Exercises-1: Intro to R- M & E data

Nathan Green, Imperial College London

07/09/2019

Read in the data

```
library(dplyr)
```

```
## Warning: package 'dplyr' was built under R version 3.5.3
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(dataPakistan)

data_name <- system.file(package = "dataPakistan", "extdata") %>% dir(full.names = TRUE) %>% .[2]
dat <- readxl::read_excel(data_name, sheet = 1) %>% as.data.frame()
```

First of all lets take a quick look at the data for a sanity check.

```
dim(dat)
```

```
## [1] 1268   30
```

```
head(dat)
```

```
##   Month Province      District Reported UC's Recall:Checked
## 1   Jan      AJK JEHLUM_VALLEY            14            286
## 2   Jan      AJK  MUZAFFARABAD            28            466
## 3   Jan      AJK        NEELUM            10            621
## 4   Jan      AJK         KOTLI            37           2272
## 5   Jan      AJK        MIRPUR            25           1597
## 6   Jan      AJK          BAGH            21           1455
##   Recall:Vaccinated Recall:Vaccinated% Total Missed Team did not visit
## 1               286               1.00            0                   0
## 2               435               0.93           31                   0
## 3               621               1.00            0                   0
## 4              2272               1.00            0                   0
## 5              1597               1.00            0                   0
## 6              1365               0.94           90                   0
##   Team did not visit% Visited but not vaccinated
## 1                   0                          0
## 2                   0                          2
## 3                   0                          0
## 4                   0                          0
## 5                   0                          0
## 6                   0                          0
##   Visited but not vaccinated% Child Away Child Away% Refusals Refusals%
```

```
## 1                       0.00          0        0.00        0         0
## 2                       0.06         29        0.94        0         0
## 3                       0.00          0        0.00        0         0
## 4                       0.00          0        0.00        0         0
## 5                       0.00          0        0.00        0         0
## 6                       0.00         90        1.00        0         0
##   Child Sleep Child Sleep% Others Others% Seen by Monitor Finger Marked
## 1           0            0      0       0             166           149
## 2           0            0      0       0             554           493
## 3           0            0      0       0             638           638
## 4           0            0      0       0            2254          2254
## 5           0            0      0       0            1595          1595
## 6           0            0      0       0            1120          1119
##   Finger Marked% Areas Monitored Poorly covered areas Missed areas
## 1           0.90              82                    0            0
## 2           0.89              52                    0            0
## 3           1.00              67                    0            0
## 4           1.00             197                    0            0
## 5           1.00             144                    0            0
## 6           1.00             156                    0            0
##   Poorly covered areas % Missed areas % Vaccinated but not Finger Marked
## 1                      0              0                                 0
## 2                      0              0                                 0
## 3                      0              0                                 0
## 4                      0              0                                 0
## 5                      0              0                                 0
## 6                      0              0                                 0
##   Vaccinated but not Finger Marked %
## 1                                  0
## 2                                  0
## 3                                  0
## 4                                  0
## 5                                  0
## 6                                  0
```

```
str(dat)
```

```
## 'data.frame':    1268 obs. of  30 variables:
##  $ Month                           : chr  "Jan" "Jan" "Jan" "Jan" ...
##  $ Province                        : chr  "AJK" "AJK" "AJK" "AJK" ...
##  $ District                        : chr  "JEHLUM_VALLEY" "MUZAFFARABAD" "NEELUM" "KOTLI" ...
##  $ Reported UC's                   : num  14 28 10 37 25 21 9 27 12 35 ...
##  $ Recall:Checked                  : num  286 466 621 2272 1597 ...
##  $ Recall:Vaccinated               : num  286 435 621 2272 1597 ...
##  $ Recall:Vaccinated%              : num  1 0.93 1 1 1 0.94 1 0.98 1 0.84 ...
##  $ Total Missed                    : num  0 31 0 0 0 ...
##  $ Team did not visit              : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ Team did not visit%             : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ Visited but not vaccinated      : num  0 2 0 0 0 0 0 0 0 34 ...
##  $ Visited but not vaccinated%     : num  0 0.06 0 0 0 0 0 0 0 0.02 ...
##  $ Child Away                      : num  0 29 0 0 0 ...
##  $ Child Away%                     : num  0 0.94 0 0 0 1 0 1 1 0.66 ...
##  $ Refusals                        : num  0 0 0 0 0 0 0 0 0 561 ...
##  $ Refusals%                       : num  0 0 0 0 0 0 0 0 0 0.3 ...
##  $ Child Sleep                     : num  0 0 0 0 0 0 0 0 0 18 ...
```

```
##  $ Child Sleep%                         : num  0 0 0 0 0 0 0 0 0 0.01 ...
##  $ Others                               : num  0 0 0 0 0 0 0 0 0 20 ...
##  $ Others%                              : num  0 0 0 0 0 0 0 0 0 0.01 ...
##  $ Seen by Monitor                      : num  166 554 638 2254 1595 ...
##  $ Finger Marked                        : num  149 493 638 2254 1595 ...
##  $ Finger Marked%                       : num  0.9 0.89 1 1 1 1 1 0.98 0.99 0.94 ...
##  $ Areas Monitored                      : num  82 52 67 197 144 156 56 178 118 436 ...
##  $ Poorly covered areas                 : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ Missed areas                         : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ Poorly covered areas %               : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ Missed areas %                       : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ Vaccinated but not Finger Marked  : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ Vaccinated but not Finger Marked %: num  0 0 0 0 0 0 0 0 0 0 ...
```

Its important to know what class each column has for when we do arithmetic and plotting witht them R will handle them differently. We can check what types each column has using the `class()` function.

We can check each column in a simple loop (other ways are possible!)

```
for (i in 1:ncol(dat)){
  print(typeof(dat[1,i]))
}
```

```
## [1] "character"
## [1] "character"
## [1] "character"
## [1] "double"
## [1] "double"
## [1] "double"
## [1] "double"
## [1] "double"
## [1] "double"
## [1] "double"
## [1] "double"
## [1] "double"
## [1] "double"
## [1] "double"
## [1] "double"
## [1] "double"
## [1] "double"
## [1] "double"
## [1] "double"
## [1] "double"
## [1] "double"
## [1] "double"
## [1] "double"
## [1] "double"
## [1] "double"
## [1] "double"
## [1] "double"
## [1] "double"
## [1] "double"
## [1] "double"
## [1] "double"
```

We see that the first 3 columns are characters and the rest numbers which seems sensible.

Now lets check for missing values. We can do this in the same way as above with a loop

```
for (i in 1:ncol(dat)){
  print(anyNA(dat[,i]))
}
```

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

Looks good, no missing data. Now we can check for some unusual values that may be typos are need further investigation. Lets generate summarys for each column.

```
for (i in 1:ncol(dat)){
  print(names(dat)[i])
  print(summary(dat[,i]))
}
```

```
## [1] "Month"
##    Length     Class      Mode
##      1268 character character
## [1] "Province"
##    Length     Class      Mode
##      1268 character character
## [1] "District"
##    Length     Class      Mode
##      1268 character character
## [1] "Reported UC's"
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1.00   12.00   25.00   57.08   40.00 4744.00
```

```
## [1] "Recall:Checked"
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##      10    1433    4714   13717    9836 1059598
## [1] "Recall:Vaccinated"
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##      10    1298    4182   12170    8700  951326
## [1] "Recall:Vaccinated%"
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.4800  0.8700  0.9000  0.8983  0.9300  1.0000
## [1] "Total Missed"
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
##      0.0    125.8    430.0   1547.0   1133.8 108272.0
## [1] "Team did not visit"
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00    0.00    1.00   16.26    9.00 1450.00
## [1] "Team did not visit%"
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.00000 0.00000 0.00000 0.01892 0.01000 0.80000
## [1] "Visited but not vaccinated"
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00    6.00   20.00   65.45   48.00 5306.00
## [1] "Visited but not vaccinated%"
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.00000 0.01000 0.04000 0.08098 0.10000 1.00000
## [1] "Child Away"
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     0.0   106.0   330.5  1165.2   825.5 85882.0
## [1] "Child Away%"
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0000  0.6400  0.8400  0.7626  0.9300  1.0000
## [1] "Refusals"
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     0.0     0.0     5.0   241.9    52.0 16804.0
## [1] "Refusals%"
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.00000 0.00000 0.01000 0.07535 0.07000 0.53000
## [1] "Child Sleep"
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00    0.00    0.00   33.73    6.00 2674.00
## [1] "Child Sleep%"
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.00000 0.00000 0.00000 0.01392 0.01000 0.26000
## [1] "Others"
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00    0.00    2.00   24.47   13.00 1836.00
## [1] "Others%"
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.00000 0.00000 0.00000 0.01155 0.02000 0.42000
## [1] "Seen by Monitor"
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##       7    1077    3494    9743    6938  760169
## [1] "Finger Marked"
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##       6    1024    3338    9343    6582  724311
```

```
## [1] "Finger Marked%"
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.3500  0.9500  0.9700  0.9619  0.9900  1.0000
## [1] "Areas Monitored"
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     2.0    85.0   309.0   859.0   682.2 65428.0
## [1] "Poorly covered areas"
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00    0.00    1.00   26.48   11.00 2572.00
## [1] "Missed areas"
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
##   0.0000   0.0000   0.0000   0.7303   0.0000 153.0000
## [1] "Poorly covered areas %"
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.00000 0.00000 0.00000 0.04319 0.05000 3.00000
## [1] "Missed areas %"
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
## 0.000000 0.000000 0.000000 0.003825 0.000000 3.500000
## [1] "Vaccinated but not Finger Marked"
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   0.000   0.000   3.093   0.000 473.000
## [1] "Vaccinated but not Finger Marked %"
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
## 0.0000000 0.0000000 0.0000000 0.0003549 0.0000000 0.0470000
```

The first column is months so lets check that there are only 12 of them using `table()`. This counts the frequencies.

```
table(dat[ ,1])
```

```
##
##  Apr  Aug  Dec  Feb  Jan July  Mar  May  Nov Sept
##  164   91  163  157  164   61  101  111   95  161
```

Notice that the order is alphabetic and not starting from January. If we come to plot this data we may want to change this using `factors` and `levels`. If we reorder then we can see if there are some missing months.

```
x <- factor(dat[,1], levels = c("Jan","Feb","Mar","Apr","May","Jun","July","Aug","Sept","Oct","Nov","De
table(x)
```

```
## x
##  Jan  Feb  Mar  Apr  May  Jun July  Aug Sept  Oct  Nov  Dec
##  164  157  101  164  111    0   61   91  161    0   95  163
```

This may need looking into further.