In [1]:

**import** numpy **as** np

**import** matplotlib.pyplot **as** plt

**%**matplotlib inline

In [2]:

**import** pandas **as** pd

In [3]:

df **=** pd.read\_csv(r"D:\College\TE\SEM-2\Practical\DSBDA\2\AcademicPerformance.csv")

In [4]:

print(df)

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | 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 |  |
| ... | ... | ... |  | ... | ... |  |
| 2235 | NaN | NaN |  | NaN | NaN |  |
| 2236 | NaN | NaN |  | NaN | NaN |  |
| 2237 | NaN | NaN |  | NaN | NaN |  |
| 2238 | NaN | NaN |  | NaN | NaN |  |
| 2239 | NaN | NaN |  | NaN | NaN |  |

test preparation course Year\_Birth math score reading score \

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 0 | none | | | 1970.0 | 72.0 | 72 |
| 1 | completed | | | 1961.0 | NaN | na |
| 2 | none | | | 1958.0 | 90.0 | 95 |
| 3 | none | | | 1967.0 | NaN | NaN |
| 4 | none | | | 1989.0 | 76.0 | 78 |
| ... | ... | | | ... | ... | ... |
| 2235 | NaN | | | NaN | NaN | NaN |
| 2236 | NaN | | | NaN | NaN | NaN |
| 2237 | NaN | | | NaN | NaN | NaN |
| 2238 | NaN | | | NaN | NaN | NaN |
| 2239 | NaN | | | NaN | NaN | NaN |
|  | writing | score | Dt\_Admission | College\_Fees | | |
| 0 |  | 74 | 6/16/14 | $84,835.00 | | |
| 1 |  | A | 6/15/14 | $57,091.00 | | |
| 2 |  | 93 | 5/13/14 | $67,267.00 | | |
| 3 |  | 44 | 05-11-2014 | $32,474.00 | | |
| 4  ... 2235 |  | 75  ...  NaN | 04-08-2014  ...  NaN | $21,474.00  ...  NaN | | |
| 2236 |  | NaN | NaN | NaN | | |
| 2237 |  | NaN | NaN | NaN | | |
| 2238 |  | NaN | NaN | NaN | | |
| 2239 |  | NaN | NaN | NaN | | |

[2240 rows x 11 columns]

In [5]:

print(df['math score'])

|  |  |  |
| --- | --- | --- |
| 0 | 72.0 |  |
| 1 | NaN |  |
| 2 | 90.0 |  |
| 3 | NaN |  |
| 4 | 76.0 |  |
|  | ... |  |
| 2235 | NaN |  |
| 2236 | NaN |  |
| 2237 | NaN |  |
| 2238 | NaN |  |
| 2239 | NaN |  |
| Name: | math score, Length: 2240, dtype: | float64 |

In [6]:

print(df['math score'].isnull())

1. False
2. True
3. False
4. True
5. False

...

2235 True

2236 True

2237 True

2238 True

2239 True

Name: math score, Length: 2240, dtype: bool

In [7]:

print(df['reading score'])

0 72

1 na

2 95

3 NaN

4 78

...

2235 NaN

2236 NaN

2237 NaN

2238 NaN

2239 NaN

Name: reading score, Length: 2240, dtype: object

In [8]:

print(df['reading score'].isnull())

|  |  |  |
| --- | --- | --- |
| 0 | False |  |
| 1 | False |  |
| 2 | False |  |
| 3 | True |  |
| 4 | False |  |
|  | ... |  |
| 2235 | True |  |
| 2236 | True |  |
| 2237 | True |  |
| 2238 | True |  |
| 2239 | True |  |
| Name: | reading | score, Length: 2240, dtype: bool |

In [9]:

missing\_values **=** ["n/a", "na", "--"]

df **=** pd.read\_csv(r"D:\College\TE\SEM-2\Practical\DSBDA\2\AcademicPerformance.csv", na\_v

In [10]:

print(df['reading score'])

|  |  |  |
| --- | --- | --- |
| 0 | 72.0 |  |
| 1 | NaN |  |
| 2 | 95.0 |  |
| 3 | NaN |  |
| 4 | 78.0 |  |
|  | ... |  |
| 2235 | NaN |  |
| 2236 | NaN |  |
| 2237 | NaN |  |
| 2238 | NaN |  |
| 2239 | NaN |  |
| Name: | reading score, Length: 2240, dtype: | float64 |

In [11]:

print(df['reading score'].isnull())

1. False
2. True
3. False
4. True
5. False

...

2235 True

2236 True

2237 True

2238 True

2239 True

Name: reading score, Length: 2240, dtype: bool

In [12]:

dataset **=** [11,41,20,3,101,55,68,97,99,6]

In [13]:

sorted(dataset)

Out[13]: [3, 6, 11, 20, 41, 55, 68, 97, 99, 101]

In [14]:

quantile1, quantile3 **=** np.percentile(dataset, [25,75])

In [15]:

print(quantile1, quantile3)

13.25 89.75

In [16]:

iqr\_value **=** (quantile3 **-** quantile1)

In [17]:

print(iqr\_value)

76.5

In [18]:

lower\_bound\_value **=** quantile1 **-** (1.5**\***iqr\_value)

In [19]:

upper\_bound\_value **=** quantile3 **+** (1.5**\***iqr\_value)

In [20]:

print(lower\_bound\_value, upper\_bound\_value)

-101.5 204.5

In [21]:

**from** datetime **import** date

df['age'] **=** date.today().year **-** df['Year\_Birth']

In [22]:

df['Year'] **=** pd.DatetimeIndex(df['Dt\_Admission']).year df['E\_L'] **=** date.today().year **-** df['Year']

In [23]:

df.head(5)

Out[23]:

**gender race/ethnicity**

**parental level of education**

**lunch**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **course** |  | | | | |
| standard | none | 1970.0 | 72.0 | 72.0 | 74 | 6 |
| standard | completed | 1961.0 | NaN | NaN | A | 6 |
| standard | none | 1958.0 | 90.0 | 95.0 | 93 | 5 |
| free/reduced | none | 1967.0 | NaN | NaN | 44 | 05-11 |
| standard | none | 1989.0 | 76.0 | 78.0 | 75 | 04-08 |

**test preparation**

**Year\_Birth**

**math score**

**reading score**

**writing Dt\_Admi score**

* 1. female group B bachelor's degree
  2. female group C some college
  3. female group B master's degree
  4. male group A associate's degree
  5. male group C some college

In [24]:

df['Fees$'] **=** df['College\_Fees'].str.replace(',', '').str.replace('$', '').str.replace( df['Fees\_M$'] **=** df['Fees$'].apply(**lambda** X:round(X**/**1000000))

In [25]:

df.head(5)

Out[25]:

**gender race/ethnicity**

**parental level of education**

**lunch**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **course** |  | | | | |
| standard | none | 1970.0 | 72.0 | 72.0 | 74 | 6 |
| standard | completed | 1961.0 | NaN | NaN | A | 6 |
| standard | none | 1958.0 | 90.0 | 95.0 | 93 | 5 |
| free/reduced | none | 1967.0 | NaN | NaN | 44 | 05-11 |
| standard | none | 1989.0 | 76.0 | 78.0 | 75 | 04-08 |

**test preparation**

**Year\_Birth**

**math score**

**reading score**

**writing Dt\_Admi score**

1. female group B bachelor's degree
2. female group C some college
3. female group B master's degree
4. male group A associate's degree
5. male group C some college

In [ ]: