­

GBC superstore:

Operational and Executive Reports

Table of Contents

[**1.** **Executive report** 3](#_Toc117259756)

[**2.** **Business assumptions** 4](#_Toc117259757)

[2.1 Business 4](#_Toc117259758)

[2.2 Audience and Stakeholder 4](#_Toc117259759)

[2.3 Technology 4](#_Toc117259760)

[2.4 Processes 4](#_Toc117259761)

[2.5 Report Objective 5](#_Toc117259762)

[**3.** **Data Governance Framework** 8](#_Toc117259763)

[Data Governance sub-domain 8](#_Toc117259764)

[3.2 Data Policies 8](#_Toc117259765)

[3.3 Sample Reports 15](#_Toc117259766)

[**4.** **Annex (Exercises)** 26](#_Toc117259767)

[4.1 Lab Exercise 1 – submission 26](#_Toc117259768)

[4.2 Lab Exercise 2 – submission A 33](#_Toc117259769)

[4.3 Lab Exercise 3 – submission A 39](#_Toc117259770)

## **Executive report**

The audience of the executive report is the CFO, CIO, CDO, CD stewards, and CEO. C-level executives use the report to oversee the company’s profits, operations, and resources for management and setting the company’s strategy.

The objective of the executive report is to address the summary of the operational report in terms of gross sales, profit, and shipment data. Aggregated data and KPIs were shown in the report to provide additional information and insights for the executive management board to make decisions.

## **Business assumptions**

### Business

GBC superstore is an online retail store that mainly sell furniture, office supplies, and technology products in the United States.

### Audience and Stakeholder

Sales and Marketing Team, Warehouse Team, Executive Officers, Data Stewards

The audience of the executive report is the CFO, CIO, CDO, CD stewards and CEO. C-level executives use the report to oversee the company’s profits, operations and resources for management and setting the company's strategy.

### Technology

A relational database storing sales transaction data, customer data, product data was built by MySQL.

Besides, GBC Superstore uses the SAP ERP system as the software streamlining the data flow across different business functions, which are the Sales and Marketing Team and the Warehouse Team in our scenario. Furthermore, as we operate as an online retail store, our e-Commerce website is in-house customized and hosted by an independent in-house server.

Better still, to ensure data security, we made good use of the 2-factor authentication, SHA-256 algorithm for encryption as well as dynamic data masking technique.

### Processes

The following processes are adopted:

1. Set up data governance sub-domain and follow the rules
2. Data cleansing and build a database from “sample - superstores.xls”
3. Set up the format and define the KPIs of monthly operational reports of sales team and warehouse and quarterly executive report
4. Generate the operational reports and executive reports by SQL script.

### Report Objective

#### Operational Report – Sales and marketing department

**Introduction**

The target audience of the above operational report is the sales team of the superstore. The structure of this report is organized by the “Region”, “Category” and “Sub-Category” in the data set and the operational activities are reported on monthly basis from Oct 2021 to Dec 2021. “Sales”, “Quantity” and 5 KPIs of each “Sub-Category” are recorded and the KPIs are calculated by the following definition.

**KPI Definition**

|  |  |
| --- | --- |
| **KPI** | **Definition** |
| Average Discount | (Sum of (Sales\*discount) in each sub-category) / (Sum of Sales in each sub-category) |
| Total Profit/Loss | The sum of profit/loss in each sub-category |
| Average Profit/Loss | (Total profit of each sub-category) / (Quantity of each sub-category) |
| Average Cost | (Sale – (Total Profit/Loss)) / (Quantity of each sub-category) |
| No. of Returned | Number of returned products by customers in each sub-category |

#### Operational Report – Warehouse department

**Introduction**

The objective of the executive report is to address the summary of the operational report in terms of gross sales, profit and shipment data. The aggregated data and KPIs were shown in the report to provide additional information and insights for the executive management board to make decisions.

The structure of this report is organized by the “Region” and “State” in the data set and the warehouse activities are reported on a monthly basis. “No. of orders”, “Sales” and 4 KPIs are recorded and the KPIs are calculated by the following definition.

**KPI Definition**

|  |  |
| --- | --- |
| **KPI** | **Definition** |
| Return Rate | 𝑟𝑒𝑡𝑢𝑟𝑛 𝑟𝑎𝑡𝑒 = 𝑁𝑜 𝑜𝑓 𝑜𝑟𝑑𝑒𝑟𝑠 𝑟𝑒𝑡𝑢𝑟𝑛/ 𝑇𝑜𝑡𝑎𝑙 𝑛𝑜 𝑜𝑓 𝑜𝑟𝑑𝑒𝑟𝑠 |
| Sales Amount by category (Office supplies, Technology, Furniture) | 𝑆𝑎𝑙𝑒𝑠 𝐴𝑚𝑜𝑢𝑛𝑡 % (𝑇𝑒𝑐ℎ𝑛𝑜𝑙𝑜𝑔𝑦) = 𝑆𝑎𝑙𝑒𝑠 𝐴𝑚𝑜𝑢𝑛𝑡 (𝑇𝑒𝑐ℎ𝑛𝑜𝑙𝑜𝑔𝑦) /𝑇𝑜𝑡𝑎𝑙 𝑆𝑎𝑙𝑒𝑠 𝐴𝑚𝑜𝑢𝑛𝑡 (𝑎𝑙𝑙 𝑐𝑎𝑡𝑒𝑔𝑜𝑟𝑦) |
| Delivery lead time (Days) | 𝐷𝑒𝑙𝑖𝑣𝑒𝑟𝑦 𝑙𝑒𝑎𝑑 𝑡𝑖𝑚𝑒 (𝑛𝑜. 𝑜𝑓 𝑑𝑎𝑦) = 𝑆ℎ𝑖𝑝 𝑑𝑎𝑡𝑒−𝑂𝑟𝑑𝑒𝑟 𝑑𝑎𝑡𝑒 |
| Average Quantity Shipped per Day | 𝐴𝑣𝑒𝑟𝑎𝑔𝑒 𝑞𝑢𝑎𝑛𝑡𝑖𝑡𝑦 𝑠ℎ𝑖𝑝𝑝𝑒𝑑 𝑝𝑒𝑟 𝑑𝑎𝑦 = 𝑇𝑜𝑡𝑎𝑙 𝑠𝑎𝑙𝑒𝑠 𝑞𝑢𝑎𝑛𝑡𝑖𝑡𝑦 𝑖𝑛 𝑡ℎ𝑒 𝑚𝑜𝑛𝑡ℎ/𝑁𝑜 𝑜𝑓 𝑑𝑎𝑦𝑠 𝑖𝑛 𝑡ℎ𝑒 𝑚𝑜𝑛𝑡ℎ |

By monitoring the delivery lead time in different regions, we could understand the efficiency of different warehouses.

By overseeing the Sales mix by Category ratio, we could have a brief idea of which category contributes to a larger proportion of the warehouse spaces, so that we could design the warehouse better and allocate resources to be tailor-made for the customers. For example, for the East region, 71% of the products are delivered as Technology, so the warehouse team may need to allocate more people to handle the fragile technology more carefully.

By monitoring the return rate in different regions, we could understand the customers’ satisfaction with the products. One of the possible reasons could be products being damaged during delivery/warehouse storage processes.

#### Executive Report

**Gross sales**

Gross sales is a Key Performance Index (KPI) to evaluate the total sales in a company without calculating the costs. It is a direct indicator to determine the performances of regional managers quarterly, which are a true picture of the top line in a company. Therefore, executives can consider the sales growth in different regions to adjust their strategies and make personnel changes. Considering the gross sales in terms of the category product, they can identify the bestselling categories in the season and deploy effective commercial strategies for the particular season. However, gross sales do not involve operational costs, taxes, or charges.

**Gross Margin**

Gross margin is a Key Performance Index (KPI) to evaluate a company's profitability. It impacts whether a company can secure financing from a bank, attract investors to fund its operations and grow its business. Also, it helps business owners, and their advisors determine whether a business can, or should, grow. Consequently, Executive officers can use profit as an indicator not only for cost control but also for expansion.

Gross margin is compared in different regions to estimate the performance of regional managers. It is calculated yearly.

**Shipment**

Quarter over quarter percentage change (QoQ % change) and Year to date percentage change (YTD % change) are the KPI used to indicate the changes in number of shipments in the four regions. With positive value, the KPI indicates the growth in the number of shipments which means the cost of shipment increases and the number of items sold increases. The executive should allocate more logistics resources and hire more staff to support the daily shipping operations. Outsourcing the shipment operation to other logistics companies is another method to relieve the operation pressure. Based on the demand and supply of shipment of different ship mode, the executive could adjust the shipment price.

## **Data Governance Framework**

### Data Governance sub-domain

Four sub-domains from Data Governance Framework were identified and chosen from DAMA wheel, which are deemed to be applicable on GBC Superstore,

The chosen sub-domains are Data Modeling and Design, Data Quality, Data Security and Data Privacy.

### Data Policies

#### Data Modeling and Design

**Policy 1: Use concise and consistent naming convention**

It is important to set up rules of the naming before creating the schema. Use logical convention and keep it consistent throughout the schema can provide meaningful messages to users and prevent them from getting frustrated. It is also easier to memorize, update and maintain the model for the one who needs to update it in the future.

* Use snake case, avoid using space in table / column names

In GBC superstore’s modeling, all table names and column names are using snake cases such as “customer\_name”, “product\_name”......

* Use “\_date” suffix for all date type columns

For columns with date type, “\_date” suffix has been included in the name such as “ship\_date”, “order\_date”

* Use “\_id” suffix for all ID type columns

For ID type data, the column name always has “\_id” suffix such as “order\_id”, “customer\_id”......

* Avoid using SQL / database reserved keywords to name the table / column
* Table / column names only contains letters and numbers
* All entity table names are in plural form
* Use meaningful name that can be easily understand by user

**Policy 2: Database design followed the normalization rule**

In order to organize the business complexity into data structures, databases with normalization rule designs are vital to increase the data consistency and reduce the data redundancy. Each level corresponds to a separate normal form. Second normal form is applied to ensure each entity has the minimal primary key and every attribute depends on the primary key. For instance, the customer table contains customer ID as the primary key and customer ID is the foreign key in the order table. There are three main characteristics to ensure the high quality in definition of entities, including clarity, accuracy and completeness. The definition is easy to read and precise with the correct description. Moreover, the cardinality captures the data rules in the correct syntax which shows how many instances of one entity are involved in the relationship with the other entity. For instance, customers have one to many relationships to orders.

#### Data Quality

**Policy1: Measure and Monitor Data Quality by Completeness and Uniqueness**

We measure the data quality of “Sample - Superstore.xls” by completeness and uniqueness and the detailed information is shown in the table below:

|  |  |  |  |
| --- | --- | --- | --- |
| **Dimension** | **Measure** | **Metrics** | **Status Indicator** |
| Completeness | Count the number of records where data is populated, compared to the total number of records. | (no. of records populated) \*100 / (total no. of records) | Unacceptable:  < 80% populated |
| Uniqueness | Count the number of duplicate records identified; report on the percentage of records that represent duplicates. | (no. of duplicate records)\*100 / (total no. of records) | Unacceptable:  > 0% |

Application of Completeness and Uniqueness

We applied the above measurement to “Sample - Superstore.xls” and the result is as follows:

Completeness:

* Total no. of records in table “Orders” of “Sample - Superstore.xls”: 9994
* Columns with missing data in table “Orders” of “Sample - Superstore.xls”: “Postal Code”
* No. of missing data in “Postal Code” in table “Orders”: 11
* Completeness of “Postal Code” in table “Orders”:

(9994 - 11)\*100 / (9994) = 99.889% (Unacceptable < 80% populated)

Uniqueness:

* Total no. of records in table “Returns” of “Sample - Superstore.xls”: 800
* Duplicate data in table “Returns” of “Sample - Superstore.xls”: “Order ID”
* Columns with missing data in table “Returns” of “Sample - Superstore.xls”: “Order ID”
* No. of duplicate data in “Order ID” in table “Returns”: 710
* Uniqueness of “Order ID” in table “Returns”:

(710)\*100/(800) = 88.75% (Unacceptable > 0%)

During the data cleansing, we applied the following solutions:

* Replace the missing data of the “Postal Code” by 10000.
* Drop the duplicate data in “Order ID” in table “Returns” and get 296 unique values in “Order ID”.

**Policy 2: Identify Critical Data and Business Rules**

Critical data is critical to success in a specific business area or business process. There are some criteria for data to become critical, such as business facts that are deemed critical and support business processes to the organization. In the operation, it is important to understand sales, quality, discount, and profit with respect to region and sub-category, because the operation team needs those KPIs to find out which products are the most and least profitable, how many discounts they make, and what is the average or marginal cost of that products. Also, critical data is for unique identifiers of things important to business, in our project, we defined the customer ID, order ID, product ID, and address ID being our primary key in the customer table, order table, product table, and address table, those are the elements that uniquely identify in our table.

#### Data Security

**Policy 1:** **Prevent Inappropriate Access to our Data Assets**

* Access:

GBC superstore employees with different levels of authorization can access different parts of the database systems.

The user account must be secured by a strong password. Any weak password that does not meet the standards for a strong password would not be accepted. The standards for a strong password should include a mix of upper case, lower case alphabet, special characters, and digits which are not shorter than 10 characters.

* Authentication

2-Factor authentication methods including security token, security questions, or SMS verification should be adopted to validate a user's identity. All transmissions during authentication are encrypted to prevent theft of the authenticating information.

* Encryption

The database should be encrypted with the SHA-256 algorithm since there is no known vulnerabilities and therefore exploited by hackers to decrypt the database.

* Masking

The data in the database should be masked with dynamic data masking so that different amounts of sensitive data is revealed for employees from different departments according to their operational requirements. For example, the customer ID would be masked for the warehouse department since they do not need such data for delivery.

**Policy 2: Reduce Risk by Reducing Exposure**

In order to minimize the risk of data leakage during data creation/ data transfer and data storage, GBC superstore has implemented the following restrictions on its employees.

Company computers, which have access to sensitive data, must not be used for personal matters.

All communication, regardless of internal or external, must be performed through company email which will be monitored by the IT department. All instant messaging, public emailing services, social networking sites, and file-sharing services should be prohibited anytime to avoid leakage of sensitive or confidential data.

To protect the company against exposure from hacking, malware and phishing, all pornography and gambling sites should be prohibited anytime. No exceptions should be allowed.

#### Data Privacy

**Policy 1: Obtain Consent before Personal Information Collection**

To comply with the principle of obtaining consent, customers must consent before performing collection, use or disclosure of any personal information. During the purchase, customers will need to provide their full name, shipping address, email address, postal code and date of birth. After getting permission from them, the purpose for the collection, use or disclosure of personal information needs to be clearly and transparently informed individually. The reason we need their full name, address and postal code is for the sake of shipping. As for the date of birth, Superstore will offer a birthday discount to customers. Email can be one of the promotion ways for Superstore to reach out customers. Apart from getting permission from customers and explaining the use of data, the agreement must not be obtained by providing false or misleading information. If users chose to withdraw consent, the consequences must be clearly explained. For instance, Superstore online shopping website will no longer be available for withdrawn users. If users decided to withdraw consent to the collection, use or disclosure of personal information, it can never be prohibited unless withdrawing consent would lead to a conflict with a legal obligation. There should be an easy procedure for users to withdraw consent like pressing ‘withdraw consent to the data collection, use, disclosure’.

**Policy 2: Identify the Purpose in Personal Information Collection**

Identifying the purpose for collecting personal information should be entirely considered during the time of collection. The purpose and the method that the Superstore uses the personal information should be determined first. It is significant to confirm that the collection of personal information fulfills the initial purpose and is accepted by the general public. Before the time of collection, the customers should receive a verbal or written agreement about the usage and how to manage their personal information. In the agreement, the basic information, such as name, address, and email address of the Superstore will be provided. The Superstore will inform and obtain consent from the customers if the company considers new and unspecial purposes for using their personal information. However, the Superstore is not able to use it incorrectly, such as sharing, editing or disclosing the information.

In individual order, the Superstore collects sufficient information, including customer name, postal code, and address, to finish the order activity. However, personal information unrelated to the order, for example, Social Insurance Number (SIN), parent personal information, bank number, will not be collected.

### Sample Reports

#### Sample Operational Sales report

**Operational Report - Sales Team**

Oct 2021 Monthly Sales





**Operational Report - Sales Team**

Nov 2021 Monthly Sales





**Operational Report - Sales Team**

Dec 2021 Monthly Sales

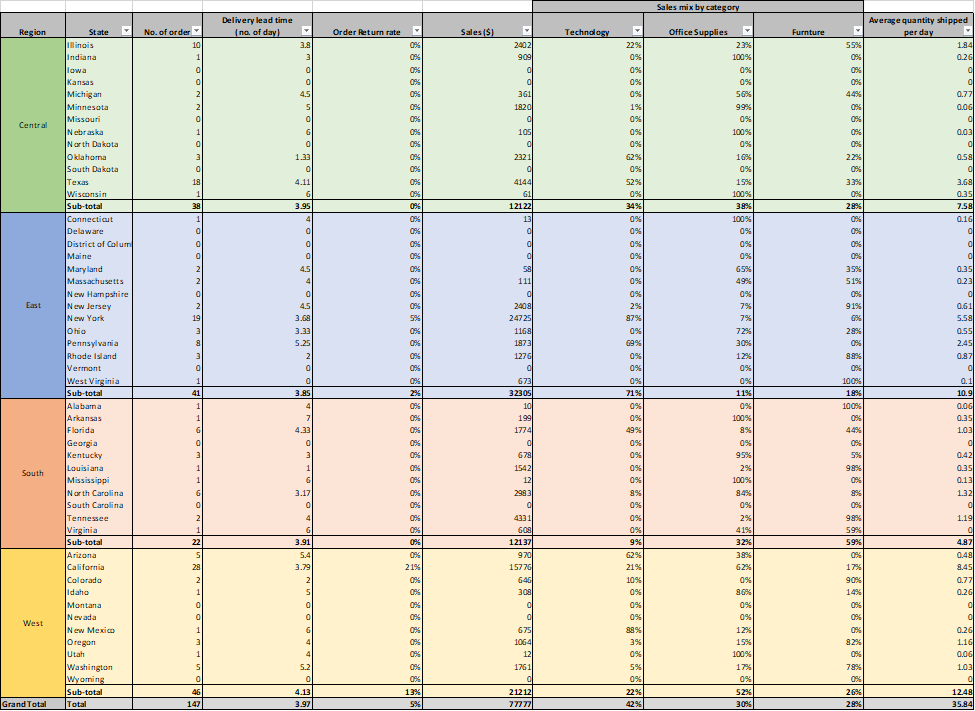




#### Sample Operational Warehouse Report

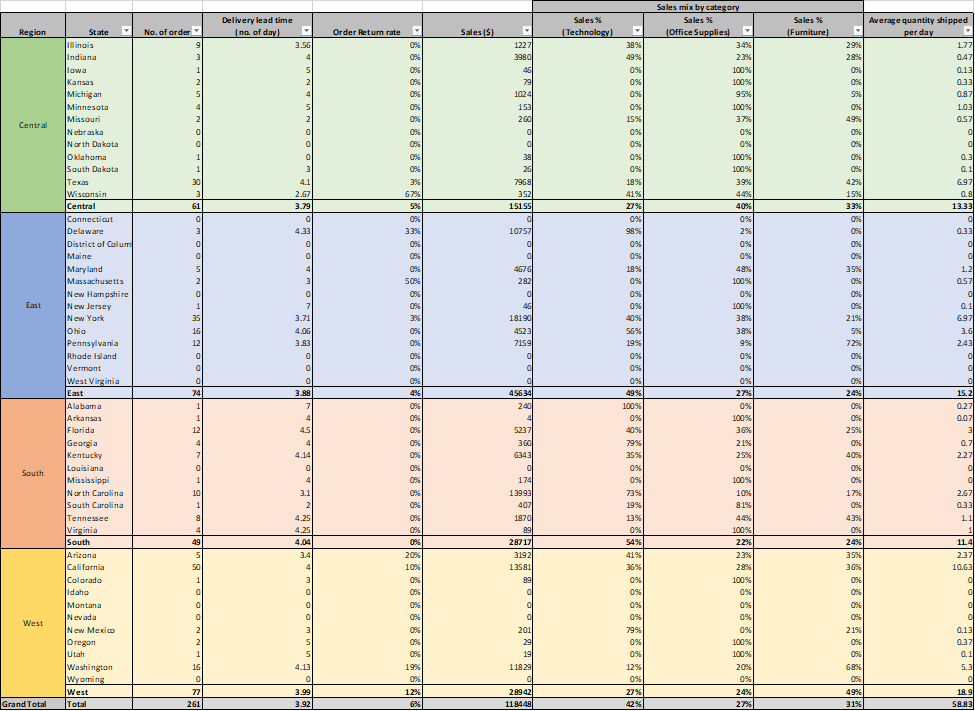
Monthly Operational Report - Warehouse Team

Oct 2021



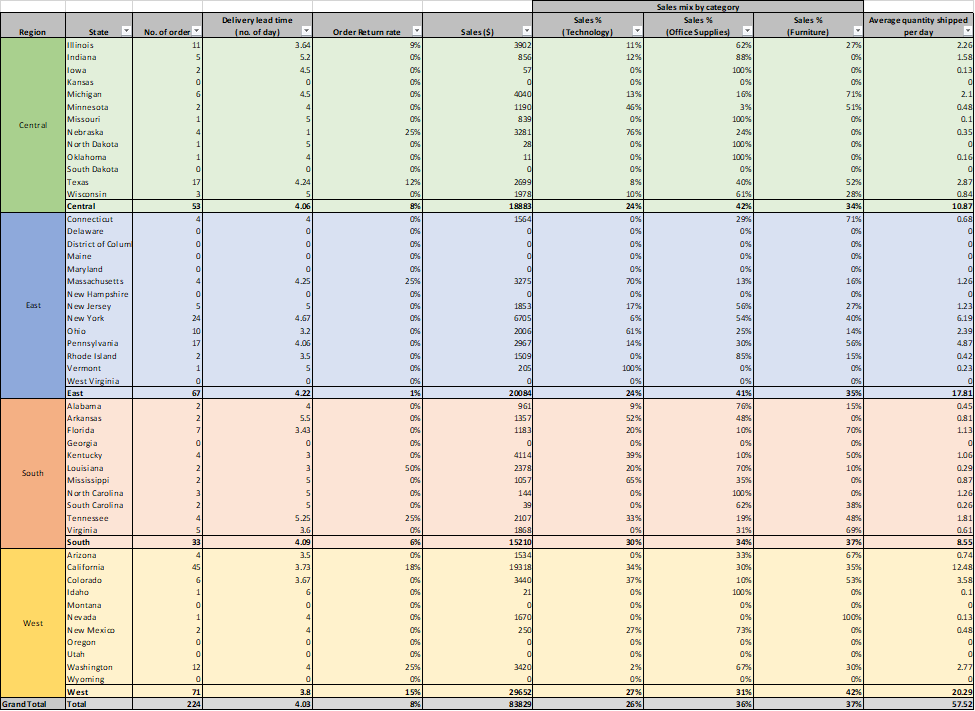
Monthly Operational Report - Warehouse Team

Nov 2021



Monthly Operational Report - Warehouse Team

Dec 2021



#### Sample Executive Report

Quarterly Executive Report

Quarter 4, 2021





## **Annex (Exercises)**

### Lab Exercise 1 – submission

1. **Analysis**
2. Completeness:
3. Some empty values (count = 11) exist in the field “Postal code”. They could be filled with previous orders from the same customer (same customer ID).

1. The “unit price before discounts” is missing in the dataset since the field “Sales” is the total price after discounts.

1. The “unit cost” is missing.

1. Inconsistencies:
2. For the field "States”, some of the records are inconsistent with the “City” input. For example, there are 3 different inputs for “States” for the “City” Concord which are “California”, “New Hampshire” and “North Carolina” while the actual input should be “New Hampshire”.

1. The field “product ID” is inconsistent with the field “Product name”, for example, the product ID “OFF-BI-10004632" appears with the product name “Ibico Hi-Tech Manual Binding System” and “GBC Binding covers” respectively.

1. For the same product, the unit price before discount is inconsistent between different orders.

1. Redundancies:
2. The fields “State”, and “Region” are dependent on the field “City” and thus redundant. The mapping information between “City” and “State” could be stored in a separate table, while the mapping information between “State” and “Region” could be stored in another mapping table to avoid such inconsistency above.

1. “Profit” could be calculated with “Unit Price”, “Quantity”, “Discount” and “Cost”.

1. Duplicates:
2. In the return table, the same “Order ID” indicates appears more than once, meaning there is either a more detailed product ID missing or there is a duplicate in the Order ID in the return table.

1. **Target Audience**

1. Operational Report:
2. Marketing and Sales/ Warehouse/ Store Manager:

The Marketing and Sales department uses it for strategy planing on sales and promotions. The warehouse uses it for monitoring delivery and dispatching stocks. The Store Manager uses it for stocktaking and outlet performance evaluation.

1. Project managers (for internal or external application):

The Project Manager will use it for building and updating internal or external applications, such as fitting the CRM systems in the company to streamline the operations.

1. Enterprise Architects:

The Enterprise Architects will use it for controlling internal workflow inclining to its planned organization structure.

1. Steering Committee:

Use the operational report to provide support, guidance and oversight of progress of a project from management level.

1. Audit teams:

Follow-up the operational activities to evaluate the risk and control environment.

1. Compliance Team:

Checking whether the employees obey the procedures and policies by reading the operational activities on a daily, weekly, biweekly or monthly basis.

1. Data steward, Data Architect, and DM professionals:

The operational activities and KPI recorded in the report may provide some useful metadata that helps them do data management.

1. Data owners:

Help monitor and keep track of the lineage of the data by the record on the operational report to protect the privacy and security.

1. Executive Report:
2. Regulatory:

The regulator can propose or update rules based on corporate activities shown in the report.

1. Maturity Assessors:

Maturity Assessors use the overview provided by the report to assess the data strategies used by the company and evaluate the data maturity level.

1. CIO, CDO, CD stewards, CEO:

C-level executives use the report to oversee the company’s profits, operations and resources for management and setting company’s strategy.

1. Data Governance Bodies:

The executive team builds the data governance to give an enterprise perspective to keep data consistency, security, and privacy. It also gives clear instructions for the operation team.

1. Compliance Team:

The compliance team uses data to keep track of data integrity, and make sure there is no illegal use of data.

1. Architecture Group:

The architecture group uses data to plan future workflow and supply chain. Optimize the shipping route and increase marginal profit.

1. Partner Organization:

The partner organization will display the KPI to other partners to find cooperation with potential companies.

1. Data Owners:

Executive-level data owners will maintain the master data with other suppliers, such as order ID and product name.

1. Data Steward, Data Entry:

The overall KPI recorded in the report provides insight for executive users to make reports and enter new information to keep track of weekly and monthly performance.

1. **Context and Additional Assumptions:**

1. Operational Report:
2. Detail, High Frequency to Update:

The report is short-term and includes details of daily operations, that will be updated in high frequency.

1. Quantity, Product Name, Product ID, Postal Code, Customer Name, Customer ID, Ship Date, Order ID:

Staff can refer to the quantity and product name/product ID for stock management. With customer name/customer ID on the report, the sales department can establish their sale strategy for qualified buyers. The logistics department needs postal code, ship data and order ID for delivery schedule planning.

1. Executive Report:
2. Summary, Quarterly Update:

The data should be summarized into a report by operational staff who will update them quarterly.

1. Trends in Different Time Intervals:

Executives need to know what is the best seller of the season before deploying any commercial strategies.

1. Regional Performances:

Executives need to know revenue growth from different regions to adjust their strategies and make personnel changes.

1. Profit and Loss:

Executives need to know the profit and loss for each product before making any commercial decisions in order to maximize the revenue.

1. Discount, Sales & Profit, Category, Sub-category, Region, Segment:

Executives need to know the sales and profit to decide which category of product will have a discount. Also, The region of the store and the promotion period need to be determined based on the reported data.

1. Country/ City/ State, Ship Mode, Order Date, Row ID:

For the state and region, the executives can use it and combined with revenue as an indicator to define the key region. For ship mode, the executives can arrange more transport on the most popular ship mode. For the order date, the executives can define the peak period of their business. For Row ID, it can count the total order number.

1. **Operational and Executive Reports:**

1. Operational Report:
2. Daily Sales Amount:

Marketing department and sales can judge the marketing strategy and understand the needs in the market and customers.

1. Number of Products:

Regional managers would estimate the number of the products in warehouses and adjust restocking goods.

1. Return Rate in Term of Category:

Regional managers would understand the qualities of the products and define sell or not in the Superstore.

1. Average Shipment time:

Regional managers and warehouse managers would evaluate the efficiency and performance of shipping strategy.

1. Executive Report:
2. Sales Amount (MTD, QTD, YTD, MoM, QoQ, YoY):

Regarding the executive reports, month to date (MTD), quarter to date (QTD), year to date (YTD) sales amount were shown on the first page to address the product sales trend. Month over month (MoM), Quarter over quarter (QoQ) and Year over year (YoY) are calculated to compare the financial performance between this month and last month, this quarter and the last quarter and this year and last year respectively.

MoM growth = ((This month’s value - Last month’s value)/Previous value)\*100

YoY growth = ((Current year value - Previous year value)/Previous year value)\*100

1. Sales Amount in Terms of Region or Category:

Furthermore, the sales performance in terms of regions, type of customers, category and sub-category were displayed to help the management board to make decisions on the future strategy to develop regional sales channel and put more budgets on a specific category of product to sell.

1. Type of Customers (B2B, B2C):

Defining type of customers is essential for a company to increase the loyal customers and transform business models. For instance, by defining which group of the customer segments such as consumer, corporate or home office have the highest purchasing power and develop key accounts. With higher demand for sales transactions, the company could develop a B2B or B2C business model.

1. Performances of Regional Managers:

Moreover, by calculating the net sales and net profit of each region, the management board could evaluate the performance of regional managers and make decisions on relocation, hiring, termination of staff.

1. **Templates:**

Operational Report (Template)

Sales Team Daily Operation (01/01/22 – 01/02/22)

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Order ID** | **Product ID** | **Product Name** | **Postal Code** | **Quantity** | **Customer Name** | **Customer ID** | **Ship Date** | **Daily Sales Amount ($)** | **Number of Products** | **Return Rate (Category) (%)** | **Average Shipment Time**  **(Days)** |
|  |  |  |  |  |  |  |  |  |  |  |  |

Executive Report (Template)

Sales Team Quarterly Executive Report (01/01/22 – 03/31/22)

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Row ID** | **Order Date** | **Country/**  **City/State** | **Ship Mode** | **Sales & Profit ($)** | **Discount(%)** | **Category** | **Sub-category** | **Region** | **Segment** | **Sales Amount ($)** | **Sales amount (Region or Category) ($)** | **Type of Customer** | **Performances of Regional Managers** |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |

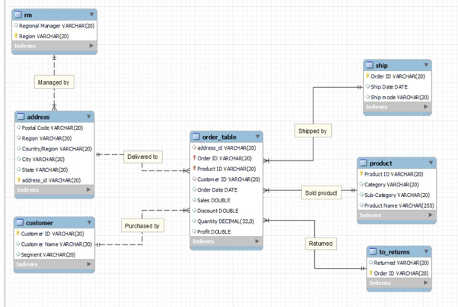
### Lab Exercise 2 – submission A

**Introduction**

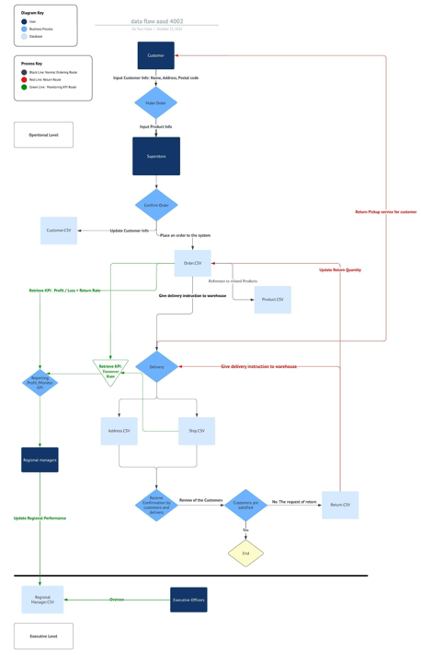
A Super Store would like to reorganize the data management for in-house usage to improve productivity and operation among departments and staff. By following the entities and normalizing the dataset, the reconstructed database would be fulfilled the operational-level support to individual departments and individual staff. With entitled checking to the dataset, the dataset was cleaned with logical skills and then the cleaned dataset would be constructed to database. Python and MySQL are the language which is applied. Our project would illustrate the logical-level ERD, Data Flow, data schema, source changes and the ELT process.

**Methodology**

**a. Logical-level ERD**



**b. Data Flow Diagram**





1. Normal Order Route

After customers enter their name, address, and postal code to make an order, the product information will pass to the superstore. Besides update the Customer.csv, the Customer Order will also update the Order.csv, placing the order to the system. With the Product.csv being referenced correctly, a delivery instruction will be made to the warehouse. After using the information as an instruction to warehouse, they will deliver to the customers’ addresses by using address.csv and ship.csv. Final confirmation will be made by the customers and delivery person. If the customers are satisfied, the process will end. However, if they are not satisfied, it will trigger the return route.

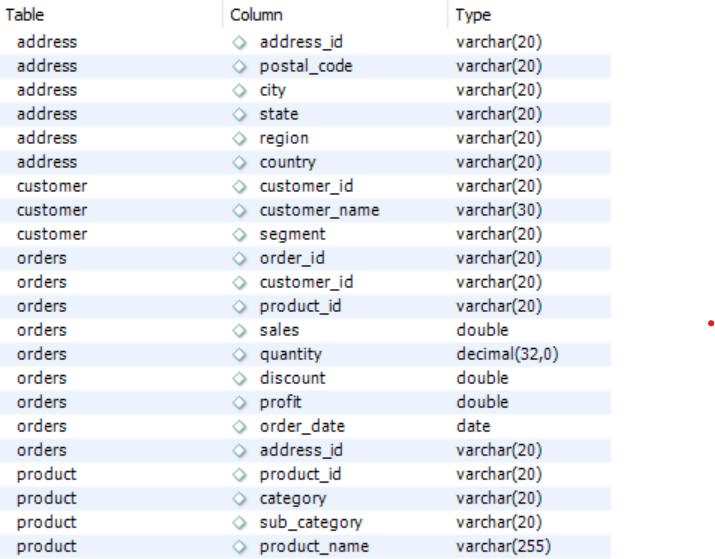
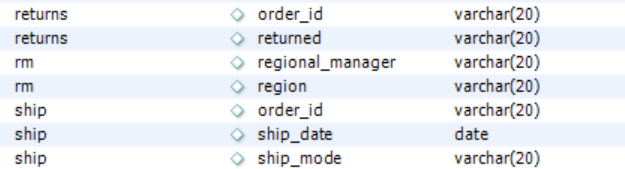
1. Return Route

We trigger the return shipping by updating the Return.csv and the Ship.csv, then giving delivery instruction to warehouse. Afterwards, a Return Pickup service will be given to the customers. On the other hand, it will update the Order.csv for recording the return rate.

1. KPI Monitoring Route

The data analyst would calculate various KPI such as Sales, Profit/Loss and return rate by retrieving data from Order.csv and Return.csv, as well as the delivery time from the Order.csv and Ship.csv to analyse the sales & inventory performance. Afterwards, it will be connected to the Regional Managers.csv so that the Executive Officers could oversee the regional performance.

**c. Database schema (include all tables, fields)**

**d. Data source changes**

The data in the excel spreadsheet is imported into a python notebook to conduct data cleaning. During the data cleansing process, the following issues were fixed.

1. Postal Code

The empty values of Country in 'United States', State in 'Vermont', and City in 'Burlington' were replaced by new value “10000”.

Some of the postal codes appear with different cities which should be an inconsistency. For example, the postal code “92024” appears with the name “San Diego” and “Encinitas” respectively. It is assumed that erroneous input would be less frequent than correct input. Therefore, "San Diego” is used to replace “Encinitas” in these records.

1. Product name and Product ID inconsistency

For the same product ID in different orders, it appears with different product names. It is assumed that erroneous input would be less frequent than correct input. Therefore, the occurrence of different product names under each product ID was counted and the one with the highest frequency was chosen as correct input to replace other records.

1. Product Name

Since the column Product Name includes utf-8 charsets that cannot be imported to MySQL database, these characters are removed with python script.

1. Order ID and product ID duplicates

For some of the records, the same Product ID appears more than once with the same Order ID. This is not allowed in the 1st normal form since the Order ID and Product ID are the composite primary key. Therefore, the duplicate purchases of the same products are grouped where sales, quantity, and profit are added up, while the discount is the weighted average by sales.

**e. ETL process**

The cleaned dataset is stored in the spreadsheet “Updated\_Dataset.xlsx”. The ETL process was partially performed in python and MySQL respectively.

1. Extract

With python, separated CSV files were created based on entities. First, based on the ERD diagram, write the create table script and execute to the database.

1. Transform

Second, import the data sheets to the relative tables through MySQL workbench’s table import wizard to the MySQL database. After data import, select the row count of each table, which should match with the number of records in the csv files.

After that, the table schema was created on the database.

1. Address

The Address table stores the information of ‘Country/region’, ‘City’, ‘State’, and ‘postal code’. Since ‘full address’ could potentially be added as a new column in this table, then the ‘postal code’ is no longer unique, therefore a new primary key ‘Address\_ID', which was auto-increment number form 0, was created in the address entity. ‘Country/Region’ is the foreign key that links to the relationship manager table. ‘Postal Code’, ‘Region’, ’Country/Region’, ‘City’, ‘State’ and ‘Address\_ID’ are stored in varchar format.

1. Customer

The Customer table was created by extracting all ‘customer ID’, ‘customer Name’, and ‘Segment’ from the cleaned dataset and then removing all duplicates records. Customer ID is the primary key of the customer table. ‘Customer ID’, ‘Customer Name’ and ‘Segment’ are stored in varchar format.

1. Order

The Order table is a master table that links other tables. Order ID and Product ID are the primary key. ‘Address ID’ and ‘Customer ID’ are stored in varchar format. ‘Order Date’ is stored as date format. ‘Sales’, ‘Discount’ and ‘Profit’ are stored as double format. ‘Quantity’ is stored in integer format.

1. Product

The Product table was created by extracting all ‘Product Name’, ‘Product ID’, ‘Category’ and ‘Sub-Category’ from the cleaned dataset and then removing all duplicate values. ‘Product ID’ is the primary key of the Product table. ‘Product Name’, ‘Category’ and ‘Sub-Category’ are stored in varchar format.

1. Regional manager

The regional manager table was created by extracting ‘Region’ and ‘Regional Manager’ from the dataset. ‘Region’ column is the primary key in this table, and being referenced in the Order table as foreign key. Both ‘Region’ and ‘Regional Manager’ are stored in varchar format.

1. Shipment

The shipment table was created by extracting all ‘Ship Date’ and ‘Ship Mode’ from the dataset. ‘Order ID’ is unique and defined as the primary key that links to order table, ‘Ship Date’ is stored in date format and ‘Ship Mode’ is stored in varchar format.

1. Returns

The Return table was created by extracting all ‘Return date’ and ‘Order ID’ from the cleaned dataset and then removing all duplicate values. The items in ‘Order ID’ are unique, and defined as primary key. The data types of all columns are variable characters with a length of 20.

1. Load

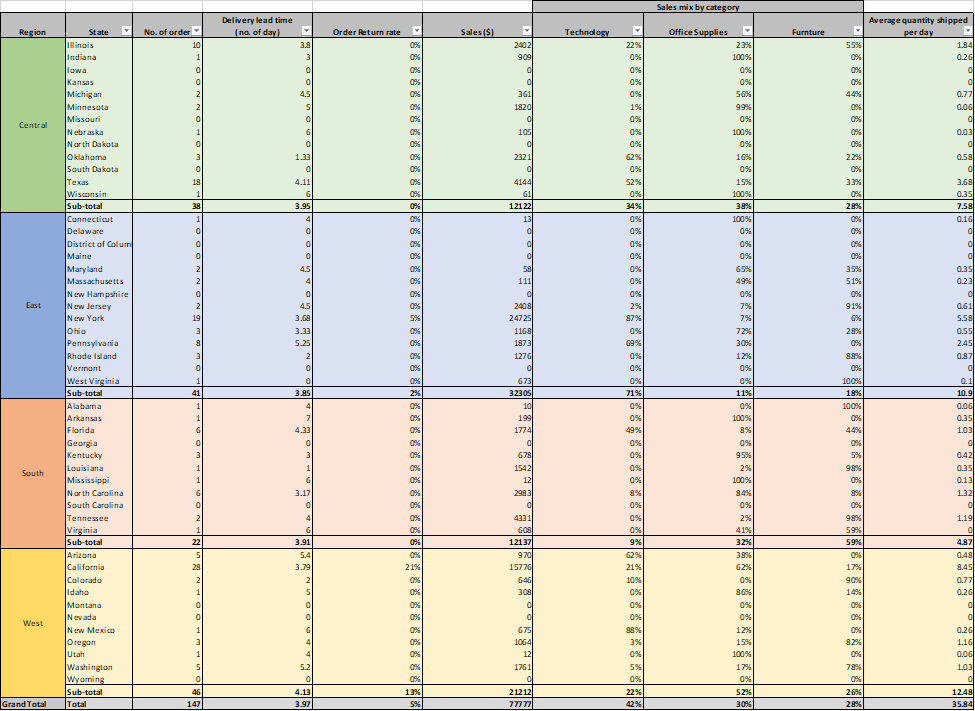
Transformed data is loaded to database table.

Referential integrity can be checked while importing data from cleaned csv files to the tables since foreign keys are declared in the table schema. Error is prompted when deleting the records from the child table but not deleting the record from the mother table. This proves that the reference integrity is complete and valid.

### Lab Exercise 3 – submission A

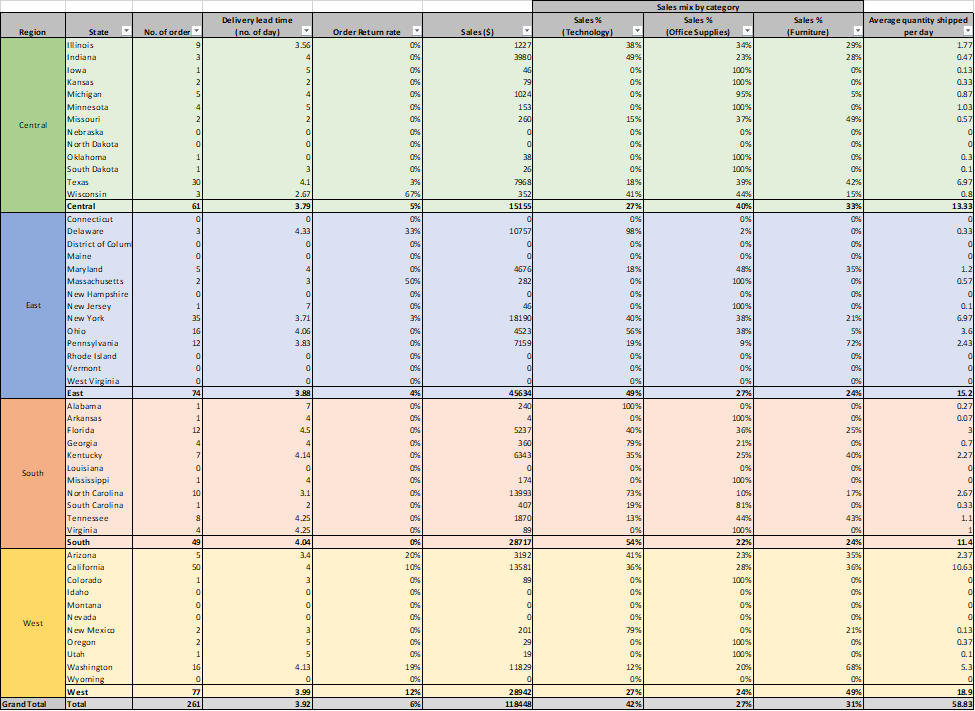
Monthly Operational Report - Warehouse Team

Oct 2021



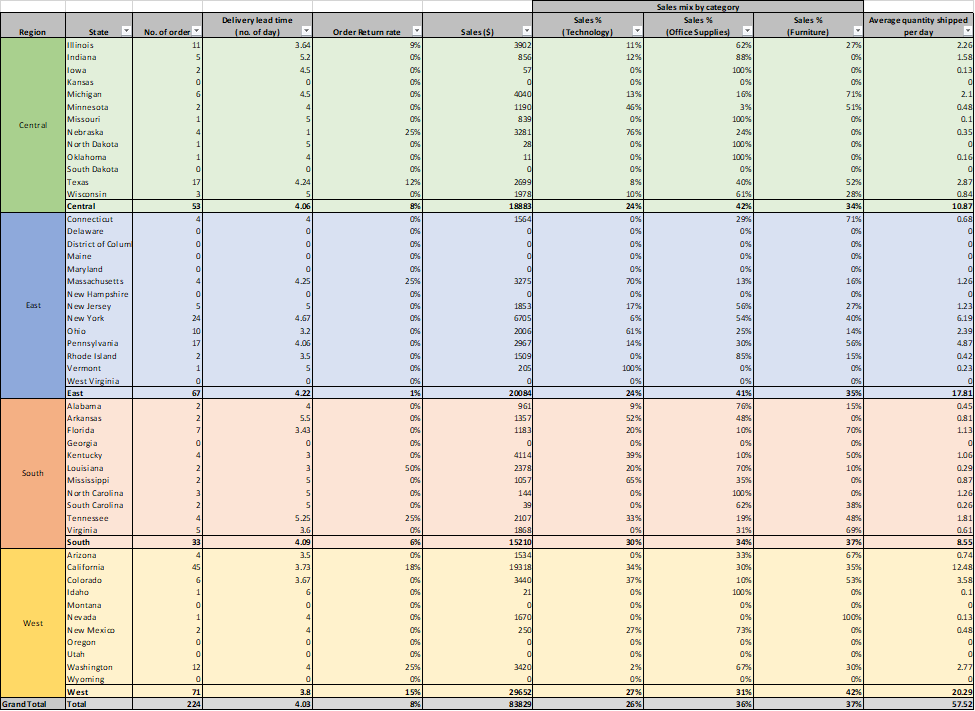
Monthly Operational Report - Warehouse Team

Nov 2021



Monthly Operational Report - Warehouse Team

Dec 2021



Introduction

The target audience of the above operational report is the warehouse team of the superstore. The structure of this report is organized by the “Region” and “State” in the data set and the warehouse activities are reported on a monthly basis from Oct 2021 to Dec 2021. “No. of orders”, “Sales” and 4 KPIs are recorded and the KPIs are calculated by the following definition.

1. **KPI Definition**

|  |  |
| --- | --- |
| **KPI** | **Definition** |
| Return Rate | return rate=No of orders return/Total no of orders*￼* |
| Sales Amount by category (Office supplies, Technology, Furniture) | Sales Amount % (Technology)=Sales Amount (Technology)Total Sales Amount (all category)*￼* |
| Delivery lead time (Days) | Delivery lead time (no. of day)=Ship date−Order date*￼* |
| Average Quantity Shipped per Day | Average quantity shipped per day=Total sales quantity in the month/No of days in the month*￼* |

1. **Note**

**Insight:**

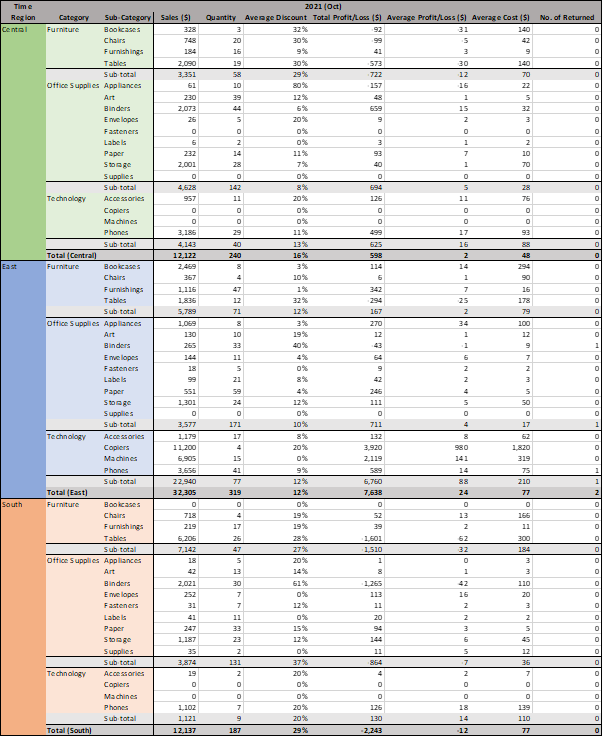
1. By monitoring the delivery lead time in different regions, we could understand the efficiency of different warehouses.

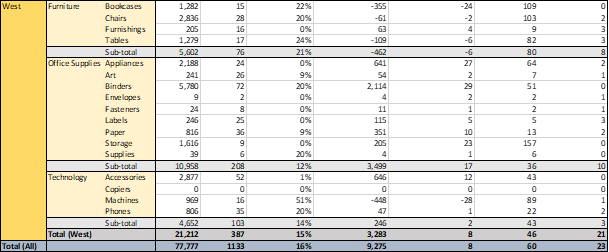
2. By overseeing the Sales mix by Category ratio, we could have a brief idea of which category contributes to a larger proportion of the warehouse spaces, so that we could design the warehouse better and allocate resources to be tailor-made for the customers. For example, for the East region, 71% of the products are delivered as Technology, so the warehouse team may need to allocate more people to handle the fragile technology more carefully.

3. By monitoring the return rate in different regions, we could understand the customers’ satisfaction with the products. One of the possible reasons could be products being damaged during delivery/warehouse storage processes.

**Monthly Operational Report – Sales Team**

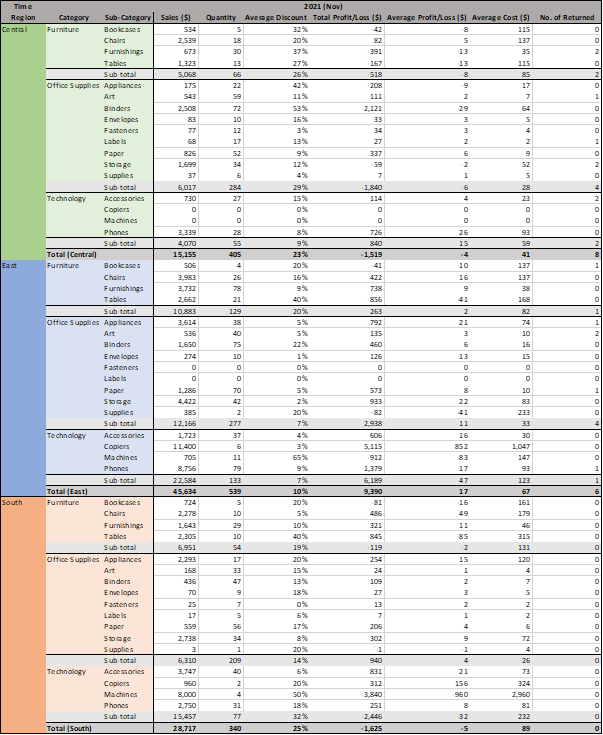
Oct 2021

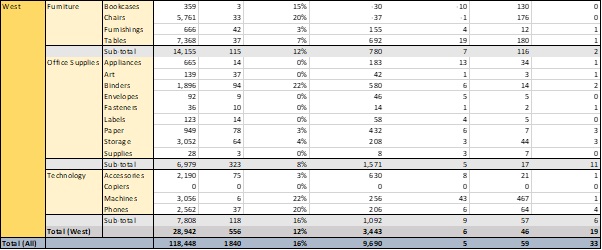




**Monthly Operational Report – Sales Team**

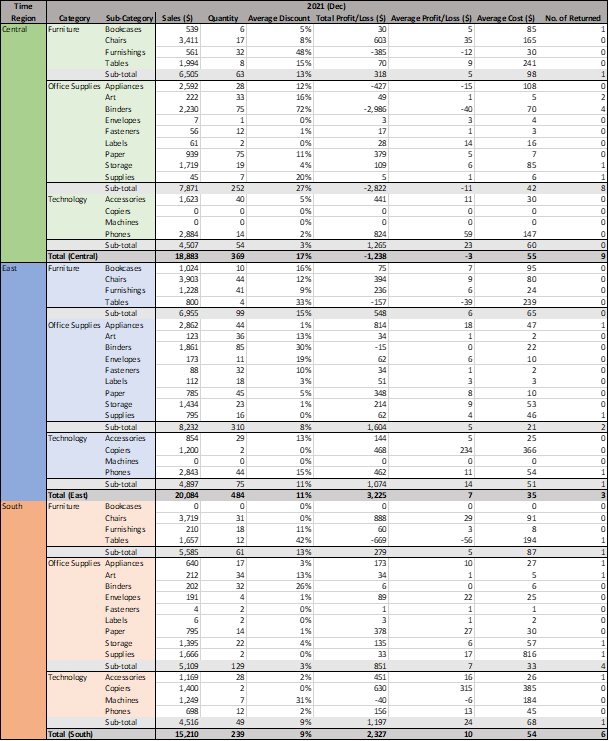
Nov 2021

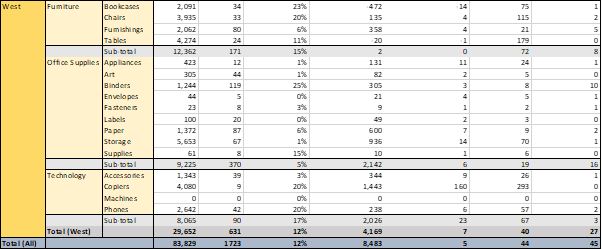




**Monthly Operational Report – Sales Team**

Dec 2021





**Summary of Operational Report (Sales Team)**

1. **Introduction**

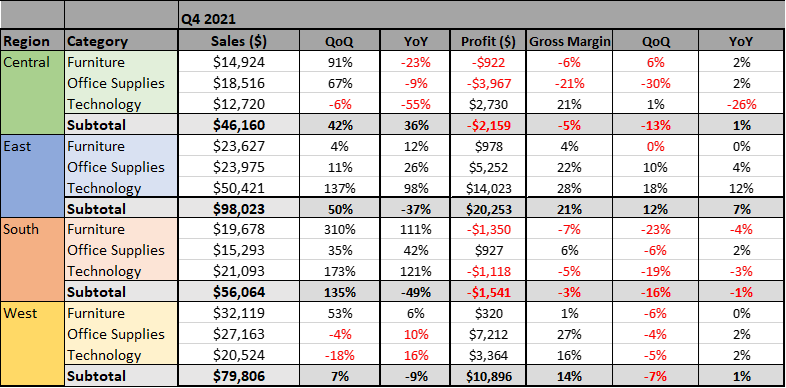
The target audience of the above operational report is the sales team of the superstore. The structure of this report is organized by the “Region”, “Category” and “Sub-Category” in the data set and the operational activities are reported on monthly basis from Oct 2021 to Dec 2021. “Sales”, “Quantity” and 5 KPIs of each “Sub-Category” are recorded and the KPIs are calculated by the following definition.

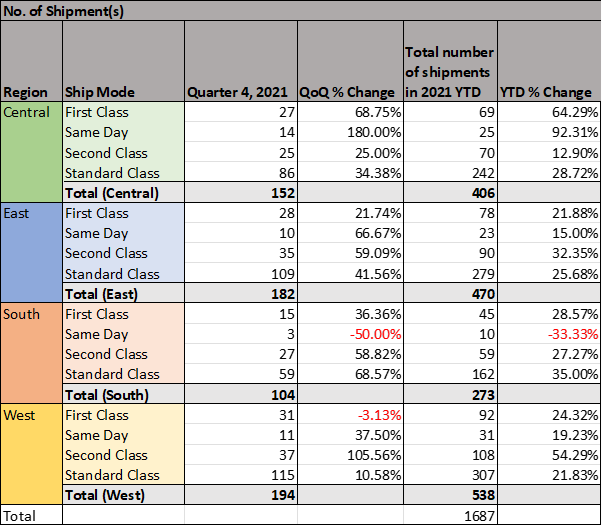
1. **KPI Definition**

|  |  |
| --- | --- |
| **KPI** | **Definition** |
| Average Discount | (Sum of (Sales\*discount) in each sub-category) / (Sum of Sales in each sub-category) |
| Total Profit/Loss | The sum of profit/loss in each sub-category |
| Average Profit/Loss | (Total profit of each sub-category) / (Quantity of each sub-category) |
| Average Cost | (Sale – (Total Profit/Loss)) / (Quantity of each sub-category) |
| No. of Returned | Number of returned products by customers in each sub-category |

Quarterly Executive Report

Quarter 4, 2021





1. **Introduction**

The report contains the sales data and shipment data which are ranging from 2021 to 2022. To illustrate difference between different period, the data is compared in quarterly and yearly.

1. **KPI**

The table below shows the KPI that we are using for our report

|  |  |
| --- | --- |
| KPI | Definition |
| QoQ sales | (Sales Current Quarter/ Sales Last Quarter - 1) |
| QoQ gross margin | (Gross margin Current Quarter - Gross margin Last Quarter) |
| YoY sales | (Sales Current Year / Sales Last Year -1) |
| YoY gross margin | (Gross margin Current Year - Gross margin Last Year) |

**Insight:**

1. By monitoring the delivery lead time in different regions, we could understand the efficiency of different warehouses.

2. By overseeing the Sales mix by Category ratio, we could have a brief idea of which category contributes to a larger proportion of the warehouse spaces, so that we could design the warehouse better and allocate resources to be tailor-made for the customers. For example, for the East region, 71% of the products are delivered as Technology, so the warehouse team may need to allocate more people to handle the fragile technology more carefully.

3. By monitoring the return rate in different regions, we could understand the customers’ satisfaction with the products. One of the possible reasons could be products being damaged during delivery/warehouse storage processes. Therefore impro