Read the dataset from the below link

https://raw.githubusercontent.com/guipsamora/pandas_exercises/master/06_Stats/US (https://raw.githubusercontent.com/guipsamora/pandas_exercises/master/06_Stats/US)_Baby_Names/US_Baby_Names_right.csv

Questions:

1. Delete unnamed columns
2. Show the distribution of male and female
3. Show the top 5 most preferred names
4. What is the median name occurence in the dataset
5. Distribution of male and female born count by states

In [1]:

```
import pandas as pd
import io
import requests
```

In [6]:

```
url="https://raw.githubusercontent.com/guipsamora/pandas_exercises/master/06_Stats/US_Baby_
s=requests.get(url).content
df=pd.read_csv(io.StringIO(s.decode('utf-8')))
```

In [7]:

```
type(df)
```

Out[7]:

```
pandas.core.frame.DataFrame
```

In [8]:

```
df.head()
```

Out[8]:

	Unnamed: 0	Id	Name	Year	Gender	State	Count
0	11349	11350	Emma	2004	F	AK	62
1	11350	11351	Madison	2004	F	AK	48
2	11351	11352	Hannah	2004	F	AK	46
3	11352	11353	Grace	2004	F	AK	44
4	11353	11354	Emily	2004	F	AK	41

In [9]:

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1016395 entries, 0 to 1016394
Data columns (total 7 columns):
Unnamed: 0    1016395 non-null int64
Id            1016395 non-null int64
Name          1016395 non-null object
Year          1016395 non-null int64
Gender        1016395 non-null object
State         1016395 non-null object
Count         1016395 non-null int64
dtypes: int64(4), object(3)
memory usage: 54.3+ MB
```

1. Delete unnamed columns

In [12]:

```
 df.loc[:, ~df.columns.str.contains('^Unnamed')]
```

...

In [15]:

```
df.drop(df.columns[df.columns.str.contains('unnamed',case = False)],axis = 1,inplace=True)
```

In [16]:

```
df.head()
```

Out[16]:

	Id	Name	Year	Gender	State	Count
0	11350	Emma	2004	F	AK	62
1	11351	Madison	2004	F	AK	48
2	11352	Hannah	2004	F	AK	46
3	11353	Grace	2004	F	AK	44
4	11354	Emily	2004	F	AK	41

2. Show the distribution of male and female

In [33]:

```
df['Gender'].value_counts()
```

Out[33]:

```
F    558846
M    457549
Name: Gender, dtype: int64
```

3.Show the top 5 most preferred names

In [59]:

```
df['Name'].value_counts().head(5)
```

Out[59]:

```
Riley     1112
Avery     1080
Jordan    1073
Peyton    1064
Hayden    1049
Name: Name, dtype: int64
```

In [62]:

```
top5=df['Name'].value_counts().head(5).index.values
```

In [63]:

```
for i in top5:
    print(i)
```

```
Riley
Avery
Jordan
Peyton
Hayden
```

4.What is the median name occurence in the dataset

In [64]:

```
df.head()
```

Out[64]:

	Id	Name	Year	Gender	State	Count
0	11350	Emma	2004	F	AK	62
1	11351	Madison	2004	F	AK	48
2	11352	Hannah	2004	F	AK	46
3	11353	Grace	2004	F	AK	44
4	11354	Emily	2004	F	AK	41

In [65]:

```
df.Name.describe()
```

Out[65]:

```
count     1016395
unique      17632
top         Riley
freq         1112
Name: Name, dtype: object
```

In [74]:

```
df['Name'].value_counts().median()
```

Out[74]:

```
8.0
```

5. Distribution of male and female born count by states

In [75]:

```
df.head()
```

Out[75]:

	Id	Name	Year	Gender	State	Count
0	11350	Emma	2004	F	AK	62
1	11351	Madison	2004	F	AK	48
2	11352	Hannah	2004	F	AK	46
3	11353	Grace	2004	F	AK	44
4	11354	Emily	2004	F	AK	41

In [95]:

```
df.groupby(['State','Gender'])['Count'].sum()
```

Out[95]:

```
State  Gender
AK     F          26250
       M          37399
AL     F         215308
       M         260114
AR     F         129712
       M         162947
AZ     F         368567
       M         439691
CA     F        2414063
       M        2670584
CO     F         260805
       M         313425
CT     F         141350
       M         171397
DC     F          35276
       M          47228
DE     F          31312
       M          41748
FL     F         915422
       M        1060957
GA     F         549637
       M         635531
HI     F          37279
       M          53127
IA     F         144764
       M         174009
ID     F          72808
       M          94320
IL     F         695312
       M         791679
                 ...
OK     F         184967
       M         228613
OR     F         172111
       M         209445
PA     F         593382
       M         682709
RI     F          35560
       M          47939
SC     F         197917
       M         237442
SD     F          34104
       M          45443
TN     F         336487
       M         398615
TX     F        1786281
       M        2005394
UT     F         202892
       M         245324
VA     F         405503
       M         466873
VT     F          15079
       M          21353
WA     F         334944
```

```
        M          395377
WI      F          264921
        M          311758
WV      F           73800
        M           93557
WY      F           14107
        M           21912
Name: Count, dtype: int64
```