

Data Analysis Practice

Michael V Cumbo

January 30, 2024

Objective

Explore the relationship between miles per gallon (*mpg*) and other variables in the `mtcars` dataset.

Tasks

1. Load the `mtcars` Dataset

- Start by loading the `mtcars` dataset into R. This dataset comes pre-loaded in R, so you don't need to download it from anywhere. Just use `data(mtcars)` to load it.

2. Basic Exploration

- Display the first few rows of the dataset using the `head()` function.
- Use the `summary()` function to get a summary of the dataset.

```
library(tidyverse)
library(dplyr)

mtcars <- mtcars

head(mtcars)

##           mpg cyl  disp  hp drat   wt  qsec vs am gear carb
## Mazda RX4      21.0   6  160 110 3.90 2.620 16.46  0  1    4    4
## Mazda RX4 Wag  21.0   6  160 110 3.90 2.875 17.02  0  1    4    4
## Datsun 710      22.8   4  108  93 3.85 2.320 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 3.440 17.02  0  0    3    2
## Valiant         18.1   6  225 105 2.76 3.460 20.22  1  0    3    1
##           mpg_category
## Mazda RX4           High MPG
## Mazda RX4 Wag       High MPG
## Datsun 710           High MPG
## Hornet 4 Drive       High MPG
## Hornet Sportabout    Low MPG
## Valiant              Low MPG
```

```
mtcars %>%
  summary()
```

##	mpg	cyl	disp	hp
##	Min. :10.40	Min. :4.000	Min. : 71.1	Min. : 52.0
##	1st Qu.:15.43	1st Qu.:4.000	1st Qu.:120.8	1st Qu.: 96.5
##	Median :19.20	Median :6.000	Median :196.3	Median :123.0
##	Mean :20.09	Mean :6.188	Mean :230.7	Mean :146.7
##	3rd Qu.:22.80	3rd Qu.:8.000	3rd Qu.:326.0	3rd Qu.:180.0
##	Max. :33.90	Max. :8.000	Max. :472.0	Max. :335.0
##	drat	wt	qsec	vs
##	Min. :2.760	Min. :1.513	Min. :14.50	Min. :0.0000
##	1st Qu.:3.080	1st Qu.:2.581	1st Qu.:16.89	1st Qu.:0.0000
##	Median :3.695	Median :3.325	Median :17.71	Median :0.0000
##	Mean :3.597	Mean :3.217	Mean :17.85	Mean :0.4375
##	3rd Qu.:3.920	3rd Qu.:3.610	3rd Qu.:18.90	3rd Qu.:1.0000
##	Max. :4.930	Max. :5.424	Max. :22.90	Max. :1.0000
##	am	gear	carb	mpg_category
##	Min. :0.0000	Min. :3.000	Min. :1.000	Length:32
##	1st Qu.:0.0000	1st Qu.:3.000	1st Qu.:2.000	Class :character
##	Median :0.0000	Median :4.000	Median :2.000	Mode :character
##	Mean :0.4062	Mean :3.688	Mean :2.812	
##	3rd Qu.:1.0000	3rd Qu.:4.000	3rd Qu.:4.000	
##	Max. :1.0000	Max. :5.000	Max. :8.000	

3. Data Analysis

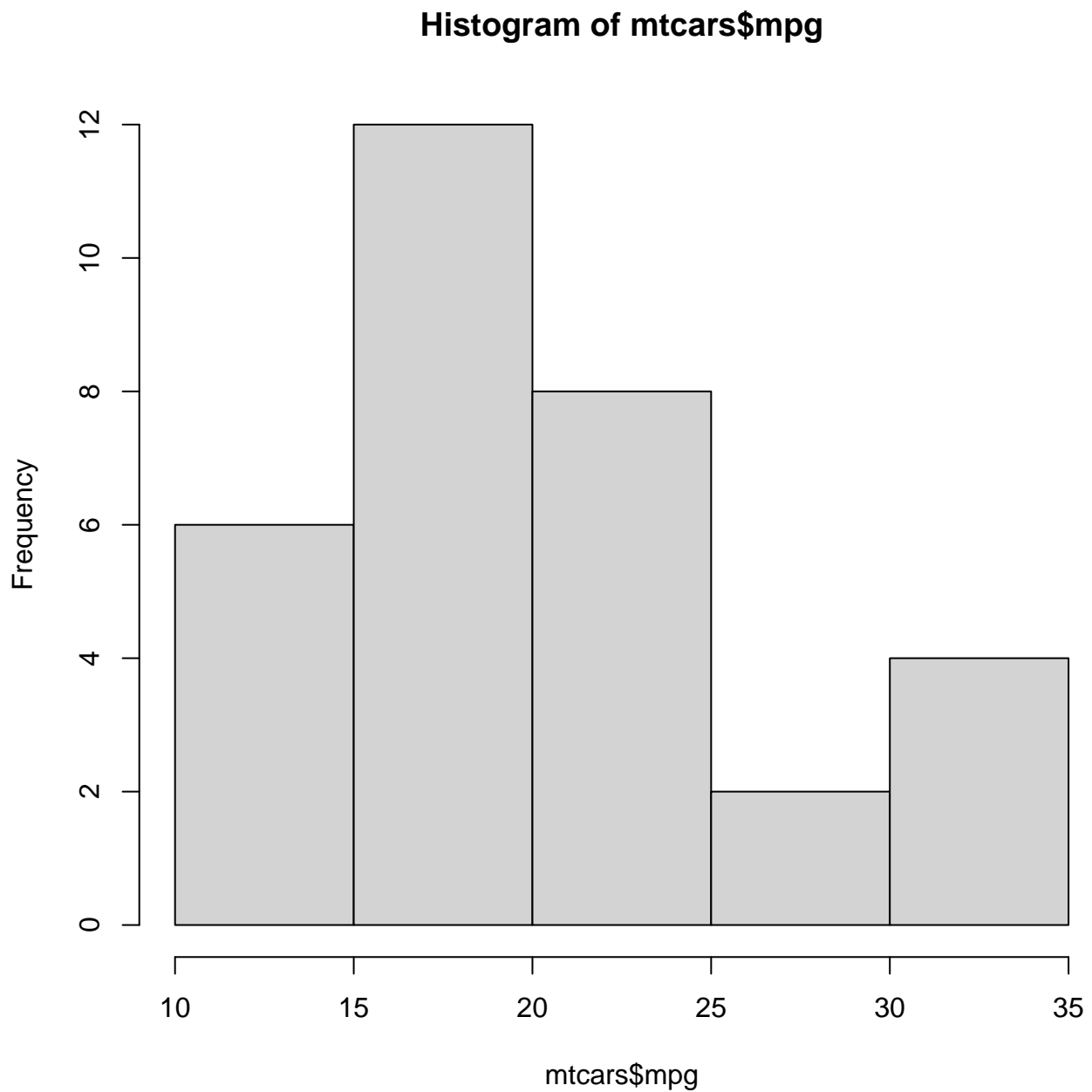
- Create a new column in the dataset that categorizes cars into "High MPG" and "Low MPG" based on whether their *mpg* is above or below the median *mpg* of all cars in the dataset. You can use the `ifelse()` function for this.

```
mtcars <- mtcars %>%
  mutate(mpg_category = ifelse(mpg > 20.09, "High MPG", "Low MPG"))
```

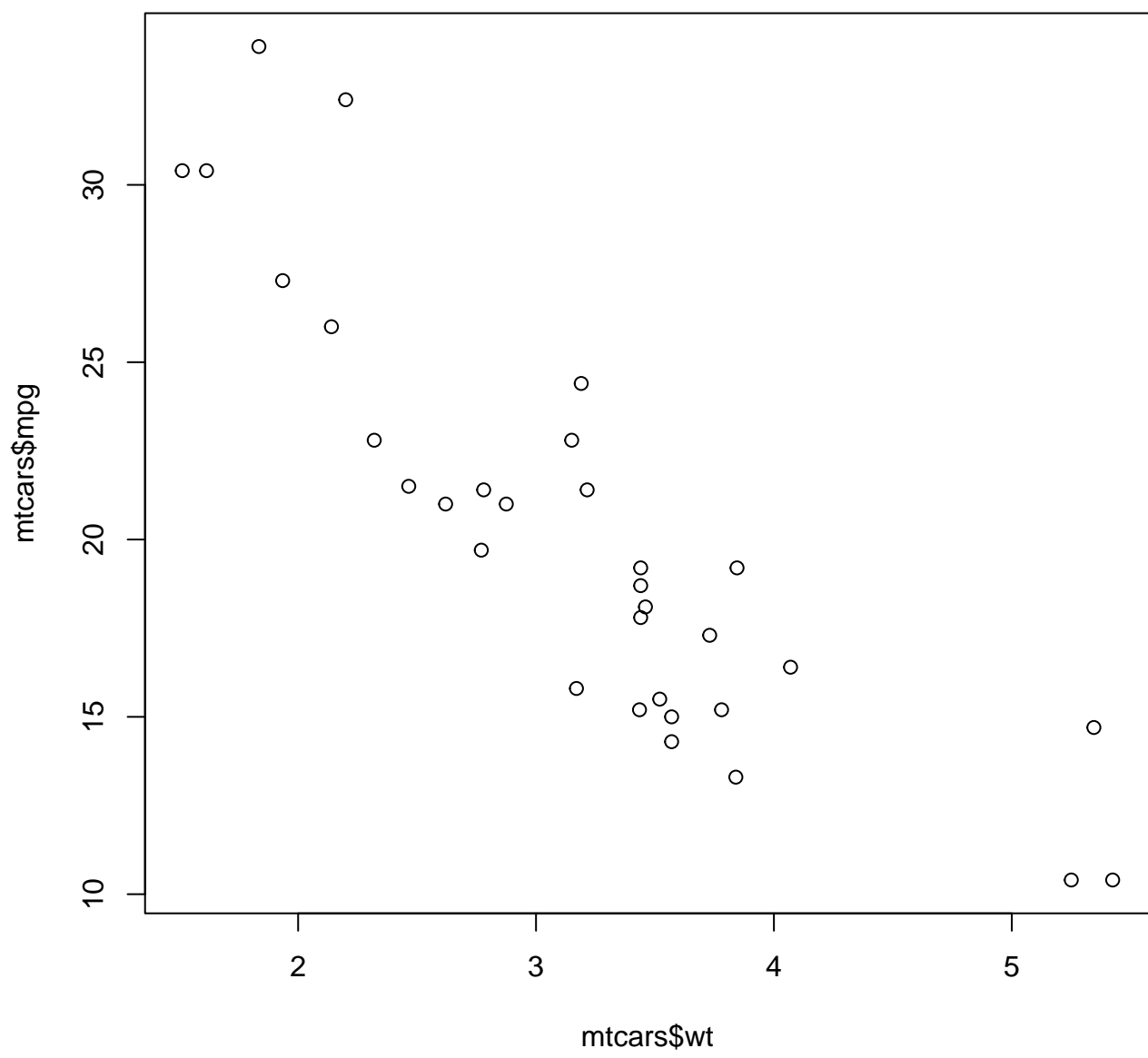
4. Visualization

- Plot a histogram of *mpg* to see its distribution.
- Create a scatter plot to examine the relationship between *mpg* and weight (*wt*).
- Bonus: Color the points in your scatter plot based on the "High MPG" and "Low MPG" categorization.

```
hist(mtcars$mpg)
```



```
plot(mtcars$wt, mtcars$mpg)
```



5. Advanced Analysis (Optional)

- Perform a linear regression analysis to study the relationship between *mpg* (as the dependent variable) and other variables like weight (*wt*), horsepower (*hp*), and number of cylinders (*cyl*). Use the `lm()` function for this.
- Summarize your linear regression model using the `summary()` function and interpret the results.

```
library(tidyverse)
# Define the independent variables
independent_vars <- c("hp", "cyl", "drat", "qsec", "vs", "carb")
# Create a list of formulas
formulas <- lapply(
  independent_vars,
  function(var) as.formula(paste("mpg ~", var))
)
# Use map() to apply lm() to each formula
models <- map(formulas, ~ lm(data = mtcars, formula = .))
# Create a tibble with model summaries
model_summaries <- tibble(variable = independent_vars, model = models) %>%
  mutate(summary = map(model, summary))
# View the tibble
print(model_summaries)

## # A tibble: 6 x 3
##   variable model    summary
##   <chr>      <list> <list>
## 1 hp        <lm>    <summary.lm>
## 2 cyl       <lm>    <summary.lm>
## 3 drat      <lm>    <summary.lm>
## 4 qsec      <lm>    <summary.lm>
## 5 vs        <lm>    <summary.lm>
## 6 carb      <lm>    <summary.lm>

# Extract and print the summary
# for the model with 'hp' as the independent variable
hp_model_summary <- model_summaries %>%
  filter(variable == "hp") %>%
  pull(summary)
# Display the summary
print(hp_model_summary)

##
## Call:
## lm(formula = ., data = mtcars)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -5.7121 -2.1122 -0.8854  1.5819  8.2360
##
## Coefficients:
##              Estimate Std. Error
## (Intercept) 30.09886    1.63392
```

```

## hp          -0.06823    0.01012
##              t value Pr(>|t|)
## (Intercept)  18.421 < 2e-16 ***
## hp          -6.742 1.79e-07 ***
## ---
## Signif. codes:
##  0 '***' 0.001 '**' 0.01 '*'
##  0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.863 on 30 degrees of freedom
## Multiple R-squared:  0.6024, Adjusted R-squared:  0.5892
## F-statistic: 45.46 on 1 and 30 DF, p-value: 1.788e-07
##
##
## Call:
## lm(formula = ., data = mtcars)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -4.9814 -2.1185  0.2217  1.0717
##  7.5186
##
## Coefficients:
##              Estimate Std. Error
## (Intercept)  37.8846     2.0738
## cyl         -2.8758     0.3224
##              t value Pr(>|t|)
## (Intercept)  18.27 < 2e-16 ***
## cyl         -8.92 6.11e-10 ***
## ---
## Signif. codes:
##  0 '***' 0.001 '**' 0.01 '*'
##  0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.206 on 30 degrees of freedom
## Multiple R-squared:  0.7262, Adjusted R-squared:  0.7171
## F-statistic: 79.56 on 1 and 30 DF, p-value: 6.113e-10
##
##
## Call:
## lm(formula = ., data = mtcars)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -9.0775 -2.6803 -0.2095  2.2976
##  9.0225
##
## Coefficients:
##              Estimate Std. Error
## (Intercept)  -7.525     5.477
## drat          7.678     1.507
##              t value Pr(>|t|)
## (Intercept)  -1.374     0.18
## drat         5.096 1.78e-05 ***

```

```
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*'
## 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.485 on 30 degrees of freedom
## Multiple R-squared: 0.464, Adjusted R-squared: 0.4461
## F-statistic: 25.97 on 1 and 30 DF, p-value: 1.776e-05
##
##
## Call:
## lm(formula = ., data = mtcars)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -9.8760 -3.4539 -0.7203  2.2774
## 11.6491
##
## Coefficients:
##              Estimate Std. Error
## (Intercept)  -5.1140    10.0295
## qsec          1.4121     0.5592
##              t value Pr(>|t|)
## (Intercept)  -0.510    0.6139
## qsec          2.525    0.0171 *
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*'
## 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.564 on 30 degrees of freedom
## Multiple R-squared: 0.1753, Adjusted R-squared: 0.1478
## F-statistic: 6.377 on 1 and 30 DF, p-value: 0.01708
##
##
## Call:
## lm(formula = ., data = mtcars)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -6.757 -3.082 -1.267  2.828  9.383
##
## Coefficients:
##              Estimate Std. Error
## (Intercept)   16.617     1.080
## vs            7.940     1.632
##              t value Pr(>|t|)
## (Intercept)  15.390 8.85e-16 ***
## vs           4.864 3.42e-05 ***
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*'
## 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.581 on 30 degrees of freedom
```

```

## Multiple R-squared:  0.4409, Adjusted R-squared:  0.4223
## F-statistic: 23.66 on 1 and 30 DF,  p-value: 3.416e-05
##
##
## Call:
## lm(formula = ., data = mtcars)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -7.250 -3.316 -1.433  3.384 10.083
##
## Coefficients:
##              Estimate Std. Error
## (Intercept)  25.8723     1.8368
## carb        -2.0557     0.5685
##              t value Pr(>|t|)
## (Intercept)  14.085 9.22e-15 ***
## carb        -3.616  0.00108 **
## ---
## Signif. codes:
##  0 '***' 0.001 '**' 0.01 '*'
##  0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.113 on 30 degrees of freedom
## Multiple R-squared:  0.3035, Adjusted R-squared:  0.2803
## F-statistic: 13.07 on 1 and 30 DF,  p-value: 0.001084

```


6. Reflection

- Write a brief summary of your findings. Which variables seem to affect *mpg* the most? Were there any surprises in your analysis?

Deliverables

- R script with your code and comments explaining each step.
- A brief report summarizing your findings.