```
# part 1
"""
This script processes a diabetes dataset that is kept in a CSV file by
using PySpark. To start,
a SparkSession called "Diabetes" is created, and the dataset is loaded
into a DataFrame called df. The script creates an updated DataFrame
called d by calculating the dataset's mean BMI (body mass index) and
replacing any zero values in the BMI column with the mean. After that,
a new DataFrame called df_rs is created, with entries in it that have
an age of at least 35. Filtering rows where the Diabetes Pedigree
Function value is
greater than or equal to 0.51 results in the generation of another
DataFrame, df_fltd. Lastly, the script produces the modified BMI
DataFrame, the DataFrame filtered according to the Diabetes Pedigree
Function threshold, and the DataFrame with age greater than or equal
to 35. When processing is finished, the Spark session is terminated.
Using PySpark, this method makes it possible to manipulate and filter
data effectively while gaining insights from the diabetes dataset.
"""
```

```
{"type":"string"}
```

```
!pip install pyspark --quiet
```

```
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 317.0/317.0 MB 2.2 MB/s eta
0:00:00
etadata (setup.py) ...
```

```
from pyspark.sql import SparkSession
from pyspark.sql.functions import mean, when

spark = SparkSession.builder \
    .appName("Diabetes") \
    .getOrCreate()


df = spark.read.csv("diabetes.csv", header=True, inferSchema=True)

mn_bmi = df.select(mean("BMI")).collect()[0][0]
d = df.withColumn("BMI", when(df["BMI"] == 0,
mn_bmi).otherwise(df["BMI"]))

# new DataFrame for rows with age >= 35
df_rs = df.filter(df["Age"] >= 35)

# Diabetes Pedigree Function value is >= 0.51
df_fltd = df.filter(df["DiabetesPedigreeFunction"] >= 0.51)


print("Updated BMI:")
```

```
d.show()

print("Age >= 35:")
df_rs.show()

print("Diabetes Pedigree Function >= 0.51:")
df_fltd.show()
spark.stop()
```

```
Updated BMI:
+-----------+-------+-------------+-------------+-------
+------------------+------------------------+---+-------+
|Pregnancies|Glucose|BloodPressure|SkinThickness|Insulin|
BMI|DiabetesPedigreeFunction|Age|Outcome|
+-----------+-------+-------------+-------------+-------
+------------------+------------------------+---+-------+
|          6|    148|           72|           35|      0|
33.6|                   0.627| 50|      1|
|          1|     85|           66|           29|      0|
26.6|                   0.351| 31|      0|
|          8|    183|           64|            0|      0|
23.3|                   0.672| 32|      1|
|          1|     89|           66|           23|     94|
28.1|                   0.167| 21|      0|
|          0|    137|           40|           35|    168|
43.1|                   2.288| 33|      1|
|          5|    116|           74|            0|      0|
25.6|                   0.201| 30|      0|
|          3|     78|           50|           32|     88|
31.0|                   0.248| 26|      1|
|         10|    115|            0|            0|      0|
35.3|                   0.134| 29|      0|
|          2|    197|           70|           45|    543|
30.5|                   0.158| 53|      1|
|          8|    125|           96|            0|      0|
31.992578124999977|                   0.232| 54|      1|
|          4|    110|           92|            0|      0|
37.6|                   0.191| 30|      0|
|         10|    168|           74|            0|      0|
38.0|                   0.537| 34|      1|
|         10|    139|           80|            0|      0|
27.1|                   1.441| 57|      0|
|          1|    189|           60|           23|    846|
30.1|                   0.398| 59|      1|
|          5|    166|           72|           19|    175|
25.8|                   0.587| 51|      1|
|          7|    100|            0|            0|      0|
30.0|                   0.484| 32|      1|
|          0|    118|           84|           47|    230|
45.8|                   0.551| 31|      1|
```

```
|          7|    107|           74|            0|      0|
29.6|                   0.254| 31|      1|
|          1|    103|           30|           38|     83|
43.3|                   0.183| 33|      0|
|          1|    115|           70|           30|     96|
34.6|                   0.529| 32|      1|
+-----------+-------+-------------+-------------+-------
+------------------+-----------------------+---+-------+
only showing top 20 rows

Age >= 35:
+-----------+-------+-------------+-------------+-------+----
+-----------------------+---+-------+
|Pregnancies|Glucose|BloodPressure|SkinThickness|Insulin| BMI|
DiabetesPedigreeFunction|Age|Outcome|
+-----------+-------+-------------+-------------+-------+----
+-----------------------+---+-------+
|          6|    148|           72|           35|      0|33.6|
0.627| 50|      1|
|          2|    197|           70|           45|    543|30.5|
0.158| 53|      1|
|          8|    125|           96|            0|      0| 0.0|
0.232| 54|      1|
|         10|    139|           80|            0|      0|27.1|
1.441| 57|      0|
|          1|    189|           60|           23|    846|30.1|
0.398| 59|      1|
|          5|    166|           72|           19|    175|25.8|
0.587| 51|      1|
|          8|     99|           84|            0|      0|35.4|
0.388| 50|      0|
|          7|    196|           90|            0|      0|39.8|
0.451| 41|      1|
|         11|    143|           94|           33|    146|36.6|
0.254| 51|      1|
|         10|    125|           70|           26|    115|31.1|
0.205| 41|      1|
|          7|    147|           76|            0|      0|39.4|
0.257| 43|      1|
|         13|    145|           82|           19|    110|22.2|
0.245| 57|      0|
|          5|    117|           92|            0|      0|34.1|
0.337| 38|      0|
|          5|    109|           75|           26|      0|36.0|
0.546| 60|      0|
|         10|    122|           78|           31|      0|27.6|
0.512| 45|      0|
|         11|    138|           76|            0|      0|33.2|
0.42| 35|      0|
```

```
|          9|    102|           76|           37|      0|32.9|
0.665| 46|      1|
|          4|    111|           72|           47|    207|37.1|
1.39| 56|      1|
|          7|    133|           84|            0|      0|40.2|
0.696| 37|      0|
|          7|    106|           92|           18|      0|22.7|
0.235| 48|      0|
+-----------+-------+-------------+-------------+-------+----
+------------------------+---+-------+
only showing top 20 rows

Diabetes Pedigree Function >= 0.51:
+-----------+-------+-------------+-------------+-------+----
+------------------------+---+-------+
|Pregnancies|Glucose|BloodPressure|SkinThickness|Insulin| BMI|
DiabetesPedigreeFunction|Age|Outcome|
+-----------+-------+-------------+-------------+-------+----
+------------------------+---+-------+
|          6|    148|           72|           35|      0|33.6|
0.627| 50|      1|
|          8|    183|           64|            0|      0|23.3|
0.672| 32|      1|
|          0|    137|           40|           35|    168|43.1|
2.288| 33|      1|
|         10|    168|           74|            0|      0|38.0|
0.537| 34|      1|
|         10|    139|           80|            0|      0|27.1|
1.441| 57|      0|
|          5|    166|           72|           19|    175|25.8|
0.587| 51|      1|
|          0|    118|           84|           47|    230|45.8|
0.551| 31|      1|
|          1|    115|           70|           30|     96|34.6|
0.529| 32|      1|
|          3|    126|           88|           41|    235|39.3|
0.704| 27|      0|
|          5|    109|           75|           26|      0|36.0|
0.546| 60|      0|
|          3|    158|           76|           36|    245|31.6|
0.851| 28|      1|
|         10|    122|           78|           31|      0|27.6|
0.512| 45|      0|
|          4|    103|           60|           33|    192|24.0|
0.966| 33|      0|
|          9|    102|           76|           37|      0|32.9|
0.665| 46|      1|
|          4|    111|           72|           47|    207|37.1|
1.39| 56|      1|
```

```
|          7|    133|           84|             0|      0|40.2|
0.696| 37|      0|
|          9|    171|          110|            24|    240|45.4|
0.721| 54|      1|
|          0|    180|           66|            39|      0|42.0|
1.893| 25|      1|
|          1|    146|           56|             0|      0|29.7|
0.564| 29|      0|
|          2|     71|           70|            27|      0|28.0|
0.586| 22|      0|
+-----------+-------+-------------+--------------+-------+----
+------------------------+---+-------+
only showing top 20 rows
```