Automobile Data - Insurance Scores

by Carlos Martinez

Problem Statement:

In the dynamic field of the insurance industry, accurate risk assessment and pricing are important for ensuring a sustainable business strategy. The existing challenge lies in the need for a refined system that analyzes comprehensive auto data to provide insurance companies with a holistic understanding of potential risks associated with diverse automobile profiles. The available dataset encompasses crucial information, including auto specifications, insurance risk ratings, and normalized losses.

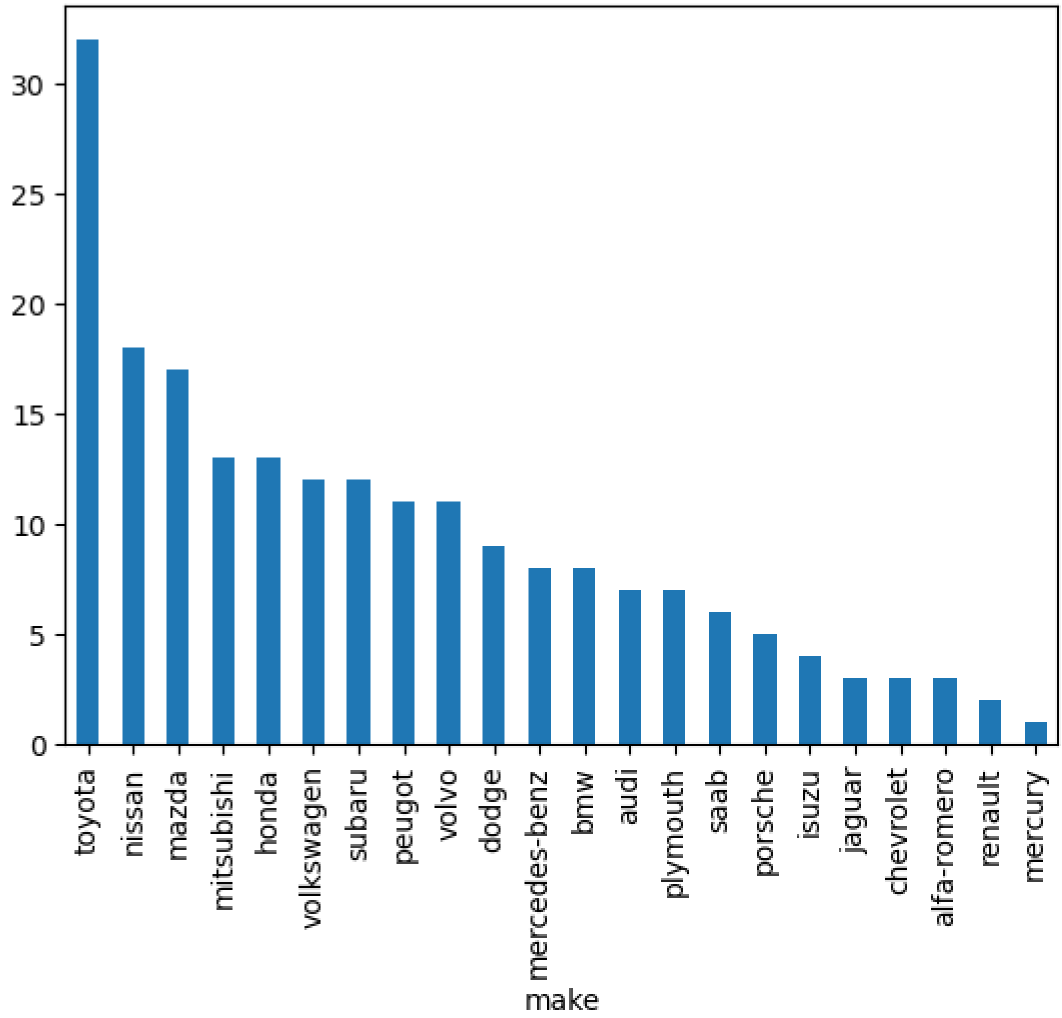
The primary objective of this project is to develop a machine-learning model solution that conducts a concise analysis of the provided auto data. The proposed solution aims to empower insurance companies with a tool that goes beyond conventional risk evaluation methods. By leveraging advanced data analysis techniques, this project looks to enhance the accuracy of risk assessments and provide a more informed basis for pricing insurance policies. Ultimately, the outcome of this analysis is anticipated to contribute to the development of a more effective risk management strategy for insurance companies operating in the auto insurance sector.

Exploratory Data Analysis

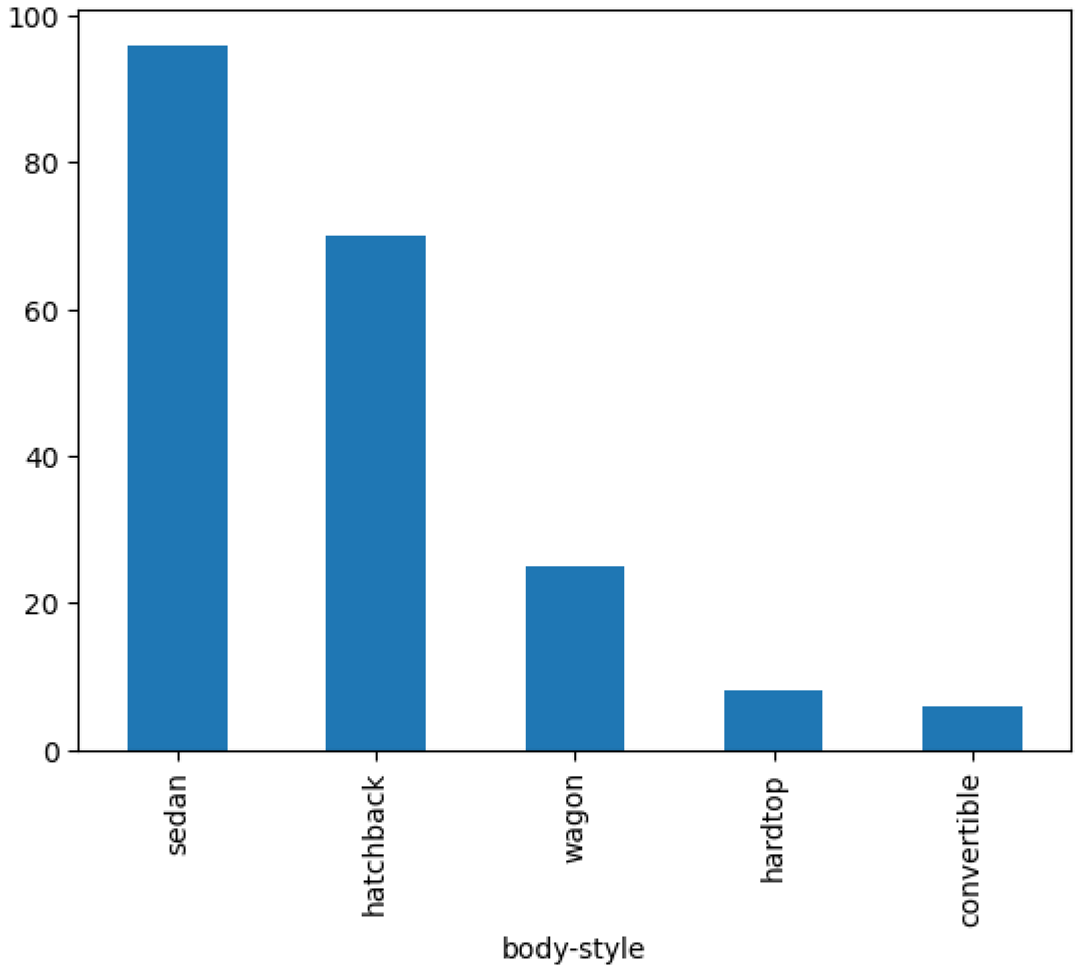
I started by importing that dataset and reformatting it to include labeled column names, and a consistent datatype format across the various value inputs. From there I looked at missing values and I identified that the “normalized-losses” column had over 41 missing values. Due to the amount of missing values, I ended up dropping the “normalized-losses” column since it would not provide any significant information to our analysis.

After the initial processing, I performed a univariate analysis to identify outliers or errors in the dataset and to plan additional processing steps and modeling approaches.

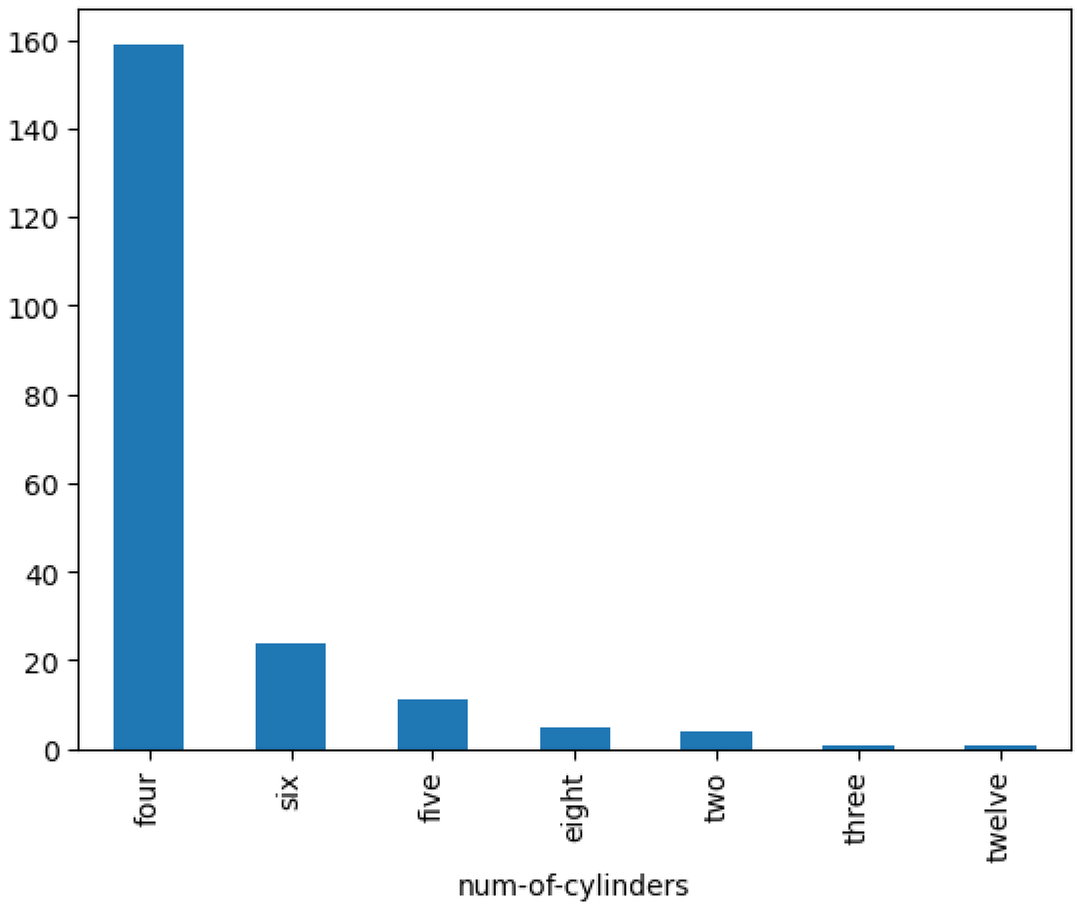
Looking at the make of the cars in the dataset: The dataset shows that Toyota is one of the most popular cars from the dataset.



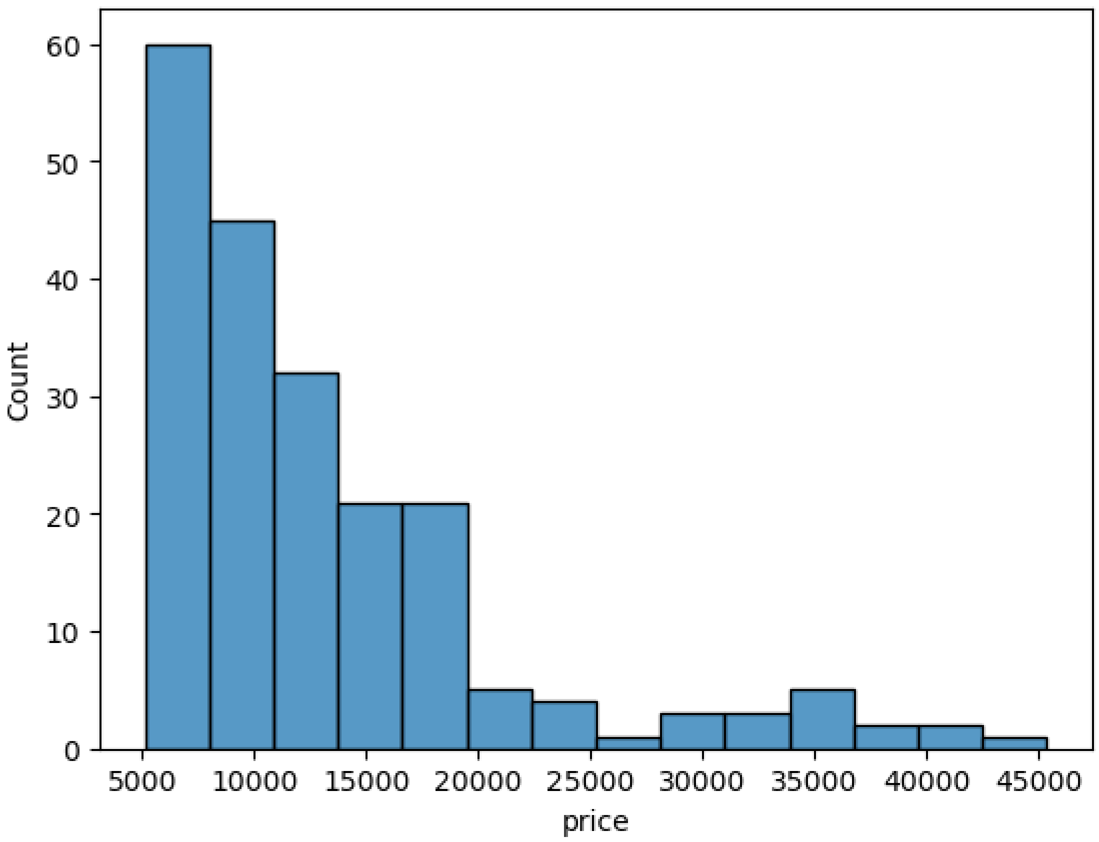
Looking at body style: The sedan car model is the most popular, followed closely by the hatchback.



Looking at num of cylinders: Cars with four cylinders were the most popular.

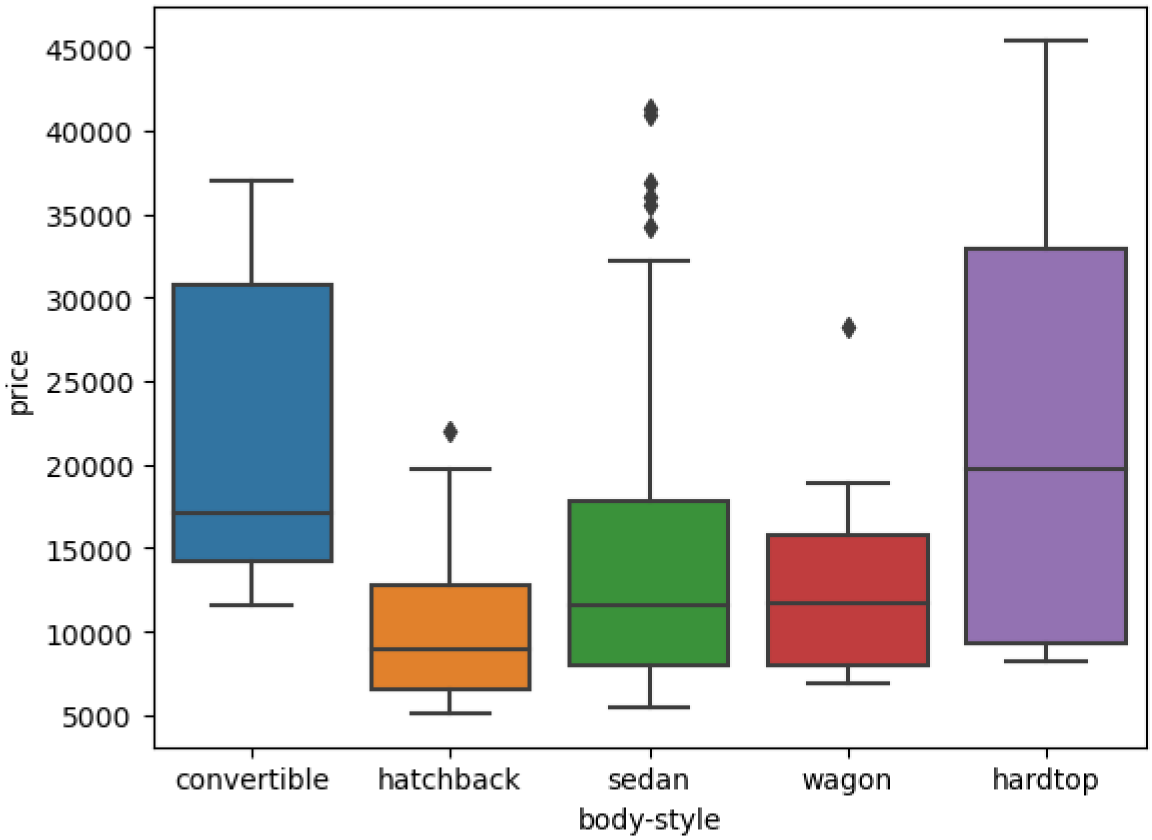


Looking at price: We observe that our dataset is right skewed. Meaning that the mean is greater than the median. With cars on the lower end of the price scale, being bought more often.

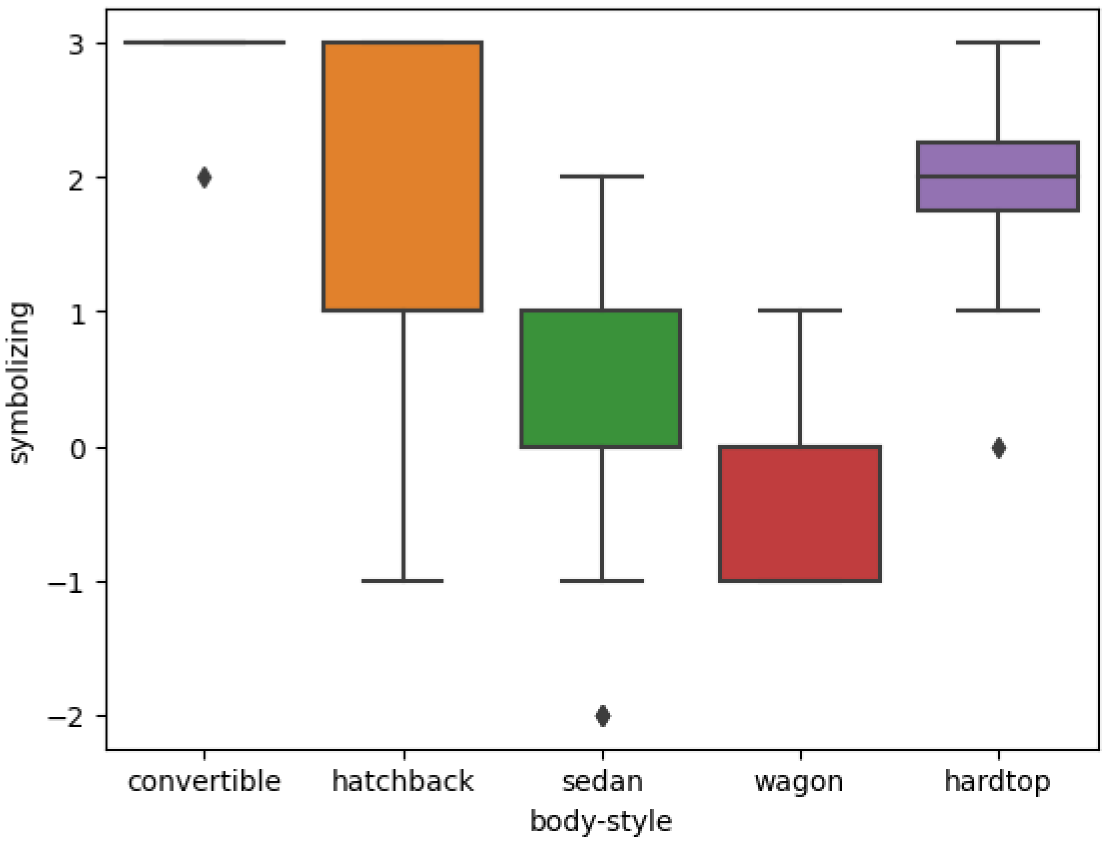


After conducting the univariate analysis, the next step in exploring this dataset was doing a bivariate analysis to look at the relationship between different variables in our dataset to identify patterns, correlations, or dependencies.

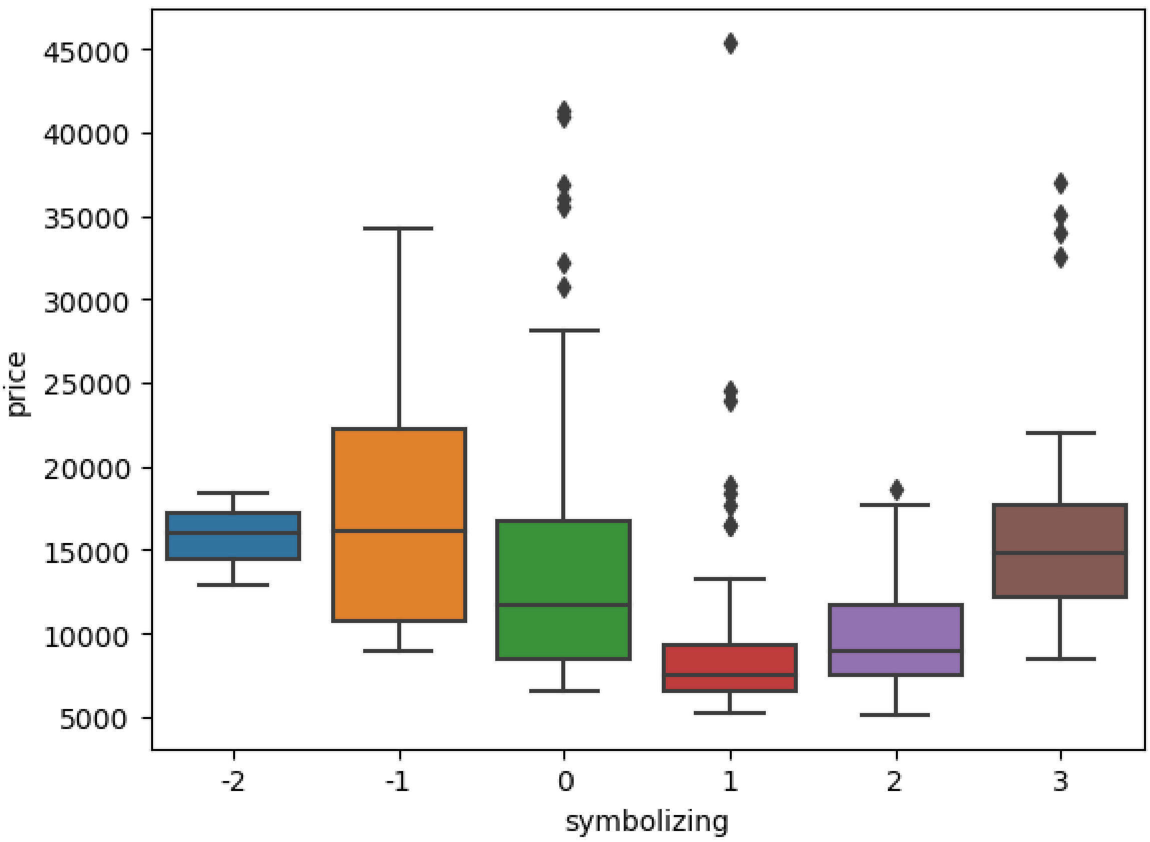
Looking at body-style and price: We observe outliers in the hatchback, sedan, and wagon body style.



Looking at body-style and symbolizing score: We observe few outliers in the sedan and hardtop bodystyle.



Looking at symbolizing score and pricing: We observe many outliers in the symbolizing scores of 0, 1, 2, and 3.



Data Wrangling

The dataset used in this project is sourced from the 1985 Model Import Car and Truck Specifications, Personal Auto Manuals from the Insurance Services Office, and the Insurance Collision Report from the Insurance Institute for Highway Safety. Previously employed in predictive modeling for car prices, the dataset, comprising 159 instances, utilized an instance-based learning algorithm.

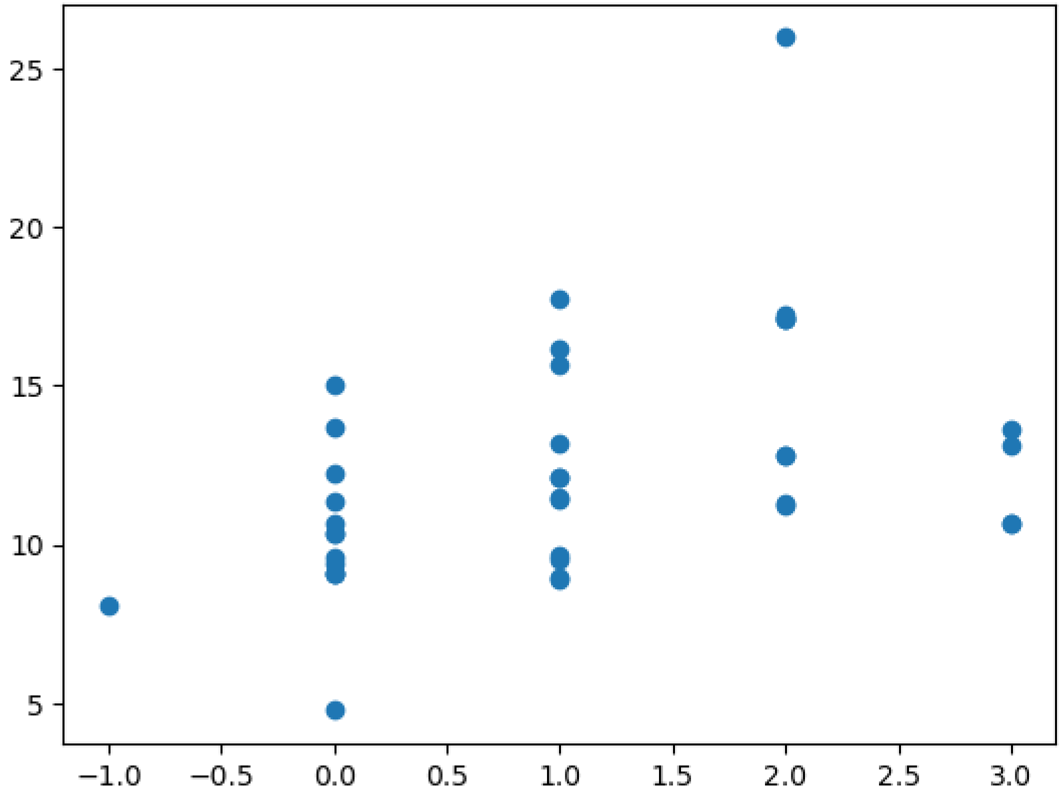
Upon thorough exploration, certain columns were identified as less useful for our analysis. Subsequently, after further investigation of the "normalized losses" column revealed a significant number of missing values, it was removed from our dataset.

Following the initial exploration, the focus shifted to examining various car factors to estimate symbolizing scores. Additionally, an area of interest emerged—analyzing how different car factors can be leveraged to estimate prices.

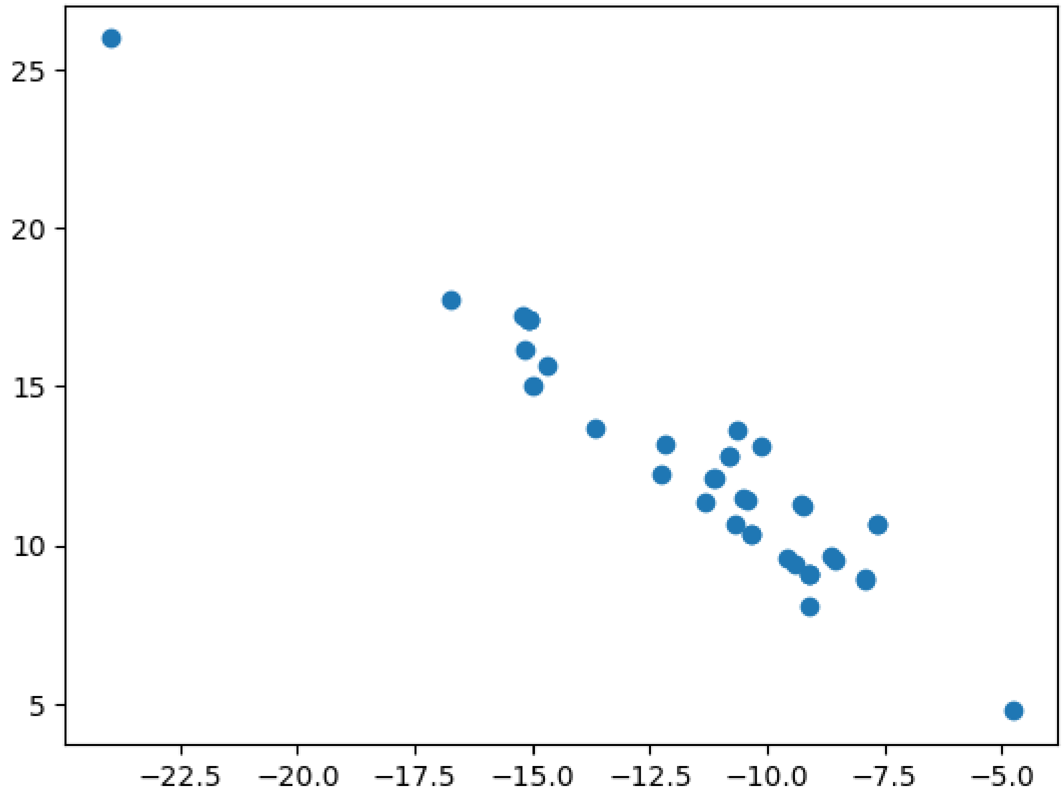
Modeling

The model selected to predict symbolizing scores for various car factors was a linear regression model. This process included first splitting our data to have both data to train and test our model. Then, scaling our data, since linear regression is dependent on a measure of distance, we used a standard scaler to scale our dataset. From there, we fit our dataset to our linear regression model. The resulting value of our linear regression model evaluation was a r_2 score of -127.64995633330477. This is indicative that our model requires refinement to further improve it. Some of the approaches could include removing outliers, errors, and columns that wont be useful in our data analysis. Below are some visual of the residual scatterplots to visualize our linear regression model:

Plotting a line through the test set:



Scatterplot:



Results and Discussion

The linear regression model was implemented to predict symbolizing scores based on various car factors. However, the obtained R-squared score of -127.65 indicates a substantial need for model refinement. To enhance the model's predictive performance, several next steps are recommended:

- **Outlier Detection:** Identify and address outliers in the dataset, particularly in variables like symbolizing scores, price, and body-style, as observed during the bivariate analysis.
- **Feature Selection:** Reevaluate the relevance of all features and consider excluding those that do not significantly contribute to the model's predictive power. This step can be informed by both exploratory data analysis and domain knowledge.
- **Error Analysis:** Conduct a detailed analysis of model errors, examining cases where predictions deviate significantly from actual values. This process will provide insights into areas where the model may be misaligned with the data.
- **Data Exploration:** Further investigate potential relationships between car factors and pricing, as well as symbolizing scores. This exploration may uncover hidden patterns or dependencies not initially considered.

By addressing these steps, I aim to refine the linear regression model, improving its accuracy and applicability in predicting symbolizing scores for diverse car profiles. This iterative process will contribute to a more effective and reliable tool for insurance companies in their risk assessment strategies within the auto insurance sector.