IIAggregation

On the GDELT dataset, perform the following

1. Create a report which shows ActionCountry_Code wise, # of events occured :combinedbykey/mapvalues+reducebykey/aggregatebykey+iterate and count for each key./countbykey.
2. Create a report which shows Month-wise, ActionCountry_Code wise, # of events occured
   :Do
3. Create a report which shows Top-10 days in a month, when the most # of events occured.
   :secondary sorting-done.
4. Create a report which shows Month-on-Month % increase or % decrease in # of events. For example, show a report as follows : Only spark-sql?

Month Year	# of events	% Change
May 2017	50,000	0
June 2017	30,000	-40%
July 2017	24,000	-20%
Aug 2017	53,000	120%
Sep 2017	75,000	41.5%
Oct 2017	23,000	-69%

5. Create a report which shows the distinct count of countries who had atleast 100 events in the month of May 2017 : countbykey
6. Write a Program which will give a report in this fashion.

May 2017 (Month Year column)	50,000
US (Actor1Geo_CountryCode)	70,000
Event Type (EventCode)	110,000
Quad Class	342,000

7. Write a Program to compute the mean of NumArticles by every Month Year

8. Write a program which will print the day with the lowest # of NumArticles for a given MonthYear
9. Write a program which will print the GlobalEventID and the Total # of events happened on that day. For example , if there were 2 events happened on that day, it should display the eventid and total # of events happened on that day.

Joins & Transformation

1. Download a file which contains the demography information of every country. Join the event dataset with the demography dataset on country_code and report the population & # of events
2. Given the following dataset,

CountryCode	MonthYear	NumArticles

convert it to the following format

CountryCode	Jan 2017	Feb 2017	Mar 2017….	Apr 2017
	NumArticles	NumArticles	NumArticles	NumArticles
	NumArticles	NumArticles	NumArticles	NumArticles

3. Write a program which will print

Month Year	Num Articles Band	Sum of Num of Articles
2017 May	1- 100	50000
2017 May	100- 1000	60000

Debugging

1. Run a Spark Command to describe the statistics for NumArticles column like min, max

2. Write a spark program to return the first & last line of a file

Data Quality

1. Write a program which will drop any duplicate events. You can identify duplicate events by the SOURCEURL field -- done
2. Write a program which will drop any record which has even one column with value NULL.-- done
3. Write a program which will drop any record which has 50% of its columns with NULL value.
4. Write a program which will drop any record which has NULL in the Actor1EthnicCode column ---- done
5. Write a program which will list the columns and datatypes of the data frames --- done
6. Write a program which will find NULLs in Actor2EthnicCode and replace NULLs with the Actor2EthnicCode that had the most # of events last month.
7. Write a program to filter out all records whose GoldStein Scale < 0 and AvgTone < 0 --- done -- done
8. Write a program which will do a Camel_Case for all values in Actor1Geo_Fullname
9. Write a program to replace all NULL values in any field with -99 -- done
10. Write a program which will create a data frame with a few selected columns from the previous dataset --- done
11. Write a program which will cast the column NumArticles to Integer   --done
12. Write a program which will show all records whose AvgTone was between 10 & 25 and whose NumArticles was > 100 ---done
13. Write a program which will show all records whose Actor1Name contains ‘modi’ or ‘trump’ -- done
14. Write a program which computes the average of NumArticles of last 13 rolling months --- done

Advanced

1. Write a program which will persist the intermediate steps of the spark pipeline
2. Write a program to fill in the previous day’s event count, if for that day event count is empty (Gap filling)