Himanshu Goyal

Intro to Spark- Assignment 4

Spark SQL

Q 1) Find birth country which has highest amount of people

```
In [1]: from pyspark import SQLContext
sqlContext = SQLContext(sc)
fake_rdd=sc.textFile("Fake_data.txt").map(lambda line: line.split(","))
header = fake_rdd.first()
log_txt = fake_rdd.filter(lambda line: line != header)
fake_df=log_txt.toDF(['No','Birth_Country','Email','First_Name','Income','Job','Last_name','Loan_Approved','SSN'])
fake_df.registerTempTable("Fake_Table")
sqlContext.sql("select  max(Birth_Country),count(*) from Fake_Table group by Birth_Country order by count(*) desc limit 1").show(
```

```
+------------------+--------+
|max(Birth_Country)|count(1)|
+------------------+--------+
|             Korea|      91|
+------------------+--------+
```

Q 2) Find average income of people who are born in **united states of America**

```
In [2]: sqlContext.sql("select avg(Income)  from Fake_Table where  Birth_Country='United States of America'").show()
```

```
+--------------------------+
|avg(CAST(Income AS DOUBLE))|
+--------------------------+
|         208759.82352941178|
+--------------------------+
```

Q 3) How many people has income over 100,000 but their loan is not approved.

```
In [3]: sqlContext.sql("select count(*) from Fake_Table where loan_approved='FALSE' and Income>100000").show()
```

```
+--------+
|count(1)|
+--------+
|    4009|
+--------+
```

```
In [165]: df.filter((df.Income>100000)&(df.Loan_Approved=='FALSE')).count()
Out[165]: 4009
```

Q 4) Find top 10 people with highest income in **United States of America**. (Print their names, income and jobs)

```
In [4]: sqlContext.sql("select  First_Name as Name,Job,Income  from Fake_Table where  Birth_Country='United States of America' order by c
```

```
+--------+--------------------+------+
|    Name|                 Job|Income|
+--------+--------------------+------+
|  Alyssa|Amenity horticult...|482588|
|  Hunter|Psychologist pris...|468946|
|    Rose|Adult guidance wo...|426115|
|Danielle|Furniture conserv...|389810|
|   Terry|       Meteorologist|380410|
|   Cindy|Research scientis...|370322|
|   Scott|       Art therapist|368913|
| Christy|       Engineer land|355150|
|   Kelly|           Press sub|341448|
|Kristina|           Herbalist|338804|
+--------+--------------------+------+
```

Q 5) How many number of distinct jobs are there?

```
In [5]: sqlContext.sql("select count(distinct job) from Fake_Table").show()
```

```
+-------------------+
|count(DISTINCT job)|
+-------------------+
|                640|
+-------------------+
```

```
In [168]: df.select('Job').distinct().count()
Out[168]: 640
```

Q 6) How many writers earn less than 100,000?

```
In [6]: sqlContext.sql("select count(*) as NumberofWriters from Fake_Table where Job='Writer' and Income<100000").show()
```

```
+---------------+
|NumberofWriters|
+---------------+
|              5|
+---------------+
```

```
In [169]: df.filter((df.Income<100000)&(df.Job=="Writer")).count()
Out[169]: 5
```