Projeto Final: Análise Exploratória de Dados de Vendas Online

Objetivo:

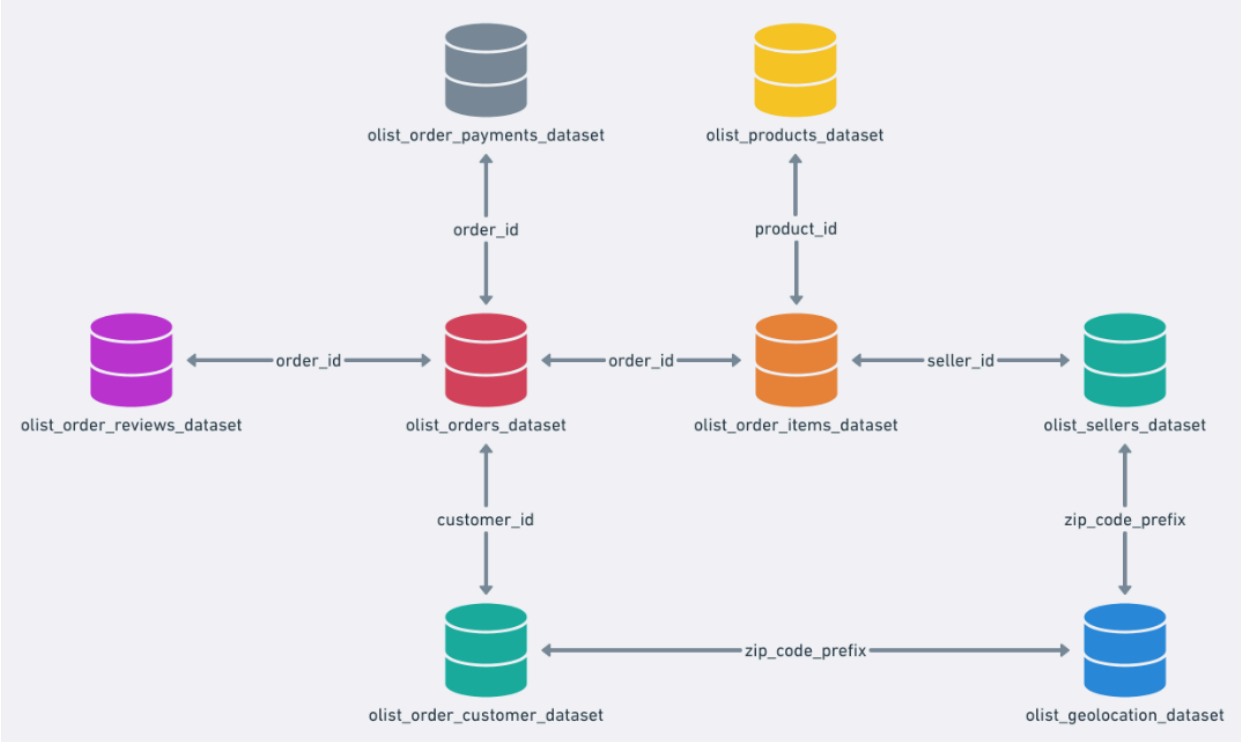
O objetivo deste projeto é proporcionar a você a oportunidade de aplicar os conhecimentos adquiridos ao longo da disciplina de Técnicas de Programação 1 em um contexto prático, relevante e *data-driven*. Sinta-se convidado a realizar uma análise exploratória de dados de vendas online, utilizando um conjunto de dados real, a fim de extrair insights e entender melhor o fenômeno das vendas e de tudo aquilo que lhe diz respeito (clientes, fornecedores, produtos, pagamentos, etc.).

Como você será avaliado:

Esta não é uma disciplina de Data Science, Estatística ou Análise Exploratória de Dados. Eu pretendo avaliar você quanto a 3 grandes conjuntos de conhecimento: Numpy, Pandas e Git. Assim, não se preocupe demasiadamente com a profundidade ou o tamanho da sua análise. Não avaliarei o seu rigor estatístico e também não avaliarei se você cobriu todas as análises relevantes possíveis de se fazer para este problema. Quero apenas avaliar o quanto você consegue aplicar do que vimos a respeito de Numpy, Pandas e Git.

Base de dados de trabalho:

Você deve trabalhar com a base de dados disponível gratuitamente aqui. Note que existem diferentes tabelas em CSV. Eu recomendo que você trabalhe com mais de uma tabela, talvez duas ou três. Recomendo que você faça o download das tabelas que considerar úteis e faça a leitura delas com o método csv_read, do objeto DataFrame da biblioteca Pandas. Para ter uma ideia mais geral do dataset, observe como as múltiplas tabelas estão organizadas. Não se preocupe se você não entender completamente esta figura agora. Isto não é estritamente necessário neste momento. Mas tente se lembrar de voltar aqui depois de ter completado o módulo de Banco de Dados I, porque a essa altura você entenderá completamente a figura 😉.



Tarefas que você não pode deixar de cumprir no projeto:

- Leitura dos datasets como objetos Pandas DataFrames
  Use a função read_csv do Pandas para ler dois ou mais arquivos csv que fazem parte da base de dados. A partir daí, trabalhe com estes objetos do tipo DataFrame.

- Análise exploratória inicial das tabelas
  Para cada tabela, produza uma análise exploratória inicial. Uma boa ideia é checar coisas como a quantidade de registros, os nomes e tipos de dados das

colunas de cada tabela, a quantidade de dados faltantes (nulos ou NaN), a proporção de cada valor para colunas categóricas, medidas estatísticas como mediana, média, mínimo, máximo e desvio padrão para cada coluna numérica.

- Levantar e responder pelo menos três questões a respeito dos dados de modo que as respostas demandem um filtro por máscara de seleção booleana. Alguns exemplos possíveis são:
    - Na tabela de pagamentos, há pagamentos do tipo “boleto” que tem mais de uma parcela (coluna payment_installments)?
    - Quais são exatamente os pagamentos que tem um valor maior ou menor do que o valor médio dos pagamentos registrados na tabela de pagamentos?
    - Na tabela de clientes, quem são os clientes que provém de uma das 3 cidades mais comuns desta tabela?

- Pelo menos duas vezes, tente criar um ndarray e adicioná-lo a alguma tabela como uma nova coluna. Para isso, tente encontrar alguma informação que você julgue parecer útil e que possa ser derivada das colunas já existentes nesta tabela. Por exemplo:
    - Na tabela de reviews, você pode ter uma coluna de booleanos que indique se a avaliação atingiu o valor mais alto (5) ou não.
    - Na tabela de produtos você pode criar uma coluna categórica com valores como “pequeno”, “médio”, e “grande” para cada produto, a depender de seus valores de height e width.

- Demonstre conhecimentos em GIT. Você pode fazer isso entregando o seu trabalho como um repositório GIT ou incluindo no seu projeto alguns comentários descrevendo quais comandos seriam necessários para levar a sua análise para um repositório GIT. Mas detalhes estão listados na seção *Entrega do Trabalho*, neste documento.

O que mais você pode fazer:

Sinta-se à vontade para incluir o que mais você quiser nas suas análises. Os itens descritos acima são o mínimo necessário para que eu possa avaliar você (e os exemplos são apenas ideias/possibilidades). Se você achar que conseguirá demonstrar melhor o seu conhecimento em Numpy/Pandas através de outras tarefas, vá em frente.

O que você não pode fazer:

Por favor, não inclua no seu trabalho respostas geradas por ferramentas de IA generativa. Acredite, se você o fizer, eu perceberei e precisarei levar isto em conta na sua avaliação. Gostaria de solicitar a sua colaboração neste sentido, por favor, para que esta fase de trabalho final seja o máximo tranquila quanto possível - tanto para mim quanto para você.

Entrega:

- No LMS, em Projetos, você encontrará disponível o projeto Entregas Finais. Sua entrega deve ser feita exclusivamente por ali.
- Você pode enviar um PDF, um arquivo .py, um notebook .ipynb ou um link do GitHub.
    - Se você escolher qualquer forma de envio que não seja um link do GitHub, inclua no seu trabalho uma seção em que você descreverá os principais comandos Git necessários para criação de um repositório e para o seu uso no dia-a-dia (coisas como enviar ou puxar novas versões de arquivos, vincular o repositório local a um repositório remoto, criar e utilizar branches, etc.).
- Você tem até 26/02/2024 23:59h para realizar o envio do seu trabalho. A plataforma não aceitará quaisquer entregas realizadas a partir deste prazo. Fique atento.