UNIVERSITI MALAYA

Centre for Continuing Education

MASTER IN DATA SCIENCE

INDIVIDUAL ASSIGNMENT

COURSE CODE : WQD7005

COURSE TITLE : DATA MINING

NAME : NUR HIDAYAH BINTI AHMAD SHAFII

MATRIC NUMBER : 22120931

LECTURER : PROF. DR. NOR LIYANA BT MOHD SHUIB

Table of Contents

1 Google AutoML

It is essential to clean and preprocess the dataset using Google DataPrep as shown in Figure 1.1 before starting the feature engineering and model-building process.

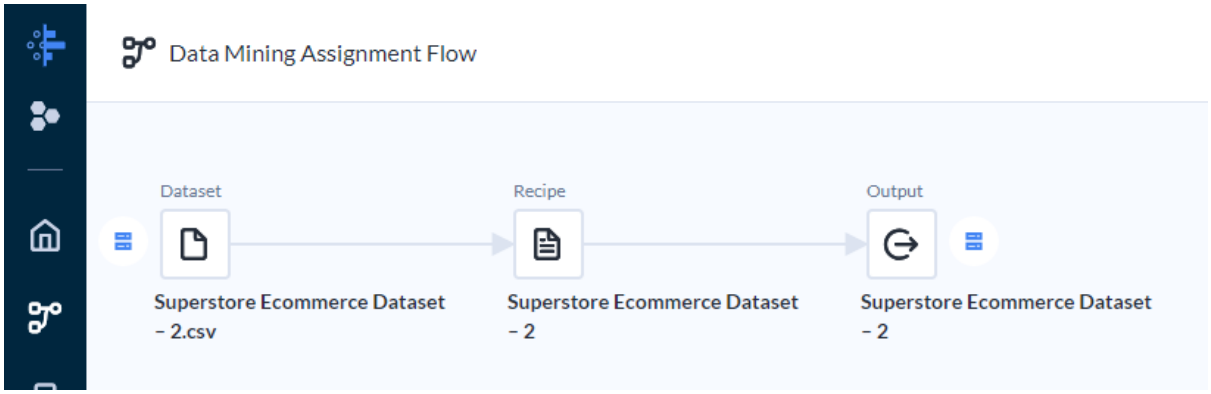

Figure1.1 DataPrep Flow

Firstly, we need to create a data flow in Google DataPrep. This is the process where we define the steps to extract, transform, and load the dataset for use in the model. After we have imported the data, we will need to explore its structure visually.

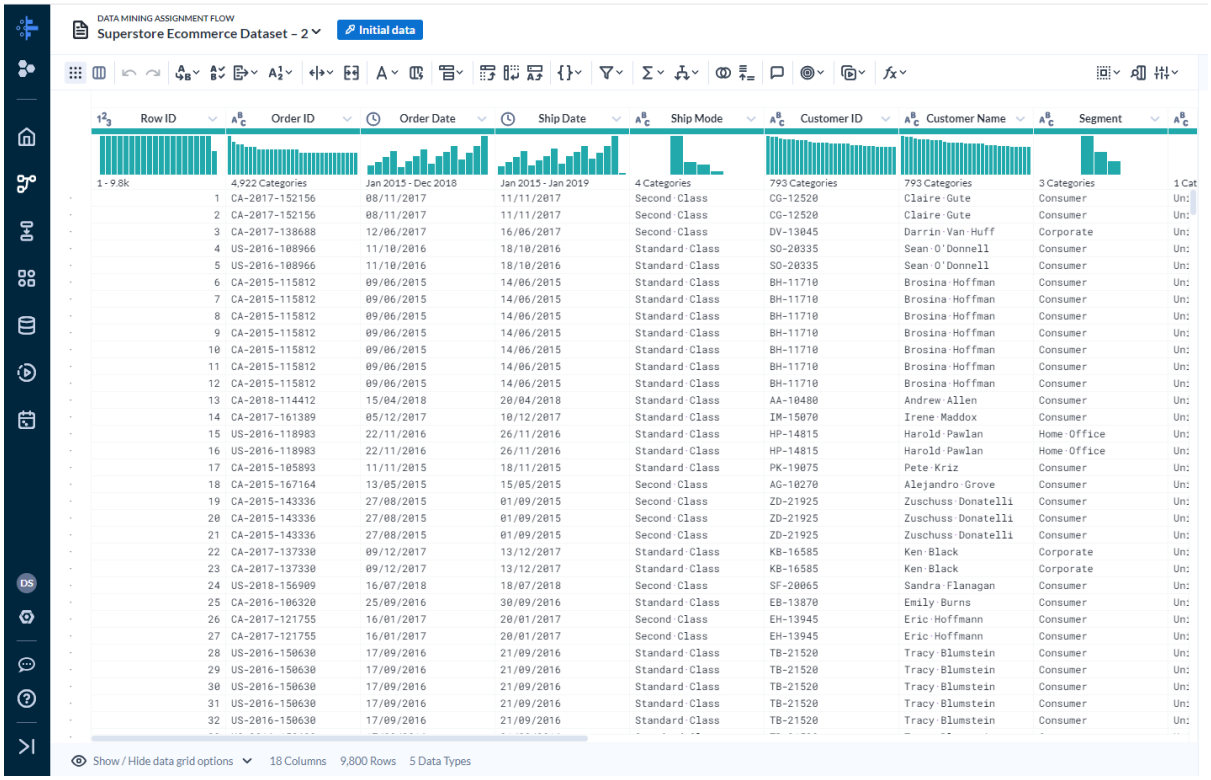

Figure 1.2 Dataset Overview

From Figure 1.2, the dataset consists of 18 column and 9800 rows. Below is the description of each column:

1. Row ID: Unique identifier for each entry in the dataset.
2. Order ID: Unique code assigned to each order.
3. Order Date: The date on which the order was placed.
4. Ship Date: The date when the order was shipped.
5. Ship Mode: The shipping method used for the order ('Second Class', 'Standard Class', 'First Class', 'Same Day').
6. Customer ID: Unique identifier for each customer.
7. Customer Name: Full name of the customer.
8. Segment: The category of the customer ('Consumer', 'Corporate', 'Home Office').
9. Country: The country where the order originated.
10. City: The city associated with the customer's location.
11. State: The state of the customer's location.
12. Postal Code: The postal code for the customer's address.
13. Region: The geographical region of the order.
14. Product ID: Unique code identifying each product.
15. Category: The category of the product.
16. Sub-Category: A more specific classification within the product category.
17. Product Name: The full name of the product.
18. Sales: The total amount of sales generated for the order.

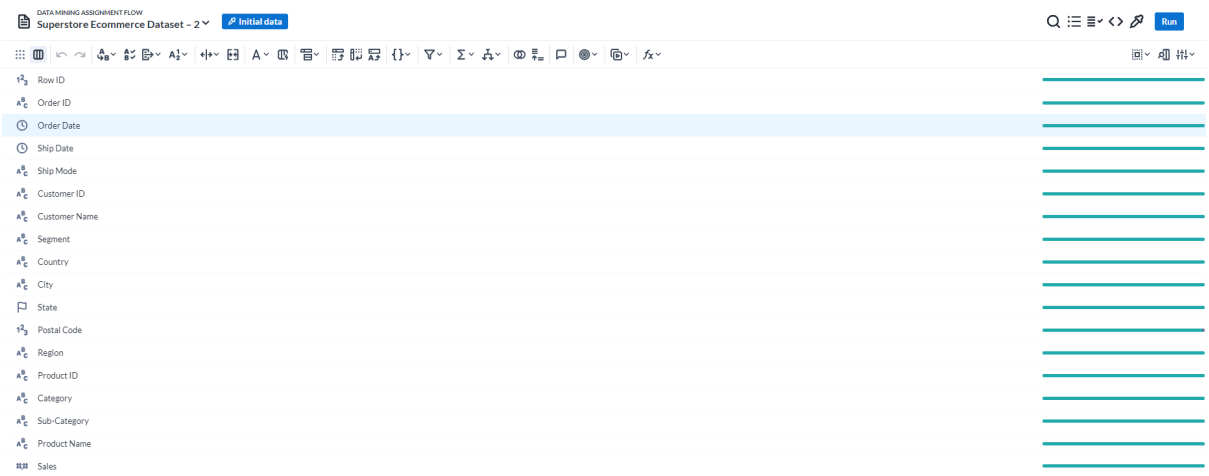

Figure 1.3 Dataset Quality

Google DataPrep automatically validated the dataset as shown in Figure 1.3, highlighting any issues with missing or invalid values. The tool's built-in data quality checks ensured all columns were correctly formatted and error-free. This automated validation helped streamline the data preparation process, ensuring the model trained on high-quality data. Each column features a horizontal bar that shows its data quality. The bars are color-coded green means the values are valid, grey means there are missing or null values, and red means the data doesn't match the expected type. In this dataset, all the columns have green bars, which means all the values are valid. Next, we need to check the data types that is automatically assigned to each column. After reviewing, we found that the data types are correct and don't need any changes.

We will now remove a few columns that aren't necessary for predicting sales. Columns like Row ID, Customer ID, Customer Name, and Product ID don't directly help with sales predictions, so we can drop them. The State column has 49 unique values, which is enough to give us geographic insights, so we'll also remove the Region, City and Postal Code columns to keep things simple. The Country column contains only one value, "United States," so it doesn't add much value and will be removed. Next, we'll

format the Sales column to show values with two decimal places. We'll also create a new column called Ship Duration, which will calculate the difference in days between the Order Date and Ship Date. Additionally, we'll extract the Month and Year from the Order Date and create new columns called Order (Month) and Order (Year). Finally, since we've created new columns for ship duration and order date details, we'll drop the original Order Date and Ship Date column.

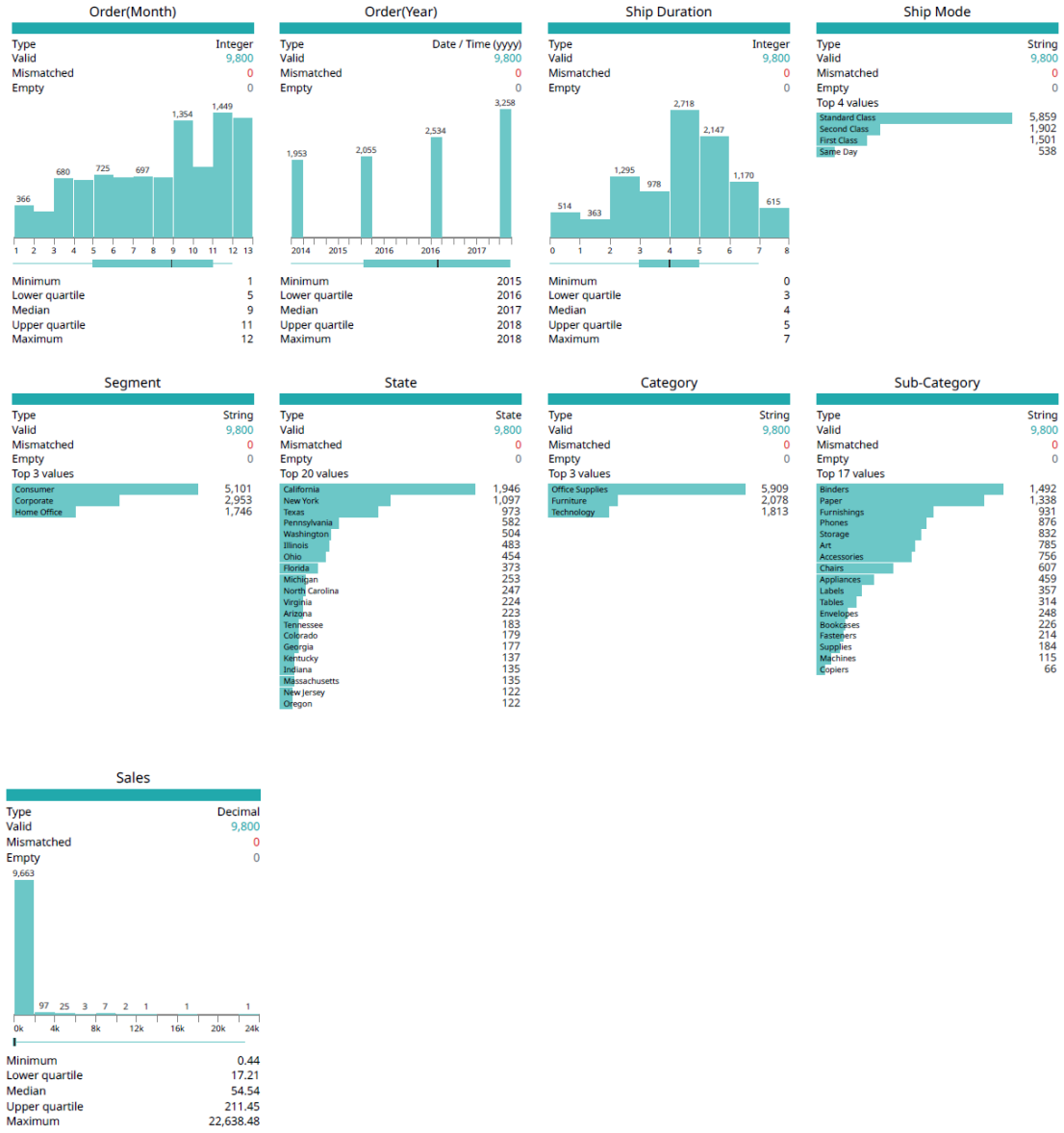

Figure 1.4 DataPrep Profile

The profile report in Figure 1.4 shows 9 columns and 9800 rows with 5 data types as shown above. The Order (Month) column shows data for all 12 months. It shows that the fourth quartile of the year has the highest number of orders, with December (Month 12) having the highest number of orders (1,449) and January (Month 1) having the lowest (366). The Order (Year) column shows an increase in sales from 2015 to 2018, with 2018 having the highest number of sales.

The Ship Duration data shows that most shipments are completed within 4 to 5 days. The minimum value is 3 days, and the maximum is 7 days. Most shipments are delivered on time, with the majority taking less than a week to ship. This suggests efficient logistics and shipping processes. The most common Ship Mode is Standard Class (5,859 orders), followed by Second Class (1,902 orders),

and First Class (1,501 orders). Same Day is the least common with 538 orders. This showed that most customers prefer standard shipping, which may be more cost-effective.

The Segment column shows that the majority of orders are from the Consumer segment (5,101 orders), followed by Home Office (2,593) and Corporate (1,746). This suggests that retail or personal purchases are more prominent than corporate sales. The State column shows that California has the highest number of orders (1,946), followed by New York (1,097), and Texas (972). California is the top contributor to sales likely due to its large population and economic activity.

The Category column shows that Office Supplies is the most popular category (5,909 orders), followed by Furniture (2,078) and Technology (1,813). Office Supplies dominate the sales, likely driven by everyday business and educational needs. In the Sub-Category column, Binders lead with 1,432 orders, followed by Furnishings (1,398) and Paper (1,331). Binders and Furnishings are the top-selling items within their respective categories, indicating that these products are in high demand for office settings, likely for organizational and furniture purposes.

The Sales graph is positively skewed, meaning that most sales are moderate to lower in value. The minimum value of 0.44 indicates that there are some very small sales amounts, which could suggest that some products are being sold in small quantities or at discounted prices. The highest sales are 23,638.48, which might indicate that some large sales transactions could be from corporate clients or large orders. Since sales distribution suggests that most transactions are of moderate value, the company should focus on understanding the needs of regular customers who make smaller purchases and larger clients who place bulk or high-value orders.

In conclusion, we can streamline the data cleaning process, ensuring that the dataset is well-prepared for feature engineering and model training by using Google DataPrep. By following the steps mentioned above, we can transform raw data into a clean and structured format, significantly improving the model's ability to make accurate predictions. This step is critical for achieving reliable results in the subsequent feature engineering and machine learning phases.

Next, we need to enable few APIs such as Vertex AI API, Cloud Storage API and Cloud AutoML Api before using the Google AutoML. In the Vertex AI, we need to upload the dataset and select Tabular data. Then, choose whether your task is Regression or Classification based on the objectives and select the region where the model to be trained. Next, click the Train the Model to initiate the training process.

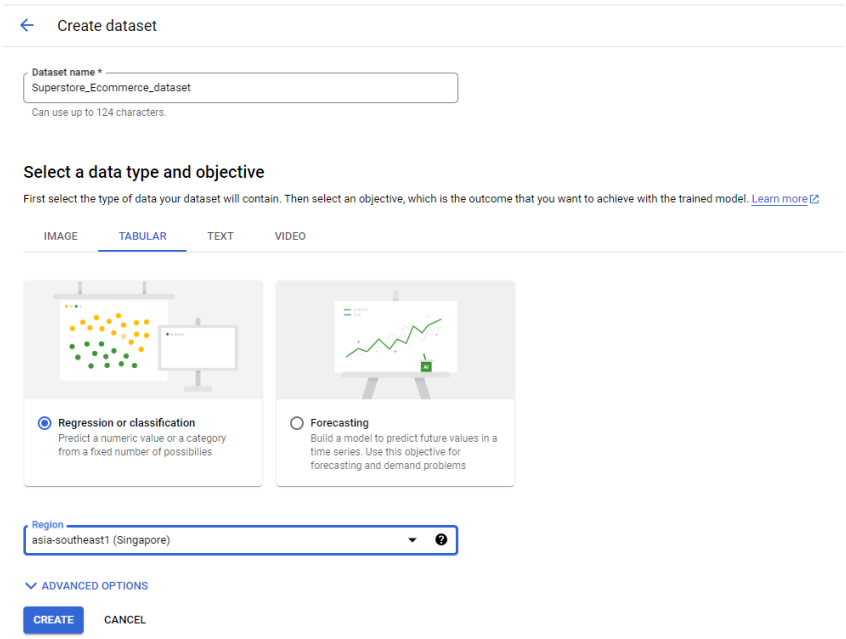

Figure 1.5 AutoML Setup

We must create a pipeline to manage the model's workflow in AutoML. In the training options, specify the Target Column for the model prediction. In the compute and pricing section, the maximum node hours were set to 1 to control the duration of the model training.

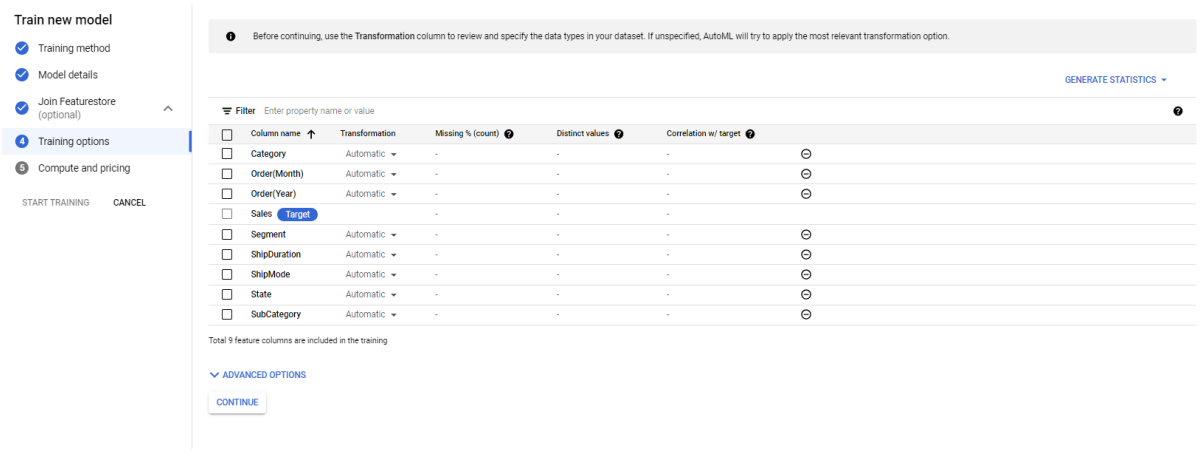

Figure 1.6 AutoML Pipeline Setup

AutoML simplifies the process of building and deploying machine learning models by automating the feature engineering, model selection, and hyperparameter tuning. This lets users quickly create high-quality models and focus on solving business problems. It is an invaluable tool for teams looking to leverage AI without deep technical knowledge. The evaluation metric of Google AutoML is shown in Figure 1.7:

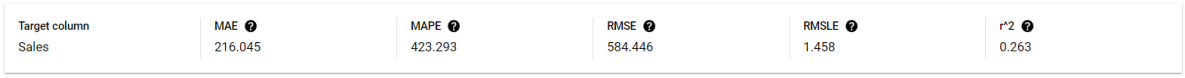

Figure 1.7 Model Performance Evaluation

The machine learning model's performance was evaluated using several key metrics:

i. Mean Absolute Error (MAE): 216.045

MAE measures the average absolute difference between the predicted and actual sales values. The MAE value shown above represents that the average of the model's predictions is off by approximately 216.045 units. This could be a significant error if the target values are smaller than 216.045.

ii. Mean Absolute Percentage Error (MAPE): 423.293

MAPE represents the average percentage difference between the predicted and actual values. A MAPE of 423.293% indicates that the model is making predictions that are, on average, off by over four times the actual value.

iii. Root Mean Squared Error (RMSE): 584.446

RMSE provides a measure of the model's prediction error that emphasis on more significant errors due to squaring. An RMSE of 584.446 indicates a substantial deviation in some predictions.

iv. Root Mean Squared Logarithmic Error (RMSLE): 1.458

RMSLE is the root of squared averages of log differences between observed and predicted values. A value of 1.458 indicates that, on average, the logarithmic difference between predicted and actual values is around this amount. It tends to penalize underestimations slightly more than overestimations, especially for datasets with skewed distributions like sales data.

v. R^2 Score: 0.263

The R^2 is the square of the Pearson correlation coefficient between the observed and predicted values. It ranges from 0 to 1, and a higher value indicates a higher-quality model. A value of 0.263 means that about 26.3% of the variance is in the target variable, which is quite low. This suggests that the model is ineffective in capturing the patterns and relationships in the data.

In conclusion, the high value of RMSE and low R^2 indicate that the model's accuracy is limited and its reliability in making predictions is relatively poor. Improvements to the model, such as better feature engineering, tuning, or using a more suitable algorithm, would be necessary to increase accuracy and reliability.

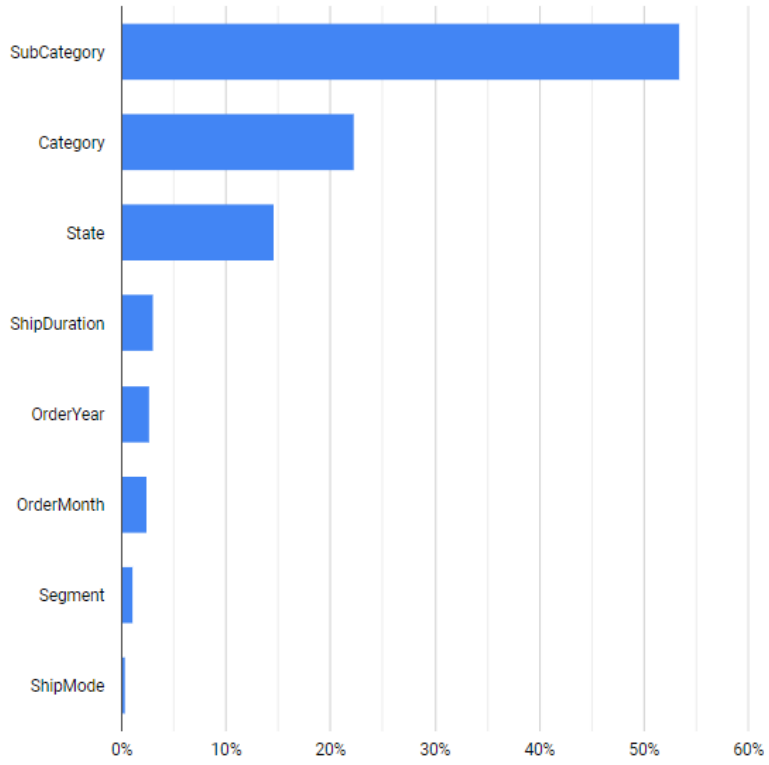

Figure 1.8 Feature Importance Graph

The feature importance graph in Figure 1.8 shows that certain features significantly influence the model's predictions. The Sub Category (53.38%), Category (23.32%) and State (14.62%) are the most important features. This suggests that product classifications play a significant role in sales prediction. The feature importance of the State shows that that different regions might have different preferences. The OrderMonth and OrderYear can help spot seasonal trends and overall changes in sales over time but it does not bring significant importance in the sales prediction. The features with the least importance, such as Ship Mode (0.4%) and Segment (1.08%) are shown to have minimal influence on sales predictions. This suggests that shipping methods and the customer category do not strongly affect the overall sales outcomes and may not be essential for improving model performance.

2 Start Schema Diagram

LucidChart is an online diagramming and visual collaboration tool that allows users to create flowcharts, process diagrams, and other types of visual representations. The Figure 2.1 below shows some of the connectors used in LucidChart when creating a star schema diagram.

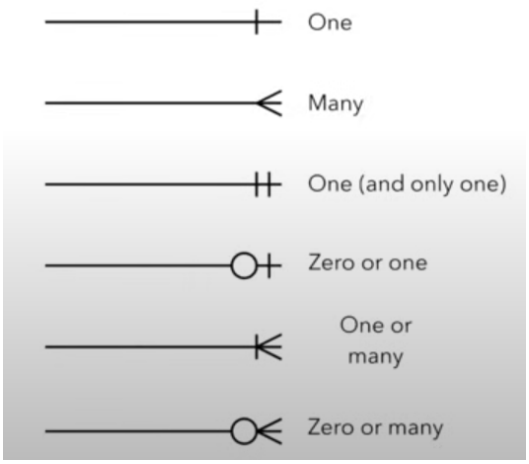

Figure 2.1 LucidChart Connectors

Using LucidChart, Figure 2.2 is created to visualize the star schema of the data before data cleaning to identify the relationships between the entities. Below is an explanation of the connectors used to represent these relationships:

i. Customer to Order Relationship:

A Customer can have zero or many orders. This is because a customer might not have placed any orders but they can place as many orders as needed. For this relationship, we use a zero-or-many connector between Customer and Order.

ii. Order to Customer Relationship:

Each Order is associated with exactly one customer. This means that for an order to exist, it must belong to a single customer. Therefore, the Order can never be linked to more than one customer. To represent this relationship, we use a one-and-only connector between Order and Customer.

iii. Order to Product Relationship:

An Order must include at least one product, but it can also contain multiple products. This showed that an order always involves at least one product, but it may consist of several products depending on the customer's purchase. For this relationship, we use a one-to-many connector between Order and Product.

iv. Product to Order Relationship:

A Product may not necessarily be part of any order but it can be part of many orders. This means a product could exist in the system without being ordered or sold multiple times across different orders. To represent this, we use a zero-or-many connector between Product and Order.

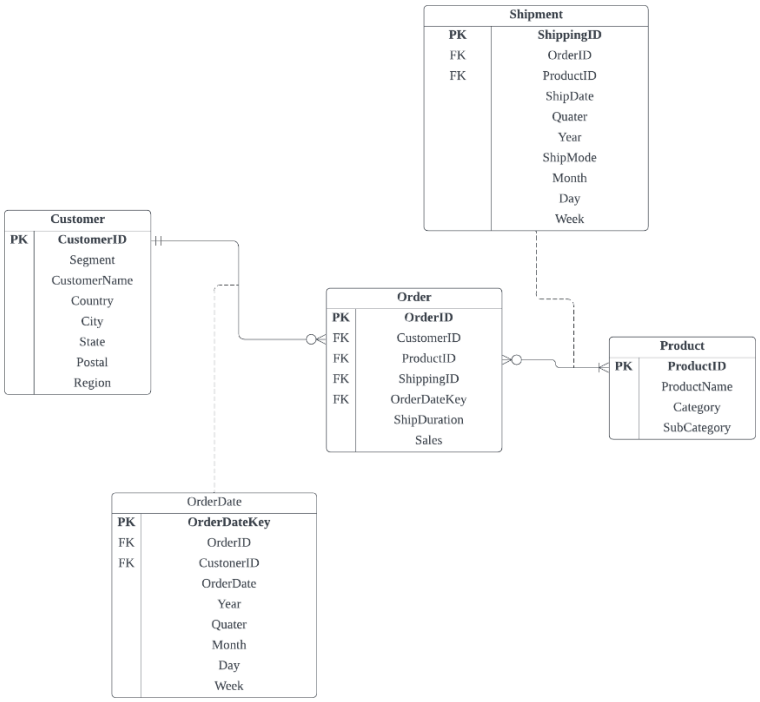

Figure 2.2 Star Schema Diagram Before Data Cleaning

The Star Schema Diagram after data cleaning and preprocessing is shown in Figure 2.3. The description of the fact table and dimensions are as such:

Fact Table

i. Order

Column Name	Description
CustomerID	Unique identifier for Order mode.
CustomerID	Foreign key to the Customer dimension.
ProductID	Foreign key to the Product dimension.
ShippingID	Foreign key to the Shipping dimension.
OrderDateKey	Foreign key to the OrderDate dimension.
ShipDuration	The duration to ship an order after it is ordered.
CategoryID	Foreign key to Category_Dimension.
Sales	The total amount of sales generated for the order.

Dimension Table:

1. Customer

Column Name	Description
CustomerID	Unique identifier for Customer dimension.
Segment	Customer segment (Consumer, Corporate, Home Office)
State	Name of the state

2. Product

Column Name	Description
ProductID	Unique identifier for Product dimension.
Category	The category of the product.
SubCategory	A more specific classification within the product category.

3. Shipment

Column Name	Description
ShipmentID	Unique identifier for Shipment dimension.
OrderID	Foreign key to the Order dimension.
ProductID	Foreign key to the Product dimension.
ShipDate	The date when the order was shipped.
ShipMode	Shipping method (Standard Class, First Class, Second Class, Same Day).

4. OrderDate

Column Name	Description
OrderDateKey	Unique identifier for OrderDate dimension.
OrderID	Foreign key to the Order dimension.
CustomerID	Foreign key to the Customer dimension.
OrderDate	The date when the product is ordered.
Year	The year when the product is ordered.
Month	The month when the product is ordered.

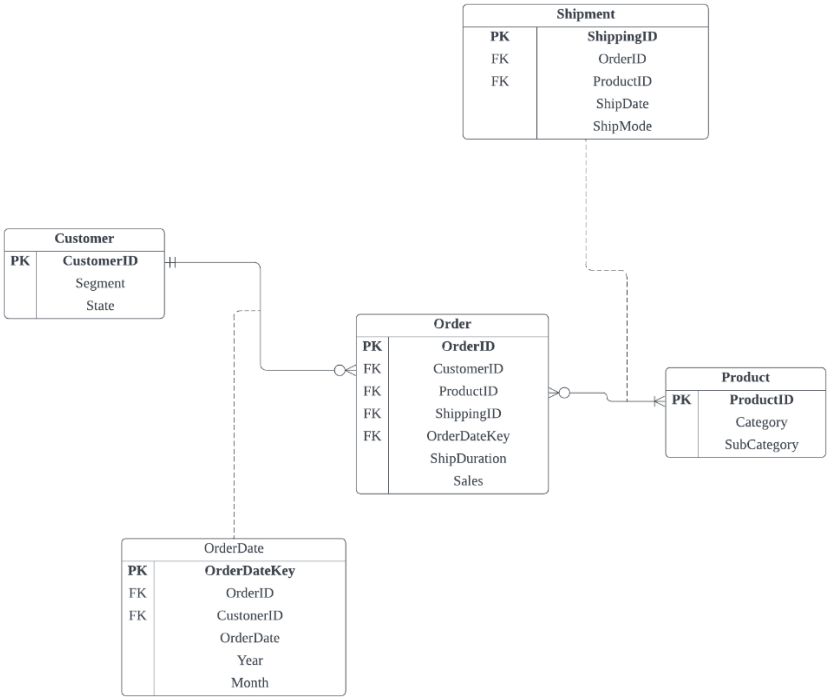

Fig: Star Schema Diagram After Data Cleaning

This star schema diagram is designed to be scalable to allow future integration of additional data sources and columns. For example, new data dimensions such as payment details, customer feedback, and discounts can easily be included in the current structure by adding corresponding foreign keys and attributes to the relevant tables. In addition, the indexing in the fact table can be optimized to handle larger datasets over time to ensure that the schema remains valid as data volume grows.

3 Data Governance and Security Plan

3.1 Security Strategy

The security strategy highlights controlling access to the dataset by applying permission settings to limit the ability to view or modify data based on user roles. This ensures that only authorized personnel can interact with sensitive information. In Google Sheets, specific permissions can be assigned to restrict access based on user roles. For example, my personal email is granted Owner access to provide me with full control over the data. On the other hand, the Editor access is granted to the siswa email to allow them to view and modify the data without the ability to change permission settings as shown in Figure 3.1. This ensures a clear division of responsibilities while maintaining data security. The lecturer or course mate might only have viewer access to prevent them from altering or deleting information.

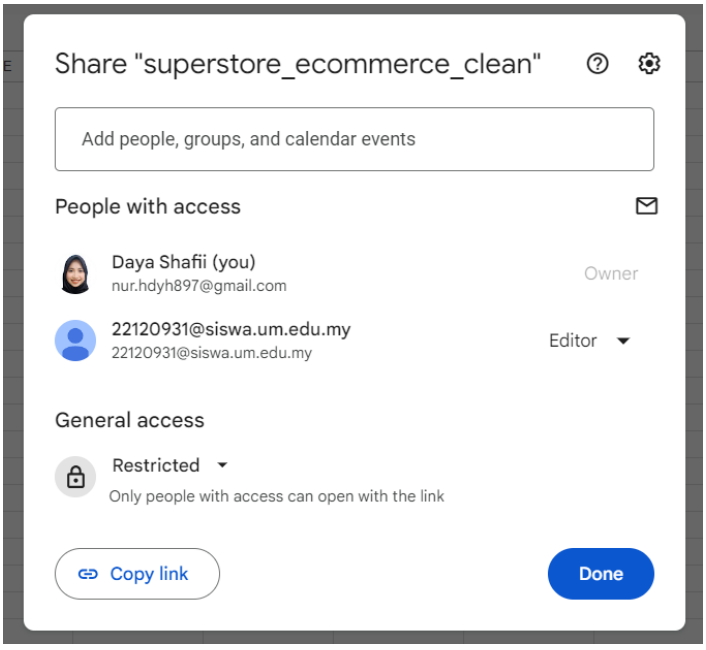

Figure 3.1: Access Control in Google Sheets

In the IAM policy of Google Cloud as illustrated in Figure 3.2, my personal email is assigned the Owner role that grants full control over all resources within the project. On the other hand, the default compute engine service account is assigned multiple roles to ensure proper access for automated tasks and machine learning operations such as:

i. Storage Admin: Allows the service account to manage storage resources that include creating and deleting storage buckets.
ii. Storage Object Admin: Grants the ability to manage objects (such as uploading and deleting files) within the storage buckets.
iii. Storage Object Viewer: The service account can view objects within the storage buckets.
iv. Vertex AI Administrator: Provides full access to Vertex AI resources to enable the service account to manage datasets, train models, and deploy machine learning solutions.
v. Vertex AI User: Enables the service account to use Vertex AI resources to create and interact with models, datasets, and predictions.
vi. Vertex AI Viewer: Allows the service account to view Vertex AI resources but without permission to make modification.

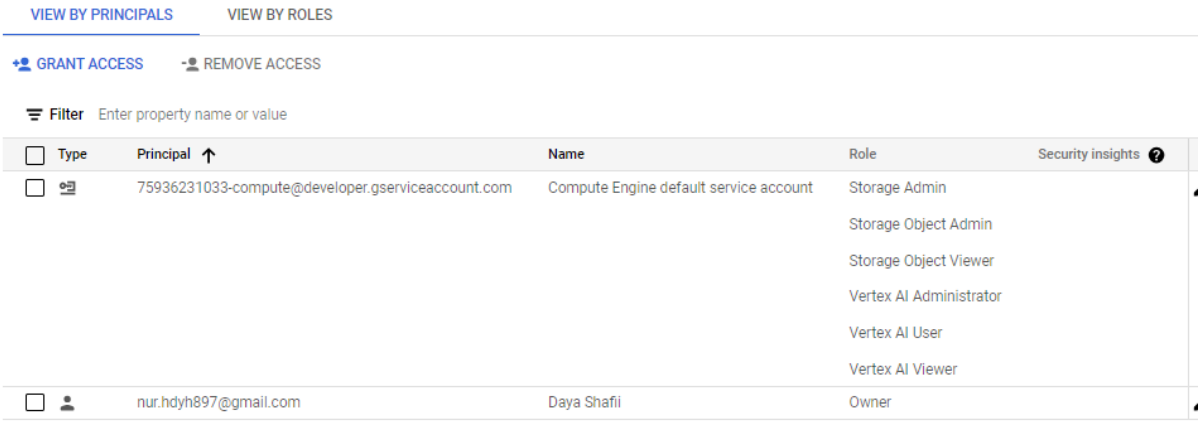

Figure 3.2: Access Control in Google Cloud IAM Policy

3.2 Data Privacy Compliance

In May 2018, the European Union (EU) instituted the General Data Protection Regulation (GDPR) to protect the privacy and personal data of all individuals. In 2020, the California Consumer Privacy Act (CCPA) was implemented in an effort to safeguard Californian residents' privacy, guarantee data transparency, and empower consumers to manage their personal information. While GDPR emphasizes permission before data collection, CCPA focuses on the individual's choice to opt up later.

To ensure that sensitive personal customer data is rigorously controlled, role-based access mechanisms are implemented. Only authorized users can access the sensitive information. We minimise data and only collect the information needed to comply with the law. In an additional effort to guarantee accountability and transparency, access to personal data is recorded and monitored. Since the data is from the United States, the CCPA would be applies if the data involves California residents. California residents have the right to request the deletion of their data or access to it per regulatory requirements. Lastly, data retention policies are implemented to prevent the data from being kept longer.

4 No-Code/Low-Code ETL Pipeline

Power Automate Visual in Power BI enables us to automate workflows in order to optimize a variety of processes. For example, the automation workflow shown in Figure 5.1 is designed to monitor shipping modes that may experience potential delays. Once a delay is identified, the system automatically sends an email notification to the affected consumer. This ensures that the consumer is promptly informed of any shipping issues. Through this approach, organizations can enhance customer satisfaction and communication. Additionally, organisation may reduce manual intervention, optimize operational efficiency, and ensure clients receive timely updates regarding their orders.

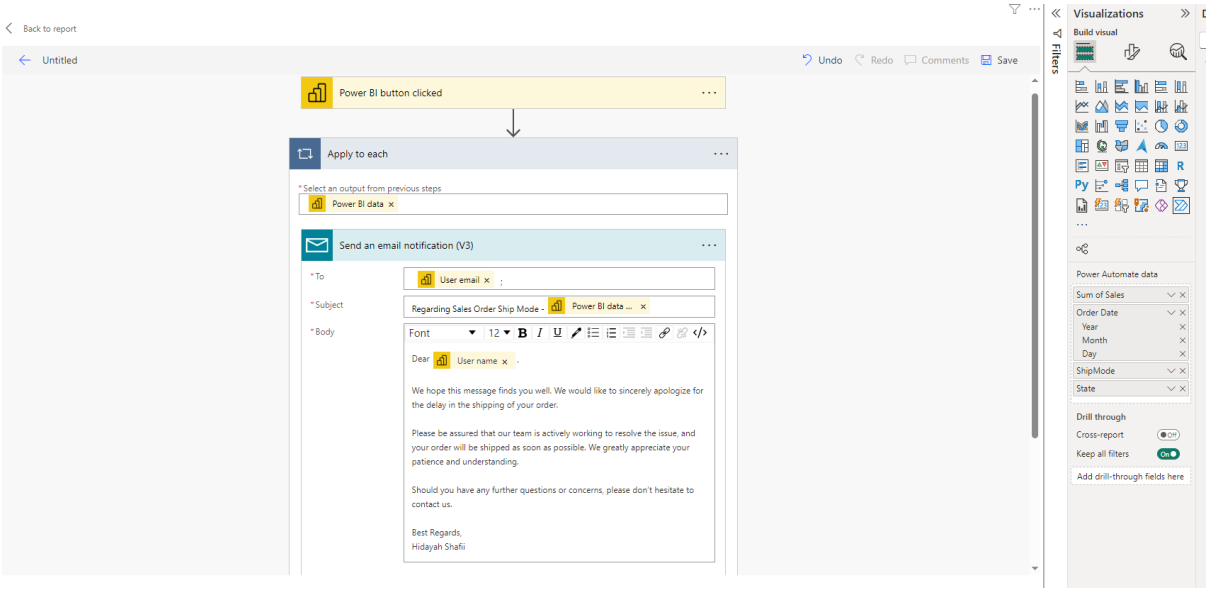

Figure 4.1 Power Automate visual in Power BI

5 Data Visualization and Dashboard

The data is connected to Google Sheets through the use of Power BI in this process, as illustrated in Figure 5.1, to facilitate real-time updates and seamless integration The data is represented in interactive dashboards in Figure 5.2, which enable users to monitor and analyze critical metrics, including total sales, sales by region, product category, and consumer segment. With dynamic visualisation, the stakeholders can quickly identify any potential issues that may affect sales and enhance the overall business operations

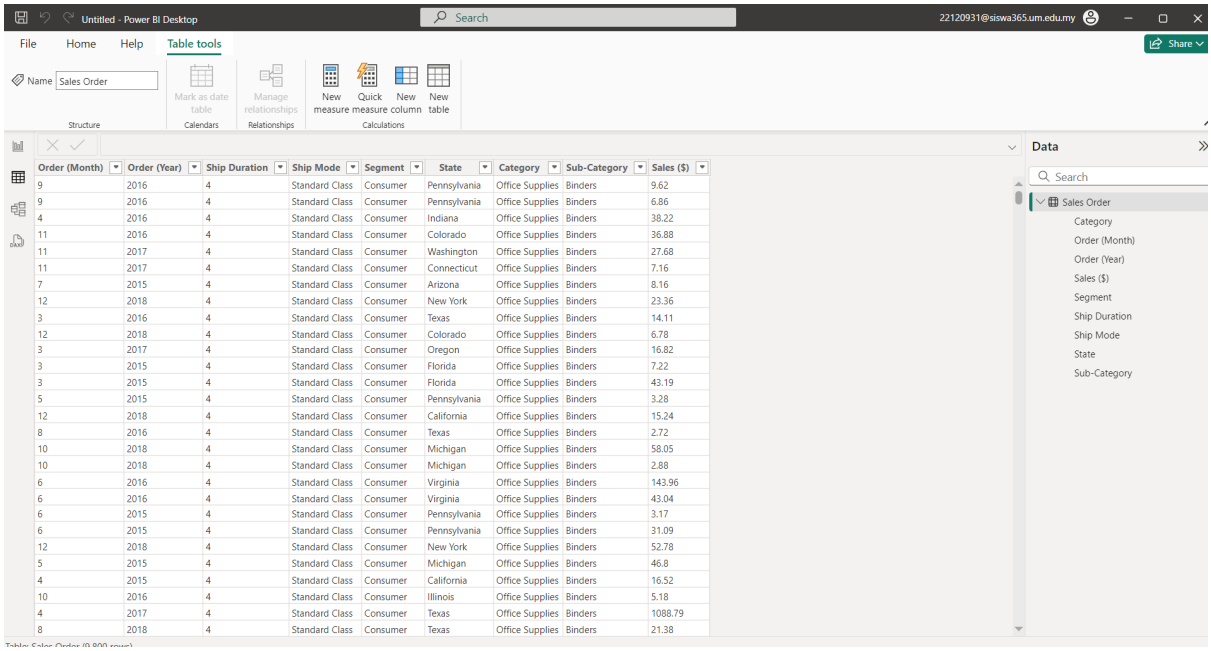

Figure 5.1 PowerBI Overview

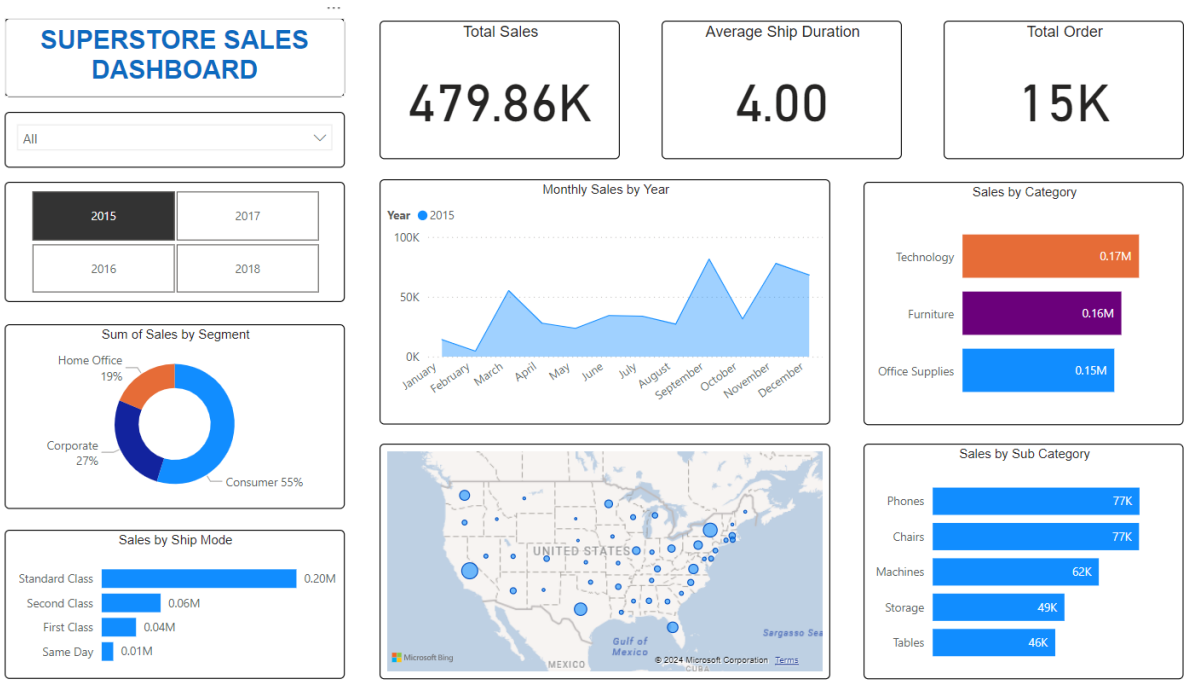

Figure 5.2 Superstore Sales Dashboard

6 Reflection on Modern Technologies

The integration of Lucidchart, Google Cloud Platform, and Power BI has changed data visualization and automation in data science practices. These technologies are crucial to data analysis, machine learning, and visualization efficiency.

Google AutoML lets enterprises use complex algorithms without extensive machine learning knowledge or infrastructure. Integration with other Google services benefits data scientists and analysts. DataPrep features in Google Cloud Platform make data preprocessing user-friendly, especially when it has an interface similar to PowerBI. The integration between Dataprep and Google AutoML allows for a more efficient pipeline. Furthermore, Using Google Sheets as the data source avoided the need for manual updates and promoted a collaborative environment. Furthermore, Lucidchart also helped create and visualize data flow and schema. It organised and streamlined the data structure for Power BI and Google Sheets interaction.This visual clarity helps stakeholders grasp complex data relationships and workflows

These modern technologies provide high data processing and visualization automation, which is essential in modern data processes. For example, the automated data preprocessing in Dataprep can be followed by model training in Google Cloud ML, with results visualized in Power BI. Besides that, we can also set triggers with Power Automate in PowerBI to automate workflows. This end-to-end automation reduces manual intervention, minimizes errors, and accelerates the overall data analysis process.

Modern technology requires effective cost management for organisations to maximize financial resources and utilise their full potential. By frequently auditing the free trial version of cloud expenditures in the Google Cloud Platform, we minimized costs while still being able to use cloud services like Dataprep and Vertex AI. This approach allowed us to conduct data preprocessing and predictive analysis without incurring significant expenses. Nevertheless, it is advisable to transition to a pay-as-you-go plan that is cost-effective in the long term. Distributing workloads among several cloud providers helps an organization choose the most cost-effective services for specific jobs.

In conclusion, these modern technologies showed the importance of integrated tools in resolving challenging business issues. The experience has helped me get better at using technology to analyze data and automate work processes.