Datasets Disponíveis para o Projeto

📊 Visão Geral

Estão disponíveis **5 datasets diferentes** para o projeto. Cada grupo deve **escolher apenas 1 dataset** para trabalhar ao longo de todas as 5 etapas.

IMPORTANTE: Todos os datasets foram criados com **problemas intencionais** (valores faltantes, outliers, inconsistências, duplicatas, etc.) para simular dados do mundo real.

🎯 Como Escolher o Dataset

Critérios de Escolha

1. **Interesse do grupo** - Qual tema é mais interessante?
2. **Aplicabilidade** - Qual tem mais relevância prática?
3. **Complexidade** - Alguns têm mais variáveis categóricas/numéricas
4. **Originalidade** - Evite escolher o mesmo que outros grupos

Distribuição Sugerida

- Máximo 2 grupos por dataset (para variedade nas apresentações)
- Comunique ao professor qual dataset escolheu na primeira semana

---

📁 Dataset 1: Desempenho Acadêmico de Estudantes

📄 Arquivo

`students_performance.csv`

🎯 Objetivo

Prever a **nota final** de estudantes universitários (0-100) com base em hábitos de estudo, condições socioeconômicas e características pessoais.

📊 Características

- **Registros:** 2.510
- **Features:** 13
- **Variável Alvo:** `final_grade` (0-100 pontos)
- **Tipo:** Regressão

📝 Variáveis

Demográficas

- `student_id`: ID único

- `age`: Idade (18-25 anos)
- `gender`: Gênero (M/F)
- `parental_education`: Educação dos pais (high_school, bachelor, master, doctorate)

Acadêmicas

- `study_hours_week`: Horas de estudo por semana
- `attendance_rate`: Taxa de frequência (%)
- `previous_scores`: Notas anteriores (0-100)
- `tutoring`: Recebe tutoria (Yes/No)
- `extracurricular`: Atividades extracurriculares (Yes/No)

Infraestrutura e Bem-estar

- `internet_quality`: Qualidade da internet (Poor/Good/Excellent)
- `family_income`: Renda familiar (Low/Medium/High)
- `sleep_hours`: Horas de sono por dia
- `health_status`: Estado de saúde (Poor/Good/Excellent)

💡 Aplicação Prática

Identificar estudantes em risco de baixo desempenho para implementar programas de apoio preventivo.

📁 Dataset 2: Vendas Mensais de E-commerce

📄 Arquivo

`ecommerce_sales.csv`

🎯 Objetivo

Prever **vendas mensais** (em R$) de uma loja online com base em métricas de marketing, tráfego e comportamento do consumidor.

📊 Características

- **Registros:** 2.510
- **Features:** 16
- **Variável Alvo:** `monthly_sales` (vendas em R$)
- **Tipo:** Regressão

📝 Variáveis

Marketing e Tráfego

- `sale_id`: ID único
- `marketing_spend`: Investimento em marketing (R$)
- `website_traffic`: Visitantes mensais
- `conversion_rate`: Taxa de conversão (%)

- `mobile_traffic_pct`: Porcentagem de tráfego mobile

Produto e Preço

- `num_products`: Número de produtos no catálogo
- `avg_price`: Preço médio dos produtos (R$)
- `discount_percentage`: Desconto médio oferecido (%)
- `product_category`: Categoria (Electronics, Fashion, Home, Books, Sports)

Experiência do Cliente

- `avg_product_rating`: Avaliação média (0-5)
- `customer_reviews`: Número de avaliações
- `return_rate`: Taxa de devolução (%)
- `free_shipping`: Frete grátis (Yes/No)
- `payment_methods`: Métodos de pagamento aceitos

Mercado

- `competition_level`: Nível de competição (Low/Medium/High)
- `seasonality`: Sazonalidade (Low/Medium/High)

💡 Aplicação Prática

Otimizar investimentos em marketing e estratégias de vendas para maximizar receita.

📁 Dataset 3: Consumo de Energia Residencial

📄 Arquivo

`energy_consumption.csv`

🎯 Objetivo

Prever **consumo mensal de energia** (kWh) de residências com base em características da casa e hábitos dos moradores.

📊 Características

- **Registros:** 2.510
- **Features:** 16
- **Variável Alvo:** `monthly_consumption_kwh` (consumo em kWh)
- **Tipo:** Regressão

📝 Variáveis

Características da Residência

- `house_id`: ID único

- `house_area_sqm`: Área construída (m²)
- `num_rooms`: Número de cômodos
- `house_age_years`: Idade da casa (anos)
- `insulation_quality`: Qualidade do isolamento (Poor/Average/Good)
- `energy_efficiency_rating`: Classificação energética (A-E)

Moradores e Uso

- `num_residents`: Número de moradores
- `num_appliances`: Número de eletrodomésticos
- `home_office_hours`: Horas de home office por dia

Sistemas e Equipamentos

- `air_conditioning`: Ar condicionado (None/1/2/3+)
- `heating_system`: Sistema de aquecimento (None/Electric/Gas/Solar)
- `solar_panels`: Painéis solares (Yes/No)
- `electric_car`: Carro elétrico (Yes/No)
- `smart_thermostat`: Termostato inteligente (Yes/No)

Ambiente

- `has_pool`: Piscina (Yes/No)
- `avg_temperature`: Temperatura média externa (°C)

💡 Aplicação Prática

Prever consumo para precificação dinâmica, identificar consumidores de alto consumo, recomendar melhorias de eficiência.

📁 Dataset 4: Preços de Imóveis

📄 Arquivo

`housing_prices.csv`

🎯 Objetivo

Prever **preço de venda** de imóveis (em R$) com base em características físicas, localização e infraestrutura.

📊 Características

- **Registros:** 2.510
- **Features:** 17
- **Variável Alvo:** `price_brl` (preço em R$)
- **Tipo:** Regressão

📝 Variáveis

Características Físicas

- `property_id`: ID único
- `built_area_sqm`: Área construída (m²)
- `bedrooms`: Número de quartos
- `bathrooms`: Número de banheiros
- `parking_spaces`: Vagas de garagem
- `num_rooms`: Total de cômodos
- `property_age_years`: Idade do imóvel (anos)
- `floor_number`: Andar (0 = térreo/casa)

Localização e Vista

- `location`: Localização (Centro, Zona Sul/Norte/Leste/Oeste, Subúrbio)
- `view_type`: Tipo de vista (None/City/Park/Sea)
- `nearby_metro_km`: Distância até metrô (km)

Infraestrutura e Amenidades

- `infrastructure`: Infraestrutura do condomínio (Básica/Completa/Premium)
- `condo_fee`: Taxa de condomínio (R$/mês)
- `has_security`: Segurança 24h (Yes/No)
- `has_pool`: Piscina (Yes/No)
- `has_elevator`: Elevador (Yes/No)

Outros

- `property_type`: Tipo (Apartamento/Casa/Cobertura)
- `furnished`: Mobiliado (Yes/No)

💡 Aplicação Prática

Precificação automática de imóveis, identificar oportunidades de investimento, avaliar tendências de mercado.

📁 Dataset 5: Tempo de Entrega de Pedidos

📄 Arquivo

`delivery_time.csv`

🎯 Objetivo

Prever **tempo de entrega** (em horas) de pedidos com base em logística, condições de tráfego e características do pedido.

📊 Características

- **Registros:** 2.510

- **Features:** 16
- **Variável Alvo:** `delivery_time_hours` (tempo em horas)
- **Tipo:** Regressão

📝 Variáveis

Características do Pedido

- `delivery_id`: ID único
- `distance_km`: Distância até destino (km)
- `package_weight_kg`: Peso do pacote (kg)
- `delivery_type`: Tipo (Express/Standard/Economy)
- `is_priority`: Prioridade (Yes/No)
- `package_fragile`: Frágil (Yes/No)

Logística

- `vehicle_type`: Veículo (Moto/Carro/Van/Caminhão)
- `driver_experience_years`: Experiência do entregador (anos)
- `num_stops`: Número de paradas na rota
- `delivery_zone`: Zona de entrega (Urbana/Suburbana/Rural)
- `fuel_cost`: Custo do combustível (R$/litro)

Condições Externas

- `traffic_condition`: Tráfego (Baixo/Médio/Alto/Congestionado)
- `weather`: Clima (Ensolarado/Nublado/Chuva Leve/Forte/Tempestade)
- `time_of_day`: Horário (Madrugada/Manhã/Tarde/Noite)
- `day_of_week`: Dia (Seg-Qui/Sexta/Sábado/Domingo)

Feedback

- `customer_rating`: Avaliação do cliente (0-5)

💡 Aplicação Prática

Otimizar rotas de entrega, prever atrasos, melhorar experiência do cliente, precificação dinâmica.

⚠️ Problemas Comuns nos Datasets

Todos os datasets contêm os seguintes problemas **intencionais**:

Problema	Quantidade Aproximada	Exemplos
Valores faltantes (NaN)	~8% dos dados	Células vazias em features numéricas e categóricas
Outliers	~40-50 registros	Valores extremos mas não impossíveis

Problema	Quantidade Aproximada	Exemplos
Valores impossíveis	~10-20 registros	Idades negativas, ratings > 5, frequência > 100%
Inconsistências	~10-15 registros	Baixo input mas alto output
Duplicatas	10 registros	Registros muito similares com IDs diferentes
Erros de digitação	~5-10 registros	Valores claramente errados
Formatação	~70 registros	Espaços extras, MAIÚSCULAS, lowercase

Por que os dados têm problemas?

✅ **Realismo:** Dados do mundo real sempre têm problemas ✅ **Aprendizado:** Praticar limpeza e pré-processamento ✅ **Habilidades:** Desenvolver senso crítico sobre qualidade de dados

---

📊 Comparação dos Datasets

Aspecto	Estudantes	E-commerce	Energia	Imóveis	Entrega
Dificuldade	⭐⭐	⭐⭐⭐	⭐⭐⭐	⭐⭐⭐⭐	⭐⭐⭐
Variáveis Categóricas	6	5	8	8	9
Variáveis Numéricas	7	11	8	9	7
Correlações Óbvias	Forte	Média	Forte	Forte	Média
Feature Engineering	Médio	Alto	Médio	Alto	Alto
Interpretabilidade	Alta	Média	Alta	Alta	Média

🚀 Como Começar

1. Escolha seu Dataset

Discuta com o grupo e escolha 1 dos 5 datasets.

2. Carregue os Dados

```
import pandas as pd

# Exemplo: Dataset de Estudantes
df = pd.read_csv('datasets/students_performance.csv')

# Ou E-commerce
df = pd.read_csv('datasets/ecommerce_sales.csv')

# Ou Energia
df = pd.read_csv('datasets/energy_consumption.csv')
```

```
# Ou Imóveis
df = pd.read_csv('datasets/housing_prices.csv')

# Ou Entrega
df = pd.read_csv('datasets/delivery_time.csv')
```

3. Explore Inicialmente

```
# Visualizar primeiras linhas
print(df.head())

# Informações sobre o dataset
print(df.info())

# Estatísticas descritivas
print(df.describe())

# Verificar valores faltantes
print(df.isnull().sum())
```

4. Siga as Instruções

Consulte `etapas/etapa1/INSTRUCOES.md` para começar a Análise Exploratória.

📞 Dúvidas Frequentes

P: Posso mudar de dataset depois? R: Não recomendado, pois você perderá tempo. Escolha com cuidado!

P: Posso combinar múltiplos datasets? R: Não. Use apenas 1 dataset.

P: Os problemas nos dados são reais? R: Sim! Foram injetados propositalmente para simular dados do mundo real.

P: Devo corrigir os problemas antes da Etapa 1? R: NÃO! A Etapa 1 é para identificar. A Etapa 2 é para corrigir.

P: Qual dataset é mais fácil? R: "Estudantes" é o mais simples. "Imóveis" é o mais desafiador.

📚 Recursos Adicionais

- Pandas Cheat Sheet
- Data Cleaning Tutorial
- EDA Guide

Boa escolha e bom trabalho! 🎓

Última atualização: Janeiro 2025