Data Wrangling II

May 1, 2022

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

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

1 Detecting Handling of Null values

[3]:
```
df
```

[3]:
```
   gender math score reading score  writing score  Placement Score  \
0  female         72            72           74.0             78.0
1  female         69            90           88.0              NaN
2  female         90            95           93.0             74.0
3    male         47            57            NaN             78.0
4    male         na            78           75.0             81.0
5  female         71            Na           78.0             70.0
6    male         12            44           52.0             12.0
7    male        NaN            65           67.0             49.0
8    male          5            77           89.0             55.0

   placement offer count  Region
0                      1    Pune
1                      2      na
2                      2  Nashik
3                      1      Na
4                      3    Pune
5                      4      na
6                      2  Nashik
7                      1    Pune
8                      0     NaN
```

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

[4]:
```
   gender  math score  reading score  writing score  Placement Score  \
0   False       False          False          False            False
```

```
1   False       False          False          False            True
2   False       False          False          False           False
3   False       False          False           True           False
4   False       False          False          False           False
5   False       False          False          False           False
6   False       False          False          False           False
7   False        True          False          False           False
8   False       False          False          False           False

   placement offer count  Region
0                  False   False
1                  False   False
2                  False   False
3                  False   False
4                  False   False
5                  False   False
6                  False   False
7                  False   False
8                  False    True
```

```
[5]: series = pd.isnull(df['math score'])
```

```
[6]: df[series]
```

```
[6]:   gender math score reading score  writing score  Placement Score  \
     7   male        NaN             65           67.0             49.0

        placement offer count Region
     7                      1   Pune
```

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

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

        placement offer count  Region
     0                   True    True
     1                   True    True
     2                   True    True
```

```
3                   True    True
4                   True    True
5                   True    True
6                   True    True
7                   True    True
8                   True   False
```

```
[8]: series = pd.notnull(df['math score'])
```

```
[9]: df[series]
```

```
[9]:
   gender math score reading score  writing score  Placement Score  \
0  female         72            72           74.0             78.0
1  female         69            90           88.0              NaN
2  female         90            95           93.0             74.0
3    male         47            57            NaN             78.0
4    male         na            78           75.0             81.0
5  female         71            Na           78.0             70.0
6    male         12            44           52.0             12.0
8    male          5            77           89.0             55.0

   placement offer count  Region
0                      1    Pune
1                      2      na
2                      2  Nashik
3                      1      Na
4                      3    Pune
5                      4      na
6                      2  Nashik
8                      0     NaN
```

```
[10]: from sklearn.preprocessing import LabelEncoder
      le = LabelEncoder()
      df['gender'] = le.fit_transform(df['gender'])
```

```
[11]: df
```

```
[11]:
   gender math score reading score  writing score  Placement Score  \
0       0         72            72           74.0             78.0
1       0         69            90           88.0              NaN
2       0         90            95           93.0             74.0
3       1         47            57            NaN             78.0
4       1         na            78           75.0             81.0
5       0         71            Na           78.0             70.0
6       1         12            44           52.0             12.0
7       1        NaN            65           67.0             49.0
8       1          5            77           89.0             55.0
```

```
   placement offer count  Region
0                      1    Pune
1                      2      na
2                      2  Nashik
3                      1      Na
4                      3    Pune
5                      4      na
6                      2  Nashik
7                      1    Pune
8                      0     NaN
```

2 filling missing values

[12]:
```
missing_values = ['Na','na']
df = pd.read_csv("StudentsPerformance.csv", na_values=missing_values)
df
```

[12]:
```
   gender  math score  reading score  writing score  Placement Score  \
0  female        72.0           72.0           74.0             78.0
1  female        69.0           90.0           88.0              NaN
2  female        90.0           95.0           93.0             74.0
3    male        47.0           57.0            NaN             78.0
4    male         NaN           78.0           75.0             81.0
5  female        71.0            NaN           78.0             70.0
6    male        12.0           44.0           52.0             12.0
7    male         NaN           65.0           67.0             49.0
8    male         5.0           77.0           89.0             55.0

   placement offer count  Region
0                      1    Pune
1                      2     NaN
2                      2  Nashik
3                      1     NaN
4                      3    Pune
5                      4     NaN
6                      2  Nashik
7                      1    Pune
8                      0     NaN
```

[13]:
```
df.fillna(0)
```

[13]:
```
   gender  math score  reading score  writing score  Placement Score  \
0  female        72.0           72.0           74.0             78.0
1  female        69.0           90.0           88.0              0.0
2  female        90.0           95.0           93.0             74.0
```

```
3    male        47.0           57.0            0.0             78.0
4    male         0.0           78.0           75.0             81.0
5  female        71.0            0.0           78.0             70.0
6    male        12.0           44.0           52.0             12.0
7    male         0.0           65.0           67.0             49.0
8    male         5.0           77.0           89.0             55.0

   placement offer count  Region
0                      1    Pune
1                      2       0
2                      2  Nashik
3                      1       0
4                      3    Pune
5                      4       0
6                      2  Nashik
7                      1    Pune
8                      0       0
```

[14]:

```
df['math score'] = df['math score'].fillna(df['math score'].mean()) #␣
 ↪mean(),median(),std(),min(),max()
df
```

[14]:

```
   gender  math score  reading score  writing score  Placement Score  \
0  female   72.000000           72.0           74.0             78.0
1  female   69.000000           90.0           88.0              NaN
2  female   90.000000           95.0           93.0             74.0
3    male   47.000000           57.0            NaN             78.0
4    male   52.285714           78.0           75.0             81.0
5  female   71.000000            NaN           78.0             70.0
6    male   12.000000           44.0           52.0             12.0
7    male   52.285714           65.0           67.0             49.0
8    male    5.000000           77.0           89.0             55.0

   placement offer count  Region
0                      1    Pune
1                      2     NaN
2                      2  Nashik
3                      1     NaN
4                      3    Pune
5                      4     NaN
6                      2  Nashik
7                      1    Pune
8                      0     NaN
```

[15]:

```
m_v = df['Placement Score'].median()
df['Placement Score'].fillna(value = m_v, inplace = True)
df
```

[15]:

```
   gender  math score  reading score  writing score  Placement Score  \
0  female   72.000000           72.0           74.0             78.0
1  female   69.000000           90.0           88.0             72.0
2  female   90.000000           95.0           93.0             74.0
3    male   47.000000           57.0            NaN             78.0
4    male   52.285714           78.0           75.0             81.0
5  female   71.000000            NaN           78.0             70.0
6    male   12.000000           44.0           52.0             12.0
7    male   52.285714           65.0           67.0             49.0
8    male    5.000000           77.0           89.0             55.0

   placement offer count  Region
0                      1    Pune
1                      2     NaN
2                      2  Nashik
3                      1     NaN
4                      3    Pune
5                      4     NaN
6                      2  Nashik
7                      1    Pune
8                      0     NaN
```

[16]:

```python
df.replace(to_replace = np.nan, value = -99)
```

[16]:

```
   gender  math score  reading score  writing score  Placement Score  \
0  female   72.000000           72.0           74.0             78.0
1  female   69.000000           90.0           88.0             72.0
2  female   90.000000           95.0           93.0             74.0
3    male   47.000000           57.0          -99.0             78.0
4    male   52.285714           78.0           75.0             81.0
5  female   71.000000          -99.0           78.0             70.0
6    male   12.000000           44.0           52.0             12.0
7    male   52.285714           65.0           67.0             49.0
8    male    5.000000           77.0           89.0             55.0

   placement offer count  Region
0                      1    Pune
1                      2     -99
2                      2  Nashik
3                      1     -99
4                      3    Pune
5                      4     -99
6                      2  Nashik
7                      1    Pune
8                      0     -99
```

3 Detecting null values

```
[17]: df.dropna() # dropping row if al least 1 null value missing
```

```
[17]:    gender  math score  reading score  writing score  Placement Score  \
      0  female   72.000000           72.0           74.0             78.0
      2  female   90.000000           95.0           93.0             74.0
      4    male   52.285714           78.0           75.0             81.0
      6    male   12.000000           44.0           52.0             12.0
      7    male   52.285714           65.0           67.0             49.0

         placement offer count  Region
      0                      1    Pune
      2                      2  Nashik
      4                      3    Pune
      6                      2  Nashik
      7                      1    Pune
```

```
[18]: df.dropna(how = 'all') # dropping row if all values in that row missing
```

```
[18]:    gender  math score  reading score  writing score  Placement Score  \
      0  female   72.000000           72.0           74.0             78.0
      1  female   69.000000           90.0           88.0             72.0
      2  female   90.000000           95.0           93.0             74.0
      3    male   47.000000           57.0            NaN             78.0
      4    male   52.285714           78.0           75.0             81.0
      5  female   71.000000            NaN           78.0             70.0
      6    male   12.000000           44.0           52.0             12.0
      7    male   52.285714           65.0           67.0             49.0
      8    male    5.000000           77.0           89.0             55.0

         placement offer count  Region
      0                      1    Pune
      1                      2     NaN
      2                      2  Nashik
      3                      1     NaN
      4                      3    Pune
      5                      4     NaN
      6                      2  Nashik
      7                      1    Pune
      8                      0     NaN
```

```
[19]: df.dropna(axis=1) # dropping column if al least 1 null value missing
```

```
[19]:    gender  math score  Placement Score  placement offer count
      0  female   72.000000             78.0                      1
      1  female   69.000000             72.0                      2
```

```
2  female  90.000000            74.0                       2
3    male  47.000000            78.0                       1
4    male  52.285714            81.0                       3
5  female  71.000000            70.0                       4
6    male  12.000000            12.0                       2
7    male  52.285714            49.0                       1
8    male   5.000000            55.0                       0
```

4 Detecting of Outliners

[20]:

```
df
```

[20]:

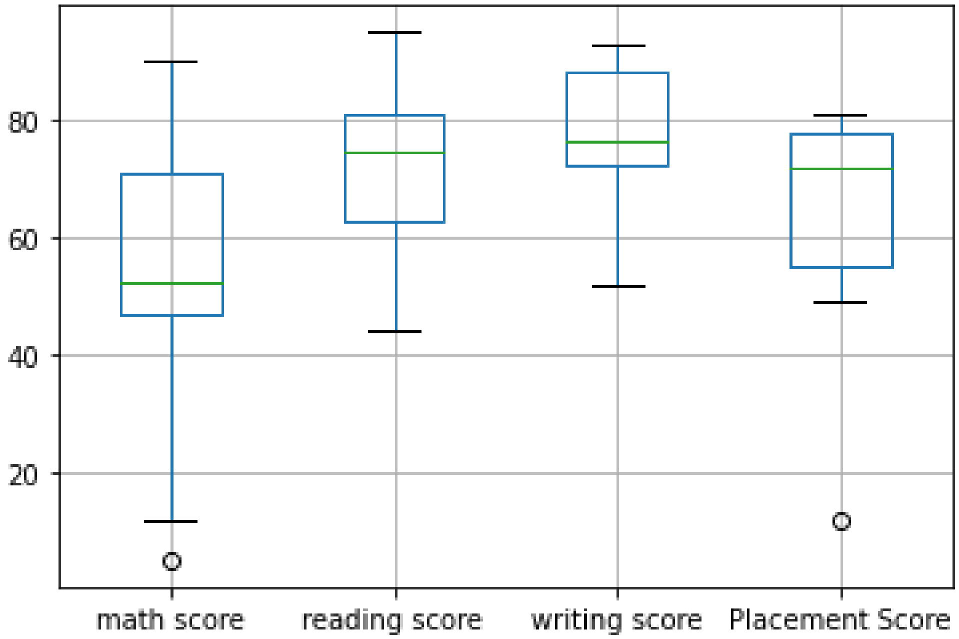
```
   gender  math score  reading score  writing score  Placement Score  \
0  female  72.000000           72.0           74.0             78.0
1  female  69.000000           90.0           88.0             72.0
2  female  90.000000           95.0           93.0             74.0
3    male  47.000000           57.0            NaN             78.0
4    male  52.285714           78.0           75.0             81.0
5  female  71.000000            NaN           78.0             70.0
6    male  12.000000           44.0           52.0             12.0
7    male  52.285714           65.0           67.0             49.0
8    male   5.000000           77.0           89.0             55.0

   placement offer count  Region
0                      1    Pune
1                      2     NaN
2                      2  Nashik
3                      1     NaN
4                      3    Pune
5                      4     NaN
6                      2  Nashik
7                      1    Pune
8                      0     NaN
```

[21]:

```
col = ['math score','reading score','writing score','Placement Score']
df.boxplot(col)
```

[21]:

```
<AxesSubplot:>
```



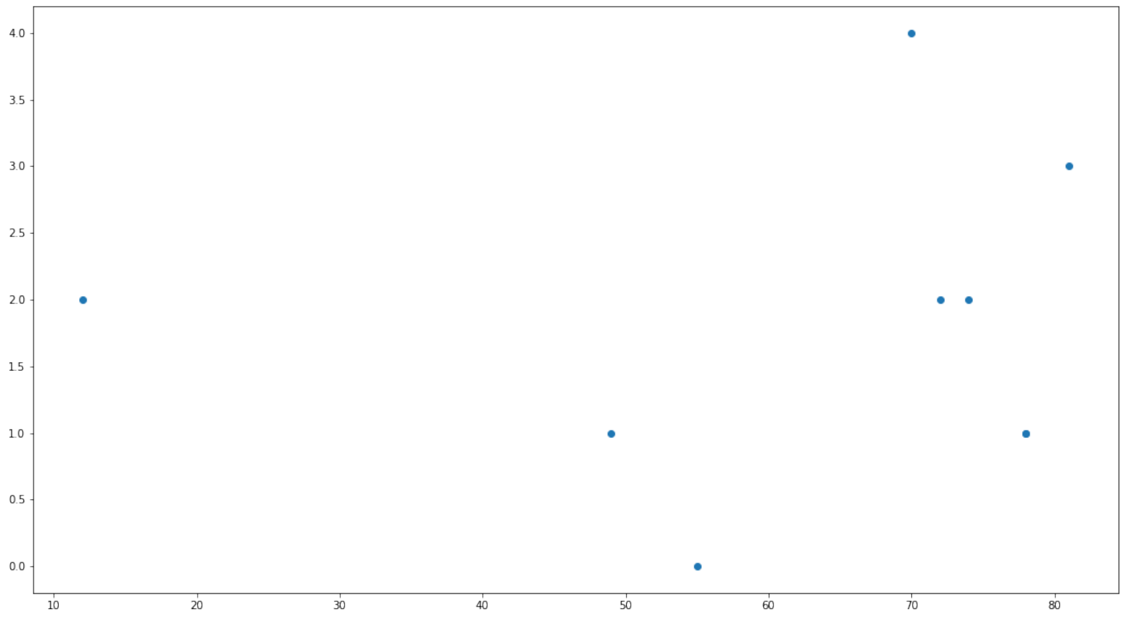
[22]:

```
print(np.where(df['math score']<25))
print(np.where(df['reading score']<25))
print(np.where(df['writing score']<30))
```

```
(array([6, 8], dtype=int64),)
(array([], dtype=int64),)
(array([], dtype=int64),)
```

[23]:

```
fig, ax = plt.subplots(figsize = (18,10))
ax.scatter(df['Placement Score'],df['placement offer count'])
plt.show()
```



```
[24]: 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([6], dtype=int64),)
      (array([0, 1, 2, 3, 6, 7, 8], dtype=int64),)
```

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

```
[26]: z = np.abs(stats.zscore(df['math score']))
```

```
[27]: print(z)

      [0.74351319 0.63036988 1.42237305 0.19934774 0.         0.70579875
       1.51935303 0.         1.78335409]
```

```
[28]: threshold = 0.18
```

```
[29]: sample_outliner = np.where(z<threshold)
      sample_outliner
```

```
[29]: (array([4, 7], dtype=int64),)
```

```
[30]: df
```

```
[30]:    gender  math score  reading score  writing score  Placement Score  \
      0  female   72.000000           72.0           74.0             78.0
      1  female   69.000000           90.0           88.0             72.0
```

```
2  female   90.000000          95.0           93.0              74.0
3    male   47.000000          57.0            NaN              78.0
4    male   52.285714          78.0           75.0              81.0
5  female   71.000000           NaN           78.0              70.0
6    male   12.000000          44.0           52.0              12.0
7    male   52.285714          65.0           67.0              49.0
8    male    5.000000          77.0           89.0              55.0

   placement offer count  Region
0                      1    Pune
1                      2     NaN
2                      2  Nashik
3                      1     NaN
4                      3    Pune
5                      4     NaN
6                      2  Nashik
7                      1    Pune
8                      0     NaN
```

[31]:

```
sorted_rscore = sorted(df['math score'])
sorted_rscore
```

[31]:

```
[5.0,
 12.0,
 47.0,
 52.285714285714285,
 52.285714285714285,
 69.0,
 71.0,
 72.0,
 90.0]
```

[32]:

```
q1 = np.percentile(sorted_rscore, 25)
q3 = np.percentile(sorted_rscore, 75)
print(q1,q3)
```

```
47.0 71.0
```

[33]:

```
IQR = q3-q1
```

[34]:

```
lower_bound = q1-(1.5*IQR)
upper_bound = q3+(1.5*IQR)
print(lower_bound,upper_bound)
```

```
11.0 107.0
```

[35]:

```
r_outliners = []
for i in sorted_rscore:
    if (i<lower_bound or i>upper_bound):
        r_outliners.append(i)
print(r_outliners)
```

```
[5.0]
```

5 Handling of Outliners

[36]:

```
new_df = df
ninetieth_percentile = np.percentile(new_df['reading score'],11)
b = np.where(new_df['math␣
 ↪score']<ninetieth_percentile,ninetieth_percentile,new_df['math score'])
print(b)
```

```
[72.         69.         90.         47.         52.28571429 71.
 12.         52.28571429  5.        ]
```

[37]:

```
new_df.insert(1,'m score',b,True)
```

[38]:

```
new_df
```

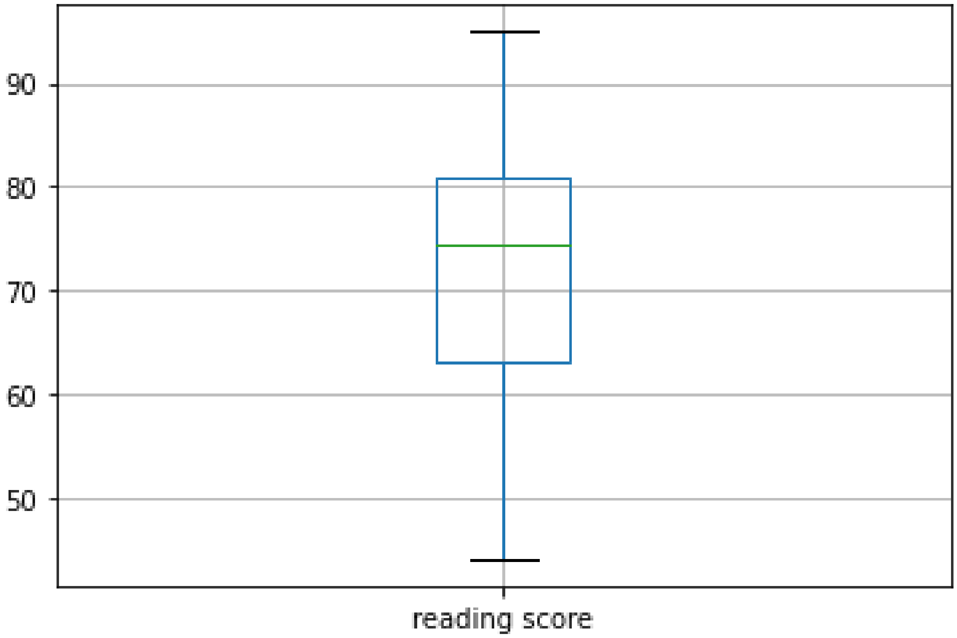
[38]:

```
   gender    m score  math score  reading score  writing score  \
0  female  72.000000   72.000000           72.0           74.0
1  female  69.000000   69.000000           90.0           88.0
2  female  90.000000   90.000000           95.0           93.0
3    male  47.000000   47.000000           57.0            NaN
4    male  52.285714   52.285714           78.0           75.0
5  female  71.000000   71.000000            NaN           78.0
6    male  12.000000   12.000000           44.0           52.0
7    male  52.285714   52.285714           65.0           67.0
8    male   5.000000    5.000000           77.0           89.0

   Placement Score  placement offer count  Region
0             78.0                      1    Pune
1             72.0                      2     NaN
2             74.0                      2  Nashik
3             78.0                      1     NaN
4             81.0                      3    Pune
5             70.0                      4     NaN
6             12.0                      2  Nashik
7             49.0                      1    Pune
8             55.0                      0     NaN
```

[39]:

```
col = ['reading score']
df.boxplot(col)
```

```
[39]: <AxesSubplot:>
```



```
[40]: median = np.median(sorted_rscore)
      median
```

```
[40]: 52.285714285714285
```

```
[41]: refined_df = df
      refined_df['reading score'] = np.where(refined_df['reading␣
       ↪score']>upper_bound,median,refined_df['reading score'])
```

```
[42]: refined_df['reading score'] = np.where(refined_df['reading␣
       ↪score']<lower_bound,median,refined_df['reading score'])
```

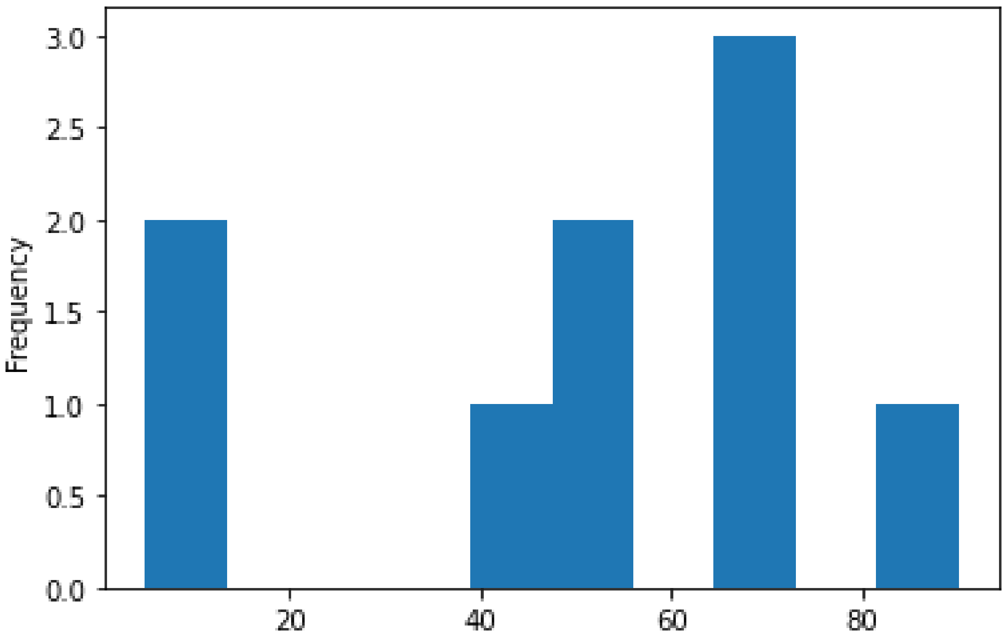
```
[43]: refined_df
```

```
[43]:    gender    m score  math score  reading score  writing score  \
      0  female  72.000000   72.000000           72.0           74.0
      1  female  69.000000   69.000000           90.0           88.0
      2  female  90.000000   90.000000           95.0           93.0
      3    male  47.000000   47.000000           57.0            NaN
      4    male  52.285714   52.285714           78.0           75.0
      5  female  71.000000   71.000000            NaN           78.0
      6    male  12.000000   12.000000           44.0           52.0
      7    male  52.285714   52.285714           65.0           67.0
      8    male   5.000000    5.000000           77.0           89.0
```

```
   Placement Score  placement offer count  Region
0             78.0                      1    Pune
1             72.0                      2     NaN
2             74.0                      2  Nashik
3             78.0                      1     NaN
4             81.0                      3    Pune
5             70.0                      4     NaN
6             12.0                      2  Nashik
7             49.0                      1    Pune
8             55.0                      0     NaN
```

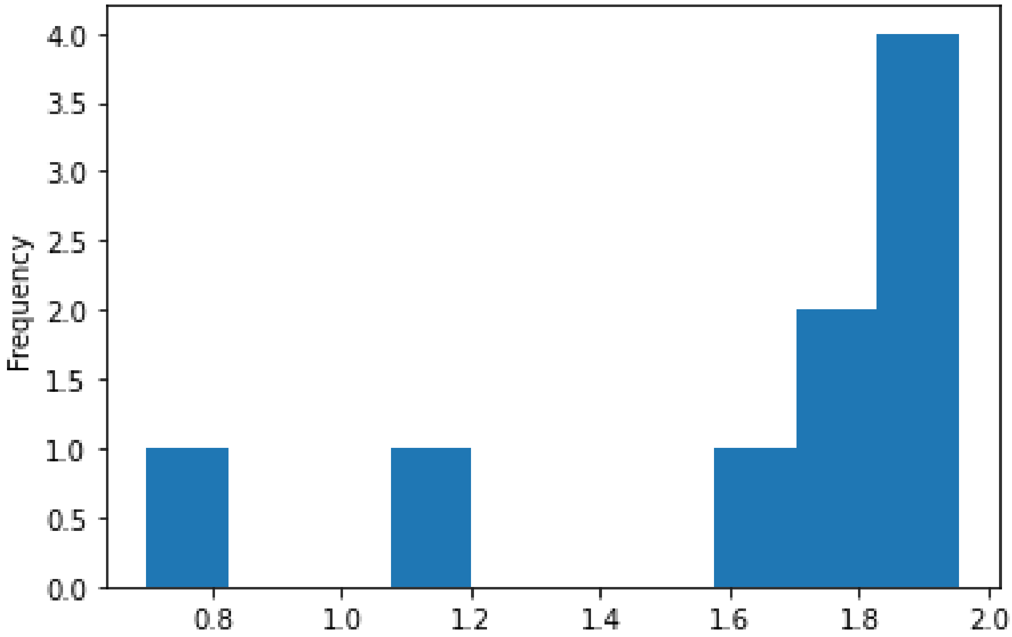
6 Data Transformation

[44]:
```
new_df['m score'].plot(kind='hist')
df['log_math'] = np.log10(df['math score'])
```



[45]:
```
df['log_math'].plot(kind='hist')
```

[45]:
```
<AxesSubplot:ylabel='Frequency'>
```


Frequency
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
0.8
1.0
1.2
1.4
1.6
1.8
2.0