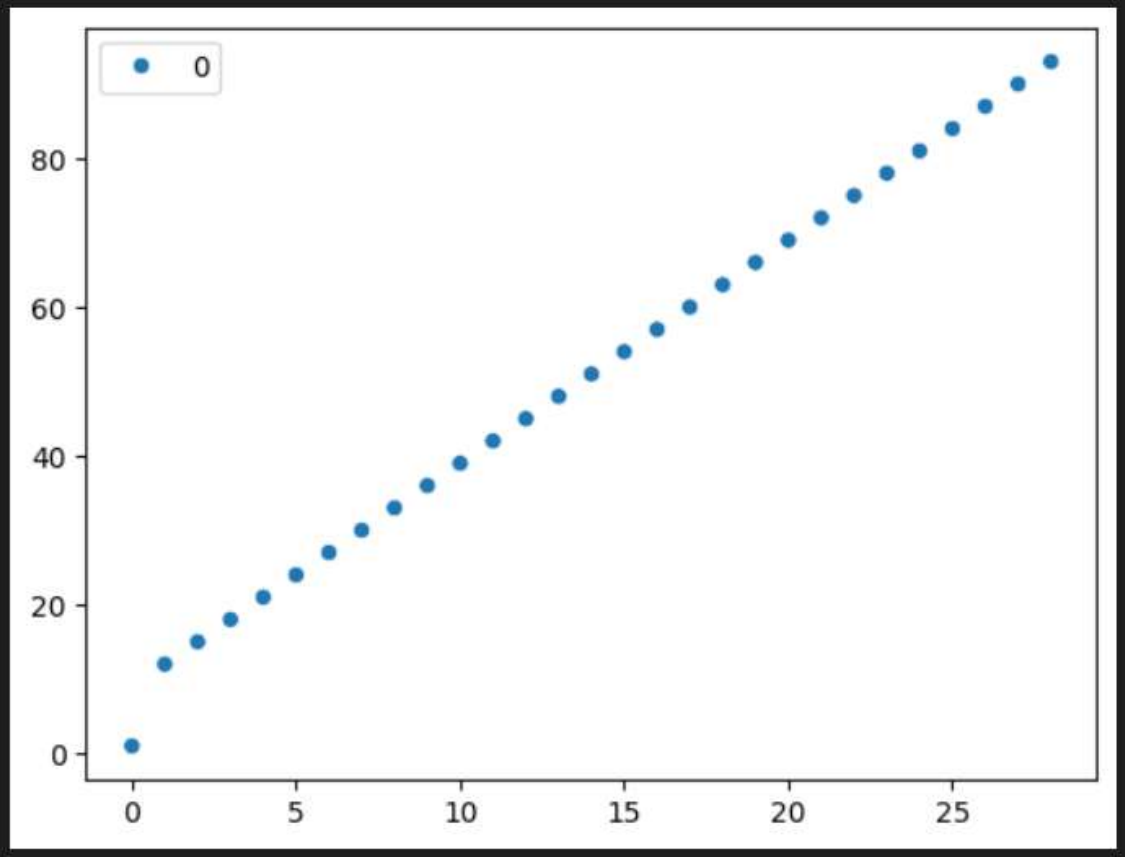
```
sns.scatterplot(df_cleaned)
```

```
<Axes: >
```



Colab Notebooks - Google Dri | ex 01.ipynb - Colab | Colab Notebooks - Google Dri | Editing exno1/README.md at

colab.research.google.com/drive/1d72PRm0D2kkX5Ja7aFR8mx4OsepTKGrb#scrollTo=4K4T_H8Qja-s

Google Chrome isn't your default browser | Set as default

ex 01.ipynb

File Edit View Insert Runtime Tools Help All changes saved

Comment | Share

+ Code + Text

T4 RAM Disk | Gemini

```
from google.colab import drive
drive.mount("/content/drive")
```

```
Drive already mounted at /content/drive; to attempt to forcibly remount, call drive.mount("/content/drive", force_remount=True).
```

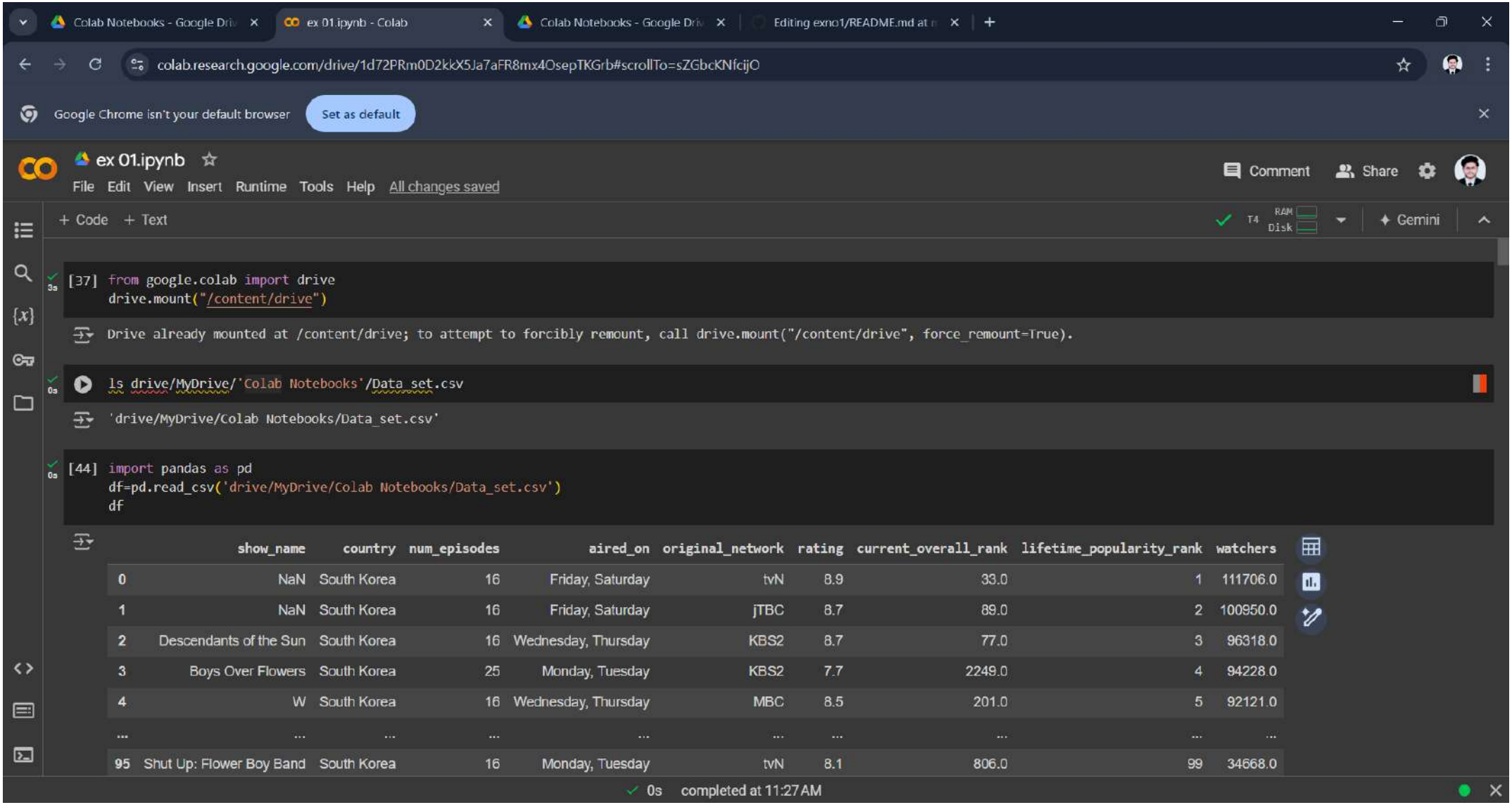
```
[42] ls drive/MyDrive/'Colab Notebooks'/Data_set.csv
```

```
'drive/MyDrive/Colab Notebooks/Data_set.csv'
```

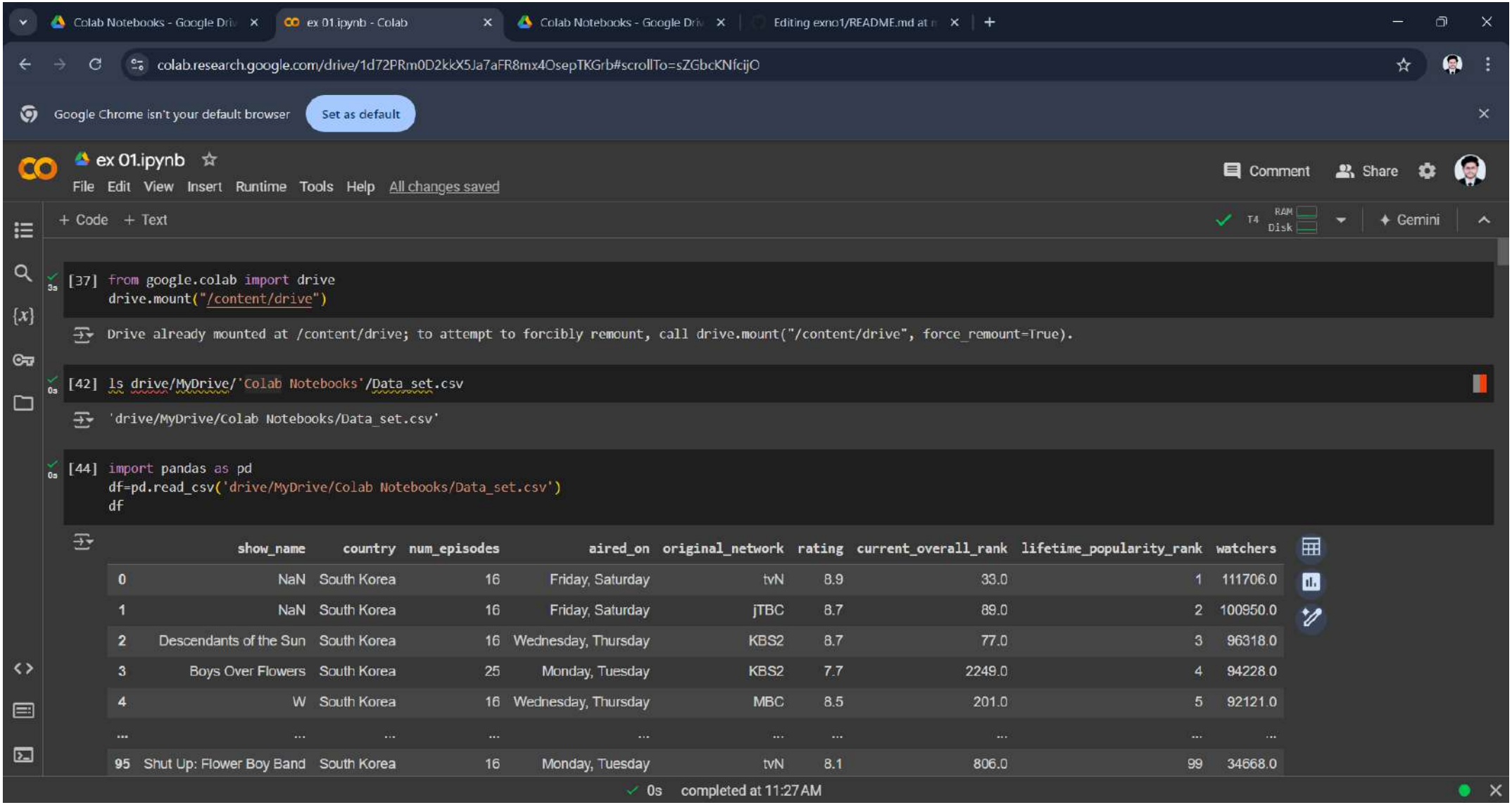
```
[44] import pandas as pd
df=pd.read_csv('drive/MyDrive/Colab Notebooks/Data_set.csv')
df
```

	show_name	country	num_episodes	aired_on	original_network	rating	current_overall_rank	lifetime_popularity_rank	watchers
0	NaN	South Korea	16	Friday, Saturday	tvN	8.9	33.0	1	111706.0
1	NaN	South Korea	16	Friday, Saturday	jTBC	8.7	89.0	2	100950.0
2	Descendants of the Sun	South Korea	16	Wednesday, Thursday	KBS2	8.7	77.0	3	96318.0
3	Boys Over Flowers	South Korea	25	Monday, Tuesday	KBS2	7.7	2249.0	4	94228.0
4	W	South Korea	16	Wednesday, Thursday	MBC	8.5	201.0	5	92121.0
...	...	...	...	...	...	...	...	...	...
95	Shut Up: Flower Boy Band	South Korea	16	Monday, Tuesday	tvN	8.1	806.0	99	34668.0

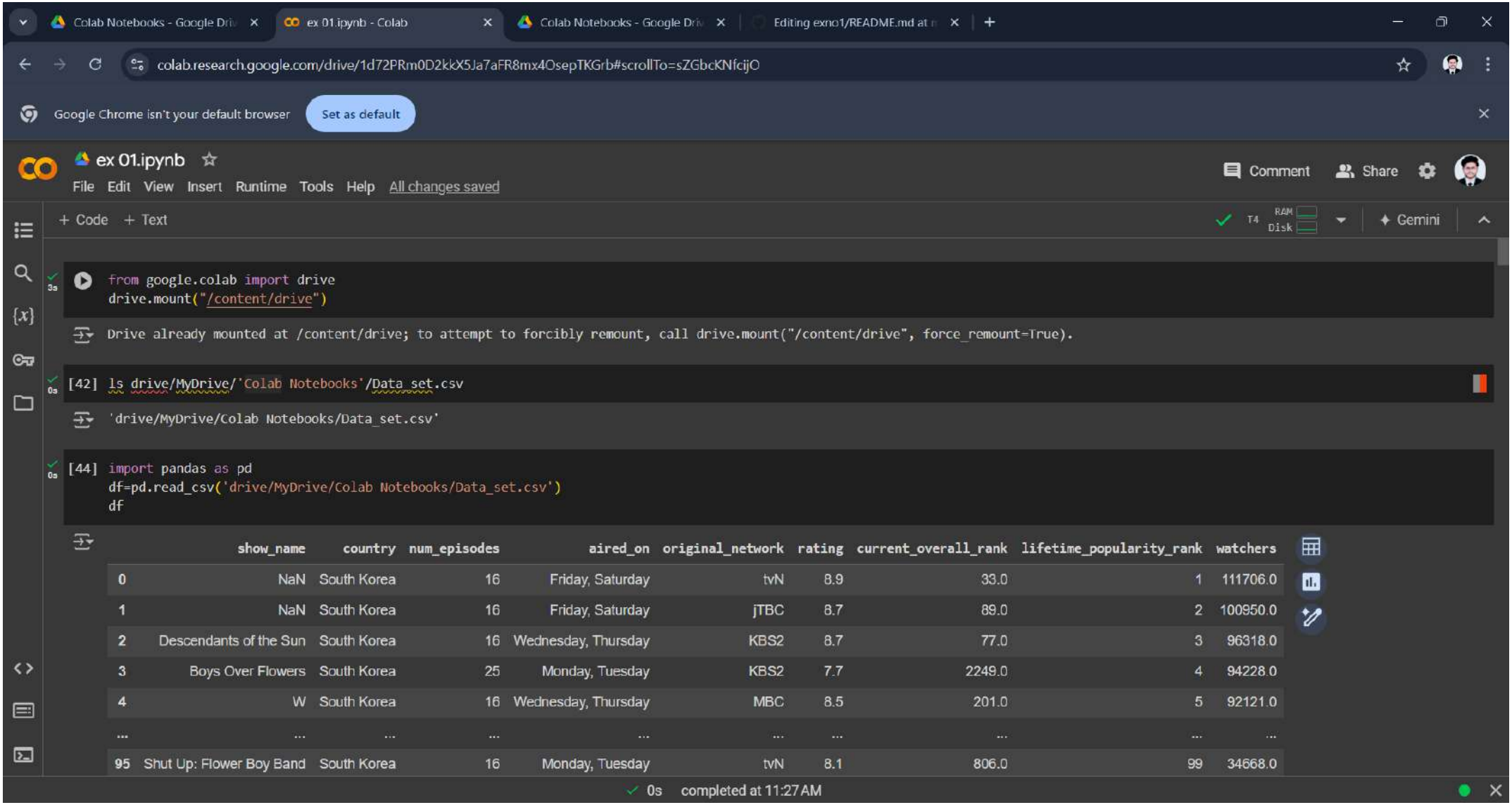
0s completed at 11:27 AM



	show_name	country	num_episodes	aired_on	original_network	rating	current_overall_rank	lifetime_popularity_rank	watchers
0	NaN	South Korea	16	Friday, Saturday	tvN	8.9	33.0	1	111706.0
1	NaN	South Korea	16	Friday, Saturday	jTBC	8.7	89.0	2	100950.0
2	Descendants of the Sun	South Korea	16	Wednesday, Thursday	KBS2	8.7	77.0	3	96318.0
3	Boys Over Flowers	South Korea	25	Monday, Tuesday	KBS2	7.7	2249.0	4	94228.0
4	W	South Korea	16	Wednesday, Thursday	MBC	8.5	201.0	5	92121.0
...	...	...	...	...	...	...	...	...	...
95	Shut Up: Flower Boy Band	South Korea	16	Monday, Tuesday	tvN	8.1	806.0	99	34668.0



	show_name	country	num_episodes	aired_on	original_network	rating	current_overall_rank	lifetime_popularity_rank	watchers
0	NaN	South Korea	16	Friday, Saturday	tvN	8.9	33.0	1	111706.0
1	NaN	South Korea	16	Friday, Saturday	jTBC	8.7	89.0	2	100950.0
2	Descendants of the Sun	South Korea	16	Wednesday, Thursday	KBS2	8.7	77.0	3	96318.0
3	Boys Over Flowers	South Korea	25	Monday, Tuesday	KBS2	7.7	2249.0	4	94228.0
4	W	South Korea	16	Wednesday, Thursday	MBC	8.5	201.0	5	92121.0
...	...	...	...	...	...	...	...	...	...
95	Shut Up: Flower Boy Band	South Korea	16	Monday, Tuesday	tvN	8.1	806.0	99	34668.0



	show_name	country	num_episodes	aired_on	original_network	rating	current_overall_rank	lifetime_popularity_rank	watchers
0	NaN	South Korea	16	Friday, Saturday	tvN	8.9	33.0	1	111706.0
1	NaN	South Korea	16	Friday, Saturday	jTBC	8.7	89.0	2	100950.0
2	Descendants of the Sun	South Korea	16	Wednesday, Thursday	KBS2	8.7	77.0	3	96318.0
3	Boys Over Flowers	South Korea	25	Monday, Tuesday	KBS2	7.7	2249.0	4	94228.0
4	W	South Korea	16	Wednesday, Thursday	MBC	8.5	201.0	5	92121.0
...	...	...	...	...	...	...	...	...	...
95	Shut Up: Flower Boy Band	South Korea	16	Monday, Tuesday	tvN	8.1	806.0	99	34668.0

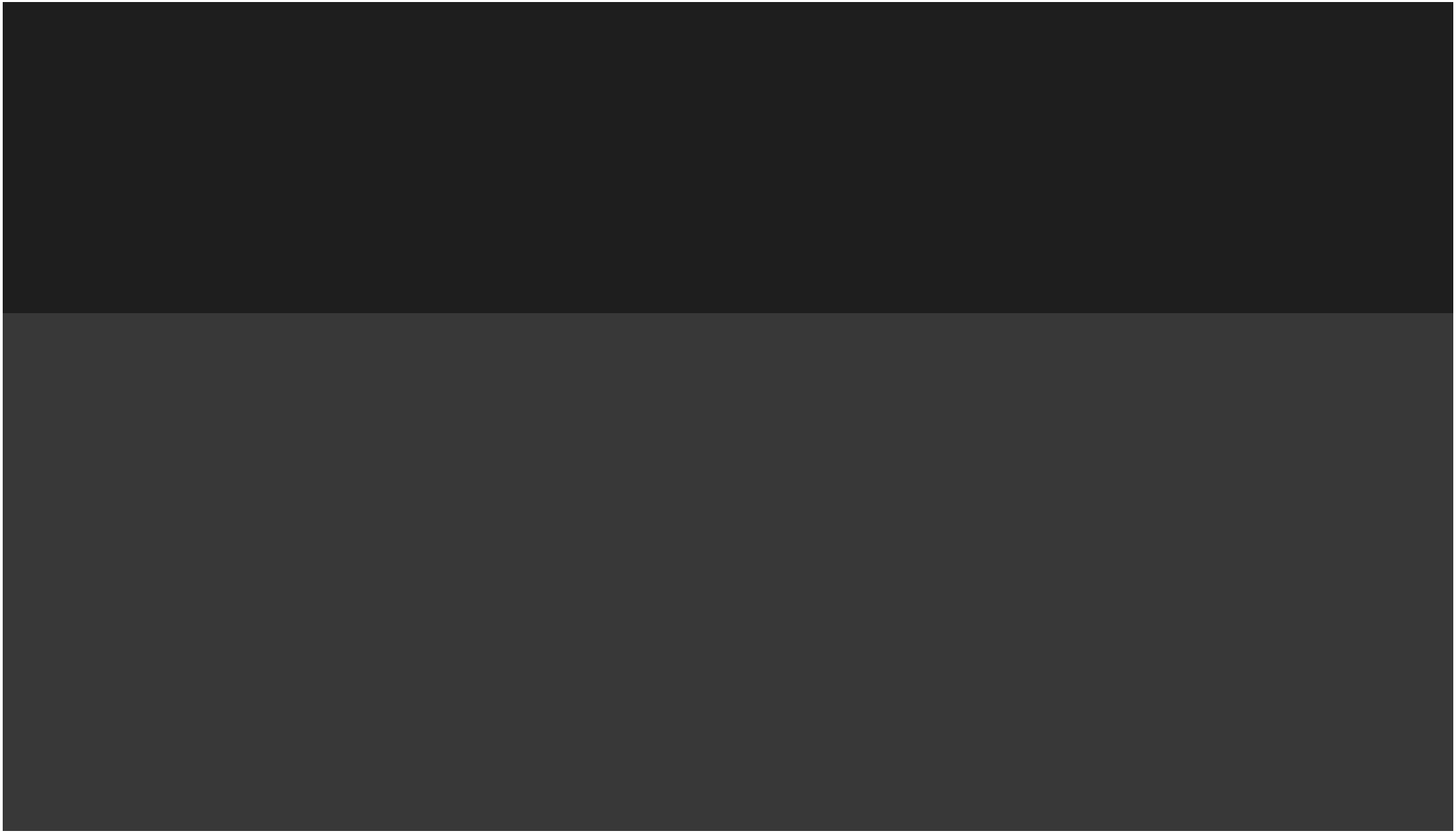
```
from google.colab import drive
drive.mount("/content/drive")
```

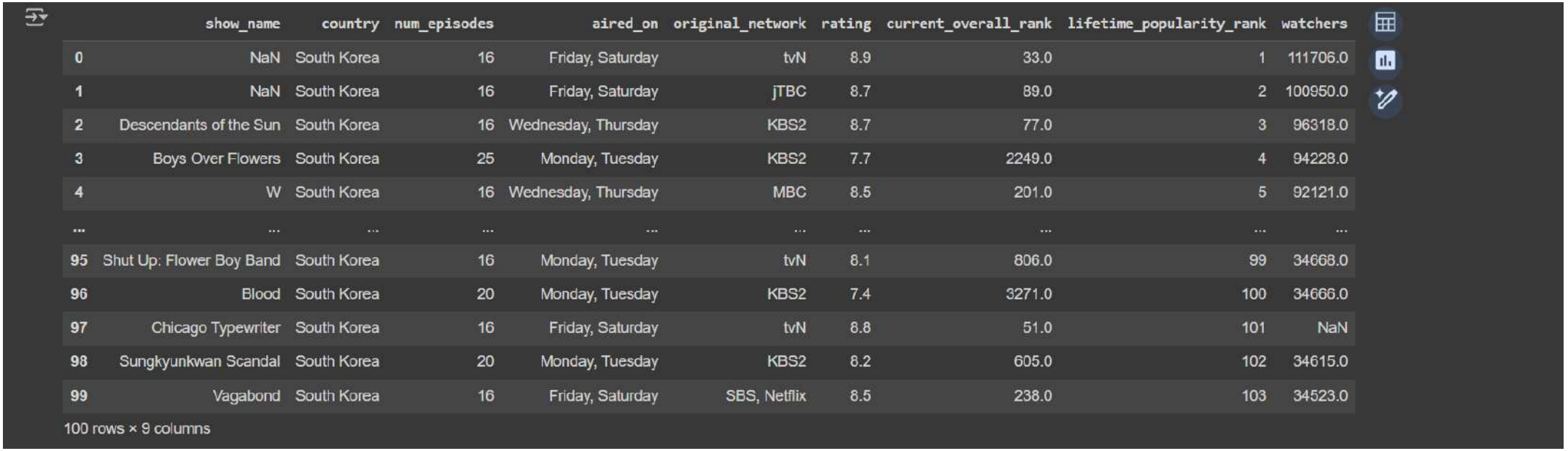
```
Drive already mounted at /content/drive; to attempt to forcibly remount, call drive.mount("/content/drive", force_remount=True).
```

```
ls drive/MyDrive/'Colab Notebooks'/Data_set.csv
```

```
'drive/MyDrive/Colab Notebooks/Data_set.csv'
```

```
import pandas as pd
df=pd.read_csv('drive/MyDrive/Colab Notebooks/Data_set.csv')
df
```





	show_name	country	num_episodes	aired_on	original_network	rating	current_overall_rank	lifetime_popularity_rank	watchers
0	NaN	South Korea	16	Friday, Saturday	tvN	8.9	33.0	1	111706.0
1	NaN	South Korea	16	Friday, Saturday	jTBC	8.7	89.0	2	100950.0
2	Descendants of the Sun	South Korea	16	Wednesday, Thursday	KBS2	8.7	77.0	3	96318.0
3	Boys Over Flowers	South Korea	25	Monday, Tuesday	KBS2	7.7	2249.0	4	94228.0
4	W	South Korea	16	Wednesday, Thursday	MBC	8.5	201.0	5	92121.0
...	...	...	...	...	...	...	...	...	...
95	Shut Up: Flower Boy Band	South Korea	16	Monday, Tuesday	tvN	8.1	806.0	99	34668.0
96	Blood	South Korea	20	Monday, Tuesday	KBS2	7.4	3271.0	100	34666.0
97	Chicago Typewriter	South Korea	16	Friday, Saturday	tvN	8.8	51.0	101	NaN
98	Sungkyunkwan Scandal	South Korea	20	Monday, Tuesday	KBS2	8.2	605.0	102	34615.0
99	Vagabond	South Korea	16	Friday, Saturday	SBS, Netflix	8.5	238.0	103	34523.0

100 rows × 9 columns

```python
# CHECK OUT NULL VALUES IN DATA SET USING FUNCTION
df_null=df.isnull()
df_null
```

Python

	show_name	country	num_episodes	aired_on	original_network	rating	current_overall_rank	lifetime_popularity_rank	watchers
0	True	False	False	False	False	False	False	False	False
1	True	False	False	False	False	False	False	False	False
2	False	False	False	False	False	False	False	False	False
3	False	False	False	False	False	False	False	False	False
4	False	False	False	False	False	False	False	False	False
...	...	...	...	...	...	...	...	...	...
95	False	False	False	False	False	False	False	False	False
96	False	False	False	False	False	False	False	False	False
97	False	False	False	False	False	False	False	False	True
98	False	False	False	False	False	False	False	False	False
99	False	False	False	False	False	False	False	False	False

100 rows × 9 columns

```python
# DISPLAY THE SUM ON NULL VALUES IN EACH ROWS
df_null_sum=df.isnull().sum()
```

```python
# DISPLAY THE SUM ON NULL VALUES IN EACH ROWS
df_null_sum=df.isnull().sum()
df_null_sum
```

Python

	0
show_name	4
country	0
num_episodes	0
aired_on	1
original_network	1
rating	4
current_overall_rank	3
lifetime_popularity_rank	0
watchers	3

dtype: int64

```python
# DROP NULL VALUES
df_dropna=df.isnull().dropna()
df_dropna
```

Python

	show_name	country	num_episodes	aired_on	original_network	rating	current_overall_rank	lifetime_popularity_rank	watchers
0	True	False	False	False	False	False	False	False	False

	show_name	country	num_episodes	aired_on	original_network	rating	current_overall_rank	lifetime_popularity_rank	watchers
0	True	False	False	False	False	False	False	False	False
1	True	False	False	False	False	False	False	False	False
2	False	False	False	False	False	False	False	False	False
3	False	False	False	False	False	False	False	False	False
4	False	False	False	False	False	False	False	False	False
...	...	...	...	...	...	...	...	...	...
95	False	False	False	False	False	False	False	False	False
96	False	False	False	False	False	False	False	False	False
97	False	False	False	False	False	False	False	False	True
98	False	False	False	False	False	False	False	False	False
99	False	False	False	False	False	False	False	False	False

100 rows × 9 columns

```python
# FILL NULL VALUES WITH CONSTANT VALUE "0"
df_nafill_0=df.fillna(0)
df_nafill_0
```

Python

	show_name	country	num_episodes	aired_on	original_network	rating	current_overall_rank	lifetime_popularity_rank	watchers
0	0	South Korea	16	Friday, Saturday	tvN	8.9	33.0	1	111706.0
1	0	South Korea	16	Friday, Saturday	jTBC	8.7	89.0	2	100950.0
2	Descendants of the Sun	South Korea	16	Wednesday, Thursday	KBS2	8.7	77.0	3	96318.0

2	Descendants of the Sun	South Korea	16	Wednesday, Thursday	KBS2	8.7	77.0	3	96318.0
3	Boys Over Flowers	South Korea	25	Monday, Tuesday	KBS2	7.7	2249.0	4	94228.0
4	W	South Korea	16	Wednesday, Thursday	MBC	8.5	201.0	5	92121.0
...	...	...	...	...	...	...	...	...	...
95	Shut Up: Flower Boy Band	South Korea	16	Monday, Tuesday	tvN	8.1	806.0	99	34668.0
96	Blood	South Korea	20	Monday, Tuesday	KBS2	7.4	3271.0	100	34666.0
97	Chicago Typewriter	South Korea	16	Friday, Saturday	tvN	8.8	51.0	101	0.0
98	Sungkyunkwan Scandal	South Korea	20	Monday, Tuesday	KBS2	8.2	605.0	102	34615.0
99	Vagabond	South Korea	16	Friday, Saturday	SBS, Netflix	8.5	238.0	103	34523.0

100 rows × 9 columns

```python
# FILL NULL VALUES WITH ffill METHOD
df_ffill=df.ffill()
df_ffill
```

Python

...

	show_name	country	num_episodes	aired_on	original_network	rating	current_overall_rank	lifetime_popularity_rank	watchers
0	NaN	South Korea	16	Friday, Saturday	tvN	8.9	33.0	1	111706.0
1	NaN	South	16	Friday, Saturday	jTBC	8.7	89.0	2	100950.0

0	NaN	South Korea	16	Friday, Saturday	tvN	8.9	33.0	1	111706.0
1	NaN	South Korea	16	Friday, Saturday	jTBC	8.7	89.0	2	100950.0
2	Descendants of the Sun	South Korea	16	Wednesday, Thursday	KBS2	8.7	77.0	3	96318.0
3	Boys Over Flowers	South Korea	25	Monday, Tuesday	KBS2	7.7	2249.0	4	94228.0
4	W	South Korea	16	Wednesday, Thursday	MBC	8.5	201.0	5	92121.0
...	...	...	...	...	...	...	...	...	...
95	Shut Up: Flower Boy Band	South Korea	16	Monday, Tuesday	tvN	8.1	806.0	99	34668.0
96	Blood	South Korea	20	Monday, Tuesday	KBS2	7.4	3271.0	100	34666.0
97	Chicago Typewriter	South Korea	16	Friday, Saturday	tvN	8.8	51.0	101	34666.0
98	Sungkyunkwan Scandal	South Korea	20	Monday, Tuesday	KBS2	8.2	605.0	102	34615.0
99	Vagabond	South Korea	16	Friday, Saturday	SBS, Netflix	8.5	238.0	103	34523.0

100 rows × 9 columns

```python
# FILL NULL VALUES WITH bfill METHOD
df_bfill=df.bfill()
df_bfill
```

Python

```python
# FILL NULL VALUES WITH bfill METHOD
df_bfill=df.bfill()
df_bfill
```

Python

	show_name	country	num_episodes	aired_on	original_network	rating	current_overall_rank	lifetime_popularity_rank	watchers
0	Descendants of the Sun	South Korea	16	Friday, Saturday	tvN	8.9	33.0	1	111706.0
1	Descendants of the Sun	South Korea	16	Friday, Saturday	jTBC	8.7	89.0	2	100950.0
2	Descendants of the Sun	South Korea	16	Wednesday, Thursday	KBS2	8.7	77.0	3	96318.0
3	Boys Over Flowers	South Korea	25	Monday, Tuesday	KBS2	7.7	2249.0	4	94228.0
4	W	South Korea	16	Wednesday, Thursday	MBC	8.5	201.0	5	92121.0
...	...	...	...	...	...	...	...	...	...
95	Shut Up: Flower Boy Band	South Korea	16	Monday, Tuesday	tvN	8.1	806.0	99	34668.0
96	Blood	South Korea	20	Monday, Tuesday	KBS2	7.4	3271.0	100	34666.0
97	Chicago Typewriter	South Korea	16	Friday, Saturday	tvN	8.8	51.0	101	34615.0
98	Sungkyunkwan Scandal	South Korea	20	Monday, Tuesday	KBS2	8.2	605.0	102	34615.0
99	Vagabond	South Korea	16	Friday, Saturday	SBS, Netflix	8.5	238.0	103	34523.0

```python
# CALCULATE MEAN VALUE OF A COLUMN AND FILL IT WITH NULL VALUES
df_mean1=df['num_episodes'].fillna(df['num_episodes'].mean())
df_mean1
```

Python

	num_episodes
0	16
1	16
2	16
3	25
4	16
...	...
95	16
96	20
97	16
98	20
99	16

100 rows × 1 columns

dtype: int64

```python
df_mean2=df['rating'].fillna(df['rating'].mean())
df_mean2
```

```python
df_mean2=df['rating'].fillna(df['rating'].mean())
df_mean2
```

Python

	rating
0	8.9
1	8.7
2	8.7
3	7.7
4	8.5
...	...
95	8.1
96	7.4
97	8.8
98	8.2
99	8.5

100 rows × 1 columns

dtype: float64

```python
df_mean3=df['current_overall_rank'].fillna(df['current_overall_rank'].mean())
df_mean3
```

Python

	current_overall_rank
0	33.0

0	33.0
1	89.0
2	77.0
3	2249.0
4	201.0
...	...
95	806.0
96	3271.0
97	51.0
98	605.0
99	238.0

100 rows × 1 columns

dtype: float64

```
df_mean4=df['lifetime_popularity_rank'].fillna(df['lifetime_popularity_rank'].mean())
df_mean4
```

Python

	lifetime_popularity_rank
0	1
1	2
2	3

0	1
1	2
2	3
3	4
4	5
...	...
95	99
96	100
97	101
98	102
99	103

100 rows × 1 columns

dtype: int64

```
df_mean5=df['watchers'].fillna(df['watchers'].mean())
df_mean5
```

Python

	watchers
0	111706.000000
1	100950.000000
2	96318.000000
3	94228.000000

95	34668.000000
96	34666.000000
97	52994.907216
98	34615.000000
99	34523.000000

100 rows × 1 columns

dtype: float64

```python
# DROP NULL VALUES
df_dropna=df.dropna()
df_dropna
```

Python

	show_name	country	num_episodes	aired_on	original_network	rating	current_overall_rank	lifetime_popularity_rank	watchers
2	Descendants of the Sun	South Korea	16	Wednesday, Thursday	KBS2	8.7	77.0	3	96318.0
3	Boys Over Flowers	South Korea	25	Monday, Tuesday	KBS2	7.7	2249.0	4	94228.0
4	W	South Korea	16	Wednesday, Thursday	MBC	8.5	201.0	5	92121.0
5	You Who Came from the Stars	South Korea	21	Wednesday, Thursday	SBS	8.6	112.0	6	91360.0
6	Weightlifting Fairy Kim Bok	South	16	Wednesday,	MBC	8.8	40.0	7	91330.0

		Korea		Thursday					
95	Shut Up: Flower Boy Band	South Korea	16	Monday, Tuesday	tvN	8.1	806.0	99	34668.0
96	Blood	South Korea	20	Monday, Tuesday	KBS2	7.4	3271.0	100	34666.0
98	Sungkyunkwan Scandal	South Korea	20	Monday, Tuesday	KBS2	8.2	605.0	102	34615.0
99	Vagabond	South Korea	16	Friday, Saturday	SBS, Netflix	8.5	238.0	103	34523.0

86 rows × 9 columns

Outlier Detection and Removal - IQR

```python
import pandas as pd
import seaborn as sns
```

Python

```python
age=[1,3,28,27,25,92,30,39,40,50,26,24,29,94]
af=pd.DataFrame(age)
af
```

Python

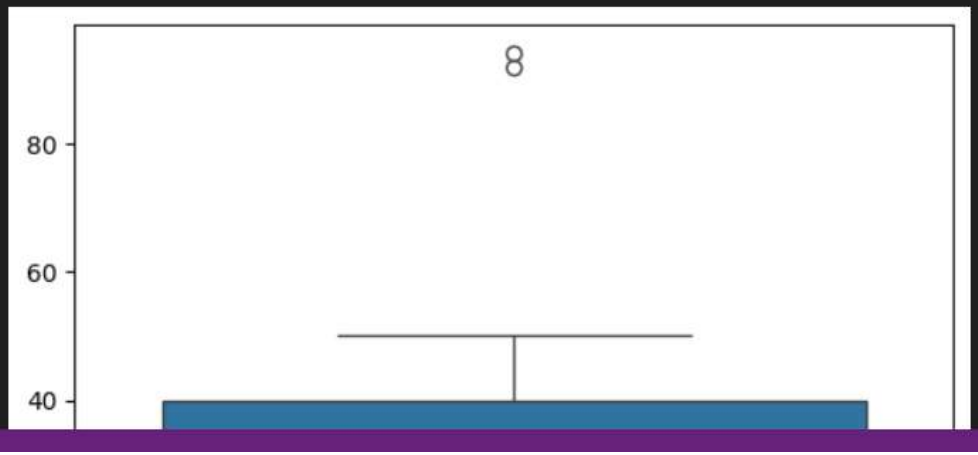
	0
0	1

4	25
5	92
6	30
7	39
8	40
9	50
10	26
11	24
12	29
13	94

```python
# USE BOXPLOT FUNCTION HERE TO DETECT OUTLIER
sns.boxplot(af)
```

Python

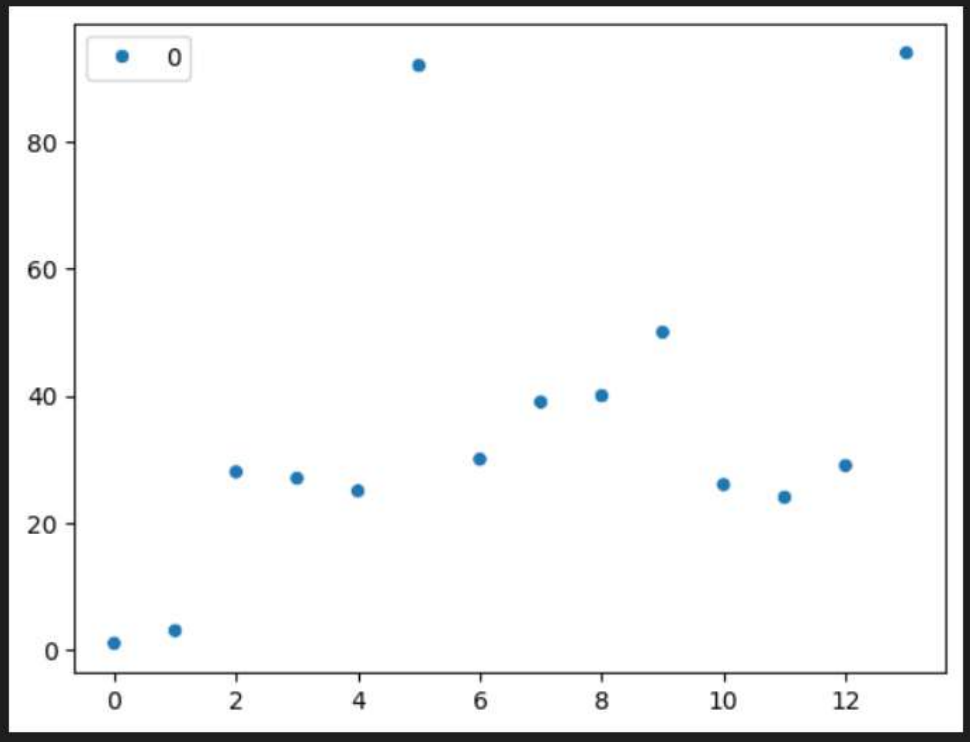
```
<Axes: >
```



```
sns.scatterplot(af)
```

Python

```
<Axes: >
```



```
q1=af.quantile(0.25)
q2=af.quantile(0.5)
q3=af.quantile(0.75)
```

```
iqr=q3-q1
iqr
```

```
q1=af.quantile(0.25)
q2=af.quantile(0.5)
q3=af.quantile(0.75)
```

```
iqr=q3-q1
iqr
```

```
...
```

	0
0	14.5

dtype: float64

```
import numpy as np
```

```
Q1=np.percentile(af,25)
Q2=np.percentile(af,50)
Q3=np.percentile(af,75)
```

```
IQR=Q3-Q1
```

```
lower_bound=Q1-1.5*IQR
upper_bound=Q3+1.5*IQR
```

```
outliers = [x for x in age if x < lower_bound or x > upper_bound]
```

+ Code + Markdown

```
print('Q1:',Q1)
print('Q3:',Q3)
print('IQR:',IQR)
print('Lower bound:',lower_bound)
print('Upper bound:',upper_bound)
print('Outliers:',outliers)
```

```
Q1: 25.25
Q3: 39.75
IQR: 14.5
Lower bound: 3.5
Upper bound: 61.5
Outliers: [1, 3, 92, 94]
```

```
af=af[((af>=lower_bound)&(af<=upper_bound))]
af
```

Python

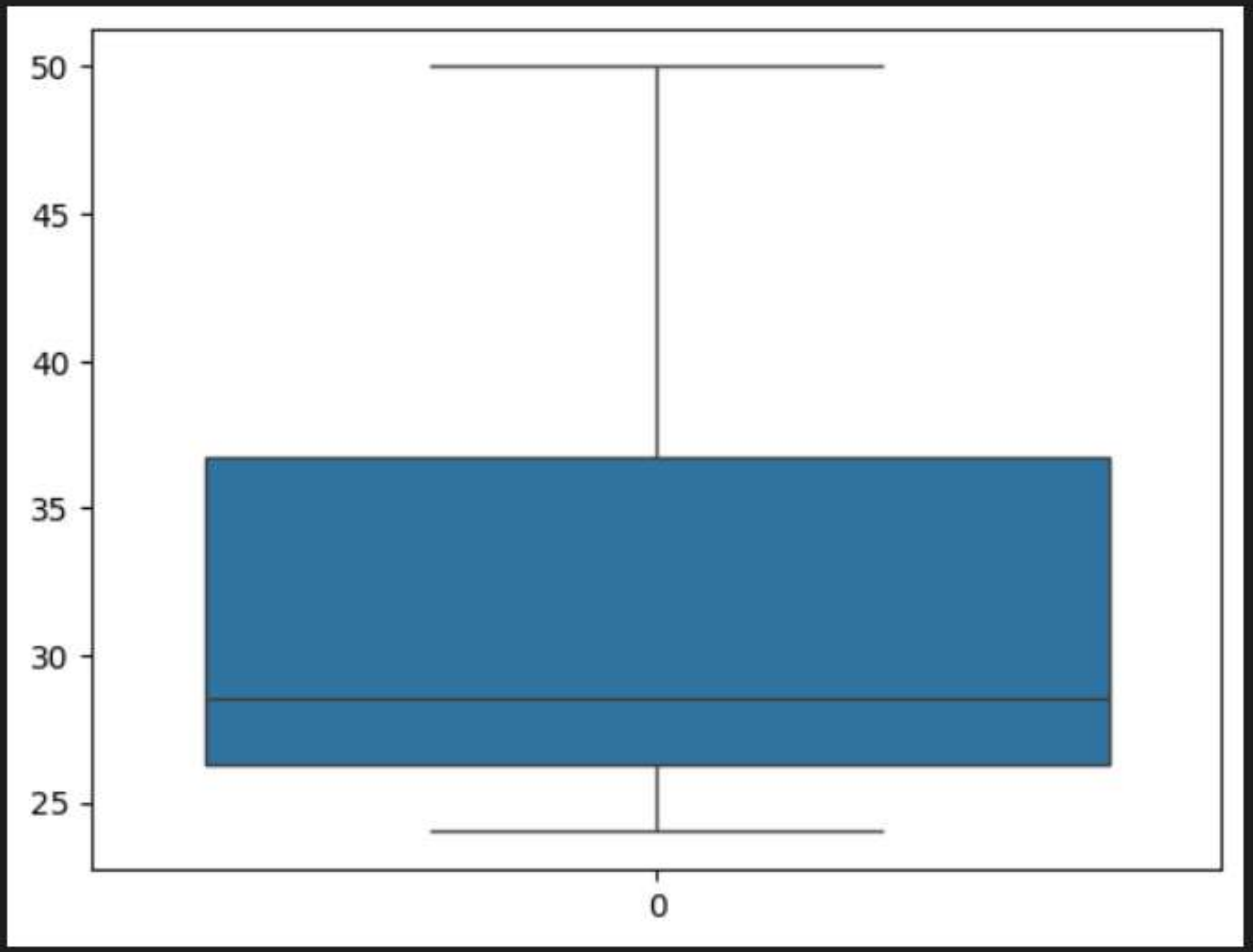
	0
0	NaN
1	NaN
2	28.0
3	27.0
4	25.0
5	NaN
6	30.0
7	39.0
8	40.0
9	50.0
10	26.0
11	24.0
12	29.0
13	NaN

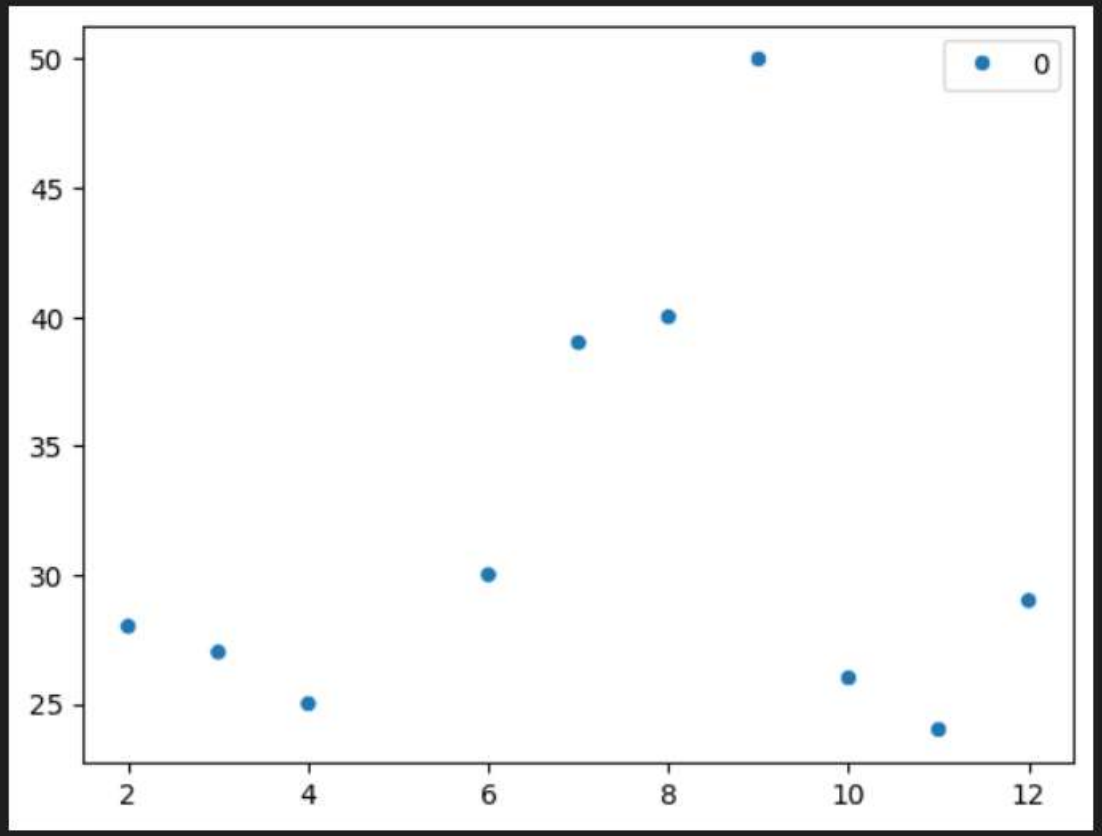
```
af.dropna()
```

Python

	0
2	28.0
3	27.0
4	25.0

```
<Axes: >
```





Z Score

```python
from scipy import stats #STATS METHOD IS USED TO IMPLEMENT Z SCORE METHOD
import numpy as np
import pandas as pd
import seaborn as sns
```

Python

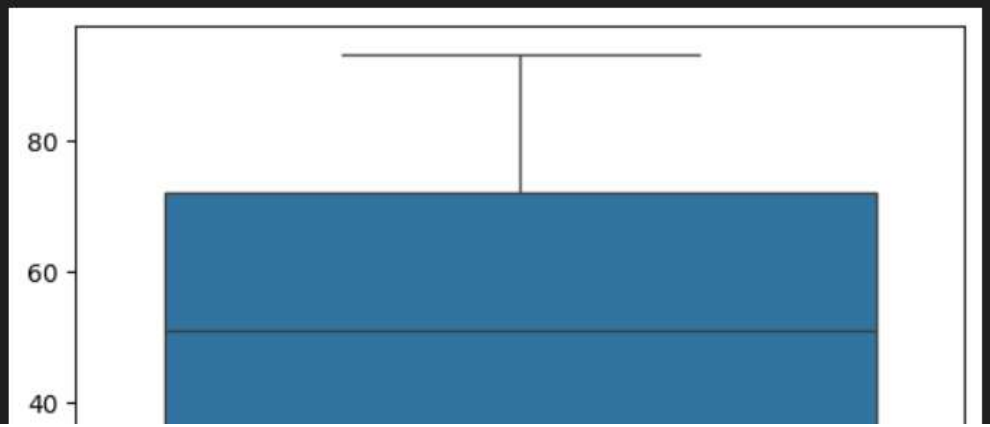
```python
data=[1,12,15,18,21,24,27,30,33,36,39,42,45,48,51,54,57,60,63,66,69,72,75,78,81,84,87,90,93]
df=pd.DataFrame(data)
```

Python

```python
# USE BOXPLOT FUNCTION HERE TO DETECT OUTLIER
sns.boxplot(df)
```

Python

```
<Axes: >
```



```python
mean=np.mean(data)
mean
```

Python

```
50.724137931034484
```

```python
std=np.std(data)
std
```

Python

```
25.59889080534025
```

```python
# PERFORM Z SCORE METHOD AND DETECT OUTLIER VALUES
z=np.abs(stats.zscore(df))
z
```

Python

	0
0	1.942433
1	1.512727
2	1.395535
3	1.278342
4	1.161149
5	1.043957
6	0.926764
7	0.809572
8	0.692379
9	0.575187

15	0.127969
16	0.245161
17	0.362354
18	0.479547
19	0.596739
20	0.713932
21	0.831124
22	0.948317
23	1.065510
24	1.182702
25	1.299895
26	1.417087
27	1.534280
28	1.651472

```python
threshold=3
outliers = df[abs(df) > 3]
print("Outliers:")
print(outliers)
```

Python

```
Outliers:
      0
0   NaN
1  12.0
2  15.0
3  18.0
```

```
7   30.0
8   33.0
9   36.0
10  39.0
11  42.0
12  45.0
13  48.0
14  51.0
15  54.0
16  57.0
17  60.0
18  63.0
19  66.0
20  69.0
21  72.0
22  75.0
...
25  84.0
26  87.0
27  90.0
28  93.0
```

Output is truncated. View as a scrollable element or open in a text editor. Adjust cell output settings...

```python
# Remove outliers
df_cleaned = df[(z <= threshold)]
df_cleaned
```

Python

```
...
```

	0
0	1

```
16  57
17  60
18  63
19  66
20  69
21  72
22  75
23  78
24  81
25  84
26  87
27  90
28  93
```

```python
# USE BOXPLOT FUNCTION HERE TO CHECK OUTLIER IS REMOVED
sns.boxplot(df_cleaned)
```

Python

```
<Axes: >
```