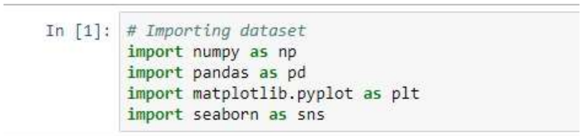
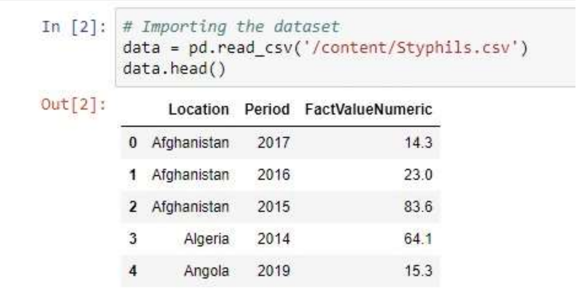
Sexually Transmitted Infections Dataset

In 38 of the 78 reporting nations, at least 1% of prenatal care participants tested positive for syphilis in 2019. An average of 3.2 percent (range 1.1 percent to 10.9 percent) of prenatal care attendance tested positive for syphilis in these 78 reporting countries. Prematurity, low birthweight, neonatal death, and infections in infants are all caused by syphilis in pregnancy, which is the world's second biggest cause of stillbirth. A simple and inexpensive fast test, followed by benzathine penicillin therapy, can avert these negative effects.

This dataset id gotten from World Health Organization data storage https://www.who.int/data/gho/data/themes/sexually-transmitted-infections



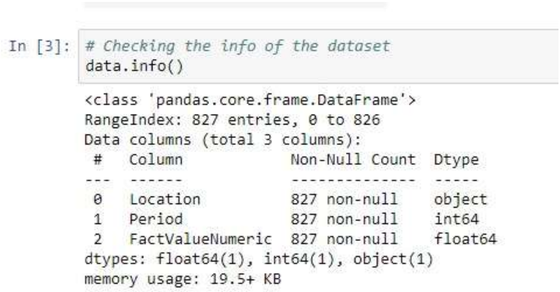
```
In [1]: # Importing dataset
        import numpy as np
        import pandas as pd
        import matplotlib.pyplot as plt
        import seaborn as sns
```



```
In [2]: # Importing the dataset
        data = pd.read_csv('/content/Styphils.csv')
        data.head()
Out[2]:
```

	Location	Period	FactValueNumeric
0	Afghanistan	2017	14.3
1	Afghanistan	2016	23.0
2	Afghanistan	2015	83.6
3	Algeria	2014	64.1
4	Angola	2019	15.3

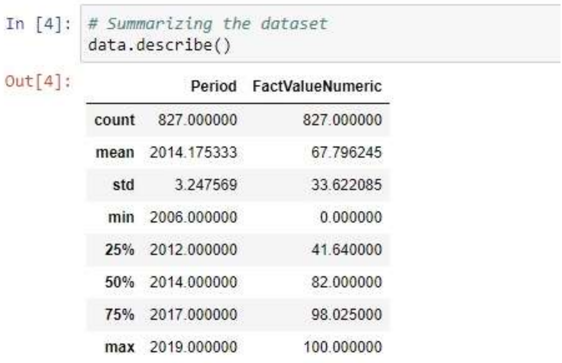
The first cell illustrates how we imported the dataset and printed the first five rows by calling the function head, and the second cell shows how we imported the dataset and printed the first five rows by calling the function head.



```
In [3]: # Checking the info of the dataset
        data.info()

        <class 'pandas.core.frame.DataFrame'>
        RangeIndex: 827 entries, 0 to 826
        Data columns (total 3 columns):
         #   Column            Non-Null Count  Dtype
        ---  ------            --------------  -----
         0   Location          827 non-null    object
         1   Period            827 non-null    int64
         2   FactValueNumeric  827 non-null    float64
        dtypes: float64(1), int64(1), object(1)
        memory usage: 19.5+ KB
```

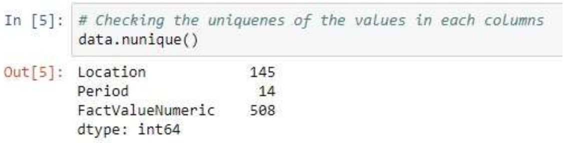
The graphic above depicts the dataset's information, summarizing the entire dataset's information by displaying the number of entries (row) of 827 and the number of columns of 3. It also reveals that there is one object, one int64, and one float datatype.



```
In [4]: # Summarizing the dataset
        data.describe()
Out[4]:
```

	Period	FactValueNumeric
count	827.000000	827.000000
mean	2014.175333	67.796245
std	3.247569	33.622085
min	2006.000000	0.000000
25%	2012.000000	41.640000
50%	2014.000000	82.000000
75%	2017.000000	98.025000
max	2019.000000	100.000000

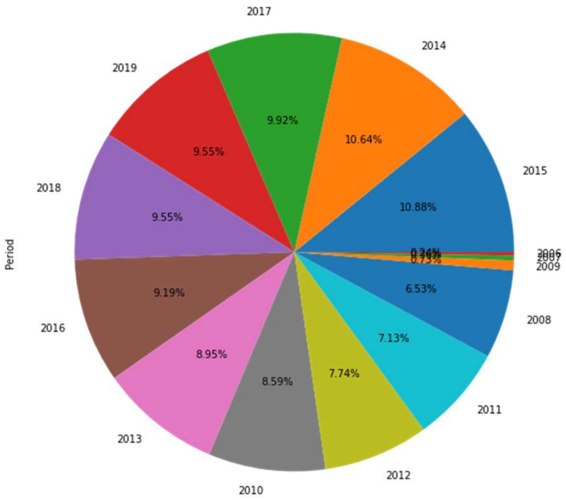
The figure above shows the statistical summary of the numerical columns on the dataset, telling us the count, mean, standard deviation, min max etc.

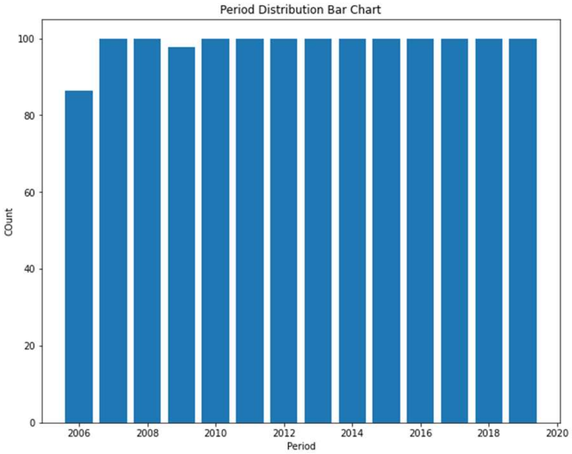


```
In [5]: # Checking the uniquenes of the values in each columns
        data.nunique()
Out[5]: Location            145
        Period               14
        FactValueNumeric    508
        dtype: int64
```

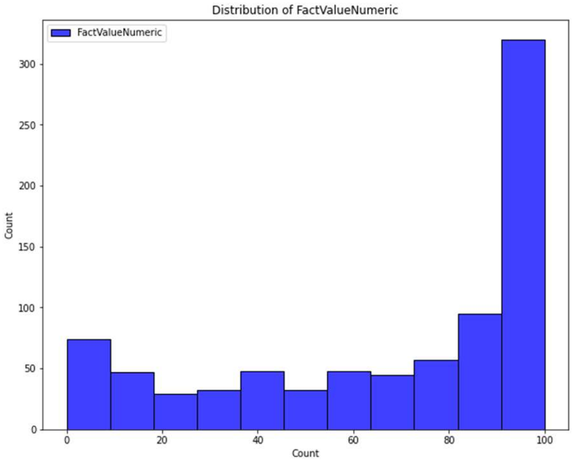
The unique count of each value in each dataset column is shown in the diagram above.

The pie chart distribution of the era with percentages is shown in the image below.

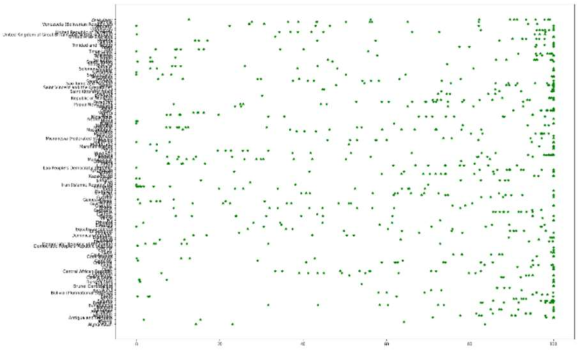




The bar chart visualization of the period count for each value count on the column in which 2014 has the lowest value count is shown in the image above.



The histogram distribution count of the FactValueNumeric is shown in the graphic above. The distribution is spaced out over the count, with a high frequency near the conclusion.



The scatter plot distribution between the FactValueNumeric and the location is shown in the image above, indicating that there is little correlation between them.