Class Survey Exercise

Eli Cook

Source files & load and clean data

```
datadir <- "C:/Users/ucg8nb/OneDrive - University of Virginia/SYS 4021/Data"
setwd(datadir)

class.data <- read.csv("Class_Survey_Data.csv")
summary(class.data)
```

```
##   Timestamp         Which.row.are.you.sitting.in. How.tall.are.you...inches.
##  Length:72          Min.   : 0.000                Min.   :56.00
##  Class :character   1st Qu.: 4.000                1st Qu.:64.75
##  Mode  :character   Median : 5.000                Median :67.60
##                     Mean   : 5.583                Mean   :68.00
##                     3rd Qu.: 7.000                3rd Qu.:72.00
##                     Max.   :10.000                Max.   :78.00
##
##  What.is.your.sex.  Are.you.from.Virginia. Do.you.want.to.go.to.grad.school.
##  Length:72          Length:72              Length:72
##  Class :character   Class :character       Class :character
##  Mode  :character   Mode  :character       Mode  :character
##
##
##
##
##  Do.you.want.to.be.a.consultant.
##  Length:72
##  Class :character
##  Mode  :character
##
##
##
##
##  How.fast.do.you.think.you.can.you.run.a.mile...seconds.
##  Min.   :  15.0
##  1st Qu.: 420.0
##  Median : 480.0
##  Mean   : 561.2
##  3rd Qu.: 540.0
##  Max.   :6000.0
##  NA’s   :1
##  Do.you.play.video.games.
##  Length:72
##  Class :character
##  Mode  :character
##
##
##
##
##  How.many.miles.do.you.live..during.the.school.year..from.Olsson.Hall...use.a.mapping.software.to.es
##  Min.   :0.1000
##  1st Qu.:0.7000
##  Median :1.0000
##  Mean   :0.9931
##  3rd Qu.:1.2000
##  Max.   :4.0000
##
##  How.do.you.get.to.Grounds. How.many.languages.do.you.speak.
##  Length:72                  Length:72
##  Class :character           Class :character
##  Mode  :character           Mode  :character
##
##
##
##
##  What.s.your.favorite.movie.genre.
##  Length:72
##  Class :character
##  Mode  :character
##
##
##
##
```

```
class.data <- class.data %>% rename(row = Which.row.are.you.sitting.in., height = How.tall.are.you...in
class.data.accurate.mile.times <- class.data %>% filter(mileTime > 223 & mileTime < 3600)
class.data.accurate.mile.times$videoGames
```

```
##  [1] "No"  "Yes" "No"  "No"  "Yes" "No"  "Yes" "No"  "Yes" "Yes" "No"  "No"
## [13] "No"  "Yes" "Yes" "No"  "No"  "Yes" "No"  "No"  "No"  "No"  "No"  "No"
## [25] "No"  "No"  "No"  "No"  "Yes" "No"  "No"  "Yes" "No"  "No"  "Yes" "Yes"
## [37] "Yes" "No"  "No"  "No"  "No"  "Yes" "Yes" "Yes" "No"  "Yes" "Yes" "No"
## [49] "No"  "No"  "Yes" "No"  "No"  "No"  "No"  "Yes" "Yes" "No"  "No"  "No"
## [61] "Yes" "No"  "No"  "Yes" "Yes" "No"  "No"
```

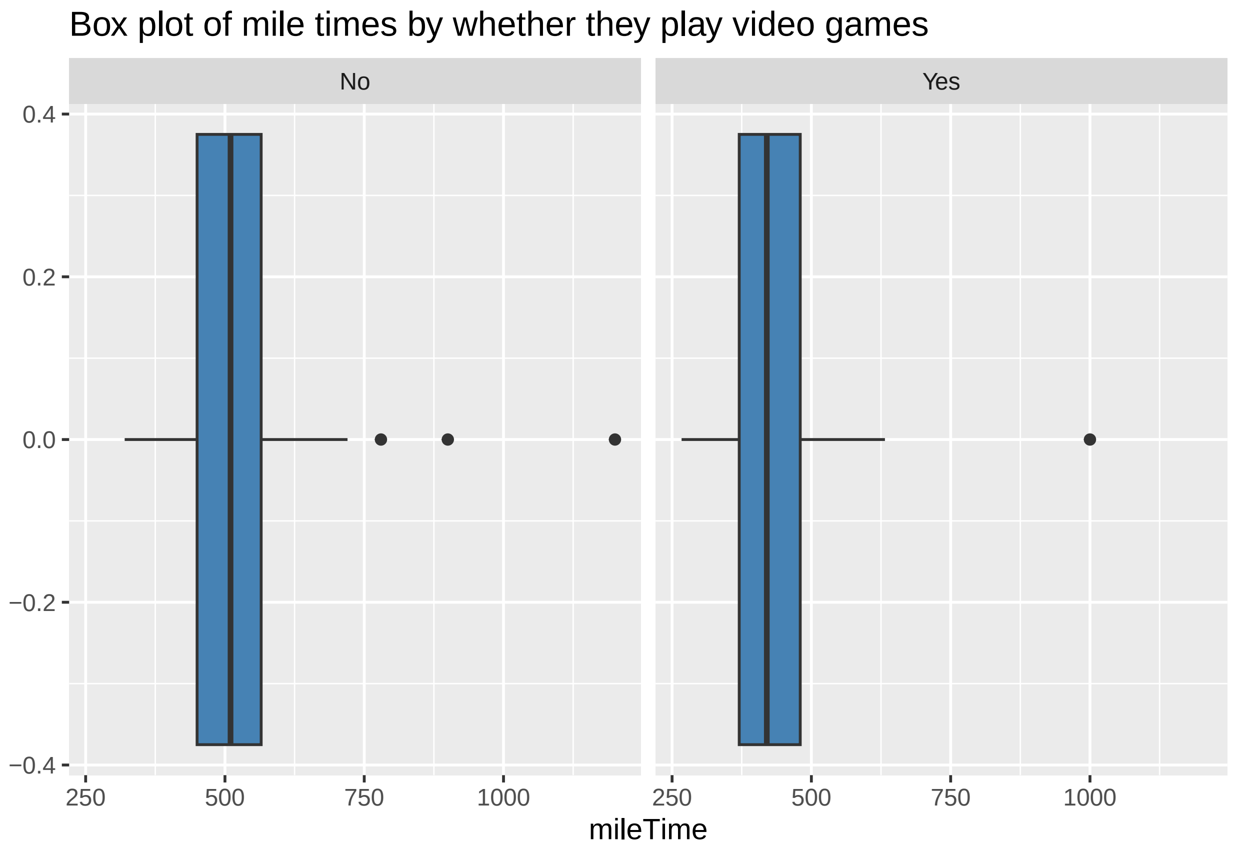
```
class.data.accurate.mile.times$videoGames <- as.factor(class.data.accurate.mile.times$videoGames)
class.data.accurate.mile.times$gender <- as.factor(class.data.accurate.mile.times$gender)
```

1. Formulate a hypothesis about this data.

I hypothesize that people who play video games run 1 mile slower

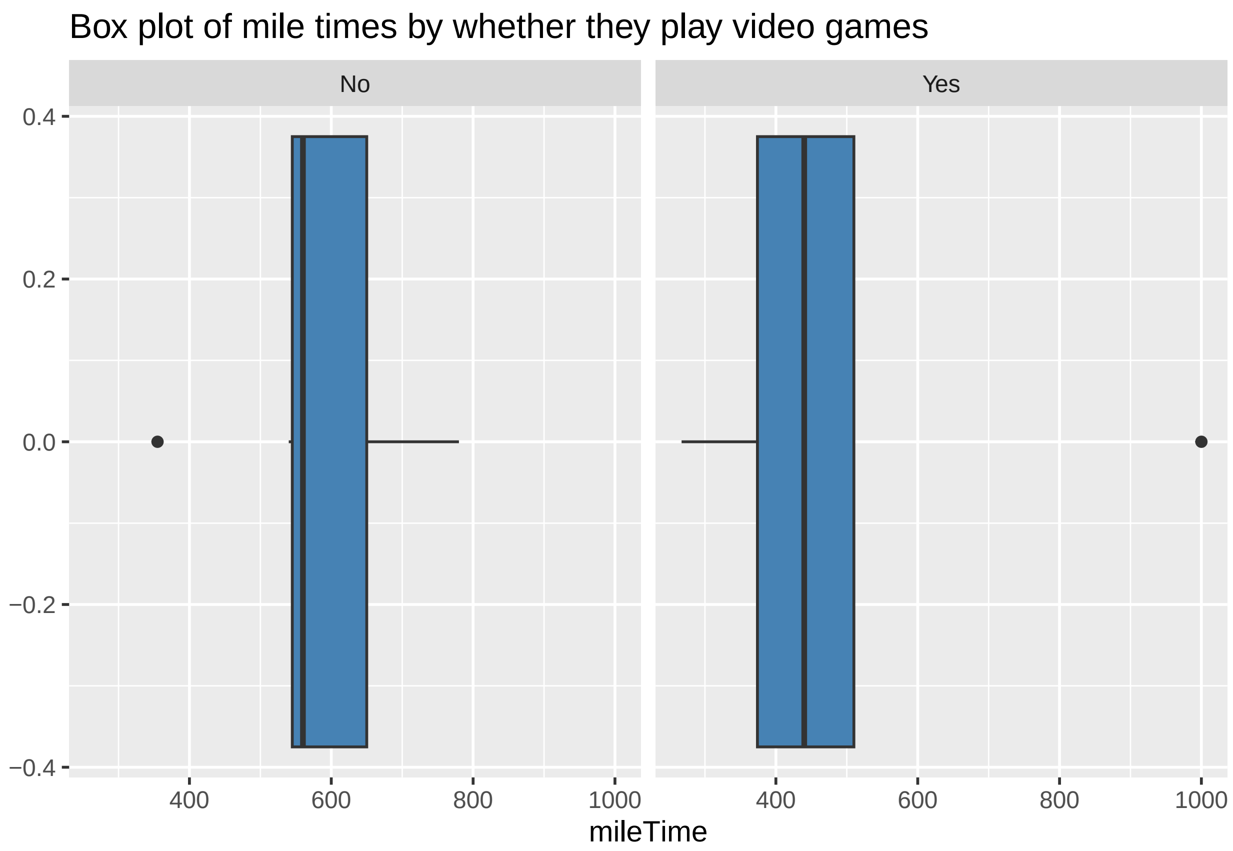
2. Now visualize your data to see if this hypothesis appears to be true. This hypthesis does apprear to be true

```
ggplot(class.data.accurate.mile.times, aes(mileTime)) +
  geom_boxplot(fill = "steelblue") +
  ggtitle("Box plot of mile times by whether they play video games") +
  facet_wrap(~videoGames)
```



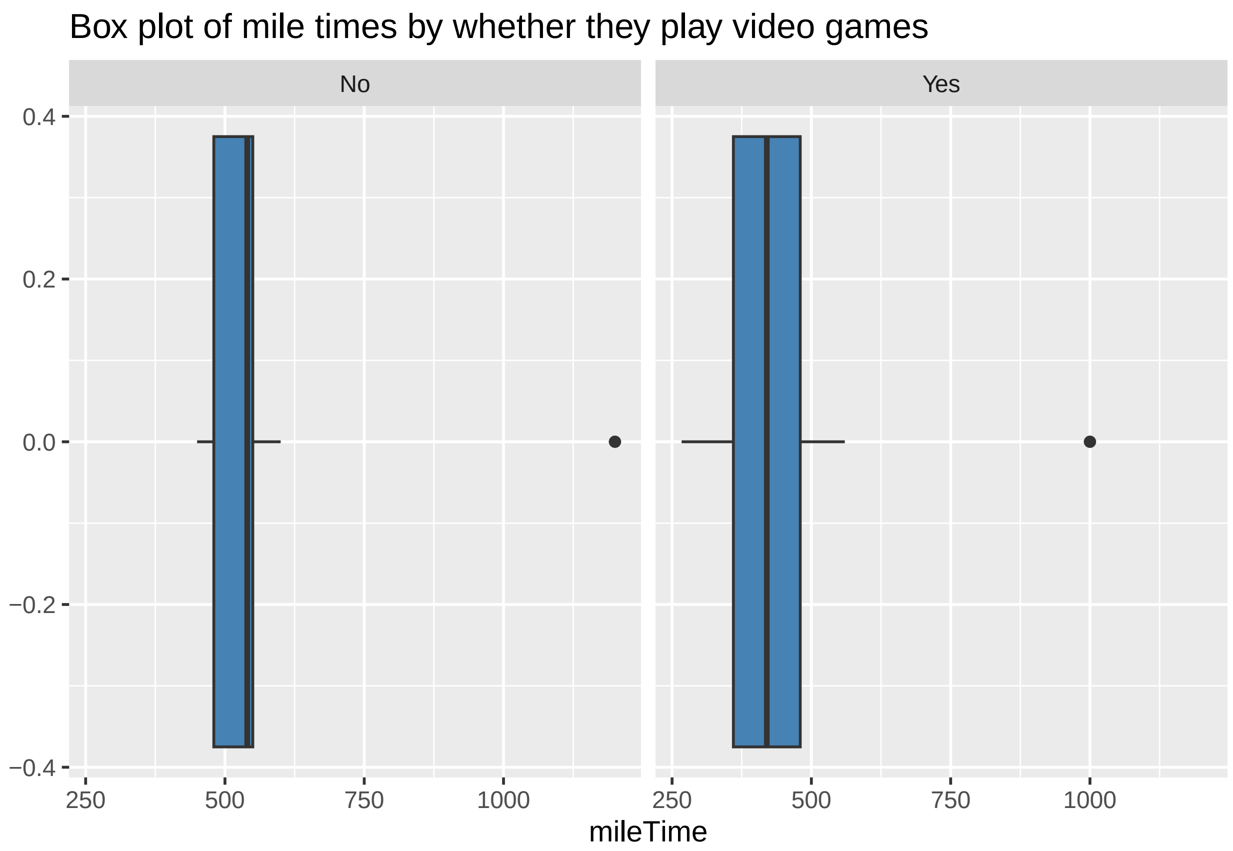
3. What if you used convenience sampling and only surveyed the first 3 rows? Would your inferences change? If we only sample the first three rows, we still find that people who play video games are faster, on average, then those who don't

```
selectedGroup <- class.data.accurate.mile.times %>% filter(row %in% c(0,1,2,3))
ggplot(selectedGroup, aes(mileTime)) +
  geom_boxplot(fill = "steelblue") +
  ggtitle("Box plot of mile times by whether they play video games") +
  facet_wrap(~videoGames)
```



4. Randomly sample 30 students. Do you come to the same conclusions as with the full dataset? While it is closer with this random sample, the conclusions don't change here either

```
set.seed(1)
selectedGroup <- class.data.accurate.mile.times %>% sample_n(30)
ggplot(selectedGroup, aes(mileTime)) +
  geom_boxplot(fill = "steelblue") +
  ggtitle("Box plot of mile times by whether they play video games") +
  facet_wrap(~videoGames)
```

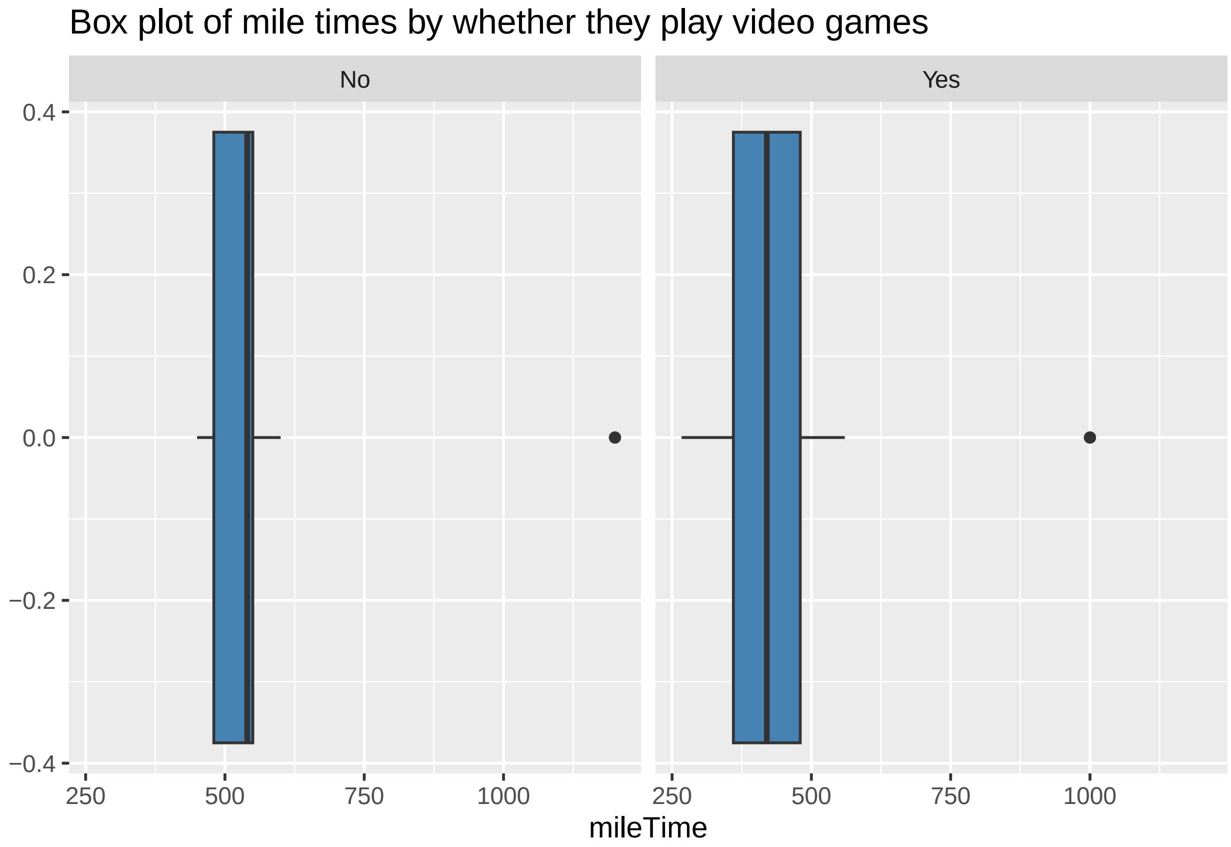


5. Randomly sample 15 male students and 15 female students. Do you come to the same conclusions as with the full dataset? Our conclusions don't change

```
class.data.accurate.mile.times$gender
```

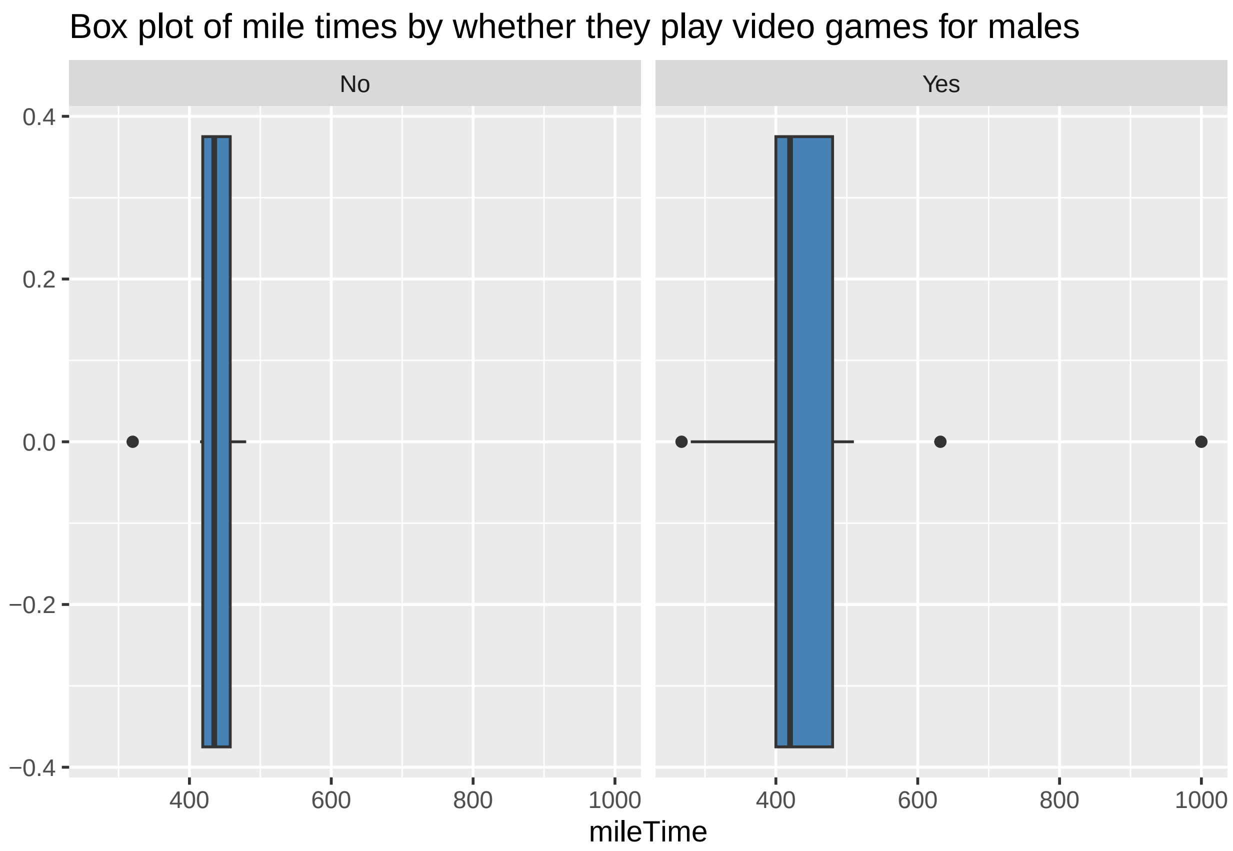
```
##  [1] Female Male   Female Female Male   Female Male   Female Male   Male  
## [11] Female Female Male   Female Female Female Female Female Female Female
## [21] Female Male   Female Female Female Female Female Female Male   Male  
## [31] Female Male   Male   Female Male   Male   Male   Male   Female Male  
## [41] Female Male   Male   Male   Male   Male   Male   Female Female Female
## [51] Male   Female Female Female Female Male   Male   Female Male   Female
## [61] Male   Female Female Male   Male   Female Female
## Levels: Female Male
```

```
maleGroup <- class.data.accurate.mile.times %>% filter(gender == "Male") %>% sample_n(15)
femaleGroup <- class.data.accurate.mile.times %>% filter(gender == "Female") %>% sample_n(15)
sampleGroup <- rbind(maleGroup, femaleGroup)
ggplot(selectedGroup, aes(mileTime)) +
  geom_boxplot(fill = "steelblue") +
  ggtitle("Box plot of mile times by whether they play video games") +
  facet_wrap(~videoGames)
```

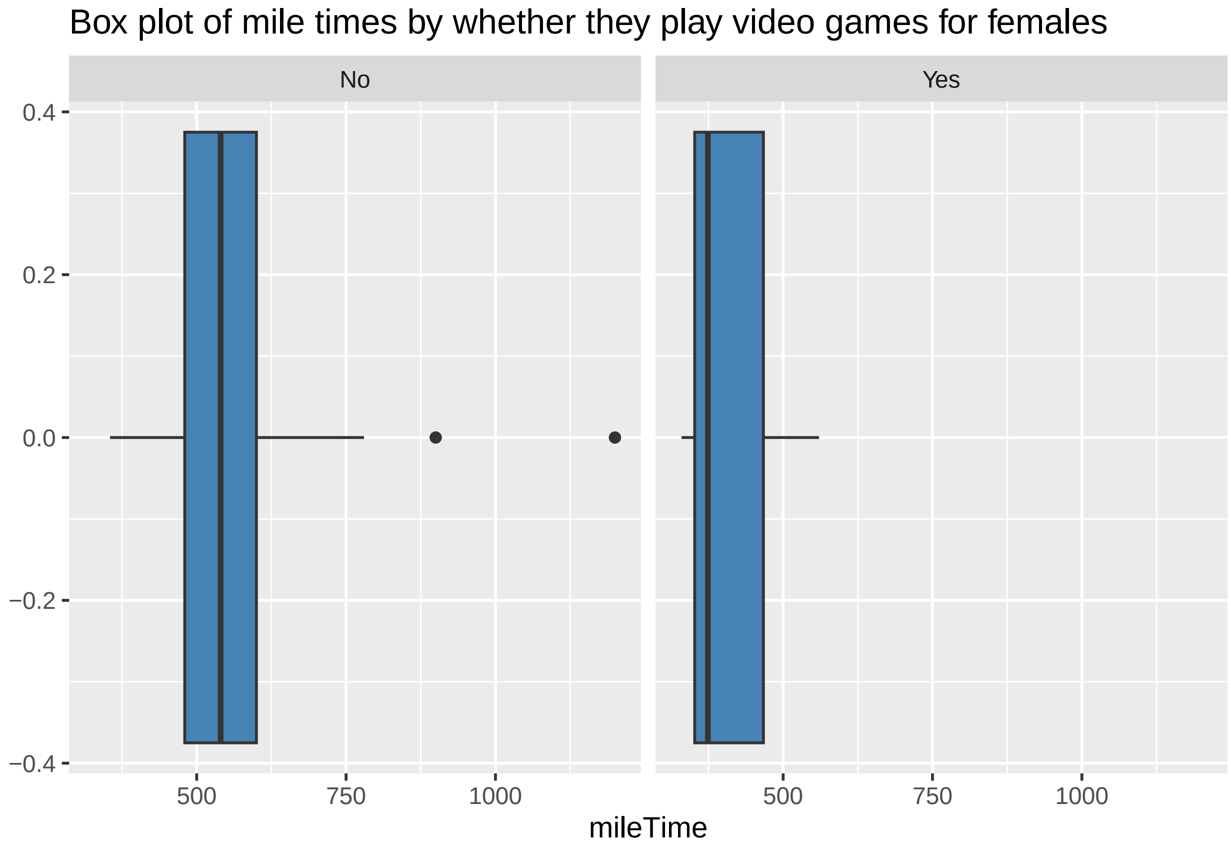


6. (Only answer if your hypothesis was not about Sex) Does your conclusion about your hypothesis differ when using the whole dataset vs. just males or just females? For males, mile times don't change much whether they play video games or not, the median time appears to be 420 for both, whereas for females playing video games it changes drastically with females who play video games running significantly faster mile times then those who don't.

```
maleGroup <- class.data.accurate.mile.times %>% filter(gender == "Male")
femaleGroup <- class.data.accurate.mile.times %>% filter(gender == "Female")
ggplot(maleGroup, aes(mileTime)) +
  geom_boxplot(fill = "steelblue") +
  ggtitle("Box plot of mile times by whether they play video games for males") +
  facet_wrap(~videoGames)
```



```r
ggplot(femaleGroup, aes(mileTime)) +
  geom_boxplot(fill = "steelblue") +
  ggtitle("Box plot of mile times by whether they play video games for females") +
  facet_wrap(~videoGames)
```


Box plot of mile times by whether they play video games for females
No
Yes
0.4
0.2
0.0
−0.2
−0.4
500
750
1000
500
750
1000
mileTime