Questions and Results

1)Find birth country which has highest amount of people

```
>>> data2='/Users/sumeetmishra/sw/spark-2.4.4-bin-hadoop2.7/examples/src/main/resources/Fake_data.csv'
>>> input3=spark.read.csv(data2,header=True)
>>> input3.createOrReplaceTempView("fake_data")
>>> sql_DF = spark.sql("SELECT Birth_Country, Count(*) PeopleCount FROM fake_data group by Birth_Country order by count(*) desc LIMIT 1")
19/12/04 02:13:42 WARN ObjectStore: Failed to get database global_temp, returning NoSuchObjectException
>>> sql_DF.show()
+-------------+-----------+
|Birth_Country|PeopleCount|
+-------------+-----------+
|        Korea|         91|
+-------------+-----------+
```

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

```
>>> sql_DF1 = spark.sql("SELECT AVG(Income) AvgIncome FROM fake_data WHERE Birth_Country = 'United States of America'")
>>> sql_DF1.show()
+------------------+
|         AvgIncome|
+------------------+
|208759.82352941178|
+------------------+
```

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

```
>>> sql_DF2 = spark.sql("SELECT Count(*) Over100KNotApprovedCt FROM fake_data WHERE Loan_Approved = False AND cast(Income as int) > 100000")
>>> sql_DF2.show()
+--------------------+
|Over100KNotApprovedCt|
+--------------------+
|                4009|
+--------------------+
```

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

```
>>> sql_DF3 = spark.sql("SELECT First_Name, Last_name, Income, Job FROM fake_data WHERE Birth_Country = 'United States of America' ORDER BY cast(Income as int)
 Desc LIMIT 10")
>>> sql_DF3.show()
+----------+---------+------+--------------------+
|First_Name|Last_name|Income|                 Job|
+----------+---------+------+--------------------+
|    Alyssa|   Miller|482588|Amenity horticult...|
|    Hunter|    Walls|468946|Psychologist, pri...|
|      Rose|Henderson|426115|Adult guidance wo...|
|  Danielle|  Leonard|389810|Furniture conserv...|
|     Terry|    Klein|380410|       Meteorologist|
|     Cindy|   Newton|370322|Research scientis...|
|     Scott| Mitchell|368913|       Art therapist|
|   Christy| Sandoval|355150|      Engineer, land|
|     Kelly| Reynolds|341448|           Press sub|
|  Kristina|    Smith|338804|           Herbalist|
+----------+---------+------+--------------------+
```

5) How many number of distinct jobs are there?

```
>>> sql_DF4 = spark.sql("SELECT count(distinct Job) FROM fake_data ")
>>> sql_DF4.show()
+--------------------+
|count(DISTINCT Job)|
+--------------------+
|                 639|
+--------------------+
```

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

```
>>> sql_DF5 = spark.sql("SELECT Count(*) WriterUnder100K FROM fake_data WHERE Job = 'Writer' AND cast(Income as int) < 100000")
>>> sql_DF5.show()
+---------------+
|WriterUnder100K|
+---------------+
|              5|
+---------------+
```