Dimensions.Length,Height,Width :These describe the physical

dimensions of a car

Engine Infromation.Driveline: Describes the type of driveline as All,Front

and Rear wheel drives.

Engine Information about the engine type

Engine Information. Transmission : Gear Transmission of the car can be

identified using this variable

Engine Tells us if the car is hybrid or not True or

False values are given

Engine Information.Number of Forward Gears

Engine Statistics.HorsePower :Engine Horse Power

Value

Engine Engine's torque

Fuel mpg&jgtyæy mpg : Gives value of the mileage of a

car in City and Highway terrains

Fuel of fuel used in the car

Identification.Classification: Used for classification on gear transmisson.

Same as of the Engine Transmission

Identification.Model Year Year ID:used to identify the car based on those

Parameters

9) Linear Regression only works with numerical values, so we need to convert categorical values into

numerical form. We need to encode those categorical values into numbers.We are gonna be doing this

at the first since we have to find the correlation between each feature and the value that have to predict.

10) No i did not use feature scaling in this we normally use feature scaling in

distance based algorithms like KNN,SVMs or PCA s. And Linear Regression model doesn't strictly

require feature scaling because here we are not using distance calculations.Hence it is not necessary

but it could improve model stability.  
  
11) I applied correlation analysis to to select the best features that would ensure the accurate prediction of

the Fuel oopg,J also used my domain knowledge to select the features mainly

used :the larger the vehicle the lesser fuel of forward

gears: the more the forward gear the better ep!jguise4 power ggljye.D/eBlse i used the numerical values

like and Torque since the more the value of them ,the engine gets more powerful and causes

high fuel consumption]

The dimension is 18. Dimension is the number of distinct features or attributes in the data

Rows – 5076

5,076 instances

9 numerical variables

Dimensions.Height

Dimensions.Length

Dimensions.Width

Engine Information.Number of Forward Gears

Fuel Information.City mpg

Fuel Information.Highway mpg

Identification.Year

Engine Information.Engine Statistics.Horsepower

Engine Information.Engine Statistics.Torque

9 categorical variables (Including Boolean):

Engine Information.Driveline

Engine Information.Engine Type

Engine Information.Transmission

Fuel Information.Fuel Type

Identification.Classification

Identification.ID

Identification.Make

Identification.Model Year

Engine Information.Hybrid (Boolean)

Dimensions.Length,Height,Width :These describe the physical

dimensions of a car

Engine Infromation.Driveline: Describes the type of driveline as All,Front

and Rear wheel drives.

Engine Information about the engine type

Engine Information. Transmission : Gear Transmission of the car can be

identified using this variable

Engine Tells us if the car is hybrid or not True or

False values are given

Engine Information.Number of Forward Gears

Engine Statistics.HorsePower :Engine Horse Power

Value

Engine Statistics. torque

Fuel Information.City,Highway mpg : Gives value of the mileage of a

car in City and Highway terrains

Identification.Classification: Used for classification on gear transmisson.

Same as of the Engine Transmission

Identification.Model Year, Year, ID:used to identify the car based on those

Parameters

The goal of this study is to predict a car’s fuel efficiency (City MPG or Highway MPG) using several car features that include engine type, horsepower and weight in addition to transmission type. By analyzing the relationships between all of these factors and fuel efficiency, we aim to create a predictive model, so we can understand which features each greatly effect a car’s mileage.

Fuel costs go down. This is because of fuel-efficient cars. Consumers save money with this action.

Effectively comparing multiple car models will be easier for buyers using a fuel efficiency prediction model.

- Car manufacturers can plan to design a car that is cost effective and more efficient.

- The study can help engineers learn which car features affect fuel consumption most and improve future designs.

My study helps:

It assists every car buyer to choose a car that has greater mileage.

For manufacturers: Guides to help design cars that use less fuel.

For Policymakers: Supports regulations that reduce fuel consumption and emissions.

In short: This study helps develop one data-driven solution for cars. Cars can be more fuel-efficient, ecologically friendly and cost-effective.

12) Outliers can significantly impact your model by distorting relationships between variables and increasing prediction errors. They can be removed or transformed to improve model accuracy. Before handling outliers, the model had a higher RMSE: 2.8759 on training and 2.9782 on testing. After removing the outliers, the RMSE decreased to 2.6830 for training and 2.7137 for testing, showing better predictive performance and stability.

- The R² score improved by approximately 4 percentage points, and the model showed less overfitting with a smaller gap between train and test scores.

- Removing outliers led to a more accurate and reliable model overall.

https://shorturl.at/HUbRD (Comparison Image)

13) Before applying Multiple Linear Regression, I ensured that key assumptions were met for proper results. I checked for linearity by using a correlation matrix and scatter plots to verify that there was a linear relationship between the independent variables and the target (Fuel Information.City mpg). we know the linear equation is y=mx+b ,where the y is the dependant variable and x is the dependant variable. I also checked for multicollinearity using the Variance Inflation Factor (VIF), ensuring that independent variables weren’t highly correlated. If VIF was greater than 5, I would remove or combine correlated features. By meeting these assumptions, I ensured the model would be more reliable.

14) I used the Variance Inflation Factor (VIF) to evaluate model performance and detect multicollinearity, which occurs when independent variables are highly correlated. High multicollinearity can lead to unstable coefficients and reduce model accuracy. A VIF greater than 5 indicates that a feature is highly correlated and should be reconsidered. To evaluate the model,

- I first checked for multicollinearity using VIF,

-Then computed the Mean Squared Error (MSE) after training the model.

- Finally, I compared the performance before and after handling outliers and multicollinearity.

Reducing multicollinearity led to a lower MSE, more stable regression coefficients, and better feature selection.

15) This project predicted city fuel efficiency (mpg) using multiple linear regression. The main points are:

I chose several relevant numerical features using correlation analysis. My knowledge of the subject let me improve this selection (for example, "larger vehicles = lower mpg," "more gears = better efficiency"). Categorical variables were handled using one-hot encoding.

The model is simpler to understand and has a reduced number of unnecessary features.

I used the IQR method to detect and remove extreme values because both horsepower and torque particularly affected the regression coefficients.

MSE was importantly lowered and prediction accuracy was considerably improved, by removing outliers.

I checked for multicollinearity. I used the Variance Inflation Factor (VIF) for all of the checks. All features with a VIF higher than 5 were removed to improve stability.

Effect: Redundancy was considerably reduced and unstable regression coefficients were completely prevented.

Before handling outliers and multicollinearity, the original MSE was 8.05 and large multicollinearity seriously affected accuracy. Because the data was cleaned, the MSE dropped to 5.76, predictions became more accurate and coefficients became more stable.

Prediction reliability was noticeably improved by large data cleaning.