Logo

Description automatically generated

**College of Professional Studies**

**Northeastern University San Jose**

**MPS Analytics**

**Course: ALY6010 - Probability Theory and Introductory Statistics**

**Assignment:**

MODULE 6 PRACTICE ASSIGNMENT 6

**Submitted on:**

December 16, 2022

**Submitted to:**  **Submitted by:**

Professor: BEHZAD AHMADI NIKSHITA RANGANATHAN

# **INTRODUCTION**

Regression analysis can be used in a variety of applications, such as predicting sales, forecasting stock prices, and analyzing customer behavior. It is also used to identify trends in data and assess the impact of certain factors on outcomes. Regression analysis is an effective data analytics approach that may be applied to provide conclusions.

Regression analysis for categorical variables is used to predict a categorical response variable based on one or more predictor variables. This type of analysis is useful when the response variable is a categorical variable such as gender, race, or religion.

Dummy variables are variables that take on the values of 0 and 1 to represent two or more categories of a single variable. In statistical analysis, they are used to represent categorical values with many groups. For example, a dummy variable might be used to represent gender, with 0 representing male and 1 representing female.

**Performing regression Analysis:**

1. Define the problem: The first step in doing a regression analysis is to define the problem. What relationship do the two variables have with one another? What is the analysis's objective?

2. Collect data: The next step is to collect data on the two variables. In order to identify trends, this data needs to be gathered over time.

3. Analyze the data: Data needs to be analyzed after it has been gathered. This can be done using a variety of methods such as linear regression, logistic regression, or polynomial regression.

4. Interpret the results: The results of the analysis need to be interpreted. This involves determining the strength of the relationship between the two variables by examining the regression equation's coefficients..

5. Make predictions: Finally, future predictions can be made using regression analysis. This can be achieved by using the regression equation to predict the value of one variable given the value of the other.

**About the dataset:**

# Title of dataset – Price Prediction -Multiple Linear Regression.

The dataset has 205 observations of car details with 26 attributes including the name of the car, fuel type, aspiration, car body type, door numbers, engine location, price, engine size, etc. The dataset contains numerical as well as categorical data (source-[*Kaggle*](https://www.kaggle.com/datasets/erolmasimov/price-prediction-multiple-linear-regression)*).*

**Purpose:**

The purpose of this project is to analyze the factors that influence used car prices, identify the strength of the relationship between the dependent (price) and independent variables, and provide insights and visualizations.

The dataset includes the following features:

|  |  |  |
| --- | --- | --- |
| **No** | **Feature** | **Dictionary** |
| 1 | ID | Unique identifier for the cars |
| 2 | symboling | Car’s insurance risk level |
| 3 | name | Car name |
| 4 | fueltype | Type of fuel (Gas/Diesel) |
| 5 | aspiration | Type of air intake system used to draw air into the engine (Standard/Turbo) |
| 6 | doornumbers | Number of doors (two/four) |
| 7 | carbody | Type of car body (sedan/convertible/wagon/hatchback/hardtop) |
| 8 | drivewheels | Wheels that are driven by the engine |
| 9 | enginelocation | Location of the engine within the car |
| 10 | wheelbase | Distance between the front and rear wheels of a car |
| 11 | carlength | Length of a car from its front to its rear bumper |
| 12 | carwidth | Width of a car |
| 13 | carheight | Height of a car |
| 14 | curbweight | Weight of a car without any passengers or cargo |
| 15 | enginetype | Type of engine in the car |
| 16 | cylindernumber | The number of cylinders in the engine |
| 17 | enginesize | Total volume of the cylinders in the engine |
| 18 | fuelsystem | Type of fuel delivery system used in the car |
| 19 | boreratio | Ratio of the diameter of the cylinder bore to the stroke of the piston |
| 20 | stroke | The distance between an engine's piston's top and bottom |
| 21 | compressionratio | Ratio of the volume of the cylinder when the piston is at the bottom of its stroke to the volume of the cylinder when the piston is at the top of its stroke |
| 22 | horsepower | Power output of an engine |
| 23 | peakrpm | The maximum rotational speed of an engine |
| 24 | citympg | Fuel economy of a vehicle when driven in an urban environment |
| 25 | highwaympg | Fuel economy of a vehicle when driven on a highway |
| 26 | price | Cost of the car |

*Table 1: Features of the Used Car Data Set with their dictionary*

**ANALYSIS & INTERPRETATION**

* Importing the scrap price CSV file

<cars> vector contains information about the specifications of cars.

* Cleaning the dataset
  + N/A Values

There are no NA and null values in this dataset.



**Chart, bar chart

Description automatically generated**

****

**Table

Description automatically generated with low confidence**

**A picture containing logo

Description automatically generated**

**Figure 1-<NA> values**

* Removing duplicate rows

Duplicate rows can be identified in R using the duplicated() functions.

anyDuplicated() function in R returns the index of the first duplicate element if any duplicates are found, and 0 otherwise. There are no duplicate rows here.

**Table

Description automatically generated**

**Figure 2- duplicated() and anyDuplicated()**

* Changing datatypes

The data type of some variables (fueltypes,enginetype,carbody etc) is changed to factor using as.factor().

* Removing the outliers

Outliers can result from measurement errors, extreme numbers, or a subset of the data that differs from the rest of the data.

Data points outside the upper and lower quartiles are considered outliers in boxplots. I have removed the extreme outliers.

Chart, box and whisker chart

Description automatically generated

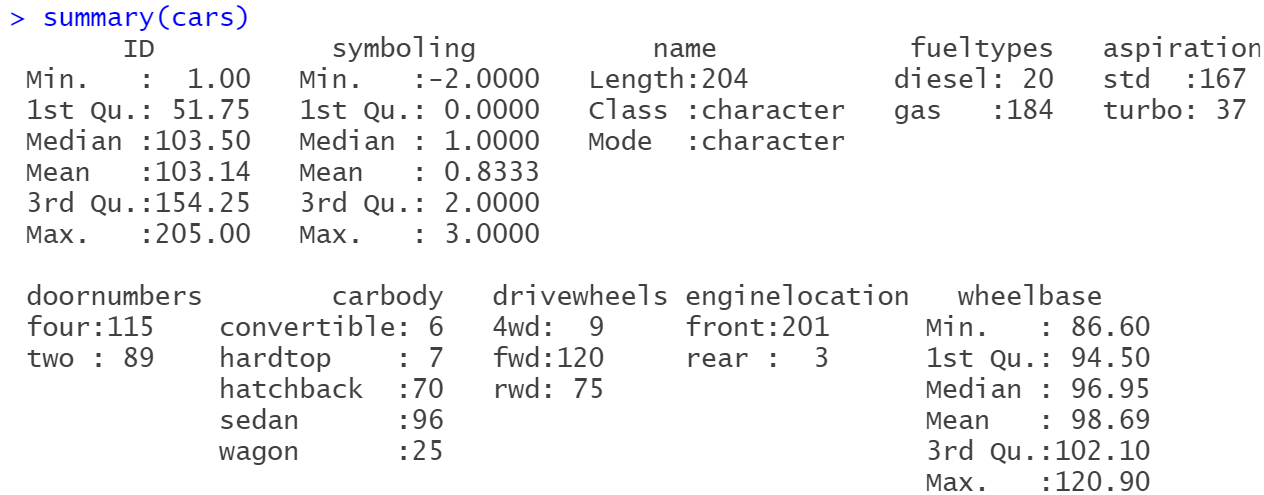
**Figure 3- Outliers**

* Understanding the cars dataset

There are 204 observations and 26 features after data cleaning. The datatypes for the variables are characters, factors, integers, and numerical.

**Text, letter

Description automatically generated**

****

**Text

Description automatically generated**

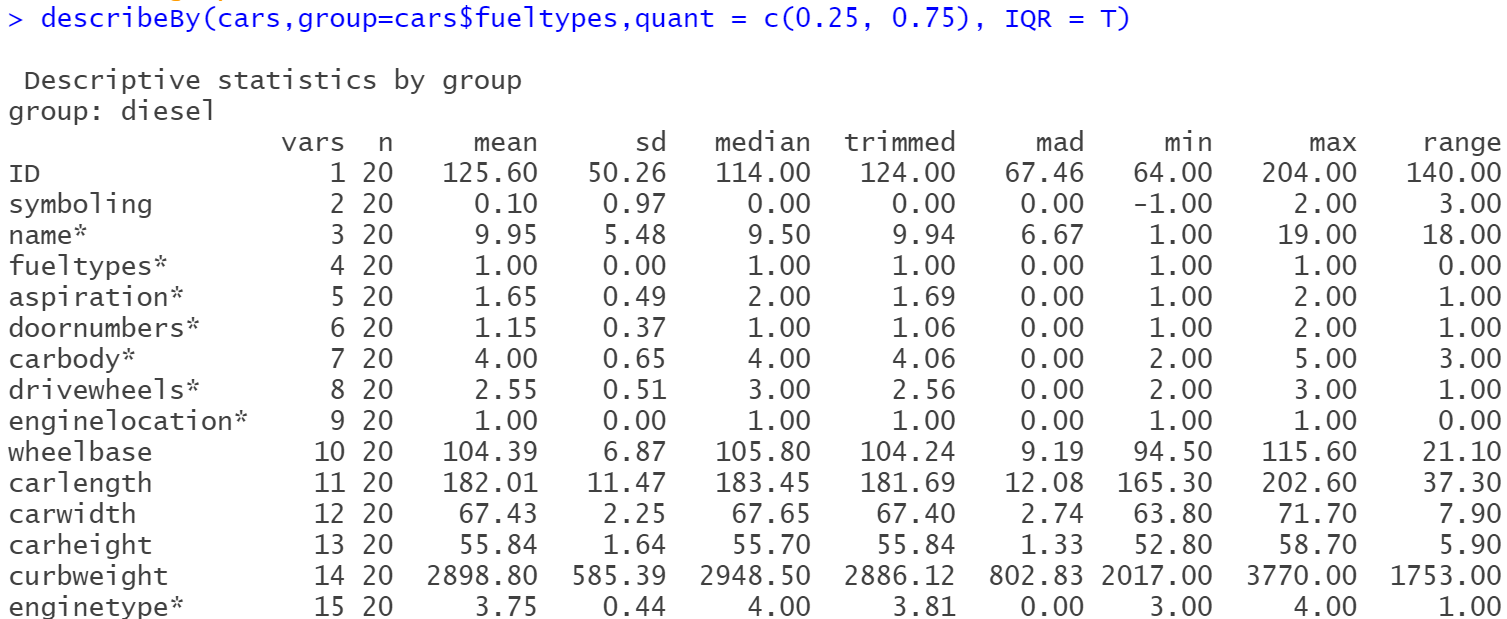
**Figure 4 – str(),summary() and headTail()**

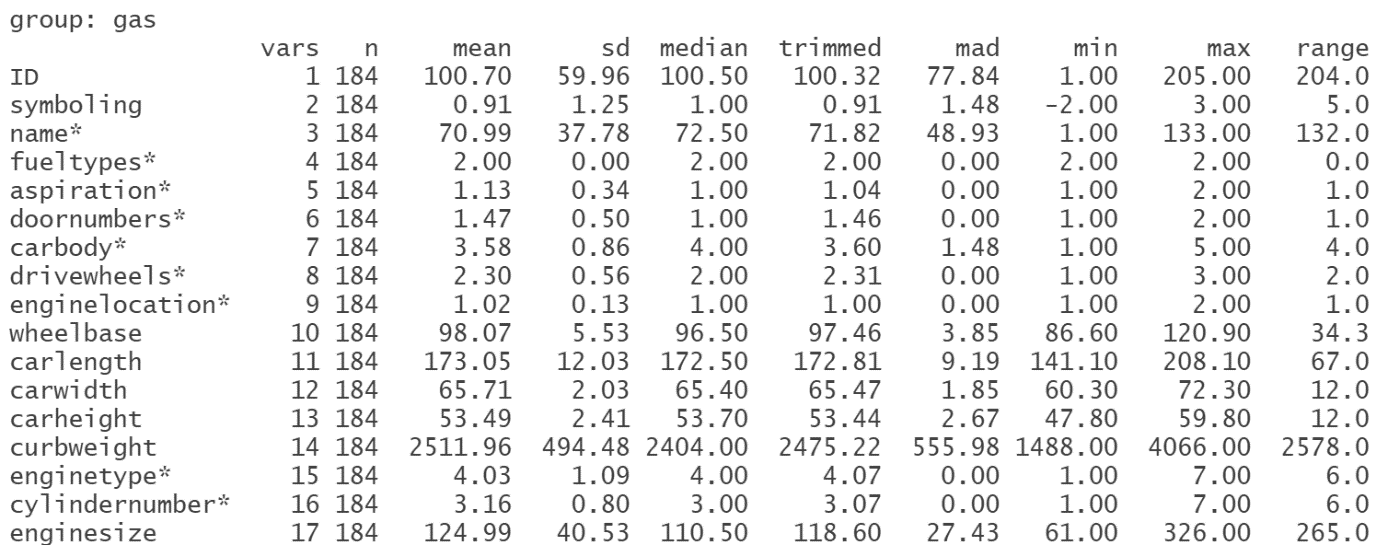
I have used the describeBy() function from the "PSYCH" package to get the descriptive statistics of the features based on groups in the data set. In this case, it is by Fuel type (Gas and Diesel). There are 184 observations in the gas category and 20 observations in the diesel category.

I have included the quartiles value (0.25 and 0.75) and IQR(Interquartile range) in the describeBy function.

The mean price(Diesel category) is approximately 15838.15, with a standard deviation of 7759.84. The minimum and maximum values are 7099 and 31600 respectively.

The mean price(Gas category) is approximately 12823.71, with a standard deviation of 7637.51. The minimum and maximum values are 5118 and 41315 respectively.



****

**Figure 5-describeBy()**

* Correlation

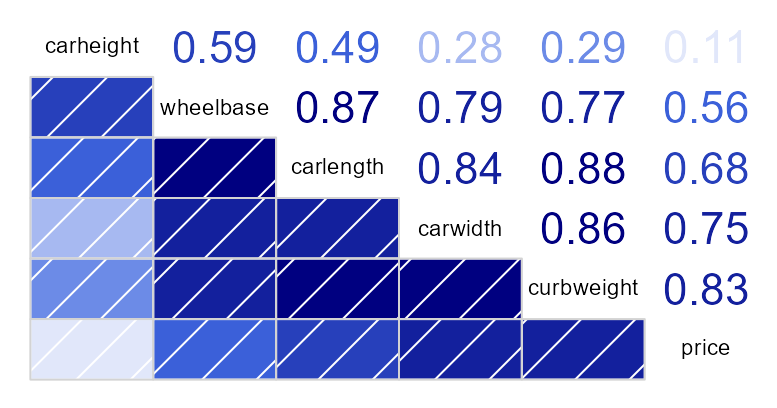
Correlation, a measure of the relationship between the two variables, is used to determine how closely two variables are related to one another. If there is a positive correlation, it means that if one variable rises, the other variable rises as well. A negative correlation means that when one variable rises, the other variable falls.

An example of a graphical display that demonstrates the correlation between various variables is a correlogram. In R, correlograms can be created using the corrplot package. The corrplot package provides a number of functions for creating correlograms, including corrplot(), corrgram(), and corrplot.matrix().

The correlation coefficients:

* -1 to -0.8: Very Strong negative
* -0.79 to -0.6: Strong negative
* -0.59 to -0.4: Moderate negative
* -0.39 to -0.2: Weak negative
* -0.19 to -0.01: Very weak negative
* 0 to 0.19: Very weak positive
* 0.2 to 0.39: Weak positive
* 0.4 to 0.59: Moderate positive
* 0.6 to 0.79: Strong positive
* 1 to 0.8: Very Strong positive

Correlation between price and car dimensions



Shape

Description automatically generated with low confidence

**Figure 6-Correlation Analysis (Price vs Car dimensions)**

Price has a very strong positive relationship with curbweight (Correlation Coefficient – 0.83)

Price has a strong positive relationship with carlength and carwidth (Correlation Coefficient – 0.68 and 0.75)

Price has a moderate positive relationship with wheelbase (Correlation Coefficient – 0.56)

Price has a very weak positive relationship with carheight(Correlation Coefficient – 0.11)

Correlation between price and engine spec

Text

Description automatically generated with low confidence

Diagram

Description automatically generated

**Figure 7 -Correlation Analysis (Price vs Engine Spec A)**

Price has a very strong positive relationship with enginesize (Correlation Coefficient – 0.86)

Price has a moderate positive relationship with boreratio (Correlation Coefficient – 0.54)

Price has a very weak positive relationship with stroke and compression ratio (Correlation Coefficient – 0.08 both)

Letter

Description automatically generated with low confidence

Diagram

Description automatically generated with medium confidence

**Figure 8 -Correlation Analysis (Price vs Engine Spec B)**

Price has a very strong positive relationship with horsepower (Correlation Coefficient – 0.81)

Price has a strong negative relationship with citympg and highwaympg (Correlation Coefficient – -0.68 and -0.69)

Price has a very weak negative relationship with stroke and compression ratio (Correlation Coefficient – -0.06)

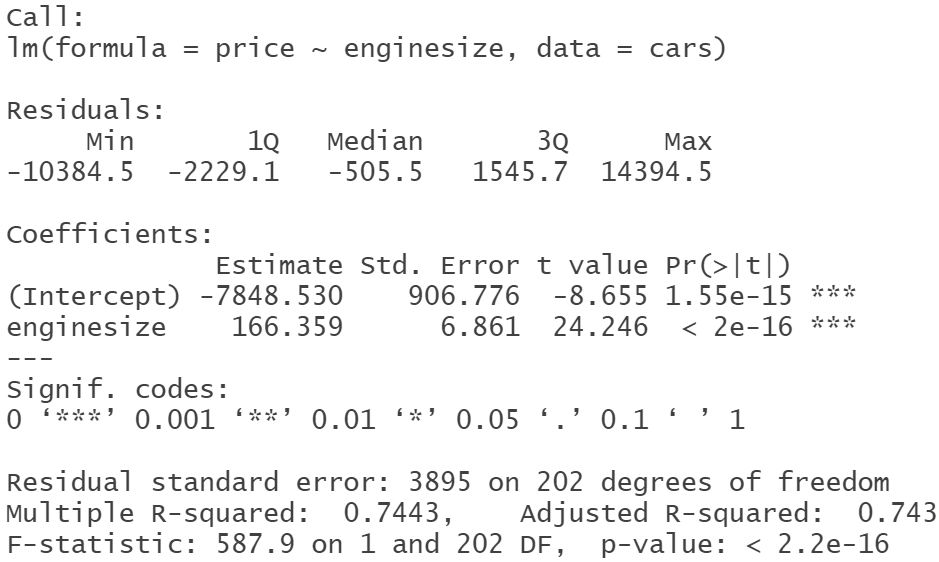
* Regression Models

In a regression model, a continuous outcome variable (or dependent variable) is predicted based on one or more predictor factors (or independent variables).

Independent variable: Engine size

Dependent variable: Price

**Regression between price and engine size:**

****

**Figure 9- Model 1**

Chart, scatter chart

Description automatically generatedThe R2 value of 0.743 indicates that the linear regression model explains 74.3% of the variance in the data. This is a relatively high R2 value, indicating that the model is a good fit for the data.

**Chart, histogram

Description automatically generated**

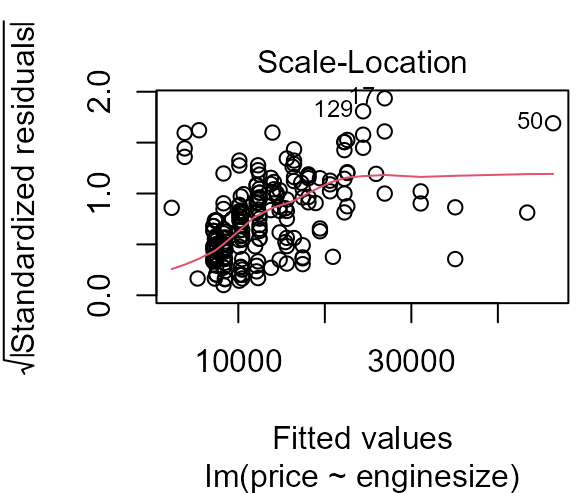
**Figure 10- Histogram of Residuals and Scatterplot with regression line (Model1)**

A histogram of residuals is a graphical representation of the distribution of the residuals (the difference between the observed values and the predicted values) of a regression analysis. If the histogram of residuals is normally distributed, it means the model is a good fit for the data. In this case, it appears to be distributed normally.

**Chart, line chart

Description automatically generatedChart, scatter chart

Description automatically generated**

****

**Chart, scatter chart

Description automatically generated**

**Figure 11- Diagnostic plots( Model 1)**

Diagnostic plots are used to assess the validity of a model. They can help identify potential issues with the model, such as outliers, non-linearity, heteroscedasticity, and other problems.

**PART-1 - Linear Regression with dummy variables**

**Regression between price, engine size, and fuel type:**

**Text, letter

Description automatically generated**

**Figure 12- Model 2**

By giving each category a numerical value, you may use the ifelse method to construct dummy variables. For example, if a variable has three categories (A, B, and C), the ifelse function can be used to assign a 0 to category A, a 1 to category B, and a 2 to category C.

The R2 value of 0.744 indicates that 74.4% of the variation in the dependent variable can be explained by the independent variable.

**Chart, histogram

Description automatically generated**Chart, scatter chart

Description automatically generated

**Figure 13- Histogram of Residuals and Scatterplot with regression line (Model2)**

For model 2, the distribution seems to be normal for the histogram of residuals. There are two lines in the regression plot.

**Chart, line chart, histogram

Description automatically generatedChart, scatter chart

Description automatically generated**

**Chart, scatter chart

Description automatically generatedChart, scatter chart

Description automatically generated**

**Figure 14- Diagnostic plots( Model 2)**

**Chart

Description automatically generated**

**Figure 15- avplots( Model 2)**

**Regression between price, engine size, and drive wheels:**

**Text

Description automatically generated**

**Figure 16- Model 3**

0.7775 is a relatively high R2 value, suggesting that the model is a good fit for the data.

**Chart, scatter chart

Description automatically generated**Chart, histogram

Description automatically generated

**Figure 17- Histogram of Residuals and Scatterplot with regression line (Model3)**

When we add drivewheels in regression model, we can again see the histogram to be normal. There are three lines in the regression plot for 3 categories in drive wheels.

**Chart, line chart, histogram

Description automatically generatedChart, scatter chart

Description automatically generated**

**Chart, scatter chart

Description automatically generated**

**Chart, scatter chart

Description automatically generated**

**Figure 18- Diagnostic plots( Model 3)**

**Chart

Description automatically generated**

**Figure 19- avplots( Model 3)**

Question - How does this impact your understanding of the impact of the categorical variable on the regression?

When we convert a categorical variable into dummy variables, it allows the regression to account for the different levels of the categorical variable. This can help the regression to better identify relationships between the categorical variable and the other variables in the regression. Additionally, it can help to reduce the amount of multicollinearity in the regression, as dummy variables are not correlated with each other

By plotting multiple regression lines, it is easier to see how different categories of the categorical variable affect the regression. This can help to identify any potential differences in the regression line based on the categorical variable.

**Part 2: Separate regression lines for each subset**

**Regression between price, engine size, and fueltypes:**

I selected a subset of rows and/or columns from a data frame or matrix using the subset() function in R. I am going to analyze two variables – first one will be fueltypes so I created 2 subsets (gas and diesel) and the second one is drive wheels for which I made 3 subsets (fourwheel, rearwheel and frontwheel).

****

**Text

Description automatically generated with medium confidence**

**Figure 20- Subsets for fuel types and drive wheels**

For model 4, I have considered gas subset as data in the lm function

**Text

Description automatically generated**

**Figure 21- Model 4**

The R2 score is significantly high (0.7775), indicating that the model is accurate.

**Chart, scatter chart

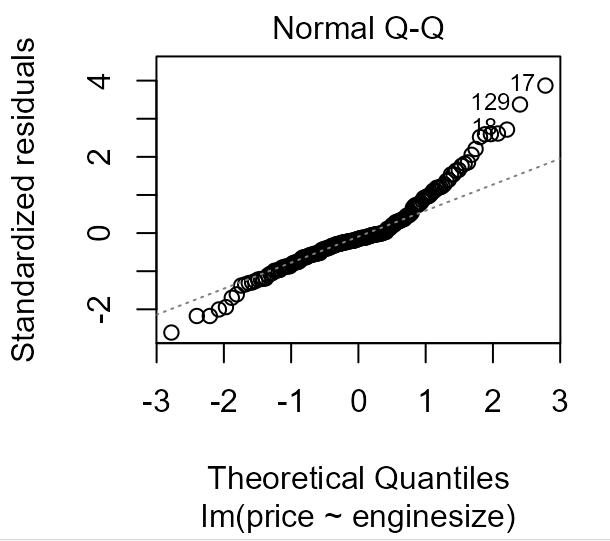
Description automatically generated**

**Chart, histogram

Description automatically generated**

**Figure 22- Histogram of Residuals and Scatterplot with regression line (Model4)**

Histogram of residuals appears normal

****

**Chart, scatter chart

Description automatically generated**

**Chart, scatter chart

Description automatically generated**Chart, scatter chart

Description automatically generated

**Figure 23- Diagnostic plots( Model 4)**

For model 5, I have considered the diesel subset as data in the lm function

**Text, letter

Description automatically generated**

**Figure 24- Model 5**

The R2 score for diesel subset is 0.8012 which is greater than gas subset.

**Chart, scatter chart

Description automatically generated**

Chart, bar chart

Description automatically generated

**Figure 25- Histogram of Residuals and Scatterplot with regression line (Model5)**

In this situation, histogram is not normal

**Chart, histogram, scatter chart

Description automatically generatedChart

Description automatically generated**

Chart, scatter chart

Description automatically generatedChart, scatter chart

Description automatically generated

**Figure 26- Diagnostic plots( Model 5)**

**Regression between price, engine size, and drive wheels:**

**Text, letter

Description automatically generated**

**Figure 27- Model 6**

The model's R2 score of 0.7526 explains 75.26% of the variation in the dependent variable.

**Chart, line chart

Description automatically generated**

**Chart, bar chart

Description automatically generated**

**Figure 28- Histogram of Residuals and Scatterplot with regression line (Model6)**

Histogram of residuals for Model 6 is not normal

Chart, scatter chart

Description automatically generatedChart, line chart

Description automatically generated

**Chart, line chart

Description automatically generatedChart

Description automatically generated**

**Figure 29- Diagnostic plots( Model 6)**

**Text, letter

Description automatically generated**

**Figure 30- Model 7**

A value of 0.6625 indicates that the regression line fits the data fairly well.

**Chart, scatter chart

Description automatically generated**

**Chart, histogram

Description automatically generated**

***Figure 31- Histogram of Residuals and Scatterplot with regression line (Model7)***

The above histogram seems to be multimodal

**Chart, histogram

Description automatically generatedChart, scatter chart

Description automatically generated**

Chart, scatter chart

Description automatically generated

Chart, scatter chart

Description automatically generated

**Figure 32- Diagnostic plots( Model 7)**

**Text, letter

Description automatically generated**

**Figure 33- Model 8**

A moderate level of accuracy is shown by the model's R2 score of 0.4768.

**Chart, scatter chart

Description automatically generated**

**Chart, histogram

Description automatically generated**

***Figure 34- Histogram of Residuals and Scatterplot with regression line (Model8)***

Model 8 Histograms of Residuals is skewed left

**Chart, histogram

Description automatically generatedChart, scatter chart

Description automatically generated**

**Chart

Description automatically generated**

Chart, scatter chart

Description automatically generated

**Figure 35- Diagnostic plots( Model 8)**

Question - How do these regression lines differ from the regression lines in step 1? How does this method of looking at the data impact your understanding of the data?

I can find any patterns or trends in the data using a second method that might not be visible when analyzing the complete dataset.. By looking at the subsetted regression lines, I can gain a better understanding of the data and how it is related to each other.

**CONCLUSION**

In this assignment, I initially performed data cleaning, EDA, then plotted the correlation matrix and used regression models to determine the factors that influence the price of the cars.

The key findings are :

* Using a correlation matrix, we could comprehend the relationship between price and other features.
* The R-square value of the all the models are able to explain a large portion of the variance in the data.It means that other variables in the model are strongly related to the dependent variable (price).

**REFERENCES**

Bluman, A. G. (2018). Elementary Statistics, 10th ed. McGraw Hill.

Kabacoff, R. I. (2011). R in action: Data analysis and graphics with R. Manning Publications Co.

Zach. (2019, May 11). How to Perform Multiple Linear Regression in R. Statology.

https://www.statology.org/multiple-linear-regression-r/

**APPENDIX: CODE**

#---------------------- Week\_6\_Module\_6 R Script ----------------------#

print("Author : Nikshita Ranganathan")

print("Week 6 Assignment - Module 6 R Practice")

print("Course Name - ALY6010: Probability Theory and Introductory Statistics")

# Importing cars dataset

getwd()

cars<-read.csv("scrap price.csv")

# Installing and loading the libraries

library(psych)

library(skimr)

library(ggplot2)

library(dplyr)

library(naniar)

library(visdat)

library(corrplot)

library(corrgram)

library(Hmisc)

library(gridExtra)

library(wesanderson)

install.packages("ggiraph")

install.packages("ggiraphExtra")

install.packages("plyr")

require(ggiraph)

require(ggiraphExtra)

library(ggpmisc)

require(plyr)

library(ggpubr)

library(car)

# Visualization and Checking NA values

vis\_dat(cars)

gg\_miss\_which(cars)

sum(is.na(cars))

sum(is.null(cars))

# Check for duplication

anyDuplicated(cars)

duplicated(cars)

# Changing the datatypes

cars$fueltypes<-as.factor(cars$fueltypes)

cars$fuelsystem<-as.factor(cars$fuelsystem)

cars$enginetype<-as.factor(cars$enginetype)

cars$enginelocation<-as.factor(cars$enginelocation)

cars$aspiration<-as.factor(cars$aspiration)

cars$carbody<-as.factor(cars$carbody)

cars$doornumbers<-as.factor(cars$doornumbers)

cars$drivewheels<-as.factor(cars$drivewheels)

cars$cylindernumber<-as.factor(cars$cylindernumber)

# Outliers

boxplot(cars$price)

cars<-subset(cars,price!=45400)

# Analysis

headTail(cars)

str(cars)

summary(cars)

dim(cars)

skim(cars)

describe(cars)

glimpse(cars)

describeBy(cars,group=cars$fueltypes,quant = c(0.25, 0.75), IQR = T)

# Correlation

# Correlation of Price with Car Dimensions

corr1<-cars %>% select(wheelbase,carlength,carwidth,carheight,curbweight,price)

corrgram(corr1, order=TRUE, upper.panel=panel.cor)

corrgram(corr1, lower.panel=panel.pts, upper.panel=panel.conf,

diag.panel=panel.density)

# Correlation of Price with Engine Specification

corr2<-cars %>% select(enginesize,boreratio,stroke,compressionratio,price)

corrgram(corr2, order=TRUE, lower.panel=panel.conf,

upper.panel=panel.pie, text.panel=panel.txt)

corrgram(corr2, lower.panel=panel.pts, upper.panel=panel.conf,

diag.panel=panel.density)

corr3<-cars %>% select(horsepower,peakrpm,citympg,highwaympg,price)

corrgram(corr3, order=TRUE, upper.panel=panel.conf,

lower.panel=panel.ellipse, text.panel=panel.txt)

corrgram(corr3, lower.panel=panel.pts, upper.panel=panel.conf,

diag.panel=panel.density)

# Models

model1<-lm(formula = price ~ enginesize, data = cars)

summary(model1)

ggplot(data = cars, aes(x = model1$residuals)) +geom\_histogram(fill = 'steelblue', color = 'black') + labs(title = 'Histogram of Residuals (Model1)', x = 'Residuals', y = 'Frequency')

ggPredict(model1,se=TRUE,interactive=TRUE)

plot(model1)

cars$fuelgas<-ifelse(cars$fueltypes=="gas",1,0)

cars$fueldiesel<-ifelse(cars$fueltypes=="diesel",1,0)

model2=lm(price~enginesize+fuelgas+fueldiesel,data=cars)

summary(model2)

ggplot(data = cars, aes(x = model2$residuals)) +geom\_histogram(fill = 'steelblue', color = 'black') + labs(title = 'Histogram of Residuals (Model2)', x = 'Residuals', y = 'Frequency')

ggplot(cars,aes(y=price,x=enginesize,color=factor(fueltypes)))+geom\_point(size=0.75)+stat\_smooth(method="lm",se=FALSE)+scale\_colour\_brewer(palette = "Accent")+labs(color="Fuel Types")+ stat\_poly\_eq(aes(label = paste(..eq.label.., ..rr.label.., sep = "~~~")))

plot(model2)

avPlots(model2)

cars$wheelsrwd<-ifelse(cars$drivewheels=="rwd",1,0)

cars$wheelsfwd<-ifelse(cars$drivewheels=="fwd",1,0)

cars$wheels4wd<-ifelse(cars$drivewheels=="4wd",1,0)

model3=lm(price~enginesize+wheels4wd+wheelsfwd+wheelsrwd,data=cars)

summary(model3)

ggplot(data = cars, aes(x = model3$residuals)) +geom\_histogram(fill = 'steelblue', color = 'black') + labs(title = 'Histogram of Residuals (Model3)', x = 'Residuals', y = 'Frequency')

ggplot(cars,aes(y=price,x=enginesize,color=factor(drivewheels)))+geom\_point()+stat\_smooth(method="lm",se=FALSE)+scale\_colour\_brewer(palette = "Set1")+labs(color="Drive Wheels")+ stat\_poly\_eq(aes(label = paste(..eq.label.., ..rr.label.., sep = "~~~")))

plot(model3)

avPlots(model3)

# Creating subsets

gas=subset(cars, fueltypes == "gas")

diesel=subset(cars, fueltypes == "diesel")

model4=lm(price~enginesize,data=gas)

summary(model4)

ggplot(data = gas, aes(x = model4$residuals)) +geom\_histogram(fill = 'steelblue', color = 'black') + labs(title = 'Histogram of Residuals (Model4)', x = 'Residuals', y = 'Frequency')

ggplot(gas,aes(y=price,x=enginesize))+geom\_point(size=0.75)+geom\_smooth(method="lm")+stat\_regline\_equation(label.x = 50, label.y = 45000)

plot(model4)

model5=lm(price~enginesize,data=diesel)

summary(model5)

ggplot(data = diesel, aes(x = model5$residuals)) +geom\_histogram(fill = 'steelblue', color = 'black') + labs(title = 'Histogram of Residuals (Model5)', x = 'Residuals', y = 'Frequency')

ggplot(diesel,aes(y=price,x=enginesize))+geom\_point(size=0.75)+geom\_smooth(method="lm")+stat\_regline\_equation(label.x = 103, label.y = 27000)

plot(model5)

fourwheel=subset(cars, drivewheels == "4wd")

rearwheel=subset(cars, drivewheels == "rwd")

frontwheel=subset(cars, drivewheels == "fwd")

model6=lm(price~enginesize,data=fourwheel)

summary(model6)

ggplot(data = fourwheel, aes(x = model6$residuals)) +geom\_histogram(fill = 'steelblue', color = 'black') + labs(title = 'Histogram of Residuals (Model6)', x = 'Residuals', y = 'Frequency')

ggplot(fourwheel,aes(y=price,x=enginesize))+geom\_point(size=0.75)+geom\_smooth(method="lm")+stat\_regline\_equation(label.x = 92, label.y = 20000)

plot(model6)

model7=lm(price~enginesize,data=rearwheel)

summary(model7)

ggplot(data = rearwheel, aes(x = model7$residuals)) +geom\_histogram(fill = 'steelblue', color = 'black') + labs(title = 'Histogram of Residuals (Model7)', x = 'Residuals', y = 'Frequency')

ggplot(rearwheel,aes(y=price,x=enginesize))+geom\_point(size=0.75)+geom\_smooth(method="lm")+stat\_regline\_equation(label.x = 80, label.y = 47000)

plot(model7)

model8=lm(price~enginesize,data=frontwheel)

summary(model8)

ggplot(data = frontwheel, aes(x = model8$residuals)) +geom\_histogram(fill = 'steelblue', color = 'black') + labs(title = 'Histogram of Residuals (Model8)', x = 'Residuals', y = 'Frequency')

ggplot(frontwheel,aes(y=price,x=enginesize))+geom\_point(size=0.75)+geom\_smooth(method="lm")+stat\_regline\_equation(label.x = 50, label.y = 25000)

plot(model8)