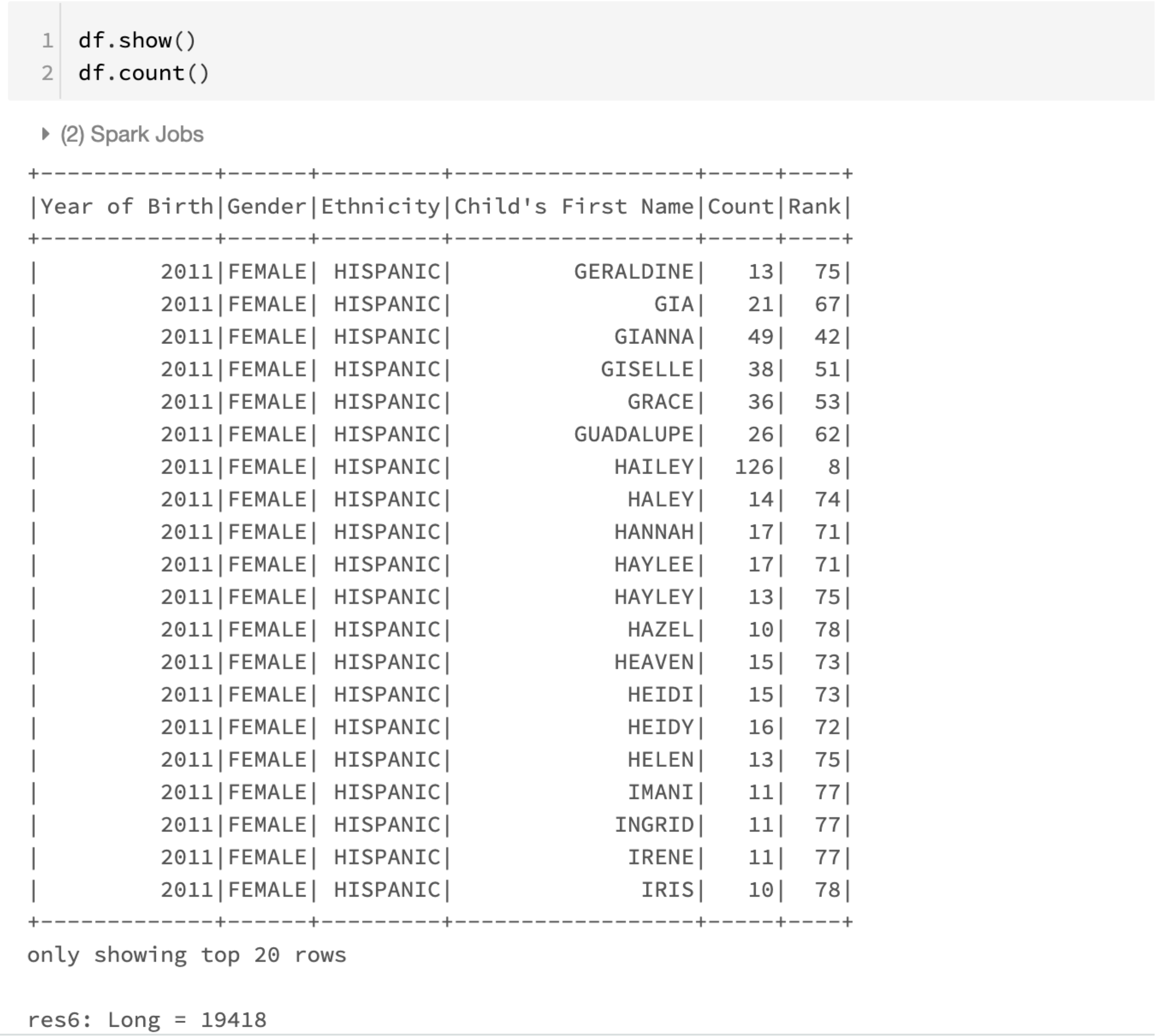
Name of the file in Databricks: Assign_1_2020

1) Read file:
import org.apache.spark.sql.functions._
val df =
spark.read.option("inferSchema","true").option("header","true").csv("dbfs:/autumn_2019/pava/Popular_Baby_Names.csv")

2) Count rows and show table –



```
df.show()
df.count()
```

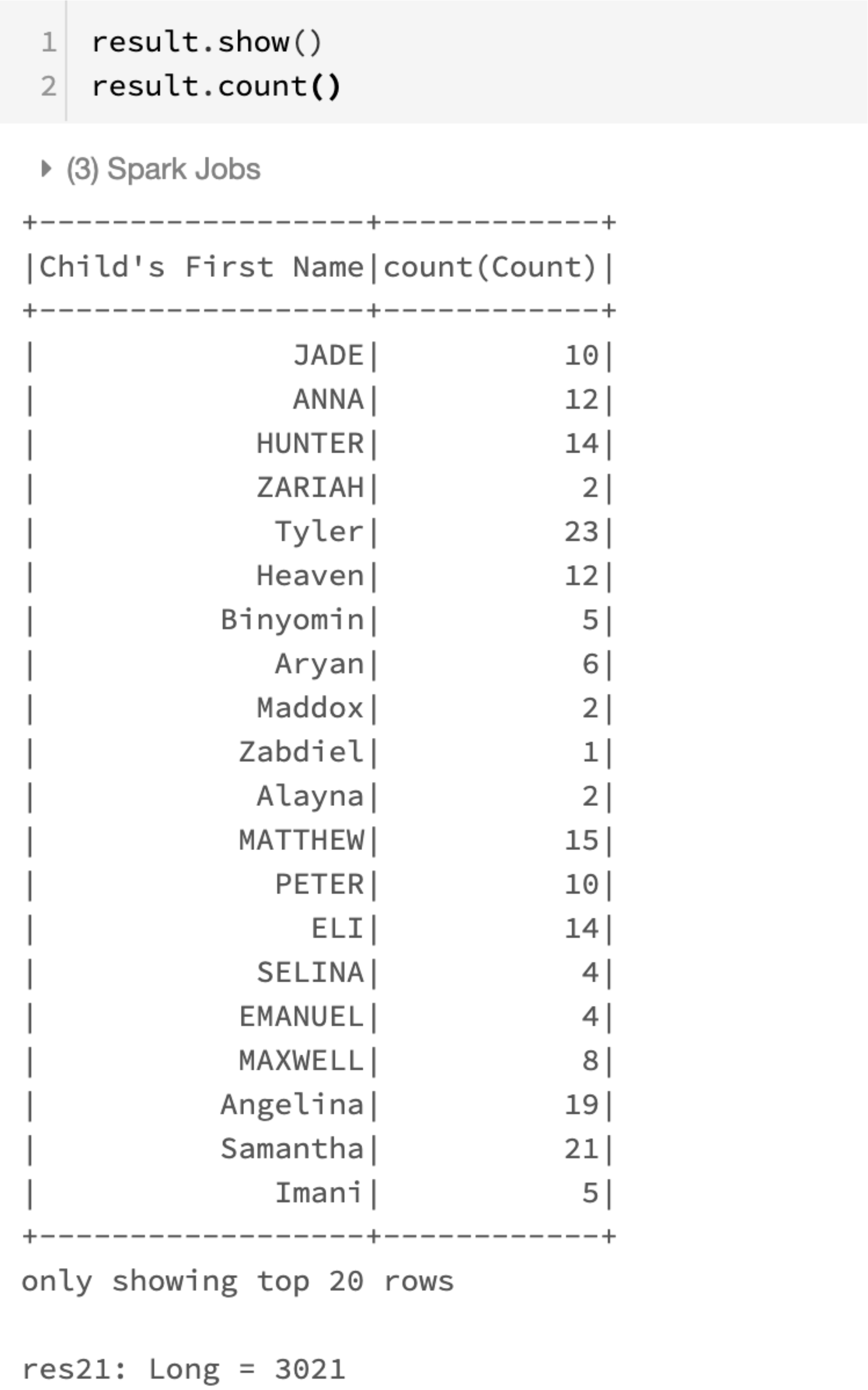
▸ (2) Spark Jobs

```
+-------------+------+---------+-----------------+-----+----+
|Year of Birth|Gender|Ethnicity|Child's First Name|Count|Rank|
+-------------+------+---------+-----------------+-----+----+
|         2011|FEMALE| HISPANIC|        GERALDINE|   13|  75|
|         2011|FEMALE| HISPANIC|              GIA|   21|  67|
|         2011|FEMALE| HISPANIC|           GIANNA|   49|  42|
|         2011|FEMALE| HISPANIC|          GISELLE|   38|  51|
|         2011|FEMALE| HISPANIC|            GRACE|   36|  53|
|         2011|FEMALE| HISPANIC|        GUADALUPE|   26|  62|
|         2011|FEMALE| HISPANIC|           HAILEY|  126|   8|
|         2011|FEMALE| HISPANIC|            HALEY|   14|  74|
|         2011|FEMALE| HISPANIC|           HANNAH|   17|  71|
|         2011|FEMALE| HISPANIC|           HAYLEE|   17|  71|
|         2011|FEMALE| HISPANIC|           HAYLEY|   13|  75|
|         2011|FEMALE| HISPANIC|            HAZEL|   10|  78|
|         2011|FEMALE| HISPANIC|           HEAVEN|   15|  73|
|         2011|FEMALE| HISPANIC|            HEIDI|   15|  73|
|         2011|FEMALE| HISPANIC|            HEIDY|   16|  72|
|         2011|FEMALE| HISPANIC|            HELEN|   13|  75|
|         2011|FEMALE| HISPANIC|            IMANI|   11|  77|
|         2011|FEMALE| HISPANIC|           INGRID|   11|  77|
|         2011|FEMALE| HISPANIC|            IRENE|   11|  77|
|         2011|FEMALE| HISPANIC|             IRIS|   10|  78|
+-------------+------+---------+-----------------+-----+----+
only showing top 20 rows

res6: Long = 19418
```

3) Aggregate name and count –
val result=df.groupBy("Child's First Name").agg(count("Count"))

4) Show aggregated value –



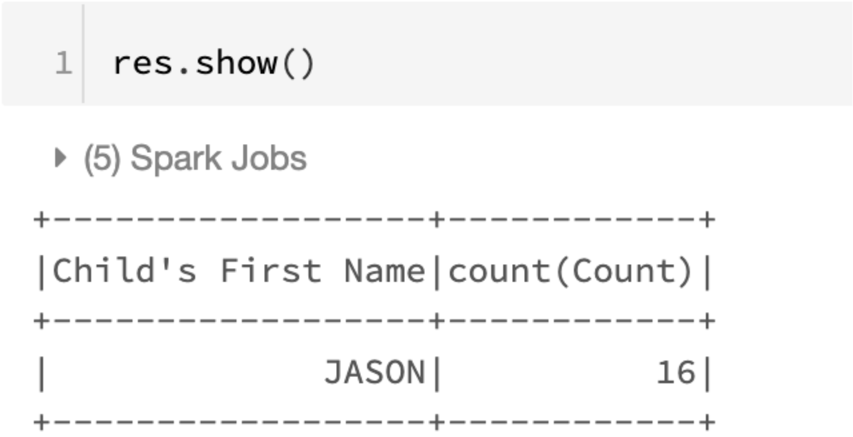
```
result.show()
result.count()
```

```
+------------------+------------+
|Child's First Name|count(Count)|
+------------------+------------+
|              JADE|          10|
|              ANNA|          12|
|            HUNTER|          14|
|            ZARIAH|           2|
|             Tyler|          23|
|            Heaven|          12|
|          Binyomin|           5|
|             Aryan|           6|
|            Maddox|           2|
|           Zabdiel|           1|
|            Alayna|           2|
|           MATTHEW|          15|
|             PETER|          10|
|               ELI|          14|
|            SELINA|           4|
|           EMANUEL|           4|
|           MAXWELL|           8|
|          Angelina|          19|
|          Samantha|          21|
|             Imani|           5|
+------------------+------------+
only showing top 20 rows

res21: Long = 3021
```

5) How frequent is my name?

val res=result.filter($"Child's First Name" === "JASON")



```
res.show()
```

```
+------------------+------------+
|Child's First Name|count(Count)|
+------------------+------------+
|             JASON|          16|
+------------------+------------+
```