**Project Title:**

Retail Promotional Discount Effectiveness

**Team Members:**

Scott Stevener and Erica Unterreiner

**Project Description:**

The company runs several promotional markdown events throughout the year. These markdowns precede prominent holidays; the four largest being the Super Bowl, Labor Day, Thanksgiving, and Christmas.

The objective of this dataset is to calculate the markdown and analyze how sales are affected by various market conditions to determine the best pricing strategies for the products sold. Understanding this information helps stores better manage their business to maintain/increase profitability.

**Data Cleanup After Importing into Python Pandas DataFrame & Analysis:**

**Steps to recreate the ETL Process:**

1. **Extract store, sales, and features data from three csv files.**

We obtained historical sales data for 45 stores located in different regions for a single company. The first data set includes discounts applied as well as market conditions information, such as the price of fuel, the consumer price index and whether there was a holiday during that week. The second data set includes aggregated weekly sales. The third data set includes information about each individual stores.

1. **Transform: what data cleaning or transformation was required.**

Once imported the data needed to be reviewed for cleanup.

1. Features data
   1. The date formats are inconsistent on a weekly basis; therefore, the dates are converted.
   2. Markdowns do not occur during all possible periods. For periods without a markdown, a NA value is placed in the field. These values are replaced with a zero.
   3. The IsHoliday field contains true/false values. These values are converted to one (1) and zero (0).
      1. 0 = False
      2. 1 = True
   4. Once changes were made, we tested the data types to aid in the setup of the table within MySQL. All data types were as expected so no additional cleanup was necessary.
2. Sales data
   1. The date formats are inconsistent on a weekly basis; therefore, the dates are converted.
   2. The IsHoliday field is an attribute of the Features table therefore, this column was removed.
   3. Once changes were made, we tested the data types to aid in the setup of the table within MySQL. All data types were as expected so no additional cleanup was necessary.
   4. We tested for incomplete rows and none were found.
3. Stores data
   1. There were no obvious changes to be made.
   2. We tested the data types to aid in the setup of the table within MySQL. All data types were as expected.
   3. We tested for duplicates and incomplete rows and none were found.
4. **Load: the final database, tables/collections, and why this was chosen.**

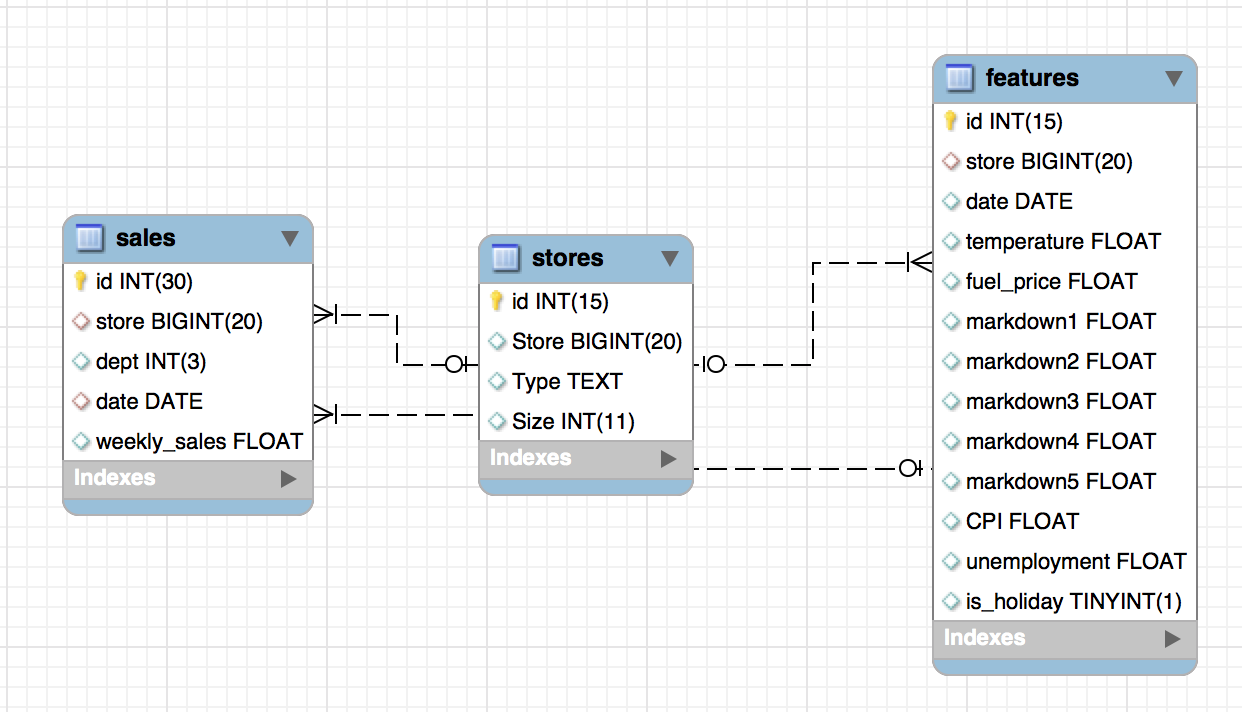
The store table is a domain table with information about the stores that made sales. The features data has implications for the store’s sales. Given the related nature, a relational database was selected to store the data.

In order to load the data from Python Pandas, we first needed to create a database and tables to hold the data. The “schema.sql” file contains the SQL to accomplish this task. Then the cleansed data was loaded into “retail\_db” via the Python Pandas “to\_sql” function. Upon completion of the load, all tables were tested using simple SELECT statements to make sure that the data loaded.

A “query.sql” file was prepared to aid an end-user with merging the data sets together for further analysis. Additionally, the entity relationship diagram and metadata were compiled to give analysts context into the meanings of the columns, usage information and a general idea of how the data is organized (see appendix).

**APPENDIX:**

**ENTITY RELATIONSHIP DIAGRAM**

****

**TABLE AND COLUMN METADATA FOR END USER ANALYSTS**

|  |  |
| --- | --- |
| Tables included in Entity Relationship Diagram | |
| Physical Name | Description |
| STORES | Anonymized information about the 45 stores, indicating the type and size of store. Stores has been divided into 3 types A, B, C, based on certain factors (size, sales, etc.). |
| FEATURES | Contains additional data related to the store, department, and regional activity for the given dates. |
| SALES | Historical sales data, which covers the time period of January 2012 through November 1, 2012. |
| RETAILED\_COMBINED | Target table with merge of stores, features and sales from January 2012 through November 2012. This table is to be used for calculating the markdown and analyzing how the profits are affected for determining the best pricing strategy for a product.  To calculate a markdown – using an item priced at $150, a 15 percent markdown is $22.50 ($150 x .15), making the sale price $127 ($150 minus $22.50). A 20 percent markdown is $30 ($150 x .20) with a sale price of $120 ($150 minus $30). Five potential markdowns means that the price of a specific product could be reduced up to five (5) times. |

|  |  |  |
| --- | --- | --- |
| Columns of table: STORES | | |
| Physical Name | Data Type | Comment |
| ID | INT(15) | The unique identifier of a record on this table |
| Store | BIGINT(20) | The identifier for each store |
| Type | TEXT | Store typing consists of categorizing stores by certain factors (e.g. size, sales, etc.) Valid values include A, B, C |
| Size | INT(11) | Square footage of each store |

|  |  |  |
| --- | --- | --- |
| Columns of table: FEATURES | | |
| Physical Name | Data Type | Comment |
| ID | INT(15) | The unique identifier of a record on this table |
| Store | BIGINT(20) | The identifier for each store |
| Date | DATE | Represents the start of the week |
| Temperature | FLOAT | Represents the average temperature in the region |
| Fuel\_Price | FLOAT | Represents the cost of fuel in the region |
| MarkDown1 | FLOAT | Represents anonymized data related to promotional markdowns. Any missing value is marked as zero. |
| MarkDown2 | FLOAT | Represents anonymized data related to promotional markdowns. Any missing value is marked as zero. |
| Markdown3 | FLOAT | Represents anonymized data related to promotional markdowns. Any missing value is marked as zero. |
| Markdown4 | FLOAT | Represents anonymized data related to promotional markdowns. Any missing value is marked as zero. |
| Markdown5 | FLOAT | Represents anonymized data related to promotional markdowns. Any missing value is marked as zero. |
| CPI | FLOAT | Represents the Consumer Price Index. |
| Unemployment | FLOAT | Represents the unemployment rate. |
| IsHoliday | TINYINT(1) | Indicator of whether or not the week contains a holiday. |

|  |  |  |
| --- | --- | --- |
| Columns of Table: SALES | | |
| Physical Name | Data Type | Comment |
| ID | INT(30) | The unique identifier of a record on this table |
| Store | BIGINT(20) | The identifier for each store |
| Dept | INT(3) | The identifier of the store department. |
| Date | DATE | Represents the week |
| Weekly\_Sales | FLOAT | Sales for a given department in a given store. |