Price and Risk Analysis: Automobile data set

COS60008 Introduction to Data Science

2024, Semester 1

Assignment 1 Report

Name: Thi Ngan Ha Do

Student ID: 103128918

Email Address: 103128918@student.swin.edu.au

Submission Date: 19/04/2024

1. Introduction

This report presents the findings of Tasks 2 and 3, focusing on examining the correlation between price and risk rating using datasets from the UCI repository. Through data cleaning and visualization, we observed a higher frequency of insurance claims for low-priced cars, contrasting with the perception of higher risk associated with high-priced cars.

2. Data Acquisition & Preparation

2.1 Data Loading

Three CSV files were loaded using pandas' "read_csv" function. Subsequently, ".equals" methods were employed to ensure that the loaded dataframes matched the raw data files. The results returned "True" for all three DataFrames, showing that the data was correctly and completely sent from the source data files.

2.2 Data Merging

The files "data1.csv" and "data2.csv" contain identical automobile sets, each with a unique ID. Merging df1 and df2 based on the "id" column using an inner join ensures that only matching IDs are merged. The generated DataFrame, "df_joined_id," matches the properties of the dataset in "data3.csv". As a result, using the 'concat' method to append df_joined_id to df3 adds additional data entries to the DataFrame.

2.3 Data Cleaning

2.3.1. Data Cleaning Issues

Error Type	Columns	Issues	Detection	Solution
Typographical errors	num-of-cylinders	‘sixth’	Used 'describe(include=[object])' to check the number of unique values within each categorical column. For columns exceeding the specified unique values in the data dictionary, the 'unique()' method was applied to identify and rectify any discrepancies.	Replaced ‘sixth’ with ‘six’
	fuel-system	‘Mpfi’		Replaced ‘Mpfi’ with ‘mpfi’
Data types	num-of-doors	4 to ‘four’ 2 to ‘two’		Changed numerical values with texts according to the data dictionary
Extra whitespaces	fuel-type	Extra whitespaces with ‘diesel’ value		Eliminated whitespace and verified the presence of two fuel types, consistent with the data dictionary

Duplicates	id	ID 10180 appeared three times in the dataset	Used “value_counts()” function to count the occurrences of each unique ID. If any ID appears more than once, it indicates a potential issue with duplicate entries	Applied `drop_duplicates`, specifying the "id" column as the subset to identify duplicates, and retained only the first instance of each unique ID. This operation directly altered the DataFrame “df”
Missing values	normalised-losses	38 null values	Used “isnull()” and “sum()” functions to check for total missing values	Used the “groupby()” function to calculate the median of each risk rating group (symboling)
	price	4 null values		Used “groupby()” to calculate the median for each make and engine size. The final missing value was replaced with the median calculated based on the make group
Impossible values	horsepower	1100, exceeding the maximum value of 288	Used the mathematical expression > (greater than) 288 horsepower	Replaced 1100 with the median horsepower of cars with similar engine sizes (136)

Table 1 summarises the detection and solution for data cleaning issues

2.3.2. Further Justification

- Missing values

From the results of the df.info() function, 'normalized losses' and 'price' are missing values, which can be filled with the median because of their non-normal distributions. However, instead of filling in the median of the entire column, categorical variables were selected based on domain expertise for grouping to ensure meaningful groups for each column.

Column 'normalized-losses': Null values for normalized losses were filled by the median of each symboling group. This is because insurance rating risk represents how hazardous a car is to insurance companies, with higher values signifying greater risk.

Column 'price’: Missing price values were filled with medians by grouping'make' and 'engine-type'. Pricing fluctuates according to car brand and maximum speed (Tsagris & Fafalios, 2022); therefore, computing median pricing within significant attributes ensures accuracy. However, when grouping by two columns, null medians are feasible, especially for make and engine type combinations with a single entry. To address this, the median of grouping just by the make column was used to fill in all null price entries completely.

- Impossible values

Column 'horsepower': Engine sizes are one of the main factors to determine horsepower of a car (Perez-Melo & Kibria, 2020). Consequently, impossible horsepower values were substituted with the median horsepower for each engine size category, enhancing data precision.

- Change data types

Column symboling: The symboling values are numerical, but they were not used for calculations. For visualisation, the data type was transformed to object and classified the risk levels.

3. Data Exploration

To go deeper into identifying similar traits among insured cars with the largest normalized losses, cross-referencing attributes and risk ratings is critical. In this study, we investigate the relationship between pricing and underwriting risks, specifically if higher prices are associated with larger risks.

3.1. Visualising categorical and numerical values

3.1.1. Numerical Column

Prior to analysing the relationship between price and risk, it is necessary to first comprehend each. Histograms were used to illustrate the distribution of insured car prices, with the aims of detecting common characteristics of insured automobiles in a car insurance company's portfolio.

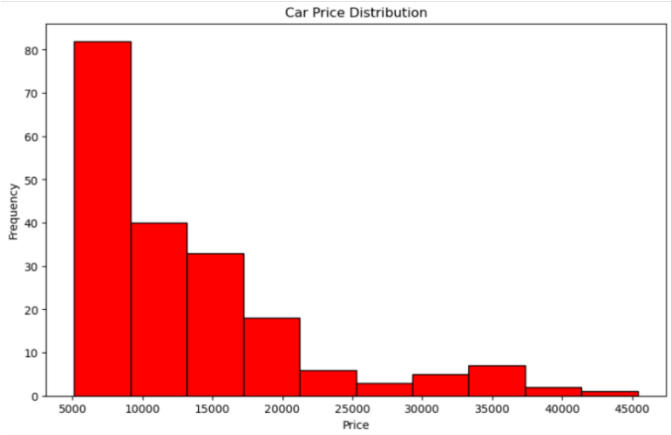

Figure 1 illustrates the distribution of car price

Figure 1: Observation
Roughly 50% of the insured cars were priced below $10,000, indicating a predominant concentration in the lower price range. The distribution displays a negative skew, implying potential outliers or extreme values towards the lower end of the price spectrum.

Key takeaway: The car insurance portfolio predominantly comprises affordably priced vehicles. However, notable disparities exist between the low-price range and the medium to high categories. As per the hypothesis, the distribution of insurance risk ratings is anticipated to primarily consist of moderate-risk and elevated-risk categories.

3.1.2. Categorical Column

Symboling	Level of Risk	Row Total	Justification
-3	Inherent Risk	0	Data type was adjusted to object to visualise frequency
-2	Minimal Risk	3	
-1	Low Risk	22	
0	Moderate Risk	65	
1	Elevated Risk	53	
2	Significant Risk	31	
3	High Risk	23	

Table 2 converts symboling to insurance risk levels

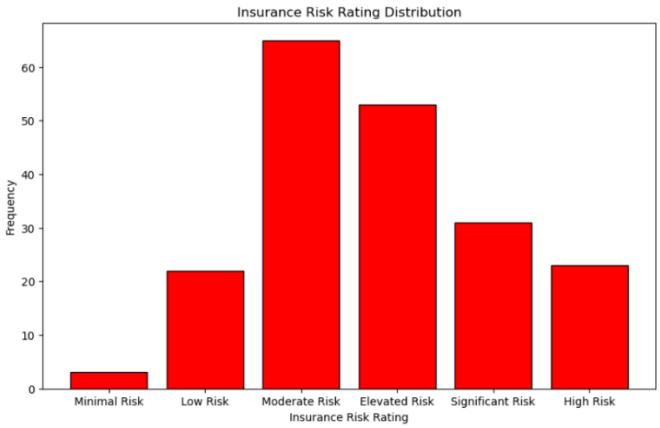

Figure 2: Observation
The most common insurance risk rating among insured cars is categorized as moderate risk and elevated risk

Figure 2 illustrates the distribution of insurance risk levels

Key takeaway: Consistent with previous findings on price distribution, the insurance risk profile consists primarily of moderate and heightened risk categories. However, this distribution does not ensure lower normalized losses for these risk categories because it is dependent on risk assessment precision. Further research is needed to analyse the relationship between normalized losses and risk rating.

3.2. Exploring relationships

3.2.1. Normalised Losses and Symboling

To evaluate risk assessment accuracy, actual expenses associated with various levels of underwriting risk are examined using the "normalised-losses" and "symboling" columns. If risk criteria were accurately assessed, normalized losses were expected to increase with higher risk levels, while outlier effects should be avoided.

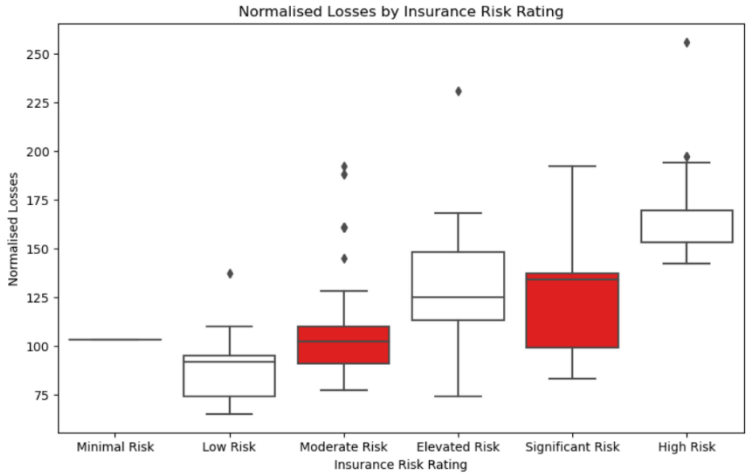

Figure 3: Observation
There is a rise in normalized losses corresponding to higher assessed risk levels
However, there appears to be significant variability in normalized losses among cars classified with elevated and significant risk levels. Additionally, there are numerous outliers in the moderate and high-risk categories

Figure 3 illustrates the relationship between risk and normalised losses

Key takeaway: Increased risk levels correspond to higher normalized losses. However, unanticipated normalized losses within the moderate and high-risk groups emphasize the need for better risk assessment in these categories.

3.2.2. Price and Symboling

Category	Make	Price Range	Row Total	Justification
Low	Chevrolet, Dodge, Honda, Isuzu, Mazda, Mitsubishi,	Under 10,000	97	Price was divided into three categories to

	Nissan, Plymouth, Subaru, Toyota, Volkswagen			investigate its association with risk rating. Price categories were added to the copy, splitting prices into bins and labelling each category.
Medium	Alfa-romero, Audi, BMW, Mercury, Peugot, Saab, Volvo	From 10,000 to under 25,000	82	
High	Jaguar, Mercedes-benz, Porsche	Over 25,000	18	

Table 3 groups car prices to price ranges

A car's brand has a considerable influence on its price. It is interesting to view the distribution of underwriting risk across different price categories.

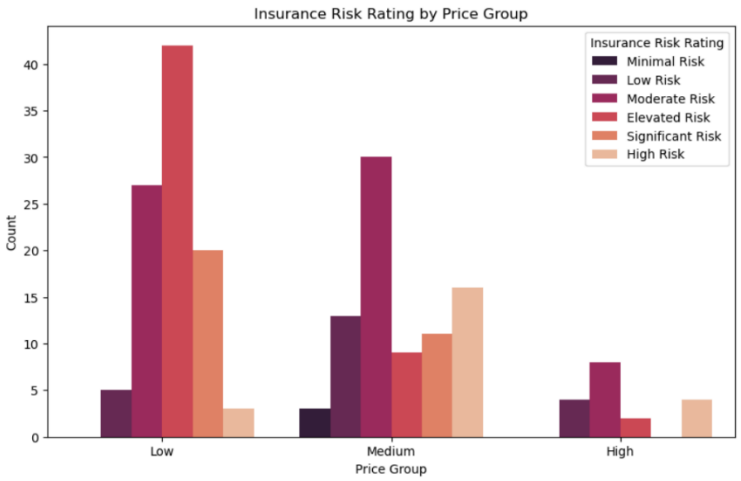

Figure 4: Observation
The predominant risk rating is moderate, except for the low-price category. Medium-priced vehicles show the highest incidence of high risk.
Normalized losses vary significantly between autos categorized as elevated or considerable risk. Furthermore, there are numerous outliers in both the moderate and high-risk categories.

Figure 4 illustrates the breakdown of price groups by risk levels

Key takeaway: At first glance, the clustered bar chart appears to contradict the idea that higher-priced vehicles pose bigger insurance risks than less expensive ones. However, given inherent inefficiencies in risk assessment (as discussed in Figure 3), it is unclear if lower-risk categories result in fewer financial losses for insurers. To learn more about which price groups contribute the most to normalized losses, consider the following relationship.

3.2.3. Normalised Losses and Make

Exploring the relationship between automobile brands and their average normalized insurance losses reveals patterns in insurance risk across manufacturers. The order of car brands was displayed in a descending order of average normalized losses, allowing for a simple comparison of insurance claims.

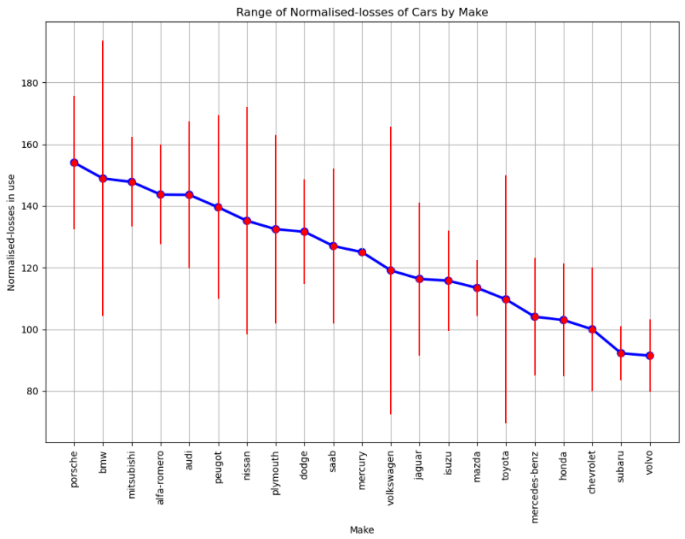

Figure 5: Observation
Normalised losses of luxury car brands (e.g, Porsche, Alfa Romeo, BMW) were recorded to be the highest. However, brands in the lower price range (e.g, Mitsubishi, Nissan) surpass some of the higher-priced brands.
Additionally, BMW, Volkswagen, and Toyota pose potential unidentified risks and unforeseen expenses in insurance, as indicated by the length of their error bars.

Figure 5 illustrates the rank of car brands based on their normalised losses

Key takeaway: Although price alone does not always imply normalised loss and risk assessment, it is critical to explore the characteristics of different price ranges, particularly the medium and low groups. The variability in normalised loss levels may make it difficult to assess insurance risk.

3.3. Scatter matrix

Previous findings indicate that, while price influences normalized losses, the relationship is not linear. To support this claim, major pricing criteria such as engine size, curb weight, horsepower, and highway miles per gallon (Tsagris & Fafalios, 2022) were chosen for comparison to normalized losses.

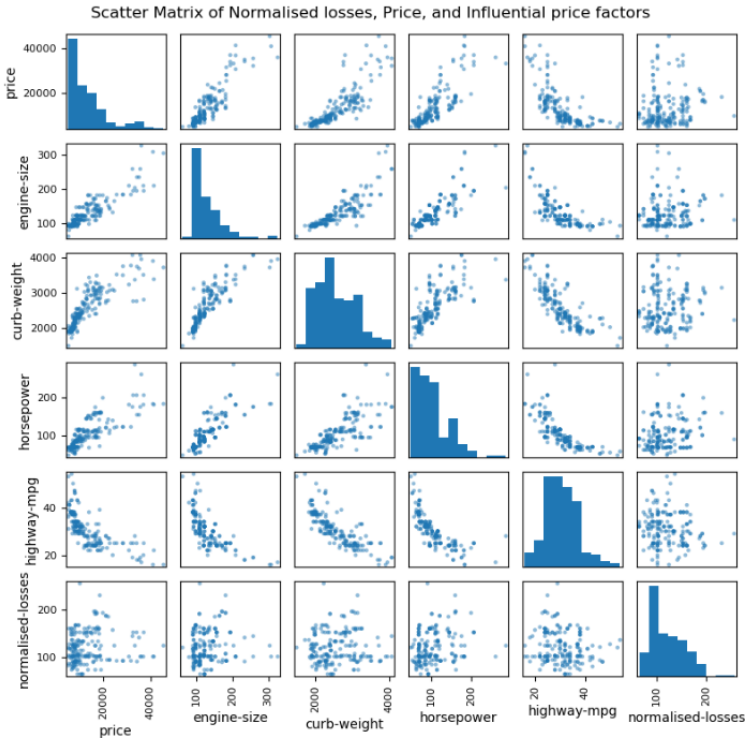

Figure 5 illustrates the relationship between normalised losses, price, and other influential price factors

- Positive correlation: The columns "engine-size," "curb-weight," and "horsepower" all showed high positive correlations (above 0.7), indicating that as these factors increase, so does the price.
- Negative correlation: Highway miles per gallon (highway-mpg) stands out as the only variable with a negative correlation. This implies a consumer preference for fuel-efficient vehicles, thereby driving up prices and elevating demand for such cars.
- No correlation: Regarding the correlation between normalised losses and other numerical columns, no discernible pattern emerges in the technical attributes of cars that could explain the amount of insurance claims.

Key takeaway: Additional data gathering on other car parameters is required to find correlations and shared traits among high-risk vehicles.

4. Conclusions

The hypothesis of this report suggests that higher-priced cars entail greater insurance risks. However, the analysis revealed that normalized losses for lower-priced cars surpassed those for higher-priced ones. There are also indications of errors in the car risk assessment.

5. Reference List

Perez-Melo, S., & Kibria, B. M. G. (2020). On some test statistics for testing the regression coefficients in presence of multicollinearity: a simulation study. *Stats*, *3*(1), 40-55.

Tsagris, M., & Fafalios, S. (2022). Advanced car price modelling and prediction. In *Advances in Econometrics, Operational Research, Data Science and Actuarial Studies: Techniques and Theories* (pp. 479-494). Springer.