dnc

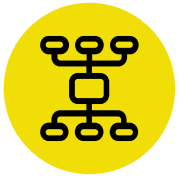
apresenta

Data Science & Machine Learning

data expert.
Data Science & Machine Learning

dnc

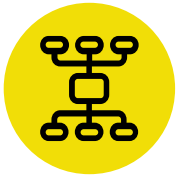
LAB
Data Cleaning



Materiais disponibilizados

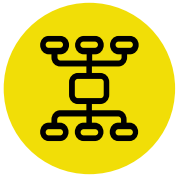
Vocês receberam o dataset pelo chat - 05-comunicados-day

- Fazer o download do dataset e colocar no colab;
- dataset.zip → dados disponibilizados pela Beauty Co;



Como vai funcionar

- Vocês terão **duas horas e meia** para finalizar o desafio;
- Vocês deverão entregar o link do colab enviado pelo co-host;
- Os facilitadores vão **auxiliar os grupos** quanto a dúvidas e dificuldades ao longo da atividade;



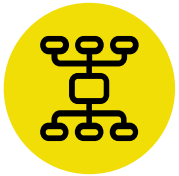
Lab

A Beauty Co, uma loja online de cosméticos, contratou seus serviços de verificar as marcas mais vendidas e quais eventos mais ocorrem em seu site.

O gerente de vendas te passou uma base com o histórico de interações no site referente ao período entre **nov de 2019**.

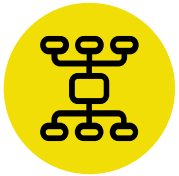
Output

- Tratar o dataset (Data cleaning/Wrangling)
- Eventos ordenados que mais ocorrem no site (Gráfico)
- Quais são as 5 marcas mais vendidas (Gráfico)



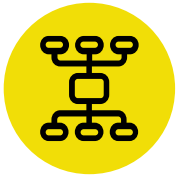
Etapa 1

- **Exploração dos Dados e Entendimento do problema:**
  - **Compreensão e Definição do Problema →** Qual o problema enfrentado pela empresa?
  - **Exploração dos Dados →** Qual o significado desses dados?



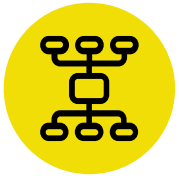
Etapa 2

- **Tratamento de duplicatas**
  - Importar bibliotecas
  - Importar dataset
  - Verificar primeiras linhas do dataset
  - Verificar estátisticas do dataset
  - Verificar tamanho (Linha x Colunas)
  - Agrupar valores (user_session, event_type, product_id, event_time por user_session para verificar se existem valores duplicados
  - Verificar tamanho do dataset retirando as duplicatas antes de executar a ação
  - Executar o tratamento de duplicatas
  - Agrupar valores (user_session, event_type, product_id, event_time por user_session para verificar se as duplicatas foram removidas



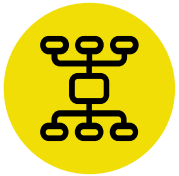
Etapa 3

- **Tratamento de datas:**
  - Transformar as datas de string para o formato correto
  - Verificar formato



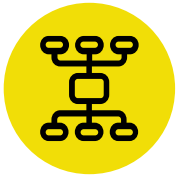
Etapa 4

- **Tratamento de nulos - Brand:**
  - Verificar quantidade de itens por marca
  - Verificar se existem nulos na coluna marcas
  - Retirar os nulos da coluna



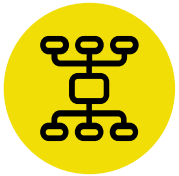
Etapa 5

- **Tratamento de nulos - price:**
  - Verificar se existem nulos na coluna price
  - Retirar os nulos da coluna
  - Calcular quartis 1%, 10%, 25%, 50%, 75%, 99%, max, min
  - Calcular IQR de 75% - 25%
  - Calcular lower_bound de 25%
  - Calcular upper_bound de 75%
  - Comparar valores com o descri percentiles de todos os valores calculados



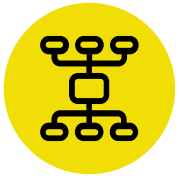
Etapa 6

- **Plotagem do gráfico de tipo de eventos por quantidade**
  - Conferir se existem valores nulos na coluna event_type
  - Definir x do gráfico → df['event_type'].value_counts().index
  - Definir y do gráfico → df['event_type'].value_counts()
  - Definir tamanho da figura
  - Definir título
  - Definir yticks
  - Definir legenda dos eixos
  - Plotar gráfico sns.barplot(x, y, saturation= aluno escolhe, order=['view', ‘cart', , 'remove_from_cart', 'purchase'], palette= aluno escolhe)



Etapa 7

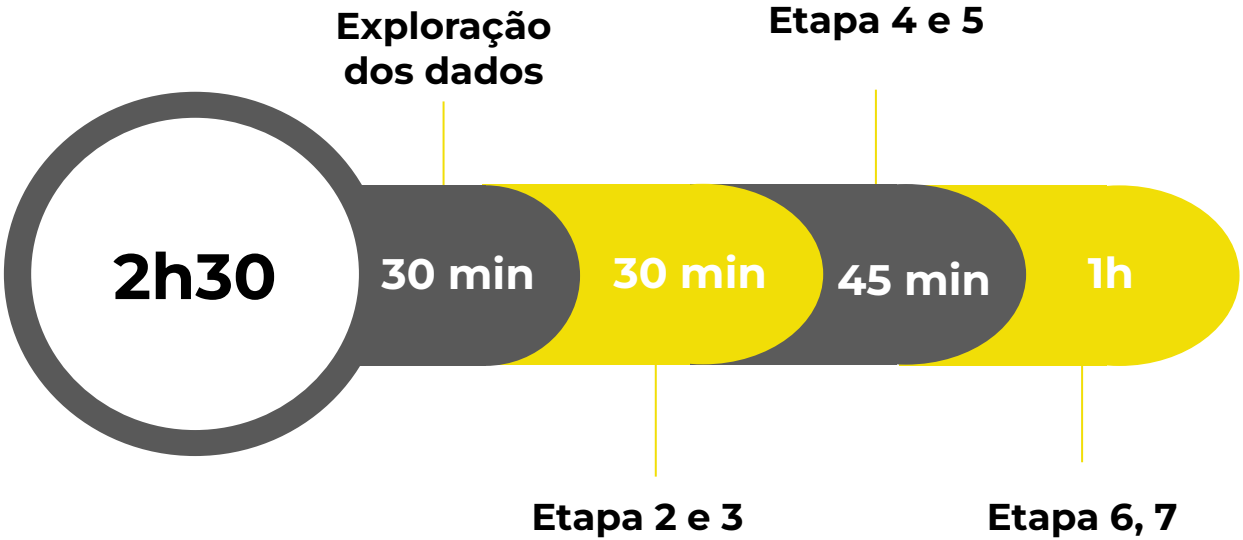
- **Plotagem das 5 marcas mais vendidas**
  - Verificar a quantidade de marcas mais vendidas
  - Redefinir colunas (Brand e Quantidade
  - Filtrar as 5 marcas mais vendidas
  - Transformar em um frame
  - Plotar o gráfico com as 5 marcas mais vendidas

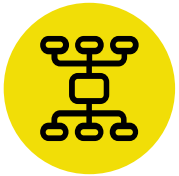


ENTREGA

- **Link de submissão**
- **Enviar o colab após a finalização da dinamica**

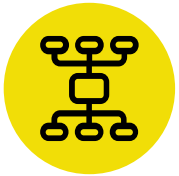
Timelog da Dinâmica





Dicionário de Variáveis

- **event_time -** Horário em que o evento aconteceu.
- **event_type -** Tipo do evento que ocorreu
- **product_id -** ID do Produto.
- **category_id -** ID da categoria* do Produto.
- **category_code -** Taxonomia da categoria* do produto (codinome). Presente para categorias significativas (normalmente), mas ignorado para alguns tipos de acessórios.



Dicionário de Variáveis

- **brand -** Caracteres referentes ao nome da marca.
- **price -** Preço flutuante de um produto.
- **user_id -** ID Permanente do Usuário.
- **** user_session** -** ID de sessão do usuário temporário. O mesmo para cada sessão de usuário. É alterado toda vez que o usuário volta à loja online após uma longa pausa.