ANÁLISE DESCRITIVA MOVIELENS

December 2, 2019

Amanda Chagas de Oliveira
MATE97 - Aspectos Recentes em Estatística, 2019.1
Professora Gecynalda Soares da Silva Gomes
Programa de Pós-Graduação em Ciência da Computação
Univesidade Federal da Bahia(UFBA)

December 2, 2019

Contents

0.1 Introdução

Esse relatório foi preparado como uma atividade da disciplina MATE97 - Aspectos Recentes em Estatística, como maneira de praticarmos o que foi abordado em sala de aula. A professora Gecynalda pediu que avaliássemos e base de dados que utilizamos em nossa pesquisa de forma descritiva. Por tanto, s fim de melhor esclarecimento da base de dados, irei relatar resumidamente a questão da minha pesquisa.

Sistemas de Recomendação tradicionalmente recomendam itens a usuários individuais, no entanto existem cenários nos quais os indivíduos se reúnem em grupos e com isso surge a necessidade de recomendar também a grupos. A maior parte desses grupos se forma naturalmente, como por exemplo para assistir um filme, almoçar em algum restaurante ou até mesmo planejar uma viagem. Em todas essas hipóteses é possível utilizar os Sistemas de Recomendação enquanto ferramentas para oferecer informações personalizadas ao grupo como um todo. Para tanto, é preciso levar em consideração as preferências individuais dos membros do grupo, com o intuito de satisfazê-lo plenamente, e, nesse sentido, utilizar técnicas de agregação dessas informações. Embora existam técnicas de consenso para agregação das informações, as recomendações podem ser repetitivas entre si pois atenderão sempre ao mesmo perfil de grupo. Esse inconveniente abre precedente para adoção de técnicas de diversidade para recomendações ao grupo. Em nossa pesquisa, nós investigamos como aplicar tais técnicas de diversificação em recomendações para grupo, com base nas preferências dos membros, a fim de evitar a superespecialização do sistema e assim manter os membros do grupo satisfeitos de forma geral. Para tanto, é preciso desenvolver um modelo de formação do grupo, e, em seguida, modelar como será realizada a recomendação, para então conseguir aplicar técnicas de diversificação nos itens a serem recomendados. Inicialmente, decidimos trabalhar com a recomendação de filmes, e por tanto, iremos utilizar a base de dados do MovieLens.

0.2 MovieLens

A base de dados utilizada nessa análise foi o MovieLens[1]. O MovieLens é composto por dados característicos dos filmes e das avaliações(***ratings***) de diversos usuários sobre esses filmes. Mais ainda, a versão escolhida para se

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

trabalhar foi a versão reduzida, chamada de *"MovieLens Latest Small"*, o qual é o mais atual disponibilizado, apresentado na página do grupo de pesquisa em questão como *ml-latest-small.zip*, possui cerca de 100.836 *ratings*, 9.742 filmes e 610 usuários e foi atualizado em Setembro de 2018. É importante frisar ainda que cada usuário que contribui na formação desse *dataset* avaliou pelo menos 20 filmes.

Na minha pesquisa, que busca criar um Sistema de Recomendação para Grupos utilizando Técnicas de Diversidade, foram utilizado dois arquivos desse *dataset*: um arquivo contendo as informações dos filmes em si, e outro contendo os *ratings* dos usuários sobre os filmes.

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

O primeiro arquivo, referente as informações dos filmes propriamente ditos, possui as colunas de informação ***MovieId*** (identificador do filme), ***Title*** (título) e ***Genres*** (gêneros). Os valores da primeira coluna são representados por números inteiros positivos, os da segunda coluna são um conjunto de caracteres, e os da terceira coluna são palavras organizadas como em uma lista, sendo essas ordenadas em ordem alfabética. A Tabela 1 ilustra esse primeiro arquivo de dados.

O segundo arquivo é composto pelos *ratings* dos usuários sobre os filmes. Esse arquivo foi utilizado de forma integral e sua estrutura é ilustrada na Tabela 2. O *UserId* e o *MovieId* são relativos aos identificadores dos usuários e filmes respectivamente e possuem valores inteiros e positivos. A coluna *Rating* contém as notas de 0 a 5 variando com apenas uma casa decimal, e a coluna *Timestamp* é o tempo de duração do filme em milisegundos.

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
...	...	...	...

0.3 Classificação

0.3.1 Formação das classes

Com o propósito de classificar os filmes sob uma determinada perspectiva, decidi separar os 9.742 filmes em cinco categorias distintas de acordo com a sua popularidade. Serão cinco classes, variando de acordo com a quantidade de avaliações recebidas pelo filme: **F1**, **F2**, **F3**, **F4** e **F5**.

- **F1**: Filmes que receberam mais do que 100 avaliações.
- **F2**: Filmes que receberam entre 51 e 100 avaliações.
- **F3**: Filmes que receberam entre 21 e 50 avaliações.
- **F4**: Filmes que receberam entre 6 e 20 avaliações.
- **F5**: Filmes que receberam 5 ou menos avaliações.

O Gráfico 1 é a expressão da relação entre a quantidade de filmes e categorias. A categoria F1 é composta por 134 filmes, seguida por F2 com 302, F3 com 799, F4 com 2033 e F5 com a quantidade mais expressiva de 6474 filmes. Assim, a grande maioria dos filmes que compõem a base de dados podem ser considerados como de baixa popularidade, visto que as categorias F4 e F5 juntas totalizam 87,32% da base de dados. Esse resultado não é

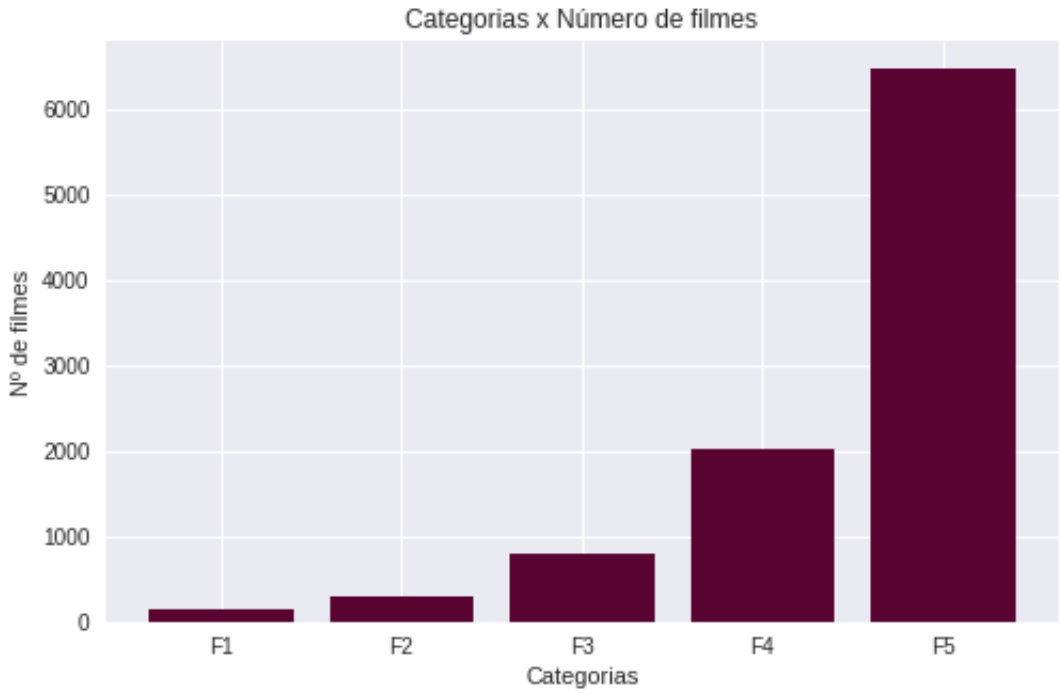


Figure 1: Gráfico com a quantidade de filmes por categoria.

surpreendente, visto que, em geral são poucos filmes que alcançam o hall da fama em Hollywood. Além disso, é preciso lembrar que essa base de dados é composta por apenas 610 usuários, dos quais não temos nenhum acesso a qualquer informação pessoal, como faixa etária ou região demográfica.

O Gráfico 2 busca expor uma outra perspectiva dessas cinco categorias, considerando agora a quantidade de avaliações que cada uma possui. Elas possuem quantidade de avaliações iguais a 19.788, 20.872, 25.998, 21.706 e 12.472 respectivamente. Apesar da grande disparidade em termos de quantidade de filmes por categorias observada anteriormente, nota-se que em termos de quantidade de avaliações, essas categorias não estão tão distantes.

Por fim, o Gráfico 3 tem por objetivo exibir o comportamento da tendência central de cada categoria. De acordo com esse gráfico, podemos afirmar então que as categorias F1 e F2 possuem uma assimetria negativa, F4 e F5 possuem uma leve assimetria negativa, ou seja, os valores em tais categorias estão mais próximos das notas mais elevadas. Entre todas as categorias, a F3 é aquela que mais se aproxima da simetria. A dispersão dos dados foi maior em F1,

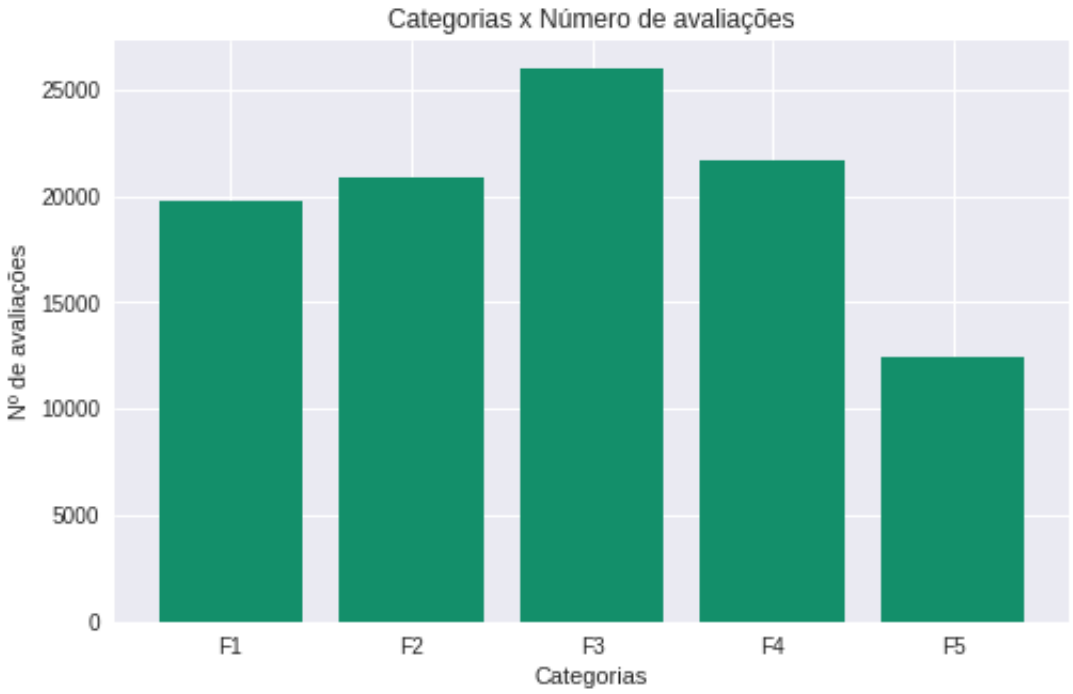


Figure 2: Gráfico com a quantidade de avaliações por categoria.

F4 e F5. As categorias F1, F2 e F3 possuem 75% das suas avaliações acima do valor 3. Além disso, todas as categorias,com exceção de F1, possuem 75% das suas avaliações até o valor 4. Como esperado, não foi encontrado nenhum *outlier* na base de dados, então todas as avaliações realmente estão contidas entre os valores 0 e 5.

0.3.2 Análise dos gêneros

Além da popularidade dos filmes, podemos analisar a base de dados sobre a vertente dos gêneros dos filmes, para assim, entendermos melhor a composição real da mesma.

A Figura 4 mostra que existem um total de 19+1 gêneros distintos na base de dados, sendo esse +1 o *(no genres listed)* que indica filmes que não tiveram seus gêneros explicitados no *dataset*. O gênero mais abundante é o *Drama*, estando presente em cerca de 45% dos filmes do *dataset*, seguido pelo gênero *Comedy*, o qual se encontra em cerca de 39% dos filmes. Os outros

gêneros categorizados não aparecem em 20% ou mais dos filmes, sendo que os gêneros mais raros são *Western* (1,71%), *IMAX* (1,62%), *Film-Noir* (0,89%) e o *(no genres listed)* (0,34%).

Considerando que, geralmente filmes possuem mais de um gênero que o definem, uma outra análise realizada diz respeito a quantidade de gêneros por filme. A Figura 5 representa a base de dados dos filmes distribuída em 4 categorias, sendo cada uma relativa a quantidade de gêneros por filme. Podemos verificar que cerca de 62,3% dos filmes no *dataset* possuem no máximo 2 gêneros, sendo que quase 30% dos filmes tem apenas 1 gênero listado. Com isso, podemos considerar isso como um indicativo que a *feature* "gênero" pode não representar tão bem os itens como o esperado.

Nas próximas figuras, iremos verificar quantos gêneros os filmes de cada uma das cinco categorias **F1**, **F2**, **F3**, **F4** e **F5** possuem, para verificarmos se pode haver algum indício de relação em atribuição de gêneros e popularidade dos filmes.

Na Figura 6 são exibidos os gráficos das categorias mais populares, **F1** e **F2**, sendo que na categoria **F2** obtivemos o maior número de filmes com 4 gêneros os mais em sua descrição. Em ambas as categorias, podemos afirmar que mais da metade dos filmes possuem 3 gêneros ou mais.

Na Figura 7 os resultados se apresentam mais homogêneos. Porém, notamos ainda em **F3** que mais da metade dos filmes possuem 3 gêneros ou mais, enquanto em **F4** mais da metade dos filmes possuem 2 gêneros ou menos.

Por fim, na Figura 8 que representa os filmes menos populares do *dataset*, constatamos a virada de tendência observada no gráfico da categoria **F4**, onde 68,44% foram pouco descritos, contendo 2 gêneros ou menos em sua estrutura.

References

http://www.datasciencemadesimple.com/box-plot-in-python/
https://towardsdatascience.com/understanding-boxplots-5e2df7bcbd51
https://matplotlib.org/3.1.1/api/$_as_gen/matplotlib.pyplot.boxplot.html$

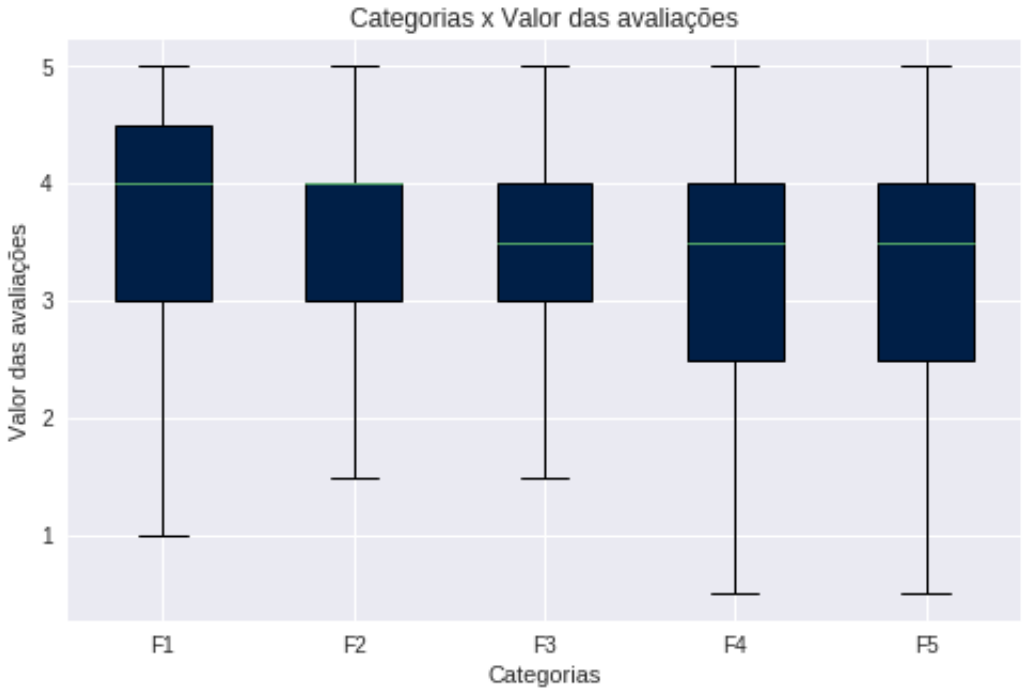


Figure 3: Gráfico com a quantidade de avaliações por categoria.

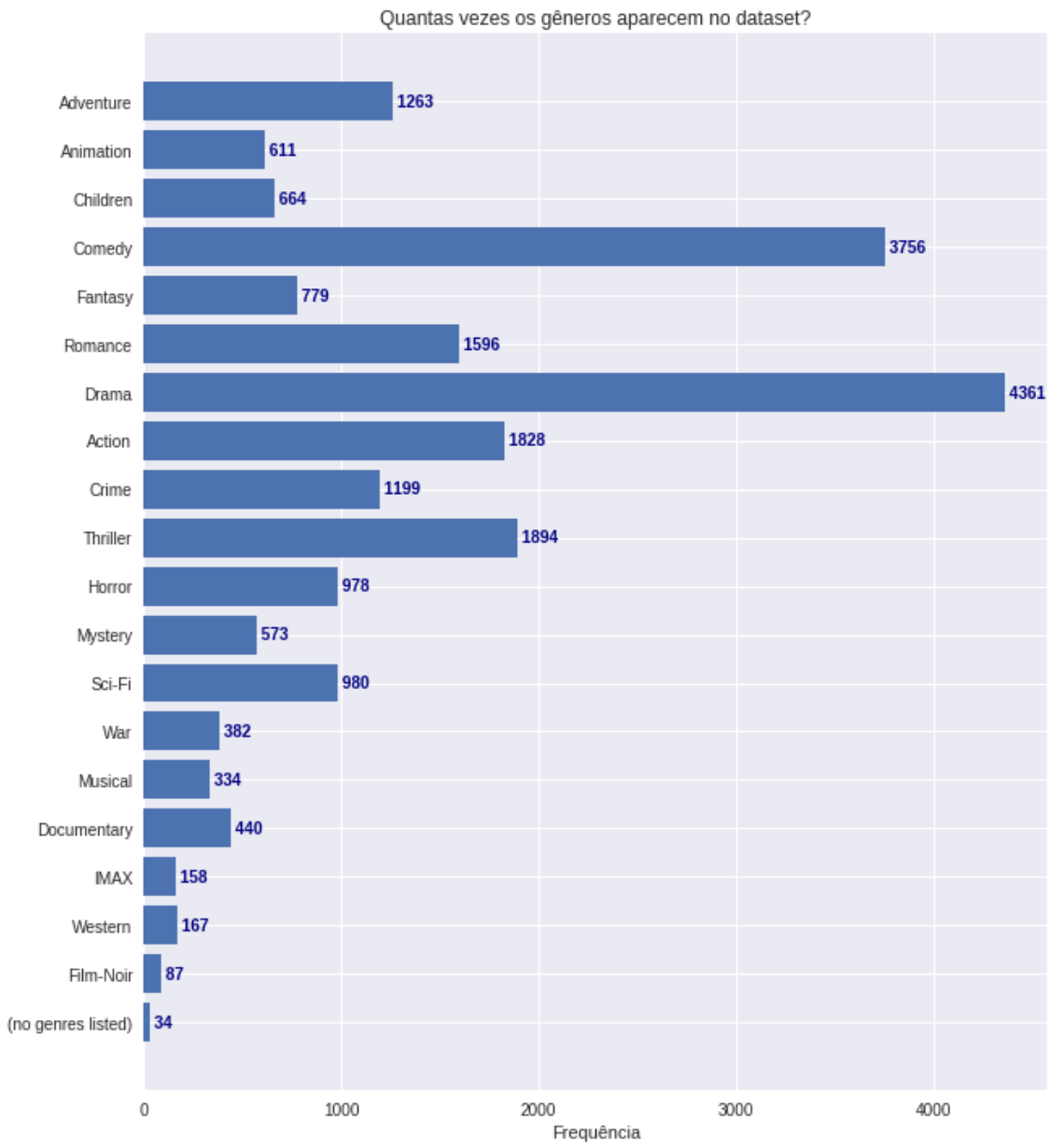


Figure 4: Gráfico com a frequência dos gêneros no *dataset*.

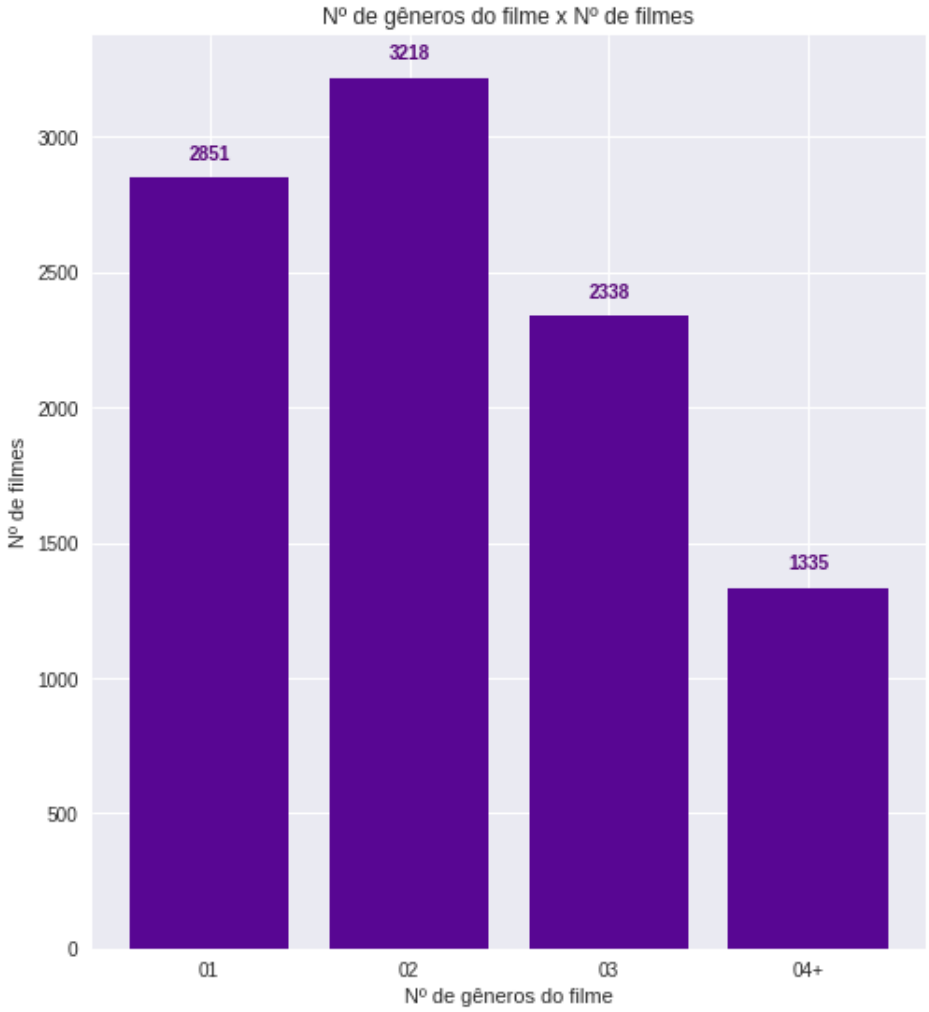
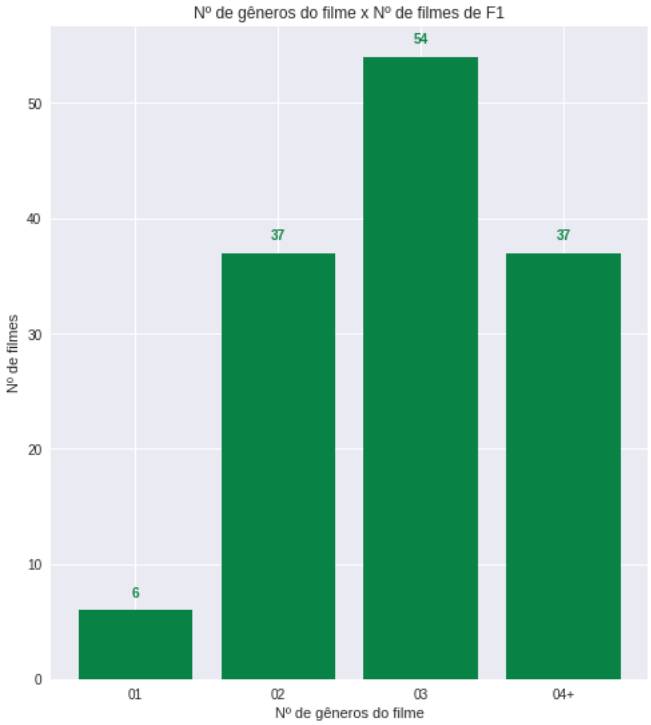


Figure 5: Gráfico com o número de gêneros por filme.

(a) Categoria F1



(b) Categoria F2

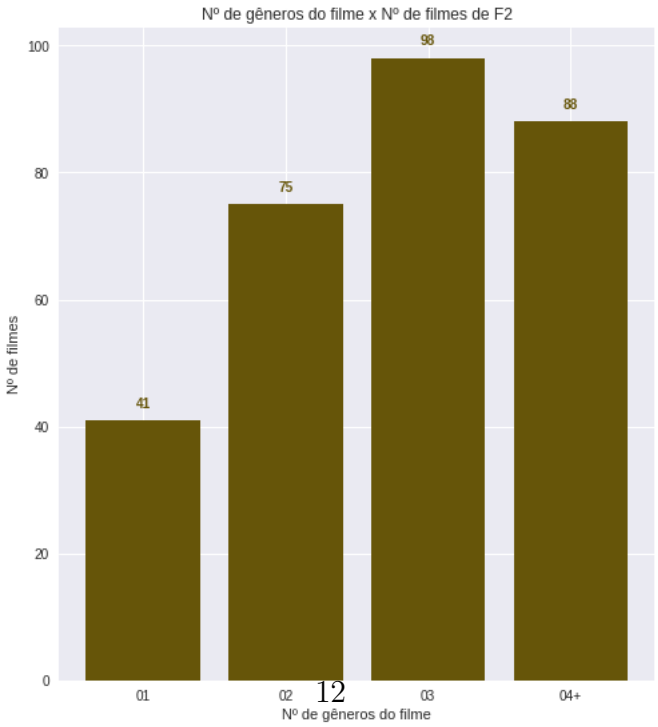
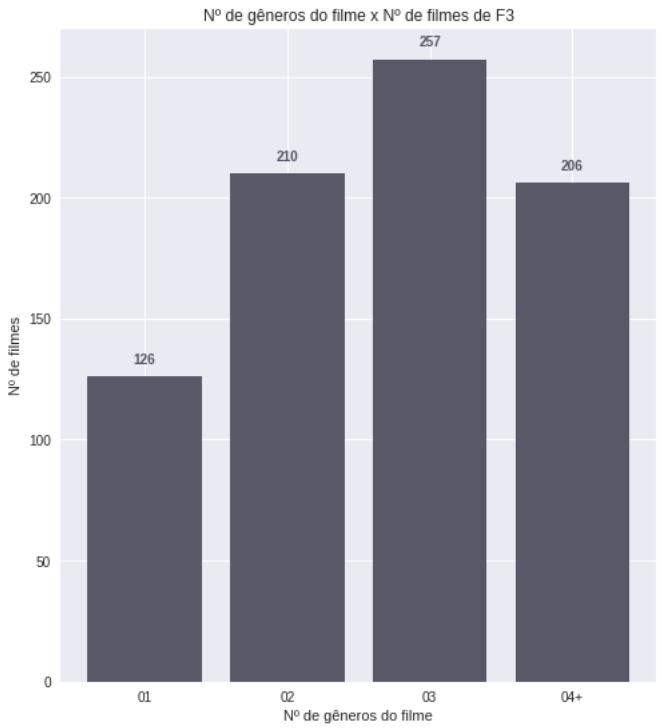


Figure 6: Gêneros por filme das Categoria de Popularidade F1 e F2.

(a) Categoria F3



(b) Categoria F4

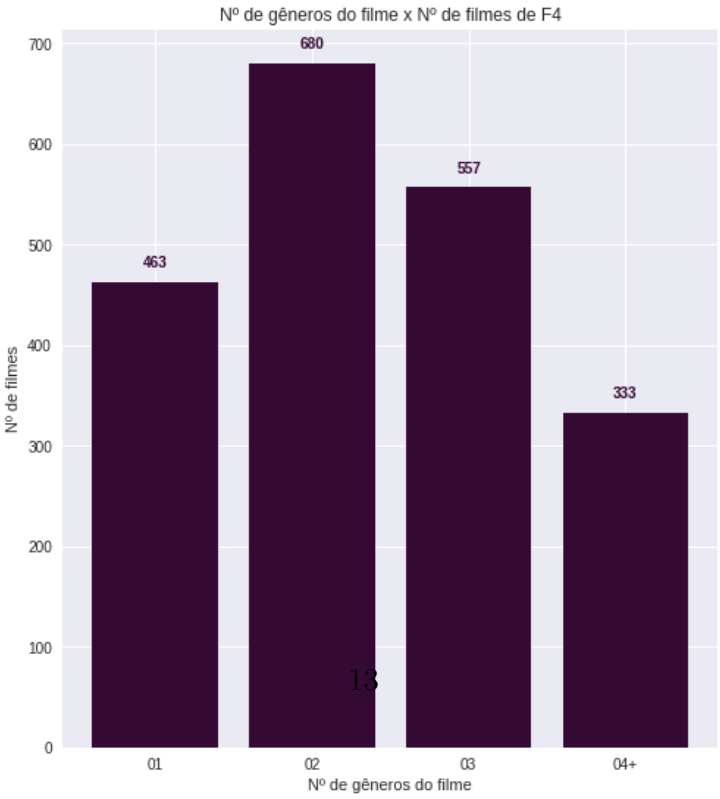


Figure 7: Gêneros por filme das Categoria de Popularidade F3 e F4.

(a) Categoria F5

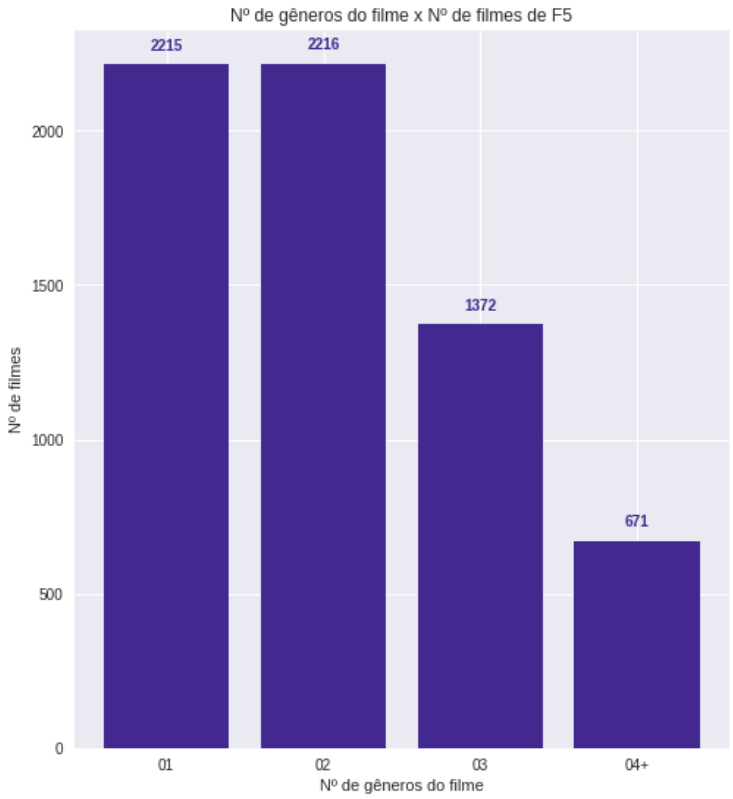


Figure 8: Gêneros por filme da Categoria de Popularidade F5.