



Bacharelado em Engenharia de Software

Disciplina: Ciência de Dados

Trabalho Final: Análise e Pré-Processamento de Dados com o Dataset Olist E-Commerce

1. Tema Geral do Projeto

“O que afeta a experiência e satisfação do cliente no e-commerce brasileiro?”

Investiguem os padrões que influenciam:

- Atrasos de entrega
- Baixa ou alta satisfação
- Diferenças de preço e frete
- Categorias de produtos problemáticas
- Variações no tempo de processamento e envio

*Obs.: NÃO se deve criar modelos de IA ou classificação. O foco é **pré-processamento e análise exploratória**.*

2. Base de dados

Base de Dados: Olist Brazilian E-Commerce Dataset (Kaggle)

[Clique aqui](#) para acessá-lo

Obs.: É um dos datasets brasileiros mais utilizados em Ciência de Dados :).

2.1 Datasets Obrigatórios

Cada dupla deve carregar e utilizar obrigatoriamente:

- **olist_orders_dataset.csv**
- **olist_order_items_dataset.csv**
- **olist_products_dataset.csv**

Esses três arquivos são suficientes para uma análise robusta.

3. Entregáveis Obrigatórios

3.1 Relatório Final em formato PDF

Deve conter:

1. **Contextualização do problema real** (com base nas percepções aprendidas com os datasets da atividade e pesquisas da equipe)
2. **Apresentação dos datasets adotados**
3. **Aplicação do Ciclo de Vida da Ciência de Dados**
4. **Exploração dos Dados (EDA)**
5. **Limpeza de Dados**
 - duplicatas
 - inconsistências
 - valores ausentes
 - outliers
6. **Conversão e padronização de tipos**
7. **Tratamento de dados categóricos e textos**
8. **Codificação de dados categóricos**
9. **Normalização e padronização (MinMax, Z-score)**
10. **Seleção de atributos**
 - correlação
 - baixa variância
 - filtros simples
11. **Criação de novos atributos (feature engineering) - pelo menos 4 técnicas**
12. **Pipeline completo de pré-processamento**
13. **Visualizações e gráficos explicativos**
14. **Insights finais**
15. **Conclusão do grupo**

Obs.: Para cada tópico do relatório relacionado ao tratamento dos dados deve-se apresentar em um breve texto com a justificativa da decisão tomada e um exemplo da situação.

3.2. Código-fonte completo em arquivo **.ipynb** ou **.py**

- Deve conter células organizadas, etapas comentadas e replicáveis.
- Histórico de merges, tratamentos e funções criadas.

3.3. Apresentação breve

- Apresentação das duplas será feita diretamente para a professora no dia **02/12/2025 entre às 07h e às 09h30**;
- Deverá ter duração máxima de 10 minutos;
- A dupla deverá apresentar a demonstração do pipeline, decisões de limpeza e principais descobertas.

4. Perguntas Norteadoras (obrigatórias no relatório)

Vocês devem responder a, pelo menos, **4 das 6 perguntas** abaixo:

1. Quais características mais se relacionam com **atrasos de entrega**?
2. Existem categorias de produtos com **maior frequência de problemas** (atrasos, preços absurdos, fretes altos)?
3. Os dados apresentam **outliers significativos**? Como foram tratados?
4. Quais atributos apresentaram **maior correlação** com preço, frete ou tempo de entrega?
5. A limpeza alterou o comportamento dos dados? Como e por quê?
6. O que pode influenciar a **satisfação do cliente**, com base nos atributos disponíveis?

5. Rubrica de Avaliação

Critério	Pontos
Exploração dos dados	2.0
Limpeza e tratamento (nulos, outliers, tipos)	2.0
Codificação, normalização e seleção de atributos	2.0
Feature engineering	1.0

Pipeline completo e código organizado	1.0
Clareza dos insights e conclusão	1.0
Organização e formatação dos relatório	1.0
TOTAL	10.0

6. Observações Importantes

- Trabalhos semelhantes ou copiados serão anulados.
- O foco é na **qualidade da limpeza, organização e justificativa de decisões**.
- Usem comentários explícitos no código.
- Preferência por boas práticas (funções, seções bem separadas).
- Podem usar Pandas, Numpy, Matplotlib, Seaborn e Scikit-Learn (somente pré-processamento).

7. Instruções de Entrega

7.1 Prazo Final

- **Data de entrega:** 01/12/2025
- **Horário limite:** 23h59 (horário de Brasília).
- **Valor:** 10 pontos.

7.2 Forma de Envio — Obrigatória

- A entrega **será realizada exclusivamente via GitHub**.
- **Criar um repositório público no GitHub**
- O nome do repositório deve seguir o padrão:

Projeto-Analise-Exploratoria-<NomeDuplaOuTrio>

- **Conteúdo mínimo do repositório**

O repositório deve conter obrigatoriamente:

- **Notebook** (**.ipynb**) ou **arquivo** **.py** com todos os passos realizados do projeto.



- **Arquivo README.md** contendo:
 - Nome dos integrantes
 - Link da base de dados utilizada
 - Objetivo do projeto
 - Descrição do processo de tratamento dos dados
 - Principais desafios encontrados
 - Principais conclusões
- **Arquivo CSV final**, já limpo e pré-processado pela equipe.

7.3 Enviar o link do repositório

O envio será feito pelo Formulário Oficial - [Clique aqui](#) para acessá-lo.

No campo destinado à entrega, enviar:

- **Relatório em PDF**
- **O link do repositório GitHub**
 - Entregas fora do GitHub ou com link incorreto não serão consideradas.

Conto com a dedicação e pontualidade de todos! Em caso de dúvidas, não hesitem em me procurar.

Bons estudos!

Profa Heloisa Guimarães :)