project

Project 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"
```

Loading mtcars dataset

```r
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
```

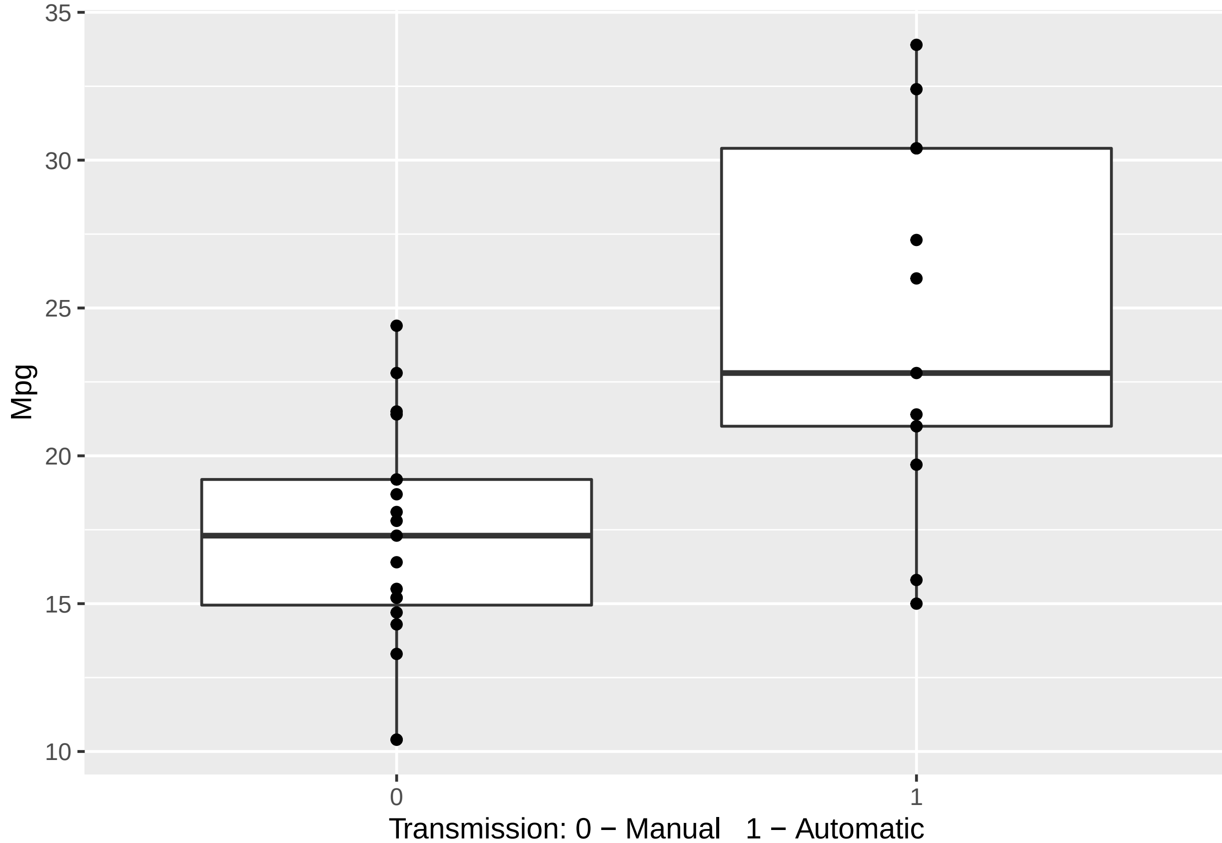
Analysis

```r
library(ggplot2)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

Exploring relationship between transmission and MPG

```r
mpg_vs_am <- mtcars %>% select(mpg, am) %>% mutate(am = as.factor(am))
ggplot(data = mpg_vs_am, aes(x=am  , y=mpg))+
  geom_boxplot() + geom_point()+
  xlab("Transmission: 0 - Manual   1 - Automatic")+
  ylab("Mpg")
```



Modeling

```
data <- mtcars %>% mutate(
                    cyl = as.factor(cyl),
                    vs = as.factor(vs),
                    am = as.factor(am),
                    gear = as.factor(am),
                    carb = as.factor(carb))

fit_all <- lm(mpg ~. , data = data)
summary(fit_all)$coef[,4]
```

```
## (Intercept)        cyl6        cyl8        disp          hp        drat
##  0.19839130  0.30216963  0.73788849  0.20273076  0.08080569  0.56166020
##          wt        qsec         vs1         am1       carb2       carb3
##  0.06114523  0.68543556  0.42300239  0.39689475  0.89500323  0.36651388
##       carb4       carb6       carb8
##  0.68216100  0.35952172  0.29327663
```

As none of the variables have a p-value less than 5%, we would have to remove most insignificant variables one by one.

```
which.max(summary(fit_all)$coef[,4])
```

```
## carb2
##    11
```

```
data <- data %>% select(-carb)
fit <- lm(mpg ~. , data = data)
summary(fit)$coef[,4]
```

```
## (Intercept)        cyl6        cyl8        disp          hp        drat
##  0.19323159  0.46992153  0.90252093  0.62551156  0.10399230  0.69939226
##          wt        qsec         vs1         am1
##  0.03690757  0.45639538  0.56269576  0.11966558
```

```
which.max(summary(fit)$coef[,4])
```

```
## cyl8
##    3
```

we now have on significant variable which is wt and we will continue this process to come up with the signifiacnt variables

```
data <- data %>% select(-cyl); fit <- lm(mpg ~. , data = data); summary(fit)$coef[,4]; which.max(summary
```

```
## (Intercept)        disp          hp        drat          wt        qsec
## 0.326616519 0.238211942 0.147781592 0.503756614 0.004567014 0.168194583
##         vs1         am1
## 0.750269228 0.082435144
```

```
## vs1
##   7
```

```
data <- data %>% select(-vs); fit <- lm(mpg ~. , data = data); summary(fit)$coef[,4]; which.max(summary
```

```
## (Intercept)        disp          hp        drat          wt        qsec
## 0.338475309 0.244054196 0.149381426 0.462401185 0.002536163 0.049550895
##         am1
## 0.079692318
```

```
## drat
##    4
```

```
data <- data %>% select(-drat); fit <- lm(mpg ~. , data = data); summary(fit)$coef[,4]; which.max(summa
```

```
## (Intercept)        disp          hp          wt        qsec         am1
## 0.152378367 0.298972150 0.156387279 0.002075008 0.043907652 0.027487809
```

```
## disp
##    2
```

```
data <- data %>% select(-disp); fit <- lm(mpg ~. , data = data); summary(fit)$coef[,4]; which.max(summa
```

```
## (Intercept)          hp          wt        qsec         am1
## 0.072149342 0.223087932 0.001141407 0.075731202 0.045790788
```

```
## hp
##  2
```

```r
data <- data %>% select(-hp); fit <- lm(mpg ~. , data = data); summary(fit)$coef[,4]; which.max(summary
```

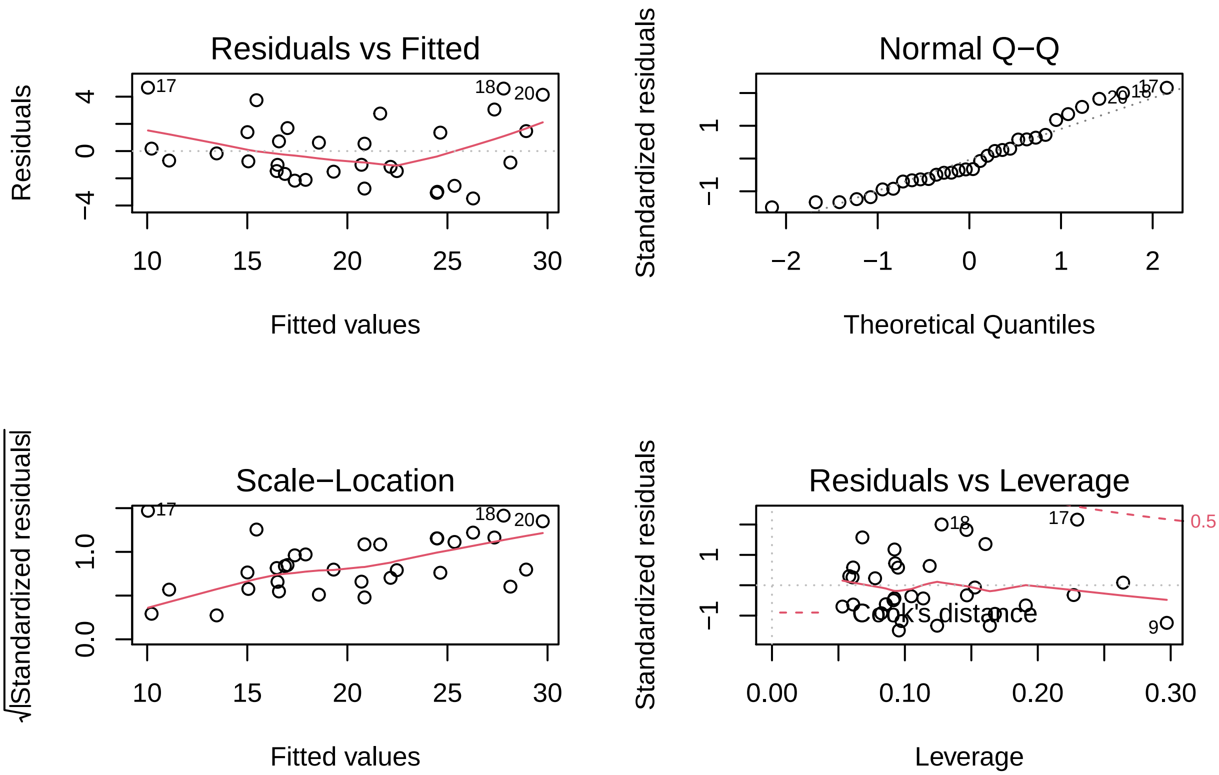
```
##  (Intercept)           wt         qsec          am1
## 1.779152e-01 6.952711e-06 2.161737e-04 4.671551e-02

## (Intercept)
##           1
```

No we have only significant variables which are wt, qsec, am

```r
summary(fit)
```

```
##
## Call:
## lm(formula = mpg ~ ., data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.4811 -1.5555 -0.7257  1.4110  4.6610
##
## Coefficients: (1 not defined because of singularities)
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   9.6178     6.9596   1.382 0.177915
## wt           -3.9165     0.7112  -5.507 6.95e-06 ***
## qsec          1.2259     0.2887   4.247 0.000216 ***
## am1           2.9358     1.4109   2.081 0.046716 *
## gear1             NA         NA      NA       NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.459 on 28 degrees of freedom
## Multiple R-squared:  0.8497, Adjusted R-squared:  0.8336
## F-statistic: 52.75 on 3 and 28 DF,  p-value: 1.21e-11
```

```r
par(mfrow = c(2, 2))
plot(fit)
```



The QQ plot shows a pretty good correlation of the standardized and theoretical residuals. There also doesn't seem to be any significant patterns in the other three plots, indicating a good fit of the selected model

Conclusion

we can conclude that if weight and 1/4 mile time are same for the two transmission, the manual transmission car will have 2.9358 higher miles/gallon than the automatic transmission car.