main

November 21, 2022

Weekly Pandas Challenge #3

```
[1]: import pandas as pd
     file = "US_Baby_Names_right.csv"
```

```
[2]: df = pd.read_csv(file)
```

0.0.1 Take A Peek At Dataset

```
[3]: df.head()
```

```
[3]:    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
```

0.0.2 Q1. See the first 10 entries

```
[4]: df.head(10)
```

```
[4]:    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
```

0.0.3 Q2. Delete the columns 'Unnamed: 0' and 'Id'.

```
[5]: #axis=1 for columns
     #inplace=True means apply this operation on the actual dataframe
     df.drop(['Unnamed: 0', 'Id'], axis=1, inplace=True)
```

```
[6]: df.head()
```

```
[6]:       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
```

0.0.4 Q3. Group the dataset by name, assign to a variable called names, and sort the dataset by highest to lowest count.

```
[7]: names = df.groupby('Name', group_keys=False).apply(lambda x : x)
```

```
[8]: names.sort_values(['Count'], ascending=False)
```

```
[8]:              Name  Year Gender State  Count
     107416     Daniel  2004      M    CA   4167
     110097     Daniel  2005      M    CA   3914
     115739     Daniel  2007      M    CA   3865
     112872     Daniel  2006      M    CA   3826
     107417    Anthony  2004      M    CA   3805
     …             …    …      …    …      …
     470218        Gus  2005      M    MI      5
     470217   Giuseppe  2005      M    MI      5
     470216   Garrison  2005      M    MI      5
     470215     Garett  2005      M    MI      5
     1016394    Waylon  2014      M    WY      5

     [1016395 rows x 5 columns]
```

```
[9]: df.groupby('Name').apply(lambda x : x.sort_values(['Count'], ascending=False))
```

```
[9]:                  Name  Year Gender State  Count
     Name
     Aaban  693699   Aaban  2013      M    NY      6
            695768   Aaban  2014      M    NY      6
     Aadan  120728   Aadan  2008      M    CA      7
            123846   Aadan  2009      M    CA      6
            138678   Aadan  2014      M    CA      5
     …                 …    …      …    …      …
     Zyriah 855869  Zyriah  2006      F    TX      6
```

```
        885288  Zyriah  2014      F     TX      6
        235986  Zyriah  2007      F     GA      5
        244816  Zyriah  2012      F     GA      5
        867512  Zyriah  2009      F     TX      5

[1016395 rows x 5 columns]
```

0.0.5 Q4. How many different names exist in the dataset?

[10]:
```
df['Name'].unique().size
```

[10]:
```
17632
```

0.0.6 Q5. What is the name with most occurrences?

[24]:
```
df['Name'].value_counts()
```

[24]:
```
Riley      1112
Avery      1080
Jordan     1073
Peyton     1064
Hayden     1049
           …
Terryn        1
Yanna         1
Zemirah       1
Emmilyn       1
Coalton       1
Name: Name, Length: 17632, dtype: int64
```

[25]:
```
print(f"This name with most occurences is Riley with {max(df['Name'].
 ↪value_counts())} counts.")
```

```
This name with most occurences is Riley with 1112 counts.
```

0.0.7 Q6. What is the standard deviation of count of names?

[30]:
```
df['Name'].value_counts().std()
```

[30]:
```
122.02996350814088
```

0.0.8 Q7. Get a summary of the dataset with the mean, min, max, std and quartiles.

[31]:
```
df.describe()
```

[31]:
```
               Year         Count
count  1.016395e+06  1.016395e+06
mean   2.009053e+03  3.485012e+01
```

```
std    3.138293e+00  9.739735e+01
min    2.004000e+03  5.000000e+00
25%    2.006000e+03  7.000000e+00
50%    2.009000e+03  1.100000e+01
75%    2.012000e+03  2.600000e+01
max    2.014000e+03  4.167000e+03
```

Challenge Completed Successfully   Ready For More