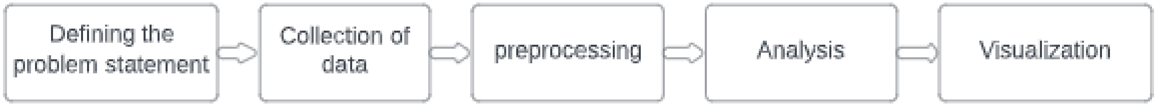
Assignment No. 2.2

Aim: Perform data analysis using R programming

Data Analysis is a subset of data analytics, it is a process where the objective has to be made clear, collect the relevant data, preprocess the data, perform analysis(understand the data, explore insights), and then visualize it. The last step visualization is important to make people understand what's happening in the firm.

Steps involved in data analysis:



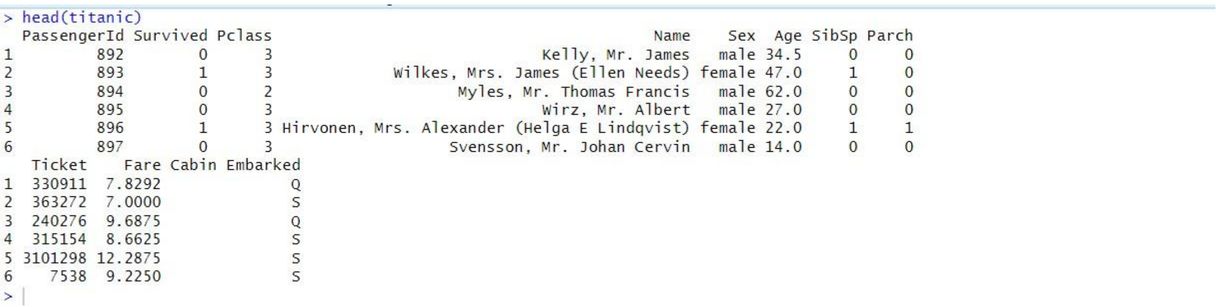
The aim of this practical is to perform data analysis on the Titanic dataset using R to explore and understand the data. The dataset contains information about passengers, including their survival status, age, gender, and other attributes. Initially, the data is loaded, and its structure is examined using functions like head(), View(), and sapply() to identify data types and inspect the first few rows. Data preprocessing steps follow, such as converting categorical variables like Survived and Sex into factors for easier analysis. The dataset is then checked for missing values, and a filtered dataset (dropnull_titanic) is created by removing rows with missing data.

The analysis is segmented into survival-based groups, dividing passengers into those who survived (slist) and those who did not (nslist). Exploratory Data Analysis (EDA) is performed using visualizations such as histograms and bar plots. The histogram provides insights into the age distribution of survivors, while the bar plot examines the gender distribution of non-survivors. These steps offer a deeper understanding of the relationships and trends within the dataset, such as survival rates by age and gender, aiding in the exploration of patterns that might have influenced survival during the Titanic disaster.

Code and Output:

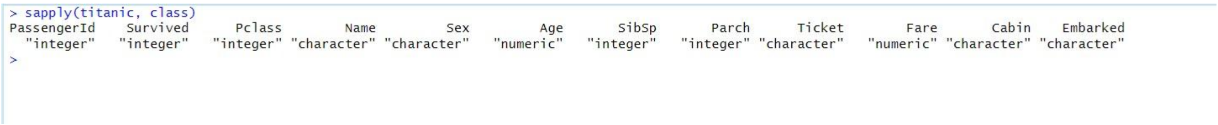
```
titanic = read.csv("titanic.csv")
```

head(titanic)



```
> head(titanic)
  PassengerId Survived Pclass                                         Name    Sex  Age SibSp Parch
1         892        0      3                             Kelly, Mr. James   male 34.5     0     0
2         893        1      3             Wilkes, Mrs. James (Ellen Needs) female 47.0     1     0
3         894        0      2                    Myles, Mr. Thomas Francis   male 62.0     0     0
4         895        0      3                             Wirz, Mr. Albert   male 27.0     0     0
5         896        1      3 Hirvonen, Mrs. Alexander (Helga E Lindqvist) female 22.0     1     1
6         897        0      3                   Svensson, Mr. Johan Cervin   male 14.0     0     0
   Ticket    Fare Cabin Embarked
1  330911  7.8292              Q
2  363272  7.0000              S
3  240276  9.6875              Q
4  315154  8.6625              S
5 3101298 12.2875              S
6    7538  9.2250              S
> 
```

sapply(titanic, class)



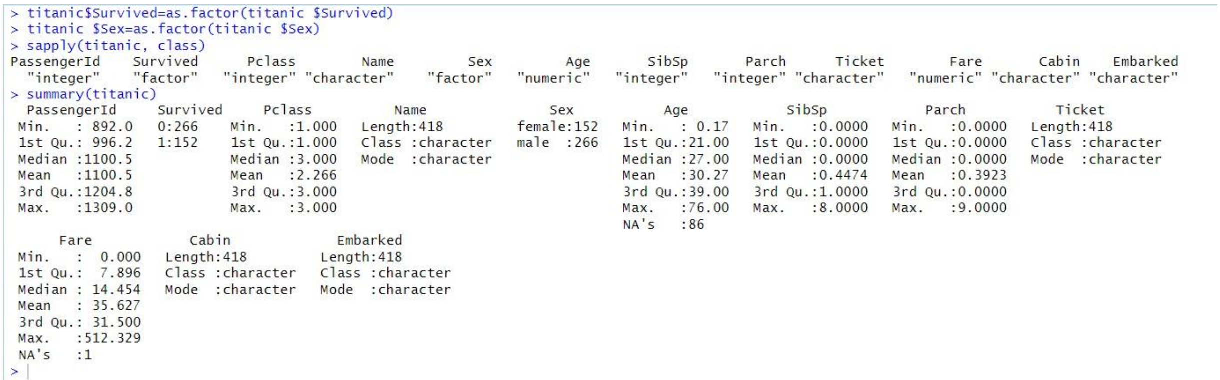
```
> sapply(titanic, class)
PassengerId    Survived      Pclass        Name         Sex         Age       SibSp       Parch      Ticket        Fare       Cabin    Embarked 
  "integer"   "integer"   "integer" "character" "character"   "numeric"   "integer"   "integer" "character"   "numeric" "character" "character" 
> 
```

titanic$Survived=as.factor(titanic $Survived)

titanic $Sex=as.factor(titanic $Sex)

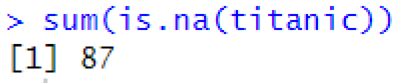
sapply(titanic, class)

summary(titanic)



```
> titanic$Survived=as.factor(titanic $Survived)
> titanic $Sex=as.factor(titanic $Sex)
> sapply(titanic, class)
PassengerId    Survived      Pclass        Name         Sex         Age       SibSp       Parch      Ticket        Fare       Cabin    Embarked 
  "integer"    "factor"   "integer" "character"    "factor"   "numeric"   "integer"   "integer" "character"   "numeric" "character" "character" 
> summary(titanic)
  PassengerId     Survived     Pclass          Name                Sex           Age            SibSp            Parch            Ticket         
 Min.   : 892.0   0:266    Min.   :1.000   Length:418         female:152   Min.   : 0.17   Min.   :0.0000   Min.   :0.0000   Length:418        
 1st Qu.: 996.2   1:152    1st Qu.:1.000   Class :character   male  :266   1st Qu.:21.00   1st Qu.:0.0000   1st Qu.:0.0000   Class :character  
 Median :1100.5            Median :3.000   Mode  :character                Median :27.00   Median :0.0000   Median :0.0000   Mode  :character  
 Mean   :1100.5            Mean   :2.266                                   Mean   :30.27   Mean   :0.4474   Mean   :0.3923                     
 3rd Qu.:1204.8            3rd Qu.:3.000                                   3rd Qu.:39.00   3rd Qu.:1.0000   3rd Qu.:0.0000                     
 Max.   :1309.0            Max.   :3.000                                   Max.   :76.00   Max.   :8.0000   Max.   :9.0000                     
                                                                           NA's   :86                                                          
      Fare             Cabin             Embarked        
 Min.   :  0.000   Length:418         Length:418        
 1st Qu.:  7.896   Class :character   Class :character  
 Median : 14.454   Mode  :character   Mode  :character  
 Mean   : 35.627                                        
 3rd Qu.: 31.500                                        
 Max.   :512.329                                        
 NA's   :1                                              
> 
```

sum(is.na(titanic))

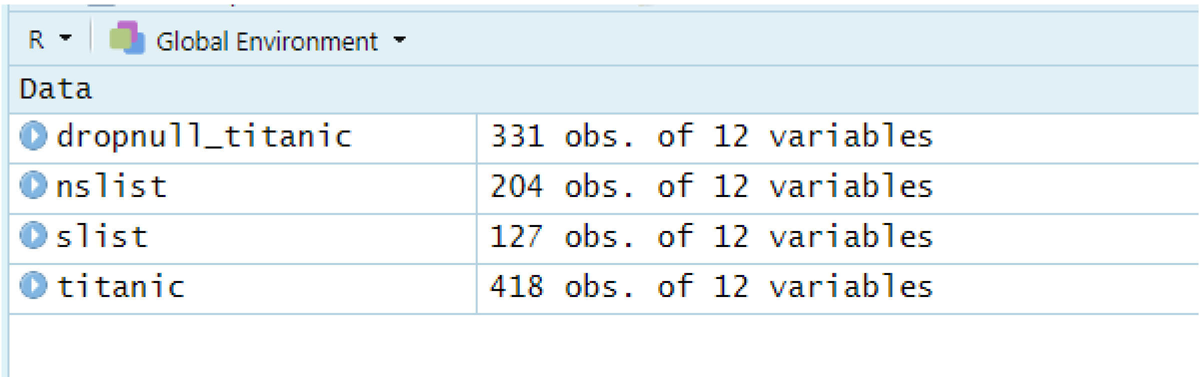


```
> sum(is.na(titanic))
[1] 87
```

dropnull_titanic = titanic[rowSums(is.na(titanic))<=0,]

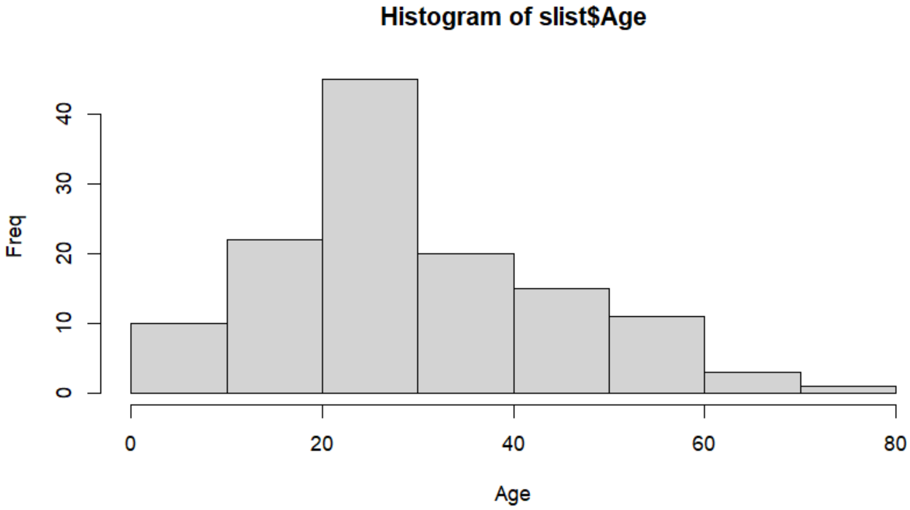
slist = dropnull_titanic[dropnull_titanic$Survived==1,]

nslist = dropnull_titanic[dropnull_titanic$Survived==0,]

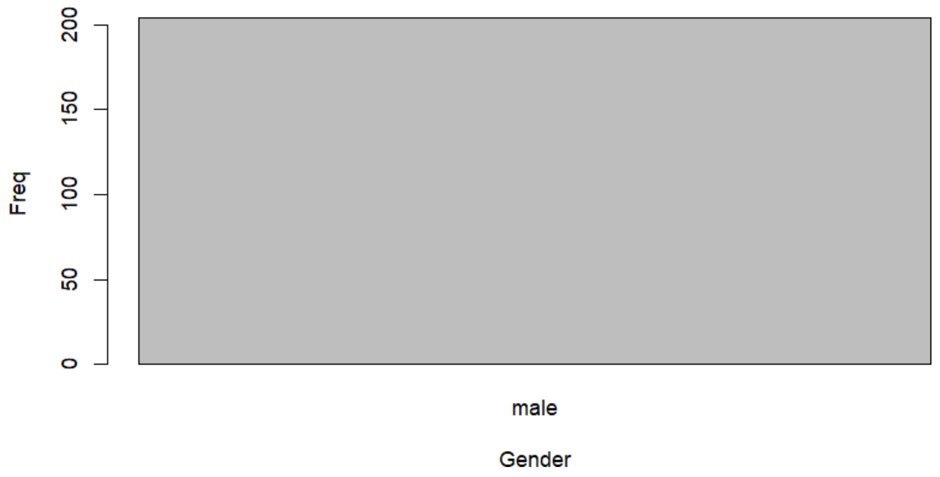


```
Data
dropnull_titanic    331 obs. of 12 variables
nslist              204 obs. of 12 variables
slist               127 obs. of 12 variables
titanic             418 obs. of 12 variables
```

```
hist(slist$Age, xlab = "Age", ylab = "Freq")
```



```
barplot(table(nslist$Sex), xlab = "Gender", ylab = "Freq")
```

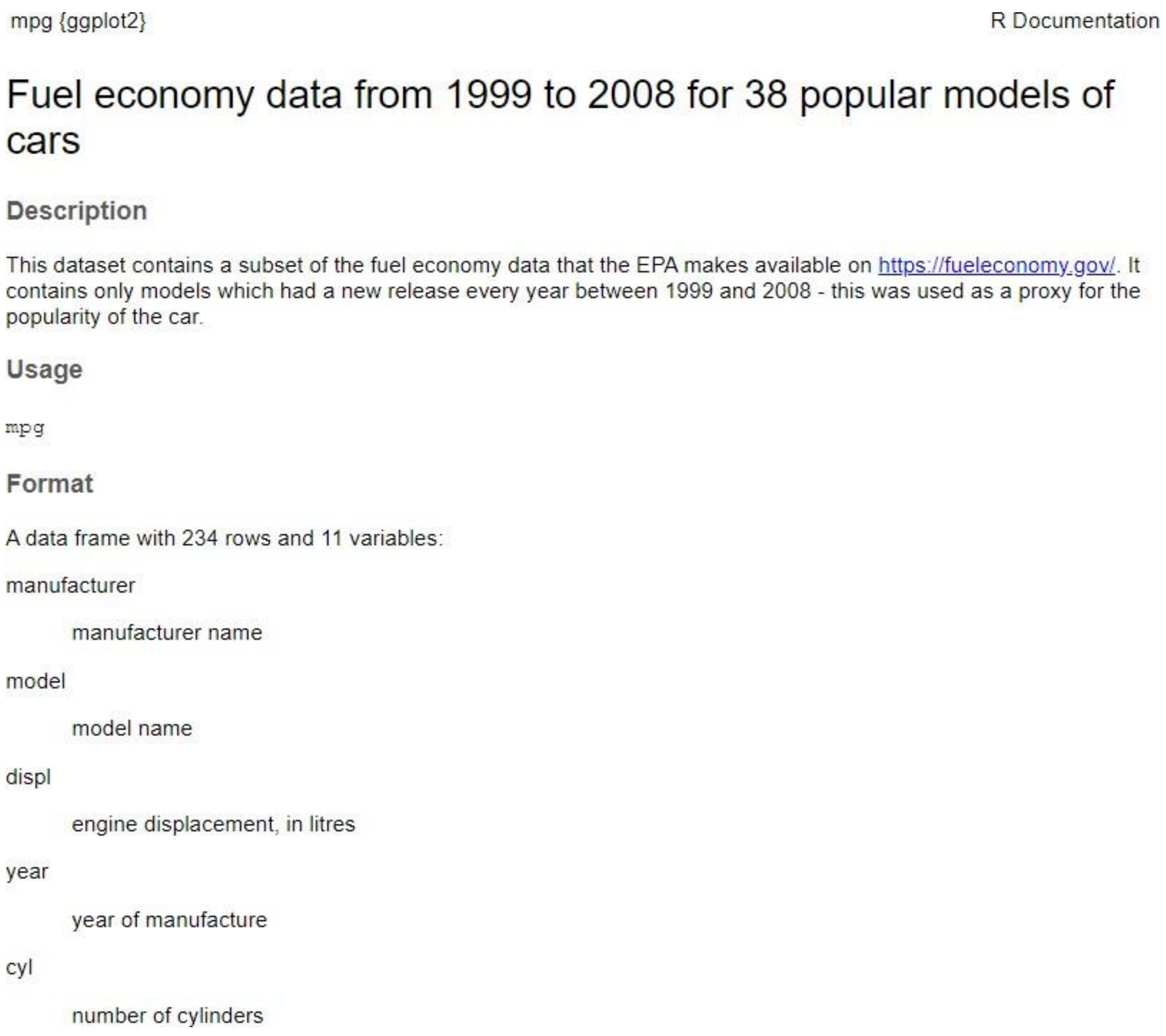


Tidyverse

library(tidyverse)

View(mpg)

?mpg



mpg {ggplot2} R Documentation

Fuel economy data from 1999 to 2008 for 38 popular models of cars

Description

This dataset contains a subset of the fuel economy data that the EPA makes available on https://fueleconomy.gov/. It contains only models which had a new release every year between 1999 and 2008 - this was used as a proxy for the popularity of the car.

Usage

```
mpg
```

Format

A data frame with 234 rows and 11 variables:

manufacturer

manufacturer name

model

model name

displ

engine displacement, in litres

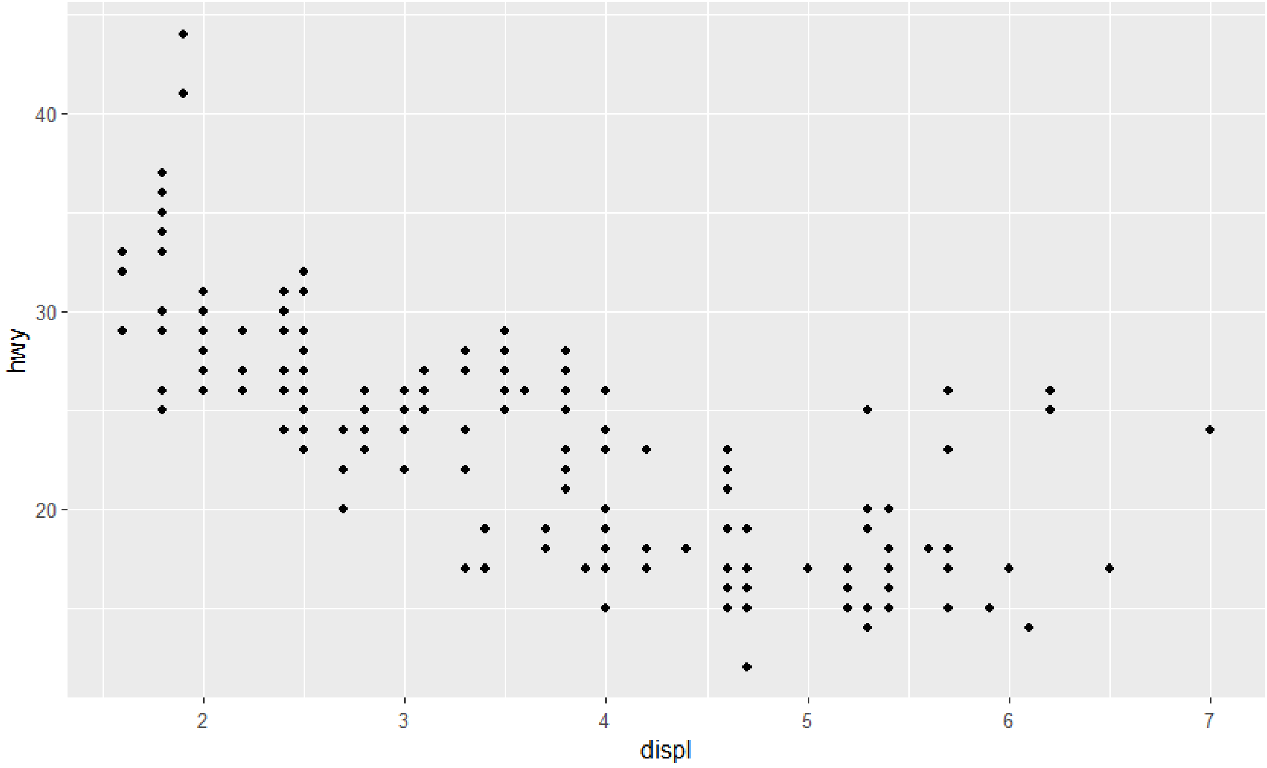
year

year of manufacture

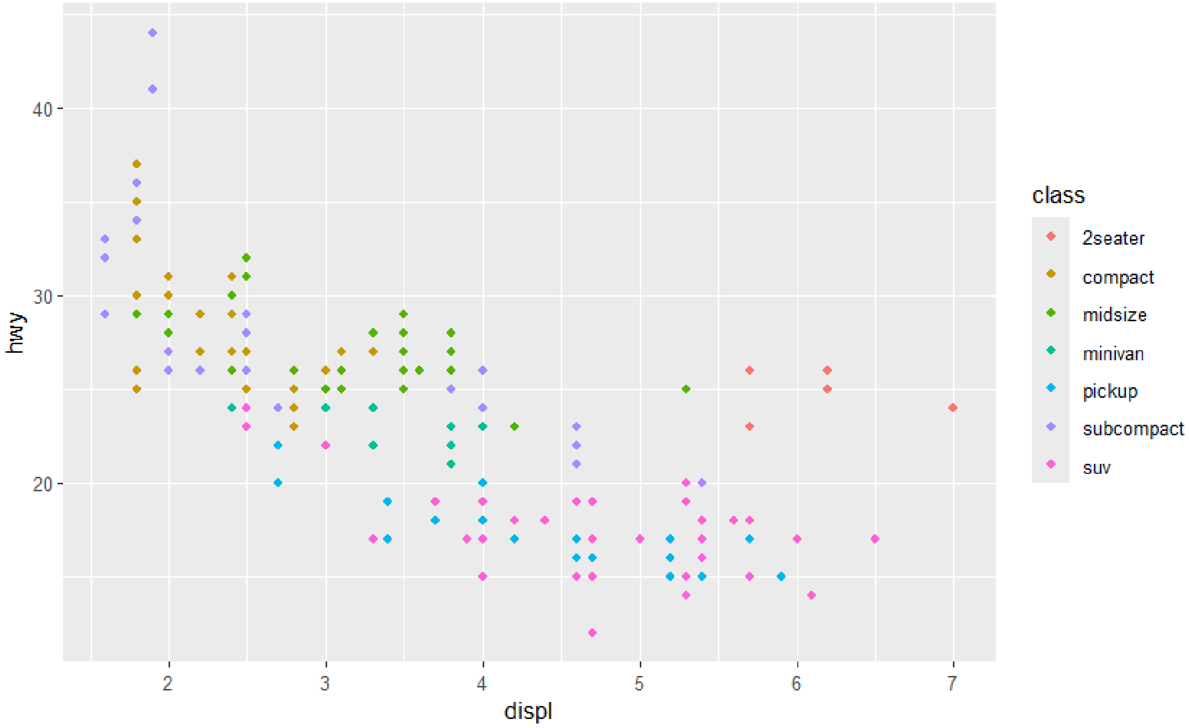
cyl

number of cylinders

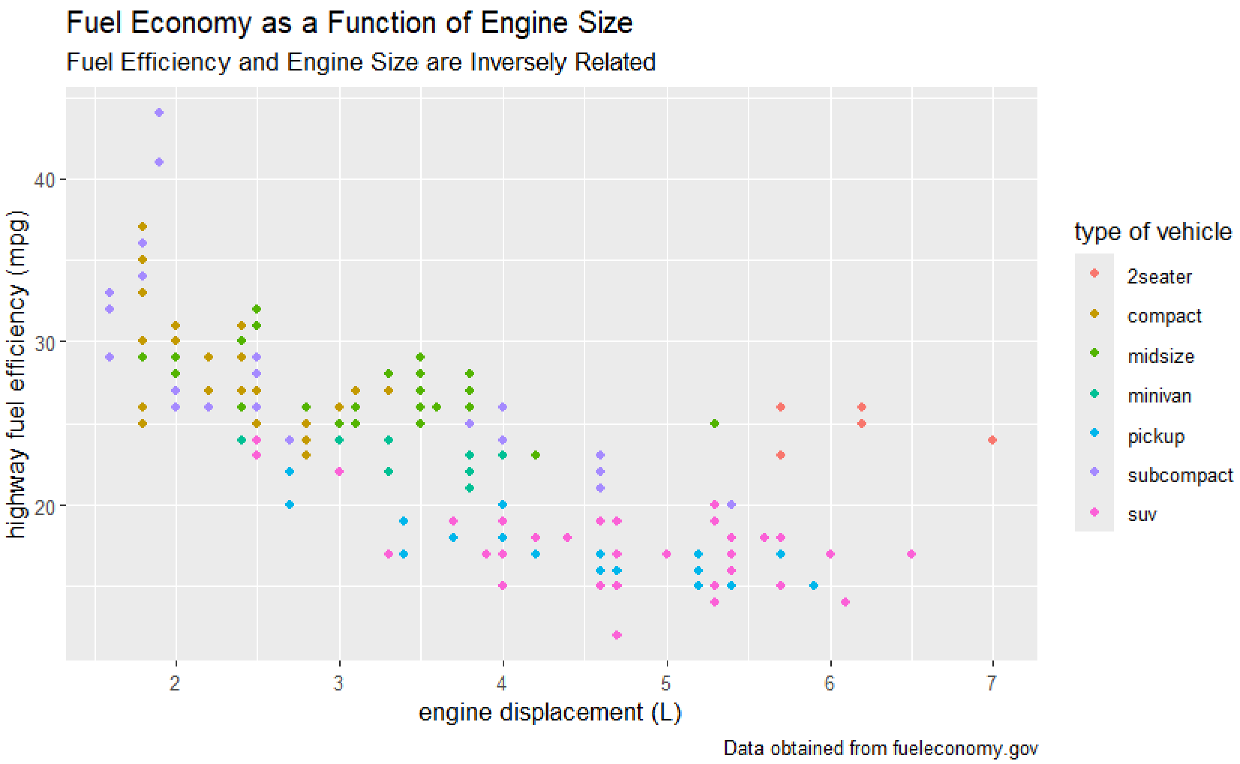
```
ggplot(data = mpg) + geom_point(mapping = aes(x = displ, y = hwy))
```



```
ggplot(data = mpg) + geom_point(mapping = aes(x = displ, y = hwy, color = class))
```



```
ggplot(mpg) +
  geom_point(mapping = aes(x = displ, y = hwy, color = class)) +
labs(x = "engine displacement (L)",
     y = "highway fuel efficiency (mpg)",
     color = "type of vehicle",
     title = "Fuel Economy as a Function of Engine Size",
     subtitle = "Fuel Efficiency and Engine Size are Inversely Related",
     caption = "Data obtained from fueleconomy.gov")
```



Conclusion:

Analyzing the Titanic dataset highlighted survival patterns based on age and gender, demonstrating the value of data cleaning and exploratory analysis.