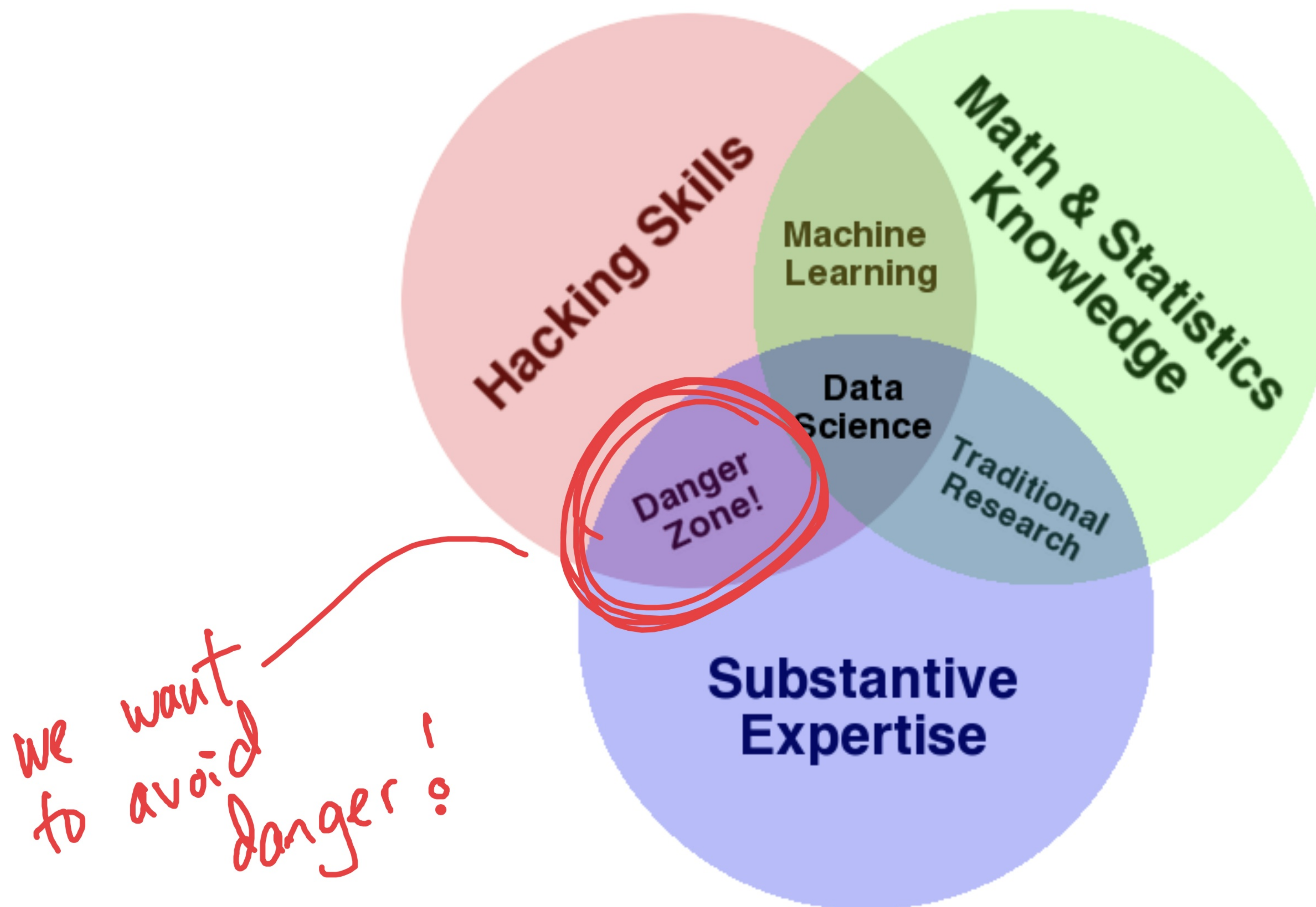


Question 🤔 (Answer at practicaldsc.org/q)

Select the **FALSE** statement below.

- A. I have size 16 feet. *True*
- B. I skipped the first grade *True*
- C. I was rejected by Michigan when I applied for undergrad. *False*
- D. Soulja Boy (the rapper) used to follow me on Twitter. *True*
- E. I was an undergrad student during COVID. *True*

way publishes his famous Data Science Venn Diagram



But there's no agreement on which "Venn diagram" is correct – see [here](#)!

Advice from last year's students

Practice really helps! I think most of the concepts are really easy to understand, but its the details that are hard to remember. As a data scientist, in practice you wouldn't want to look up every little thing, so practice will help in memorizing the "language" of this class and the concepts as well.

Stay consistent (boring, but still important). Buckle in because it covers a wide range of material. The second half is very different from the first half and transitions into a different kind of class.

Plan on starting the HW assignments as early as you can- some took a lot longer than others, and I never really knew how long one would take me until I began. Don't stress too much though, the generous late day policy was incredibly helpful!

My biggest piece of advice is to definitely make sure to keep up with the discussion worksheets every week. I felt like working on those was best way to study for the exam and the reason why I did well on the midterm. In addition, don't rely too much on gpt for the homeworks because it can hurt you in the long run.

study.practicaldisc.org

In [7]: 1 min(1, 2, 3) + max(4, 5, 6)

Out [7]: 7

In [8]: 1 # Why do we only see one line

2 3 + 5

3 4 * 10

Out [8]: 40

by default, only
shows value
of the last
expression
in cell!

In []: 1 # Strings can be created usi

In [7]: 1 min(1, 2, 3) + max(4, 5, 6)

Out[7]: 7

In [9]: 1 # Why do we only see one line

2 print(3 + 5)

3 4 * 10

9th time
we ran
code cell
→ 8
↓

Out[9]: 40

ut[13]: 40

n [14]:

1	# Strings can be created
2	# There's no difference
3	'679' + '888'

concatenation .

ut[14]: '679888'

In []:

1	...
---	-----

2 baby

Out[20]:

	Name	Sex	Count	Year
0	Liam	M	20456	2022
1	Noah	M	18621	2022
2	Olivia	F	16573	2022
...	...	...	...	...
2085155	Wright	M	5	1880
2085156	York	M	5	1880
2085157	Zachariah	M	5	1880

Olivia M 5. 2022

2085158 rows x 4 columns



How many babies were recorded in the dataset each year?

In [24]:

```
1 baby
```

Out [24]:

	Name	Sex	Count	Year
0	Liam	M	20456	2022
1	Noah	M	18621	2022
2	Olivia	F	16573	2022
...	...	...	...	...
2085155	Wright	M	5	1880
2085156	York	M	5	1880
2085157	Zachariah	M	5	1880

2085158 rows x 4 columns

sum count col,
separately
for each year



In []:

```
1 ...
```





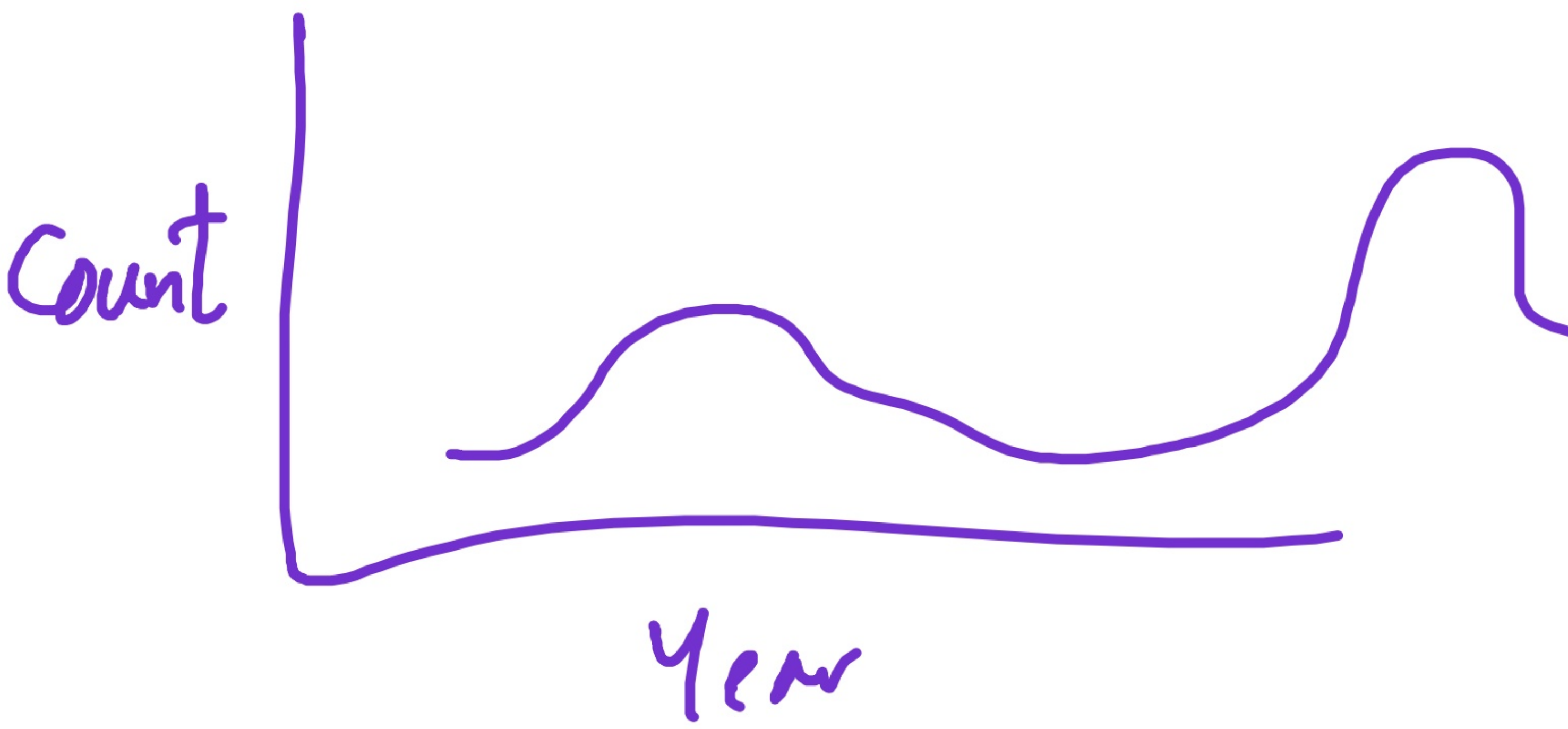
What about trends in individual names?

```
In [35]: 1 baby[baby['Name'] == 'Olivia']
```

Out[35]:

	Name	Sex	Count	Year
2	Olivia	F	16573	2022
13017	Olivia	M	16	2022
31917	Olivia	F	17798	2021
...	...	...	...	...
2079563	Olivia	F	52	1882
2081652	Olivia	F	51	1881
2083640	Olivia	F	44	1880

202 rows x 4 columns



```
In [ ]: 1 ...
```





What about trends in individual names?

In [37]: `1 baby[baby['Name'] == 'Olivia']`

Out[37]:

	Name	Sex	Count	Year
2	Olivia	F	16573	2022
13017	Olivia	M	16	2022
31917	Olivia	F	17798	2021
...	...	...	...	...
2079563	Olivia	F	52	1882
2081652	Olivia	F	51	1881
2083640	Olivia	F	44	1880

202 rows x 4 columns

In [36]: `1 baby[baby['Name'] == 'Olivia'].groupby('Year')['Count'].sum()`

Out[36]:

Year	Count
1880	44
1881	51
1882	52
...	...
2020	17677
2021	17887
2022	16589

Name: Count, Length: 143, dtype: int64

In []: `1 ...`



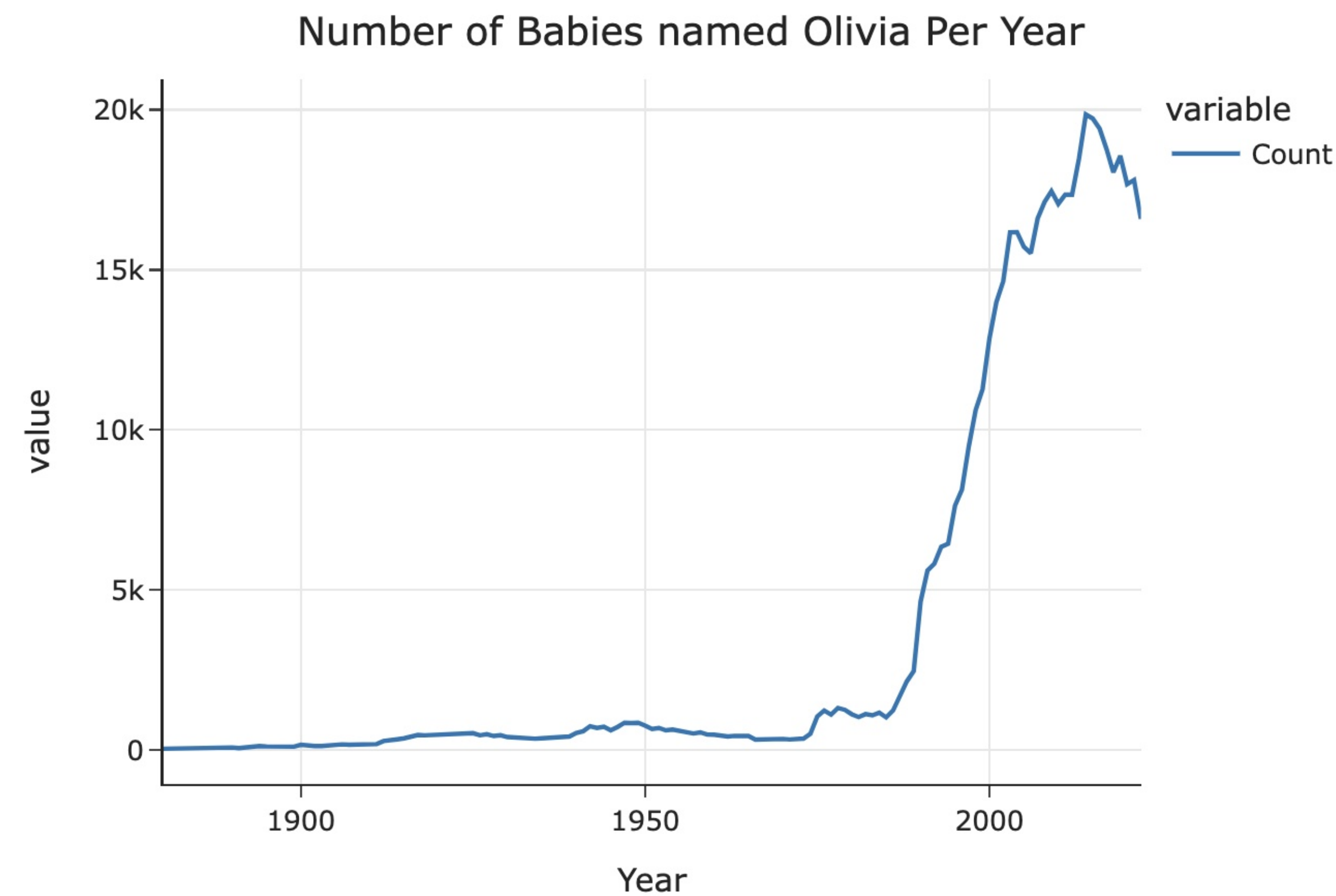


2081652	Olivia	F	51	1881
2083640	Olivia	F	44	1880

202 rows x 4 columns

In [39]:

```
1 (
2     baby[baby['Name'] == 'Olivia']
3     .groupby('Year')
4     ['Count']
5     .sum()
6     .plot(kind='line', title='Number of Babies named Olivia Per Year')
7 )
```



In []:

1 ...

