STAT112 - INTRODUCTION TO DATA PROCESSING AND VISUALISATION TERM PROJECT

BY

BUSE NUR AÇIKGÖZ

1. INTRODUCTION

The aim of this project is to analyze the sales of an automobile company and understand some factors affecting sales, using two datasets. The first dataset covers information about an automobile company’s sales. The second dataset includes demographic and economic statistics of countries. Both datasets are combined to provide an effective observation.

1. DATA PREPROCESSING

Before analyzing the data, data cleaning and preparing are a crucial step to have a reliable outcome. Since two datasets were used by inner joining with them by country, there was a valuable chance for data to be duplicated or mislabeled or incorrect. First, Tableau’s “data interpreter” feature which is useful for detecting whether there exists any missing data was used, there was not any null data. Also, there was not found any duplicate. Some of the observations were irrelevant for this project, so these irrelevant data were hided. Some fields such as “Gasoline Price”, “GDP” and “Ordernumber” were misclassified. For “Gasoline Price” column, since it contained a string “$”, it should be removed. First a calculated field was created and by using “replace()” code, “$” sign was removed and it’s data type was converted as number. For “GDP” column, same steps were made. For “Ordernumber”, since it represents the unique identification number, and any operation can be done, it was better to convert it to a string type. Lastly, to make the dataset much understandable, some fields were renamed such as “Ordernumber” to “Order Number” and “Days\_Since\_Last\_Order” to “Days Since Last Order”.

1. EXPLORATORY DATA ANALYSIS (EDA)

After data preprocessing process, EDA is a critical stage in data analysis process, allowing to summarize the main characteristics of a dataset using visualizations. In this step, 5 questions were asked and drawn some conclusions. Regarding the two datasets, one of them consist of 20 variables and 2747 observations and it is about an automobile company’s sales data. Second dataset consists of 14 variables and 195 observations which are related to all countries’ demographic statistics, economic indicators etc. Before moving on to the research questions, some crucial variables are given below.

|  |  |
| --- | --- |
| **Categorical** | **Numerical** |
| **Order Date** (Ordinal) | **Quantity Ordered** (Discrete) |
| **Status** (Nominal) | **Sales** (Discrete) |
| **Product Line** (Nominal) | **Co2 Emissions** (Continuous) |
| **Customer Name** (Nominal) | **Total Tax Rate** (Continuous) |
| **Country** (Geographical) | **CPI** (Continuous) |
| **Deal Size** (Ordinal) | **GDP** (Continuous) |
|  | **Urban Population** (Discrete) |

In this project, some variables like GDP or Urban Population were used since these variables are crucial to reach some conclusions about sales performance which is naturally affected by external factors.

3.1. Research Questions

3.1.1. How is the distribution of sales and order quantities across different time periods?

A graph with blue squares and lines

Description automatically generated

Visualization 1: Sales by Year

A graph of a graph

Description automatically generated with medium confidence

Visualization 2: Number of Orders Across Order Date

It is essential to know when their products sold better, since each time’s trends were different. For instance, observing growth whether or not it is a positive growth in sales during a certain period may help the company about determining market dynamics or economic conditions, by considering some strategies. If there is a significant rise in sales during a particular time, it may indicate the success of some marketing strategies or product launches. On contrary, declines may be an alert for company. A search about selling prices, customer satisfaction or competition might be questioned. Additionally, it may help detect seasonal patterns. Some factors in certain years such as economic crises or pandemics might also impact sales. For instance, if we have sales data after Covid-19 officially announced as pandemic, we would be able to draw some conclusions how it affected the sales. However, from the data we had, it can be searched why the sales were low between December 2018 and This kind of analysis let the company to consider past performance, make more accurate predictions for the future sales, and develop supportable growth strategies.

From Visualization 1, the year that median of the sales covers greater portion and broader distribution is 2019. From Visualization 2, there are jumps in December 2019 and August 2019, which affects previous sentence. Also, line chart shows that there exist two absolute peaks of ordered quantities, December 2018 and December 2019. This may be explained by seasonal cycling or utilize from tax deductions. There are small increases in August 2019 and May 2020 also. Furthermore, both line chart and box plot indicate company’s growth but not that high.

3.1.2. How do different order statuses contribute to overall sales performance?

A graph with different colored bars

Description automatically generated

Visualization 3: Total Amount of Sales by Status

The bar chart highlights a great part of the sales, which is shown as blue color, caused by successfully shipped orders. It shows that company is handled with sales very well. Small part of sales lies under the “In Progress, On Hold, Resolved” category which is represented with yellow color and from this, we can conclude that company should focus on lowering the delays and quickening the orders. The “Cancelled and Disputed” category with red color represent the lowest piece of sales, but it emphasizes dissatisfaction of customers. Understanding the reason lies behind this category is important, and, in this part, customer services should communicate those customers directly. By focusing on reducing the rate of cancellations and disputes as much as possible, company will be improving overall effectiveness and maximize sales performance.

3.1.3. If company decides to launch a personalized campaign targeting five customers, which customers should be selected?

A graph of blue squares

Description automatically generated with medium confidence

Visualization 4: Five high value customers and Their Sales

Customers in the graph are top five customers which have higher purchasing amount. From the graph, it can be understood clearly that the sales amount is almost between 150K and 900K. Euro Shopping Channel’s sales more than the other worldwide customers also its sales much more than the average sales of all countries. While top five customers mostly prefer “Medium” deal size among their sales, the less preferred deal size is “Large”. Company should prioritize these five customers for a personalized campaign due to the high sales contributions and potential growth. By understanding their preferences and business needs, the company can design campaigns to enhance loyalty.

3.1.4. If the company wants to upsize in some countries, which countries can they be? A graph with many dots

Description automatically generated

Visualization 5: Relationship Between Sales, Percentage of Urban Population and CO2 emissions by Country

A group of different colored squares

Description automatically generated

Visualization 6: GDP per country

In Visualization 5, x-axis is calculated by the ratio of urban population and population, since this will give more accurate information rather than population. The sizes of the bubbles correspond the amount of that country’s sales. From the scatter plot, USA has the highest and extraordinary CO2 emissions, reason of it might be the fact that almost eight out of ten people live in urban area. However, to be able to make this conclusion, it is not solely enough. The reason is that when considering other countries with higher urban population rate such as Singapore, Norway, Denmark, they have not high CO2 emissions like USA. Other conditions might suppose to be affect CO2 emissions, the number of automobiles per person or the number of industrial facilities. Also, Japan and Germany have a little bit higher CO2 rate compared to other countries. Regarding the sales, in general USA, Spain and France contribute sales more.

In Visualization 6, each area represents GDP of that country. Again, USA has the greatest value. That means in USA, there is a strong economic activity. This indicates customer’s demand and competitiveness will be high, too. The least area belongs to Finland. That means businesses cannot invest that country. For the company, low GDP countries would not be preferable to expanding, since in those countries customers spend less and because of that product demands will be less.

As for the question 4, in countries with high GDP such as USA, Japan and Germany, company may consider upsizing. Also upsizing by manufacturing new products might be a valuable strategy. For instance, in USA, Japan and Germany, i.e. high CO2 emissions countries, manufacturing sustainable and eco-friendly products boosts company’s profits while contributing those countries with lowering CO2 emissions.

3.1.5. How does the total ordered quantity vary across different product lines?

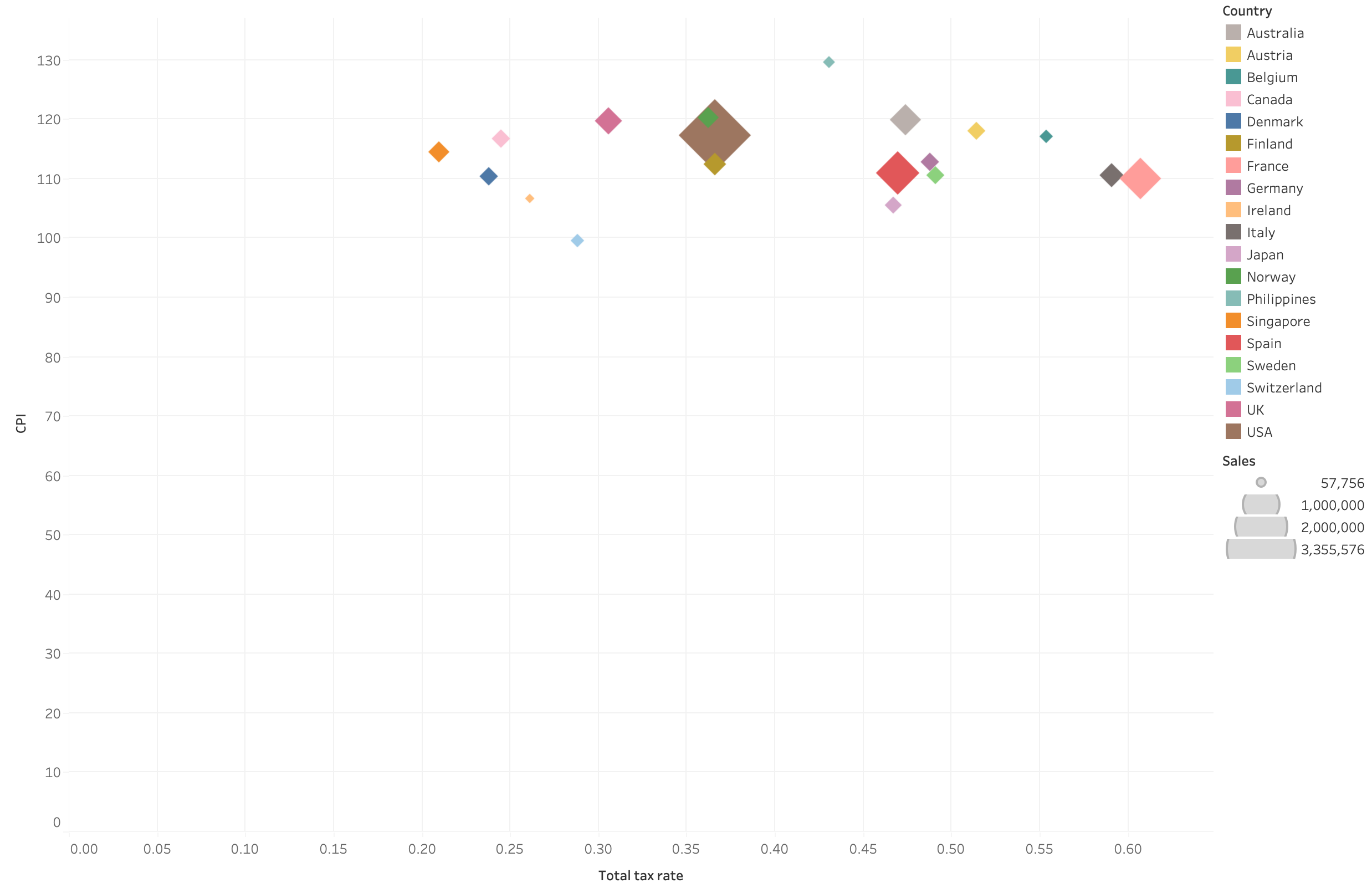
A graph of a graph showing different types of objects

Description automatically generated with medium confidence

Visualization 7: Ordered Quantity Across Productlines

From lollipop chart, the most preferred product is classic cars with an ordered quantity of almost 35K. The less preferred one is trains with an ordered quantity of nearly 4K. The reason train sales is less than the others may be customers’ prefer another company for purchasing. Focusing on that area and improving weaknesses might be clearly affect company positively in a long term.

3.1.6. How are the sales affected by CPI and Total Tax Rate? Is there a relationship between them?



Visualization 8: Relationship Among Total Tax Rate, CPI and Sales by Country

CPI and Total Tax Rate are two of the external factors which affect sales indirectly. The higher GPI means the higher inflation and expenses. If prices change drastically, customers cannot afford to buy products. Additionally, when total tax rate increases, sales might decrease. However, while CPI and Total Tax Rate are low, sales are not as much as high. That means sales are not solely affected by them. When a country’s GPI is lower for example Switzerland or Austria, its sales are also lower.

Most of the countries spread between the range of 20% and 60% tax rate and clustered around 115 CPI. USA seems to be an outlier, since sales are excessively high despite its total tax rate and CPI position. Singapore has lowest and France has the highest tax rate. For the CPI, Switzerland has the lowest and Philippines has the highest value, and both countries has little number of sales.

1. CONCLUSSION AND DISCUSSION

This project analyzed an automobile company’s sales based on two datasets. Analyses are made with six research questions by visualization in Tableau. Analyses give some insights which can be helpful for the company to take some measurements to potential issues, understanding sales relationships between economic indicators or demographic indicators or highlight some important patterns etc. Some conclusions are given below.

Sales are increasing during specific periods-December months.

Overall sales mainly consist of successful sales, and this indicates that company is managed well.

Two customers among five high value customers have noticeable number of sales.

USA has extreme values such as high CO2 emissions, GDP and the number of sales, the reason might be its economy and great number of population.

There is no strong correlation among CO2 emissions and Urban Population, which is not an expected property.

Classical cars are the most sold product over three years and trains are the less.

One expected aspect is higher CPI and Tax Rate bring greatest sales, however countries with low CPI and low Tax Rate have not the greatest sales.

Links

<https://github.com/buseacikgozz>