 non-null     int64
 7   First year:   Sem 1    64 non-null     float64
 8   First year:   Sem 2    64 non-null     float64
 9   Second year:   Sem 1   64 non-null     float64
 10  Second year:   Sem 2   64 non-null     float64
dtypes: float64(4), int64(1), object(6)
memory usage: 5.6+ KB
```

```
print("The column names of the dataset are as follows:")
data.columns
```

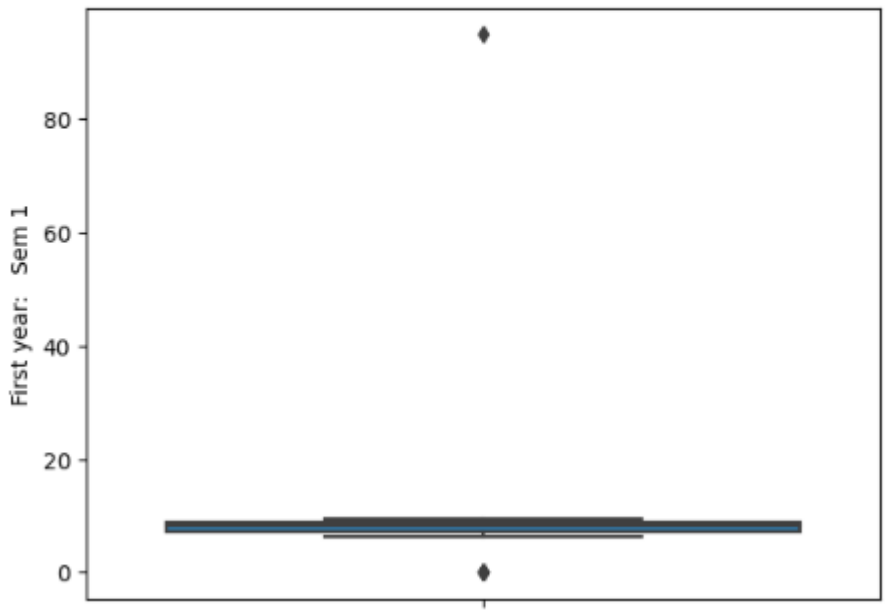
```
The column names of the dataset are as follows:

]: Index(['Timestamp', 'Email Address', 'Name', 'Email', 'Roll no ', 'PR
       'Mobile No.', 'First year:   Sem 1', 'First year:   Sem 2',
       'Second year:   Sem 1', 'Second year:   Sem 2'],
      dtype='object')
```

```
data.isnull().sum()
```

```
]: Timestamp              0
   Email Address          0
   Name                   0
   Email                  0
   Roll no                0
   PRN No.                0
   Mobile No.             0
```

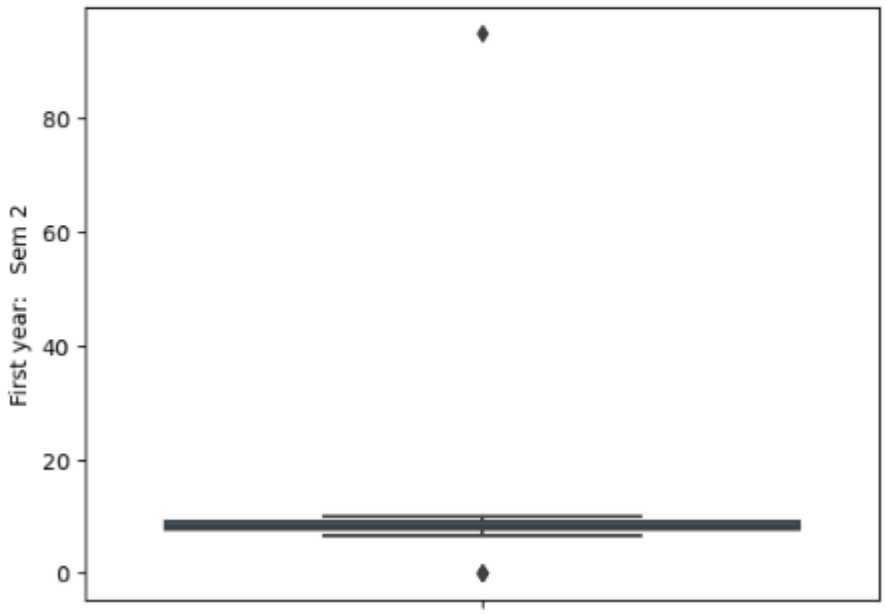
```
sns.boxplot(y=data['First year:   Sem 1'])
```

```
15]: <Axes: ylabel='First year:   Sem 1'>
```



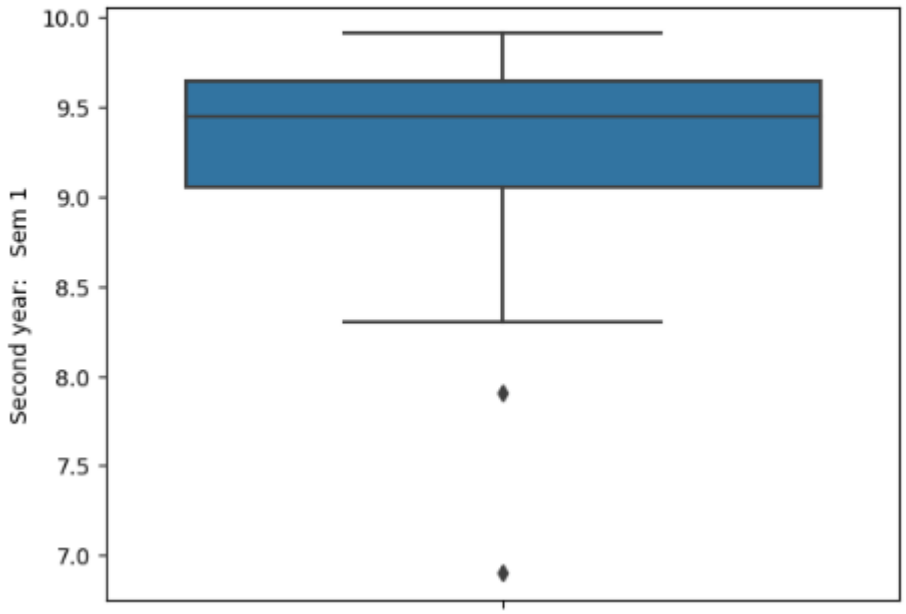
```
sns.boxplot(y=data['First year:   Sem 2'])
```

```
8]: <Axes: ylabel='First year:   Sem 2'>
```




```
sns.boxplot(y=data['Second year:   Sem 1'])
```

```
9]: <Axes: ylabel='Second year:   Sem 1'>
```



```
sns.boxplot(y=data['Second year:   Sem 2'])
```

```
0]: <Axes: ylabel='Second year:   Sem 2'>
```