Dataset & Problem Statement

This is a compulsory assignment on SparkSQL and SparkR based data processing. You will get a document with instructions to code. You need to write the SparkSQL, SparkR commands and your comments about those commands below each instruction and submit the document (only the commands and your explanation about the commands used is needed).

The deadline for the submission of this assignment is **January 29, 2017, 11:59 PM**. For submissions obtained within 1 week of the deadline, there will be a 30% penalty. Submissions beyond 1 week of the deadline will not be accepted.

Mixpanel Event's Dataset

In this assignment, we will use UpGrad's ad events dataset. As you know, UpGrad provides different courses in different domains. The marketing team of UpGrad conduct various ad campaigns to promote these courses. Online digital marketing methods are used to host ads on various channels. The typical medias are Social networking websites, ads on search engines, links on IIITB's and UpGrad's websites, etc. The prospective student can reach the intended program web page through any of these ads/links. The prospective student may further visit various pages hosted on UpGrad's website. He/She may go through the course details, faculty pages, raise a query for a particular course, fill a registration form, and perform other actions. UpGrad collects and tracks all this website access data to decide their future marketing actions. The data help them identify the effective ad channels, ad campaigns, and the web page portability on various operating systems and web browsers. This assignment is aimed to help UpGrad's internal marketing team in deciding their future marketing actions by analysing the given dataset using SparkSQL and SparkR. As a first step towards this aim, let us understand the dataset. The data dictionary is given below:

Metadata/Data Dictionary

Sr. No.	Field Name	Description
1	event	This field lists all the possible events that can happen on the web page.
2	time	This is the timestamp of the event. Which specifically gives the timestamp value of when this event happened.
3	distinct_id	This is a random number generated by the system for each user.
4	city	This gives the location of the user.
5	current_url	This field gives the URL of the current webpage the user is on
6	initial_referrer	This field gives the URL of the website through which the user reached UpGrad's website for the first time.
7	initial_referring_domain	The domain name for the first visit of the user.
8	OS	Name of the operating system through which the user visited the website.
9	referrer	This field gives the URL of the website through which the current visit occurred.
10	referring_domain	The domain name for the current visit.
11	region	Gives the state in a country.
12	course	The name of the course whose web page is visited.

13	latest_utm_campaign	The tag used for the campaign.
14	latest_utm_content	The tag used for the method of reach out.
15	latest_utm_medium	The name of the communication in the campaign.
16	latest_utm_source	The source of the campaign.
17	page_title	The name of the page visited.
18	user_requested_DA_info	This is a boolean field, specifying the user's choice of information on the DA programs.
19	user_viewed_DA_prog_details	This is a boolean field, indicating the view of DA program page.
20	utm_campaign	The tag for initial campaign.
21	utm_content	The tag used in the initial campaign for the method of reach out.
22	utm_medium	The name of the communication used in the initial campaign.
23	utm_source	The source of the campaign.
24	utm_term	The URL of the page visited.

The data for this assignment is available in our s3 bucket. The path for the dataset is **'s3://ughdfsdemo/Spark_Assignment_data/ugmixp_data'**. Using this dataset, you need to perform some analysis using SparkSQL and SparkR. The general checkpoints are given below:

Checkpoint 1: Understanding the data

This is the most important stage of solving any problem. At this checkpoint, you are supposed to get well versed with the dataset. There is a question in your assignment which is based on this understanding.

Checkpoint 2: Creating the data tables and data frames for analysis

At this checkpoint, you are expected to use DDL statements/SparkR commands and create the necessary data tables/data frames. This step needs very keen observation about the datatypes to be used while creating tables/data frames. With the thorough understanding of data in hand, you can easily decide and formulate the commands for the same.

Checkpoint 3: Performing basic analysis

At this checkpoint, you are expected to use DML statements/basic SparkR commands to get basic insights from this data. In this section, the commands will be simple and straightforward. These commands will form the base for complex analysis that is the focus of the next checkpoint.

Checkpoint 4: Performing complex analysis

At this checkpoint, you are expected to use DDL/DML statements/SparkR commands and your prior knowledge about SQL/R commands to get more complex insights. The tasks at this checkpoint will include the results from multiple commands to get final insights. This checkpoint will demand the use of a chain of commands to arrive at the expected solution.

NOTE: Keep on writing about each and every command that you are using to arrive at the final insight. Your comments/explanation about each command will help you in doing so.

Assignment Submission

The deadline for the submission of this assignment is **January 29, 2017, 11:59 PM**. For submissions obtained within 1 week of the deadline, there will be a 30% penalty. Submissions beyond 1 week after the deadline will not be accepted.

Instructions

Download the SparkAssignment.docx file from the bottom of this page and follow the instructions to complete your submission.

Marks Distribution

This assignment is **worth 700 points.**

There are 12 tasks in total in this assignment for which you need to write the applicable SparkSQL or SparkR command.

Submission

Submit the completed **.docx** file through the submission link below. Name it **DDA_Spark.docx**.

IMPORTANT NOTE: Please name your submission file correctly using the above naming convention to avoid confusion during the grading process.

SUBMIT CODE HERE (filename = DDA_Spark.docx)