E-shopping Clickstream Dataset

**Group 4**

**Ali Al Hady**

**Hemanth Reddy Musuku**

**Siddhartha Setty**

**Venkata Dharma Teja Nunna**

**About the Dataset**

The dataset contains information on clickstream from online store offering clothing for pregnant women. Data are from five months of 2008 and include, among others, product category, location of the photo on the page, country of origin of the IP address and product price in US dollars.

The original dataset itself contains numerical values for many categorical in the 14 columns that needs to be handled in pre-processing.

The columns include year, month, day (number),order, country(number of ip address of location), session ID, main category(of product), clothing model, colour, location (of the photo in the page), photography, price, price 2(Boolean indicating if the price is greater than average price for the category.).

**Objectives**

Some basic questions(tentative) that we wish to answer include:

When do sales peak?

What type/categories of clothing sells most, and is there any correlation between the time of the year and sales?

Which countries, according to the IPs, have contributed to more sales, and when?

How much effect, if any, does the placement of the photo on the page have on its sales?

How accommodating are people towards products that are priced above the average price in their category, and how are their sales affected because of this?

Overall, what factors contributed to a particular product being successful in sales?

**Initial Thoughts and Direction**

We plan to use Python to perform pre-processing, and then use the cleaned dataset in Tableau.

We anticipate we would need around 4-5 dashboards, each having 2-3 visualizations, to answer the above questions. In addition to this, since these questions were surmised from an initial glance at the data, we expect to refine the questions further as we start our project.