DSBDAL 02

Aditya T. Salagare

Roll no : 13320 (C1 Batch)

In [45]:
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
```

In [46]:
```
df=pd.read_csv("C:/Users/Welcome/Downloads/DSBDA02.csv")
df
```

Out[46]:

	Math_Score	Reading_Score	Writing_Score	Placement_Score	Club_Join_Date	Placement_Offer_
0	72.0	86	64.0	75.0	2020.0	
1	65.0	91	67.0	111.0	2020.0	
2	66.0	81	88.0	95.0	NaN	
3	59.0	77	64.0	90.0	2000.0	
4	66.0	89	76.0	81.0	2018.0	
5	61.0	92	76.0	82.0	2019.0	
6	NaN	93	72.0	66.0	2018.0	
7	85.0	85	59.0	93.0	2019.0	
8	73.0	95	68.0	NaN	NaN	
9	65.0	81	67.0	86.0	2023.0	
10	79.0	91	NaN	100.0	2018.0	
11	75.0	80	77.0	89.0	2018.0	
12	78.0	81	73.0	69.0	2002.0	
13	65.0	93	76.0	NaN	2019.0	
14	81.0	84	17.0	76.0	2021.0	
15	62.0	80	64.0	81.0	NaN	
16	70.0	77	68.0	86.0	2021.0	
17	75.0	86	70.0	76.0	2020.0	
18	71.0	95	55.0	93.0	2011.0	
19	NaN	79	70.0	NaN	2021.0	
20	70.0	86	71.0	94.0	2018.0	
21	78.0	92	67.0	78.0	NaN	
22	66.0	82	74.0	75.0	2019.0	
23	68.0	77	77.0	100.0	2022.0	
24	65.0	75	NaN	101.0	2021.0	
25	55.0	89	73.0	91.0	2018.0	
26	73.0	88	79.0	77.0	2020.0	
27	80.0	80	68.0	83.0	2019.0	
28	74.0	92	60.0	88.0	2019.0	
29	75.0	78	66.0	97.0	2021.0	

In [47]:
```
df.columns
```

Out[47]:
```
Index(['Math_Score  ', ' Reading_Score', 'Writing_Score ', 'Placement_Score',
       'Club_Join_Date', 'Placement_Offer_Count'],
      dtype='object')
```

In [48]:
```
df.isnull()
```

Out[48]:

	Math_Score	Reading_Score	Writing_Score	Placement_Score	Club_Join_Date	Placement_Offer_
0	False	False	False	False	False	
1	False	False	False	False	False	
2	False	False	False	False	True	
3	False	False	False	False	False	
4	False	False	False	False	False	
5	False	False	False	False	False	
6	True	False	False	False	False	
7	False	False	False	False	False	
8	False	False	False	True	True	
9	False	False	False	False	False	
10	False	False	True	False	False	
11	False	False	False	False	False	
12	False	False	False	False	False	
13	False	False	False	True	False	
14	False	False	False	False	False	
15	False	False	False	False	True	
16	False	False	False	False	False	
17	False	False	False	False	False	
18	False	False	False	False	False	
19	True	False	False	True	False	
20	False	False	False	False	False	
21	False	False	False	False	True	
22	False	False	False	False	False	
23	False	False	False	False	False	
24	False	False	True	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	
29	False	False	False	False	False	

In [49]:
```
series = pd.isnull(df['Math_Score  '])
df[series]
```

Out[49]:

	Math_Score	Reading_Score	Writing_Score	Placement_Score	Club_Join_Date	Placement_Offer_
6	NaN	93	72.0	66.0	2018.0	
19	NaN	79	70.0	NaN	2021.0	

In [50]:
```
df.notnull()
```

Out[50]:

	Math_Score	Reading_Score	Writing_Score	Placement_Score	Club_Join_Date	Placement_Offer_
0	True	True	True	True	True	
1	True	True	True	True	True	
2	True	True	True	True	False	
3	True	True	True	True	True	
4	True	True	True	True	True	
5	True	True	True	True	True	
6	False	True	True	True	True	
7	True	True	True	True	True	
8	True	True	True	False	False	
9	True	True	True	True	True	
10	True	True	False	True	True	
11	True	True	True	True	True	
12	True	True	True	True	True	
13	True	True	True	False	True	
14	True	True	True	True	True	
15	True	True	True	True	False	
16	True	True	True	True	True	
17	True	True	True	True	True	
18	True	True	True	True	True	
19	False	True	True	False	True	
20	True	True	True	True	True	
21	True	True	True	True	False	
22	True	True	True	True	True	
23	True	True	True	True	True	
24	True	True	False	True	True	
25	True	True	True	True	True	
26	True	True	True	True	True	
27	True	True	True	True	True	
28	True	True	True	True	True	
29	True	True	True	True	True	

In [51]:
```
series1 = pd.notnull(df['Math_Score  '])
df[series1]
```

Out[51]:

	Math_Score	Reading_Score	Writing_Score	Placement_Score	Club_Join_Date	Placement_Offer_
0	72.0	86	64.0	75.0	2020.0	
1	65.0	91	67.0	111.0	2020.0	
2	66.0	81	88.0	95.0	NaN	
3	59.0	77	64.0	90.0	2000.0	
4	66.0	89	76.0	81.0	2018.0	
5	61.0	92	76.0	82.0	2019.0	
7	85.0	85	59.0	93.0	2019.0	
8	73.0	95	68.0	NaN	NaN	
9	65.0	81	67.0	86.0	2023.0	
10	79.0	91	NaN	100.0	2018.0	
11	75.0	80	77.0	89.0	2018.0	
12	78.0	81	73.0	69.0	2002.0	
13	65.0	93	76.0	NaN	2019.0	
14	81.0	84	17.0	76.0	2021.0	
15	62.0	80	64.0	81.0	NaN	
16	70.0	77	68.0	86.0	2021.0	
17	75.0	86	70.0	76.0	2020.0	
18	71.0	95	55.0	93.0	2011.0	
20	70.0	86	71.0	94.0	2018.0	
21	78.0	92	67.0	78.0	NaN	
22	66.0	82	74.0	75.0	2019.0	
23	68.0	77	77.0	100.0	2022.0	
24	65.0	75	NaN	101.0	2021.0	
25	55.0	89	73.0	91.0	2018.0	
26	73.0	88	79.0	77.0	2020.0	
27	80.0	80	68.0	83.0	2019.0	
28	74.0	92	60.0	88.0	2019.0	
29	75.0	78	66.0	97.0	2021.0	

In [53]:

```
missing_values = ["Na", "na"]
df = pd.read_csv("C:/Users/Welcome/Downloads/DSBDA02.csv", na_values =missing_value
df
```

Out[53]:

	Math_Score	Reading_Score	Writing_Score	Placement_Score	Club_Join_Date	Placement_Offer_
0	72.0	86	64.0	75.0	2020.0	
1	65.0	91	67.0	111.0	2020.0	
2	66.0	81	88.0	95.0	NaN	
3	59.0	77	64.0	90.0	2000.0	
4	66.0	89	76.0	81.0	2018.0	
5	61.0	92	76.0	82.0	2019.0	
6	NaN	93	72.0	66.0	2018.0	
7	85.0	85	59.0	93.0	2019.0	
8	73.0	95	68.0	NaN	NaN	
9	65.0	81	67.0	86.0	2023.0	
10	79.0	91	NaN	100.0	2018.0	
11	75.0	80	77.0	89.0	2018.0	
12	78.0	81	73.0	69.0	2002.0	
13	65.0	93	76.0	NaN	2019.0	
14	81.0	84	17.0	76.0	2021.0	
15	62.0	80	64.0	81.0	NaN	
16	70.0	77	68.0	86.0	2021.0	
17	75.0	86	70.0	76.0	2020.0	
18	71.0	95	55.0	93.0	2011.0	
19	NaN	79	70.0	NaN	2021.0	
20	70.0	86	71.0	94.0	2018.0	
21	78.0	92	67.0	78.0	NaN	
22	66.0	82	74.0	75.0	2019.0	
23	68.0	77	77.0	100.0	2022.0	
24	65.0	75	NaN	101.0	2021.0	
25	55.0	89	73.0	91.0	2018.0	
26	73.0	88	79.0	77.0	2020.0	
27	80.0	80	68.0	83.0	2019.0	
28	74.0	92	60.0	88.0	2019.0	
29	75.0	78	66.0	97.0	2021.0	

In [54]:
```
df=df.head(11)
```

In [55]:
```
ndf=df
ndf.fillna(0)
```

Out[55]:

	Math_Score	Reading_Score	Writing_Score	Placement_Score	Club_Join_Date	Placement_Offer_
0	72.0	86	64.0	75.0	2020.0	
1	65.0	91	67.0	111.0	2020.0	
2	66.0	81	88.0	95.0	0.0	
3	59.0	77	64.0	90.0	2000.0	
4	66.0	89	76.0	81.0	2018.0	
5	61.0	92	76.0	82.0	2019.0	
6	0.0	93	72.0	66.0	2018.0	
7	85.0	85	59.0	93.0	2019.0	
8	73.0	95	68.0	0.0	0.0	
9	65.0	81	67.0	86.0	2023.0	
10	79.0	91	0.0	100.0	2018.0	

In [57]:
```
m_v=df['Math_Score  '].mean()
df['Math_Score  '].fillna(value=m_v, inplace=True)
df
```

```
C:\Users\Welcome\AppData\Local\Temp\ipykernel_10212\2348491271.py:2: SettingWithCo
pyWarning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stabl
e/user_guide/indexing.html#returning-a-view-versus-a-copy
  df['Math_Score  '].fillna(value=m_v, inplace=True)
```

Out[57]:

	Math_Score	Reading_Score	Writing_Score	Placement_Score	Club_Join_Date	Placement_Offer_
0	72.0	86	64.0	75.0	2020.0	
1	65.0	91	67.0	111.0	2020.0	
2	66.0	81	88.0	95.0	NaN	
3	59.0	77	64.0	90.0	2000.0	
4	66.0	89	76.0	81.0	2018.0	
5	61.0	92	76.0	82.0	2019.0	
6	69.1	93	72.0	66.0	2018.0	
7	85.0	85	59.0	93.0	2019.0	
8	73.0	95	68.0	NaN	NaN	
9	65.0	81	67.0	86.0	2023.0	
10	79.0	91	NaN	100.0	2018.0	

In [58]:
```
ndf.replace(to_replace = np.nan, value = -99)
```

Out[58]:

	Math_Score	Reading_Score	Writing_Score	Placement_Score	Club_Join_Date	Placement_Offer_
0	72.0	86	64.0	75.0	2020.0	
1	65.0	91	67.0	111.0	2020.0	
2	66.0	81	88.0	95.0	-99.0	
3	59.0	77	64.0	90.0	2000.0	
4	66.0	89	76.0	81.0	2018.0	
5	61.0	92	76.0	82.0	2019.0	
6	69.1	93	72.0	66.0	2018.0	
7	85.0	85	59.0	93.0	2019.0	
8	73.0	95	68.0	-99.0	-99.0	
9	65.0	81	67.0	86.0	2023.0	
10	79.0	91	-99.0	100.0	2018.0	

In [59]:
```
ndf.dropna()
```

Out[59]:

	Math_Score	Reading_Score	Writing_Score	Placement_Score	Club_Join_Date	Placement_Offer_C
0	72.0	86	64.0	75.0	2020.0	
1	65.0	91	67.0	111.0	2020.0	
3	59.0	77	64.0	90.0	2000.0	
4	66.0	89	76.0	81.0	2018.0	
5	61.0	92	76.0	82.0	2019.0	
6	69.1	93	72.0	66.0	2018.0	
7	85.0	85	59.0	93.0	2019.0	
9	65.0	81	67.0	86.0	2023.0	

In [60]:
```
#To Drop rows if all values in that row are missing
ndf.dropna(how = 'all')
```

Out[60]:

	Math_Score	Reading_Score	Writing_Score	Placement_Score	Club_Join_Date	Placement_Offer_
0	72.0	86	64.0	75.0	2020.0	
1	65.0	91	67.0	111.0	2020.0	
2	66.0	81	88.0	95.0	NaN	
3	59.0	77	64.0	90.0	2000.0	
4	66.0	89	76.0	81.0	2018.0	
5	61.0	92	76.0	82.0	2019.0	
6	69.1	93	72.0	66.0	2018.0	
7	85.0	85	59.0	93.0	2019.0	
8	73.0	95	68.0	NaN	NaN	
9	65.0	81	67.0	86.0	2023.0	
10	79.0	91	NaN	100.0	2018.0	

In [61]:
```
#To Drop columns with at least 1 null value.
ndf.dropna(axis = 1)
```

Out[61]:

	Math_Score	Reading_Score	Placement_Offer_Count
0	72.0	86	2
1	65.0	91	3
2	66.0	81	3
3	59.0	77	3
4	66.0	89	2
5	61.0	92	2
6	69.1	93	1
7	85.0	85	3
8	73.0	95	3
9	65.0	81	3
10	79.0	91	3

In [62]:
```
new_data = ndf.dropna(axis = 0, how ='any')
new_data
```

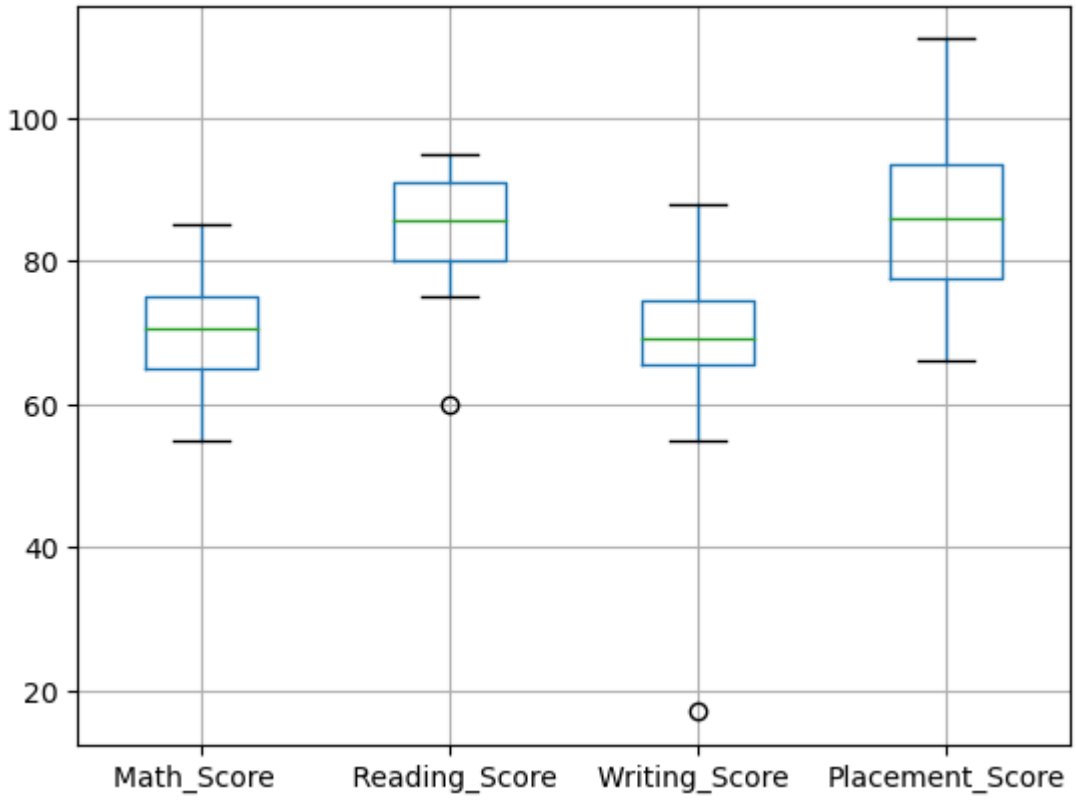
Out[62]:

	Math_Score	Reading_Score	Writing_Score	Placement_Score	Club_Join_Date	Placement_Offer_C
0	72.0	86	64.0	75.0	2020.0	
1	65.0	91	67.0	111.0	2020.0	
3	59.0	77	64.0	90.0	2000.0	
4	66.0	89	76.0	81.0	2018.0	
5	61.0	92	76.0	82.0	2019.0	
6	69.1	93	72.0	66.0	2018.0	
7	85.0	85	59.0	93.0	2019.0	
9	65.0	81	67.0	86.0	2023.0	

In [ ]:

In [108…
```
df=pd.read_csv("C:/Users/Welcome/Downloads/DSBDA02.csv")
```

In [109…
```
col = ['Math_Score  ', ' Reading_Score', 'Writing_Score ', 'Placement_Score']
df.boxplot(col)
plt.show()
```



In [110…
```
print(np.where(df['Math_Score  ']>90))
```

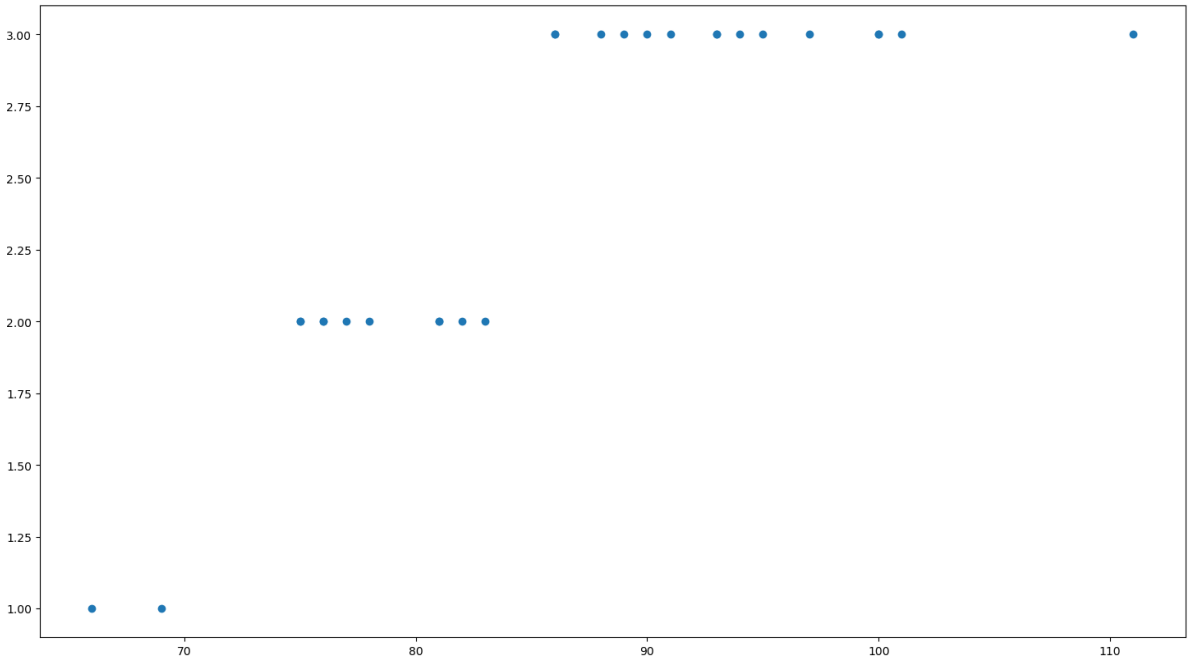
```
(array([], dtype=int64),)
```

In [111…
```
print(np.where(df[' Reading_Score']<25))
```

```
(array([], dtype=int64),)
```

In [112…
```
fig, ax = plt.subplots(figsize = (18,10))
```

In [113…
```
ax.scatter(df['Placement_Score'], df['Placement_Offer_Count'])
plt.show()
```



In [114…
```
print(np.where((df['Placement_Score']<50) & (df['Placement_Offer_Count']>1)))
```

```
(array([], dtype=int64),)
```

In [115…
```
print(np.where((df['Placement_Score']>85) & (df['Placement_Offer_Count']<3)))
```

```
(array([], dtype=int64),)
```

In [116…
```
from scipy import stats
```

In [117…
```
z = np.abs(stats.zscore(df[' Reading_Score']))
print(z)
```

```
0     0.188365
1     0.861099
2     0.484368
3     3.309850
4     0.592006
5     0.995646
6     1.130193
7     0.053819
8     1.399286
9     0.484368
10    0.861099
11    0.618915
12    0.484368
13    1.130193
14    0.080728
15    0.618915
16    1.022555
17    0.188365
18    1.399286
19    0.753462
20    0.188365
21    0.995646
22    0.349822
23    1.022555
24    1.291649
25    0.592006
26    0.457459
27    0.618915
28    0.995646
29    0.888008
Name:  Reading_Score, dtype: float64
```

In [118…
```
sorted_rscore= sorted(df[' Reading_Score'])
print(sorted_rscore)
```

```
[60, 75, 77, 77, 78, 79, 80, 80, 80, 81, 81, 81, 82, 84, 85, 86, 86, 86, 88, 89, 8
9, 91, 91, 92, 92, 92, 93, 93, 95, 95]
```

In [119…
```
q1 = np.percentile(sorted_rscore, 25)
q3 = np.percentile(sorted_rscore, 75)
print(q1,q3)
IQR = q3-q1
IQR
```

```
80.0 91.0
```

Out[119]:
```
11.0
```

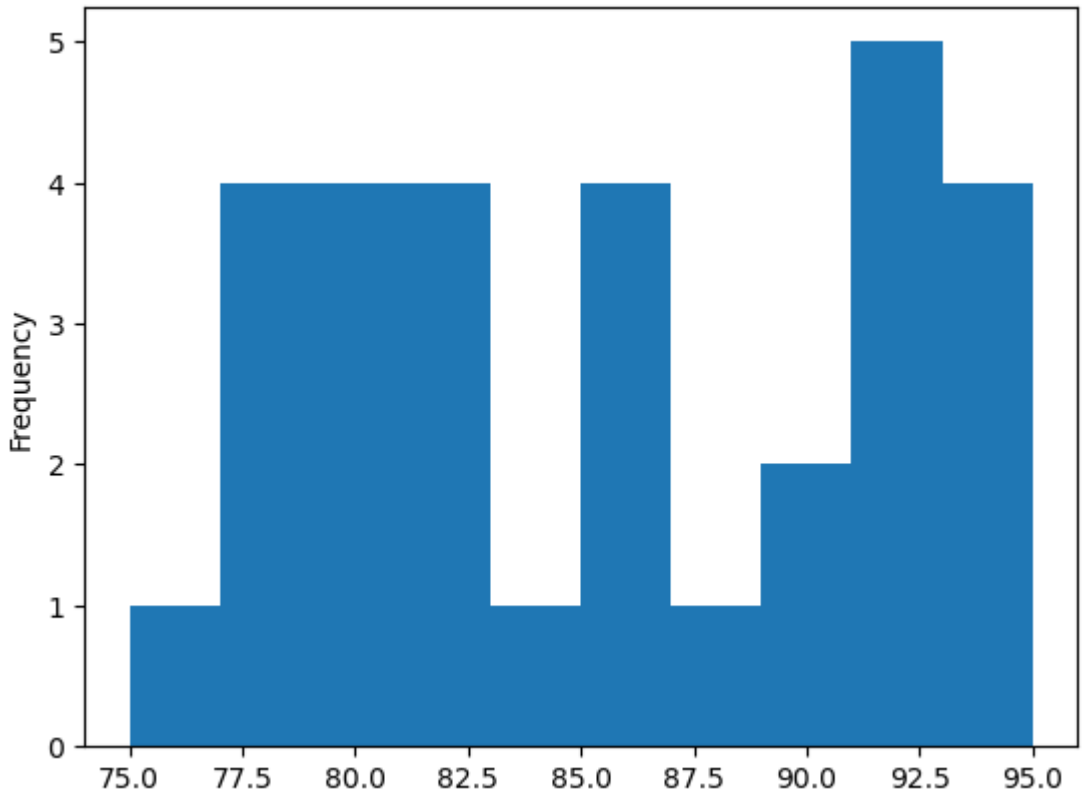
In [120…
```
lwr_bound = q1-(1*IQR)
upr_bound = q3+(1*IQR)
print(lwr_bound, upr_bound)
```

```
69.0 102.0
```

In [121…
```
r_outliers = []
for i in sorted_rscore:
    if (i<lwr_bound or i>upr_bound):
        r_outliers.append(i)
print(r_outliers)
```

```
[60]
```

In [26]:
```
df[' Reading_Score'].plot(kind = 'hist')
plt.show()
```



In [ ]:
```
df['log_math'] = np.log10(df['math score'])
df['log_math'].plot(kind = 'hist')
```