```
pip install pyspark
```

```
Requirement already satisfied: pyspark in /usr/local/lib/python3.11/dist-packages (3.5.1)
Requirement already satisfied: py4j==0.10.9.7 in /usr/local/lib/python3.11/dist-packages (from pyspark) (0.10
```

```
from pyspark.sql import SparkSession


spark = SparkSession.builder.appName("uber data analysis").getOrCreate()


sc = spark.sparkContext


data_with_headerRDD = sc.textFile("/content/sample_data/uber_data")


data_with_headerRDD.count()
```

```
355
```

```
for line in data_with_headerRDD.take(5):
  print(line)
```

```
dispatching_base_number,date,active_vehicles,trips
B02512,1/1/2015,190,1132
B02765,1/1/2015,225,1765
B02764,1/1/2015,3427,29421
B02682,1/1/2015,945,7679
```

```
data_with_headerRDD.getNumPartitions()
```

```
2
```

```
header = data_with_headerRDD.first()


print(header)
```

```
dispatching_base_number,date,active_vehicles,trips
```

```
uberRDD = data_with_headerRDD.filter(lambda a : a != header)


uberRDD.count()
```

```
354
```

```
for line in uberRDD.take(5):
  print(line)
```

```
B02512,1/1/2015,190,1132
B02765,1/1/2015,225,1765
B02764,1/1/2015,3427,29421
B02682,1/1/2015,945,7679
B02617,1/1/2015,1228,9537
```

```
import datetime


format_date = "%m/%d/%Y"


print(datetime.datetime.strptime("05/11/2025", format_date).strftime("%A"))
```

```
Sunday
```

```
tupleRDD = uberRDD.map(lambda a : (a.split(",")[0],datetime.datetime.strptime(a.split(",")[1], format_date).strft
```

```
for line in tupleRDD.take(5):
  print(line)
```

```
('B02512', 'Thursday', '1132')
('B02765', 'Thursday', '1765')
('B02764', 'Thursday', '29421')
('B02682', 'Thursday', '7679')
('B02617', 'Thursday', '9537')
```

```
tupleRDD.count()
```

```
354
```

```
kvPairRDD = tupleRDD.map(lambda a : (a[0] + " " + a[1], int(a[2])))
```

```
for line in kvPairRDD.take(5):
  print(line)
```

```
('B02512 Thursday', 1132)
('B02765 Thursday', 1765)
('B02764 Thursday', 29421)
('B02682 Thursday', 7679)
('B02617 Thursday', 9537)
```

```
totalTripsRDD = kvPairRDD.reduceByKey(lambda a,b : a+b)
```

```
for line in totalTripsRDD.collect():
  print(line)
```

```
('B02512 Thursday', 15809)
('B02764 Thursday', 304200)
('B02682 Thursday', 106643)
('B02617 Thursday', 118254)
('B02598 Friday', 93126)
('B02617 Friday', 125067)
('B02512 Friday', 16435)
('B02682 Friday', 114662)
('B02765 Friday', 34934)
('B02764 Friday', 326968)
('B02512 Saturday', 15026)
('B02617 Sunday', 91722)
('B02764 Sunday', 249896)
('B02512 Monday', 11297)
('B02682 Monday', 74939)
('B02617 Monday', 80591)
('B02764 Monday', 214116)
('B02765 Monday', 21974)
('B02765 Tuesday', 22741)
('B02617 Wednesday', 94887)
('B02682 Wednesday', 86252)
('B02764 Wednesday', 241137)
('B02598 Wednesday', 71956)
('B02765 Thursday', 30408)
('B02598 Thursday', 90333)
('B02765 Saturday', 36737)
('B02617 Saturday', 127902)
('B02598 Saturday', 94588)
('B02682 Saturday', 120283)
('B02764 Saturday', 356789)
('B02512 Sunday', 10487)
('B02682 Sunday', 82825)
('B02598 Sunday', 66477)
('B02765 Sunday', 22536)
('B02598 Monday', 60882)
('B02764 Tuesday', 221343)
```

```
('B02682 Tuesday', 76905)
('B02617 Tuesday', 86602)
('B02512 Tuesday', 12041)
('B02598 Tuesday', 63429)
('B02765 Wednesday', 24340)
('B02512 Wednesday', 12691)
```

```
sortbyval = totalTripsRDD.sortBy(lambda a : -a[1])
```

```
for line in sortbyval.collect():
  print(line)
```

```
('B02764 Saturday', 356789)
('B02764 Friday', 326968)
('B02764 Thursday', 304200)
('B02764 Sunday', 249896)
('B02764 Wednesday', 241137)
('B02764 Tuesday', 221343)
('B02764 Monday', 214116)
('B02617 Saturday', 127902)
('B02617 Friday', 125067)
('B02682 Saturday', 120283)
('B02617 Thursday', 118254)
('B02682 Friday', 114662)
('B02682 Thursday', 106643)
('B02617 Wednesday', 94887)
('B02598 Saturday', 94588)
('B02598 Friday', 93126)
('B02617 Sunday', 91722)
('B02598 Thursday', 90333)
('B02617 Tuesday', 86602)
('B02682 Wednesday', 86252)
('B02682 Sunday', 82825)
('B02617 Monday', 80591)
('B02682 Tuesday', 76905)
('B02682 Monday', 74939)
('B02598 Wednesday', 71956)
('B02598 Sunday', 66477)
('B02598 Tuesday', 63429)
('B02598 Monday', 60882)
('B02765 Saturday', 36737)
('B02765 Friday', 34934)
('B02765 Thursday', 30408)
('B02765 Wednesday', 24340)
('B02765 Tuesday', 22741)
('B02765 Sunday', 22536)
('B02765 Monday', 21974)
('B02512 Friday', 16435)
('B02512 Thursday', 15809)
('B02512 Saturday', 15026)
('B02512 Wednesday', 12691)
('B02512 Tuesday', 12041)
('B02512 Monday', 11297)
('B02512 Sunday', 10487)
```

```
mylist = [100,200,300]
```

```
print(mylist)
```

```
[100, 200, 300]
```

```
listRDD = sc.parallelize(mylist)
```

```
for line in listRDD.collect():
  print(line)
```

```
100
200
300
```

```
rdd1 = sc.parallelize([("a", 1), ("b", 4)])
```

```
rdd1.count()
```

```
2
```

```
for line in rdd1.collect():
  print(line)
```

```
('a', 1)
('b', 4)
```

```
rdd2 = sc.parallelize([("a", 2), ("a", 3)])
```

```
for line in rdd2.collect():
  print(line)
```

```
('a', 2)
('a', 3)
```

```
#inner join
rdd1.join(rdd2).collect()
```

```
[('a', (1, 2)), ('a', (1, 3))]
```

```
#left outer join
rdd1.leftOuterJoin(rdd2).collect()
```

```
[('b', (4, None)), ('a', (1, 2)), ('a', (1, 3))]
```

```
#right outer join
rdd1.rightOuterJoin(rdd2).collect()
```

```
[('a', (1, 2)), ('a', (1, 3))]
```

```
#full outer join
rdd1.fullOuterJoin(rdd2).collect()
```

```
[('b', (4, None)), ('a', (1, 2)), ('a', (1, 3))]
```