

Segue a correção do exercício, por algum motivo não me atentei para o final dos ultimos tópicos.

Exercício

Filtraremos os dados correspondentes ao valor 'GOLD', na coluna 'ticker',
Calcularemos o valor médio da coluna 'open' e o desvio padrão,
Calcularemos o valor médio da coluna 'low' e o desvio padrão.

Filtramos os dados pelo valor de ticker == 'GOLD'

```
df_gold = df.filter(df.ticker == 'GOLD')
```

Agrupamos por 'GOLD' e calculamos a média de 'open' e 'low'

```
df_both = df_gold.groupBy(df_gold.ticker == 'GOLD').avg('open','low')
```

Agrupamos por 'GOLD' e calculamos a média de 'open' e 'low', e já mostramos a tabela

```
df_both = df_gold.groupBy(df_gold.ticker == 'GOLD').avg('open','low').show()
```

Calculamos o desvio padrão para a coluna 'open' e 'low'

```
open_stddev = df_gold.select(stddev('open')).first()
low_stddev = df_gold.select(stddev('low')).first()
```

mostramos os valores de desvio padrão calculados

```
print(f'O desvio padrão relativo a coluna open é: {open_stddev[0]}')
print(f'O desvio padrão relativo a coluna low é: {low_stddev[0]}')
```

```
main.py x Prática
main.py > _
1 from pyspark.sql import SparkSession
2 from pyspark.sql.functions import stddev
3
4 spark = SparkSession.builder.appName("my_app").getOrCreate()
5
6 df = spark.read.option("delimiter", ",").csv("C:/Users/USER/Downloads/CODE/TOKIO/bigdata/modulo 5/EDA/stocks_2021.csv", header=True, inferSchema=True)
7
8 df_gold = df.filter(df.ticker == 'GOLD')
9
10 df_both = df_gold.groupBy(df_gold.ticker == 'GOLD').avg('open','low').show()
11
12
13 ## Calculamos o desvio padrão para a coluna open
14 open_stddev = df_gold.select(stddev('open')).first()
15
16 ## Calculamos o desvio padrão para a coluna low
17 low_stddev = df_gold.select(stddev('low')).first()
18
19 ## mostramos os valores de desvio padrão calculados
20 print(f'O desvio padrão relativo a coluna open é: {open_stddev[0]}')
21 print(f'O desvio padrão relativo a coluna low é: {low_stddev[0]}')
22
23
24
25
26
27
28
```

PROBLEMS OUTPUT DEBUG CONSOLE TERMINAL SQL CONSOLE Python + - + - + -

```
PS C:\Users\USER\Downloads\CODE\TOKIO\bigdata\modulo 5\EDA> & C:\Users\USER\AppData\Local\Microsoft\WindowsApps\python3.10.exe "c:/Users/USER/Downloads/CODE/TOKIO/bigdata/modulo 5/EDA/main.py"
23/08/21 16:10:05 WARN Shell: Did not find winutils.exe: java.io.FileNotFoundException: java.io.FileNotFoundException: HADOOP_HOME and hadoop.home.dir are unset. -see https://wiki.apache.org/hado
op/WindowsProblems
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).
23/08/21 16:10:06 WARN NativeCodeLoader: Unable to load native-heapoop library for your platform... using builtin-java classes where applicable
+-----+-----+-----+
+|(ticker = GOLD)|      avg(open)|      avg(low)|
+-----+-----+-----+
+|               |true|20.338983488223743|20.073286365783812|
+-----+-----+-----+

O desvio padrão relativo a coluna open é: 1.6754837208831312
O desvio padrão relativo a coluna low é: 1.64961849493862755
O desvio padrão relativo a coluna low é: 1.64961849493862755
SUCCESS: The process with PID 15824 (child process of PID 20528) has been terminated.
SUCCESS: The process with PID 20528 (child process of PID 26212) has been terminated.
```