Joining and Merging Data

Prof. Roger D. Peng
Department of Statistics and Data Sciences
University of Texas at Austin

Why Join?

- Sometimes data are more efficiently stored in separate data frames or datasets

- Data are created by different / independent people and are not directly related to each other

- New data can be created to answer new questions by **joining** multiple tables together to create the dataset that we want

- Relational databases can consist of multiple **tables** that can be used and re-used for different purposes

R Functions for Joins

- The **dplyr** package provides a set of functions for joining two data frames into a single data frame based on a set of **key** columns.
    - left_join()
    - inner_join()
    - right_join()
- There are other functions for joining but they are less commonly used.

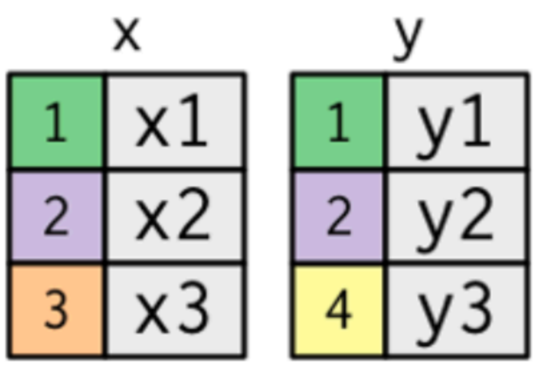
R Functions for Joins

- **left_join()** is useful for merging a “large” data frame (left) with a “smaller” one (right) while retaining all the rows of the “large” data frame

- **inner_join()** gives you the intersection of the rows between two data frames

- **right_join()** is like **left_join()** with the arguments reversed (likely only useful at the end of a pipeline)

Left Join

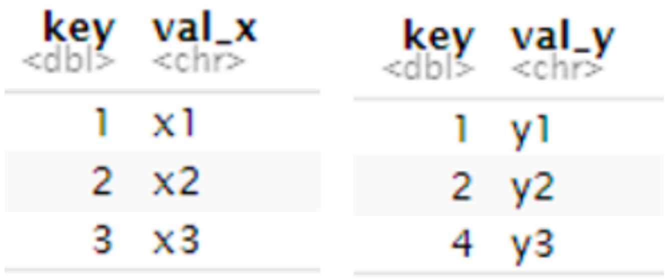


x

1	x1
2	x2
3	x3

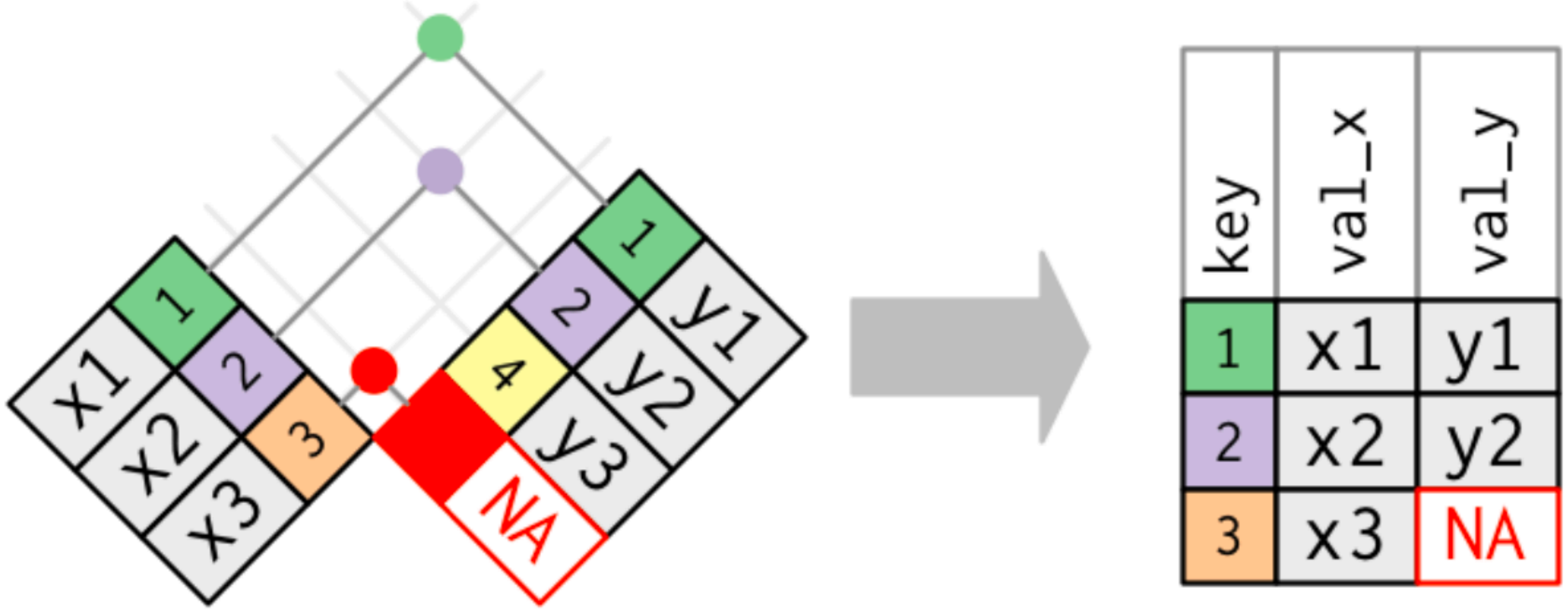
y

1	y1
2	y2
4	y3



key <dbl>	val_x <chr>
1	x1
2	x2
3	x3

key <dbl>	val_y <chr>
1	y1
2	y2
4	y3



key	val_x	val_y
1	x1	y1
2	x2	y2
3	x3	NA

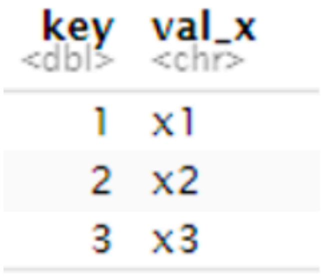
Inner Join

x

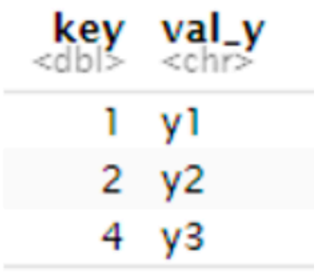
1	x1
2	x2
3	x3

y

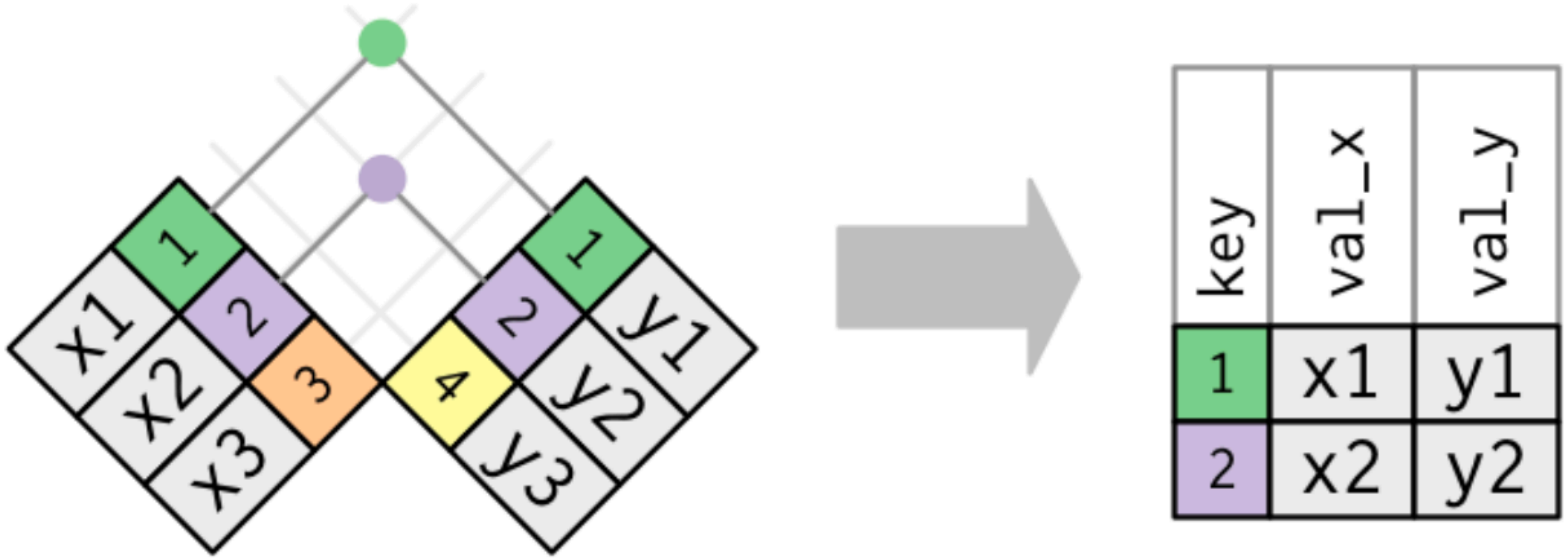
1	y1
2	y2
4	y3



key <dbl>	val_x <chr>
1	x1
2	x2
3	x3

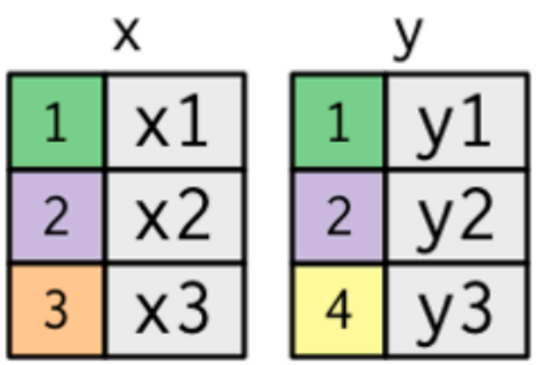


key <dbl>	val_y <chr>
1	y1
2	y2
4	y3



key	val_x	val_y
1	x1	y1
2	x2	y2

Right Join

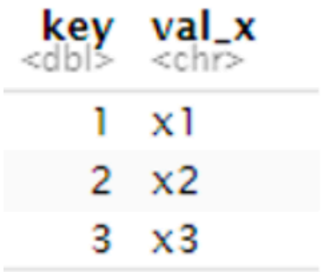


x

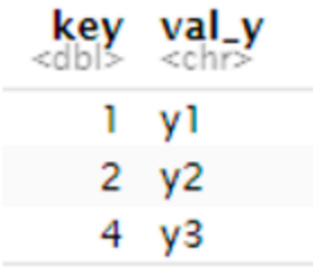
1	x1
2	x2
3	x3

y

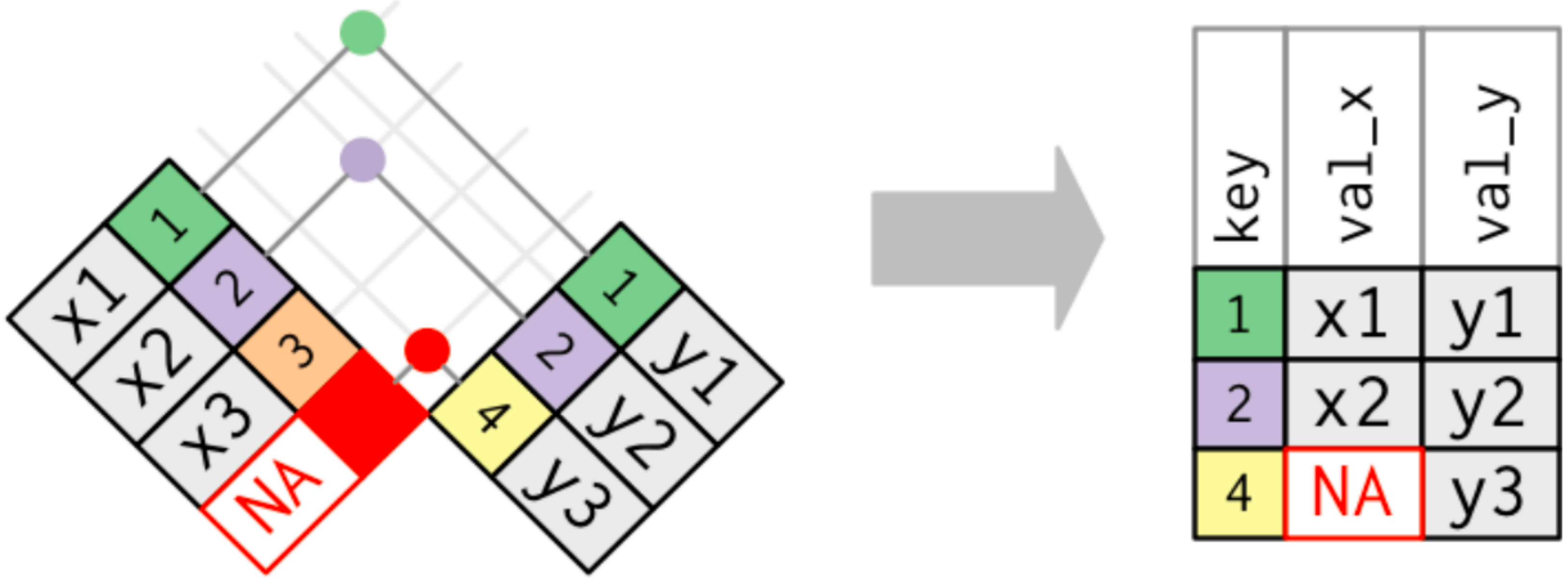
1	y1
2	y2
4	y3



key <dbl>	val_x <chr>
1	x1
2	x2
3	x3



key <dbl>	val_y <chr>
1	y1
2	y2
4	y3



key	val_x	val_y
1	x1	y1
2	x2	y2
4	NA	y3

Example: MAAIT Longitudinal Study

- Enrolled multiple participants and followed them for a year
- Clinic / home visits every 3 months for a total of 5 visits over the year
- Some information collected at **baseline**
- Some information collected at **each visit**
- Some information collected at a **subset of visits**

Fully Merged Dataset

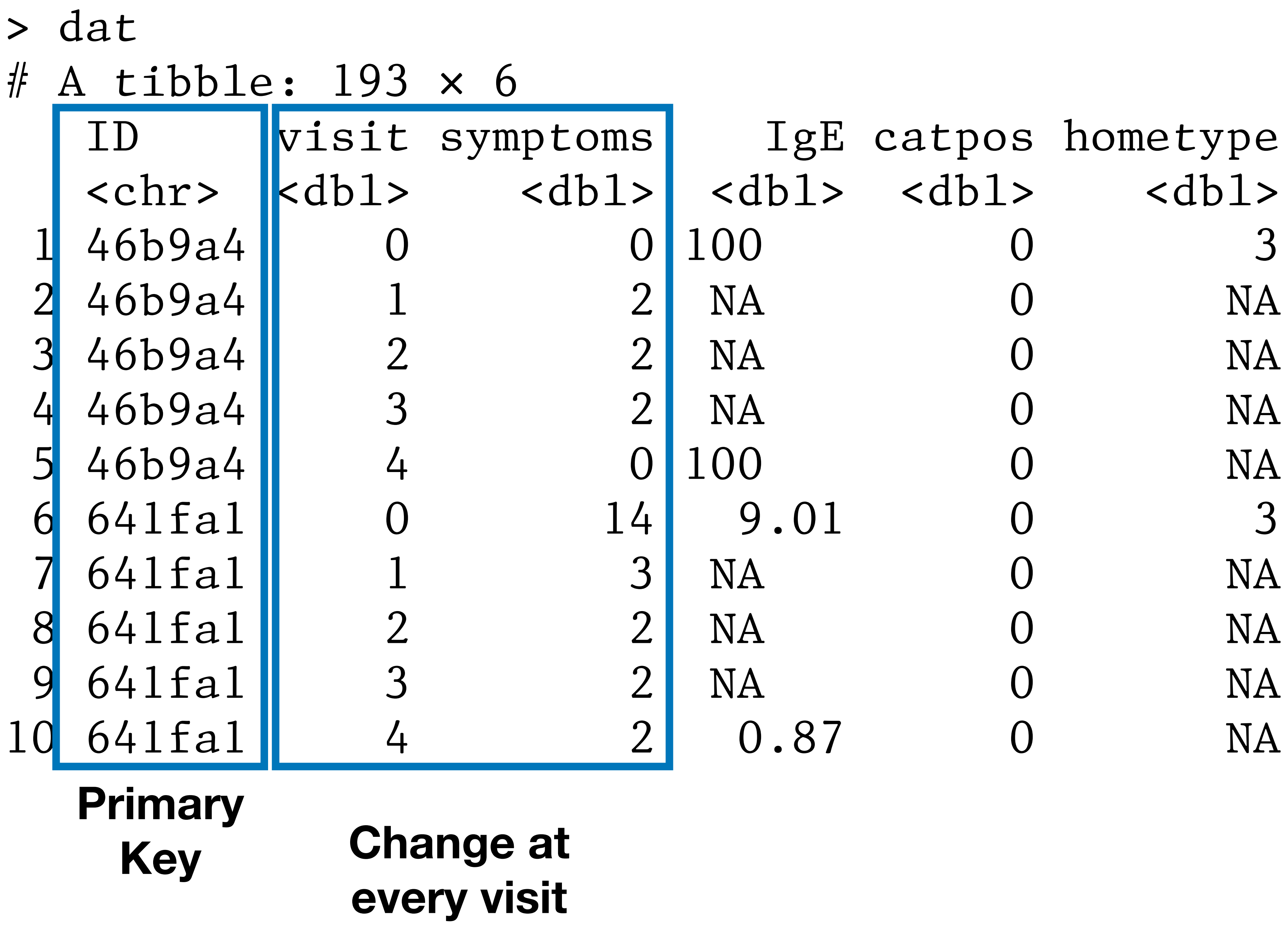
```
> dat
# A tibble: 193 × 6
   ID      visit symptoms    IgE catpos hometype
   <chr>   <dbl>    <dbl>  <dbl>  <dbl>    <dbl>
 1 46b9a4      0        0 100         0        3
 2 46b9a4      1        2  NA         0       NA
 3 46b9a4      2        2  NA         0       NA
 4 46b9a4      3        2  NA         0       NA
 5 46b9a4      4        0 100         0       NA
 6 641fa1      0       14   9.01      0        3
 7 641fa1      1        3  NA         0       NA
 8 641fa1      2        2  NA         0       NA
 9 641fa1      3        2  NA         0       NA
10 641fa1      4        2   0.87      0       NA
```

Fully Merged Dataset

```
> dat
# A tibble: 193 × 6
   ID     visit symptoms    IgE catpos hometype
   <chr>  <dbl>    <dbl>  <dbl>  <dbl>    <dbl>
 1 46b9a4     0        0 100          0        3
 2 46b9a4     1        2  NA          0       NA
 3 46b9a4     2        2  NA          0       NA
 4 46b9a4     3        2  NA          0       NA
 5 46b9a4     4        0 100          0       NA
 6 641fa1     0       14   9.01       0        3
 7 641fa1     1        3  NA          0       NA
 8 641fa1     2        2  NA          0       NA
 9 641fa1     3        2  NA          0       NA
10 641fa1     4        2   0.87       0       NA
```

Primary Key

Fully Merged Dataset

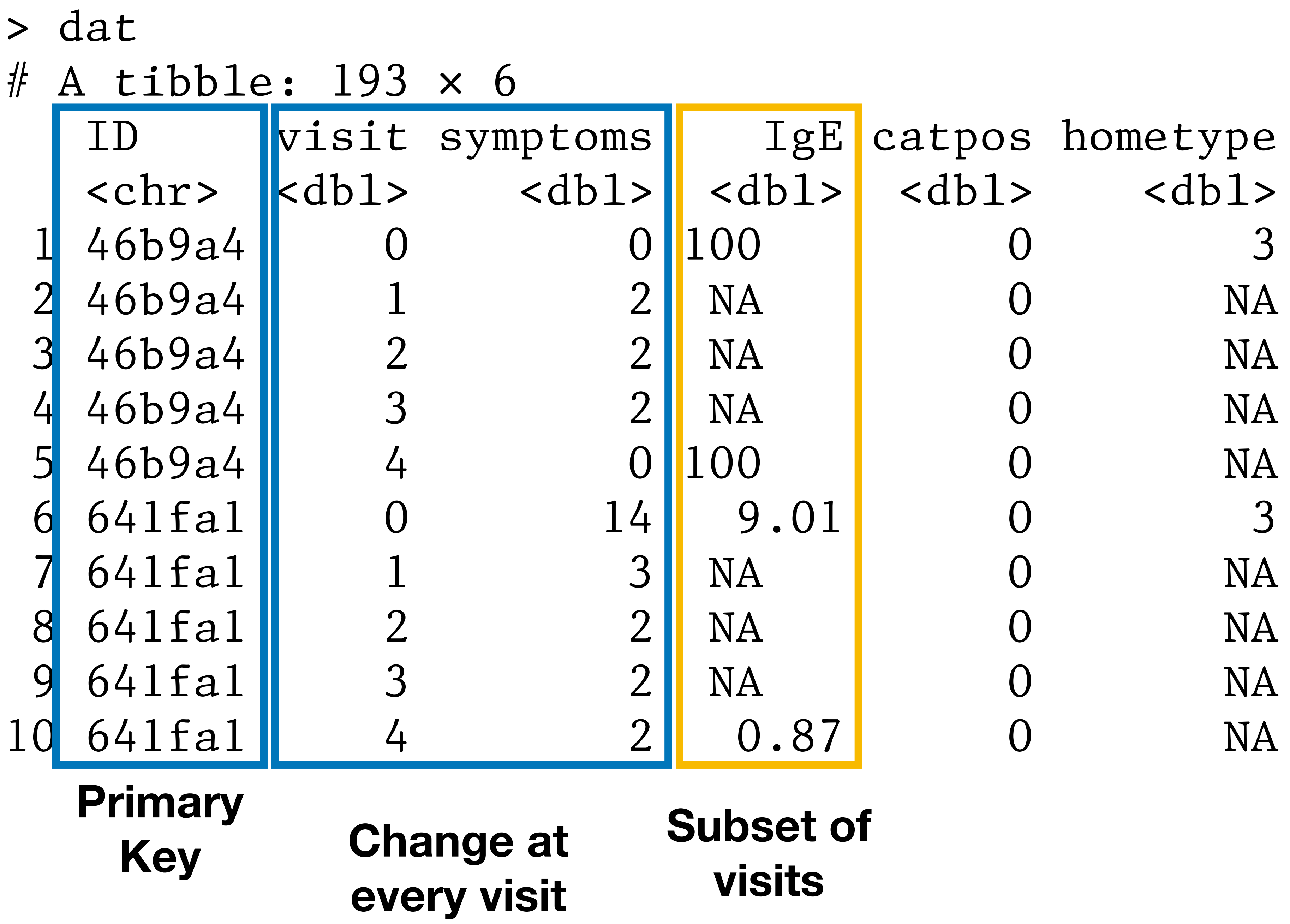


```
> dat
# A tibble: 193 × 6
   ID     visit symptoms   IgE catpos hometype
   <chr>  <dbl>    <dbl> <dbl>  <dbl>    <dbl>
 1 46b9a4     0        0 100        0        3
 2 46b9a4     1        2  NA        0       NA
 3 46b9a4     2        2  NA        0       NA
 4 46b9a4     3        2  NA        0       NA
 5 46b9a4     4        0 100        0       NA
 6 641fa1     0       14   9.01     0        3
 7 641fa1     1        3  NA        0       NA
 8 641fa1     2        2  NA        0       NA
 9 641fa1     3        2  NA        0       NA
10 641fa1     4        2   0.87     0       NA
```

Primary Key

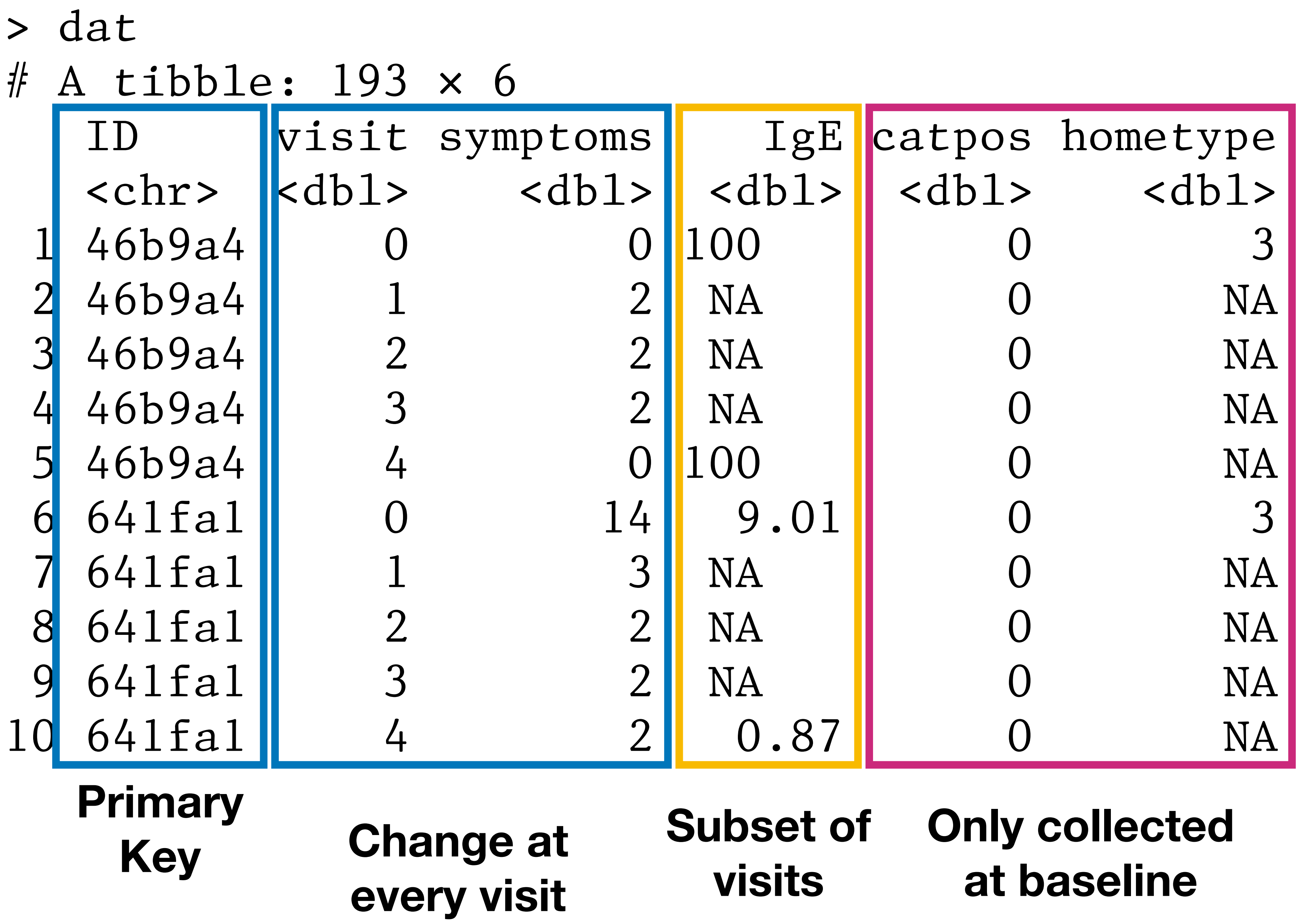
Change at every visit

Fully Merged Dataset



```
> dat
# A tibble: 193 × 6
   ID     visit symptoms     IgE catpos hometype
   <chr>  <dbl>    <dbl>  <dbl>  <dbl>    <dbl>
 1 46b9a4     0        0 100         0        3
 2 46b9a4     1        2  NA         0       NA
 3 46b9a4     2        2  NA         0       NA
 4 46b9a4     3        2  NA         0       NA
 5 46b9a4     4        0 100         0       NA
 6 641fa1     0       14   9.01      0        3
 7 641fa1     1        3  NA         0       NA
 8 641fa1     2        2  NA         0       NA
 9 641fa1     3        2  NA         0       NA
10 641fa1     4        2   0.87      0       NA
```

Fully Merged Dataset



```
> dat
# A tibble: 193 × 6
   ID     visit symptoms   IgE catpos hometype
   <chr>  <dbl>    <dbl> <dbl>  <dbl>    <dbl>
 1 46b9a4     0        0 100        0        3
 2 46b9a4     1        2  NA        0       NA
 3 46b9a4     2        2  NA        0       NA
 4 46b9a4     3        2  NA        0       NA
 5 46b9a4     4        0 100        0       NA
 6 641fa1     0       14   9.01     0        3
 7 641fa1     1        3  NA        0       NA
 8 641fa1     2        2  NA        0       NA
 9 641fa1     3        2  NA        0       NA
10 641fa1     4        2   0.87     0       NA
```

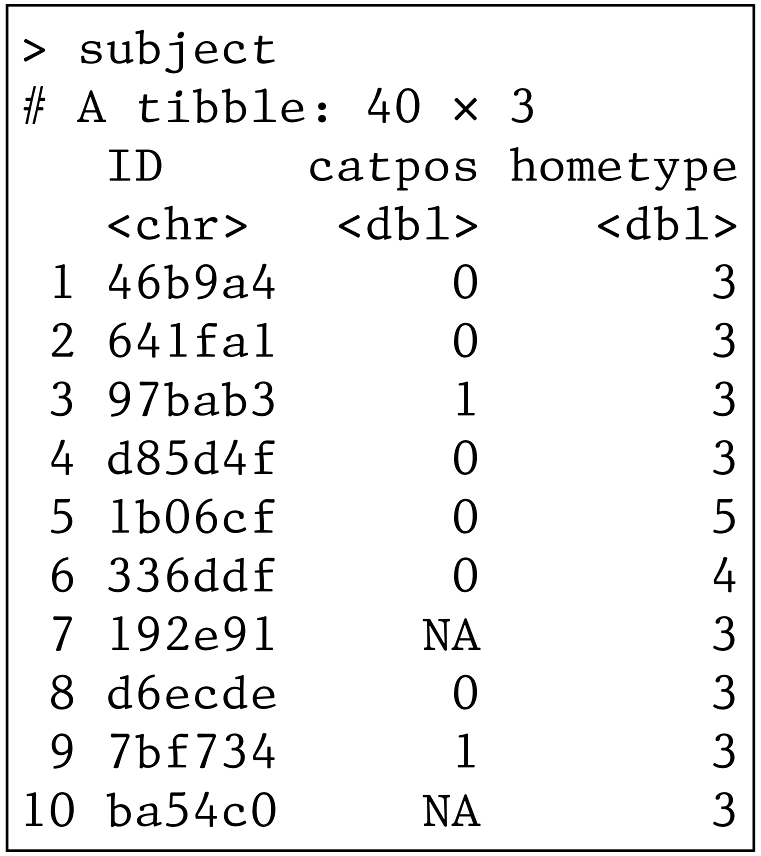
Primary Key

Change at every visit

Subset of visits

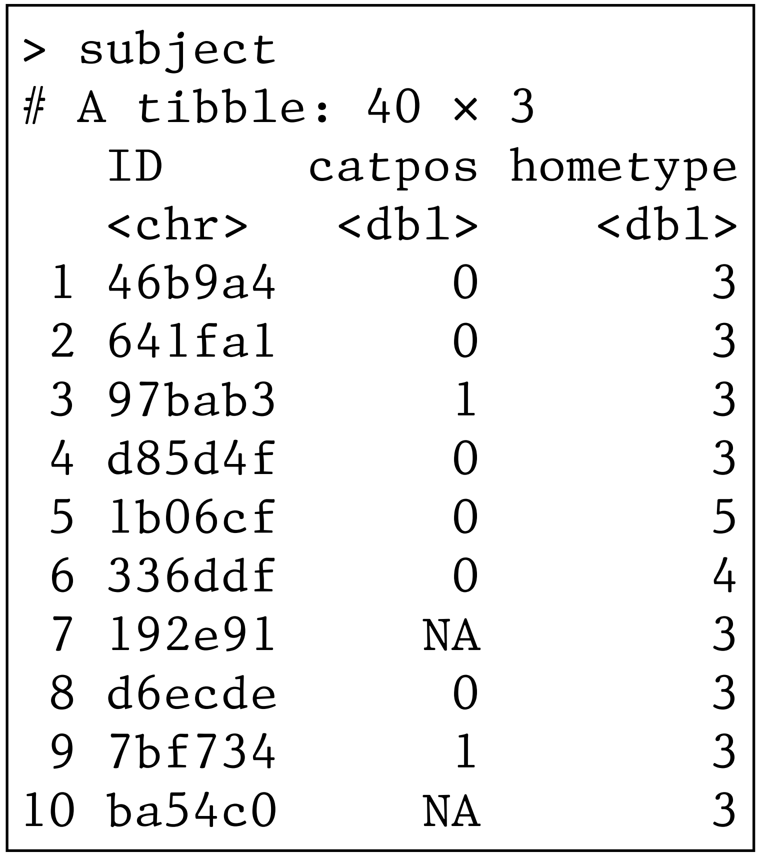
Only collected at baseline

Alternate Unmerged Form

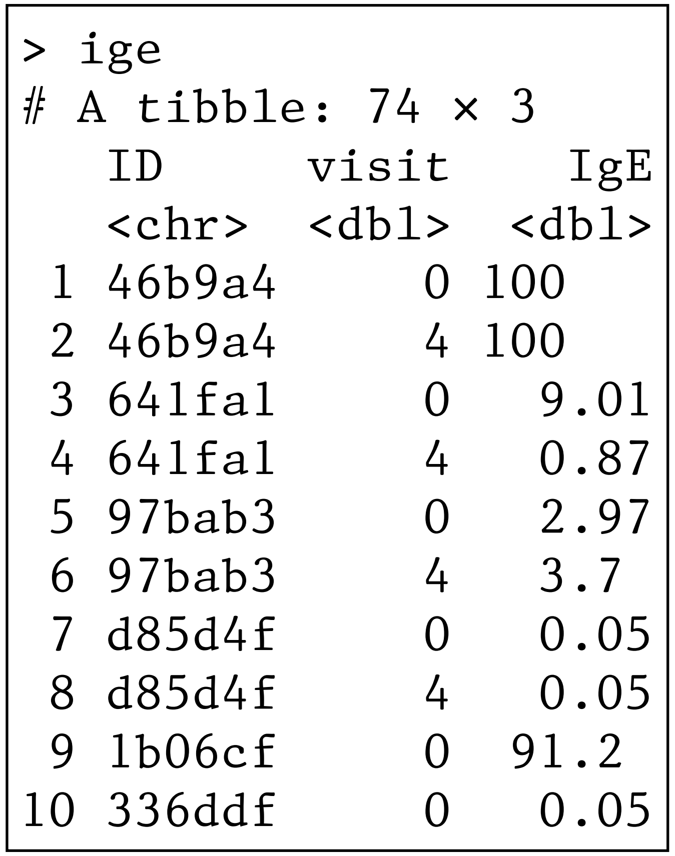


```
> subject
# A tibble: 40 × 3
   ID     catpos hometype
   <chr>   <dbl>    <dbl>
 1 46b9a4      0        3
 2 641fa1      0        3
 3 97bab3      1        3
 4 d85d4f      0        3
 5 1b06cf      0        5
 6 336ddf      0        4
 7 192e91     NA        3
 8 d6ecde      0        3
 9 7bf734      1        3
10 ba54c0     NA        3
```

Alternate Unmerged Form

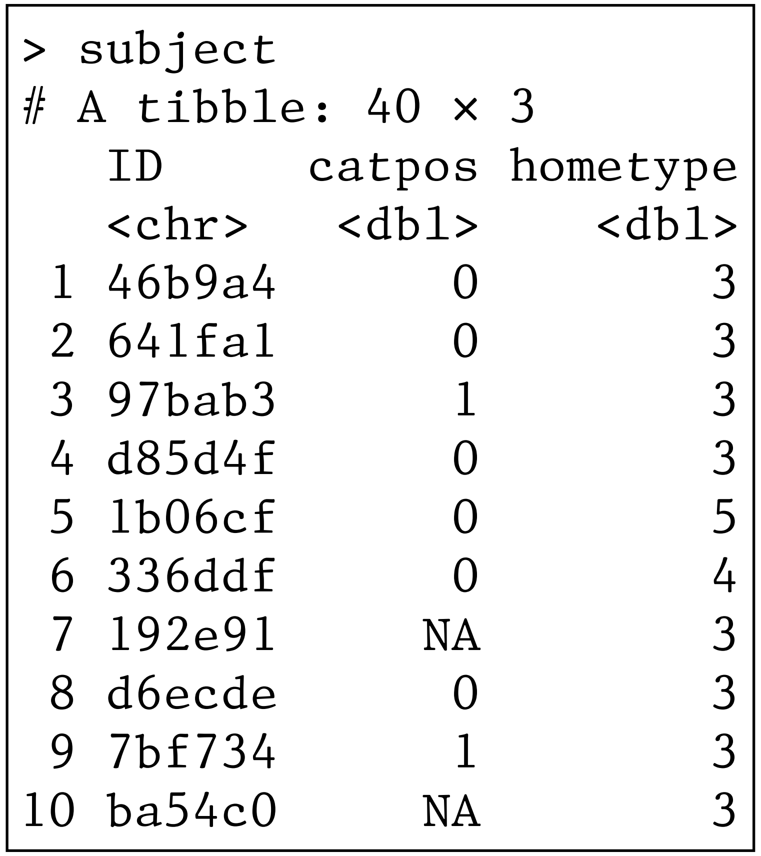


```
> subject
# A tibble: 40 × 3
   ID     catpos hometype
   <chr>   <dbl>    <dbl>
 1 46b9a4      0        3
 2 641fa1      0        3
 3 97bab3      1        3
 4 d85d4f      0        3
 5 1b06cf      0        5
 6 336ddf      0        4
 7 192e91     NA        3
 8 d6ecde      0        3
 9 7bf734      1        3
10 ba54c0     NA        3
```

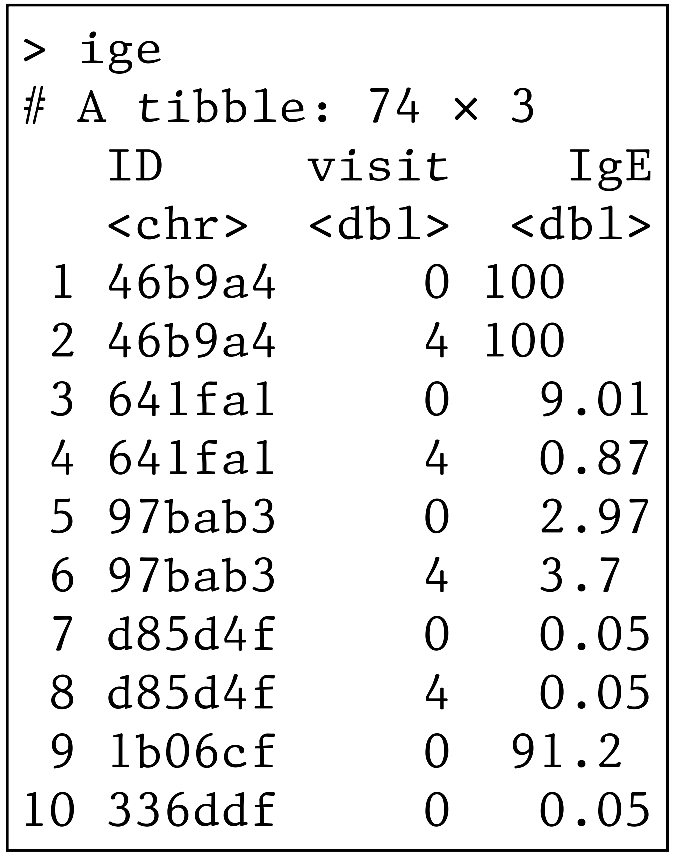


```
> ige
# A tibble: 74 × 3
   ID     visit    IgE
   <chr>  <dbl>  <dbl>
 1 46b9a4     0 100   
 2 46b9a4     4 100   
 3 641fa1     0   9.01
 4 641fa1     4   0.87
 5 97bab3     0   2.97
 6 97bab3     4   3.7 
 7 d85d4f     0   0.05
 8 d85d4f     4   0.05
 9 1b06cf     0  91.2 
10 336ddf     0   0.05
```

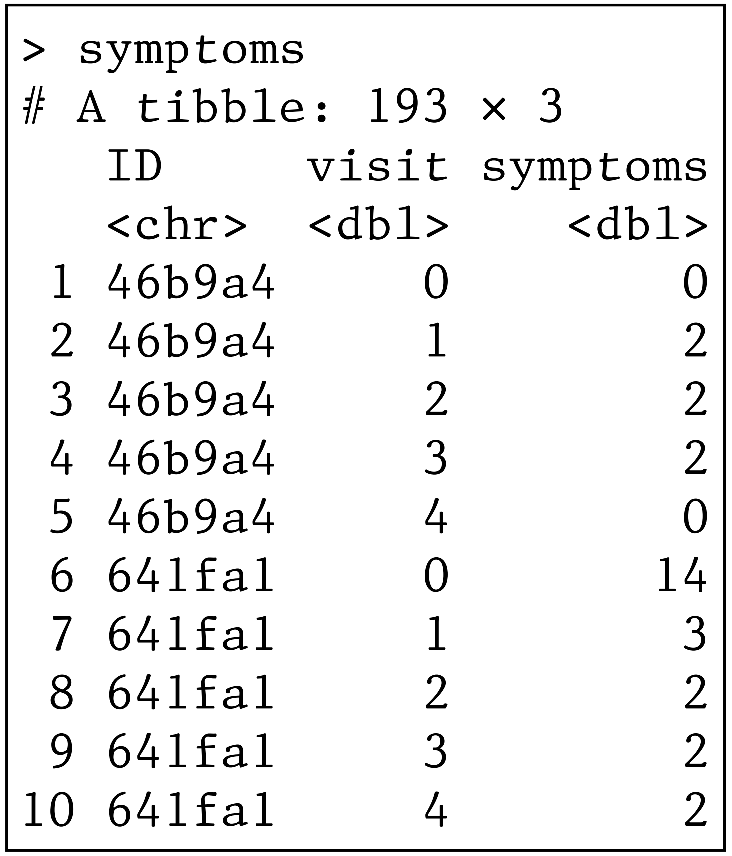
Alternate Unmerged Form



```
> subject
# A tibble: 40 × 3
   ID     catpos hometype
   <chr>   <dbl>    <dbl>
 1 46b9a4      0        3
 2 641fa1      0        3
 3 97bab3      1        3
 4 d85d4f      0        3
 5 1b06cf      0        5
 6 336ddf      0        4
 7 192e91     NA        3
 8 d6ecde      0        3
 9 7bf734      1        3
10 ba54c0     NA        3
```

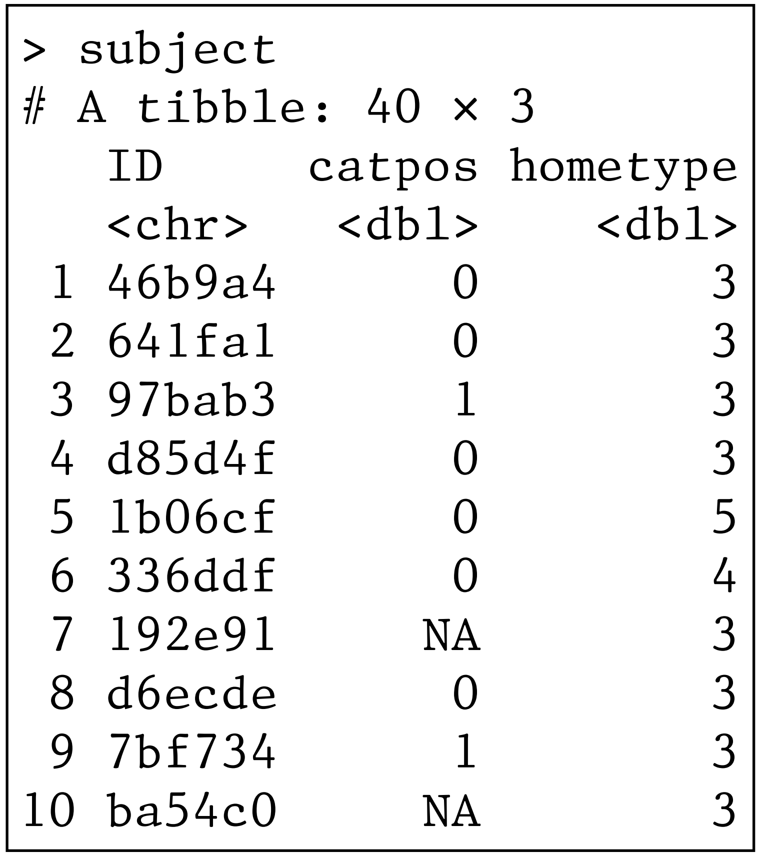


```
> ige
# A tibble: 74 × 3
   ID     visit    IgE
   <chr>  <dbl>  <dbl>
 1 46b9a4     0 100   
 2 46b9a4     4 100   
 3 641fa1     0   9.01
 4 641fa1     4   0.87
 5 97bab3     0   2.97
 6 97bab3     4   3.7 
 7 d85d4f     0   0.05
 8 d85d4f     4   0.05
 9 1b06cf     0  91.2 
10 336ddf     0   0.05
```

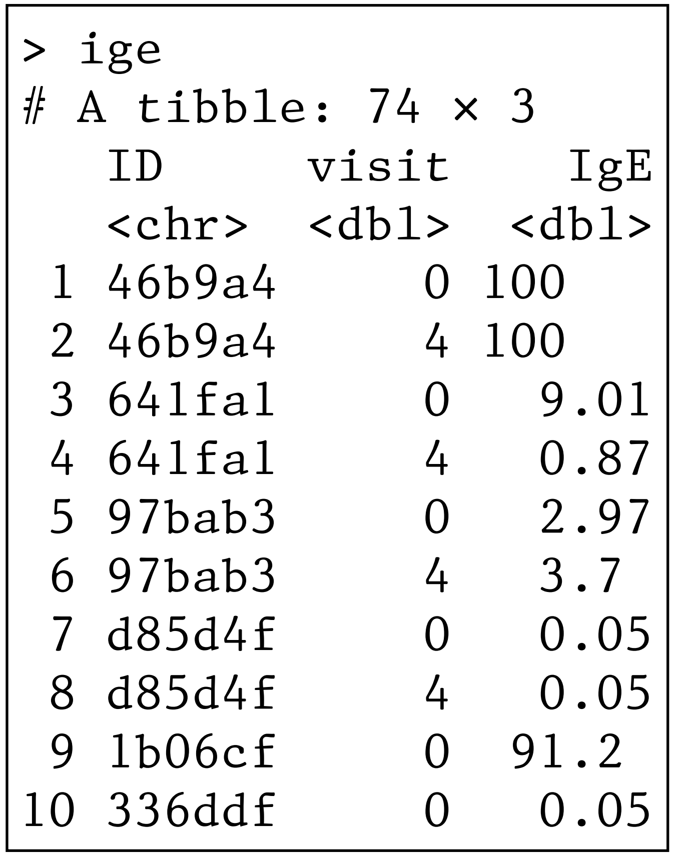


```
> symptoms
# A tibble: 193 × 3
   ID     visit symptoms
   <chr>  <dbl>    <dbl>
 1 46b9a4     0        0
 2 46b9a4     1        2
 3 46b9a4     2        2
 4 46b9a4     3        2
 5 46b9a4     4        0
 6 641fa1     0       14
 7 641fa1     1        3
 8 641fa1     2        2
 9 641fa1     3        2
10 641fa1     4        2
```

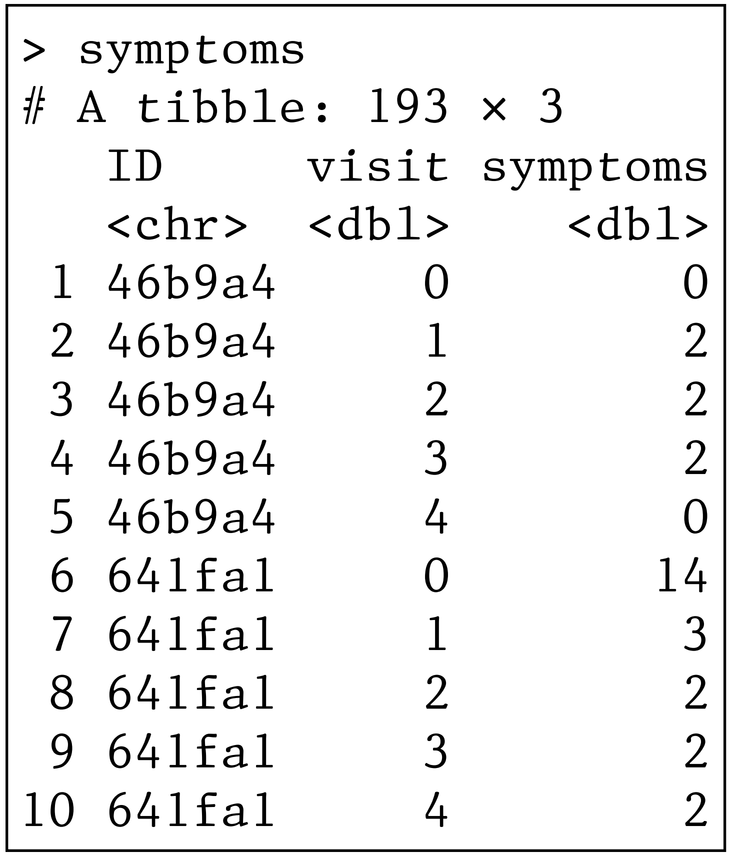
Alternate Unmerged Form



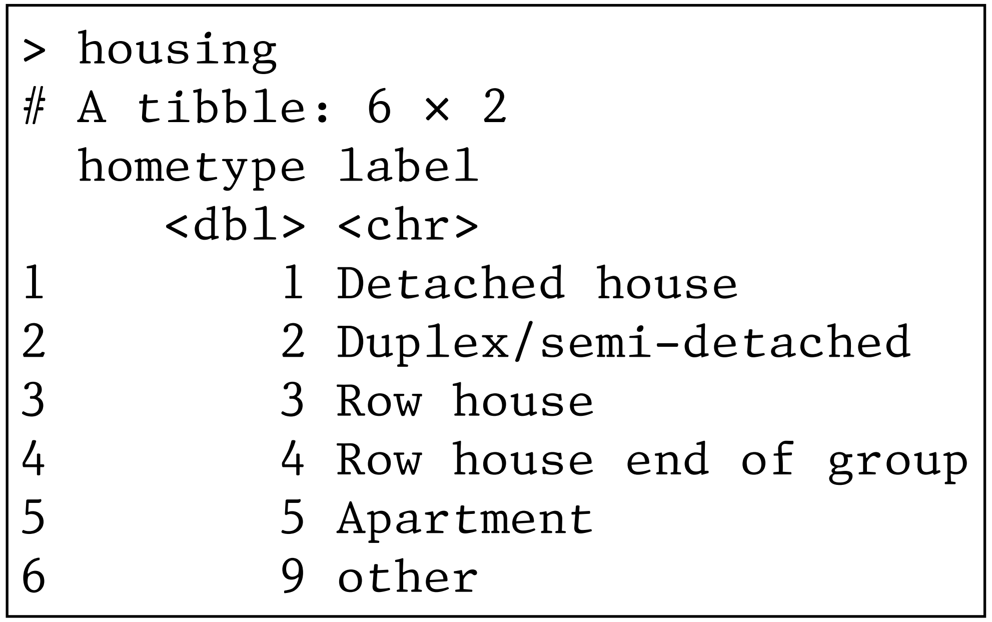
```
> subject
# A tibble: 40 × 3
   ID     catpos hometype
   <chr>   <dbl>    <dbl>
 1 46b9a4      0        3
 2 641fa1      0        3
 3 97bab3      1        3
 4 d85d4f      0        3
 5 1b06cf      0        5
 6 336ddf      0        4
 7 192e91     NA        3
 8 d6ecde      0        3
 9 7bf734      1        3
10 ba54c0     NA        3
```



```
> ige
# A tibble: 74 × 3
   ID     visit    IgE
   <chr>  <dbl>  <dbl>
 1 46b9a4     0 100
 2 46b9a4     4 100
 3 641fa1     0   9.01
 4 641fa1     4   0.87
 5 97bab3     0   2.97
 6 97bab3     4   3.7
 7 d85d4f     0   0.05
 8 d85d4f     4   0.05
 9 1b06cf     0  91.2
10 336ddf     0   0.05
```

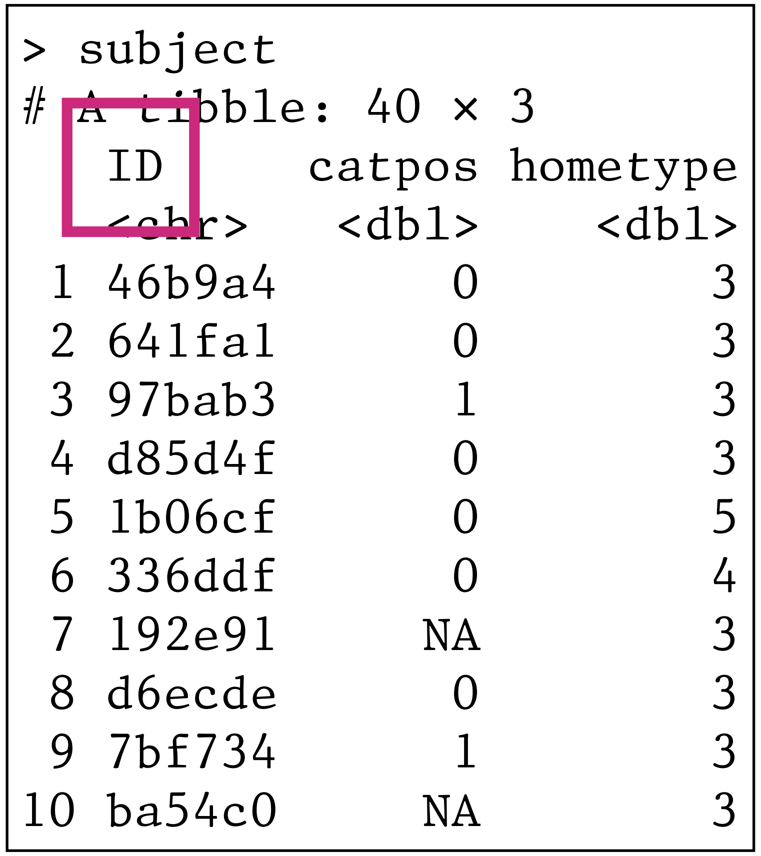


```
> symptoms
# A tibble: 193 × 3
   ID     visit symptoms
   <chr>  <dbl>    <dbl>
 1 46b9a4     0        0
 2 46b9a4     1        2
 3 46b9a4     2        2
 4 46b9a4     3        2
 5 46b9a4     4        0
 6 641fa1     0       14
 7 641fa1     1        3
 8 641fa1     2        2
 9 641fa1     3        2
10 641fa1     4        2
```

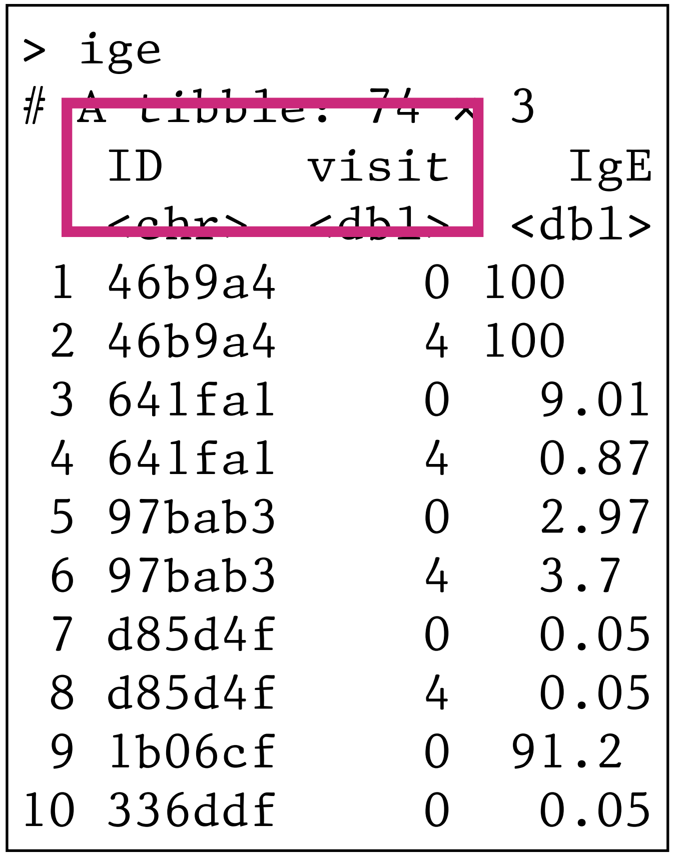


```
> housing
# A tibble: 6 × 2
  hometype label
     <dbl> <chr>
1        1 Detached house
2        2 Duplex/semi-detached
3        3 Row house
4        4 Row house end of group
5        5 Apartment
6        9 other
```

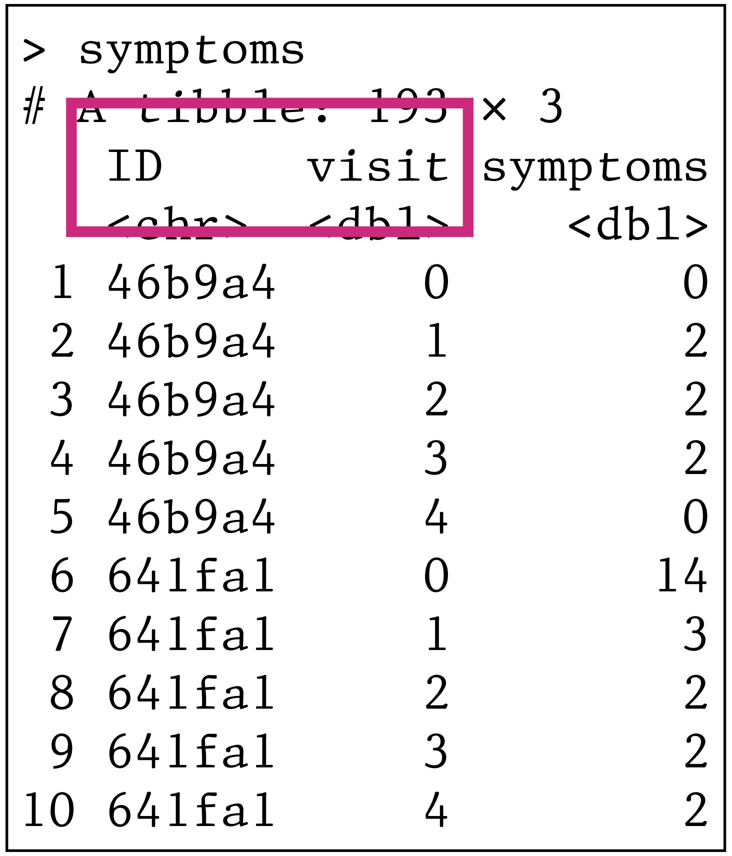
Alternate Unmerged Form



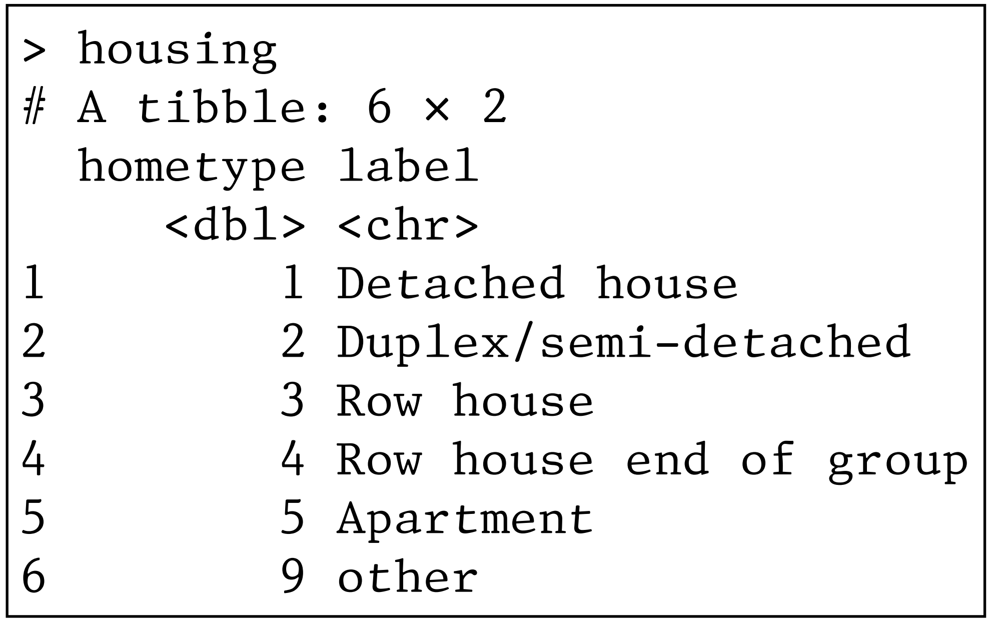
```
> subject
# A tibble: 40 × 3
   ID      catpos hometype
   <chr>    <dbl>    <dbl>
 1 46b9a4       0        3
 2 641fa1       0        3
 3 97bab3       1        3
 4 d85d4f       0        3
 5 1b06cf       0        5
 6 336ddf       0        4
 7 192e91      NA        3
 8 d6ecde       0        3
 9 7bf734       1        3
10 ba54c0      NA        3
```



```
> ige
# A tibble: 74 × 3
   ID      visit    IgE
   <chr>   <dbl>  <dbl>
 1 46b9a4      0 100
 2 46b9a4      4 100
 3 641fa1      0   9.01
 4 641fa1      4   0.87
 5 97bab3      0   2.97
 6 97bab3      4   3.7
 7 d85d4f      0   0.05
 8 d85d4f      4   0.05
 9 1b06cf      0  91.2
10 336ddf      0   0.05
```

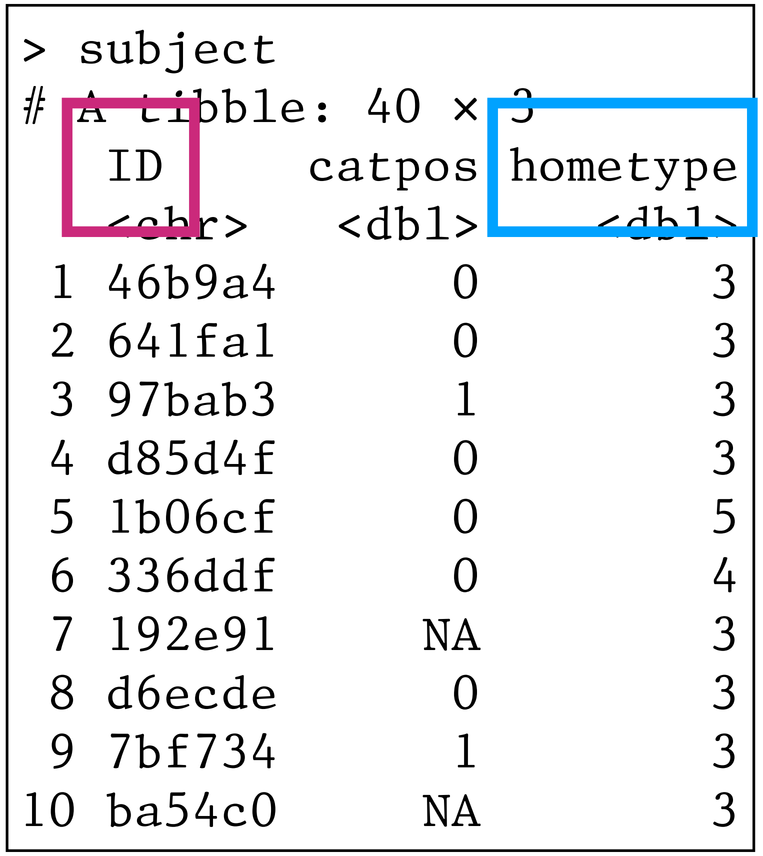


```
> symptoms
# A tibble: 193 × 3
   ID      visit symptoms
   <chr>   <dbl>    <dbl>
 1 46b9a4      0        0
 2 46b9a4      1        2
 3 46b9a4      2        2
 4 46b9a4      3        2
 5 46b9a4      4        0
 6 641fa1      0       14
 7 641fa1      1        3
 8 641fa1      2        2
 9 641fa1      3        2
10 641fa1      4        2
```

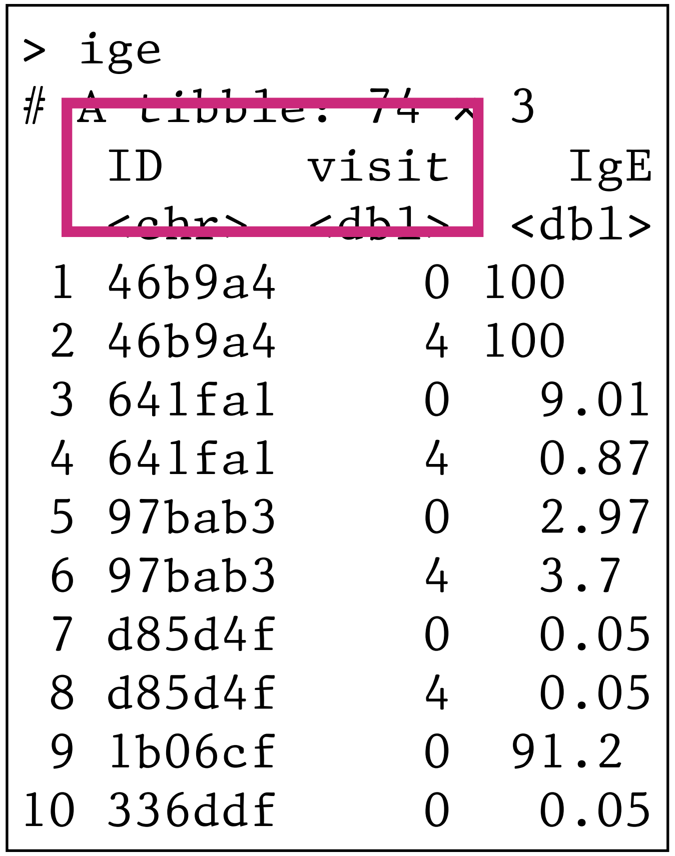


```
> housing
# A tibble: 6 × 2
  hometype label
     <dbl> <chr>
1        1 Detached house
2        2 Duplex/semi-detached
3        3 Row house
4        4 Row house end of group
5        5 Apartment
6        9 other
```

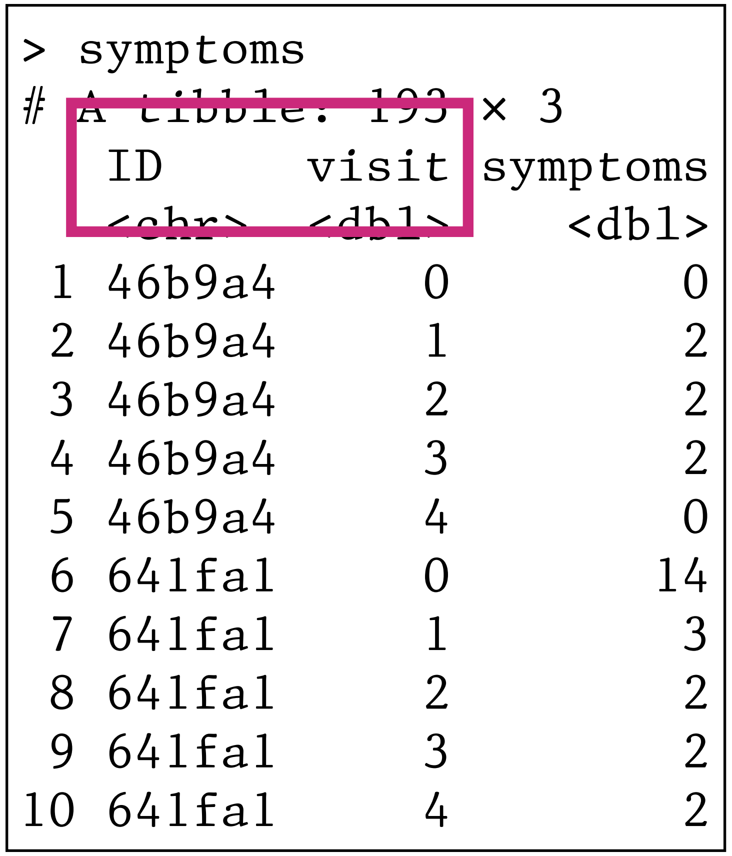
Alternate Unmerged Form



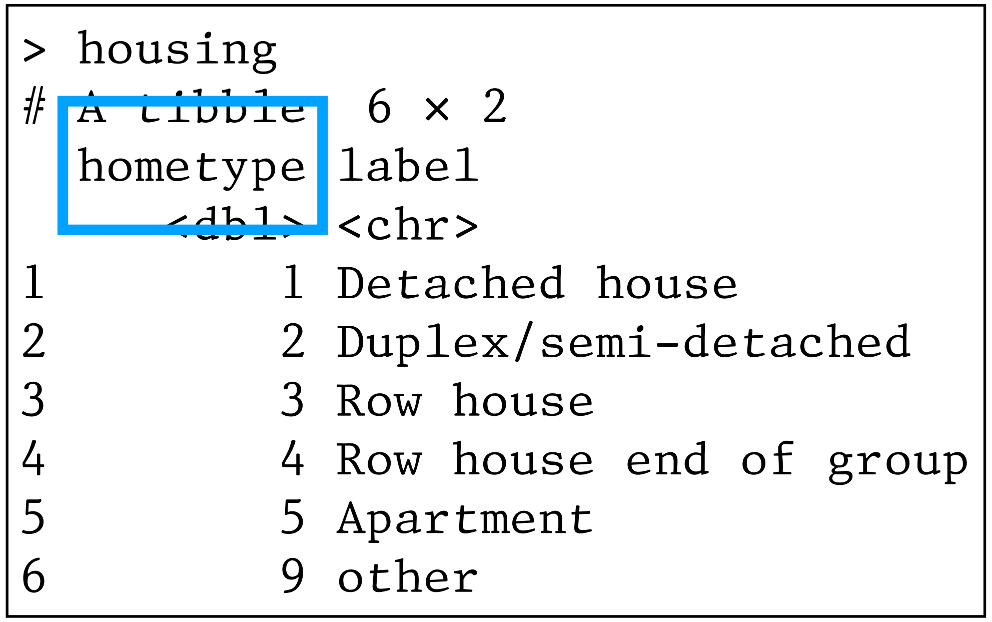
```
> subject
# A tibble: 40 × 3
   ID     catpos hometype
   <chr>   <dbl>    <dbl>
 1 46b9a4      0        3
 2 641fa1      0        3
 3 97bab3      1        3
 4 d85d4f      0        3
 5 1b06cf      0        5
 6 336ddf      0        4
 7 192e91     NA        3
 8 d6ecde      0        3
 9 7bf734      1        3
10 ba54c0     NA        3
```



```
> ige
# A tibble: 74 × 3
   ID     visit    IgE
   <chr>  <dbl>  <dbl>
 1 46b9a4     0 100   
 2 46b9a4     4 100   
 3 641fa1     0   9.01
 4 641fa1     4   0.87
 5 97bab3     0   2.97
 6 97bab3     4   3.7 
 7 d85d4f     0   0.05
 8 d85d4f     4   0.05
 9 1b06cf     0  91.2 
10 336ddf     0   0.05
```



```
> symptoms
# A tibble: 193 × 3
   ID     visit symptoms
   <chr>  <dbl>    <dbl>
 1 46b9a4     0        0
 2 46b9a4     1        2
 3 46b9a4     2        2
 4 46b9a4     3        2
 5 46b9a4     4        0
 6 641fa1     0       14
 7 641fa1     1        3
 8 641fa1     2        2
 9 641fa1     3        2
10 641fa1     4        2
```



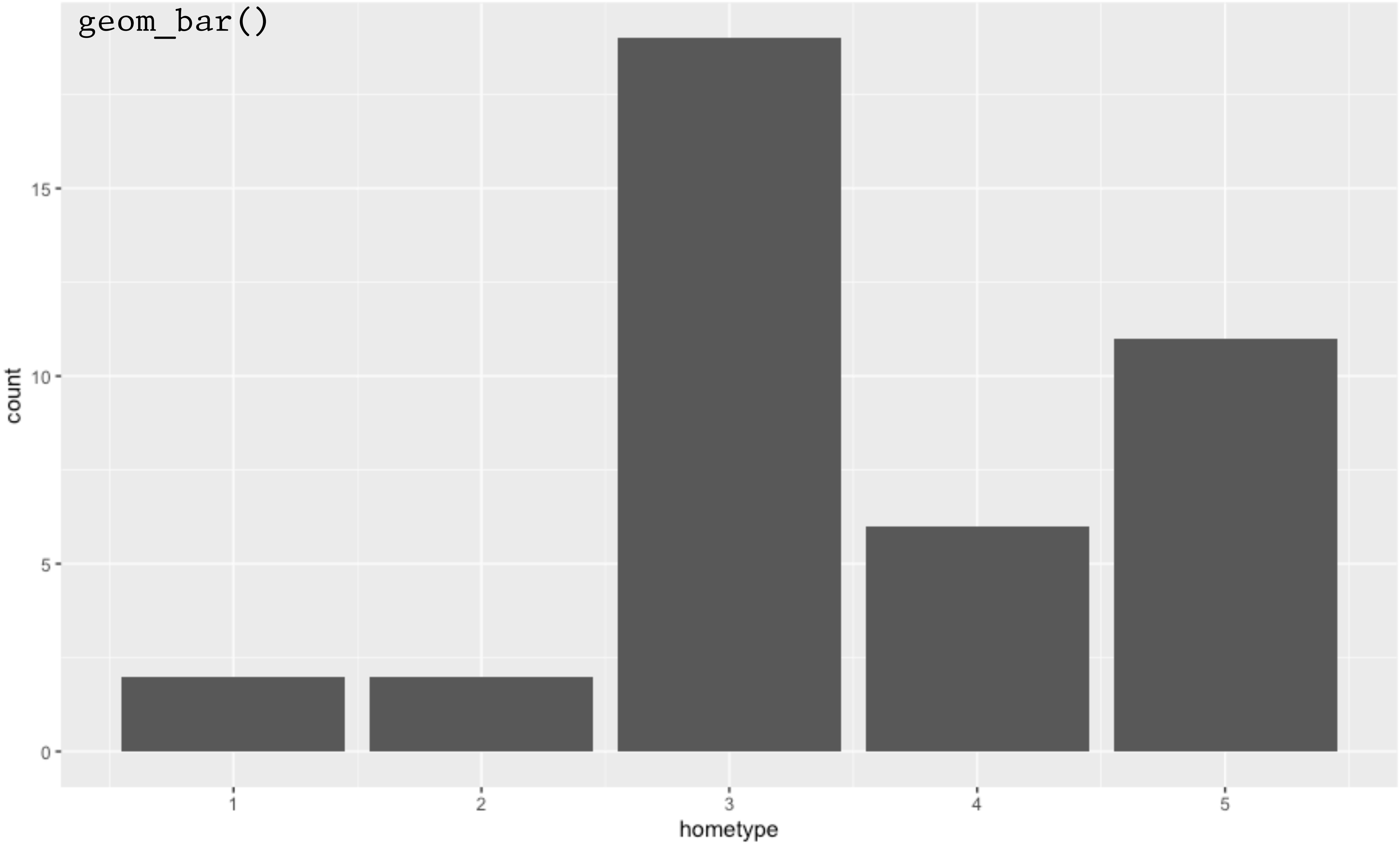
```
> housing
# A tibble: 6 × 2
  hometype label
     <dbl> <chr>
1        1 Detached house
2        2 Duplex/semi-detached
3        3 Row house
4        4 Row house end of group
5        5 Apartment
6        9 other
```

Home Types

```
subject |>
        ggplot(aes(x = hometype)) +
        geom_bar()
```

Home Types

```
subject |>
      ggplot(aes(x = hometype)) +
      geom_bar()
```



Cat Allergy by Home Type

```
subject |>
        ggplot(aes(x = hometype, y = catpos)) +
        geom_bar(stat = "summary", fun = "mean")
```

Cat Allergy by Home Type

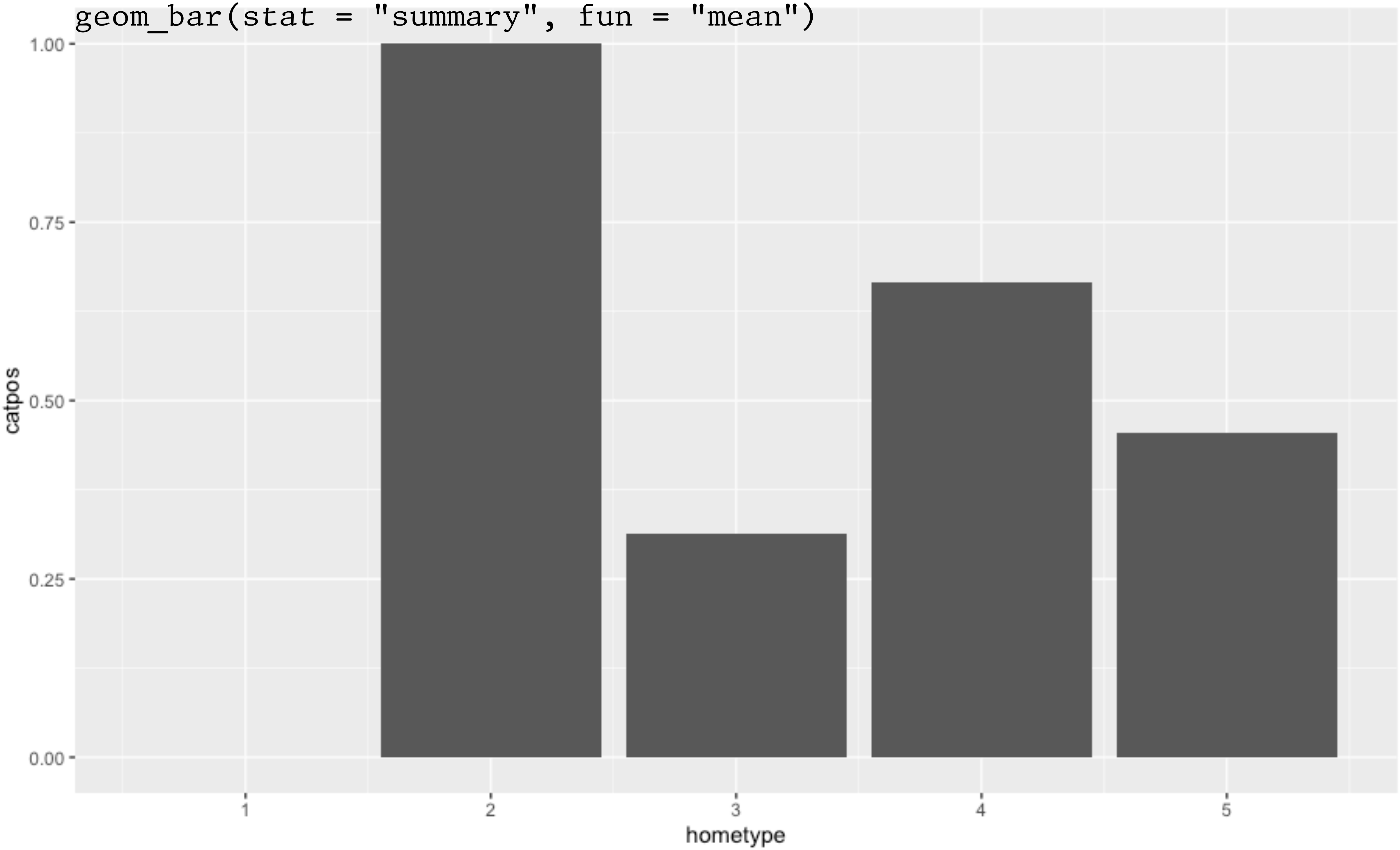
```
subject |>
        ggplot(aes(x = hometype, y = catpos)) +
        geom_bar(stat = "summary", fun = "mean")
```

Takes a 0 or 1 value

Cat Allergy by Home Type

```
subject |>
      ggplot(aes(x = hometype, y = catpos)) +
      geom_bar(stat = "summary", fun = "mean")
```

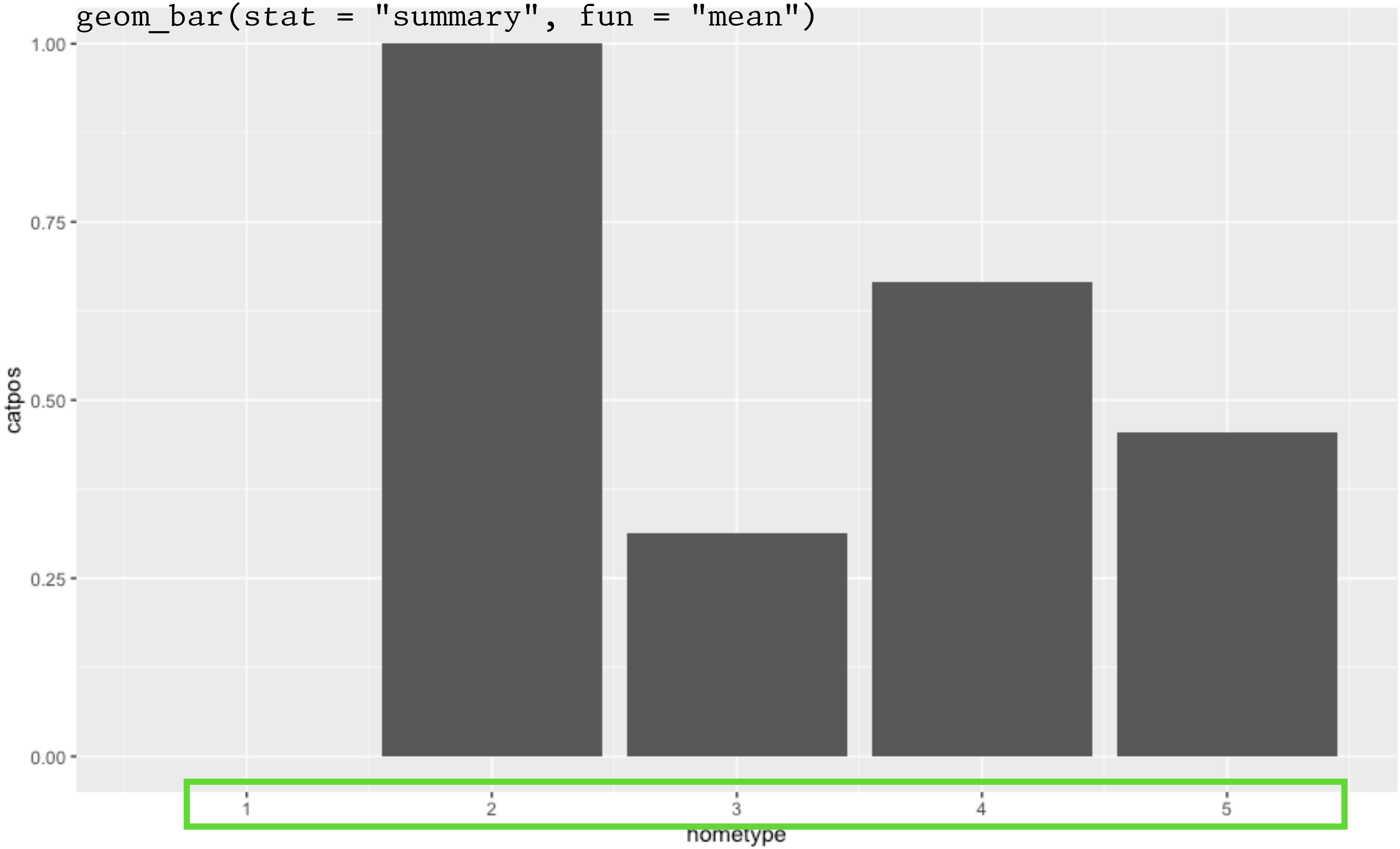
Takes a 0 or 1 value



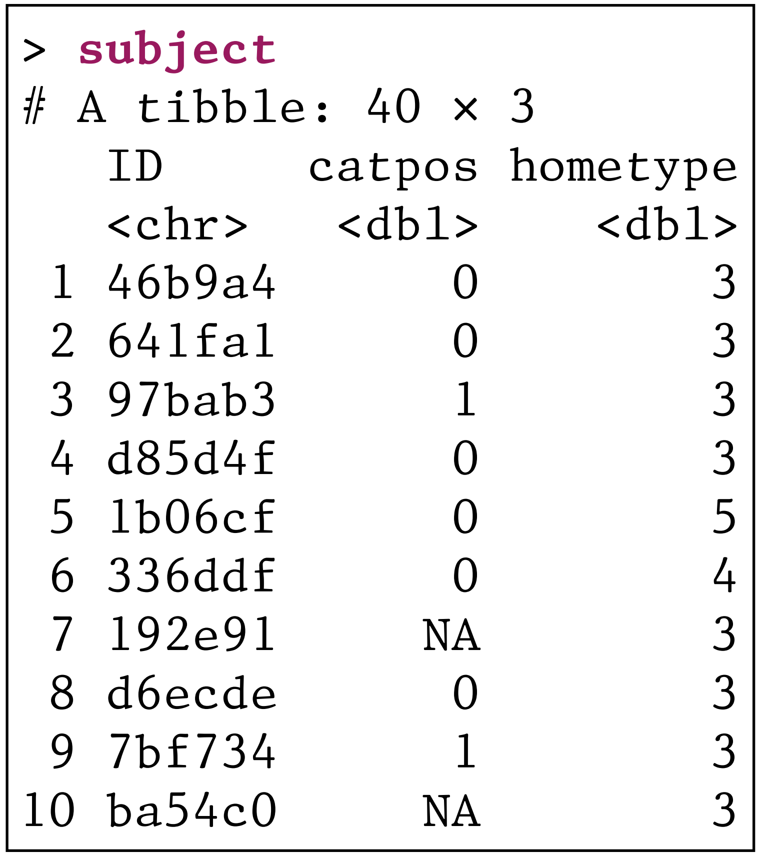
Cat Allergy by Home Type

```
subject |>
    ggplot(aes(x = hometype, y = catpos)) +
    geom_bar(stat = "summary", fun = "mean")
```

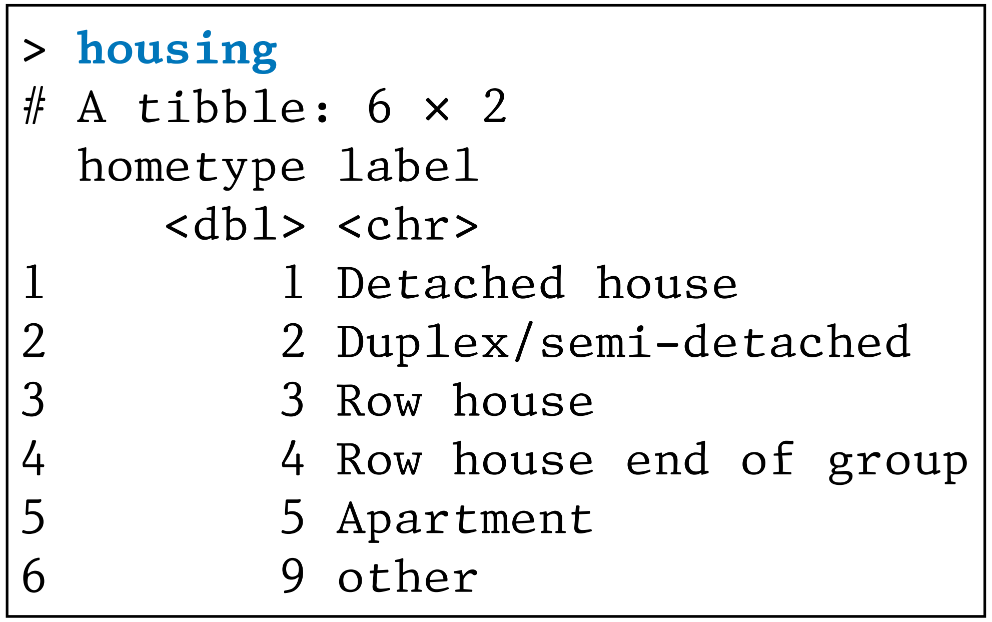
Takes a 0 or 1 value



Joining Tables



```
> subject
# A tibble: 40 × 3
   ID     catpos hometype
   <chr>   <dbl>    <dbl>
 1 46b9a4      0        3
 2 641fa1      0        3
 3 97bab3      1        3
 4 d85d4f      0        3
 5 1b06cf      0        5
 6 336ddf      0        4
 7 192e91     NA        3
 8 d6ecde      0        3
 9 7bf734      1        3
10 ba54c0     NA        3
```



```
> housing
# A tibble: 6 × 2
  hometype label
     <dbl> <chr>
1        1 Detached house
2        2 Duplex/semi-detached
3        3 Row house
4        4 Row house end of group
5        5 Apartment
6        9 other
```

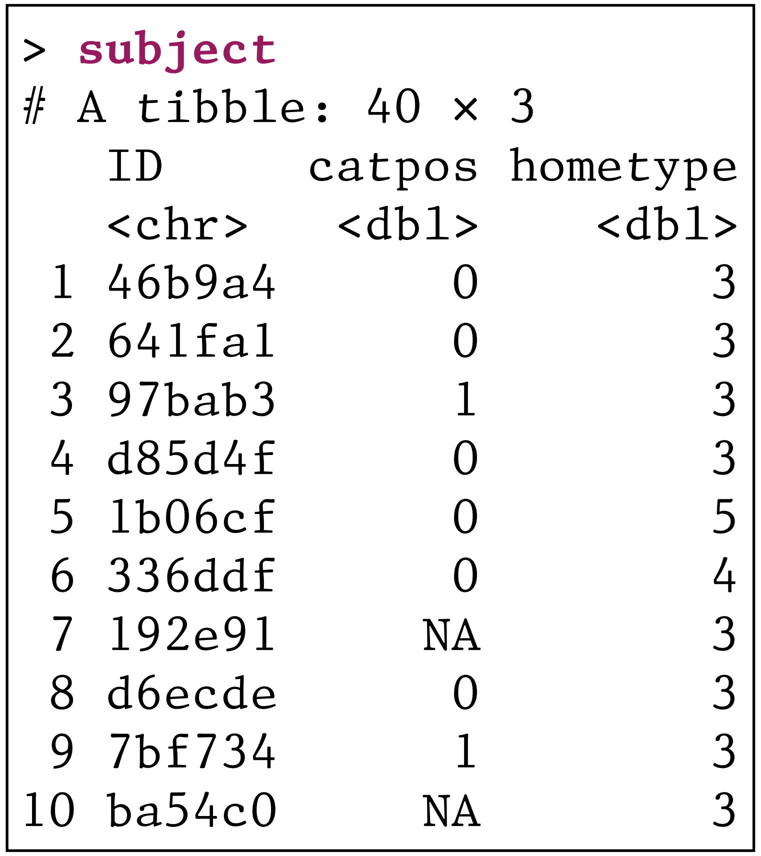
Joining Tables

```
> subject
# A tibble: 40 × 3
   ID      catpos hometype
   <chr>    <dbl>    <dbl>
 1 46b9a4       0        3
 2 641fa1       0        3
 3 97bab3       1        3
 4 d85d4f       0        3
 5 1b06cf       0        5
 6 336ddf       0        4
 7 192e91      NA        3
 8 d6ecde       0        3
 9 7bf734       1        3
10 ba54c0      NA        3
```

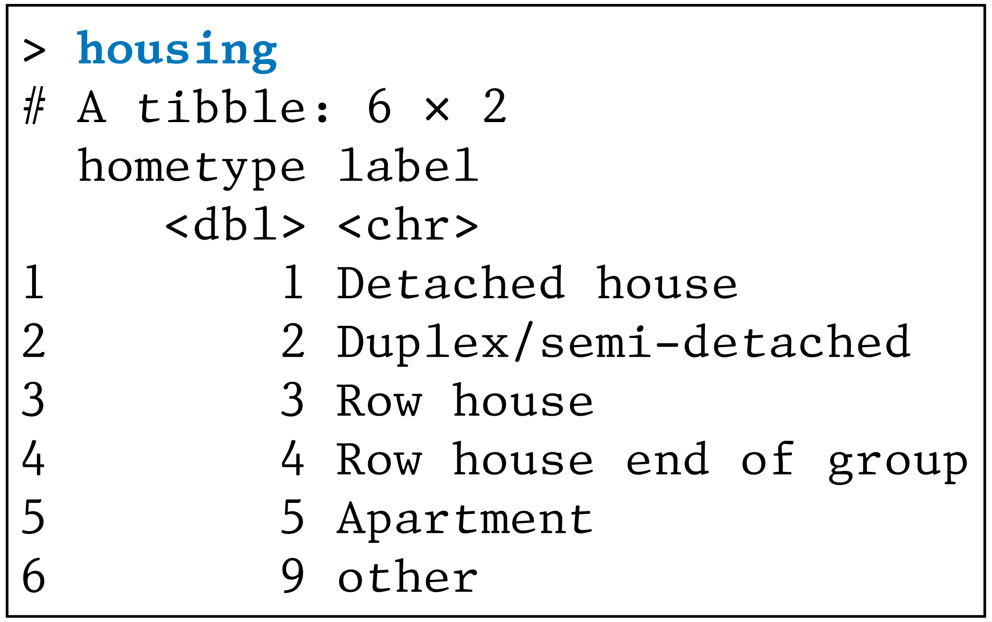
```
subject |>
        left_join(housing, by = "hometype")
```

```
> housing
# A tibble: 6 × 2
  hometype label
     <dbl> <chr>
1        1 Detached house
2        2 Duplex/semi-detached
3        3 Row house
4        4 Row house end of group
5        5 Apartment
6        9 other
```

Joining Tables



```
> subject
# A tibble: 40 × 3
   ID     catpos hometype
   <chr>   <dbl>    <dbl>
 1 46b9a4      0        3
 2 641fa1      0        3
 3 97bab3      1        3
 4 d85d4f      0        3
 5 1b06cf      0        5
 6 336ddf      0        4
 7 192e91     NA        3
 8 d6ecde      0        3
 9 7bf734      1        3
10 ba54c0     NA        3
```

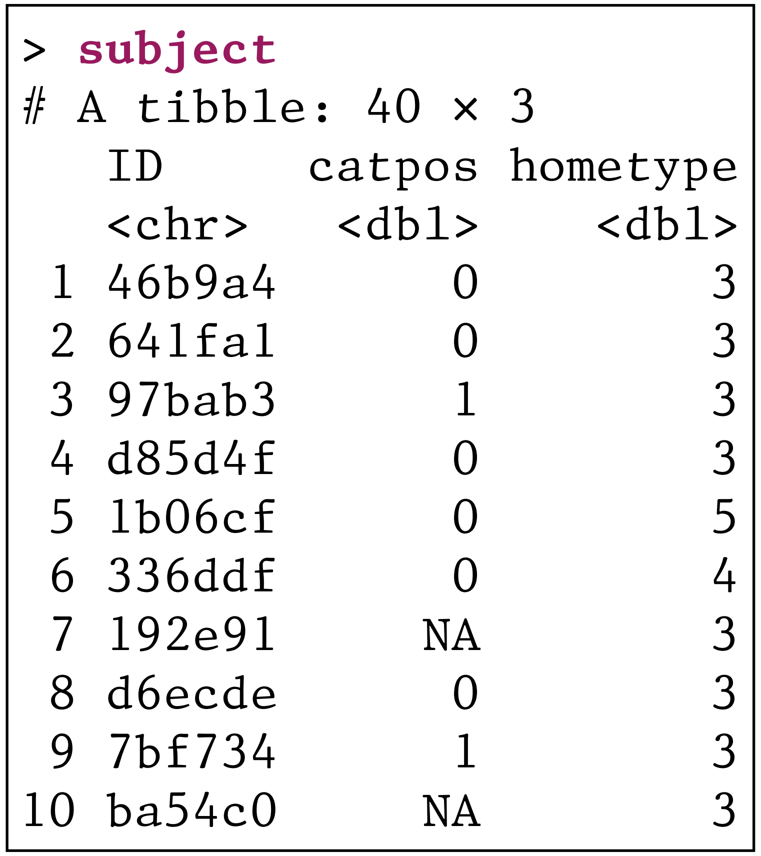


```
> housing
# A tibble: 6 × 2
  hometype label
     <dbl> <chr>
1        1 Detached house
2        2 Duplex/semi-detached
3        3 Row house
4        4 Row house end of group
5        5 Apartment
6        9 other
```

```
subject |>
      left_join(housing, by = "hometype")
```

Column that both tables have in common (must have same name)

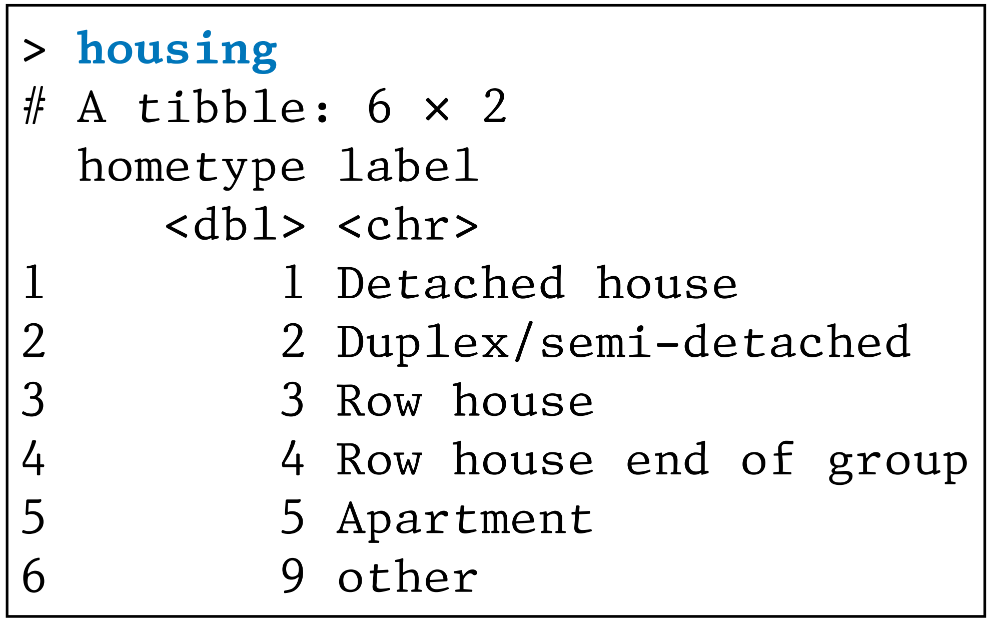
Joining Tables



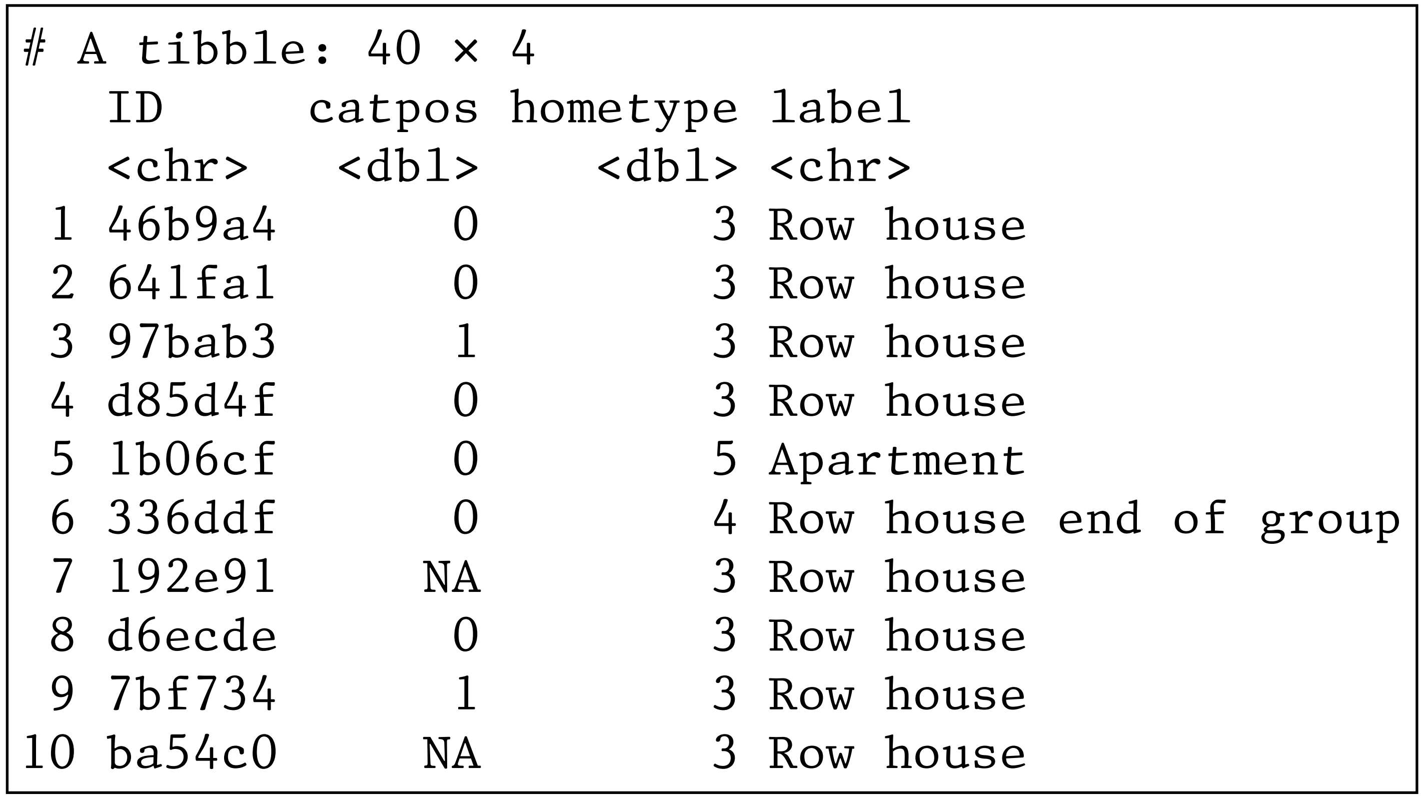
```
> subject
# A tibble: 40 × 3
   ID     catpos hometype
   <chr>   <dbl>    <dbl>
 1 46b9a4      0        3
 2 641fa1      0        3
 3 97bab3      1        3
 4 d85d4f      0        3
 5 1b06cf      0        5
 6 336ddf      0        4
 7 192e91     NA        3
 8 d6ecde      0        3
 9 7bf734      1        3
10 ba54c0     NA        3
```

```
subject |>
        left_join(housing, by = "hometype")
```

Column that both tables have in common (must have same name)



```
> housing
# A tibble: 6 × 2
  hometype label
     <dbl> <chr>
1        1 Detached house
2        2 Duplex/semi-detached
3        3 Row house
4        4 Row house end of group
5        5 Apartment
6        9 other
```



```
# A tibble: 40 × 4
   ID     catpos hometype label
   <chr>   <dbl>    <dbl> <chr>
 1 46b9a4      0        3 Row house
 2 641fa1      0        3 Row house
 3 97bab3      1        3 Row house
 4 d85d4f      0        3 Row house
 5 1b06cf      0        5 Apartment
 6 336ddf      0        4 Row house end of group
 7 192e91     NA        3 Row house
 8 d6ecde      0        3 Row house
 9 7bf734      1        3 Row house
10 ba54c0     NA        3 Row house
```

Cat Allergy by Home Type (labelled)

```
subject |>
    left_join(housing, by = "hometype") |>
    ggplot(aes(x = label, y = catpos)) +
    geom_bar(stat = "summary", fun = "mean") +
    labs(x = NULL,
         y = "Proportion Cat Allergic")
```

Cat Allergy by Home Type (labelled)

```
subject |>
    left_join(housing, by = "hometype") |>
    ggplot(aes(x = label, y = catpos)) +
    geom_bar(stat = "summary", fun = "mean") +
    labs(x = NULL,
         y = "Proportion Cat Allergic")
```

Cat Allergy by Home Type (labelled)

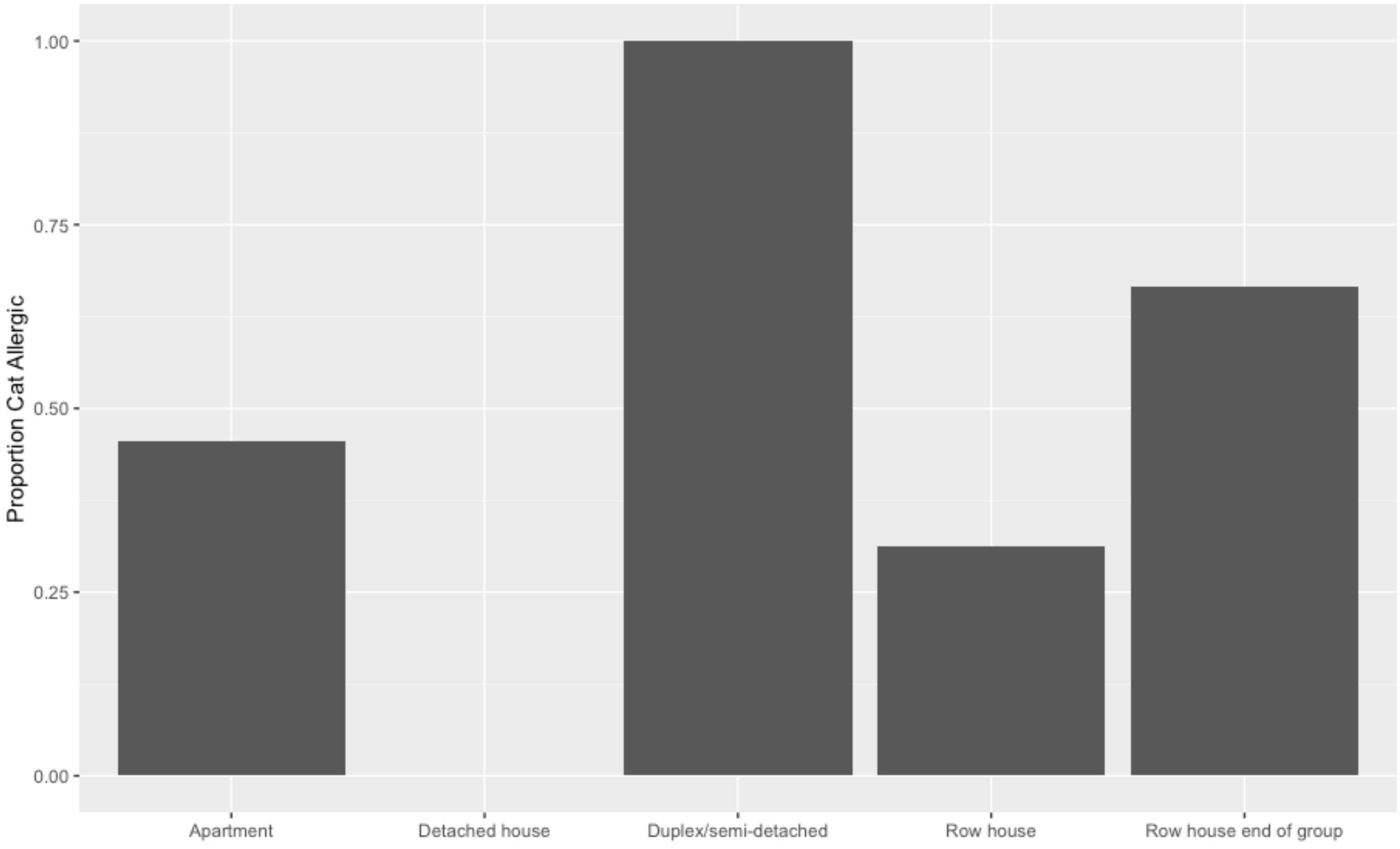
```
subject |>
    left_join(housing, by = "hometype") |>
    ggplot(aes(x = label, y = catpos)) +
    geom_bar(stat = "summary", fun = "mean") +
    labs(x = NULL,
         y = "Proportion Cat Allergic")
```

Remove x-axis label

Cat Allergy by Home Type (labelled)

```
subject |>
    left_join(housing, by = "hometype") |>
    ggplot(aes(x = label, y = catpos)) +
    geom_bar(stat = "summary", fun = "mean") +
    labs(x = NULL,
         y = "Proportion Cat Allergic")
```

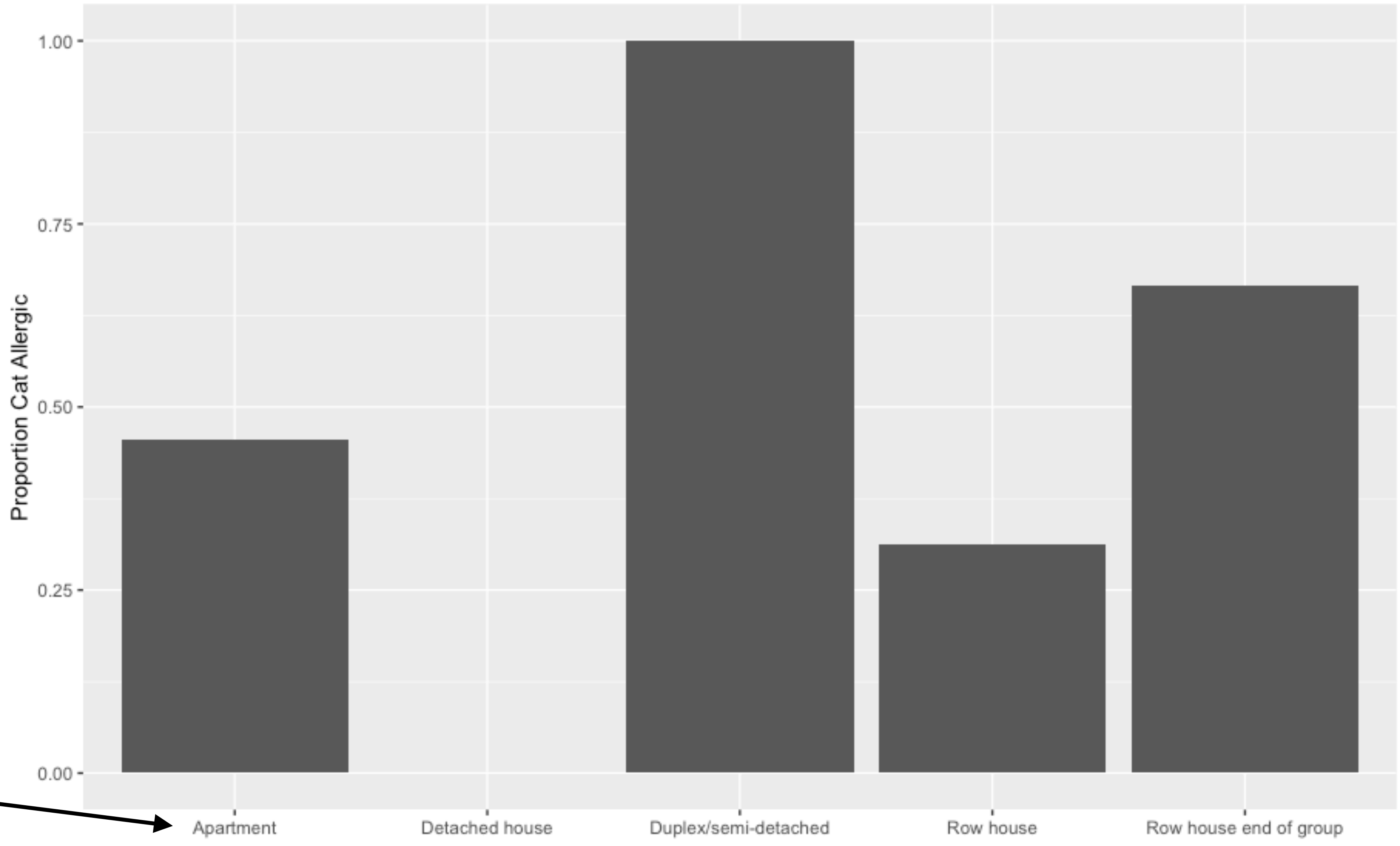
Remove x-axis label



Cat Allergy by Home Type (labelled)

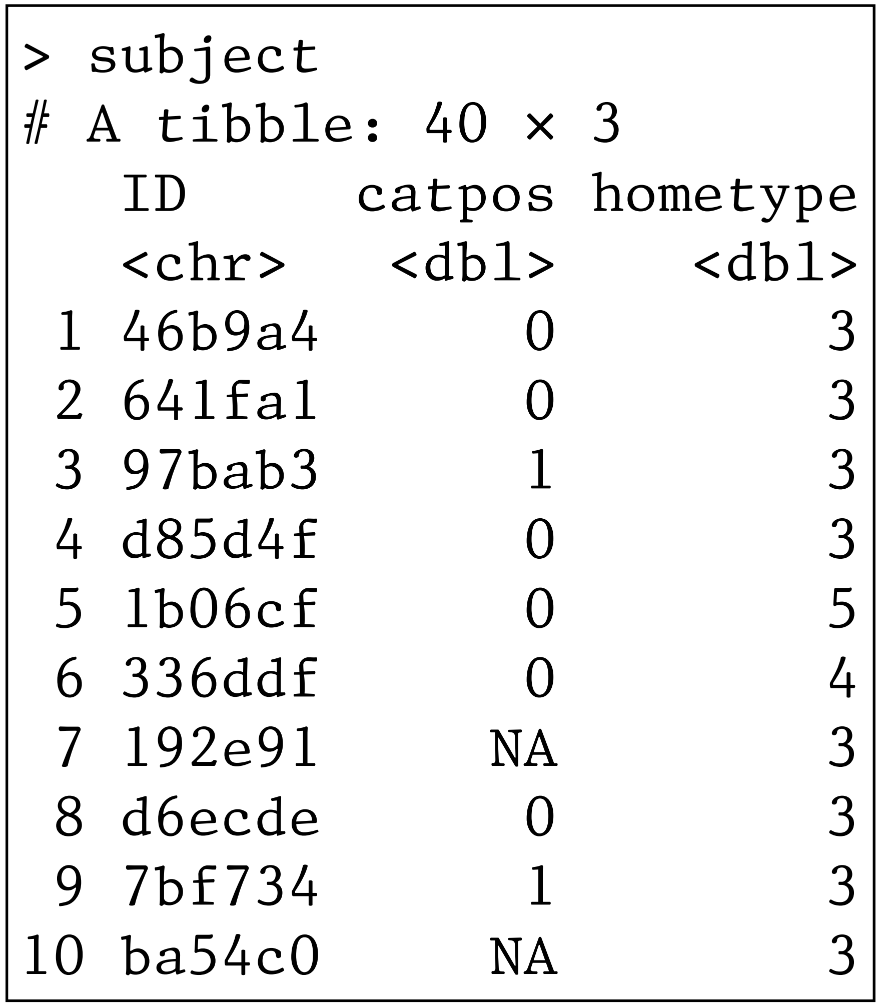
```
subject |>
    left_join(housing, by = "hometype") |>
    ggplot(aes(x = label, y = catpos)) +
    geom_bar(stat = "summary", fun = "mean") +
    labs(x = NULL,
         y = "Proportion Cat Allergic")
```

Remove x-axis label

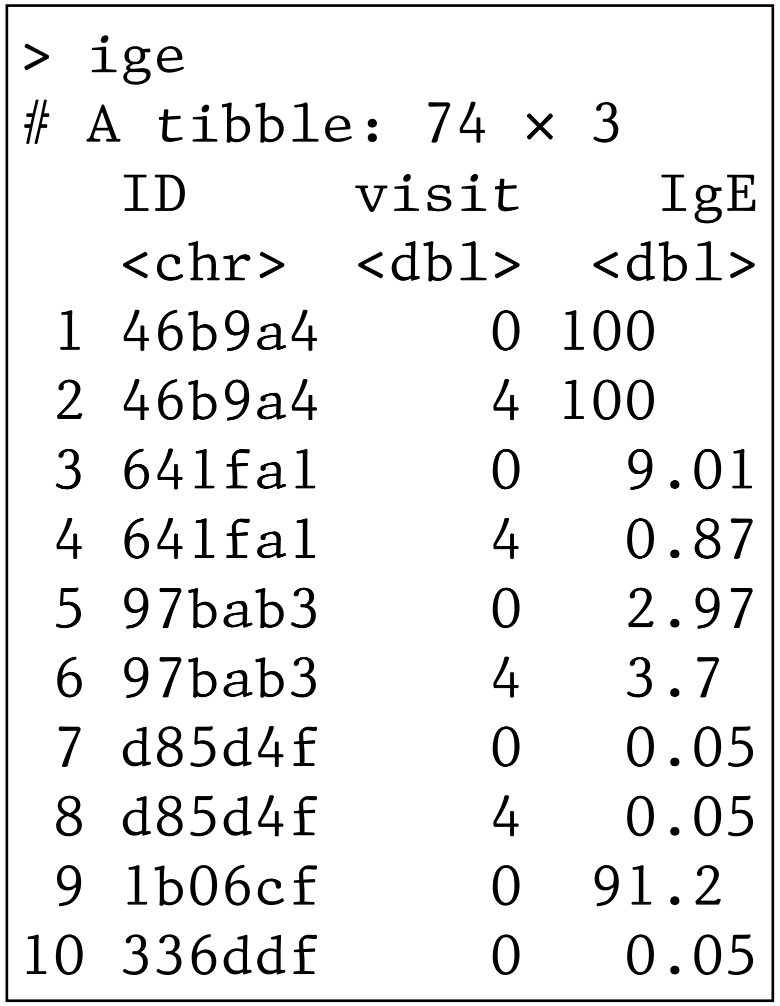


Labels in alphabetical order

Housing Type vs. Baseline IgE

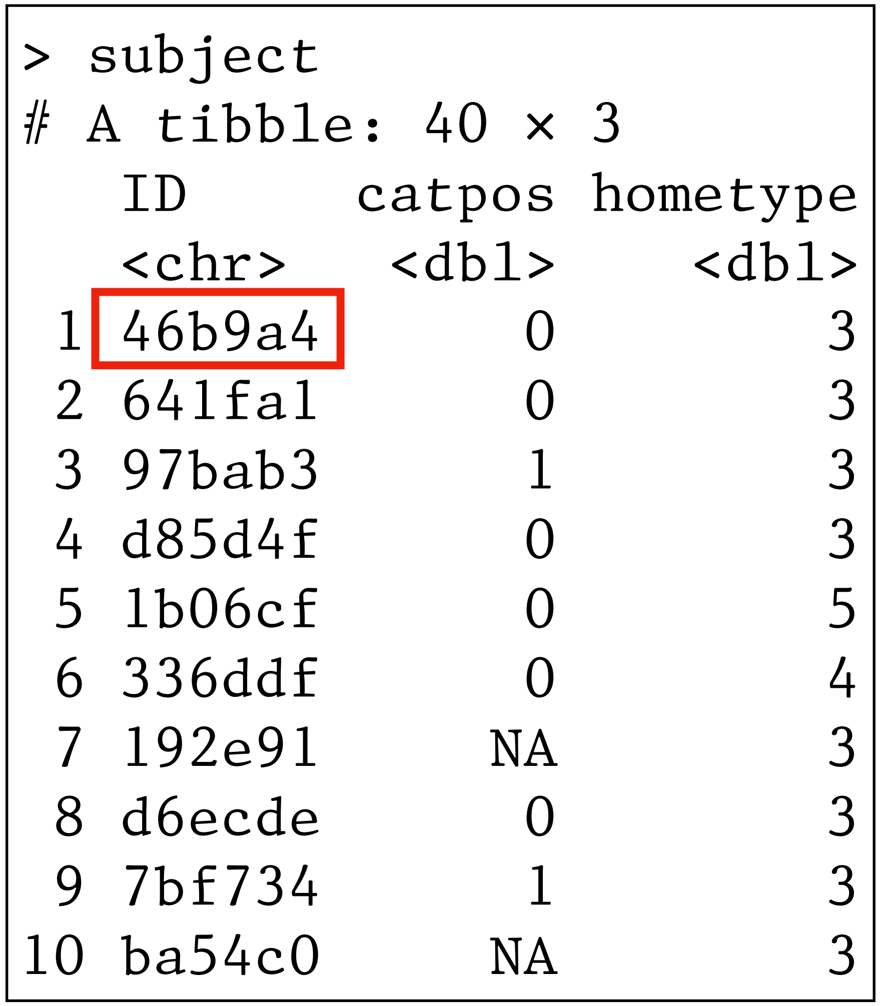


```
> subject
# A tibble: 40 × 3
   ID     catpos hometype
   <chr>   <dbl>    <dbl>
 1 46b9a4      0        3
 2 641fa1      0        3
 3 97bab3      1        3
 4 d85d4f      0        3
 5 1b06cf      0        5
 6 336ddf      0        4
 7 192e91     NA        3
 8 d6ecde      0        3
 9 7bf734      1        3
10 ba54c0     NA        3
```

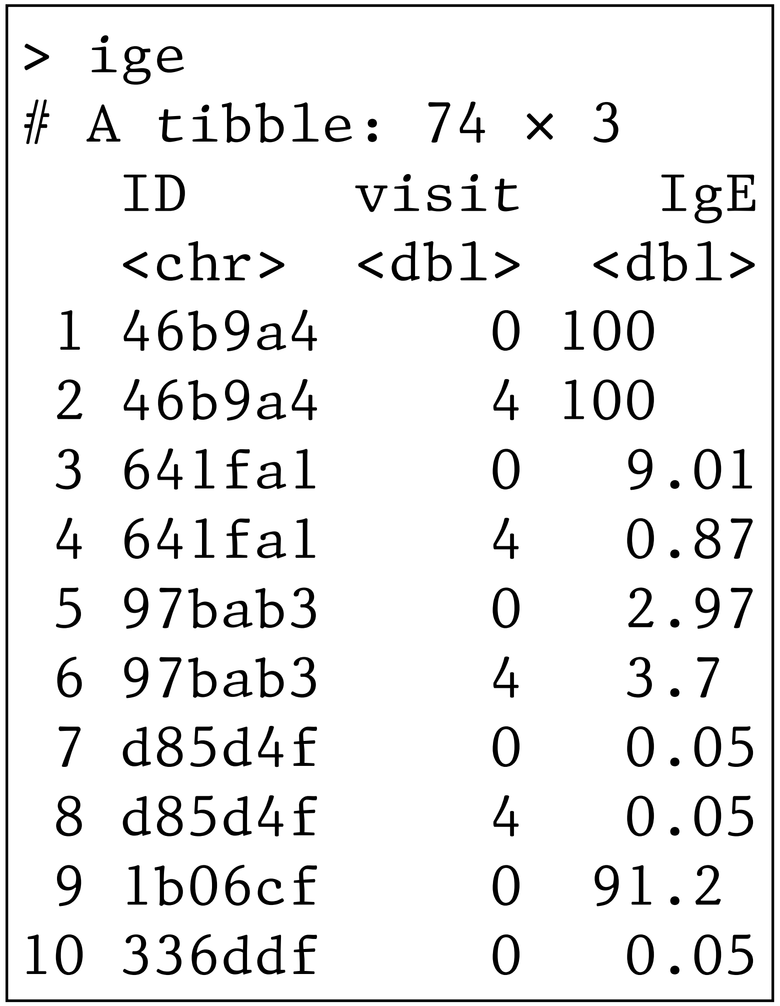


```
> ige
# A tibble: 74 × 3
   ID     visit   IgE
   <chr>  <dbl> <dbl>
 1 46b9a4     0 100  
 2 46b9a4     4 100  
 3 641fa1     0   9.01
 4 641fa1     4   0.87
 5 97bab3     0   2.97
 6 97bab3     4   3.7 
 7 d85d4f     0   0.05
 8 d85d4f     4   0.05
 9 1b06cf     0  91.2 
10 336ddf     0   0.05
```

Housing Type vs. Baseline IgE

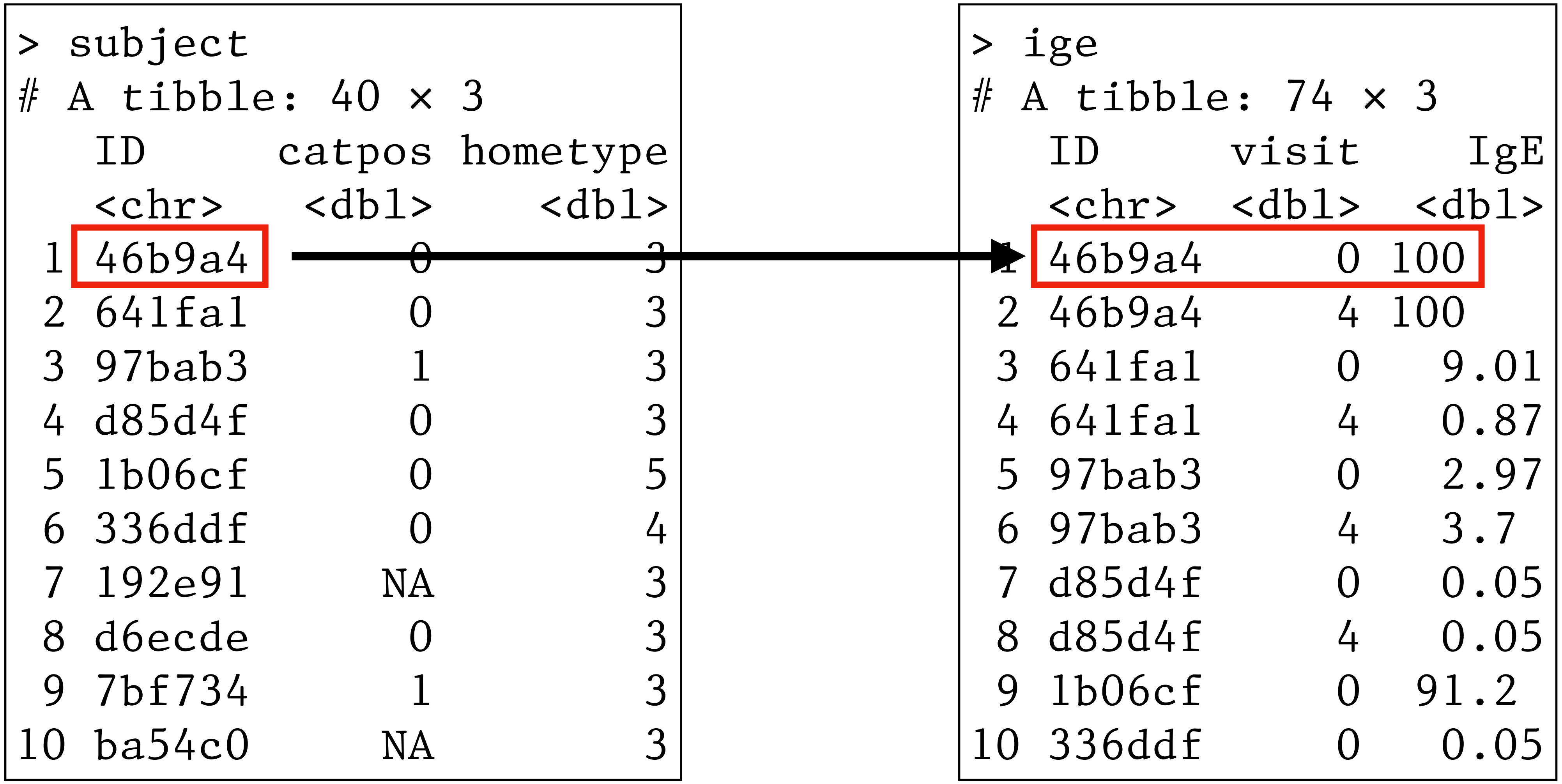


```
> subject
# A tibble: 40 × 3
   ID     catpos hometype
   <chr>   <dbl>    <dbl>
 1 46b9a4      0        3
 2 641fa1      0        3
 3 97bab3      1        3
 4 d85d4f      0        3
 5 1b06cf      0        5
 6 336ddf      0        4
 7 192e91     NA        3
 8 d6ecde      0        3
 9 7bf734      1        3
10 ba54c0     NA        3
```



```
> ige
# A tibble: 74 × 3
   ID     visit    IgE
   <chr>  <dbl>  <dbl>
 1 46b9a4     0 100   
 2 46b9a4     4 100   
 3 641fa1     0   9.01
 4 641fa1     4   0.87
 5 97bab3     0   2.97
 6 97bab3     4   3.7 
 7 d85d4f     0   0.05
 8 d85d4f     4   0.05
 9 1b06cf     0  91.2 
10 336ddf     0   0.05
```

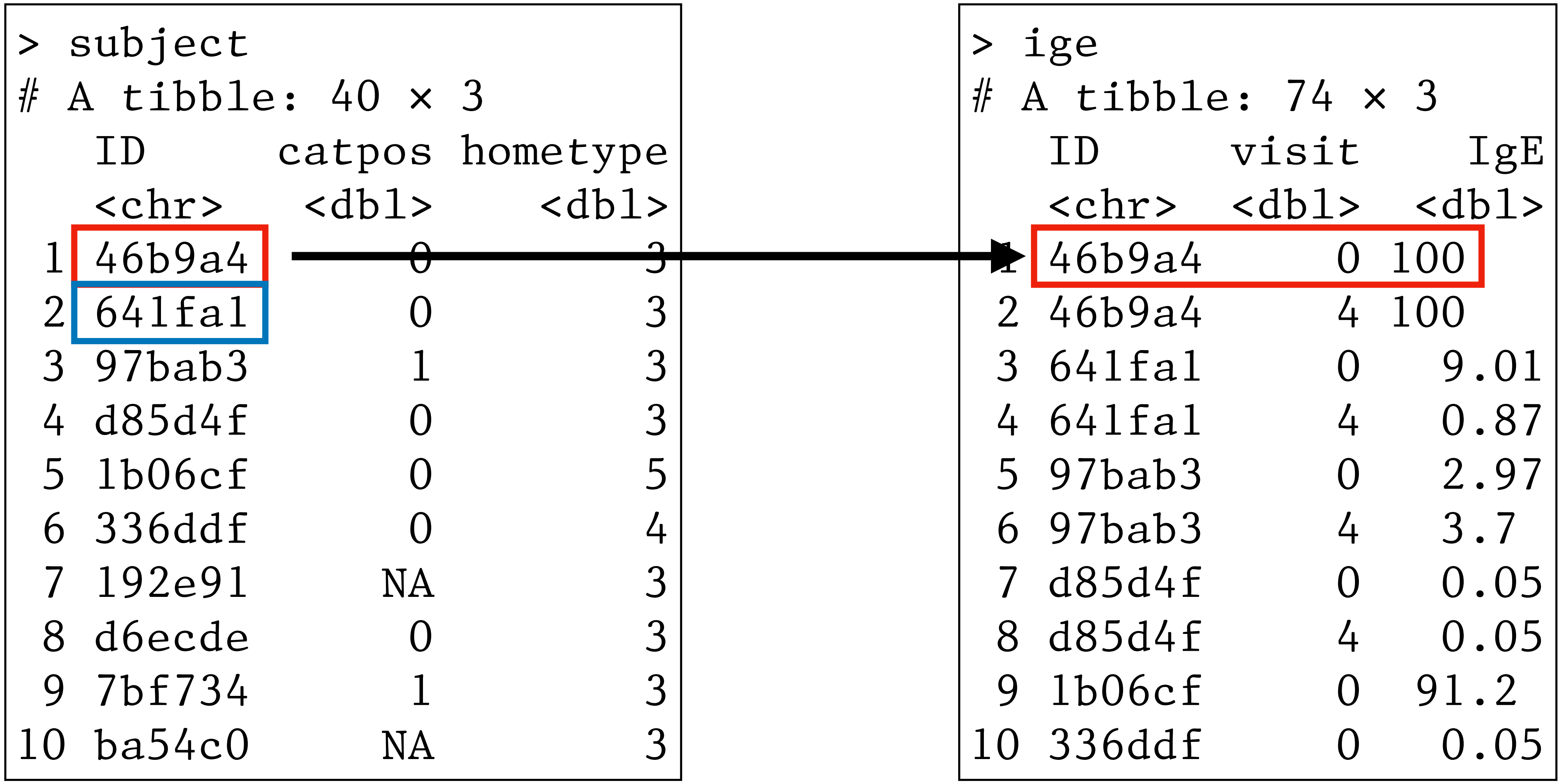
Housing Type vs. Baseline IgE



```
> subject
# A tibble: 40 × 3
   ID     catpos hometype
   <chr>   <dbl>    <dbl>
 1 46b9a4      0        3
 2 641fa1      0        3
 3 97bab3      1        3
 4 d85d4f      0        3
 5 1b06cf      0        5
 6 336ddf      0        4
 7 192e91     NA        3
 8 d6ecde      0        3
 9 7bf734      1        3
10 ba54c0     NA        3
```

```
> ige
# A tibble: 74 × 3
   ID     visit   IgE
   <chr>  <dbl> <dbl>
 1 46b9a4     0 100
 2 46b9a4     4 100
 3 641fa1     0   9.01
 4 641fa1     4   0.87
 5 97bab3     0   2.97
 6 97bab3     4   3.7
 7 d85d4f     0   0.05
 8 d85d4f     4   0.05
 9 1b06cf     0  91.2
10 336ddf     0   0.05
```

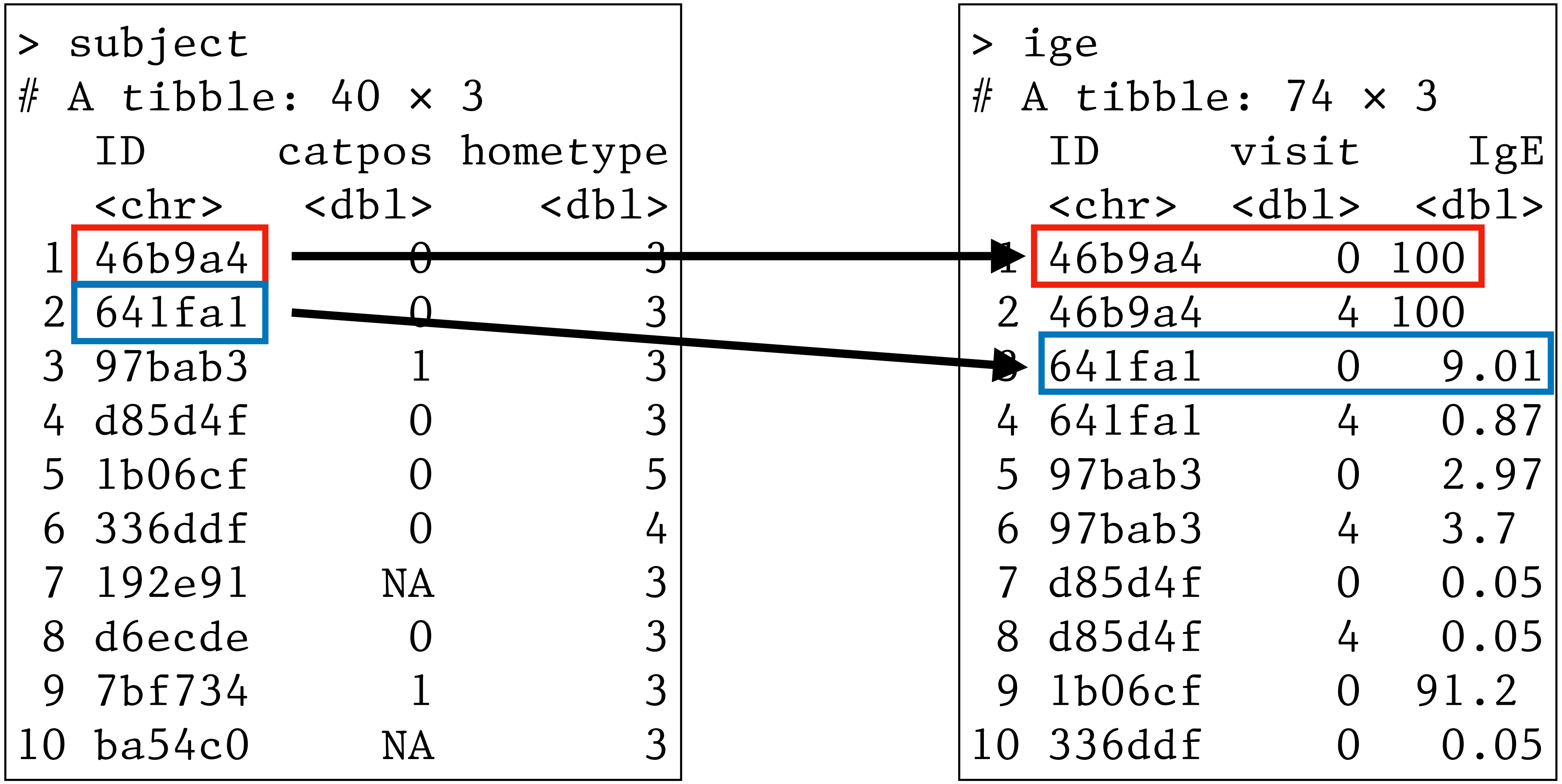
Housing Type vs. Baseline IgE



```
> subject
# A tibble: 40 × 3
   ID     catpos hometype
   <chr>   <dbl>    <dbl>
 1 46b9a4      0        3
 2 641fa1      0        3
 3 97bab3      1        3
 4 d85d4f      0        3
 5 1b06cf      0        5
 6 336ddf      0        4
 7 192e91     NA        3
 8 d6ecde      0        3
 9 7bf734      1        3
10 ba54c0     NA        3
```

```
> ige
# A tibble: 74 × 3
   ID     visit    IgE
   <chr>  <dbl>  <dbl>
 1 46b9a4     0 100
 2 46b9a4     4 100
 3 641fa1     0   9.01
 4 641fa1     4   0.87
 5 97bab3     0   2.97
 6 97bab3     4   3.7
 7 d85d4f     0   0.05
 8 d85d4f     4   0.05
 9 1b06cf     0  91.2
10 336ddf     0   0.05
```

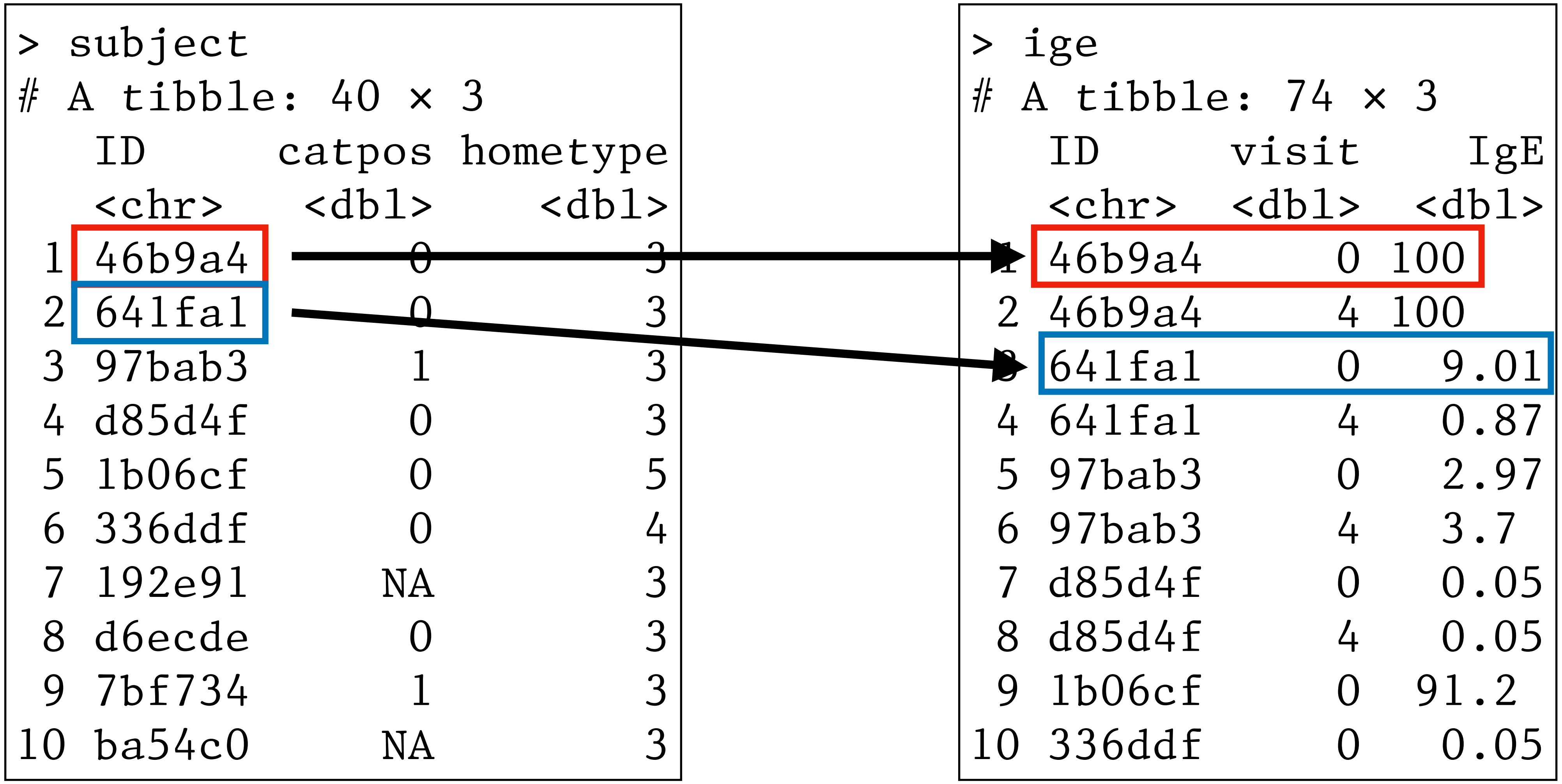
Housing Type vs. Baseline IgE



```
> subject
# A tibble: 40 × 3
   ID       catpos hometype
   <chr>     <dbl>    <dbl>
 1 46b9a4        0        3
 2 641fa1        0        3
 3 97bab3        1        3
 4 d85d4f        0        3
 5 1b06cf        0        5
 6 336ddf        0        4
 7 192e91       NA        3
 8 d6ecde        0        3
 9 7bf734        1        3
10 ba54c0       NA        3
```

```
> ige
# A tibble: 74 × 3
   ID       visit    IgE
   <chr>    <dbl>  <dbl>
 1 46b9a4       0 100
 2 46b9a4       4 100
 3 641fa1       0   9.01
 4 641fa1       4   0.87
 5 97bab3       0   2.97
 6 97bab3       4   3.7
 7 d85d4f       0   0.05
 8 d85d4f       4   0.05
 9 1b06cf       0  91.2
10 336ddf       0   0.05
```

Housing Type vs. Baseline IgE



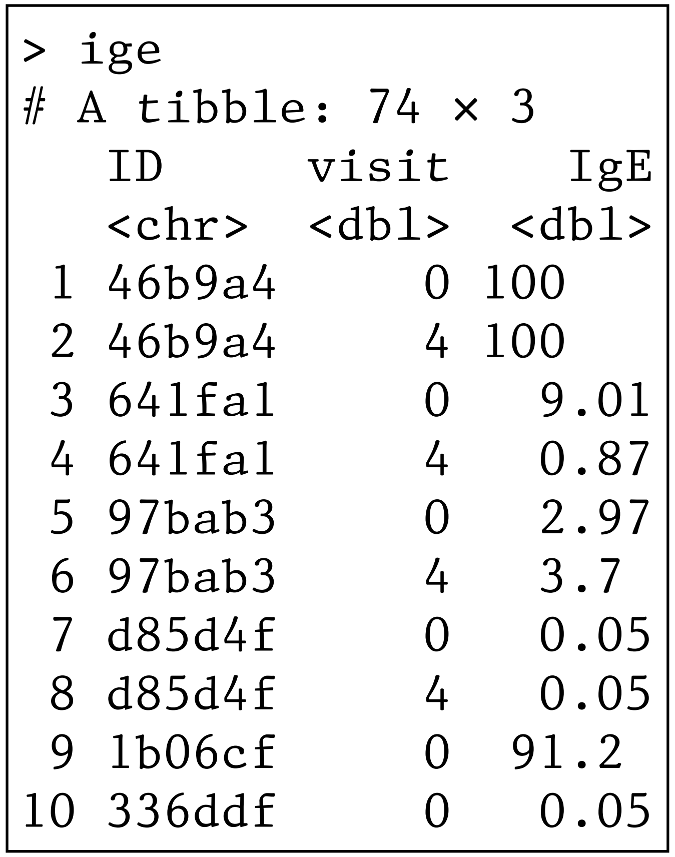
```
> subject
# A tibble: 40 × 3
   ID     catpos hometype
   <chr>   <dbl>    <dbl>
 1 46b9a4      0        3
 2 641fa1      0        3
 3 97bab3      1        3
 4 d85d4f      0        3
 5 1b06cf      0        5
 6 336ddf      0        4
 7 192e91     NA        3
 8 d6ecde      0        3
 9 7bf734      1        3
10 ba54c0     NA        3
```

```
> ige
# A tibble: 74 × 3
   ID     visit    IgE
   <chr>  <dbl>  <dbl>
 1 46b9a4     0 100
 2 46b9a4     4 100
 3 641fa1     0   9.01
 4 641fa1     4   0.87
 5 97bab3     0   2.97
 6 97bab3     4   3.7
 7 d85d4f     0   0.05
 8 d85d4f     4   0.05
 9 1b06cf     0  91.2
10 336ddf     0   0.05
```

1. **Filter rows for visit == 0**
2. **Remove visit column**
3. **left join with 'subject' table by ID column**

Housing Type vs. Baseline IgE

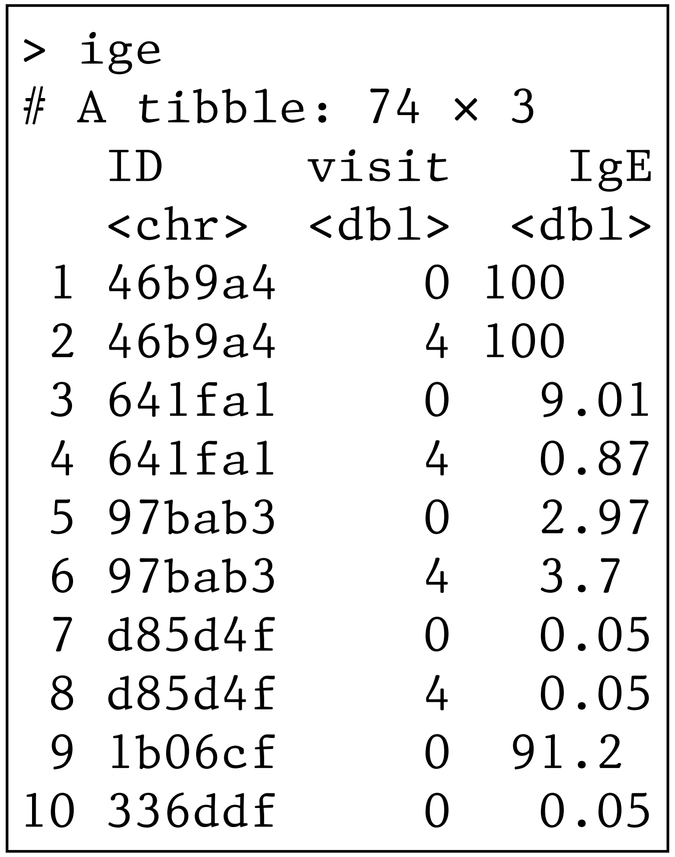
```
ige |>
    filter(visit == 0) |>
    select(-visit) |>
    left_join(subject, by = "ID") |>
    left_join(housing, by = "hometype")
```



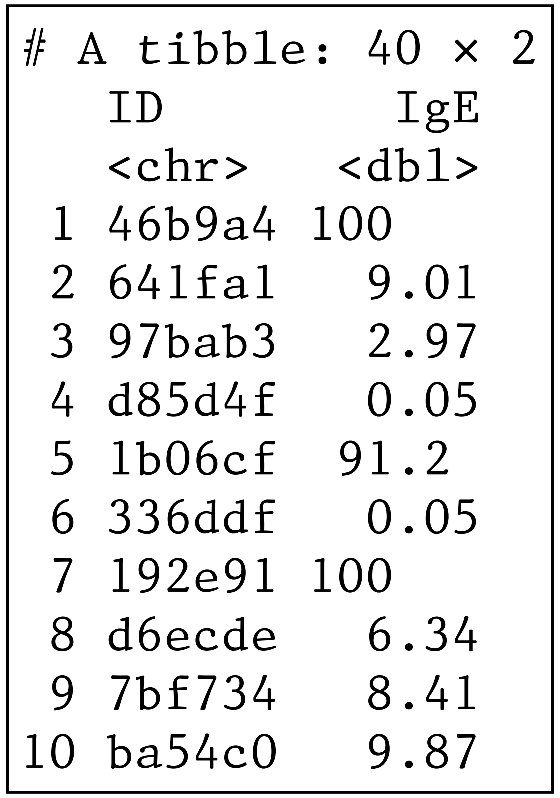
```
> ige
# A tibble: 74 × 3
   ID     visit    IgE
   <chr>  <dbl>  <dbl>
 1 46b9a4     0 100   
 2 46b9a4     4 100   
 3 641fa1     0   9.01
 4 641fa1     4   0.87
 5 97bab3     0   2.97
 6 97bab3     4   3.7 
 7 d85d4f     0   0.05
 8 d85d4f     4   0.05
 9 1b06cf     0  91.2 
10 336ddf     0   0.05
```

Housing Type vs. Baseline IgE

```
ige |>
    filter(visit == 0) |>
    select(-visit) |>
    left_join(subject, by = "ID") |>
    left_join(housing, by = "hometype")
```



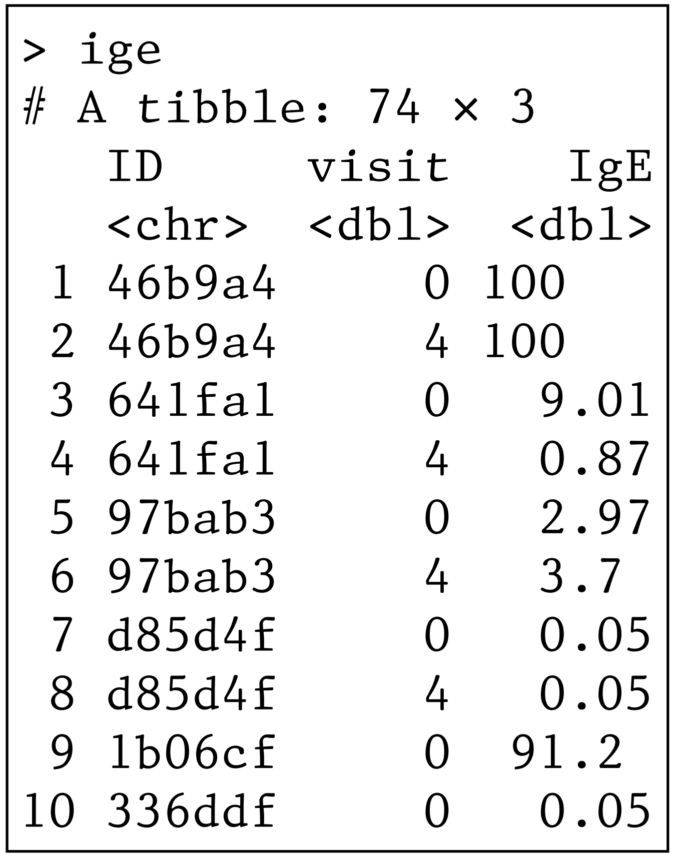
```
> ige
# A tibble: 74 × 3
   ID     visit    IgE
   <chr>  <dbl>  <dbl>
 1 46b9a4     0 100
 2 46b9a4     4 100
 3 641fa1     0   9.01
 4 641fa1     4   0.87
 5 97bab3     0   2.97
 6 97bab3     4   3.7
 7 d85d4f     0   0.05
 8 d85d4f     4   0.05
 9 1b06cf     0  91.2
10 336ddf     0   0.05
```



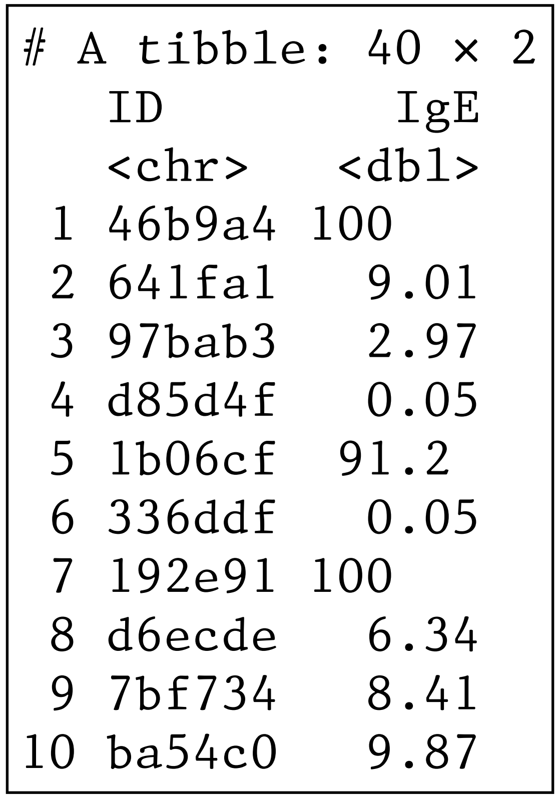
```
# A tibble: 40 × 2
   ID        IgE
   <chr>   <dbl>
 1 46b9a4 100
 2 641fa1   9.01
 3 97bab3   2.97
 4 d85d4f   0.05
 5 1b06cf  91.2
 6 336ddf   0.05
 7 192e91 100
 8 d6ecde   6.34
 9 7bf734   8.41
10 ba54c0   9.87
```

Housing Type vs. Baseline IgE

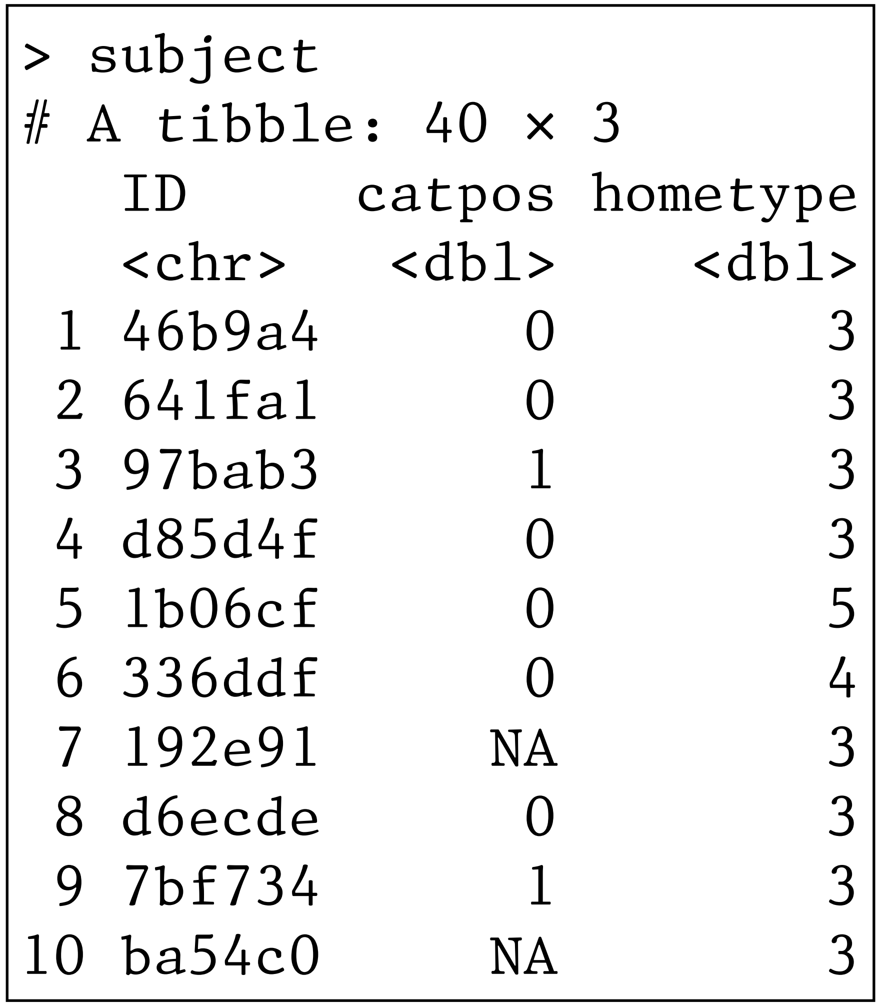
```
ige |>
    filter(visit == 0) |>
    select(-visit) |>
    left_join(subject, by = "ID") |>
    left_join(housing, by = "hometype")
```



```
> ige
# A tibble: 74 × 3
   ID      visit   IgE
   <chr>   <dbl> <dbl>
 1 46b9a4      0 100  
 2 46b9a4      4 100  
 3 641fa1      0   9.01
 4 641fa1      4   0.87
 5 97bab3      0   2.97
 6 97bab3      4   3.7 
 7 d85d4f      0   0.05
 8 d85d4f      4   0.05
 9 1b06cf      0  91.2 
10 336ddf      0   0.05
```



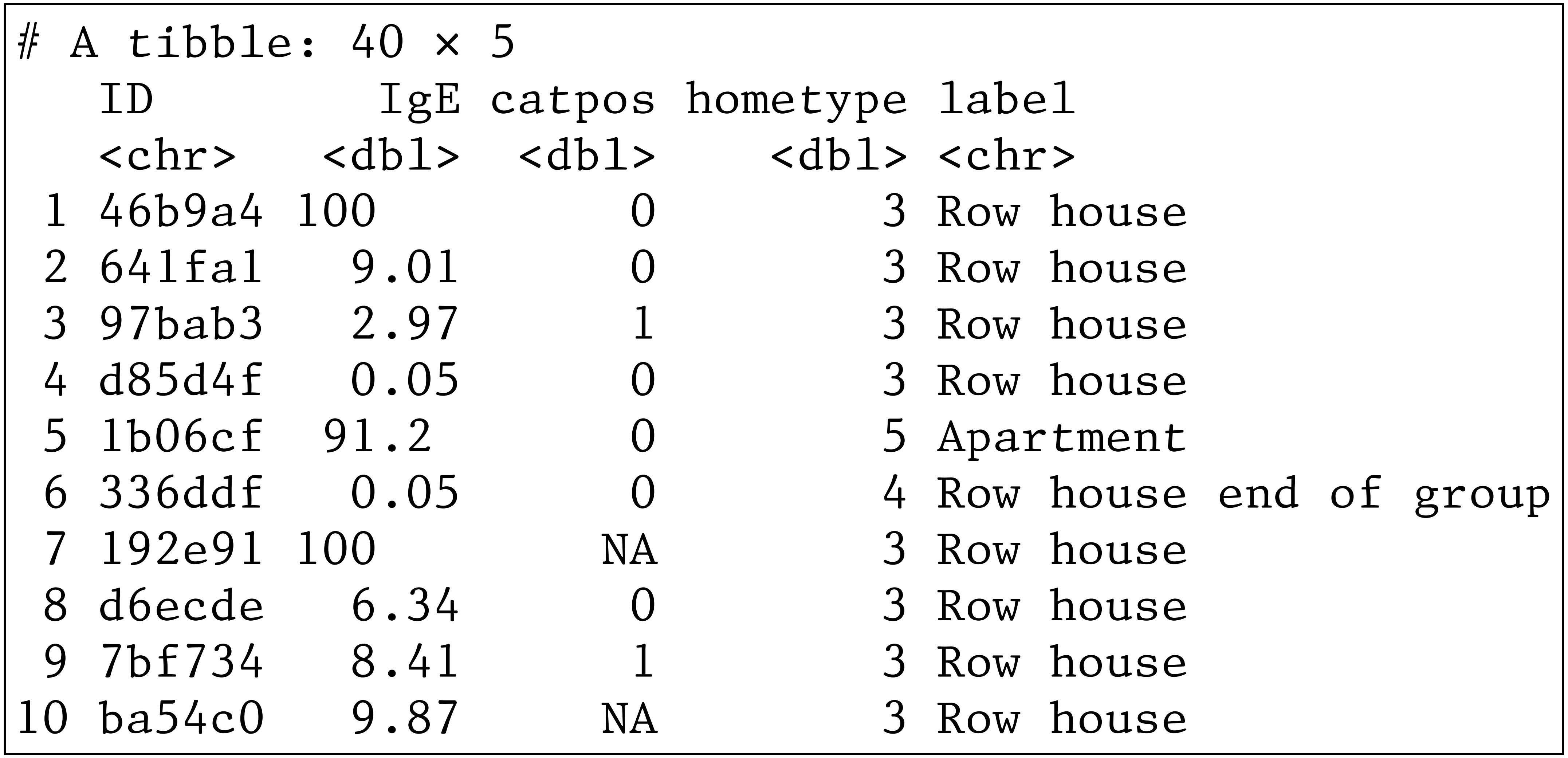
```
# A tibble: 40 × 2
   ID        IgE
   <chr>   <dbl>
 1 46b9a4 100   
 2 641fa1   9.01
 3 97bab3   2.97
 4 d85d4f   0.05
 5 1b06cf  91.2 
 6 336ddf   0.05
 7 192e91 100   
 8 d6ecde   6.34
 9 7bf734   8.41
10 ba54c0   9.87
```



```
> subject
# A tibble: 40 × 3
   ID     catpos hometype
   <chr>   <dbl>    <dbl>
 1 46b9a4      0        3
 2 641fa1      0        3
 3 97bab3      1        3
 4 d85d4f      0        3
 5 1b06cf      0        5
 6 336ddf      0        4
 7 192e91     NA        3
 8 d6ecde      0        3
 9 7bf734      1        3
10 ba54c0     NA        3
```

Housing Type vs. Baseline IgE

```
ige |>
    filter(visit == 0) |>
    select(-visit) |>
    left_join(subject, by = "ID") |>
    left_join(housing, by = "hometype")
```



```
# A tibble: 40 × 5
   ID        IgE catpos hometype label
   <chr>   <dbl>  <dbl>    <dbl> <chr>
 1 46b9a4 100         0        3 Row house
 2 641fa1   9.01      0        3 Row house
 3 97bab3   2.97      1        3 Row house
 4 d85d4f   0.05      0        3 Row house
 5 1b06cf  91.2       0        5 Apartment
 6 336ddf   0.05      0        4 Row house end of group
 7 192e91 100        NA        3 Row house
 8 d6ecde   6.34      0        3 Row house
 9 7bf734   8.41      1        3 Row house
10 ba54c0   9.87     NA        3 Row house
```

Housing Type vs. Baseline IgE

```
ige |>
    filter(visit == 0) |>
    select(-visit) |>
    left_join(subject, by = "ID") |>
    left_join(housing, by = "hometype")
```

ID column no longer needed

```
# A tibble: 40 × 5
   ID         IgE catpos hometype label
   <chr>    <dbl>  <dbl>    <dbl> <chr>
 1 46b9a4 100          0        3 Row house
 2 641fa1   9.01       0        3 Row house
 3 97bab3   2.97       1        3 Row house
 4 d85d4f   0.05       0        3 Row house
 5 1b06cf  91.2        0        5 Apartment
 6 336ddf   0.05       0        4 Row house end of group
 7 192e91 100         NA        3 Row house
 8 d6ecde   6.34       0        3 Row house
 9 7bf734   8.41       1        3 Row house
10 ba54c0   9.87      NA        3 Row house
```

Housing Type vs. Baseline IgE

```
ige |>
    filter(visit == 0) |>
    select(-visit) |>
    left_join(subject, by = "ID") |>
    left_join(housing, by = "hometype") |>
    ggplot(aes(x = label, y = IgE)) +
    geom_bar(stat = "summary", fun = "mean") +
    geom_errorbar(stat = "summary", fun.data = "mean_se", width = 0.3)
```

Compute average IgE

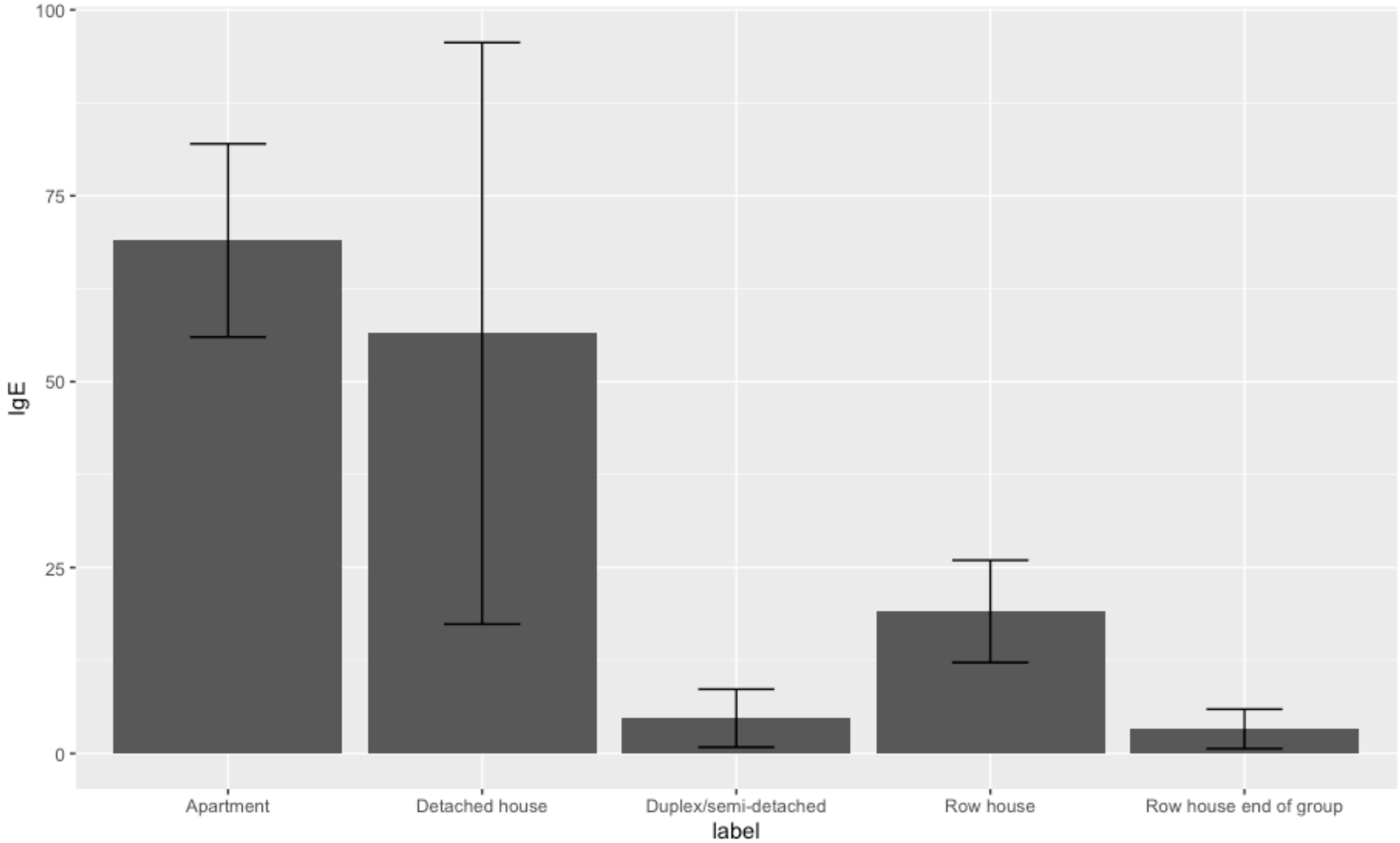
Compute ±1 Std. Error

Housing Type vs. Baseline IgE

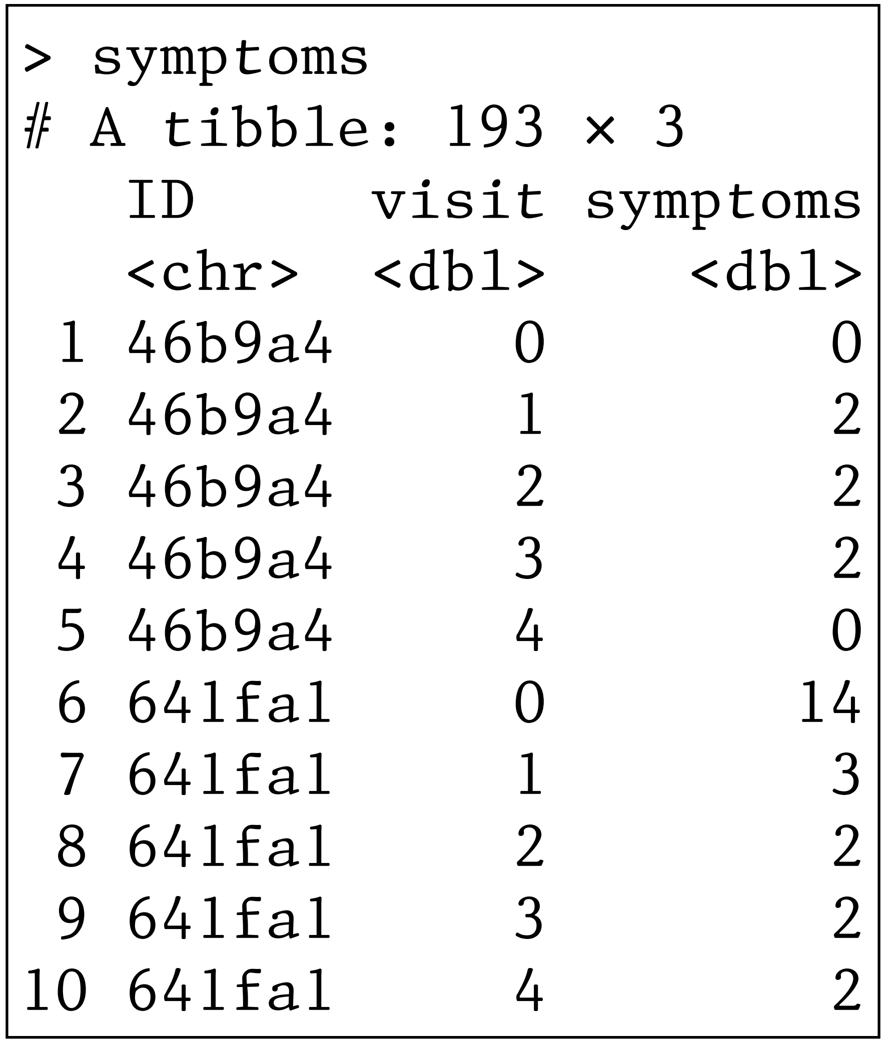
```
ige |>
    filter(visit == 0) |>
    select(-visit) |>
    left_join(subject, by = "ID") |>
    left_join(housing, by = "hometype") |>
    ggplot(aes(x = label, y = IgE)) +
    geom_bar(stat = "summary", fun = "mean") +
    geom_errorbar(stat = "summary", fun.data = "mean_se", width = 0.3)
```

Compute average IgE

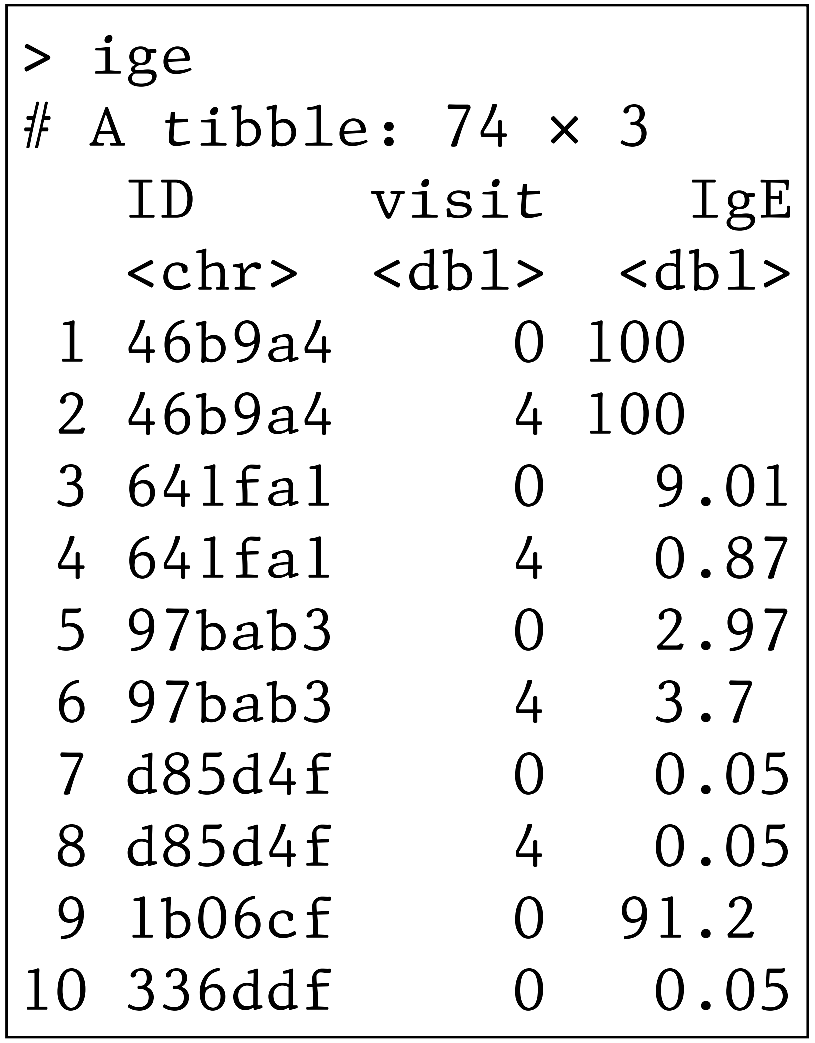
Compute ± 1 Std. Error



Symptoms vs. IgE



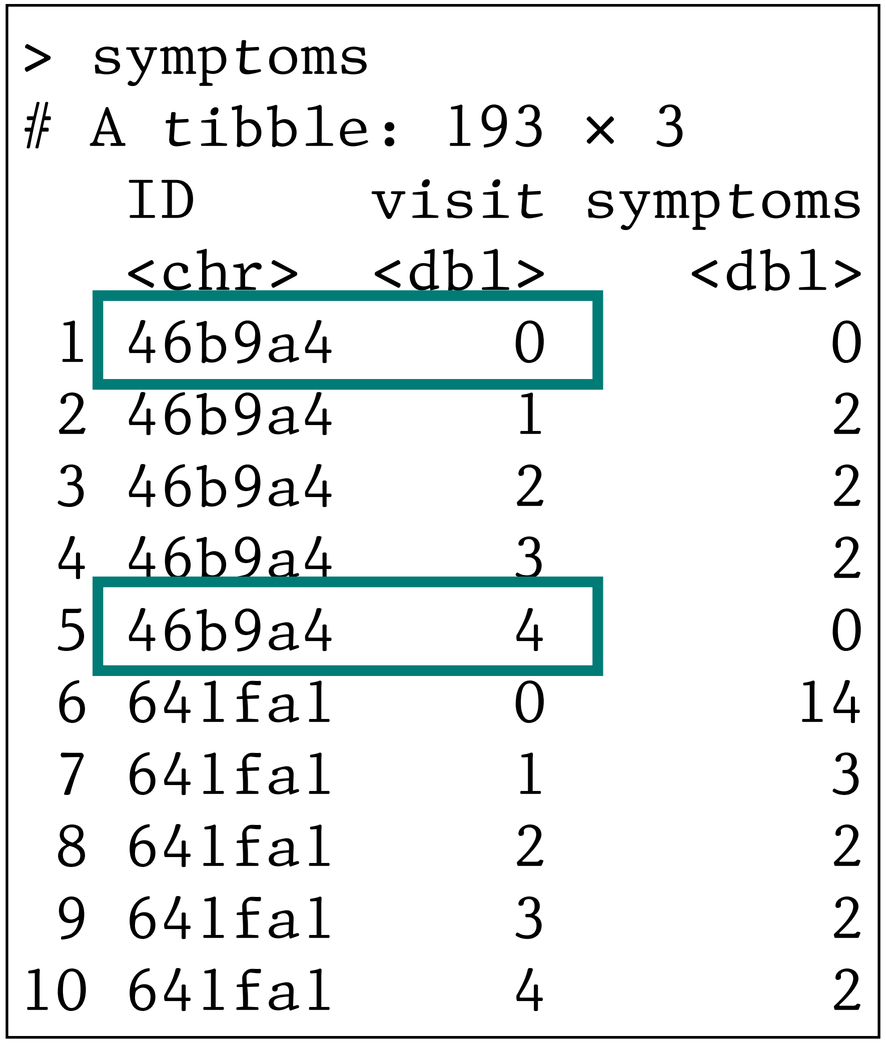
```
> symptoms
# A tibble: 193 × 3
   ID     visit symptoms
   <chr>  <dbl>    <dbl>
 1 46b9a4     0        0
 2 46b9a4     1        2
 3 46b9a4     2        2
 4 46b9a4     3        2
 5 46b9a4     4        0
 6 641fa1     0       14
 7 641fa1     1        3
 8 641fa1     2        2
 9 641fa1     3        2
10 641fa1     4        2
```



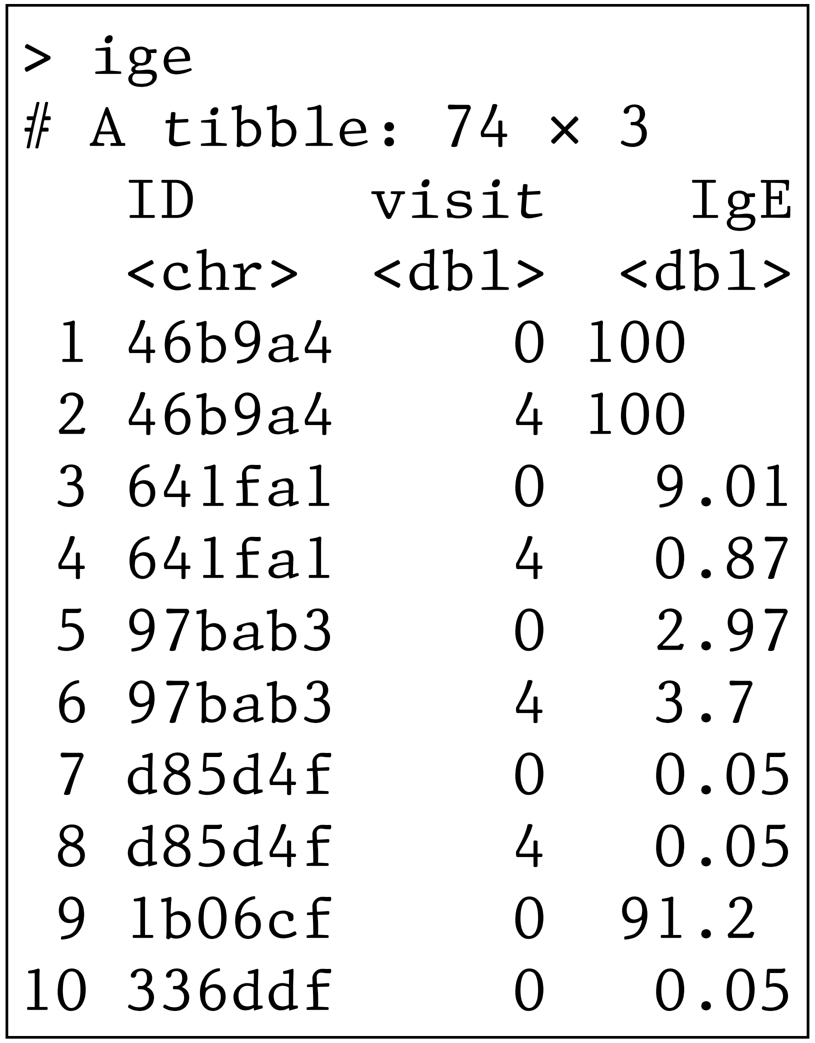
```
> ige
# A tibble: 74 × 3
   ID     visit    IgE
   <chr>  <dbl>  <dbl>
 1 46b9a4     0 100   
 2 46b9a4     4 100   
 3 641fa1     0   9.01
 4 641fa1     4   0.87
 5 97bab3     0   2.97
 6 97bab3     4   3.7 
 7 d85d4f     0   0.05
 8 d85d4f     4   0.05
 9 1b06cf     0  91.2 
10 336ddf     0   0.05
```

Must join by both "ID" and "visit"

Symptoms vs. IgE



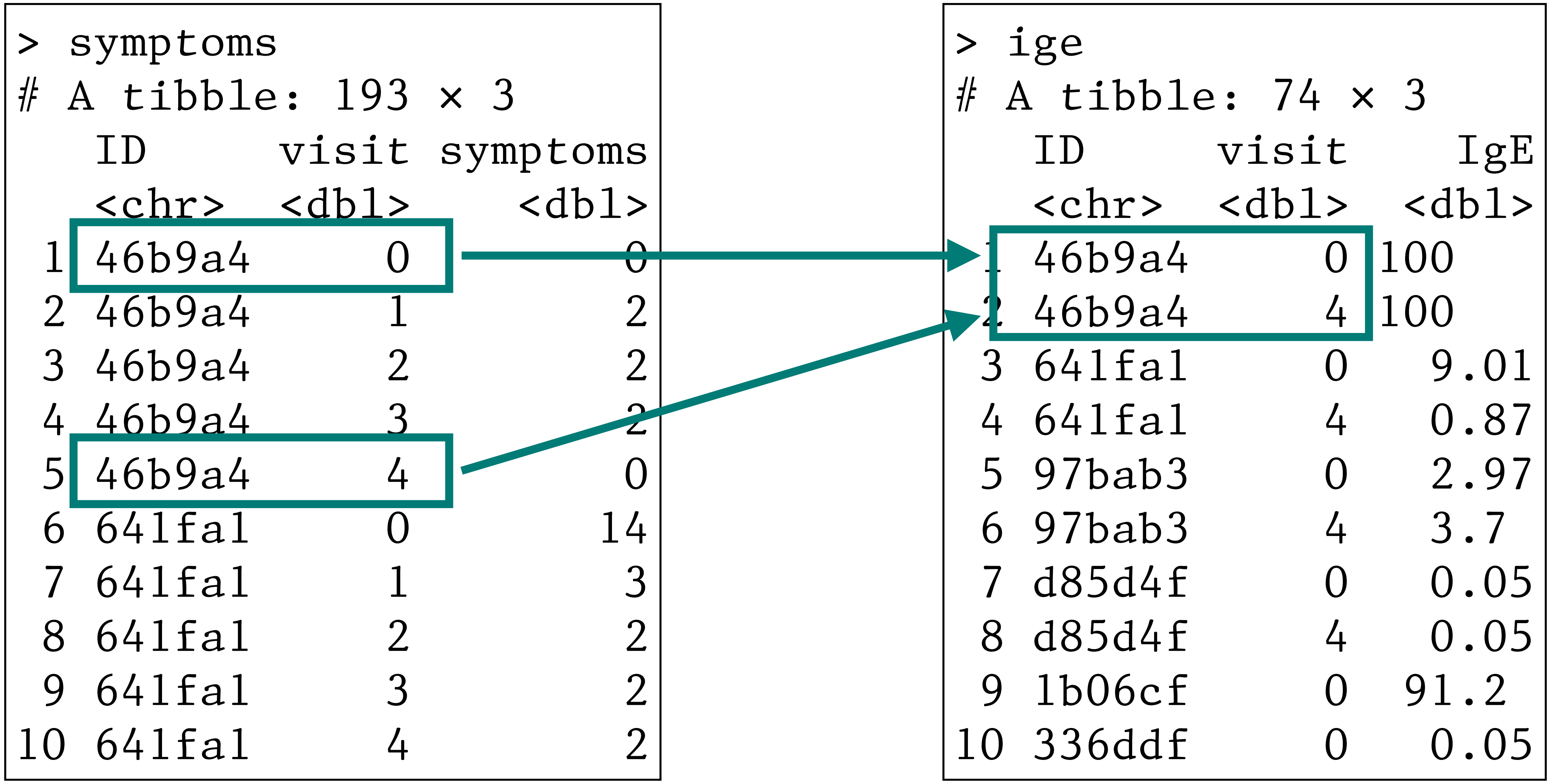
```
> symptoms
# A tibble: 193 × 3
   ID     visit symptoms
   <chr>  <dbl>    <dbl>
 1 46b9a4     0        0
 2 46b9a4     1        2
 3 46b9a4     2        2
 4 46b9a4     3        2
 5 46b9a4     4        0
 6 641fa1     0       14
 7 641fa1     1        3
 8 641fa1     2        2
 9 641fa1     3        2
10 641fa1     4        2
```



```
> ige
# A tibble: 74 × 3
   ID     visit    IgE
   <chr>  <dbl>  <dbl>
 1 46b9a4     0 100   
 2 46b9a4     4 100   
 3 641fa1     0   9.01
 4 641fa1     4   0.87
 5 97bab3     0   2.97
 6 97bab3     4   3.7 
 7 d85d4f     0   0.05
 8 d85d4f     4   0.05
 9 1b06cf     0  91.2 
10 336ddf     0   0.05
```

Must join by both "ID" and "visit"

Symptoms vs. IgE



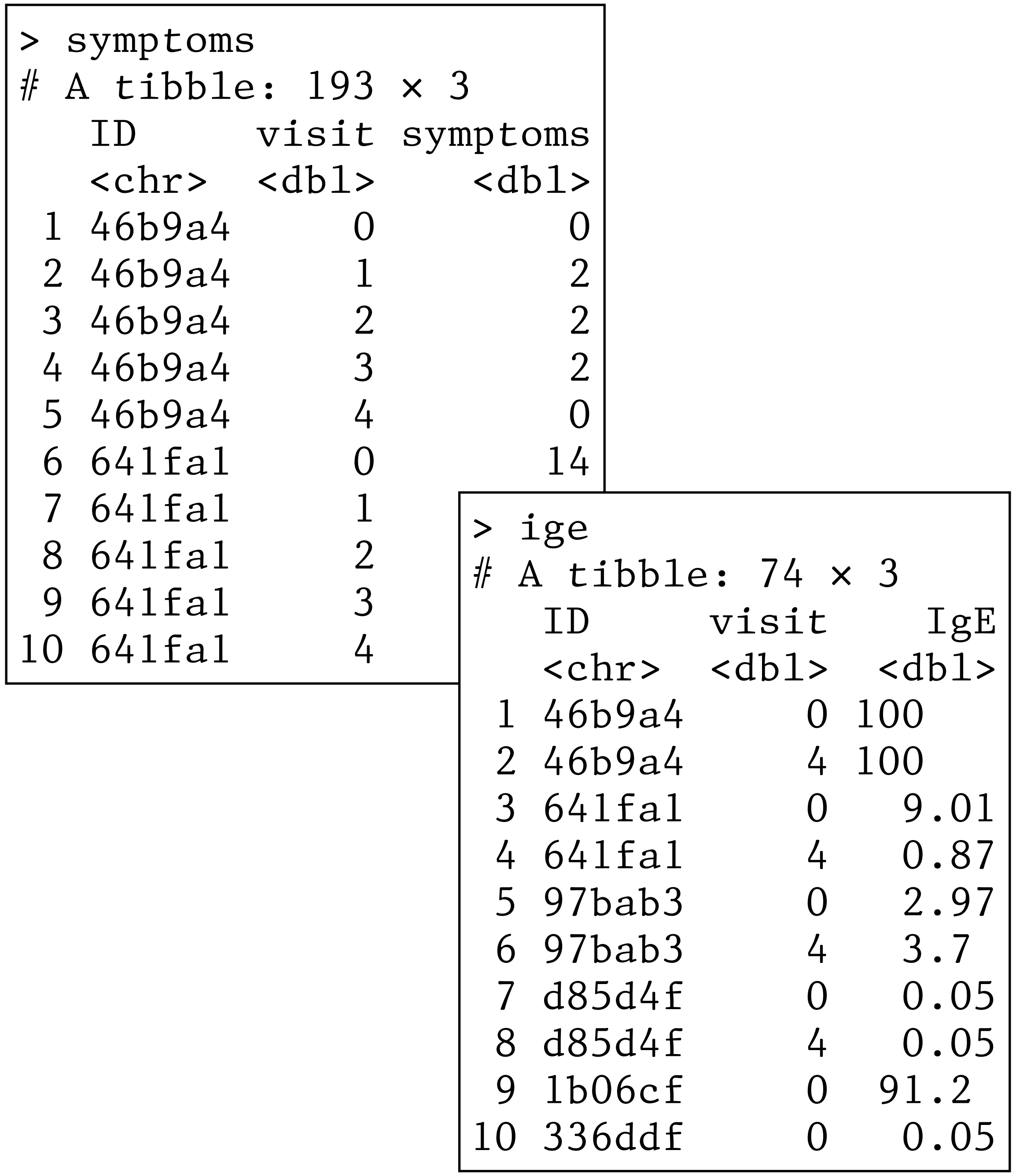
```
> symptoms
# A tibble: 193 × 3
   ID     visit symptoms
   <chr>  <dbl>    <dbl>
 1 46b9a4     0        0
 2 46b9a4     1        2
 3 46b9a4     2        2
 4 46b9a4     3        2
 5 46b9a4     4        0
 6 641fa1     0       14
 7 641fa1     1        3
 8 641fa1     2        2
 9 641fa1     3        2
10 641fa1     4        2
```

```
> ige
# A tibble: 74 × 3
   ID     visit    IgE
   <chr>  <dbl>  <dbl>
 1 46b9a4     0 100   
 2 46b9a4     4 100   
 3 641fa1     0   9.01
 4 641fa1     4   0.87
 5 97bab3     0   2.97
 6 97bab3     4   3.7 
 7 d85d4f     0   0.05
 8 d85d4f     4   0.05
 9 1b06cf     0  91.2 
10 336ddf     0   0.05
```

Must join by both "ID" and "visit"

Symptoms vs. IgE

```
symptoms |>
    left_join(ige, by = c("ID", "visit"))
```

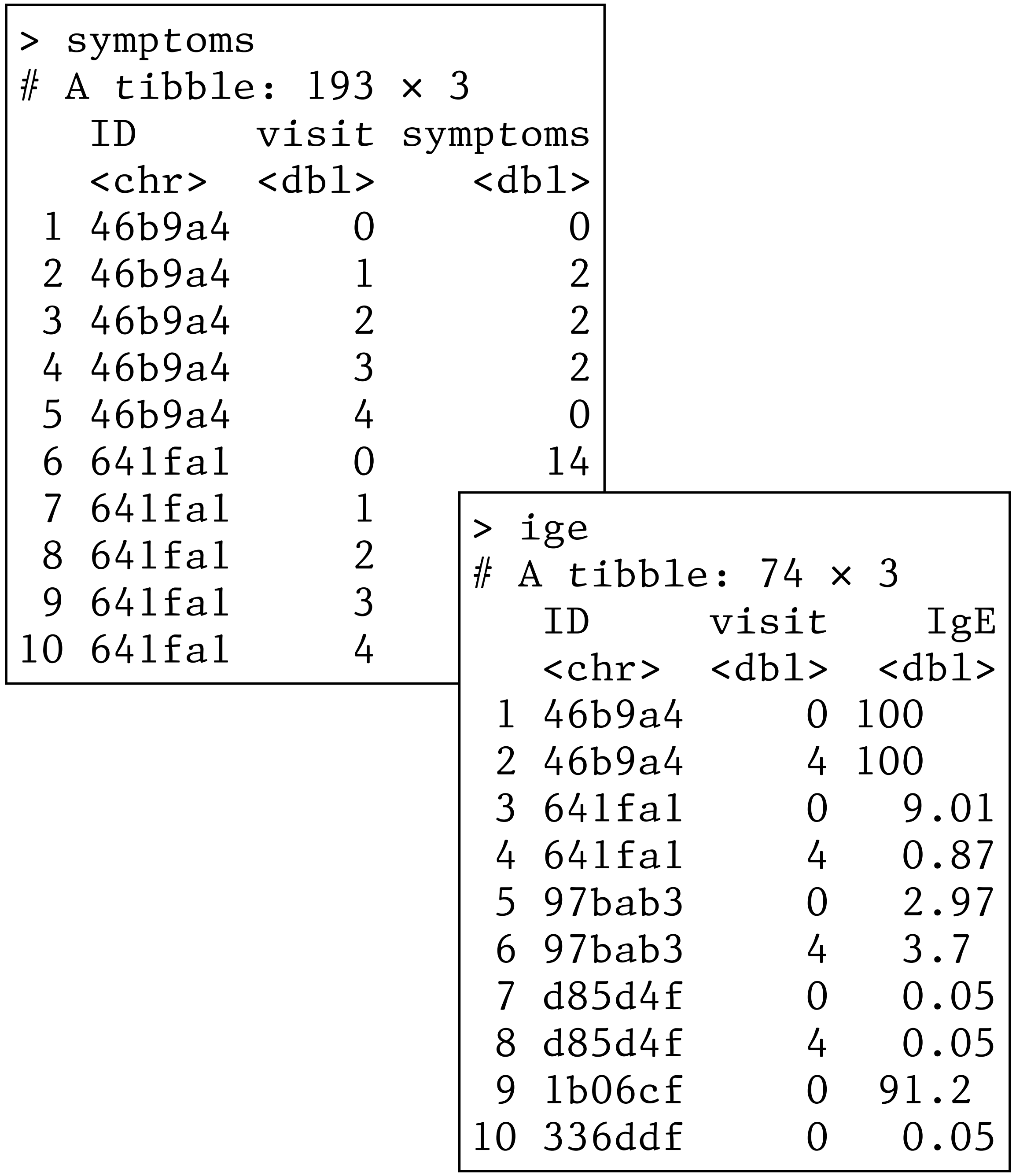


```
> symptoms
# A tibble: 193 × 3
   ID     visit symptoms
   <chr>  <dbl>    <dbl>
 1 46b9a4     0        0
 2 46b9a4     1        2
 3 46b9a4     2        2
 4 46b9a4     3        2
 5 46b9a4     4        0
 6 641fa1     0       14
 7 641fa1     1
 8 641fa1     2
 9 641fa1     3
10 641fa1     4
```

```
> ige
# A tibble: 74 × 3
   ID     visit    IgE
   <chr>  <dbl>  <dbl>
 1 46b9a4     0 100   
 2 46b9a4     4 100   
 3 641fa1     0   9.01
 4 641fa1     4   0.87
 5 97bab3     0   2.97
 6 97bab3     4   3.7 
 7 d85d4f     0   0.05
 8 d85d4f     4   0.05
 9 1b06cf     0  91.2 
10 336ddf     0   0.05
```

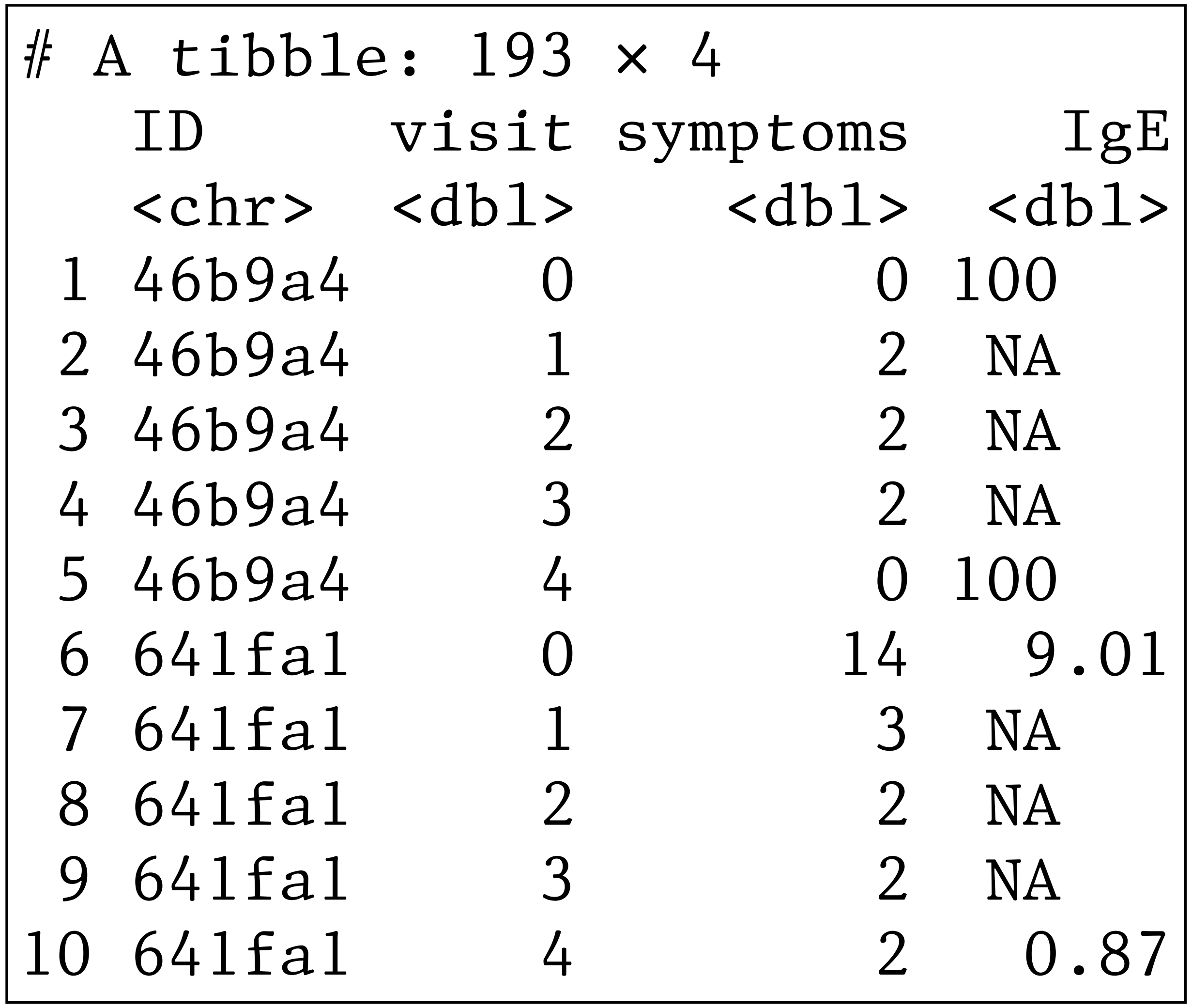
Symptoms vs. IgE

```
symptoms |>
    left_join(ige, by = c("ID", "visit"))
```



```
> symptoms
# A tibble: 193 × 3
   ID     visit symptoms
   <chr>  <dbl>    <dbl>
 1 46b9a4     0        0
 2 46b9a4     1        2
 3 46b9a4     2        2
 4 46b9a4     3        2
 5 46b9a4     4        0
 6 641fa1     0       14
 7 641fa1     1
 8 641fa1     2
 9 641fa1     3
10 641fa1     4
```

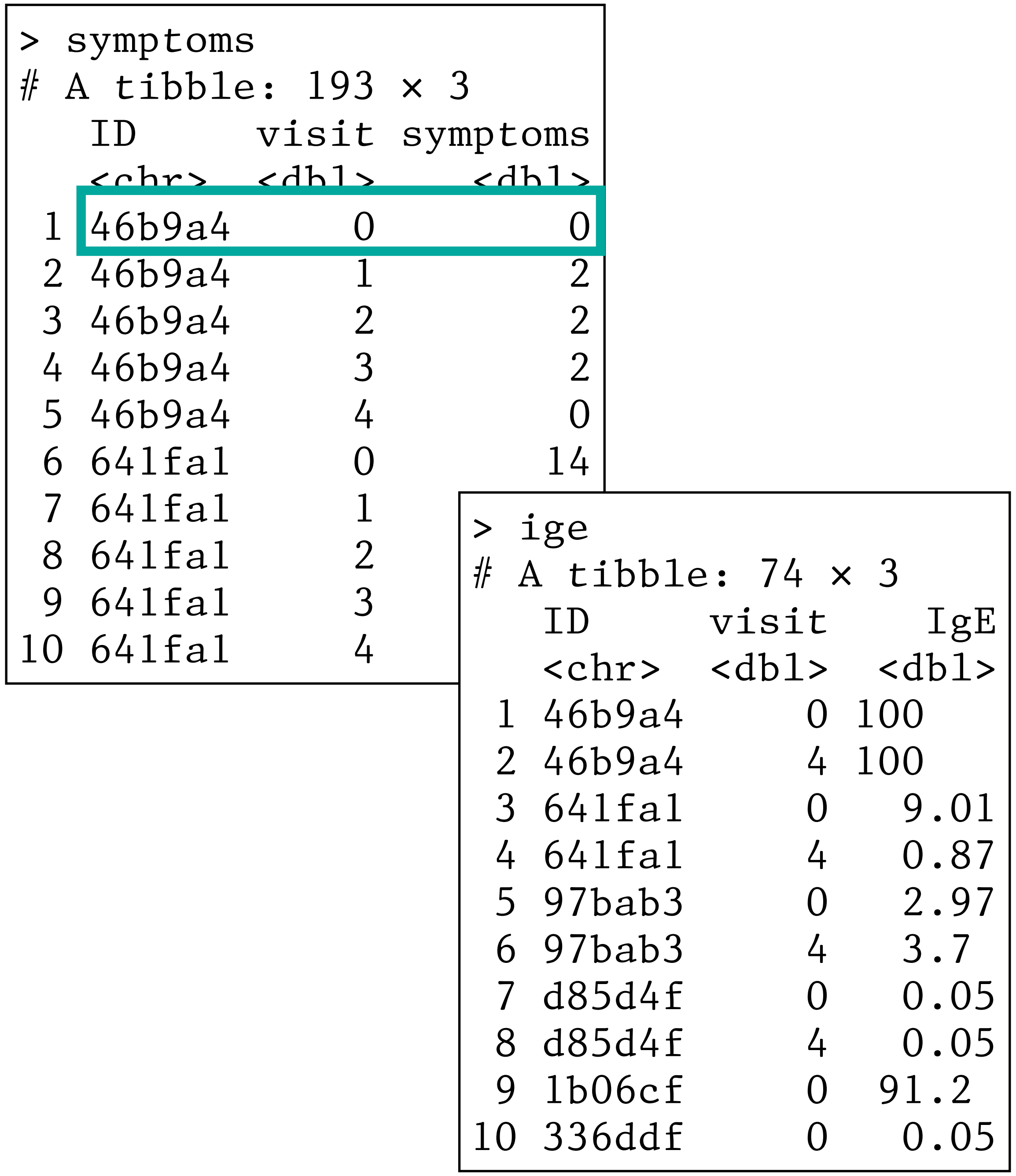
```
> ige
# A tibble: 74 × 3
   ID     visit    IgE
   <chr>  <dbl>  <dbl>
 1 46b9a4     0 100
 2 46b9a4     4 100
 3 641fa1     0   9.01
 4 641fa1     4   0.87
 5 97bab3     0   2.97
 6 97bab3     4   3.7
 7 d85d4f     0   0.05
 8 d85d4f     4   0.05
 9 1b06cf     0  91.2
10 336ddf     0   0.05
```



```
# A tibble: 193 × 4
   ID     visit symptoms    IgE
   <chr>  <dbl>    <dbl>  <dbl>
 1 46b9a4     0        0 100
 2 46b9a4     1        2  NA
 3 46b9a4     2        2  NA
 4 46b9a4     3        2  NA
 5 46b9a4     4        0 100
 6 641fa1     0       14   9.01
 7 641fa1     1        3  NA
 8 641fa1     2        2  NA
 9 641fa1     3        2  NA
10 641fa1     4        2   0.87
```

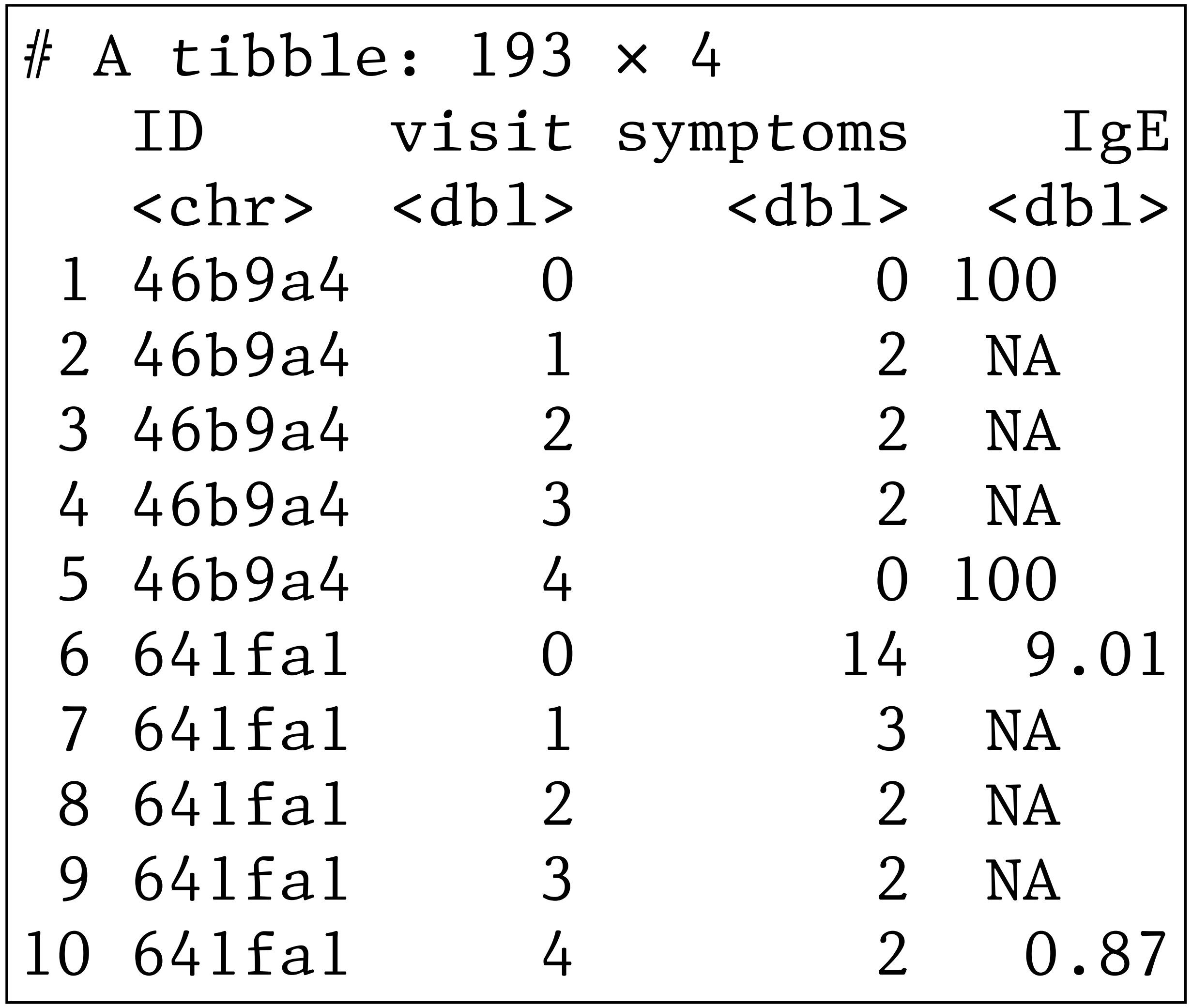
Symptoms vs. IgE

```
symptoms |>
    left_join(ige, by = c("ID", "visit"))
```



```
> symptoms
# A tibble: 193 × 3
   ID      visit symptoms
   <chr>   <dbl>    <dbl>
 1 46b9a4      0        0
 2 46b9a4      1        2
 3 46b9a4      2        2
 4 46b9a4      3        2
 5 46b9a4      4        0
 6 641fa1      0       14
 7 641fa1      1
 8 641fa1      2
 9 641fa1      3
10 641fa1      4
```

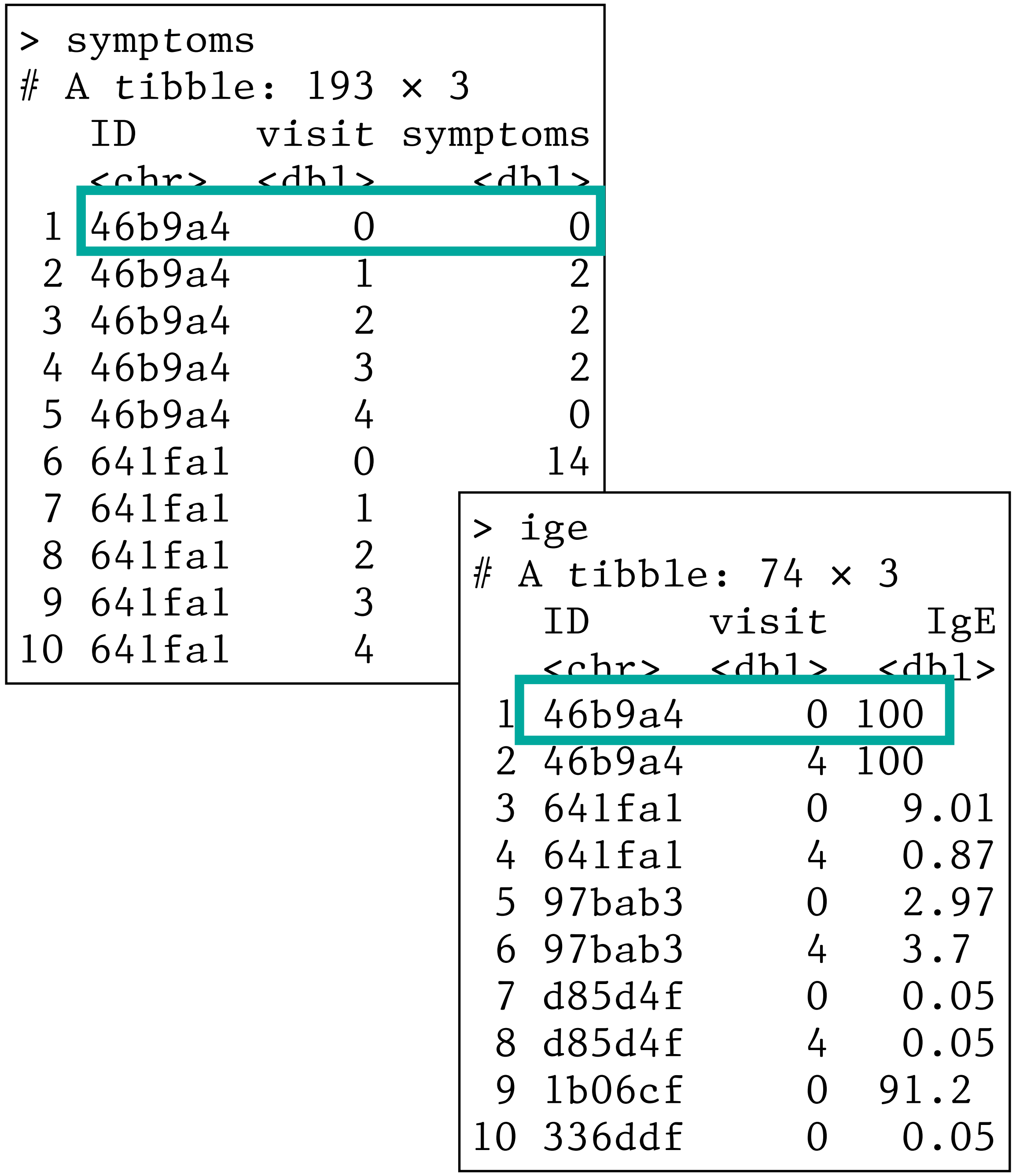
```
> ige
# A tibble: 74 × 3
   ID      visit    IgE
   <chr>   <dbl>  <dbl>
 1 46b9a4      0 100
 2 46b9a4      4 100
 3 641fa1      0   9.01
 4 641fa1      4   0.87
 5 97bab3      0   2.97
 6 97bab3      4   3.7
 7 d85d4f      0   0.05
 8 d85d4f      4   0.05
 9 1b06cf      0  91.2
10 336ddf      0   0.05
```



```
# A tibble: 193 × 4
   ID      visit symptoms    IgE
   <chr>   <dbl>    <dbl>  <dbl>
 1 46b9a4      0        0 100
 2 46b9a4      1        2  NA
 3 46b9a4      2        2  NA
 4 46b9a4      3        2  NA
 5 46b9a4      4        0 100
 6 641fa1      0       14   9.01
 7 641fa1      1        3  NA
 8 641fa1      2        2  NA
 9 641fa1      3        2  NA
10 641fa1      4        2   0.87
```

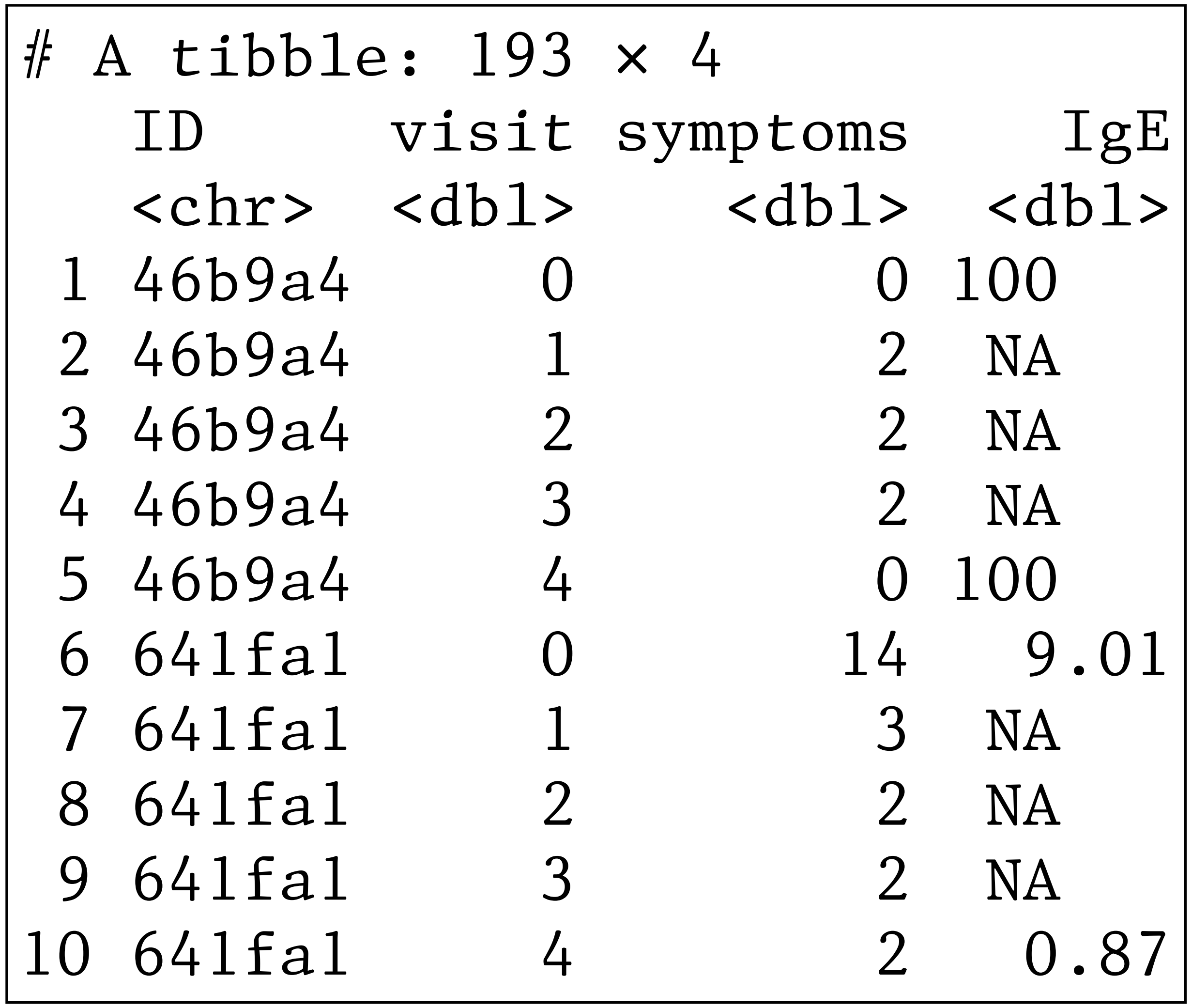
Symptoms vs. IgE

```
symptoms |>
    left_join(ige, by = c("ID", "visit"))
```



```
> symptoms
# A tibble: 193 × 3
   ID      visit symptoms
   <chr>   <dbl>    <dbl>
 1 46b9a4      0        0
 2 46b9a4      1        2
 3 46b9a4      2        2
 4 46b9a4      3        2
 5 46b9a4      4        0
 6 641fa1      0       14
 7 641fa1      1
 8 641fa1      2
 9 641fa1      3
10 641fa1      4
```

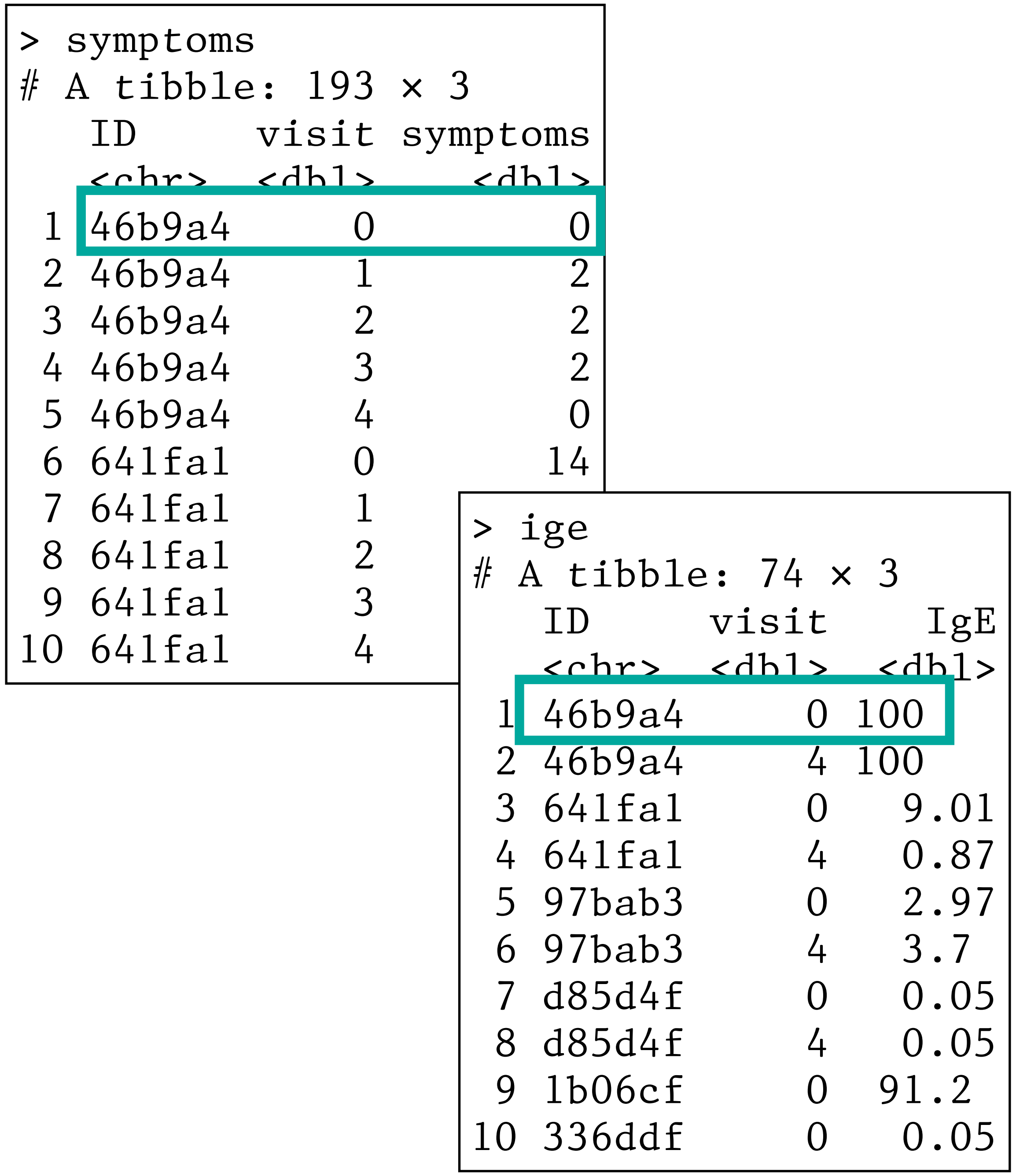
```
> ige
# A tibble: 74 × 3
   ID      visit    IgE
   <chr>   <dbl>  <dbl>
 1 46b9a4      0 100
 2 46b9a4      4 100
 3 641fa1      0   9.01
 4 641fa1      4   0.87
 5 97bab3      0   2.97
 6 97bab3      4   3.7
 7 d85d4f      0   0.05
 8 d85d4f      4   0.05
 9 1b06cf      0  91.2
10 336ddf      0   0.05
```



```
# A tibble: 193 × 4
   ID      visit symptoms    IgE
   <chr>   <dbl>    <dbl>  <dbl>
 1 46b9a4      0        0 100
 2 46b9a4      1        2  NA
 3 46b9a4      2        2  NA
 4 46b9a4      3        2  NA
 5 46b9a4      4        0 100
 6 641fa1      0       14   9.01
 7 641fa1      1        3  NA
 8 641fa1      2        2  NA
 9 641fa1      3        2  NA
10 641fa1      4        2   0.87
```

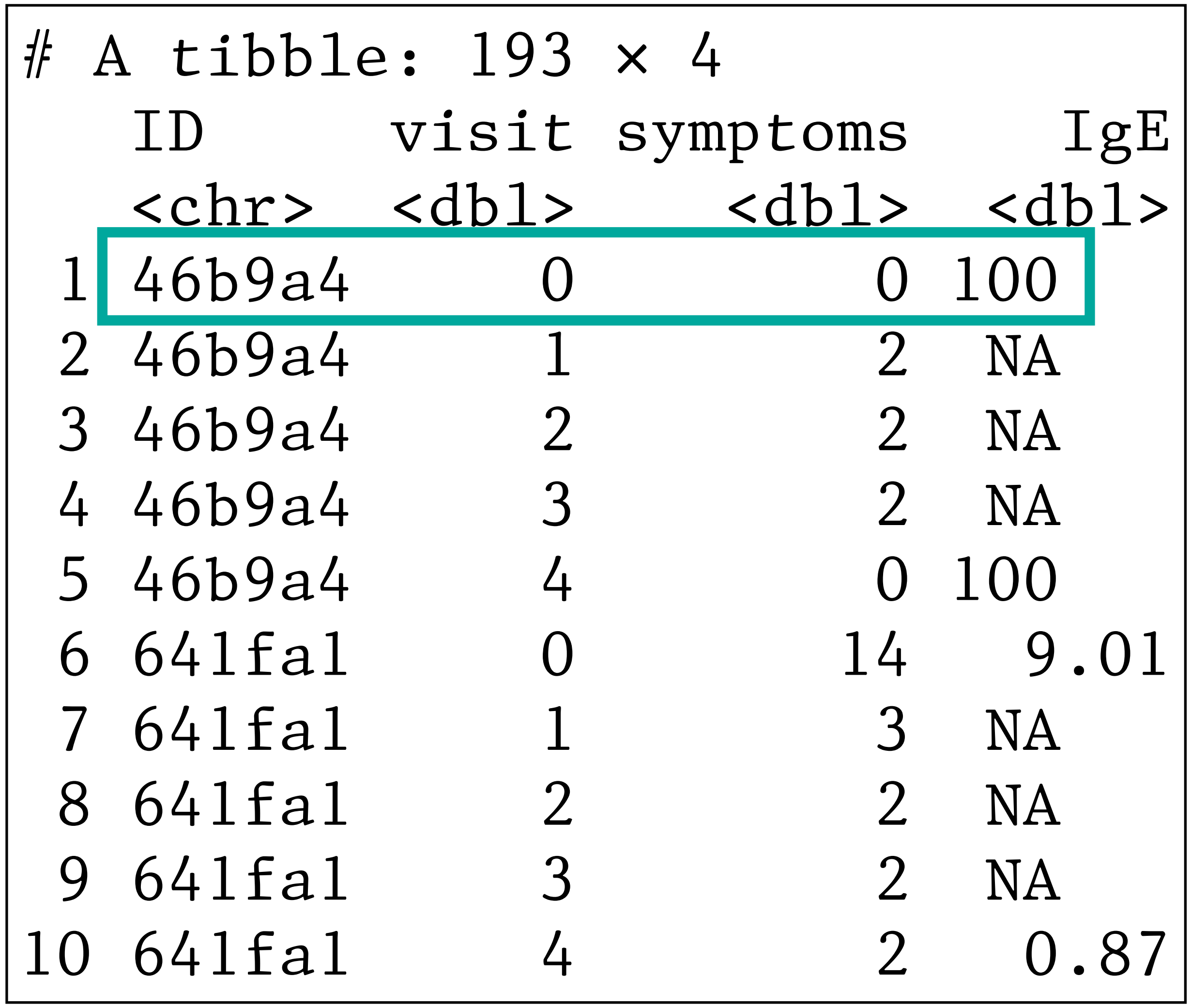
Symptoms vs. IgE

```
symptoms |>
    left_join(ige, by = c("ID", "visit"))
```



```
> symptoms
# A tibble: 193 × 3
   ID     visit symptoms
   <chr>  <dbl>    <dbl>
 1 46b9a4     0        0
 2 46b9a4     1        2
 3 46b9a4     2        2
 4 46b9a4     3        2
 5 46b9a4     4        0
 6 641fa1     0       14
 7 641fa1     1
 8 641fa1     2
 9 641fa1     3
10 641fa1     4
```

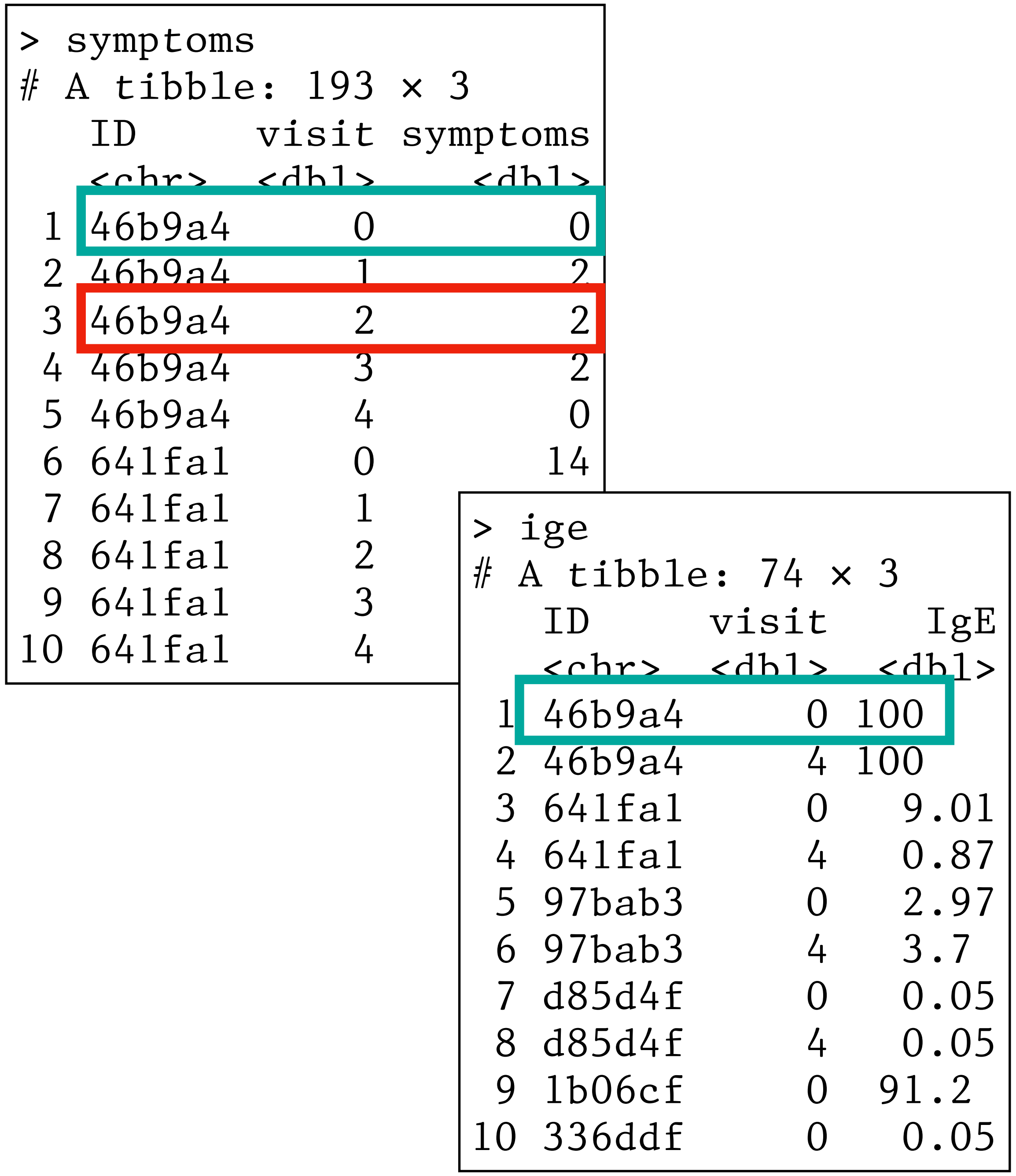
```
> ige
# A tibble: 74 × 3
   ID     visit   IgE
   <chr>  <dbl> <dbl>
 1 46b9a4     0 100
 2 46b9a4     4 100
 3 641fa1     0   9.01
 4 641fa1     4   0.87
 5 97bab3     0   2.97
 6 97bab3     4   3.7
 7 d85d4f     0   0.05
 8 d85d4f     4   0.05
 9 1b06cf     0  91.2
10 336ddf     0   0.05
```



```
# A tibble: 193 × 4
   ID     visit symptoms    IgE
   <chr>  <dbl>    <dbl>  <dbl>
 1 46b9a4     0        0 100
 2 46b9a4     1        2  NA
 3 46b9a4     2        2  NA
 4 46b9a4     3        2  NA
 5 46b9a4     4        0 100
 6 641fa1     0       14   9.01
 7 641fa1     1        3  NA
 8 641fa1     2        2  NA
 9 641fa1     3        2  NA
10 641fa1     4        2   0.87
```

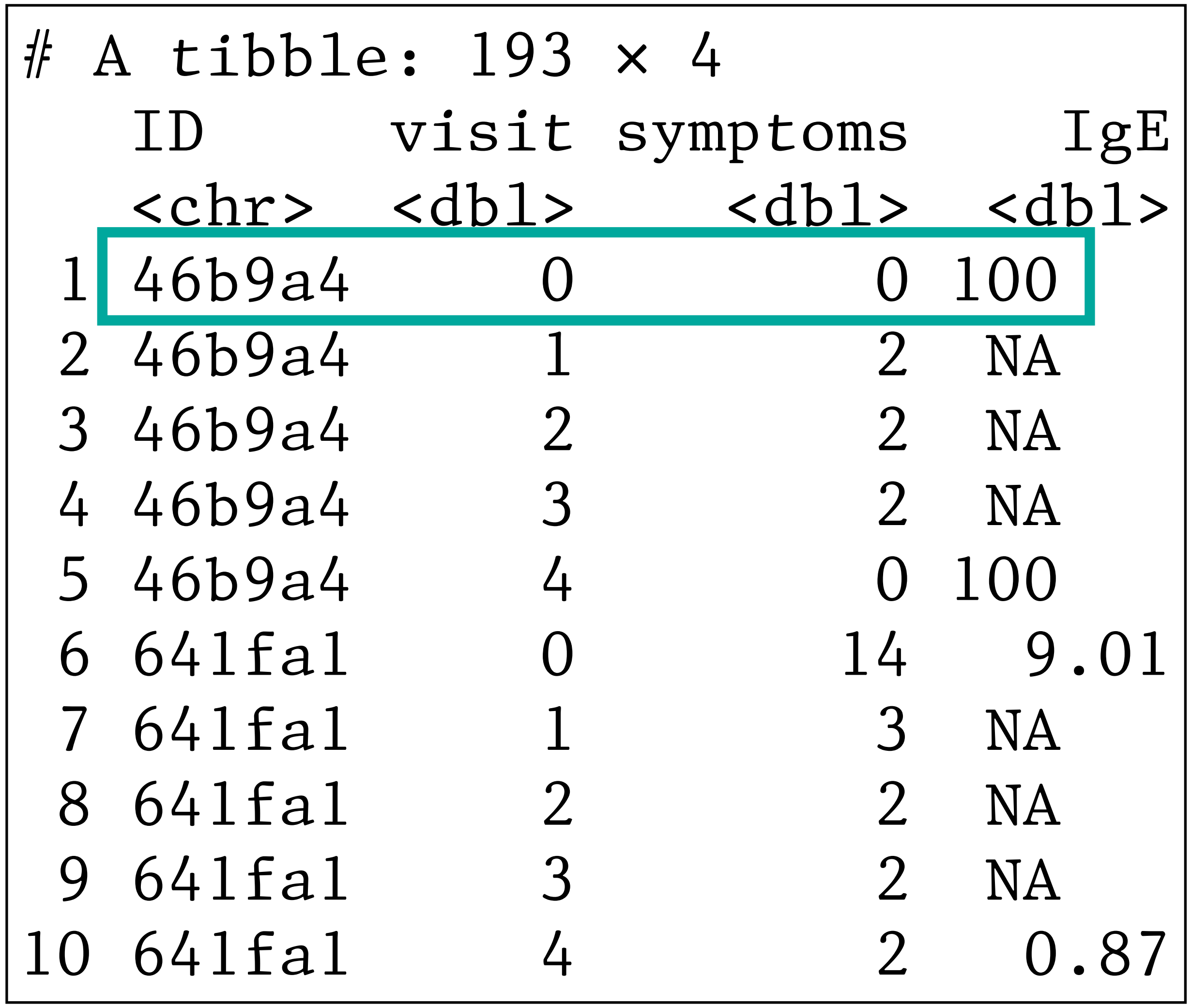
Symptoms vs. IgE

```
symptoms |>
    left_join(ige, by = c("ID", "visit"))
```



```
> symptoms
# A tibble: 193 × 3
   ID     visit symptoms
   <chr>  <dbl>    <dbl>
 1 46b9a4     0        0
 2 46b9a4     1        2
 3 46b9a4     2        2
 4 46b9a4     3        2
 5 46b9a4     4        0
 6 641fa1     0       14
 7 641fa1     1
 8 641fa1     2
 9 641fa1     3
10 641fa1     4
```

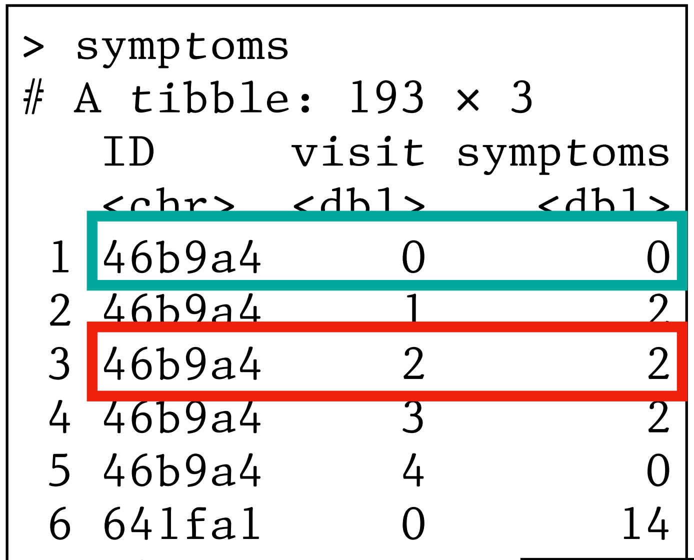
```
> ige
# A tibble: 74 × 3
   ID     visit    IgE
   <chr>  <dbl>  <dbl>
 1 46b9a4     0 100   
 2 46b9a4     4 100   
 3 641fa1     0   9.01
 4 641fa1     4   0.87
 5 97bab3     0   2.97
 6 97bab3     4   3.7 
 7 d85d4f     0   0.05
 8 d85d4f     4   0.05
 9 1b06cf     0  91.2 
10 336ddf     0   0.05
```



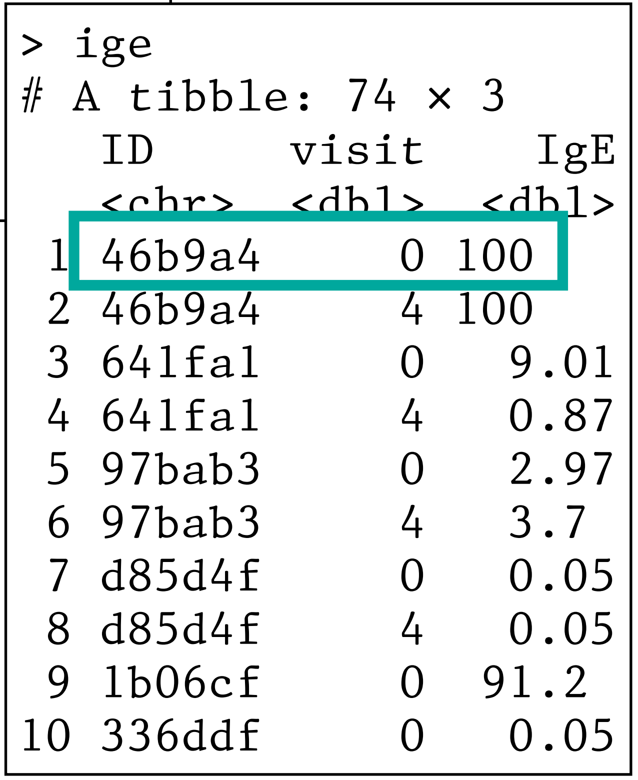
```
# A tibble: 193 × 4
   ID     visit symptoms    IgE
   <chr>  <dbl>    <dbl>  <dbl>
 1 46b9a4     0        0 100   
 2 46b9a4     1        2  NA   
 3 46b9a4     2        2  NA   
 4 46b9a4     3        2  NA   
 5 46b9a4     4        0 100   
 6 641fa1     0       14   9.01
 7 641fa1     1        3  NA   
 8 641fa1     2        2  NA   
 9 641fa1     3        2  NA   
10 641fa1     4        2   0.87
```

Symptoms vs. IgE

```
symptoms |>
    left_join(ige, by = c("ID", "visit"))
```

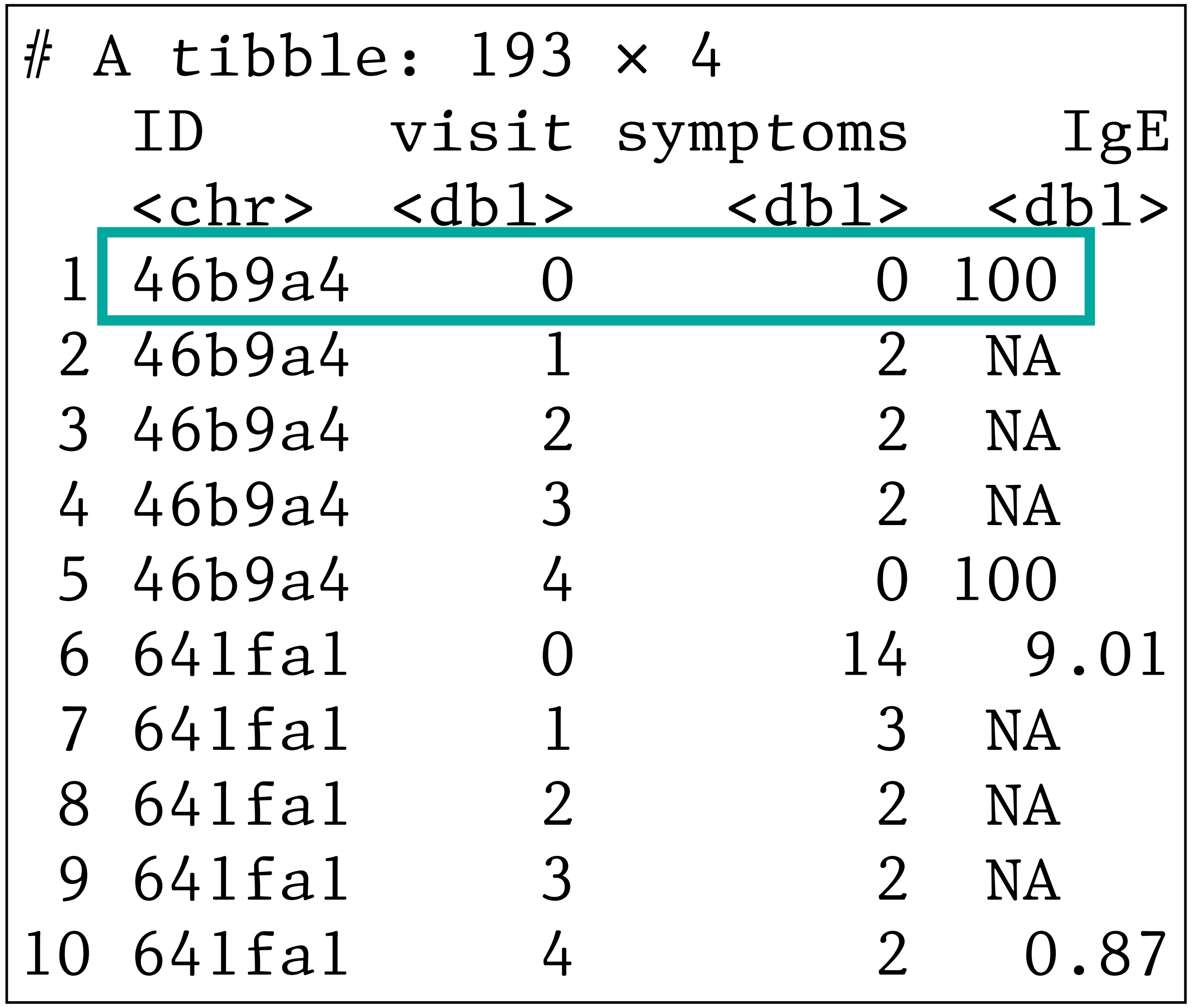


```
> symptoms
# A tibble: 193 × 3
   ID      visit symptoms
   <chr>   <dbl>    <dbl>
 1 46b9a4      0        0
 2 46b9a4      1        2
 3 46b9a4      2        2
 4 46b9a4      3        2
 5 46b9a4      4        0
 6 641fal      0       14
 7 641fal      1
 8 641fal      2
 9 641fal      3
10 641fal      4
```



```
> ige
# A tibble: 74 × 3
   ID      visit    IgE
   <chr>   <dbl>  <dbl>
 1 46b9a4      0 100
 2 46b9a4      4 100
 3 641fal      0   9.01
 4 641fal      4   0.87
 5 97bab3      0   2.97
 6 97bab3      4   3.7
 7 d85d4f      0   0.05
 8 d85d4f      4   0.05
 9 1b06cf      0  91.2
10 336ddf      0   0.05
```

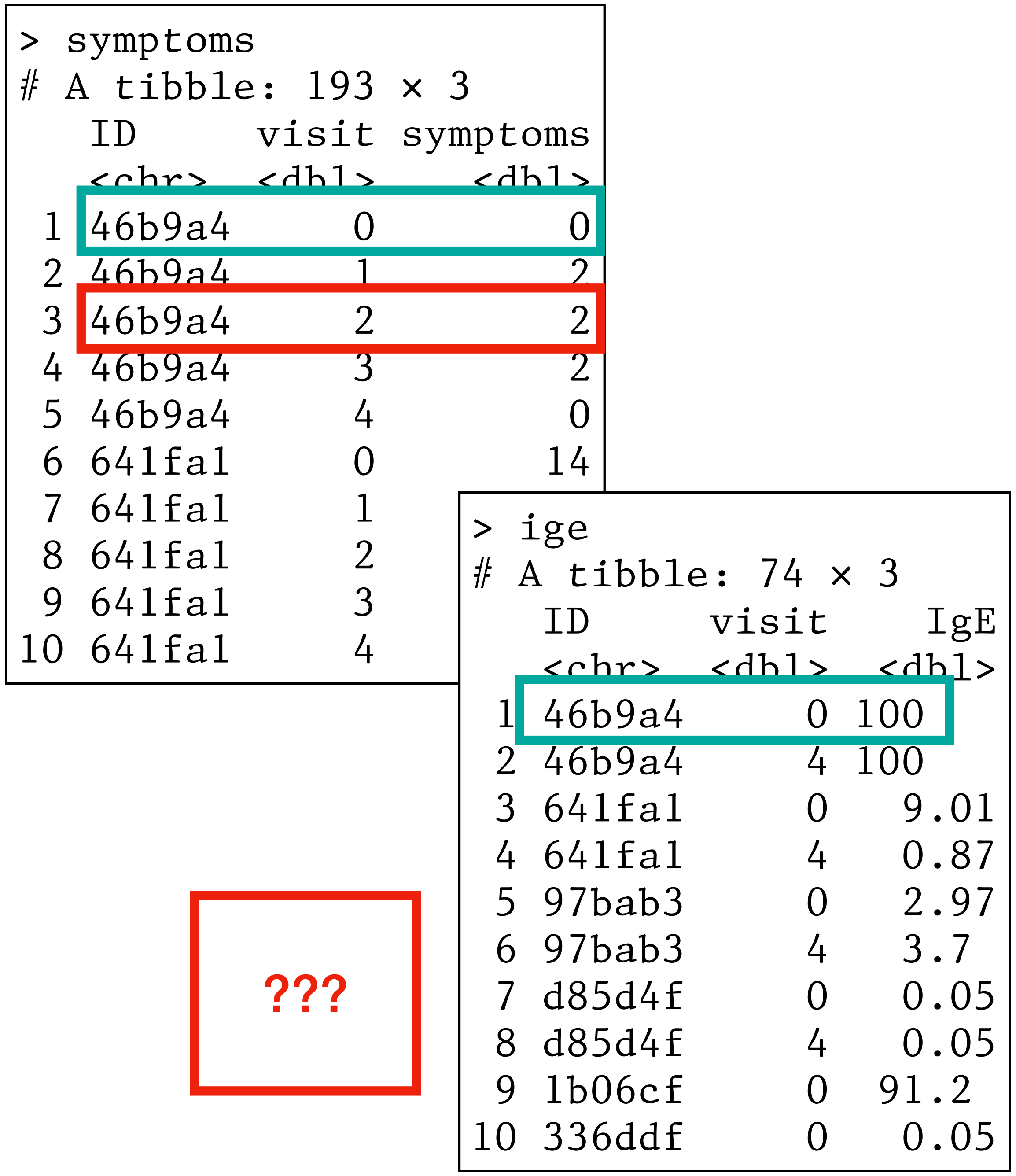
???



```
# A tibble: 193 × 4
   ID      visit symptoms    IgE
   <chr>   <dbl>    <dbl>  <dbl>
 1 46b9a4      0        0 100
 2 46b9a4      1        2  NA
 3 46b9a4      2        2  NA
 4 46b9a4      3        2  NA
 5 46b9a4      4        0 100
 6 641fal      0       14   9.01
 7 641fal      1        3  NA
 8 641fal      2        2  NA
 9 641fal      3        2  NA
10 641fal      4        2   0.87
```

Symptoms vs. IgE

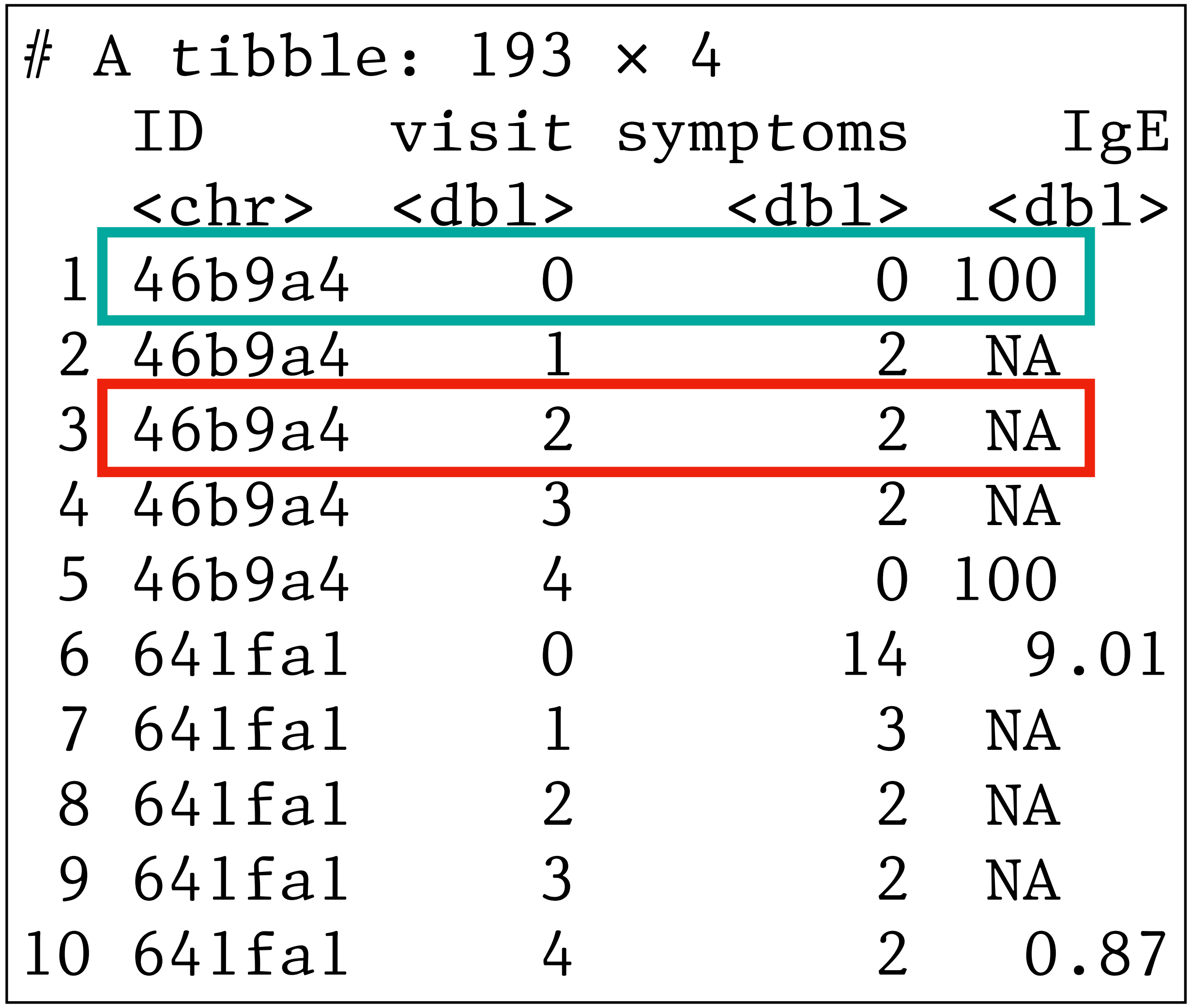
```
symptoms |>
    left_join(ige, by = c("ID", "visit"))
```



```
> symptoms
# A tibble: 193 × 3
   ID      visit symptoms
   <chr>   <dbl>    <dbl>
 1 46b9a4      0        0
 2 46b9a4      1        2
 3 46b9a4      2        2
 4 46b9a4      3        2
 5 46b9a4      4        0
 6 641fa1      0       14
 7 641fa1      1
 8 641fa1      2
 9 641fa1      3
10 641fa1      4
```

```
> ige
# A tibble: 74 × 3
   ID      visit    IgE
   <chr>   <dbl>  <dbl>
 1 46b9a4      0  100
 2 46b9a4      4  100
 3 641fa1      0    9.01
 4 641fa1      4    0.87
 5 97bab3      0    2.97
 6 97bab3      4    3.7
 7 d85d4f      0    0.05
 8 d85d4f      4    0.05
 9 1b06cf      0   91.2
10 336ddf      0    0.05
```

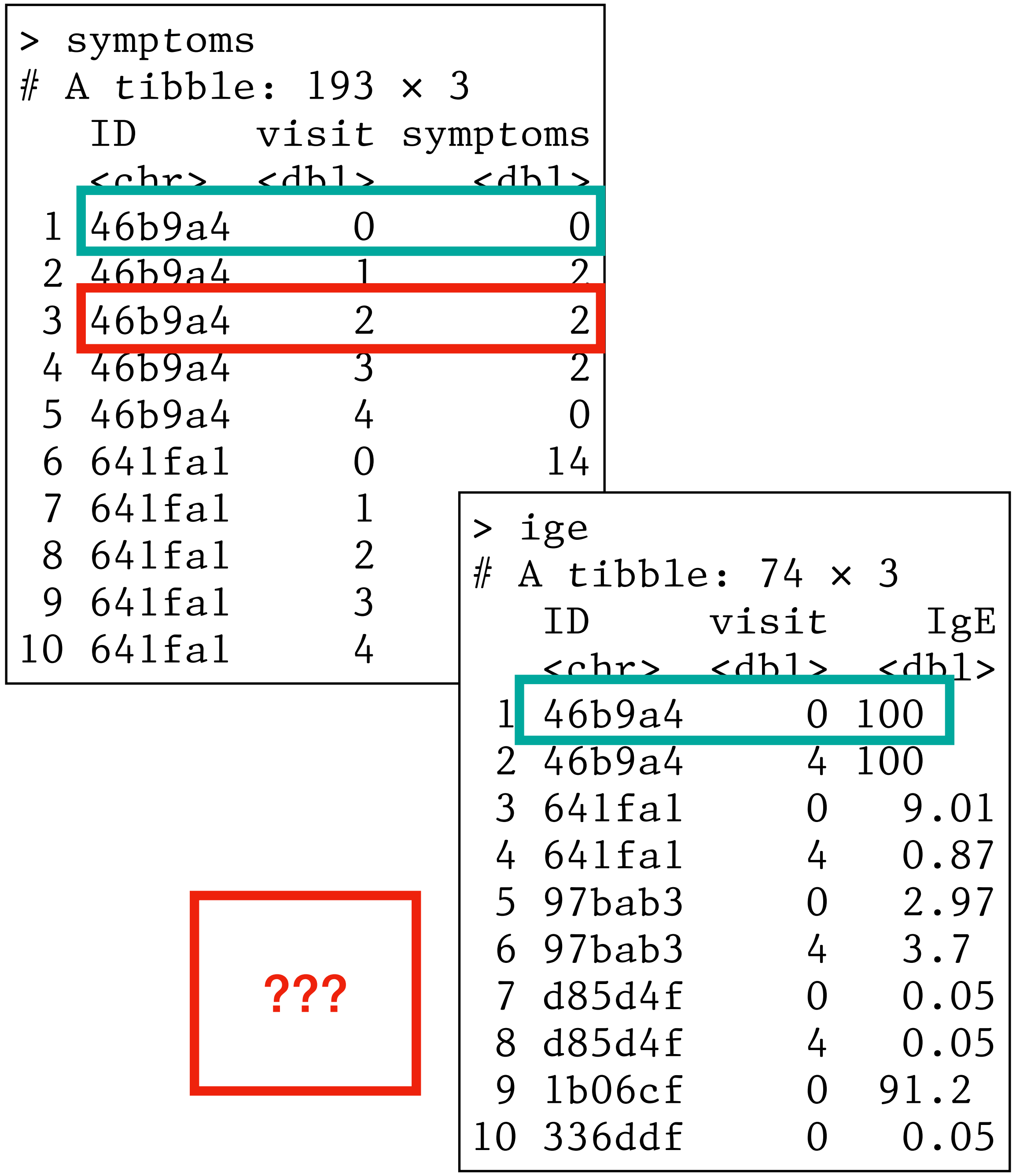
???



```
# A tibble: 193 × 4
   ID      visit symptoms    IgE
   <chr>   <dbl>    <dbl>  <dbl>
 1 46b9a4      0        0 100
 2 46b9a4      1        2  NA
 3 46b9a4      2        2  NA
 4 46b9a4      3        2  NA
 5 46b9a4      4        0 100
 6 641fa1      0       14   9.01
 7 641fa1      1        3  NA
 8 641fa1      2        2  NA
 9 641fa1      3        2  NA
10 641fa1      4        2   0.87
```

Symptoms vs. IgE

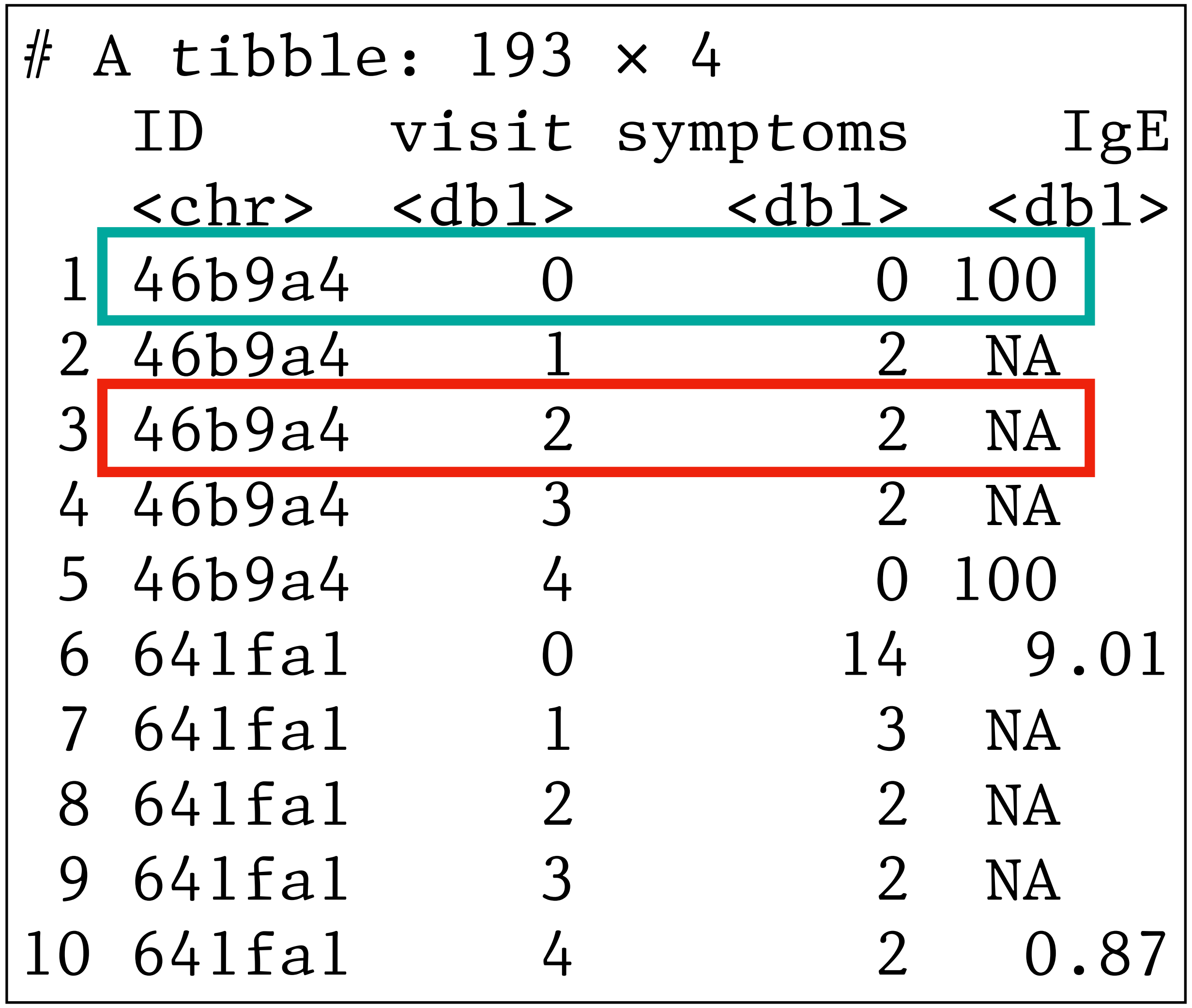
```
symptoms |>
    left_join(ige, by = c("ID", "visit"))
```



```
> symptoms
# A tibble: 193 × 3
   ID      visit symptoms
   <chr>   <dbl>    <dbl>
 1 46b9a4      0        0
 2 46b9a4      1        2
 3 46b9a4      2        2
 4 46b9a4      3        2
 5 46b9a4      4        0
 6 641fa1      0       14
 7 641fa1      1
 8 641fa1      2
 9 641fa1      3
10 641fa1      4
```

```
> ige
# A tibble: 74 × 3
   ID      visit    IgE
   <chr>   <dbl>  <dbl>
 1 46b9a4      0 100
 2 46b9a4      4 100
 3 641fa1      0   9.01
 4 641fa1      4   0.87
 5 97bab3      0   2.97
 6 97bab3      4   3.7
 7 d85d4f      0   0.05
 8 d85d4f      4   0.05
 9 1b06cf      0  91.2
10 336ddf      0   0.05
```

???

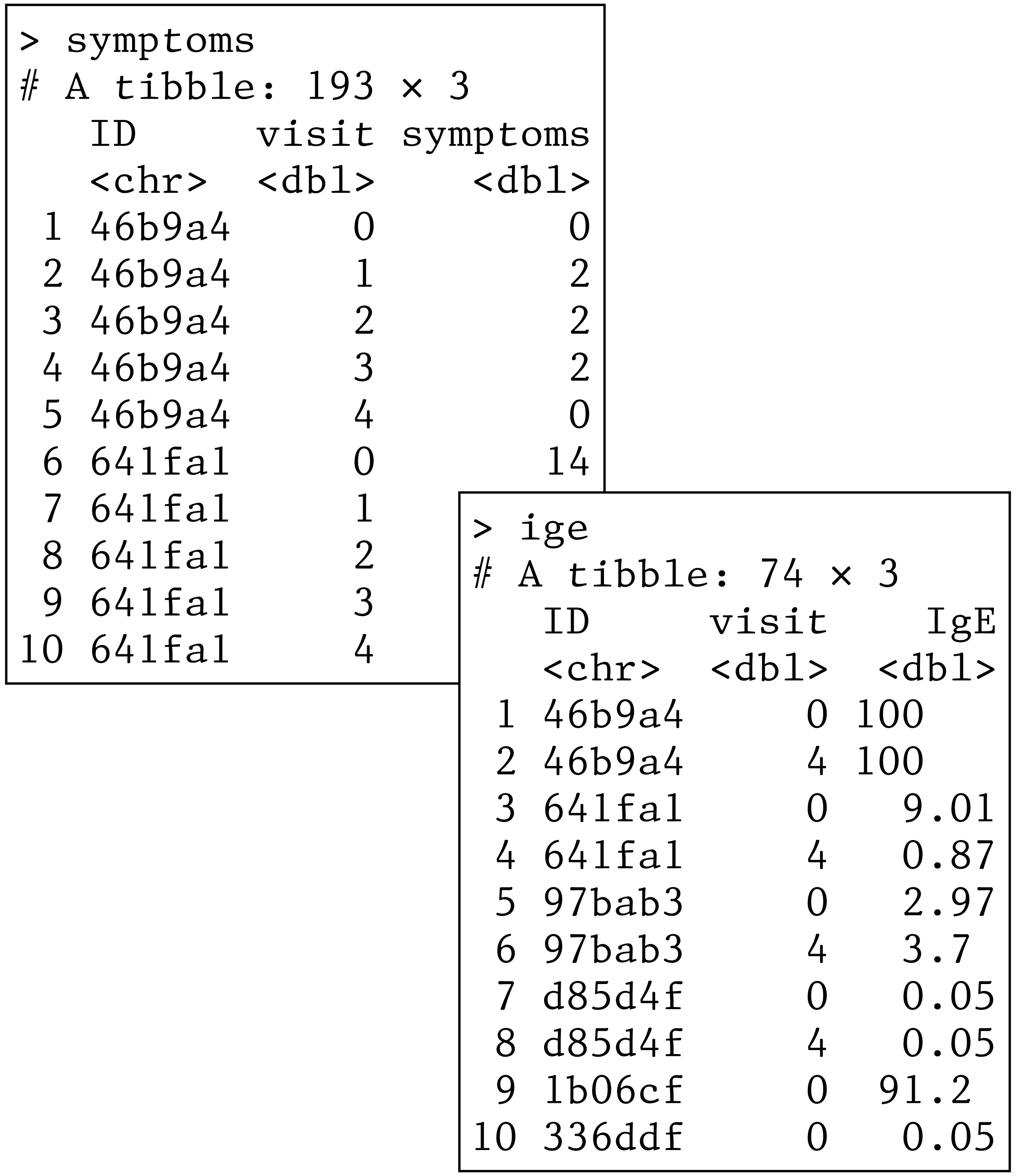


```
# A tibble: 193 × 4
   ID      visit symptoms    IgE
   <chr>   <dbl>    <dbl>  <dbl>
 1 46b9a4      0        0 100
 2 46b9a4      1        2  NA
 3 46b9a4      2        2  NA
 4 46b9a4      3        2  NA
 5 46b9a4      4        0 100
 6 641fa1      0       14   9.01
 7 641fa1      1        3  NA
 8 641fa1      2        2  NA
 9 641fa1      3        2  NA
10 641fa1      4        2   0.87
```

Left join fills in NA values when there is no match in the 'ige' table

Symptoms vs. IgE

```
symptoms |>
    inner_join(ige, by = c("ID", "visit"))
```



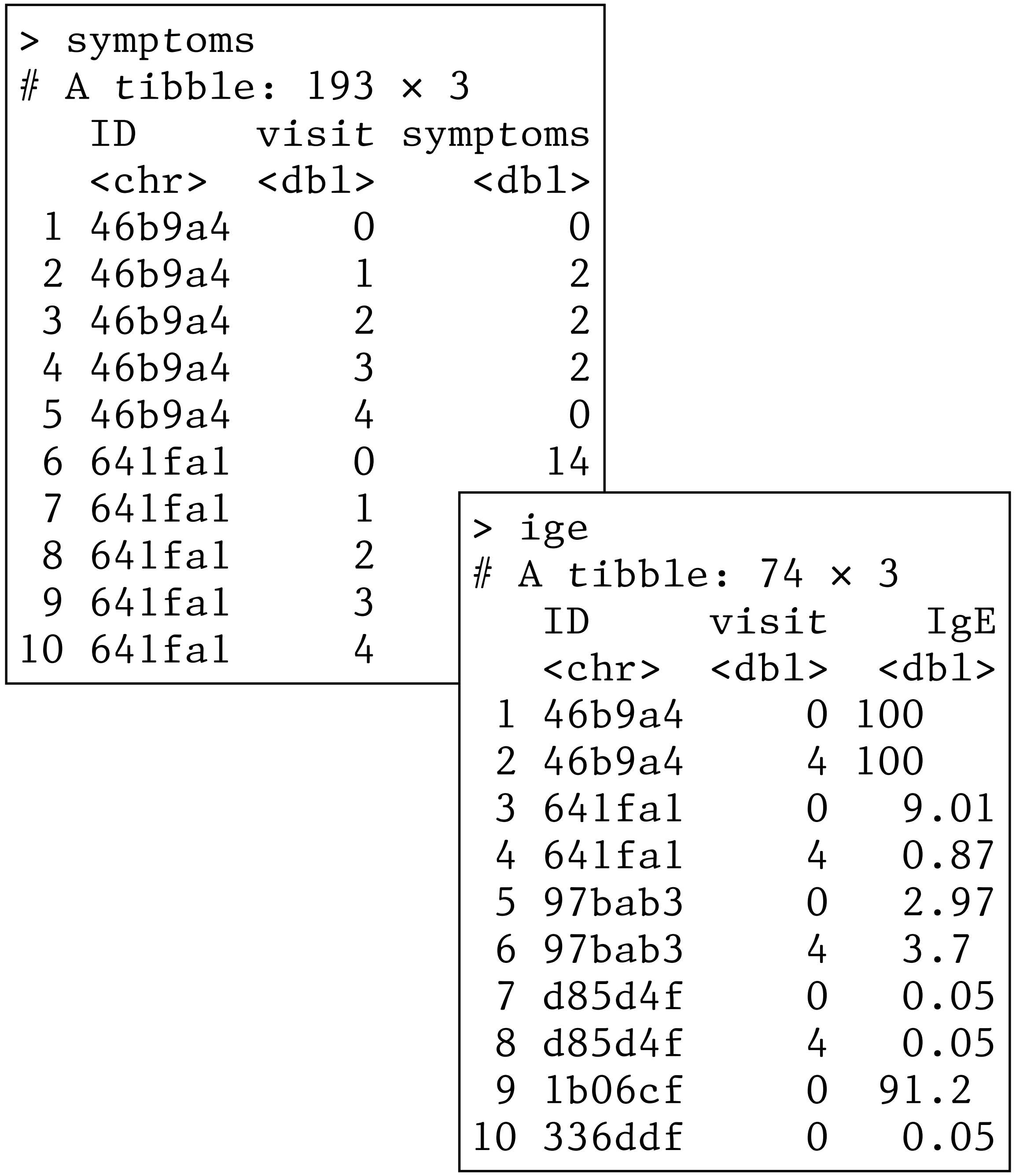
```
> symptoms
# A tibble: 193 × 3
   ID     visit symptoms
   <chr>  <dbl>    <dbl>
 1 46b9a4     0        0
 2 46b9a4     1        2
 3 46b9a4     2        2
 4 46b9a4     3        2
 5 46b9a4     4        0
 6 641fa1     0       14
 7 641fa1     1
 8 641fa1     2
 9 641fa1     3
10 641fa1     4
```

```
> ige
# A tibble: 74 × 3
   ID     visit    IgE
   <chr>  <dbl>  <dbl>
 1 46b9a4     0 100
 2 46b9a4     4 100
 3 641fa1     0   9.01
 4 641fa1     4   0.87
 5 97bab3     0   2.97
 6 97bab3     4   3.7
 7 d85d4f     0   0.05
 8 d85d4f     4   0.05
 9 1b06cf     0  91.2
10 336ddf     0   0.05
```

inner join only keeps rows where there is a match in BOTH tables

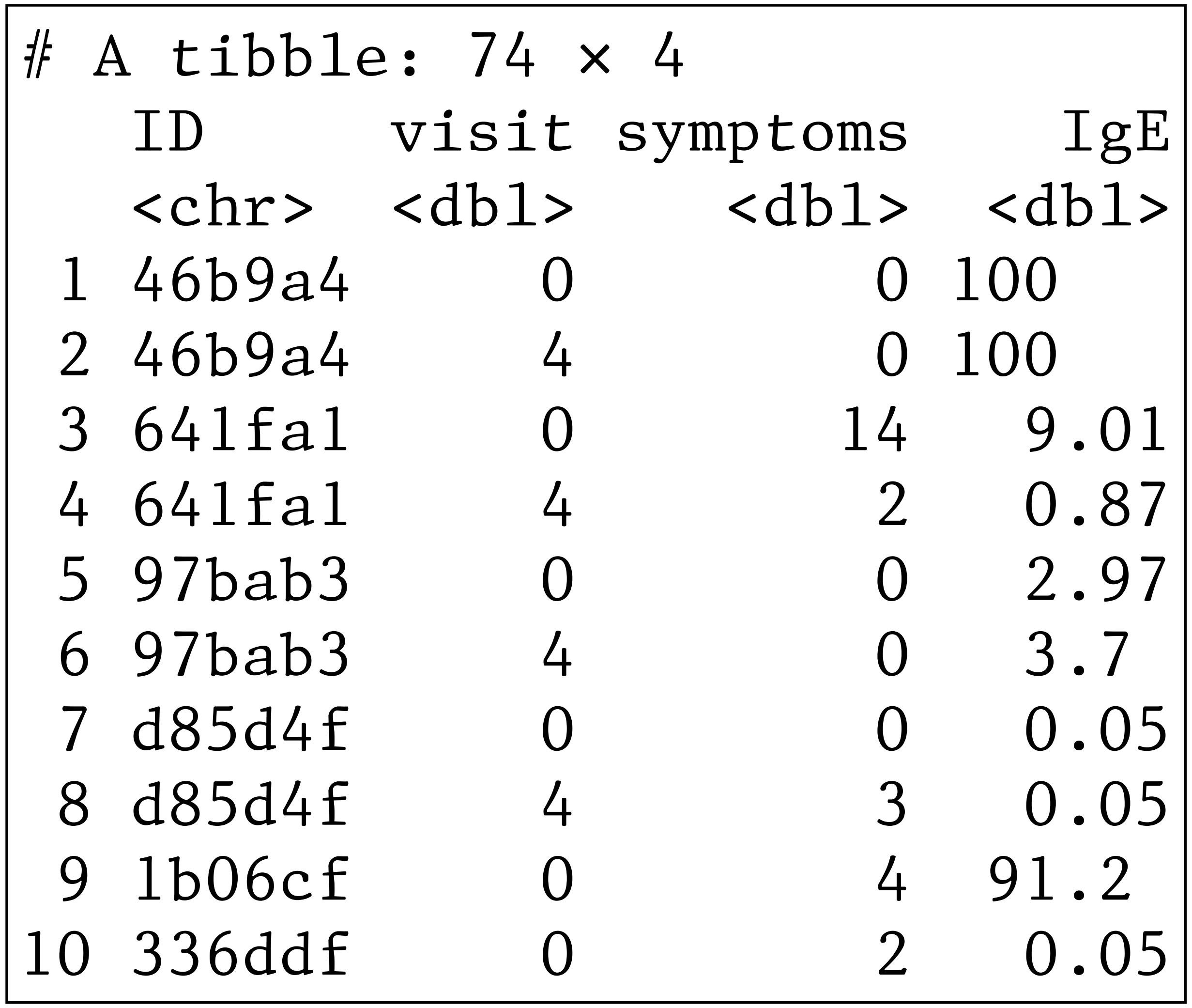
Symptoms vs. IgE

```
symptoms |>
    inner_join(ige, by = c("ID", "visit"))
```



```
> symptoms
# A tibble: 193 × 3
   ID      visit symptoms
   <chr>   <dbl>    <dbl>
 1 46b9a4      0        0
 2 46b9a4      1        2
 3 46b9a4      2        2
 4 46b9a4      3        2
 5 46b9a4      4        0
 6 641fa1      0       14
 7 641fa1      1
 8 641fa1      2
 9 641fa1      3
10 641fa1      4
```

```
> ige
# A tibble: 74 × 3
   ID      visit    IgE
   <chr>   <dbl>  <dbl>
 1 46b9a4      0 100
 2 46b9a4      4 100
 3 641fa1      0   9.01
 4 641fa1      4   0.87
 5 97bab3      0   2.97
 6 97bab3      4   3.7
 7 d85d4f      0   0.05
 8 d85d4f      4   0.05
 9 1b06cf      0  91.2
10 336ddf      0   0.05
```

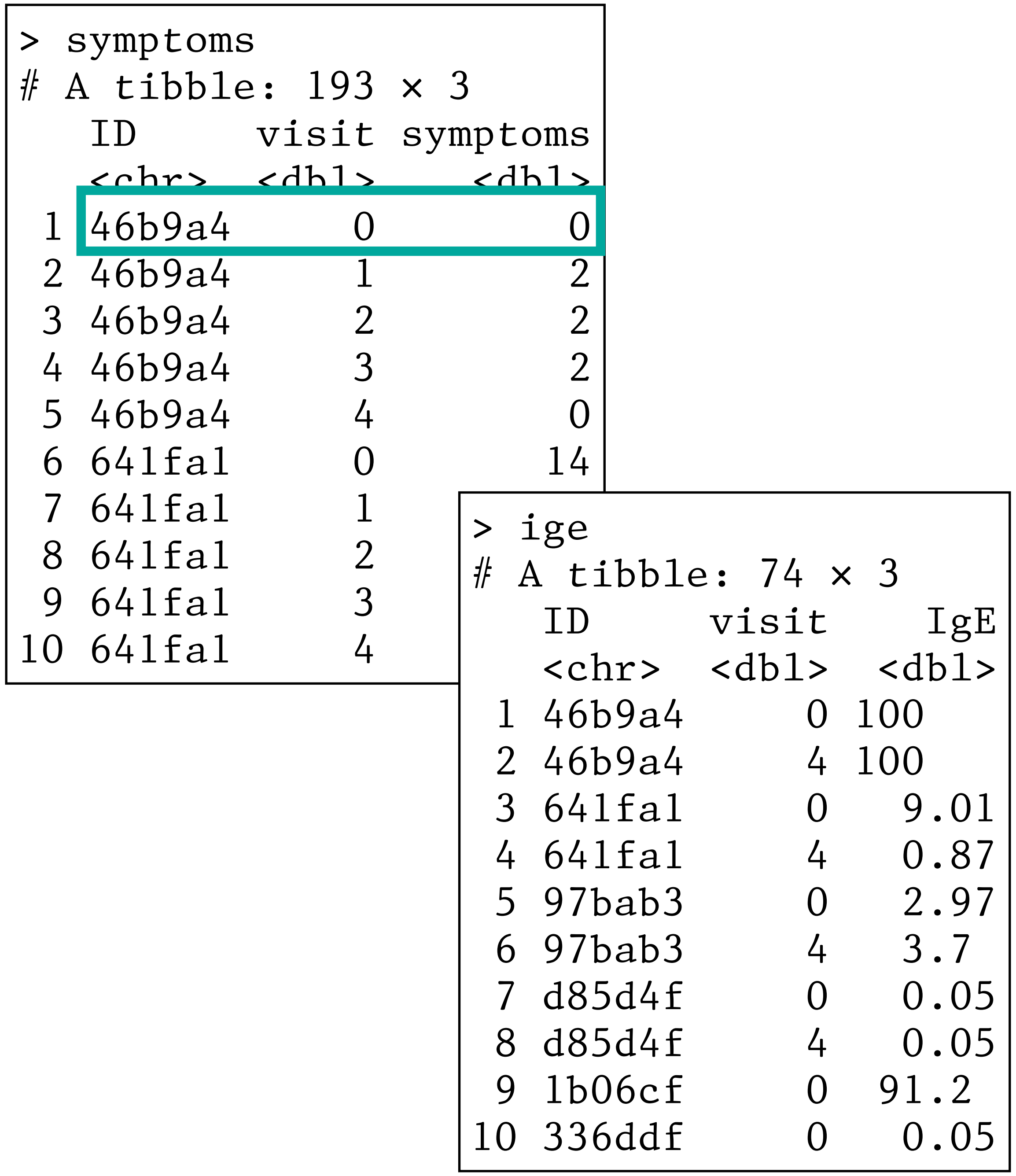


```
# A tibble: 74 × 4
   ID      visit symptoms    IgE
   <chr>   <dbl>    <dbl>  <dbl>
 1 46b9a4      0        0 100
 2 46b9a4      4        0 100
 3 641fa1      0       14   9.01
 4 641fa1      4        2   0.87
 5 97bab3      0        0   2.97
 6 97bab3      4        0   3.7
 7 d85d4f      0        0   0.05
 8 d85d4f      4        3   0.05
 9 1b06cf      0        4  91.2
10 336ddf      0        2   0.05
```

inner join only keeps rows where there is a match in BOTH tables

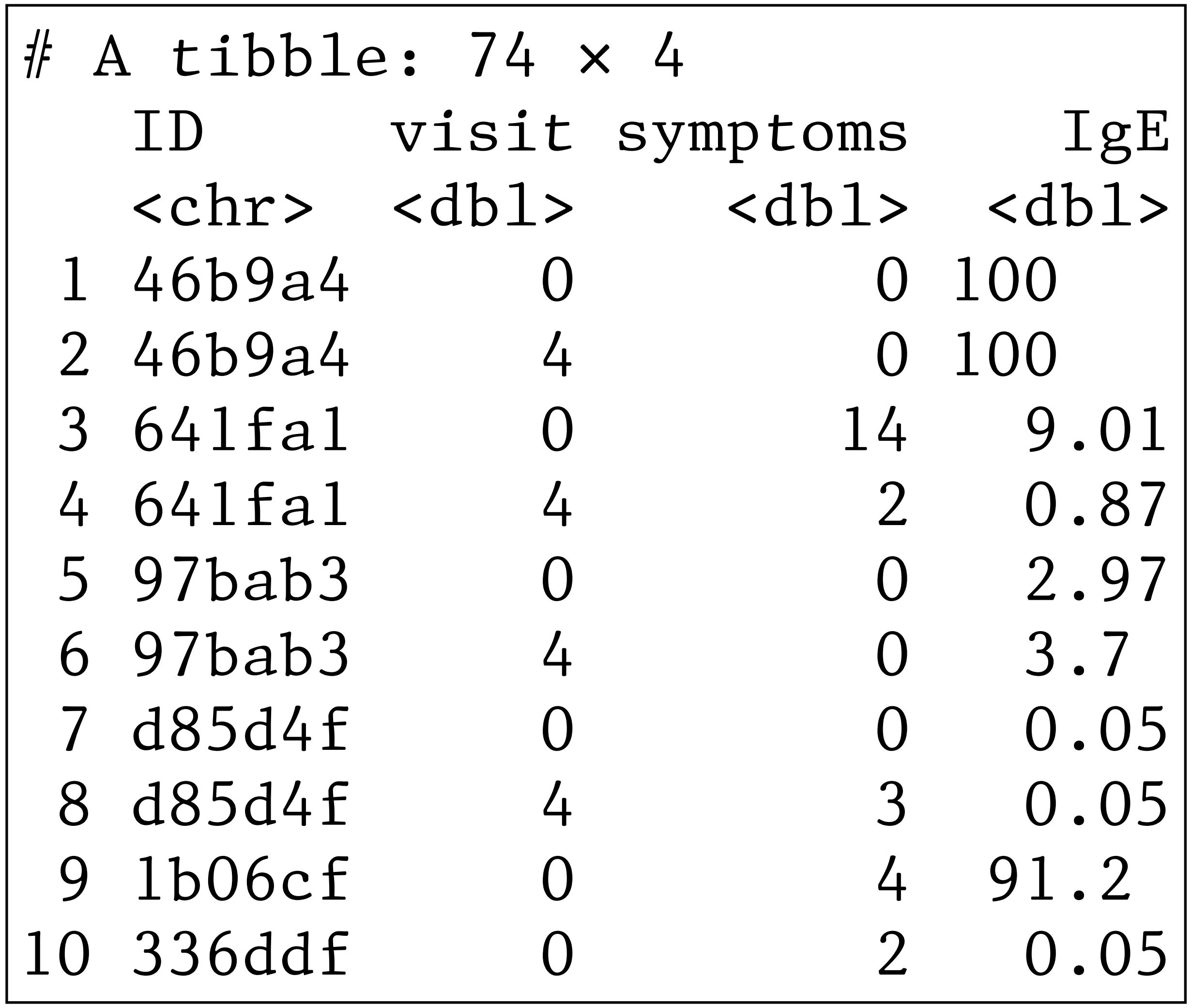
Symptoms vs. IgE

```
symptoms |>
    inner_join(ige, by = c("ID", "visit"))
```



```
> symptoms
# A tibble: 193 × 3
   ID     visit symptoms
   <chr>  <dbl>    <dbl>
 1 46b9a4     0        0
 2 46b9a4     1        2
 3 46b9a4     2        2
 4 46b9a4     3        2
 5 46b9a4     4        0
 6 641fa1     0       14
 7 641fa1     1
 8 641fa1     2
 9 641fa1     3
10 641fa1     4
```

```
> ige
# A tibble: 74 × 3
   ID     visit   IgE
   <chr>  <dbl> <dbl>
 1 46b9a4     0 100
 2 46b9a4     4 100
 3 641fa1     0   9.01
 4 641fa1     4   0.87
 5 97bab3     0   2.97
 6 97bab3     4   3.7
 7 d85d4f     0   0.05
 8 d85d4f     4   0.05
 9 1b06cf     0  91.2
10 336ddf     0   0.05
```

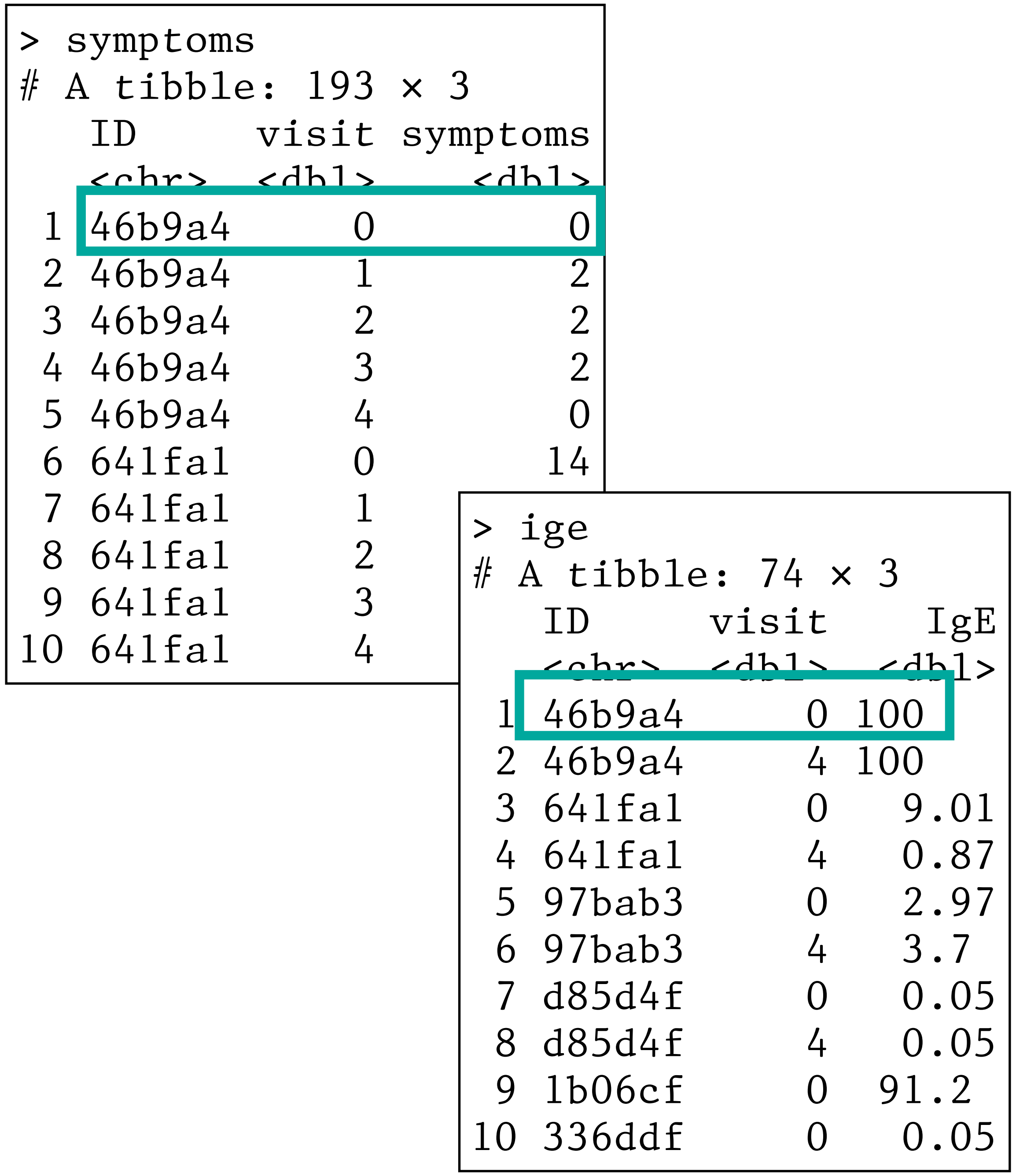


```
# A tibble: 74 × 4
   ID     visit symptoms   IgE
   <chr>  <dbl>    <dbl> <dbl>
 1 46b9a4     0        0 100
 2 46b9a4     4        0 100
 3 641fa1     0       14   9.01
 4 641fa1     4        2   0.87
 5 97bab3     0        0   2.97
 6 97bab3     4        0   3.7
 7 d85d4f     0        0   0.05
 8 d85d4f     4        3   0.05
 9 1b06cf     0        4  91.2
10 336ddf     0        2   0.05
```

inner join only keeps rows where there is a match in BOTH tables

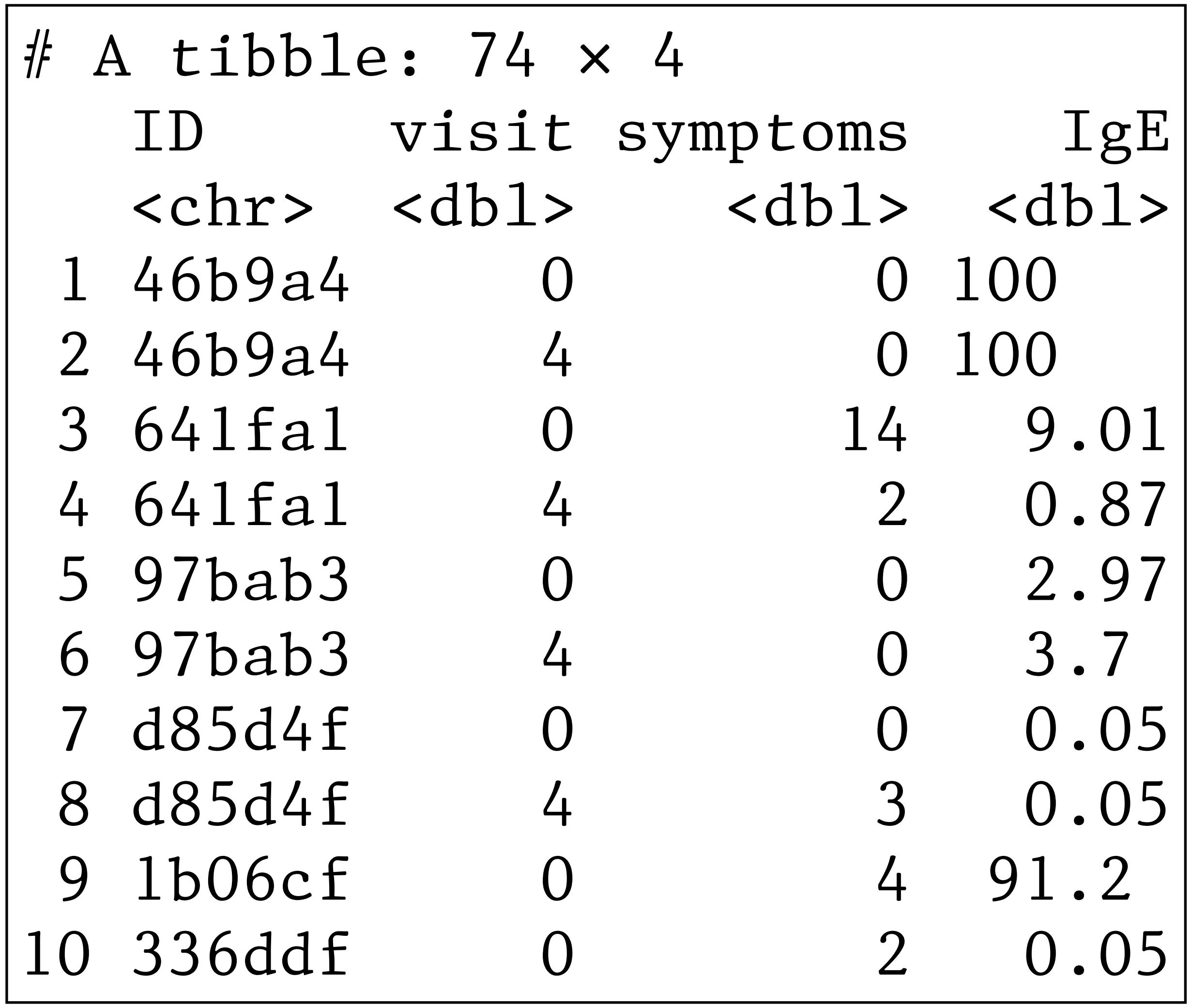
Symptoms vs. IgE

```
symptoms |>
    inner_join(ige, by = c("ID", "visit"))
```



```
> symptoms
# A tibble: 193 × 3
   ID      visit symptoms
   <chr>   <dbl>    <dbl>
 1 46b9a4      0        0
 2 46b9a4      1        2
 3 46b9a4      2        2
 4 46b9a4      3        2
 5 46b9a4      4        0
 6 641fa1      0       14
 7 641fa1      1
 8 641fa1      2
 9 641fa1      3
10 641fa1      4
```

```
> ige
# A tibble: 74 × 3
   ID      visit    IgE
   <chr>   <dbl>  <dbl>
 1 46b9a4      0 100
 2 46b9a4      4 100
 3 641fa1      0   9.01
 4 641fa1      4   0.87
 5 97bab3      0   2.97
 6 97bab3      4   3.7
 7 d85d4f      0   0.05
 8 d85d4f      4   0.05
 9 1b06cf      0  91.2
10 336ddf      0   0.05
```

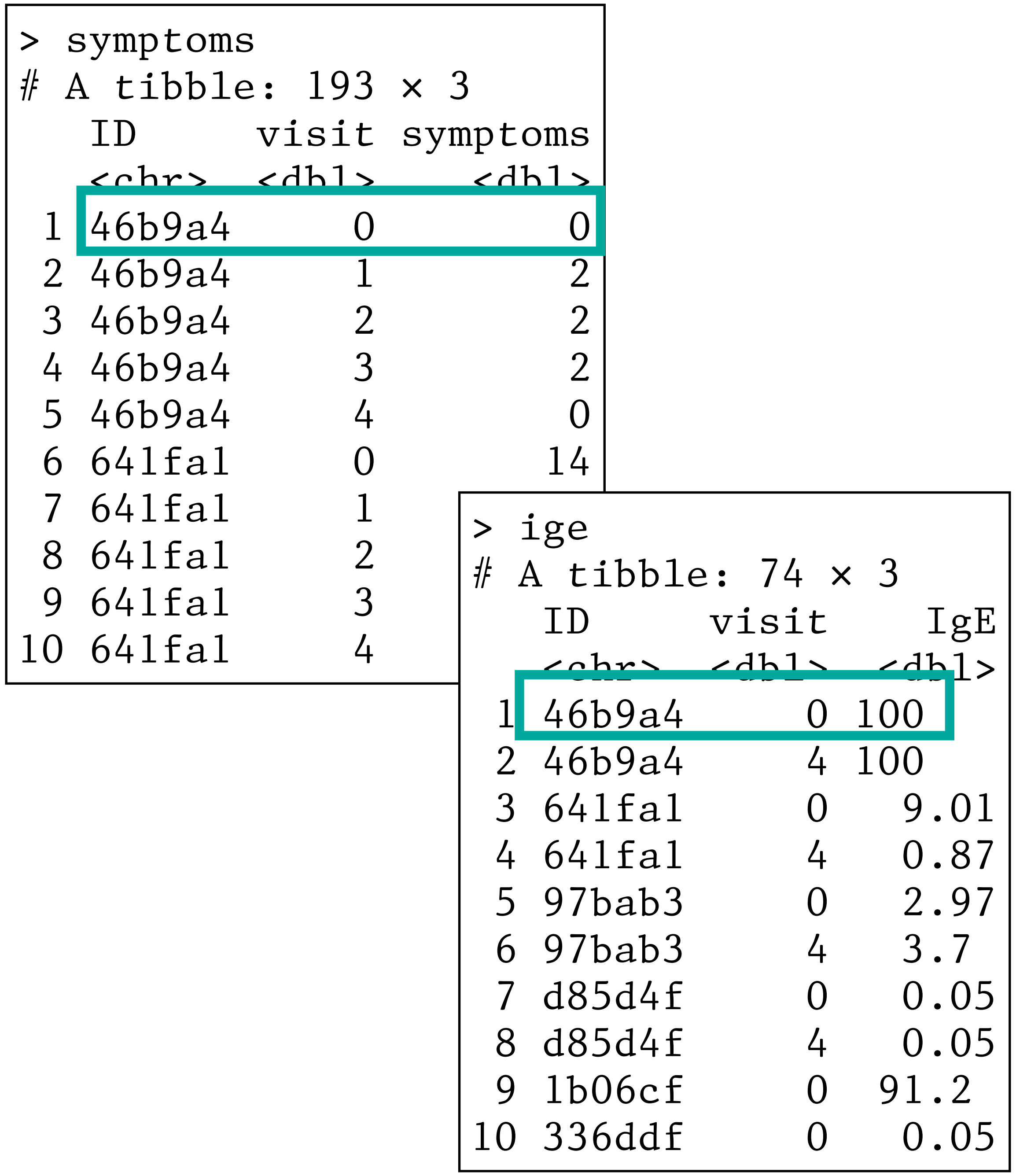


```
# A tibble: 74 × 4
   ID      visit symptoms    IgE
   <chr>   <dbl>    <dbl>  <dbl>
 1 46b9a4      0        0 100
 2 46b9a4      4        0 100
 3 641fa1      0       14   9.01
 4 641fa1      4        2   0.87
 5 97bab3      0        0   2.97
 6 97bab3      4        0   3.7
 7 d85d4f      0        0   0.05
 8 d85d4f      4        3   0.05
 9 1b06cf      0        4  91.2
10 336ddf      0        2   0.05
```

inner join only keeps rows where there is a match in BOTH tables

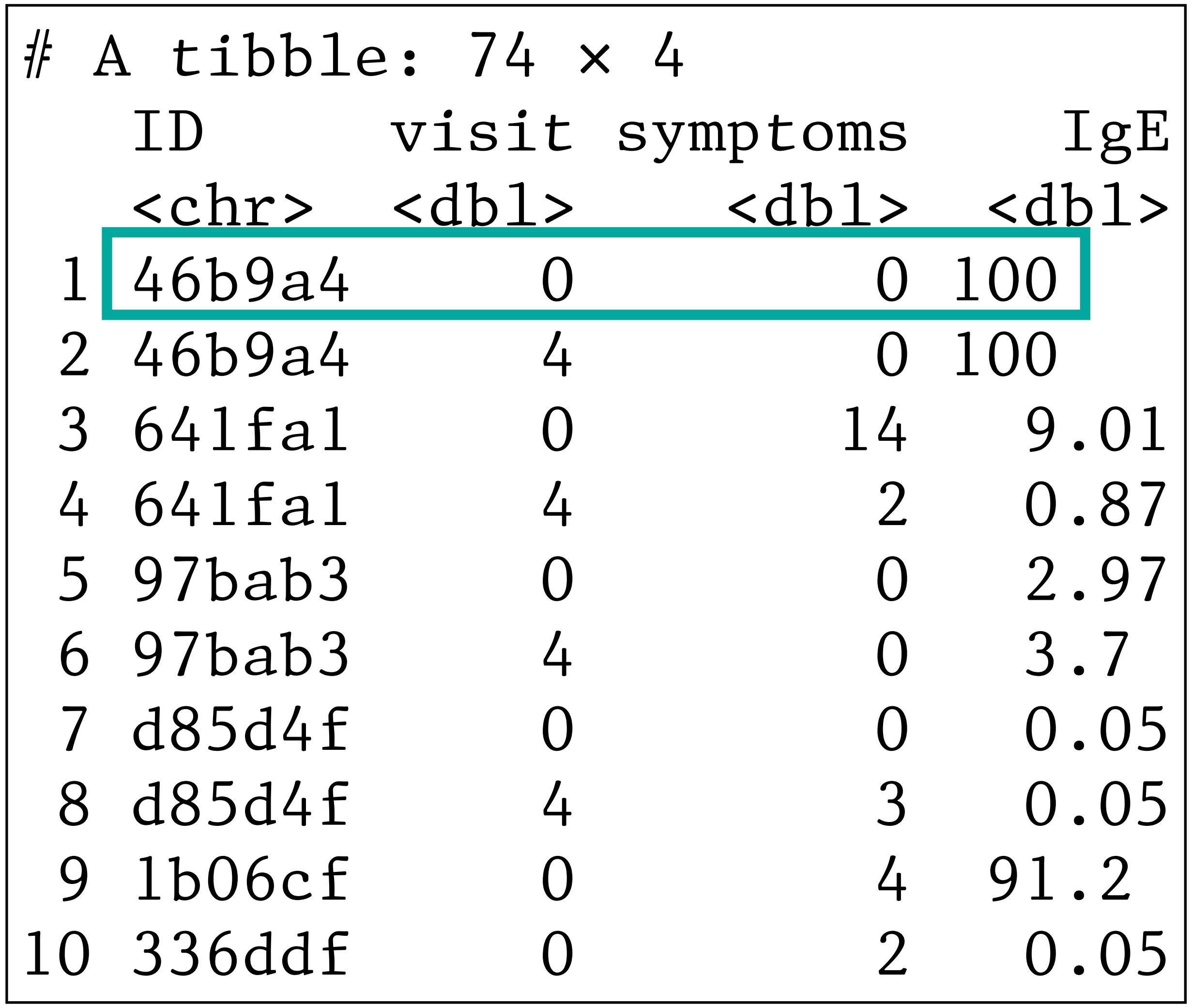
Symptoms vs. IgE

```
symptoms |>
    inner_join(ige, by = c("ID", "visit"))
```



```
> symptoms
# A tibble: 193 × 3
   ID      visit symptoms
   <chr>   <dbl>    <dbl>
 1 46b9a4      0        0
 2 46b9a4      1        2
 3 46b9a4      2        2
 4 46b9a4      3        2
 5 46b9a4      4        0
 6 641fa1      0       14
 7 641fa1      1
 8 641fa1      2
 9 641fa1      3
10 641fa1      4
```

```
> ige
# A tibble: 74 × 3
   ID      visit    IgE
   <chr>   <dbl>  <dbl>
 1 46b9a4      0 100
 2 46b9a4      4 100
 3 641fa1      0   9.01
 4 641fa1      4   0.87
 5 97bab3      0   2.97
 6 97bab3      4   3.7
 7 d85d4f      0   0.05
 8 d85d4f      4   0.05
 9 1b06cf      0  91.2
10 336ddf      0   0.05
```

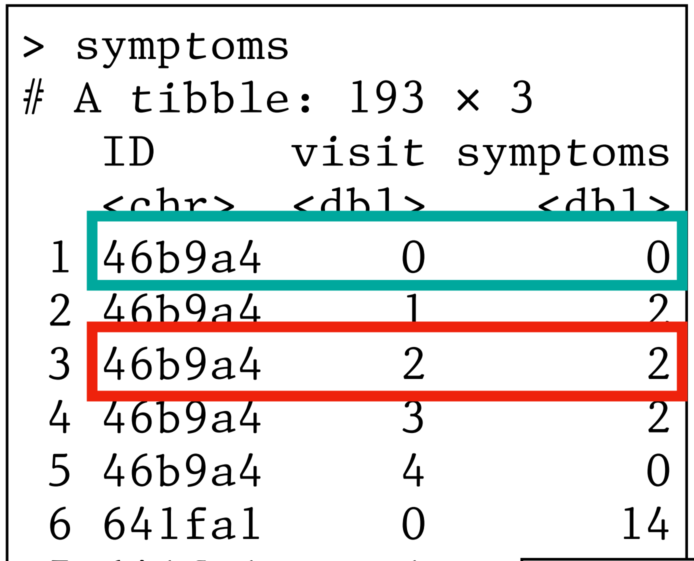


```
# A tibble: 74 × 4
   ID      visit symptoms    IgE
   <chr>   <dbl>    <dbl>  <dbl>
 1 46b9a4      0        0 100
 2 46b9a4      4        0 100
 3 641fa1      0       14   9.01
 4 641fa1      4        2   0.87
 5 97bab3      0        0   2.97
 6 97bab3      4        0   3.7
 7 d85d4f      0        0   0.05
 8 d85d4f      4        3   0.05
 9 1b06cf      0        4  91.2
10 336ddf      0        2   0.05
```

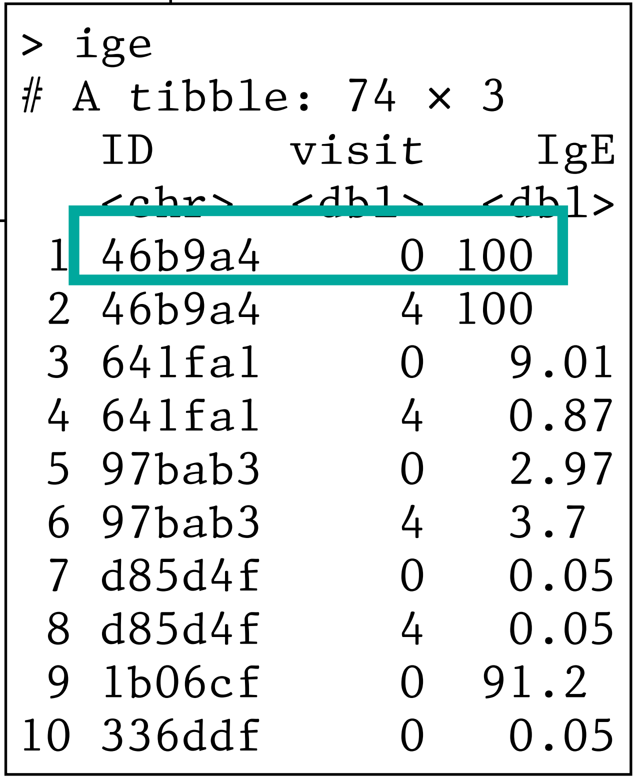
inner join only keeps rows where there is a match in BOTH tables

Symptoms vs. IgE

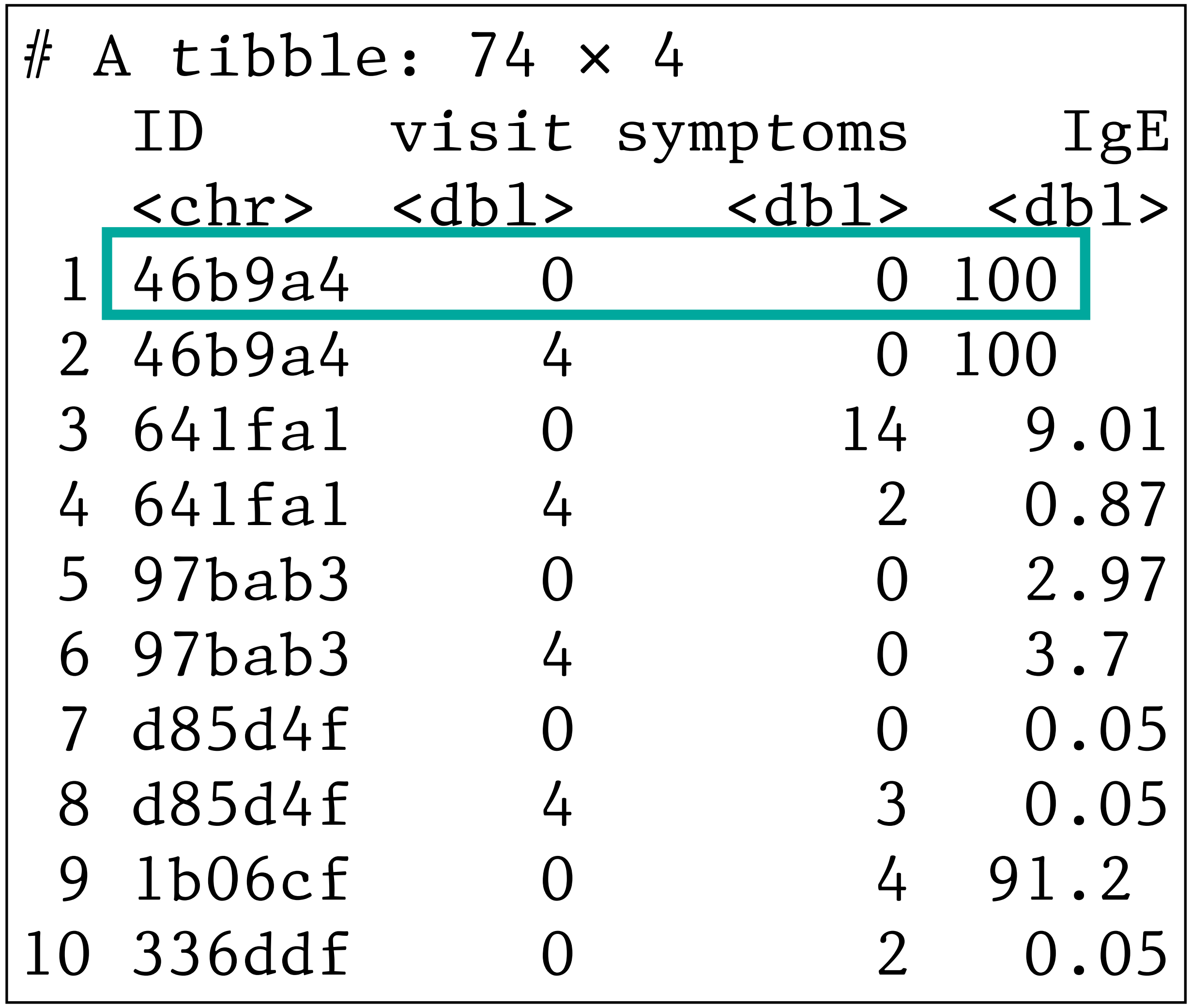
```
symptoms |>
    inner_join(ige, by = c("ID", "visit"))
```



```
> symptoms
# A tibble: 193 × 3
   ID      visit symptoms
   <chr>   <dbl>    <dbl>
 1 46b9a4      0        0
 2 46b9a4      1        2
 3 46b9a4      2        2
 4 46b9a4      3        2
 5 46b9a4      4        0
 6 641fa1      0       14
 7 641fa1      1
 8 641fa1      2
 9 641fa1      3
10 641fa1      4
```



```
> ige
# A tibble: 74 × 3
   ID      visit    IgE
   <chr>   <dbl>  <dbl>
 1 46b9a4      0 100
 2 46b9a4      4 100
 3 641fa1      0   9.01
 4 641fa1      4   0.87
 5 97bab3      0   2.97
 6 97bab3      4   3.7
 7 d85d4f      0   0.05
 8 d85d4f      4   0.05
 9 1b06cf      0  91.2
10 336ddf      0   0.05
```

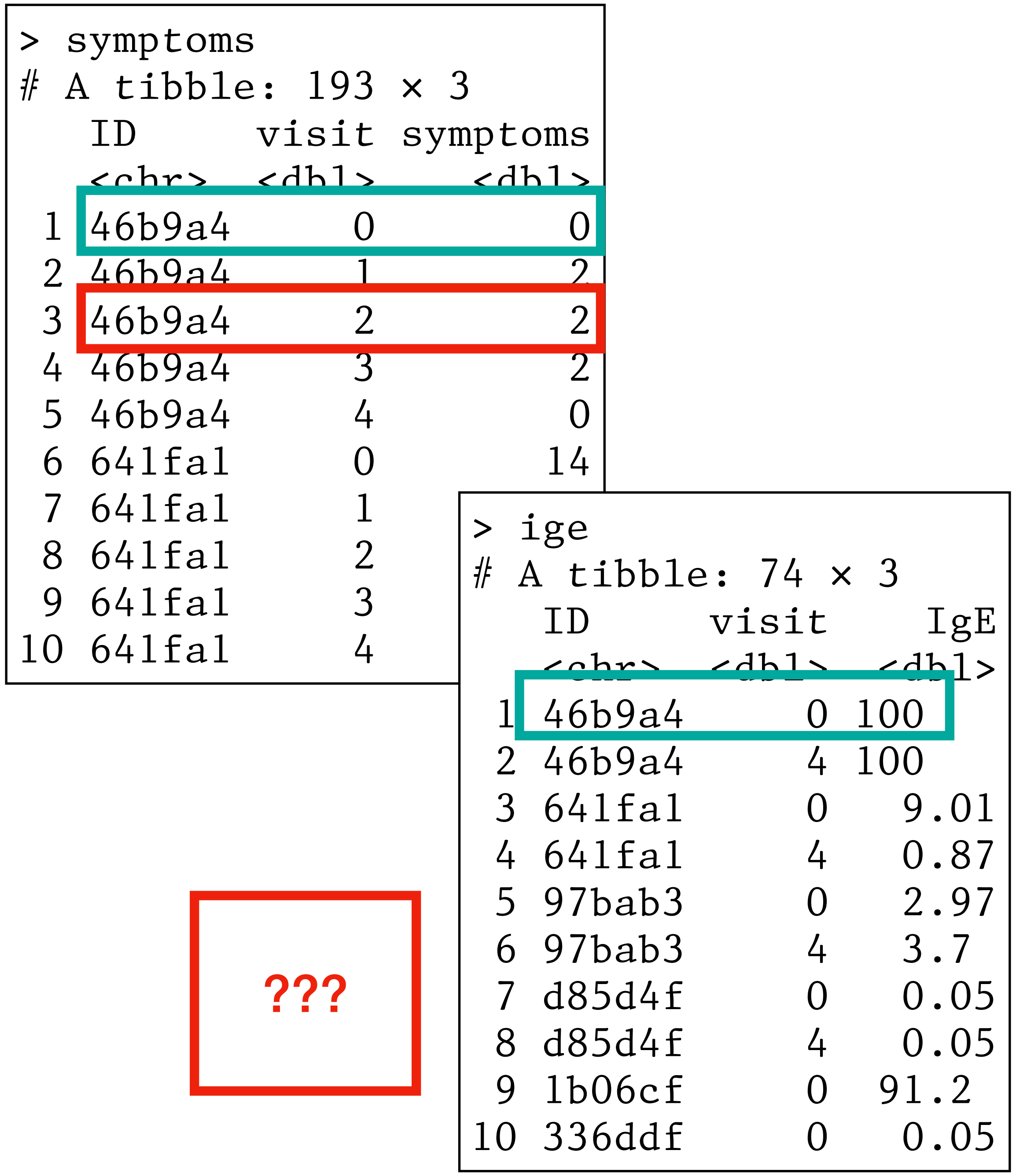


```
# A tibble: 74 × 4
   ID      visit symptoms    IgE
   <chr>   <dbl>    <dbl>  <dbl>
 1 46b9a4      0        0 100
 2 46b9a4      4        0 100
 3 641fa1      0       14   9.01
 4 641fa1      4        2   0.87
 5 97bab3      0        0   2.97
 6 97bab3      4        0   3.7
 7 d85d4f      0        0   0.05
 8 d85d4f      4        3   0.05
 9 1b06cf      0        4  91.2
10 336ddf      0        2   0.05
```

inner join only keeps rows where there is a match in BOTH tables

Symptoms vs. IgE

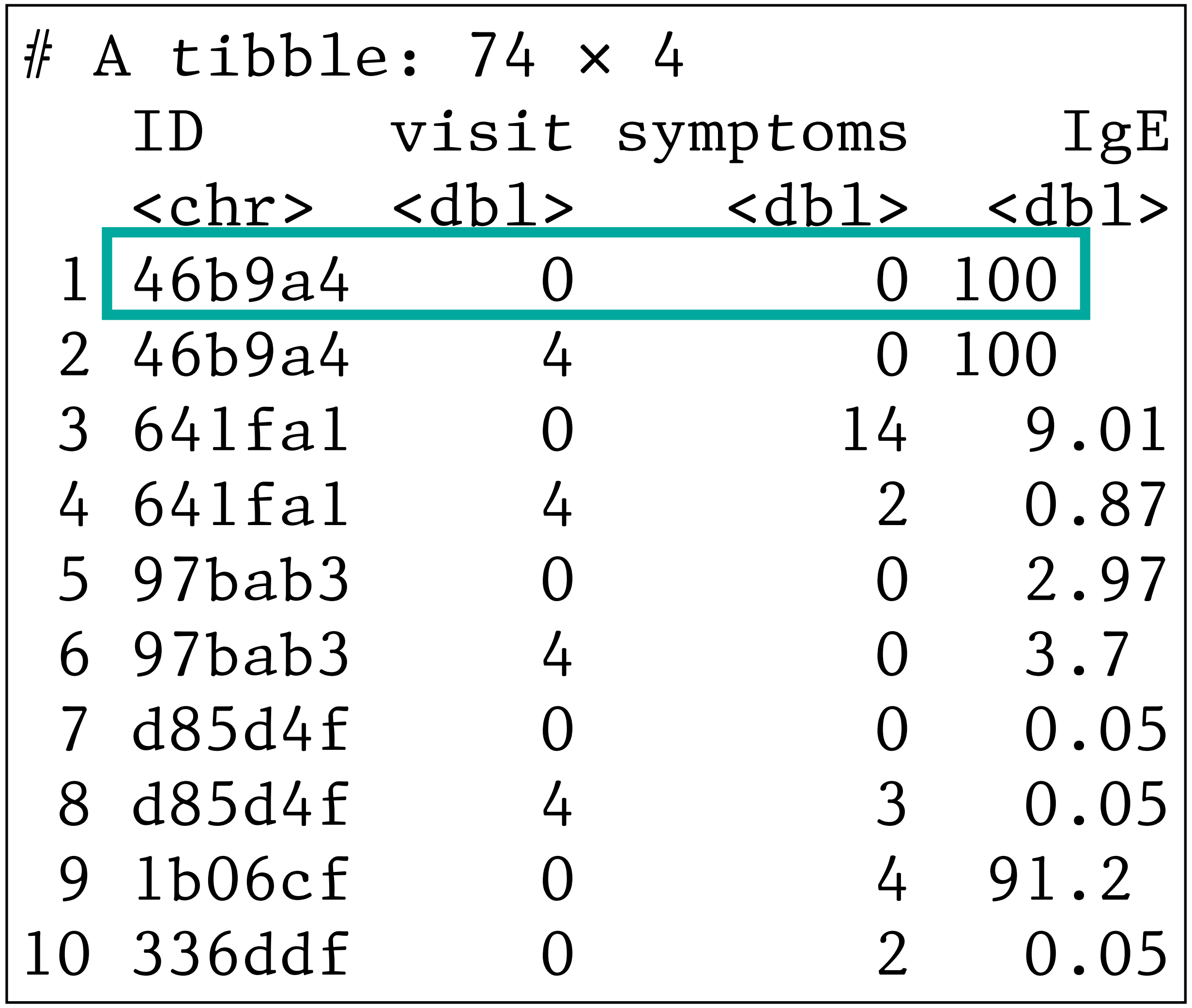
```
symptoms |>
    inner_join(ige, by = c("ID", "visit"))
```



```
> symptoms
# A tibble: 193 × 3
   ID      visit symptoms
   <chr>   <dbl>    <dbl>
 1 46b9a4      0        0
 2 46b9a4      1        2
 3 46b9a4      2        2
 4 46b9a4      3        2
 5 46b9a4      4        0
 6 641fa1      0       14
 7 641fa1      1
 8 641fa1      2
 9 641fa1      3
10 641fa1      4
```

```
> ige
# A tibble: 74 × 3
   ID      visit   IgE
   <chr>   <dbl> <dbl>
 1 46b9a4      0 100
 2 46b9a4      4 100
 3 641fa1      0   9.01
 4 641fa1      4   0.87
 5 97bab3      0   2.97
 6 97bab3      4   3.7
 7 d85d4f      0   0.05
 8 d85d4f      4   0.05
 9 1b06cf      0  91.2
10 336ddf      0   0.05
```

???

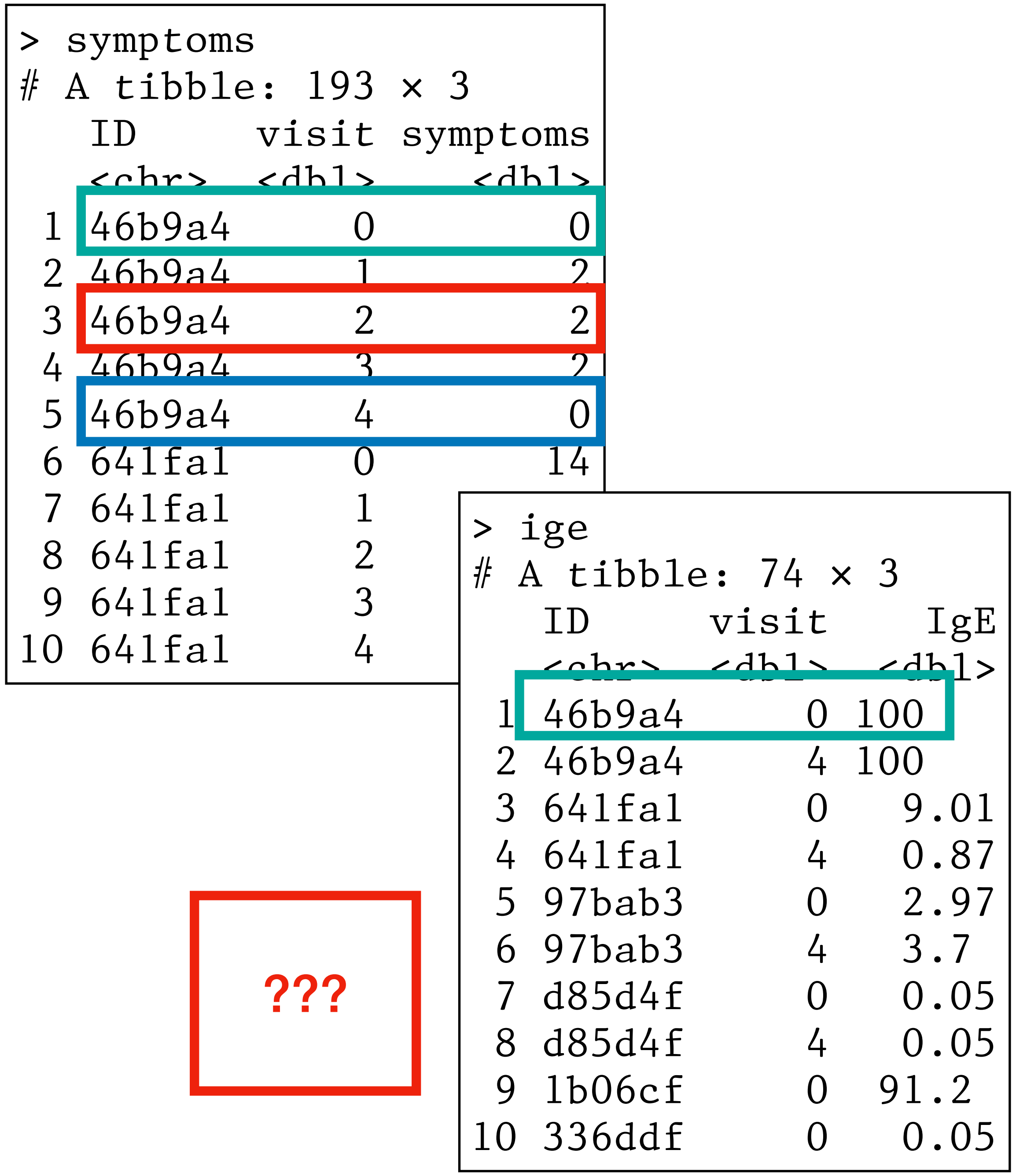


```
# A tibble: 74 × 4
   ID      visit symptoms    IgE
   <chr>   <dbl>    <dbl>  <dbl>
 1 46b9a4      0        0 100
 2 46b9a4      4        0 100
 3 641fa1      0       14   9.01
 4 641fa1      4        2   0.87
 5 97bab3      0        0   2.97
 6 97bab3      4        0   3.7
 7 d85d4f      0        0   0.05
 8 d85d4f      4        3   0.05
 9 1b06cf      0        4  91.2
10 336ddf      0        2   0.05
```

inner join only keeps rows where there is a match in BOTH tables

Symptoms vs. IgE

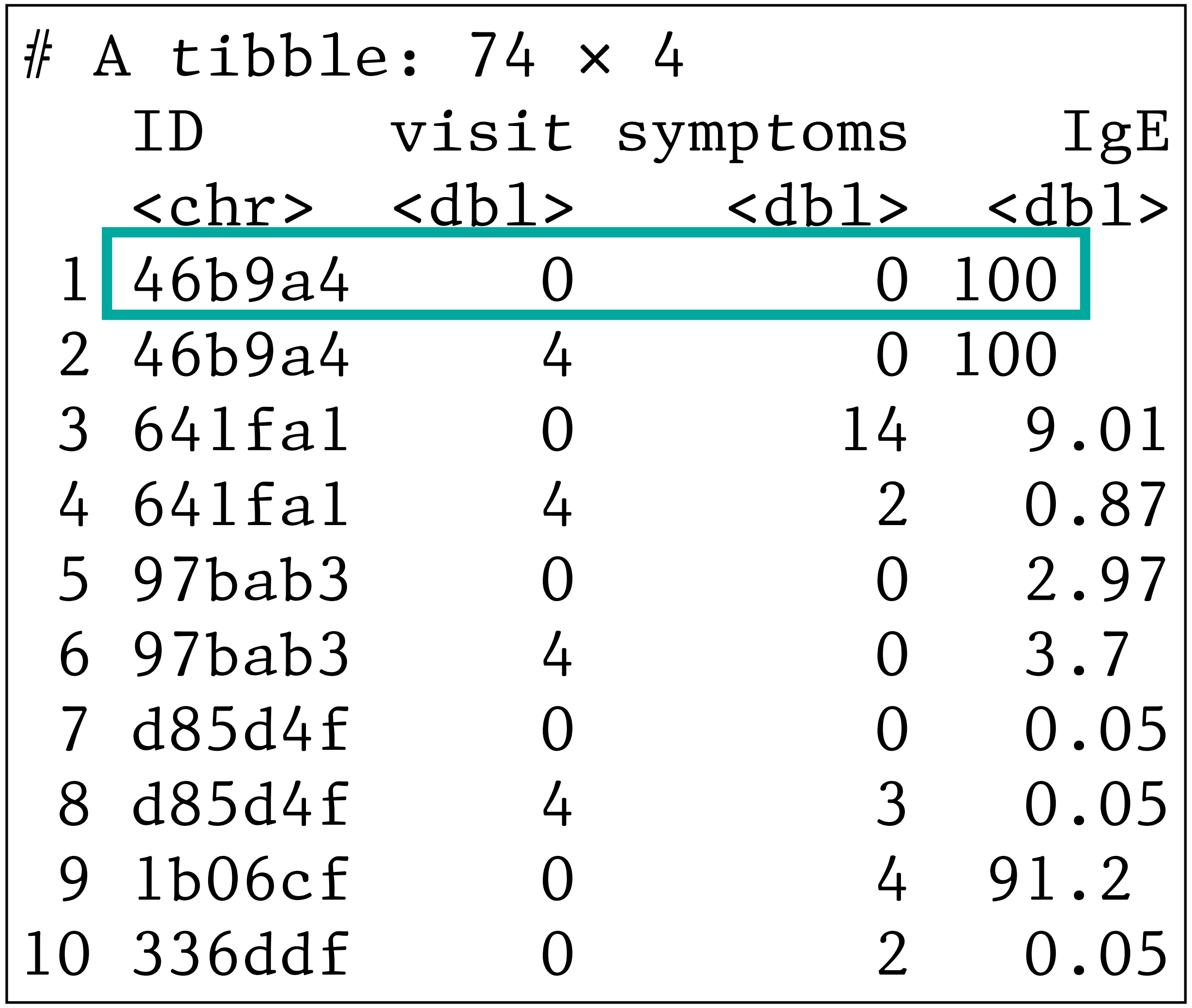
```
symptoms |>
    inner_join(ige, by = c("ID", "visit"))
```



```
> symptoms
# A tibble: 193 × 3
   ID      visit symptoms
   <chr>   <dbl>    <dbl>
 1 46b9a4      0        0
 2 46b9a4      1        2
 3 46b9a4      2        2
 4 46b9a4      3        2
 5 46b9a4      4        0
 6 641fa1      0       14
 7 641fa1      1
 8 641fa1      2
 9 641fa1      3
10 641fa1      4
```

```
> ige
# A tibble: 74 × 3
   ID      visit   IgE
   <chr>   <dbl> <dbl>
 1 46b9a4      0 100
 2 46b9a4      4 100
 3 641fa1      0   9.01
 4 641fa1      4   0.87
 5 97bab3      0   2.97
 6 97bab3      4   3.7
 7 d85d4f      0   0.05
 8 d85d4f      4   0.05
 9 1b06cf      0  91.2
10 336ddf      0   0.05
```

???

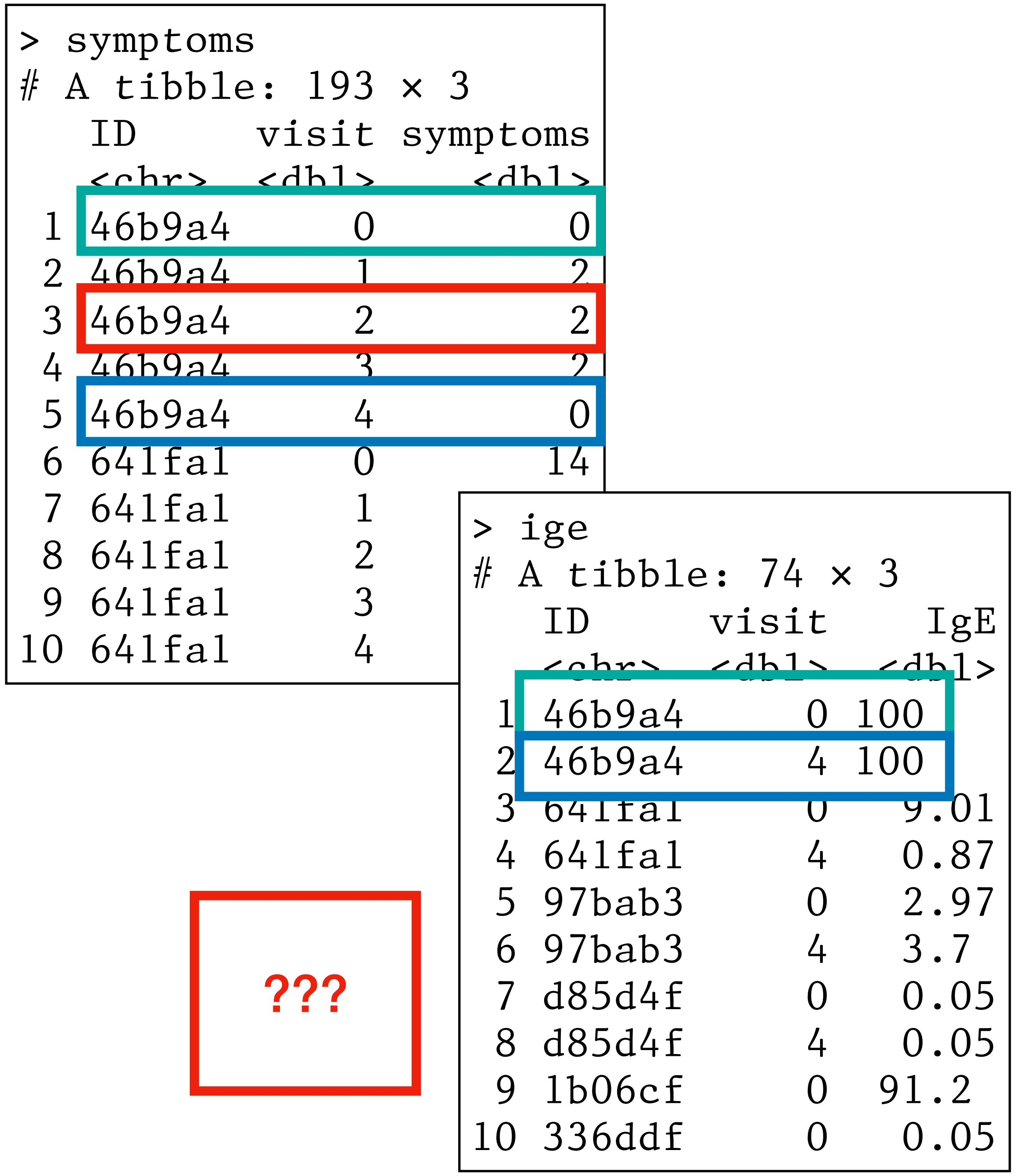


```
# A tibble: 74 × 4
   ID      visit symptoms    IgE
   <chr>   <dbl>    <dbl>  <dbl>
 1 46b9a4      0        0 100
 2 46b9a4      4        0 100
 3 641fa1      0       14   9.01
 4 641fa1      4        2   0.87
 5 97bab3      0        0   2.97
 6 97bab3      4        0   3.7
 7 d85d4f      0        0   0.05
 8 d85d4f      4        3   0.05
 9 1b06cf      0        4  91.2
10 336ddf      0        2   0.05
```

inner join only keeps rows where there is a match in BOTH tables

Symptoms vs. IgE

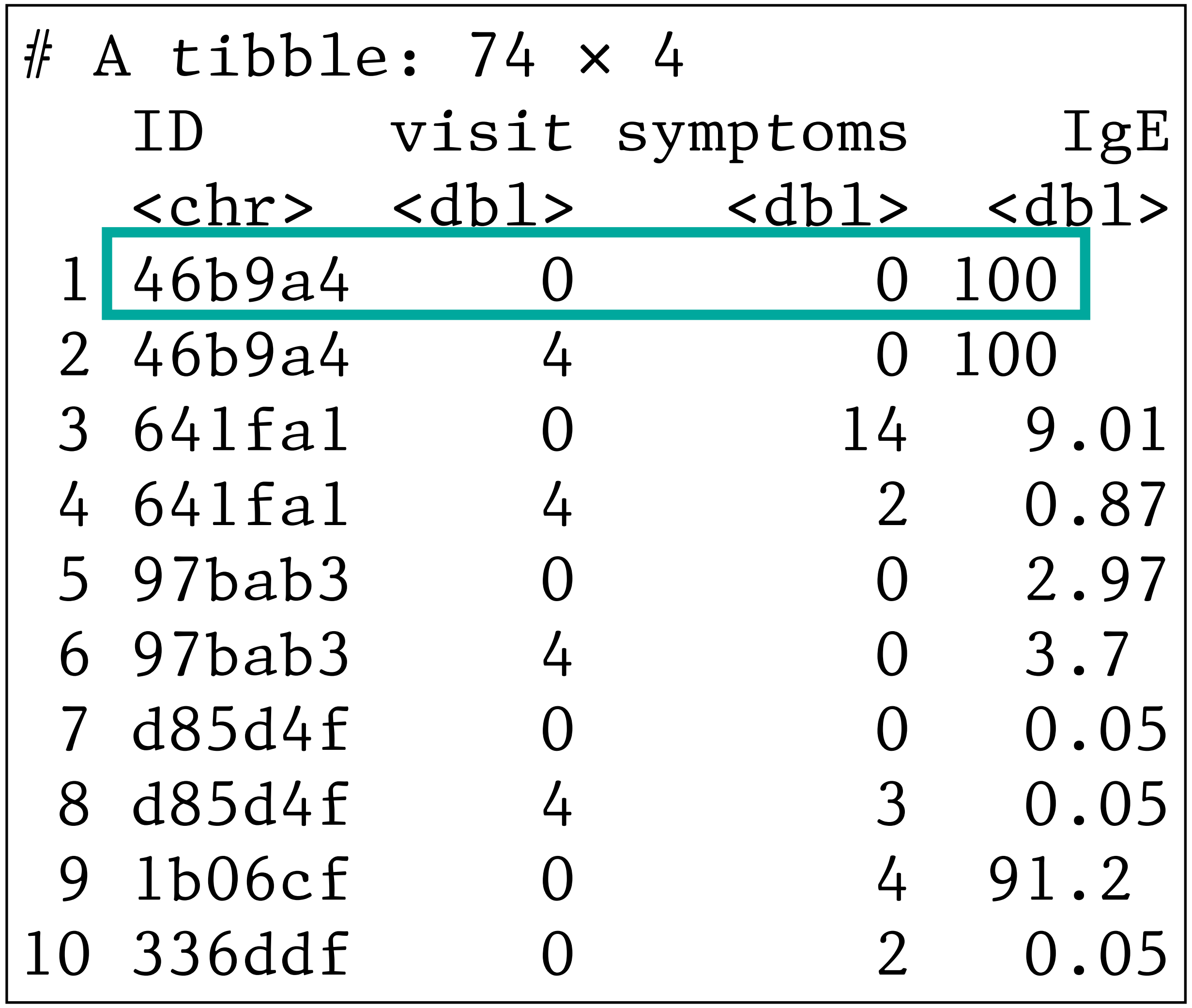
```
symptoms |>
    inner_join(ige, by = c("ID", "visit"))
```



```
> symptoms
# A tibble: 193 × 3
   ID      visit symptoms
   <chr>   <dbl>    <dbl>
 1 46b9a4      0        0
 2 46b9a4      1        2
 3 46b9a4      2        2
 4 46b9a4      3        2
 5 46b9a4      4        0
 6 641fa1      0       14
 7 641fa1      1
 8 641fa1      2
 9 641fa1      3
10 641fa1      4
```

```
> ige
# A tibble: 74 × 3
   ID      visit    IgE
   <chr>   <dbl>  <dbl>
 1 46b9a4      0 100
 2 46b9a4      4 100
 3 641fa1      0   9.01
 4 641fa1      4   0.87
 5 97bab3      0   2.97
 6 97bab3      4   3.7
 7 d85d4f      0   0.05
 8 d85d4f      4   0.05
 9 1b06cf      0  91.2
10 336ddf      0   0.05
```

???

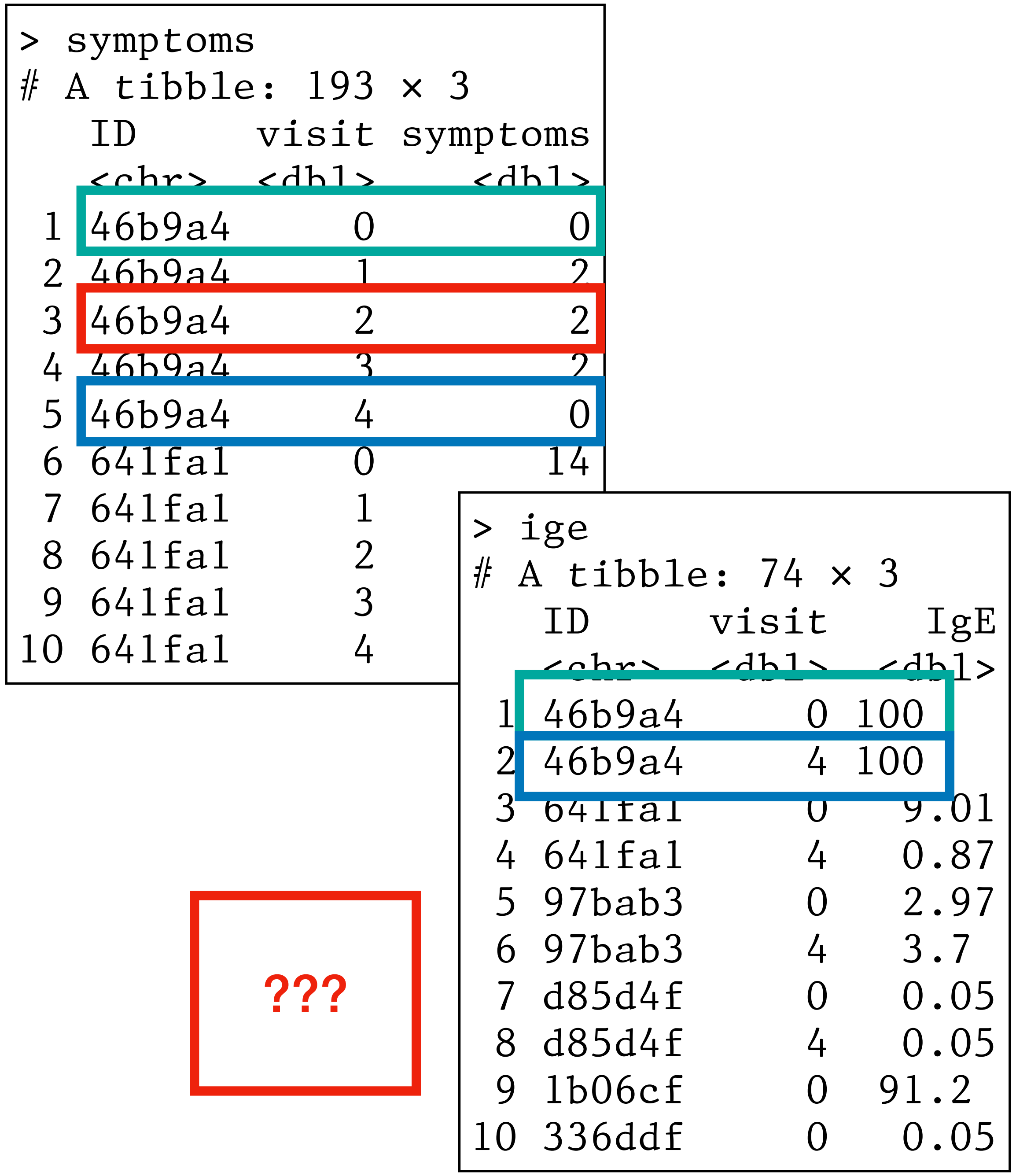


```
# A tibble: 74 × 4
   ID      visit symptoms    IgE
   <chr>   <dbl>    <dbl>  <dbl>
 1 46b9a4      0        0 100
 2 46b9a4      4        0 100
 3 641fa1      0       14   9.01
 4 641fa1      4        2   0.87
 5 97bab3      0        0   2.97
 6 97bab3      4        0   3.7
 7 d85d4f      0        0   0.05
 8 d85d4f      4        3   0.05
 9 1b06cf      0        4  91.2
10 336ddf      0        2   0.05
```

inner join only keeps rows where there is a match in BOTH tables

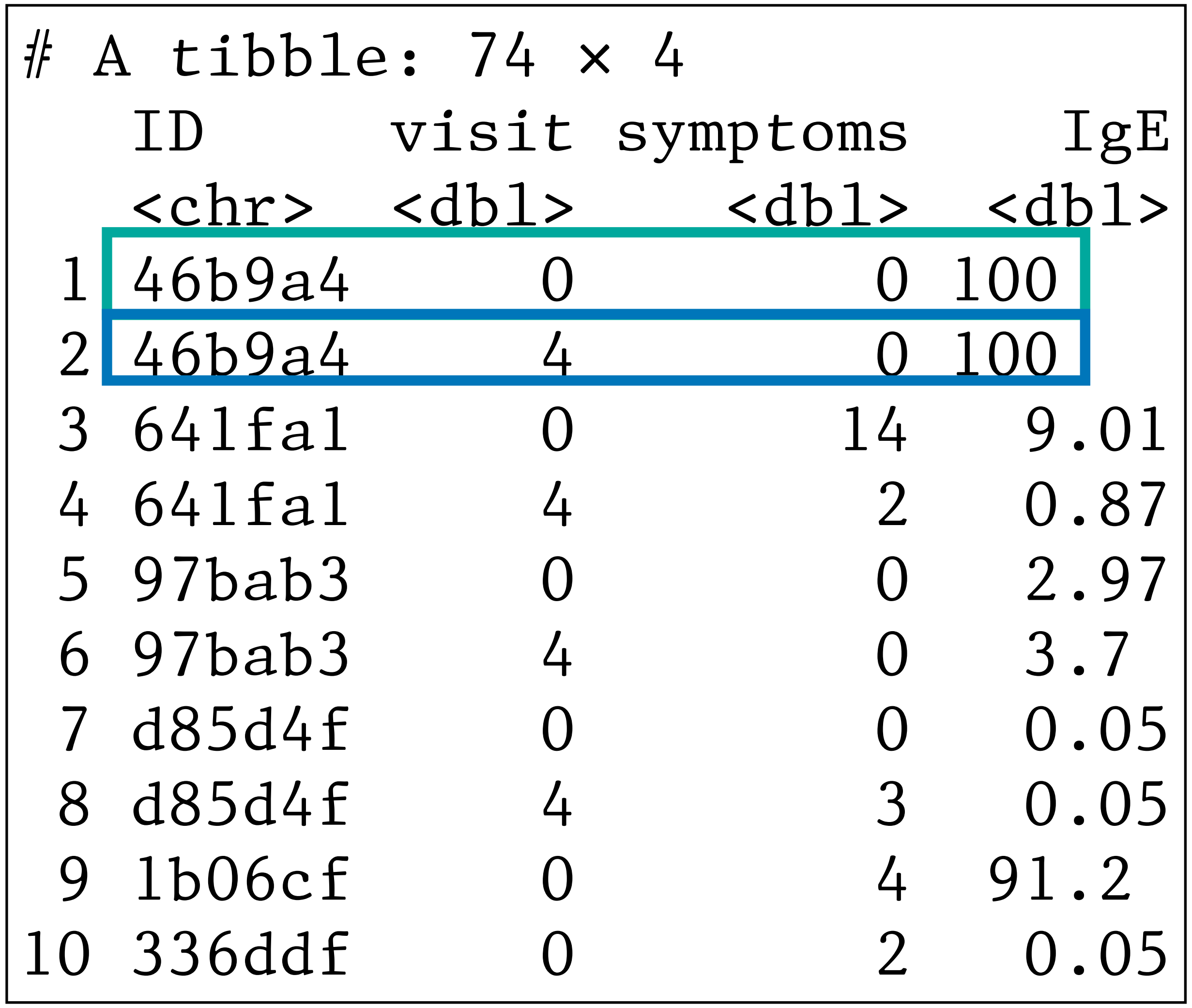
Symptoms vs. IgE

```
symptoms |>
    inner_join(ige, by = c("ID", "visit"))
```



```
> symptoms
# A tibble: 193 × 3
   ID      visit symptoms
   <chr>   <dbl>    <dbl>
 1 46b9a4      0        0
 2 46b9a4      1        2
 3 46b9a4      2        2
 4 46b9a4      3        2
 5 46b9a4      4        0
 6 641fa1      0       14
 7 641fa1      1
 8 641fa1      2
 9 641fa1      3
10 641fa1      4
```

```
> ige
# A tibble: 74 × 3
   ID      visit   IgE
   <chr>   <dbl> <dbl>
 1 46b9a4      0 100
 2 46b9a4      4 100
 3 641fa1      0  9.01
 4 641fa1      4  0.87
 5 97bab3      0  2.97
 6 97bab3      4  3.7
 7 d85d4f      0  0.05
 8 d85d4f      4  0.05
 9 1b06cf      0 91.2
10 336ddf      0  0.05
```

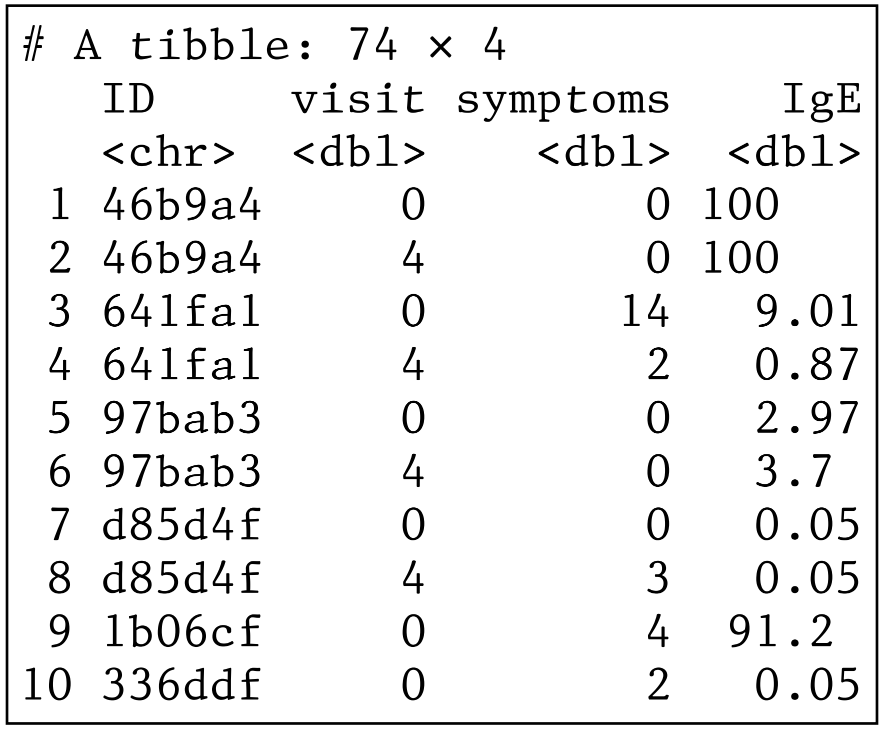


```
# A tibble: 74 × 4
   ID      visit symptoms   IgE
   <chr>   <dbl>    <dbl> <dbl>
 1 46b9a4      0        0 100
 2 46b9a4      4        0 100
 3 641fa1      0       14  9.01
 4 641fa1      4        2  0.87
 5 97bab3      0        0  2.97
 6 97bab3      4        0  3.7
 7 d85d4f      0        0  0.05
 8 d85d4f      4        3  0.05
 9 1b06cf      0        4 91.2
10 336ddf      0        2  0.05
```

inner join only keeps rows where there is a match in BOTH tables

Symptoms vs. IgE

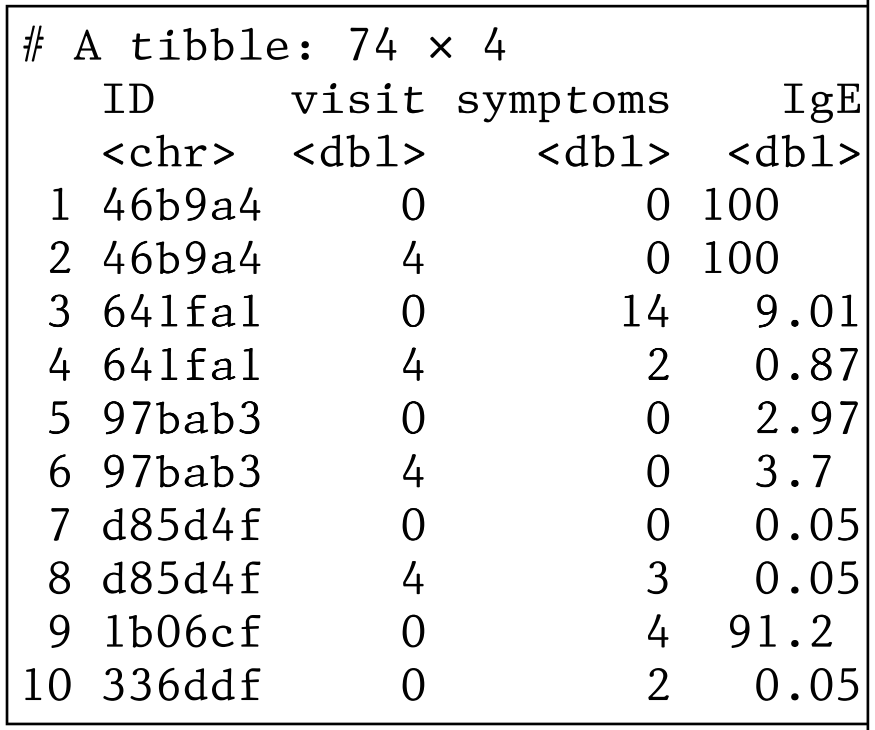
```
symptoms |>
    inner_join(ige, by = c("ID", "visit")) |>
    ggplot(aes(x = IgE, y = symptoms)) +
    geom_point(size = 3)
```



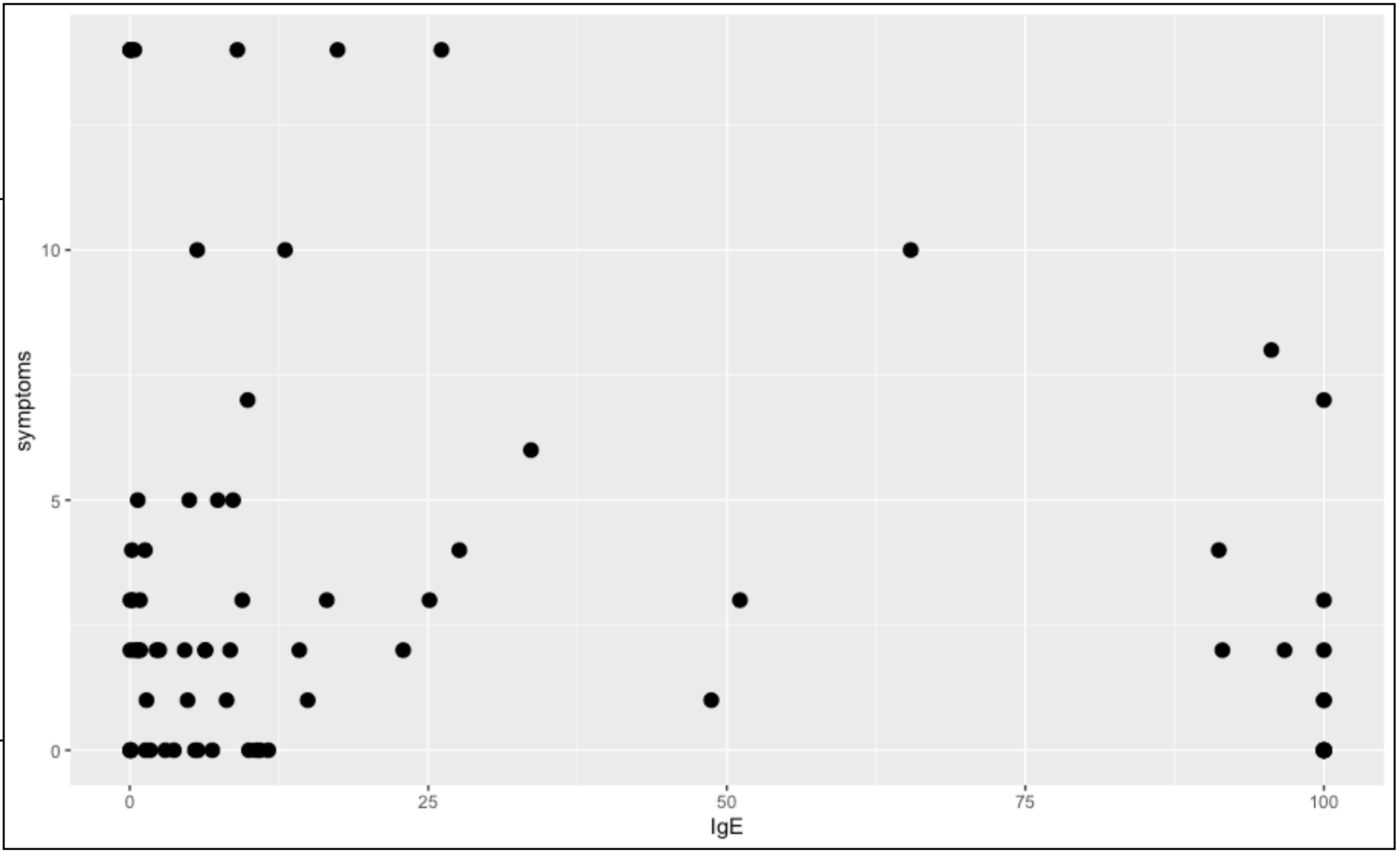
```
# A tibble: 74 × 4
   ID     visit symptoms    IgE
   <chr>  <dbl>    <dbl>  <dbl>
 1 46b9a4     0        0 100   
 2 46b9a4     4        0 100   
 3 641fa1     0       14   9.01
 4 641fa1     4        2   0.87
 5 97bab3     0        0   2.97
 6 97bab3     4        0   3.7 
 7 d85d4f     0        0   0.05
 8 d85d4f     4        3   0.05
 9 1b06cf     0        4  91.2 
10 336ddf     0        2   0.05
```

Symptoms vs. IgE

```
symptoms |>
    inner_join(ige, by = c("ID", "visit")) |>
    ggplot(aes(x = IgE, y = symptoms)) +
    geom_point(size = 3)
```



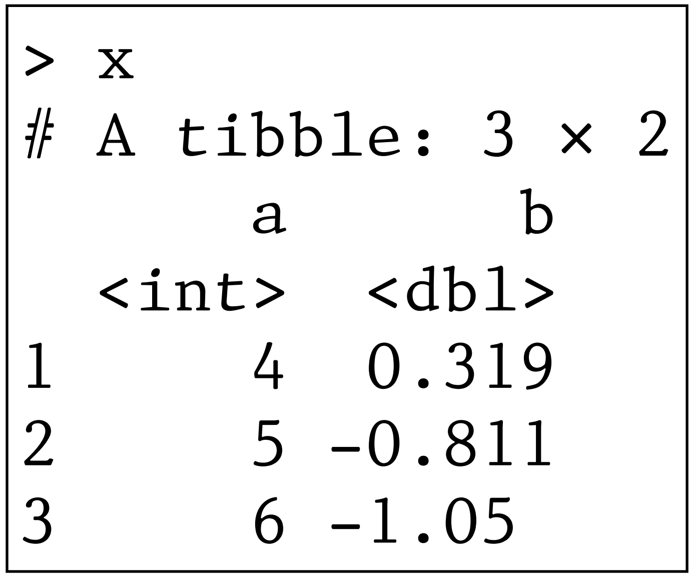
```
# A tibble: 74 × 4
   ID     visit symptoms    IgE
   <chr>  <dbl>    <dbl>  <dbl>
 1 46b9a4     0        0 100
 2 46b9a4     4        0 100
 3 641fa1     0       14   9.01
 4 641fa1     4        2   0.87
 5 97bab3     0        0   2.97
 6 97bab3     4        0   3.7
 7 d85d4f     0        0   0.05
 8 d85d4f     4        3   0.05
 9 1b06cf     0        4  91.2
10 336ddf     0        2   0.05
```



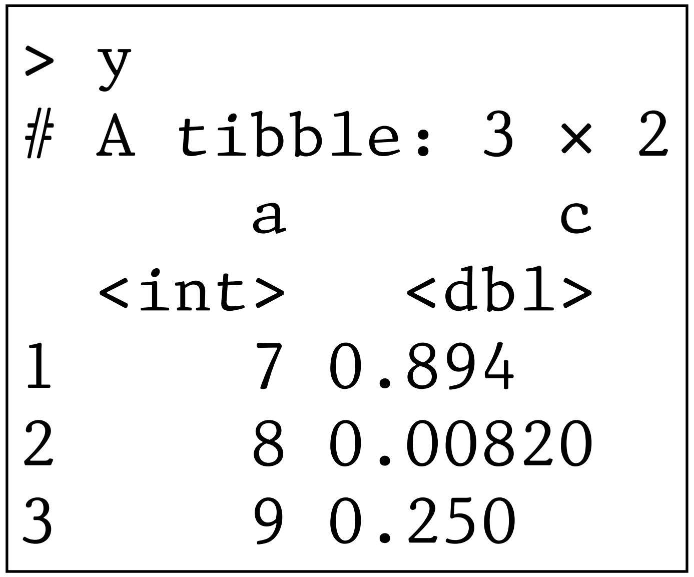
Joins Summary (So Far)

- `x |> left_join(y, by = "column name")`
  - Always keeps the rows of x and adds the columns in y that match the rows of x; Add NAs for the rows that don't match
- `x |> inner_join(y, by = "column name")`
  - Only keep the rows of x that match the rows of y
- `x |> right_join(y, by = "column name")`
  - Always keeps the rows of y and adds the columns in x that match the rows of y; Add NAs for the rows that don't match

Joins Summary (So Far)

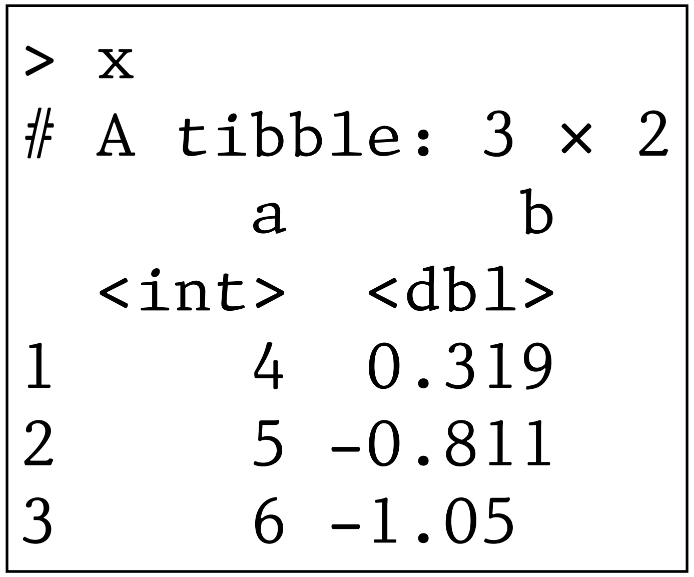


```
> x
# A tibble: 3 × 2
      a      b
  <int>  <dbl>
1     4  0.319
2     5 -0.811
3     6 -1.05
```

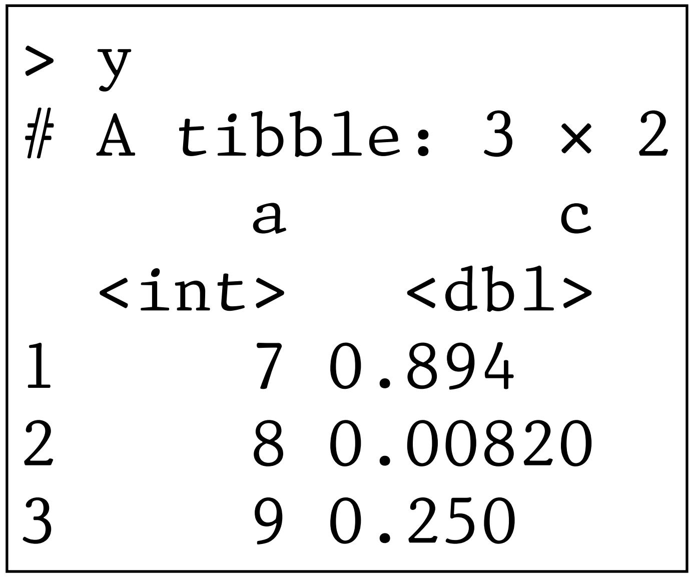


```
> y
# A tibble: 3 × 2
      a       c
  <int>   <dbl>
1     7 0.894
2     8 0.00820
3     9 0.250
```

Joins Summary (So Far)



```
> x
# A tibble: 3 × 2
      a      b
  <int>  <dbl>
1     4  0.319
2     5 -0.811
3     6 -1.05
```



```
> y
# A tibble: 3 × 2
      a       c
  <int>   <dbl>
1     7 0.894
2     8 0.00820
3     9 0.250
```

```
x |>
    left_join(y, by = "a")
```

?

```
x |>
    right_join(y, by = "a")
```

?

```
x |>
    inner_join(y, by = "a")
```

?

Less Common Joins

- **full_join(x, y)**
  - Keep all observations from both data frames and add NA values for unmatched rows
- **anti_join(x, y)**
  - Remove the rows from x that match in y

Full Join

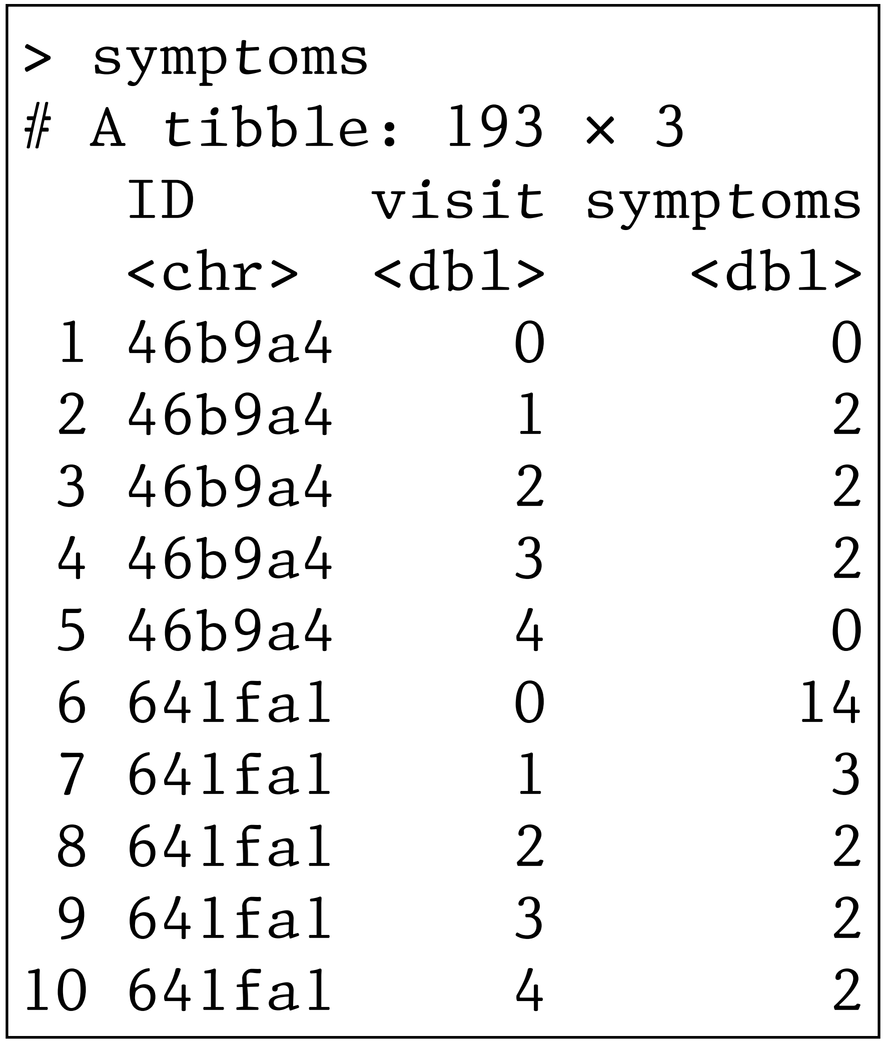
```
> x
# A tibble: 3 × 2
      a      b
  <int>  <dbl>
1     4  0.319
2     5 -0.811
3     6 -1.05
```

```
> y
# A tibble: 3 × 2
      a       c
  <int>   <dbl>
1     7 0.894
2     8 0.00820
3     9 0.250
```

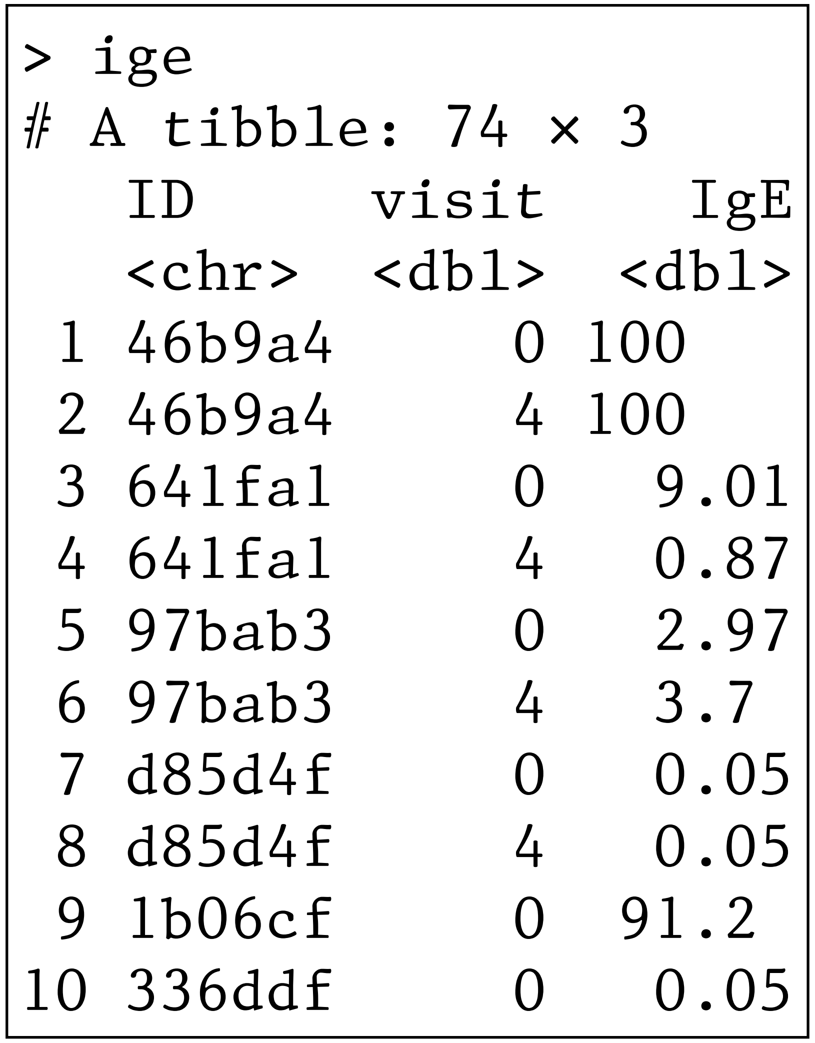
```
x |>
    full_join(y, by = "a")
```

```
# A tibble: 6 × 3
      a      b       c
  <int>  <dbl>   <dbl>
1     4  0.319 NA
2     5 -0.811 NA
3     6 -1.05  NA
4     7 NA       0.894
5     8 NA       0.00820
6     9 NA       0.250
```

Anti Join



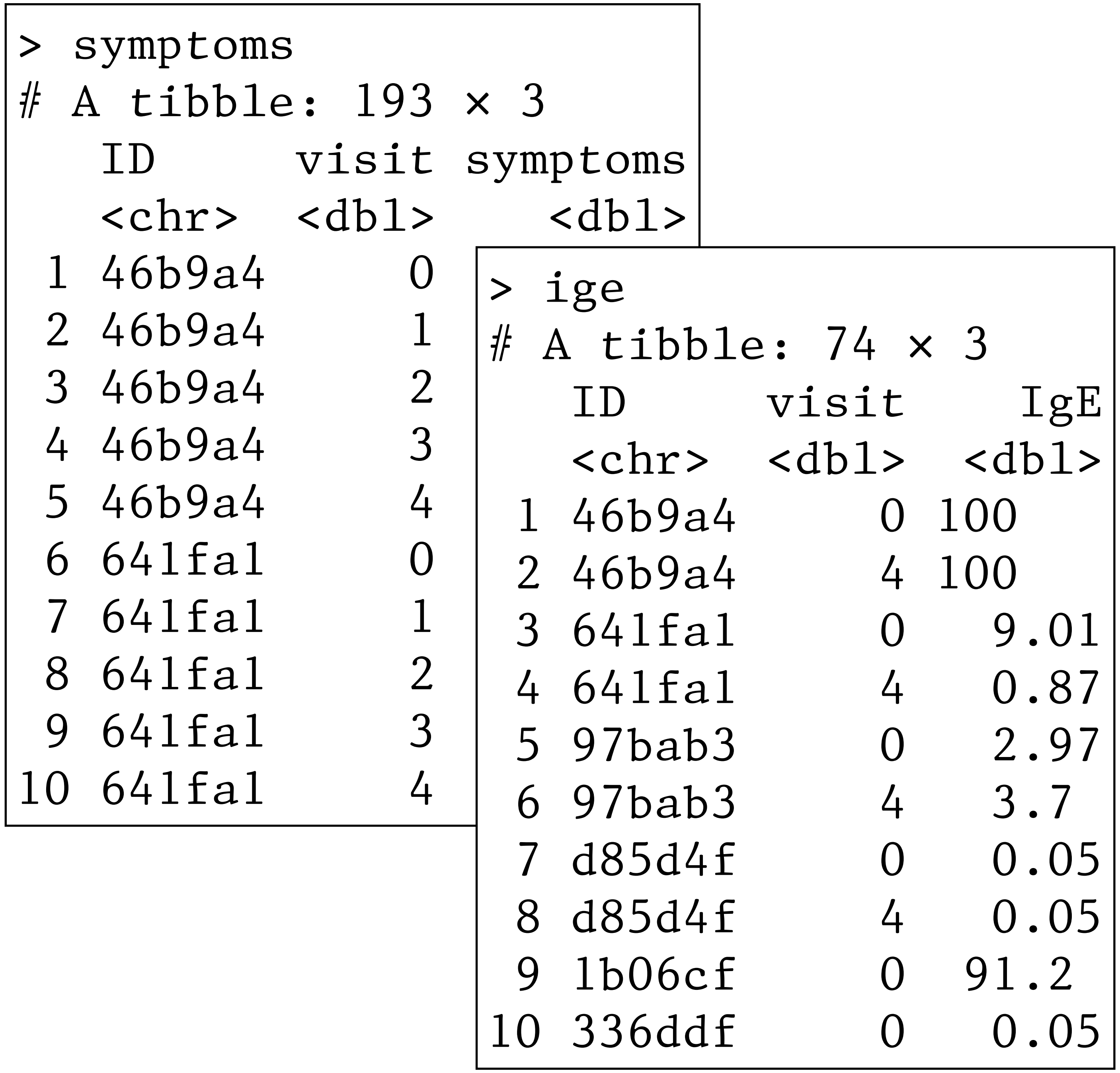
```
> symptoms
# A tibble: 193 × 3
   ID     visit symptoms
   <chr>  <dbl>    <dbl>
 1 46b9a4     0        0
 2 46b9a4     1        2
 3 46b9a4     2        2
 4 46b9a4     3        2
 5 46b9a4     4        0
 6 641fa1     0       14
 7 641fa1     1        3
 8 641fa1     2        2
 9 641fa1     3        2
10 641fa1     4        2
```



```
> ige
# A tibble: 74 × 3
   ID     visit    IgE
   <chr>  <dbl>  <dbl>
 1 46b9a4     0 100   
 2 46b9a4     4 100   
 3 641fa1     0   9.01
 4 641fa1     4   0.87
 5 97bab3     0   2.97
 6 97bab3     4   3.7 
 7 d85d4f     0   0.05
 8 d85d4f     4   0.05
 9 1b06cf     0  91.2 
10 336ddf     0   0.05
```

Anti Join

```
symptoms |>
    anti_join(ige, by = c("ID", "visit"))
```

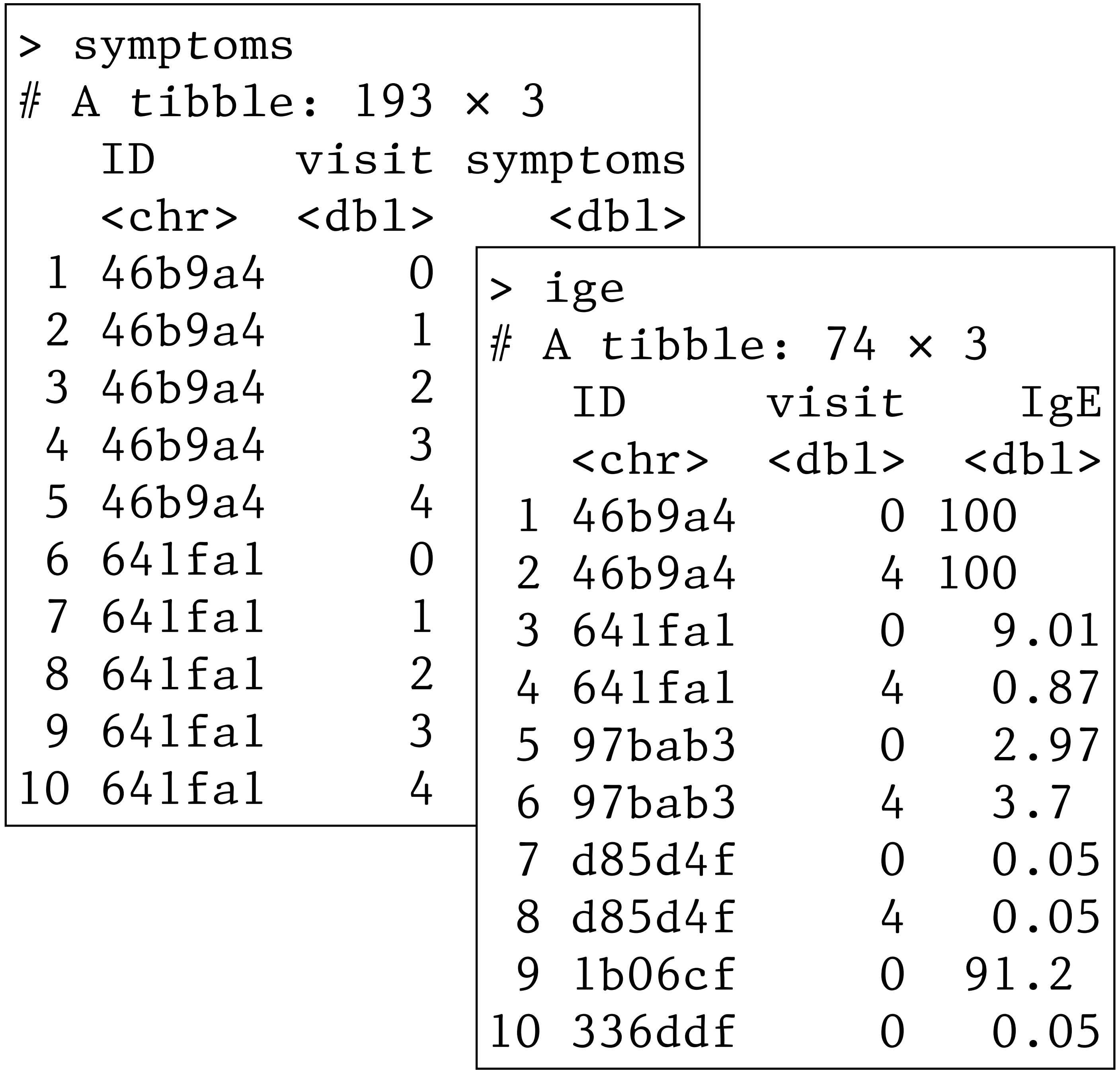


```
> symptoms
# A tibble: 193 × 3
   ID     visit symptoms
   <chr>  <dbl>    <dbl>
 1 46b9a4     0
 2 46b9a4     1
 3 46b9a4     2
 4 46b9a4     3
 5 46b9a4     4
 6 641fa1     0
 7 641fa1     1
 8 641fa1     2
 9 641fa1     3
10 641fa1     4
```

```
> ige
# A tibble: 74 × 3
   ID     visit    IgE
   <chr>  <dbl>  <dbl>
 1 46b9a4     0 100
 2 46b9a4     4 100
 3 641fa1     0   9.01
 4 641fa1     4   0.87
 5 97bab3     0   2.97
 6 97bab3     4   3.7
 7 d85d4f     0   0.05
 8 d85d4f     4   0.05
 9 1b06cf     0  91.2
10 336ddf     0   0.05
```

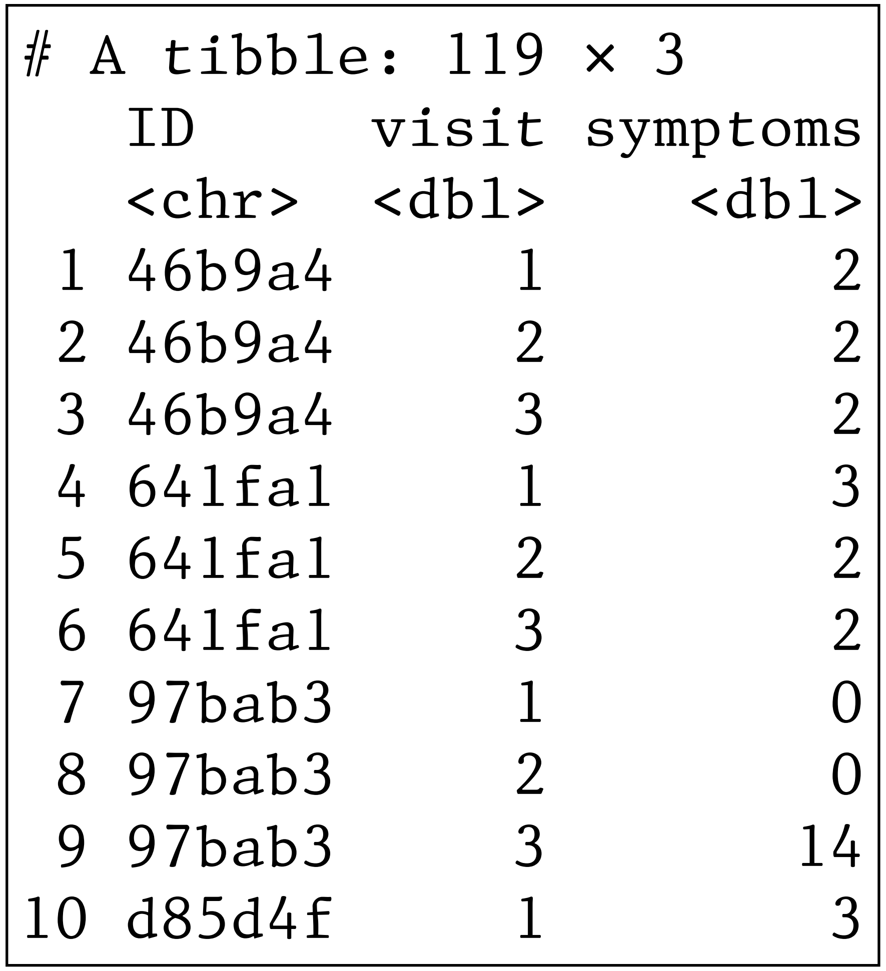
Anti Join

```
symptoms |>
    anti_join(ige, by = c("ID", "visit"))
```



```
> symptoms
# A tibble: 193 × 3
   ID     visit symptoms
   <chr>  <dbl>    <dbl>
 1 46b9a4     0
 2 46b9a4     1
 3 46b9a4     2
 4 46b9a4     3
 5 46b9a4     4
 6 641fa1     0
 7 641fa1     1
 8 641fa1     2
 9 641fa1     3
10 641fa1     4
```

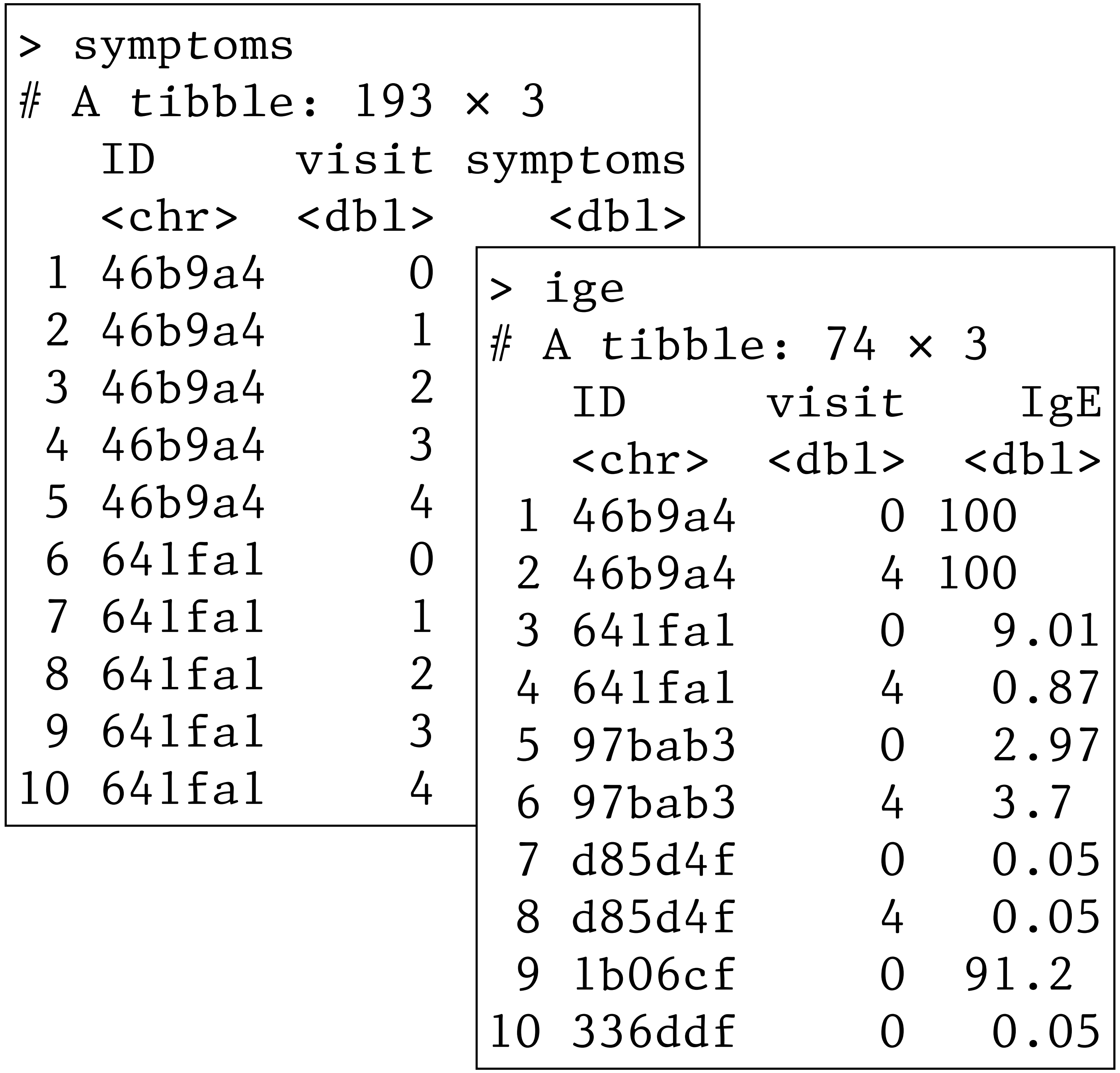
```
> ige
# A tibble: 74 × 3
   ID     visit    IgE
   <chr>  <dbl>  <dbl>
 1 46b9a4     0 100   
 2 46b9a4     4 100   
 3 641fa1     0   9.01
 4 641fa1     4   0.87
 5 97bab3     0   2.97
 6 97bab3     4   3.7 
 7 d85d4f     0   0.05
 8 d85d4f     4   0.05
 9 1b06cf     0  91.2 
10 336ddf     0   0.05
```



```
# A tibble: 119 × 3
   ID     visit symptoms
   <chr>  <dbl>    <dbl>
 1 46b9a4     1        2
 2 46b9a4     2        2
 3 46b9a4     3        2
 4 641fa1     1        3
 5 641fa1     2        2
 6 641fa1     3        2
 7 97bab3     1        0
 8 97bab3     2        0
 9 97bab3     3       14
10 d85d4f     1        3
```

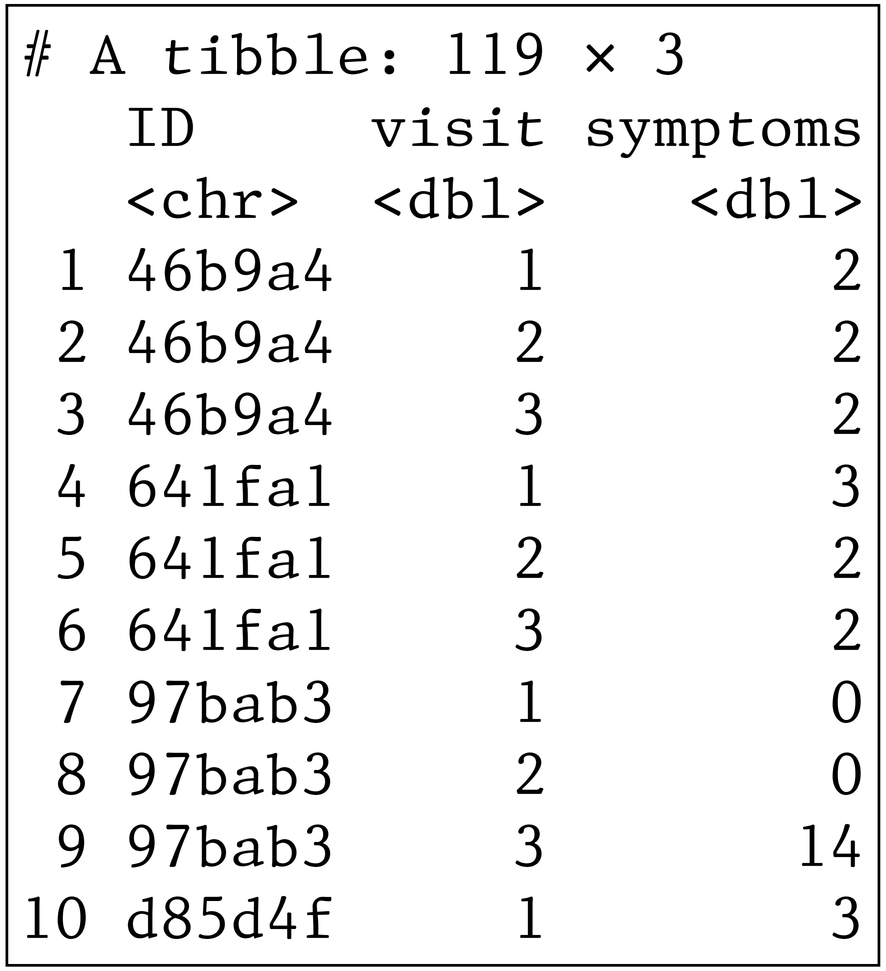
Anti Join

```
symptoms |>
    anti_join(ige, by = c("ID", "visit"))
```



```
> symptoms
# A tibble: 193 × 3
   ID     visit symptoms
   <chr>  <dbl>    <dbl>
 1 46b9a4     0
 2 46b9a4     1
 3 46b9a4     2
 4 46b9a4     3
 5 46b9a4     4
 6 641fa1     0
 7 641fa1     1
 8 641fa1     2
 9 641fa1     3
10 641fa1     4
```

```
> ige
# A tibble: 74 × 3
   ID     visit    IgE
   <chr>  <dbl>  <dbl>
 1 46b9a4     0 100
 2 46b9a4     4 100
 3 641fa1     0   9.01
 4 641fa1     4   0.87
 5 97bab3     0   2.97
 6 97bab3     4   3.7
 7 d85d4f     0   0.05
 8 d85d4f     4   0.05
 9 1b06cf     0  91.2
10 336ddf     0   0.05
```



```
# A tibble: 119 × 3
   ID     visit symptoms
   <chr>  <dbl>    <dbl>
 1 46b9a4     1        2
 2 46b9a4     2        2
 3 46b9a4     3        2
 4 641fa1     1        3
 5 641fa1     2        2
 6 641fa1     3        2
 7 97bab3     1        0
 8 97bab3     2        0
 9 97bab3     3       14
10 d85d4f     1        3
```

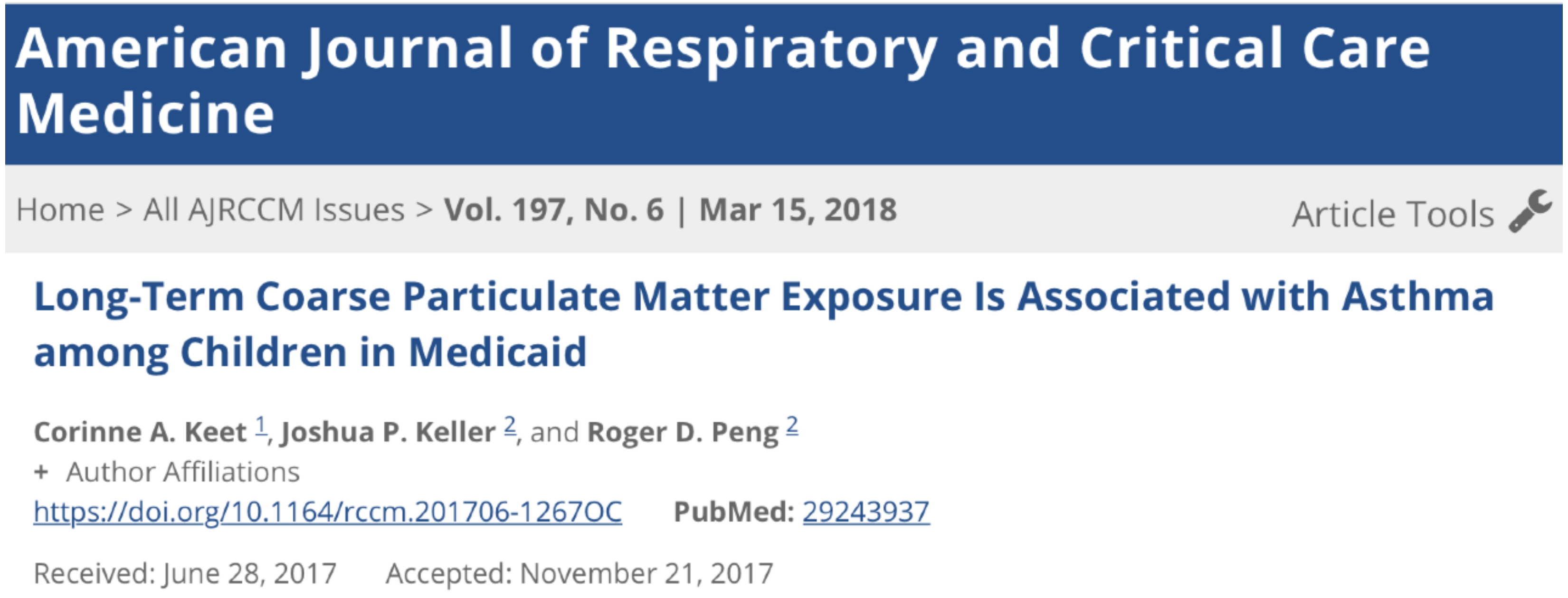
Why not just filter()?

An abstraction of the filtering process

Join Summary

- Datasets store different types of information that can be joined together to be even more informative
- **left_join()** and **inner_join()** are probably the most common forms of joining datasets in data analysis
- For all joins a key column (or columns) must be identified that serves as the connection between two datasets
- Joining in R is analogous to relational database operations (w/SQL statements) where information is stored in separate tables

Example: Asthma and Air Pollution in Medicaid



American Journal of Respiratory and Critical Care Medicine

Home > All AJRCCM Issues > **Vol. 197, No. 6 | Mar 15, 2018**

Article Tools

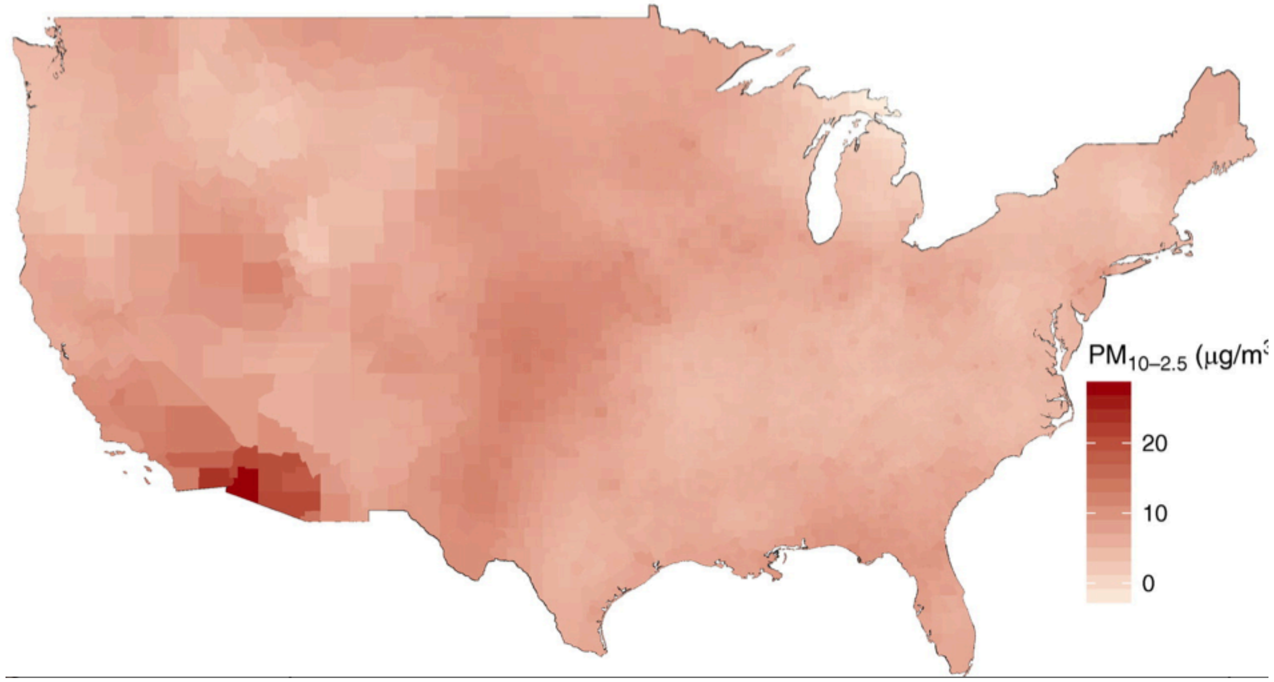
Long-Term Coarse Particulate Matter Exposure Is Associated with Asthma among Children in Medicaid

Corinne A. Keet [1], **Joshua P. Keller** [2], and **Roger D. Peng** [2]

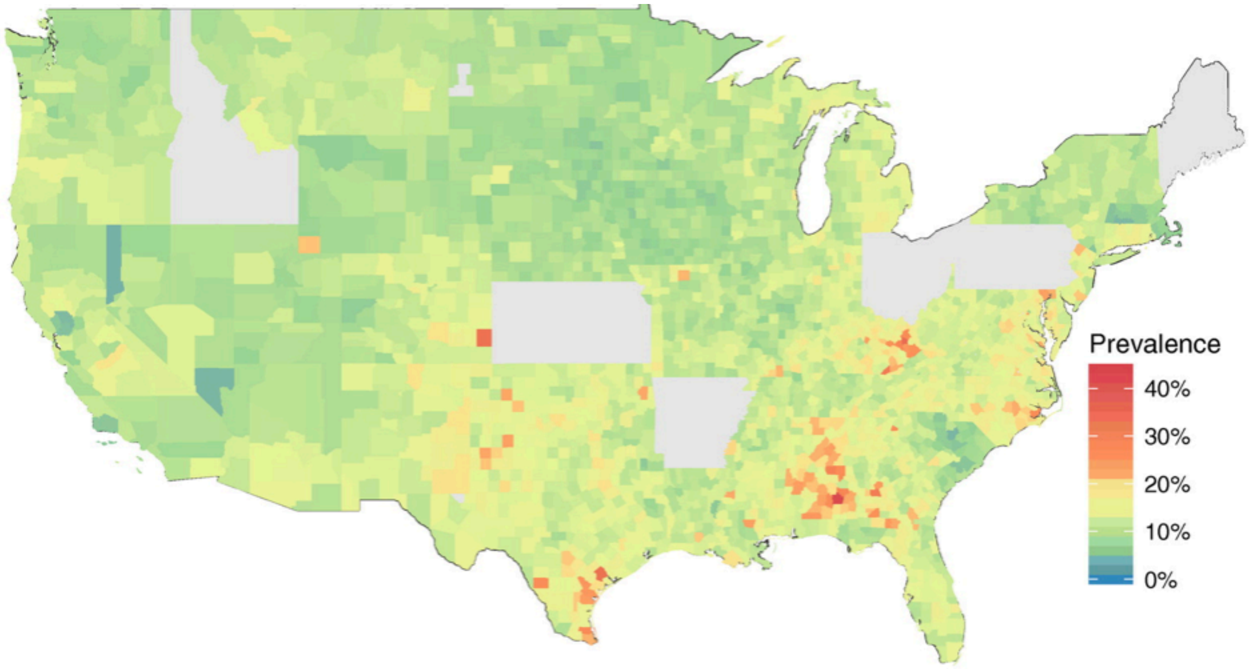
+ Author Affiliations

https://doi.org/10.1164/rccm.201706-1267OC **PubMed:** 29243937

Received: June 28, 2017 Accepted: November 21, 2017



Asthma Prevalence



Example: Asthma and Air Pollution in Medicaid

- Are outdoor coarse particulate matter concentrations associated with the rate of asthma hospitalizations, emergency room visits, or doctor's visits?
- Hospitalization, ER, doctor's visits from Medicaid claims
- Particulate matter data from EPA national monitoring network

Joining Data

Medicaid Personnel File

ID	ZIP	DOB	Months
11111108E1G0EwD	79502	20000424	12
111111018Gmuw0u	78227	20041108	5
1111110GDDww0Dg	76028	20070827	12
1111110Dm11EgGu	75212	19900906	12
1111118DuwE8DEG	78570	20101201	12
1111118um0D8EDE	76306	20111219	12
11111100gu1gGGm	75040	19971125	12
1111110wmE8ugmG	75167	19960808	12
1111110GEwDmuw1	78539	19961209	8
1111118w1DE1gDu	77084	20101127	10
1111118G0GDm0gE	77803	19940716	1
1111110D8gwmGgg	78251	19970816	12
111111wEgu1wgwg	78593	20011127	9
111111wEGw8GGww	75068	19920922	3
1111118G0D1E100	76164	20120611	7

4,400,000 rows

NOTE: Medicaid data is simulated

Joining Data

Medicaid Personnel File

ID	ZIP	DOB	Months
11111108E1G0EwD	79502	20000424	12
111111018Gmuw0u	78227	20041108	5
1111110GDDww0Dg	76028	20070827	12
1111110Dm11EgGu	75212	19900906	12
1111118DuwE8DEG	78570	20101201	12
1111118um0D8EDE	76306	20111219	12
11111100gu1gGGm	75040	19971125	12
1111110wmE8ugmG	75167	19960808	12
1111110GEwDmuw1	78539	19961209	8
1111118w1DE1gDu	77084	20101127	10
1111118G0GDm0gE	77803	19940716	1
1111110D8gwmGgg	78251	19970816	12
111111wEgu1wgwg	78593	20011127	9
111111wEGw8GGww	75068	19920922	3
1111118G0D1E100	76164	20120611	7

4,400,000 rows

Medicaid Hospitalizations

ID	Date	ICD9
1111118uGEE80mG	20120221	46611
1111118118GDm1E	20120511	380
11111108uu0E8EE	20120917	78900
1111118G0DgD1ug	20120517	V3001
111111wEEg01mw1	20120517	29633
1111110Gm8mDGuD	20120904	650
1111118um0gG1GG	20120110	V3000
1111110GGm8g0w0	20120916	650

62,000 rows

NOTE: Medicaid data is simulated

Joining Data

Medicaid Personnel File

ID	ZIP	DOB	Months
11111108E1G0EwD	79502	20000424	12
111111018Gmuw0u	78227	20041108	5
1111110GDDww0Dg	76028	20070827	12
1111110Dm11EgGu	75212	19900906	12
1111118DuwE8DEG	78570	20101201	12
1111118um0D8EDE	76306	20111219	12
11111100gu1gGGm	75040	19971125	12
1111110wmE8ugmG	75167	19960808	12
1111110GEwDmuw1	78539	19961209	8
1111118w1DE1gDu	77084	20101127	10
1111118G0GDm0gE	77803	19940716	1
1111110D8gwmGgg	78251	19970816	12
111111wEgu1wgwg	78593	20011127	9
111111wEGw8GGww	75068	19920922	3
1111118G0D1E100	76164	20120611	7

4,400,000 rows

Medicaid Hospitalizations

ID	Date	ICD9
1111118uGEE80mG	20120221	46611
1111118118GDm1E	20120511	380
11111108uu0E8EE	20120917	78900
1111118G0DgD1ug	20120517	V3001
111111wEEg01mw1	20120517	29633
1111110Gm8mDGuD	20120904	650
1111118um0gG1GG	20120110	V3000
1111110GGm8g0w0	20120916	650

62,000 rows

Particulate Matter

ZIP	Date	value
<chr>	<chr>	<dbl>
76306	20120113	32.4
75040	20120117	2.1
75167	20120307	24.6
78539	20120330	8.9
77084	20120413	5.9
77803	20120415	15.9
78251	20120508	11.1
78593	20120617	4.8
75068	20120803	8.7
76164	20121231	15.8

NOTE: Medicaid data is simulated

Joining Data

Medicaid Personnel File

ID	ZIP	DOB	Months
11111108E1G0EwD	79502	20000424	12
111111018Gmuw0u	78227	20041108	5
1111110GDDww0Dg	76028	20070827	12
1111110Dm11EgGu	75212	19900906	12
1111118DuwE8DEG	78570	20101201	12
1111118um0D8EDE	76306	20111219	12
11111100gu1gGGm	75040	19971125	12
1111110wmE8ugmG	75167	19960808	12
1111110GEwDmuw1	78539	19961209	8
1111118w1DE1gDu	77084	20101127	10
1111118G0GDm0gE	77803	19940716	1
1111110D8gwmGgg	78251	19970816	12
111111wEgu1wgwg	78593	20011127	9
111111wEGw8GGww	75068	19920922	3
1111118G0D1E100	76164	20120611	7

4,400,000 rows

Medicaid Hospitalizations

ID	Date	ICD9
1111118uGEE80mG	20120221	46611
1111118118GDm1E	20120511	380
11111108uu0E8EE	20120917	78900
1111118G0DgD1ug	20120517	V3001
111111wEEg01mw1	20120517	29633
1111110Gm8mDGuD	20120904	650
1111118um0gG1GG	20120110	V3000
1111110GGm8g0w0	20120916	650

62,000 rows

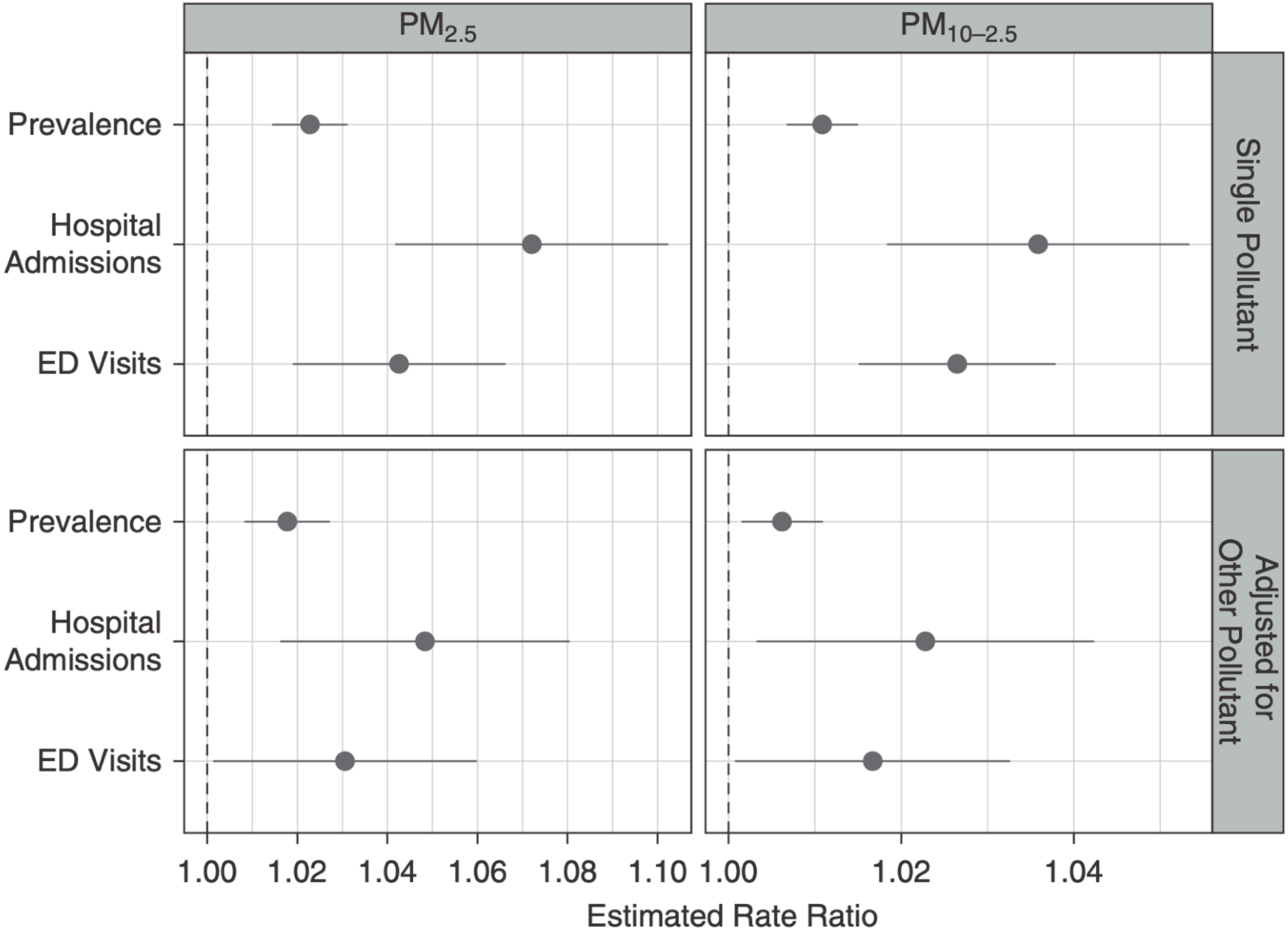
Particulate Matter

ZIP	Date	value
<chr>	<chr>	<dbl>
76306	20120113	32.4
75040	20120117	2.1
75167	20120307	24.6
78539	20120330	8.9
77084	20120413	5.9
77803	20120415	15.9
78251	20120508	11.1
78593	20120617	4.8
75068	20120803	8.7
76164	20121231	15.8

Additional tables for weather, Census data...

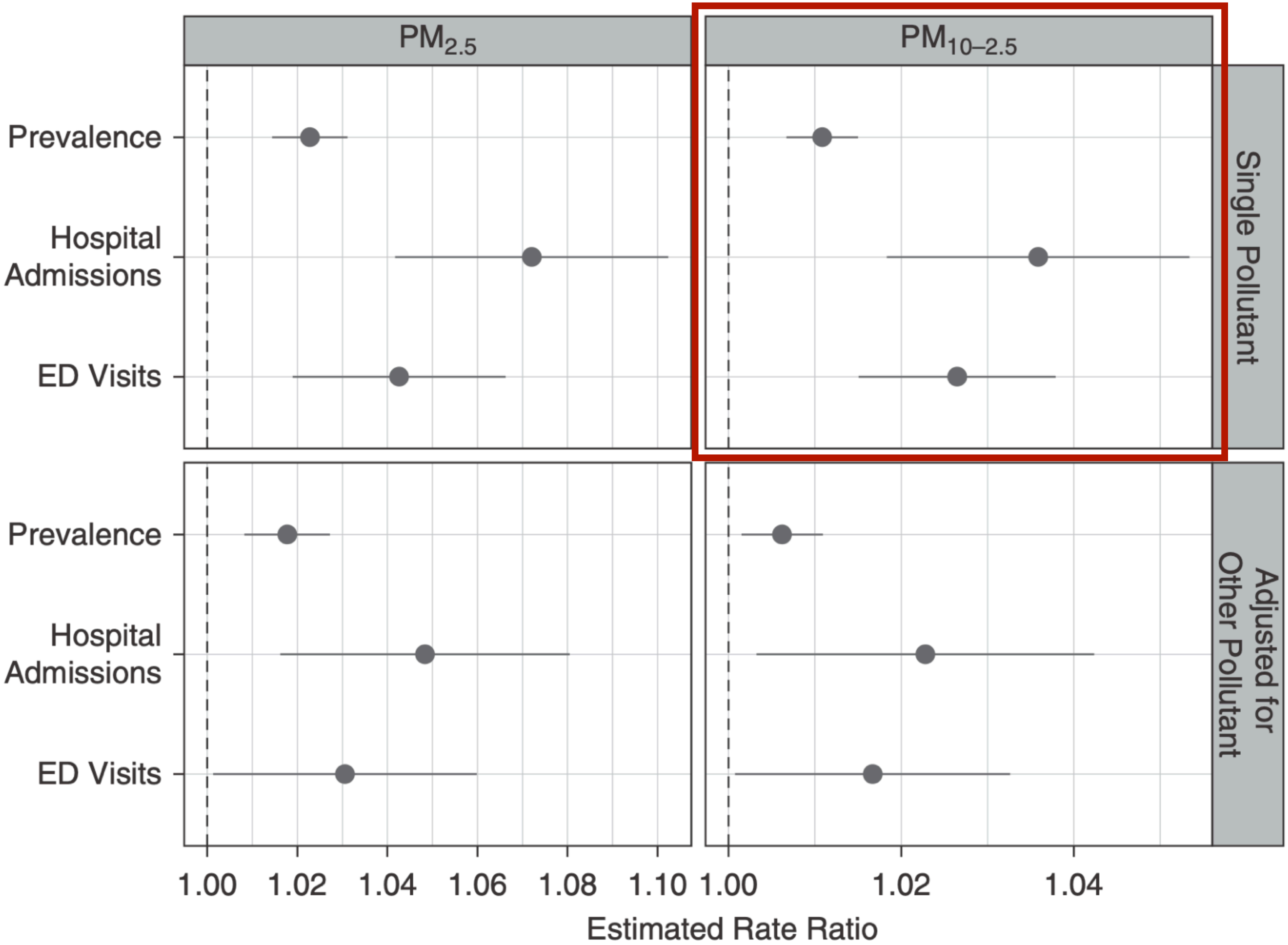
NOTE: Medicaid data is simulated

Results



$PM_{2.5}$
$PM_{10-2.5}$
Single Pollutant
Adjusted for Other Pollutant
Prevalence
Hospital Admissions
ED Visits
Prevalence
Hospital Admissions
ED Visits
1.00
1.02
1.04
1.06
1.08
1.10
1.00
1.02
1.04
Estimated Rate Ratio

Results


PM$_{2.5}$
PM$_{10-2.5}$
Single Pollutant
Adjusted for Other Pollutant
Prevalence
Hospital Admissions
ED Visits
Prevalence
Hospital Admissions
ED Visits
1.00
1.02
1.04
1.06
1.08
1.10
1.00
1.02
1.04
Estimated Rate Ratio

The Road So Far

- Tidy Data via the Tidyverse
  - One observation per row; Each column represents a variable or measure or characteristic; Every row / column combination is a single value
- Reading data (readr) - read_csv, read_tsv
- Plotting and Visualization with ggplot2
- Data Wrangling (dplyr, tidyr)
  - Data transformation - select, filter, arrange, rename, mutate group_by, summarize
  - Pivoting functions - pivot_longer, pivot_wider
  - Joining functions - left_join, inner_join, right_join, full_join, anti_join