Big Data Integration Final Paper

Sidney Schaeper

2/20/2020

### Understanding Our Topic

### Describe The Tools Used

### Data Sources

The data that we wanted to evaluate for this project came from two sources. The first source was from [84.51’s website](https://www.8451.com/). The folder of data that we used from 84.51 are called the Carbo Loading datasets. There are a total of four datasets within this folder. The four datasets are called the casual\_lookup, product\_lookup, store\_lookup, and transactions. According to the Carbo Loading guide provided, these datasets came from a relational database. Also, these datasets contain purchases at a household level over two years, and the datasets are filtered to only contain products from four categories. These four categories are pasta, pasta sauce, syrup, and pancake mix. In addition, the guide contained information pertaining to the variables within each of the datasets. The tables below show these variables.

**Transactions Dataset Table**

|  |  |
| --- | --- |
| Variable | Description |
| upc | It is a standard 10 digit code assigned to products. This is the product that was purchased. |
| dollar\_sales | The amount of money spent on this product by the customer. These are recorded in dollars. |
| units | The quantity of this product purchased. |
| time\_of\_transaction | The time the transaction occurred. This is recorded in military time. |
| geography | This label tells you where it was purchased out of the two large regions. These two regions consist of multiple values. The value can either be 1 or 2. |
| week | This notifies the week that the transaction occured. The range of values is from 1 to 104. These numbers are assigned chronologically. |
| household | This value is a unique number assigned to a household. This is the purchaser of the product. |
| store | This value is a unique number assigned to each store. This is where the product was purchased. |
| basket | This is a unique number assigned to a trip to the store. This is the trip that this product was assigned to. |
| day | This is the day that this product was purchased. The range of values is 1 to 728. |
| coupon | This is dummy variable to notify whether a coupon was used. The possible value is 1 or 0. 1 means a coupon was used. 0 means a coupon wasn’t used. |

**Store Lookup Dataset Table**

|  |  |
| --- | --- |
| Variable | Description |
| store | This value is a unique number assigned to each store. |
| store\_zip\_code | This is the 5 digit zip code for the store. |

**Product Lookup Dataset Table**

|  |  |
| --- | --- |
| Variable | Description |
| upc | This is the standard 10 digit code assigned to this product. |
| product\_description | This details the product. This likely contains the name of the product. |
| commodity | This is the category of the product. The four possibly values are pasta, pasta sauce, pancake mix, or syrup. |
| brand | This is the brand name of the product. |
| product\_size | This is the size of the product. These aren’t all in the same measurement. |

**Casual Lookup Dataset Table**

### Data Integration Process

#### Schema Alignment

#### Record Linkage

#### Data Fusion

### Business Insights

### Analyzing Results

### Future Steps

### Bibliography

### Additonal Significant Code