Análise:

Após ter acesso ao banco de dados pelo Kaggle e entender o que significava cada coluna, importei a base de dados para meu computador para começar as análises. Dando uma rápida olhada, reparei que havia muitos valores ausentes ('missing values'), então esse foi o primeiro passo a ser tomado. Utilizando as bibliotecas numpy e pandas, eu substituí os valores 'SEM' por 'NaN' (Not a number) por meio da linha:

'ecommerce_df.replace("SEM", np.nan, inplace=True)'.

Após isso, apaguei todas as linhas que apresentavam o valor 'NaN' com o auxílio do pandas, retirei as colunas que não seriam utilizadas para ficar mais organizado. Então, com o auxílio do pandas, selecionei os últimos 3 anos de compras e as utilizei para responder as seguintes perguntas:

Quais os produtos mais vendidos?

Parâmetros utilizados: "Product Category" (PC), "Quantity"(Q).

Organizados pelos campos com o mesmo "valor" e somando cada quantidade de produto vendido, cheguei à conclusão de que o produto mais vendido foi da categoria "clothing".

Qual o produto mais caro e o mais barato?

Parâmetros utilizados: "Product Category"(PC), "Product Price" (PP).

Como há muitos produtos com o mesmo valor, decidi selecionar e fazer a média da categoria de produtos. Criando variáveis para armazenar o preço médio por categoria de produto, foi feita uma média entre PC e PP. Para descobrir qual a PC mais cara, por meio do resultado entre a média, foi utilizada o pandas para pegar o ID mais alto e o ID mais baixo para a PC mais barata. Desta forma, chegando à conclusão de que o produto mais caro está na categoria "Home" com a média de preço de 255.13 e o mais barato está na categoria "Books" com a média de preço de 254.49.

Qual categoria de produto mais vendida e menos vendida?

Parâmetros utilizados: "Product Category" (PC), "Quantity"(Q).

Para achar a resposta para essa pergunta, decidi somar a quantidade de produtos de cada categoria para chegar à conclusão de que a categoria de produtos mais vendida foi "Clothing" com 165027 produtos vendidos e a categoria de produtos menos vendida sendo a "Home" com 110016 produtos vendidos.

Qual categoria mais e menos cara?

Parâmetros utilizados: "Product Category"(PC), "Product Price" (PP).

Para encontrar a categoria mais cara e a mais barata, foi calculado o preço médio das quatro categorias e foi buscado o maior ID entre as quatro categorias e o menor ID entre as quatro categorias, chegando ao resultado de categoria mais cara sendo "Home" e categoria mais barata sendo "Books".

Qual o produto com o melhor e pior NPS?

Parâmetros utilizados: "Product Category"(PC), "NPS"(N).

Procurei fazer a média do NPS de todas as categorias de produtos para ficar mais organizado e simples, e com isso tivemos acesso às categorias com melhor e pior NPS, sendo a melhor "Home" com 5.02 de NPS médio e "Electronics" com o pior NPS de 4.96.

Pensar em um modelo de dados para descobrir qual o melhor tipo de público (considerando gênero e idade) e o canal ideal para vender determinado tipo de produto?

Parâmetros utilizados: "Product Category"(PC), "Customer Age"(CA), "Gender"(G), "Source"(S).

Neste momento, foi feita uma tentativa de trabalhar com o modelo de regressão logística para encontrar o público ideal de cada PC por meio dos dados CA, G e S. Foi utilizado 20% dos dados para testes do modelo e o restante para treinamento, porém

ao fim do treinamento e inúmeros testes, foi visível que o modelo não ficou muito preditivo, tendo apenas 30% de precisão, sendo bastante abaixo do que eu gostaria.