LECTURE 3

Pandas, Part I

Introduction to `pandas` syntax, operators, and functions

Data Science| Fall 23@ Knowledge Stream

Sana Jabbar

Useful Utility Functions

Lecture 03

- Data extraction with `loc`, `iloc`, and `[]`
- Conditional selection
- Adding, removing, and modifying columns
- **Useful utility functions**
- Custom sorts
- Grouping

Built-In `pandas` Methods

In addition to its rich syntax for indexing and support for other libraries (**`NumPy`**, native Python functions), **`pandas`** provides an enormous number of useful utility functions. Today, we'll discuss just a few:

- `size/shape`
- `describe`
- `sample`
- `value_counts`
- `unique`
- **`sort_values`**

The **`pandas`** library is rich in utility functions (we could spend the entire Course talking about them)! We encourage you to explore as you complete your assignments by Googling and reading documentation just as data scientists do.

`.describe()`

- `.describe()` returns a "description" of a `DataFrame` or `Series` that lists summary statistics of the data

`babynames`

	State	Sex	Year	Name	Count
0	CA	F	1910	Mary	295
1	CA	F	1910	Helen	239
2	CA	F	1910	Dorothy	220
3	CA	F	1910	Margaret	163
4	CA	F	1910	Frances	134
...	...	...	...	...	...
407423	CA	M	2022	Zayvier	5
407424	CA	M	2022	Zia	5
407425	CA	M	2022	Zora	5
407426	CA	M	2022	Zuriel	5
407427	CA	M	2022	Zylo	5

407428 rows × 5 columns

`babynames.describe()`

	Year	Count
count	407428.000000	407428.000000
mean	1985.733609	79.543456
std	27.007660	293.698654
min	1910.000000	5.000000
25%	1969.000000	7.000000
50%	1992.000000	13.000000
75%	2008.000000	38.000000
max	2022.000000	8260.000000

`.describe()`

- A different set of statistics will be reported if `.describe()` is called on a `Series`.

```
babynames["Sex"].describe()
```

```
count     407428
unique         2
top            F
freq      239537
Name: Sex, dtype: object
```

`.sample()`

To sample a random selection of rows from a `DataFrame`, we use the `.sample()` method.

- By default, ***it is without replacement***. Use `replace=True` for **replacement**.
- Naturally, can be chained with other methods and operators (`iloc`, etc).

```
babynames.sample()
```

	State	Sex	Year	Name	Count
121141	CA	F	1992	Shanelle	28

```
babynames.sample(5).iloc[:, 2:]
```

	Year	Name	Count
44448	1961	Karyn	36
260410	1948	Carol	7
397541	2019	Arya	11
4767	1921	Sumiko	16
104369	1987	Thomas	11

```
babynames[babynames["Year"] == 2000]
         .sample(4, replace=True)
         .iloc[:, 2:]
```

	Year	Name	Count
151749	2000	Iridian	7
343560	2000	Maverick	14
149491	2000	Stacy	91
149212	2000	Angel	307

`.value_counts()`

The **`Series.value_counts`** method counts the number of occurrences of each unique value in a **`Series`** (it *counts* the number of times each *value* appears).

- Return value is also a **`Series`**.

```
babyname["Name"].value_counts()
```

```
Name
Jean         223
Francis      221
Guadalupe    218
Jessie       217
Marion       214
             ...
Renesme        1
Purity         1
Olanna         1
Nohea          1
Zayvier        1
Name: count, Length: 20437, dtype: int64
```

.unique()

The `Series.unique` method returns an array of every unique value in a `Series`.

```
babynames["Name"].unique()
```

```
array(['Mary', 'Helen', 'Dorothy', ..., 'Zae', 'Zai', 'Zayvier'],
      dtype=object)
```

`.sort_values()`

The `DataFrame.sort_values` and `Series.sort_values` methods sort a `DataFrame` (or `Series`).

- **`Series.sort_values( )` will automatically sort all values in the `Series`.**
- `DataFrame.sort_values(column_name)` must specify the name of the column to be used for sorting.

```
babynames["Name"].sort_values()
```

```
366001      Aadan
384005      Aadan
369120      Aadan
398211    Aadarsh
370306      Aaden
           ...
220691      Zyrah
197529      Zyrah
217429      Zyrah
232167      Zyrah
404544      Zyrus
Name: Name, Length: 407428, dtype: object
```

`.sort_values()`

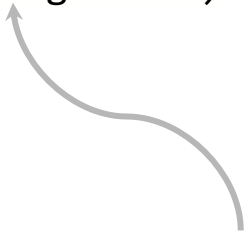
The `DataFrame.sort_values` and `Series.sort_values` methods sort a `DataFrame` (or `Series`).

- `Series.sort_values( )` will automatically sort all values in the `Series`.
- **`DataFrame.sort_values(column_name)` must specify the name of the column to be used for sorting.**

```
babynames.sort_values(by = "Count", ascending=False)
```

	State	Sex	Year	Name	Count
268041	CA	M	1957	Michael	8260
267017	CA	M	1956	Michael	8258
317387	CA	M	1990	Michael	8246
281850	CA	M	1969	Michael	8245
283146	CA	M	1970	Michael	8196
...	...	...	...	...	...
317292	CA	M	1989	Olegario	5
317291	CA	M	1989	Norbert	5
317290	CA	M	1989	Niles	5
317289	CA	M	1989	Nikola	5
407427	CA	M	2022	Zylo	5

407428 rows × 5 columns



By default, rows are sorted in *ascending* order.

Custom Sorts

Lecture 03

- Data extraction with `loc`, `iloc`, and `[]`
- Conditional selection
- Adding, removing, and modifying columns
- Useful utility functions
- **Custom sorts**
- Grouping

Sorting By Length

Let’s try to solve the sorting problem with different approaches:

- We will create a temporary column, then sort on it.

Approach 1: Create a Temporary Column and Sort Based on the New Column

Sorting the `DataFrame` as usual:

```
# Create a Series of the length of each name
babyname_lengths = babynames["Name"].str.len()
```

```
# Add a column named "name_lengths" that includes the length of each name
babynames["name_lengths"] = babyname_lengths
```

```
babynames = babynames.sort_values(by = "name_lengths", ascending=False)
babynames.head(5)
```

	State	Sex	Year	Name	Count	name_lengths
334166	CA	M	1996	Franciscojavier	8	15
337301	CA	M	1997	Franciscojavier	5	15
339472	CA	M	1998	Franciscojavier	6	15
321792	CA	M	1991	Ryanchristopher	7	15
327358	CA	M	1993	Johnchristopher	5	15

Approach 2: Sorting Using the `key` Argument

```
babynames.sort_values("Name", key=lambda x: x.str.len(), ascending=False)
                 .head()
```

	State	Sex	Year	Name	Count
334166	CA	M	1996	Franciscojavier	8
327472	CA	M	1993	Ryanchristopher	5
337301	CA	M	1997	Franciscojavier	5
337477	CA	M	1997	Ryanchristopher	5
312543	CA	M	1987	Franciscojavier	5

Approach 3: Sorting Using the `map` Function

Suppose we want to sort by the number of occurrences of "dr" and “ea”s.

- Use the `Series.map` method.

```
def dr_ea_count(string):
    return string.count('dr') + string.count('ea')

# Use map to apply dr_ea_count to each name in the "Name" column
babynames["dr_ea_count"] = babynames["Name"].map(dr_ea_count)
babynames = babynames.sort_values(by = "dr_ea_count", ascending=False)
babynames.head()
```

	State	Sex	Year	Name	Count	dr_ea_count
115957	CA	F	1990	Deandrea	5	3
101976	CA	F	1986	Deandrea	6	3
131029	CA	F	1994	Leandrea	5	3
108731	CA	F	1988	Deandrea	5	3
308131	CA	M	1985	Deandrea	6	3

Grouping

Lecture 03

- Data extraction with `loc`, `iloc`, and `[]`
- Conditional selection
- Adding, removing, and modifying columns
- Useful utility functions
- Custom sorts
- **Grouping**

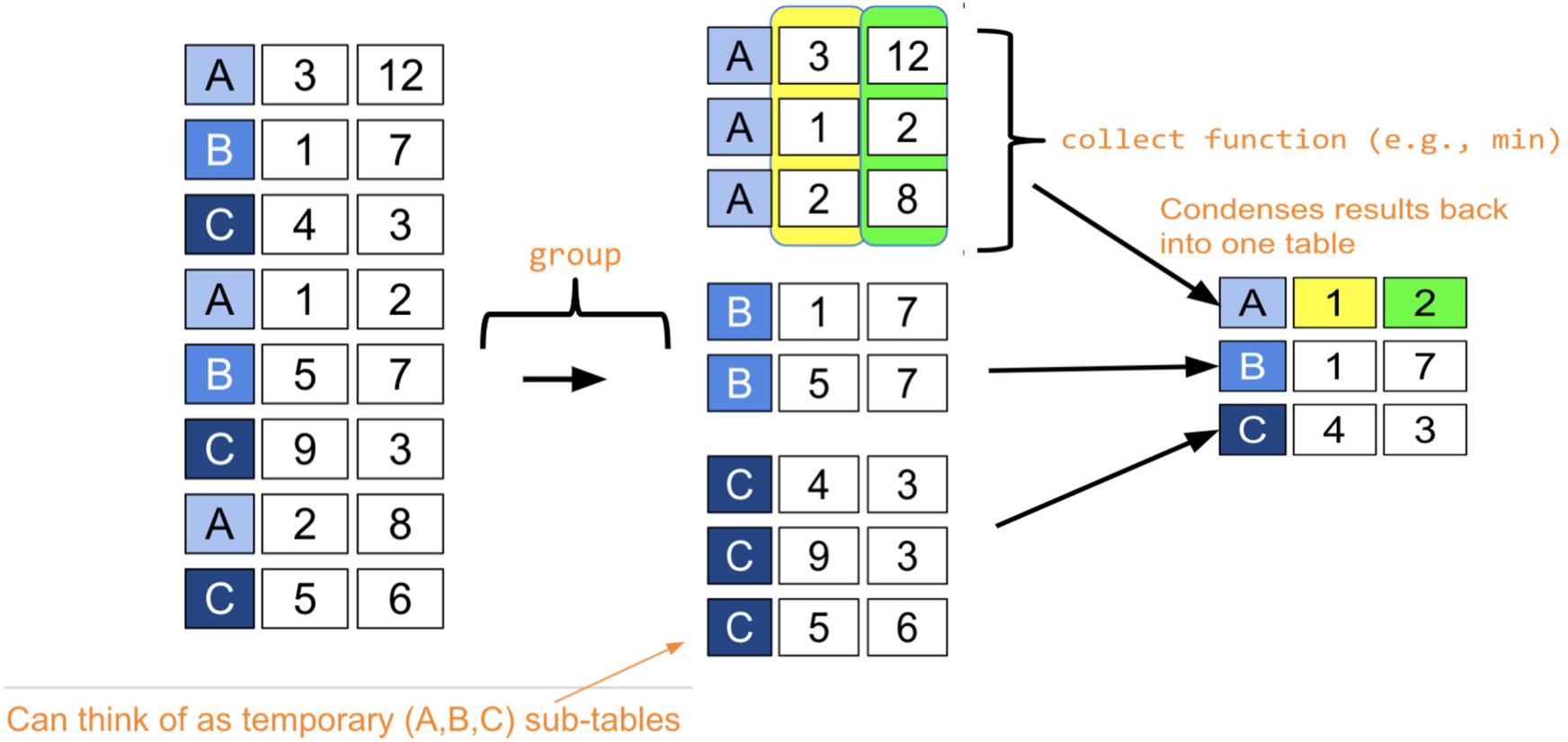
Why Group?

Our goal:

- Group together rows that fall under the same category.
  - For example, group together all rows from the same year.
- Perform an operation that *aggregates* across all rows in the category.
  - For example, sum up the total number of babies born in that year.

Grouping is a powerful tool to 1) perform large operations, all at once and 2) summarize trends in a dataset.

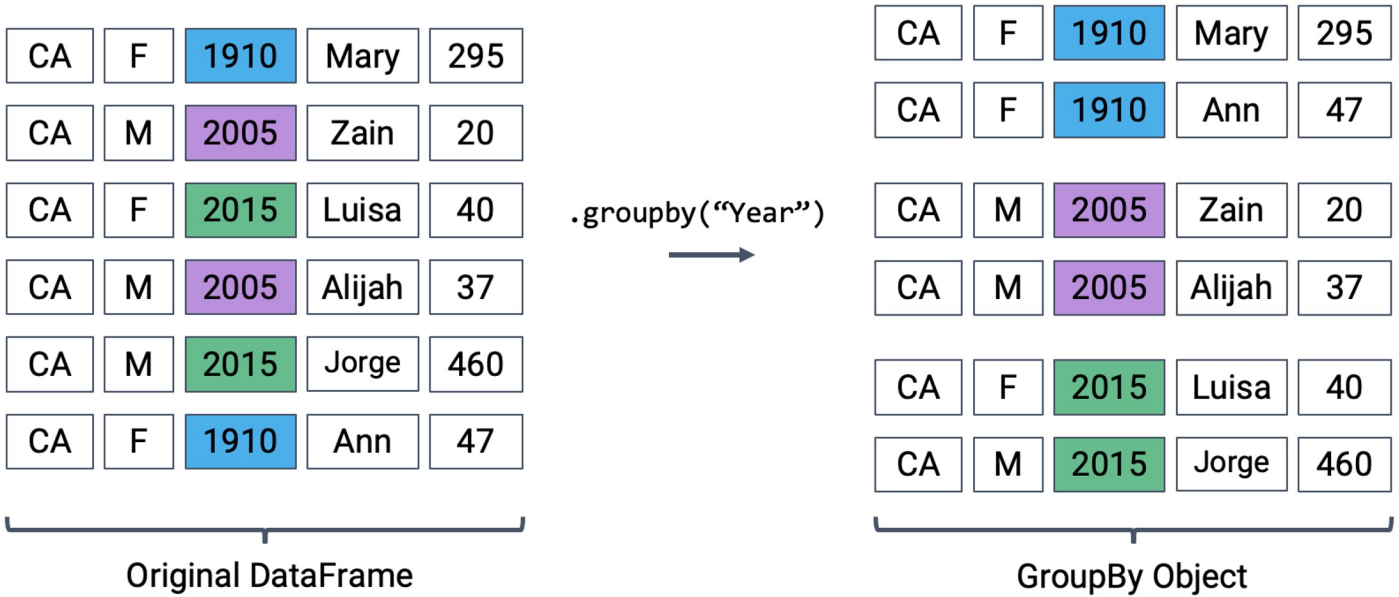
Visual Review of Grouping and Collection



`.groupby()`

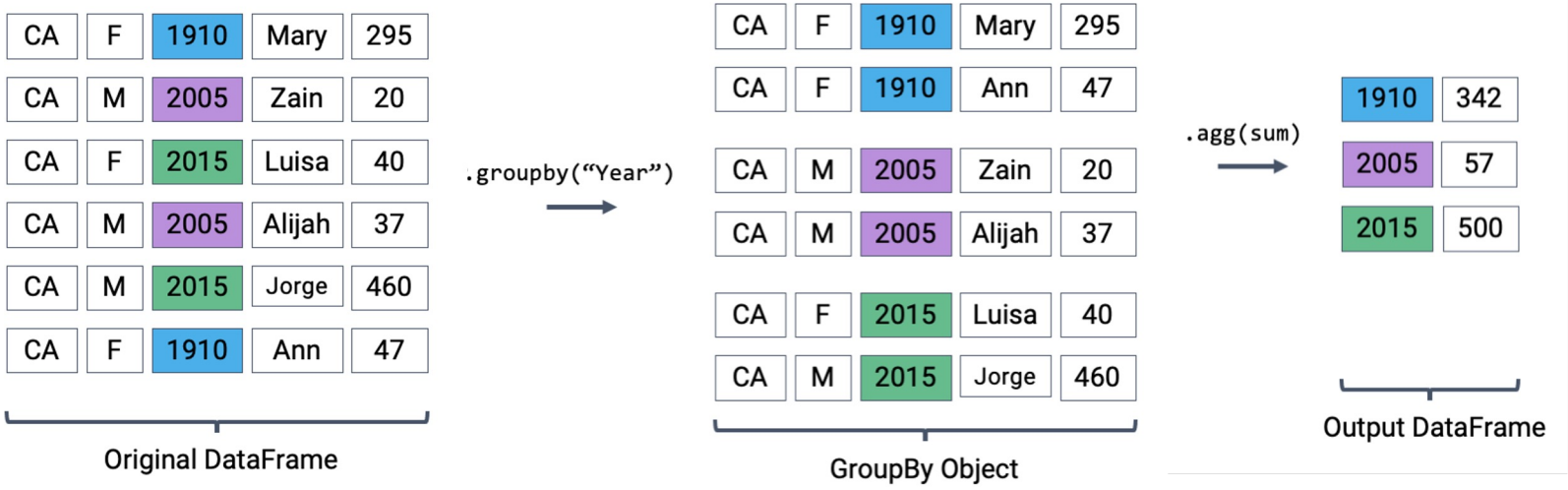
A `.groupby()` operation involves some combination of **splitting the object, applying a function**, and **combining the results**.

- Calling `.groupby()` generates `DataFrameGroupBy` objects → "mini" sub-DataFrames
- Each subframe contains all rows that correspond to the same group (here, a particular year)



`.groupby().agg()`

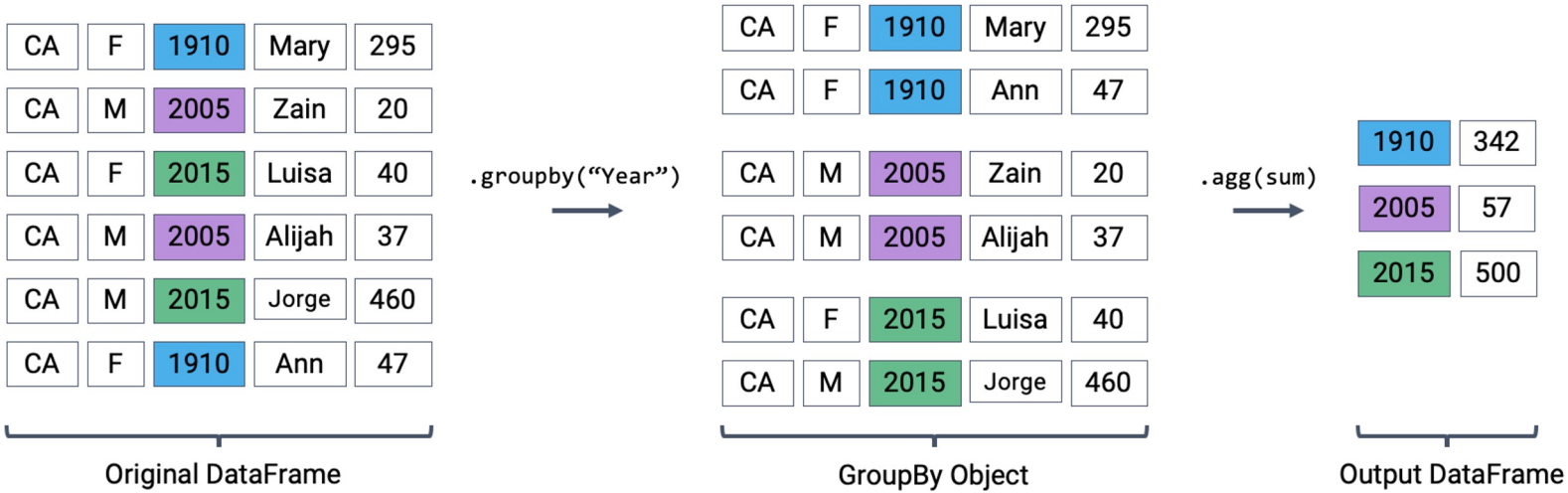
- We cannot work directly with **`DataFrameGroupBy`** objects! The diagram below is to help understand what goes on conceptually – in reality, we can't "see" the result of calling `.groupby`.
- Instead, we transform a **`DataFrameGroupBy`** object back into a DataFrame using `.agg`
  - `.agg` is how we apply an aggregation operation to the data.



Putting It All Together

```
dataframe.groupby(column_name).agg(aggregation_function)
```

- `babynames[["Year", "Count"]].groupby("Year").agg(sum)` computes the total number of babies born in each year.



Aggregation Functions

What goes inside of `.agg( )`?

- Any function that aggregates several values into one summary value
- Common examples:

In-Built Python Functions	NumPy Functions	In-Built `pandas` functions
`.agg(sum)`	`.agg(np.sum)`	`.agg("sum")`
`.agg(max)`	`.agg(np.max)`	`.agg("max")`
`.agg(min)`	`.agg(np.min)`	`.agg("min")`
	`.agg(np.mean)`	`.agg("mean")`
		`.agg("first")`
		`.agg("last")`

Some commonly-used aggregation functions can even be called directly, without the explicit use of `.agg( )`

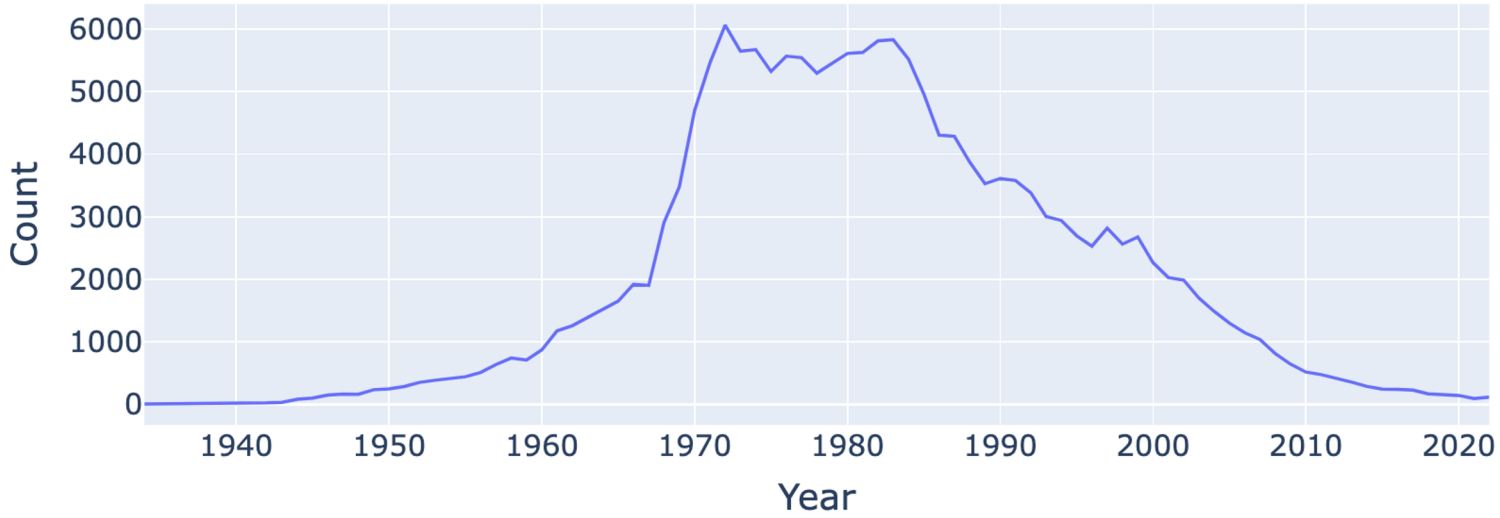
```
babynames.groupby("Year").mean()
```

Putting Things Into Practice

Goal: Find the baby name with sex "F" that has fallen in popularity the most.

```
f_babynames = babynames[babynames["Sex"] == "F"]
f_babynames = f_babynames.sort_values(["Year"])
jenn_counts_series = f_babynames[f_babynames["Name"] == "Jennifer"]["Count"]
```

Number of Jennifers Born in California Per Year



What Is "Popularity"?

Goal: Find the baby name with sex "F" that has fallen in popularity the most.

How do we define "fallen in popularity?"

- Let's create a metric: "ratio to peak" (RTP).
- The RTP is the ratio of babies born with a given name in 2022 to the *maximum* number of babies born with that name in *any* year.

Example for "Jennifer":

- In 1972, we hit peak Jennifer. 6,065 Jennifers were born.
- In 2022, there were only 114 Jennifers.
- RTP is 114 / 6065 = 0.018796372629843364.

Calculating RTP

```
max_jenn = max(f_babynames[f_babynames["Name"] == "Jennifer"]["Count"])
```

```
6065
```

```
curr_jenn = f_babynames[f_babynames["Name"] == "Jennifer"]["Count"].iloc[-1]
```

```
114
```

Remember: `f_babynames` is sorted by year.
`.iloc[-1]` means "grab the latest year"

```
rtp = curr_jenn / max_jenn
```

```
0.018796372629843364
```

```
def ratio_to_peak(series):
    return series.iloc[-1] / max(series)
```

```
jenn_counts_ser = f_babynames[f_babynames["Name"] == "Jennifer"]["Count"]
ratio_to_peak(jenn_counts_ser)
```

```
0.018796372629843364
```

Calculating RTP Using `.groupby()`

`.groupby()` makes it easy to compute the RTP for all names at once!

```
rtp_table = f_babynames.groupby("Name")[["Year","Count"]].agg(ratio_to_peak)
```

Name	Year	Count
Aadhini	1.0	1.000000
Aadhira	1.0	0.500000
Aadhya	1.0	0.660000
Aadya	1.0	0.586207
Aahana	1.0	0.269231
...	...	...
Zyanya	1.0	0.466667
Zyla	1.0	1.000000
Zylah	1.0	1.000000
Zyra	1.0	1.000000
Zyrah	1.0	0.833333

13782 rows × 2 columns

Here, `collect = rtp`.

we'll implement it using `pandas`.

Answer

In the five rows shown, note the Year is 1.0 for every value.

Are there any rows for which Year is **not** 1.0?

A. Yes, names that appeared for the first time in 2022.
B. Yes, names that did not appear in 2022.
C. Yes, names whose peak Count was in 2022.
D. No, every row has a Year value of 1.0.

```
rtp_table = (
    f_babynames
    .groupby("Name")[["Year","Count"]]
    .agg(ratio_to_peak)
)
```

Name	Year	Count
Aadhini	1.0	1.000000
Aadhira	1.0	0.500000
Aadhya	1.0	0.660000
Aadya	1.0	0.586207
Aahana	1.0	0.269231
...	...	...
Zyanya	1.0	0.466667
Zyla	1.0	1.000000
Zylah	1.0	1.000000
Zyra	1.0	1.000000
Zyrah	1.0	0.833333

13782 rows × 2 columns

Answer

In the five rows shown, note the Year is 1.0 for every value.

Are there any rows for which Year is **not** 1.0?

A. Yes, names that appeared for the first time in 2022.
B. Yes, names that did not appear in 2022.
C. Yes, names whose peak Count was in 2022.
D. No, every row has a Year value of 1.0.

```
rtp_table = (
    f_babynames
    .groupby("Name")[["Year","Count"]]
    .agg(ratio_to_peak)
)
```

Name	Year	Count
Aadhini	1.0	1.000000
Aadhira	1.0	0.500000
Aadhya	1.0	0.660000
Aadya	1.0	0.586207
Aahana	1.0	0.269231
...	...	...
Zyanya	1.0	0.466667
Zyla	1.0	1.000000
Zylah	1.0	1.000000
Zyra	1.0	1.000000
Zyrah	1.0	0.833333

13782 rows × 2 columns

A Note on Nuisance Columns

At least as of the time of this slide creation executing our agg call results in a `TypeError`.

```
f_babynames.groupby("Name").agg(ratio_to_peak)
```

```
Cell In[110], line 5, in ratio_to_peak(series)
      1 def ratio_to_peak(series):
      2     """
      3     Compute the RTP for a Series containing the counts per year for a single name
      4     """
----> 5     return series.iloc[-1] / np.max(series)

TypeError: unsupported operand type(s) for /: 'str' and 'str'
```

A Note on Nuisance Columns

Below, we explicitly select the column(s) we want to apply our aggregation function to **BEFORE** calling **agg**. This avoids the warning (and can prevent unintentional loss of data).

```
rtp_table = f_babynames.groupby("Name")[["Count"]].agg(ratio_to_peak)
```

	Count
Name	
Aadhini	1.000000
Aadhira	0.500000
Aadhya	0.660000
Aadya	0.586207
Aahana	0.269231
...	...
Zyanya	0.466667
Zyla	1.000000
Zylah	1.000000
Zyra	1.000000
Zyrah	0.833333

13782 rows × 1 columns

Renaming Columns After Grouping

By default, `.groupby` will not rename any aggregated columns (the column is still named "Count", even though it now represents the RTP.

For better readability, we may wish to rename “Count” to “Count RTP”

```
rtp_table = female_babynames.groupby("Name")[["Count"]].agg(ratio_to_peak)
rtp_table = rtp_table.rename(columns = {"Count": "Count RTP"})
```

Name	Count
Aadhini	1.000000
Aadhira	0.500000
Aadhya	0.660000
Aadya	0.586207
Aahana	0.269231
...	...

Name	Count RTP
Aadhini	1.000000
Aadhira	0.500000
Aadhya	0.660000
Aadya	0.586207
Aahana	0.269231
...	...

Some Data Science Payoff

By sorting `rtp_table` we can see the names whose popularity has decreased the most.

```
rtp_table.sort_values("Count RTP")
```

Name	Count RTP
Debra	0.001260
Debbie	0.002815
Carol	0.003180
Tammy	0.003249
Susan	0.003305
...	...
Fidelia	1.000000
Naveyah	1.000000
Finlee	1.000000
Roseline	1.000000
Aadhini	1.000000

13782 rows × 1 columns

Some Data Science Payoff

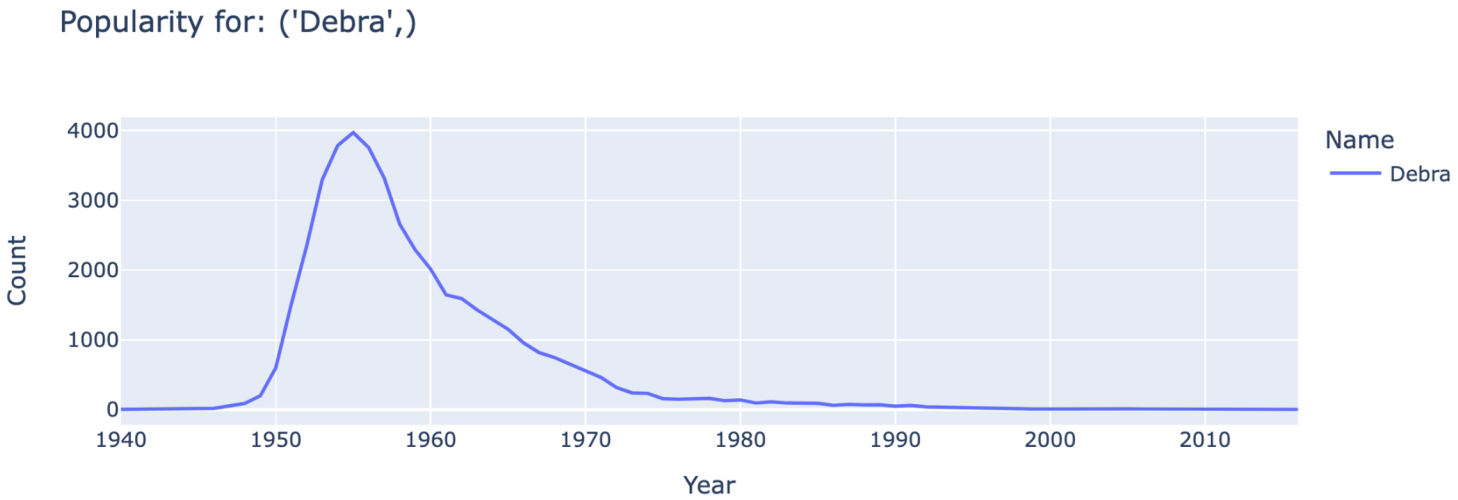
By sorting `rtp_table` we can see the names whose popularity has decreased the most.

```
rtp_table.sort_values("Count RTP")
```

Name	Count RTP
Debra	0.001260
Debbie	0.002815
Carol	0.003180
Tammy	0.003249
Susan	0.003305
...	...
Fidelia	1.000000
Naveyah	1.000000
Finlee	1.000000
Roseline	1.000000
Aadhini	1.000000

13782 rows × 1 columns

```
px.line(f_babynames[f_babyname["Name"] == "Debra")],
                        x = "Year", y = "Count")
```



We'll learn about plotting in week 3.

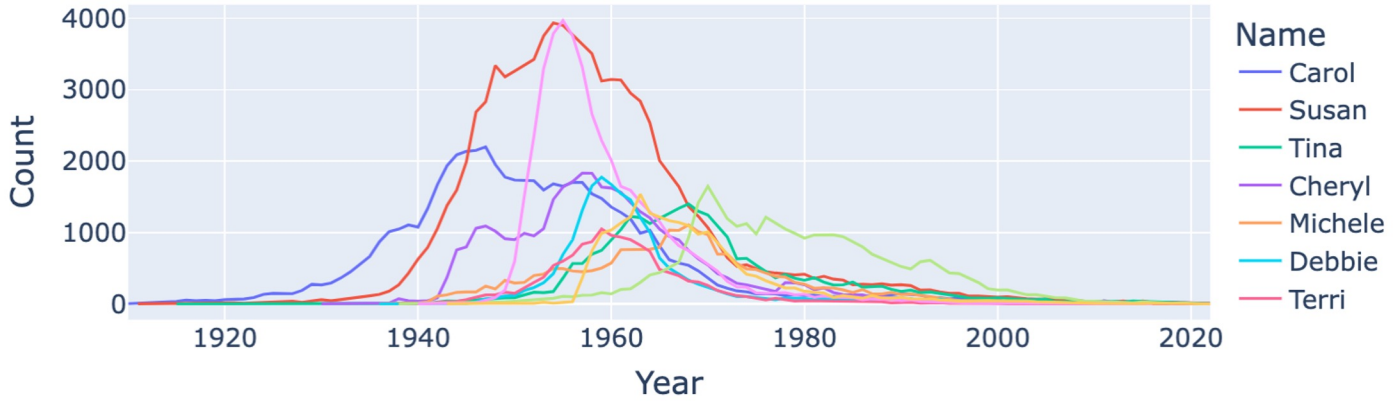
Some Data Science Payoff

We can get the list of the top 10 names and then plot popularity with::

```
top10 = rtp_table.sort_values("Count RTP").head(10).index
```

```
ndex(['Debra', 'Debbie', 'Carol', 'Tammy', 'Susan', 'Cheryl', 'Shannon',
       'Tina', 'Michele', 'Terri'],
      dtype='object', name='Name')
```

```
px.line(f_babynames[f_babyname["Name"].isin(top10)],
                          x = "Year", y = "Count", color = "Name")
```



Let’s Start Work on Notebook