```r
## Cross-validation

## Coffee example  (Coffee Quality Institute, 2018) continued
coffee <- read.csv("csv/coffee_arabica.csv")
# 1 = wet, 0 otherwise
coffee$wet <-
  ifelse(coffee$Processing.Method == 'Washed / Wet', 1, 0)
# 1 = semi/dry, 0 otherwise
coffee$semi <-
  ifelse(coffee$Processing.Method == 'Semi-washed / Semi-pulped', 1, 0)
coffee$Processing.Method <- NULL

N <- nrow(coffee)

## Train and validation set split
set.seed(12345678)
trainInd <- sample(1:N, round(N * 0.8), replace = F)
trainSet <- coffee[trainInd,]
validSet <- coffee[-trainInd,]

# Calculate RMSE on three models each with different variables included
m1 <-
  lm(Flavor ~ wet + semi + Aroma + Aftertaste + Body, dat = trainSet)
pred1 <- predict(m1, newdata = validSet)
sqrt(mean((validSet$Flavor - pred1) ^ 2)) # RMSE
```

```
## [1] 0.1577479
```

```r
mean(abs(validSet$Flavor - pred1)) # MAE
```

```
## [1] 0.113643
```

```r
m2 <- lm(
  Flavor ~ wet + Aroma + Aftertaste +
    Body + Acidity + Balance + Sweetness + Uniformity + Moisture,
  dat = trainSet
)
pred2 <- predict(m2, newdata = validSet)
sqrt(mean((validSet$Flavor - pred2) ^ 2))
```

```
## [1] 0.1426565
```

```r
m3 <- lm(Flavor ~ Aroma + Aftertaste, dat = trainSet)
pred3 <- predict(m3, newdata = validSet)
sqrt(mean((validSet$Flavor - pred3) ^ 2))
```

```
## [1] 0.1615385
```

```r
# K fold cross validation
K <- 5
validSetSplits <- sample((1:N) %% K + 1)
RMSE1 <- c()
RMSE2 <- c()
RMSE3 <- c()
for (k in 1:K) {
  validSet <- coffee[validSetSplits == k,]
  trainSet <- coffee[validSetSplits != k,]

  m1 <-
    lm(Flavor ~ wet + semi + Aroma + Aftertaste + Body, dat = trainSet)
  pred1 <- predict(m1, newdata = validSet)
  RMSE1[k] <- sqrt(mean((validSet$Flavor - pred1) ^ 2))

  m2 <- lm(
    Flavor ~ wet + Aroma + Aftertaste +
      Body + Acidity + Balance + Sweetness + Uniformity + Moisture,
    dat = trainSet
  )
  pred2 <- predict(m2, newdata = validSet)
  RMSE2[k] <- sqrt(mean((validSet$Flavor - pred2) ^ 2))

  m3 <- lm(Flavor ~ Aroma + Aftertaste, dat = trainSet)
  pred3 <- predict(m3, newdata = validSet)
  RMSE3[k] <- sqrt(mean((validSet$Flavor - pred3) ^ 2))
}

RMSE1
```

```
## [1] 0.1479415 0.1653329 0.1556385 0.1656876 0.1482716
```

```r
RMSE2
```

```
## [1] 0.1427025 0.1525461 0.1478815 0.1620440 0.1384244
```

```r
RMSE3
```

```
## [1] 0.1513836 0.1667202 0.1616626 0.1675113 0.1532496
```

```r
mean(RMSE1)
```

```
## [1] 0.1565744
```

```r
mean(RMSE2)
```

```
## [1] 0.1487197
```

```r
mean(RMSE3)
```

```
## [1] 0.1601055
```