**M1BIT1\_AI3\_EricssonMarc72**

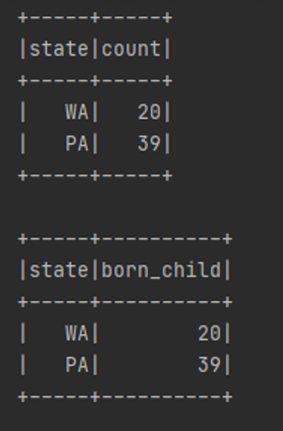
**EJERCICIO 1:** Obtén en qué 10 estados nacieron más niños y niñas en 2003.

**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)



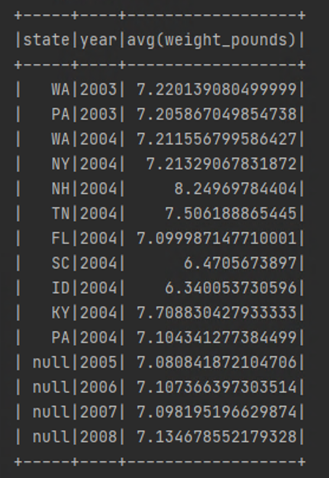
**EJERCICIO 2:** Obtén la media de peso de los niños y niñas por año y estado

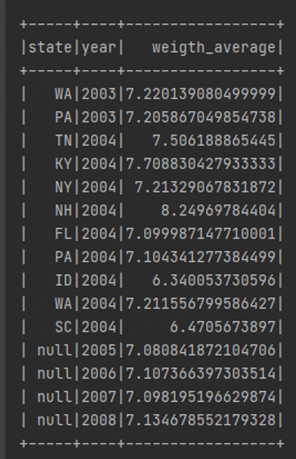
**API**

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()





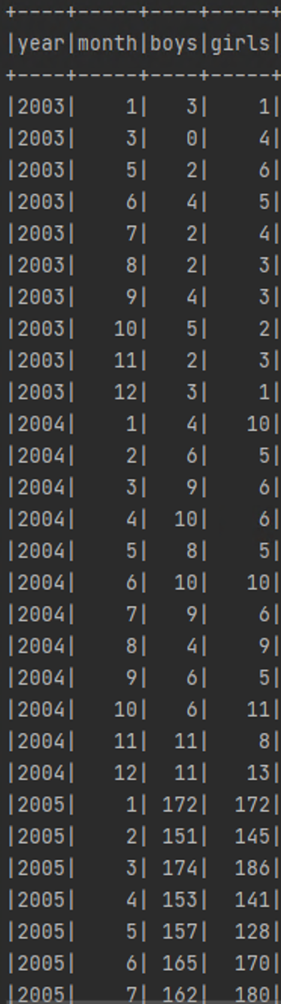
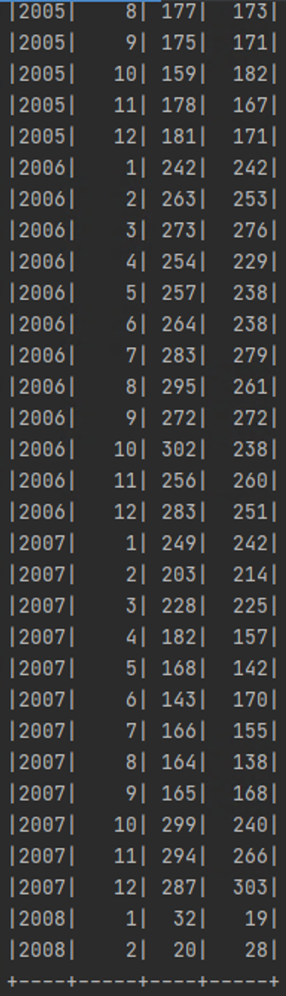
**EJERCICIO 3:** Evolución por año y por mes del número de niños y niñas nacidas

**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)

****

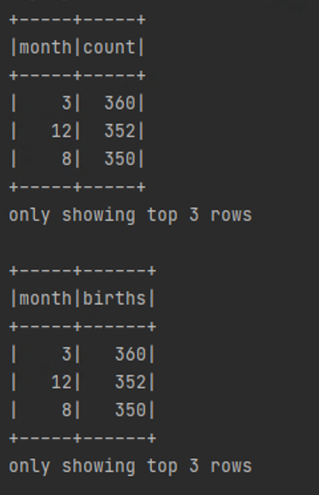
**EJERCICIO 4:** Obtén los tres meses de 2005 en que nacieron más niños y niñas.

**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)



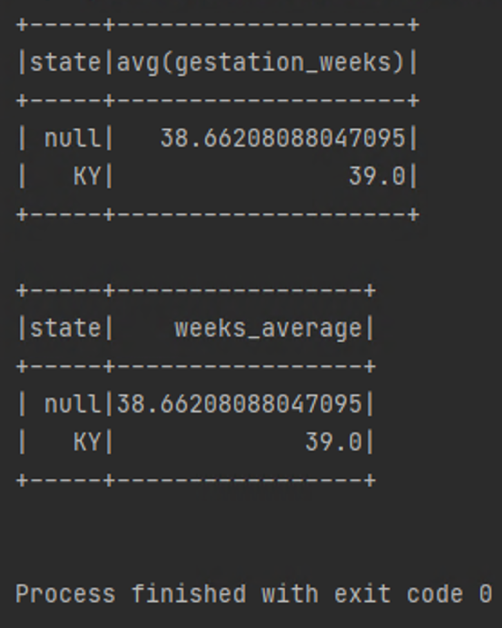
**EJERCICIO 5:** Obtén los estados donde las semanas de gestación son superiores a la media de EE. UU.

**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()



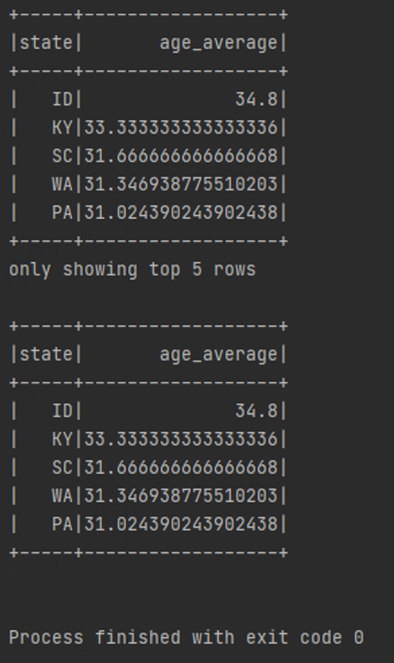
**EJERCICIO 6:** Obtén los cinco estados donde la media de edad de las madres ha sido mayor.

**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()

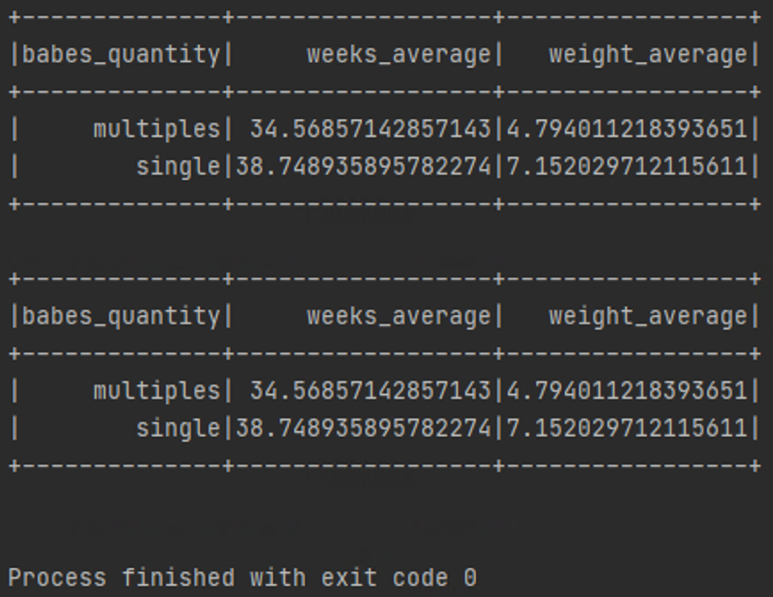


**EJERCICIO 7:** Indica cómo influye en el peso y las semanas de gestación que la madre haya bebido y/o fumado respecto a las que no lo han hecho.

**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()

****