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

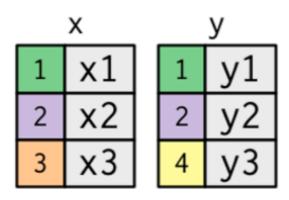
R Functions for Joins

- 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.

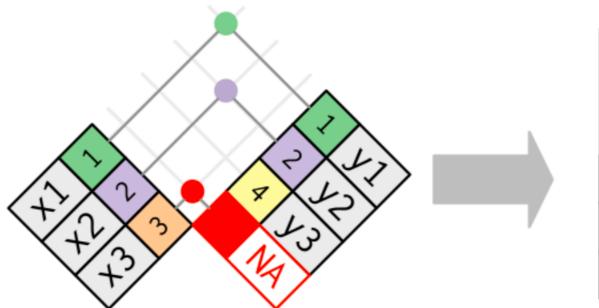
R Functions for Joins

- 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

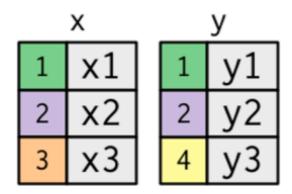


key <dbl></dbl>	val_x <chr></chr>	key <dbl></dbl>	val_y <chr></chr>
1	x1	1	y1
2	x2	2	y2
3	х3	4	у3

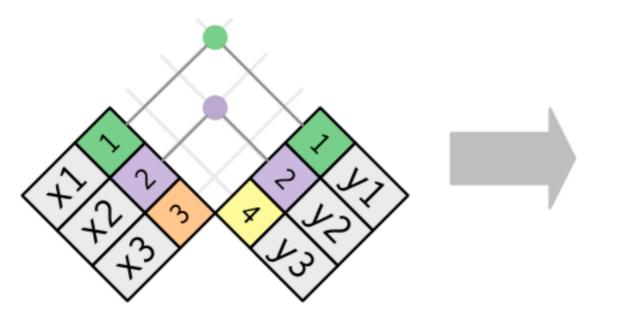


key	val_x	val_y
1	x1	у1
2	x2	y2
3	х3	NA

Inner Join

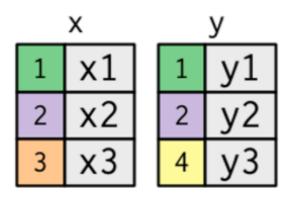


key <dbl></dbl>	val_x <chr></chr>	key <dbl></dbl>	val_y <chr></chr>
1	x1	1	y1
2	x2	2	y2
3	х3	4	у3

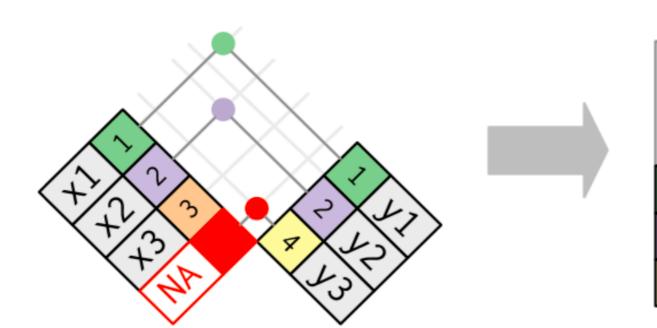


key	val_x	val_y
1	x1	у1
2	x2	y2

Right Join



key <dbl></dbl>	val_x <chr></chr>	key <dbl></dbl>	val_y <chr></chr>
1	x1	1	y1
2	x2	2	y2
3	x3	4	у3



key	val_x	val_y
1	x1	у1
2	x2	y2
4	NA	у3

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

```
> dat
# A tibble: 193 \times 6
  ID
         visit symptoms IgE catpos hometype
  <chr> <dbl>
                  <db1> <db1> <db1>
                                         < db1 >
1 46b9a4
                                             3
                      0 100
2 46b9a4
                         NA
                                            NA
3 46b9a4
                         NA
                                            NA
4 46b9a4
                         NA
                                            NA
5 46b9a4
                        100
                                            NA
6 641fa1
                     14
                          9.01
                                             3
 7 641fa1
                         NA
                                            NA
8 641fa1
                         NA
                                            NA
9 641fa1
                         NA
                                            NA
10 641fa1
                          0.87
                                            NA
```

```
> dat
# A tibble: 193 \times 6
          visit symptoms
   ID
                             IgE catpos hometype
          <db1>
                    <db1>
                            <db1>
                                   <db1>
                                             <db1>
   <chr>
   46b9a4
                         0 100
   46b9a4
                            NA
                                                NA
   46b9a4
                            NA
                                                NA
   46b9a4
                            NA
                                                NA
   46b9a4
                          100
                                                NA
   641fa1
                        14
                             9.01
                                                  3
   641fa1
                            NA
                                                NA
   641fa1
                            NA
                                                NA
   641fa1
                            NA
                                                NA
   641fa1
                             0.87
                                                NA
```

Primary Key

```
> dat
# A tibble: 193 \times 6
           visit symptoms
   ID
                                IgE catpos hometype
           <db1>
                     <db1>
                             <db1>
                                     <db1>
                                                < db1 >
   <chr>
   46b9a4
                            100
                          0
   46b9a4
                             NA
                                                   NA
   46b9a4
                             NA
                                                   NA
   46b9a4
                             NA
                                                   NA
   46b9a4
                            100
                                                   NA
   641fa1
                              9.01
                                                    3
                         14
   641fa1
                             NA
                                                   NA
   641fa1
                             NA
                                                   NA
   641fa1
                             NA
                                                   NA
   641fa1
                              0.87
                                                   NA
```

Primary Key

Change at every visit

```
> dat
# A tibble: 193 \times 6
           visit symptoms
                                IgE catpos hometype
   ID
                      <db1>
           <db1>
                              <db1>
                                      <db1>
                                                < db1 >
   <chr>
                             100
   46b9a4
                          0
   46b9a4
                              NA
                                                    NA
   46b9a4
                              NA
                                                    NA
   46b9a4
                              NA
                                                    NA
   46b9a4
                             100
                                                    NA
                               9.01
   641fa1
                         14
                                                     3
   641fa1
                              NA
                                                    NA
   641fa1
                              NA
                                                    NA
   641fa1
                              NA
                                                    NA
   641fa1
                               0.87
                                                    NA
   Primary
```

Change at

every visit

Key

Subset of

visits

```
> dat
# A tibble: 193 \times 6
                                     catpos hometype
           visit symptoms
                                IgE
   ID
           <db1>
                      <db1>
                              <db1>
                                      <db1>
                                                 <db1>
   <chr>
                             100
   46b9a4
                          0
   46b9a4
                              NA
                                                    NA
   46b9a4
                                                    NA
                              NA
   46b9a4
                                                    NA
                              NA
   46b9a4
                             100
                                                    NA
                               9.01
   641fa1
                         14
   641fa1
                              NA
                                                    NA
                2
                          2
   641fa1
                              NA
                                                    NA
   641fa1
                              NA
                                                    NA
   641fa1
                               0.87
                                           0
                                                    NA
```

Change at

every visit

Subset of

visits

Only collected

at baseline

Primary

Key

```
> subject
# A tibble: 40 \times 3
          catpos hometype
   ID
            <db1>
                      <db1>
   <chr>
                          3
 1 46b9a4
 2 641fa1
 3 97bab3
 4 d85d4f
                          5
 5 1b06cf
 6 336ddf
 7 192e91
               NA
 8 d6ecde
                0
 9 7bf734
10 ba54c0
               NA
```

```
> subject
# A tibble: 40 \times 3
           catpos hometype
   ID
            < db1 >
                      <db1>
   <chr>
                           3
 1 46b9a4
 2 641fa1
 3 97bab3
 4 d85d4f
 5 1b06cf
 6 336ddf
 7 192e91
               NA
 8 d6ecde
                0
 9 7bf734
10 ba54c0
               NA
```

> ige				
# A	A tibble	e: 74 >	< 3	
	ID	visit	IgE	
	<chr></chr>	<db1></db1>	<db1></db1>	
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	

```
> subject
# A tibble: 40 \times 3
           catpos hometype
   ID
            < db1 >
   <chr>
                      <db1>
 1 46b9a4
                           3
                 0
 2 641fa1
   97bab3
 4 d85d4f
 5 1b06cf
                 0
 6 336ddf
                 0
 7 192e91
               NA
 8 d6ecde
                 0
 9 7bf734
                           3
10 ba54c0
               NA
```

```
> ige
# A tibble: 74 \times 3
   TD
          visit
                     IgE
          <db1>
                  <db1>
   <chr>
 1 46b9a4
               0 100
 2 46b9a4
               4 100
 3 641fa1
                    9.01
                   0.87
 4 641fa1
 5 97bab3
                   2.97
                   3.7
 6 97bab3
 7 d85d4f
                   0.05
 8 d85d4f
                   0.05
  1b06cf
                  91.2
10 336ddf
                    0.05
```

```
symptoms
# A tibble: 193 \times 3
           visit symptoms
   ID
          <db1>
                     <db1>
   <chr>
 1 46b9a4
                          0
 2 46b9a4
  46b9a4
 4 46b9a4
                          0
 5 46b9a4
 6 641fa1
                        14
 7 641fa1
                          3
 8 641fa1
  641fa1
10 641fa1
```

```
> subject
# A tibble: 40 \times 3
   ID
           catpos hometype
            < db1 >
   <chr>
                      <db1>
 1 46b9a4
                           3
                 0
 2 641fa1
 3 97bab3
 4 d85d4f
 5 1b06cf
                 0
 6 336ddf
 7 192e91
               NA
 8 d6ecde
                 0
 9 7bf734
10 ba54c0
               NA
```

```
> ige
# A tibble: 74 \times 3
   TD
          visit
                     IgE
          <db1>
                  <db1>
   <chr>
 1 46b9a4
               0 100
 2 46b9a4
               4 100
 3 641fa1
                   9.01
                   0.87
 4 641fa1
  97bab3
                   2.97
                   3.7
 6 97bab3
 7 d85d4f
                   0.05
 8 d85d4f
                   0.05
  1b06cf
                  91.2
10 336ddf
                   0.05
```

```
> symptoms
# A tibble: 193 \times 3
   ID
          visit symptoms
          <db1>
                     <db1>
   <chr>
 1 46b9a4
                         0
 2 46b9a4
  46b9a4
 4 46b9a4
 5 46b9a4
                         0
 6 641fa1
                        14
 7 641fa1
 8 641fa1
  641fa1
10 641fa1
```

```
> subject
# A tibble: 40 x 3
   ID
           catpos hometype
            < db1 >
                      <db1>
   cobr>
 1 46b9a4
                           3
                0
 2 641fa1
   97bab3
 4 d85d4f
  1b06cf
                0
 6 336ddf
 7 192e91
               NA
 8 d6ecde
                0
 9 7bf734
                           3
10 ba54c0
               NA
```

```
> ige
# A tibble: 74
          visit
   TD
                    IgE
          /dh1>
                  <db1>
               0 100
 1 46b9a4
 2 46b9a4
               4 100
   641fa1
                   9.01
                   0.87
 4 641fa1
  97bab3
                   2.97
                   3.7
 6 97bab3
 7 d85d4f
                   0.05
  d85d4f
                   0.05
  1b06cf
                  91.2
10 336ddf
                   0.05
```

```
symptoms
 A tibble: 193 x 3
   TD
          visit symptoms
          /dh1>
                    <db1>
 1 46b9a4
                         0
 2 46b9a4
  46b9a4
4 46b9a4
5 46b9a4
                         0
6 641fa1
                        14
 7 641fa1
  641fa1
  641fa1
10 641fa1
```

```
> subject
# A tibble: 40 \times 3
   TD
           catpos hometype
            < db1 >
                      /dh1\
 1 46b9a4
                           3
                 0
 2 641fa1
   97bab3
 4 d85d4f
  1b06cf
 6 336ddf
 7 192e91
               NA
 8 d6ecde
                 0
 9 7bf734
10 ba54c0
               NA
```

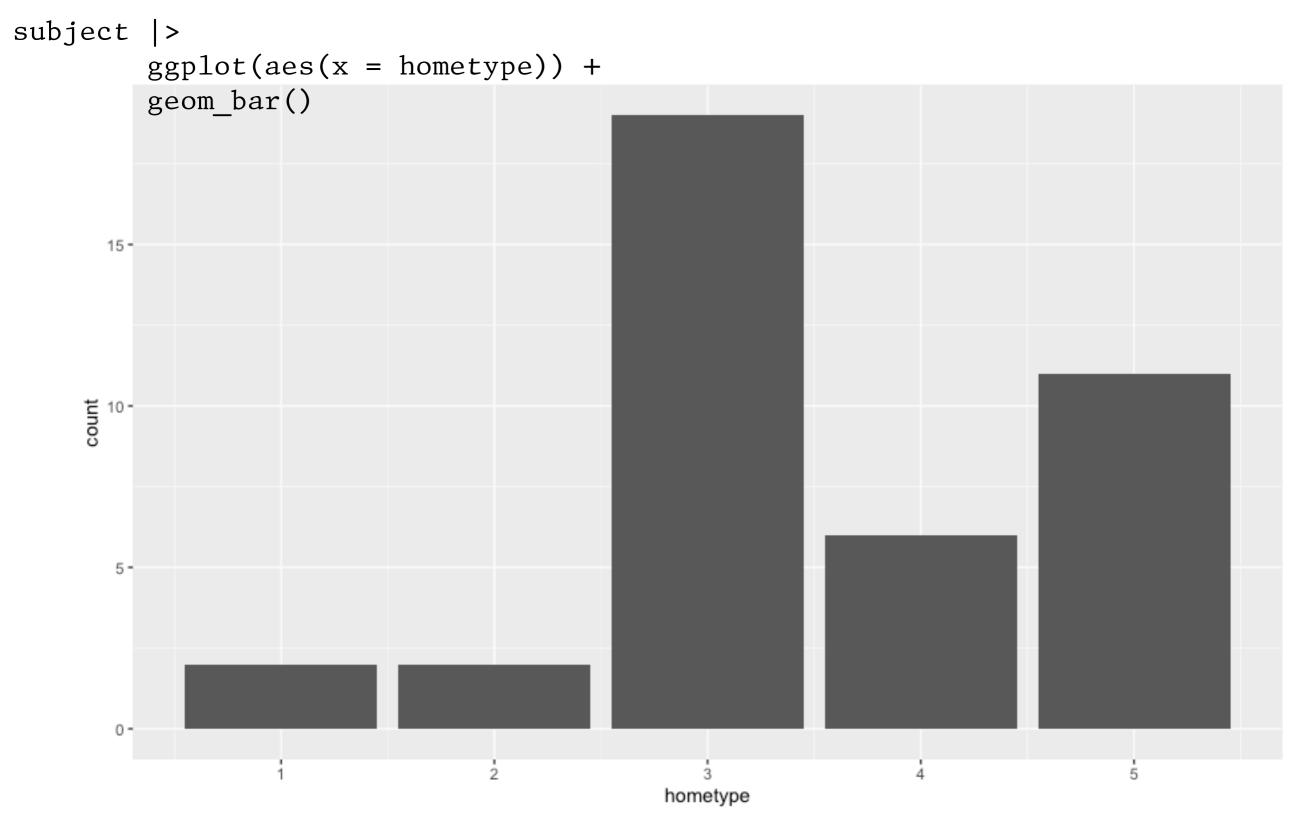
```
> ige
# A tibble: 74
          visit
   TD
                    IgE
           /dh1\
                  <db1>
               0 100
 1 46b9a4
 2 46b9a4
               4 100
   641fa1
                   9.01
 4 641fa1
                   0.87
  97bab3
                   2.97
                   3.7
 6 97bab3
 7 d85d4f
                   0.05
  d85d4f
                   0.05
  1b06cf
                  91.2
10 336ddf
                   0.05
```

```
symptoms
 A tibble: 193 × 3
          visit symptoms
   TD
          /dh1>
                    <db1>
 1 46b9a4
                         0
 2 46b9a4
  46b9a4
4 46b9a4
5 46b9a4
                         0
6 641fa1
                        14
 7 641fa1
  641fa1
  641fa1
10 641fa1
```

Home Types

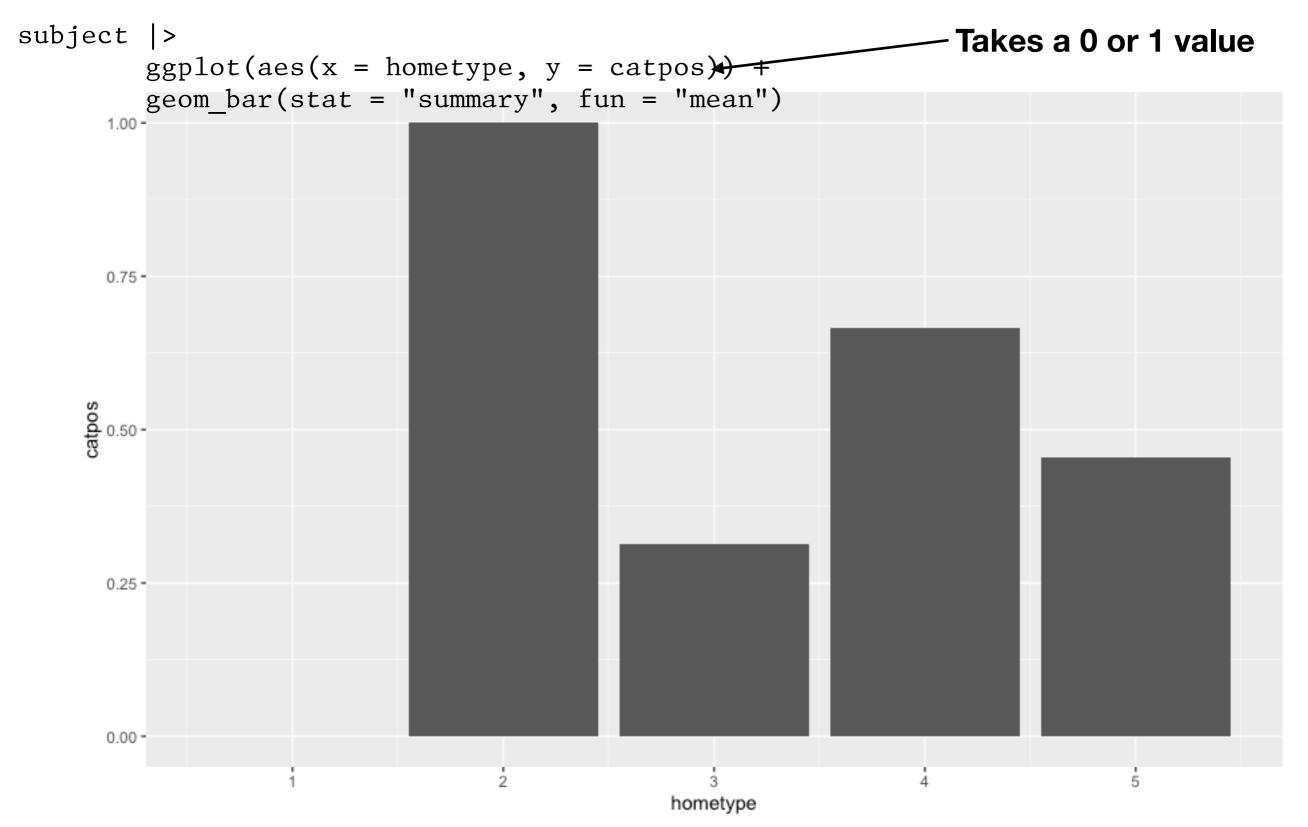
```
subject |>
    ggplot(aes(x = hometype)) +
    geom_bar()
```

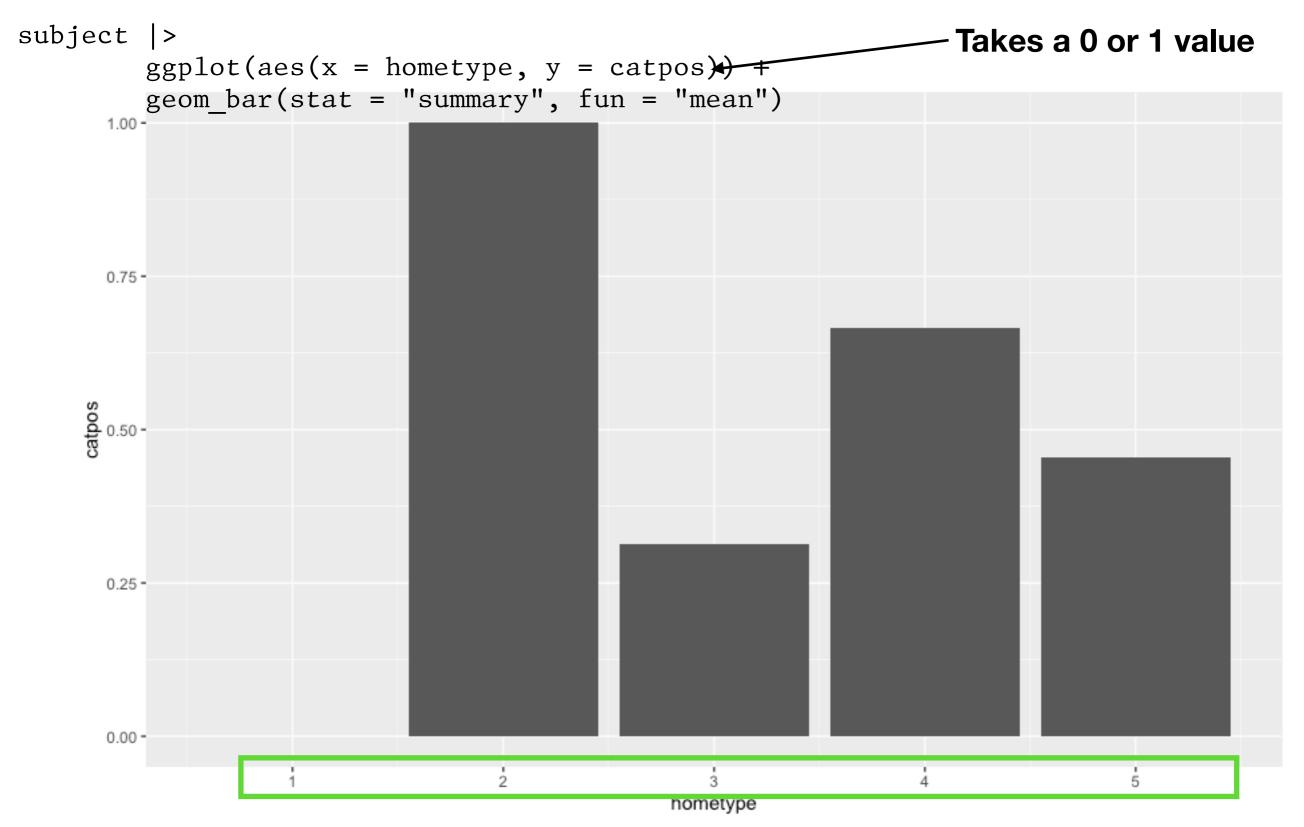
Home Types



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

```
subject |>
    ggplot(aes(x = hometype, y = catpos)+) +
    geom_bar(stat = "summary", fun = "mean")
Takes a 0 or 1 value
```





```
> subject
# A tibble: 40 \times 3
           catpos hometype
   ID
            < db1 >
                      <db1>
   <chr>
 1 46b9a4
                           3
 2 641fa1
 3 97bab3
 4 d85d4f
 5 1b06cf
 6 336ddf
 7 192e91
               NA
                           3
 8 d6ecde
                0
 9 7bf734
10 ba54c0
               NA
```

```
> subject
# A tibble: 40 \times 3
          catpos hometype
   ID
           <db1>
   <chr>
                     <db1>
1 46b9a4
                          3
 2 641fa1
 3 97bab3
 4 d85d4f
 5 1b06cf
 6 336ddf
7 192e91
               NA
8 d6ecde
                0
 9 7bf734
10 ba54c0
               NA
```

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

```
> subject
# A tibble: 40 \times 3
          catpos hometype
   ID
           <db1>
   <chr>
                     <db1>
 1 46b9a4
                          3
 2 641fa1
 3 97bab3
 4 d85d4f
 5 1b06cf
 6 336ddf
 7 192e91
               NA
8 d6ecde
 9 7bf734
10 ba54c0
               NA
```

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

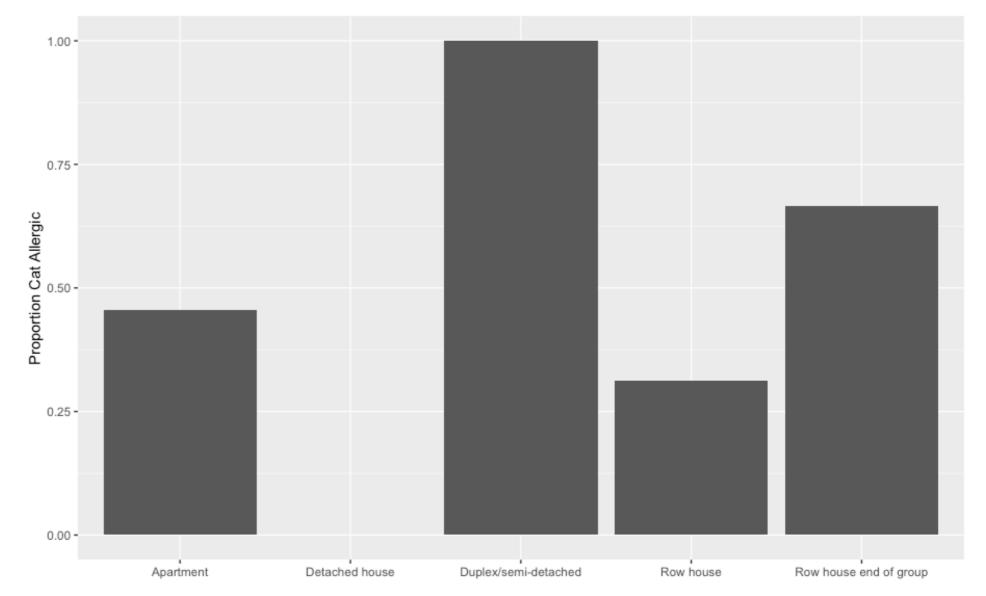
Column that both tables have in common (must have same name)

```
> subject
# A tibble: 40 \times 3
          catpos hometype
   ID
            <db1>
                      <db1>
   <chr>
 1 46b9a4
                          3
 2 641fa1
 3 97bab3
 4 d85d4f
 5 1b06cf
 6 336ddf
 7 192e91
               NA
 8 d6ecde
                0
 9 7bf734
10 ba54c0
               NA
```

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

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

```
\# A tibble: 40 x 4
   ID
          catpos hometype label
   <chr>
           <db1>
                     <db1> <chr>
 1 46b9a4
                         3 Row house
 2 641fa1
                         3 Row house
 3 97bab3
                         3 Row house
 4 d85d4f
                         3 Row house
 5 1b06cf
                         5 Apartment
 6 336ddf
                         4 Row house end of group
 7 192e91
                         3 Row house
              NA
 8 d6ecde
                         3 Row house
 9 7bf734
                         3 Row house
10 ba54c0
              NA
                         3 Row house
```



```
subject |>
    left join(housing, by = "hometype") |>
    ggplot(aes(x = label, y = catpos)) +
    geom_bar(stat = "summary", fun = "mean") +
                                                                                   Remove x-axis label
    labs(x = NULL, \leftarrow
          y = "Proportion Cat Allergic")
                                   1.00 -
                                   0.75 -
                                 Proportion Cat Allergic
                                   0.25 -
      Labels in
                                   0.00 -
alphabetical order
```

Detached house

Duplex/semi-detached

Row house

Row house end of group

Housing Type vs. Baseline IgE

```
> subject
# A tibble: 40 \times 3
          catpos hometype
   ID
            <db1>
                      <db1>
   <chr>
 1 46b9a4
 2 641fa1
 3 97bab3
 4 d85d4f
 5 1b06cf
 6 336ddf
 7 192e91
               NA
8 d6ecde
  7bf734
10 ba54c0
               NA
```

```
> ige
# A tibble: 74 \times 3
   ID
          visit
                    IgE
          <db1>
                  <db1>
   <chr>
 1 46b9a4
               0 100
 2 46b9a4
               4 100
 3 641fa1
                   9.01
               4 0.87
 4 641fa1
                   2.97
 5 97bab3
                   3.7
 6 97bab3
 7 d85d4f
               0 0.05
 8 d85d4f
               4 0.05
  1b06cf
                  91.2
10 336ddf
                   0.05
```

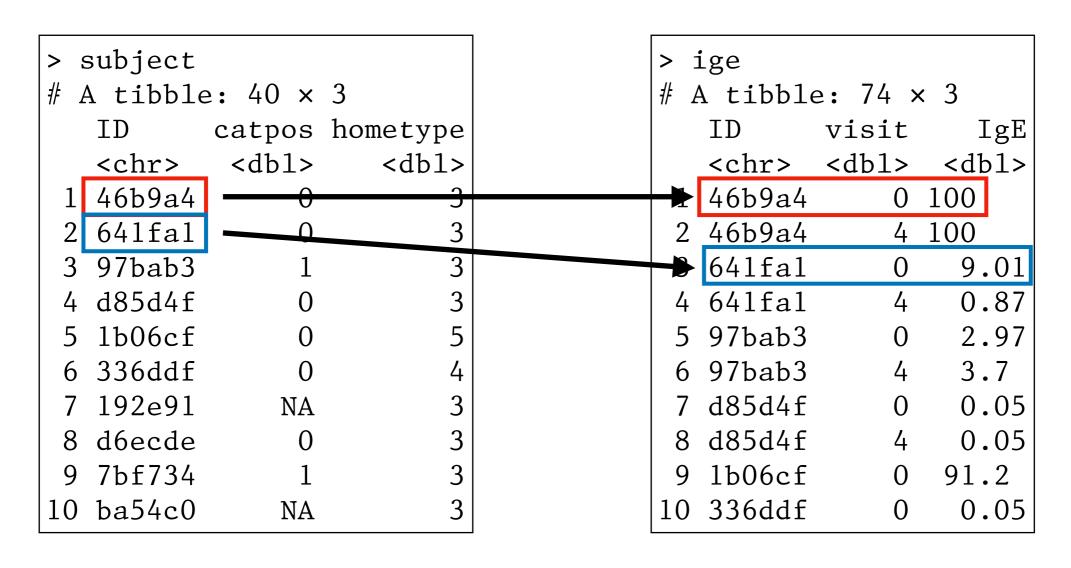
Housing Type vs. Baseline IgE

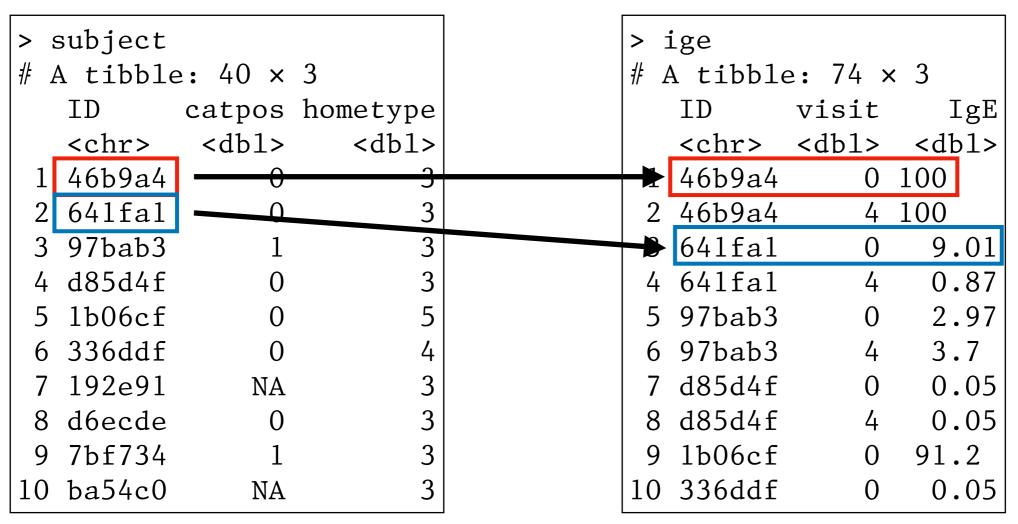
```
> subject
# A tibble: 40 \times 3
          catpos hometype
   ID
            <db1>
                      <db1>
   <chr>
 1 46b9a4
 2 641fa1
  97bab3
 4 d85d4f
  1b06cf
 6 336ddf
 7 192e91
               NA
8 d6ecde
  7bf734
10 ba54c0
               NA
```

```
> ige
# A tibble: 74 \times 3
   ID
          visit
                    IgE
   <chr>
          <db1>
                  <db1>
 1 46b9a4
              0 100
 2 46b9a4
               4 100
 3 641fa1
                   9.01
               4 0.87
 4 641fa1
                   2.97
  97bab3
                   3.7
 6 97bab3
 7 d85d4f
                0.05
 8 d85d4f
               4 0.05
  1b06cf
                  91.2
10 336ddf
                   0.05
```

> subject				> ige				
# A tibble: 40 × 3					# A tibble: 74 × 3			
	ID	catpos	hometype			ID	visit	IgE
Ι.	<chr></chr>	<db1></db1>	<db1></db1>			<chr></chr>	<db1></db1>	<db1></db1>
1	46b9a4	0	3		→	46b9a4	0	100
2	641fa1	0	3		2	46b9a4	4	100
3	97bab3	1	3		3	641fa1	0	9.01
4	d85d4f	0	3		4	641fa1	4	0.87
5	1b06cf	0	5		5	97bab3	0	2.97
6	336ddf	0	4		6	97bab3	4	3.7
7	192e91	NA	3		7	d85d4f	0	0.05
8	d6ecde	0	3		8	d85d4f	4	0.05
9	7bf734	1	3		9	1b06cf	0	91.2
10	ba54c0	NA	3		10	336ddf	0	0.05

> subject				> ige				
# A tibble: 40 × 3				# A tibble: 74 × 3				
	ID	catpos	hometype		ID	visit	IgE	
.	<chr></chr>	<db1></db1>	<db1></db1>		<chr></chr>	<db1></db1>	<db1></db1>	
1	46b9a4	0	3	+	46b9a4	0	100	
2	641fa1	0	3	2	46b9a4	4	100	
3	97bab3	1	3	3	641fal	0	9.01	
4	d85d4f	0	3	4	641fal	4	0.87	
5	1b06cf	0	5	5	97bab3	0	2.97	
6	336ddf	0	4	6	97bab3	4	3.7	
7	192e91	NA	3	7	d85d4f	0	0.05	
8	d6ecde	0	3	8	d85d4f	4	0.05	
9	7bf734	1	3	9	1b06cf	0	91.2	
10	ba54c0	NA	3	10	336ddf	0	0.05	





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

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

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

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

```
> ige
# A tibble: 74 \times 3
          visit
                   IgE
   TD
          <db1>
                <db1>
   <chr>
 1 46b9a4
              0 100
 2 46b9a4
              4 100
 3 641fa1
                9.01
 4 641fa1
              4 0.87
                2.97
 5 97bab3
                3.7
 6 97bab3
 7 d85d4f
                0.05
                  0.05
 8 d85d4f
 9 1b06cf
                91.2
10 336ddf
                  0.05
```

```
# A tibble: 40 \times 2
   ID
              IgE
            <db1>
   <chr>
 1 46b9a4 100
 2 641fa1
           9.01
            2.97
 3 97bab3
            0.05
 4 d85d4f
 5 1b06cf
           91.2
 6 336ddf
            0.05
 7 192e91 100
            6.34
8 d6ecde
 9 7bf734
            8.41
10 ba54c0
             9.87
```

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

```
> ige
# A tibble: 74 \times 3
   TD
          visit
                    IgE
          <db1>
                  <db1>
   <chr>
 1 46b9a4
               0 100
               4 100
 2 46b9a4
 3 641fa1
                   9.01
 4 641fa1
                   0.87
                  2.97
 5 97bab3
                   3.7
 6 97bab3
 7 d85d4f
                   0.05
 8 d85d4f
                   0.05
  1b06cf
                  91.2
10 336ddf
                   0.05
```

```
# A tibble: 40 \times 2
              IgE
   TD
            <db1>
   <chr>
 1 46b9a4 100
 2 641fa1
             9.01
             2.97
 3 97bab3
             0.05
 4 d85d4f
 5 1b06cf
            91.2
 6 336ddf
             0.05
 7 192e91 100
             6.34
 8 d6ecde
 9 7bf734
             8.41
10 ba54c0
             9.87
```

```
> subject
# A tibble: 40 \times 3
           catpos hometype
   ID
            <db1>
   <chr>>
                      <db1>
 1 46b9a4
 2 641fa1
 3 97bab3
 4 d85d4f
 5 1b06cf
 6 336ddf
 7 192e91
               NA
 8 d6ecde
                0
 9 7bf734
10 ba54c0
               NA
```

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

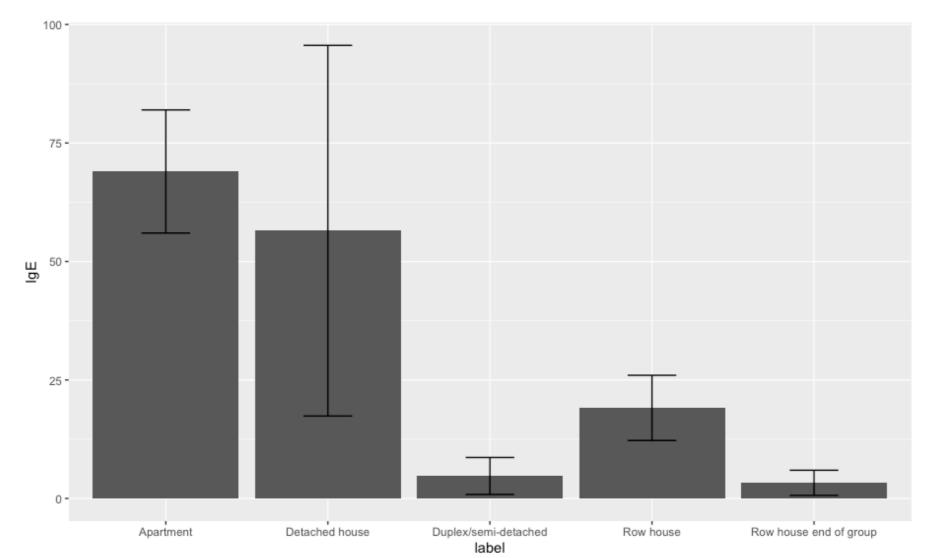
```
# A tibble: 40 \times 5
            IgE catpos hometype label
  <chr> <dbl> <dbl> <dbl> <dbl> <dr>
1 46b9a4 100
                              3 Row house
2 641fa1 9.01
                              3 Row house
3 97bab3 2.97
                              3 Row house
4 d85d4f 0.05
                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

```
ige |>
   filter(visit == 0) |>
   select(-visit) |>
   left_join(subject, by = "ID") |>
   left_join(housing, by = "hometype")
ID column no longer needed
```

```
# A tibble: 40
            IgE catpos hometype label
   TD 🗲
          <db1> <db1> <db1> <chr>
  <chr>
1 46b9a4 100
                              3 Row house
2 641fa1 9.01
                              3 Row house
3 97bab3 2.97
                              3 Row house
4 d85d4f 0.05
                              3 Row house
5 1b06cf 91.2
                              5 Apartment
6 336ddf 0.05
                              4 Row house end of group
 7 192e91 100
                    NA
                              3 Row house
8 d6ecde 6.34
                     0
                              3 Row house
  7bf734 8.41
                              3 Row house
         9.87
10 ba54c0
                              3 Row house
                    NA
```

```
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
        geom_bar(stat = "summary", fun = "mean_se", width = 0.3)
```



```
symptoms
# A tibble: 193 \times 3
          visit symptoms
   ID
         <db1>
                    <db1>
   <chr>
 1 46b9a4
 2 46b9a4
                         2
 3 46b9a4
 4 46b9a4
 5 46b9a4
 6 641fa1
                        14
 7 641fa1
 8 641fa1
 9 641fa1
10 641fa1
```

```
> ige
# A tibble: 74 \times 3
          visit
   TD
                    IgE
          <db1>
                 <db1>
   <chr>
 1 46b9a4
              0 100
 2 46b9a4
              4 100
  641fa1
                  9.01
                  0.87
 4 641fa1
                2.97
  97bab3
 6 97bab3
              4 3.7
  d85d4f
                  0.05
  d85d4f
                0.05
  1b06cf
                 91.2
10 336ddf
                  0.05
```

```
symptoms
# A tibble: 193 \times 3
           visit symptoms
   ID
   <chr> <db1>
                      <db1>
   46b9a4
                           0
   46b9a4
                           2
2
   46b9a4
 4<u>46b9a4</u>
   46b9a4
 6 641fa1
                0
                          14
   641fa1
 8 641fa1
   641fa1
10 641fa1
```

```
> ige
# A tibble: 74 \times 3
          visit
   TD
                    IgE
          <db1>
                 <db1>
   <chr>
 1 46b9a4
              0 100
 2 46b9a4
              4 100
  641fa1
                  9.01
                  0.87
 4 641fa1
                2.97
  97bab3
  97bab3
              4 3.7
   d85d4f
                  0.05
  d85d4f
                 0.05
  1b06cf
                 91.2
10 336ddf
                   0.05
```

```
> ige
  symptoms
# A tibble: 193 \times 3
                                       # A tibble: 74 \times 3
          visit symptoms
                                                  visit
                                           TD
                                                            IgE
   ID
   <chr> <dbl>
                     <db1>
                                                  <db1>
                                                          <db1>
                                          <chr>
   46b9a4
                                          46b9a4
                                                       0 100
                                                         100
   46b9a4
                                          46b9a4
                         2
                                          641fa1
   46b9a4
                                                           9.01
                                        4 641fa1
                                                           0.87
 4<u>46b9a4</u>
                                                           2.97
   46b9a4
               4
                                          97bab3
 6 641fa1
               0
                        14
                                          97bab3
                                                       4 3.7
   641fa1
                                          d85d4f
                                                           0.05
                                                         0.05
                                          d85d4f
 8 641fa1
                                          1b06cf
 9 641fa1
                                                         91.2
10 641fa1
                                       10 336ddf
                                                           0.05
```

```
symptoms
     left_join(ige, by = c("ID", "visit"))
> symptoms
# A tibble: 193 \times 3
          visit symptoms
   TD
   <chr> <dbl>
                   <db1>
 1 46b9a4
              0
                       0
 2 46b9a4
 3 46b9a4
 4 46b9a4
 5 46b9a4
              0
 6 641fa1
                      14
 7 641fa1
                   > ige
              2
 8 641fa1
                   # A tibble: 74 \times 3
 9 641fa1
                             visit
                                       IgE
                      ID
10 641fa1
                             <db1>
                                    <db1>
                      <chr>
                    1 46b9a4
                                 0 100
                    2 46b9a4
                                 4 100
                    3 641fa1
                                     9.01
                    4 641fa1
                                     0.87
                                   2.97
                    5 97bab3
                    6 97bab3
                                     3.7
                    7 d85d4f
                                   0.05
                    8 d85d4f
                                     0.05
                    9 1b06cf
                                    91.2
                   10 336ddf
                                      0.05
```

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

```
visit symptoms
   TD
   <chr> <dbl>
                    <db1>
1 46b9a4
               0
                         0
2 46b9a4
 3 46b9a4
4 46b9a4
 5 46b9a4
6 641fa1
               0
                        14
7 641fa1
                    > ige
8 641fa1
                    # A tibble: 74 \times 3
9 641fa1
                                         IgE
                               visit
                        ID
10 641fa1
                               <db1>
                                       <db1>
                       <chr>
                     1 46b9a4
                                   0 100
                     2 46b9a4
                                   4 100
                     3 641fa1
                                        9.01
                     4 641fa1
                                        0.87
                                        2.97
                     5 97bab3
                     6 97bab3
                                        3.7
                     7 d85d4f
                                        0.05
                     8 d85d4f
                                        0.05
                     9 1b06cf
                                       91.2
                    10 336ddf
                                        0.05
```

> symptoms

A tibble: 193×3

```
# A tibble: 193 \times 4
          visit symptoms
   ID
                              IgE
   <chr> <dbl>
                    <db1>
                            <db1>
 1 46b9a4
                           100
  46b9a4
                            NA
  46b9a4
                            NA
 4 46b9a4
                            NA
  46b9a4
                           100
 6 641fa1
                        14
                             9.01
  641fa1
                            NA
 8 641fa1
                            NA
  641fa1
                            NA
10 641fa1
                             0.87
```

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

```
# A tibble: 193 \times 3
          visit symptoms
   TD
   cchrs cdh1s cdh1s
 1 46b9a4
               0
2 46b9a4
 3 46b9a4
4 46b9a4
 5 46b9a4
6 641fa1
               0
                        14
7 641fa1
                    > ige
8 641fa1
                    # A tibble: 74 \times 3
9 641fa1
                               visit
                                         IgE
                        ID
10 641fa1
                                       <db1>
                       <chr>
                               <db1>
                     1 46b9a4
                                   0 100
                     2 46b9a4
                                   4 100
                     3 641fa1
                                        9.01
                     4 641fa1
                                        0.87
                                        2.97
                     5 97bab3
                     6 97bab3
                                        3.7
                     7 d85d4f
                                        0.05
                     8 d85d4f
                                        0.05
                     9 1b06cf
                                       91.2
                    10 336ddf
                                        0.05
```

```
# A tibble: 193 \times 4
          visit symptoms
   ID
                              IgE
   <chr> <dbl>
                    <db1>
                            <db1>
 1 46b9a4
                           100
 2 46b9a4
                            NA
  46b9a4
                            NA
 4 46b9a4
                            NA
  46b9a4
                           100
 6 641fa1
                        14
                             9.01
  641fa1
                            NA
 8 641fa1
                            NA
  641fa1
                            NA
10 641fa1
                             0.87
```

```
symptoms
     left join(ige, by = c("ID", "visit"))
> symptoms
# A tibble: 193 \times 3
          visit symptoms
   TD
   cchrs cdh1s cdh1s
 1 46b9a4
              0
 2 46b9a4
 3 46b9a4
 4 46b9a4
 5 46b9a4
 6 641fa1
              0
                      14
 7 641fa1
                   > ige
 8 641fa1
                   # A tibble: 74 \times 3
 9 641fa1
                             visit
                                      IgE
10 641fa1
                      <chr> <dh1> <dh1>
                    1 46b9a4
                                 0 100
                    2 46b9a4
                                 4 100
                    3 641fa1
                                      9.01
                    4 641fa1
                                      0.87
                                      2.97
                    5 97bab3
                    6 97bab3
                                      3.7
                    7 d85d4f
                                     0.05
                    8 d85d4f
                                     0.05
```

9 1b06cf

10 336ddf

91.2

```
# A tibble: 193 \times 4
          visit symptoms
   ID
                              IgE
   <chr> <dbl>
                            <db1>
                    <db1>
 1 46b9a4
                           100
 2 46b9a4
                            NA
  46b9a4
                            NA
 4 46b9a4
                            NA
  46b9a4
                           100
 6 641fa1
                        14
                             9.01
  641fa1
                            NA
 8 641fa1
                            NA
  641fa1
                            NA
10 641fa1
                             0.87
```

```
symptoms
     left join(ige, by = c("ID", "visit"))
> symptoms
# A tibble: 193 \times 3
          visit symptoms
   TD
   cchrs <dh1> <dh1>
 1 46b9a4
              0
 2 46b9a4
 3 46b9a4
 4 46b9a4
 5 46b9a4
 6 641fa1
              0
                      14
 7 641fa1
                   > ige
 8 641fa1
                   # A tibble: 74 \times 3
              3
 9 641fa1
                             visit
                      ID
                                       IgE
10 641fa1
                      <chr> <dh1> <dh1>
                    1 46b9a4
                                  0 100
                    2 46b9a4
                                  4 100
                    3 641fa1
                                      9.01
                    4 641fa1
                                      0.87
                                      2.97
                    5 97bab3
                    6 97bab3
                                      3.7
                    7 d85d4f
                                      0.05
                    8 d85d4f
                                      0.05
```

9 1b06cf

10 336ddf

91.2

```
# A tibble: 193 \times 4
          visit symptoms
                              IgE
   <chr> <db1>
                    <db1>
                            <db1>
  46b9a4
                         0 100
               0
  46b9a4
                            NA
  46b9a4
                            NA
 4 46b9a4
                            NA
  46b9a4
                           100
  641fa1
                        14
                             9.01
  641fa1
                            NA
  641fa1
                            NA
  641fa1
                            NA
10 641fa1
                             0.87
```

```
symptoms
     left join(ige, by = c("ID", "visit"))
> symptoms
# A tibble: 193 \times 3
          visit symptoms
   TD
   cchrs <dh1> <dh1>
 1 46b9a4
              0
 2 46b9a4
 3 46b9a4
 4 46b9a4
 5 46b9a4
              4
 6 641fa1
              0
                      14
 7 641fa1
                   > ige
 8 641fa1
                   # A tibble: 74 \times 3
              3
 9 641fa1
                             visit
                                       IgE
10 641fa1
                      <chr> <dh1> <dh1>
                    1 46b9a4
                                  0 100
                    2 46b9a4
                                  4 100
                    3 641fa1
                                      9.01
                    4 641fa1
                                      0.87
                                      2.97
                    5 97bab3
                    6 97bab3
                                      3.7
                    7 d85d4f
                                      0.05
                    8 d85d4f
                                      0.05
                    9 1b06cf
                                     91.2
```

10 336ddf

```
# A tibble: 193 \times 4
          visit symptoms
                              IgE
   <chr> <db1>
                    <db1>
                            <db1>
  46b9a4
                         0 100
               0
  46b9a4
                            NA
  46b9a4
                            NA
 4 46b9a4
                            NA
  46b9a4
                           100
  641fa1
                        14
                             9.01
  641fa1
                            NA
  641fa1
                            NA
  641fa1
                            NA
10 641fa1
                             0.87
```

```
symptoms
     left join(ige, by = c("ID", "visit"))
> symptoms
# A tibble: 193 \times 3
          visit symptoms
   TD
   cchrs <dh1> <dh1>
 1 46b9a4
              0
 2 46b9a4
 3 46b9a4
 4 46b9a4
 5 46b9a4
 6 641fa1
              0
                       14
 7 641fa1
                   > ige
 8 641fa1
                   # A tibble: 74 \times 3
              3
 9 641fa1
                             visit
                                       IgE
10 641fa1
                      <chr> <dh1> <dh1>
                    1 46b9a4
                                  0 100
                    2 46b9a4
                                  4 100
                    3 641fa1
                                      9.01
                    4 641fa1
                                      0.87
                                      2.97
                    5 97bab3
                    6 97bab3
                                      3.7
          ???
                    7 d85d4f
                                      0.05
                    8 d85d4f
                                      0.05
                    9 1b06cf
                                     91.2
```

10 336ddf

```
# A tibble: 193 \times 4
          visit symptoms
                              IgE
   <chr> <db1>
                    <db1>
                            <db1>
   46b9a4
                         0 100
               0
  46b9a4
                            NA
  46b9a4
                            NA
  46b9a4
                            NA
  46b9a4
                           100
  641fa1
                        14
                             9.01
  641fa1
                            NA
  641fa1
                            NA
  641fa1
                            NA
10 641fa1
                             0.87
```

```
symptoms
     left join(ige, by = c("ID", "visit"))
> symptoms
# A tibble: 193 \times 3
          visit symptoms
   TD
   cchrs <dh1> <dh1>
 1 46b9a4
              0
 2 46b9a4
 3 46b9a4
 4 46b9a4
 5 46b9a4
 6 641fa1
              0
                       14
 7 641fa1
                   > ige
 8 641fa1
                   # A tibble: 74 \times 3
              3
 9 641fa1
                             visit
                                       IgE
10 641fa1
                      <chr> <dh1> <dh1>
                    1 46b9a4
                                  0 100
                    2 46b9a4
                                  4 100
                    3 641fa1
                                      9.01
                    4 641fa1
                                      0.87
                                      2.97
                    5 97bab3
                    6 97bab3
                                      3.7
          ???
                    7 d85d4f
                                      0.05
                    8 d85d4f
                                      0.05
                    9 1b06cf
                                     91.2
```

10 336ddf

```
# A tibble: 193 \times 4
           visit symptoms
                               IgE
   <chr> <db1>
                     <db1>
                             <db1>
   46b9a4
                         0 100
               0
  46b9a4
                             NA
  46b9a4
                             NA
   46b9a4
                             NA
                            100
   46b9a4
  641fa1
                              9.01
                        14
  641fa1
                             NA
  641fa1
                            NA
  641fa1
                             NA
10 641fa1
                              0.87
```

```
symptoms
     left join(ige, by = c("ID", "visit"))
> symptoms
# A tibble: 193 \times 3
          visit symptoms
   TD
   cchr> <dh1>
                cdh1s
 1 46b9a4
              0
 2 46b9a4
 3 46b9a4
 4 46b9a4
 5 46b9a4
              0
 6 641fa1
                       14
 7 641fa1
                   > ige
 8 641fa1
                   # A tibble: 74 \times 3
              3
 9 641fa1
                              visit
                                       IgE
10 641fa1
                       <chr> <dh1> <dh1>
                     1 46b9a4
                                  0 100
                     2 46b9a4
                                  4 100
                     3 641fa1
                                      9.01
                                      0.87
                    4 641fa1
                                      2.97
                     5 97bab3
                                      3.7
                    6 97bab3
          ???
                                      0.05
                    7 d85d4f
                    8 d85d4f
                                      0.05
                      1b06cf
                                     91.2
                   10 336ddf
                                      0.05
```

```
A tibble: 193 \times 4
          visit symptoms
                               IgE
   <chr> <db1>
                     <db1>
                            <db1>
  46b9a4
                         0 100
  46b9a4
                            NA
  46b9a4
                            NA
  46b9a4
                            NA
  46b9a4
                           100
                              9.01
  641fa1
                        14
  641fa1
                            NA
  641fa1
                            NA
  641fa1
                            NA
10 641fa1
                              0.87
```

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

```
symptoms
     inner_join(ige, by = c("ID", "visit"))
> symptoms
# A tibble: 193 \times 3
          visit symptoms
   TD
   <chr> <dbl>
                   <db1>
 1 46b9a4
                       0
 2 46b9a4
 3 46b9a4
 4 46b9a4
 5 46b9a4
 6 641fa1
              0
 7 641fa1
                   > ige
 8 641fa1
                   # A tibble: 74 \times 3
 9 641fa1
                             visit
                                      IgE
10 641fa1
                                    <db1>
                      <chr>
                             <db1>
                    1 46b9a4
                                 0 100
                    2 46b9a4
                                 4 100
                    3 641fa1
                                     9.01
                    4 641fa1
                                     0.87
                                   2.97
                    5 97bab3
                    6 97bab3
                                     3.7
                    7 d85d4f
                                   0.05
                    8 d85d4f
                                     0.05
```

91.2

0.05

9 1b06cf

10 336ddf

3 46b9a4

4 46b9a4

5 46b9a4

6 641fa1

7 641fa1

8 641fa1

9 641fa1

10 641fa1

0

```
14
> ige
# A tibble: 74 \times 3
           visit
                     IgE
   ID
                  <db1>
   <chr>
           <db1>
               0 100
 1 46b9a4
 2 46b9a4
               4 100
 3 641fa1
                    9.01
                    0.87
 4 641fa1
                    2.97
 5 97bab3
 6 97bab3
                    3.7
 7 d85d4f
                   0.05
 8 d85d4f
                   0.05
 9 1b06cf
                  91.2
```

0.05

10 336ddf

```
# A tibble: 74 \times 4
          visit symptoms
                               IgE
   <chr> <dbl>
                     <db1>
                            <db1>
 1 46b9a4
                           100
  46b9a4
                           100
                             9.01
  641fa1
                        14
 4 641fa1
                             0.87
  97bab3
                             2.97
  97bab3
                             3.7
  d85d4f
               0
                             0.05
  d85d4f
                             0.05
   1b06cf
                            91.2
10 336ddf
               0
                             0.05
```

```
symptoms
     inner_join(ige, by = c("ID", "visit"))
> symptoms
# A tibble: 193 \times 3
          visit symptoms
   TD
   cchr> <dh1> <dh1>
 1 46b9a4
              0
 2 46b9a4
 3 46b9a4
 4 46b9a4
 5 46b9a4
 6 641fa1
              0
 7 641fa1
                   > ige
 8 641fa1
                   # A tibble: 74 \times 3
 9 641fa1
                             visit
                                       IgE
                      ID
10 641fa1
                                     <db1>
                      <chr>
                             <db1>
                                  0 100
                    1 46b9a4
                    2 46b9a4
                                  4 100
                    3 641fa1
                                      9.01
                    4 641fa1
                                      0.87
                                      2.97
                    5 97bab3
                    6 97bab3
                                      3.7
```

7 d85d4f

8 d85d4f

9 1b06cf

10 336ddf

0.05

0.05

0.05

91.2

```
# A tibble: 74 \times 4
          visit symptoms
                              IgE
   <chr> <dbl>
                     <db1>
                            <db1>
 1 46b9a4
                           100
  46b9a4
                           100
                             9.01
  641fa1
                        14
 4 641fa1
                             0.87
  97bab3
                             2.97
  97bab3
                             3.7
  d85d4f
               0
                             0.05
  d85d4f
                             0.05
   1b06cf
                            91.2
10 336ddf
                             0.05
```

```
symptoms
     inner_join(ige, by = c("ID", "visit"))
> symptoms
# A tibble: 193 \times 3
          visit symptoms
   TD
   cchr> <dh1> <dh1>
 1 46b9a4
              0
 2 46b9a4
 3 46b9a4
 4 46b9a4
 5 46b9a4
 6 641fa1
              0
 7 641fa1
                   > ige
 8 641fa1
                    # A tibble: 74 \times 3
 9 641fa1
                              visit
                                       IgE
10 641fa1
                              \frac{dh1}{dh1}
                     1 46b9a4
                                  0 100
                                  4 100
                     2 46b9a4
                     3 641fa1
                                      9.01
                                      0.87
                     4 641fa1
                                      2.97
                     5 97bab3
                     6 97bab3
                                      3.7
                     7 d85d4f
                                      0.05
                    8 d85d4f
                                      0.05
                     9 1b06cf
                                     91.2
```

10 336ddf

0.05

```
# A tibble: 74 \times 4
          visit symptoms
                               IgE
   <chr> <dbl>
                     <db1>
                             < db1 >
 1 46b9a4
                           100
  46b9a4
                           100
                              9.01
  641fa1
                        14
 4 641fa1
                              0.87
  97bab3
                              2.97
  97bab3
                              3.7
  d85d4f
                              0.05
  d85d4f
                             0.05
   1b06cf
                             91.2
10 336ddf
                              0.05
```

```
symptoms
     inner_join(ige, by = c("ID", "visit"))
> symptoms
# A tibble: 193 \times 3
          visit symptoms
   TD
   cchr> <dh1> <dh1>
 1 46b9a4
              0
 2 46b9a4
 3 46b9a4
 4 46b9a4
 5 46b9a4
 6 641fa1
              0
 7 641fa1
                    > ige
 8 641fa1
                    # A tibble: 74 \times 3
 9 641fa1
                              visit
                                       IgE
10 641fa1
                              \frac{dh1}{dh1}
                     1 46b9a4
                                  0 100
                                  4 100
                     2 46b9a4
                     3 641fa1
                                      9.01
                                      0.87
                     4 641fa1
                                      2.97
                     5 97bab3
                     6 97bab3
                                      3.7
                     7 d85d4f
                                      0.05
                     8 d85d4f
                                      0.05
                     9 1b06cf
                                     91.2
```

10 336ddf

0.05

```
# A tibble: 74 \times 4
           visit symptoms
                               IgE
   <chr> <db1>
                     <db1>
                             <db1>
                         0 100
  46b9a4
               0
  46b9a4
               4
                            100
                              9.01
   641fa1
                        14
  641fa1
                              0.87
  97bab3
                              2.97
  97bab3
                              3.7
  d85d4f
               0
                              0.05
  d85d4f
                              0.05
   1b06cf
                             91.2
10 336ddf
                              0.05
```

```
symptoms
     inner_join(ige, by = c("ID", "visit"))
> symptoms
# A tibble: 193 \times 3
          visit symptoms
   TD
   cchr> <dh1>
                cdh1s
              0
 1 46b9a4
 2 46b9a4
  46b9a4
 4 46b9a4
 5 46b9a4
 6 641fa1
              0
                       14
 7 641fa1
                    > ige
 8 641fa1
                    # A tibble: 74 \times 3
 9 641fa1
                              visit
                                       IgE
10 641fa1
                              \frac{dh1}{dh1}
                     1 46b9a4
                                  0 100
                                  4 100
                     2 46b9a4
                     3 641fa1
                                      9.01
                                      0.87
                     4 641fa1
                                      2.97
                     5 97bab3
                                      3.7
                     6 97bab3
                                      0.05
                     7 d85d4f
                     8 d85d4f
                                      0.05
                     9 1b06cf
                                     91.2
                    10 336ddf
                                      0.05
```

```
A tibble: 74 \times 4
          visit symptoms
                               IgE
   <chr> <db1>
                     <db1>
                            <db1>
                         0 100
  46b9a4
               0
  46b9a4
               4
                           100
                              9.01
  641fa1
                        14
  641fa1
                              0.87
  97bab3
                              2.97
  97bab3
                              3.7
  d85d4f
               0
                              0.05
  d85d4f
                             0.05
  1b06cf
                            91.2
10 336ddf
                              0.05
```

```
symptoms
     inner_join(ige, by = c("ID", "visit"))
> symptoms
# A tibble: 193 \times 3
          visit symptoms
   TD
   cchr> <dh1>
                cdh1s
 1 46b9a4
              0
 2 46b9a4
  46b9a4
 4 46b9a4
 5 46b9a4
 6 641fa1
              0
 7 641fa1
                   > ige
 8 641fa1
                    # A tibble: 74 \times 3
 9 641fa1
                              visit
                                       IgE
10 641fa1
                              \frac{dh1}{dh1}
                                  0 100
                     1 46b9a4
                                  4 100
                     2 46b9a4
                      641fa1
                                      9.01
                                      0.87
                     4 641fa1
                                      2.97
                     5 97bab3
                                      3.7
                     6 97bab3
          ???
                                      0.05
                     7 d85d4f
                    8 d85d4f
                                      0.05
                      1b06cf
                                     91.2
```

10 336ddf

0.05

```
A tibble: 74 \times 4
          visit symptoms
                               IgE
   <chr> <db1>
                     <db1>
                            <db1>
                         0 100
  46b9a4
               0
  46b9a4
               4
                           100
                              9.01
  641fa1
  641fa1
                              0.87
  97bab3
                              2.97
  97bab3
                              3.7
  d85d4f
               0
                              0.05
  d85d4f
                             0.05
  1b06cf
                            91.2
10 336ddf
                              0.05
```

```
symptoms
     inner_join(ige, by = c("ID", "visit"))
> symptoms
# A tibble: 193 \times 3
          visit symptoms
   TD
   cchr> <dh1>
                cdh1s
 1 46b9a4
              0
 2 46b9a4
  46b9a4
 4 46b9a4
  46b9a4
              4
 6 641fa1
              0
 7 641fa1
                    > ige
 8 641fa1
                    # A tibble: 74 \times 3
              3
 9 641fa1
                              visit
                                       IgE
10 641fa1
                              \frac{dh1}{dh1}
                     1 46b9a4
                                  0 100
                                  4 100
                     2 46b9a4
                      641fa1
                                      9.01
                                      0.87
                     4 641fa1
                                      2.97
                     5 97bab3
                                      3.7
                     6 97bab3
          ???
                                      0.05
                     7 d85d4f
                     8 d85d4f
                                      0.05
                      1b06cf
                                     91.2
```

10 336ddf

0.05

```
A tibble: 74 \times 4
          visit symptoms
                               IgE
   <chr> <db1>
                    <db1>
                            <db1>
                         0 100
  46b9a4
  46b9a4
               4
                           100
                             9.01
  641fa1
  641fa1
                             0.87
  97bab3
                             2.97
  97bab3
                             3.7
  d85d4f
               0
                             0.05
  d85d4f
                             0.05
  1b06cf
                            91.2
10 336ddf
                             0.05
```

```
symptoms
     inner_join(ige, by = c("ID", "visit"))
> symptoms
# A tibble: 193 \times 3
          visit symptoms
   TD
   cchr> <dh1>
                cdh1s
 1 46b9a4
              0
 2 46b9a4
  46b9a4
 4 46b9a4
  46b9a4
              4
 6 641fa1
              0
 7 641fa1
                    > ige
 8 641fa1
                    # A tibble: 74 \times 3
               3
 9 641fa1
                              visit
                                        IgE
10 641fa1
                              \frac{2dh1}{2dh1}
                      46b9a4
                                  0 100
                                  4 100
                      46b9a4
                      641fa1
                                       9.01
                                       0.87
                     4 641fa1
                                       2.97
                     5 97bab3
                                       3.7
                     6 97bab3
           ???
                                      0.05
                     7 d85d4f
                     8 d85d4f
                                       0.05
                      1b06cf
                                      91.2
                    10 336ddf
                                       0.05
```

```
A tibble: 74 \times 4
          visit symptoms
                              IgE
   <chr> <db1>
                    <db1>
                            <db1>
  46b9a4
                         0 100
  46b9a4
               4
                           100
  641fa1
                             9.01
  641fa1
                             0.87
  97bab3
                             2.97
  97bab3
                             3.7
  d85d4f
               0
                             0.05
  d85d4f
                             0.05
  1b06cf
                            91.2
10 336ddf
                             0.05
```

```
inner_join(ige, by = c("ID", "visit"))
> symptoms
# A tibble: 193 \times 3
          visit symptoms
   TD
   cchr> <dh1>
                cdh1s
 1 46b9a4
              0
 2 46b9a4
  46b9a4
 4 46b9a4
  46b9a4
              4
 6 641fa1
              0
 7 641fa1
                   > ige
8 641fa1
                   # A tibble: 74 \times 3
              3
9 641fa1
                             visit
                                      IgE
10 641fa1
                             46b9a4
                                 0 100
                                 4 100
                     46b9a4
                     641fa1
                                     9.01
                                     0.87
                    4 641fa1
                                     2.97
                    5 97bab3
                                     3.7
                    6 97bab3
          ???
                    7 d85d4f
                                     0.05
                    8 d85d4f
                                     0.05
                     1b06cf
                                    91.2
                   10 336ddf
                                     0.05
```

symptoms

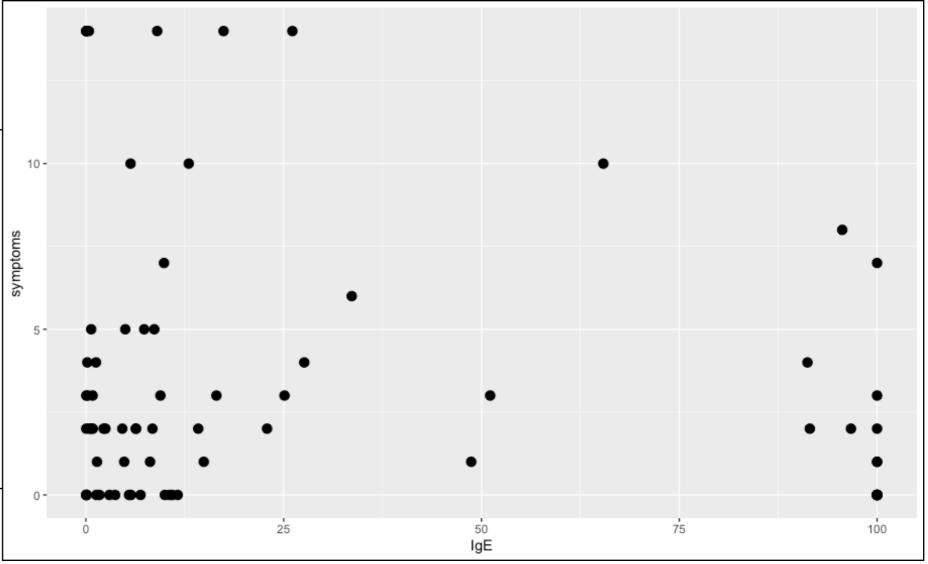
```
A tibble: 74 \times 4
          visit symptoms
                               IgE
   <chr> <db1>
                     <db1>
                             <db1>
  46b9a4
                         0 100
  46b9a4
               4
                            100
               0
  641fa1
                        14
                              9.01
  641fa1
                              0.87
  97bab3
                              2.97
  97bab3
                              3.7
  d85d4f
               0
                              0.05
  d85d4f
                              0.05
  1b06cf
                             91.2
10 336ddf
                              0.05
```

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

```
# A tibble: 74 \times 4
          visit symptoms
                            IgE
   ID
   <chr> <dbl>
                   <db1>
                          <db1>
 1 46b9a4
                       0 100
 2 46b9a4
                       0 100
 3 641fa1
              0
                           9.01
                      14
                         0.87
 4 641fa1
                          2.97
 5 97bab3
                          3.7
 6 97bab3
 7 d85d4f
                          0.05
 8 d85d4f
                         0.05
 9 1b06cf
                         91.2
10 336ddf
                           0.05
```

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

```
# A tibble: 74 \times 4
                             IgE
          visit symptoms
   <chr>
         <db1>
                    <db1>
                           <db1>
 1 46b9a4
                        0 100
 2 46b9a4
                        0 100
              0
                            9.01
 3 641fa1
                       14
                            0.87
 4 641fa1
                            2.97
 5 97bab3
                           3.7
 6 97bab3
 7 d85d4f
                           0.05
 8 d85d4f
                           0.05
                           91.2
 9 1b06cf
                            0.05
10 336ddf
```



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
```

Joins Summary (So Far)

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

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 x 2
          a          b
          <int> <dbl>
1          4      0.319
2          5      -0.811
3          6      -1.05
```

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

```
# A tibble: 6 \times 3
            b
      a
  <int> <dbl>
               <db1>
      4 0.319 NA
      5 -0.811 NA
      6 - 1.05
               NA
4
      7 NA 0.894
               0.00820
      8 NA
6
      9 NA
                0.250
```

```
symptoms
# A tibble: 193 \times 3
   ID
           visit symptoms
   <chr>
           <db1>
                      <db1>
 1 46b9a4
                          0
   46b9a4
                           2
2
 3 46b9a4
 4 46b9a4
   46b9a4
 6 641fa1
                         14
   641fa1
                           2
2
 8 641fa1
   641fa1
10 641fa1
```

```
> ige
# A tibble: 74 \times 3
           visit
   ID
                     IgE
   <chr>
           <db1>
                   <db1>
 1 46b9a4
                 100
  46b9a4
               4 100
   641fa1
                    9.01
 4 641fa1
                    0.87
   97bab3
                    2.97
   97bab3
                    3.7
   d85d4f
                    0.05
   d85d4f
                    0.05
               4
   1b06cf
                   91.2
               0
  336ddf
                    0.05
               0
10
```

```
anti_join(ige, by = c("ID", "visit"))
  symptoms
# A tibble: 193 \times 3
          visit symptoms
   ID
         <db1>
                    <db1>
   <chr>
 1 46b9a4
                  > ige
 2 46b9a4
                  # A tibble: 74 \times 3
 3 46b9a4
                             visit
                     ID
                                      IgE
 4 46b9a4
                             <db1>
                                    <db1>
                     <chr>
 5 46b9a4
                   1 46b9a4
                                 0 100
 6 641fa1
                   2 46b9a4
                                 4 100
 7 641fa1
                   3 641fa1
                                    9.01
8 641fa1
                   4 641fa1
                                    0.87
 9 641fa1
                   5 97bab3
                                    2.97
10 641fa1
                   6 97bab3
                                    3.7
                     d85d4f
                                    0.05
                     d85d4f
                                     0.05
                     1b06cf
                                    91.2
                                     0.05
                     336ddf
```

symptoms |>

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

```
symptoms
# A tibble: 193 \times 3
           visit symptoms
   ID
           <db1>
                     <db1>
   <chr>
 1 46b9a4
                   > ige
  46b9a4
                   # A tibble: 74 \times 3
   46b9a4
                              visit
                                        IgE
                      ID
 4 46b9a4
                                      <db1>
                      <chr>
                              <db1>
  46b9a4
                    1 46b9a4
                                    100
 6 641fa1
                    2 46b9a4
                                  4 100
   641fa1
                    3 641fa1
                                       9.01
 8 641fa1
                    4 641fa1
                                       0.87
   641fa1
                                       2.97
                    5 97bab3
10 641fa1
                                       3.7
                    6 97bab3
                      d85d4f
                                       0.05
                      d85d4f
                                       0.05
                      1b06cf
                                      91.2
                      336ddf
                                       0.05
```

```
A tibble: 119 \times 3
          visit symptoms
   TD
           <db1>
                     <db1>
   <chr>
  46b9a4
  46b9a4
               3
   46b9a4
  641fa1
  641fa1
               3
  641fa1
  97bab3
  97bab3
  97bab3
                        14
10 d85d4f
```

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

```
symptoms
 A tibble: 193 \times 3
           visit symptoms
   ID
                     <db1>
           <db1>
   <chr>
 1 46b9a4
                   > ige
   46b9a4
                  # A tibble: 74 \times 3
   46b9a4
                              visit
                      TD
                                        IgE
 4 46b9a4
                      <chr>>
                                      <db1>
                              <db1>
  46b9a4
                    1 46b9a4
                                    100
 6 641fa1
                    2 46b9a4
                                  4 100
   641fa1
                                       9.01
                    3 641fa1
 8 641fa1
                    4 641fa1
                                       0.87
   641fa1
                                       2.97
                    5 97bab3
10 641fa1
                                      3.7
                    6 97bab3
                      d85d4f
                                      0.05
                      d85d4f
                                       0.05
                      1b06cf
                                      91.2
                      336ddf
                                       0.05
```

```
A tibble: 119 \times 3
          visit symptoms
   TD
          <db1>
   <chr>
                     <db1>
  46b9a4
  46b9a4
               3
  46b9a4
  641fa1
  641fa1
  641fa1
  97bab3
  97bab3
  97bab3
                        14
10 d85d4f
```

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



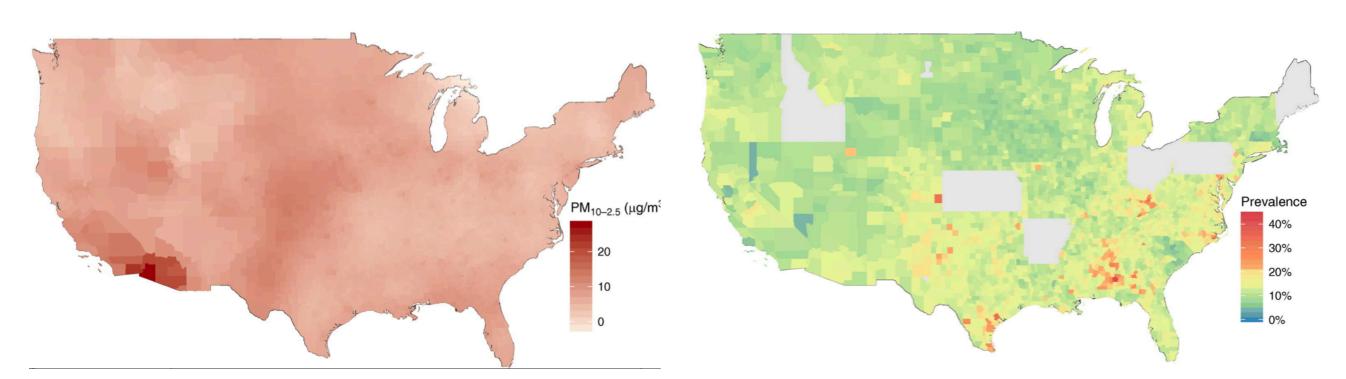
Long-Term Coarse Particulate Matter Exposure Is Associated with Asthma among Children in Medicaid

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

+ Author Affiliations

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

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
1111111wEgu1wgwg	78593	20011127	9
111111wEGw8GGww	75068	19920922	3
1111118G0D1E100	76164	20120611	7

4,400,000 rows

NOTE: Medicaid data is simulated

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

Medicaid Hospitalizations

ID	Date	ICD9
1111118uGEE80mG	20120221	46611
1111118118GDm1E	20120511	380
11111108uu0E8EE	20120917	78900
1111118G0DgD1ug	20120517	V3001
1111111wEEg01mw1	20120517	29633
1111110Gm8mDGuD	20120904	650
1111118um0gG1GG	20120110	V3000
1111110GGm8g0w0	20120916	650

62,000 rows

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
11111118w1DE1gDu	77084	20101127	10
1111118G0GDm0gE	77803	19940716	1
1111110D8gwmGgg	78251	19970816	12
111111wEgu1wgwg	78593	20011127	9
111111wEGw8GGww	75068	19920922	3
1111118G0D1E100	76164	20120611	7

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

T di tiodiato iviattoi			
ZIP	Date	value	
<chr></chr>	<chr></chr>	<dbl></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	

4,400,000 rows

NOTE: Medicaid data is simulated

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
1111111wEgu1wgwg	78593	20011127	9
111111wEGw8GGww	75068	19920922	3
1111118G0D1E100	76164	20120611	7

Medicaid Hospitalizations

ID	Date	ICD9
1111118uGEE80mG	20120221	46611
1111118118GDm1E	20120511	380
11111108uu0E8EE	20120917	78900
1111118G0DgD1ug	20120517	V3001
1111111wEEg01mw1	20120517	29633
1111110Gm8mDGuD	20120904	650
1111118um0gG1GG	20120110	V3000
1111110GGm8g0w0	20120916	650

62,000 rows

Particulate Matter

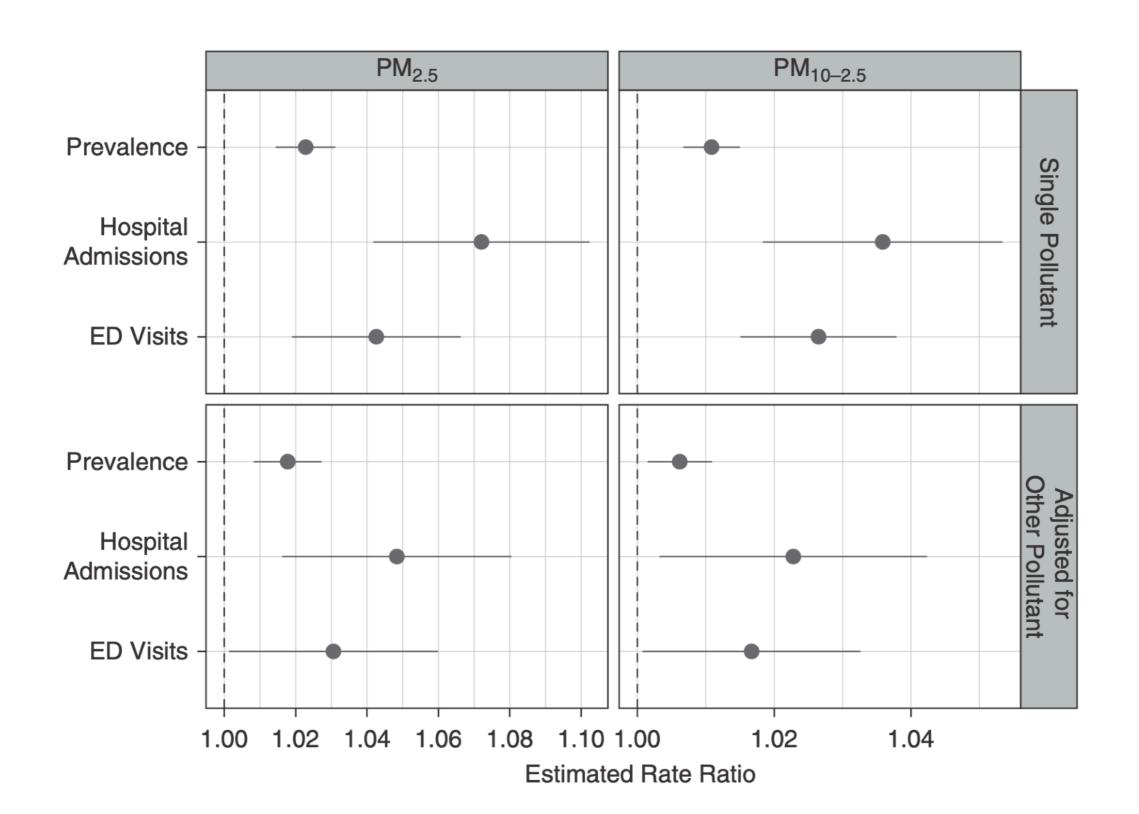
- artiodiato iviattor			
ZIP	Date	value	
<chr></chr>	<chr></chr>	<dbl></dbl>	
76306	20120113	32.4	
75040	20120117	2.1	
75167	20120307	24.6	
78539	20120330	8.9	
77084	20120413	5.9	
77803	20120415	15.9	
78251	20120508	11.1	
78593	20120617	4.8	
75068	20120803	8.7	
76164	20121231	15.8	

Additional tables for weather, Census data...

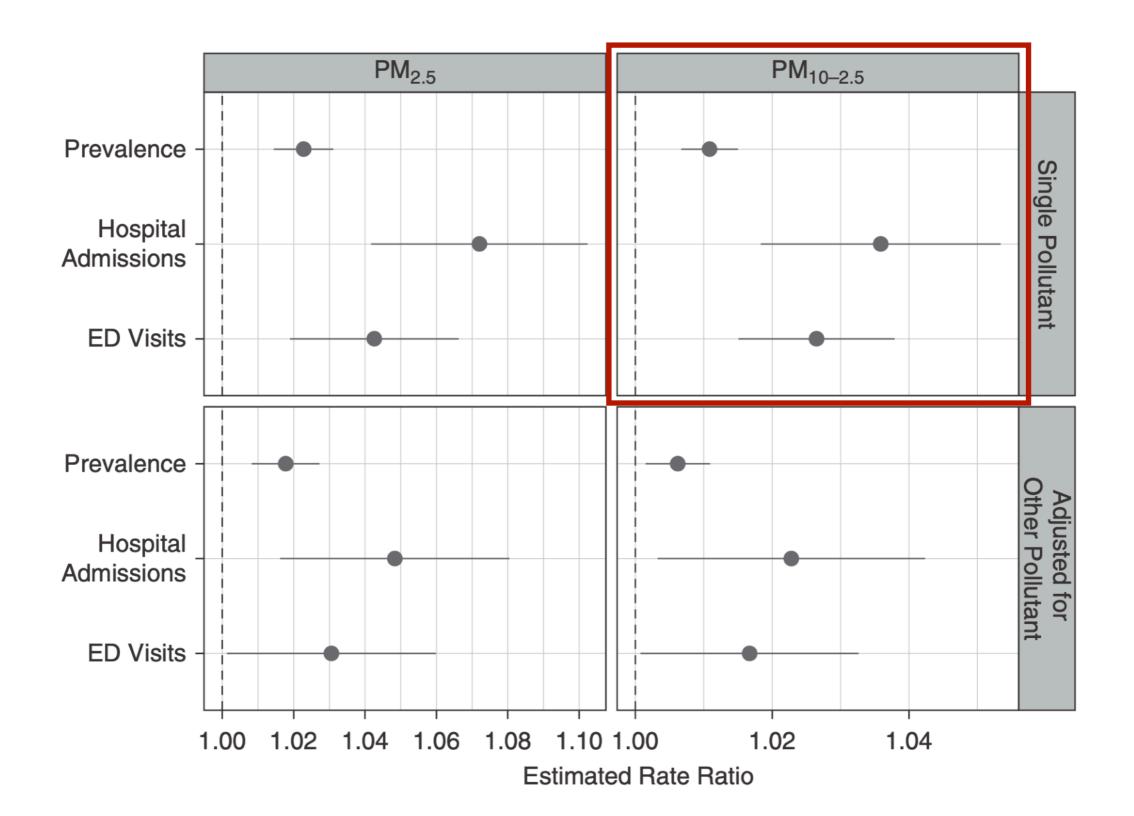
4,400,000 rows

NOTE: Medicaid data is simulated

Results



Results



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