Practical 5

Adele Valeria

November 9, 2019

Getting and Cleaning Data

West Nile Virus (WNV) Mosquito Test

1. Import packages

```
library(tidyr)
```

2. Import WNV data

Use *na.strings* to convert missing values to NA, then check the summary. We also want to know if there is any missing value in the data, use *anyNA* to check.

```
wnv <- read.csv("WNV_mosquito_test_results.csv", na.strings = "")
```

```
summary(wnv)
```

```
##   SEASON.YEAR      TEST.ID                            BLOCK
##  Min.   :2007   Min.   :20000   100XX W OHARE AIRPORT   : 2949
##  1st Qu.:2009   1st Qu.:27718   127XX S DOTY AVE        :  787
##  Median :2012   Median :35150   101XX S STONY ISLAND AVE:  627
##  Mean   :2013   Mean   :35156   41XX N OAK PARK AVE     :  597
##  3rd Qu.:2016   3rd Qu.:42641   52XX S KOLMAR AVE       :  514
##  Max.   :2019   Max.   :50029   70XX W ARMITAGE AVE     :  484
##                                 (Other)                 :23531
##       TRAP          TRAP_TYPE                  TEST.DATE
##  T115   :  787   CDC     : 1256   08/15/2007 12:08:00:  276
##  T002   :  593   GRAVID  :27956   08/03/2012 12:08:00:  245
##  T138   :  555   OVI     :    1   08/21/2014 12:08:00:  238
##  T114   :  504   SENTINEL:  276   07/27/2012 12:07:00:  237
##  T151   :  484                    08/14/2014 12:08:00:  227
##  T008   :  475                    07/09/2012 12:07:00:  215
##  (Other):26091                    (Other)            :28051
##  NUMBER.OF.MOSQUITOES                  SPECIES
##  Min.   : 1.00        CULEX PIPIENS/RESTUANS:13354
##  1st Qu.: 2.00        CULEX RESTUANS        :10058
##  Median : 5.00        CULEX PIPIENS         : 4864
##  Mean   :12.35        CULEX TERRITANS       :  910
##  3rd Qu.:16.00        CULEX SALINARIUS      :  218
##  Max.   :77.00        CULEX TARSALIS        :   48
##                       (Other)               :   37
##                                    LOCATION
##  (41.66238672759086, -87.59017972751752) :  787
##  (41.956298856118664, -87.79751744482932):  593
##  (41.71054240215372, -87.58455893336821) :  555
##  (41.79821072626856, -87.73692496319906) :  504
##  (41.91613471854847, -87.80109280863755) :  484
##  (Other)                                 :22150
```

```
##  NA's                                  : 4416
```

```
anyNA(wnv)
```

```
## [1] TRUE
```

3. Drop missing values

How do we find NA?

```
apply(is.na(wnv), 2, which)
```

Use *drop_na* command from *tidyr* package to drop the incomplete records.

```
dim(wnv) #Check the dimension of the data
```

```
## [1] 29489     9
```

```
wnv <- drop_na(wnv, LOCATION)
dim(wnv)
```

```
## [1] 25073     9
```

Before deletion, there were 29489 rows in wnv and 4416 of them were NA. Thus, R has successfully deleted all the missing records because 29489 - 4416 = 25073.

4. Check data types(class) of the variables in data frame

```
typeof()
class()
```

5. Change variable name

More info: rprogramming.net/rename-columns-in-r/

```
names(wnv)
```

```
## [1] "SEASON.YEAR"          "TEST.ID"              "BLOCK"
## [4] "TRAP"                 "TRAP_TYPE"            "TEST.DATE"
## [7] "NUMBER.OF.MOSQUITOES" "SPECIES"              "LOCATION"
```

```
names(wnv)[names(wnv) == "SEASON.YEAR"] <- "YEAR"
names(wnv)
```

```
## [1] "YEAR"                 "TEST.ID"              "BLOCK"
## [4] "TRAP"                 "TRAP_TYPE"            "TEST.DATE"
## [7] "NUMBER.OF.MOSQUITOES" "SPECIES"              "LOCATION"
```

6. Convert TEST.DATE type to POSIXct format

```
class(wnv$TEST.DATE)
```

```
## [1] "factor"
```

```
wnv$TEST.DATE <- as.POSIXct(wnv$TEST.DATE, format = "%m/%d/%Y %H:%M:%S" ,tz="America/Chicago")
class(wnv$TEST.DATE)
```

```
## [1] "POSIXct" "POSIXt"
```

7. Convert timezone

Assign the first datetime to "dat1"

```
dat1 <- wnv$TEST.DATE[1]
dat1
```

```
## [1] "2019-09-26 12:09:00 CDT"
```

```
attributes(dat1)
```

```
## $class
## [1] "POSIXct" "POSIXt"
##
## $tzone
## [1] "America/Chicago"
```

```
attributes(dat1)$tzone <- "America/Los_Angeles"
dat1
```

```
## [1] "2019-09-26 10:09:00 PDT"
```

8. Use *gsub*

More info: www.programmingr.com/tutorial/gsub-in-r/

```
gsub(search_term, replacement_term, string_searched, ignore.case = FALSE, perl = FALSE, fixed = FALSE, u
```

More info: www.rdocumentation.org/packages/tidyr/versions/0.8.3/topics/separate

```
separate(data, col, into, sep = "[^[:alnum:]]+", remove = TRUE,
  convert = FALSE, extra = "warn", fill = "warn", ...)
```

```
wnv$LOCATION <- gsub("[()]","", wnv$LOCATION, perl = T)
wnv <- separate(wnv,LOCATION, into = c("LATITUDE", "LONGITUDE"), sep = ",", remove = F, fill = "left" ,c
head(wnv)
```

```
##   YEAR TEST.ID                BLOCK TRAP TRAP_TYPE           TEST.DATE
## 1 2019   49933 62XX N MCCLELLAN AVE T236    GRAVID 2019-09-26 12:09:00
## 2 2019   49952    17XX N PULASKI RD T039    GRAVID 2019-09-26 12:09:00
## 3 2019   49966   11XX W CHICAGO AVE T049    GRAVID 2019-09-26 12:09:00
## 4 2019   49984       63XX W 64TH ST T155    GRAVID 2019-09-26 12:09:00
## 5 2019   50009       17XX W 95TH ST T094    GRAVID 2019-09-26 12:09:00
## 6 2019   49929    71XX N HARLEM AVE T233    GRAVID 2019-09-26 12:09:00
##   NUMBER.OF.MOSQUITOES               SPECIES
## 1                    3         CULEX RESTUANS
## 2                    2 CULEX PIPIENS/RESTUANS
## 3                   12 CULEX PIPIENS/RESTUANS
## 4                    4 CULEX PIPIENS/RESTUANS
## 5                    6 CULEX PIPIENS/RESTUANS
## 6                   23 CULEX PIPIENS/RESTUANS
##                              LOCATION LATITUDE LONGITUDE
## 1  41.99496630402897, -87.77083721987879 41.99497 -87.77084
## 2  41.91356758228873, -87.72630030176042 41.91357 -87.72630
## 3 41.896131092623506, -87.65676212387862 41.89613 -87.65676
## 4  41.77600539167921, -87.77940766760916 41.77601 -87.77941
## 5  41.72128749967918, -87.66523570170051 41.72129 -87.66524
## 6   42.0106432736568, -87.80679730045945 42.01064 -87.80680
```

Tests for Antibodies to Trachoma PGP3 Antigen

1. Import data

```
pgp3 <- read.csv("Tests_PGP3.csv", na.strings = c("","NA"))
summary(pgp3)
```

```
##     SampleID          sex            age.f        elisa.od
##  Min.   :  1.0   Min.   :1.000   (0,10] : 97   Min.   :0.0460
##  1st Qu.:145.8   1st Qu.:1.000   (10,20]: 78   1st Qu.:0.1860
##  Median :290.5   Median :1.000   (20,30]: 76   Median :0.5180
##  Mean   :290.5   Mean   :1.356   (30,40]: 67   Mean   :0.9929
##  3rd Qu.:435.2   3rd Qu.:2.000   (40,50]: 63   3rd Qu.:1.7040
##  Max.   :580.0   Max.   :2.000   (50,90]: 92   Max.   :4.0880
##                  NA's   :103     NA's   :107
##   elisa.pre.od
##  Min.   :0.0220
##  1st Qu.:0.1760
##  Median :0.4665
##  Mean   :0.9174
##  3rd Qu.:1.6387
##  Max.   :3.2820
##
```

2. Make variables more readable

Convert SampleID, age.F, sex to readable factor.

```
pgp3$SampleID <- as.character(pgp3$SampleID)
pgp3$age.f <- as.factor(pgp3$age.f)
pgp3$sex <- gsub("1","M",as.character(pgp3$sex))
pgp3$sex <- gsub("2","F",as.character(pgp3$sex))
pgp3$sex <- as.factor(pgp3$sex)
summary(pgp3)
```

```
##    SampleID           sex          age.f        elisa.od
##  Length:580         F   :170   (0,10] : 97   Min.   :0.0460
##  Class :character   M   :307   (10,20]: 78   1st Qu.:0.1860
##  Mode  :character   NA's:103   (20,30]: 76   Median :0.5180
##                                (30,40]: 67   Mean   :0.9929
##                                (40,50]: 63   3rd Qu.:1.7040
##                                (50,90]: 92   Max.   :4.0880
##                                NA's   :107
##   elisa.pre.od
##  Min.   :0.0220
##  1st Qu.:0.1760
##  Median :0.4665
##  Mean   :0.9174
##  3rd Qu.:1.6387
##  Max.   :3.2820
##
```

3. Drop incomplete records

```
anyNA(pgp3)
```

```
## [1] TRUE
```

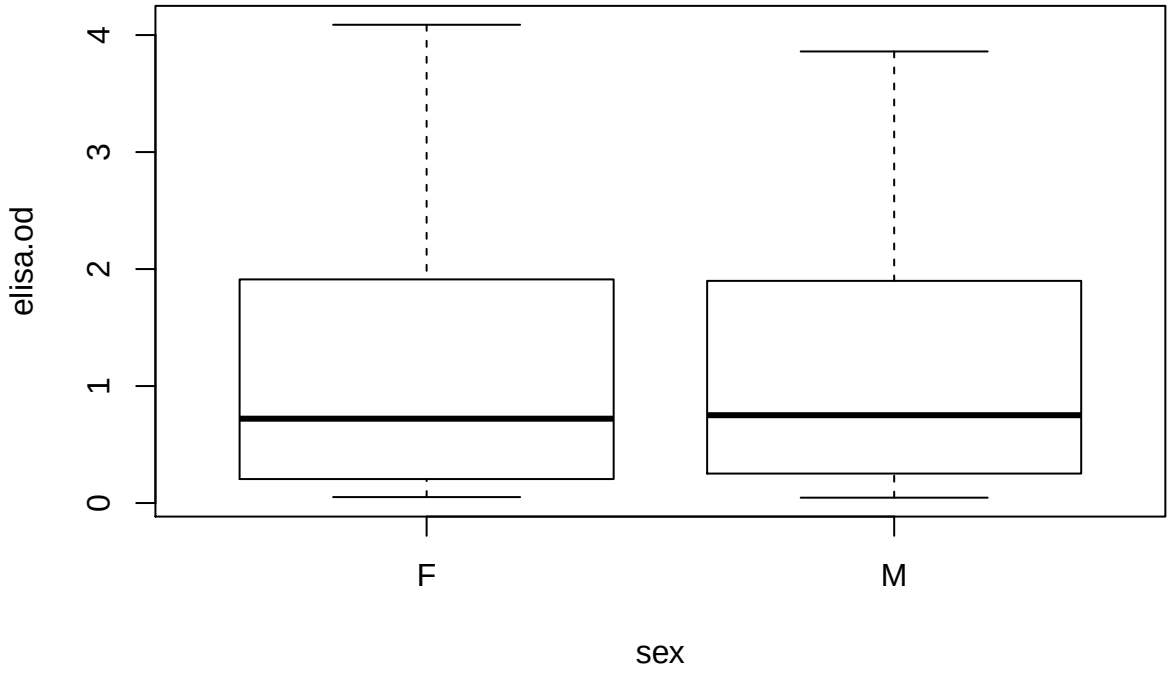
```
pgp3<- drop_na(pgp3)
anyNA(pgp3)
```

```
## [1] FALSE
```

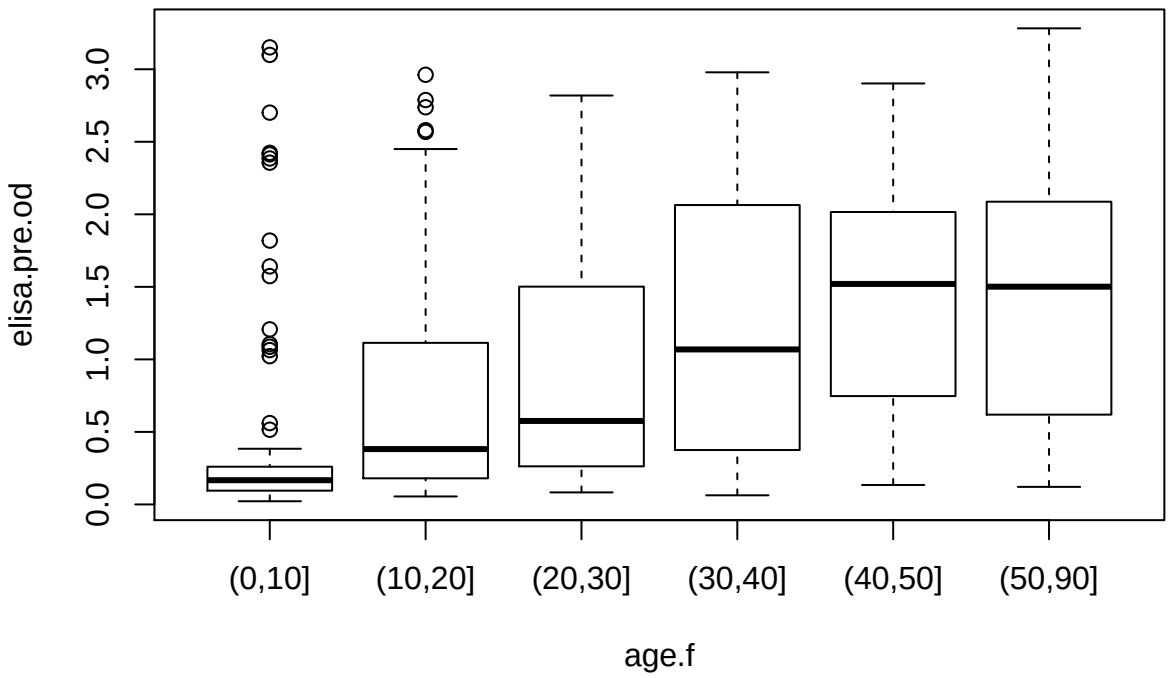
4. Boxplots

View the relationship between elisa.od and sex plus elisa.od and age.f

```
boxplot(data=pgp3, elisa.od~sex)
```




```
boxplot(data=pgp3, elisa.pre.od~age.f)
```



5. Use *gather* to combine two measurements into one variable

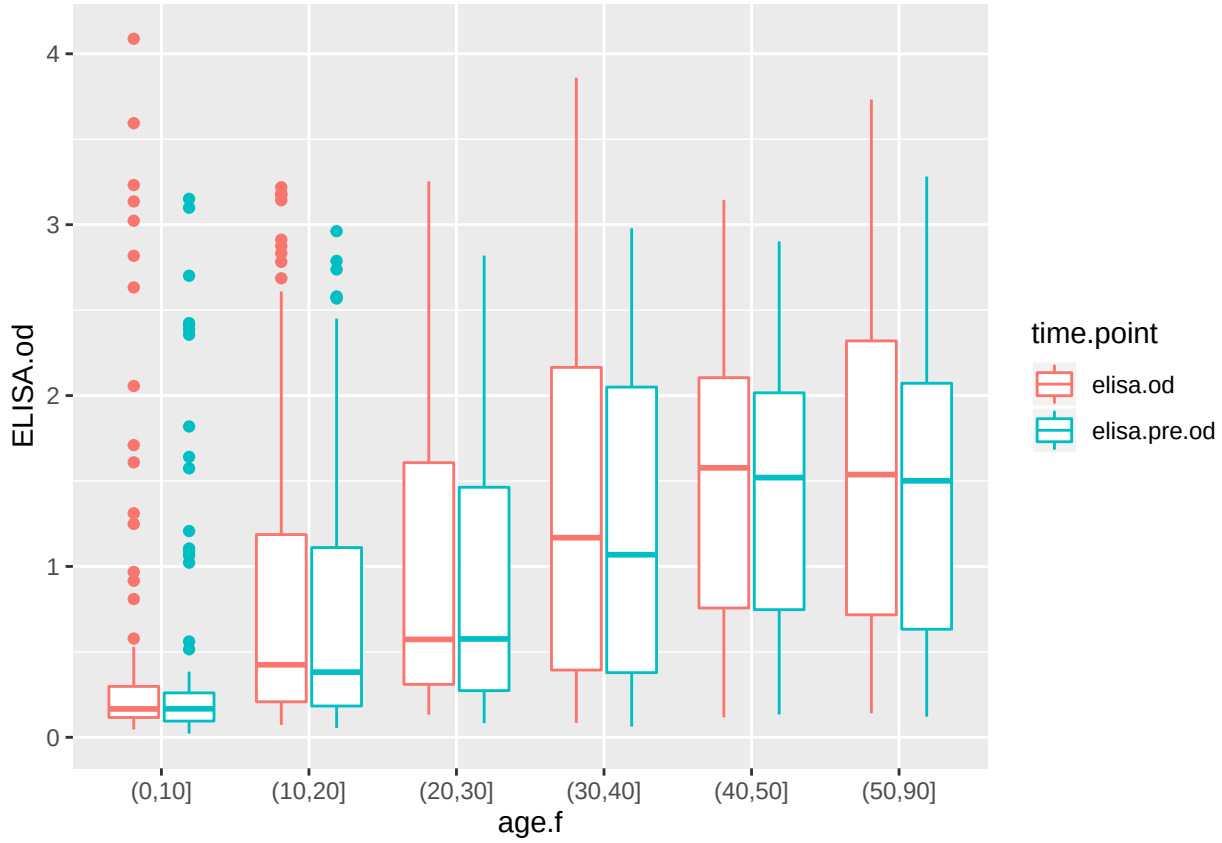
Combine the variables "elisa.od" and "elisa.pre.od" into one variable "ELISA.od" because both are measurements at different time points with ELISA.

```
pgp3 <- gather(pgp3, key = "time.point", value = "ELISA.od", elisa.od:elisa.pre.od, factor_key = T)
summary(pgp3)
```

```
##    SampleID         sex         age.f            time.point
##  Length:944         F:334   (0,10] :194   elisa.od    :472
##  Class :character   M:610   (10,20]:156   elisa.pre.od:472
##  Mode  :character           (20,30]:152
##                             (30,40]:132
##                             (40,50]:126
##                             (50,90]:184
##     ELISA.od
##  Min.   :0.0220
##  1st Qu.:0.2167
##  Median :0.7230
##  Mean   :1.0644
##  3rd Qu.:1.8455
##  Max.   :4.0880
```

6. ggplot

```
library(ggplot2)
ggplot(data= pgp3, aes(age.f, ELISA.od, color=time.point)) + geom_boxplot()
```



7. Use *spread* to reverse the reshaping

```
pgp3 <- spread(pgp3, key = time.point, value = ELISA.od )
summary(pgp3)
```

```
##    SampleID          sex         age.f       elisa.od      elisa.pre.od
##  Length:472         F:167   (0,10] :97   Min.   :0.046   Min.   :0.0220
##  Class :character   M:305   (10,20]:78   1st Qu.:0.235   1st Qu.:0.2018
##  Mode  :character           (20,30]:76   Median :0.735   Median :0.7055
##                             (30,40]:66   Mean   :1.111   Mean   :1.0183
##                             (40,50]:63   3rd Qu.:1.904   3rd Qu.:1.8120
##                             (50,90]:92   Max.   :4.088   Max.   :3.2820
```