Data Science Masters :Assignment 9

```
Read the dataset from the below link

https://raw.githubusercontent.com/guipsamora/pandas_exercises/master/06_Stats/US_Baby_Nam
es/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 [62]:

```
import pandas as pd
df = pd.read_csv("https://raw.githubusercontent.com/guipsamora/pandas_exercises/master/06_S
df
```

Out[62]:

	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
10	11359	11360	Alexis	2004	F	AK	27
11	11360	11361	Elizabeth	2004	F	AK	27
12	11361	11362	Hailey	2004	F	AK	27
13	11362	11363	Anna	2004	F	AK	26
14	11363	11364	Natalie	2004	F	AK	25
15	11364	11365	Sarah	2004	F	AK	25
16	11365	11366	Sydney	2004	F	AK	25
17	11366	11367	Ava	2004	F	AK	23
18	11367	11368	Trinity	2004	F	AK	22
19	11368	11369	Haley	2004	F	AK	21
20	11369	11370	Kaylee	2004	F	AK	21
21	11370	11371	Taylor	2004	F	AK	21
22	11371	11372	Chloe	2004	F	AK	20
23	11372	11373	Ella	2004	F	AK	20
24	11373	11374	Mackenzie	2004	F	AK	20
25	11374	11375	Sierra	2004	F	AK	19
26	11375	11376	Kayla	2004	F	AK	18
27	11376	11377	Samantha	2004	F	AK	18
28	11377	11378	Zoe	2004	F	AK	18
29	11378	11379	Jessica	2004	F	AK	17
...	...	...	...	...	...	...	...
1016365	5647396	5647397	Brooks	2014	M	WY	5

	Unnamed: 0	Id	Name	Year	Gender	State	Count
1016366	5647397	5647398	Calvin	2014	M	WY	5
1016367	5647398	5647399	Cameron	2014	M	WY	5
1016368	5647399	5647400	Dalton	2014	M	WY	5
1016369	5647400	5647401	Dawson	2014	M	WY	5
1016370	5647401	5647402	Edward	2014	M	WY	5
1016371	5647402	5647403	Elias	2014	M	WY	5
1016372	5647403	5647404	Gage	2014	M	WY	5
1016373	5647404	5647405	Hayden	2014	M	WY	5
1016374	5647405	5647406	Jasper	2014	M	WY	5
1016375	5647406	5647407	Jose	2014	M	WY	5
1016376	5647407	5647408	Kaiden	2014	M	WY	5
1016377	5647408	5647409	Kaleb	2014	M	WY	5
1016378	5647409	5647410	Kasen	2014	M	WY	5
1016379	5647410	5647411	Kyson	2014	M	WY	5
1016380	5647411	5647412	Lukas	2014	M	WY	5
1016381	5647412	5647413	Myles	2014	M	WY	5
1016382	5647413	5647414	Nathaniel	2014	M	WY	5
1016383	5647414	5647415	Nolan	2014	M	WY	5
1016384	5647415	5647416	Oakley	2014	M	WY	5
1016385	5647416	5647417	Odin	2014	M	WY	5
1016386	5647417	5647418	Paxton	2014	M	WY	5
1016387	5647418	5647419	Raymond	2014	M	WY	5
1016388	5647419	5647420	Richard	2014	M	WY	5
1016389	5647420	5647421	Rowan	2014	M	WY	5
1016390	5647421	5647422	Seth	2014	M	WY	5
1016391	5647422	5647423	Spencer	2014	M	WY	5
1016392	5647423	5647424	Tyce	2014	M	WY	5
1016393	5647424	5647425	Victor	2014	M	WY	5
1016394	5647425	5647426	Waylon	2014	M	WY	5

1016395 rows × 7 columns

In [63]:

```
# Solution for question 1
df = df.drop(df.columns[df.columns.str.contains('unnamed',case = False)],axis = 1)
df
```

Out[63]:

	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
10	11360	Alexis	2004	F	AK	27
11	11361	Elizabeth	2004	F	AK	27
12	11362	Hailey	2004	F	AK	27
13	11363	Anna	2004	F	AK	26
14	11364	Natalie	2004	F	AK	25
15	11365	Sarah	2004	F	AK	25
16	11366	Sydney	2004	F	AK	25
17	11367	Ava	2004	F	AK	23
18	11368	Trinity	2004	F	AK	22
19	11369	Haley	2004	F	AK	21
20	11370	Kaylee	2004	F	AK	21
21	11371	Taylor	2004	F	AK	21
22	11372	Chloe	2004	F	AK	20
23	11373	Ella	2004	F	AK	20
24	11374	Mackenzie	2004	F	AK	20
25	11375	Sierra	2004	F	AK	19
26	11376	Kayla	2004	F	AK	18
27	11377	Samantha	2004	F	AK	18
28	11378	Zoe	2004	F	AK	18
29	11379	Jessica	2004	F	AK	17
...	...	...	...	...	...	...
1016365	5647397	Brooks	2014	M	WY	5
1016366	5647398	Calvin	2014	M	WY	5

	Id	Name	Year	Gender	State	Count
1016367	5647399	Cameron	2014	M	WY	5
1016368	5647400	Dalton	2014	M	WY	5
1016369	5647401	Dawson	2014	M	WY	5
1016370	5647402	Edward	2014	M	WY	5
1016371	5647403	Elias	2014	M	WY	5
1016372	5647404	Gage	2014	M	WY	5
1016373	5647405	Hayden	2014	M	WY	5
1016374	5647406	Jasper	2014	M	WY	5
1016375	5647407	Jose	2014	M	WY	5
1016376	5647408	Kaiden	2014	M	WY	5
1016377	5647409	Kaleb	2014	M	WY	5
1016378	5647410	Kasen	2014	M	WY	5
1016379	5647411	Kyson	2014	M	WY	5
1016380	5647412	Lukas	2014	M	WY	5
1016381	5647413	Myles	2014	M	WY	5
1016382	5647414	Nathaniel	2014	M	WY	5
1016383	5647415	Nolan	2014	M	WY	5
1016384	5647416	Oakley	2014	M	WY	5
1016385	5647417	Odin	2014	M	WY	5
1016386	5647418	Paxton	2014	M	WY	5
1016387	5647419	Raymond	2014	M	WY	5
1016388	5647420	Richard	2014	M	WY	5
1016389	5647421	Rowan	2014	M	WY	5
1016390	5647422	Seth	2014	M	WY	5
1016391	5647423	Spencer	2014	M	WY	5
1016392	5647424	Tyce	2014	M	WY	5
1016393	5647425	Victor	2014	M	WY	5
1016394	5647426	Waylon	2014	M	WY	5

1016395 rows × 6 columns

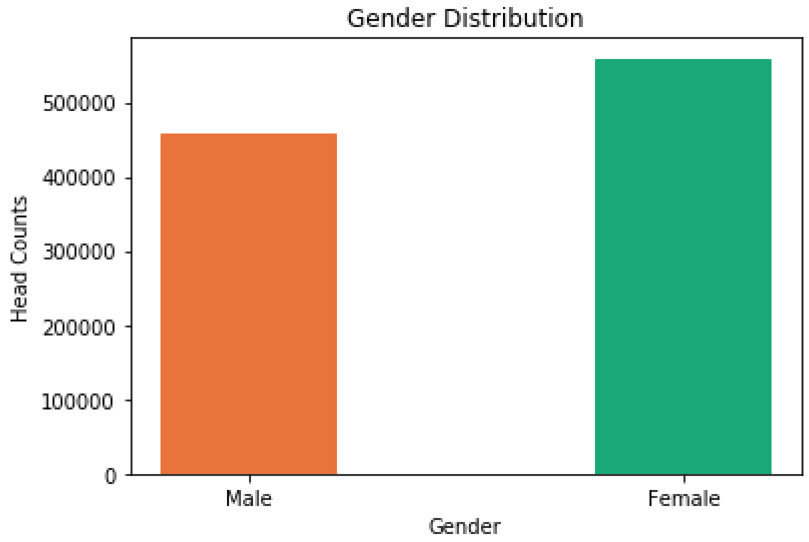
In [65]:

```
# Solution for question 2...

NoOfMale=df[df['Gender']=='M']['Gender'].count()
NoOfFemale=df[df['Gender']=='F']['Gender'].count()
print("Total no of Male =",NoOfMale)
print("Total no of Female =",NoOfFemale)

# Graphical Distribution
import matplotlib.pyplot as plt
male=df[df['Gender']=='M']['Gender'].count().astype(float)
female=df[df['Gender']=='F']['Gender'].count().astype(float)
x = [2,3]
y = [male,female]
bar=plt.bar(x, y, align='center',width=0.4)
bar[0].set_color('#E8743B')
bar[1].set_color('#19A979')
plt.title('Gender Distribution')
plt.ylabel('Head Counts')
plt.xlabel('Gender')
plt.xticks(x , ['Male','Female'])
plt.show()
plt.rcParams['figure.figsize'] = [5,5]
```

```
Total no of Male = 457549
Total no of Female = 558846
```



In [66]:

```
# Solution for question 3...
preferredNames = df['Name'].value_counts().head().index.values
print("Top 5 most preferred names are",','.join(preferredNames))
```

```
Top 5 most preferred names are Riley,Avery,Jordan,Peyton,Hayden
```

In [77]:

```
# Solution for question 4...
# As Name column is a character, sorting it and finding the middle name for maiden name occ
df2 = df.sort_values(by=['Name'])
df3 = df2['Name'].reset_index(drop=True)
middleIndex = (len(df2['Name']) - 1)/2  # finding the middle index in the Name column...
print("Median name occurence in the given dataset is",df3.loc[middleIndex])
```

```
Median name occurence in the given dataset is Jocelyn
```

In [69]:

```
# Solution for question 5...
df1 = pd.DataFrame(df.groupby(['State','Gender'])["Count"].sum())
df1
```

Out[69]:

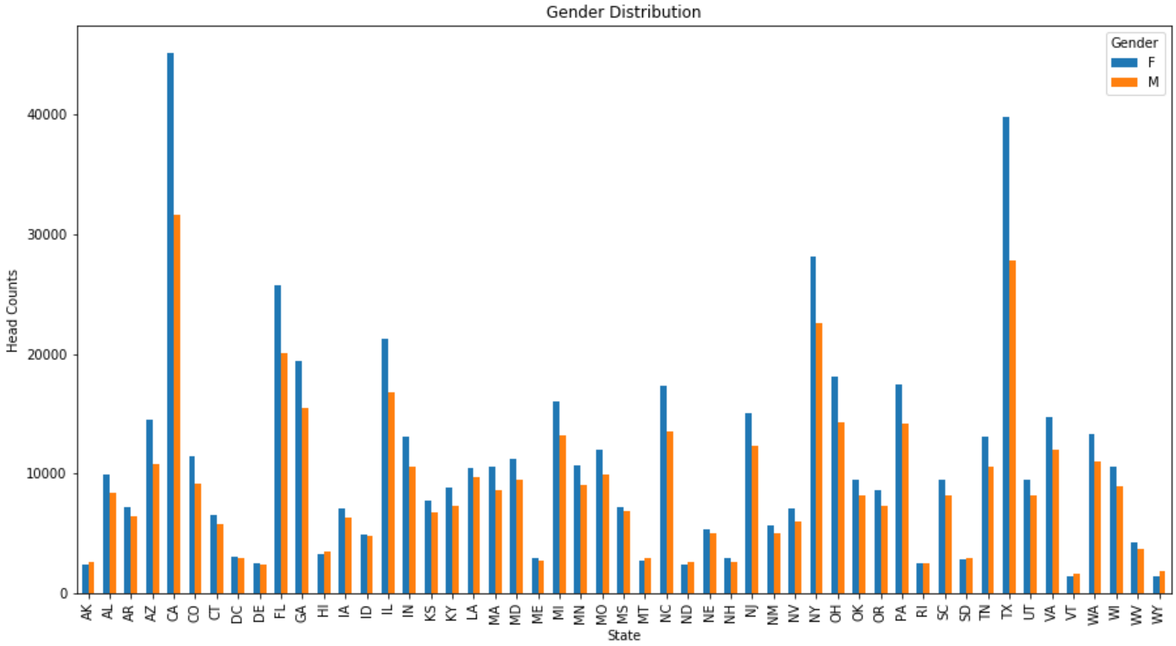
		Count
State	Gender	
AK	F	26250
	M	37399
AL	F	215308
	M	260114
AR	F	129712
	M	162947
AZ	F	368567
	M	439691
CA	F	2414063
	M	2670584
CO	F	260805
	M	313425
CT	F	141350
	M	171397
DC	F	35276
	M	47228
DE	F	31312
	M	41748
FL	F	915422
	M	1060957
GA	F	549637
	M	635531
HI	F	37279
	M	53127
IA	F	144764
	M	174009
ID	F	72808
	M	94320
IL	F	695312
	M	791679
...	...	...
OK	F	184967

		Count
State	Gender	
	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

In [76]:

```
# Graphical distribution...
df.groupby(['State','Gender']).size().unstack(fill_value=0).plot.bar(width=0.6)
plt.title('Gender Distribution')
plt.ylabel('Head Counts')
plt.xlabel('State')
plt.show()
plt.rcParams['figure.figsize'] = [15,8]
```


Gender Distribution
Gender
F
M
Head Counts
State
AK
AL
AR
AZ
CA
CO
CT
DC
DE
FL
GA
HI
IA
ID
IL
IN
KS
KY
LA
MA
MD
ME
MI
MN
MO
MS
MT
NC
ND
NE
NH
NJ
NM
NV
NY
OH
OK
OR
PA
RI
SC
SD
TN
TX
UT
VA
VT
WA
WI
WV
WY