* **Sarvesh Patankar**

Capstone Project

Sales Analytics: Northwind Dataset

The Northwind dataset is a sample database representing the sales and operations of a fictitious specialty food import/export company, "Northwind Traders." It comprises several interconnected tables detailing customers, products, orders, order details, employees, suppliers, shippers, and product categories. These tables contain information ranging from customer contact details and product pricing to order shipment information and employee roles. The dataset's structure allows for comprehensive analysis of various aspects of the business, including sales trends, customer behaviour, product performance, and operational efficiency. The project utilizes this data to derive key insights through cleaning, loading into different platforms, answering specific business questions, and creating visualizations to aid in data-driven decision-making for Northwind Traders.

Overview

The Process

**Data Acquisition from GitHub:**

The Northwind dataset was obtained from a GitHub repository, providing a reliable source for the project's data files. The process involved locating and downloading these files, making them ready for analysis.

**MECE Breakdown for Self-Analysis:**

To ensure a comprehensive understanding of the project's scope and my approach, a Mutually Exclusive, Collectively Exhaustive (MECE) breakdown was conducted. This involved systematically categorizing the different aspects of the project to ensure all relevant areas were addressed without overlap, facilitating a structured self-assessment of my work.

**Connecting with SQL and Power BI, Data Transformation & Cleaning:**

The downloaded data was connected to SQL and Power BI. Within these tools, the data was examined for quality, and necessary cleaning and transformations were performed to prepare it for effective analysis and visualization.

**Power BI Problem Solving and Visualization:**

Using Power BI, the provided business questions were addressed. This involved creating calculations, linking data, and developing visuals like charts and maps to clearly present insights for each question.

**SQL Solving and Excel Visualization for EDA:**

SQL queries were used to analyse the data for the exploratory questions. The results were then visualized in Excel using various charts to identify patterns and correlations within the data.

**PowerPoint Presentation:**

Key findings and visualizations from both Power BI and EDA were compiled into a PowerPoint presentation. This aimed to clearly and engagingly communicate the project's core aspects and insights.

**Documentation:**

Finally, a comprehensive document was created to record the entire project process. This included data sources, cleaning steps, problem-solving methods, visualization types, and a summary of the main findings and conclusions.

Objective and Significance

The primary objective of this project was to develop an insightful and visually engaging Power BI dashboard for Northwind Traders. This dashboard aims to effectively communicate key performance metrics related to customer behaviour, sales patterns, and employee performance. By consolidating data from various tables within the Northwind dataset, the project sought to provide a comprehensive understanding of the company's operations.

The significance of this project lies in its potential to revolutionize how Northwind Traders utilizes its data for decision-making. The interactive visualizations and dynamic filters within the Power BI report empower stakeholders to explore data, uncover valuable insights, and ultimately make more informed business decisions. This data-driven approach is crucial for Northwind Traders to maintain a competitive edge and foster business growth within the wholesale market. Furthermore, the exploratory data analysis conducted using SQL and Excel aimed to uncover deeper patterns and correlations that could further inform strategic initiatives.

Data Dictionary

**Customers Table:**

* **CustomerID:** Unique identifier for each customer.
* **CompanyName**: The name of the customer's organization.
* **ContactName:** The name of the individual contact at the customer.
* **ContactTitle:** The job title of the contact person.
* **Address:** The street address of the customer**.**
* **City:** The city where the customer is located.
* **Region:** The region or state where the customer is located.
* **PostalCode:** The postal code of the customer's location.
* **Country:** The country where the customer is located.
* **Phone:** The customer's phone number**.**
* **Fax**: The customer's fax number.

**Employees Table:**

* **EmployeeID:** Unique identifier for each employee.
* **LastName:** The last name of the employee.
* **FirstName:** The first name of the employee.
* **Title:** The employee's job title.
* **TitleOfCourtesy:** The courtesy title used for the employee (e.g., Mr., Ms.).
* **BirthDate:** The date the employee was born.
* **HireDate:** The date the employee was hired.
* **Address:** The employee's street address.
* **City:** The city where the employee lives.
* **Region**: The region or state where the employee lives.
* **PostalCode:** The postal code of the employee's residence.
* **Country:** The country where the employee lives.
* **HomePhone:** The employee's home phone number.
* **Extension**: The employee's office phone extension**.**
* **Notes:** General notes or information about the employee.
* **ReportsTo:** The EmployeeID of the employee's direct manager.

**Orders Table:**

* **OrderID:** Unique identifier for each order**.**
* **CustomerID:** The identifier of the customer who placed the order.
* **EmployeeID:** The identifier of the employee who handled the order.
* **OrderDate:** The date the order was placed.
* **RequiredDate:** The date the order was requested to be delivered.
* **ShippedDate:** The date the order was actuallyshipped.
* **ShipVia:** The identifier of the shipping company used.
* **Freight:** The shipping charges for the order.
* **ShipName:** The name to be used for shipping.
* **ShipAddress:** The shipping street address.
* **ShipCity:** The shipping city.
* **ShipRegion:** The shipping region or state.
* **ShipPostalCode:** The shipping postal code.
* **ShipCountry:** The shipping country.

**Order Details Table:**

* **OrderID:** The identifier of the order this detail belongs to.
* **ProductID:** The identifier of the product included in this order detail.
* **UnitPrice:** The price per unit of the product at the time of the order.
* **Quantity:** The number of units of the product ordered.
* **Discount:** The discount applied to this product in this order.

**Products Table:**

* **ProductID:** Unique identifier for each product.
* **ProductName:** The name of the product.
* **SupplierID:** The identifier of the company that supplies this product**.**
* **CategoryID:** The identifier of the category this product belongs to.
* **QuantityPerUnit:** The quantity of the product in each sales unit.
* **UnitPrice:** The selling price per unit of the product.
* **UnitsInStock:** The number of units of the product currently in stock.
* **UnitsOnOrder:** The number of units of the product currently on order from suppliers.
* **ReorderLevel:** The minimum stock level before reordering is needed.
* **Discontinued:** Indicates whether the product has been discontinued (1 for yes, 0 for no).

**Suppliers Table:**

* **SupplierID:** Unique identifier for each supplier.
* **CompanyName**: The name of the supplier's company.
* **ContactName:** The name of the primary contact person at the supplier.
* **ContactTitle:** The job title of the contact person.
* **Address:** The supplier's street address.
* **City:** The city where the supplier is located.
* **Region:** The region or state where the supplier is located.
* **PostalCode:** The postal code of the supplier'slocation.
* **Country:** The country where the supplier is located.
* **Phone:** The supplier's phone number.
* **Fax:** The supplier's fax number.
* **HomePage:** The supplier's website or homepage URL.

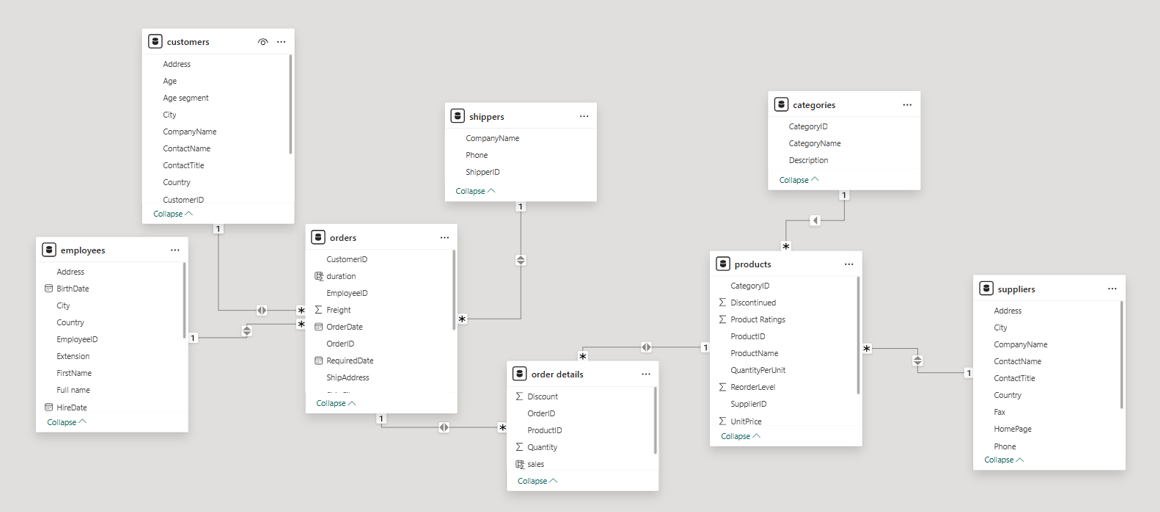
**Shippers Table:**

* **ShipperID:** Unique identifier for each shipper.
* **CompanyName:** The name of the shipping company.
* **Phone:** The shipping company's phone number.

**Categories Table:**

* **CategoryID:** Unique identifier for each category.
* **CategoryName:** The name of the product category.
* **Description:** A brief description of the product category.

Data Model – ER Diagram

****

Additionally, more calculated columns and measures are added in the data model during the solving of problem statements, if necessary.

Power BI dashboard and problem statements

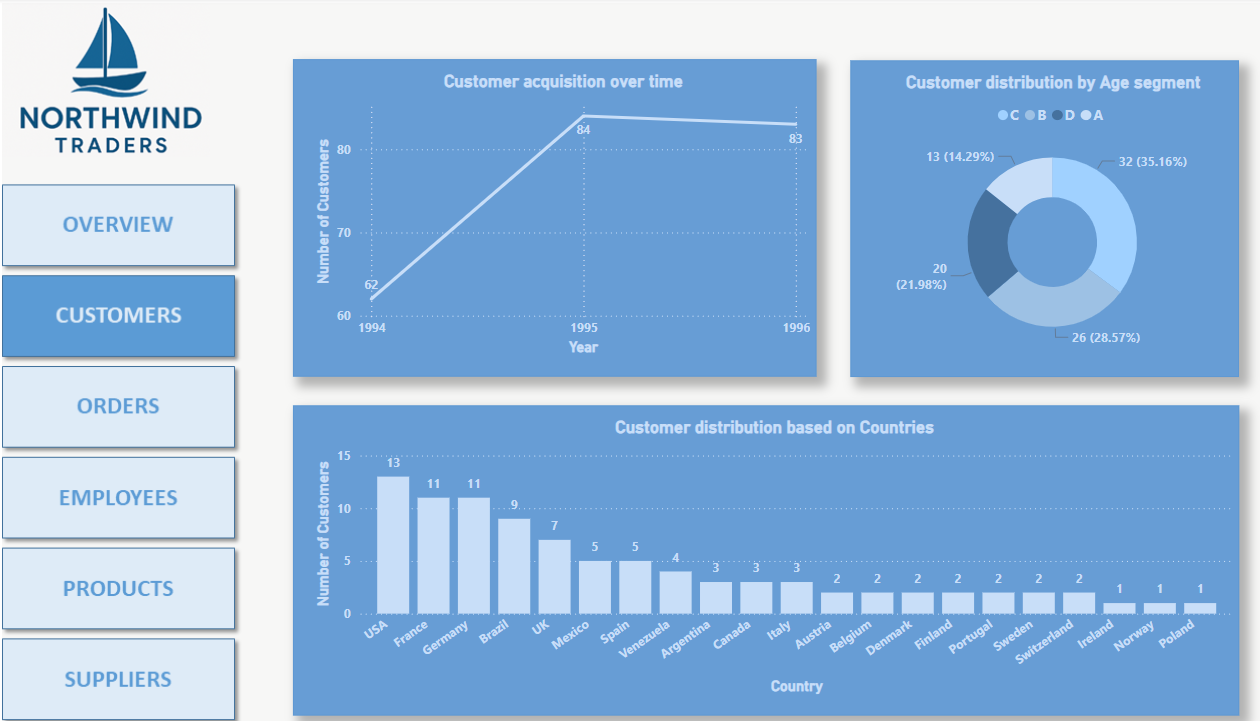
To really make sense of the Northwind data and uncover valuable insights, I designed an interactive Power BI dashboard that's both easy to use and packed with information. Imagine it as a well-structured digital storybook, divided into six clear chapters, each just a click away thanks to simple navigation buttons. The first page gives you the big picture, showing overall performance through key metrics like total sales and the number of orders we've handled. Then, you can delve into five specific areas – think of them as focused lenses on Customers, Orders, Employees, Products, and Suppliers. I've taken the 15 questions provided and organized them within these relevant sections, using clear and engaging visuals like charts and maps to bring the data to life and answer those questions directly. My goal was to create a dashboard that not only presents the data but also tells a compelling story, empowering anyone who uses it to understand Northwind Traders' business at a glance and make smarter decisions.

# **Overview Page –**

****

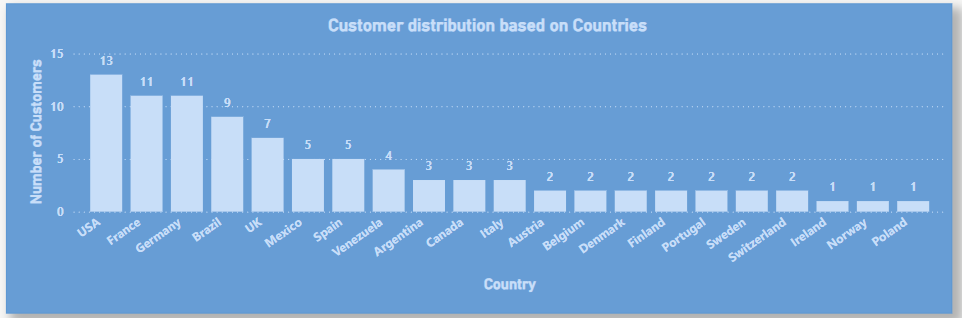
The Overview page of the Power BI dashboard provides a high-level snapshot of Northwind Traders' key performance indicators. At the top, slicers for 'Category,' 'Country,' and 'Date' enable users to filter the data for a more focused analysis across the entire dashboard. The central section displays aggregated metrics, including 'Total Sales,' 'Total Orders,' 'Total Employees,' 'Total Customers,' 'Total Suppliers,' and 'Total Available Products,' offering an immediate understanding of the company's scale and activity. Complementing these KPIs are visualizations for 'Top 5 Countries by Sales,' presenting a geographical perspective on revenue generation, and 'Sales by Year,' illustrating the trend of sales performance over time. This page serves as an initial point of reference, providing a comprehensive yet concise summary of Northwind's overall business performance.

# **Customers Page -**

****

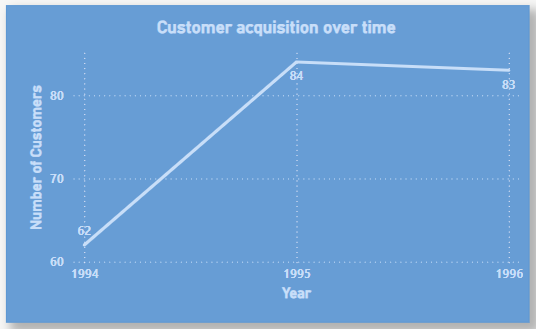
This is our customers page, where three problem statements are solved. The problem statements are -

**1. How does customer distribution vary across different regions or customer segments? Can we visualize it on a map or bar chart?**

****

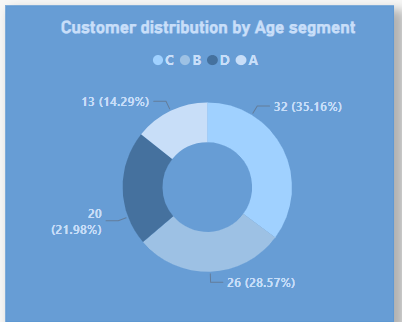
This vertical bar chart, 'Customer distribution based on Countries,' visually compares the number of Northwind Traders' customers across different countries. It was created by taking the **Customers table** and counting the distinct **CustomerID** for each **Country**. The x-axis represents the different countries, and the y-axis displays the count of customers. The varying heights of the bars directly indicate the customer concentration in each country. This highlights key markets for focused marketing and identifies potential areas for expansion. Analyzing this distribution also provides insights into regional business performance and informs resource allocation strategies.

**2. What is the trend in customer acquisition over time? Can we create a line chart or area chart to display it?**

****

The 'Customer acquisition over time' line chart tracks the yearly trend of unique customers acquired by Northwind Traders. It was generated by taking data from the **Orders table**, extracting the year from the **OrderDate**, and then counting the distinct **CustomerID** for each year. The x-axis represents the year, and the y-axis shows the number of new customers acquired in that year. The line connecting these points reveals the company's customer growth trajectory, the impact of past initiatives, and aids in forecasting future growth and identifying key acquisition periods for strategic learning.

**3. Can we visualize the distribution of customer demographics such as age, gender, or income using histograms or pie charts?**

****

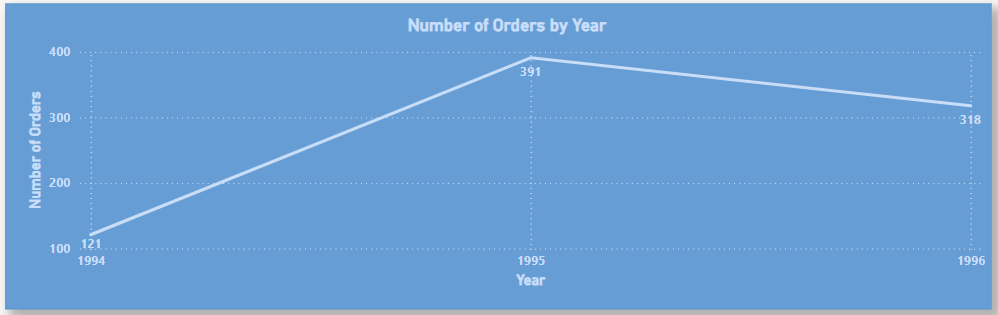
The 'Customer distribution by Age segment' donut chart illustrates the proportion of customers within four distinct age groups (A, B, C, D). This visual was created by first adding a new 'Age' column with random values to the **Customers table**. Subsequently, an 'Age segment' column was created using a conditional formula (*if [Age] <= 30 then "A" else if [Age] <= 40 then "B" else if [Age] <= 50 then "C" else "D"*). The donut chart then displays the count of **CustomerID** for each 'Age segment' as a percentage of the total customer count. While the underlying age data is not real, this demonstrates the potential for demographic understanding to inform segmentation and targeted marketing if real age information were available.

# **Orders Page –**

# 

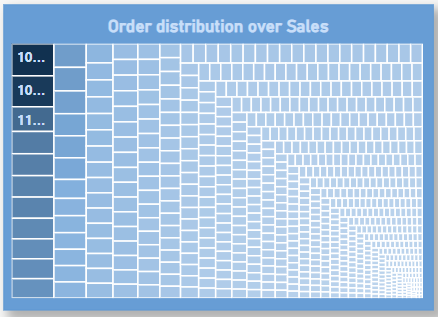
This is the orders page , 3 problem statements are solved here as well -

# **4. How does order volume change over time? Can we create a time series chart or stacked bar chart to visualize it?**

****

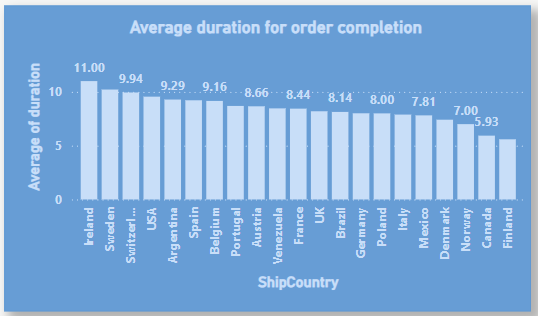
This line chart, titled 'Number of Orders by Year,' illustrates the trend of order volume over the years present in the dataset. It was created by taking data from the **Orders table**, extracting the year from the **OrderDate**, and then counting the number of **OrderID** for each year. The x-axis represents the year, and the y-axis displays the total number of orders placed in that year. The line connecting these points visually demonstrates how the order volume has fluctuated over time, highlighting periods of high or low order activity. This helps Northwind understand its order trends and potentially correlate them with external factors or internal initiatives.

**5. What is the distribution of order values? Can we create a histogram or box plot to display it?**

****

The treemap, titled 'Order distribution over Sales,' visualizes the distribution of sales for individual orders. It was generated by using the **Order Details table** to calculate the 'Sales' for each order. The 'Sales' was a calculated column derived using the formula*: Sales = 'Order Details'[Quantity] \* 'Order Details'[UnitPrice] \* (1 - 'Order Details'[Discount])*. The treemap then displays each **OrderID** as a rectangle, with the size of the rectangle proportional to the total 'Sales' value for that order. This visualization allows for a quick identification of high-value orders and provides an overview of the distribution of order sizes.

# **6. Can we visualize the average order processing time or shipping duration using a bar chart or box plot?**

****

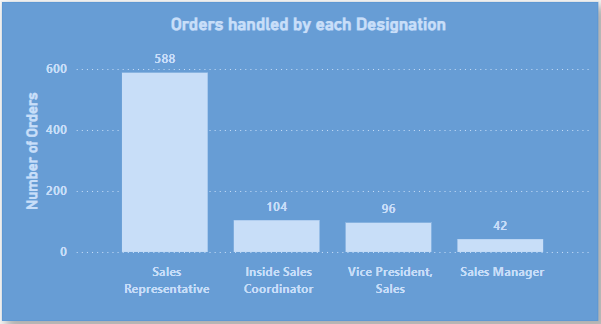
The bar chart, titled 'Average duration for order completion,' displays the average number of days it took to complete an order (from order date to shipped date) for each country. This was calculated using the **Orders table** by creating a new column 'Duration' with the formula: *Duration = DATEDIFF(Orders[OrderDate], Orders[ShippedDate], DAY)*. The chart then shows the average of this 'Duration' for all orders originating from each unique country in the **Orders table**. The x-axis represents the country, and the y-axis represents the average duration in days. This visualization allows Northwind to compare the efficiency of order processing and shipping across different geographical regions.

# **Employees Page –**

****

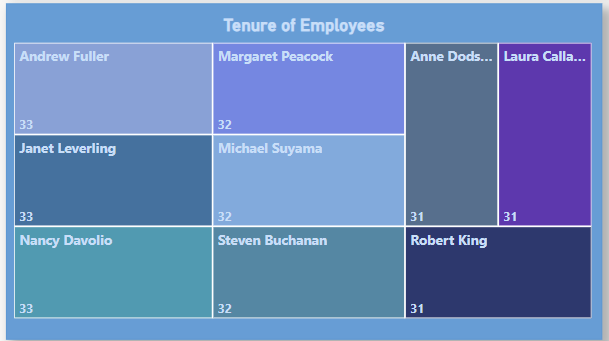
This is the Employees page , 3 problem statements are solved here, which are -

**7. How does employee productivity vary across different departments or job roles? Can we create a stacked bar chart or grouped column chart to visualize it?**

****

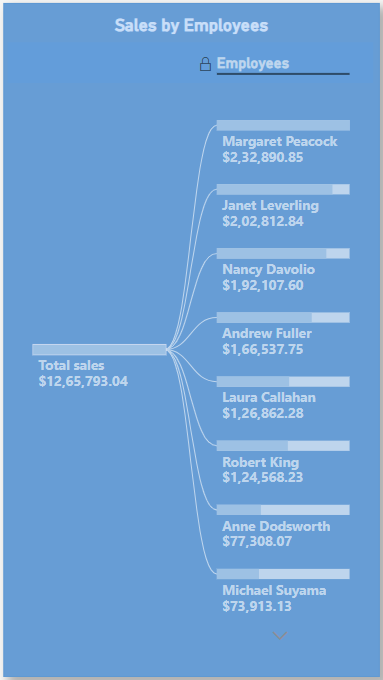
This bar chart, displays the number of orders handled by each employee job role. It was created by using the Title column from **Employees table** to determine job roles, and then counting the **OrderID** from the **Orders table** associated with each employee. The x-axis represents the different job roles, and the y-axis shows the number of orders. This visualization allows for a comparison of productivity across different roles, highlighting which roles handle the most orders and potentially indicating areas where workload distribution or efficiency could be improved. We have defined productivity as number of orders handled as it is a key factor in defining productivity.

**8. What is the distribution of employee tenure? Can we create a histogram or box plot to display it?**

****

The treemap, titled 'Tenure of Employees', visualizes the distribution of employee tenure in years. It was generated by using the **Employees table** and calculating a 'Tenure' column. The 'Tenure' was a calculated column derived using the formula: *Tenure = DATEDIFF(Employees[HireDate], TODAY(), YEAR)*. The treemap then displays each employee as a rectangle, with the size of the rectangle representing their tenure in years. This visualization provides an overview of employee longevity within the company, highlighting both long-term and recently hired employees.

**9. Can we visualize employee performance ratings or KPIs using a radar chart or bullet graph?**

****

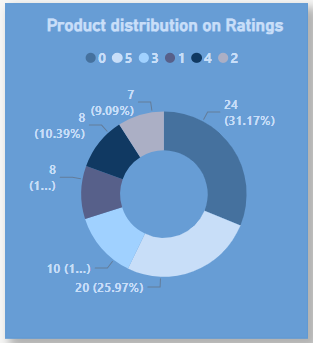
The decomposition tree, visualizes the total sales and how it breaks down across individual employees. It was created by using the **Orders table** and **Order Details table** to calculate the total sales for each employee. The root of the tree represents the total sales for the company, and it branches down to show the contribution of each employee to that total. This visualization provides a clear picture of individual employee sales performance and their relative contribution to the overall company revenue.

# **Products Page –**

****

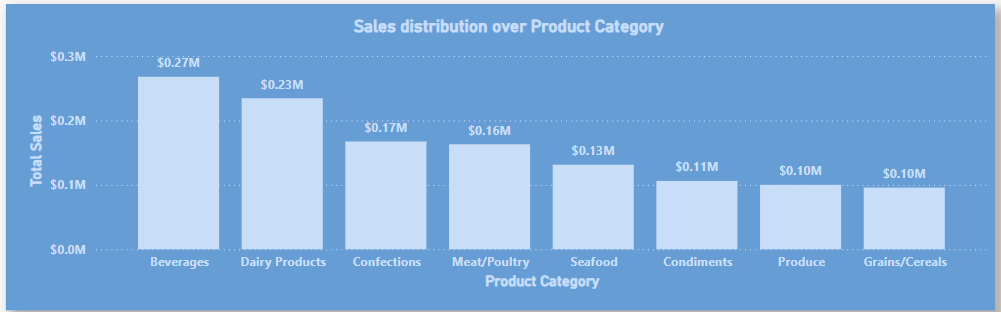
This is the Products page, 3 problem statements are solved here, which are -

**10. What is the distribution of product ratings or reviews? Can we create a histogram or stacked bar chart to visualize it?**

****

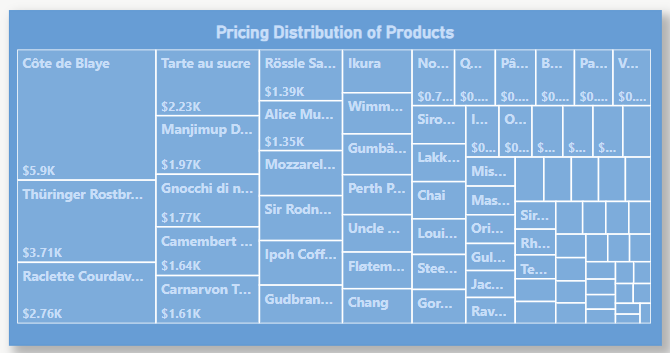
The donut chart, displays the count of products for each defined product rating. This visual was created by first generating a **'Product Rating'** column in the **Products table**. This rating was conditionally created based on the **'ReorderLevel'** column, assumed to be directly proportional to the rating. The donut chart then shows the count of **ProductID** for each distinct **'Product Rating'**. This provides a visual distribution of product ratings based on their reorder levels, potentially indicating the perceived popularity or demand for different products.

**11. How does the sales volume vary across different product categories? Can we create a bar chart or treemap to display it?**

****

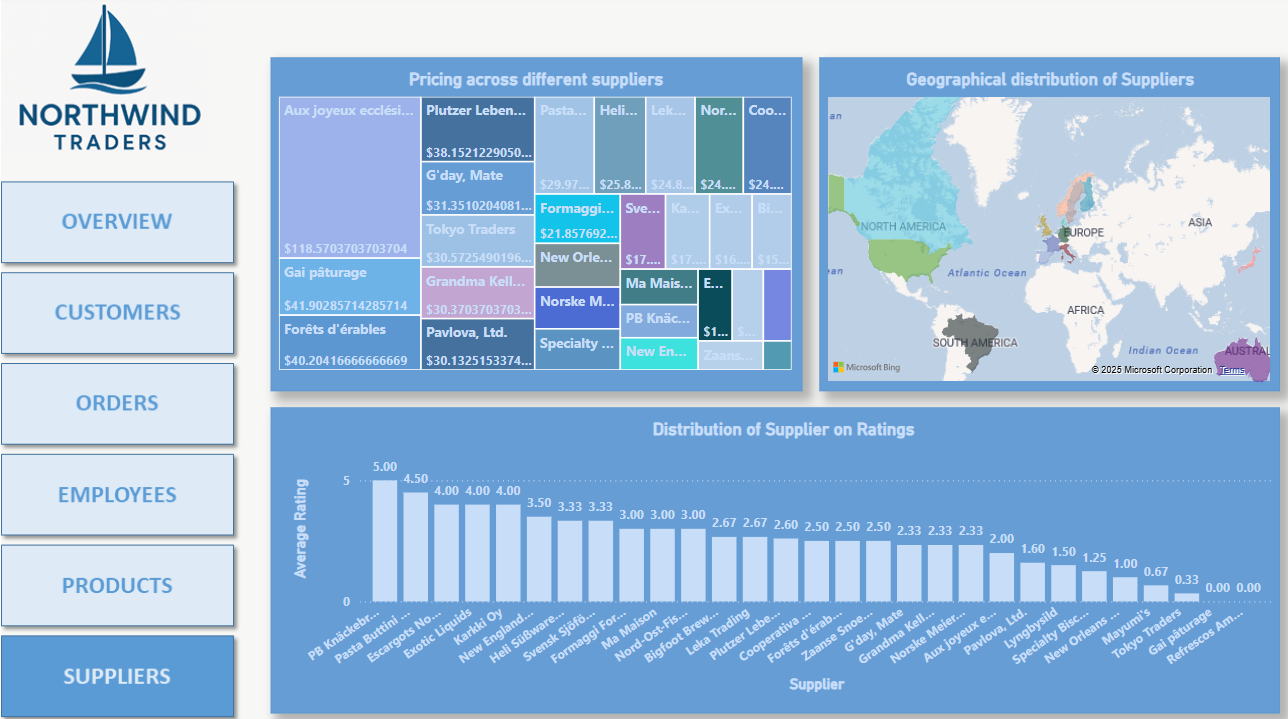
This bar chart, titled 'Sales distribution over Product Category,' illustrates the sales volume for each product category. It was created by linking the **Categories table** to the **Order Details table** through the **Products table** (ProductID links Order Details and Products, CategoryID links Products and Categories). The 'Sales' for each order detail line was calculated ('Order Details'[Quantity] \* 'Order Details'[UnitPrice] \* (1 - 'Order Details'[Discount])), and then aggregated by 'CategoryName' from the **Categories table**. The x-axis represents the product categories, and the y-axis displays the total sales for each category. This visualization allows Northwind to identify its top-performing product categories in terms of sales revenue.

**12. Can we visualize the pricing distribution of products using a box plot or violin plot?**

****

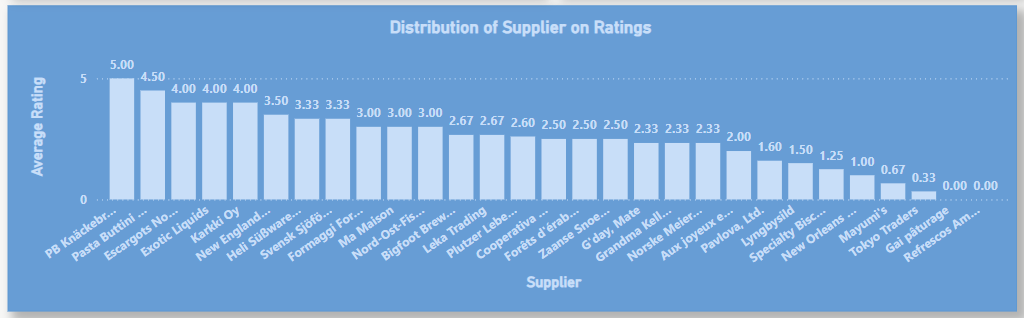
The treemap, titled 'Pricing Distribution of Products,' visualizes the distribution of product prices. It was generated by using the **Products table** and displaying each product as a rectangle, with the size of the rectangle representing its 'UnitPrice' from UnitPrice column. This visualization provides a clear overview of the pricing range and distribution across Northwind's product catalog, allowing for a quick understanding of the price points of different products.

# **Suppliers Page –**

****

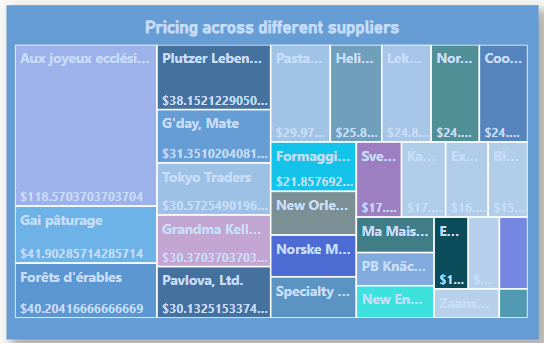
This is the Products page, last 3 problem statements are solved here, which are -

**13. What is the distribution of supplier ratings or performance metrics? Can we create a bar chart or radar chart to visualize it?**

****

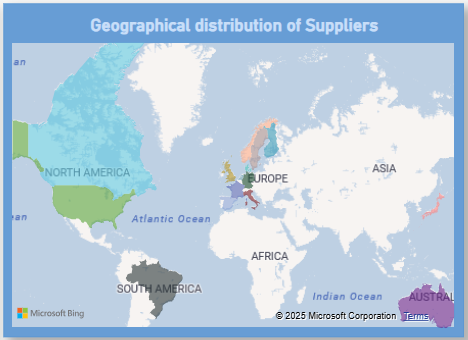
The bar chart, displays the average supplier ratings for different suppliers. To derive this, the 'Product Rating' (created based on 'ReorderLevel' in the **Products table**) was used as a proxy for supplier rating, under the assumption that supplier performance influences product rating. The average of these 'Product Ratings' was then calculated and grouped by 'SupplierID' from the **Suppliers table** (linked via the **Products table**). The x-axis represents the suppliers, and the y-axis displays their average rating. This visualization provides a comparative view of supplier performance based on the indirect metric of product ratings.

**14. How does the cost or pricing structure vary across different suppliers? Can we create a box plot or stacked bar chart to display it?**

****

The treemap, visualizes the average unit price of products supplied by each supplier. It was generated by linking the **Suppliers table** to the **Products table** using 'SupplierID' and then calculating the average 'UnitPrice' for products associated with each supplier. The treemap displays each supplier as a rectangle, with the size potentially representing the number of products they supply or another relevant metric, and the label indicating the average unit price of their products. This allows for a comparison of the general pricing levels of different suppliers.

**15. Can we visualize the geographical distribution of suppliers using a map or bubble chart?**

****

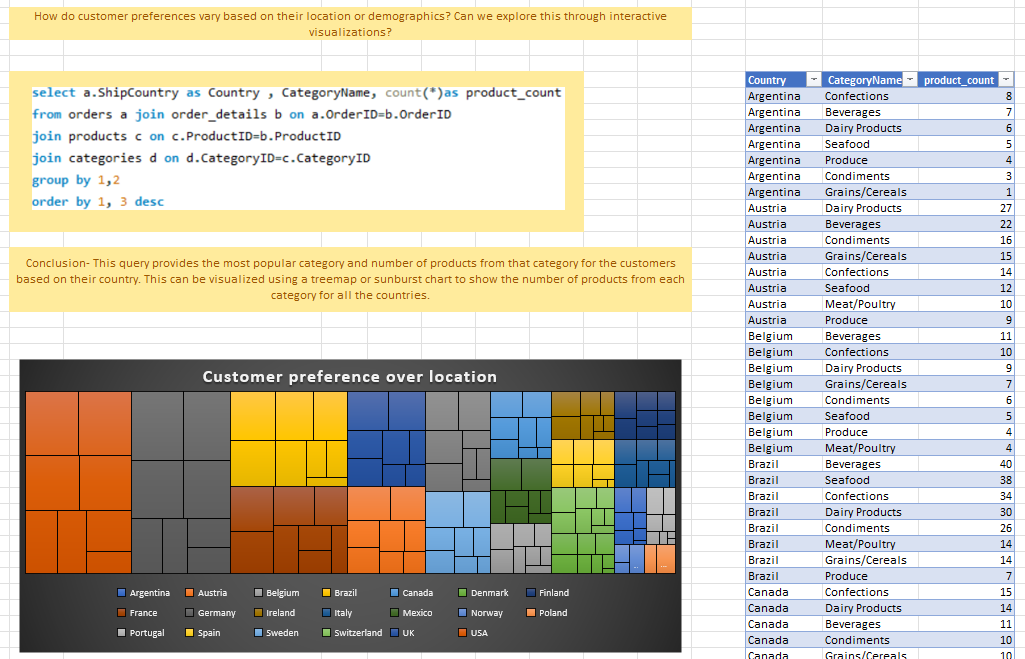
The map, titled 'Geographical distribution of Suppliers,' displays the countries where Northwind Traders' suppliers are located. This visualization was created by using the 'Country' information from the **Suppliers table** to plot the locations on a world map. The map highlights the geographical spread of the current supplier network. An observation noted is the limited presence of suppliers in Asia and Africa, suggesting a potential opportunity for Northwind Traders to explore partnerships in these continents, which could potentially lead to reduced shipping costs and delivery times.

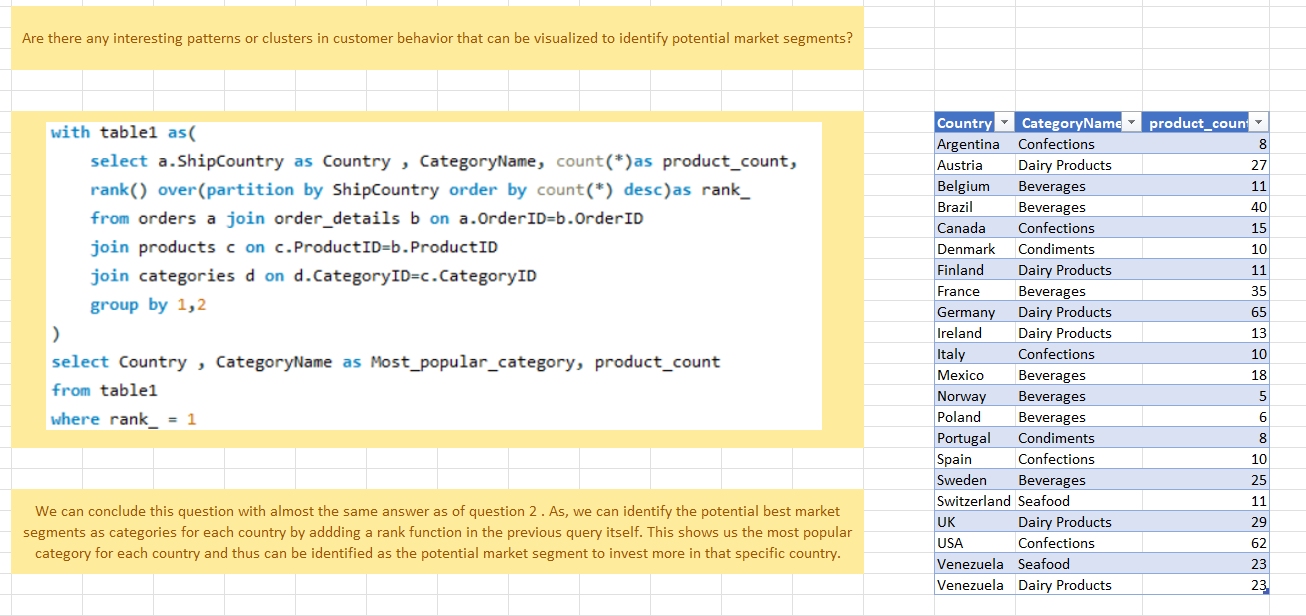
Beyond the structured problem-solving in Power BI, an Exploratory Data Analysis (EDA) was conducted using SQL to delve deeper into the Northwind dataset and uncover potential hidden patterns and relationships. The following section outlines the 15 EDA questions that were investigated. For each question, the SQL queries employed to extract and analyse the relevant data are described, followed by an explanation of the Excel visualizations created to represent the findings. Finally, a summary of the key insights gleaned from each exploration is provided, offering a more nuanced understanding of the underlying dynamics within the Northwind business data.

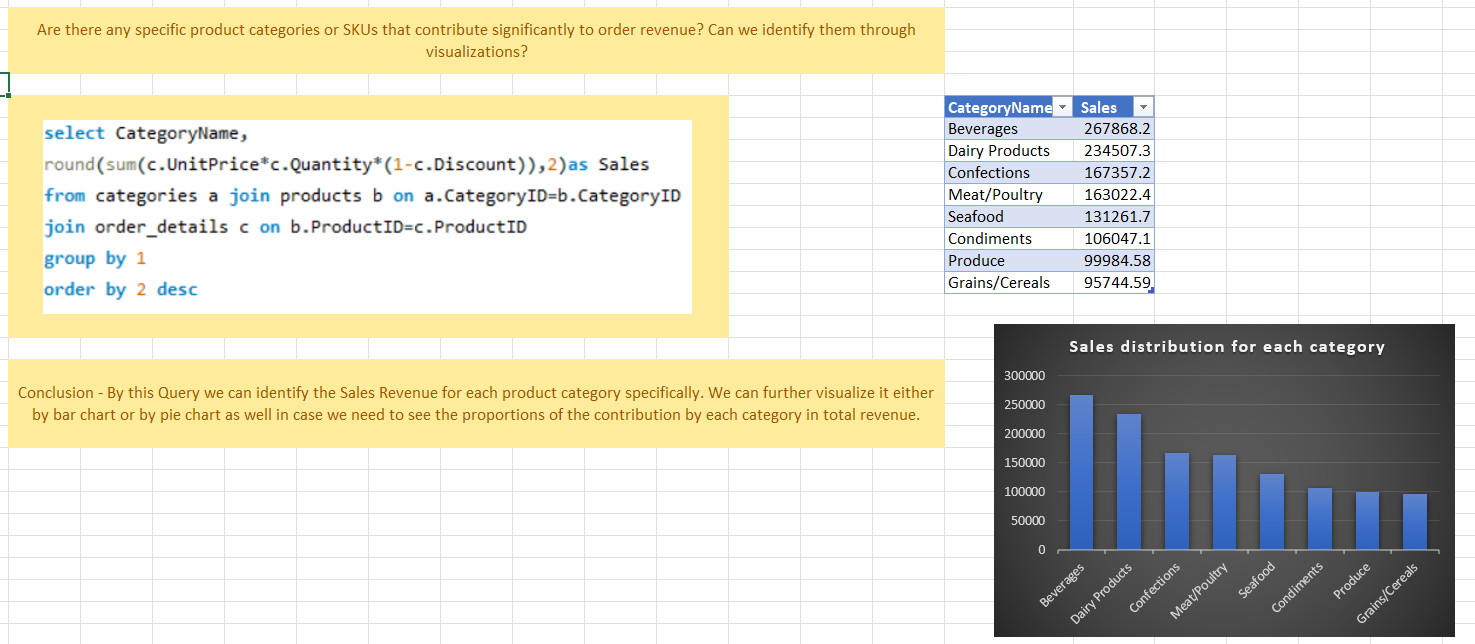
EDA Problem Statements

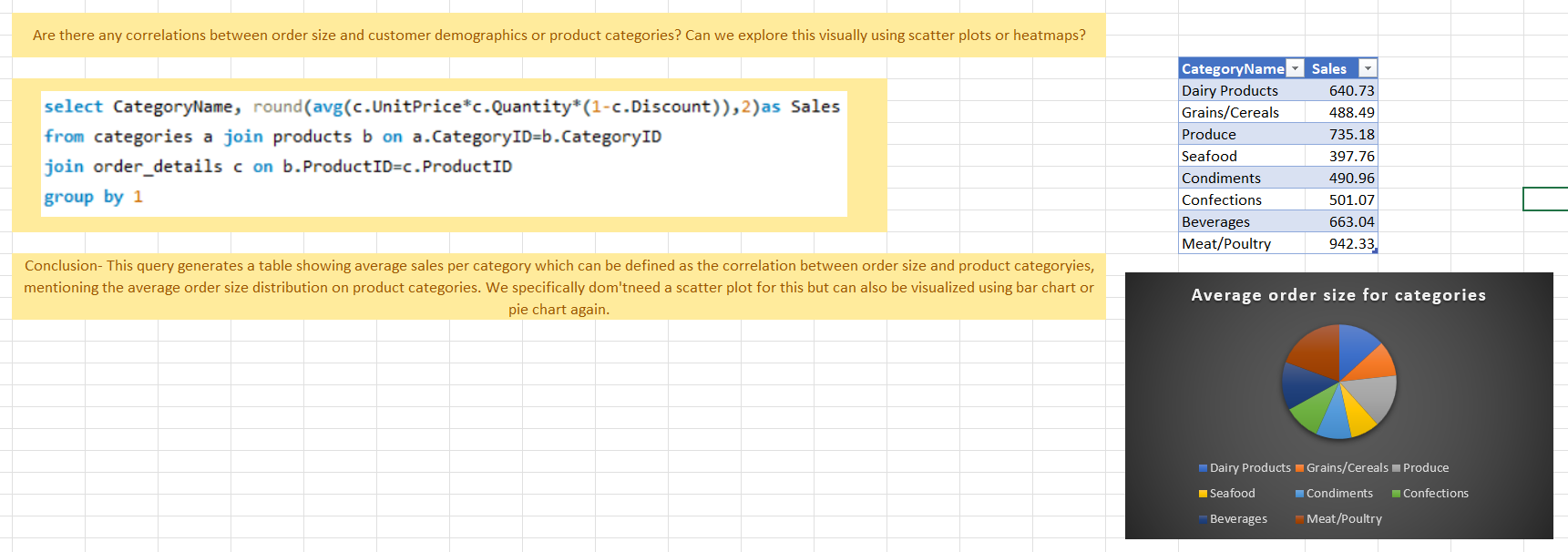
Below are the 15 EDA Solutions –

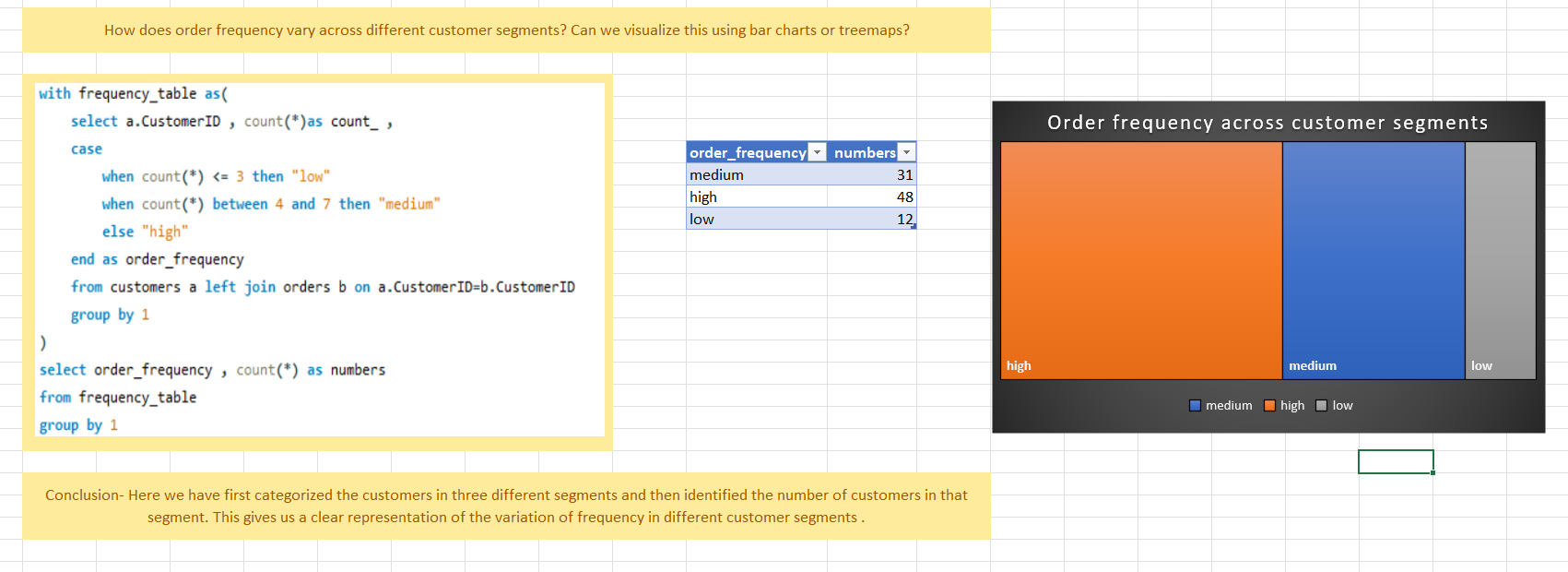
****

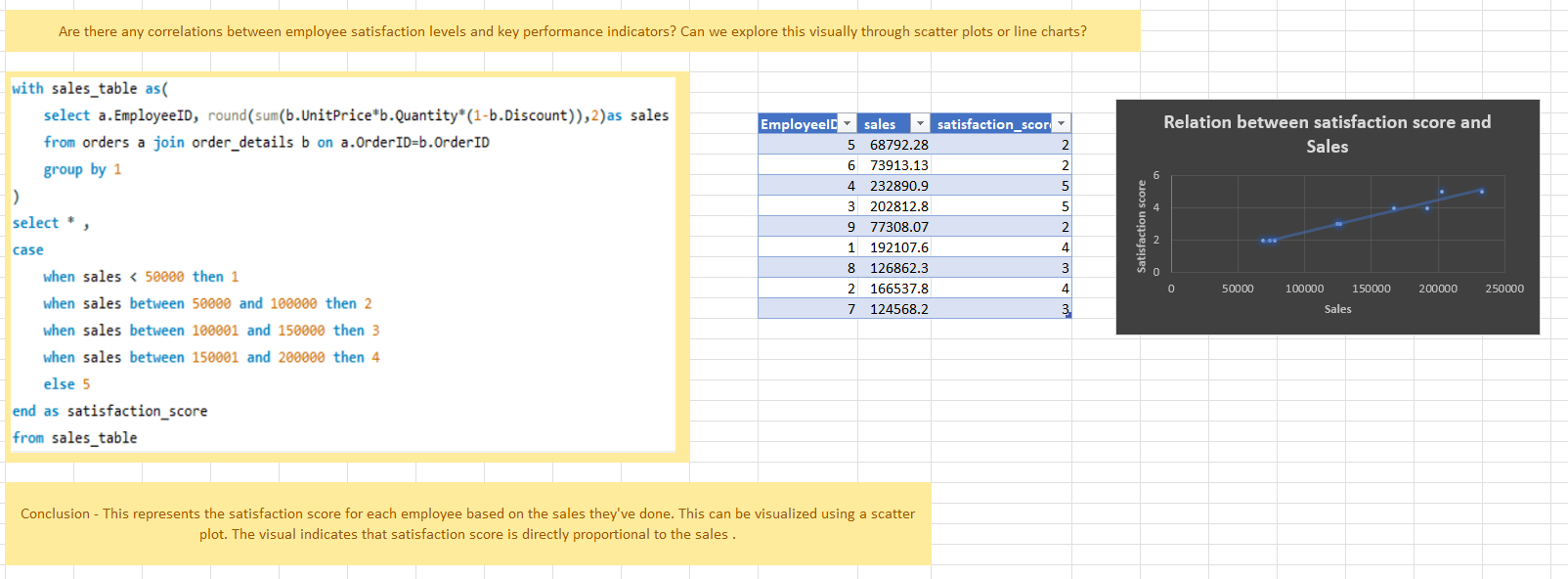


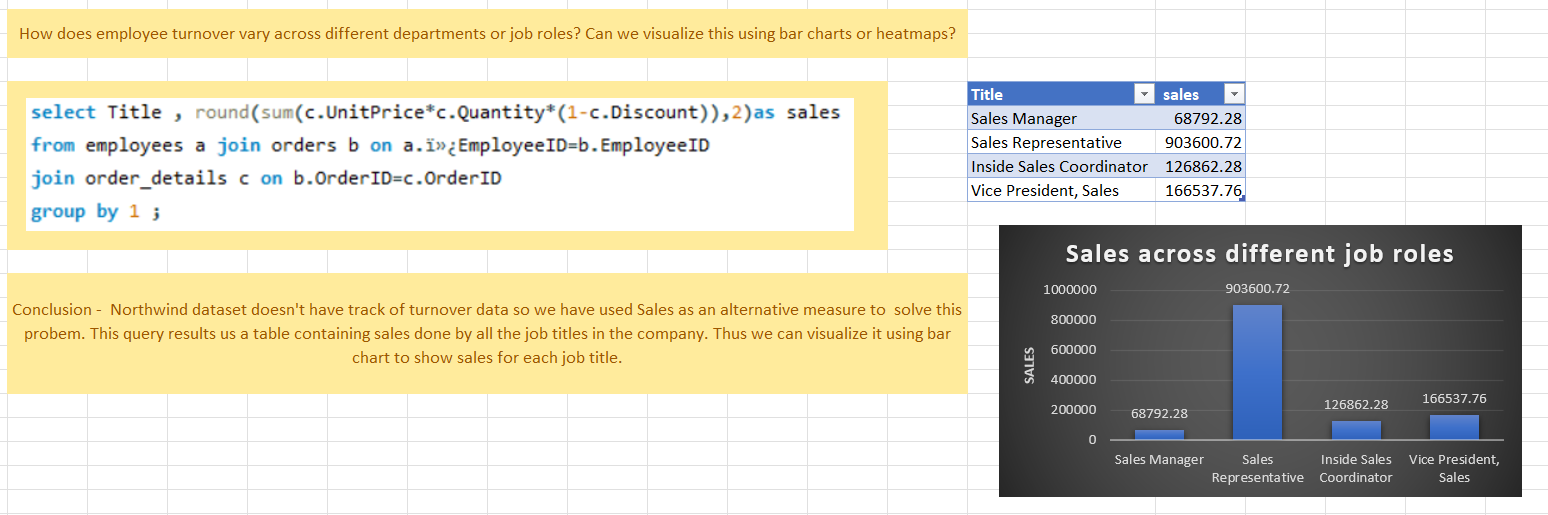


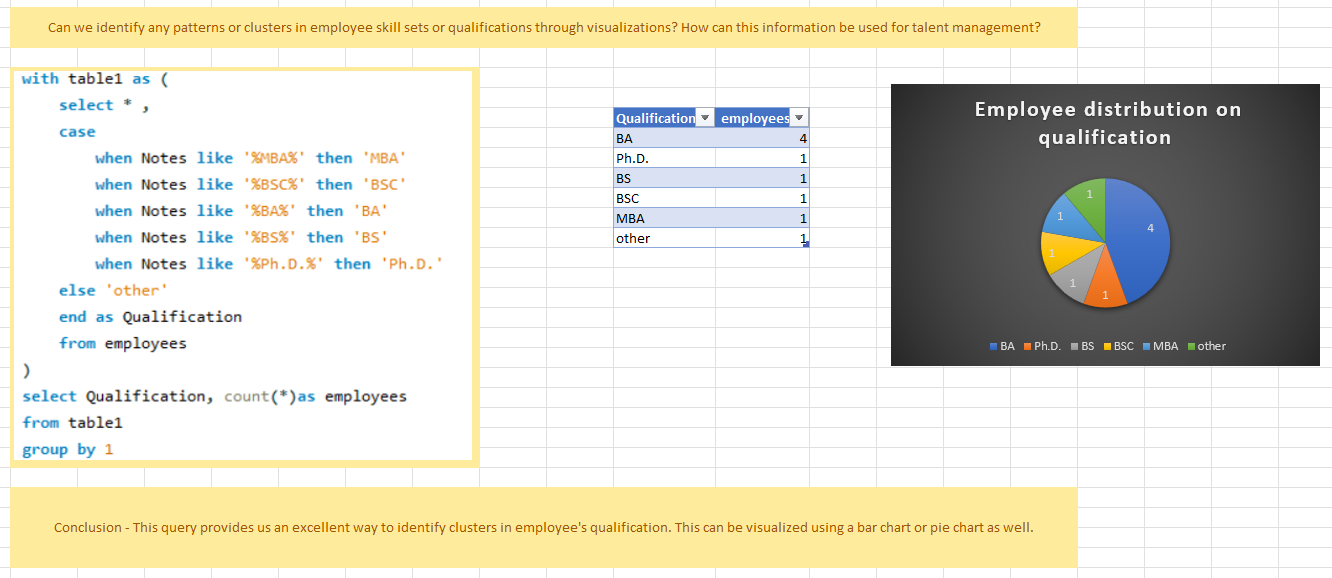


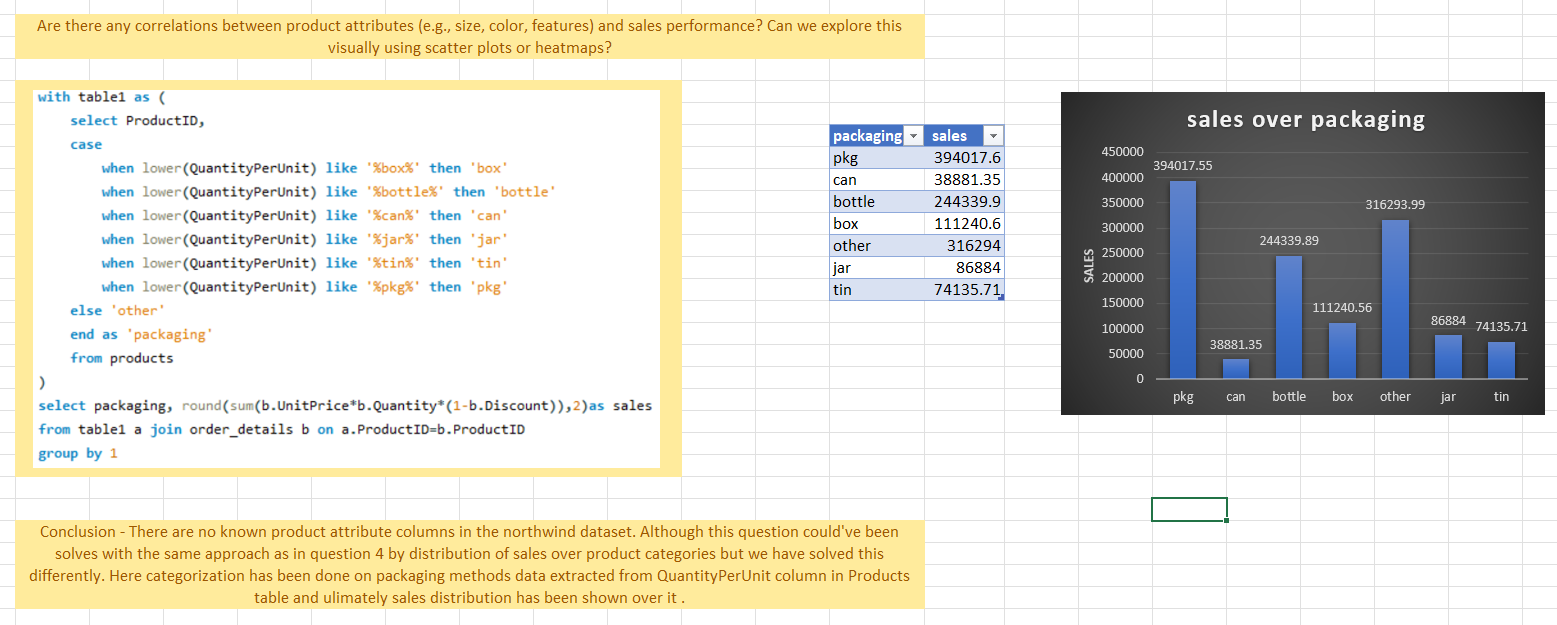


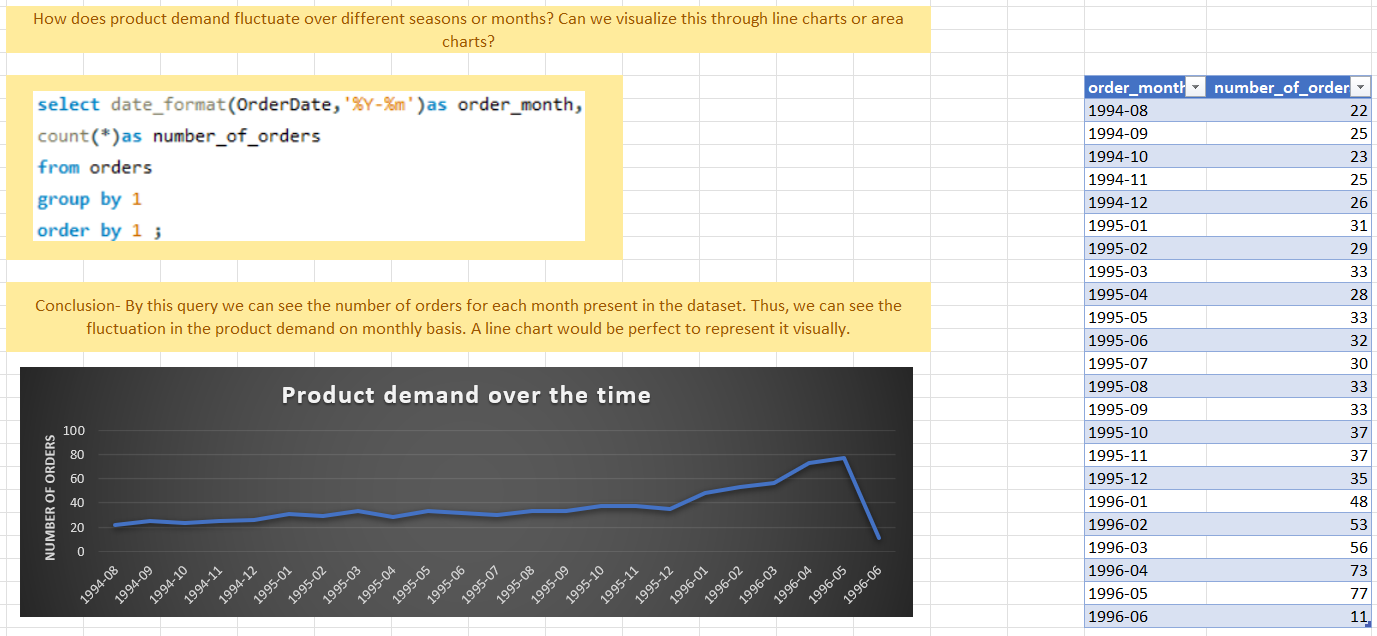


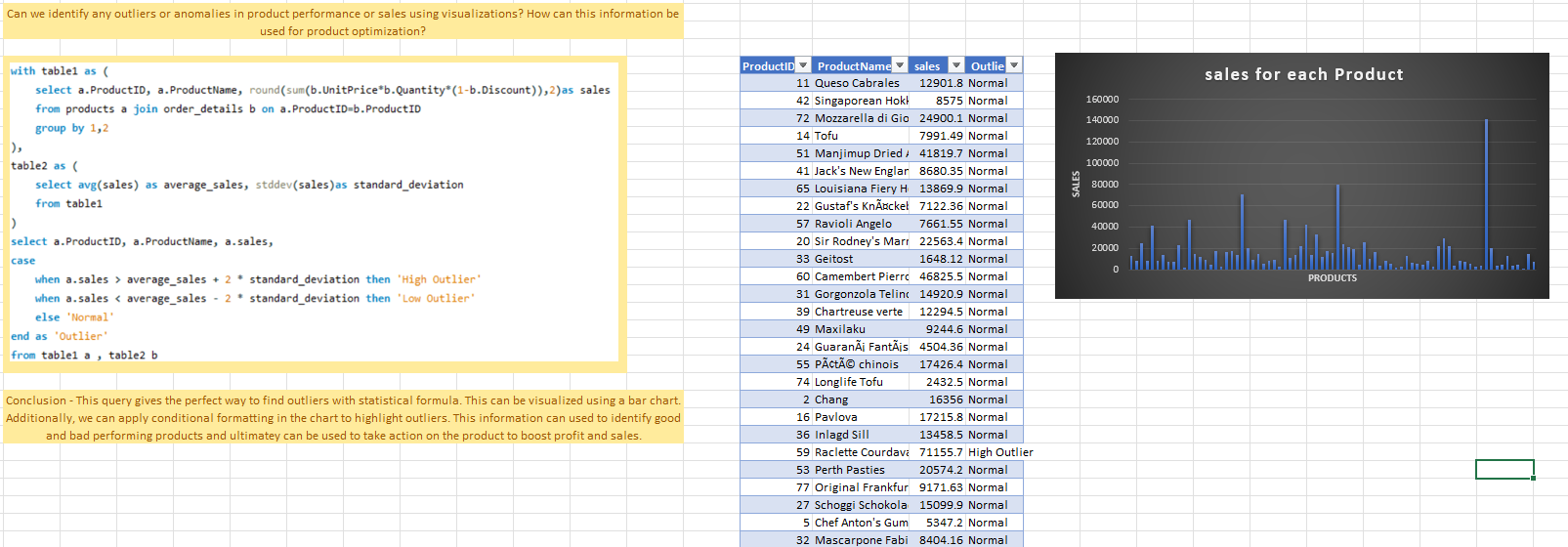


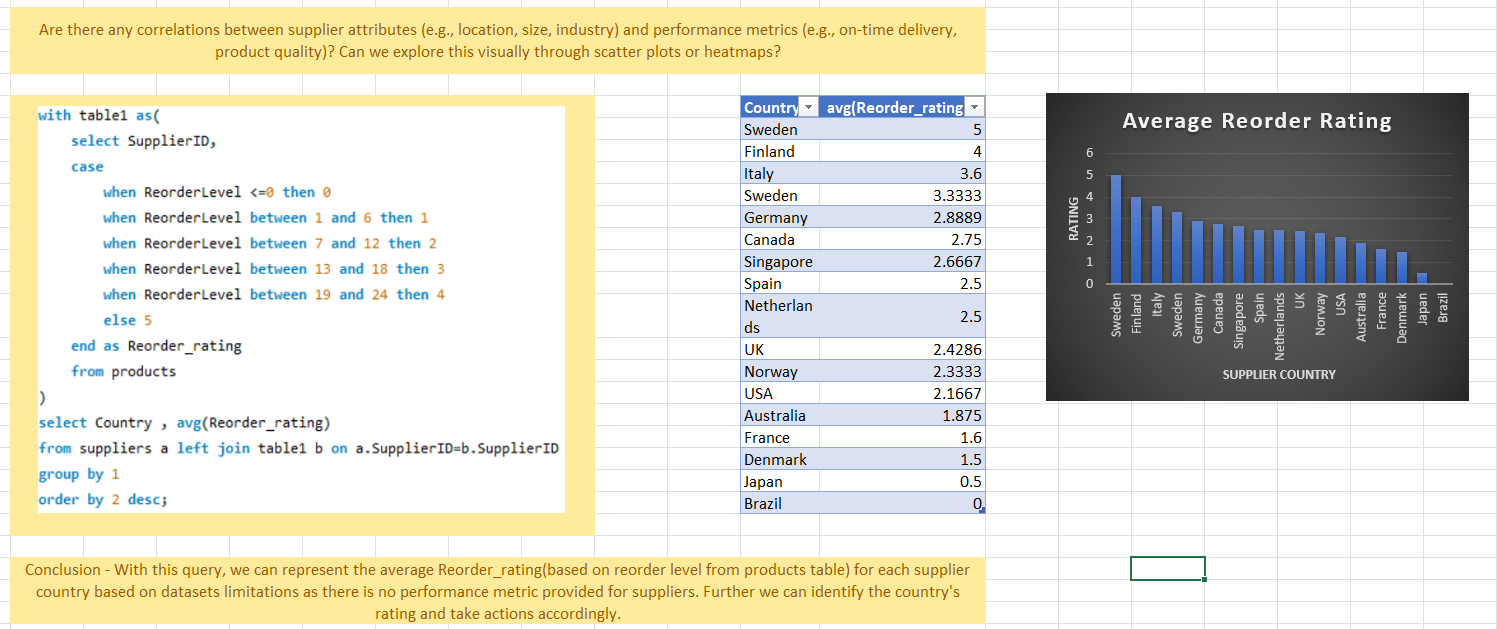


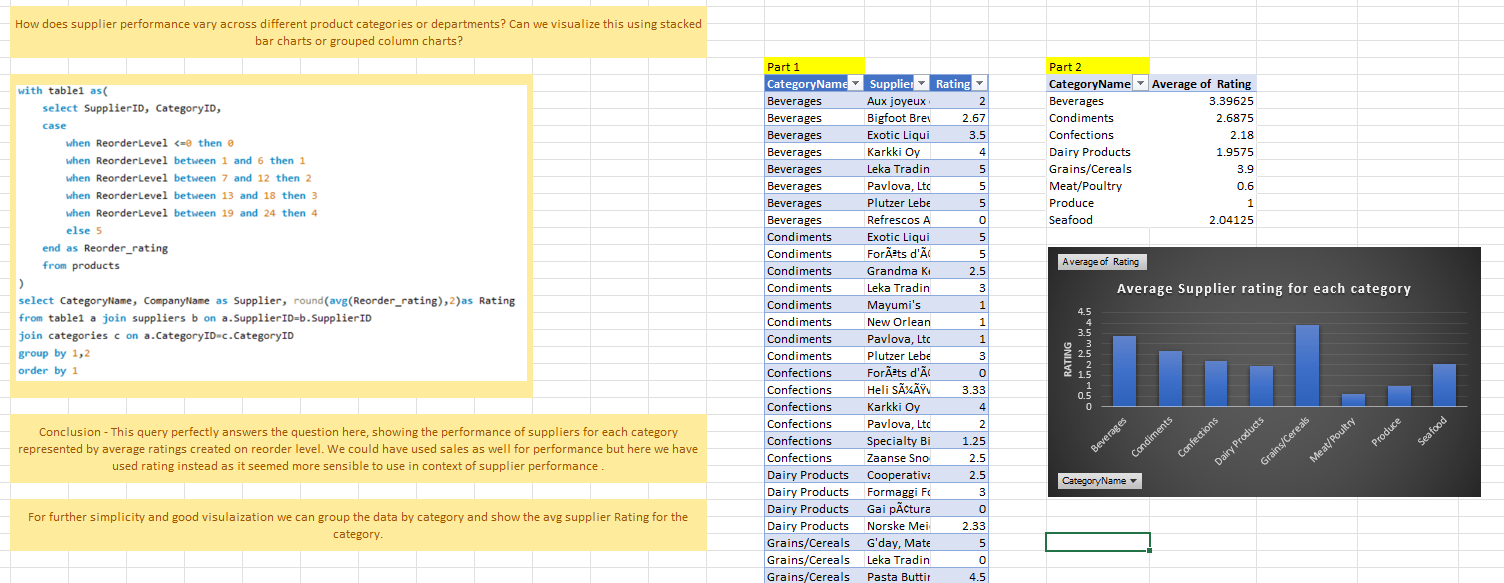








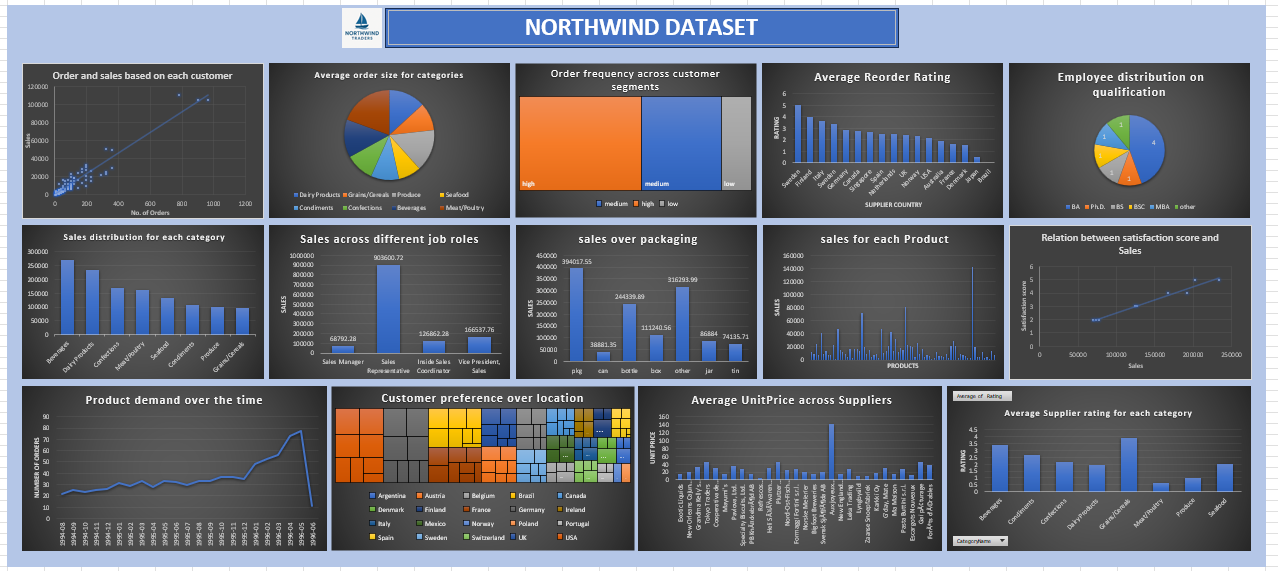






Excel Dashboard

A supplementary Excel dashboard was created to visually represent the findings from the Exploratory Data Analysis (EDA) conducted using SQL. This dashboard compiles the various charts and graphs generated to address the 15 EDA questions. Each visual on the dashboard corresponds to a specific exploration, providing a consolidated view of the patterns, correlations, and distributions uncovered within the Northwind dataset during the EDA process.



The PowerPoint presentation accompanying this project serves to communicate the key problem statements, the methodologies employed to solve them using Power BI and SQL, and the resulting insights in a clear and accessible format. Designed for enhanced readability and presentation, the slides incorporate the problem statements directly alongside the corresponding visualizations and concise summaries of the findings. This format aims to facilitate a better understanding of the analysis and its conclusions for a broader audience.

Presentation