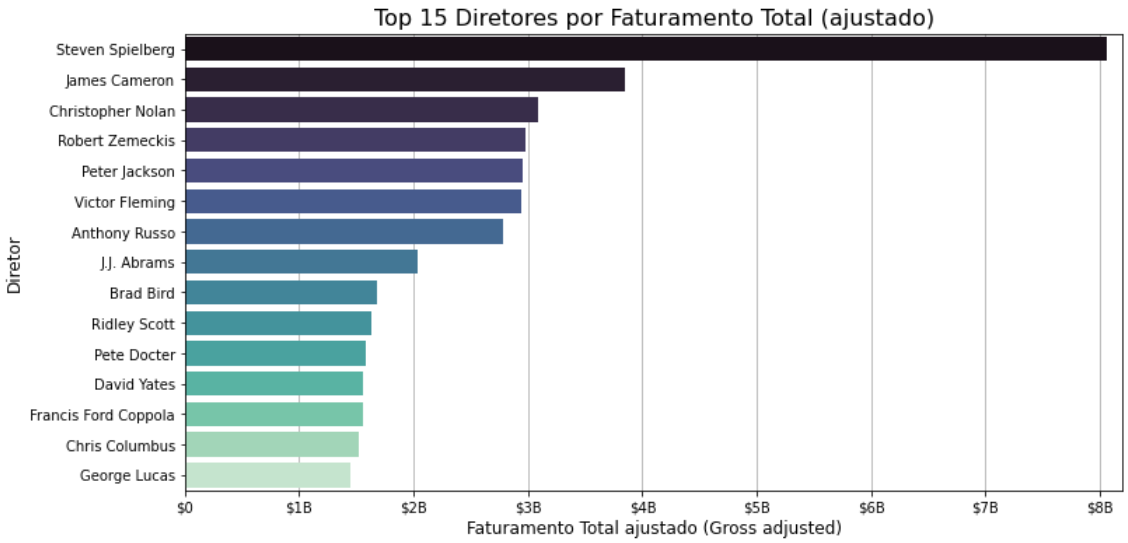
Desafio Cientista de Dados

Análise exploratória dos dados

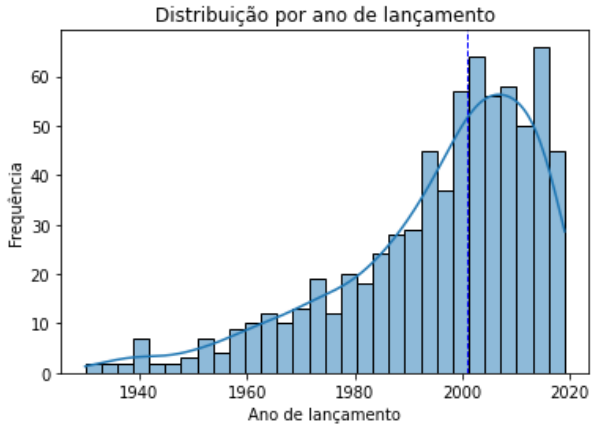
A seguir, serão apresentados alguns gráficos, seguidos de uma discussão sobre as observações realizadas.

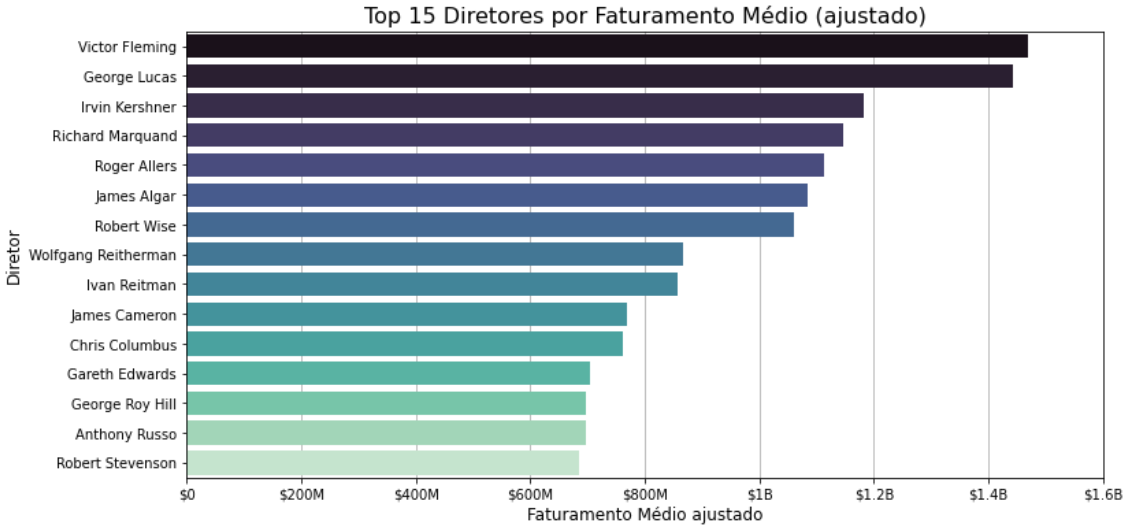
Devido a base apresentada conter títulos datados a partir de 1930, adicionei uma nova coluna chamada Gross_adjusted, que é o campo Gross ajustado pela inflação. O método utilizado foi baseado na inflação de 1930 até 2025, resultando em uma inflação média de 3,17% ano. Basicamente, foi criada uma coluna representando quanto o dinheiro da época do lançamento do título valeria atualmente, atualizado pela inflação média.

Spielberg sempre foi um ícone no cinema, seus números no gráfico a seguir, comprovam isso. O diretor de vários de meus filmes prediletos também figura a lista na terceira posição.



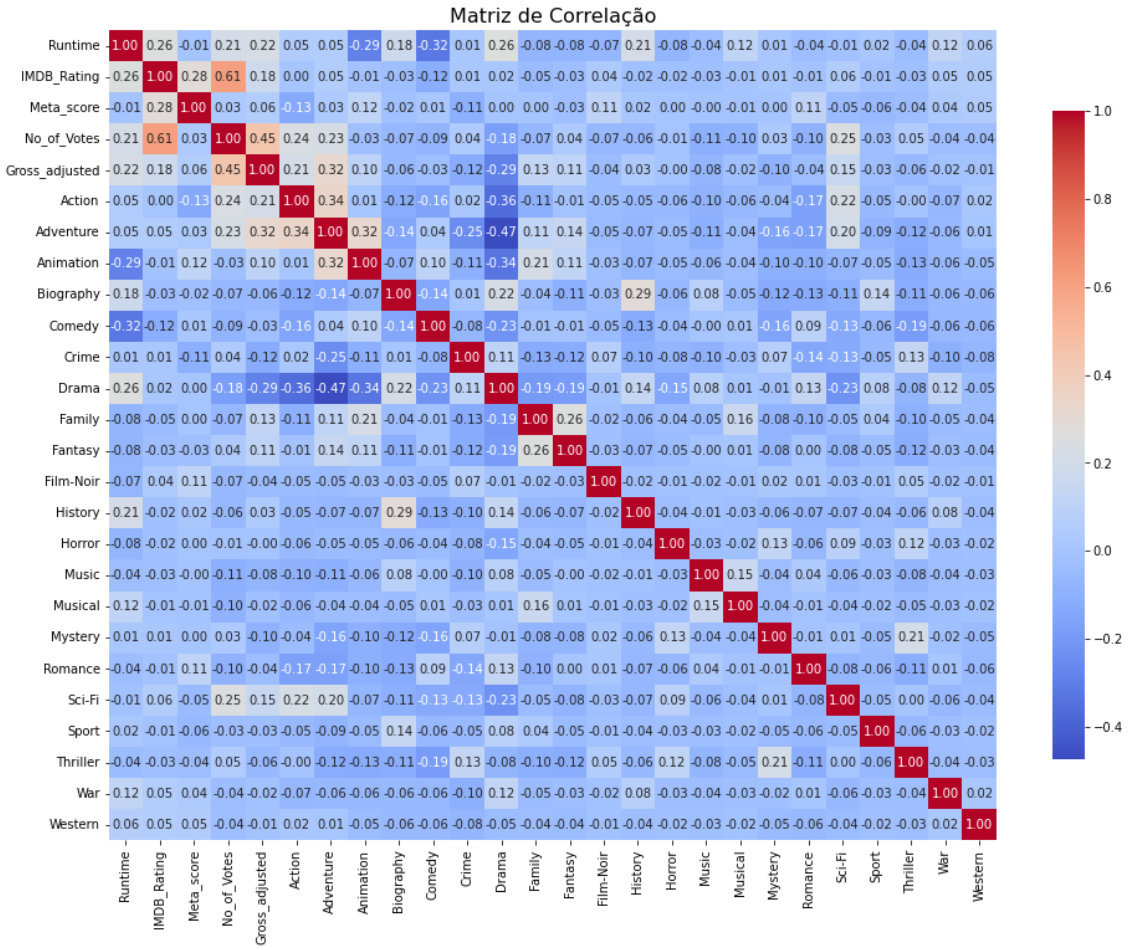
O gráfico abaixo me fez notar um padrão interessante nos títulos antes de 2000 (em sua esmagadora maioria). Na média, os diretores de títulos anterior a 2000 foram os que figuraram o top 15. O gráfico abaixo comprova ainda mais, uma vez que a base de dados contém mais filmes lançados no século XXI (363 após 2000 contra 350 no século anterior.





Isso leva a acreditar que o cinema era mais rentável no século passado que o atual. Graças a tecnologia, a ida ao cinema neste século apresentou ser menos interessante que no século passado – minha opinião – isso pode ser uma prova que o streaming fez com que pessoas fosse ao cinema com menos frequência, preferindo a comodidade do lar.

E por fim é sempre interessante olhar a matriz de correlação. Quanto maior a bilheteria do filme, maior o número de votos recebidos, o que é de se esperar. Os gêneros Drama e Aventura dificilmente irão fazer parte de um mesmo filme. Animação, Aventura e Ação tem uma correlação apreciável de modo que estes podem aparecer juntos compondo o gênero de um filme.



Respostas às perguntas

a) Pela correlação de gênero e faturamento eu recomendaria para uma pessoa um filme de Aventura.
b) Pela matriz de correlação, um filme de Aventura ou Sci-Fi espera-se ter um bom retorno no faturamento.
c) É possível ter uma ideia do gênero do filme realizando uma análise de sentimento do campo Overview.

Modelo de previsão da nota IMDB

Primeiramente eu removi todas as linhas que continha NA e aquelas que o dado no campo não corresponde ao esperado o que reduziu em quase 1/3 a base de dados (que já é pequena). O modelo de previsão que decidi leva em consideração as seguintes variáveis: Runtime, Meta_score, No_of_Votes, Gross_adjusted e campo Genre. No entanto como Genre é uma variável categórica, empreguei o *MultiLabelBinarizer* para transformar cada categoria em uma coluna. Quando se tem variáveis categóricas (não da ideia de quantidade nem ordem) o algoritmo escolhido é o ideal.

O modelo selecionado foi o RandomForestRegressor porque pode capturar relações não lineares, é robusto a outliers e não requer muita preparação dos dados (como normalização) e se comporta bem com variáveis categóricas e numéricas. Como a base de dados é pequena o modelo é treinado muito rápido.

Para medir a performance utilizei o R^2 que se próximo de 1 mede o quão bem o modelo explica a variância dos dados.

Para o filme *Sawshank Redemption* o modelo previu uma nota de 8,8 de 10.