DS- Project Group 54 submission

Edition: 2023 2A
Project: Business Intelligence
Primary Topic: DEP
Secondary Topic: PM
Members: Manojkumar Muthukumaran, Piyush Singh
Last update: 26/4/2024, 12:28:30

Motivation

Business Intelligence enhances and helps an organization by enabling data-driven decision-making. It gives management and stakeholders an idea of where to improve, what the next plan of action should be, where they are performing well, if the process followed is yielding results, and if it will deliver the expected results in the future. It will even bring to light areas that were previously hidden. In this project, we have dealt with the case of Classic Models Inc., a company that sells miniature models of various vehicles. We received data about employees, transactions, offices, products, and customers, and we had to come up with insights to examine current practices, to determine if they are effective and pinpoint specific areas in which they are lacking. We used the Balanced Scorecard approach to come up with business questions. This approach introduces four different perspectives: financial, internal process, customer, and learning and development. The idea is that focusing only on financial performance indicators (KPIs) will not benefit an organization in the long term. A company needs to focus on all four aspects mentioned above to grow in a holistic way and sustain that growth. Ensuring that all the other aspects are covered will result in better financial performance (1). We have used various data preparation, processing, analysis, exploration, and visualization techniques to effectively handle the data and to create an ETL pipeline to answer the various business questions catering to the four different business perspectives.

(Business/Research) questions

The project has followed the guidelines of the business scorecard perspectives and has focused on the four aspects the business scorecard mentions, namely learning and development, customer, internal business/process, and financial. We have identified focus areas under each of this perspective and have tried to answer questions under each focus area. Further, multiple Key Performance Indicators (KPIs) have been chosen to analyze and back the solutions quantitatively (2). The business questions are as follows:
1) Learning and Development:
Focus area: “Improve Sales Rep skills and performance”
Specific question to be addressed - Find the country with the least performing Sales Reps.
KPIs:

- Average revenue per Sales Rep for each location.
- Average number of sales performed per Sales Rep by country.

2) Customer:
Focus area: "Increase Customer Profitability."
Specific question to be addressed - Analyze customer purchase trends over a period and identify problems.

KPIs:

- $ Revenue per customer (QoQ)
- No. of customers serviced (QoQ)

3) Internal Business/Process:
a) Focus area: “Streamline product distribution and logistics.”
Specific question to be addressed - Identify and reduce lead times and improve order fulfilment.
KPIs:

- % of orders shipped late per product line.

b) Focus area: “Indicate Potential Overstocking”
Specific question to be addressed - Compare the selling price with the MSRP and the buy price to identify if any price adjustments can be made.
KPIs:

- # products in stock vs # products sold or ordered per product line.

4)Financial Perspective
Focus are: “Identify pricing issues”
Specific question to be addressed - Compare the selling price with the MSRP and the buy price to identify if any price adjustments can be made.
KPIs:

- % deviation of selling price from MSRP.
- % deviation of selling price from buyPrice

Source data

The source data used is the ClassicModels dataset. There were six files in total, namely - Products, Payments, Orders, Orderdetails, Employees, Offices, and Customers. Initially, all these text files were read and converted into a dataframe. In this process, headers had to be added to each dataframe as the files which were consumed to create these dataframes had no headers. Headers make it simpler to access values in a dataframe. This was followed by data cleaning. Each dataframe was checked for the existence of null values. Null values for different data frames were handled differently. Empty values in the Orders dataframe were ignored as we weren't interested in those particular columns. Columns addressLine2, state, and territory in Offices, and addressLine2, state, salesRepEmployeeNumber, and postalCode in Customers had blanks. Columns addressLine2, state, and postalCode were dropped from each of these tables as they had numerous null values. Logically looking at it, not all countries have states and not every customer will have the patience to fill addressLine2. Moreover, we were not interested in these columns. Null values in the territory column were filled with 'NA', denoting North America, as upon exploration, it was found that only the North American offices had this blank. Customers having blank salesRepEmployeeNumbers were removed as we were only interested in customers who had done some business. Further, each dataframe was checked for duplicates. There were no duplicates in any of the dataframes. We also checked if all products had at least one order detail associated and if every order had an order detail. Every order had an order detail, and every product was associated with an order detail. The distribution and summary statistics of each dataframe were also visualized. This was done not only to just get an idea about the dataset and to get a general overview of the overall trend but also to explore the unique values of every categorical column to get an idea of all the values a particular column can take. All categorical columns like product line, scale, shipping status, job role, etc., were explored in this manner. The distribution of products across different scales and different product lines was visualized. The distribution of paycheck amounts was plotted. Moreover, the distributions of price each and quantity ordered were also visualized. Following this, a separate dataframe for date values starting from January 1, 2003, to July 1, 2005, was created. This table had attributes for each day like which year it was, which quarter it was, and which day of the week it was. This would be used later when normalizing dataframes when integrating multiple data frames. Later, data integration was done where multiple tables were joined accordingly to get the required fact tables as mentioned in the star schema.

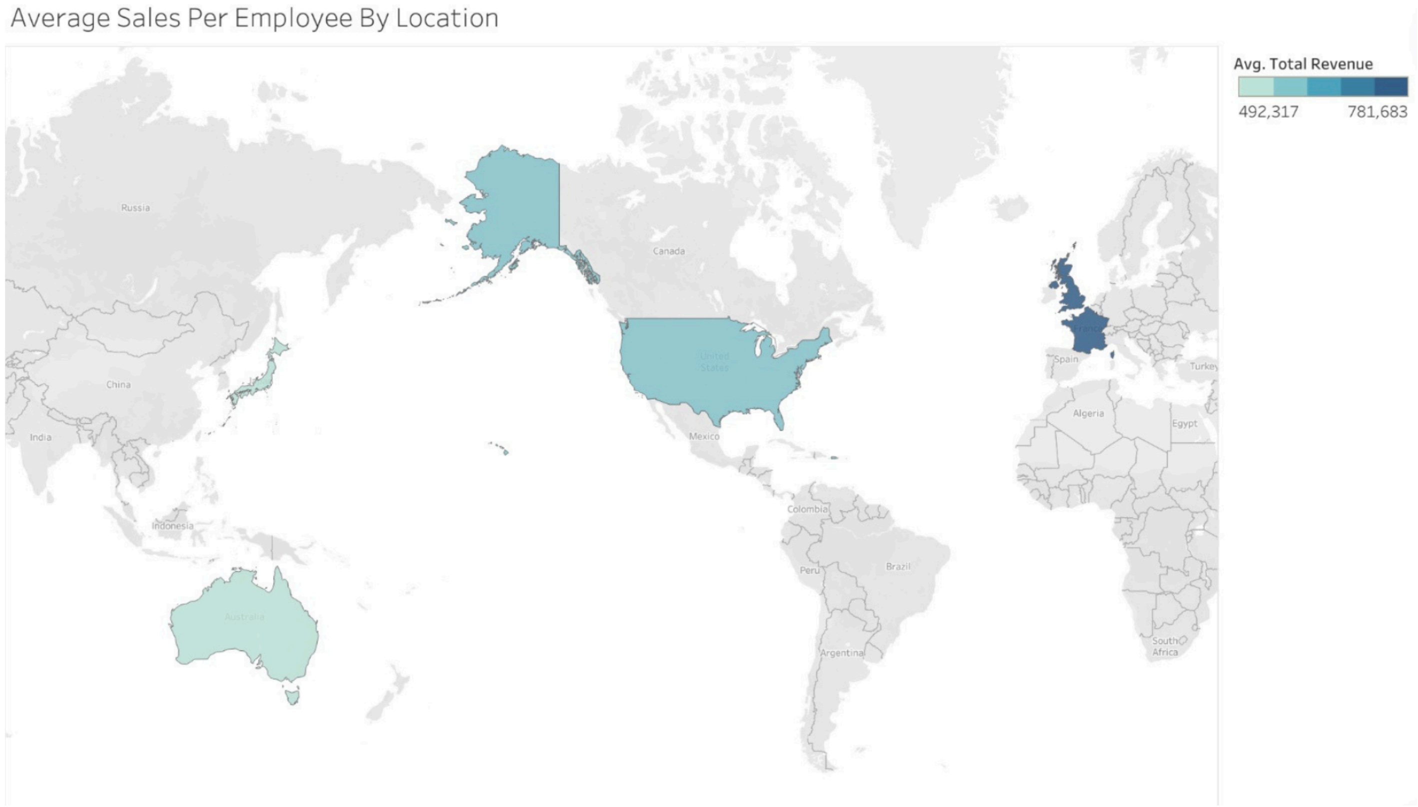
Method

All the fundamental concepts, methods, and measures of an extract, transform, and load (ETL) process were followed. We began with the extraction of data. The dataset was provided in the form of text files. I/O operations were to be performed to load the text files into the processing environment, which was a Python notebook. Python was specifically selected for the entire data processing because of the availability of robust libraries to handle data like pandas and NumPy. Further, there are a number of libraries even for visualization. Pandas was used to input the text files into the environment. The generated data frames were then profiled to see their structure and the kind of data they contained. The attribute names and their types were all analyzed to understand what each attribute is trying to convey and what information a dataframe as a whole is giving. Then we stepped into the data transformation phase. Data cleaning was done. Every dataframe was checked for duplicates. Null values were handled. A new table for dates was created. We then performed data integration, to combine different tables and to use it to answer the business questions we want to address in an efficient manner. Further, data aggregation was done to group the data in these fact tables based on categories to a level suitable for getting meaningful answers to our questions. Now we enter the load phase. We chose to go with PostgreSQL as our database management system. We particularly chose PostgreSQL because it allows relational modeling of data. Since pandas use a tabular format for storing and visualizing data, the transition would be easier. Moreover, PostgreSQL is open source and is compatible with multiple systems, programming languages, and third-party apps like Tableau. All the data frames were loaded into the PostgreSQL database using the psycopg2 library. Data integrity was ensured. Checks were performed to see if data had been loaded correctly into the database. There were datatype mismatches that occurred during the loading process. For instance, integer values were stored as floats. These datatype mismatches had to be rectified. To answer the business questions and to communicate the results to stakeholders, visualizations were required. Various charts and graphs were plotted, analyzing trends and numbers that can effectively answer each business question. This was accomplished using Tableau, as it is efficient and effective in connecting with PostgreSQL, providing interactive and appealing visualizations just through simple drag-and-drop functionalities. In each of these steps, constant efforts were taken to ensure that data is intact, maintaining its accuracy, consistency, and reliability throughout the process. The link of the GitHub repository is mentioned in the appendix.

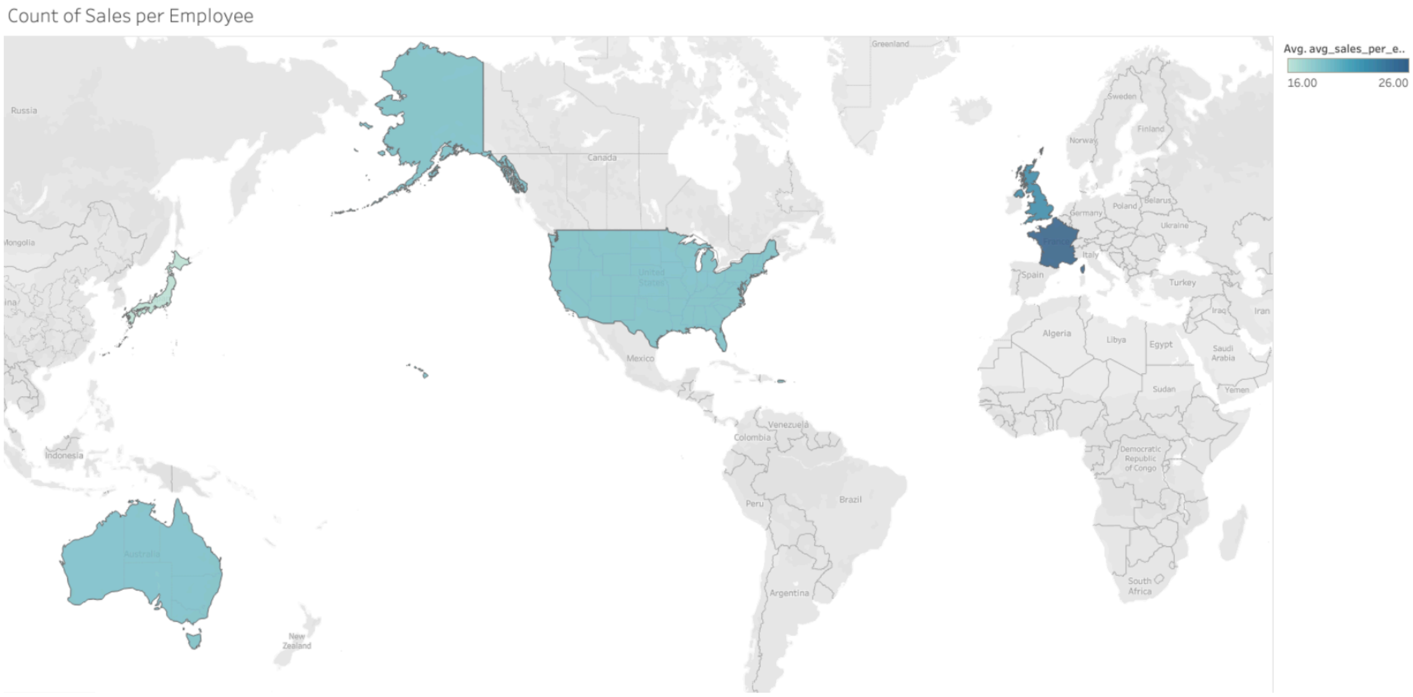
Results

Multiple graphs and charts were used to provide answer to single business question. Results from each of these visualizations were combined to come to a conclusion as we wanted to give a solution which looked at different aspects of a problem.

Learning and Development - “Improve Sales Rep skills and performance”
Find the country with the least performing Sales Reps.


Average Sales Per Employee By Location

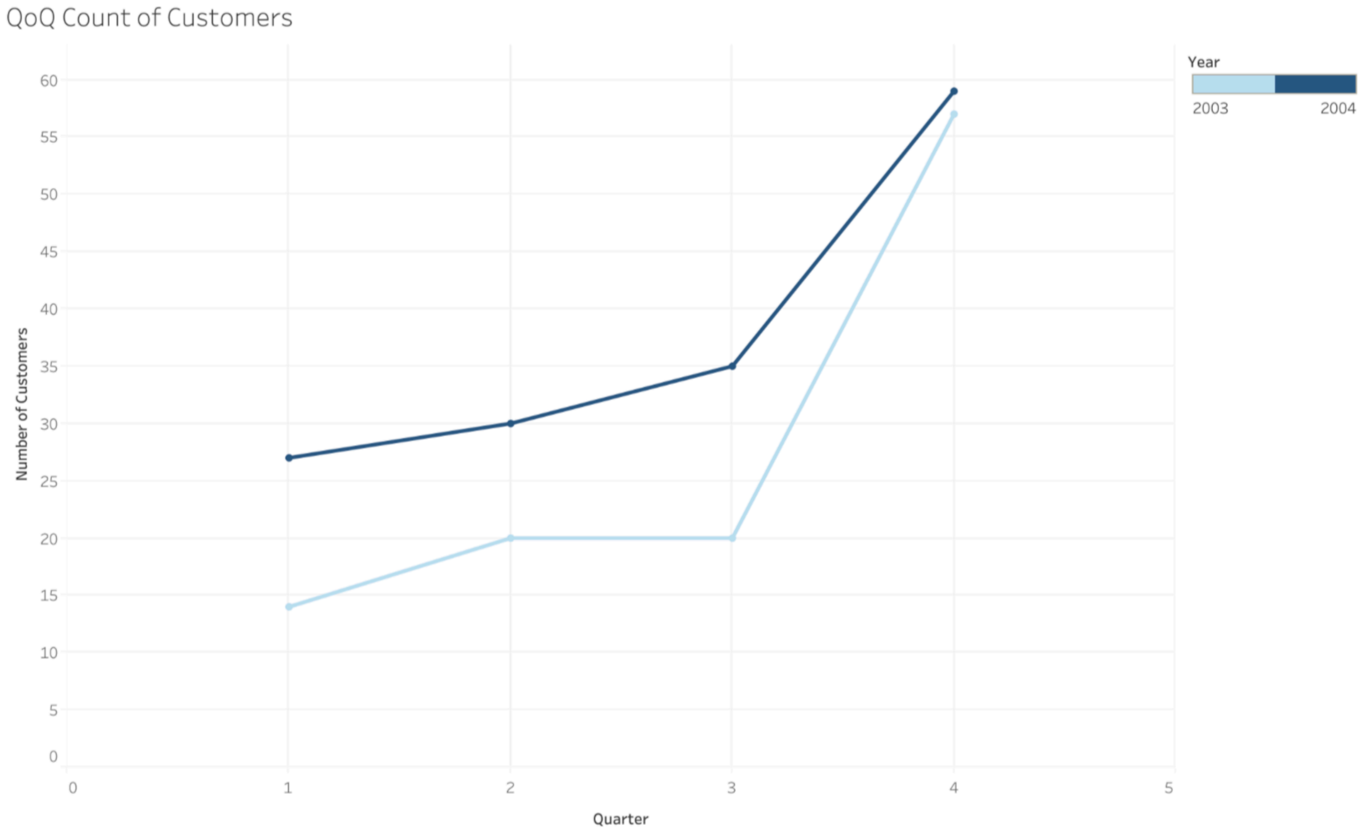
In this visualization, the darker the color of the country the higher is the average sales per employee in that country. With this we can understand that Sales Reps in the UK and France get orders that are higher in value compared to US, Japan and Australia.


Count of Sales per Employee

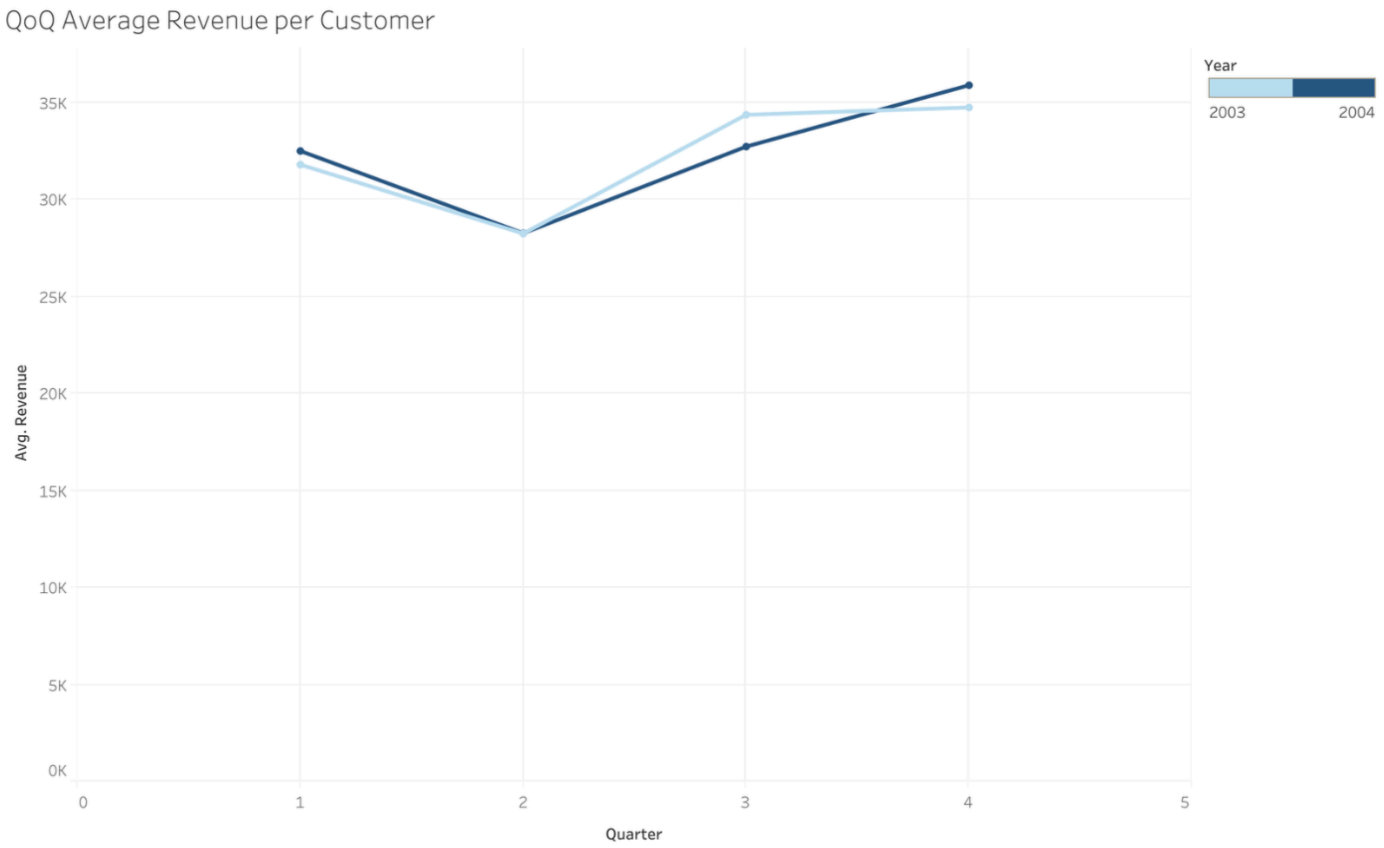
Again in this visualization, the darker the color of the country the higher is the count of sales per

employee in that country. With this we can understand that a Sales Rep in the UK and France makes more number of sales compared to US, Japan and Australia.
With these two visualizations we can clearly see that Sales Reps in Australia are performing the worst in both the aspects.

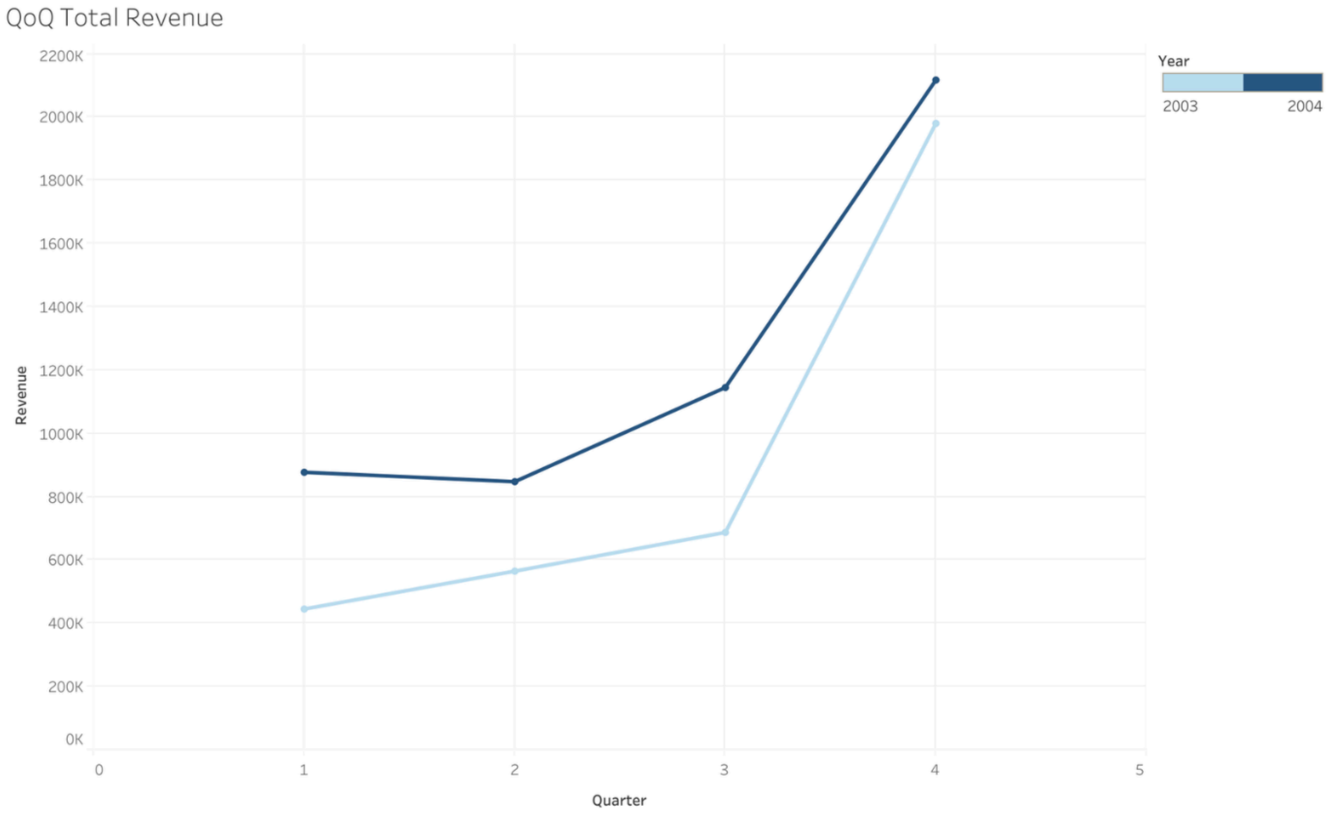
Customer Perspective **-** Increase Customer Profitability.
Analyze customer purchase trends over a period.



The dark blue line is the graph for 2004 while the other one is the graph for 2003. This graph shows us that the number of customers have increased from 2003 to 2004. There is also a steep increase in the number of customers in the third quarter which might be due to the holiday season.

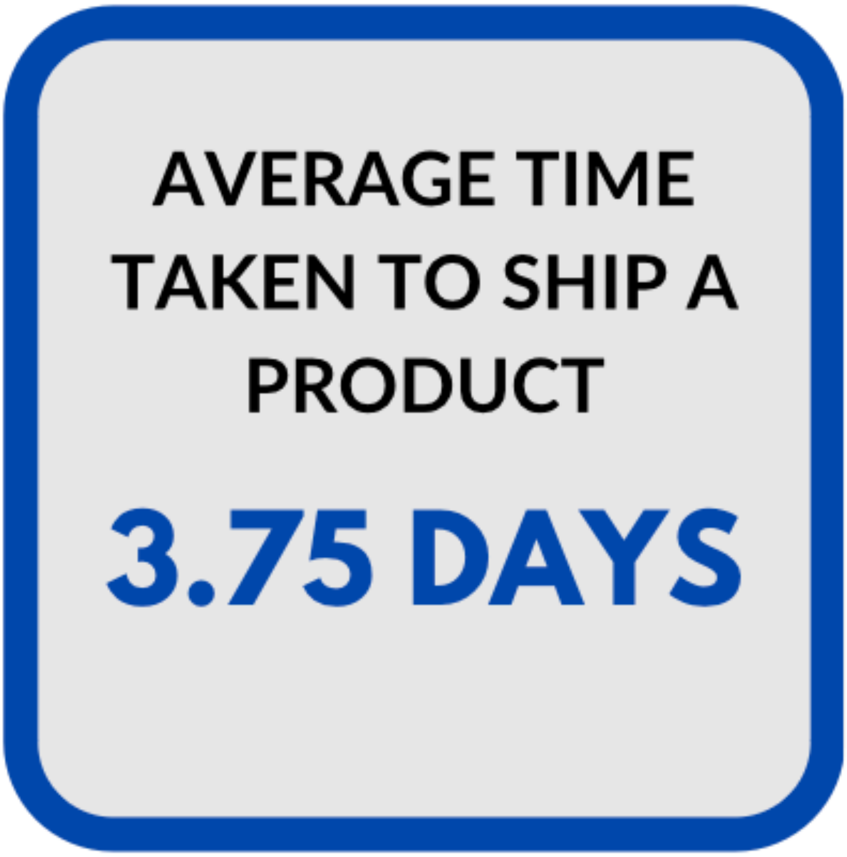


This graph shows us that the average revenue per customer has stayed nearly the same over the two years which is a positive sign as generally as the number of customers increase the average revenue per customer goes down.



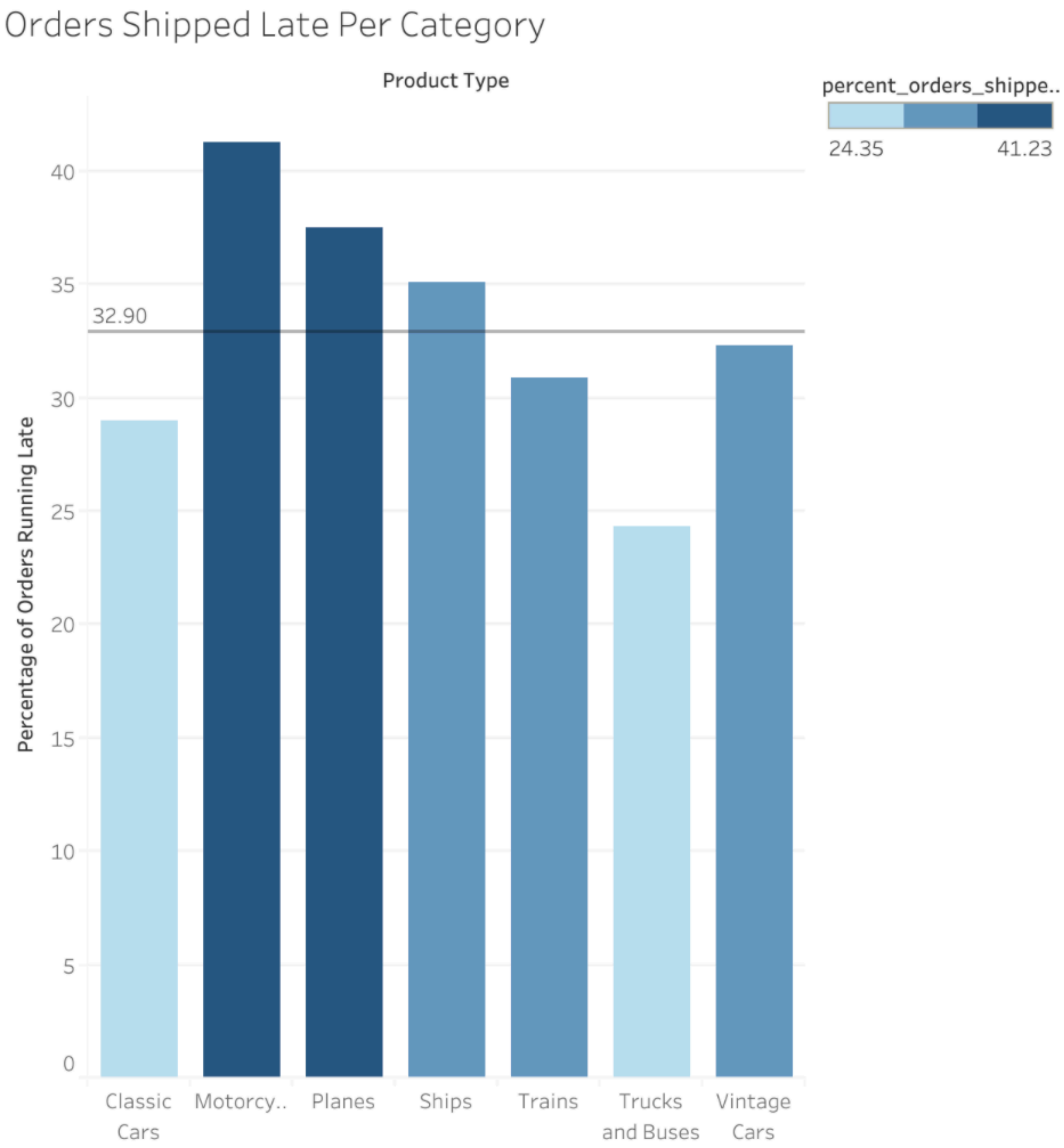
This graph is just for reference and is obtained when both the above graphs are combined.

Internal Business
“Streamline product distribution and logistics.” **-** Reduce lead times and improve order fulfilment.



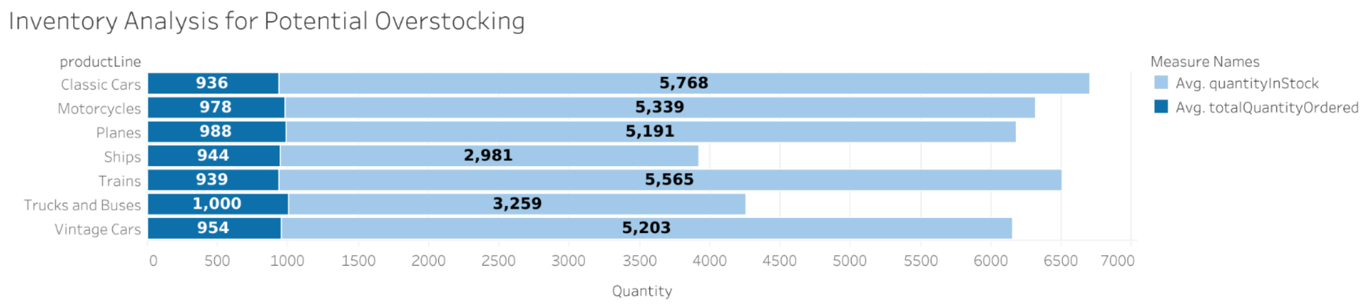
AVERAGE TIME TAKEN TO SHIP A PRODUCT

3.75 DAYS



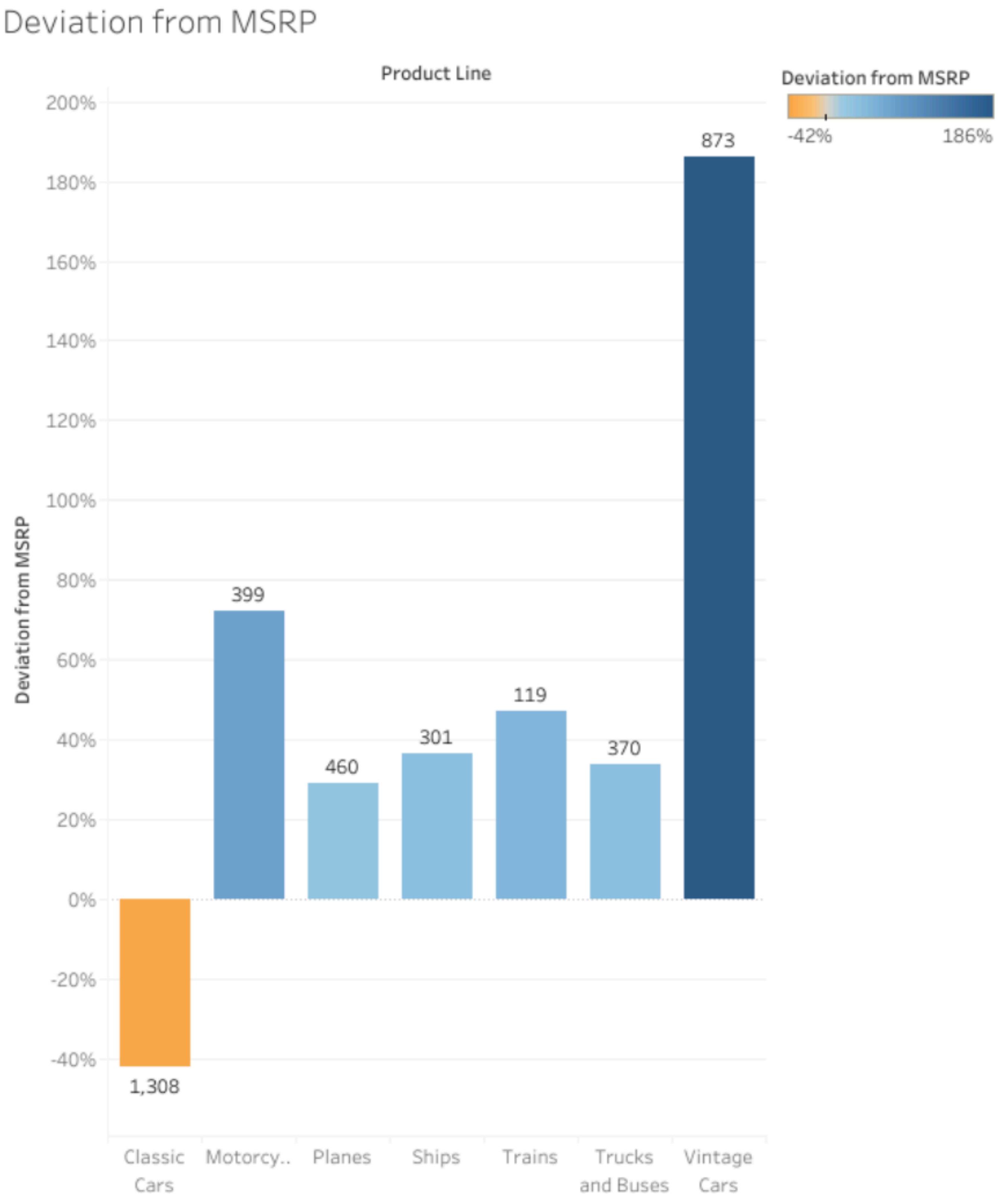
We considered products shipped after 4 days to be shipped late. This chart shows the percentage of products in each product line that has been shipped late. On average 33% of all products are shipped late with Motorcycles performing the worst.

“Indicate potential overstocking.” - Are there any product lines with high quantities in stock but low sales volume?

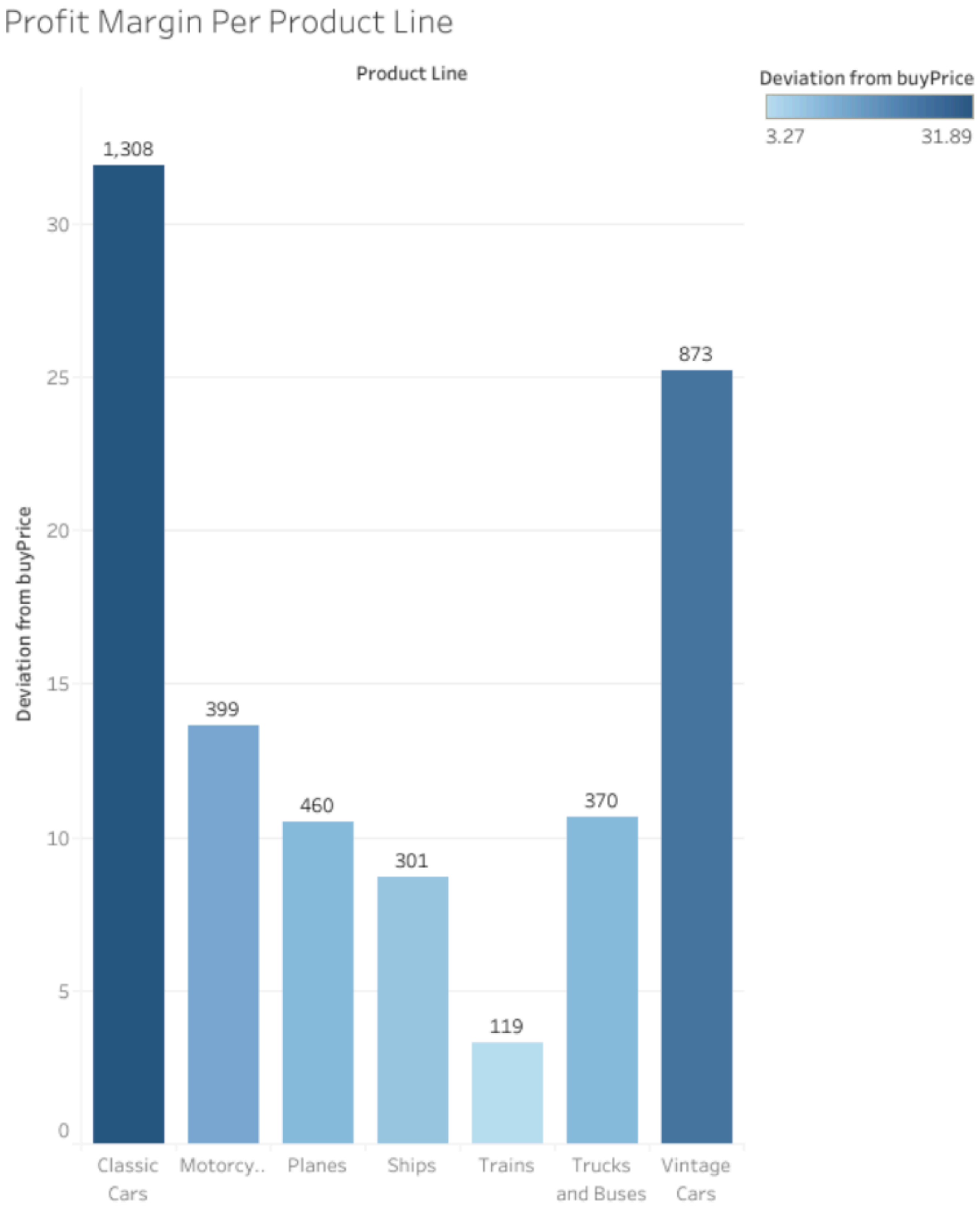


This chart compares the number of products sold and products in stock across product line. Generally all product lines are excessively overstocked.

Financial Perspective - “Identify pricing issues”
Compare the selling price with the MSRP and the buy price to identify if any price adjustments can be made.



This bar-graph depicts the percentage difference between the selling price of products and the MSRP across product lines. The number on top of each bar is the number of products sold in each category whereas the height of the bar is the deviation from MSRP.



This bar-graph depicts the percentage difference between the selling price of products and the buy price across product lines.

Reliability of results

We use a data warehouse with PostgreSQL to serve as the backbone for answering a variety of business questions, which in itself features integrated mechanisms for reliability and stability of data. Throughout the visualization and ETL process, we rigorously cross-checked the accuracy of data values and fields to ensure they were relevant and correct with regards to the key performance indicators we wanted to analyze. While visualizing results in Tableau, we took extra care in representing the correct quantities, which we also independently verified and analyzed within the original dataset to check for possible issues. Our data is sourced from the ClassicModels dataset provided by UT, which should be suitable for maintaining integrity. When sourcing from the dataset, we incorporate data cleaning processes to handle missing values and anomalies, to strive for a high degree of data quality. Alongside this, while we checked for potential outliers in the dataset to avoid negatively affecting the quality of our inferences when assessing the KPI for the business questions, we do not feel this can be completely guaranteed. Furthermore, while this problem wasn't encountered on our end, frequent use of foreign keys can result in added computational load for querying, which can possibly affect performance on lower spec hardware. While we have implemented several measures to enhance data quality, it should be understood that the quality of the underlying data source plays a pivotal role in determining the richness and reliability of the results obtained. Improving the quality of the inherent data source can significantly increase the depth and accuracy of the information available for analysis.

Technical depth

Since we are using the ETL methodology, we realised that containerisation of each module would be beneficial. We used Docker and Docker Compose to containerize the Python scripts, the database, and the database client. This ensures consistency and portability as the ETL process runs the expected way across different environments, such as while testing on multiple systems. We kept alternating between Mac and Windows systems, and the Docker image really simplified the setup process. Further, when the pipeline reaches the deployment stage, Docker images are easier to deploy and manage using Kubernetes. This also greatly facilitates Continuous Integration and Development (CI/CD). Additionally, since we have resolved the entire process into small functionalities (microservices), it is easier in terms of management and fault handling. Furthermore, continuous integration of data can be introduced, where data can be ingested continuously and analytics can be presented in real-time (4).

Conclusions & recommendations

Problems in the company falling in the four different perspectives of the business scorecard approach were solved and comprehensive conclusions were given. The following is the summary for each problem and our recommendation:
1) Learning and Development - “Improve Sales Rep skills and performance”
SalesReps from Australia are performing the worst across the average revenue per employee and the average count of sales per employee. This can see improvement with training and cross-regional knowledge exchange.
2) Customer Perspective - "Increase Customer Profitability."
Customer count and average sales increase in Q4 possibly due to the holiday season. Average revenue has almost been constant. Need to check why the average drops in Q2. Try to attract diverse customers with purchase behaviours to further increase average sales per customer. Scaling measures should be brought in to handle the influx of customers.
3) Internal Business -
a) “Streamline product distribution and logistics.”
b) “Indicate potential overstocking.”
Waiting time to ship is generally high despite abundant products available in stock. Open inventories/warehouses in multiple locations to reduce shipping time even though the dispatch

duration is large, effectively reducing the delay in delivery.
4) Financial Perspective - “Identify pricing issues”
Classic Cars exhibits the highest profitability among the product lines despite the current negative deviation from MSRP. Increase the price of this product line to at least match the MSRP to further increase profit.

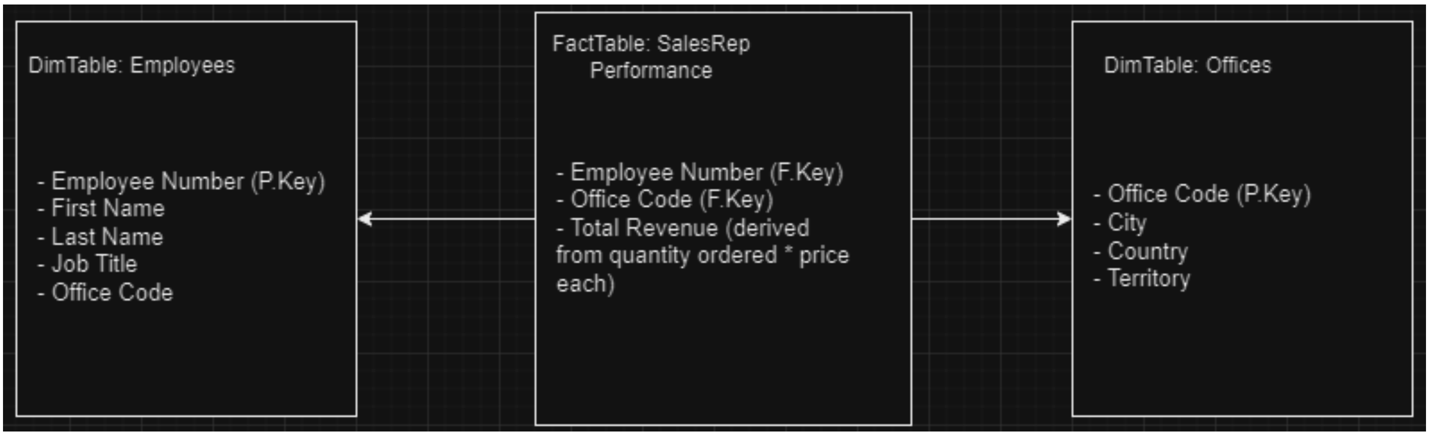
Since AI, explainableAI, Long Language Models are now performing tasks effectively, an automated model for data preparation can be brought that identifies what techniques are to be performed for a given dataset to effectively prepare the data for processing. This is being researched extensively and some level of implementation is already available for XML data in particular (5,6).

Reflection

The skills from the course have genuinely helped in this project, starting from the creation of the star schema to selecting the adequate colors for the visualizations. We used ChatGPT to gain a better understanding of the Balanced Scorecard approach and the different types of questions that are relevant for each perspective. We further used ChatGPT to search for python syntax pertaining to particular functions and operations if we got stuck in between. It also helped with the syntax for executing necessary queries in the process of exploring the data and getting an understanding of it. The method for data preparation taught in the section of multidimensional modeling was key in understanding the process flow to follow to get the desired results-from formulating questions to drawing inferences from the visualizations themselves. Assignment 2 in the course guide served as a good reference and practice for executing the ETL process and subsequently generating visualizations in Tableau. Working through the project we identified a need for better proficiency in analytics, a more thorough understanding of general business process analytics would have proved beneficial. Recognizing possible metrics that could be generated using the information from the dataset required meticulous data exploration and recurring discussions. Overall, we feel that this has provided valuable experience in dealing with problems that a data scientist might face in the real world.

Cube Data

We use multidimensional modeling to design star schemas that are able to address the analysis needed to provide accurate answers to the business questions raised in the project. For the learning and development perspective of the balanced scorecard-with focus on finding the country with the least performing sales reps, which needed analysis on average revenue and number of sales per sales rep for each location, we design the following schema:



DimTable: Employees

- Employee Number (P.Key)
- First Name
- Last Name
- Job Title
- Office Code

FactTable: SalesRep Performance

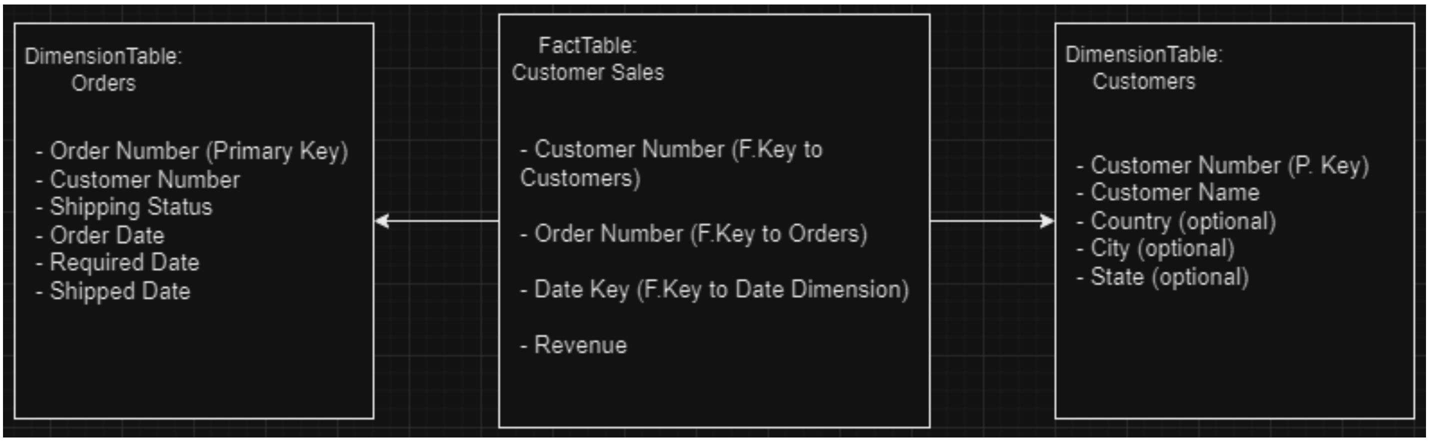
- Employee Number (F.Key)
- Office Code (F.Key)
- Total Revenue (derived from quantity ordered * price each)

DimTable: Offices

- Office Code (P.Key)
- City
- Country
- Territory

this enables us to correctly identify the revenue across different territories and sales rep, leading us to the conclusion that Australia is primarily the territory where sales rep training would be most beneficial. The Sales Rep fact table links to the existing "Employees" and "Offices" dimension table in the database using foreign keys while "Total Revenue" acts as the central fact in the fact table.

For customer perspective, we needed to be able to analyze customer purchase trends over a period and identify problems for finding ways to increase customer profitability. We use the following schema to accomplish this:



DimensionTable:
Orders

- Order Number (Primary Key)
- Customer Number
- Shipping Status
- Order Date
- Required Date
- Shipped Date

FactTable:
Customer Sales

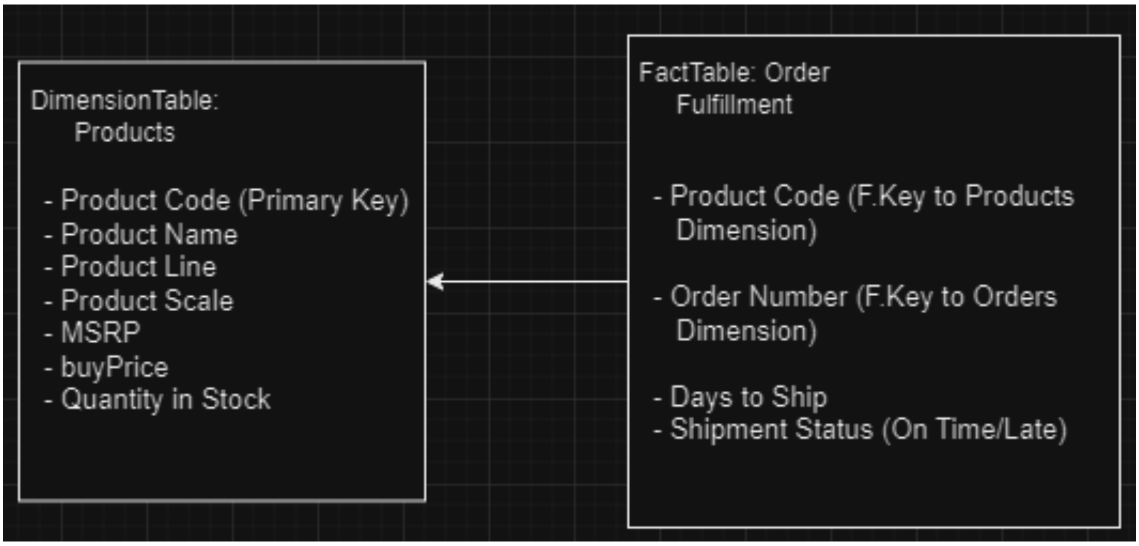
- Customer Number (F.Key to Customers)
- Order Number (F.Key to Orders)
- Date Key (F.Key to Date Dimension)
- Revenue

DimensionTable:
Customers

- Customer Number (P. Key)
- Customer Name
- Country (optional)
- City (optional)
- State (optional)

Using this we can accurately chart revenue per customer and the number of customers serviced MoM and QoQ across a specified period of time. The "Customer Sales" fact table links to the "Orders" and "Customers" dimension in the database using foreign keys while the "Revenue" field, which is our central fact, stores the revenue per customer per order. The date key links to the date dimension using a foreign key, the date dimension is detailed in the fourth figure of this section.

The Internal Business Process perspective requires determining the percent of orders shipped late and comparing products in stock vs sold. The first is handled by the schema given below:



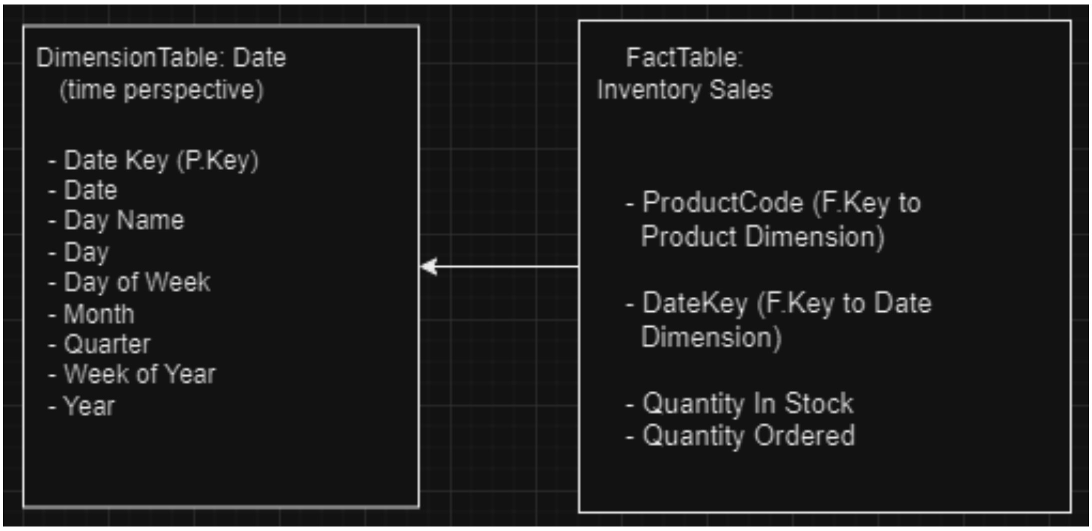
DimensionTable:
Products

- Product Code (Primary Key)
- Product Name
- Product Line
- Product Scale
- MSRP
- buyPrice
- Quantity in Stock

FactTable: Order
Fulfillment

- Product Code (F.Key to Products Dimension)
- Order Number (F.Key to Orders Dimension)
- Days to Ship
- Shipment Status (On Time/Late)

Here, we can use "days to ship", which is the central fact to discern the number of orders that are running late, which is assumed to be 4 days or later, based on the knowledge that customers want their orders delivered in 6 days from order placement. Order number links to the "Order" dimension detailed in the 2nd schema. For the second business question in this perspective which requires indicating potential overstocking, we use the subsequent schema:



DimensionTable: Date
(time perspective)

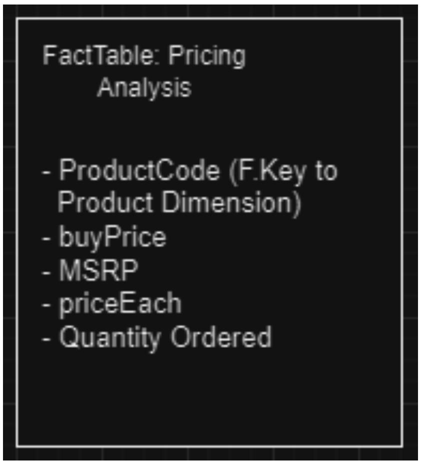
- Date Key (P.Key)
- Date
- Day Name
- Day
- Day of Week
- Month
- Quarter
- Week of Year
- Year

FactTable:
Inventory Sales

- ProductCode (F.Key to Product Dimension)
- DateKey (F.Key to Date Dimension)
- Quantity In Stock
- Quantity Ordered

Linking current stock and quantity of sales to product code enables tracking of inventory and sorting by product type or scale if the need arises.

Finally, for the financial perspective we needed to identify pricing issues in the current offering of products, which we chose to address by visualizing percent deviation of product selling price from the MSRP and profit margin per product line



FactTable: Pricing
Analysis

- ProductCode (F.Key to Product Dimension)
- buyPrice
- MSRP
- priceEach
- Quantity Ordered

In this schema product code links to the "Products" dimension using a foreign key. "buyPrice" represents the price at which the item was purchased by the company from the supplier, MSRP is the suggested retail price by the manufacturer and "priceEach" is the the price at which the item was bought by the customers. By using these measures we are able to determine potential pricing changes that can be made, particularly for the classic cars product line.

References

1) Kaplan, R. S., & Norton, D. P. (1996). The Balanced Scorecard: Translating Strategy into Action. Harvard Business School Press.
2) Parmenter, D. (2015). Key Performance Indicators: Developing, Implementing, and Using Winning KPIs. John Wiley & Sons.
3) Tableau Documentation. (n.d.). Tableau Software.
4) https://www.researchgate.net/publication/221598461_Container-Managed_ETL_Applications_for_Integrating_Data_in_Near_Real-Time
5) Alhassan Mumuni, Fuseini Mumuni, Automated data processing and feature engineering for deep learning and big data applications: A survey, Journal of Information and Intelligence, 2024, ISSN

2949-7159, https://doi.org/10.1016/j.jiixd.2024.01.002.
6) Juan A. Lara, David Lizcano, M. Aurora Martínez, Juan Pazos, Data preparation for KDD through automatic reasoning based on description logic, Information Systems, Volume 44, 2014, Pages 54-72, ISSN 0306-4379, https://doi.org/10.1016/j.is.2014.03.002.

Appendix

Source Code - The end-to-end code can be found in the following GitHub repository.
https://github.com/mmkumar5401/Business_Intelligence.git