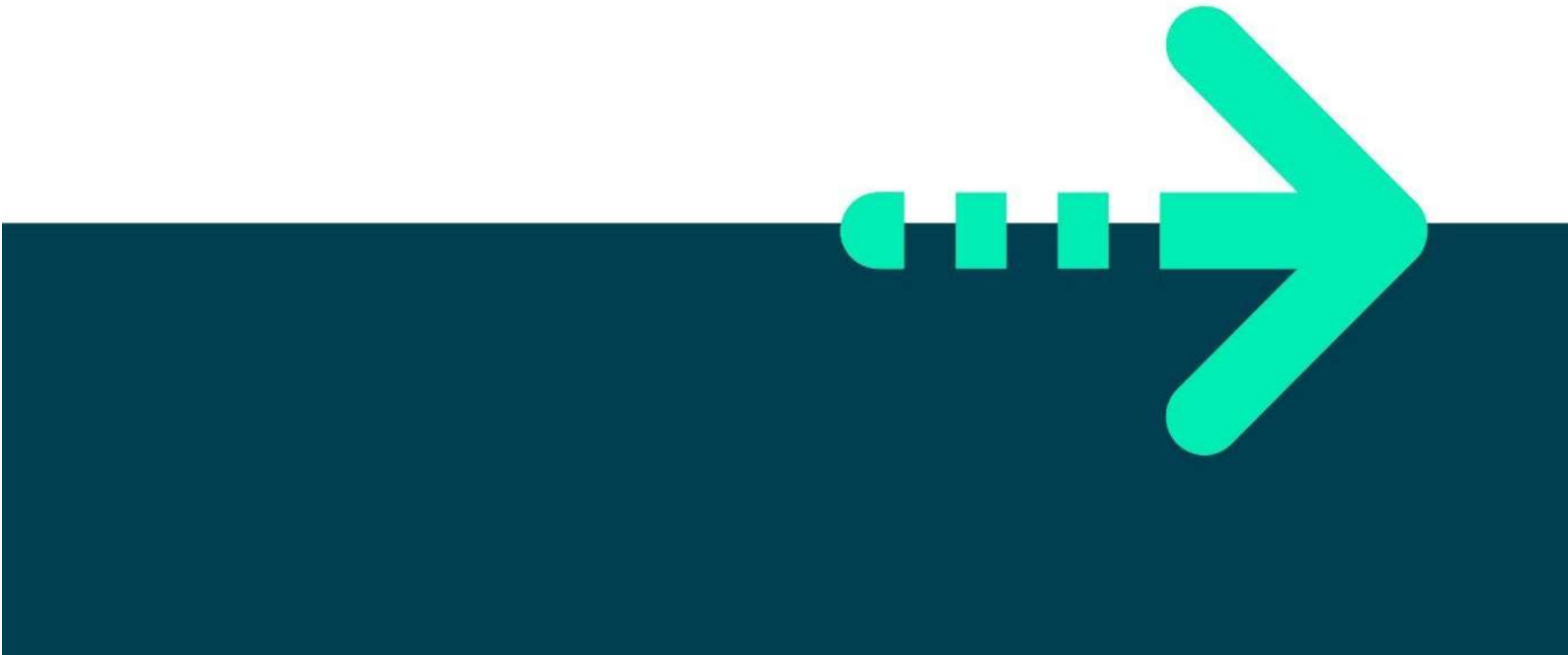
Superstore Sales Project

Creating a Data Warehouse Prototype



Contents

Introduction

I am a data analyst at QA Learning. I have been approached by a large retailer in America that would like me to help them understand how their business is performing. They are finding it hard to make data driven decisions as their data is stored in multiple operational databases and there is no single source of truth. Betty from finance helps the business but uses spreadsheets and sometimes forgets to ask Nigel for an up-to-date copy. Nigel does not have access to all the databases. Fezzan from IT has access to all the databases but always seems to have a sport injury. If Betty and other colleagues had a place for them to carry out analysis, where they all shared the same data and didn't have to rely on other colleagues to get datasets, the business can make data driven decisions to help the business:

- Become more profitable
- Stop stocking products that are not popular
- Invest in advertising in regions that are under performing

Problem Statement

The problem that the retailer has is that they do not have a central place to storethe data and want to create a data warehouse. The retailer would like to see a proof of concept of a data warehouse to demonstrate ROI (return on investment). The retailer has given me a sample of their data from their sales source system to showcase what can be done.

This project needs to produce a data mart that can be connected to by end users to create some data visualisations. The CEO, CFO and Betty are my stakeholders,who will help me if I have any questions about the data. I will also have to present my findings to them.

The dataset needs to be in a good format and cleansed to make sure the data mart produces correct results. The last stage is to visualise the data to help the retailer gain insights to make strategical and tactical business decisions.

Objectives

The retailer wants to see the benefits of using data warehouse design techniques before the start a project to implement one. They want to answer some specific questions:

- What is the total sales revenue by region by quarter?
- What is the total units sold by product category by month?
- What is the total profit by customer segment by month?

By presenting the retailer with reports, dashboards, analysis that can help them answer these questions and others, this project will be classified as successful.

Method

Data

The data used was provided to me by the retailer. The dataset does have personal information such as the Customer's name and location. This report will not contain any sensitive information, due to GDPR.

When cleansing the dataset, some duplicate orders were found:

CA-2016-129714	**US-2014-150119**
US-2016-123750	**CA-2015-103135**
CA-2016-137043	**CA-2017-118017**
CA-2017-152912	**CA-2016-140571**

Out of the 8 orders, only one could be de-duplicated (US-2014-150119) as the sales total matched. The other orders had different sales values and both records had to be removed from the dataset. The anomalies have been passed onto Betty to investigate which row is the correct row.

This project is about using a data warehousing technique. A data warehouse is effectively a database that is separated by time delay from any operational system databases. A data warehouse's sole purpose is to be used for analytical purposes. The data can come from multiple source systems and live in the same data warehouse so users can have a unified and consistent view of the underlying data. The data is structured in such a way that makes it easy for business users to use the data indata visualisation software such as PowerBI or Tableau. The data needs to follow astar schema which is made up of fact tables and dimensions.

Database Design

This project is about using a data warehousing technique. A data warehouse effectively a database that is separate from any operational system databases. A data warehouse's sole purpose is to be used for analytical purposes. The data can come from multiple source systems and live in the same data warehouse so users can have a unified and consistent view of the underlying data. The data is structured in such a way that makes it easy for business users to use the data in data visualisation software such as PowerBI or Tableau. The data needs to follow a star schema which is made up of fact tables and dimensions.

Facts: Facts are numbers and are also called measures. A fact relating to sales could be "Sales in US Dollars". Facts also have a defined granularity, which is the level of detail. For example, "Sales in US Dollars" may have a granularity of "Sales in US Dollars by day", "Sales in US Dollars by month".

Dimension: Dimensions are used to filter, categorise, and label facts. Going back to the questions provided by the retailer:

- What is the total sales revenue by region by quarter?
- What is the total units sold by product category by month?
- What is the total profit by customer segment by month?

A sales fact table needs to be created and the dimensions that need to be created are Date dimensions, Geography dimension, Product dimension and a Customer dimension.

Now that we know our facts and dimensions, we need to organise our data. Data warehoused are normally organised using a star schema. Facts are stored in Fact tables at the centre of the star and the dimensions surround the fact table. Facts have foreign keys to each dimension table.

The data provided came in a spreadsheet with the following columns:

Row ID
Order ID
Order Date
Ship Date
Ship Mode
Customer ID
Customer Name
Segment
Country
City
State
Postal Code
Region
Product ID
Category
Sub-Category
Product Name Sales
Quantity
Discount
Profit

As you can see, there was no mention of the ship mode by the stakeholder questions, but because it is in the dataset, I have decided to make a Ship Mode dimension as well.

Here is what the designed schema looks like:

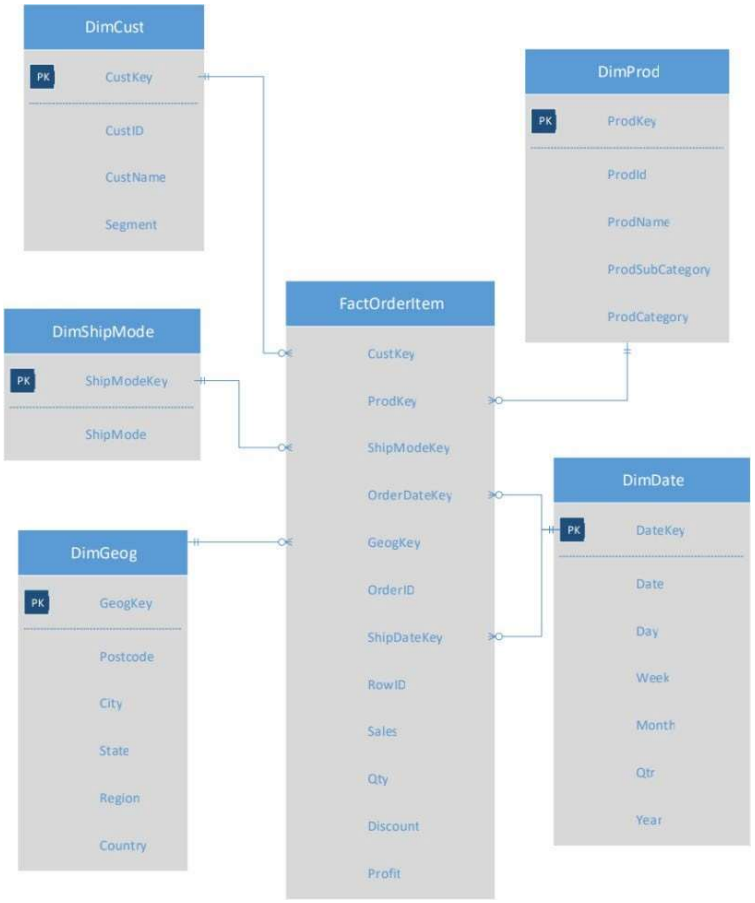


Figure 1: Star Schema

Database Creation

After the data was cleansed the data was split out into the dimensions and fact table described earlier on. **SQL Server Management Studio** was used to create a database with the following tables:

orders

cust

prod

dim_date

geog

shipmode

The import wizard was used to import the data into each table:

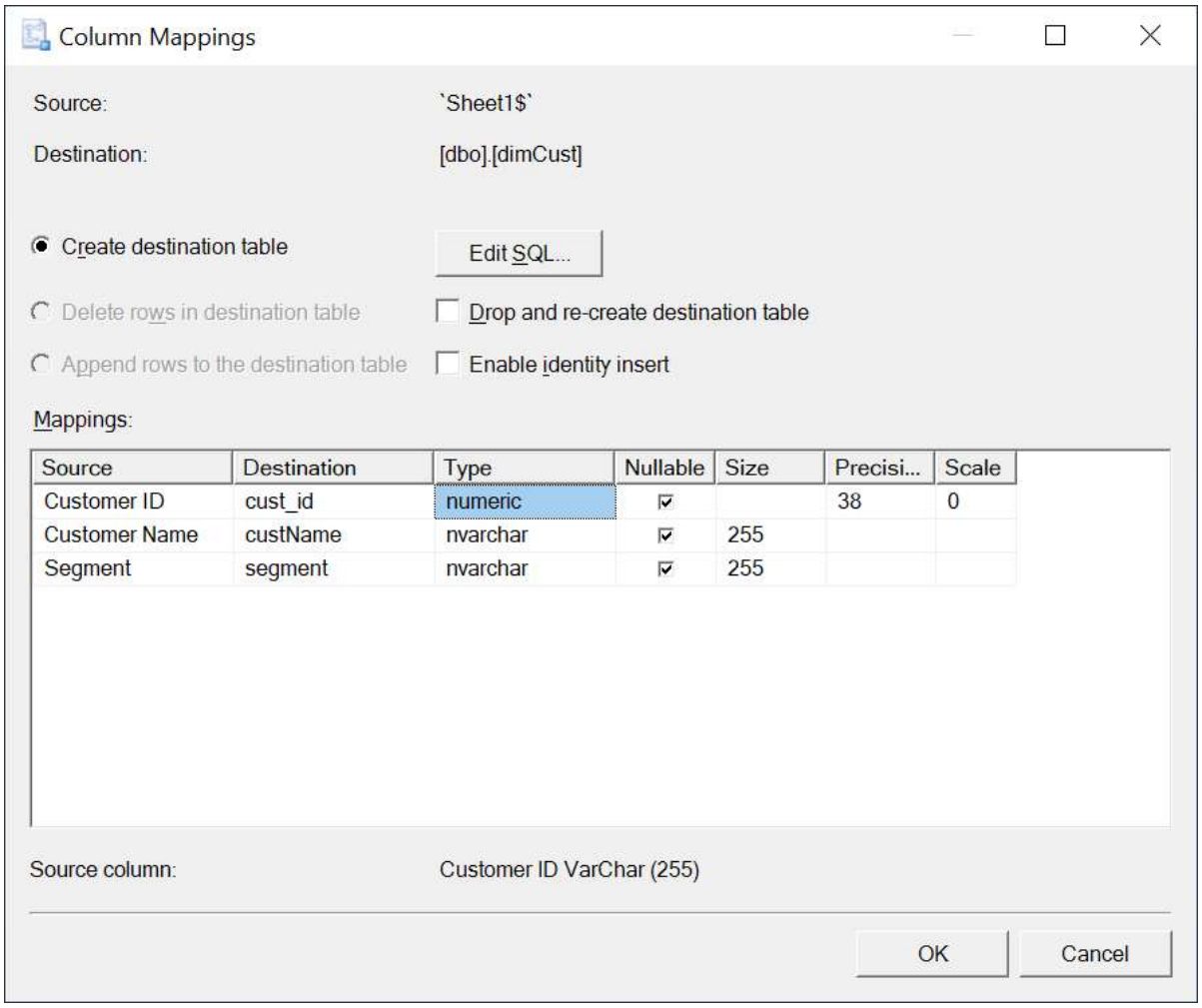


Figure 2: Import Wizard

Once each table had been populated. I created the relationships between each dimension and the fact table using the Database designer.

The last step is to create visualisations by connecting to the data from our chosen visualisation tool. This was quite easy using the 'Connect to SQL Server' functionality in Tableau:

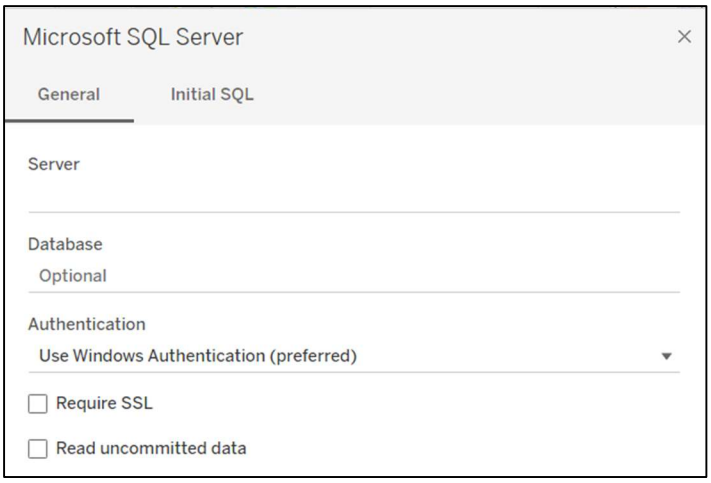


Figure 3: Connect to Data Source -SQL (Tableau)

Once we have connected the tables in Tableau this is how the model looks:

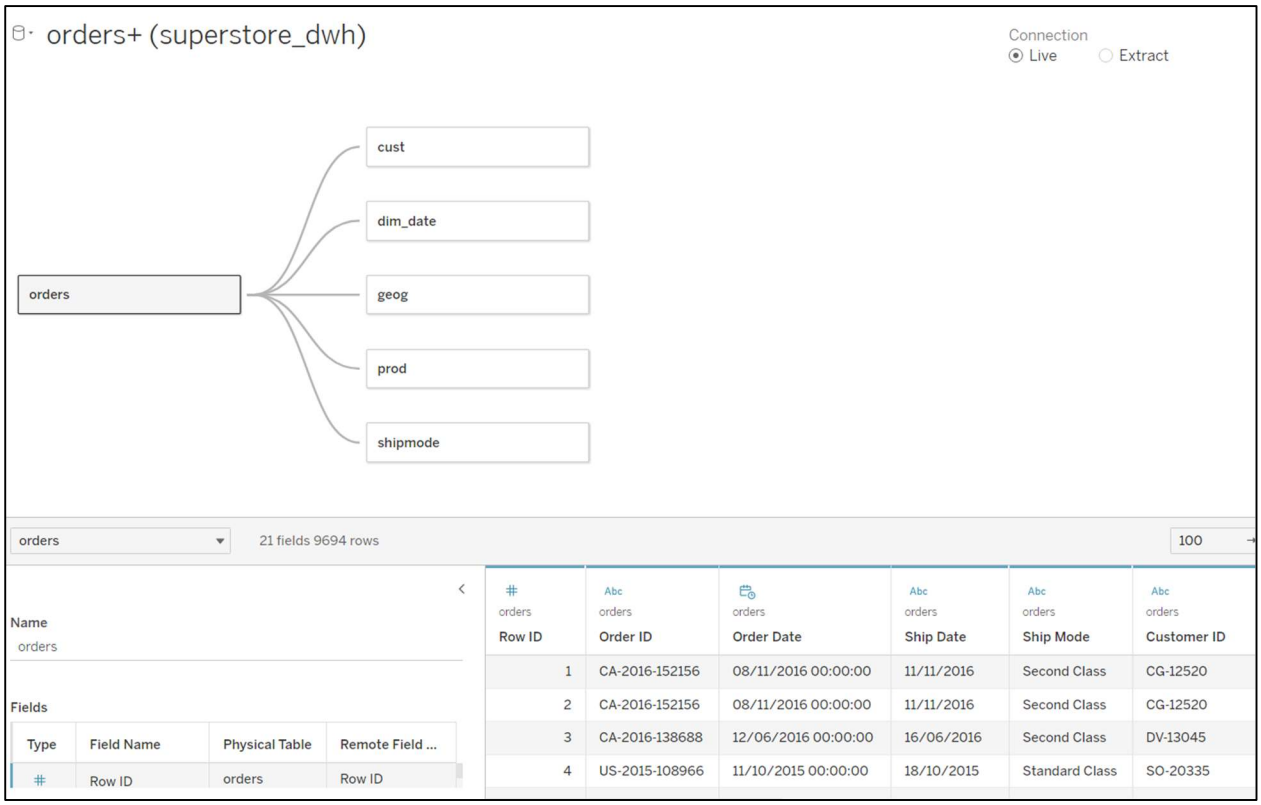


Figure 4: Data relationships logical layer in Tableau

The data visualisations will be displayed in the next section.

Results and Analysis

What is the total sales revenue by region by quarter?

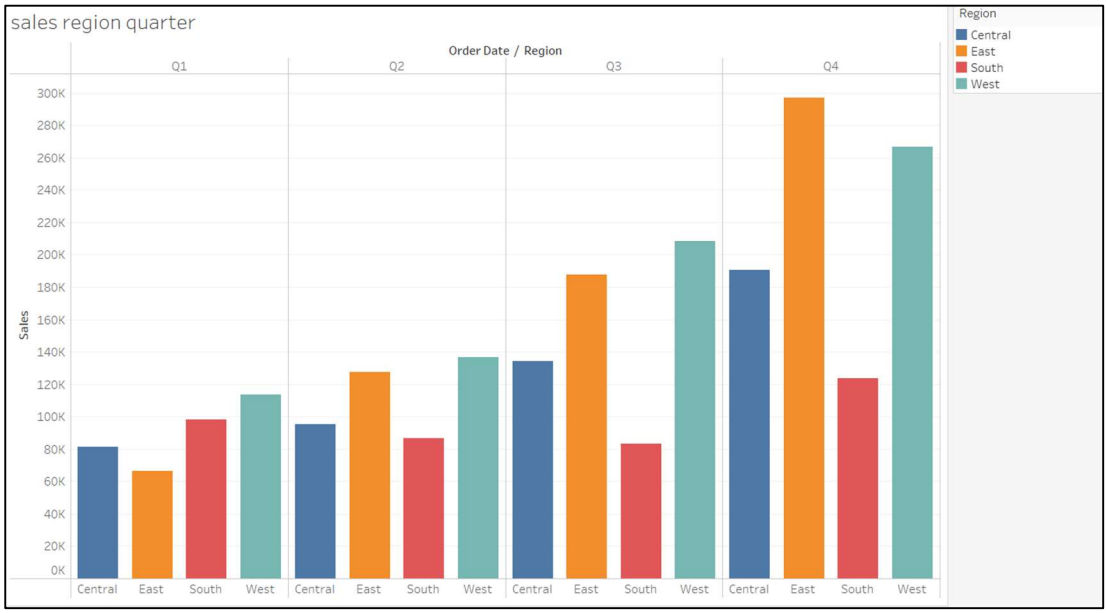


Figure 5: Total sales revenue by region and by quarter

This visualisation was created very quickly using Tableau

What is the total units sold by product category by month?

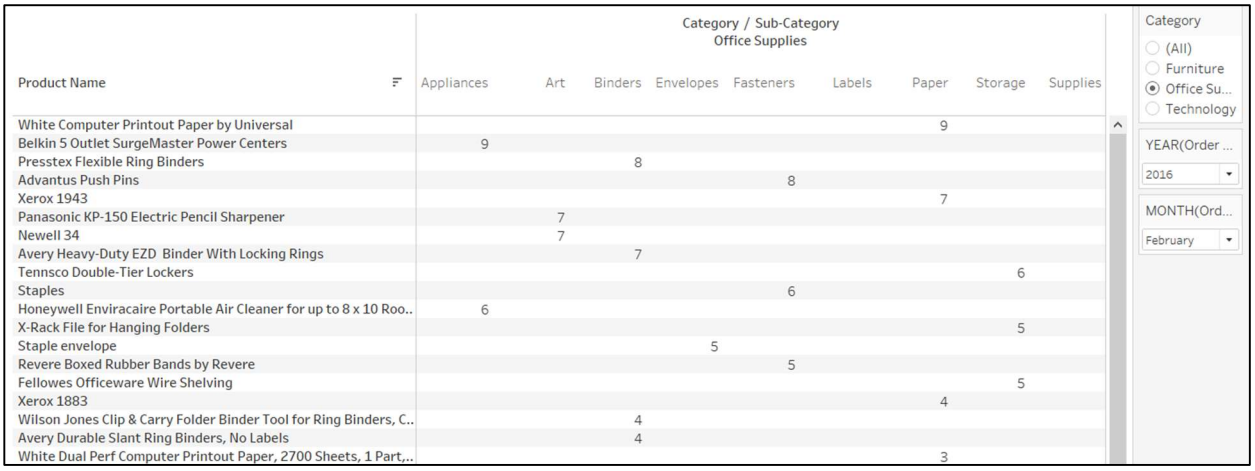


Figure 6: units sold by product and filtered by product category / year/ month

This is just showing how the data has been manipulated, some drill down functionality is required and this is hard to show in a snapshot of this report. But the data isavailable and can be manipulated by the end user quite easily.

What is the total profit by customer segment by month?

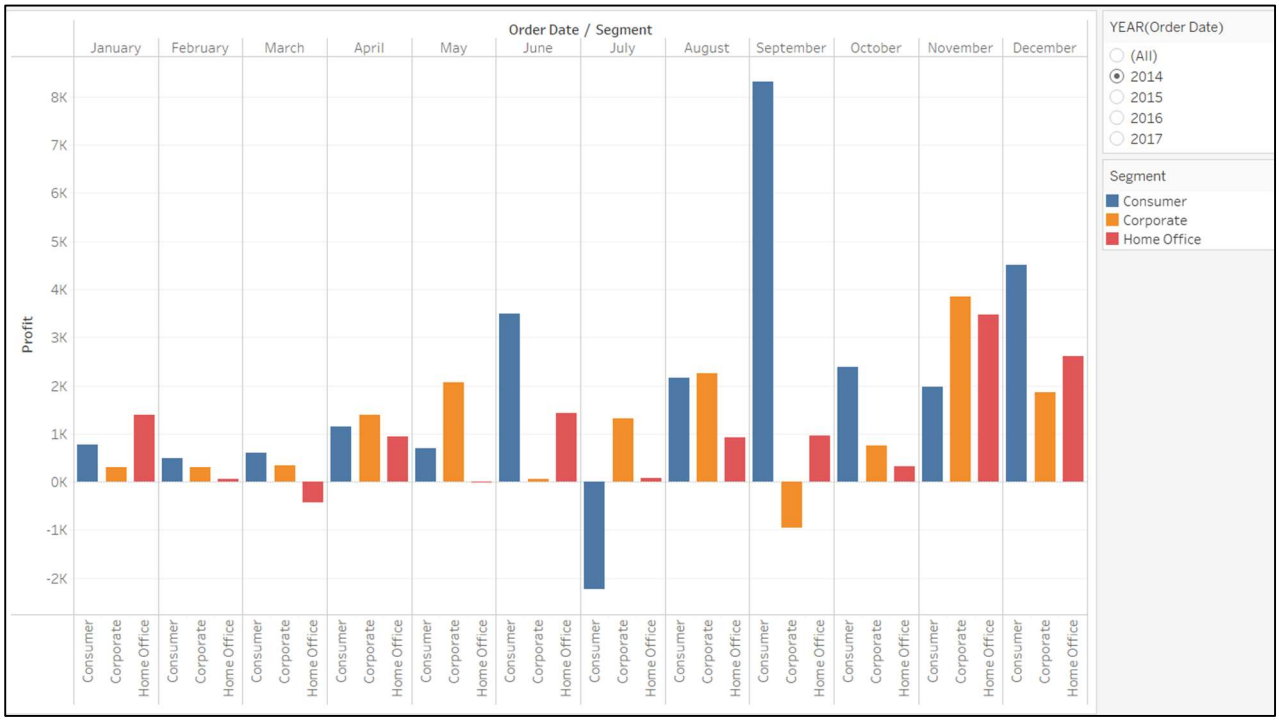


Figure 7: Total profit by month and customer segment

This chart shows the breakdown of profit by each customer segment during each month, and can be filtered by year

Conclusion and Recommendation

This project has been successful as I was able to demonstrate how to create a data mart with the dataset provided to me by the American retailer. I have showed my visualisations to the retailer, and they are astonished how they canmanipulate the data so easily. Once I answered their original questions, they had a lot more questions, but because the data was structured so well, I was able to answer their question in real time.

A data warehouse is a big undertaking and will require a lot more time and effort invested in the creation of it. You must gather requirements from the stakeholders to find out all the data sources required and the use cases of the data. Not only do you have to create the star schema with the facts and dimensions, but you also need to come up with a strategy of keeping the data up to date with scheduled tasks. However, the retailer is extremely happy with the prototype and wants to use the visualisations in the upcoming board report.

I recommend expanding the team to help complete the task of the data warehouse. There are other databases on the market that can import data from a wide range of databases, not just spreadsheets or Microsoft SQL Server, such as Snowflake, through the use of Java Database Connectivity (JDBC).

Also, an ETL tool might be useful to use as I had to search for data anomalies manually.

KSBs achieved

I feel I have achieved the following KSBs:

Knowledge

K6: the fundamentals of data structures, database system design, implementation and maintenance

I have designed a star schema for my data warehouse. Once the design phase wasover, I added the data into Microsoft SQL database.

K11: approaches to organisational tools and methods for data analysis

The company I work for, the tools used for databases are Microsoft SQL server, Tableau and Excel. These tools were used in the implementation of this project.

Skills

S4: analyse data sets taking account of different data structures and database designs

This project was focussed on the implementation of a data mart using a star schema. Tableau visualisations were produced using both types of data structures.

S7: undertake customer requirements analysis and implement findings in data analytics planning and outputs

I was given a brief from stakeholders. I had to plan the data structures required for the data analytics required. Output visualisations were produced to help stakeholders answer questions in the brief.

S12: collaborate and communicate with a range of internal and external stakeholders using appropriate styles and behaviours to suit the audience

I communicated with internal stakeholders Betty, Nigel and Fezaan. The data visualisations were required by Betty, the CEO and the CFO. A professional presentation was given to Betty, the CEO and CFO. However, before the presentation I gave Betty a tutorial, as she had a good knowledge of the data and could then help explain the insights.

S15: select and apply the most appropriate data tools to achieve the optimum outcome

Tableau is very powerful and has functionality to import data from databases. I

feel this was the best data tool to use for this project. In my recommendation, I don't think Microsoft SQL Management Studio would be able to handle non-Microsoft databases very well, so a different tool should be used, such as Snowflake.

Behaviours

B4: logical and analytical

I feel that I have showcased logical and analytical approach to this project.

Appendix A: Data Analyst Apprenticeship Standard KSBs

Knowledge

K1: current relevant legislation and its application to the safe use of data

K2: organisational data and information security standards, policies and procedures relevant to data management activities

K3: principles of the data life cycle and the steps involved in carrying out routine data analysis tasks

K4: principles of data, including open and public data, administrative data, and research data

K5: the differences between structured and unstructured data

K6: the fundamentals of data structures, database system design, implementation and maintenance

K7: principles of user experience and domain context for data analytics

K8: quality risks inherent in data and how to mitigate or resolve these

K9: principal approaches to defining customer requirements for data analysis

K10: approaches to combining data from different sources

K11: approaches to organisational tools and methods for data analysis

K12: organisational data architecture

K13: principles of statistics for analysing datasets

K14: the principles of descriptive, predictive and prescriptive analytics

K15: the ethical aspects associated with the use and collation of data

Skills

S1: use data systems securely to meet requirements and in line with organisational procedures and legislation including principles of Privacy by Design

S2: implement the stages of the data analysis lifecycle

S3: apply principles of data classification within data analysis activity

S4: analyse data sets taking account of different data structures and database designs

S5: assess the impact on user experience and domain context on data analysis activity

S6: identify and escalate quality risks in data analysis with suggested mitigation or resolutions as appropriate

S7: undertake customer requirements analysis and implement findings in data analytics planning and outputs

S8: identify data sources and the risks and challenges to combination within data analysis activity

S9: apply organisational architecture requirements to data analysis activities

S10: apply statistical methodologies to data analysis tasks

S11: apply predictive analytics in the collation and use of data

S12: collaborate and communicate with a range of internal and external stakeholders using appropriate styles and behaviours to suit the audience

S13: use a range of analytical techniques such as data mining, time series forecasting and modelling techniques to identify and predict trends and patterns in data

S14: collate and interpret qualitative and quantitative data and convert into infographics, reports, tables, dashboards and graphs

S15: select and apply the most appropriate data tools to achieve the optimum outcome

Behaviours

B1: maintain a productive, professional and secure working environment

B2: show initiative, being resourceful when faced with a problem and taking responsibility for solving problems within their own remit

B3: work independently and collaboratively

B4: logical and analytical

B5: identify issues quickly, investigating and solving complex problems and applying appropriate solutions. Ensures the true root cause of any problem is found and a solution is identified which prevents recurrence.

B6: resilient - viewing obstacles as challenges and learning from failure.

B7: adaptable to changing contexts within the scope of a project, direction of the organisation or Data Analyst role.