**Documentação do Projeto Módulo 6**

**Grupo:** G4

**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 (name), a condição que se encontra o produto (item\_condition\_id), a categoria que o produto se encontra (category\_name), a marca do produto (brand\_name), o preço que foi vendido (price), forma de envio (shipping), descrição do item (item\_description), data da venda (date) e o estoque de quando foi vendido (stock).

[Link do dicionário](https://github.com/flaviowu/btc-c14-g4/blob/main/dados/dicionario_rev_1.md)

**Insights.**

**Sobre dados nulos:**

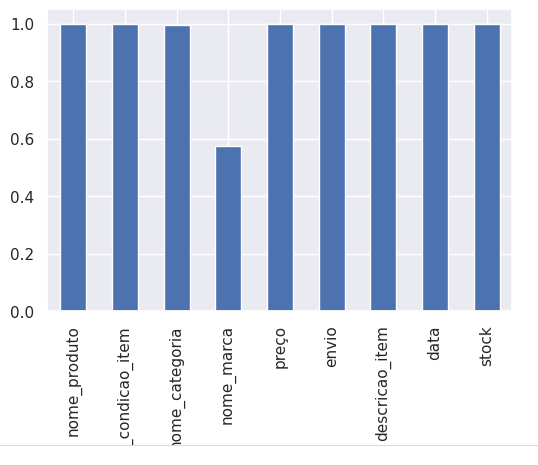
Ao iniciar as análises do DataSet, identificado a ausência de valores em três colunas distintas, que são, nome\_marca(**brand\_name)** , data(**date)** e nome\_categoria(**category\_name).**

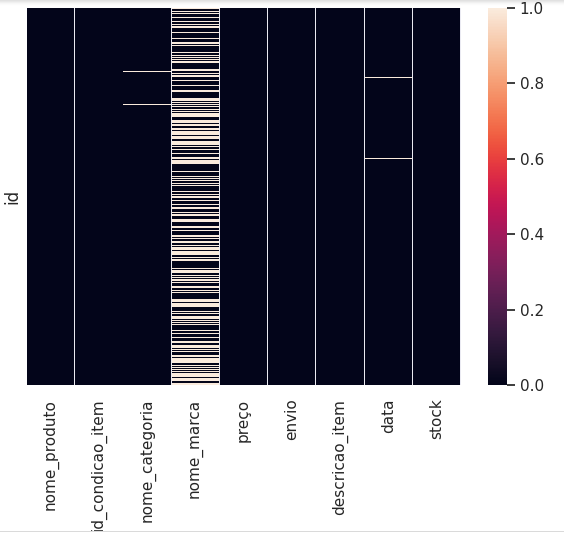
Nas colunas nome\_marca(**brand\_name)** e data(**date),** verificamos que os dados ausentes representam uma quantidade muito pequena, que não afetaria o DataSet se os excluirmos.

Na coluna **brand\_name,** 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.



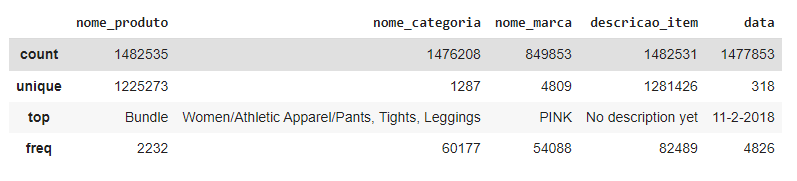


**Sobre a coluna category\_name:**

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.

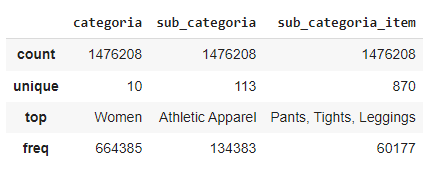
Um dos pontos que notamos é grande quantidade de dados únicos na coluna nome\_categoria (category\_name), sobre esse fato resolvemos tratar, assim facilitando sua visualização e entendimento.

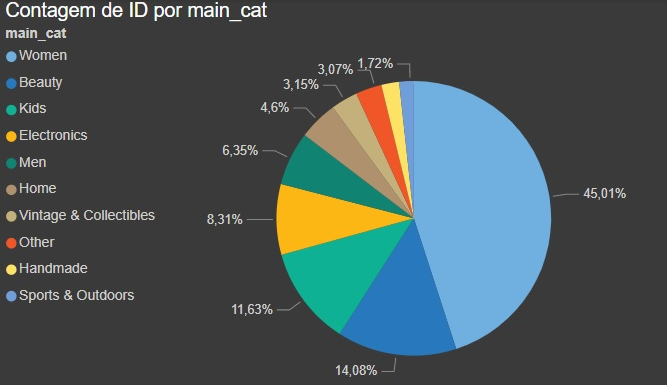


Dividimos a coluna **category\_name** em níveis, a primeira palavra representa a **main\_cat**, a segunda palavra representa a **sub\_cat1** e a terceira palavra é a sub\_cat2.

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”.



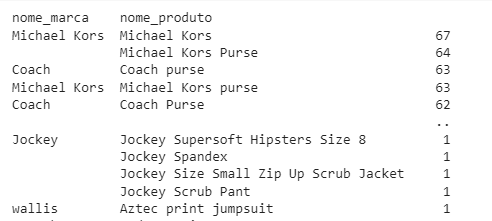


**Sobre a coluna item\_description:**

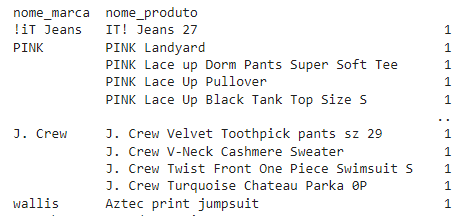
Como observado, no setor Infantil e de Maquiagem a maior parte dos produtos são comercializados “Novos” (condição 1), e as Categorias Vintage, Eletrônicos e Mulher contém maior quantidade de itens Semi-novos.

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.

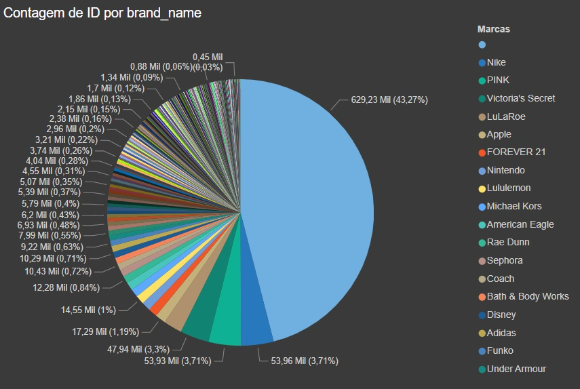
Antes:



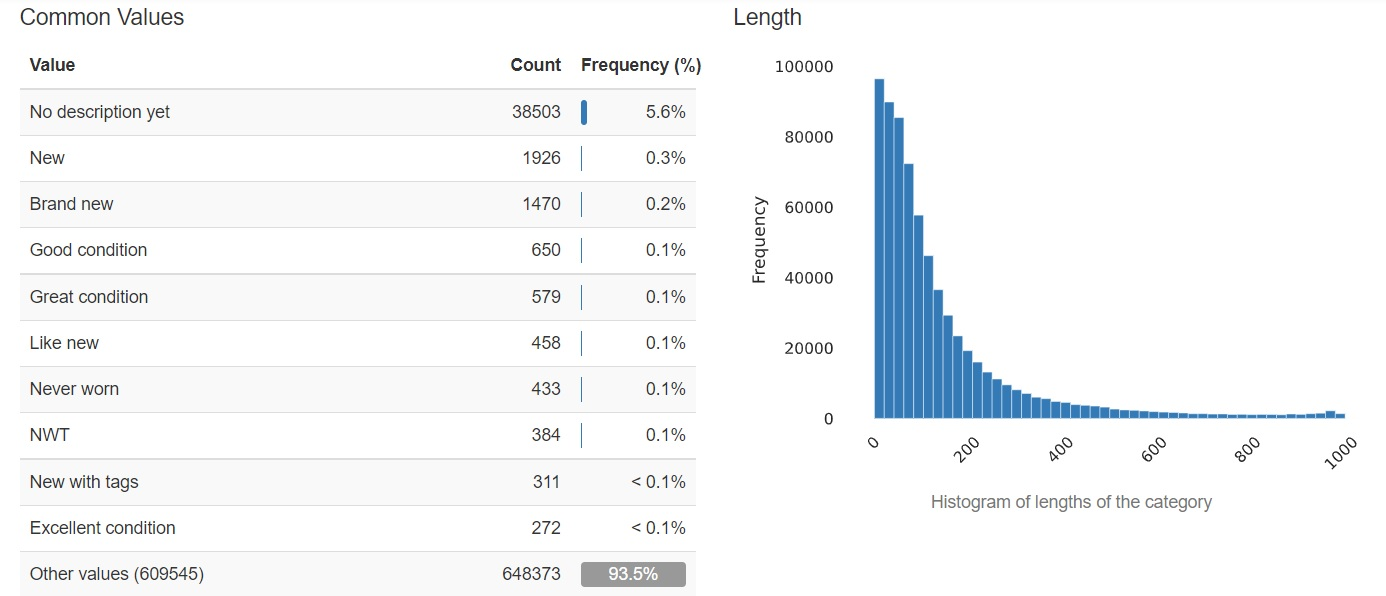
Depois:



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.

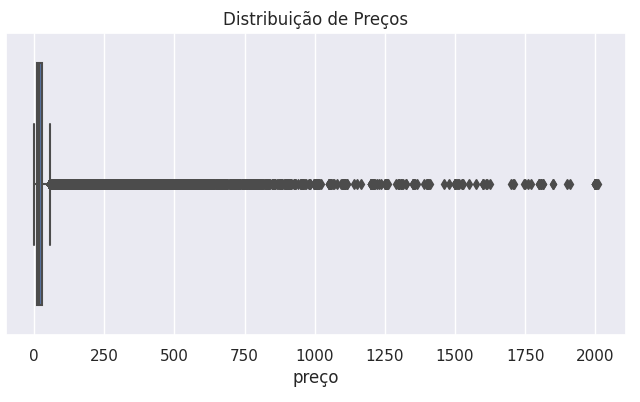


A coluna **item\_description**, a descrição mais usada é a : ‘No description yet”, seguidas por “NEW” e “Brand new”.

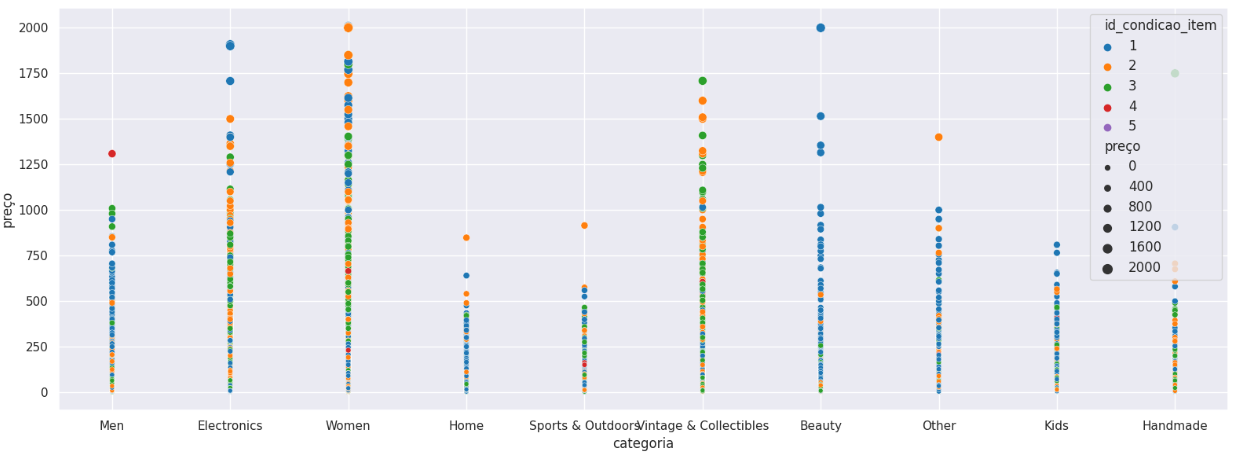


**Sobre a coluna price, item\_condition\_id e shipping:**

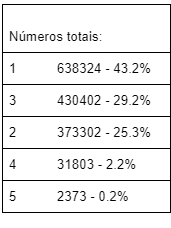
Referente a coluna **price**, no qual o conteúdo é referente ao valor que o produto foi vendido. Observamos que 75% dos produtos custam menos de $100,00 Dólares, enquanto outros alcançam a faixa de $ 2.000,00 Dólares.



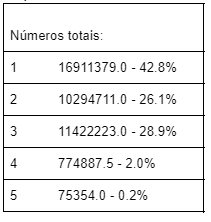
A **category\_name** influi diretamente nos limites de preços dos produtos e segmenta a maior concentração de diferentes Condição do item:



Notamos que a **item\_condition\_id** dos itens que mais aparecem, é a condição 1, com 43% do total.



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



#### Preço médiopor **item\_condition\_id** dos itens. Os itens com condição 5 se destacam com a maior média de preços.

#### 

#### 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.

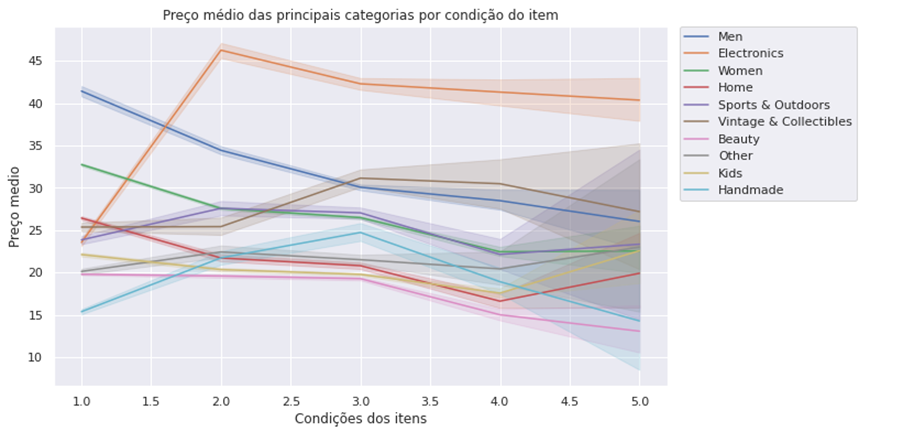
|  | main\_cat | sub\_cat1 | sub\_cat2 |
| --- | --- | --- | --- |
| count | 638324 | 638324 | 638324 |
| unique | 10 | 113 | 829 |
| top | Women | Makeup | Face |
| freq | 228077 | 89130 | 33787 |

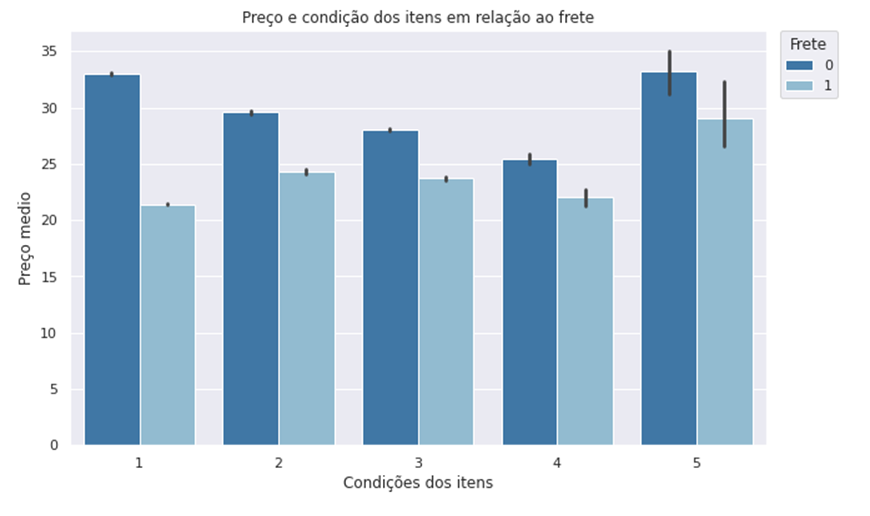
#### Women e Beauty são as principais categorias da coluna **main\_cat**, com produtos identificados com a **item\_condition\_id** 1, chegando a mais de 59% do total dos itens.

| 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.

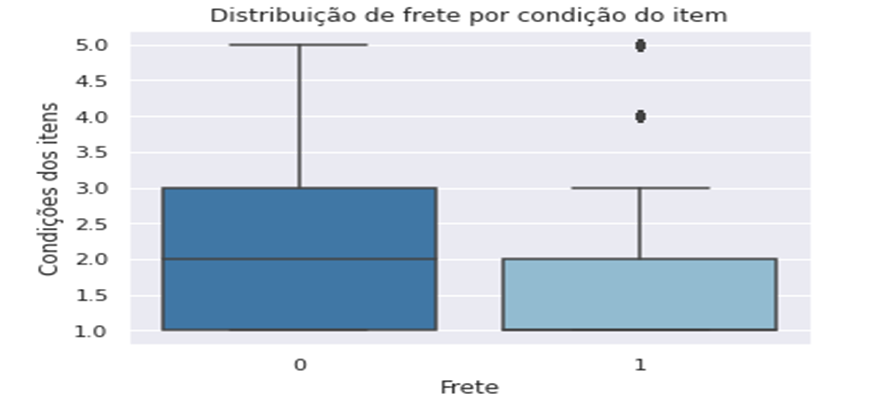
| Categoria referente a itens com condição 1 |
| --- |
| Women 44.2% |
| Beauty 17.6% |
| Kids 7.5% |
| Electronics 7.3% |
| Men 7.1% |
| Home 6.8% |
| Other 3.3% |
| Vintage & Collectibles 2.8% |
| Handmade 1.8% |
| Sports & Outdoors 1.7% |





#### Visualização do preço médio das principais categorias por condição do item

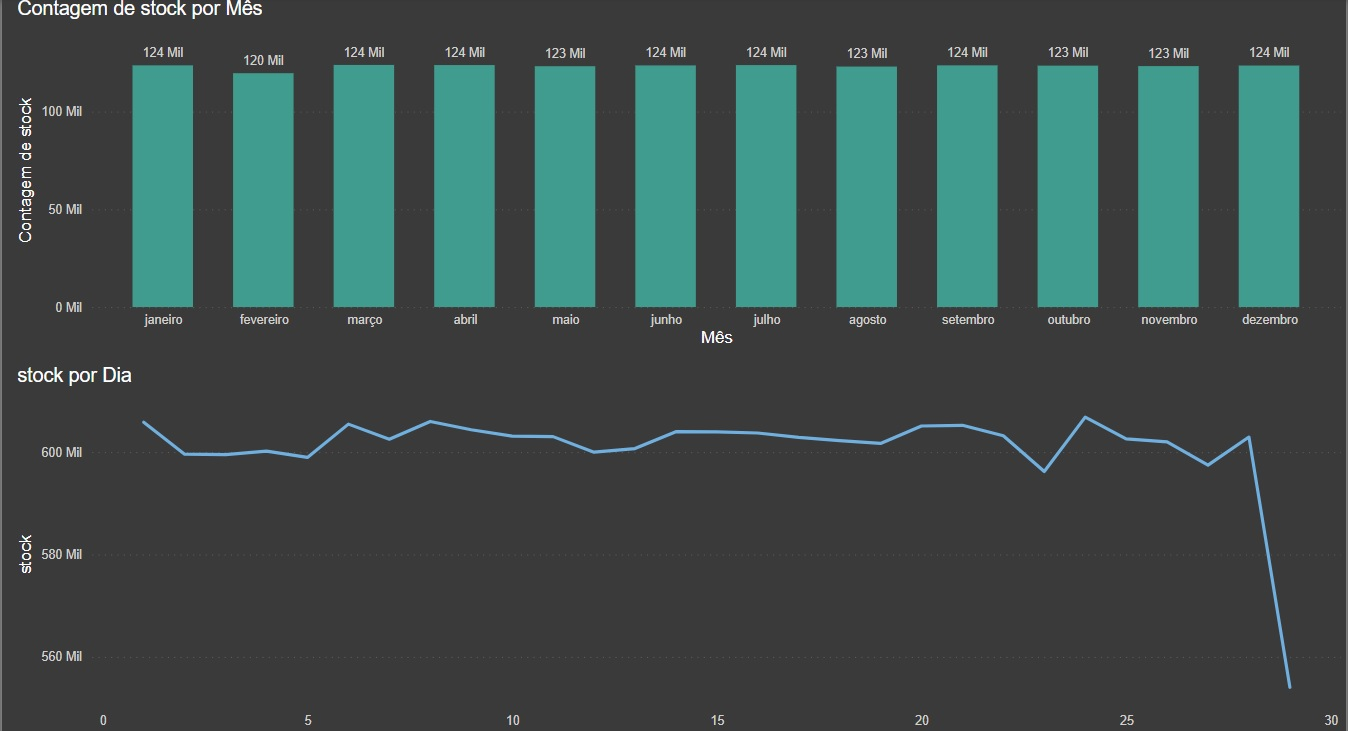
O preço médio, independente da condição do item, corresponde ao esperado em relação ao frete. Os itens com o preço médio maior possuem a maior parte dos itens com frete grátis.

****

**Sobre a coluna stock e date:**

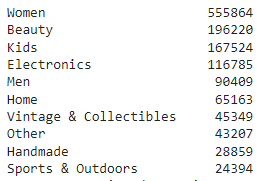
Analisando as que a coluna **stock**, notamos que elas seguem um padrão que se matem o ano todo, não identificando sazonalidade.

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.

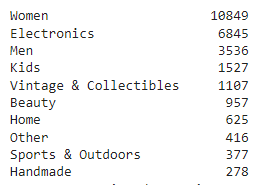


**Sobre dados repetidos e correção:**

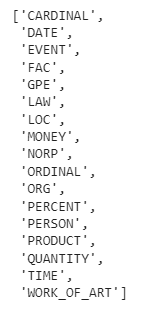
Conforme apresentado inicialmente, alguns valores estão duplicados no DataSet. Eles foram retirados com o mesmo objetivo de diminuir o viés. A quantidade de dados únicos, e exposto por Categoria são:



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



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:



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>