FINAL PROJECT PROPOSAL

GROUP 8

Car price Prediction

1. Introduction:

In the automotive sector, pricing analytics play an essential role for both companies and individuals to assess the market price of a vehicle before putting it on sale or buying it. There are many automobile companies aspires to enter the US market by setting up their manufacturing unit in US. Predicting car prices involves determining a vehicle's market worth based on various attributes like brand, model, year of manufacture, mileage, and overall state. This prediction holds significant value for the auto industry, aiding both prospective buyers and sellers in understanding pricing and making educated decisions regarding vehicle transactions.

Objective:

This project aims to predict car prices based on various car features in the dataset, exploring relationships between car specifications and their market prices. The attributes used in the dataset provide valuable insights into the factors influencing car prices which help the management to understand how exactly the prices vary with the independent variables and can be used to develop predictive models for estimating the selling price of cars. They can accordingly manipulate the design of the cars, the business strategy etc. to meet certain price levels.

Research Questions:

- Which variables are significant in predicting the price of a car?
- What features significantly impact the price of a car?
- Can car price differ with number of cylinders, engine type, car body?
- Can car body be dependent on number of cylinders in the car?

Motivation:

Understanding price determinants can aid manufacturers in setting prices and help consumers make informed purchasing decisions.

Hypothesis:

We hypothesize that factors such as Higher engine size, horsepower, and specific body types will positively impact car prices. Features like fuel economy may negatively correlate with price due to a trade-off with performance.

2. Analysis Plan

Dataset description:
Dataset is sourced from Kaggle and comprises the information which is gathered from other automobile web sources.
The data consists of 205 entries with 26 attributes, representing various car features and their prices. The source of the dataset has not been provided, but the dataset appears to be preprocessed for analysis.

Variables:

- Observed Variable: Price of the car (target variable to be predicted).
- Predictor Variables:

  Quantitative: Various car specifications including wheelbase, carlength, carwidth, carheight, curbweight, enginesize, horsepower, citympg, highwaympg.

  Qualitative: fueltype, aspiration, doornumber, carbody, drivewheel, enginelocation etc.

Potential Models:

EDA: Identifying relationships between price and numerical features.

Hypothesis testing: T-test, ANOVA, Chi-square tests.

Predictive modeling:

Linear Regression- For baseline analysis to examine how individual variables impact price.

Multiple Linear Regression- For multivariate relationships between price and predictors, to handle multicollinearity.

Polynomial Regression- For fitting a regression model when the relationship between the predictor variables and the response variable is nonlinear.

Model Evaluation:

Coefficient of determination R^2, MSE, AIC metrics.

Model selection:

Forward selection and Backward elimination stepwise procedures will be performed.

Assumptions:

We consider few assumptions for the analysis.

- Assumption 1- Linearity: There is a Linear Relationship Between independent variables ‘X’ and dependent variables ‘Y’.

  Verification: Plots predicting variables against observed variable.

- Assumption 2- Multicollinearity: Minimum collinearity amongst the independent variable’s ‘X’.

  Verification: Verifying VIF (variation inflation factor).

- Assumption 3- Residuals: Error terms are Normally distributed and there are no Patterns among them.

  Verification: Durbin-Watson test or residual plots.

- Assumption 4: Homoscedasticity -Variance around the regression line is same for all independent variable’s ‘X’.

  Verification: Residual vs. fitted values plot or Breusch-Pagan test

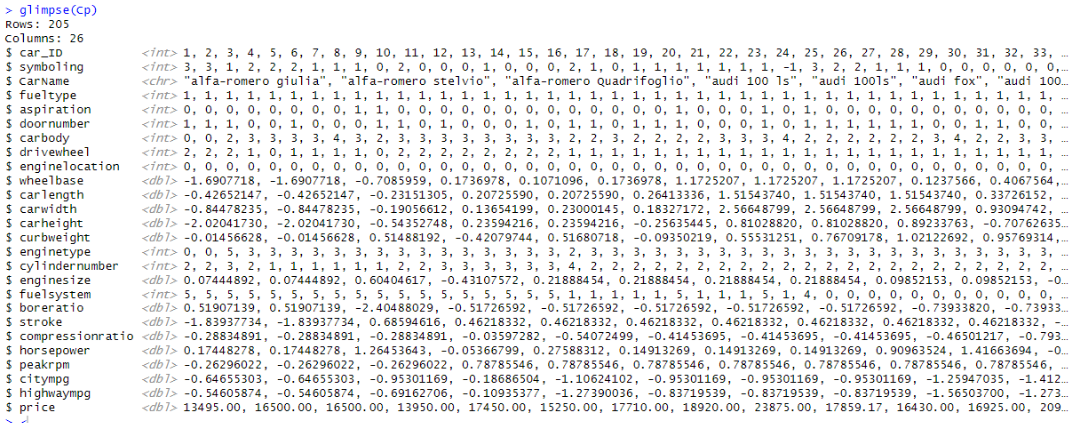
3. Data.

Dimensions: 205 observations and 26 variables.

Data Dictionary:

- car_ID: Unique identifier for each car.
- symboling: A risk factor assigned to the car (-3 to +3), where higher values indicate more risky cars.
- carName: Name of the car (brand and model).
- fueltype: Type of fuel (e.g., gas or diesel).
- aspiration: Type of aspiration (e.g., standard or turbocharged).
- doornumber: Number of doors. (e.g., two-door, four-door).
- carbody: Type of car body (e.g., sedan, hatchback).
- drivewheel: Type of drive wheel (e.g., front-wheel drive).
- enginelocation: Location of the engine. (e.g., front, rear).

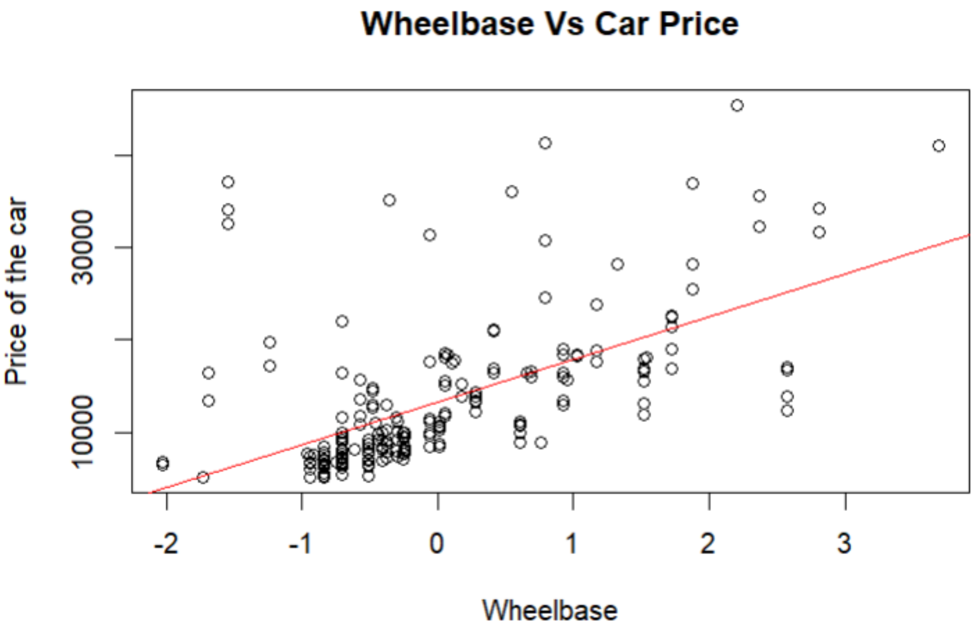
- wheelbase: The distance between the front and rear axles of the car, typically measured in inches or millimeters.
- carlength, carwidth, carheight : Length, width, height of the car.
- curbweight: combines weight of the car
- enginetype: Type of the engine (e.g., Turbo charged, Turbo engine, Buick, Internal combustion engine, Flat engine, V8 engine SBC , rotary)
- cynlindernumber: Number of cylinders.
- enginesize: size of the engine in cubic cm.
- fuelsystem: system that delivers the proper amount of fuel to the engine.
- boreratio: determines engine’s power & torque characteristics.
- stroke: movement of piston in the cylinder.
- compressionratio: minimum to maximum cylinder volume.
- horsepower: measurement of engine power.
- peakrpm: speak of the car in revolution per minute.
- citympg: mileage in cities
- highwaympg: mileage on highway roads.
- price: price of the car.



```
> glimpse(Cp)
Rows: 205
Columns: 26
$ car_ID           <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, …
$ symboling        <int> 3, 3, 1, 2, 2, 2, 1, 1, 1, 0, 2, 0, 0, 0, 1, 0, 0, 0, 2, 1, 0, 1, 1, 1, 1, 1, 1, 1, -1, 3, 2, 2, 1, 1, 1, 0, 0, 0, 0, 0, 0,…
$ CarName          <chr> "alfa-romero giulia", "alfa-romero stelvio", "alfa-romero Quadrifoglio", "audi 100 ls", "audi 100ls", "audi fox", "audi 100…
$ fueltype         <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, …
$ aspiration       <int> 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 1, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, …
$ doornumber       <int> 1, 1, 1, 0, 0, 1, 0, 0, 0, 1, 1, 0, 1, 0, 0, 0, 1, 0, 1, 1, 0, 1, 1, 1, 0, 0, 0, 1, 0, 1, 1, 1, 1, 1, 1, 0, 0, 1, 1, 0, 0, …
$ carbody          <int> 0, 0, 2, 3, 3, 3, 3, 4, 3, 2, 3, 3, 3, 3, 3, 3, 3, 3, 2, 2, 3, 2, 2, 2, 2, 3, 3, 3, 4, 2, 2, 2, 2, 2, 2, 3, 4, 2, 2, 3, 3, …
$ drivewheel       <int> 2, 2, 2, 1, 0, 1, 1, 1, 1, 0, 2, 2, 2, 2, 2, 2, 2, 2, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, …
$ enginelocation   <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, …
$ wheelbase        <dbl> -1.6907718, -1.6907718, -0.7085959, 0.1736978, 0.1071096, 0.1736978, 1.1725207, 1.1725207, 1.1725207, 0.1237566, 0.4067564,…
$ carlength        <dbl> -0.42652147, -0.42652147, -0.23151305, 0.20725590, 0.20725590, 0.26413336, 1.51543740, 1.51543740, 1.51543740, 0.33726152, …
$ carwidth         <dbl> -0.84478235, -0.84478235, -0.19056612, 0.13654199, 0.23000145, 0.18327172, 2.56648799, 2.56648799, 2.56648799, 0.93094742, …
$ carheight        <dbl> -2.02041730, -2.02041730, -0.54352748, 0.23594216, 0.23594216, -0.25635445, 0.81028820, 0.81028820, 0.89233763, -0.70762635…
$ curbweight       <dbl> -0.01456628, -0.01456628, 0.51488192, -0.42079744, 0.51680718, -0.09350219, 0.55531251, 0.76709178, 1.02122692, 0.95769314,…
$ enginetype       <int> 0, 0, 5, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 2, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, …
$ cylindernumber   <int> 2, 2, 3, 2, 1, 1, 1, 1, 1, 1, 2, 2, 3, 3, 3, 3, 3, 3, 4, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, …
$ enginesize       <dbl> 0.07444892, 0.07444892, 0.60404617, -0.43107572, 0.21888454, 0.21888454, 0.21888454, 0.21888454, 0.09852153, 0.09852153, -0…
$ fuelsystem       <int> 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 1, 1, 1, 1, 1, 5, 1, 1, 1, 5, 1, 4, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, …
$ boreratio        <dbl> 0.51907139, 0.51907139, -2.40488029, -0.51726592, -0.51726592, -0.51726592, -0.51726592, -0.51726592, -0.73933820, -0.73933…
$ stroke           <dbl> -1.83937734, -1.83937734, 0.68594616, 0.46218332, 0.46218332, 0.46218332, 0.46218332, 0.46218332, 0.46218332, 0.46218332, -…
$ compressionratio <dbl> -0.28834891, -0.28834891, -0.28834891, -0.03597282, -0.54072499, -0.41453695, -0.41453695, -0.41453695, -0.46501217, -0.793…
$ horsepower       <dbl> 0.17448278, 0.17448278, 1.26453643, -0.05366799, 0.27588312, 0.14913269, 0.14913269, 0.14913269, 0.90963524, 1.41663694, -0…
$ peakrpm          <dbl> -0.26296022, -0.26296022, -0.26296022, 0.78785546, 0.78785546, 0.78785546, 0.78785546, 0.78785546, 0.78785546, 0.78785546, …
$ citympg          <dbl> -0.64655303, -0.64655303, -0.95301169, -0.18686504, -1.10624102, -0.95301169, -0.95301169, -0.95301169, -1.25947035, -1.412…
$ highwaympg       <dbl> -0.54605874, -0.54605874, -0.69162706, -0.10935377, -1.27390036, -0.83719539, -0.83719539, -0.83719539, -1.56503700, -1.273…
$ price            <dbl> 13495.00, 16500.00, 16500.00, 13950.00, 17450.00, 15250.00, 17710.00, 18920.00, 23875.00, 17859.17, 16430.00, 16925.00, 209…
>
```

Initial Visualizations:

To visualize the dataset's relationships, created scatter plots and distribution plots of the key predictors against price.



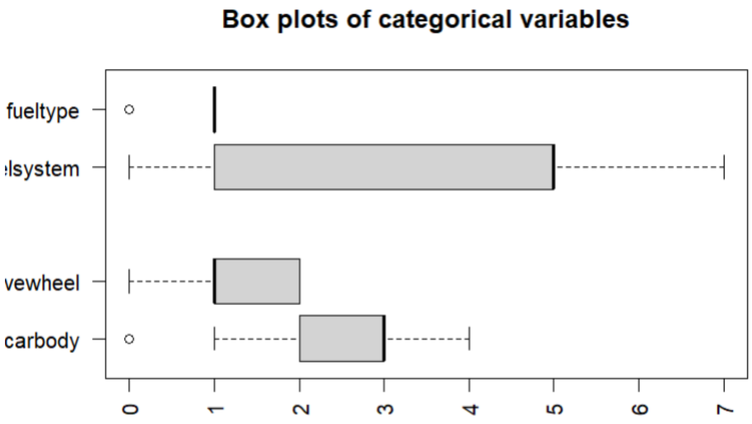


Fig: Boxplots of categorical variables

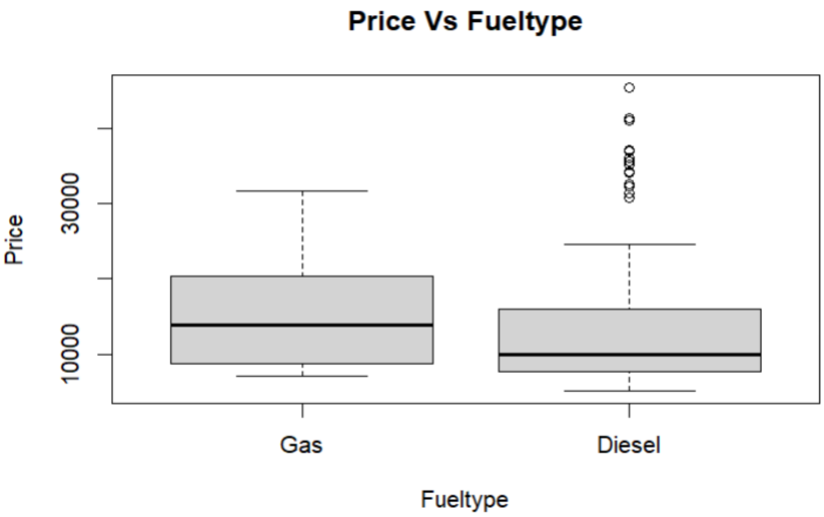


Fig: Box plot of categorical variable against price

Few categorical variables (carbody, drivewheel, fuelsystem, fueltype) are considered for comparison.

Numerical variables against price(y) scatter plot to check for collinearity.

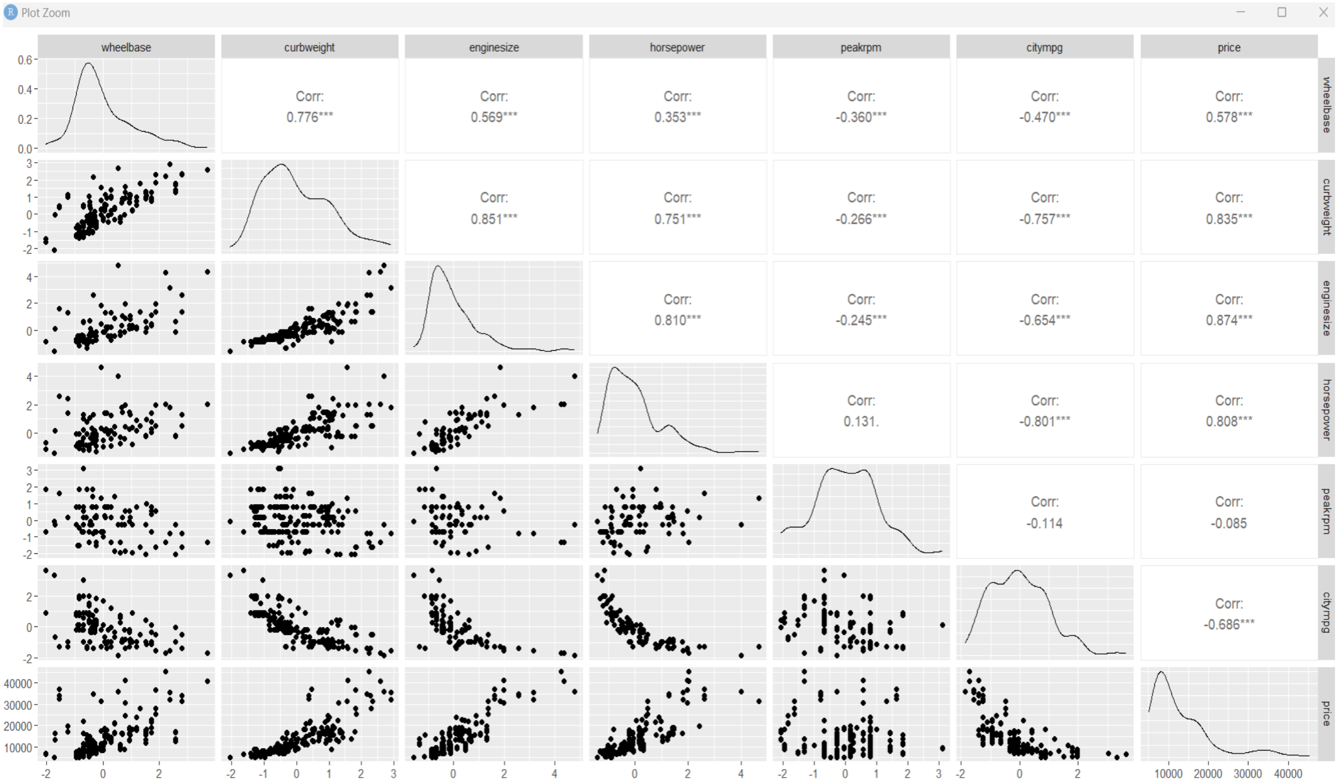


Fig: Scatter plot of few selected predictor variables with observed variable.

Correlation of target variable “Price” with independent variables:

- Price is highly (positively) correlated with wheelbase, curbweight, enginesize, horsepower, mainly enginesize(0.87) (these variables represent the size/weight/engine power of the car)
- Price is negatively correlated to ‘peakrpm’, ‘citympg’. This suggest that cars having high mileage may fall in the ‘economy’ cars category, and are priced lower.

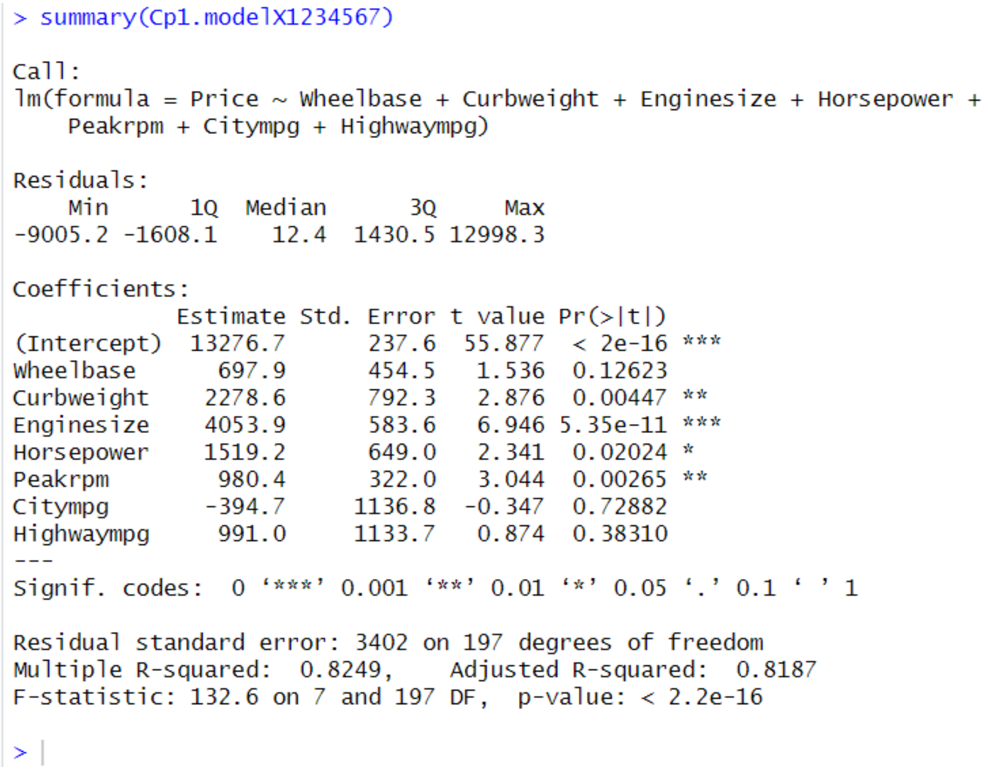
We can see that most of the X are correlated with Y i.e They have a Linear Relationship, so we can consider this Assumption as Satisfied.

Model Building:

If data cleaning is done then we will start with data model building. Creating dummy variables for categorical variables. Scaling the features and getting the final list of columns in data frame for model building.

To identify and retain the most relevant predictors for feature selection we will try Stepwise selection (Forward, Backward or both).

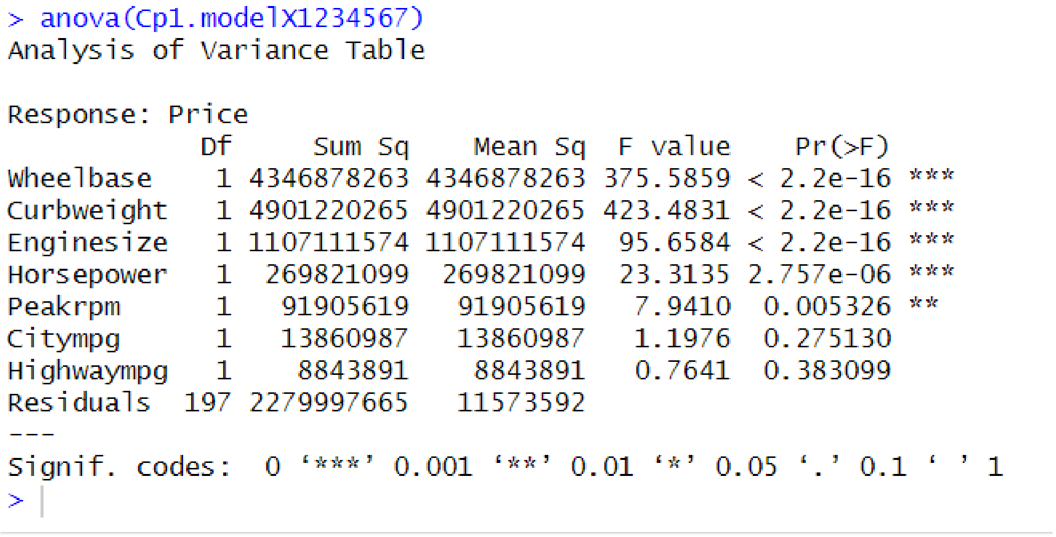
Forward selection: We will add predictors iteratively based on significance p-value < 0.05.
Backward selection: We will start with all predictor variables and remove the least significant ones.
During model building we will check if the error terms are normally distributed. To handle any assumptions violations for data transformation logarithmic, square root will be used for non-linear relationships.



```
> summary(Cp1.modelX1234567)

Call:
lm(formula = Price ~ Wheelbase + Curbweight + Enginesize + Horsepower +
    Peakrpm + Citympg + Highwaympg)

Residuals:
    Min      1Q  Median      3Q     Max
-9005.2 -1608.1    12.4  1430.5 12998.3

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  13276.7      237.6  55.877  < 2e-16 ***
Wheelbase      697.9      454.5   1.536  0.12623
Curbweight    2278.6      792.3   2.876  0.00447 **
Enginesize    4053.9      583.6   6.946 5.35e-11 ***
Horsepower    1519.2      649.0   2.341  0.02024 *
Peakrpm        980.4      322.0   3.044  0.00265 **
Citympg       -394.7     1136.8  -0.347  0.72882
Highwaympg     991.0     1133.7   0.874  0.38310
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3402 on 197 degrees of freedom
Multiple R-squared:  0.8249,	Adjusted R-squared:  0.8187
F-statistic: 132.6 on 7 and 197 DF,  p-value: < 2.2e-16

>
```

Fig: model building with all predictor variables against price.

By above figure Coefficient of multiple determination R^2 coming around 0.825 approx. with selected variables (wheelbase, curbweight, enginesize, horsepower, citympg) which is not bad.



```
> anova(Cp1.modelX1234567)
Analysis of Variance Table

Response: Price
            Df     Sum Sq    Mean Sq  F value    Pr(>F)
Wheelbase    1 4346878263 4346878263 375.5859 < 2.2e-16 ***
Curbweight   1 4901220265 4901220265 423.4831 < 2.2e-16 ***
Enginesize   1 1107111574 1107111574  95.6584 < 2.2e-16 ***
Horsepower   1  269821099  269821099  23.3135 2.757e-06 ***
Peakrpm      1   91905619   91905619   7.9410  0.005326 **
Citympg      1   13860987   13860987   1.1976  0.275130
Highwaympg   1    8843891    8843891   0.7641  0.383099
Residuals  197 2279997665   11573592
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
>
```

Fig: ANOVA table for all numeric predictor variables

Conclusion:

This project proposal outlines an approach to understanding the key factors influencing car prices. Through the analysis of a dataset containing car attributes and prices, we aim to determine which variables have the most significant impact on price and identify any trends or patterns.

Our initial visualizations suggest that attributes such as curbweight, enginesize, and horsepower are likely to correlate positively with price, whereas other factors, like fuel efficiency (citympg), may have a negative or minimal effect. From scatter plot describes that there is a non-linear(curve) relationships like price (dependent variable) vs citympg (independent variable) this might result in constructing polynomial regression model to explain nonlinear relationship.

By validating the hypothesis through these models, we expect to develop a predictive model that accurately estimates car prices based on specific attributes.

4. References

- https://www.kaggle.com/
- https://archive.ics.uci.edu/