CptS 475/575: Data Science, Fall 2018

Sukhjinder Singh

20 September 2018

Load the data into R, and check the first few rows for abnormalities. You will likely notice several.

```
msleep <- read.csv("https://scads.eecs.wsu.edu/wp-content/uploads/2017/10/msleep_ggplot2.csv")
```

```
library(dplyr)
```

```
##
## 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
```

```
summary(msleep)
```

```
##                          name             genus         vore
##  African elephant          : 1   Panthera    : 3   carni  :19
##  African giant pouched rat: 1   Spermophilus: 3   herbi  :32
##  African striped mouse     : 1   Equus       : 2   insecti: 5
##  Arctic fox                : 1   Vulpes      : 2   omni   :20
##  Arctic ground squirrel    : 1   Acinonyx    : 1   NA's   : 7
##  Asian elephant            : 1   Aotus       : 1
##  (Other)                   :77   (Other)     :71
##          order          conservation  sleep_total      sleep_rem
##  Rodentia    :22   cd          : 2    Min.   : 1.90   Min.   :0.100
##  Carnivora   :12   domesticated:10    1st Qu.: 7.85   1st Qu.:0.900
##  Primates    :12   en          : 4    Median :10.10   Median :1.500
##  Artiodactyla: 6   lc          :27    Mean   :10.43   Mean   :1.875
##  Soricomorpha: 5   nt          : 4    3rd Qu.:13.75   3rd Qu.:2.400
##  Cetacea     : 3   vu          : 7    Max.   :19.90   Max.   :6.600
##  (Other)     :23   NA's        :29                    NA's   :22
##   sleep_cycle         awake          brainwt            bodywt
##  Min.   :0.1167   Min.   : 4.10   Min.   :0.00014   Min.   :   0.005
##  1st Qu.:0.1833   1st Qu.:10.25   1st Qu.:0.00290   1st Qu.:   0.174
##  Median :0.3333   Median :13.90   Median :0.01240   Median :   1.670
##  Mean   :0.4396   Mean   :13.57   Mean   :0.28158   Mean   : 166.136
##  3rd Qu.:0.5792   3rd Qu.:16.15   3rd Qu.:0.12550   3rd Qu.:  41.750
##  Max.   :1.5000   Max.   :22.10   Max.   :5.71200   Max.   :6654.000
##  NA's   :51                       NA's   :27
```

Use select() to print the head of the columns with a title including “sleep”.

```
head(msleep)
```

```
##                          name      genus  vore        order conservation
## 1                     Cheetah   Acinonyx carni    Carnivora           lc
## 2                  Owl monkey      Aotus  omni     Primates         <NA>
## 3            Mountain beaver Aplodontia herbi     Rodentia           nt
## 4 Greater short-tailed shrew    Blarina  omni Soricomorpha           lc
## 5                        Cow        Bos herbi Artiodactyla domesticated
## 6           Three-toed sloth   Bradypus herbi       Pilosa         <NA>
##   sleep_total sleep_rem sleep_cycle awake brainwt  bodywt
## 1        12.1        NA          NA  11.9      NA  50.000
## 2        17.0       1.8          NA   7.0 0.01550   0.480
## 3        14.4       2.4          NA   9.6      NA   1.350
## 4        14.9       2.3   0.1333333   9.1 0.00029   0.019
## 5         4.0       0.7   0.6666667  20.0 0.42300 600.000
## 6        14.4       2.2   0.7666667   9.6      NA   3.850
```

a). Use filter() to count the number of animals which weigh over 50 kilograms and sleep more than 6 hours a day.

```
filter(msleep, sleep_total >6, bodywt > 50)
```

```
##               name       genus    vore        order conservation
## 1        Gray seal Haliochoerus   carni    Carnivora           lc
## 2            Human        Homo    omni     Primates         <NA>
## 3       Chimpanzee         Pan    omni     Primates         <NA>
## 4            Tiger    Panthera   carni    Carnivora           en
## 5           Jaguar    Panthera   carni    Carnivora           nt
## 6             Lion    Panthera   carni    Carnivora           vu
## 7 Giant armadillo   Priodontes insecti    Cingulata           en
## 8              Pig         Sus    omni Artiodactyla domesticated
##   sleep_total sleep_rem sleep_cycle awake brainwt  bodywt
## 1         6.2       1.5          NA  17.8   0.325  85.000
## 2         8.0       1.9    1.500000  16.0   1.320  62.000
## 3         9.7       1.4    1.416667  14.3   0.440  52.200
## 4        15.8        NA          NA   8.2      NA 162.564
## 5        10.4        NA          NA  13.6   0.157 100.000
## 6        13.5        NA          NA  10.5      NA 161.499
## 7        18.1       6.1          NA   5.9   0.081  60.000
## 8         9.1       2.4    0.500000  14.9   0.180  86.250
```

b).Use piping (%>%), select() and arrange() to print the name, order, sleep time and bodyweight of the animals with the top 6 sleep times, in order of sleep time.

```
msleep %>%
    select(name, order, sleep_total,bodywt) %>%
    arrange(order, sleep_total) %>%
    filter(sleep_total >= 17.4)
```

```
##                        name           order sleep_total bodywt
## 1             Big brown bat      Chiroptera        19.7  0.023
## 2          Little brown bat      Chiroptera        19.9  0.010
## 3      Long-nosed armadillo       Cingulata        17.4  3.500
## 4           Giant armadillo       Cingulata        18.1 60.000
## 5 North American Opossum Didelphimorphia        18.0  1.700
## 6    Thick-tailed opposum Didelphimorphia        19.4  0.370
```

c).Use mutate to add two new columns to the dataframe; wt_ratio with the ratio of brain size to body weight, rem_ratio with the ratio of rem sleep to sleep time. If you think they might be useful, feel free to extract more features than these, and describe what they are

```
msleep %>%
    mutate(rem_ratio = sleep_rem / sleep_total,
           wt_ratio = brainwt/bodywt ) %>%
    head
```

```
##                          name      genus  vore        order conservation
## 1                     Cheetah   Acinonyx carni    Carnivora           lc
## 2                  Owl monkey      Aotus  omni     Primates         <NA>
## 3             Mountain beaver Aplodontia herbi     Rodentia           nt
## 4 Greater short-tailed shrew    Blarina  omni Soricomorpha           lc
## 5                         Cow        Bos herbi Artiodactyla domesticated
## 6            Three-toed sloth   Bradypus herbi       Pilosa         <NA>
##   sleep_total sleep_rem sleep_cycle awake brainwt  bodywt rem_ratio
## 1        12.1        NA          NA  11.9      NA  50.000        NA
## 2        17.0       1.8          NA   7.0 0.01550   0.480 0.1058824
## 3        14.4       2.4          NA   9.6      NA   1.350 0.1666667
## 4        14.9       2.3   0.1333333   9.1 0.00029   0.019 0.1543624
## 5         4.0       0.7   0.6666667  20.0 0.42300 600.000 0.1750000
## 6        14.4       2.2   0.7666667   9.6      NA   3.850 0.1527778
##     wt_ratio
## 1         NA
## 2 0.03229167
## 3         NA
## 4 0.01526316
## 5 0.00070500
## 6         NA
```

d).Use group_by() and summarize() to display the average, min and max sleep times for each order. Remember to use ungroup() when you are done.

```
msleep %>%
    group_by(order) %>%
    summarise(avg_sleep = mean(sleep_total),
              min_sleep = min(sleep_total),
              max_sleep = max(sleep_total),
              total = n())
```

```
## # A tibble: 19 x 5
##    order           avg_sleep min_sleep max_sleep total
##    <fct>               <dbl>     <dbl>     <dbl> <int>
##  1 Afrosoricida        15.6       15.6      15.6     1
##  2 Artiodactyla         4.52       1.9       9.1     6
##  3 Carnivora           10.1        3.5      15.8    12
##  4 Cetacea              4.5        2.7       5.6     3
##  5 Chiroptera          19.8       19.7      19.9     2
##  6 Cingulata           17.8       17.4      18.1     2
##  7 Didelphimorphia     18.7       18        19.4     2
##  8 Diprotodontia       12.4       11.1      13.7     2
##  9 Erinaceomorpha      10.2       10.1      10.3     2
## 10 Hyracoidea           5.67       5.3       6.3     3
## 11 Lagomorpha           8.4        8.4       8.4     1
## 12 Monotremata          8.6        8.6       8.6     1
## 13 Perissodactyla       3.47       2.9       4.4     3
## 14 Pilosa              14.4       14.4      14.4     1
## 15 Primates            10.5        8        17      12
## 16 Proboscidea           3.6        3.3        3.9     2
## 17 Rodentia             12.5        7         16.6    22
## 18 Scandentia            8.9        8.9        8.9     1
## 19 Soricomorpha         11.1        8.4       14.9     5
```

```
ungroup(msleep)
```

```
##                              name         genus    vore           order
## 1                         Cheetah      Acinonyx   carni       Carnivora
## 2                      Owl monkey         Aotus    omni        Primates
## 3                 Mountain beaver    Aplodontia   herbi        Rodentia
## 4      Greater short-tailed shrew       Blarina    omni    Soricomorpha
## 5                             Cow           Bos   herbi    Artiodactyla
## 6                Three-toed sloth      Bradypus   herbi          Pilosa
## 7               Northern fur seal   Callorhinus   carni       Carnivora
## 8                    Vesper mouse       Calomys    <NA>        Rodentia
## 9                             Dog         Canis   carni       Carnivora
## 10                       Roe deer     Capreolus   herbi    Artiodactyla
## 11                           Goat         Capri   herbi    Artiodactyla
## 12                     Guinea pig         Cavis   herbi        Rodentia
## 13                         Grivet Cercopithecus    omni        Primates
## 14                     Chinchilla    Chinchilla   herbi        Rodentia
## 15                Star-nosed mole     Condylura    omni    Soricomorpha
## 16      African giant pouched rat    Cricetomys    omni        Rodentia
## 17      Lesser short-tailed shrew     Cryptotis    omni    Soricomorpha
## 18           Long-nosed armadillo       Dasypus   carni       Cingulata
## 19                     Tree hyrax   Dendrohyrax   herbi      Hyracoidea
## 20         North American Opossum     Didelphis    omni Didelphimorphia
## 21                 Asian elephant       Elephas   herbi     Proboscidea
## 22                  Big brown bat     Eptesicus insecti      Chiroptera
## 23                          Horse         Equus   herbi  Perissodactyla
## 24                         Donkey         Equus   herbi  Perissodactyla
## 25              European hedgehog     Erinaceus    omni  Erinaceomorpha
## 26                   Patas monkey  Erythrocebus    omni        Primates
## 27      Western american chipmunk      Eutamias   herbi        Rodentia
## 28                   Domestic cat         Felis   carni       Carnivora
## 29                         Galago        Galago    omni        Primates
## 30                        Giraffe       Giraffa   herbi    Artiodactyla
## 31                    Pilot whale Globicephalus   carni         Cetacea
## 32                      Gray seal  Haliochoerus   carni       Carnivora
## 33                     Gray hyrax   Heterohyrax   herbi      Hyracoidea
## 34                          Human          Homo    omni        Primates
## 35                 Mongoose lemur         Lemur   herbi        Primates
## 36               African elephant     Loxodonta   herbi     Proboscidea
## 37           Thick-tailed opposum    Lutreolina   carni Didelphimorphia
## 38                        Macaque        Macaca    omni        Primates
## 39               Mongolian gerbil      Meriones   herbi        Rodentia
## 40                 Golden hamster  Mesocricetus   herbi        Rodentia
## 41                          Vole       Microtus   herbi        Rodentia
## 42                    House mouse           Mus   herbi        Rodentia
## 43               Little brown bat        Myotis insecti      Chiroptera
## 44           Round-tailed muskrat      Neofiber   herbi        Rodentia
## 45                     Slow loris     Nyctibeus   carni        Primates
## 46                           Degu       Octodon   herbi        Rodentia
## 47     Northern grasshopper mouse     Onychomys   carni        Rodentia
## 48                         Rabbit   Oryctolagus   herbi      Lagomorpha
## 49                          Sheep          Ovis   herbi    Artiodactyla
## 50                     Chimpanzee           Pan    omni        Primates
## 51                          Tiger      Panthera   carni       Carnivora
## 52                         Jaguar      Panthera   carni       Carnivora
## 53                           Lion      Panthera   carni       Carnivora
## 54                         Baboon         Papio    omni        Primates
## 55                Desert hedgehog   Paraechinus    <NA>  Erinaceomorpha
## 56                          Potto  Perodicticus    omni        Primates
## 57                     Deer mouse    Peromyscus    <NA>        Rodentia
## 58                      Phalanger     Phalanger    <NA>   Diprotodontia
## 59                   Caspian seal         Phoca   carni       Carnivora
## 60                Common porpoise      Phocoena   carni         Cetacea
## 61                        Potoroo      Potorous   herbi   Diprotodontia
## 62                Giant armadillo    Priodontes insecti       Cingulata
## 63                     Rock hyrax      Procavia    <NA>      Hyracoidea
## 64                 Laboratory rat        Rattus   herbi        Rodentia
## 65          African striped mouse     Rhabdomys    omni        Rodentia
## 66                Squirrel monkey       Saimiri    omni        Primates
## 67           Eastern american mole      Scalopus insecti    Soricomorpha
## 68                     Cotton rat      Sigmodon   herbi        Rodentia
## 69                       Mole rat        Spalax    <NA>        Rodentia
## 70         Arctic ground squirrel  Spermophilus   herbi        Rodentia
## 71 Thirteen-lined ground squirrel  Spermophilus   herbi        Rodentia
## 72 Golden-mantled ground squirrel  Spermophilus   herbi        Rodentia
## 73                     Musk shrew        Suncus    <NA>    Soricomorpha
## 74                            Pig           Sus    omni    Artiodactyla
## 75            Short-nosed echidna  Tachyglossus insecti     Monotremata
## 76      Eastern american chipmunk        Tamias   herbi        Rodentia
## 77                Brazilian tapir       Tapirus   herbi  Perissodactyla
## 78                         Tenrec        Tenrec    omni    Afrosoricida
## 79                     Tree shrew        Tupaia    omni      Scandentia
## 80           Bottle-nosed dolphin      Tursiops   carni         Cetacea
## 81                          Genet       Genetta   carni       Carnivora
## 82                     Arctic fox        Vulpes   carni       Carnivora
## 83                        Red fox        Vulpes   carni       Carnivora
##    conservation sleep_total sleep_rem sleep_cycle awake brainwt   bodywt
## 1            lc        12.1        NA          NA 11.90      NA   50.000
## 2          <NA>        17.0       1.8          NA  7.00 0.01550    0.480
## 3            nt        14.4       2.4          NA  9.60      NA    1.350
## 4            lc        14.9       2.3   0.1333333  9.10 0.00029    0.019
## 5  domesticated         4.0       0.7   0.6666667 20.00 0.42300  600.000
## 6          <NA>        14.4       2.2   0.7666667  9.60      NA    3.850
## 7            vu         8.7       1.4   0.3833333 15.30      NA   20.490
## 8          <NA>         7.0        NA          NA 17.00      NA    0.045
## 9  domesticated        10.1       2.9   0.3333333 13.90 0.07000   14.000
## 10           lc         3.0        NA          NA 21.00 0.09820   14.800
## 11           lc         5.3       0.6          NA 18.70 0.11500   33.500
## 12 domesticated         9.4       0.8   0.2166667 14.60 0.00550    0.728
## 13           lc        10.0       0.7          NA 14.00      NA    4.750
## 14 domesticated        12.5       1.5   0.1166667 11.50 0.00640    0.420
## 15           lc        10.3       2.2          NA 13.70 0.00100    0.060
## 16         <NA>         8.3       2.0          NA 15.70 0.00660    1.000
## 17           lc         9.1       1.4   0.1500000 14.90 0.00014    0.005
## 18           lc        17.4       3.1   0.3833333  6.60 0.01080    3.500
## 19           lc         5.3       0.5          NA 18.70 0.01230    2.950
## 20           lc        18.0       4.9   0.3333333  6.00 0.00630    1.700
## 21           en         3.9        NA          NA 20.10 4.60300 2547.000
## 22           lc        19.7       3.9   0.1166667  4.30 0.00030    0.023
## 23 domesticated         2.9       0.6   1.0000000 21.10 0.65500  521.000
## 24 domesticated         3.1       0.4          NA 20.90 0.41900  187.000
## 25           lc        10.1       3.5   0.2833333 13.90 0.00350    0.770
## 26           lc        10.9       1.1          NA 13.10 0.11500   10.000
## 27         <NA>        14.9        NA          NA  9.10      NA    0.071
## 28 domesticated        12.5       3.2   0.4166667 11.50 0.02560    3.300
## 29         <NA>         9.8       1.1   0.5500000 14.20 0.00500    0.200
## 30           cd         1.9       0.4          NA 22.10      NA  899.995
## 31           cd         2.7       0.1          NA 21.35      NA  800.000
## 32           lc         6.2       1.5          NA 17.80 0.32500   85.000
## 33           lc         6.3       0.6          NA 17.70 0.01227    2.625
## 34         <NA>         8.0       1.9   1.5000000 16.00 1.32000   62.000
## 35           vu         9.5       0.9          NA 14.50      NA    1.670
## 36           vu         3.3        NA          NA 20.70 5.71200 6654.000
## 37           lc        19.4       6.6          NA  4.60      NA    0.370
## 38         <NA>        10.1       1.2   0.7500000 13.90 0.17900    6.800
## 39           lc        14.2       1.9          NA  9.80      NA    0.053
## 40           en        14.3       3.1   0.2000000  9.70 0.00100    0.120
## 41         <NA>        12.8        NA          NA 11.20      NA    0.035
## 42           nt        12.5       1.4   0.1833333 11.50 0.00040    0.022
## 43         <NA>        19.9       2.0   0.2000000  4.10 0.00025    0.010
## 44           nt        14.6        NA          NA  9.40      NA    0.266
## 45         <NA>        11.0        NA          NA 13.00 0.01250    1.400
## 46           lc         7.7       0.9          NA 16.30      NA    0.210
## 47           lc        14.5        NA          NA  9.50      NA    0.028
## 48 domesticated         8.4       0.9   0.4166667 15.60 0.01210    2.500
## 49 domesticated         3.8       0.6          NA 20.20 0.17500   55.500
## 50         <NA>         9.7       1.4   1.4166667 14.30 0.44000   52.200
## 51           en        15.8        NA          NA  8.20      NA  162.564
## 52           nt        10.4        NA          NA 13.60 0.15700  100.000
## 53           vu        13.5        NA          NA 10.50      NA  161.499
## 54         <NA>         9.4       1.0   0.6666667 14.60 0.18000   25.235
## 55           lc        10.3       2.7          NA 13.70 0.00240    0.550
## 56           lc        11.0        NA          NA 13.00      NA    1.100
## 57         <NA>        11.5        NA          NA 12.50      NA    0.021
## 58         <NA>        13.7       1.8          NA 10.30 0.01140    1.620
## 59           vu         3.5       0.4          NA 20.50      NA   86.000
## 60           vu         5.6        NA          NA 18.45      NA   53.180
## 61         <NA>        11.1       1.5          NA 12.90      NA    1.100
## 62           en        18.1       6.1          NA  5.90 0.08100   60.000
## 63           lc         5.4       0.5          NA 18.60 0.02100    3.600
## 64           lc        13.0       2.4   0.1833333 11.00 0.00190    0.320
## 65         <NA>         8.7        NA          NA 15.30      NA    0.044
## 66         <NA>         9.6       1.4          NA 14.40 0.02000    0.743
## 67           lc         8.4       2.1   0.1666667 15.60 0.00120    0.075
## 68         <NA>        11.3       1.1   0.1500000 12.70 0.00118    0.148
## 69         <NA>        10.6       2.4          NA 13.40 0.00300    0.122
## 70           lc        16.6        NA          NA  7.40 0.00570    0.920
## 71           lc        13.8       3.4   0.2166667 10.20 0.00400    0.101
## 72           lc        15.9       3.0          NA  8.10      NA    0.205
## 73         <NA>        12.8       2.0   0.1833333 11.20 0.00033    0.048
## 74 domesticated         9.1       2.4   0.5000000 14.90 0.18000   86.250
## 75         <NA>         8.6        NA          NA 15.40 0.02500    4.500
## 76         <NA>        15.8        NA          NA  8.20      NA    0.112
## 77           vu         4.4       1.0   0.9000000 19.60 0.16900  207.501
## 78         <NA>        15.6       2.3          NA  8.40 0.00260    0.900
## 79         <NA>         8.9       2.6   0.2333333 15.10 0.00250    0.104
## 80         <NA>         5.2        NA          NA 18.80      NA  173.330
## 81         <NA>         6.3       1.3          NA 17.70 0.01750    2.000
## 82         <NA>        12.5        NA          NA 11.50 0.04450    3.380
## 83         <NA>         9.8       2.4   0.3500000 14.20 0.05040    4.230
```

e).Make a copy of your dataframe, and use group_by() and mutate() to impute the missing brain weights as the average wt_ratio for that animal's order times the animal's weight. Make a 2 second copy of your dataframe, but this time use group_by() and mutate() to impute missing brain weights with the average brain weight for that animal's order. What assumptions do these data filling methods make? Which is the best way to impute the data, or do you see a better way, and why? You may impute or remove other variables as you find appropriate. Briefly explain your decisions.

```
Firstcopy=data.frame(msleep)
head(Firstcopy)
```

```
##                          name      genus  vore        order conservation
## 1                     Cheetah   Acinonyx carni    Carnivora           lc
## 2                  Owl monkey      Aotus  omni     Primates         <NA>
## 3             Mountain beaver Aplodontia herbi     Rodentia           nt
## 4 Greater short-tailed shrew    Blarina  omni Soricomorpha           lc
## 5                         Cow        Bos herbi Artiodactyla domesticated
## 6            Three-toed sloth   Bradypus herbi       Pilosa         <NA>
##   sleep_total sleep_rem sleep_cycle awake brainwt  bodywt
## 1        12.1        NA          NA  11.9      NA  50.000
## 2        17.0       1.8          NA   7.0 0.01550   0.480
## 3        14.4       2.4          NA   9.6      NA   1.350
## 4        14.9       2.3   0.1333333   9.1 0.00029   0.019
## 5         4.0       0.7   0.6666667  20.0 0.42300 600.000
## 6        14.4       2.2   0.7666667   9.6      NA   3.850
```

```
Firstcopy %>%
            group_by(order) %>%
            mutate(brainwt= ifelse(is.na(brainwt), mean(brainwt, na.rm=TRUE),brainwt))
```

```
## # A tibble: 83 x 11
## # Groups:   order [19]
##    name  genus vore  order conservation sleep_total sleep_rem sleep_cycle
##    <fct> <fct> <fct> <fct> <fct>              <dbl>     <dbl>       <dbl>
##  1 Chee~ Acin~ carni Carn~ lc                  12.1      NA          NA
##  2 Owl ~ Aotus omni  Prim~ <NA>                17         1.8        NA
##  3 Moun~ Aplo~ herbi Rode~ nt                  14.4       2.4        NA
##  4 Grea~ Blar~ omni  Sori~ lc                  14.9       2.3         0.133
##  5 Cow   Bos   herbi Arti~ domesticated         4         0.7         0.667
##  6 Thre~ Brad~ herbi Pilo~ <NA>                14.4       2.2         0.767
##  7 Nort~ Call~ carni Carn~ vu                   8.7       1.4         0.383
##  8 Vesp~ Calo~ <NA>  Rode~ <NA>                 7        NA          NA
##  9 Dog   Canis carni Carn~ domesticated        10.1       2.9         0.333
## 10 Roe ~ Capr~ herbi Arti~ lc                   3        NA          NA
## # ... with 73 more rows, and 3 more variables: awake <dbl>, brainwt <dbl>,
## #   bodywt <dbl>
```

```
secondcopy=data.frame(msleep)

secondcopy %>%
            group_by(order) %>%
            mutate(bodywt= ifelse(is.na(bodywt), mean(bodywt, na.rm=TRUE), bodywt))
```

```
## # A tibble: 83 x 11
## # Groups:   order [19]
##    name  genus vore  order conservation sleep_total sleep_rem sleep_cycle
##    <fct> <fct> <fct> <fct> <fct>              <dbl>     <dbl>       <dbl>
##  1 Chee~ Acin~ carni Carn~ lc                  12.1      NA        NA
##  2 Owl ~ Aotus omni  Prim~ <NA>                17         1.8      NA
##  3 Moun~ Aplo~ herbi Rode~ nt                  14.4       2.4      NA
##  4 Grea~ Blar~ omni  Sori~ lc                  14.9       2.3       0.133
##  5 Cow   Bos   herbi Arti~ domesticated         4         0.7       0.667
##  6 Thre~ Brad~ herbi Pilo~ <NA>                14.4       2.2       0.767
##  7 Nort~ Call~ carni Carn~ vu                   8.7       1.4       0.383
##  8 Vesp~ Calo~ <NA>  Rode~ <NA>                 7        NA        NA
##  9 Dog   Canis carni Carn~ domesticated        10.1       2.9       0.333
## 10 Roe ~ Capr~ herbi Arti~ lc                   3        NA        NA
## # ... with 73 more rows, and 3 more variables: awake <dbl>, brainwt <dbl>,
## #   bodywt <dbl>
```

Exercise 2

For this question, you will first need to read section 12.6 in the R for Data Science book, here (http://r4ds.had.co.nz/tidy-data.html#case-study). Grab the dataset from the tidyr package, and tidy it as shown in the case study before answering the following questions

```
readdata = read.csv("assignment 3/TB_notification.csv")
summary(readdata)
```

```
##            country          iso2           iso3       iso_numeric
##  Afghanistan   :  38   AD     :  38   ABW    :  38   Min.   :  4.0
##  Albania       :  38   AE     :  38   AFG    :  38   1st Qu.:212.0
##  Algeria       :  38   AF     :  38   AGO    :  38   Median :430.0
##  American Samoa:  38   AG     :  38   AIA    :  38   Mean   :431.7
##  Andorra       :  38   AI     :  38   ALB    :  38   3rd Qu.:646.0
##  Angola        :  38   (Other):7842   AND    :  38   Max.   :894.0
##  (Other)       :7842   NA's   :  38   (Other):7842
##  g_whoregion      year          new_sp          new_sn
##  AFR:1755    Min.   :1980   Min.   :     0   Min.   :     0.0
##  AMR:1688    1st Qu.:1989   1st Qu.:    99   1st Qu.:    60.0
##  EMR: 836    Median :1999   Median :  1054   Median :   521.5
##  EUR:2027    Mean   :1999   Mean   :  9880   Mean   :  8252.9
##  SEA: 396    3rd Qu.:2008   3rd Qu.:  5012   3rd Qu.:  2423.5
##  WPR:1368    Max.   :2017   Max.   :642321   Max.   :932998.0
##                             NA's   :4166     NA's   :4504
##     new_su           new_ep          new_oth           ret_rel
##  Min.   :     0   Min.   :     0   Min.   :  0.00   Min.   :     0.0
##  1st Qu.:     0   1st Qu.:    46   1st Qu.:  0.00   1st Qu.:     3.0
##  Median :     3   Median :   373   Median :  0.00   Median :    92.5
##  Mean   :  1230   Mean   :  3228   Mean   : 61.07   Mean   :  1122.4
##  3rd Qu.:   205   3rd Qu.:  1843   3rd Qu.:   0.00   3rd Qu.:   442.8
##  Max.   :787338   Max.   :298831   Max.   :7342.00   Max.   :112508.0
##  NA's   :5235     NA's   :3473     NA's   :6674      NA's   :4708
##     ret_taf            ret_tad           ret_oth         newret_oth
##  Min.   :    0.00   Min.   :    0.0   Min.   :     0   Min.   :    0.0
##  1st Qu.:    0.00   1st Qu.:    0.0   1st Qu.:     0   1st Qu.:    0.0
##  Median :    5.00   Median :   15.0   Median :    12   Median :    0.0
##  Mean   :  272.86   Mean   :  614.9   Mean   :  1437   Mean   :  207.8
##  3rd Qu.:   76.75   3rd Qu.:  125.2   3rd Qu.:   241   3rd Qu.:    2.0
##  Max.   :39840.00   Max.   :77618.0   Max.   :101832   Max.   :40659.0
##  NA's   :5852       NA's   :5838      NA's   :5828     NA's   :6487
##   new_labconf     new_clindx         ret_rel_labconf  ret_rel_clindx
##  Min.   :     0   Min.   :     0.0   Min.   :     0   Min.   :     0.0
##  1st Qu.:   123   1st Qu.:    28.5   1st Qu.:     3   1st Qu.:     0.0
##  Median :  1154   Median :   357.0   Median :    92   Median :     1.0
##  Mean   : 12656   Mean   : 10288.0   Mean   :  1410   Mean   :   716.6
##  3rd Qu.:  5278   3rd Qu.:  2240.5   3rd Qu.:   459   3rd Qu.:    67.0
##  Max.   :817239   Max.   :599786.0   Max.   :124679   Max.   :140820.0
##  NA's   :4838     NA's   :7039       NA's   :7079     NA's   :7178
##    ret_rel_ep         ret_nrel        notif_foreign       c_newinc
##  Min.   :   0.00   Min.   :     0.0   Min.   :   0.0   Min.   :      0
##  1st Qu.:   0.00   1st Qu.:     1.0   1st Qu.:   0.0   1st Qu.:    234
##  Median :   1.00   Median :    56.0   Median :   5.0   Median :   2162
##  Mean   :  67.81   Mean   :  1446.6   Mean   : 304.8   Mean   :  21488
##  3rd Qu.:  25.00   3rd Qu.:   413.8   3rd Qu.: 111.0   3rd Qu.:   9616
##  Max.   :2734.00   Max.   :172282.0   Max.   :9527.0   Max.   :1786681
##  NA's   :7189      NA's   :7070       NA's   :6363     NA's   :539
##    new_sp_m04       new_sp_m514       new_sp_m014       new_sp_m1524
##  Min.   :  0.000   Min.   :   0.00   Min.   :   0.00   Min.   :    0.0
##  1st Qu.:  0.000   1st Qu.:   0.00   1st Qu.:   0.00   1st Qu.:    9.0
##  Median :  0.000   Median :   1.00   Median :   5.00   Median :   90.0
##  Mean   :  8.287   Mean   :  42.23   Mean   :  83.77   Mean   : 1016.3
##  3rd Qu.:  3.000   3rd Qu.:  13.00   3rd Qu.:  37.00   3rd Qu.:  503.5
##  Max.   :655.000   Max.   :1594.00   Max.   :5001.00   Max.   :78278.0
##  NA's   :6996      NA's   :6985      NA's   :4899      NA's   :4863
##   new_sp_m2534      new_sp_m3544    new_sp_m4554    new_sp_m5564
##  Min.   :    0.0   Min.   :    0   Min.   :    0   Min.   :    0.0
##  1st Qu.:   14.0   1st Qu.:   13   1st Qu.:   12   1st Qu.:    8.0
##  Median :  150.5   Median :  131   Median :  102   Median :   63.0
##  Mean   : 1404.7   Mean   : 1317   Mean   : 1105   Mean   :  801.2
##  3rd Qu.:  716.0   3rd Qu.:  584   3rd Qu.:  440   3rd Qu.:  279.2
##  Max.   :84003.0   Max.   :90830   Max.   :82921   Max.   :63814.0
##  NA's   :4866      NA's   :4853    NA's   :4849    NA's   :4854
##    new_sp_m65        new_sp_mu         new_sp_f04       new_sp_f514
##  Min.   :    0.0   Min.   :   0.00   Min.   :  0.000   Min.   :   0.0
##  1st Qu.:    8.0   1st Qu.:   0.00   1st Qu.:  0.000   1st Qu.:   0.0
##  Median :   53.0   Median :   0.00   Median :  0.000   Median :   2.0
##  Mean   :  683.2   Mean   :  10.85   Mean   :  6.511   Mean   :  59.2
##  3rd Qu.:  233.0   3rd Qu.:   0.00   3rd Qu.:  2.000   3rd Qu.:  24.0
##  Max.   :70376.0   Max.   :7417.00   Max.   :620.000   Max.   :3132.0
##  NA's   :4863      NA's   :7153      NA's   :6995      NA's   :6982
##   new_sp_f014      new_sp_f1524      new_sp_f2534      new_sp_f3544
##  Min.   :   0.0   Min.   :    0.0   Min.   :    0.0   Min.   :    0.0
##  1st Qu.:   1.0   1st Qu.:    7.0   1st Qu.:    9.0   1st Qu.:    6.0
##  Median :   7.0   Median :   66.0   Median :   84.0   Median :   57.0
##  Mean   : 114.4   Mean   :  826.4   Mean   :  917.6   Mean   :  640.6
##  3rd Qu.:  51.0   3rd Qu.:  421.0   3rd Qu.:  476.5   3rd Qu.:  308.0
##  Max.   :8576.0   Max.   :53975.0   Max.   :49887.0   Max.   :34698.0
##  NA's   :4897     NA's   :4877      NA's   :4871      NA's   :4872
##   new_sp_f4554      new_sp_f5564       new_sp_f65      new_sp_fu
##  Min.   :    0.0   Min.   :    0.0   Min.   :    0   Min.   :   0.000
##  1st Qu.:    4.0   1st Qu.:    3.0   1st Qu.:    4   1st Qu.:   0.000
##  Median :   38.0   Median :   25.0   Median :   30   Median :   0.000
##  Mean   :  445.9   Mean   :  314.0   Mean   :  284   Mean   :   4.137
##  3rd Qu.:  211.0   3rd Qu.:  146.8   3rd Qu.:  129   3rd Qu.:   0.000
##  Max.   :23977.0   Max.   :18203.0   Max.   :21339   Max.   :2559.000
##  NA's   :4867      NA's   :4876      NA's   :4874    NA's   :7155
##    new_sn_m04       new_sn_m514      new_sn_m014       new_sn_m1524
##  Min.   :    0.0   Min.   :   0.0   Min.   :    0.0   Min.   :    0.0
##  1st Qu.:    0.0   1st Qu.:   0.0   1st Qu.:    1.0   1st Qu.:    2.0
##  Median :    3.0   Median :   4.0   Median :    9.0   Median :   15.5
##  Mean   :  155.6   Mean   : 144.4   Mean   :  308.8   Mean   :  513.0
##  3rd Qu.:   18.0   3rd Qu.:  29.0   3rd Qu.:   61.0   3rd Qu.:  102.0
##  Max.   :15147.0   Max.   :8438.0   Max.   :22355.0   Max.   :60246.0
##  NA's   :7115      NA's   :7116     NA's   :7025      NA's   :7040
##   new_sn_m2534      new_sn_m3544       new_sn_m4554      new_sn_m5564
##  Min.   :    0.0   Min.   :      0.0   Min.   :    0.0   Min.   :    0.0
##  1st Qu.:    2.0   1st Qu.:      2.0   1st Qu.:    2.0   1st Qu.:    2.0
##  Median :   23.0   Median :     19.0   Median :   19.0   Median :   16.0
##  Mean   :  653.7   Mean   :    837.9   Mean   :  520.8   Mean   :  448.6
##  3rd Qu.:  135.5   3rd Qu.:    132.0   3rd Qu.:  127.5   3rd Qu.:  102.0
##  Max.   :50282.0   Max.   :250051.0   Max.   :57181.0   Max.   :64972.0
##  NA's   :7048      NA's   :7045       NA's   :7043      NA's   :7049
##    new_sn_m65      new_sn_m15plus       new_sn_mu         new_sn_f04
##  Min.   :    0.0   Min.   :      0.0   Min.   :    0.0   Min.   :    0.0
##  1st Qu.:    2.0   1st Qu.:     15.0   1st Qu.:    0.0   1st Qu.:    0.0
##  Median :   20.5   Median :    124.5   Median :    0.0   Median :    3.0
##  Mean   :  460.4   Mean   :   3480.0   Mean   :  246.7   Mean   :  139.4
##  3rd Qu.:  111.8   3rd Qu.:    886.0   3rd Qu.:    0.0   3rd Qu.:   14.0
##  Max.   :74282.0   Max.   :361435.0   Max.   :66885.0   Max.   :14084.0
##  NA's   :7050      NA's   :7014       NA's   :7294      NA's   :7118
##   new_sn_f514    new_sn_f014     new_sn_f1524      new_sn_f2534
##  Min.   :   0   Min.   :    0   Min.   :    0.0   Min.   :    0.0
##  1st Qu.:   0   1st Qu.:    1   1st Qu.:    1.0   1st Qu.:    2.0
##  Median :   5   Median :    8   Median :   12.0   Median :   18.0
##  Mean   : 146   Mean   :  292   Mean   :  407.9   Mean   :  466.3
##  3rd Qu.:  27   3rd Qu.:   58   3rd Qu.:   89.0   3rd Qu.:  103.2
##  Max.   :7322   Max.   :21406   Max.   :35518.0   Max.   :28753.0
##  NA's   :7119   NA's   :7030    NA's   :7048      NA's   :7054
##   new_sn_f3544        new_sn_f4554       new_sn_f5564
##  Min.   :     0.00   Min.   :    0.00   Min.   :    0.0
##  1st Qu.:     1.00   1st Qu.:    1.00   1st Qu.:    1.0
##  Median :    11.00   Median :   10.00   Median :    8.0
##  Mean   :   506.60   Mean   :  271.16   Mean   :  213.4
##  3rd Qu.:    82.25   3rd Qu.:   76.75   3rd Qu.:   56.0
##  Max.   :148811.00   Max.   :23869.00   Max.   :26085.0
##  NA's   :7050        NA's   :7052       NA's   :7053
##    new_sn_f65      new_sn_f15plus     new_sn_fu       new_sn_sexunk04
##  Min.   :    0.0   Min.   :     0   Min.   :    0.0   Min.   :   0.00
##  1st Qu.:    1.0   1st Qu.:     9   1st Qu.:    0.0   1st Qu.:  22.75
##  Median :   13.0   Median :    80   Median :    0.0   Median : 135.50
##  Mean   :  230.8   Mean   :  2184   Mean   :  179.4   Mean   : 467.33
##  3rd Qu.:   74.0   3rd Qu.:   608   3rd Qu.:    0.0   3rd Qu.: 752.75
##  Max.   :29630.0   Max.   :170327   Max.   :47305.0   Max.   :1667.00
##  NA's   :7051      NA's   :7026     NA's   :7297      NA's   :8046
##  new_sn_sexunk514 new_sn_sexunk014 new_sn_sexunk15plus   new_ep_m04
##  Min.   :  0.0   Min.   :    0    Min.   :     0.0    Min.   :0
##  1st Qu.: 31.0   1st Qu.:   66    1st Qu.:   639.8    1st Qu.:0
##  Median : 112.5  Median :  328    Median :  1445.5    Median :0
##  Mean   : 517.2  Mean   : 3010    Mean   : 20955.2    Mean   :0
##  3rd Qu.: 598.8  3rd Qu.: 2133    3rd Qu.:  4650.8    3rd Qu.:0
##  Max.   :4438.0  Max.   :36673    Max.   :303530.0    Max.   :0
##  NA's   :8046    NA's   :8035     NA's   :8036        NA's   :8062
##   new_ep_m514       new_ep_m014      new_ep_m1524     new_ep_m2534
##  Min.   :   0.00   Min.   :   0.0   Min.   :   0.0   Min.   :    0.0
##  1st Qu.:   0.00   1st Qu.:   0.0   1st Qu.:   1.0   1st Qu.:    1.0
##  Median :   3.00   Median :   6.0   Median :  11.0   Median :   13.0
##  Mean   :  82.31   Mean   : 128.6   Mean   : 158.3   Mean   :  201.2
##  3rd Qu.:  32.00   3rd Qu.:  56.5   3rd Qu.:  88.0   3rd Qu.:  124.0
##  Max.   :4369.00   Max.   :7869.0   Max.   :8558.0   Max.   :11843.0
##  NA's   :7122      NA's   :7032     NA's   :7044     NA's   :7050
##   new_ep_m3544        new_ep_m4554      new_ep_m5564       new_ep_m65
##  Min.   :     0.00   Min.   :   0.00   Min.   :   0.00   Min.   :   0.00
##  1st Qu.:     1.00   1st Qu.:   1.00   1st Qu.:   1.00   1st Qu.:   1.00
##  Median :    10.50   Median :   8.50   Median :   7.00   Median :  10.00
##  Mean   :   272.73   Mean   : 108.12   Mean   :  72.17   Mean   :  78.94
##  3rd Qu.:    91.25   3rd Qu.:  63.25   3rd Qu.:  46.50   3rd Qu.:  55.00
##  Max.   :105825.00   Max.   :5875.00   Max.   :3957.00   Max.   :3061.00
##  NA's   :7046        NA's   :7050      NA's   :7055      NA's   :7052
##  new_ep_m15plus       new_ep_mu          new_ep_f04       new_ep_f514
##  Min.   :     0.0   Min.   :    0.00   Min.   :   0.00   Min.   :   0.00
##  1st Qu.:     7.0   1st Qu.:    0.00   1st Qu.:   0.00   1st Qu.:   0.00
##  Median :    64.0   Median :    0.00   Median :   1.00   Median :   3.00
##  Mean   :   939.5   Mean   :   46.04   Mean   :  34.31   Mean   :  76.49
##  3rd Qu.:   576.5   3rd Qu.:    0.00   3rd Qu.:  14.00   3rd Qu.:  28.00
##  Max.   :105825.0   Max.   :16676.00   Max.   :3300.00   Max.   :4055.00
##  NA's   :7019       NA's   :7290       NA's   :7124      NA's   :7124
##   new_ep_f014       new_ep_f1524     new_ep_f2534      new_ep_f3544
##  Min.   :   0.00   Min.   :   0.0   Min.   :    0.0   Min.   :     0.0
##  1st Qu.:   0.00   1st Qu.:   1.0   1st Qu.:    1.0   1st Qu.:     1.0
##  Median :   5.00   Median :   9.0   Median :   12.0   Median :     9.0
##  Mean   : 112.89   Mean   : 149.2   Mean   :  189.5   Mean   :   241.7
##  3rd Qu.:  50.25   3rd Qu.:  78.0   3rd Qu.:   95.0   3rd Qu.:    77.0
##  Max.   :6960.00   Max.   :7866.0   Max.   :10759.0   Max.   :101015.0
##  NA's   :7038      NA's   :7049     NA's   :7049      NA's   :7049
##   new_ep_f4554      new_ep_f5564       new_ep_f65       new_ep_f15plus
##  Min.   :   0.00   Min.   :   0.00   Min.   :   0.00   Min.   :     0.0
##  1st Qu.:   1.00   1st Qu.:   1.00   1st Qu.:   0.00   1st Qu.:     6.0
##  Median :   8.00   Median :   6.00   Median :  10.00   Median :    58.0
##  Mean   :  93.78   Mean   :  63.04   Mean   :  72.31   Mean   :   863.8
##  3rd Qu.:  56.00   3rd Qu.:  42.00   3rd Qu.:  51.00   3rd Qu.:   463.0
##  Max.   :6759.00   Max.   :4684.00   Max.   :2548.00   Max.   :101015.0
##  NA's   :7053      NA's   :7053      NA's   :7056      NA's   :7024
##    new_ep_fu        new_ep_sexunk04  new_ep_sexunk514 new_ep_sexunk014
##  Min.   :    0.00   Min.   :   0.0   Min.   :   0.0   Min.   :    0.0
##  1st Qu.:    0.00   1st Qu.:  16.0   1st Qu.:  35.0   1st Qu.:   74.5
##  Median :    0.00   Median :  66.0   Median : 157.0   Median :  383.0
##  Mean   :   40.81   Mean   : 455.4   Mean   : 911.4   Mean   : 3003.1
##  3rd Qu.:    0.00   3rd Qu.: 694.0   3rd Qu.: 774.0   3rd Qu.: 1454.0
##  Max.   :21246.00   Max.   :2604.0   Max.   :5376.0   Max.   :34062.0
##  NA's   :7297       NA's   :8045     NA's   :8045     NA's   :8035
##  new_ep_sexunk15plus new_ep_sexunkageunk rel_in_agesex_flg
##  Min.   :     0      Min.   :0           Min.   :0.00
##  1st Qu.:   548      1st Qu.:0           1st Qu.:1.00
##  Median :  1119      Median :0           Median :1.00
##  Mean   : 14838      Mean   :0           Mean   :0.82
##  3rd Qu.:  4771      3rd Qu.:0           3rd Qu.:1.00
##  Max.   :200528      Max.   :0           Max.   :1.00
##  NA's   :8036        NA's   :8062        NA's   :7051
##    newrel_m04       newrel_m514       newrel_m014       newrel_m1524
##  Min.   :    0.0   Min.   :    0.0   Min.   :    0.0   Min.   :     0.0
##  1st Qu.:    1.0   1st Qu.:    2.0   1st Qu.:    4.0   1st Qu.:    14.0
##  Median :   12.0   Median :   23.0   Median :   36.5   Median :   162.5
##  Mean   :  375.3   Mean   :  616.8   Mean   :  956.0   Mean   :  2530.9
##  3rd Qu.:  101.2   3rd Qu.:  157.5   3rd Qu.:  247.5   3rd Qu.:   883.2
##  Max.   :15727.0   Max.   :38668.0   Max.   :52377.0   Max.   :197736.0
##  NA's   :7142      NA's   :7143      NA's   :7098      NA's   :7126
##   newrel_m2534        newrel_m3544     newrel_m4554
##  Min.   :     0.00   Min.   :     0   Min.   :     0.0
##  1st Qu.:    24.75   1st Qu.:    21   1st Qu.:    20.0
##  Median :   249.00   Median :   226   Median :   208.5
##  Mean   :  3269.35   Mean   :  3181   Mean   :  3010.7
##  3rd Qu.:  1148.25   3rd Qu.:   952   3rd Qu.:   805.5
##  Max.   :208747.00   Max.   :219012   Max.   :206433.0
##  NA's   :7126        NA's   :7125     NA's   :7124
##   newrel_m5564        newrel_m65       newrel_m15plus
##  Min.   :     0.0   Min.   :     0.0   Min.   :      0
##  1st Qu.:    16.0   1st Qu.:    13.5   1st Qu.:    136
##  Median :   158.0   Median :   128.0   Median :   1373
##  Mean   :  2459.7   Mean   :  2167.0   Mean   :  16502
##  3rd Qu.:   593.8   3rd Qu.:   576.0   3rd Qu.:   5331
##  Max.   :168943.0   Max.   :127287.0   Max.   :1109208
##  NA's   :7124       NA's   :7123       NA's   :7085
##    newrel_mu          newrel_f04       newrel_f514       newrel_f014
##  Min.   :    0.00   Min.   :    0.0   Min.   :    0.0   Min.   :    0.0
##  1st Qu.:    0.00   1st Qu.:    1.0   1st Qu.:    2.0   1st Qu.:    4.0
##  Median :    0.00   Median :   10.0   Median :   24.0   Median :   39.0
##  Mean   :   98.03   Mean   :  302.3   Mean   :  654.9   Mean   :  921.3
##  3rd Qu.:    0.00   3rd Qu.:   83.0   3rd Qu.:  159.0   3rd Qu.:  230.5
##  Max.   :23928.00   Max.   :11212.0   Max.   :47757.0   Max.   :55828.0
##  NA's   :7231       NA's   :7145      NA's   :7144      NA's   :7099
##   newrel_f1524     newrel_f2534       newrel_f3544       newrel_f4554
##  Min.   :     0   Min.   :     0.0   Min.   :    0.00   Min.   :    0.0
##  1st Qu.:    11   1st Qu.:    15.0   1st Qu.:   10.25   1st Qu.:    9.0
##  Median :   119   Median :   172.0   Median :  124.50   Median :   93.0
##  Mean   :  2070   Mean   :  2217.2   Mean   : 1682.52   Mean   : 1360.4
##  3rd Qu.:   670   3rd Qu.:   780.5   3rd Qu.:  579.50   3rd Qu.:  450.5
##  Max.   :176341   Max.   :141461.0   Max.   :95103.00   Max.   :70606.0
##  NA's   :7125     NA's   :7123       NA's   :7124       NA's   :7123
##   newrel_f5564     newrel_f65      newrel_f15plus       newrel_fu
##  Min.   :    0   Min.   :    0.0   Min.   :     0.0   Min.   :    0.00
##  1st Qu.:    8   1st Qu.:    9.0   1st Qu.:    72.5   1st Qu.:    0.00
##  Median :   72   Median :   78.0   Median :   766.0   Median :    0.00
##  Mean   : 1059   Mean   : 1015.2   Mean   :  9470.0   Mean   :   64.72
##  3rd Qu.:  339   3rd Qu.:  400.2   3rd Qu.:  3568.5   3rd Qu.:    0.00
##  Max.   :54259   Max.   :53551.0   Max.   :571905.0   Max.   :14494.00
##  NA's   :7124    NA's   :7124      NA's   :7083       NA's   :7237
##  newrel_sexunk04  newrel_sexunk514 newrel_sexunk014 newrel_sexunk15plus
##  Min.   :   0.0   Min.   :   0.0   Min.   :    0    Min.   :      0
##  1st Qu.:   0.0   1st Qu.:   0.0   1st Qu.:    0    1st Qu.:      0
##  Median :   0.0   Median :   0.0   Median :    0    Median :      0
##  Mean   : 431.6   Mean   : 506.1   Mean   : 3657    Mean   :  55454
##  3rd Qu.: 229.2   3rd Qu.: 347.2   3rd Qu.:  641    3rd Qu.:   7094
##  Max.   :3239.0   Max.   :3720.0   Max.   :64726    Max.   :1179179
##  NA's   :8050     NA's   :8050     NA's   :8047     NA's   :8047
##  newrel_sexunkageunk rdx_data_available   newinc_rdx     rdxsurvey_newinc
##  Min.   :    0       Min.   : 0.00      Min.   :     0   Min.   :   14.0
##  1st Qu.:    0       1st Qu.: 0.00      1st Qu.:    35   1st Qu.:  200.8
##  Median :    0       Median :60.00      Median :   236   Median :  870.5
##  Mean   : 1232       Mean   :37.58      Mean   :  7110   Mean   : 4517.5
##  3rd Qu.:  301       3rd Qu.:60.00      3rd Qu.:  1206   3rd Qu.: 5187.2
##  Max.   :11716       Max.   :61.00      Max.   :720051   Max.   :16315.0
##  NA's   :8050        NA's   :7471       NA's   :7709     NA's   :8066
##  rdxsurvey_newinc_rdx    rdst_new         rdst_ret         rdst_unk
##  Min.   :  14.0       Min.   :     0   Min.   :     0   Min.   :     0.0
##  1st Qu.:  74.0       1st Qu.:    13   1st Qu.:     3   1st Qu.:     0.0
##  Median : 561.5       Median :   233   Median :    69   Median :     0.0
##  Mean   :1263.2       Mean   :  3978   Mean   :  2056   Mean   :  1339.6
##  3rd Qu.:1750.8       3rd Qu.:  1291   3rd Qu.:   461   3rd Qu.:    31.5
##  Max.   :3916.0       Max.   :537180   Max.   :283400   Max.   :218231.0
##  NA's   :8066         NA's   :7173     NA's   :7184     NA's   :7219
##    conf_rrmdr         conf_mdr           rr_sldst         all_conf_xdr
##  Min.   :    0.0   Min.   :    0.00   Min.   :    0.00   Min.   :   0.00
##  1st Qu.:    1.0   1st Qu.:    0.00   1st Qu.:    0.00   1st Qu.:   0.00
##  Median :   27.0   Median :   12.50   Median :    6.00   Median :   0.00
##  Mean   :  724.9   Mean   :  393.01   Mean   :  343.57   Mean   :  46.05
##  3rd Qu.:  178.5   3rd Qu.:   87.75   3rd Qu.:   58.75   3rd Qu.:   3.00
##  Max.   :39009.0   Max.   :25971.00   Max.   :26832.00   Max.   :3661.00
##  NA's   :7286      NA's   :6528       NA's   :7524       NA's   :7509
##  unconf_rrmdr_tx   conf_rrmdr_tx     unconf_mdr_tx      conf_mdr_tx
##  Min.   :   0.00   Min.   :    0.0   Min.   :   0.00   Min.   :    0.0
##  1st Qu.:   0.00   1st Qu.:    1.0   1st Qu.:   0.00   1st Qu.:    0.0
##  Median :   0.00   Median :   19.0   Median :   0.00   Median :    7.0
##  Mean   :  28.45   Mean   :  623.5   Mean   :  36.74   Mean   :  277.4
##  3rd Qu.:   0.00   3rd Qu.:  135.0   3rd Qu.:   1.00   3rd Qu.:   62.0
##  Max.   :5301.00   Max.   :35950.0   Max.   :3344.00   Max.   :21093.0
##  NA's   :7328      NA's   :7294      NA's   :7400      NA's   :7030
##   conf_xdr_tx      mdrxdr_bdq_used mdrxdr_bdq_tx    mdrxdr_dlm_used
##  Min.   :   0.00   Min.   :0.000   Min.   :   0.0   Min.   :0.000
##  1st Qu.:   0.00   1st Qu.:0.000   1st Qu.:   1.0   1st Qu.:0.000
##  Median :   0.00   Median :0.000   Median :   6.0   Median :0.000
##  Mean   :  33.88   Mean   :0.324   Mean   : 128.5   Mean   :0.247
##  3rd Qu.:   1.00   3rd Qu.:0.000   3rd Qu.:  22.0   3rd Qu.:0.000
##  Max.   :2882.00   Max.   :3.000   Max.   :8240.0   Max.   :3.000
##  NA's   :6774      NA's   :7285    NA's   :7928     NA's   :7685
##  mdrxdr_dlm_tx    mdr_shortreg_used mdr_shortreg_tx
##  Min.   :  0.00   Min.   :0.000     Min.   :   0.00
##  1st Qu.:  2.00   1st Qu.:0.000     1st Qu.:   5.75
##  Median :  8.00   Median :0.000     Median :  33.00
##  Mean   : 27.77   Mean   :0.354     Mean   : 116.92
##  3rd Qu.: 40.50   3rd Qu.:1.000     3rd Qu.:  80.00
##  Max.   :140.00   Max.   :3.000     Max.   :3474.00
##  NA's   :8031     NA's   :7486      NA's   :7950
##  mdr_tx_adverse_events newrel_tbhiv_flg newrel_hivtest    newrel_hivpos
##  Min.   :   0.00       Min.   :0.000    Min.   :      0   Min.   :     0
##  1st Qu.:   0.00       1st Qu.:0.000    1st Qu.:    107   1st Qu.:     2
##  Median :   0.00       Median :1.000    Median :   1894   Median :    67
##  Mean   :  72.31       Mean   :0.702    Mean   :  19310   Mean   :  2618
##  3rd Qu.:   9.00       3rd Qu.:1.000    3rd Qu.:   8428   3rd Qu.:   724
##  Max.   :3635.00       Max.   :1.000    Max.   :1271416   Max.   :157505
##  NA's   :7653          NA's   :7691     NA's   :7514      NA's   :7519
##    newrel_art          hivtest         hivtest_pos
##  Min.   :     0.0   Min.   :      0   Min.   :     0.0
##  1st Qu.:     1.0   1st Qu.:     32   1st Qu.:     1.0
##  Median :    67.5   Median :    482   Median :    28.0
##  Mean   :  2350.8   Mean   :   9528   Mean   :  2091.6
##  3rd Qu.:   581.8   3rd Qu.:   4175   3rd Qu.:   413.5
##  Max.   :133116.0   Max.   :1034712   Max.   :211128.0
##  NA's   :7572       NA's   :6019      NA's   :6048
##     hiv_cpt            hiv_art           hiv_tbscr
##  Min.   :     0.0   Min.   :     0.0   Min.   :      0.0
##  1st Qu.:     0.0   1st Qu.:     0.0   1st Qu.:     18.2
##  Median :     3.0   Median :     8.0   Median :    317.0
##  Mean   :  2029.7   Mean   :  1286.3   Mean   :  35321.0
##  3rd Qu.:   194.8   3rd Qu.:   190.5   3rd Qu.:   4697.8
##  Max.   :161561.0   Max.   :141755.0   Max.   :1324386.0
##  NA's   :6514       NA's   :6419       NA's   :7312
##     hiv_reg           hiv_ipt        hiv_reg_new      hiv_ipt_reg_all
##  Min.   :      0   Min.   :     0   Min.   :      0   Min.   :     0
##  1st Qu.:    114   1st Qu.:     0   1st Qu.:    164   1st Qu.:   110
##  Median :   1429   Median :    11   Median :    986   Median :  1158
##  Mean   :  65784   Mean   :  6011   Mean   :  33983   Mean   : 39795
##  3rd Qu.:  13659   3rd Qu.:   571   3rd Qu.:   5804   3rd Qu.: 41156
##  Max.   :4277683   Max.   :551787   Max.   :1091549   Max.   :170022
##  NA's   :7145      NA's   :7167     NA's   :7659      NA's   :8062
##   hiv_tbdetect      hiv_reg_new2
##  Min.   :      9   Min.   :      0.0
##  1st Qu.:   1086   1st Qu.:    166.2
##  Median :   3916   Median :    893.0
##  Mean   : 207674   Mean   :  24412.9
##  3rd Qu.:  50609   3rd Qu.:   5602.2
##  Max.   :1480908   Max.   :1091549.0
##  NA's   :8062      NA's   :7666
```

a)Explain why this line > mutate(key = stringr::str_replace(key, “newrel”, “new_rel”)) is necessary to properly tidy the data.

This dataset contains 1). It looks like country, iso2 and iso3 are three variables that redundantly specify the country. 2). We dont know about what all other columns are yet, but given the structure in the variables name (new_sp_m014, new_ep_m014, new_ep_f014) these are likely to be values, not variables.

In this dataset, we need to make some minor changes to fix the format of the columns name because the names are slightly inconsistent. As we seen in the statement we have newrel instead of new_rel (its difficult to spot this here but if we don’t fix it we will get the errors in subsequent steps). So, using the idea of replacing the characters “newrel” with “new_rel”. This makes all variable names consistent.

What happens if you neglect the mutate() step?

First Solution We can neglect the mutate step only if we know that all cases are new and we just parse the case type after the 3rd character. But we may not know that so better to mutate.

Second Solution

The separate() function emits the warning “too few values”. If we check the rows for keys beginning with “newrel_”, we see that sexage is missing, and type = m014.

b) How many entries are removed from the dataset when you set na.rm to true in the gather command (in this dataset). How else could those NA values be handled? Among these options, which do you think is the best way to handle those missing values for this dataset, and why?

How many entries are removed from the dataset when you set na.rm to true in the gather command (in this dataset)

To give this question answer, i would need to know more about the data generation process. There are zero’s in the data, which means they may explicitly be indicating no cases. To get the zero’s in the dataset, below is the r command.

How else could those NA values be handled? Among these options, which do you think is the best way to handle those missing values for this dataset, and why?

There are Two R functions which deal with the NA values using Fill argument.

1). In Spread(), all NA values are replaced by the fill value. The fill argument only takes in one value. 2). In complete(), all NA values are under different variables can be replaced by different values. The fill argument takes in a list that specifies the values to replace NA for different variables.

Considering the best way to handle missing values for this dataset is using Gather() and Spread(0 function because we have the count for the indiviuals columns who has TRUE(NA) and False(value) counts. Now, these missing value could be informative. After analysing the dataset, I have found that most countries have loads of missing values ! we can decide to remove all the missing values from dataset using readdata very easily with na.omit(). In the following commands, I showed the whole process for getting the NA values and omiting the NA values.

c) Explain the difference between an explicit and implicit missing value, in general. Can you find any implicit missing values in this dataset, if so where?

In the dataset, a value can be missing in the two possible ways. 1). Explicitly which means dataset flagged with “NA” values which we have in this dataset as i showed in the previous example. 2). Implicitly which means simply nothing present in the dataset. for example, it could be one or more empty row or has zero in the country column in this dataset.

Implicity missing values in this dataset

d) Looking at the features (country, year, var, sex, age, cases) in the tidied data, are they all appropriately typed? Are there any features you think would be better suited as a different type? Why or why not?

e) Explain in your own words what a gather operation is, and give an example of a situation when it might be useful. Do the same for spread.

Gather operation will take multiple columns and collapse them into key-value pairs, duplicating all other column needed.

Spread operation function spreads a key-value pair across multiple columns.

f) Generate an informative visualization, which shows something about the data. Give a brief description of what it shows, and why you thought it was interesting.