Practical 2

Title: Data Wrangling II

[ ]:
```
#import required libraries
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
%matplotlib inline
from scipy import stats
```

[ ]:
```
#load dataset
df=pd.read_csv("D:\StudentPerformance.csv")
```

[ ]:
```
#display dataframe
df
```

[ ]:

	gender	math score	reading score	writing score	placement score	club join year	placement offer
0	female	63.0	84.0	64	84.0	2020	2
1	female	71.0	80.0	76	86.0	2018	3
2	female	64.0	81.0	66	81.0	2020	2
3	male	71.0	85.0	77	96.0	2018	1
4	male	68.0	86.0	76	NaN	2021	3
5	female	94.0	86.0	61	100.0	2019	1
6	male	75.0	79.0	66	-99.0	2020	1
7	female	NaN	NaN	66	95.0	2019	3
8	male	66.0	88.0	66	88.0	2020	3
9	male	70.0	79.0	61			
10	female	-99.0	80.0	65			
11	male	76.0	84.0	-99			
12	female	74.0	79.0	79			

```
9              87.0            2021               2
10             85.0            2021               1
11              NaN            2020               2
12             98.0            2019               2
```

Identification and Handling of Missing Values

Checking for missing values

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

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

    placement score  club join year  placement offer
0             False           False            False
1             False           False            False
2             False           False            False
3             False           False            False
4              True           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             True           False            False
12            False           False            False
```

[ ]:
```
#check null value in specific column
series = pd.isnull(df["reading score"])
df[series]
```

[ ]:
```
    gender  math score                   reading score  writing score  \
7  female                          NaN            NaN             66
```

```
   placement score  club join year  placement offer
7             95.0            2019                3
```

[ ]:
```
#checking for missing values using notnull()
df.notnull()
```

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

    placement score  club join year  placement offer
0              True            True             True
1              True            True             True
2              True            True             True
3              True            True             True
4             False            True             True
5              True            True             True
6              True            True             True
7              True            True             True
8              True            True             True
9              True            True             True
10             True            True             True
11            False            True             True
12             True            True             True
```

[ ]:
```
seseries = pd.notnull(df["reading score"])
df[series]
```

[ ]:
```
   gender  math score                    reading score  writing score  \
7  female                           NaN            NaN             66

   placement score  club join year  placement offer
7             95.0            2019                3
```

[ ]:
```
#categorical variable to quantitative variable
from sklearn.preprocessing import LabelEncoder
```

```
le = LabelEncoder ()
df['gender'] = le.fit_transform(df['gender'])
newdf=df
df
```

[ ]:

	gender	math score	reading score	writing score	\
0	0	63.0	84.0	64	
1	0	71.0	80.0	76	
2	0	64.0	81.0	66	
3	1	71.0	85.0	77	
4	1	68.0	86.0	76	
5	0	94.0	86.0	61	
6	1	75.0	79.0	66	
7	0	NaN	NaN	66	
8	1	66.0	88.0	66	
9	1	70.0	79.0	61	
10	0	-99.0	80.0	65	
11	1	76.0	84.0	-99	
12	0	74.0	79.0	79	

	placement score	club join year	placement offer
0	84.0	2020	2
1	86.0	2018	3
2	81.0	2020	2
3	96.0	2018	1
4	NaN	2021	3
5	100.0	2019	1
6	-99.0	2020	1
7	95.0	2019	3
8	88.0	2020	3
9	87.0	2021	2
10	85.0	2021	1
11	NaN	2020	2
12	98.0	2019	2

[ ]:
```
df['gender'].replace({1: "male", 0: "female"}, inplace=True)
df
```

[ ]:

	gender	math score	reading score	writing score	\
0	female	63.0	84.0	64	
1	female	71.0	80.0	76	
2	female	64.0	81.0	66	
3	male	71.0	85.0	77	
4	male	68.0	86.0	76	
5	female	94.0	86.0	61	
6	male	75.0	79.0	66	
7	female	NaN	NaN	66	

```
8     male                          66.0           88.0             66
9     male                          70.0           79.0             61
10  female                         -99.0           80.0             65
11    male                          76.0           84.0            -99
12  female                          74.0           79.0             79

    placement score  club join year  placement offer
0              84.0            2020                2
1              86.0            2018                3
2              81.0            2020                2
3              96.0            2018                1
4               NaN            2021                3
5             100.0            2019                1
6             -99.0            2020                1
7              95.0            2019                3
8              88.0            2020                3
9              87.0            2021                2
10             85.0            2021                1
11              NaN            2020                2
12             98.0            2019                2
```

Filling Misssing Values

[ ]:
```
#filling missing values
ndf=df
ndf.fillna(0)
```

[ ]:
```
    gender  math score                      reading score  writing score  \
0   female                          63.0           84.0             64
1   female                          71.0           80.0             76
2   female                          64.0           81.0             66
3     male                          71.0           85.0             77
4     male                          68.0           86.0             76
5   female                          94.0           86.0             61
6     male                          75.0           79.0             66
7   female                           0.0            0.0             66
8     male                          66.0           88.0             66
9     male                          70.0           79.0             61
10  female                         -99.0           80.0             65
11    male                          76.0           84.0            -99
12  female                          74.0           79.0             79

    placement score  club join year  placement offer
0              84.0            2020                2
1              86.0            2018                3
2              81.0            2020                2
3              96.0            2018                1
4               0.0            2021                3
```

```
5              100.0            2019                1
6              -99.0            2020                1
7               95.0            2019                3
8               88.0            2020                3
9               87.0            2021                2
10              85.0            2021                1
11               0.0            2020                2
12              98.0            2019                2
```

[ ]:
```
#filling missing values with mean
m_v=df['reading score'].mean()
df['reading score'].fillna(value=m_v, inplace=True)
df
```

[ ]:
```
     gender  math score                    reading score  writing score  \
0   female                          63.0      84.000000             64
1   female                          71.0      80.000000             76
2   female                          64.0      81.000000             66
3     male                          71.0      85.000000             77
4     male                          68.0      86.000000             76
5   female                          94.0      86.000000             61
6     male                          75.0      79.000000             66
7   female                           NaN      82.583333             66
8     male                          66.0      88.000000             66
9     male                          70.0      79.000000             61
10  female                         -99.0      80.000000             65
11    male                          76.0      84.000000            -99
12  female                          74.0      79.000000             79

    placement score  club join year  placement offer
0              84.0            2020                2
1              86.0            2018                3
2              81.0            2020                2
3              96.0            2018                1
4               NaN            2021                3
5             100.0            2019                1
6             -99.0            2020                1
7              95.0            2019                3
8              88.0            2020                3
9              87.0            2021                2
10             85.0            2021                1
11              NaN            2020                2
12             98.0            2019                2
```

Replacing null values

[ ]:
```
#replacing null values
ndf.replace(to_replace = np.nan, value = -99)
```

[ ]:

	gender	math score		reading score	writing score	\
0	female		63.0	84.000000	64	
1	female		71.0	80.000000	76	
2	female		64.0	81.000000	66	
3	male		71.0	85.000000	77	
4	male		68.0	86.000000	76	
5	female		94.0	86.000000	61	
6	male		75.0	79.000000	66	
7	female		-99.0	82.583333	66	
8	male		66.0	88.000000	66	
9	male		70.0	79.000000	61	
10	female		-99.0	80.000000	65	
11	male		76.0	84.000000	-99	
12	female		74.0	79.000000	79	

	placement score	club join year	placement offer
0	84.0	2020	2
1	86.0	2018	3
2	81.0	2020	2
3	96.0	2018	1
4	-99.0	2021	3
5	100.0	2019	1
6	-99.0	2020	1
7	95.0	2019	3
8	88.0	2020	3
9	87.0	2021	2
10	85.0	2021	1
11	-99.0	2020	2
12	98.0	2019	2

Deleting null values

[ ]:
```
#dropping rows with null values
ndf.dropna()
```

[ ]:

	gender	math score		reading score	writing score	\
0	female		63.0	84.0	64	
1	female		71.0	80.0	76	
2	female		64.0	81.0	66	
3	male		71.0	85.0	77	
5	female		94.0	86.0	61	
6	male		75.0	79.0	66	
8	male		66.0	88.0	66	
9	male		70.0	79.0	61	
10	female		-99.0	80.0	65	
12	female		74.0	79.0	79	

	placement score	club join year	placement offer

```
0              84.0            2020                2
1              86.0            2018                3
2              81.0            2020                2
3              96.0            2018                1
5             100.0            2019                1
6             -99.0            2020                1
8              88.0            2020                3
9              87.0            2021                2
10             85.0            2021                1
12             98.0            2019                2
```

[ ]:
```
#drop rows if all values in that row is missing
ndf.dropna(how = 'all')
```

[ ]:
```
    gender  math score                    reading score  writing score  \
0   female                         63.0      84.000000             64
1   female                         71.0      80.000000             76
2   female                         64.0      81.000000             66
3     male                         71.0      85.000000             77
4     male                         68.0      86.000000             76
5   female                         94.0      86.000000             61
6     male                         75.0      79.000000             66
7   female                          NaN      82.583333             66
8     male                         66.0      88.000000             66
9     male                         70.0      79.000000             61
10  female                        -99.0      80.000000             65
11    male                         76.0      84.000000            -99
12  female                         74.0      79.000000             79

    placement score  club join year  placement offer
0              84.0            2020                2
1              86.0            2018                3
2              81.0            2020                2
3              96.0            2018                1
4               NaN            2021                3
5             100.0            2019                1
6             -99.0            2020                1
7              95.0            2019                3
8              88.0            2020                3
9              87.0            2021                2
10             85.0            2021                1
11              NaN            2020                2
12             98.0            2019                2
```

[ ]:
```
#drop column with at least 1 null value
ndf.dropna(axis = 1)
```

[ ]:

	gender	reading score	writing score	club join year	placement offer
0	female	84.000000	64	2020	2
1	female	80.000000	76	2018	3
2	female	81.000000	66	2020	2
3	male	85.000000	77	2018	1
4	male	86.000000	76	2021	3
5	female	86.000000	61	2019	1
6	male	79.000000	66	2020	1
7	female	82.583333	66	2019	3
8	male	88.000000	66	2020	3
9	male	79.000000	61	2021	2
10	female	80.000000	65	2021	1
11	male	84.000000	-99	2020	2
12	female	79.000000	79	2019	2

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

[ ]:

	gender	math score	reading score	writing score	\
0	female	63.0	84.0	64	
1	female	71.0	80.0	76	
2	female	64.0	81.0	66	
3	male	71.0	85.0	77	
5	female	94.0	86.0	61	
6	male	75.0	79.0	66	
8	male	66.0	88.0	66	
9	male	70.0	79.0	61	
10	female	-99.0	80.0	65	
12	female	74.0	79.0	79	

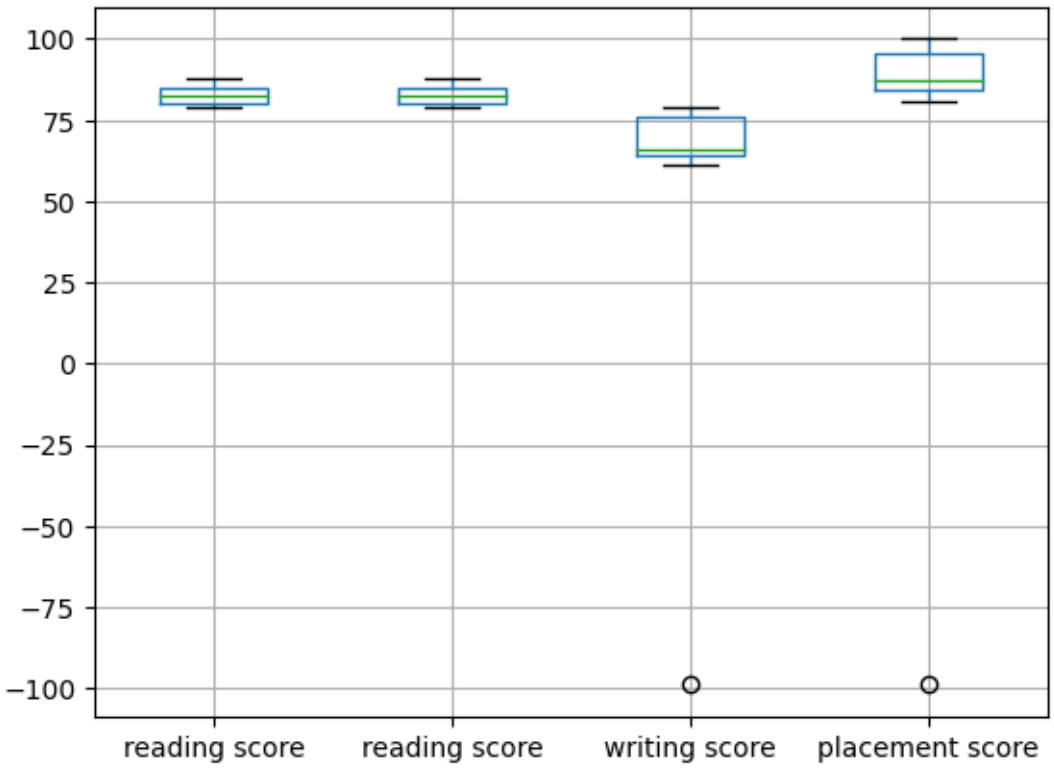
	placement score	club join year	placement offer
0	84.0	2020	2
1	86.0	2018	3
2	81.0	2020	2
3	96.0	2018	1
5	100.0	2019	1
6	-99.0	2020	1
8	88.0	2020	3
9	87.0	2021	2
10	85.0	2021	1
12	98.0	2019	2

Detecting Otliers

Detecting outliers using Boxplot

[ ]:
```
col =['reading score', 'reading score', 'writing score', 'placement score']
df.boxplot(col)
```

[ ]: <AxesSubplot:>

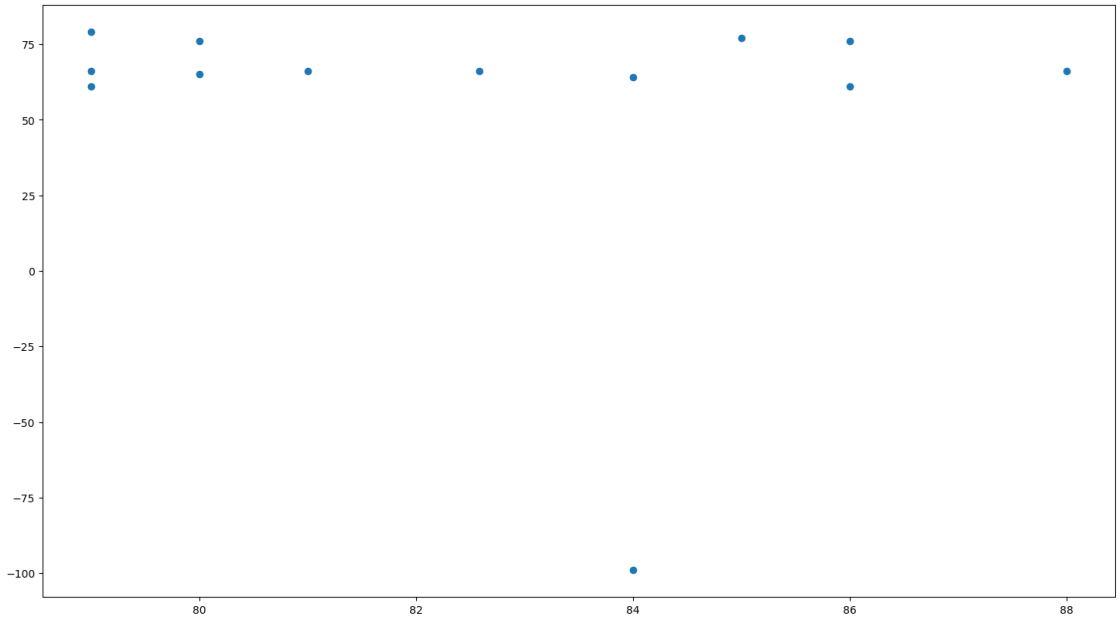


[ ]:
```
print(np.where(df['reading score']>90))
print(np.where(df['writing score']>90))
```

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

Detecting outliers using scatterplot

[ ]:
```
fig, ax =plt.subplots(figsize = (18,10))
ax.scatter(df['reading score'], df['writing score'])
plt.show()
ax.set_xlabel('(Proportion non-retail business acres)/(town)')
ax.set_ylabel('(Full-value property-tax rate)/($10,000)')
```



[ ]:
```
Text(4.444444444444452, 0.5, '(Full-value property-tax rate)/($10,000)')
```

[ ]:
```
print(np.where((df['reading score']<50) & (df['writing score']>1)))
print(np.where((df['reading score']>85) & (df['writing score']<3)))
```

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

Detecting outliers using Z-score

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

[ ]:
```
print(z)
```

```
0     0.472390
1     0.861418
2     0.527966
3     0.805843
4     1.139295
5     1.139295
6     1.194870
7     0.000000
8     1.806199
9     1.194870
10    0.861418
11    0.472390
12    1.194870
Name: reading score, dtype: float64
```

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

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

[ ]:
```
sample_outliers
```

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

Detecting outliers using IQR

[ ]:
```
sorted_rscore= sorted(df['reading score'])
```

[ ]:
```
sorted_rscore
```

[ ]:
```
[79.0,
 79.0,
 79.0,
 80.0,
 80.0,
 81.0,
 82.58333333333333,
 84.0,
 84.0,
 85.0,
 86.0,
 86.0,
 88.0]
```

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

```
80.0 85.0
```

[ ]:
```
IQR = q3-q1
```

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

```
72.5 92.5
```

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

[ ]:

	gender	math score	reading score	writing score	\
0	female	63.0	84.0	64	
1	female	71.0	80.0	76	

```
2   female                          64.0           81.0            66
3     male                          71.0           85.0            77
4     male                          68.0           86.0            76
5   female                          94.0           86.0            61
6     male                          75.0           79.0            66
8     male                          66.0           88.0            66
9     male                          70.0           79.0            61
10  female                         -99.0           80.0            65
11    male                          76.0           84.0           -99
12  female                          74.0           79.0            79
```

```
    placement score  club join year  placement offer
0              84.0            2020                2
1              86.0            2018                3
2              81.0            2020                2
3              96.0            2018                1
4               NaN            2021                3
5             100.0            2019                1
6             -99.0            2020                1
8              88.0            2020                3
9              87.0            2021                2
10             85.0            2021                1
11              NaN            2020                2
12             98.0            2019                2
```

Handling of Outliers

Quantile based flooring and capping

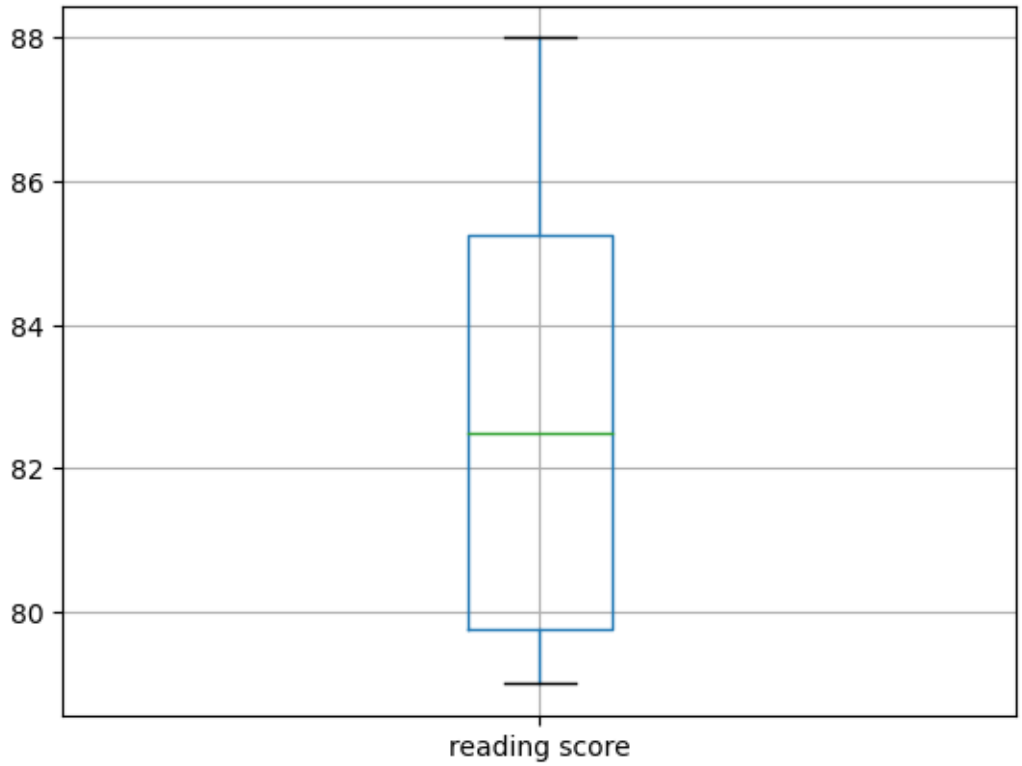
[ ]:
```
df=pd.read_csv("D:\StudentPerformance.csv")
```

[ ]:
```
df_stud=df
ninetieth_percentile = np.percentile(df_stud['reading score'], 90)
b = np.where(df_stud['reading score']>ninetieth_percentile,
ninetieth_percentile, df_stud['reading score'])
print("New array:",b)
```

```
New array: [84. 80. 81. 85. 86. 86. 79. nan 88. 79. 80. 84. 79.]
```

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

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



Median Imputation

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

[ ]:
```
82.58333333333333
```

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

[ ]:
```
df
```

[ ]:

	gender	math score	reading score	writing score	\
0	female	63.0	84.0	64	
1	female	71.0	80.0	76	
2	female	64.0	81.0	66	
3	male	71.0	85.0	77	
4	male	68.0	86.0	76	
5	female	94.0	86.0	61	
6	male	75.0	79.0	66	
7	female	NaN	NaN	66	
8	male	66.0	88.0	66	

```
9     male                          70.0           79.0             61
10  female                         -99.0           80.0             65
11    male                          76.0           84.0            -99
12  female                          74.0           79.0             79

    placement score  club join year  placement offer
0              84.0            2020                2
1              86.0            2018                3
2              81.0            2020                2
3              96.0            2018                1
4               NaN            2021                3
5             100.0            2019                1
6             -99.0            2020                1
7              95.0            2019                3
8              88.0            2020                3
9              87.0            2021                2
10             85.0            2021                1
11              NaN            2020                2
12             98.0            2019                2
```

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

[ ]:
```
df
```

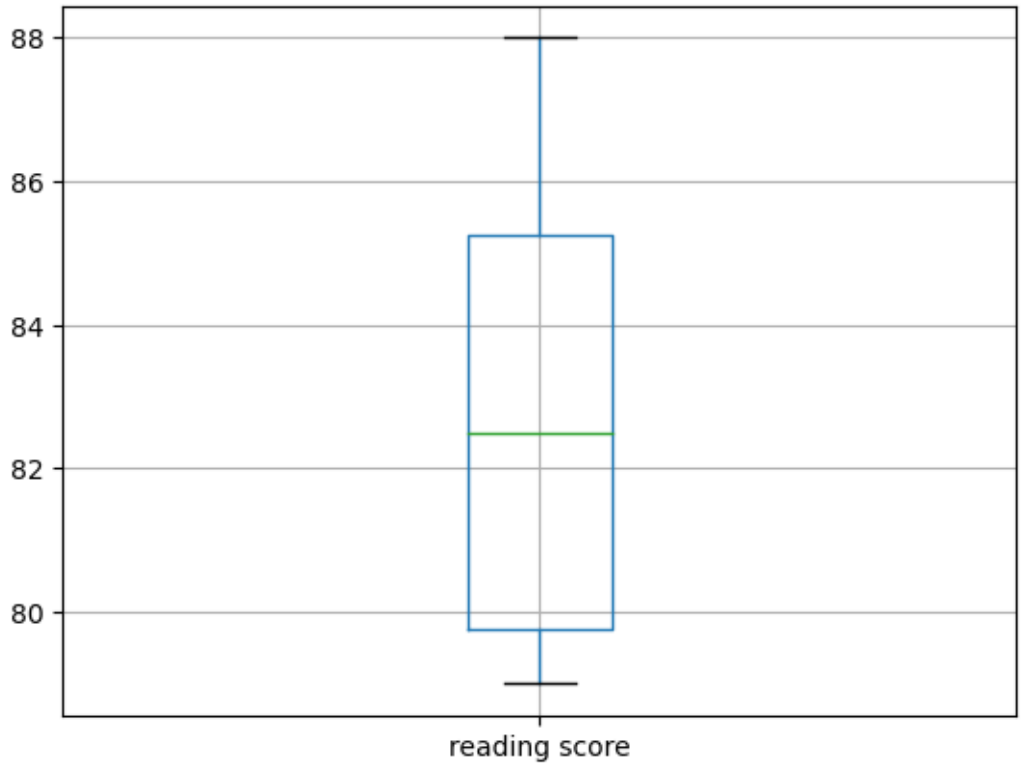
[ ]:
```
    gender  math score                      reading score  writing score  \
0   female                          63.0           84.0             64
1   female                          71.0           80.0             76
2   female                          64.0           81.0             66
3     male                          71.0           85.0             77
4     male                          68.0           86.0             76
5   female                          94.0           86.0             61
6     male                          75.0           79.0             66
7   female                           NaN            NaN             66
8     male                          66.0           88.0             66
9     male                          70.0           79.0             61
10  female                         -99.0           80.0             65
11    male                          76.0           84.0            -99
12  female                          74.0           79.0             79

    placement score  club join year  placement offer
0              84.0            2020                2
1              86.0            2018                3
2              81.0            2020                2
3              96.0            2018                1
4               NaN            2021                3
5             100.0            2019                1
```

6	-99.0	2020	1
7	95.0	2019	3
8	88.0	2020	3
9	87.0	2021	2
10	85.0	2021	1
11	NaN	2020	2
12	98.0	2019	2

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

[ ]: <AxesSubplot:>



[ ]:
```
df
```

[ ]:

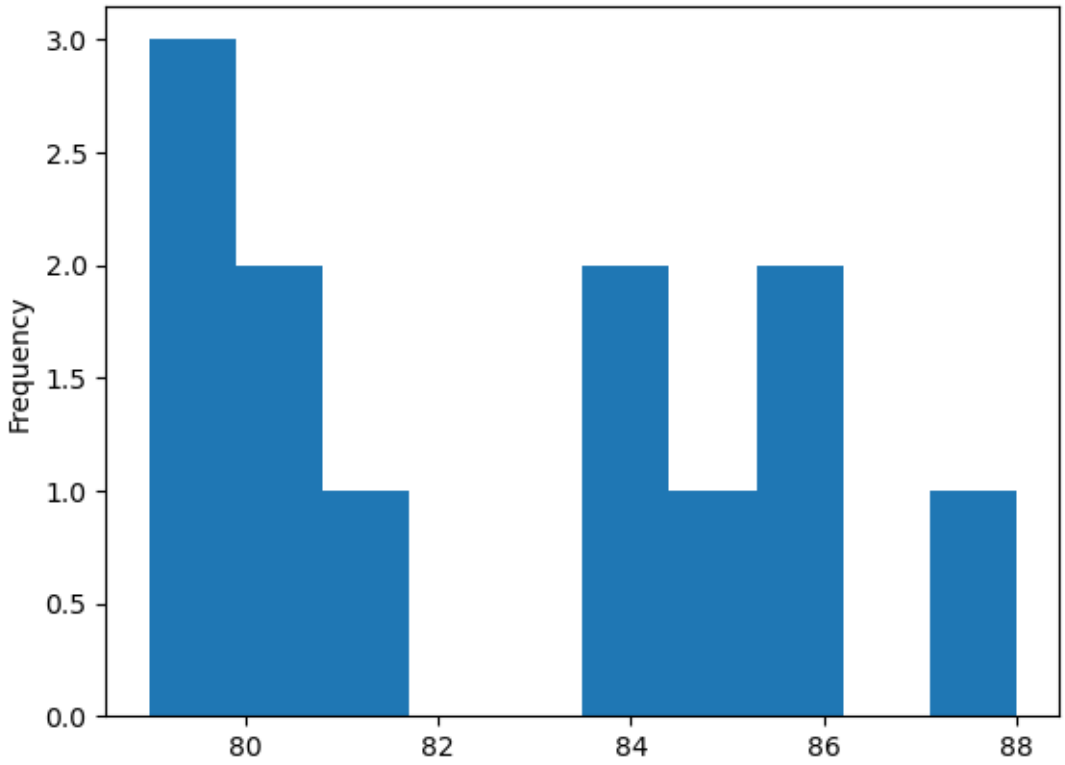
	gender	math score	reading score	writing score	\
0	female	63.0	84.0	64	
1	female	71.0	80.0	76	
2	female	64.0	81.0	66	
3	male	71.0	85.0	77	
4	male	68.0	86.0	76	
5	female	94.0	86.0	61	

```
6     male                          75.0           79.0             66
7   female                           NaN            NaN             66
8     male                          66.0           88.0             66
9     male                          70.0           79.0             61
10  female                         -99.0           80.0             65
11    male                          76.0           84.0            -99
12  female                          74.0           79.0             79
```

```
    placement score  club join year  placement offer
0              84.0            2020                2
1              86.0            2018                3
2              81.0            2020                2
3              96.0            2018                1
4               NaN            2021                3
5             100.0            2019                1
6             -99.0            2020                1
7              95.0            2019                3
8              88.0            2020                3
9              87.0            2021                2
10             85.0            2021                1
11              NaN            2020                2
12             98.0            2019                2
```

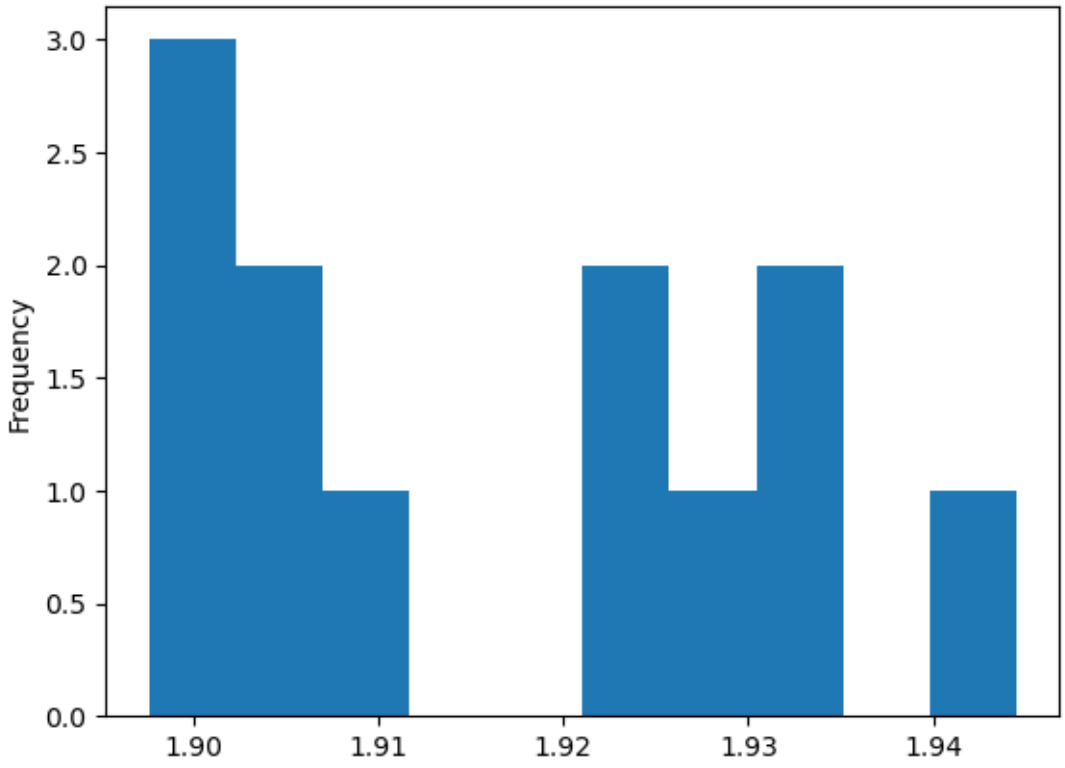
Reducing Skewness

[ ]:
```
import matplotlib.pyplot as plt
new_df['reading score'].plot(kind ='hist')
df['log_math'] = np.log10(df['reading score'])
```



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

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


Frequency
3.0
2.5
2.0
1.5
1.0
0.5
0.0
1.90
1.91
1.92
1.93
1.94