# **Capstone Project – Sales Analysis**

**Shiraj Ali Ahmad**

Data Analytics

# **Overview**

The dashboard provides meaningful insights into the various aspects of the market which focuses on customer analysis, sales analysis, employee analysis, order analysis, shipper analysis and supplier analysis. Customer analysis provides customer-related metrics such as location, number of orders etc. Sales analysis provides customer-related metrics such as total sales over the period, top-selling products in terms of sales etc. Employee analysis provides information about employees such as job roles, location etc. Order analysis provides information like monthly orders, time to process orders etc. Shipper analysis gives detailed information about the shipper such as the time takes to deliver the order, the total number of shippers working for the company etc. Last but not least supplier analysis, provides information about the quantity and the price of the products.

# **Process**

1. Data acquisition from GitHub:

I acquired the dataset from the GitHub repository which contains essential data in SQL and CSV files that will help us to perform the desired visualisation. The data consists of multiple tables like customer, employee, order, order details, products, suppliers and shipper.

1. Data transformation and Enhancement:

I have to transform the data when and where necessary. So, I have a measure table to find the total sales, total customers and total order count made by the customers. I have also extracted the data from the table wherever necessary by the use of calculated columns example: To find the start of the month, to find sales of the product etc.

1. Connecting With Tools:

To connect with data I have used SQL and CSV files. I have imported the SQL files into the MySQL server to perform various operations by generating queries to find the answers related to EDA problem statements. For Power BI problem statements we have to use the get data function key which helps to import the data in the Power BI. I have used CSV files to get connected with data in Power BI and then transform it with the help of power query. Making the use of ER diagram I have connected the table to get the desired data.

# **Objective**

The objective of this project is to conduct an analysis that helps the NorthWind traders find

customer behaviour, sales patterns, and employee performance to aid decision-making processes. The main aim of this project is to find insights that communicate key performance metrics for Northwind Traders effectively. The process involves extracting the data from the traders and performing various tasks such as cleaning the data and transformation, making visually appealing charts etc. By offering valuable analysis and insights that will empower stakeholders to make data-driven decisions.

# **Significance**

The significance of Northwind traders is to provide the user with a real-world business environment as the data closely resembles real-world business operations. This will enable the user to gain experience on how to analyse and find the insights of the business data. The business data contains various tables like customers, products, employees, shippers, suppliers, and orders with help of these tables users can find the relationship between them which further helps in making engaging visualizations.

The Northwind traders dataset has many various aspects of business scenarios such as sales analysis, customer analysis, employee analysis, order analysis and supplier analysis that help the users to make effective dashboards that help the stakeholders find the insights quickly.

# **Data Dictionary**

Table: Categories

* CategoryID: Unique Identifier for each category.
* CategoryName: Name of each category.
* Description: Description of the categories.

Table: Customers

* CustomerID: Unique identifier for each customer.
* CompanyName: Company name the customers are working for.
* ContactName: Name of the customers.
* ContactTitle: The position they hold in their respective company.
* Address: Address of the customers.
* City: The city to which the customers belong.
* Country: Country of the customers
* Phone: Phone number of the customer.

Table: Employees

* EmployeeID: Unique Identifier for each employee.
* Nameofemployee: A calculated column emerged by last name and first name of the employee.
* Title: Department of the Employee.
* BirthDate: Birth date of the employee.
* HireDate: Date when the employee got hired.
* Address: Address of the employee.
* City: The city to which the employee belongs.
* PostalCode: Post address of employee.
* Country: Country where employees come from.
* HomePhone: Phone number of employee.
* TitleofCourtesy: Title given to employee.
* EmployeeTenure: A calculated column. (DATEDIFF(employees[HireDate],TODAY(),YEAR)) Tenure completed by the employee.

Table: Order Details

* OrderID: A foreign key referencing OrderID from the Orders table.
* ProductID: A foreign key referencing ProductID from the product table.
* UnitPrice: Price for each single product.
* Quantity: Number of product quantities purchased.
* Discount: Discount given on the orders.
* Sales: A calculated column, (unitprice\*Quantity) to find sales for each order.

Table: Orders

* OrderID: Unique identifier for each order.
* CustomerID: A foreign key referencing CustomerID from the Customer table.
* EmployeeID: A foreign key referencing EmployeeID from the Employee table
* OrderDate: The date on which the order was made.
* RequiredDate: The Date on which the order should be placed.
* ShippedDate: The date on which the order was placed.
* ShipVia: A foreign key referencing ShipperID from the shipper table.
* ShipAdress: Address where the order is shipped.
* ShipCountry: The country where the order is shipped.
* Startofyearoforderdate: Calculated column to find the start of the order date in date format
* Startofmonthoforderdate: Calculated column to find the start of the month of the order date in date format.
* Yearoforderdate: Calculated column to find the year of the order date.
* Monthoforderdate: Calculated column to find the month of the order date.
* Monthofshippeddate: Calculated column to find when the order is shipped.
* Duration date of shipped: Calculated column to find days required to ship the orders when the order was placed.

Table: Products

* ProductID: Unique identifier for each product.
* ProductName: Name of the products.
* SupplierID: A foreign key referencing SupplierID from the Supplier table.
* CategoryID: A foreign key referencing CategoryID from the Category table.
* QuantityPerUnit: Number of quantities per unit.
* UnitPrice: price of a single unit of product.
* UnitsInStock: Units present in stocks.
* UnitsOnOrder: units we can order.
* Discontinued: The products that are no longer available.

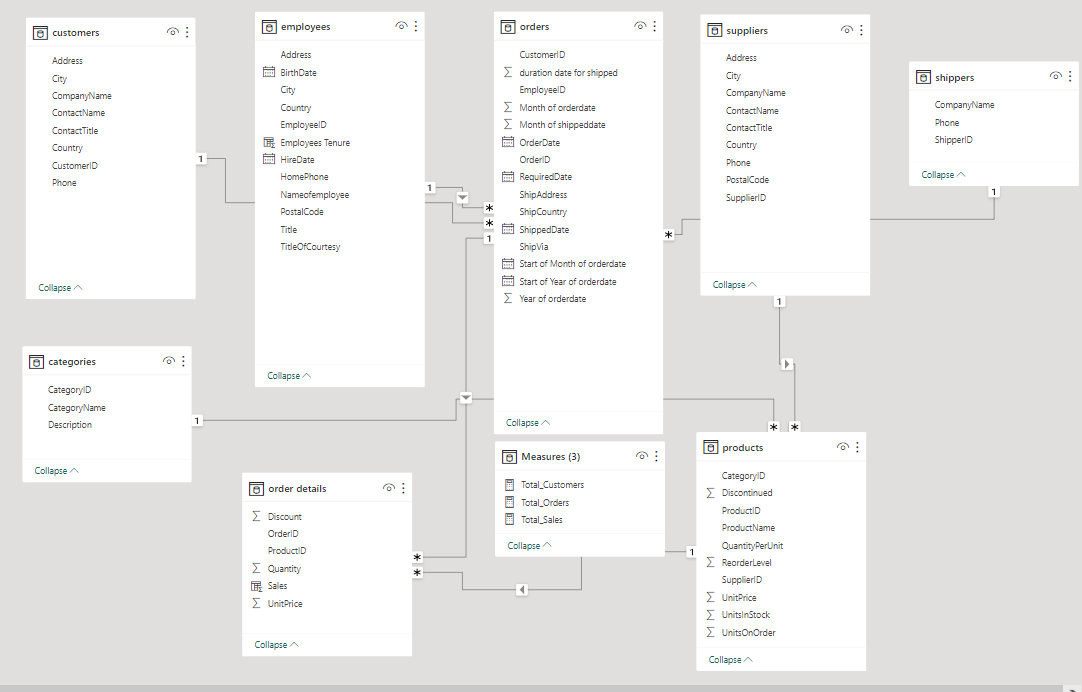
Table: Shippers

* ShipperID: Unique identifier for each shipper.
* CompanyName: Name of the company.
* Phone: Phone number of shipper.

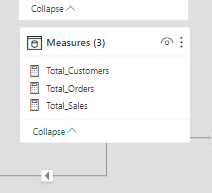
Table: Suppliers

* SupplierID: Unique identifier for each supplier.
* CompanyName: name of company.
* ContactName: name of the supplier.
* ContactTitle: Department supplier works in.
* Address: Address of supplier.
* City: City supplier lives in.
* PostalCode: post address of the supplier.
* Country: Country supplier belongs to.
* Phone: Phone Number of supplier.

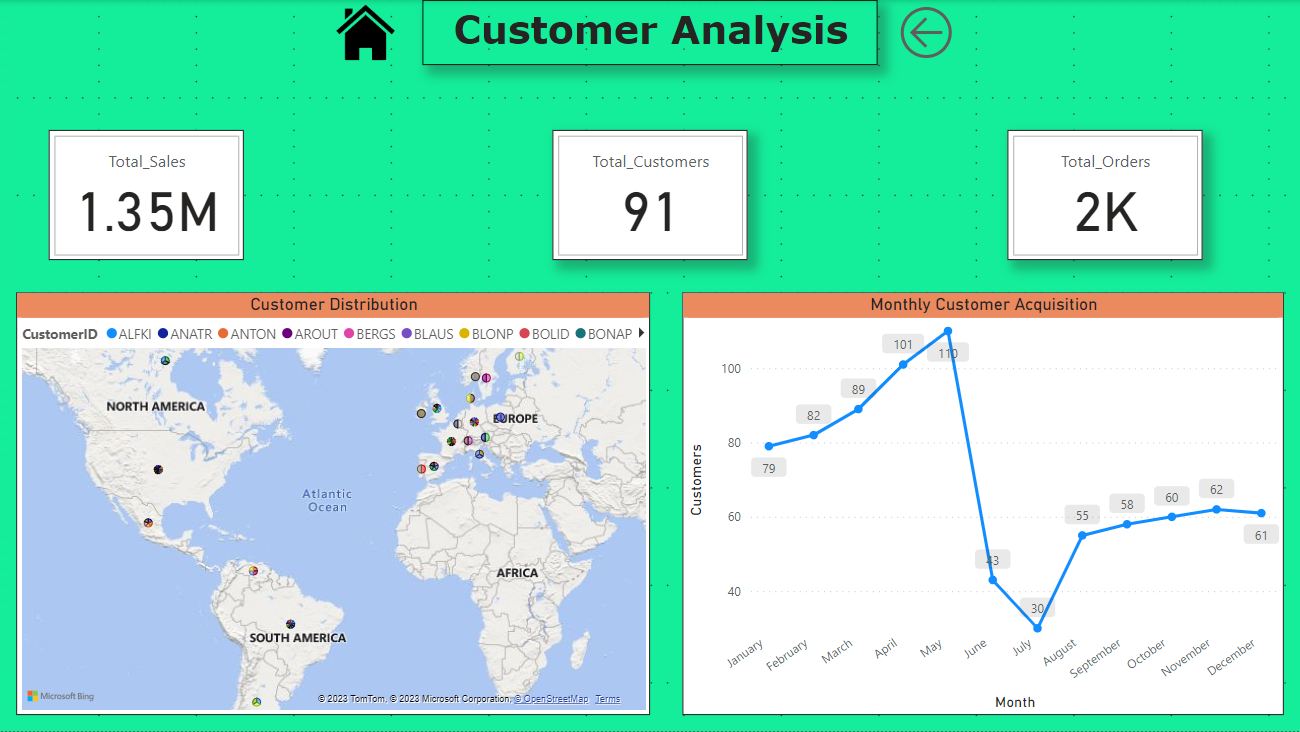
# **ER Diagram**



I have made a measure table which includes total sales, count of total orders and count of total customers. This makes Calculation more easy.

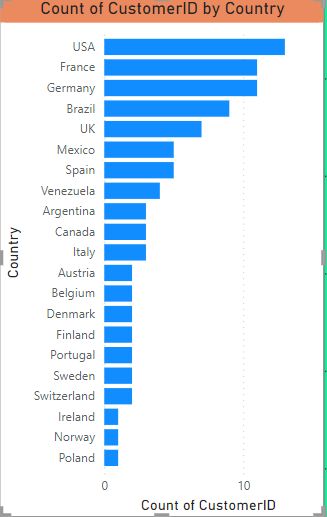
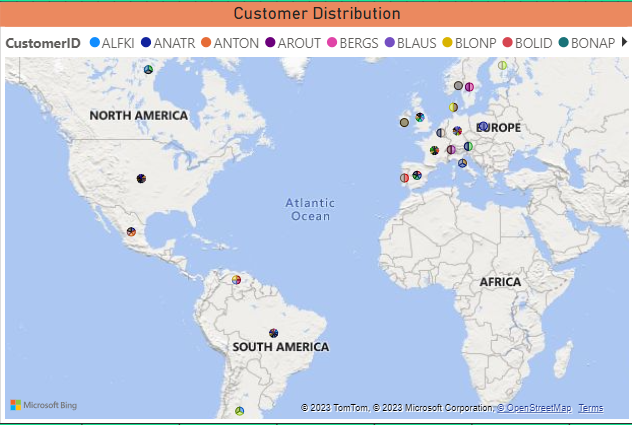


# **Power BI problem Statements**



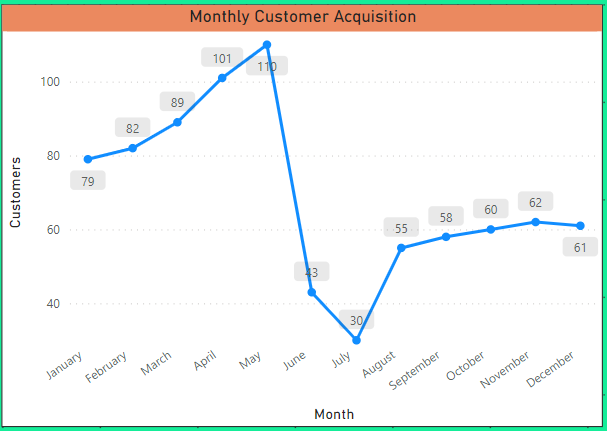
**How does customer distribution vary across different regions or customer segments?**

Analysing this visualization we can identify that countries like the USA, France, Germany, Brazil and the UK have a maximum number of customers located which says that customers are well aware of our company and like to order from our company. Whereas countries like Ireland, Norway and Poland have a minimum number of customers which means that the people of these countries are not aware of the company.



**What is the trend in customer acquisition over time?**

From this Visualization, we can identify that in the first 5 months that is from January to May number of customers increased but in June and July, the numbers drastically decreased which reflects that customers don’t order products from the company which affects customer acquisition and from august to december customer acquisiyion is decent.



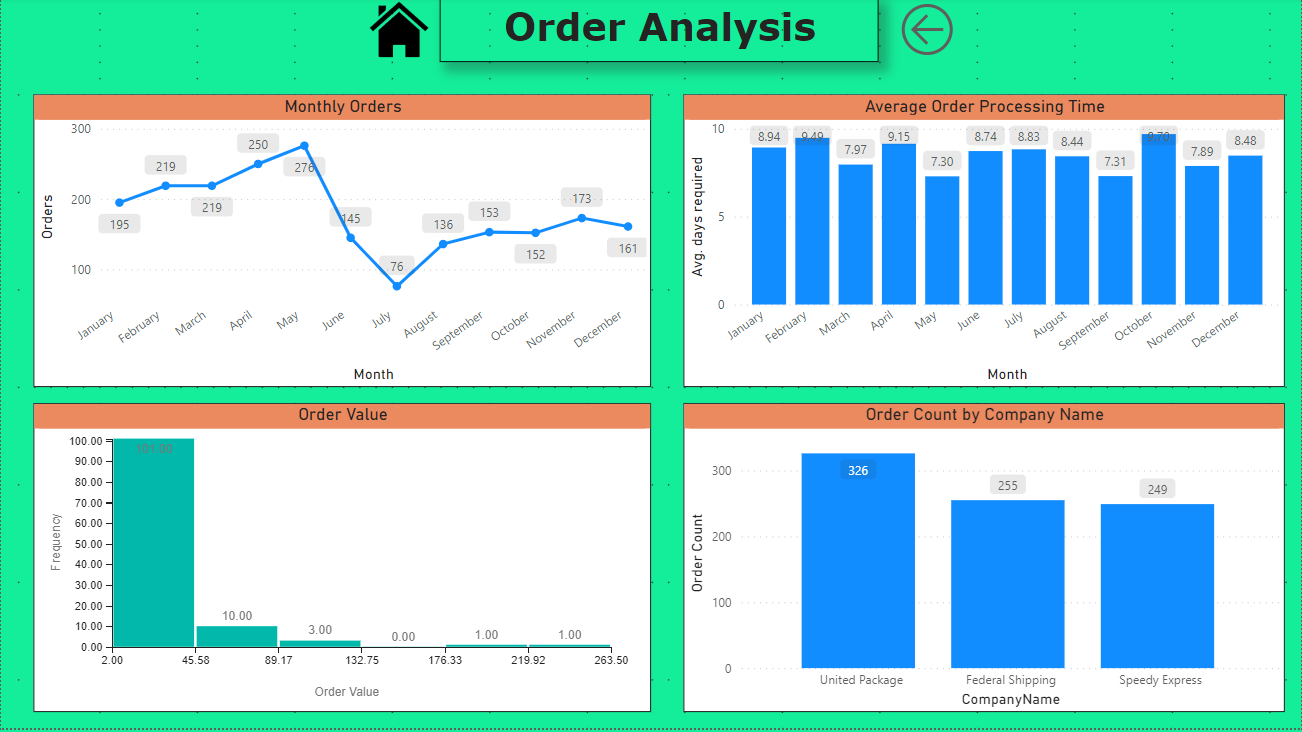
**Can we visualize the distribution of customer demographics such as age, gender, or income?**

By analysing the data we come to know that we don’t have data related to age, gender or income hence we won’t be able to get any insights.

**Can we find the KPI related to customers?**

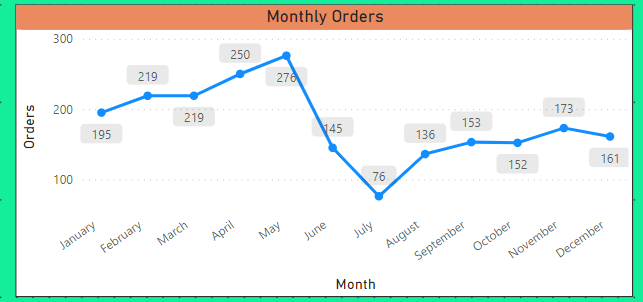
As we see in the below visualization we come to know that it’s a start-up company with a total number of customers of just 91 most of which are from countries like the USA and France which reflects that the company is not that famous. Total sales are also pretty decent as compared to other market companies.





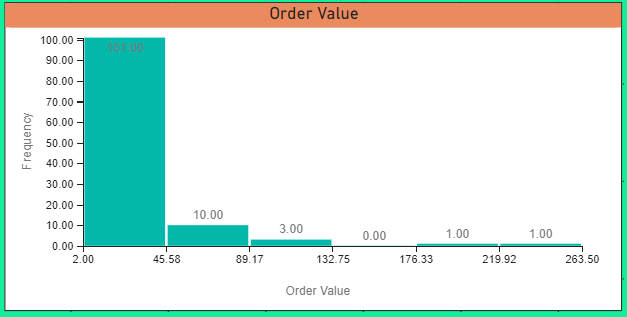
**How does order volume change over time?**

As we can see, from January to May number of customer orders increases and decreases as we move forward. May has the maximum number of order and July have the minimum order volume. This can be because of many reasons like customers ordering the product in bulk in the starting phase of the year hence they don’t have to buy them again.



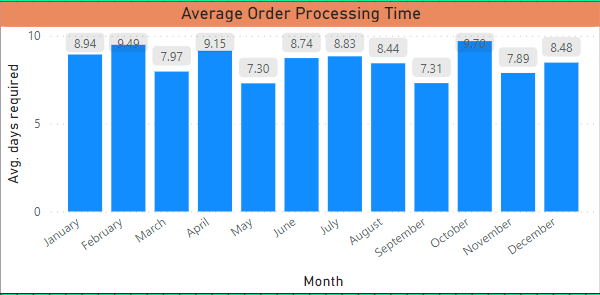
**What is the distribution of order values?**

From this visualization, we come to know that most of the order values lie between the price range of 2 to 45. The most expensive price range is from 219 to 263 which has only 1 product which reflects that the company focuses on products that are affordable for every customer. Also, many product lies in the 2 to 45 range of price customers can think of reordering the products.



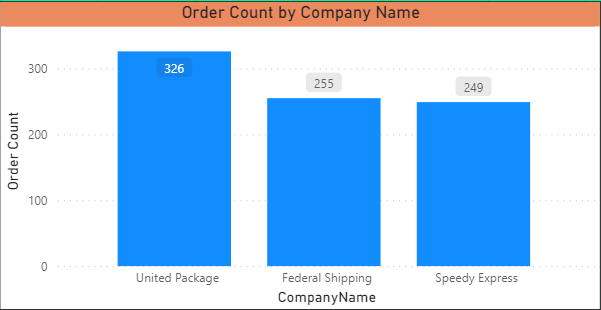
**Can we visualize the average order processing time or shipping duration?**

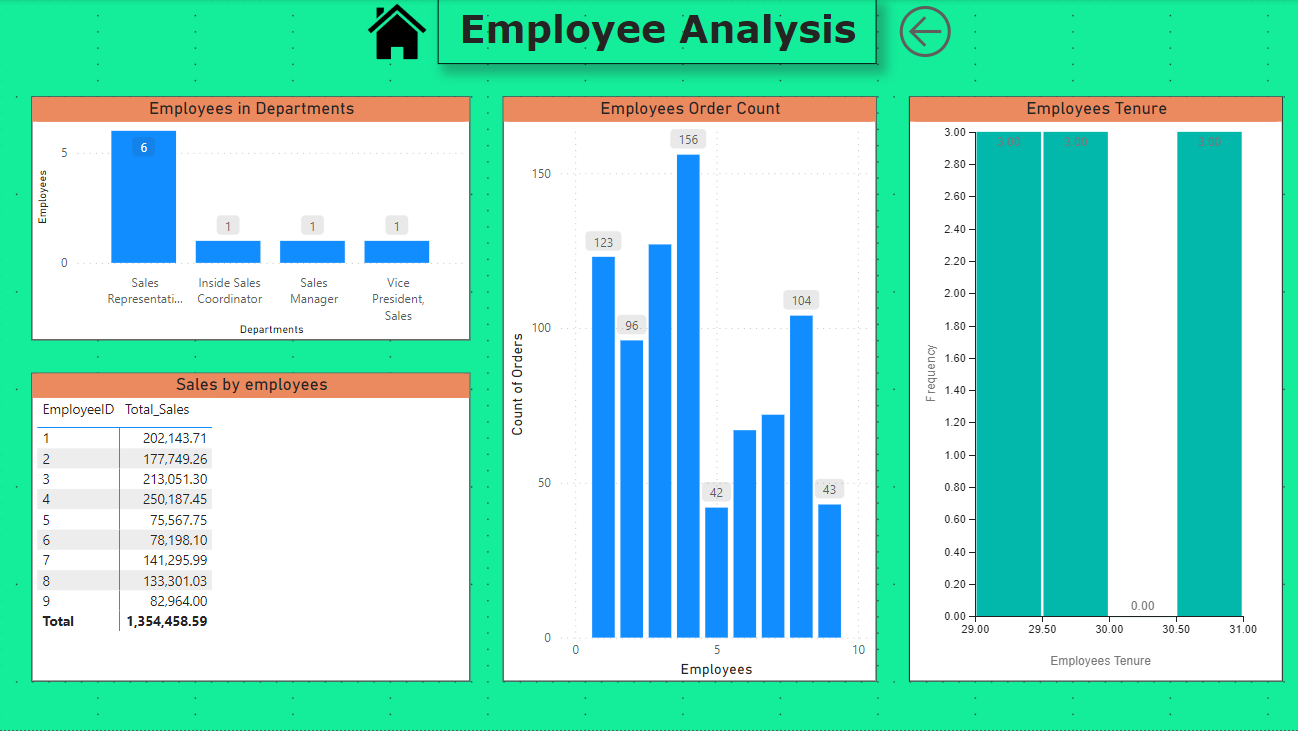
From the above visualization, we can see that the average order processing time for every month is nearly about same. It reflects that there is no delay in shipping the orders.

****

**Can we find the Order Count each shipping company have done?**

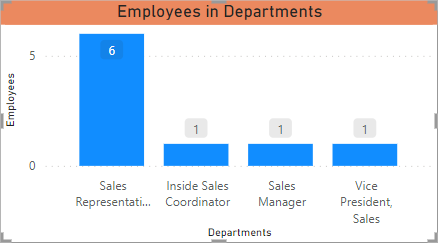
In this visualization, we find that the Northwind Traders company interact with many shipping companies which reflects that they have a good relationship with shipping companies. The United Package company have delivered the most number of orders(326) followed by Federal Shipping and Speedy Express.





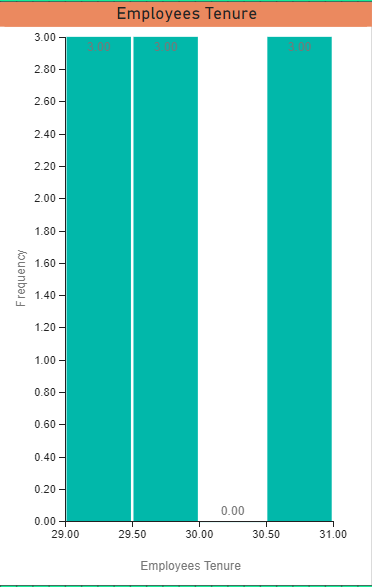
**How does employee productivity vary across different departments or job roles?**

From the above Visualization, we come to know that the organization focuses more on employees in the Sales Representative department than the others, as it has the most number of employees. It also reflects that the company can hire more employees in other departments which can help the other employees to distribute work which can lead to an increase in the work process.



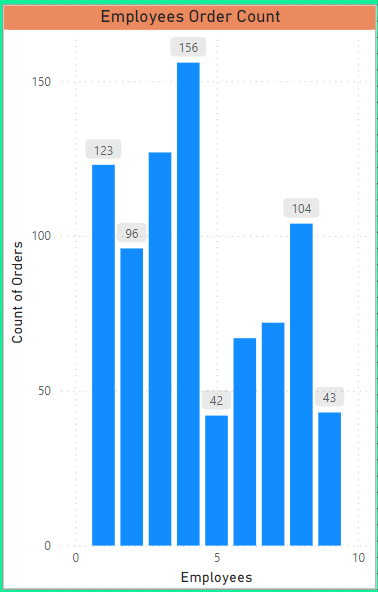
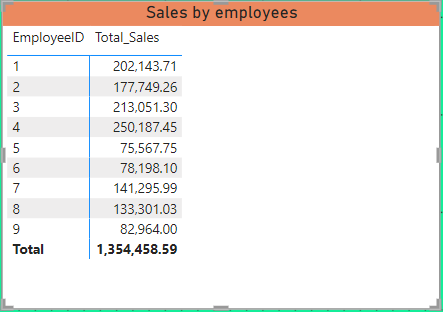
**What is the distribution of employee tenure?**

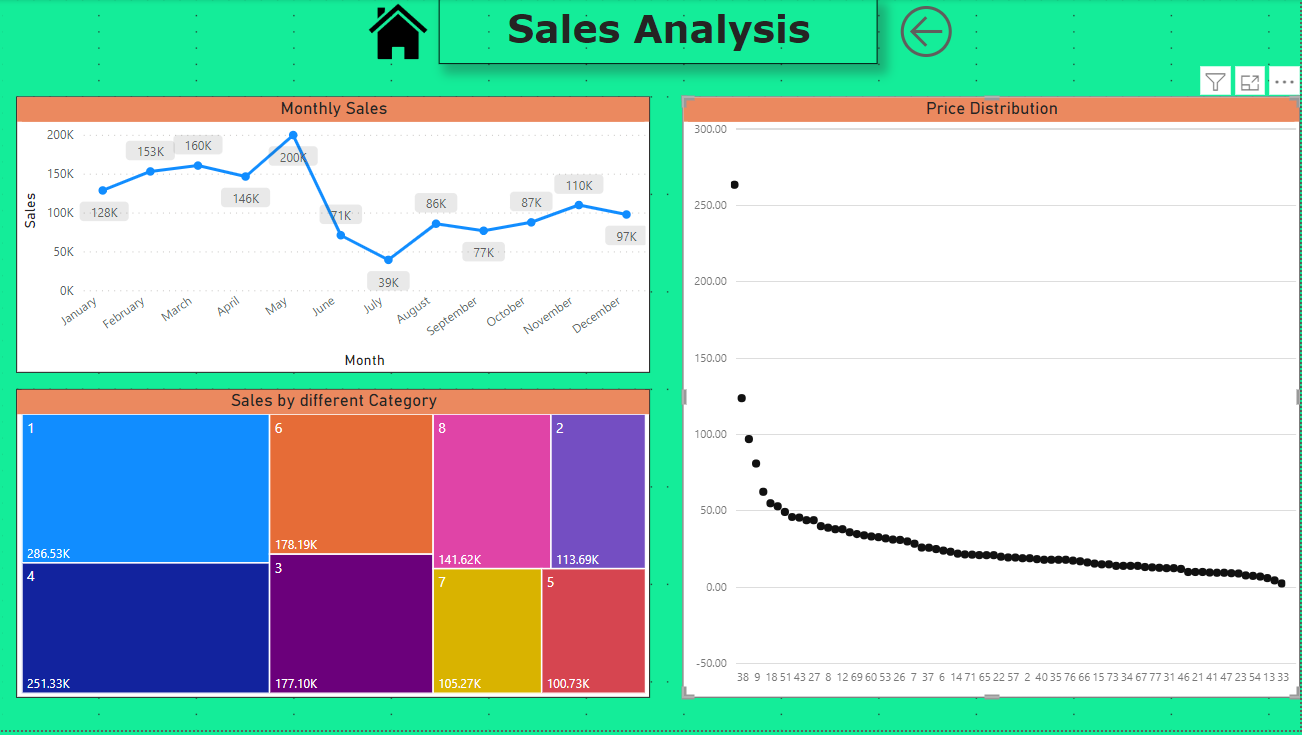
As we don’t have related data on which date the employees got separated from the organization, we assume that they are still working for the company that is (2023). So we can analyse that most of the employees have completed 30 years for the company which shows loyalty towards the company and have a considerable amount of experience in their job role.

****

**Can we visualize employee performance ratings or KPIs?**

As the employees have worked for a long period in the company, I can rate them on their performance such as contribution to sales, a contribution to orders that were made by the customers, we can say employees help in order count. In this visualization, we come to know that employees 1,3,4 and 8 have the maximum contribution in order count whereas Employees 5 and 9 have the minimum order count. Employees 1,2,3 and 4 have the maximum sales contribution whereas Employees 5,6 and 9 have the minimum sales contribution.



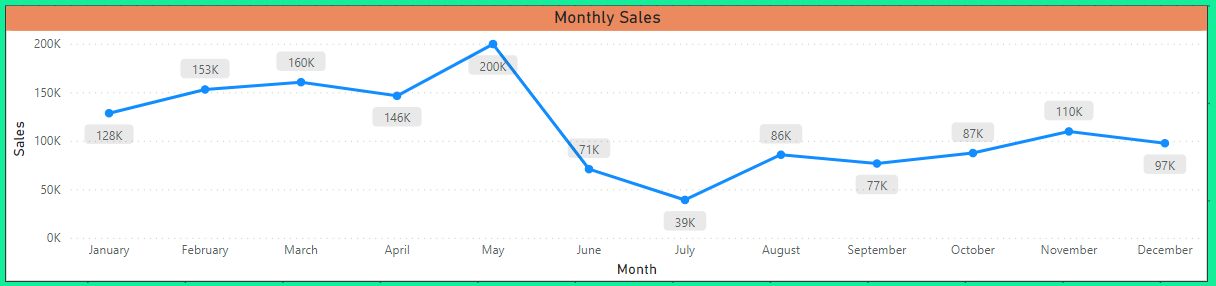


**What is the distribution of product ratings or reviews?**

By analysing the data we come to know that we don’t have data related to product rating hence we won’t be able to get any insights.

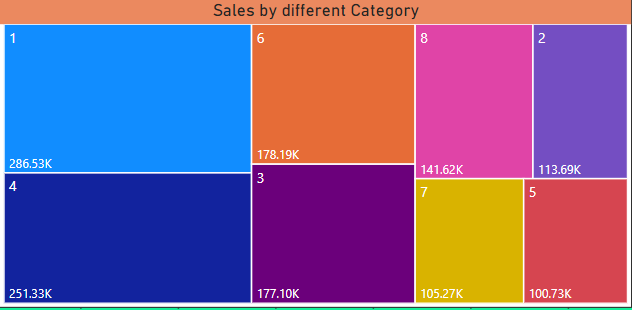
**How does Sales change over time?**

From this visualization, we find that in May, the sales were the most(200K) and the least sales in July(39K). We can also analyse that in the first 5 months, the sales are promising but after that it decreases. This pattern is similar to customer acquisition and order count which reflects that these three are correlated to each other.



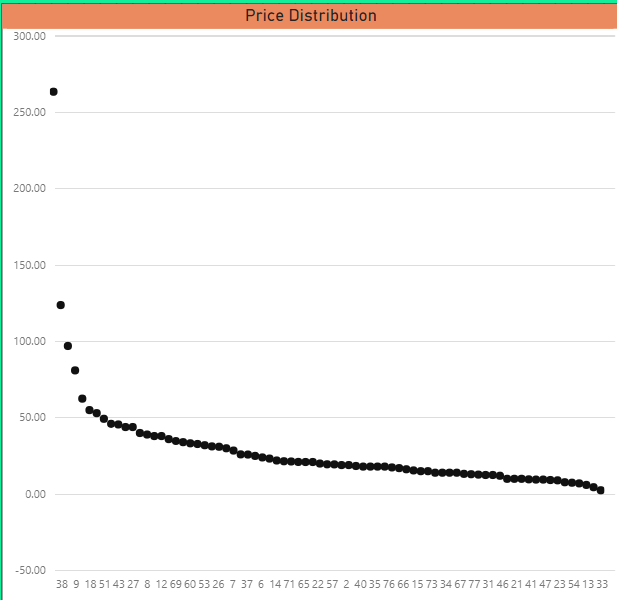
**What is the distribution sales in different product category?**

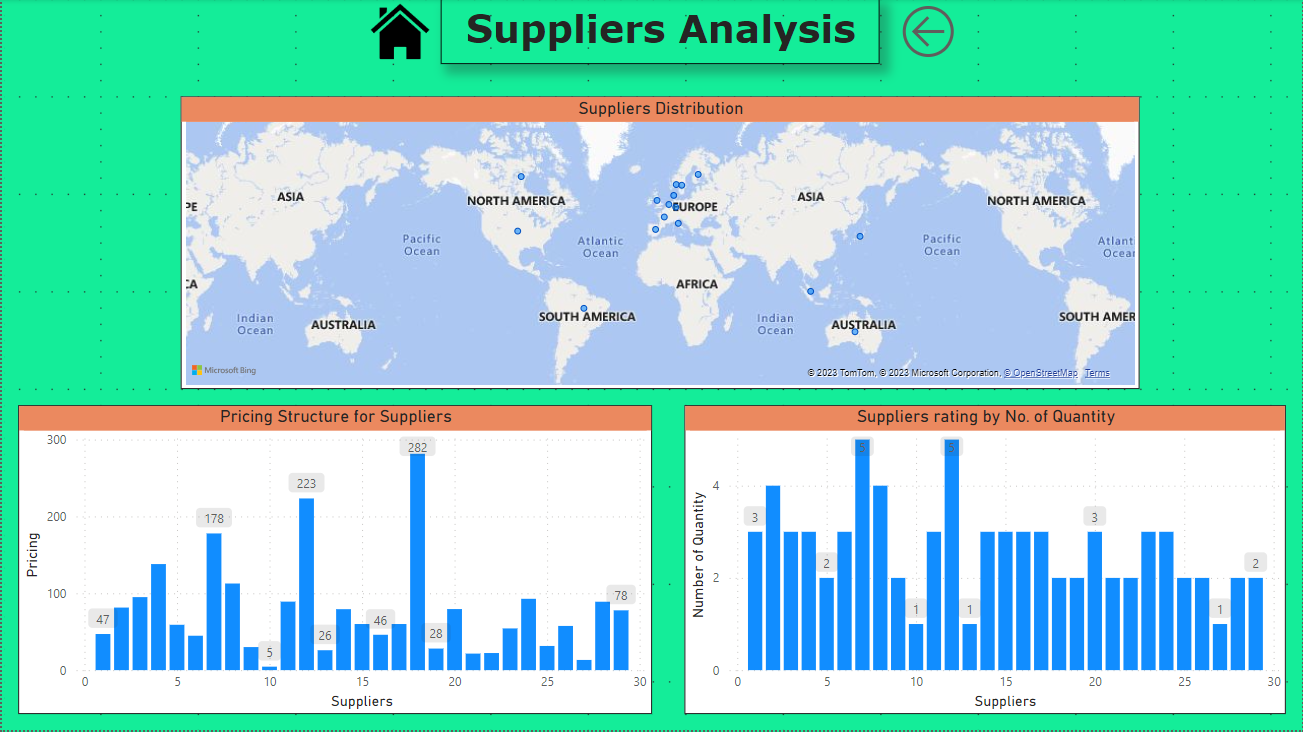
We can analyse that Category 1,2,6 and 3 are the leading categories in terms of sales which contribute half of the sales of products. It also depicts that customers are more interested in buying the products from these categories only.

****

**Can we visualize the pricing distribution of products?**

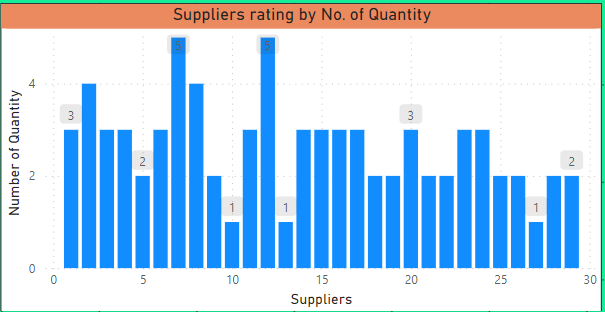
From this visualization, we come to that there are 2 to 4 outliers in the pricing of the products. With the help of this, we can help the company to manage the price range of the products from the suppliers.



****

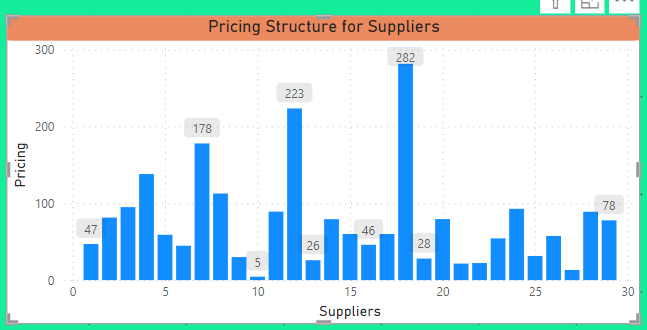
**What is the distribution of supplier ratings or performance metrics?**

We can analyse supplier ratings by the number of product quantities supplied by the supplier. From this visualization, we come to know that suppliers 7 and 12 have the most product quantities followed by suppliers 2 and 8 which have 4 product quantities.



**How does the cost or pricing structure vary across different suppliers?**

From this visualization, we can see that most of the prices of products are in the range between 0 to 100. There are 3 to 4 outliers. The most expensive product price is 282 which is not affordable for every customer so we can think of optimizing the prices of the outliers.

****

**Can we visualize the geographical distribution of suppliers?**

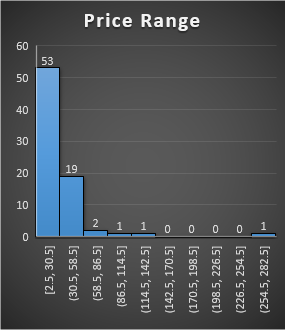
From this visualization, we can see that most suppliers are from Europe and the earlier visualization of customers’ geographical distribution we come to know that they are also from Europe which makes it easy for the company to distribute orders on or before delivery time.

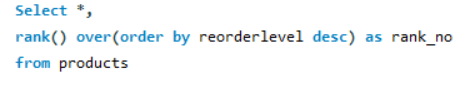
****

# **EDA Problem Statements**

**What are the key factors influencing customer retention or loyalty based on the dataset?**

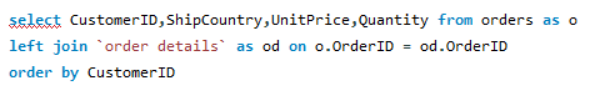
From this visualization, we can analyse the maximum product price range between 2.5 to 58.5 which is affordable for all the customer segments. With the help of reorder analysis, we can say that many customers reorder our products which indicates the company provides good services to the customers. We can also relate that as the price range is affordable for all the customer segments they reorder the products from the company.

****



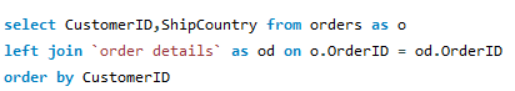
**How do customer preferences vary based on their location or demographics? Can we explore this through interactive visualizations?**

We can see that Germany and France customers order the most products from the company whereas Italy and Mexico order the least.



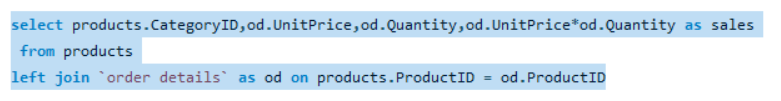
**Are there any interesting patterns or clusters in customer behaviour that can be visualized to identify potential market segments?**

We can see that most of our company’s customers are from Germany and France and Italy has the lowest number of customers.



**Are there any specific product categories or SKUs that contribute significantly to order revenue? Can we identify them through visualizations?**

From this visualization, we come to see that category 1 contributes maximum in sales of the Northwind Traders. Category 8 contributes a minimum in sales.



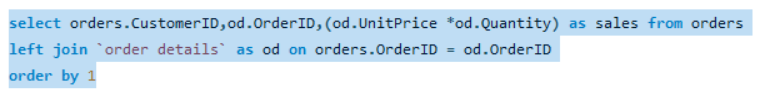
**Are there any correlations between order size and customer demographics or product categories?**

From the above visualization, we see that most of the order sales lie between 0 to 2000.

**How does order frequency vary across different customer segments?**

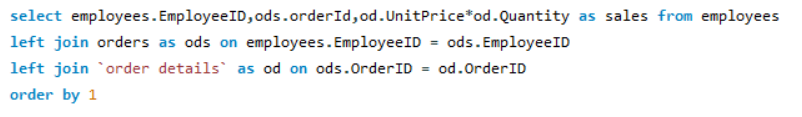
As we don’t have data related to the customer segment we can relate the customer segment to the country that they belong to. With the help of that, we can say that customers who belong to Germany order the most followed by France UK

Austria and Sweden. Portugal, Argentina and Italy have the least number of customers.



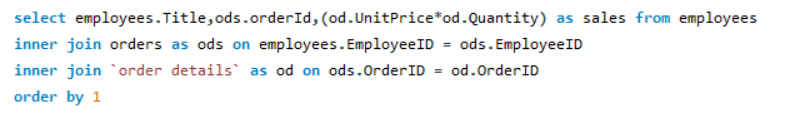
**Are there any correlations between employee satisfaction levels and key performance indicators?**

Employee satisfaction can treated as how they help customers to buy a product which leads to sales of the company. So we can relate this to how many customers each employee handles and we can correlate that sales done by each employee. So, with the help of this, we can say that employee 4 has helped in sales of the company more than others.



**How does employee turnover vary across different departments or job roles?**

We can see that the Sales Representative department has more turnover around 1000000. This is because the department has 6 employees compared to the other departments that have only 1 employee in each department. Northwind Traders should focus on hiring employees in other departments to optimize sales.



**Can we identify any patterns or clusters in employee skill sets or qualifications through visualizations?**

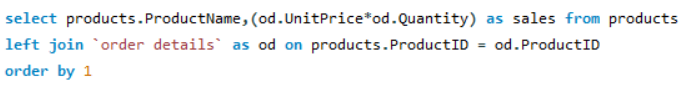
From the above Visualization, we come to know that the organization focuses more on employees in the Sales Representative department than the others, as it has the most number of employees. It also reflects that the company can hire more employees in other departments which can help the other employees to distribute work which can lead to an increase in the work process.



**Are there any correlations between product attributes (e.g., size, colour, features) and sales performance?**

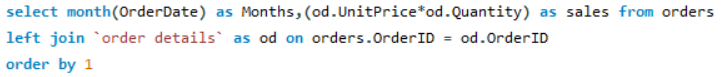
We can consider the names of the products as the product attribute and total sales as sales performance.

With the help of this correlation, we can analyse that many product sales are between 0 to 40000. The product sales that are above 40000 can be considered outliers. The most stand-out value from the outliers is between 140000 to 160000.



**How does product demand fluctuate over different seasons or months?**

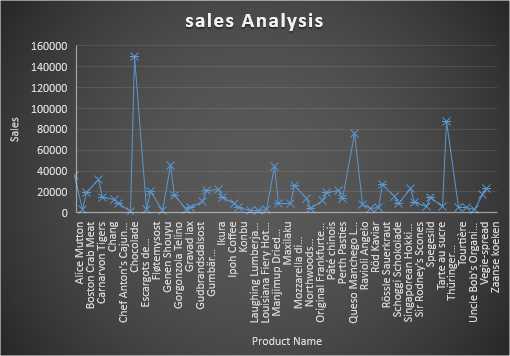
From this visualization, we find that in May, the sales were the most(200K) and the least sales in July(39K). We can also analyse that in the first 5 months, the sales are promising but after that it decreases. This pattern is similar to customer acquisition and order count which reflects that these three are correlated to each other.

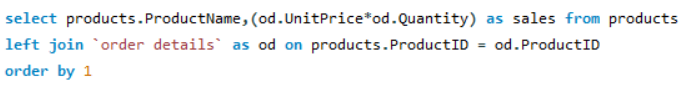


**Can we identify any outliers or anomalies in product performance or sales using visualizations? How can this information be used for product optimization?**

Outliers are the values that stand out greatly from the pattern formed.

From this visualization, we understand that there are 5 to 8 product values which can be considered as outliers in the sales analysis. This can be optimised when the high unit price of the products decreases. This can happen only when the supplier sells products in an optimised way to the Northwind Traders.

****



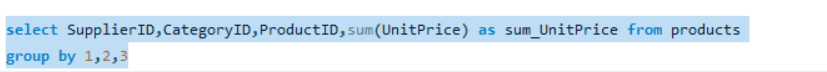
**Are there any correlations between supplier attributes (e.g., location, size, industry) and performance metrics (e.g., on-time delivery, product quality)?**

In the dataset, we have one entity that is related to the attribute of supplier which is the company in which the supplier works. For Performance metrics, we can take the number of products each supplier sells to the company.

We can see that 2 companies have 5 products and 2 companies have 4 products that they sell to the Northwind Traders. These are the main companies for Northwind Traders which delivers the most number of products.

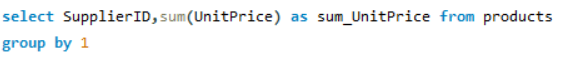
**How does supplier performance vary across different product categories or departments?**

As we don’t have data related to the performance of the suppliers we can relate performance to the number of categories of products they sell to the company. We can see that suppliers 7 and 12 have the most categories which is 5 which indicates that they have products that lie in 5 different categories. Then we have suppliers 2 and 7 with 4 categories. These are the main assets of the company as compared to others.



**Can we identify any trends or patterns in supplier costs or pricing structures through visualizations? How can this information be used for procurement optimization?**

From this visualization, we can identify that most of the supplier’s pricing structure is between 0 to 100 and some are above 100. so the Northwind traders can approach the suppliers and discuss how can the suppliers optimise the high-range products such that many customers afford the products which will result in customers buying the high product on a repeated basis which eventually increases the sales of the company.



# **Conclusion**

* Here, we have come to the end of the capstone project with the topic name as Sales Analysis. I would like to share my experience while doing this project. I have learned the working of many tools like MySQL, Excel and Power BI.
* How to find the insights of the problem statements and make charts which helps us to visualize the problem statements more effectively.
* Making effective dashboards in Power BI helps the viewer to have a better understanding of the Problem statements.
* With the help of MySQL, we learned how to write a query on a workbench and to get the desired records from the tables.
* With the help of Excel, we learned how to make effective charts with the help of pivot tables and Excel functions.