Exploring the Link Between Safety Ratings and Car Price

MARTA BRASOLA 905305

SETTEMBRE 2023

Index

Introduction

Certainly, when it comes to buying a car, consumers take various factors into account, such as brand reputation, specific model features, engine specifications, and fuel efficiency. However, one critical consideration that often emerges is safety. Are people willing to pay a premium for a vehicle that offers superior safety features and performance? As the Figure 1 shows, 55% of the people surveyed said that safety is one of the most important factors when buying a car.

From an economic standpoint, prices in a competitive market should ideally incorporate all available information, including safety ratings. With this project, I aim to delve into this relationship between car prices and safety tests. I want to explore whether there's a discernible connection between the two variables and if a straightforward model can help us quantify this relationship.

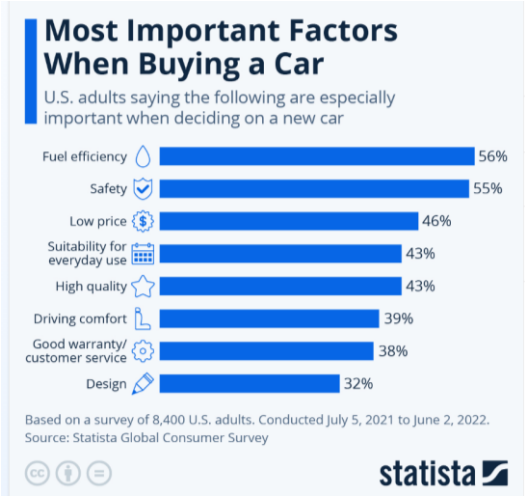


Figure 1 – source statista

What makes this project particularly intriguing is that, as today, there was no existing dataset that directly linked car prices and safety test results. Therefore, by using web scraping techniques I tried to build this dataset from scratch, to gain insights into the complex interplay between car safety and pricing in the automotive market.

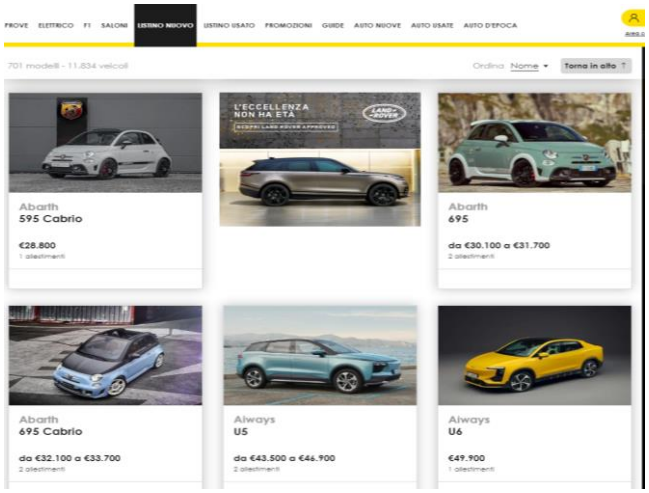
Data Acquisition

As previously said, there is not any ready-to-use dataset in which all the informations I need are stored. For this reason, I needed to choose different sources of data and create the dataset myself. To analyze the car prices I used a website called Automoto.it which stores prices of new and used cars and by comparing with other websites it was the one with more car's models. Regarding the safety tests I choose to scrape data from Euro NCAP reasonable because it is one of the most authoritative websites in the field of automotive safety testing and aligns with European safety standards.

Automoto.it

Automoto.it was initially established as a motorcycle classifieds website, but over the years, it has evolved to become a platform for the resale of used motorcycles and cars. However, it's not limited to that; Automoto has also transformed into a hub for gathering news, events, and reviews related to the world of vehicles. Their list of new cars is constantly updated with a wealth of information. When comparing this website to other vehicle portals, it appeared to me as the most comprehensive in terms of information. That's why I decided to perform web scraping on Automoto.it.

To extract information from the website, I used Python along with libraries like BeautifulSoup and Selenium. First, I loaded the page with Selenium and then created an automated action to click on the banner that appears on the freshly opened URL. Subsequently, still utilizing Selenium, I executed an automated script to load all the pages and allow the retrieval of all the information. To achieve this, I had to identify the "container" where the car banners were located and ensure that the page scrolled to the end of its height, then paused for a second (*time.sleep*) to load, and proceeded indefinitely (*while loop*) until the *new_height* and *last_height* were equal. At that point, the loop terminated because it had reached the condition.



Once all the pages were loaded, I utilized BeautifulSoup and HTML tags to extract all the information I was interested in: the brand and model name of the car, the price range, and the number of trim levels. At the end of the scraping process, my dataset consists of 701 rows and 3 columns.

Naturally, the price of a car is influenced by various factors, and it's not solely determined by its safety test results. While my analysis represents a simplified snapshot of reality, I've taken deliberate measures to maintain as much impartiality as possible. For instance, I made the decision to focus on scraping data for new cars exclusively. This approach was chosen to minimize the impact of wear and tear on a vehicle's condition, ensuring that the safety test results remain a significant factor.

Another critical consideration was the choice of trim levels. Each trim level often commands a different price point, leading to a sometimes substantial, price range for the same model. To reduce the influence of these variations on my analysis, I opted to gather data exclusively for the base model. This way, any mark-up associated with additional features or enhancements found in higher trim levels was not factored into the analysis.

Furthermore, I conducted a careful verification of the safety tests for base trim levels. This involved a random sampling check based on the official reports provided by Euro NCAP. The positive outcome of this verification process reinforced my confidence in the validity of safety test results for base trim levels.

Given these considerations, I concluded the data scraping from the Automoto.it website, ensuring that the dataset used for analysis maintains a high degree of reliability and consistency. This approach enables a more accurate evaluation of the relationship between safety ratings and vehicle prices, allowing for a nuanced understanding of this complex interplay in the automotive market.

This is the final scraped dataset:

- “brand_model”;
- “price_range”;
- “setups”.

Euro NCAP

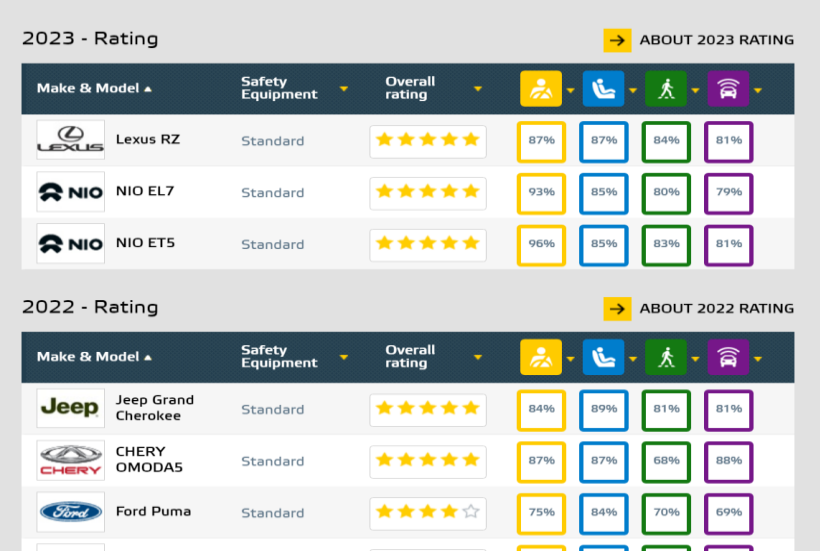
Euro NCAP is a private company that conducts safety tests according to European standards and provides a final rating in terms of stars, which aggregates points from individual categories. The star rating system is described in this manner on their website:

- **5 stars:** overall excellent performance in rash protection and well equipped with comprehensive and robust crash avoidance technology;

- **4 stars:** overall good performance in crash protection and all-round; additional crash avoidance technology may be present;
- **3 stars:** at least average occupant protection but not always equipped with the latest crash avoidance features;
- **2 stars:** nominal crash protection but lacking crash avoidance technology
- **1 star:** marginal crash protection and little in the way of crash avoidance technology;
- **0 stars:** meeting type-approval standards so can legally be sold but lacking critical modern safety technology.

It's important to note that if a car only meets the minimum legal requirements for safety equipment, it is evaluated as having 0 stars. This means that cars that do not receive excellent ratings are not necessarily unsafe, but they are not as safe as competing cars that have received better ratings.

In this case as well, I proceeded with web scraping using Python libraries. I used Selenium only to close a banner since the information was already fully loaded. To conclude the scraping process, I continued with BeautifulSoup, and once the appropriate HTML tags were identified, I scraped all the information present in the table.



The dataset consists of the following columns:

- `brand_model`: This column contains the brand and model of the car;
- `safety_equipment`: This column indicates the type of safety equipment used for the safety test. It can have values "Standard" and "Safety Pack." For your analysis of the base trim, you will focus on cars with "Standard" safety equipment, as it does not increase the base price;

- `rating_year`: This column represents the year in which the rating was conducted. It will be useful for dataset cleaning, as ratings for models are often updated over the years;
- `star_rating`: This column contains the number of stars assigned by the website according to the rating system described earlier;
- `adult_occupant_safety`: This column represents the percentage of safety for an adult occupant;
- `child_occupant_safety`: This column represents the percentage of safety for a child occupant;
- `road_users_safety`: This column represents the percentage of safety for road users;
- `safety_assist`: This column indicates the level of safety assistance provided by sensors (detecting other cars, objects, etc.).

These columns provide comprehensive information about the safety ratings and features of the tested vehicles.

On the Euro NCAP website, there are also ratings for cars produced before 2009, which I did not include in the scraping process. This decision was made because safety tests conducted prior to that year followed a completely different methodology. Including these older ratings in the dataset would have made the data incomparable due to the methodological differences. For these reasons, I chose not to initially include them in the analysis. This approach ensures that the analysis is based on consistent and comparable data, making it more meaningful and relevant for the modern automotive market.

Data Integration

Data Cleaning

Before proceeding with the actual data integration through their union, I had to perform an initial data cleaning to avoid potential conflicts or inconsistencies. Data cleaning is a crucial step to ensure that the dataset is accurate, complete, and free from errors, making subsequent analysis more reliable and meaningful.

Automoto.it

The data collected from the Automoto.it website initially appeared in the following format:

	brand_model	price_range	setups
0	Abarth 500e	da €37.950\n a €43.000	3 allestimenti
1	Abarth 500e Cabrio	da €37.950\n a €43.000	3 allestimenti
2	Abarth 595	€26.800	1 allestimenti
3	Abarth 595 Cabrio	€28.800	1 allestimenti
4	Abarth 695	da €30.100\n a €31.700	2 allestimenti

As you can observe, the "price_range" column contains not only the minimum and maximum prices for a particular car model but also other symbols and letters. To make the dataset cleaner, I performed the following actions:

- Cleaned the "price_range" column using a regex pattern to split it into two columns: "min_price" and "max_price" for the range of trim levels. In cases where the scraper found only one price, it indicates that there is only one trim level, and thus, the price is singular. However, even in situations with multiple trim levels, there might be a single price. In such cases, I decided to impute the minimum price into the "max_price" column.
- Separated the "brand_model" column into "brand" and "model" of the car.
- Ensured that brands such as Alfa Romeo and Aston Martin were correctly assigned.
- Removed rows containing keywords like "Furgone" and "Telaio" because I am specifically interested in passenger models.

After completing the data cleaning phase on the dataset extracted from the Automoto.it website, the dataset had a shape of (627, 6). This cleaning process helps ensure that the data is ready for further analysis and maintains its integrity and accuracy. The dataset extracted from EuroNCAP required formatting of the columns containing percentages to remove the symbols. The "brand_model" column also needed the same formatting.

Euroncap.com

The issue with this dataset primarily revolved around duplicates. When a safety test is no longer valid, it is re-conducted and reintroduced as the same model but with a different year. Hence, the year in which the test was conducted was essential for identifying duplicates. Naturally, to maintain data validity among duplicates, I chose to retain the car model with the most recent safety test. This approach ensures that the dataset contains the latest and most relevant safety information for each car model, enhancing the quality and accuracy of the analysis.

lj:

	brand_model	safety_equipement	rating_year
201	mazda 6	Standard	2018
408	mazda 6	no_info	2013
580	mazda 6	no_info	2009

Data Merging

To perform the actual merge of the two datasets, I utilized the pandas library's "merge" function. I used "brand_model," "brand," and "model" as the identification keys for the merge operation, and I performed a "left" merge, with the dataset containing car prices placed on the left side.

After the merge operation, it's expected that the number of rows in the merged dataset matches the number of rows in the Automoto dataset, given that duplicates were eliminated during the Data Cleaning process. However, this merging process might have resulted in several instances of null values.

Therefore, the final dataset that I subsequently loaded into the database and used for exploratory analysis comprises 169 car models. These models provide a comprehensive basis for conducting various analyses and gaining insights into the relationships between safety ratings, car prices, and other factors.

- **“brand_model”:** model and brand of a car;
- **“setups”:** number of setups;
- **“min_price” :** minimum price;
- **“max_price”:** maximum price;
- **“brand”:** brand;
- **“model”:** model;
- **“safety_equipement”:** safety equipment of the car, can be Standard or no_info;
- **“rating_year”:** year the rating was done;
- **“star_rating”:** how many stars;
- “**adult_occupant_safety”:** percentage of points in this category;

- **“child_occupant_safety”:** percentage of points in this category;
- **“road_users_safety”:** percentage of points in this category;
- **“safety_assist”:** percentage of points in this category;

Data Storage

To store the data, I opted for a NoSQL database for several compelling reasons. Firstly, NoSQL databases offer unmatched flexibility when it comes to schema design. Although the dataset is relatively simple at this stage, it's essential to plan for future expansion. Expanding the dataset with a more extensive range of data points would enable more in-depth analysis. For example, incorporating historical data on the number of cars sold per model or data that can be used to gauge brand awareness. By incorporating these potential future data points, starting with a flexible schema makes the storage of additional data more straightforward.

Furthermore, NoSQL databases are well-suited for handling unstructured or semi-structured data, which might be encountered in future data additions. This adaptability ensures that the database can accommodate various data formats seamlessly.

In addition to flexibility, NoSQL databases excel in scalability, making them a robust choice for handling large volumes of data efficiently. As the dataset grows and new data dimensions are introduced, the NoSQL database will continue to perform optimally, facilitating extensive analyses and insights.

In summary, the decision to utilize a NoSQL database aligns with a forward-thinking approach, allowing for the storage and analysis of not only the current dataset but also potential future data expansions. This adaptability and scalability ensure that the database remains a valuable resource for conducting in-depth and comprehensive analyses in the ever-evolving automotive domain.

Data Quality

Data quality is an essential aspect that must be taken into consideration because poor data quality can lead to incorrect analysis. That's why, to assess the quality of the data I collected, I decided to employ measures of *completeness* and *consistency*.

Completeness

Regarding data completeness, my final dataset does not contain any null values within it. However, I believe it's essential to consider how many null values were generated during the merging of the two datasets. In this case, out of 627 rows, 457 have null values in the section related to safety test information. The final dataset comprises only 169 unique values, which was somewhat expected because safety tests conducted annually cannot keep pace with the number of cars introduced to the market in the same time frame. Therefore, the final dataset contains only 27% of the initial values.

This assessment of data completeness highlights that while the initial dataset was more extensive, the merging process resulted in a significant reduction due to differences in data availability and updates in safety test information over time. Understanding these dynamics is crucial for interpreting the dataset accurately and accounting for any potential biases or limitations in the analysis.

Consistency

The purpose of the consistency analysis is to understand why 627 values did not find a match in the final dataset. To achieve this, I randomly sampled 10% of the dataset and conducted manual verification.

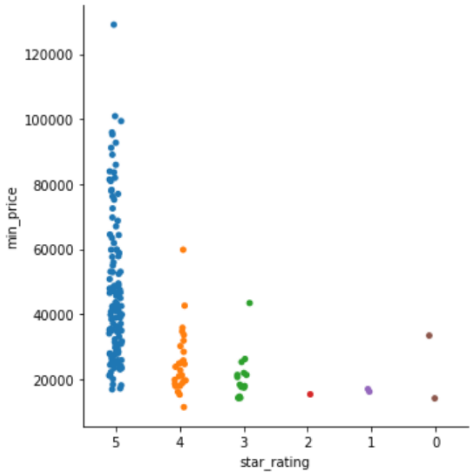
reason	
not_present	0.767442
version	0.162791
simbols	0.069767

In 76% of cases, the reason why a data point was not present in the final dataset was that a safety test was never conducted for that specific car model. In just over 16% of cases, it was related to different versions of the same model. However, in all of these cases, the base model was present in the final dataset. In less than 1% of cases, the issue was related to symbols, particularly in the case of Citroen. This was because, at times, the brand name was written with symbols, while other times it was written without symbols. This variation in representation hindered the matching of elements.

These insights from the consistency analysis provide valuable information about the reasons behind data mismatches, helping to clarify the causes and contributing to a more thorough understanding of the dataset's characteristics and limitations.

Exploratory Data Analysis

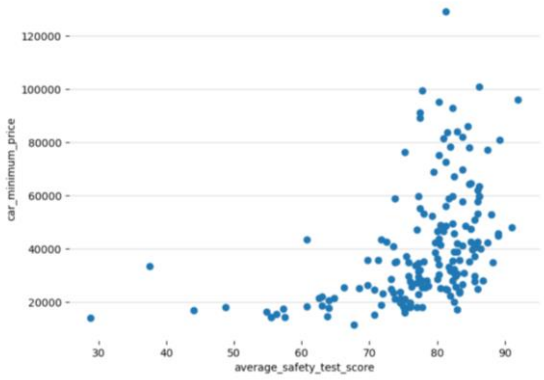
In this section, I will attempt to provide a brief analysis of the collected data and explore whether there is indeed a relationship between safety tests and car prices. The price under consideration is always the minimum price for each model.



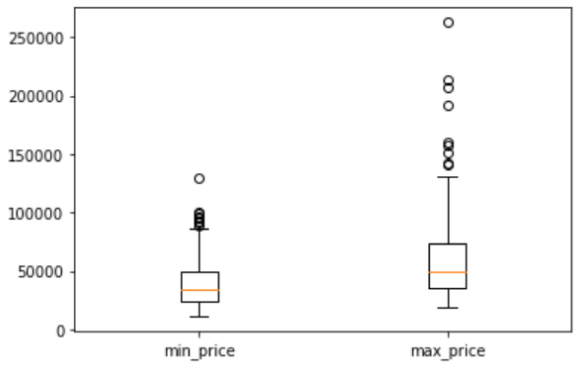
In this case, it's evident that the data is not evenly distributed. Most of the cars are classified with 5-star safety ratings, and within this category, there is a high variability in prices.

To better analyze the relationship, I calculated the average safety statistics and created a scatter plot to illustrate the connection. While the relationship is not strictly linear, the correlation coefficient between the two variables is approximately 40%. To explain the relationship between these two variables more comprehensively, more detailed analyses and the inclusion of additional factors in the model would be necessary.

This suggests that safety ratings play a role in determining car prices, but the relationship is influenced by various other factors that require a more in-depth investigation. Further analysis with a richer dataset and a more extensive set of variables would provide a clearer understanding of this complex relationship.



	min_price	average_safety
min_price	1.000000	0.445513
average_safety	0.445513	1.000000



This represents the distribution of minimum and maximum prices for different trim levels. As one might expect, the maximum price exhibits greater variability compared to the minimum price. While the minimum price also includes some outliers, they are fewer in number compared to the outliers in the maximum price range.

Conclusion and future development

To conclude, after collecting, cleaning, integrating, and analyzing all the data, the relationship between a car's safety test and its price does not appear to be very strong. However, it's important to note that considered the limited amount of data I have been able to gather, it would indeed be valuable to expand the dataset and incorporate additional factors that influence car prices. This expansion could include historical sales data for car models, other characteristics of the model or brand in general, and even an analysis of used cars. By collecting more data points, a more comprehensive analysis can be conducted, enabling a more meaningful understanding of the factors that impact car prices.

Incorporating additional information into the dataset would help in better isolating the influence of safety features on a car's price and discerning how consumer purchasing habits are affected. This broader dataset would allow for a more in-depth exploration of the complex interplay between safety characteristics, market trends, and consumer preferences.

Expanding the dataset in this manner would not only enhance the depth of analysis but also provide more robust insights, making it a valuable resource for researchers and analysts in the automotive industry.