Removing invalid data in R

Cong Wang, Ph.D.

First, let's import the data. The name of the data in R is "mydata.csv."

```
setwd("/Users/congwang/Desktop/Real-world Data/sample data")
mydata.csv<-read.csv("1_Understanding_your_data.csv", header=TRUE)
```

1. Removing non-data rows

```
dim(mydata.csv) # before we drop the first two rows, let's check the size of the data. It should include 69 rows and 138 columns
```

```
## [1]  69 138
```

```
mydata.csv1<-mydata.csv[-c(1,2),] # the first two rows are removed from the data. I get it a different name -- mydata.csv1
dim(mydata.csv1) #now the data includes 67 rows and 138 columns
```

```
## [1]  67 138
```

2. Removing testing data

```
mydata.csv2<-subset(mydata.csv1, StartDate > "3/20/19 08:00") # I only keep responses with StartDate after 3/20/19 08:00. The new data is named mydata.csv2
dim(mydata.csv2) #mydata.csv2 includes 51 rows
```

```
## [1]  51 138
```

3. Removing incomplete responses

```
table(mydata.csv2$Progress) #running a contingency table
```

```
##
## 100   4
##  50   1
```

```
mydata.csv3<-subset(mydata.csv2, as.numeric(Progress) == 100) #keeping cases with Progress value of 100
dim(mydata.csv3) #mydata.csv3 includes 50 rows
```

```
## [1]  50 138
```

4. Removing uncommon responses

```
mydata.csv4<-subset(mydata.csv3, as.numeric(Duration..in.seconds.) > 900 & as.numeric(Duration..in.seconds.) < 3600) #keeping cases with response time greater than 900 seconds and less than 3600 seconds
dim(mydata.csv4) #mydata.csv4 includes 46 rows
```

```
## [1]  46 138
```

```
saveRDS(mydata.csv4, file="mydata.valid") # Saving the cleaned data as an R file called mydata.valid
```