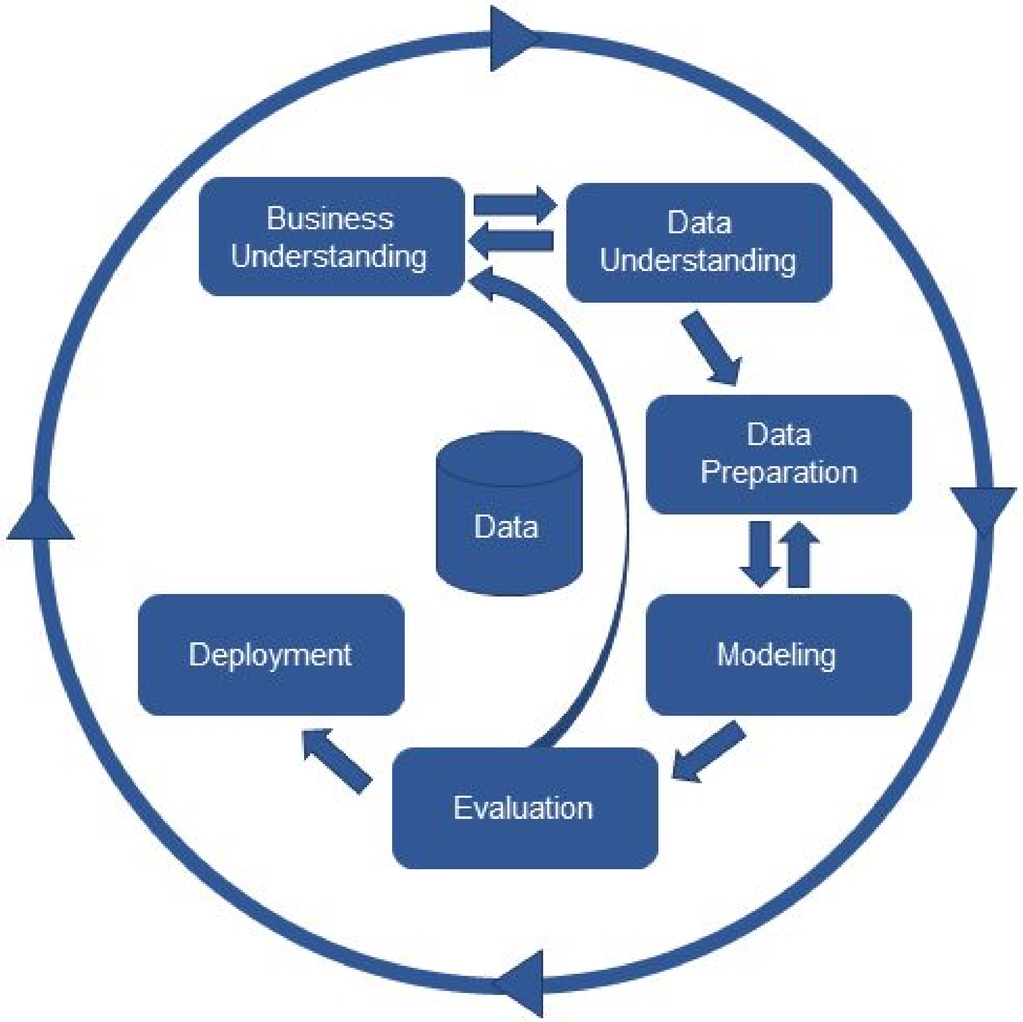
IMDB 5000 MOVIE DATASET

Gil Gonçalves - A67738
Luis Paulo Ferreira Pedro - A70415
José Pedro Santos Monteiro - A73014
Bruno Manuel Gonçalves Ribeiro - A73269

INTRODUÇÃO

CONTEXTUALIZAÇÃO E APRESENTAÇÃO DO CASO DE ESTUDO

METODOLOGIA CRISP-DM

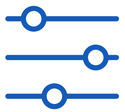


Business Understanding

Business Understanding

DETERMINAR OBJETIVOS DO NEGÓCIO

SITUAÇÃO ATUAL



DEFINIÇÃO DOS OBJETIVOS DO NEGÓCIO

PLANO DO PROJETO

1. Prever se um filme vai ser bem recebido pela crítica
2. Prever se um filme vai ser um sucesso a nível monetário
3. Encontrar grupos de filmes com impacto semelhante
4. Perceber se os atores e realizadores influenciam a classificação de um filme
5. Perceber se os atores e realizadores influenciam a classificação de um filme

Data Understanding

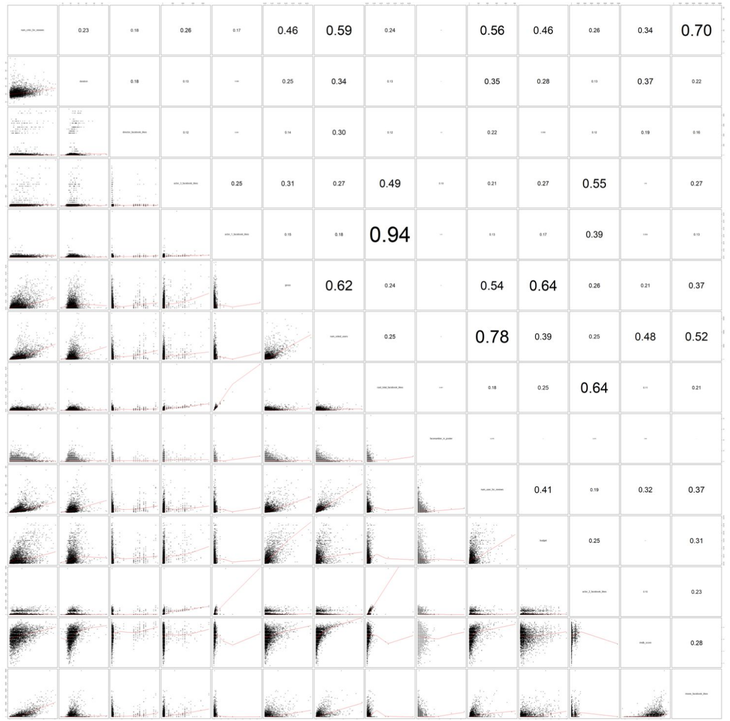
Data Understanding

RECOLHA INICIAL DOS DADOS

DESCRIÇÃO DOS DADOS

EXPLORAÇÃO DOS DADOS

VALIDAÇÃO DA QUALIDADE DOS DADOS

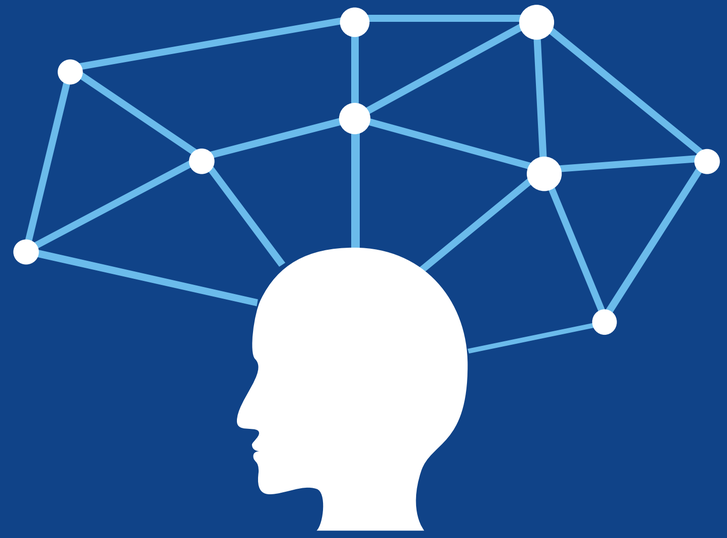


Data Preparation

SELEÇÃO DOS DADOS

Foram analisados os dados e posteriormente descartamos dados que não eram relevantes para o projeto:

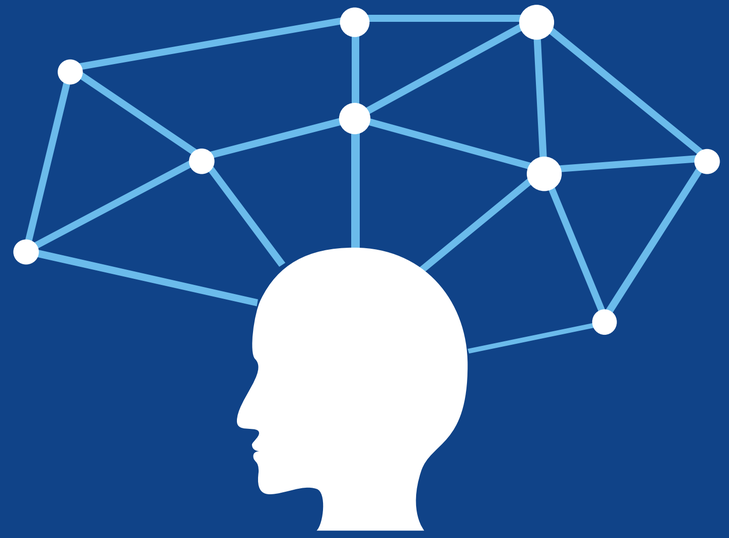
- color
- facenumer_in_poster
- plot_keywords
- movie_imdb_link
- aspect_radio



LIMPEZA DOS DADOS

Foi então necessário realizar a limpeza dos dados para posterior aplicação dos modelos, para isso identificamos e tratamos:

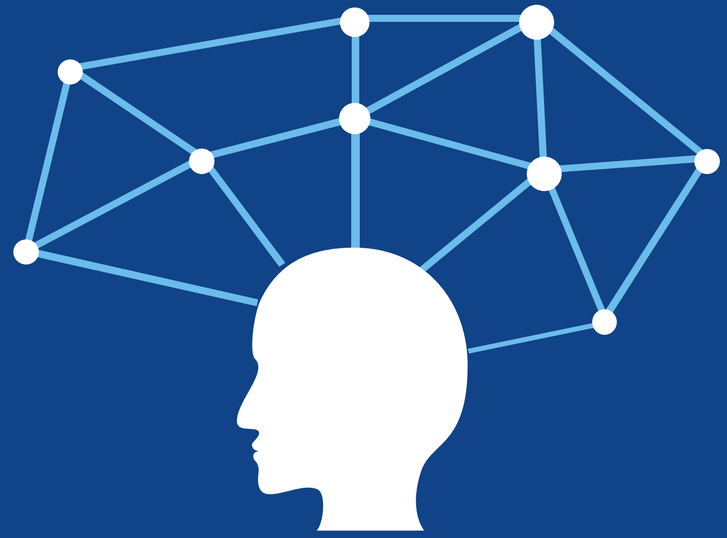
- Valores Nulos
- Dados duplicados



CONSTRUÇÃO DE NOVOS DADOS

Para o atributo gênero foi necessário criar novos campos e para isso aplicamos duas abordagens:

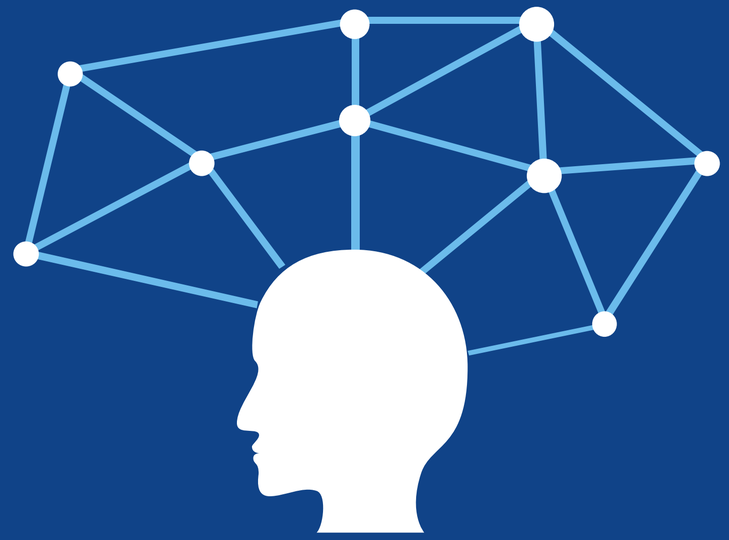
- 1-of-C (python)
- Lista (biblioteca reshape->melt)



FORMATAÇÃO DOS DADOS

Procedemos então à seguinte formatação dos dados:

- Uniformização da moeda
- Titulos dos filmes delimitados com um caracter especial

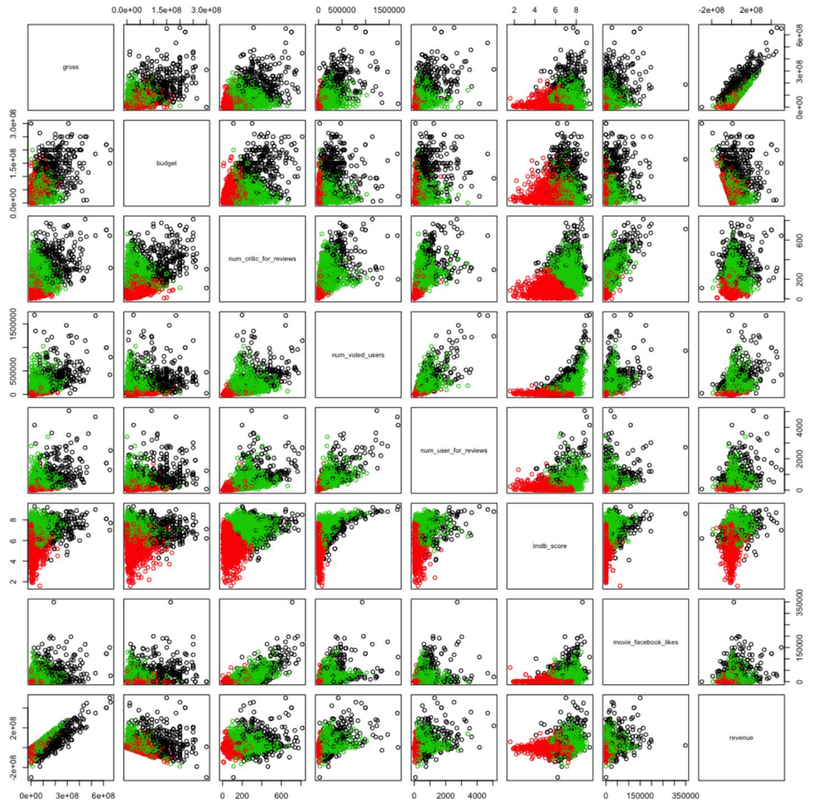


Modeling

Encontrar filmes com impacto semelhante *(Clustering)*

Remoção dos outliers

Kmeans



Modelo de previsão para o IMDB score de um filme
(Regressão)

Atributos utilizados:

- imdb_scores
- budget
- movie_facebook_likes
- duration
- director_facebook_likes

Resultados:

Modelo	RMSE	R2
Regressão Linear	0.037	0.204
Decision Tree	0.055	0.236
Random Forest	0.026	0.363

Modelo de previsão para a receita bruta de um filme *(Regressão)*

Atributos utilizados:

- gross
- budget
- imdb_score
- num_critics_for_reviews
- num_voted_users
- num_users_for_reviews

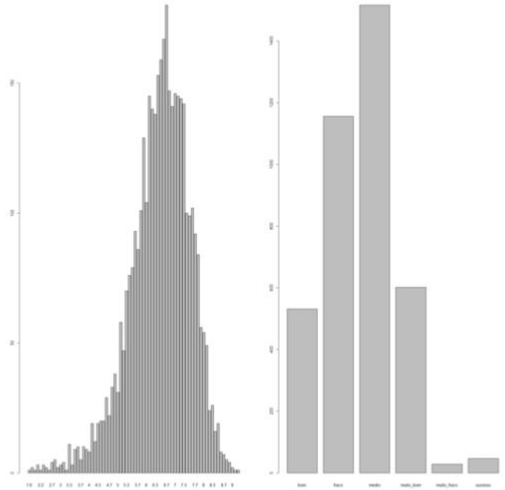
Resultados:

Modelo	RMSE	R2
Regressão Linear	0.548	-15.019
Decision Tree	0.549	-15.057
Random Forest	0.550	-15.089

Perceber se os atores e realizadores influenciam a classificação de um filme (Associação)

Foram então criadas seis escalas:

- muito_fraco → varia entre [0;3]
- fraco → varia entre]3;6]
- medio → varia entre]6;7]
- bom → varia entre]7;7.5[
- muito_bom → varia entre [7.5;8.5]
- sucesso → varia entre]8.5;10]



Resultados:

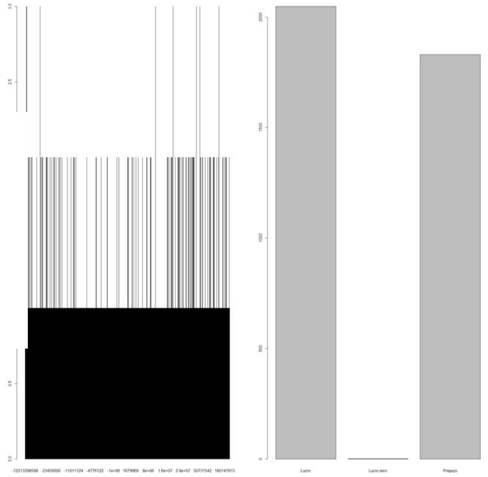
	lhs		rhs	support	confidence	lift
[1]	{Brian Levant}	=>	{fraco}	0.001545993	1.0000000	3.357266
[2]	{Michael Winterbottom}	=>	{medio}	0.001545993	1.0000000	2.560026
[3]	{Jonathan Liebesman}	=>	{fraco}	0.001545993	1.0000000	3.357266
[4]	{Tyler Perry}	=>	{fraco}	0.001545993	1.0000000	3.357266
[5]	{Raja Gosnell}	=>	{fraco}	0.001545993	1.0000000	3.357266
[6]	{Michael Jai White}	=>	{fraco}	0.001803659	1.0000000	3.357266
[7]	{Jon Turteltaub}	=>	{medio}	0.001803659	1.0000000	2.560026
[8]	{Jay Roach}	=>	{medio}	0.001803659	1.0000000	2.560026
[9]	{Peter Segal}	=>	{medio}	0.001545993	0.8571429	2.194308
[10]	{Sarah Michelle Gellar}	=>	{fraco}	0.001803659	0.8750000	2.937608
[11]	{Garry Marshall}	=>	{fraco}	0.001545993	0.7500000	2.517950
[12]	{Christopher Nolan}	=>	{sucesso}	0.001545993	0.7500000	61.930851
[13]	{Amy Poehler}	=>	{fraco}	0.001545993	0.7500000	2.517950
[14]	{Zooey Deschanel}	=>	{medio}	0.001803659	0.7777778	1.991132
[15]	{Robin Wright}	=>	{medio}	0.001545993	0.6666667	1.706684

Atributos utilizados: Director_name , actor_1_name, imdb_score

Perceber se os atores e realizadores influenciam o lucro de um filme (Associação)

Foram então criadas três classes:

- Valores menor que zero → Prejuízo
- Valores igual a zero → Lucro zero
- Valores maior que zero → Lucro



Resultados:

```
     lhs                          rhs          support     confidence lift
[1]  {Michael Winterbottom}    => {Prejuizo} 0.001545993 1.0000000  2.119607
[2]  {Lin Shaye}               => {Lucro}    0.001545993 1.0000000  1.894095
[3]  {Tim Story}               => {Lucro}    0.001545993 1.0000000  1.894095
[4]  {Jada Pinkett Smith}      => {Lucro}    0.001545993 1.0000000  1.894095
[5]  {Tyler Perry}             => {Lucro}    0.001545993 1.0000000  1.894095
[6]  {Andy Fickman}            => {Lucro}    0.001545993 1.0000000  1.894095
[7]  {Nia Long}                => {Lucro}    0.001545993 1.0000000  1.894095
[8]  {Catherine Deneuve}       => {Prejuizo} 0.001545993 1.0000000  2.119607
[9]  {Alyson Hannigan}         => {Lucro}    0.001803659 1.0000000  1.894095
[10] {James Wan}               => {Lucro}    0.001545993 0.8571429  1.623510
```

Atributos utilizados: Director_name , actor_1_name, grosse o budget

Evaluation

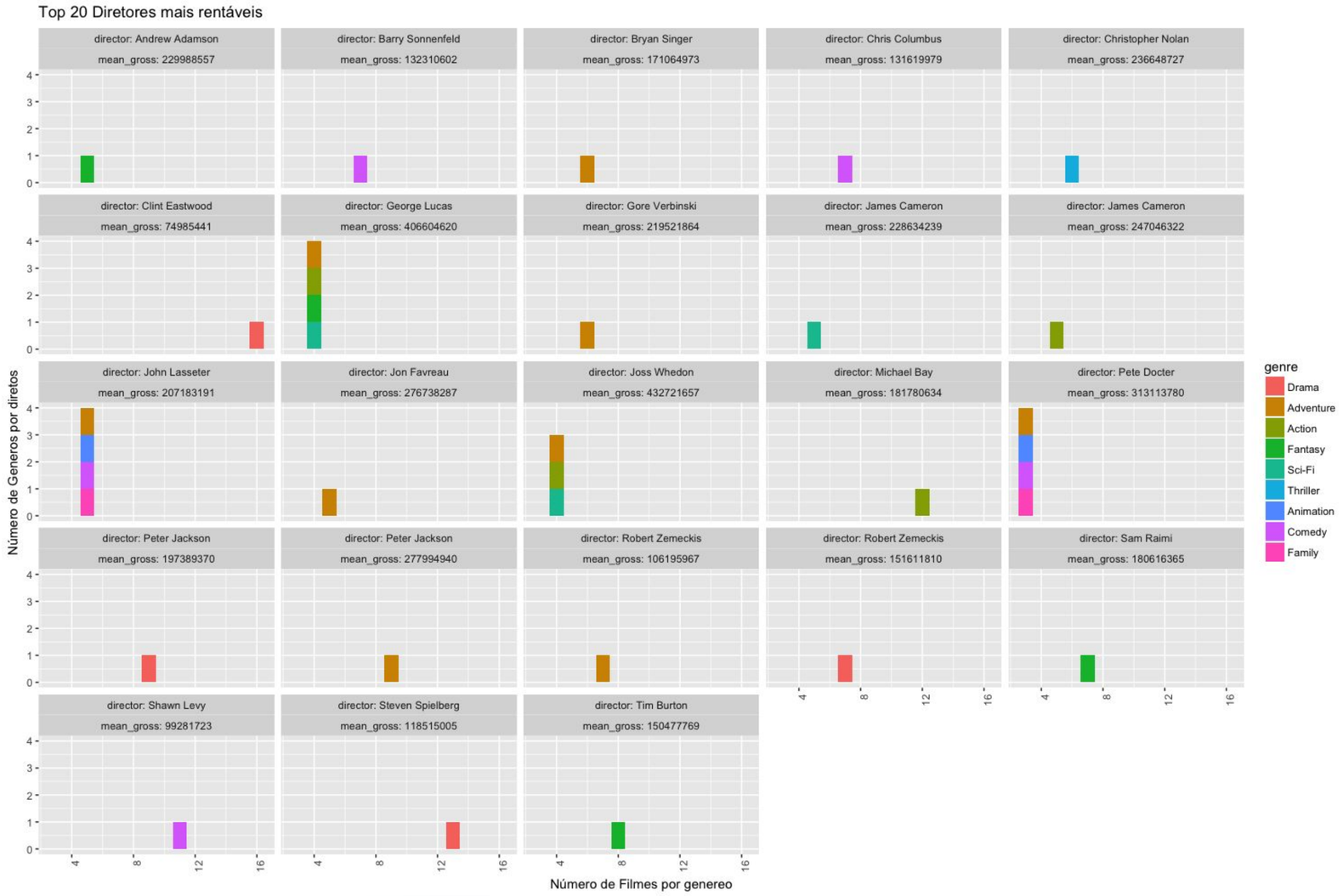
Evaluation

- Encontrar filmes com impacto semelhante
- Modelo de previsão para o IMDB score de um filme
- Modelo de previsão para o receita bruta de um filme
- Perceber se os atores e realizadores influenciam a classificação de um filme
- Perceber se os atores e realizadores influenciam o lucro de um filme

Deployment

Genres

Foi feita uma análise por atributo, e vimos quais os realizadores mais rentáveis na área:



IMDB 5000 MOVIE DATASET

Gil Gonçalves - A67738
Luis Paulo Ferreira Pedro - A70415
José Pedro Santos Monteiro - A73014
Bruno Manuel Gonçalves Ribeiro - A73269