Analyze data RDD

FINISHED

```
%pyspark

# Columns in input file:
# 0 Year
# 1 Month
# 2 DayofMonth
# 3 DepTime
# 4 UniqueCarrier
# 5 FlightNum
# 6 ArrDelay
# 7 Origin
# 8 Dest

input_file = 'hdfs:///user/maria_dev/flights/flightdelays_clean_rdd'
total_output_file = 'hdfs:///user/maria_dev/flights/flightdelays_cleaned_total_rdd'
denver_output_file = 'hdfs:///user/maria_dev/flights/flightdelays_cleaned_denver_rdd'
denver_late_output_file = 'hdfs:///user/maria_dev/flights/flightdelays_cleaned_denver_late_rdd'

flight_delays = sc.textFile(input_file)
```

Took 0 sec. Last updated by anonymous at March 03 2018, 1:16:46 PM. (outdated)

Total Count

FINISHED

```
%pyspark

rdd = sc.parallelize([flight_delays.count()])
rdd.saveAsTextFile(total_output_file)
```

Took 0 sec. Last updated by anonymous at March 03 2018, 1:17:51 PM.

Denver Count

FINISHED

```
%pyspark

c = flight_delays \
```

```
    .map(lambda l: l.split(',')) \
    .filter(lambda c: c[8] == 'DEN') \
    .count()

rdd = sc.parallelize([c])
```

Took 0 sec. Last updated by anonymous at March 03 2018, 1:32:28 PM.

Denver Late Count

FINISHED

```
%pyspark

c = flight_delays \
    .map(lambda l: l.split(',')) \
    .filter(lambda c: c[8] == 'DEN' and int(c[6]) >= 60) \
    .count()

rdd = sc.parallelize([c])
rdd.saveAsTextFile(denver_late_output_file)
```

Took 1 sec. Last updated by anonymous at March 03 2018, 1:35:00 PM.

```
%pyspark
```

READY