Relatório do teste estágio em dados no Itaú

Para a realização deste teste utilizei as seguintes bibliotecas: Numpy, Pandas, Scipy, Matplotlib, Seaborn, Kmodes e Statsmodelsapi.

As principais técnicas estatísticas utilizadas foram análises e filtragens simples dos dados para responder às questões iniciais, ANOVA para estudar a relação entre Customer Age, Product Category e Source e análise de clusters para a relação entre as colunas Gender, Product Category.

1) **Quais os produtos mais vendidos considerando os últimos 3 anos?**

Para responder essa pergunta, limitei os dados àqueles referentes aos últimos 3 anos, os agrupei por Product Category e somei a quantidade vendida em cada categoria. A categoria com o maior número de vendas no período foi Clothing, seguida de Books e depois Electronics.

2) **Qual o produto mais caro e o mais barato?**

Para esta resposta, localizei os registros que continham o maior e o menor valor da coluna Product Price e obtive os dados das compras. O produto mais caro era da categoria Electronics, com Product Price igual a 500 e o mais barato era da mesma categoria com Product Price igual a 10. (Os dados completos de cada produto desta resposta são exibidos na execução do código)

3) **Qual a categoria de produto mais vendida e menos vendida? Qual a categoria mais e menos cara?**

Nesta questão agrupei os dados em cada Product Category, somei as quantidades totais de cada uma e calculei a média de preços de cada categoria. A categoria mais vendida foi Clothing, a menos vendida foi Home, a mais cara também Home e a mais barata Clothing, em média.

4) **Qual o produto com melhor e pior NPS?**

Nesta questão agrupei os dados por Product Category e calculando as médias de NPS por categoria. Depois localizei as categorias com melhor e pior NPS médio: Home e Electronics, respectivamente.

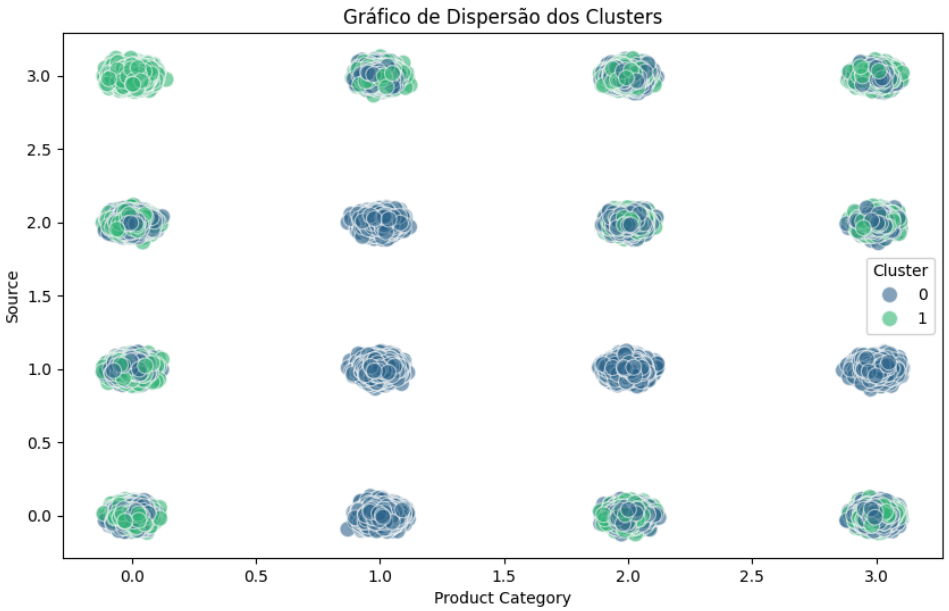
Para analisar o tipo de público e canal ideal para vender determinado tipo de produto, julguei necessário estudar a correlação entre as variáveis envolvidas nesta questão. Para isso, utilizei a estatística qui-quadrado. Os valores obtidos de qui-quadrado e p-value geraram as seguintes conclusões:

- Não deve haver correlação entre Gender e Source
- Deve haver forte correlação entre Customer Age e Source
- Não Deve haver correlação entre Gender e Product Category
- Não deve haver correlação entre Customer Age e Product Category

Para uma análise mais aprofundada da relação entre Gender, Product Category e Source, escolhi realizar uma análise de clusters. Por se tratar de 3 variáveis categóricas, utilizei o algoritmo kmodes para definir os clusters. As variáveis foram mapeadas para valores numéricos da seguinte forma:

Product Category: 0 = 'Books', 1 = 'Clothing', 2 = 'Electronics', 3 = 'Home'
Gender: 0 = ‘Female’, 1 = ‘Male’
Source: 0 = 'FaceBook campaign', 1 = 'Instagram Campign', 2 = 'Organic Search', 3 = 'SEM'

O gráfico de dispersão dos dados clusterizados foi gerado da seguinte forma:



Este gráfico permite visualizar de forma eficiente como os grupos se distribuem entre os valores de Source e Product Category e com isso fornece uma boa visão sobre o público ideal para cada par (Product Category X Source) levando em conta a variável Gender. Por exemplo, percebe-se que para valores de Product Category iguais a 1 (que correspondem à categoria Clothing) há uma grande concentração de observações pertencentes ao cluster 0. O mesmo acontece para valores de Source iguais a 1.

Já na análise da relação entre Customer Age e as variáveis Product Category e Source, escolhi aplicar a técnica ANOVA, por se tratar da relação entre duas variáveis categóricas e uma numérica. Defini o modelo ANOVA com estas variáveis e a tabela resultante é exibida a seguir:

	sum_sq	df	F	PR(>F)
C(df["Product Category"])	8.524248e+02	3.0	1.199428	0.308239
C(df["Source"])	6.039318e+03	3.0	8.497788	0.000012
C(df["Product Category"]):C(df["Source"])	1.258751e+03	9.0	0.590386	0.806169
Residual	5.922063e+07	249984.0	NaN	NaN

Os resultados obtidos na última coluna referem-se aos p-values de cada fator na análise. O baixo p-value (< 0.05) obtido para o fator Source mostra que Source tem um efeito significativo sobre o valor de Customer Age. Por outro lado os valores obtidos para Product Category e para a relação Product Category X Source mostram que estes não têm efeito significativo sobre Customer Age.

Gustavo Santana Santos