STAT-112 PROJECT

Automobile Sales Analysis

Elif Sude Kürkçü  
 *Department of Statistics*  
*Middle East Technical University*Ankara/Türkiye  
elifsudekurkcu@gmail.com

*Abstract*— This document is an analysis of auto sales of an automobile company. There are two datasets containing information about auto sales of a company and countries' attributes. Visualization techniques employed on merged data to get insights into the tendencies of the company’s sales across the globe.

Keywords— auto sales, automobile, worldwide, analysis, automobile

# Introduction

This analysis focuses on six questions:

* Which deal size is more profitable for the company?
* Which vehicle category has generated the most revenue for the company?
* What are the proportions of revenues by countries?
* Is there a correlation between a country's GDP and the sales made to that country?
* What are the respective proportions of vehicle types selected by countries?
* What are the countries’ sales proportions for each vehicle category?

20 variables and 2,747 rows are present in the automobile sales dataset and 14 variables and 195 rows, meaning 195 distinct countries are present in the world data dataset..

# Data Preprocessing

There are no missing or duplicated values in the automobile sales dataset however, some missing values present in the world data dataset.

## Automobile sales dataset variables:

### Qualitative:

Ordernumber: Unique code of each order.

Productcode: Unique code of each product.

Customername: Name of customer.

Phone: Phone number of each customer.

Addressline: Adress of each customer.

City: City where customers are located.

Postalcode: Postal code of the city.

Country: Country where customer located.

Contactlastname: Last name of the contact person associated with the customer .

Contactfirstname: First name of the contact person associated with the customer.

Dealsize: Size of the order.

Status: Current situation of the order. .

Productline: Category of the product.

Orderlinenumber: Line number of each item within an order.

Orderdate: Date of order placed.

### Quantitative:

Quantityordered: Number of items ordered in each order. (discrete, ratio).

Priceeach: Price of each product (continuous, ratio).

Sales: Total revenue of each order (continuous ratio).

Days\_since\_lastorder: Number of days that have passed since the last order for each customer (discrete, ratio).

Msrp: Suggested selling price for each item. (continuous, ratio).

## World data dataset variables:

### Qualitative:

Country: Name of the country.

Latitude: Latitude coordinate.

Longitude: Longitude coordinate.

### Quantitative:

Birth Rate: Number of births per 1,000 population per year (continuous, ratio).

CO2-Emission: Co2 emission in tons (continuous, ratio).

CPI: Consumer price index, measure of purchasing power. (continuous, interval).

Gasoline Price: Price of gasoline per liter in local currency (continuous, ratio).

GDP: Gross Domestic Product, the total value of goods and services produced in the country (continuous, ratio).

Life expectancy: Average years of a newborn expected to live (continuous, ratio).

Population: Population of the country (continuous, ratio).

Tax Revenue (%): Tax revenue as a percentage of GDP (continuous ratio).

Total tax rate: overall tax burden as a percentage of business profits. (continuous, ratio).

Unemployment rate: Percentage of labor force unemployed (continuous, ratio).

Urban\_population: Percentage of population living in urban areas. (continuous, ratio).

Since the analysis only focuses on the countries involve in the auto sales dataset, datasets merged based on the “country” variable. The merged dataset has no missing or duplicated values. Therefore, data cleaning is not required. The focus of interest of research questions includes the following variables: *Deal Size, Sales, Product Line, GDP, Price each, Quantity ordered,* and *status*. Therefore, exploratory data analysis will be utilized on these variables.

# Exploratory Data Analysıs

Histogram and box plots are effective techniques for exploratory data analysis and detecting outliers.

## Sales Variable

Sales histogram and box plot distributions illustrated in Fig. 1 and Fig. 2.

TABLE I. SALES HISTOGRAM

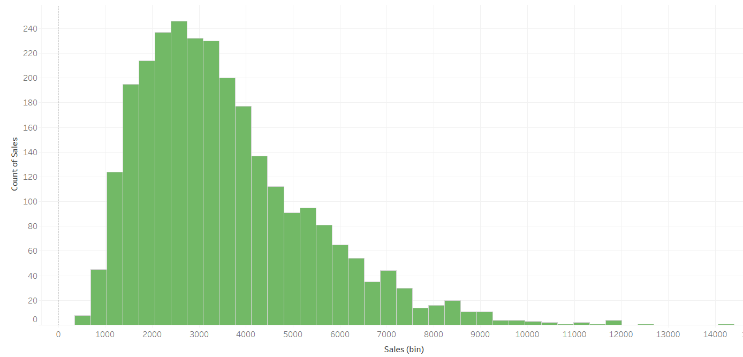


Fig.1 .

TABLE II. SALES BOX PLOT

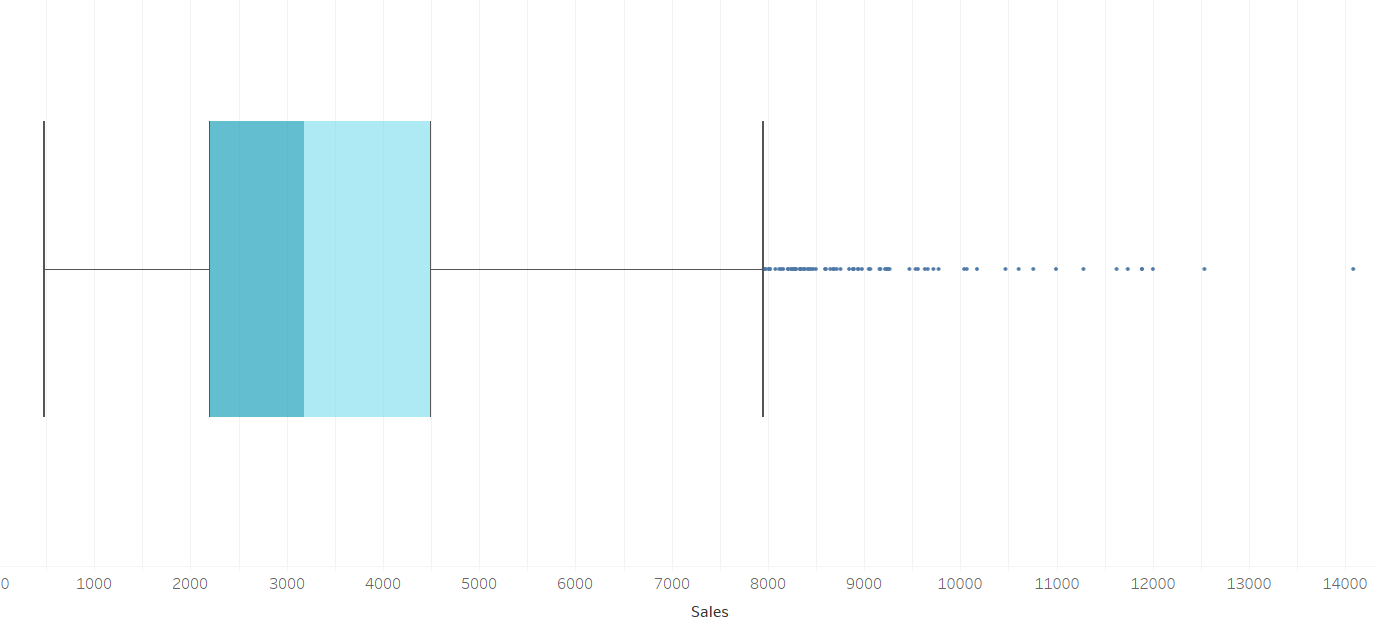


Fig.2

As shown in Fig. 1 and Fig.2, Sales distribution is positively skewed and unimodal with the median value of 3185 and mean value of 3553. There are several outliers based on the 1.5 IQR rule. Since the research questions are country-based, it is aimed to preserve the characteristics of each country by not removing outliers from the overall distribution plot. Therefore, country-based distributions are utilized to eliminate outliers, as illustrated in Fig. 3.

TABLE III. SALES BOX PLOT OF EACH COUNTRY

metin, ekran görüntüsü, diyagram, çizgi içeren bir resim

Açıklama otomatik olarak oluşturuldu

Fig 3. Sales distribution box plot of each country.

As seen in Fig. 3, there are outliers for each country. Since all the outliers are placed on the right side of the plot, they represent extreme sales. To preserve their meanings, outliers adjusted to be closer to Q3 by replacing them with the upper whisker. In this manner, upper outliers preserved their significance while reducing their impact on the analysis

## Price Each Variable

Category-based distributions are utilized to preserve each category's characteristics while eliminating outliers, as illustrated in Fig. 4.

TABLE IV. “PRICE EACH” BOX PLOT OF EACH CATEGORY

metin, diyagram, plan, teknik çizim içeren bir resim

Açıklama otomatik olarak oluşturuldu

Fig.4. Price each distribution box plot of each category

There are outliers on the “price each” variable for each vehicle category. Since the data size is small and all the outliers are on the higher side of the plot, outliers are eliminated by replacing with the upper whisker of the plot in order to preserve their meanings.

## Quantity Ordered Variable

Quantity ordered variable illustrated in Fig 5.

TABLE V. “QUANTITY ORDERED” VARIABLE BOX PLOT

metin, ekran görüntüsü, çizgi, yazı tipi içeren bir resim

Açıklama otomatik olarak oluşturuldu

Fig. 5. Quantity ordered variable box plot.

Distribution is nearly symmetric, with a median value of 35 and the mean value of 35.1. There are a few outliers; therefore, they are replaced with the median value.

## GDP Variable

GDP data’s scale of measure could be recognized as categorical due to the currency sign before the number. It has been changed by deleting currency signs and converting the data type to numerical. GDP distribution shown with box plot in Fig.6.

TABLE VI. GDP BOX PLOT

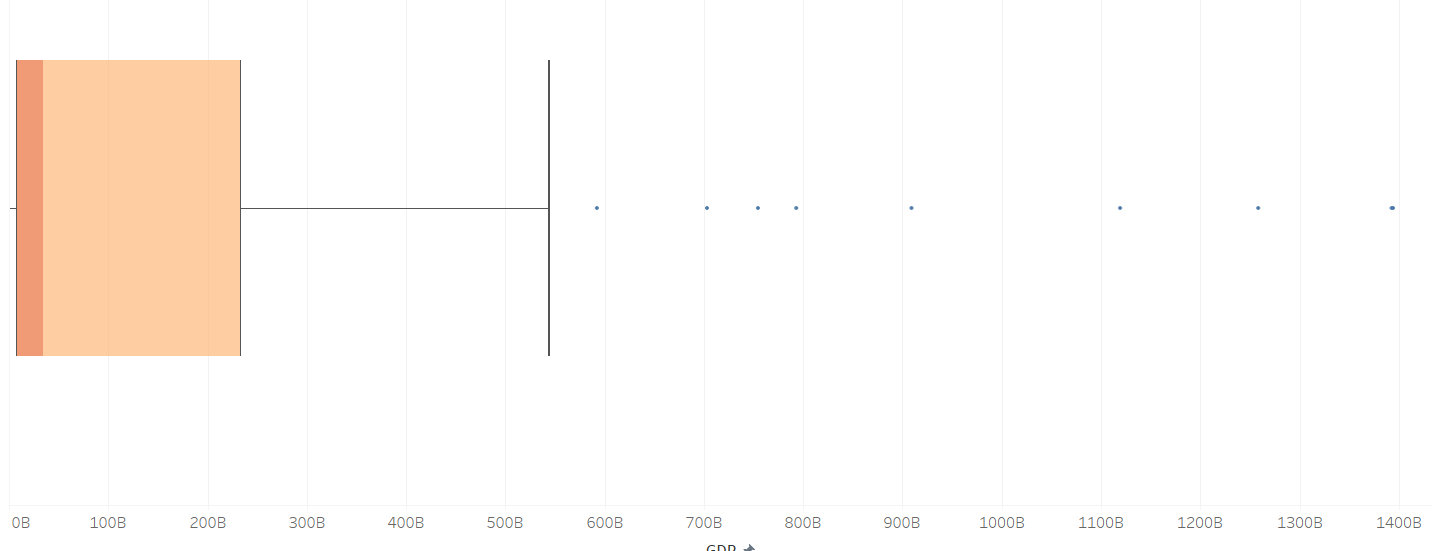


Fig.6. GDP box plot.

There are several outliers, especially the USA’s GDP, which is 21.43 trillion. The distribution is positively skewed. Most of the data is concentrated on the lower part of the distribution. This analysis solely focuses on the countries where the company has made sales. The histogram of GDP filtered to show only these countries shown in Fig.7.

TABLE VII. GDP HISTOGRAM

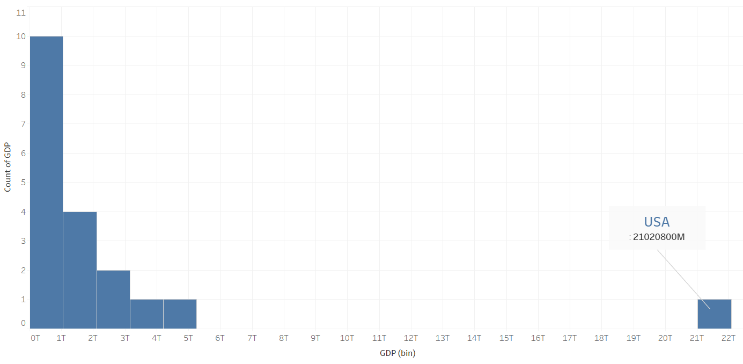


Fig.7. GDP histogram.

Therefore, when examining the distribution of the GDP of these countries, as seen in Fig.7, the USA appeared as the sole outlier. Since the GDP of the USA is a real and legitimate value, no adjustments have been made.

# Research Questions

Research questions will be answered and interpreted by several multivariate visualization techniques.

## Which deal size is more profitable for company?

A simple table shown in Fig. 8 is adequate for this question. Data is filtered by setting *status* to “Shipped” to obtain more precise information.

TABLE VIII. SALE VALUES OF EACH DEAL SIZE

metin, ekran görüntüsü, yazı tipi, makbuz içeren bir resim

Açıklama otomatik olarak oluşturuldu

Fig.8.

Although the average sales of large-sized sales are the highest value, it is the one providing the least profit to the company due to its count. Medium-sized sales are the most profitable, with small-sized sales coming in second.

## Which vehicle category has generated the most revenue for the company?

Pie chart shown in fig. 9 utilized for this question (Data is filtered to *status = Shipped*).

TABLE IX. TOTAL SALE PROPORTIONS OF CATEGORIES

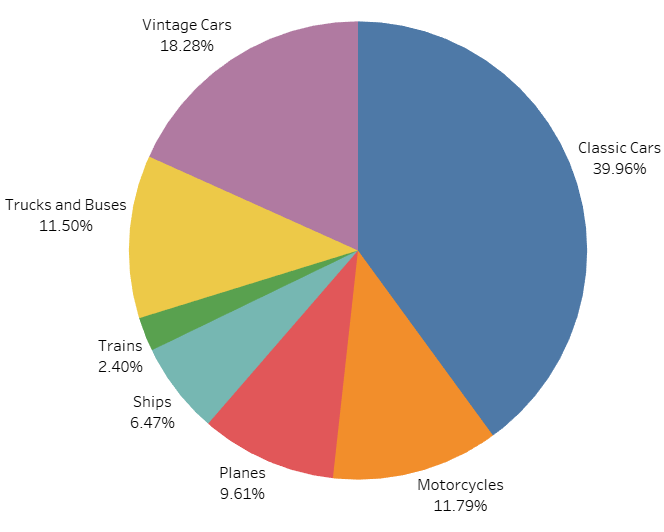


Fig.9 Sale proportions of categories

As seen in the fig. 9, most revenue is earned by selling *Classical Cars,* after that *Vintage Cars* comes. Minimum revenue is earned by selling *Trains*. To get a deeper insight into the data, we calculate the expected revenue for each item by multiplying the variables: average of *price each*, average of *quantity ordered*, and *number of sales*. After that, we compare the actual revenue (*Sum(Sales)*) and expected revenue, shown in fig 10.

TABLE X. DIFFERENCE OF EXPECTED AND REAL REVENUES ACROSS CATEGORIES

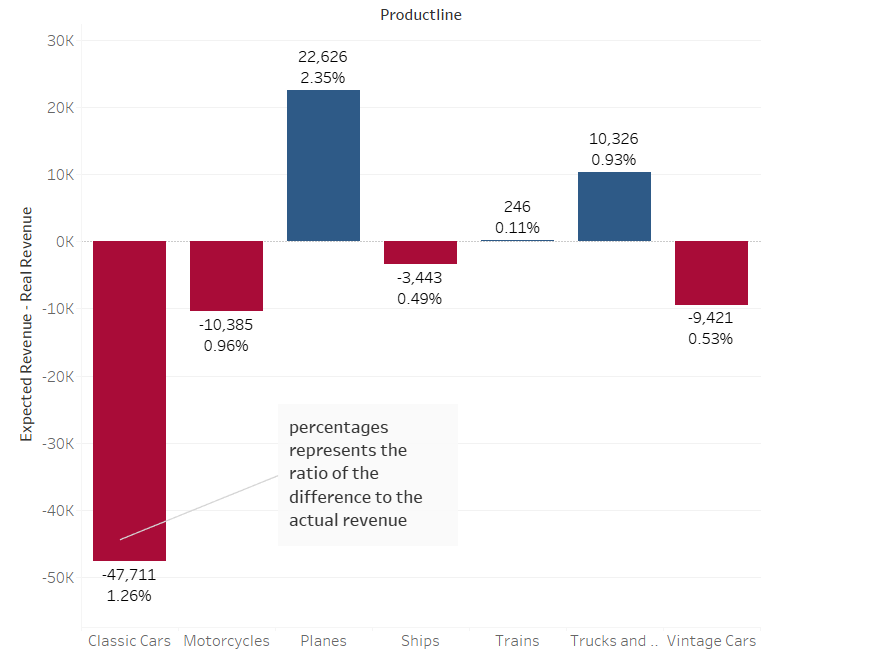


Fig.10

As seen in fig. 10, the biggest difference across the categories is *Classical Cars*. A 1.26% loss percentage is not negligible for an automobile company. Therefore, the company should increase the prices of *Classical Cars*. The others can be considered negligible. Moreover, the actual revenue of the *Planes* is quite higher than the expected.

## What are the proportions of revenues of each country?

A treemap shown in fig.11 is utilized to show proportions, and the data is filtered to *status = Shipped.*

TABLE XI. SALE PROPORTIONS OF COUNTRIES

metin, ekran görüntüsü, diyagram, sayı, numara içeren bir resim

Açıklama otomatik olarak oluşturuldu

Fig.11

As seen in fig. 11, the USA appears to be by far the largest market for the company, followed by Spain and France. Proportions for each vehicle category could be seen by using the dashboard linked at the end of the document

## What are the respective proportions of vehicle categories selected by countries to examine each country’s needs?

A stacked bar chart shown in fig.12 is utilized for this question. However, using the dashboard provides a large-scale treemap for each category, which offers a more comprehensive and detailed visualization.

TABLE XII. SALES OF CATEGORIES ACROSS COUNTRIES

metin, ekran görüntüsü, renklilik, çizgi içeren bir resim

Açıklama otomatik olarak oluşturuldu

Fig.12.

In many countries, *Classic Cars* have the highest sales rate. Especially in the USA, C*lassic Cars* had the highest rate with 34.23%. Moreover, *Classic Cars* attracts attention with 30.90% in France. However, *Vintage Cars* sales are ahead with 39.86% in Spain. Therefore, it would make sense for the company to focus on *Vintage Cars* in sales for Spain and *Classic Cars* for France*.* Another striking factor is that there are only *Classic Cars* in Switzerland sales. However, this should not lead to the misconception that Switzerland only chooses *Classic Cars*. As it seen in the bar chart, the company's Switzerland market is quite low, which is a confounding factor. As illustrated in the previous treemap, the majority of the sales market consists of the USA, Spain, France, and Australia. It would be logical for the company to consider the rates of these countries in production stages.

## What is the sale proportion of each country for each vehicle category?

Proportions of countries for each vehicle category could be interpreted by filtering the dashboard linked at the end of the document. A stacked bar chart again appears as a logical visualization technique for the research question. As seen in fig.13, it is evident that the sales for the USA dominate across all vehicle categories. It results from both the company concentrated on the USA most and the USA is the biggest market.

TABLE XIII. SALES FOR COUNTRIES ACROSS CATEGORIES

metin, ekran görüntüsü, renklilik, tasarım içeren bir resim

Açıklama otomatik olarak oluşturuldu

Fig.13.

For a more comprehensive and detailed approach, USA excluded from the table and shown in fig.14.This approach made the table easier to interpret.

TABLE XIV. SALES FOR COUNTRIES ACROSS CATEGORIES WITHOUT THE USA

metin, ekran görüntüsü, renklilik, tasarım içeren bir resim

Açıklama otomatik olarak oluşturuldu

Fig.14.

Visualization in fig.14 makes a lot of sense for the company since it is a clear indication of which country’s market the sales of a ready-made product can be concentrated on after the USA. Motorcycles and Planes are mostly preferred by France, and the sales for Spain dominate the other fields. Therefore, focusing on *Vintage Cars* in Spain or *Planes* in France could maximize profitability. Another remarkable thing is that Australia has an extremely low rate in ship sales, although it is an ocean country. It may indicate a potential gap in the market. This highlights how regional preferences vary significantly across.

## Is there a correlation between a country's GDP and the sales made to that country?

A scatter plot is a favorable way to show the relationship between two different variables. According to the trend line of the scatter plot in fig 15, average GDP and sales have a positive correlation.

TABLE XV. SALES vs AVG. GDP

metin, ekran görüntüsü, diyagram, çizgi içeren bir resim

Açıklama otomatik olarak oluşturuldu

Fig.15.

Since the USA is an outlier of GDP data, interpreting after excluding the USA would give a clearer insight. According to the USA excluded plot shown in fig.16, There is still a positive correlation; however, the slope is less steep.

TABLE XVI. SALES VS AVG. GDP WITHOUT THE USA

metin, ekran görüntüsü, diyagram, çizgi içeren bir resim

Açıklama otomatik olarak oluşturuldu

Fig.16

Since auto sales data size is small, countries where the company's market is small may be confusing. To eliminate this problem, only the top 10 countries where the company achieved the highest sales are displayed. It is displayed with and without the USA in fig.17 and fig.18.

TABLE XVII. SALES VS AVG. GDP FOR TOP 10 COUNTRIES

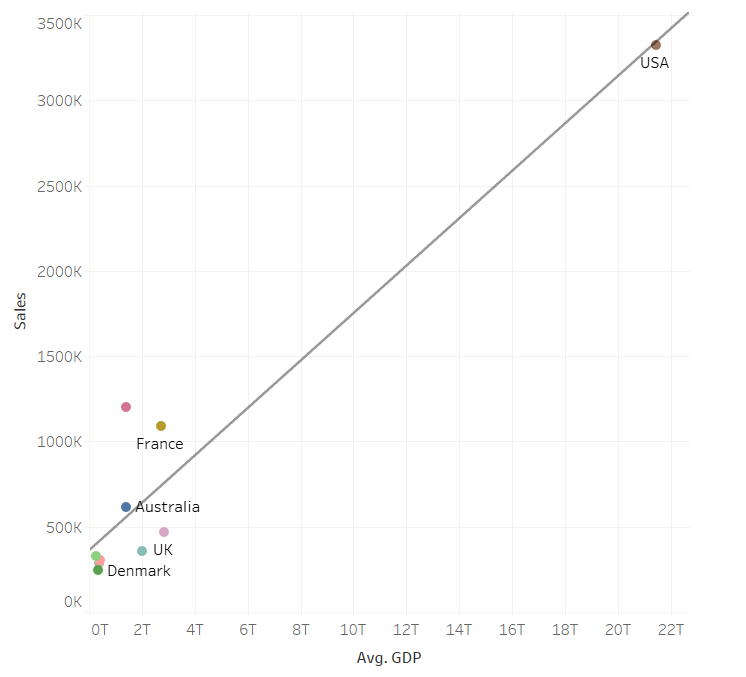


Fig.17

TABLE XVIII. SALES VS AVG. GDP FOR TOP 10 COUNTRIES WITHOUT THE USA

metin, ekran görüntüsü, diyagram, çizgi içeren bir resim

Açıklama otomatik olarak oluşturuldu

Fig.18.

According to fig.17 and fig.18 it is certain that there is a positive correlation between the average GDP and sales. Since GDP reflects a country's economic size, this correlation is unsurprising. If the company intends to expand into new markets, it must consider the GDP of the country to maximize potential revenue. Fig.19 is a list of 10 countries where the company could establish new markets.

TABLE IXX. TOP 10 COUNTRIES FOR NEW MARKETS

metin, ekran görüntüsü, yazı tipi, sayı, numara içeren bir resim

Açıklama otomatik olarak oluşturuldu

Fig.19.

# CONCLUSION

In conclusion, throughout the analysis, it is aimed to provide the company to identify opportunities and challenges in specific markets, making it easier to allocate resources effectively. Giving comprehensive analysis of data both at the product and country level allowed the company to gain insight into optimizing existing sales and suggest strategic improvements where necessary. This analysis provided valuable options by analyzing the correlation of economic size and sales to entering new markets that could contribute to the company's growth. Overall, this analysis provided a comprehensive framework for improving sales strategies.

*DASHBOARD~* [*https://public.tableau.com/views/Stat-1121stPROJECT/AutoSalesAnalysis?:language=en-US&:sid=&:redirect=auth&:display\_count=n&:origin=viz\_share\_link*](https://public.tableau.com/views/Stat-1121stPROJECT/AutoSalesAnalysis?:language=en-US&:sid=&:redirect=auth&:display_count=n&:origin=viz_share_link)

*GITHUB~* [*https://github.com/elifsude/STAT-112.git*](https://github.com/elifsude/STAT-112.git)