In [1]:
```
import pandas as pd
import numpy as np
```

In [3]:
```
df = pd.read_csv("D:\\Summer Training Video\\ML\\newplacementdata.csv")
```

In [4]:
```
df
```

Out[4]:

	cgpa	placement_exam_marks	placed
0	7.19	26	1
1	7.46	38	1
2	7.54	40	1
3	6.42	8	1
4	7.23	17	0
...	...	...	...
995	8.87	44	1
996	9.12	65	1
997	4.89	34	0
998	8.62	46	1
999	4.90	10	1

1000 rows × 3 columns

In [5]:
```
df.head()
```

Out[5]:

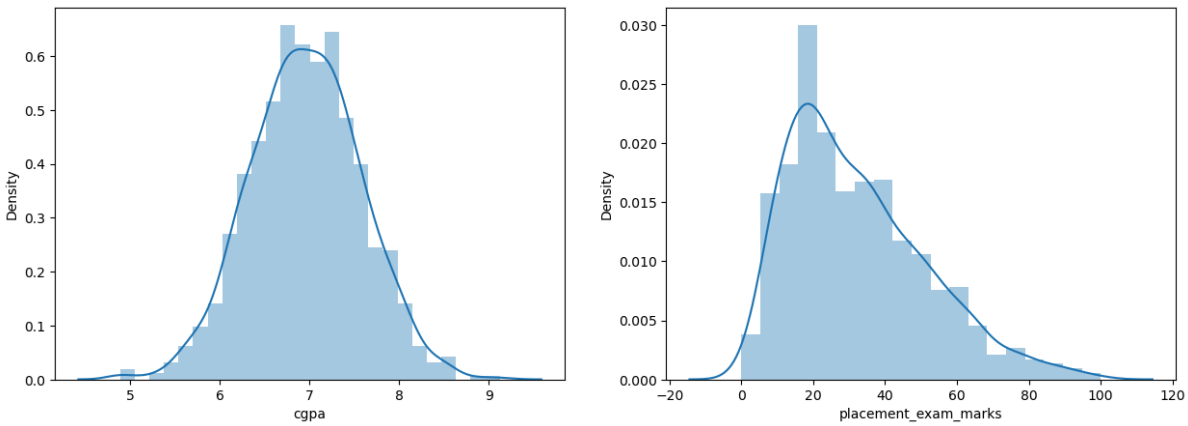
	cgpa	placement_exam_marks	placed
0	7.19	26	1
1	7.46	38	1
2	7.54	40	1
3	6.42	8	1
4	7.23	17	0

In [6]:
```
# even = ((n/2) + ((n/2)+1))/2
# odd = ((n/2) + 1)
```

In [7]:
```
import matplotlib.pyplot as plt
import seaborn as sns
```

In [9]:
```
plt.figure(figsize = (15,5))
plt.subplot(121)
sns.distplot(df['cgpa'])
plt.subplot(122)
sns.distplot(df['placement_exam_marks'])
plt.show()
```

```
C:\Users\yashs\AppData\Local\Temp\ipykernel_19528\48147761.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.github.com/mwaskom/de44147ed2974457ad6372750bbe5751)

  sns.distplot(df['cgpa'])
C:\Users\yashs\AppData\Local\Temp\ipykernel_19528\48147761.py:5: 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.github.com/mwaskom/de44147ed2974457ad6372750bbe5751)

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

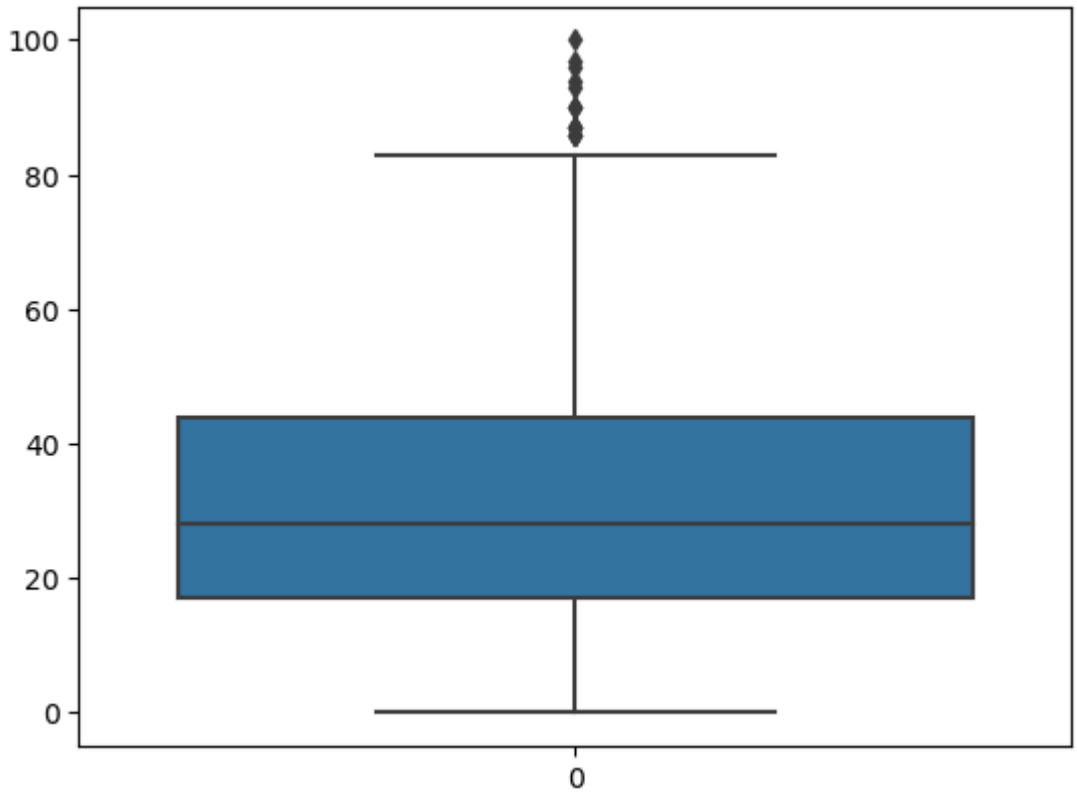


In [10]:
```
df['placement_exam_marks'].describe()
```

Out[10]:
```
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
```

In [11]:
```
sns.boxplot(df['placement_exam_marks'])
```

Out[11]:
```
<Axes: >
```



In [12]:
```
# Finding the IRQ

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

In [15]:
```
IQR = percentile75 - percentile25
```

In [16]:
```
IQR
```

Out[16]: 27.0

In [17]:
```
upper_limit = percentile25 + 1.5*IQR
upper_limit
```

Out[17]: 57.5

In [18]:
```
lower_limit = percentile25 - 1.5*IQR
lower_limit
```

Out[18]: -23.5

Finding Our Outliers

In [19]:
```
df[df['placement_exam_marks']> upper_limit]
```

Out[19]:

	cgpa	placement_exam_marks	placed
9	7.75	94	1
25	6.28	58	1
40	6.60	86	1
42	7.46	71	1
43	7.85	63	0
...	...	...	...
966	6.24	72	1
967	7.35	59	0
987	6.77	62	0
994	6.48	63	0
996	9.12	65	1

114 rows × 3 columns

In [20]:
```
df[df['placement_exam_marks']< lower_limit]
```

Out[20]:

	cgpa	placement_exam_marks	placed

In [21]:
```
# Trimming (Outlier Removing technique 1)
```

In [22]:
```
new_df = df[df['placement_exam_marks'] < upper_limit]
```

In [23]:
```
new_df
```

Out[23]:

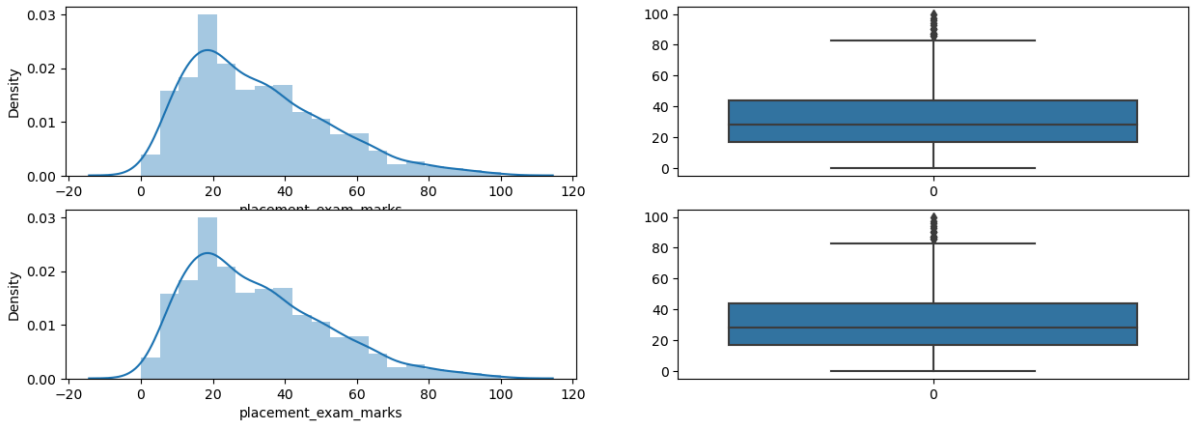
	cgpa	placement_exam_marks	placed
0	7.19	26	1
1	7.46	38	1
2	7.54	40	1
3	6.42	8	1
4	7.23	17	0
...	...	...	...
993	6.73	21	1
995	8.87	44	1
997	4.89	34	0
998	8.62	46	1
999	4.90	10	1

886 rows × 3 columns

In [24]:
```
# Comparision
```

In [28]:
```
plt.figure(figsize = (15,5))
plt.subplot(221)
sns.distplot(df['placement_exam_marks'])

plt.subplot(222)
sns.boxplot(df['placement_exam_marks'])

plt.subplot(223)
sns.distplot(df['placement_exam_marks'])

plt.subplot(224)
sns.boxplot(df['placement_exam_marks'])
plt.show()
```

```
C:\Users\yashs\AppData\Local\Temp\ipykernel_19528\3534830507.py:3: UserWarnin
g:

`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://gis
t.github.com/mwaskom/de44147ed2974457ad6372750bbe5751)

  sns.distplot(df['placement_exam_marks'])
C:\Users\yashs\AppData\Local\Temp\ipykernel_19528\3534830507.py:9: UserWarnin
g:

`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://gis
t.github.com/mwaskom/de44147ed2974457ad6372750bbe5751)

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



In [29]:
```
# Capping(Outlier Removing technique 2)
```

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

In [31]:
```
# min = 5 , max  15

# min 4 , 3 , 1
# max = 20 , 30 , 50

# updated_min_value = 1
# updated_max_value = 50
```

In [32]:
```
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 [33]:
```
new_df_cap
```

Out[33]:

	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	57.5	1
997	4.89	34.0	0
998	8.62	46.0	1
999	4.90	10.0	1

1000 rows × 3 columns

In [34]:
```
new_df_cap.shape
```

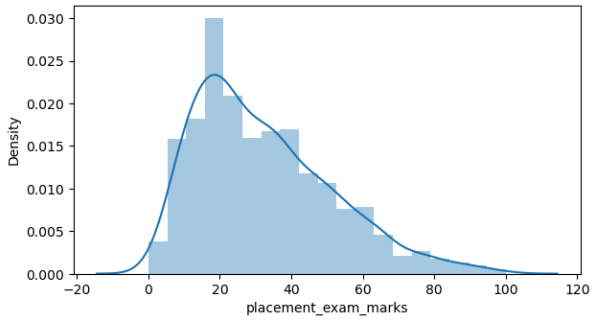
Out[34]:
```
(1000, 3)
```

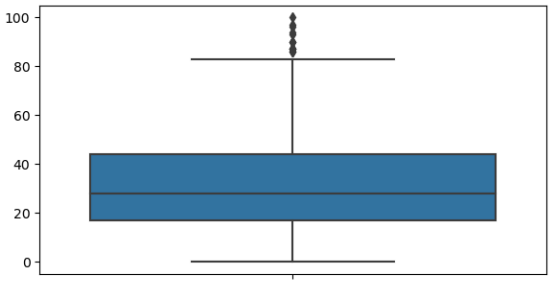
In [35]:
```
# Comparision
```

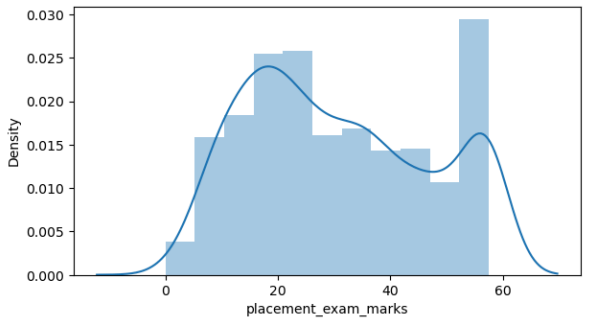
In [37]:
```
plt.figure(figsize = (15,8))
plt.subplot(221)
sns.distplot(df['placement_exam_marks'])

plt.subplot(222)
sns.boxplot(df['placement_exam_marks'])

plt.subplot(223)
sns.distplot(new_df_cap['placement_exam_marks'])

plt.subplot(224)
sns.boxplot(new_df_cap['placement_exam_marks'])
```

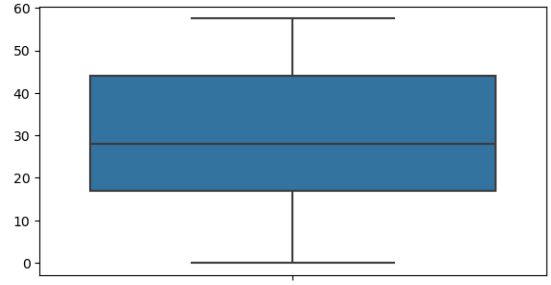
```
C:\Users\yashs\AppData\Local\Temp\ipykernel_19528\1993446638.py:3: UserWarnin
g:

`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://gis
t.github.com/mwaskom/de44147ed2974457ad6372750bbe5751)

  sns.distplot(df['placement_exam_marks'])
C:\Users\yashs\AppData\Local\Temp\ipykernel_19528\1993446638.py:9: UserWarnin
g:

`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://gis
t.github.com/mwaskom/de44147ed2974457ad6372750bbe5751)

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

Out[37]:
```
<Axes: >
```


Density
placement_exam_marks




Density
placement_exam_marks



In [ ]: