39_Project3_Part3

Tilak Poudel

2025-04-19

Use "mpg" data of ggplot2 library and test the hypothesis that large engine cars (displ) have less highway milage (hwy) using scatterplot (geom_point) for overall plot and subcategories of disp variable with ggplot2 package. Then, perform smoothing on this scatterplot and explain the algorithm/method used to do smoothing by the ggplot2 package. Finally, use facet wrap and facet grid layers to show scatterplots for cyl and drv variables in ggplot2 and explain the results carefully.

```r
library(ggplot2)

data(mpg)
# Check the structure of the dataset
str(mpg)
```

```
## tibble [234 x 11] (S3: tbl_df/tbl/data.frame)
##  $ manufacturer: chr [1:234] "audi" "audi" "audi" "audi" ...
##  $ model       : chr [1:234] "a4" "a4" "a4" "a4" ...
##  $ displ       : num [1:234] 1.8 1.8 2 2 2.8 2.8 3.1 1.8 1.8 2 ...
##  $ year        : int [1:234] 1999 1999 2008 2008 1999 1999 2008 1999 1999 2008 ...
##  $ cyl         : int [1:234] 4 4 4 4 6 6 6 4 4 4 ...
##  $ trans       : chr [1:234] "auto(l5)" "manual(m5)" "manual(m6)" "auto(av)" ...
##  $ drv         : chr [1:234] "f" "f" "f" "f" ...
##  $ cty         : int [1:234] 18 21 20 21 16 18 18 18 16 20 ...
##  $ hwy         : int [1:234] 29 29 31 30 26 26 27 26 25 28 ...
##  $ fl          : chr [1:234] "p" "p" "p" "p" ...
##  $ class       : chr [1:234] "compact" "compact" "compact" "compact" ...
```

```r
# Check the first few rows of the dataset
head(mpg)
```

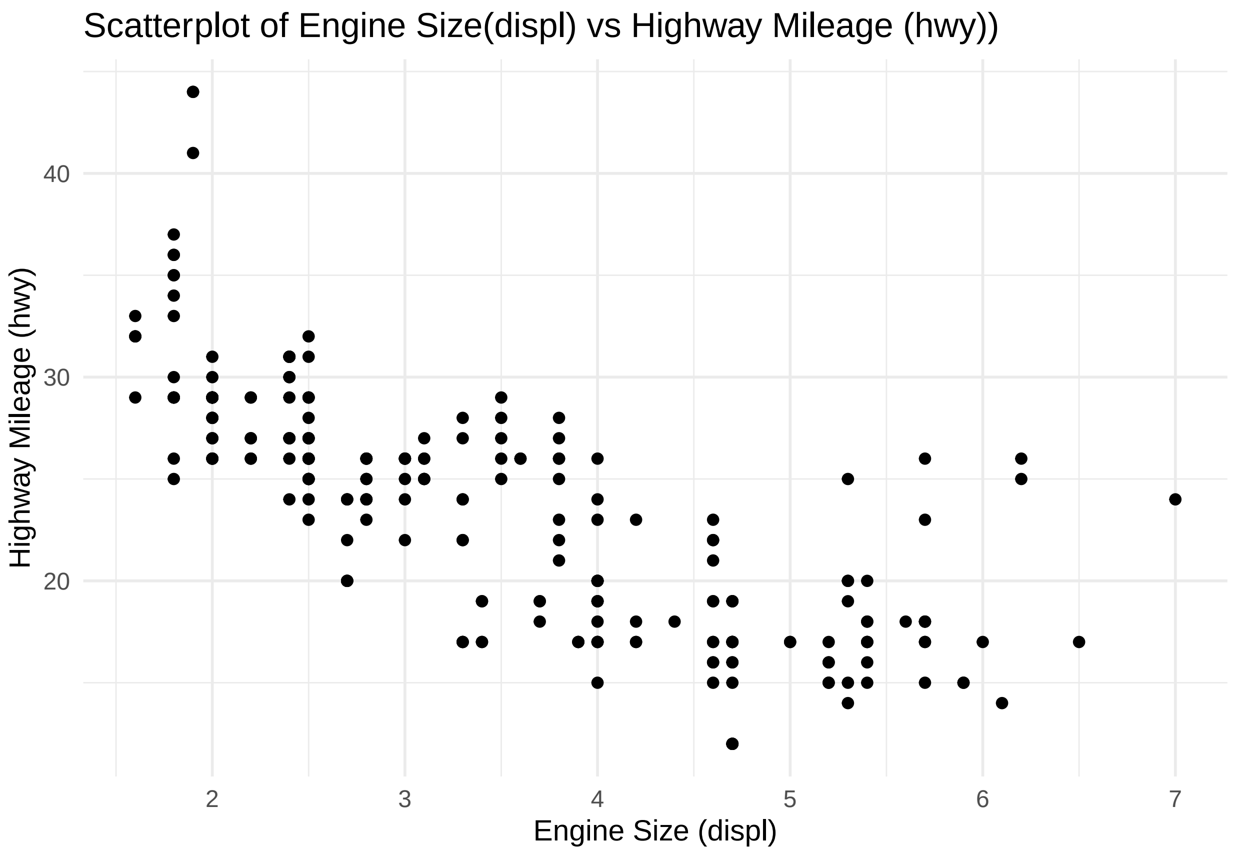
```
## # A tibble: 6 x 11
##   manufacturer model displ  year   cyl trans      drv     cty   hwy fl    class
##   <chr>        <chr> <dbl> <int> <int> <chr>      <chr> <int> <int> <chr> <chr>
## 1 audi         a4      1.8  1999     4 auto(l5)   f        18    29 p     compa~
## 2 audi         a4      1.8  1999     4 manual(m5) f        21    29 p     compa~
## 3 audi         a4      2    2008     4 manual(m6) f        20    31 p     compa~
## 4 audi         a4      2    2008     4 auto(av)   f        21    30 p     compa~
## 5 audi         a4      2.8  1999     6 auto(l5)   f        16    26 p     compa~
## 6 audi         a4      2.8  1999     6 manual(m5) f        18    26 p     compa~
```

Hypothesis testing

Hypothesis: Large engine cars (displ) have less highway mileage (hwy)

- Null Hypothesis (H0): There is no relationship between engine size (displ) and highway mileage (hwy).
- Alternative Hypothesis (H1): There is a negative relationship between engine size (displ) and highway mileage (hwy).

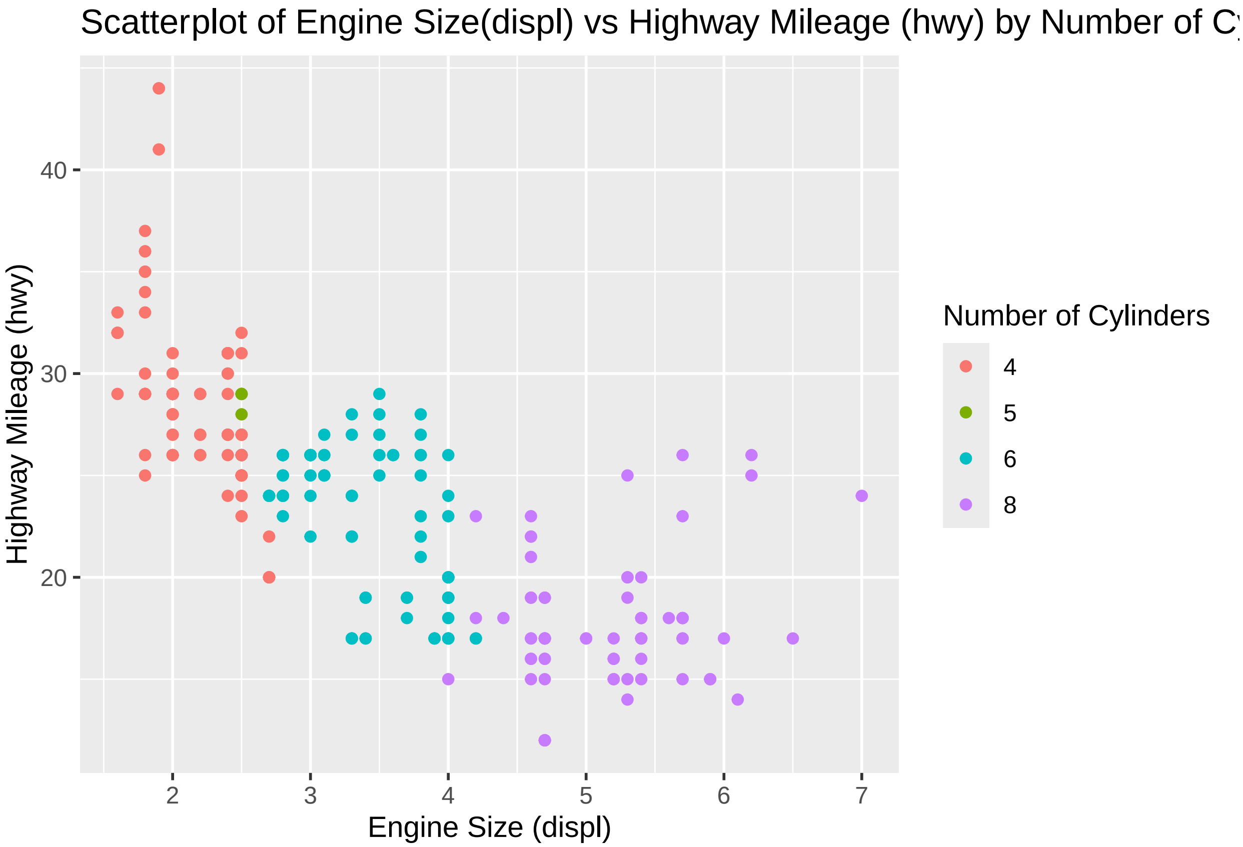
```r
# Scatter plot of displ vs hwy
ggplot(mpg, aes(x = displ, y = hwy)) +
  geom_point() +
  labs(title = "Scatterplot of Engine Size(displ) vs Highway Mileage (hwy))",
       x = "Engine Size (displ)",
       y = "Highway Mileage (hwy)") +
  theme_minimal()
```



We observe a negative correlation — as `displ` increases, `hwy` tends to decrease. `This supports the hypothesis that larger engines have lower fuel efficiency.`

Plot the scatter plot with subcategories of displ

```r
ggplot(mpg, aes(x = displ, y = hwy)) +
  geom_point(aes(color = factor(cyl))) +
  labs(title = "Scatterplot of Engine Size(displ) vs Highway Mileage (hwy) by Number of Cylinders",
       x = "Engine Size (displ)",
       y = "Highway Mileage (hwy)",
       color = "Number of Cylinders")
```

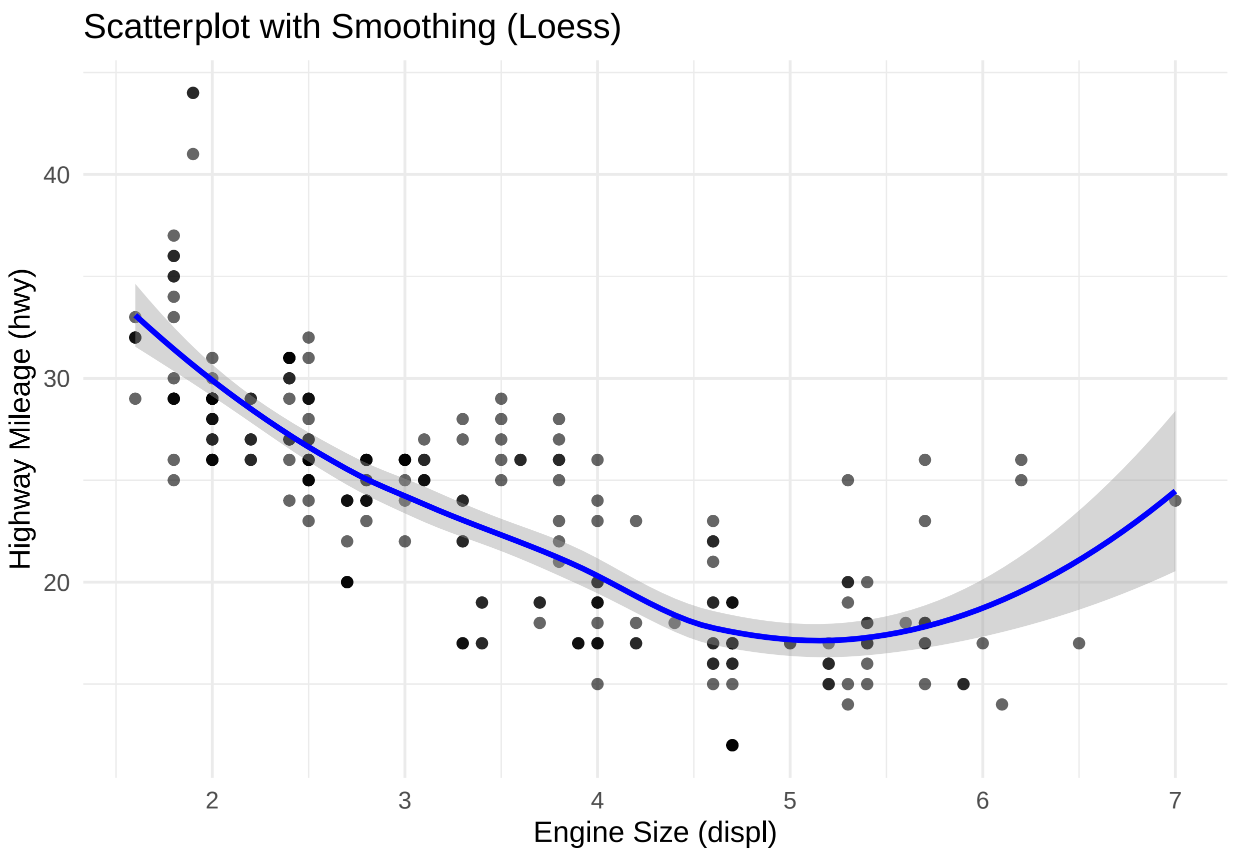


The scatter plot shows that cars with more cylinders (`cyl`) tend to have larger engine sizes (`displ`) and lower highway mileage (`hwy`).

Smoothing

```r
# Smoothing using geom_smooth
ggplot(mpg, aes(x = displ, y = hwy)) +
  geom_point(alpha=0.6) +
  geom_smooth(method = "loess", se = TRUE, color = "blue") +
  labs(title = "Scatterplot with Smoothing (Loess)",
       x = "Engine Size (displ)",
       y = "Highway Mileage (hwy)") +
  theme_minimal()
```

```
## ‘geom_smooth()‘ using formula = ’y ~ x’
```

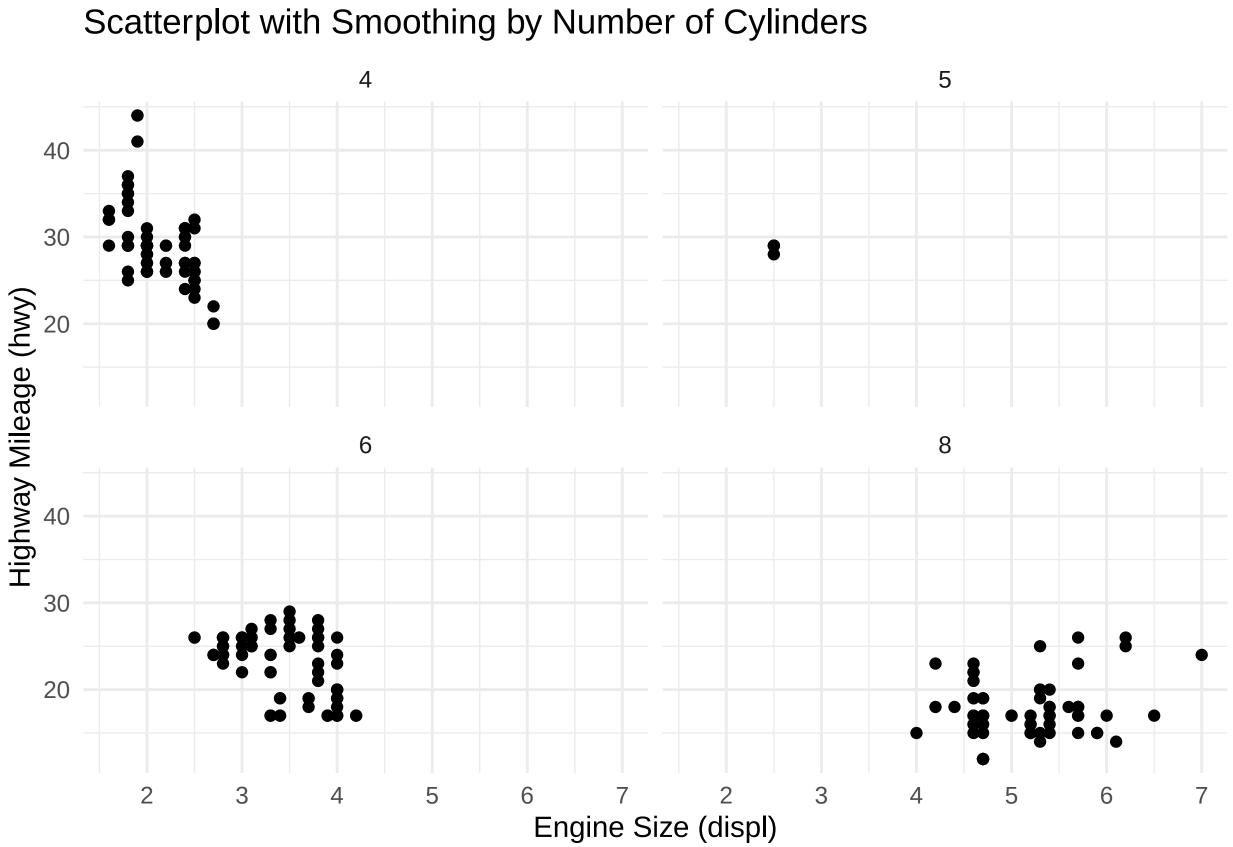


The `geom_smooth` function in ggplot2 uses a method called LOESS (Locally Estimated Scatterplot Smoothing) by default.LOESS is a non-parametric regression method that fits multiple regressions in localized subsets of the data. It is particularly useful for capturing non-linear relationships in scatterplots.

- The blue line represents the smoothed relationship between `displ` and `hwy`.
- The shaded area around the line represents the confidence interval for the smoothed line.

Facet Wrap

```r
# Facet wrap for cyl variable
ggplot(mpg, aes(x = displ, y = hwy)) +
  geom_point() +
  facet_wrap(~ cyl) +
  labs(title = "Scatterplot with Smoothing by Number of Cylinders",
       x = "Engine Size (displ)",
       y = "Highway Mileage (hwy)") +
  theme_minimal()
```

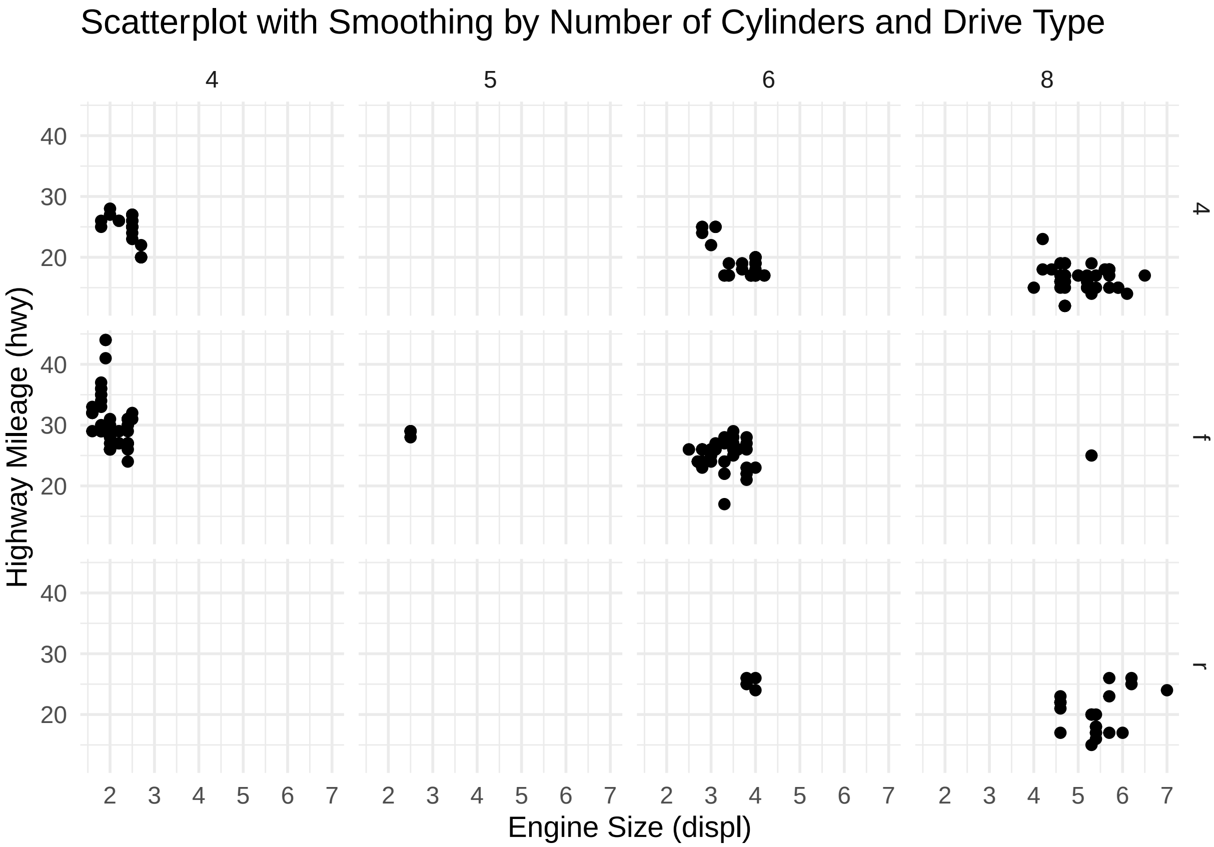


The facet wrap creates separate scatterplots for each level of the `cyl` variable. This allows us to see how the relationship between `displ` and `hwy` varies with the number of cylinders.

- The vechiles with less than 4 cylinders have higher hwy mileage while vechile with 8 cylinders have the least hwy mileage. Vechiles with 6 cylinders have moderate hwy mileage.

Facet Grid

```r
# Facet grid for cyl and drv variables
ggplot(mpg, aes(x = displ, y = hwy)) +
  geom_point() +
  facet_grid(drv ~ cyl) +
  labs(title = "Scatterplot with Smoothing by Number of Cylinders and Drive Type",
       x = "Engine Size (displ)",
       y = "Highway Mileage (hwy)") +
  theme_minimal()
```



The facet grid creates a grid of scatterplots based on the `cyl` and `drv` variables. This allows us to see how the relationship between `displ` and `hwy` varies with both the number of cylinders and the drive type.

- The scatterplots show that the relationship between displ and hwy is similar across different drive types (fwd, rwd, 4wd),but the overall trend is that larger engines (displ) tend to have lower highway mileage (hwy).
- Conclusion: The analysis shows that larger engine cars tend to have lower highway mileage.