BonusProblem_2

2023-04-23

Bonus Problem (5 points)

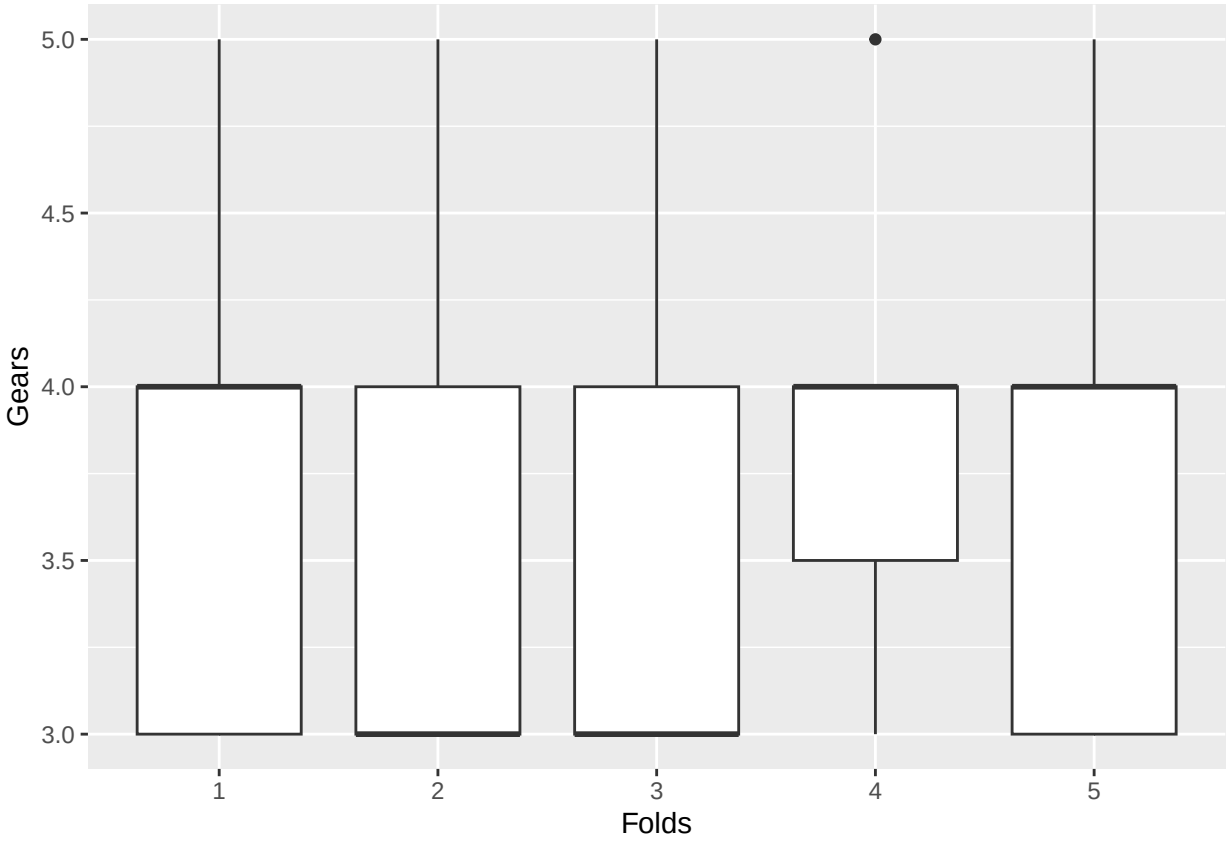
To understand just how much different subsets can differ, create a 5 fold partitioning of the cars data included in R (mtcars) and visualize the distribution of the gears variable across the folds. Rather than use the fancy trainControl methods for making the folds, create them directly so you actually can keep track of which data points are in which fold. This is not covered in the tutorial, but it is quick. Here is code to create 5 folds and a variable in the data frame that contains the fold index of each point. Use that resulting data frame to create your visualization.

```
library(ggplot2)
library(caret)
```

```
## Loading required package: lattice
```

```
mycars <- mtcars
mycars$folds <- 0
no_flds <- createFolds(1:nrow(mycars), k=5, list = TRUE)

for (i in 1:5) {
  mycars$folds[no_flds[[i]]] <- i
}
ggplot(mycars, aes(x = factor(folds), y = gear)) +
  geom_boxplot() +
  xlab("Folds") +
  ylab("Gears")
```



Through this box plot we can also visualize an outlier. In addition to this, distribution of each fold can help us to identify the variability in the number of gears across different folds. As in folds mean is shifting at the top and bottom which means that the data is skewed at the right a and left sometimes and most of the values are skwed to the right and left in these folds.