In [31]:
```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
```

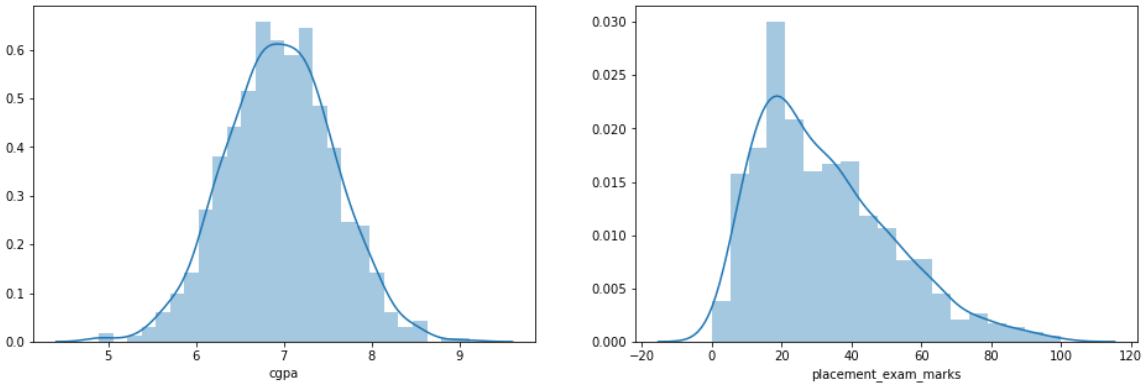
In [32]:
```
df = pd.read_csv('placement.csv')
```

In [33]:
```
df.head()
```

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

In [34]:
```
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()
```



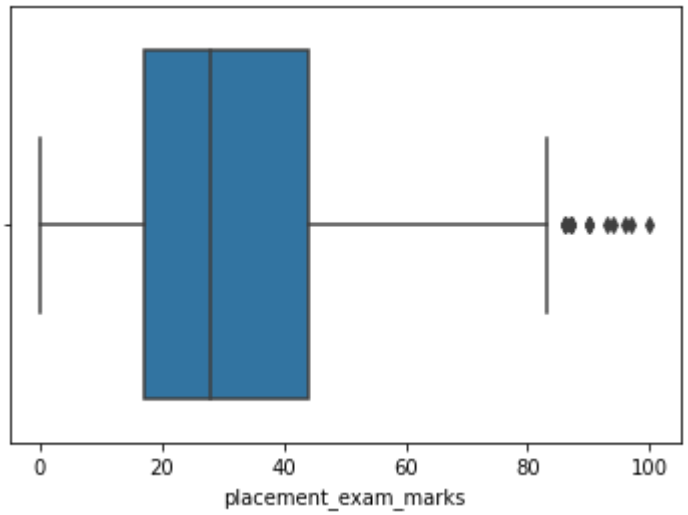
In [35]:
```
df['placement_exam_marks'].describe()
```

Out[35]:
```
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 [36]:
```
sns.boxplot(df['placement_exam_marks'])
```

Out[36]:



In [37]:
```
# Finding the IQR
percentile25 = df['placement_exam_marks'].quantile(0.25)
percentile75 = df['placement_exam_marks'].quantile(0.75)
```

In [39]:
```
percentile75
```

Out[39]: 44.0

In [40]:
```
iqr = percentile75 - percentile25
```

In [41]:
```
iqr
```

Out[41]: 27.0

In [42]:
```
upper_limit = percentile75 + 1.5 * iqr
lower_limit = percentile25 - 1.5 * iqr
```

In [43]:
```
print("Upper limit",upper_limit)
print("Lower limit",lower_limit)
```

```
Upper limit 84.5
Lower limit -23.5
```

Finding Outliers

In [44]:
```
df[df['placement_exam_marks'] > upper_limit]
```

Out[44]:

	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 [45]:
```
df[df['placement_exam_marks'] < lower_limit]
```

Out[45]:

cgpa	placement_exam_marks	placed

Trimming

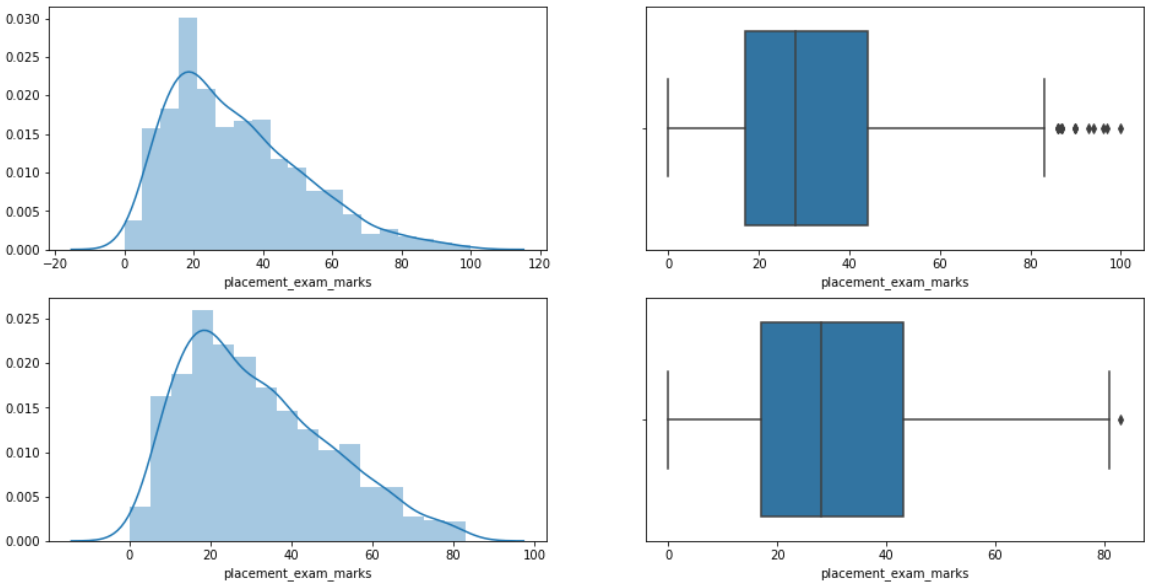
In [46]:
```
new_df = df[df['placement_exam_marks'] < upper_limit]
```

In [47]:
```
new_df.shape
```

Out[47]: (985, 3)

In [48]:
```
# Comparing

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()
```



Capping

In [49]:
```
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']
    )
)
```

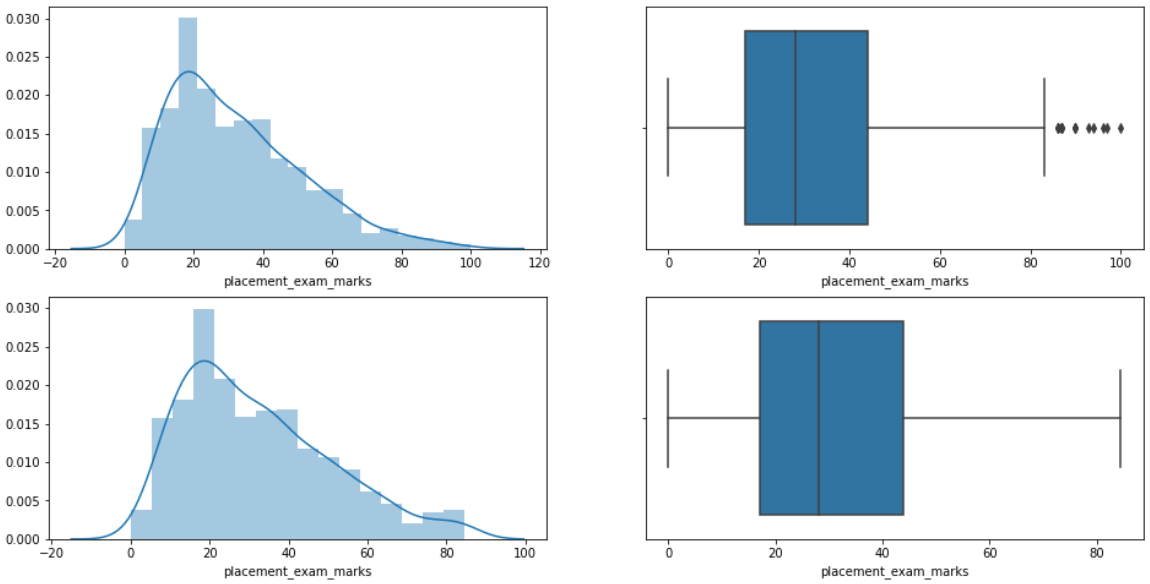
In [ ]:
```
np.where(condtion,true,false)
```

In [50]:
```
new_df_cap.shape
```

Out[50]: (1000, 3)

In [51]:
```
# Comparing

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()
```



In [ ]: