M1BIT1_AI3_EricssonMarc72

EJERCICIO 1: Obtén en qué 10 estados nacieron más niños y niñas en 2003.

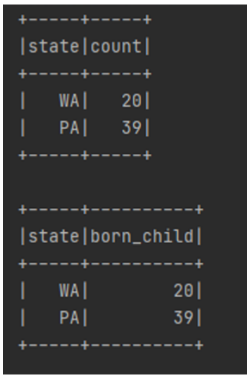
API

```
df.where(df.year==2003)\
   .select(df.state,df.is_male)\
   .groupBy("state")\
   .count().alias("born_child")\
   .show()
```

SQL

```
df.createOrReplaceTempView("natality")

spark.sql("SELECT state,COUNT(is_male) FROM natality WHERE year=2003 GROUP BY state").show(5)
```



```
+-----+-----+
|state|count|
+-----+-----+
|   WA|   20|
|   PA|   39|
+-----+-----+

+-----+----------+
|state|born_child|
+-----+----------+
|   WA|        20|
|   PA|        39|
+-----+----------+
```

EJERCICIO 2: Obtén la media de peso de los niños y niñas por año y estado

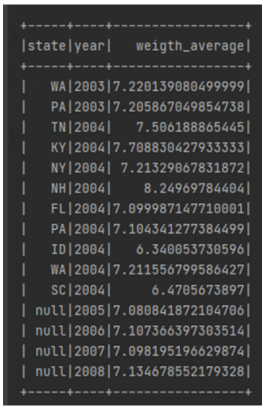
API

```
df.groupBy("state","year")\
   .agg({"weight_pounds":"avg"})\
   .show()
```

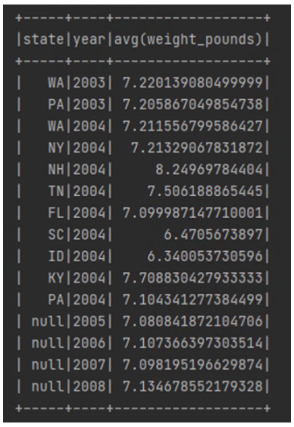
SQL

```
df.createOrReplaceTempView("natality")

spark.sql("SELECT state, year, AVG(weight_pounds) AS weigth_average FROM natality GROUP BY state, year ORDER BY year").show()
```



```
+-----+----+-----------------+
|state|year|   weigth_average|
+-----+----+-----------------+
|   WA|2003|7.220139080499999|
|   PA|2003|7.205867049854738|
|   TN|2004|   7.506188865445|
|   KY|2004|7.708830427933333|
|   NY|2004| 7.21329067831872|
|   NH|2004|    8.24969784404|
|   FL|2004|7.099987147710001|
|   PA|2004|7.104341277384499|
|   ID|2004|   6.340053730596|
|   WA|2004|7.211556799586427|
|   SC|2004|     6.4705673897|
| null|2005|7.080841872104706|
| null|2006|7.107366397303514|
| null|2007|7.098195196629874|
| null|2008|7.134678552179328|
+-----+----+-----------------+
```



```
+-----+----+------------------+
|state|year|avg(weight_pounds)|
+-----+----+------------------+
|   WA|2003| 7.220139080499999|
|   PA|2003| 7.205867049854738|
|   WA|2004| 7.211556799586427|
|   NY|2004|  7.21329067831872|
|   NH|2004|     8.24969784404|
|   TN|2004|    7.506188865445|
|   FL|2004| 7.099987147710001|
|   SC|2004|      6.4705673897|
|   ID|2004|    6.340053730596|
|   KY|2004| 7.708830427933333|
|   PA|2004| 7.104341277384499|
| null|2005| 7.080841872104706|
| null|2006| 7.107366397303514|
| null|2007| 7.098195196629874|
| null|2008| 7.134678552179328|
+-----+----+------------------+
```

EJERCICIO 3: Evolución por año y por mes del número de niños y niñas nacidas

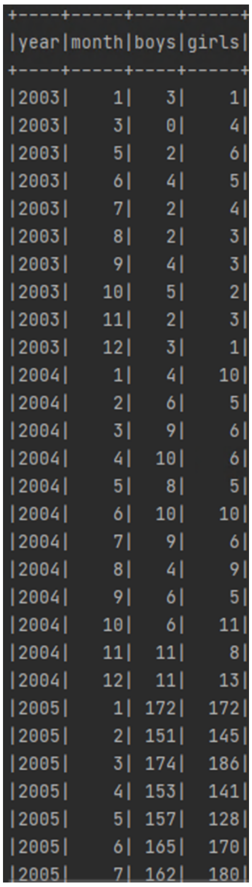
API

```
df.select(df.year,df.month,when(df.is_male == 'true', 1).otherwise(0).alias('boys'),when(df.is_male == 'false', 1).otherwise(0).alias('girls'))\
        .groupBy(df.year,df.month)\
        .agg(sum('boys').alias('boys'),sum('girls').alias('girls'))\
        .orderBy(df.year,df.month)\
        .show(60)
```

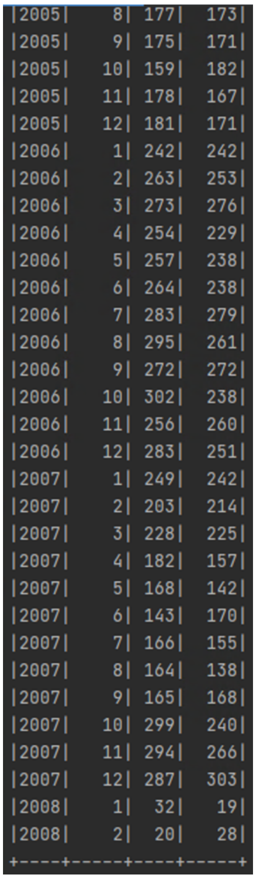
SQL

```
df.createOrReplaceTempView("natality")

spark.sql("SELECT year, month, SUM(CASE WHEN is_male = 'true' THEN 1 ELSE 0 END) AS boys, "
        "SUM(CASE WHEN is_male = 'false' THEN 1 ELSE 0 END) AS girls "
        "FROM natality "
        "GROUP BY year, month "
        "ORDER BY year, month").show(60)
```



```
+----+-----+----+-----+
|year|month|boys|girls|
+----+-----+----+-----+
|2003|    1|   3|    1|
|2003|    3|   0|    4|
|2003|    5|   2|    6|
|2003|    6|   4|    5|
|2003|    7|   2|    4|
|2003|    8|   2|    3|
|2003|    9|   4|    3|
|2003|   10|   5|    2|
|2003|   11|   2|    3|
|2003|   12|   3|    1|
|2004|    1|   4|   10|
|2004|    2|   6|    5|
|2004|    3|   9|    6|
|2004|    4|  10|    6|
|2004|    5|   8|    5|
|2004|    6|  10|   10|
|2004|    7|   9|    6|
|2004|    8|   4|    9|
|2004|    9|   6|    5|
|2004|   10|   6|   11|
|2004|   11|  11|    8|
|2004|   12|  11|   13|
|2005|    1| 172|  172|
|2005|    2| 151|  145|
|2005|    3| 174|  186|
|2005|    4| 153|  141|
|2005|    5| 157|  128|
|2005|    6| 165|  170|
|2005|    7| 162|  180|
```



```
|2005|    8| 177|  173|
|2005|    9| 175|  171|
|2005|   10| 159|  182|
|2005|   11| 178|  167|
|2005|   12| 181|  171|
|2006|    1| 242|  242|
|2006|    2| 263|  253|
|2006|    3| 273|  276|
|2006|    4| 254|  229|
|2006|    5| 257|  238|
|2006|    6| 264|  238|
|2006|    7| 283|  279|
|2006|    8| 295|  261|
|2006|    9| 272|  272|
|2006|   10| 302|  238|
|2006|   11| 256|  260|
|2006|   12| 283|  251|
|2007|    1| 249|  242|
|2007|    2| 203|  214|
|2007|    3| 228|  225|
|2007|    4| 182|  157|
|2007|    5| 168|  142|
|2007|    6| 143|  170|
|2007|    7| 166|  155|
|2007|    8| 164|  138|
|2007|    9| 165|  168|
|2007|   10| 299|  240|
|2007|   11| 294|  266|
|2007|   12| 287|  303|
|2008|    1|  32|   19|
|2008|    2|  20|   28|
+----+-----+----+-----+
```

EJERCICIO 4: Obtén los tres meses de 2005 en que nacieron más niños y niñas.

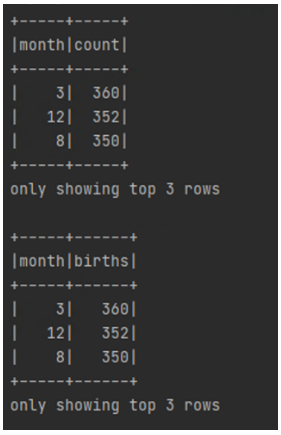
API

```
df.where(df.year==2005)\
   .select(df.month,df.is_male)\
   .groupBy("month")\
   .count()\
   .orderBy("count", ascending = False)\
   .show(3)
```

SQL

```
df.createOrReplaceTempView("natality")

spark.sql("SELECT month,COUNT(is_male) FROM natality WHERE year=2005 GROUP BY month ORDER BY COUNT(is_male) DESC ").show(3)
```



```
+-----+-----+
|month|count|
+-----+-----+
|    3|  360|
|   12|  352|
|    8|  350|
+-----+-----+
only showing top 3 rows

+-----+------+
|month|births|
+-----+------+
|    3|   360|
|   12|   352|
|    8|   350|
+-----+------+
only showing top 3 rows
```

EJERCICIO 5: Obtén los estados donde las semanas de gestación son superiores a la media de EE. UU.

API

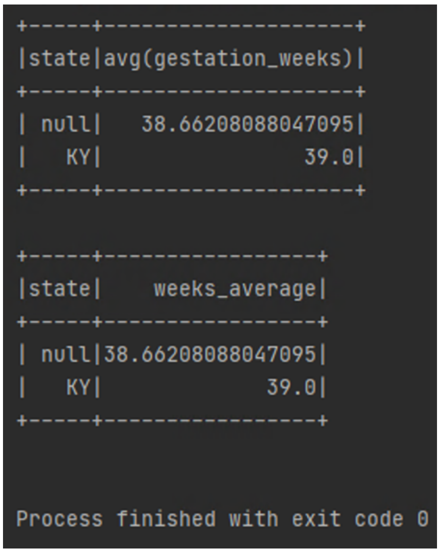
```
val = df.agg({"gestation_weeks":"avg"}).collect()

df.groupBy("state")\
.agg({"gestation_weeks":"avg"})\
.where(col('avg(gestation_weeks)') > val[0][0])\
.show()
```

SQL

```
df.createOrReplaceTempView("natality")

spark.sql("SELECT state, AVG(gestation_weeks) AS weeks_average FROM natality GROUP BY state HAVING AVG(gestation_weeks) > (SELECT AVG(gestation_weeks) FROM natality)").show()
```



```
+-----+--------------------+
|state|avg(gestation_weeks)|
+-----+--------------------+
| null|   38.66208088047095|
|   KY|                39.0|
+-----+--------------------+


+-----+-----------------+
|state|    weeks_average|
+-----+-----------------+
| null|38.66208088047095|
|   KY|             39.0|
+-----+-----------------+


Process finished with exit code 0
```

EJERCICIO 6: Obtén los cinco estados donde la media de edad de las madres ha sido mayor.

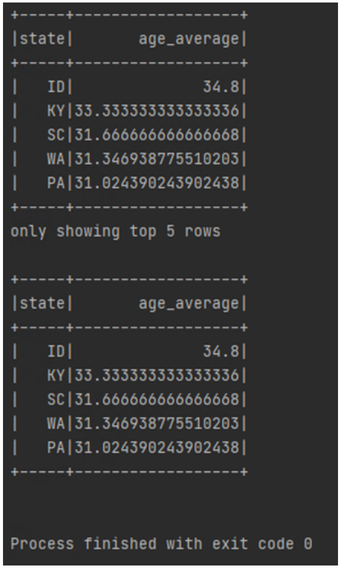
API

```
df.groupBy(df.state)\
    .agg(avg(df.mother_age).alias("age_average"))\
    .sort(desc("age_average"))\
    .show(5)
```

SQL

```
df.createOrReplaceTempView("natality")

spark.sql("SELECT state, AVG(mother_age) AS age_average FROM natality GROUP BY state ORDER BY 2 DESC LIMIT 5").show()
```



```
+-----+------------------+
|state|       age_average|
+-----+------------------+
|   ID|              34.8|
|   KY|33.333333333333336|
|   SC|31.666666666666668|
|   WA|31.346938775510203|
|   PA|31.024390243902438|
+-----+------------------+
only showing top 5 rows

+-----+------------------+
|state|       age_average|
+-----+------------------+
|   ID|              34.8|
|   KY|33.333333333333336|
|   SC|31.666666666666668|
|   WA|31.346938775510203|
|   PA|31.024390243902438|
+-----+------------------+


Process finished with exit code 0
```

EJERCICIO 7: Indica cómo influye en el peso y las semanas de gestación que la madre haya bebido y/o fumado respecto a las que no lo han hecho.

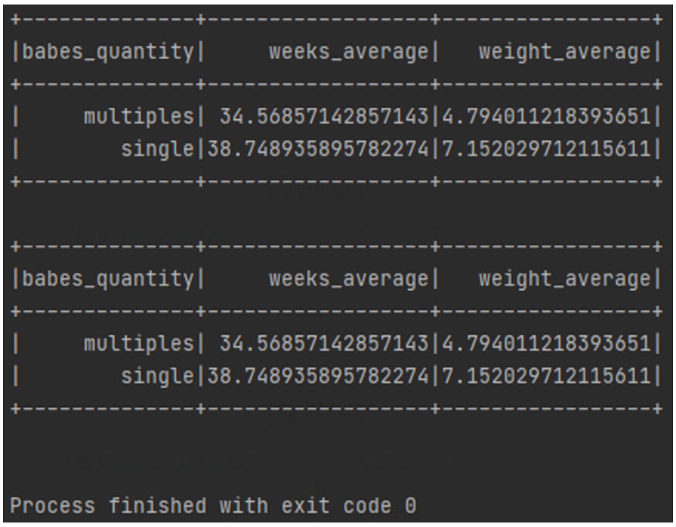
API

```
df.select(df.weight_pounds, df.gestation_weeks, when(df.plurality == 1,
'single').otherwise('multiples').alias('babes_quantity'))\
.groupBy('babes_quantity')\
.agg(avg(df.gestation_weeks).alias('weeks_average'),avg(df.weight_pounds).alias('weight_average')\
.show()
```

SQL

```
df.createOrReplaceTempView("natality")

spark.sql("SELECT CASE WHEN plurality = 1 THEN 'single' ELSE 'multiples' END num_babes,
AVG(gestation_weeks) AS weeks_average, AVG(weight_pounds) AS weight_average FROM natality GROUP BY
babes_quantity").show()
```



```
+--------------+------------------+-----------------+
|babes_quantity|     weeks_average|   weight_average|
+--------------+------------------+-----------------+
|     multiples| 34.56857142857143|4.794011218393651|
|        single|38.748935895782274|7.152029712115611|
+--------------+------------------+-----------------+


+--------------+------------------+-----------------+
|babes_quantity|     weeks_average|   weight_average|
+--------------+------------------+-----------------+
|     multiples| 34.56857142857143|4.794011218393651|
|        single|38.748935895782274|7.152029712115611|
+--------------+------------------+-----------------+



Process finished with exit code 0
```