MPG data exploring

maesh

5/2/2021

Introduction

In this course I explore the relationship between a set of variables and miles per gallon (MPG)

Particularly I focus on the following two questions:

1. Is an automatic or manual transmission better for MPG?
2. Quantify the MPG difference between automatic and manual transmissions.

exploratory data analysis

processing data

```
## tibble[,11] [234 x 11] (S3: tbl_df/tbl/data.frame)
##  $ manufacturer: chr [1:234] "audi" "audi" "audi" "audi" ...
##  $ model       : chr [1:234] "a4" "a4" "a4" "a4" ...
##  $ displ       : num [1:234] 1.8 1.8 2 2 2.8 2.8 3.1 1.8 1.8 2 ...
##  $ year        : int [1:234] 1999 1999 2008 2008 1999 1999 2008 1999 1999 2008 ...
##  $ cyl         : int [1:234] 4 4 4 4 6 6 6 4 4 4 ...
##  $ trans       : chr [1:234] "auto(l5)" "manual(m5)" "manual(m6)" "auto(av)" ...
##  $ drv         : chr [1:234] "f" "f" "f" "f" ...
##  $ cty         : int [1:234] 18 21 20 21 16 18 18 18 16 20 ...
##  $ hwy         : int [1:234] 29 29 31 30 26 26 27 26 25 28 ...
##  $ fl          : chr [1:234] "p" "p" "p" "p" ...
##  $ class       : chr [1:234] "compact" "compact" "compact" "compact" ...
```

```
##
##   auto manual
##    157     77
```

compare hwy,cty vs other valuables

(Graphs are shown in appendices)

As the graph shows, type of transmission,engine displacement, number of cylinders, drive type, fuel type and vehicle class seems to affect on city and highway mileage. It cannot be judged whether Year of manufacturing affect on mileage from the graph.

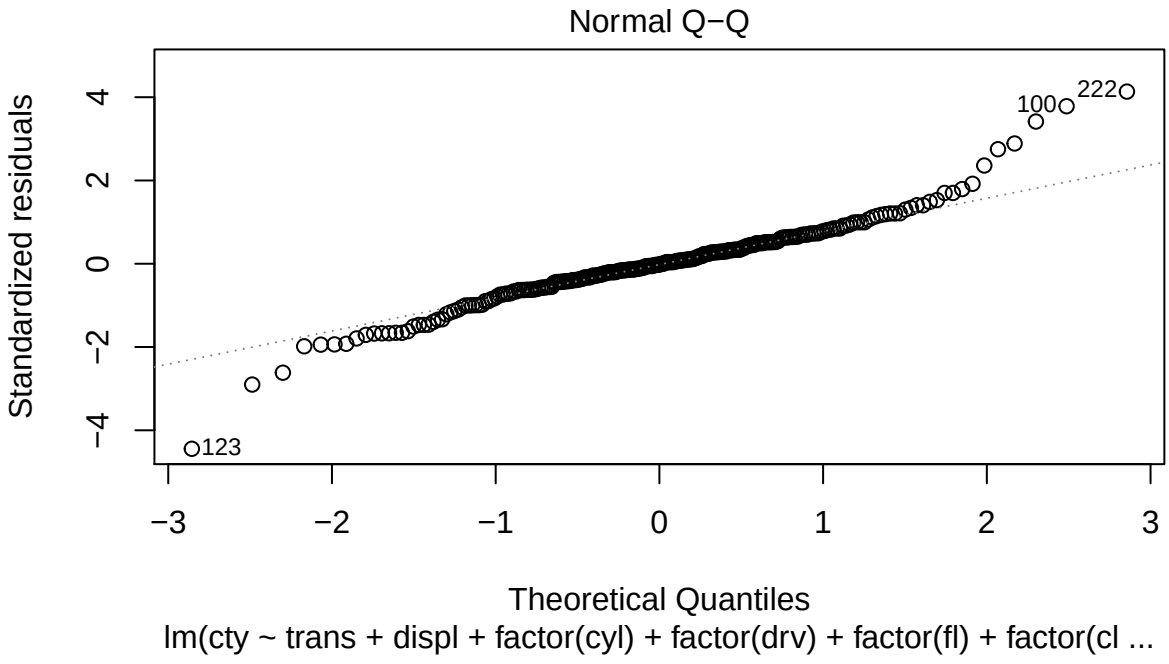
According to the graphs below, city mileage and highway mileage has the similar characteristics. Here I focused on city mileage.

model selection

I created several fitted model which contains variables of mpg datasets. Then I compared these fitted model by using analysis of variance (anova) to tell which variables are necessary to explain mileage change.

```
## Analysis of Variance Table
##
## Model 1: cty ~ trans
## Model 2: cty ~ trans + displ
## Model 3: cty ~ trans + displ + factor(cyl)
## Model 4: cty ~ trans + displ + factor(cyl) + factor(drv)
## Model 5: cty ~ trans + displ + factor(cyl) + factor(drv) + factor(fl)
## Model 6: cty ~ trans + displ + factor(cyl) + factor(drv) + factor(fl) +
##     factor(class)
## Model 7: cty ~ trans + displ + factor(cyl) + factor(drv) + factor(fl) +
##     factor(class) + factor(year)
##   Res.Df    RSS Df Sum of Sq        F    Pr(>F)
## 1    232 3841.7
## 2    231 1499.9  1   2341.82 976.9626 < 2.2e-16 ***
## 3    228 1340.0  3    159.94  22.2416 1.394e-12 ***
## 4    226 1043.0  2    296.99  61.9484 < 2.2e-16 ***
## 5    222  676.5  4    366.44  38.2183 < 2.2e-16 ***
## 6    216  567.4  6    109.12   7.5873 2.137e-07 ***
## 7    215  515.4  1     52.05  21.7145 5.546e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

## Warning: not plotting observations with leverage one:
##   107
```



As the analysis of variance shows, it can be said that fit7c (contain 7 variables: trans,displ,cyl,drv,fl,class and year) is the most representative fitted model.

From the Q-Q plot, there are several outliers (row no. 100,107,123,222) These outliers are removed and fit7c is calculated again.

```
##             (Intercept)            transmanual                  displ
##                  32.508                  0.260                 -1.151
##            factor(cyl)5            factor(cyl)6            factor(cyl)8
##                  -1.292                 -2.037                 -2.680
##            factor(drv)f            factor(drv)r             factor(fl)e
##                   1.505                  0.615                -11.642
##             factor(fl)p             factor(fl)r    factor(class)compact
##                  -8.827                 -7.776                 -2.813
##    factor(class)midsize    factor(class)minivan     factor(class)pickup
##                  -3.039                 -4.856                 -4.772
## factor(class)subcompact        factor(class)suv        factor(year)2008
##                  -2.976                 -4.467                  1.198
```

Coeffissient of transmanual was 0.260. It suggests that Manual transmission is better for MPG by 0.260 miles per gallon.

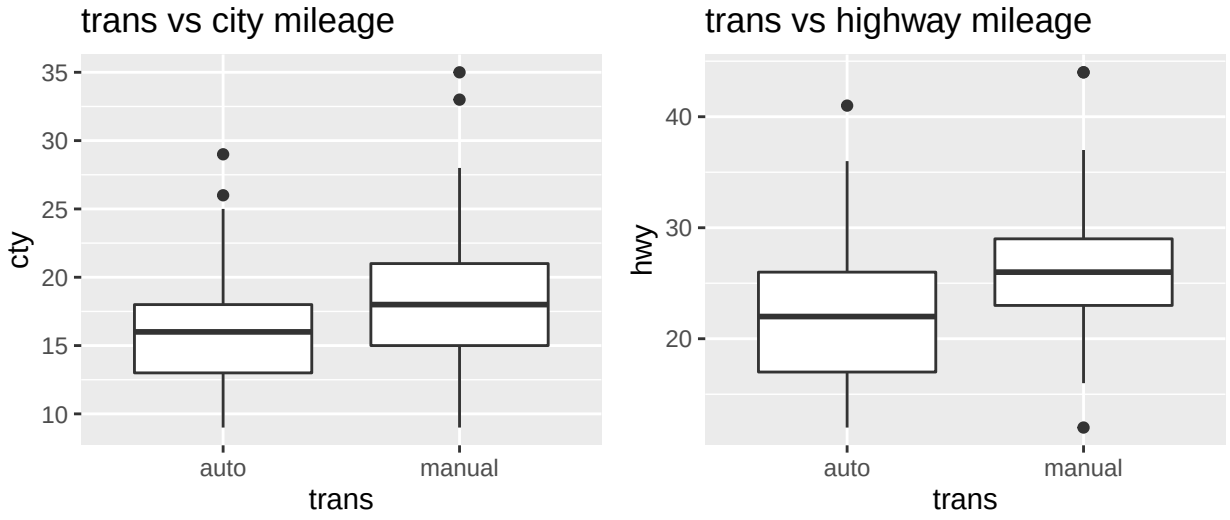
Conclusion

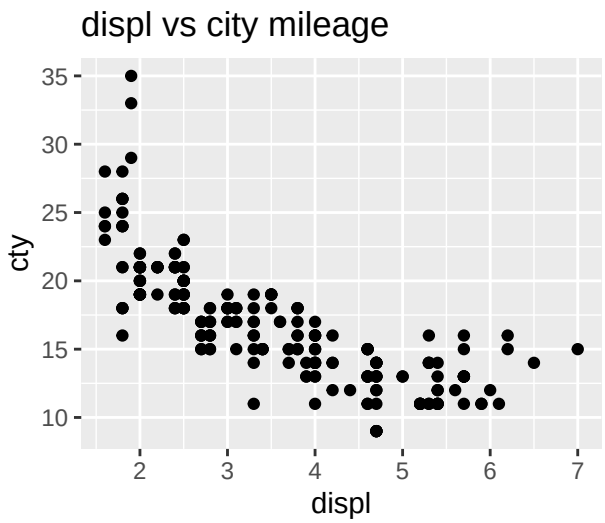
I conducted exploratory analysis and showed that manual transmission is better for MPG. Then I fitted the models to infer how much manual transmission is better than automatic transmission. As the model showed, manual transmission is better for MPG by 0.260 miles per gallon in city.

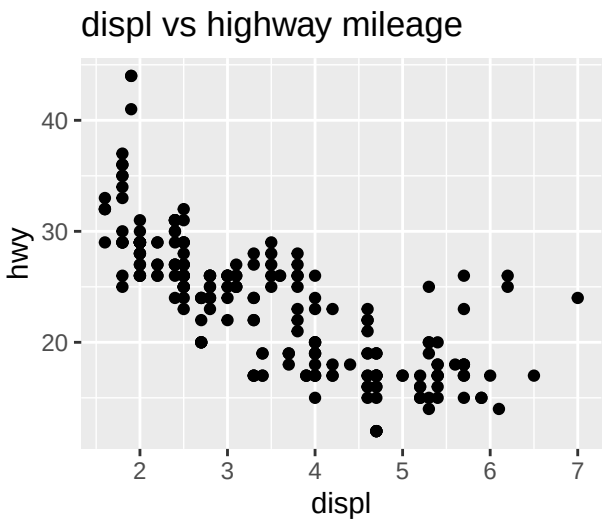
It is necessary to analyze how much manual transmission is better in highway in next study.

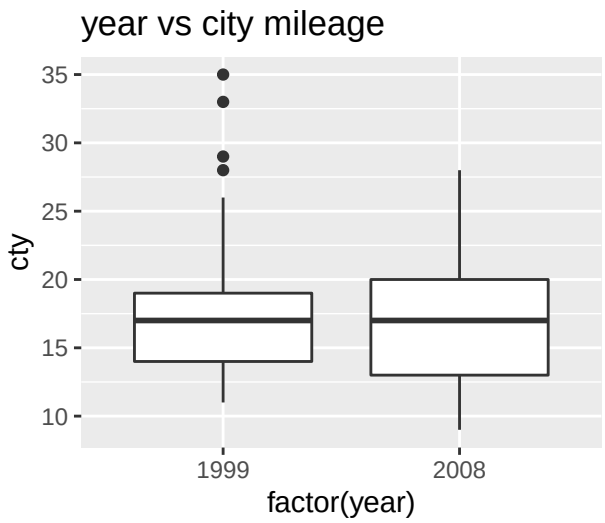
Appendices

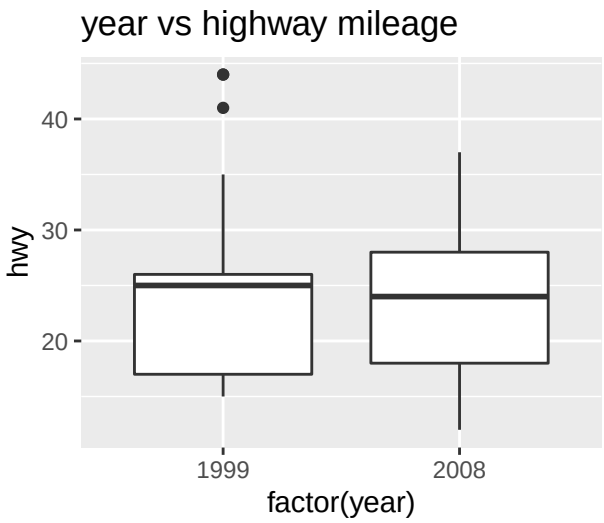
compare hwy,cty vs other valuables in graph

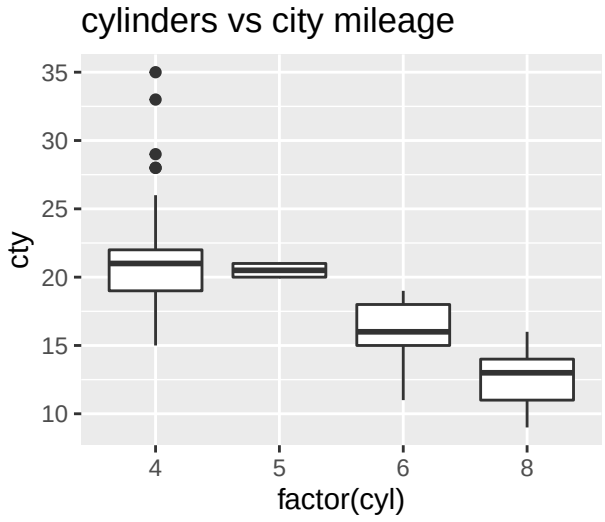


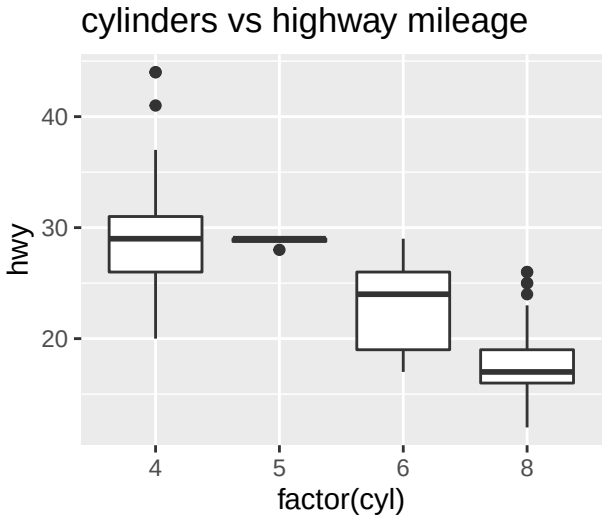

displ vs city mileage
cty
displ

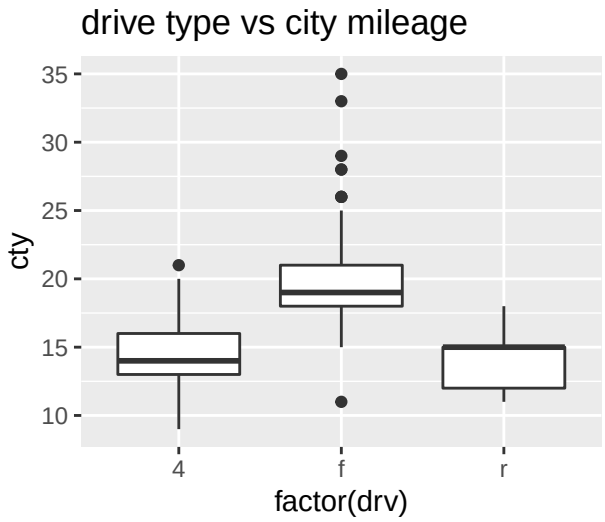

displ vs highway mileage
hwy
displ

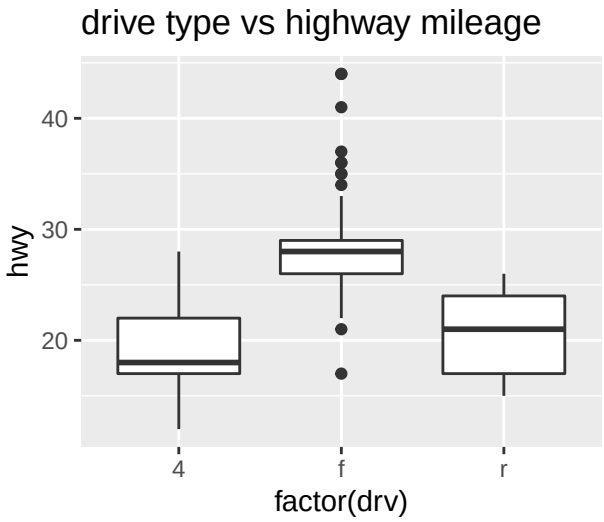

year vs city mileage
cty
factor(year)

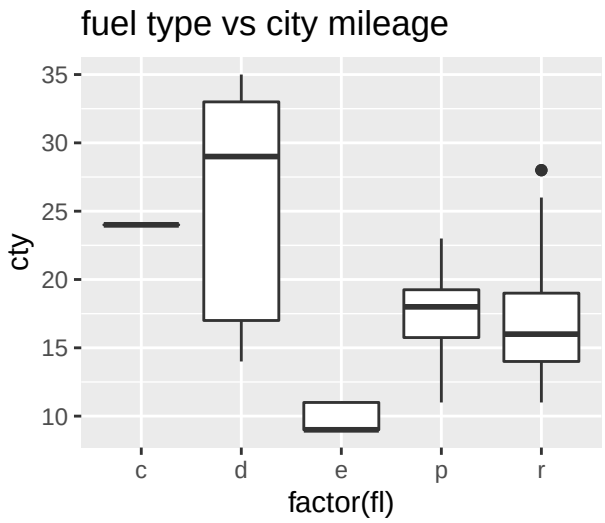

year vs highway mileage
hwy
factor(year)

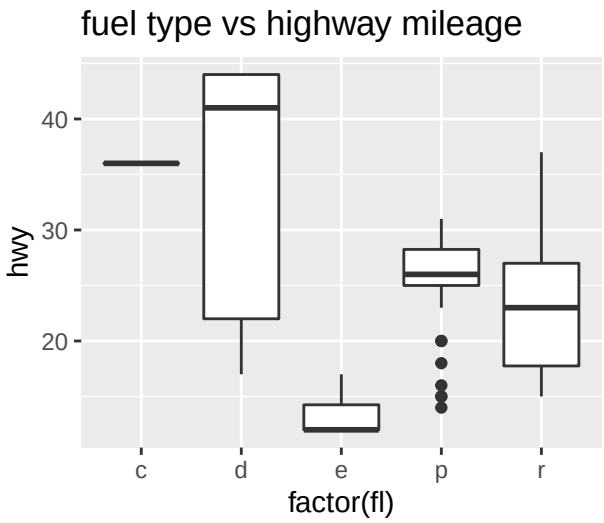

cylinders vs city mileage
cty
factor(cyl)

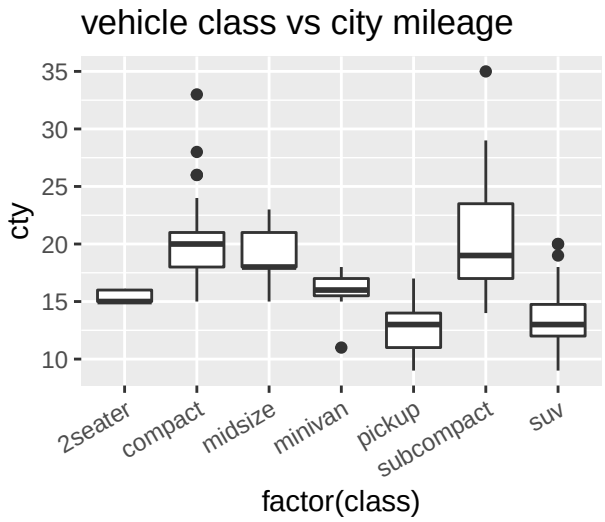

cylinders vs highway mileage
hwy
factor(cyl)

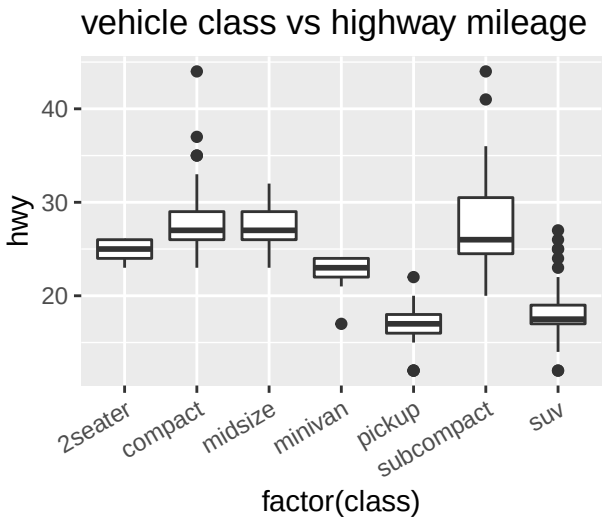

drive type vs city mileage
cty
35
30
25
20
15
10
4
f
r
factor(drv)


drive type vs highway mileage
hwy
40
30
20
4
f
r
factor(drv)


fuel type vs city mileage
cty
35
30
25
20
15
10
c
d
e
p
r
factor(fl)


fuel type vs highway mileage
hwy
40
30
20
c
d
e
p
r
factor(fl)


vehicle class vs city mileage
cty
35
30
25
20
15
10
2seater
compact
midsize
minivan
pickup
subcompact
suv
factor(class)


vehicle class vs highway mileage
hwy
40
30
20
2seater
compact
midsize
minivan
pickup
subcompact
suv
factor(class)