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')
```

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

	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?

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

print("female" if gender_counts['F'] > gender_counts['M'] else "male")
```

```
female
```

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

```
names = baby_names.groupby('Name')['Count'].sum()
```

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

```
int(names.count())
```

```
17632
```

Step 9. What is the name with most occurrences?

```
names.idxmax()
```

```
'Jacob'
```

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

```
min_ocr = names.min()
num_least_ocr = (names == min_ocr).sum()
print(f"number of names w least occurrences ({min_ocr}): {num_least_ocr}")
```

```
number of names w least occurrences (5): 2578
```

Step 11. What is the median name occurrence?

```
names.median()
```

```
49.0
```

Step 12. What is the standard deviation of names?

```
names.std()
```

```
11006.069467891111
```

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

```
names.describe()
```

	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

dtype: float64

Start coding or generate with AI.