practical-2-dsbda

May 4, 2025

[1]:
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
```

[2]:
```
df = pd.read_csv("AcademicPerformanceNEW.csv")
```

[3]:
```
df.head()
```

[3]:
```
   gender  math score  reading score  writing score  placement score  \
0    Male          66           65.0           76.0               97
1  Female          91           70.0           76.0               76
2  Female          72           66.0           75.0               79
3    Male          99           75.0           67.0               84
4  Female          62           65.0           68.0               87

   club join year  placement count    region
0          2020.0                3      Pune
1          2019.0                2  Baramati
2          2019.0                2    Satara
3          2018.0                2     Karad
4          2018.0                3    Mulshi
```

[4]:
```
df.tail()
```

[4]:
```
    gender  math score  reading score  writing score  placement score  \
24    Male          77           61.0           74.0               84
25  Female          75           73.0           74.0               86
26    Male          65           64.0           64.0               96
27    Male          64           68.0           65.0               95
28    Male          69           68.0           80.0               75

    club join year  placement count  region
24          2021.0                2  Mulshi
25          2021.0                3    Pune
26          2021.0                3  Mulshi
27          2020.0                3  Mulshi
```

```
28          2019.0                  2    Pune
```

[5]:
```python
df.isnull()
```

[5]:
```
    gender  math score  reading score  writing score  placement score  \
0    False       False          False          False            False
1    False       False          False          False            False
2    False       False          False          False            False
3    False       False          False          False            False
4    False       False          False          False            False
5     True       False           True          False            False
6    False       False          False          False            False
7    False       False          False          False            False
8    False       False          False           True            False
9    False       False          False          False            False
10   False       False          False          False            False
11    True       False          False          False            False
12   False       False          False          False            False
13   False       False          False          False            False
14   False       False          False          False            False
15   False       False          False          False            False
16   False       False          False          False            False
17   False       False           True          False            False
18    True       False          False          False            False
19   False       False          False          False            False
20   False       False          False          False            False
21   False       False          False          False            False
22   False       False          False          False            False
23   False       False          False          False            False
24   False       False          False          False            False
25   False       False          False          False            False
26   False       False          False          False            False
27   False       False          False          False            False
28   False       False          False          False            False

    club join year  placement count  region
0            False            False   False
1            False            False   False
2            False            False   False
3            False            False   False
4            False            False   False
5             True            False   False
6            False            False   False
7            False            False   False
8            False            False   False
9            False            False   False
10           False            False    True
```

```
11           False            False   False
12            True            False   False
13           False            False   False
14           False            False   False
15           False            False   False
16           False            False   False
17           False            False   False
18           False            False   False
19           False            False    True
20           False            False   False
21           False            False   False
22            True            False   False
23           False            False   False
24           False            False   False
25           False            False   False
26           False            False   False
27           False            False   False
28           False            False   False
```

[6]:
```
df.isnull().sum()
```

[6]:
```
gender             3
math score         0
reading score      2
writing score      1
placement score    0
club join year     3
placement count    0
region             2
dtype: int64
```

[8]:
```
#handle missing values
for col in df.select_dtypes(include=np.number).columns:
    df[col].fillna(df[col].mean(), inplace=True)
```

[9]:
```
df.isnull()
```

[9]:
```
    gender  math score  reading score  writing score  placement score  \
0    False       False          False          False            False
1    False       False          False          False            False
2    False       False          False          False            False
3    False       False          False          False            False
4    False       False          False          False            False
5     True       False          False          False            False
6    False       False          False          False            False
7    False       False          False          False            False
8    False       False          False          False            False
```

9	False	False	False	False	False
10	False	False	False	False	False
11	True	False	False	False	False
12	False	False	False	False	False
13	False	False	False	False	False
14	False	False	False	False	False
15	False	False	False	False	False
16	False	False	False	False	False
17	False	False	False	False	False
18	True	False	False	False	False
19	False	False	False	False	False
20	False	False	False	False	False
21	False	False	False	False	False
22	False	False	False	False	False
23	False	False	False	False	False
24	False	False	False	False	False
25	False	False	False	False	False
26	False	False	False	False	False
27	False	False	False	False	False
28	False	False	False	False	False

	club	join year	placement	count	region
0		False		False	False
1		False		False	False
2		False		False	False
3		False		False	False
4		False		False	False
5		False		False	False
6		False		False	False
7		False		False	False
8		False		False	False
9		False		False	False
10		False		False	True
11		False		False	False
12		False		False	False
13		False		False	False
14		False		False	False
15		False		False	False
16		False		False	False
17		False		False	False
18		False		False	False
19		False		False	True
20		False		False	False
21		False		False	False
22		False		False	False
23		False		False	False
24		False		False	False

```
25           False            False   False
26           False            False   False
27           False            False   False
28           False            False   False
```

[11]:
```
for col in df.select_dtypes(include='object').columns:
    df[col].fillna(df[col].mode()[0],inplace=True)
```

[12]:
```
df.isnull()
```

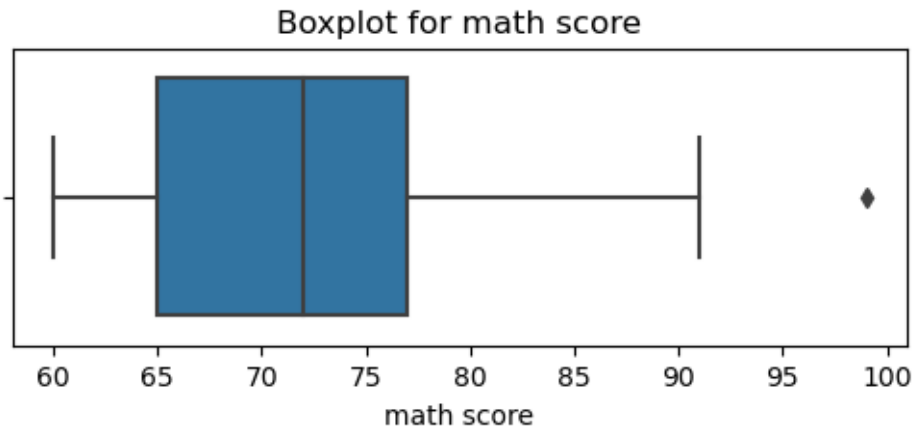
[12]:
```
    gender  math score  reading score  writing score  placement score  \
0    False       False          False          False            False
1    False       False          False          False            False
2    False       False          False          False            False
3    False       False          False          False            False
4    False       False          False          False            False
5    False       False          False          False            False
6    False       False          False          False            False
7    False       False          False          False            False
8    False       False          False          False            False
9    False       False          False          False            False
10   False       False          False          False            False
11   False       False          False          False            False
12   False       False          False          False            False
13   False       False          False          False            False
14   False       False          False          False            False
15   False       False          False          False            False
16   False       False          False          False            False
17   False       False          False          False            False
18   False       False          False          False            False
19   False       False          False          False            False
20   False       False          False          False            False
21   False       False          False          False            False
22   False       False          False          False            False
23   False       False          False          False            False
24   False       False          False          False            False
25   False       False          False          False            False
26   False       False          False          False            False
27   False       False          False          False            False
28   False       False          False          False            False

    club join year  placement count  region
0           False            False   False
1           False            False   False
2           False            False   False
3           False            False   False
4           False            False   False
```

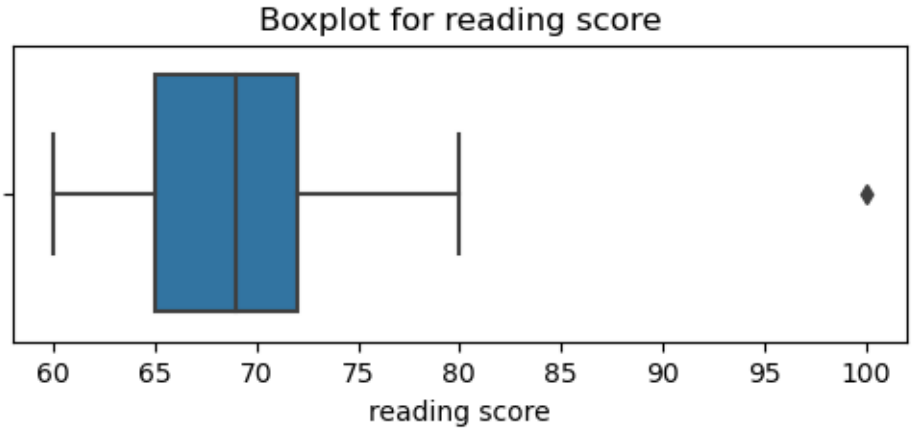
```
5           False           False   False
6           False           False   False
7           False           False   False
8           False           False   False
9           False           False   False
10          False           False   False
11          False           False   False
12          False           False   False
13          False           False   False
14          False           False   False
15          False           False   False
16          False           False   False
17          False           False   False
18          False           False   False
19          False           False   False
20          False           False   False
21          False           False   False
22          False           False   False
23          False           False   False
24          False           False   False
25          False           False   False
26          False           False   False
27          False           False   False
28          False           False   False
```

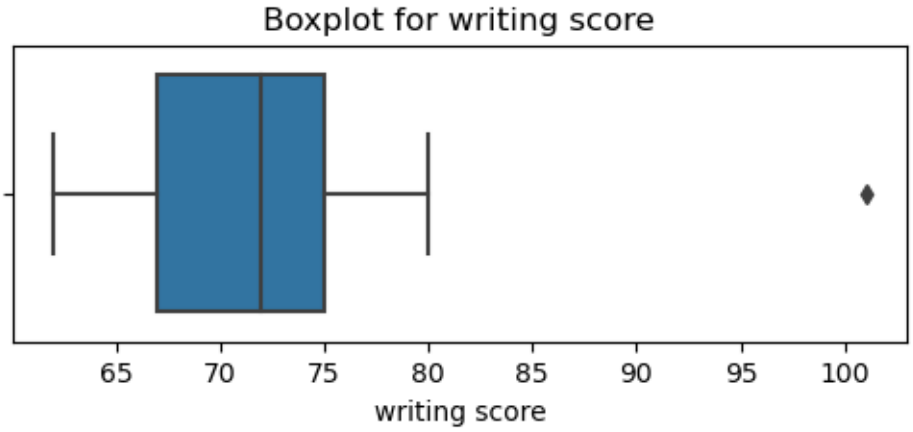
[13]:
```
df.isnull().sum()
```

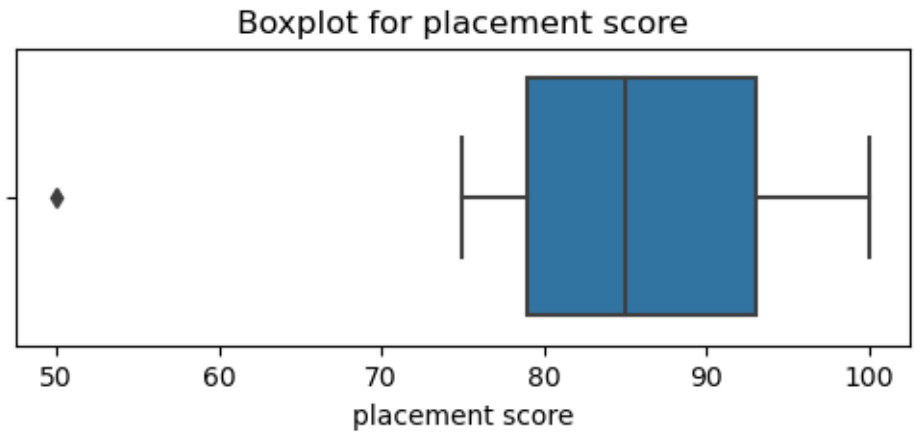
[13]:
```
gender              0
math score          0
reading score       0
writing score       0
placement score     0
club join year      0
placement count     0
region              0
dtype: int64
```

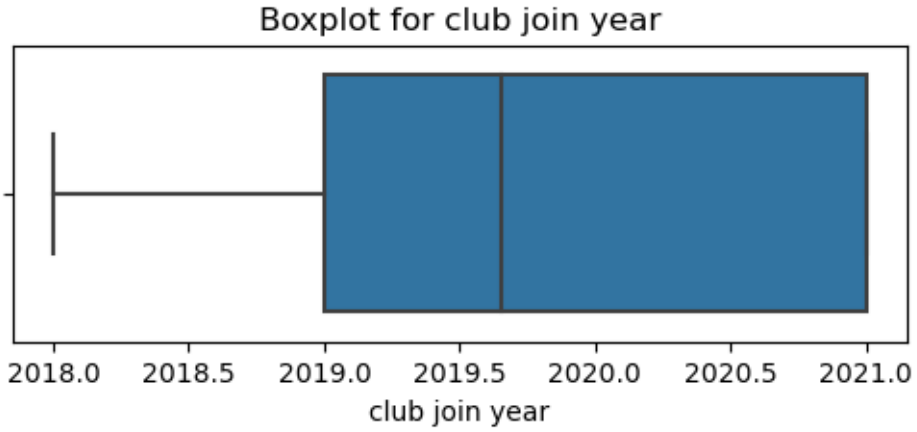
[14]:
```
#Detect outliers
numeric_cols = df.select_dtypes(include=np.number).columns
for col in numeric_cols:
    plt.figure(figsize=(6, 2))
    sns.boxplot(data=df, x=col)
    plt.title(f"Boxplot for {col}")
    plt.show()
```

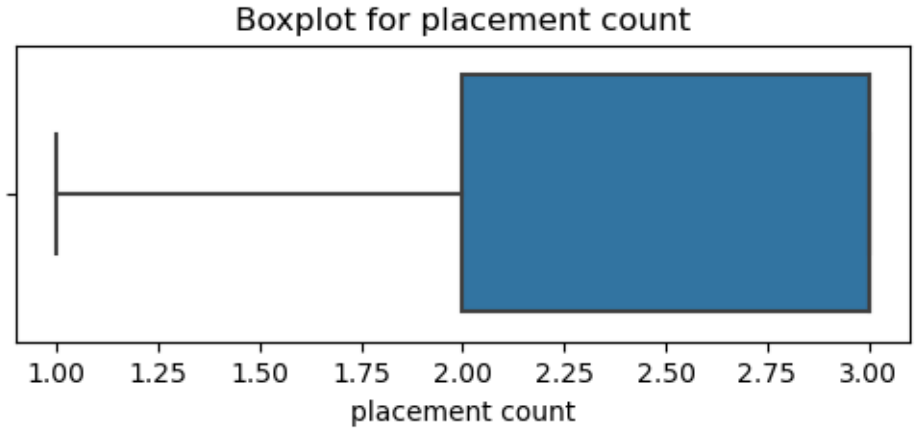

Boxplot for math score
60
65
70
75
80
85
90
95
100
math score


Boxplot for reading score
60
65
70
75
80
85
90
95
100
reading score


Boxplot for writing score
65
70
75
80
85
90
95
100
writing score

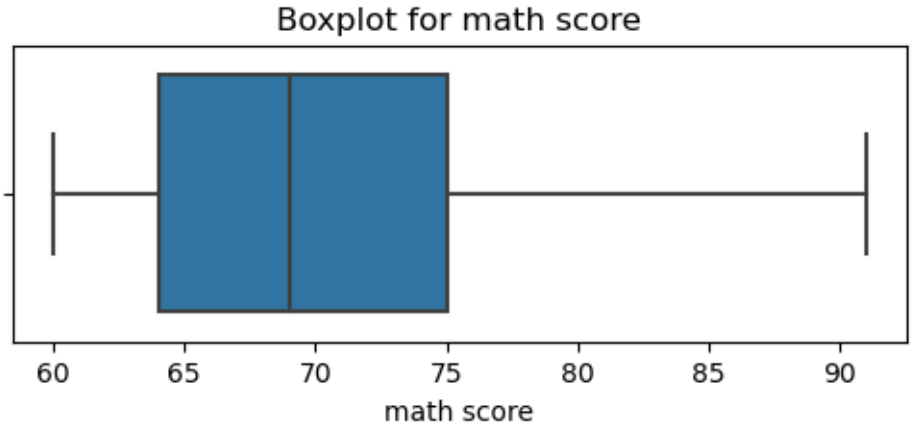

Boxplot for placement score
50
60
70
80
90
100
placement score

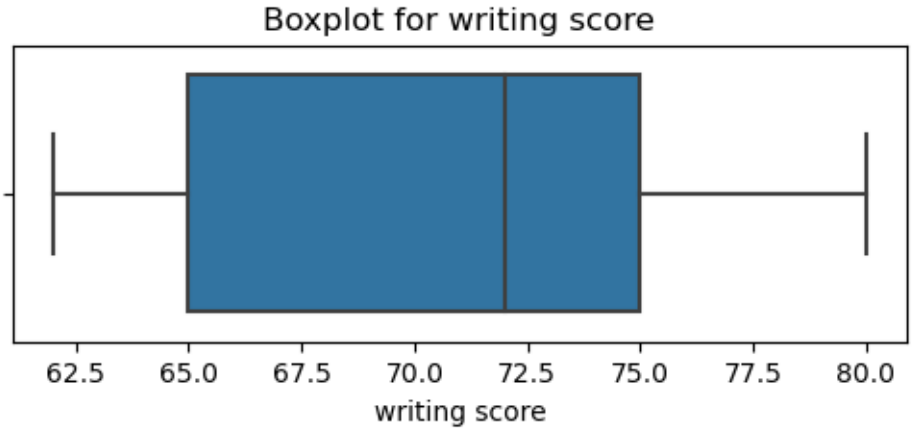

Boxplot for club join year
2018.0
2018.5
2019.0
2019.5
2020.0
2020.5
2021.0
club join year

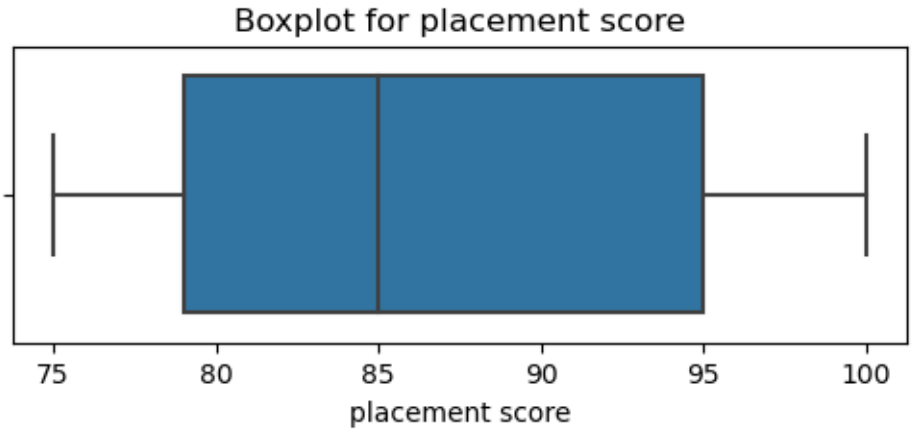

Boxplot for placement count
1.00
1.25
1.50
1.75
2.00
2.25
2.50
2.75
3.00
placement count

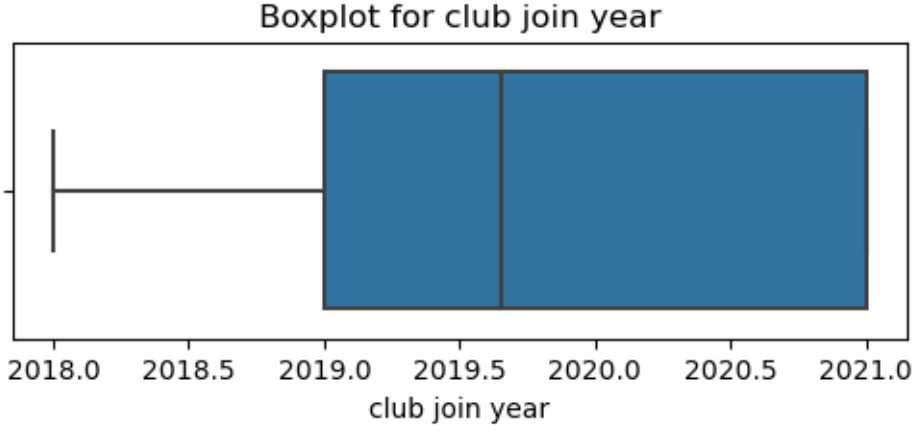
[16]:
```
#Handle the outliers
def remove_outliers(df, columns):
    Q1 = df[columns].quantile(0.25)
    Q3 = df[columns].quantile(0.75)
    IQR = Q3 - Q1
    lower = Q1 - 1.5*IQR
    upper = Q3 + 1.5*IQR
    return df[(df[columns]>= lower) & (df[columns]<=upper)]

for col in numeric_cols:
    df = remove_outliers(df, col)
```

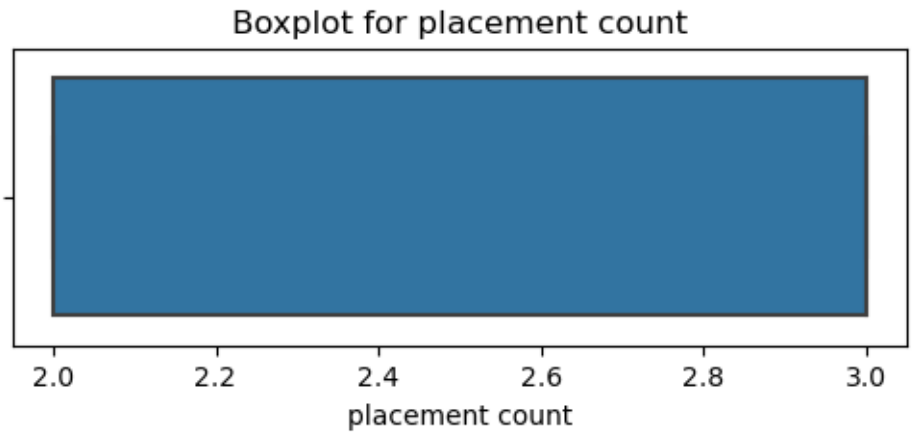
[18]:
```
for col in numeric_cols:
    plt.figure(figsize=(6, 2))
    sns.boxplot(data=df, x=col)
    plt.title(f"Boxplot for {col}")
    plt.show()
```




Boxplot for writing score
62.5
65.0
67.5
70.0
72.5
75.0
77.5
80.0
writing score


Boxplot for placement score
75
80
85
90
95
100
placement score


Boxplot for club join year
2018.0
2018.5
2019.0
2019.5
2020.0
2020.5
2021.0
club join year

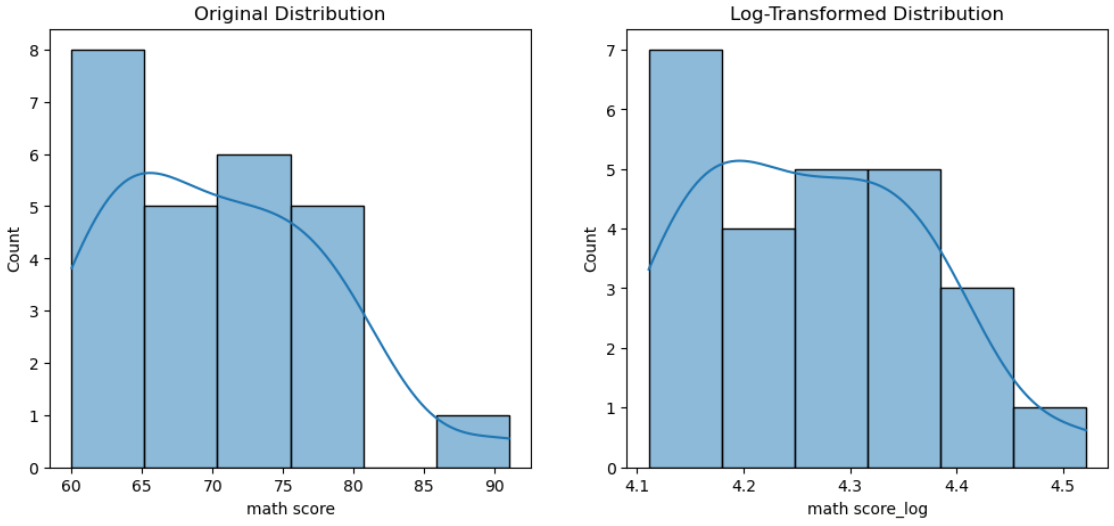


[19]:
```
df['math score'].skew()
```

[19]:
```
0.6859227337618937
```

[20]:
```
df['math score_log'] = np.log1p(df['math score'])
```

[23]:
```
plt.figure(figsize=(12, 5))
plt.subplot(1, 2, 1)
sns.histplot(df['math score'], kde=True)
plt.title("Original Distribution")

plt.subplot(1, 2, 2)
sns.histplot(df['math score_log'], kde=True)
plt.title("Log-Transformed Distribution")

plt.show()
```



[ ]: