**PCA**

Principal Component Analysis to the Toyota Corolla dataset in order to explore the dimensionality reduction techniques for structured data.

Goal was to identify redundant or highly correlated variables, convert categorical features into appropriate numeric representations, and ultimately reduce the feature space while retaining the majority of the dataset’s variance.

Steps included:

Dummy variable creation

Correlation Analysis

Standardization of Inputs

Applying PCA

Objective was to determine how many principal components are sufficient to explain a high percentage of the total variance in the data.

1. Identify the categorical variables.
2. In the Toyota corolla dataset, the following variables are identified as categorical based on their data types and unique value counts:
   1. Fuel Type: Indicated the type of fuel used by the car. It has 3 distinct categories.
   2. Color: Represents the specific color of the vehicle, It contains 10 unique categories.
   3. Model: Represents the specific model name or ID of each car. This variable has 319 unique values and is considered high cardinality.
3. Explain the relationship between a categorical variable and the series of binary dummy variables derived from it.
4. When a categorical variable is converted into a numerical format suitable for statistical analysis or modeling, it is typically transformed into a series of binary (dummy) variables. Each dummy variable represents one category of the original variable and takes the value.
   1. 1 if the observation belongs to the category.
   2. 0 otherwise.
5. For example, if the Fuel\_Type variable contains the categories Petrol, Diesel, and CNG, it can be represented by two dummy variables:
   1. Fuel\_Type\_Diesel
   2. Fuel\_Type\_Petrol

Here, the absence of both the dummy variables (0, 0) would imply the car uses CNG, the dropped baseline category.

1. How many dummy binary variables are required to capture the information in a categorical variable with N categories?
2. For a categorical variable with N distinct categories, you need (N-1) dummy binary variables to fully represent the information.

This is because one of the categories is treated as the reference or baseline, and its presence is implied when all the other dummy variables are 0.

Dropping one category avoids perfect multicollinearity (also known as dummy variable trap) in models like linear regression and PCA.

Example:

If a variable Color has 4 categories:

Red, Blue, Black, White

We would create 3 dummy variables:

Color\_Blue

Color\_Black

Color\_White

If all three are 0, the color is assumed to be the baseline category, in this case – Red.

Hence, (N-1) dummy variables are sufficient and statistically recommended.

1. Use R to convert the categorical variables in this dataset into dummy variables, and explain in words, for one record, the values in the derived binary dummies.
2. In R programming, categorical variables can be converted into binary (dummy) variables using functions such as:

Model.matrix( ~ Fuel\_Type + Color – 1, data = Toyota\_data)

Library(fastDummies)

Dummy\_data 🡨 dummy\_cols(Toyota\_data, select\_columns = c(“Fuel\_Type”, “Color”), remove\_first\_dummy = TRUE)

The remove\_first\_dummy = TRUE option avoids the dummy variable trap by dropping one category per variable to serve as a reference.

Since the analysis was performed in Python, the equivalent process was done using pandas.get\_dummies() with drop\_first=True, which serves the same statistical purpose.

1. Use R to produce a correlation matrix and matrix plot. Comment on the relationships among variables.
2. The correlation matrix was computed in R using the cor() function, and the matrix plot was generated using the corrplot package with color gradients to visualize the strength and direction of relationships among numeric variables.

* Price is negatively correlated with Age and positively correlated with HP, which aligns with the intuition – older cars depreciate, and more powerful cars cost more.
* Age and KM (Mileage) also show a moderate positive correlation, indicating that older cars tend to be more driven.
* Features like Quarterly\_Tax, Weight, and Guarantee\_Period show weaker but noticeable correlations with Price, possibly hinting at value-add features or newness.
* Most binary accessory features (e.g., Tow\_bar, CD\_Player, Sport\_model) show very weak correlations, which is expected due to their binary nature.

This analysis has confirmed the presence of multicollinearity and redundant information across several numeric variables.

A diagram of a number of words

AI-generated content may be incorrect.