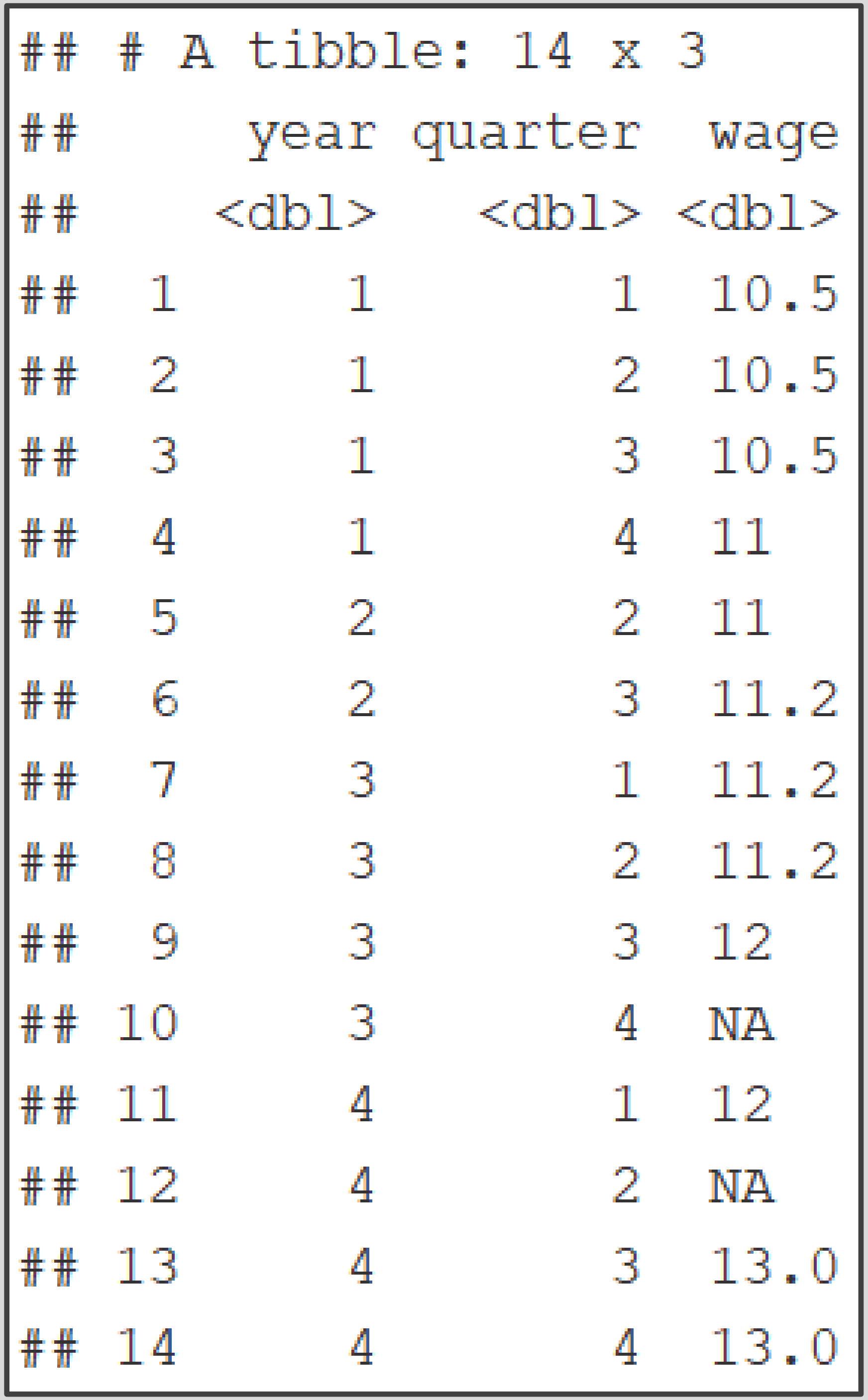
Tidy Data II

Missing Values

- Two Ways
  - Explicitly: Defined to Be Missing Using NA
  - Implicitly: Absent From Data
- There is not a Uniform Way to Handle Either of These Problems
- Rule: Either Convert All Explicitly Missing to Implicitly Missing or Convert All Implicitly Missing to Explicitly Missing

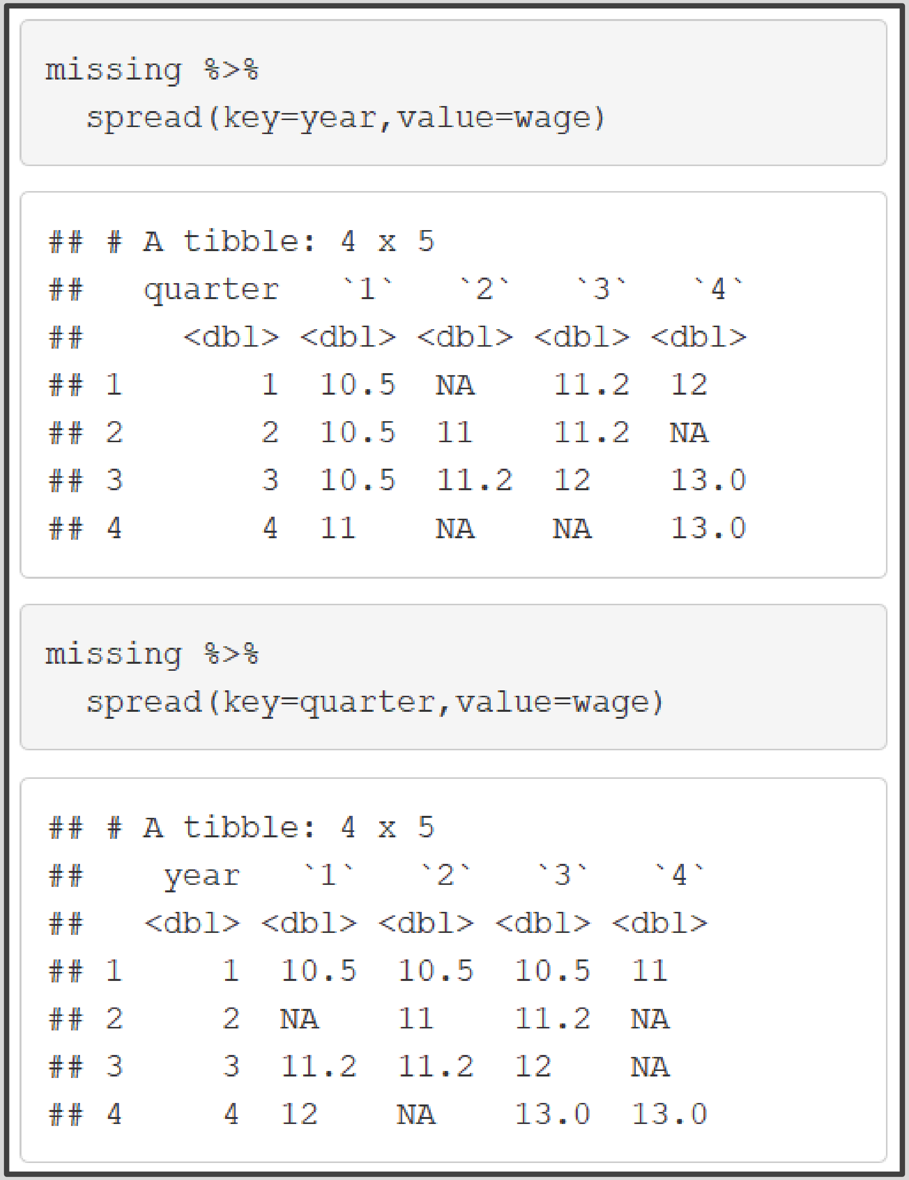
Missing Example



```
## # A tibble: 14 x 3
##       year quarter  wage
##      <dbl>   <dbl> <dbl>
##  1       1       1  10.5
##  2       1       2  10.5
##  3       1       3  10.5
##  4       1       4  11  
##  5       2       2  11  
##  6       2       3  11.2
##  7       3       1  11.2
##  8       3       2  11.2
##  9       3       3  12  
## 10       3       4  NA  
## 11       4       1  12  
## 12       4       2  NA  
## 13       4       3  13.0
## 14       4       4  13.0
```

Missing Values

- Notice:



```
missing %>%
  spread(key=year,value=wage)
```

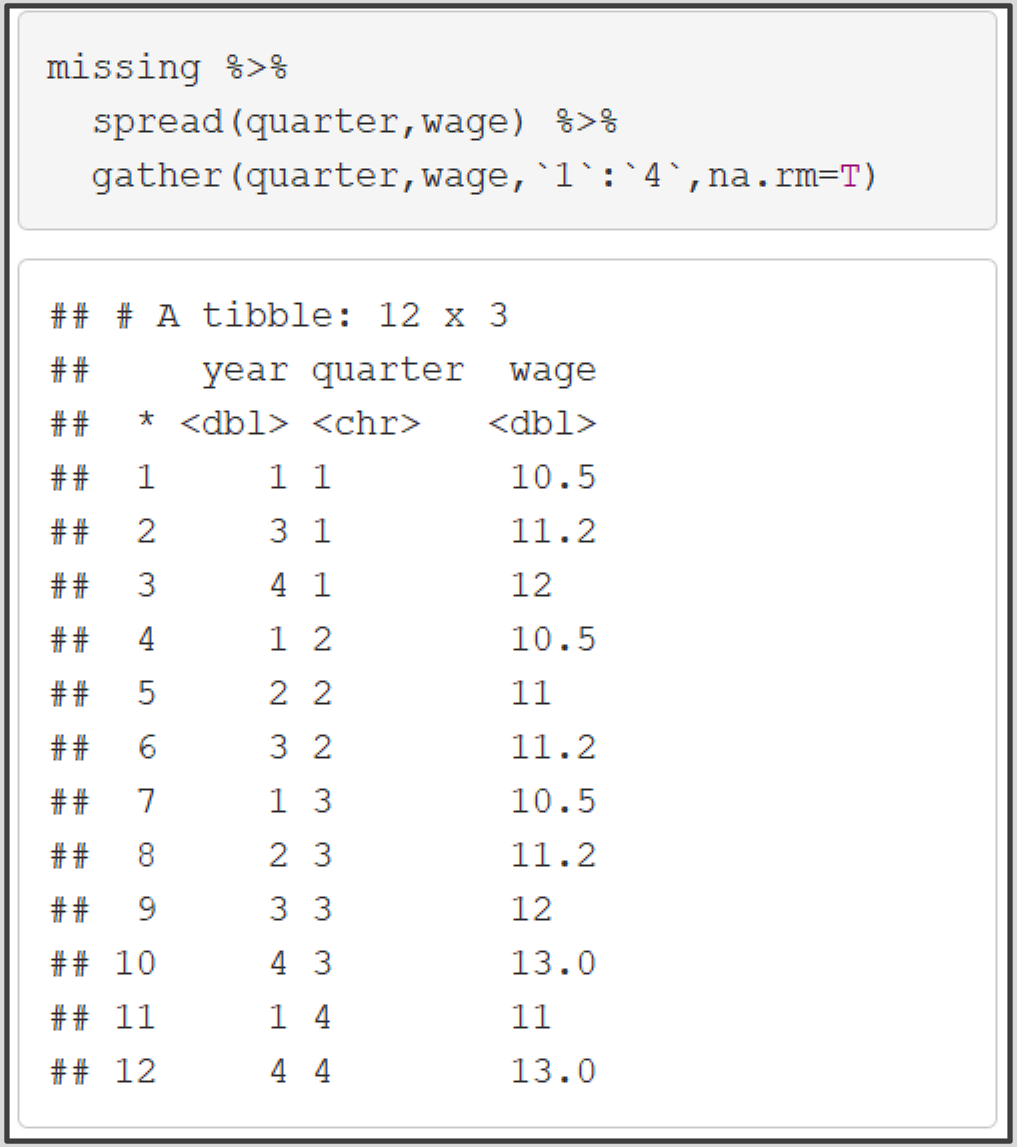
```
## # A tibble: 4 x 5
##   quarter   `1`   `2`   `3`   `4`
##     <dbl> <dbl> <dbl> <dbl> <dbl>
## 1       1  10.5  NA    11.2  12  
## 2       2  10.5  11    11.2  NA  
## 3       3  10.5  11.2  12    13.0
## 4       4  11    NA    NA    13.0
```

```
missing %>%
  spread(key=quarter,value=wage)
```

```
## # A tibble: 4 x 5
##    year   `1`   `2`   `3`   `4`
##   <dbl> <dbl> <dbl> <dbl> <dbl>
## 1     1  10.5  10.5  10.5  11  
## 2     2  NA    11    11.2  NA  
## 3     3  11.2  11.2  12    NA  
## 4     4  12    NA    13.0  13.0
```

Missing Values

- Explicit to Implicit

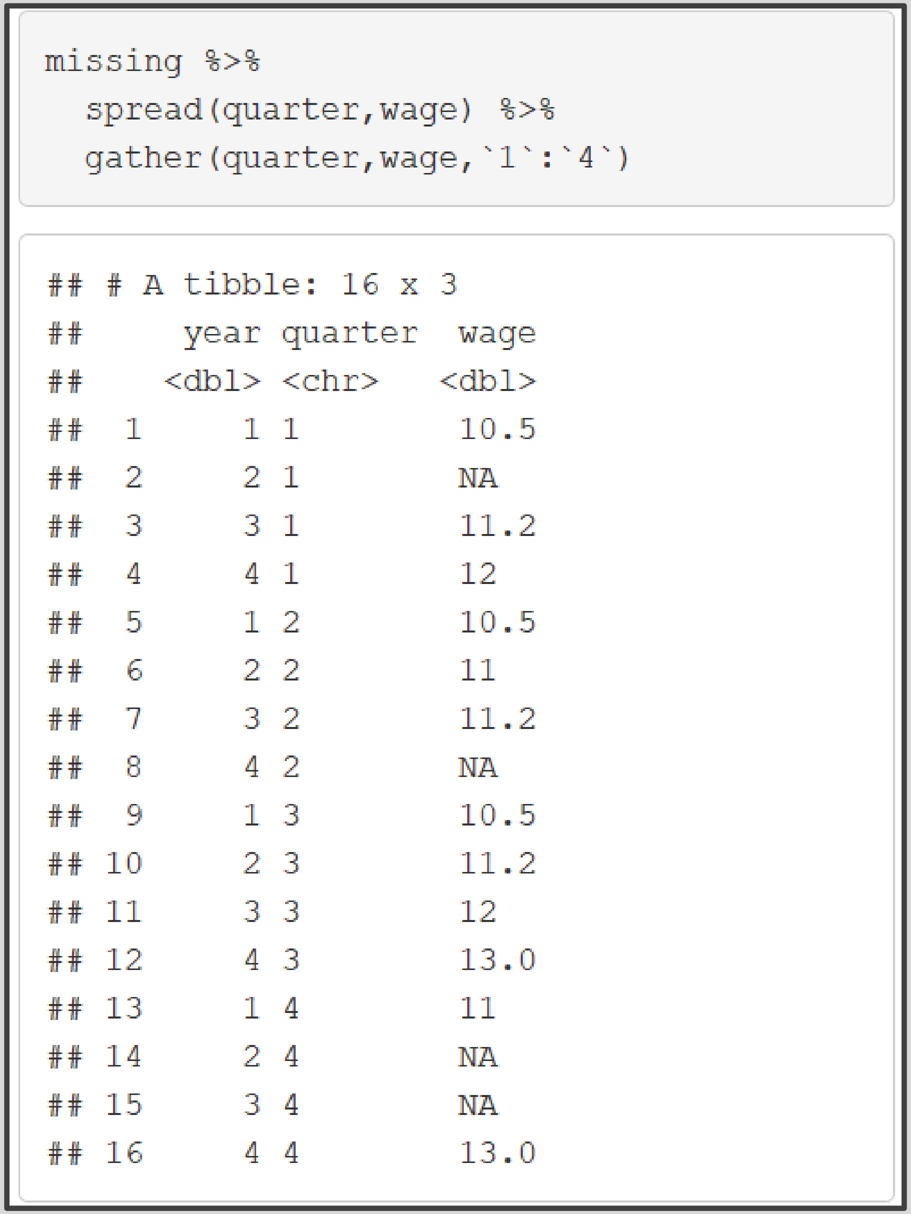


```
missing %>%
  spread(quarter,wage) %>%
  gather(quarter,wage,`1`:`4`,na.rm=T)
```

```
## # A tibble: 12 x 3
##     year quarter  wage
##   * <dbl> <chr>   <dbl>
##   1     1 1        10.5
##   2     3 1        11.2
##   3     4 1        12
##   4     1 2        10.5
##   5     2 2        11
##   6     3 2        11.2
##   7     1 3        10.5
##   8     2 3        11.2
##   9     3 3        12
## 10     4 3        13.0
## 11     1 4        11
## 12     4 4        13.0
```

Missing Values

- Implicit to Explicit

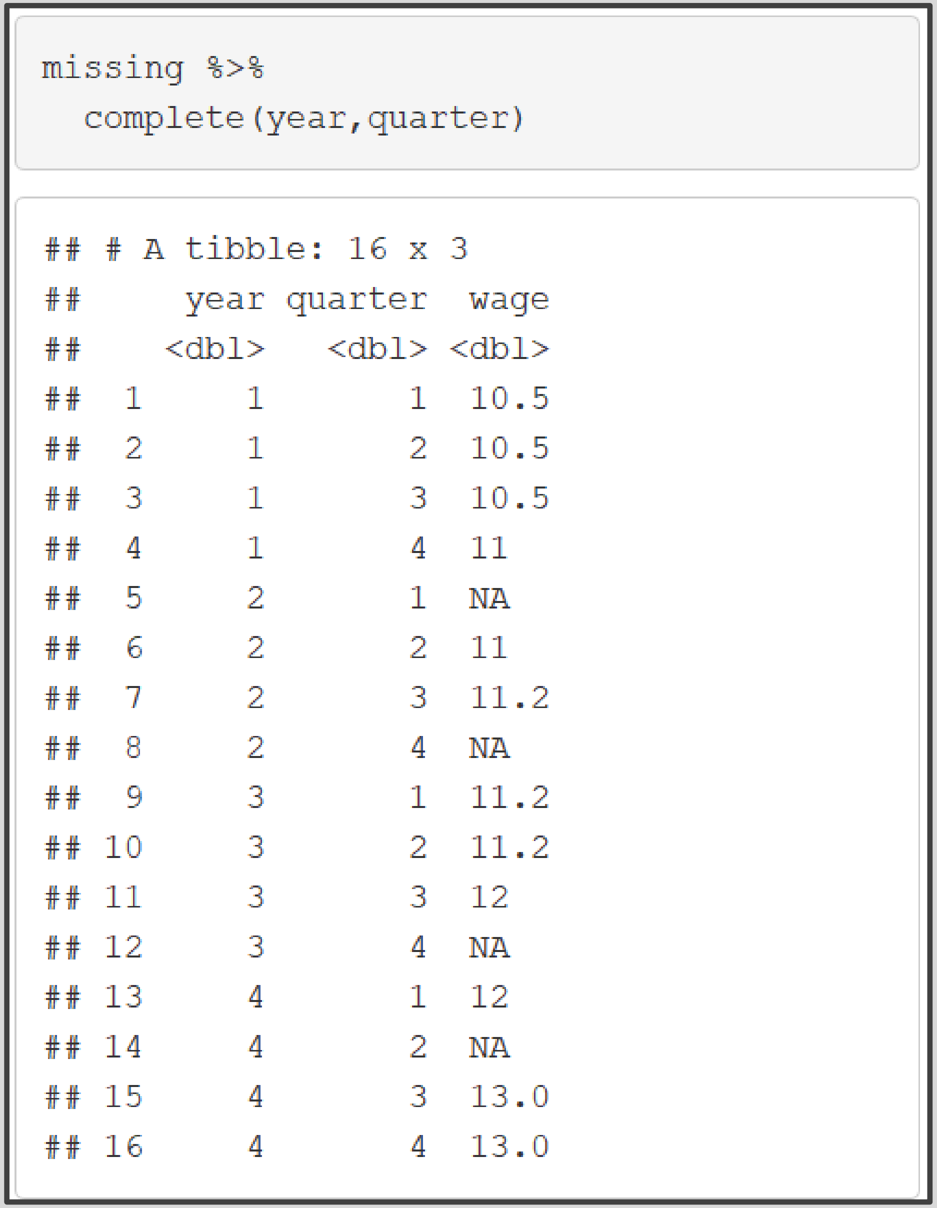


```
missing %>%
  spread(quarter,wage) %>%
  gather(quarter,wage,`1`:`4`)
```

```
## # A tibble: 16 x 3
##     year quarter  wage
##    <dbl> <chr>   <dbl>
##  1     1 1        10.5
##  2     2 1        NA
##  3     3 1        11.2
##  4     4 1        12
##  5     1 2        10.5
##  6     2 2        11
##  7     3 2        11.2
##  8     4 2        NA
##  9     1 3        10.5
## 10     2 3        11.2
## 11     3 3        12
## 12     4 3        13.0
## 13     1 4        11
## 14     2 4        NA
## 15     3 4        NA
## 16     4 4        13.0
```

Missing Values

- Complete Function



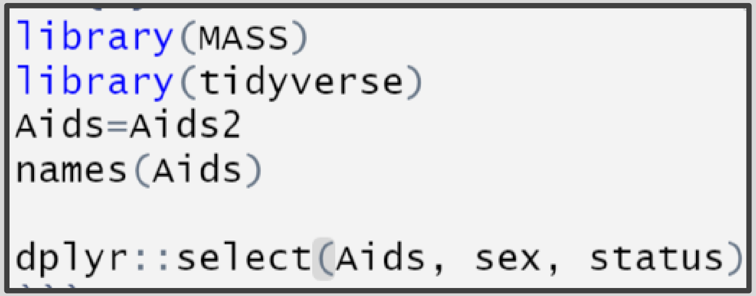
```
missing %>%
  complete(year,quarter)
```

```
## # A tibble: 16 x 3
##      year quarter  wage
##     <dbl>   <dbl> <dbl>
##  1      1       1  10.5
##  2      1       2  10.5
##  3      1       3  10.5
##  4      1       4  11
##  5      2       1  NA
##  6      2       2  11
##  7      2       3  11.2
##  8      2       4  NA
##  9      3       1  11.2
## 10      3       2  11.2
## 11      3       3  12
## 12      3       4  NA
## 13      4       1  12
## 14      4       2  NA
## 15      4       3  13.0
## 16      4       4  13.0
```

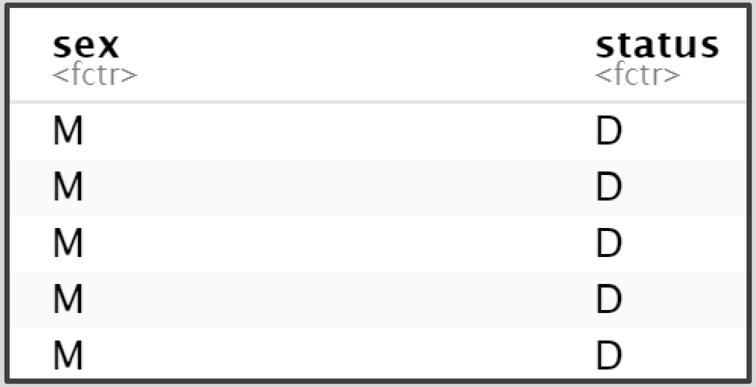
Contingency Tables

- Contingency Tables
  - Frequencies for Combination of 2 Categorical Variables
  - Relative Frequencies
  - Summarize() + Spread()

- AIDS Data from MASS Package
  - Data from 2,843 Patients



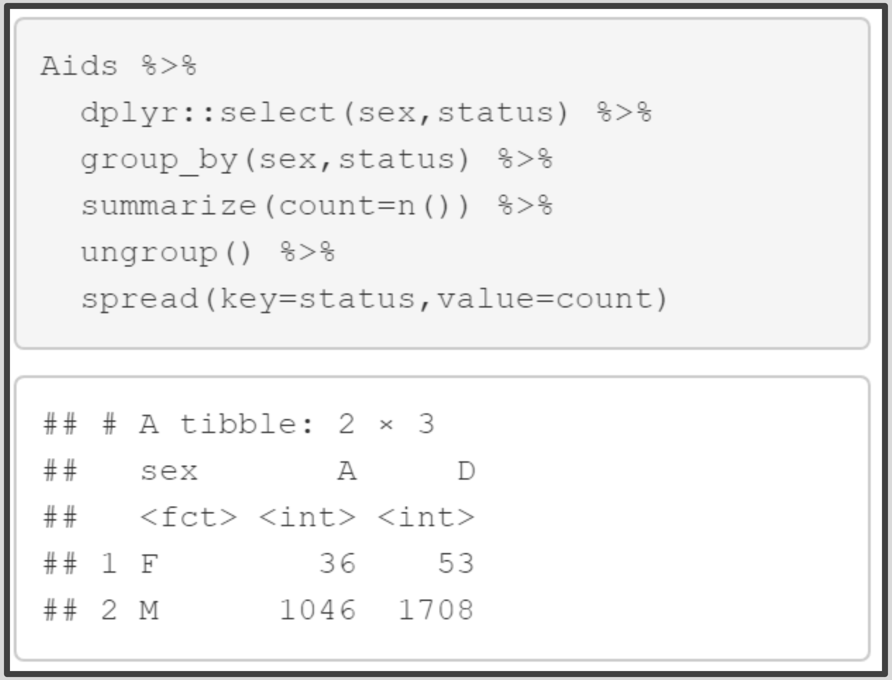
```
library(MASS)
library(tidyverse)
Aids=Aids2
names(Aids)

dplyr::select(Aids, sex, status)
```



sex <fctr>	status <fctr>
M	D
M	D
M	D
M	D
M	D

Contingency Tables

- Create Table of Frequencies
  - Used message=FALSE



```
Aids %>%
  dplyr::select(sex,status) %>%
  group_by(sex,status) %>%
  summarize(count=n()) %>%
  ungroup() %>%
  spread(key=status,value=count)
```

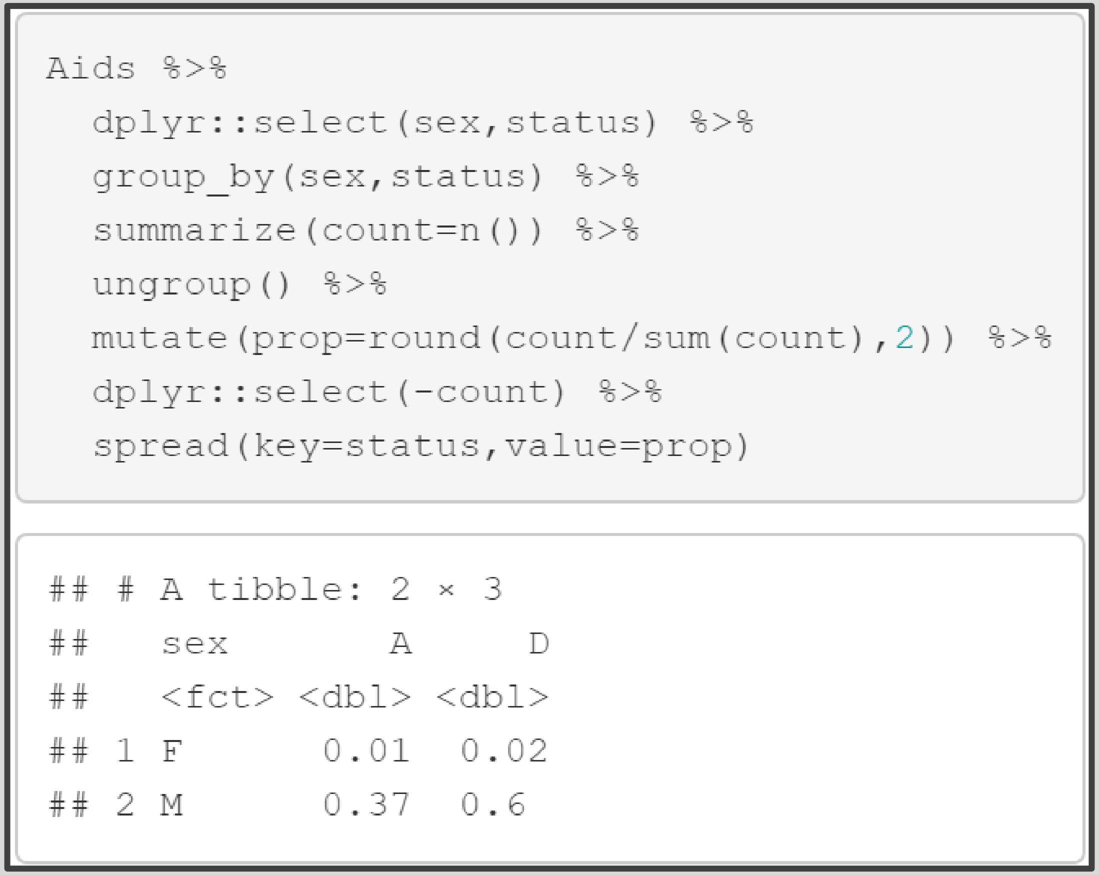
```
## # A tibble: 2 × 3
##   sex       A     D
##   <fct> <int> <int>
## 1 F        36    53
## 2 M      1046  1708
```

- Check:

$$36 + 53 + 1046 + 1708 = 2843$$

Contingency Tables

- Create Table of Proportions

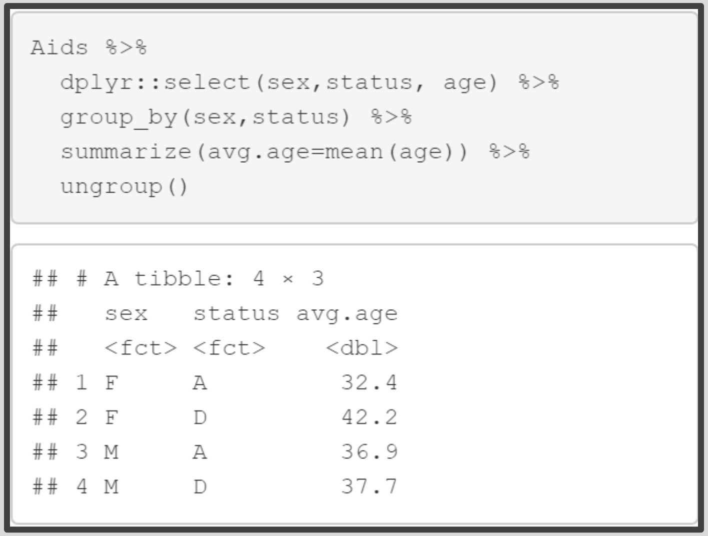


```
Aids %>%
  dplyr::select(sex,status) %>%
  group_by(sex,status) %>%
  summarize(count=n()) %>%
  ungroup() %>%
  mutate(prop=round(count/sum(count),2)) %>%
  dplyr::select(-count) %>%
  spread(key=status,value=prop)
```

```
## # A tibble: 2 × 3
##   sex       A     D
##   <fct> <dbl> <dbl>
## 1 F      0.01  0.02
## 2 M      0.37  0.6
```

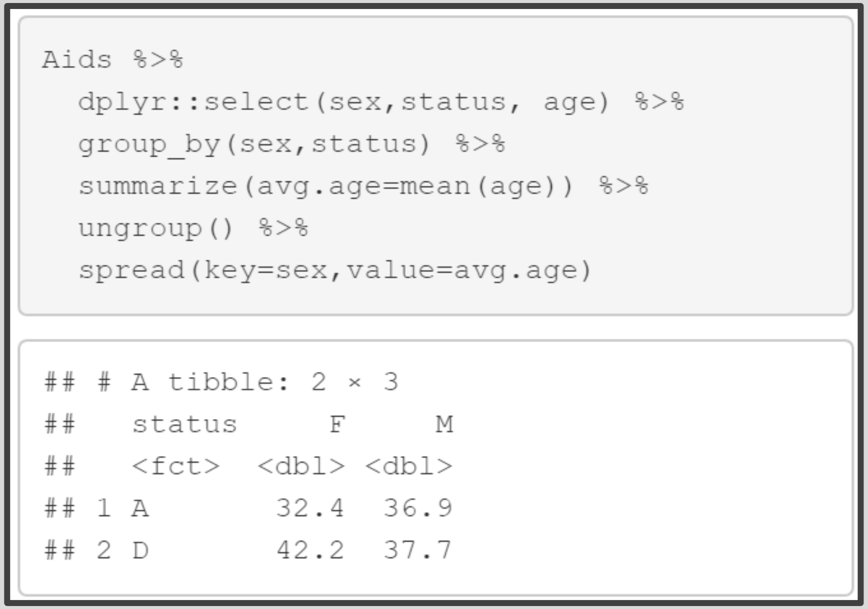
Contingency Tables

- Create Table of Average Age



```
Aids %>%
  dplyr::select(sex,status, age) %>%
  group_by(sex,status) %>%
  summarize(avg.age=mean(age)) %>%
  ungroup()
```

```
## # A tibble: 4 × 3
##   sex    status avg.age
##   <fct> <fct>     <dbl>
## 1 F     A          32.4
## 2 F     D          42.2
## 3 M     A          36.9
## 4 M     D          37.7
```



```
Aids %>%
  dplyr::select(sex,status, age) %>%
  group_by(sex,status) %>%
  summarize(avg.age=mean(age)) %>%
  ungroup() %>%
  spread(key=sex,value=avg.age)
```

```
## # A tibble: 2 × 3
##   status     F     M
##   <fct>  <dbl> <dbl>
## 1 A       32.4  36.9
## 2 D       42.2  37.7
```

Closing

Disperse and Make Reasonable Decisions