

Olá Dev,

Estamos muito felizes que você chegou até aqui!!! Parabéns pelo seu desempenho ao longo do curso.

A última proposta que temos para você é a realização de um pequeno projeto com uma temática de seu interesse (Ex. Animais, time de Futebol, Spotify, Filmes, Olimpíadas....) para analisar algo relacionado a ela (Ex. As músicas estão diminuindo de duração ao longo do tempo? A chance de um time perder em casa é menor ? Um país com menos imposto é mais feliz?) utilizando tudo que você aprendeu de Python no curso até então.

A ideia é que você encontre alguma base de dados que contenha todas as informações necessárias para responder sua hipótese e faça uma análise bem detalhada a fim de chegar a uma conclusão considerável, com gráficos e tudo.

Instruções gerais para realização do trabalho final do curso de Python

Esse arquivo contém as orientações para a realização do trabalho final do curso. Siga todas as orientações para que o resultado da atividade esteja apto para avaliação. Não seguir alguma delas implica em nota 0 na atividade.

- Assim como consta no [programa do curso](#), o trabalho resultante deve ser entregue via Forms até as 23h59 do dia 10/12/2023. Forms de envio: <https://forms.gle/RA5Gc2Dg2kmUznoT8>
- A entrega deve ser composta de um único arquivo em formato .ipynb contendo todo o código construído e o resultado desse código. Arquivos entregues em outros formatos serão desconsiderados.
- A atividade deve ser realizada no **Python**. Embora o **Jupyter Notebook** seja a opção mais indicada para essa atividade, fique a vontade para utilizar a **IDE** que preferir.
- Todas as etapas do trabalho devem ser comentadas. Apresentar apenas as linhas de código não é suficiente para uma avaliação do conhecimento adquirido ao longo do curso. **Fazer não é suficiente, é preciso deixar claro que se sabe o que se está fazendo.**
- Ao realizar o Projeto Final você estará apto para ganhar AAC (Somente alunos FEA-USP) e participar do processo seletivo (Somente alunos USP - Butantã) mediante às outras regras do nosso regulamento.
- Haverá duas monitorias para tratar sobre o projeto final (30/11 e 02/12)

1. Base de Dados

Alguns sites são recomendados para que você encontre a base de dados perfeita para você, com temáticas diversas.

Portal Brasileiro de Dados Abertos (dados.gov.br):

dados.gov.br: Portal brasileiro que reúne diversas bases de dados governamentais.

World Bank Open Data:

[World Bank Data](https://data.worldbank.org/): Oferece uma ampla variedade de dados econômicos e sociais globais.

Google Dataset Search:

[Google Dataset Search](https://datasetsearch.research.google.com/): Ferramenta do Google para pesquisar conjuntos de dados em uma variedade de tópicos.

Kaggle Datasets:

[Kaggle Datasets](https://www.kaggle.com/): Comunidade online de cientistas de dados que compartilham conjuntos de dados, competem em desafios e colaboram em projetos.

IPEA:

[IPEA Data](https://repositorio.ipea.gov.br/data/) - Instituto de Pesquisa Econômica Aplicada (IPEA) - Oferece uma variedade de dados econômicos e sociais, incluindo pesquisas e análises governamentais.

Alguns temas que podem interessar:

Business and Industry Reports Kaggle
Boston House Prices-Advanced Regression Techniques Kaggle
World Development Indicators 2022 Kaggle
Premier League Matches 1993-2022 Kaggle
Netflix popular movies dataset Kaggle
Loneliness and Social Connections Kaggle
Global Economic Indicators Kaggle
Maven Pizza Challenge Dataset Kaggle
U.S. Incomes by Occupation and Gender Kaggle
Gender Development Index 2019 Kaggle

Impact of Covid-19 Pandemic on the Global Economy Kaggle
FIFA World Cup 2022 🏆 Kaggle
Historical Military Battles Kaggle
International football results from 1872 to 2022 Kaggle
The Movies Dataset Kaggle

2. Estrutura do arquivo

O arquivo final entregue deve ter uma estrutura específica e ser composto por 5 seções, **nessa ordem**:

1. **INTRODUÇÃO**: apresentação do contexto e do modelo formal por trás da hipótese a ser testada, seguido da descrição **clara** da própria hipótese;
2. **DADOS**: descrição da base de dados e suas principais características.
3. **RESULTADOS**: apresentação das evidências empíricas relacionadas à hipótese (gráficos e estatísticas descritivas).
4. **CONCLUSÃO**: considerações finais e sugestão de passos futuros para enriquecimento da análise.

3. Método de avaliação

	0-50%	50%-75%	75-100%
Questões	Questões simplistas, não motivadas	Questões apropriadas, coerentes e motivadas	Questões bem motivadas, interessantes, perspicazes e novas
Apresentação	<p>A apresentação verbal é ilógica, incorreta ou incoerente.</p> <p>A apresentação visual é confusa, desconexa ou ilegível</p> <p>Apresentação verbal e visual não relacionada</p>	<p>Apresentação verbal parcialmente correta, mas incompleta ou pouco convincente.</p> <p>A apresentação visual é legível e clara</p> <p>Apresentação verbal e visual relacionada</p>	<p>A apresentação verbal é correta, completa e convincente</p> <p>A apresentação visual é atraente, informativa e nítida</p> <p>Apresentação verbal e visual claramente relacionada</p>

Resultados	Faltam conclusões, são incorretas ou não são baseadas em análises Escolha inadequada de gráficos; gráficos mal rotulados; gráficos faltando	Conclusões relevantes, mas parcialmente corretas ou parcialmente completas Os gráficos transmitem informações, mas carecem de contexto para interpretação	Conclusões relevantes explicitamente vinculadas à análise e ao contexto Os gráficos transmitem informações corretamente com informações de referência adequadas e apropriadas
Legibilidade	O código é confuso e mal organizado; código não utilizado ou irrelevante distrai ao ler o código. Nomes de variáveis e funções não ajudam a entender o código.	O código é razoavelmente bem organizado. Há pouco código não utilizado ou irrelevante, ou esse código foi removido dos arquivos principais do projeto. Nomes de variáveis e funções geralmente significativos e úteis para compreensão.	Código muito bem organizado. Nenhum código irrelevante ou confuso. Os nomes de variáveis e funções têm uma relação clara com sua finalidade no código. O código é fácil de ler e entender.

4. Exemplo de Projeto Final

Análise sobre a mudança da duração dos filmes na Era Digital

Introdução

De acordo com um estudo realizado pela Microsoft entre 2000 e 2013, o estilo de vida pautado pelo uso excessivo de dispositivos com telas afetou negativamente a concentração e o foco das pessoas. No relatório da pesquisa, diz-se que "Usuários de múltiplas telas acham difícil filtrar estímulos irrelevantes – são mais facilmente distraídos por múltiplos fluxos de mídia" e "Estilos de vida digitais afetam a capacidade de manter o foco por longos períodos de tempo." Assim sendo, **este trabalho tem como objetivo, a partir do auxílio da base de dados Netflix Data: Cleaning, Analysis and Visualization, verificar se esse declínio na atenção das pessoas causou algum impacto na duração dos filmes ao longo dos últimos anos, tendo em vista a potencial dificuldade de se assistir a filmes muito longos. Verificaremos, também, se existem diferenças de duração entre os gêneros dos filmes no mesmo período.**

```
# Importando as bibliotecas
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
import os
import plotly.express as px
```

```
#Importando a tabela
os.chdir(r'C:\Users\isarus\Downloads\Projeto comput final')
data = pd.read_csv("netflix1.csv")
```

Dados

```
[4]: #visualização 5 primeiras linhas
data.head()
```

	type	title	director	country	date_added	release_year	rating	duration	listed_in
0	Movie	Dick Johnson Is Dead	Kirsten Johnson	United States	9/25/2021	2020	PG-13	90 min	Documentaries
1	TV Show	Ganglands	Julien Leclercq	France	9/24/2021	2021	TV-MA	1 Season	Crime TV Shows, International TV Shows, TV Act...
2	TV Show	Midnight Mass	Mike Flanagan	United States	9/24/2021	2021	TV-MA	1 Season	TV Dramas, TV Horror, TV Mysteries
3	Movie	Confessions of an Invisible Girl	Bruno Garotti	Brazil	9/22/2021	2021	TV-PG	91 min	Children & Family Movies, Comedies
4	Movie	Sankofa	Haile Gerima	United States	9/24/2021	1993	TV-MA	125 min	Dramas, Independent Movies, International Movies

```
[5]: # quantidade de linhas e colunas
data.shape
```

```
t[5]: (8790, 9)
```

```
[6]: #informações gerais da base de dados
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 8790 entries, 0 to 8789
Data columns (total 9 columns):
```

...

Resultados:

Gráfico 1: Relação entre duração média dos filmes e ano de lançamento

```
[6]: px.line(data_mean, title = "Duração em minutos de filmes, de acordo com data de lançamento ", x = data_mean["lançamento"], y
```

Conclusão 1:

Observa-se pelo gráfico que entre os anos 1980 e 2000 há uma oscilação muito grande na duração média dos filmes. Podemos dizer que isso ocorre devido ao tamanho das amostras nesse período, que é consideravelmente menor do que no período seguinte; assim, estas não representam fielmente a população de todos os filmes - dentro e fora da Netflix. Por outro lado, a partir dos anos 2000, verificamos uma oscilação menor, já que temos uma amostra maior. Nesse período, que foi justamente o da consolidação do estilo de vida digital no mundo, marcado pelo crescimento exponencial do uso de computadores e criação da internet, observamos uma tendência significativa de redução na duração média dos filmes. Podemos garantir, portanto, que existe, no mínimo, uma correlação positiva entre o crescente uso de telas pelas pessoas e a redução na duração dos filmes, o que corrobora a nossa hipótese inicial.

2. Como essa mudança é observada para cada gênero de filme?

A nossa base de dados dividiu os gêneros em uma quantidade absurda, então, nós simplificamos as categorias desconsiderando especificidades que julgamos irrelevantes para a análise. Assim, ficamos com os 7 gêneros mais relevantes do cinema de todos os tempos, de acordo com uma pesquisa realizada pela Netflix em 2017. São esses: comédia, ação, drama, documentário, tv kids, terror e romance.

...

Resultados:

Gráfico 2.1: Relação duração em minutos e ano de lançamento de cada gênero

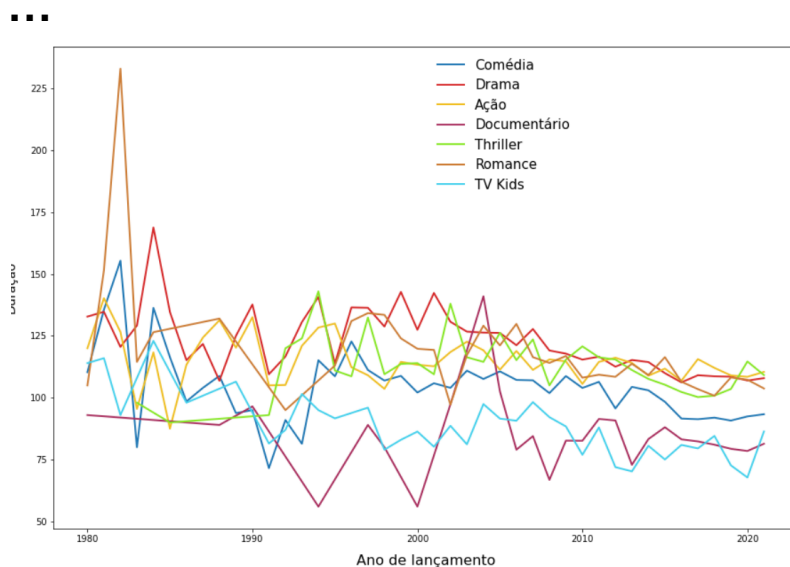
```
40]: #usando a função "for" selecionamos os radicais de cada palavra que desejamos ter no nosso dicio, a fim de incluir todos os
for i in ('comed',
        'drama',
        'action',
        'documentar',
        'thriller',
        'romant',
        'children'):
    dicio[f'df_{i}'] = data_mi.loc[data_mi.gênero.str.lower().str.contains(i)].groupby(by=['lançamento'], as_index=False).me

41]: dicio.keys()

41]: dict_keys(['df_comed', 'df_drama', 'df_action', 'df_documentar', 'df_thriller', 'df_romant', 'df_children'])

42]: #Para uma melhor compreensão dos gráficos, vamos alterar o nome no dicionário

dicio["df_comédia"]=dicio['df_comed']
del dicio['df_comed']
dicio["df_ação"]=dicio["df_action"]
del dicio['df_action']
dicio["df_documentário"]=dicio["df_documentar"]
del dicio['df_documentar']
dicio["df_romance"]=dicio["df_romant"]
del dicio['df_romant']
```



Conclusão 2:

Conclusão Final:

Por fim, podemos garantir que existe uma correlação positiva entre a intensificação do estilo de vida digital e a redução da duração dos filmes, com algum nível de dependência dos gêneros, abordada na conclusão 2. Apesar de verificarmos essa correlação, não podemos afirmar com 100% de certeza que a causa dessa redução foi o estilo de vida digital, pois correlação não implica causalidade. Como meio de resolver este problema, poderia-se buscar mais estudos e dados. Um possível estudo seria uma pesquisa sobre as motivações das grandes produtoras de filmes para diminuir a duração das obras; provavelmente, o retorno financeiro passou a ser maior com filmes não muito longos. Outro estudo que poderia ser feito é uma análise de preferência dos telespectadores. poderia-se selecionar pessoas que não levam um estilo de vida digital e pessoas que levam, para obter dados sobre a duração dos últimos filmes a que elas assistiram ou sobre qual duração é preferível a elas. Também como forma de separar, mesmo que parcialmente, pessoas que vivem ou não um estilo de vida digital, poderiam ser obtidos dados sobre a relação entre a idade das pessoas e a média de duração dos filmes a que elas já assistiram. Para a análise dos gêneros, poderiam ser utilizados base de dados sobre a quantidade de filmes de cada gênero produzida e suas bilheterias, a fim de verificar se os filmes que exigem uma atenção mais profunda tiveram seu sucesso reduzido na Era Digital. Esses mecanismos iriam no caminho de reforçar a nossa hipótese de causalidade.

Para ler o exemplo completo:

https://drive.google.com/file/d/1zVy3rAlyZ9T2LnorBXWdd6hmmHptGzIU/view?usp=drive_link

Dúvidas

Caso haja alguma outra dúvida acerca do Projeto Final, por favor utilizar o momento de Monitorias via Discord <https://discord.gg/DmE95Jme> nos dias 30/11 e 02/12 das 17h às 18h.