**Analyzing Auto Sales Across Europe and Global Trends**

Mehmet Tatbak   
 *Department of Statistics*   
 *Middle East Technical University* Ankara, Türkiye  
 mehmet.tatbak@metu.edu.tr

*Abstract*—This research is about automobiles sales and some world data in different locations. The data required some tidying and cleaning to ensure suitability for analysis.Research questions primarily centered on examining orders and sales details across Europe countries. As a result of our research questions interpreted relationships between several numerical and categorical data.

Keywords—data tidying, regression, skewness, outlier

# Introduction

In this research, ı worked with two datasets: auto-sales and world-data-2023, both in Excel format. The auto-sales dataset includes 20 variables, such as Ordernumber and Sales, while the world-data-2023 dataset has 14 variables.

To analyze the data, I combined the two datasets using the joining method in Tableau. During the initial review, I noticed that some variables had incorrect data types. I corrected these issues to ensure the data was accurate and ready for analysis.

The steps and details of the data cleaning process are explained in the following section.

1.1 *Data Description*

|  |  |
| --- | --- |
| **Automobile-Sales Dataset** | **Country-Information** |
| ORDERNUMBER(Numerical) | Country(Geographical) |
| PRICEEACH(Numerical) | Birth Rate(Numerical) |
| MSRP(Numerical) | CO2-Emission(Numerical) |
| COUNTRY(Geographical) | CPI(Numerical) |
| ORDERDATE(Date) | Gasoline Price(Numerical) |
| PRODUCTCODE(Categorical) | GDP(Numerical) |
| QUANTITYORDERED(Numerical) | Life expectancy(Numerical) |
| SALES(Numerical) | Population(Numerical) |
| ORDERLINENUMBER(Numeric) | Total tax rate(Numerical) |
| PRODUCTLINE(Categorical) | TaxRevenue(%)(Numerical) |
|  | Unemployment rate(Numerical) |

*Note*: It only includes the variables I used in my analysis. It does not include all the variables.

# data tidying and cleaning

In the data pre-processing phase, my primary goal was to thoroughly examine all variables and address any duplicates, missing values, or incorrect data types. Irrelevant data that did not impact my research or not attached with my research questions was excluded from the source. This section outlines the steps taken during preprocessing:

-I corrected data types.

-I identified and removed duplicate rows, followed by an assessment of outliers; any necessary outlier removal was performed.

-Finally, for missing values, I employed imputation techniques to fill in the gaps where appropriate.

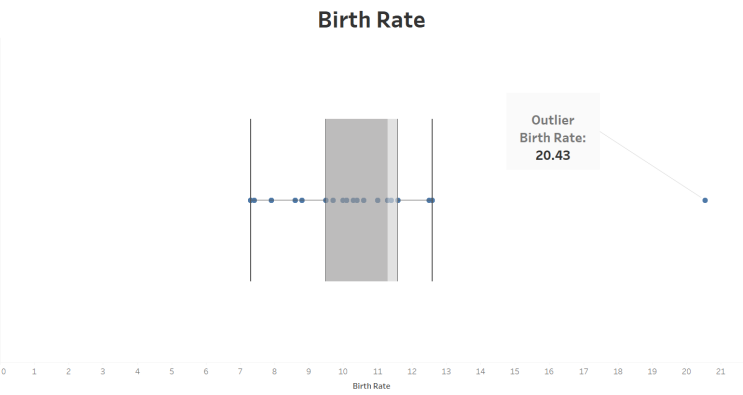
## 2.1 Data Types

During the data analysis, I identified that some variables had incorrect data types. In the world-data-2023 dataset, for instance, the Gasoline Price column was incorrectly categorized as a categorical variable due to the presence of a dollar ($) sign. To address this, ı created a calculated field and applied the *“REPLACE”* method to remove the dollar sign. Following this adjustment, I converted the Gasoline Price column to a numerical data type. A similar issue was observed with the GDP variable, which was also categorized as a categorical variable. Using the same method, I corrected its data type to numerical with “*REPLACE”* method. In contrast, the auto-sales dataset did not contain any incorrect data types.

## 2.2 Duplicate Rows

To check for duplicate rows, I used Excel's *conditional formatting feature*. I selected all the cells in the dataset, then navigated to the *Conditional Formatting* menu, chose *Highlight Cells Rules*, and selected *Duplicate Values*. However, upon applying this method, I did not observe any duplicate values in either of the datasets.

## 2.3 Outliers

Some variables in the dataset contain outliers, which ı identified through visual analysis. To check for outliers, ı used Tableau's visualization tools, specifically the box plot and histogram. These visualizations revealed several points that were identified as outliers. 

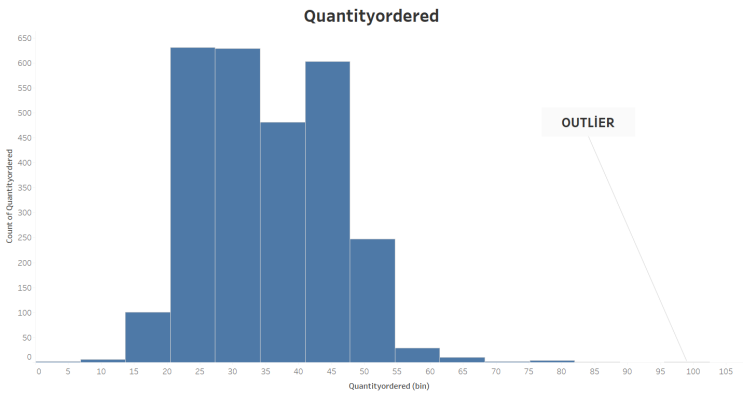
When analyzing the birth rates of various countries using a box plot, I observed the following:

-The median birth rate falls within the 11–12 interval.

-The minimum birth rate is within the 7–7.5 interval.

-The maximum birth rate is within the 12.5–13 interval.

-One country, however, has a birth rate of 20.43, which is identified as an outlier.



Based on the histogram analysis of the "Quantity Ordered" variable:

-The majority of orders fall within the range of 20–50 units.

-There are relatively few orders below 20 or above 50 units.

-An outlier is observed which is significantly higher than the other values in the dataset.

However, after further examination, I determined that these outliers were logical and consistent with the nature of the data. As such, I chose not to remove any of the outliers, as they did not appear to distort the overall analysis or the integrity of the data.

## 2.4 Missing Values

The first dataset (automobile sales) does not contain any missing values; however, the world dataset has missing values in the birth rate column.

-The maximum value for the birth rate is 46.08,

-The minimum value is 5.90,

-The median is 17.95,

-The average is 20.21.

The birth rate distribution is right-skewed, and the range (maximum - minimum) is quite large. Therefore, I preferred imputing the missing values with the median. The same approach was applied to handle missing values in other columns of the dataset as well.

*Note*: Although missing values were no longer visible after performing the inner join operation in Tableau, I updated the dataset in Excel format to ensure accuracy and consistency.

# Exploratory Data Analysis

*3.1 Descriptive Statistics*

The Autosale dataset contains sales-related data, including variables such as Sales, Price, and Order Number.

Below is the summary of key variables:

* Sales:

Mean: 70.44

Median: 34.00

Distribution:Right-skewed

Interpretation: Most sales transactions are smaller in value, but a few high-value sales increase the average significantly.

* Price:

Mean: 101.1

Median: 95.6

Distribution: Slightly right-skewed

Interpretation: Prices are generally concentrated around the median, with some higher-priced items stretching the distribution to the right.

* Order Number:

Mean: 10,259.1

Median: 10,264

Distribution: Slightly left-skewed

Interpretation: Most order numbers are close to the median, with a slight concentration toward lower values.

The World Data 2023 dataset includes country-level statistics such as Birth Rate, Life Expectancy, and Tax Revenue.

Below is the summary of key variables:

* Life-Expectancy:

Mean: 72.28

Median: 73.20

Distribution: Slightly left-skewed

Interpretation: Life expectancy is concentrated around the median, with a slight emphasis on higher values.

* Birth Rate:

Mean: 20.21

Median: 17.95

Distribution: Right-skewed

Interpretation: While most countries have moderate birth rates, a few countries with very high birth rates drive the average upward.

* Tax Revenue:

Mean: 0.1657

Median: 0.1630

Distribution: Slightly right-skewed

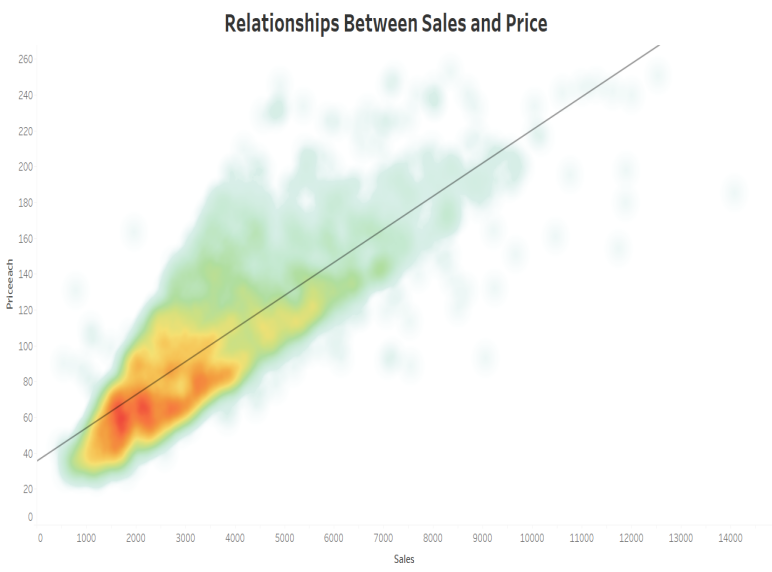
Interpretation: Tax revenues are primarily concentrated around the median, with a small number of countries showing higher values.

*3.2 Research Questions*

*3.2.1 How does the relationship between item Sales and Price ?*

Understanding the relationship between price and sales is very important for making smart business decisions. If a company or producer knows how price affects the number of products sold, it can plan better and increase its sales. Sometimes, people might think that expensive products always sell less, but this is not always true. By looking at the data, we can see the real pattern.

To check this, I used a scatter plot. The scatter plot shows price on the y-axis and sales on the x-axis. Each point on the plot represents a product, with its price and the number of units sold. By looking at this chart, we can better understand if there’s a connection between price and sales, and whether higher prices lead to fewer sales or not. This helps the company make better decisions about pricing and sales strategies.



Visualization 1 : Relationships Between Sales and Price

*Interpreting the Visualization 1:*

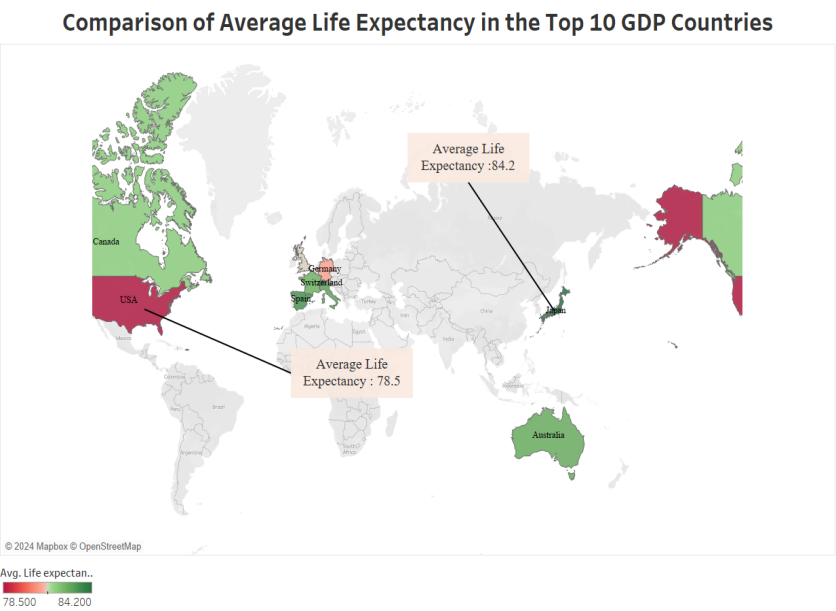
Firstly, both sales and price are continuous variables, and the number of sales tends to cluster within the 40-80 price range. When the price is more affordable, individual sales may be lower, but the overall sales volume can still be quite high.

When we look at the scatter plot, we can see that as the number of sales goes up, the total sales volume goes down. However, it’s also clear that as the number of sales increases, the price per unit tends to go up. This shows a positive linear relationship between the two variables, which we can see clearly in the trend line on the graph. The trend line supports this, suggesting that more sales are usually linked to higher prices.

*3.2.2 How does the average life expectancy compare among the top 10 countries with the highest GDP?*

I want to analyze whether the average life expectancy of rich countries is similar or quite different by using the top 10 countries with the highest GDP. The countries I will analyze are the USA, Canada, the UK, Spain, Switzerland, Italy, France, Germany, Australia, and Japan.

I used a map for this analysis and colored it based on average life expectancy. The most important thing is that the top 10 GDP countries have a small range in life expectancy. On the color scale, red represents below average, and green represents above average life expectancy.

**

Visualization 2 : Comparison of Average Life Expectancy in the Top 10 GDP Countries

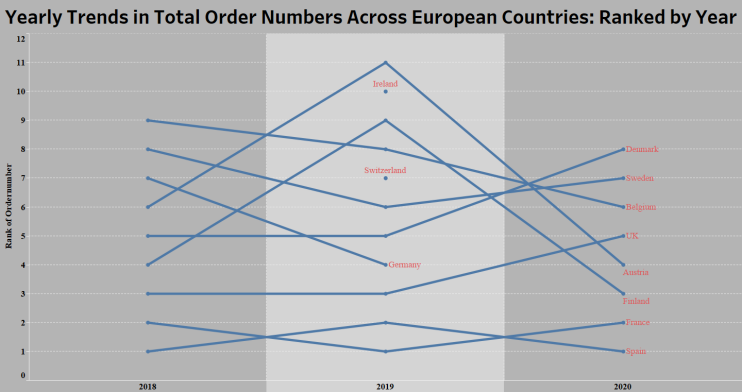
*Interpreting the Visualization 2:*

When we analyze the map, some countries are shown in red and some in green. The country with the most red is the USA, which has an average life expectancy of 78.5 years, the lowest in the top 10. On the other hand, Japan has the highest average life expectancy at 84.2 years. In Europe, the values are quite similar, but Germany’s life expectancy is slightly lower than other European countries. The range of life expectancy is close, with the minimum being 78.5 and the maximum 84.2. However, due to the color divergence, we can still see that there are noticeable differences in life expectancy among the top 10 countries.

*3.2.3 How do car orders in Europe vary over time?*

I want to study how the rankings of car orders change for European countries over three years (2018, 2019 and 2020). Instead of looking at the total number of car orders, I will focus on how countries are ranked each year.

I used a slope chart for this analysis because it is very useful for checking trends by year. The y-axis shows the ranking, and the x-axis shows the years 2018, 2019 and 2020. I used a different color for the 2019 column to make the analysis easier.



Visualization 3: Yearly Trend in Total Order Numbers Across European Countries:Ranked by Year

*Interpreting Visualization 3:*

In 2018, Spain had the highest number of orders among European countries, ranking number one. However, in 2019, Spain moved to rank two, and France increased its rank from two to one. Finland’s order numbers decreased in 2019 compared to other European countries, but in 2020, its order numbers increased, putting Finland in rank four.

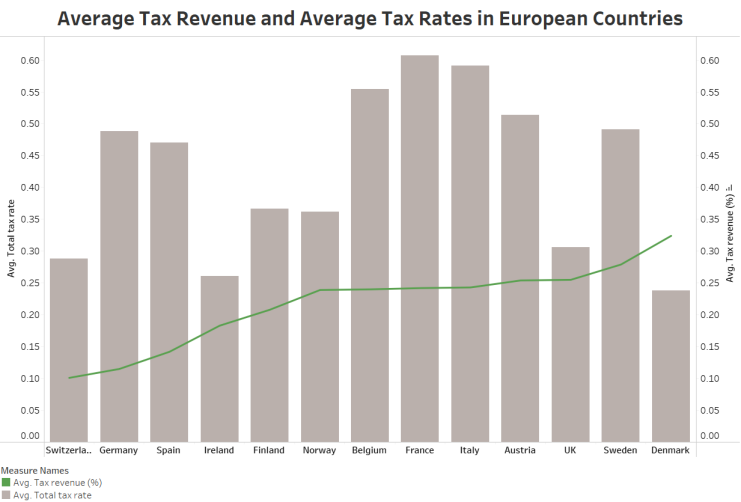
Germany’s order numbers increased in 2019, but there were no orders in 2020. Another observation is that Switzerland and Ireland only had orders in 2019. We have no data for 2018 or 2020 for these countries, or they simply made orders only in 2019.

We can say that France and Spain are the most stable in order numbers based on the ranking system.

*3.2.4 Comparison of Average Tax Revenue and Average Tax Rate Among European Countries*

I want to compare the average tax revenue and average tax rate among European countries. My goal is to see how these countries differ in their tax systems. The countries I will analyze are France, Germany, Spain, Italy, the UK, Sweden, Switzerland, Denmark, Ireland and Finland.

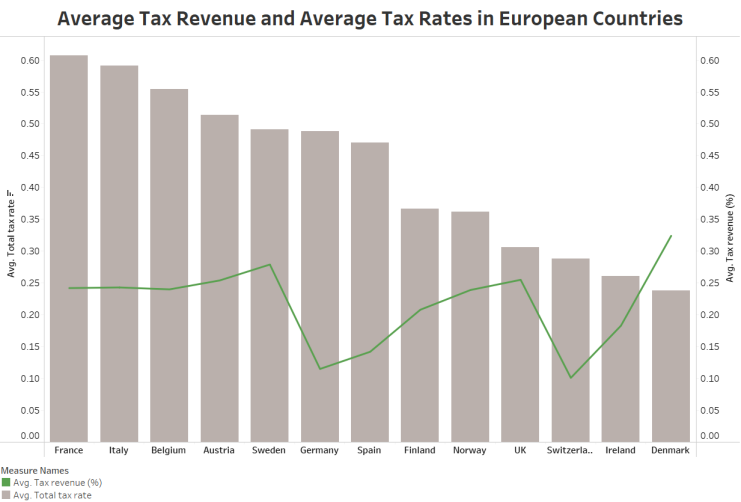
I used both a bar chart and a line chart with dual axes and synchronized the axes. The bars represent the average tax rate, and the line represents the average tax revenue.

**

Visualization 4: Average Tax Revenue and Average Tax Rates in European Countries

*Interpreting Visualization 4:*

When I analyze the visualization, it is easy to see that Belgium, Italy, and France have very similar tax revenue rates. However, France has the highest total tax. In general, as tax revenue increases, total tax also increases. Denmark has the highest tax revenue in Europe, but at the same time, it has the lowest total tax. This suggests that other taxes in Denmark are lower compared to other European countries.

**

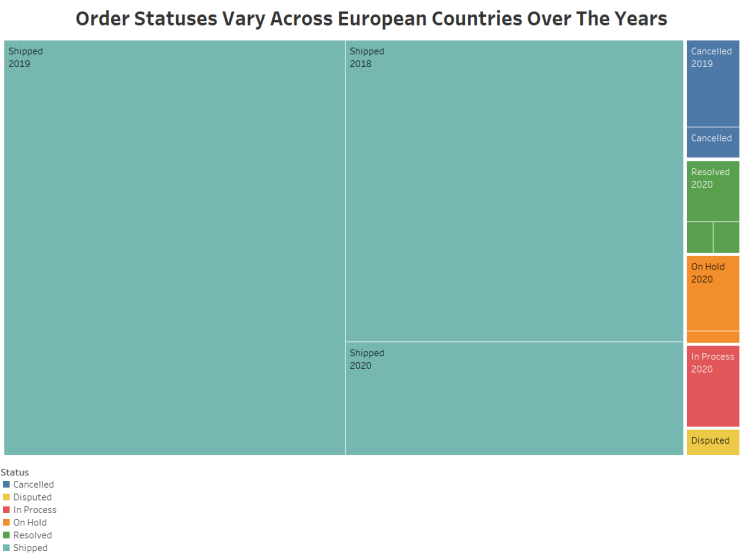
Visualization 5: Average Tax Revenue and Average Tax Rates in European Countries

When you look at visualization 5, we can see that as tax revenue decreases, total tax also decreases. However, there are some countries that do not follow this pattern, such as Ireland and Denmark.

*3.2.5 How do the order statuses vary across European countries over the years in the vehicle sales data?*

I chose this question to investigate the order statuses in European countries. There are a total of six statuses: cancelled, disputed, in process, on hold, resolved, and shipped.

For this analysis, I used a heatmap, with statuses and dates as the main parameters, and total orders in European countries as the label.



Visualization 6: Order Statuses Vary Across European Countries Over The Years

*Interpreting Visualization 6:*

When we examine the heatmap, we can see that between 2018 and 2020, the majority of orders in Europe were in the shipped status, followed by the cancelled status. In 2020, there were no orders with the cancelled status. After that, the resolved status came next, followed by on hold. In 2018, there were no orders in the on hold status. Then came the in process status, which only appeared in the orders of 2020. The least common status was disputed, which only appeared in orders from 2020.

1. *Conclusion and Discussion*

This analysis delves into the European automotive market from 2018 to 2020, examining vehicle sales figures and order statuses. By employing a range of visual tools like scatter plots, heatmaps, and trend charts the study aimed to uncover significant trends and patterns in the data. The focus was on gaining a comprehensive understanding of the market dynamics across various European countries.

Based on the analysis, we found that sales and price have a positive linear relationship, meaning that as the price increases, sales also increase. We also found that the top 10 GDP countries have similar life expectancy, with the USA having an average of 78.5 years, which is the lowest among the top 10 GDP countries. Japan has the highest average life expectancy at 84.2 years. The result of the yearly trend in order numbers in European countries shows that France and Spain have the most orders compared to other European countries. We also saw that when average tax revenue increases, the total tax rate generally increases, but Denmark does not fit this pattern. Lastly, we observed that in European countries, the most common order status by year is shipped.

* 1. *Discussion*

While the results give a general overview of the order statuses, further research could delve deeper into the causes of the cancelled, resolved, and disputed statuses to better understand the factors behind these orders. Understanding why orders are marked as on hold or in process could also provide insights into potential improvements in order management and customer service.

Overall, this analysis highlights the importance of monitoring and understanding order statuses, especially in the context of unexpected global events like the pandemic, which might have affected order fulfillment and customer satisfaction across Europe.

*For Dashboard Please Follow Link*

https://public.tableau.com/app/profile/mehmet.tatbak/viz/ExploringVehicleOrdersAcrossEurope/ExploringVehicleOrdersAcrossEurope?publish=yes

*For Github Please Follow Link*

https://github.com/mtatbak

*REFERENCES*

Verma, A. (2020, May 20).Medium. <https://medium.com/@ajayverma23/data-imputation-a-comprehensive-guide-to-handling-missing-values-b5c7d11c3488>

LinkedIn. (n.d.). How can data visualization identify outliers? LinkedIn. Retrieved December 7, 2024,from https://www.linkedin.com/advice/1/how-can-data-visualization-identify-outliers?utm\_source=share&utm\_medium=member\_android&utm\_campaign=share\_via