DALHOUSIE UNIVERSITY

Faculty of Computer Science

CSCI 5408 Data Warehousing, Management and Analytics

Sales Data Case Study

March 24, 2020

Case Study Group-4

Sasidharan Srikrishna	B00835818
Bhalerao Ketan	B00839791
Rachakonda Sumith Sai	B00851825

Table of Contents

1. Introduction……………………………………………………………………………..3
2. The concept of data warehouse…………………………………………………………3
3. A closer look: Sales dataset……………………………………………………………..3
4. Schema design…………………………………………………………………..........5
5. Cognos BI visualizations and Analytics………………………………………………...7
6. Future predictions through neural networks…………………………………………….13

1. Introduction

The business today requires Internet as a primary medium for communication. A lot of business transactions are performed over Internet which has set up a new challenge of analyzing customers in a deeper fashion. However, the incoming data from Internet transactions has to be stored in a structured way so that the future steps of data analysis can easily fetch the required data. Thus, all the incoming data pieces need to be merged somewhere in order to complete the overall insight puzzle, which provides the ground for **"data warehouses".** Thus, a data warehouse is analogous to a big data storage system which stores all the past data related to every process domain of an organization. It can be centralized as well as distributed. A data warehouse can be effectively utilized for improving the future steps of business analytics as it connects all the data pieces from different sources. Section 2 discusses the concept of data warehouse in detail.

2. The concept of data warehouse

A data warehouse is a data management system designed to support business intelligence and business analytics and enable them. On huge amounts of historical data, we perform analytics and queries by using data warehouses. Typically, it is used for connection and analysis of business data obtained from different data sources. In this process the structured data from various sources is aggregated such that it is useful for comparison and analysis of business intelligence. Moreover, a data warehouse contains large amounts of data from various multiple sources. Decision making for organizations can be achieved by using this data warehouse. Gradually, Historical records are built from time to time. These records are very much useful for data scientists & business analysts. In a Business Intelligence (BI) system, the data warehouse is considered as the core. This core material is used for analysis and reporting of data. In an organization, operational database and data warehouse is maintained individually. We use data warehouse for making designs that helps in response time reduction and query performance enhancer for data analysis and data reporting.

Data warehouses maintain a central repository, in which the data can be structured/semi-structured/unstructured. Users may use this data with BI tools/spreadsheets/SQL clients because this data is processed and can be used with these tools. This leads to data mining, which is a process of finding patterns and knowledge in data. Thus, data warehouses are extremely essential as a first step to storing data in a structured manner.

3. A closer look: Sales dataset

The given dataset is a sample of a much larger dataset of Sales originally written by María Carina Roldán, who is a BI consultant at Assert Solutions in Argentina [1]. It is licensed under the Creative Commons Attribution-Non-commercial-Share Alike 3.0 Unreported License [1]. It was modified by Gus Segura in June 2014[1].

Dataset File Name: sales_data_sample.csv
Number of rows: 2823
Number of columns: 25

Dataset Columns

1) ORDERNUMBER: It holds the identification of orders in a unique fashion. However, in the given dataset, the numbers are getting repeated. It is present in all rows. It is numeric.

2) QUANTITYORDERED: It holds the count of same products ordered in a particular order. It is numeric. It is also present in all rows.
3) PRICEEACH: This column holds the price of single quantity of a product. It is numeric. It is also present in all rows.
4) ORDERLINENUMBER*:* This column holds the id of order-lines of organization. It is numeric. It is also present in all rows.
5) SALES: It is *PRICEEACH * QUANTITYORDERED*. Several values are incorrect and need to be cleaned. It is numeric. It is also present in all rows.
6) ORDERDATE: Stores the date on which the product was ordered. It is descriptive. It is also present in all rows.
7) STATUS: Stores the status of order. It is descriptive. It is also present in all rows.
8) QTR_Id: Stores the Quarter ID from 1 to 4. It is descriptive. It is also present in all rows. A quarter is a three-month period on a company's financial calendar that acts as a basis for periodic financial reports and the paying of dividends [2]. 1 means (Jan, Feb, and March). 2 means (April, May, June). 3 means (July, August, September). 4 means (October, November, December).
9) MONTH_ID: Stores months from 1 to 12. It is numeric but is actually descriptive in nature since every number represents a month. It is also present in all rows.
10) YEAR_ID: Stores the year of the order. It is numeric but is actually descriptive in nature since every number represents a month. It is also present in all rows.
11) PRODUCTLINE*:* Stores the product-line name. One product-line can have several products. However, one product always belongs to a single product line. It is descriptive. It is also present in all rows.
12) MSRP: Stores the Manufacturer Suggested Retail Price (MSRP). It is numeric and present in all rows.
13) PRODUCTCODE: Stores primary key for Products. It is alpha-numeric and present in all rows.
14) CUSTOMERNAME: Stores the name of the organization who ordered the product.
15) PHONE: Stores phone number of the customer. It is descriptive and present in all rows. The phone numbers are in different formats and are needed to be cleaned.
16) ADDRESSLINE1*:* Stores the first part of address of customers. It is descriptive. It is present in all rows.
17) ADDRESSLINE2: Stores the second part of the customer's address. It is descriptive. It is absent in certain rows.
18) CITY: Stores the city name of the customer. It is descriptive. It is present in all rows.
19) STATE: Stores the state name of the customer. It is descriptive. It is absent in certain rows.
20) POSTALCODE*:* Stores the postal code of area. It is numeric. It is absent in certain rows.
21) COUNTRY: Stores the country name of customer. It is descriptive. It is present in all rows.
22) TERRITORY: Stores the territories in short forms. EMEA means Europe Middle East and Africa. APAC means Asia Pacific Region. NA is North America. Japan needs to be replaced with APAC since it is a part of Asia.
23) CONTACTLASTNAME*:* Last Name of customer who will act as a contact point for accepting the product. It is descriptive. It is present in all rows.
24) CONTACTFIRSTNAME*:* First Name of customer who will act as a contact point for accepting the product. It is descriptive. It is present in all rows.
25) DEALSIZE*:* Holds 3 values: LARGE, MEDIUM and SMALL. It is descriptive and present in all columns.

4. Schema design

Sale- Fact table

The given dataset reports sales record of vehicles such as (Car, Trucks, Planes and Ships, etc.). From the dataset we could infer that a sale which is placed may include following facts to identify each sale record.

- PRODUCTCODE
- QUANTITY (Total number of products ordered)
- PRICEEACH (Price of each product)
- SALES(Total amount of this sale)
- ORDERNUMBER

These attributes are the key performance indicators for the sale during 2003-2005. While exploring the dataset, we figured that a sale can have multiple dimensions like Order, SalePeriod, Product, DealType, Customer, State, Country and Territory. The fact table will have foreign keys of all these dimensions as attributes.

Dimension tables

[1] Order- Dimension Table

As a part of data analysis, a business may consider tracking number of shipped orders, number of in process orders or number of on hold sales. All these factors help in understanding the business better. An order entity will have attributes like Order number, and order status. Each order will have one sale record (i.e. order has one to one relationship with the sale fact table).

[2] Sale Period- Dimension Table

The sale data warehouse is a central repository from where sales of a company can be analyzed based on year, month, or quarter of a year. Analyzing each of these report gives the business new insights on the current year's (or month's or quarter's) profit. This helps the statistician to perform prediction analysis of how the business will grow in next quarter or year. Sale Period dimension table will have attributes like order number, order date, order month, year, and the quarter. Each sale fact table has one to one cardinality with the SalePeriod table.

[3] Product- Dimension Table

Each product of purchase has a product line. If in any case, a company wants to identify which product line has made highest sales this year then this dimensional table can be used to analyze the sales. Each Product entity has attributes such as product line, MSRP (Maximum sale retail price) and the product code. The sale fact table has one to one cardinality with the product dimension table.

[4] DealType- Dimension Table

Based on size and amount of purchase each sale is categorized as small, medium and large. Suppose a business wants to look the number of small sales it made in the year 2004. In this case having a separate dimension for deal type will help retrieve the records efficiently. Each sale can have one deal type.

[5] Customer- Dimension Table

Let’s consider a scenario where a business wants to know which client has made the maximum number of sales. This can be done by analyzing a dataset having separate dimension for client will be helpful. A Customer entity will have attributes like Customer Name, customer Id, Phone, Address Line1, Address Line2, Postal code, City, state id, and Contact person first name and last name. Each customer can have one or more order which means customer dimension table can have one to many relationships with the fact table.

[6] State- Dimension Table

Since one state can have multiple client location, this dimension table will have one to many relationships with customer address entity. By connecting two or more dimensions to one another a snow-flake schema approach is followed. State table has attributes such as state name, state id and country id.

[7] Country- Dimension Table

A Country entity table will have attributes like country name, country id, and territory id. One country can have many states with client locations.

[8] Territory- Dimension Table

A Territory comprises of a list of countries which are grouped under certain norms. So, any analysis based on territories will be easier if it has a separate dimension.

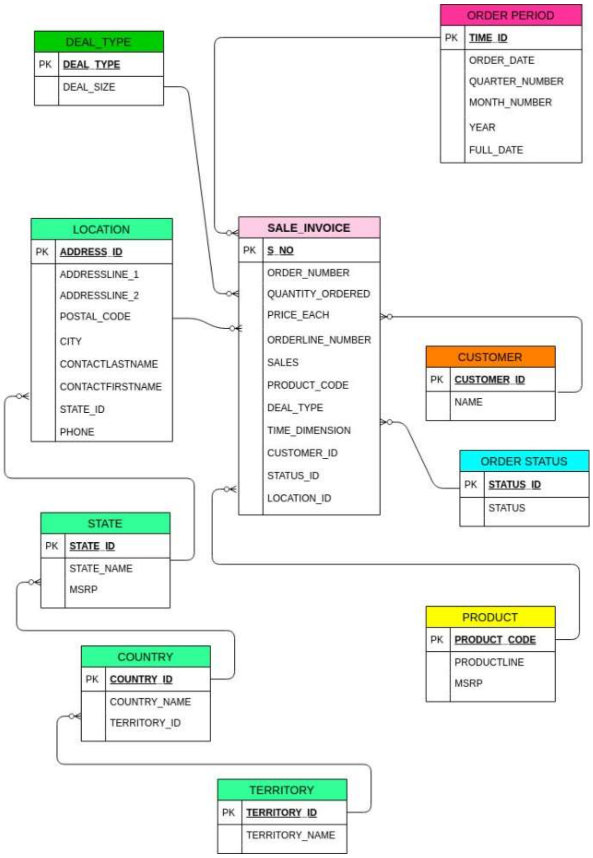
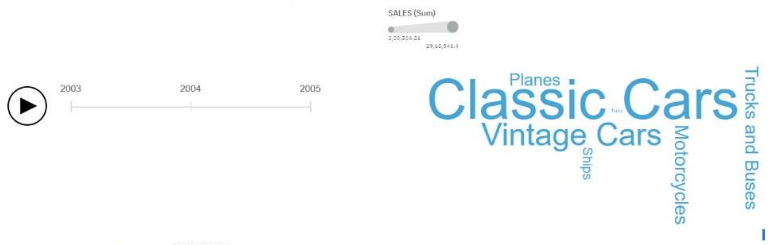


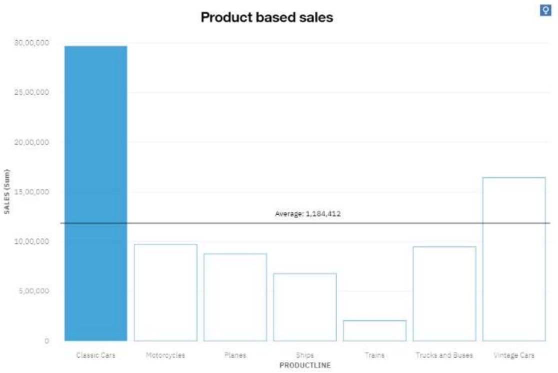
Fig.1 the Snowflake schema of sales data warehouse

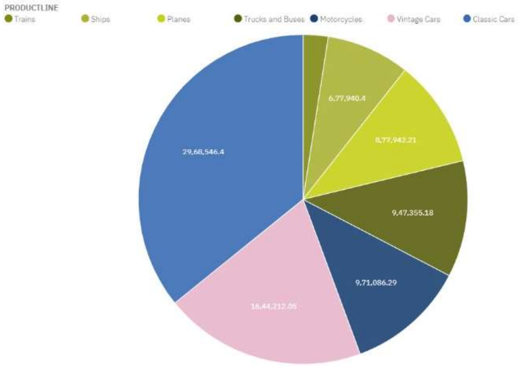
5. Cognos BI visualizations and analytics

Product-based profit

- **Query 2**: To determine the most profitable product lines.
- **Insights**: Classic cars and vintage cars made the maximum profit. And Trains are appearing in minimal orders.

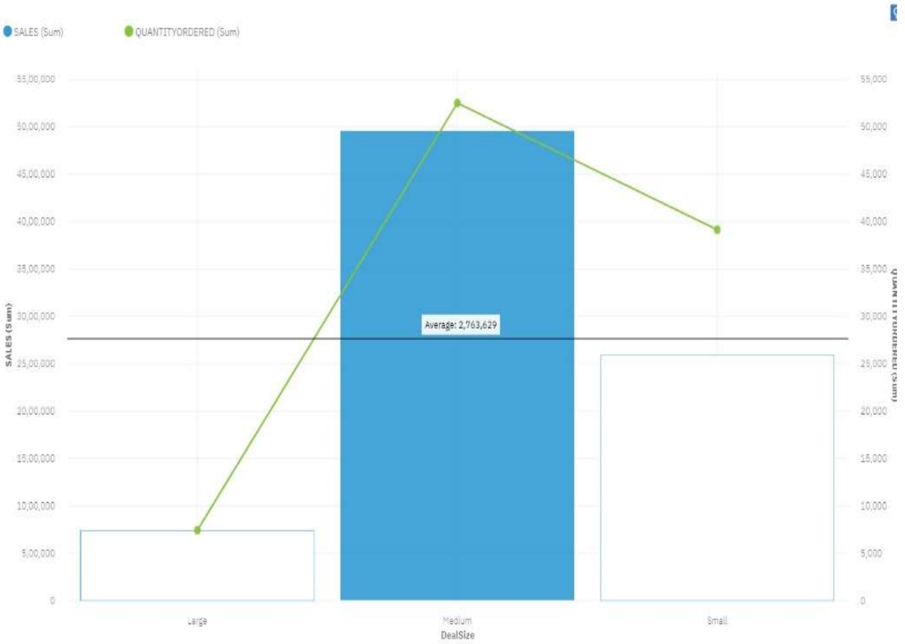






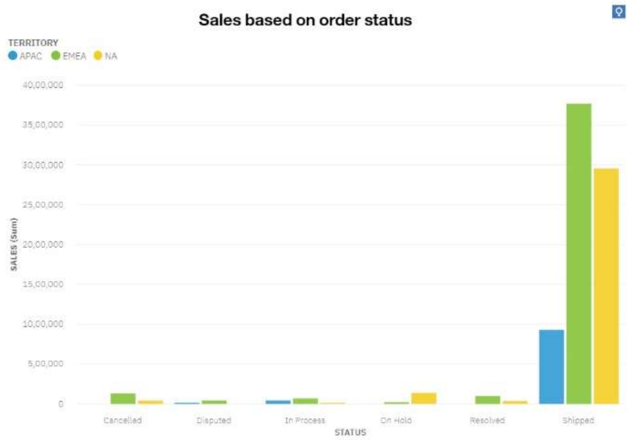
Sale based on deal size

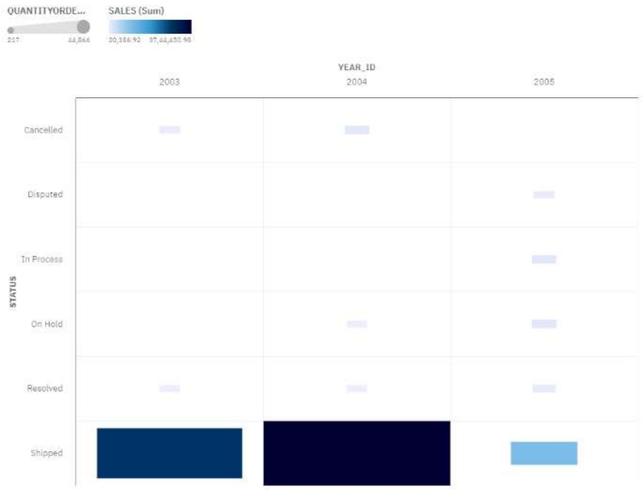
- **Query 3**: To determine sales with respect to deal size.
- **Insights**: Majority of orders are making Medium deals and have sold 5929 products. More focus should be given over Large deals which sold only 916 products.



Sale based on order status

- **Query 4**: To determine sales based on order status.
- **Insights**: In QTR_4, the majority of products were successfully shipped. Compared to 2003 and 2004 more orders were in the hold as well as the disputed state in 2005. This is one of the reasons for decline in sales for 2005.



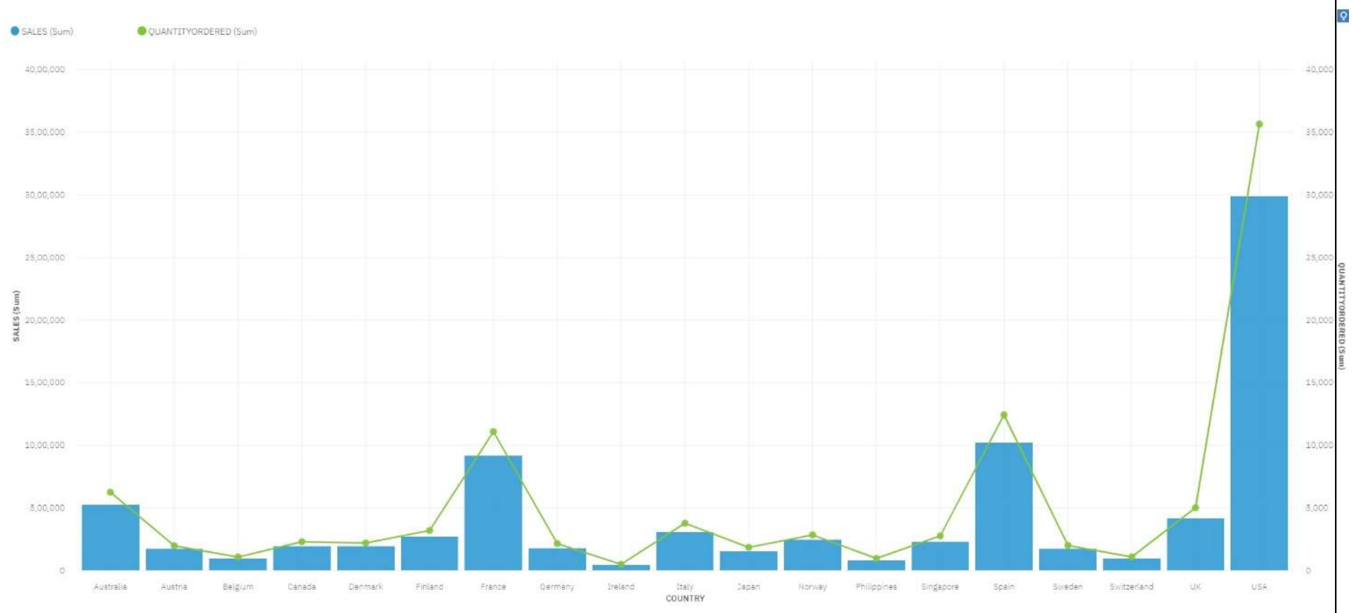


307
order count

8.29M
SALES

Sales as per country

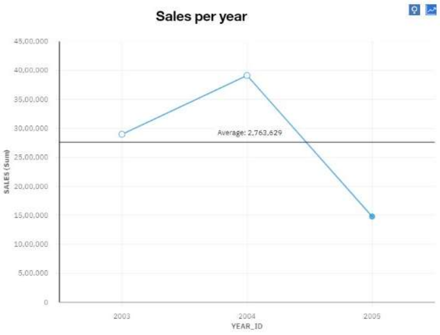
- **Query 5**: To determine sales as per country.

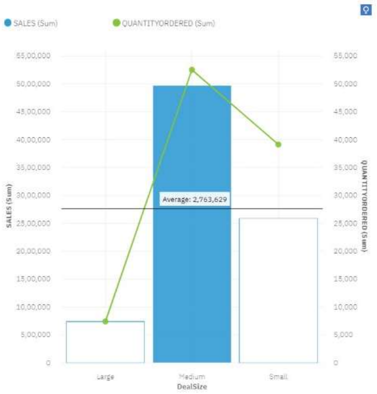


Sales per year

- **Query 1:** To determine sales made based on the year.
- **Insights:** From 2003 – 2004 the profit from sales hiked up. But the profit dropped sharply between 2004 – 2005.

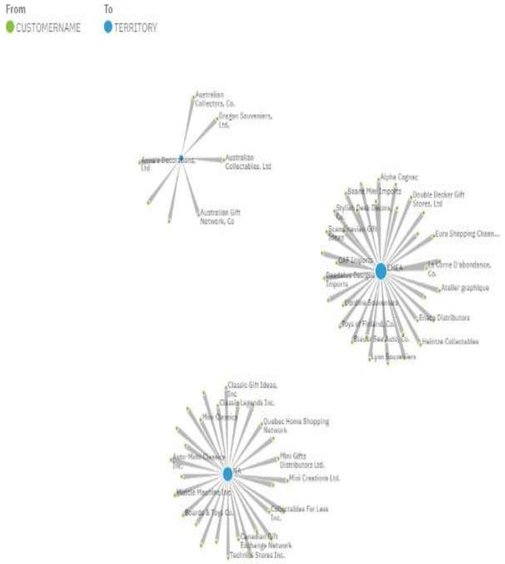
  Although there were ups and downs in the sales graph, the fourth quarter always showed maximum sales.

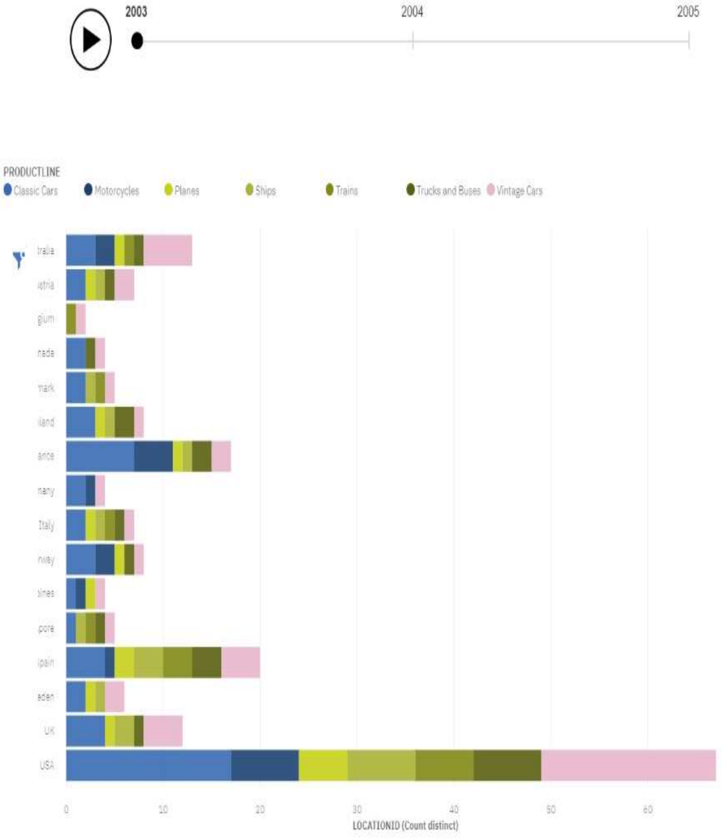




Sales with respect to country

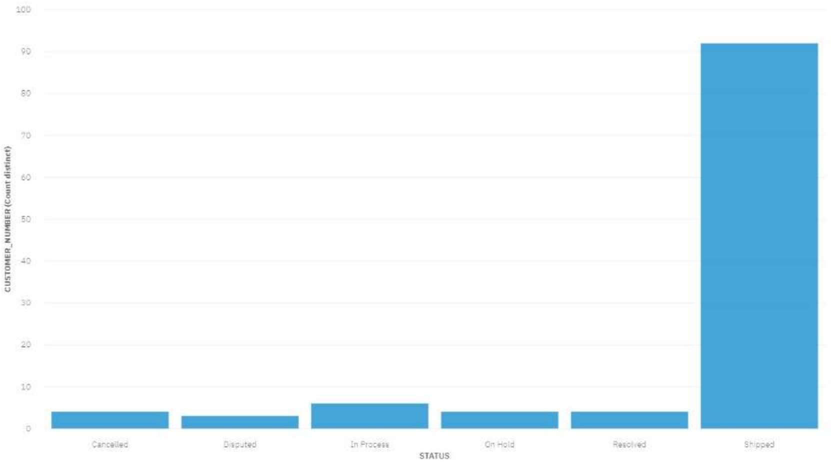
- **Query 7**: To determine sales with respect to country
- **Insights**: Keep maximum products of all product-lines in USA. Reduce the product stock in Switzerland. Keep uniform distribution of products in other countries.





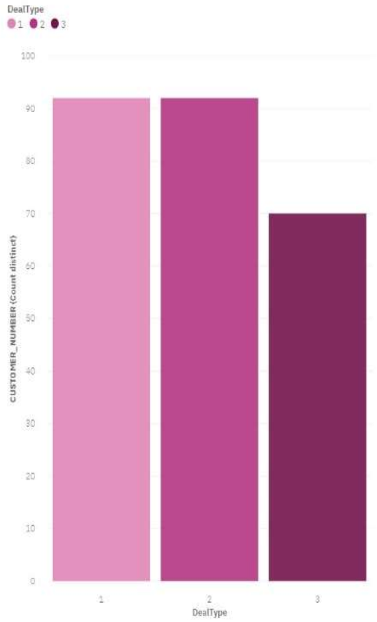
Analysing Status with respect to Customers

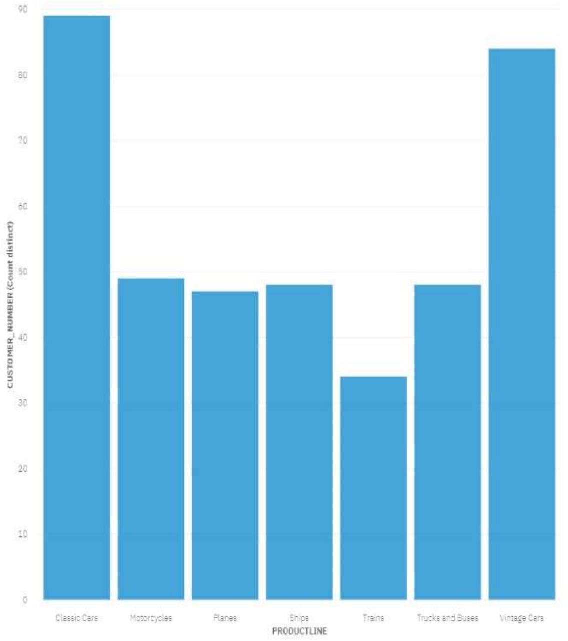
- **Query 8**: Analysing Status with respect to Customers
- **Insights**: Majority of orders were shipped successfully. In 2005, customers faced maximum number of disputes, holds and pending orders.

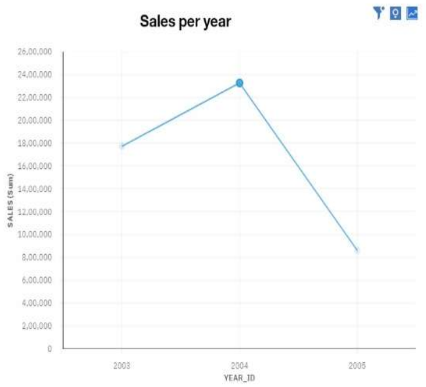


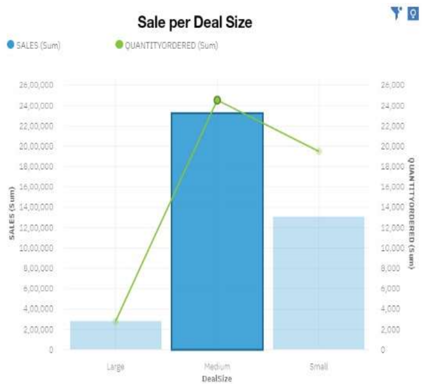
Customer orders with respect to Deal size

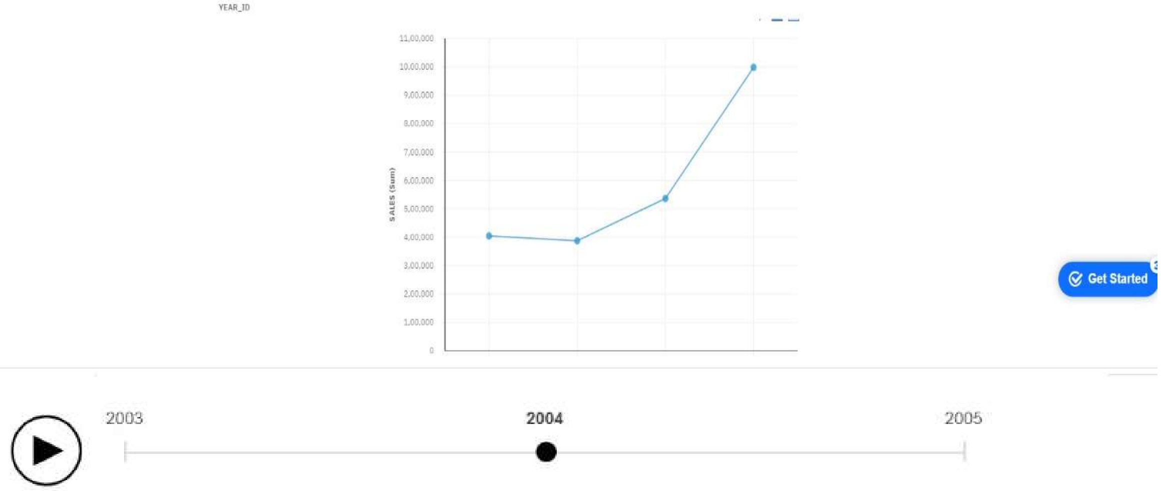
- **Query 6**: To determine the count of deal sizes with respect to customers.
- **Insights**: 90 customers have engaged in small deals and 92 customers have made medium deals.

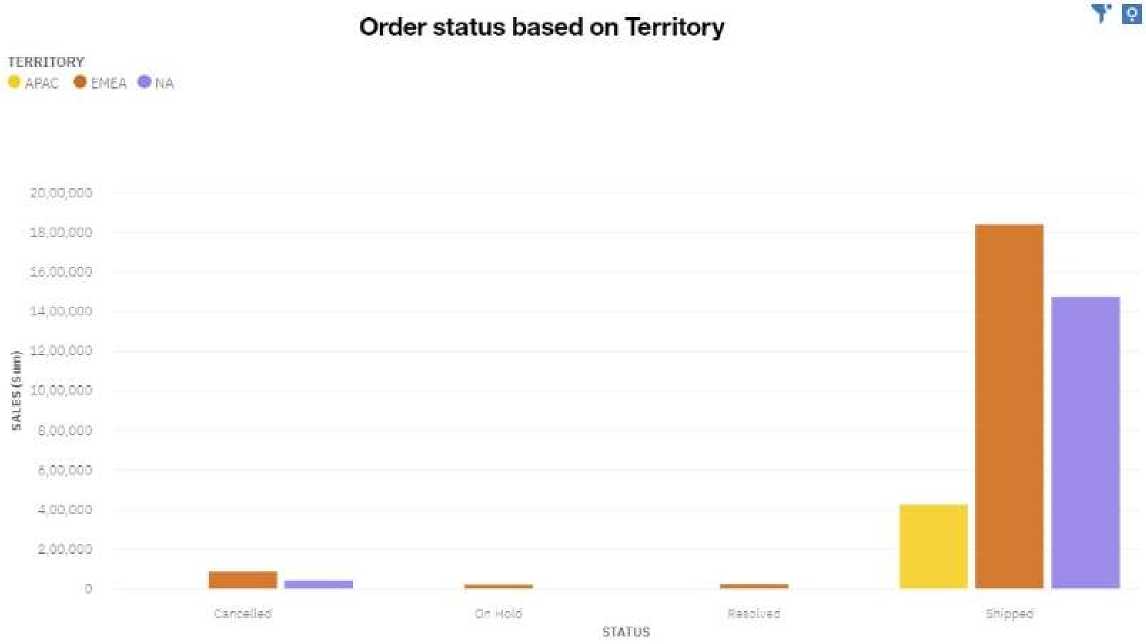




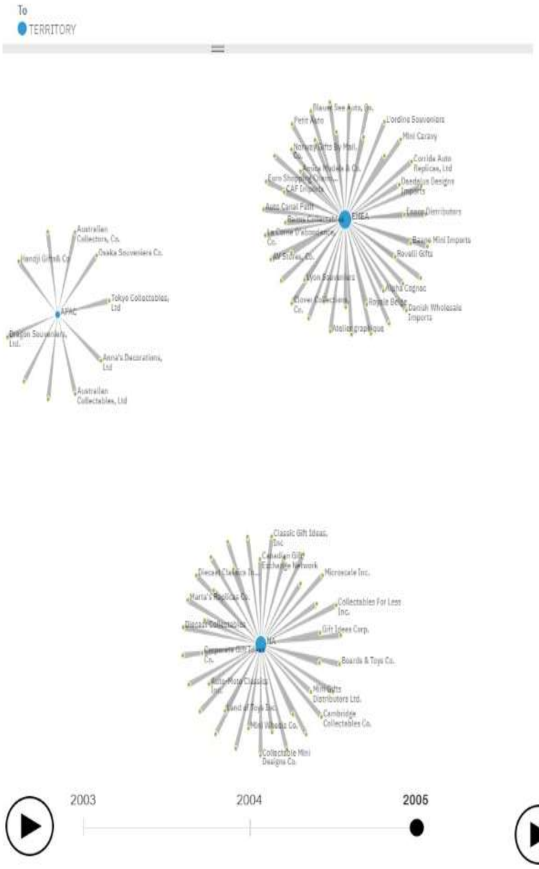

Sales per year
26,00,000
24,00,000
22,00,000
20,00,000
18,00,000
16,00,000
14,00,000
12,00,000
10,00,000
8,00,000
6,00,000
4,00,000
2,00,000
0
SALES (Sum)
2003
2004
2005
YEAR_ID


Sale per Deal Size
● SALES (Sum)
● QUANTITYORDERED (Sum)
26,00,000
24,00,000
22,00,000
20,00,000
18,00,000
16,00,000
14,00,000
12,00,000
10,00,000
8,00,000
6,00,000
4,00,000
2,00,000
0
SALES (Sum)
26,000
24,000
22,000
20,000
18,000
16,000
14,000
12,000
10,000
8,000
6,000
4,000
2,000
0
QUANTITYORDERED (Sum)
Large
Medium
Small
DealSize

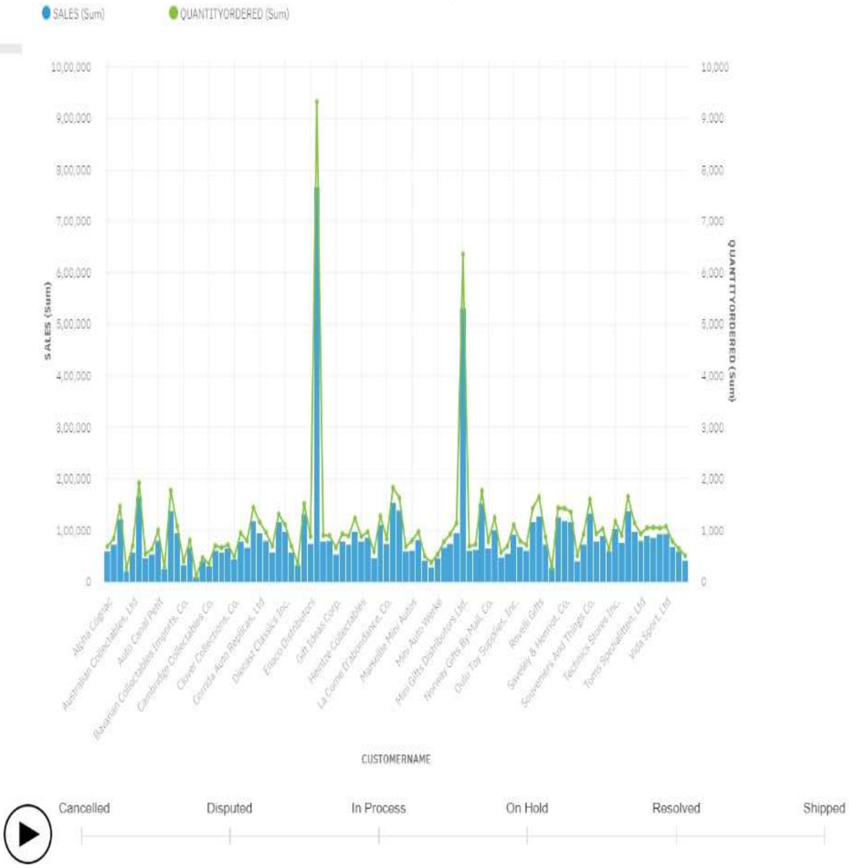

11,00,000
10,00,000
9,00,000
8,00,000
7,00,000
6,00,000
5,00,000
4,00,000
3,00,000
2,00,000
1,00,000
0
SALES (Sum)
Get Started
2003
2004
2005

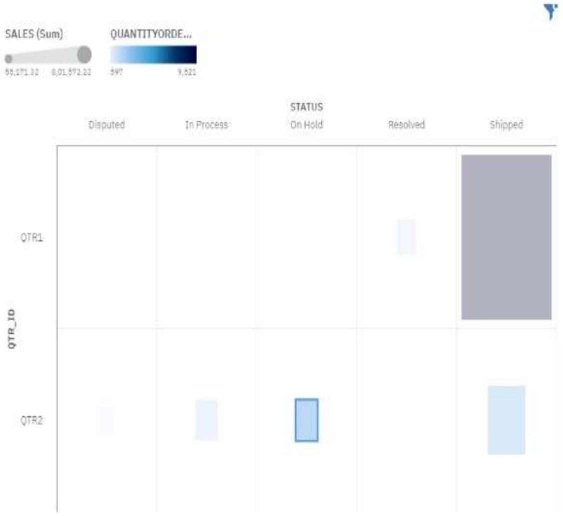

Order status based on Territory
TERRITORY
● APAC ● EMEA ● NA
20,00,000
18,00,000
16,00,000
14,00,000
12,00,000
10,00,000
8,00,000
6,00,000
4,00,000
2,00,000
0
SALES (Sum)
Cancelled
On Hold
Resolved
Shipped
STATUS

Region based customer networks



Customer Sales and Quantity ordered





Sales per year

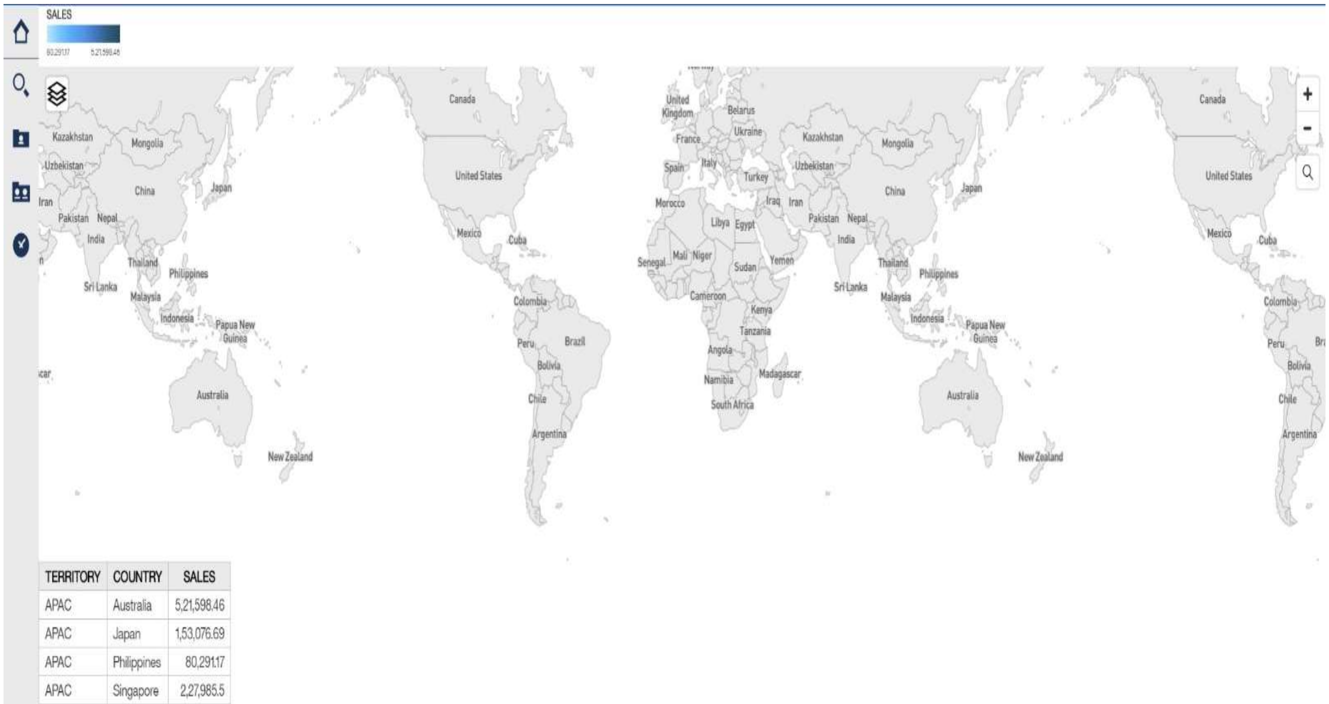
139K

SALES

Total Orders

3

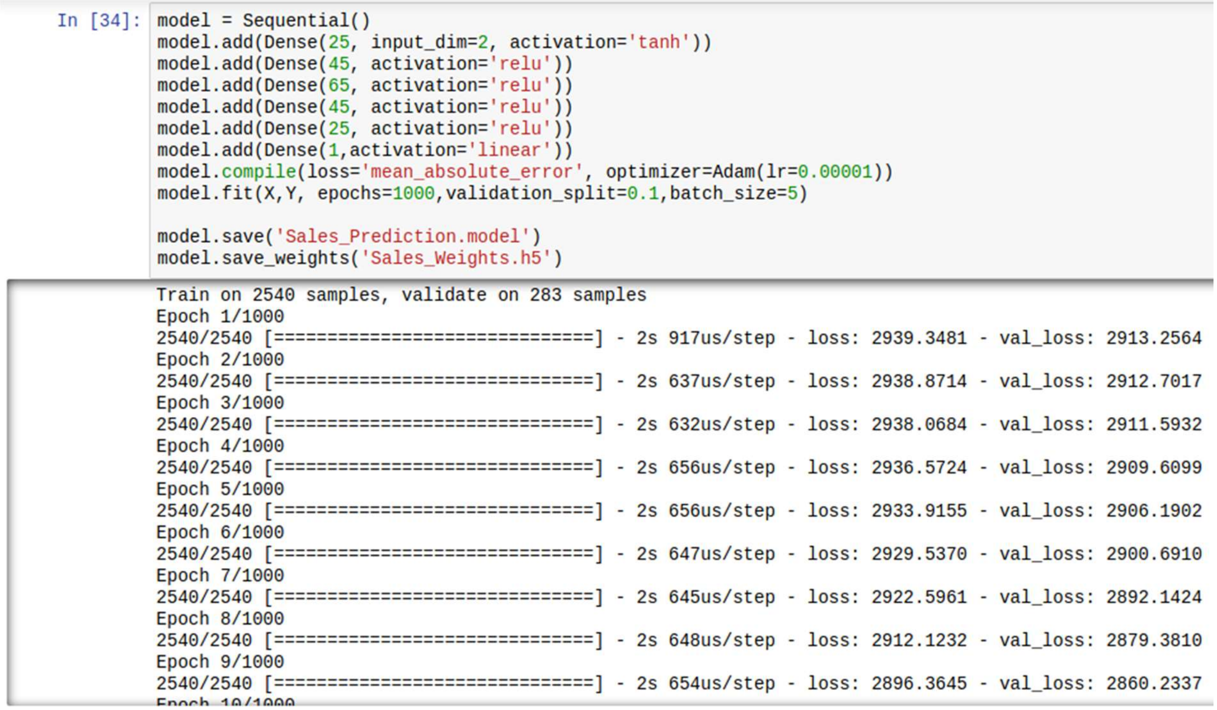
count_ORDERNUMBER



TERRITORY	COUNTRY	SALES
APAC	Australia	5,21,598.46
APAC	Japan	1,53,076.69
APAC	Philippines	80,291.17
APAC	Singapore	2,27,985.5

6. Future Predictions through neural networks

We have trained a neural network in order to predict sales for 2020-2022 (upcoming 3 years). It was trained on **Tensorflow** (Machine learning library of Google). The prediction error is -30 to +30 dollars. The training time was approximately 45 minutes over the local machine.



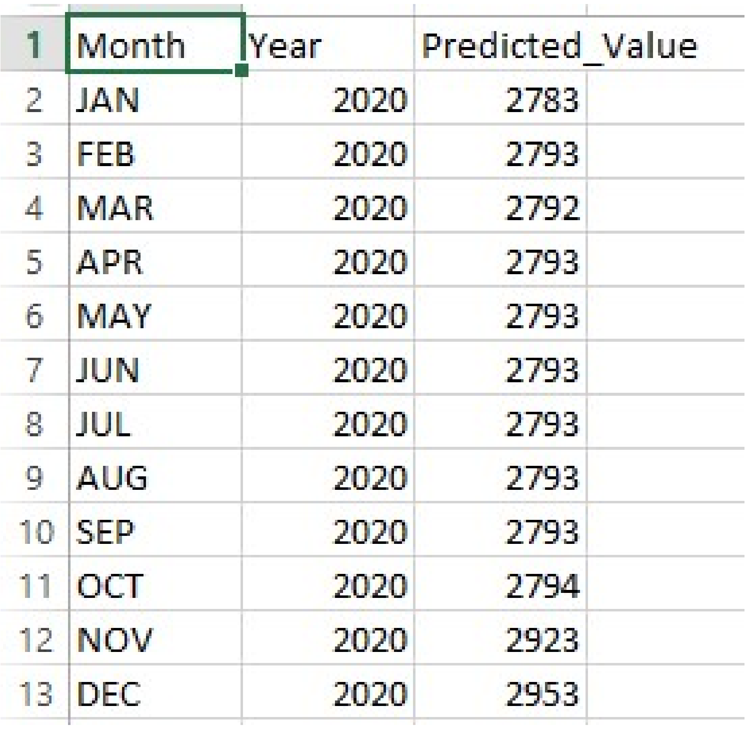
```
In [34]: model = Sequential()
         model.add(Dense(25, input_dim=2, activation='tanh'))
         model.add(Dense(45, activation='relu'))
         model.add(Dense(65, activation='relu'))
         model.add(Dense(45, activation='relu'))
         model.add(Dense(25, activation='relu'))
         model.add(Dense(1,activation='linear'))
         model.compile(loss='mean_absolute_error', optimizer=Adam(lr=0.00001))
         model.fit(X,Y, epochs=1000,validation_split=0.1,batch_size=5)

         model.save('Sales_Prediction.model')
         model.save_weights('Sales_Weights.h5')
```

```
Train on 2540 samples, validate on 283 samples
Epoch 1/1000
2540/2540 [==============================] - 2s 917us/step - loss: 2939.3481 - val_loss: 2913.2564
Epoch 2/1000
2540/2540 [==============================] - 2s 637us/step - loss: 2938.8714 - val_loss: 2912.7017
Epoch 3/1000
2540/2540 [==============================] - 2s 632us/step - loss: 2938.0684 - val_loss: 2911.5932
Epoch 4/1000
2540/2540 [==============================] - 2s 656us/step - loss: 2936.5724 - val_loss: 2909.6099
Epoch 5/1000
2540/2540 [==============================] - 2s 656us/step - loss: 2933.9155 - val_loss: 2906.1902
Epoch 6/1000
2540/2540 [==============================] - 2s 647us/step - loss: 2929.5370 - val_loss: 2900.6910
Epoch 7/1000
2540/2540 [==============================] - 2s 645us/step - loss: 2922.5961 - val_loss: 2892.1424
Epoch 8/1000
2540/2540 [==============================] - 2s 648us/step - loss: 2912.1232 - val_loss: 2879.3810
Epoch 9/1000
2540/2540 [==============================] - 2s 654us/step - loss: 2896.3645 - val_loss: 2860.2337
Epoch 10/1000
```

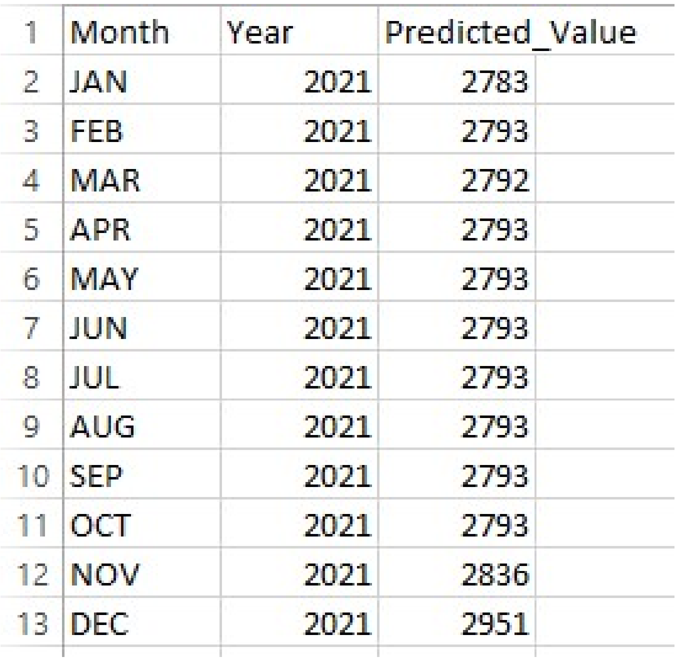
Outputs:

For 2020:



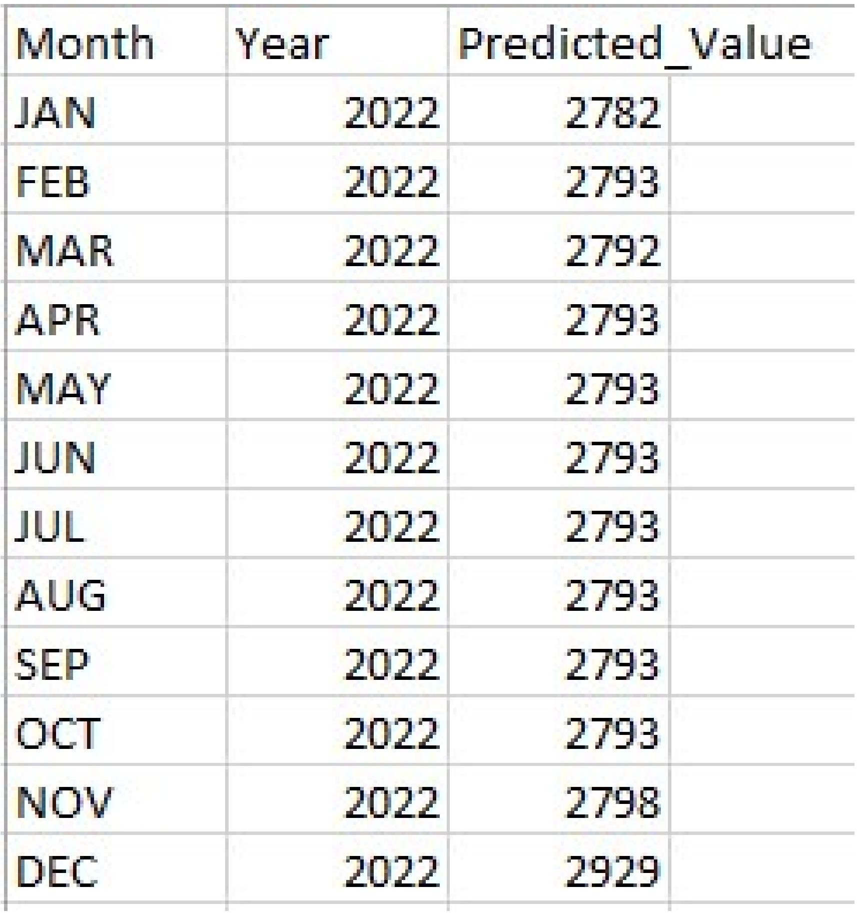
1	Month	Year	Predicted_Value
2	JAN	2020	2783
3	FEB	2020	2793
4	MAR	2020	2792
5	APR	2020	2793
6	MAY	2020	2793
7	JUN	2020	2793
8	JUL	2020	2793
9	AUG	2020	2793
10	SEP	2020	2793
11	OCT	2020	2794
12	NOV	2020	2923
13	DEC	2020	2953

For 2021



1	Month	Year	Predicted_Value
2	JAN	2021	2783
3	FEB	2021	2793
4	MAR	2021	2792
5	APR	2021	2793
6	MAY	2021	2793
7	JUN	2021	2793
8	JUL	2021	2793
9	AUG	2021	2793
10	SEP	2021	2793
11	OCT	2021	2793
12	NOV	2021	2836
13	DEC	2021	2951

For 2022:



Month	Year	Predicted_Value
JAN	2022	2782
FEB	2022	2793
MAR	2022	2792
APR	2022	2793
MAY	2022	2793
JUN	2022	2793
JUL	2022	2793
AUG	2022	2793
SEP	2022	2793
OCT	2022	2793
NOV	2022	2798
DEC	2022	2929

The predicted outputs for all the three upcoming years are almost same because if the organization won’t change its way of selling products, then the profit will remain almost same every year.