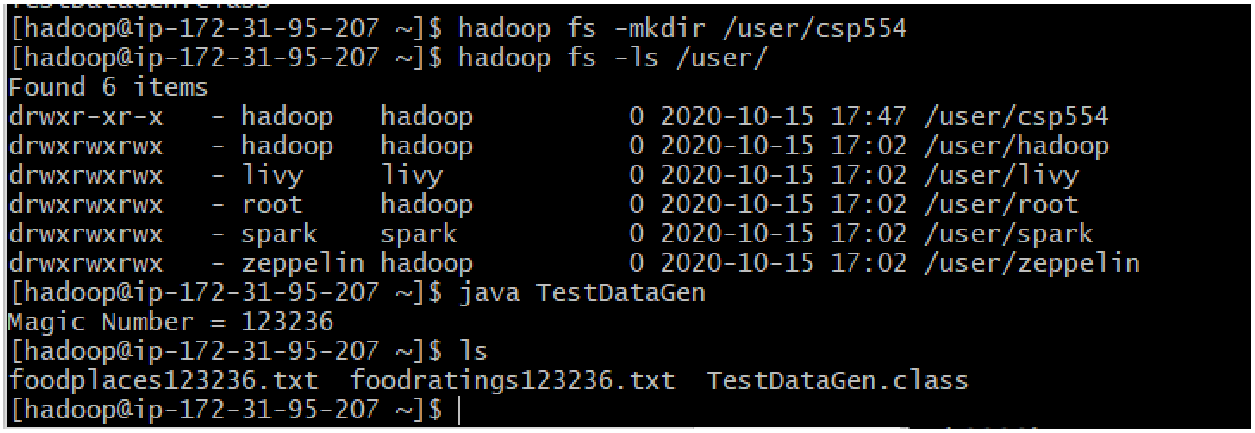
Creating an EMR Cluster

```
EEEEEEEEEEEEEEEEEEEE MMMMMMMM           MMMMMMMM RRRRRRRRRRRRRRR
E::::::::::::::::::E M:::::::M         M:::::::M R::::::::::::::R
EE:::::EEEEEEEEE:::E M::::::::M       M::::::::M R:::::RRRRRR:::::R
  E::::E       EEEEE M:::::::::M     M:::::::::M RR::::R      R::::R
  E::::E             M::::::M:::M   M:::M::::::M   R:::R      R::::R
  E:::::EEEEEEEEEE   M:::::M M:::M M:::M M:::::M   R:::RRRRRR:::::R
  E::::::::::::::E   M:::::M  M:::M:::M  M:::::M   R:::::::::::RR
  E:::::EEEEEEEEEE   M:::::M   M:::::M   M:::::M   R:::RRRRRR::::R
  E::::E             M:::::M    M:::M    M:::::M   R:::R      R::::R
  E::::E       EEEEE M:::::M     MMM     M:::::M   R:::R      R::::R
EE:::::EEEEEEEE::::E M:::::M             M:::::M   R:::R      R::::R
E::::::::::::::::::E M:::::M             M:::::M RR::::R      R::::R
EEEEEEEEEEEEEEEEEEEE MMMMMMM             MMMMMMM RRRRRRR      RRRRRR
```

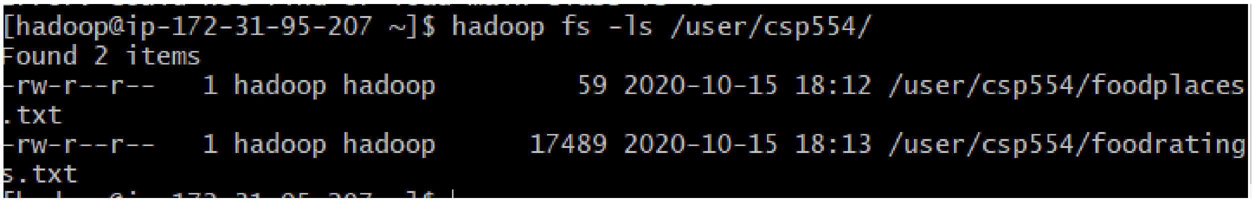
Running TestDataGen.Class

Magic Number: 123236



```
[hadoop@ip-172-31-95-207 ~]$ hadoop fs -mkdir /user/csp554
[hadoop@ip-172-31-95-207 ~]$ hadoop fs -ls /user/
Found 6 items
drwxr-xr-x   - hadoop   hadoop          0 2020-10-15 17:47 /user/csp554
drwxrwxrwx   - hadoop   hadoop          0 2020-10-15 17:02 /user/hadoop
drwxrwxrwx   - livy     livy            0 2020-10-15 17:02 /user/livy
drwxrwxrwx   - root     hadoop          0 2020-10-15 17:02 /user/root
drwxrwxrwx   - spark    spark           0 2020-10-15 17:02 /user/spark
drwxrwxrwx   - zeppelin hadoop          0 2020-10-15 17:02 /user/zeppelin
[hadoop@ip-172-31-95-207 ~]$ java TestDataGen
Magic Number = 123236
[hadoop@ip-172-31-95-207 ~]$ ls
foodplaces123236.txt  foodratings123236.txt  TestDataGen.class
[hadoop@ip-172-31-95-207 ~]$ |
```

Moving the files created using testDatagen.class to newly created Directory csp554 using copyFropmLocal command



```
[hadoop@ip-172-31-95-207 ~]$ hadoop fs -ls /user/csp554/
Found 2 items
-rw-r--r--   1 hadoop hadoop         59 2020-10-15 18:12 /user/csp554/foodplaces
.txt
-rw-r--r--   1 hadoop hadoop      17489 2020-10-15 18:13 /user/csp554/foodrating
s.txt
```

Exercise 1:

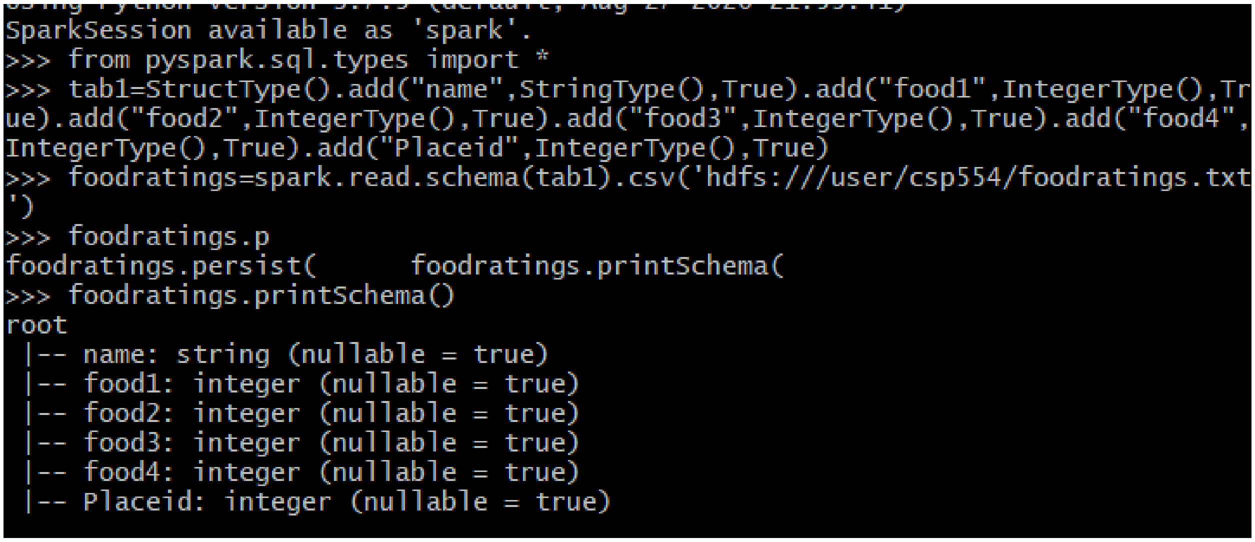
Loading the foodratings.txt file to foodratings DataFrame

Command Used:

from pyspark.sql.types import *

tab1=StructType().add("name",StringType(),True).add("food1",IntegerType(),True).add("food2",IntegerType(),True).add("food3",IntegerType(),True).add("food4",IntegerType(),True).add("Placeid",IntegerType(),True)

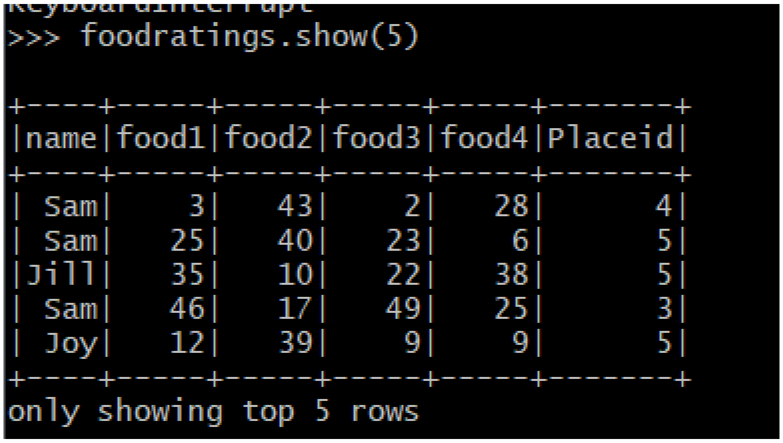
foodratings=spark.read.schema(tab1).csv('hdfs:///user/csp554/foodratings.txt')



```
SparkSession available as 'spark'.
>>> from pyspark.sql.types import *
>>> tab1=StructType().add("name",StringType(),True).add("food1",IntegerType(),Tr
ue).add("food2",IntegerType(),True).add("food3",IntegerType(),True).add("food4",
IntegerType(),True).add("Placeid",IntegerType(),True)
>>> foodratings=spark.read.schema(tab1).csv('hdfs:///user/csp554/foodratings.txt
')
>>> foodratings.p
foodratings.persist(      foodratings.printSchema(
>>> 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)
```

Showing top 5 Rows

Command Used: foodratings.show(5)



```
KeyboardInterrupt
>>> foodratings.show(5)

+----+-----+-----+-----+-----+-------+
|name|food1|food2|food3|food4|Placeid|
+----+-----+-----+-----+-----+-------+
| Sam|    3|   43|    2|   28|      4|
| Sam|   25|   40|   23|    6|      5|
|Jill|   35|   10|   22|   38|      5|
| Sam|   46|   17|   49|   25|      3|
| Joy|   12|   39|    9|    9|      5|
+----+-----+-----+-----+-----+-------+
only showing top 5 rows
```

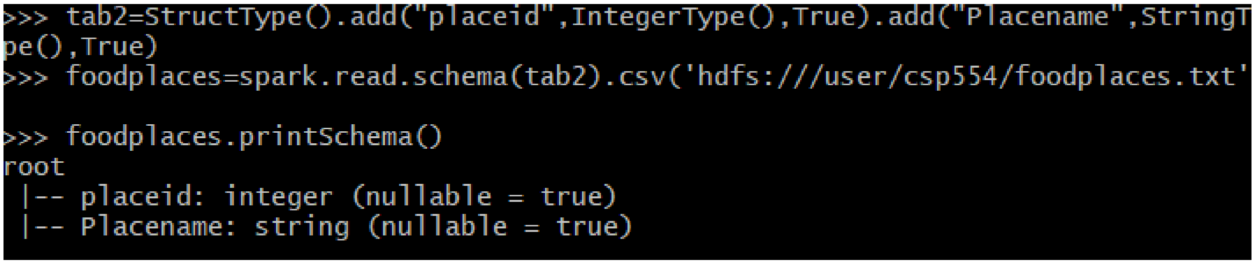
Exercise 2:

Loading the foodplaces.txt file to foodplaces DataFrame

Command Used:
tab2=StructType().add("placeid",IntegerType(),True).add("Placename",StringType(),True)

foodplaces=spark.read.schema(tab2).csv('hdfs:///user/csp554/foodplaces.txt')

foodplaces.printSchema()

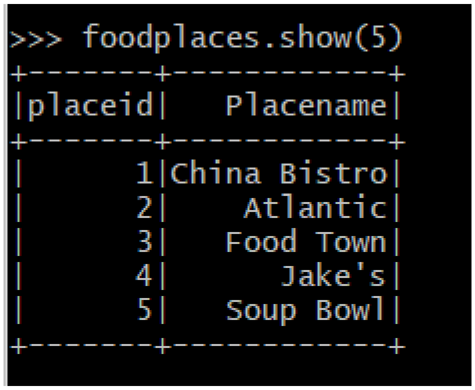


```
>>> tab2=StructType().add("placeid",IntegerType(),True).add("Placename",StringT
pe(),True)
>>> foodplaces=spark.read.schema(tab2).csv('hdfs:///user/csp554/foodplaces.txt'

>>> foodplaces.printSchema()
root
 |-- placeid: integer (nullable = true)
 |-- Placename: string (nullable = true)
```

Showing top 5 Rows

Command Used: foodplaces.show(5)



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

Exercise 3:

a) Creating a table using the below command

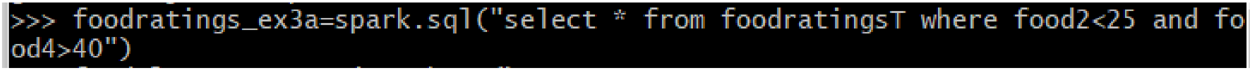
Command used:

foodratings.createOrReplaceTempView("foodratingsT")
foodplaces.createOrReplaceTempView("foodplacesT")

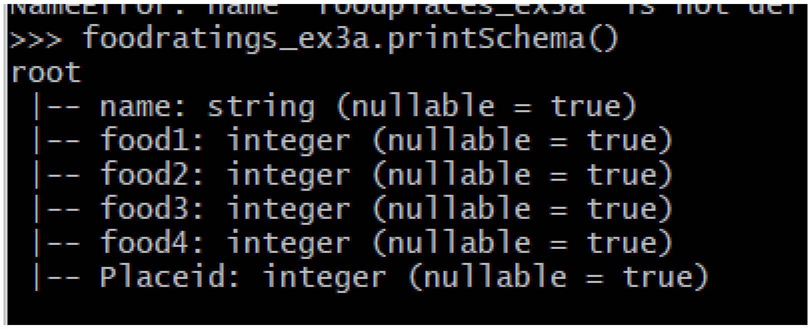
b) Creating a new table from the fooodratingsT created at above step

Command Used:

foodratings_ex3a=spark.sql("select * from foodratingsT where food2<25 and food4>40")



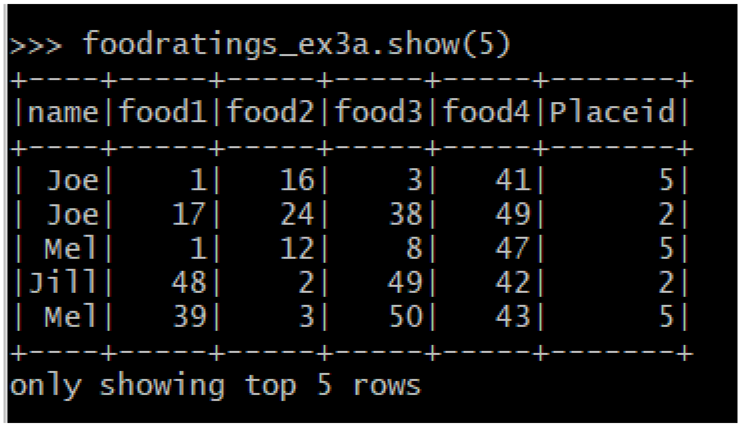
```
>>> foodratings_ex3a=spark.sql("select * from foodratingsT where food2<25 and fo
od4>40")
```



```
NameError: name 'foodplaces_ex3a' is not def
>>> 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)
```

Showing top 5 Rows

Command Used: foodpratings_ex3a.show(5)

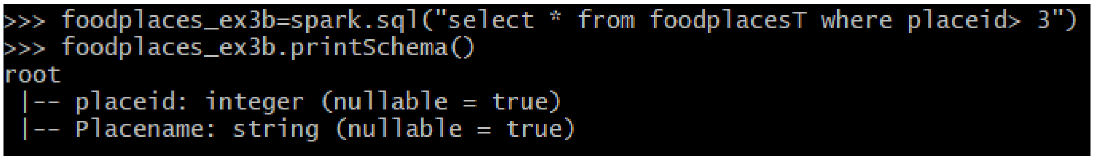


```
>>> foodratings_ex3a.show(5)
+----+-----+-----+-----+-----+-------+
|name|food1|food2|food3|food4|Placeid|
+----+-----+-----+-----+-----+-------+
| Joe|    1|   16|    3|   41|      5|
| Joe|   17|   24|   38|   49|      2|
| Mel|    1|   12|    8|   47|      5|
|Jill|   48|    2|   49|   42|      2|
| Mel|   39|    3|   50|   43|      5|
+----+-----+-----+-----+-----+-------+
only showing top 5 rows
```

c) Creating a new table from the fooodplacesT created at above step

Command Used:

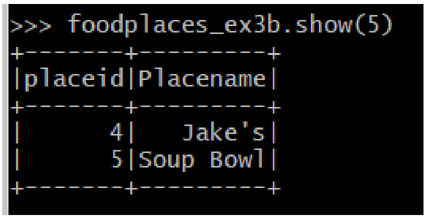
foodplaces_ex3b=spark.sql("select * from foodplacesT where placeid> 3")
foodplaces_ex3b.printSchema()



```
>>> foodplaces_ex3b=spark.sql("select * from foodplacesT where placeid> 3")
>>> foodplaces_ex3b.printSchema()
root
 |-- placeid: integer (nullable = true)
 |-- Placename: string (nullable = true)
```

Showing top 5 Rows

Command Used: foodplaces.show(5)



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

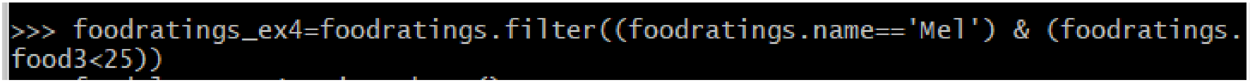
Exercise 4:

Creating a new DataFrame using the below command

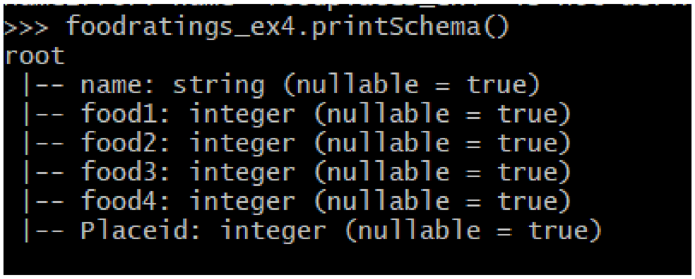
Command Used:

foodratings_ex4=foodratings.filter((foodratings.name=='Mel') & (foodratings.food3<25))

foodratings_ex4.printSchema()



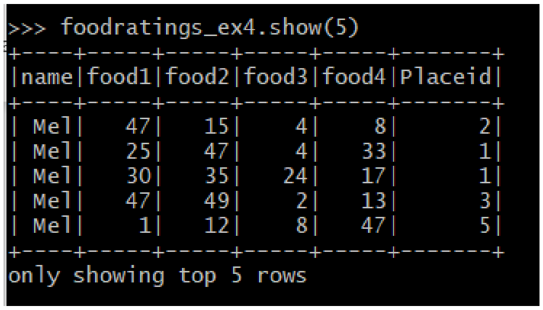
```
>>> foodratings_ex4=foodratings.filter((foodratings.name=='Mel') & (foodratings.
food3<25))
```



```
>>> 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)
```

Showing top 5 Rows

Command Used: foodratings_ex4.show(5)

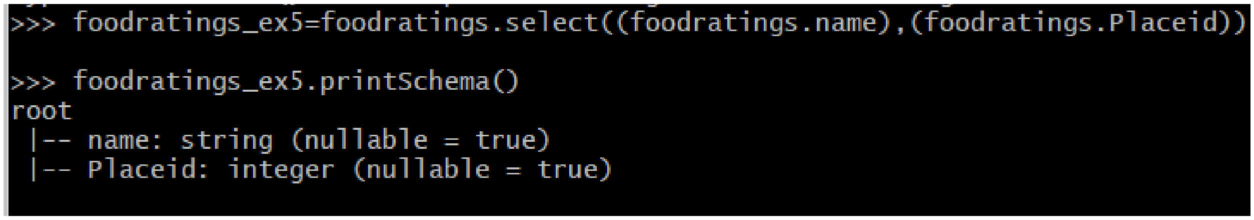


```
>>> foodratings_ex4.show(5)
+----+-----+-----+-----+-----+-------+
|name|food1|food2|food3|food4|Placeid|
+----+-----+-----+-----+-----+-------+
| Mel|   47|   15|    4|    8|      2|
| Mel|   25|   47|    4|   33|      1|
| Mel|   30|   35|   24|   17|      1|
| Mel|   47|   49|    2|   13|      3|
| Mel|    1|   12|    8|   47|      5|
+----+-----+-----+-----+-----+-------+
only showing top 5 rows
```

Exercise 5: Creating a new DataFrame using columns name and PlaceId

Command Used:

foodratings_ex5=foodratings.select((foodratings.name),(foodratings.Placeid))

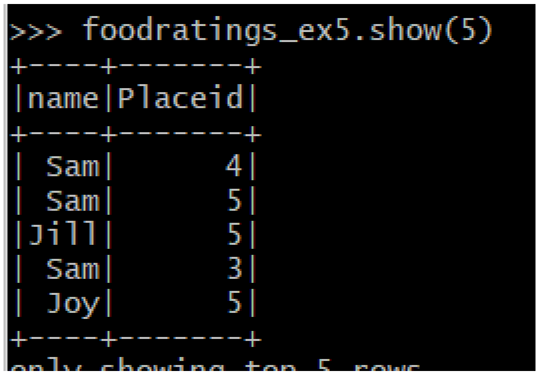


```
>>> foodratings_ex5=foodratings.select((foodratings.name),(foodratings.Placeid))

>>> foodratings_ex5.printSchema()
root
 |-- name: string (nullable = true)
 |-- Placeid: integer (nullable = true)
```

Showing top 5 Rows

Command Used: foodratings_ex5.show(5)

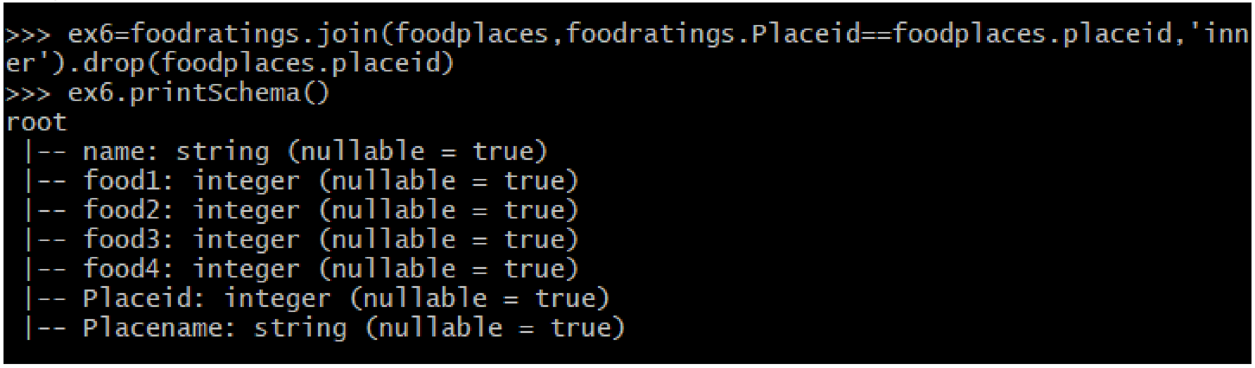


```
>>> foodratings_ex5.show(5)
+----+-------+
|name|Placeid|
+----+-------+
| Sam|      4|
| Sam|      5|
|Jill|      5|
| Sam|      3|
| Joy|      5|
+----+-------+
only showing top 5 rows
```

Exercise 6: Creating a new Dataframe using below command

Command Used:

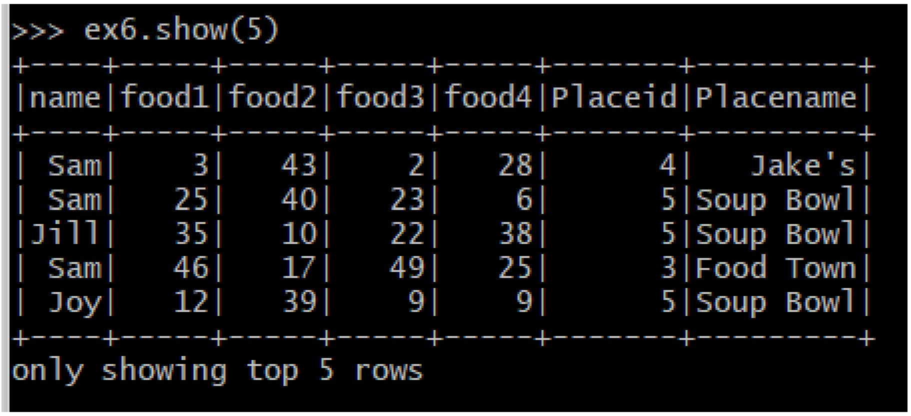
ex6=foodratings.join(foodplaces,foodratings.Placeid==foodplaces.placeid,'inner').drop(foodplaces.placeid)
ex6.printSchema()



```
>>> ex6=foodratings.join(foodplaces,foodratings.Placeid==foodplaces.placeid,'inn
er').drop(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)
 |-- Placename: string (nullable = true)
```

Showing top 5 Rows

Command Used: ex6.show(5)



```
>>> ex6.show(5)
+----+-----+-----+-----+-----+-------+---------+
|name|food1|food2|food3|food4|Placeid|Placename|
+----+-----+-----+-----+-----+-------+---------+
| Sam|    3|   43|    2|   28|      4|   Jake's|
| Sam|   25|   40|   23|    6|      5|Soup Bowl|
|Jill|   35|   10|   22|   38|      5|Soup Bowl|
| Sam|   46|   17|   49|   25|      3|Food Town|
| Joy|   12|   39|    9|    9|      5|Soup Bowl|
+----+-----+-----+-----+-----+-------+---------+
only showing top 5 rows
```