College of Computing and Informatics

Artificial Intelligence

DS-363

Project/Case Study

In this project, you will learn how to manually classify and cluster data based on a specific variable using R programming language. The project consists of retrieving data, classifying, plot a graph and interpretation of the outcomes.

The dataset that are going to be used in this project is “mtcar” that is available in R studio.

Methodology:

Using the number of cylinders (cyl column) in mtcar dataset(below), classify cars as 4-, 6- or 8-cylinders car. For each class, give a color to the values that belong to that class. Then plot a graph for using “mpg” column as x-axis and “qsec” column as y-axis.

mtcar

	mpg	cyl	disp	hp	drat	wt	qsec	vs	am	gear	carb
Mazda RX4	21	6	160	110	3.9	2.62	16.46	0	1	4	4
Mazda RX4 Wag	21	6	160	110	3.9	2.875	17.02	0	1	4	4
Datsun 710	22.8	4	108	93	3.85	2.32	18.61	1	1	4	1
Hornet 4 Drive	21.4	6	258	110	3.08	3.215	19.44	1	0	3	1
Hornet Sportabout	18.7	8	360	175	3.15	3.44	17.02	0	0	3	2
Valiant	18.1	6	225	105	2.76	3.46	20.22	1	0	3	1
Duster 360	14.3	8	360	245	3.21	3.57	15.84	0	0	3	4
Merc 240D	24.4	4	146.7	62	3.69	3.19	20	1	0	4	2
Merc 230	22.8	4	140.8	95	3.92	3.15	22.9	1	0	4	2

Steps:

1- Read the file using R programming language.
2- Read values in column “cyl”.
3- If x =8, then paint as red, if x =6, then paint as green, If x =4, then paint as blue.
4- Plot the values of the columns mpg and qsec into an x&y graph.
5- Plot painted values of “cyl” on the same graph.

Finally, you have to add your interpretations of the results. This is very important!

Marking criteria:

5 marks on your implementation.

5 marks on interpretation

4 marks on Presentation

Problem Statement:

The problem focuses on classifying cars based on the number of cylinders in their engines. This classification helps show how engine size affects both fuel efficiency and acceleration performance.

By grouping the cars into three-cylinder categories, we can visualize the relationship between MPG and QSEC in a simple scatter plot and observe how engine design impacts fuel economy and acceleration.

Interpretation

In this section, the main variables used in the analysis are defined to help interpret the results correctly, before analyzing the results, it is essential to define the main metrics.

- **MPG**

  - A higher value indicates better fuel efficiency.
  - Cars with higher MPG consume less fuel and are more economical.
  - Lower MPG values represent cars that are less efficient and have poorer fuel economy.

- **QSEC**

  - A higher value indicates worse acceleration performance.
  - Cars with lower QSEC values can accelerate faster and perform better.
  - Therefore, a lower QSEC generally corresponds to better engine power and responsiveness.

Figure 1. R code and scatter plot:

In this figure I show the R script I used to work with the mtcars dataset. First, I load the data and select the variables mpg (Miles per Gallon) and qsec (Quarter Mile Time), along with cyl for the number of cylinders. Then I create a color vector so that cars with 4 cylinders are blue, 6 cylinders are green, and 8 cylinders are red. The scatter plot on the right display's mpg on the x-axis and qsec on the y-axis, with each point colored by its cylinder group and a legend that explains the colors.

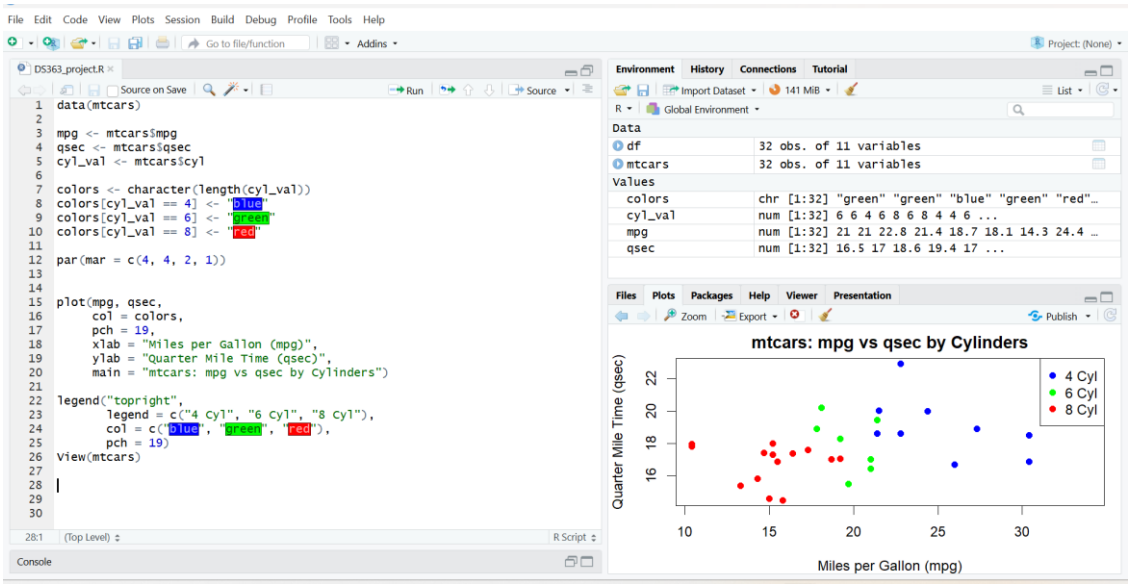
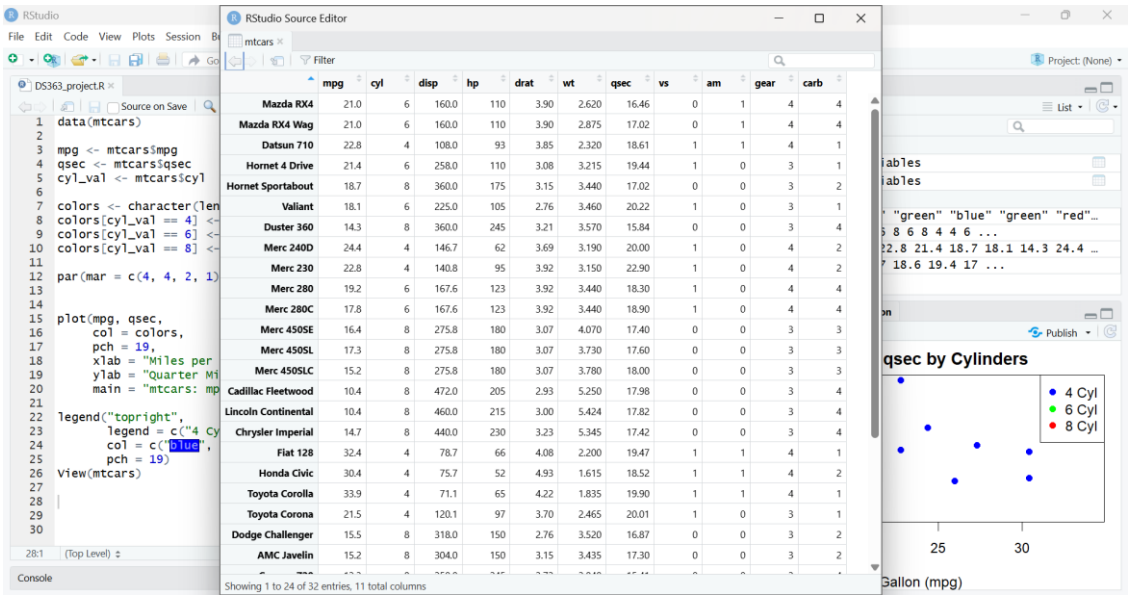


Figure 2. mtcars data table and colored groups:

This figure shows the mtcars data table that I used in the analysis, The table contains 32 cars (rows) and 11 variables (columns), including mpg, cyl, qsec, and other engine and performance measurements. This is the raw data that my R code reads before creating the scatter plot. The plot at the bottom is the same mpg vs qsec graph, where the points are grouped by the number of cylinders (4, 6, and 8) using the color scheme defined in the script.



Full explanation for the R code:

1) Loading the dataset:

data(mtcars)

Explanation:

Here I simply load the mtcars dataset that is already inside R, so that I can use it in the remaining steps.

2) Extracting the variables I need:

mpg <- mtcars$mpg

qsec <- mtcars$qsec

cyl_val <- mtcars$cyl

Explanation:

Pulling only three columns of data:

-mpg for fuel consumption

-qsec for quarter-mile time

-cyl_val for the number of cylinders

These are the variables I need to plot the required relationship without any additional information.

3) Creating the color vector for cylinder groups:

```
colors <- character(length(cyl_val))

colors[cyl_val == 4] <- "blue"

colors[cyl_val == 6] <- "green"

colors[cyl_val == 8] <- "red"
```

Explanation:

Create a blank list, then color each car according to the number of cylinders:

-Blue for 4-cylinder cars

-Green for 6-cylinder cars

-Red for 8-cylinder cars

This step helps me distinguish each group by color in the drawing.

4) Adjusting the plot margins:

```
par(mar = c(4, 4, 2, 1))
```

Explanation:

I change the margins so that the drawing elements, such as the title and axes, are clear and do not clash with the edges of the shape.

5) Creating the scatter plot:

```
plot(mpg, qsec,
    col = colors,
    pch = 19,
    xlab = "Miles per Gallon (mpg)",
    ylab = "Quarter Mile Time (qsec)",
    main = "mtcars: mpg vs qsec by Cylinders")
```

Explanation:

Here I create the main diagram.

I place mpg on the X-axis and qsec on the Y-axis, and each point takes its color according to the set of cylinders.

I use a full point (pch = 19) and add a title and labels to the axes to make the diagram understandable.

6) Adding the legend:

```
legend("topright",
       legend = c("4 Cyl", "6 Cyl", "8 Cyl"),
       col = c("blue", "green", "red"),
       pch = 19)
```

Explanation:

I added a small box explaining the meaning of each color, so the reader knows which color corresponds to which number of cylinders.

I placed it in the upper right corner so it wouldn't obscure the dots.

7) Viewing the data table:

```
View(mtcars)
```

Explanation:

This line opens the table inside RStudio, so I can see the raw data like the one I attached in the table image.

Initial Conceptual Analysis:

Before looking at the plot, I wanted to remind myself how these variables usually behave in real cars. This makes it easier to understand whether the patterns in the dataset are reasonable.

1) Cylinders and Fuel Efficiency:

Cars with more cylinders usually use more fuel.

In contrast, 4-cylinder cars tend to be the most fuel-efficient.

The idea is simple: bigger engines need more power to run, so they burn more gasoline.

2) Cylinders and Performance (qsec):

More cylinders usually mean more power.

So, cars with 6 or 8 cylinders can accelerate faster, which gives them a lower qsec time.

3) Fuel Efficiency vs Performance:

Fast cars are almost never the most fuel-efficient, and the cars that save fuel aren't usually the quickest.

Because of that, mpg and qsec often move in opposite directions.

Overall Expectation:

From this, I expect:

4-cyl cars: higher mpg but slower acceleration

6-cyl cars: somewhere in the middle

8-cyl cars: fastest acceleration but lowest mpg

These basic ideas help explain the differences we see later in the scatter plot.

Summary:

This case study provided a clear understanding of how engine size affects both mpg and acceleration, which helped highlight the basic trade-off between performance and fuel efficiency when comparing different cars.

References:

- R Core Team (2024). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL: https://www.R-project.org/
- Henderson, H. V., & Velleman, P. F. (1981). Building multiple regression models interactively. Biometrics, 37(2), 391–411.
- R Core Team. (2024). R Documentation. https://www.rdocumentation.org/
- https://www.w3schools.com/r/