US - Baby Names

Introduction:

We are going to use a subset of US Baby Names (https://www.kaggle.com/kaggle/us-baby-names) from Kaggle.
In the file it will be names from 2004 until 2014

Step 1. Import the necessary libraries

In [73]:

```
1 import pandas as pd
2 import seaborn as sns
3 import matplotlib.pyplot as plt
```

Step 2. Import the dataset from this address (https://raw.githubusercontent.com/guipsamora/pandas_exercises/master/06_Stat

Step 3. Assign it to a variable called baby_names.

In [74]:

```
1 baby_names = pd.read_csv('https://raw.githubusercontent.com/guipsamora/pandas_exercises
```

Step 4. See the first 10 entries

In [75]:

```
1 baby_names.head(10)
```

Out[75]:

	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

Step 5. Delete the column 'Unnamed: 0' and 'Id'

In [76]:

```
1 del baby_names['Unnamed: 0']
2 del baby_names['Id']
```

In [77]:

```
1 baby_names.head()
```

Out[77]:

	Name	Year	Gender	State	Count
0	Emma	2004	F	AK	62
1	Madison	2004	F	AK	48
2	Hannah	2004	F	AK	46
3	Grace	2004	F	AK	44
4	Emily	2004	F	AK	41

Step 6. Is there more male or female names in the dataset?

In [78]:

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

Out[78]:

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

Step 7. Group the dataset by name and assign to names

In [79]:

```
del baby_names['Year']
```

In [80]:

```
names = baby_names.groupby("Name").sum()
```

In [81]:

```
names.head()
```

Out[81]:

Name	Count
Aaban	12
Aadan	23
Aadarsh	5
Aaden	3426
Aadhav	6

In [82]:

```
print(names.shape)
```

```
(17632, 1)
```

In [83]:

```
1 names.sort_values("Count", ascending = 0).head()
```

Out[83]:

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

Step 8. How many different names exist in the dataset?

In [84]:

```
1 len(names)
```

Out[84]:

```
17632
```

Step 9. What is the name with most occurrences?

In [85]:

```
1 names.Count.idxmax()
```

Out[85]:

```
'Jacob'
```

Step 10. How many different names have the least occurrences?

In [91]:

```
1 len(names[names.Count == names.Count.min()])
```

Out[91]:

```
2578
```

Step 11. What is the median name occurrence?

In [95]:

```
1 names[names.Count == names.Count.median()]
```

Out[95]:

Name	Count
Aishani	49
Alara	49
Alysse	49
Ameir	49
Anely	49
...	...
Sriram	49
Trinton	49
Vita	49
Yoni	49
Zuleima	49

66 rows × 1 columns

Step 12. What is the standard deviation of names?

In [96]:

```
1 names.Count.std()
```

Out[96]:

```
11006.069467891111
```

Step 13. Get a summary with the mean, min, max, std and quartiles.

In [99]:

```
1 names.describe()
```

Out[99]:

	Count
count	17632.000000
mean	2008.932169
std	11006.069468
min	5.000000
25%	11.000000
50%	49.000000
75%	337.000000
max	242874.000000

In [ ]:

```
1
```