2nd

import pandas as pd import numpy as np

import seaborn as sns

from scipy import stats

import matplotlib.pyplot as plt

from sklearn.preprocessing import MinMaxScaler, StandardScaler, RobustScaler

data=pd.read\_csv('/content/student-dataset (1).csv') df=pd.DataFrame(data)

df

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **id** | **name** | **nationality** | **city** | **latitude** | **longitude** | **gender** | **ethnic.group** |
| **0** | 0 | Kiana Lor | China | Suzhou | 31.31 | 120.62 | F | NaN |

Joshua Lonaker

|  |  |
| --- | --- |
| **1** | 1 |
| **2** | 2 |
| **3** | 3 |
| **4** | 4 |
| **...** | ... |
| **302** | 302 |
| **303** | 303 |
| **304** | 304 |
| **305** | 305 |
| **306** | 306 |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| United States  of America | Santa  Clarita | 34.39 | -118.54 | M | NaN |
| United States of America | Oakland | 37.80 | -122.27 | F | NaN |
| United States | Castro | 37.69 | -122.09 | F | NaN |

Dakota Blanco

Natasha Yarusso

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| of America | Valley  São José |  | | | |
| Brazil | dos | -23.18 | -45.88 | F | NaN |
|  | Campos |  |  |  |  |
| ... | ... | ... | ... | ... | .. |
| United States of America | Columbus | 39.96 | -83.00 | M | NaN |
| United States of America | Los Angeles | 34.05 | -118.24 | F | NaN |
| United States of America | Los Angeles | 34.05 | -118.24 | M | NaN |
| United States of America | Oakland | 37.80 | -122.27 | F | NaN |
| Canada | Toronto | 43.67 | -79.42 | M | NaN |

Brooke Cazares

...

Austin Haas

Madison Fithian

Zachary Mulvahill

Eliana Michelsen

Dane Whittemore

307 rows × 16 columns

Start coding or ge nerate with AI.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| df.describe() | **id** | **latitude** | **longitude** | **ethnic.group** | **age** | **english.grade ma** |
| **count** | 307.000000 | 307.000000 | 307.000000 | 0.0 | 307.000000 | 307.000000 30 |
| **mean** | 153.000000 | 32.863388 | -64.539121 | NaN | 21.964169 | 3.369707 |
| **std** | 88.767487 | 13.498582 | 81.249146 | NaN | 1.248013 | 0.538724 |
| **min** | 0.000000 | -33.450000 | -123.130000 | NaN | 19.000000 | 1.500000 |
| **25%** | 76.500000 | 32.720000 | -118.240000 | NaN | 21.000000 | 3.100000 |
| **50%** | 153.000000 | 34.390000 | -99.140000 | NaN | 22.000000 | 3.500000 |
| **75%** | 229.500000 | 38.960000 | -73.855000 | NaN | 23.000000 | 3.800000 |
| **max** | 306.000000 | 59.890000 | 139.750000 | NaN | 26.000000 | 4.000000 |
|  |  |  |  |  |  |  |
| df.head(10) |  |  |  |  |  |  |

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **id** | **name** | **nationality** | **city** | **latitude** | **longitude** | **gender** | **ethnic.group** | **ag** |

1. 0 Kiana Lor
2. 1 Joshua Lonaker
3. 2 Dakota Blanco

China Suzhou 31.31 120.62 F NaN 2

United States of America

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Santa Clarita | 34.39 | -118.54 | M | NaN | 2 |
| Oakland | 37.80 | -122.27 | F | NaN | 2 |

United States of America

1. 3 Natasha Yarusso

Brooke

United States of America

Castro Valley

São José

37.69 -122.09 F NaN 2

1. 4 Cazares Brazil

dos Campos

-23.18 -45.88 F NaN 2

1. 5 Rochelle Johnson
2. 6 Joey

United States of America

Indianapolis 39.77 -86.16 F NaN 2

Abreu China Shenyang 41.79 123.43 M NaN 2

1. 7 Preston Brazil São Paulo -23.47 -46.67 M NaN 2 Suarez
2. 8 Lee Philippines Manila 14.60 120.98 F NaN 2 Dong
3. 9 Maa'iz Turkey Istanbul 41.02 28.96 M NaN 2 al-Dia

Whittemore

# lity

States erica

|  |  |
| --- | --- |
| df.tail() | **id name nationa** |
| **302** | 302 Austin United Haas of Am |
| **303** | 303 Madison United  Fithian of Am |
| **304** | 304 Zachary United Mulvahill of Am |
| **305** | 305 Eliana United Michelsen of Am |
| **306** | 306 Dane Ca |

States erica

States erica

States erica

nada

# city

Columbus

|  |  |  |  |
| --- | --- | --- | --- |
| **latitude** | **longitude** | **gender** | **ethnic.group** |
| 39.96 | -83.00 | M | NaN |
| 34.05 | -118.24 | F | NaN |
| 34.05 | -118.24 | M | NaN |
| 37.80 | -122.27 | F | NaN |
| 43.67 | -79.42 | M | NaN |

Los Angeles

Los Angeles

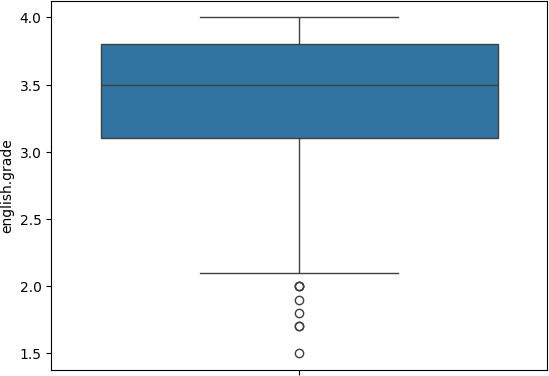
Oakland

Toronto

df.isnull().sum()

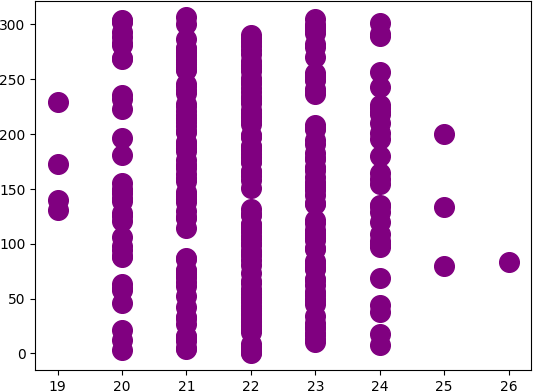
|  |  |
| --- | --- |
| id | 0 |
| name | 0 |
| nationality | 0 |
| city | 0 |
| latitude | 0 |
| longitude | 0 |
| gender | 0 |
| ethnic.group | 307 |
| age | 0 |
| english.grade | 0 |
| math.grade | 0 |
| sciences.grade | 0 |
| language.grade | 0 |
| portfolio.rating | 0 |
| coverletter.rating | 0 |
| refletter.rating | 0 |
| dtype: int64 |  |

sns.boxplot(df['english.grade'])

<Axes: ylabel='english.grade'>

x=df.age y=df.id

plt.scatter(x,y,c="Purple",s=200) plt.show()



data=df.age

m=np.mean(data) sd=np.std(data)

print('Mean is:',m)

print('Standard Deviation is:',sd)

 Mean is: 21.964169381107492

Standard Deviation is: 1.2459783686278594

threshold = 3 outlier = []

for i in data:

z = (i-m)/sd

if z > threshold:

outlier.append(i)

print('Outlier in dataset is', outlier) Outlier in dataset is [26]

data=np.sort(df.age)

Q1 = np.percentile(data, 25, interpolation = 'midpoint') Q2 = np.percentile(data, 50, interpolation = 'midpoint') Q3 = np.percentile(data, 75, interpolation = 'midpoint') print('Q1 25 percentile of the given data is, ', Q1)

print('Q1 50 percentile of the given data is, ', Q2) print('Q1 75 percentile of the given data is, ', Q3)

IQR = Q3 - Q1

print('Interquartile range is', IQR)

Q1 25 percentile of the given data is, 21.0 Q1 50 percentile of the given data is, 22.0 Q1 75 percentile of the given data is, 23.0 Interquartile range is 2.0

low\_lim = Q1 - 1.5 \* IQR up\_lim = Q3 + 1.5 \* IQR

print('low\_limit is', low\_lim) print('up\_limit is', up\_lim) outlier =[]

for x in data:

if ((x> up\_lim) or (x<low\_lim)): outlier.append(x)

print('Outlier in the dataset is', outlier)

low\_limit is 18.0

up\_limit is 26.0

Outlier in the dataset is []

df.drop(10, inplace = True) df

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **id** | **name** | **nationality** | **city** | **latitude** | **longitude** | **gender** | **ethnic.group** |
| **0** | 0 | Kiana Lor | China | Suzhou | 31.31 | 120.62 | F | NaN |

Joshua Lonaker

|  |  |
| --- | --- |
| **1** | 1 |
| **2** | 2 |
| **3** | 3 |
| **4** | 4 |
| **...** | ... |
| **302** | 302 |
| **303** | 303 |
| **304** | 304 |
| **305** | 305 |
| **306** | 306 |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| United States  of America | Santa  Clarita | 34.39 | -118.54 | M | NaN |
| United States of America | Oakland | 37.80 | -122.27 | F | NaN |
| United States | Castro | 37.69 | -122.09 | F | NaN |

Dakota Blanco

Natasha Yarusso

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| of America | Valley  São José |  | | | |
| Brazil | dos | -23.18 | -45.88 | F | NaN |
|  | Campos |  |  |  |  |
| ... | ... | ... | ... | ... | .. |
| United States of America | Columbus | 39.96 | -83.00 | M | NaN |
| United States of America | Los Angeles | 34.05 | -118.24 | F | NaN |
| United States of America | Los Angeles | 34.05 | -118.24 | M | NaN |
| United States of America | Oakland | 37.80 | -122.27 | F | NaN |
| Canada | Toronto | 43.67 | -79.42 | M | NaN |

Brooke Cazares

...

Austin Haas

Madison Fithian

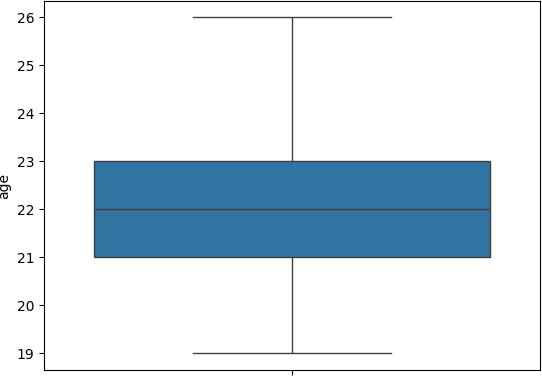
Zachary Mulvahill

Eliana Michelsen

Dane Whittemore

306 rows × 16 columns

sns.boxplot(df['age'])

<Axes: ylabel='age'>

df['log\_refletter.rating'] = np.log2(df['refletter.rating']) df

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **y latitude** | **longitude** | **gender** | **ethnic.group** | **age** | **english.grade** | **math.grade** | **sciences.** |
| u 31.31 | 120.62 | F | NaN | 22 | 3.5 | 3.7 |  |
| a 34.39 | -118.54 | M | NaN | 22 | 2.9 | 3.2 |  |
| d 37.80 | -122.27 | F | NaN | 22 | 3.9 | 3.8 |  |
| o 37.69 | -122.09 | F | NaN | 20 | 3.3 | 2.8 |  |
| é  s -23.18  s | -45.88 | F | NaN | 21 | 3.7 | 2.6 |  |
| .. ... | ... | ... | ... | ... | ... | ... |  |
| s 39.96 | -83.00 | M | NaN | 20 | 3.6 | 3.7 |  |
| s 34.05 | -118.24 | F | NaN | 20 | 3.6 | 3.9 |  |
| s 34.05 | -118.24 | M | NaN | 20 | 3.2 | 3.4 |  |
| d 37.80 | -122.27 | F | NaN | 23 | 3.0 | 2.8 |  |
| o 43.67 | -79.42 | M | NaN | 21 | 3.8 | 3.2 |  |

a

y

s

s

New Section