Hyperiondev

TASK

Exploratory Data Analysis on the Automobile Data Set

Visit our website

Introduction

*source : https://rpubs.com/Arnav_Jain/DSP_AutomobileDatasetAnalysis#:~:text=The%20second%20attribute%2C%20“normalized-,loss%20per%20vehicle%20per%20year. *

The dataset describes used automobiles in 3 ways:

- The technical specification of the automobile

- The loss per vehicle per year given as “normalized-losses”

- The insurance risk rating of the automobile given as “symboling”

“symboling”, corresponds to a car’s insurance risk level. Cars are initially assigned a risk factor symbol that corresponds to their price. If an automobile is more dangerous, this symbol is adjusted by increasing it. A value of +3 indicates that the vehicle is risky, while -3 indicates that it is likely safe to insure.

The second attribute, “normalized-losses,” is the relative average loss payment per insured vehicle year. This figure is normalised for all vehicles within a given size category (two-door, small, station wagons, sports/specialty, etc…) and represents the average loss per vehicle per year.

Key attributes include:

- symboling : int64
- normalized-losses : object
- make : object
- fuel-type : object
- aspiration : object
- num-of-doors : object
- body-style : object
- drive-wheels : object
- engine-location : object
- wheel-base : float64
- length : float64
- width : float64
- height : float64
- curb-weight : int64
- engine-type : object
- num-of-cylinders : object
- engine-size : int64
- fuel-system : object
- bore : object

- stroke : object
- compression-ratio : float64
- horsepower : object
- peak-rpm : object
- city-mpg : int64
- highway-mpg : int64
- price : object

To conduct this analysis, we will utilize Python's data analysis and visualization libraries, including pandas, numpy, matplotlib, and seaborn. Techniques such as summary statistics, correlation analysis, and data visualization will be employed to uncover insights and highlight significant trends.

DATA CLEANING

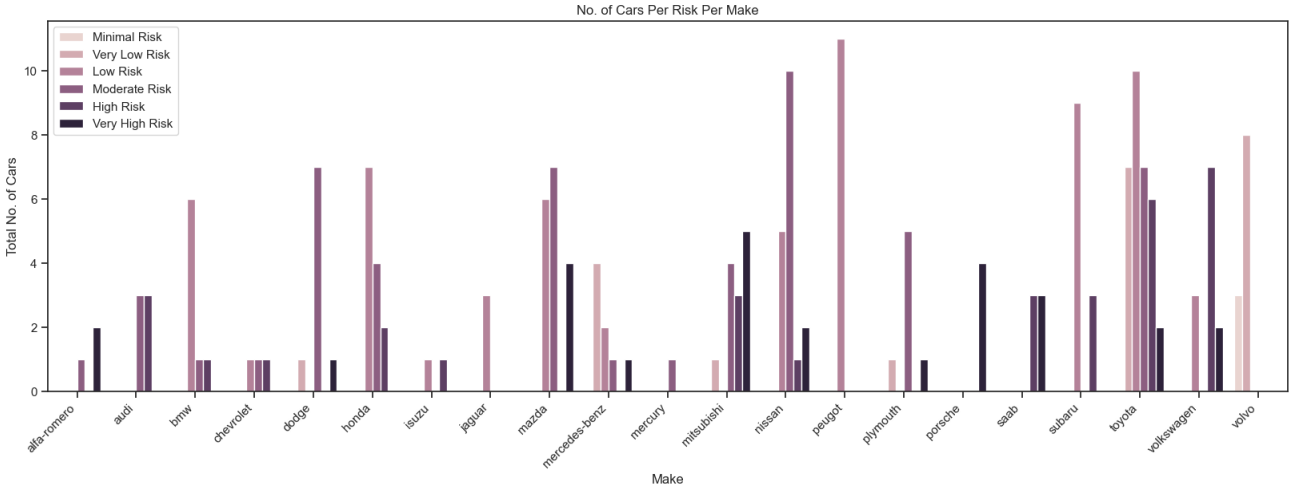
The following steps were taken to clean the data:

1. Redundant and unnecessary data was removed from the dataset. The columns removed were identified as not being relevant for the decision-making and evaluation of the data. ('normalized-losses','num-of-doors','wheel-base','length','width','height','curb-weight','engine-type','fuel-system','bore','stroke','compression-ratio','engine-size','num-of-cylinders')
2. The removal of missing/null values from the dataset.
3. The horsepower column data type was changed from string to int64.

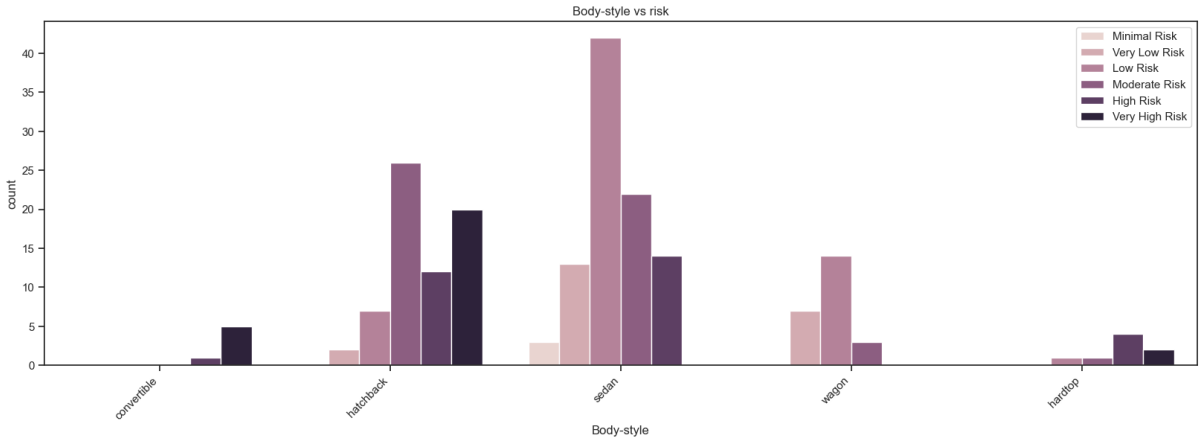
MISSING DATA

Upon observation of the data, it appeared that the missing data was represented as a question mark character. A conversion from the question mark character to null allowed for the drop of the null values. This is how the missing data was dealt with.

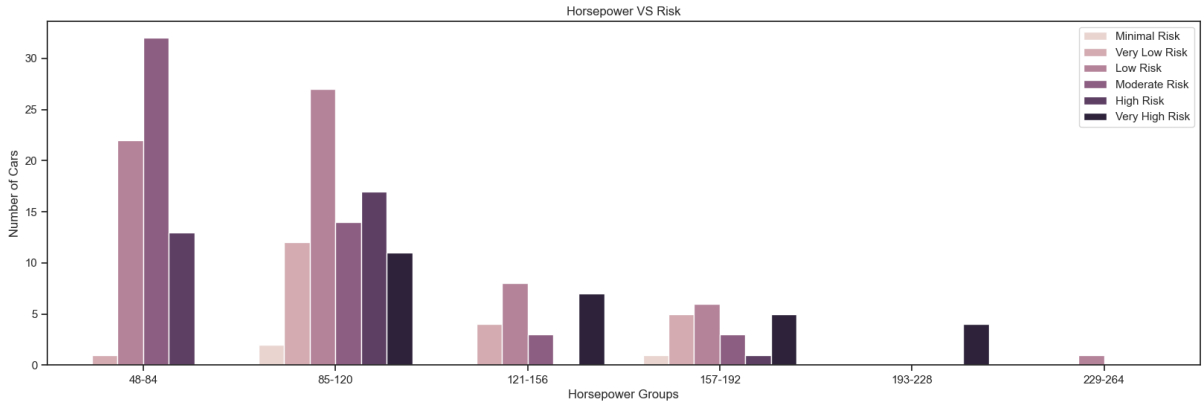
DATA STORIES AND VISUALISATIONS



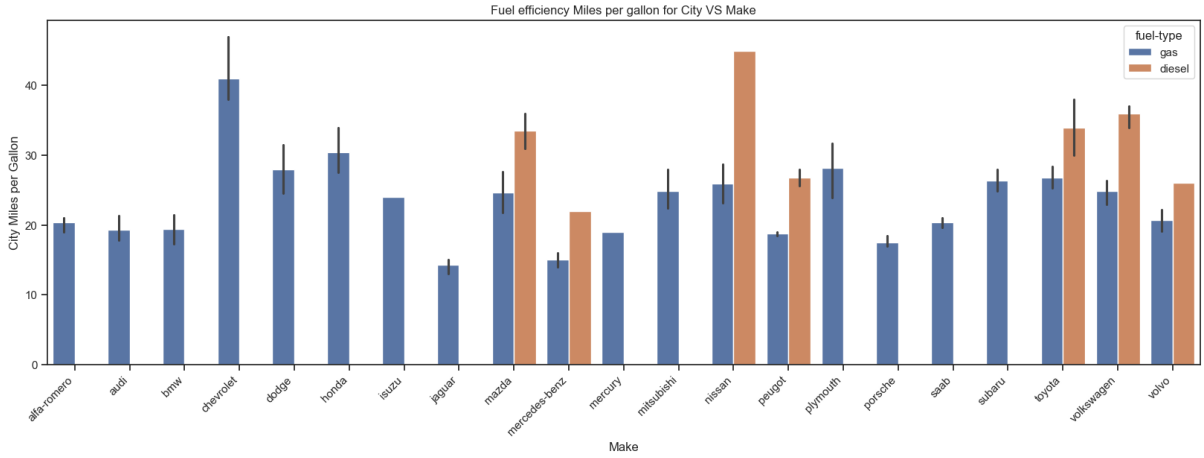
This graph displays the relationship between brand and total number of cars counted and their risk factor. The more popular brands of cars have cars in most of the risk factors while the less popular car brands don't feature in across as many risk categories. It can also be noted that certain makes of cars do display what we have come to expect from them with regards to risk for example Volvo is not to be one of the safest cars. Porsche could be deemed 'Very High Risk' due to the nature of it being a sports/ performance car and speeding leads to a lot of accidents.



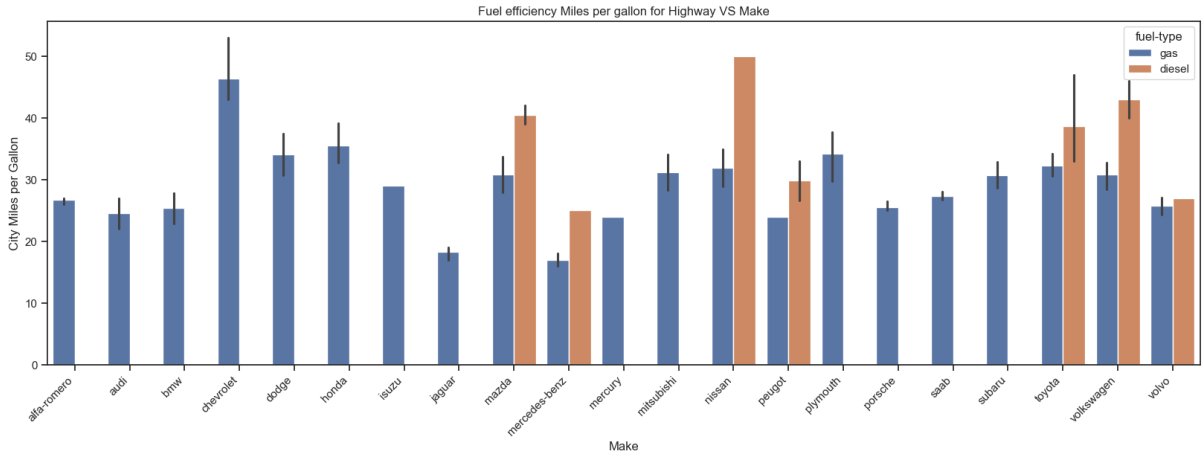
From the graph above it can be noted that the most popular car body style might be sedan but the highest risk body-style car is hatchback. We could deduce that most performance cars tend to be hatchbacks and with that comes speeding and reckless driving leading to more accidents. Convertibles however are also high risk but this could be due to their lack of crash protection in a rollover incident is significantly less than that of a hard-top car.



The most popular horsepower groupings are the lower horsepower groupings, we could assume that this is due to these cars being cheaper than cars of a higher horsepower. There also does not appear to be a clear relationship between horsepower and the risk factor because the grouping with the highest risk factor is 84 - 120 this is the second lowest horsepower grouping. Based on the graph you could assume that higher-risk cars are due to popularity and not horsepower.



In the comparison between fuel efficiency of city driving versus make versus fuel type it can be not from the graph Chevrolet is the most fuel efficient gas make for city driving but diesel and the Nissan brand beats this and can be said to be the most fuel efficient option for city driving.



Just Like the previous graph the same results can be seen to be true about fuel efficiency and highway driving in the Chevrolet would be the best choice should you want gas as your fuel type. Nissan is still the most fuel efficient when choosing diesel as fuel type.

THIS REPORT WAS WRITTEN BY : SHEENA WEBER