Universidade Federal da Bahia
Departamento de Ciência da Computação
Mestrado em Ciência da Computação

Análise Descritiva Movielens

Amanda Chagas de Oliveira
MATE97 - Aspectos Recentes em Estatítica, Profa. Gecynalda Soares, 2019.2

Agenda

- Introdução
- O que é o Movielens?
- Abordagem de Popularidade
- Abordagem de Qualidade Descritiva

Introdução

- RecSys: necessidade de filtrar conteúdo de forma **personalizada**
  - Baseado em Filtragem Colaborativa
  - **Baseado em Conteúdo**
  - Fatorização de Matriz

- Primeiro aplicar alguma heurística de formação do grupo e então recomendar itens ao perfil de tal grupo

Movielens[1]

É composto por dados característicos dos filmes e das avaliações(**ratings**)

- Filmes
  - 9.742 itens distintos
- Ratings
  - 100.836 ratings
  - 610 usuários

[1] https://grouplens.org/datasets/movielens/

Movielens

Table 1: Exemplo do arquivo do *dataset* com as informações dos filmes.

MovieId	Title	Genres
1	Toy Story	Adventure,Animation,Children,Comedy,Fantasy
2	Jumanji	Adventure,Children,Fantasy
3	Grumpier Old Men	Comedy,Romance
4	Waiting to Exhale	Comedy,Drama,Romance

Table 2: Exemplo do arquivo com os *ratings* dos usuários sobre os filmes.

UserId	MovieId	Rating	Timestamp
1	3	4,0	964981247
1	6	4,0	964982224
1	47	5,0	964983815
...	...	...	...
2	318	3,0	1445714835
2	333	4,0	1445715029
...	...	...	...
3	31	0,5	1306463578
3	849	5,0	1306463611
...	...	...	...

Sob quais abordagens podemos analisar melhor esses dados?

- Abordagem de Popularidade
- Abordagem de Qualidade Descritiva

Abordagem de Popularidade

Serão cinco classes variando de acordo com a quantidade de avaliações recebidas pelos filmes:

- **F1** > 100
- 51 > **F2** >= 100
- 21 > **F3** >= 50
- 6 > **F4** >= 20
- 5 >= **F5**

Abordagem de Popularidade

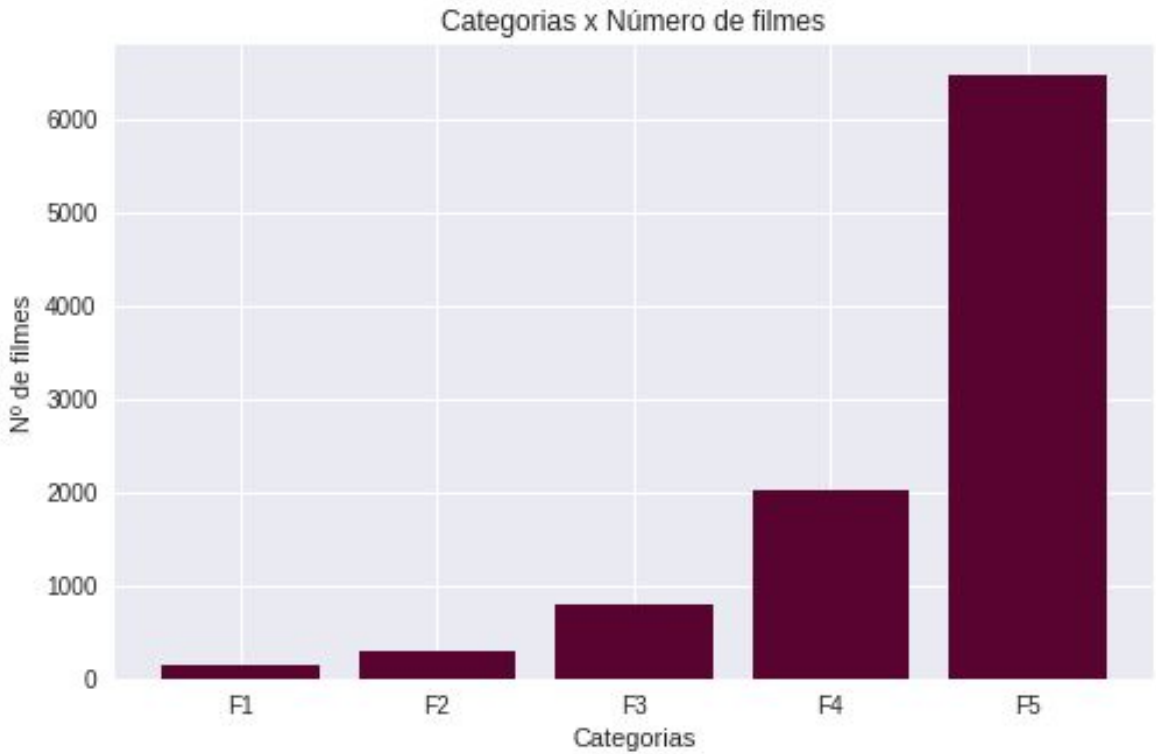
F1 = 134 filmes

F2 = 302 filmes

F3 = 799 filmes

F4 = 2033 filmes

F5 = 6474 filmes



Abordagem de Popularidade

Apesar da quantidade de itens ser bem discrepante por categoria, investigando a quantidade de avaliações também por categoria esse gráfico foi obtido:

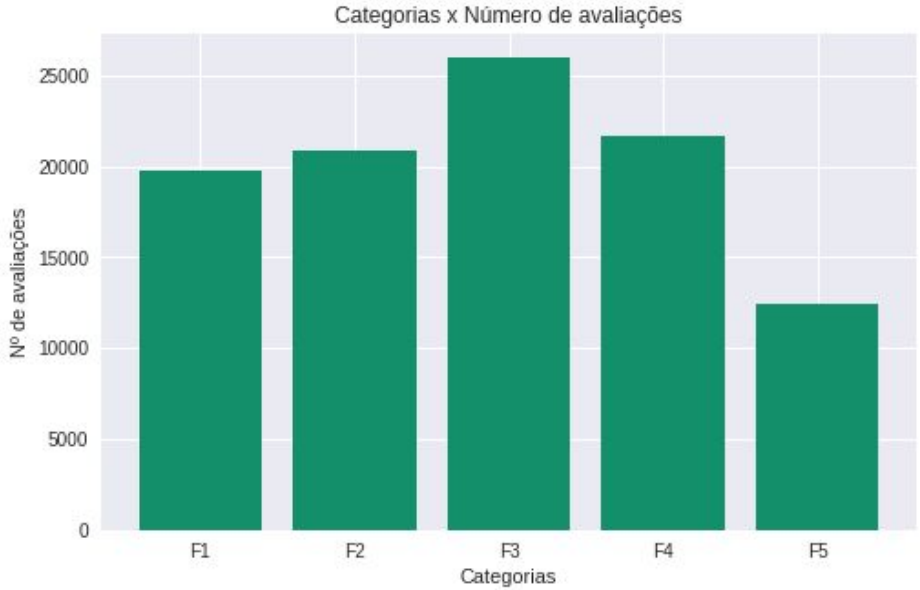
F1 = 19.788

F2 = 20.872

F3 = 25.998

F4 = 21.706

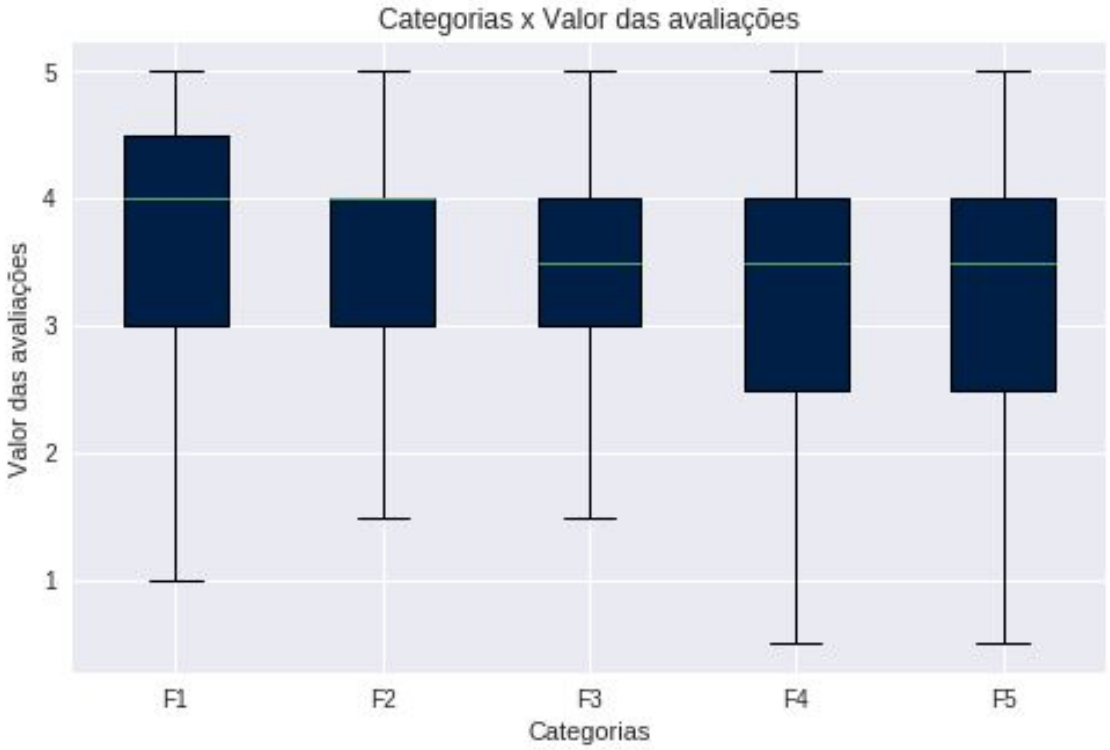
F5 = 12.472



Abordagem de Popularidade

Medidas de tendência central:

- F1 e F2 possuem assimetria negativa, F4 e F5 também, porém mais leve;
- Maior dispersão dos dados em F1, F4 e F5;
- F1, F2 e F3 possuem 75% das suas avaliações acima da nota 3;
- Todas as categorias, com exceção de F1, possuem 75% de avaliações até a nota 4.



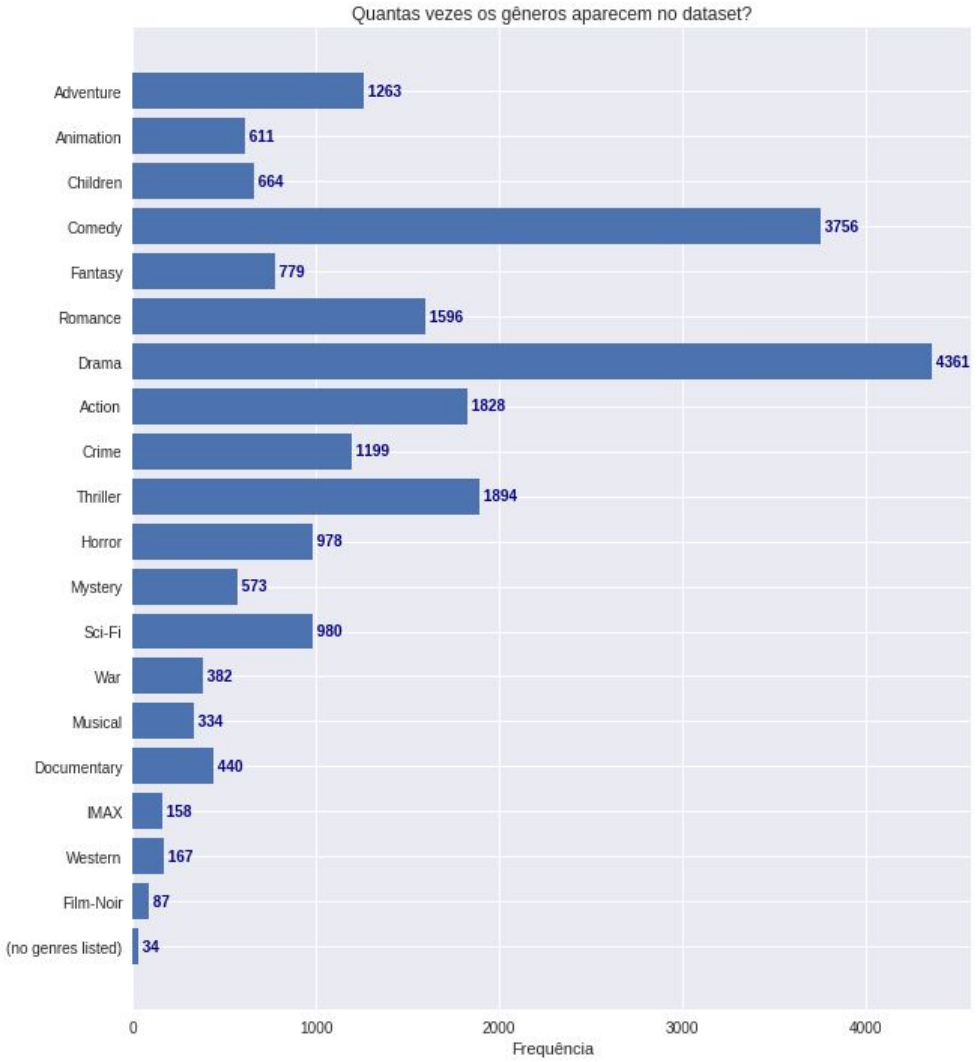
Abordagem de Qualidade Descritiva

Informações sobre os itens: Título e Gênero.

- Quão bem os gêneros descrevem um filme?
- Quantos gêneros esse dataset possue?
- Quais os gêneros mais populares nele?
- Quantos gêneros os filmes tem?

Abordagem de Qualidade Descritiva

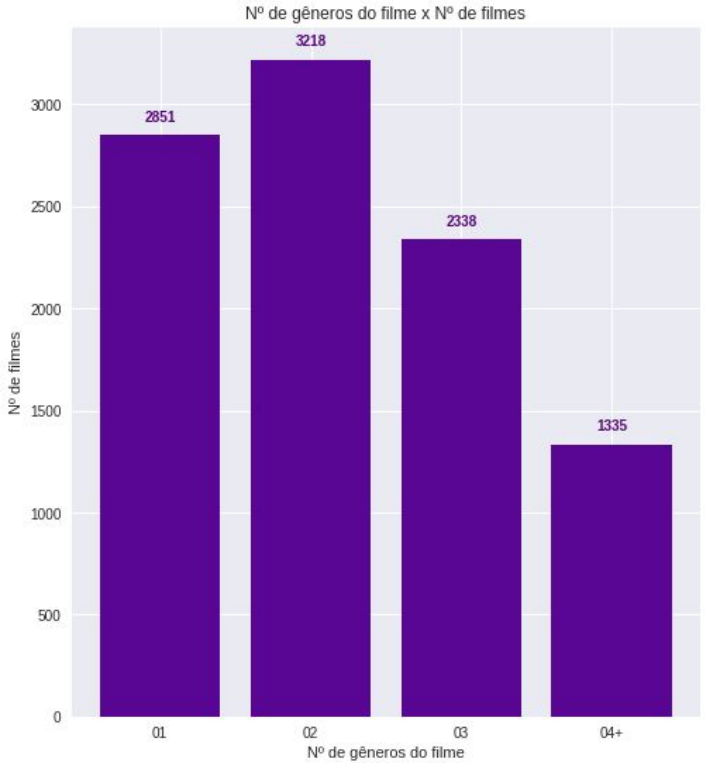
- 19 + 1 gêneros
- Mais frequentes: Drama (45%) e Comedy (39%)
- Os outros gêneros não atingem 20% dos itens
- Menos frequentes: Western (1,71%), IMAX (1,62%), Film Noir (0,89%) e (no genres listed) (0,34%)



Abordagem de Qualidade Descritiva

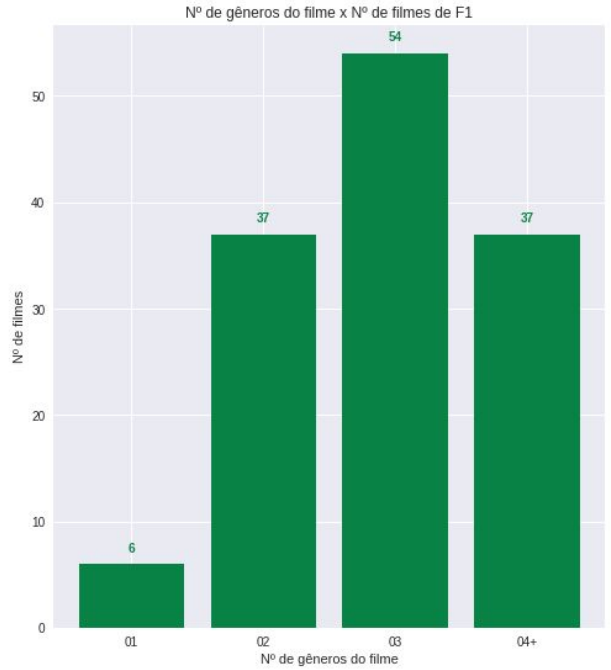
Geralmente, um filme possue múltiplos gêneros:

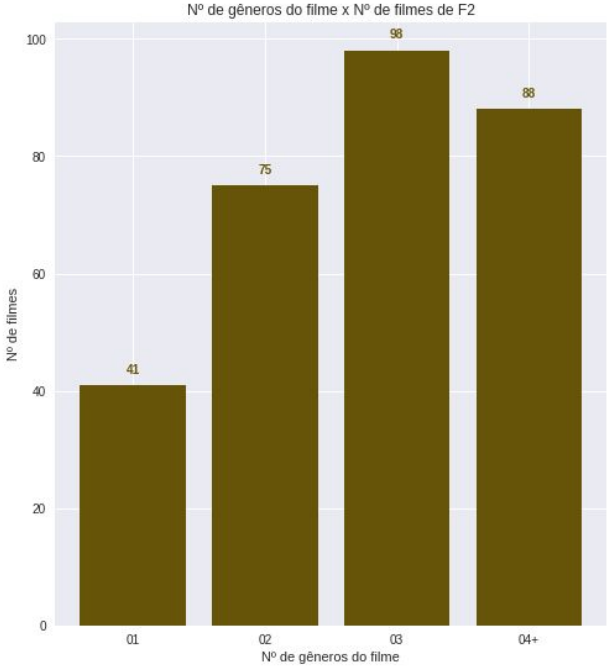
- Cerca de 62,3% dos filmes no dataset possuem no máximo 2 gêneros;
- Quase 30% dos filmes tem apenas 1 gênero listado;
- Será que o gênero descreve tão bem os filmes como imaginado?



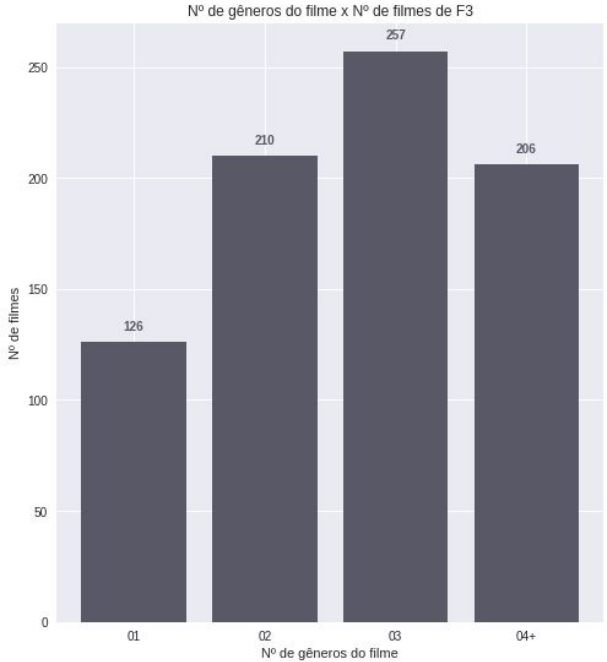
Abordagem de Qualidade Descritiva

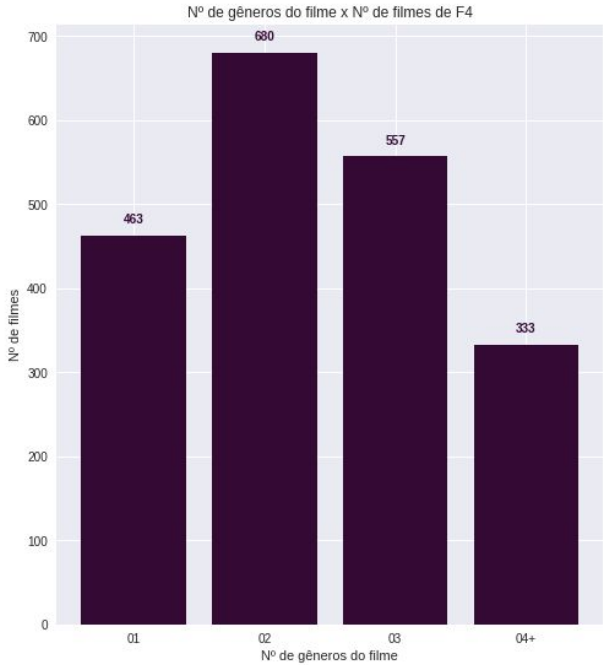
O próximo passo da investigação será verificar quantos gêneros os filmes de cada uma das cinco categorias **F1**, **F2**, **F3**, **F4** e **F5** possuem.

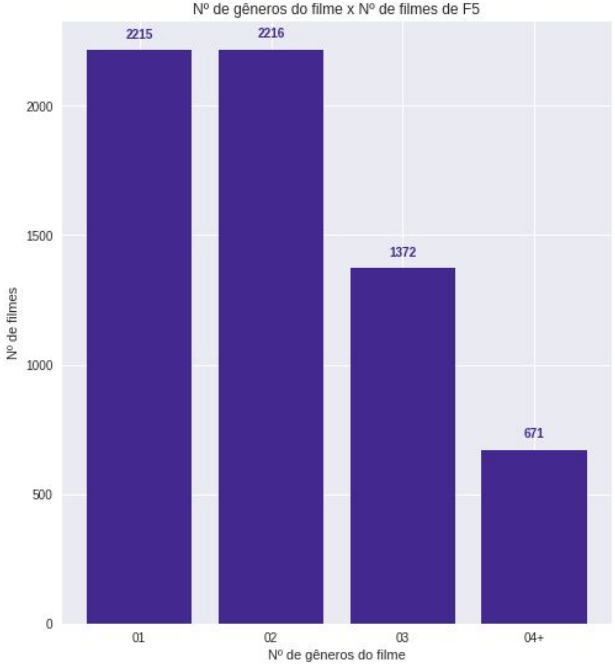




Abordagem de Qualidade Descritiva







Abordagem de Qualidade Descritiva

- Em **F2** foi observado o maior valor percentual de filmes com 4 ou mais gêneros;
- Em **F1** e **F2** pode-se afirmar que mais da metade dos filmes possuem 3 ou mais gêneros;
- Ainda em **F3**, mais da metade dos filmes possuem 3 ou mais gêneros;
- No entanto, em **F4** observa-se uma mudança de comportamento, onde mais da metade dos filmes possuem 2 gêneros ou menos;
- Em **F5** constatamos a virada de tendência observada no gráfico de **F4**, onde 68,44% dos filmes foram pouco descritos, contendo 2 gêneros ou menos.

Obrigada!

Código fonte, Relatório completo em .pdf e essa apresentação disponibilizados em:

https://github.com/amandachagas/surprise-test/tree/master/Descritive_Analyze