**2. Details about Data**

**A. Acknowledging source of data**

As a part of the project we are planning to do analysis on data collected from James M. Klits Center, University of Chicago Booth School of Business. Total size of this data add up to 4.76 GB needs lot of changes to convert this dirty data to perform meaningful analysis. The data comprises of 9 years of store level data of more than 3500 UPCs which was sold through around 100 stores across United States. Most of the Stores are located in Chicago area. We need to analyze the data from 1989-1994 for same number of stores. The complete product line is classified in to 29 different categories.

**B. Understanding of the data**

In total there are 4 data files and they can be classified in to 2 categories namely General files and Category files. These 4 data files are available in .csv format for the analysis of the sales of Dominick’s FF

a) Customer Count Files

* This file contains the information about in-store traffic of all the stores which was compiled on a daily basis from registers and scanners located at each store in DFF.
* This table contains the information about the sale information of each product categories of Beer, Meat, grocery and Diary etc. store wise on weekly basis.
* This file also contains the separate information of each above product purchased using Coupons at each store in DFF chain on weekly basis.

b) Store-Specific Demographics

* This file contains the store wise information of all the customers purchased the products from DFF on weekly and demography basis.
* Information present in this file is obtained by mapping the customer information with the Census information collected by US government for Chicago Metropolitan area.
* Various demography information available in this file include age groups, household income, number of dependent members, employment status, retired status of every customer.
* Demographic information is the most important in later stages of the project for Data Warehousing and building different store wise strategies targeting different demography of people. This file have very importance in framing the questions in this phase of the project.

c) UPC Files

* As the name indicates UPC means Unique Product Code. Each Product is mapped to a UPC and other information related to that particular product.
* Complete mapping of UPC code to product is available in the “List of all UPCs in the category ” table . This information can be used for product specific strategies for DFF.

d) Movement Files

* This file contain category wise weekly movement of each product in DFF.
* Information in this file give clear idea about profit margin on each product. This in turn will give idea about strategies need to be adopted by DFF to reduce losses and increase profit. This very important in Business point of view.
* If the Sale quantity is predictable after analyzing weekly sale data , it will give strategic advantage to the inventory department of DFF in preparing for peak and off seasons.

In addition to the above data sources we have “Weeks Decode Table” which gives the week to date mappings useful for analysis. This is very useful in formulating different business strategies .

Metadata

Need to add ½ tables from pdf.