3 Exploratory Data Analysis (EDA)

Mohan Khanal

2025-01-08

Table of contents

1 Chapter 3: Exploratory Data Analysis (EDA)

1.1 Agenda

- **Part 1 (30 mins)** – Descriptive Statistics
- **Part 2 (50 mins)** – Data Visualization with ggplot2
- **Part 3 (30 mins)** – Creating Basic Plots (Scatter, Bar, Histogram, Boxplot)
- **Part 4 (10 mins)** – Practice and Q&A

1.2 Part 1: Descriptive Statistics (30 mins)

1.2.1 Loading Built-in Datasets

R comes with several built-in datasets perfect for practice.

```
# Load iris dataset (flower measurements)
data(iris)
head(iris)
```

```
  Sepal.Length Sepal.Width Petal.Length Petal.Width Species
1          5.1         3.5          1.4         0.2  setosa
2          4.9         3.0          1.4         0.2  setosa
3          4.7         3.2          1.3         0.2  setosa
4          4.6         3.1          1.5         0.2  setosa
5          5.0         3.6          1.4         0.2  setosa
6          5.4         3.9          1.7         0.4  setosa
```

```
# Load mtcars dataset (car specifications)
data(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
```

```r
# See all available datasets
# data()
```

1.2.2 Summary Statistics

```r
# Quick overview
summary(iris)
```

```
  Sepal.Length    Sepal.Width     Petal.Length    Petal.Width
 Min.   :4.300   Min.   :2.000   Min.   :1.000   Min.   :0.100
 1st Qu.:5.100   1st Qu.:2.800   1st Qu.:1.600   1st Qu.:0.300
 Median :5.800   Median :3.000   Median :4.350   Median :1.300
 Mean   :5.843   Mean   :3.057   Mean   :3.758   Mean   :1.199
 3rd Qu.:6.400   3rd Qu.:3.300   3rd Qu.:5.100   3rd Qu.:1.800
 Max.   :7.900   Max.   :4.400   Max.   :6.900   Max.   :2.500
       Species
 setosa    :50
 versicolor:50
 virginica :50
```

```r
# Summary for mtcars
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
```

1.2.3 Measures of Central Tendency

```
# Mean
mean(iris$Sepal.Length)
```

```
[1] 5.843333
```

```
mean(mtcars$mpg)
```

```
[1] 20.09062
```

```
# Median
median(iris$Sepal.Length)
```

```
[1] 5.8
```

```
median(mtcars$mpg)
```

```
[1] 19.2
```

```
# Mode (R doesn't have built-in mode function)
# We can find it using table
table(iris$Species)
```

```
    setosa versicolor  virginica
        50         50         50
```

1.2.4 Measures of Dispersion

```
# Range
range(iris$Sepal.Length)
```

```
[1] 4.3 7.9
```

```
range(mtcars$mpg)
```

```
[1] 10.4 33.9
```

```
# Variance
var(iris$Sepal.Length)
```

```
[1] 0.6856935
```

```
var(mtcars$mpg)
```

```
[1] 36.3241
```

```
# Standard Deviation
sd(iris$Sepal.Length)
```

```
[1] 0.8280661
```

```
sd(mtcars$mpg)
```

```
[1] 6.026948
```

```
# Interquartile Range
IQR(iris$Sepal.Length)
```

```
[1] 1.3
```

```r
IQR(mtcars$mpg)
```

```
[1] 7.375
```

1.2.5 Quantiles and Percentiles

```r
# Quartiles
quantile(iris$Sepal.Length)
```

```
  0%  25%  50%  75% 100%
 4.3  5.1  5.8  6.4  7.9
```

```r
# Specific percentiles
quantile(mtcars$mpg, probs = c(0.25, 0.50, 0.75, 0.90, 0.95))
```

```
   25%    50%    75%    90%    95%
15.425 19.200 22.800 30.090 31.300
```

1.2.6 Grouped Statistics with dplyr

```r
library(dplyr)

# Summary by species
iris %>%
  group_by(Species) %>%
  summarize(
    mean_sepal = mean(Sepal.Length),
    sd_sepal = sd(Sepal.Length),
    min_sepal = min(Sepal.Length),
    max_sepal = max(Sepal.Length),
    n = n()
  )
```

```
# A tibble: 3 x 6
  Species    mean_sepal sd_sepal min_sepal max_sepal     n
  <fct>           <dbl>    <dbl>     <dbl>     <dbl> <int>
```

```
1 setosa           5.01    0.352       4.3       5.8    50
2 versicolor       5.94    0.516       4.9       7      50
3 virginica        6.59    0.636       4.9       7.9    50
```

```r
# Summary by number of cylinders
mtcars %>%
  group_by(cyl) %>%
  summarize(
    mean_mpg = mean(mpg),
    sd_mpg = sd(mpg),
    mean_hp = mean(hp),
    count = n()
  )
```

```
# A tibble: 3 x 5
    cyl mean_mpg sd_mpg mean_hp count
  <dbl>    <dbl>  <dbl>   <dbl> <int>
1     4     26.7   4.51    82.6    11
2     6     19.7   1.45   122.      7
3     8     15.1   2.56   209.     14
```

1.2.7 Correlation Analysis

```r
# Correlation between two variables
cor(iris$Sepal.Length, iris$Sepal.Width)
```

```
[1] -0.1175698
```

```r
cor(mtcars$mpg, mtcars$wt)
```

```
[1] -0.8676594
```

```r
# Correlation matrix (numeric columns only)
cor(iris[, 1:4])
```

```
             Sepal.Length Sepal.Width Petal.Length Petal.Width
Sepal.Length    1.0000000  -0.1175698    0.8717538   0.8179411
Sepal.Width    -0.1175698   1.0000000   -0.4284401  -0.3661259
Petal.Length    0.8717538  -0.4284401    1.0000000   0.9628654
Petal.Width     0.8179411  -0.3661259    0.9628654   1.0000000
```

```r
cor(mtcars[, c("mpg", "hp", "wt", "qsec")])
```

```
            mpg         hp         wt       qsec
mpg   1.0000000 -0.7761684 -0.8676594  0.4186840
hp   -0.7761684  1.0000000  0.6587479 -0.7082234
wt   -0.8676594  0.6587479  1.0000000 -0.1747159
qsec  0.4186840 -0.7082234 -0.1747159  1.0000000
```

```r
# Round for better readability
round(cor(iris[, 1:4]), 2)
```

```
             Sepal.Length Sepal.Width Petal.Length Petal.Width
Sepal.Length         1.00       -0.12         0.87        0.82
Sepal.Width         -0.12        1.00        -0.43       -0.37
Petal.Length         0.87       -0.43         1.00        0.96
Petal.Width          0.82       -0.37         0.96        1.00
```

1.3 Part 2: Data Visualization with ggplot2 (50 mins)

1.3.1 Introduction to ggplot2

ggplot2 is the most popular visualization package in R, following the "grammar of graphics" philosophy.

```r
library(ggplot2)
```

1.3.2 Basic ggplot2 Syntax

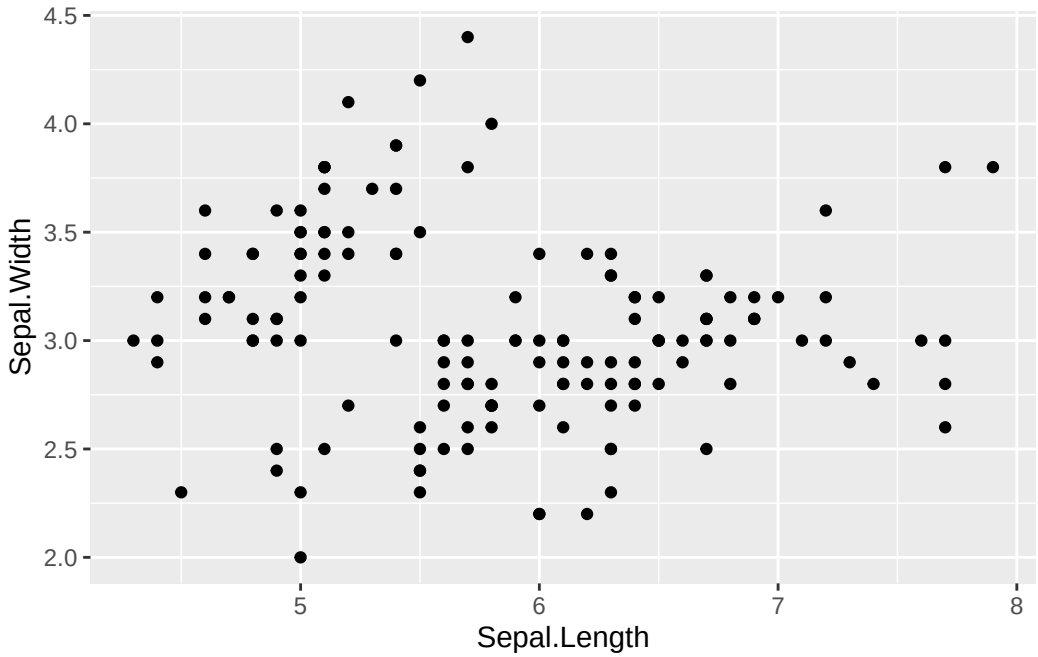
```
ggplot(data = <DATA>, aes(x = <X-VARIABLE>, y = <Y-VARIABLE>)) +
  geom_<TYPE>()
```

Components: - `ggplot()`: Initialize the plot - `aes()`: Define aesthetics (what goes on x, y, color, etc.) - `geom_*()`: Add layers (points, lines, bars, etc.)

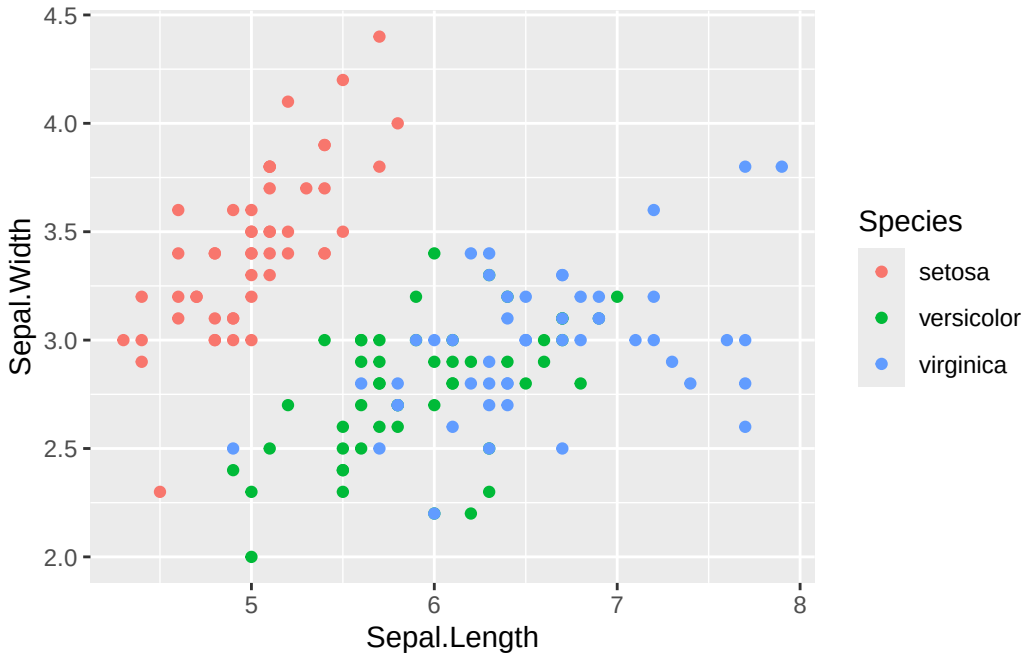
1.4 Part 3: Creating Basic Plots (30 mins)

1.4.1 1. Scatter Plots (Relationship between two numeric variables)

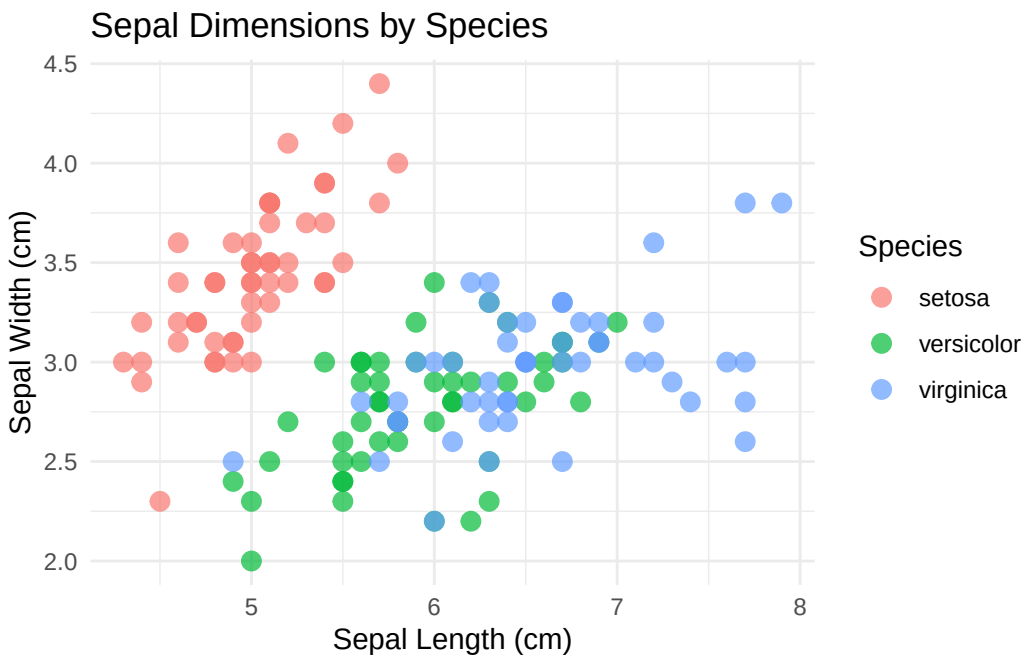
```
# Basic scatter plot
ggplot(iris, aes(x = Sepal.Length, y = Sepal.Width)) +
  geom_point()
```



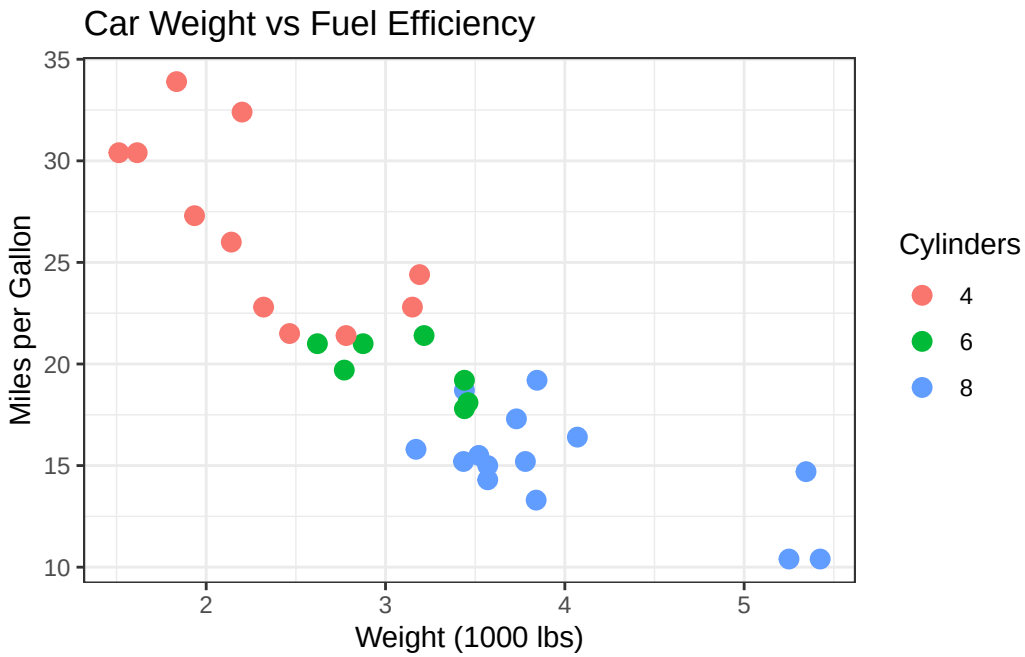

```
# Add color by species
ggplot(iris, aes(x = Sepal.Length, y = Sepal.Width, color = Species)) +
  geom_point()
```




```
# Add labels and title
ggplot(iris, aes(x = Sepal.Length, y = Sepal.Width, color = Species)) +
  geom_point(size = 3, alpha = 0.7) +
  labs(
    title = "Sepal Dimensions by Species",
    x = "Sepal Length (cm)",
    y = "Sepal Width (cm)"
  ) +
  theme_minimal()
```

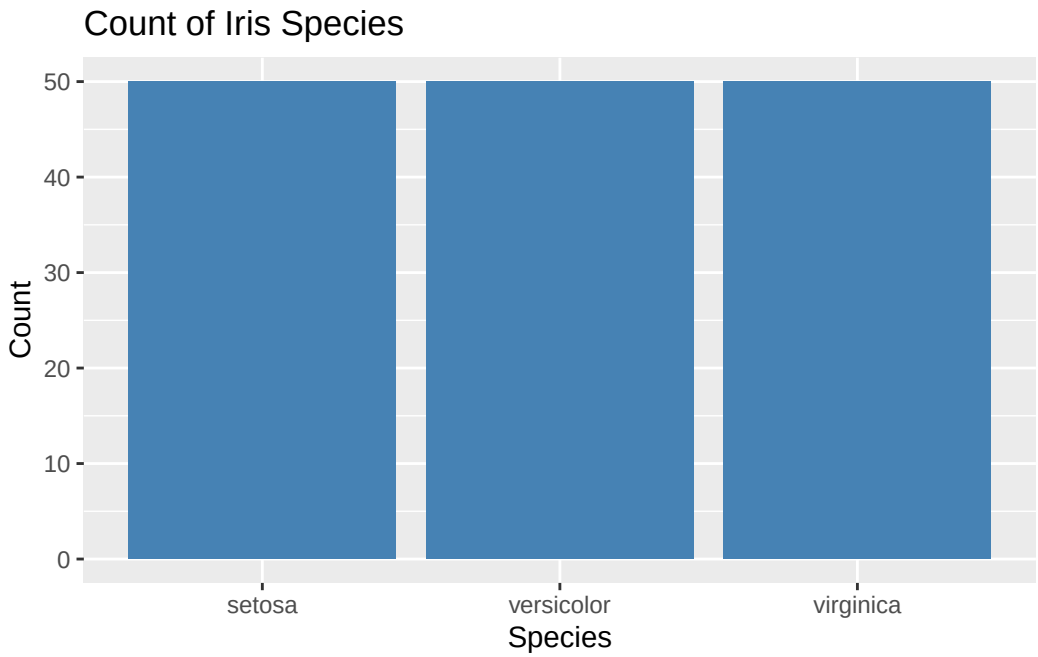



```r
# Scatter plot with mtcars
ggplot(mtcars, aes(x = wt, y = mpg)) +
  geom_point(aes(color = factor(cyl)), size = 3) +
  labs(
    title = "Car Weight vs Fuel Efficiency",
    x = "Weight (1000 lbs)",
    y = "Miles per Gallon",
    color = "Cylinders"
  ) +
  theme_bw()
```

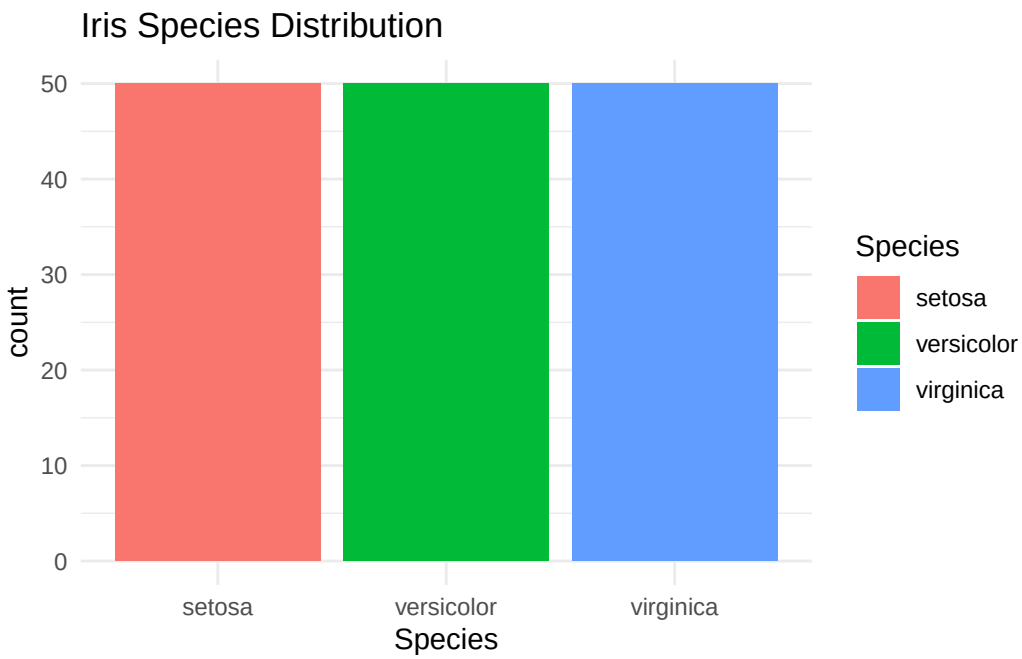


1.4.2 2. Bar Plots (Categorical data)

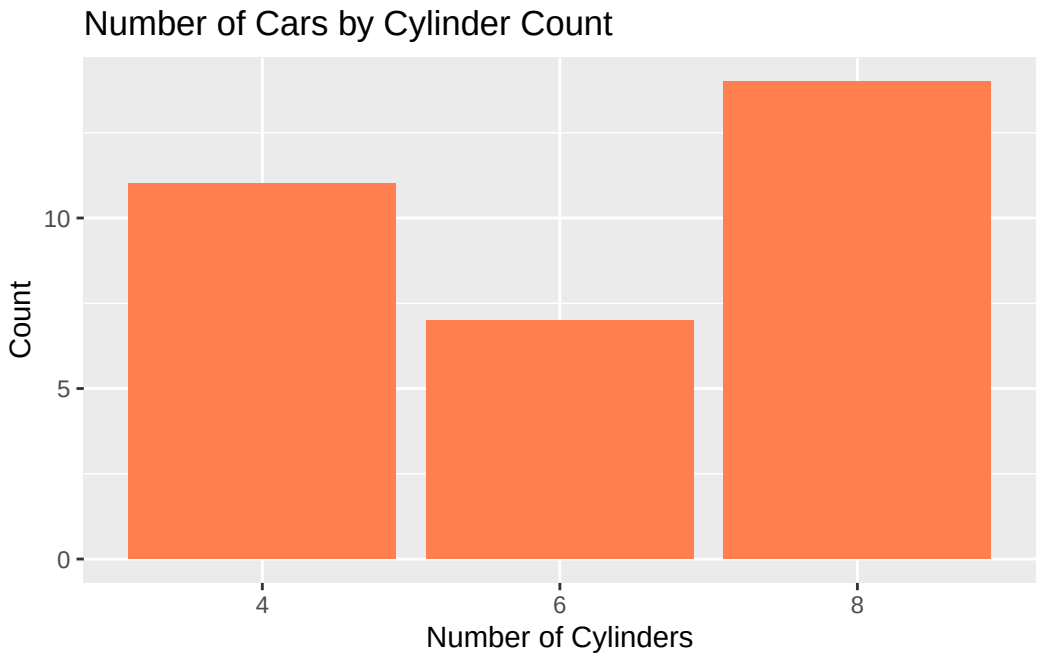
```
# Count of each species
ggplot(iris, aes(x = Species)) +
  geom_bar(fill = "steelblue") +
  labs(
    title = "Count of Iris Species",
    x = "Species",
    y = "Count"
  )
```



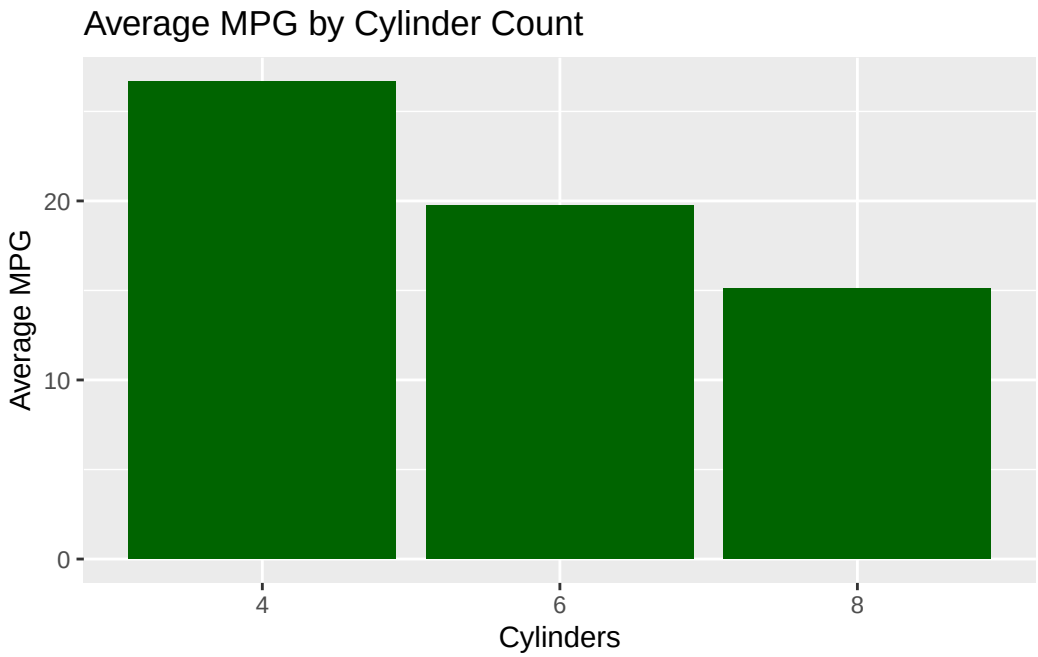

```
# Bar plot with custom colors
ggplot(iris, aes(x = Species, fill = Species)) +
  geom_bar() +
  labs(title = "Iris Species Distribution") +
  theme_minimal()
```




```
# Bar plot for mtcars (count by cylinders)
ggplot(mtcars, aes(x = factor(cyl))) +
  geom_bar(fill = "coral") +
  labs(
    title = "Number of Cars by Cylinder Count",
    x = "Number of Cylinders",
    y = "Count"
  )
```

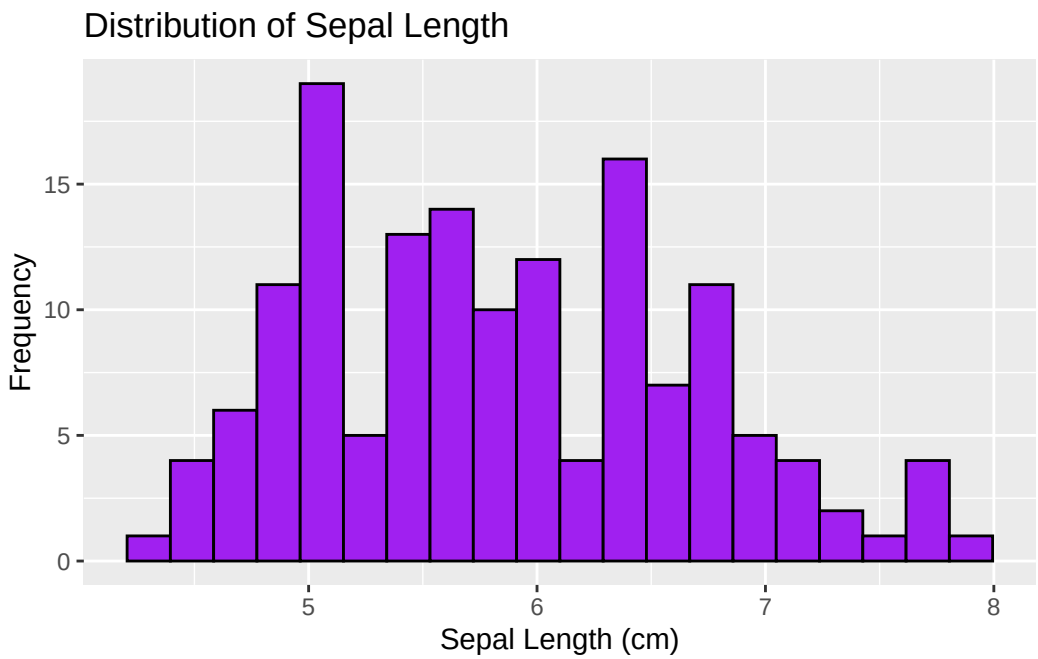



```
# Bar plot with aggregated data
mtcars %>%
  group_by(cyl) %>%
  summarize(avg_mpg = mean(mpg)) %>%
  ggplot(aes(x = factor(cyl), y = avg_mpg)) +
  geom_col(fill = "darkgreen") +
  labs(
    title = "Average MPG by Cylinder Count",
    x = "Cylinders",
    y = "Average MPG"
  )
```

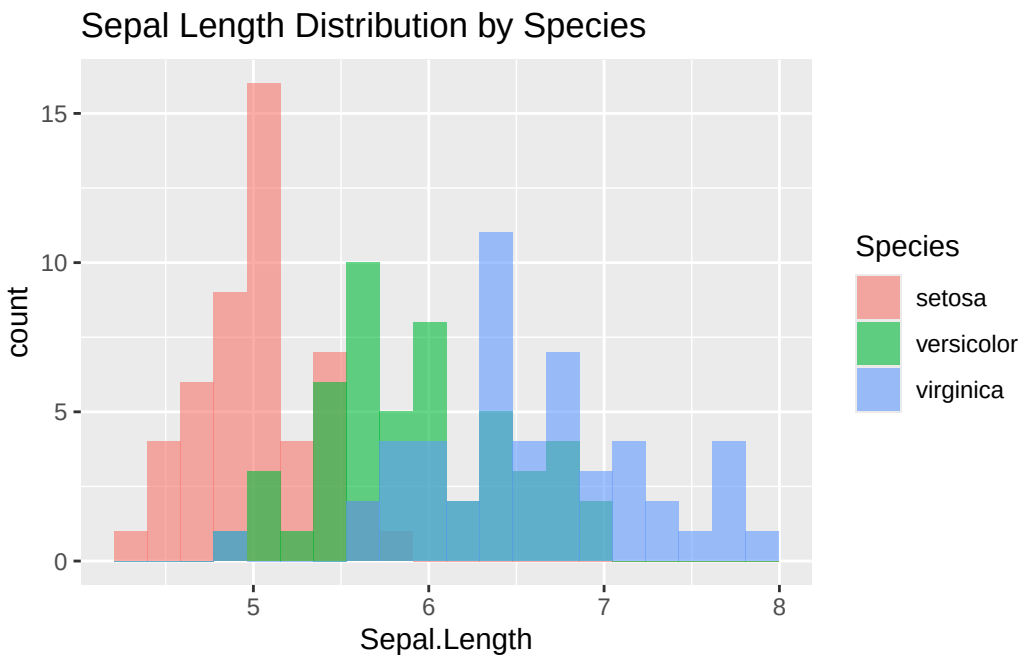


1.4.3 3. Histograms (Distribution of numeric variable)

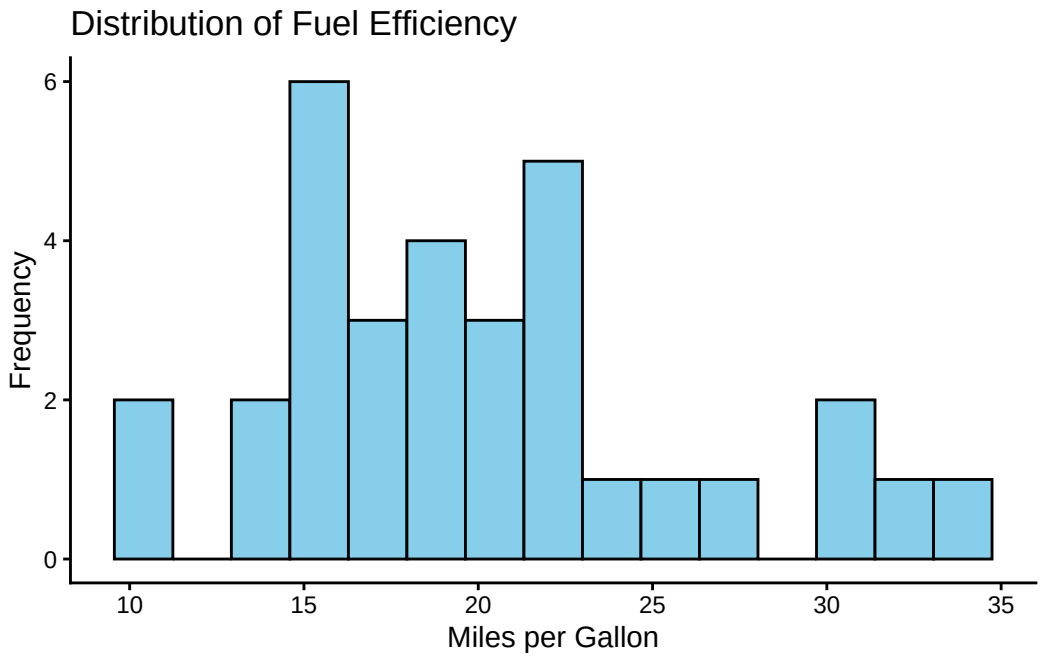
```
# Basic histogram
ggplot(iris, aes(x = Sepal.Length)) +
  geom_histogram(bins = 20, fill = "purple", color = "black") +
  labs(
    title = "Distribution of Sepal Length",
    x = "Sepal Length (cm)",
    y = "Frequency"
  )
```




```
# Histogram by species
ggplot(iris, aes(x = Sepal.Length, fill = Species)) +
  geom_histogram(bins = 20, alpha = 0.6, position = "identity") +
  labs(title = "Sepal Length Distribution by Species")
```



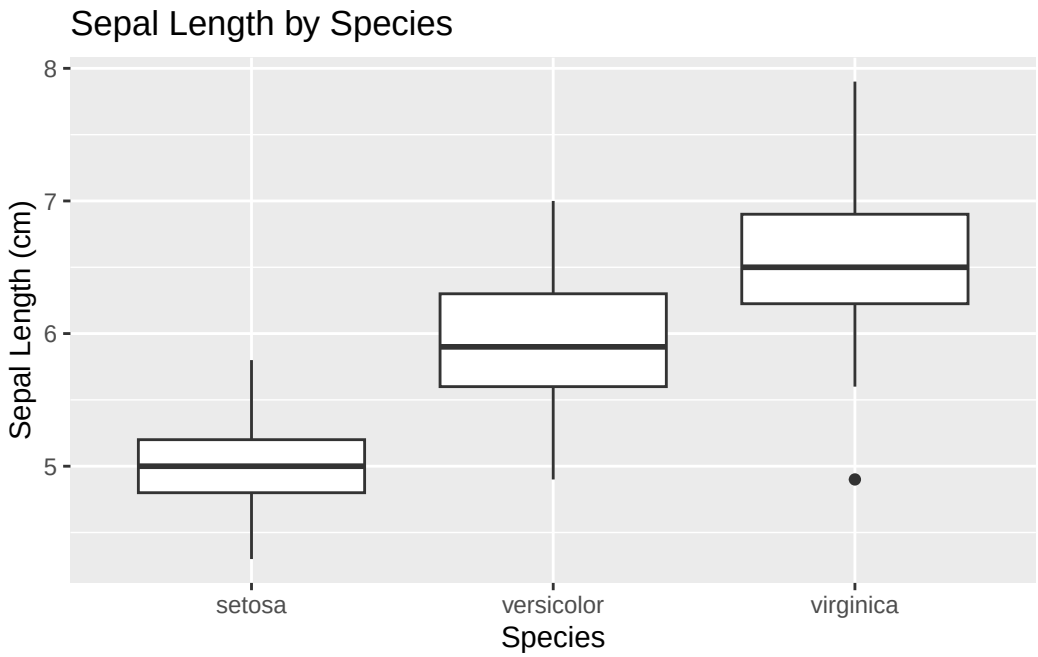

```r
# Histogram for mtcars
ggplot(mtcars, aes(x = mpg)) +
  geom_histogram(bins = 15, fill = "skyblue", color = "black") +
  labs(
    title = "Distribution of Fuel Efficiency",
    x = "Miles per Gallon",
    y = "Frequency"
  ) +
  theme_classic()
```



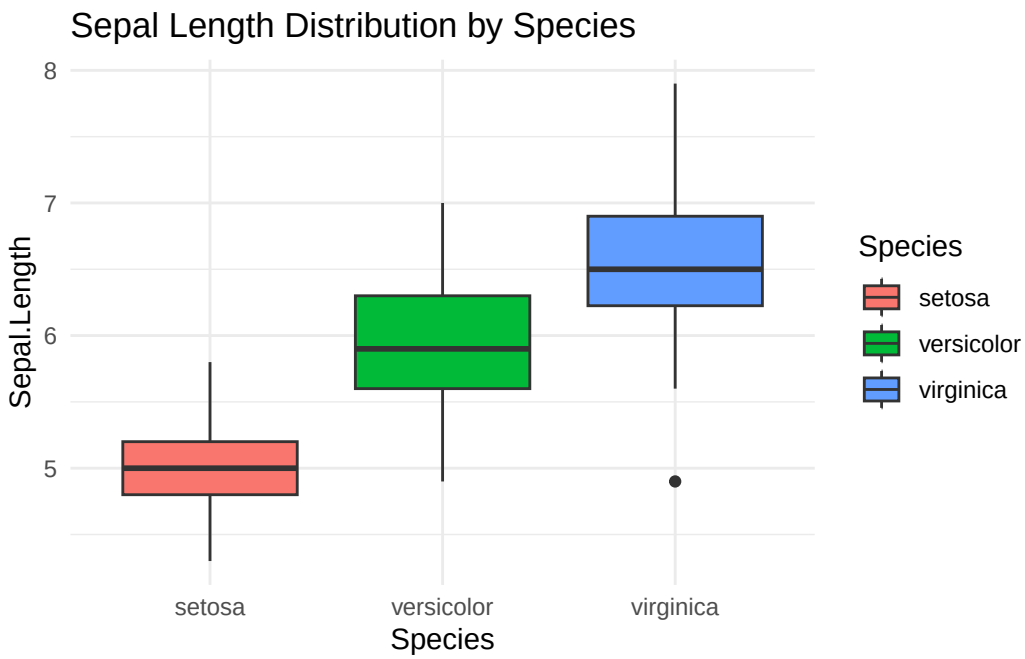

```
# Density plot (smooth histogram)
ggplot(iris, aes(x = Petal.Length, fill = Species)) +
  geom_density(alpha = 0.5) +
  labs(title = "Petal Length Density by Species")
```

1.4.4 4. Box Plots (Distribution and outliers)

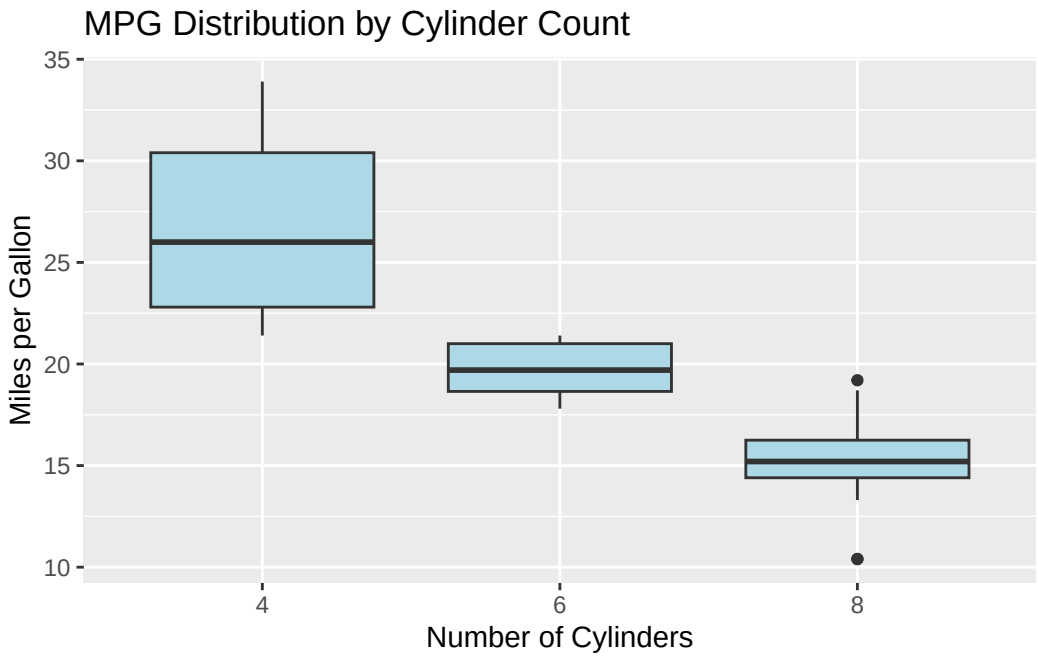
```
# Basic boxplot
ggplot(iris, aes(x = Species, y = Sepal.Length)) +
  geom_boxplot() +
  labs(
    title = "Sepal Length by Species",
    x = "Species",
    y = "Sepal Length (cm)"
  )
```



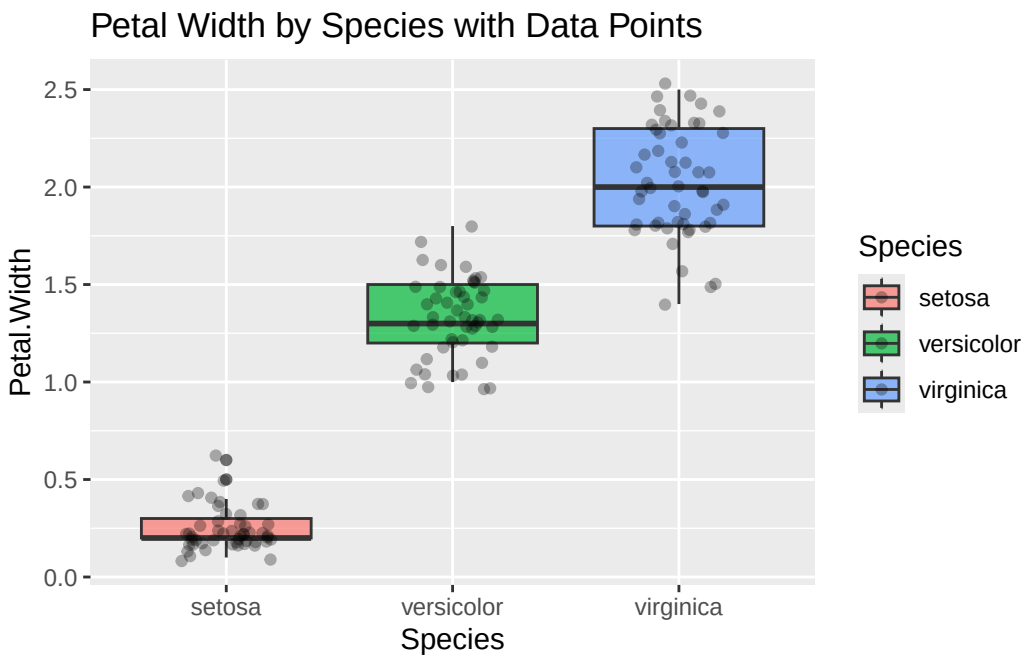

```
# Boxplot with color
ggplot(iris, aes(x = Species, y = Sepal.Length, fill = Species)) +
  geom_boxplot() +
  labs(title = "Sepal Length Distribution by Species") +
  theme_minimal()
```




```
# Boxplot for mtcars
ggplot(mtcars, aes(x = factor(cyl), y = mpg)) +
  geom_boxplot(fill = "lightblue") +
  labs(
    title = "MPG Distribution by Cylinder Count",
    x = "Number of Cylinders",
    y = "Miles per Gallon"
  )
```



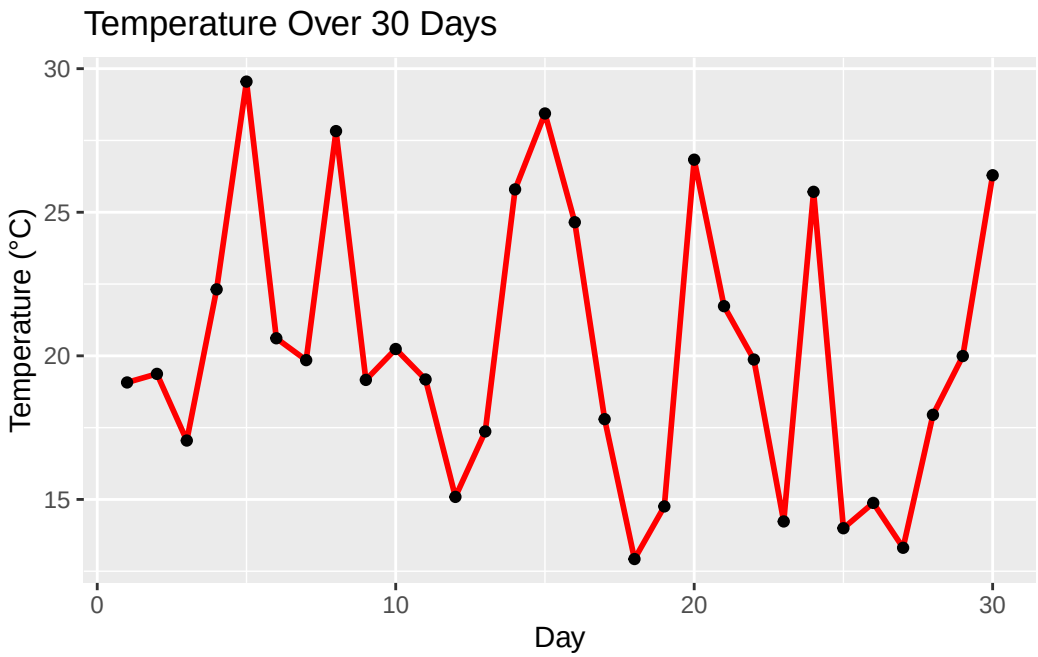

```
# Boxplot with points overlay
ggplot(iris, aes(x = Species, y = Petal.Width, fill = Species)) +
  geom_boxplot(alpha = 0.7) +
  geom_jitter(width = 0.2, alpha = 0.3) +
  labs(title = "Petal Width by Species with Data Points")
```



1.4.5 5. Line Plots (Trends over time or ordered data)

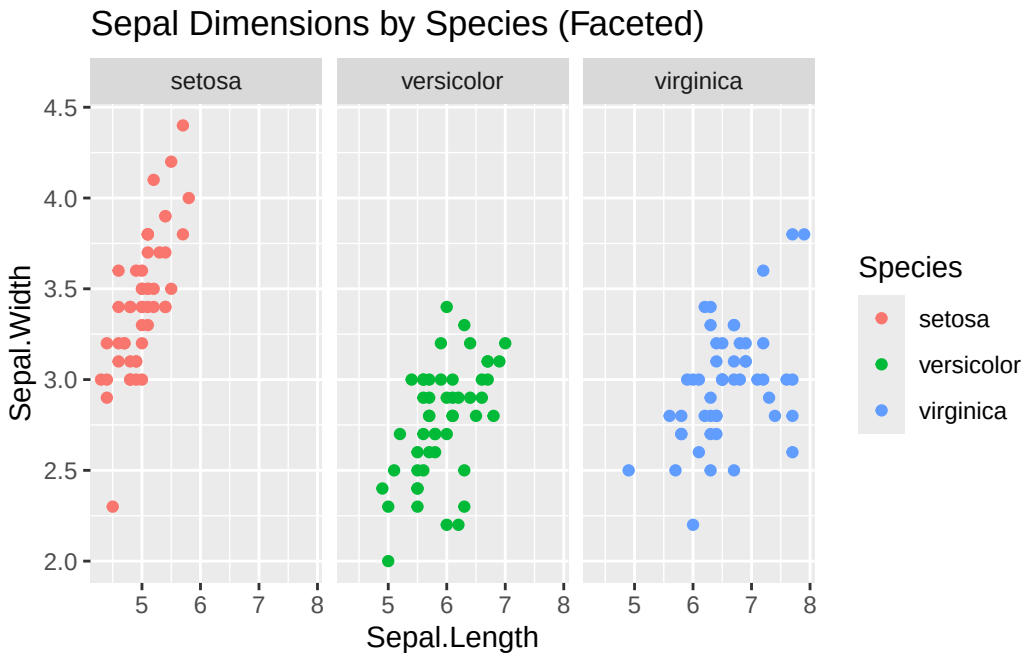
```r
# Create sample time series data
time_data <- data.frame(
  day = 1:30,
  temperature = rnorm(30, mean = 20, sd = 5)
)

ggplot(time_data, aes(x = day, y = temperature)) +
  geom_line(color = "red", size = 1) +
  geom_point() +
  labs(
    title = "Temperature Over 30 Days",
    x = "Day",
    y = "Temperature (°C)"
  )
```

```
Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.
i Please use `linewidth` instead.
```

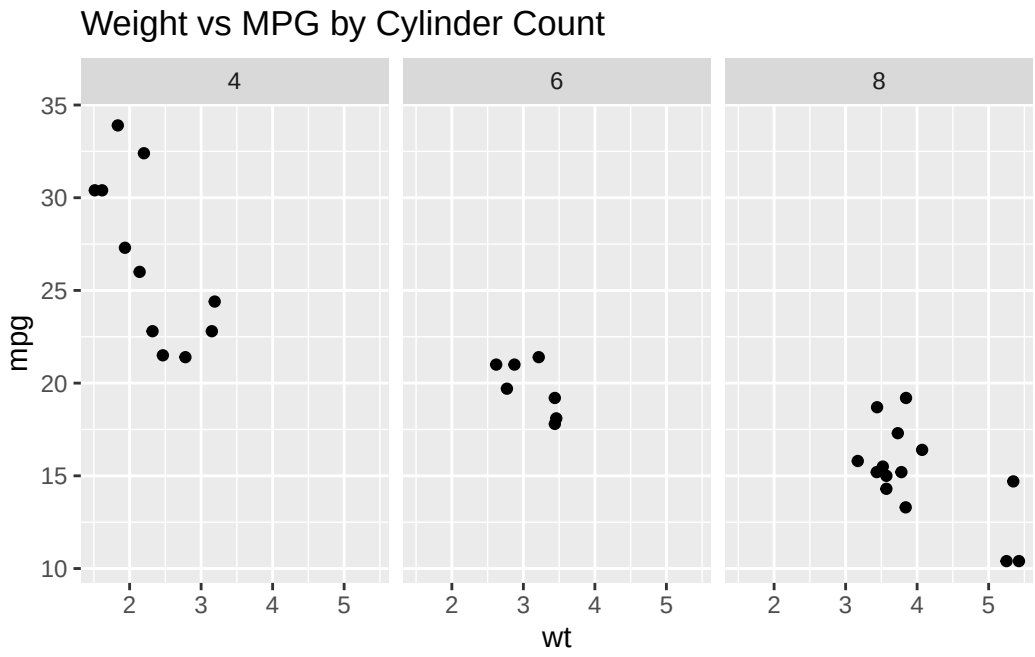


1.4.6 Multiple Plots in One (Faceting)

```
# Facet by species
ggplot(iris, aes(x = Sepal.Length, y = Sepal.Width)) +
  geom_point(aes(color = Species)) +
  facet_wrap(~ Species) +
  labs(title = "Sepal Dimensions by Species (Faceted)")
```

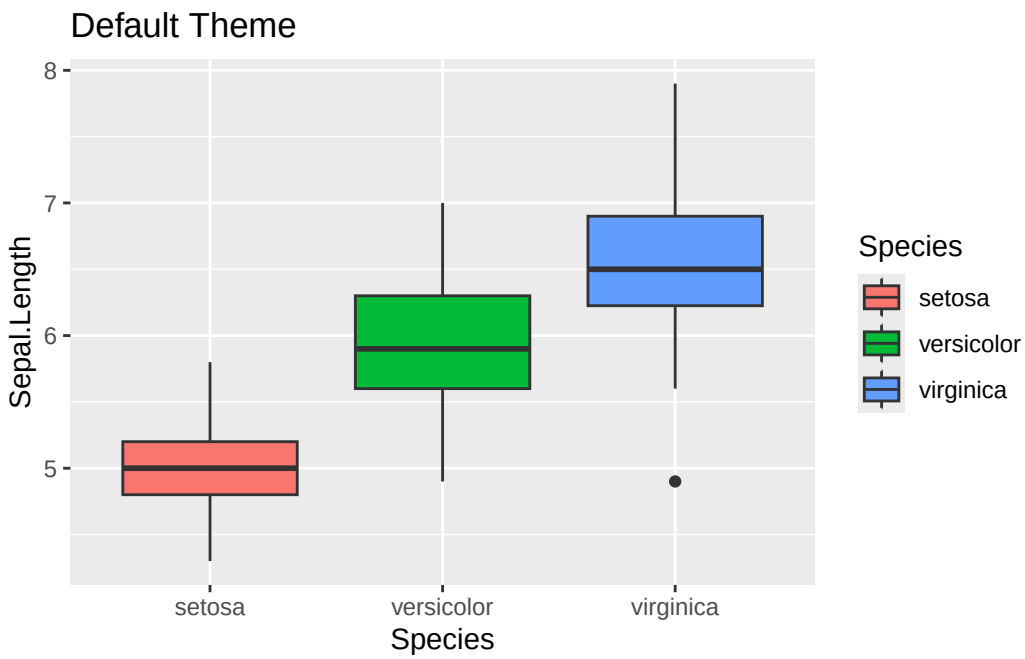



```
# Facet for mtcars
ggplot(mtcars, aes(x = wt, y = mpg)) +
  geom_point() +
  facet_wrap(~ cyl) +
  labs(title = "Weight vs MPG by Cylinder Count")
```

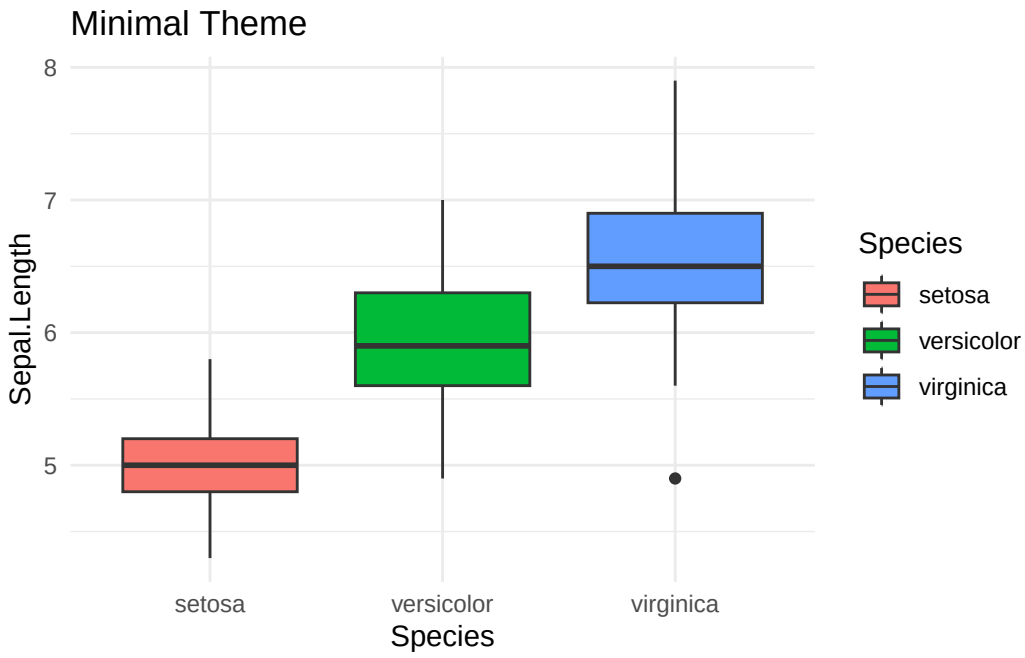


1.4.7 Customizing Themes

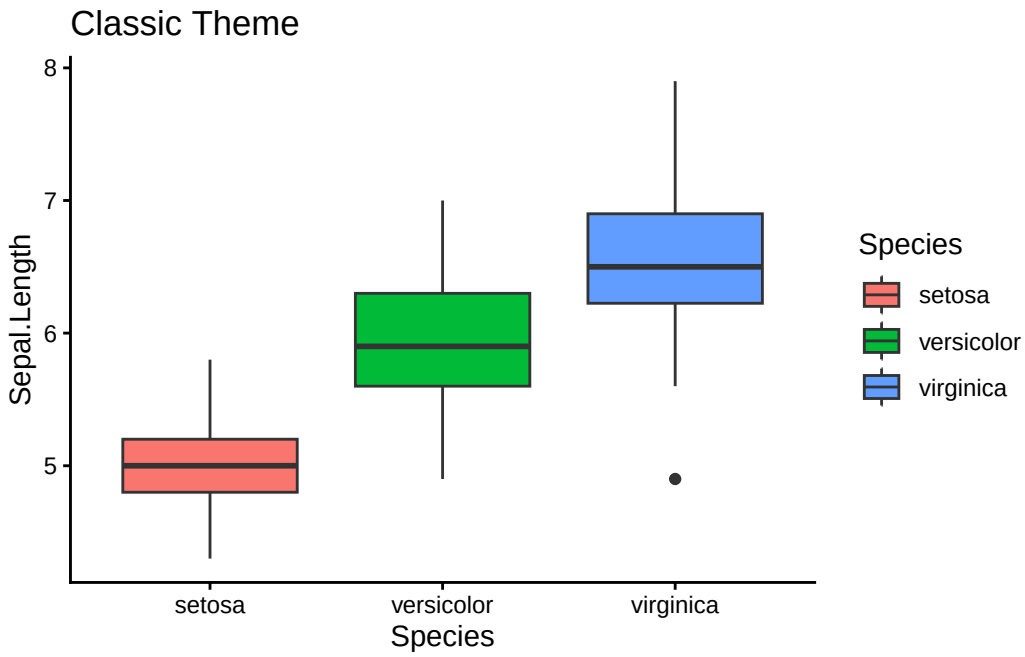
```
# Different themes
p <- ggplot(iris, aes(x = Species, y = Sepal.Length, fill = Species)) +
  geom_boxplot()

# Default
p + labs(title = "Default Theme")
```



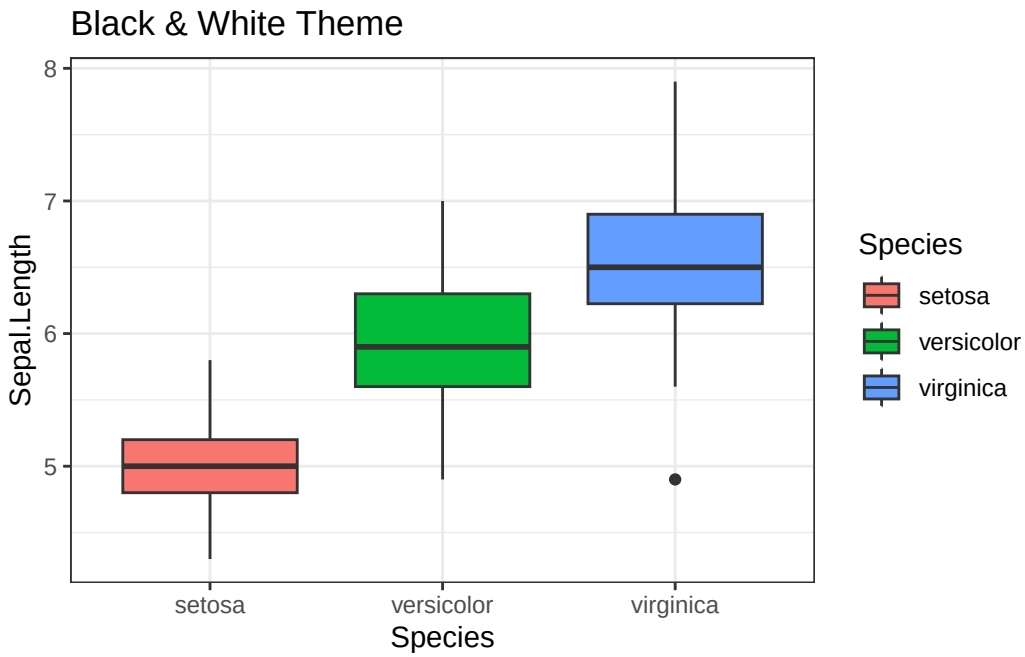

```
# Minimal
p + theme_minimal() + labs(title = "Minimal Theme")
```



```
# Classic
p + theme_classic() + labs(title = "Classic Theme")
```




```
# Black and white
p + theme_bw() + labs(title = "Black & White Theme")
```



1.5 Part 4: Practice and Q&A (10 mins)

1.5.1 Practice Tasks

1.5.1.1 Task 1: Descriptive Statistics

Calculate mean, median, and standard deviation for `mtcars$hp` (horsepower).

```
# Solution
mean(mtcars$hp)
```

```
[1] 146.6875
```

```
median(mtcars$hp)
```

```
[1] 123
```

```r
sd(mtcars$hp)
```

```
[1] 68.56287
```

```r
# Summary
summary(mtcars$hp)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   52.0    96.5   123.0   146.7   180.0   335.0
```

1.5.1.2 Task 2: Grouped Summary

Find the average petal length for each iris species.

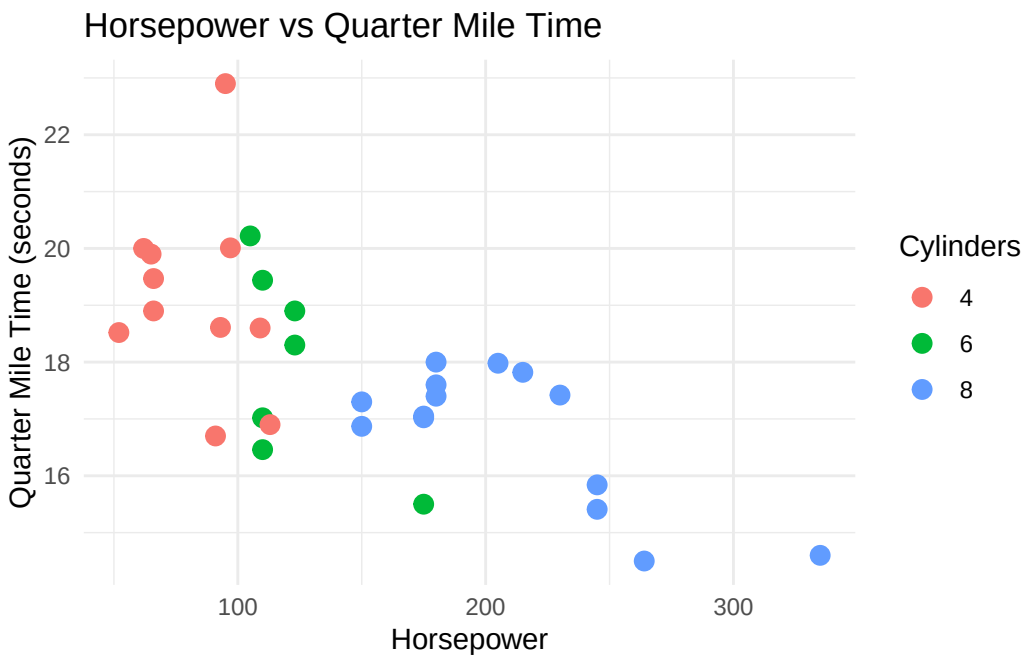
```r
# Solution
iris %>%
  group_by(Species) %>%
  summarize(avg_petal_length = mean(Petal.Length))
```

```
# A tibble: 3 x 2
  Species    avg_petal_length
  <fct>                 <dbl>
1 setosa                 1.46
2 versicolor             4.26
3 virginica              5.55
```

1.5.1.3 Task 3: Create a Scatter Plot

Plot the relationship between `mtcars$hp` and `mtcars$qsec` (quarter mile time).

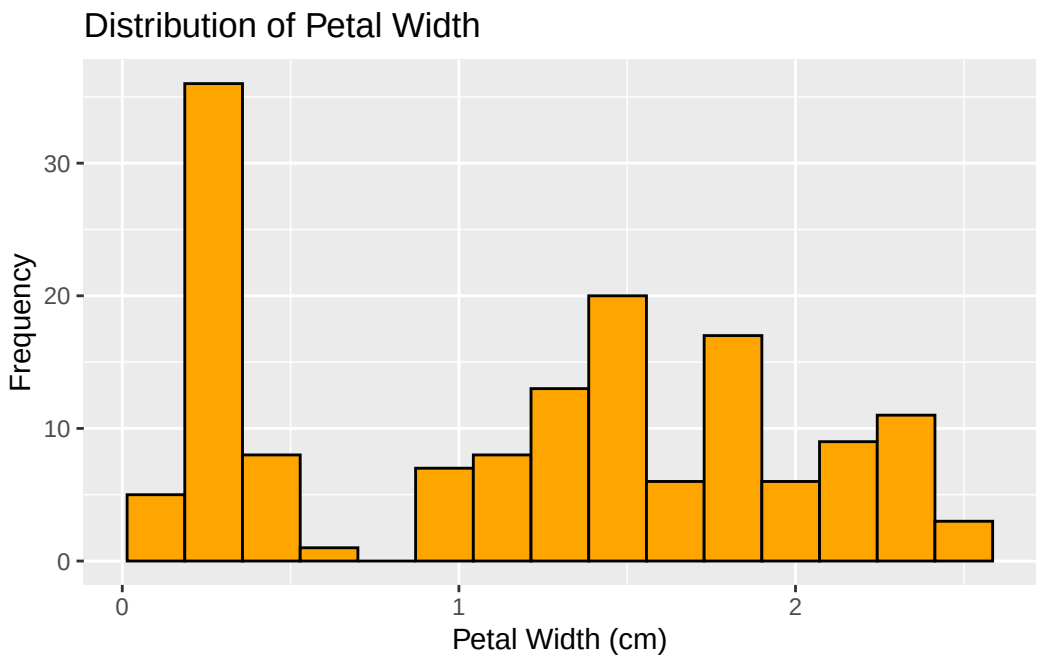
```r
# Solution
ggplot(mtcars, aes(x = hp, y = qsec)) +
  geom_point(aes(color = factor(cyl)), size = 3) +
  labs(
    title = "Horsepower vs Quarter Mile Time",
    x = "Horsepower",
    y = "Quarter Mile Time (seconds)",
    color = "Cylinders"
  ) +
  theme_minimal()
```



1.5.1.4 Task 4: Create a Histogram

Create a histogram of `iris$Petal.Width` with 15 bins.

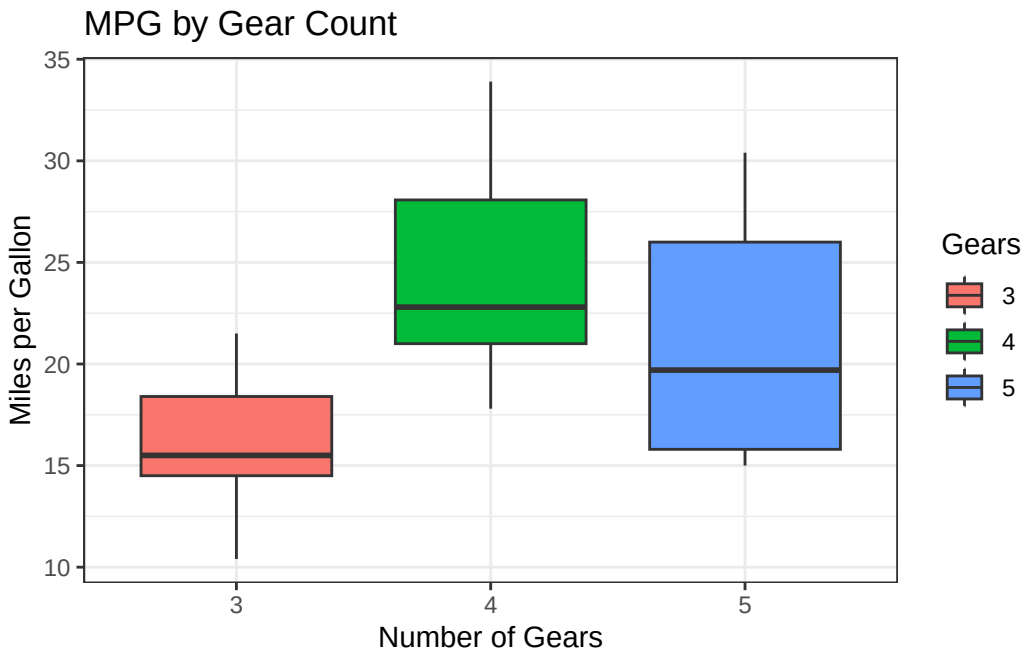
```
# Solution
ggplot(iris, aes(x = Petal.Width)) +
  geom_histogram(bins = 15, fill = "orange", color = "black") +
  labs(
    title = "Distribution of Petal Width",
    x = "Petal Width (cm)",
    y = "Frequency"
  )
```



1.5.1.5 Task 5: Create a Boxplot

Create a boxplot comparing `mpg` across different gear counts in `mtcars`.

```
# Solution
ggplot(mtcars, aes(x = factor(gear), y = mpg, fill = factor(gear))) +
  geom_boxplot() +
  labs(
    title = "MPG by Gear Count",
    x = "Number of Gears",
    y = "Miles per Gallon",
    fill = "Gears"
  ) +
  theme_bw()
```



1.5.2 Challenge: Combined Analysis

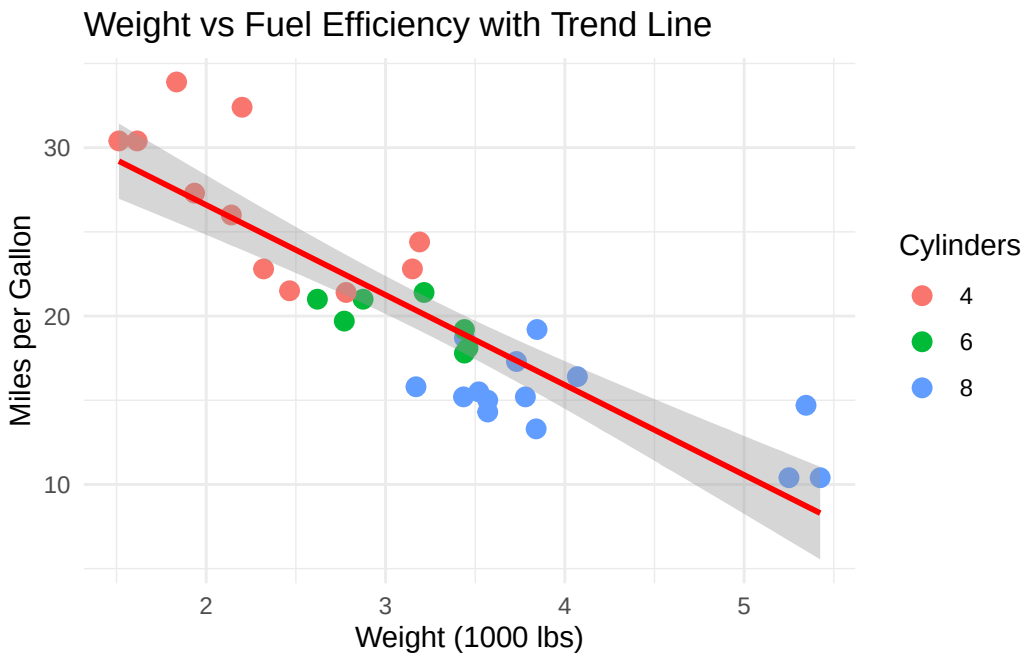
Analyze the relationship between car weight and fuel efficiency: 1. Calculate correlation 2. Create a scatter plot 3. Add a trend line

```
# 1. Correlation
cor(mtcars$wt, mtcars$mpg)
```

```
[1] -0.8676594
```

```
# 2 & 3. Scatter plot with trend line
ggplot(mtcars, aes(x = wt, y = mpg)) +
  geom_point(aes(color = factor(cyl)), size = 3) +
  geom_smooth(method = "lm", se = TRUE, color = "red") +
  labs(
    title = "Weight vs Fuel Efficiency with Trend Line",
    x = "Weight (1000 lbs)",
    y = "Miles per Gallon",
    color = "Cylinders"
  ) +
  theme_minimal()
```

```
`geom_smooth()` using formula = 'y ~ x'
```



1.6 Key Takeaways

Descriptive statistics summarize data: mean, median, sd, range
`summary()` gives quick overview of data
`cor()` measures linear relationships
ggplot2 follows: `ggplot(data, aes()) + geom_*()`
Common plots: scatter, bar, histogram, boxplot
Use `facet_wrap()` for multiple plots
Customize with themes and labels

1.7 Resources

- ggplot2 Documentation: https://ggplot2.tidyverse.org/
- R Graph Gallery: https://r-graph-gallery.com/
- ggplot2 Cheatsheet: https://posit.co/resources/cheatsheets/