Report

Comprehensive Analysis of Sales Performance and Customer Behavior by Country and Overall

Data Visualization Report for STAT 112

Cemile Bilgir

METU

December 2024

Instructor: Prof. Dr. Ceylan Yozgatlıgil

1. Introduction

In this report, visualizations and evaluations regarding General Sales Performance and Comprehensive Analysis of Customer Behaviors by Countries were conducted. With these visualizations and evaluations, the aim was to achieve better sales performance by assessing the distribution characteristics of sales and the adequacy of existing strategies.

1.1 Data Description: The following data was used while creating the visualizations.

Data	Description
Country (Categorical, Nominal)	Name of the country.
Latitude (Numerical, Continuous)	Latitude coordinates the country's location.
Longitude (Numerical, Continuous)	Longitude coordinate of the country's location.
Birth Rate (Numerical, Continuous)	Number of births per 1,000 population per year.
CO2-Emission (Numerical, Continuous)	Carbon dioxide emissions in tons.
CPI (Numerical, Continuous)	Consumer Price Index, a measure of inflation and purchasing power.
Gasoline Price (Numerical, Continuous)	Price of gasoline per liter in local currency.
GDP (Numerical, Continuous)	Gross Domestic Product, the total value of goods and services produced in the country.
Life expectancy (Numerical, Continuous)	The average number of years a newborn is expected to live.
Population (Numerical, Discrete)	Total population of the country.
Tax Revenue (%) (Numerical, Continuous)	Tax revenue as a percentage of GDP.
Total Tax Rate (Numerical, Continuous)	Overall tax burden as a percentage of commercial profits.

Data	Description
Unemployment Rate (Numerical, Continuous)	Percentage of the labor force that is unemployed.
Urban Population (Numerical, Continuous)	Percentage of the population living in urban areas.
Order Number (Categorical, Nominal)	This column represents the unique identification number assigned to each order.
Quantity Ordered (Numerical, Discrete)	It indicates the number of items ordered in each order.
Price Each (Numerical, Continuous)	This column specifies the price of each item in the order.
Order Line Number (Categorical, Ordinal)	It represents the line number of each item within an order.
Sales (Numerical, Continuous)	This column denotes the total sales amount for each order, which is calculated by multiplying the quantity ordered by the price of each item.
Order Date (Categorical, Nominal)	It denotes the date on which the order was placed.
Days Since Last Order (Numerical, Discrete)	This column represents the number of days that have passed since the last order for each customer. It can be used to analyze customer purchasing patterns.
Status (Categorical, Nominal)	It indicates the status of the order, such as "Shipped," "In Process," "Cancelled," "Disputed," "On Hold," or "Resolved."
Product Line (Categorical, Nominal)	This column specifies the product line categories to which each item belongs.
MSRP (Numerical, Continuous)	It stands for Manufacturer's Suggested Retail Price and represents the suggested selling price for each item.
Product Code (Categorical, Nominal)	This column represents the unique code assigned to each product.
Customer Name (Categorical, Nominal)	It denotes the name of the customer who placed the order.
Phone (Categorical, Nominal)	This column contains the contact phone number for the customer.
Address Line 1 (Categorical, Nominal)	It represents the first line of the customer's address.
City (Categorical, Nominal)	This column specifies the city where the customer is located.

Data	Description
Postal Code (Categorical, Nominal)	It denotes the postal code or ZIP code associated with the customer's address.
Country (Categorical, Nominal)	This column indicates the country where the customer is located.
Contact Last Name (Categorical, Nominal)	It represents the last name of the contact person associated with the customer.
Contact First Name (Categorical, Nominal)	This column denotes the first name of the contact person associated with the customer.
Deal Size (Categorical, Ordinal)	It indicates the size of the deal or order, which are the categories "Small," "Medium," or "Large."

2. Data Preprocessing

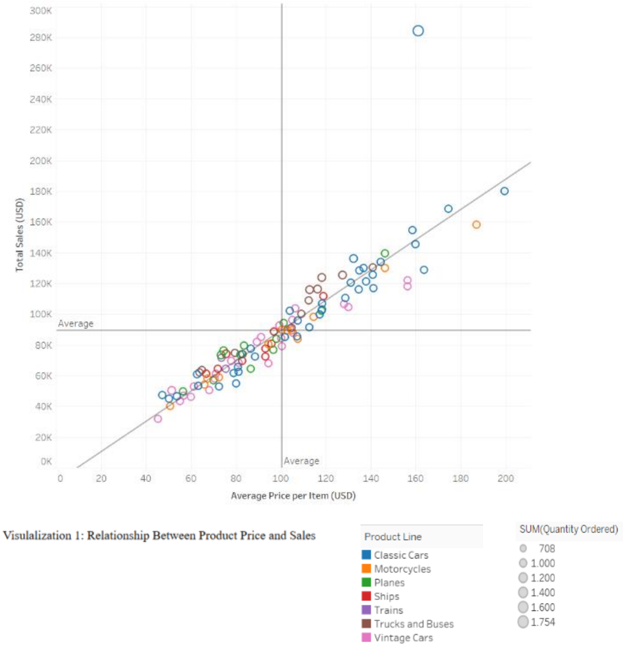
For data preprocessing, data cleaning was required as the first step to ensure more reliable outcomes. In this data analysis, we had two datasets, and *country* was the common variable between them. The datasets were initially merged using an inner join in Tableau, eliminating any resulting null values. Subsequently, misclassified data (e.g., *gasoline price* and *GDP* being classified as strings) was corrected by converting them to their proper class, *numerical*.

3. Exploratory Data Analysis

The data analysis section was conducted using 5 research questions and visualizations.

3.1 What is the relationship between product unit price and total sales?

Scatter plot is a suitable graph to understand the relationship between product prices and total sales amounts. Because in this direction, correlation analysis can be made about the impact of product pricing on sales, and improvements regarding pricing can be made because of the correlation analysis.



Visulalization 1: Relationship Between Product Price and Sales

Interpreting the Scatter Plot

The product line is shown in different colors in the chart, thus allowing the sales performance per category to be examined.

Balloon sizes represent the number of products ordered, and which category has the most sales is visualized with increasing balloon sizes. For example, in this visualization, classic cars and vintage cars are shown with larger balloons than trains and planes, which indicates that they have more sales.

The trend line in the chart shows that there is a positive correlation between product prices and total sales. This shows that as the Price increases, total sales increase. The reason for this may be the result of customers thinking that higher-priced products are better.

The average line on the chart shows the average of total sales. The average line divides the scatter plot into four parts. These are: top-right (high sales / high price), bottom-right (low sales / high price), upper-left (high sales / low price), and bottom-left (low sales / low price). Thanks to this, it is determined whether the pricing is correct or incorrect. It can be interpreted that the current strategy can be continued for products in the upper right (high sales / high price) region (e.g. Classic Cars). Because although its price is higher than other categories, its sales are also quite good. However, for products that are in the Alt-Right (Low Sell / High Price) zone (e.g. Planes) it is necessary to change the sales strategy. The low number of sales indicates that the price of this product is higher than it should be. When the products in the Upper-Left (High Sales / Low Price) region (e.g. Motorcycles) are evaluated, it is observed that their sales are quite high, but their prices are low. Here, a strategy can be developed to increase the income potential by increasing the prices of the products. Finally, when the Alt-Left (Low Sales / Low Price) region is interpreted, products that are below the average price and sales are observed (e.g. Trains). Changes such as stopping production or improving products may be made for this region. Additionally, there is a clutter in the middle parts of the chart. Regarding this accumulation, it can be inferred that products in the middle segment are much more in demand.

Finally, there is a classic car outlier coded S18_3232 in the chart. Although this model has a very high price, it has a high total sales volume. This shows that a study needs to be done to find out why this product is contradictory.

3.2 Is there a relationship between total sales and the percentage of tax revenue across different countries?

One of the most effective graphs for comparing categorical data, such as countries, is a bar graph. Using the length and colors of the bars, the relationship between countries and two different variables can be easily analyzed simultaneously. Additionally, the balance of bar length and color gradient provides an easy comparison opportunity.

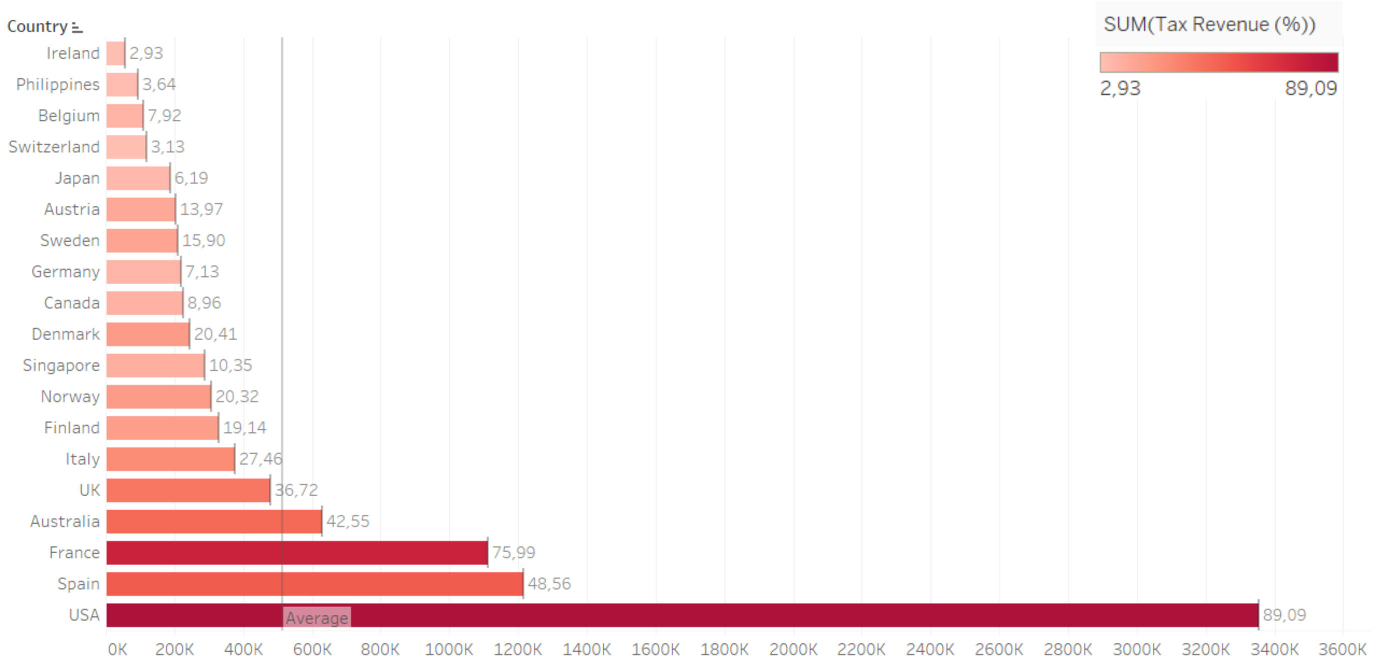
Interpreting Bar Graphs

In the graph, the color tones of the bars progress from light to dark, representing the percentage of tax revenue from the lowest to the highest. For instance, since the percentage of tax revenue in the USA and France is much higher than in Ireland and Switzerland, the bars representing these countries are darker in color. When the bar lengths are considered to relate to total sales, it is observed that the USA has both the darkest and the longest bar, while Ireland has the lightest and the shortest bar.

However, it would be incorrect to conclude that there is a direct proportionality between the percentage of tax revenue and total sales. This is because for direct or inverse proportionality to exist, it must be consistent. Instead of looking at only the first and last data points, all data should be considered.

When analyzing the graph, some countries disrupt the consistency. For example, Spain's total sales are higher than France's, but when looking at the percentage of tax revenue, Spain's (48.56%) is lower than France's (75.99%). For direct proportionality, the opposite would need to be true. Generally, when observing the colors, although the bars increase in length, the colors do not progressively darken but instead vary.

The lack of a relationship between total sales and the percentage of tax revenue can be attributed to additional variables such as the country's population, economic size, and tax rates



Visualization 2: Countries by Total Sales and Tax Revenue (%)

3.3 How has the distribution of average sales and outliers changed over the years?

To identify outliers in the dataset, a box plot is one of the most effective graphs. Additionally, the distribution boundaries (lower and upper limits) of yearly sales data for each year can be easily determined using a box plot visualization.

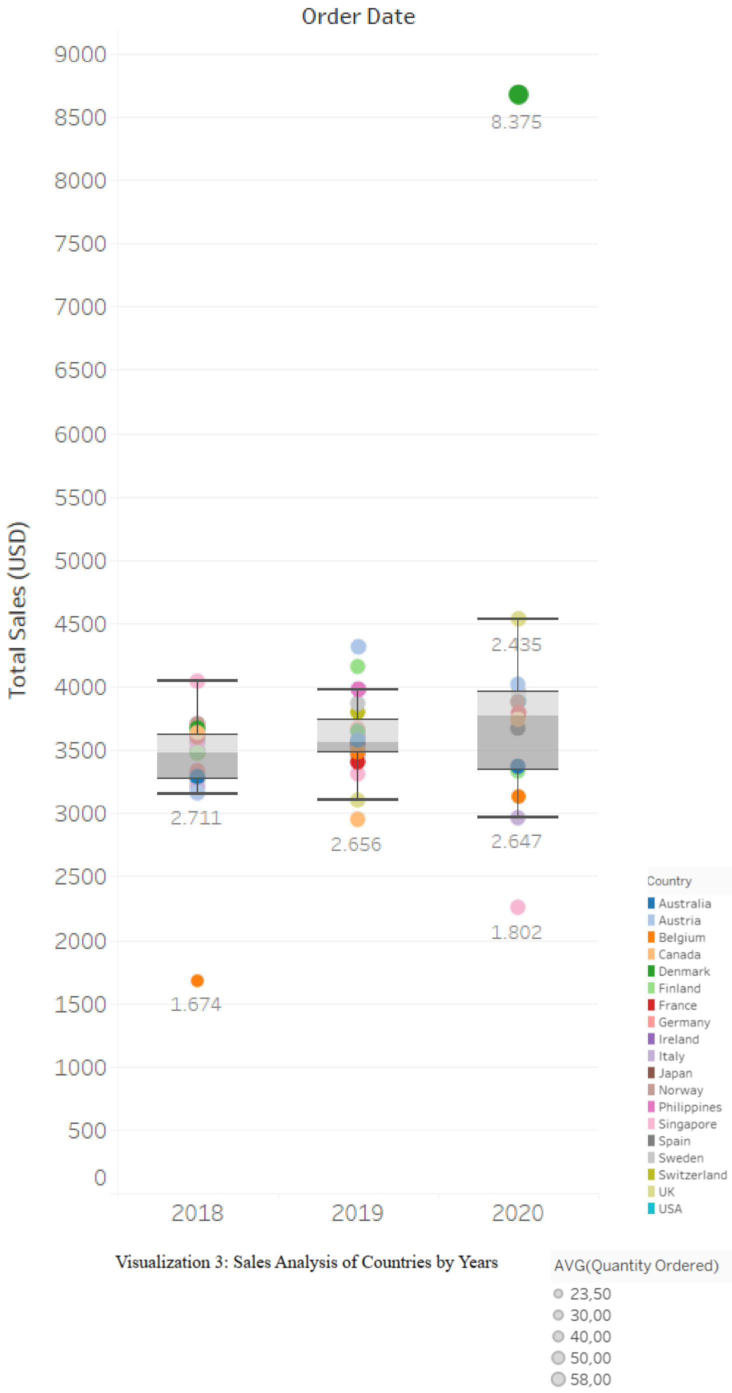
Interpreting the Box Plot

In 2018, the average sales value was approximately 2,711K USD, decreasing slightly to 2,656K USD in 2019 and 2,647K USD in 2020. Although the decline is small, it indicates a slight downward trend.

Regarding outliers, in 2018, a low sales value close to the lower limit, 1,674K USD (Belgium), was observed. However, this situation was corrected in 2019, and the sales value aligned with the average. The outliers in 2019 were very close to the lower and upper limits, whereas in 2020, two outliers were identified: a low sales value of 1,802K USD (Singapore) and a very high sales value of 8,375K USD (Denmark). The high outlier in 2020 suggests a significant increase in demand, which could be analyzed and potentially integrated into the sales strategies of other countries.

When analyzing the distribution, a wide range is observed in 2018. Although some products showed lower sales, the fact that most sales were concentrated around the median indicates consistent performance. In 2019, the distribution became more compressed, suggesting a more balanced sales performance compared to the previous year. The distribution in 2020 also narrowed, reducing sales inconsistency.

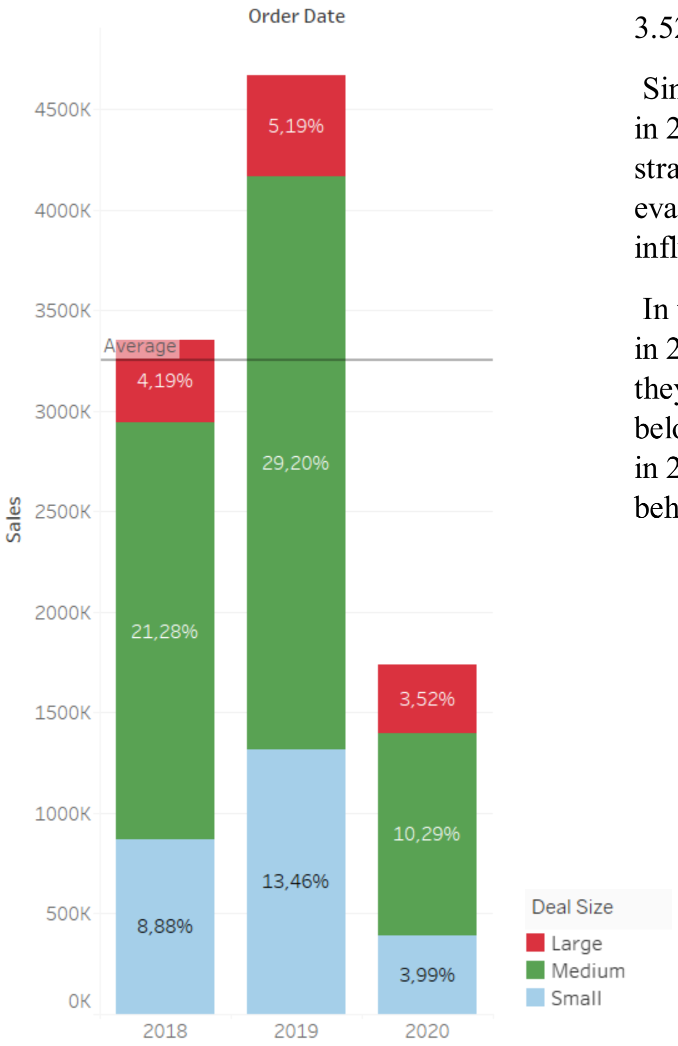
However, despite this consistency, the presence of a very high outlier in 2020 remains an exceptional situation that should be considered.



Visualization 3: Sales Analysis of Countries by Years

3.4 How has the sales performance of different Deal Size categories changed over the years?

A stacked bar chart is an excellent visualization tool for this analysis as it shows both the total performance and the distribution among categories. This visualization allows for easy examination of both total sales by year and changes in deal size proportions. Therefore, a stacked bar chart is the best choice for this question.



Visualization 4: Yearly Sales Performance by Deal Size Categories

Interpreting the Stacked Bar Chart

When analyzing deal sizes, it is immediately apparent that the most preferred deal size is *medium*, while the least preferred is *large*.

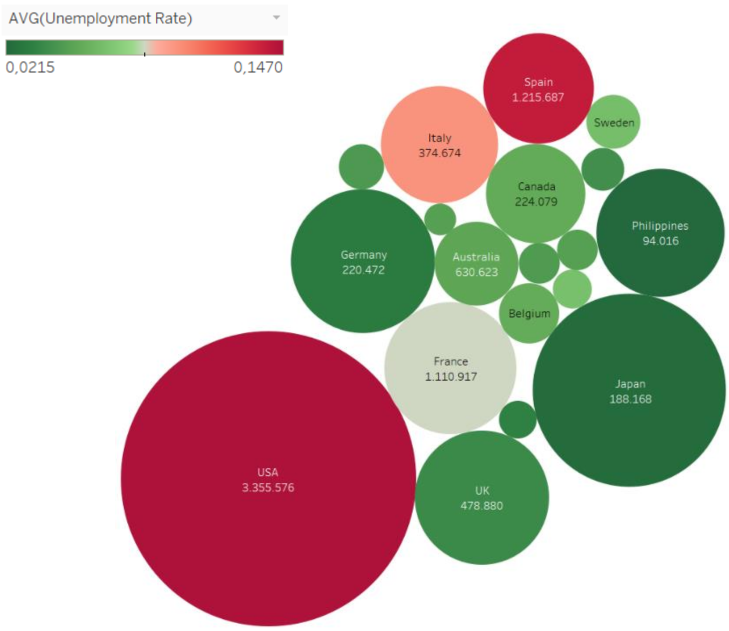
On a category basis, it is observed that the proportion of small deal size dropped significantly from 13.46% in 2019 to 3.99% in 2020. This indicates the need for a different sales strategy for small deal sizes. The *medium* deal size, which has the largest sales proportion in all years, also experienced a decline from 29.20% in 2019 to 10.29% in 2020. Similarly, the *large* category showed a decrease during the same period, dropping from 5.19% to 3.52%.

Since all categories experienced a decline in 2020, instead of focusing solely on sales strategies, it may be more insightful to evaluate whether any extraordinary factors influenced these results.

In the overall yearly sales evaluation, sales in 2018 were slightly above the average, they peaked in 2019 but fell significantly below the average in 2020. The pandemic in 2020 could be the extraordinary factor behind this unexpected decline.

3.5 What is the impact of Urban Population and Unemployment Rate on Total Sales by Country?

For this question, it is necessary to visualize more than two variables, and a bubble plot is one of the most suitable graphs for such multi-variable visualizations. The increasing bubble size represents total sales, color changes represent unemployment rates, and the positioning categorizes countries based on urban population.



Visualization 5: Urban Population, Unemployment Rate and Total Sales Relationship by Country

Interpreting the Bubble Plot

When examining the graph, it is evident that the USA plays a dominant role, as indicated by its significantly larger bubble size.

For medium-sized bubbles (e.g., Spain and France), it is observed that they occupy a middle position in terms of sales. However, compared to the USA, a country with high sales, the urban population in these countries is much lower. Similarly, in Belgium and the Philippines, where the urban population is quite low, the sales are also significantly lower.

Regarding the unemployment rate, the USA shows a low unemployment rate combined with high sales, suggesting a possible inverse relationship. However, when evaluating other countries with low unemployment rates (e.g., Spain or Germany), it is observed that their sales are not as high as expected. This indicates that the relationship between these two variables is not consistently applicable across all cases.

5. Conclusion and Discussion

In the five questions analyzed and visualized, various evaluations regarding sales were made. These include how product prices, countries and percentage tax rates, deal sizes, rural areas, and unemployment rates affect sales, as well as the annual sales distribution of countries.

A positive correlation was found between product prices and sales, while no definitive impact of percentage tax rates on sales was identified on a country basis. Similarly, urban areas and unemployment rates did not exhibit a significant relationship with sales.

When analyzing sales over the years, it was determined that the unexpected decrease in 2020 could be attributed to the economic impact of the coronavirus pandemic. Additionally, the yearly sales distribution of countries was analyzed, and countries with outliers were identified.

https://public.tableau.com/app/profile/cemile.bilgir/viz/FNALdasboard/Dashboard1?publish=yes