Prediction of Fuel Efficiency using the Auto-MPG Dataset

Group # 14

Syeda Shanzay Shah 23I-2016

Sarita Sangrez 23I-2088

Laiba Nasir 23I-2079

CY-4A

Task 1: Introduction

- **Topic Chosen**: Prediction of Fuel Efficiency using the Auto-MPG Dataset.
- **Objective**: To find out how different features of a car (like weight, horsepower, etc.) affect its fuel efficiency (MPG).
- **Importance of Study**: Helps in understanding which car features save fuel.
- Useful for car manufacturers and customers to make better decisions Promotes environmentally friendly vehicle choices.
- **Approach**: Use data analysis and basic prediction techniques on a real-world dataset.

Task 2: Dataset Selection

- **Dataset Name:** Auto-MPG (Miles Per Gallon)
- **Source:** UCI Machine Learning Repository — a reliable and public source of datasets
- **Type:** Secondary data (already collected and made available online)
- **Total Records:** 398 car entries
- **Main Features** in the Dataset:
  - *mpg:* Miles per gallon (target variable)
  - *cylinders:* Number of cylinders in the engine
  - *horsepower:* Engine power
  - *weight:* Vehicle weight
  - *acceleration:* Time to reach a certain speed
  - *model year:* Year of car model
  - *origin:* Country of origin (1 = USA, 2 = Europe, 3 = Japan)

Task 3: Descriptive statistics on dataset

Displaying dataset:

RGui

File Edit View Misc Packages Windows Help

R Console

```
> # Read the CSV file
> data <- read.csv(file_path)
>
> ?data
> data
    mpg.Y. Cylinders Displacement HorsePower Weight acceleration ModelYear
1     18.0         8        307.0        130   3504         12.0         70
2     15.0         8        350.0        165   3693         11.5         70
3     18.0         8        318.0        150   3436         11.0         70
4     16.0         8        304.0        150   3433         12.0         70
5     17.0         8        302.0        140   3449         10.5         70
6     15.0         8        429.0        198   4341         10.0         70
7     14.0         8        454.0        220   4354          9.0         70
8     14.0         8        440.0        215   4312          8.5         70
9     15.0         8        350.0        165   3693         11.5         70
10    16.0         8        304.0        150   3433         12.0         70
11      NA         4        133.0        115   3090         17.5         70
12      NA         8        350.0        165   4142         11.5         70
13      NA         8        351.0        153   4034         11.0         70
14      NA         8        383.0        175   4166         10.5         70
15      NA         8        360.0        175   3850         11.0         70
16    15.0         8        383.0        170   3563         10.0         70
17    14.0         8        340.0        160   3609          8.0         70
18      NA         8        302.0        140   3353          8.0         70
19    15.0         8        400.0        150   3761          9.5         70
```

C:\Users\User\Downloads\Lecture 22.R - R Editor

```
# Load the dataset into the R environment
# Set the file path
file_path <- "C:/Users/User/Downloads/autompg.csv"

# Read the CSV file
data <- read.csv(file_path)

?data
data

# Display the default number of rows(lst 6 rows) from the dataset
head(data)
# Display the first 3 rows of the dataset
head(data,3)

# Display the default number of rows(last 6 rows) from the dataset
tail(data)
# Display the first 4 rows of the dataset
tail(data,4)

# Use the 'dim()' function to obtain the dimensions of the dataset
# This returns a vector with the number of rows and columns in dataset
dim(data)

# Use the 'names()' function to get the names of the columns in the da
```

Type here to search

Rain...

7:11 PM
5/7/2025

ANALYSIS OF DATASET BEFORE CLEANING

Load dataset in R environment:

RGui

File Edit View Misc Packages Windows Help

R Console

```
> file_path <- "C:/Users/User/Downloads/autompg.csv"
>
> # Read the CSV file
> data <- read.csv(file_path)
>
> ?data
> data
```

	mpg.Y.	Cylinders	Displacement	HorsePower	Weight	acceleration	ModelYear
1	18.0	8	307.0	130	3504	12.0	70
2	15.0	8	350.0	165	3693	11.5	70
3	18.0	8	318.0	150	3436	11.0	70
4	16.0	8	304.0	150	3433	12.0	70
5	17.0	8	302.0	140	3449	10.5	70
6	15.0	8	429.0	198	4341	10.0	70
7	14.0	8	454.0	220	4354	9.0	70
8	14.0	8	440.0	215	4312	8.5	70
9	15.0	8	350.0	165	3693	11.5	70
10	16.0	8	304.0	150	3433	12.0	70
11	NA	4	133.0	115	3090	17.5	70
12	NA	8	350.0	165	4142	11.5	70
13	NA	8	351.0	153	4034	11.0	70
14	NA	8	383.0	175	4166	10.5	70
15	NA	8	360.0	175	3850	11.0	70
16	15.0	8	383.0	170	3563	10.0	70
17	14.0	8	340.0	160	3609	8.0	70

C:\Users\User\Downloads\Lecture 22.R - R Editor

```
# Load the dataset into the R environment
# Set the file path
file_path <- "C:/Users/User/Downloads/autompg.csv"

# Read the CSV file
data <- read.csv(file_path)

?data
data

# Display the default number of rows(1st 6 rows) from the dataset
head(data)
# Display the first 3 rows of the dataset
head(data,3)

# Display the default number of rows(last 6 rows) from the dataset
tail(data)
# Display the first 4 rows of the dataset
tail(data,4)

# Use the 'dim()' function to obtain the dimensions of the dataset
# This returns a vector with the number of rows and columns in dataset
dim(data)

# Use the 'names()' function to get the names of the columns in the da
```

Descriptive stats on MPG:

RGui

File Edit Packages Windows Help

R Console

```
> # descriptive stats for mpg
> mean(data$ mpg.Y., na.rm = TRUE)
[1] 23.50207
> median(data$mpg.Y, na.rm = TRUE)
[1] 22.75
> sd(data$mpg.Y, na.rm = TRUE)          # Standard deviation
[1] 7.821489
> var(data$mpg.Y, na.rm = TRUE)         # Variance
[1] 61.17568
> min(data$mpg.Y, na.rm = TRUE)
[1] 9
> max(data$mpg.Y, na.rm = TRUE)
[1] 46.6
> range(data$mpg.Y, na.rm = TRUE)
[1]  9.0 46.6
> quantile(data$mpg.Y, na.rm = TRUE)
   0%   25%   50%   75%  100%
 9.00 17.50 22.75 29.00 46.60
>
> # descriptive stats for cylinders
> mean(data$Cylinders, na.rm = TRUE)
[1] 5.463731
> median(data$Cylinders, na.rm = TRUE)
[1] 4
>
```

C:\Users\User\Downloads\Lecture 22.R - R Editor

```
# Load the dataset into the R environment
# Set the file path
file_path <- "C:/Users/User/Downloads/autompg.csv"

# Read the CSV file
data <- read.csv(file_path)

?data
data

# descriptive stats for mpg
mean(data$ mpg.Y., na.rm = TRUE)
median(data$mpg.Y, na.rm = TRUE)
sd(data$mpg.Y, na.rm = TRUE)          # Standard deviation
var(data$mpg.Y, na.rm = TRUE)         # Variance
min(data$mpg.Y, na.rm = TRUE)
max(data$mpg.Y, na.rm = TRUE)
range(data$mpg.Y, na.rm = TRUE)
quantile(data$mpg.Y, na.rm = TRUE)

# descriptive stats for cylinders
mean(data$Cylinders, na.rm = TRUE)
median(data$Cylinders, na.rm = TRUE)
sd(data$Cylinders, na.rm = TRUE)
var(data$Cylinders, na.rm = TRUE)
```

Descriptive stats on Cylinders:

RGui

File Edit View Misc Packages Windows Help

R Console

```
 9.00 17.50 22.75 29.00 46.60
>
> # descriptive stats for cylinders
> mean(data$Cylinders, na.rm = TRUE)
[1] 5.463731
> median(data$Cylinders, na.rm = TRUE)
[1] 4
> # descriptive stats for horsepower
> mean(data$HorsePower, na.rm = TRUE)
[1] 103.7617
> median(data$HorsePower, na.rm = TRUE)
[1] 92.5
> sd(data$HorsePower, na.rm = TRUE)
[1] 37.60264
> var(data$HorsePower, na.rm = TRUE)
[1] 1413.959
> min(data$HorsePower, na.rm = TRUE)
[1] 46
> max(data$HorsePower, na.rm = TRUE)
[1] 230
> range(data$HorsePower, na.rm = TRUE)
[1]  46 230
> quantile(data$HorsePower, na.rm = TRUE)
   0%   25%   50%   75%  100%
 46.0  75.0  92.5 125.0 230.0
```

C:\Users\User\Downloads\Lecture 22.R - R Editor

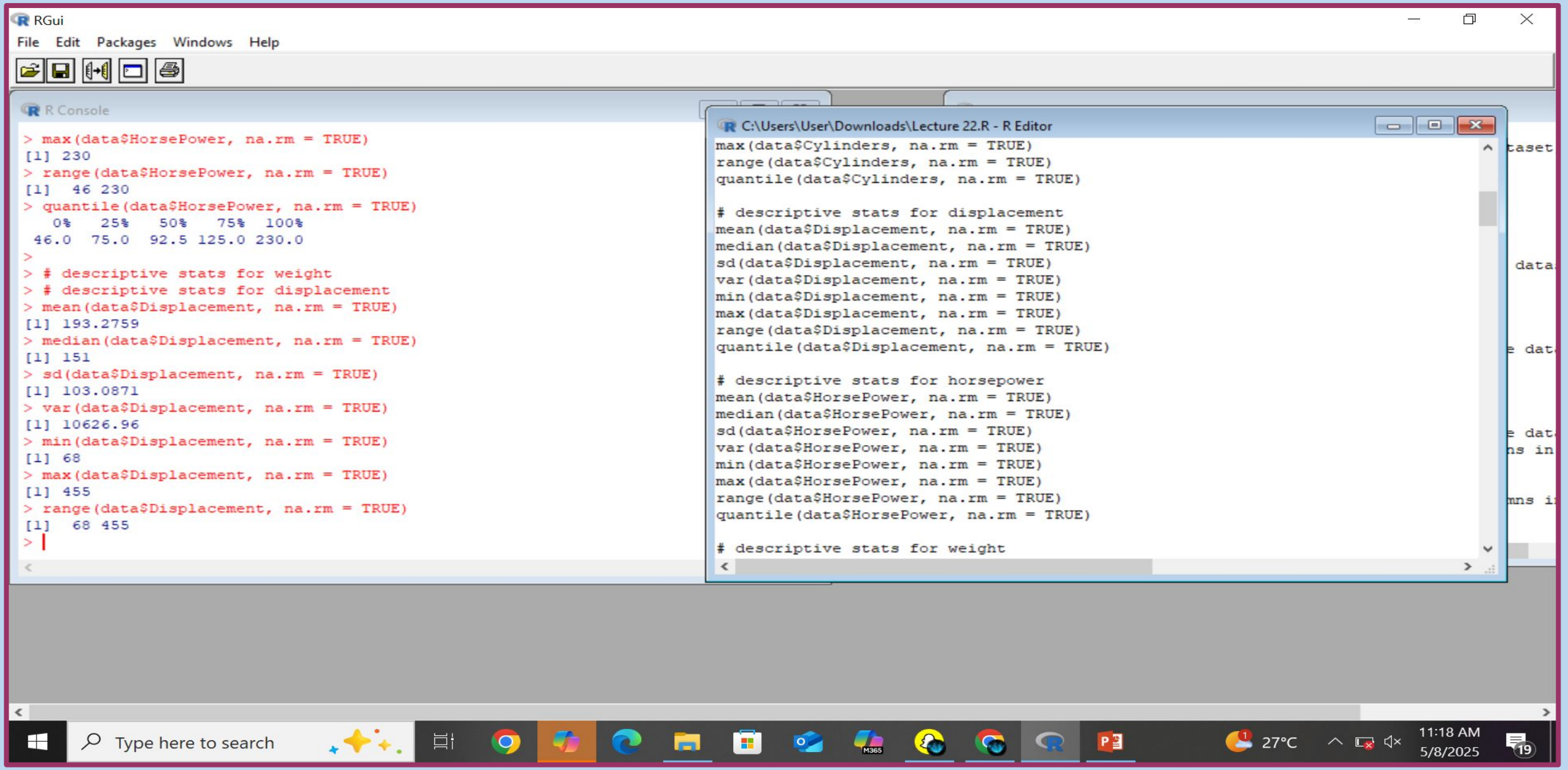
```
# descriptive stats for displacement
mean(data$Displacement, na.rm = TRUE)
median(data$Displacement, na.rm = TRUE)
sd(data$Displacement, na.rm = TRUE)
var(data$Displacement, na.rm = TRUE)
min(data$Displacement, na.rm = TRUE)
max(data$Displacement, na.rm = TRUE)
range(data$Displacement, na.rm = TRUE)
quantile(data$Displacement, na.rm = TRUE)

# descriptive stats for horsepower
mean(data$HorsePower, na.rm = TRUE)
median(data$HorsePower, na.rm = TRUE)
sd(data$HorsePower, na.rm = TRUE)
var(data$HorsePower, na.rm = TRUE)
min(data$HorsePower, na.rm = TRUE)
max(data$HorsePower, na.rm = TRUE)
range(data$HorsePower, na.rm = TRUE)
quantile(data$HorsePower, na.rm = TRUE)

# descriptive stats for weight
mean(data$Weight, na.rm = TRUE)
median(data$Weight, na.rm = TRUE)
sd(data$Weight, na.rm = TRUE)
var(data$Weight, na.rm = TRUE)
```

Type here to search

27°C

11:17 AM
5/8/2025

Descriptive stats on Displacement:



Descriptive stats on Horsepower:

RGui

File Edit Packages Windows Help

R Console

```
> min(data$Displacement, na.rm = TRUE)
[1] 68
> max(data$Displacement, na.rm = TRUE)
[1] 455
> range(data$Displacement, na.rm = TRUE)
[1]  68 455
> # descriptive stats for horsepower
> mean(data$HorsePower, na.rm = TRUE)
[1] 103.7617
> median(data$HorsePower, na.rm = TRUE)
[1] 92.5
> sd(data$HorsePower, na.rm = TRUE)
[1] 37.60264
> var(data$HorsePower, na.rm = TRUE)
[1] 1413.959
> min(data$HorsePower, na.rm = TRUE)
[1] 46
> max(data$HorsePower, na.rm = TRUE)
[1] 230
> range(data$HorsePower, na.rm = TRUE)
[1]  46 230
> quantile(data$HorsePower, na.rm = TRUE)
   0%   25%   50%   75%  100%
 46.0  75.0  92.5 125.0 230.0
>
```

C:\Users\User\Downloads\Lecture 22.R - R Editor

```
max(data$Cylinders, na.rm = TRUE)
range(data$Cylinders, na.rm = TRUE)
quantile(data$Cylinders, na.rm = TRUE)

# descriptive stats for displacement
mean(data$Displacement, na.rm = TRUE)
median(data$Displacement, na.rm = TRUE)
sd(data$Displacement, na.rm = TRUE)
var(data$Displacement, na.rm = TRUE)
min(data$Displacement, na.rm = TRUE)
max(data$Displacement, na.rm = TRUE)
range(data$Displacement, na.rm = TRUE)
quantile(data$Displacement, na.rm = TRUE)

# descriptive stats for horsepower
mean(data$HorsePower, na.rm = TRUE)
median(data$HorsePower, na.rm = TRUE)
sd(data$HorsePower, na.rm = TRUE)
var(data$HorsePower, na.rm = TRUE)
min(data$HorsePower, na.rm = TRUE)
max(data$HorsePower, na.rm = TRUE)
range(data$HorsePower, na.rm = TRUE)
quantile(data$HorsePower, na.rm = TRUE)

# descriptive stats for weight
```

Type here to search

27°C 11:18 AM 5/8/2025

Descriptive stats on Weight:

RGui

File Edit Packages Windows Help

R Console

```
> range(data$HorsePower, na.rm = TRUE)
[1]  46 230
> quantile(data$HorsePower, na.rm = TRUE)
   0%   25%   50%   75%  100%
 46.0  75.0  92.5 125.0 230.0
> # descriptive stats for weight
> mean(data$Weight, na.rm = TRUE)
[1] 2972.674
> median(data$Weight, na.rm = TRUE)
[1] 2803.5
> sd(data$Weight, na.rm = TRUE)
[1] 847.1126
> var(data$Weight, na.rm = TRUE)
[1] 717599.8
> min(data$Weight, na.rm = TRUE)
[1] 1613
> max(data$Weight, na.rm = TRUE)
[1] 5140
> range(data$Weight, na.rm = TRUE)
[1] 1613 5140
> quantile(data$Weight, na.rm = TRUE)
     0%     25%     50%     75%    100%
1613.00 2223.75 2803.50 3608.00 5140.00
>
>
```

C:\Users\User\Downloads\Lecture 22.R - R Editor

```
max(data$Displacement, na.rm = TRUE)
range(data$Displacement, na.rm = TRUE)
quantile(data$Displacement, na.rm = TRUE)

# descriptive stats for horsepower
mean(data$HorsePower, na.rm = TRUE)
median(data$HorsePower, na.rm = TRUE)
sd(data$HorsePower, na.rm = TRUE)
var(data$HorsePower, na.rm = TRUE)
min(data$HorsePower, na.rm = TRUE)
max(data$HorsePower, na.rm = TRUE)
range(data$HorsePower, na.rm = TRUE)
quantile(data$HorsePower, na.rm = TRUE)

# descriptive stats for weight
mean(data$Weight, na.rm = TRUE)
median(data$Weight, na.rm = TRUE)
sd(data$Weight, na.rm = TRUE)
var(data$Weight, na.rm = TRUE)
min(data$Weight, na.rm = TRUE)
max(data$Weight, na.rm = TRUE)
range(data$Weight, na.rm = TRUE)
quantile(data$Weight, na.rm = TRUE)

# descriptive stats for acceleration
```

Type here to search

27°C

11:18 AM
5/8/2025

Descriptive stats on Acceleration:

RGui

File Edit Packages Windows Help

R Console

```
[1] 1613 5140
> quantile(data$Weight, na.rm = TRUE)
     0%     25%     50%     75%    100%
1613.00 2223.75 2803.50 3608.00 5140.00
>
> # descriptive stats for acceleration
> mean(data$acceleration, na.rm = TRUE)
[1] 15.59119
> median(data$acceleration, na.rm = TRUE)
[1] 15.5
> sd(data$acceleration, na.rm = TRUE)
[1] 2.715989
> var(data$acceleration, na.rm = TRUE)
[1] 7.376598
> min(data$acceleration, na.rm = TRUE)
[1] 8
> max(data$acceleration, na.rm = TRUE)
[1] 24.8
> range(data$acceleration, na.rm = TRUE)
[1]  8.0 24.8
> quantile(data$acceleration, na.rm = TRUE)
    0%    25%    50%    75%   100%
 8.000 13.825 15.500 17.075 24.800
>
>
```

C:\Users\User\Downloads\Lecture 22.R - R Editor

```
max(data$HorsePower, na.rm = TRUE)
range(data$HorsePower, na.rm = TRUE)
quantile(data$HorsePower, na.rm = TRUE)

# descriptive stats for weight
mean(data$Weight, na.rm = TRUE)
median(data$Weight, na.rm = TRUE)
sd(data$Weight, na.rm = TRUE)
var(data$Weight, na.rm = TRUE)
min(data$Weight, na.rm = TRUE)
max(data$Weight, na.rm = TRUE)
range(data$Weight, na.rm = TRUE)
quantile(data$Weight, na.rm = TRUE)

# descriptive stats for acceleration
mean(data$acceleration, na.rm = TRUE)
median(data$acceleration, na.rm = TRUE)
sd(data$acceleration, na.rm = TRUE)
var(data$acceleration, na.rm = TRUE)
min(data$acceleration, na.rm = TRUE)
max(data$acceleration, na.rm = TRUE)
range(data$acceleration, na.rm = TRUE)
quantile(data$acceleration, na.rm = TRUE)

# descriptive stats for modelyear
```

Type here to search

27°C 11:18 AM 5/8/2025

Descriptive stats on Model year:

RGui

File Edit Packages Windows Help

R Console

```
> range(data$acceleration, na.rm = TRUE)
[1]  8.0 24.8
> quantile(data$acceleration, na.rm = TRUE)
    0%    25%    50%    75%   100%
 8.000 13.825 15.500 17.075 24.800
>
> # descriptive stats for modelyear
> mean(data$ModelYear, na.rm = TRUE)
[1] 76.07254
> median(data$ModelYear, na.rm = TRUE)
[1] 76
> sd(data$ModelYear, na.rm = TRUE)
[1] 3.635313
> var(data$ModelYear, na.rm = TRUE)
[1] 13.2155
> min(data$ModelYear, na.rm = TRUE)
[1] 70
> max(data$ModelYear, na.rm = TRUE)
[1] 82
> range(data$ModelYear, na.rm = TRUE)
[1] 70 82
> quantile(data$ModelYear, na.rm = TRUE)
  0%  25%  50%  75% 100%
  70   73   76   79   82
>
```

C:\Users\User\Downloads\Lecture 22.R - R Editor

```
min(data$Weight, na.rm = TRUE)
max(data$Weight, na.rm = TRUE)
range(data$Weight, na.rm = TRUE)
quantile(data$Weight, na.rm = TRUE)

# descriptive stats for acceleration
mean(data$acceleration, na.rm = TRUE)
median(data$acceleration, na.rm = TRUE)
sd(data$acceleration, na.rm = TRUE)
var(data$acceleration, na.rm = TRUE)
min(data$acceleration, na.rm = TRUE)
max(data$acceleration, na.rm = TRUE)
range(data$acceleration, na.rm = TRUE)
quantile(data$acceleration, na.rm = TRUE)

# descriptive stats for modelyear
mean(data$ModelYear, na.rm = TRUE)
median(data$ModelYear, na.rm = TRUE)
sd(data$ModelYear, na.rm = TRUE)
var(data$ModelYear, na.rm = TRUE)
min(data$ModelYear, na.rm = TRUE)
max(data$ModelYear, na.rm = TRUE)
range(data$ModelYear, na.rm = TRUE)
quantile(data$ModelYear, na.rm = TRUE)
```

Type here to search 27°C 11:19 AM 5/8/2025

Descriptive stats on origin:

RGui

File Edit Packages Windows Help

R Console

```
> range(data$ModelYear, na.rm = TRUE)
[1] 70 82
> quantile(data$ModelYear, na.rm = TRUE)
  0%  25%  50%  75% 100%
  70   73   76   79   82
>
> # descriptive stats for origin
> mean(data$origin, na.rm = TRUE)
[1] 1.572539
> median(data$origin, na.rm = TRUE)
[1] 1
> sd(data$origin, na.rm = TRUE)
[1] 0.8033516
> var(data$origin, na.rm = TRUE)
[1] 0.6453738
> min(data$origin, na.rm = TRUE)
[1] 1
> max(data$origin, na.rm = TRUE)
[1] 3
> range(data$origin, na.rm = TRUE)
[1] 1 3
> quantile(data$origin, na.rm = TRUE)
  0%  25%  50%  75% 100%
   1    1    1    2    3
>
```

C:\Users\User\Downloads\Lecture 22.R - R Editor

```
min(data$acceleration, na.rm = TRUE)
max(data$acceleration, na.rm = TRUE)
range(data$acceleration, na.rm = TRUE)
quantile(data$acceleration, na.rm = TRUE)

# descriptive stats for modelyear
mean(data$ModelYear, na.rm = TRUE)
median(data$ModelYear, na.rm = TRUE)
sd(data$ModelYear, na.rm = TRUE)
var(data$ModelYear, na.rm = TRUE)
min(data$ModelYear, na.rm = TRUE)
max(data$ModelYear, na.rm = TRUE)
range(data$ModelYear, na.rm = TRUE)
quantile(data$ModelYear, na.rm = TRUE)

# descriptive stats for origin
mean(data$origin, na.rm = TRUE)
median(data$origin, na.rm = TRUE)
sd(data$origin, na.rm = TRUE)
var(data$origin, na.rm = TRUE)
min(data$origin, na.rm = TRUE)
max(data$origin, na.rm = TRUE)
range(data$origin, na.rm = TRUE)
quantile(data$origin, na.rm = TRUE)
```

Type here to search

27°C

11:19 AM
5/8/2025

Summary of Descriptive Stats:

RGui

File Edit View Misc Packages Windows Help

R Console

```
> # ==================================
> # Task 3: Descriptive Statistics
> # ==================================
>
> library(psych)
>
> #Shows Summary of all Data - detailed descriptive stats
> describe(data)
              vars   n    mean     sd  median trimmed    mad  min    max  range
mpg.Y.           1 386   23.50   7.82   22.75   23.06   8.60    9   46.6   37.6
Cylinders        2 386    5.46   1.70    4.00    5.35   0.00    3    8.0    5.0
Displacement     3 386  193.28 103.09  151.00  183.20  90.44   68  455.0  387.0
HorsePower       4 386  103.76  37.60   92.50   99.44  28.91   46  230.0  184.0
Weight           5 386 2972.67 847.11 2803.50 2911.62 945.16 1613 5140.0 3527.0
acceleration     6 386   15.59   2.72   15.50   15.51   2.45    8   24.8   16.8
ModelYear        7 386   76.07   3.64   76.00   76.07   4.45   70   82.0   12.0
origin           8 386    1.57   0.80    1.00    1.47   0.00    1    3.0    2.0
CarName*         9 386  146.05  88.33  147.00  145.82 118.61    1  299.0  298.0
              skew kurtosis    se
mpg.Y.        0.45    -0.55  0.40
Cylinders     0.51    -1.39  0.09
Displacement  0.68    -0.82  5.25
HorsePower    1.05     0.61  1.91
Weight        0.52    -0.80 43.12
acceleration  0.36     0.37  0.14
ModelYear     0.01    -1.17  0.19
origin        0.92    -0.84  0.04
CarName*      0.03    -1.26  4.50
>
> # Quick Summary descriptive stats
> summary(data)
     mpg.Y.        Cylinders      Displacement     HorsePower        Weight
 Min.   : 9.00   Min.   :3.000   Min.   : 68.0   Min.   : 46.0   Min.   :1613
 1st Qu.:17.50   1st Qu.:4.000   1st Qu.:105.0   1st Qu.: 75.0   1st Qu.:2224
 Median :22.75   Median :4.000   Median :151.0   Median : 92.5   Median :2804
 Mean   :23.50   Mean   :5.464   Mean   :193.3   Mean   :103.8   Mean   :2973
 3rd Qu.:29.00   3rd Qu.:8.000   3rd Qu.:262.0   3rd Qu.:125.0   3rd Qu.:3608
 Max.   :46.60   Max.   :8.000   Max.   :455.0   Max.   :230.0   Max.   :5140
  acceleration     ModelYear        origin       CarName
 Min.   : 8.00   Min.   :70.00   Min.   :1.000   Length:386
 1st Qu.:13.82   1st Qu.:73.00   1st Qu.:1.000   Class :character
 Median :15.50   Median :76.00   Median :1.000   Mode  :character
```

C:\Users\HP\Downloads\Lecture 22.R - R Editor

```
# ==================================
# Task 3: Descriptive Statistics
# ==================================

library(psych)

#Shows Summary of all Data - detailed descriptive stats
describe(data)

# Quick Summary descriptive stats
summary(data)


# descriptive stats for mpg
mean(data$ mpg.Y., na.rm = TRUE)
median(data$mpg.Y, na.rm = TRUE)
sd(data$mpg.Y, na.rm = TRUE)          # Standard deviation
var(data$mpg.Y, na.rm = TRUE)         # Variance
min(data$mpg.Y, na.rm = TRUE)
max(data$mpg.Y, na.rm = TRUE)
range(data$mpg.Y, na.rm = TRUE)
quantile(data$mpg.Y, na.rm = TRUE)

# descriptive stats for cylinders
mean(data$Cylinders, na.rm = TRUE)
median(data$Cylinders, na.rm = TRUE)
sd(data$Cylinders, na.rm = TRUE)
var(data$Cylinders, na.rm = TRUE)
min(data$Cylinders, na.rm = TRUE)
max(data$Cylinders, na.rm = TRUE)
range(data$Cylinders, na.rm = TRUE)
quantile(data$Cylinders, na.rm = TRUE)

# descriptive stats for displacement
mean(data$Displacement, na.rm = TRUE)
median(data$Displacement, na.rm = TRUE)
sd(data$Displacement, na.rm = TRUE)
```

Search

8:03 PM
5/8/2025

Display the default number of rows(1st 6 rows) from the dataset:

RGui

File Edit Packages Windows Help

R Console

```
398      1                    ford granada l
399      3                  toyota celica gt
400      1                 dodge charger 2.2
401      1                  chevrolet camaro
402      1                   ford mustang gl
403      2                         vw pickup
404      1                    dodge rampage
405      1                      ford ranger
406      1                       chevy s-10
> head(data)
  mpg.Y. Cylinders Displacement HorsePower Weight acceleration ModelYear origin
1     18         8          307        130   3504         12.0         70      1
2     15         8          350        165   3693         11.5         70      1
3     18         8          318        150   3436         11.0         70      1
4     16         8          304        150   3433         12.0         70      1
5     17         8          302        140   3449         10.5         70      1
6     15         8          429        198   4341         10.0         70      1
                  CarName
1 chevrolet chevelle malibu
2         buick skylark 320
3        plymouth satellite
4             amc rebel sst
5               ford torino
6          ford galaxie 500
>
```

C:\Users\User\Downloads\Lecture 22.R - R Editor

```
# Load the dataset into the R environment
# Set the file path
file_path <- "C:/Users/User/Downloads/autompg.csv"

# Read the CSV file
data <- read.csv(file_path)

?data
data

# Display the default number of rows(1st 6 rows) from the dataset
head(data)
# Display the first 3 rows of the dataset
head(data,3)

# Display the default number of rows(last 6 rows) from the dataset
tail(data)
# Display the first 4 rows of the dataset
tail(data,4)

# Use the 'dim()' function to obtain the dimensions of the dataset
# This returns a vector with the number of rows and columns in dataset
dim(data)

# Use the 'names()' function to get the names of the columns in the da
```

Type here to search

Rain...

7:12 PM
5/7/2025

Display the default number of rows(last 6 rows) from the dataset

RGui

File Edit Packages Windows Help

R Console

```
5     17         8          302         140   3449         10.5         70       1
6     15         8          429         198   4341         10.0         70       1
                   CarName
1 chevrolet chevelle malibu
2         buick skylark 320
3        plymouth satellite
4             amc rebel sst
5               ford torino
6          ford galaxie 500
> tail(data)
    mpg.Y. Cylinders Displacement HorsePower Weight acceleration ModelYear
401     27         4          151         90   2950         17.3         82
402     27         4          140         86   2790         15.6         82
403     44         4           97         52   2130         24.6         82
404     32         4          135         84   2295         11.6         82
405     28         4          120         79   2625         18.6         82
406     31         4          119         82   2720         19.4         82
    origin          CarName
401      1 chevrolet camaro
402      1  ford mustang gl
403      2        vw pickup
404      1    dodge rampage
405      1      ford ranger
406      1       chevy s-10
> 
```

C:\Users\User\Downloads\Lecture 22.R - R Editor

```
# Load the dataset into the R environment
# Set the file path
file_path <- "C:/Users/User/Downloads/autompg.csv"

# Read the CSV file
data <- read.csv(file_path)

?data
data

# Display the default number of rows(1st 6 rows) from the dataset
head(data)
# Display the first 3 rows of the dataset
head(data,3)

# Display the default number of rows(last 6 rows) from the dataset
tail(data)
# Display the first 4 rows of the dataset
tail(data,4)

# Use the 'dim()' function to obtain the dimensions of the dataset
# This returns a vector with the number of rows and columns in dataset
dim(data)

# Use the 'names()' function to get the names of the columns in the da
```

Type here to search

Rain...

7:13 PM
5/7/2025

Dimensions:

RGui

File Edit Packages Windows Help

R Console

```
                    CarName
1 chevrolet chevelle malibu
2          buick skylark 320
3         plymouth satellite
4              amc rebel sst
5                ford torino
6           ford galaxie 500
> tail(data)
    mpg.Y. Cylinders Displacement HorsePower Weight acceleration ModelYear
401     27         4          151         90   2950         17.3         82
402     27         4          140         86   2790         15.6         82
403     44         4           97         52   2130         24.6         82
404     32         4          135         84   2295         11.6         82
405     28         4          120         79   2625         18.6         82
406     31         4          119         82   2720         19.4         82
    origin           CarName
401      1 chevrolet camaro
402      1  ford mustang gl
403      2         vw pickup
404      1     dodge rampage
405      1       ford ranger
406      1        chevy s-10
> dim(data)
[1] 406   9
>
```

C:\Users\User\Downloads\Lecture 22.R - R Editor

```
head(data)
# Display the first 3 rows of the dataset
head(data,3)

# Display the default number of rows(last 6 rows) from the dataset
tail(data)
# Display the first 4 rows of the dataset
tail(data,4)

# Use the 'dim()' function to obtain the dimensions of the dataset
# This returns a vector with the number of rows and columns in dataset
dim(data)

# Use the 'names()' function to get the names of the columns in the da
names(data)

# Check the structure of the iris dataset
str(data)

# Use the 'summary()' function to generate a statistical summary of th
# This provides a quick overview of the data, including min, max, mean
summary(data)

# Missing Values in iris dataset
# Total number of missing values in iris dataset
```

Type here to search

Rain...

7:14 PM
5/7/2025

Names of variables in dataset:

RGui

File Edit Packages Windows Help

R Console

```
3         plymouth satellite
4              amc rebel sst
5                ford torino
6           ford galaxie 500
> tail(data)
    mpg.Y. Cylinders Displacement HorsePower Weight acceleration ModelYear
401     27         4          151         90   2950         17.3         82
402     27         4          140         86   2790         15.6         82
403     44         4           97         52   2130         24.6         82
404     32         4          135         84   2295         11.6         82
405     28         4          120         79   2625         18.6         82
406     31         4          119         82   2720         19.4         82
    origin           CarName
401      1 chevrolet camaro
402      1  ford mustang gl
403      2         vw pickup
404      1    dodge rampage
405      1      ford ranger
406      1       chevy s-10
> dim(data)
[1] 406   9
> names(data)
[1] "mpg.Y."       "Cylinders"    "Displacement" "HorsePower"   "Weight"
[6] "acceleration" "ModelYear"    "origin"       "CarName"
>
```

C:\Users\User\Downloads\Lecture 22.R - R Editor

```
head(data)
# Display the first 3 rows of the dataset
head(data,3)

# Display the default number of rows(last 6 rows) from the dataset
tail(data)
# Display the first 4 rows of the dataset
tail(data,4)

# Use the 'dim()' function to obtain the dimensions of the dataset
# This returns a vector with the number of rows and columns in dataset
dim(data)

# Use the 'names()' function to get the names of the columns in the da
names(data)

# Check the structure of the iris dataset
str(data)

# Use the 'summary()' function to generate a statistical summary of th
# This provides a quick overview of the data, including min, max, mean
summary(data)

# Missing Values in iris dataset
# Total number of missing values in iris dataset
```

Type here to search

Rain...

7:15 PM
5/7/2025

Structure of dataset:

RGui

File Edit Packages Windows Help

R Console

```
406     31         4          119         82   2720         19.4         82
    origin          CarName
401      1 chevrolet camaro
402      1  ford mustang gl
403      2         vw pickup
404      1    dodge rampage
405      1      ford ranger
406      1       chevy s-10
> dim(data)
[1] 406   9
> names(data)
[1] "mpg.Y."       "Cylinders"    "Displacement" "HorsePower"   "Weight"
[6] "acceleration" "ModelYear"    "origin"       "CarName"
> str(data)
'data.frame':   406 obs. of  9 variables:
 $ mpg.Y.       : num  18 15 18 16 17 15 14 14 15 16 ...
 $ Cylinders    : int  8 8 8 8 8 8 8 8 8 8 ...
 $ Displacement: num  307 350 318 304 302 429 454 440 350 304 ...
 $ HorsePower   : int  130 165 150 150 140 198 220 215 165 150 ...
 $ Weight       : int  3504 3693 3436 3433 3449 4341 4354 4312 3693 3433 ...
 $ acceleration: num  12 11.5 11 12 10.5 10 9 8.5 11.5 12 ...
 $ ModelYear    : int  70 70 70 70 70 70 70 70 70 70 ...
 $ origin       : int  1 1 1 1 1 1 1 1 1 1 ...
 $ CarName      : chr  "chevrolet chevelle malibu" "buick skylark 320" "plymouth$
>
```

C:\Users\User\Downloads\Lecture 22.R - R Editor

```
head(data)
# Display the first 3 rows of the dataset
head(data,3)

# Display the default number of rows(last 6 rows) from the dataset
tail(data)
# Display the first 4 rows of the dataset
tail(data,4)

# Use the 'dim()' function to obtain the dimensions of the dataset
# This returns a vector with the number of rows and columns in dataset
dim(data)

# Use the 'names()' function to get the names of the columns in the da
names(data)

# Check the structure of the iris dataset
str(data)

# Use the 'summary()' function to generate a statistical summary of th
# This provides a quick overview of the data, including min, max, mean
summary(data)

# Missing Values in iris dataset
# Total number of missing values in iris dataset
```

Task 4: Data Prepossessing

Displaying dataset:

RGui

File Edit View Misc Packages Windows Help

R Console

```
> # Read the CSV file
> data <- read.csv(file_path)
>
> ?data
> data
    mpg.Y. Cylinders Displacement HorsePower Weight acceleration ModelYear
1     18.0         8        307.0        130   3504         12.0         70
2     15.0         8        350.0        165   3693         11.5         70
3     18.0         8        318.0        150   3436         11.0         70
4     16.0         8        304.0        150   3433         12.0         70
5     17.0         8        302.0        140   3449         10.5         70
6     15.0         8        429.0        198   4341         10.0         70
7     14.0         8        454.0        220   4354          9.0         70
8     14.0         8        440.0        215   4312          8.5         70
9     15.0         8        350.0        165   3693         11.5         70
10    16.0         8        304.0        150   3433         12.0         70
11      NA         4        133.0        115   3090         17.5         70
12      NA         8        350.0        165   4142         11.5         70
13      NA         8        351.0        153   4034         11.0         70
14      NA         8        383.0        175   4166         10.5         70
15      NA         8        360.0        175   3850         11.0         70
16    15.0         8        383.0        170   3563         10.0         70
17    14.0         8        340.0        160   3609          8.0         70
18      NA         8        302.0        140   3353          8.0         70
19    15.0         8        400.0        150   3761          9.5         70
```

C:\Users\User\Downloads\Lecture 22.R - R Editor

```
# Load the dataset into the R environment
# Set the file path
file_path <- "C:/Users/User/Downloads/autompg.csv"

# Read the CSV file
data <- read.csv(file_path)

?data
data

# Display the default number of rows(lst 6 rows) from the dataset
head(data)
# Display the first 3 rows of the dataset
head(data,3)

# Display the default number of rows(last 6 rows) from the dataset
tail(data)
# Display the first 4 rows of the dataset
tail(data,4)

# Use the 'dim()' function to obtain the dimensions of the dataset
# This returns a vector with the number of rows and columns in dataset
dim(data)

# Use the 'names()' function to get the names of the columns in the da
```

Type here to search

Rain...

7:11 PM
5/7/2025

Missing values in dataset:

RGui

File Edit Packages Windows Help

R Console

```
 1st Qu.:17.00   1st Qu.:4.000   1st Qu.:105   1st Qu.: 75.75   1st Qu.:2228
 Median :22.45   Median :4.000   Median :151   Median : 95.00   Median :2845
 Mean   :23.44   Mean   :5.505   Mean   :196   Mean   :105.36   Mean   :2986
 3rd Qu.:29.00   3rd Qu.:8.000   3rd Qu.:302   3rd Qu.:130.00   3rd Qu.:3628
 Max.   :46.60   Max.   :8.000   Max.   :455   Max.   :230.00   Max.   :5140
 NA's   :8                                     NA's   :6
  acceleration     ModelYear         origin        CarName
 Min.   : 8.00   Min.   :70.00   Min.   :1.000   Length:406
 1st Qu.:13.53   1st Qu.:73.00   1st Qu.:1.000   Class :character
 Median :15.50   Median :76.00   Median :1.000   Mode  :character
 Mean   :15.50   Mean   :75.92   Mean   :1.557
 3rd Qu.:17.07   3rd Qu.:79.00   3rd Qu.:2.000
 Max.   :24.80   Max.   :82.00   Max.   :3.000

> # Missing Values in dataset
> # Total number of missing values in  dataset
> sum(is.na(data))
[1] 14
> # Missing values per column in  dataset
> colSums(is.na(data))
      mpg.Y.    Cylinders Displacement   HorsePower       Weight acceleration
           8            0            0            6            0            0
   ModelYear       origin      CarName
           0            0            0
>
```

C:\Users\User\Downloads\Lecture 22.R - R Editor

```
str(data)

# Use the 'summary()' function to generate a statistical summary of th
# This provides a quick overview of the data, including min, max, mean
summary(data)

# Missing Values in dataset
# Total number of missing values in  dataset
sum(is.na(data))
# Missing values per column in  dataset
colSums(is.na(data))

# Duplicate rows in
# Number of duplicate rows in  dataset
sum(duplicated(data))
# Shows duplicate rows in
duplicated_data <- data[duplicated(data), ]
duplicated_data

# # Assuming your dataset is called data
data$CarName_numeric <- as.numeric(factor(data$CarName))

# Check the result
head(data$CarName_numeric)
```

Duplicates in dataset:

RGui

File Edit Packages Windows Help

R Console

```
> colSums(is.na(data))
      mpg.Y.    Cylinders Displacement   HorsePower       Weight acceleration
           8            0            0            6            0            0
   ModelYear       origin      CarName
           0            0            0
> # Duplicate rows in
> # Number of duplicate rows in  dataset
> sum(duplicated(data))
[1] 5
> # Shows duplicate rows in
> duplicated_data <- data[duplicated(data), ]
> duplicated_data
   mpg.Y. Cylinders Displacement HorsePower Weight acceleration ModelYear
9      15         8          350        165   3693         11.5        70
10     16         8          304        150   3433         12.0        70
21     15         8          350        165   3693         11.5        70
25     15         8          350        165   3693         11.5        70
27     16         8          304        150   3433         12.0        70
   origin           CarName
9       1 buick skylark 320
10      1      amc rebel sst
21      1 buick skylark 320
25      1 buick skylark 320
27      1      amc rebel sst
>
```

C:\Users\User\Downloads\Lecture 22.R - R Editor

```
str(data)

# Use the 'summary()' function to generate a statistical summary of th
# This provides a quick overview of the data, including min, max, mean
summary(data)

# Missing Values in dataset
# Total number of missing values in  dataset
sum(is.na(data))
# Missing values per column in  dataset
colSums(is.na(data))

# Duplicate rows in
# Number of duplicate rows in  dataset
sum(duplicated(data))
# Shows duplicate rows in
duplicated_data <- data[duplicated(data), ]
duplicated_data

# # Assuming your dataset is called data
data$CarName_numeric <- as.numeric(factor(data$CarName))

# Check the result
head(data$CarName_numeric)
```

Convert categorial variables to numeric:

RGui - [R Console]

File Edit View Misc Packages Windows Help

```
[306] "vw dasher (diesel)"
[307] "vw pickup"
[308] "vw rabbit"
[309] "vw rabbit c (diesel)"
[310] "vw rabbit custom"
> # Assuming your dataset is called data
> data$CarName_numeric <- as.numeric(factor(data$CarName))
>
> # Check how car names are mapped to numbers
> levels(factor(data$CarName))
  [1] "amc ambassador brougham"
  [2] "amc ambassador sst"
  [3] "amc concord"
  [4] "amc concord d/l"
  [5] "amc concord dl"
  [6] "amc concord dl 6"
  [7] "amc gremlin"
  [8] "amc hornet"
  [9] "amc hornet sportabout (sw)"
 [10] "amc matador"
 [11] "amc matador (sw)"
 [12] "amc pacer"
 [13] "amc pacer d/l"
 [14] "amc rebel sst"
 [15] "amc rebel sst (sw)"
 [16] "amc spirit dl"
 [17] "audi 100 ls"
 [18] "audi 100ls"
 [19] "audi 4000"
 [20] "audi 5000"
 [21] "audi 5000s (diesel)"
 [22] "audi fox"
 [23] "bmw 2002"
 [24] "bmw 320i"
 [25] "buick century"
```

RGui - [R Console]

File Edit View Misc Packages Windows Help

```
81      1    chevrolet chevelle concours (sw)           49
82      1            ford gran torino (sw)             145
83      1      plymouth satellite custom (sw)          238
84      2                  volvo 145e (sw)             301
85      2              volkswagen 411 (sw)             288
86      2                peugeot 504 (sw)              216
87      2                renault 12 (sw)               258
88      1                ford pinto (sw)               160
89      3                datsun 510 (sw)               87
90      3         toyouta corona mark ii (sw)          284
91      1                dodge colt (sw)               106
92      3           toyota corolla 1600 (sw)           274
93      1                buick century 350             26
94      1                    amc matador               10
95      1               chevrolet malibu               56
96      1               ford gran torino               144
97      1             dodge coronet custom             111
98      1          mercury marquis brougham            197
99      1          chevrolet caprice classic           45
100     1                      ford ltd                 150
101     1          plymouth fury gran sedan            226
102     1        chrysler new yorker brougham          75
103     1          buick electra 225 custom            30
104     1           amc ambassador brougham            1
105     1                plymouth valiant              240
106     1             chevrolet nova custom            63
107     1                    amc hornet                8
108     1                ford maverick                 152
109     1               plymouth duster                224
110     2            volkswagen super beetle           297
111     1                chevrolet impala              55
112     1                ford country                  130
113     1             plymouth custom suburb           223
114     1            oldsmobile vista cruiser          211
115     1                    amc gremlin               7
```

Standardization:

RGui

File Edit View Misc Packages Windows Help

R Console

```
> # View the first 10 rows
> head(data_scaled, 10)
       mpg.Y. Cylinders Displacement HorsePower    Weight acceleration
1  -0.6943935  1.449815     1.062007  0.6396594 0.6143182    -1.251712
2  -1.0770342  1.449815     1.473392  1.5481743 0.8382717    -1.430403
3  -0.6943935  1.449815     1.167245  1.1588108 0.5337423    -1.609093
4  -0.9494873  1.449815     1.033305  1.1588108 0.5301875    -1.251712
5  -0.8219404  1.449815     1.014171  0.8992351 0.5491465    -1.787783
6  -1.0770342  1.449815     2.229193  2.4047741 1.6061125    -1.966473
7  -1.2045811  1.449815     2.468370  2.9758407 1.6215167    -2.323854
8  -1.2045811  1.449815     2.334431  2.8460528 1.5717493    -2.502544
9  -1.0770342  1.449815     1.473392  1.5481743 0.8382717    -1.430403
10 -0.9494873  1.449815     1.033305  1.1588108 0.5301875    -1.251712
   ModelYear     origin
1  -1.579514 -0.7026267
2  -1.579514 -0.7026267
3  -1.579514 -0.7026267
4  -1.579514 -0.7026267
5  -1.579514 -0.7026267
6  -1.579514 -0.7026267
7  -1.579514 -0.7026267
8  -1.579514 -0.7026267
9  -1.579514 -0.7026267
10 -1.579514 -0.7026267
>
```

C:\Users\User\Downloads\Lecture 22.R - R Editor

```
# Check how car names are mapped to numbers
levels(factor(data$CarName))

# View the result
print(data)

# Standardize all 4 numeric features
data_scaled <- as.data.frame(scale(data[, 1:8]))
# View the result (standardized values for all 150 rows)
head(data_scaled, 406)  # View first 10 rows

# Standardize by column names
data_scaled <- as.data.frame(scale(data[, c("mpg.Y.", "Cylinders", "D
                                            "HorsePower", "Weight",
                                            "acceleration", "ModelYe

# View the first 10 rows
head(data_scaled, 10)

# Function to detect outliers using IQR method
find_outliers <- function(data) {
  Q1 <- quantile(data, 0.25)
  Q3 <- quantile(data, 0.75)
```

Type here to search

31°C

8:07 PM
5/7/2025

Outliers:

RGui

File Edit Packages Windows Help

R Console

```
> length(outliers_data_Weight)
[1] 0
> length(outliers_data_acceleration)
[1] 6
> length(outliers_data_ModelYear)
[1] 0
> length(outliers_data_origin)
[1] 0
>
>
> data[outliers_data_mpg.Y., "mpg.Y."]
Error: object 'outliers_data_mpg.Y.' not found
> data[outliers_data_Cylinders, "Cylinders"]
integer(0)
> data[outliers_data_Displacement, "Displacement"]
numeric(0)
> data[outliers_data_HorsePower, "HorsePower"]
Error: object 'outliers_data_HorsePower' not found
> data[outliers_data_Weight, "Weight"]
integer(0)
> data[outliers_data_acceleration, "acceleration"]
[1]  8.5  8.5 15.5 16.0 15.5 16.0
> data[outliers_data_ModelYear, "ModelYear"]
integer(0)
>
```

C:\Users\User\Downloads\Lecture 22.R - R Editor

```
outliers_data_acceleration <- find_outliers(data$acceleration)
outliers_data_ModelYear <- find_outliers(data$ModelYear)
outliers_data_origin <- find_outliers(data$origin)



length(outliers_data_mpg.Y.)
length(outliers_data_Cylinders)
length(outliers_data_Displacement)
length(outliers_data_HorsePower)
length(outliers_data_Weight)
length(outliers_data_acceleration)
length(outliers_data_ModelYear)
length(outliers_data_origin)


data[outliers_data_mpg.Y., "mpg.Y."]
data[outliers_data_Cylinders, "Cylinders"]
data[outliers_data_Displacement, "Displacement"]
data[outliers_data_HorsePower, "HorsePower"]
data[outliers_data_Weight, "Weight"]
data[outliers_data_acceleration, "acceleration"]
data[outliers_data_ModelYear, "ModelYear"]
data[outliers_data_origin, "origin"]
```

Type here to search

31°C

8:08 PM
5/7/2025

Correlation coefficient:

RGui

File Edit Packages Windows Help

R Console

```
>
> data[outliers_data_mpg.Y., "mpg.Y."]
Error: object 'outliers_data_mpg.Y.' not found
> data[outliers_data_Cylinders, "Cylinders"]
integer(0)
> data[outliers_data_Displacement, "Displacement"]
numeric(0)
> data[outliers_data_HorsePower, "HorsePower"]
Error: object 'outliers_data_HorsePower' not found
> data[outliers_data_Weight, "Weight"]
integer(0)
> data[outliers_data_acceleration, "acceleration"]
[1]  8.5  8.5 15.5 16.0 15.5 16.0
> data[outliers_data_ModelYear, "ModelYear"]
integer(0)
> # Correlation Coefficient
> # Calculate the Pearson correlation coefficient
> # Calculate Pearson correlation matrix for selected columns
> correlation_matrix <- cor(data[c("mpg.Y.", "Cylinders", "Displacement", "Hors$
> print(correlation_coefficient)
[1] 0.02509152
>
>
>
>
```

C:\Users\User\Downloads\Lecture 22.R - R Editor

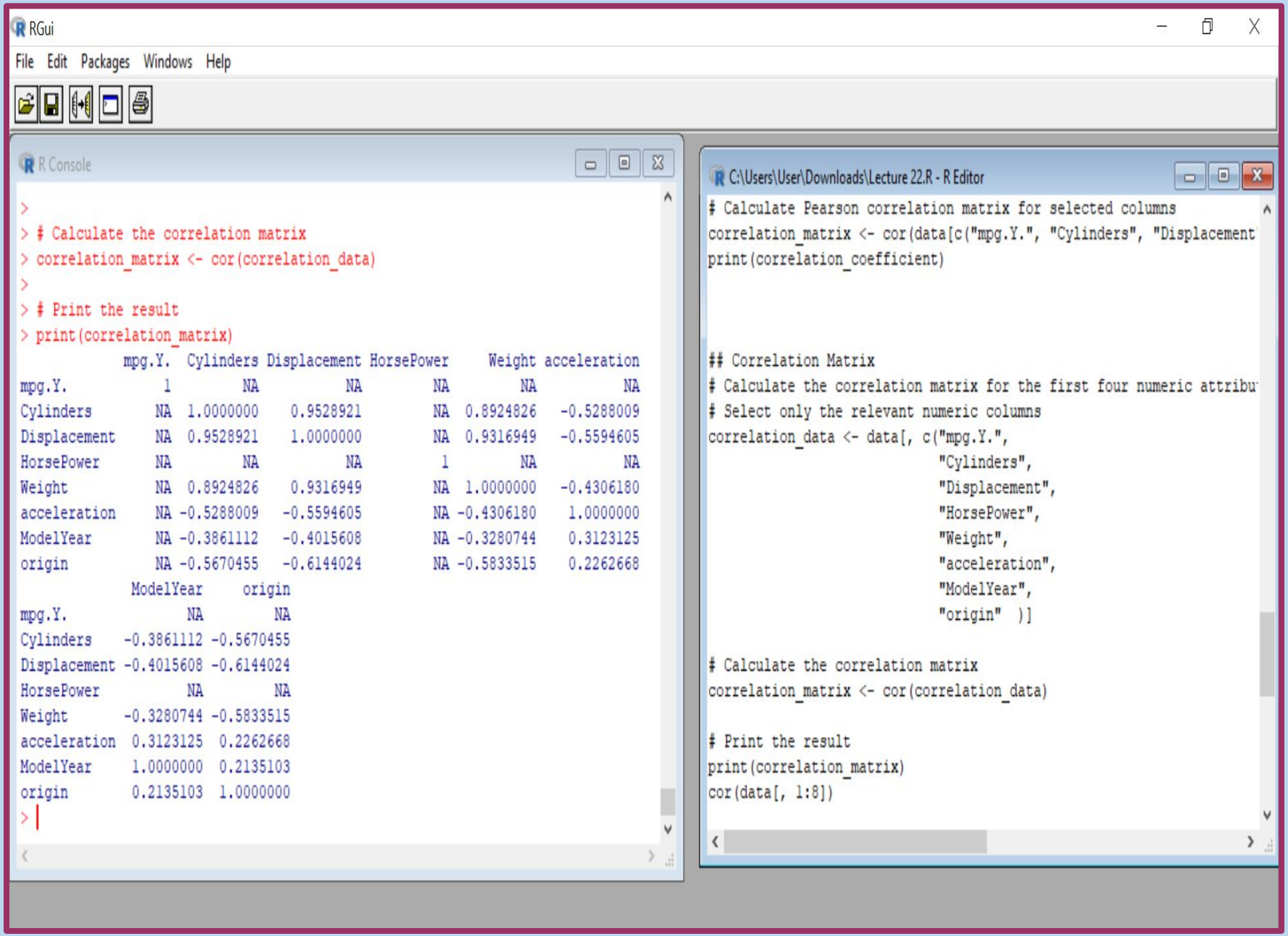
```
data[outliers_data_mpg.Y., "mpg.Y."]
data[outliers_data_Cylinders, "Cylinders"]
data[outliers_data_Displacement, "Displacement"]
data[outliers_data_HorsePower, "HorsePower"]
data[outliers_data_Weight, "Weight"]
data[outliers_data_acceleration, "acceleration"]
data[outliers_data_ModelYear, "ModelYear"]
data[outliers_data_origin, "origin"]




# Correlation Coefficient
# Calculate the Pearson correlation coefficient
# Calculate Pearson correlation matrix for selected columns
correlation_matrix <- cor(data[c("mpg.Y.", "Cylinders", "Displacement"
print(correlation_coefficient)



## Correlation Matrix
# Calculate the correlation matrix for the first four numeric attribu
# Select only the relevant numeric columns
```

Type here to search

31°C

8:08 PM
5/7/2025

Correlation Matrix:



The correlation matrix returned NA values due to missing data. To fix this, we used cor(data, use = "complete.obs") to ignore rows with missing values when calculating correlations.

Cleaning of dataset:

Dimensions changes as rows were deleted to handle missing values.

RGui

File Edit Packages Windows Help

R Console

```
    mpg.Y. Cylinders Displacement HorsePower Weight acceleration ModelYear
382     27         4          151         90   2950         17.3         82
383     27         4          140         86   2790         15.6         82
384     44         4           97         52   2130         24.6         82
385     32         4          135         84   2295         11.6         82
386     28         4          120         79   2625         18.6         82
387     31         4          119         82   2720         19.4         82
    origin          CarName
382      1 chevrolet camaro
383      1  ford mustang gl
384      2         vw pickup
385      1    dodge rampage
386      1      ford ranger
387      1       chevy s-10
> # Display the first 4 rows of the dataset
> tail(data,4)
    mpg.Y. Cylinders Displacement HorsePower Weight acceleration ModelYear
384     44         4           97         52   2130         24.6         82
385     32         4          135         84   2295         11.6         82
386     28         4          120         79   2625         18.6         82
387     31         4          119         82   2720         19.4         82
    origin       CarName
384      2     vw pickup
385      1 dodge rampage
386      1   ford ranger
387      1    chevy s-10
>
> # Use the 'dim()' function to obtain the dim
> # This returns a vector with the number of row
> dim(data)
[1] 387   9
>
```

Before Cleaning, it was 406.

C:\Users\User\Downloads\Lecture 22.R - R Editor

```
# Load the dataset into the R environment
# Set the file path
file_path <- "C:/Users/User/Downloads/autompg.csv"

# Read the CSV file
data <- read.csv(file_path)

?data
data

# Display the default number of rows(1st 6 rows) from the dataset
head(data)
# Display the first 3 rows of the dataset
head(data,3)

# Display the default number of rows(last 6 rows) from the dataset
tail(data)
# Display the first 4 rows of the dataset
tail(data,4)

# Use the 'dim()' function to obtain the dimensions of the dataset
# This returns a vector with the number of rows and columns in datase
dim(data)

# Use the 'names()' function to get the names of the columns in the d
```

Summary of dataset:

RGui

File Edit Packages Windows Help

R Console

```
 $ HorsePower  : int  130 165 150 150 140 198 220 150 170 160 ...
 $ Weight      : int  3504 3693 3436 3433 3449 4341 4354 3433 3563 3609 ...
 $ acceleration: num  12 11.5 11 12 10.5 10 9 12 10 8 ...
 $ ModelYear   : int  70 70 70 70 70 70 70 70 70 70 ...
 $ origin      : int  1 1 1 1 1 1 1 1 1 1 ...
 $ CarName     : chr  "chevrolet chevelle malibu" "buick skylark 320" "plymo$
>
> # Use the 'summary()' function to generate a statistical summary of the da$
> # This provides a quick overview of the data, including min, max, mean, an$
> summary(data)
     mpg.Y.        Cylinders      Displacement     HorsePower
 Min.   : 9.00   Min.   :3.00   Min.   : 68.0   Min.   : 46.0
 1st Qu.:17.25   1st Qu.:4.00   1st Qu.:105.0   1st Qu.: 75.0
 Median :22.50   Median :4.00   Median :151.0   Median : 93.0
 Mean   :23.48   Mean   :5.47   Mean   :193.6   Mean   :103.9
 3rd Qu.:29.00   3rd Qu.:8.00   3rd Qu.:264.5   3rd Qu.:125.0
 Max.   :46.60   Max.   :8.00   Max.   :455.0   Max.   :230.0
     Weight      acceleration     ModelYear        origin
 Min.   :1613   Min.   : 8.00   Min.   :70.00   Min.   :1.000
 1st Qu.:2224   1st Qu.:13.80   1st Qu.:73.00   1st Qu.:1.000
 Median :2807   Median :15.50   Median :76.00   Median :1.000
 Mean   :2974   Mean   :15.58   Mean   :76.06   Mean   :1.571
 3rd Qu.:3607   3rd Qu.:17.05   3rd Qu.:79.00   3rd Qu.:2.000
 Max.   :5140   Max.   :24.80   Max.   :82.00   Max.   :3.000
   CarName
 Length:387
 Class :character
 Mode  :character


>
```

C:\Users\User\Downloads\Lecture 22.R - R Editor

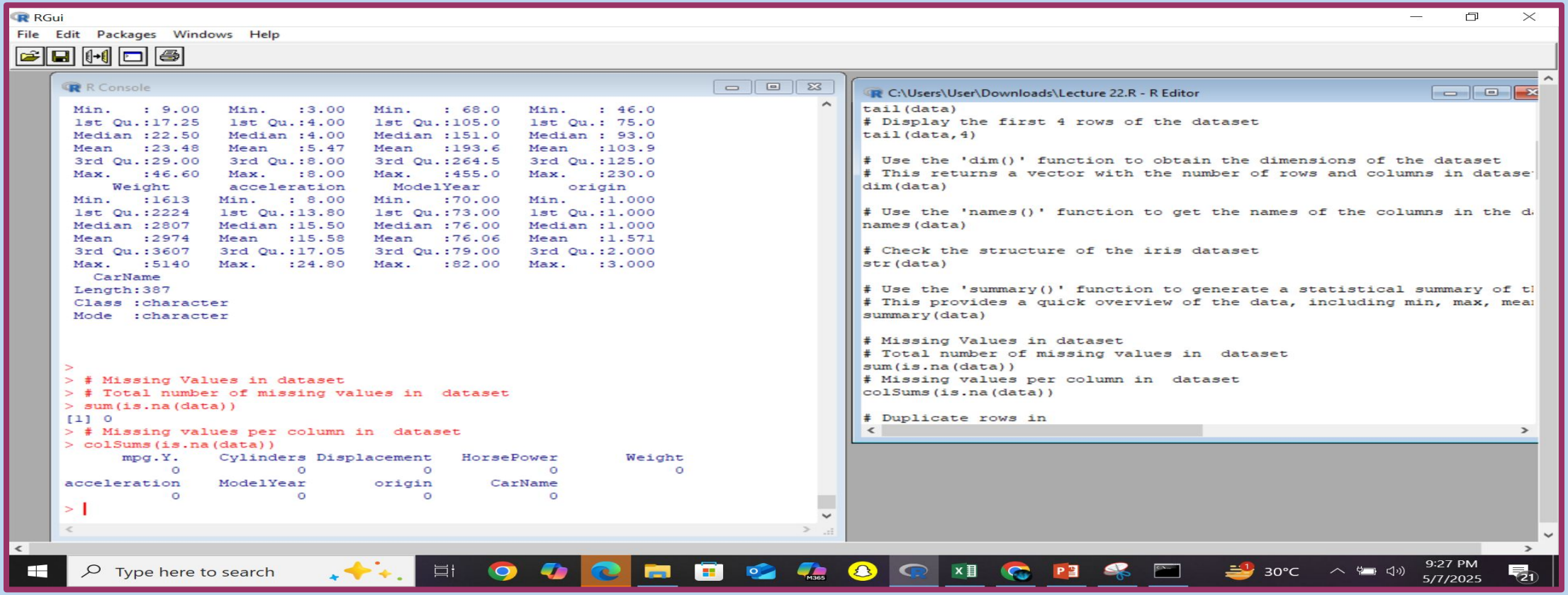
```
# Display the default number of rows(1st 6 rows) from the dataset
head(data)
# Display the first 3 rows of the dataset
head(data,3)

# Display the default number of rows(last 6 rows) from the dataset
tail(data)
# Display the first 4 rows of the dataset
tail(data,4)

# Use the 'dim()' function to obtain the dimensions of the dataset
# This returns a vector with the number of rows and columns in datase
dim(data)

# Use the 'names()' function to get the names of the columns in the d
names(data)

# Check the structure of the iris dataset
str(data)

# Use the 'summary()' function to generate a statistical summary of t
# This provides a quick overview of the data, including min, max, mea
summary(data)

# Missing Values in dataset
```

Type here to search

9:27 PM
5/7/2025

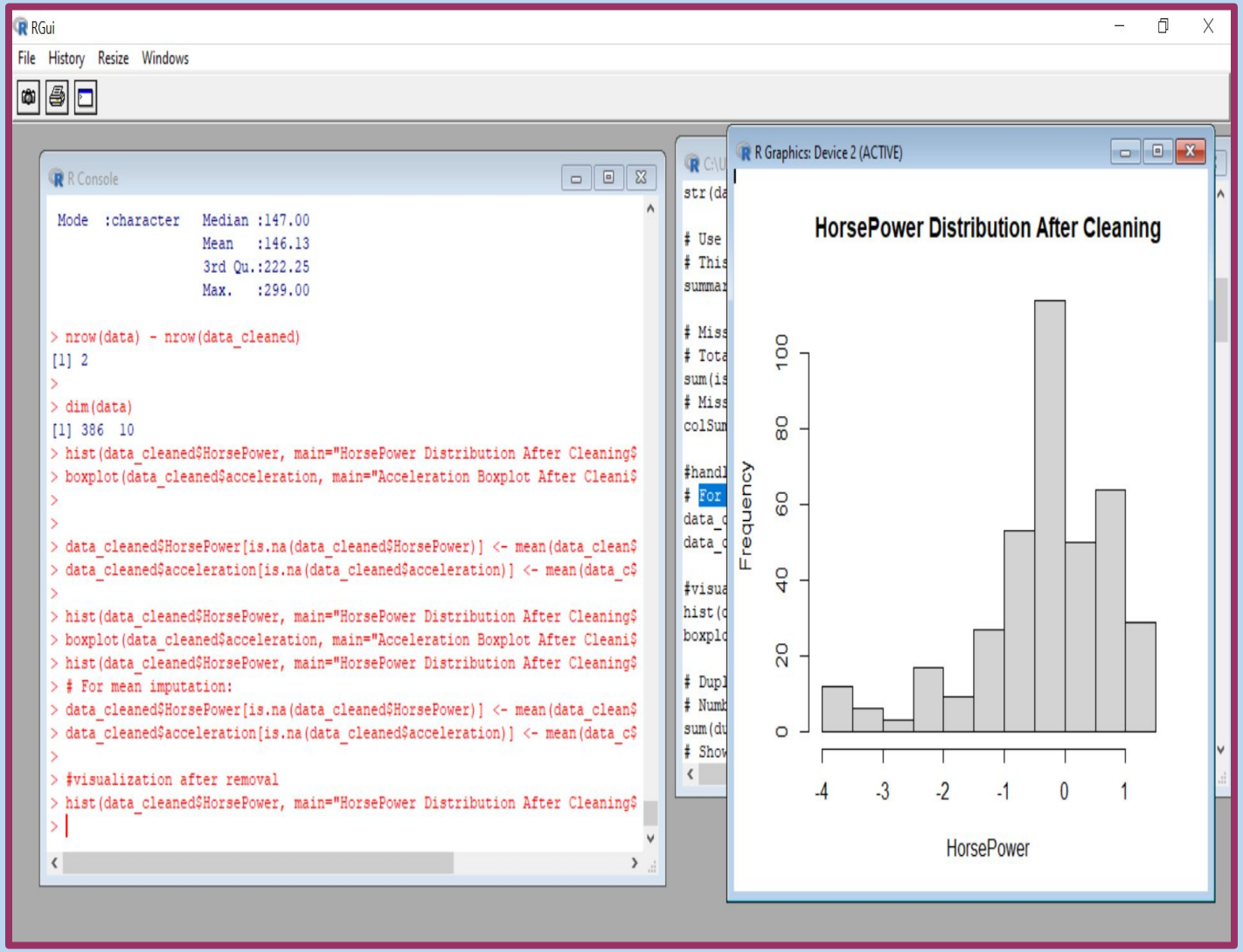
Handling missing values:

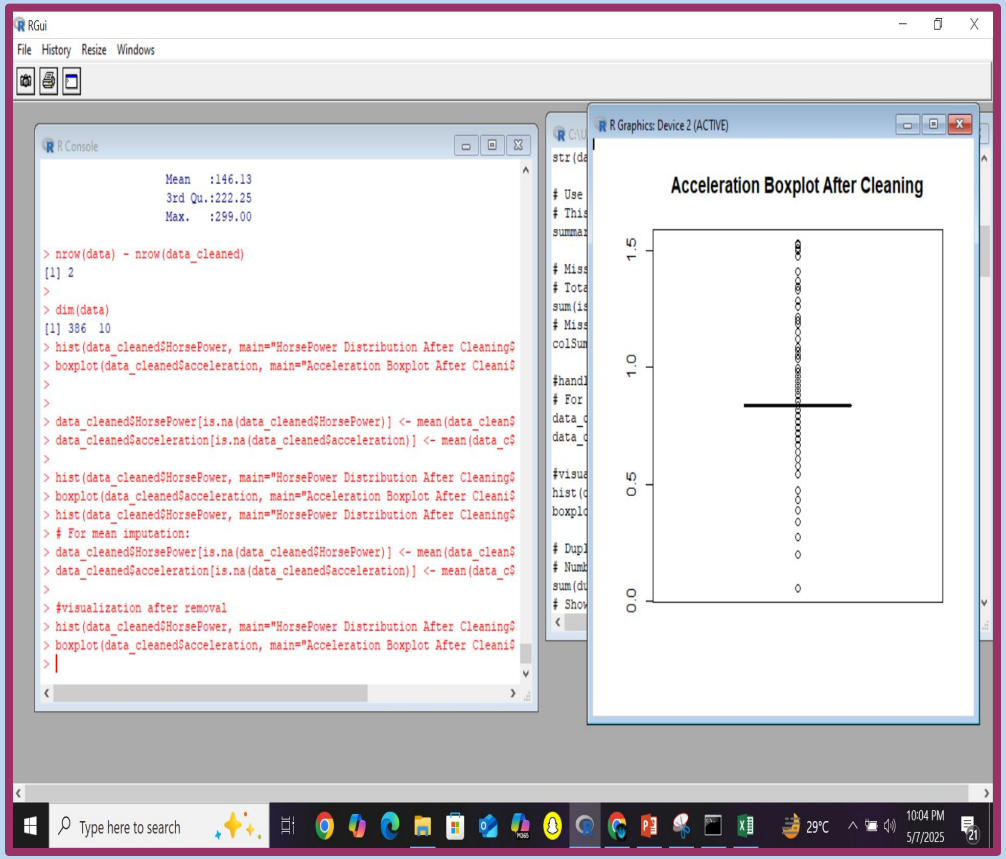
- Techniques used:
  - Delete rows.
- Mean imputation:
- Commands(# For mean imputation:
  - data_cleaned$HorsePower[is.na(data_cleaned$HorsePower)] <- mean(data_cleaned$HorsePower, na.rm = TRUE)
  - data_cleaned$acceleration[is.na(data_cleaned$acceleration)] <- mean(data_cleaned$acceleration, na.rm = TRUE)



Visualization after removing missing values.

hist(data_cleaned$HorsePower, main="HorsePower Distribution After Cleaning", xlab="HorsePower")





boxplot(data_cleaned$acceleration, main="Acceleration Boxplot After Cleaning")

Handling Duplicates:

Deleted rows having duplicate values.

RGui

File Edit View Misc Packages Windows Help

C:\Users\User\Downloads\Lecture 22.R - R Editor

```
# Use the 'summary()' function to generate a statistical summary of tl
# This provides a quick overview of the data, including min, max, mea
summary(data)

# Missing Values in dataset
# Total number of missing values in  dataset
sum(is.na(data))
# Missing values per column in  dataset
colSums(is.na(data))

# Duplicate rows in
# Number of duplicate rows in  dataset
sum(duplicated(data))
# Shows duplicate rows in
duplicated_data <- data[duplicated(data), ]
duplicated_data

# Assuming your dataset is called data
data$CarName_numeric <- as.numeric(factor(data$CarName))


# Check how car names are mapped to numbers
levels(factor(data$CarName))

# View the result
```

R Console

```
>
> sum(duplicated(data))
[1] 0
> # Shows duplicate rows in
> duplicated_data <- data[duplicated(data), ]
> duplicated_data
[1] mpg.Y.        Cylinders     Displacement HorsePower    Weight
[6] acceleration ModelYear     origin       CarName
<0 rows> (or 0-length row.names)
>
>
```

Type here to search

9:33 PM
5/7/2025

Removed Outlier by Increasing 1.5 to 2.4

RGui

File Edit Packages Windows Help

R Console

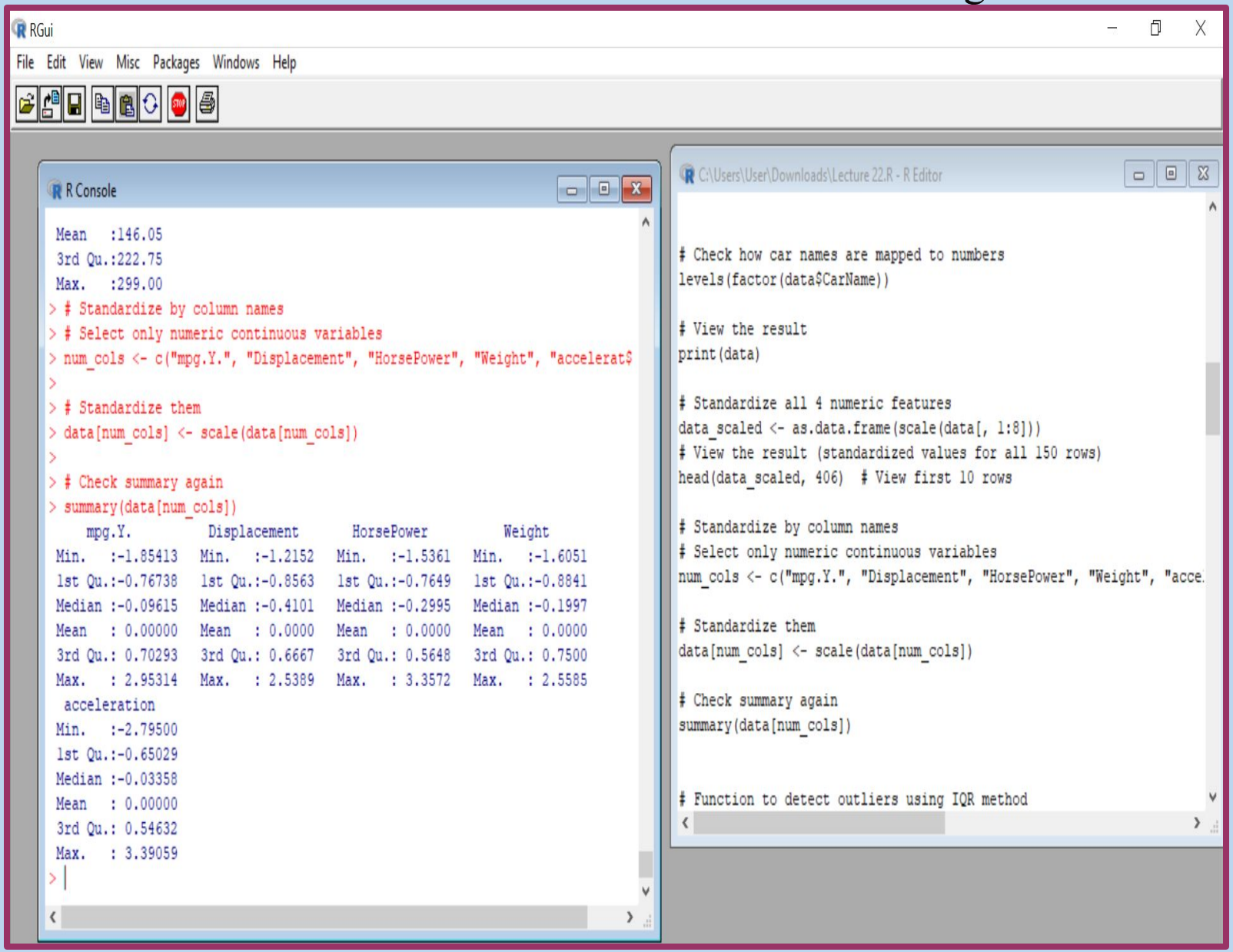
```
>
>
>
> length(outliers_data_mpg.Y.)
[1] 0
> length(outliers_data_Cylinders)
[1] 0
> length(outliers_data_Displacement)
[1] 0
> length(outliers_data_HorsePower)
[1] 0
> length(outliers_data_Weight)
[1] 0
> length(outliers_data_acceleration)
[1] 0
> length(outliers_data_ModelYear)
[1] 0
> length(outliers_data_origin)
[1] 0
>
>
> data[outliers_data_mpg.Y., "mpg.Y."]
numeric(0)
> data[outliers_data_Cylinders, "Cylinders"]
integer(0)
> data[outliers_data_Displacement, "Displacement"]
numeric(0)
> data[outliers_data_HorsePower, "HorsePower"]
numeric(0)
> data[outliers_data_Weight, "Weight"]
numeric(0)
> data[outliers_data_acceleration, "acceleration"]
numeric(0)
> data[outliers_data_ModelYear, "ModelYear"]
integer(0)
> data[outliers_data_origin, "origin"]
integer(0)
>
>
>
>
```

C:\Users\HP\Downloads\Lecture 22.R - R Editor

```
num_cols <- c("mpg.Y.", "Displacement", "HorsePower", "Weight", "acceleration")

# Standardize them
data[num_cols] <- scale(data[num_cols])

# Check summary again
summary(data[num_cols])


# Function to detect outliers using IQR method
find_outliers <- function(data) {
  Q1 <- quantile(data, 0.25)
  Q3 <- quantile(data, 0.75)
  IQR <- Q3 - Q1
  lower_bound <- Q1 - 2.4 * IQR
  upper_bound <- Q3 + 2.4 * IQR
  which(data < lower_bound | data > upper_bound)
}

# Outliers in  dataset
outliers_data_mpg.Y. <- find_outliers(data$mpg.Y.)
outliers_data_Cylinders <- find_outliers(data$Cylinders)
outliers_data_Displacement <- find_outliers(data$Displacement)
outliers_data_HorsePower <- find_outliers(data$HorsePower)
outliers_data_Weight <- find_outliers(data$Weight)
outliers_data_acceleration <- find_outliers(data$acceleration)
outliers_data_ModelYear <- find_outliers(data$ModelYear)
outliers_data_origin <- find_outliers(data$origin)



length(outliers_data_mpg.Y.)
length(outliers_data_Cylinders)
length(outliers_data_Displacement)
length(outliers_data_HorsePower)
length(outliers_data_Weight)
length(outliers_data_acceleration)
length(outliers_data_ModelYear)
```

Standardization:

should not be done on categorial and discrete values.



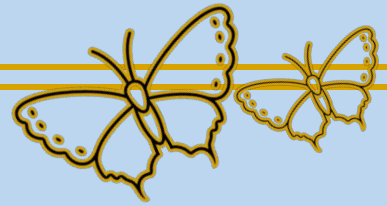
Standardization:

- **Standardize only continuous numeric variables** (e.g., mpg.Y., Displacement, HorsePower, Weight, acceleration).
- These variables have meaningful magnitudes and benefit from mean = 0, SD = 1 transformation.

Do not standardize:

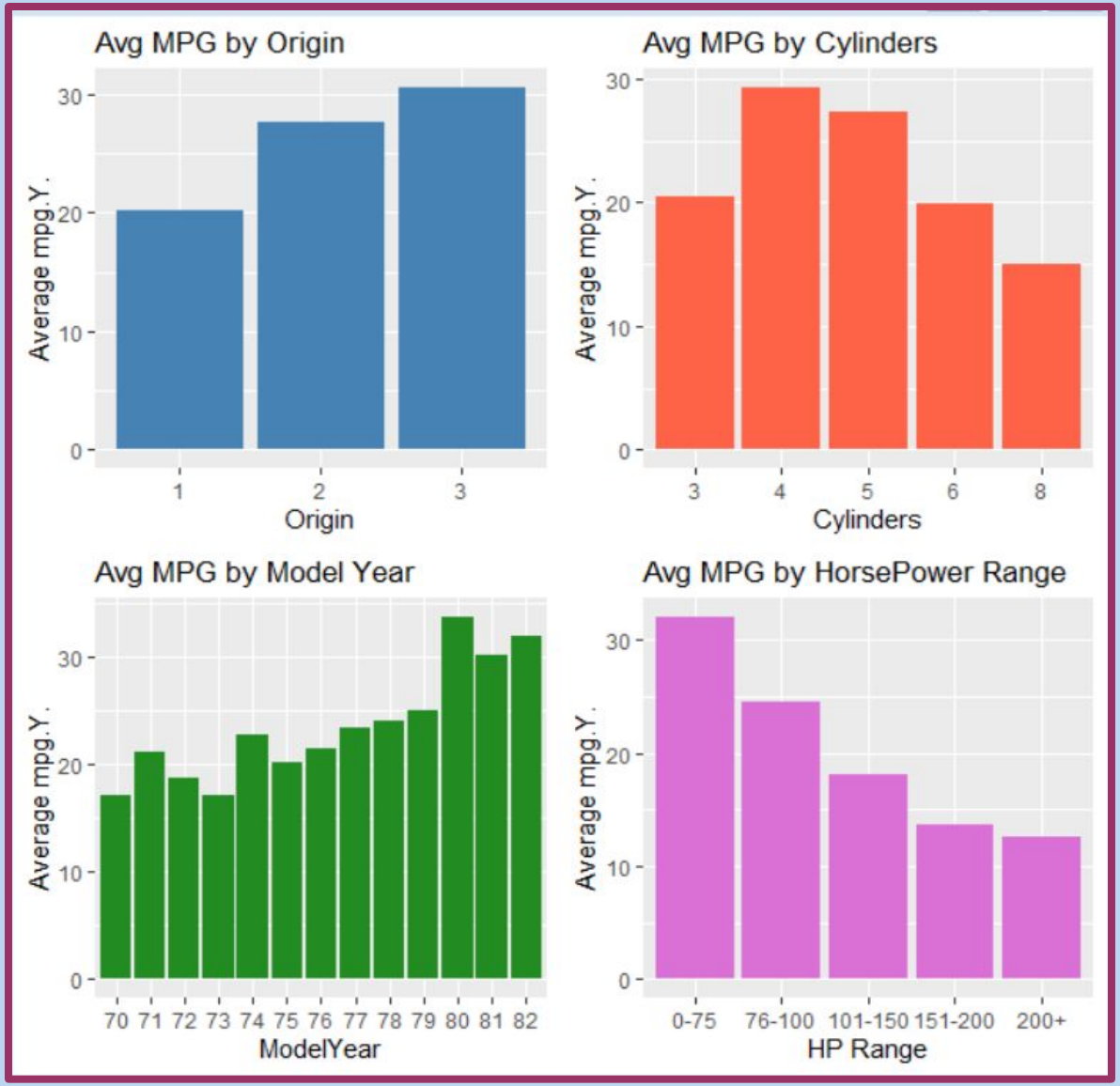
- **Categorical variables** like origin → use one-hot encoding or factors.
- **Text variables** like CarName → extract brand or encode separately.
- **Discrete or identifier-like variables**:
  - Cylinders: Discrete, may be better as category.
  - ModelYear: Represents time, not magnitude.
  - CarName_numeric: Just an ID, not meaningful.

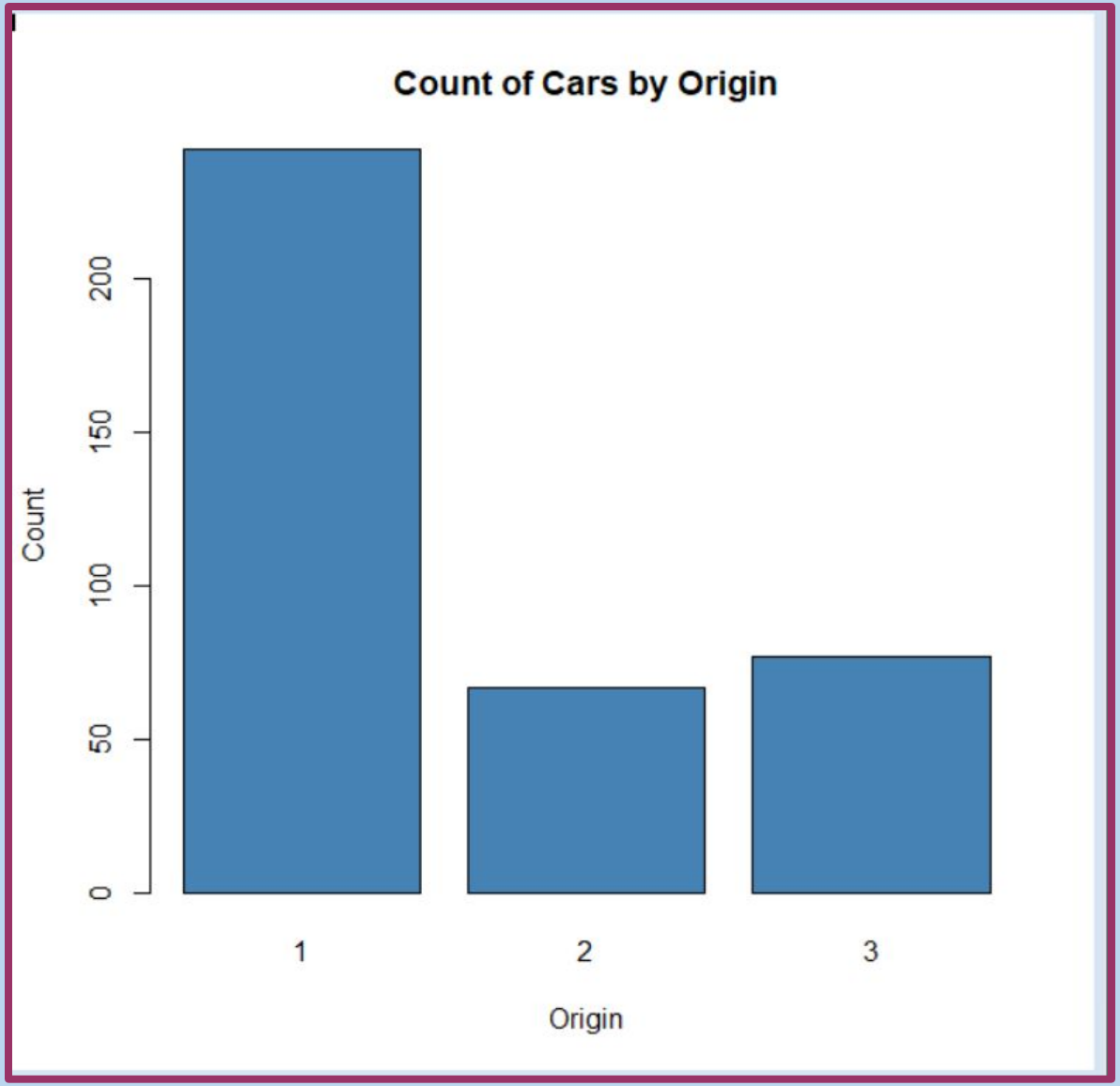
Only standardize variables where numeric relationships are meaningful.



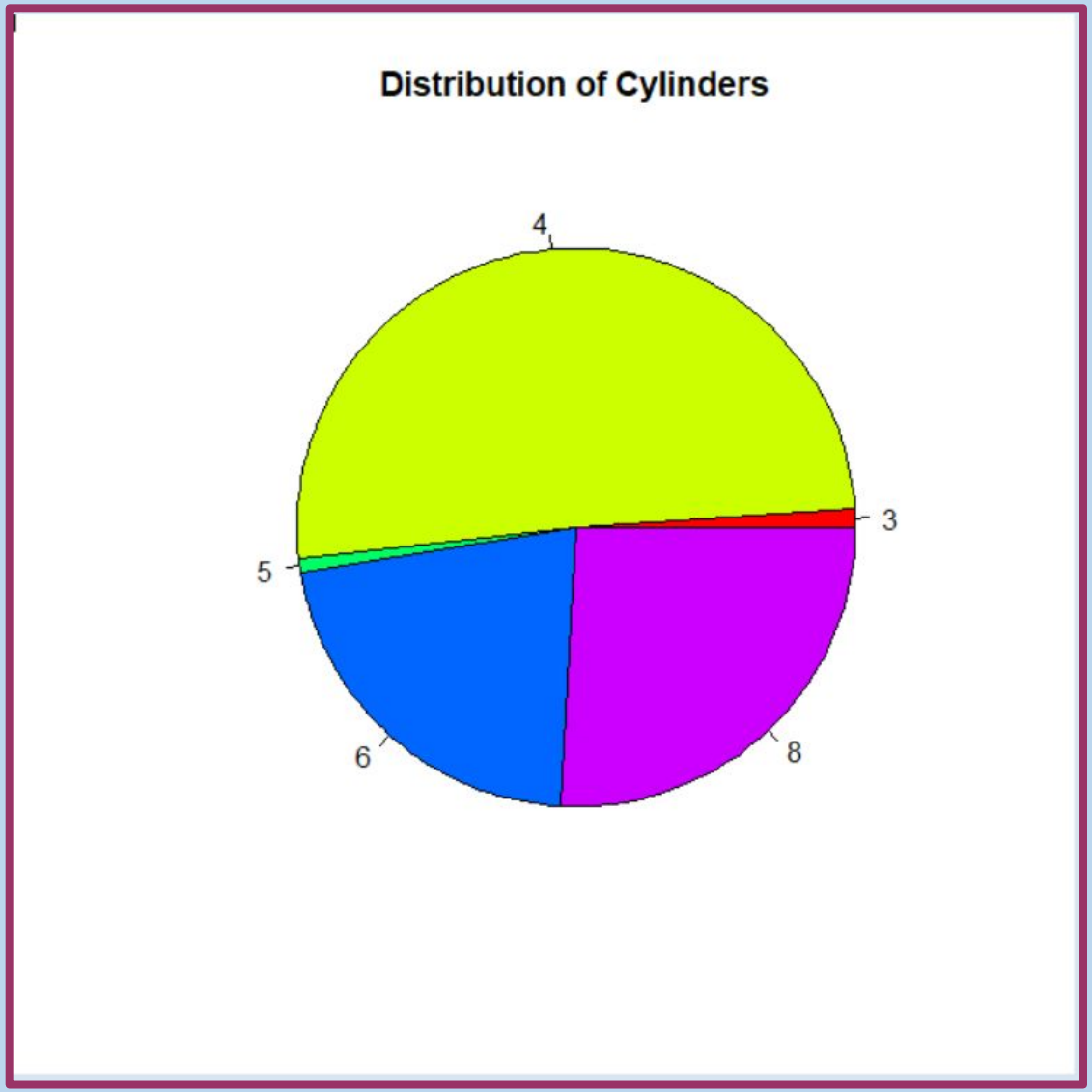
Task 5 : Exploratory Data Analysis (EDA)

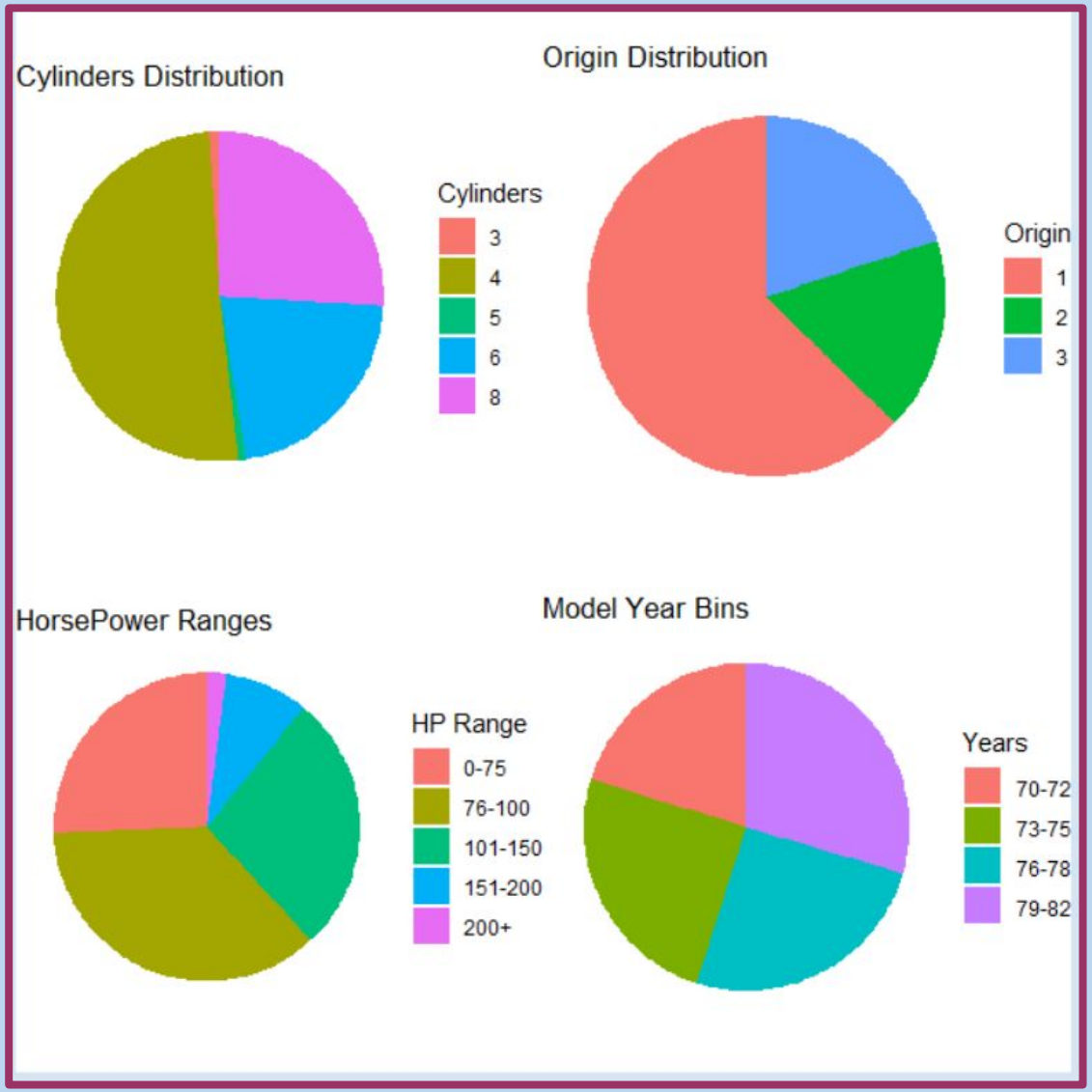
Bar Graphs


Avg MPG by Origin
Average mpg.Y.
0
10
20
30
1
2
3
Origin
Avg MPG by Cylinders
Average mpg.Y.
0
10
20
30
3
4
5
6
8
Cylinders
Avg MPG by Model Year
Average mpg.Y.
0
10
20
30
70 71 72 73 74 75 76 77 78 79 80 81 82
ModelYear
Avg MPG by HorsePower Range
Average mpg.Y.
0
10
20
30
0-75
76-100
101-150
151-200
200+
HP Range

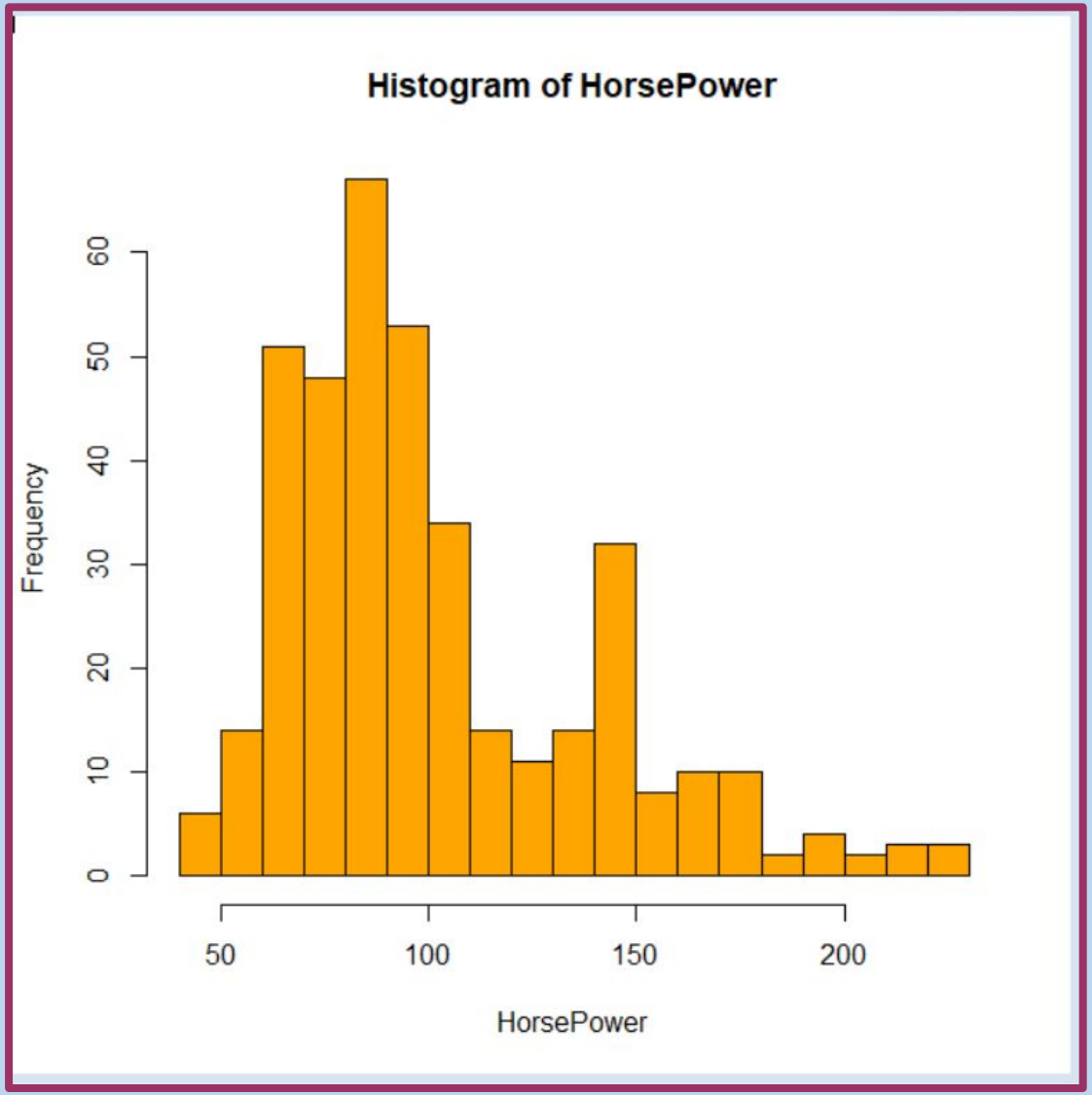

Count of Cars by Origin
Count
0
50
100
150
200
1
2
3
Origin

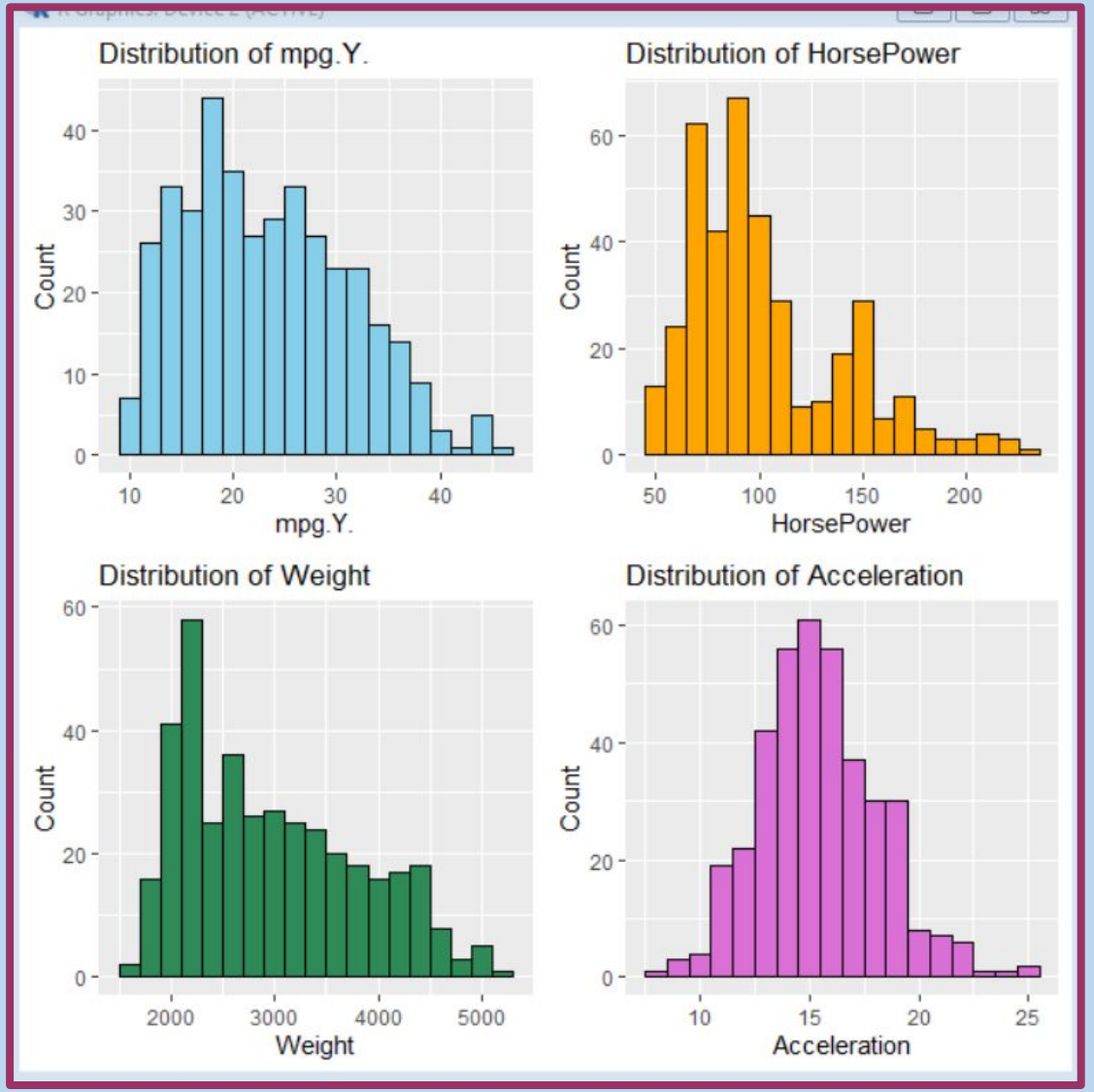
Pie Charts


Distribution of Cylinders
4
3
5
6
8

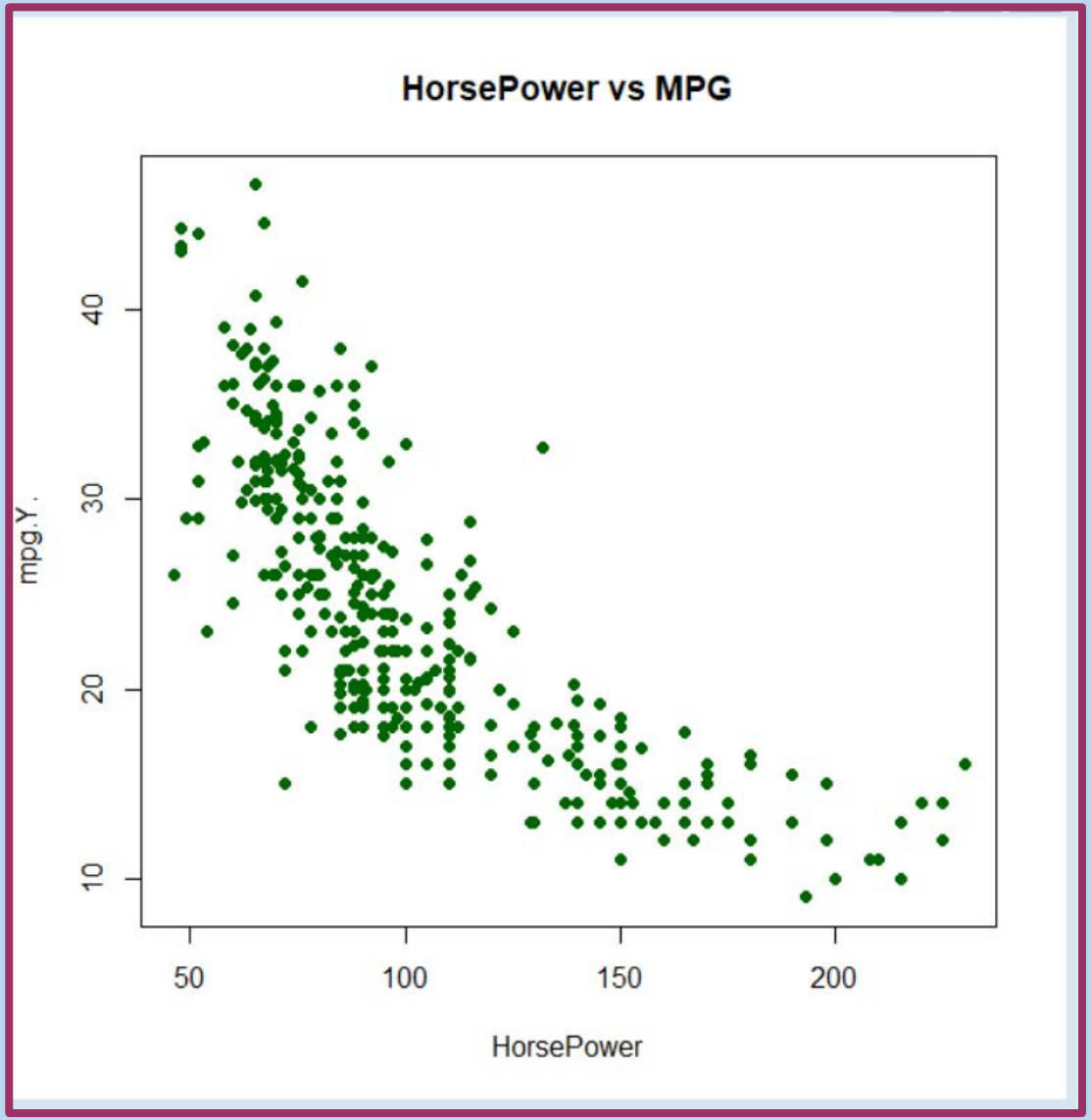

Cylinders Distribution
Cylinders
3
4
5
6
8
Origin Distribution
Origin
1
2
3
HorsePower Ranges
HP Range
0-75
76-100
101-150
151-200
200+
Model Year Bins
Years
70-72
73-75
76-78
79-82

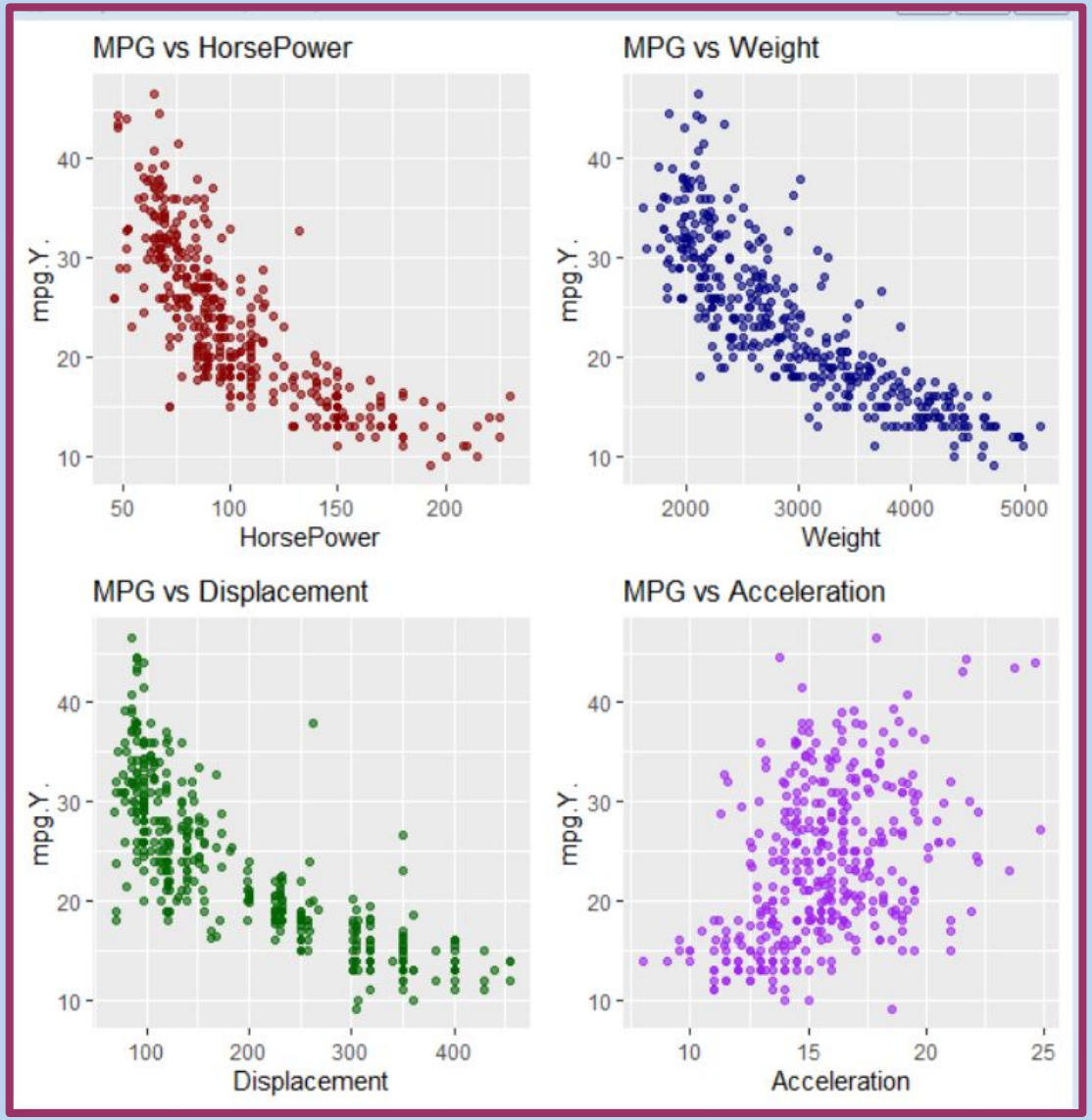
Histogram


Histogram of HorsePower
Frequency
0
10
20
30
40
50
60
50
100
150
200
HorsePower

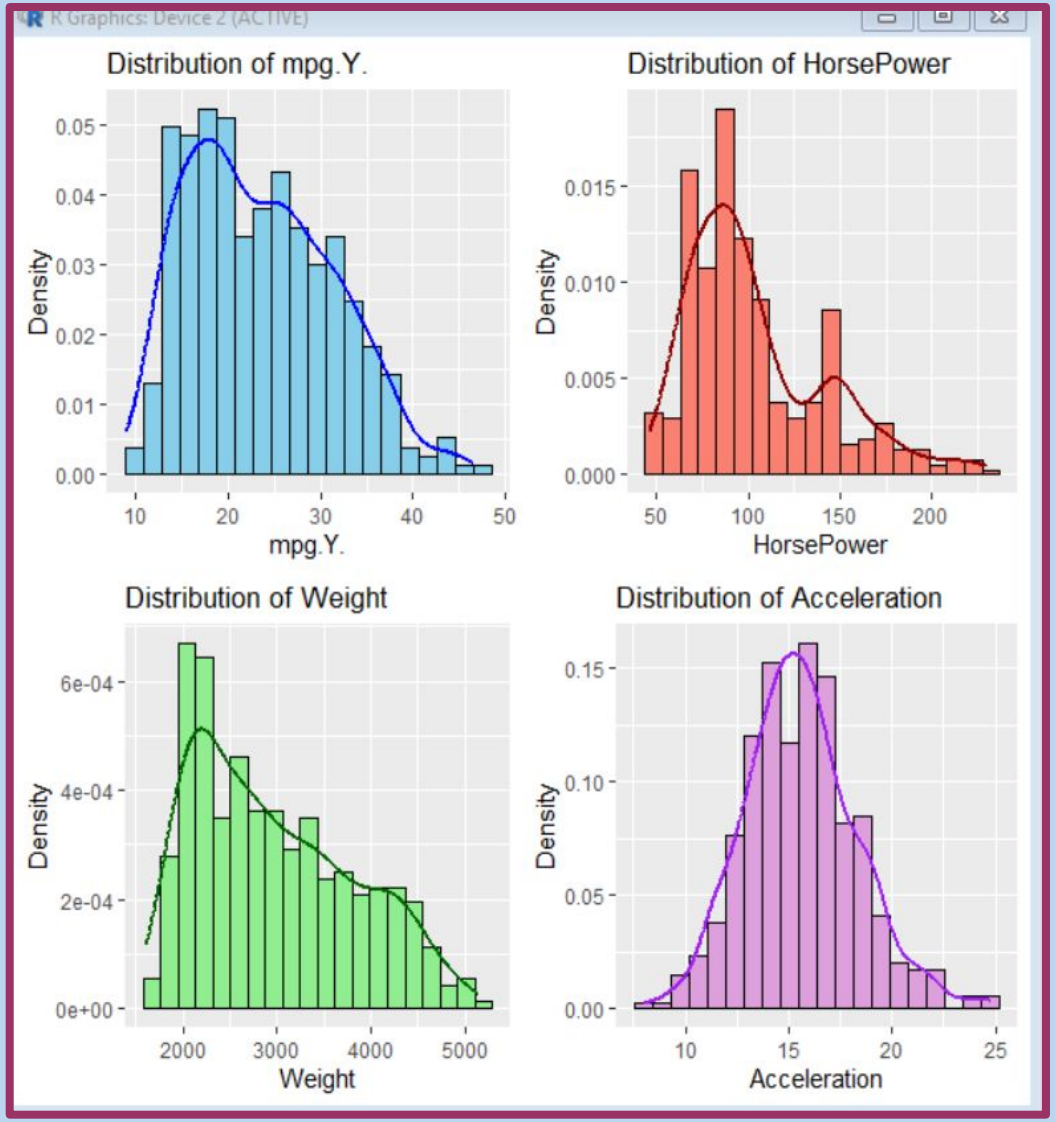

Distribution of mpg.Y.
Count
mpg.Y.
Distribution of HorsePower
Count
HorsePower
Distribution of Weight
Count
Weight
Distribution of Acceleration
Count
Acceleration

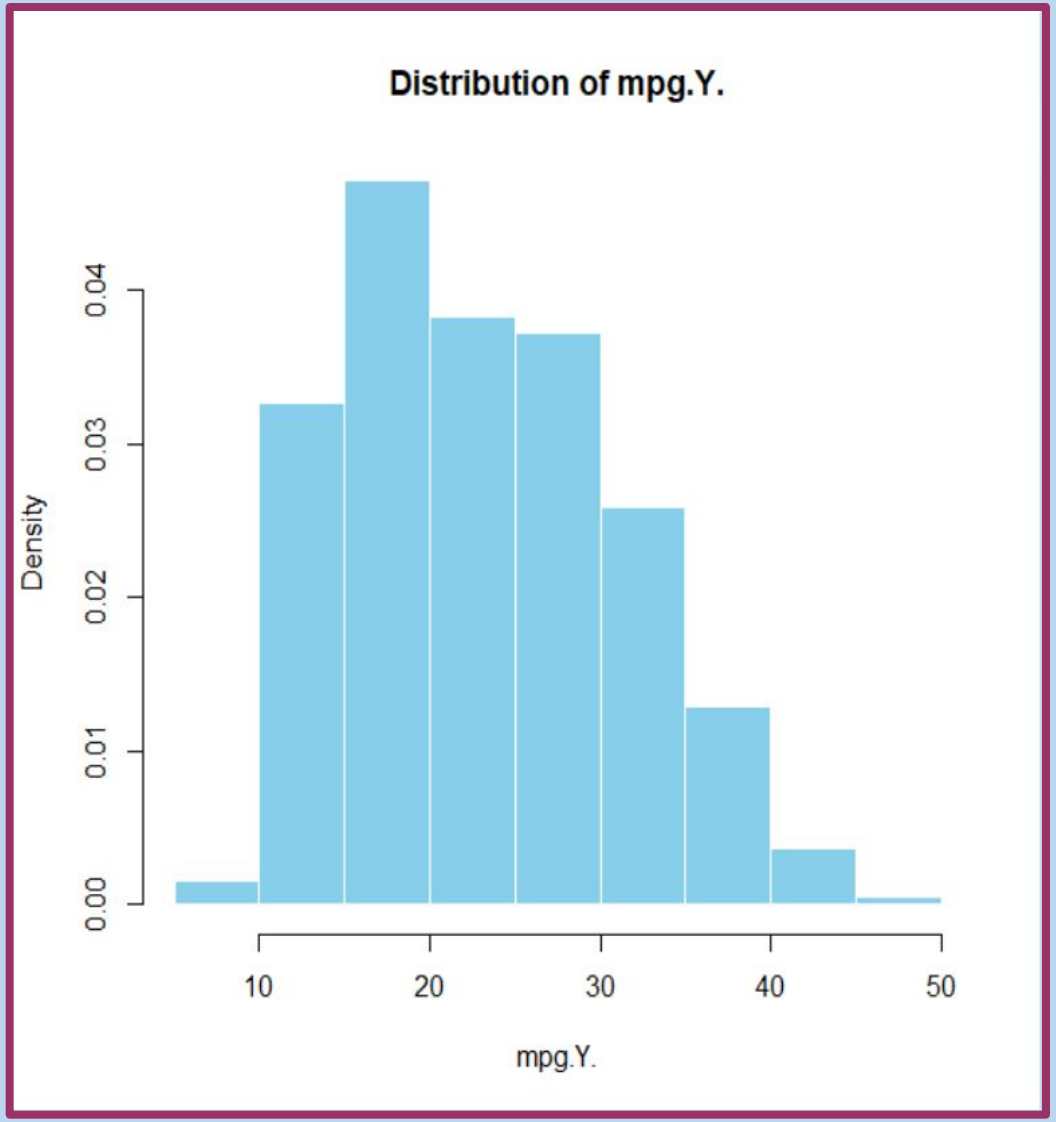
Scatter Plots



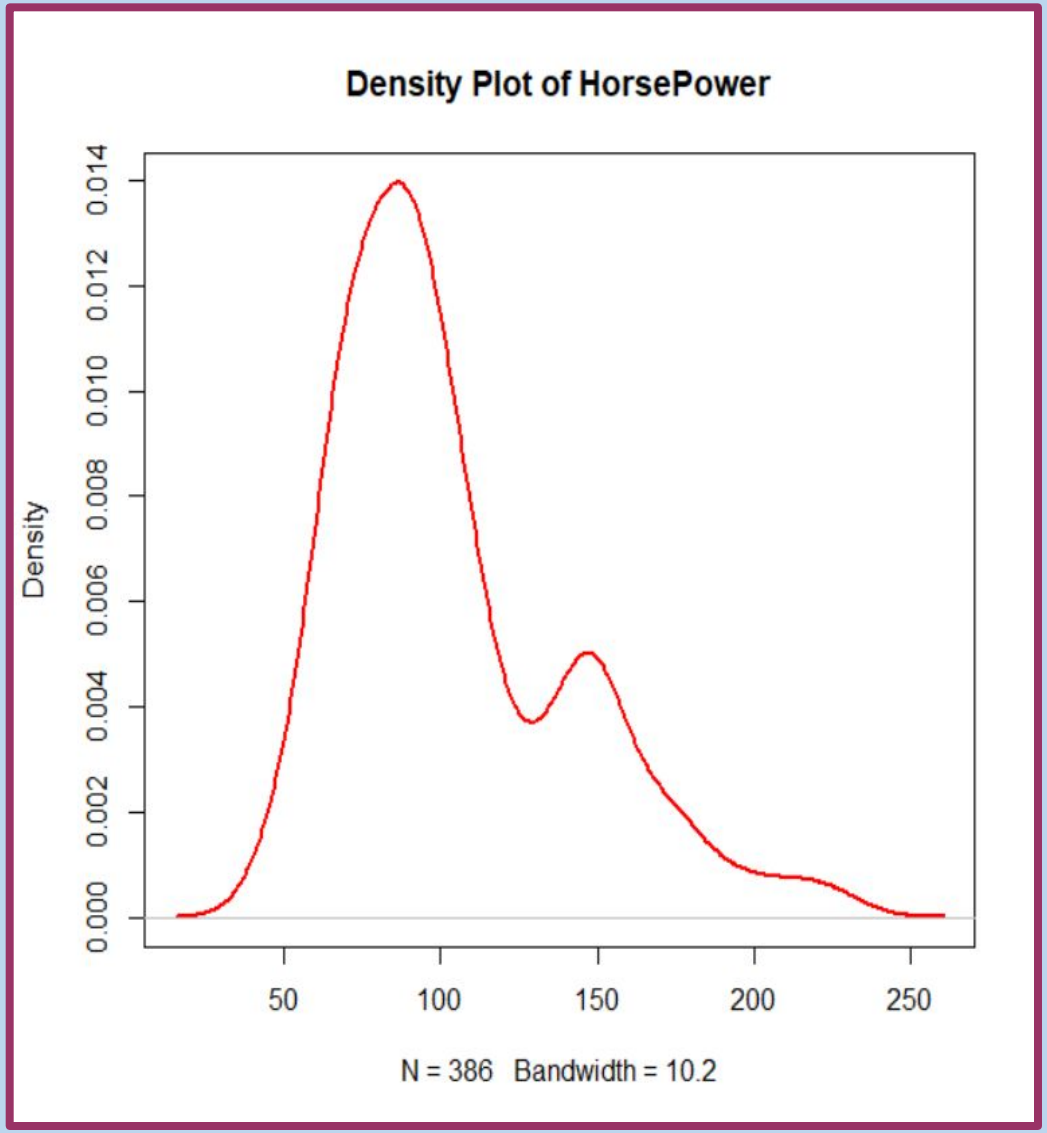


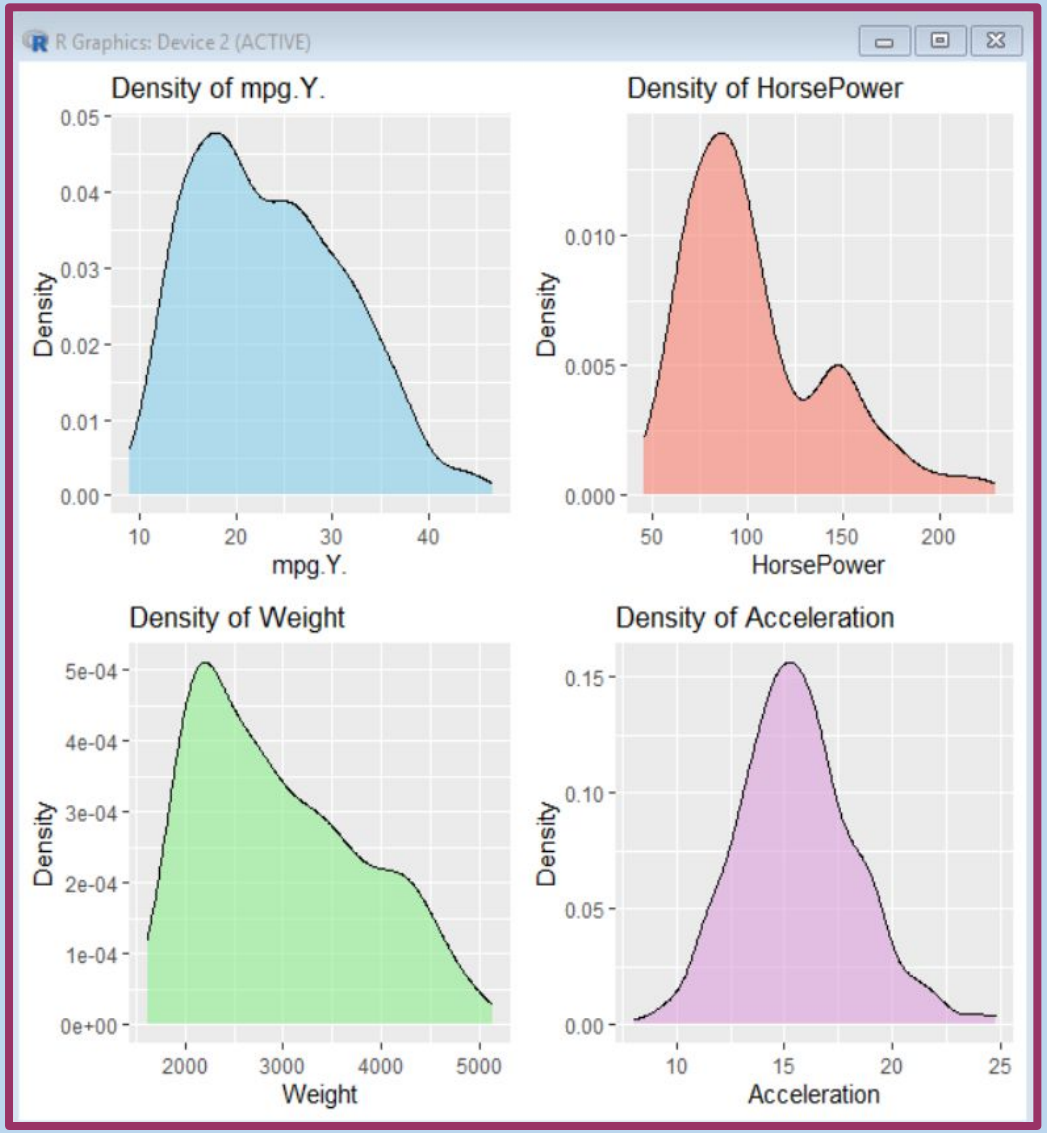
Distribution Plot



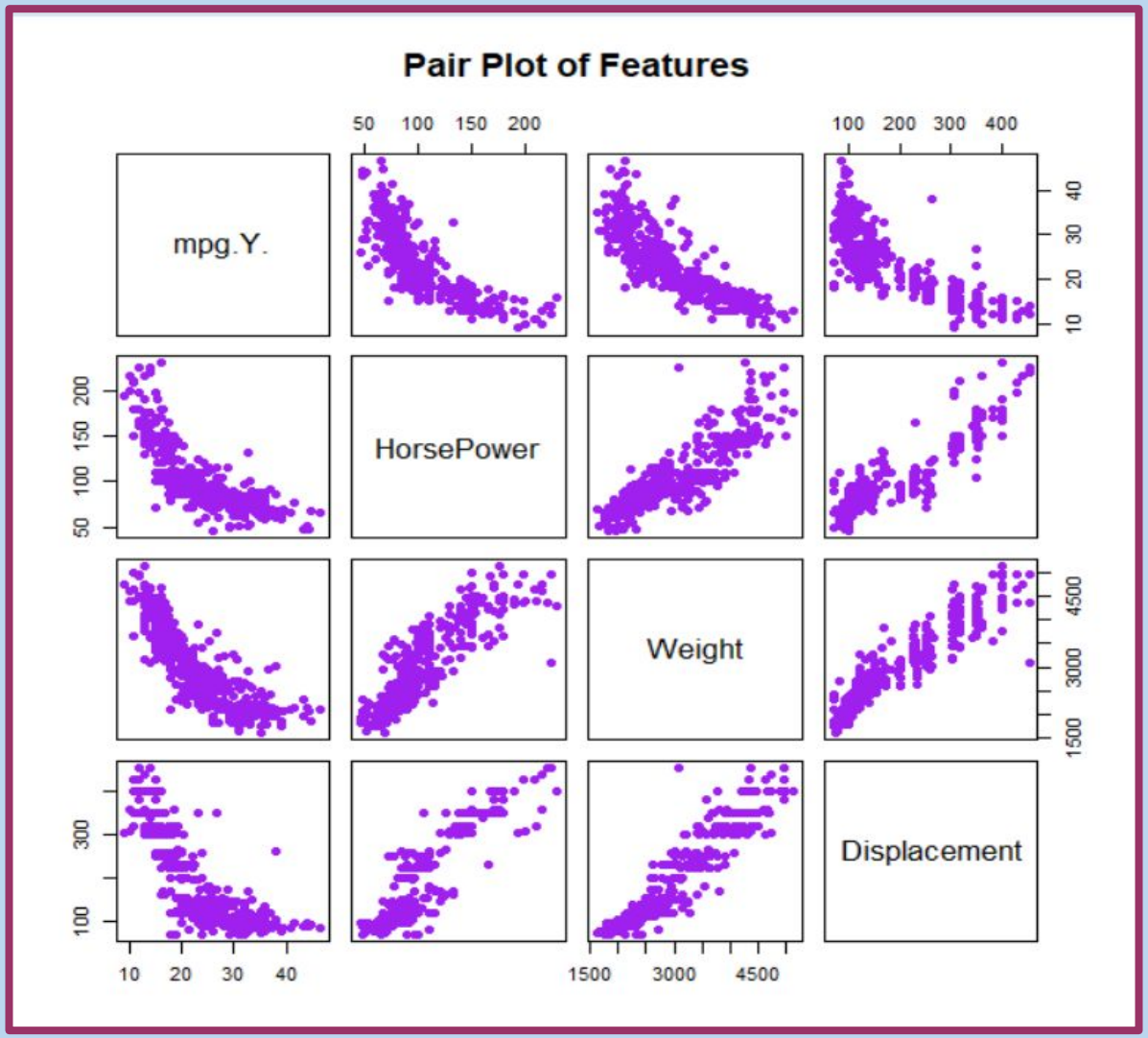


Density Plot

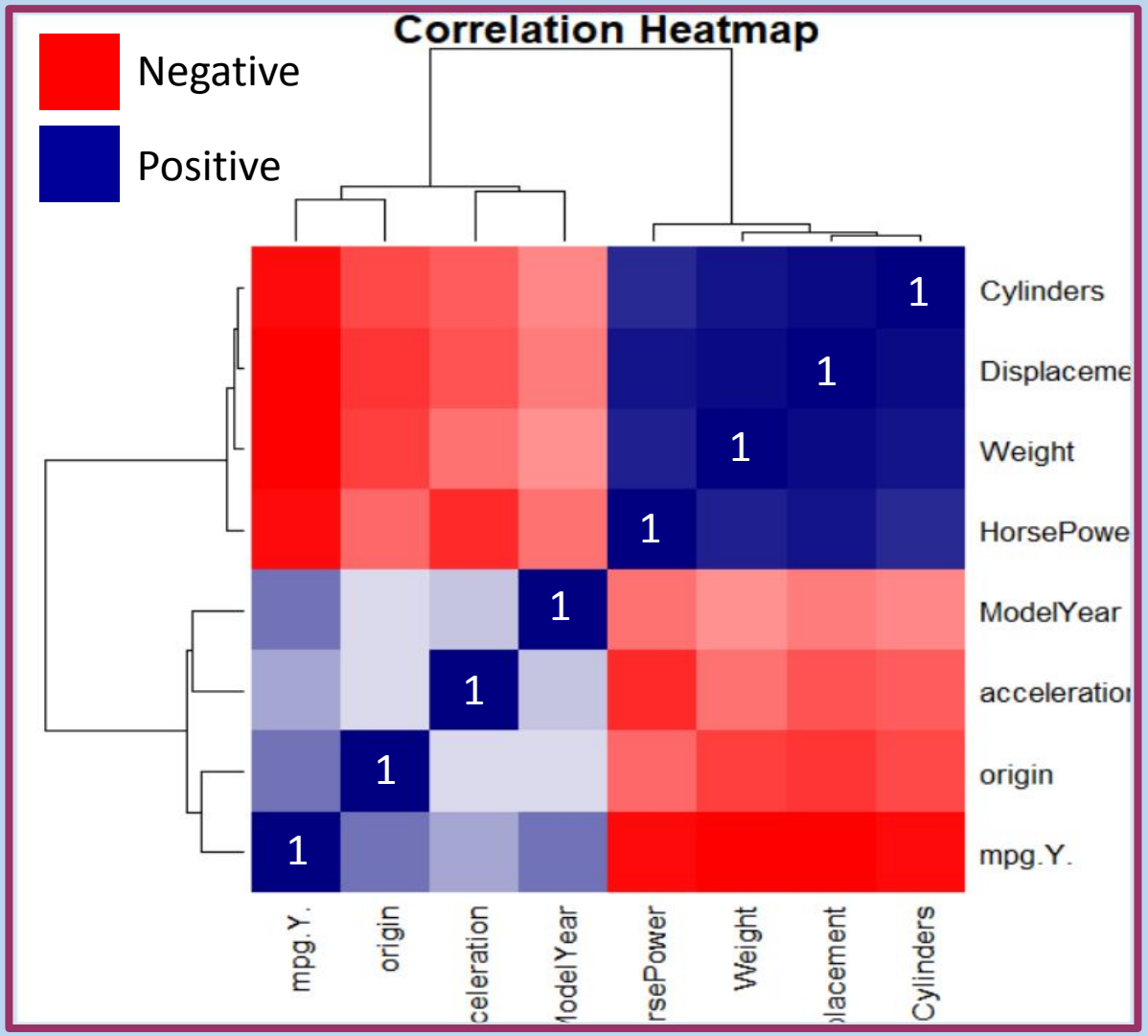




Pair Plot

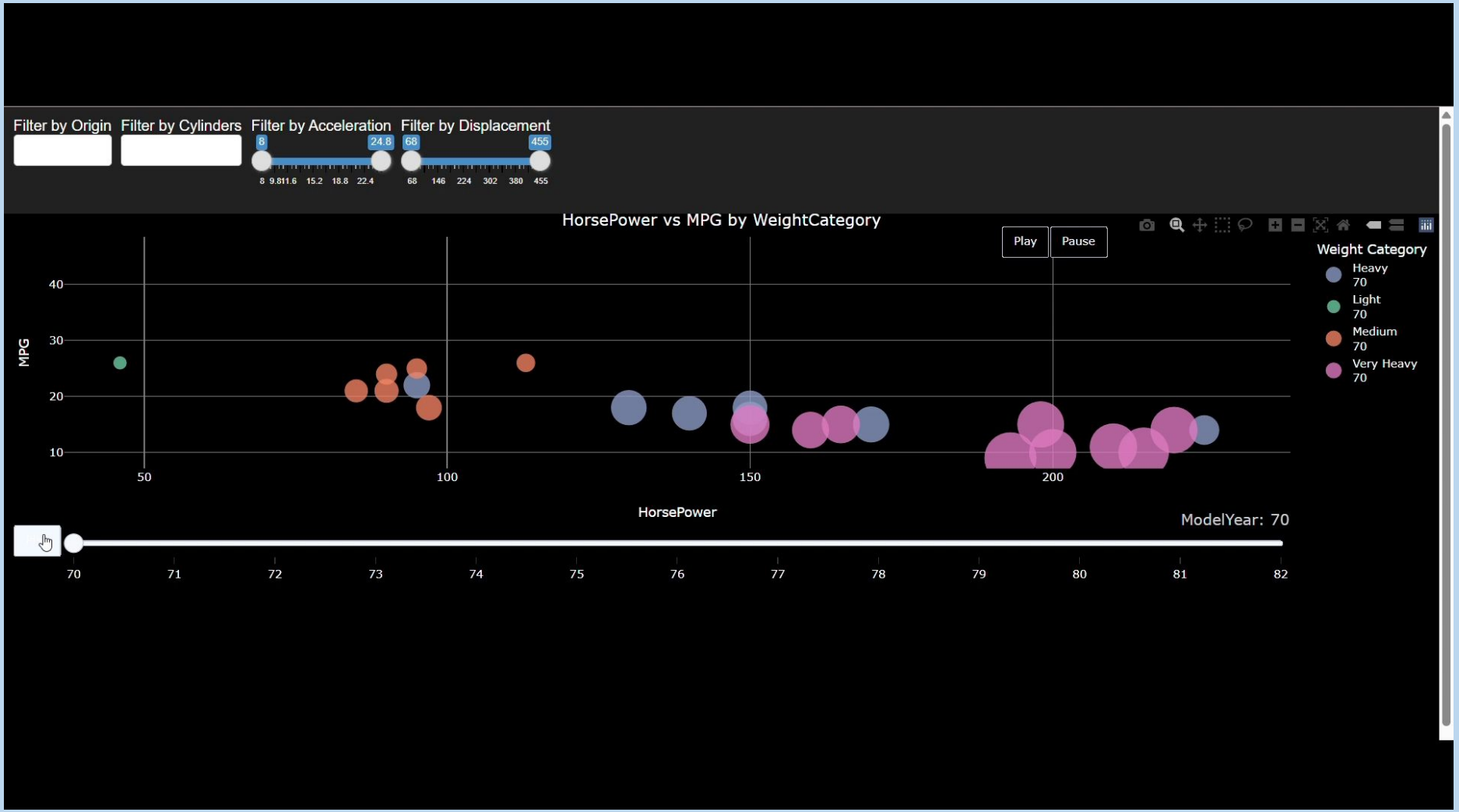


Heatmap



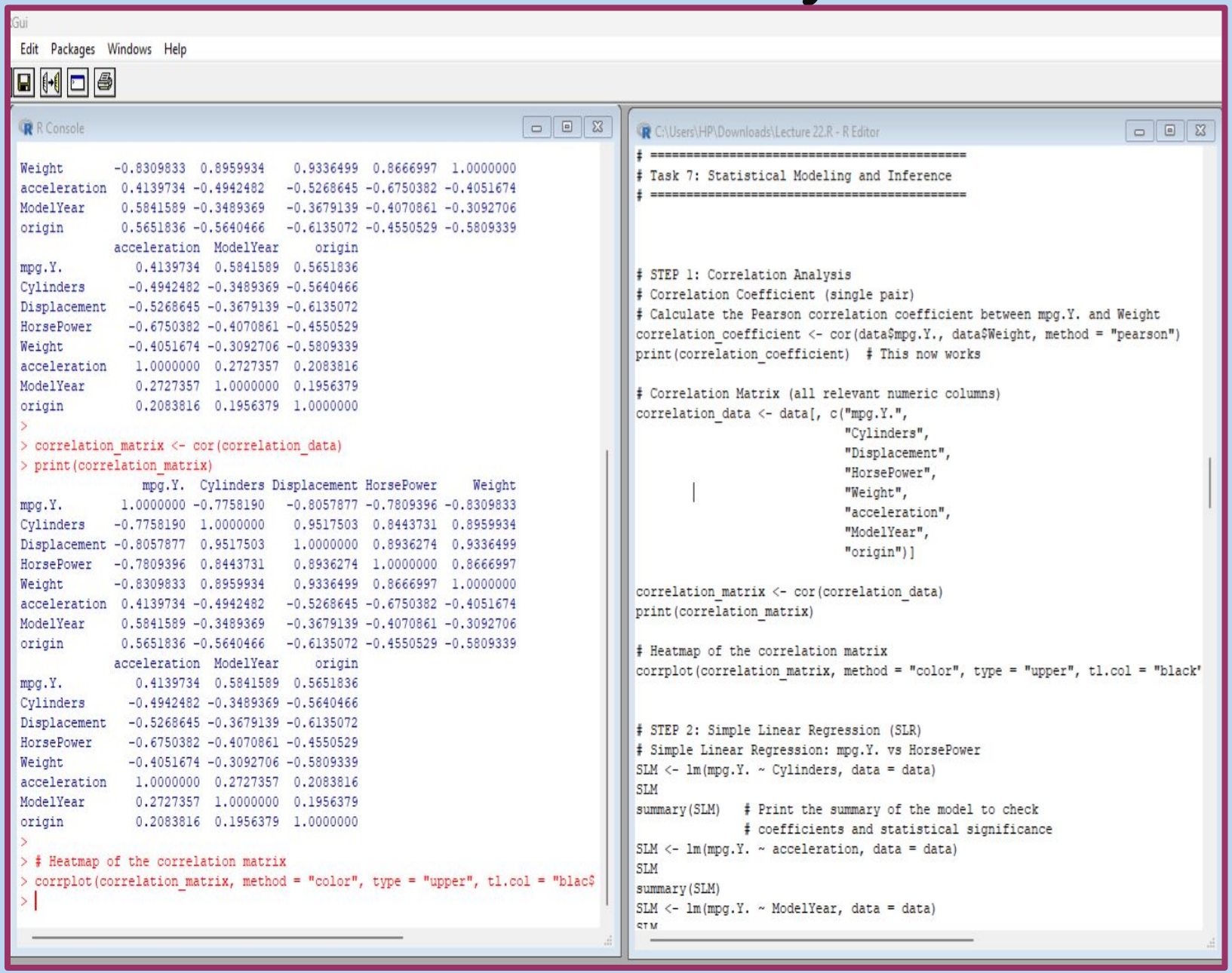
Task 6 : Interactive Visualizations

Interactive Scatter Plot



Task 7 : Statistical Modeling and Inference

Correlation Analysis:

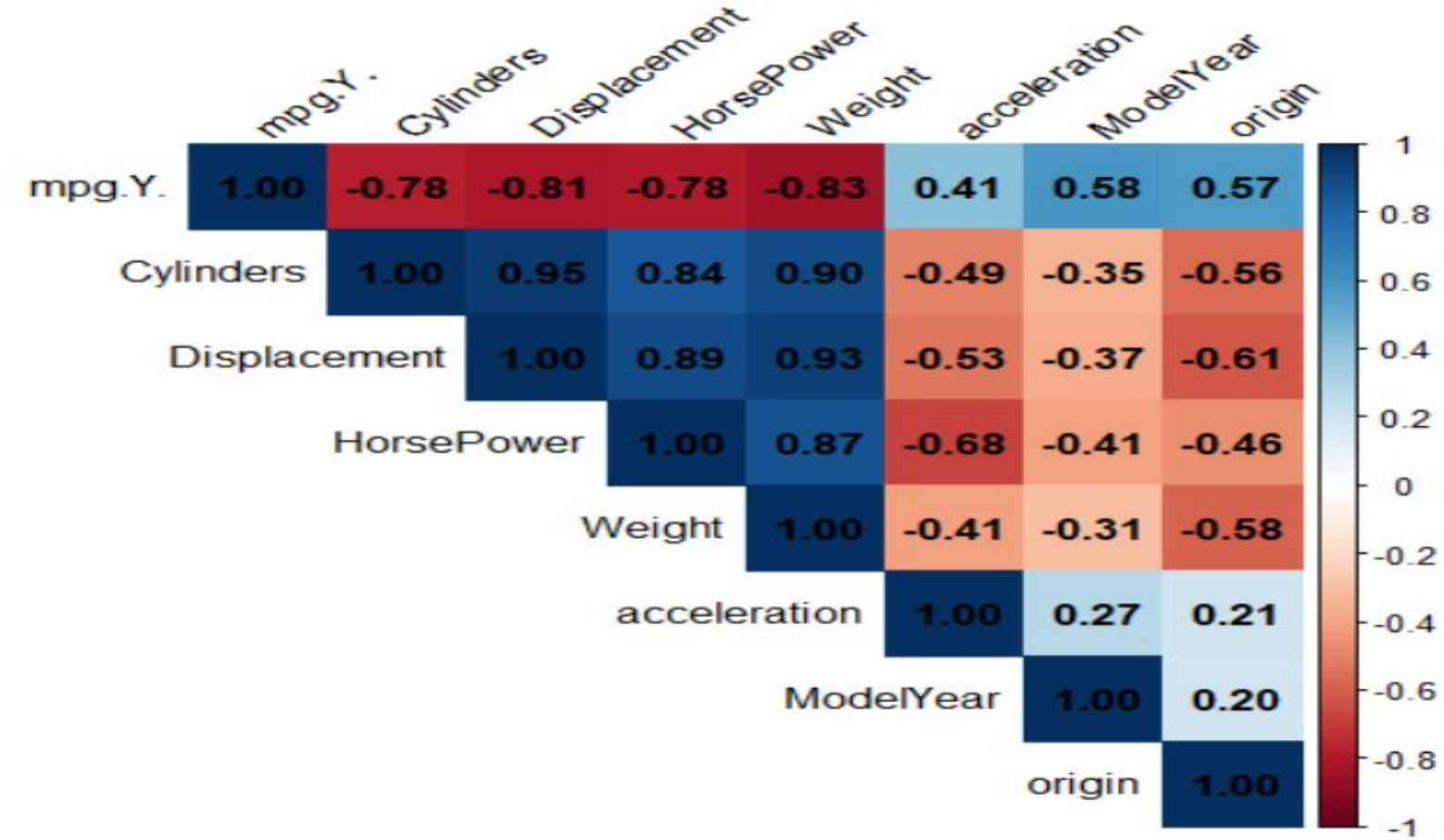


- We analyzed how car attributes relate to fuel efficiency (mpg.Y.).
- Pearson correlation coefficients were calculated between mpg.Y. and other numeric variables.

Top Findings:

- Weight and mpg.Y.: -0.83 → Heavier cars are less fuel efficient.
- HorsePower and mpg.Y.: -0.78 → More powerful engines = lower mpg.
- Model Year and mpg.Y.: +0.58 → Newer models are more efficient.
- Cylinders, Displacement, and HorsePower are highly interrelated (r > 0.85), suggesting multicollinearity risk in modeling.

Coefficient Correlation

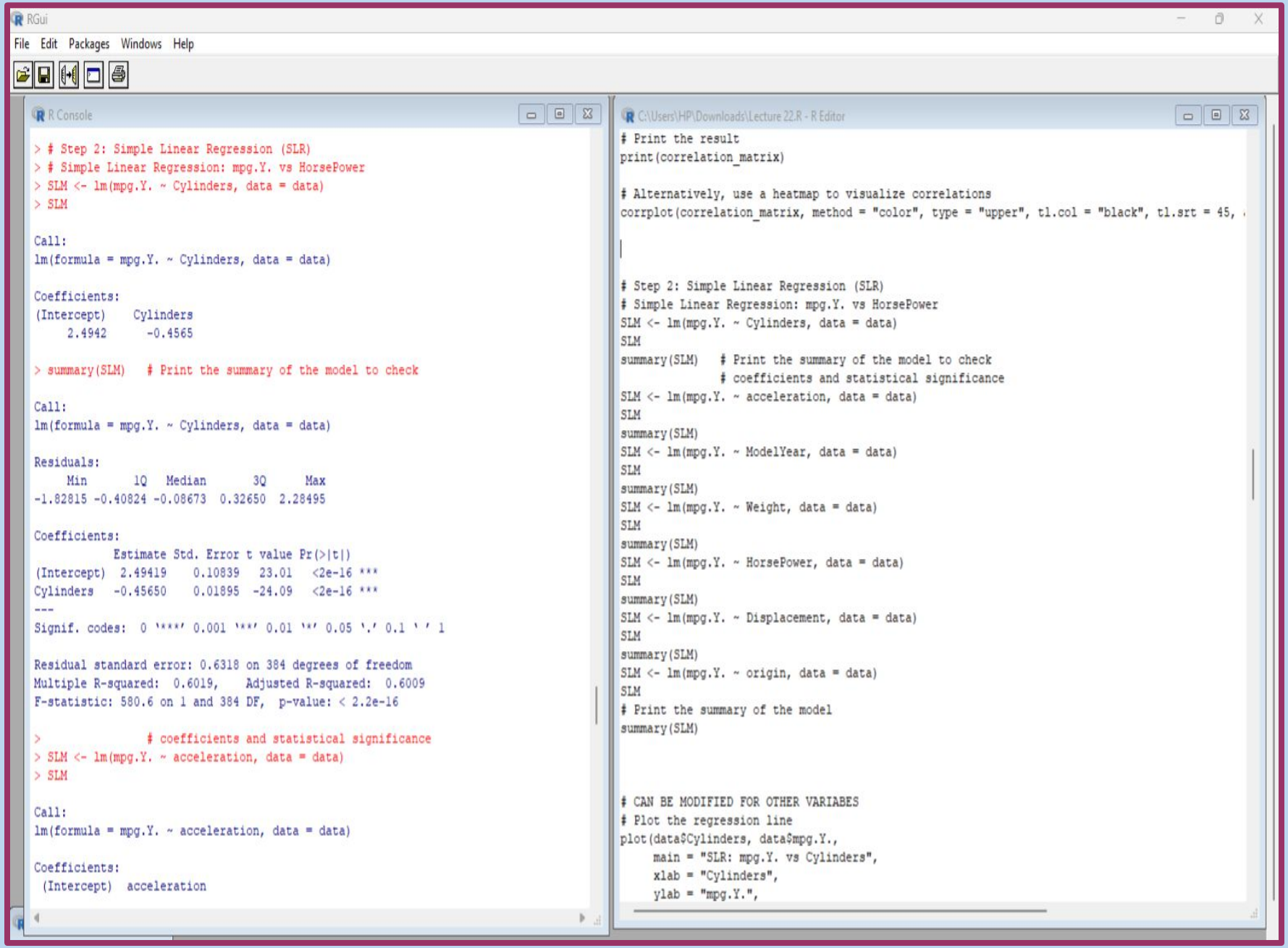


Summary

Variable	Correlation with mpg.Y.	Interpretation
Weight	-0.83	Heavier cars = lower fuel efficiency. Strong **negative** correlation.
Cylinders	-0.78	More cylinders = less fuel efficient. Strong **negative** correlation.
Displacement	-0.81	Bigger engine size = worse mileage. Strong **negative** correlation.
HorsePower	-0.78	Powerful engines = poor fuel efficiency. Strong **negative** correlation.
Acceleration	+0.41	Slight trend: quicker acceleration = better mpg, but it's weak.
ModelYear	+0.58	Newer cars tend to have **better mileage**. Medium positive correlation.
Origin	+0.56	Cars from some origins (e.g., Japan) have better mpg.

Variable Pair	Correlation	Meaning
Cylinders vs. Displacement	+0.95	More cylinders → bigger engines. VERY highly correlated.
HorsePower vs. Displacement	+0.89	More powerful cars tend to have larger engines.
Weight vs. Displacement	+0.93	Heavier cars often have bigger engines.

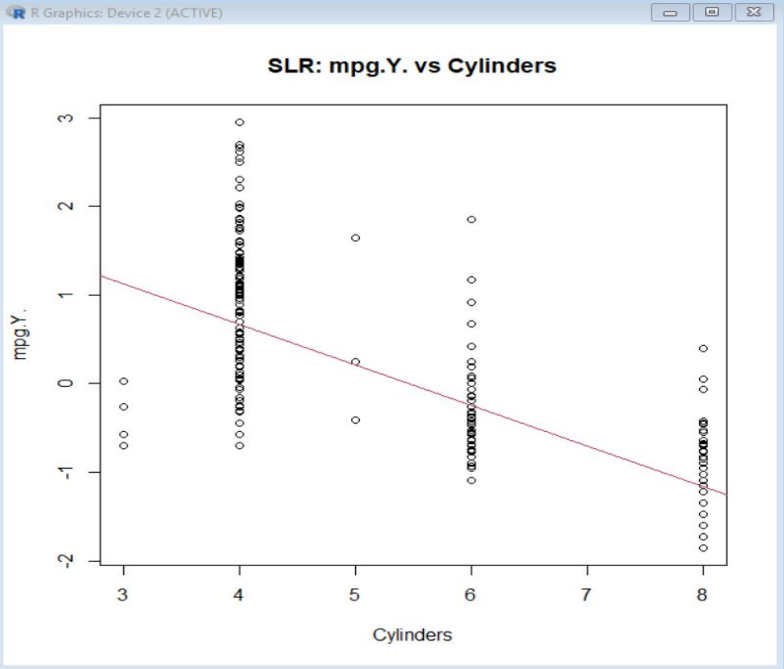
Simple linear Regression:

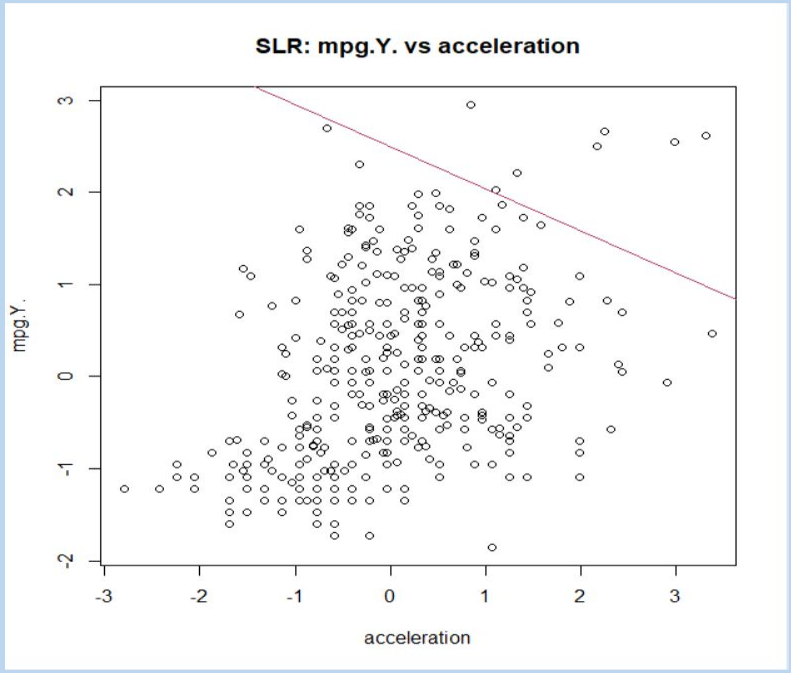


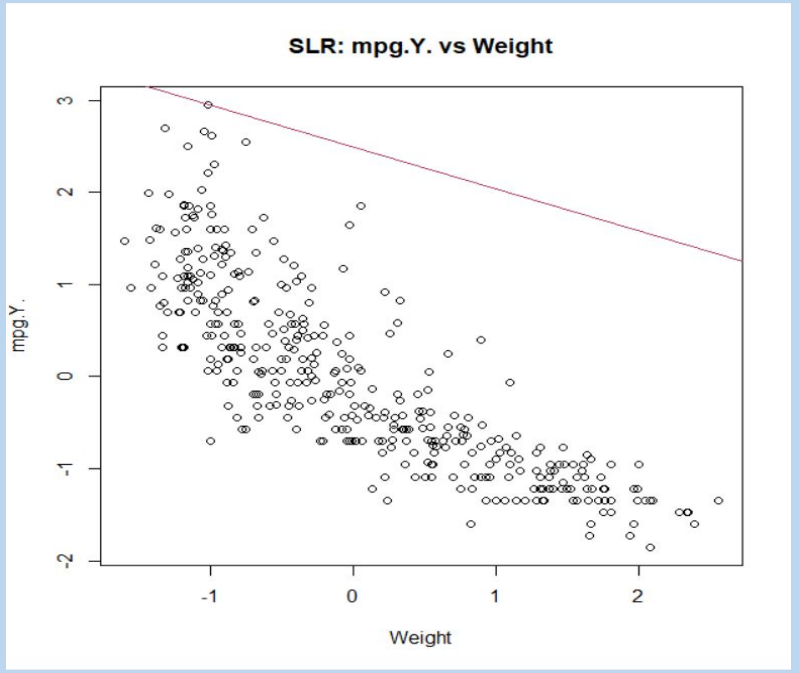
A series of simple linear regressions were run to predict mpg using different variables. **Weight** had the highest predictive power ($\mathbf{R^2 = 0.693}$) with a strong negative impact on mpg. **Displacement**, **HorsePower**, and **Cylinders** also showed strong negative relationships ($\mathbf{R^2}$ between 0.605–0.652). **Acceleration** had the weakest relationship ($\mathbf{R^2 = 0.181}$) and a positive coefficient. **ModelYear** and **Origin** positively influenced mpg, with $\mathbf{R^2}$ values of 0.351 and 0.322, respectively. All predictors were statistically significant ($p < 0.001$).

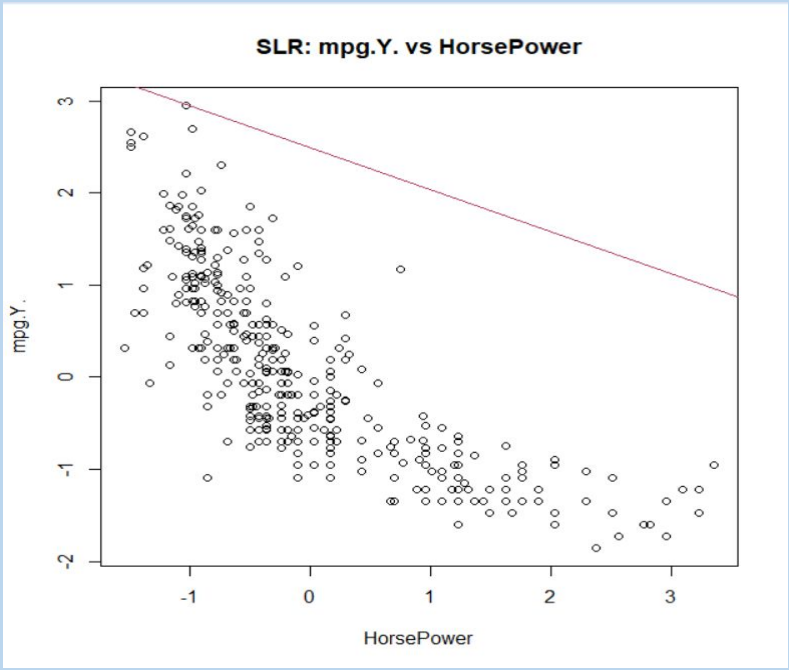
Simple Linear Regression Summary:

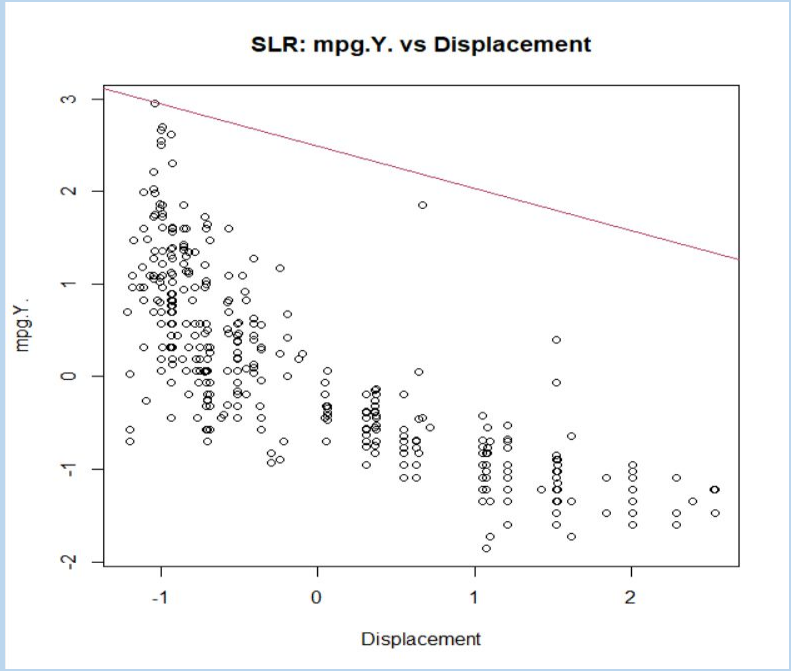
Predictor	Estimate (Slope)	R^2 Value	Adjusted R^2	Significance (p-value)	Strength of Relationship
Cylinders	-0.4565	0.6019	0.6009	< 2e-16	Strong Negative
Acceleration	0.4140	0.1714	0.1692	< 2e-16	Weak Positive
Model Year	0.1607	0.3412	0.3395	< 2e-16	Weak Positive
Weight	-0.8310	0.6905	0.6897	< 2e-16	Strong Negative
Horsepower	-0.7809	0.6099	0.6089	< 2e-16	Strong Negative
Displacement	-0.8058	0.6493	0.6484	< 2e-16	Strong Negative
Origin	0.7035	0.3194	0.3177	< 2e-16	Moderate Positive

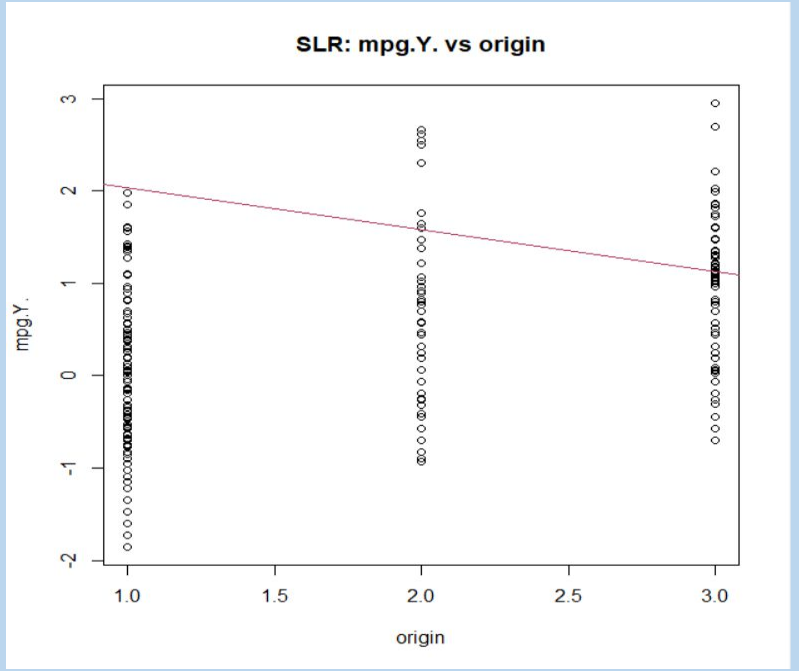

R Graphics: Device 2 (ACTIVE)
SLR: mpg.Y. vs Cylinders
mpg.Y.
Cylinders


SLR: mpg.Y. vs acceleration
mpg.Y.
acceleration

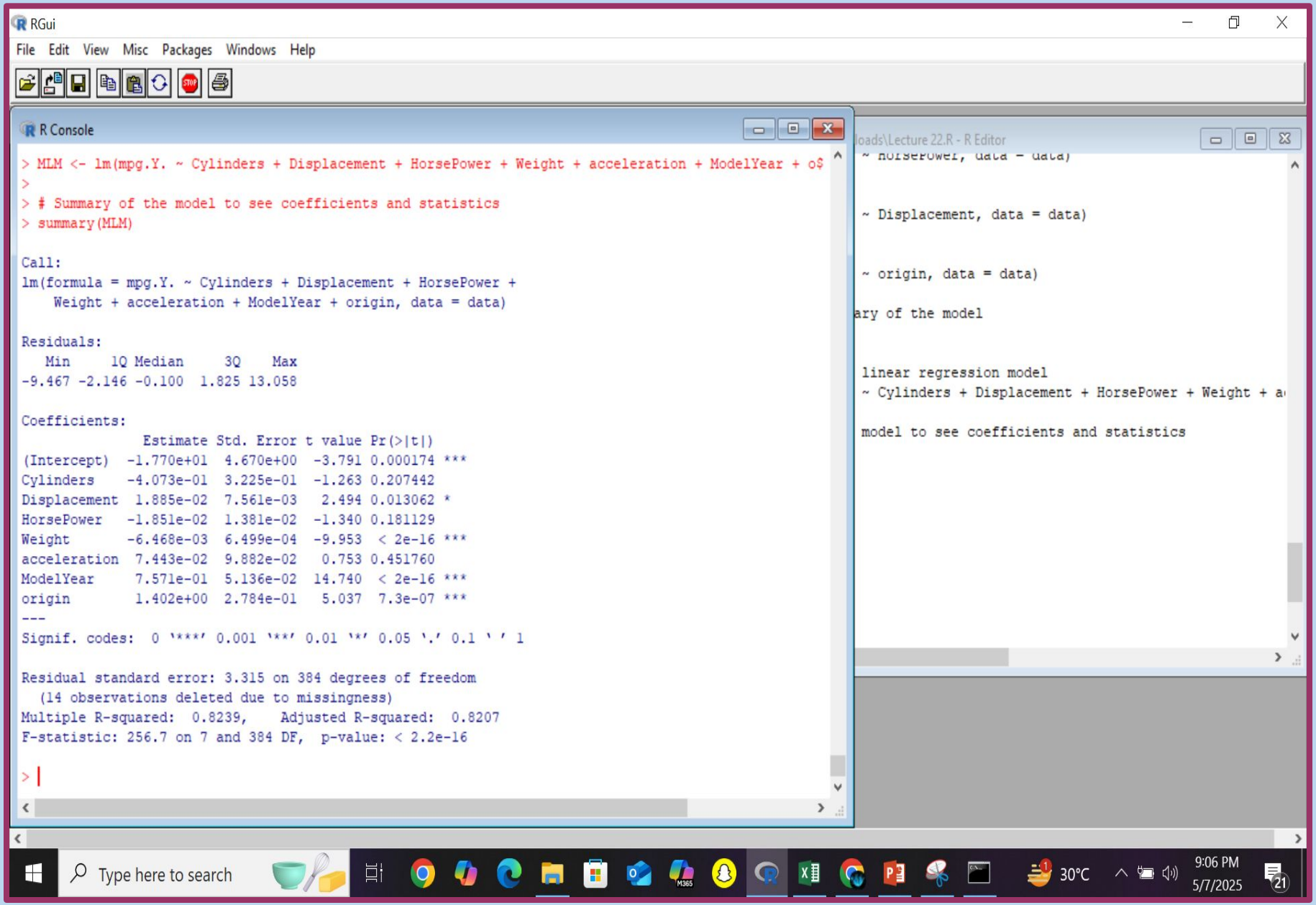

SLR: mpg.Y. vs Weight
mpg.Y.
Weight


SLR: mpg.Y. vs HorsePower
mpg.Y.
HorsePower

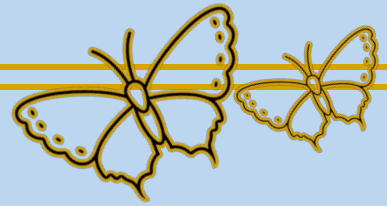

SLR: mpg.Y. vs Displacement
mpg.Y.
Displacement


SLR: mpg.Y. vs origin
mpg.Y.
origin

Multiple linear regression model:



A multiple linear regression was run to predict mpg. The model explained **82.07%** of the variance (**$R^2 = 0.8207$**). **Weight**, **ModelYear**, and **Origin** were highly significant predictors ($p < 0.001$), with **Weight** showing a strong negative effect and **ModelYear** a moderate positive one. **Displacement** was also significant ($p = 0.013$), while **Cylinders**, **HorsePower**, and **Acceleration** were not statistically significant.



Conclusion

- **MPG** (Miles Per Gallon) is a measure of a vehicle's fuel efficiency, the higher the MPG, the more distance a car can travel using less fuel.
- Using **Multiple Linear Regression**, we analyzed how various car features affect MPG.
- **Variables included:**
  - Cylinders
  - Displacement
  - HorsePower
  - Weight
  - Acceleration
  - ModelYear
  - Origin

A1 | mpg(Y)

	A	B	C	D	E	F	G	H	I	J	K
1	mpg(Y)	Cylinders	Displacem	HorsePow	Weight	accelerati	ModelYear	origin	CarName		
2	18	8	307	130	3504	12	70	1	chevrolet chevelle malibu		
3	15	8	350	165	3693	11.5	70	1	buick skylark 320		
4	18	8	318	150	3436	11	70	1	plymouth satellite		
5	17	8	302	140	3449	10.5	70	1	ford torino		
6	15	8	429	198	4341	10	70	1	ford galaxie 500		
7	14	8	454	220	4354	9	70	1	chevrolet impala		
8	16	8	304	150	3433	12	70	1	amc rebel sst		
9	15	8	383	170	3563	10	70	1	dodge challenger se		
10	14	8	340	160	3609	8	70	1	plymouth 'cuda 340		
11	15	8	400	150	3761	9.5	70	1	chevrolet monte carlo		
12	14	8	455	225	3086	10	70	1	buick estate wagon (sw)		
13	22	6	198	95	2833	15.5	70	1	plymouth duster		
14	18	6	199	97	2774	15.5	70	1	amc hornet		
15	21	6	200	85	2587	16	70	1	ford maverick		

📊 Key Findings:

- Cylinders: -0.4565→ Each additional cylinder in the engine decreases MPG by approximately 0.4565. (Interpretation: Strong Negative Relationship — more cylinders = less fuel efficiency.)
- Acceleration: 0.4140→ For each additional unit of acceleration, MPG increases by about 0.4140. (Interpretation: Weak Positive Relationship — slightly better MPG with faster acceleration.)
- Model Year: 0.1607→ Each year increase in model year improves MPG by 0.1607. (Interpretation: weak Positive — newer cars tend to be more fuel-efficient.)
- Weight: -0.8310→ For every additional unit of weight, MPG drops by 0.8310. (Interpretation: Very Strong Negative — heavier cars burn more fuel.)
- Horsepower: -0.7809→ Each unit increase in horsepower reduces MPG by 0.7809. (Interpretation: Strong Negative — more power, more fuel consumption.)
- Displacement: -0.8058→ For every extra unit of engine displacement, MPG decreases by 0.8058. (Interpretation: Very Strong Negative — bigger engines are less efficient.)
- Origin: 0.7035→ Vehicles from different origins (regions) increase MPG by about 0.7035 per unit change. (Interpretation: Moderate Positive — cars from some origins (e.g., Japan) may be more fuel-efficient.)

CONCLUSION:

The multiple linear regression model developed to predict miles per gallon (MPG) based on various vehicle characteristics—including weight, horsepower, displacement, cylinders, acceleration, and model year, demonstrates strong predictive performance.

With an R^2 value of **0.8184**, the model explains approximately **82%** of the **variability in MPG**, indicating a strong linear relationship between the predictors and the response variable.

Among the predictors, Weight and Cylinders showed the most significant individual impact on fuel efficiency. The model suggests that **lighter** and **newer** vehicles generally offer better MPG.

Overall, this regression analysis provides valuable insights into the key factors affecting a car's fuel economy.