

Aula 5

Aula 5: Análise de Coorte

1. Introdução à Análise de Coorte

Na última aula, abordamos o conceito de séries temporais e como poderíamos utilizar dados de tempo para determinar tendências e realizar análises. É a partir dessa ideia que podemos, por exemplo, determinar quantos usuários ativos uma plataforma possui em cada mês de um ano, identificando, em seguida, quedas e aumentos. Contudo, observe que essa é uma visão macro da estatística: ela apenas informa o comportamento geral de uma tendência, não auxiliando a entender o porquê disso. Para tanto, uma abordagem seria analisar quão ativo é um grupo que começou na plataforma no mês X durante alguns meses, buscando analisar se a tendência se deve a alguns grupos específicos. É com base nisso que surge a análise de Coorte.

Um coorte é um grupo de entidades (indivíduos, empresas, produtos, etc.) que compartilham uma característica comum em um período de tempo. No exemplo anterior, um conjunto de pessoas que começaram na plataforma em um mês específico (janeiro, por exemplo) é um coorte. Em uma análise de coorte, buscamos entender como cada grupo se comporta.

No geral, uma análise coorte possui três elementos:

- **Agrupamento de Coorte:** a junção de entidades, em geral, é feita com base na data inicial de uma característica, como data de primeira compra, data de inscrição e data de início de um curso. Em outros casos, podemos usar características intrínsecas aos dados (data de aniversário, ano de fundação da empresa) e características que se alteram com o tempo (cidade de residência).
- **Série Temporal:** dados no intervalo de tempo de análise. Se os agrupamento de coorte são feitos com base em meses, a série temporal deverá ter diversos meses.
- **Métrica de Agregação:** métrica de objetivo para análise dos dados, como continuidade de compra de produtos por consumidores.

A análise de coorte é importante na medida em que permite verificar o comportamento de subgrupos de entidades em torno de uma tendência. No exemplo inicial, somente a análise de tendências não é capaz de explicar a razão de uma queda ou de um crescimento. Uma análise de coorte permitiria, por exemplo, observar se a tendência se deu por conta de movimentações de usuários mais antigos ou mais novos.

As análises de coorte podem ser utilizadas em diferentes situações. Dentre os tipos de análises de coorte, podemos citar:

- **Retenção:** ajuda a entender se uma entidade do coorte está presente numa data, a qual é estipulada depois de X períodos de tempo da data inicial. Isso pode ser relevante para entender como subgrupos participam de atividades de negócio repetitivas, como assinaturas de plataformas.
- **Sobrevivência:** relacionada a quantas entidades permanecem ativas por um período de tempo determinado ou mais.
- **Retorno:** relacionado a determinar se uma ação foi realizada por um número mínimo de vezes em um janela fixa de tempo. Pode ser importante em situações de comportamento imprevisível.
- **Cumulativo:** relacionados ao número total ou aos valores totais medidos em janelas de tempo fixas.

2. Conceitos Fundamentais e Preparação

Para os exemplos práticos desta aula, utilizaremos o The Legislators Data Set, um conjunto de dados que contém informações sobre membros passados e presentes do Congresso dos Estados Unidos. Os dados vieram do livro que usamos de base (SQL for Data Analysis, Cathy Tanimura), disponíveis no github: https://github.com/cathytanimura/sql_book/tree/master/Chapter%204%3A%20Cohorts. Os arquivos CSV podem ser importados para o PostgreSQL usando o pgAdmin4.

A tabela *legislators* contém uma lista de todas as pessoas no conjunto de dados, com informações como data de nascimento (*birthday*), gênero (*gender*) e um conjunto de campos de identificação. A tabela *legislators_terms* armazena um registro para cada mandato de um legislador, incluindo a data de início e fim, a câmara a que pertenceu (Senado ou Câmara de Representantes) e seu partido político. Entre eles, há uma coluna *id_bioguide* usada como o identificador único para cada legislador, permitindo a conexão entre as duas tabelas.

A principal métrica que exploraremos nesta análise é a retenção. Em essência, a retenção mede quantos membros de uma coorte permanecem "ativos" ao longo de sucessivos períodos de tempo. Esse tipo de métrica permite analisar, por exemplo, a porcentagem de legisladores que começaram seus mandatos em determinado ano, e que ainda estão ativos depois de digamos 3 anos. Outro exemplo seria ver quantos clientes que se assinaram meu serviço em uma promoção continuam usando ele depois de um certo período de tempo.

3. Calculando a Retenção com SQL

Nesta seção, vamos construir a análise de retenção passo a passo usando SQL. Começaremos com uma abordagem inicial, identificaremos suas falhas e as corrigiremos para, em seguida, segmentar nossa análise em diferentes coortes.

SQL para uma Curva de Retenção Básica

O primeiro passo é identificar quando cada legislador iniciou sua trajetória no Congresso. Para isso, precisamos encontrar a data do **primeiro mandato** de cada indivíduo. Isso nos dá a *linha de partida* de cada coorte.

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

Essa consulta agrupa todos os mandatos por legislador e retorna apenas a menor data de início. Assim, criamos uma tabela auxiliar com a data do primeiro mandato de cada pessoa.



	id_bioguide character varying 	first_term date 
1	A000118	1975-01-14
2	P000281	1933-03-09
3	K000039	1933-03-09
4	A000306	1907-12-02
5	O000095	1949-01-03
6	B000937	1913-04-07
7	S000038	1912-01-01

Figura 1: Tabela de primeira eleição de cada legislador.

Criamos a base de coorte definindo o ano de entrada de cada legislador. Em seguida, calculamos a retenção: queremos saber **quantos legisladores permaneceram em atividade ao longo dos anos**. Para isso, comparamos o ano de cada mandato com o ano do primeiro mandato e contamos o número de legisladores distintos em cada período. Usamos a função de janela *FIRST_VALUE* para capturar o tamanho da coorte no início e calcular a porcentagem de retenção em cada período.

```

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
    ) AS b
    JOIN legislators_terms AS l
    ON l.id_bioguide = b.id_bioguide
    GROUP BY period
    ) AS sub
ORDER BY period;

```

	period double precision 🔒	cohort_size bigint 🔒	cohort_retained bigint 🔒	pct_retained numeric 🔒
1	0	12518	12518	1.00000000000000000000
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

Figura 2: Tabela de retenção inicial.

Passo Explícito: Contamos os legisladores ativos em cada período e dividimos pelo tamanho inicial da coorte para obter a taxa de retenção.

O resultado é uma curva de retenção inicial. Contudo, essa curva apresenta oscilações bruscas, pois um legislador só aparece no período em que iniciou um novo mandato, não em todos os anos em que efetivamente estava ativo. Precisamos então corrigir essa limitação.

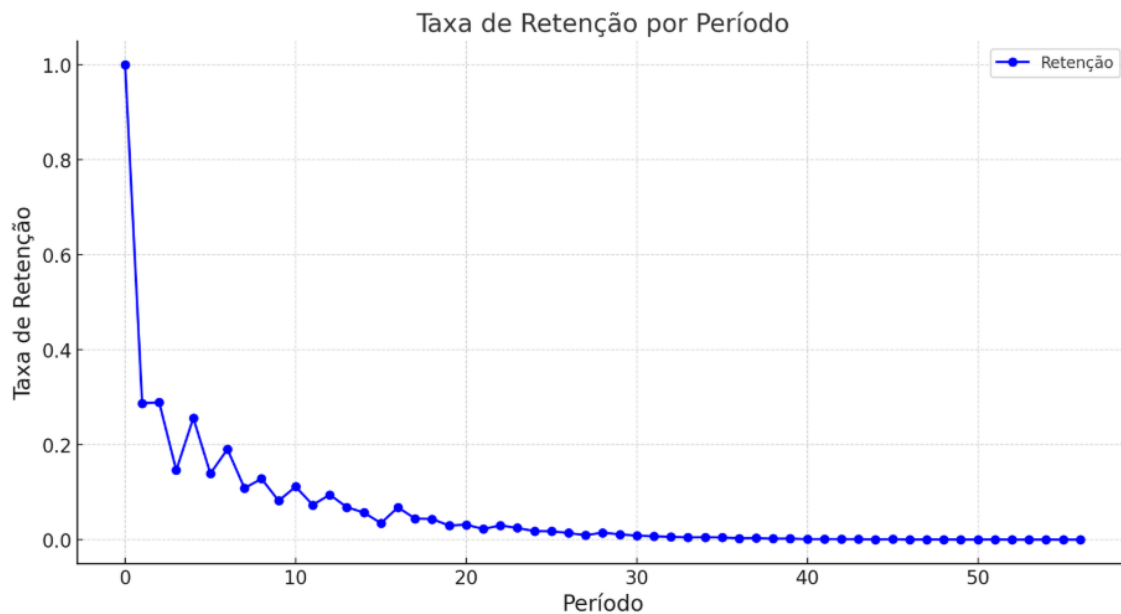


Figura 3: Retenção.

Lidando com Coortes Esparsas

Para resolver o problema, criamos uma tabela de **dimensão de datas**, que contém um registro para cada ano. Isso nos permite verificar se um legislador estava em mandato ativo mesmo que não tenha iniciado um novo mandato naquele ano específico.

```
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;
```

Construímos uma tabela de datas que gera um marco temporal anual para verificar atividade contínua de cada legislador.

A ideia é simples: cada linha representa o último dia de um ano. Assim, podemos comparar se a data de referência de cada ano estava dentro do intervalo de mandato de um legislador.

Para isso, usamos um `LEFT JOIN` com a condição:

```
LEFT JOIN date_dim AS d
ON d.date BETWEEN l.term_start AND l.term_end
```

Garantimos que o legislador seja contado em todos os anos de atividade, mesmo sem novos mandatos.

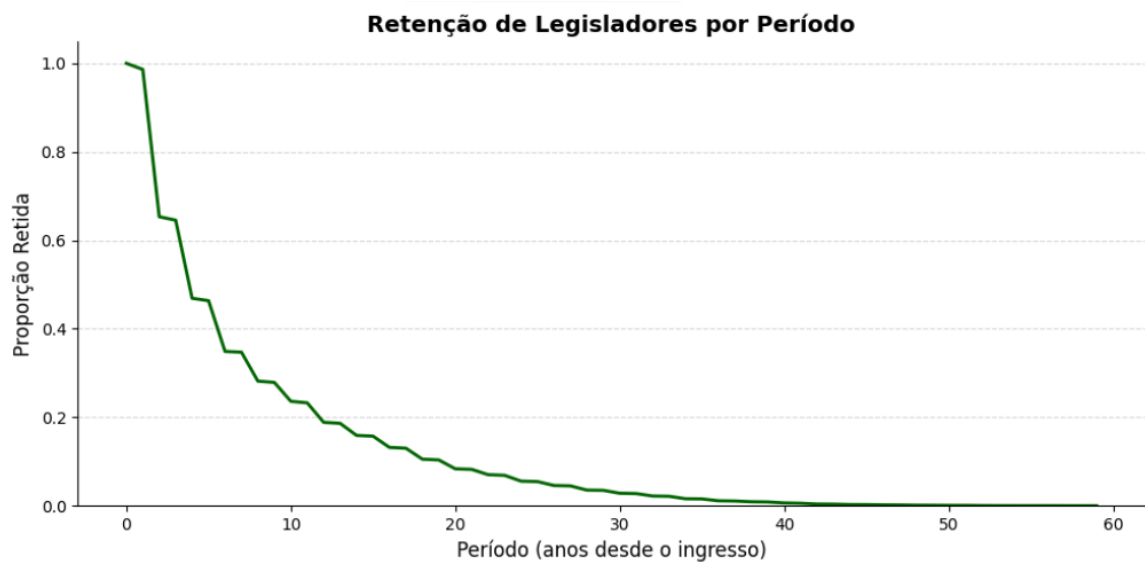


Figura 4: Retenção corrigida.

	period double precision	cohort_size bigint	cohort_retained bigint	pct_retained numeric
1	0	12501	12501	1.00000000000000000000
2	1	12501	12328	0.98616110711143108551
3	2	12501	8166	0.65322774178065754740
4	3	12501	8069	0.64546836253099752020
5	4	12501	5862	0.46892248620110391169
6	5	12501	5795	0.46356291496680265579
7	6	12501	4361	0.34885209183265338773
8	7	12501	4339	0.34709223262139028878
9	8	12501	3521	0.28165746740260779138
10	9	12501	3485	0.27877769778417726582
11	10	12501	2952	0.23614110871130309575
12	11	12501	2908	0.23262139028877689785

Figura 5: Tabela de retenção corrigida.

Definindo Coortes pelo Século de Entrada

Uma vez corrigida a métrica de retenção, podemos **segmentar** os legisladores em coortes. Um critério natural é a época em que começaram seus mandatos. No exemplo abaixo, usamos o século de entrada:

```
SELECT
  start_century,
  period,
  FIRST_VALUE(cohort_retained) OVER (
    PARTITION BY start_century ORDER BY period
```

```

) AS cohort_size,
cohort_retained,
cohort_retained::decimal /
FIRST_VALUE(cohort_retained) OVER (
PARTITION BY start_century ORDER BY period
) AS pct_retained
FROM (
SELECT
DATE_PART('century', b.first_term) AS start_century,
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
) AS 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
) AS sub
ORDER BY 1, 2;

```

Passo Explícito: Segmentamos os legisladores em coortes por século de entrada e comparamos padrões de retenção ao longo do tempo.

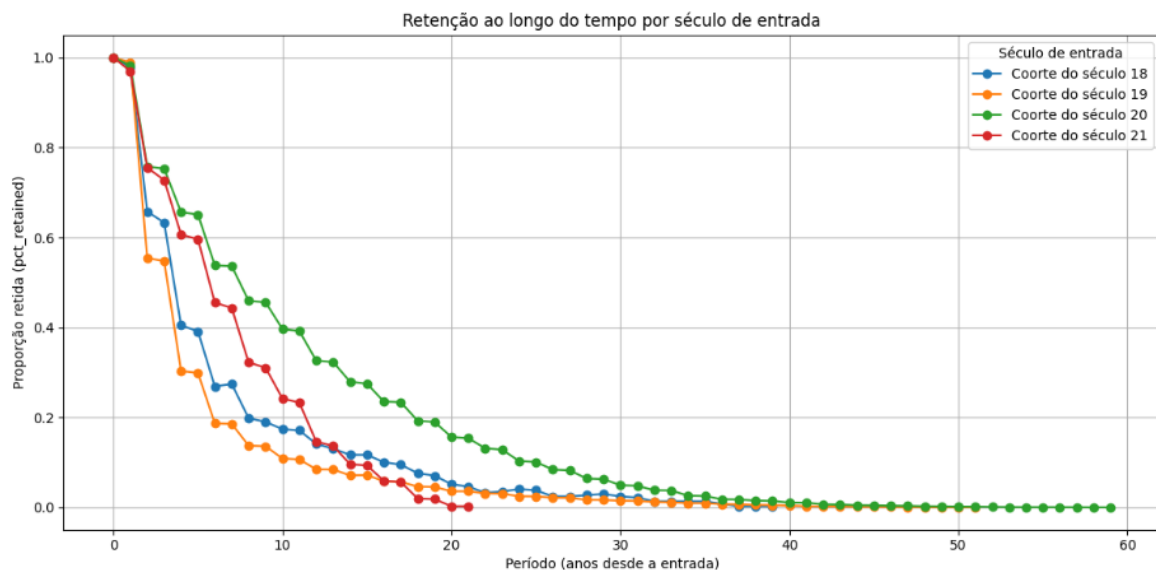


Figura 6: Retenção ao longo do tempo por século de entrada.

Definindo Coortes por Estado

Outra forma de análise é segmentar as coortes a partir de **atributos dos dados**, como o estado que o legislador representou. Assim, podemos identificar diferenças regionais nos padrões de retenção.

```
SELECT
    term_state,
    period,
    FIRST_VALUE(cohort_retained) OVER (
        PARTITION BY term_state ORDER BY period
    ) AS cohort_size,
    cohort_retained,
    cohort_retained::decimal /
    FIRST_VALUE(cohort_retained) OVER (
        PARTITION BY term_state ORDER BY period
    ) AS pct_retained
FROM (
    SELECT
        l.state AS term_state,
        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
    ) AS 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
) AS sub
ORDER BY 1, 2;
```

Comparamos a retenção entre estados, observando variações regionais no tempo.

Esse tipo de análise nos permite responder perguntas como: “*A retenção dos legisladores é mais alta em alguns estados do que em outros?*”. Dessa forma, a análise de coortes se torna uma ferramenta poderosa para explorar padrões específicos em diferentes dimensões dos dados.

	term_state character varying	period double precision	cohort_size bigint	cohort_retained bigint	pct_retained numeric
1	AK	0	19	19	1.00000000000000000000
2	AK	1	19	19	1.00000000000000000000
3	AK	2	19	15	0.78947368421052631579
4	AK	3	19	15	0.78947368421052631579
5	AK	4	19	13	0.68421052631578947368
6	AK	5	19	13	0.68421052631578947368
7	AK	6	19	11	0.57894736842105263158
8	AK	7	19	11	0.57894736842105263158
9	AK	8	19	10	0.52631578947368421053
10	AK	9	19	10	0.52631578947368421053
11	AK	10	19	8	0.42105263157894736842
12	AK	11	19	8	0.42105263157894736842
13	AK	12	19	5	0.26315789473684210526
14	AK	13	19	5	0.26315789473684210526

Figura 7: Tabela de retenção por estado.