p0akzvvgq

January 22, 2025

[14]:
```
print(df. describe())
```

```
              Age  Admission Test Score  High School Percentage
count  147.000000             146.000000              146.000000
mean    19.680272              77.657534               75.684726
std      4.540512              16.855343               17.368014
min     -1.000000              -5.000000              -10.000000
25%     18.000000              68.250000               65.052500
50%     20.000000              79.000000               77.545000
75%     22.000000              89.000000               88.312500
max     24.000000             150.000000              110.500000
```

[7]:
```
import pandas as pd
df = pd.read_csv('student_admission_record_dirty.csv')
```

[15]:
```
print(df.head(10))
```

```
      Name   Age  Gender  Admission Test Score  High School Percentage  \
0  Shehroz  24.0  Female                   50.0                   68.90
1    Waqar  21.0  Female                   99.0                   60.73
2   Bushra  17.0    Male                   89.0                     NaN
3    Aliya  17.0    Male                   55.0                   85.29
4    Bilal  20.0    Male                   65.0                   61.13
5  Murtaza  23.0  Female                    NaN                     NaN
6     Asad  18.0    Male                    NaN                   97.31
7    Rabia  20.0  Female                   82.0                   55.67
8   Rohail  17.0    Male                   64.0                     NaN
9   Kamran  18.0    Male                   53.0                   98.98

        City Admission Status
0     Quetta         Rejected
1    Karachi              NaN
2  Islamabad         Accepted
3    Karachi         Rejected
4     Lahore              NaN
5  Islamabad         Accepted
6     Multan         Accepted
7     Lahore         Accepted
```

```
8    Karachi         Accepted
9     Multan         Rejected
```

[16]:
```
print(df.isnull())
```

```
      Name    Age  Gender  Admission Test Score  High School Percentage  \
0    False  False   False                 False                   False
1    False  False   False                 False                   False
2    False  False   False                 False                    True
3    False  False   False                 False                   False
4    False  False   False                 False                   False
..     …    …     …                     …                       …
152  False  False   False                 False                   False
153  False  False   False                 False                   False
154  False  False   False                 False                   False
155  False  False   False                 False                   False
156  False  False   False                 False                   False

      City  Admission Status
0    False             False
1    False              True
2    False             False
3    False             False
4    False              True
..     …                …
152  False             False
153  False             False
154  False             False
155  False             False
156  False             False

[157 rows x 7 columns]
```

[17]:
```
print(df.notnull())
```

```
     Name   Age  Gender  Admission Test Score  High School Percentage  City  \
0    True  True    True                  True                    True  True
1    True  True    True                  True                    True  True
2    True  True    True                  True                   False  True
3    True  True    True                  True                    True  True
4    True  True    True                  True                    True  True
..    …    …      …                     …                       …    …
152  True  True    True                  True                    True  True
153  True  True    True                  True                    True  True
154  True  True    True                  True                    True  True
155  True  True    True                  True                    True  True
156  True  True    True                  True                    True  True
```

```
     Admission Status
0                True
1               False
2                True
3                True
4               False
..                …
152              True
153              True
154              True
155              True
156              True

[157 rows x 7 columns]
```

[18]:
```
print(df.dropna())
```

```
        Name   Age  Gender  Admission Test Score  High School Percentage  \
0    Shehroz  24.0  Female                  50.0                   68.90
3      Aliya  17.0    Male                  55.0                   85.29
7      Rabia  20.0  Female                  82.0                   55.67
9     Kamran  18.0    Male                  53.0                   98.98
10    Shafiq  17.0    Male                  78.0                  -10.00
..       …   …      …                     …                      …
152      Ali  19.0  Female                  85.0                   78.09
153    Bilal  17.0  Female                  81.0                   84.40
154   Fatima  21.0  Female                  98.0                   50.86
155   Shoaib  -1.0    Male                  91.0                   80.12
156     Maaz  17.0    Male                  88.0                   86.85

          City Admission Status
0       Quetta         Rejected
3      Karachi         Rejected
7       Lahore         Accepted
9       Multan         Rejected
10      Quetta         Rejected
..         …               …
152     Quetta         Accepted
153  Islamabad         Rejected
154     Multan         Accepted
155     Quetta         Accepted
156     Lahore         Accepted

[100 rows x 7 columns]
```

[19]:
```
print(df.replace())
```

```
        Name   Age  Gender  Admission Test Score  High School Percentage  \
```

```
0    Shehroz  24.0  Female                  50.0                    68.90
1      Waqar  21.0  Female                  99.0                    60.73
2     Bushra  17.0    Male                  89.0                    60.73
3      Aliya  17.0    Male                  55.0                    85.29
4      Bilal  20.0    Male                  65.0                    61.13
..       …   …     …                     …                       …
152      Ali  19.0  Female                  85.0                    78.09
153    Bilal  17.0  Female                  81.0                    84.40
154   Fatima  21.0  Female                  98.0                    50.86
155   Shoaib  -1.0    Male                  91.0                    80.12
156     Maaz  17.0    Male                  88.0                    86.85

          City Admission Status
0       Quetta         Rejected
1      Karachi         Rejected
2    Islamabad         Accepted
3      Karachi         Rejected
4       Lahore         Rejected
..         …              …
152     Quetta         Accepted
153  Islamabad         Rejected
154     Multan         Accepted
155     Quetta         Accepted
156     Lahore         Accepted

[157 rows x 7 columns]
```

```
C:\Users\Dell\AppData\Local\Temp\ipykernel_10720\365866856.py:1: FutureWarning:
DataFrame.replace without 'value' and with non-dict-like 'to_replace' is
deprecated and will raise in a future version. Explicitly specify the new values
instead.
  print(df.replace())
```

[42]:
```
print(df.interpolate())
```

```
        Name   Age  Gender  Admission Test Score  High School Percentage  \
0    Shehroz  24.0  Female                  50.0                    68.90
1      Waqar  21.0  Female                  99.0                    60.73
2     Bushra  17.0    Male                  89.0                    73.01
3      Aliya  17.0    Male                  55.0                    85.29
4      Bilal  20.0    Male                  65.0                    61.13
..       …   …     …                     …                       …
152      Ali  19.0  Female                  85.0                    78.09
153    Bilal  17.0  Female                  81.0                    84.40
154   Fatima  21.0  Female                  98.0                    50.86
155   Shoaib  -1.0    Male                  91.0                    80.12
156     Maaz  17.0    Male                  88.0                    86.85
```

```
          City Admission Status
0       Quetta         Rejected
1      Karachi              NaN
2    Islamabad         Accepted
3      Karachi         Rejected
4       Lahore              NaN
..         …               …
152     Quetta         Accepted
153  Islamabad         Rejected
154     Multan         Accepted
155     Quetta         Accepted
156     Lahore         Accepted

[157 rows x 7 columns]
```

```
C:\Users\Dell\AppData\Local\Temp\ipykernel_10720\796038140.py:1: FutureWarning:
DataFrame.interpolate with object dtype is deprecated and will raise in a future
version. Call obj.infer_objects(copy=False) before interpolating instead.
  print(df.interpolate())
```

[4]:
```
import pandas as pd
import seaborn as sns
```

[6]:
```
df_boston = pd.read_csv('student_admission_record_dirty.csv')
```

[27]:
```
print(df_boston.head())
```

```
      Name   Age  Gender  Admission Test Score  High School Percentage  \
0  Shehroz  24.0  Female                   50.0                   68.90
1    Waqar  21.0  Female                   99.0                   60.73
2   Bushra  17.0    Male                   89.0                     NaN
3    Aliya  17.0    Male                   55.0                   85.29
4    Bilal  20.0    Male                   65.0                   61.13

        City Admission Status
0     Quetta         Rejected
1    Karachi              NaN
2  Islamabad         Accepted
3    Karachi         Rejected
4     Lahore              NaN
```

[29]:
```
print(df_boston.columns)
```

```
Index(['Name', 'Age', 'Gender', 'Admission Test Score',
       'High School Percentage', 'City', 'Admission Status'],
      dtype='object')
```

[40]:
```
import seaborn as sns
```

[8]:
```
df_boston = pd.read_csv('student_admission_record_dirty.csv')
```

[9]:
```
df_boston = pd.read_csv('student_admission_record_dirty.csv')
print(df_boston.head())
```

```
      Name   Age  Gender  Admission Test Score  High School Percentage  \
0  Shehroz  24.0  Female                  50.0                   68.90
1    Waqar  21.0  Female                  99.0                   60.73
2   Bushra  17.0    Male                  89.0                     NaN
3    Aliya  17.0    Male                  55.0                   85.29
4    Bilal  20.0    Male                  65.0                   61.13

        City Admission Status
0     Quetta         Rejected
1    Karachi              NaN
2  Islamabad         Accepted
3    Karachi         Rejected
4     Lahore              NaN
```

[50]:
```
df_boston.columns = df_boston.columns.str.strip()
```

[51]:
```
print(df_boston.columns)
```

```
Index(['DIS'], dtype='object')
```

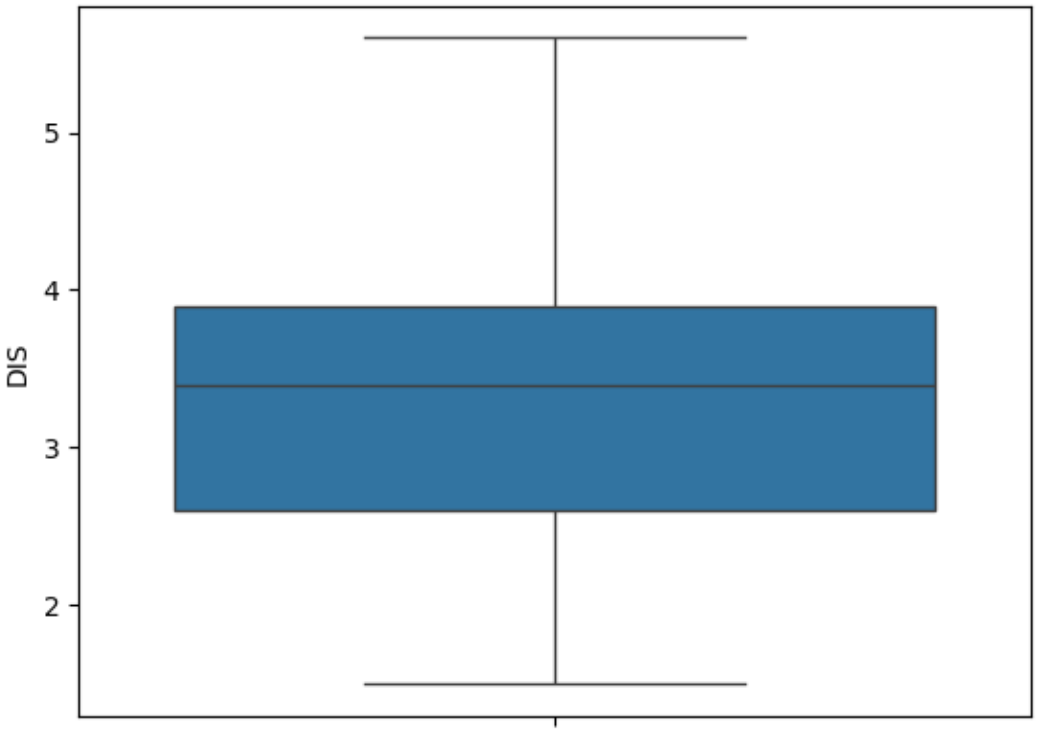
[53]:
```
df_boston.columns = df_boston.columns.str.strip()

print(df_boston.columns)
```

```
Index(['DIS'], dtype='object')
```

[55]:
```
import seaborn as sns
sns.boxplot(df_boston['DIS'])
```

[55]:
```
<Axes: ylabel='DIS'>
```



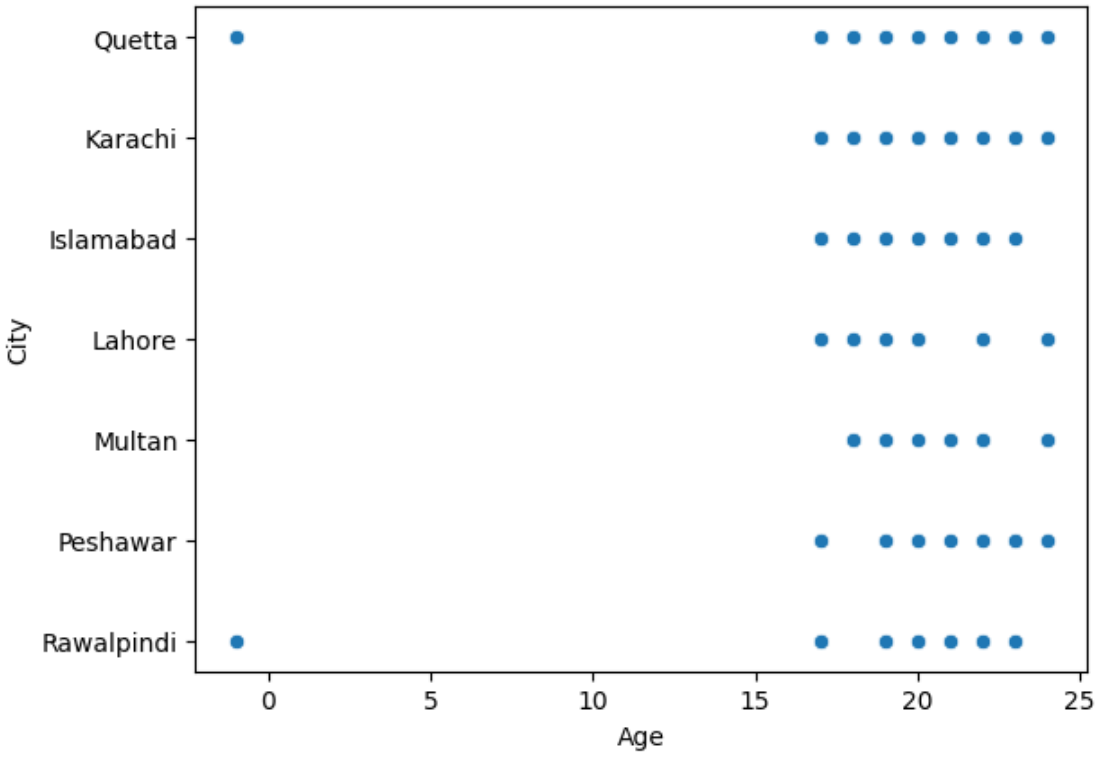
[11]:
```
import seaborn as sns
```

[12]:
```
import matplotlib.pyplot as plt
```

[15]:
```
df_boston = pd.read_csv('student_admission_record_dirty.csv')
```

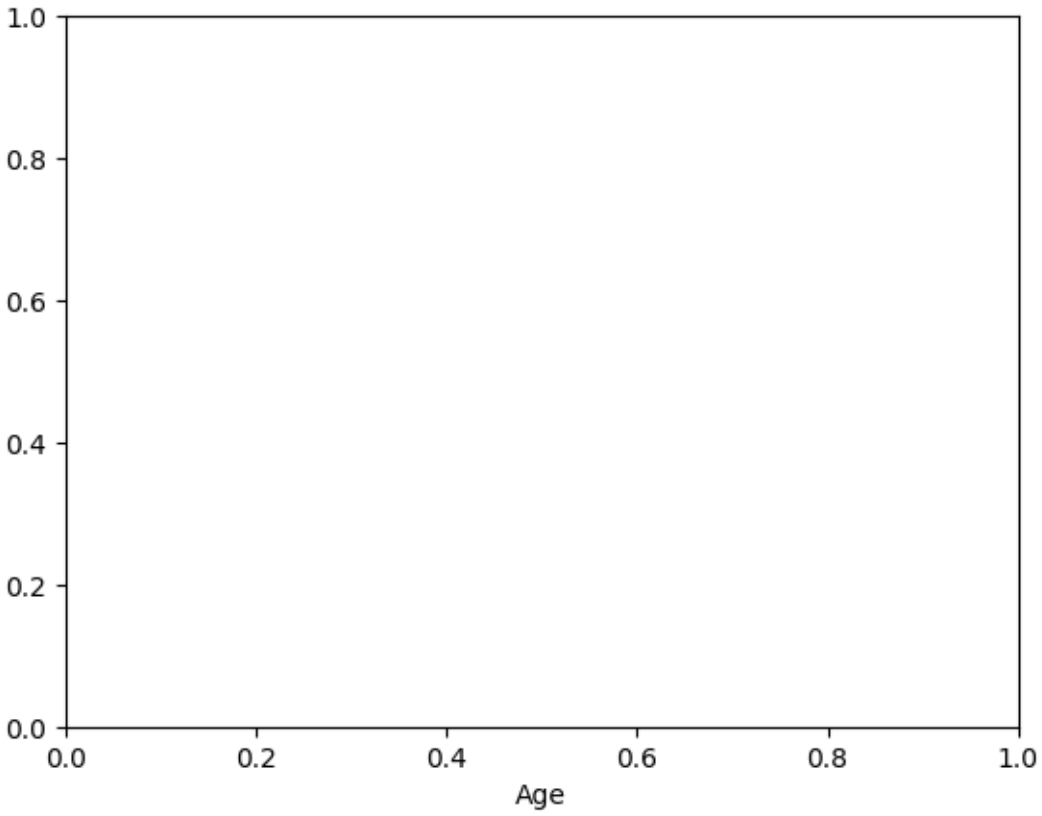
[17]:
```
sns.scatterplot(x=df_boston['Age'], y=df_boston['City'])
```

[17]:
```
<Axes: xlabel='Age', ylabel='City'>
```



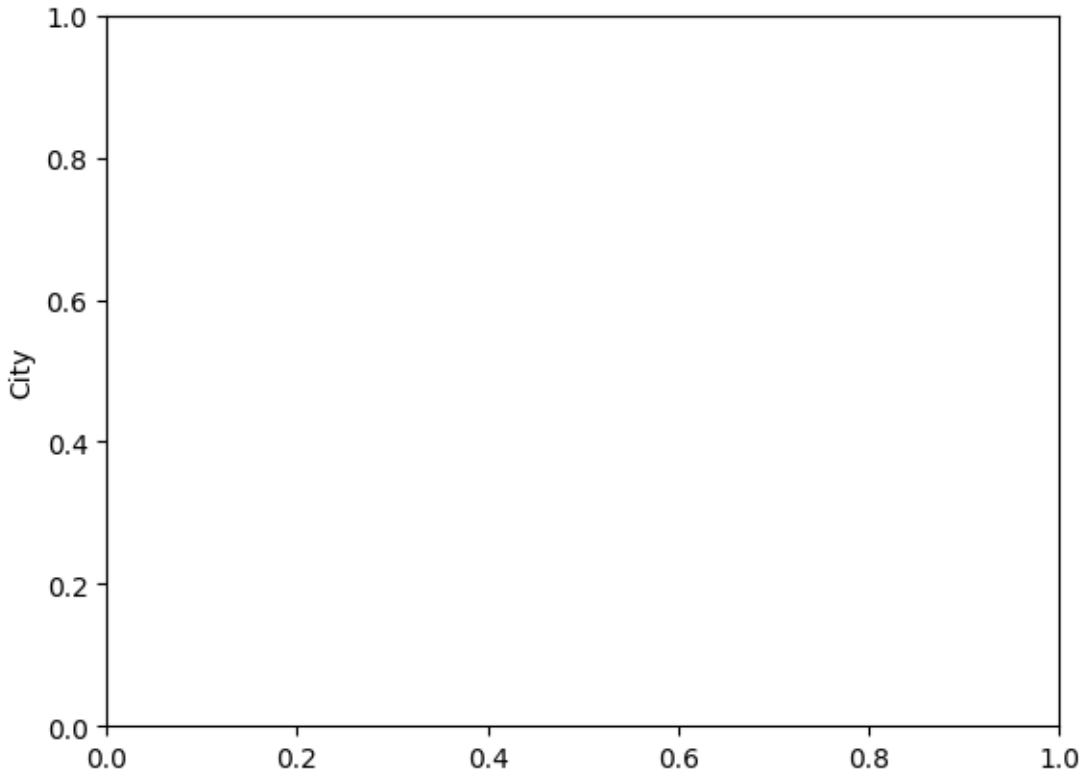
[21]:
```
plt.xlabel('Age')
```

[21]:
```
Text(0.5, 0, 'Age')
```



[22]:
```
plt.ylabel('City')

plt.show()
```



[23]:
```
import pandas as pd
from scipy import stats
```

[24]:
```
df = pd.read_csv('student_admission_record_dirty.csv')
```

[25]:
```
numeric_columns = df.select_dtypes(include='number').columns
```

[26]:
```
z_scores = stats.zscore(df[numeric_columns])
```

[27]:
```
z_scores_df = pd.DataFrame(z_scores, columns=numeric_columns)
print(z_scores_df)
```

```
     Age  Admission Test Score  High School Percentage
0    NaN                   NaN                     NaN
1    NaN                   NaN                     NaN
2    NaN                   NaN                     NaN
3    NaN                   NaN                     NaN
4    NaN                   NaN                     NaN
..   …                     …                       …
152  NaN                   NaN                     NaN
153  NaN                   NaN                     NaN
154  NaN                   NaN                     NaN
```

```
155  NaN                  NaN                  NaN
156  NaN                  NaN                  NaN

[157 rows x 3 columns]
```

[28]:
```
import pandas as pd
df = pd.read_csv('student_admission_record_dirty.csv')
print(df.head())
```

```
      Name   Age  Gender  Admission Test Score  High School Percentage  \
0  Shehroz  24.0  Female                  50.0                   68.90
1    Waqar  21.0  Female                  99.0                   60.73
2   Bushra  17.0    Male                  89.0                     NaN
3    Aliya  17.0    Male                  55.0                   85.29
4    Bilal  20.0    Male                  65.0                   61.13

        City Admission Status
0     Quetta         Rejected
1    Karachi              NaN
2  Islamabad         Accepted
3    Karachi         Rejected
4     Lahore              NaN
```

[29]:
```
df = df.dropna()
```

[30]:
```
df = df.fillna(0)
```

[31]:
```
from scipy import stats

numeric_columns = df.select_dtypes(include='number').columns

z_scores = stats.zscore(df[numeric_columns])

z_scores_df = pd.DataFrame(z_scores, columns=numeric_columns)
print(z_scores_df)
```

```
         Age  Admission Test Score  High School Percentage
0   0.965581             -1.625981               -0.439615
1  -0.643721             -1.320001                0.522809
2   0.045980              0.332295               -1.216483
3  -0.413820             -1.442393                1.326689
4  -0.643721              0.087510               -5.072637
..       …                     …                       …
95 -0.183920              0.515883                0.100024
96 -0.643721              0.271099                0.470548
97  0.275880              1.311433               -1.498927
98 -4.781925              0.883060                0.219226
99 -0.643721              0.699472                0.614413
```

```
[100 rows x 3 columns]
```

[32]:
```
for col in numeric_columns:
    Q1 = df[col].quantile(0.25)
    Q3 = df[col].quantile(0.75)
    IQR = Q3 - Q1
    print(f"{col} - Q1: {Q1}, Q3: {Q3}, IQR: {IQR}")
```

```
Age - Q1: 18.0, Q3: 22.0, IQR: 4.0
Admission Test Score - Q1: 68.75, Q3: 89.5, IQR: 20.75
High School Percentage - Q1: 66.75, Q3: 90.125, IQR: 23.375
```

[33]:
```
for col in numeric_columns:
    Q1 = df[col].quantile(0.25)
    Q3 = df[col].quantile(0.75)
    IQR = Q3 - Q1


    lower_bound = Q1 - 1.5 * IQR
    upper_bound = Q3 + 1.5 * IQR

    print(f"{col} - Outliers: Lower Bound: {lower_bound}, Upper Bound:␣
 ↪{upper_bound}")

    df_clean = df[(df[col] >= lower_bound) & (df[col] <= upper_bound)]
    print(f"Cleaned {col} data without outliers:\n{df_clean[col].head()}")
```

```
Age - Outliers: Lower Bound: 12.0, Upper Bound: 28.0
Cleaned Age data without outliers:
0     24.0
3     17.0
7     20.0
9     18.0
10    17.0
Name: Age, dtype: float64
Admission Test Score - Outliers: Lower Bound: 37.625, Upper Bound: 120.625
Cleaned Admission Test Score data without outliers:
0     50.0
3     55.0
7     82.0
9     53.0
10    78.0
Name: Admission Test Score, dtype: float64
High School Percentage - Outliers: Lower Bound: 31.6875, Upper Bound: 125.1875
Cleaned High School Percentage data without outliers:
0     68.90
3     85.29
```

```
7     55.67
9     98.98
13    79.03
Name: High School Percentage, dtype: float64
```

[ ]: