Spark SQL - Exercise

Exercise #47

- Input:
  - A CSV file containing a list of user profiles
    - Header
      - name,age,gender
    - Each line of the file contains the information about one user
- Output:
  - Select male users (gender="male"), increase by one their age, and store in the output folder name and age of these users sorted by decreasing age and ascending name (if the age value is the same)
  - The output does not contain the header line

Exercise #47

- Example of input data:

```
name,age,gender
Paul,40,male
John,40,male
David,15,male
Susan,40,female
Karen,34,female
```

- Example of expected output:

```
John,41
Paul,41
David,16
```

3

Exercise #47

- Implement two different solutions for this exercise
  - A solution based only on DataFrames
  - A solution based on SQL like queries executed on a temporary table associated with the input data

4

Exercise #48

- Input:
  - A CSV file containing a list of user profiles
    - Header
      - name,age,gender
    - Each line of the file contains the information about one user
- Output:
  - Select the names occurring at least two times and store in the output folder name and average(age) of the selected names
  - The output does not contain the header line

5

Exercise #48

- Example of input data:

  name,age,gender
  Paul,40,male
  Paul,38,male
  David,15,male
  Susan,40,female
  Susan,34,female
- Example of expected output:

  Paul,39
  Susan,37

6

Exercise #48

- Implement two different solutions for this exercise
  - A solution based only on DataFrames
  - A solution based on SQL like queries executed on a temporary table associated with the input data