Exploratory Data Analysis

Motor Trends Project

JJC

1 Reflect on the Research Question

1. "Is an automatic or manual transmission better for MPG"
2. "How different is the MPG between automatic and manual transmission?"

In question 1, there is a comparison between a categorical variable (transmission) and the effect on response variable (MPG). "better" in this context means higher miles per gallon.

Question 2 calls for a quantitative comparision between the two transmission categories.

1.1 Data Narrative

1.1.1 Univariable numerical summaries

Table 1. Types of variables.
Inspect first 6 observations to determine types of variables and their values.

```
##       mpg       cyl      disp        hp      drat        wt      qsec
## "numeric" "numeric" "numeric" "numeric" "numeric" "numeric" "numeric"
##        vs        am      gear      carb
## "numeric" "numeric" "numeric" "numeric"
```

- Convert categorical variables to factor.

Table 2. Quartile summaries
Quartile summaries for *mpg*:

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    10.4    15.4    19.2    20.1    22.8    33.9
```

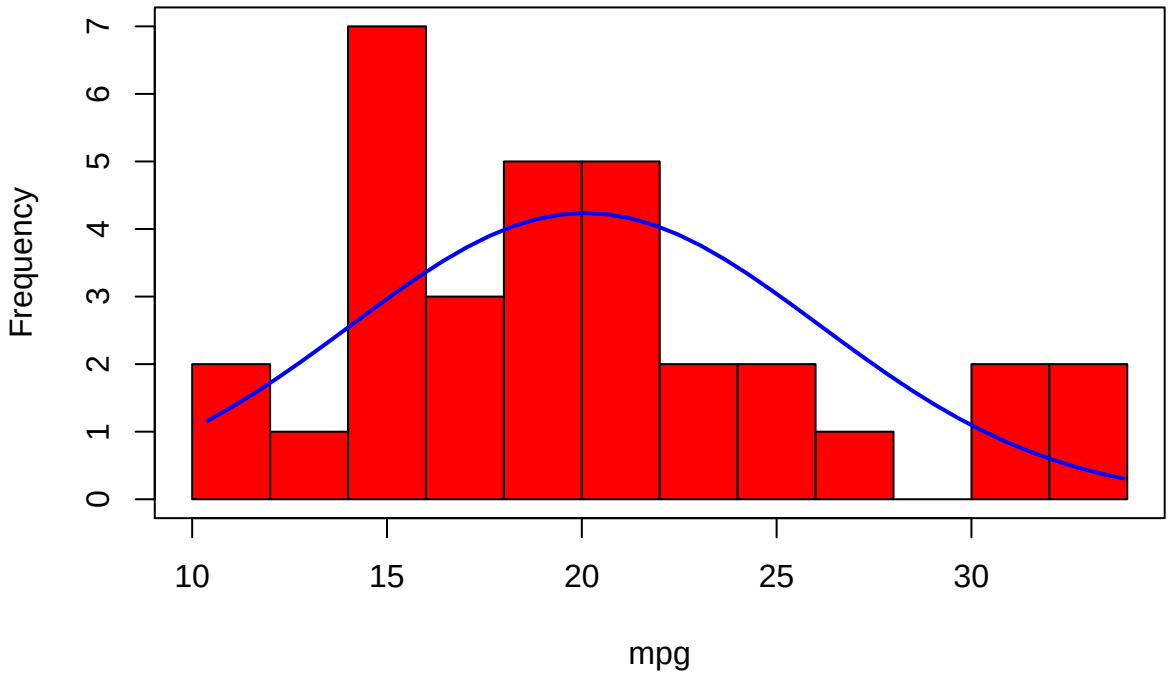
Counts for each transmission type:

```
##  auto manual
##    19     13
```

1.1.2 Distribution of Response Variable

Linear regression requires that the mean of the response is normal.

Figure 1. Histogram of mpg with Normal curve



By superimposing a normal curve, figure 1 shows that the distribution of mpg values is approximately normal.

1.1.3 Bivariate associations

Numeric Variables

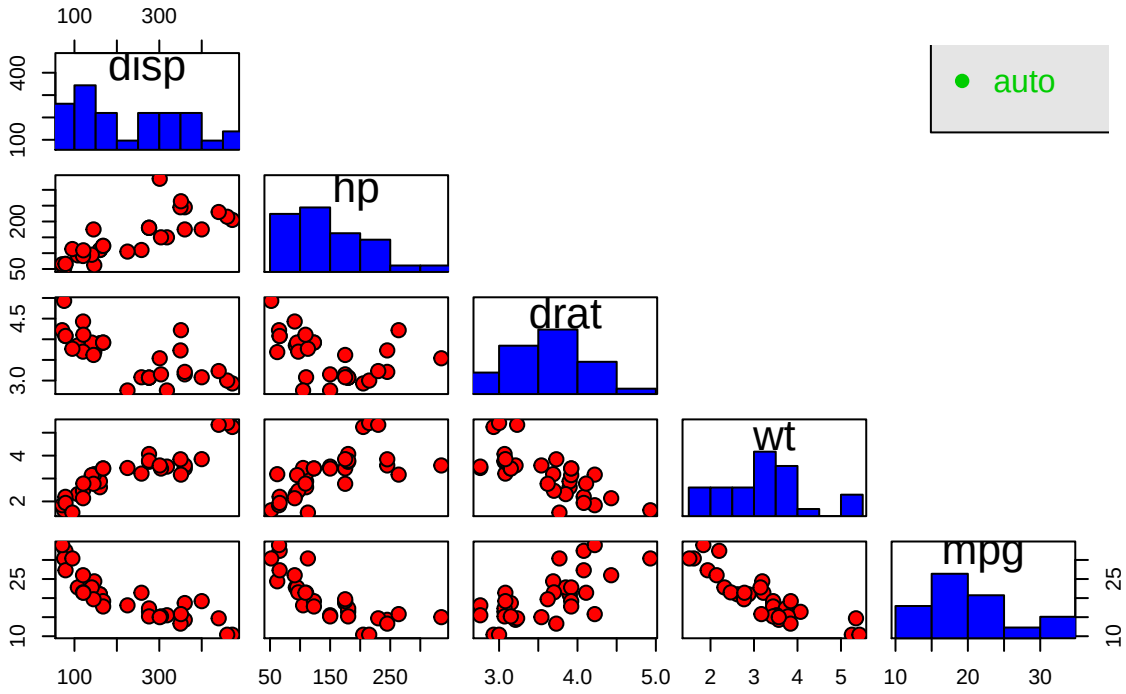
1. Correlation with respect to *mpg*

```
##         mpg
## disp -0.848
## hp   -0.776
## drat  0.681
## wt   -0.868
## qsec  0.419
```

Three variables (*disp*, *hp* and *wt*) have high negative correlation with *mpg*. So the more the powerful engine and the heavier the car leads to lower fuel efficiency. *qsec* has only moderate positive correlation with mpg and so, it does not appear in figure 2. The canon for high correlation here is an absolute value greater than 0.75.

2. Scatterplots

Figure 2. Motor Trends Measures − 2 transmission types

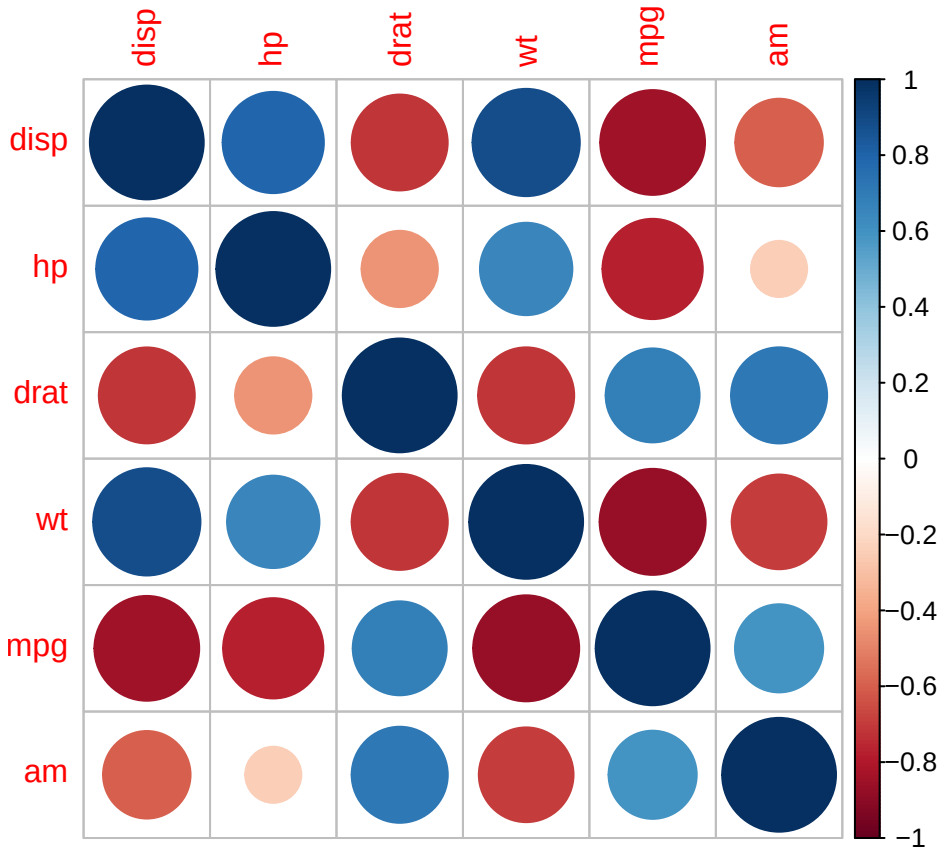


* In figure 2, the lowest row confirms the correlation between *mpg* and the four numerical explanatory variables. * Automatic transmission (green points) tends to have miles per gallon except for the *drat* measure. * Note that here is high correlation between the explanatory variables themselves. So checking for multicollinearity in linear regression diagnostics is important.

3. Correlation

```
## corrplot 0.84 loaded
```

```
##        disp     hp   drat     wt    mpg     am
## disp  1.000  0.791 -0.710  0.888 -0.848 -0.591
## hp    0.791  1.000 -0.449  0.659 -0.776 -0.243
## drat -0.710 -0.449  1.000 -0.712  0.681  0.713
## wt    0.888  0.659 -0.712  1.000 -0.868 -0.692
## mpg  -0.848 -0.776  0.681 -0.868  1.000  0.600
## am   -0.591 -0.243  0.713 -0.692  0.600  1.000
```




```
## [1] 0.888
```

high correlation between wt and disp

Categorical Variables

Check transmission groups by wt and by disp

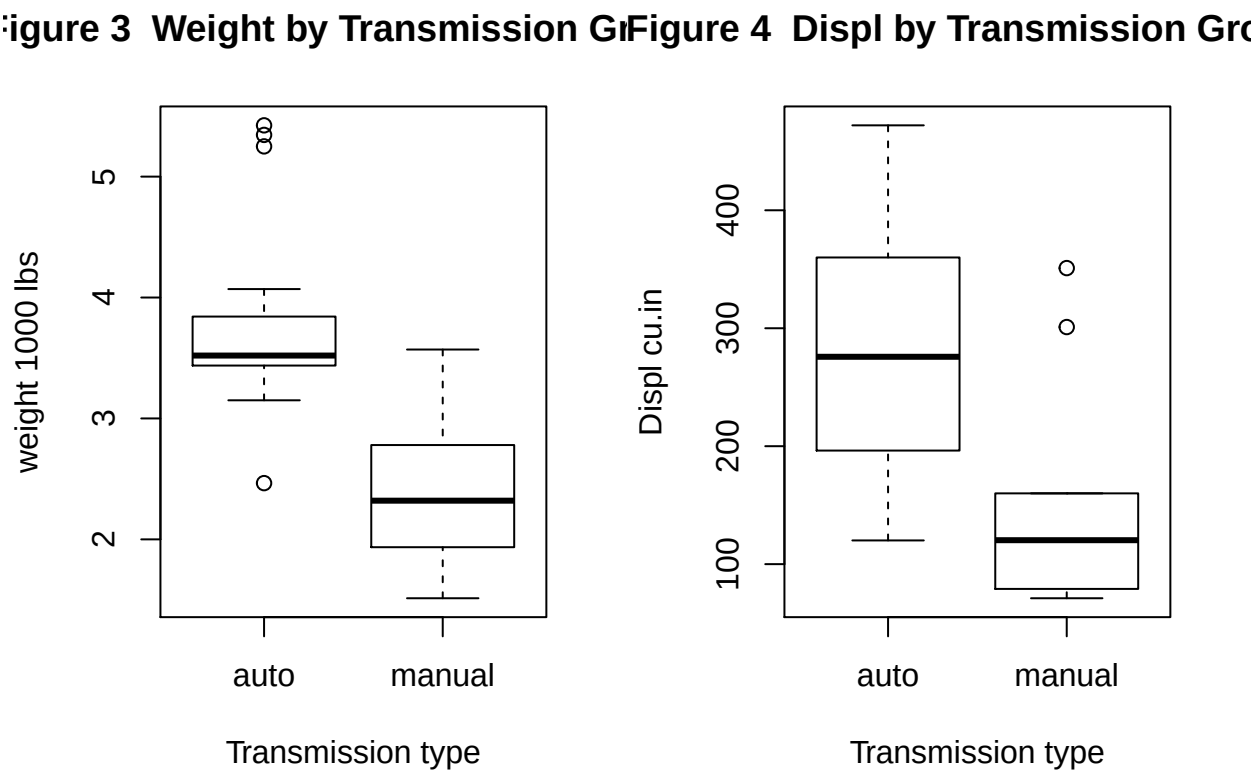
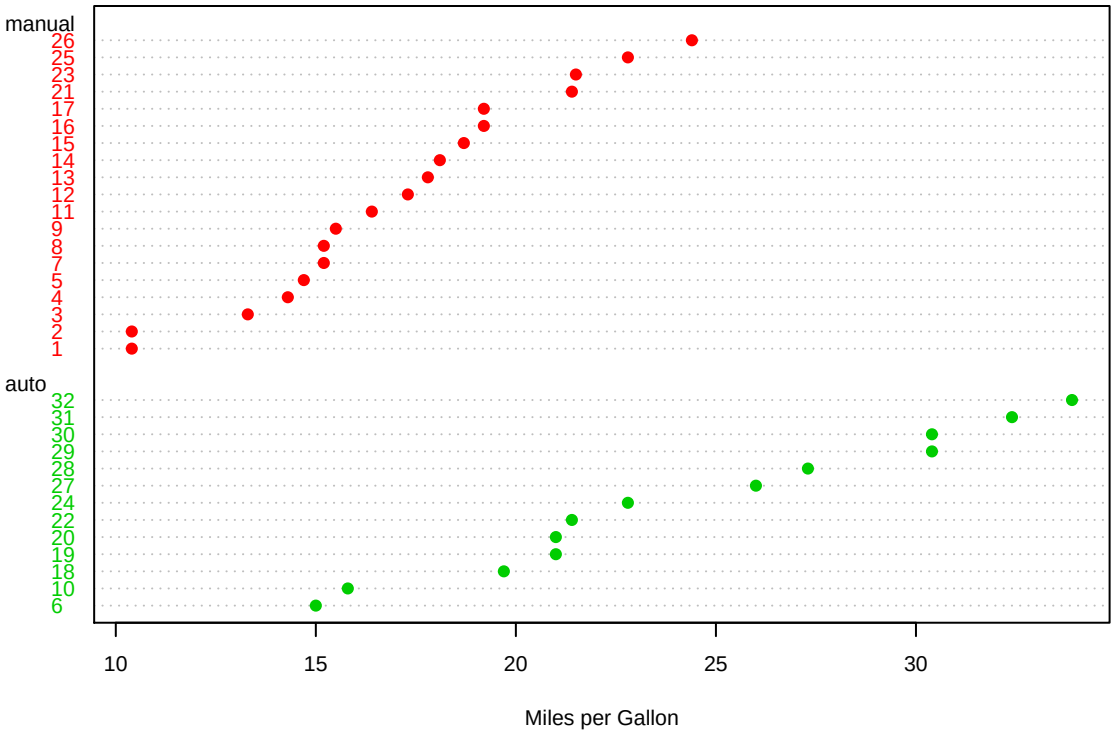


Figure 3 orders each car model by fuel mileage for the two transmission groups. Again, the *auto* group has higher *mpg*. However, there is crossover between the two groups where models have similar mileage values.

Figure 3, Car mileage for two transmission categories.



Question 1 asks if one group is "better". This seeks to determine if the mean of one group is significantly

higher than the other. One solution is to perform a single-tail t-test.

1.1.4 Assumptions for t-test

1. **Independent Sample**
   Differentiating by transmission category results in two independent samples since the manual group does not effect the automatic group.

2. **Independent Observations**
   The selection of one car model does not effect the selection of another model.
3. **Normality Assumption**
   Figure 2 shows that the manual group has a normal distribution. However, the automatic group has right skew. Also, it only contains 13 observations.

Table 1. *mpg* numeric summaries by transmission group.

n

mean

sd

min

max

manual

19

17.1

3.83

10.4

24.4

auto

13

24.4

6.17

15.0

33.9

However, Table 1 shows that both groups are within 2.5 times standard deviation of the mean. This suggests that a t-test is valid with a caveat about the distribution of the *am* group.

For question 2, a linear model with a single categorical regressor variable would quantify the difference between the two transmission categories. It would be worth comparing a model with only *am* as regressor to a model with *am* that accounts for all the other numeric variables. Accounting for the remaining categorical variables would require too many comparisons due to the number of levels involved.

1.1.5 mpg and binary categorical variables

t-test, Independent Samples mpg by transmission category

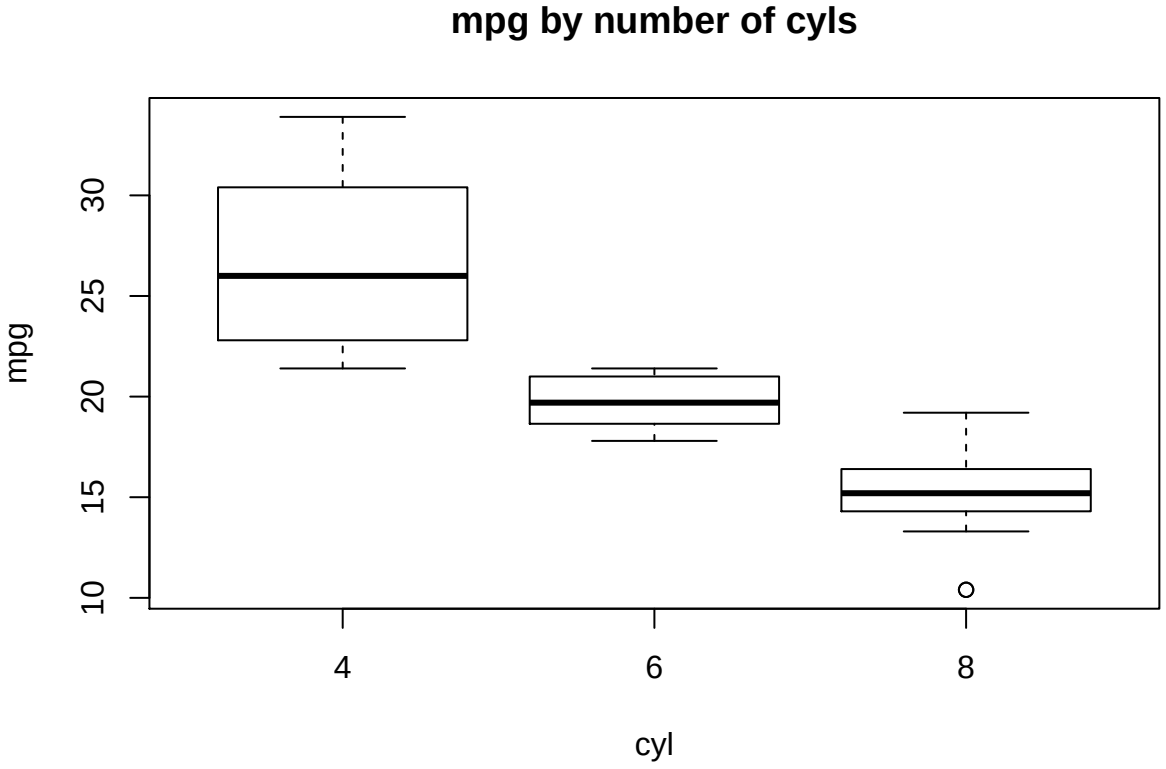
```
##
##  Welch Two Sample t-test
##
## data:  mpg_m and mpg_a
## t = 4, df = 20, p-value = 0.001
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##   3.21 11.28
## sample estimates:
## mean of x mean of y
##      24.4      17.1
```

mpg by vs

```
##
##  Welch Two Sample t-test
##
## data:  mpg_vs0 and mpg_vs1
## t = -5, df = 20, p-value = 1e-04
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -11.46  -4.42
## sample estimates:
## mean of x mean of y
##      16.6      24.6
```

1.1.6 mpg by multi-level categorical variables using ANOVA

mpg and cyl



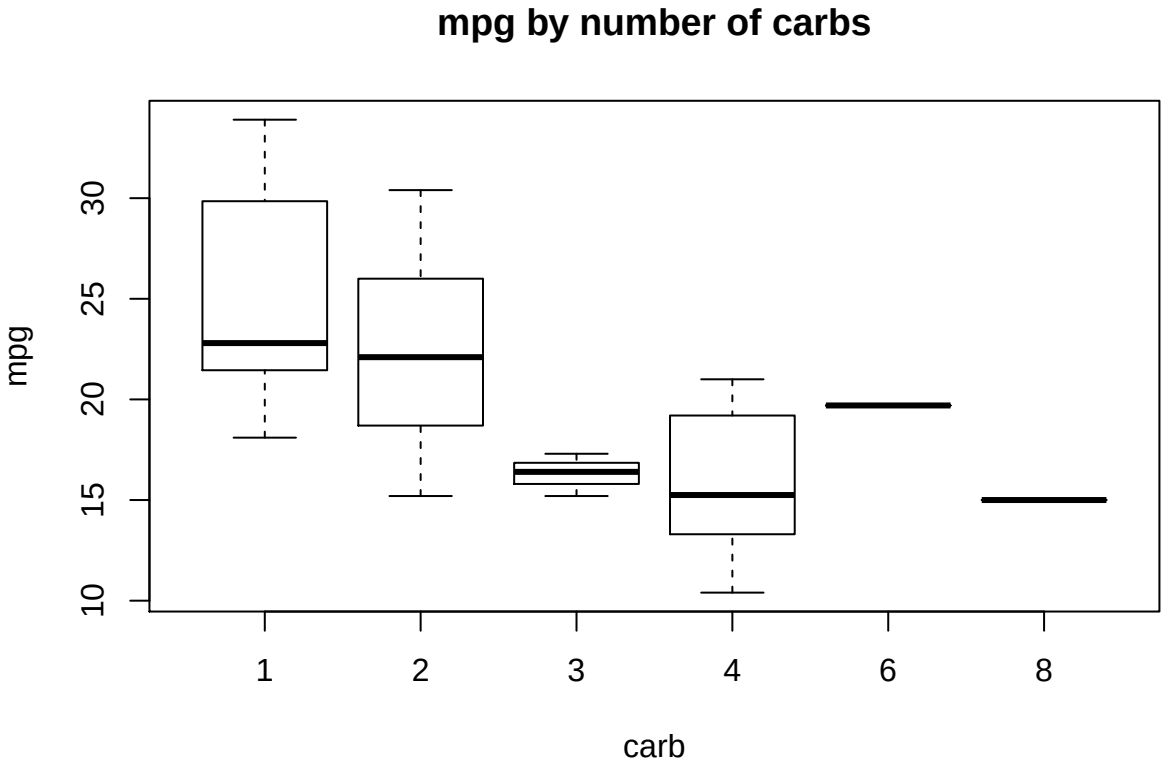
```
##   cyl  mpg
## 1   4 26.7
## 2   6 19.7
## 3   8 15.1
```

```
##   cyl  mpg
## 1   4 4.51
## 2   6 1.45
## 3   8 2.56
```

```
##             Df Sum Sq Mean Sq F value Pr(>F)
## cyl          2    825     412    39.7  5e-09 ***
## Residuals   29    301      10
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = mpg ~ cyl)
##
## $cyl
##       diff    lwr    upr p adj
## 6-4  -6.92 -10.77 -3.072 0.000
## 8-4 -11.56 -14.77 -8.356 0.000
## 8-6  -4.64  -8.33 -0.958 0.011
```

mpg and carb




```
## [1] 15
```

```
##   carb  mpg
## 1    1 25.3
```

```
## 2    2 22.4
## 3    3 16.3
## 4    4 15.8
## 5    6 19.7
## 6    8 15.0
```

```
##   carb  mpg
## 1    1 6.00
## 2    2 5.47
## 3    3 1.05
## 4    4 3.91
## 5    6   NA
## 6    8   NA
```

```
##             Df Sum Sq Mean Sq F value Pr(>F)
## carb         5    501   100.1    4.16 0.0065 **
## Residuals   26    625    24.1
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = mpg ~ carb)
##
## $carb
##       diff   lwr    upr p adj
## 2-1  -2.94 -10.4  4.484 0.824
## 3-1  -9.04 -19.4  1.356 0.116
## 4-1  -9.55 -17.0 -2.126 0.006
## 6-1  -5.64 -21.8 10.467 0.886
## 8-1 -10.34 -26.5  5.767 0.384
## 3-2  -6.10 -16.0  3.820 0.431
## 4-2  -6.61 -13.3  0.129 0.057
## 6-2  -2.70 -18.5 13.105 0.995
## 8-2  -7.40 -23.2  8.405 0.704
## 4-3  -0.51 -10.4  9.410 1.000
## 6-3   3.40 -14.0 20.801 0.990
## 8-3  -1.30 -18.7 16.101 1.000
## 6-4   3.91 -11.9 19.715 0.972
## 8-4  -0.79 -16.6 15.015 1.000
## 8-6  -4.70 -26.0 16.612 0.983
```