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('us_baby.tsv', sep='\t')
```

Step 4. See the first 10 entries

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

```
     ,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(columns=[col for col in ['Unnamed: 0', 'Id'] if col in baby_names.columns], inplace=True)
```

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

```
baby_names = pd.read_csv('us_baby.tsv', sep=',', header=None, skiprows=1)
baby_names.columns = ['Index', 'Id', 'Name', 'Year', 'Gender', 'State', 'Count']

print(baby_names.head())
baby_names.drop(columns=['Index', 'Id'], inplace=True)

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

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

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

```
names = baby_names.groupby('Name').agg({'Count': 'sum'}).reset_index()
```

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

```
print("Number of different names:", names['Name'].nunique())
```

```
Number of different names: 17632
```

Step 9. What is the name with most occurrences?

```
most_common_name = names[names['Count'] == names['Count'].max()]
print("Most common name:\n", most_common_name)
```

```
Most common name:
         Name   Count
7198  Jacob  242874
```

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

```
least_common = names[names['Count'] == names['Count'].min()]
print("Number of least common names:", len(least_common))
```

```
Number of least common names: 2578
```

Step 11. What is the median name occurrence?

```
print("Median of name occurrences:", names['Count'].median())
```

```
Median of name occurrences: 49.0
```

Step 12. What is the standard deviation of names?

```
print("Standard deviation:", names['Count'].std())
```

```
Standard deviation: 11006.069467891111
```

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

```
print("Summary statistics:")
print(names['Count'].describe())
```

```
Summary statistics:
count     17632.000000
mean       2008.932169
std       11006.069468
min           5.000000
25%          11.000000
50%          49.000000
75%         337.000000
max      242874.000000
Name: Count, dtype: float64
```