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.

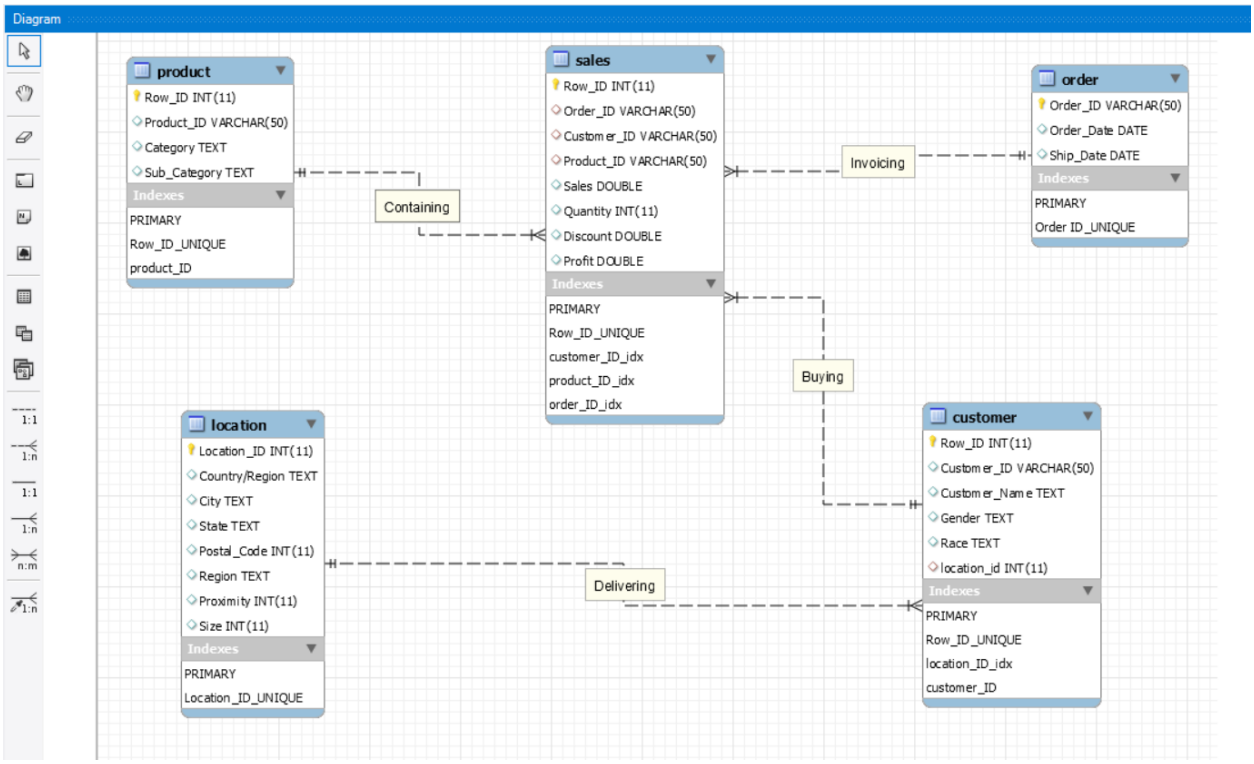
While working with the data, there were some adjustments made from the original idea of this report due to my skill level and the data given. For my product table, I left out the product name as importing the product name into MySQL caused issues. A great deal of information would not get imported due 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.

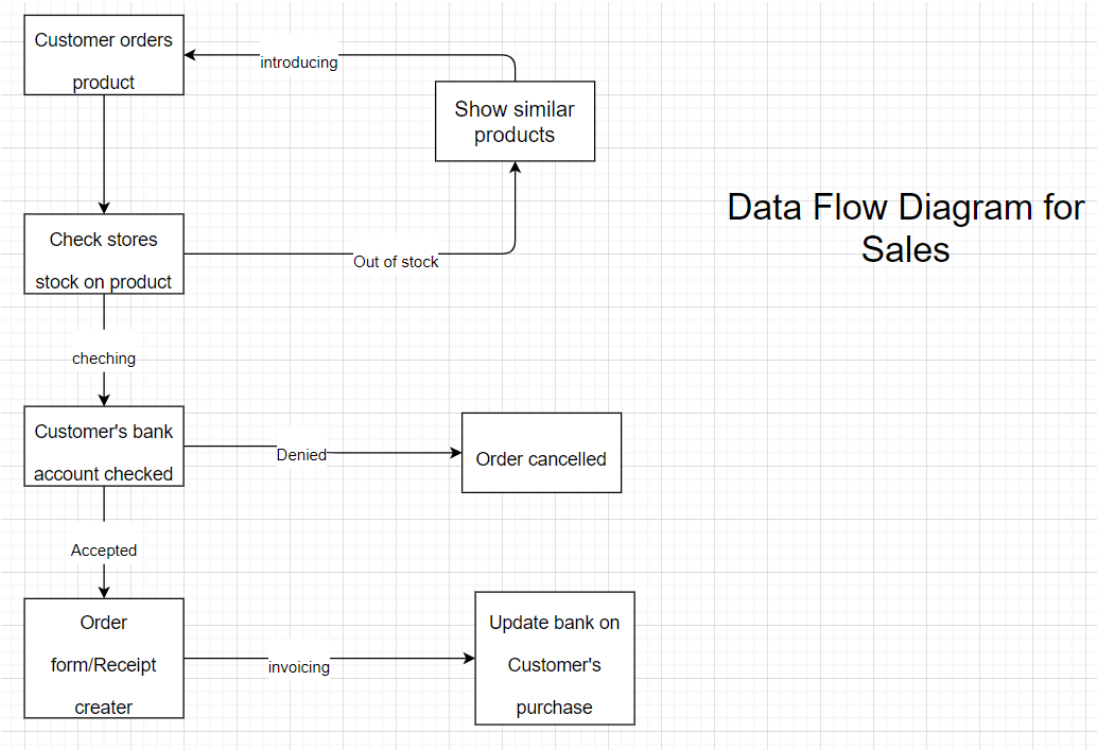
Appendix:

ERD:



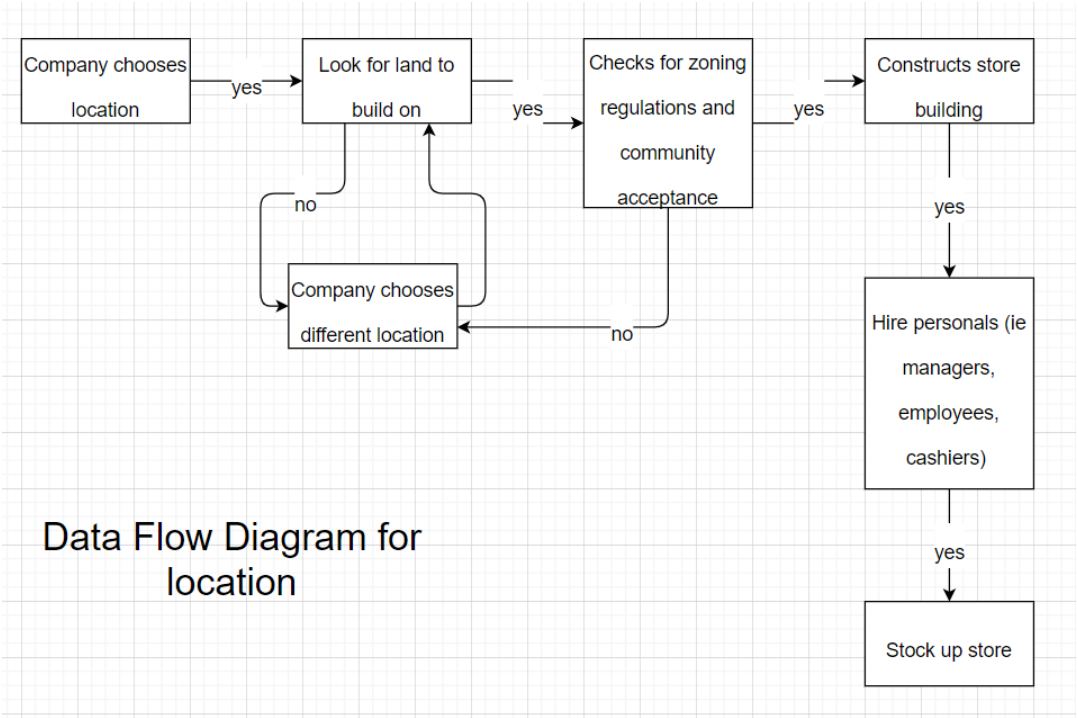
ERD 1: Superstore ERD

Data Flow Diagram 1:



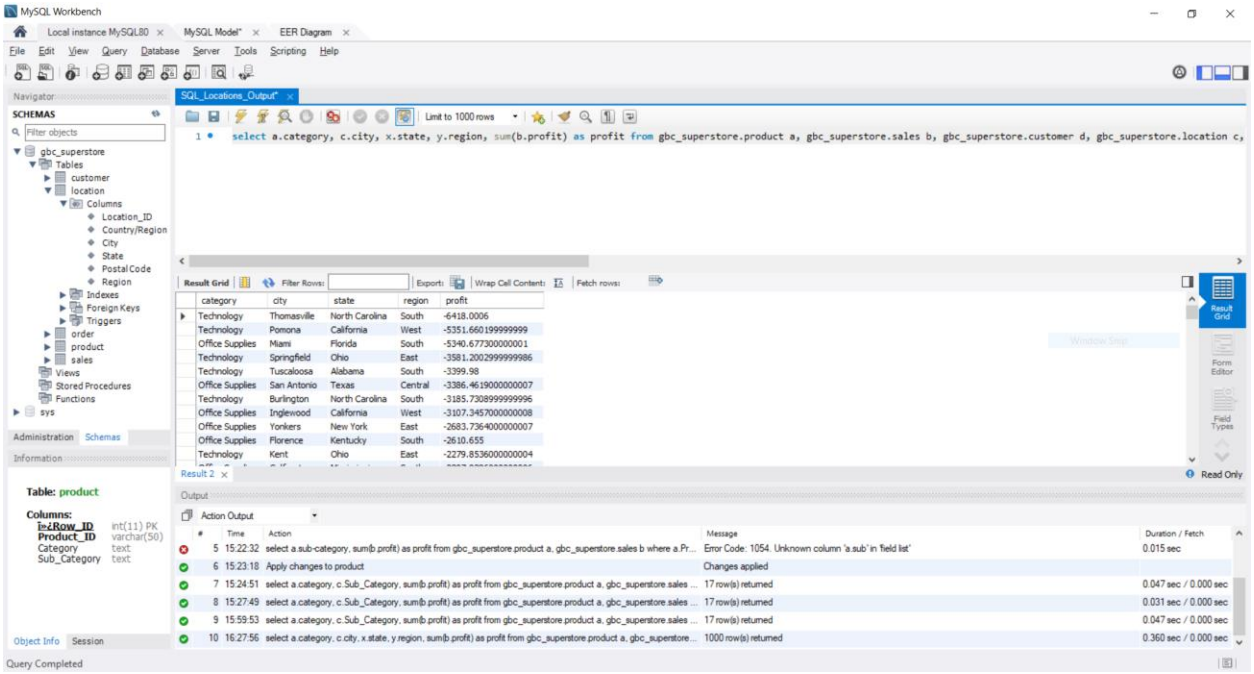
Data Flow Diagram 1: Superstore Sales data flow diagram

Data Flow Diagram 2:



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

MySQL Script 1:

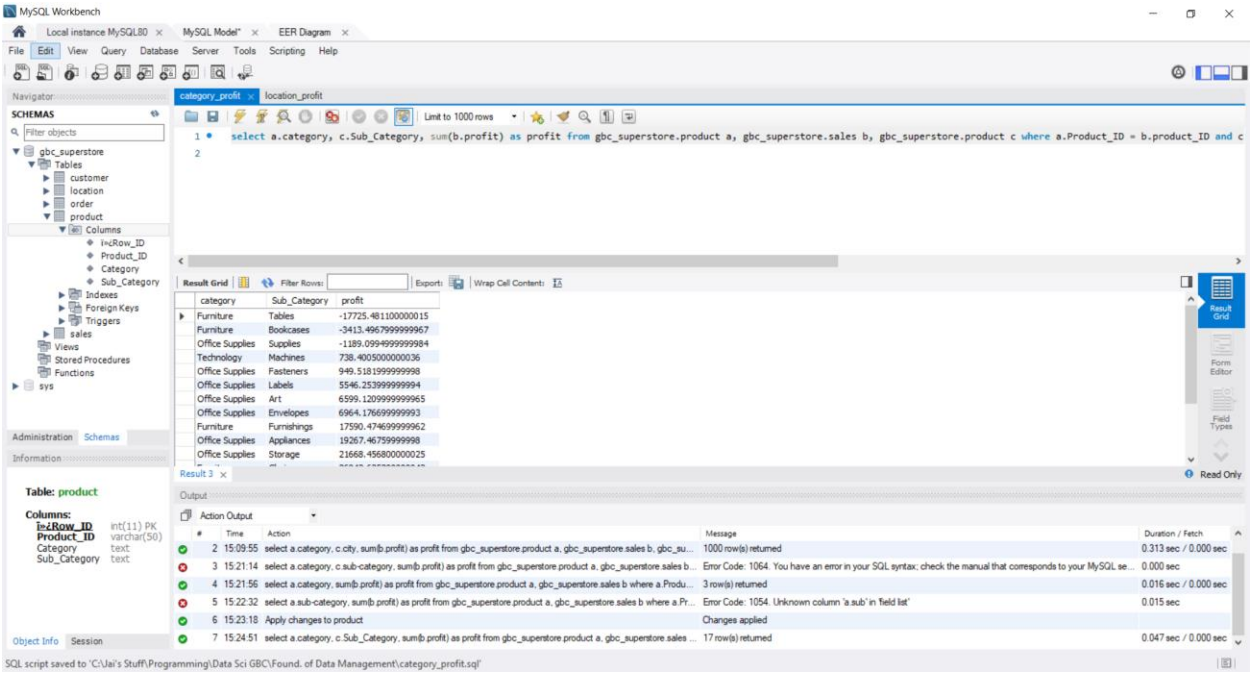


MySQL Script 1: Table shows use all the cities, their profits and the category that sold the most

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:

	A	B	C	D	E	F
1	Row_ID	Customer	Customer Name	Gender	Race	location_id
2	1	CG-12520	Claire Gute	F	Caucasian	1
3	2	DV-13045	Darrin Van Huff	M	South Asia	2
4	3	SO-20335	Sean O'Donnell	M	Caucasian	3
5	4	BH-11710	Brosina Hoffman	M	African	4
6	5	AA-10480	Andrew Allen	M	Caucasian	5
7	6	IM-15070	Irene Maddox	F	Asian	6
8	7	HP-14815	Harold Pawlan	M	Caucasian	7
9	8	PK-19075	Pete Kriz	M	Caucasian	8
10	9	AG-10270	Alejandro Grove	M	Caucasian	9
11	10	ZD-21925	Zuschuss Donate	M	African	10
12	11	KB-16585	Ken Black	M	Caucasian	11
13	12	SF-20065	Sandra Flanagan	F	Caucasian	12
14	13	EB-13870	Emily Burns	F	Caucasian	13
15	14	EH-13945	Eric Hoffmann	M	Caucasian	14
16	15	TB-21520	Tracy Blumstein	F	Caucasian	12
17	16	MA-17560	Matt Abelman	M	Caucasian	15
18	17	GH-14485	Gene Hale	M	Caucasian	16
19	18	SN-20710	Steve Nguyen	M	Asian	17
20	19	LC-16930	Linda Cazamias	F	Caucasian	18
21	20	RA-19885	Ruben Ausman	M	Caucasian	14
22	21	ES-14080	Erin Smith	M	Caucasian	19
23	22	ON-18715	Odella Nelson	F	Caucasian	20
24	23	PO-18865	Patrick O'Donnel	M	Caucasian	21
25	24	LH-16900	Lena Hernandez	F	Caucasian	22
26	25	DP-13000	Darren Powers	M	South Asia	23
27	26	JM-15265	Janet Molinari	F	Caucasian	24
28	27	TB-21055	Ted Butterfield	M	Caucasian	25
29	28	KM-16720	Kunst Miller	M	Caucasian	26

Table 1: Customer data from file "customer_v1.csv" which includes the metadata features gender and race

Table 2:

	A	B	C	D	E	F	G	H
1	Location_ID	Country/R	City	State	Postal Cod	Region	Proximity	Size
2	1	United Sta	Hendersor	Kentucky	42420	South	64	2
3	2	United Sta	Los Angele	California	90036	West	128	1
4	3	United Sta	Fort Laude	Florida	33311	South	86	3
5	4	United Sta	Los Angele	California	90032	West	71	3
6	5	United Sta	Concord	North Car	28027	South	149	2
7	6	United Sta	Seattle	Washingtc	98103	West	95	1
8	7	United Sta	Fort Worth	Texas	76106	Central	10	3
9	8	United Sta	Madison	Wisconsin	53711	Central	51	2
10	9	United Sta	West Jorda	Utah	84084	West	286	3
11	10	United Sta	San Franci	California	94109	West	280	2
12	11	United Sta	Fremont	Nebraska	68025	Central	227	1
13	12	United Sta	Philadelph	Pennsylvar	19140	East	110	1
14	13	United Sta	Orem	Utah	84057	West	211	3
15	14	United Sta	Los Angele	California	90049	West	105	1
16	15	United Sta	Houston	Texas	77095	Central	130	3
17	16	United Sta	Richardsor	Texas	75080	Central	237	1
18	17	United Sta	Houston	Texas	77041	Central	62	3
19	18	United Sta	Naperville	Illinois	60540	Central	191	2
20	19	United Sta	Melbourne	Florida	32935	South	186	2
21	20	United Sta	Eagan	Minnesota	55122	Central	284	1
22	21	United Sta	Westland	Michigan	48185	Central	252	1
23	22	United Sta	Dover	Delaware	19901	East	297	2
24	23	United Sta	New Albar	Indiana	47150	Central	216	2
25	24	United Sta	New York	New York	10024	East	138	1
26	25	United Sta	Troy	New York	12180	East	34	2
27	26	United Sta	Los Angele	California	90004	West	74	3
28	27	United Sta	Chicago	Illinois	60610	Central	300	2
29	28	United Sta	Gilbert	Arizona	85234	West	95	3

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:

	A	B	C
1	Order ID	Order Date	Ship Date
2	CA-2019-152156	2019-11-08	2019-11-11
3	CA-2019-138688	2019-06-12	2019-06-16
4	US-2018-108966	2018-10-11	2018-10-18
5	CA-2017-115812	2017-06-09	2017-06-14
6	CA-2020-114412	2020-04-15	2020-04-20
7	CA-2019-161389	2019-12-05	2019-12-10
8	US-2018-118983	2018-11-22	2018-11-26
9	CA-2017-105893	2017-11-11	2017-11-18
10	CA-2017-167164	2017-05-13	2017-05-15
11	CA-2017-143336	2017-08-27	2017-09-01
12	CA-2019-137330	2019-12-09	2019-12-13
13	US-2020-156909	2020-07-16	2020-07-18
14	CA-2018-106320	2018-09-25	2018-09-30
15	CA-2019-121755	2019-01-16	2019-01-20
16	US-2018-150630	2018-09-17	2018-09-21
17	CA-2020-107727	2020-10-19	2020-10-23
18	CA-2019-117590	2019-12-08	2019-12-10
19	CA-2018-117415	2018-12-27	2018-12-31
20	CA-2020-120999	2020-09-10	2020-09-15
21	CA-2019-101343	2019-07-17	2019-07-22
22	CA-2020-139619	2020-09-19	2020-09-23
23	CA-2019-118255	2019-03-11	2019-03-13
24	CA-2017-146703	2017-10-20	2017-10-25
25	CA-2019-169194	2019-06-20	2019-06-25
26	CA-2018-115742	2018-04-18	2018-04-22
27	CA-2019-105816	2019-12-11	2019-12-17
28	CA-2019-111682	2019-06-17	2019-06-18
29	CA-2018-135545	2018-11-24	2018-11-30

Table 3: Orders data from file "Order.csv"

Table 4:

	A	B	C	D
1	Row_ID	Product ID	Category	Sub-Category
2	1112	OFF-PA-10003424	Office Supplies	Paper
3	761	OFF-EN-10000461	Office Supplies	Envelopes
4	1225	OFF-EN-10000781	Office Supplies	Envelopes
5	594	OFF-EN-10001219	Office Supplies	Envelopes
6	320	OFF-EN-10001137	Office Supplies	Envelopes
7	1496	OFF-EN-10002312	Office Supplies	Envelopes
8	1375	OFF-EN-10004483	Office Supplies	Envelopes
9	37	OFF-EN-10002986	Office Supplies	Envelopes
10	1680	OFF-EN-10003068	Office Supplies	Envelopes
11	78	OFF-AP-10002118	Office Supplies	Appliances
12	1876	OFF-PA-10001583	Office Supplies	Paper
13	1036	OFF-AR-10002221	Office Supplies	Art
14	566	FUR-FU-10002759	Furniture	Furnishings
15	1724	OFF-PA-10004381	Office Supplies	Paper
16	354	OFF-ST-10000798	Office Supplies	Storage
17	1700	OFF-ST-10003805	Office Supplies	Storage
18	899	FUR-FU-10000076	Furniture	Furnishings
19	1229	OFF-AP-10002534	Office Supplies	Appliances
20	1499	FUR-FU-10003247	Furniture	Furnishings
21	1674	TEC-MA-10001047	Technology	Machines
22	1815	TEC-MA-10002073	Technology	Machines
23	347	FUR-FU-10001935	Furniture	Furnishings
24	411	OFF-AP-10002472	Office Supplies	Appliances
25	813	OFF-BI-10003669	Office Supplies	Binders
26	1636	FUR-FU-10001085	Furniture	Furnishings
27	1386	FUR-FU-10004665	Furniture	Furnishings
28	993	OFF-AP-10004980	Office Supplies	Appliances
29	626	OFF-BI-10004187	Office Supplies	Binders

Table 4: Product data from file "Products_v3.csv"

Table 5:

	A	B	C	D	E	F	G	H
1	Row ID	Order ID	Customer ID	Product ID	Sales	Quantity	Discount	Profit
2	1	CA-2019-152156	CG-12520	FUR-BO-10001798	261.96	2	0	41.9136
3	2	CA-2019-152156	CG-12520	FUR-CH-10000454	731.94	3	0	219.582
4	3	CA-2019-138688	DV-13045	OFF-LA-10000240	14.62	2	0	6.8714
5	4	US-2018-108966	SO-20335	FUR-TA-10000577	957.5775	5	0.45	-383.031
6	5	US-2018-108966	SO-20335	OFF-ST-10000760	22.368	2	0.2	2.5164
7	6	CA-2017-115812	BH-11710	FUR-FU-10001487	48.86	7	0	14.1694
8	7	CA-2017-115812	BH-11710	OFF-AR-10002833	7.28	4	0	1.9656
9	8	CA-2017-115812	BH-11710	TEC-PH-10002275	907.152	6	0.2	90.7152
10	9	CA-2017-115812	BH-11710	OFF-BI-10003910	18.504	3	0.2	5.7825
11	10	CA-2017-115812	BH-11710	OFF-AP-10002892	114.9	5	0	34.47
12	11	CA-2017-115812	BH-11710	FUR-TA-10001539	1706.184	9	0.2	85.3092
13	12	CA-2017-115812	BH-11710	TEC-PH-10002033	911.424	4	0.2	68.3568
14	13	CA-2020-114412	AA-10480	OFF-PA-10002365	15.552	3	0.2	5.4432
15	14	CA-2019-161389	IM-15070	OFF-BI-10003656	407.976	3	0.2	132.5922
16	15	US-2018-118983	HP-14815	OFF-AP-10002311	68.81	5	0.8	-123.858
17	16	US-2018-118983	HP-14815	OFF-BI-10000756	2.544	3	0.8	-3.816
18	17	CA-2017-105893	PK-19075	OFF-ST-10004186	665.88	6	0	13.3176
19	18	CA-2017-167164	AG-10270	OFF-ST-10000107	55.5	2	0	9.99
20	19	CA-2017-143336	ZD-21925	OFF-AR-10003056	8.56	2	0	2.4824
21	20	CA-2017-143336	ZD-21925	TEC-PH-10001949	213.48	3	0.2	16.011
22	21	CA-2017-143336	ZD-21925	OFF-BI-10002215	22.72	4	0.2	7.384
23	22	CA-2019-137330	KB-16585	OFF-AR-10000246	19.46	7	0	5.0596
24	23	CA-2019-137330	KB-16585	OFF-AP-10001492	60.34	7	0	15.6884
25	24	US-2020-156909	SF-20065	FUR-CH-10002774	71.372	2	0.3	-1.0196
26	25	CA-2018-106320	EB-13870	FUR-TA-10000577	1044.63	3	0	240.2649
27	26	CA-2019-121755	EH-13945	OFF-BI-10001634	11.648	2	0.2	4.2224
28	27	CA-2019-121755	EH-13945	TEC-AC-10003027	90.57	3	0	11.7741
29	28	US-2018-150630	TB-21520	FUR-BO-10004834	3083.43	7	0.5	-1665.05

Table 5: Sales data from file "Sales_v1.csv"

Table 6:

	A	B	C	D	E
1	category	city	state	region	profit
2	Technology	Thomasville	North Carolina	South	-6418
3	Technology	Pomona	California	West	-5351.66
4	Office Supplies	Miami	Florida	South	-5340.68
5	Technology	Springfield	Ohio	East	-3581.2
6	Technology	Tuscaloosa	Alabama	South	-3399.98
7	Office Supplies	San Antonio	Texas	Central	-3386.46
8	Technology	Burlington	North Carolina	South	-3185.73
9	Office Supplies	Inglewood	California	West	-3107.35
10	Office Supplies	Yonkers	New York	East	-2683.74
11	Office Supplies	Florence	Kentucky	South	-2610.66
12	Technology	Kent	Ohio	East	-2279.85
13	Office Supplies	Gulfport	Mississippi	South	-2207.93
14	Furniture	Concord	North Carolina	South	-2189.32
15	Furniture	Waterbury	Connecticut	East	-2170.56
16	Office Supplies	Santa Maria	California	West	-2159.94
17	Technology	Florence	South Carolina	South	-2149.45
18	Office Supplies	Memphis	Tennessee	South	-2143.89
19	Furniture	Aurora	Illinois	Central	-2122.82
20	Office Supplies	Mesquite	Texas	Central	-2121.77
21	Office Supplies	Royal Oak	Michigan	Central	-2058.8
22	Technology	Louisville	Colorado	West	-2040.37
23	Furniture	Fairfield	Ohio	East	-2011.94
24	Office Supplies	Colorado Springs	Colorado	West	-1880.07
25	Furniture	Redondo Beach	California	West	-1760.43
26	Furniture	Sierra Vista	Arizona	West	-1752.93
27	Office Supplies	Hickory	North Carolina	South	-1662.94
28	Furniture	Denver	Colorado	West	-1637.6
29	Furniture	Beaumont	Texas	Central	-1634.85

Table 6: MySQL query output table from file "SQL_Locations_Output.csv"

Table 7:

	A	B	C
1	category	Sub_Category	profit
2	Furniture	Tables	-17725.5
3	Furniture	Bookcases	-3413.5
4	Office Supplies	Supplies	-1189.1
5	Technology	Machines	738.4005
6	Office Supplies	Fasteners	949.5182
7	Office Supplies	Labels	5546.254
8	Office Supplies	Art	6599.121
9	Office Supplies	Envelopes	6964.177
10	Furniture	Furnishings	17590.47
11	Office Supplies	Appliances	19267.47
12	Office Supplies	Storage	21668.46
13	Furniture	Chairs	26942.63
14	Office Supplies	Binders	30676.08
15	Office Supplies	Paper	42876.46
16	Technology	Phones	51777.1
17	Technology	Copiers	55617.82
18	Technology	Accessories	61203.89
19			

Table 7: MySQL query output table from file "SQL_Sales_Output.csv"