

Importing one file and understanding Data Structure

In [8]: 1 `import pandas as pd`

In [9]: 1 `pd.read_csv(r"X:\Data Science\Projects\Work with Pandas, SQL Databases, JS`

Out[9]:

	Mary	F	7065
--	------	---	------

0	Anna	F	2604
1	Emma	F	2003
2	Elizabeth	F	1939
3	Minnie	F	1746
4	Margaret	F	1578
...
1994	Woodie	M	5
1995	Worthy	M	5
1996	Wright	M	5
1997	York	M	5
1998	Zachariah	M	5

1999 rows × 3 columns

In [10]: 1 `df = pd.read_csv(r"X:\Data Science\Projects\Work with Pandas, SQL Database`

In [11]: 1 `df.head()`

Out[11]:

	Name	Gender	Count
--	------	--------	-------

0	Mary	F	7065
1	Anna	F	2604
2	Emma	F	2003
3	Elizabeth	F	1939
4	Minnie	F	1746

In [12]:

```
1 df.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2000 entries, 0 to 1999
Data columns (total 3 columns):
#   Column  Non-Null Count  Dtype
---  -
0   Name    2000 non-null     object
1   Gender  2000 non-null     object
2   Count   2000 non-null     int64
dtypes: int64(1), object(2)
memory usage: 47.0+ KB
```

Importing & Merging many files

```
In [13]: 1 df_1880 = pd.read_csv('yob1880.txt', header = None, names = ['Names', 'Gen
2 df_1880.head()
```

Out[13]:

	Names	Gender	Count
0	Mary	F	7065
1	Anna	F	2604
2	Emma	F	2003
3	Elizabeth	F	1939
4	Minnie	F	1746

```
In [14]: 1 df_1881 = pd.read_csv('yob1881.txt', header = None, names = ['Names', 'Gen
2 df_1881
```

Out[14]:

	Names	Gender	Count
0	Mary	F	6919
1	Anna	F	2698
2	Emma	F	2034
3	Elizabeth	F	1852
4	Margaret	F	1658
...
1930	William	M	5
1931	Wilton	M	5
1932	Wing	M	5
1933	Wood	M	5
1934	Wright	M	5

1935 rows × 3 columns

In [17]:

```
1 pd.concat(objs = [df_1880, df_1881], axis = 0)
```

Out[17]:

	Names	Gender	Count
0	Mary	F	7065
1	Anna	F	2604
2	Emma	F	2003
3	Elizabeth	F	1939
4	Minnie	F	1746
...
1930	William	M	5
1931	Wilton	M	5
1932	Wing	M	5
1933	Wood	M	5
1934	Wright	M	5

3935 rows × 3 columns

In [21]:

```
1 pd.concat(objs = [df_1880, df_1881], axis = 0, keys = [1880,1881],
2           names = ["Year"]).droplevel(-1).reset_index()
```

Out[21]:

	Year	Names	Gender	Count
0	1880	Mary	F	7065
1	1880	Anna	F	2604
2	1880	Emma	F	2003
3	1880	Elizabeth	F	1939
4	1880	Minnie	F	1746
...
3930	1881	William	M	5
3931	1881	Wilton	M	5
3932	1881	Wing	M	5
3933	1881	Wood	M	5
3934	1881	Wright	M	5

3935 rows × 4 columns

In [25]: 1 `pd.read_csv("yob{}.txt".format(1880), header = None, names = ["Name", "Gen`

Out[25]:

	Name	Gender	Count
0	Mary	F	7065
1	Anna	F	2604
2	Emma	F	2003
3	Elizabeth	F	1939
4	Minnie	F	1746
...
1995	Woodie	M	5
1996	Worthy	M	5
1997	Wright	M	5
1998	York	M	5
1999	Zachariah	M	5

2000 rows × 3 columns

In [28]: 1 `years = list(range(1880,2019))`

In [29]: 1 `years`

Out[29]: [1880,
1881,
1882,
1883,
1884,
1885,
1886,
1887,
1888,
1889,
1890,
1891,
1892,
1893,
1894,
1895,
1896,
1897,
1898,
1899,
1900,
1901,
1902,
1903,
1904,
1905,
1906,
1907,
1908,
1909,
1910,
1911,
1912,
1913,
1914,
1915,
1916,
1917,
1918,
1919,
1920,
1921,
1922,
1923,
1924,
1925,
1926,
1927,
1928,
1929,
1930,
1931,
1932,
1933,
1934,
1935,
1936,
1937,
1938,
1939,
1940,
1941,
1942,
1943,
1944,
1945,
1946,
1947,
1948,
1949,
1950,
1951,
1952,
1953,
1954,
1955,
1956,
1957,
1958,
1959,
1960,
1961,
1962,
1963,
1964,
1965,
1966,
1967,
1968,
1969,
1970,
1971,
1972,
1973,
1974,
1975,
1976,
1977,
1978,
1979,
1980,
1981,
1982,
1983,
1984,
1985,
1986,
1987,
1988,
1989,
1990,
1991,
1992,
1993,
1994,
1995,
1996,
1997,
1998,
1999,
2000,
2001,
2002,
2003,
2004,
2005,
2006,
2007,
2008,
2009,
2010,
2011,
2012,
2013,
2014,
2015,
2016,
2017,
2018,
2019]

```
In [40]: 1 dataframes = []
2 for year in years:
3     data = pd.read_csv("yob{}.txt".format(year), header = None, names = ["
4     dataframes.append(data)
```

```
In [41]: 1 dataframes

2294 Winston M 5
2295 York M 5
2296 Zachariah M 5

[2297 rows x 3 columns],
      Name Gender Count
0      Mary F 9128
1      Anna F 3994
2      Emma F 2728
3 Elizabeth F 2582
4 Margaret F 2204
...      ...      ...
2289 Wallie M 5
2290 Willian M 5
2291 Wirt M 5
2292 Yee M 5
2293 Zeb M 5

[2294 rows x 3 columns],
      Name Gender Count
```

```
In [42]: 1 len(dataframes)
```

```
Out[42]: 139
```

```
In [44]: 1 df = pd.concat(dataframes, axis = 0, keys = years, names = ["Year"]).drop1
```

In [45]:

1 df

Out[45]:

	Year	Name	Gender	Count
0	1880	Mary	F	7065
1	1880	Anna	F	2604
2	1880	Emma	F	2003
3	1880	Elizabeth	F	1939
4	1880	Minnie	F	1746
...
1957041	2018	Zylas	M	5
1957042	2018	Zyran	M	5
1957043	2018	Zyrie	M	5
1957044	2018	Zyron	M	5
1957045	2018	Zzyzx	M	5

1957046 rows × 4 columns

In [47]:

1 df.info()

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1957046 entries, 0 to 1957045
Data columns (total 4 columns):
#   Column  Dtype
---  -
0    Year   int64
1    Name   object
2    Gender object
3    Count  int64
dtypes: int64(2), object(2)
memory usage: 59.7+ MB
```

In [48]:

1 df.to_csv("us_baby_names.csv", index = False)

```
In [49]: 1 pd.read_csv("us_baby_names.csv")
```

Out[49]:

	Year	Name	Gender	Count
0	1880	Mary	F	7065
1	1880	Anna	F	2604
2	1880	Emma	F	2003
3	1880	Elizabeth	F	1939
4	1880	Minnie	F	1746
...
1957041	2018	Zylas	M	5
1957042	2018	Zyran	M	5
1957043	2018	Zyrie	M	5
1957044	2018	Zyron	M	5
1957045	2018	Zzyzx	M	5

1957046 rows × 4 columns

```
In [ ]: 1
```