Lab3 - Checkpoint 2

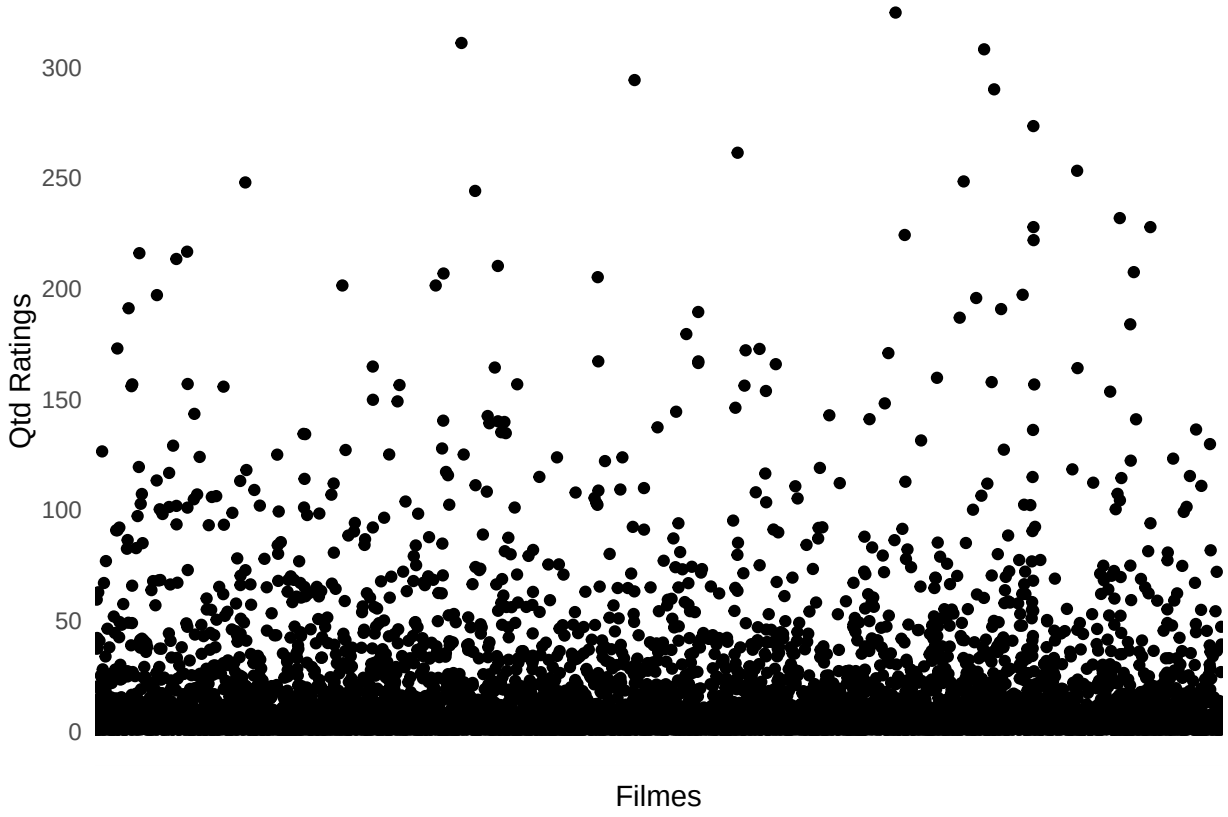
12 de maio de 2016

Análise exploratória

Os dados que iremos explorar fazem parte de um *dataframe* com um conjunto de filmes, seus generos e suas avaliações de qualidade. Abaixo temos o resumo da estrutura dos nossos dados.

- **moviesId**: identificador único do filme
- **title**: Titulo do filme
- **numGenres**: Quantidade total de gêneros do filme.
- **gêneros (vários)**: os genêros que o filme tem, quando for 0 o gênero não está contido no filme, quando for 1 o mesmo está contido.
- **rating**: nota de avaliação do filme.

O *dataset* tem uma quantidade de avaliações por filme muito baixa como podemos ver no gráfico abaixo:



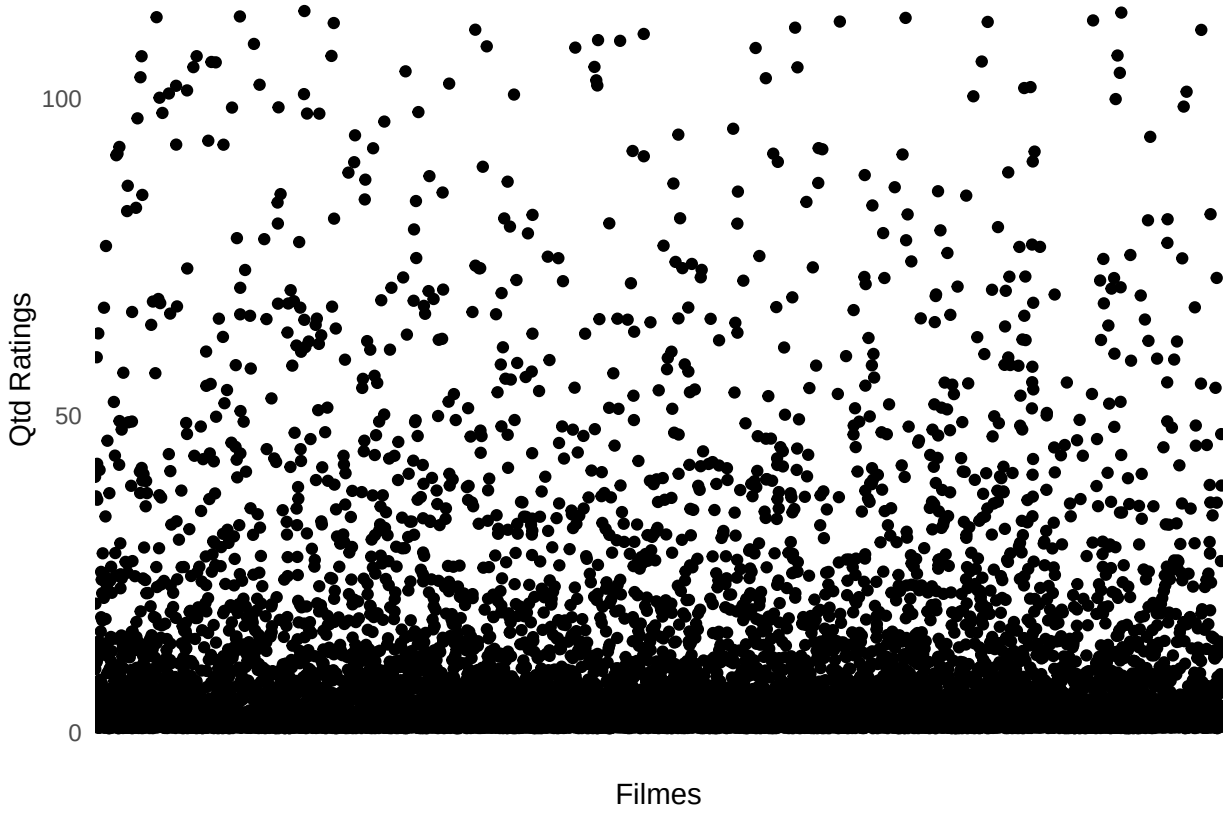
Percebendo isso, foi gerado um sumário para entender melhor a distribuição dos dados.

```
Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 1.0     1.0     3.0    10.2     8.0   325.0
```

Podemos perceber que existe uma concentração muito alta de filmes com **oito** avaliações ou menos. Iremos trabalhar com a 99% dos filmes, eles tem 114 avaliações ou menos. Os filmes com mais que essa quantidade, consideramos como *outliers*.

A preocupação foi retirar os *outliers* do gênero, para que eles não impactem na avaliação da relação entre a quantidade de gêneros e a avaliação que um filme tem.

Com a retirada dos filmes que tem mais de 114 avaliações, Nosso *dataset* resultante tem agora cerca de 44 mil observações contra as 53 mil avaliações iniciais.



Abaixo é possível ver a média de avaliação por quantidade de gêneros dos filmes. Percebemos algo estranho que é o resultado inteiro nos filmes com oito gêneros, ao verificar o *dataset*, percebe-se que só há um filme com 8 gêneros e esse teve apenas uma avaliação. Também há poucas avaliações para filmes com 6 e 7 gêneros. Logo, foram descartados os filmes com 6 gêneros ou mais por não terem pelo menos 300 avaliações.

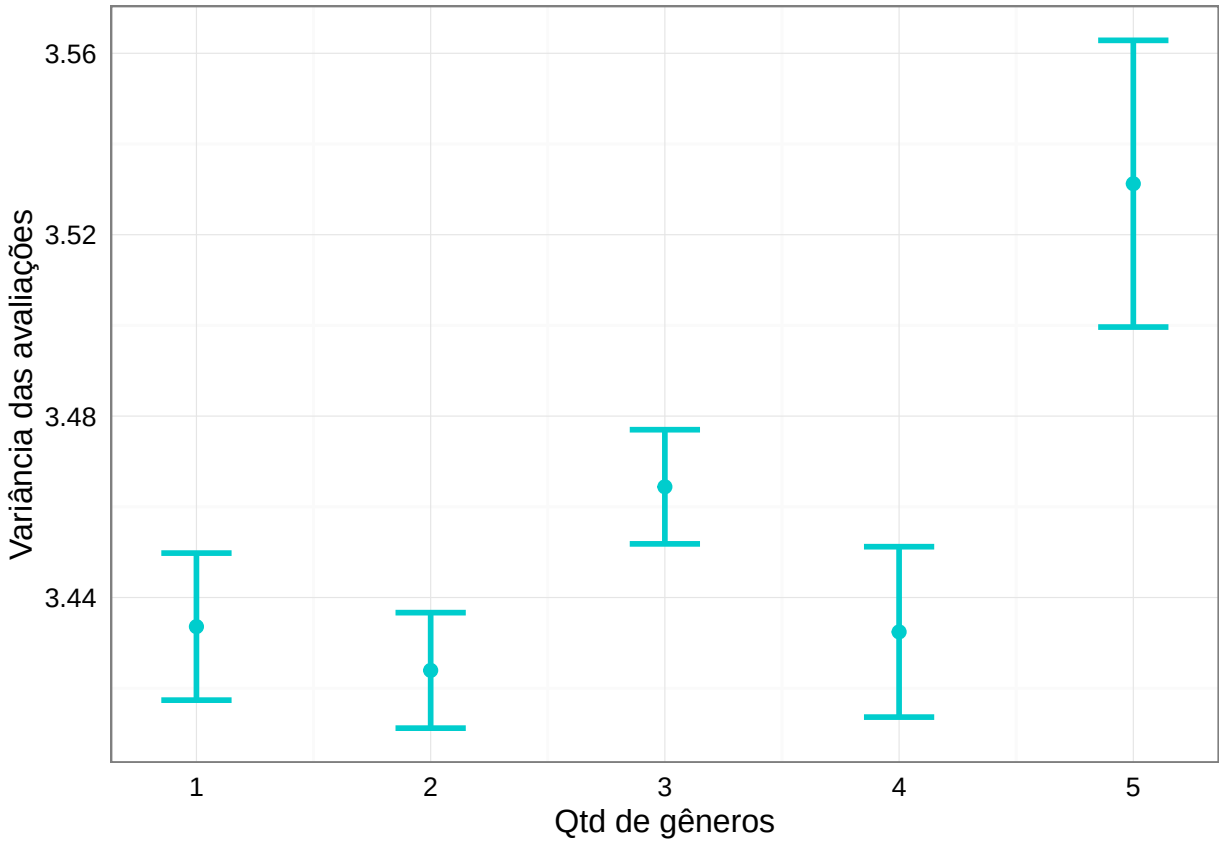
```
Source: local data frame [9 x 3]

  numGenres mean(rating) length(rating)
      (int)        (dbl)          (int)
1         1     3.433778          16792
2         2     3.423884          26131
3         3     3.464377          26612
4         4     3.431785          13025
5         5     3.531060           4121
6         6     3.401389            720
7         7     3.872928            181
8         8     3.166667              3
9        10     2.250000              2
```

Relação entre quantidade de gêneros e avaliações médias

Normalmente os filmes têm vários gêneros. Existe uma relação em quantos gêneros os filmes se encaixam e a avaliação média que os filmes recebem? Mais especificamente: se consideramos a média dos filmes com 1, 2, 3 . . . gêneros, existe alguma quantidade de gêneros num mesmo filme que em média recebe avaliações melhores? Caso exista, estime a diferença nas médias entre essa combinação e filmes com apenas um gênero.

Como definido anteriormente, iremos utilizar para essa análise somente filmes que tem até cinco gêneros. No gráfico abaixo podemos perceber que filmes com 5 gêneros recebe em média uma avaliações melhores que os demais.



Tendo em vista que existe uma quantidade de gêneros que em média recebe notas maiores, agora iremos calcular a diferença nas médias entre os filmes com **um** gênero e com **cinco** gêneros. Para isso, iremos aplicar a técnica *bootstrap* de comparação entre duas estatísticas.

```
filmes_5genres <- filmes.result %>% filter(numGenres == 5) %>% select(rating)
filmes_1genres <- filmes.result %>% filter(numGenres == 1) %>% select(rating)
diff_genres <- bootstrap2(
  data=filmes_5genres$rating,
  data2=filmes_1genres$rating,
  statistic = mean)
```

Abaixo podemos ver o intervalo de confiança da diferença entre as duas médias de avaliações.

```
                        2.5%     97.5%
mean: data-data2 0.06207871 0.1318784
```

Quais os genêros que tem maior variação nas notas

Entre os 10 gêneros que têm mais filmes, quais possuem maior variação nas notas atribuídas a seus filmes?

Como podemos ver abaixo, temos a lista de gêneros mais analisados. Os 10 primeiros são Drama, Action, Comedy, Thriller, Adventure, Romance, Crime, SciFi, Fantasy, Mystery.

```
     Drama      Action      Comedy    Thriller   Adventure     Romance
     46960       39308       38055       29288       23076       19094
     Crime       SciFi     Fantasy     Mystery      Horror   Animation
     18291       16795       10889        8320        7983        5966
       War     Musical     Western   FilmNoir Documentary     NoGenre
      5828        4287        2314        1210        1206           7
 Childrens
         0
```

Logo, iremos utilizar a técnica de *resample* para encontrar os intervalos de confiança e plotar o resultado em um gráfico do tipo error bar. Por fim, podemos ver que os filmes que tem maior variância nas avaliações são os que tem o gênero Drama. É interessante confirmar isso, já que filmes de drama tem diversas categorias e se relaciona com vários outros gêneros, assim é natural que esse tipo de gênero tenha uma variância maior nas avaliações.