UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ
Câmpus Campo Mourão

Aluno(a)			**RA**:
Curso		**Ano:**	
Disciplina	Ciência de Dados		
Professor	Prof. Eduardo Pena		
Nota			
Armazenamento/Processamento de Dados			

Orientações gerais:

1 - Todos os materiais (códigos fontes, documentos, diagramas, etc) deverão ser entregues em um arquivo .zip identificado (Nome/RA/Ano).

2 - A interpretação das questões é parte do processo de avaliação.

Dataset: NYC Yellow Taxi Trip Records (2023-2024)

Entrega: Notebook com dashboard (viz) de resultados + relatório (2-3 páginas)

Individual ou duplas

Contexto

Somos analistas de dados de uma empresa de mobilidade urbana. Precisamos analisar grandes volumes de dados de táxis de NYC para otimizar infraestrutura de dados e orientar decisões sobre ferramentas de processamento.

Tarefa 1: Setup e Consolidação dos Dados

Precisamos consolidar dados históricos. Desenvolva no notebook:

1. **Download automatizado:**
   - Códigos para baixar arquivos mensais de 2023 e 2024 (meta: 24 meses, mínimo: 12 meses)
   - Encontrar os dados e o padrão da URL em: `https://www.nyc.gov/site/tlc/about/tlc-trip-record-data.page`
   - Focar apenas nos **Yellow Taxi Trip Records**
   - Verificar integridade dos arquivos e tratar erros
2. **Análise de schema:**
   - Examine estrutura dos arquivos Parquet
   - Identifique e cuide de mudanças de schema entre diferentes meses
3. **Limpeza de dados:**

- Remover registros com duração de viagem negativa ou zero
- Filtrar coordenadas inválidas (fora dos ranges de latitude/longitude)
- Eliminar valores negativos em campos monetários
- Remover registros com `passenger_count` = 0
- Validar datas dentro do período esperado

4. **Consolidação:**
    - Concatene arquivos em dataset único
    - Exporte em: CSV, Parquet Snappy, Parquet ZSTD
    - Registre os tamanhos de arquivo de cada formato

Responda:

- Houve mudanças de schema entre os meses?
- Qual estratégia para dados ausentes/inconsistentes?
- Como garantir reprodutibilidade do processo? Ou seja, outros analistas chegariam no mesmo resultado que você?

Tarefa 2: Benchmark de Performance

Precisamos escolher entre Pandas e DuckDB para processamento.

Hipótese: "DuckDB com Parquet é mais rápido que Pandas com CSV"

1. **Medidas de armazenamento:**
    - Compare tamanhos: CSV vs Parquet Snappy vs Parquet ZSTD
    - Calcule taxas de compressão relativas
    - Gráfico de barras com diferenças percentuais
2. **Benchmark de carregamento:**
    - Meça tempo de leitura:
        - CSV → Pandas
        - CSV → DuckDB
        - Parquet → Pandas
        - Parquet → DuckDB
    - Use `time.time()` com pelo menos 5 execuções, tire a média e desvio padrão

3. **Benchmark de consultas (mínimo 5 consultas):**

   Escolha pelo menos 5 das seguintes sugestões de consultas, garantindo o uso de agregação, subconsultas, CTEs, window functions e pivot:

   - **Receita por zona:** TOP 10 zonas de pickup por receita total média
   - **Padrões temporais:** Agregação por mês/semana (COUNT viagens, SUM receita, AVG distância)
   - **Análise de gorjeta:** Taxa de gorjeta média por borough usando CASE WHEN para categorizar
   - **Horário de pico:** Ranking de horários mais movimentados usando window functions (RANK/ROW_NUMBER)
   - **Rotas populares:** TOP rotas origem-destino com CTE para calcular distâncias
   - **Análise de velocidade:** Velocidade média por zona usando subconsulta para filtrar outliers
   - **Sazonalidade:** Comparar receita mensal vs média anual usando window functions
   - **Análise de passageiros:** Distribuição de passenger_count com PIVOT
   - **Duração de viagens:** Mediana de duração por borough
   - **Eficiência por motorista:** Receita/hora por vendor usando CTEs aninhadas
   - **Análise de pagamento:** Distribuição de tipos de pagamento por zona
   - **Viagens longas vs curtas:** Comparativo usando CASE para categorizar distâncias
   - **Padrões de fim de semana:** Diferenças entre weekday/weekend usando window functions
   - **Crescimento temporal:** Taxa de crescimento mensal usando LAG/LEAD
   - **Zonas mais rentáveis:** Receita por quilômetro por zona com subconsultas complexas

4. Para cada consulta escolhida, você deve:

   - **Implementar em SQL** (DuckDB) e **validar os resultados** inspecionando os dados
   - **Implementar equivalente em Pandas** e verificar se os resultados são consistentes
   - Caso o Pandas não consiga processar o dataset completo, **usar uma amostra** e documentar esta limitação
   - **Anotar quais consultas falharam** no Pandas (memória, tempo, erro)
   - **Medir tempo de execução** rodando cada consulta 5 vezes em cada ferramenta
   - **Registrar apenas consultas que executaram com sucesso** para comparação de performance

5. **Visualizações:**

   - Gráficos (apropriados, a seu critério) comparando desempenho do Pandas vs DuckDB para cada query

Responda:

- Qual combinação de ferramenta e formato foi a vencedora? Justifique sua escolha e desenvolva a resposta.

Especificações do Relatório PDF

O relatório deve ter **2-3 páginas** e incluir:

Conteúdo Obrigatório

- Todas as visualizações geradas (gráficos de barras, comparações de performance)
- Respostas fundamentadas às questões propostas nas tarefas
- **Justificativa detalhada das escolhas visuais** (tipos de gráficos, paletas de cores, escalas, layout)

Formatação Recomendada

- **Layout profissional** usando (preferencialmente) LaTeX com tipografia limpa
- **Imagens de alta qualidade** bem posicionadas com legendas descritivas
- **Tabelas bem formatadas** para dados numéricos (usar booktabs)
- **Paleta de cores consistente** e profissional nas visualizações
- **Seções numeradas** com hierarquia clara e espaçamento adequado
- **Cabeçalho/rodapé** com identificação do autor

Entrega: Notebook + Relatório PDF com nome(s) do(s) autor(es)