METU Department of Statistics

STAT 112 - Introduction to Data Processing and Visualization

Project 1 — Sales & Country Data

Aliha Shahid
2703486

Abstract

This report is based on datasets covering an automobile company's sales and various global indicators; data analysis methods are applied to ascertain how these datasets relate to each other. Five research questions are examined ranging through various indicators and sales data from both datasets after joining and cleaning the data. The research questions and findings are presented below.

Introduction

The scope of this report covers a description of the datasets, detailing the process of data cleaning and Exploratory Data Analysis (EDA), and finally using this data to answer questions pertaining to the influence of global factors on automobile sales where applicable.

The first dataset used in this project is the "Automobile Sales" dataset, consisting of data including but not limited to quantity, price, sales, order date, status, product line (or category), and country of order. These variables are put next to the second dataset, "Global Country Information", including indicators by country, such as GDP, Carbon dioxide emissions, gasoline price, population and so on. This data was imported into Tableau and preprocessed.

Data Cleaning/Preprocessing

The two datasets were imported into Tableau, after checking for the existence of duplicates in Excel and finding none. In the Data Source Pane of Tableau, both datasets were connected using an Inner Join operation that left only the data relevant to the analysis. Column names were rewritten for enhanced readability, and some variables were manually converted to categorical dimensions instead of measures. The data was thus ready for processing.

Exploratory Data Analysis

After the data preprocessing, the main variables of interest for this project were as follows:

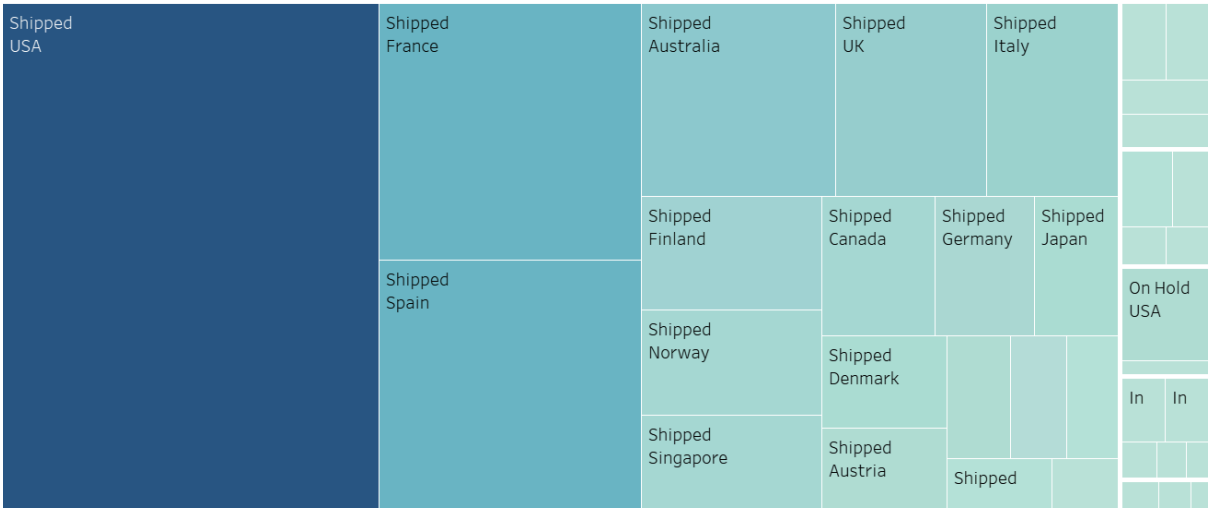
Variable Name	Variable Description	Data Type & Scale
Country	Name of the country	Qualitative, Nominal
Status	Status of the order i.e. Shipped, In Process, Cancelled, Disputed, On Hold, Resolved	Qualitative, Nominal
CO2 Emissions	Carbon dioxide emissions in tons	Quantitative, Ratio
Gasoline Price	Price of gasoline per liter in local currency	Quantitative, Ratio
Quantity Ordered	Number of items ordered in each order	Quantitative, Ratio

Sales	Total sales amount for each order (Quantity Ordered * Price Per Item)	Quantitative, Ratio
Price Per Item	Price of each item in the order	Quantitative, Ratio
Product Line	Product line categories to which each item belongs	Qualitative, Nominal

Table 1: Variable names, descriptions, with data types and scales.

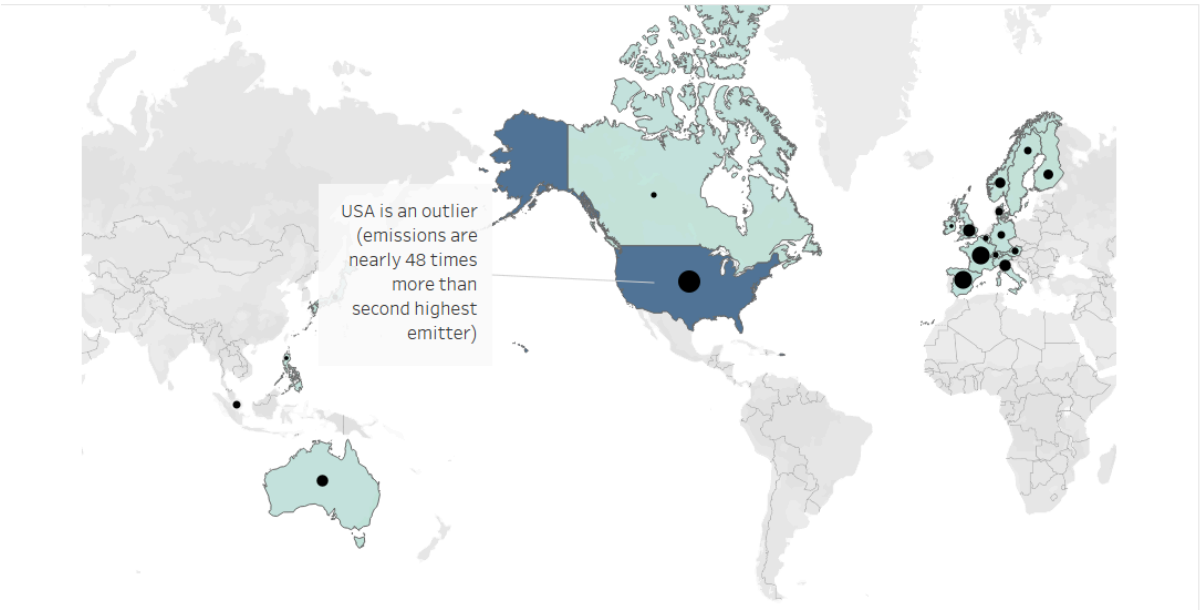
Hence, it is seen that in this particular analysis, the variables are either qualitative and nominal-scaled, or quantitative and ratio-scaled. How they affect or are linked with each other becomes clearer through in-depth exploration with the five research questions (RQs) in mind.

RQ1: What is the distribution of order status by country?



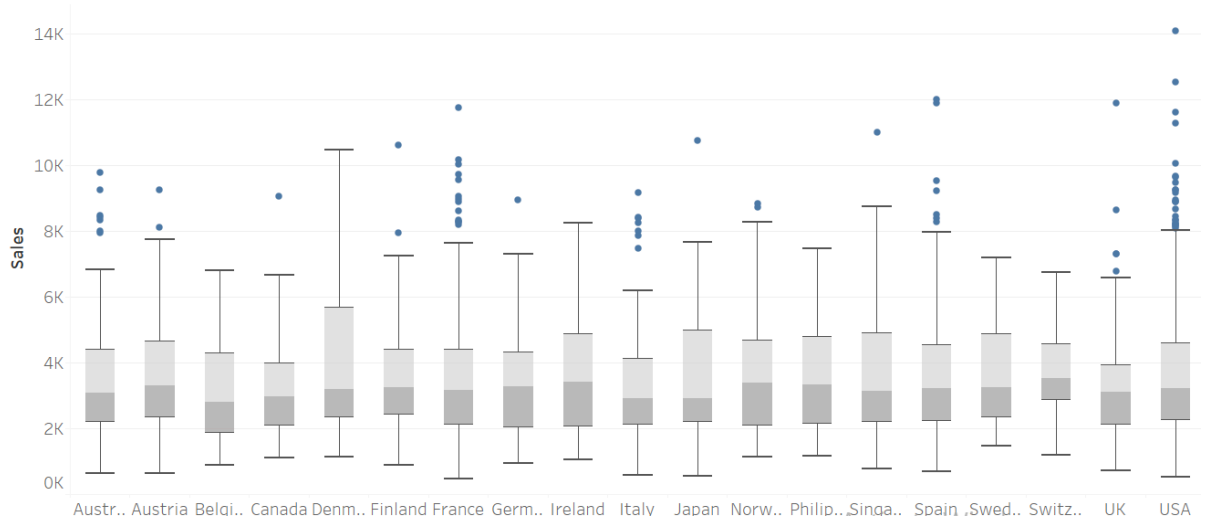
A mosaic plot was chosen to answer this question as it is most apt in showing the part-to-whole relationship of order status to all orders. It also makes it easier to understand the segregation of each order status by country. As can be simply inferred from the graph, above 80% of orders are shipped and delivered to their respective countries, with the other statuses sharing an almost equal percentage of total orders. This shows that the manufacturer is quite reliable in making sure the orders are received, no matter where in the world the customer may be, as there is no biased data which may indicate any particular countries where problems arise in the order delivery status. It is further seen that most orders are shipped to the USA, followed almost equally by Spain and France, which may suggest that these countries are highly prized consumers of this manufacturer, and more automobiles may be located in these countries.

RQ2: What is the relationship between CO2 emissions and gas prices in each country?



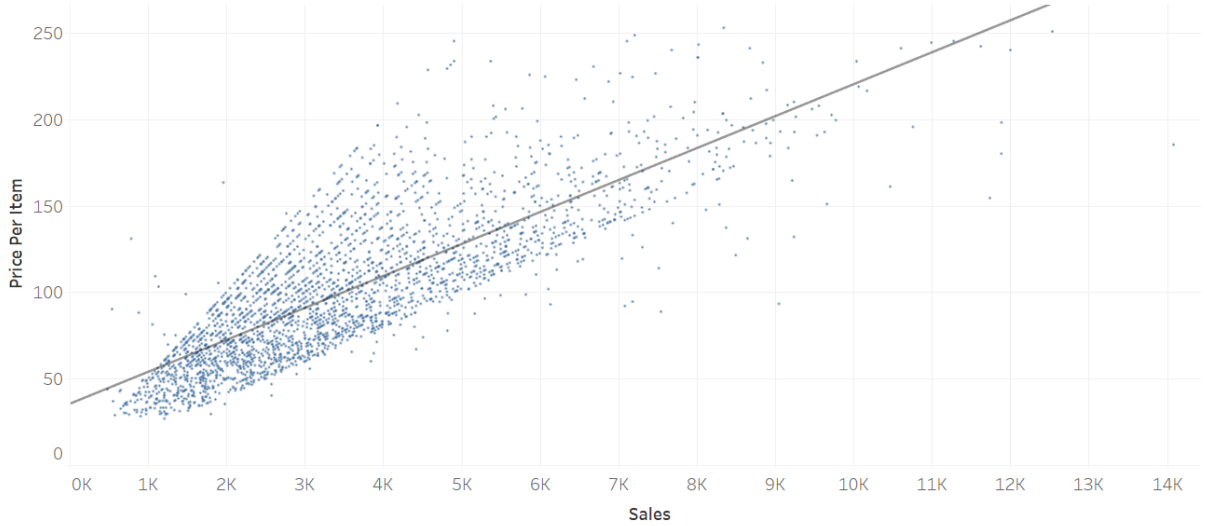
A map is used to show the relationship by country. The countries which are buyers of this manufacturer are shaded in accordance to their CO2 emissions, with a darker blue shade indicating a higher rate of emissions. The size of the black dots over each country indicates the prices of gasoline in each country. The USA is an immediate outlier as it emits 4,645,848,256 tons of carbon dioxide, which is nearly 48 times more than the second highest emitter (France). This can be seen by the intense difference in shade of the USA and other countries. However, this may be due to the larger size of the USA compared to other countries. Generally in the graph, higher gas prices tend to correlate with higher CO2 emissions. This is seen as the black dots of largest sizes correspond to the USA, Spain, and France. In the previous mosaic plot, it was observed that these three countries are the largest consumers of the automobile manufacturer. The observation makes sense as when the amounts of automobiles increase in a country, more gasoline will be demanded to run them, leading to a shortage of supply hence gasoline prices will increase in these countries.

RQ3: What is the distribution of sales per country?
(cont. on next page)



A box-and-whisker plot is useful for visualising the spread of items and showing outliers. In this graph, the distribution of sales per country is shown, and it is seen that the average sale amount is nearly the same for all countries, i.e. there is very little variation in the median sale amounts. However, a lot of outliers are observed, especially for the USA as there seems to have been a lot of sales of amounts high above the threshold, which may imply that the USA places an order of a higher number of items so as to contribute so heavily to the manufacturer's sale amount. Other observations from the graph are that Denmark has the highest range of sales, and all sales are restricted to developed countries.

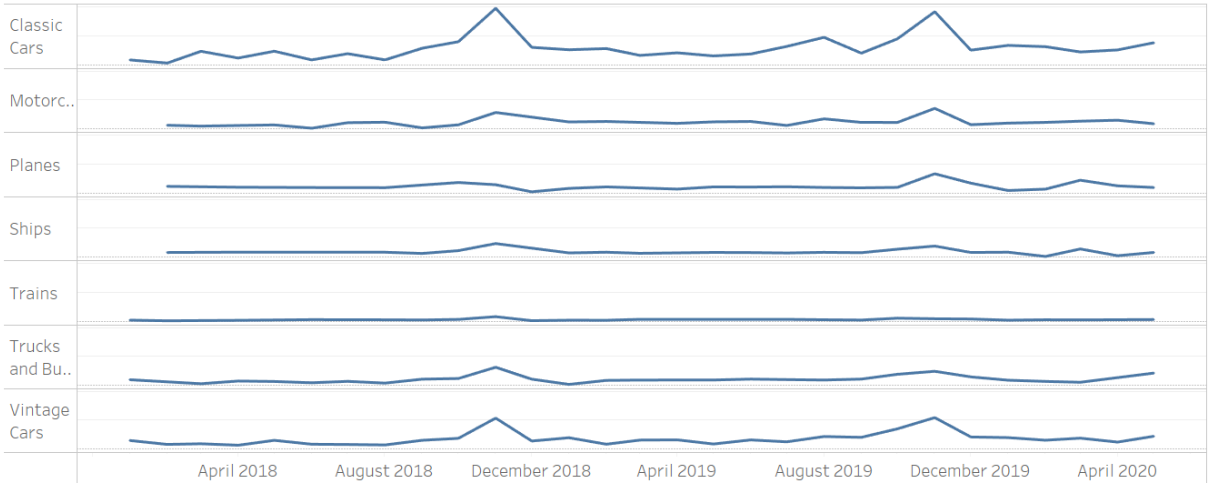
RQ4: How do item prices and sales correlate with each other?



A scatter plot is ideally used for plotting the relationship between two variables, hence why it is used for this type of question. Price per item and Sales are plotted on the axes and the data aggregation is removed to show each point. A trend line is also fitted in and it can be seen that the data has a strong positive correlation. The correlation is stronger nearer to the origin and the data tends to spread out as the values on both axes increase. The graph strongly suggests that item prices and sales are positively related, which may seem counterintuitive at

first. However, demand for automobiles from this supplier may be more inelastic so consumers tend to buy the product no matter if the price gets higher and higher; automobiles may be a necessity in some consuming countries where alternatives such as a public transport system may not exist or are not properly functional. The wider spread of the graph as the price gets higher may imply that even though a positive correlation exists, values for sales tend to scatter as the price increases so the product is not fully inelastic.

RQ5: Is there a trend of sales according to product lines?



Trends are usually shown through line graphs. This graph is divided into rows based on the product line (i.e. category of product). We can make a few observations from the graph: most sales belong to classic cars, while the least sales belong to trains. The highest and second highest data points of almost all the product lines are found in November 2018 and November 2019. As the data is only between 2018 to 2020, it is not possible to say with a lot of confidence why this peak occurs at the same time for almost all product lines. If more years' worth of data could also be collected and analysed, it may be possible to infer whether the trend truly exists and why. However, from this data we can conclude that cars have the most sales among all product lines and maximum sales happen in November across almost all categories.

Conclusion

Through exploring the data, multiple observations and trends were unearthed. The manufacturer sells all their products to developed nations, out of which the USA is a stark outlier in many respects, e.g. CO2 emissions, gasoline prices etc. A positive correlation was found between both sales and item price, and gasoline prices per CO2 emissions. The underlying socioeconomic factors were briefly brought to light as hypotheses on why these trends have emerged. It is observed that sales distribution per country is relatively constant, however sales distribution per product type over years carries a surprising result; namely, the peak of sales every year in or around November. Through the data analysis above, it is possible to identify trends, as observed above; it is also possible to predict future patterns or optimise sales and marketing strategies to yield a higher profit.

Project materials:
Github Link
Tableau Dashboard