STAT 112 – Introduction to Data Processing and Visualization

Tableau Project Report

Yiğit Efe Özdemir

2740835

December 7, 2024

**1. Introduction**

In this report, i analyzed two datasets which are about automobile sales and 2023 world data. In total, datasets contain 34 variables and 2942 observations ( 20 variables and 2747 observations for automobile sales data, 14 variables and 195 observations for 2023 world data.) In the raw data 14 of the variables are categorical and 20 of the variables are numerical. In this report, firstly i will examine the datasets and variables. Then, in Data Preprocessing part i will do some cleaning with the data and prepare the data for Exploraritary Data Analysis part. After that, in the Exploraritary Data Analysis part, firstly i will interpret some of the variables in the datasets. Then i will prepare 5 research questions about the variables. Next, i will do visualizations for these research questions and interpret these visualizations. Finally in the Conclusion and Discussion part, i will summarize all the report and i will explain some important findings from research questions.

**2. Data Preprocessing**

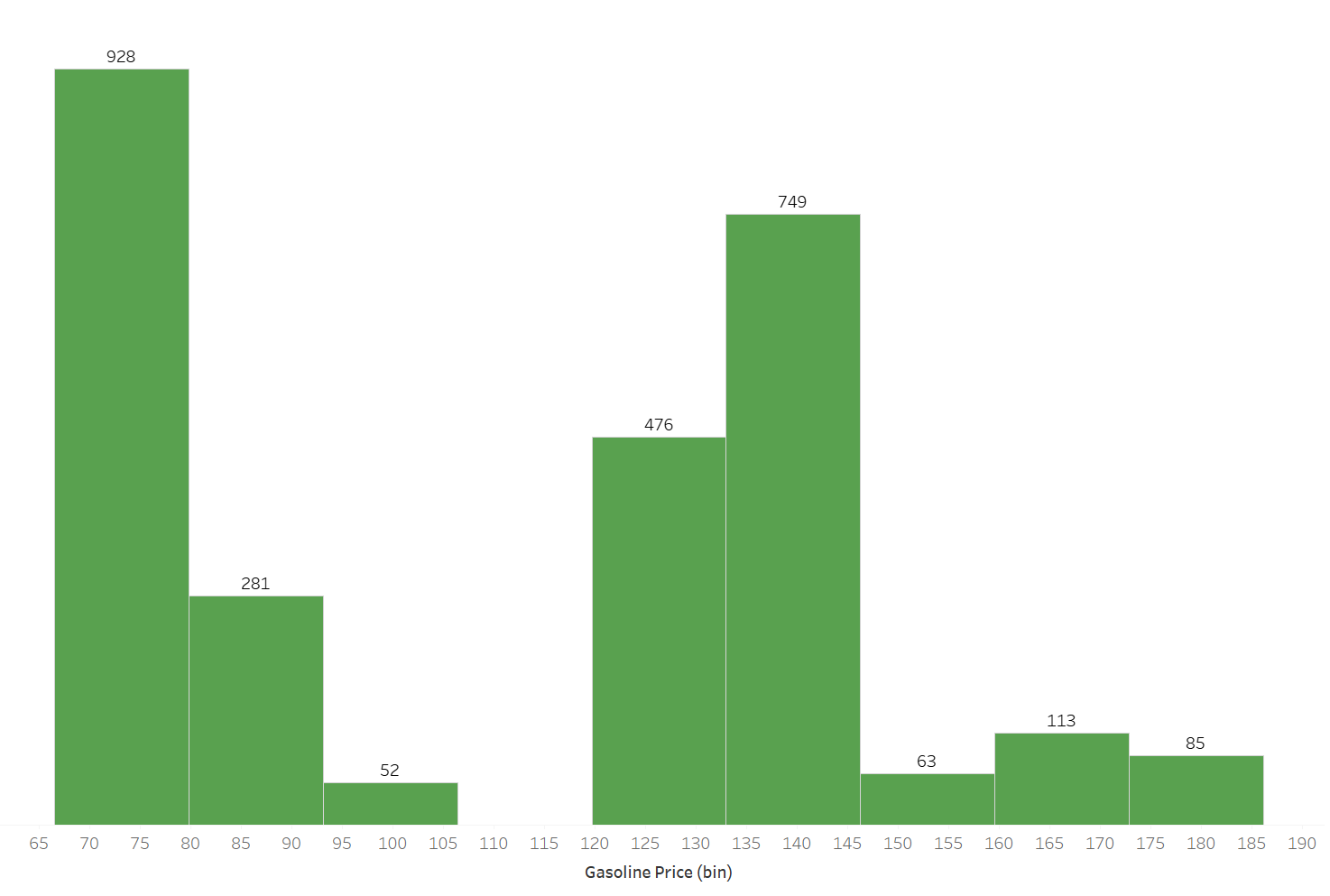
Tidying and cleaning data is one of the most important steps. Firstly, in order to get rid of null data and missing data, i used inner join method for automobile sales and world data datasets and match “country” variables in the datasets. After that, there were no null or missing data but some of the variables’ data types were incorrect. “Gasoline Price” and “GDP” variables were shown categorical because of the dollar sign in front of the numbers. In order to correct them i used calculated fields and remove dollar sign from the variables. Finally, i changed these variables’ data type to numerical and made the data ready for the Exploraritary Data Analysis ( EDA ) part.

**3. Exploraritary Data Analysis ( EDA )**

Exploraritary Data Analysis is the most important part of the report. There will be two parts : Interpretation of variables, Research Questions and Visualizations.

**3.1 Interpretation of Variables**

To be more familiar with the datasets and check if there is outliers, firstly i interpret some of the variables that i plan to use. I used Box Plot and Histogram visualizations for the interpretation.



Histogram of Gasoline Price

As we can see that, Gasoline Price variable has bimodal distribution. Median value for gasoline price is 125 and mean value for gasoline price is 111.1. Q1 of gasoline price variable is 71, Q3 of gasoline price variable is 139. IQR is 68. There are no possible outliers because there aren’t any values 1.5 IQR above Q3 or 1.5 IQR below Q1.

A graph with numbers and letters

Description automatically generated

Box Plot of Co2-Emissions

Co2-Emissions variable has right skewed distribution. Median value for Co2-Emissions is 375K and mean value for Co2-Emissions is around 1800K. Q1 of Co2-Emissions is 244K, Q3 is 5000K, IQR is 4756K. Even if it seems there are outliers at around 5000K, the values are not outliers because they are not 1.5IQR above from Q3.

A screenshot of a graph

Description automatically generated

Box Plot of Life Expectancy

Life Expectancy variable has slightly right skewed distribution. Median value for Life Expectancy is 81.70 and mean value for Life expectancy is 81.04. Although it seems not slightly right skewed, when we look at median and mean values it is really close to symmetric distribution. Also there are possible outliers at interval (70-75)

A graph of a number of people

Description automatically generated

Histogram of Sales

Sales variable has right skewed distribution. Median value for sales variable is 3185, mean value vor sales variable is 3553. Q1 of sales variable is 2204, Q3 of sales variable is 4503 and IQR is 2299. As we can see, there are values that 1.5 IQR above from Q3 which is 6802. Values that above 6802 are possible outliers.

A graph with numbers and a bar

Description automatically generated

Box Plot of Birth Rate

Birth Rate variable has left skewed distribution. Median value for Birth Rate is 11.30 and mean value for Birth Rate is 10.57. Also as we can see in that box plot there are possible outliers at interval (20-22).

A graph with numbers and a bar graph

Description automatically generated

Histogram of CPI

CPI variable has slightly left skewed distribution. Median value for CPI is 117.2 and mean value for CPI is 114.7. Q1 of CPI is 111, Q3 of CPI is 117,2 and IQR is 6,2. There are some different values at interval (96-100) and (128-132). And as the values at interval (128-132) are 1.5 IQR above from Q3 which is 126,5 and values at interval (96-100) are 1.5 IQR below from Q1 which is 101,7, they are possible outliers.

**3.2 Research Questions and Visualizations**

For these report, i prepared 5 research questions related to the datasets and 7 different visualizations to interpret and analyze the research questions.( 5 of them will be used in the dashboard.)My research questions are :

1) How does vehicle sales change over years?

2) What is the relationship between Life Expectancy and Birth Rate?

3) How does average GDP affect vehicle sales among countries?

4) Which vehicle types have most sales by regions?  
5) How does Average CO2 Emissions and Gasoline Price vary across Europe?

**3.2.1 How does vehicle sales change over years?**

A graph with blue lines and dots

Description automatically generated

Visualization 1 : Line Chart

I used line chart to show how does vehicle sales change over years because line chart is proper visualization for showing the change for a variable and this line chart adds interactivity to the dashboard with “ use as filter “ option for specific months of the years. For this line chart, there are 2 peak points at November 2018 and November 2019. About this peak points, we can say that people usually prefer buying their vehicles in November. On the other hand, although generally there is an increase in vehicles sales over years, at some times vehicle sales decreased for instance in December 2018, March 2019, September 2019, December 2019 and April 2020. Reason of the increase in December 2018 and December 2019 can be sudden increase at November 2018 and November 2019.

**3.2.2 What is the relationship between Life Expectancy and Birth Rate?**

A graph with dots and lines

Description automatically generated

Visualizaton 2 : Scatter Plot

In order to examine the relationship between Life Expectancy and Birth Rate, I used scatterplot because scatter plot is the best visualization type for interpreting the relationship between two numerical variables. Correlation coefficient for this scatter plot is approximately -0.68. It means there is a strong negative relationship between Life Expectancy and Birth Rate. As we can also see from the trend line, generally when Life Expectancy increases, Birth Rate decreases. Possible reason for this situation is when Life Expectancy increases death rate of children decreases and it causes lower birth rates.

**3.2.3 How does average GDP affect vehicle sales among countries?**

A blue rectangle with black text

Description automatically generated

Visualization 3 : Tree Map

In order to explain this research question i used tree map because tree map is efficient when visualizing two numerical and one categorical variable. As i have GDP ( Gross Domestic Product ), Vehicle Sales and Country variables, i preferred tree map. I filter the countries with top 10 because too much country can cause really crowded visualization. Also this visualization adds interactivity to the dashboard with use as filter option for specific countries. In this tree map size of the boxes represent average GDP and color represents total sales. ( darker blue means more sales and lighter blue means less sales ) As we can see, USA have both darkest color and biggest size meaning that the country that has most average GDP also has most vehicle sales. On the other hand, except for France and Spain, generally bigger sized countries also have darker colors, it means average GDP and total amount of sales have positive relationship.

A graph with a line and text

Description automatically generated with medium confidence

Visualization 4 : Bubble Plot

I also used bubble plot for this research question because with tree map we can understand which countries have most average GDP and total vehicle sales but in order to examine actual relatonship between average GDP and total vehicle sales, bubble plot is a proper technique. In this bubble plot, as we can see in the filter, colors represent countries and bubble size represents total sales. Correlation coefficient for this bubble plot is approximately 0.94. It means there is really strong positive relationship between average GDP and Vehicle Sales. Possible reasons for this situation can be when GDP increases economic activity increases and higher economic activity causes more vehicle sales.

**3.2.4 Which vehicle types have most Sales by Region?**

**A colorful bar chart with text

Description automatically generated with medium confidence**

Visualization 5 : Clustered Bar Chart

For this research question, i use Clustered Bar Chart because in order to group total sales by both region and vehicle type, Clustered Bar Chart is the most suitable visualization technique. Also with this Clustered Bar chart i added interactivity to the dashboard with use as filter option for both regions and vehicle types. As we can see in this Clustered Bar Chart, classic cars are the most sold vehicle type among vehicle types and trains are the least sold vehicle type. For all vehicle types europe is on top according to total sales. In some vehicle types for example motorcycles, planes and vintage cars; Asia is the least seller region but in classic cars, ships, trains, trucks and busses; Australia is the least seller region. In every vehicle type, North America is second best seller among other regions. The difference between vehicle types and sales shows that, requests and needs of regions are different.

**3.2.5 How does Average CO2 Emissions and Gasoline Price vary across Europe?**

A graph of graphing with red and blue bars

Description automatically generated with medium confidence

Visualization 6 : Dual Bar Chart

I used Dual Bar Chart for this research question because examining two different variable for one country is possible with Dual Bar Chart. In this dual bar chart, red bars represent average CO2 emissions and blue bars represent Gasoline Price for countries. Germany has the highest average CO2 emission among countries but Gasoline Price of Germany is low. On the other hand although France has the highest Gasoline Price, average CO2 emission of France is not as high as its Gasoline Price. With this observations, we can conclude that average CO2 emission and gasoline price have negative relationship but as Germany doesn’t have the least Gasoline Price and France doesn’t have the least CO2 Emissions, we can not say that there is a strong negative relationship.

A graph with a line and dots

Description automatically generated

Visualization 7 : Scatter Plot

I also used scatter plot for this research question because even if Dual Bar Chart shows the relationship between average CO2 emissions and Gasoline Price, scatter plot is the most suitable visualization for showing exact relationship. Correlation coefficient for this scatter plot is – 0.17. It means there is weak negative relationship between average CO2 emissions and Gasoline Price. There are some possible reasons of this weak negative relationship. For example when gasoline prices increase, people use less vehicles and less vehicle usage decreases the average CO2 emission. On the other hand, there is too much reasons for CO2 emissions, so that even if gasoline prices increase CO2 emissions might not increase as much as gasoline prices.

**4. Discussion and Conclusion**

In the discussion part i will conclude the whole EDA Part and give some important findings we get from visualizations. In the conclusion part i will summarize whole report and finish.

**4.1 Discussion**

According to the research questions and visualizations of the data, i got important findings :

- Most people prefer buying their vehicles in November ( from 1st research question )

- There is negative relationship between Life Expectancy and Birth Rate and decrease of children death rate when life expectancy increases can be a reason of this negative relationship. ( from 2nd research question )

- There is positive relationship between GDP and vehicle sales and increase of economic activity when GDP increases can be a possible reason for this positive relationship. ( from 3rd research question

- Best seller vehicle type is classic cars and best seller region is Europe. ( from 4th research question )

- Germany has the most CO2 emission and France has the most Gasoline Price among countries in Europe. ( from 5th research question )

- There is a weak negative relationship between average CO2 emission and Gasoline Price. Increase of gasoline price decreases vehicle usage. It can be the possible reason of this weak negative relationship. ( from 5th research question )

**4.2 Conclusion**

In conclusion, first i analyzed the raw data and cleaned the data.Then I examined distributions of some variables, visualize the data according to the research questions. Finally i interpret the visualizations and get some important findings from these visualizations as explained in the discussion part.

Link to the Dashboard : <https://public.tableau.com/app/profile/yi.it.efe.zdemir/viz/STAT112PROJECT2740835YTEFEZDEMR/Dashboard?publish=yes>

Link to the GitHub Repository:

<https://github.com/YigitEfeOzdemir/STAT112-Tableau-Project-2740835-Yi-it-Efe-zdemir>