Exercise 1 :

Magic Number : 62216

Code :

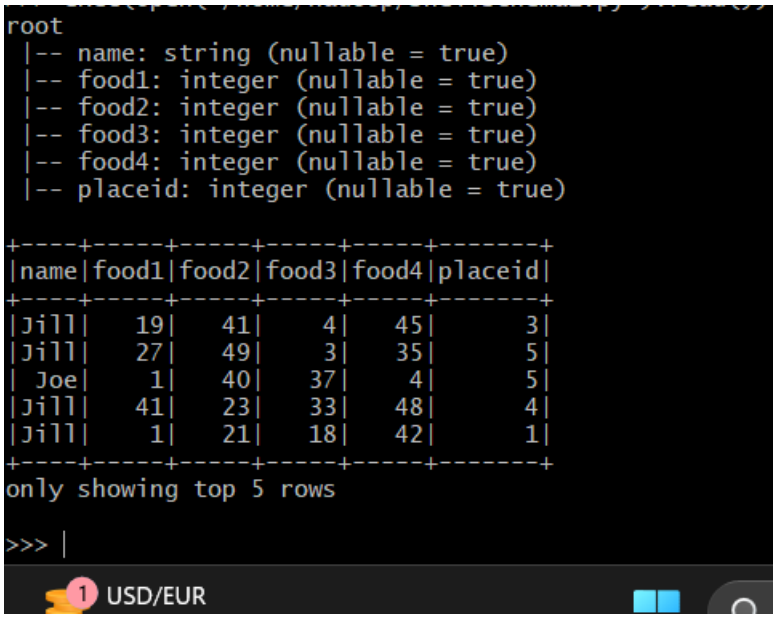
```
from pyspark.sql import *
food_schema = StructType([
    StructField("name", StringType(), True),
    StructField("food1", IntegerType(), True),
    StructField("food2", IntegerType(), True),
    StructField("food3", IntegerType(), True),
    StructField("food4", IntegerType(), True),
    StructField("placeid", IntegerType(), True)
])


foodratings = spark.read.csv("/user/hadoop/foodratings62216.txt",
schema=food_schema, header=True)


foodratings.printSchema()


foodratings.show(5)
```

output :



Exercise 2:

Code :

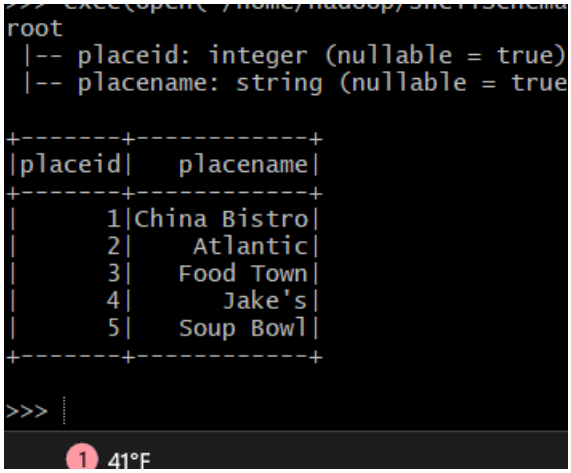
```
from pyspark.sql import *

struct1 = StructType().add("placeid", IntegerType(), True).add("placename", StringType(), True)
```

foodplaces = spark.read.schema(struct1).csv('/user/hadoop/foodplaces62216.txt')

foodplaces.printSchema()

foodplaces.show(5)



Exercise 3:

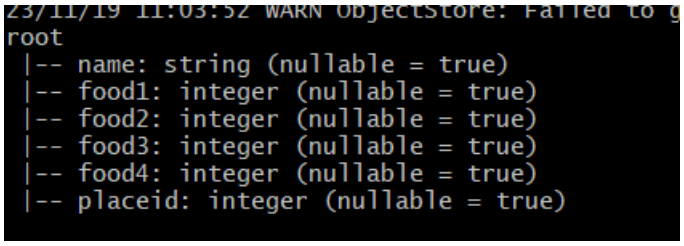
3.1:

Code :

foodratings.createOrReplaceTempView("foodratingsT")

foodplaces.createOrReplaceTempView("foodplacesT")

output :



3.2

Code :

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

foodratings_ex3a.printSchema()

foodratings_ex3a.show(5)

output :

```
+----+-----+-----+-----+-----+-------+
|name|food1|food2|food3|food4|placeid|
+----+-----+-----+-----+-----+-------+
|Jill|   41|   23|   33|   48|      4|
|Jill|    1|   21|   18|   42|      1|
| Joy|    5|   19|   35|   43|      5|
| Sam|    8|   12|    7|   50|      3|
| Joy|    8|   24|    9|   47|      4|
+----+-----+-----+-----+-----+-------+
only showing top 5 rows
```

3.3

Code :

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

foodplaces_ex3b.printSchema()

foodplaces_ex3b.show(5)

output :

```
root
 |-- placeid: integer (nullable = true)
 |-- placename: string (nullable = true)

+-------+---------+
|placeid|placename|
+-------+---------+
|      4|   Jake's|
|      5|Soup Bowl|
+-------+---------+

>>>
```

Exercise 4:

Code :

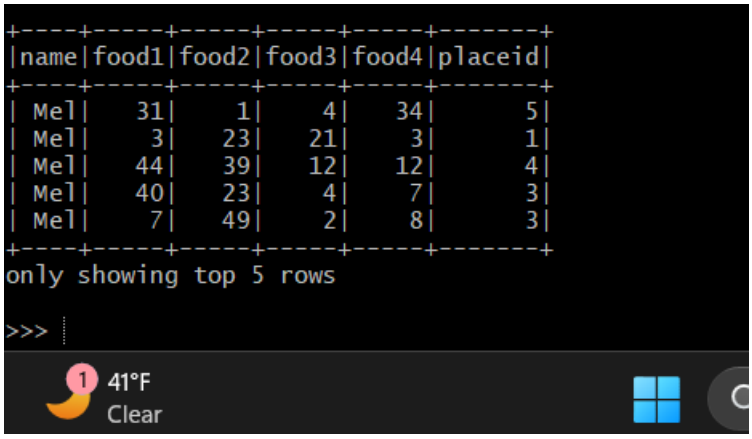
foodratings_ex4 = foodratings.filter((foodratings['name'] == "Mel") & (foodratings['food3'] < 25))

foodratings_ex4.printSchema()

foodratings_ex4.show(5)

output :

```
>>> exec(open( /home/hadoop/shellSchema4.py ).r
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)
```



Exercise 5:

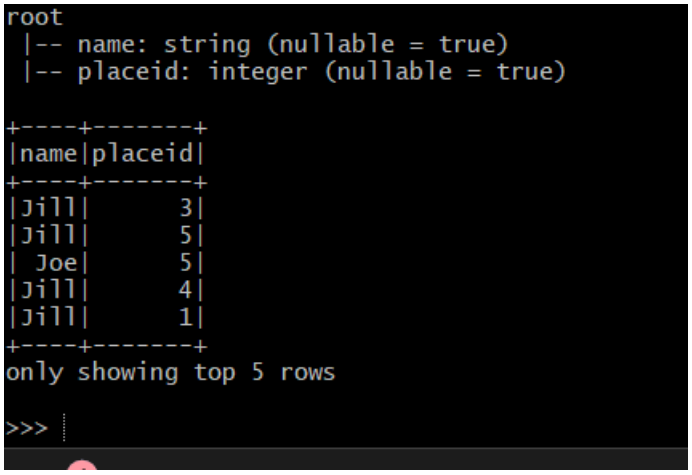
Code :

foodratings_ex5 = foodratings.select(foodratings['name'],foodratings['placeid'])

foodratings_ex5.printSchema()

foodratings_ex5.show(5)

output:



Exercise 6:

Code :

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

ex6.printSchema()

ex6.show(5)

output :

```
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)
```

```
+----+-----+-----+-----+-----+-------+-------+------------+
|name|food1|food2|food3|food4|placeid|placeid|   placename|
+----+-----+-----+-----+-----+-------+-------+------------+
|Jill|   19|   41|    4|   45|      3|      3|   Food Town|
|Jill|   27|   49|    3|   35|      5|      5|   Soup Bowl|
| Joe|    1|   40|   37|    4|      5|      5|   Soup Bowl|
|Jill|   41|   23|   33|   48|      4|      4|      Jake's|
|Jill|    1|   21|   18|   42|      1|      1|China Bistro|
+----+-----+-----+-----+-----+-------+-------+------------+
only showing top 5 rows

>>>
```