Problem Statement

Read the dataset from the below link

https://raw.githubusercontent.com/guipsamora/pandas_exe (https://raw.githubusercontent.com/guipsamora/pandas_exe

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 [2]:
```
import pandas as pd
us_baby_names = pd.read_csv('https://raw.githubusercontent.com/guipsamora/pandas_
us_baby_names.head()
```

Out[2]:

	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

1. Delete unnamed columns¶

In [3]:
```
# deletes Unnamed: 0
del us_baby_names['Unnamed: 0']


us_baby_names.head()
```

Out[3]:

	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 [4]:
```
us_baby_names['Gender'].value_counts()
```

Out[4]:
```
F    558846
M    457549
Name: Gender, dtype: int64
```

3. Show the top 5 most preferred names

In [5]:
```
# Select the names with count only
names = us_baby_names[["Name", "Count"]]
# print(names)

# group by names and sum it
names_sum = names.groupby("Name").sum()

# print the first 5 observations
# print(names_sum.head(5))
#print(names_sum.shape)


# sort it from the biggest value to the smallest one
names_sum.sort_values("Count", ascending = 0).head()
```

Out[5]:

Name	Count
Jacob	242874
Emma	214852
Michael	214405
Ethan	209277
Isabella	204798

4. What is the median name occurence in the dataset

In [6]:
```
names_sum[names_sum.Count == names_sum.Count.median()]
```

Out[6]:

Name	Count
Aishani	49
Alara	49
Alysse	49
Ameir	49
Anely	49
Antonina	49
Aveline	49
Aziah	49
Baily	49
Caleah	49
Carlota	49
Cristine	49
Dahlila	49
Darvin	49
Deante	49
Deserae	49
Devean	49
Elizah	49
Emmaly	49
Emmanuela	49
Envy	49
Esli	49
Fay	49
Gurshaan	49
Hareem	49
Iven	49
Jaice	49
Jaiyana	49
Jamiracle	49
Jelissa	49
...	...
Kyndle	49

Name	Count
Kynsley	49
Leylanie	49
Maisha	49
Malillany	49
Mariann	49
Marquell	49
Maurilio	49
Mckynzie	49
Mehdi	49
Nabeel	49
Nalleli	49
Nassir	49
Nazier	49
Nishant	49
Rebecka	49
Reghan	49
Ridwan	49
Riot	49
Rubin	49
Ryatt	49
Sameera	49
Sanjuanita	49
Shalyn	49
Skylie	49
Sriram	49
Trinton	49
Vita	49
Yoni	49
Zuleima	49

66 rows × 1 columns

5. Distribution of male and female born count by states

In [7]:
```
gender_grouping_state = us_baby_names [["State","Gender","Count"]]
gender_grouping_state.groupby(["State", "Gender"]).sum()
```

Out[7]:

		Count
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

State	Gender	Count
	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

102 rows × 1 columns