Escola Korú
Projeto de Aprendizagem: Análise e Engenharia de Dados no E-commerce Brasileiro
Curso: Engenharia de Dados
Módulo: Big Data

1. Situação problema

Utilizando o 'Brazilian E-Commerce Public Dataset by Olist', os grupos irão simular o papel de engenheiros de dados em uma empresa de e-commerce, com o objetivo de extrair insights valiosos para aprimorar as operações de negócios, otimizar a logística e melhorar a experiência do cliente.

Problemas

Como os dados podem ser usados para melhorar a recomendação de produtos e personalizar a experiência do cliente?

Quais insights podem ser obtidos para otimizar a gestão de inventário e operações logísticas?

De que maneira a visualização de dados pode auxiliar na decisão estratégica e operacional?

Logo, ao responderem às seguintes perguntas, vocês devem gerar insights valiosos para o negócio e para isso vocês devem utilizar técnicas de manipulação/visualização de dados com python e/ou pyspark.

2. Objetivos de aprendizagem do projeto:

- Compreender e praticar os fundamentos do Python para análise de dados.
- Explorar a integração de Python com bancos de dados e APIs.
- Aplicar técnicas de análise de dados usando Pandas, visualização com Streamlit e análise de dados em larga escala com PySpark.

3. Atividades e cronograma:

Atividade	Divulgação da Tarefa (tutor)	Apresentação da Tarefa (alunos)
Instalação do Python, VSCode	13/11	-

e bibliotecas necessárias para desenvolvimento do projeto.		
Atividade I Integração de Python com APIs e Bancos de Dados	20/11/2023	27/11/2023
Atividade I Análise de Dados com Pandas	27/11/2023	04/12/2023
Atividade II Visualização de Dados com Streamlit	04/12/2023	11/12/2023
Atividade II Processamento de Dados em Larga Escala com PySpark	11/12/2023	18/12/2023
Preparação para Pitch	11/12/2023	18/12/2023

➢ Atividade I: Integração de Python com APIs e Bancos de Dados

Descrição da Atividade:

A primeira tarefa é a Integração de Python com APIs e Bancos de Dados, ou seja, a criação das soluções para resolução dos problemas. E, para isso, vocês terão que integrar dados da base OLIST via API e armazenar os resultados em um banco de dados, para então posteriormente todas as análises serem desenvolvidas.

Lembrando que, nesta primeira atividade devemos também executar a limpeza, preparação e análise de nossos dados onde tal análise possui o objetivo de nos trazer algumas respostas das perguntas apresentadas anteriormente.

Vale ressaltar que, nesta primeira etapa, o principal objetivo é termos uma integração bem estruturada e seguindo boas práticas nos códigos, porém não se deve esquecer que um dos objetivos do projeto é a entrega de um repositório no GitHub estruturado e bem documentado. Por tanto, durante o desenvolvimento da primeira atividade é importante que a documentação e o Github andem em paralelo a isso.

Recursos Necessários:

- **Os datasets (bases de dados):** referência de dados disponibilizados pela Base dos Dados: Link
- Python, pandas, seaborn e matplotlib: documentação das bibliotecas disponíveis nestes links: Pandas, **Seaborn**, **Matplotlib**

Objetivos específicos de aprendizagem:

➽ Entender e aplicar técnica de extração de dados via integração com API.

➽ Analisar distribuições, correlações e possíveis anomalias presentes nos dados.

➽ Fazer uma análise exploratória destacando características da base de dados, explicação das variáveis e mapeamento de objetivos para análises futuras.

Rubrica de avaliação:

➽ Objetivo: Identificar variáveis e características relevantes da base de dados

Insuficiente	Não identificou nenhuma variável ou característica relevante da base de dados.
Regular	Identificou variáveis ou características, mas com pouca relevância para a análise proposta.
Bom	Identificou tecnologias ou soluções com relevância para o problema escolhido.
Ótimo	Identificou variáveis e características com relevância e apresentou como elas se relacionam com os objetivos da análise exploratória.

➽ Objetivo: Fazer uma análise exploratória destacando características da base de dados, explicação das variáveis e mapeamento de objetivos para análises futuras.

Insuficiente	Não realizou uma análise exploratória adequada e não destacou características relevantes da base de dados.
Regular	Realizou uma análise exploratória, mas com pouca profundidade nas características da base de dados ou na explicação das variáveis.
Bom	Realizou uma análise exploratória destacando características relevantes da base de dados e forneceu uma explicação adequada das variáveis.
Ótimo	Realizou uma análise exploratória abrangente, destacando características da base de dados, explicando as variáveis de forma clara e mapeando objetivos de forma estratégica para análises futuras.

➽ Objetivo: Integração da Base de Dados via API

Insuficiente	Não integrou os dados via API.
Regular	Integrou os dados via API mas não aplicou boas práticas no desenvolvimento do código.
Bom	Integrou os dados via API e aplicou boas práticas no desenvolvimento do código.
Ótimo	Integrou os dados via API, aplicou boas práticas no desenvolvimento do código e aplicou o conceito de funções para isso.

Habilidades a serem aprendidas/desenvolvidas:

Ao final da atividade, os participantes terão adquirido as habilidades de:

- **Análise Exploratória**: Compreender e selecionar as variáveis e características relevantes na base de dados.
- **Análise Estatística**: Utilizar Python para analisar distribuições, correlações e detectar anomalias nos dados.
- **Programação em Python**: Desenvolver códigos afim de integrar dados via API.
- **Análise Exploratória de Dados (EAD)**: Realizar uma análise exploratória completa, utilizando Pandas para destacar características essenciais, explicar variáveis e mapear objetivos para análises futuras.

➢ Atividade II: Análise de Dados com Pandas

Descrição da Atividade:

A segunda atividade diz respeito de desenvolver visualizações utilizando python, afim de enriquecer ainda mais a EDA desenvolvida na primeira atividade. Vale lembrar também que, nesta atividade iremos utilizar PySpark para replicarmos algumas de nossas análises desenvolvida com Pandas na primeira atividade, para então assim, entedermos suas semelhanças.

Ressaltando que, no final dessa tarefas os grupos deverão consolidar em uma apresentação (PowerPoint ou similar), as principais descobertas e recomendações baseadas na análise. Criar e apresentar também as referências de códigos utilizados.

Recursos Necessários:

- **Os datasets (bases de dados):** referência de dados disponibilizados pela Base dos Dados: Link
- **Python, pandas, seaborn e matplotlib:** documentação das bibliotecas disponíveis nestes links: Pandas**, Seaborn, Matplotlib**
- **PySpark: PySpark**

Objetivos específicos de aprendizagem

➽ Compreender as técnicas e ferramentas para visualização de dados.
➽ Manipulação de dados utilizando pyspark.
➽ Criação, estruturação e documentação de um repositório no Github.

Rubrica de avaliação:

➳ Qualidade técnica e organização do código.

Insuficiente	Códigos desorganizados e sem aplicações nenhuma de boas práticas.
Regular	Códigos organizados e sem aplicações nenhuma de boas práticas.
Bom	Códigos organizados e com aplicações de boas práticas.
Ótimo	Códigos organizados e com aplicações de boas práticas e repositório no github bem estruturado.

➳ Visualizações estruturadas

Insuficiente	Não apresentou visualizações claras.
Regular	Apresentou visualizações claras mas sem utilizar boas práticas (cores, gráficos, entre outros)
Bom	Apresentou visualizações claras utilizando boas práticas (cores, gráficos, entre outros)
Ótimo	Apresentou visualizações claras utilizando boas práticas (cores, gráficos, entre outros) e apresentou insights ricos sobre as visualizações

Habilidades a serem aprendidas/desenvolvidas:

Ao final da atividade, os participantes terão adquirido as habilidades de:

- **Planejamento e Estruturação de um Repositório**: Organização e estruturação de um repositório no Github.
- **Técnicas e Ferramentas visualização de dados**: Compreender e aplicar técnicas avançadas para desenvolver visualizações gráficas.