DATAFRAME REVIEW QUESTIONS

1. Download following datasets and then upload them onto Databricks cluster

http://www.utdallas.edu/~axn112530/cs6350/data/flight-data/csv/2015-summary.csv

2. Let's analyze the flights data using a PySpark notebook

Create a dataframe using the above file with following options:
- "inferSchema" set to "true"
- "header" set to "true"

- Display the contents of the dataframe
- Display the first 3 rows of the dataframe
- Sort the dataframe on the count field in a descending order
- Display the summary statistics of each of the columns
- Import the following library
  from pyspark.sql.functions import max
  Find out the maximum value of the counts field.
- Import the following library
  from pyspark.sql.functions import desc
  We would like to find the top 5 countries that have the largest number of incoming flights.
  Hint: group by the DEST_COUNTRY_NAME field and generate a count of the flights by each group and then sort this count in a descending order and take top 5.

3. Let's download more files
   http://www.utdallas.edu/~axn112530/cs6350/data/orders/NW-Orders-01.csv
   http://www.utdallas.edu/~axn112530/cs6350/data/orders/NW-Order-Details.csv

   Upload them into Databricks cluster and answer the following commands

4.

- Find out the count of orders placed by each customer and then return the top 5 customers with the highest count of orders.

- Find out the count of orders placed by each customer and then return the top 5 customers with the highest count of orders.

- Join the orders and orderDetails data on the orderID column and display the results.

- Join the orders and orderDetails data on the orderID column and then group the data by ShipCountry field and find the sum of quantity purchased by each country. Then sort by the sum of quantity field and list the top 10 countries.