**Executive Report**

By: Jaideep Prasad

Summary:

This is a brief section responsible for summarizing the report. The detailed report can be found on page 2.

The purpose of this report is to see which products and locations sell the most. The four main issues that challenged us with the data are the formatting of product names, understanding the relationship between customers and locations, the issue with product ID and using VLOOKUP in excel to make sure the foreign keys have the right values when deleting duplicates in the tables.

Would suggest implementing the following:

* + - Data quality policies:
      * + More focus on metadata
        + Software to standardize records of transactions
    - Data privacy policies:
      * + Encode sensitive data to everyone other than departments that needed unencoded data such as IT, marketing, analytics/business departments
        + Follow PIPEDA standards

Data Management Framework: DMBOK Pyramid

* + - Stage 1: Look at data and understand what each feature represents
    - Stage 2: Isolate relevant features and extract any useful data/metadata
    - Stage 3: Make assumption, data policies and clear documentation
    - Stage 4: Extract conclusions for our sales questions listed above

General list of content in appendix:

* + - ERD
    - Data Flow Diagram
    - MySQL script
    - Tabular data

Conclusion

* + - Technology is the bulk of out profits while furniture such as tables and bookcases are the bulk of loss in profits
    - Stop selling technology in Thomasville, Pomona, Springfield and Tuscaloosa. Stop selling office supplies in Miami at the rate we do as these are the top 5 losses in profits. Location either doesn’t affect sales or we don’t have enough data to form a reliable conclusion

Report:

The purpose of this report is to look at the products that are selling as well as pre-existing store locations to figure out what type of products sell the most and where we sell the most. By looking at the categories that sell the most/least, we will have a good idea on what to stock up on and what products we have too much of. By analysing which store locations are doing good, we have an idea of where to expand our business locations should the need arise.

As we are a big company with many departments, it is a priority to focus on data privacy and data quality. As such, I would like to suggest a few changes which would benefit the corporation. Data quality should introduce two need features, metadata, and standardized forms for data entry. Metadata will allow us to gain more insights which can accelerate our abilities to capitalize on trends. Though we only record receipts and transactions, information such as gender, race, store proximity to other stores and store location size can help shape the decisions of our marketing, analytics, shipping, and other departments. These insights and new data will not help us unless it is organized in its early stages, hence why we should invest in a software that gives clerks and employees more of a drop-down menu style of recording transactions rather than letting employees write whatever they seem fit. Something as simple as gender can be written in several ways, not to mention the potential typos. The software will prevent any type of miscommunication when we analyse the data later on.

We have had a lot of success with the superstore however that means now we have more eyes on us. Due to this we need to step up the quality of our data privacy policies as any potential incidents can cost us millions of dollars and more importantly, brand loyalty. For starters, we should encode all personal data such that departments who do not need access to the data will not have access to the data. People like store managers will only be able to see basic transaction information while departments who need access to the data (such as IT, marketing, analytics/business departments) will have access to unencoded data. Our higher level of data governance will also have access to the unencoded data and can pursue any inquiries employees have, should the need for unencoded data access arise. Apart from this, Canada has a standard data privacy framework called PIPEDA which comes with a lot of resources for understanding and planning on data privacy subject matters. We can implement this to prevent potential incidents and give our customers a piece of mind.

For this report, I used the DMBOK Pyramid as my data management framework. For the first stage, I requested access to the unencoded data from the IT department. Having access allowed me to see what data we have and what the data represented. After gaining a good idea of what the data represented, I moved on to the second stage. I started populating the new excel files with relevant features/data along with its relevant metadata. Once populated, I made sure that none of the primary keys were duplicated. With the data ready to go, stage three required me to document my process/reasoning as well as think about any suggestions that could benefit our data governance policies. After some research and the above-mentioned data policy changes, I started up MySQL, created a new schema, imported my tables using the table wizard, edited the tables to create the required primary and foreign keys, reverse engineered the schema to create the ERD, verified the ERD (as the first few attempts yielded some errors) and wrote my queries.

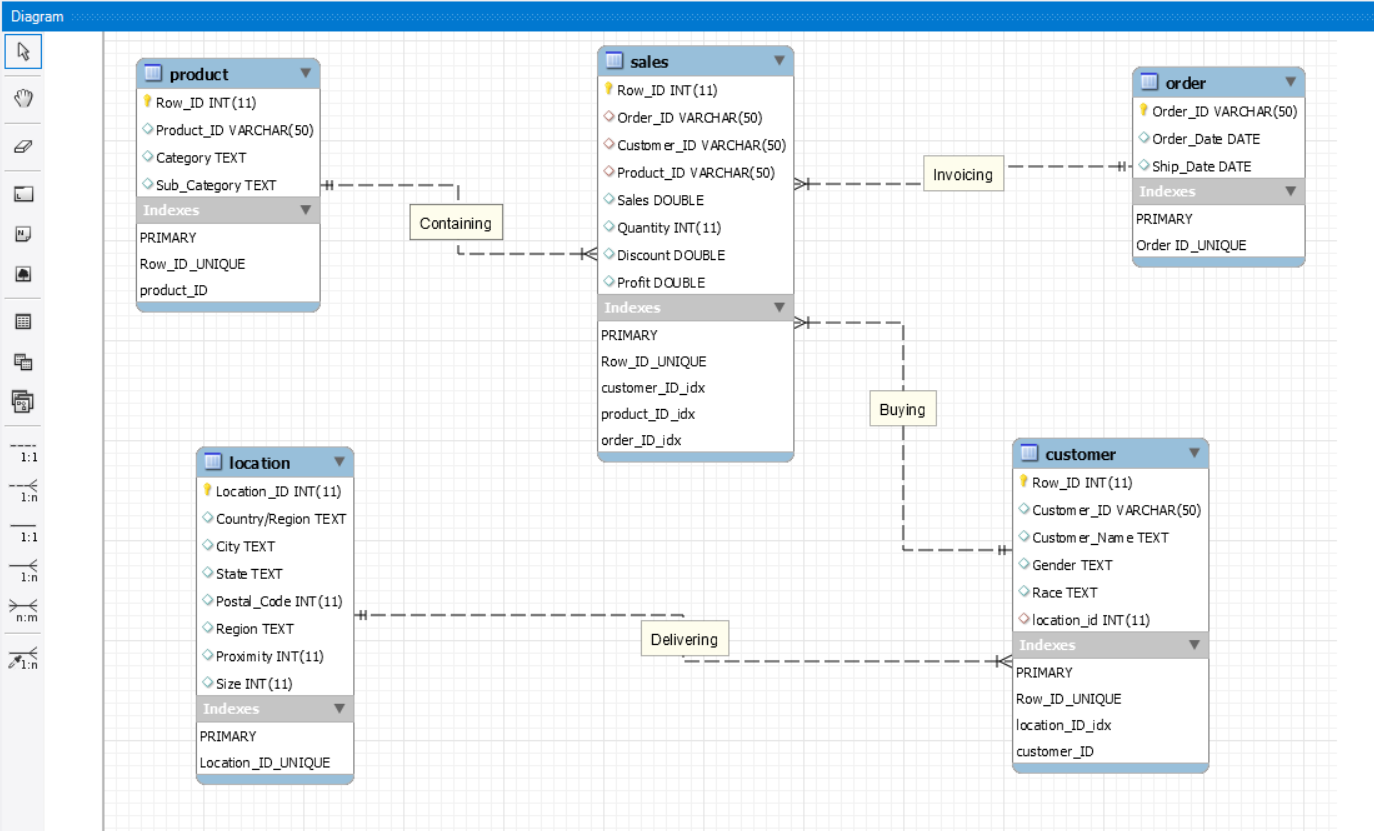
While working with the data, there were some adjustments made from the original idea of this report due to my skill level and the data given. For my product table, I left out the product name as importing the product name into MySQL caused issues. A great deal of information would not get imported due to the formatting of the product names, dropping the imported records from 1894 to 95. It is also important to mention that the product ID “FUR-BO-10002213” refers to multiple product names, “DMI Eclipse Executive Suite Bookcases” and “Sauder Forest Hills Library, Woodland Oak Finish”. We created a new feature called Row ID and instead of looking at the profits of products, we now looked at profits of categories and sub-categories. Customers had multiple locations attached to them, this changed our initial understanding of the location table from representing a customers’ address to a customers’ delivery location. The last issue was learning how to using VLOOKUP in excel to make sure the foreign keys have the right values when deleting duplicates in the tables. While the product ID is a simple fix (give one of the products a unique product ID), the product name issue is an interesting one for me. I’m not sure what the formatting error was as I don’t have too much knowledge on SQL and the parsing for csv files. With a little time however, I am sure I myself or our IT department can figure out a standard format for product names that will allow us to use this data column later on. In the appendix below, you will be able to see the data tables used for the database.

When looking at the output for the sales query (SQL\_Sales\_Output.csv), there is a clear distinction between products that are doing good and products that are doing bad. In general, it looks like technology products do good especially phones, copiers, and tech accessories. Office supplies tend to never make us loss profits and furniture is sort of all over the place. We notice that chairs and furnishings net us a decent profit while tables and bookcases net us a negative profit. Looking at locations (SQL\_Locations\_Output.csv), the data seems to be erratic. There seems to be no real pattern as neither city, state or region tend to repeat in a certain order.

Based on the provided research we can come up with two conclusions. We can safely say that technology is the bulk of out profits while furniture such as tables and bookcases are the bulk of loss in profits. As if right now however, either we do not have enough data to find a relationship that would let us know which locations are increasing and decreasing our profits or, location does not affect our profits. We should stop selling technology in Thomasville, Pomona, Springfield, and Tuscaloosa and stop selling office supplies in Miami at the rate we do as these are the top 5 losses in profits.

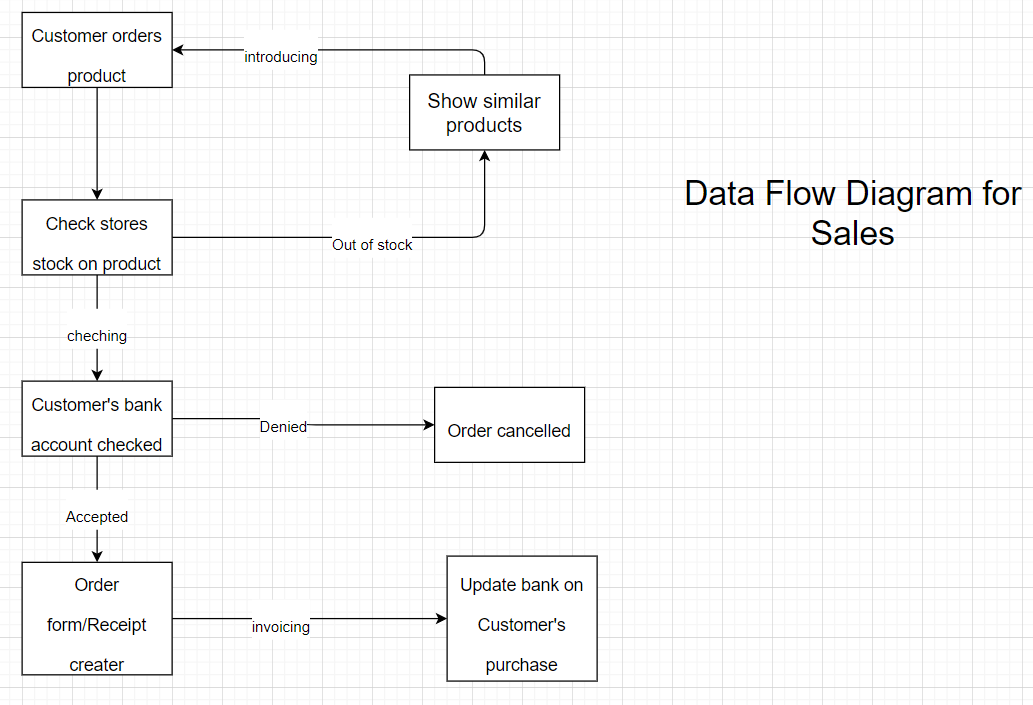
Appendix:

ERD:



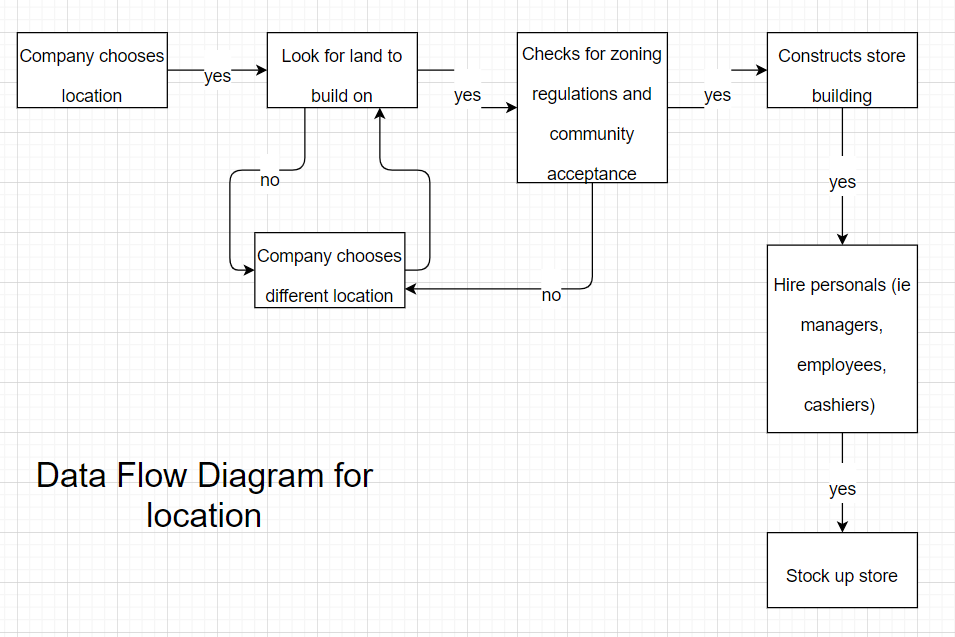
*ERD 1: Superstore ERD*

Data Flow Diagram 1:



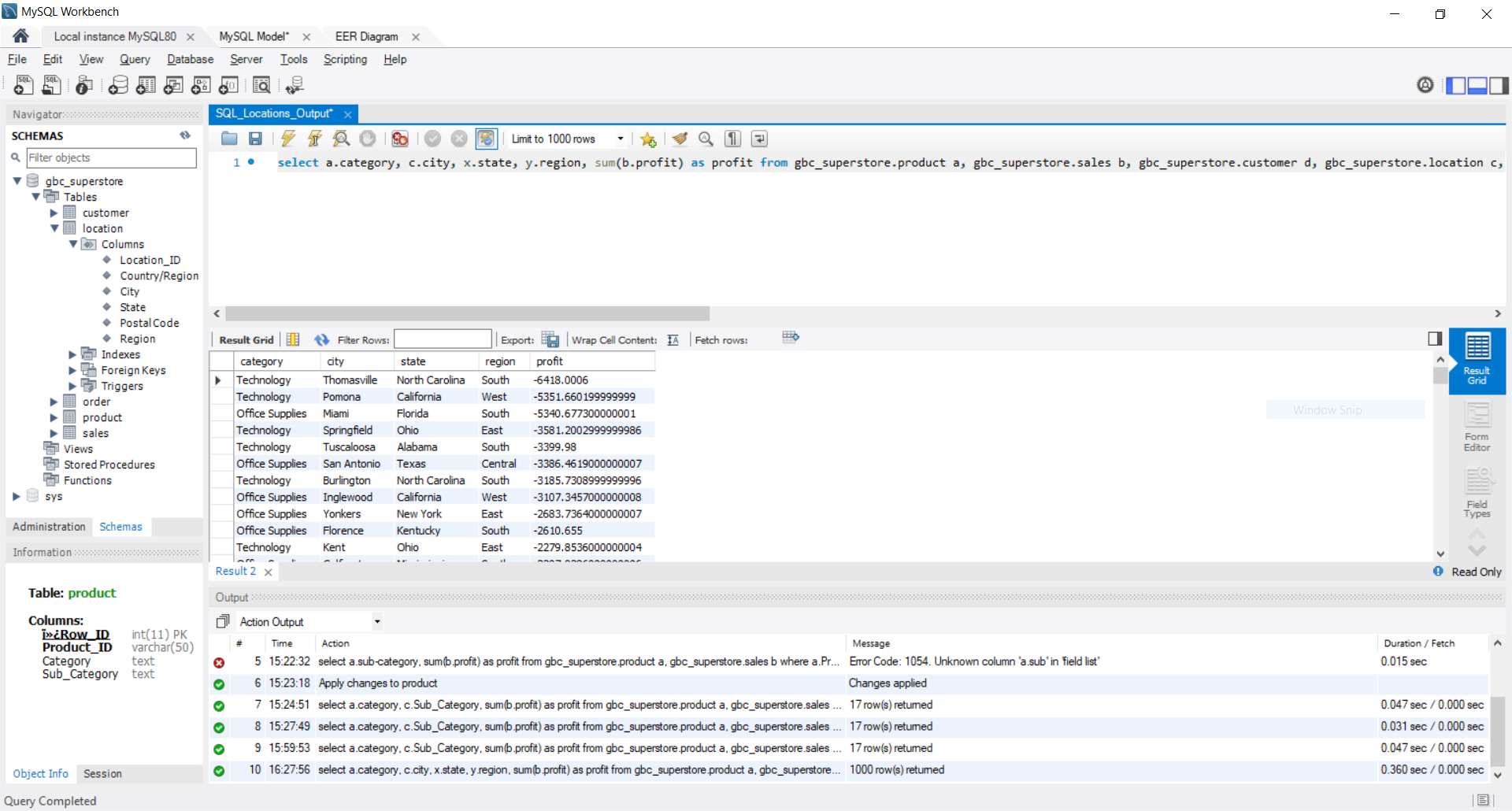
*Data Flow Diagram 1: Superstore Sales data flow diagram*

Data Flow Diagram 2:



*Data Flow Diagram 2: Superstore choosing new location data flow diagram*

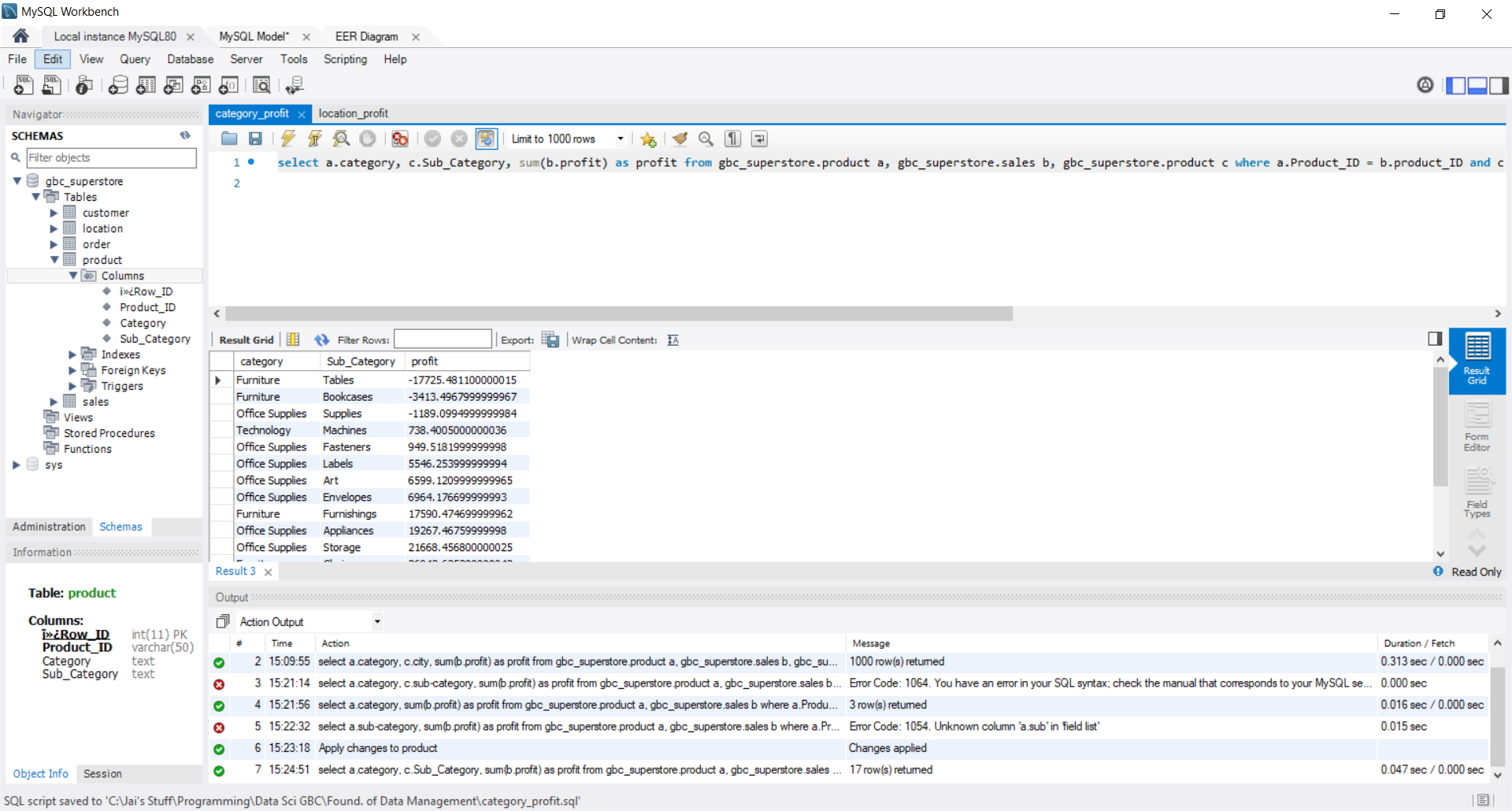
MySQL Script 1:

 *MySQL Script 1: Table shows use all the cities, their profits and the category that sold the most*

*The output table can be found listed with the tables in this report named “SQL\_Locations\_Output”*

*Code: select a.category, c.city, x.state, y.region, sum(b.profit) as profit from gbc\_superstore.product a, gbc\_superstore.sales b, gbc\_superstore.customer d, gbc\_superstore.location c, gbc\_superstore.location x, gbc\_superstore.location y where a.Product\_ID = b.product\_ID and b.customer\_ID = d.customer\_ID and d.location\_ID = c.location\_ID and d.location\_ID = x.location\_ID and d.location\_ID = y.location\_ID group by a.category, c.city, x.State, y.Region order by profit;*

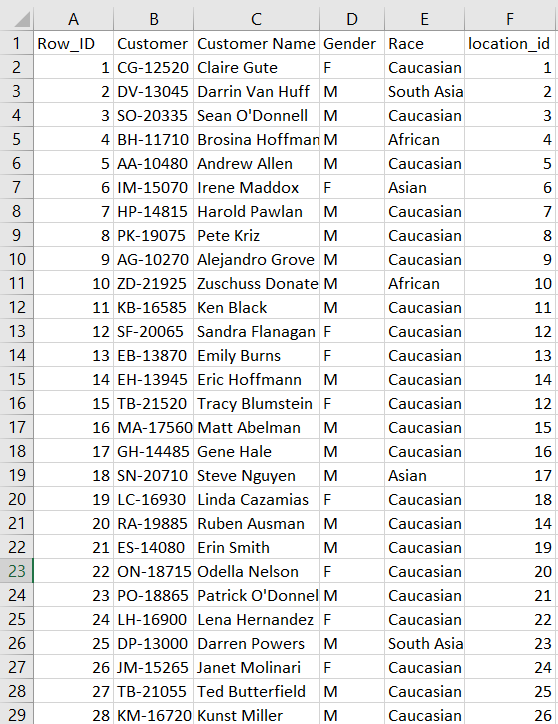
MySQL Script 2:

 *MySQL Script 2: Table shows use all sub-categories with their respective categories and profit*

*The output table can be found listed with the tables in this report named “SQL\_Sales\_Output”*

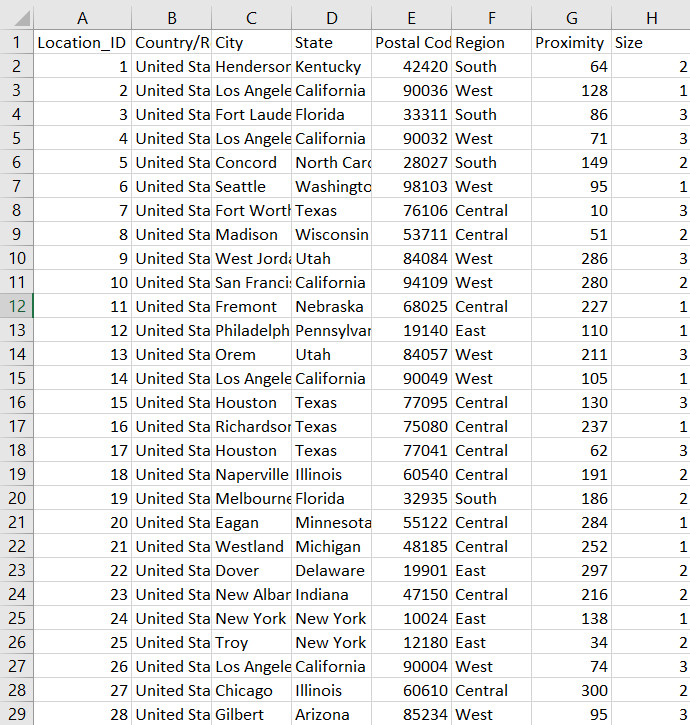
*Code: select a.category, c.Sub\_Category, sum(b.profit) as profit from gbc\_superstore.product a, gbc\_superstore.sales b, gbc\_superstore.product c where a.Product\_ID = b.product\_ID and c.Product\_ID = b.product\_ID group by a.category, c.Sub\_Category order by profit;*

Table 1:



*Table 1: Customer data from file “customer\_v1.csv” which includes the metadata features gender and race*

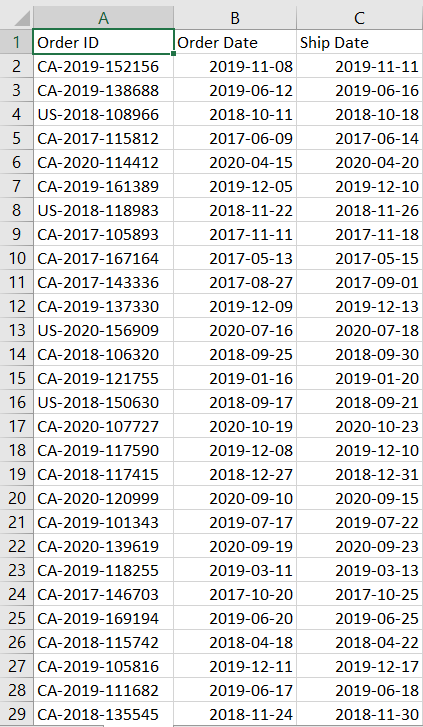
Table 2:



*Table 2: Location data from file “location.csv” which includes the metadata features proximity and size*

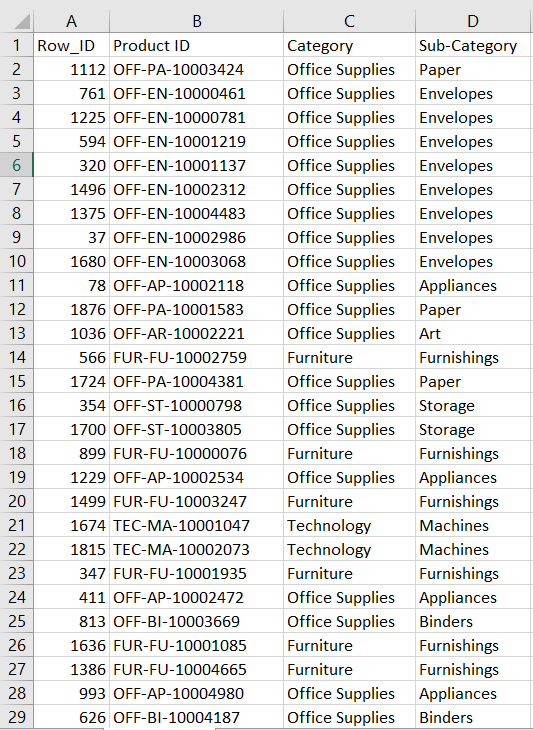
*Proximity represents the proximity of our Superstore in relations to any other general store in kilometer. Size represents the size of our superstore where 1 is small and 3 is large.*

Table 3:



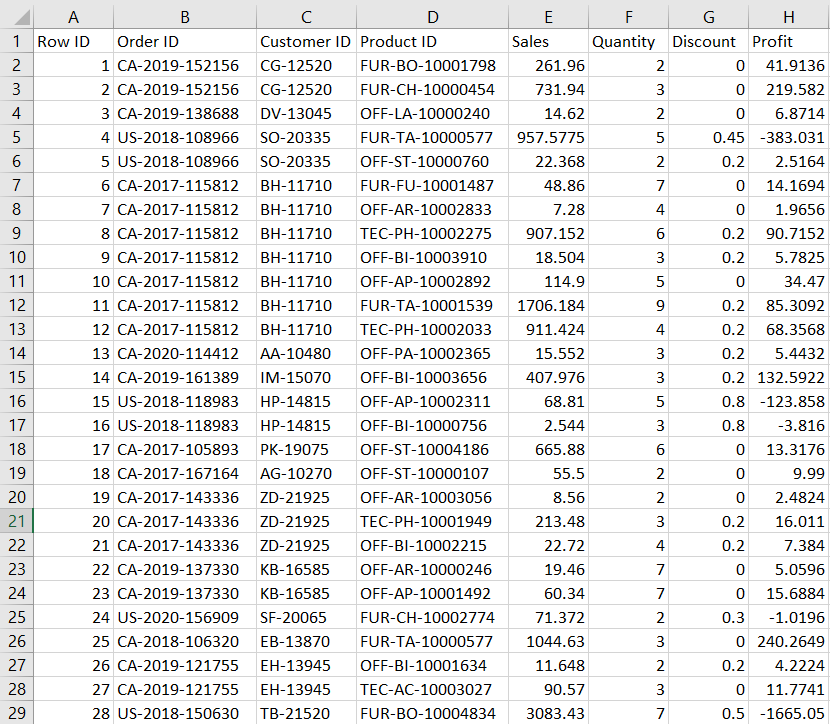
*Table 3: Orders data from file “Order.csv”*

Table 4:



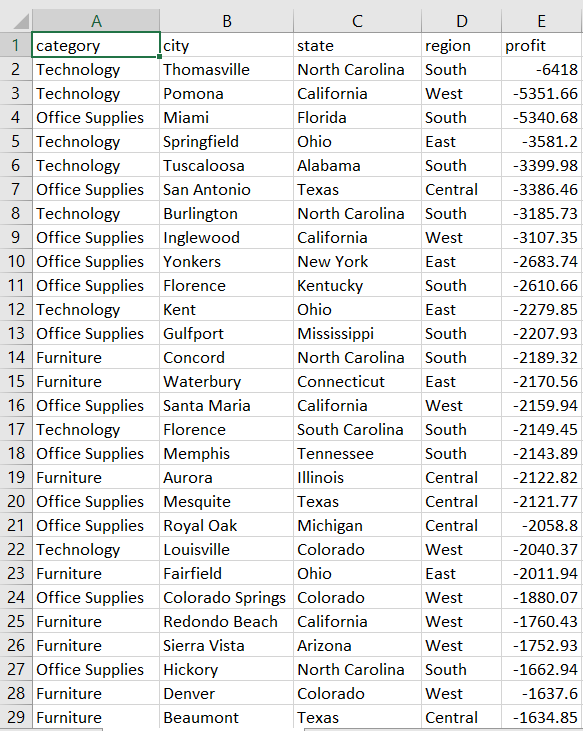
*Table 4: Product data from file “Products\_v3.csv”*

Table 5:



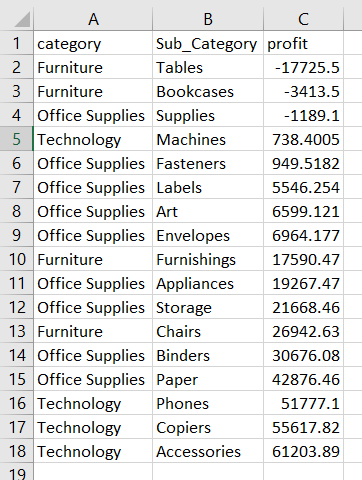
*Table 5: Sales data from file “Sales\_v1.csv”*

Table 6:



*Table 6: MySQL query output table from file “SQL\_Locations\_Output.csv”*

Table 7:



*Table 7: MySQL query output table from file “SQL\_Sales\_Output.csv”*