trabalho_jony

September 30, 2020

1 ENEM 2016

O objetivo deste caderno é utilizar o Pyspark para ler a grande base de dados e realizar uma limpeza nos dados, de forma que apenas sejam mantidas as informações que são importante para a realização da pesquisa, filtrando linhas e selecionando colunas.

[1]:
```
import findspark
import pyspark

from pyspark.sql import SparkSession
```

Iniciar sessão do Spark

[2]:
```
findspark.init()
```

[3]:
```
spark = SparkSession.builder.getOrCreate()
```

Ler arquivo

[4]:
```
df = spark.read.csv("microdados_enem_2016.csv", sep=";", header=True,␣
↪encoding="Windows-1252")
```

[5]:
```
df.printSchema()
```

```
root
 |-- NU_INSCRICAO: string (nullable = true)
 |-- NU_ANO: string (nullable = true)
 |-- CO_MUNICIPIO_RESIDENCIA: string (nullable = true)
 |-- NO_MUNICIPIO_RESIDENCIA: string (nullable = true)
 |-- CO_UF_RESIDENCIA: string (nullable = true)
 |-- SG_UF_RESIDENCIA: string (nullable = true)
 |-- NU_IDADE: string (nullable = true)
 |-- TP_SEXO: string (nullable = true)
 |-- TP_ESTADO_CIVIL: string (nullable = true)
 |-- TP_COR_RACA: string (nullable = true)
 |-- TP_NACIONALIDADE: string (nullable = true)
 |-- CO_MUNICIPIO_NASCIMENTO: string (nullable = true)
 |-- NO_MUNICIPIO_NASCIMENTO: string (nullable = true)
 |-- CO_UF_NASCIMENTO: string (nullable = true)
```

```
|-- SG_UF_NASCIMENTO: string (nullable = true)
|-- TP_ST_CONCLUSAO: string (nullable = true)
|-- TP_ANO_CONCLUIU: string (nullable = true)
|-- TP_ESCOLA: string (nullable = true)
|-- TP_ENSINO: string (nullable = true)
|-- IN_TREINEIRO: string (nullable = true)
|-- CO_ESCOLA: string (nullable = true)
|-- CO_MUNICIPIO_ESC: string (nullable = true)
|-- NO_MUNICIPIO_ESC: string (nullable = true)
|-- CO_UF_ESC: string (nullable = true)
|-- SG_UF_ESC: string (nullable = true)
|-- TP_DEPENDENCIA_ADM_ESC: string (nullable = true)
|-- TP_LOCALIZACAO_ESC: string (nullable = true)
|-- TP_SIT_FUNC_ESC: string (nullable = true)
|-- IN_BAIXA_VISAO: string (nullable = true)
|-- IN_CEGUEIRA: string (nullable = true)
|-- IN_SURDEZ: string (nullable = true)
|-- IN_DEFICIENCIA_AUDITIVA: string (nullable = true)
|-- IN_SURDO_CEGUEIRA: string (nullable = true)
|-- IN_DEFICIENCIA_FISICA: string (nullable = true)
|-- IN_DEFICIENCIA_MENTAL: string (nullable = true)
|-- IN_DEFICIT_ATENCAO: string (nullable = true)
|-- IN_DISLEXIA: string (nullable = true)
|-- IN_DISCALCULIA: string (nullable = true)
|-- IN_AUTISMO: string (nullable = true)
|-- IN_VISAO_MONOCULAR: string (nullable = true)
|-- IN_OUTRA_DEF: string (nullable = true)
|-- IN_SABATISTA: string (nullable = true)
|-- IN_GESTANTE: string (nullable = true)
|-- IN_LACTANTE: string (nullable = true)
|-- IN_IDOSO: string (nullable = true)
|-- IN_ESTUDA_CLASSE_HOSPITALAR: string (nullable = true)
|-- IN_SEM_RECURSO: string (nullable = true)
|-- IN_BRAILLE: string (nullable = true)
|-- IN_AMPLIADA_24: string (nullable = true)
|-- IN_AMPLIADA_18: string (nullable = true)
|-- IN_LEDOR: string (nullable = true)
|-- IN_ACESSO: string (nullable = true)
|-- IN_TRANSCRICAO: string (nullable = true)
|-- IN_LIBRAS: string (nullable = true)
|-- IN_LEITURA_LABIAL: string (nullable = true)
|-- IN_MESA_CADEIRA_RODAS: string (nullable = true)
|-- IN_MESA_CADEIRA_SEPARADA: string (nullable = true)
|-- IN_APOIO_PERNA: string (nullable = true)
|-- IN_GUIA_INTERPRETE: string (nullable = true)
|-- IN_MACA: string (nullable = true)
|-- IN_COMPUTADOR: string (nullable = true)
|-- IN_CADEIRA_ESPECIAL: string (nullable = true)
```

```
|-- IN_CADEIRA_CANHOTO: string (nullable = true)
|-- IN_CADEIRA_ACOLCHOADA: string (nullable = true)
|-- IN_PROVA_DEITADO: string (nullable = true)
|-- IN_MOBILIARIO_OBESO: string (nullable = true)
|-- IN_LAMINA_OVERLAY: string (nullable = true)
|-- IN_PROTETOR_AURICULAR: string (nullable = true)
|-- IN_MEDIDOR_GLICOSE: string (nullable = true)
|-- IN_MAQUINA_BRAILE: string (nullable = true)
|-- IN_SOROBAN: string (nullable = true)
|-- IN_MARCA_PASSO: string (nullable = true)
|-- IN_SONDA: string (nullable = true)
|-- IN_MEDICAMENTOS: string (nullable = true)
|-- IN_SALA_INDIVIDUAL: string (nullable = true)
|-- IN_SALA_ESPECIAL: string (nullable = true)
|-- IN_SALA_ACOMPANHANTE: string (nullable = true)
|-- IN_MOBILIARIO_ESPECIFICO: string (nullable = true)
|-- IN_MATERIAL_ESPECIFICO: string (nullable = true)
|-- IN_NOME_SOCIAL: string (nullable = true)
|-- IN_CERTIFICADO: string (nullable = true)
|-- NO_ENTIDADE_CERTIFICACAO: string (nullable = true)
|-- CO_UF_ENTIDADE_CERTIFICACAO: string (nullable = true)
|-- SG_UF_ENTIDADE_CERTIFICACAO: string (nullable = true)
|-- CO_MUNICIPIO_PROVA: string (nullable = true)
|-- NO_MUNICIPIO_PROVA: string (nullable = true)
|-- CO_UF_PROVA: string (nullable = true)
|-- SG_UF_PROVA: string (nullable = true)
|-- TP_PRESENCA_CN: string (nullable = true)
|-- TP_PRESENCA_CH: string (nullable = true)
|-- TP_PRESENCA_LC: string (nullable = true)
|-- TP_PRESENCA_MT: string (nullable = true)
|-- CO_PROVA_CN: string (nullable = true)
|-- CO_PROVA_CH: string (nullable = true)
|-- CO_PROVA_LC: string (nullable = true)
|-- CO_PROVA_MT: string (nullable = true)
|-- NU_NOTA_CN: string (nullable = true)
|-- NU_NOTA_CH: string (nullable = true)
|-- NU_NOTA_LC: string (nullable = true)
|-- NU_NOTA_MT: string (nullable = true)
|-- TX_RESPOSTAS_CN: string (nullable = true)
|-- TX_RESPOSTAS_CH: string (nullable = true)
|-- TX_RESPOSTAS_LC: string (nullable = true)
|-- TX_RESPOSTAS_MT: string (nullable = true)
|-- TP_LINGUA: string (nullable = true)
|-- TX_GABARITO_CN: string (nullable = true)
|-- TX_GABARITO_CH: string (nullable = true)
|-- TX_GABARITO_LC: string (nullable = true)
|-- TX_GABARITO_MT: string (nullable = true)
|-- TP_STATUS_REDACAO: string (nullable = true)
```

```
|-- NU_NOTA_COMP1: string (nullable = true)
|-- NU_NOTA_COMP2: string (nullable = true)
|-- NU_NOTA_COMP3: string (nullable = true)
|-- NU_NOTA_COMP4: string (nullable = true)
|-- NU_NOTA_COMP5: string (nullable = true)
|-- NU_NOTA_REDACAO: string (nullable = true)
|-- Q001: string (nullable = true)
|-- Q002: string (nullable = true)
|-- Q003: string (nullable = true)
|-- Q004: string (nullable = true)
|-- Q005: string (nullable = true)
|-- Q006: string (nullable = true)
|-- Q007: string (nullable = true)
|-- Q008: string (nullable = true)
|-- Q009: string (nullable = true)
|-- Q010: string (nullable = true)
|-- Q011: string (nullable = true)
|-- Q012: string (nullable = true)
|-- Q013: string (nullable = true)
|-- Q014: string (nullable = true)
|-- Q015: string (nullable = true)
|-- Q016: string (nullable = true)
|-- Q017: string (nullable = true)
|-- Q018: string (nullable = true)
|-- Q019: string (nullable = true)
|-- Q020: string (nullable = true)
|-- Q021: string (nullable = true)
|-- Q022: string (nullable = true)
|-- Q023: string (nullable = true)
|-- Q024: string (nullable = true)
|-- Q025: string (nullable = true)
|-- Q026: string (nullable = true)
|-- Q027: string (nullable = true)
|-- Q028: string (nullable = true)
|-- Q029: string (nullable = true)
|-- Q030: string (nullable = true)
|-- Q031: string (nullable = true)
|-- Q032: string (nullable = true)
|-- Q033: string (nullable = true)
|-- Q034: string (nullable = true)
|-- Q035: string (nullable = true)
|-- Q036: string (nullable = true)
|-- Q037: string (nullable = true)
|-- Q038: string (nullable = true)
|-- Q039: string (nullable = true)
|-- Q040: string (nullable = true)
|-- Q041: string (nullable = true)
|-- Q042: string (nullable = true)
```

```
|-- Q043: string (nullable = true)
|-- Q044: string (nullable = true)
|-- Q045: string (nullable = true)
|-- Q046: string (nullable = true)
|-- Q047: string (nullable = true)
|-- Q048: string (nullable = true)
|-- Q049: string (nullable = true)
|-- Q050: string (nullable = true)
```

[6]:
```
len(df.columns)
```

[6]: 166

Filtrar os dados segundo os critérios do contratante: - Apensar a região nordeste; - Participantes que estavam no último ano do ensino médio ou já concluíram; - Participantes que não eram “treineiros”; - Participantes que estavam presentes em todas as provas; - Participantes que não tiveram a redação anulada.

[7]:
```
df = df.filter(df.CO_MUNICIPIO_RESIDENCIA.startswith("2"))\
        .filter(df.TP_ST_CONCLUSAO != "3")\
        .filter(df.TP_ST_CONCLUSAO != "4")\
        .filter(df.IN_TREINEIRO == 0)\
        .filter(df.TP_PRESENCA_CN == 1)\
        .filter(df.TP_PRESENCA_CH == 1)\
        .filter(df.TP_PRESENCA_LC == 1)\
        .filter(df.TP_PRESENCA_MT == 1)\
        .filter(df.TP_STATUS_REDACAO == 1)
```

[8]:
```
df = df.select('NU_INSCRICAO',
               'CO_MUNICIPIO_RESIDENCIA',
               'TP_COR_RACA',
               'TP_ST_CONCLUSAO',
               'IN_TREINEIRO',
               'TP_ESCOLA',
               'TP_PRESENCA_CN',
               'TP_PRESENCA_CH',
               'TP_PRESENCA_LC',
               'TP_PRESENCA_MT',
               'TP_STATUS_REDACAO',
               'Q006',
               'Q005',
               'Q047',
               'NU_NOTA_CN',
               'NU_NOTA_CH',
               'NU_NOTA_LC',
               'NU_NOTA_MT',
               'NU_NOTA_REDACAO')
```

[9]:
```
len(df.columns)
```

[9]:
```
19
```

Salva arquivo com menor dimensão

[10]:
```
df.toPandas().to_csv('dados_enem_filtrado.csv')
```