

Projeto Integrador Final

Programação para Ciência de Dados

Arthur Casals

14 de Novembro de 2025

Introdução

Visão Geral do Problema

Seu Papel: Cientista de Dados Consultor

O Cenário:

- ▶ Você foi contratado por uma **consultoria em e-commerce**
- ▶ Cliente: varejista internacional operando há 2 anos
- ▶ Ele precisa de **insights açãoáveis** para crescer

Preocupações do CEO:

- ▶ "Dependemos muito de poucos produtos ou países?"
- ▶ "Onde investir em marketing?"
- ▶ "Como reter nossos melhores clientes?"
- ▶ "Que produtos merecem mais atenção?"

Nota Importante

Sua missão: Analisar dados transacionais e entregar recomendações estratégicas que o CEO possa implementar nos próximos 3-6 meses.

O Dataset: Online Retail

Fonte: UCI Machine Learning Repository / Kaggle

URL: <https://archive.ics.uci.edu/ml/machine-learning-databases/00352/Online%20Retail.xlsx>

Descrição:

- ▶ Transações de e-commerce UK
- ▶ Período: 01/12/2010 - 09/12/2011
- ▶ Varejo especializado em presentes
- ▶ **541.000 transações**
- ▶ 8 colunas de dados

Principais colunas:

- ▶ InvoiceNo (ID transação)
- ▶ StockCode (ID produto)
- ▶ Description (nome produto)
- ▶ Quantity, UnitPrice
- ▶ InvoiceDate
- ▶ CustomerID
- ▶ Country

Estrutura do Projeto

O que você vai entregar

Visão Geral do Projeto

Formato e Prazo:

- ▶ **Tempo:** 5-6 horas de trabalho
- ▶ **Entrega:** Notebook Jupyter (.ipynb)
- ▶ **Plataforma:** Google Colab
- ▶ **Peso:** 30% da nota final
- ▶ **Deadline:** 17/11/2025, 18:00 (São Paulo)

Pontuação Total:

- ▶ **100 pontos** (base)
- ▶ 70 pts: Aspectos técnicos
- ▶ 12 pts: Documentação e
- ▶ 18 pts: Soft skills

⚠ Atenção

IMPORTANTE: Este é um trabalho INDIVIDUAL. Discussões são permitidas, mas o código e análises devem ser seus.

As 8 Seções do Projeto

1. Estruturação do Problema e Contexto (5 pts)

- ▶ Introdução clara do problema de negócio
- ▶ Objetivos e perguntas a responder

2. Carregamento e Inspeção (5 pts)

- ▶ Carregamento correto, análise preliminar

3. Análise Exploratória - EDA (18 pts)

- ▶ Univariada (6 pts), Bivariada (7 pts), Multidimensional (5 pts)

4. Pré-processamento (20 pts)

- ▶ Limpeza (10 pts), Feature Engineering (10 pts)

As 8 Seções do Projeto (cont.)

5. Análise de Negócio (22 pts)

- ▶ **5.1 Análise Temporal OBRIGATÓRIA** (10 pts)
- ▶ **5.2 Outras 3 análises** - escolha 3 de 5 (12 pts)

6. Insights e Recomendações (18 pts)

- ▶ Priorização (6 pts), Comunicação (6 pts), Decisões (6 pts)

7. Qualidade Técnica do Código (8 pts)

- ▶ Organização, eficiência, reproduzibilidade

8. Visualizações (4 pts)

- ▶ Mínimo 12 visualizações de qualidade



Total: 100 pontos

Seção 3:

Análise Exploratória - EDA

EDA: O que você precisa fazer

3.1 Análise Univariada (6 pontos)

- ▶ Distribuição de variáveis numéricas (Quantity, UnitPrice, Revenue)
- ▶ Frequência de categóricas (Top produtos, Top países)
- ▶ Identificação de outliers
- ▶ **Mínimo 4 visualizações**

3.2 Análise Bivariada/Multivariada (7 pontos)

- ▶ Relações entre variáveis (correlações)
- ▶ Análise cruzada: País × Produto, Mês × Revenue
- ▶ Identificação de segmentos naturais
- ▶ **Mínimo 4 visualizações**

EDA: Análise Multidimensional

3.3 Análise Multidimensional (5 pontos)

- ▶ Criar **pelo menos 2 análises multidimensionais**:
 - ▶ Pivot tables (Produtos × Países × Métricas)
 - ▶ Agregações complexas (Top 10 produtos por país)
 - ▶ Heatmaps multidimensionais
 - ▶ FacetGrids ou visualizações condicionais
- ▶ Interpretação clara dos padrões encontrados

💡 Nota Importante

Use todo o arsenal do curso: `groupby()`, `pivot_table()`, `crosstab()`, heatmaps do Seaborn, etc.

Seção 4:

Pré-processamento

Limpeza de Dados (10 pontos)

Dados Faltantes:

- ▶ Análise de padrões de missing (percentual, tipo)
- ▶ Estratégia documentada: remoção vs imputação
- ▶ Justificativa clara das decisões

Outliers e Inconsistências:

- ▶ Identificação de valores negativos, zeros, extremos
- ▶ Tratamento de cancelamentos (InvoiceNo com 'C')
- ▶ Remoção/correção com justificativa

Validação:

- ▶ Verificação de tipos corretos
- ▶ Ranges válidos (Quantity > 0, UnitPrice ≥ 0)
- ▶ Remoção de duplicatas

Feature Engineering (10 pontos)

Features Obrigatórias:

</> Python

```
1 # 1. Revenue
2 df['Revenue'] = df['Quantity'] * df['UnitPrice']
3
4 # 2. Features temporais
5 df['Year'] = df['InvoiceDate'].dt.year
6 df['Month'] = df['InvoiceDate'].dt.month
7 df['DayOfWeek'] = df['InvoiceDate'].dt.dayofweek
8
9 # 3. Flag de cancelamento
10 df['Is_Cancellation'] = df['InvoiceNo'].str.startswith('C')
11
```

Feature Engineering (10 pontos)

Features Obrigatórias (devem estar presentes):

- ▶ Revenue = Quantity × UnitPrice
- ▶ Year, Month, DayOfWeek, Hour
- ▶ Is_Cancellation (flag booleana)

Features Criativas (mínimo 3):

- ▶ Exemplos: Order_Size, Customer_Value, Product_Category
- ▶ Justifique por que cada feature é relevante para o negócio

Seção 5:

Análise de Negócio

Estrutura da Análise de Negócio

5.1 Análise Temporal

- ▶ **OBRIGATÓRIA**
- ▶ **10 pontos**
- ▶ Mais detalhada e profunda
- ▶ Inclui técnicas avançadas

5.2 Outras Análises

- ▶ **Escolha 3 de 5**
- ▶ **12 pontos** (4 pts cada)
- ▶ Opções disponíveis no próximo slide



Total da Seção 5: 22 pontos ($10 + 12$)

Análise Temporal OBRIGATÓRIA (10 pontos)

Requisitos Mínimos:

- ▶ Tendência de receita ao longo dos meses
- ▶ Identificação de sazonalidade (mensal/semanal)
- ▶ Cálculo de crescimento MoM (Month-over-Month)

Análises Avançadas Obrigatórias:

- ▶ **Decomposição temporal:** tendência + sazonalidade + resíduo
- ▶ **Médias móveis:** 7 dias e 30 dias
- ▶ **Detecção de anomalias:** Z-score ou IQR
- ▶ **Visualizações temporais:** mínimo 4 gráficos

Recomendação Final:

- ▶ Quando intensificar campanhas de marketing?

Outras Análises (escolha 3 de 5)

Opções Disponíveis (4 pontos cada):

1. Análise de Produtos

- ▶ Top 20 produtos, concentração 80/20, margem
- ▶ Recomendação: produtos para destacar

2. Análise Geográfica

- ▶ Distribuição por país, crescimento por região
- ▶ Recomendação: onde investir em expansão

3. Análise de Clientes

- ▶ CLV, segmentação Top 10%, frequência
- ▶ Recomendação: estratégias de retenção

Outras Análises (cont.)

4. Análise de Comportamento de Compra

- ▶ Tamanho médio de pedido, frequência
- ▶ Recomendação: como aumentar ticket médio

5. Análise de Cancelamentos

- ▶ Taxa por produto/país/mês, impacto financeiro
- ▶ Recomendação: como reduzir cancelamentos

💡 Nota Importante

Escolha as 3 análises que fazem mais sentido para responder às preocupações do CEO mencionadas no início.

Seção 6:

Insights e Recomendações

Soft Skill #1: Priorização (6 pontos)

O que fazer:

- ▶ Identifique os **5 insights mais importantes**
- ▶ Para cada insight:
 - ▶ **O que você descobriu** (dados)
 - ▶ **Por que isso importa** (impacto no negócio)
 - ▶ **Nível de urgência** (alta/média/baixa)
- ▶ Ordene por prioridade considerando:
 - ▶ Impacto potencial na receita
 - ▶ Facilidade de implementação
 - ▶ Alinhamento com objetivos estratégicos

💡 Nota Importante

Pense como um **empreendedor**: o que gera mais valor com menos esforço?

Soft Skill #2: Comunicação de Valor (6 pontos)

Resumo Executivo para o CEO:

- ▶ **Máximo 300 palavras**
- ▶ Escrito para público **não-técnico**
- ▶ Deve incluir:
 1. Contexto resumido (1-2 frases)
 2. Top 3 descobertas mais importantes
 3. Implicações para o negócio
 4. Próximos passos recomendados
- ▶ Linguagem clara, objetiva, **sem jargões técnicos**

Atenção

Evite: "correlação de Pearson", "outliers IQR", "pivot table"

Use: "forte relação entre X e Y", "valores extremos", "análise cruzada"

Soft Skill #3: Tomada de Decisão (6 pontos)

Para cada um dos Top 3 insights, forneça:

1. **Recomendação específica** (ação concreta)
 - ▶ Ex: "Investir 20% do budget de marketing em..."
2. **Métrica de sucesso** (KPI para acompanhar)
 - ▶ Ex: "Aumentar receita internacional de 13% para 25% em 12 meses"
3. **Risco/limitação** da recomendação
 - ▶ Ex: "Custos de logística podem reduzir margem inicial"
4. **Timeline estimado** (curto/médio/longo prazo)
 - ▶ Ex: "Médio prazo (6-12 meses)"

Seções 7 e 8

Qualidade Técnica e Visualizações

Qualidade Técnica do Código (8 pontos)

Organização (2 pts):

- ▶ Estrutura lógica do notebook
- ▶ Código limpo e legível
- ▶ Imports no início
- ▶ Seções bem definidas

Reprodutibilidade (3 pts):

- ▶ Notebook executa sem erros **do início ao fim**
- ▶ Seeds definidos se necessário
- ▶ Resultados consistentes

Eficiência (3 pts):

- ▶ Operações vetorizadas
- ▶ Evitar loops desnecessários
- ▶ Código não redundante

Qualidade Técnica do Código (8 pontos)

Atenção

CRÍTICO: Teste seu notebook completo antes de entregar! Execute "Executar tudo" no Colab.

Visualizações (4 pontos)

Requisitos de Quantidade e Qualidade:

Quantidade:

- ▶ **Mínimo 12 visualizações**
- ▶ Distribuídas ao longo do notebook
- ▶ Variedade de tipos

Qualidade Técnica:

- ▶ Título descritivo
- ▶ Eixos rotulados
- ▶ Legendas quando necessário
- ▶ Cores apropriadas
- ▶ Tamanho adequado

Interpretação:

- ▶ Cada gráfico acompanhado de texto em markdown
- ▶ Insight claro derivado
- ▶ Conexão com objetivos

Tipos sugeridos:

- ▶ Histogramas, box plots
- ▶ Scatter plots, line plots
- ▶ Bar charts, heatmaps
- ▶ Pair plots, violin plots

Entrega e Avaliação

O que você precisa saber

Como Será Avaliado

Avaliação Automática (Otter-Grader):

- ▶ **65-70 pontos** verificados automaticamente
- ▶ Estrutura do notebook
- ▶ Shape dos datasets
- ▶ Existência de features
- ▶ Análises obrigatórias
- ▶ Execução sem erros
- ▶ Número de visualizações
- ▶ Técnicas avançadas (temporal)

Avaliação Manual (Professor):

- ▶ **30-35 pontos** avaliados manualmente
- ▶ Profundidade das análises
- ▶ Qualidade das interpretações
- ▶ Insights e recomendações
- ▶ Comunicação e documentação
- ▶ Soft skills
- ▶ Criatividade

Como Será Avaliado

Nota Importante

A avaliação automática garante que você atendeu os requisitos mínimos. A nota final depende da **qualidade** do seu trabalho.

Checklist Antes da Entrega

Checklist Técnico:

- Notebook executa sem erros
- Todas células executadas em ordem
- Mínimo 12 visualizações
- Features obrigatórias criadas
- Análise temporal COMPLETA
- Decomposição + médias móveis
- Código limpo e comentado

Checklist de Conteúdo:

- Seção de Introdução completa
- Resumo Executivo presente
- Top 5 insights priorizados
- Top 3 recomendações detalhadas
- Interpretações em todas visualizações

Atenção

Formato de entrega: SeuNome_ProjetoIntegrador.ipynb

Deadline: 17/11/2025, 18:00 (São Paulo), via Moodle

Recursos e Suporte

Recursos Permitidos:

- ▶ Documentação oficial
- ▶ Material do curso
- ▶ Stack Overflow
- ▶ Tutoriais online
- ▶ Discussões conceituais com colegas
- ▶ LLMs para dúvidas pontuais

NÃO Permitido:

- ▶ Copiar código de colegas
- ▶ Usar soluções prontas
- ▶ Plágio de análises publicadas
- ▶ LLMs para gerar notebook completo



Atenção

Plágio = Nota ZERO

Onde buscar ajuda:

- ▶ Fórum do Moodle (dúvidas gerais)
- ▶ Email (questões individuais)

Dicas para Nota Máxima

Como se destacar

5 Critérios de Excelência

1. Profundidade Analítica

- ▶ Vá além do óbvio
- ▶ Identifique padrões não triviais
- ▶ Conecte diferentes dimensões dos dados

2. Comunicação Clara

- ▶ Resumo executivo impactante e conciso
- ▶ Visualizações autoexplicativas
- ▶ Narrativa coerente

3. Insights Acionáveis

- ▶ Recomendações específicas e implementáveis
- ▶ Justificativas baseadas em dados sólidos
- ▶ Consideração de trade-offs e riscos

5 Critérios de Excelência (cont.)

4. Excelência Técnica

- ▶ Código limpo, eficiente e bem documentado
- ▶ Uso adequado de todas as ferramentas do curso
- ▶ Reprodutibilidade garantida

5. Visão de Negócio

- ▶ Priorização alinhada com impacto
- ▶ Linguagem adaptada ao público-alvo
- ▶ Foco em resultados mensuráveis (KPIs)

Nota Importante

Lembre-se: Você não é apenas um programador, você é um profissional capaz de transformar dados em decisões estratégicas!

Gestão de Tempo Sugerida (6 horas)

Fase 1: Exploração (1.5h)

- ▶ Carregamento
- ▶ Inspeção inicial
- ▶ EDA univariada e bivariada

Fase 2: Preparação (1.5h)

- ▶ Limpeza de dados
- ▶ Tratamento de missing/outliers
- ▶ Feature engineering

Fase 3: Análise (1.5h)

- ▶ 4 análises de negócio
- ▶ Análise multidimensional
- ▶ Visualizações

Fase 4: Síntese (1.5h)

- ▶ Insights prioritizados
- ▶ Resumo executivo
- ▶ Recomendações
- ▶ Documentação final

Gestão de Tempo Sugerida (6 horas)

Atenção

Dica: Não deixe para a última hora! Comece cedo e vá refinando.