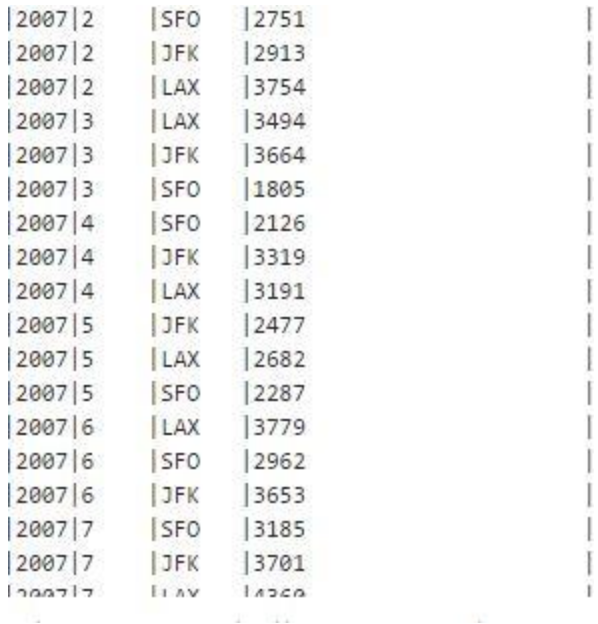
1. Delay Analysis of flights from JFK , LAX, SFO in 2006-2007

a. Distributed by month

```
import scala.collection.mutable.WrappedArray
import spark.implicits._
import org.apache.spark.sql.functions._

 val spark = org.apache.spark.sql.SparkSession.builder
      .master("local")
      .appName("Spark CSV Reader")
      .getOrCreate;

val ds2006 = spark.read
      .format("com.databricks.spark.csv")
      .option("header", "true") //reading the headers
      .option("mode", "DROPMALFORMED")
      .load("/usr/nishchal/Nishchal/bigData/flight/2006.csv");
val ds2007 = spark.read
      .format("com.databricks.spark.csv")
      .option("header", "true") //reading the headers
      .option("mode", "DROPMALFORMED")
      .load("/usr/nishchal/Nishchal/bigData/flight/2007.csv");

ds2006.registerTempTable("flightDelays_2006");
ds2007.registerTempTable("flightDelays_2007");

val df2=spark.sql("select year,month,origin, count(*) as number_of_delayed_flights from
flightDelays_2006 where DepDelay>=15 and Origin in ('JFK','LAX','SFO') group by year,Month,origin
UNION ALL select year,month,origin, count(*) as number_of_delayed_flights from flightDelays_2007
where DepDelay>=15 and Origin in ('JFK','LAX','SFO') group by year,Month,origin order by year,month");

df2.show(72,false);
```

2006	8	LAX	3931
2006	8	JFK	2595
2006	9	LAX	2970
2006	9	SFO	1898
2006	9	JFK	2166
2007	1	LAX	3655
2007	1	JFK	2865
2007	1	SFO	2080
2007	10	JFK	2110
2007	10	LAX	3077
2007	10	SFO	2823
2007	11	JFK	1967
2007	11	SFO	2524
2007	11	LAX	3422
2007	12	LAX	5439
2007	12	JFK	2947
2007	12	SFO	3432
2007	2	SFO	2751
2007	7	JFK	3761
2007	7	LAX	4360
2007	8	JFK	3752
2007	8	LAX	4212
2007	8	SFO	3007
2007	9	SFO	2173
2007	9	JFK	1728
2007	9	LAX	2716

+----+-----+------+-------------------------+
|year|month|origin|number_of_delayed_flights|
+----+-----+------+-------------------------+
2006	1	LAX	3008
2006	1	JFK	1924
2006	1	SFO	2128
2006	10	LAX	3722
2006	10	JFK	2444
2006	10	SFO	2174
2006	11	LAX	3210
2006	11	SFO	2559
2006	11	JFK	2618
2006	12	SFO	3016
2006	12	LAX	4401
2006	12	JFK	3257
2006	2	LAX	3436
2006	2	SFO	2474
2006	2	JFK	1548
2006	2	JFK	1548
2006	3	SFO	3403
2006	3	JFK	1470
2006	3	LAX	4201
2006	4	JFK	1321
2006	4	LAX	3518
2006	4	SFO	2649
2006	5	SFO	1949
2006	5	JFK	1255
2006	5	LAX	3007
2006	6	JFK	2047
2006	6	SFO	2530
2006	6	LAX	3462
2006	7	JFK	2860
2006	7	SFO	2287
2006	7	LAX	3798
2006	8	SFO	2512
2006	8	LAX	3931



```
|2007|2    |SFO   |2751                |
|2007|2    |JFK   |2913                |
|2007|2    |LAX   |3754                |
|2007|3    |LAX   |3494                |
|2007|3    |JFK   |3664                |
|2007|3    |SFO   |1805                |
|2007|4    |SFO   |2126                |
|2007|4    |JFK   |3319                |
|2007|4    |LAX   |3191                |
|2007|5    |JFK   |2477                |
|2007|5    |LAX   |2682                |
|2007|5    |SFO   |2287                |
|2007|6    |LAX   |3779                |
|2007|6    |SFO   |2962                |
|2007|6    |JFK   |3653                |
|2007|7    |SFO   |3185                |
|2007|7    |JFK   |3701                |
```

b. Distributed by hour of day

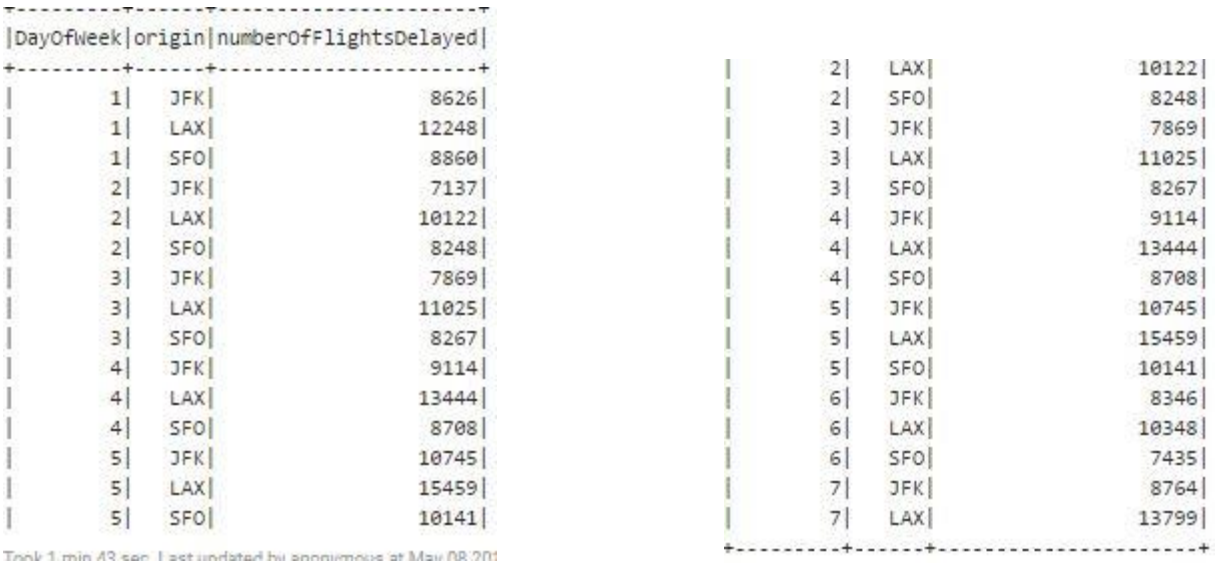
```
import scala.collection.mutable.WrappedArray
import spark.implicits._
import org.apache.spark.sql.functions._

 val spark = org.apache.spark.sql.SparkSession.builder
      .master("local")
      .appName("Spark CSV Reader")
      .getOrCreate;

val ds2006 = spark.read
      .format("com.databricks.spark.csv")
      .option("header", "true") //reading the headers
      .option("mode", "DROPMALFORMED")
      .load("/usr/nishchal/Nishchal/bigData/flight/2006.csv");
val ds2007 = spark.read
      .format("com.databricks.spark.csv")
      .option("header", "true") //reading the headers
      .option("mode", "DROPMALFORMED")
      .load("/usr/nishchal/Nishchal/bigData/flight/2007.csv");

ds2006.registerTempTable("flightDelays_2006");
ds2007.registerTempTable("flightDelays_2007");
```

```
val df2=spark.sql("select cast(CRSDepTime/100 as int) as timeOfDay,origin from flightDelays_2006 where DepDelay>=15 and Origin in ('JFK','LAX','SFO') union all select cast(CRSDepTime/100 as int) as timeOfDay,origin from flightDelays_2007 where DepDelay>=15 and Origin in ('JFK','LAX','SFO')");
df2.registerTempTable("timeOfDayAndOrigin");
val df3=spark.sql("select timeOfDay,origin,count(*) as numberOfDelayedFlights from timeOfDayAndOrigin group by timeOfDay,origin order by timeOfDay,origin");
df3.show(72,false);
```

```
+---------+------+----------------------+
|timeOfDay|origin|numberOfDelayedFlights|
+---------+------+----------------------+
|0        |JFK   |2                     |
|0        |LAX   |699                   |
|0        |SFO   |369                   |
|1        |LAX   |104                   |
|1        |SFO   |12                    |
|2        |JFK   |1                     |
|5        |JFK   |165                   |
|5        |LAX   |31                    |
|5        |SFO   |3                     |
|6        |JFK   |986                   |
|6        |LAX   |1451                  |
|6        |SFO   |977                   |
|7        |JFK   |1297                  |
|7        |LAX   |2607                  |
|7        |SFO   |1745                  |
|8        |JFK   |4001                  |
|8        |LAX   |3516                  |
|8        |SFO   |2496                  |
|9        |JFK   |3108                  |
|9        |LAX   |3736                  |
|9        |SFO   |2441                  |
|10       |JFK   |1763                  |
|10       |LAX   |5392                  |
|10       |SFO   |4883                  |
|11       |JFK   |1693                  |
|11       |LAX   |5360                  |
|11       |SFO   |4614                  |
|12       |JFK   |2337                  |
|12       |LAX   |6739                  |
|12       |SFO   |6067                  |
|13       |JFK   |2920                  |
|13       |LAX   |5820                  |
|13       |SFO   |5265                  |
|14       |JFK   |1911                  |
|14       |LAX   |5701                  |
|14       |SFO   |4090                  |
|15       |JFK   |3281                  |
|15       |LAX   |5369                  |
|15       |SFO   |4421                  |
|16       |JFK   |7596                  |
|16       |LAX   |6293                  |
|16       |SFO   |3643                  |
|17       |JFK   |6982                  |
|17       |LAX   |5596                  |
|17       |SFO   |2624                  |
|18       |JFK   |4540                  |
|18       |LAX   |7170                  |
|18       |SFO   |4302                  |
|19       |JFK   |6366                  |
|19       |LAX   |4972                  |
|19       |SFO   |2534                  |
|20       |JFK   |6326                  |
|20       |LAX   |5324                  |
|20       |SFO   |3497                  |
|21       |JFK   |3275                  |
|21       |LAX   |4070                  |
|21       |SFO   |2007                  |
|22       |JFK   |1774                  |
|22       |LAX   |4114                  |
|22       |SFO   |3613                  |
|23       |JFK   |277                   |
|23       |LAX   |2381                  |
|23       |SFO   |1131                  |
+---------+------+----------------------+
```

c. Distributed by Day of week

```
import scala.collection.mutable.WrappedArray
import spark.implicits._
import org.apache.spark.sql.functions._

 val spark = org.apache.spark.sql.SparkSession.builder
       .master("local")
       .appName("Spark CSV Reader")
       .getOrCreate;

val ds2006 = spark.read
       .format("com.databricks.spark.csv")
       .option("header", "true") //reading the headers
       .option("mode", "DROPMALFORMED")
       .load("/usr/nishchal/Nishchal/bigData/flight/2006.csv");
val ds2007 = spark.read
       .format("com.databricks.spark.csv")
       .option("header", "true") //reading the headers
       .option("mode", "DROPMALFORMED")
       .load("/usr/nishchal/Nishchal/bigData/flight/2007.csv");

ds2006.registerTempTable("flightDelays_2006");
ds2007.registerTempTable("flightDelays_2007");
val df2=spark.sql("select DayOfWeek,origin from flightDelays_2006 where DepDelay>=15 and Origin in
('JFK','LAX','SFO') union all select DayOfWeek,origin from flightDelays_2007 where DepDelay>=15 and
Origin in ('JFK','LAX','SFO')");

df2.registerTempTable("dayWiseDelays");

val df3=spark.sql("select DayOfWeek,origin,count(*) as numberOfFlightsDelayed from dayWiseDelays
group by DayOfWeek,origin order by DayOfWeek,origin");
df3.show();
```



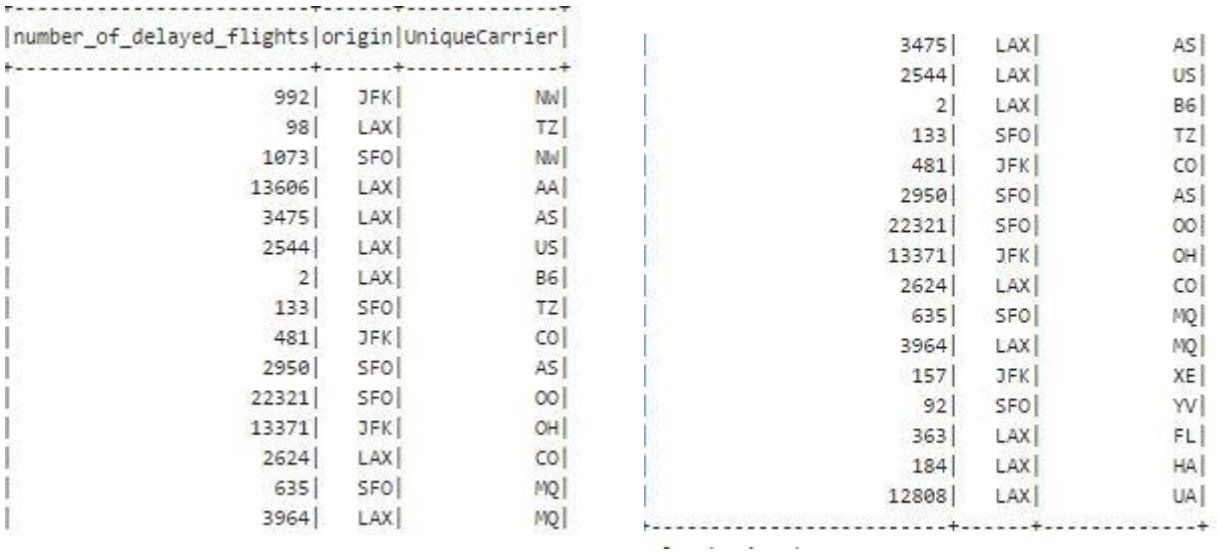
DayOfWeek	origin	numberOfFlightsDelayed
1	JFK	8626
1	LAX	12248
1	SFO	8860
2	JFK	7137
2	LAX	10122
2	SFO	8248
3	JFK	7869
3	LAX	11025
3	SFO	8267
4	JFK	9114
4	LAX	13444
4	SFO	8708
5	JFK	10745
5	LAX	15459
5	SFO	10141
6	JFK	8346
6	LAX	10348
6	SFO	7435
7	JFK	8764
7	LAX	13799

Took 1 min 43 sec. Last updated by anonymous at May 08 201

d. Distributed by Carrier

```
import scala.collection.mutable.WrappedArray
import spark.implicits._
import org.apache.spark.sql.functions._

 val spark = org.apache.spark.sql.SparkSession.builder
      .master("local")
      .appName("Spark CSV Reader")
      .getOrCreate;

val ds2006 = spark.read
      .format("com.databricks.spark.csv")
      .option("header", "true") //reading the headers
      .option("mode", "DROPMALFORMED")
      .load("/usr/nishchal/Nishchal/bigData/flight/2006.csv");
val ds2007 = spark.read
      .format("com.databricks.spark.csv")
      .option("header", "true") //reading the headers
      .option("mode", "DROPMALFORMED")
      .load("/usr/nishchal/Nishchal/bigData/flight/2007.csv");

ds2006.registerTempTable("flightDelays_2006");
ds2007.registerTempTable("flightDelays_2007");
```

```
val df2=spark.sql("select UniqueCarrier,origin, count(*) as number_of_delayed_flights from flightDelays_2006 where DepDelay>=15 and Origin in ('JFK','LAX','SFO') group by UniqueCarrier,origin UNION ALL select UniqueCarrier,origin, count(*) as number_of_delayed_flights from flightDelays_2007 where DepDelay>=15 and Origin in ('JFK','LAX','SFO') group by UniqueCarrier,origin");

df2.registerTempTable("temp");

val df3=spark.sql("select sum(number_of_delayed_flights)as number_of_delayed_flights,origin,UniqueCarrier from temp group by origin,UniqueCarrier");

df3.show();
```



number_of_delayed_flights	origin	UniqueCarrier
992	JFK	NW
98	LAX	TZ
1073	SFO	NW
13606	LAX	AA
3475	LAX	AS
2544	LAX	US
2	LAX	B6
133	SFO	TZ
481	JFK	CO
2950	SFO	AS
22321	SFO	OO
13371	JFK	OH
2624	LAX	CO
635	SFO	MQ
3964	LAX	MQ
3475	LAX	AS
2544	LAX	US
2	LAX	B6
133	SFO	TZ
481	JFK	CO
2950	SFO	AS
22321	SFO	OO
13371	JFK	OH
2624	LAX	CO
635	SFO	MQ
3964	LAX	MQ
157	JFK	XE
92	SFO	YV
363	LAX	FL
184	LAX	HA
12808	LAX	UA