Preprocessing your data with R

Susan Holmes (c)

Missing data.

```
?is.na
example <- c("A",1,6,7,NA,"B")
example
```

```
## [1] "A" "1" "6" "7" NA  "B"
```

```
mean(example)
```

```
## Warning in mean.default(example): argument is not numeric or logical:
## returning NA
```

```
## [1] NA
```

```
is.na(example)
```

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

```
example2 <- c(2,1,6,7,NA,4)
example2
```

```
## [1]  2  1  6  7 NA  4
```

```
is.na(example2)
```

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

```
mean(example2)
```

```
## [1] NA
```

```
length(example2)
```

```
## [1] 6
```

```
mean(example2,na.rm=TRUE)
```

```
## [1] 4
```

```
median(example2,na.rm=TRUE)
```

```
## [1] 4
```

Replacing just the missing values

```
example3 <- example2
example3
```

```
## [1]  2  1  6  7 NA  4
```

```
example3[is.na(example3)] <- 0
example3
```

```
## [1] 2 1 6 7 0 4
```

```
length(example3)
```

```
## [1] 6
```

Missing values can behave strangely

```
NA == NA
```

```
## [1] NA
```

```
NA+8
```

```
## [1] NA
```

```
NA^0
```

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

```
1/NA
```

```
## [1] NA
```

Other strange values…

```r
1/0
```

```
## [1] Inf
```

```r
1/0-1/0
```

```
## [1] NaN
```

Data imputation

```
library("mice")
data(mammalsleep)
?mammalsleep
dim(mammalsleep)
```

```
## [1] 62 11
```

```
nic(mammalsleep)
```

```
## [1] 20
```

```
md.pattern(mammalsleep)
```

```
##    species bw brw pi sei odi ts mls gt ps sws
## 42       1  1   1  1   1   1  1   1  1  1   1  0
##  2       1  1   1  1   1   1  1   0  1  1   1  1
##  3       1  1   1  1   1   1  1   1  0  1   1  1
##  9       1  1   1  1   1   1  1   1  1  0   0  2
##  2       1  1   1  1   1   1  0   1  1  1   0  2
##  1       1  1   1  1   1   1  1   0  0  1   1  2
##  2       1  1   1  1   1   1  0   1  1  0   0  3
##  1       1  1   1  1   1   1  1   0  1  0   0  3
##         0  0   0  0   0   0  4   4  4 12  14 38
```

Missing at random (MCAR) versus systematic patterns (MNAR).

```
?mice
```

Outlier detection

```
summary(mammalsleep)
```

```
##                          species         bw                brw
##  African elephant           : 1   Min.   :   0.005   Min.   :   0.14
##  African giant pouched rat: 1   1st Qu.:   0.600   1st Qu.:   4.25
##  Arctic Fox                 : 1   Median :   3.342   Median :  17.25
##  Arctic ground squirrel     : 1   Mean   : 198.790   Mean   : 283.13
##  Asian elephant             : 1   3rd Qu.:  48.203   3rd Qu.: 166.00
##  Baboon                     : 1   Max.   :6654.000   Max.   :5712.00
##  (Other)                    :56
##       sws               ps              ts             mls
##  Min.   : 2.100   Min.   :0.000   Min.   : 2.60   Min.   :  2.000
##  1st Qu.: 6.250   1st Qu.:0.900   1st Qu.: 8.05   1st Qu.:  6.625
##  Median : 8.350   Median :1.800   Median :10.45   Median : 15.100
##  Mean   : 8.673   Mean   :1.972   Mean   :10.53   Mean   : 19.878
##  3rd Qu.:11.000   3rd Qu.:2.550   3rd Qu.:13.20   3rd Qu.: 27.750
##  Max.   :17.900   Max.   :6.600   Max.   :19.90   Max.   :100.000
##  NA's   :14       NA's   :12      NA's   :4       NA's   :4
##        gt               pi             sei             odi
##  Min.   : 12.00   Min.   :1.000   Min.   :1.000   Min.   :1.000
##  1st Qu.: 35.75   1st Qu.:2.000   1st Qu.:1.000   1st Qu.:1.000
##  Median : 79.00   Median :3.000   Median :2.000   Median :2.000
##  Mean   :142.35   Mean   :2.871   Mean   :2.419   Mean   :2.613
##  3rd Qu.:207.50   3rd Qu.:4.000   3rd Qu.:4.000   3rd Qu.:4.000
##  Max.   :645.00   Max.   :5.000   Max.   :5.000   Max.   :5.000
##  NA's   :4
```

```
which.max(mammalsleep$bw)
```

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

```
mammalsleep[which.max(mammalsleep$bw),]
```

```
##            species   bw  brw sws ps  ts  mls  gt pi sei odi
## 1 African elephant 6654 5712  NA NA 3.3 38.6 645  3   5   3
```

Document them, find the reason they occurred, then remove them.

Make the data easier to look at interactively

```
View(pressure)
View(iris)
```

Grouping Data

```
load('births.RData')
head(births)
```

```
##   year month date_of_month day_of_week births
## 1 2000     1             1           6   9083
## 2 2000     1             2           7   8006
## 3 2000     1             3           1  11363
## 4 2000     1             4           2  13032
## 5 2000     1             5           3  12558
## 6 2000     1             6           4  12466
```

```
birthn <- births
save(birthn,file="birthn.RData")
```

Different ways of filtering the data

Choosing only the Saturday births.

```
###Subsetting
Sat <-birthn[birthn$day_of_week==6,]
Sat[1:5,]
```

```
##    year month date_of_month day_of_week births
## 1  2000     1             1           6   9083
## 8  2000     1             8           6   8934
## 15 2000     1            15           6   8525
## 22 2000     1            22           6   8855
## 29 2000     1            29           6   8805
```

```
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
```

```
Sat1 <-filter(birthn, day_of_week == 6)
Sat1[1:5,]
```

```
##   year month date_of_month day_of_week births
## 1 2000     1             1           6   9083
## 2 2000     1             8           6   8934
## 3 2000     1            15           6   8525
## 4 2000     1            22           6   8855
## 5 2000     1            29           6   8805
```

```
Sat2 <- birthn %>% filter(day_of_week == 6)
Sat2[1:5,]
```

```
##   year month date_of_month day_of_week births
## 1 2000     1             1           6   9083
## 2 2000     1             8           6   8934
## 3 2000     1            15           6   8525
## 4 2000     1            22           6   8855
## 5 2000     1            29           6   8805
```

Another way of looking at data is to make them into what is called a tibble: (tbl: tibble).

`tbl` s have the advantage of always showing themselves in the console optimally.

`tbl_df` gives similar information as `str` we have been using.

```
tbl_df(Sat1)
```

```
## # A tibble: 783 × 5
##     year month date_of_month day_of_week births
##    <int> <int>         <int>       <int>  <int>
## 1   2000     1             1           6   9083
## 2   2000     1             8           6   8934
## 3   2000     1            15           6   8525
## 4   2000     1            22           6   8855
## 5   2000     1            29           6   8805
## 6   2000     2             5           6   8624
## 7   2000     2            12           6   8836
## 8   2000     2            19           6   8861
## 9   2000     2            26           6   9026
## 10  2000     3             4           6   9054
## # ... with 773 more rows
```

```
str(Sat1)
```

```
## 'data.frame':    783 obs. of  5 variables:
##  $ year         : int  2000 2000 2000 2000 2000 2000 2000 2000 2000 2000 ...
##  $ month        : int  1 1 1 1 1 2 2 2 2 3 ...
##  $ date_of_month: int  1 8 15 22 29 5 12 19 26 4 ...
##  $ day_of_week  : int  6 6 6 6 6 6 6 6 6 6 ...
##  $ births       : int  9083 8934 8525 8855 8805 8624 8836 8861 9026 9054 ...
```

Sequences of Transformations

The %>% operator helps when we are doing several nested operations.

Here is an example

```
GroupBirths <- group_by(birthn,day_of_week)
GroupMeans <- summarise(GroupBirths,mean(births))
SortedBirths <- arrange(GroupMeans, `mean(births)`)
SortedBirths
```

```
## # A tibble: 7 × 2
##   day_of_week `mean(births)`
##         <int>          <dbl>
## 1           7       7518.377
## 2           6       8562.573
## 3           1      11897.830
## 4           5      12596.162
## 5           4      12845.826
## 6           3      12910.766
## 7           2      13122.444
```

```
str(SortedBirths)
```

```
## Classes 'tbl_df', 'tbl' and 'data.frame':    7 obs. of  2 variables:
##  $ day_of_week : int  7 6 1 5 4 3 2
##  $ mean(births): num  7518 8563 11898 12596 12846 ...
```

```
birthn %>%
 group_by(day_of_week) %>%
 summarise(avg = mean(births)) %>%
 arrange(avg)
```

```
## # A tibble: 7 × 2
##   day_of_week       avg
##         <int>     <dbl>
## 1           7  7518.377
## 2           6  8562.573
## 3           1 11897.830
## 4           5 12596.162
## 5           4 12845.826
## 6           3 12910.766
## 7           2 13122.444
```

```
####More succintly
birthn %>%
 group_by(day_of_week) %>%
 summarise(mean(births)) %>%
 arrange()
```

```
## # A tibble: 7 × 2
##   day_of_week `mean(births)`
##         <int>          <dbl>
## 1           1      11897.830
## 2           2      13122.444
```

```
## 3            3      12910.766
## 4            4      12845.826
## 5            5      12596.162
## 6            6       8562.573
## 7            7       7518.377
```

x %>% f(y) is equivalent to just executing f(x,y)

If we need to execute a sequence of functions: h(g(f(x,y),z),m) can be hard to parse and read.

x %>% f(y) %>% g(z) %>% h(m) gives the same answer.

To find out the average of Friday 13th births:

```
birthn %>%
  filter(day_of_week == 5) %>%
  filter(date_of_month == 13) %>%
  summarise(mean(births))
```

```
##   mean(births)
## 1     11949.96
```

```
birthn %>%
  filter(day_of_week < 5) %>%
  filter(date_of_month != 13) %>%
  summarise(mean(births))
```

```
##   mean(births)
## 1     12700.61
```

Bad Drivers Data

Five Thirty Eight Article

You need an internet connection for this to work:

```
drivers <- read.csv(url("https://raw.githubusercontent.com/fivethirtyeight/data/master/bad-drivers/bad-drivers.csv"))
head(drivers)
```

```
##        State
## 1    Alabama
## 2     Alaska
## 3    Arizona
## 4   Arkansas
## 5 California
## 6   Colorado
##   Number.of.drivers.involved.in.fatal.collisions.per.billion.miles
## 1                                                             18.8
## 2                                                             18.1
## 3                                                             18.6
## 4                                                             22.4
## 5                                                             12.0
## 6                                                             13.6
##   Percentage.Of.Drivers.Involved.In.Fatal.Collisions.Who.Were.Speeding
## 1                                                                   39
## 2                                                                   41
## 3                                                                   35
## 4                                                                   18
## 5                                                                   35
## 6                                                                   37
##   Percentage.Of.Drivers.Involved.In.Fatal.Collisions.Who.Were.Alcohol.Impaired
## 1                                                                           30
## 2                                                                           25
## 3                                                                           28
## 4                                                                           26
## 5                                                                           28
## 6                                                                           28
##   Percentage.Of.Drivers.Involved.In.Fatal.Collisions.Who.Were.Not.Distracted
## 1                                                                         96
## 2                                                                         90
## 3                                                                         84
## 4                                                                         94
## 5                                                                         91
## 6                                                                         79
##   Percentage.Of.Drivers.Involved.In.Fatal.Collisions.Who.Had.Not.Been.Involved.In.Any.Previous.Accidents
## 1                                                                                                     80
## 2                                                                                                     94
## 3                                                                                                     96
## 4                                                                                                     95
## 5                                                                                                     89
## 6                                                                                                     95
##   Car.Insurance.Premiums....
## 1                     784.55
## 2                    1053.48
## 3                     899.47
## 4                     827.34
## 5                     878.41
## 6                     835.50
##   Losses.incurred.by.insurance.companies.for.collisions.per.insured.driver....
## 1                                                                       145.08
## 2                                                                       133.93
## 3                                                                       110.35
## 4                                                                       142.39
```

```
## 5                                                                  165.63
## 6                                                                  139.91
```

```
tbl_df(drivers)
```

```
## # A tibble: 51 × 8
##                   State
##                  <fctr>
## 1               Alabama
## 2                Alaska
## 3               Arizona
## 4              Arkansas
## 5            California
## 6              Colorado
## 7           Connecticut
## 8              Delaware
## 9  District of Columbia
## 10              Florida
## # ... with 41 more rows, and 7 more variables:
## #   Number.of.drivers.involved.in.fatal.collisions.per.billion.miles <dbl>,
## #   Percentage.Of.Drivers.Involved.In.Fatal.Collisions.Who.Were.Speeding <int>,
## #   Percentage.Of.Drivers.Involved.In.Fatal.Collisions.Who.Were.Alcohol.Impaired <int>,
## #   Percentage.Of.Drivers.Involved.In.Fatal.Collisions.Who.Were.Not.Distracted <int>,
## #   Percentage.Of.Drivers.Involved.In.Fatal.Collisions.Who.Had.Not.Been.Involved.In.Any.Previous.Accident
## #   Car.Insurance.Premiums.... <dbl>,
## #   Losses.incurred.by.insurance.companies.for.collisions.per.insured.driver.... <dbl>
```

```
glimpse(drivers)
```

```
## Observations: 51
## Variables: 8
## $ State
## $ Number.of.drivers.involved.in.fatal.collisions.per.billion.miles
## $ Percentage.Of.Drivers.Involved.In.Fatal.Collisions.Who.Were.Speeding
## $ Percentage.Of.Drivers.Involved.In.Fatal.Collisions.Who.Were.Alcohol.Impaired
## $ Percentage.Of.Drivers.Involved.In.Fatal.Collisions.Who.Were.Not.Distracted
## $ Percentage.Of.Drivers.Involved.In.Fatal.Collisions.Who.Had.Not.Been.Involved.In.Any.Previous.Accidents
## $ Car.Insurance.Premiums....
## $ Losses.incurred.by.insurance.companies.for.collisions.per.insured.driver....
```

```
summary(drivers)
```

```
##          State
##  Alabama   : 1
##  Alaska    : 1
##  Arizona   : 1
##  Arkansas  : 1
##  California: 1
##  Colorado  : 1
##  (Other)   :45
##  Number.of.drivers.involved.in.fatal.collisions.per.billion.miles
##  Min.   : 5.90
##  1st Qu.:12.75
##  Median :15.60
##  Mean   :15.79
##  3rd Qu.:18.50
##  Max.   :23.90
##
##  Percentage.Of.Drivers.Involved.In.Fatal.Collisions.Who.Were.Speeding
##  Min.   :13.00
##  1st Qu.:23.00
##  Median :34.00
##  Mean   :31.73
```

```
##   3rd Qu.:38.00
##   Max.   :54.00
##
##   Percentage.Of.Drivers.Involved.In.Fatal.Collisions.Who.Were.Alcohol.Impaired
##   Min.   :16.00
##   1st Qu.:28.00
##   Median :30.00
##   Mean   :30.69
##   3rd Qu.:33.00
##   Max.   :44.00
##
##   Percentage.Of.Drivers.Involved.In.Fatal.Collisions.Who.Were.Not.Distracted
##   Min.   : 10.00
##   1st Qu.: 83.00
##   Median : 88.00
##   Mean   : 85.92
##   3rd Qu.: 95.00
##   Max.   :100.00
##
##   Percentage.Of.Drivers.Involved.In.Fatal.Collisions.Who.Had.Not.Been.Involved.In.Any.Previous.Accidents
##   Min.   : 76.00
##   1st Qu.: 83.50
##   Median : 88.00
##   Mean   : 88.73
##   3rd Qu.: 95.00
##   Max.   :100.00
##
##   Car.Insurance.Premiums....
##   Min.   : 642.0
##   1st Qu.: 768.4
##   Median : 859.0
##   Mean   : 887.0
##   3rd Qu.:1007.9
##   Max.   :1301.5
##
##   Losses.incurred.by.insurance.companies.for.collisions.per.insured.driver....
##   Min.   : 82.75
##   1st Qu.:114.64
##   Median :136.05
##   Mean   :134.49
##   3rd Qu.:151.87
##   Max.   :194.78
##
```

```
colnames(drivers)=
  c("State","NperB","PrcSpeed","PrcAlco","PrcNotDist","PrcNoPrev","Premium","Loss")
sort(drivers[,2])
```

```
##  [1]  5.9  8.2  9.6 10.6 10.8 11.1 11.2 11.3 11.6 12.0 12.3 12.5 12.7 12.8
## [15] 12.8 13.6 13.6 13.8 14.1 14.1 14.5 14.7 14.9 15.1 15.3 15.6 15.7 16.1
## [29] 16.2 16.8 17.4 17.5 17.6 17.8 17.9 18.1 18.2 18.4 18.6 18.8 19.4 19.4
## [43] 19.5 19.9 20.5 21.4 21.4 22.4 23.8 23.9 23.9
```

```
drivers[1:10,1:3]
```

```
##                   State NperB PrcSpeed
## 1               Alabama  18.8       39
## 2                Alaska  18.1       41
## 3               Arizona  18.6       35
## 4              Arkansas  22.4       18
## 5            California  12.0       35
## 6              Colorado  13.6       37
## 7           Connecticut  10.8       46
## 8              Delaware  16.2       38
```

```
## 9  District of Columbia   5.9       34
## 10              Florida  17.9       21
```

```
drivers[order(drivers[,2]),1:3]
```

```
##                    State NperB PrcSpeed
## 9  District of Columbia   5.9       34
## 22        Massachusetts   8.2       23
## 24            Minnesota   9.6       23
## 48           Washington  10.6       42
## 7           Connecticut  10.8       46
## 40         Rhode Island  11.1       34
## 31           New Jersey  11.2       16
## 45                 Utah  11.3       43
## 30        New Hampshire  11.6       35
## 5            California  12.0       35
## 33             New York  12.3       32
## 21             Maryland  12.5       34
## 47             Virginia  12.7       19
## 14             Illinois  12.8       36
## 38               Oregon  12.8       33
## 6              Colorado  13.6       37
## 46              Vermont  13.6       30
## 50            Wisconsin  13.8       36
## 23             Michigan  14.1       24
## 36                 Ohio  14.1       28
## 15              Indiana  14.5       25
## 29               Nevada  14.7       37
## 28             Nebraska  14.9       13
## 20                Maine  15.1       38
## 13                Idaho  15.3       36
## 11              Georgia  15.6       19
## 16                 Iowa  15.7       17
## 26             Missouri  16.1       43
## 8              Delaware  16.2       38
## 34       North Carolina  16.8       39
## 51              Wyoming  17.4       42
## 12               Hawaii  17.5       54
## 25          Mississippi  17.6       15
## 17               Kansas  17.8       27
## 10              Florida  17.9       21
## 2                Alaska  18.1       41
## 39         Pennsylvania  18.2       50
## 32           New Mexico  18.4       19
## 3               Arizona  18.6       35
## 1               Alabama  18.8       39
## 42         South Dakota  19.4       31
## 44                Texas  19.4       40
## 43            Tennessee  19.5       21
## 37             Oklahoma  19.9       32
## 19            Louisiana  20.5       35
## 18             Kentucky  21.4       19
## 27              Montana  21.4       39
## 4              Arkansas  22.4       18
## 49        West Virginia  23.8       34
## 35         North Dakota  23.9       23
## 41       South Carolina  23.9       38
```

```
arrange(drivers,NperB)
```

```
##                    State NperB PrcSpeed PrcAlco PrcNotDist PrcNoPrev
## 1  District of Columbia   5.9       34      27        100       100
## 2         Massachusetts   8.2       23      35         87        80
## 3             Minnesota   9.6       23      29         88        88
## 4            Washington  10.6       42      33         82        86
```

```
## 5          Connecticut 10.8        46      36          87        82
## 6         Rhode Island 11.1        34      38          92        79
## 7           New Jersey 11.2        16      28          86        78
## 8                 Utah 11.3        43      16          88        96
## 9        New Hampshire 11.6        35      30          87        83
## 10          California 12.0        35      28          91        89
## 11            New York 12.3        32      29          88        80
## 12            Maryland 12.5        34      32          71        99
## 13            Virginia 12.7        19      27          87        88
## 14            Illinois 12.8        36      34          94        96
## 15              Oregon 12.8        33      26          67        90
## 16            Colorado 13.6        37      28          79        95
## 17             Vermont 13.6        30      30          96        95
## 18           Wisconsin 13.8        36      33          39        84
## 19            Michigan 14.1        24      28          95        77
## 20                Ohio 14.1        28      34          99        82
## 21             Indiana 14.5        25      29          95        95
## 22              Nevada 14.7        37      32          95        99
## 23            Nebraska 14.9        13      35          93        90
## 24               Maine 15.1        38      30          87        84
## 25               Idaho 15.3        36      29          85        98
## 26             Georgia 15.6        19      25          95        93
## 27                Iowa 15.7        17      25          97        87
## 28            Missouri 16.1        43      34          92        84
## 29            Delaware 16.2        38      30          87        99
## 30      North Carolina 16.8        39      31          94        81
## 31             Wyoming 17.4        42      32          81        90
## 32              Hawaii 17.5        54      41          82        87
## 33         Mississippi 17.6        15      31          10       100
## 34              Kansas 17.8        27      24          77        85
## 35             Florida 17.9        21      29          92        94
## 36              Alaska 18.1        41      25          90        94
## 37        Pennsylvania 18.2        50      31          96        88
## 38          New Mexico 18.4        19      27          67        98
## 39             Arizona 18.6        35      28          84        96
## 40             Alabama 18.8        39      30          96        80
## 41        South Dakota 19.4        31      33          98        86
## 42               Texas 19.4        40      38          91        87
## 43           Tennessee 19.5        21      29          82        81
## 44            Oklahoma 19.9        32      29          92        94
## 45           Louisiana 20.5        35      33          73        98
## 46            Kentucky 21.4        19      23          78        76
## 47             Montana 21.4        39      44          84        85
## 48            Arkansas 22.4        18      26          94        95
## 49       West Virginia 23.8        34      28          97        87
## 50        North Dakota 23.9        23      42          99        86
## 51      South Carolina 23.9        38      41          96        81
##    Premium   Loss
## 1  1273.89 136.05
## 2  1011.14 135.63
## 3   777.18 133.35
## 4   890.03 111.62
## 5  1068.73 167.02
## 6  1148.99 148.58
## 7  1301.52 159.85
## 8   809.38 109.48
## 9   746.54 120.21
## 10  878.41 165.63
## 11 1234.31 150.01
## 12 1048.78 192.70
## 13  768.95 153.72
## 14  803.11 139.15
## 15  804.71 104.61
## 16  835.50 139.91
## 17  716.20 109.61
## 18  670.31 106.62
## 19 1110.61 152.26
## 20  697.73 133.52
## 21  710.46 108.92
## 22 1029.87 138.71
```

```
## 23  732.28 114.82
## 24  661.88  96.57
## 25  641.96  82.75
## 26  913.15 142.80
## 27  649.06 114.47
## 28  790.32 144.45
## 29 1137.87 151.48
## 30  708.24 127.82
## 31  791.14 122.04
## 32  861.18 120.92
## 33  896.07 155.77
## 34  780.45 133.80
## 35 1160.13 144.18
## 36 1053.48 133.93
## 37  905.99 153.86
## 38  869.85 120.75
## 39  899.47 110.35
## 40  784.55 145.08
## 41  669.31  96.87
## 42 1004.75 156.83
## 43  767.91 155.57
## 44  881.51 178.86
## 45 1281.55 194.78
## 46  872.51 137.13
## 47  816.21  85.15
## 48  827.34 142.39
## 49  992.61 152.56
## 50  688.75 109.72
## 51  858.97 116.29
```

```
arrange(drivers,desc(PrcSpeed))
```

```
##                   State NperB PrcSpeed PrcAlco PrcNotDist PrcNoPrev
## 1                Hawaii  17.5       54      41         82        87
## 2          Pennsylvania  18.2       50      31         96        88
## 3           Connecticut  10.8       46      36         87        82
## 4              Missouri  16.1       43      34         92        84
## 5                  Utah  11.3       43      16         88        96
## 6            Washington  10.6       42      33         82        86
## 7               Wyoming  17.4       42      32         81        90
## 8                Alaska  18.1       41      25         90        94
## 9                 Texas  19.4       40      38         91        87
## 10              Alabama  18.8       39      30         96        80
## 11              Montana  21.4       39      44         84        85
## 12       North Carolina  16.8       39      31         94        81
## 13             Delaware  16.2       38      30         87        99
## 14                Maine  15.1       38      30         87        84
## 15       South Carolina  23.9       38      41         96        81
## 16             Colorado  13.6       37      28         79        95
## 17               Nevada  14.7       37      32         95        99
## 18                Idaho  15.3       36      29         85        98
## 19             Illinois  12.8       36      34         94        96
## 20            Wisconsin  13.8       36      33         39        84
## 21              Arizona  18.6       35      28         84        96
## 22           California  12.0       35      28         91        89
## 23            Louisiana  20.5       35      33         73        98
## 24        New Hampshire  11.6       35      30         87        83
## 25 District of Columbia   5.9       34      27        100       100
## 26             Maryland  12.5       34      32         71        99
## 27         Rhode Island  11.1       34      38         92        79
## 28        West Virginia  23.8       34      28         97        87
## 29               Oregon  12.8       33      26         67        90
## 30             New York  12.3       32      29         88        80
## 31             Oklahoma  19.9       32      29         92        94
## 32         South Dakota  19.4       31      33         98        86
## 33              Vermont  13.6       30      30         96        95
## 34                 Ohio  14.1       28      34         99        82
```

```
## 35         Kansas 17.8    27    24     77      85
## 36        Indiana 14.5    25    29     95      95
## 37       Michigan 14.1    24    28     95      77
## 38  Massachusetts  8.2    23    35     87      80
## 39      Minnesota  9.6    23    29     88      88
## 40   North Dakota 23.9    23    42     99      86
## 41        Florida 17.9    21    29     92      94
## 42      Tennessee 19.5    21    29     82      81
## 43        Georgia 15.6    19    25     95      93
## 44       Kentucky 21.4    19    23     78      76
## 45     New Mexico 18.4    19    27     67      98
## 46       Virginia 12.7    19    27     87      88
## 47       Arkansas 22.4    18    26     94      95
## 48           Iowa 15.7    17    25     97      87
## 49     New Jersey 11.2    16    28     86      78
## 50    Mississippi 17.6    15    31     10     100
## 51       Nebraska 14.9    13    35     93      90
##    Premium   Loss
## 1   861.18 120.92
## 2   905.99 153.86
## 3  1068.73 167.02
## 4   790.32 144.45
## 5   809.38 109.48
## 6   890.03 111.62
## 7   791.14 122.04
## 8  1053.48 133.93
## 9  1004.75 156.83
## 10  784.55 145.08
## 11  816.21  85.15
## 12  708.24 127.82
## 13 1137.87 151.48
## 14  661.88  96.57
## 15  858.97 116.29
## 16  835.50 139.91
## 17 1029.87 138.71
## 18  641.96  82.75
## 19  803.11 139.15
## 20  670.31 106.62
## 21  899.47 110.35
## 22  878.41 165.63
## 23 1281.55 194.78
## 24  746.54 120.21
## 25 1273.89 136.05
## 26 1048.78 192.70
## 27 1148.99 148.58
## 28  992.61 152.56
## 29  804.71 104.61
## 30 1234.31 150.01
## 31  881.51 178.86
## 32  669.31  96.87
## 33  716.20 109.61
## 34  697.73 133.52
## 35  780.45 133.80
## 36  710.46 108.92
## 37 1110.61 152.26
## 38 1011.14 135.63
## 39  777.18 133.35
## 40  688.75 109.72
## 41 1160.13 144.18
## 42  767.91 155.57
## 43  913.15 142.80
## 44  872.51 137.13
## 45  869.85 120.75
## 46  768.95 153.72
## 47  827.34 142.39
## 48  649.06 114.47
## 49 1301.52 159.85
## 50  896.07 155.77
## 51  732.28 114.82
```

Make new variables

```
driversp=mutate(drivers,prem_c=Loss/Premium)
select(arrange(driversp,prem_c),State,prem_c)
```

```
##                    State    prem_c
## 1                Montana 0.1043236
## 2   District of Columbia 0.1067989
## 3               New York 0.1215335
## 4                Arizona 0.1226834
## 5             New Jersey 0.1228179
## 6                Florida 0.1242792
## 7             Washington 0.1254115
## 8                 Alaska 0.1271310
## 9                  Idaho 0.1289021
## 10          Rhode Island 0.1293136
## 11                Oregon 0.1299971
## 12              Delaware 0.1331259
## 13         Massachusetts 0.1341357
## 14                Nevada 0.1346869
## 15                  Utah 0.1352640
## 16        South Carolina 0.1353831
## 17              Michigan 0.1370958
## 18            New Mexico 0.1388170
## 19                Hawaii 0.1404120
## 20          South Dakota 0.1447311
## 21                 Maine 0.1459026
## 22             Louisiana 0.1519878
## 23               Vermont 0.1530438
## 24               Indiana 0.1533091
## 25         West Virginia 0.1536958
## 26               Wyoming 0.1542584
## 27                 Texas 0.1560886
## 28           Connecticut 0.1562789
## 29               Georgia 0.1563818
## 30              Nebraska 0.1567979
## 31              Kentucky 0.1571673
## 32             Wisconsin 0.1590607
## 33          North Dakota 0.1593031
## 34         New Hampshire 0.1610229
## 35              Colorado 0.1674566
## 36          Pennsylvania 0.1698253
## 37                Kansas 0.1714396
## 38             Minnesota 0.1715819
## 39              Arkansas 0.1721058
## 40              Illinois 0.1732639
## 41           Mississippi 0.1738369
## 42                  Iowa 0.1763627
## 43        North Carolina 0.1804755
## 44              Missouri 0.1827741
## 45              Maryland 0.1837373
## 46               Alabama 0.1849213
## 47            California 0.1885566
## 48                  Ohio 0.1913634
## 49              Virginia 0.1999090
## 50             Tennessee 0.2025888
## 51              Oklahoma 0.2029018
```

Document all the changes you make using a script.

The best way to make a report is to put everything into an `.Rmd` document and then `knit` into an html file using the knitr package.

Summary of this Session:

- Careful data preprocessing is necessary at the beginning of any data exploration exercise.
- Missing data and outliers need to be identified.
- Missing data may be imputed if there are only a few in a column or row and if their occurrence patterns are random.
- We saw how to use the package `dplyr` that allows us to easily do a sequence of actions on data using the %>% operator.
- Some of the possible actions are:
  - *filter()*
  - *arrange()*
  - *select()*
  - *mutate()*
  - *summarise()*
  - *sample_n()*
- We saw that preprocessing your data should be documented with scripts you save. A good way to do this is to use RStudio's Rmd editor and html generator.

Question: Look up the RStudio data wrangling cheatsheet: R Data Wrangling Cheatsheet

Activity: Re-analyze the `drivers` data and make your own Rmd and html reports.