Big Cities Health

Description

Illustrates health status of 26 of the nation’s largest and most urban cities.

Summary

This dataset illustrates health status of 26 of the nation’s largest and most urban citie

Attribution:U.S. Centers for Disease Control and Prevention

Source: Big Cities Health Inventory Data

Aim

To perform Data cleaning on the Data set

Importing Libraries

```
#Python
import numpy as np
import pandas as pd

#Visualization
import matplotlib.pyplot as plt
import seaborn as sns

#Data cleaning
import missingno
```

Importing Dataset

The dataset could be found using the following link

```
health_df=pd.read_csv('https://query.data.world/s/nnsiif4n3gg3oj6o6efdiuqfr2eded')
health_df.head()
```

	Indicator Category	Indicator	Year	Gender	Race/ Ethnicity	Value	Place	BCHC Requested Methodology	D
0	HIV/AIDS	AIDS Diagnoses Rate (Per 100,000 people)	2013	Both	All	30.4	Atlanta (Fulton County), GA	AIDS cases diagnosed in 2012, 2013, 2014 (as a...	
1	HIV/AIDS	AIDS Diagnoses Rate (Per 100,000 people)	2012	Both	All	39.6	Atlanta (Fulton County), GA	AIDS cases diagnosed in 2012, 2013, 2014 (as a...	

```
health_df.sample()
```

	Indicator Category	Indicator	Year	Gender	Race/ Ethnicity	Value	Place	BCHC Requested Methodology
731	Life Expectancy and Death	Life Expectancy	2008-	Male	All	77.1	Boston,	Three most recent years

Column removal

Here all the columns seems to be essential, so there is no need to remove any of them

Changing Index

```
health_df1=health_df
health_df1['Identifier']=list(range(1,13513))
health_df1.head()
```

	Indicator Category	Indicator	Year	Gender	Race/ Ethnicity	Value	Place	BCHC Requested Methodology	
0	HIV/AIDS	AIDS Diagnoses Rate (Per 100,000 people)	2013	Both	All	30.4	Atlanta (Fulton County), GA	AIDS cases diagnosed in 2012, 2013, 2014 (as a...	D
1	HIV/AIDS	AIDS Diagnoses Rate (Per 100,000 people)	2012	Both	All	39.6	Atlanta (Fulton County), GA	AIDS cases diagnosed in 2012, 2013, 2014 (as a...	D

```
health_df1.set_index("Identifier",inplace=True)
health_df1.head()
```

Identifier	Indicator Category	Indicator	Year	Gender	Race/ Ethnicity	Value	Place	Re Meth
1	HIV/AIDS	AIDS Diagnoses Rate (Per 100,000 people)	2013	Both	All	30.4	Atlanta (Fulton County), GA	AI diag 20 201
2	HIV/AIDS	AIDS Diagnoses Rate (Per 100,000 people)	2012	Both	All	39.6	Atlanta (Fulton County), GA	AI diag 20 201
3	HIV/AIDS	AIDS Diagnoses Rate (Per 100,000	2011	Both	All	41.7	Atlanta (Fulton County), GA	AI diag 20 201

Tidying up fields

```
health_df1["Indicator Category"].value_counts()
```

```
HIV/AIDS                                  2177
Injury and Violence                       1916
Nutrition, Physical Activity, & Obesity   1841
Infectious Disease                        1486
Cancer                                    1432
Maternal and Child Health                 1323
Behavioral Health/Substance Abuse          983
```

```
Food Safety                                    874
Life Expectancy and Death Rate (Overall)     544
Demographics                                   504
Tobacco                                        432
Name: Indicator Category, dtype: int64
```

```
health_df1["Year"].value_counts()
```

```
2012         3950
2013         3657
2011         3501
2010         1357
2014         1020
2008-2012       7
2011-2012       6
2015            6
2007-2012       3
2003-2012       2
2011-2013       1
2003-2013       1
2004-2013       1
Name: Year, dtype: int64
```

```
health_df1["Gender"].value_counts()
```

```
Both      9409
Female    2423
Male      1680
Name: Gender, dtype: int64
```

```
health_df1["Race/ Ethnicity"].value_counts()
```

```
All                              5757
White                            1914
Black                            1869
Hispanic                         1688
Asian/PI                         1015
Other                             570
Native American                   371
Multiracial                       270
American Indian/Alaska Native      58
Name: Race/ Ethnicity, dtype: int64
```

```
No value seems to be out of context
```

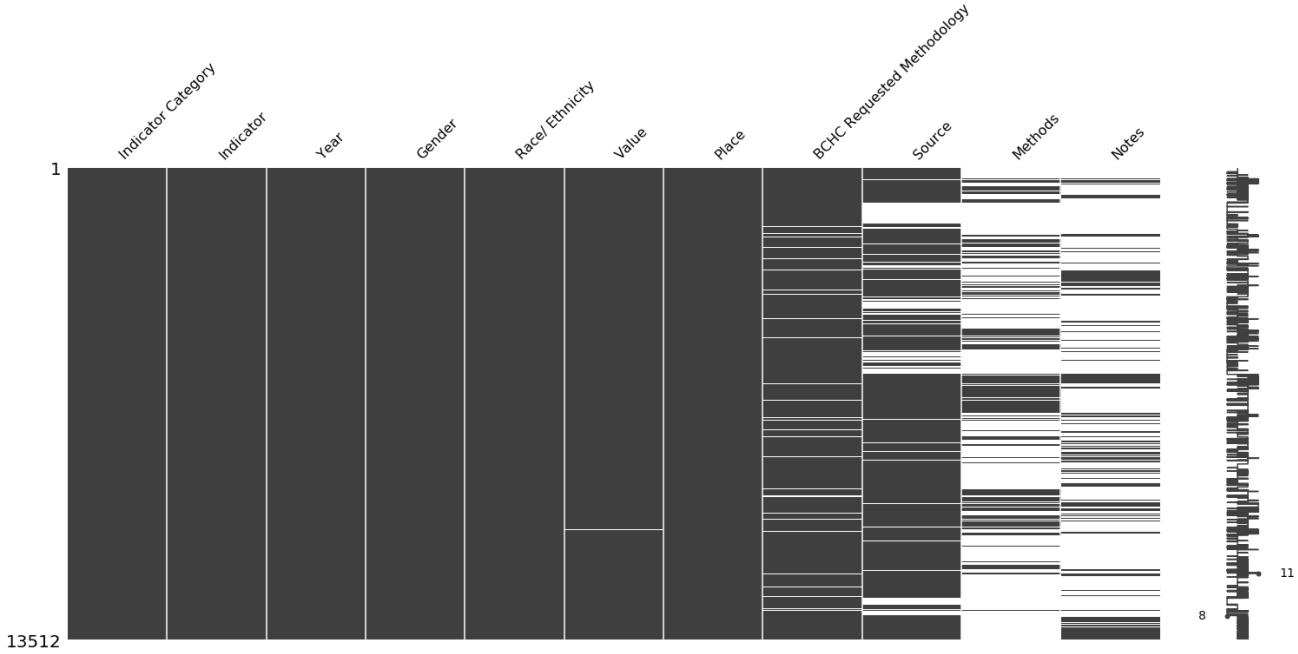
Missing values and Treatment

```
health_df1.isnull().sum()
```

```
Indicator Category              0
Indicator                       0
Year                            0
```

```
Gender                          0
Race/ Ethnicity                 0
Value                          13
Place                           0
BCHC Requested Methodology    508
Source                       2290
Methods                      9280
Notes                        9971
dtype: int64
```

```
missingno.matrix(health_df1)
plt.show()
```



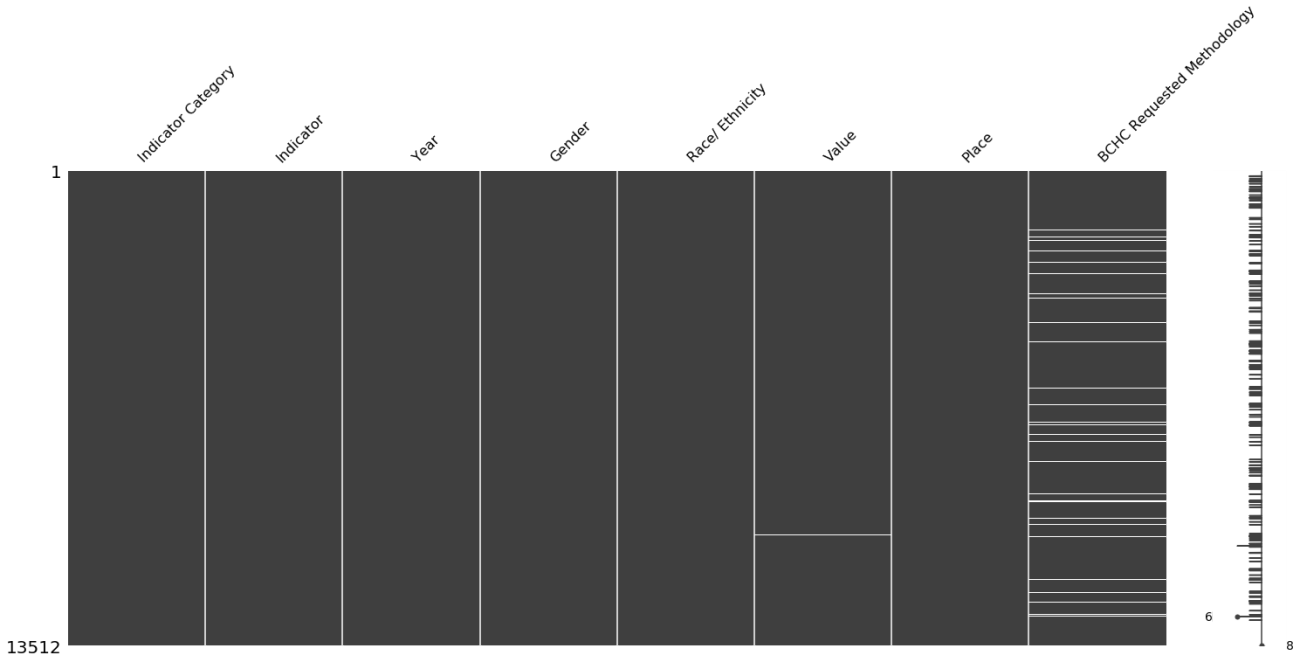
We can clearly see that columns - Notes and Methods are almost null

Also column - Source can also be considered having many null values

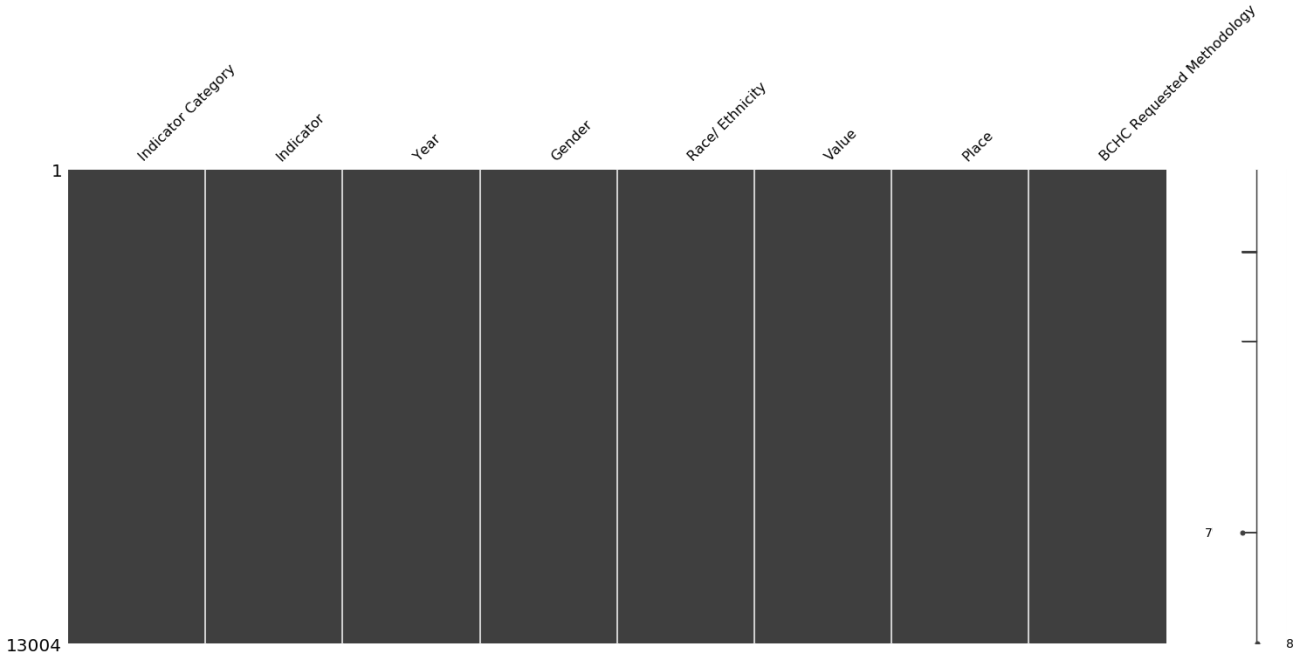
Looking at column - BCHC Requested Methodology, it surely do contain null values, but we

Column Value contains very few null value, but being numerical value, we can replace it w

```
health_df2=health_df1.iloc[:,0:8]
missingno.matrix(health_df2)
plt.show()
```




```
health_df3=health_df2[health_df2["BCHC Requested Methodology"].isnull()==False]
missingno.matrix(health_df3)
plt.show()
```




```
health_df3.isnull().sum()
```

```
Indicator Category            0
Indicator                     0
Year                          0
Gender                        0
Race/ Ethnicity               0
Value                        10
Place                         0
BCHC Requested Methodology    0
dtype: int64
```

Value still contains some null values, we can replace it with 1st business moment

```
health_df3.describe()
```

	Value
count	12994.000000
mean	96.447853
std	286.261235
min	0.000000
25%	7.000000
50%	15.900000
75%	44.900000
max	4199.600000

```
health_df4=health_df3.fillna(15.9)
health_df4.isnull().sum()
```

```
Indicator Category           0
Indicator                    0
Year                         0
Gender                       0
Race/ Ethnicity              0
Value                        0
Place                        0
BCHC Requested Methodology   0
dtype: int64
```

With our data cleaned we will now save it

```
health_df4.to_csv("Big_Cities_Health__cleaned.csv")
```