Objetivos do projeto

Neste projeto cobrimos todas as etapas de um projeto real de Ciência de Dados e respondemos a algumas questoes importantes sobre o negócio utilizando dados com intuito de permitir que a empresa **Olist** tenha conhecimento sobre:

- Qual a distribuição dos pedidos por estado?
- Qual foi o ano com mais vendas e quais o meses que os clientes mais compram?
- Qual é a distribuição dos score dos reviews?
- Quais os 5 maiores vendedores?
- Quais as 5 categorias de produtos mais vendidos?
- Qual é a média de produtos vendidos num determinado pedido?
- Qual a relação entre o preço de venda e o valor do frete?
- Qual é a forma de pagamento mais utilizada?
- A maior parte dos pagamentos são feitos na totalidade ou de forma parcelada?
- Quantos métodos de pagamentos em média são escolhidos pelos clientes em um determinado pedido?

O objetivo é dar a conhecer o estado actual do negócio a Olist usando uma abordagem descritiva e criar uma solução para que a empresa possa prever as vendas diárias ao longo do tempo.

Por fim, realçar que o grande impacto que o projeto terá sobre o negócio da Olist, é que a esta ferramenta permitirá que a empresa investigue com mais detalhes o que contribuiu para que em determinados dias e/ou meses o número de vendas seja reduzido, de modo a definir-se estratégias a médio e longo prazo para aumentar as vendas da empresas.

Solução Proposta

Tecnologias Utilizadas

Para resolver este problema foi construida uma solução completa para armazenamento e gestão no **databricks** (usou o Google Colab como estrutura de armazenamento dos datasets para os notebooks de análise de dados) além de explorar uma suite de tecnologias e/ou bibliotecas para análise, visualização de dados e *machine learning* tais como: **pandas, matplotlib, seaborn, plotly, scikit-learn, statsmodels, fbprophet, streamlit e pyspark.**

Pandas – biblioteca usada para manipulação de dados

Matplotlib – biblioteca usada para visualização de dados.

Seaborn – biblioteca usada para visualização de dados baseada no matplotlib, permitindo construir graficos mais profissonais.

Plotly – biblioteca utilizada para desenvolver gráficos iterativos e construir *dashboards* usando chart studio.

Scikit-learn – biblioteca usada para implementar os algoritmos de *machine learning* (utilizou-se o pickle para serializar o modelo em disco)

Statsmodels – biblioteca usada para implementar métodos estatisticos e alguns algoritmos de séries temporaris como o ARIMA e SARIMAX.

Facebook Prophet – biblioteca usada para implementar algoritmos de *machine learning* para resolver problemas de séries tempiorais desenvolvido pela Microsoft.

Streamlit – biblioteca utilizada para desenvolver a aplicação e/ou fórmulário para testar o modelo em ambiente de produção.

Ferramentas de auxiliares:

Pyspark – para processamento de grandes volumes de dados em ambientes distribuído.

Python – liguagem de programação utilizada para desenvolver o projeto de ciência de dados.

Google colab – editor de código online que geralmente é organizado por células.

Github – ferramenta que permite versionar, partilhar o código desenvolvido e também atribuir acesso a outros profissinais para colaborarem nos artefatos do projeto.

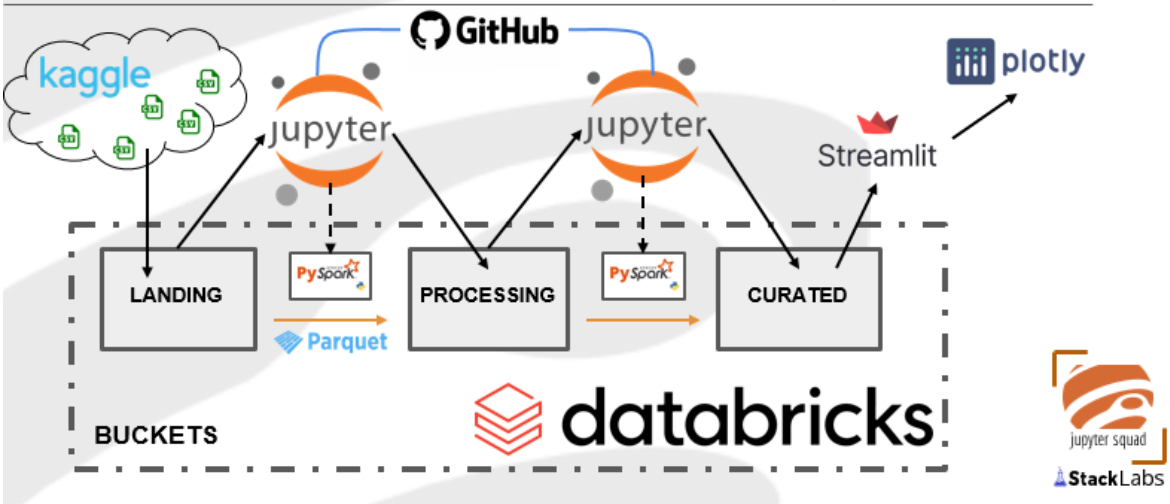
Por fim, realçar que utilizou-se como infraestrutura de armazenamento do **Google Colab** para poder analisar as diversas fontes de dados que estavam em arquivos no formato **csv** (dados disponibilizados pela Olist extraídos do Kaggle).

Overview Geral de Tecnologias Utilizadas no Projeto

Arquitecturas

Em seguida é ilustrada o overview da solução desde a **coleta** até ao **deploy** da solução desenvolvida**.**

Projetada pela squad



Utilizada no notebook de análise de dados devido aos desafios enfrentados e deadline do projeto

Os principais desafios enfrentados foram:

- **Integrar** o notebook do Google Colab com o ambiente do Databricks que serviria como infraestrutura para armazenamento de dados permitindo desata forma utilizar um ambiente distribuido aumentando a eficiência do processo de limpeza e transformação dos dados (não tivemos sucesso).
- **Disponilizar** a app na *cloud* da Streamlit, visto que, tivemos dificuldades em fazer upload dos arquivos do projeto armazenados na máquina local via terminal de comando do Windows para Github e erros de instalação de biblioteca na *cloud* da streamlit.
- **Instalar** algumas bibliotecas na máquina local, como por exemplo, o fbprophet.

Resultados

Insights e Conhecimento Gerado

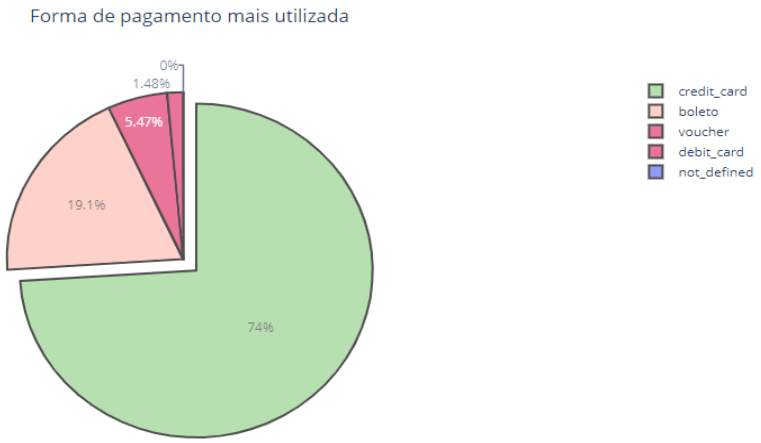
Na etapa de Análise Exploratória dos Dados foram descobertos vários insights importantes abaixo:

- As vendas tem uma tendência positiva ou de cresceimento ao longo dos anos.

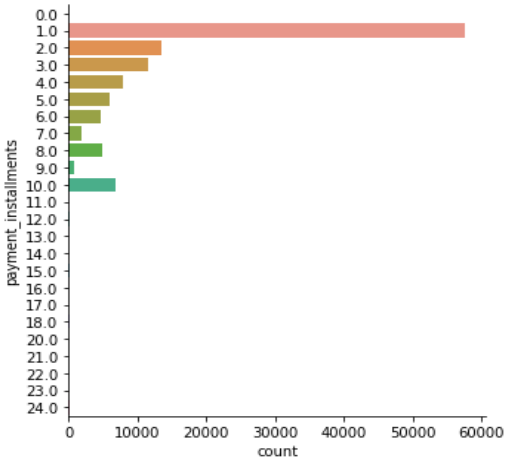
- 97% dos produtos são entregues aos clientes, ou seja, os pedidos estão no estado *delivered*.
- A maior parte dos reviews dos produtos é **positiva (média de 4.08)** e um dos títulos mais usado é o **Recomendo** e a mensagem mais escrita é o **Muito Bom**.
- Uma das categorias de produtos mais comprados é **cama_mesa_banho**.

product_category_name	price
cama_mesa_banho	11814
beleza_saude	9816
esporte_lazer	8791
moveis_decoracao	8643
informatica_acessorios	7963

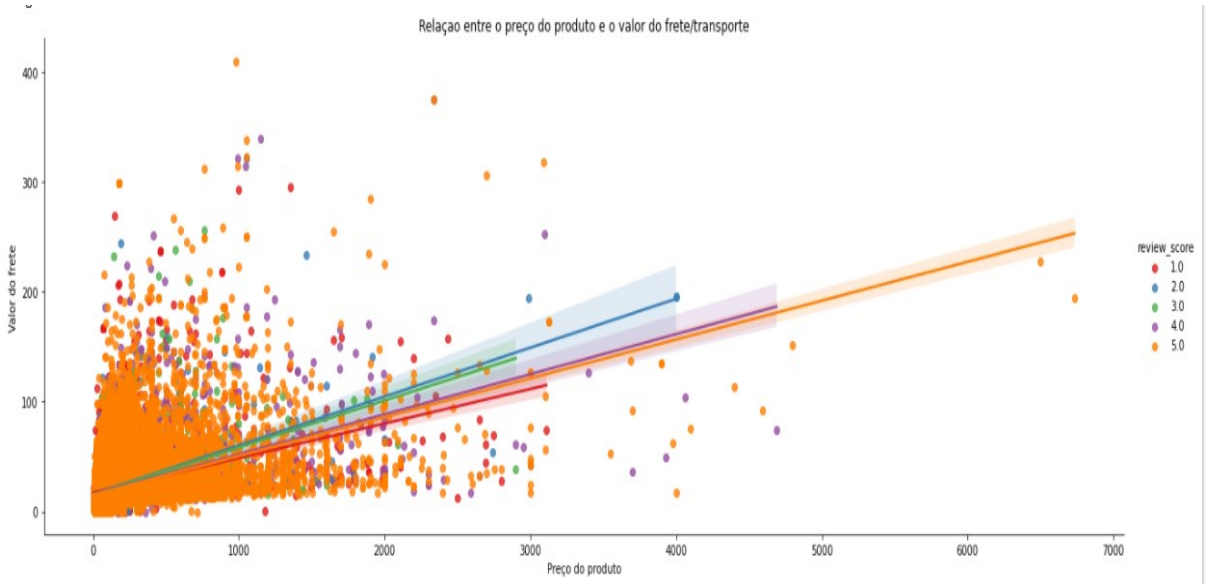
- Em **média** um pedido tem apenas 1 items/produto.
- Os clientes utilizam **5 formas de pagamento** das compras por si efectuadas e a forma de pagamento mais usada é **credit_card**



- A maior partes do pagamentos é feita em uma única prestação mas os pagamentos feitos em parcelas variam de **2** à **24 tranches.** No entanto, em média os clientes pagam em **3 prestações.**



- Existe uma correlação **positiva fraca (0.41)** entre o preço de venda e o valor de frete e também constatamos que a maior parte das vendas teve um **score** de **5** (classificação máxima), isto é, um review positivo.



- Na cadeia de lojas existentes **São Paulo** é a cidade que mais vende.

seller_city	price
sao paulo	2743479.78
ibitinga	664028.51
curitiba	485468.31
rio de janeiro	352680.77
guarulhos	329228.72

- Existem vendas sem **não pagas** ou com **valor de pagamento igual a zero**.
    - Os preços dos produtos variam de 0.85 a 6 735 reais.
    - Os preços dos fretes variam de 0 a 409.68 reais.
    - Perguntas que necessitam de ser aprofundadas em relação ao frete 0: Será que houve desconto de frete? O cliente foi levar o produto a loja?
    - Perguntas que necessitam de ser aprofundadas em relação a vendas nao pagas: Será que o cliente cancelou o pedido? O produto teve promoção de 100%?

Métricas de Performance

Para a estimativa de predizer o valor de venda diário foi implementado um modelo utilizando o **Facebook Prophet** que atingiu uma performance **RMSE** de aproximadamente **9 612.24** reais (superando um pouco um modelo de base que é de **11 848.30 reais**), isto é, em média o valor previsto **pode** diferir do valor real em **+/- 9 mil reais**.

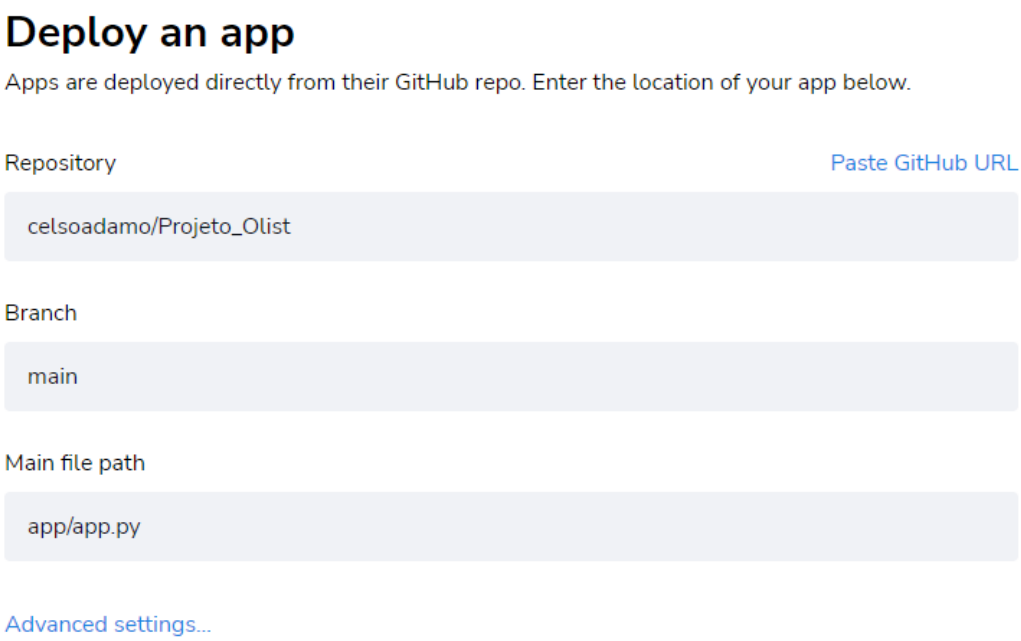
Conclusão

Através deste projeto foi possivel praticar e implementar conceitos importantes de Ciência e Engenharia de Dados e propor uma solução para um problema latente em qualquer empresa de e-commerce que é a análise da **situação actual** do negócio e a **previsão de vendas** dos meses subsequentes.

A resolução dos problemas acima mencionados, permitirá obter insights para desenhar e/ou alterar a estratégia do negócio da empresa, visto que, o e-commerce é um ambiente muito dinâmico e os clientes podem ter comportamentos diferentes de compras ao longo do tempo. Com a implementação desta solução teremos como beneficio, um recurso organizacional que permitirá identificar o comportamento de compra dos clientes e contribuirá para ajustar a estratégia de venda e retenção de clientes num horizonte temporal.

Por fim, como um processo de melhoria continua podemos reduzir o erro de previsão, desenvolver uma **automação** para executar o pipeline de coleta e transformação dos dados e automatizar a etapa de Machine Learning e Deploy.

Anexos



Link do Streamlit: https://share.streamlit.io/celsoadamo/projeto_olist/main/app/app.py