Data Cleaning - Day 1

William F. Lamberti [1]

George Mason University

February 21, 2018

[1] MS Statistical Science and
PhD Student Computational Sciences and Informatics

Outline

Introduction - Course

- 1 hour talk (12:00 PM - 1:00 PM)
- 3 Sessions Goal: Introduce R concepts in Data Cleaning
- Assuming little to no experience in coding languages like C or Java
- Will reference `RGalleon.com` pages for additional information
- Will *not* go over the "why" of statistics (simply do not have the time)
- If you have any questions, please feel free to ask at any point

Introduction - Session Themes

- Day 1 - Partitioning and Loading Data
- Day 2 - Unique R Object Types
- Day 3 - Using Cleaned Data in ggplot2

Data Partitioning - Matrices

- Matrices can only handle one type
- Utilizes [,] to get parts of matrix

Data Partitioning - Example Code Matrices

```
1  #creating data
2
3  x<-c(1, 2, 3, 4, 6)
4  x
5
6  is.vector(x) #checking if a vector
7  is.data.frame(x) #checking if a data frame
8
9  x2<-as.matrix(x) #convert to data frame
10 is.matrix(x2)
11 x2
12 x2[,1:3]
```

Data Partitioning - Example Code Matrices

```
1  #creating character matrix
2
3  y<-matrix(nrow=2, data=c("1", "2", "3", "6"))
4  y
5
6  q<-matrix(nrow=2, data=c(100, 99))
7
8  #combine by column
9  new<-cbind(y, q)
10 new
```

Exercise

$$\begin{matrix} 3 & 4 & 22 \\ 7 & 100 & 678 \\ 432 & 1 & 76 \\ 0 & 14 & 22 \end{matrix}$$

- Create the above matrix
- Create a new object via partitioning the 3^{rd} and 4^{th} rows and 3^{rd} column
- You have 5 minutes

Exercise - Solution

```
1  #create matrix
2  x<-matrix(nrow=4, ncol=3, data=c(3, 7, 432,
       0, 4, 100, 1, 14, 22, 678, 76, 22))
3  x
4
5  #create new object
6  new<-x[3:4,3]
7  new<-as.matrix(new)
8
9  #create new column/row names
10 colnames(new)<-c("3rd")
11 rownames(new)<-c("3rd", "4th")
```

Data Partitioning - Data Frames

- Data frames can handle both numeric and character data in one object
- Vectors and matrices cannot
- Data frames are heavily used in ggplot2
- Data frames are more flexible and unclear

Data Partitioning - Example Code Data Frames

```
1  #creating data
2
3  x<-c(1, 2, 3, 4, 6)
4  x
5
6  is.vector(x) #checking if a vector
7  is.data.frame(x) #checking if a data frame
8
9  x2<-as.data.frame(x) #convert to data frame
10 is.data.frame(x2)
```

Object Type - Example Code Data Frames

```
1  #creating character matrix
2
3  y<-matrix(nrow=2, data=c("1", "2", "3", "6"))
4  y
5
6  q<-matrix(nrow=2, data=c(100, 99))
7
8  #combine by column
9  new<-cbind(y, q)
10 new
```

Object Type - Example Code Data Frames

```
#creating data frame

y2<-as.data.frame(y)
q2<-as.data.frame(q)

new2<-cbind(y2, q2)
new2

#dataframes can be unclear
sum(new2)
sum(new2[,3])

#dataframes are more flexible
colnames(new2)<-c("V1", "V2", "V3")

sum(new2$V3)
```

Loading Data - General Tips

- R can work with a variety of file types such as:
  - .csv
  - .txt
  - .xlsx
  - .xpt
- Some require packages
  - Excel and SAS files require different packages
- More info at `https://www.statmethods.net/input/importingdata.html`

Loading Data - Example: CSV

```
#creating csv

#create some data
x<- c(-0.6, -2.3, -0.4)

#saving as csv
write.csv(x, "mydata.csv")

#clear workspace
rm(list=ls())
ls()
```

Loading Data - Example: CSV

```
1  #loading .csv
2  y<-read.table("mydata.csv", header=TRUE, sep=
       ",")
3
4  y
5
6  as.matrix(unname(y))
7
8  #converting y to be like x
9  y2<-y[,-1]
10 y2
```

Exercise

- Use

```
#load economics data from ggplot2
install.packages('ggplot2')
library(ggplot2)
data(economics)
mydf<-as.data.frame(economics)

#extract year
dates<-as.character(mydf[,1])
mon<-substring(dates, 6, 7)
```

- Add a month column
- Save mydf as a CSV
- Bonus: Save mydf as as a DTA file
- You have 15 minutes

Exercise - Solution

```
1  #add month to data frame
2  mydf$month<-mon
3
4  #saving as csv
5  write.csv(mydf, "mydata.csv")
6
7  #saving as dta
8  install.packages("foreign")
9  library(foreign)
10 write.dta(as.data.frame(mydf), "mydata.dta")
11
12 #loading in STATA file
13 new<-read.dta("mydata.dta")
```

Any Questions?