Welcome to 30 Days | ML | Day 19 ¶

Import Library

In [1]:

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

Import Dataset

In [2]:

```
df = pd.read_csv('placement.csv')
```

In [3]:

```
df
```

Out[3]:

	cgpa	placement_exam_marks	placed
0	7.19	26.0	1
1	7.46	38.0	1
2	7.54	40.0	1
3	6.42	8.0	1
4	7.23	17.0	0
...	...	...	...
995	8.87	44.0	1
996	9.12	65.0	1
997	4.89	34.0	0
998	8.62	46.0	1
999	4.90	10.0	1

1000 rows × 3 columns

In [4]:

```
df.sample(10)
```

Out[4]:

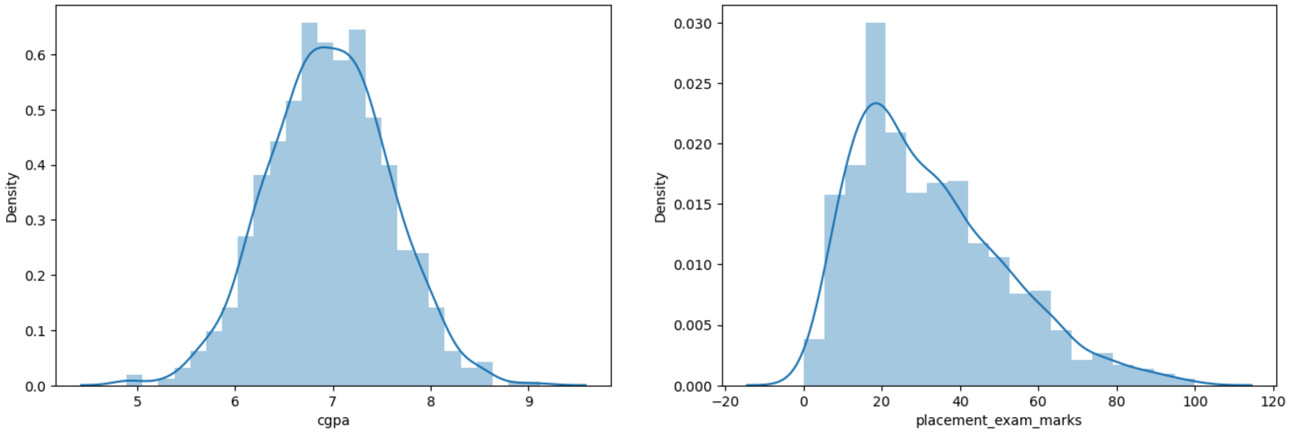
	cgpa	placement_exam_marks	placed
636	6.39	43.0	1
146	6.75	22.0	1
369	6.69	36.0	1
542	7.06	22.0	0
557	6.47	25.0	0
280	6.62	55.0	0
317	7.47	19.0	0
809	6.39	22.0	1
916	6.88	11.0	1
704	6.91	45.0	1

Plot Show in CGPA and Placement Marks

In [6]:

```
plt.figure(figsize=(16,5))
plt.subplot(1,2,1)
sns.distplot(df['cgpa'])

plt.subplot(1,2,2)
sns.distplot(df['placement_exam_marks'])

plt.show()
```

```
C:\Users\ASUS\AppData\Local\Temp\ipykernel_16968\40043834.py:3: UserWarning:

`distplot` is a deprecated function and will be removed in seaborn v0.14.0.

Please adapt your code to use either `displot` (a figure-level function with
similar flexibility) or `histplot` (an axes-level function for histograms).

For a guide to updating your code to use the new functions, please see
https://gist.github.com/mwaskom/de44147ed2974457ad6372750bbe5751 (https://gist.githu
b.com/mwaskom/de44147ed2974457ad6372750bbe5751)

  sns.distplot(df['cgpa'])
C:\Users\ASUS\AppData\Local\Temp\ipykernel_16968\40043834.py:6: UserWarning:

`distplot` is a deprecated function and will be removed in seaborn v0.14.0.

Please adapt your code to use either `displot` (a figure-level function with
similar flexibility) or `histplot` (an axes-level function for histograms).

For a guide to updating your code to use the new functions, please see
https://gist.github.com/mwaskom/de44147ed2974457ad6372750bbe5751 (https://gist.githu
b.com/mwaskom/de44147ed2974457ad6372750bbe5751)

  sns.distplot(df['placement_exam_marks'])
```



#Describe placement marks

In [9]:

```
df['placement_exam_marks'].describe()
```

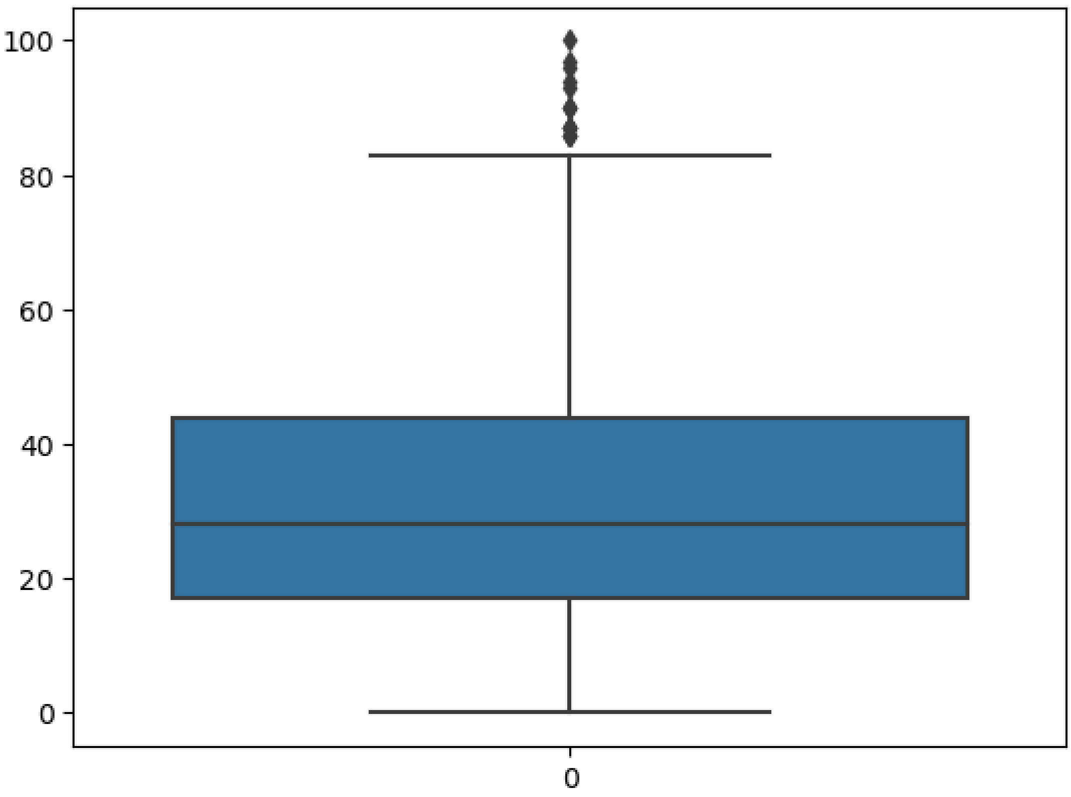
Out[9]:

```
count    1000.000000
mean       32.225000
std        19.130822
min         0.000000
25%        17.000000
50%        28.000000
75%        44.000000
max       100.000000
Name: placement_exam_marks, dtype: float64
```

Draw Box Plot

In [10]:
```
sns.boxplot(df['placement_exam_marks'])
```

Out[10]:
```
<Axes: >
```



Finding IQR Value

In [12]:
```
percentile25 = df['placement_exam_marks'].quantile(0.25)
percentile75 = df['placement_exam_marks'].quantile(0.75)
```

In [13]:
```
percentile25
```

Out[13]:
```
17.0
```

In [14]:
```
percentile75
```

Out[14]:
```
44.0
```

Calculate IQR (Q3-Q1)

In [16]:
```
iqr = percentile75 - percentile25
```

In [17]:
```
iqr
```

Out[17]:
```
27.0
```

Calculate Upper and Lower Limit:

In [18]:

```
upper_limit = percentile75 + 1.5 * iqr
lower_limit = percentile25 - 1.5 * iqr
```

In [19]:

```
print("Upper limit",upper_limit)
```

```
Upper limit 84.5
```

In [20]:

```
print("Lower limit",lower_limit)
```

```
Lower limit -23.5
```

Finding Outliers in Upper Limit:

In [21]:

```
df[df['placement_exam_marks'] > upper_limit]
```

Out[21]:

	cgpa	placement_exam_marks	placed
9	7.75	94.0	1
40	6.60	86.0	1
61	7.51	86.0	0
134	6.33	93.0	0
162	7.80	90.0	0
283	7.09	87.0	0
290	8.38	87.0	0
311	6.97	87.0	1
324	6.64	90.0	0
630	6.56	96.0	1
685	6.05	87.0	1
730	6.14	90.0	1
771	7.31	86.0	1
846	6.99	97.0	0
917	5.95	100.0	0

In [22]:

```
df[df['placement_exam_marks'] > upper_limit].shape
```

Out[22]:

```
(15, 3)
```

Finding Outliers in Lower Limit:

In [23]:

```
df[df['placement_exam_marks'] < lower_limit]
```

Out[23]:

cgpa	placement_exam_marks	placed

Apply Trimming Method - 1:

```
In [25]: new_df = df[df['placement_exam_marks'] < upper_limit]
```

```
In [26]: new_df.shape
```

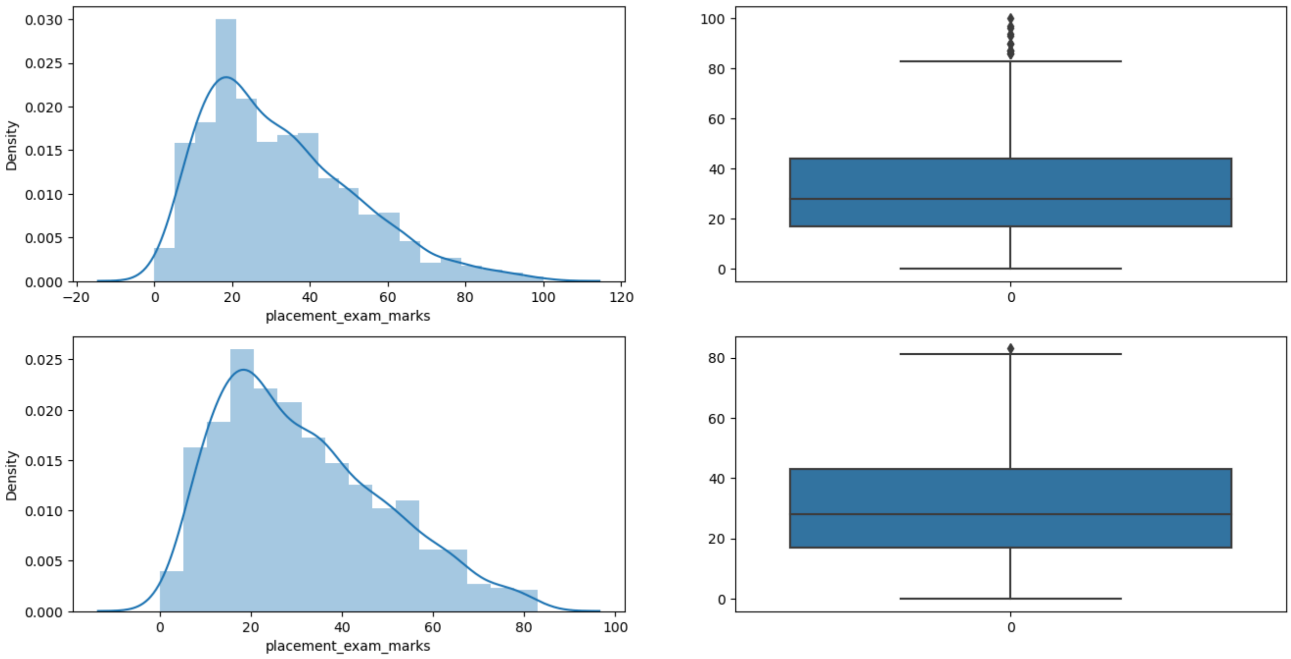
```
Out[26]: (985, 3)
```

Compare Before and After (After Trimming):

In [27]:

```
plt.figure(figsize=(16,8))
plt.subplot(2,2,1)
sns.distplot(df['placement_exam_marks'])

plt.subplot(2,2,2)
sns.boxplot(df['placement_exam_marks'])

plt.subplot(2,2,3)
sns.distplot(new_df['placement_exam_marks'])

plt.subplot(2,2,4)
sns.boxplot(new_df['placement_exam_marks'])

plt.show()
```

```
C:\Users\ASUS\AppData\Local\Temp\ipykernel_16968\3858278419.py:3: UserWarning:

`distplot` is a deprecated function and will be removed in seaborn v0.14.0.

Please adapt your code to use either `displot` (a figure-level function with
similar flexibility) or `histplot` (an axes-level function for histograms).

For a guide to updating your code to use the new functions, please see
https://gist.github.com/mwaskom/de44147ed2974457ad6372750bbe5751 (https://gist.githu
b.com/mwaskom/de44147ed2974457ad6372750bbe5751)

  sns.distplot(df['placement_exam_marks'])
C:\Users\ASUS\AppData\Local\Temp\ipykernel_16968\3858278419.py:9: UserWarning:

`distplot` is a deprecated function and will be removed in seaborn v0.14.0.

Please adapt your code to use either `displot` (a figure-level function with
similar flexibility) or `histplot` (an axes-level function for histograms).

For a guide to updating your code to use the new functions, please see
https://gist.github.com/mwaskom/de44147ed2974457ad6372750bbe5751 (https://gist.githu
b.com/mwaskom/de44147ed2974457ad6372750bbe5751)

  sns.distplot(new_df['placement_exam_marks'])
```



Apply Capping Method - 2:

In [30]:
```
new_df_cap = df.copy()

new_df_cap['placement_exam_marks'] = np.where(
    new_df_cap['placement_exam_marks'] > upper_limit,
    upper_limit,
    np.where(
        new_df_cap['placement_exam_marks'] < lower_limit,
        lower_limit,
        new_df_cap['placement_exam_marks']
    )
)
```

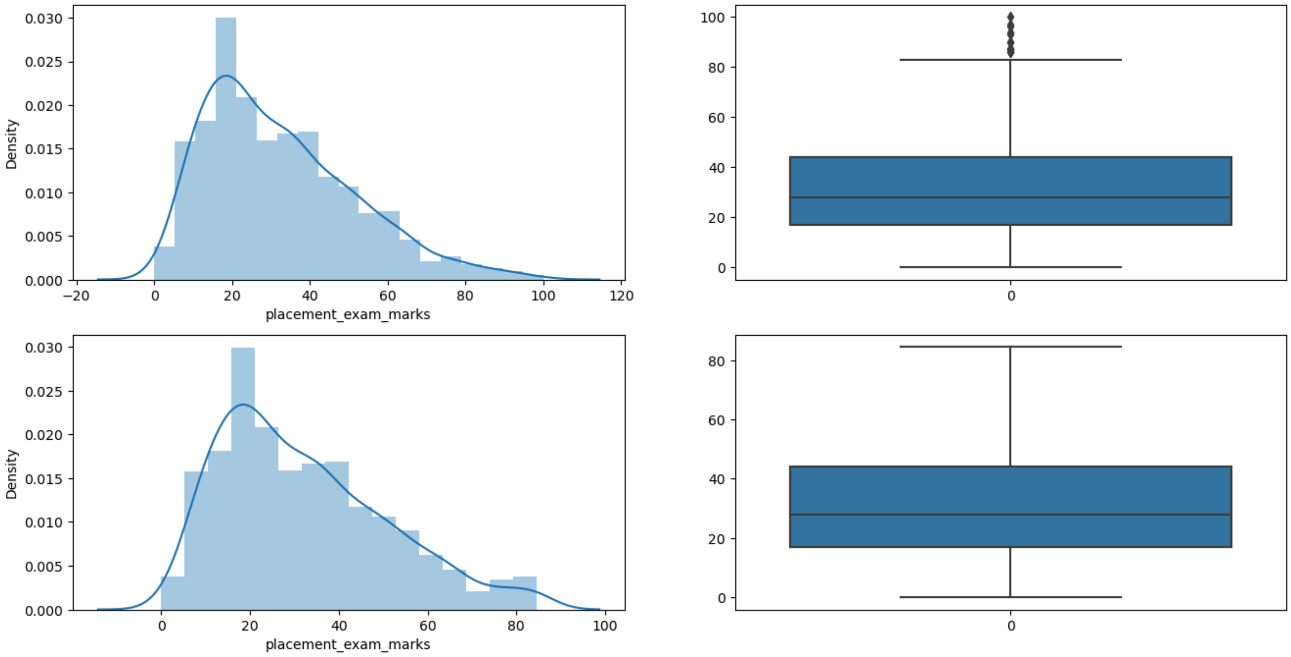
In [31]:
```
new_df_cap.shape
```

Out[31]:
```
(1000, 3)
```

Compare Before and After (After Capping):

In [32]:
```
plt.figure(figsize=(16,8))
plt.subplot(2,2,1)
sns.distplot(df['placement_exam_marks'])

plt.subplot(2,2,2)
sns.boxplot(df['placement_exam_marks'])

plt.subplot(2,2,3)
sns.distplot(new_df_cap['placement_exam_marks'])

plt.subplot(2,2,4)
sns.boxplot(new_df_cap['placement_exam_marks'])

plt.show()
```

```
C:\Users\ASUS\AppData\Local\Temp\ipykernel_16968\1476363708.py:3: UserWarning:

`distplot` is a deprecated function and will be removed in seaborn v0.14.0.

Please adapt your code to use either `displot` (a figure-level function with
similar flexibility) or `histplot` (an axes-level function for histograms).

For a guide to updating your code to use the new functions, please see
https://gist.github.com/mwaskom/de44147ed2974457ad6372750bbe5751 (https://gist.githu
b.com/mwaskom/de44147ed2974457ad6372750bbe5751)

  sns.distplot(df['placement_exam_marks'])
C:\Users\ASUS\AppData\Local\Temp\ipykernel_16968\1476363708.py:9: UserWarning:

`distplot` is a deprecated function and will be removed in seaborn v0.14.0.

Please adapt your code to use either `displot` (a figure-level function with
similar flexibility) or `histplot` (an axes-level function for histograms).

For a guide to updating your code to use the new functions, please see
https://gist.github.com/mwaskom/de44147ed2974457ad6372750bbe5751 (https://gist.githu
b.com/mwaskom/de44147ed2974457ad6372750bbe5751)

  sns.distplot(new_df_cap['placement_exam_marks'])
```



In [ ]: