Group HW 3
Digital Marketing Analytics

(NO extensions---please start as soon as possible)

This constitutes 15% of your total course grade

E-commerce datasets are difficult to obtain in general. One of the few public datasets is one from gazelle.com from an old KDDCup contest from the early period of e-commerce. There are a few proposed solutions to the challenge on the internet, but not necessarily to the questions I ask below.

Background (not all of it may be relevant):

It is helpful to know the following background information about the Gazelle.com webstore:

• The home page contained more than 70 images. This made downloads extremely slow for modem-based visitors.

• As with many dot-coms, Gazelle.com's initial goal was to attract customers, even if it meant losing money in the short term. They had many promotions that are relevant for analytics, because promotions affected traffic to the site, the type of customers, etc. The important promotions were

- FREE - Free shipping ($3.95 value) active from March 20 to April 30 (shipping was normally free if sale was above $40)
- MARCH1 - $10 off from March 1 to April 1
- FRIEND - $10 off from March 1 to April 30
- FREEBAG – A free bag from March 30 to April 30.

Note that both the MARCH1 and FRIEND promotions offered $10 off. They were used for different purposes, and were run with different promotion codes.

• Gazelle.com ran a TV advertisement during a prime-time episode of the popular comedy show, Ally McBeal, on February 28.

• Gazelle.com changed their registration form significantly on February 26, so some customer attributes were only collected prior to this date and some were collected only after this date.

Questions:

a) Based on the dataset q3, what can you say about the high-spenders (define them as > some cut-off)?
b) (Based on q1 and q2 datasets) This is in the nature of recommendation engines, but not quite (there is no need to choose the most advanced method, but something that is transparent and gives insight will do). Analyse the customer journey on the website and which groups of pages they are viewing in one session (define this). Develop some insight based on the data that you can present to management.

Submit your code, as well as a report of up to 6 pages, detailing (a) what you did on the data cleaning and manipulation part (b) how you analysed (c) Your findings.

Your report will also be graded for clarity and exposition.

Data *(the .names file contain the description; please infer what the columns mean from them):*

Gazelle.com went live on January 30, 2000 with soft-launch to friends and families. On the webstore, an application server generates web pages from Java based templates. Among the other things, the architecture logs customer transactions and clickstreams at the application server layer.

Since the application server generates the content (e.g., images, products and articles), it has detailed knowledge of the content being served. This is true even when the content is dynamically generated or encrypted for transmission commonly used for checkout.

Weblog data is not needed. Application servers use cookies (or URL encoding in the absence of cookies) to keep track of a user's session, so there is no need for "sessionising" clickstreams as there is for standard weblogs. Since the application server also keeps track of users using login mechanisms or cookies, it is easy to associate individual page views with a particular visitor.

Among the data collected by the server, the following three categories are relevant:

• Customer information, which includes customer ID, registration information, and registration form questionnaire responses.

• Order information at two levels of granularity: 1) Order header, which includes date/time, discount, tax, total amount, payment, shipping, status, and session ID; and 2) Order line, which includes quantity, price, product, date/time, assortment, and status.

• Clickstream information at two levels of granularity: 1) Session, which includes starting and ending date/time, cookie, browser, referrer, visit count, and user agent; and 2) Page view, which includes date/time, sequence number, URL, processing time, product, and assortment.

In general, each customer can have multiple sessions. Each session can have multiple page views and multiple orders. Each order can have multiple order lines. Each order line is a purchase record of one product with a quantity of one or more.

Data is given in an aggregated format *(please understand the data from the names files)*

The aggregated data consists of three datasets. These datasets are derived by aggregating the two unaggregated datasets containing clicks and orders to the level of granularity appropriate for mining.

At the same time, they have some new attributes based on examination of existing attributes. For example, added are the session browser family names, the browser names, and the top three browser family names.

In the two aggregated datasets for q1, q2, each session is a single record. In the aggregated dataset for q3, each customer is a single record.

The aggregation operations generated 151 and 153 new attributes for q1 and q2, respectively. Examples include the number of views of individual top products which were selected based on the statistics of the datasets, the number of views of assortments, the number of views of different templates, and information about the last page, which includes information appearing on it and its date/time information.

For q2, we defined three numeric attributes indicating the number of views of the respective brands (Hanes, Donna Karan, American Essentials) in the remainder of the session. In addition, we also defined a Boolean attribute that was set to true if none of the brands were viewed in the remainder of the session and false otherwise.

When generating the aggregated dataset q3, we joined clickstream data to the order lines data since we believed that clickstream data is hard to join them after aggregation. The aggregation for this dataset was carried out at two levels: first to the session level and then to the customer level, generating 434 new attributes in total such as “Average Session Request Count”, “First Session First Referrer Top 5”, and “Percent of Products Purchased on Sunday”.