```python
from pyspark.sql import SparkSession
from pyspark import SparkContext, SQLContext
from pyspark.sql.functions import lit, rand, col
from pyspark.sql.types import IntegerType
import random

# Inicializar Spark Session
spark = SparkSession \
    .builder \
    .master("local[*]") \
    .appName("Exercicio Intro") \
    .getOrCreate()

# Ler o arquivo nomes_aleatorios.txt
df_nomes = spark.read.text("nomes_aleatorios.txt")

# Renomear a coluna para "Nomes"
df_nomes = df_nomes.withColumnRenamed("value", "Nomes")

# Mostrar esquema e 10 linhas do dataframe
df_nomes.printSchema()
df_nomes.show(10, truncate=False)

# Adicionar coluna "Escolaridade" com valores aleatórios
df_nomes = df_nomes.withColumn("Escolaridade",
lit(random.choice(["Fundamental", "Medio", "Superior"])))

# Adicionar coluna "Pais" com nomes de países da América do Sul de forma
aleatória
paises_am_sul = ["Argentina", "Bolivia", "Brasil", "Chile", "Colombia",
"Equador", "Guiana", "Paraguai", "Peru", "Suriname", "Uruguai",
"Venezuela"]
df_nomes = df_nomes.withColumn("Pais", lit(random.choice(paises_am_sul)))

# Adicionar coluna "AnoNascimento" com valores aleatórios entre 1945 e
2010
df_nomes = df_nomes.withColumn("AnoNascimento", (rand() * (2010 - 1945 +
1) + 1945).cast(IntegerType()))

# Selecionar pessoas que nasceram neste século
df_select = df_nomes.filter(df_nomes["AnoNascimento"] >= 2000)
df_select.show(10, truncate=False)

# Registrar tabela temporária para uso com Spark SQL
df_nomes.createOrReplaceTempView("pessoas")

# Usar Spark SQL para selecionar pessoas que nasceram neste século
df_select_sql = spark.sql("SELECT * FROM pessoas WHERE AnoNascimento >=
2000")
```

```
df_select_sql.show(10, truncate=False)

# Contar o número de pessoas da geração Millennials usando o método
select do dataframe
count_millennials_df = df_nomes.filter((df_nomes["AnoNascimento"] >=
1980) & (df_nomes["AnoNascimento"] <= 1994)).count()
print("Número de Millennials (usando DataFrame):", count_millennials_df)

# Contar o número de pessoas da geração Millennials usando Spark SQL
count_millennials_sql = spark.sql("SELECT COUNT(*) FROM pessoas WHERE
AnoNascimento BETWEEN 1980 AND 1994").collect()[0][0]
print("Número de Millennials (usando Spark SQL):", count_millennials_sql)

# Obter a quantidade de pessoas por país e geração usando Spark SQL
df_quantidade_paises_geracao = spark.sql("""
    SELECT Pais,
           CASE
               WHEN AnoNascimento BETWEEN 1944 AND 1964 THEN 'Baby
Boomers'
               WHEN AnoNascimento BETWEEN 1965 AND 1979 THEN 'Geração X'
               WHEN AnoNascimento BETWEEN 1980 AND 1994 THEN
'Millennials'
               WHEN AnoNascimento BETWEEN 1995 AND 2015 THEN 'Geração Z'
               ELSE 'Outra'
           END AS Geracao,
           COUNT(*) AS Quantidade
    FROM pessoas
    GROUP BY Pais, Geracao
    ORDER BY Pais, Geracao
""")
df_quantidade_paises_geracao.show(truncate=False)
```