

Análise de dados com PySpark

Humberto camara marriel
Douglas Lopes Amora
Ismael Wesley Neves de Brito
Mateus do Nascimento Magalhães da Silva



OBJETIVO:

O presente trabalho tem como objetivo, demonstrar como uma ferramenta python chamada de PySpark pode ser utilizada para se fazer análise de dados, tais como sua eventual manipulação e sua transformação de dados para informação.

INTRODUÇÃO

- Foi utilizado neste trabalho o Pyshark para a importação de dados que estavam em CSV para que se pudesse ser realizado a tarefa de análise dos dados.
- Foram criadas algumas funções no Python para podermos utilizar alguns tipos de gráficos e realizar outras rotinas, como a importação.
- Para a visualização, utilizamos gráficos de barras e de pizza, com o principal intuito de mostrar as diferenças entre as informações.
- Para poder correlacionar esses dados e utilizá-los com gráficos visuais e intuitivos, foi utilizado uma biblioteca python chamada de Matplotlib

OBSERVAÇÕES IMPORTANTES:

- Para fins de tornar mais fácil a reprodução deste trabalho e seus eventuais testes, foi utilizada a ferramenta Google Colab como ambiente de execução Python.
- As análises foram feitas utilizando-se apenas os 10 resultados com maiores quantidades na base de dados, e caso a quantidade esteja repetida os valores serão exibidos ordenados pelo nome.
- BASE DE DADOS UTILIZADA:
base_info_produtos.csv - (Informações sobre produtos de um sistema)

Instalação do Ambiente PySPark

```
# instalar as dependências
!apt-get install openjdk-8-jdk-headless -qq > /dev/null
!wget -q https://archive.apache.org/dist/spark/spark-2.4.4/spark-2.4.4-bin-hadoop2.7.tgz
!tar xf spark-2.4.4-bin-hadoop2.7.tgz

# configura as variáveis de ambiente e o Spark
import os
os.environ["JAVA_HOME"] = "/usr/lib/jvm/java-8-openjdk-amd64"
os.environ["SPARK_HOME"] = "/content/spark-2.4.4-bin-hadoop2.7"
!pip install -q findspark

# tornar o pyspark "importável"
import findspark
findspark.init('spark-2.4.4-bin-hadoop2.7')

# iniciar uma sessão local
from pyspark.sql import SparkSession

# importar biblioteca para gerar gráficos
import matplotlib.pyplot as plt
spark = SparkSession.builder.master("local[*]").appName("DadosLoja").getOrCreate()
```

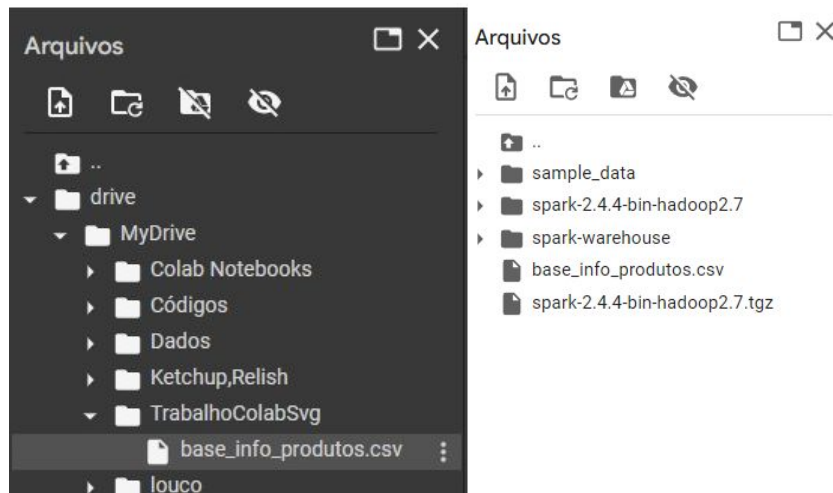
Importando os dados da planilha do excel

```
#Cria o DataFrame utilizando uma planilha localizada neste endereço
df=data("/content/drive/MyDrive/TrabalhoColabSvg/base_info_produtos.csv",separador=' ')
```

```
#Exibe as 50 primeiras linhas do DataFrame
df.show(50,True)
```

```
#Conferir se o separador da planilha é o separador ;
df=data(dados,separador=';')
```

Os dados podem ser importados direto do Google Drive



nome	tipo	marca	categoria	cor	modelo
Samsung UN40C6900...	TV	Samsung	Eletrônicos	None	None
Sapateira Limeira...	Armario	Politorno	Casa e Decoração	None	None
Faqueiro Tramonti...	Faqueiro	Tramontina	Casa e Decoração	Inox	None
Cartucho de tinta...	Cartucho	HP	None	None	None
Bolsa Pure Evo Me...	Bolsa	Puma	Moda e Acessórios	None	None
Docking Station 8...	Docking Station	Maxprint	Eletrônicos	None	None
Bandeirante Ecojipe	Mini Veículo	Bandeirante	Brinquedos	None	None
Câmera Digital Sa...	Câmera Digital	Samsung	Eletrônicos	Preto	42lw4500
Ducha Manual Lore...	Chuveiro	Lorenzetti	None	None	None
Espagueteira de A...	Espagueteira	MTA	None	None	None
Americanflex Post...	Colchão	Americanflex	Casa e Decoração	None	None
Philips SA2VBE16K...	MP Player	Philips	Eletrônicos	None	None
Mondial C-06 Elét...	Cafeteira	Mondial	Eletrodomésticos	None	c-06
Máquina Profissio...	Máquina de cortar...	WAHL	Cosméticos e Perf...	None	None
Caneta Esferográ...	Caneta	Pentel	None	None	None
Guitarra LP Nashv...	Guitarra	Shelter	None	None	None
Grill Cotherm 1301	Churrasqueira	Cotherm	Esporte e Lazer	Preto	None
Adesivo p/ Motoro...	Adesivo	iSkin	None	None	None
Ventilador/Exaust...	Grafite	Tron	None	Grafite	None
Rochedo Elegance ...	Conjunto de Painelas	Rochedo	Casa e Decoração	Inox	None
Electrolux FE26 V...	Freezer	Electrolux	Eletrodomésticos	None	fe26
Burigotto AT6 Berço	Carrinho de Bebê	Burigotto	Brinquedos	None	None
Relógio Masculino...	Relógio	Backer	Moda e Acessórios	None	None
Smart Trike Recli...	Carrinho de Bebê	Dican	Bebês e Cia	Vermelho	None

Funções criadas para análise dos dados

```
def data(caminho, separador=';') :
```

Importa um CSV para o programa através do parâmetro **caminho** separando os valores através do parâmetro **separador**, que por padrão é “ ; “ quando omitido.

Cria um novo DataFrame e o utiliza como dados principal para o programa.

Retorna esse DataFrame para ser utilizado quando necessário.

```
def tabela(var, total=10) :
```

Exibe uma coluna de quantidades em formato de tabela da base de dados, sendo o nome dessa coluna o parâmetro **var**.

O valor mostrado é agrupado por nome, e ordenado pela quantidade de produtos e depois pelo nome, ou seja, mostrará sempre os valores com maior quantidade na base de dados.

A quantidade máxima de valores a serem mostrados vem do parâmetro **total** que por padrão é 10 caso omitido.

```
def pizza(var, total=10) :
```

Exibe um gráfico de pizza com uma coluna da base de dados, sendo o nome dessa coluna o parâmetro **var**.

O valor mostrado é agrupado por nome, e ordenado pela quantidade de produtos e depois pelo nome, ou seja, mostrará sempre os valores com maior quantidade na base de dados.

O percentual mostrado é em relação aos dados obtidos durante a consulta e não em relação a base total de dados, por exemplo, se a variável **total** estiver com valor 5, então o percentual será dos valores entre esses 5 dados.

Funções criadas para análise dos dados

```
def grafico(var, total=10) :
```

Exibe um gráfico de barras com uma coluna da base de dados, sendo o nome dessa coluna o parâmetro **var**.

O valor mostrado é agrupado por nome, e ordenado pela quantidade de produtos e depois pelo nome, ou seja, mostrará sempre os valores com maior quantidade na base de dados.

A quantidade máxima de valores a serem mostrados vem do parâmetro **total** que por padrão é 10 caso omitido.

```
def lista(var, total=10) :
```

Busca uma coluna de quantidades em formato de tabela da base de dados e retorna o resultado em uma lista, sendo o nome dessa coluna o parâmetro **var**.

O valor mostrado é agrupado por nome, e ordenado pela quantidade de produtos e depois pelo nome, ou seja, mostrará sempre os valores com maior quantidade na base de dados.

A quantidade máxima de valores a serem retornados vem do parâmetro **total** que por padrão é 10 caso omitido.

```
def busca(Categoria, Buscar, Total) :
```

Busca linhas da base de dados e exibe em formato de tabela.

Essa busca precisa de 2 parâmetros, o primeiro **Categoria** é o nome do campo que vai ser pesquisado na base de dados e o **Buscar** é o conteúdo que essa coluna tem que conter.

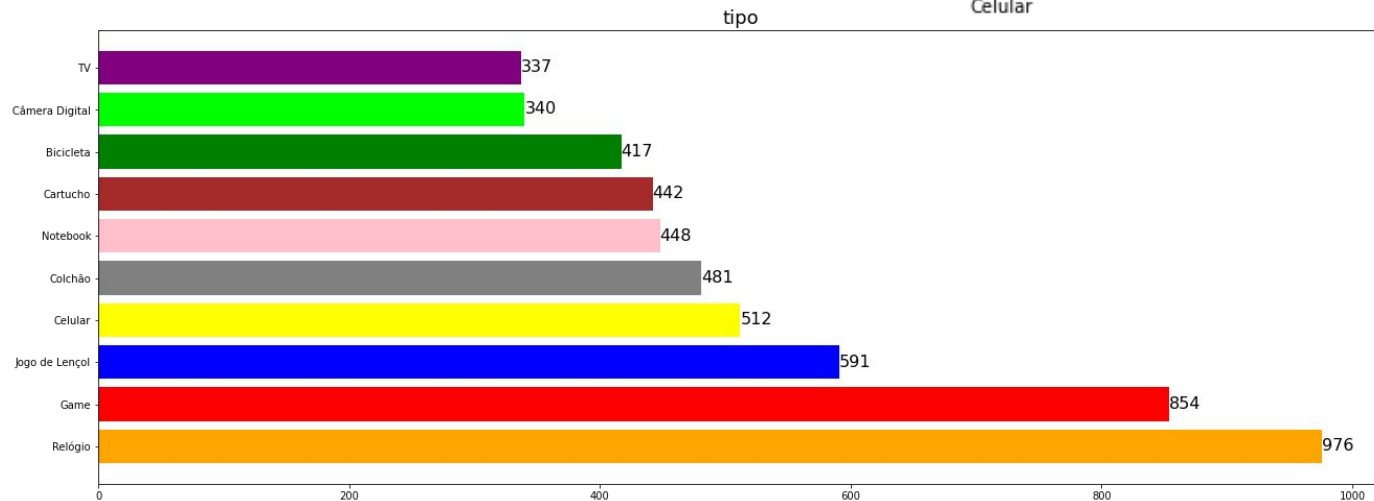
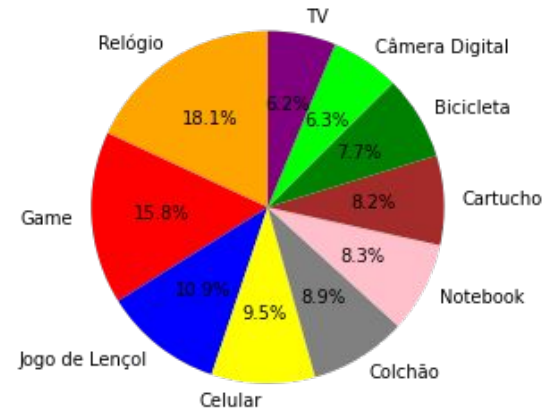
A quantidade máxima de valores a serem retornados vem do parâmetro **Total** que por padrão é 10 caso omitido.



Análises realizadas com a base de dados “base_info_produtos”

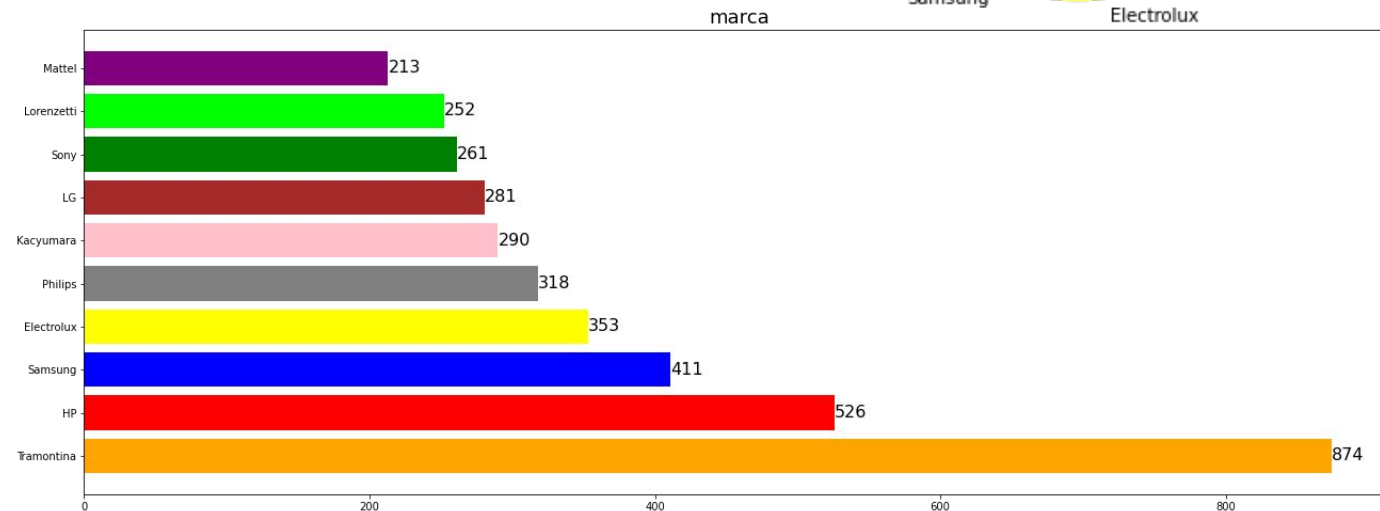
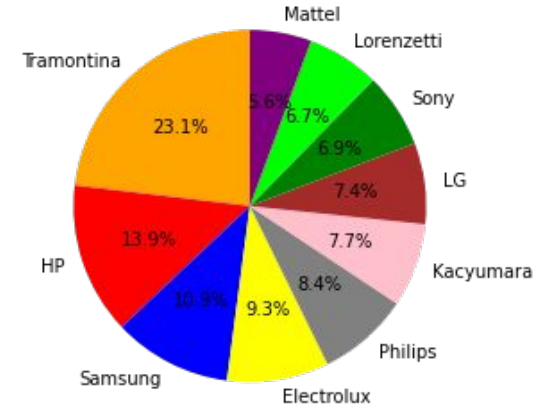
ANÁLISES RELACIONADAS AO TIPO

tipo	qtde
Relógio	976
Game	854
Jogo de Lençol	591
Celular	512
Colchão	481
Notebook	448
Cartucho	442
Bicicleta	417
Câmera Digital	340
TV	337



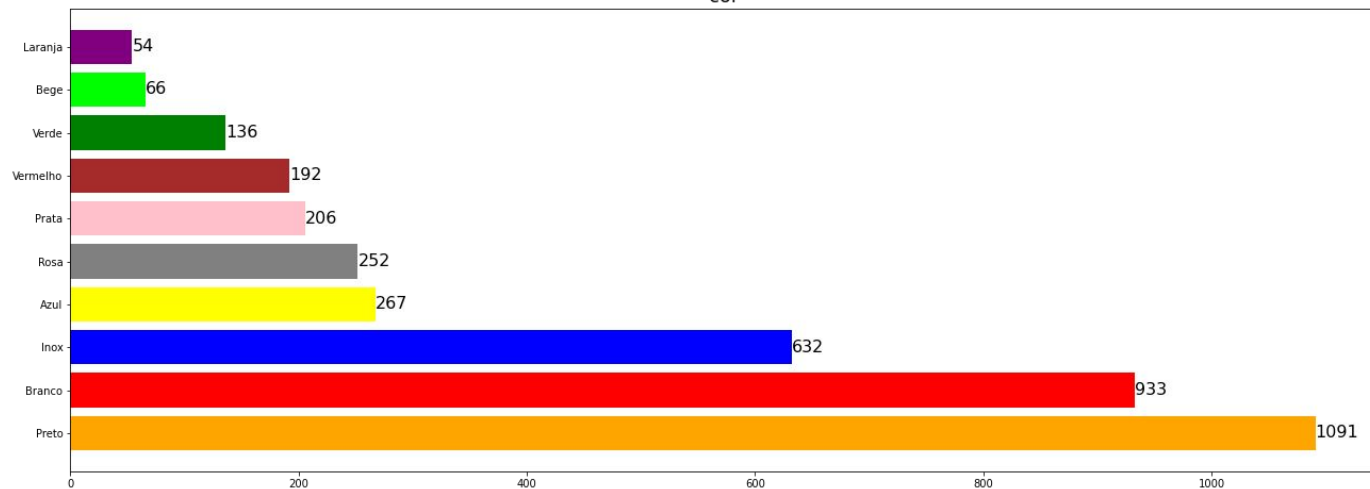
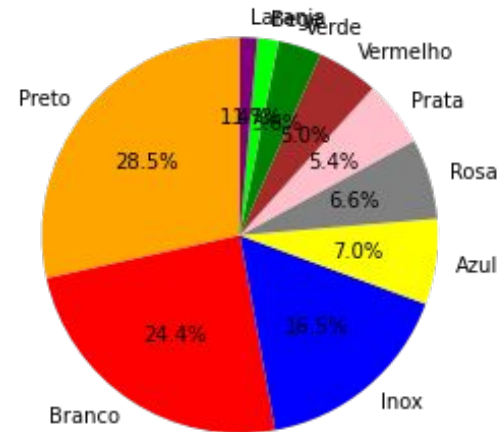
ANÁLISES RELACIONADAS A MARCA

+-----+-----+	
marca	qtde
+-----+-----+	
Tramontina	874
HP	526
Samsung	411
Electrolux	353
Philips	318
Kacyumara	290
LG	281
Sony	261
Lorenzetti	252
Mattel	213
+-----+-----+	



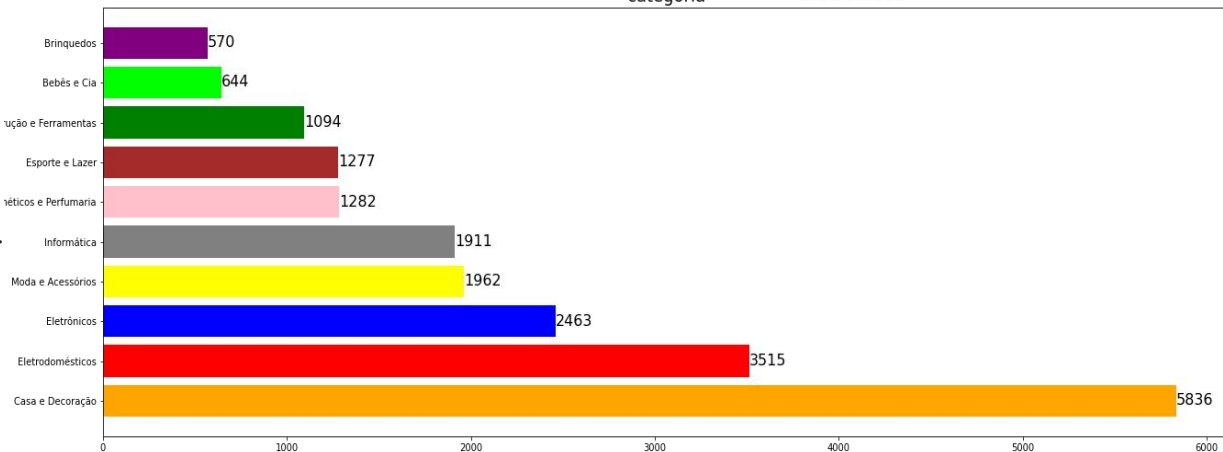
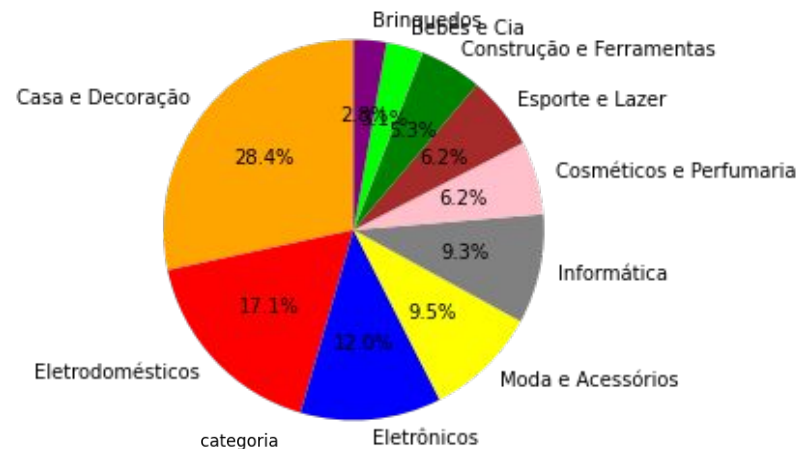
ANÁLISES RELACIONADAS A COR

cor	qtds
Preto	1091
Branco	933
Inox	632
Azul	267
Rosa	252
Prata	206
Vermelho	192
Verde	136
Bege	66
Laranja	54



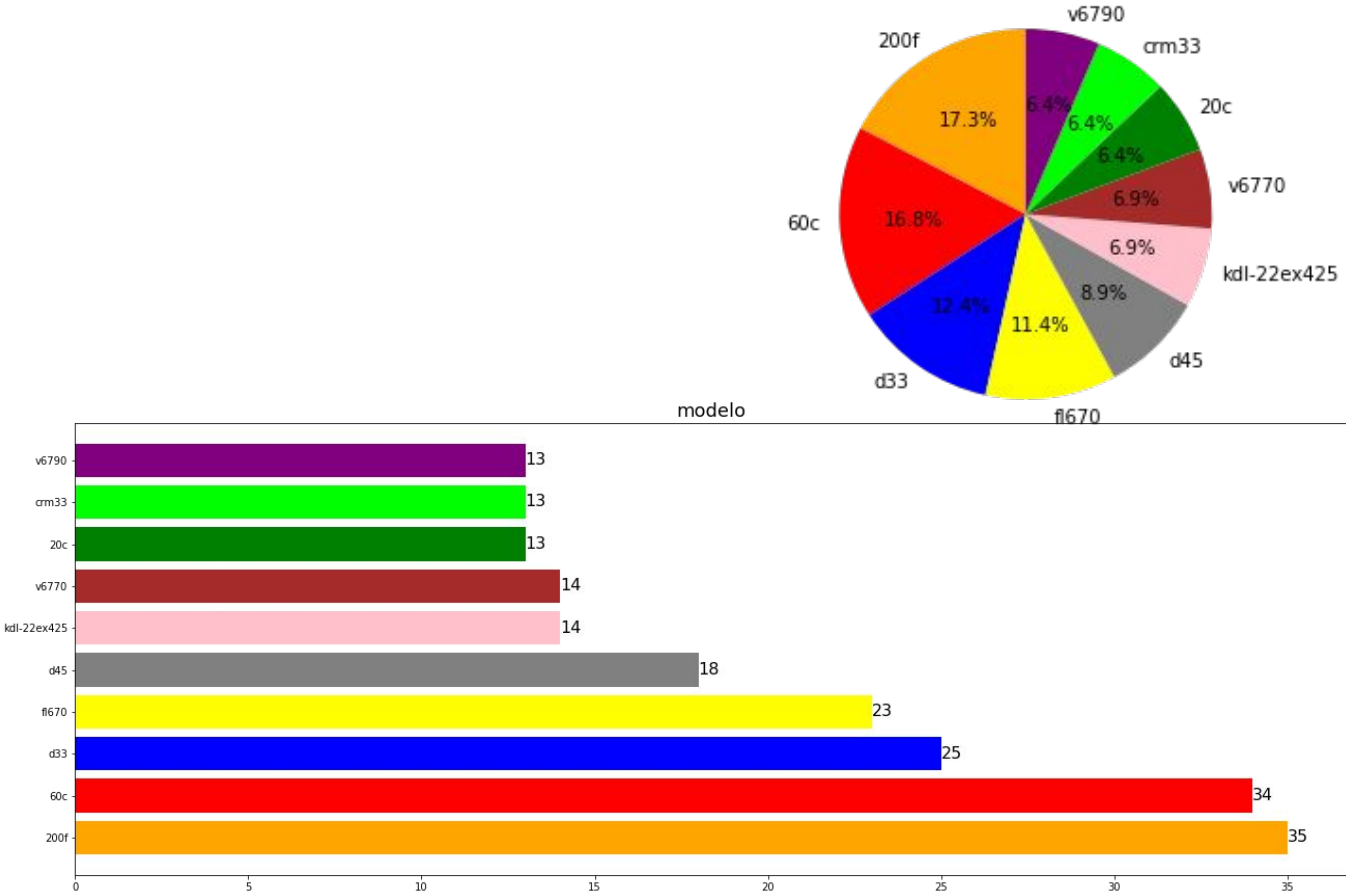
ANÁLISES RELACIONADAS A CATEGORIA

categoria	qtde
Casa e Decoração	5836
Eletrodomésticos	3515
Eletrônicos	2463
Moda e Acessórios	1962
Informática	1911
Cosméticos e Perfumaria	1282
Esporte e Lazer	1277
Construção e Ferramentas	1094
Bebês e Cia	644
Brinquedos	570



ANÁLISES RELACIONADAS AO MODELO

+-----+-----+	
modelo	qtds
+-----+-----+	
200f	35
60c	34
d33	25
f1670	23
d45	18
kdl-22ex425	14
v6770	14
20c	13
crm33	13
v6790	13
+-----+-----+	





Realizando uma Busca nos dados

FUNÇÃO PARA PESQUISAR NOS DADOS

- Para se fazer rápidas observações sobre os dados, foi criada uma função para se fazer eventuais pesquisas na base de dados, através do módulo sql do PySpark:

```
def busca(Categoria, Buscar, Total):  
    spark.sql("SELECT ROW_NUMBER() OVER (ORDER BY nome) AS n,* from df where lower("+Categoria+") LIKE  
    '%" + Buscar + "%' order by nome, tipo, marca, categoria, cor, modelo").show(Total, False)
```


```
#@title Buscar registro nos dados  
Categoria = 'modelo' #@param ["nome", "tipo", "marca", "categoria", "cor", "modelo"]  
Buscar = '200f' #@param {type:"string"}  
Buscar=str(Buscar).lower()  
Limite = 60 #@param {type:"slider", min:10, max:500, step:1}
```



```
busca(Categoria, Buscar, Limite)
```


FUNÇÃO PARA PESQUISAR NOS DADOS

- Através de uma interface fácil, essa função pode ser executada para buscar através das colunas disponíveis na base de dados.
- Também é possível especificar um limite de resultados que será exibido quando a função for executada.

Buscar registro nos dados

Categoria:  

Buscar:  

Limite: 

FUNÇÃO PARA PESQUISAR NOS DADOS

Parte do resultado de uma busca na coluna **Marca** pela marca **Brastemp**:

id	descricao	tipo	marca	categoria	cor	modelo
1	Ar Condicionado Split Brastemp BBF18 18000BTUS 220V	Ar-Condicionado	Brastemp	None	None	None
2	Adega de Vinhos Wine Cooler BZC12AE para 12 Garrafas All Black - Brastemp	Adega	Brastemp	Eletrodomésticos	Preto	ccx21db
3	Ar Condicionado 12000Btus BBV12 Split Unidade Interna Frio 220v - Brastemp	Ar-Condicionado	Brastemp	Eletrodomésticos	None	None
4	Ar Condicionado 12000Btus Split Clean BBZ12B Quente e Frio Unidade Externa 220v - Brastemp	Ar-Condicionado	Brastemp	Eletrodomésticos	None	None
5	Ar Condicionado 9000 BTUs Split BBF09 Frio Un. Interna 220v - Brastemp	Ar-Condicionado	Brastemp	Eletrodomésticos	None	None
6	Ar Condicionado Split 12000Btus BBy12 Unidade Externa Frio 220v - Brastemp	Ar-Condicionado	Brastemp	Eletrodomésticos	None	None
7	Ar Condicionado Split Clean CBV/BBY09BBBNA 9000 BTUs 220v - Brastemp	Ar-Condicionado	Brastemp	Eletrodomésticos	None	None
8	Ar-Condicionado Split Clean BBV09BBBNA/BBY09BBBNA Frio 9.000 BTUs 220V - Brastemp	Ar-Condicionado	Brastemp	Eletrodomésticos	None	bbv09bbbna
9	Aspirador de Pó / Água Brastemp Ative B7M16A4	Aspirador de Pó	Brastemp	Eletrodomésticos	None	ative-b7m16
10	Aspirador de Pó / Água Brastemp Clean B7B14A4	Aspirador de Pó	Brastemp	Eletrodomésticos	None	lse11
11	Aspirador de Pó Clean Com Filtro Hepa B7B14 - Brastemp	Aspirador de Pó	Brastemp	Eletrodomésticos	None	clean-b7b14
12	Aspirador pó brastemp ative-b7m16 220 vinho	Aspirador de Pó	Brastemp	Eletrodomésticos	None	fgct005psgda0br
13	Brastemp Allblack BAI60AE 60 cm Parede Aço Inox	Exaustor	Brastemp	Eletrodomésticos	Preto	None
14	Brastemp Ative B7D12 12 W	Aspirador de Pó	Brastemp	Eletrodomésticos	None	None

CONCLUSÃO

Através deste trabalho foi possível ver na prática, como funciona a instalação da ferramenta PySpark no Python através do google colab e algumas de suas funcionalidades para análise de dados, utilizando-se dos seus módulos para tratamento dos dados.

Sendo assim, foi possível aferir as quantidades dos itens de cada coluna individual e gerar para isso, um elemento visual que pudesse representar melhor a cada coluna extraída de uma pesquisa, gerando então, informação sobre a base de dados.

Obrigado pela atenção!

REFERÊNCIAS

<https://alyx.com.br/como-instalar-o-pyspark-no-seu-computador/>

<https://sparkbyexamples.com/pyspark/pyspark-read-csv-file-into-dataframe/>

https://www.w3schools.com/python/matplotlib_bars.asp

https://spark.apache.org/docs/latest/api/python/getting_started/quickstart_df.html#Viewing-Data