US - Baby Names

Introduction:

We are going to use a subset of US Baby Names from Kaggle.
In the file it will be names from 2004 until 2014

Step 1. Import the necessary libraries

```
import pandas as pd
```

Step 2. Import the dataset from this address.

Step 3. Assign it to a variable called baby_names.

```
baby_names = pd.read_csv('https://raw.githubusercontent.com/thieu1995/csv-files/main/data/pandas/US_Baby_Names_right.csv')
baby_names
```

	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
...	...	...	...	...	...	...	...
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

Step 4. See the first 10 entries

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

	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'

```
baby_names.drop(['Unnamed: 0', 'Id'], axis=1, inplace=True)
baby_names
```

	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
...	...	...	...	...	...
1016390	Seth	2014	M	WY	5
1016391	Spencer	2014	M	WY	5
1016392	Tyce	2014	M	WY	5
1016393	Victor	2014	M	WY	5
1016394	Waylon	2014	M	WY	5

1016395 rows × 5 columns

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

```
baby_names['Gender'].value_counts()
if baby_names['Gender'].value_counts()['F'] > baby_names['Gender'].value_counts()['M']:
  print('There is more female than male')
else:
  print('There is more male than female')
```

```
There is more female than male
```

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

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

```
<pandas.core.groupby.generic.DataFrameGroupBy object at 0x7b2574284590>
```

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

```
len(names)
```

```
17632
```

Step 9. What is the name with most occurrences?

```
names.size().sort_values(ascending = False).head(1)
```

Name	0
Riley	1112

dtype: int64

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

```
least_occurrence_count = names.size().min()
print(len(names.size()[names.size() == least_occurrence_count]))
```

```
3682
```

Step 11. What is the median name occurrence?

```
names_median = names.size()[names.size() == names.size().median()]
print(names_median)
```

```
Name
Abubakar    8
Adelie      8
Adira       8
Adylene     8
Aerial      8
           ..
Zamaya      8
Zanaya      8
Zari        8
Zaylie      8
Zyanna      8
Length: 360, dtype: int64
```

Step 12. What is the standard deviation of names?

```
names.size().std()
```

```
122.02996350814125
```

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

```
names.size().describe()
```

	0
count	17632.000000
mean	57.644907
std	122.029964
min	1.000000
25%	2.000000
50%	8.000000
75%	39.000000
max	1112.000000

dtype: float64