M1B1T1_AI3: Procesamiento de datos con Spark 2.x. Natalidad EE.UU.

Autor: Leandro Gutierrez

Este documento intenta dar respuesta a las actividades propuestas en el Modulo 1 Bloque 1 Actividad Individual 3. En él se describirán cada uno de los enunciados postulados y los resultados obtenidos a través del uso de Spark y sus APIs DataFrame y SQL

Abril 20, 2024

Descripción

Has sido contratado por una empresa consultora como Data Engineer y te proporcionan un fichero CSV con datos reales sobre la natalidad en EE. UU.

El esquema del fichero es el siguiente:

Field name	Type	Mode	Description
source_year	INTEGER	REQUIRED	Four-digit year of the birth. Example: 1975.
year	INTEGER	NULLABLE	Four-digit year of the birth. Example: 1975.
month	INTEGER	NULLABLE	Month index of the date of birth, where 1=January.
day	INTEGER	NULLABLE	Day of birth, starting from 1.
wday	INTEGER	NULLABLE	Day of the week, where 1 is Sunday and 7 is Saturday.
state	STRING	NULLABLE	The two character postal code for the state. Entries after 2004 do not include this value.
is_male	BOOLEAN	REQUIRED	TRUE if the child is male, FALSE if female.
child_race	INTEGER	NULLABLE	The race of the child. One of the following numbers: 1 - White 2 - Black 3 - American Indian 4 - Chinese 5 - Japanese 6 - Hawaiian 7 - Filipino 9 - Unknown/Other 18 - Asian Indian 28 - Korean 39 - Samoan 48 - Vietnamese
weight_pounds	FLOAT	NULLABLE	Weight of the child, in pounds.
plurality	INTEGER	NULLABLE	How many children were born as a result of this pregnancy. twins=2, triplets=3, and so on.
apgar_1min	INTEGER	NULLABLE	Apgar scores measure the health of a newborn child on a scale from 0-10. Value after 1 minute. Available from 1978-2002.
apgar_5min	INTEGER	NULLABLE	Apgar scores measure the health of a newborn child on a scale from 0-10. Value after 5 minutes. Available from 1978-2002.
mother_residence_state	STRING	NULLABLE	The two-letter postal code of the mother's state of residence when the child was born.
mother_race	INTEGER	NULLABLE	Race of the mother. Same values as child_race.
mother_age	INTEGER	NULLABLE	Reported age of the mother when giving birth.
gestation_weeks	INTEGER	NULLABLE	The number of weeks of the pregnancy.
lmp	STRING	NULLABLE	Date of the last menstrual period in the format MMDDYYYY. Unknown values are recorded as "99" or "9999".
mother_married	BOOLEAN	NULLABLE	True if the mother was married when she gave birth.
mother_birth_state	STRING	NULLABLE	The two-letter postal code of the mother's birth state.
cigarette_use	BOOLEAN	NULLABLE	True if the mother smoked cigarettes. Available starting 2003.
cigarettes_per_day	INTEGER	NULLABLE	Number of cigarettes smoked by the mother per day. Available starting 2003.
alcohol_use	BOOLEAN	NULLABLE	True if the mother used alcohol. Available starting 1989.
drinks_per_week	INTEGER	NULLABLE	Number of drinks per week consumed by the mother. Available starting 1989.
weight_gain_pounds	INTEGER	NULLABLE	Number of pounds gained by the mother during pregnancy.
born_alive_alive	INTEGER	NULLABLE	Number of children previously born to the mother who are now living.
born_alive_dead	INTEGER	NULLABLE	Number of children previously born to the mother who are now dead.
born_dead	INTEGER	NULLABLE	Number of children who were born dead (i.e. miscarriages)
ever_born	INTEGER	NULLABLE	Total number of children to whom the woman has ever given birth (includes the current birth).
father_race	INTEGER	NULLABLE	Race of the father. Same values as child_race.
father_age	INTEGER	NULLABLE	Age of the father when the child was born.
record_weight	INTEGER	NULLABLE	1 or 2, where 1 is a row from a full-reporting area, and 2 is a row from a 50% sample area.

Ejercicio 1

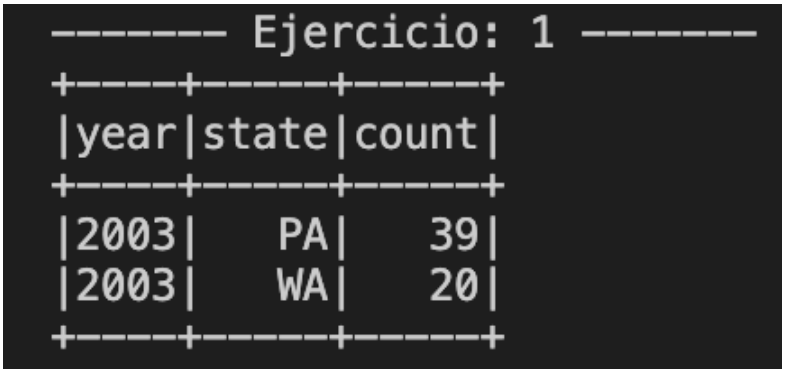
> Obtén en qué 10 estados nacieron más bebés en 2003.

Query

```
# DataFrame
df.filter(df.year == 2003)\
    .groupBy("year", "state")\
    .count()\
    .orderBy("count", ascending=False)\
    .limit(10)\
    .show()

# SQL
df.createOrReplaceTempView("natality")
sqlDF = spark.sql('''SELECT year, state, COUNT(*) as count
                     FROM natality
                     WHERE year = 2003
                     GROUP BY year, state
                     ORDER BY count desc
                     LIMIT 10''')
sqlDF.show()
```

Resultados



Respuesta

Durante el año 2003 solo se cuentan con datos de dos estados PA (Pennsylvania) y WA (Washington), donde se registraron 39 y 20 naciemientos respectivamente.

Ejercicio 2

> Obtén la media de peso de los bebés por año y estado.

Query

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

# SQL
df.createOrReplaceTempView("natality")
sql = spark.sql('''
                SELECT year, state, AVG(weight_pounds)
                FROM natality
                GROUP BY year, state
                ORDER BY year, state ASC
                ''')
sql.show()
```

Resultados

Respuesta

La imagen superior muestra el desarrollo temporal del peso promedio de los bebes nacidos por año y por estado. El promedio mas alto lo encontramos en NH (Nuevo Hampshire) en el año 2004 con unos 8.24 lb. Mientras que el promedio mas bajo, tambien durante el 2004, en ID (Idaho) de al rededor de unos 6.3 lb.

Ejercicio 3

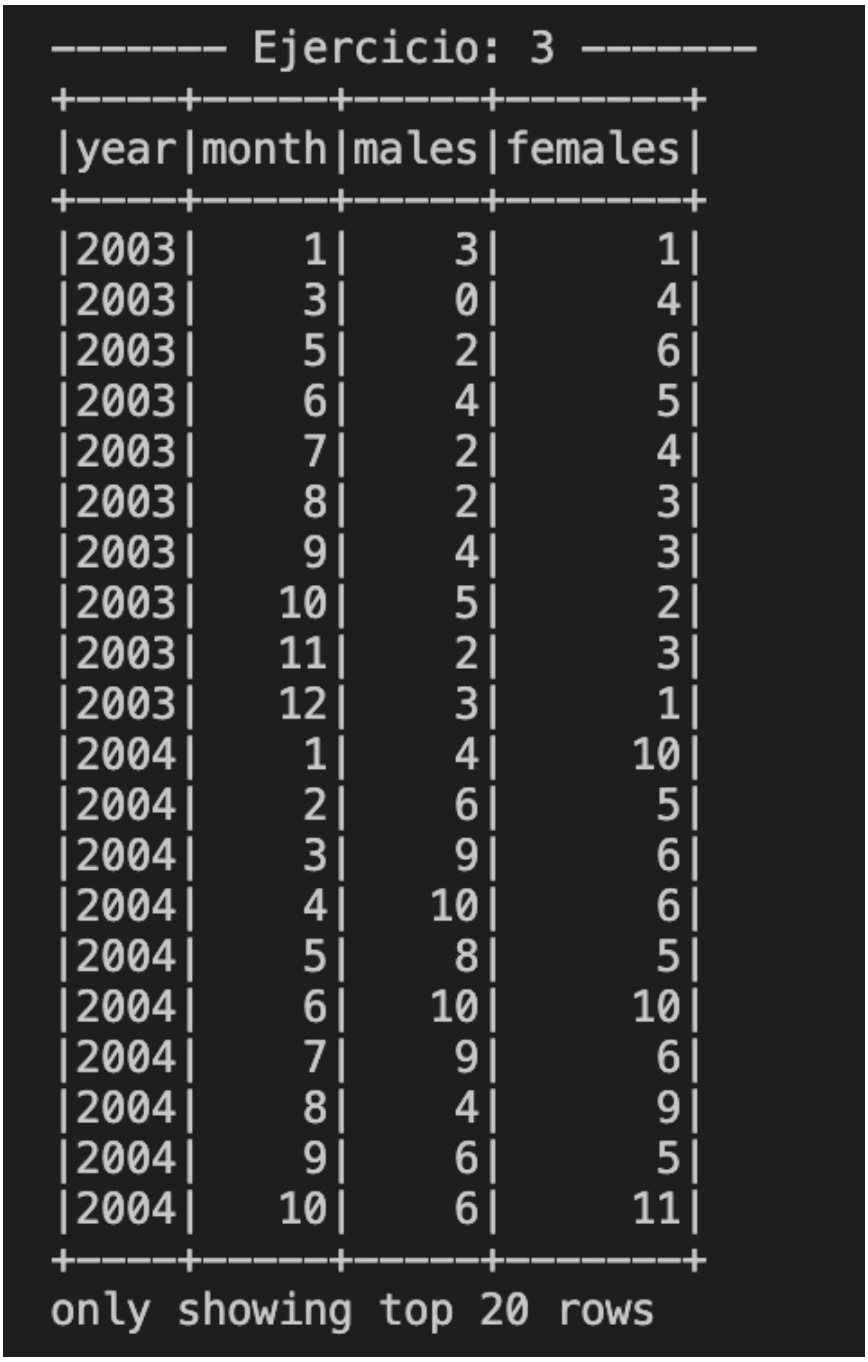
> Evolución por año y por mes del número de niños y niñas nacidas (Resultado por separado con una sola consulta cada registro debe tener 4 columnas: año, mes, numero de niños nacidos, numero de niñas nacidas).

Query

```
    # DataFrame
    df.groupBy(df.year, "month")\
        .agg(
            F.count(F.when(df.is_male, 1)).alias("males"),
            F.count(F.when(df.is_male == "false", 1)).alias("females"))\
        .sort("year", "month")\
        .show()

    # SQL
    df.createOrReplaceTempView("natality")
    sql = spark.sql('''
                    SELECT  year, month,
                            SUM(IF(is_male = true, 1, 0)) AS males,
                            SUM(IF(is_male = true, 0, 1)) AS females
                    FROM natality
                    GROUP BY year, month
                    ORDER BY year, month ASC
                    ''')
    sql.show()
```

Resultados



```
------- Ejercicio: 3 -------
+----+-----+-----+-------+
|year|month|males|females|
+----+-----+-----+-------+
|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|
+----+-----+-----+-------+
only showing top 20 rows
```

Nota: solo se muestran las primeras 20 filas

Respuesta

La imagen superior muestra la evolución mensual de la tasa de natalidad de Hombres y Mujeres. Se percibe un paulatino crecimiento en el registro de información, probablemente dado a causa de la implementación del registro digital y no a una variacion en la cantidad de nacidos promedio.

Ejercicio 4

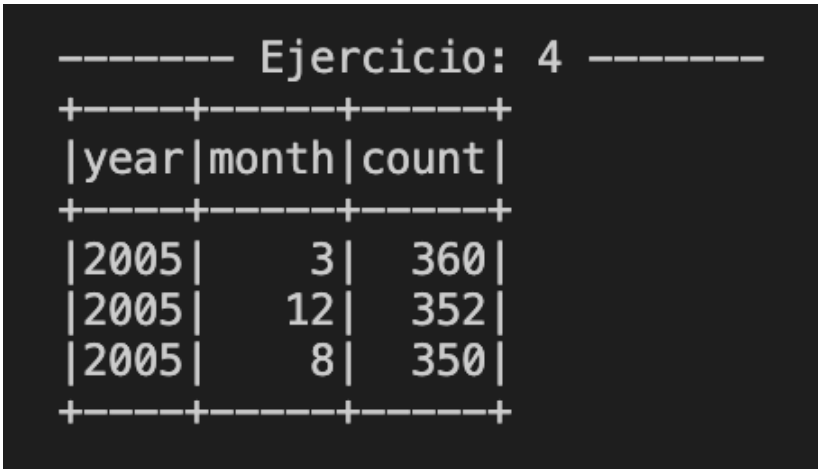
> Obtén los tres meses de 2005 en que nacieron más bebés.

Query

```python
# DataFrame
df.where(F.col("year") == "2005")\
    .groupBy("year", "month")\
    .count()\
    .sort("count", ascending=False)\
    .limit(3).show()

# SQL
df.createOrReplaceTempView("natality")
sql = spark.sql('''
                SELECT year, month, COUNT(*) as cantidad
                FROM natality
                WHERE year = 2005
                GROUP BY year, month
                ORDER BY cantidad DESC
                LIMIT 3
                ''')
sql.show()
```

Resultados



Respuesta

Durante el año 2005 los meses donde mas bebés nacieron fueron Marzo (360), Diciembre (352) y Agosto (350).

Ejercicio 5

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

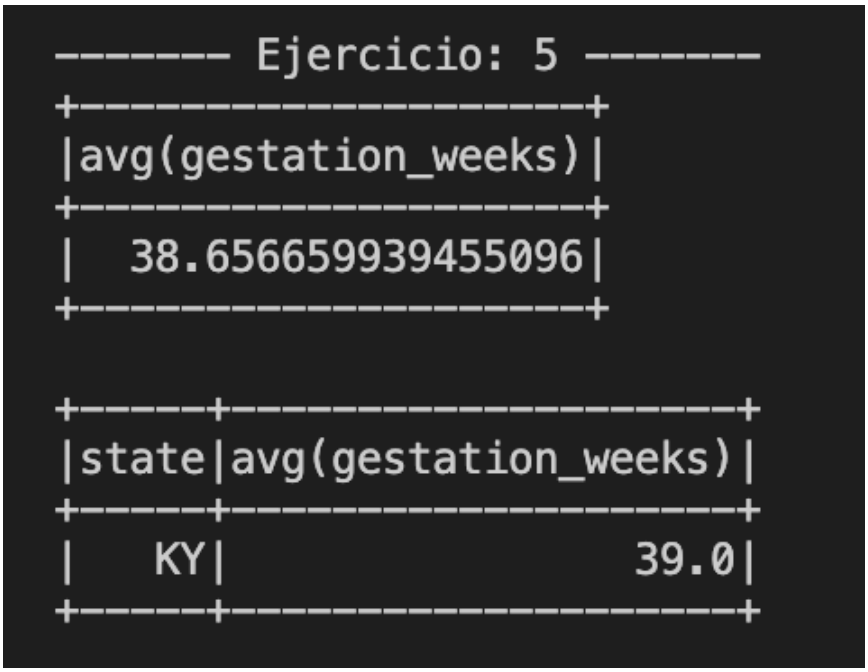
Query

```
# DataFrame
avg_gest = df.select(F.avg("gestation_weeks")).collect()[0][0]

df.groupBy("state")\
    .agg({"gestation_weeks": "avg"})\
    .where(F.col("avg(gestation_weeks)").cast("float") > avg_gest)\
    .show()

# SQL
df.createOrReplaceTempView("natality")
sql = spark.sql('''
                SELECT state, AVG(gestation_weeks)
                FROM natality
                GROUP BY state
                HAVING AVG(gestation_weeks) > (SELECT AVG(gestation_weeks)
                ''')
sql.show()
```

Resultados



Respuesta

Con una media de gestación Nacional de 38.6 semanas, solo el estado de Kentucky (KY) se encuentra por encima de esta media con 39 semanas de gestación.

Ejercicio 6

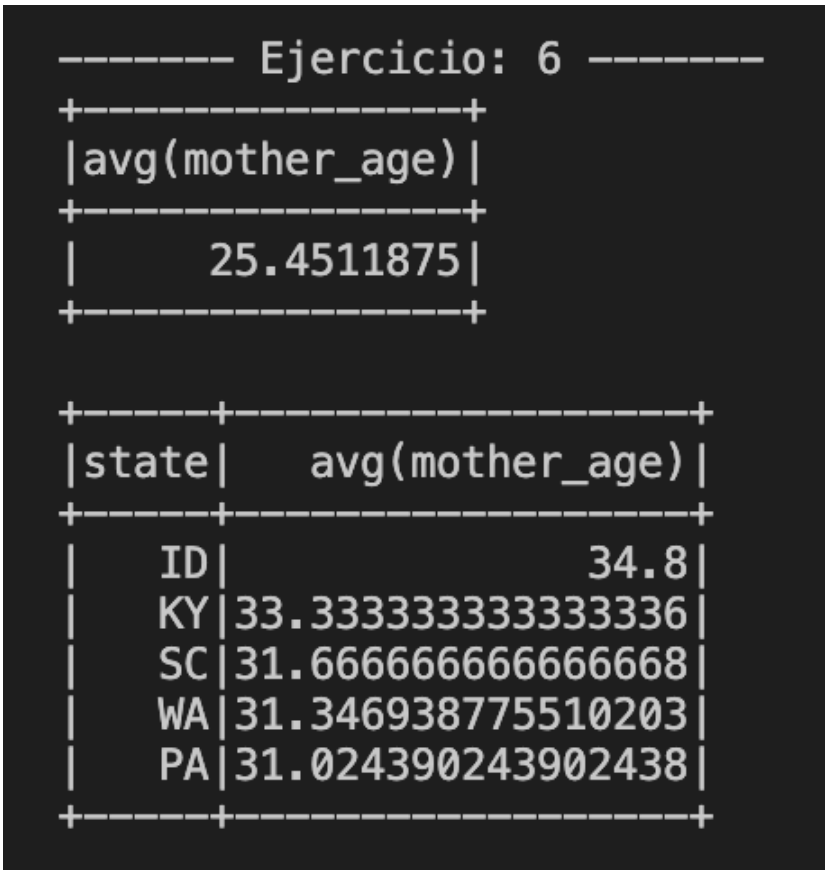
> Obtén los cinco estados donde la media de edad de las madres ha sido mayor.

Query

```python
# DataFrame
df.groupBy("state")\
    .agg({"mother_age": "avg"})\
    .orderBy("avg(mother_age)", ascending=False)\
    .limit(5)\
    .show()

# SQL
df.createOrReplaceTempView("natality")
sql = spark.sql('''
                SELECT state, AVG(mother_age)
                FROM natality
                GROUP BY state
                ORDER BY 2 DESC
                LIMIT 5
                ''')
sql.show()
```

Resultados



Respuesta

Los estado de Idaho (ID), Kentucky (KY), Carolina de Sur (SC), WA (Washington), PA (Pennsylvania) son los 5 estados con mayor promedio de edad de las madres. Idaho lidera el listado con un promedio de casi 35 años.

Ejercicio 7

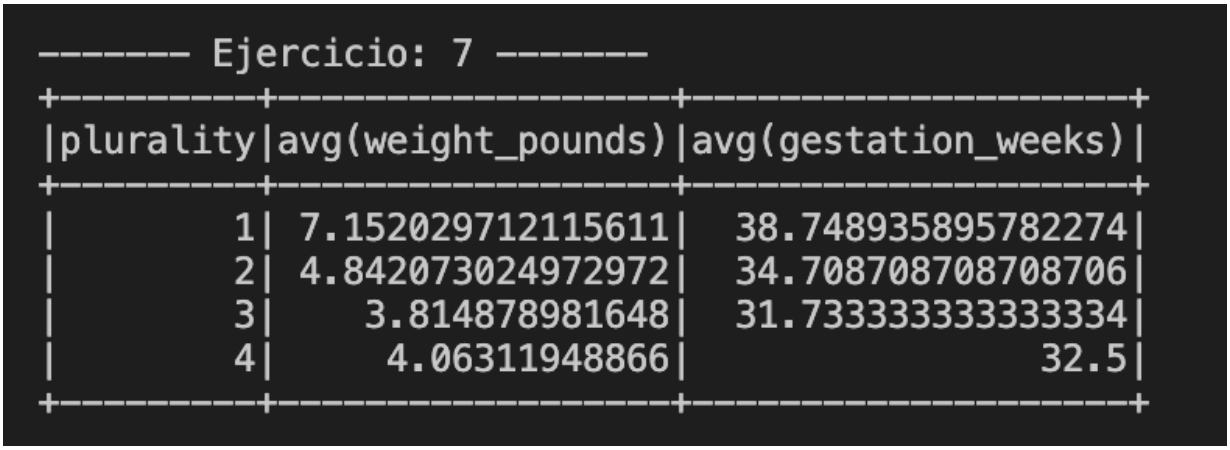
> Indica cómo influye en el peso del bebé y las semanas de gestación que la madre haya tenido un parto múltiple (campo plurality) a las que no lo han tenido.

Query

```
    # DataFrame
    df.groupBy("plurality")\
        .agg({
            "weight_pounds": "avg",
            "gestation_weeks": "avg"
            })\
        .orderBy("plurality")\
        .show()

    # SQL
    df.createOrReplaceTempView("natality")
    sql = spark.sql('''
                    SELECT plurality, AVG(weight_pounds), AVG(gestation_weeks)
                    FROM natality
                    GROUP BY plurality
                    ORDER BY plurality
                    ''')
    sql.show()
```

Resultados



```
------- Ejercicio: 7 -------
+---------+------------------+--------------------+
|plurality|avg(weight_pounds)|avg(gestation_weeks)|
+---------+------------------+--------------------+
|        1| 7.152029712115611|  38.748935895782274|
|        2| 4.842073024972972|  34.708708708708706|
|        3|    3.814878981648|  31.733333333333334|
|        4|     4.06311948866|                32.5|
+---------+------------------+--------------------+
```

Respuesta

Se puede apreciar en los resultados una correlación entre la cantidad de bebés nacientes en el parto, con el peso de cada individuo y con la cantidad de semanas de gestación. Los resultados indican que un parto doble implica en promedio que cada bebé pese un 32% menos de lo que pesa en un bebe nacido en parto simple, además que de media el parto doble acarrea 1 mes menos de gestación. Mientras que si vamos al caso de trillizos, estos pesan en promedio un 53% menos cada uno que un bebé nacido en parto simple y llevan casi 2 meses menos de gestación que el caso simple.