Introdução

Olá, seja bem-vinda e bem-vindo ao notebook da **aula 03**! A partir desta aula iremos analisar e disc
dados junto com você. Por isso, será **importante que as discussões nos vídeos sejam acompanha**
todos os processos das análises.

Nessa aula utilizaremos uma base totalmente nova, que nós também não conhecíamos até o mor
Você vai acompanhar a exploração e, principalmente, as dificuldades ao analisar uma base de dac

Vamos começar importando a nossa base de dados! Nessa aula iremos trabalhar com a IMBD 50
contém uma série de informações sobre filmes, sendo uma pequena amostra da famosa base de

```
import pandas as pd
imdb = pd.read_csv("https://gist.githubusercontent.com/guilhermesilveira/24e271e68afe8fd25
imdb.head()
```

	color	director_name	num_critic_for_reviews	duration	director_facebook_likes	a
0	Color	James Cameron	723.0	178.0	0.0	
1	Color	Gore Verbinski	302.0	169.0	563.0	
2	Color	Sam Mendes	602.0	148.0	0.0	
3	Color	Christopher Nolan	813.0	164.0	22000.0	
4	NaN	Doug Walker	NaN	NaN	131.0	

Como você acompanhou, iniciamos a aula tentando conhecer as diversas colunas de cada filme e
chamou mais a atenção foi a color. Vamos conhecer quais valores temos nesta colunas?!

```
imdb["color"].unique()
```

Verificamos que essa coluna **color** informa se o filme é colorido ou é preto e branco. Vamos desc
filmes de cada tipo nós temos:

```
imdb["color"].value_counts()
```

```
imdb["color"].value_counts(normalize=True)
```

Agora já descobrimos quantos filmes coloridos e preto e branco temos, e também sabemos que h filmes na base. Fizemos algo novo, que foi chamar o `value_counts()`, passando o parâmetro **norı** Desse modo, já calculamos qual é a participação de cada um dos tipos de filmes (**95% são filmes**

Excelente! Agora vamos explorar outra coluna a fim de conhecer os diretores que tem mais filmes dados (**lembrando que nossa base é uma amostra muito pequena da realidade**)

```
imdb["director_name"].value_counts()
```

Steven Spielberg e Woody Allen são os diretores com mais filmes no **IMDB 5000**.

Continuando com nossa exploração de algumas informações, vamos olhar para o número de críti

```
imdb["num_critic_for_reviews"]
```

```
imdb["num_critic_for_reviews"].describe()
```

Veja que as colunas **color** e **director_name** são *strings*, não fazendo sentido olhar para médias, me
Olhar para o número de avaliações já pode ser interessante, por isso usamos o `.describe()`.

Agora podemos até plotar um histograma para avaliar o número de review.

```
import seaborn as sns
sns.set_style("whitegrid")
imdb["num_critic_for_reviews"].plot(kind='hist')
```

Verificamos que poucos filmes tem mais de 500 votos, por isso um paralelo que podemos fazer é
muitos votos são mais populares e filmes com poucos votos não são tão populares. Logo, pelo hi
evidente que poucos filmes fazem muito muito sucesso. Claro que não conseguimos afirmar isso
pois, novamente, estamos lidando com um número restrito de dados, mas são pontos interessant

Outra informação interessante de se analisar, são os orçamentos e receitas de um filme, ou seja o
financeiro. Vamos começar pelo gross:

```
imdb["gross"].hist()
```

Como você deve ter reparado, essa é a primeira vez que as escalas estão totalmente diferentes, p valores tão altos que a escala teve que ser de centena de milhões. Veja como pouquíssimos filme **faturamento**, o que nos acende um primeiro alerta de que tem algo estranho (ou temos filmes que dinheiro neste dataset).

Vamos tentar conhecer quais são esses filmes com faturamento astronômico.

```
imdb.sort_values("gross", ascending=False).head()
```

Nessa lista temos **Avatar, Titanic, Jurassic World e The Avengers**, o que parece fazer sentido para sabemos que esses foram filmes com bilheterias gigantescas. Analisando esses dados consegui os maiores faturamentos fazem sentido, mas encontramos um problema nos dados, dado que en linhas diplicadas. Podemos usar o pandas para remover esses dados, mas por enquanto vamos n informações (Se estiver curioso em saber como se faz, consulte o `.drop_duplicates()`).

Maravilha, agora temos o faturamento e parece estar OK. Queremos começar a responder alguma delas é: será que filmes coloridos tem faturamento maior que filmes preto e branco?

Para começar a responder essa pergunta precisamos transformar a coluna Color:

```
color_or_bw = imdb.query("color in ['Color', ' Black and White']")
```

```
color_or_bw["color_0_ou_1"] = (color_or_bw["color"]=="Color") * 1
color_or_bw["color_0_ou_1"].value_counts()
```

```
color_or_bw.head()
```

Veja que agora nós temos uma última coluna em nosso dataframe com valores 0 e 1. Agora pode gráficos com essa informação de filmes coloridos ou não.

P.S: Em aula tivemos problemas porque Black and White tinha um espaço no início, vou cortar ess no notebook, mas reforço a importância de acompanhar este processo no vídeo.

```
sns.scatterplot(data=color_or_bw, x="color_0_ou_1", y="gross")
```

Então plotamos nossos dados com um displot! Existem várias formas de visualizar essa informaç
essa nos ajuda a comparar os resultados. Repare como filmes coloridos tem valores bem maiores
esperado), mas também temos pontos bem altos em filmes preto e branco, chamando muito aten

Vamos explorar algumas estatísticas destes filmes:

```
color_or_bw.groupby("color").mean()["gross"]
```

```
color_or_bw.groupby("color").mean()["imdb_score"]
```

```
color_or_bw.groupby("color").median()["imdb_score"]
```

Das estatísticas temos duas bem interessantes, a média e mediana das notas de filmes preto e br
Há várias possíveis explicações sobre o porquê disso, reflita aí sobre algumas delas e compartilhe

A partir de agora, vamos fazer uma investigação melhor em relação às finanças dos filmes (fatura
orçamento). Vamos iniciar plotando e interpretando um gráfico de **gross** por **budget**:

```
budget_gross=  imdb[["budget", "gross"]].dropna().query("budget >0 | gross > 0")

sns.scatterplot(x="budget", y="gross", data = budget_gross)
```

Para plotar os dados, primeiro removemos as linhas com informações de faturamento e orçamen
com valores igual a 0, para então gerar o gráfico.

Agora vamos analisar esse gráfico juntos, veja que a escala de **budget** mudou, agora é **e10**. Repar
poucos filmes tem orçamentos tão grandes assim, e seus faturamentos são muito baixos. Será qu
problema nos dados? Vamos investigar melhor!

```
imdb.sort_values("budget", ascending=False).head()
```

Ordenando os dados pelo **budget** percebemos que as primeiras posições são de filmes asiáticos.
trouxe um ponto interessante para a investigação, pois países como a Coreia usam moedas que t
decimais a mais que o dólar. Então provavelmente o que está ocorrendo é que os dados de orçam
na moeda local, por isso detectamos valores tão discrepantes.

Como não temos garantia dos números, vamos precisar trabalhar apenas com filmes americanos
que tanto gross e budget estão em dólares. Então vamos iniciar esse processo:

Não esqueca de compartilhar a solucão dos seus desafios c

```
imdb["country"].unique()
```

Veja que temos filmes de diversos locais de origem:

```
imdb = imdb.drop_duplicates()
imdb_usa = imdb.query("country == 'USA'")
imdb_usa.sort_values("budget", ascending=False).head()
```

Agora temos os dados para fazer uma análise melhor entre gross e budget. Vamos plotar o gráfic

```
budget_gross = imdb_usa[["budget", "gross"]].dropna().query("budget >0 | gross > 0")

sns.scatterplot(x="budget", y="gross", data = budget_gross)
```

Veja que interessante, aparentemente temos uma relação entre orçamento e faturamento. Quanto orçamento, maior o faturamento.

Já que estamos trabalhando com orçamento e faturamento, podemos construir uma nova informa analisar. De forma bem simplista esse processo de construir novas informações a partir das exist conhecido como feature engineering.

```
imdb_usa['lucro'] = imdb_usa['gross'] - imdb_usa['budget']
```

```
budget_gross = imdb_usa.query("budget >0 | gross > 0")[["budget", "lucro"]].dropna()
```

```
sns.scatterplot(x="budget", y="lucro", data = budget_gross)
```

MUito bom! Nós construímos nossa coluna lucro na base de dados e plotamos o orçamento cont

Repare que temos pontos interessantes nesta visualização, um deles são esses filmes com muito
Isso pode ser um prejuizo real, mas também podem ser filmes que ainda não tiveram tempo de re
investimento (lançamentos recentes). Outros pontos interessantes de se anlisar seriam os filmes
orçamentos e muito lucro, será que são estão corretos ou pode ser algum erro da base? Parece qu
gastar uma tonelada de dinheiro vai gerar lucros absurdos, será que é isso é verdade?

Esse gráfico é muito rico em informações, vale a pena você gastar um tempo criando hipóteses.

Já que essa nova feature (lucro) parace ser interessante de se analisar, vamos continuar! Mas ago
lucro em relação ao ano de produção:

```
budget_gross = imdb_usa.query("budget >0 | gross > 0")[["title_year", "lucro"]].dropna()

sns.scatterplot(x="title_year", y="lucro", data = budget_gross)
```

Olha que legal esse gráfico, veja como alguns pontos mais recentes reforça a teoria de que alguns
ainda não ter recuperado o dinheiro investido (Claro que temos muitas variáveis para se analisar, r
relevante).

Outro ponto que chama muito atenção, são os filmes da década de 30 e 40 com lucros tão altos. (
filmes? Bom, essa pergunta você vai responder no desafio do Paulo, que está louco para descobri

Falando em Paulo, ele sugeriu uma análise com os nome dos diretores e o orçamento de seus filn
conseguimos concluir alguma coisa:

```
filmes_por_diretor = imdb_usa["director_name"].value_counts()
gross_director = imdb_usa[["director_name", "gross"]].set_index("director_name").join(film
gross_director.columns=["dindin", "filmes_irmaos"]
gross_director = gross_director.reset_index()
gross_director.head()
```

```
sns.scatterplot(x="filmes_irmaos", y="dindin", data = gross_director)
```

Essa imagem aparentemente não é muito conclusiva, então não conseguimos inferir tantas inforn

Esse processo de gerar dados, visualizações e acabar não sendo conclusivo é muito comum na vi de dados, pode ir se acostumando =P.

Para finalizar, que tal realizar uma análise das correlações dos dados? EXistem várias formas de c correlação, esse é um assunto denso.Você pode ler mais sobre essas métricas neste link.

Vamos então inciar a análise das correlações plotando o pairplot.

```
sns.pairplot(data = imdb_usa[["gross", "budget", "lucro", "title_year"]])
```

O pairplot mostra muita informação e a melhor forma de você entender é assistindo as conclusõe
sobre esses gráficos na vídeoaula.

Embora plotamos um monte de informação, não necessariamente reduzimos a correlação em um
simplificar a análise. Vamos fazer isso com a ajuda do `.corr()` do pandas.

```
imdb_usa[["gross", "budget", "lucro", "title_year"]].corr()
```

Com o pandas é simples de se calcular a correlação, mas precisamos saber interpretar os resulta
isso?

A correlação é uma métrica que vai de 1 a -1. Quando a correlação é 1, dizemos que é totalmente
(relação linear perfeita e positiva), ou seja se uma variável aumenta em 10 a outra também irá aur
Quando o valor da correlação é -1, também temos variáveis totalmente correlacionda, só que de m
(relação linear perfeita negativa), neste caso, se uma variável aumenta em 10 a outra reduz em 10
correlação é 0 temos a inexistência de correlação, ou seja, uma variável não tem influêcia sobre a

Agora sim, entendido sobre a correlação vamos analisar as nossas. Veja que lucro e gross tem un
o que indica que quanto maior o orçamento maior o lucro (mas repare que a correlação não é perf
e lucro tem correlação negativa, mas muito perto de zero (ou seja quase não tem correlação). Viu
conseguimos analisar muitas coisas com a correlação?! Pense e tente analisar os outros casos ta

Com isso chegamos ao final de mais uma aula da #quarentenadados. E aí, o que está achando, ca
e ao mesmo tempo mais complexo né?

O que importa é estar iniciando e entendendo o que fazemos para analisar os dados! **Continue ate
que vai valer a pena.** Vamos praticar?

Crie seu próprio notebook, reproduza nossa aula e resolva os desafios que deixamos para vocês

Até a próxima aula!

**P.S: A partir de agora teremos muitos desafios envolvendo mais análises e conclusões, então nã
"gabarito". O importante é você compartilhar suas soluções com os colegas e debater os seus re
outras pessoas**

Desafio 1 do Thiago Gonçalves

Plotar e analisar o Boxplot da média (coluna imbd_score) dos filmes em preto e branco e colorido

Desafio 2 do Guilherme Silveira

No gráfico de **budget por lucro** temos um ponto com muito custo e prejuizo, descubra com é esse
próximo de 2.5).

Desafio 3 do Guilherme Silveira

Em aula falamos que talvez, filmes mais recentes podem ter prejuizo pois ainda não tiveram temp
investimento. Analise essas informações e nos conte quais foram suas conclusões.

Desafio 4 do Paulo Silveira

Quais foram os filmes da decada pré 2° guerra que tiveram muito lucro.

Desafio 5 do Paulo Silveira

No gráfico de **filmes_irmaos por dindin** temos alguns pontos estranhos entre 15 e 20. Confirme a
Paulo que o cidadão estranho é o Woody Allen. (Se ele tiver errado pode cornete nas redes sociais

Desafio 6 do Thiago Gonçalves

Analise mais detalhadamente o gráfico pairplot, gaste um tempo pensando e tentando enteder os

Desafio 7 do Thiago Gonçalves

Calcular a correlação apenas dos filmes pós anos 2000 (Jogar fora filmes antes de 2000) e interp
correlação.

Desafio 8 do Allan Spadini

Tentar encontrar uma reta, pode ser com uma régua no monitor (não faça isso), com o excel/goog
python, no gráfico que parece se aproximar com uma reta (por exemplo budget/lucro, gross/lucro

Desafio 9 da Thais André

Analisar e interpretar a correlação de outras variáveis além das feitas em sala (notas é uma boa).
avaliações por ano pode ser também uma feature.

Não esqueça de compartilhar a solução dos seus desafios c instrutores, seja no Twitter, seja LinkedIn. Boa sorte!