

# **PostgreSQL**

## **Aula 05**



# Presença

- Linktree: Presente na bio do nosso instagram
- Presença ficará disponível até 1 hora antes da próxima aula
- É necessário 70% de presença para obter o certificado



# Presença



# **Análise de Coorte**

## **Tópicos Principais**

1. Introdução à Análise de Coorte
2. Conceitos Fundamentais e Preparação
3. Calculando a Retenção com SQL



# **Introdução à Análise de Coorte**



# Relembrando - Séries Temporais

Sequência de dados obtidos num intervalo de tempo regular

- **Análise de tendências**
- Geração de relatórios e análises

# Relembrando - Séries Temporais

Sequência de dados obtidos num intervalo de tempo regular

- **Análise de tendências**
- Geração de relatórios e análises

No caso de uma análise de assinaturas ativas de uma plataforma, as séries temporais podem fornecer informações de tendências de queda/aumento, por exemplo

- **Problema:** entender o motivo da tendência
- **Possível Solução:** analisar a tendência de subgrupos de usuários (do mais antigo ao mais novo)



# Análise de Coorte

## Coorte

- Grupo de entidades que compartilham uma característica comum em um período de tempo
- Exemplo anterior
  - Conjunto de pessoas que iniciaram uma assinatura no mês X pertencem ao mesmo subgrupo/coorte
- Ajuda a entender o **porquê de uma tendência**
  - Em uma tendência geral de queda, são os assinantes mais antigos ou mais novos que estão saindo?



# Análise de Coorte

## Elementos de uma Análise de Coorte

- **Agrupamento de Coorte**
  - Característica de Agrupamento
  - Data de inscrição, primeira compra, cidade de residência, etc



# Análise de Coorte

## Elementos de uma Análise de Coorte

- **Agrupamento de Coorte**
  - Característica de Agrupamento
  - Data de inscrição, primeira compra, cidade de residência, etc
- **Série Temporal**
  - Representa os intervalos de tempo em que o comportamento das coortes será analisado.

# Análise de Coorte

## Elementos de uma Análise de Coorte

- **Agrupamento de Coorte**
  - Característica de Agrupamento
  - Data de inscrição, primeira compra, cidade de residência, etc
- **Série Temporal**
  - Representa os intervalos de tempo em que o comportamento das coortes será analisado.
- **Métrica de Avaliação**
  - Métrica de objetivo da análise



# Análise de Coorte

## Tipos de Análises

- **Retenção:** uma entidade retornou para realizar uma ação específica após X períodos de tempo da sua data inicial (tempo discreto)



# Análise de Coorte

## Tipos de Análises

- **Retenção:** uma entidade retornou para realizar uma ação específica após X períodos de tempo da sua data inicial (tempo discreto)
- Exemplo
  - Serviço de streaming registra um aumento no número de assinaturas mensais (**tendência da série temporal**)
  - Coorte por mês: porcentagem de assintantes ativos mais recentes da plataforma, após 1 e 2 meses, cai mais rápido do que os assinantes mais antigos (**Análise de Coorte**)



# Análise de Coorte

## Tipos de Análises

- **Sobrevivência:** Duração do tempo que uma entidade permanece ativa antes de um evento terminal, a partir da sua data inicial (tempo contínuo)

# Análise de Coorte

## Tipos de Análises

- **Sobrevivência:** Duração do tempo que uma entidade permanece ativa antes de um evento terminal, a partir da sua data inicial (tempo contínuo)
- Exemplo (serviço de streaming anterior)
  - Análise de tendências: taxa de cancelamento mensal constante nos últimos meses
  - Análise de coorte: coortes de meses mais recentes permanecem por menos tempo ativas do que coortes antigas



# Análise de Coorte

## Tipos de Análises

- **Retorno:** determina se uma entidade realizou uma ação por um número mínimo de vezes numa janela de tempo



# Análise de Coorte

## Tipos de Análises

- **Retorno:** determina se uma entidade realizou uma ação por um número mínimo de vezes numa janela de tempo
- Exemplo: para um comércio, fazendo um coorte por mês da primeira compra, podemos observar, depois de X dias, quantos usuários compraram novamente, assim classificando as coortes.



# Análise de Coorte

## Tipos de Análises

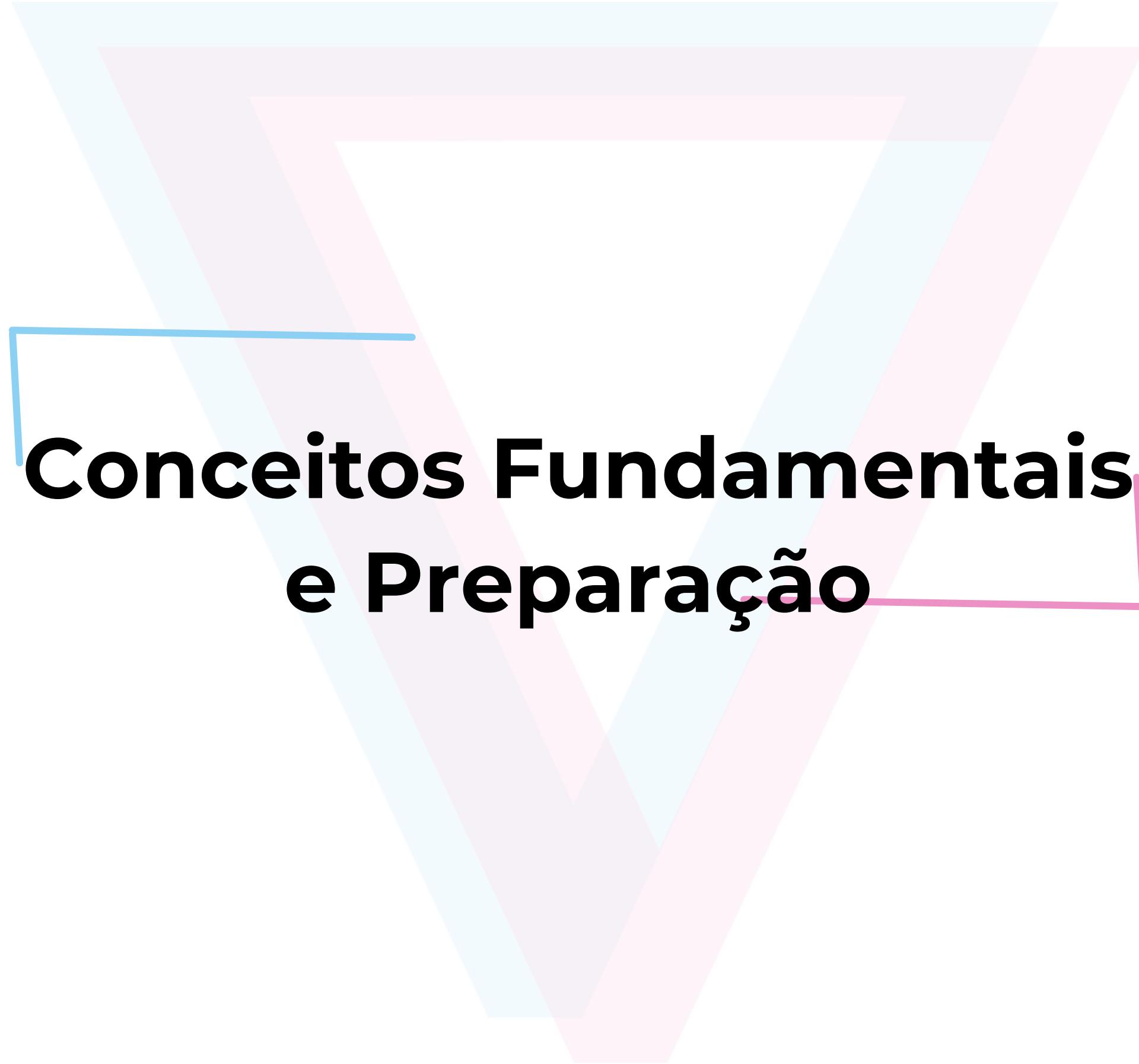
- **Cumulativo:** valores totais obtidos em uma estatística em janelas de tempo fixas



# Análise de Coorte

## Tipos de Análises

- **Cumulativo:** valores totais obtidos em uma estatística em janelas de tempo fixas
- Exemplo: valor gasto por jogadores que se registraram em um mês X (coorte) durante um período seguinte de Y dias (janela de tempo)



# **Conceitos Fundamentais e Preparação**

# O Conjunto de Dados

Nesta aula, usaremos o The Legislators Data Set, que contém informações sobre membros do Congresso dos EUA. Ela é dividida em duas tabelas que se conectam por um ID.

- legislators: Dados de cada legislador (como nome, gênero, etc).
- legislators\_terms: Dados dos mandatos (inicio/fim, partido, etc).

[https://github.com/cathytanimura/sql\\_book/tree/master/Chapter%204%3A%20Cohorts](https://github.com/cathytanimura/sql_book/tree/master/Chapter%204%3A%20Cohorts)

# O Conjunto de Dados

1	id_bioguide	term_number	term_id	term_type	term_start	term_end	state	district	class	party	how	url
2	B000944	0	B000944-0	rep	1993-01-05	1995-01-03	OH	13	NULL	Democrat	NULL	NULL
3	C000127	0	C000127-0	rep	1993-01-05	1995-01-03	WA	1	NULL	Democrat	NULL	NULL
4	C000141	0	C000141-0	rep	1987-01-06	1989-01-03	MD	3	NULL	Democrat	NULL	NULL
5	C000174	0	C000174-0	rep	1983-01-03	1985-01-03	DE	0	NULL	Democrat	NULL	NULL
6	C001070	0	C001070-0	sen	2007-01-04	2013-01-03	PA	NULL	1	Democrat	NULL	<a href="http://casey.senate.gov/">http://casey.senate.gov/</a>
7	F000062	0	F000062-0	sen	1992-11-10	1995-01-03	CA	NULL	1	Democrat	NULL	NULL
8	F000469	0	F000469-0	rep	2019-01-03	2021-01-03	ID	1	NULL	Republican	NULL	<a href="https://fulcher.house.gov">https://fulcher.house.gov</a>
9	K000367	0	K000367-0	sen	2007-01-04	2013-01-03	MN	NULL	1	Democrat	NULL	<a href="http://klobuchar.senate.gov">http://klobuchar.senate.gov</a>
10	M000639	0	M000639-0	rep	1993-01-05	1995-01-03	NJ	13	NULL	Democrat	NULL	NULL
11	S000033	0	S000033-0	rep	1991-01-03	1993-01-03	VT	0	NULL	Independent	NULL	NULL
12	S000770	0	S000770-0	rep	1997-01-07	1999-01-03	MI	8	NULL	Democrat	NULL	NULL
13	T000464	0	T000464-0	sen	2007-01-04	2013-01-03	MT	NULL	1	Democrat	NULL	<a href="http://tester.senate.gov">http://tester.senate.gov</a>
14	W000802	0	W000802-0	sen	2007-01-04	2013-01-03	RI	NULL	1	Democrat	NULL	<a href="http://whitehouse.senate.gov">http://whitehouse.senate.gov</a>
15	B001300	0	B001300-0	rep	2017-01-03	2019-01-03	CA	44	NULL	Democrat	NULL	<a href="https://barragan.house.gov">https://barragan.house.gov</a>
16	B001261	0	B001261-0	sen	2007-06-25	2013-01-03	WY	NULL	1	Republican	appointment	<a href="http://barrasso.senate.gov">http://barrasso.senate.gov</a>
17	W000437	0	W000437-0	rep	1995-01-04	1997-01-03	MS	1	NULL	Republican	NULL	NULL
18	A000360	0	A000360-0	sen	2003-01-07	2009-01-03	TN	NULL	2	Republican	NULL	<a href="http://alexander.senate.gov">http://alexander.senate.gov</a>
19	C001035	0	C001035-0	sen	1997-01-07	2003-01-03	ME	NULL	2	Republican	NULL	NULL
20	C001056	0	C001056-0	sen	2002-11-30	2003-01-03	TX	NULL	2	Republican	NULL	NULL
21	D000563	0	D000563-0	rep	1983-01-03	1985-01-03	IL	20	NULL	Democrat	NULL	NULL
22	S001194	0	S001194-0	sen	2012-12-27	2015-01-03	HI	NULL	3	Democrat	appointment	<a href="http://www.schatz.senate.gov">http://www.schatz.senate.gov</a>

# Análise de Retenção



# Retenção

A principal métrica da análise será a retenção. O objetivo é medir quantos membros de um grupo inicial, chamado de coorte, continuam "ativos" com o passar do tempo.



# Análise de Retenção

Queremos identificar quantos dos representantes e legisladores que ingressaram a partir de uma determinada data (nossa coorte) ainda permanecem em exercício de mandato.

# Análise de Retenção

```
SELECT  
    id_bioguide,  
    MIN(term_start) AS first_term  
FROM legislators_terms  
GROUP BY id_bioguide;
```

# Análise de Retenção

	<code>id_bioguide</code> character varying	<code>first_term</code> date
1	A000118	1975-01-14
2	P000281	1933-03-09
3	K000039	1933-03-09
4	A000306	1907-12-02
5	0000095	1949-01-03
6	B000937	1913-04-07
7	S000038	1912-01-01

# Análise de Retenção

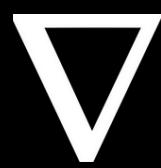
```
SELECT
    DATE_PART('year', AGE(l.term_start, b.first_term)) AS period,
    COUNT(DISTINCT l.id_bioguide) AS cohort_retained
FROM (
    SELECT
        id_bioguide,
        MIN(term_start) AS first_term
    FROM legislators_terms
    GROUP BY id_bioguide
) b
JOIN legislators_terms AS l
ON l.id_bioguide = b.id_bioguide
GROUP BY period;
```

# Análise de Retenção

	period double precision 	cohort_retained bigint 
1	0	12518
2	1	3600
3	2	3619
4	3	1831
5	4	3210
6	5	1744

# Análise de Retenção

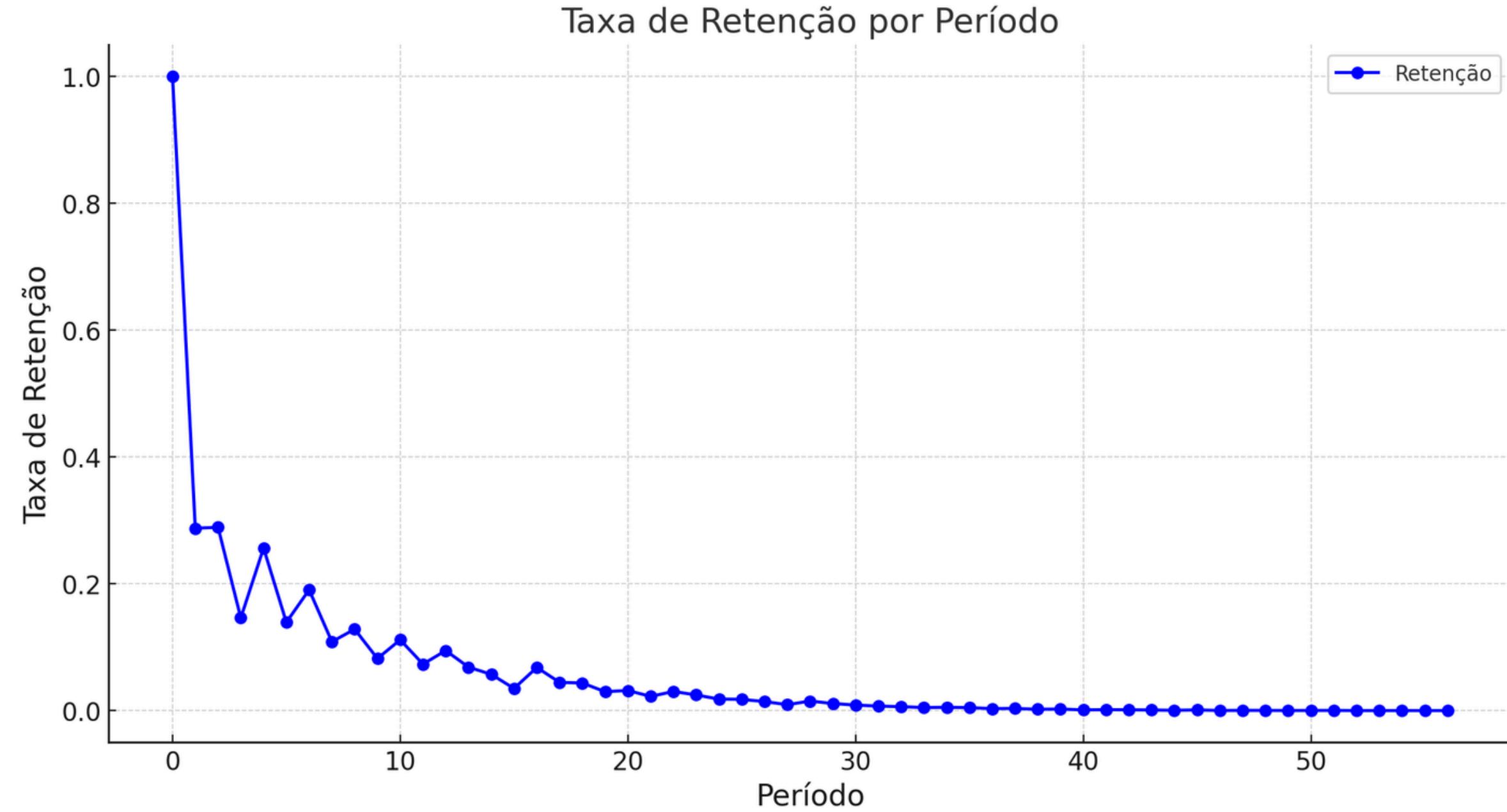
```
SELECT
    period,
    FIRST_VALUE(cohort_retained) OVER (ORDER BY period) AS cohort_size,
    cohort_retained,
    cohort_retained::decimal /
    FIRST_VALUE(cohort_retained) OVER (ORDER BY period) AS pct_retained
FROM (
    SELECT
        DATE_PART('year', AGE(l.term_start, b.first_term)) AS period,
        COUNT(DISTINCT l.id_bioguide) AS cohort_retained
    FROM (
        SELECT
            id_bioguide,
            MIN(term_start) AS first_term
        FROM legislators_terms
        GROUP BY id_bioguide
    ) b
    JOIN legislators_terms AS l
    ON l.id_bioguide = b.id_bioguide
    GROUP BY period
) sub
ORDER BY period.
```



# Análise de Retenção

	period double precision	cohort_size bigint	cohort_retained bigint	pct_retained numeric
1	0	12518	12518	1.00000000000000000000000000000000
2	1	12518	3600	0.28758587633807317463
3	2	12518	3619	0.28910369068541300527
4	3	12518	1831	0.14626937210416999521
5	4	12518	3210	0.25643073973478191404
6	5	12518	1744	0.13931938009266656015
7	6	12518	2385	0.19052564307397347819
8	7	12518	1360	0.10864355328327208819
9	8	12518	1607	0.12837513979868988656

# Análise de Retenção





# Análise de Retenção

Atualmente um legislador apenas faz parte de um período se ele foi eleito naquele período, porém queremos que ele faça parte para todo ano que estava em mandato, como corrigir?

# Análise de Retenção

- Criar uma tabela com uma linha por ano
- Fazer um join entre os mandatos dos legisladores e essa tabela de datas.
- Para cada ano, verificamos se o legislador ainda estava em mandato 31 de dezembro.
- A partir disso, conseguimos marcar presença ativa ano a ano, mesmo que ele não tenha iniciado um mandato novo naquele ano.

# Análise de Retenção

```
CREATE TABLE date_dim (
    date DATE PRIMARY KEY
);
INSERT INTO date_dim (date)
SELECT
    MAKE_DATE(ano, 12, 31)
FROM
    generate_series(
        1701,
        EXTRACT(YEAR FROM NOW())::integer
    ) AS ano
```

# Análise de Retenção

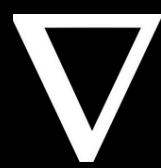
```
1  SELECT
2      DATE_PART('year', AGE(d.date, b.first_term)) AS period,
3      COUNT(DISTINCT l.id_bioguide) AS cohort_retained
4  FROM (
5      SELECT
6          id_bioguide,
7          MIN(term_start) AS first_term
8      FROM legislators_terms
9      GROUP BY id_bioguide
10     ) b
11 JOIN legislators_terms AS l
12     ON l.id_bioguide = b.id_bioguide
13 JOIN date_dim as d
14     ON d.date between l.term_start and l.term_end
15 GROUP BY 1;
```

# Análise de Retenção

	period double precision 	cohort_retained bigint 
1	0	12501
2	1	12328
3	2	8166
4	3	8069
5	4	5862
6	5	5795
7	6	4361
8	7	4339
9	8	3521

# Análise de Retenção

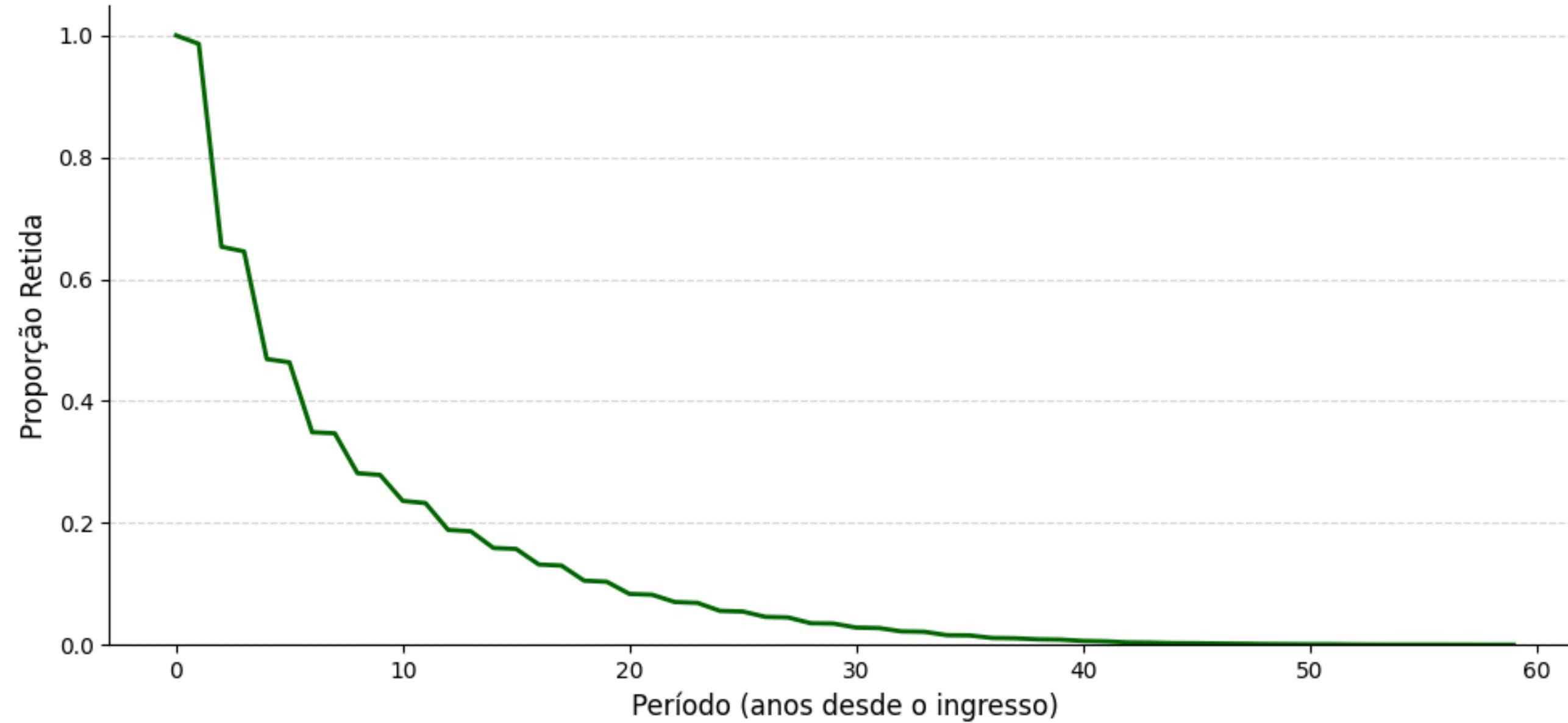
```
1  SELECT
2      period,
3      FIRST_VALUE(cohort_retained) OVER (ORDER BY period) AS cohort_size,
4      cohort_retained,
5      cohort_retained::decimal
6      / FIRST_VALUE(cohort_retained) OVER (ORDER BY period) AS pct_retained
7  FROM (
8      SELECT
9          DATE_PART('year', AGE(d.date, b.first_term)) AS period,
10         COUNT(DISTINCT l.id_bioguide) AS cohort_retained
11     FROM (
12         SELECT
13             id_bioguide,
14             MIN(term_start) AS first_term
15         FROM legislators_terms
16         GROUP BY id_bioguide
17     ) b
18     JOIN legislators_terms AS l
19     ON l.id_bioguide = b.id_bioguide
20     JOIN date_dim AS d
21     ON d.date between l.term_start and l.term_end
22     GROUP BY 1
23 ) sub
24 ORDER BY 1
25 ;
```



# Análise de Retenção

# Análise de Retenção

**Retenção de Legisladores por Período**





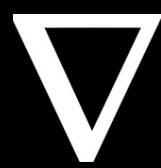
# Análise de Retenção

E as coortes? Como usá-las na nossa análise de retenção?

Basta agrupar nossa rentenção pelo tipo de coorte que queremos

# Análise de Retenção

```
SELECT
    start_year,
    period,
    FIRST_VALUE(cohort_retained) OVER (PARTITION BY start_year ORDER BY period) AS cohort_size,
    cohort_retained,
    cohort_retained::decimal
    / FIRST_VALUE(cohort_retained) OVER (PARTITION BY start_year ORDER BY period) AS pct_retained
FROM (
    SELECT
        DATE_PART('year', b.first_term) AS start_year
        ,COALESCE(DATE_PART('year', AGE(d.date, b.first_term)), 0) AS period,
        COUNT(DISTINCT l.id_bioguide) AS cohort_retained
    FROM (
        SELECT
            id_bioguide,
            MIN(term_start) AS first_term
        FROM legislators_terms
        GROUP BY id_bioguide
    ) b
    JOIN legislators_terms AS l
        ON l.id_bioguide = b.id_bioguide
    LEFT JOIN date_dim AS d
        ON d.date between l.term_start and l.term_end
    GROUP BY 1, 2
) sub
ORDER BY 1, 2
;
```



# Análise de Retenção



# Análise de Retenção

Temos mais de 200 anos, o que torna o número de coortes muito extenso, como resolver? Escolher uma coorte maior

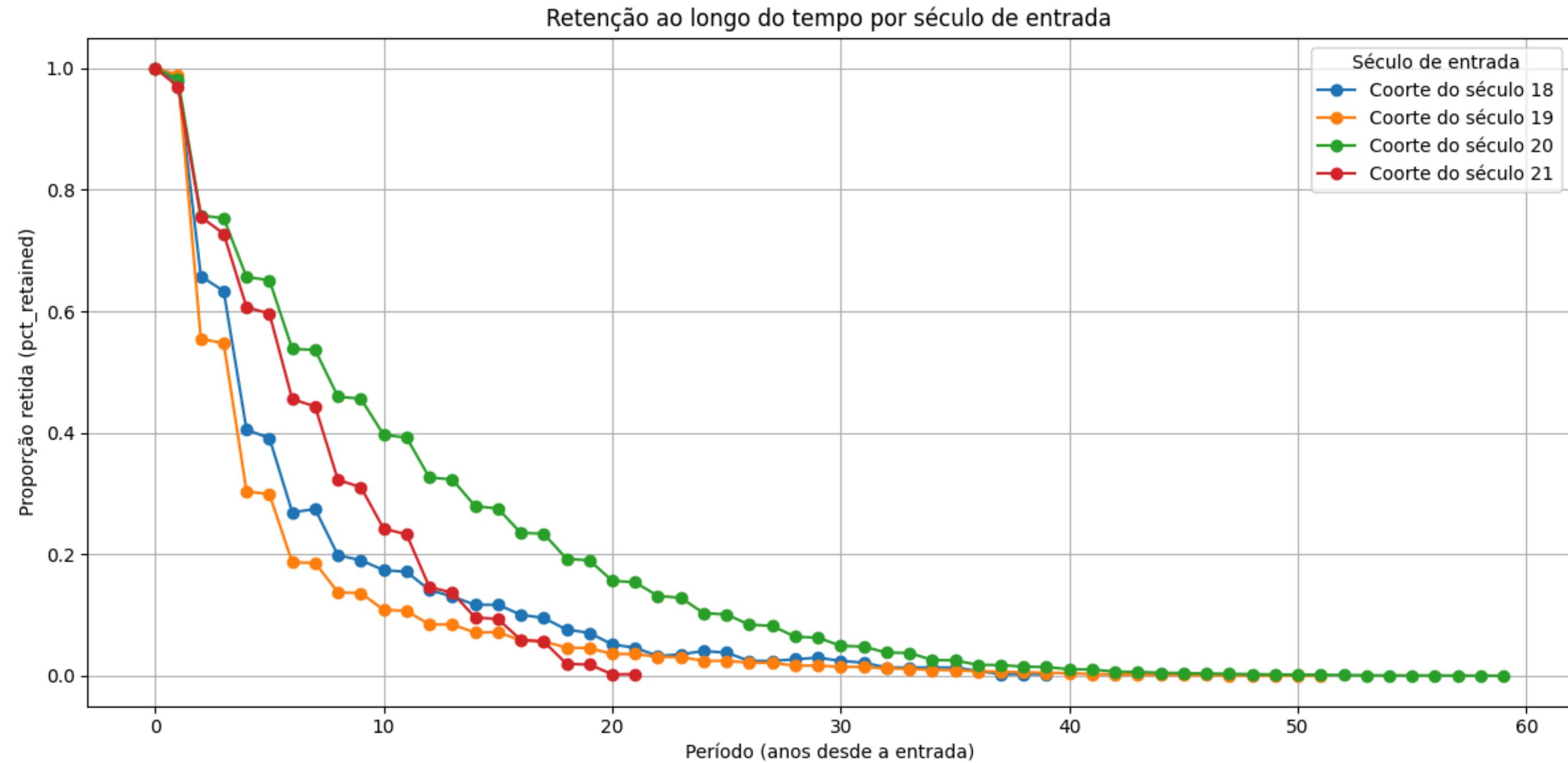
# Análise de Retenção

```
1 ▾ SELECT
2     start_century,
3     period,
4     FIRST_VALUE(cohort_retained) OVER (PARTITION BY start_century ORDER BY period) AS cohort_size,
5     cohort_retained,
6     cohort_retained::decimal
7     / FIRST_VALUE(cohort_retained) OVER (PARTITION BY start_century ORDER BY period) AS pct_retained
8 FROM (
9     SELECT
10         DATE_PART('century', b.first_term) AS start_century
11         ,COALESCE(DATE_PART('year', AGE(d.date, b.first_term)), 0) AS period,
12         COUNT(DISTINCT l.id_bioguide) AS cohort_retained
13     FROM (
14         SELECT
15             id_bioguide,
16             MIN(term_start) AS first_term
17             FROM legislators_terms
18             GROUP BY id_bioguide
19         ) b
20     JOIN legislators_terms AS l
21         ON l.id_bioguide = b.id_bioguide
22     LEFT JOIN date_dim AS d
23         ON d.date between l.term_start and l.term_end
24     GROUP BY 1, 2
25 ) sub
26 ORDER BY 1, 2
27 ;
```

# Análise de Retenção

	start_century double precision	period double precision	cohort_size bigint	cohort_retained bigint	pct_retained numeric
31	18	30	368	9	0.02445652173913043478
32	18	31	368	8	0.02173913043478260870
33	18	32	368	5	0.01358695652173913043
34	18	33	368	5	0.01358695652173913043
35	18	34	368	5	0.01358695652173913043
36	18	35	368	5	0.01358695652173913043
37	18	36	368	3	0.00815217391304347826
38	18	37	368	1	0.00271739130434782609
39	18	38	368	1	0.00271739130434782609
40	18	39	368	1	0.00271739130434782609
41	19	0	6299	6299	1.00000000000000000000
42	19	1	6299	6231	0.98920463565645340530
43	19	2	6299	3492	0.55437371011271630418
44	19	3	6299	3449	0.54754722971900301635
45	19	4	6299	1911	0.30338148912525797746
46	19	5	6299	1883	0.29893633910144467376

# Análise de Retenção



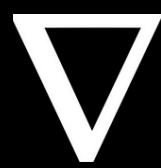


# Análise de Retenção

Até agora as coortes são baseadas em data, podemos fazer coortes a partir de outras características?

# Análise de Retenção

```
1 ▾ SELECT
2     term_state,
3     period,
4     FIRST_VALUE(cohort_retained) OVER (PARTITION BY term_state ORDER BY period) AS cohort_size,
5     cohort_retained,
6     cohort_retained::decimal
7     / FIRST_VALUE(cohort_retained) OVER (PARTITION BY term_state ORDER BY period) AS pct_retained
8 FROM (
9     SELECT
10        l.state AS term_state
11        ,COALESCE(DATE_PART('year', AGE(d.date, b.first_term)), 0) AS period,
12        COUNT(DISTINCT l.id_bioguide) AS cohort_retained
13    FROM (
14        SELECT
15            id_bioguide,
16            MIN(term_start) AS first_term
17        FROM legislators_terms
18        GROUP BY id_bioguide
19    ) b
20    JOIN legislators_terms AS l
21        ON l.id_bioguide = b.id_bioguide
22    LEFT JOIN date_dim AS d
23        ON d.date BETWEEN l.term_start AND l.term_end
24    GROUP BY 1, 2
25 ) sub
26 ORDER BY 1, 2
27 ;
28 ;
```



# Análise de Retenção



# Análise de Retenção

E se a informação da coorte está em uma tabela separada da série temporal?

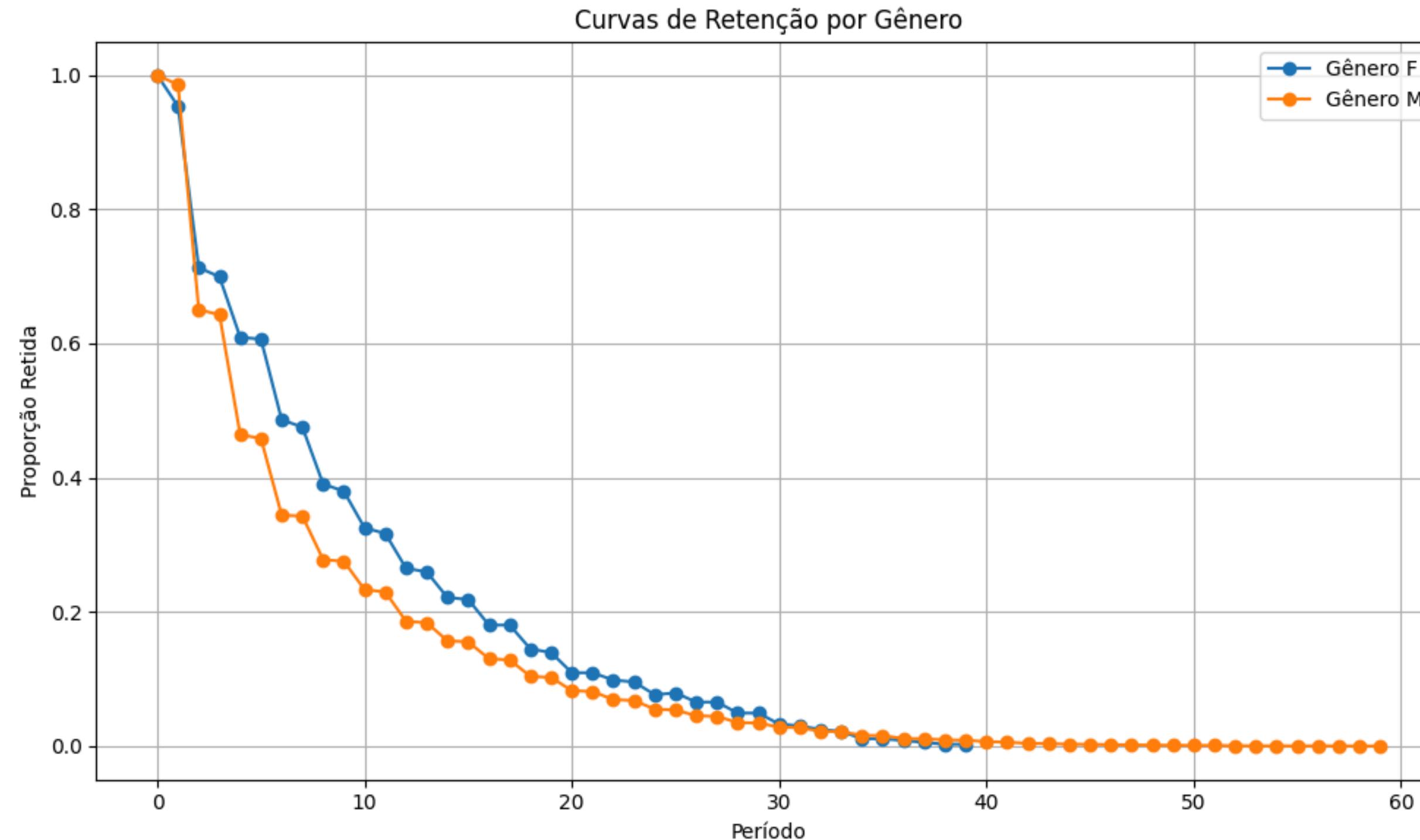
# Análise de Retenção

```
1  SELECT
2      gender,
3      period,
4      FIRST_VALUE(cohort_retained) OVER (PARTITION BY gender ORDER BY period) AS cohort_size,
5      cohort_retained,
6      cohort_retained::decimal
7          / FIRST_VALUE(cohort_retained) OVER (PARTITION BY gender ORDER BY period) AS pct_retained
8  FROM (
9      SELECT
10         leg.gender,
11         COALESCE(DATE_PART('year', AGE(d.date, b.first_term)), 0) AS period,
12         COUNT(DISTINCT ltrs.id_bioguide) AS cohort_retained
13     FROM (
14         SELECT
15             id_bioguide,
16             MIN(term_start) AS first_term
17         FROM legislators_terms
18         GROUP BY id_bioguide
19     ) b
20     JOIN legislators_terms ltrs
21         ON ltrs.id_bioguide = b.id_bioguide
22     JOIN legislators leg
23         ON leg.id_bioguide = b.id_bioguide
24     LEFT JOIN date_dim d
25         ON d.date BETWEEN ltrs.term_start AND ltrs.term_end
26     GROUP BY leg.gender, period
27 ) sub
28 ORDER BY gender, period;
```

# Análise de Retenção

	gender character varying 	period double precision 	cohort_size bigint 	cohort_retained bigint 	pct_retained numeric 
1	F	0	366	366	1.0000000000000000000000000000000
2	F	1	366	349	0.95355191256830601093
3	F	2	366	261	0.71311475409836065574
4	F	3	366	256	0.69945355191256830601
5	F	4	366	223	0.60928961748633879781
6	F	5	366	222	0.60655737704918032787
7	F	6	366	178	0.48633879781420765027
8	F	7	366	174	0.47540983606557377049
9	F	8	366	143	0.39071038251366120219
10	F	9	366	139	0.37978142076502732240
11	F	10	366	119	0.32513661202185792350
12	F	11	366	116	0.31693989071038251366
13	F	12	366	97	0.26502732240437158470
14	F	13	366	95	0.25956284153005464481
15	F	14	366	81	0.22131147540983606557
16	F	15	366	80	0.21857923497267759563

# Análise de Retenção





# Análise de Retenção

Mulheres permanecem mais tempo em seus cargos, mas será que é verdade?



# Análise de Retenção

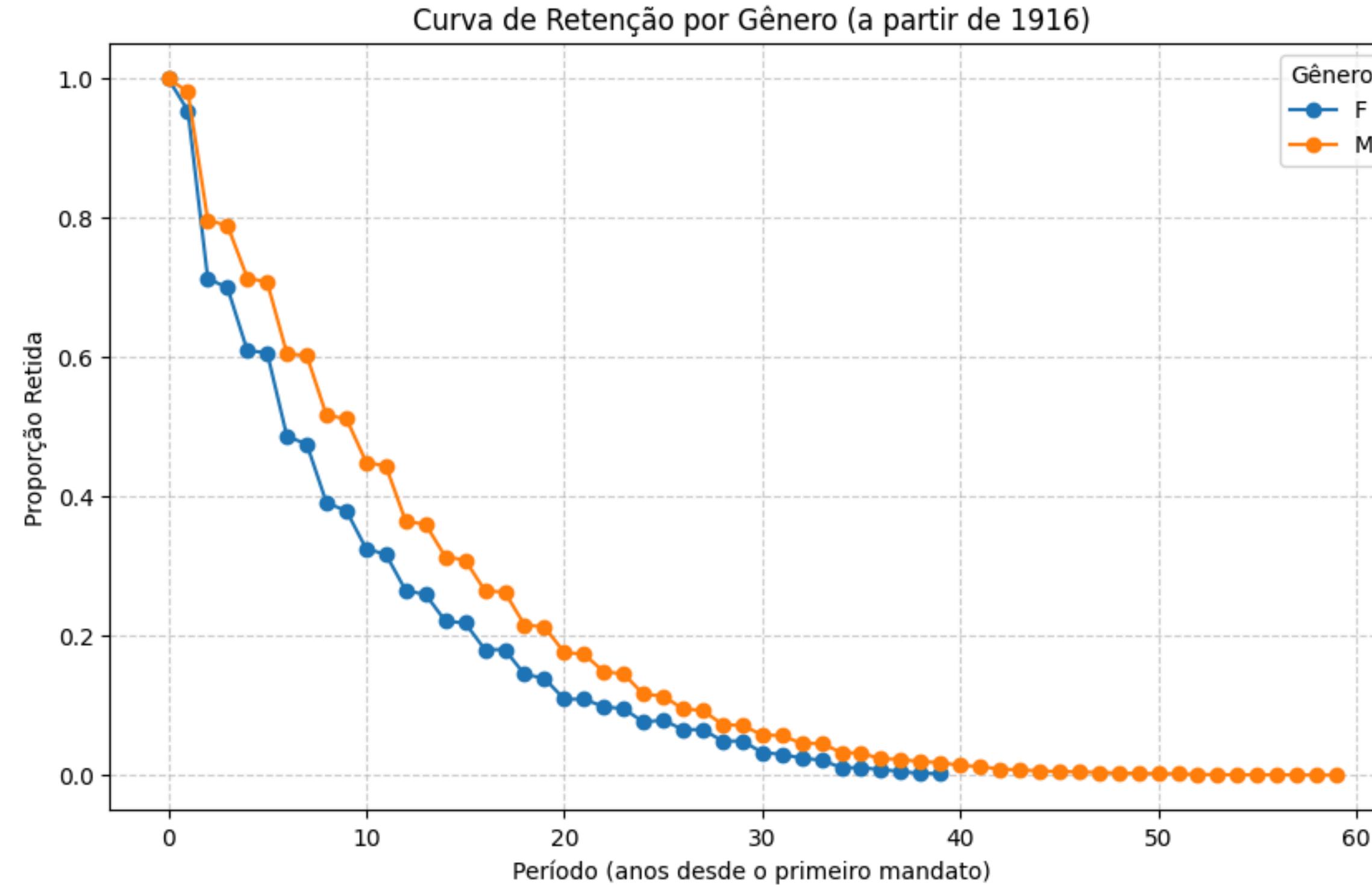
Mulheres permanecem mais tempo em seus cargos, mas será que é verdade?

No século 20 a retenção aumentou muito e as primeiras mulheres foram eleitas justamente neste século. Vamos filtrar nossos dados a partir da primeira mulher eleita e avaliar novamente a retenção por gênero.

# Análise de Retenção

```
1  SELECT
2      gender,
3      period,
4      FIRST_VALUE(cohort_retained) OVER (PARTITION BY gender ORDER BY period) AS cohort_size,
5      cohort_retained,
6      cohort_retained::decimal
7          / FIRST_VALUE(cohort_retained) OVER (PARTITION BY gender ORDER BY period) AS pct_retained
8  FROM (
9      SELECT
10         leg.gender,
11         COALESCE(DATE_PART('year', AGE(d.date, b.first_term)), 0) AS period,
12         COUNT(DISTINCT ltrs.id_bioguide) AS cohort_retained
13     FROM (
14         SELECT
15             id_bioguide,
16             MIN(term_start) AS first_term
17         FROM legislators_terms
18         GROUP BY id_bioguide
19     ) b
20     JOIN legislators_terms ltrs
21         ON ltrs.id_bioguide = b.id_bioguide
22     JOIN legislators leg
23         ON leg.id_bioguide = b.id_bioguide
24     LEFT JOIN date_dim d
25         ON d.date BETWEEN ltrs.term_start AND ltrs.term_end
26     WHERE term_start >= '1916-01-01'
27     GROUP BY leg.gender, period
28 ) sub
29 ORDER BY gender, period;
```

# Análise de Retenção





[data@icmc.usp.br](mailto:data@icmc.usp.br)



[@data.icmc](https://www.instagram.com/@data.icmc)



[/c/DataICMC](https://www.youtube.com/c/DataICMC)



[/icmc-data](https://github.com/icmc-data)



[data.icmc.usp.br](http://data.icmc.usp.br)

|| obrigado!