|  |
| --- |
| AIML231 |
| Assignment Three |
| 300601546 |

|  |
| --- |
| Shemaiah Rangitaawa  [Date] |

**Part One | Linear Regression**

Exploratory Data Analysis (EDA)

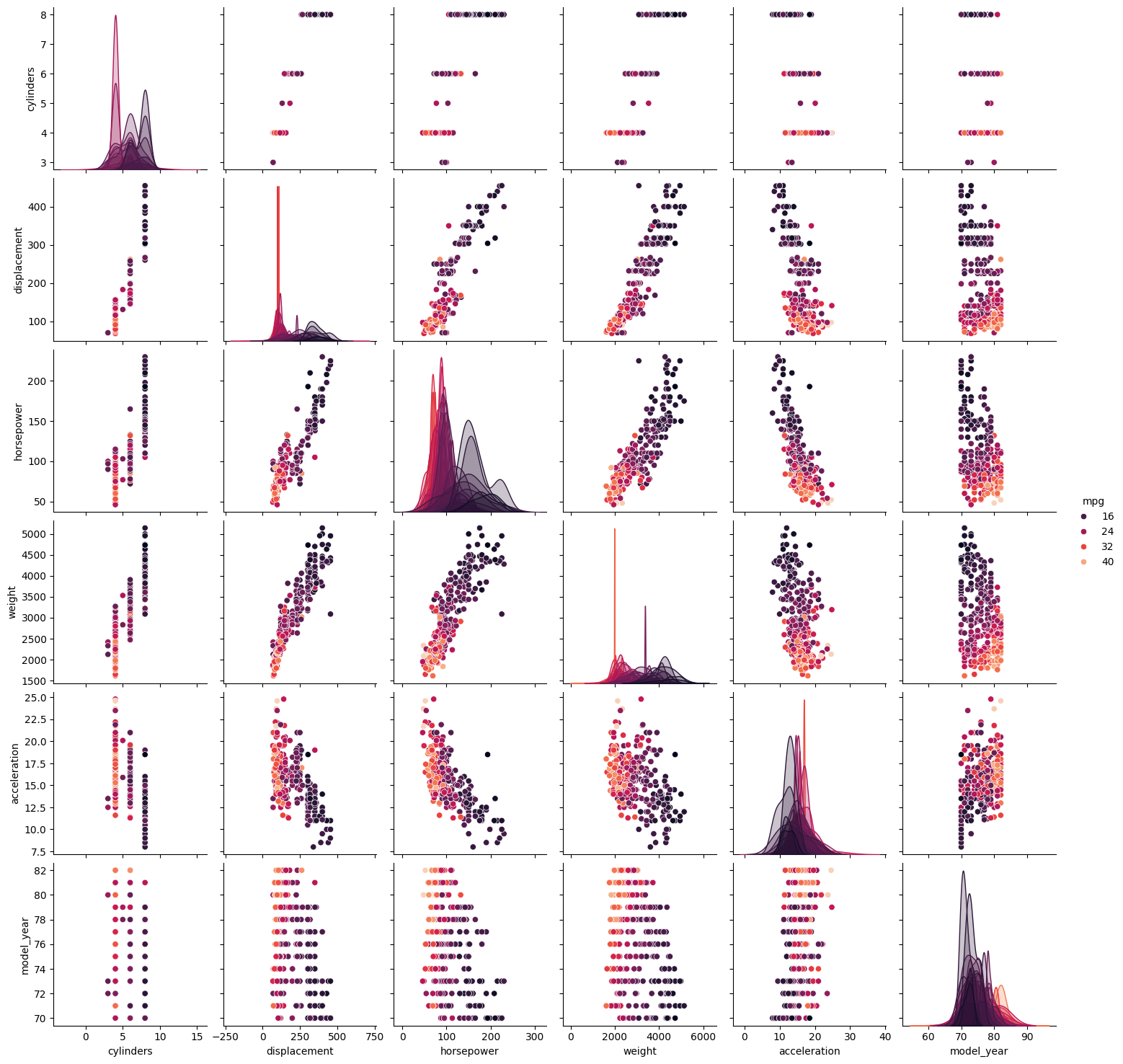


Figure : MPG Train Split Pair Plot

The pair plot of the training split of the MPG data set reveals key insights into how different vehicle attributes relate to fuel efficiency and to each other:

**Weight and MPG:** There is a clear negative correlation between MPG and vehicle weight. As weight increases, MPG decreases, suggesting that heavier cars tend to have lower fuel efficiency.

**MPG and Displacement:** Similarly, there's a strong negative correlation between MPG and engine displacement. Larger engine sizes are associated with lower MPG.

**MPG and Horsepower**: The scatter plot shows a negative correlation between MPG and horsepower. Vehicles with higher horsepower generally have lower MPG.

**Model Year and MPG**: There is a trend indicating that newer models tend to have better MPG, showing an improvement in fuel efficiency over time.

A group of blue and white bars

Description automatically generated

Figure : MPG Histograms

Analyzing the histogram distributions, we find several informative features. Miles per gallon (MPG) is notably right skewed, highlighting a significant number of vehicles with high fuel efficiency, which are likely newer or more technologically advanced models. This skewness in MPG is an essential indicator for identifying factors that contribute to enhanced fuel efficiency.

In terms of engine characteristics, the data shows a prevalence of vehicles equipped with 4-cylinder engines, which are typically associated with lower displacement and horsepower. This commonality suggests a standardization around engines that optimize fuel efficiency, as fewer cylinders generally correlate with lower displacement and horsepower, thereby improving MPG. Additionally, the relationship between displacement, horsepower, and the number of cylinders indicates a strong correlation that is pivotal for modeling the impact of engine specifications on fuel efficiency.

The vehicle weight distribution exhibits a moderate right skew, with fewer instances of very heavy vehicles. This pattern suggests that heavier vehicles often have lower MPG, underscoring the inverse relationship between vehicle weight and fuel efficiency. Acceleration, on the other hand, shows a normal distribution across the dataset, suggesting that this attribute varies less between vehicles and may not be as critical to fuel efficiency as engine size or weight.

Moreover, the gradual increase in the number of vehicle models from the early 1970s to the early 1980s likely reflects advancements in automotive technology and changes in market availability. This trend could be associated with improvements in fuel efficiency and emission standards, which are significant for understanding historical advancements in vehicle technology.

Preprocessing Methodology

I chose to utilize the Median strategy for imputing missing values in the training and test MPG datasets.

Median imputation is an effective method for handling missing values in the horsepower attribute of automotive datasets, offering several distinct advantages. Firstly, the median is robust to outliers and skewed distributions, making it a more representative measure of central tendency than the average.

Categorical variables are Cylinders (e.g. 4, 6, 8) represents distinct categories of engine types, Model Year as each year represents a separate group of models with potentially different characteristics, Origin the region or country where the vehicle was manufactured, and Name of the car model, is also categorical as it represents different categories of car models.

Cylinders will be encoded using the Ordinal method, there is an inherent order (4 cylinders, 6 cylinders, etc., where the number of cylinders usually relates to engine power).

Model year and Origin will use One-hot encoding. One-hot encoding transforms each unique year and origin into a separate feature, allowing models to interpret each features impact independently.