zpexv4whb

February 19, 2025

[1]:
```
import pandas as pd
```

[3]:
```
import numpy as np
```

[5]:
```
df=pd.read_csv('StudentsPerformance.csv')
```

[7]:
```
df
```

[7]:

	Math_Score	Reading_Score	Writing_Score	Placement_Score	Club_Join_Date	Placement_Offer_Count
0	71	88	61.0	100	2018	4
1	61	90	68.0	80	2018	2
2	63	90	70.0	88	2019	2
3	65	95	63.0	96	2020	
4	66	81	63.0	92	2019	
5	80	88	70.0	89	2021	
6	75	90	NaN	89	2020	
7	78	98	75.0	100	2021	
8	67	94	63.0	83	2020	
9	69	98	78.0	91	2021	
10	80	78	89.0	78	2021	
11	77	89	64.0	98	2021	
12	75	93	80.0	95	2018	
13	67	89	72.0	93	2020	
14	63	80	61.0	80	2018	
15	80	82	66.0	83	2021	
16	67	85	74.0	95	2018	
17	73	81	72.0	84	2019	
18	75	86	76.0	76	2020	
19	67	86	74.0	89	2021	
20	74	76	71.0	89	2018	
21	73	88	77.0	85	2021	
22	60	91	67.0	78	2018	

```
3                     3
4                     3
5                     2
6                     3
7                     4
8                     2
9                     3
10                    1
11                    3
12                    3
13                    3
14                    2
15                    2
16                    3
17                    2
18                    1
19                    2
20                    2
21                    2
22                    1
```

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

[9]:

	Math_Score	Reading_Score	Writing_Score	Placement_Score	Club_Join_Date	\
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	True	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	

```
    Placement_Offer_Count
0                   False
1                   False
2                   False
3                   False
4                   False
5                   False
6                   False
7                   False
8                   False
9                   False
10                  False
11                  False
12                  False
13                  False
14                  False
15                  False
16                  False
17                  False
18                  False
19                  False
20                  False
21                  False
22                  False
```

[11]:
```
series=pd.isnull(df["Writing_Score"])
df[series]
```

[11]:
```
   Math_Score  Reading_Score  Writing_Score  Placement_Score  Club_Join_Date  \
6          75             90            NaN               89            2020

   Placement_Offer_Count
6                      3
```

[13]:
```
df.notnull()
```

[13]:
```
    Math_Score  Reading_Score  Writing_Score  Placement_Score  Club_Join_Date  \
0         True           True           True             True            True
1         True           True           True             True            True
2         True           True           True             True            True
3         True           True           True             True            True
4         True           True           True             True            True
5         True           True           True             True            True
6         True           True          False             True            True
7         True           True           True             True            True
8         True           True           True             True            True
```

```
9         True           True           True           True           True
10        True           True           True           True           True
11        True           True           True           True           True
12        True           True           True           True           True
13        True           True           True           True           True
14        True           True           True           True           True
15        True           True           True           True           True
16        True           True           True           True           True
17        True           True           True           True           True
18        True           True           True           True           True
19        True           True           True           True           True
20        True           True           True           True           True
21        True           True           True           True           True
22        True           True           True           True           True

    Placement_Offer_Count
0                    True
1                    True
2                    True
3                    True
4                    True
5                    True
6                    True
7                    True
8                    True
9                    True
10                   True
11                   True
12                   True
13                   True
14                   True
15                   True
16                   True
17                   True
18                   True
19                   True
20                   True
21                   True
22                   True
```

[15]:
```
series=pd.notnull(df["Writing_Score"])
df[series]
```

[15]:
```
    Math_Score  Reading_Score  Writing_Score  Placement_Score  Club_Join_Date  \
0           71             88           61.0              100            2018
1           61             90           68.0               80            2018
2           63             90           70.0               88            2019
```

3	65	95	63.0	96	2020
4	66	81	63.0	92	2019
5	80	88	70.0	89	2021
7	78	98	75.0	100	2021
8	67	94	63.0	83	2020
9	69	98	78.0	91	2021
10	80	78	89.0	78	2021
11	77	89	64.0	98	2021
12	75	93	80.0	95	2018
13	67	89	72.0	93	2020
14	63	80	61.0	80	2018
15	80	82	66.0	83	2021
16	67	85	74.0	95	2018
17	73	81	72.0	84	2019
18	75	86	76.0	76	2020
19	67	86	74.0	89	2021
20	74	76	71.0	89	2018
21	73	88	77.0	85	2021
22	60	91	67.0	78	2018

	Placement_Offer_Count
0	4
1	2
2	2
3	3
4	3
5	2
7	4
8	2
9	3
10	1
11	3
12	3
13	3
14	2
15	2
16	3
17	2
18	1
19	2
20	2
21	2
22	1

[17]:
```
ndf=df
ndf.fillna(0)
```

[17]:

	Math_Score	Reading_Score	Writing_Score	Placement_Score	Club_Join_Date	Placement_Offer_Count
0	71	88	61.0	100	2018	4
1	61	90	68.0	80	2018	2
2	63	90	70.0	88	2019	2
3	65	95	63.0	96	2020	3
4	66	81	63.0	92	2019	3
5	80	88	70.0	89	2021	2
6	75	90	0.0	89	2020	3
7	78	98	75.0	100	2021	4
8	67	94	63.0	83	2020	2
9	69	98	78.0	91	2021	3
10	80	78	89.0	78	2021	1
11	77	89	64.0	98	2021	3
12	75	93	80.0	95	2018	3
13	67	89	72.0	93	2020	3
14	63	80	61.0	80	2018	2
15	80	82	66.0	83	2021	2
16	67	85	74.0	95	2018	3
17	73	81	72.0	84	2019	2
18	75	86	76.0	76	2020	1
19	67	86	74.0	89	2021	2
20	74	76	71.0	89	2018	2
21	73	88	77.0	85	2021	
22	60	91	67.0	78	2018	

```
21                      2
22                      1
```

[19]:
```
m_v=df['Writing_Score'].mean()
df['Writing_Score'].fillna(value=m_v, inplace=True)
df
```

```
C:\Users\Deep Shelke\AppData\Local\Temp\ipykernel_23740\3781148174.py:2:
FutureWarning: A value is trying to be set on a copy of a DataFrame or Series
through chained assignment using an inplace method.
The behavior will change in pandas 3.0. This inplace method will never work
because the intermediate object on which we are setting values always behaves as
a copy.

For example, when doing 'df[col].method(value, inplace=True)', try using
'df.method({col: value}, inplace=True)' or df[col] = df[col].method(value)
instead, to perform the operation inplace on the original object.


  df['Writing_Score'].fillna(value=m_v, inplace=True)
```

[19]:

	Math_Score	Reading_Score	Writing_Score	Placement_Score	Club_Join_Date	\
0	71	88	61.000000	100	2018	
1	61	90	68.000000	80	2018	
2	63	90	70.000000	88	2019	
3	65	95	63.000000	96	2020	
4	66	81	63.000000	92	2019	
5	80	88	70.000000	89	2021	
6	75	90	70.636364	89	2020	
7	78	98	75.000000	100	2021	
8	67	94	63.000000	83	2020	
9	69	98	78.000000	91	2021	
10	80	78	89.000000	78	2021	
11	77	89	64.000000	98	2021	
12	75	93	80.000000	95	2018	
13	67	89	72.000000	93	2020	
14	63	80	61.000000	80	2018	
15	80	82	66.000000	83	2021	
16	67	85	74.000000	95	2018	
17	73	81	72.000000	84	2019	
18	75	86	76.000000	76	2020	
19	67	86	74.000000	89	2021	
20	74	76	71.000000	89	2018	
21	73	88	77.000000	85	2021	
22	60	91	67.000000	78	2018	

	Placement_Offer_Count

```
0                    4
1                    2
2                    2
3                    3
4                    3
5                    2
6                    3
7                    4
8                    2
9                    3
10                   1
11                   3
12                   3
13                   3
14                   2
15                   2
16                   3
17                   2
18                   1
19                   2
20                   2
21                   2
22                   1
```

[21]:
```
ndf.dropna()
```

[21]:

	Math_Score	Reading_Score	Writing_Score	Placement_Score	Club_Join_Date	\
0	71	88	61.000000	100	2018	
1	61	90	68.000000	80	2018	
2	63	90	70.000000	88	2019	
3	65	95	63.000000	96	2020	
4	66	81	63.000000	92	2019	
5	80	88	70.000000	89	2021	
6	75	90	70.636364	89	2020	
7	78	98	75.000000	100	2021	
8	67	94	63.000000	83	2020	
9	69	98	78.000000	91	2021	
10	80	78	89.000000	78	2021	
11	77	89	64.000000	98	2021	
12	75	93	80.000000	95	2018	
13	67	89	72.000000	93	2020	
14	63	80	61.000000	80	2018	
15	80	82	66.000000	83	2021	
16	67	85	74.000000	95	2018	
17	73	81	72.000000	84	2019	
18	75	86	76.000000	76	2020	
19	67	86	74.000000	89	2021	

```
20          74             76      71.000000              89            2018
21          73             88      77.000000              85            2021
22          60             91      67.000000              78            2018

    Placement_Offer_Count
0                       4
1                       2
2                       2
3                       3
4                       3
5                       2
6                       3
7                       4
8                       2
9                       3
10                      1
11                      3
12                      3
13                      3
14                      2
15                      2
16                      3
17                      2
18                      1
19                      2
20                      2
21                      2
22                      1
```

[23]:
```
new_data = ndf.dropna(axis = 0, how ='any')
new_data
```

[23]:
```
    Math_Score  Reading_Score  Writing_Score  Placement_Score  Club_Join_Date  \
0           71             88      61.000000              100            2018
1           61             90      68.000000               80            2018
2           63             90      70.000000               88            2019
3           65             95      63.000000               96            2020
4           66             81      63.000000               92            2019
5           80             88      70.000000               89            2021
6           75             90      70.636364               89            2020
7           78             98      75.000000              100            2021
8           67             94      63.000000               83            2020
9           69             98      78.000000               91            2021
10          80             78      89.000000               78            2021
11          77             89      64.000000               98            2021
12          75             93      80.000000               95            2018
13          67             89      72.000000               93            2020
```

```
14          63             80      61.000000              80            2018
15          80             82      66.000000              83            2021
16          67             85      74.000000              95            2018
17          73             81      72.000000              84            2019
18          75             86      76.000000              76            2020
19          67             86      74.000000              89            2021
20          74             76      71.000000              89            2018
21          73             88      77.000000              85            2021
22          60             91      67.000000              78            2018

    Placement_Offer_Count
0                       4
1                       2
2                       2
3                       3
4                       3
5                       2
6                       3
7                       4
8                       2
9                       3
10                      1
11                      3
12                      3
13                      3
14                      2
15                      2
16                      3
17                      2
18                      1
19                      2
20                      2
21                      2
22                      1
```

[25]:
```
df
```

[25]:
```
    Math_Score  Reading_Score  Writing_Score  Placement_Score  Club_Join_Date  \
0           71             88      61.000000              100            2018
1           61             90      68.000000               80            2018
2           63             90      70.000000               88            2019
3           65             95      63.000000               96            2020
4           66             81      63.000000               92            2019
5           80             88      70.000000               89            2021
6           75             90      70.636364               89            2020
7           78             98      75.000000              100            2021
8           67             94      63.000000               83            2020
```

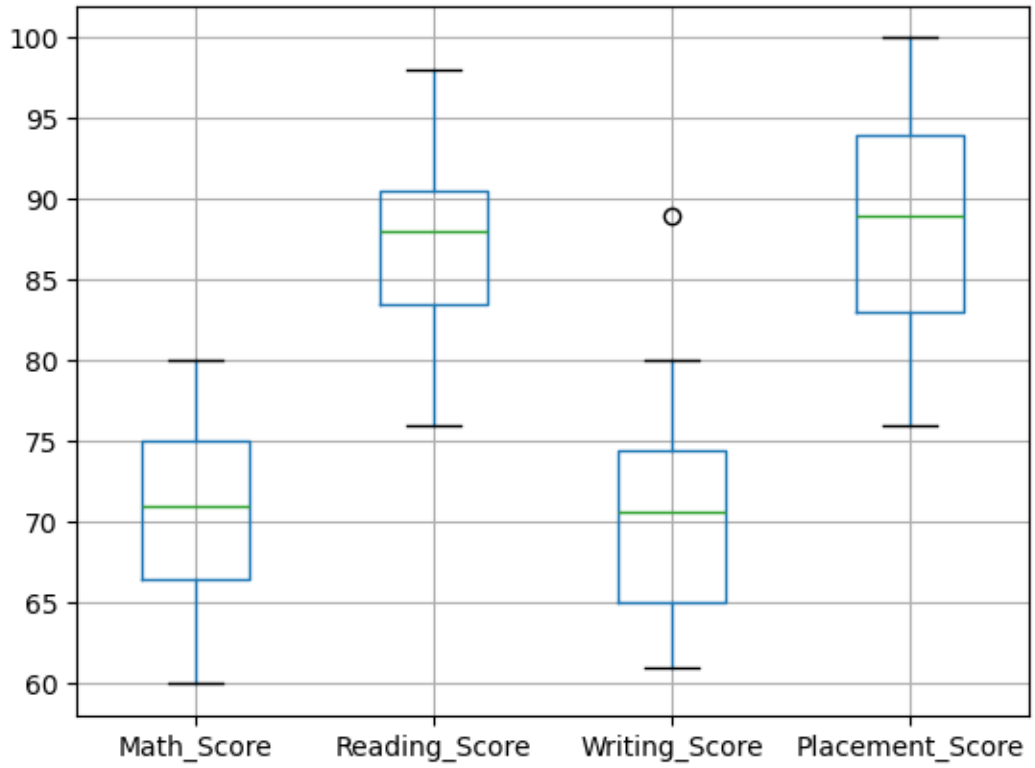
9	69	98	78.000000	91	2021
10	80	78	89.000000	78	2021
11	77	89	64.000000	98	2021
12	75	93	80.000000	95	2018
13	67	89	72.000000	93	2020
14	63	80	61.000000	80	2018
15	80	82	66.000000	83	2021
16	67	85	74.000000	95	2018
17	73	81	72.000000	84	2019
18	75	86	76.000000	76	2020
19	67	86	74.000000	89	2021
20	74	76	71.000000	89	2018
21	73	88	77.000000	85	2021
22	60	91	67.000000	78	2018

	Placement_Offer_Count
0	4
1	2
2	2
3	3
4	3
5	2
6	3
7	4
8	2
9	3
10	1
11	3
12	3
13	3
14	2
15	2
16	3
17	2
18	1
19	2
20	2
21	2
22	1

[27]:
```
col = ['Math_Score', 'Reading_Score' , 'Writing_Score','Placement_Score']
```

[29]:
```
df.boxplot(col)
```

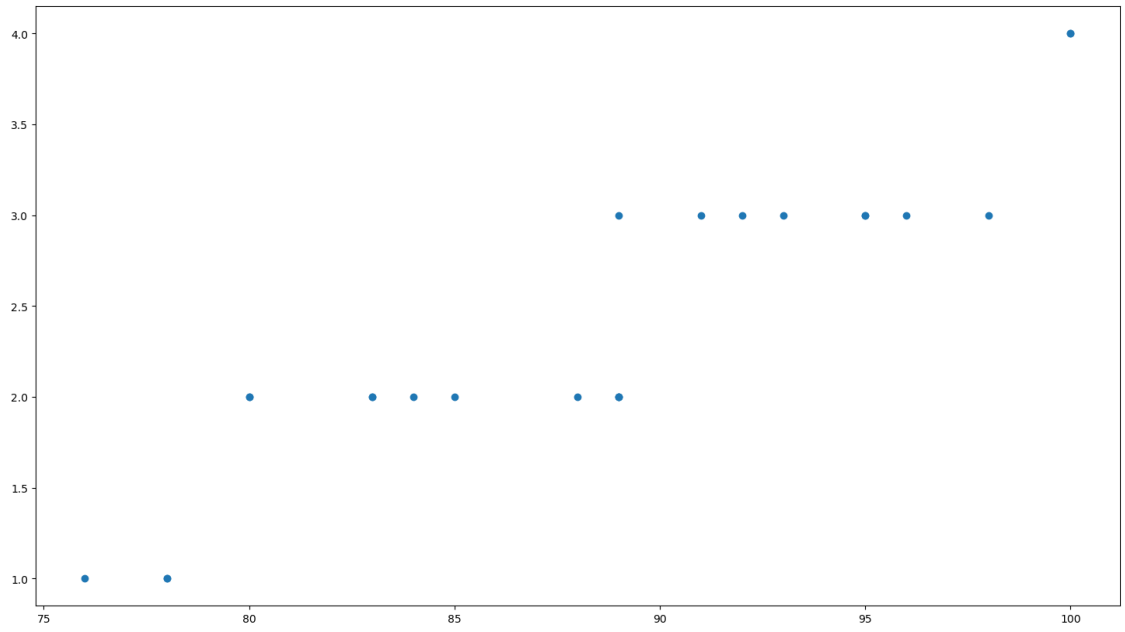
[29]: <Axes: >



col = [‘reading_score’] df.boxplot(col)

[32]:
```
import matplotlib.pyplot as plt
```

[34]:
```
fig, ax = plt.subplots(figsize = (18,10))
ax.scatter(df['Placement_Score'], df['Placement_Offer_Count'])
plt.show()
print(np.where((df['Placement_Score']<50) & (df['Placement_Offer_Count']>1)))
print(np.where((df['Placement_Score']>85) & (df['Placement_Offer_Count']<3)))
(array([], dtype=int64),)
(array([], dtype=int64),)
```




```
(array([], dtype=int64),)
(array([ 2,  5, 19, 20], dtype=int64),)
```

```
---------------------------------------------------------------------------
NameError                                 Traceback (most recent call last)
Cell In[34], line 6
      4 print(np.where((df['Placement_Score']<50) &␣
 ↪(df['Placement_Offer_Count']>1)))
      5 print(np.where((df['Placement_Score']>85) &␣
 ↪(df['Placement_Offer_Count']<3)))
----> 6 (array([], dtype=int64),)
      7 (array([], dtype=int64),)

NameError: name 'array' is not defined
```

[38]:
```
from scipy import stats
```

[39]:
```
z = np.abs(stats.zscore(df['Math_Score']))
```

[42]:
```
z
```

[42]:
```
0    0.049517
1    1.577479
2    1.252080
3    0.926681
4    0.763981
```

```
5     1.513814
6     0.700316
7     1.188415
8     0.601281
9     0.275882
10    1.513814
11    1.025715
12    0.700316
13    0.601281
14    1.252080
15    1.513814
16    0.601281
17    0.374917
18    0.700316
19    0.601281
20    0.537616
21    0.374917
22    1.740179
Name: Math_Score, dtype: float64
```

[64]:
```
threshold = 0.18
sample_outliers = np.where(z <threshold)
sample_outliers
```

[64]:
```
(array([0], dtype=int64),)
```

[44]:
```
#iqr
sorted_rscore= sorted(df['Reading_Score'])
```

[46]:
```
sorted_rscore
```

[46]:
```
[76,
 78,
 80,
 81,
 81,
 82,
 85,
 86,
 86,
 88,
 88,
 88,
 89,
 89,
 90,
 90,
```

```
 90,
 91,
 93,
 94,
 95,
 98,
 98]
```

[48]:
```
q1 = np.percentile(sorted_rscore, 25)
q3 = np.percentile(sorted_rscore, 75)
```

[50]:
```
print(q1,q3)
```

```
83.5 90.5
```

[52]:
```
IQR = q3-q1
lwr_bound = q1-(1.5*IQR)
upr_bound = q3+(1.5*IQR)
print(lwr_bound, upr_bound)
```

```
73.0 101.0
```

[68]:
```
#Handling Outliers
new_df=df
for i in sample_outliers:
  new_df.drop(i,inplace=True)
new_df
```

[68]:

	Math_Score	Reading_Score	Writing_Score	Placement_Score	Club_Join_Date	\
1	61	90	68.000000	80	2018	
2	63	90	70.000000	88	2019	
3	65	95	63.000000	96	2020	
4	66	81	63.000000	92	2019	
5	80	88	70.000000	89	2021	
6	75	90	70.636364	89	2020	
7	78	98	75.000000	100	2021	
8	67	94	63.000000	83	2020	
9	69	98	78.000000	91	2021	
10	80	78	89.000000	78	2021	
11	77	89	64.000000	98	2021	
12	75	93	80.000000	95	2018	
13	67	89	72.000000	93	2020	
14	63	80	61.000000	80	2018	
15	80	82	66.000000	83	2021	
16	67	85	74.000000	95	2018	
17	73	81	72.000000	84	2019	
18	75	86	76.000000	76	2020	

19	67	86	74.000000	89	2021
20	74	76	71.000000	89	2018
21	73	88	77.000000	85	2021
22	60	91	67.000000	78	2018

	Placement_Offer_Count
1	2
2	2
3	3
4	3
5	2
6	3
7	4
8	2
9	3
10	1
11	3
12	3
13	3
14	2
15	2
16	3
17	2
18	1
19	2
20	2
21	2
22	1

[72]:
```
#Quantile based flooring and capping:
df_stud=df
ninetieth_percentile = np.percentile(df_stud['Math_Score'], 90)
b = np.where(df_stud['Math_Score']>ninetieth_percentile,
ninetieth_percentile, df_stud['Math_Score'])
print("New array:",b)
```

```
New array: [61.  63.  65.  66.  79.8 75.  78.  67.  69.  79.8 77.  75.  67.  63.
 79.8 67.  73.  75.  67.  74.  73.  60. ]
```

[74]:
```
df_stud.insert(1,"m score",b,True)
df_stud
```

[74]:

	Math_Score	m score	Reading_Score	Writing_Score	Placement_Score	\
1	61	61.0	90	68.000000	80	
2	63	63.0	90	70.000000	88	
3	65	65.0	95	63.000000	96	
4	66	66.0	81	63.000000	92	

5	80	79.8	88	70.000000	89
6	75	75.0	90	70.636364	89
7	78	78.0	98	75.000000	100
8	67	67.0	94	63.000000	83
9	69	69.0	98	78.000000	91
10	80	79.8	78	89.000000	78
11	77	77.0	89	64.000000	98
12	75	75.0	93	80.000000	95
13	67	67.0	89	72.000000	93
14	63	63.0	80	61.000000	80
15	80	79.8	82	66.000000	83
16	67	67.0	85	74.000000	95
17	73	73.0	81	72.000000	84
18	75	75.0	86	76.000000	76
19	67	67.0	86	74.000000	89
20	74	74.0	76	71.000000	89
21	73	73.0	88	77.000000	85
22	60	60.0	91	67.000000	78

	Club_Join_Date	Placement_Offer_Count
1	2018	2
2	2019	2
3	2020	3
4	2019	3
5	2021	2
6	2020	3
7	2021	4
8	2020	2
9	2021	3
10	2021	1
11	2021	3
12	2018	3
13	2020	3
14	2018	2
15	2021	2
16	2018	3
17	2019	2
18	2020	1
19	2021	2
20	2018	2
21	2021	2
22	2018	1

[76]:
```
#Mean/Median Inputation
median=np.median(sorted_rscore)
median
```

[76]: 88.0

[ ]: