Programming for Data Science

Section C

Group 4 Project

Armen Madoyan, Hmayak Paravyan, Vahagn Hakobyan, , Zorik Ghazaryan,

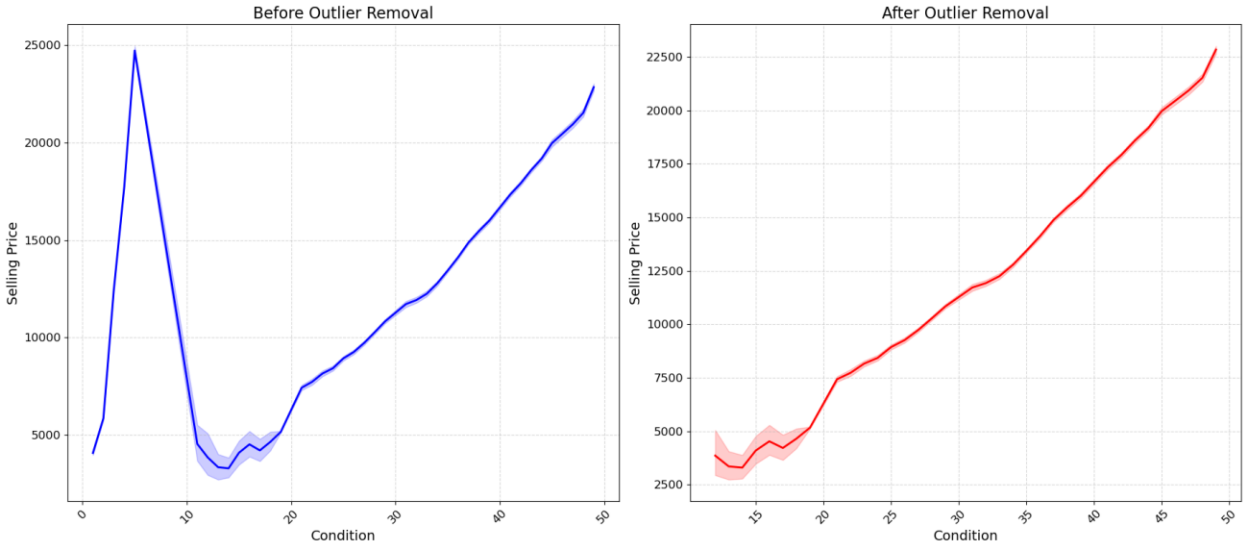
REPORT

The dataset under analysis comprises 558,837 rows and 16 columns. The "Vehicle Sales and Market Trends Dataset" is a comprehensive compilation of data concerning the sales activities of diverse vehicles. It includes extensive information such as the year, make, model, trim, body type, transmission type, VIN (Vehicle Identification Number), state of registration, condition rating, odometer reading, exterior and interior colors, seller information, Manheim Market Report (MMR) values, selling prices, and sale dates The goal of the project is to analyze the trends and patterns within the vehicle sales market to gain insights that can inform business strategies, marketing decisions, and pricing strategies for both manufacturers and sellers.

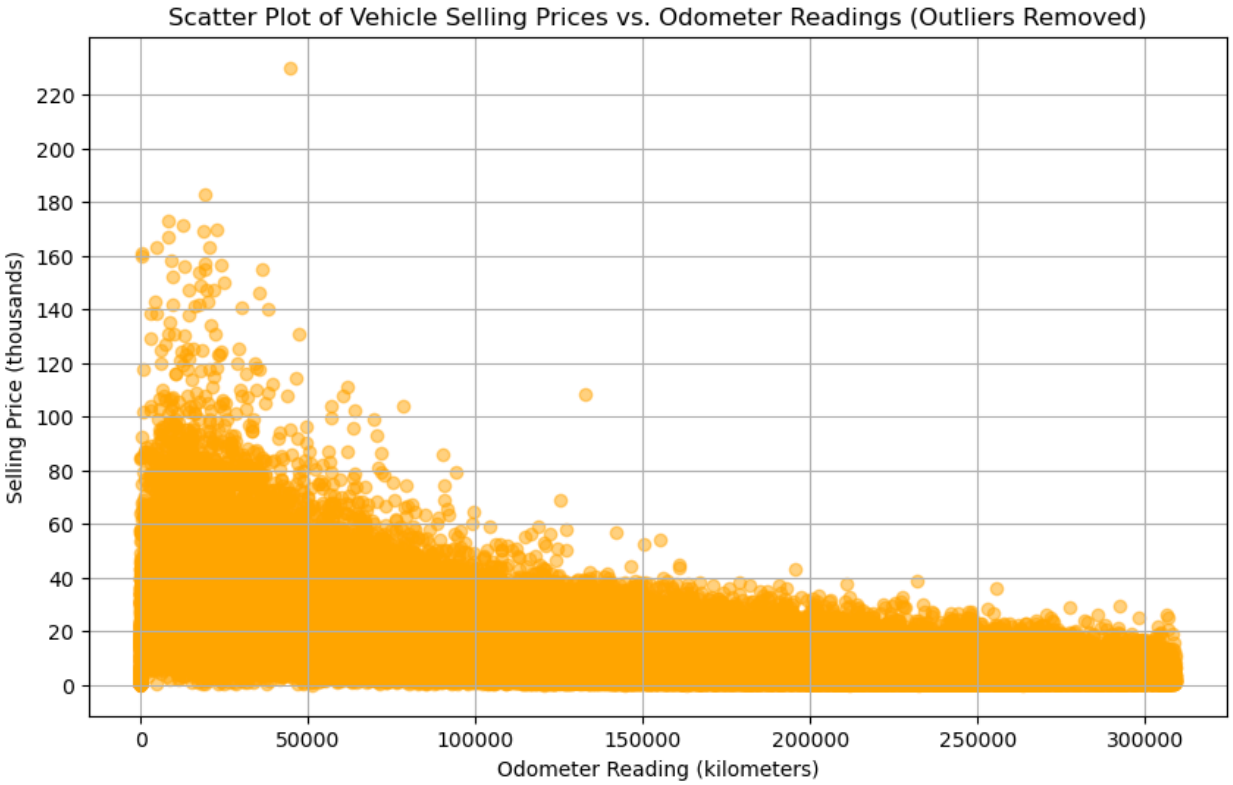
THE PROCESS

We started with a data cleaning process with our dataset. Initially comprising 558,837 rows and 16 columns, the dataset has been streamlined to 465,973 rows and 9 columns. 92,864 rows were removed during the cleaning process, primarily due to missing values and outliers; also 7 columns were eliminated as they were deemed irrelevant to the analysis or contained redundant information. There were also some repetitions related to car names, which was handled with mapping.

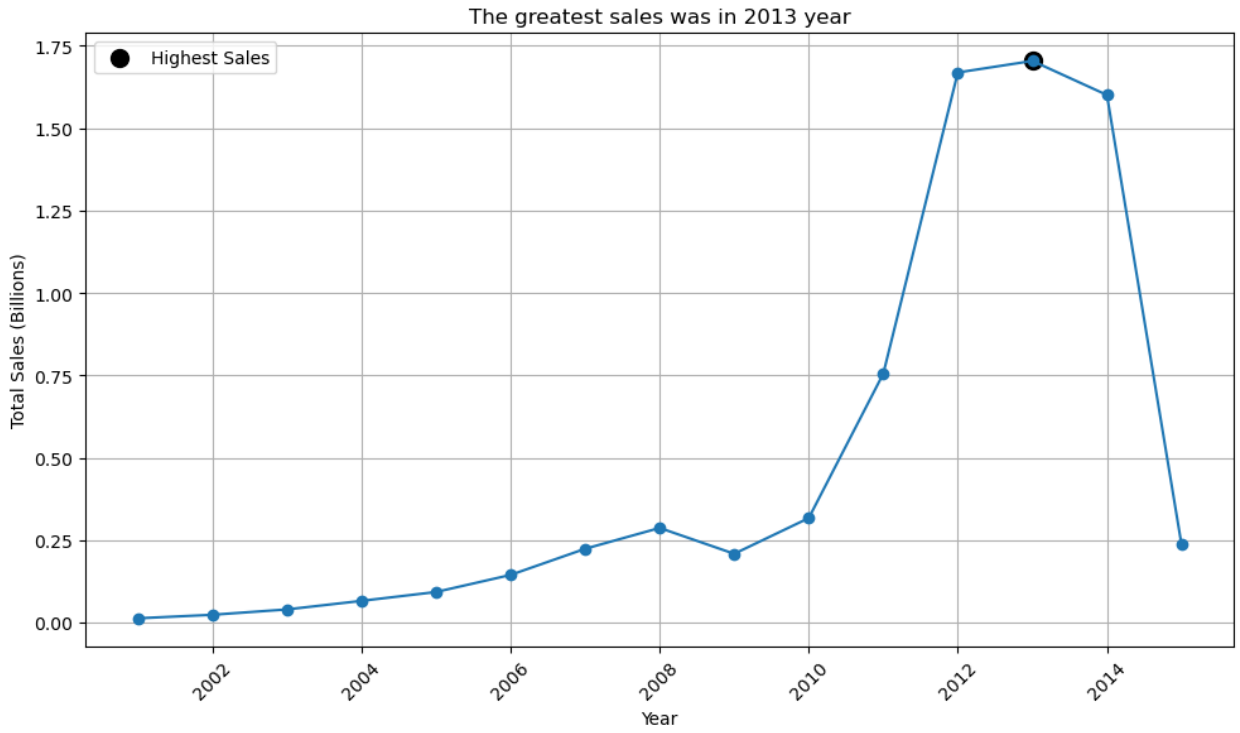
Descriptions of Graphs and Conclusions



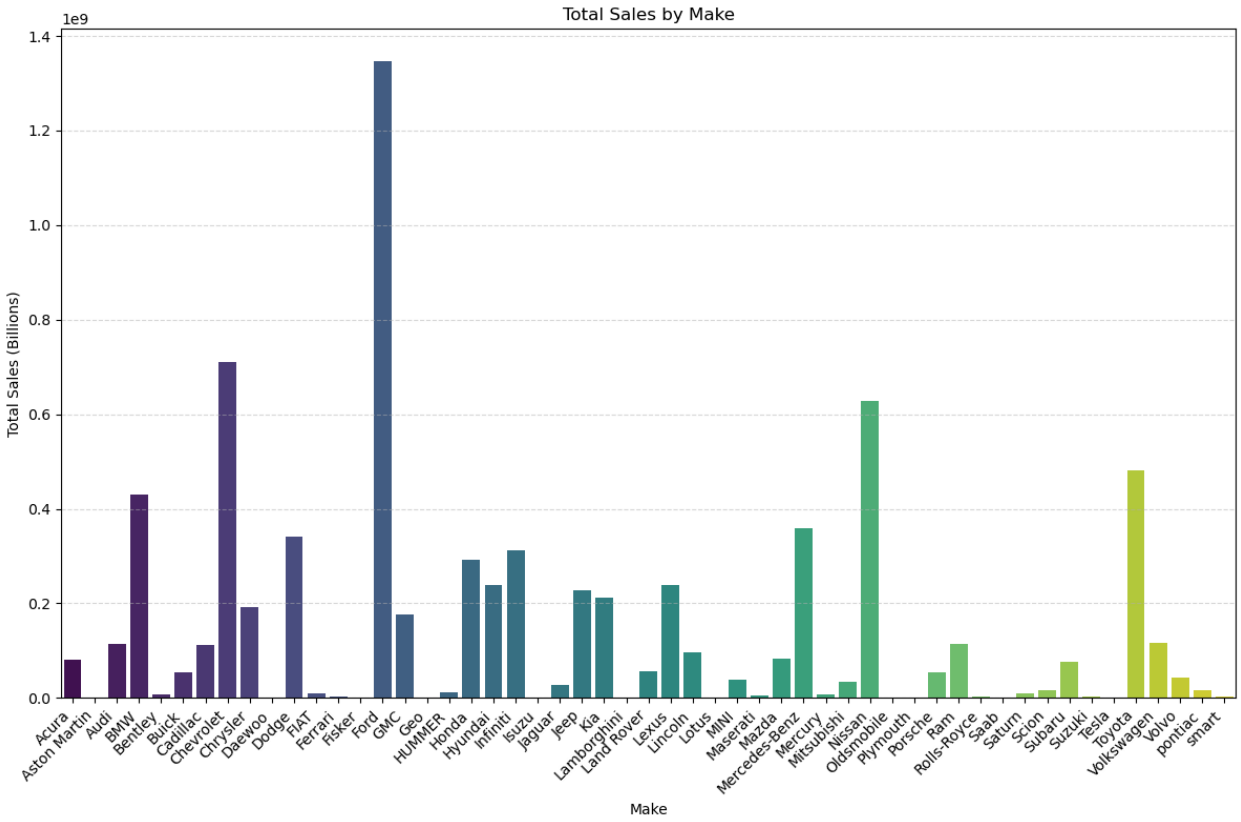
This graph represents the relationship between a vehicle's condition, rated on a scale of 1 to 50, and its selling price. The graph on the left shows the data before any adjustments were made. We can see that there are data points for vehicles with a condition rate between 1 and 12 that have a significantly higher selling price than other vehicles with similar scores. These data points were considered outliers and removed during data cleaning. The graph on the right shows the data after the outliers were removed. There is now a positive correlation between vehicle condition rate and selling price, which means that cars with better conditions tend to sell for a higher price. However, it is important to note that correlation does not necessarily imply causation. Other factors, such as the make, model, and year of the car, could also affect the selling price.



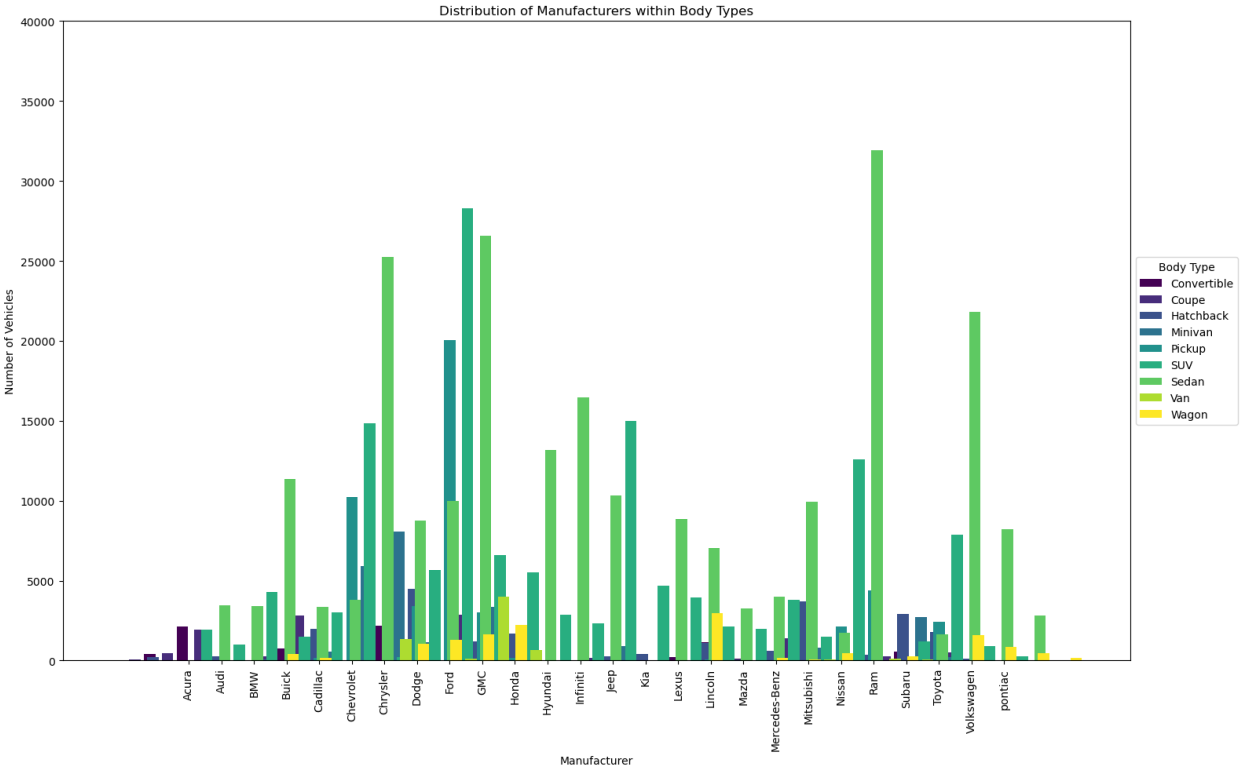
Afterward, we tried to understand the relationship between a vehicle's odometer reading and its selling price. We had outliers, that were causing difficulties with the graph readability. Data points considered outliers were removed during data cleaning and are not included in this graph. There is a negative correlation between odometer readings and selling price. This means that cars with higher odometer readings tend to have lower selling prices.



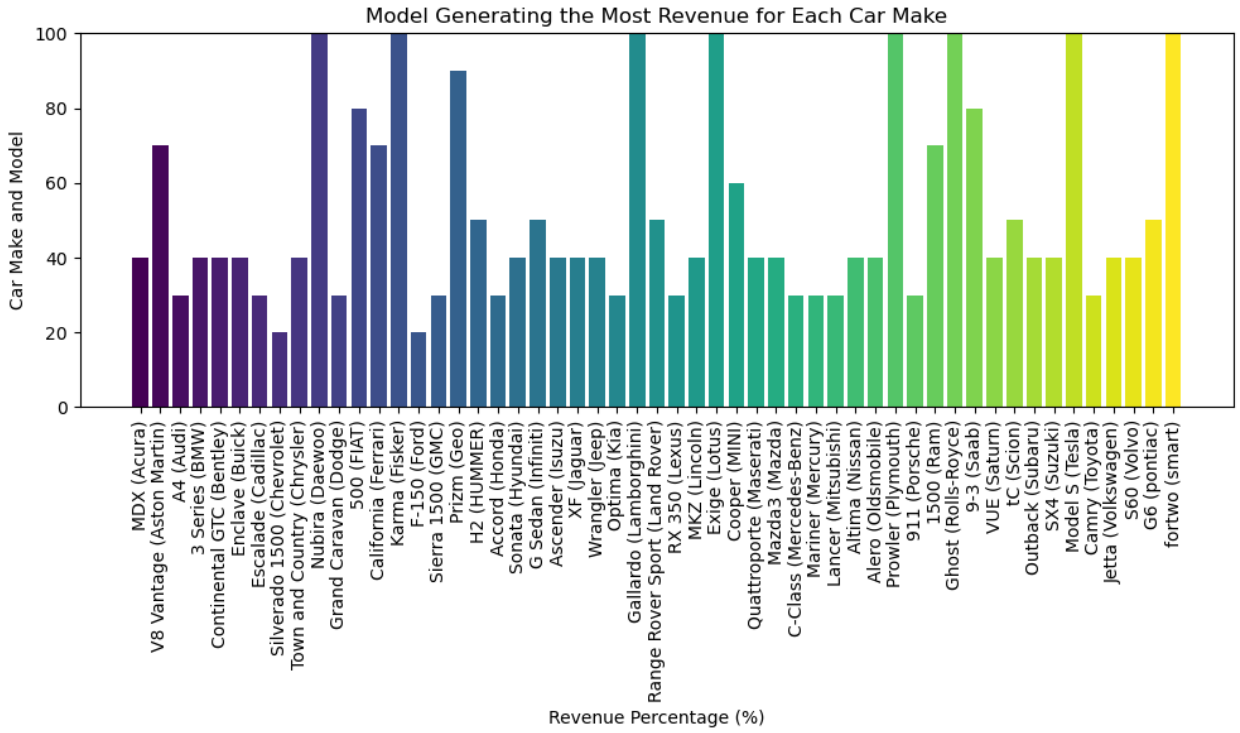
This graph analyzes the trend in total vehicle sales between 2001 and 2015, based on the information from the data set we have. This graph suggests a possible boom period for vehicle sales between 2001 and 2013, followed by a potential decline in 2014. As we can see from the graph, the highest sales were in 2013, and the lowest sales in 2001.



This graph represents the total sales (in billions) for various car makes. While analyzing the original data, duplicate car names were identified, and as a part of the data cleaning, as mentioned above, they were mapped together to ensure an accurate representation of sales figures. According to the graph, the cars makes with the highest total sales are Ford, Chevrolet, Nissan, and Toyota accordingly.



The bar chart presents how different car makers are spread out among different car body types. The graph helps to understand which types of cars each maker focuses on. For example, Ram's main concentration is on Sedans, while Ford is concentrated on Pickups. By looking at the heights of the bars, we can see which body types are popular for each maker. This helps to understand what kinds of cars are in demand and how makers position themselves in the market. Overall, this chart gives us a clear picture of the car industry, helping us see trends and understand how makers operate in the market.



The visualization depicts the revenue percentages generated by different car models across various car makes. The data is represented through a bar chart, with each bar indicating the revenue percentage associated with a specific car model. The bar chart illustrates the distribution of revenue percentages among different car models for each car make. Each bar represents a car model, with its height indicating the revenue percentage it contributes to its respective car make. From the visualization, it is evident that certain car models contribute significantly more to their respective car makes' revenue compared to others. For example, the Gallardo model of Lamborghini has the most significant influence compared to others.