Week 5

Felipe Rodriguez

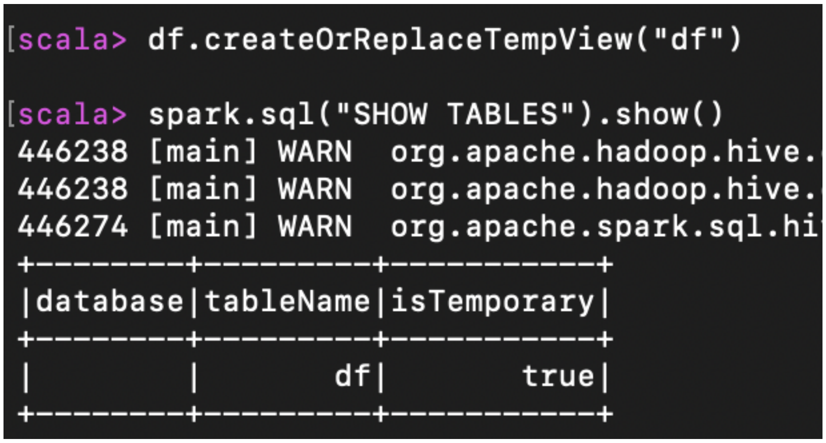
Bellevue University

DSC 650 Big Data

Professor Nasheb Ismaily

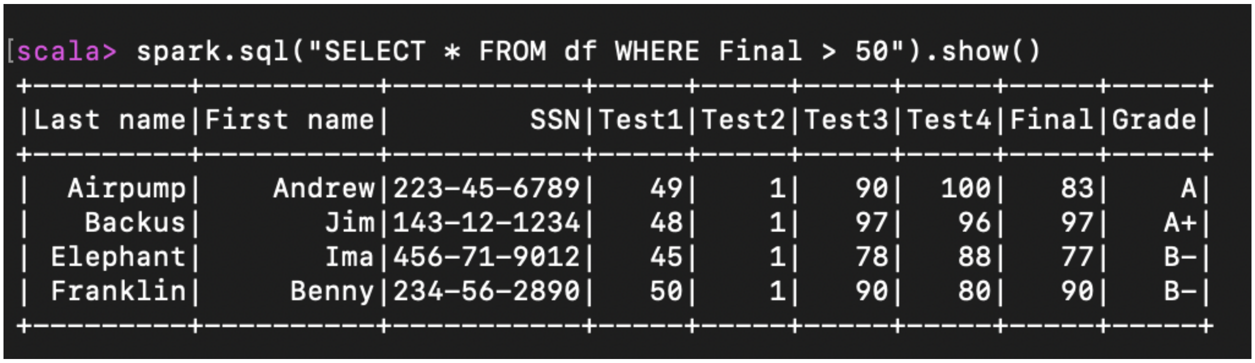
January 13, 2024

SparkSQL commands in Scala

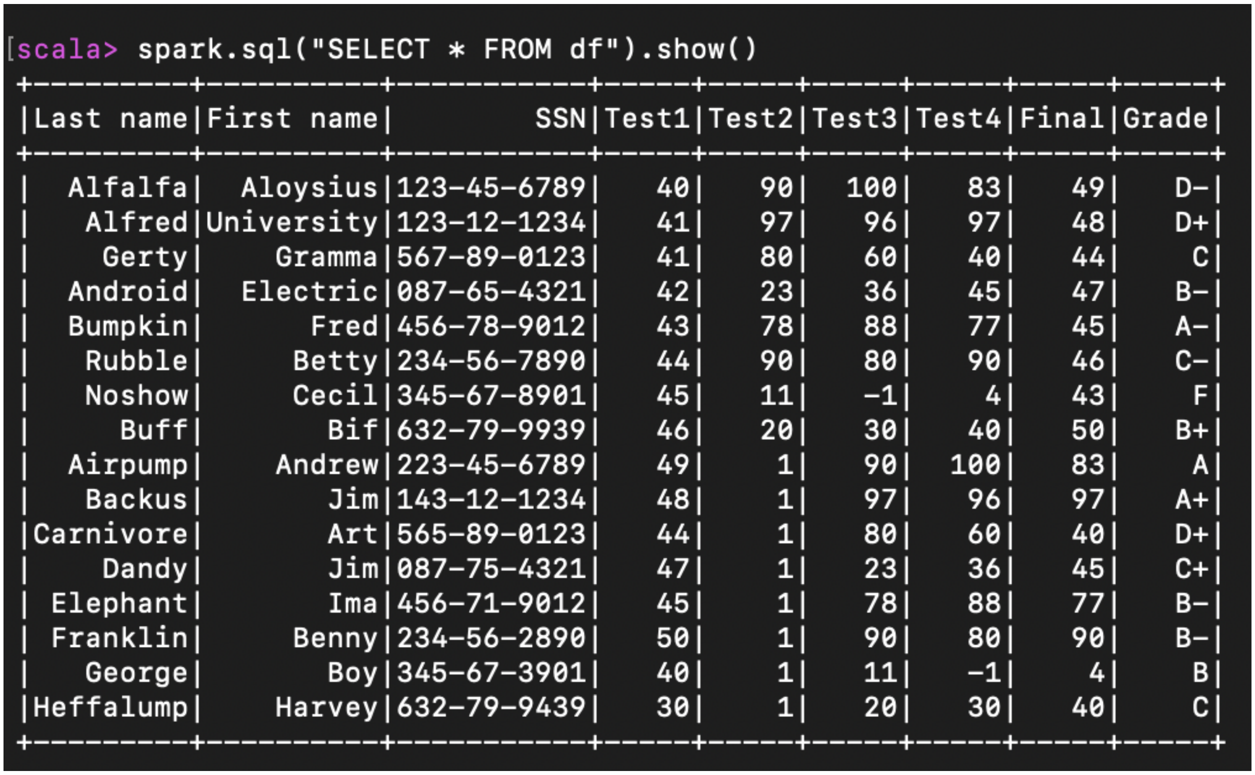


```
[scala> df.createOrReplaceTempView("df")

[scala> spark.sql("SHOW TABLES").show()
446238 [main] WARN  org.apache.hadoop.hive.
446238 [main] WARN  org.apache.hadoop.hive.
446274 [main] WARN  org.apache.spark.sql.hi
+--------+---------+-----------+
|database|tableName|isTemporary|
+--------+---------+-----------+
|        |       df|       true|
+--------+---------+-----------+
```



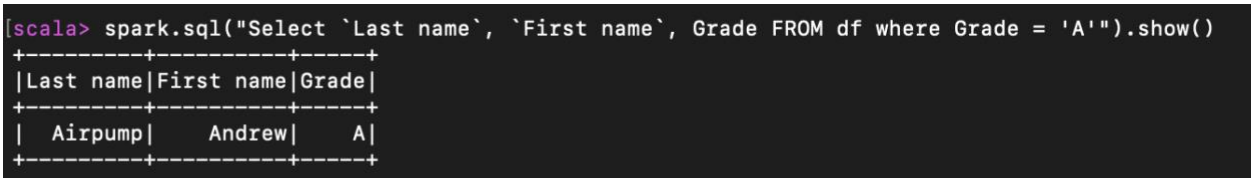
```
[scala> spark.sql("SELECT * FROM df WHERE Final > 50").show()
+---------+----------+-----------+-----+-----+-----+-----+-----+-----+
|Last name|First name|        SSN|Test1|Test2|Test3|Test4|Final|Grade|
+---------+----------+-----------+-----+-----+-----+-----+-----+-----+
|  Airpump|    Andrew|223-45-6789|   49|    1|   90|  100|   83|    A|
|   Backus|       Jim|143-12-1234|   48|    1|   97|   96|   97|   A+|
| Elephant|       Ima|456-71-9012|   45|    1|   78|   88|   77|   B-|
| Franklin|     Benny|234-56-2890|   50|    1|   90|   80|   90|   B-|
+---------+----------+-----------+-----+-----+-----+-----+-----+-----+
```



```
[scala> spark.sql("SELECT * FROM df").show()
+---------+----------+-----------+-----+-----+-----+-----+-----+-----+
|Last name|First name|        SSN|Test1|Test2|Test3|Test4|Final|Grade|
+---------+----------+-----------+-----+-----+-----+-----+-----+-----+
|  Alfalfa|  Aloysius|123-45-6789|   40|   90|  100|   83|   49|   D-|
|   Alfred|University|123-12-1234|   41|   97|   96|   97|   48|   D+|
|    Gerty|    Gramma|567-89-0123|   41|   80|   60|   40|   44|    C|
|  Android|  Electric|087-65-4321|   42|   23|   36|   45|   47|   B-|
|  Bumpkin|      Fred|456-78-9012|   43|   78|   88|   77|   45|   A-|
|   Rubble|     Betty|234-56-7890|   44|   90|   80|   90|   46|   C-|
|   Noshow|     Cecil|345-67-8901|   45|   11|   -1|    4|   43|    F|
|     Buff|       Bif|632-79-9939|   46|   20|   30|   40|   50|   B+|
|  Airpump|    Andrew|223-45-6789|   49|    1|   90|  100|   83|    A|
|   Backus|       Jim|143-12-1234|   48|    1|   97|   96|   97|   A+|
|Carnivore|       Art|565-89-0123|   44|    1|   80|   60|   40|   D+|
|    Dandy|       Jim|087-75-4321|   47|    1|   23|   36|   45|   C+|
| Elephant|       Ima|456-71-9012|   45|    1|   78|   88|   77|   B-|
| Franklin|     Benny|234-56-2890|   50|    1|   90|   80|   90|   B-|
|   George|       Boy|345-67-3901|   40|    1|   11|   -1|    4|    B|
|Heffalump|    Harvey|632-79-9439|   30|    1|   20|   30|   40|    C|
+---------+----------+-----------+-----+-----+-----+-----+-----+-----+
```

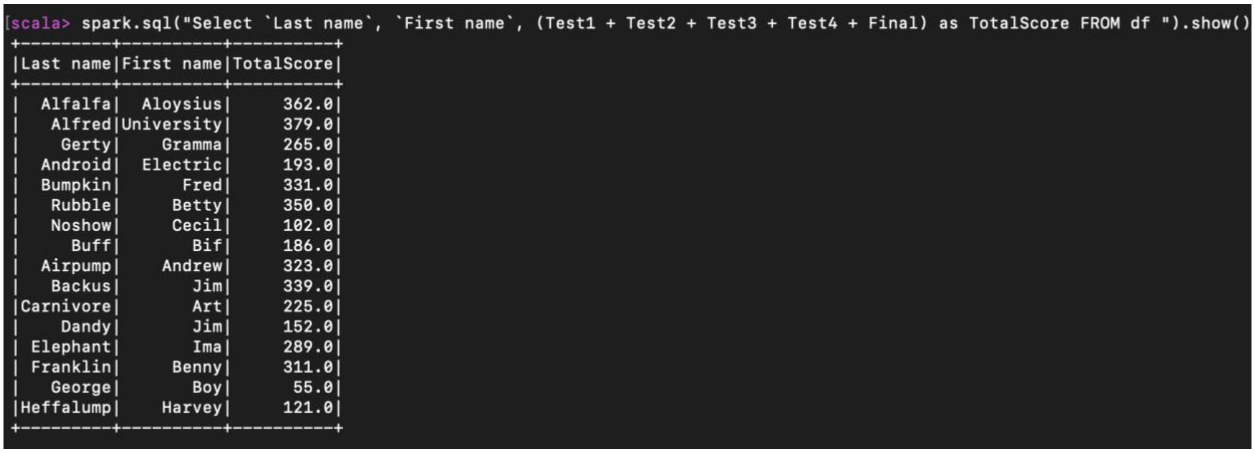
3 Spark SQL Queries in Scala

Query 1



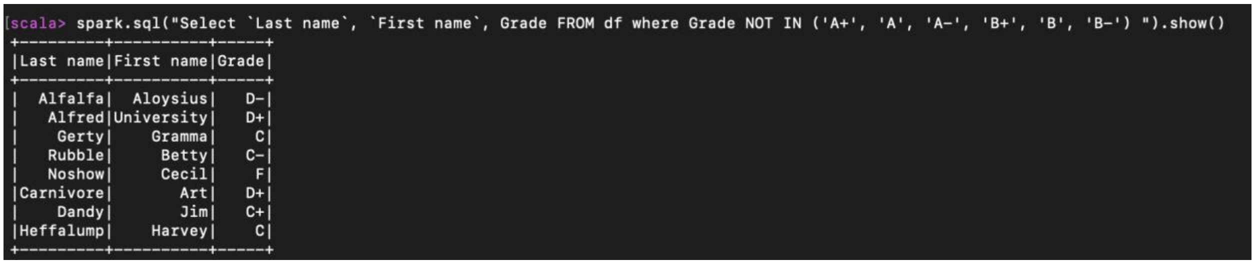
```
[scala> spark.sql("Select `Last name`, `First name`, Grade FROM df where Grade = 'A'").show()
+---------+----------+-----+
|Last name|First name|Grade|
+---------+----------+-----+
|  Airpump|    Andrew|    A|
+---------+----------+-----+
```

Query 2



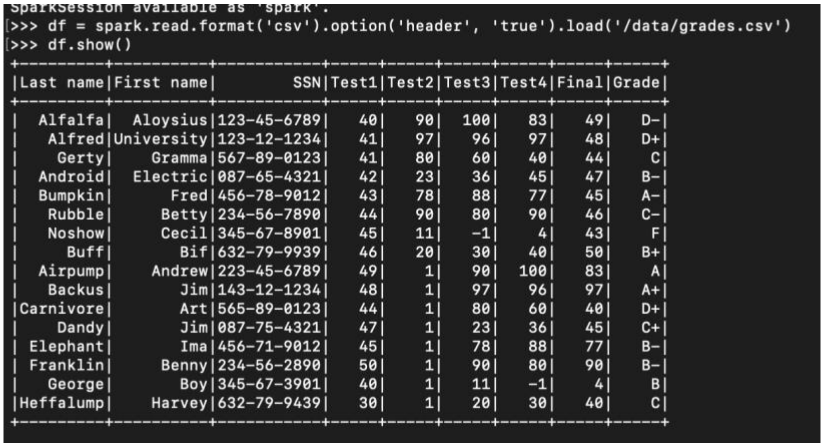
```
[scala> spark.sql("Select `Last name`, `First name`, (Test1 + Test2 + Test3 + Test4 + Final) as TotalScore FROM df ").show()
+---------+----------+----------+
|Last name|First name|TotalScore|
+---------+----------+----------+
|  Alfalfa|  Aloysius|     362.0|
|   Alfred|University|     379.0|
|    Gerty|    Gramma|     265.0|
|  Android|  Electric|     193.0|
|  Bumpkin|      Fred|     331.0|
|   Rubble|     Betty|     350.0|
|   Noshow|     Cecil|     102.0|
|     Buff|       Bif|     186.0|
|  Airpump|    Andrew|     323.0|
|   Backus|       Jim|     339.0|
|Carnivore|       Art|     225.0|
|    Dandy|       Jim|     152.0|
| Elephant|       Ima|     289.0|
| Franklin|     Benny|     311.0|
|   George|       Boy|      55.0|
|Heffalump|    Harvey|     121.0|
+---------+----------+----------+
```

Query 3

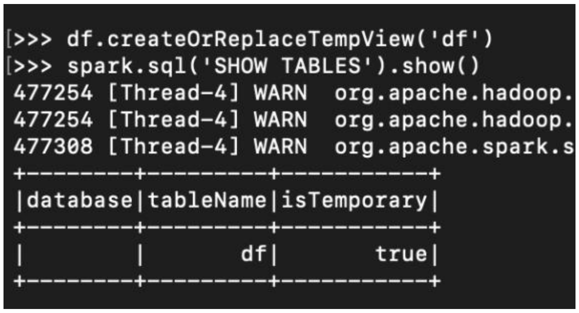


```
[scala> spark.sql("Select `Last name`, `First name`, Grade FROM df where Grade NOT IN ('A+', 'A', 'A-', 'B+', 'B', 'B-') ").show()
+---------+----------+-----+
|Last name|First name|Grade|
+---------+----------+-----+
|  Alfalfa|  Aloysius|   D-|
|   Alfred|University|   D+|
|    Gerty|    Gramma|    C|
|   Rubble|     Betty|   C-|
|   Noshow|     Cecil|    F|
|Carnivore|       Art|   D+|
|    Dandy|       Jim|   C+|
|Heffalump|    Harvey|    C|
+---------+----------+-----+
```

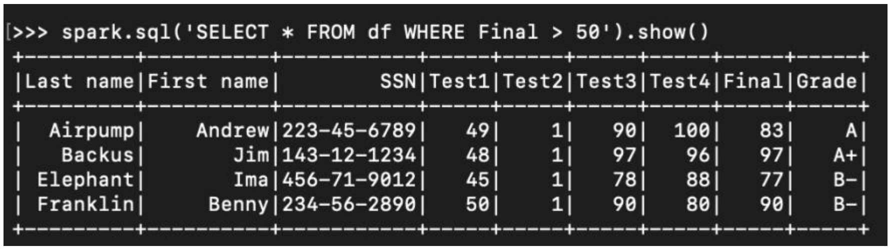
SparkSQL with Python (PySpark)



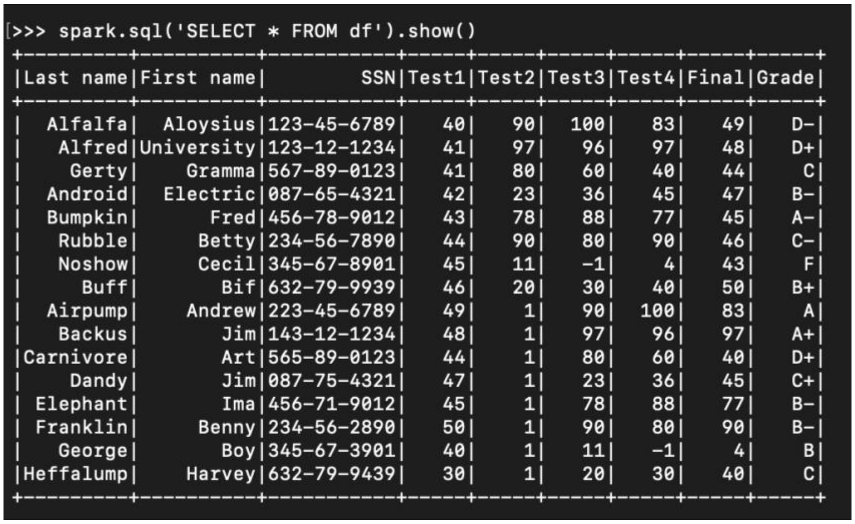
```
SparkSession available as 'spark'.
[>>> df = spark.read.format('csv').option('header', 'true').load('/data/grades.csv')
[>>> df.show()
+---------+----------+-----------+-----+-----+-----+-----+-----+-----+
|Last name|First name|        SSN|Test1|Test2|Test3|Test4|Final|Grade|
+---------+----------+-----------+-----+-----+-----+-----+-----+-----+
|  Alfalfa|  Aloysius|123-45-6789|   40|   90|  100|   83|   49|   D-|
|   Alfred|University|123-12-1234|   41|   97|   96|   97|   48|   D+|
|    Gerty|    Gramma|567-89-0123|   41|   80|   60|   40|   44|    C|
|  Android|  Electric|087-65-4321|   42|   23|   36|   45|   47|   B-|
|  Bumpkin|      Fred|456-78-9012|   43|   78|   88|   77|   45|   A-|
|   Rubble|     Betty|234-56-7890|   44|   90|   80|   90|   46|   C-|
|   Noshow|     Cecil|345-67-8901|   45|   11|   -1|    4|   43|    F|
|     Buff|       Bif|632-79-9939|   46|   20|   30|   40|   50|   B+|
|  Airpump|    Andrew|223-45-6789|   49|    1|   90|  100|   83|    A|
|   Backus|       Jim|143-12-1234|   48|    1|   97|   96|   97|   A+|
|Carnivore|       Art|565-89-0123|   44|    1|   80|   60|   40|   D+|
|    Dandy|       Jim|087-75-4321|   47|    1|   23|   36|   45|   C+|
| Elephant|       Ima|456-71-9012|   45|    1|   78|   88|   77|   B-|
| Franklin|     Benny|234-56-2890|   50|    1|   90|   80|   90|   B-|
|   George|       Boy|345-67-3901|   40|    1|   11|   -1|    4|    B|
|Heffalump|    Harvey|632-79-9439|   30|    1|   20|   30|   40|    C|
+---------+----------+-----------+-----+-----+-----+-----+-----+-----+
```



```
[>>> df.createOrReplaceTempView('df')
[>>> spark.sql('SHOW TABLES').show()
477254 [Thread-4] WARN  org.apache.hadoop.
477254 [Thread-4] WARN  org.apache.hadoop.
477308 [Thread-4] WARN  org.apache.spark.s
+--------+---------+-----------+
|database|tableName|isTemporary|
+--------+---------+-----------+
|        |       df|       true|
+--------+---------+-----------+
```



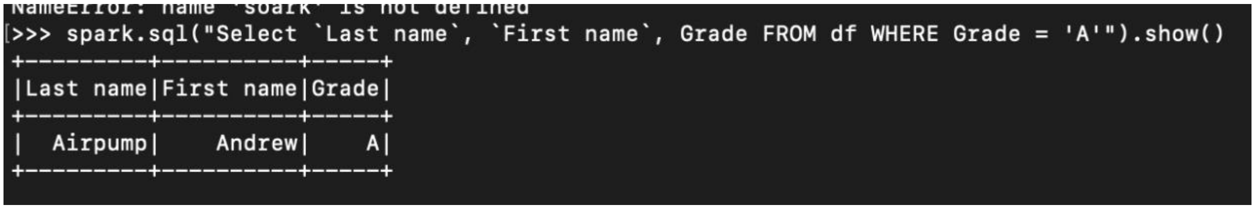
```
[>>> spark.sql('SELECT * FROM df WHERE Final > 50').show()
+---------+----------+-----------+-----+-----+-----+-----+-----+-----+
|Last name|First name|        SSN|Test1|Test2|Test3|Test4|Final|Grade|
+---------+----------+-----------+-----+-----+-----+-----+-----+-----+
|  Airpump|    Andrew|223-45-6789|   49|    1|   90|  100|   83|    A|
|   Backus|       Jim|143-12-1234|   48|    1|   97|   96|   97|   A+|
| Elephant|       Ima|456-71-9012|   45|    1|   78|   88|   77|   B-|
| Franklin|     Benny|234-56-2890|   50|    1|   90|   80|   90|   B-|
+---------+----------+-----------+-----+-----+-----+-----+-----+-----+
```



```
[>>> spark.sql('SELECT * FROM df').show()
+---------+----------+-----------+-----+-----+-----+-----+-----+-----+
|Last name|First name|        SSN|Test1|Test2|Test3|Test4|Final|Grade|
+---------+----------+-----------+-----+-----+-----+-----+-----+-----+
|  Alfalfa|  Aloysius|123-45-6789|   40|   90|  100|   83|   49|   D-|
|   Alfred|University|123-12-1234|   41|   97|   96|   97|   48|   D+|
|    Gerty|    Gramma|567-89-0123|   41|   80|   60|   40|   44|    C|
|  Android|  Electric|087-65-4321|   42|   23|   36|   45|   47|   B-|
|  Bumpkin|      Fred|456-78-9012|   43|   78|   88|   77|   45|   A-|
|   Rubble|     Betty|234-56-7890|   44|   90|   80|   90|   46|   C-|
|   Noshow|     Cecil|345-67-8901|   45|   11|   -1|    4|   43|    F|
|     Buff|       Bif|632-79-9939|   46|   20|   30|   40|   50|   B+|
|  Airpump|    Andrew|223-45-6789|   49|    1|   90|  100|   83|    A|
|   Backus|       Jim|143-12-1234|   48|    1|   97|   96|   97|   A+|
|Carnivore|       Art|565-89-0123|   44|    1|   80|   60|   40|   D+|
|    Dandy|       Jim|087-75-4321|   47|    1|   23|   36|   45|   C+|
| Elephant|       Ima|456-71-9012|   45|    1|   78|   88|   77|   B-|
| Franklin|     Benny|234-56-2890|   50|    1|   90|   80|   90|   B-|
|   George|       Boy|345-67-3901|   40|    1|   11|   -1|    4|    B|
|Heffalump|    Harvey|632-79-9439|   30|    1|   20|   30|   40|    C|
+---------+----------+-----------+-----+-----+-----+-----+-----+-----+
```

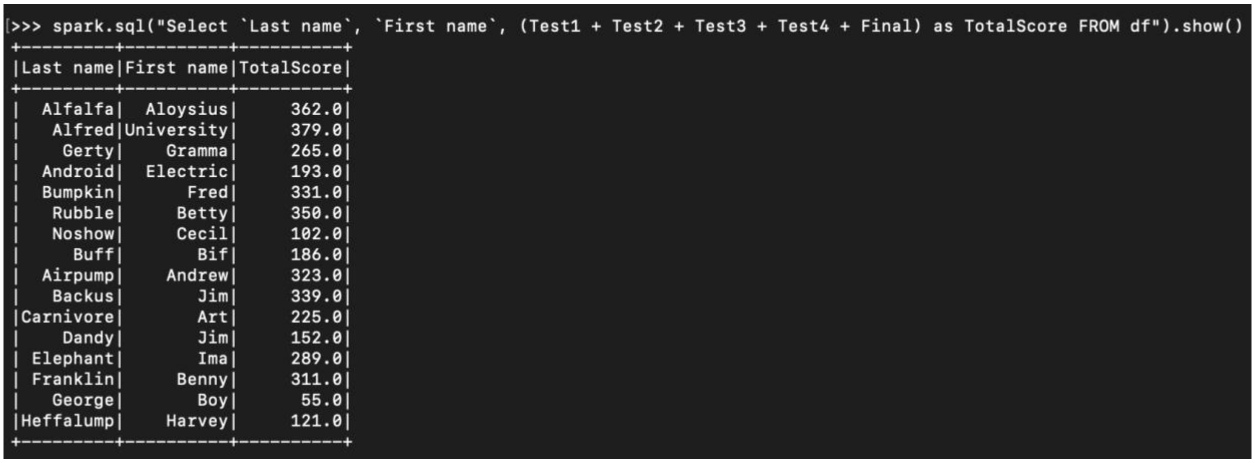
3 Spark SQL Queries in PySpark

Query 1



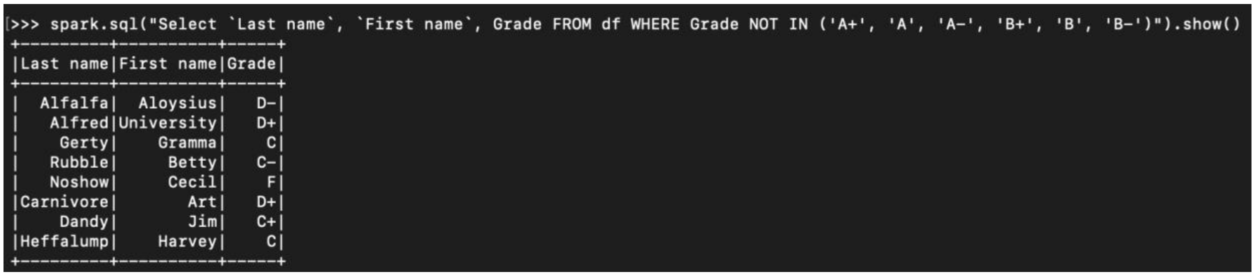
```
NameError: name 'soark' is not defined
>>> spark.sql("Select `Last name`, `First name`, Grade FROM df WHERE Grade = 'A'").show()
+---------+----------+-----+
|Last name|First name|Grade|
+---------+----------+-----+
|  Airpump|    Andrew|    A|
+---------+----------+-----+
```

Query 2



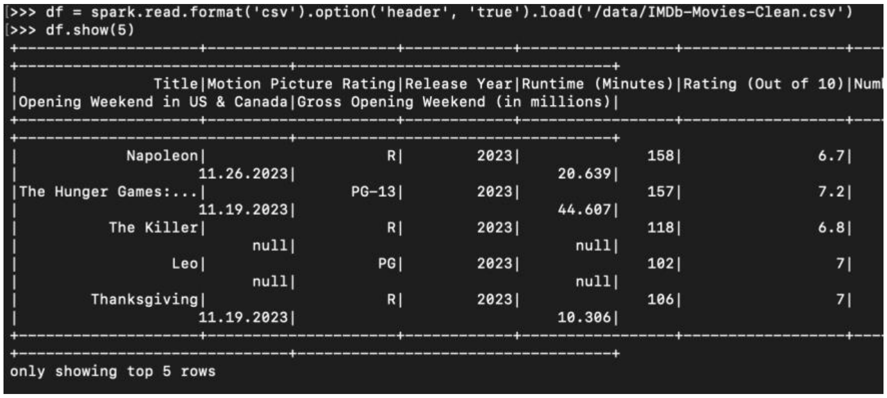
```
>>> spark.sql("Select `Last name`, `First name`, (Test1 + Test2 + Test3 + Test4 + Final) as TotalScore FROM df").show()
+---------+----------+----------+
|Last name|First name|TotalScore|
+---------+----------+----------+
|  Alfalfa|  Aloysius|     362.0|
|   Alfred|University|     379.0|
|    Gerty|    Gramma|     265.0|
|  Android|  Electric|     193.0|
|  Bumpkin|      Fred|     331.0|
|   Rubble|     Betty|     350.0|
|   Noshow|     Cecil|     102.0|
|     Buff|       Bif|     186.0|
|  Airpump|    Andrew|     323.0|
|   Backus|       Jim|     339.0|
|Carnivore|       Art|     225.0|
|    Dandy|       Jim|     152.0|
| Elephant|       Ima|     289.0|
| Franklin|     Benny|     311.0|
|   George|       Boy|      55.0|
|Heffalump|    Harvey|     121.0|
+---------+----------+----------+
```

Query 3



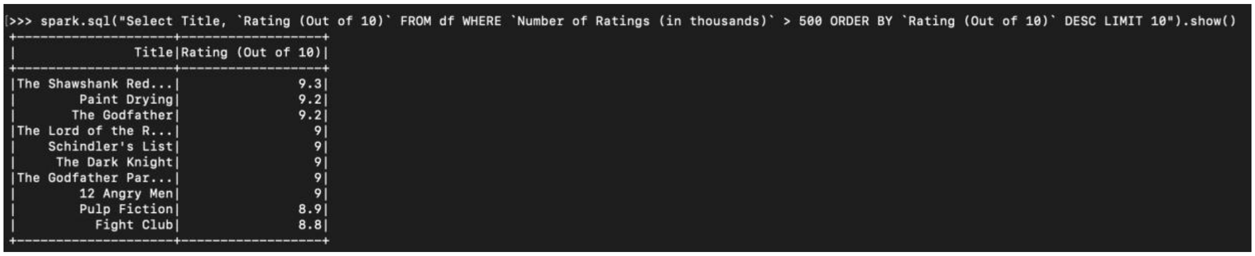
```
>>> spark.sql("Select `Last name`, `First name`, Grade FROM df WHERE Grade NOT IN ('A+', 'A', 'A-', 'B+', 'B', 'B-')").show()
+---------+----------+-----+
|Last name|First name|Grade|
+---------+----------+-----+
|  Alfalfa|  Aloysius|   D-|
|   Alfred|University|   D+|
|    Gerty|    Gramma|    C|
|   Rubble|     Betty|   C-|
|   Noshow|     Cecil|    F|
|Carnivore|       Art|   D+|
|    Dandy|       Jim|   C+|
|Heffalump|    Harvey|    C|
+---------+----------+-----+
```

PySpark Code



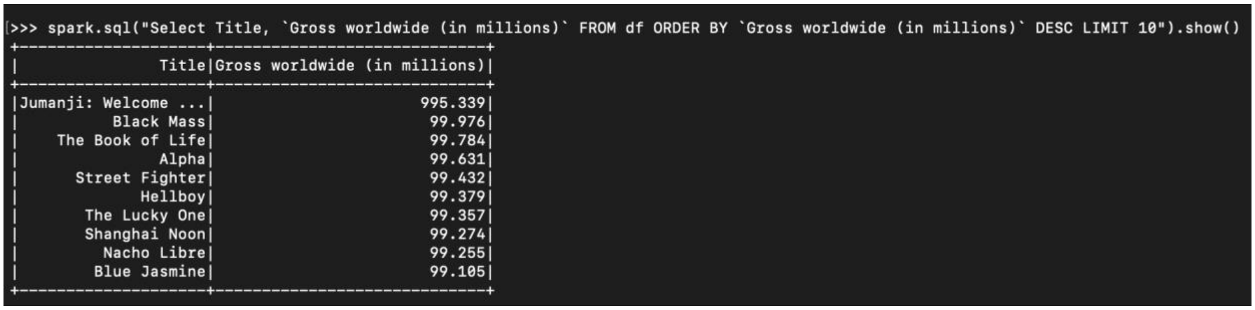
```
>>> df = spark.read.format('csv').option('header', 'true').load('/data/IMDb-Movies-Clean.csv')
>>> df.show(5)
+--------------------+---------------------+------------+-----------------+------------------+---
+------------------------------+-----------------------------------+
|               Title|Motion Picture Rating|Release Year|Runtime (Minutes)|Rating (Out of 10)|Num
|Opening Weekend in US & Canada|Gross Opening Weekend (in millions)|
+--------------------+---------------------+------------+-----------------+------------------+---
+------------------------------+-----------------------------------+
|            Napoleon|                    R|        2023|              158|               6.7|
|                    11.26.2023|                             20.639|
|The Hunger Games:...|                PG-13|        2023|              157|               7.2|
|                    11.19.2023|                             44.607|
|          The Killer|                    R|        2023|              118|               6.8|
|                          null|                               null|
|                 Leo|                   PG|        2023|              102|                 7|
|                          null|                               null|
|        Thanksgiving|                    R|        2023|              106|                 7|
|                    11.19.2023|                             10.306|
+--------------------+---------------------+------------+-----------------+------------------+---
+------------------------------+-----------------------------------+
only showing top 5 rows
```

Query 1



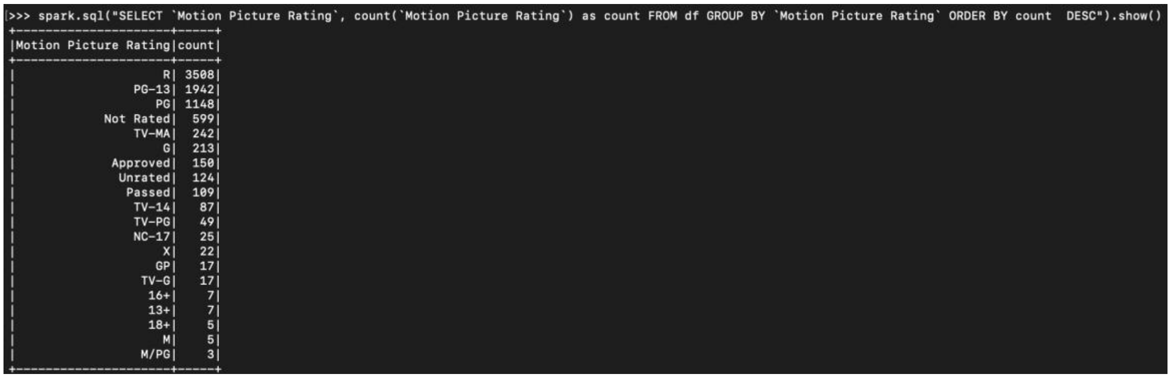
```
>>> spark.sql("Select Title, `Rating (Out of 10)` FROM df WHERE `Number of Ratings (in thousands)` > 500 ORDER BY `Rating (Out of 10)` DESC LIMIT 10").show()
+--------------------+------------------+
|               Title|Rating (Out of 10)|
+--------------------+------------------+
|The Shawshank Red...|               9.3|
|        Paint Drying|               9.2|
|       The Godfather|               9.2|
|The Lord of the R...|                 9|
|    Schindler's List|                 9|
|     The Dark Knight|                 9|
|The Godfather Par...|                 9|
|        12 Angry Men|                 9|
|        Pulp Fiction|               8.9|
|          Fight Club|               8.8|
+--------------------+------------------+
```

Query 2



```
>>> spark.sql("Select Title, `Gross worldwide (in millions)` FROM df ORDER BY `Gross worldwide (in millions)` DESC LIMIT 10").show()
+--------------------+-----------------------------+
|               Title|Gross worldwide (in millions)|
+--------------------+-----------------------------+
|Jumanji: Welcome ...|                      995.339|
|          Black Mass|                       99.976|
|    The Book of Life|                       99.784|
|               Alpha|                       99.631|
|      Street Fighter|                       99.432|
|             Hellboy|                       99.379|
|       The Lucky One|                       99.357|
|       Shanghai Noon|                       99.274|
|         Nacho Libre|                       99.255|
|        Blue Jasmine|                       99.105|
+--------------------+-----------------------------+
```

Query 3



```
>>> spark.sql("SELECT `Motion Picture Rating`, count(`Motion Picture Rating`) as count FROM df GROUP BY `Motion Picture Rating` ORDER BY count  DESC").show()
+---------------------+-----+
|Motion Picture Rating|count|
+---------------------+-----+
|                    R| 3508|
|                PG-13| 1942|
|                   PG| 1148|
|            Not Rated|  599|
|                TV-MA|  242|
|                    G|  213|
|             Approved|  150|
|              Unrated|  124|
|               Passed|  109|
|                TV-14|   87|
|                TV-PG|   49|
|                NC-17|   25|
|                    X|   22|
|                   GP|   17|
|                 TV-G|   17|
|                  16+|    7|
|                  13+|    7|
|                  18+|    5|
|                    M|    5|
|                 M/PG|    3|
+---------------------+-----+
```