Regression Analysis: MotorTrend_MPG

EHarris

7/18/2017

Packages for Course 7 Project: Motor Trend MPG

Overview

You work for Motor Trend, a magazine about the automobile industry. Looking at a data set of a collection of cars, they are interested in exploring the relationship between a set of variables and miles per gallon (MPG) (outcome). They are particularly interested in the following two questions:

- Is an automatic or manual transmission better for MPG?
- Quantify the MPG difference between automatic and manual transmissions"

A first step in the analysis is to look at the effect of transmission type, automatic vs. manual, on miles per gallon (MPG). Without much review of the dataset, it may be reasonable to expect that a manual transmission vehicle is associated with higher MPG. Further review will uncover other items, such as vehicle weight and the number of cylinders (e.g. 4, 6, 8), that may contribute more significantly to the mileage of a vehicle.

Load Data

Given that this is an R dataset we can simply use data(ToothGrowth) to load the dataset into the current enviromnent.

```
data(mtcars)
```

Exploratory Data Analysis

Our exploratory analysis provides some quick insigtht to the data of ‘mtcars’ data set. We provide a chart of the correlation for each variable and two panel plots. One of the panel plots, representing line graphs of the relationship for each variable, is illustrated in the Appendix. The other panel plot, illustrated below in Exhibit 2, provides additional information about the relationship of variables to mpg. The correlation matrix is used to determine which variables are included in the panel plot, based on correlation factor with absolute value > 0.5.

Summary view of dataset

```
str(mtcars)
```

```
## 'data.frame':    32 obs. of  11 variables:
##  $ mpg : num  21 21 22.8 21.4 18.7 18.1 14.3 24.4 22.8 19.2 ...
##  $ cyl : num  6 6 4 6 8 6 8 4 4 6 ...
##  $ disp: num  160 160 108 258 360 ...
##  $ hp  : num  110 110 93 110 175 105 245 62 95 123 ...
##  $ drat: num  3.9 3.9 3.85 3.08 3.15 2.76 3.21 3.69 3.92 3.92 ...
##  $ wt  : num  2.62 2.88 2.32 3.21 3.44 ...
##  $ qsec: num  16.5 17 18.6 19.4 17 ...
##  $ vs  : num  0 0 1 1 0 1 0 1 1 1 ...
##  $ am  : num  1 1 1 0 0 0 0 0 0 0 ...
##  $ gear: num  4 4 4 3 3 3 3 4 4 4 ...
##  $ carb: num  4 4 1 1 2 1 4 2 2 4 ...
```

```
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
```

Relationship of Variables

Three figures or exhibits are provided in the Appendix to provide insight to the data in ‘mtcars’. Figure 1 and Figure 2 compare each of the variables, or terms, in the data set against one another. Figure 1 provides a panel plot of the relationship between each variable. Similarly, Figure 2 shows the relationship of each variable in a correlation matrix.

Correlation Matrix (Upper Triangle)

This table provides an easy way to look at the relationship between the different variables of the ‘mtcars’ dataset. Along the diagonal, we see a value of 1, representing the correlation of a variable(e.g. mpg) and itself. A correlation factor, or value, is between -1 and 1. A positive correlation factor indicates that a variable changes directionally with another defined variable. A negative correlation suggests that two variables move in opposite directions. The closer the correlation factor is to zero indicates lower correlation. As correlation moves closer to -1 or 1, the stroger is the correlation or relationship.

Plot MPG Relationships

As the focus of the analysis is evaluating drivers or predictors of mpg, Figure 3 plots the relationship of each variable to mpg. Only 8 variables, rather than 10 available, are illustrated. The variables excluded from the plots have correlation factor <0.50 (absolute value). This provides a focus on those variables creating the most influence.

Regression Analysis

The focus of our analysis is miles per gallon (MPG). More specifically, we want to understand what variables or terms influence the mpg of a car. It might be reasonable and/or appropriate to include all variables within the 'mtcars' dataset to predict the mpg of a car. However, we likely want to explore and understand whether each variable contributes to the estimate. If we look at the correlation matrix, you can see that the variables 'cyl', 'disp', and 'wt' are similarly correlated with mpg, -0.85, -0.85, and -0.87 respectively. This strong relationship is further suggested when we look at the correlation factor between these variables: 'cyl'~'disp' = 0.9; 'cyl'~'wt' = 0.78; and 'disp'~'wt' = 0.89. We'll start with looking at 'mpg' related to the variable 'am', transmission (automatic vs. manual).

Automatic versus Manual Regression

The initial model looks at the effect of transmission type (automatic vs. manual) on the miles per gallon (MPG) of a car. It is important to note that the variable 'am' contains a 0 (automatic) or 1 (manual). So the baseline, or intercept, reflects the average mpg for a car with automatic transmission. The 'am' coefficient represents the average change in mpg for a car with manual transmission.

Exhibit 1: Simple Linear Regression (mpg ~ am)

This simple linear regression produces an intercept of 17.147 mpg. If the car has a manual transmission, we would estimate an additional 7.25 mpg. As previously stated, it is idealistic to assume that the type of transmission, automatic vs. manual, could define the mpg of a car. This is further validated by a relatively low adjusted r.square of 0.3385, meaning that roughly 34% of the change in mpg is explained by transmission type alone.

```
fit1 <- lm(mpg ~ factor(am), data = mtcars)
fit1coef <- round(summary(fit1)$coef, 4)
fit1glance <- round(glance(fit1)[1:6], 4)
fit1coef
```

```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  17.1474     1.1246 15.2475    0e+00
## factor(am)1   7.2449     1.7644  4.1061    3e-04
```

```
fit1glance
```

```
##   r.squared adj.r.squared sigma statistic p.value df
## 1    0.3598        0.3385 4.902   16.8603   3e-04  2
```

Automatic versus Manual (Multivariable Regression)

To build a multivariable regression that best fits the data, we could add one variable, or term, at a time. The goal is to avoid underfitting or overfitting the model. Therefore, we want to include variables that minimizes the residual, maximizes adjusted r.square, and statistically significant. We will review our exploratory analysis to help select which variables to add. As transmission (automatic vs. manual) is a key variable of the analysis, it is retained in all regression models developed.

Exhibit 2: Multivariable Regression (mpg ~ am + wt)

The next variable we add to the regression model is ‘wt’. This variable has highest correlation, -0.87, with mpg. This model produces an adjusted r.square of 0.7358, explaining more than twice as much of the mpg variation than transmission alone. The residual squared error (sigma) is also reduced from 4.9 to 3.1. The addition of ‘wt’ also changes the coefficients for the intercept and transmission (manual), as anticipated

```
fit2 <- lm(mpg ~ factor(am) + wt, data = mtcars)
fit2coef <- round(summary(fit2)$coef, 4)
fit2glance <- round(glance(fit2)[1:6], 4)
fit2coef
```

```
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  37.3216     3.0546 12.2180   0.0000
## factor(am)1  -0.0236     1.5456 -0.0153   0.9879
## wt           -5.3528     0.7882 -6.7908   0.0000
```

```
fit2glance
```

```
##   r.squared adj.r.squared  sigma statistic p.value df
## 1    0.7528        0.7358 3.0979   44.1652       0  3
```

Exhibit 3: Multivariable Regression (mpg ~ am + wt + cyl)

In the following regression model, we added the variable ‘cyl’ as a factor, due to limited number of discrete values (4, 6, 8). This regression model improves the adjusted r.square to 0.8134 and reduces the residual (sigma) to 2.6032 while maintaining p-value = 0 (rounded to 4 decimal places). Each of the additional variables (wt, cyl6, cyl8) is significant at the 0.05 threshold.

```
fit3 <- lm(mpg ~ factor(am) + wt + factor(cyl), data = mtcars)
fit3coef <- round(summary(fit3)$coef, 4)
fit3glance <- round(glance(fit3)[1:6], 4)
fit3coef
```

```
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   33.7536     2.8135 11.9971   0.0000
## factor(am)1    0.1501     1.3002  0.1154   0.9089
## wt            -3.1496     0.9080 -3.4685   0.0018
## factor(cyl)6  -4.2573     1.4112 -3.0167   0.0055
## factor(cyl)8  -6.0791     1.6837 -3.6105   0.0012
```

```
fit3glance
```

```
##   r.squared adj.r.squared  sigma statistic p.value df
## 1    0.8375        0.8134 2.6032   34.7917       0  5
```

Exhibit 4: Multivariable Regression (mpg ~ am + wt + cyl + hp)

In this regression model, we add the variable ‘hp’. Besides ‘drat’ this variable has the most significant correlation with mpg of the remaining variables. The variable ‘drat’ was considered. However, it failed to meet several important criteria, specifically: input p-value 0.8613 not significant; adjusted r.square decreases rather than increase; and the residual (sigma) increases rather than decrease.

For reasons indicated above, we selected to add the variable ‘hp’ to the regression model instead of ‘drat’. The addition of ‘hp’ improves the adjusted r.square to 0.8401 and reduces the residual to 2.4101. The only concern is the effect it has on factor(cyl)8. Intuitively, we would expect the coeffient for ‘cyl8’ to reduce mpg more than ‘cyl6’. This implies a different interaction between ‘hp’ and ‘cyl8’. As a result, the p-value, 0.3523, for factor(cyl)8 is no longer significant at 0.05 level.

```
fit4 <- lm(mpg ~ factor(am) + wt + factor(cyl) + hp, data = mtcars)
fit4coef <- round(summary(fit4)$coef, 4)
fit4glance <- round(glance(fit4)[1:6], 4)
fit4coef
```

```
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    33.7083     2.6049 12.9404   0.0000
## factor(am)1     1.8092     1.3963  1.2957   0.2065
## wt             -2.4968     0.8856 -2.8194   0.0091
## factor(cyl)6   -3.0313     1.4073 -2.1540   0.0407
## factor(cyl)8   -2.1637     2.2843 -0.9472   0.3523
## hp             -0.0321     0.0137 -2.3450   0.0269
```

```
fit4glance
```

```
##   r.squared adj.r.squared  sigma statistic p.value df
## 1    0.8659        0.8401 2.4101   33.5712       0  6
```

Exhibit 8: Multivariable Regression (include all independent variables)

The purpose of this exhibit is to explore a regression model that includes all variables, or terms, in the ‘mtcars’ data set. Although we did not include an variables as a factor, it is clear that including all variables does not produce a better model. The adjusted r.square and residual are both worse than including only the vaiables am, wt, cyl (factor), and hp. Additionally, each of the variables has a p-value that is not significant at 0.05 level.

```
fit10 <- lm(mpg ~ ., data = mtcars)
fit10coef <- round(summary(fit10)$coef, 4)
fit10glance <- round(glance(fit10)[1:6], 4)
fit10coef
```

```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  12.3034    18.7179  0.6573   0.5181
## cyl          -0.1114     1.0450 -0.1066   0.9161
## disp          0.0133     0.0179  0.7468   0.4635
## hp           -0.0215     0.0218 -0.9868   0.3350
## drat          0.7871     1.6354  0.4813   0.6353
## wt           -3.7153     1.8944 -1.9612   0.0633
## qsec          0.8210     0.7308  1.1234   0.2739
## vs            0.3178     2.1045  0.1510   0.8814
## am            2.5202     2.0567  1.2254   0.2340
## gear          0.6554     1.4933  0.4389   0.6652
## carb         -0.1994     0.8288 -0.2406   0.8122
```

```
fit10glance
```

```
##   r.squared adj.r.squared  sigma statistic p.value df
## 1     0.869        0.8066 2.6502   13.9325       0 11
```

Exhibit 9: ANOVA

The findings provided below for Model 4: mpg ~ factor(am) + wt + factor(cyl) + hp produce a p-value of 0.026935, which indicates that this model is significant. These variables further contribute to the accuracy of the model.

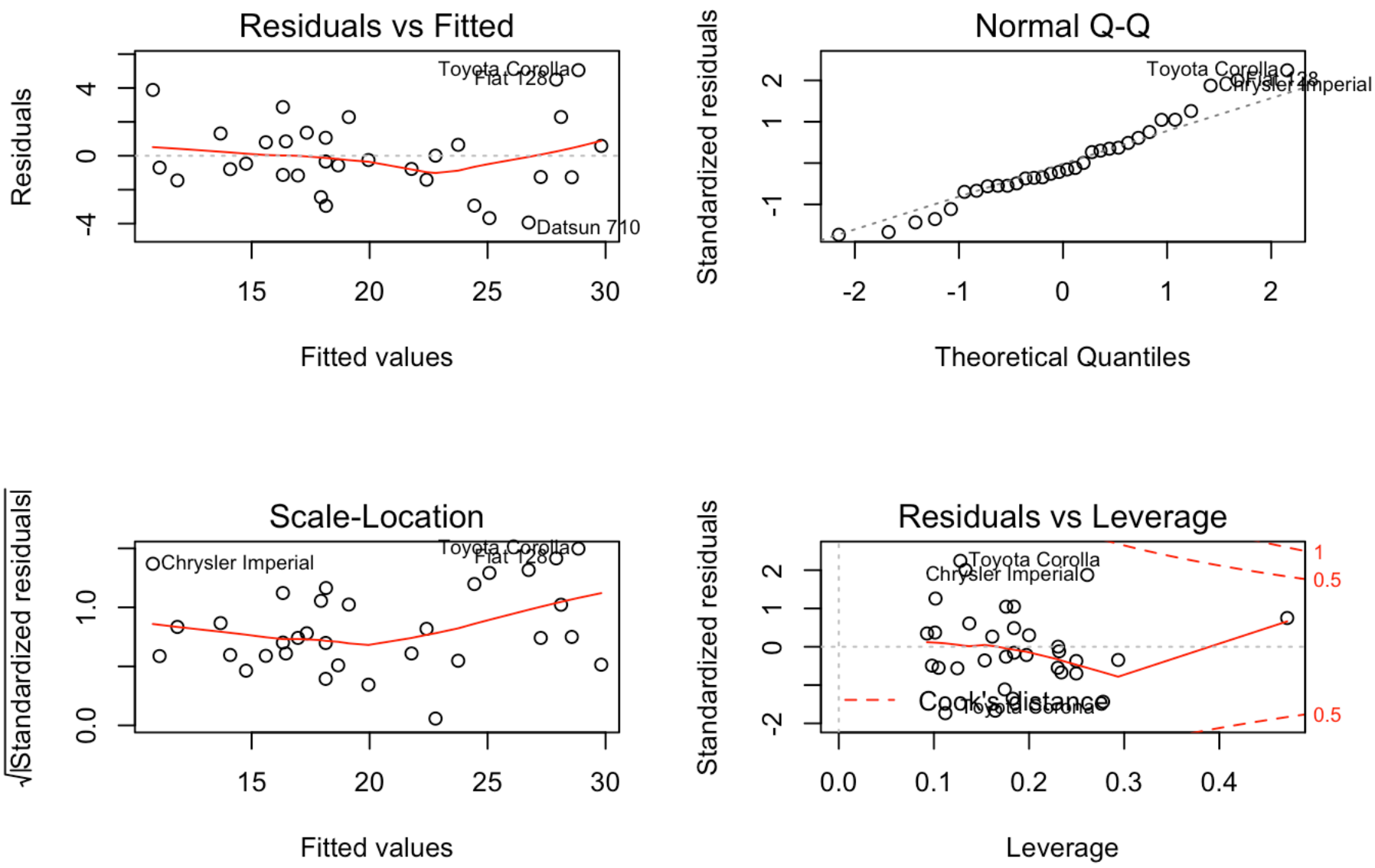
```
mpg.anova <- anova(fit1, fit2, fit3, fit4)
mpg.anova
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ factor(am)
## Model 2: mpg ~ factor(am) + wt
## Model 3: mpg ~ factor(am) + wt + factor(cyl)
## Model 4: mpg ~ factor(am) + wt + factor(cyl) + hp
##   Res.Df    RSS Df Sum of Sq       F   Pr(>F)
## 1     30 720.90
## 2     29 278.32  1    442.58 76.1924 3.32e-09 ***
## 3     27 182.97  2     95.35  8.2077 0.001725 **
## 4     26 151.03  1     31.94  5.4991 0.026935 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Residuals & Diagnostics

Below is a panel plot of residuals. Understanding residuals is critical to understanding how well the regression model fits the data, goodness of fit. The different plots provide some insight to

```
par(mfrow = c(2,2))
plot(fit4)
```



Residuals vs Fitted

This plot shows if residuals have non-linear patterns. The fact that the residuals are fairly well-distributed around the fitted line, and no particular pattern, we are fairly confident that the selected regression model is a relatively good fit.

Normal Q-Q

This plot shows if residuals are normally distributed. It's good fit if residuals are lined well on the straight dashed line. In our plot, the residuals are fairly close to the straight line. However, there appears to be a little greater distance for outliers, particularly upper right corner. We may want to explore outliers a little further.

Scale-Location

We use the scale-location plot to check the assumption of equal variance (homoscedasticity). Ideally, we would see points that are equally spread around a horizontal line along the range of predictors. Although not a horizontal line, the fitted line does not show much slope up or down. Additionally, the residual points are fairly evenly distributed about the line, no narrowing or spreading.

Residuals vs. Leverage

This is the final plot to examine. The purpose of this plot is identify “influential” outliers. If an outlier does not change, maybe minimally changes, the regression line whether it is included or excluded, the outlier is not “influential”. If outliers, residual points, appear in the upper right or lower right corner of the plot, thee points may be influential against the regression line.

Additional Regression Diagnostics

In addition to plots provided above, we can look evaluate our model through some other tests. Below, we look at a couple and explain the findings.

Hat Values

Conclusion

Following our review of Exploratory Data Analysis, it was decided to test two assumptions. From the boxplot, it appeared that the supplement ‘OJ’ promoted greater tooth growth than the supplement ‘VC’. It was also apparent that a higher dose promoted greater tooth growth, regardless of supplement introduced. We chose to test these assumptions by setting NULL hypothesis that there was no difference in the tooth growth for supplement ‘OJ’ versus ‘VC’ and a higher dose versus lower dose. The alternative hypothesis being that ‘OJ’ promotes greater tooth growth than ‘VC’ and a higher dose promotes greater tooth growth than a lower dose.

From our t.test, the p-values were rather small, with the exception of testing whether ‘OJ’ promotes greater tooth growth than ‘VC’ at a dose = 2.0, indicating that we reject the NULL hypothesis. In summary, the analysis suggests that ‘OJ’ promotes greater tooth growth than ‘VC’. Additionally, regardless of the supplement selected, ‘OJ’ or ‘VC’, a higher dose of the supplement promotes greater tooth growth than lower dose.

Appendix

The information provided below is an outline of the request or question driving this assingment and analysis of data for different cars and related miles per gallon (MPG).

Instructions

You work for Motor Trend, a magazine about the automobile industry. Looking at a data set of a collection of cars, they are interested in exploring the relationship between a set of variables and miles per gallon (MPG) (outcome). They are particularly interested in the following two questions:

“Is an automatic or manual transmission better for MPG” “Quantify the MPG difference between automatic and manual transmissions”

Review criteria

Peer Grading

The criteria that your classmates will use to evaluate and grade your work are shown below. Each criteria is binary: (1 point = criteria met acceptably; 0 points = criteria not met acceptably)

Criteria

1. Did the student interpret the coefficients correctly?
2. Did the student do some exploratory data analyses?
3. Did the student fit multiple models and detail their strategy for model selection?
4. Did the student answer the questions of interest or detail why the question(s) is (are) not answerable?
5. Did the student do a residual plot and some diagnostics?
6. Did the student quantify the uncertainty in their conclusions and/or perform an inference correctly?
7. Was the report brief (about 2 pages long) for the main body of the report and no longer than 5 with supporting appendix of figures?
8. Did the report include an executive summary?
9. Was the report done in Rmd (knitr)?

Figure 1 (Panel Plot): Illustrate the relationship of each variable

This panel plot provides a quick, high-level, illustration of a variable within the dataset against each other variable in the dataset. The pu

```
pairs(mtcars, panel = panel.smooth, main = "MT Cars Data", col = 3)
```

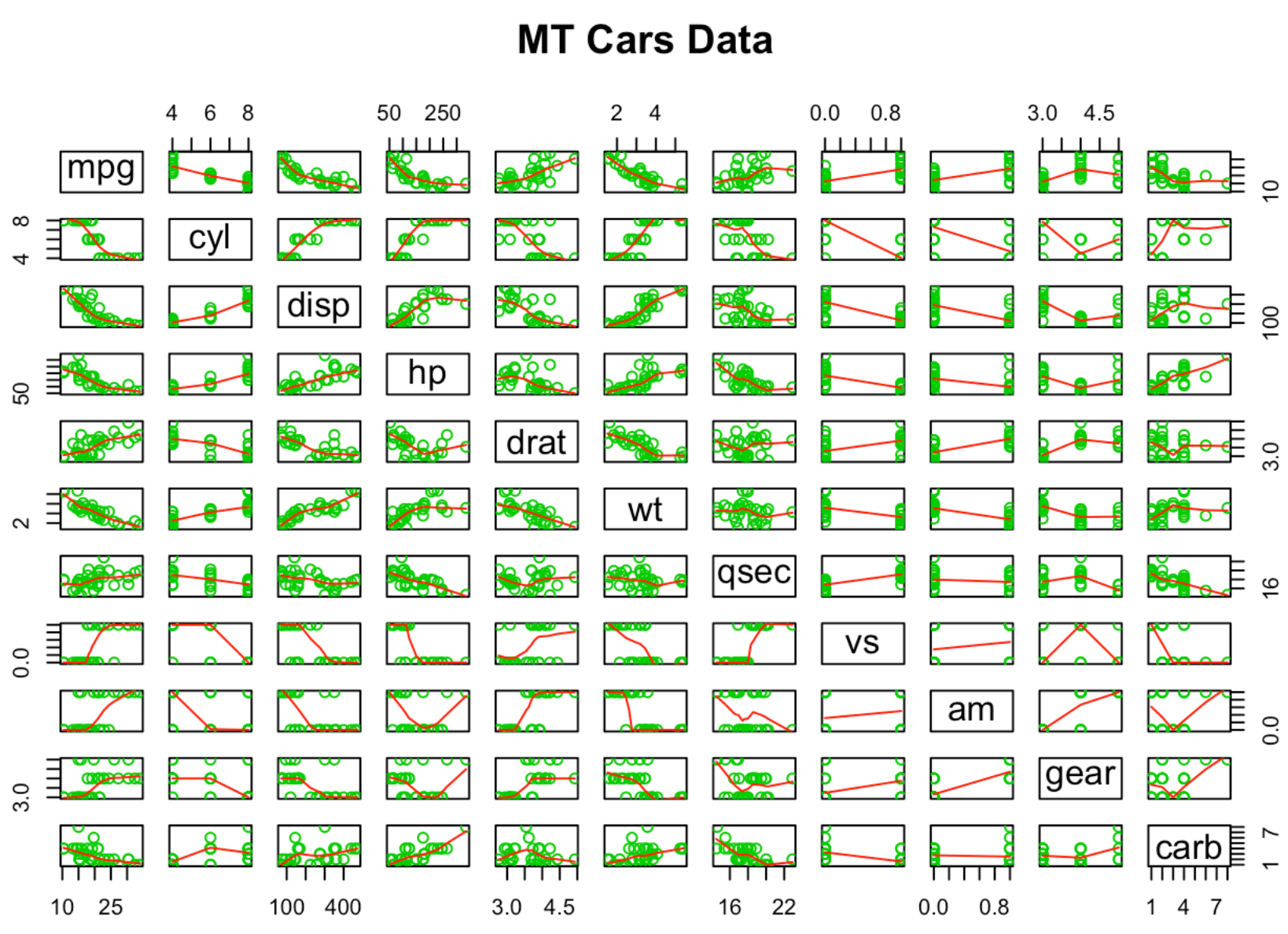


Figure 2: Correlation Factor Matrix

```
mtcars.cor <- cor(mtcars)
corrplot(mtcars.cor, method = "number", type = "upper", add = FALSE,
                  order = "original", is.corr = TRUE)
```

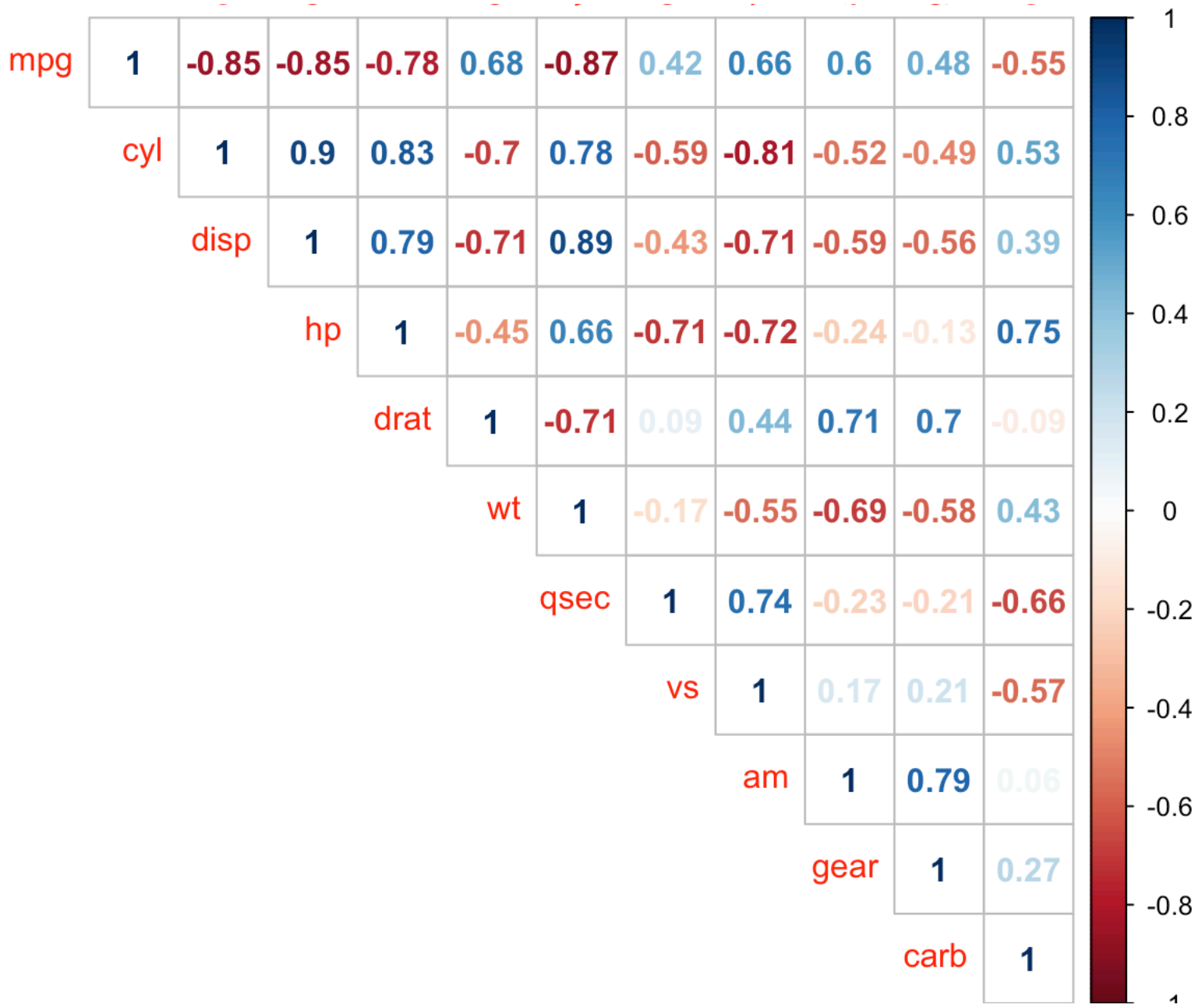
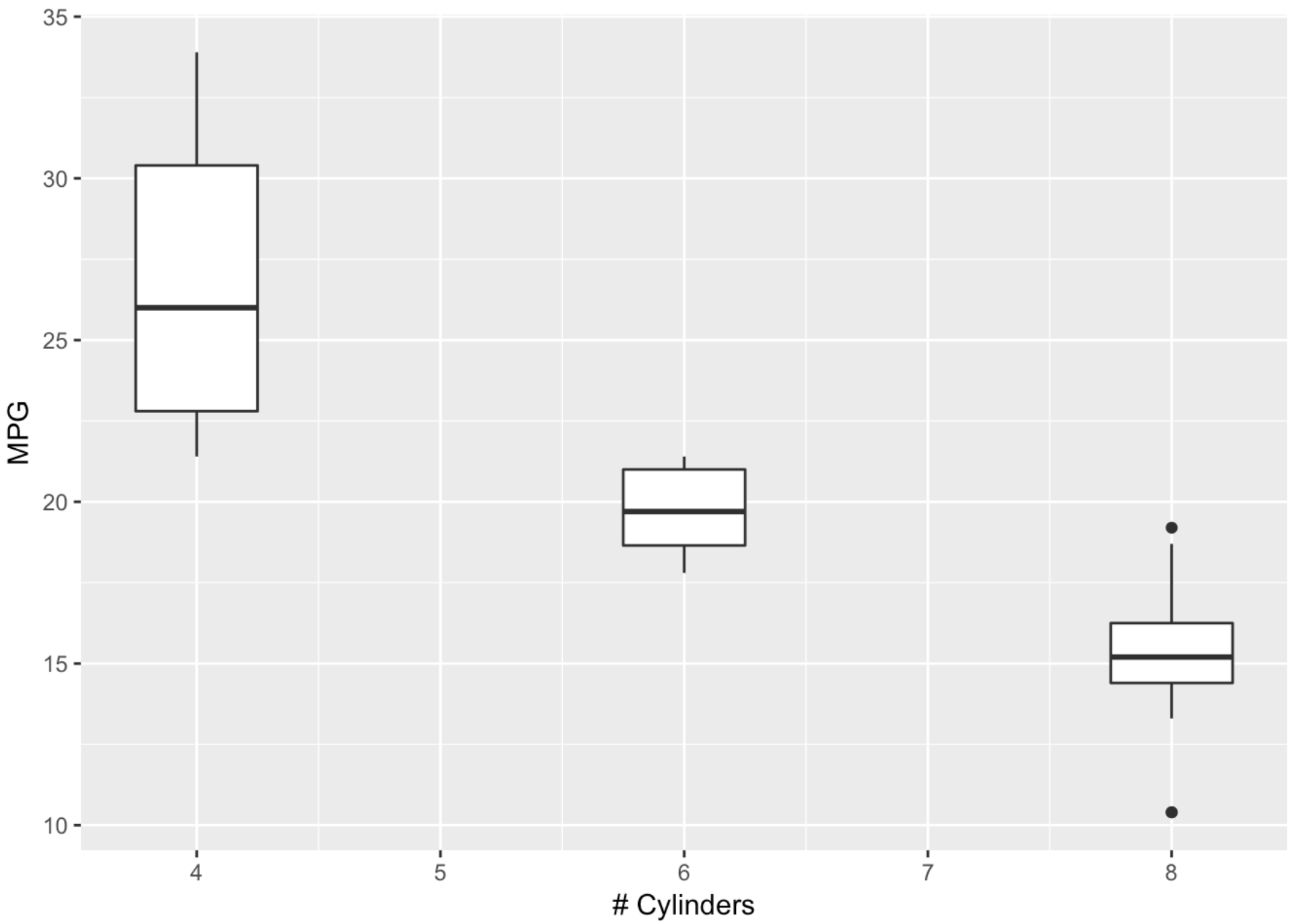


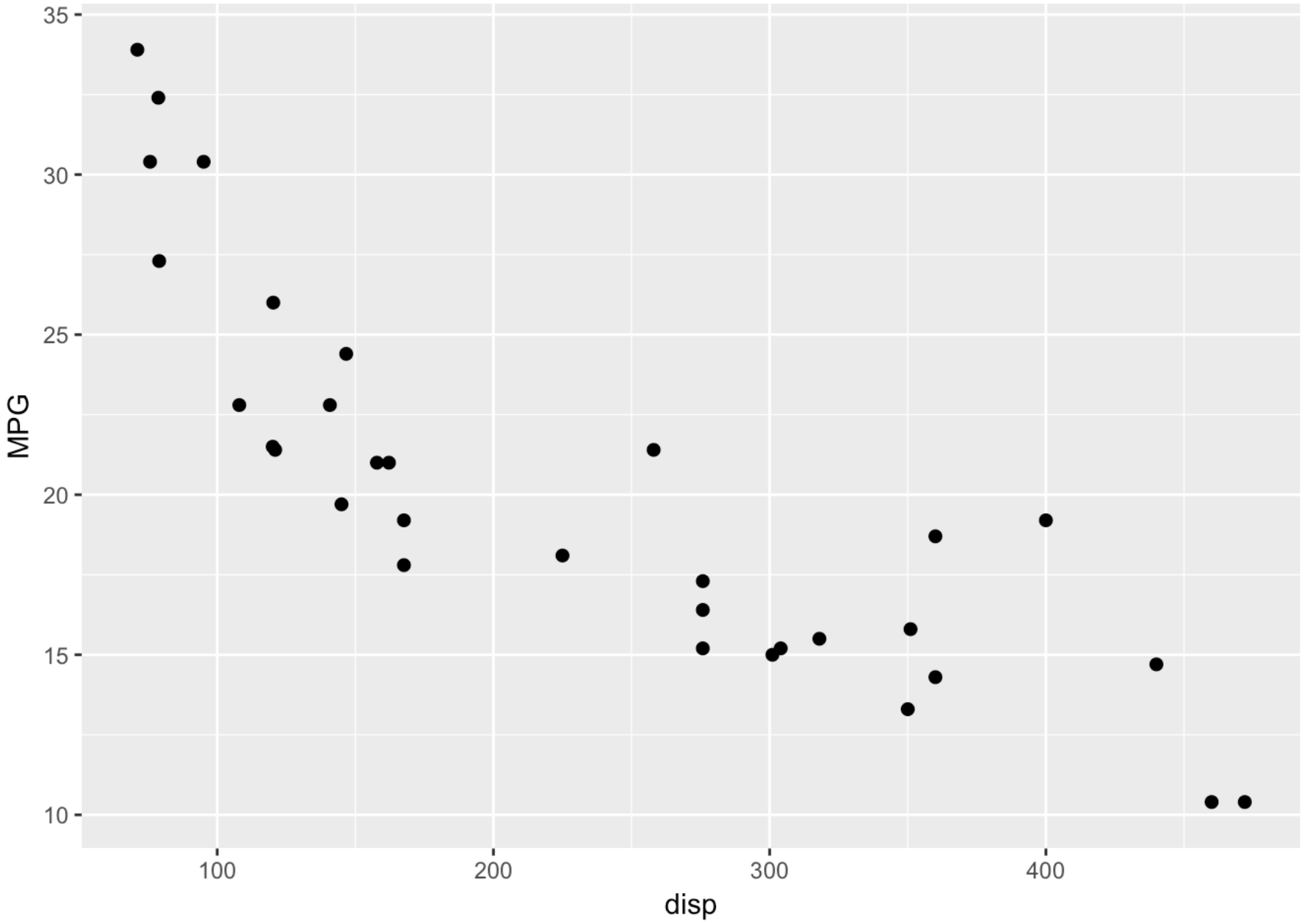
Figure 3: Variable Relationship to Miles per Gallon (MPG)

This panel plot illustrates the relationship of variables with a high correlation, absolute value > 0.50, with the mpg. Where it made sense boxplots are used, dotplots in other instances.

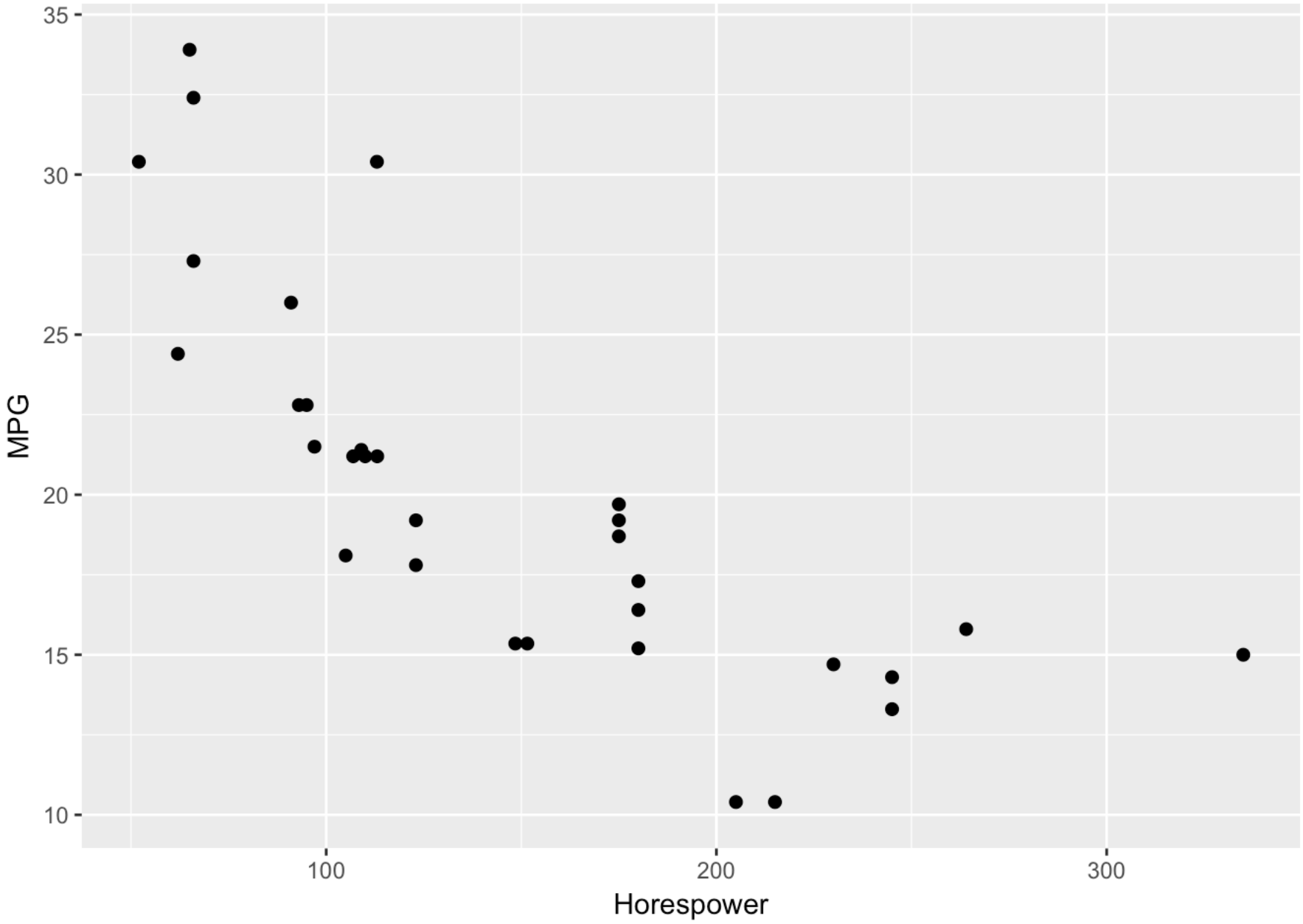
```
par(mfrow = c(2,4))
g1 <- ggplot(mtcars, aes(x = cyl, y = mpg, group = cyl)) +
            geom_boxplot(width = 0.5) +
            labs(x = "# Cylinders", y = "MPG")
g1
```



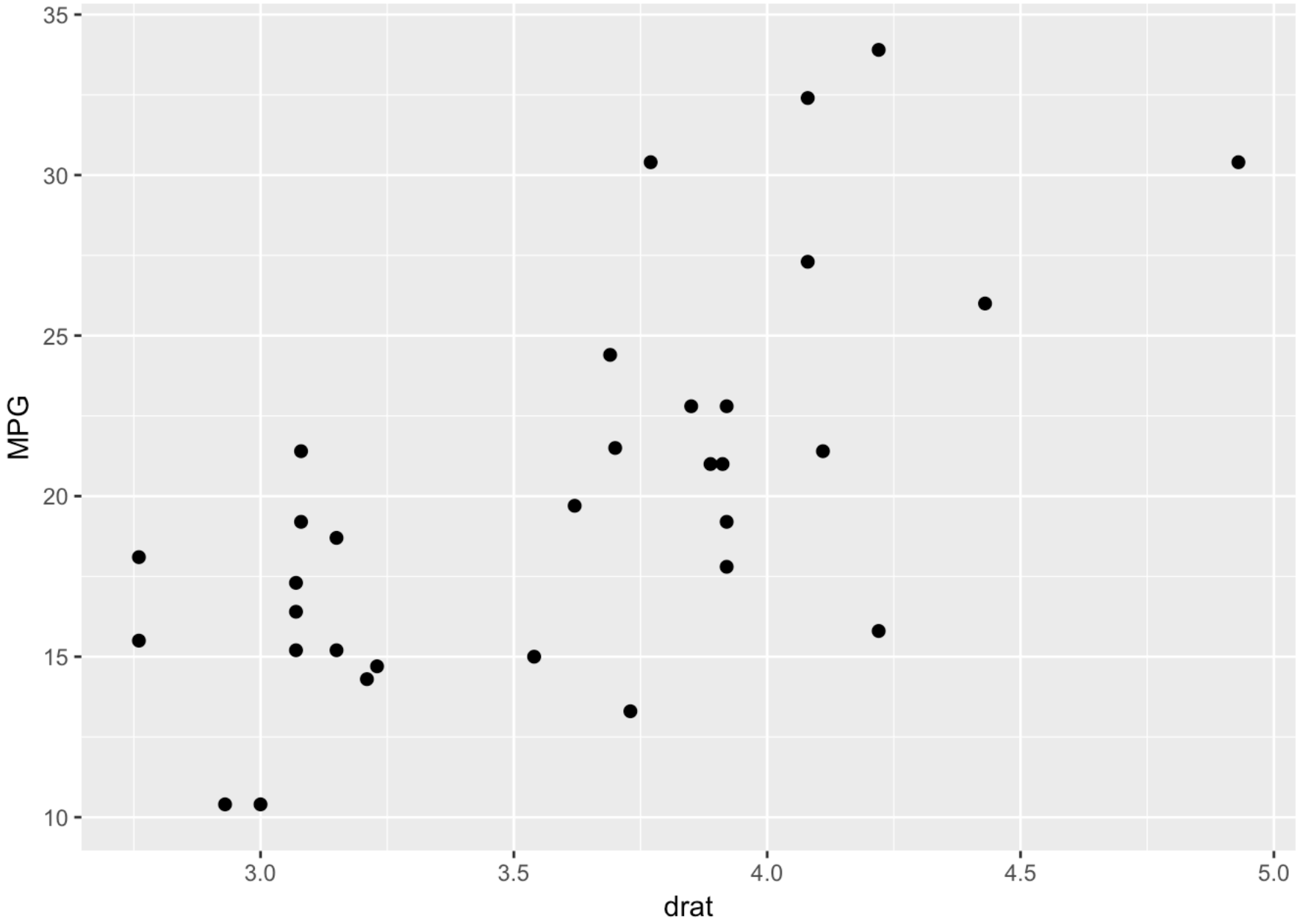

```
g2 <- ggplot(mtcars, aes(x = disp, y = mpg, group = disp)) +
            geom_dotplot(binaxis = "y", stackdir = "center", dotsize = 0.75, binwidth
= 0.5) +
            labs(x = "disp", y = "MPG")
g2
```



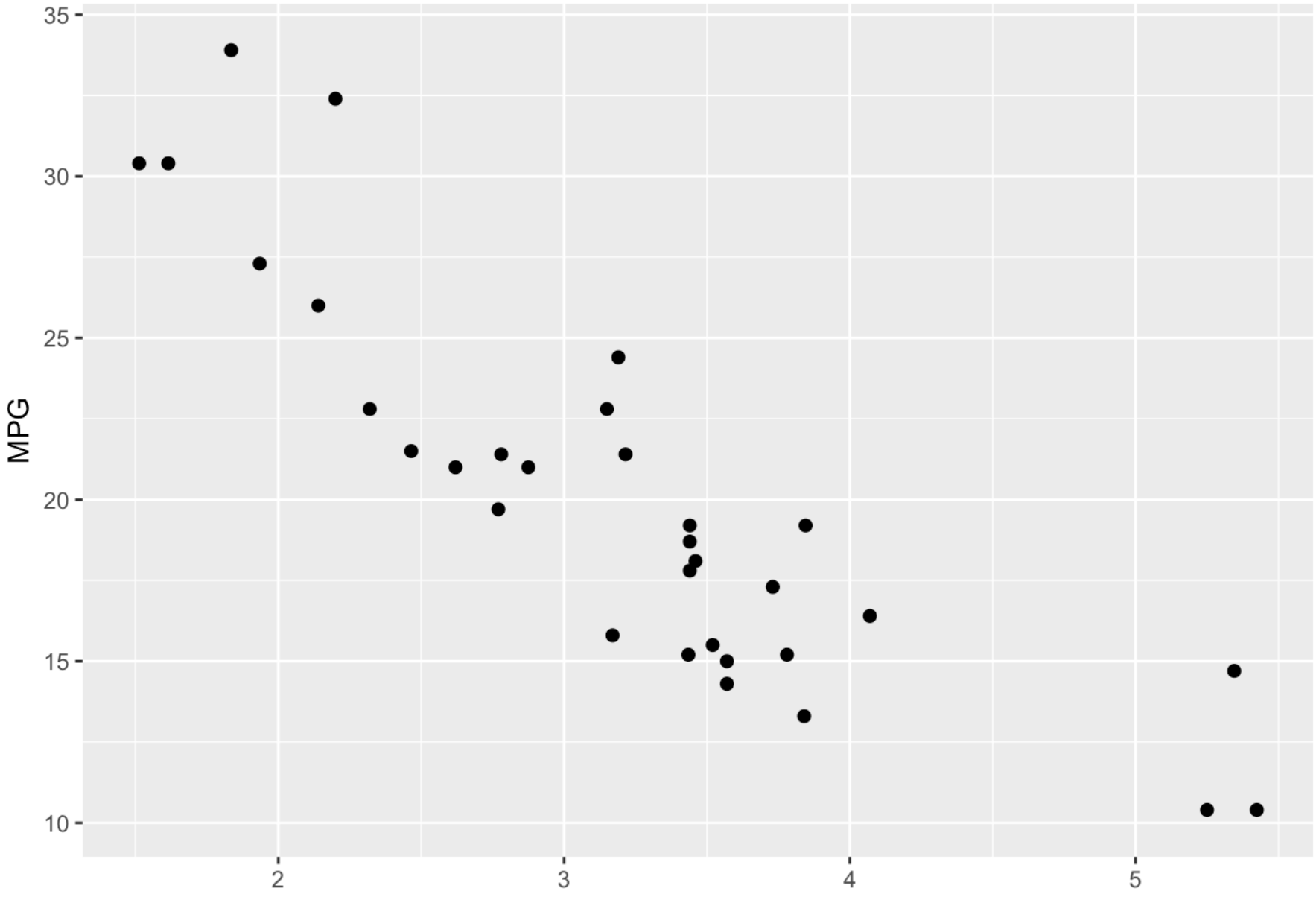

```
g3 <- ggplot(mtcars, aes(x = hp, y = mpg, group = hp)) +
            geom_dotplot(binaxis = "y", stackdir = "center", dotsize = 0.75, binwidth
= 0.5) +
            labs(x = "Horespower", y = "MPG")
g3
```



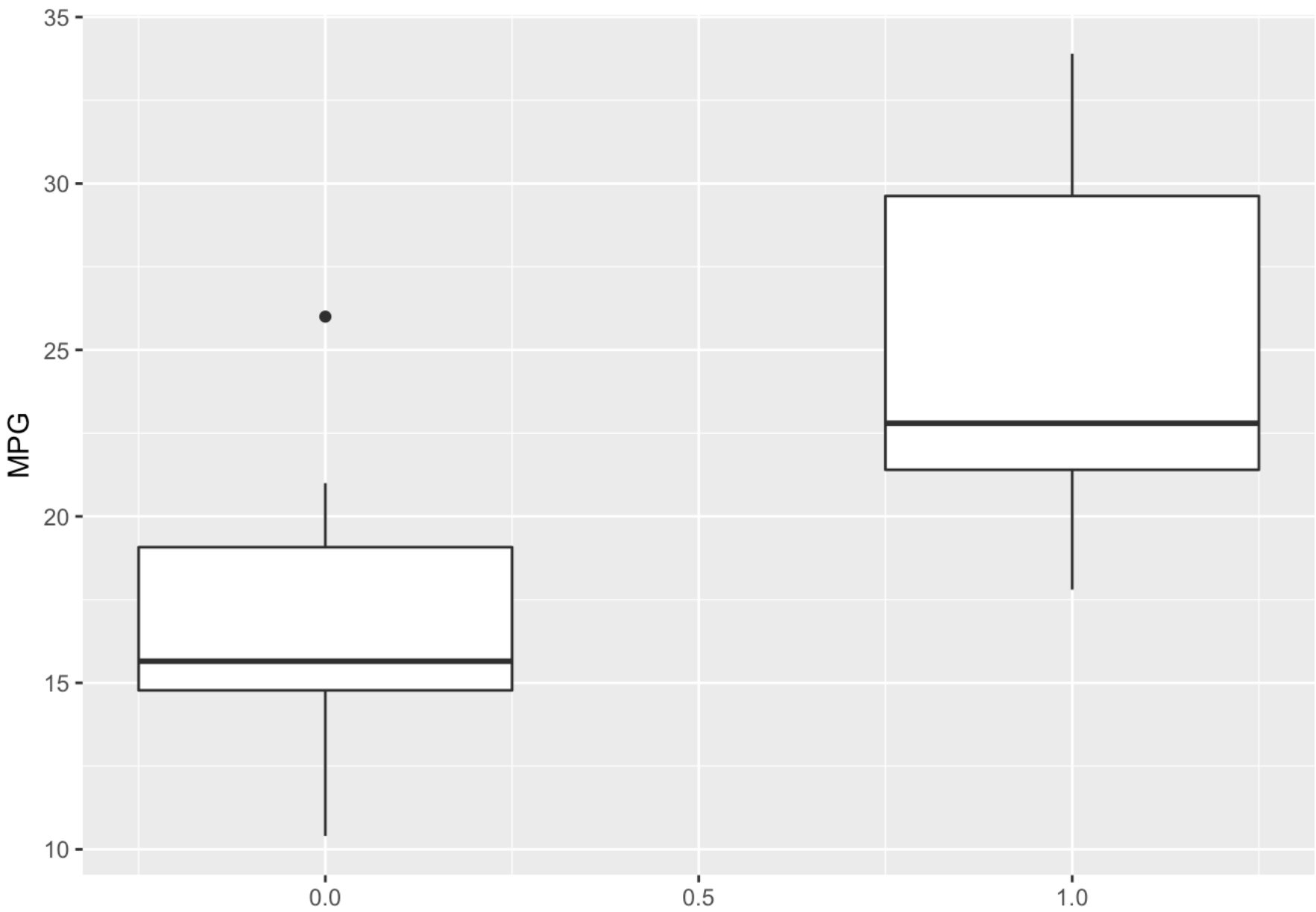

```
g4 <- ggplot(mtcars, aes(x = drat, y = mpg, group = drat)) +
            geom_dotplot(binaxis = "y", stackdir = "center", dotsize = 0.75, binwidth
= 0.5) +
            labs(x = "drat", y = "MPG")
g4
```




```
g5 <- ggplot(mtcars, aes(x = wt, y = mpg, group = wt)) +
            geom_dotplot(binaxis = "y", stackdir = "center", dotsize = 0.75, binwidth
= 0.5) +
            labs(x = "Weight", y = "MPG")
g5
```

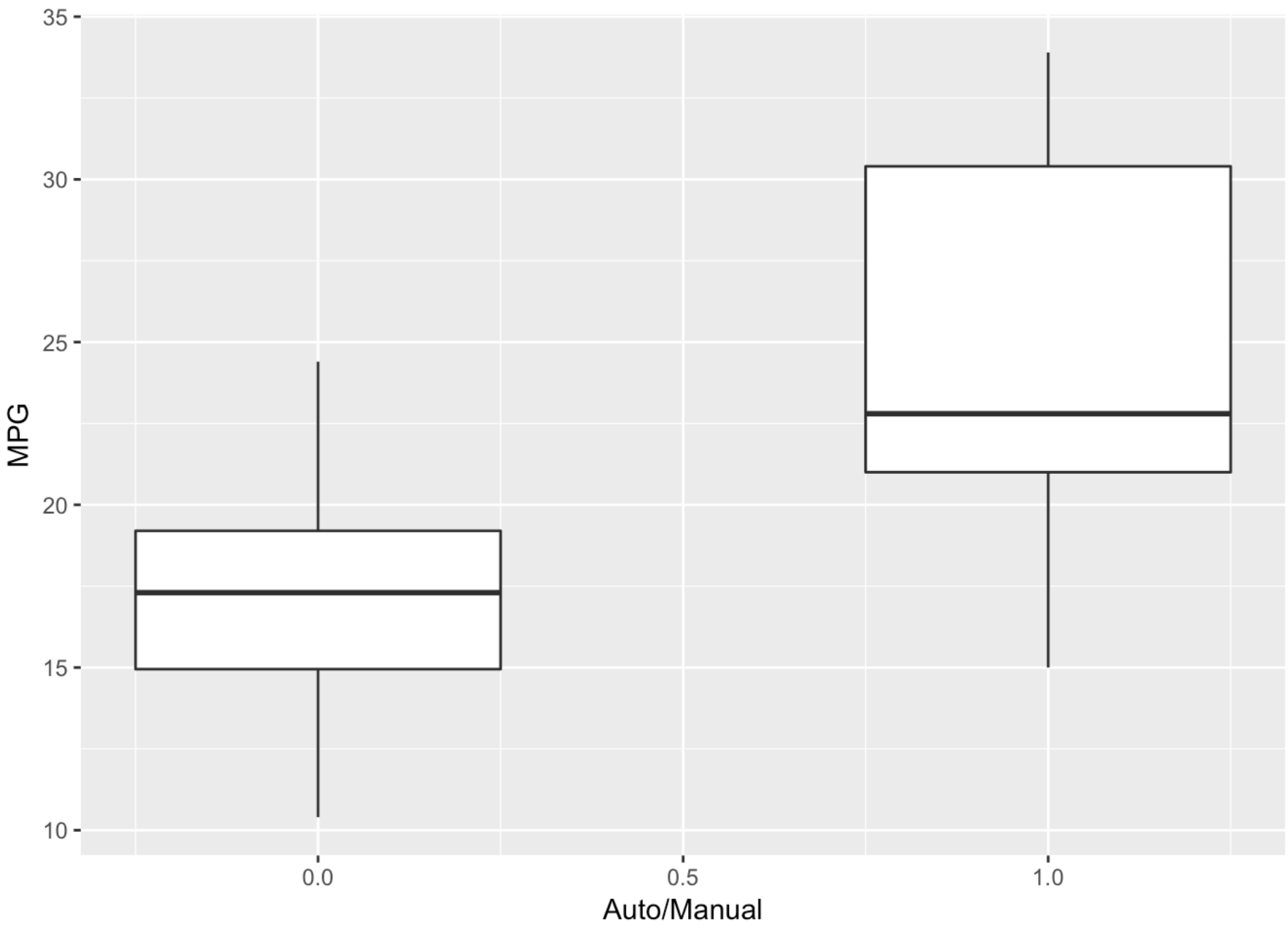



```
g6 <- ggplot(mtcars, aes(x = vs, y = mpg, group = vs)) +
            geom_boxplot(width = 0.5) +
            labs(x = "vs", y = "MPG")
g6
```

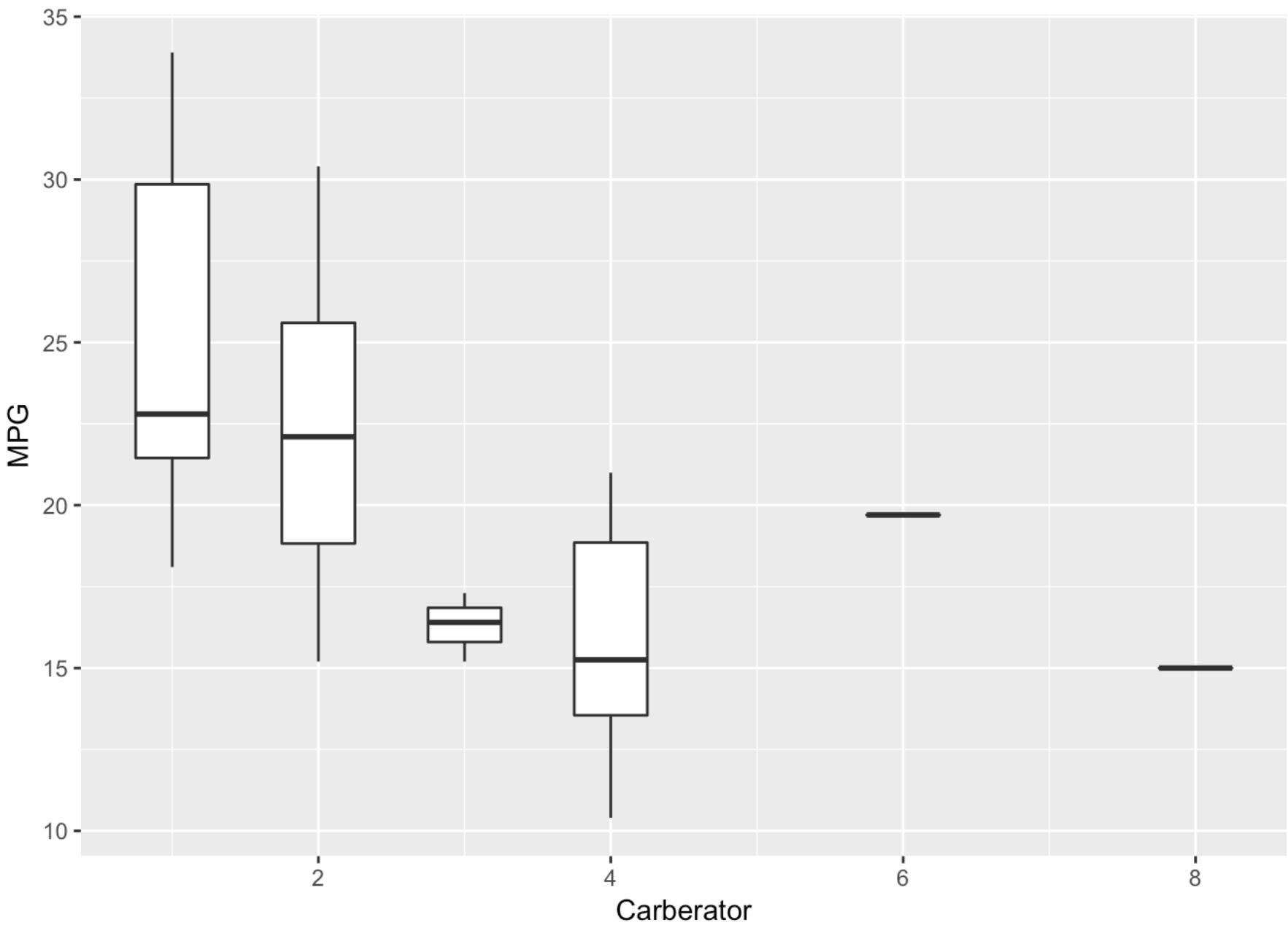


vs

```
g7 <- ggplot(mtcars, aes(x = am, y = mpg, group = am)) +
            geom_boxplot(width = 0.5) +
            labs(x = "Auto/Manual", y = "MPG")
g7
```




```
g8 <- ggplot(mtcars, aes(x = carb, y = mpg, group = carb)) +
            geom_boxplot(width = 0.5) +
            labs(x = "Carberator", y = "MPG")
g8
```


35
30
25
20
15
10
MPG
2
4
6
8
Carberator