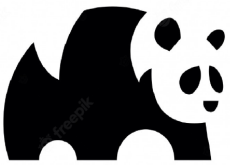


Adestradores de Pandas©

Analytics Company

Documentação do Projeto Módulo 6

Integrantes:

Bruno Rodrigues, Flavio WU, Lucas Bezerra, Thiago Charles

Sobre o DataSet.

O DataSet é referente aos dados distribuídos pela empresa Mercari.

A Mercari é o maior aplicativo de compras baseado na comunidade do Japão, fundada em 2013.

O seu serviço é o fornecimento de um local on-line, no qual o “cliente/vendedor” possa realizar o anúncio e venda do seu produto.

O DataSet consiste nas seguintes informações: train_id/test_id referente a listagem do produto em uma ordem numeral crescente. O nome do produto, a condição que se encontra o produto, a categoria que o produto se encontra , a marca do produto, o preço que foi vendido , forma de envio, descrição do item, data da venda e o estoque de quando foi vendido.

Link do dicionário

Insights.

Sobre dados nulos:

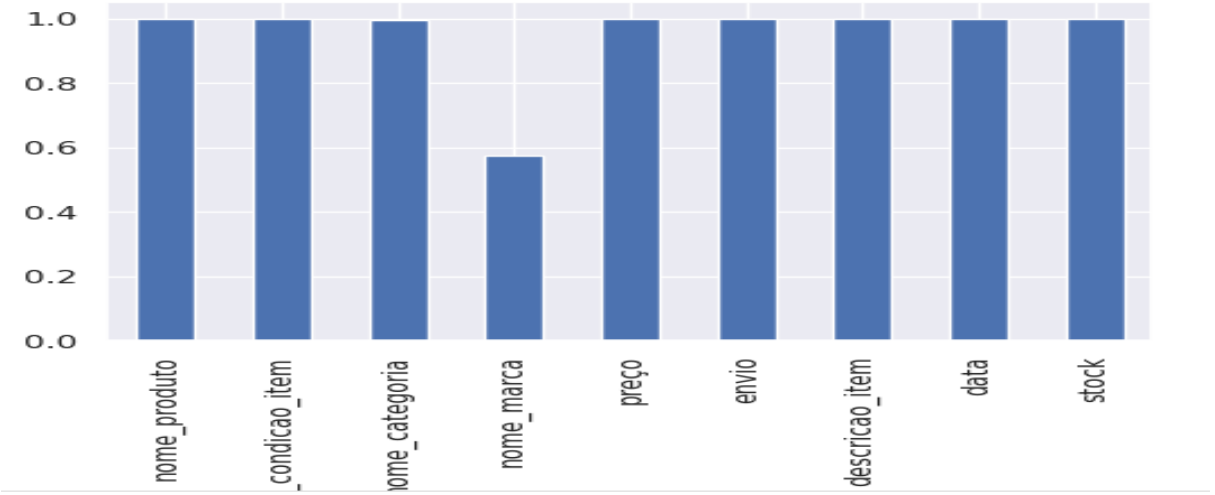
Ao iniciar as análises do Dataset, foi identificado a ausência de valores em três colunas distintas, que são, nome_marca, data e nome_categoria.

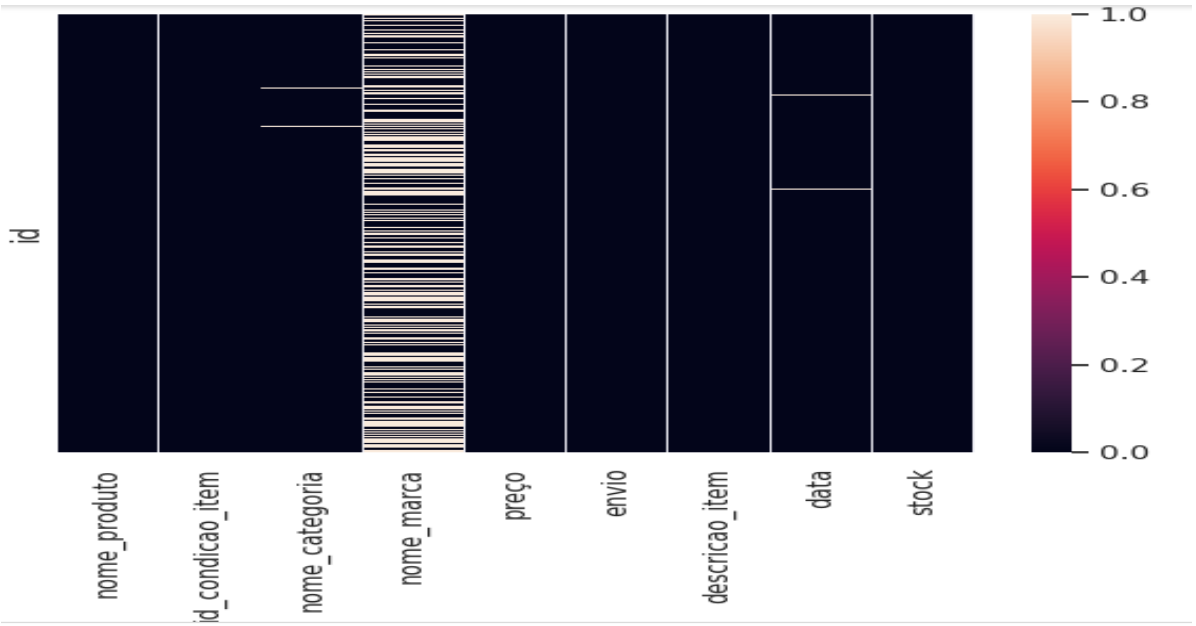
Nas colunas nome_marca e data, verificamos que os dados ausentes representam uma quantidade muito pequena, que não afetaria o DataSet se os excluirmos.

Na coluna nome_marca, a tratativa de exclusão dos dados tem que ser melhor estudada, já que a quantidade de dado ausente chega a ser 40% do DataSet.

1. Remover a coluna e manter as linhas;
2. Remover as linhas onde os dados são nulos;
3. Usar métodos de preenchimento de dados nulos.

Cada iniciativa descrita deve ser analisada pelo ponto de vista de performance e custo, onde técnicas de experimentação e comparação contribuem para tomada de decisão.





Sobre a coluna category_name:

A análise do DataSet revelou uma quantidade muito grande de produtos, categorias e descrições, no ano de 2018 foi realizado 1482535 vendas(linhas).

Dentro dessas vendas, encontramos 2232 itens com o mesmo nome.

Um dos pontos que notamos é grande quantidade de dados únicos na coluna nome_categoria, sobre esse fato resolvemos tratar, assim facilitando sua visualização e entendimento.

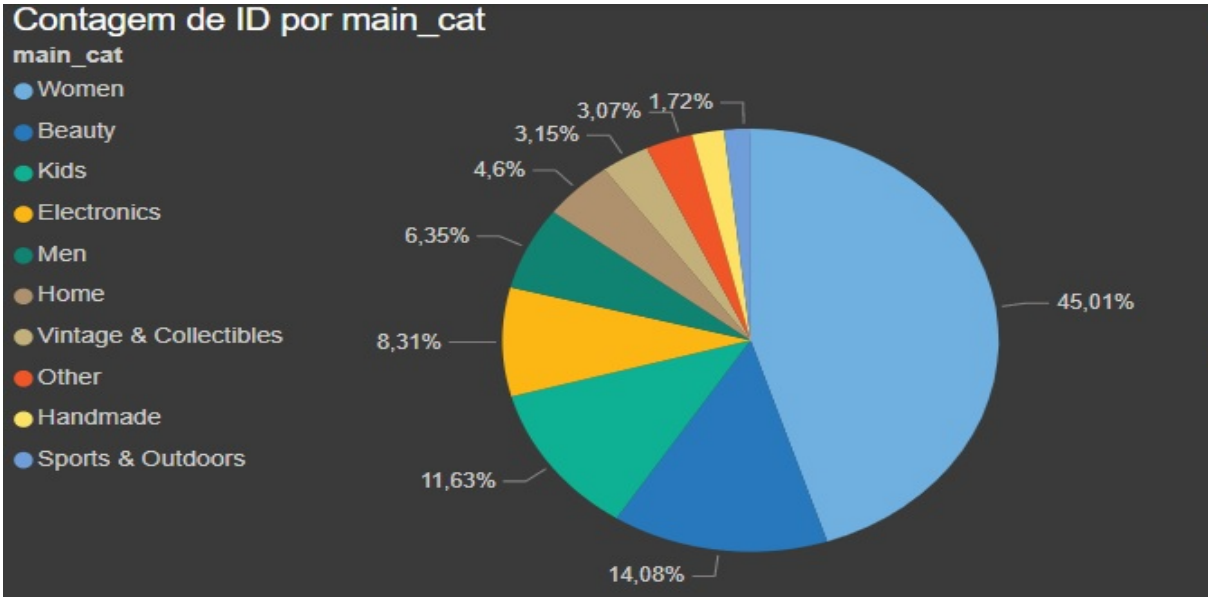
	nome_produto	nome_categoria	nome_marca	descricao_item	data
count	1482535	1476208	849853	1482531	1477853
unique	1225273	1287	4809	1281426	318
top	Bundle	Women/Athletic Apparel/Pants, Tights, Leggings	PINK	No description yet	11-2-2018
freq	2232	60177	54088	82489	4826

Dividimos a coluna **category_name** em níveis, a primeira palavra representa a **main_cat**, a segunda palavra representa a **sub_cat1** e a terceira palavra é a sub_cat2.

Na **main_cat**, encontramos 10 Categorias, que são: *Beauty, Electronics,Handmade, Home, Kids, Men, Other, Sports & Outdoors, Vintage & Collectibles e Women*. Encontramos 113 **sub_cat1** e 870 na **sub_cat2**.

Na categoria principal os produtos listados na categoria “Mulher” é a maior sendo responsável por 43%. Na **sub_categoria1** é a “Vestuário esportivo” e na **sub_cat2** é “Calças, Meias, Leggings”.

	categoria	sub_categoria	sub_categoria_item
count	1476208	1476208	1476208
unique	10	113	870
top	Women	Athletic Apparel	Pants, Tights, Leggings
freq	664385	134383	60177



Sobre a coluna item_description:

Como observado, no setor Infantil e de Maquiagem a maior parte dos produtos são comercializados “Novos” (condição 1), e as Categorias Vintage, Eletrônicos e Mulher contém maior quantidade de itens Semi-novos.

Na coluna **nome_marca**, foram encontrados registros de produtos repetidos. Com o objetivo de garantir a uniformidade dos dados, e diminuir o viés de classificação, manteve-se um exemplar de produto para cada marca.

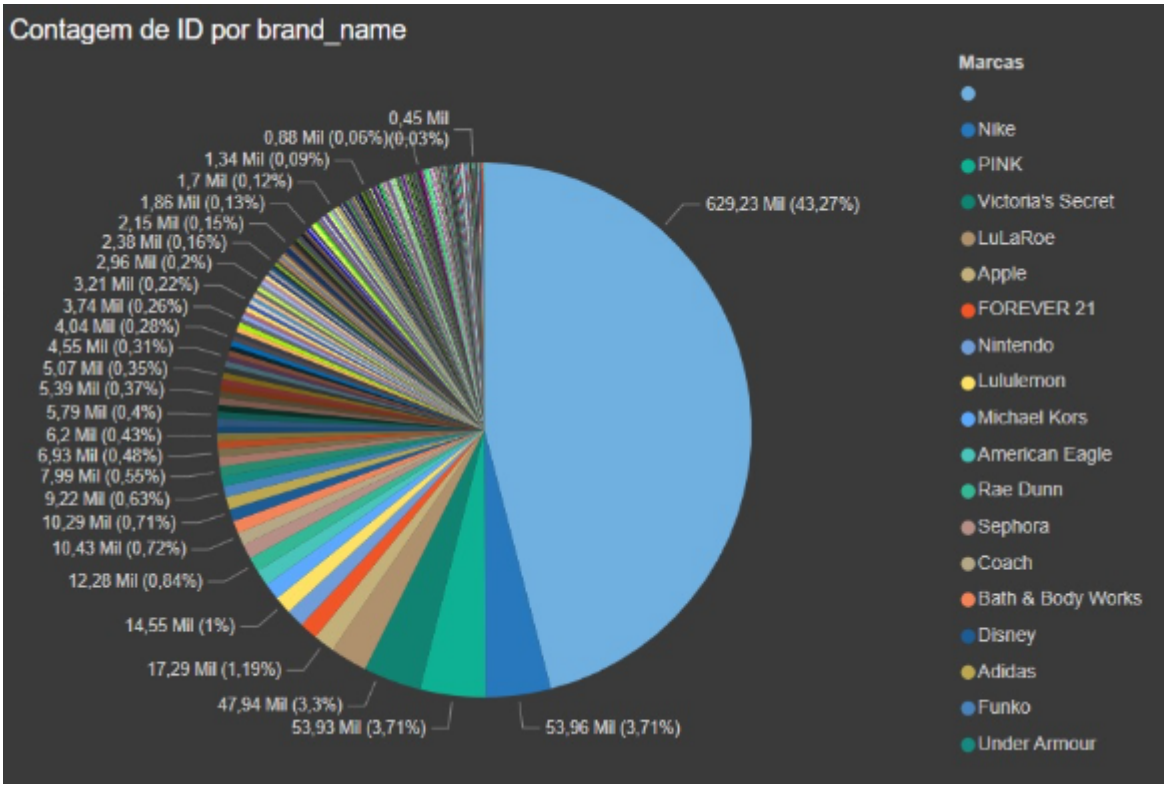
Antes:

nome_marca	nome_produto	
Michael Kors	Michael Kors	67
	Michael Kors Purse	64
Coach	Coach purse	63
Michael Kors	Michael Kors purse	63
Coach	Coach Purse	62
		..
Jockey	Jockey Supersoft Hipsters Size 8	1
	Jockey Spandex	1
	Jockey Size Small Zip Up Scrub Jacket	1
	Jockey Scrub Pant	1
wallis	Aztec print jumpsuit	1

Depois:

nome_marca	nome_produto	
!iT Jeans	IT! Jeans 27	1
PINK	PINK Landyard	1
	PINK Lace up Dorm Pants Super Soft Tee	1
	PINK Lace Up Pullover	1
	PINK Lace Up Black Tank Top Size S	1
		..
J. Crew	J. Crew Velvet Toothpick pants sz 29	1
	J. Crew V-Neck Cashmere Sweater	1
	J. Crew Twist Front One Piece Swimsuit S	1
	J. Crew Turquoise Chateau Parka 0P	1
wallis	Aztec print jumpsuit	1

Conforme indicamos anteriormente, a coluna **brand_name**, 67% dos dados são válidos e contêm as marcas e 43% dos dados são nulos. As marcas mais informadas são: NIKE, PINK e Victoria's Secret.

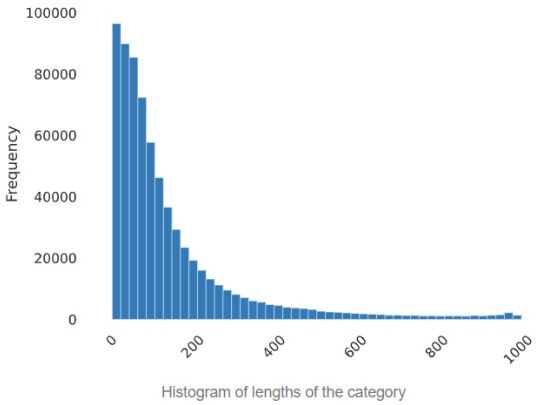


A coluna **item_description**, a descrição mais usada é a : ‘No description yet”, seguidas por “NEW” e “Brand new”.

Common Values

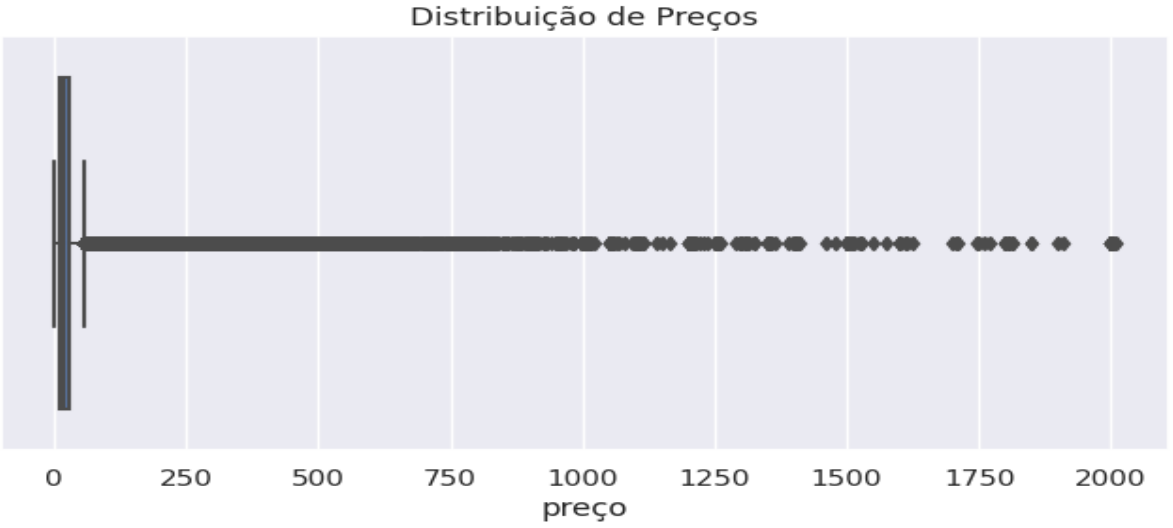
Value	Count	Frequency (%)
No description yet	38503	5.6%
New	1926	0.3%
Brand new	1470	0.2%
Good condition	650	0.1%
Great condition	579	0.1%
Like new	458	0.1%
Never worn	433	0.1%
NWT	384	0.1%
New with tags	311	< 0.1%
Excellent condition	272	< 0.1%
Other values (609545)	648373	93.5%

Length

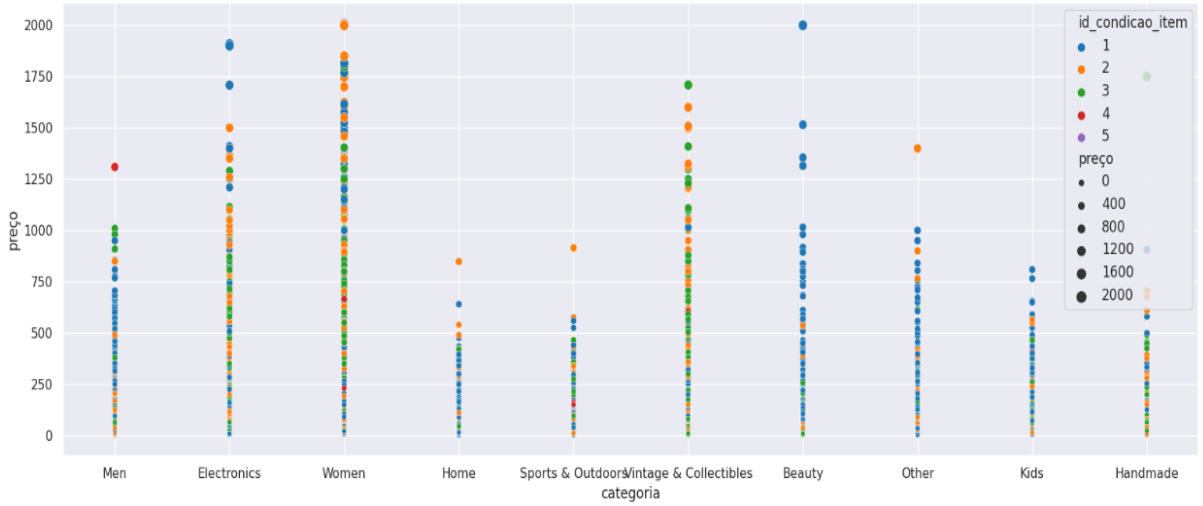


Sobre a coluna price, item_condition_id e shipping:

Referente a coluna **price**, no qual o conteúdo é referente ao valor que o produto foi vendido. Observamos que 75% dos produtos custam menos de $100,00 Dólares, enquanto outros alcançam a faixa de $ 2.000,00 Dólares.



A **category_name** influi diretamente nos limites de preços dos produtos e segmenta a maior concentração de diferentes Condição do item:



Notamos que a **item_condition_id** dos itens que mais aparecem, é a condição 1, com 43% do total.

Números totais:	
1	638324 - 43.2%
3	430402 - 29.2%
2	373302 - 25.3%
4	31803 - 2.2%
5	2373 - 0.2%

E os itens com as condições entre 1 e 3 se destacam com os maiores valores totais. Representando 97% do DataSet.

Números totais:	
1	16911379.0 - 42.8%
2	10294711.0 - 26.1%
3	11422223.0 - 28.9%
4	774887.5 - 2.0%
5	75354.0 - 0.2%

Preço médio por **item_condition_id** dos itens. Os itens com condição 5 se destacam com a maior média de preços.

Média do preço por categoria:
1 26.49
2 27.58
3 26.54
4 24.37
5 31.75

Verificamos qual é a Categoria que mais se repete nas colunas **main_cat**, **sub_cat_1** e **sub_cat_2**, referente a produtos com a condição 1.

	main_cat	sub_cat1	sub_cat2
count	638324	638324	638324
unique	10	113	829
top	Women	Makeup	Face
freq	228077	89130	33787

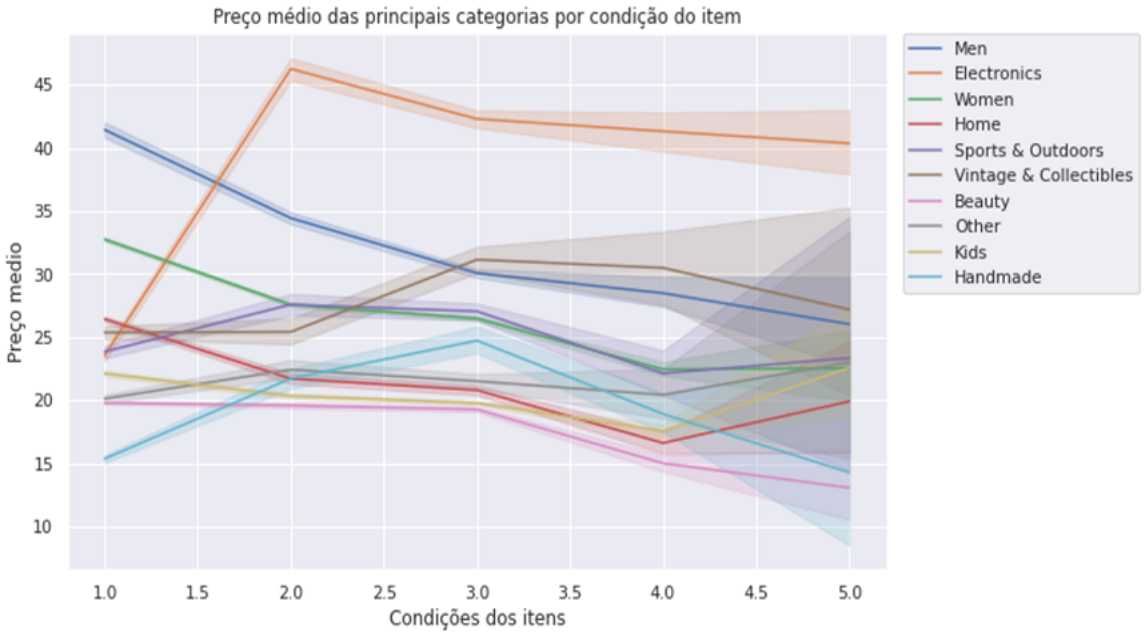
Women e Beauty são as principais categorias da coluna **main_cat**, com produtos identificados com a **item_condition_id** 1, chegando a mais de 59% do total dos itens.

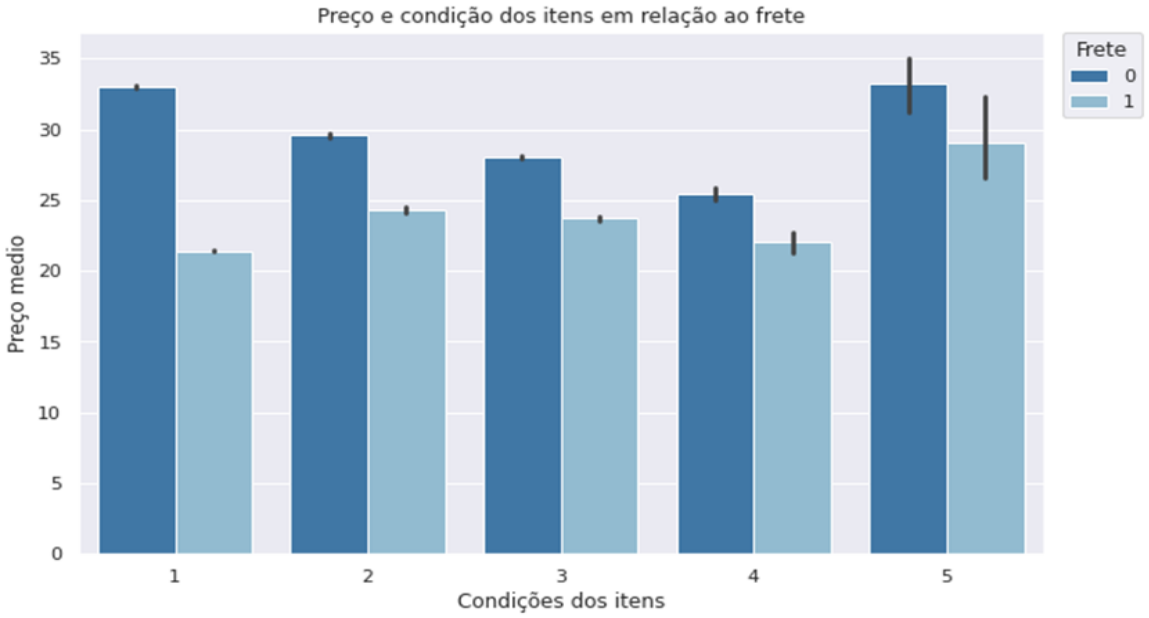
Frequência relativa das categorias com condição 1	
Women	35.7%
Beauty	23.5%
Kids	9.0%
Electronics	8.2%
Home	6.8%
Men	4.5%

Other	4.4%
Handmade	3.1%
Vintage & Collectibles	2.9%
Sports & Outdoors	1.9%

A categoria Women representa mais de 44% dos preços totais da **item_condition_id** 1, somando quase 7.5 milhões de dólares. Logo em seguida, com mais de 17% dos preços totais vem a categoria Beauty, somando quase 3 milhões de dólares.

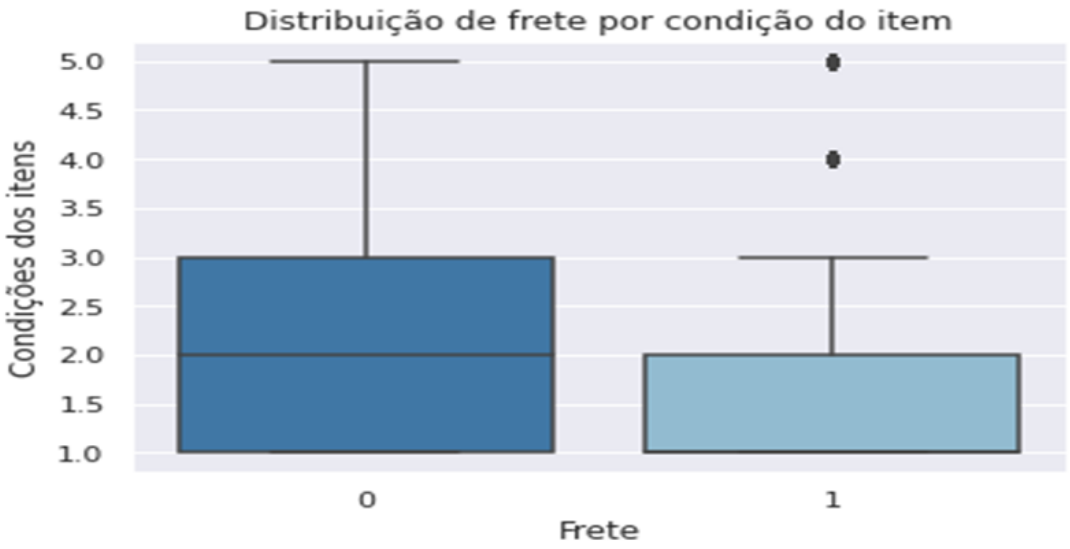
Categoria referente a itens com condição 1	
Women	44.2%
Beauty	17.6%
Kids	7.5%
Electronics	7.3%
Men	7.1%
Home	6.8%
Other	3.3%
Vintage & Collectibles	2.8%
Handmade	1.8%
Sports & Outdoors	1.7%





Visualização do preço médio das principais categorias por condição do item

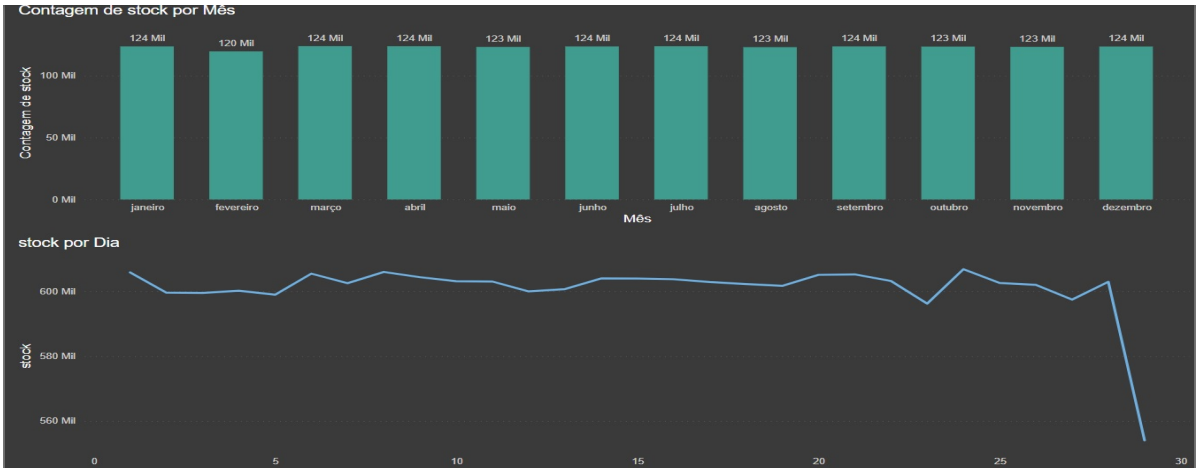
O preço médio, independente da condição do item, corresponde ao esperado em relação ao frete. Os itens com o preço médio maior possuem a maior parte dos itens com frete grátis.



Sobre a coluna stock e date:

Analisando as que a coluna **stock**, notamos que elas seguem um padrão que se matem o ano todo, não identificando sazonalidade.

Como não temos a identificação do vendedor e da saída desse material não conseguimos identificar se tem reposição de **stock**, com isso não podemos saber se o item foi vendido.



Sobre dados repetidos e correção:

Conforme apresentado inicialmente, alguns valores estão duplicados no DataSet. Eles foram retirados com o mesmo objetivo de diminuir o viés. A quantidade de dados únicos, e exposto por Categoria são:

```
Women                    555864
Beauty                   196220
Kids                     167524
Electronics              116785
Men                       90409
Home                      65163
Vintage & Collectibles    45349
Other                     43207
Handmade                  28859
Sports & Outdoors         24394
```

Outliers representam 1.9881 % do conjunto de dados e o número por “Categoria” é demonstrado a seguir:

```
Women                    10849
Electronics               6845
Men                       3536
Kids                      1527
Vintage & Collectibles    1107
Beauty                     957
Home                       625
Other                      416
Sports & Outdoors          377
Handmade                   278
```

Foi retirada uma pequena amostra de 7.000 valores da coluna “Descrição do Item", a fim de analisar suas características. Verificou-se diferentes tipos de Entidades, conforme a imagem abaixo:

```
['CARDINAL',
 'DATE',
 'EVENT',
 'FAC',
 'GPE',
 'LAW',
 'LOC',
 'MONEY',
 'NORP',
 'ORDINAL',
 'ORG',
 'PERCENT',
 'PERSON',
 'PRODUCT',
 'QUANTITY',
 'TIME',
 'WORK_OF_ART']
```

Foram encontradas 292 Empresas na coluna Descrição nesta amostra de: 7000 registros. E as palavras mais frequentes são.

A coluna de Descrição apresenta informações como Cor, Condição do Item, Marca, Data... Através desses dados o algoritmo poderá reconhecer padrões e ajudar a classificar melhor.
Pode-se criar um atributo que agrupe produtos com preço acima, abaixo ou dentro da média de consumo e verificar se as palavras comuns - para diferentes categorias de itens - também são distintas. Se forem, esses dados poderão melhorar a eficiência do Modelo.

Hipótese

Os nomes das empresas aparecem na “Descrição”.Podemos usar essa informação para preencher os dados nulos de marca? Ou utilizando a coluna de Descrição junto aos demais atributos a coluna de “Marca” torna-se irrelevante?

O modelo utilizado tem resultado melhor sem utilizar a coluna **brand_name** ?Ou o modelo tem um resultado melhor com a coluna mas excluindo as linhas sem informação.

A condição do item tem um fator mais ponderante em qual categoria?

A coluna date e stock podem ser excluídas por não apresentarem utilidade na análise, os stocks são apresentam as mesmas movimentações o ano todo.

Qual é a coluna mais importante para precipitação do preço?

Se fossemos vender um produto,qual seria o melhor?

Link:
Repositório Github: https://github.com/flaviowu/btc-c14-g4
Miro : https://miro.com/app/board/uXjVPZ-bhT0=/
Trello :.https://trello.com/b/Y6SpaCOm/tarefas-btc