Data Engineering Coursework 1

Name: Om Sadigale

ID: W1943544

Lecturer: Ingrid Membrillo Solis

Introduction:

RDBMS is Relational Database Management System, and it used to organise the information by storing data in columns and rows in one or more tables to understand the relationship between each entity.

The business scenario we are using here is of transaction data of the purchases made by customer in a store with assistance from Cashier. The transaction data contains the information about the products, customers, and cashiers with a reference to individual datasets. This relationship between the transactions and other datasets can be easily explained in the RDBMS.

1. Data Extraction (CSV Files)

For the project, I have selected 5 data files from the Kaggle website.

**Transaction.csv:**

The main data set consist of transactions for the purchase of various products by Customers assisted by Cashiers. There are total 9994 records in the file however only 6 months data is used for project. This data set consist of 11 attributes.

**Product.csv:**

The list of products available for sale is given in this data set. In total there are 1862 unique productid and ProductName combination with 6 attributes.

**Customer.csv:**

This data set is of the customers who purchase the products from the store. There are total 793 customers listed in the file without any duplication. There are 5 attributes of the Customer in the file.

**Staff.csv:**

This dataset is the list of Cashiers working on the counter of the store. There are only 3 Cashiers with 4 attributes in the dataset.

The Transaction.csv has the reference to Product, Customer and Cashier IDs in the dataset thus beneficial to establish relationship between the datasets.

**Categories.csv:**

This data set categories the Products and set of 17 different categories and has 2 attributes.

The transaction dataset does not directly refer to Categories and it establishes the relationship through product dataset.

1. Data Exploration:

The data sets selected in the Data Extraction process are further examined and analysed to understand it’s structure and contents.

For each data set below properties are noted:

* + Number of datapoints
  + Name of attributes
  + Type of attributes
  + Number of missing values
  + Entry Errors

**Transaction**:

* This dataset did not have any unique identifier attribute hence I had to add a Transaction\_ID attribute and set is as primary key.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Attributes (Columns) | Total No. of data points (Rows) | Type | Missing Values | Errors |
| Transaction\_ID | 9994 | VARCHAR | 0 | 0 |
| Transaction\_date | DATE | 0 | 0 |
| Quantity | INT | 0 | 0 |
| Subtotal | DOUBLE | 0 | 0 |
| Taxes | DOUBLE | 0 | 0 |
| DiscountPrice | DOUBLE | 0 | 0 |
| TotalAmount | DOUBLE | 0 | 0 |
| Payment\_Method | VARCHAR | 0 | 0 |
| Products\_ProductID | VARCHAR | 0 | 0 |
| CustomerCart\_CustomerID | INT | 0 | 0 |
| Cashier\_CashierID | INT | 0 | 0 |

Product:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Attributes (Columns) | Total.No of data points (Rows) | Type | Missing Values | Errors |
| Product\_ID | 1862 | VARCHAR | 0 | 0 |
| ProductName | VARCHAR | 0 | 0 |
| ProductStock | INT | 0 | 0 |
| Price | DOUBLE | 0 | 0 |
| Discount | INT | 0 | 0 |
| Category\_CategoryID | INT | 0 | 0 |

Customer:

CustomerID attribute is referenced in the Transaction dataset however it was not available in the Customer dataset hence I have added it.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Attributes (Columns) | Total No of data points (Rows) | Type | Missing Values | Errors |
| CustomerID | 793 | INT | 0 | 0 |
| CustomerName | VARCHAR | 0 | 0 |
| CustomerPhNo | INT | 0 | 0 |
| CustomerAddress | VARCHAR | 0 | 0 |
| Password | VARCHAR | 0 | 0 |

Category:

CategoryID attribute is referenced in the Product dataset however it was not available in the Categories dataset hence I have added it.

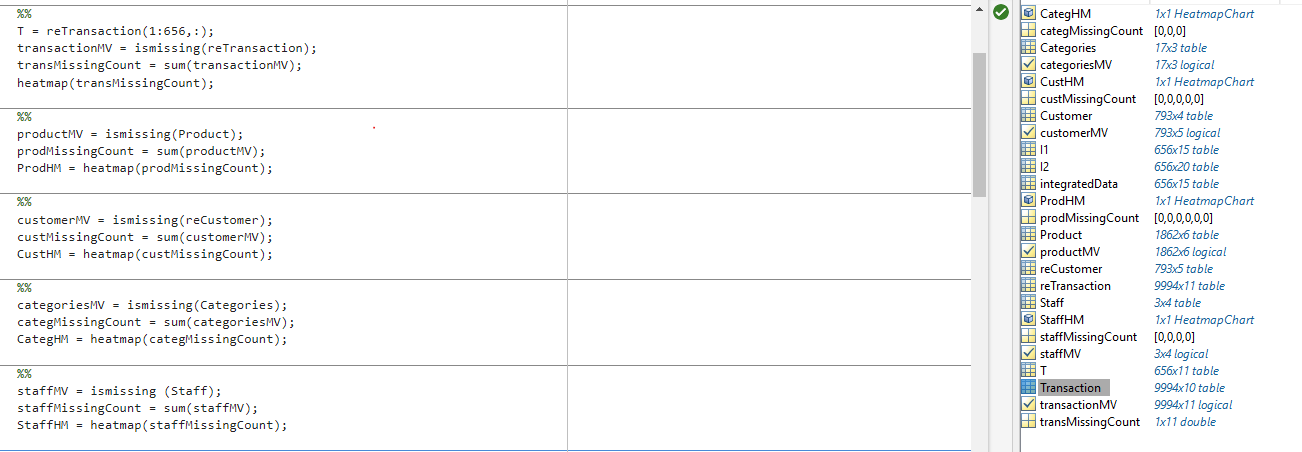
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Attributes (Columns) | Total No of data points (Rows) | Type | Missing Values | Errors |
| CategoryID | 17 | INT | 0 | 0 |
| Category | VARCHAR | 0 | 0 |
| SubCategory | VARCHAR | 0 | 0 |

Staff:

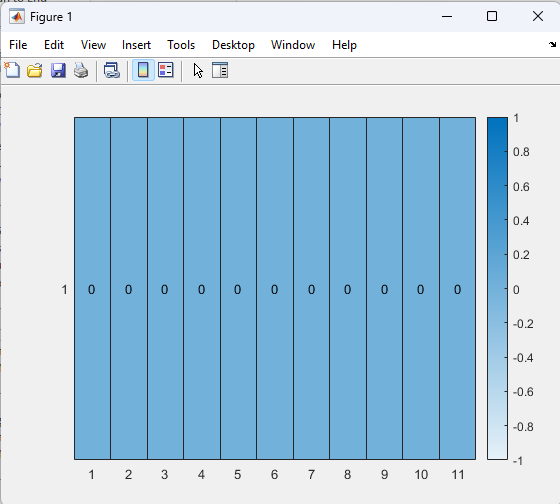
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Attributes (Columns) | Total No of data points (Rows) | Type | Missing Values | Errors |
| CashierID | 3 | INT | 0 | 0 |
| CashierName | VARCHAR | 0 | 0 |
| CashierPhoneNo | INT | 0 | 0 |
| CashierAddress | VARCHAR | 0 | 0 |

**Heatmap to check missing values:**

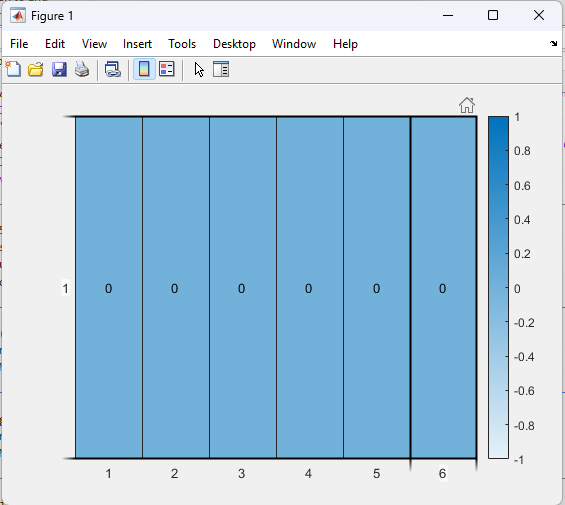
* In Matlab, the missing function was used to search for the missing data in a specific csv file.
* For further elaboration I used a heatmap to find any missing values. However, there were no missing values in any of csv files.



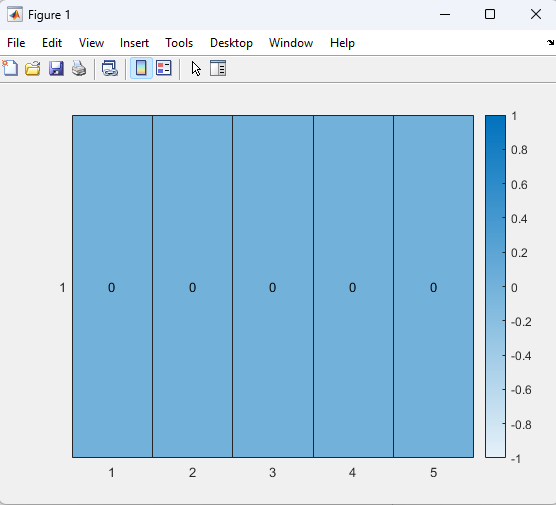
reTransaction Heatmap



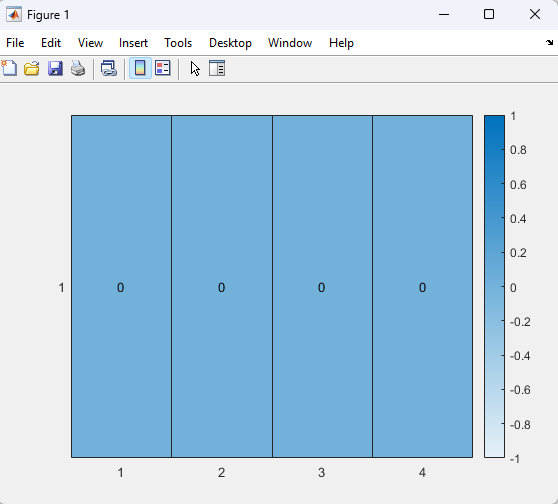
Product Heatmap



Customer Heatmap



Staff Heatmap



Data Transformation:

Data Cleaning:

As shown in the heatmap above there are no missing values in any of the attributes of all the data sets hence no handling of missing data was required.

Feature Engineering:

The four tasks chosen were:

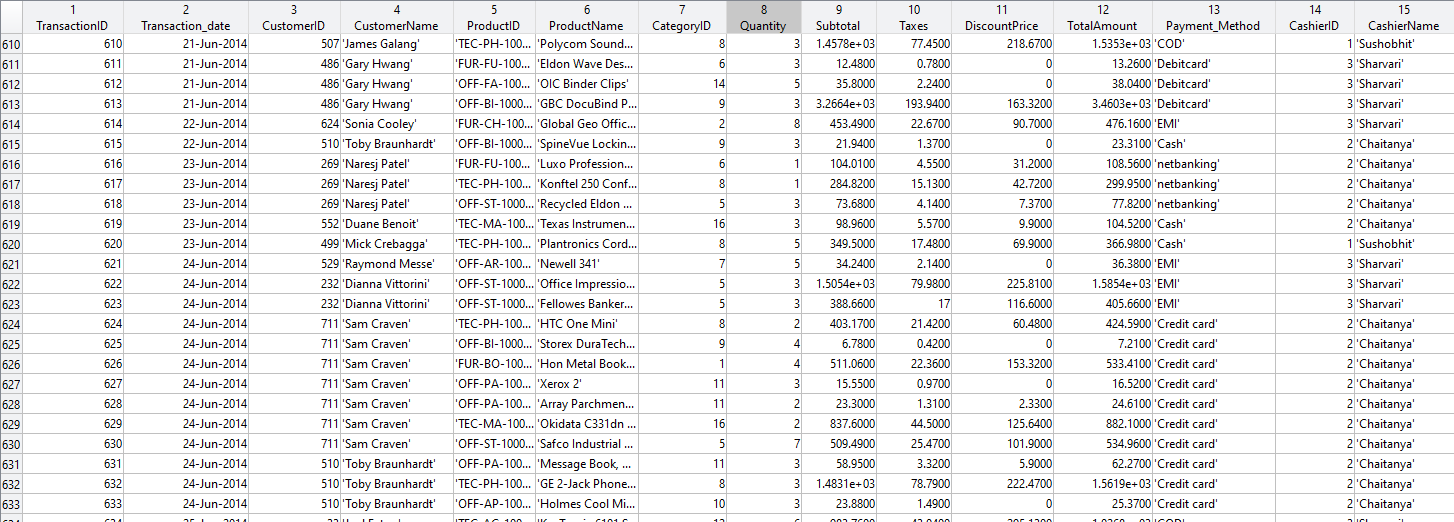
1. Integrating data from multiple CSV files
2. Removing columns of redundant features.
3. Aggregating purchase data
4. Applying data enrichment

For the data transformation, the transaction csv file is the main dataset and I have selected 6 months (07/Jan/2014 – 30 Jun/2014) of transaction data.

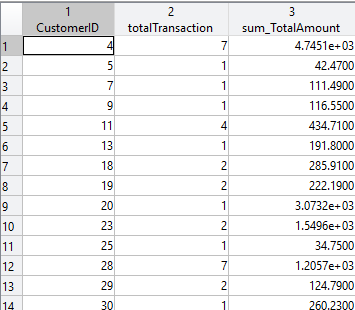
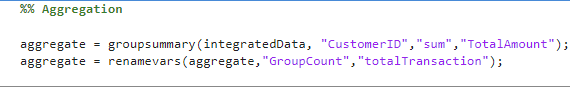
I have integrated 4 cvs files (Transaction, Product, Customer and Staff) to generate one single dataset. It was possible as the Transaction data set references Product, Customer and Staff Ids.

For integration I am using functions such as ‘join’ to join the 4 csv files together based on the references and using ‘removevars’, ‘movevars’ to arrange and remove any redundant columns.

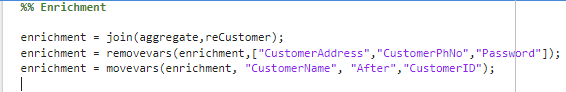


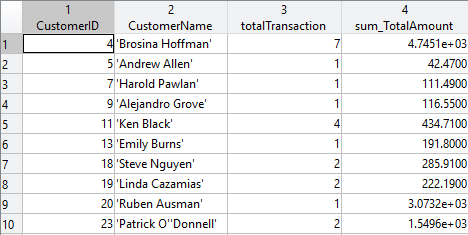


The resulting single dataset from integration was used for aggregating transaction/purchase data for each customer within the selected 6-months data.



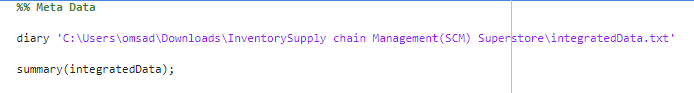
For data enrichment, I added CustomerName to ‘aggregate’ table as it helps to better understand the data and becomes more specific.

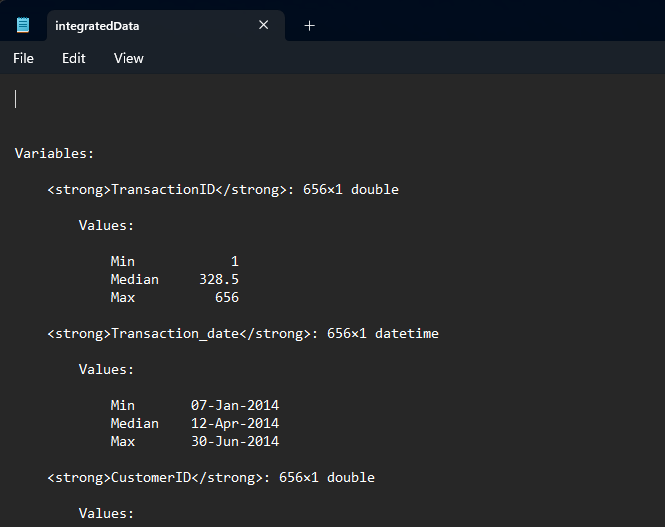




Meta Data:

Since the main file for Transformation was ‘integratedData’. As many variables were already combined into one csv file. I created only one Meta Data file. The file generated various basic attribute properties and statistics by using the function ‘summary’ in Matlab.

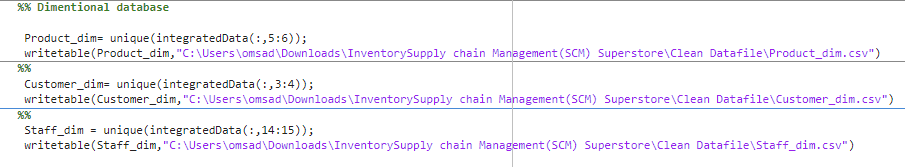


’

Data Loading:

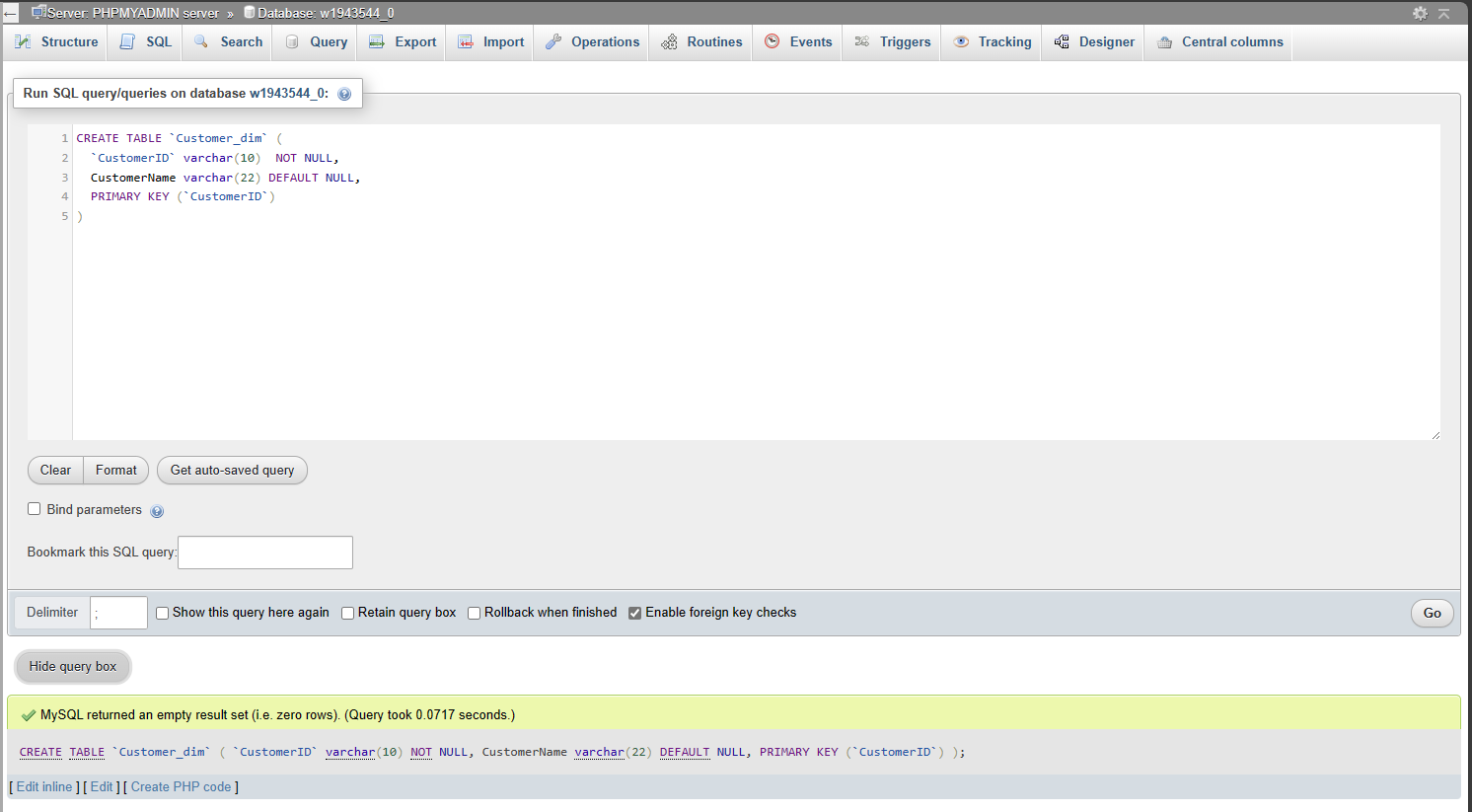
For data loading and storage, I am using MySQL. A database w1943544 is used.

For loading the data from MATLAB to MYSQL, I was not able to setup connection due to installation error. Hence I have created 4 sets of data files ( IntegratedData, Product\_dim, Customer\_dim and Staff\_dim) through matlab.

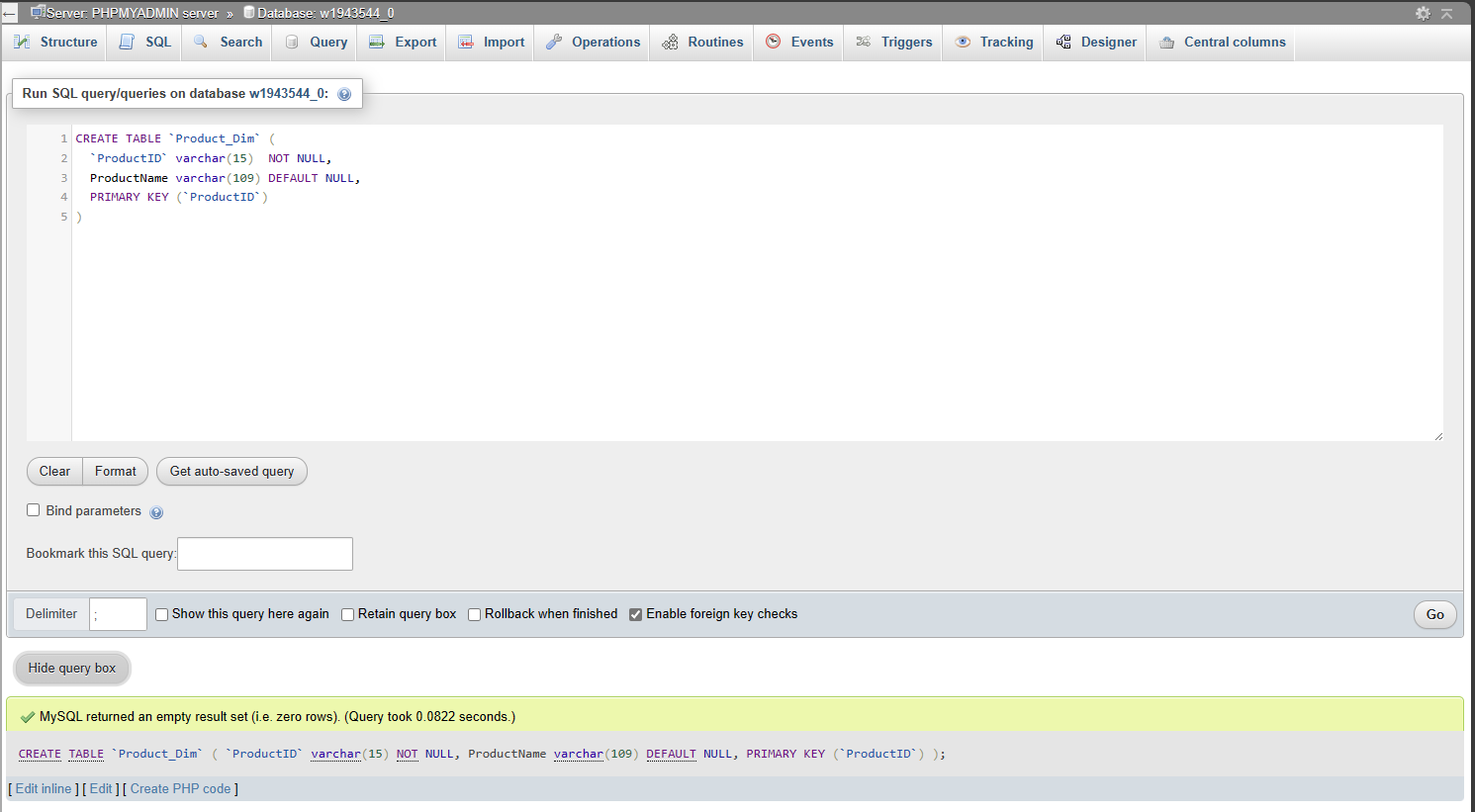


Four tables ( FACT\_Transactions, Product\_Dim, Customer\_Dim and Staff\_Dim ) are created and the data from csv files generated from matlab( IntegratedData, Product\_dim, Customer\_dim and Staff\_dim ) is imported in respective tables.

Create Table for Customer:



Create Table for Product:



Create Table for Staff:

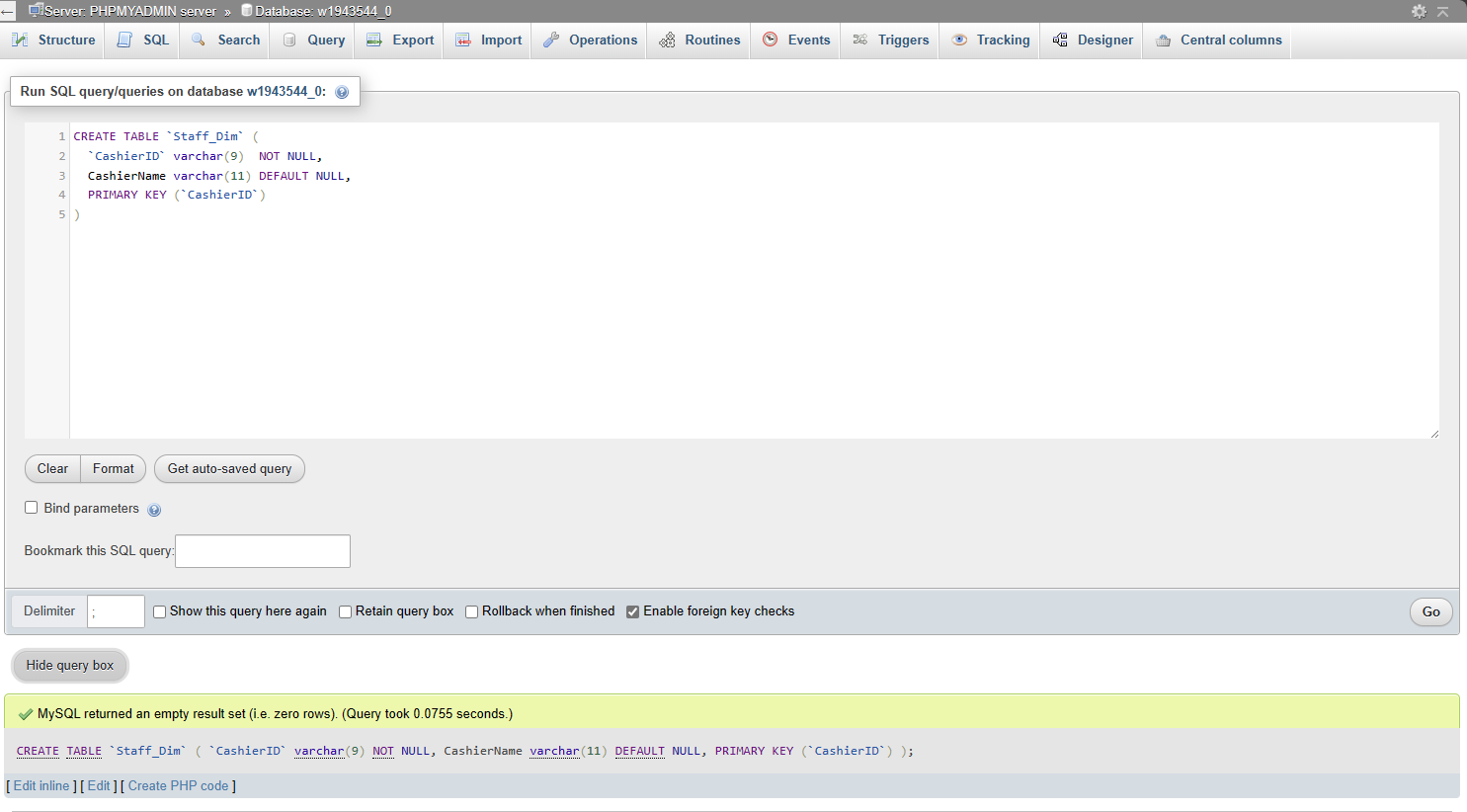
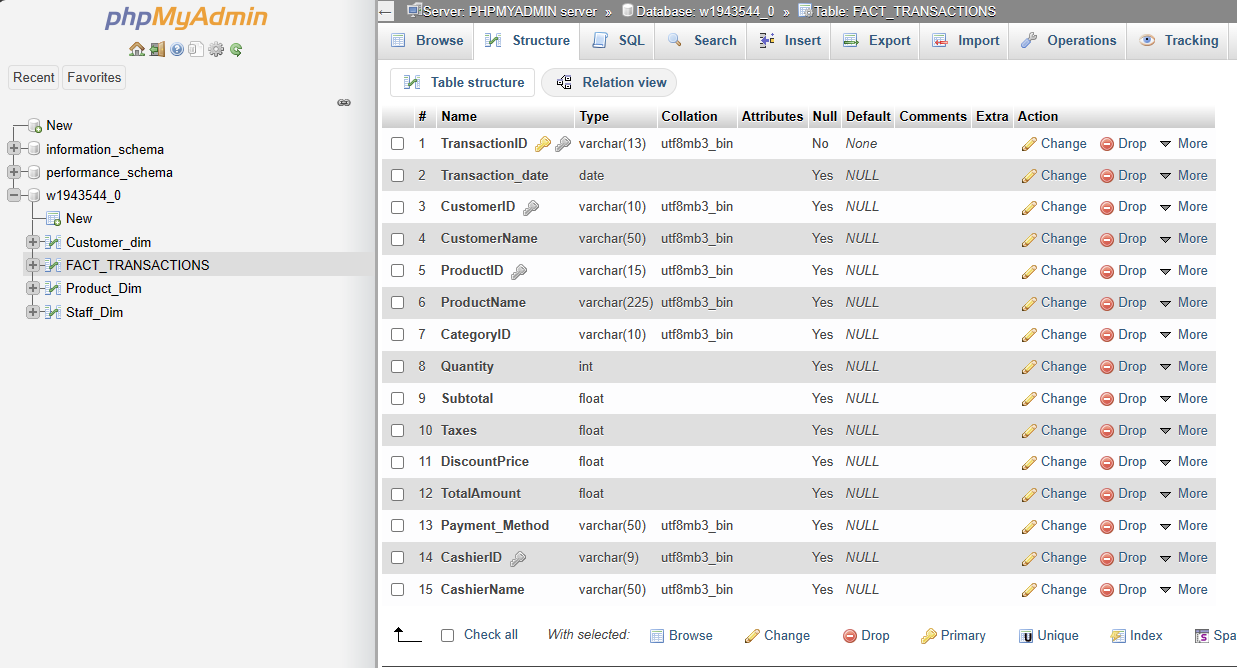
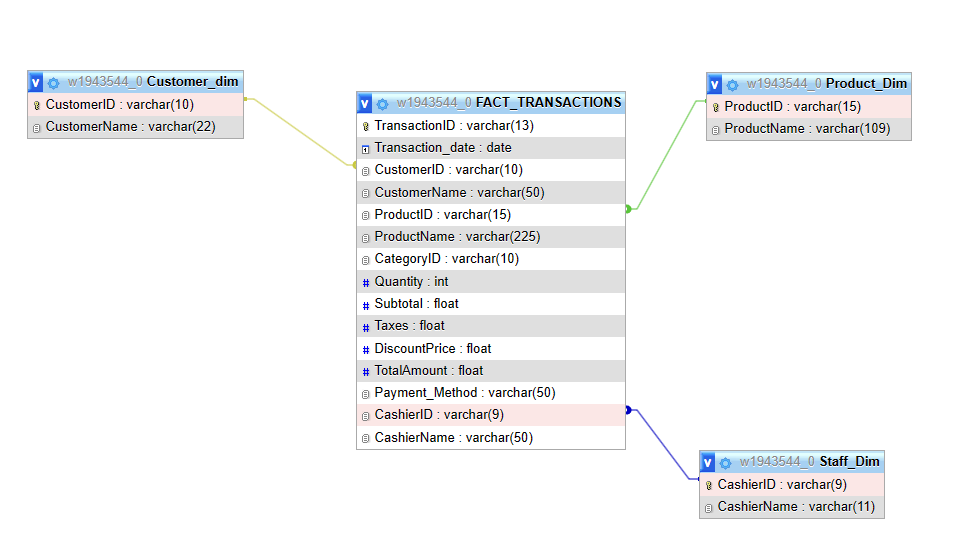


Table for Fact Transaction:



Dimensional Table

I used the sql databases to create the dimensional table.



Self-Reflection:

It was challenging to get a valid dataset with required datapoints and attributes.

The data set which I found had no missing or error value. Due to this I was not able to explore more into methods for data cleaning.

Learning and using a new tool such as Matlab for data transformation was challenging and had to search for functions with correct parameters to get the desired output.

There was Data Toolbox installation issue so I was not able to establish connection between Matlab and MySQL Server. As a work around I exported the files from Matlab and imported in MYSQL Database.