lbbyif3w7

January 31, 2025

#2 #Name :Rudrani Grirsh Jangale #Roll No. : 12

[146]:
```
import pandas as pd
import numpy as np
import seaborn as sns
from sklearn.preprocessing import OneHotEncoder
```

[147]:
```
from google.colab import drive
drive.mount('/content/drive')
file_path = '/content/drive/My Drive/StudentsPerformance.csv'
```

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

[148]:
```
#df = pd.read_csv('penguins_size.csv)
```

[149]:
```
df = pd.read_csv(file_path)
```

#Algorithm

[150]:
```
df.head()
```

[150]:
```
   gender race/ethnicity parental level of education         lunch  \
0  female        group B           bachelor's degree      standard
1  female        group C                some college      standard
2  female        group B             master's degree      standard
3    male        group A          associate's degree  free/reduced
4    male        group C                some college      standard

  test preparation course  math score  reading score  writing score
0                    none          72             72             74
1               completed          69             90             88
2                    none          90             95             93
3                    none          47             57             44
4                    none          76             78             75
```

[151]:
```
df.isnull()
```

[151]:

```
     gender  race/ethnicity  parental level of education  lunch  \
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
..      …             …                           …    …
995   False           False                        False  False
996   False           False                        False  False
997   False           False                        False  False
998   False           False                        False  False
999   False           False                        False  False

     test preparation course  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
..                       …         …             …            …
995                    False       False          False          False
996                    False       False          False          False
997                    False       False          False          False
998                    False       False          False          False
999                    False       False          False          False

[1000 rows x 8 columns]
```

[152]:

```
series = pd.isnull(df["math score"])
df[series]
```

[152]:

```
Empty DataFrame
Columns: [gender, race/ethnicity, parental level of education, lunch, test
preparation course, math score, reading score, writing score]
Index: []
```

#Algorithm

[153]:

```
df.notnull()
```

[153]:

```
     gender  race/ethnicity  parental level of education  lunch  \
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
..      …             …                           …    …
```

```
995    True            True                            True   True
996    True            True                            True   True
997    True            True                            True   True
998    True            True                            True   True
999    True            True                            True   True

     test preparation course  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
..                       …         …            …            …
995                     True        True           True           True
996                     True        True           True           True
997                     True        True           True           True
998                     True        True           True           True
999                     True        True           True           True

[1000 rows x 8 columns]
```

[154]:
```
series1 = pd.notnull(df["math score"])
df[series1]
```

[154]:
```
     gender race/ethnicity parental level of education         lunch  \
0    female        group B           bachelor's degree      standard
1    female        group C                some college      standard
2    female        group B             master's degree      standard
3      male        group A          associate's degree  free/reduced
4      male        group C                some college      standard
..      …            …                         …           …
995  female        group E             master's degree      standard
996    male        group C                 high school  free/reduced
997  female        group C                 high school  free/reduced
998  female        group D                some college      standard
999  female        group D                some college  free/reduced

    test preparation course  math score  reading score  writing score
0                      none          72             72             74
1                 completed          69             90             88
2                      none          90             95             93
3                      none          47             57             44
4                      none          76             78             75
..                      …         …            …            …
995               completed          88             99             95
996                    none          62             55             55
997               completed          59             71             65
```

```
998             completed          68          78          77
999                  none          77          86          86

[1000 rows x 8 columns]
```

```
[155]: print(df.isnull().sum())
```

```
gender                          0
race/ethnicity                  0
parental level of education     0
lunch                           0
test preparation course        31
math score                      0
reading score                   0
writing score                   0
dtype: int64
```

```
[156]: df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1000 entries, 0 to 999
Data columns (total 8 columns):
 #   Column                       Non-Null Count  Dtype
---  ------                       --------------  -----
 0   gender                       1000 non-null   object
 1   race/ethnicity               1000 non-null   object
 2   parental level of education  1000 non-null   object
 3   lunch                        1000 non-null   object
 4   test preparation course      969 non-null    object
 5   math score                   1000 non-null   int64
 6   reading score                1000 non-null   int64
 7   writing score                1000 non-null   int64
dtypes: int64(3), object(5)
memory usage: 62.6+ KB
```

```
[157]: print(df.isnull().values.any())  # Returns True if there are missing values
```

```
True
```

```
[158]: df.fillna(0, inplace=True)
```

```
[159]: print(df.isnull().sum())
```

```
gender                         0
race/ethnicity                 0
parental level of education    0
lunch                          0
test preparation course        0
```

```
math score                     0
reading score                  0
writing score                  0
dtype: int64
```

Here we have 30+ null value but now have o null value

[160]:
```
from sklearn.preprocessing import LabelEncoder
le = LabelEncoder()
df['gender'] = le.fit_transform(df['gender'])
newdf=df
df
```

[160]:
```
     gender race/ethnicity parental level of education         lunch  \
0         0        group B           bachelor's degree      standard
1         0        group C                some college      standard
2         0        group B             master's degree      standard
3         1        group A          associate's degree  free/reduced
4         1        group C                some college      standard
..      …            …                         …           …
995       0        group E             master's degree      standard
996       1        group C                 high school  free/reduced
997       0        group C                 high school  free/reduced
998       0        group D                some college      standard
999       0        group D                some college  free/reduced

    test preparation course  math score  reading score  writing score
0                      none          72             72             74
1                 completed          69             90             88
2                      none          90             95             93
3                      none          47             57             44
4                      none          76             78             75
..                      …         …             …             …
995               completed          88             99             95
996                    none          62             55             55
997               completed          59             71             65
998               completed          68             78             77
999                    none          77             86             86

[1000 rows x 8 columns]
```

[161]:
```
#For replacing null values with NaN
missing_values = ["Na", "na"]
```

[162]:
```
df
```

[162]:
```
     gender race/ethnicity parental level of education         lunch  \
0         0        group B           bachelor's degree      standard
```

```
1         0        group C                some college      standard   
2         0        group B             master's degree      standard   
3         1        group A          associate's degree  free/reduced   
4         1        group C                some college      standard   
..      …            …                         …           …   
995       0        group E             master's degree      standard   
996       1        group C                 high school  free/reduced   
997       0        group C                 high school  free/reduced   
998       0        group D                some college      standard   
999       0        group D                some college  free/reduced   

    test preparation course  math score  reading score  writing score  
0                      none          72             72             74  
1                 completed          69             90             88  
2                      none          90             95             93  
3                      none          47             57             44  
4                      none          76             78             75  
..                      …         …            …            …  
995               completed          88             99             95  
996                    none          62             55             55  
997               completed          59             71             65  
998               completed          68             78             77  
999                    none          77             86             86  

[1000 rows x 8 columns]
```

#Filling null values with a single value

[163]:

```
ndf=df
ndf.fillna(0)
```

[163]:

```
     gender race/ethnicity parental level of education         lunch  \
0         0        group B           bachelor's degree      standard   
1         0        group C                some college      standard   
2         0        group B             master's degree      standard   
3         1        group A          associate's degree  free/reduced   
4         1        group C                some college      standard   
..      …            …                         …           …   
995       0        group E             master's degree      standard   
996       1        group C                 high school  free/reduced   
997       0        group C                 high school  free/reduced   
998       0        group D                some college      standard   
999       0        group D                some college  free/reduced   

    test preparation course  math score  reading score  writing score  
0                      none          72             72             74  
1                 completed          69             90             88  
```

```
2                  none         90          95          93
3                  none         47          57          44
4                  none         76          78          75
..                    …      …          …          …
995           completed         88          99          95
996                none         62          55          55
997           completed         59          71          65
998           completed         68          78          77
999                none         77          86          86

[1000 rows x 8 columns]
```

[164]:
```
df['math score'] = df['math score'].fillna(df['math score'].mean())
```

[165]:
```
df['math score'] = df['math score'].fillna(df['math score'].median())
```

[166]:
```
df['math score'] = df['math score'].fillna(df['math score'].std())
```

[167]:
```
df['math score'] = df['math score'].fillna(df['math score'].min())
```

[168]:
```
df['math score'] = df['math score'].fillna(df['math score'].max())
```

#Filling null values in dataset

[169]:
```
m_v=df['math score'].mean()
df['math score'].fillna(value=m_v, inplace=True)
df
```

```
<ipython-input-169-0ff51d643ba7>: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['math score'].fillna(value=m_v, inplace=True)
```

[169]:
```
     gender race/ethnicity parental level of education         lunch  \
0         0        group B           bachelor's degree      standard
1         0        group C                some college      standard
2         0        group B             master's degree      standard
3         1        group A          associate's degree  free/reduced
```

```
4         1         group C              some college      standard
..      …          …                       …          …
995       0         group E            master's degree      standard
996       1         group C                high school  free/reduced
997       0         group C                high school  free/reduced
998       0         group D               some college      standard
999       0         group D               some college  free/reduced

    test preparation course  math score  reading score  writing score
0                      none          72             72             74
1                 completed          69             90             88
2                      none          90             95             93
3                      none          47             57             44
4                      none          76             78             75
..                      …          …             …             …
995               completed          88             99             95
996                    none          62             55             55
997               completed          59             71             65
998               completed          68             78             77
999                    none          77             86             86

[1000 rows x 8 columns]
```

#Filling a null values using replace() method

[170]:
```
ndf.replace(to_replace = np.nan, value = -99)
```

[170]:
```
     gender race/ethnicity parental level of education         lunch  \
0         0         group B           bachelor's degree      standard
1         0         group C                some college      standard
2         0         group B             master's degree      standard
3         1         group A          associate's degree  free/reduced
4         1         group C                some college      standard
..      …          …                       …          …
995       0         group E             master's degree      standard
996       1         group C                 high school  free/reduced
997       0         group C                 high school  free/reduced
998       0         group D                some college      standard
999       0         group D                some college  free/reduced

    test preparation course  math score  reading score  writing score
0                      none          72             72             74
1                 completed          69             90             88
2                      none          90             95             93
3                      none          47             57             44
4                      none          76             78             75
..                      …          …             …             …
```

```
995               completed          88             99             95
996                    none          62             55             55
997               completed          59             71             65
998               completed          68             78             77
999                    none          77             86             86

[1000 rows x 8 columns]
```

#Algorithm

[170]:

```

```

[171]:

```
df
```

[171]:

```
     gender race/ethnicity parental level of education         lunch  \
0         0        group B           bachelor's degree      standard
1         0        group C                some college      standard
2         0        group B             master's degree      standard
3         1        group A          associate's degree  free/reduced
4         1        group C                some college      standard
..      …            …                          …           …
995       0        group E             master's degree      standard
996       1        group C                 high school  free/reduced
997       0        group C                 high school  free/reduced
998       0        group D                some college      standard
999       0        group D                some college  free/reduced

    test preparation course  math score  reading score  writing score
0                      none          72             72             74
1                 completed          69             90             88
2                      none          90             95             93
3                      none          47             57             44
4                      none          76             78             75
..                      …          …            …            …
995               completed          88             99             95
996                    none          62             55             55
997               completed          59             71             65
998               completed          68             78             77
999                    none          77             86             86

[1000 rows x 8 columns]
```

[172]:

```
ndf.dropna()
```

[172]:

```
     gender race/ethnicity parental level of education         lunch  \
0         0        group B           bachelor's degree      standard
1         0        group C                some college      standard
```

```
2         0        group B             master's degree      standard
3         1        group A          associate's degree  free/reduced
4         1        group C                some college      standard
..      …             …                         …           …
995       0        group E             master's degree      standard
996       1        group C                 high school  free/reduced
997       0        group C                 high school  free/reduced
998       0        group D                some college      standard
999       0        group D                some college  free/reduced

    test preparation course  math score  reading score   writing score
0                      none          72             72              74
1                 completed          69             90              88
2                      none          90             95              93
3                      none          47             57              44
4                      none          76             78              75
..                       …         …            …            …
995               completed          88             99              95
996                    none          62             55              55
997               completed          59             71              65
998               completed          68             78              77
999                    none          77             86              86

[1000 rows x 8 columns]
```

[173]:

```
ndf.dropna(how = 'all')
```

[173]:

```
     gender race/ethnicity parental level of education         lunch  \
0         0        group B           bachelor's degree      standard
1         0        group C                some college      standard
2         0        group B             master's degree      standard
3         1        group A          associate's degree  free/reduced
4         1        group C                some college      standard
..      …             …                         …           …
995       0        group E             master's degree      standard
996       1        group C                 high school  free/reduced
997       0        group C                 high school  free/reduced
998       0        group D                some college      standard
999       0        group D                some college  free/reduced

    test preparation course  math score  reading score   writing score
0                      none          72             72              74
1                 completed          69             90              88
2                      none          90             95              93
3                      none          47             57              44
4                      none          76             78              75
..                       …         …            …            …
```

```
995             completed          88            99             95
996                  none          62            55             55
997             completed          59            71             65
998             completed          68            78             77
999                  none          77            86             86

[1000 rows x 8 columns]
```

[174]:
```
ndf.dropna(axis = 1)
```

[174]:
```
     gender race/ethnicity parental level of education         lunch  \
0         0        group B           bachelor's degree      standard
1         0        group C                some college      standard
2         0        group B             master's degree      standard
3         1        group A          associate's degree  free/reduced
4         1        group C                some college      standard
..      …            …                        …            …
995       0        group E             master's degree      standard
996       1        group C                 high school  free/reduced
997       0        group C                 high school  free/reduced
998       0        group D                some college      standard
999       0        group D                some college  free/reduced

    test preparation course  math score  reading score  writing score
0                      none          72             72             74
1                 completed          69             90             88
2                      none          90             95             93
3                      none          47             57             44
4                      none          76             78             75
..                      …          …            …             …
995             completed          88            99             95
996                  none          62            55             55
997             completed          59            71             65
998             completed          68            78             77
999                  none          77            86             86

[1000 rows x 8 columns]
```

[175]:
```
new_data = ndf.dropna(axis = 0, how ='any')
```

[176]:
```
new_data
```

[176]:
```
     gender race/ethnicity parental level of education         lunch  \
0         0        group B           bachelor's degree      standard
1         0        group C                some college      standard
2         0        group B             master's degree      standard
3         1        group A          associate's degree  free/reduced
```

```
4         1        group C                some college      standard   
..      …            …                         …           …   
995       0        group E             master's degree      standard   
996       1        group C                 high school  free/reduced   
997       0        group C                 high school  free/reduced   
998       0        group D                some college      standard   
999       0        group D                some college  free/reduced   

    test preparation course  math score  reading score  writing score  
0                      none          72             72             74  
1                 completed          69             90             88  
2                      none          90             95             93  
3                      none          47             57             44  
4                      none          76             78             75  
..                      …         …            …            …  
995               completed          88             99             95  
996                    none          62             55             55  
997               completed          59             71             65  
998               completed          68             78             77  
999                    none          77             86             86  

[1000 rows x 8 columns]
```

[177]:
```
df
```

[177]:
```
     gender race/ethnicity parental level of education         lunch  \
0         0        group B           bachelor's degree      standard   
1         0        group C                some college      standard   
2         0        group B             master's degree      standard   
3         1        group A          associate's degree  free/reduced   
4         1        group C                some college      standard   
..      …            …                         …           …   
995       0        group E             master's degree      standard   
996       1        group C                 high school  free/reduced   
997       0        group C                 high school  free/reduced   
998       0        group D                some college      standard   
999       0        group D                some college  free/reduced   

    test preparation course  math score  reading score  writing score  
0                      none          72             72             74  
1                 completed          69             90             88  
2                      none          90             95             93  
3                      none          47             57             44  
4                      none          76             78             75  
..                      …         …            …            …  
995               completed          88             99             95  
996                    none          62             55             55  
```

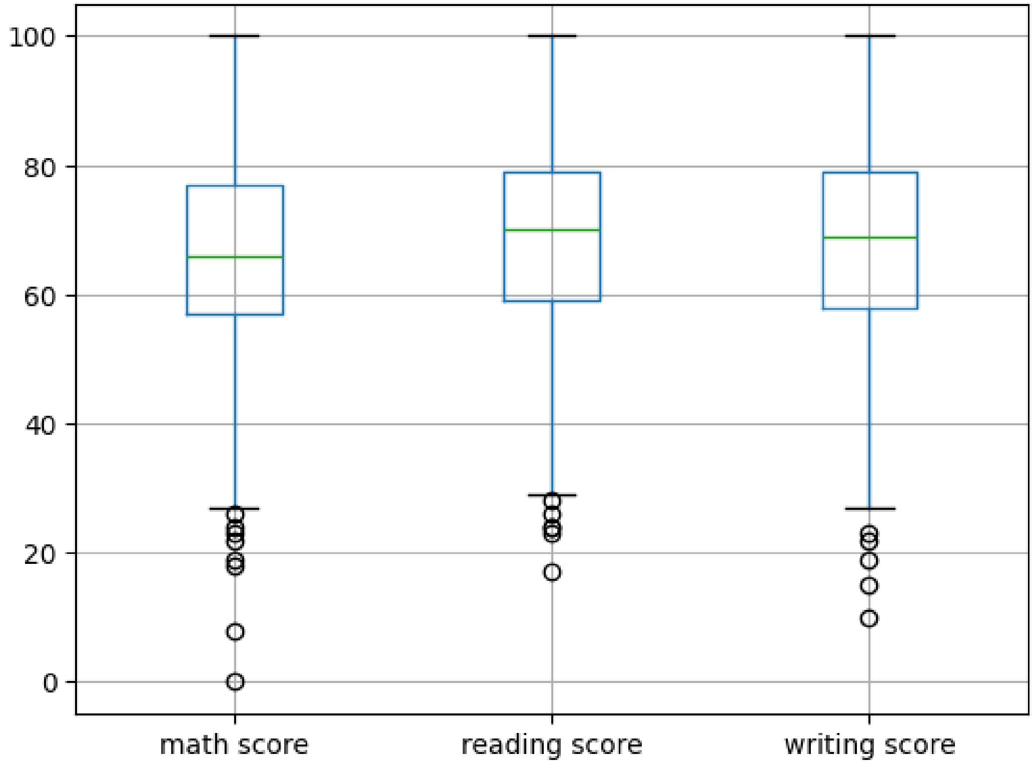
```
997                  completed          59             71             65
998                  completed          68             78             77
999                       none          77             86             86

[1000 rows x 8 columns]
```

#Identification outliers

```
[178]: df = pd.read_csv(file_path)
```

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

```
[179]: <Axes: >
```



```
[180]: print(np.where(df['math score']>90))
```

```
(array([ 34, 104, 114, 121, 149, 165, 171, 179, 233, 263, 286, 306, 451,
       458, 469, 501, 503, 521, 539, 546, 562, 566, 571, 594, 612, 618,
       623, 625, 685, 689, 710, 712, 717, 719, 736, 779, 784, 815, 846,
       855, 864, 886, 903, 916, 919, 934, 950, 957, 962, 979]),)
```

[181]:

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

```
(array([ 59, 327, 596, 980]),)
```

[182]:

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

```
(array([ 17,  59,  76, 211, 327, 338, 596, 896, 980]),)
```

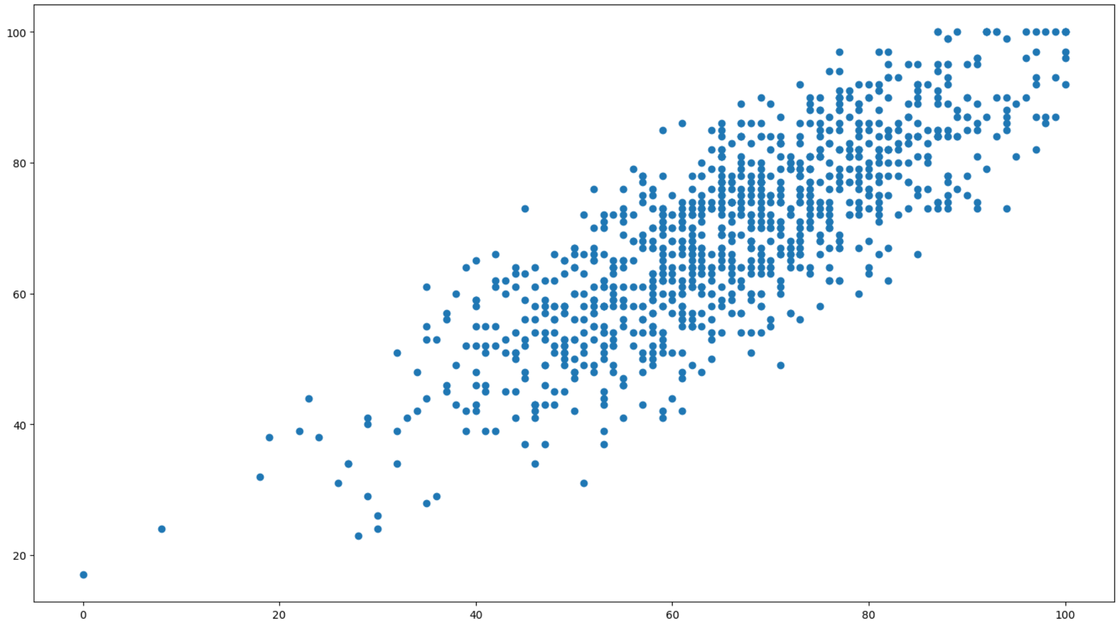
#Algorithm

[183]:

```
import matplotlib.pyplot as plt
```

[184]:

```
fig, ax = plt.subplots(figsize = (18,10))
ax.scatter(df['math score'], df['reading score'])
plt.show()
```



[185]:

```
print(np.where((df['math score']<50) & (df['reading score']>1)))
```

```
(array([  3,   7,   9,  11,  17,  18,  22,  33,  55,  59,  61,  66,  69,
         72,  74,  75,  76,  80,  81,  84,  91,  93, 142, 145, 162, 174,
        176, 181, 184, 188, 198, 211, 212, 217, 225, 231, 250, 262, 272,
        281, 284, 296, 298, 309, 323, 324, 327, 329, 331, 337, 338, 339,
        353, 357, 363, 365, 368, 371, 375, 383, 384, 395, 402, 422, 424,
        433, 448, 455, 466, 484, 504, 527, 528, 531, 552, 555, 564, 565,
        575, 578, 587, 589, 596, 601, 607, 616, 620, 628, 629, 640, 658,
        683, 690, 694, 706, 709, 724, 741, 761, 775, 777, 780, 785, 787,
        790, 794, 807, 811, 816, 822, 824, 840, 842, 844, 862, 869, 874,
        889, 895, 896, 902, 913, 914, 917, 921, 928, 929, 947, 948, 958,
        961, 973, 980, 986, 988]),)
```

[186]:

```
print(np.where((df['math score']>85) & (df['writing score']<3)))
```

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

#algorithm

[187]:

```
import numpy as np
from scipy import stats
```

[188]:

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

[189]:

```
print(z)
```

```
0      0.390024
1      0.192076
2      1.577711
3      1.259543
4      0.653954
         …
995    1.445746
996    0.269803
997    0.467751
998    0.126093
999    0.719937
Name: math score, Length: 1000, dtype: float64
```

[190]:

```
threshold = 0.18
```

[191]:

```
sample_outliers = np.where(z <threshold)
sample_outliers
```

[191]:

```
(array([  8,  12,  20,  21,  27,  45,  47,  65,  90,  96,  99, 101, 107,
        127, 148, 156, 159, 168, 169, 183, 185, 190, 200, 201, 218, 228,
        232, 237, 248, 249, 256, 259, 260, 273, 278, 287, 293, 295, 302,
        311, 312, 313, 320, 343, 351, 379, 385, 386, 394, 406, 418, 428,
        429, 430, 440, 445, 450, 452, 453, 472, 482, 488, 490, 491, 495,
        498, 506, 511, 517, 518, 519, 525, 530, 535, 544, 569, 585, 592,
        599, 613, 619, 630, 632, 636, 645, 647, 651, 653, 663, 670, 673,
        680, 692, 699, 707, 726, 727, 730, 735, 751, 768, 774, 776, 788,
        792, 800, 806, 827, 829, 832, 839, 841, 847, 857, 879, 882, 898,
        899, 904, 908, 915, 926, 927, 930, 936, 963, 966, 968, 975, 989,
        991, 998]),)
```

#Algorithm

[192]:

```
import numpy as np
```

[193]:

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

[194]:

```
first_ten = sorted_rscore[:10]
last_ten = sorted_rscore[-10:]
```

[195]:

```
print("First 10:", first_ten)
print("Last 10:", last_ten)
```

```
First 10: [17, 23, 24, 24, 26, 28, 29, 29, 31, 31]
Last 10: [100, 100, 100, 100, 100, 100, 100, 100, 100, 100]
```

[196]:

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

```
59.0 79.0
```

[197]:

```
IQR = q3-q1
```

[198]:

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

```
29.0 109.0
```

#Handling Outliers

[199]:

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

```
[17, 23, 24, 24, 26, 28]
```

[200]:

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

[200]:

```
     gender race/ethnicity parental level of education         lunch  \
0    female        group B           bachelor's degree      standard
1    female        group C                some college      standard
2    female        group B             master's degree      standard
3      male        group A          associate's degree  free/reduced
4      male        group C                some college      standard
..      …            …                           …            …
994    male        group A                 high school      standard
```

```
995  female         group E            master's degree      standard
996    male         group C                high school  free/reduced
997  female         group C                high school  free/reduced
999  female         group D               some college  free/reduced

    test preparation course  math score  reading score  writing score
0                      none          72             72             74
1                 completed          69             90             88
2                      none          90             95             93
3                      none          47             57             44
4                      none          76             78             75
..                      …          …            …            …
994                    none          63             63             62
995               completed          88             99             95
996                    none          62             55             55
997               completed          59             71             65
999                    none          77             86             86

[868 rows x 8 columns]
```

```
[201]: df = pd.read_csv(file_path)
```

```
[202]: 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: [72. 69. 86. 47. 76. 71. 86. 40. 64. 38. 58. 40. 65. 78. 50. 69. 86.
18.
 46. 54. 66. 65. 44. 69. 74. 73. 69. 67. 70. 62. 69. 63. 56. 40. 86. 81.
 74. 50. 75. 57. 55. 58. 53. 59. 50. 65. 55. 66. 57. 82. 53. 77. 53. 86.
 71. 33. 82. 52. 58.  0. 79. 39. 62. 69. 59. 67. 45. 60. 61. 39. 58. 63.
 41. 61. 49. 44. 30. 80. 61. 62. 47. 49. 50. 72. 42. 73. 76. 71. 58. 73.
 65. 27. 71. 43. 79. 78. 65. 63. 58. 65. 79. 68. 85. 60. 86. 58. 86. 66.
 52. 70. 77. 62. 54. 51. 86. 84. 75. 78. 51. 55. 79. 86. 86. 63. 83. 86.
 72. 65. 82. 51. 86. 53. 86. 75. 74. 58. 51. 70. 59. 71. 76. 59. 42. 57.
 86. 22. 86. 73. 68. 86. 62. 77. 59. 54. 62. 70. 66. 60. 61. 66. 82. 75.
 49. 52. 81. 86. 53. 58. 68. 67. 72. 86. 79. 63. 43. 81. 46. 71. 52. 86.
 62. 46. 50. 65. 45. 65. 80. 62. 48. 77. 66. 76. 62. 77. 69. 61. 59. 55.
 45. 78. 67. 65. 69. 57. 59. 74. 82. 81. 74. 58. 80. 35. 42. 60. 86. 84.
 83. 34. 66. 61. 56. 86. 55. 86. 52. 45. 72. 57. 68. 86. 76. 46. 67. 86.
 83. 80. 63. 64. 54. 84. 73. 80. 56. 59. 75. 85. 86. 58. 65. 68. 47. 71.
 60. 80. 54. 62. 64. 78. 70. 65. 64. 79. 44. 86. 76. 59. 63. 69. 86. 71.
 69. 58. 47. 65. 86. 83. 85. 59. 65. 73. 53. 45. 73. 70. 37. 81. 86. 67.
 86. 77. 76. 86. 63. 65. 78. 67. 46. 71. 40. 86. 81. 56. 67. 80. 74. 69.
 86. 51. 53. 49. 73. 66. 67. 68. 59. 71. 77. 83. 63. 56. 67. 75. 71. 43.
 41. 82. 61. 28. 82. 41. 71. 47. 62. 86. 83. 61. 76. 49. 24. 35. 58. 61.
 69. 67. 79. 72. 62. 77. 75. 86. 52. 66. 63. 46. 59. 61. 63. 42. 59. 80.
 58. 85. 52. 27. 59. 49. 69. 61. 44. 73. 84. 45. 74. 82. 59. 46. 80. 85.
 71. 66. 80. 86. 79. 38. 38. 67. 64. 57. 62. 73. 73. 77. 76. 57. 65. 48.
 50. 85. 74. 60. 59. 53. 49. 86. 54. 63. 65. 82. 52. 86. 70. 84. 71. 63.
 51. 84. 71. 74. 68. 57. 82. 57. 47. 59. 41. 62. 86. 69. 65. 68. 64. 61.
 61. 47. 73. 50. 75. 75. 70. 86. 67. 78. 59. 73. 79. 67. 69. 86. 47. 81.
 64. 86. 65. 65. 53. 37. 79. 53. 86. 72. 53. 54. 71. 77. 75. 84. 26. 72.
 77. 86. 83. 63. 68. 59. 86. 71. 76. 80. 55. 76. 73. 52. 68. 59. 49. 70.
 61. 60. 64. 79. 65. 64. 83. 81. 54. 68. 54. 59. 66. 76. 74. 86. 63. 86.
 40. 82. 68. 55. 79. 86. 76. 64. 62. 54. 77. 76. 74. 66. 66. 67. 71. 86.
 69. 54. 53. 68. 56. 36. 29. 62. 68. 47. 62. 79. 73. 66. 51. 51. 85. 86.
 75. 79. 81. 82. 64. 78. 86. 72. 62. 79. 79. 86. 40. 77. 53. 32. 55. 61.
 53. 73. 74. 63. 86. 63. 48. 48. 86. 61. 63. 68. 71. 86. 53. 50. 74. 40.
 61. 81. 48. 53. 81. 77. 63. 73. 69. 65. 55. 44. 54. 48. 58. 71. 68. 74.
 86. 56. 30. 53. 69. 65. 54. 29. 76. 60. 84. 75. 85. 40. 61. 58. 69. 58.
 86. 65. 82. 60. 37. 86. 86. 65. 35. 62. 58. 86. 61. 86. 69. 61. 49. 44.
 67. 79. 66. 75. 84. 71. 67. 80. 86. 76. 41. 74. 72. 74. 70. 65. 59. 64.
 50. 69. 51. 68. 85. 65. 73. 62. 77. 69. 43. 86. 74. 73. 55. 65. 80. 50.
 63. 77. 73. 81. 66. 52. 69. 65. 69. 50. 73. 70. 81. 63. 67. 60. 62. 29.
 62. 86. 85. 77. 53. 86. 49. 73. 66. 77. 49. 79. 75. 59. 57. 66. 79. 57.
 86. 63. 59. 62. 46. 66. 86. 42. 86. 80. 86. 81. 60. 76. 73. 86. 76. 86.
 62. 55. 74. 50. 47. 81. 65. 68. 73. 53. 68. 55. 86. 55. 53. 67. 86. 53.
 81. 61. 80. 37. 81. 59. 55. 72. 69. 69. 50. 86. 71. 68. 79. 77. 58. 84.
 55. 70. 52. 69. 53. 48. 78. 62. 60. 74. 58. 76. 68. 58. 52. 75. 52. 62.
 66. 49. 66. 35. 72. 86. 46. 77. 76. 52. 86. 32. 72. 19. 68. 52. 48. 60.
 66. 86. 42. 57. 70. 70. 69. 52. 67. 76. 86. 82. 73. 75. 64. 41. 86. 59.
 51. 45. 54. 86. 72. 86. 45. 61. 60. 77. 85. 78. 49. 71. 48. 62. 56. 65.
 69. 68. 61. 74. 64. 77. 58. 60. 73. 75. 58. 66. 39. 64. 23. 74. 40. 86.
 86. 64. 59. 80. 71. 61. 86. 82. 62. 86. 75. 65. 52. 86. 53. 81. 39. 71.
 86. 82. 59. 61. 78. 49. 59. 70. 82. 86. 43. 80. 81. 57. 59. 64. 63. 71.
 64. 55. 51. 62. 86. 54. 69. 44. 86. 85. 50. 86. 59. 32. 36. 63. 67. 65.
 85. 73. 34. 86. 67. 86. 57. 79. 67. 70. 50. 69. 52. 47. 46. 68. 86. 44.
 57. 86. 69. 35. 72. 54. 74. 74. 64. 65. 46. 48. 67. 62. 61. 70. 86. 70.
 67. 57. 85. 77. 72. 78. 81. 61. 58. 54. 82. 49. 49. 57. 86. 75. 74. 58.
 62. 72. 84. 86. 45. 75. 56. 48. 86. 65. 72. 62. 66. 63. 68. 75. 86. 78.
 53. 49. 54. 64. 60. 62. 55. 86.  8. 81. 79. 78. 74. 57. 40. 81. 44. 67.
 86. 65. 55. 62. 63. 86. 62. 59. 68. 77.]
```

[203]:

```
df_stud.insert(1,"m score",b,True)
df_stud
```

[203]:

```
  gender  m score race/ethnicity parental level of education         lunch  \
0  female     72.0        group B           bachelor's degree      standard
1  female     69.0        group C                some college      standard
2  female     86.0        group B             master's degree      standard
3    male     47.0        group A          associate's degree  free/reduced
```

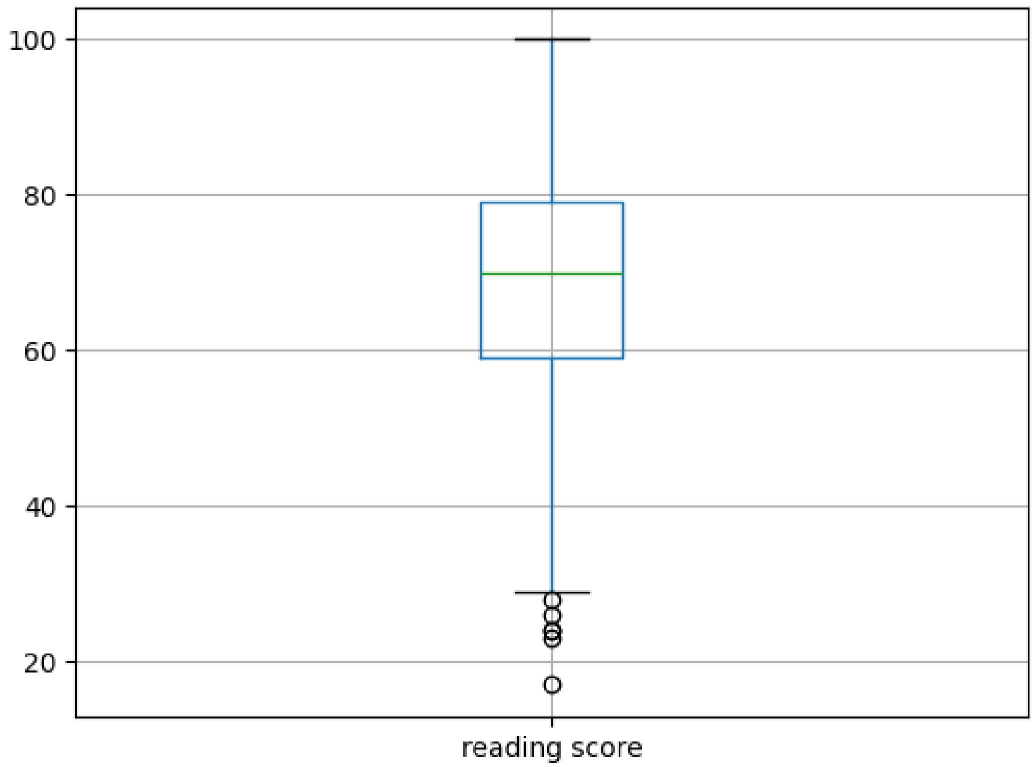
```
4      male     76.0        group C                some college      standard
..      …      …            …                         …           …
995  female     86.0        group E             master's degree      standard
996    male     62.0        group C                 high school  free/reduced
997  female     59.0        group C                 high school  free/reduced
998  female     68.0        group D                some college      standard
999  female     77.0        group D                some college  free/reduced

    test preparation course  math score  reading score  writing score
0                      none          72             72             74
1                 completed          69             90             88
2                      none          90             95             93
3                      none          47             57             44
4                      none          76             78             75
..                      …         …            …            …
995               completed          88             99             95
996                    none          62             55             55
997               completed          59             71             65
998               completed          68             78             77
999                    none          77             86             86

[1000 rows x 9 columns]
```

[204]:

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

[204]:

```
<Axes: >
```



[205]:
```
median=np.median(sorted_rscore)
median
```

[205]:
```
70.0
```

[206]:
```
refined_df=df
refined_df['reading score'] = np.where(refined_df['reading score'] >upr_bound,␣
  ↪median,refined_df['reading score'])
```

[207]:
```
refined_df
```

[207]:
```
     gender  m score race/ethnicity parental level of education         lunch  \
0    female     72.0        group B           bachelor's degree      standard
1    female     69.0        group C                some college      standard
2    female     86.0        group B             master's degree      standard
3      male     47.0        group A          associate's degree  free/reduced
4      male     76.0        group C                some college      standard
..      …        …              …                         …           …
995  female     86.0        group E             master's degree      standard
996    male     62.0        group C                 high school  free/reduced
997  female     59.0        group C                 high school  free/reduced
998  female     68.0        group D                some college      standard
```

```
999  female     77.0        group D                some college  free/reduced

    test preparation course  math score  reading score  writing score
0                      none          72           72.0             74
1                 completed          69           90.0             88
2                      none          90           95.0             93
3                      none          47           57.0             44
4                      none          76           78.0             75
..                      …         …            …            …
995               completed          88           99.0             95
996                    none          62           55.0             55
997               completed          59           71.0             65
998               completed          68           78.0             77
999                    none          77           86.0             86

[1000 rows x 9 columns]
```

[208]:
```
refined_df['reading score'] = np.where(refined_df['reading score'] <lwr_bound,␣
 ↪median,refined_df['reading score'])
```

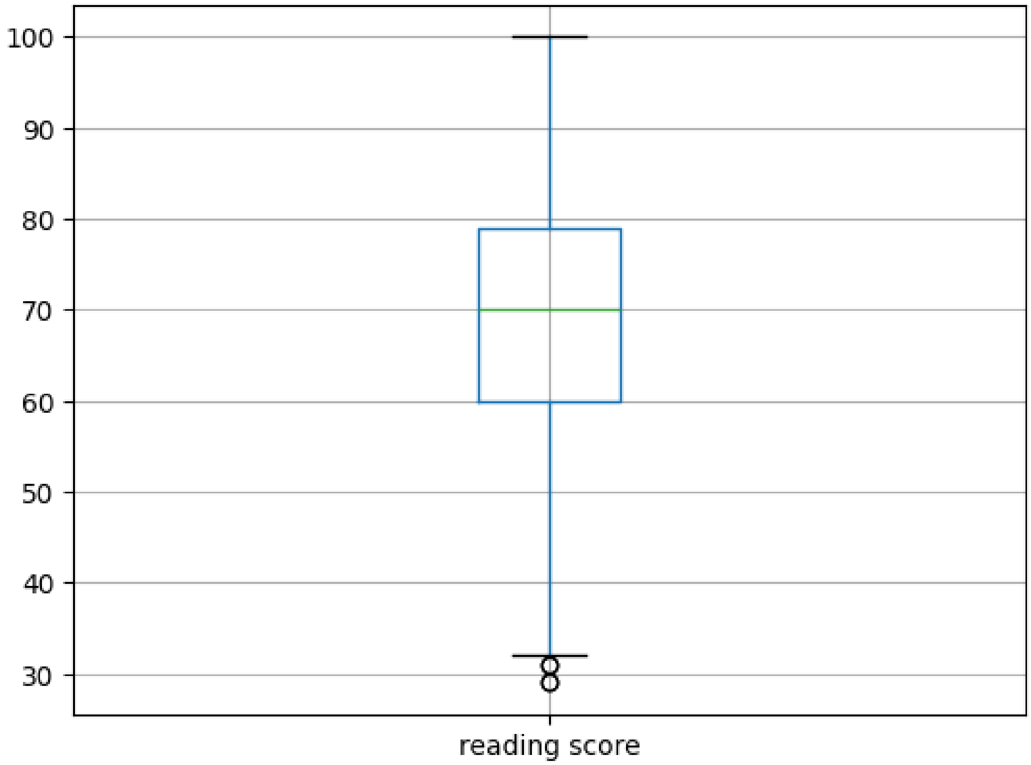
[209]:
```
refined_df
```

[209]:
```
     gender  m score race/ethnicity parental level of education         lunch  \
0    female     72.0        group B           bachelor's degree      standard
1    female     69.0        group C                some college      standard
2    female     86.0        group B             master's degree      standard
3      male     47.0        group A          associate's degree  free/reduced
4      male     76.0        group C                some college      standard
..      …       …             …                          …           …
995  female     86.0        group E             master's degree      standard
996    male     62.0        group C                 high school  free/reduced
997  female     59.0        group C                 high school  free/reduced
998  female     68.0        group D                some college      standard
999  female     77.0        group D                some college  free/reduced

    test preparation course  math score  reading score  writing score
0                      none          72           72.0             74
1                 completed          69           90.0             88
2                      none          90           95.0             93
3                      none          47           57.0             44
4                      none          76           78.0             75
..                      …         …            …            …
995               completed          88           99.0             95
996                    none          62           55.0             55
997               completed          59           71.0             65
998               completed          68           78.0             77
999                    none          77           86.0             86
```

```
[1000 rows x 9 columns]
```

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

```
[210]: <Axes: >
```



#Algorithm

```
[211]: df = pd.read_csv(file_path)
```

```
[212]: df
```

```
[212]:     gender race/ethnicity parental level of education         lunch  \
       0    female        group B           bachelor's degree      standard
       1    female        group C                some college      standard
       2    female        group B             master's degree      standard
       3      male        group A          associate's degree  free/reduced
       4      male        group C                some college      standard
       ..      …            …                          …            …
       995  female        group E             master's degree      standard
```

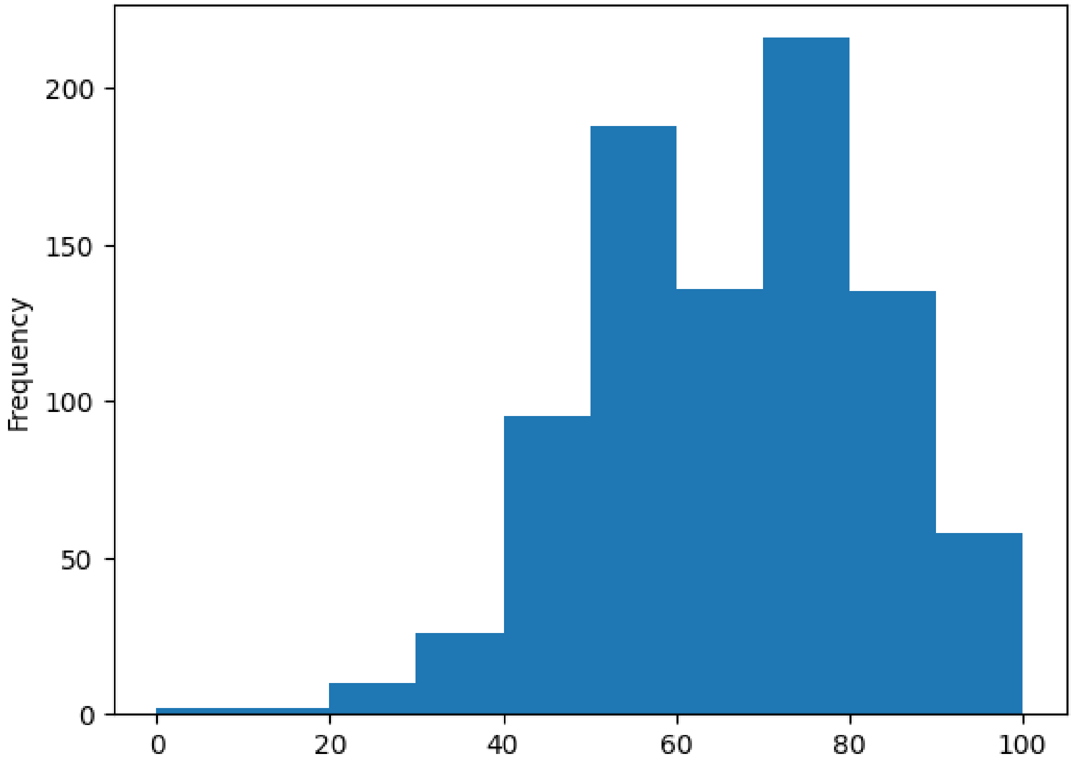
```
996    male        group C                 high school  free/reduced
997  female        group C                 high school  free/reduced
998  female        group D                some college      standard
999  female        group D                some college  free/reduced

    test preparation course  math score  reading score  writing score
0                      none          72             72             74
1                 completed          69             90             88
2                      none          90             95             93
3                      none          47             57             44
4                      none          76             78             75
..                      …          …            …            …
995               completed          88             99             95
996                    none          62             55             55
997               completed          59             71             65
998               completed          68             78             77
999                    none          77             86             86

[1000 rows x 8 columns]
```

[213]:

```
import matplotlib.pyplot as plt
new_df['math score'].plot(kind = 'hist')
```

[213]:

```
<Axes: ylabel='Frequency'>
```

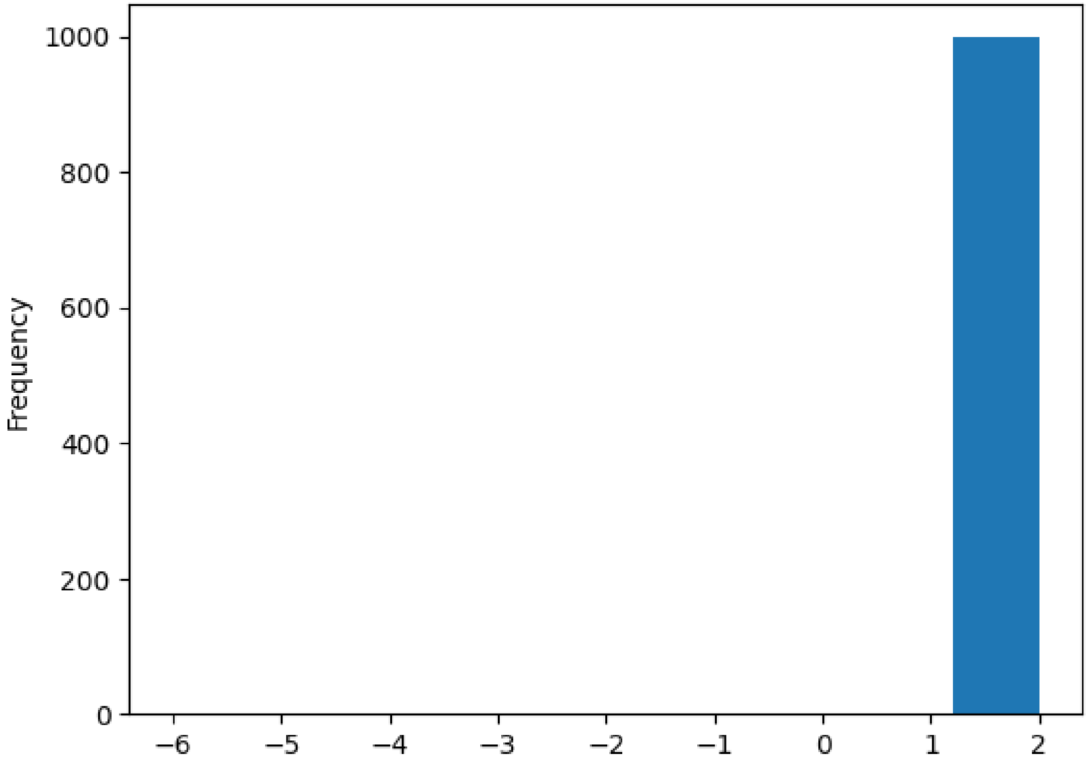


[214]:
```
df['log_math'] = np.log10(df['math score'])
```

```
/usr/local/lib/python3.11/dist-packages/pandas/core/arraylike.py:399:
RuntimeWarning: divide by zero encountered in log10
  result = getattr(ufunc, method)(*inputs, **kwargs)
```

[215]:
```
import numpy as np
import pandas as pd

df['math score'] = df['math score'].apply(lambda x: 1e-6 if x <= 0 else x)


df['log_math'] = np.log10(df['math score'])

# Create the histogram
df['log_math'].plot(kind='hist')
```

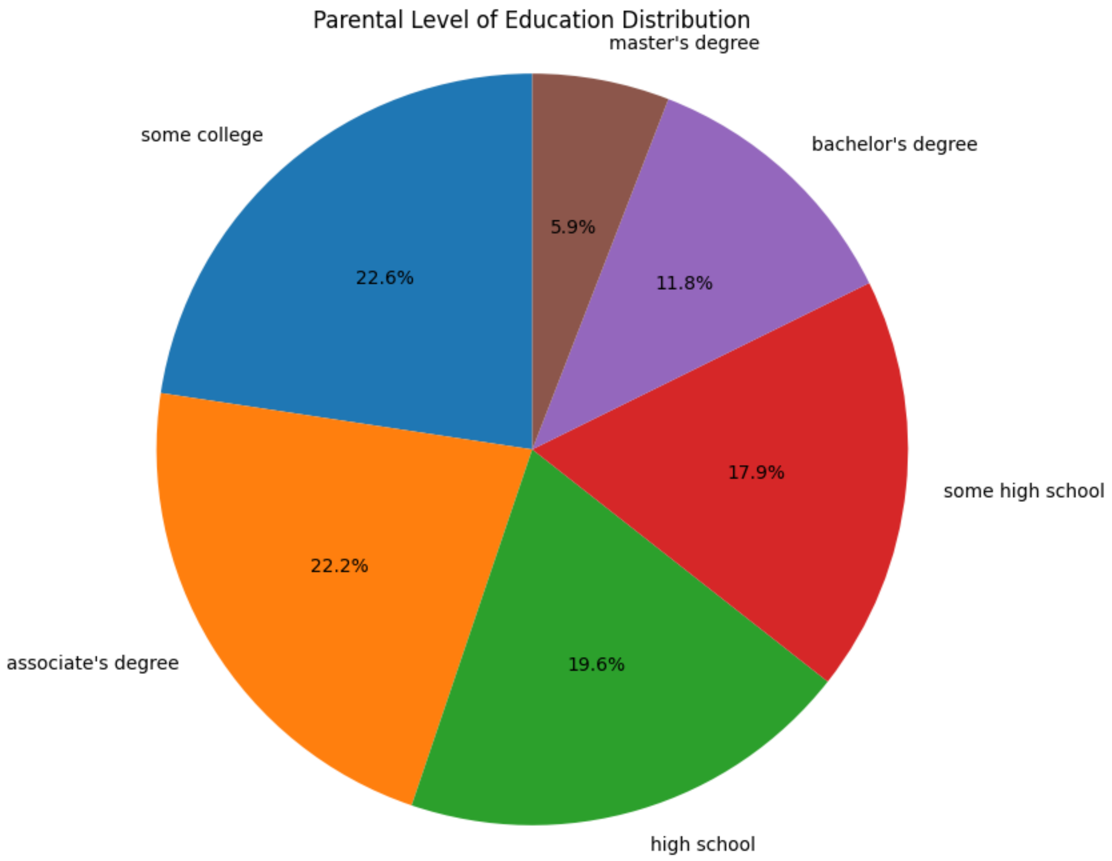
[215]:
```
<Axes: ylabel='Frequency'>
```



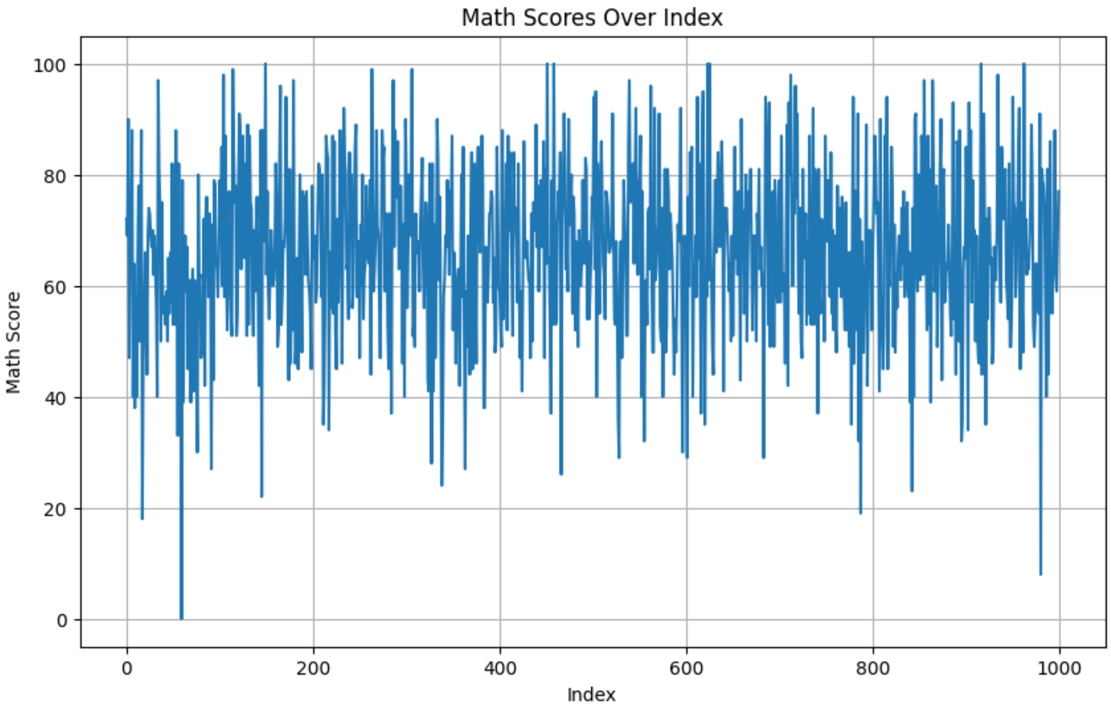
[216]:
```
parental_education_counts = df['parental level of education'].value_counts()
```

```
[217]: plt.figure(figsize=(8, 8))
       plt.pie(parental_education_counts, labels=parental_education_counts.index,␣
        ↪autopct='%1.1f%%', startangle=90)
       plt.title('Parental Level of Education Distribution')
       plt.axis('equal')
```

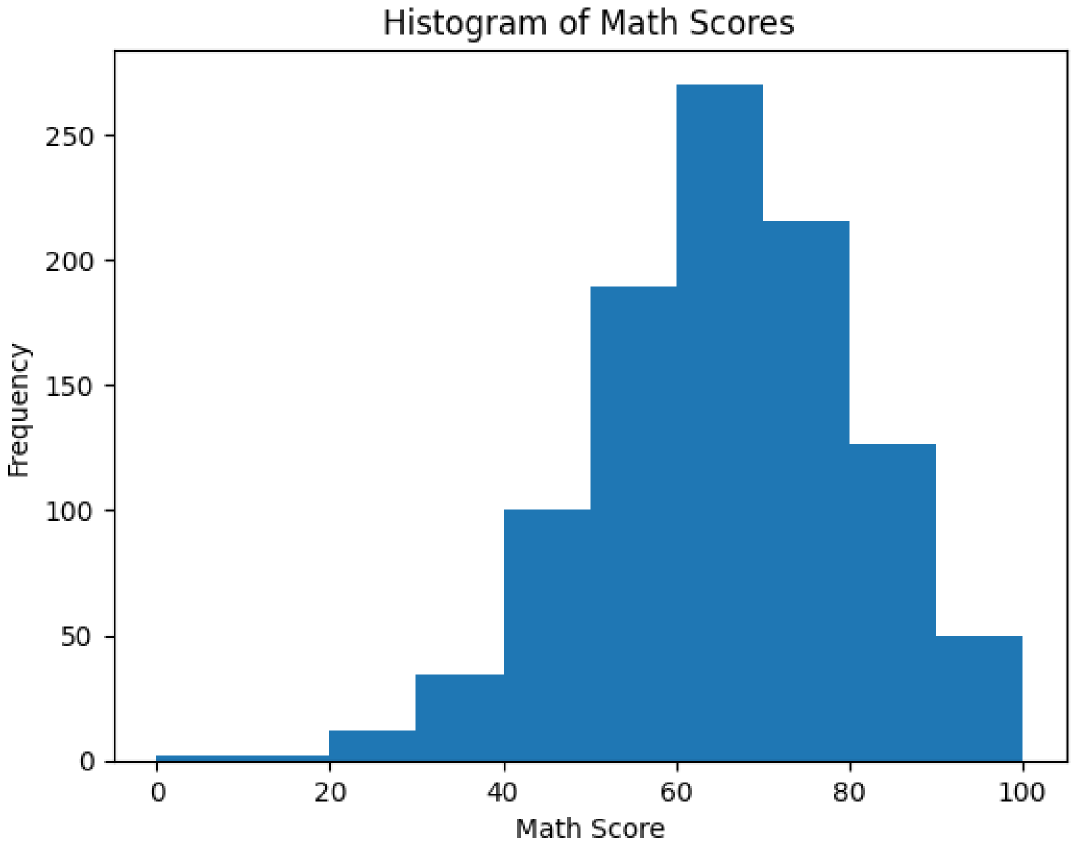
```
[217]: (-1.0999953405307346,
        1.0999988965017011,
        -1.0999930118454864,
        1.0999996672307375)
```



```
[218]: plt.figure(figsize=(10, 6))
       plt.plot(df.index, df['math score'])
       plt.xlabel('Index')
       plt.ylabel('Math Score')
       plt.title('Math Scores Over Index')
       plt.grid(True)

       plt.show()
```



[219]:
```
plt.hist(df['math score'], bins=10)  # Adjust the number of bins as needed
plt.xlabel('Math Score')
plt.ylabel('Frequency')
plt.title('Histogram of Math Scores')
plt.show()
```


Histogram of Math Scores
Frequency
0
50
100
150
200
250
Math Score
0
20
40
60
80
100