— Report —

Global Economy and Automobile Sales Analysis

by Eylül Zeynep Dalbudak
2024 December
METU

1. Abstract: This study examines worldwide economic statistics and car sales to identify significant trends. Madrid is the city with the most orders, according to the results, and Euro Shopping Channel is the top customer. San Marino has the longest life expectancy while China has the most CO2 emissions, according to global data. These results demonstrate the relationship between sales patterns and economic conditions. These patterns are explained and helpful insights are provided for better decision-making through the use of charts such as pie charts and histograms.

2. Introduction: The goal of this study is to comprehend car sales trends and how they connect to international economic variables. To address important concerns regarding sales patterns, consumer behavior, and variables like GDP and urbanization, two datasets were used, and the data was cleaned and visualized. The findings are displayed in a narrative and dashboard to offer lucid insights for improved decision-making.

3. Data Tidying and Cleaning Steps: These data have 2747 observations. Also, the data have 34 variables. There are 20 numerical and 14 categorical data.

First flaw that is found in the data is Birth Rate, CPI, Life Expectancy columns was classified as string,
When the data is inspected it was seen that the data delimiter was comma instead of dots
This problem was resolved by creating a calculated field.
Using the function

```
FLOAT(REPLACE([Birth Rate], ",", "."))
```

The data before and after modification is shown below.

Abc world-data-2023.csv Birth Rate	=# Calculation Birth Rate Fixed
11,6	11.60
11,3	11.30
11,3	11.30
11,6	11.60
11,6	11.60
11,3	11.30
10,4	10.40
11,3	11.30
12,6	12.60

Besides these, Population and Urban population Column was also string but the number was shown in a format as “XX.XXX.XXX” which would be interpreted as wrong since the float delimiter is “.” for tableau

So the dots were removed using

```
REPLACE([Population], "." , "")
```

Then it was casted to float.

The data before and after modification is shown below.

Abc world-data-2023.csv Population	=# Calculation Population Fixed
328.239.523	328,239,523.00
67.059.887	67,059,887.00
67.059.887	67,059,887.00
328.239.523	328,239,523.00
328.239.523	328,239,523.00
67.059.887	67,059,887.00

Also it is observed that Co2 emissions column contains null values that can cause some issues while creating visualizations. Hence, it is decided to replace null values to zeros using the following formula;

```
ZN(SUM([Co2-Emissions]))
```

# world-data-2023.csv Co2-Emissions	=# Calculation Co2 Emissio
null	0.000
303.276	303.276
303.276	303.276
null	0.000
null	0.000
303.276	303.276
41.023	41.023
303.276	303.276
375.908	375.908
null	0.000
null	0.000
null	0.000

The last flaw that was found is that Rate columns (Tax Revenue, Total Tax Rate, Unemployment Rate) were also strings that contained % on the front, so in addition to replacing commas it was needed to remove '%' . This was achieved using the following formula.

```
FLOAT(REPLACE(REPLACE([Tax revenue (%)], ",", "."), "%", ""))
```

Note that Float casting was first made when we used data type modification. So it is added to the newly added formula to bypass making one more operation.
The resulting conversion can be observed below.

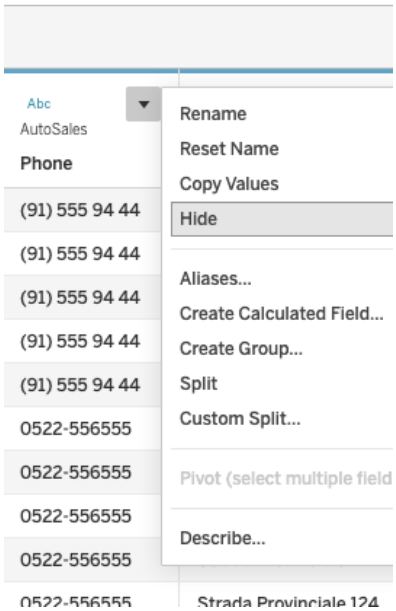
Abc world-data-2023.csv Tax revenue (%)	=# Calculation Tax revenue (%) Fixed
%9,60	9.6000
%24,20	24.2000
%24,20	24.2000
%9,60	9.6000
%9,60	9.6000

When we checked the data after fixing the data types, it is observed that there are NO duplicate data points. Although there are same order id numbers these rows are not identical which indicates there exists multiple purchases under the same order id so I decided not to eliminate these values.

Also using sort functionality of the tableau it is observed that there are no out of range values for different quantitative measures.

The final state of the data can be used for visualizations and the pre-processing can be said to be completed.

There is some data that is irrelevant and shouldnt be used for the visualization such as phone number row. These data are hidden for a simpler view of data.

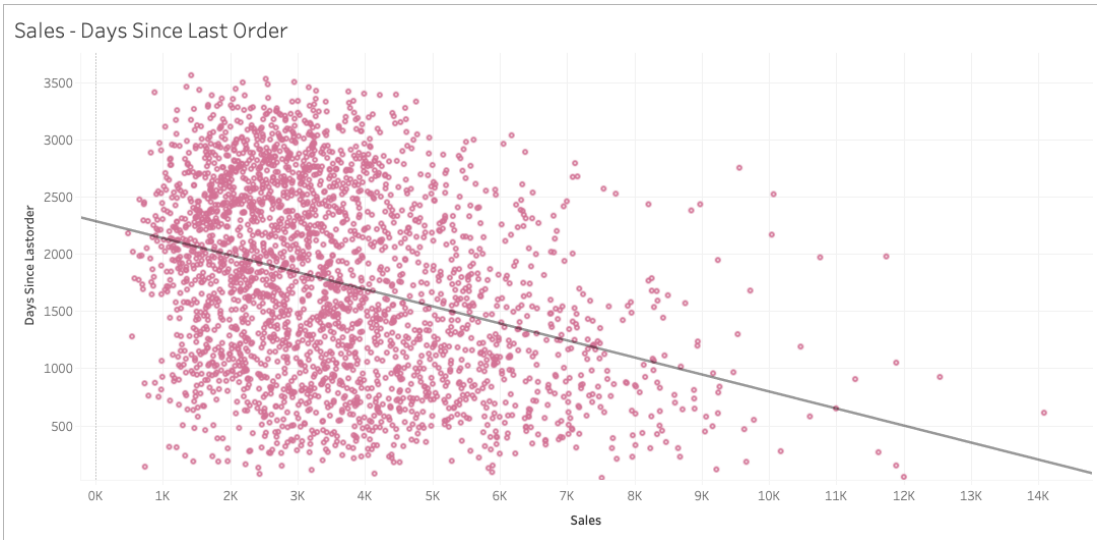


4. Exploratory Data Analysis: We conducted five research questions for this project.

4.1.What is the relationship between the days since last order and sales amount?

As the data describes, the sales amount is the actual spending of that order since the data is multiplication of order quantity and the price of each item.

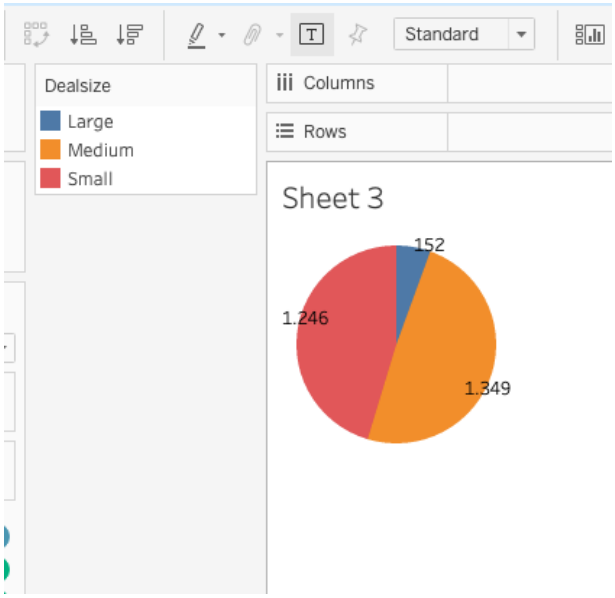
For this research question it is observed that the customers who are ordering in a more frequent manner tend to spend more on each order. A scatter plot is a good way to show the observation and the trend line supports the hypothesis.



4.2. What is the ratio between deal sizes?

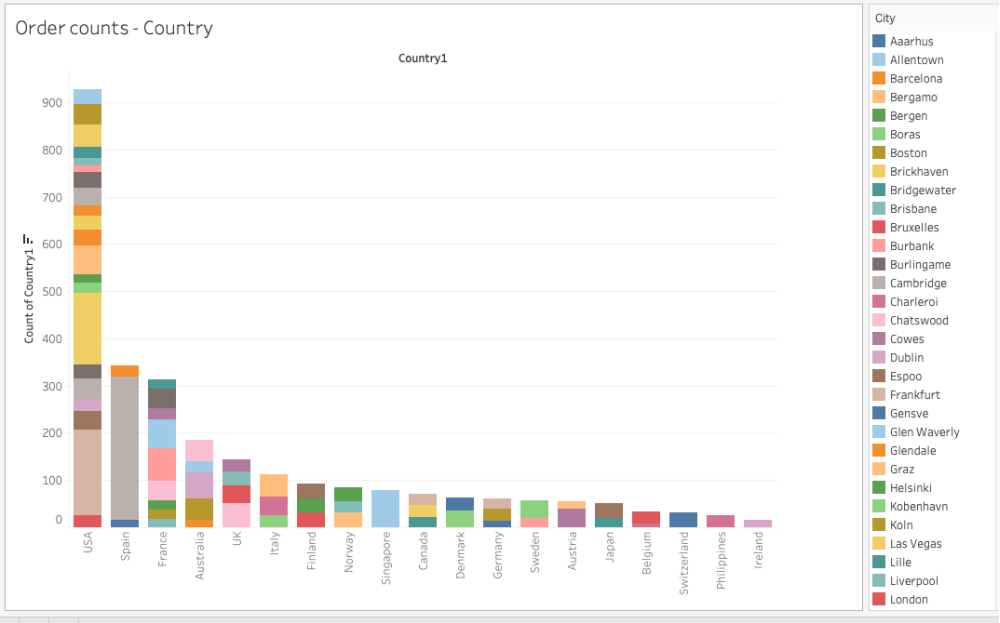
Although this would be a univariate plot, I think it is important to check the ratios of small, medium and large deal sizes. Since our class count is 3 for this case, it is appropriate to use

bar chart, As a summary it was expected for the large deals should be smaller in amount and the bar chart emphasizes that.



4.3. What is the relation between order count vs countries?

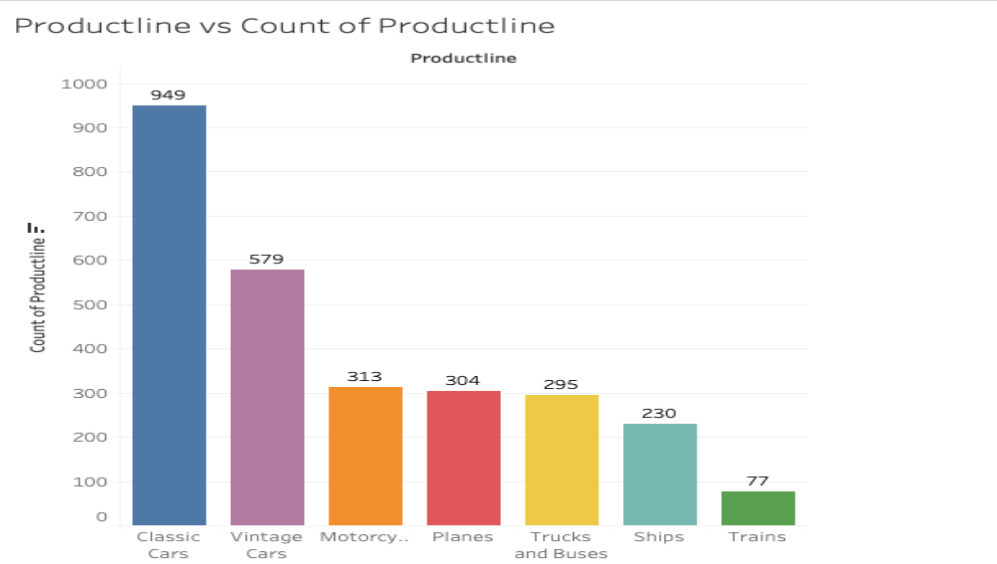
It is shown that the orders from the United States overtook the other countries. Also as an extra information the breakdown of cities are integrated to our distribution chart. At The second best country Spain it is observed that most of the purchases are made by a single city which is Madrid. Which is higher than the all individual cities in United states.



4.4.What is the distribution between product lines?

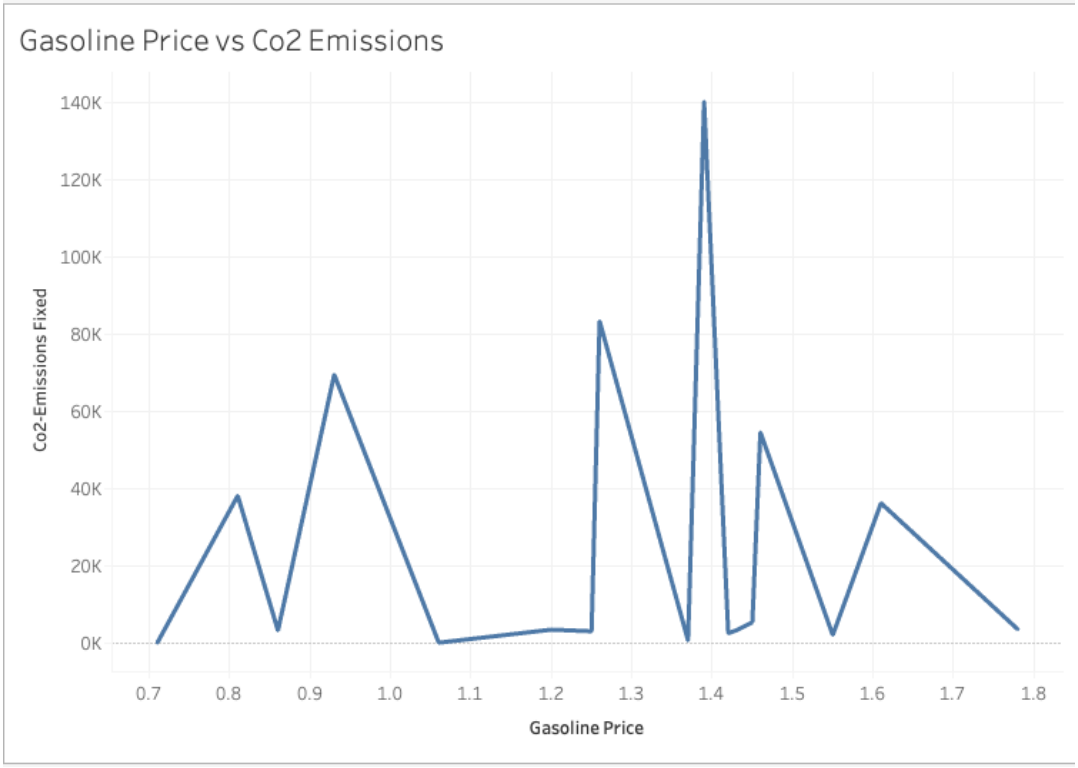
The distribution of product lines demonstrates the strong demand for automobiles, with classic cars being the most popular category. Vintage cars come next, accounting for a sizeable portion of sales as well. The market is dominated by these two groups together,

which shows how strongly consumers prefer cars. Other categories, such trucks and buses, motorcycles, and airplanes, also make significant contributions to the distribution as a whole and represent a wide variety of consumer interests.



4.5. What is the relation between Co2 emission and Gas prices?

Although it was expected to observe higher Co2 emissions with lower gas prices, the visualization did not support this hypothesis. There seems to be no correlation between these data. This can be observed from the visualization below.



5. Conclusion: To conclude, product categories and deal sizes had an impact on sales. Classic cars and vintage cars were the most popular product lines, and medium-sized bargains were the most prevalent. Additionally, the data revealed that sales in some nations were significantly higher than those in others.

The results did not show the anticipated correlation between carbon dioxide emissions and gas prices. This implies that other elements, such as energy use or legislation, have a greater impact on emissions.

This research ultimately demonstrated how data analysis may provide insightful answers to challenging topics.

https://github.com/eyluldalbudak/tableau-projecteylul