# Projeto 1

# Análise de Desempenho de Vendas

Este é um projeto que mostra minha capacidade de consultar dados do banco de dados Adventure Works DW usando SQL, depois usando Power BI para importar, analisar e visualizar.

\* (A imagem abaixo é obtida do site freepiks) \*

![](Salesman.jpg)



## Introdução

Esta é uma análise do desempenho de vendas da AdventureWorksDW. Isso é feito analisando dados de diferentes esquemas no AdventureWorks DWdatabase. Analisei os dados das tabelas dim\_customers, dim\_date, dim\_product, dim\_territory sales e fact\_internet sales. Usei o SQL para consultar os dados do banco de dados e importei para o Power BI, onde a análise real foi realizada.

\*\*\_Isenção de responsabilidade\_\*\* Esta não é uma empresa real, pois sabemos que o banco de dados DW da Adventure Works é compilado pela Microsoft para fins de aprendizado.

## Declaração do problema

O objetivo desta análise é

- Determinar a demanda atual do mercado por produtos e como ela varia por região/localização

- Conheça o desempenho de vendas atual e como ele se compara ao desempenho passado e aos benchmarks do setor

- Identificar as áreas onde o desempenho de vendas está faltando ou poderia ser melhorado.

- Para descobrir mercados-alvo ou produtos que tenham o maior impacto no sucesso de vendas.

- Finalmente, o objetivo é criar planos baseados em dados que possam ajudar a aumentar as vendas, melhorar a satisfação do cliente e impulsionar o crescimento corporativo.

## Habilidades e conceitos demonstrados:

- SQL (Select, Alias, Views)

- Conceitos de Power BI como:

- Criação de colunas para cálculos de datas, hora e duração

- Criação de indicadores-chave de desempenho (KPIs) e outros cálculos de negócios,

- Desenvolver cálculos DAX gerais que lidam com texto e números,

- Realizou cálculos DAX avançados para resolver medidas estatísticas e outras fórmulas matemáticas

- Modelagem de Dados,

- Medidas,

- Filtros,

- Dicas de ferramentas,

- Botões de página,

- Visualização de dados

---

## Fonte de dados:

Os dados utilizados para este trabalho são obtidos do banco de dados AdventureWorks 2019 2019. Estudei o Schema, Objetos relacionados ao Schema, dicionário de dados e encontrei as tabelas certas para a análise.

- Você pode encontrar um link para iniciar a instalação e restauração do banco de dados em sua máquina local. [here:](<https://youtu.be/VpY0Q_kwtIw>)

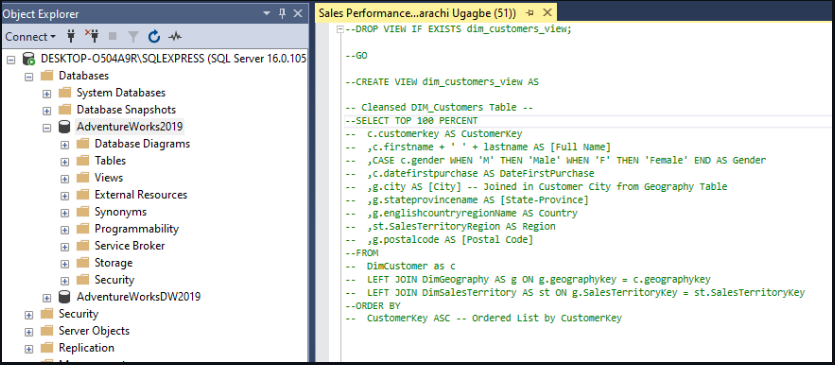
---

## Transformação de dados:

- Várias consultas estruturadas foram escritas para obter as tabelas corretas e, em seguida, salvas como visualizações no estúdio de gerenciamento de banco de dados em SQL. As tabelas e visualizações são:

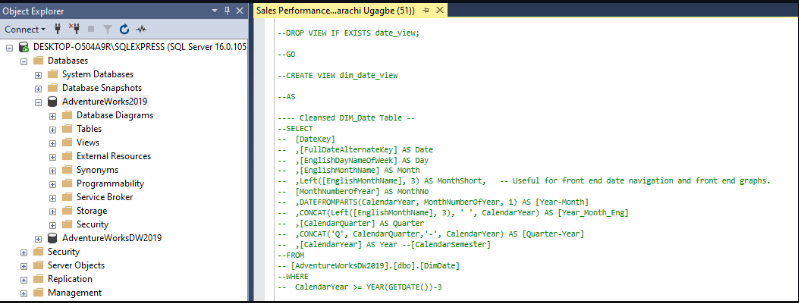
- Tabela de Clientes

![](CustomersView.png)



- Tabela de datas

![](DateView.png)



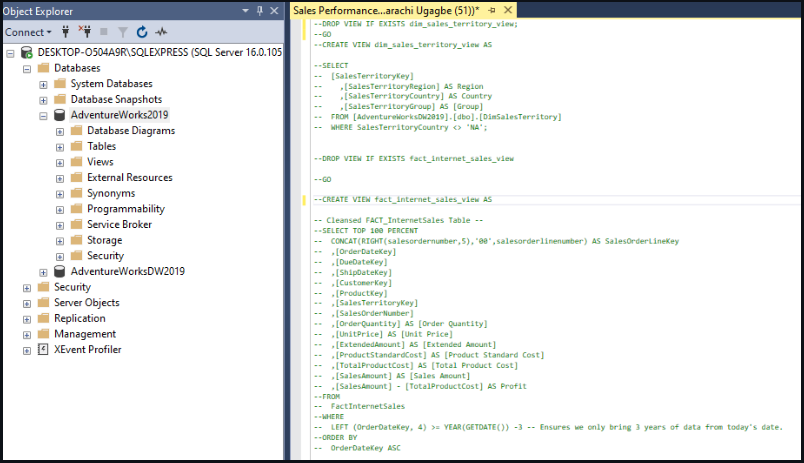
- Tabela de produtos

![](ProductView.png)



- Território e Tabela de Vendas

![](TerritoryandSalesView.png)



Observe que encontrei um script que atualizou as datas. Parte dela é mostrada aqui:

![](UpdateDate.png)



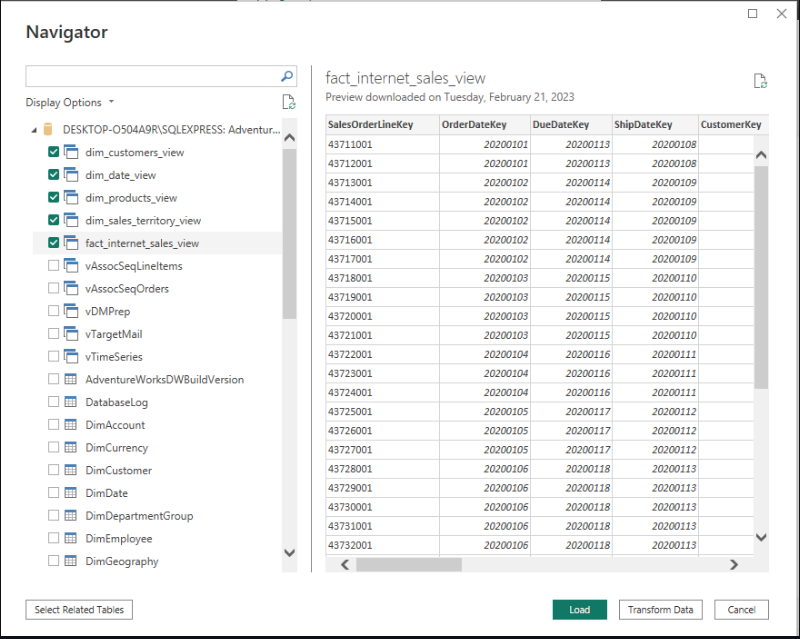
- Essas tabelas foram importadas para a área de trabalho do Power BI para análise.

![](SQLtoPowerBI.png)



- Eu escolhi as tabelas necessárias

![](NecessaryTables.png)



Transformei os dados no Power Query, verifiquei a qualidade da coluna, tipos de dados consistentes ou apropriados.

Comecei escrevendo vários DAX e criando medidas e colunas calculadas para obter as métricas certas para a análise de desempenho de vendas.

Também criei Medidas em uma tabela chamada "Medidas Chave".

As consultas que escrevi no SQL são compiladas e também carregadas aqui neste repositório como

[SalesPerformanceScript.sql](<https://github.com/Amarabright/Sales-Performance-Analysis/blob/main/SalesPerformanceScript.sql>)

---

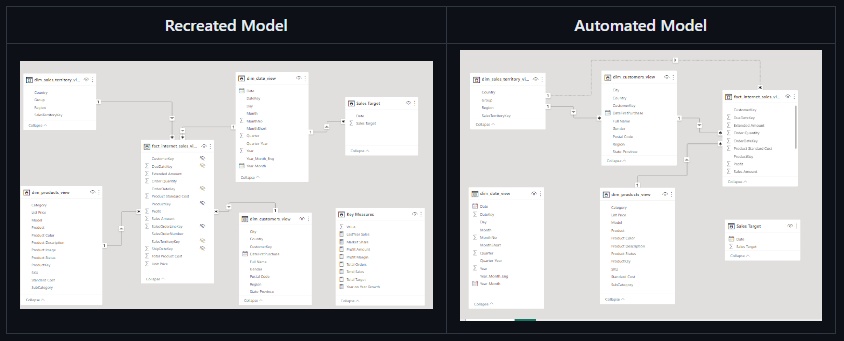
## Modelagem de Dados:

As tabelas foram unidas automaticamente criando relacionamentos com elas, o Power BI faz isso de forma inteligente. No entanto, como alguém que entende o conjunto de dados e deseja obter insights e informações específicas. Tive que criar outras relações e medidas para me capacitar. então fiz outro modelo. Criei um Star Schema que possui 1 tabela de fatos e 5 tabelas de dimensões.

Modelo recriado | Modelo Automatizado

:--------------------------: | :-----------------------------------:

![](RecreatedModel.png) | ![](AutomatedModel.png)



## Análise e visualização de dados

Várias expressões e funções foram feitas para chegar ao KPI ou Métrica desejado. Cheguei a um relatório com três (3) páginas denominadas:

- Página de vendas

- Página do produto e depois

- Página de informações

Por alguns motivos que não entendo, talvez motivos administrativos no serviço Power BI, o principal Influencers Visual que usei não estava visível ou talvez permitisse o uso em minhas configurações de administrador. Tentei achar um jeito, mas não consegui. Se eu encontrar uma maneira, deixarei o link para que você possa interagir com o relatório.

## Características do Relatório

A primeira página é denominada Página de vendas e contém informações sobre vendas. A página possui um botão Navegação na página na parte superior, logo à direita do título do relatório, para facilitar a navegação para as demais páginas. O botão também está na segunda página, mas a terceira página tem um botão "Voltar" que retorna à página de vendas ou produto.

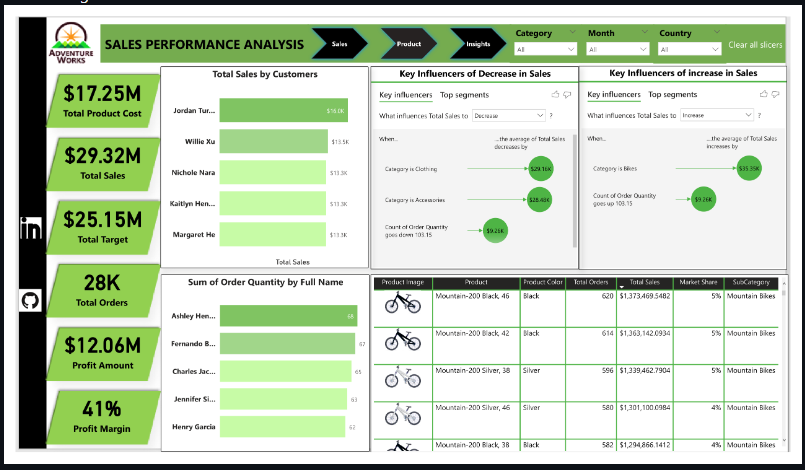
- Página de vendas

![](SalesPage.png)



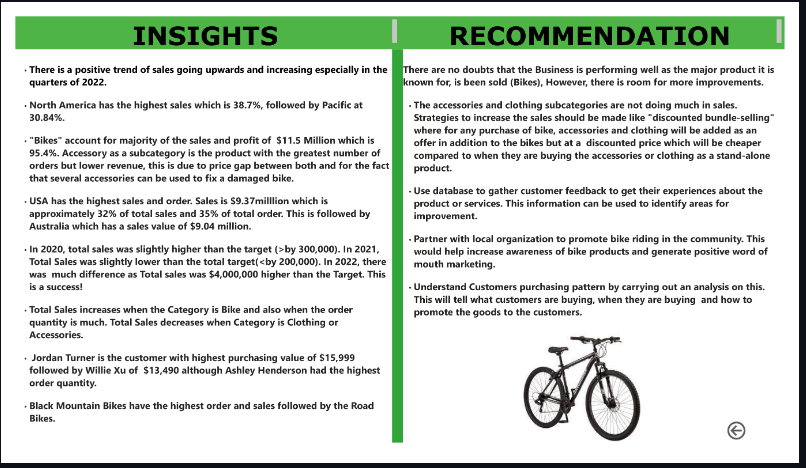
- Página do produto

![](ProductPage.png)



- Página de informações

![](InsightPage.png)



### Você pode interagir com o relatório

[here.](<https://app.powerbi.com/view?r=eyJrIjoiZTZjY2RiNTMtYTViYS00ZDUzLTkzOWUtMWEyOTg2MDgxNDUwIiwidCI6IjY0NTI5N2QxLWVkNmMtNDk0Ni1iNzg1LTE1MTI5MThlYmU2YiJ9>)

---

## Análise

- Há uma tendência positiva de vendas subindo e aumentando principalmente nos trimestres de 2022.

- A América do Norte é o local (grupo) que tem as maiores vendas com 38,7%, seguido pelo Pacífico com 30,84%.

- "Bicicletas" respondem pela maioria das vendas e lucro de US$ 11,5 milhões, o que representa 95,4%. Acessório como subcategoria é o produto com maior número de encomendas, mas menor faturamento, isso se deve a diferença de preço entre ambos e pelo fato de vários acessórios poderem ser usados ​​para consertar uma bicicleta danificada.

- Os EUA têm as maiores vendas e pedidos. As vendas são de US$ 9,37 milhões, o que representa aproximadamente 32% das vendas totais e 35% do pedido total. Isso é seguido pela Austrália, que tem um valor de vendas de $ 9,04 milhões.

- Em 2020, as vendas totais foram ligeiramente superiores à meta (> em 300.000). Em 2021, as vendas totais ficaram ligeiramente abaixo da meta total (< em 200.000). Em 2022, houve muita diferença, pois as vendas totais foram $ 4.000.000 acima da meta. Isso é um sucesso!

- As Vendas Totais aumentam quando a Categoria é Bicicleta e também quando a quantidade do pedido é grande. As vendas totais diminuem quando a categoria é roupas ou acessórios.

- Jordan Turner é o cliente com maior valor de compra de $ 15.999, seguido por Willie Xu de $ 13.490, embora Ashley Henderson tenha a maior quantidade de pedidos.

- As Black Mountain Bikes têm o maior pedido e vendas seguidas pelas Road Bikes.

---

## Recomendação

Não há dúvidas de que o Negócio está tendo um bom desempenho, pois o principal produto pelo qual é conhecido já foi vendido (Bicicletas). No entanto, há espaço para mais melhorias.

- As subcategorias de acessórios e roupas não estão ganhando muito nas vendas. Estratégias para aumentar as vendas devem ser feitas como "venda de pacotes com desconto" onde para qualquer compra de bicicleta, acessórios e roupas serão adicionados como oferta além das bicicletas, mas a um preço com desconto que será mais barato em comparação com quando são comprando os acessórios ou roupas como um produto autônomo.

- Use o banco de dados para coletar feedback do cliente para obter suas experiências sobre o produto ou serviço. Essas informações podem ser usadas para identificar áreas de melhoria.

- Faça parceria com uma organização local para promover o ciclismo na comunidade. Isso ajudaria a aumentar a conscientização sobre os produtos de bicicleta e geraria marketing boca a boca positivo.

- Compreender o padrão de compra dos Clientes realizando uma análise sobre o mesmo. Isso informará o que os clientes estão comprando, quando estão comprando e como promover os produtos para os clientes.

---

### Obrigado por ler.

Estou aberto para a função de analista de dados de nível básico.

Deixe-nos ter uma discussão sobre sua empresa e indústria agora!

<https://github.com/Amarabright/Sales-Performance-Analysis/tree/main>

# Projeto 2

# Análise do desempenho da produção

![](produção\_overview.png)



O total de unidades de itens produzidos é 2961

---

## Introdução

Este é um Projeto Power Bi de Análise de Produção de uma \*\*Empresa de Manufatura\*\*

O Projeto é analisar e obter insights para responder a perguntas cruciais e fazer a empresa toma decisões baseadas em dados.

\_Isenção de responsabilidade\_: \_Datasets do banco de dados Mircosoft Adevtureworks. Usado para demonstrar meus aprendizados e capacidades com a comprovação de decisões orientadas por dados com poder de BI\_

## Objetivo:

Para fornecer informações sobre o desempenho da produção por várias categorias e períodos de tempo

## Declaração do problema

1. Quais categorias de produtos contribuíram para o maior desempenho de produção

2. Em que período ocorreu o maior desempenho de produção?

## Habilidades e Conceitos demonstrados:

Os seguintes recursos do Power BI foram incorporados

- Obtendo dados do servidor (baseado em Microsoft)

- Modelar relacionamentos com tabelas

- DAX e Medida Calculada

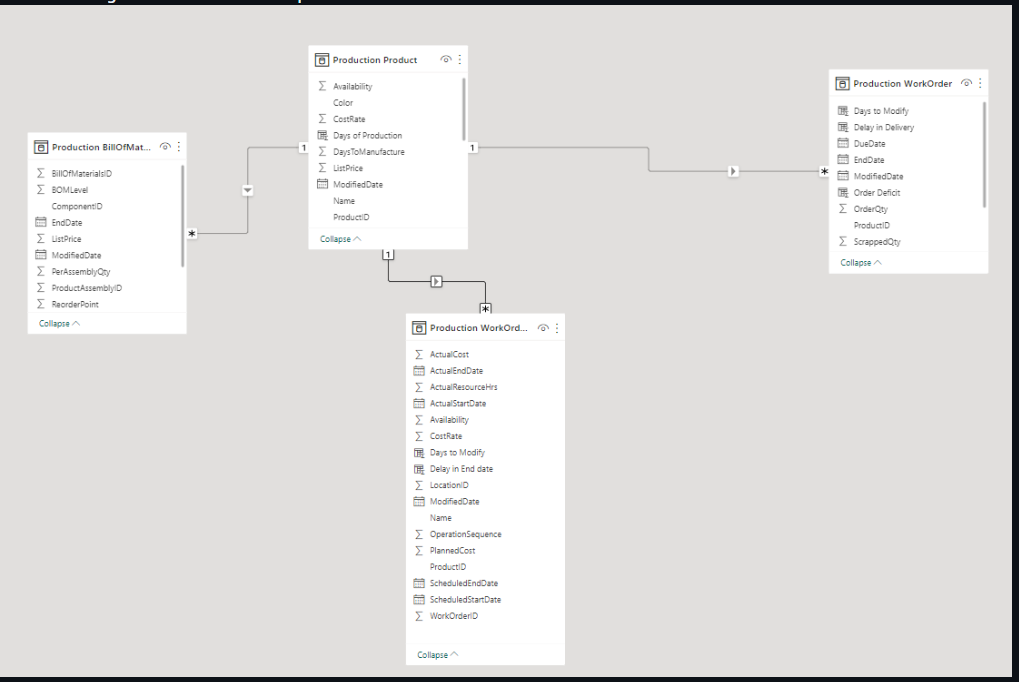
- Navegação da página e botão

\*\*Adquirir dados\*\*

Os dados foram solicitados usando o método do servidor por meio do computador host local \*\*O modelo é um esquema estelar\*\*

\*\*Modelagem\*\*D

![](data\_model.png)



## Visualização

O relatório é composto por 5 páginas:

1. Visão Geral

2. Produção por produto

3. Ordem de Serviço da Produção B

4. Roteamento de Produção por Ordem de Serviço

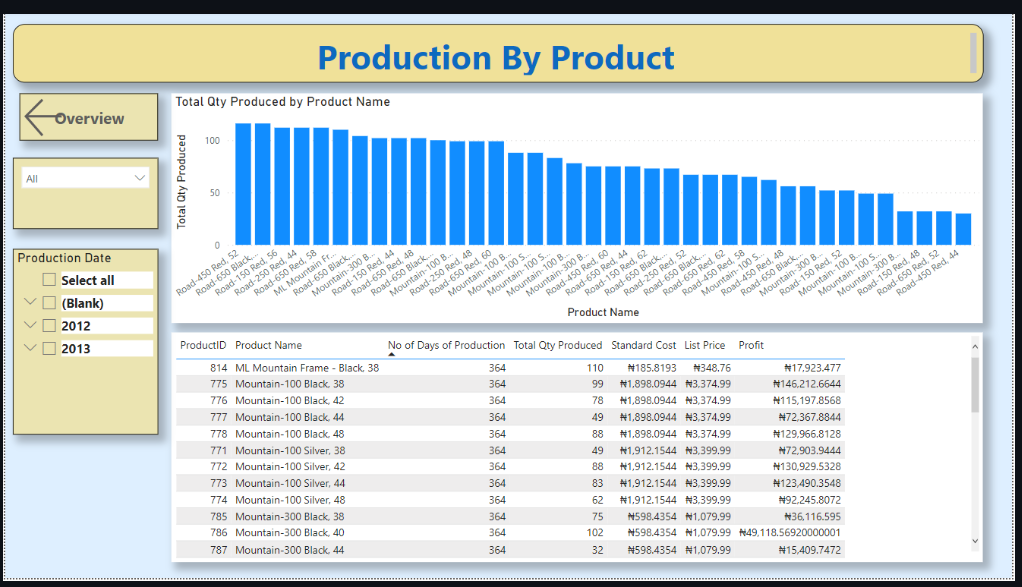
5. Produção por data

Você pode interagir com o relatório

[here](<https://app.powerbi.com/groups/me/reports/1a8e3174-c32d-4335-8106-e1f66041eeb4/ReportSection1ae8ff6bbf1920352db2?experience=power-bi>)

## Produção Por Produto

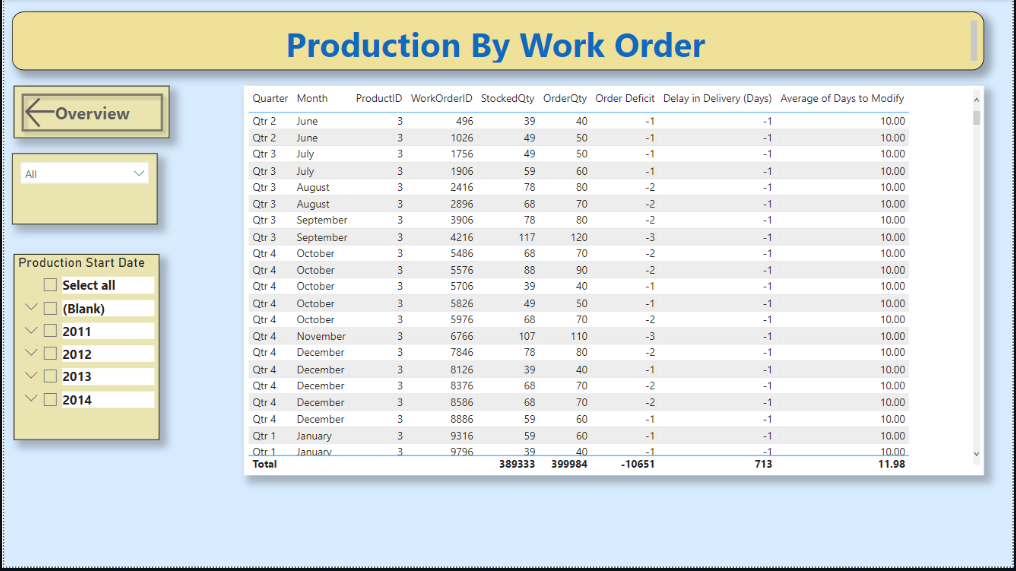
![](produção\_por\_produto.png)



Os itens com Produc ID 758 e 766 foram os mais produzidos com 116 unidades

## Produção por ordem de serviço

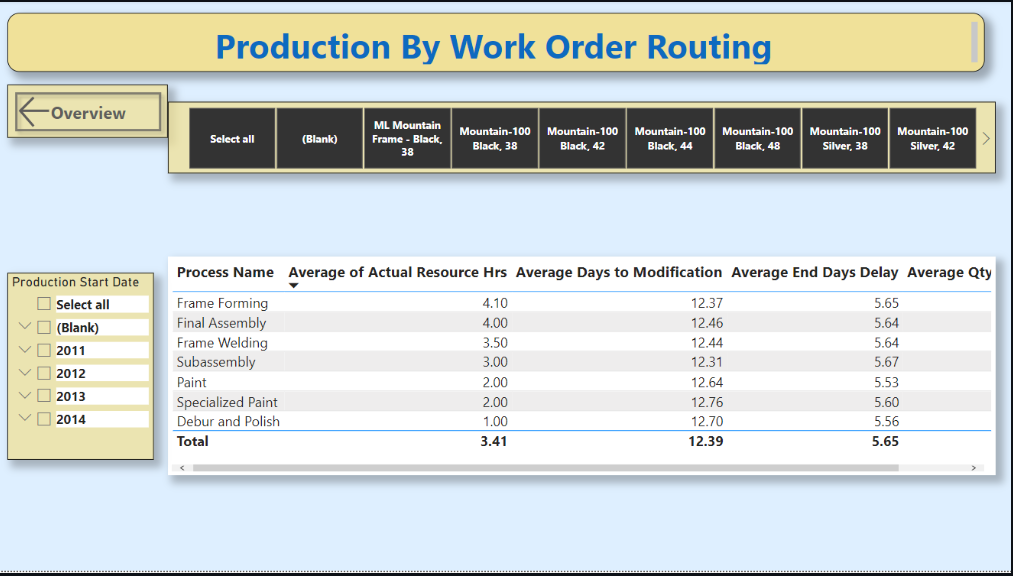
![](produção\_by\_work\_order.png)



Número médio de dias para fazer modificação nas unidades foi de 10 dias

## Roteamento de Produção por Ordem de Serviço

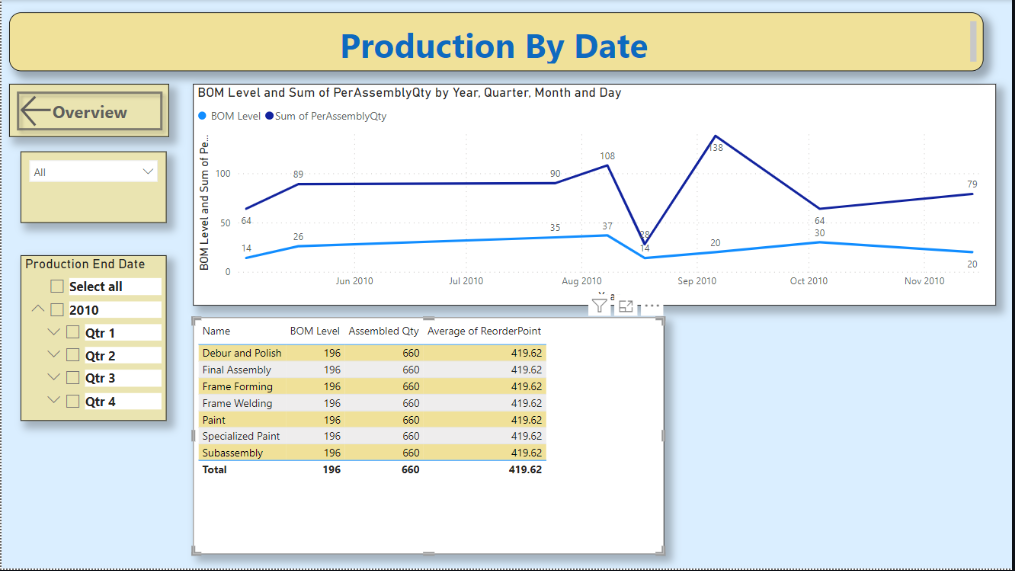
![](production\_by\_work\_order\_routing.png)



Em média, houve 5,65 atrasos para a data de término programada

## Produção por Data

![](produção\_por\_data.png)



A maior quantidade reunida ocorreu por volta de setembro de 2010

<https://github.com/omotox77/Project1/tree/main>

# Projeto 3

# IMBD-Limpeza-de-dados-com-R

---

![image](<https://user-images.githubusercontent.com/128150171/230795269-40d5cdc1-1f31-40a8-a2cc-d0ed3b54377c.png>)



## INTRODUÇÃO

Este é um projeto R de limpeza de dados de uma \*\*lista de filmes IMBD\*\*. Este projeto visa garantir que os dados sejam precisos, completos e consistentes para que possam ser utilizados para análise ou outra aplicação com confiança em sua qualidade.

## DECLARAÇÃO DO PROBLEMA

O conjunto de dados contém informações sobre filmes, como ID do título, ano de lançamento, gênero, duração, país, classificação do conteúdo, diretor, receita, votos e pontuação. Os dados não são limpos e contêm várias inconsistências, como formatos de data diferentes, valores ausentes e valores inconsistentes. O problema é limpar e pré-processar os dados para torná-los adequados para análises posteriores. Especificamente, os dados precisam ser padronizados para um formato comum, os valores ausentes precisam ser tratados e os valores inconsistentes precisam ser corrigidos. O objetivo é criar um conjunto de dados limpo que possa ser usado para analisar os dados \*\*.IMBD\*\*. Por exemplo:

- A coluna \*\*Ano de lançamento\*\* (por exemplo, '09 21 1972', '22 de fevereiro de 04', '10-29-99'), que precisa ser convertida em um formato consistente (por exemplo, 'AAAA-MM- DD') para análise.

- Algumas linhas contêm dados ausentes ou inválidos na coluna \*\*Duration\*\* (por exemplo, 'Nan', 'Inf'), que precisam ser limpos ou imputados para garantir a integridade dos dados.

- A coluna \*\*Gênero\*\* contém vários gêneros separados por vírgulas, que precisam ser divididos em colunas separadas para facilitar a análise.

- A coluna \*\*Classificação de conteúdo\*\* contém caracteres abreviados que precisam estar em sua forma completa para aqueles que não sabem o significado (por exemplo, 'R', 'Não classificado', 'PG-13', 'Aprovado') que precisam ser padronizados para análise.

- A coluna \*\*Receita\*\* contém cifrões e vírgulas que precisam ser removidos e os valores convertidos em números inteiros para análise.

- A coluna \*\*Votes\*\* contém vírgulas que precisam ser removidas e os valores convertidos em números inteiros para análise.

- A coluna \*\*Pontuação\*\* contém separadores decimais inconsistentes (por exemplo, ',' e '.'), que precisam ser padronizados para análise.

- Algumas linhas contêm dados ausentes ou inválidos em várias colunas, que precisam ser limpas ou imputadas para garantir a integridade dos dados.

## HABILIDADES DEMONSTRADAS

A organização e limpeza de dados são habilidades usadas para este projeto, como manipulação, remodelação e transformação de dados em vários formatos e estruturas. Este projeto utiliza vários pacotes R, incluindo stringr, dplyr, visdat, alignr, scales e lubridate para realizar as tarefas de limpeza de dados desejadas.

## FONTE DE DADOS

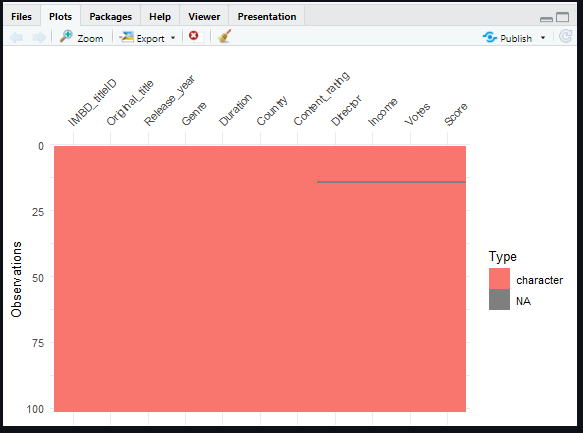
Este conjunto de dados foi adquirido de um usuário do Twitter que compartilhou um link para um Google Drive contendo os dados.

## AVALIANDO OS DADOS

O conjunto de dados foi avaliado quanto a anormalidades e qualidade dos dados. Observou-se que todos os dados estão na classe de caracteres e há alguns valores ausentes.

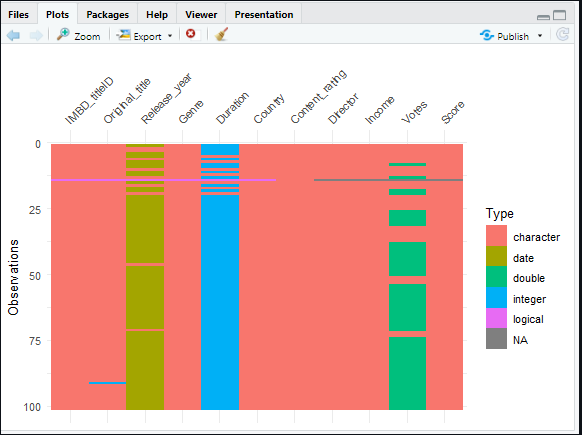
\*\*Usando visdat para nos mostrar visualmente como nossa coluna se parece\*\*

![05 messy\_visdat](<https://user-images.githubusercontent.com/128150171/230800032-6e5033bc-4526-4d08-a434-a4f639cccae5.png>)



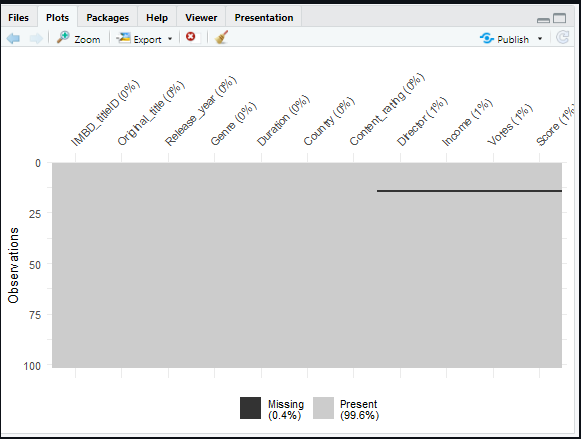
\*\*Usando vis\_guess para pedir a R para nos ajudar a examinar os dados e adivinhar a qual classe eles deveriam pertencer\*\*

![06 visguess](https://user-images.githubusercontent.com/128150171/230800034-82a637a3-07c8-470b-a85f-5eecad3f185e.png)



\*\*Usando vis\_miss para visualizar a coluna com observações ausentes\*\*

![07 vismiss](https://user-images.githubusercontent.com/128150171/230800039-289fff34-0843-4135-acef-32d0973606b1.png)



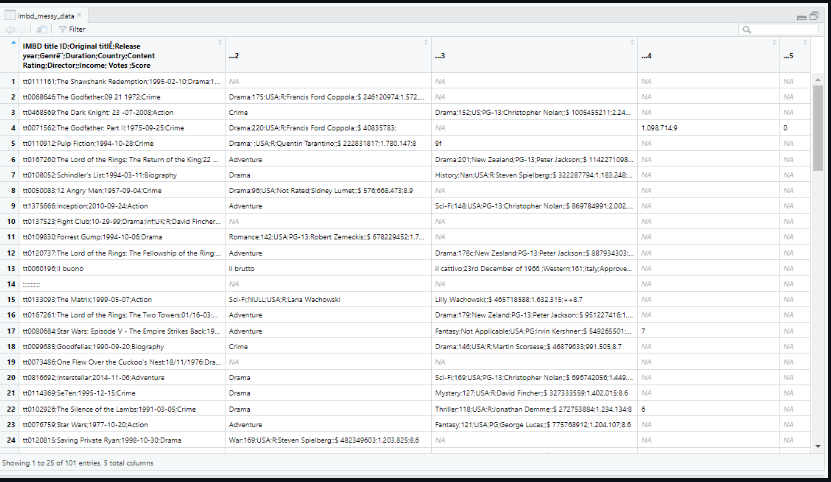
---

## TRANSFORMAÇÃO DE DADOS

Quando baixei inicialmente o conjunto de dados, visualizei-o no Excel antes de importá-lo para R. Como o conjunto de dados estava em um formato de arquivo CSV, converti-o em um arquivo Excel. Antes da limpeza, o conjunto de dados tinha aproximadamente 101 linhas e 5 colunas. No entanto, após a limpeza, uma linha não tinha dados e foi removida, resultando em um total de 100 linhas. Além disso, o número de colunas aumentou para 11 porque as colunas iniciais não foram devidamente distribuídas em suas variáveis ​​apropriadas. Além disso, o conjunto de dados foi separado apenas por ponto e vírgula em vez de ser separado por dois pontos, o que exigia etapas adicionais de limpeza. Você pode ver os dados originais [here](Imbd\_messy\_data.xlsx)

### Mesa bagunçada

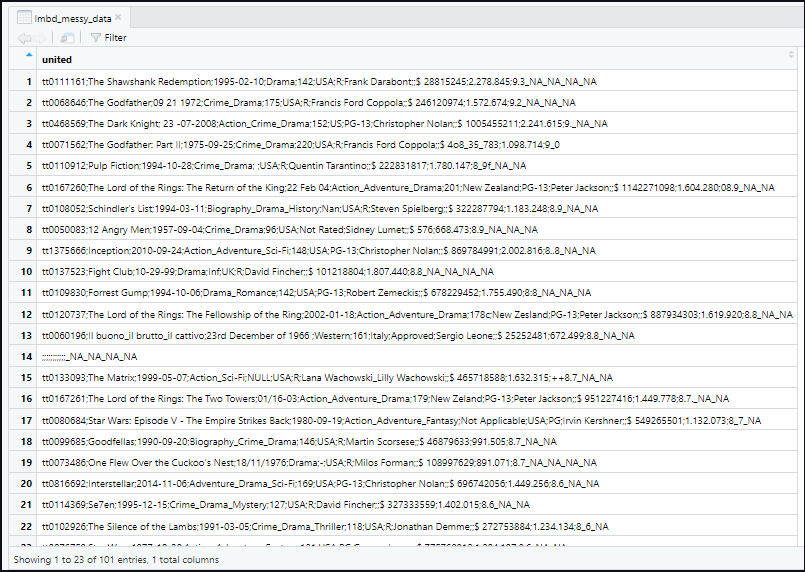
![01 Imessy\_MBD table](https://user-images.githubusercontent.com/128150171/230797854-53b921a6-e41d-4f32-800c-620a6b00467d.png)



\*Combinei todas as colunas em uma única coluna usando a função 'unite' e usei a função 'mutate' para garantir que todos os delimitadores fossem consistentes. Em seguida, separei os dados em suas respectivas colunas .\*

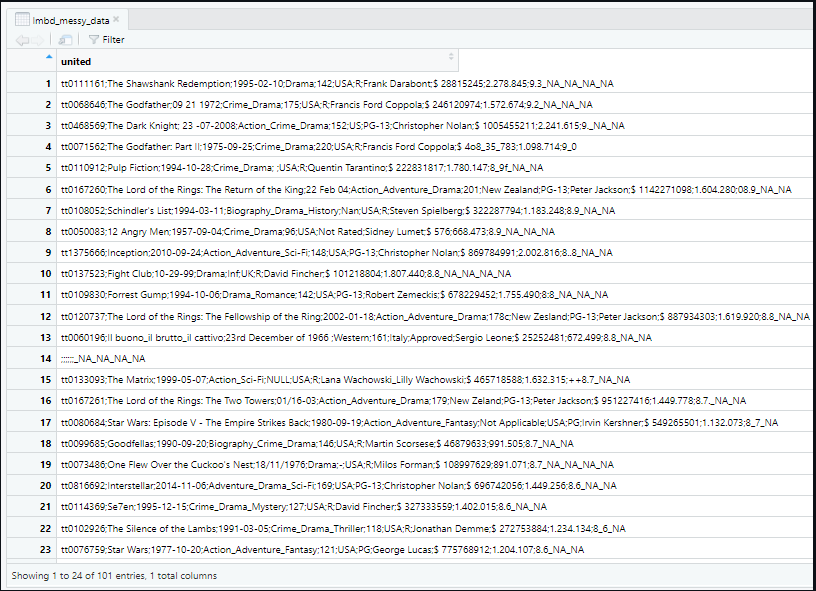
### UNIR

![02 unite](https://user-images.githubusercontent.com/128150171/230800743-6f13156f-6f4a-4419-b6d5-65d3fb675bea.png)



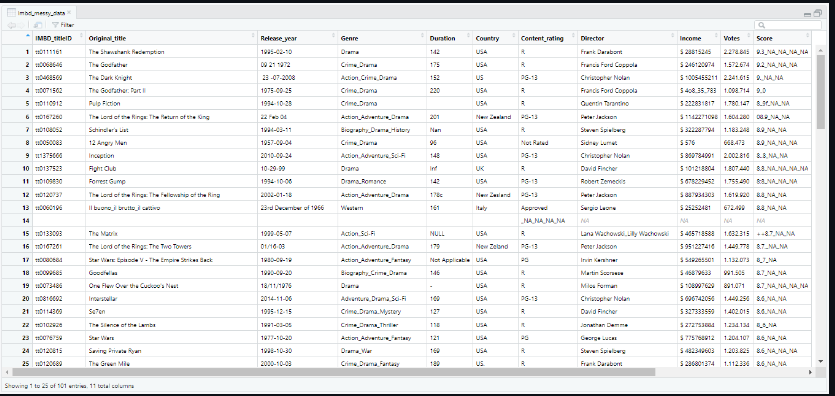
### MUTAÇÃO

![03 mutate](https://user-images.githubusercontent.com/128150171/230800762-ea7763aa-4db9-4f8b-aa02-78784c9dfff9.png)



### SEPARADO

![04 seperate](https://user-images.githubusercontent.com/128150171/230800872-8d7fd039-508a-4020-87bf-162130c08bdd.png)



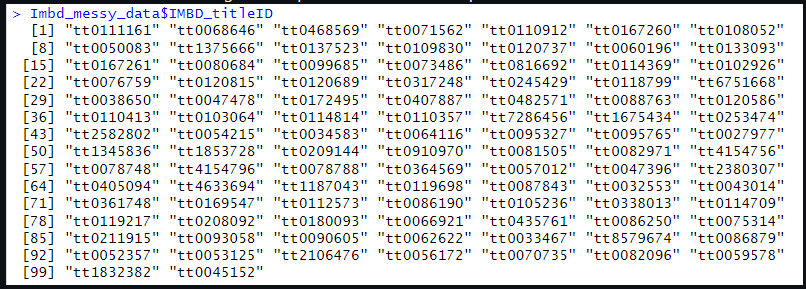
---

## TRANSFORMANDO AS COLUNAS UMA APÓS A OUTRA

### ID DO TÍTULO IMBD

- Possui um caractere de 9 strings de comprimento e nenhuma duplicata.

![09 imbd\_titleid](<https://user-images.githubusercontent.com/128150171/230798123-481ffa2c-570e-4c30-8116-e97b3220e264.png>)



### TÍTULO ORIGINAL

- Corrigido algumas palavras com erros ortográficos com alfabetos acentuados

Bagunçado | limpo

:-------------------------------------------: |:-------------------------------:

![10 messy\_Imbdoriginaltitle](https://user-images.githubusercontent.com/128150171/230798215-8f93d3ad-c786-42f1-a99f-40e90c0f85e1.png) | ![11 cleaned\_Imbdoriginaltitle](<https://user-images.githubusercontent.com/128150171/230798236-ba467bff-4110-447c-a8b0-83de007dc098.png>)



### ANO DE LANÇAMENTO

- Tornou o formato de data consistente

- Ano extraído a partir da data

- Inserir manualmente a data correta obtida no site imbd online para preencher a lacuna

- Convertido para uma classe numérica

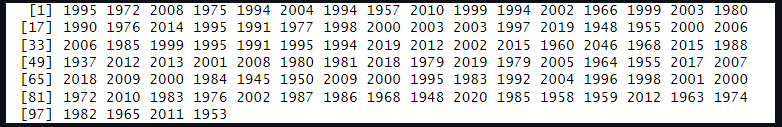
Bagunçado | limpo

:-------------------------------------------: |:-------------------------------:

![12 messy\_Imbdrelease\_year](https://user-images.githubusercontent.com/128150171/230798305-000429fd-5dba-469b-b0bb-20a5abbeeef5.png) |![13 consistency in year](https://user-images.githubusercontent.com/128150171/230798819-1c386d62-3c18-445c-9e41-a5528320dc4f.png)



![14 cleaned\_Imbdrelease\_year](<https://user-images.githubusercontent.com/128150171/230798340-3edb205e-14ed-4330-ad3b-59b1f178c1e6.png>)



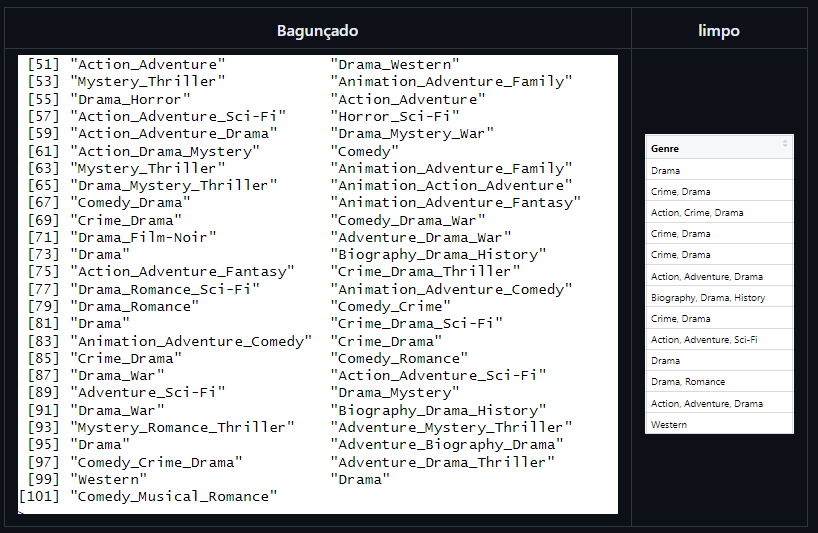
### GÊNERO

- Mudou os sublinhados para vírgulas

Bagunçado | limpo

:-------------------------------------------: |:-------------------------------:

![15 messy\_Imbdgenre](https://user-images.githubusercontent.com/128150171/230798874-cfc2db2d-4b69-4afe-bad8-f5269c180e6f.png) | ![16 cleaned\_Imbdgenre](https://user-images.githubusercontent.com/128150171/230798902-df394574-8643-4af7-b5db-5a3a5714c451.png)



### DURAÇÃO

- Teve 5 NAs, um NaN e um Inf

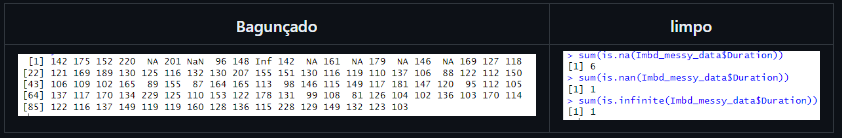
- Inseriu manualmente a duração correta do site \*\*IMBD\*\* on-line para preencher as observações nulas

- Convertido de caractere de classe para numérico de classe

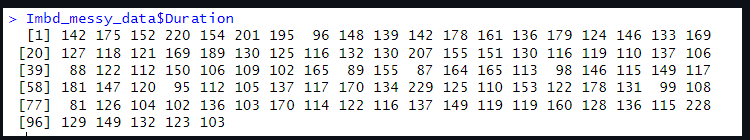
Bagunçado | limpo

:-------------------------------------------: |:-------------------------------:

![17 messy\_Imbdduration](https://user-images.githubusercontent.com/128150171/230798920-3f23e054-90fe-4657-821c-6f8d16bf6ddf.png) |![18 na, nana inf of Duration](<https://user-images.githubusercontent.com/128150171/230799411-52a4719f-d6bd-437e-9837-2b2b5e50b5ca.png>)



![19 cleaned\_imbdduration](https://user-images.githubusercontent.com/128150171/230798932-b14ef38f-051a-43e4-82da-51e671d00257.png)



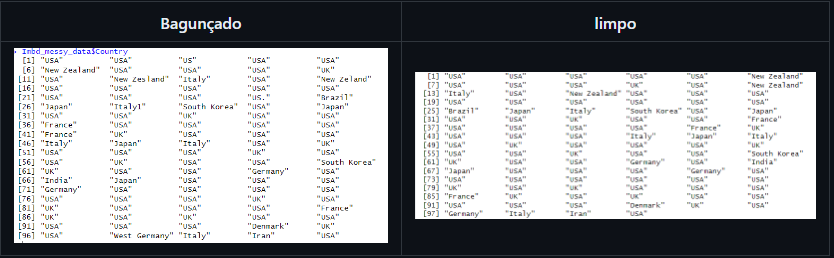
### PAÍS

- Corrigidos os nomes dos países com erros ortográficos

Bagunçado | limpo

:-------------------------------------------: |:-------------------------------:

![20 messy\_Imbdcountry](https://user-images.githubusercontent.com/128150171/230798567-b79a7a7e-6f69-45db-8b97-87be409602b0.png) | ![21 cleaned\_Imbdcountry](<https://user-images.githubusercontent.com/128150171/230798575-4458ad82-bc7a-4c7c-babd-e4f7e9ced7a1.png>)



### CLASSIFICAÇÃO DO CONTEÚDO

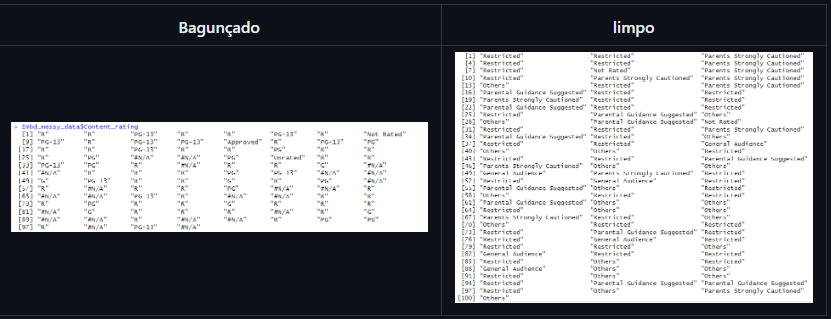
- Substituída a classificação de conteúdo abreviada para sua forma completa

- Convertido em um fator, pois é um dado categórico com um total de \*\*6 níveis\*\*

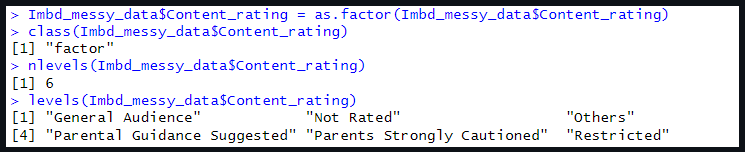
Bagunçado | limpo

:-------------------------------------------: |:-------------------------------:

![22 messy\_Imbdcontent\_rating](https://user-images.githubusercontent.com/128150171/230798590-bcb654b5-cae8-414c-99ad-2978345dc33d.png) | ![23 cleaned\_Imbdcontent\_rating](<https://user-images.githubusercontent.com/128150171/230798594-a3d1053d-18be-440a-a9f0-d454cfc1e46e.png>)



![24 factor level content rating](https://user-images.githubusercontent.com/128150171/230799492-21dd0b8d-8b28-43d7-b9e2-c0d568ede1d8.png)



### DIRETOR

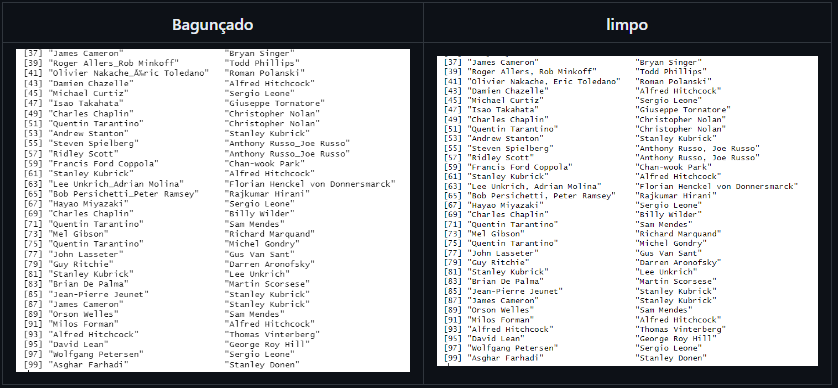
- Alfabeto acentuado alterado para um alfabeto adequado,

- Substituído sublinhado por vírgula onde temos mais de um diretor

Bagunçado | limpo

:-------------------------------------------: |:-------------------------------:

![25 messy\_Imbddirector](https://user-images.githubusercontent.com/128150171/230798632-c8f633ec-d245-4d10-ac14-3f8f2cd94800.png) | ![26 cleaned\_Imbddirector](<https://user-images.githubusercontent.com/128150171/230798637-5038ba33-a366-4d85-b197-91e9ebc2c882.png>)



### RENDA

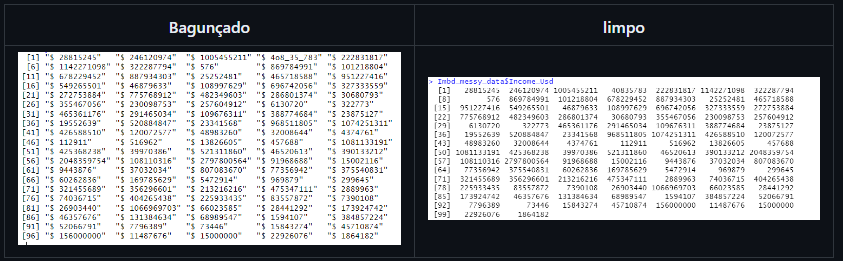
- Removido o cifrão para poder converter para uma classe numérica,

- Renomeou a coluna para \*\*income\_usd\*\* para indicar que está no formato de dólar

Bagunçado | limpo

:-------------------------------------------: |:-------------------------------:

![27 messy\_Imbdincome](https://user-images.githubusercontent.com/128150171/230798659-1b98d391-2d76-47fa-81c9-094d86abc75a.png) | ![28 cleaned\_Imbdincome](<https://user-images.githubusercontent.com/128150171/230798664-54e592f7-b405-4ede-9b02-a847b0de6591.png>)



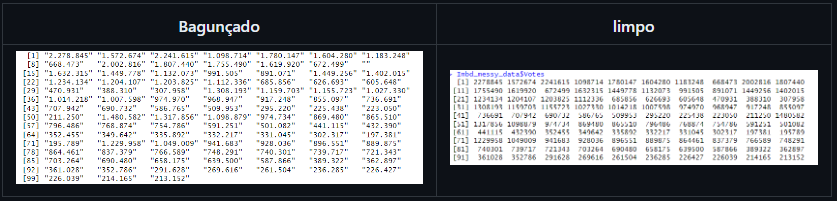
### VOTOS

- Removidos caracteres não numéricos (.) e convertidos em dados numéricos

Bagunçado | limpo

:-------------------------------------------: |:-------------------------------:

![29 messy\_Imbdvote](https://user-images.githubusercontent.com/128150171/230798696-b86d4822-851d-418a-903e-0cd6dd299a93.png) | ![30 cleaned\_Imbdvotes](<https://user-images.githubusercontent.com/128150171/230798719-725d656f-1d92-4ba9-8f17-88eefdd776fb.png>)



### PONTUAÇÃO

- Caracteres não alfabéticos removidos,

- Removido 0 inicial,

- Adicionado pontos decimais para aqueles sem,

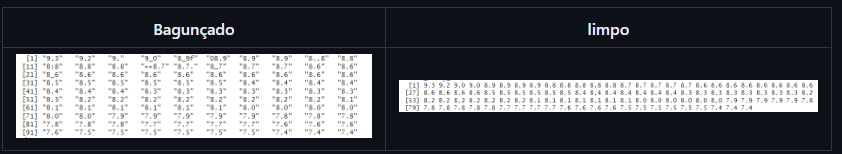
- Ponto decimal extra removido,,

- Convertido em dados numéricos

Bagunçado | limpo

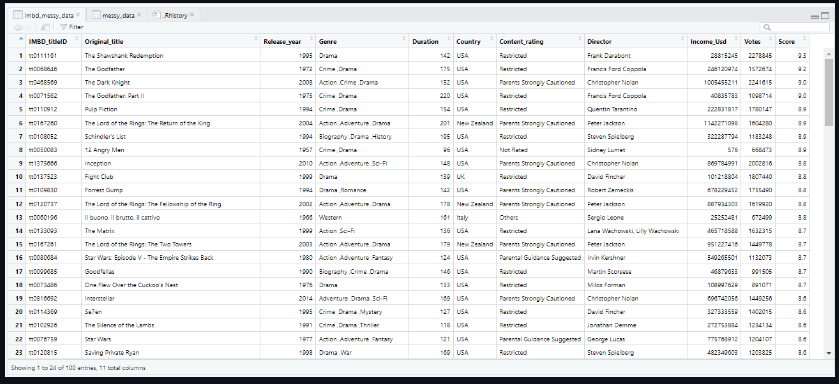
:-------------------------------------------: |:-------------------------------:

![31 messy\_Imbdscore](https://user-images.githubusercontent.com/128150171/230798979-1a280b92-624c-43f6-9fd5-0c73a71df285.png) | ![32 cleaned\_Imbdscore](<https://user-images.githubusercontent.com/128150171/230799016-8aa6bf32-f0b9-4a5f-a14b-074abe5fc685.png>)



### TABELA LIMPA

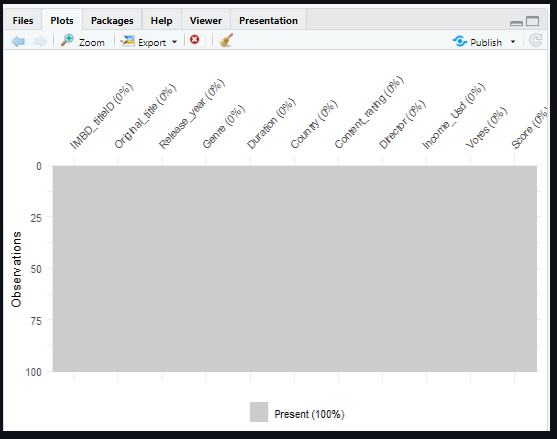
![33 cleaned\_table](<https://user-images.githubusercontent.com/128150171/230797913-88339da8-6dba-41cb-8be1-ca3d1855f138.png>)



Você pode ver os dados limpos [here](Imbd\_cleaned\_data.xlsx)

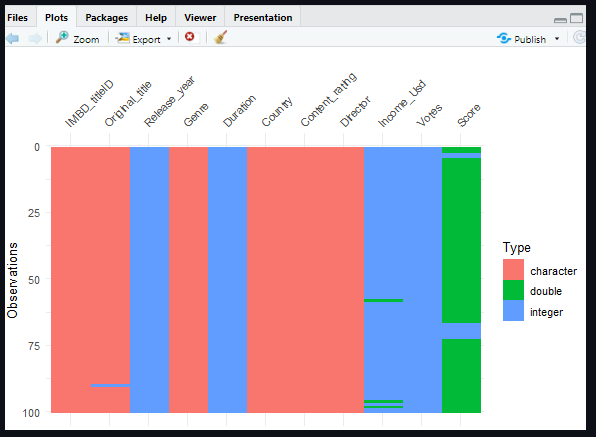
\*\*Agora estamos verificando para confirmar se não há nenhuma observação ausente\*\*

![34 cleaned vismiss](https://user-images.githubusercontent.com/128150171/230800242-dc27c161-f696-4c20-8390-e59ec65d2ac5.png)



\*\*Agora estamos usando o vis\_guess para verificar se eles estão nas classes corretas\*\*

![35 cleaned visguess](<https://user-images.githubusercontent.com/128150171/230800249-10b890e3-5f72-47f9-92a9-a172fced4854.png>)



## LIMITAÇÕES

O processo de limpeza do conjunto de dados tem algumas limitações que devem ser observadas. Inserir dados manualmente torna o processo de limpeza mais lento e pode introduzir erros ou vieses nos dados se não for feito com cuidado. Além disso, algumas colunas no conjunto de dados possuem valores ausentes ou dados incompletos, o que pode limitar a análise que pode ser realizada. Em termos da natureza do conjunto de dados, também há algumas limitações a serem consideradas. O conjunto de dados inclui apenas informações sobre os filmes com melhor classificação no \*\*IMBD\*\*, que é um subconjunto relativamente pequeno de todos os filmes já feitos, o que o torna limitado em seu escopo. Além disso, o conjunto de dados é baseado em avaliações de usuários, o que pode afetar a representatividade da amostra e a precisão das avaliações devido a vieses de auto-seleção, não resposta e manipulação de avaliações. Por fim, o conjunto de dados inclui apenas filmes até um determinado ponto no tempo e pode não refletir tendências ou mudanças mais recentes na indústria cinematográfica

## CONCLUSÃO E RECOMENDAÇÃO

Depois de limpar o conjunto de dados, alguns dos muitos insights que podem ser obtidos, por exemplo, são: \*A trilogia O Senhor dos Anéis tem três filmes no top 15, com O Retorno do Rei tendo a pontuação mais alta de 8,9. Os filmes mais votados são principalmente crimes e drama, seguidos de ação e aventura, e são principalmente dos Estados Unidos. As classificações de conteúdo dos filmes com melhor classificação variam, sendo R o mais comum, seguido por PG-13 e PG. Curiosamente, não há uma correlação clara entre a receita de um filme e sua classificação, pois alguns filmes de alta classificação tiveram uma receita baixa, enquanto alguns filmes de classificação mais baixa tiveram uma receita alta. Além disso, o número de votos para um filme está positivamente correlacionado com sua classificação, indicando que filmes populares tendem a ter classificações mais altas. No geral, os filmes com melhor classificação no conjunto de dados são diversos em termos de estilos, gêneros e origens, sugerindo que não existe uma fórmula única para fazer um filme com alta classificação.\* Por fim, o objetivo da limpeza de dados é garantir a precisão , consistência e integridade dos dados, o que pode ajudar a evitar erros e permitir análises e previsões mais precisas.

---

Você pode ver o processo passo a passo que segui para limpar os dados [here](IMBD\_Cleaning.R)

<https://github.com/tolamoye/IMBD-Data-Cleaning-with-R/tree/main>

# Projeto 4

# Análise do Twitter: o desafio da limpeza de dados

<img src="https://github.com/Ebuka456/Twitter-Analysis-The-Data-Cleaning-Challenge/blob/main/Data%20Cleaning%20Challenge/istockphoto-1473433691-612x612.jpg" alt="Alt text" style= "width: 900px; height: 450px"/>



## Introdução

O #datacleaningchallenge é um evento do twitter destinado a promover as melhores práticas de limpeza de dados. O desafio incentiva os participantes a compartilhar suas experiências, dicas e truques na limpeza de dados, trabalhando em dados sujos. Também serviu como um meio para os entusiastas terem uma ideia do que é a limpeza de dados enquanto recebiam orientação dos organizadores. O evento online foi lançado em 9 de março de 2023 por meio de um espaço no Twitter e durou todo o mês de março.

Este projeto é uma análise baseada em Python de tweets relacionados ao #datacleaningchallenge. A visualização de dados foi feita usando o Power BI. Escrevi um post no Medium explicando meus insights com mais detalhes. Confira

[HERE](https://medium.com/@okonkwoebuka456/the-data-cleaning-challenge-a-twitter-data-analysis-project-c25ae4a32dd3)

## Declaração do problema

Os organizadores do #datacleaningchallenge que aconteceu em março pretendem iniciar outro desafio em abril. Pretendo fornecer informações sobre os dados obtidos no desafio, como as pessoas percebem a limpeza de dados, as ferramentas mais comentadas que podem dar uma dica sobre as ferramentas usadas pelos participantes e as estratégias de como tornar o próximo desafio ainda maior.

## Fornecimento de dados

Reuni meus dados usando a biblioteca python Snscrape. Raspei meus dados do Twitter usando a hashtag #DataCleaningChallenge de 1º de março a 31 de março de 2023. A função que usei para raspar os dados é mostrada abaixo.

# to define a function that scrapes tweet from tweet

def scrape\_hashtag\_tweets(hashtag, start\_date, end\_date):

"""

Scrapes all tweets containing a certain hashtag within a specified time frame,

ignoring case sensitivity.

Args:

hashtag (str): the hashtag to scrape, without the "#" symbol

start\_date (str): the start date in "YYYY-MM-DD" format

end\_date (str): the end date in "YYYY-MM-DD" format

"""

# Convert the start and end dates to datetime objects

start\_dt = dt.datetime.strptime(start\_date, "%Y-%m-%d")

end\_dt = dt.datetime.strptime(end\_date, "%Y-%m-%d")

# Create a list to store the scraped tweets

tweets = []

# Iterate over all tweets containing the specified hashtag

for tweet in sntwitter.TwitterSearchScraper(f"#{hashtag} since:{start\_date} until:{end\_date}").get\_items():

# Ignore tweets that don't match the hashtag (ignoring case sensitivity)

if hashtag.lower() not in tweet.content.lower():

continue

# Add the relevant information about the tweet to the list

tweets.append({

"id": tweet.id, #

"content": tweet.content,

"timestamp": tweet.date,

"username": tweet.user.username,

"userdisplayname": tweet.user.displayname,

"userlocation": tweet.user.location,

"retweetCount": tweet.retweetCount,

"likeCount": tweet.likeCount,

"language": tweet.lang,

"source": tweet.source

})

return tweets

return go(f, seed, [])

}

No total, raspei 922 tweets e 11 colunas.

## Limpeza e pré-processamento de dados

A limpeza e pré-processamento de dados foi realizada usando python. Os poucos passos que dei para conseguir isso são

- Garantindo que não haja id de tweet duplicado

- Verificar valores nulos e manipulá-los, se existirem

- Garantir que os tipos de dados sejam consistentes

- Extraindo o conteúdo da coluna `source` que apareceu no formato de tag html

```python

# python function extract the content between the HTML tags using str.extract()

def extract\_html\_tags(df, column\_name):

content = df[column\_name].str.extract(r'>(.\*?)<')

return content

# apply the extract\_html\_tags function to the 'source' column of the tweets dataframe

tweets['platform'] = extract\_html\_tags(tweets, 'source')

```

- Presença de código de idioma de três letras na coluna `idioma`. O código do idioma não deve exceder duas letras, exceto de Indefinido (und)

- Também extraí a data (aaaa/mm/dd) da coluna `timestamp`.

## Análise e visualização de dados

O conjunto de dados limpo foi explorado usando python e visualizado usando o Power BI. Não há modelo de dados, exceto o relacionamento entre minha tabela Calendat e meu conjunto de dados limpo. O relatório que fiz é mostrado abaixo

<img src="https://github.com/Ebuka456/Twitter-Analysis-The-Data-Cleaning-Challenge/blob/main/Data%20Cleaning%20Challenge/Twitter%20Analysis%20Report\_page-0001.jpg" alt="Alt text" style= "width: 600px; height: 1000px"/>

Realizei uma análise de sentimento que diria como as pessoas percebiam o desafio da limpeza de dados. O sentimento positivo pode indicar que os usuários estão achando o desafio envolvente, informativo ou útil, enquanto o sentimento negativo pode sugerir que os usuários não estão gostando do desafio ou tendo dificuldades com ele.

![Sentiment](https://github.com/Ebuka456/Twitter-Analysis-The-Data-Cleaning-Challenge/blob/main/Data%20Cleaning%20Challenge/sentiment.png)

I then created a word cloud to visualize the qualitative data in order find out the most commonly used words. The word cloud was done using [Word Art](https://wordart.com/)

![Word Cloud](https://github.com/Ebuka456/Twitter-Analysis-The-Data-Cleaning-Challenge/blob/main/Data%20Cleaning%20Challenge/word%20cloud.png)

Para o desafio de limpeza de dados, os participantes puderam usar as ferramentas populares de análise de dados, como Power BI, Python, Excel, SQL, R e Tableau. O número de vezes que as ferramentas de limpeza de dados são mencionadas nos tweets pode fornecer informações valiosas sobre quais ferramentas são populares entre os usuários do Twitter, preferidas pelos usuários para executar tarefas de limpeza de dados e ferramentas de tendências na comunidade de limpeza de dados.

![tools](<https://github.com/Ebuka456/Twitter-Analysis-The-Data-Cleaning> Challenge/blob/main/Data%20Cleaning%20Challenge/tools.png)

### The insights I got from the Analysis and visualization are documented in my [GitHub Repository](https://github.com/Ebuka456/Twitter-Analysis-The-Data-Cleaning-Challenge/blob/main/Data%20Cleaning%20Challenge/%23DataCleaningChallenge%20Twitter%20Analysis.ipynb) and my [Medium Post](https://medium.com/@okonkwoebuka456/the-data-cleaning-challenge-a-twitter-data-analysis-project-c25ae4a32dd3)

## Recommendations

Before I make recommendations, I would like to congratulate the Organizers and the speakers for a job well done. On the 11th of March 2023, #DataCleaningChallenge was one of the trending topics in Nigeria.

![Trending](https://github.com/Ebuka456/Twitter-Analysis-The-Data-Cleaning-Challenge/blob/main/Data%20Cleaning%20Challenge/Trending.jpg)

Some recommendations I can make are

- For wider reach and more online presence, I recommend that the organizers encourages the participants to write their experience, the things they learnt and the struggles they faced when participating in subsequent challenges organized. This way it gives the Organizers the chance to help them and improve on the sentiments from the participants.

- Looking at the most talked about tool, we can see that the users have preference tending towards Excel, SQL and Python for data cleaning. For subsequent data cleaning challenge, I recommend teaching sessions be held on how to effectively use these three tools and take their skillset to the next level. This should be done before commencement of subsequent challenges so that the participants would not be stuck when the challenges begin.

- Concerning tools that are the least frequently talked about, it could be related to the fact that the users find them difficult to use. So to help individuals get familiar with more data analytics tool, The organizers could fix session to train all the participants on one tool at a time, maybe Power BI since it has similarities to Excel.

- I analyzed for the top mentioned Twitter handles and the most common hashtags. For future challenges and to foster collaboration, I would recommend the Organizers collaborate with anyone from the names below. This could help for wider reach and bring in more users to partake in the challenge.

![handles](https://github.com/Ebuka456/Twitter-Analysis-The-Data-Cleaning-Challenge/blob/main/Data%20Cleaning%20Challenge/Top%20handles.png)

- I lastly recommend that a broadcasting strategy is put in place to make this challenge known to more users. For example, a flier could be made as a means of making the challenge look more official. You can even plead with Data Influencers to spread the word and encourage those who participated in the previous challenge to spread the word as well.

Generally, I would say the challenge was a success and it did well in terms of social media outreach but there is room for improvement.

For subsequent challenges, I recommend a target to be set on the number of users tweeting about the challenge. A total of 502 twitter users were discovered for this challenge, 1000 could be the target for the next challenge.

![thanks](https://github.com/Ebuka456/Twitter-Analysis-The-Data-Cleaning-Challenge/blob/main/Data%20Cleaning%20Challenge/Thanks.gif)

Thank you for reading!

<https://github.com/Ebuka456/Twitter-Analysis-The-Data-Cleaning-Challenge/tree/main>

<https://github.com/kemiscut/TECA-Data-Analysis>

<https://github.com/Isoken00/Air-Data-Quality-In-Beijing>

<https://github.com/PascalBurume/United-Nation-Report>

<https://github.com/bellaTHEanalyst/powerbi-report-weather-quality-analysis->

<https://github.com/dimeji268/Park-n-Ride>

<https://github.com/Amarabright/Inventory-Management-Analysis>

<https://github.com/JusttOdosa/Pizzalytics>

<https://github.com/Olorlade/Road-Accident-Analysis>