Himanshu Goyal

Intro to Spark- Assignment 4

Spark Transformation:

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

```
In [163]: from pyspark.sql.types import StructType
          from pyspark.sql.types import StructField
          from pyspark.sql.types import StringType
          log_txt=sc.textFile("Fake_data.txt")
          header = log_txt.first()
          log_txt = log_txt.filter(lambda line: line != header)
          emp_var = log_txt.map(lambda k: k.split(","))
          schema = StructType(
              [
               StructField('No', StringType(), True),
               StructField('Birth_Country', StringType(), True),
               StructField('Email', StringType(), True),
               StructField('First_Name', StringType(), True),
               StructField('Income', StringType(), True),
               StructField('Job', StringType(), True),
               StructField('Last_name', StringType(), True),
               StructField('Loan_Approved', StringType(), True),
               StructField('SSN', StringType(), True)]
          )
          df = sqlContext.createDataFrame(emp_var, schema)
          df.groupby('Birth_Country').count().sort(desc('count')).limit(1).show()

          +-------------+-----+
          |Birth_Country|count|
          +-------------+-----+
          |        Korea|   91|
          +-------------+-----+
```

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

```
In [164]: df.filter(df.Birth_Country =="United States of America").agg({'Income':'avg'}).show()

          +------------------+
          |       avg(Income)|
          +------------------+
          |208759.82352941178|
          +------------------+
```

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

```
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 [167]: df.sort(("Income")).filter(df.Birth_Country == "United States of America").select('First_Name','Last_name','Income','Job').show(1

          +----------+---------+------+--------------------+
          |First_Name|Last_name|Income|                 Job|
          +----------+---------+------+--------------------+
          |     Sherri|  Aguilar|101810|     Therapist drama|
          |     Justin|   Murphy|102305|Secretary/adminis...|
          |      Sarah| Harrison|112140|           Barrister|
          |  Stephanie|   Harris|127482|            Best boy|
          |Christopher|     Ward|153716|Manufacturing sys...|
          |    Adriana| Mcdonald|183976|Psychotherapist c...|
          |       John|   Martin|190090|       Haematologist|
          |       Seth| Campbell| 19296|        Metallurgist|
          |    Allison|    Price|195395|Television floor ...|
          |    Kristin| Reynolds| 20076|         Ship broker|
          +----------+---------+------+--------------------+
          only showing top 10 rows
```

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

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

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

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