STAT 112 - Introduction to Data Processing and Visualization Project

Elin Aliev

Department of Statistics,

Faculty of Arts and Sciences,

Ankara, Turkey,

elinaliev2006@gmail.com

Introduction

In this research I prepared an analysis of two datasets (world-data-2023 and autosale) and their union. I have prepared multiple questions, which can occur by examining this data. This research can be interesting to some individuals, who are looking forward to analyzing Auto Company activity, their sales, customers, approximate profit and some reliances between some data of the different countries and the company activity in these countries.

In this research I faced multiple data variables from both data sets.

Autosale variables:

ORDER NUMBER - This column represents the unique identification number assigned to each order.

QUANTITY ORDERED - It indicates the number of items ordered in each order.

PRICE EACH - This column specifies the price of each item in the order.

ORDER LINE NUMBER - It represents the line number of each item within an order.

SALES - This column denotes the total sales amount for each order, which is calculated by multiplying the quantity ordered by the price of each item.

ORDERDATE - It denotes the date on which the order was placed.

DAYS\_SINCE\_LASTORDER - 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 - It indicates the status of the order, such as "Shipped," "In Process," "Cancelled," "Disputed," "On Hold," or "Resolved."

PRODUCT LINE - This column specifies the product line categories to which each item belongs.

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

PRODUCTCODE - This column represents the unique code assigned to each product.

CUSTOMER NAME - It denotes the name of the customer who placed the order.

PHONE - This column contains the contact phone number for the customer.

ADDRESSLINE1 - It represents the first line of the customer's address.

CITY - This column specifies the city where the customer is located.

POSTAL CODE - It denotes the postal code or ZIP code associated with the customer's address.

COUNTRY - This column indicates the country where the customer is located.

CONTACT LASTNAME - It represents the last name of the contact person associated with the customer.

CONTACT FIRSTNAME - This column denotes the first name of the contact person associated with the customer.

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

World data variables:

**Country** - Name of the country

**Latitude** - Latitude coordinate of the country's location

**Longitude** - Longitude coordinate of the country's location

**Birth Rate** - Number of births per 1,000 population per year

**CO2-Emission** - Carbon dioxide emissions in tons

**CPI** - Consumer Price Index, a measure of inflation and purchasing power.

**Gasoline Price** - Price of gasoline per liter in local currency

**GDP** - Gross Domestic Product, the total value of goods and services produced in the country

**Life expectancy** - Average number of years a newborn is expected to live

**Population** - Total population of the country

**Tax Revenue(%)** - Tax revenue as a percentage of GDP

**Total tax rate** - Overall tax burden as a percentage of commercial profits

**Unemployment rate** - Percentage of the labor force that is unemployed

**Urban population** - Percentage of the population living in urban areas

# Data Tidying and Cleaning Process

* Firstly, I changed the data type of variables GDP, Gasoline Price, Total Tax revenue, Life expectancy and Population variables from strings to the numbers (decimal).
* Also, I changed latitude and longitude to dimensions and converted them to discrete variables.
* From the analysis in the dataset world-data-2023 some errors were found:

1. In the 152 row of the table there wasn't information about the country name, its latitude and longitude, missing information was pasted.
2. There was some missing information in several columns about some countries such as Vatican City, Palestinian National Authority, Tuvalu, Eswatini, Andorra etc. In the case of the difficulty of finding appropriate and accurate information on the net, I decided to not use information from not appropriate sources, so I pasted ‘no info’ in all empty cells.

* From the analysis in the dataset autosale some errors were found:

1. I found a great number of duplicates with sorting by Order Number, deleted them and left only the original data.

From studying the variables, we have:

* Categorical : Contact First Name, Contact Last Name, Customer Name, Deal Size, Phone, Postal Code, Product Code, Product Line, Status, Order line number, Order number,
* Numerical: Gasoline Price, GDP, Days Since Last Order, MSRP, Price each, Quantity ordered, Sales, Birth Rate, Co2-Emissions, CPI, Life expectancy, Population, Tax revenue, Total tax revenue, Unemployment rate, Urban\_Population
* Date: Order Date
* Location: Address Line1, City, Country, Latitude, Longitude

1. How does urban population, tax revenue and MSRP influence the sales amount of the Auto Company in different countries?

To research this question, I created a scatter plot, where I chose such values as Avg. Tax Revenue(%), Avg. Urban Population and Avg. MSRP to be variables and a variable Sales to be a function value:



As Figure 1.1

We can see from the colour range, the number of sales of the company increases while blue colour transforms to purple, pink, yellow, orange and finally, to red. According to the colour range and function value we can see, the USA is the country with the greatest number of sales and Ireland has the least number of sales.

By investigating the scatter plot, we can not conclude that we do not have any relationship between Avg. Tax revenue(%) and Sales, for example, in the country with the highest number of sales (USA) we can see that there is also lowest Avg. Tax revenue. On the other hand, in France, Avg. Tax Revenue is quite high, otherwise, France is also on the third place by sales.

Observing the second case, we can easily see the direct correlation between Sales and Avg. Urban Population, such countries as USA, Spain, France, Italy etc., which have impressive urban populations, also they have a great number of sales. So, we can conclude that urban population number influences the transport usage rate.

In the third example, we can consider that there is a slightly inverse relationship between Avg. MSRP and Sales. Most of the countries with the high sales amount have Avg. MSRP vault less than 160.

2.How the number of items ordered in each order was changed from 2018 to 2020?

This question can be interesting for people who are analyzing the condition of Auto Company and all transport markets throughout the 3 years.

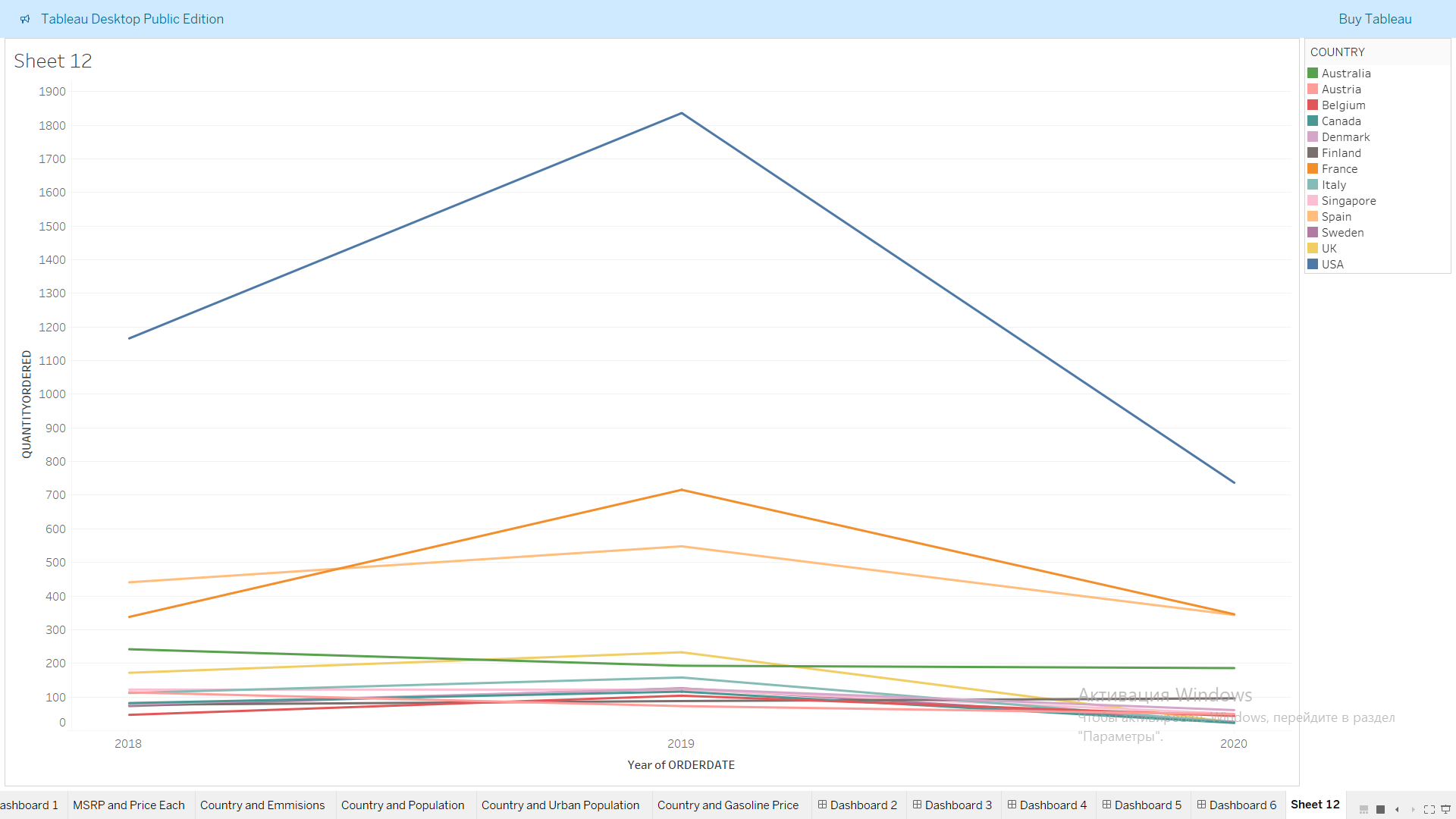


Figure 2.1

From this visualization, we can consider that sales of Auto Company in each country decreased significantly, the greatest fall of sales was observed in the USA. I consider that Auto Company in particular and the whole market in general are in a state of decline and crisis for some reasons. In my opinion several reasons can lead to these circumstances. Firstly, it can be Covid-19 that took place in 2020 and resulted in a great world economic crisis. Secondly, the development of the electrical transport industry could lead to such circumstances. Last but not least, the continuing city urbanisation and the development of the public transport system in highly urbanized countries, such as the USA or France, can result in the decrease of the private transport demand.

3. In the next question I wanted to examine the relationship between Country, its GDP per person (I created this variable by dividing GDP variable by Population variable using calculated field in Tableau) and Life expectancy.

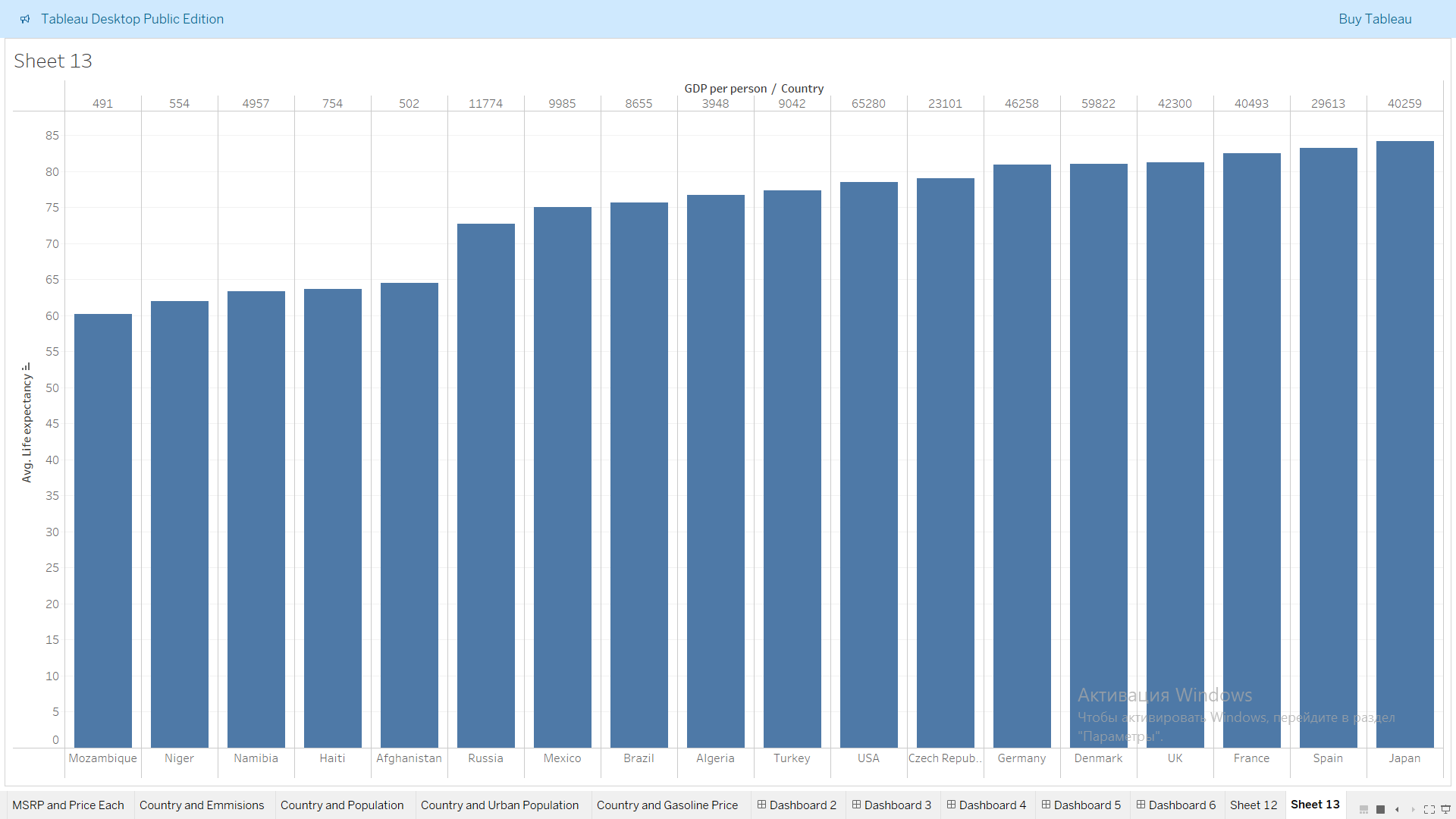


Figure 3.1

By examining the bar chart we can be sure about positive correlation between these variables. The Avg. Life Expectancy value increases with the amount of GDP per Person in different countries. We can consider a positive trend, that in general as the wealth of citizens increases, their life expectancy rises too.

4. In the fourth question I wanted to understand why some countries, which have a significant population can produce less CO2-Emissions than those, where population is less.

As we can see from the graphs in figure 4.1 below, we can observe that Japan, Philippines and Germany are population leaders in this data set, but they are not leaders when it comes to CO2 production.

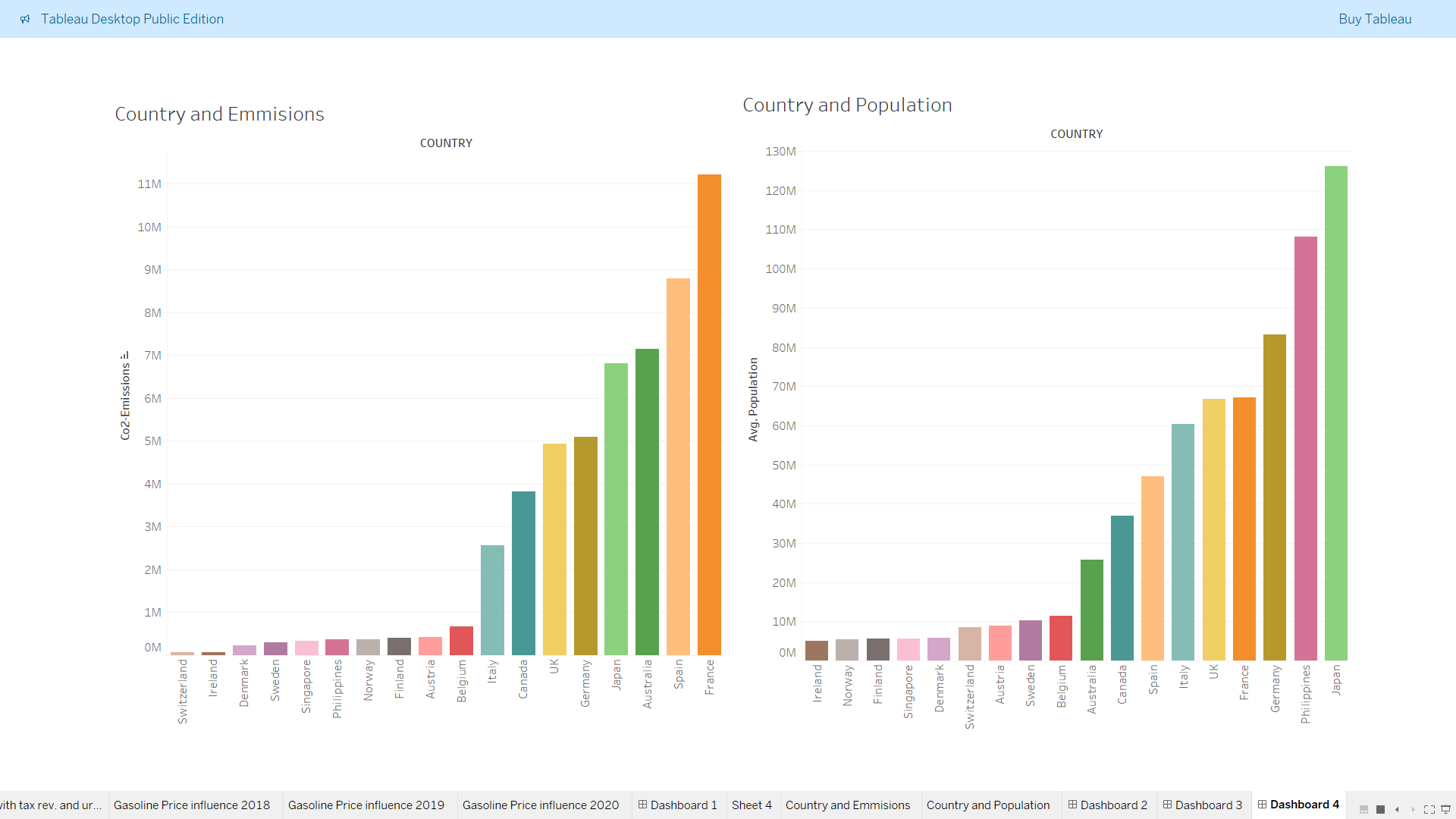


Figure 4.1

Firstly, we can observe the bar chart in figure 4.2 about urban population in these countries. Now the leaders are Japan, UK and Germany. For example in the Philippines more than half of the population does not live in big cities, their demand for their own vehicles is much lower than the population which lives in towns. According to this information and the fact that there are much fewer factories and enterprises which can produce CO2 in the Philippines than in France or Japan, the Philippines and similar countries produce less CO2.

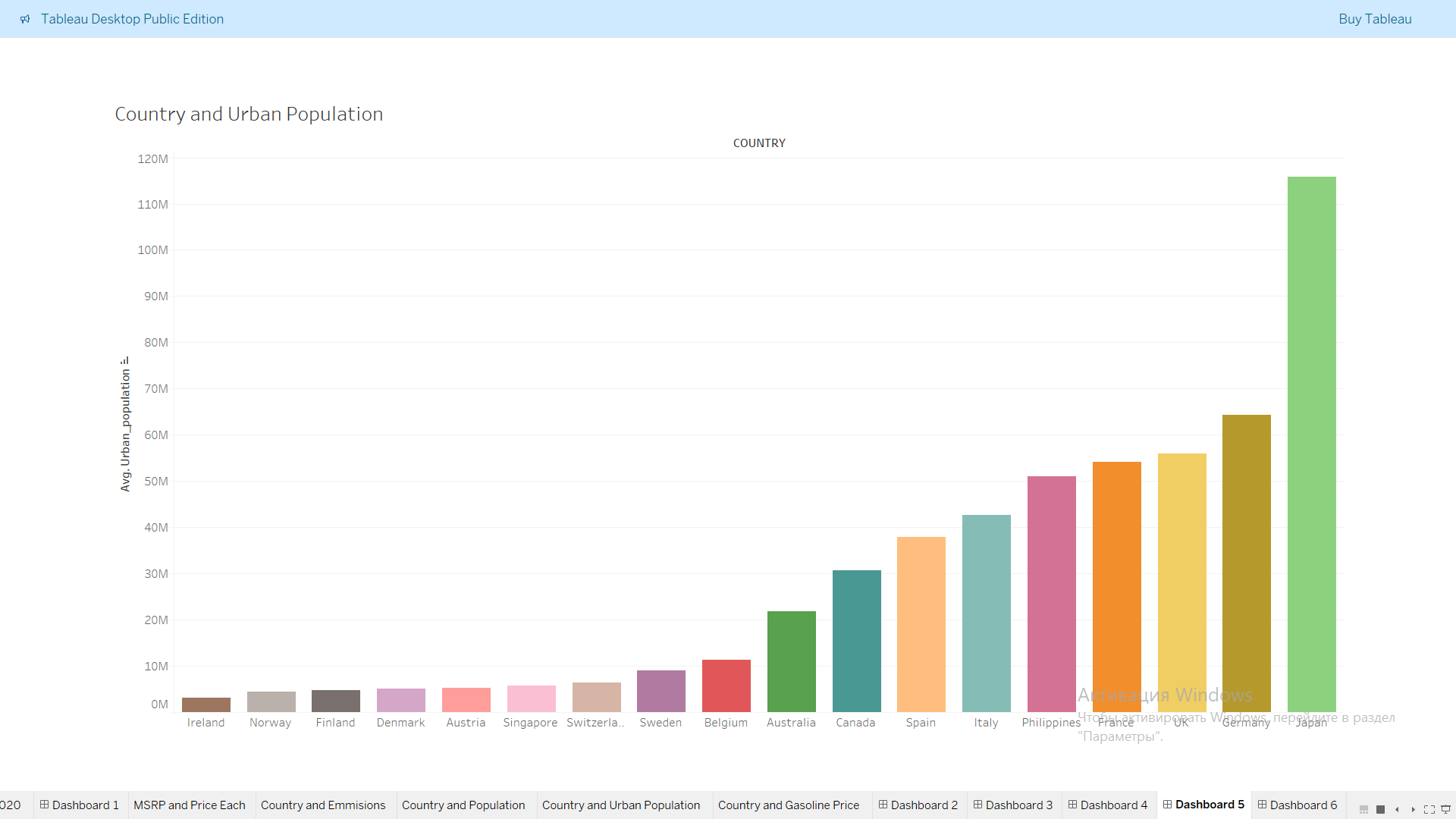


Figure 4.2

Secondly, let's consider the bar chart in figure 4.3 about Gasoline Price in these states. The gasoline price is an important factor that is taken into account by any person who is planning to buy a vehicle. As we can see, in this chart leaders are Norway, Italy and Denmark, countries which have lowest rates of CO2-Emission. One of the reasons for such prices can be the economic policy of the States of Norway, Italy, Denmark and similar countries, but the fact that these states are having the lowest CO2-Emission rates is clear.

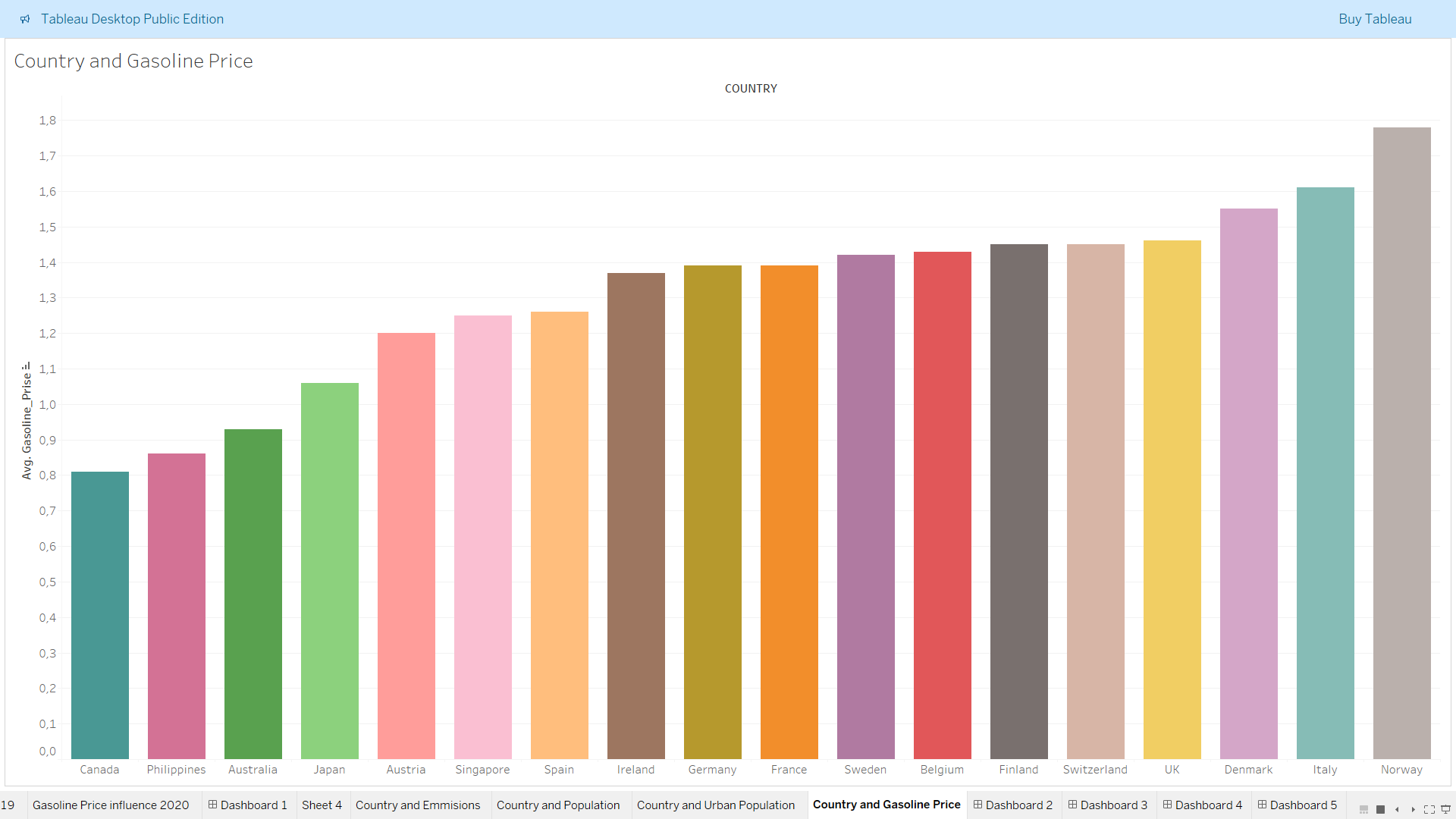


Figure 4.3

So, while answering the fourth question we are convinced that a population is not the only factor that influences CO2-Emission, there are more other factors, such as urban population and gasoline price.

5. In the fifth question let's examine the relationship between MSRP and price for each item in several countries in 2019.

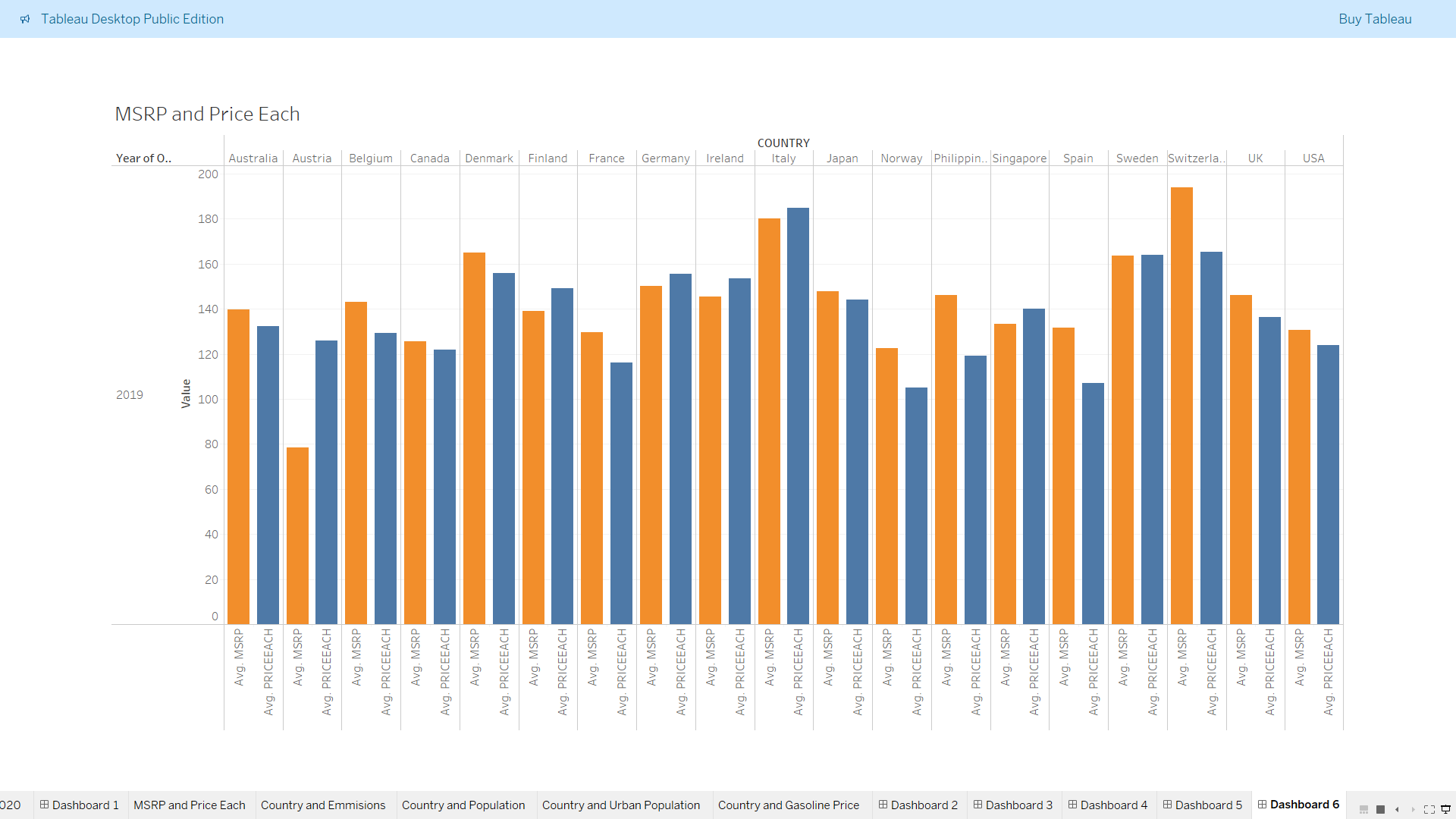


Figure 5.1

In 2019 the values of MSRP and prices for each item varied from country to country, in some states MSRP value was greater than PRICE EACH, in some PRICE EACH was greater than MSRP, in some countries these figures were almost equal. After the analysis I considered that in countries where PRICE EACH was greater than MSRP, the demand for Auto Company production was greater than its supply, which is why the company raised the price(Austria, Finland, Germany, Ireland and Singapore). On the other hand, in countries where MSRP value was greater than PRICE EACH demand for the company production was less than its supply, because of that company dropped prices (Australia, Belgium, Denmark, France, Norway, Philippines, Spain, Switzerland and UK). Last but not least, market equilibrium has almost been achieved in such countries as Canada, Italy, USA, Japan and Sweden.

Conclusion

In conclusion, we can see this data from different perspectives, and use it for several researches. By analyzing the world dataset we can answer different questions about macroeconomics, demography, geography, human living standards etc.. By analyzing the autosale data set, we can create some conclusions about Auto Company activity. If we are observing auto sales and world data datasets together, we can answer some questions about all vehicle sales markets.

My dashboard:

https://public.tableau.com/app/profile/elin.aliev/viz/Project\_17336842698620/Dashboard1?publish=yes