Homework 1: Solution

TA: Xiaomeng Yan (xiaomengyan@stat.tamu.edu)

8/29/2021

Problem 1

Load the file car.csv into RStudio as a data frame which represents more than most want to know about 2018 cars courtesy of the US Environmental Protection Agency. We will use it to make some comparisons of fuel efficiency (FE) across car companies. Some of the cars listed are electric so cannot directly be compared with gas cars so we will first clean up the spreadsheet. The columns of interest are

- Column B = 2 = “Vehicle Manufacturer Name”,
- AT = 46 = “RND ADJ FE”,
- V = 22 = “Weight”
- AK = 37 = “Test Fuel Type Description”

a. Make a new data frame by extracting the columns that are most informative.

Suggestion:

```
car <- read.csv("car.csv")
carReduce <- car[,c(2,22,37,46)]
```

Solution:

```
data <- read.csv("car.csv")
carReduce<-data[,c(2,22,37,46)]
```

If you like you can also eliminate row 3017, since Mercedes has suspiciously high FE.

b. You might want to rename columns with something easier such as

```
colnames(carReduce)<-c("Maker","Weight","Type","FE")
```

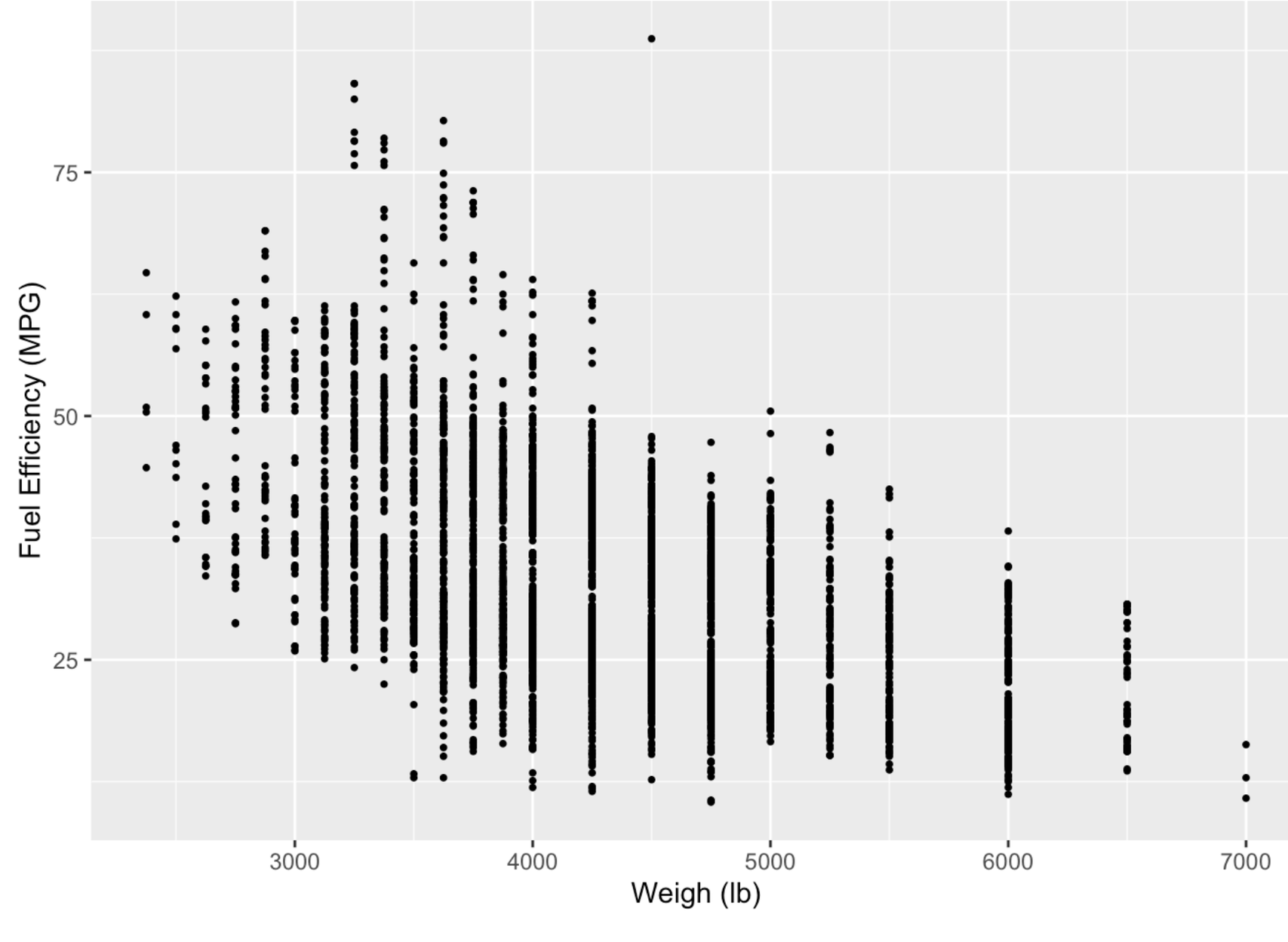
c. Make a new data frame by extracting only those car models with gas engines:

```
carReduceGas<-subset(carReduce,Type =="Tier 2 Cert Gasoline")
carReduceGas <- carReduceGas[-3017,]
```

d. Plot (Weight, FE) to get an idea of the effect of Weight on FE. Is the plot informative? (Meh)

Solution:

```
library(ggplot2)
ggplot(data = carReduceGas,aes(x = Weight, y = FE))+geom_point(size = 0.7)+xlab("Weigh (lb)")+ylab("Fuel Efficien
cy (MPG)")
```



- There is a trend between Weight and Fuel Efficienty. In general, heavier vehicles have lower fuel efficiency.

e. Use the summary(FE) command to get an idea of overall FE; do this also for Weight.

Solution:

```
summary(carReduceGas$FE)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   10.40   24.90   31.50   33.32   40.33   88.70       1
```

```
summary(carReduceGas$Weight)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    2375    3625    4250    4222    4750    7000
```

f. Using the subset command, we can extract from the data frame carReduceGas to compare 4 companies Honda, Toyota, GM, Ford (FOMOCO) the combined FE per company. For instance, for Honda we write

```
honda<-subset(CarReduceGas,Maker =="Honda")
```

to get the data frame corresponding only to Maker==“Honda”. This keeps all the columns but only if Make==“Honda”. Then honda$FE gives the column Honda cars. Do this for the 4 car companies and compare the 4 mean FE’s or use the summary() command.

Solution:

```
honda<-subset(carReduceGas,Maker =="Honda")
toyota<-subset(carReduceGas,Maker =="Toyota")
gm<-subset(carReduceGas,Maker =="GM")
ford<-subset(carReduceGas,Maker =="FOMOCO")
knitr::kable(data.frame(Maker = c("Honda","Toyota","GM","FOMOCO"),Mean_FE = c(mean(honda$FE),mean(toyota$FE),mean
(gm$FE),mean(ford$FE))),align = "c")
```

Maker	Mean_FE
Honda	39.35652
Toyota	38.38131
GM	31.32156
FOMOCO	30.45668

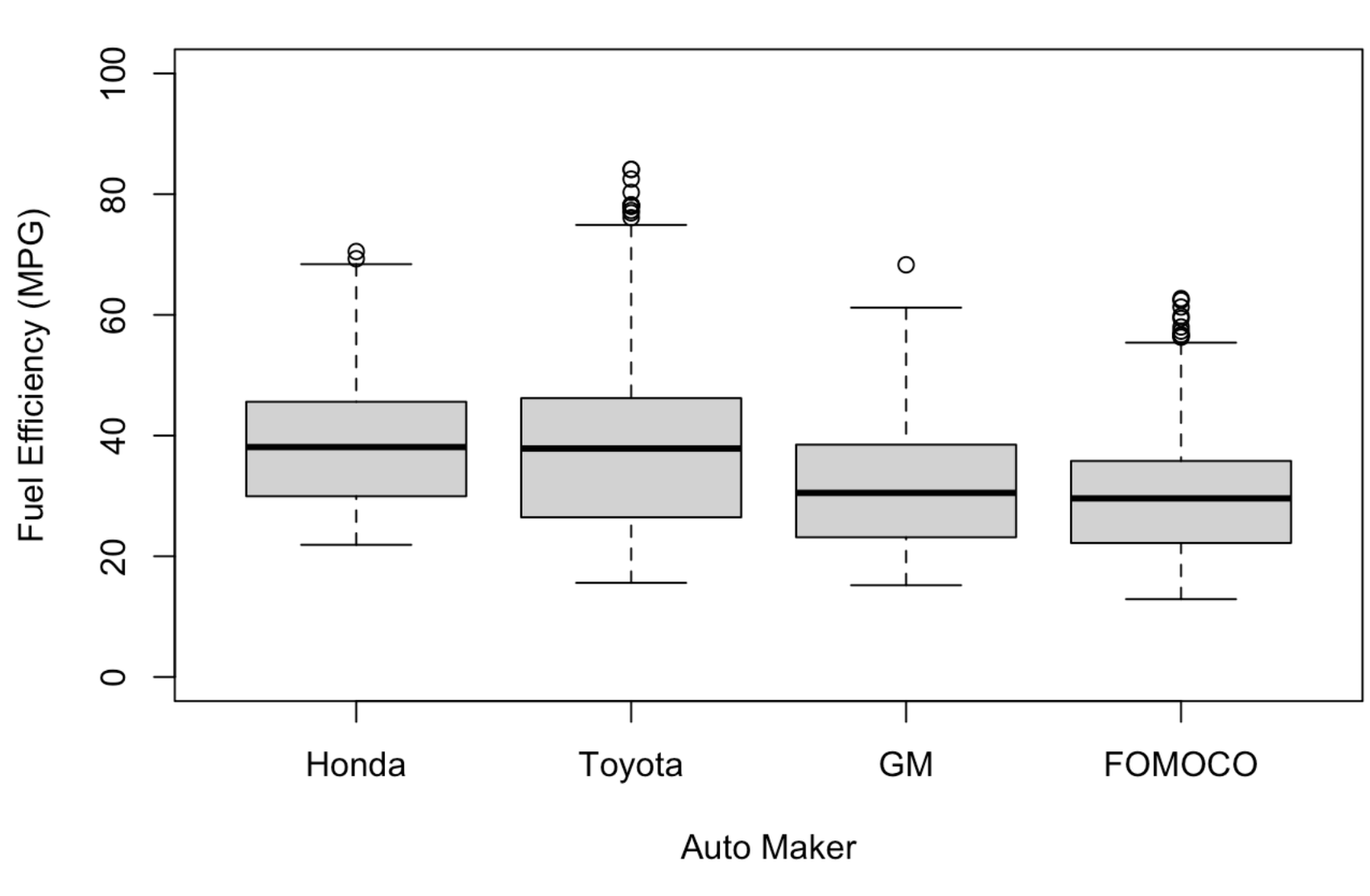
g. Make comparative boxplots for the fuel effciency of the 4 makers, for instance by typing,

```
boxplot(subset(carReduceGas2,Maker=="Honda")$FE,
subset(carReduceGas2,Maker=="Toyota")$FE,
subset(carReduceGas2,Maker=="GM")$FE,
subset(carReduceGas2,Maker=="FOMOCO")$FE)
```

and make a title or use xlab=“blah”. Note the use of quotes; fussy, fussy.

Solution:

```
boxplot(subset(carReduceGas,Maker=="Honda")$FE,
subset(carReduceGas,Maker=="Toyota")$FE,
subset(carReduceGas,Maker=="GM")$FE,
subset(carReduceGas,Maker=="FOMOCO")$FE,
names=c("Honda", "Toyota", "GM", "FOMOCO"),
ylim=c(0,100),
ylab="Fuel Efficiency (MPG)",
xlab = "Auto Maker")
```

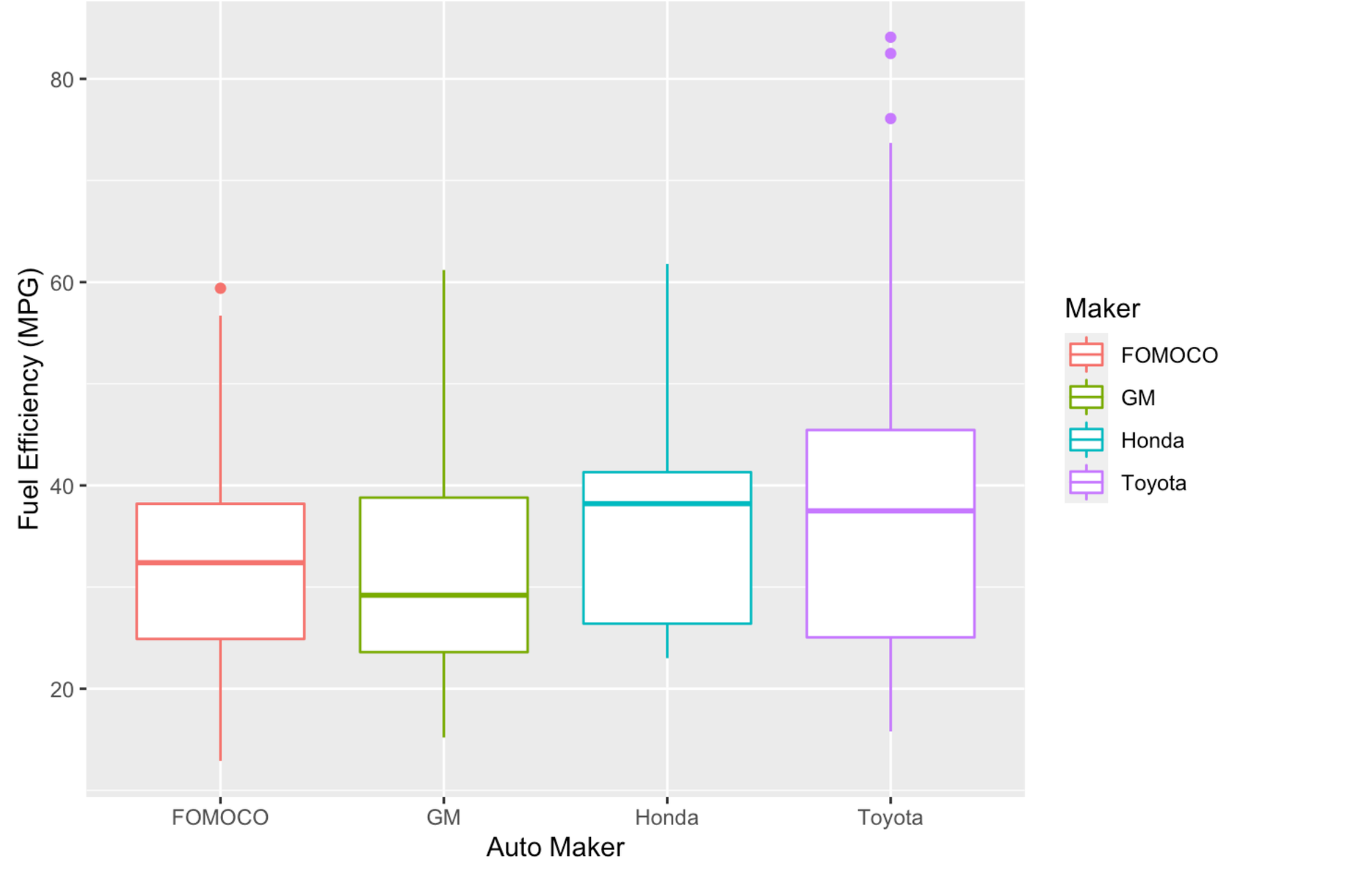


Suggestion: ggplot2 is a data visualization package for the statistical programming language R.

```
library(ggplot2)
carReduceGas_sub = carReduceGas[carReduceGas$Maker == c("Honda","Toyota","GM","FOMOCO"),]
```

```
## Warning in carReduceGas$Maker == c("Honda", "Toyota", "GM", "FOMOCO"): longer
## object length is not a multiple of shorter object length
```

```
ggplot(data = carReduceGas_sub,aes(x = Maker, y = FE,color = Maker))+geom_boxplot()+ ylab("Fuel Efficiency (MPG)"
)+xlab("Auto Maker")
```



h. After making your beautiful plot, what do you see that is interesting. Who is best positioned for the next oil crisis? If you have other favorite car makers (BMW? Tesla?), feel free to choose your favorites instead of the four I chose. Say something interesting.

Solution:

```
summary(honda$FE)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   21.90   29.95   38.10   39.36   45.60   70.50
```

```
summary(toyota$FE)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   15.60   26.48   37.85   38.38   46.15   84.10
```

```
summary(ford$FE)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   12.90   22.20   29.60   30.46   35.80   62.70
```

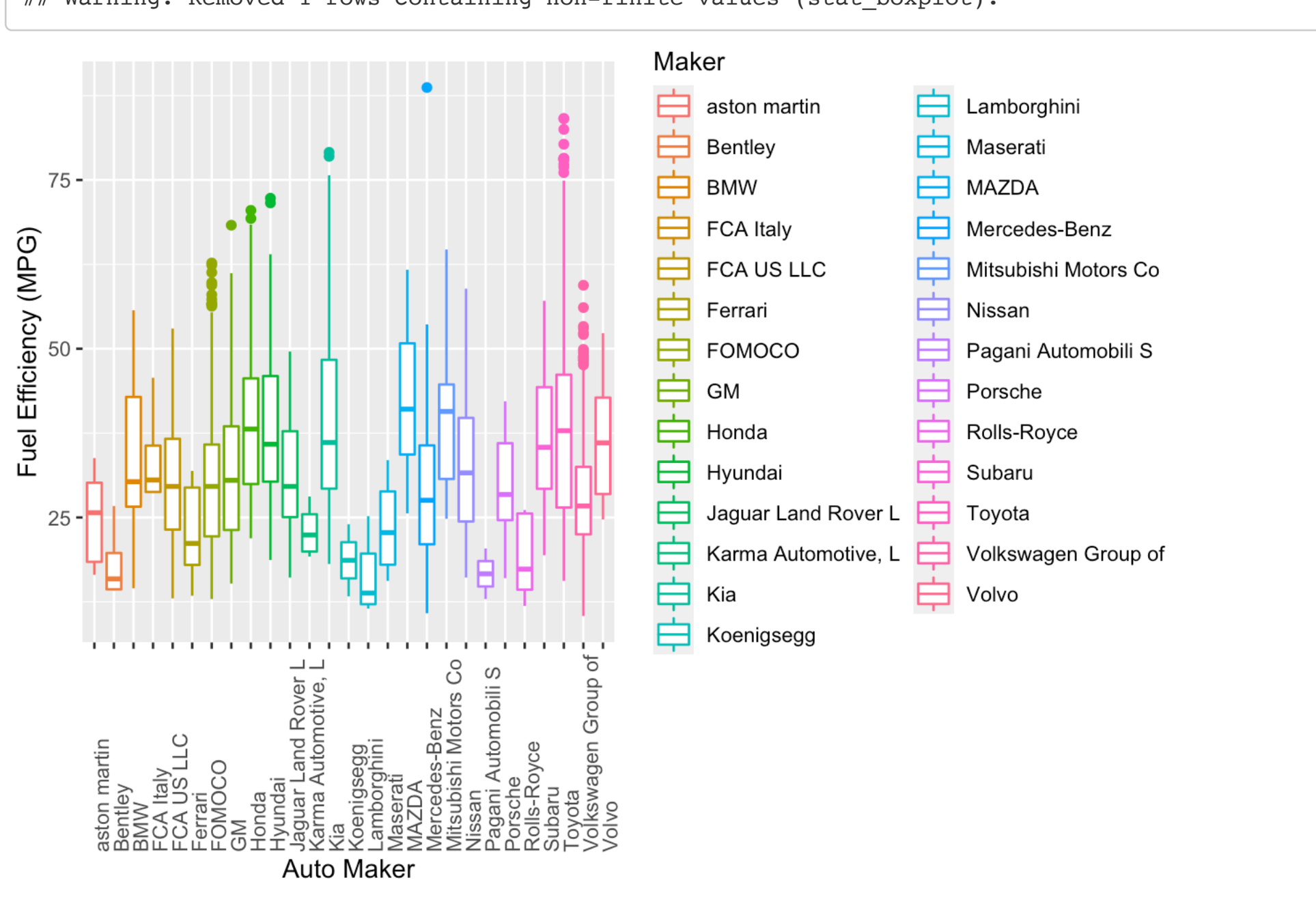
```
summary(gm$FE)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   15.20   23.15   30.50   31.32   38.50   68.30
```

- In general, both Toyota and Honda have higher fuel effiency. 50% of Honda have FE ranged betwwen 29.95 and 45.60, positioning itself well for a fuel crisis whereas 50% of Toyota have FE ranged between 26.48 and 46.15. Though the 3rd Qu. of Toyota is slightly higher than Honda, 1st Qu. of Honda is a lot higher than Toyota. These features positions Honda well for a fuel crisis.
- Compared with Toyota and Honda, Ford and GM have lower overall fuel efficiency, especially Ford. This makes sense since Ford focuses on Trucks, SUVs and Mustang.

```
library(ggplot2)
ggplot(data = carReduceGas,aes(x = Maker, y = FE,color = Maker))+geom_boxplot()+ ylab("Fuel Efficiency (MPG)")+xl
ab("Auto Maker")+theme(axis.title.x = )+theme(axis.text.x = element_text(angle = 90, hjust = 0))
```

```
## Warning: Removed 1 rows containing non-finite values (stat_boxplot).
```



Problem 2

(The Monty Hall problem) Suppose there are three doors, labeled A, B, and C. A new car is behind one the the three doors, but you don’t know which. You select one of the doors, say door A. Thr host then opens one of doors B or C, as follows: If the car is behind Bm then they open C; if car is behind C, then they open B; if the car is behind A, then they open either B or C with probability 1/2 each. (In any case, the door opened by the host will not have the car behind it.) The host then gives you the option of whether sticking with your original door choice (i.e., A), or switching to the remaining unopened door (i.e., whichever of B or C the host did not open). You then win (i.e., get to keep the car) if and only if the car is behind your final door slection. (Source: Parade Magazine, “Ask Marilyn” column, September 9. 1990.) Suppose for definiteness that the host opens door B.

a. If you stick with your original choice (i.e., door A), conditional on the host having opened door B, then what is your probability of winning?

Solution:

Since a car is randomly selected to be behind door A,B or C, then $P(\text{car behind A}) = P(\text{car behind B}) = P(\text{car behind C}) = \frac{1}{3}$. If you choose door A, then

$$\begin{aligned} P(\text{open B}) &= P(\text{open B} \cap \text{car behind A}) + P(\text{open B} \cap \text{car behind B}) \\ &+ P(\text{open B} \cap \text{car behind A}) \\ &= P(\text{open B}|\text{car behind A})P(\text{car behind A}) + P(\text{open B}|\text{car behind B})P(\text{car behind B}) \\ &+ P(\text{open B}|\text{car behind C})P(\text{car behind C}) \\ &= 1/2 * 1/3 + 0 * 1/3 + 1 * 1/3 = 1/2 \end{aligned}$$

Follow the same formula, we also have $P(\text{open C}) = 1/2$. If you choose door A and do not change your choice when host opens door B, then the probability to win is

$$\begin{aligned} P(\text{car behind A}|\text{open B}) &= \frac{P(\text{car behind A} \cap \text{open B})}{P(\text{open B})} \\ &= \frac{P(\text{open B}|\text{ car behind A})P(\text{car behind A})}{P(\text{open B})} = \frac{1/2 * 1/3}{1/2} = 1/3 \end{aligned}$$

b. If you switch to the remaining door (i.e., door C), conditional on the host having opened door B, then what is your probability of winning?

Solution: Since conditional on car is behind C the probability of opening B is 1, thus if you change your choice to C when the hos opens B, the probability to win is

$$\begin{aligned} P(\text{car behind C}|\text{open B}) &= \frac{P(\text{car behind C} \cap \text{open B})}{P(\text{open B})} \\ &= \frac{P(\text{open B}|\text{ car behind C})P(\text{car behind C})}{P(\text{open B})} = \frac{1 * 1/3}{1/2} = 2/3 \end{aligned}$$

c. Do you find the result of part (a) and part (b) surprising? How could you desin a physical experiment to verify the result?

Solution:

To do the experiment, you can hide a key in one of three small cups. Then let one person to point out one of them in which he believes the key is hidden. Then you reveal one of the cups without the key and ask the person whether he wants to change his mind. Repeat the above steps for several times and count the fraction of the time they win with or without changing their first choice.

Problem 3

Consider the simple ALOHA network model, run for two stages with $X_0 = 2$. Say we know that there have been 2 transmission attempts (regardless of whether they are successful or not).

a. Find $P(X_1 = 2 \text{ or } X_2 = 2)$.

Solution:

$$P(X_1 = 2 \text{ or } X_2 = 2|2 \text{ total attempts}) = \frac{P((X_1 = 2 \cup X_2 = 2) \cap 2 \text{ total attempts})}{P(2 \text{ total attempts})}$$

The event “$(X_1 = 2 \cup X_2 = 2) \cap 2$ total attempts” happens only if both attempts happen in epoch 1 or epoch 2, which has the probability $2p^2(1-p)^2$.

The event in the denominator **2 total transmission attempts** includes three subevents:

- both attempts occurred during stage 1. This event indicates the proccess that: both nodes try to send the message at stage 1, but they collide and two nodes remain active. In stage 2, they don’t try to send the messsage. Therefore the probability is: $(1-p)^2p^2$
- both attempts occurred during stage 2. This event means that both nodes don’t try to send the message at stage 1 and remain active at the begaining of the stage 2. Finally they try to send the message at stage 2. Therefore the probability is: $(1-p)^2p^2$
- one attempt occurred during stage 1 and one attempt occurred during stage 2. (Let’s assume the first node (A) tries to send the message at stage 1.) This event that node A send the message at stage 1 and node B doesn’t try to send the message at stage 1, hence the node A becomes inactive (no message stored) whereas node B remain active. At stage 2, A with probability q to get a new information. There are three possibilites: a.If A get a new message, A with probability p sends the message, and under this assumption, B can not send a message.
    b. If A get a message and A doesn’t send the message at stage 2, B tries to send a message.
    c. If A doesn’t get a new message, then B tries to send the message. Therefore, the probability is: $p(1-p)[qp(1-p) + q(1-p)p + (1-q)p]$. We will get the same probability if we assume the second node tries to send the message at stage 1 which is $p(1-p)[qp(1-p) + q(1-p)p + (1-q)p]$.

Therefore, the solution is

$$\frac{2p^2(1-p)^2}{2p^2(1-p)^2 + 2p(1-p)[qp(1-p) + q(1-p)p + (1-q)p]} = \frac{p(1-p)}{p(1-p) + 2qp(1-p) + (1-q)p}$$

b. Find the probability that at least one of those attempts occurred during stage 2. Give you analytical answer for general p and q.

Solution:

$$\begin{aligned} &P(\text{At least one of two attempts occurred during stage 2}|\text{2 total transmission attempts}) \\ &= 1 - P(\text{no attempts occurred during stage 2}|\text{2 total transmission attempts}) \\ &= 1 - \frac{P(\text{both attempts occurred during stage 1})}{P(\text{2 total transmission attempts})} \end{aligned}$$

The event in the numerator **both attempts occurred during stage 1** indicates the proccess that: both nodes try to send the message at stage 1, but they collide and two nodes remain active. In stage 2, they don’t try to send the messsage. Therefore the probability is: $p^2(1-p)^2$

Therefore,

$$\begin{aligned} &P(\text{At least one of two attempts occurred during stage 2}|\text{2 total transmission attempts}) \\ &= 1 - \frac{p^2(1-p)^2}{2p^2(1-p)^2 + 2p(1-p)[2qp(1-p) + (1-q)p]} \end{aligned}$$

Problem 4

Consider a three-node version of the ALOHA network example, with all nodes active at time 0. One of the users tells us at the end of epoch 1 that her node was involved in a collision during that epoch. (We have no information from the other two users.) What is the probability that all three nodes were involved in that collision? How will the probability change if none of the nodes is active at time 0?

- What is the probability that all three nodes were involved in that collision?

Solution: This problem can be interpreted as given a collision happens during epoch 1 (at least two nodes attempt to send the message), what is the probability that all three nodes attempt to send the message. The probability can be expressed as,

$$\begin{aligned} &P(\text{All three nodes attempt to send the message}|\text{At least two nodes attempt to send the message}) \\ &= \frac{P(\text{All three nodes attempt to send the message})}{P(\text{At least two nodes attempt to send the message})} \\ &= \frac{p^3}{\binom{3}{2}p^2(1-p) + p^3} \\ &= \frac{p}{3-2p} \end{aligned}$$

- How will the probability change if none of the nodes is active at time 0?

Solution: This problem is more complicated than the previous one because we have to take the new message receiving information into consideration.

$$\begin{aligned} &P(\text{All three nodes attempt to send the message}|\text{At least two nodes attempt to send the message}) \\ &= \frac{P(\text{All three nodes attempt to send the message})}{P(\text{At least two nodes attempt to send the message})} \end{aligned}$$

The event in the numerator **All three nodes attempt to send the message** can happen only if **all three nodes receive new message** and **they attempt to send these messages**. The probability is : q^3p^3.

The even in the denominator **At least two nodes attempt to send the message** includes three subevent:

a. Only two nodes receive new message and they attempt to send them. The probability is: $\binom{3}{2}q^2(1-q)p^2$.
b. All three nodes receive new message and only two of them attempt to send them. The probability is: $q^3\binom{3}{2}p^2(1-p)$.
c. All three nodes receive new message and all of them attempt to send them (the same as numerator). The probability is: q^3p^3.

Therefore,

$$\begin{aligned} &P(\text{All three nodes attempt to send the message}|\text{At least two nodes attempt to send the message}) \\ &= \frac{q^3p^3}{\binom{3}{2}q^2(1-q)p^2 + q^3\binom{3}{2}p^2(1-p) + q^3p^3} = \frac{pq}{3-2pq} \end{aligned}$$

Rubric

- The total points: 100
- Problem 1: 22
- Problem 2: 28
- Problem 3: 25
- Problem 4: 25

Problem 1 (22)

- a. 2 points
- b. 2 points
- c. 2 points
- d. 4 points
    - 2 points for plot
    - 2 for analysis of the plot
- e. 2 points
- f. 2 points
- g. 4 points
- h. 4 points

Problem 2 (28)

- a. 10 points
- b. 10 points
- c. 8 points

Problem 3 (25)

- a. 20 points
    - The numerator: 5 points
    - both attempts during stage 1: 5 points
    - both attempts during stage 2: 5 points
    - one in stage 1 one in stage 2: 5 points
- b. The formula of the conditional probability: 5 points.

Problem 4 (25)

- a. 10 points
- b. 15 points.