AVIATION DATA ANALYSIS

Dataset Used:

Delayed_Flights.csv
https://drive.google.com/file/d/0B_Qjau8wv1KoWTVDUVFOdzlJNWM/view?usp=sharing

Delayed_Flights.csv Datasets
There are 29 columns in this dataset. Some of them have been mentioned below:
• Year: 1987 – 2008

• Month: 1 – 12

• FlightNum: Flight number

• Canceled: Was the flight canceled?

• CancelleationCode: The reason for cancellation.

For complete details, refer to this link.

- Creating a Spark Session Object

```
package com.spark.streaming

import org.apache.log4j.{Level, Logger}
import org.apache.spark.sql.SparkSession
import org.apache.spark.sql.functions._

object SparkMLIB {
  def main(args: Array[String]): Unit = {

    val spark = SparkSession
      .builder()
      .master("local")
      .appName("Spark MLIB")
      .config("spark.some.config.option", "some-value")
      .getOrCreate()

    println("Spark Session Object created")
```

- Removing All INFO Logs from terminal

```
// Removing all INFO logs in console printing only result sets
val rootLogger = Logger.getRootLogger()
rootLogger.setLevel(Level.ERROR)
```

- Creating Data frame out of CSV file
- Create Temporary table for SQL queries

```
val Flight = spark.read.format("CSV").option("header", true).load("C:\\Users\\lenovo\\Downloads\\DelayedFlights.csv")
val Fl = Flight.toDF()
// Flight.show()
Fl.registerTempTable("Flights_Table")
println("Flights_Table is registered!")
```

AVIATION DATA ANALYSIS

Problem Statement 1

Find out the top 5 most visited destinations

```
//Find out the top 5 most visited destinations.
val dest = spark.sql("""select Dest,count(dest) as Visits from Flights_Table  group by Dest """).toDF()
dest.sort(desc("Visits")).show(5)
println("Top 5 most visited destinations are as above!")
```

Ans:

```
18/08/29 00:58:13 INFO BlockManagerMaster: Registered BlockManager BlockManagerId(driver, DESKTOP-RQKF6HV, 64422,
18/08/29 00:58:13 INFO BlockManager: Initialized BlockManager: BlockManagerId(driver, DESKTOP-RQKF6HV, 64422, None
Spark Session Object created
Flights_Table is registered!
+----+------+
|Dest|Visits|
+----+------+
| ORD|108984|
| ATL|106898|
| DFW| 70657|
| DEN| 63003|
| LAX| 59969|
+----+------+
only showing top 5 rows

Top 5 most visited destinations are as above!
```

Problem Statement 2

Which month has seen the most number of cancellations due to bad weather?

```
//Which month has seen the most number of cancellations due to bad weather?
val cancel = spark.sql("""select Month,count(FlightNum) as Num_Flights_Cancelled from Flights_Table WHERE Cancelled =1 AND
CancellationCode="B" group by Month """).toDF()
cancel.show()
println("Month which has seen the most of the number of cancellations due to bad weather are as above!")
```

Ans:

```
+-----+---------------------+
|Month|Num_Flights_Cancelled|
+-----+---------------------+
|   11|                   40|
|   10|                   17|
|   12|                  250|
+-----+---------------------+

Month which has seen the most of the number of cancellations due to bad weather are as above!
```

AVIATION DATA ANALYSIS

Problem Statement 3

Which route (origin & destination) has seen the maximum diversion?

```
//Which route (origin & destination) has seen the maximum diversion?
val diversion = spark.sql("""select Origin,Dest,count(FlightNum) as Max_Diversion from Flights_Table where Diverted =1 group
by Origin,Dest """)
diversion.toDF().sort(desc("Max_Diversion")).show()
println("List of Origin and destination that has seen the maximum diversion")
```

Ans:

```
+-----+---------------------+

Month which has seen the most of the number of cancellations due to bad weather are as above!
+------+----+-------------+
|Origin|Dest|Max_Diversion|
+------+----+-------------+
|   ORD| LGA|           39|
|   DAL| HOU|           35|
|   DFW| LGA|           33|
|   ATL| LGA|           32|
|   SLC| SUN|           31|
|   ORD| SNA|           31|
|   MIA| LGA|           31|
|   BUR| JFK|           29|
|   HRL| HOU|           28|
|   BUR| DFW|           25|
|   PHX| ORD|           24|
|   DEN| LGA|           23|
|   LAX| JFK|           23|
|   MCI| MDW|           23|
|   LGA| ORD|           22|
|   OTZ| OME|           22|
|   IAH| LGA|           21|
|   JFK| SFO|           20|
|   HLN| MSP|           20|
|   ATL| DFW|           20|
+------+----+-------------+
only showing top 20 rows

List of Origin and destination that has seen the maximum diversion
```

4: Run 6: TODO sbt shell Terminal Build 0: Messages

Compilation completed successfully with 1 warning in 13 s 453 ms (13 minutes ago)