EDA with ggplot2 on mtcars Dataset

Ken Wood

2024-11-24

Objective:

This assignment aims to guide you through exploratory data analysis (EDA) using the ggplot2 package in R, focusing on the mtcars dataset. By completing this assignment, you will enhance your proficiency in visualization.

Data Overview:

The mtcars dataset comprises various automobile characteristics such as miles per gallon (mpg), number of cylinders (cyl), horsepower (hp), and other performance metrics.

Instructions:

In your own R script file, please complete the following tasks:

1. Select the mtcars dataset for analysis.
2. Perform EDA to comprehend the dataset's structure and characteristics thoroughly.
3. Identify continuous and discrete variables within the mtcars dataset.
4. Create insightful visualizations using ggplot2 to uncover patterns and relationships within the data.

```r
# Load the datasets package (usually not necessary as it's loaded by default)
library(datasets)
library(ggplot2)

# Import the CO2 dataset
data(mtcars)

# Display the first few rows of the dataset
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
```

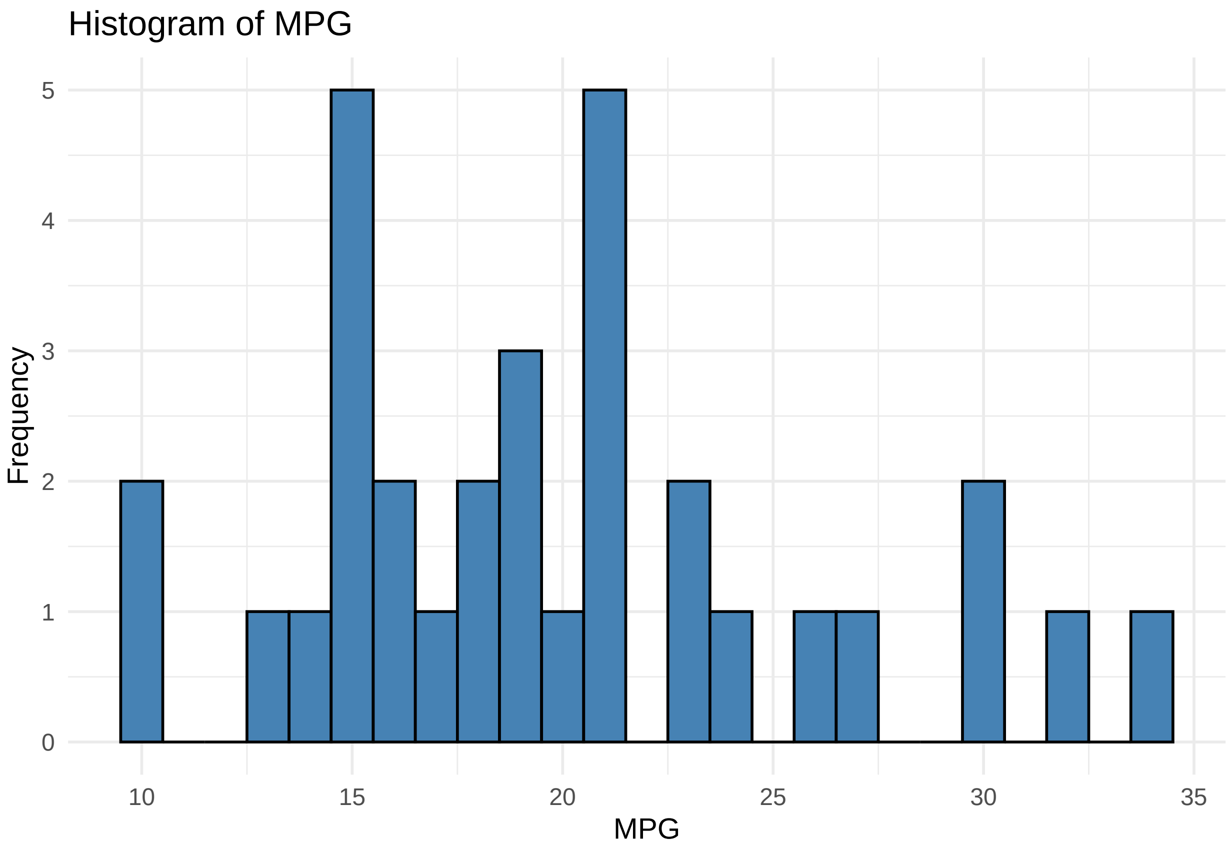
```r
# Get a set of summary stats for the dataset
summary(mtcars)
```

```
##       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      
##  Min.   :0.0000   Min.   :3.000   Min.   :1.000  
##  1st Qu.:0.0000   1st Qu.:3.000   1st Qu.:2.000  
##  Median :0.0000   Median :4.000   Median :2.000  
##  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
```

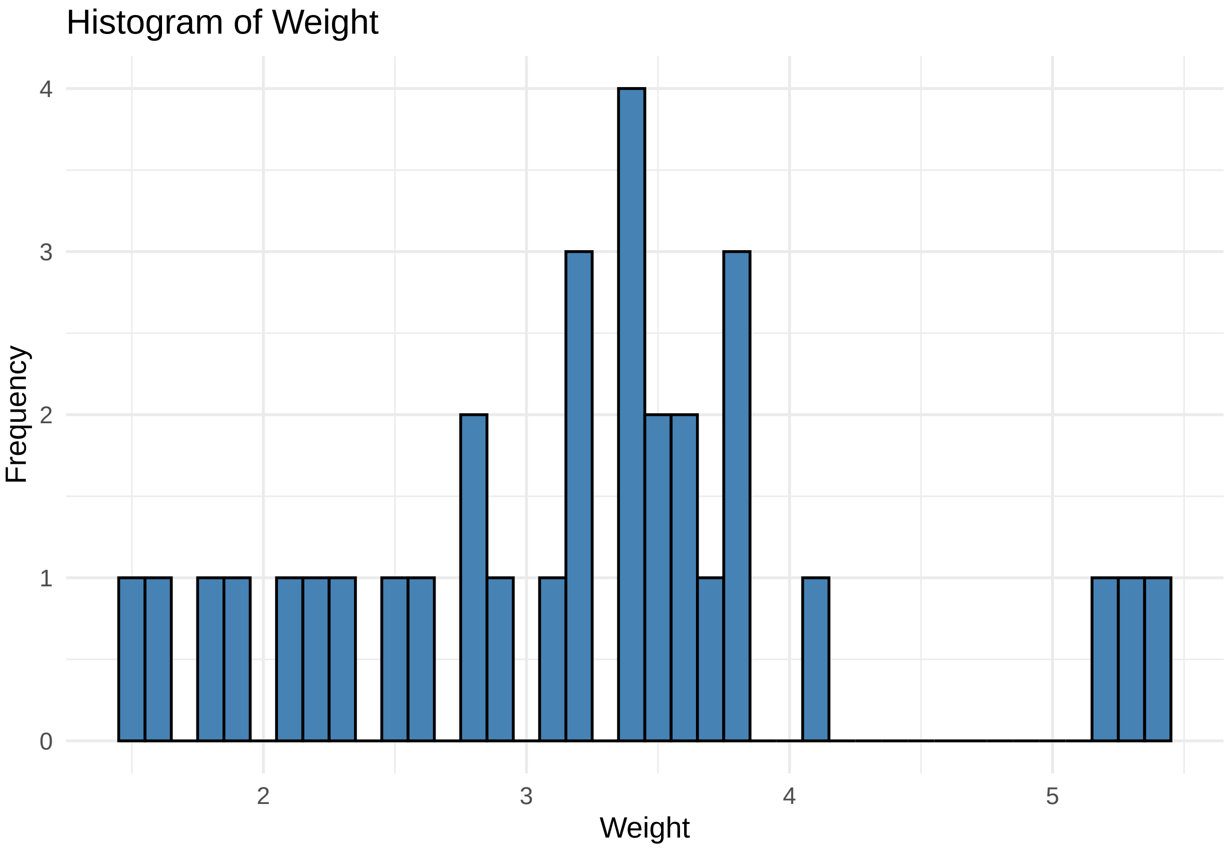
Data Visualization:

1. Histograms or density plots to visualize the distribution of continuous variables (mpg, hp, etc.).

```
# Create a histogram
ggplot(mtcars, aes(x=mpg)) +
  geom_histogram(color = "black", fill = "steelblue", binwidth = 1) +
  labs(x = "MPG", y = "Frequency") +
  ggtitle("Histogram of MPG") +
  theme_minimal()
```



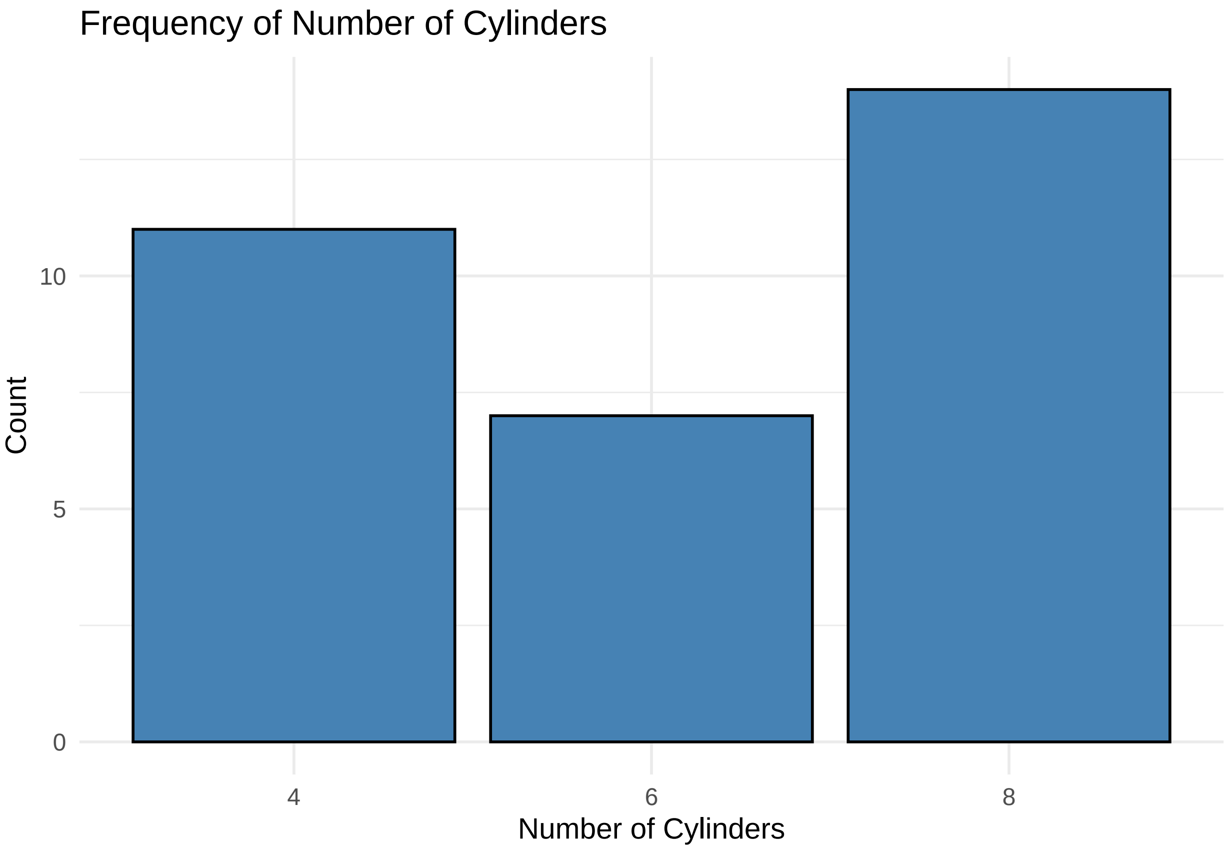
```
# Create a histogram
ggplot(mtcars, aes(x=wt)) +
  geom_histogram(color = "black", fill = "steelblue", binwidth = 0.1) +
  labs(x = "Weight", y = "Frequency") +
  ggtitle("Histogram of Weight") +
  theme_minimal()
```



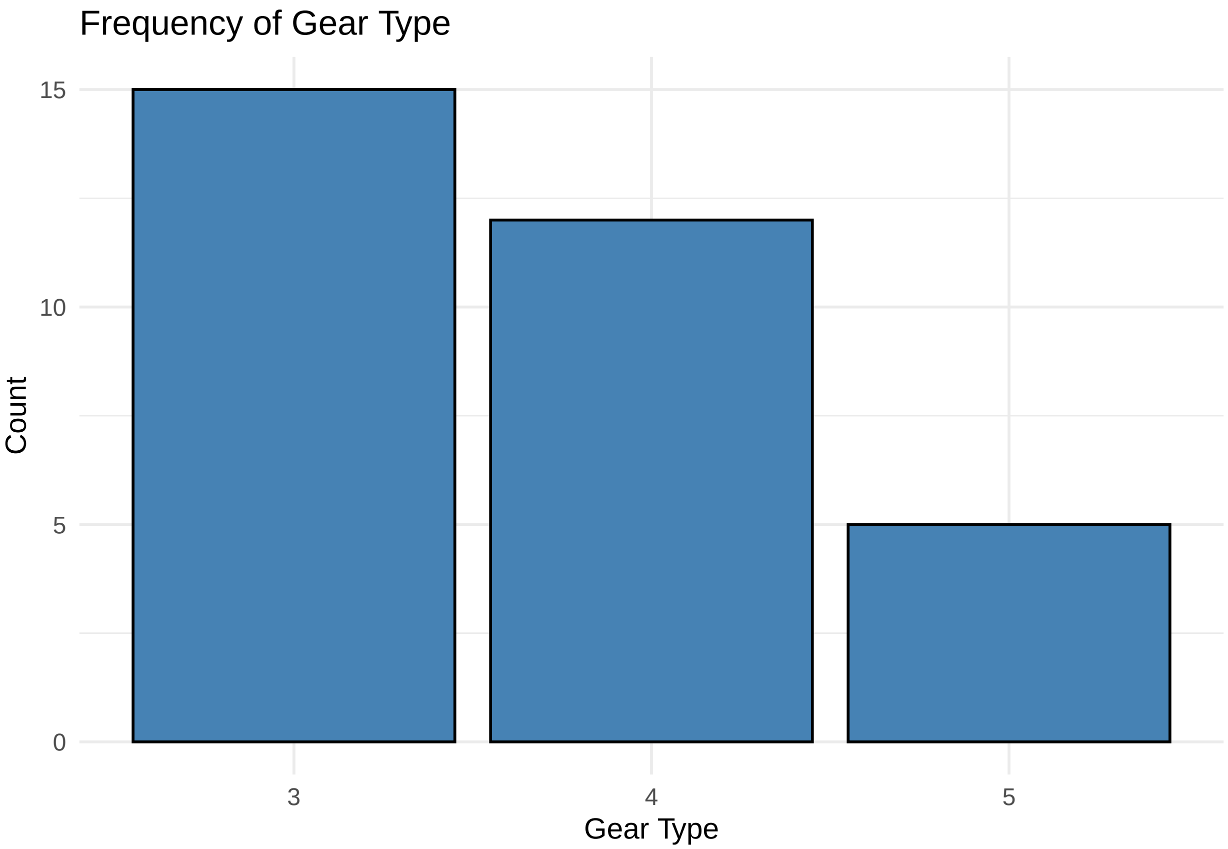
2.

Bar plots to display the frequency of discrete variables (number of cylinders, gear type).

```
ggplot(mtcars, aes(x = factor(cyl))) +
  geom_bar(color = "black", fill = "steelblue") +
  labs(title = "Frequency of Number of Cylinders", x = "Number of Cylinders", y = "Count") + theme_mini
```

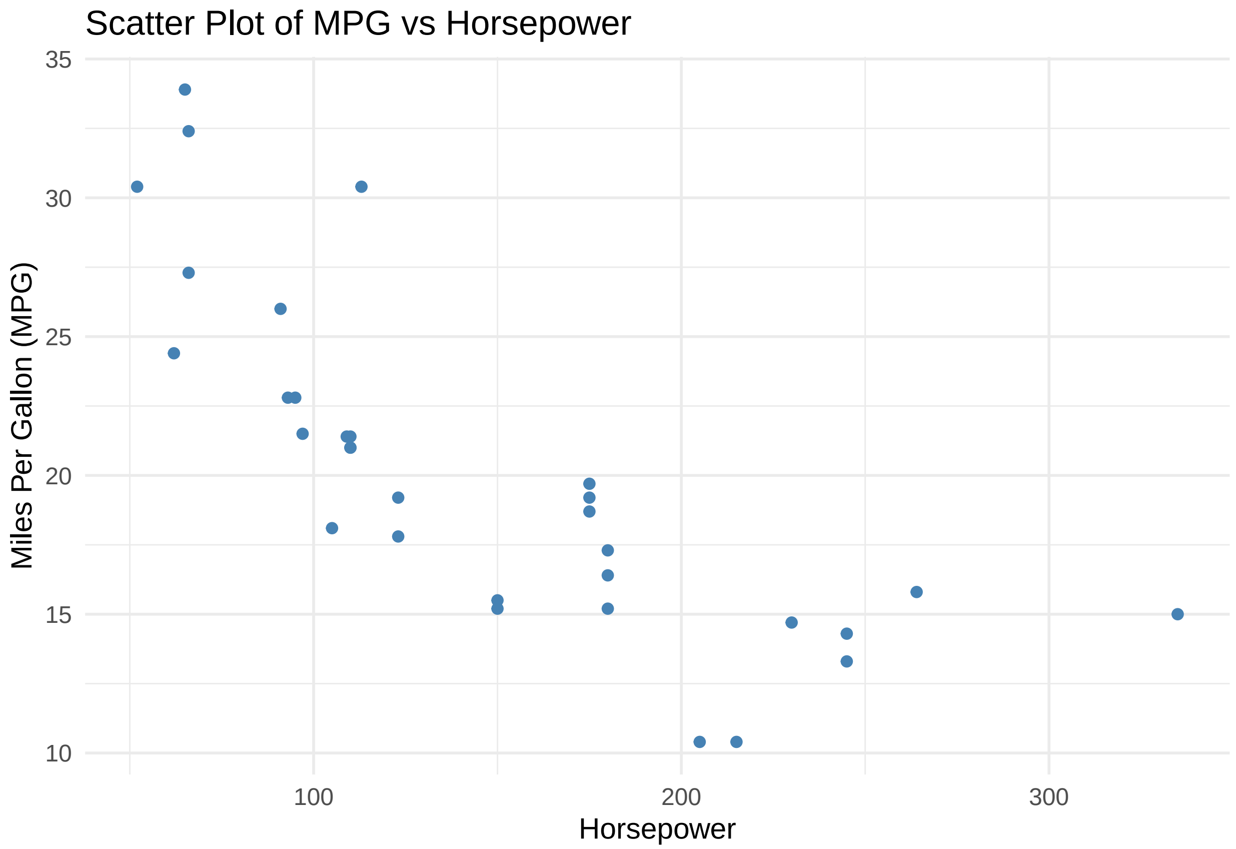


```
ggplot(mtcars, aes(x = factor(gear))) +
  geom_bar(color = "black", fill = "steelblue") +
  labs(title = "Frequency of Gear Type", x = "Gear Type", y = "Count") + theme_minimal()
```



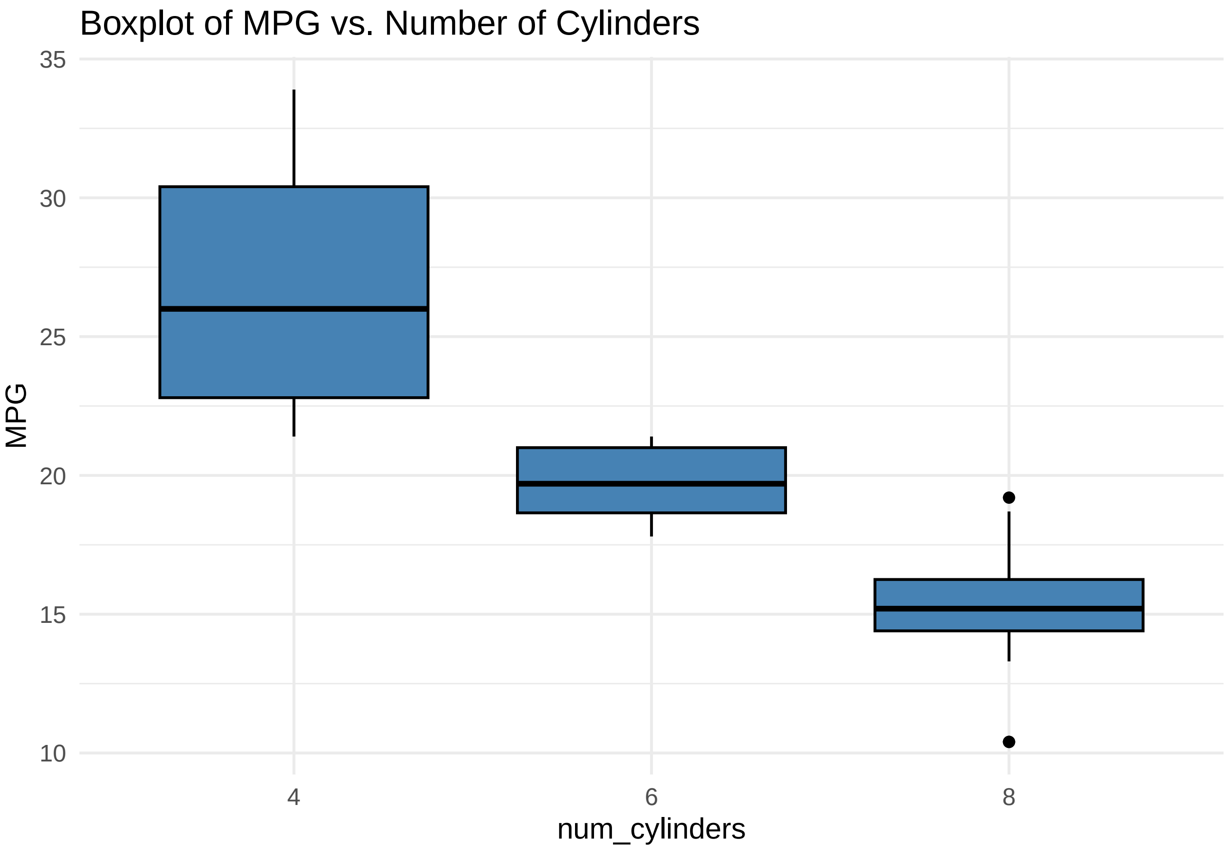
3. Scatter plots to explore relationships between two continuous variables.

```
ggplot(mtcars, aes(x = hp, y = mpg)) +
  geom_point(color = "steelblue") +
  labs(title = "Scatter Plot of MPG vs Horsepower", x = "Horsepower", y = "Miles Per Gallon (MPG)") + th
```

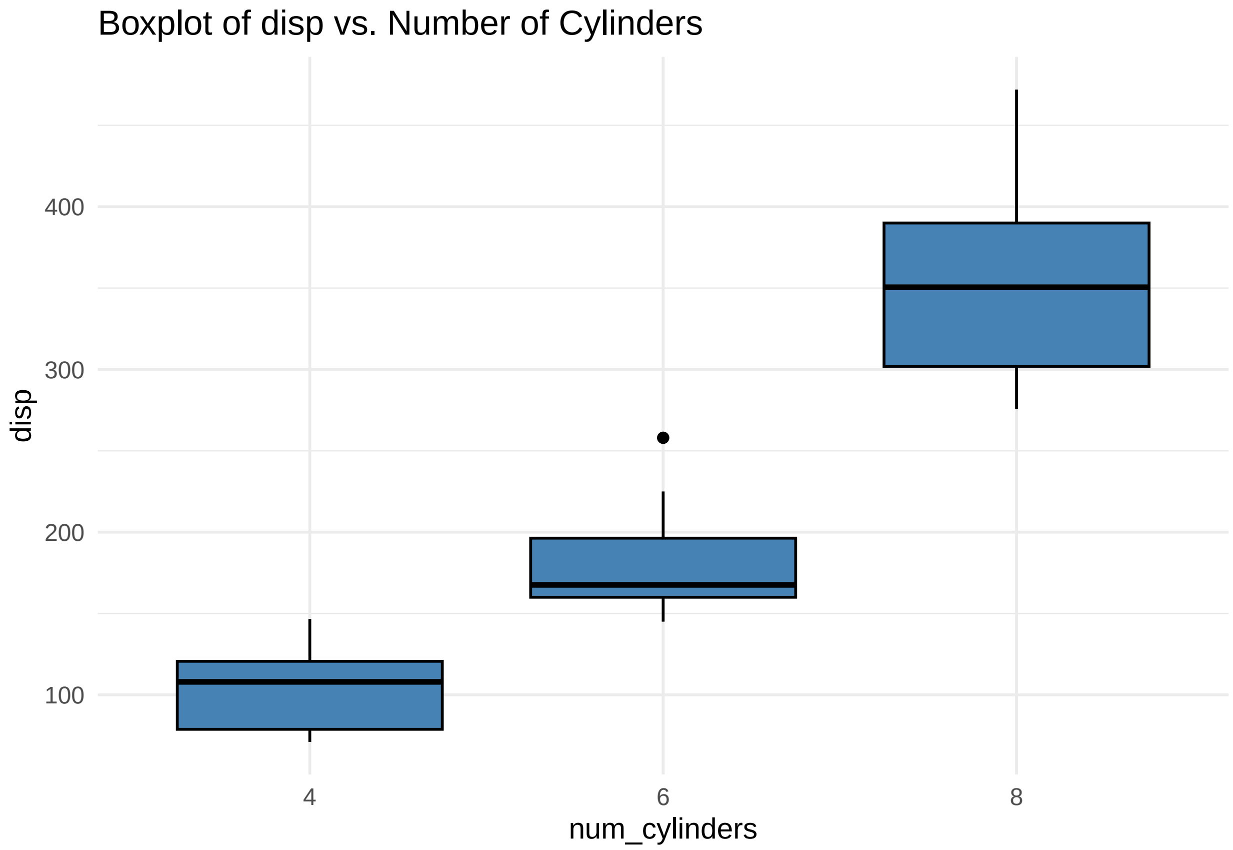


4. Box plots or violin plots to compare the distribution of a continuous variable across different levels of a categorical variable (e.g., cylinders).

```
# Create a boxplot
ggplot(data=mtcars, aes(x=factor(cyl), y=mpg)) +
  geom_boxplot(color = "black", fill = "steelblue")  +
  labs(x = "num_cylinders", y = "MPG") +
  ggtitle("Boxplot of MPG vs. Number of Cylinders") +
  theme_minimal()
```



```
# Create a boxplot
ggplot(data=mtcars,aes(x=factor(cyl), y=disp)) +
  geom_boxplot(color = "black", fill = "steelblue") +
  labs(x = "num_cylinders", y = "disp") +
  ggtitle("Boxplot of disp vs. Number of Cylinders") +
  theme_minimal()
```


Boxplot of disp vs. Number of Cylinders
disp
400
300
200
100
4
6
8
num_cylinders