**STAT 112- Introduction to Data Processing and Visualization Project**

Interpretation of Auto Sales and World Data Information

by

**Arda Ergüven**

Department of Statistics of Middle East Technical University

Ankara, Turkey

**Abstract**

This research aims to search and find any correlation between given dataset. In order to combine these two datasets, using joints were necessary. This has been done by joining the Country columns of each dataset in Tableau. In this dataset there were some defective data. In order to continue to the research, these irregularities had to be cleaned. This data cleaning part consists of changing the data type in Tableau of some of the data, creating specified calculated fields and moving a few data from dimensions to the measures. These two datasets are explored by using appropriate statistical and visualization methods. By exploring the data deeply, some questions have been tried to answer.

**Introduction**

The main purpose of this analysis is to examine the relationship between autosales information and world data, so that the autosales company can make more reasonable decisions in their sales. This datasets includes many variables. However, 10 of them are more important and therefore used in this research. These variables are:

- Quantities of Orders: It indicates the number of items ordered in each order.

- Order Dates: It denotes the date on which the order was placed.

- Total Sales: Denotes the total sales amount for each order, which is calculated by multiplying the quantity ordered by the price of each item.

- Price of Each Order: This column specifies the price of each item in the order.

- Deal Size: It indicates the size of the deal or order, which are the categories "Small," "Medium," or "Large."

- MSRP: It stands for Manufacturer's Suggested Retail Price and represents the suggested selling price for each item.

- Status: It indicates the status of the order, such as "Shipped," "In Process," "Cancelled," "Disputed," "On Hold," or "Resolved."

- Unemployment Rates: Percentage of the labor force that is unemployed

- Gasoline Prices: Price of gasoline per liter in local currency

- CO2 Emissions: Carbon dioxide emissions in tons

**Data Tidying and Cleaning**

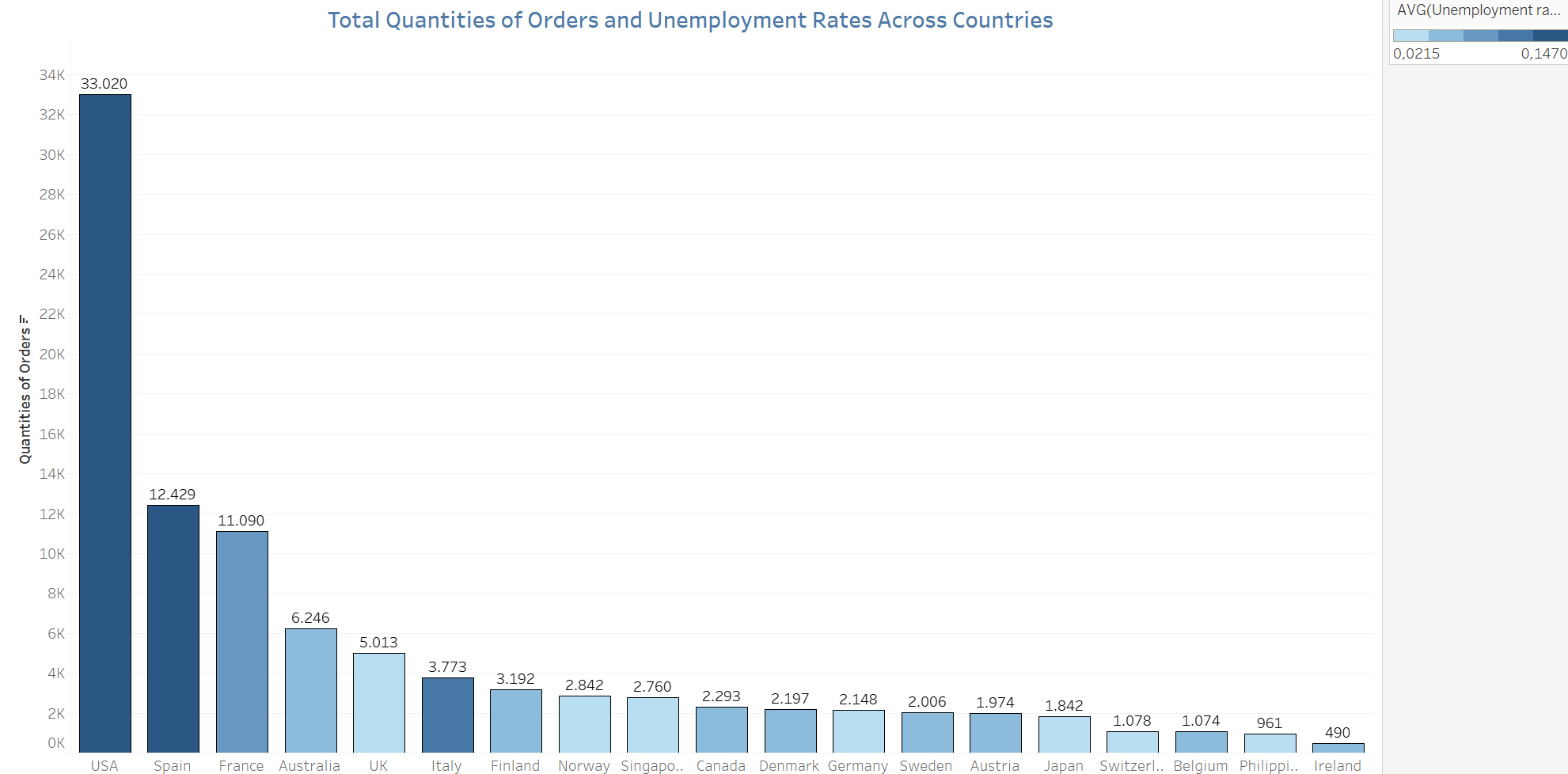
In this dataset there are a few points to be cleaned but not too many. There were no duplicates in the dataset. The steps that are taken to tidy and clean the dataset are:

1. Firstly, the inner joint of the two datasets have been provided in order to create a combined dataset. The joint was done by matching “Country” variables with an inner join.
2. Secondly, by using gasoline prices, a new calculated field have been created because its data type was string, and when it changed to float the decimal places were getting multiplied by 100. Therefore, a new calculated field have been created by using FLOAT(REPLACE([Gasoline Price], "$", "")) this code.
3. Same step in the second step have been applied for the GDP variable. Because it was an also a string variable with “$” in it. So the same code in the second step have been applied to GDP variable by changing the [Gasoline Price] to [GDP] to create a new calculated field.
4. The Order Number have been changed to a string because it is an ID variable not a numeric variable. Thus, it is moved from numerical to dimensions.

**Exploratory Data Analysis**

After the data cleaning part, 5 research questions were rise to be answered. By diving into the dataset and making some attempts to create proper visualization techniques, the answers were found.

1. **Are there any correlation between quantities of orders and unemployment rates across countries ?**



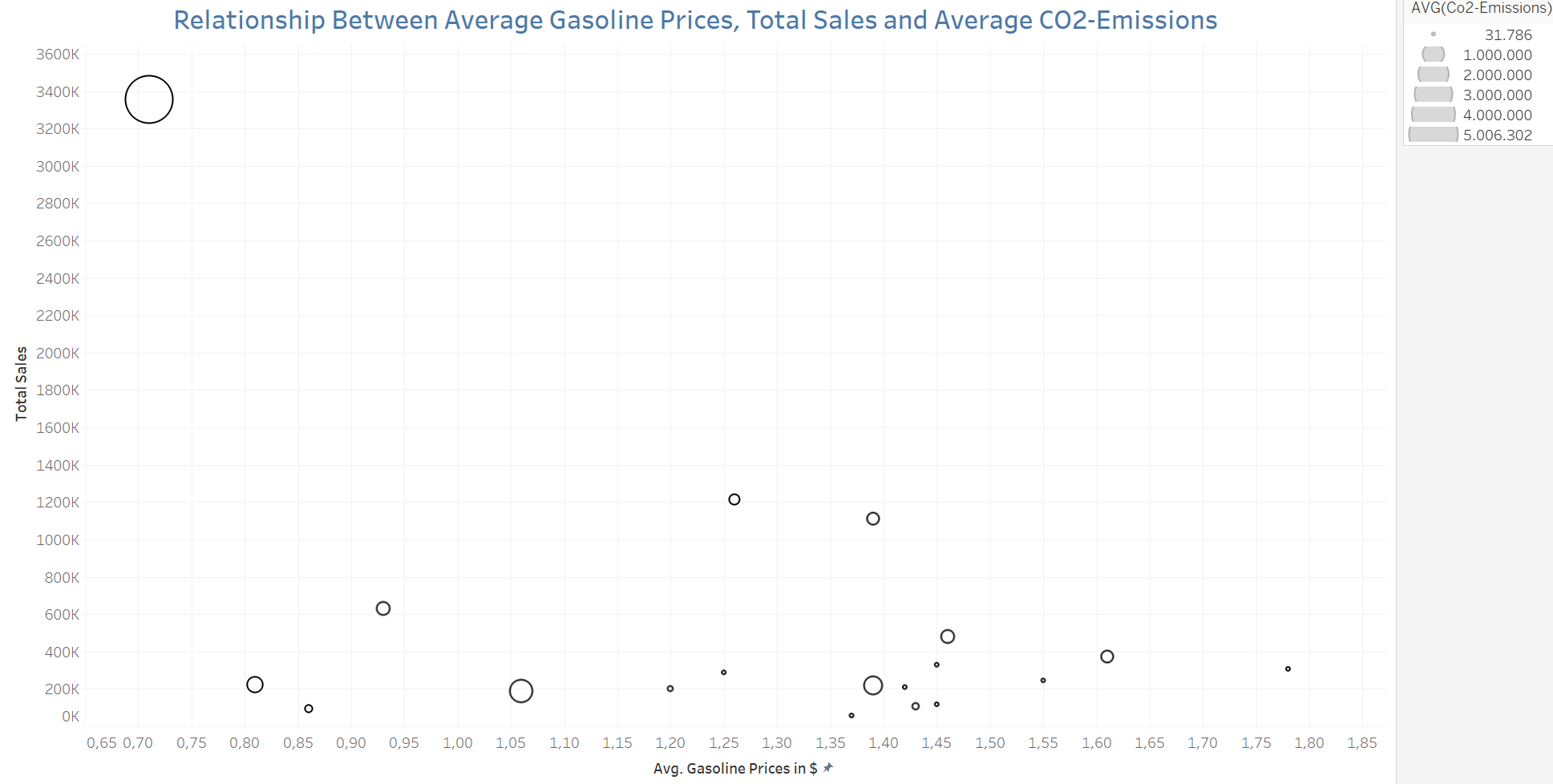
Best way to show the relationship between these variables is bar chart. As we can see the total quantities of orders on the rows and countries on the columns and the unemployment rates are showed as diverging color palette.

As we can see that there is not a strong relationship between unemployment rates and total quantities of orders. There is a small relationship of course it is tend to be a positive relationship. However, some countries obstruct us to come up with a strong relationship like Italy. We can think that “is not this related with the population of countries?”. However, the indicator we are working on is not the number of unemployeed people, it is the unemployment rate of people. So no, this most likely is not about the population whether it is about the rates of unemployeed people.

As we can observe that USA has the most quantities of orders which is 33.020 and the highest unemployment rate which is 0,147 and Japan has one of the lowest unemployment rate with a lower quantities of orders. However, as we mentioned at the top, it is hard to say that there is a positive or negative relationship between unemployment rates and order quantities. At the left side of the bar graph it gives a impression that it would be an positive relationship but Italy breaks this forecasting. And right side of Italy seems like has no relationship between unemployment rates and total quantities of orders.

In conclusion even though there is a slight relationship, it is not reliable to say there is definitely a relationship between unemployment rates and total quantities of orders.

1. **Are there any relationship among total sales, average gasoline prices and average CO2 emissions?**



As we have 3 numerical variables for this question, the best visualization technique is the bubble plot. The total sales are on the rows, average gasoline prices are on the columns and the average CO2 emissions are displayed as size of the bubbles.

We can observe that there is a slight relationship between the total sales and the CO2 emissions. As the total sales decrease, the CO2 emissions decrease too. This is normal and expected because if a country has less individual cars the CO2 emissions get lower as we expect.

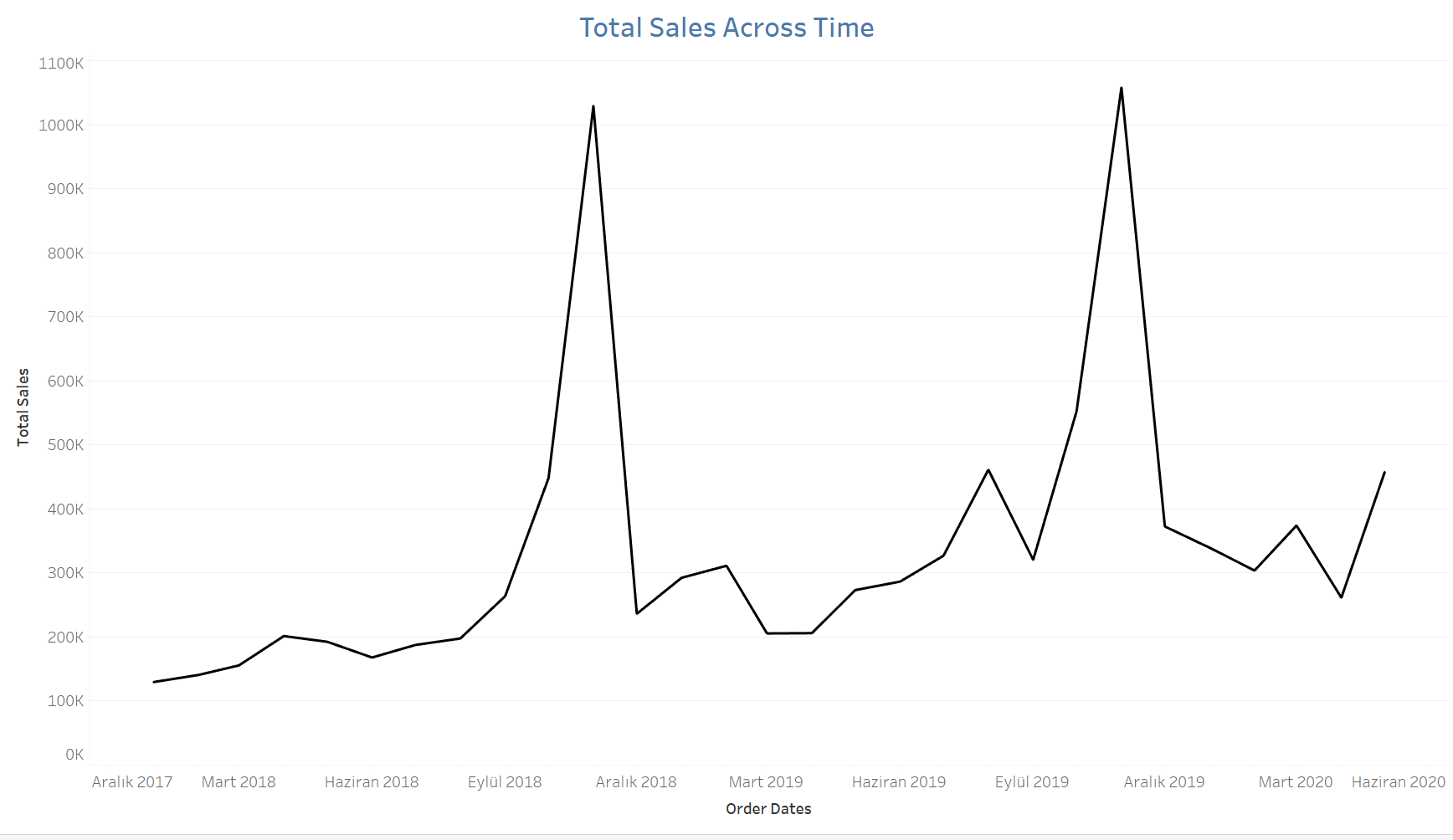
However, it is hard to say that there is a relationship between gasoline prices and total sales or CO2 emissions.

The USA has the lowest gasoline price with the highest CO2 emissions and Total Sales.

Denmark has the lowest CO2 emissions with the 245,637 sales and 1.55$ gasoline price.

In conclusion, there is a relationship between total sales and the CO2 emissions in ton. However, it is hard to say that there is a relationship between gasoline prices with CO2 emissions or Total Sales.

1. **How do the sales differ in periods of months ?**



The best way to show the distribution of total sales in periods of months is to using a line graph.

As we can see in the line graph, the sales get to its peak on November every year. It also has a major difference between December and November. The sales suddenly drop down nearly to the bottom after November.

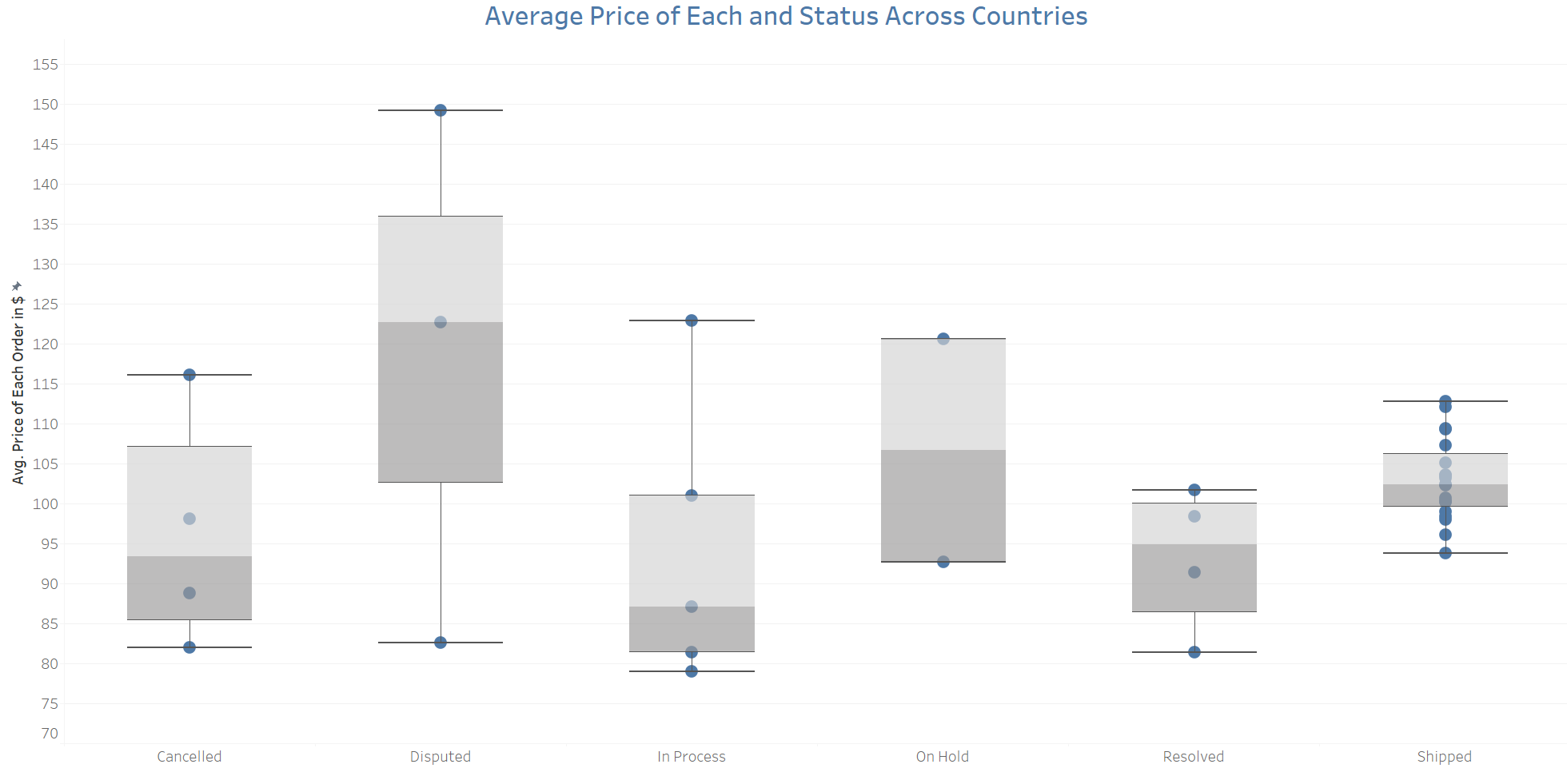
In November of all years the total sales exceed the 1,000,000$

In 2018; It is mostly increasing on January 2018 to November 2018 and decreasing on November to December

In 2019; It decreases on March despite of 2018, and it increases on March 2019 to August 2019 then it goes down on September, after that it goes to its peak again on November and crashes down on December again.

In conclusion, the total sales are increasing over the years, but in a year it always crashes down on December. However, the previous December sales are always less than the next December sales. So in general, the total sales are increasing.

1. **How is the distribution of average price of each order and status of orders ?**



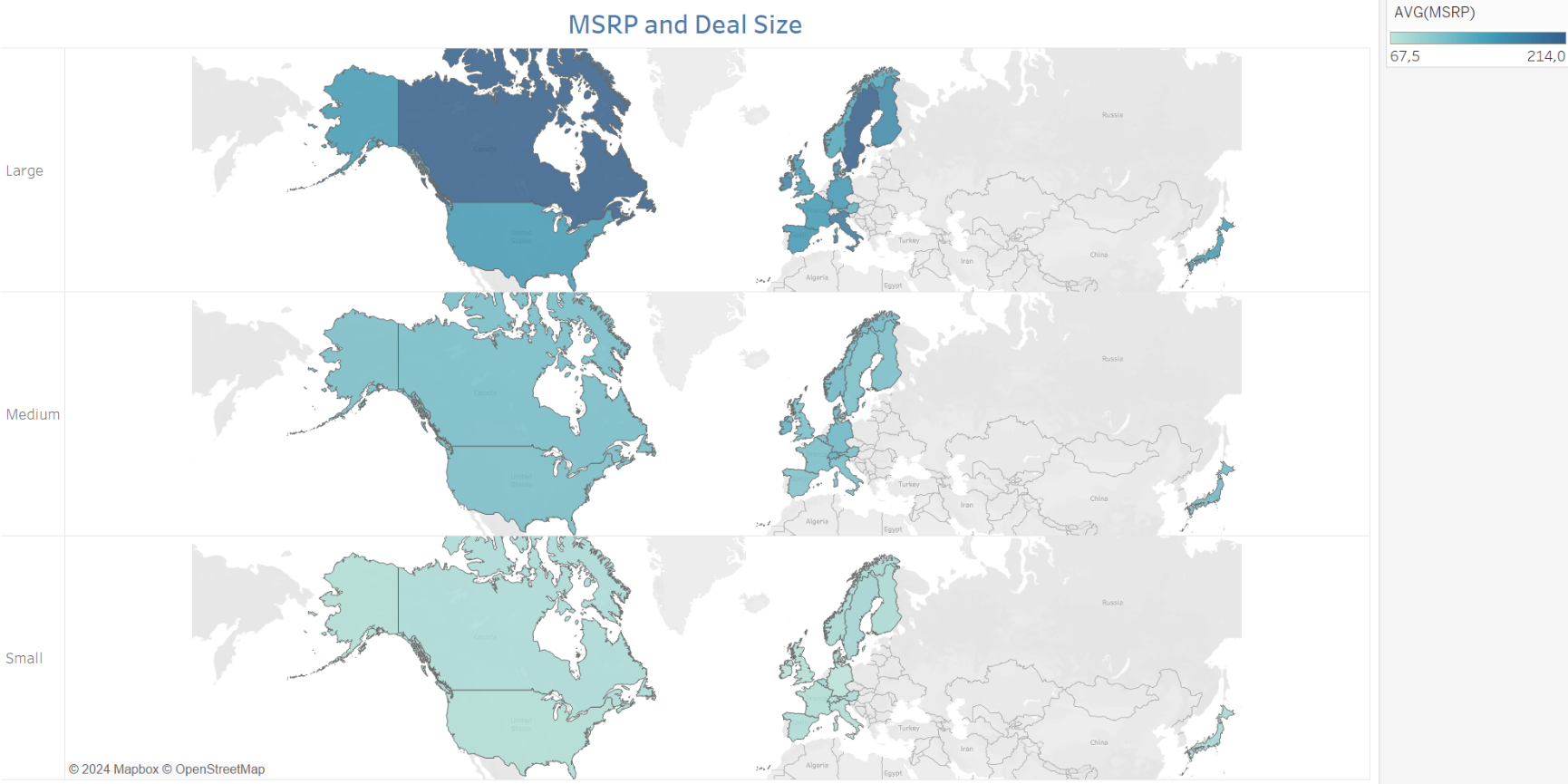
We choose to use box plot to show the distribution of average price of each order and the status of orders.

We can easily see that orders that are in the shipped status are more compiled and have nearly a symmetric distribution with a median of 102.40 .

On the other hand, the disputed status has more spreaded distribution with a median of 122.69 .

There are no outliers in any kind of status of orders. Each prices are on the boundaries between the minimum and maximum values.

1. **How the average MSRP and Deal Size are related with each other?**



We can see the deal sizes on the rows and average MSRP values as a diverging color palette on three world maps.

As we can observe that there is a high relationship between the deal size and the MSRP values.

The large deal size indicates that higher MSRP values. Therefore, small deal size indicates the lower MSRP values.

Canada and Sweden have the highest MSRP value on large deal size which is 214.0 .

Austria has the lowest MSRP value on large deal size which is 129.3 .

On medium deal size almost every country have a MSRP price that is close to other countries’ MSRP values. The maximum is on Switzerland and it is 131.0 . The minimum is on Spain and it is 108.5 .

The same result is prevailing on the small deal size. Almost every country have a MSRP price which is a close value to other countries’ MSRP values. The maximum is on Belgium and it is 85.5 . The minimum is on Germany and it is 69.5 .

**Conclusion**

To conclude, this project explored the relationships between auto sales data and global indicators such as unemployment rates, gasoline prices, and CO2 emissions. The analysis combined two datasets through data tidying and cleaning, followed by comprehensive exploratory data analysis using visualization techniques.

•We find the suggestion that unemployment rates alone are insufficient to predict order quantities reliably.

•A clear relationship was identified between total sales and CO2 emissions, indicating that lower sales correlate with reduced emissions. However, no consistent pattern was observed between gasoline prices between total sales or CO2 emissions.

• Sales always peak in November and then drop in December. Overall sales were higher compared to the previous year, reflecting an upward trend.

•Shipped orders exhibited a more consistent and symmetrical price distribution, whereas the prices of disputed orders demonstrated greater variability. There were no outliers in either category.

• There was a strong correlation between deal sizes and their corresponding MSRP values; large deal sizes commanded high values of MSRP. This trend was consistent across all countries.

<https://public.tableau.com/views/Project1Arda_Ergven_2666550/Dashboard?:language=en-US&:sid=&:redirect=auth&:display_count=n&:origin=viz_share_link>

<https://github.com/ArdaErguven/TableauDashboard>