PANDAS CONTINUED - Assignment

By Prakash Ghosh

Read the dataset from the below link

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

In [1]:
```
import pandas as pd
df_US_Baby_Names=pd.read_csv('https://raw.githubusercontent.com/guipsamora/pandas_exercises/master/06_Stats/US_Baby_Names/US_Baby_Names_right.csv')
```

In [2]:
```
df_US_Baby_Names.head(10)
```

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
5	11354	11355	Abigail	2004	F	AK	37
6	11355	11356	Olivia	2004	F	AK	33
7	11356	11357	Isabella	2004	F	AK	30
8	11357	11358	Alyssa	2004	F	AK	29
9	11358	11359	Sophia	2004	F	AK	28

Problem 1. Delete unnamed columns

In [3]:
```
df_US_Baby_Names.drop(labels='Unnamed: 0',axis=1,inplace=True)

df_US_Baby_Names.head(10)
```

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
5	11355	Abigail	2004	F	AK	37
6	11356	Olivia	2004	F	AK	33
7	11357	Isabella	2004	F	AK	30
8	11358	Alyssa	2004	F	AK	29
9	11359	Sophia	2004	F	AK	28

Problem 2. Show the distribution of male and female

In [4]:
```
df_US_Baby_Names.groupby(['Gender']).count()
```

Out[4]:

	Id	Name	Year	State	Count
Gender					
F	558846	558846	558846	558846	558846
M	457549	457549	457549	457549	457549

Problem 3. Show the top 5 most preferred names

In [6]:
```
# Most count of the Name is considered as most preferred
# Sort the DataFrame based on the Count (descending) take top five rows

df_US_Baby_Names.sort_values(['Count'],ascending=[False]).head(5)
```

Out[6]:

	Id	Name	Year	Gender	State	Count
107416	678594	Daniel	2004	M	CA	4167
110097	681275	Daniel	2005	M	CA	3914
115739	686917	Daniel	2007	M	CA	3865
112872	684050	Daniel	2006	M	CA	3826
107417	678595	Anthony	2004	M	CA	3805

Problem 4. What is the median name occurence in the dataset

In [7]:
```
# Median of the count
print('Median name occurence:\t' ,df_US_Baby_Names['Count'].median())
```

```
Median name occurence:	 11.0
```

Problem 5. Distribution of male and female born count by states

In [8]:
```
# group by State and Gender
df_US_Baby_Names.groupby(['State','Gender']).count()
```

Out[8]:

		Id	Name	Year	Count
State	**Gender**				
AK	**F**	2404	2404	2404	2404
	M	2587	2587	2587	2587
AL	**F**	9878	9878	9878	9878
	M	8419	8419	8419	8419
AR	**F**	7171	7171	7171	7171
	M	6475	6475	6475	6475
AZ	**F**	14518	14518	14518	14518
	M	10820	10820	10820	10820
CA	**F**	45144	45144	45144	45144
	M	31637	31637	31637	31637
CO	**F**	11424	11424	11424	11424
	M	9183	9183	9183	9183
CT	**F**	6575	6575	6575	6575
	M	5733	5733	5733	5733
DC	**F**	3053	3053	3053	3053
	M	3000	3000	3000	3000
DE	**F**	2549	2549	2549	2549
	M	2440	2440	2440	2440
FL	**F**	25781	25781	25781	25781
	M	20070	20070	20070	20070
GA	**F**	19385	19385	19385	19385
	M	15454	15454	15454	15454

		Id	Name	Year	Count
State	**Gender**				
HI	**F**	3255	3255	3255	3255
	M	3546	3546	3546	3546
IA	**F**	7131	7131	7131	7131
	M	6307	6307	6307	6307
ID	**F**	4918	4918	4918	4918
	M	4833	4833	4833	4833
IL	**F**	21268	21268	21268	21268
	M	16828	16828	16828	16828
...	**...**	...	...	...	...
OK	**F**	9519	9519	9519	9519
	M	8138	8138	8138	8138
OR	**F**	8604	8604	8604	8604
	M	7333	7333	7333	7333
PA	**F**	17480	17480	17480	17480
	M	14171	14171	14171	14171
RI	**F**	2558	2558	2558	2558
	M	2468	2468	2468	2468
SC	**F**	9465	9465	9465	9465
	M	8195	8195	8195	8195
SD	**F**	2838	2838	2838	2838
	M	2908	2908	2908	2908
TN	**F**	13063	13063	13063	13063

		Id	Name	Year	Count
State	**Gender**				
	M	10588	10588	10588	10588
TX	**F**	39760	39760	39760	39760
	M	27791	27791	27791	27791
UT	**F**	9515	9515	9515	9515
	M	8233	8233	8233	8233
VA	**F**	14759	14759	14759	14759
	M	11997	11997	11997	11997
VT	**F**	1398	1398	1398	1398
	M	1618	1618	1618	1618
WA	**F**	13329	13329	13329	13329
	M	11049	11049	11049	11049
WI	**F**	10549	10549	10549	10549
	M	8940	8940	8940	8940
WV	**F**	4305	4305	4305	4305
	M	3733	3733	3733	3733
WY	**F**	1456	1456	1456	1456
	M	1904	1904	1904	1904

102 rows × 4 columns