Configurações iniciais

Instalação dos pacotes

```
1 !apt-get install openjdk-8-jdk-headless -qq > /dev/null
2 !wget -q http://archive.apache.org/dist/spark/spark-3.1.1/spark-3.1.1-bin-hadoop3.2.tgz
3 !tar xf spark-3.1.1-bin-hadoop3.2.tgz
4 !pip install -q findspark
```

Criando o ambiente virtual do PySpark

```
1 import os
2 os.environ["JAVA_HOME"] = "/usr/lib/jvm/java-8-openjdk-amd64"
3 os.environ["SPARK_HOME"] = "/content/spark-3.1.1-bin-hadoop3.2"
```

Criando a sessão PySpark

```
1 import findspark
2 findspark.init()
3 from pyspark.sql import SparkSession
4 spark = SparkSession.builder.master("local[*]").getOrCreate()
5 from pyspark.sql.functions import regexp_replace
6 spark.conf.set("spark.sql.repl.eagerEval.enabled", True) # Para deixar a visualização das tabelas mais amigável
7 spark
SparkSession - in-memory
    SparkContext
    Spark UI
    Version
         v3.1.1
    Master
         local[*]
    AppName
         pyspark-shell
```

2 - Lendo um arquivo a partir do disco

```
1 df = spark.read.csv('/content/Estatística Descritiva - Organização Gráfica.csv',header=True)
1 df.show()
```

ota	Escolaridade No	Salário (R\$)	Cargo	Departamento	Idade	País	Estado	Cidade	Nome	Matrícula
	Superior	25000	Gerente	Compras	35	Brasil	SP	Atibaia	Lucas	1
(Superior	12000	Coordenador	Vendas	29	Brasil	SP	São Paulo	Ana	2
9	MBA	28000	Gerente	Finanças	38	Brasil	SP	Santos	Luiza	3
-	Mestrado	40000	Diretor	Marketing	36	Brasil	SP	Atibaia	Fernando	4
	Superior	23000	Analista	Produção	28	Brasil	SP	Atibaia	Sandra	5
9	Superior	11000	Analista	Finanças	29	Brasil	SP	Bragança	Douglas	6
4	MBA	12000	Gerente	Marketing	30	Brasil	MG	Extrema	Eduardo	7
2	Superior	10000	Analista	Compras	29	Brasil	MG	Itapeva	Ester	8
-	Superior	13000	Analista	Marketing	30	Brasil	MG	Extrema	Pedro	9
	MBA	12000	Analista	Produção	40	Brasil	MG	Extrema	Maria	10

1 df.printSchema()

```
root
|-- Matrícula: string (nullable = true)
|-- Nome: string (nullable = true)
|-- Cidade: string (nullable = true)
|-- Estado: string (nullable = true)
|-- País: string (nullable = true)
|-- Idade: string (nullable = true)
|-- Departamento: string (nullable = true)
|-- Cargo: string (nullable = true)
|-- Salário (R$): string (nullable = true)
|-- Escolaridade: string (nullable = true)
|-- Nota : string (nullable = true)
```

Criando um arquivo a partir do zero

```
1 from pyspark.sql.types import *
 2 from pyspark.sql.types import StructType,StructField,StringType,IntegerType,DoubleType # importando estrutura da tabela (structType)
 3 schema = StructType([ \
              StructField('nome',StringType(),True), \
              StructField('idade',IntegerType(),True), \
StructField('altura',DoubleType(),True), \
 6
              StructField('peso', DoubleType(), True), \
 8 ])
 1
 2 '''
 3 Criação do objeto (nível Python) que contém as tuplas com os regitros que farão
 6 dados = [('Douglas',45,1.85,70.),
             ('Daniela',7,1.23,22.),
               ('Pedro',65,1.75,87.),
 9
               ('Maria',64,1.67,64.),
             ('Eduardo',37,1.82,96.),
              ('Ester',37,1.73,68.)
11
12 ]
 1 df = spark.createDataFrame(schema=schema, data=dados)
 1 df.show()
           nome|idade|altura|peso|
           |Douglas| 45| 1.85|70.0|
                                  7 | 1.23 | 22.0 |
65 | 1.75 | 87.0 |
64 | 1.67 | 64.0 |
           |Daniela|
               Pedro
                                 37 | 1.82 | 96.0 |
37 | 1.73 | 68.0 |
            Eduardo
              Ester
 1 df.printSchema()
              |-- nome: string (nullable = true)
              |-- idade: integer (nullable = true)
|-- altura: double (nullable = true)
              |-- peso: double (nullable = true)
     Filtro com operadores lógicos AND (&) OR (|)
 1 filtro01 = df.filter((df['idade']>=40)&(df['peso']>=50)) # em pyspark o operador lógico and deve ser usado pelo símbolo &
 2 filtro01.show()
           | nome|idade|altura|peso|
          |Douglas| 45| 1.85|70.0|
| Pedro| 65| 1.75|87.0|
                 Maria 64 1.67 64.0
  1 \; filtro02 = df.filter((df['idade']>=40)|(df['peso']>=50)) \; \# \; em \; pyspark \; o \; operador \; l\'ogico \; ou \; deve \; ser \; usado \; pelo \; símbolo \; | \; l'allore | l'allore
 2 filtro02.show()
           nome idade altura peso
                                 45 | 1.85 | 70.0 |
65 | 1.75 | 87.0 |
           Douglas
                Pedro
                Marial
                                  64 | 1.67 | 64.0 |
37 | 1.82 | 96.0 |
            Eduardo
           Ester
                                    37 | 1.73 | 68.0 |
 1 filtro03 = df.filter('idade >= 40 AND peso >= 50')
 2 filtro03.show()
          +-----
           | nome|idade|altura|peso|
            |Douglas| 45| 1.85|70.0|
           | Pedro| 65| 1.75|87.0|
| Maria| 64| 1.67|64.0|
```

```
2 filtro04.show()
→ +----+
      | nome|idade|altura|peso|
      Douglas
                 45 1.85 70.0
       Douglas | 45 | 1.63 | 70.0 |
Pedro | 65 | 1.75 | 87.0 |
Maria | 64 | 1.67 | 64.0 |
Eduardo | 37 | 1.82 | 96.0 |
Ester | 37 | 1.73 | 68.0 |
      | Eduardo |
      Ester
  1 \; \text{filtro05 = df.filter($\sim$(df['nome'] == 'Douglas')$) \# Filtro com operador l\'ogico simb\'olico de negação } 
 2 filtro05.show()
→ +----+
      | nome|idade|altura|peso|
      Daniela
                   7 | 1.23 | 22.0 |
                  65 | 1.75 | 87.0 |
64 | 1.67 | 64.0 |
37 | 1.82 | 96.0 |
        Pedro
        Maria
      Eduardo
      Ester 37 1.73 68.0

    Filtro com funções

 1 from pyspark.sql.functions import *
 1 filtro06 = df.filter(col('nome').startswith('D'))
 2 filtro06.show()
      | nome|idade|altura|peso|
      |Douglas| 45| 1.85|70.0|
|Daniela| 7| 1.23|22.0|
 1 # selecionando coluna única e armazenando em um df spark
 2 idade = df.select('idade')
 3 idade.show()
      idade
          45
          65
          64
          37
          37
 1 \# selecionando mais de uma coluna e as armazenando em um df spark
 2 filtro07 = df.select('altura','peso')
 3 filtro07.show()
      |altura|peso|
        1.85 70.0
         1.23 22.0
        1.75 87.0
        1.67 64.0
        1.82 96.0
        1.73 68.0
 1 # Renomear a coluna peso para massa
 2 df = df.withColumnRenamed('peso', 'massa')
 3 df.show()
      | nome|idade|altura|massa|
      Douglas
                  45 | 1.85 | 70.0 |
                  7 | 1.23 | 22.0 |
65 | 1.75 | 87.0 |
64 | 1.67 | 64.0 |
37 | 1.82 | 96.0 |
      |Daniela|
        Pedro
         Maria
      Eduardo
                  37 | 1.73 | 68.0
       Ester
```

```
2 df = df.withColumn('imc',df['massa']/df['altura']**2)
2 Agora imagine que você deseja escolher uma determina linha por seu pseudo-índice
4 tio = df.collect()[0]
1 print(tio)

→ Row(nome='Douglas', idade=45, altura=1.85, massa=70.0, imc=20.45288531775018)

1 '''
2 Agora imagine que você deseja extrair e armazena apenas o valor 'Douglas'
4 doug = df.collect()[0][0]
1 print(doug)
→ Douglas
1 type(doug)
→ str
1 elemento = [('Tobias',18,1.77,80.,25.535446)]
1 nova_linha = spark.createDataFrame(elemento)
1 nova_linha.show()
      _1| _2| _3| _4| _5|
     |Tobias| 18|1.77|80.0|25.535446|
1 df = df.union(nova_linha)
2 df.show()
     | nome|idade|altura|massa|
                 45| 1.85| 70.0| 20.45288531775018|
     Douglas
                 7 | 1.23 | 22.0 | 14.541608830722454
65 | 1.75 | 87.0 | 28.408163265306122
64 | 1.67 | 64.0 | 22.948115744558788
37 | 1.82 | 96.0 | 28.982007003985025
     |Daniela|
        Pedrol
        Maria
      Eduardo
      Ester | 37 | 1.73 | 68.0 | 22.720438370810918
Tobias | 18 | 1.77 | 80.0 | 25.535446
                                              25.535446
2 Agora imagine que você deseja calcular as principais medidas descritivas do df
4 df.describe().show()
                                      idade
     |summary| nome|
       count
                  null
                                        39.0 | 1.6885714285714286 | 69.57142857142857 | 23.369809219019068 |
         mean
                  null 21.641010450839243 0.21058874387682508 23.831751529259886 4.976893829941699
       stddev
                                                               1.23
                                                                                     22.0 14.541608830722454
         max Tobias
                                          65
                                                               1.85
                                                                                     96.0 28.982007003985025
1 '''
2 Agora imagine que você deseja criar uma nova coluna com os nomes só com letras
3 em caixa baixa
4 '''
5 df = df.withColumn('minu',lower(col('nome')))
6 df.show()
     nome|idade|altura|massa|
                                                    imc minu
     Douglas
                 45| 1.85| 70.0| 20.45288531775018|douglas|
                 7| 1.33 | 22.0 | 14.541608830722454 | daniela
65 | 1.75 | 87.0 | 28.408163265306122 | pedro
64 | 1.67 | 64.0 | 22.948115744558788 | maria
     |Daniela|
        Pedro
        Maria
      Eduardo
                 37 | 1.82 | 96.0 | 28.982007003985025 | eduardo |
                37 | 1.73 | 68.0 | 22.720438370810918 | ester | 18 | 1.77 | 80.0 | 25.535446 | tobias |
       Ester
       Tobias
```

1 # Criando uma nova coluna

```
1 '''
 2 Agora imagine que você deseja criar uma nova coluna com os nomes só com letras
  3 em caixa alta
 4 ' ' '
5 df = df.withColumn('maiu',upper(col('nome')))
6 df.show()
                 | nome|idade|altura|massa| imc| minu| maiu|
                  Douglas
                                                            45| 1.85| 70.0| 20.45288531775018|douglas|DOUGLAS|
                   |Daniela|
                                                                  7 |
                                                                               1.23 | 22.0 | 14.541608830722454 | daniela | DANIELA
                                                              Pedro
                                                                                                                                                                                                                                                PEDRO
                            Maria
                     Eduardo
                                                               37 | 1.82 | 96.0 | 28.982007003985025 | eduardo | EDUARDO
                         Ester
                                                               37 | 1.73 | 68.0 | 22.720438370810918 | ester | ESTER |
                                                         18 | 1.77 | 80.0 | 25.535446 | tobias | TOBIAS |
                   | Tohias|
 1 '''
  2 Medidas agregadas por coluna
  3 Exemplo descobrir o menor valor da coluna idade
 4 ''
5 idade_min = df.agg({'idade':'min'})  # como resultado teremos um df spark com uma linha e uma coluna
6 print(idade_min)
                  |min(idade)|
                  1 71
 1 '''
 2 Medidas agregadas por coluna
  3 Exemplo descobrir o maior valor da coluna idade
 4 ' '
 \texttt{5} \; \texttt{idade\_max} \; = \; \texttt{df.agg(\{'idade':'max'\})} \; \; \texttt{\#} \; \texttt{como} \; \; \texttt{resultado} \; \; \texttt{teremos} \; \; \texttt{um} \; \; \texttt{df} \; \texttt{spark} \; \; \texttt{com} \; \; \texttt{uma} \; \; \texttt{linha} \; \; \texttt{e} \; \; \texttt{uma} \; \; \texttt{columna} \; \; \texttt{columna} \; \; \texttt{linha} \; \; \texttt{e} \; \; \texttt{uma} \; \; \texttt{columna} \; \; \texttt{linha} \; \; \texttt{e} \; \; \texttt{uma} \; \; \texttt{columna} \; \; \texttt{linha} \; \; \texttt{e} \; \; \texttt{uma} \; \; \texttt{columna} \; \; \texttt{linha} \; \; \texttt{e} \; \; \texttt{uma} \; \; \texttt{columna} \; \; \texttt{linha} \; \; \texttt{e} \; \; \texttt{uma} \; \; \texttt{columna} \; \; \texttt{linha} \; \; \texttt{e} \; \; \texttt{uma} \; \; \texttt{linha} \; \; \texttt{e} \; \; \texttt{uma} \; \; \texttt{linha} \; \; \texttt{e} \; \; \texttt{uma} \; \; \texttt{linha} \; \; \texttt{e} \; \; \texttt{uma} \; \; \texttt{linha} \; \; \texttt{e} \; \; \texttt{uma} \; \; \texttt{linha} \;
6 print(idade_max)
                  |max(idade)|
                                            65
1 '''
 2 Medidas agregadas por coluna
 3 Exemplo descobrir o maior valor da coluna idade
 4 '
 5 \; idade\_media = df.agg(\{'idade':'avg'\}) \; \# \; como \; resultado \; teremos \; um \; df \; spark \; com \; uma \; linha \; e \; uma \; coluna \; description | linha |
 6 print(idade_media)
                +----
\exists \forall
                  |avg(idade)|
                                         39.0
2 Agora vamos supor que você deseja ordenar de forma crescente por nome
4 df.orderBy('nome')
                                                idade altura massa
                        nome
                                                                                                                                                      imc
                                                                                                                                                                                                      minu
                                                                                                                                                                                                                                        maiu
                                                                                                  22.0
                                                                                                                        14.541608830722454
                                                                                                                                                                                                  daniela
                                                                                                                                                                                                                                DANIELA
                                                                                                                        20.45288531775018
                   Douglas 45
                                                                        1 85
                                                                                                                                                                                                  douglas
                                                                                                                                                                                                                               DOUGLAS
                   Eduardo 37
                                                                        1.82
                                                                                                 96.0
                                                                                                                        28.982007003985025
                                                                                                                                                                                                  eduardo EDUARDO
                   Ester
                                                                        1.73
                                                                                                 68.0
                                                                                                                        22.720438370810918 ester
                                                                                                                                                                                                                                ESTER
                   Maria
                                                  64
                                                                        1.67
                                                                                                 64.0
                                                                                                                        22.948115744558788 maria
                                                65
                                                                        1.75
                                                                                                87.0
                                                                                                                       28.408163265306122 pedro
                   Tobias
                                                                                                                       25.535446
                                                                                                                                                                                                 tobias
                                                 18
                                                                        1.77
                                                                                                80.0
```

1 '''
2 Agora vamos supor que você deseja ordenar de forma decrescente por nome
3 '''
4 df.orderBy(df['nome'].desc())

 \equiv

```
nome idade altura massa
                                               minu
                                                        maiu
Tobias
                    0.08
                          25.535446
                                                      TOBIAS
Pedro
              1.75
                    87.0
                          28.408163265306122
                                              pedro
                                                     PEDRO
Maria
       64
              1.67
                    64.0
                          22.948115744558788 maria
                                                     MARIA
                    0.86
Ester
Douglas 45
              1 85
                          20.45288531775018
                                              douglas
                                                      DOLIGI AS
                          14.541608830722454 daniela DANIELA
Daniela 7
             1 23
                   22 0
```

```
1 '''
2 Agora imagine que você deseja agrupar por idade
3 '''
4 df.groupBy('idade').agg({'idade':'count'})
```



idade	count(idade)
65	1
7	1
37	2
18	1
64	1
45	1

1 df = spark.read.csv('/content/Estatística Descritiva - Organização Gráfica.csv',header=True)

1 df.show()

 \rightarrow |Matrícula| Nome | Cidade | Estado | País | Idade | Departamento | Cargo|Salário (R\$)|Escolaridade|Nota | Lucas Atibaia SP|Brasil Compras Gerente 25000 Superior Ana São Paulo 2 SP|Brasil 29 Vendas Coordenador 12000 Superior Luiza Santos SP|Brasil 38 Finanças Gerente 28000 MBA 3 Atibaia 4 Fernando SP Brasil Marketing Diretor 40000 Mestrado 5 Sandra Atibaia SP Brasil 28 Produção Analista 23000 Superior 6 Douglas Bragança SP|Brasil 29 Financas Analistal 11000 Superior 7 | Eduardol 30 Marketing 12000 MBA Extrema MG Brasil Gerente 29 Compras Analista 10000 Superior Ester MG|Brasil 8 Itapeva 13000 Pedro Extrema MG|Brasil Marketing Analista Superior 10| Maria| Extrema| MG|Brasil| 40 Produção Analista 12000 MBA

1 df.printSchema()

```
root

-- Matrícula: string (nullable = true)
-- Nome: string (nullable = true)
-- Cidade: string (nullable = true)
-- Estado: string (nullable = true)
-- País: string (nullable = true)
-- Idade: string (nullable = true)
-- Departamento: string (nullable = true)
-- Cargo: string (nullable = true)
-- Salário (R$): string (nullable = true)
-- Escolaridade: string (nullable = true)
-- Nota : string (nullable = true)
```

```
1 '''
 2 Neste bloco vamos converter idade em inteiro, matrículaem inteiro, salári em
 3 double e nota em Double
 5 from pyspark.sql.types import *
 6 from pyspark.sql.types import StructType,StructField,StringType,IntegerType,DoubleType
   .withColumn("Matrícula" ,
               df["Matrícula"]
                .cast(IntegerType())) \
10
11
    .withColumn("Idade",
               df["Idade"]
12
                .cast(IntegerType()))
13
    .withColumn("Salário (R$)"
14
               df["Salário (R$)"]
15
16
                .cast(DoubleType()))
17
    .withColumn("Nota "
               df["Nota "]
18
19
                .cast(DoubleType()))
20 df.printSchema()
→ root
      |-- Matrícula: integer (nullable = true)
```

```
root
|-- Matricula: integer (nullable = true)
|-- Nome: string (nullable = true)
|-- Cidade: string (nullable = true)
|-- Estado: string (nullable = true)
|-- País: string (nullable = true)
|-- Idade: integer (nullable = true)
|-- Departamento: string (nullable = true)
|-- Cargo: string (nullable = true)
```

```
|-- Salário (R$): double (nullable = true)
       -- Escolaridade: string (nullable = true)
      |-- Nota : double (nullable = true)
1 df.show()
     | Matrícula | Nome | Cidade | Estado | País | Idade | Departamento |
                                                                         Cargo|Salário (R$)|Escolaridade|Nota|
                 Lucas Atibaia
                                       SP|Brasil|
                                                                       Gerente
                                                                                     25000.0
                                                                                                  Superior
              2
                    Ana|São Paulo|
                                       SP Brasil
                                                    29
                                                             Vendas Coordenador
                                                                                     12000.0
                                                                                                             6.0
              3 |
                  Luizal
                          Santos
                                       SP Brasil
                                                    38
                                                          Financas
                                                                       Gerente
                                                                                     28000.0
                                                                                                     MBA
                                                                                                             9.0
                          Atibaia
                                                                                     40000.0
                                                                                                  Mestrado
              4 | Fernando |
                                       SPIBrasil
                                                    361
                                                         Marketing
                                                                        Diretor
                                                                                                             7.0
                                                                                     23000.0
              5
                 Sandra
                           Atibaia
                                       SP Brasil
                                                    28
                                                          Produção
                                                                       Analista
                                                                                                  Superior
                                                                                                             5.0
              6 Douglas
                         Bragança
                                       SP|Brasil
                                                    29
                                                           Finanças
                                                                       Analista|
                                                                                     11000.0
                                                                                                  Superior
                                                                                                             9.0
              7 Eduardo
                           Extrema
                                       MG|Brasil
                                                    30
                                                          Marketing
                                                                        Gerente
                                                                                     12000.0
                                                                                                      MRΔ
                                                                                                            4.0
                                                    29
                                                                                     10000.0
                                                                                                  Superior
              8
                  Ester
                           Itapeva
                                       MG Brasil
                                                           Compras
                                                                       Analista
                                                                                                            2.0
                                       MG Brasil
                                                    30 İ
                                                          Marketing
                                                                                     13000.0
                                                                                                  Superior
                   Pedro
                                                                       Analista
                                                                                                             1.0
                           Extrema
                  Maria
                          Extrema
                                       MG|Brasil|
                                                    40
                                                           Produção
                                                                       Analista
                                                                                     12000.0
                                                                                                     MBA
1 '''
2 Agora você vai aplicar seus conhecimentos de Pyspark para para as seguintes
 3 transformações no DF:
 5 1 - Renomear todas as colunas de forma que se mantenha os nomes, mas apenas com
 6 letras minúsculas sem espaços, e/ou caracteres especiais
 8 2 - Fazer livremente no mínimo 10 filtros que tragam recortes a partir de colunas
 9 diferentes
10
11 3 - Considerando que todos os funcionários tem um desconto de previdência complementar
12 de 10 % do salário Bruto crie e calcule os valores de uma nova coluna chamada
13 previdencia.
14
15 4 - Considerando que todos os funcionários nesse mês epecífico terão uma bonificaccao
16 de 20% sobre o salário bruto, crie e calcule os valores de uma nova coluna chamada
17 bonus
18
19 5 - Crie uma nova coluna com todos os nomes dos funcionários em letras maiúsculas
20
21 6 - Crie uma nova coluna com os nomes dos funcionários todos em letras minúsculas
22
23 7 - Crie um novo dataframe chamado quant apenas com as colunas idade,
24 salário e notas
26 8 - calcule as principais medidas descritivas do df quant (describe)
28 9 - Ordene o df original por notas de forma crescente
30 10 - faça as contagens de frequências, usando groupby para as seguintes colunas:
31 a) Cargo
32 b) Cidade
33 c) Estado
34 d) Escolaridade
35 e) Departamento
36 '
1 df.printSchema()
→ root
      |-- Matrícula: integer (nullable = true)
       -- Nome: string (nullable = true)
       -- Cidade: string (nullable = true)
       |-- Estado: string (nullable = true)
       -- País: string (nullable = true)
       -- Idade: integer (nullable = true)
       -- Departamento: string (nullable = true)
       -- Cargo: string (nullable = true)
      |-- Salário (R$): double (nullable = true)
|-- Escolaridade: string (nullable = true)
       -- Nota : double (nullable = true)
 1 ...
 21 - Renomear todas as colunas de forma que se mantenha os nomes, mas apenas com
 3 letras minúsculas sem espaços, e/ou caracteres especiais
 5 print("atual: ", df.columns)
 6 df = df.toDF(*('matricula','nome','cidade','estado','pais','idade',
                  'departamento', 'cargo', 'salario', 'escolaridade', 'nota'))
 8 df.show()
🚁 atual: ['Matrícula', 'Nome', 'Cidade', 'Estado', 'País', 'Idade', 'Departamento', 'Cargo', 'Salário (R$)', 'Escolaridade', 'Nota ']
```

cargo salario escolaridade nota

Superior 8.0

Superior | 6.0

MBA 9.0

Compras | Gerente 25000.0

Vendas | Coordenador | 12000.0 |

Gerente 28000.0

25

|matricula| nome | cidade estado | pais | idade | departamento |

SP|Brasil| 35|

SP|Brasil|

SP|Brasil

29

Finanças

38

1 Lucas Atibaia

Ana|São Paulo|

Luiza Santos

```
Eduardo
                       Extrema
                                                 Marketing
                                                             Gerente 12000.0
                                                                               Superior 2.0
                                                   Compras
                                                             Analista 10000.0
           8
               Ester
                       Itapeva
                                 MG Brasil
                                            29
                                            30
                                                 Marketing
                                                             Analista 13000.0
                                                                               Superior | 1.0
           9
               Pedro
                       Extrema
                                 MG Brasil
                                                            Analista 12000.0
           10| Maria| Extrema|
                                 MG|Brasil|
                                                 Produção
                                                                                MBA 7.0
1 '''
2 2 - Fazer livremente no mínimo 10 filtros que tragam recortes a partir de colunas
3 diferentes
5 from pyspark.sql.functions import col
6 f1 = df.filter(df['cidade']=='Atibaia')
7 f1.show()
8 f2 = df.filter(df['idade']>=40)
9 f2.show()
10 f3 = df.filter("idade >= 40 AND salario >= 35000")
11 f3.show()
12 f4 = df.filter("idade >= 40 OR salario >= 35000")
13 f4.show()
14 f5 = df.filter(col('nome').startswith('D'))
15 f5.show()
16 f6 = df.select('idade')
17 f6.show()
18 f7 = df.select('cargo')
19 f7.show()
20 f8 = df.select('cidade','estado')
21 f8.show()
22 f9 = df.select('cidade','estado')
23 f9.show()
24 f10 = df.select('escolaridade','salario')
25 f10.show()
    |matricula| nome| cidade|estado| pais|idade|departamento| cargo|salario|escolaridade|nota|
           1 Lucas Atibaia SP Brasil 35
                                                Compras | Gerente 25000.0
           4 Fernando Atibaia
                               SP|Brasil|
                                           36
                                                Marketing | Diretor | 40000.0 |
                                                                          Mestrado
                             SP|Brasil| 28|
           5 | Sandra Atibaia
                                              Produção Analista 23000.0
                                                                         Superior 5.0
    |matricula| nome| cidade|estado| pais|idade|departamento| cargo|salario|escolaridade|nota|
             10|Maria|Extrema| MG|Brasil| 40| Produção|Analista|12000.0|
    |matricula|nome|cidade|estado|pais|idade|departamento|cargo|salario|escolaridade|nota|
    +-----
    |matricula| nome| cidade|estado| pais|idade|departamento| cargo|salario|escolaridade|nota|
                             SP|Brasil| 36| Marketing|Diretor|40000.0|
MG|Brasil| 40| Produção|Analista|12000.0|
           4|Fernando|Atibaia|
          10| Maria|Extrema|
                                                                             MBA 7.0
     |matricula| nome| cidade|estado| pais|idade|departamento| cargo|salario|escolaridade|nota|
    | 6|Douglas|Bragança| SP|Brasil| 29| Finanças|Analista|11000.0| Superior| 9.0|
    lidadel
       35
       29
       38
       36
       28
       29
       30
       29
       30
       40
         cargo
     Coordenador
        Gerente
        Diretor
       Analista
       Analista
        Gerente
       Analista
```

4|Fernando| Atibaia|

Atibaia

Bragança

Sandra

6| Douglas|

SP|Brasil|

SP Brasil

SP|Brasil|

MG|Brasil|

Marketing

Produção

Finanças

36

28

29

30

Diretor 40000.0

Analista 23000.0

Analista 11000.0

Mestradol 7.01

Superior 5.0 Superior 9.0

MBA 4.0

```
3 de 10 % do salário Bruto crie e calcule os valores de uma nova coluna chamada
4 previdencia.
6 df = df.withColumn('previdencia',df['salario']*0.1)
7 df.show()
    |matricula| nome| cidade|estado| pais|idade|departamento|
                                                                   cargo|salario|escolaridade|nota|previdencia|
            11
               Lucas | Atihaia
                                   SP[Rrasil]
                                                35
                                                       Compras
                                                                 Gerente 25000 0
                                                                                     Superior 8 0
                                                                                                       2500 01
            2
                 AnalSão Paulol
                                   SPIBrasil
                                                29
                                                       Vendas | Coordenador | 12000.0|
                                                                                     Superior 6.0
                                                                                                       1200.0
                Luiza
                        Santos
                                   SP Brasil
                                                38
                                                      Finanças
                                                                  Gerente 28000.0
                                                                                         MBA 9.0
                                                                                                       2800.0
            3
            4 Fernando
                        Atibaia
                                                                   Diretor 40000.0
                                                                                     Mestrado 7.0
                                   SP Brasil
                                                36
                                                     Marketing
                                                                                                       4000.0
               Sandral
                       Atibaia
                                   SP|Brasil
                                                     Produção
                                                                  Analista 23000.0
                                                                                     Superior 5.0
                                                                                                       2300.0
                                                28
            6| Douglas| Bragança|
7| Eduardo| Extrema|
                                   SP[Brasil]
                                                291
                                                     Finanças
                                                                 Analista | 11000.0 |
                                                                                     Superior | 9.0|
                                                                                                       1100.0
                                   MG Brasil
                                                30
                                                     Marketing
                                                                  Gerente 12000.0
                                                                                        MBA 4.0
                                                                                                       1200.0
                Ester
                                                       Compras
                                                                 Analista 10000.0
                                                                                     Superior 2.0
                                                                                                       1000.0
            8
                        Itapeva
                                   MG|Brasil|
                                                29
                                                30
                                                     Marketing
                                                                  Analista 13000.0
                Pedro
                                   MG Brasil
                                                                                     Superior 1.0
                                                                                                       1300.0
                                               40
           10
                Maria
                       Extrema
                                   MG|Brasil
                                                      Produção
                                                                 Analista 12000.0
                                                                                         MBA 7.0
                                                                                                       1200.0
1 '''
2 4 - Considerando que todos os funcionários nesse mês epecífico terão uma bonificaccao
3 de 20% sobre o salário bruto, crie e calcule os valores de uma nova coluna chamada
4 honus
6 df = df.withColumn('bonus',df['salario']*0.2)
7 df.show()
    |matricula| nome| cidade|estado| pais|idade|departamento|
                                                                   cargo|salario|escolaridade|nota|previdencia| bonus|
                                                                  Gerente 25000.0
                                                                                      Superior 8.0
               Lucas | Atibaia
                                   SP Brasil
                                                                                                       2500.0[5000.0]
                                                                                     Superior 6.0
            2
                  AnalSão Paulol
                                   SP|Brasil|
                                                29
                                                        Vendas | Coordenador | 12000.0 |
                                                                                                       1200 0 2400 0
                Luiza
                                                     Finanças
                                   SP|Brasil|
                                                                  Gerente 28000.0
                                                                                        MBA 9.0
                                                                                                       2800.0 | 5600.0
            3
                        Santos
                                                38
                       Atibaia
                                   SP Brasil
                                                     Marketing
                                                                  Diretor 40000.0
                                                                                     Mestrado 7.0
                                                                                                       4000.0 8000.0
            4 Fernando
                                                36
                                                                  Analista 23000.0
                        Atibaia
                                   SP|Brasil|
                                                28
                                                     Produção
                                                                                     Superior | 5.0|
                                                                                                       2300.0 4600.0
            6
              Douglas
                       Bragança
                                   SP Brasil
                                                29
                                                      Finanças
                                                                  Analista 11000.0
                                                                                     Superior 9.0
                                                                                                       1100.0 2200.0
            7
              Eduardo
                        Extrema
                                   MG|Brasil|
                                                301
                                                     Marketing
                                                                  Gerente | 12000.0|
                                                                                         MBA 4.0
                                                                                                       1200.0 2400.0
                                                                                     Superior 2.0
                                                29
                                                      Compras
                                                                  Analista 10000.0
                                                                                                       1000.0 2000.0
            8
               Ester
                        Itapeva
                                   MG Brasil
                Pedro
                                                     Marketing|
                                                                 Analista 13000.0
                                                                                     Superior 1.0
                                                                                                       1300.0 2600.0
                                   MG|Brasil
                                                30
                        Extrema
                                                                                         MBA 7.0
               Maria
                                   MG Brasil
                                                40
                                                      Produção
                                                                  Analista 12000.0
                                                                                                       1200.0 2400.0
2 5 - Crie uma nova coluna com todos os nomes dos funcionários em letras maiúsculas
4 from pyspark.sql.functions import *
5 df = df.withColumn('maiu',upper(col('nome')))
6 df.show()
    |matricula| nome| cidade|estado| pais|idade|departamento|
                                                                   cargo|salario|escolaridade|nota|previdencia| bonus| maiu|
               Lucas Atibaia
                                                                  Gerente 25000.0
                                   SP Brasil
                                                                                      Superior 8.0
                                                                                                       2500.0 5000.0 LUCAS
                                                                                     Superior 6.0
            2
                  Ana São Paulo
                                   SP|Brasil|
                                               29
                                                        Vendas | Coordenador | 12000.0 |
                                                                                                       1200.0 2400.0
                                                                                                                        ΔΝΔ
            3 [
                Luizal
                       Santos
                                   SPI Brasill
                                               38
                                                     Financas
                                                                  Gerente | 28000.0|
                                                                                         MRAL 9.01
                                                                                                       2800.0[5600.0]
                                                                                                                       Ι ΠΤΖΔ
                                                                                     Mestrado 7.0
                                                                                                       4000.0 8000.0 FERNANDO
                       Atibaia
                                                     Marketing
                                                                  Diretor 40000.0
            4|Fernando|
                                   SPIBrasil
                                               36
                        Atibaia
               Sandra
                                   SP Brasil
                                               28
                                                     Produção
                                                                  Analista 23000.0
                                                                                     Superior 5.0
                                                                                                       2300.0 4600.0 SANDRA
            5
              Douglas
                       Bragança
                                   SP Brasil
                                                29
                                                      Finanças
                                                                  Analista 11000.0
                                                                                     Superior 9.0
                                                                                                       1100.0 2200.0 DOUGLAS
            71
              Eduardo
                        Extrema
                                   MG|Brasil
                                                30
                                                     Marketing
                                                                  Gerente 12000.0
                                                                                         MBA 4.0
                                                                                                       1200.0 2400.0 EDUARDO
            8
                Ester
                        Itapeval
                                   MG|Brasil|
                                               29
                                                      Compras
                                                                  Analista 10000.0
                                                                                     Superior 2.0
                                                                                                       1000.0 2000.0
                                                                                                                       ESTER
                                                     Marketing
                                                                 Analista 13000.0
                                                                                                       1300.0 2600.0
                                   MG Brasil
                                                30
                                                                                     Superior | 1.0|
                                                                                                                       PEDRO
            9
                Pedro
                        Extrema
                Maria
                        Extrema
                                   MG Brasil
                                                      Produção
                                                                 Analista 12000.0
                                                                                        MBA 7.0
                                                                                                       1200.0 2400.0
                                                                                                                       MARIA
1 '''
2 6 - Crie uma nova coluna com os nomes dos funcionários todos em letras minúsculas
4 df = df.withColumn('minu',lower(col('nome')))
5 df.show()
   |matricula| nome| cidade|estado| pais|idade|departamento|
                                                                    cargo|salario|escolaridade|nota|previdencia| bonus| maiu| minu|
            1 Lucas Atibaia
                                   SP|Brasil
                                                     Compras
                                                                   Gerente 25000.0
                                                                                     Superior 8.0
                                                                                                       2500.0|5000.0| LUCAS|
                                   SP Brasil
                                                        Vendas | Coordenador | 12000.0 |
                                                                                                       1200.0 2400.0
            2
                  Ana São Paulo
                                                29
                                                                                     Superior 6.0
                                                                                                                         ANA
                                                                                                                                 ana
               Luizal
                                                                                                                       LUIZAİ
                                                                                                                                luiza
            3 |
                        Santos
                                   SP|Brasil|
                                               38
                                                     Financas
                                                                   Gerente 28000.0
                                                                                         MBA | 9.0
                                                                                                       2800.0 | 5600.0 |
                                                                  Diretor 40000.01
                                                                                     Mestradol 7.01
                                                                                                       4000.018000.01FERNANDO1fernando
            4|Fernando|
                       Atibaia
                                   SP|Brasil
                                                36
                                                     Marketing
                                                                  Analista 23000.0
                                                                                                       2300.0 4600.0
                                                                                                                      SANDRA
            5
               Sandra
                       Atibaia
                                   SP Brasil
                                                28
                                                      Produção
                                                                                     Superior 5.0
                                                                                                                               sandra
            6 Douglas
                       Bragança
                                                29
                                                                  Analista 11000.0
                                                                                                       1100.0 2200.0 DOUGLAS
                                   SP|Brasil|
                                                     Financas
                                                                                     Superior | 9.0
                                                                                                                              douglas
            7 Eduardo
                                   MG Brasil
                                                30
                                                                  Gerente 12000.0
                                                                                         MBA 4.0
                                                                                                       1200.0 2400.0 EDUARDO
                        Extrema
                                                     Marketing
                                                                                                                              eduardo
               Ester
                                                                                     Superior 2.0
            81
                        Itapeva
                                   MG|Brasil|
                                               29
                                                      Compras
                                                                  Analistal10000.01
                                                                                                       1000.0[2000.0]
                                                                                                                       ESTER
                                                                                                                                ester
```

MG Brasil

MG|Brasil|

9

Pedro

10 Maria Extrema

Extrema

30

40 İ

Marketing

Produção

Analista 13000.0

Analista 12000.0

1300.0 2600.0

1200.0 2400.0

Superior 1.0

MBA 7.0

PEDRO

MARTA

pedro

marial

```
1 '''
2 7 - Crie um novo dataframe chamado quant apenas com as colunas idade,
3 salário e notas
4 '''
5 quanti = df.select('idade','salario','nota')
6 quanti.show()
   |idade|salario|nota|
       35 | 25000.0 | 8.0 |
       29 | 12000.0 | 6.0 |
       38 28000.01 9.01
       36 40000.0 7.0
       28 23000.0 5.0
       29 11000.0 9.0
       30 12000.0 | 4.0 |
       29 10000.0 | 2.0 |
       30 13000.0 1.0
       40 12000.0 7.0
1 '''
2 8 - calcule as principais medidas descritivas do df quant (describe)
3 '''
4 quanti.describe().show()
₹ +-----+
   Isummary
               idade| salario|
                                                      nota
                       10|
                                             10|
     mean 32.4 18600.0 5.8 stddev 4.402019738458447 10002.221975363496 2.780887148615228
             28 10000.0
                                               1.0
      min
                                    40000.0
        max
2 9 - Ordene o df original por notas de forma crescente 3 '''
4 df.orderBy('nota')
```

 $\overline{\Rightarrow}$

|São Paulo|

matricula	nome	cidade	estado	pais	idade	departamento	cargo	salario	escolaridade	nota	previdencia	bonus	maiu	minu
9	Pedro	Extrema	MG	Brasil	30	Marketing	Analista	13000.0	Superior	1.0	1300.0	2600.0	PEDRO	pedro
8	Ester	Itapeva	MG	Brasil	29	Compras	Analista	10000.0	Superior	2.0	1000.0	2000.0	ESTER	ester
7	Eduardo	Extrema	MG	Brasil	30	Marketing	Gerente	12000.0	MBA	4.0	1200.0	2400.0	EDUARDO	eduardo
5	Sandra	Atibaia	SP	Brasil	28	Produção	Analista	23000.0	Superior	5.0	2300.0	4600.0	SANDRA	sandra
2	Ana	São Paulo	SP	Brasil	29	Vendas	Coordenador	12000.0	Superior	6.0	1200.0	2400.0	ANA	ana
10	Maria	Extrema	MG	Brasil	40	Produção	Analista	12000.0	MBA	7.0	1200.0	2400.0	MARIA	maria
4	Fernando	Atibaia	SP	Brasil	36	Marketing	Diretor	40000.0	Mestrado	7.0	4000.0	8000.0	FERNANDO	fernando
1	Lucas	Atibaia	SP	Brasil	35	Compras	Gerente	25000.0	Superior	8.0	2500.0	5000.0	LUCAS	lucas
3	Luiza	Santos	SP	Brasil	38	Finanças	Gerente	28000.0	MBA	9.0	2800.0	5600.0	LUIZA	luiza
6	Douglas	Bragança	SP	Brasil	29	Finanças	Analista	11000.0	Superior	9.0	1100.0	2200.0	DOUGLAS	douglas

```
1 '''
2 10 - faça as contagens de frequências, usando groupby para as seguintes colunas:
3 a) Cargo
4 b) Cidade
5 c) Estado
6 d) Escolaridade
7 e) Departamento
8
9 '''
10 a = df.groupBy('cargo').agg({'cargo':'count'})
11 a.show()
12 b = df.groupBy('cidade').agg({'cidade':'count'})
13 b.show()
14 c = df.groupBy('estado').agg({'estado':'count'})
15 c.show()
16 d = df.groupBy('escolaridade').agg({'escolaridade':'count'})
17 d.show()
18 e = df.groupBy('departamento').agg({'departamento':'count'})
19 e.show()
      cargo|count(cargo)|
    +-----
                          3
       Gerente
     |Coordenador
      Diretor
       Analista
                          5
    | cidade|count(cidade)|
     ------
      Santos
                         1|
      Extrema
```

Atibaia Bragança Itapeva +	3 1 1
estado count(es SP MG	tado)
	unt(escolaridade) 6 1 3
	unt(departamento) 1 2 3 2 2