MH6142 Group Project Report

On

Global Super Store Dataset

By:

Leong Yew Wing G1902405D (Part-time)

Li Fengzhi G1901809H (Part-time)

Sim Poh Swee G1111111A (Part-time)

Sim Shu Biao Desmond G1111111A (Full-time)

Table of Contents

[Chapter 1 Introduction 3](#_Toc21902133)

[Chapter 2 Project Setup 4](#_Toc21902134)

[2.1 Dataset Description 4](#_Toc21902135)

[2.2 SQL Server and SQL Developer 4](#_Toc21902136)

[2.3 Visualization with Tableau 4](#_Toc21902137)

[2.4 Python with Jupyter Notebook 4](#_Toc21902138)

[2.5 Google Cloud Platform 5](#_Toc21902139)

[Chapter 3 Loading Data into SQL Server 6](#_Toc21902140)

[3.1 Setting up database server and Creating the database 6](#_Toc21902141)

[3.2 Loading global\_superstore.xlsx into Database 7](#_Toc21902142)

[Chapter 4 SQL Server Data Query and Analytics 10](#_Toc21902143)

[Chapter 5 Data Analytics and Visualization with Tableau 11](#_Toc21902144)

[Chapter 6 Google Cloud Platform with Hive and Spark 12](#_Toc21902145)

# Chapter 1 Introduction

We have applied the basic

# Chapter 2 Project Setup

## 2.1 Dataset Description

For this project, we’ve chosen the Global Super Store dataset because it has over 50000 records of sale transactions with more than 24 different features which provides us sufficient data to carry out data analytic tasks.

## 2.2 SQL Server and SQL Developer

For this project we’ve set up MySql server to host the dataset because it’s freely available Each of us has MySql server installed and configured on our computer, we’ve used both command line interface as well as SQL Developer to interact with the database server. Oracle’s SQL Developer provides a much more user-friendly interaction experience than the command line based interface.

## 2.3 Visualization with Tableau

Tableau is a powerful and fastest growing data visualization software tool used primarily in the business intelligence industry. All of us have learnt to use Tableau in another course of the program. It helps in simplifying raw data into very easily understandable format. Moreover, Tableau is able to connect to various types of data sources including MySql server which is the SQL Server that we are using for the project.

## 2.4 Python with Jupyter Notebook

There is no doubt that Python has become the most popular programming language in the data analytics community. There are various Python libraries available to connect to and execute SQL queries on MySql server. In this project, we’ve used Jupyter Notebook to develop our Python codes.

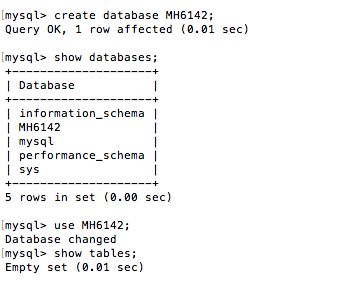
## 2.5 Google Cloud Platform

Besides MySql Server, we have also imported the Global Super Store dataset into Google Cloud Platform, we then explored similar SQL queries with Hive and tried various data exploration techniques with PySpark.

# Chapter 3 Loading Data into SQL Server

## 3.1 Setting up database server and Creating the database

After installation of MySql Server, from the command line, we’ve created MH6142 database.



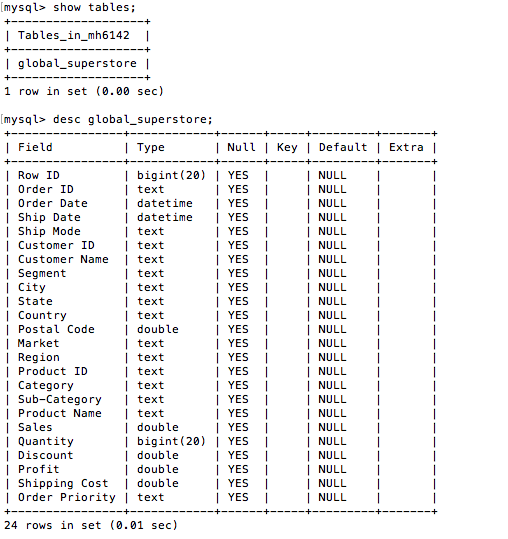
After the database MH6142 is created, it’s an empty database with no data.

## 3.2 Loading global\_superstore.xlsx into Database

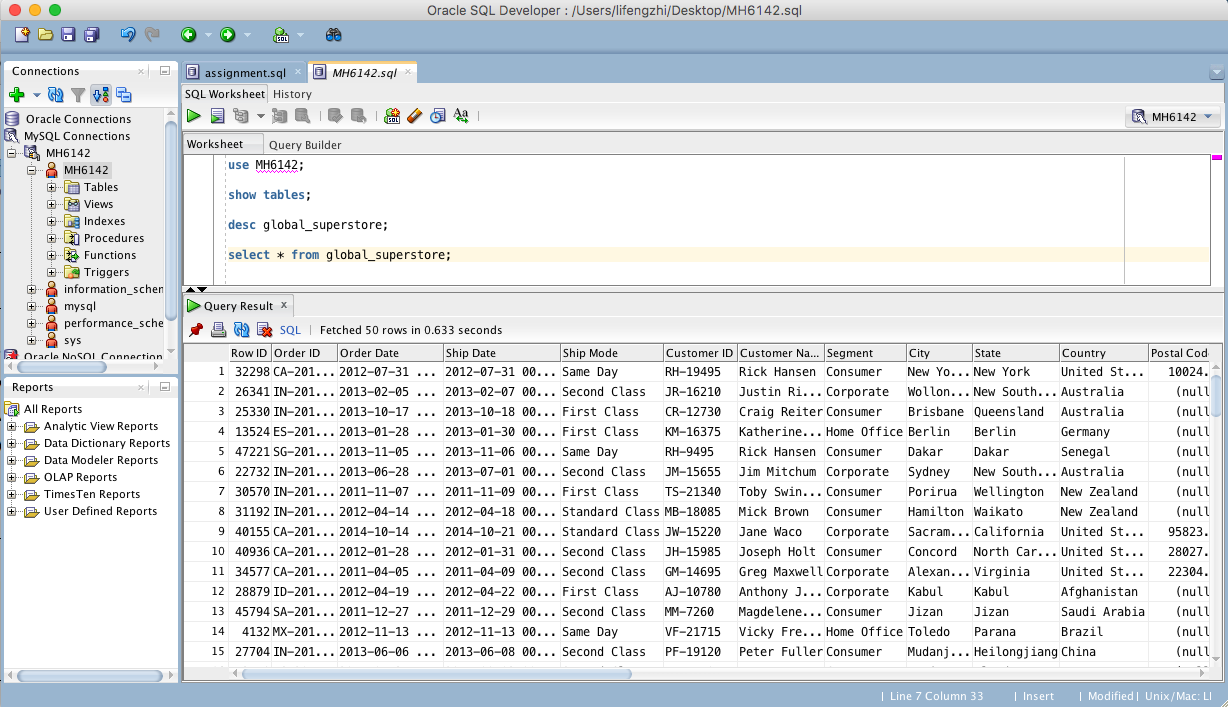
We’ve set up the connection to the MySql Server’s MH6142 database created earlier, and loaded the global\_superstore.csv into the database,



We then verified the “global\_superstore” table from MySql command line interface, and we can see that the table has been created successfully,

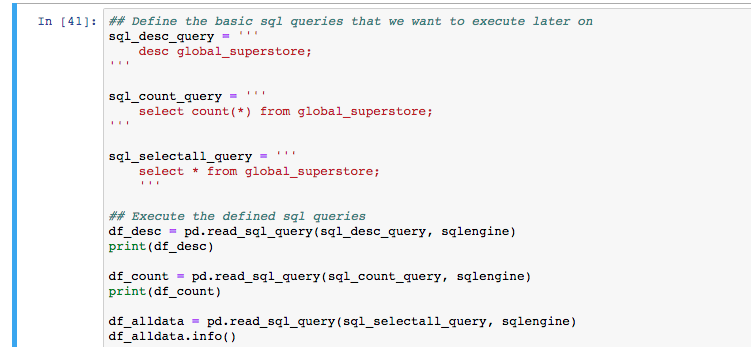


We also verified through SQL Developer interface,



# Chapter 4 SQL Server Data Query and Analytics

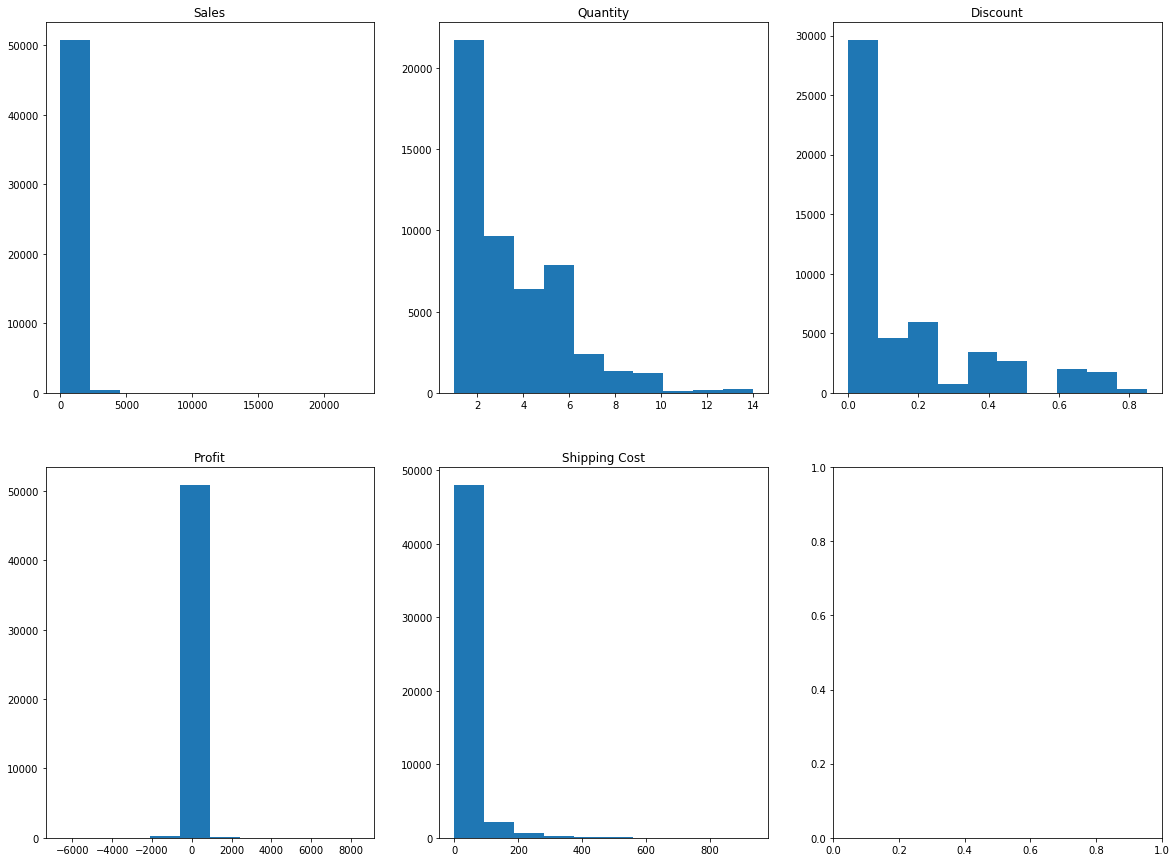
In this section, we used Python with Jupyter Notebook to explore the different patterns and trends in order to gain some insights into the transaction data, we can explore and load the data with the following simple SQL queries,



Note that the Python Source Codes Jupyter Notebook is attached separately, hence they are not repeated entirely in this section.

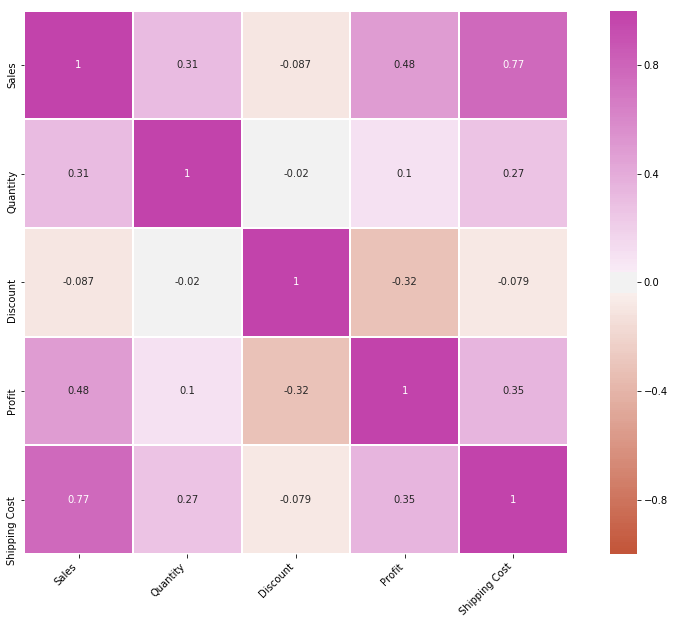
## 4.1 Value distribution of Numerical features

From the figures below, we can easily observe that profit can be either positive or negative, while most discount is below 10%.



## 4.2 Correlations (Profit vs Discount, Profit vs Sales)

We can see that Profit has a negative correlation with Discount which makes sense. In addition, Profit has a positive correlation with Sales.



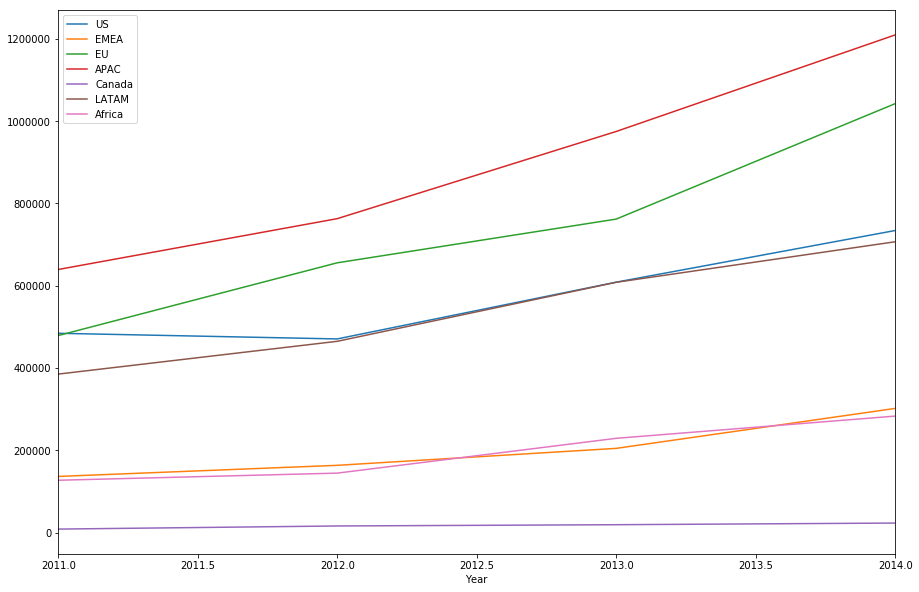
## 4.2 Total Sales Among Market, Category and Sub-Category

Next we studied the aggregated Sales based on Category, Market and Sub-Category features,

|  |  |
| --- | --- |
| /var/folders/mm/wp35_1_j0_g8pksrj2ql92s40000gn/T/com.microsoft.Word/Content.MSO/BD55199C.tmp | /var/folders/mm/wp35_1_j0_g8pksrj2ql92s40000gn/T/com.microsoft.Word/Content.MSO/309B5E0A.tmp |
| /var/folders/mm/wp35_1_j0_g8pksrj2ql92s40000gn/T/com.microsoft.Word/Content.MSO/64A612A8.tmp | |

## 4.2 Total Sales Among Market, Category and Sub-Category

We also explored the Sales growth tread between 2011 and 2014 among the 7 major markets, we can see that US market’s Sales dipped around 2011 year end. APAC market have achieved tremendous growth.



# Chapter 5 Data Analytics and Visualization with Tableau

# Chapter 6 Google Cloud Platform with Hive and Spark