Rakshith Churchagundi Amarnath
A20424771
CSP 554 – Assignment 8

Exercise 1:

Magic Number – 228154

```
hadoop@ip-172-31-7-238:~
[hadoop@ip-172-31-7-238 ~]$ java TestDataGen
Magic Number = 228154
[hadoop@ip-172-31-7-238 ~]$
```

```
hadoop@ip-172-31-7-238:~
[hadoop@ip-172-31-7-238 ~]$ hadoop fs -copyFromLocal foodratings228154.txt /user
/hadoop
[hadoop@ip-172-31-7-238 ~]$ hadoop fs -copyFromLocal foodplaces228154.txt /user/
hadoop
[hadoop@ip-172-31-7-238 ~]$ hadoop fs -ls /user/hadoop
Found 2 items
-rw-r--r--   1 hadoop hadoop         59 2020-03-09 18:34 /user/hadoop/foodplaces
228154.txt
-rw-r--r--   1 hadoop hadoop      17435 2020-03-09 18:34 /user/hadoop/foodrating
s228154.txt
[hadoop@ip-172-31-7-238 ~]$
```

Step C:

```
>>> from pyspark.sql.types import *
>>> ratings=StructType(
... [
... StructField("name",StringType(),True),
... StructField("food1",IntegerType(),True),
... StructField("food2",IntegerType(),True),
... StructField("food3",IntegerType(),True),
... StructField("food4",IntegerType(),True),
... StructField("placeid",IntegerType(),True),
... ])
>>>foodratings=spark.read.schema(ratings).csv('hdfs:///user/hadoop/foodratings228154.t
xt')
>>>foodratings.printSchema()
>>>foodratings.show(5)
```

```
>>> from pyspark.sql.types import *
>>> ratings=StructType(
... [
... StructField("name",StringType(),True),
... StructField("food1",IntegerType(),True),
... StructField("food2",IntegerType(),True),
... StructField("food3",IntegerType(),True),
... StructField("food4",IntegerType(),True),
... StructField("placeid",IntegerType(),True),
... ])
>>> foodratings=spark.read.schema(ratings).csv('hdfs:///user/hadoop/foodratings228154.txt')
>>> foodratings.printSchema()
root
 |-- name: string (nullable = true)
 |-- food1: integer (nullable = true)
 |-- food2: integer (nullable = true)
 |-- food3: integer (nullable = true)
 |-- food4: integer (nullable = true)
 |-- placeid: integer (nullable = true)

>>> foodratings.show(5)
+----+-----+-----+-----+-----+-------+
|name|food1|food2|food3|food4|placeid|
+----+-----+-----+-----+-----+-------+
|Jill|   22|   41|   31|   25|      1|
| Joy|    2|   25|    7|    4|      2|
| Joe|   38|   45|    6|    7|      2|
|Jill|    2|    9|   25|   40|      2|
| Mel|   24|   45|   42|   20|      1|
+----+-----+-----+-----+-----+-------+
only showing top 5 rows

>>> |
```

Exercise 2:

```
>>> from pyspark.sql.types import *
>>> places=StructType(
... [
... StructField("placeid", IntegerType(), True),
... StructField("placename", StringType(), True),
... ])
>>>
foodplaces=spark.read.schema(places).csv('hdfs:///user/hadoop/foodplaces228154.txt')
>>> foodplaces.printSchema()
>>> foodplaces.show(5)
```

```
>>> from pyspark.sql.types import *
>>> places=StructType(
... [
... StructField("placeid", IntegerType(), True),
... StructField("placename", StringType(), True),
... ])
>>> foodplaces=spark.read.schema(places).csv('hdfs:///user/hadoop/foodplaces228154.txt')
>>> foodplaces.printSchema()
root
 |-- placeid: integer (nullable = true)
 |-- placename: string (nullable = true)

>>> foodplaces.show(5)
+-------+------------+
|placeid|   placename|
+-------+------------+
|      1|China Bistro|
|      2|    Atlantic|
|      3|   Food Town|
|      4|      Jake's|
|      5|   Soup Bowl|
+-------+------------+

>>> |
```

Exercise 3:

Step A:

```
>>> from pyspark.sql.types import *
>>> foodratings.createOrReplaceTempView("foodratingsT")
>>> foodplaces.createOrReplaceTempView("foodplacesT")
```

Step B:

```
>>> foodratings_ex3a=spark.sql("SELECT * FROM foodratingsT WHERE food2<25 AND
food4>40")
>>> foodratings_ex3a.printSchema()
>>> foodratings_ex3a.show(5)
```

```
>>> from pyspark.sql.types import *
>>>
>>> foodratings.createOrReplaceTempView("foodratingsT")
>>> foodplaces.createOrReplaceTempView("foodplacesT")
>>> foodratings_ex3a=spark.sql("SELECT * FROM foodratingsT WHERE food2<25 AND food4>40")
>>> foodratings_ex3a.printSchema()
root
 |-- name: string (nullable = true)
 |-- food1: integer (nullable = true)
 |-- food2: integer (nullable = true)
 |-- food3: integer (nullable = true)
 |-- food4: integer (nullable = true)
 |-- placeid: integer (nullable = true)

>>> foodratings_ex3a.show(5)
+----+-----+-----+-----+-----+-------+
|name|food1|food2|food3|food4|placeid|
+----+-----+-----+-----+-----+-------+
| Sam|   23|   24|   48|   46|      1|
| Sam|   13|   15|   35|   50|      1|
| Sam|   16|   20|   27|   44|      4|
| Sam|   13|   18|    8|   45|      5|
| Joe|    4|   14|   49|   49|      2|
+----+-----+-----+-----+-----+-------+
only showing top 5 rows

>>>
```

Step C:

```
>>>
>>> foodplaces_ex3b=spark.sql("SELECT * FROM foodplacesT WHERE placeid>3")
>>> foodplaces_ex3b.printSchema()
>>> foodplaces_ex3b.show(5)
```

```
>>>
>>> foodplaces_ex3b=spark.sql("SELECT * FROM foodplacesT WHERE placeid>3")
>>> foodplaces_ex3b.printSchema()
root
 |-- placeid: integer (nullable = true)
 |-- placename: string (nullable = true)

>>> foodplaces_ex3b.show(5)
+-------+---------+
|placeid|placename|
+-------+---------+
|      4|   Jake's|
|      5|Soup Bowl|
+-------+---------+

>>>
```

Exercise 4:

```
>>> foodratings_ex4=foodratings.filter("name='Mel' and food3<25")
>>> foodratings_ex4.printSchema()
>>> foodratings_ex4.show(5)
```

```
>>> foodratings_ex4=foodratings.filter("name='Mel' and food3<25")
>>> foodratings_ex4.printSchema()
root
 |-- name: string (nullable = true)
 |-- food1: integer (nullable = true)
 |-- food2: integer (nullable = true)
 |-- food3: integer (nullable = true)
 |-- food4: integer (nullable = true)
 |-- placeid: integer (nullable = true)

>>> foodratings_ex4.show(5)
+----+-----+-----+-----+-----+-------+
|name|food1|food2|food3|food4|placeid|
+----+-----+-----+-----+-----+-------+
| Mel|   41|   21|   23|   36|      3|
| Mel|   49|    3|   14|   28|      5|
| Mel|   39|   41|   14|   26|      2|
| Mel|   20|   11|    2|   45|      5|
| Mel|   29|   28|   13|    5|      3|
+----+-----+-----+-----+-----+-------+
only showing top 5 rows

>>> |
```

Exercise 5:

```
>>> foodratings_ex5 = foodratings.select('name','placeid')
>>> foodratings_ex5.printSchema()
>>> foodratings_ex5.show(5)
```

```
>>> foodratings_ex5 = foodratings.select('name','placeid')
>>> foodratings_ex5.printSchema()
root
 |-- name: string (nullable = true)
 |-- placeid: integer (nullable = true)

>>> foodratings_ex5.show(5)
+----+-------+
|name|placeid|
+----+-------+
|Jill|      1|
| Joy|      2|
| Joe|      2|
|Jill|      2|
| Mel|      1|
+----+-------+
only showing top 5 rows

>>> |
```

Exercise 6:

```
>>> ex6 = foodratings.join(foodplaces,foodratings.placeid==foodplaces.placeid)
>>> ex6.printSchema()
>>> ex6.show(5)
```

```
>>> ex6 = foodratings.join(foodplaces,foodratings.placeid==foodplaces.placeid)
>>> ex6.printSchema()
root
 |-- name: string (nullable = true)
 |-- food1: integer (nullable = true)
 |-- food2: integer (nullable = true)
 |-- food3: integer (nullable = true)
 |-- food4: integer (nullable = true)
 |-- placeid: integer (nullable = true)
 |-- placeid: integer (nullable = true)
 |-- placename: string (nullable = true)

>>> ex6.show(5)
+----+-----+-----+-----+-----+-------+-------+------------+
|name|food1|food2|food3|food4|placeid|placeid|   placename|
+----+-----+-----+-----+-----+-------+-------+------------+
|Jill|   22|   41|   31|   25|      1|      1|China Bistro|
| Joy|    2|   25|    7|    4|      2|      2|    Atlantic|
| Joe|   38|   45|    6|    7|      2|      2|    Atlantic|
|Jill|    2|    9|   25|   40|      2|      2|    Atlantic|
| Mel|   24|   45|   42|   20|      1|      1|China Bistro|
+----+-----+-----+-----+-----+-------+-------+------------+
only showing top 5 rows

>>> |
```