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. |
| 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. |

### Data Integration Process

#### Schema Alignment

#### Record Linkage

#### Data Fusion

### Business Insights

### Analyzing Results

### Future Steps

### Bibliography

### Additonal Significant Code