Analysis of Variance (ANOVA) in R Using the mtcars Dataset

This tutorial provides a step-by-step guide on performing Analysis of Variance (ANOVA) in R using the built-in "mtcars" dataset. ANOVA is a statistical method used to compare means between two or more groups.

Step 1: Load the mtcars Dataset

Mtcars is included in base R so there is no need to download it from another source.

```
3  #Step 1: Load the mtcars Dataset
4  data(mtcars)
5  head(mtcars)
```

The 'data()' function will open the dataset in your workspace.
The 'head()' function will display the first few rows of the dataset. Here is what will appear when you run this line:

```
                   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
```

Step 2: Clean and Transform the Data (if needed)

Check for any missing values and clean if needed.

```
7  #Step 2: Clean and Transform the Data (if needed)
8  any(is.na(mtcars))
```

The 'any(is.na())' function will check for any missing values. Because there are no missing values in mtcars, no cleaning is needed.

Step 3: View and Summarize the mtcars Dataset

Examine the structure of the dataset and load a summary of statistics for each variable.

```
10 #Step 3: View and Summarize the mtcars Dataset
11 str(mtcars)
12 summary(mtcars)
```

The 'str()' function will display information about the variables and their types. This is what will appear when you run the line:

```
'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 ...
```

The 'summary()' function will display the summary statistics of your variables. This includes information such as minimum and maximum, median, and mean. This is what will appear when you run the line:

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

Step 4: Perform ANOVA

Fit the ANOVA model and summarize the results of the ANOVA.

```
#Step 4: Perform ANOVA
anova_model <- aov(mpg ~ cyl, data = mtcars)
summary(anova_model)
```

The 'aov()' function will perform ANOVA.

The 'summary()' function will display a summary of the ANOVA results. This is what will appear when you run the line:

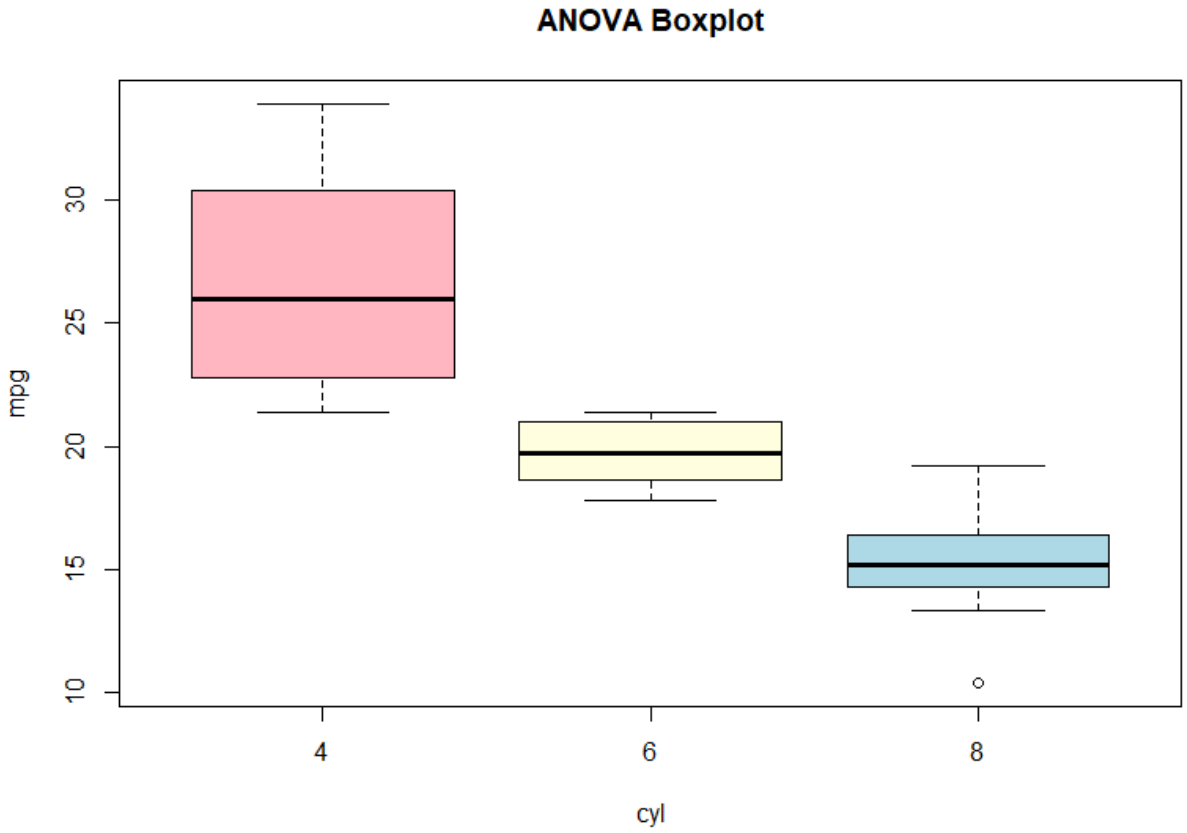
```
             Df Sum Sq Mean Sq F value   Pr(>F)    
cyl           1  817.7   817.7   79.56 6.11e-10 ***
Residuals    30  308.3    10.3                     
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Step 5: Visualize the Data

Use a boxplot to visualize the differences in the data.

```r
#Step 5: Visualize the Data
boxplot(mpg ~ cyl, data = mtcars,
        col = c("lightpink", "lightyellow", "lightblue"),
        main = "ANOVA Boxplot")
```

The 'boxplot()' function generates a boxplot using the provided variables. 'mpg ~ cyl' specifies that the variable being plotted against the cyl will be mpg. 'col' selects the color that the plot will be. 'main' provides a title for the plot. Here is how the graph will look:



Step 6: Interpret the Data and Make a Conclusion

Explain the ANOVA results, including the F-statistic, p-value, and any post-hoc tests if applicable. Interpret the findings and draw conclusions.

References:

- Example Tutorial: http://betsymccall.net/edu/CDSE/coding/R/bar_graphs.pdf