Poli-301: The Political Science Discipline

TODAY'S AGENDA

1 Data Wrangling: summary, group_by

2 Data Wrangling: join, smaller stuff

3 Practice

summarise

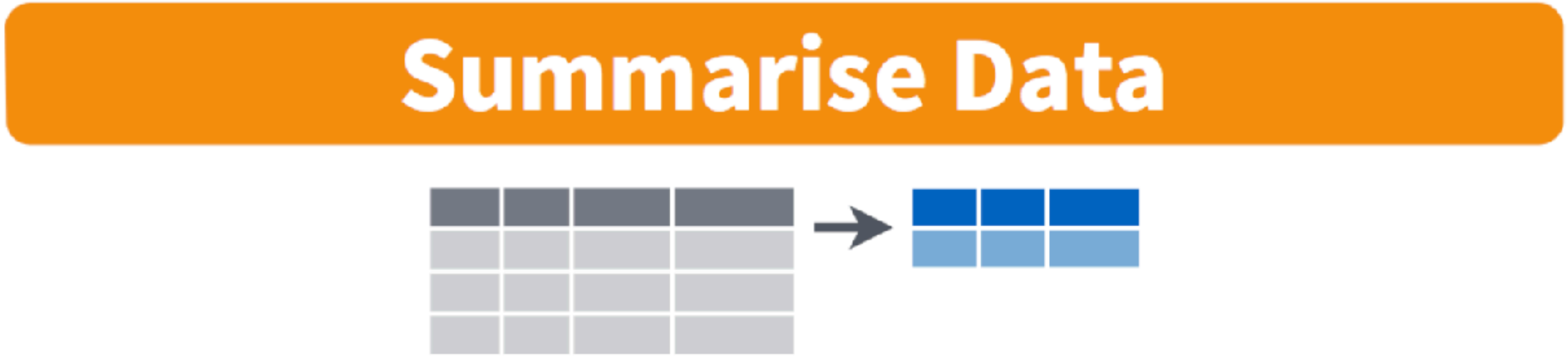


FIGURE 3.3: Diagram of summarize() rows.

Calculate summary statistics —> Numbers that describe the data

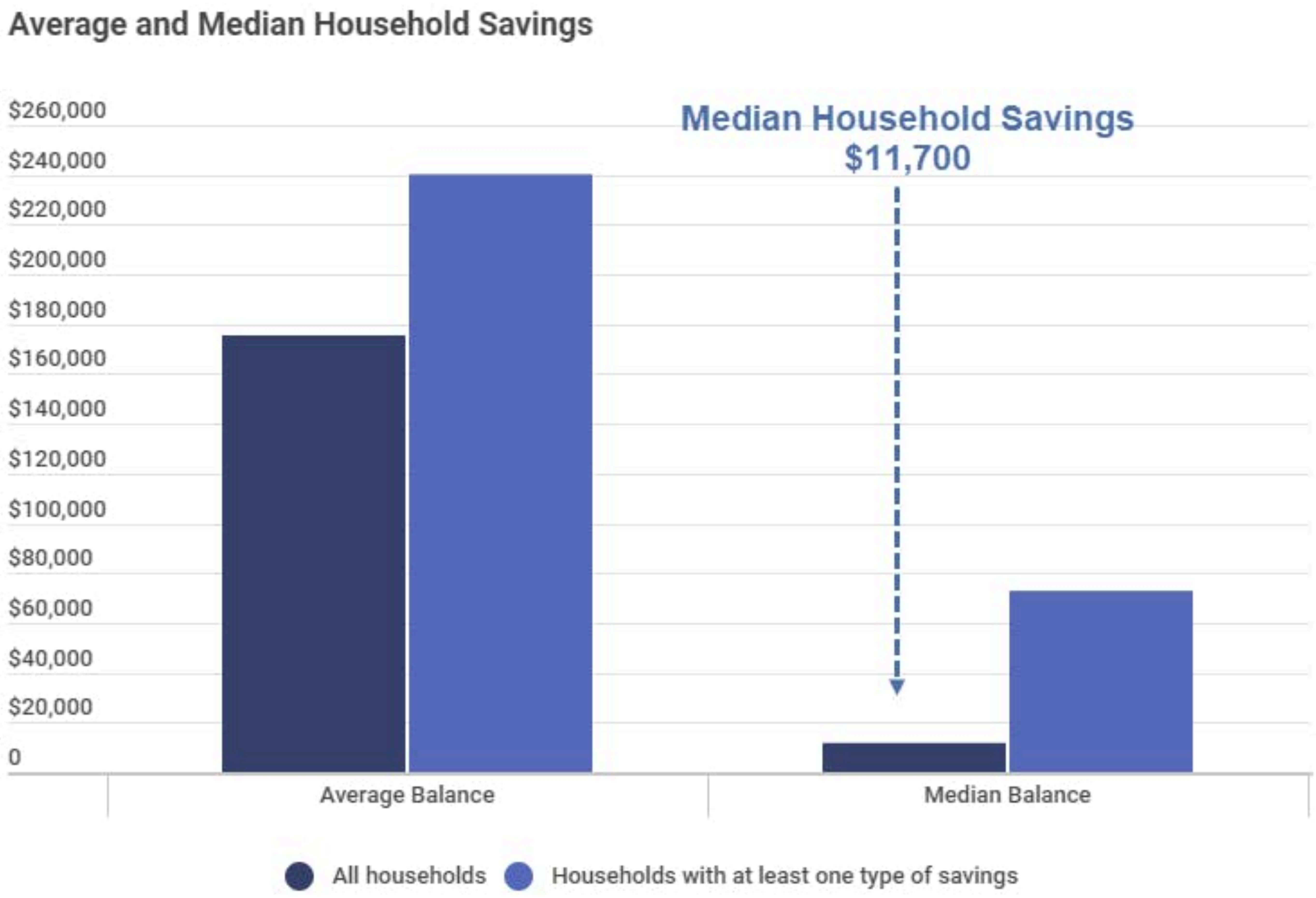
Four quantities

Min

mean

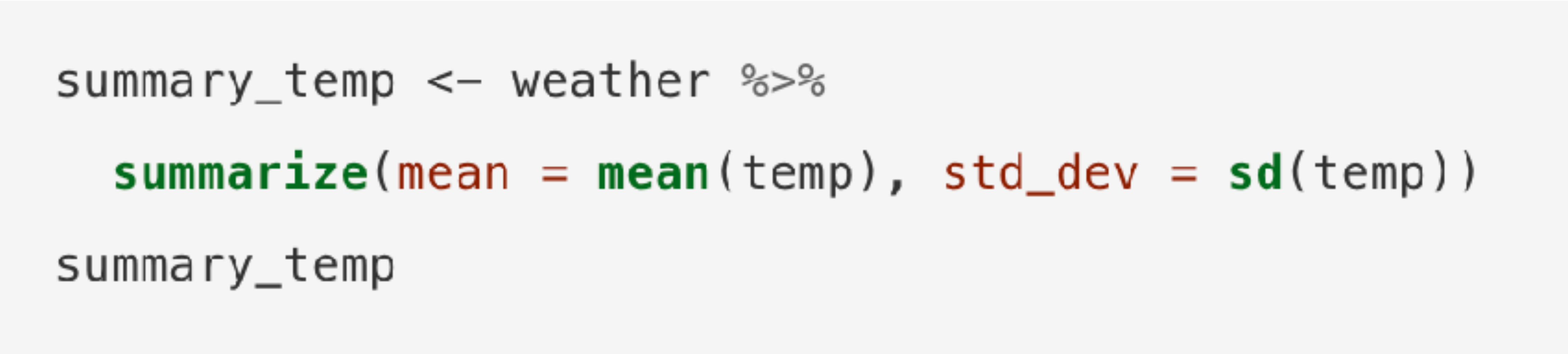
Median

Max

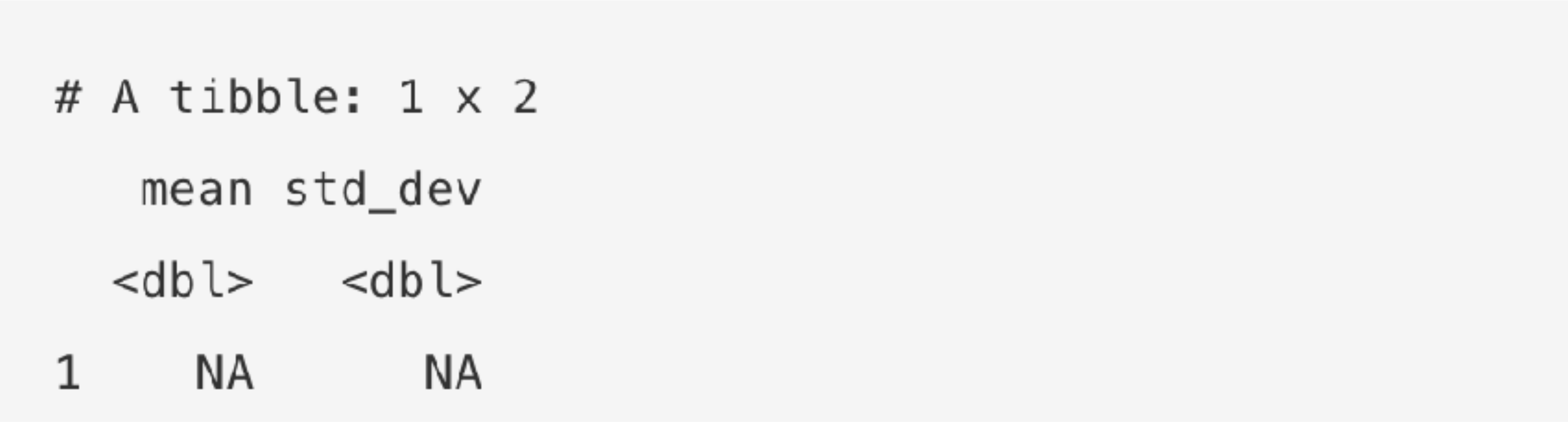


Code

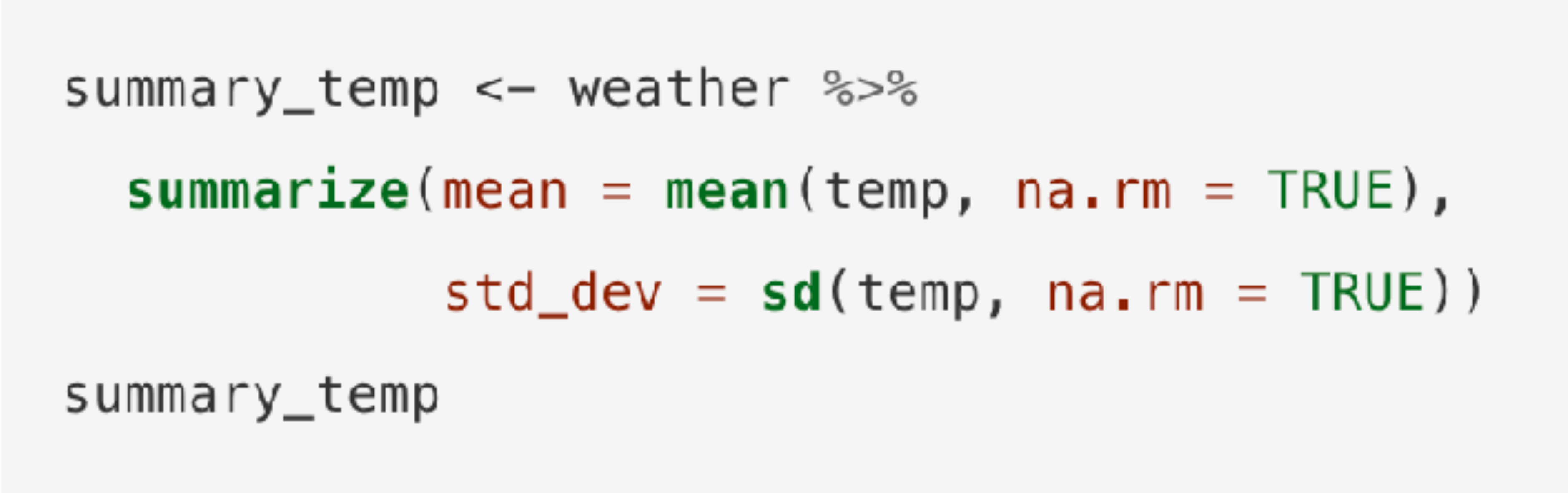
If there’s any missing data (NA) mean(), min(), etc., will return NA



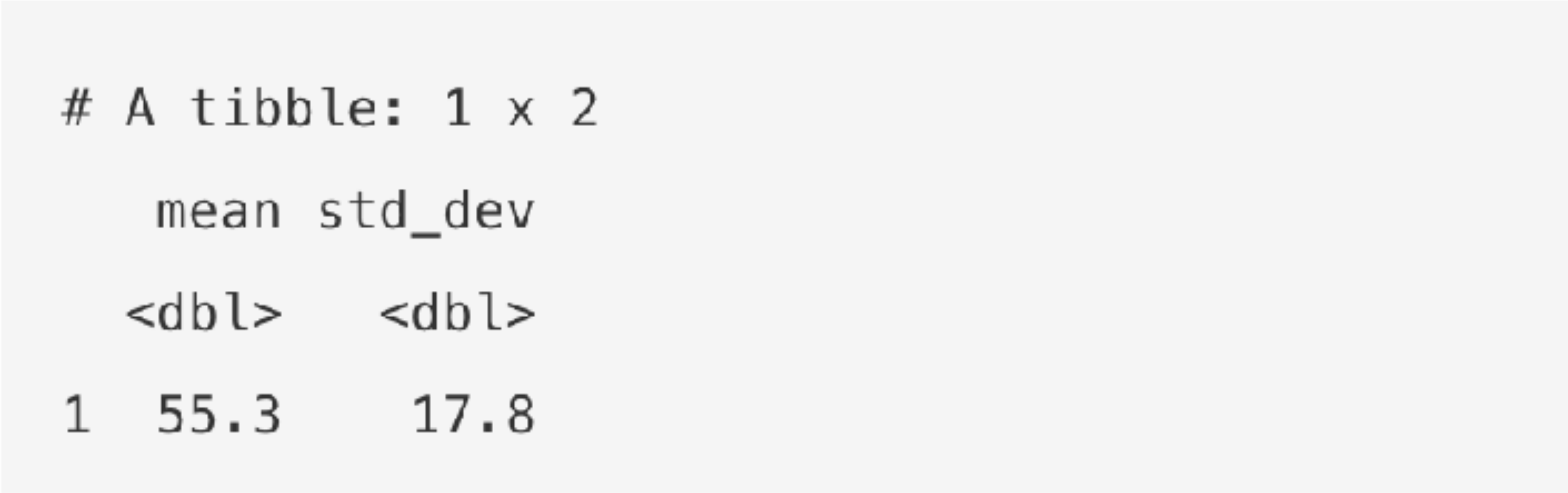
```
summary_temp <- weather %>%
  summarize(mean = mean(temp), std_dev = sd(temp))
summary_temp
```



```
# A tibble: 1 x 2
   mean std_dev
  <dbl>   <dbl>
1    NA      NA
```



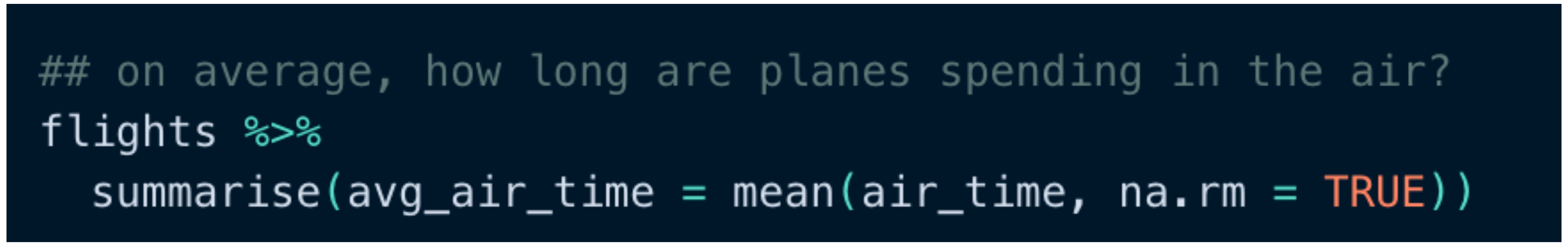
```
summary_temp <- weather %>%
  summarize(mean = mean(temp, na.rm = TRUE),
            std_dev = sd(temp, na.rm = TRUE))
summary_temp
```



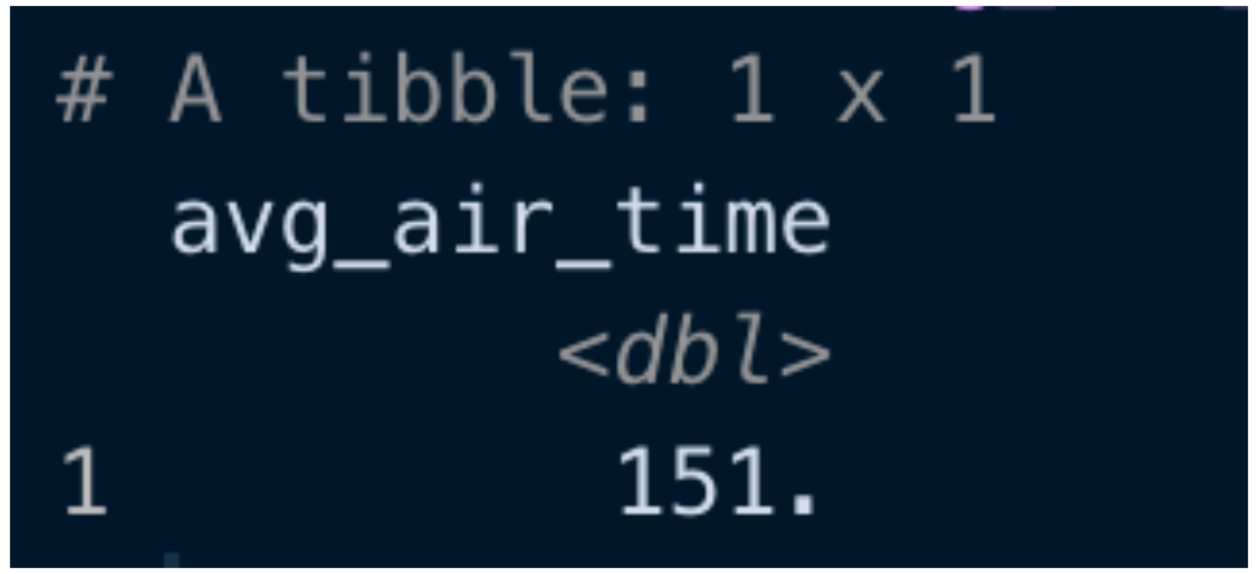
```
# A tibble: 1 x 2
   mean std_dev
  <dbl>   <dbl>
1  55.3    17.8
```

why?

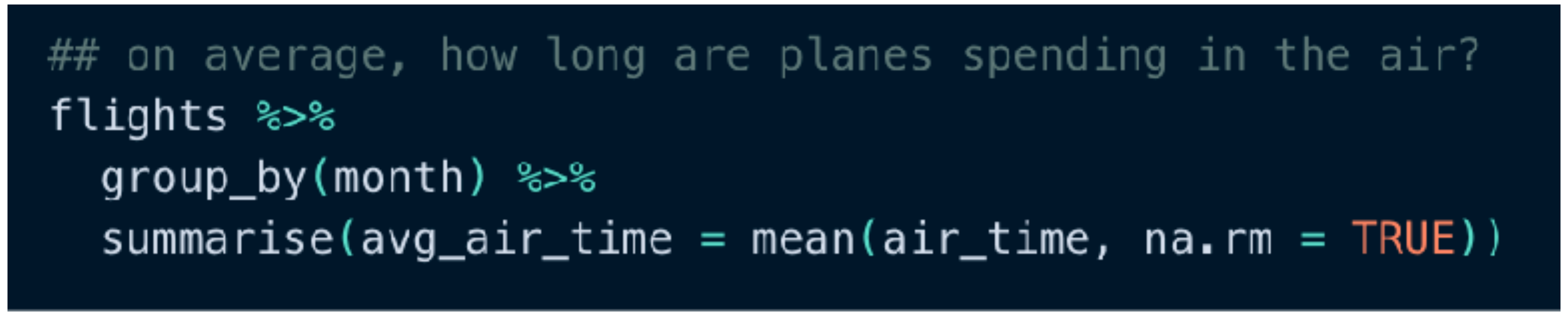
group_by



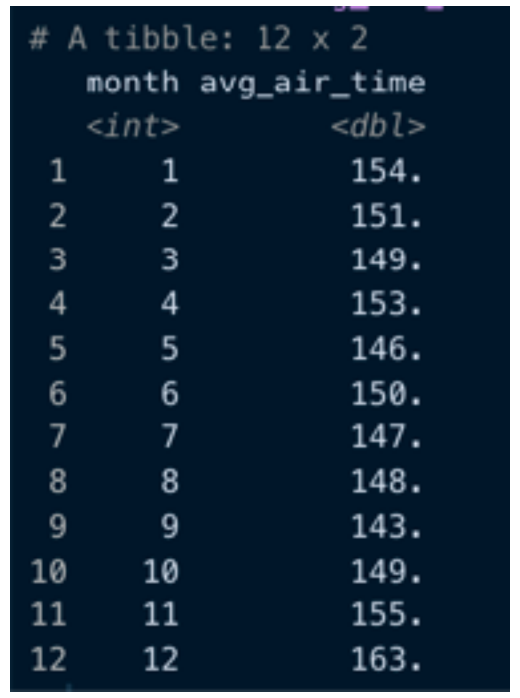
```
## on average, how long are planes spending in the air?
flights %>%
  summarise(avg_air_time = mean(air_time, na.rm = TRUE))
```



```
# A tibble: 1 x 1
  avg_air_time
         <dbl>
1         151.
```



```
## on average, how long are planes spending in the air?
flights %>%
  group_by(month) %>%
  summarise(avg_air_time = mean(air_time, na.rm = TRUE))
```



```
# A tibble: 12 x 2
   month avg_air_time
   <int>        <dbl>
 1     1         154.
 2     2         151.
 3     3         149.
 4     4         153.
 5     5         146.
 6     6         150.
 7     7         147.
 8     8         148.
 9     9         143.
10    10         149.
11    11         155.
12    12         163.
```

Instead of one summary statistic per column, group_by gives us summaries *per* column value

inner_join

merge **two data frames that share a common *key***

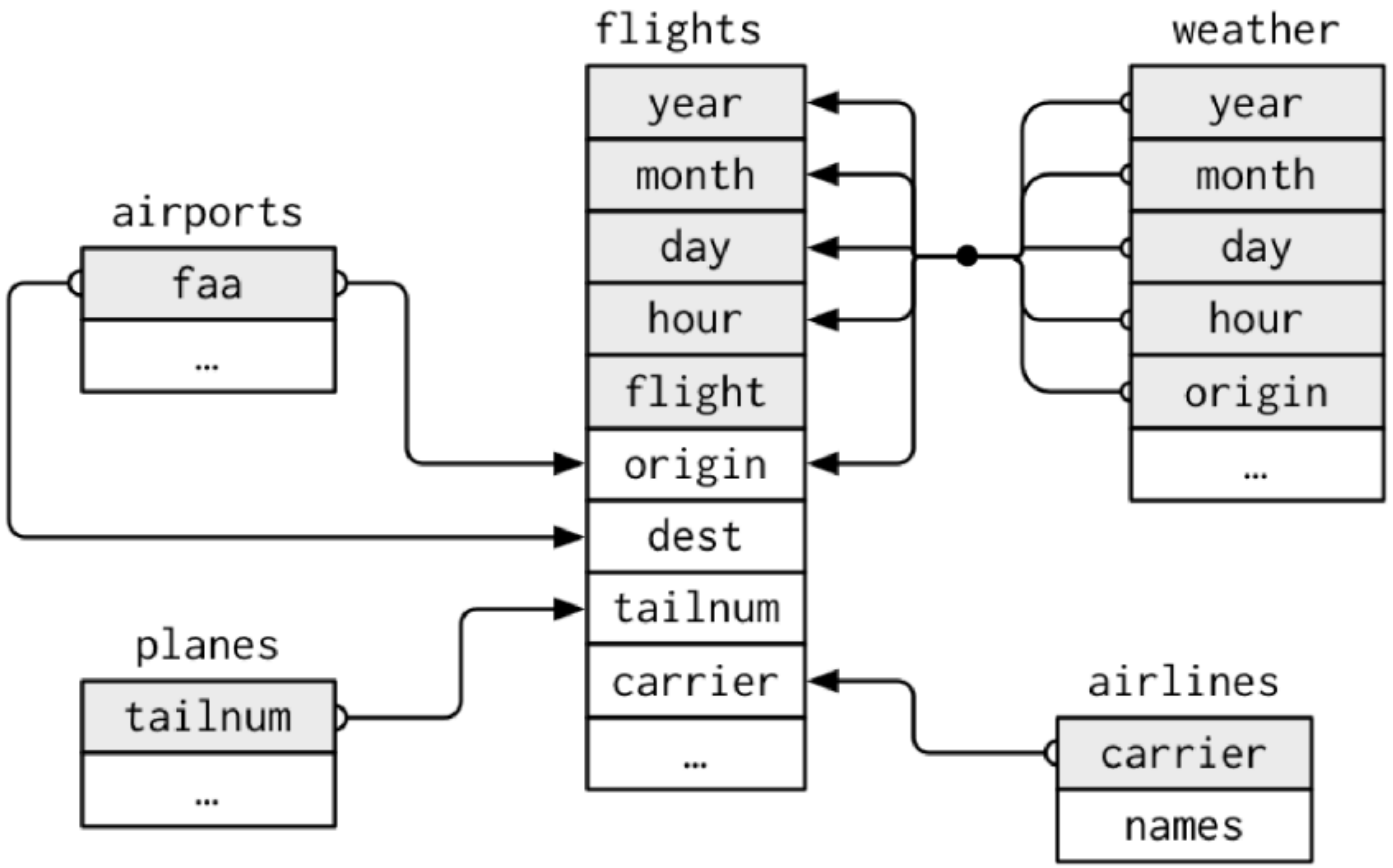


FIGURE 3.7: Data relationships in nycflights13 from R for Data Science.

Subset variables

Subset Variables (Columns)

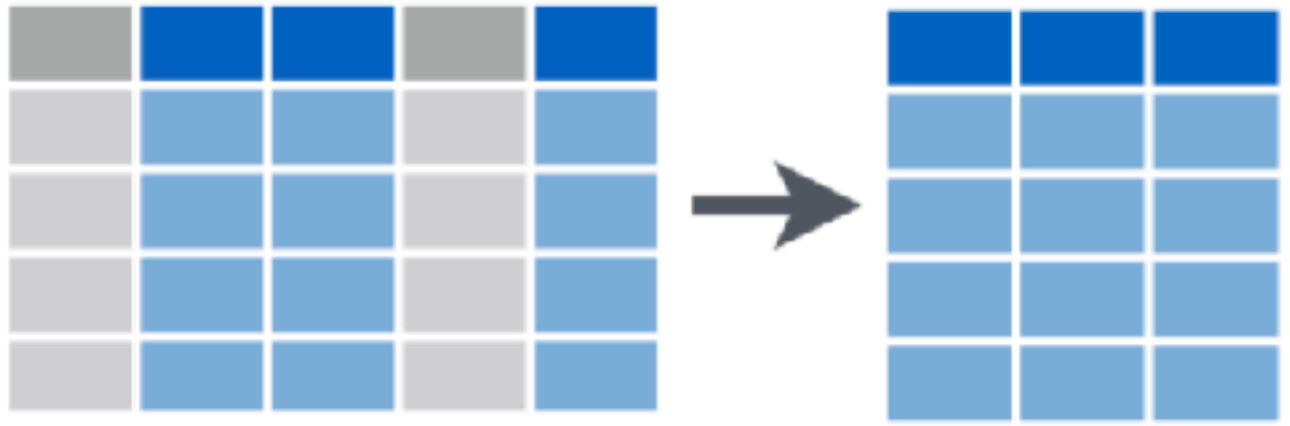


FIGURE 3.9: Diagram of select() columns.

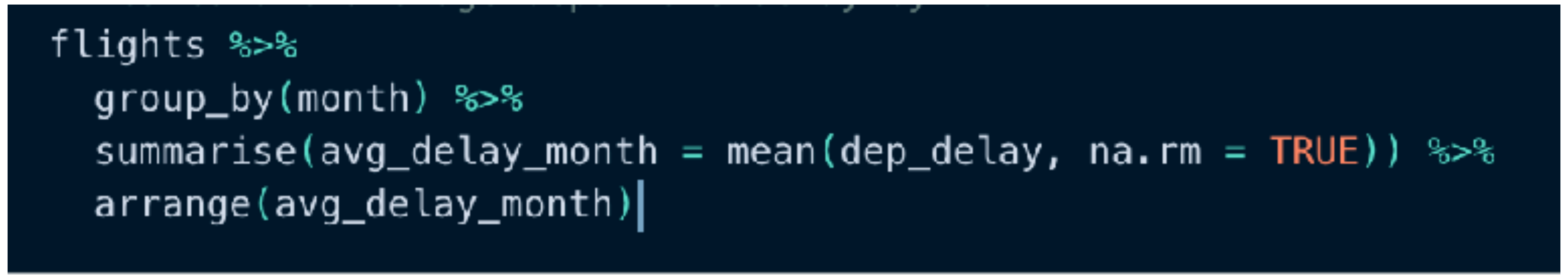
```
flights %>%
  select(carrier, flight)
```

```
flights_no_year <- flights %>%
  select(-year)
```

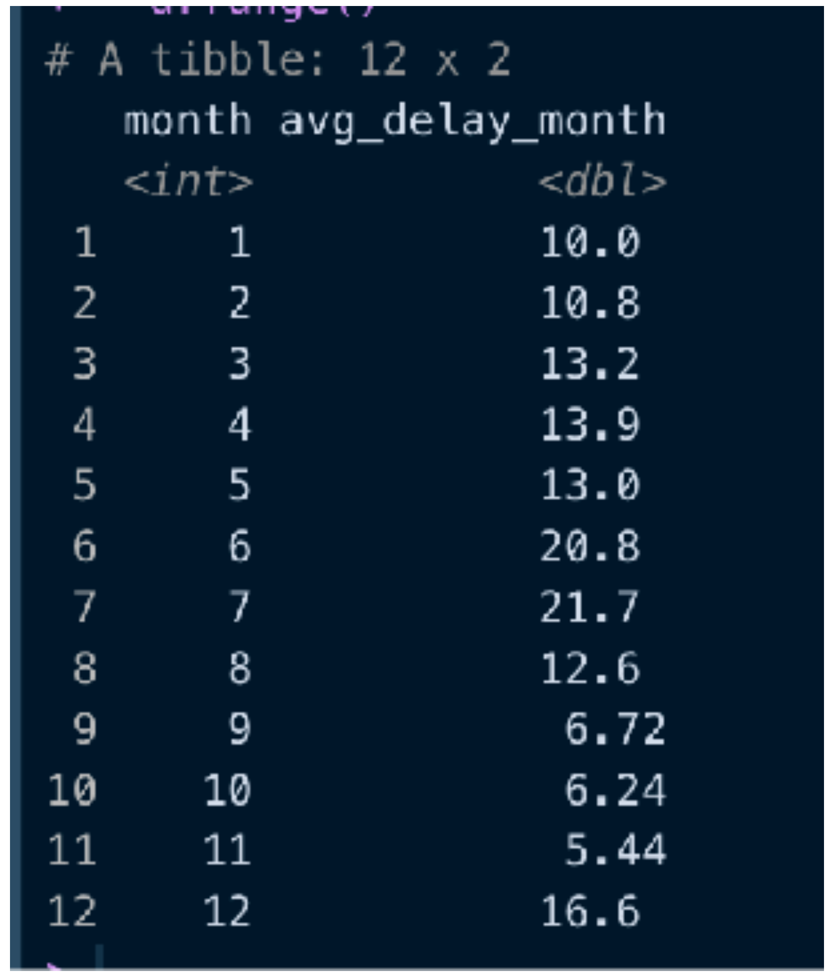
arrange()

Sort data frames by one or more variables

Increasing

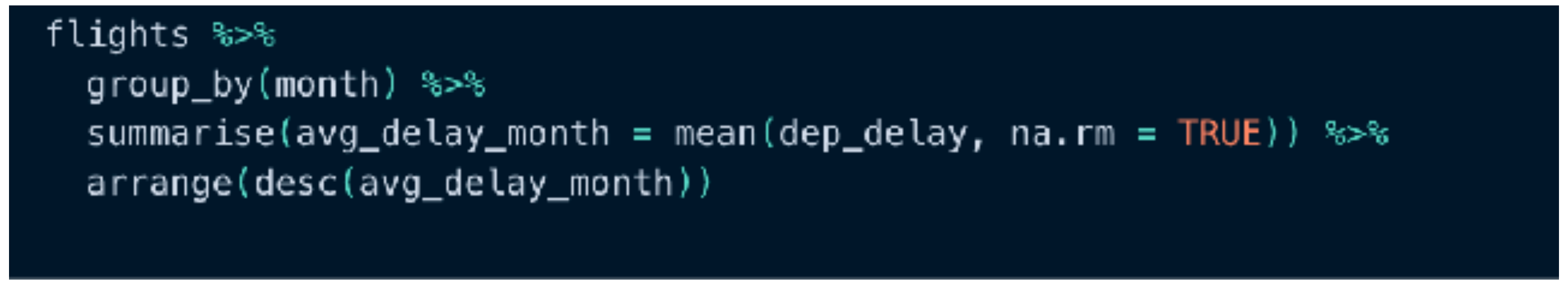


```
flights %>%
  group_by(month) %>%
  summarise(avg_delay_month = mean(dep_delay, na.rm = TRUE)) %>%
  arrange(avg_delay_month)
```

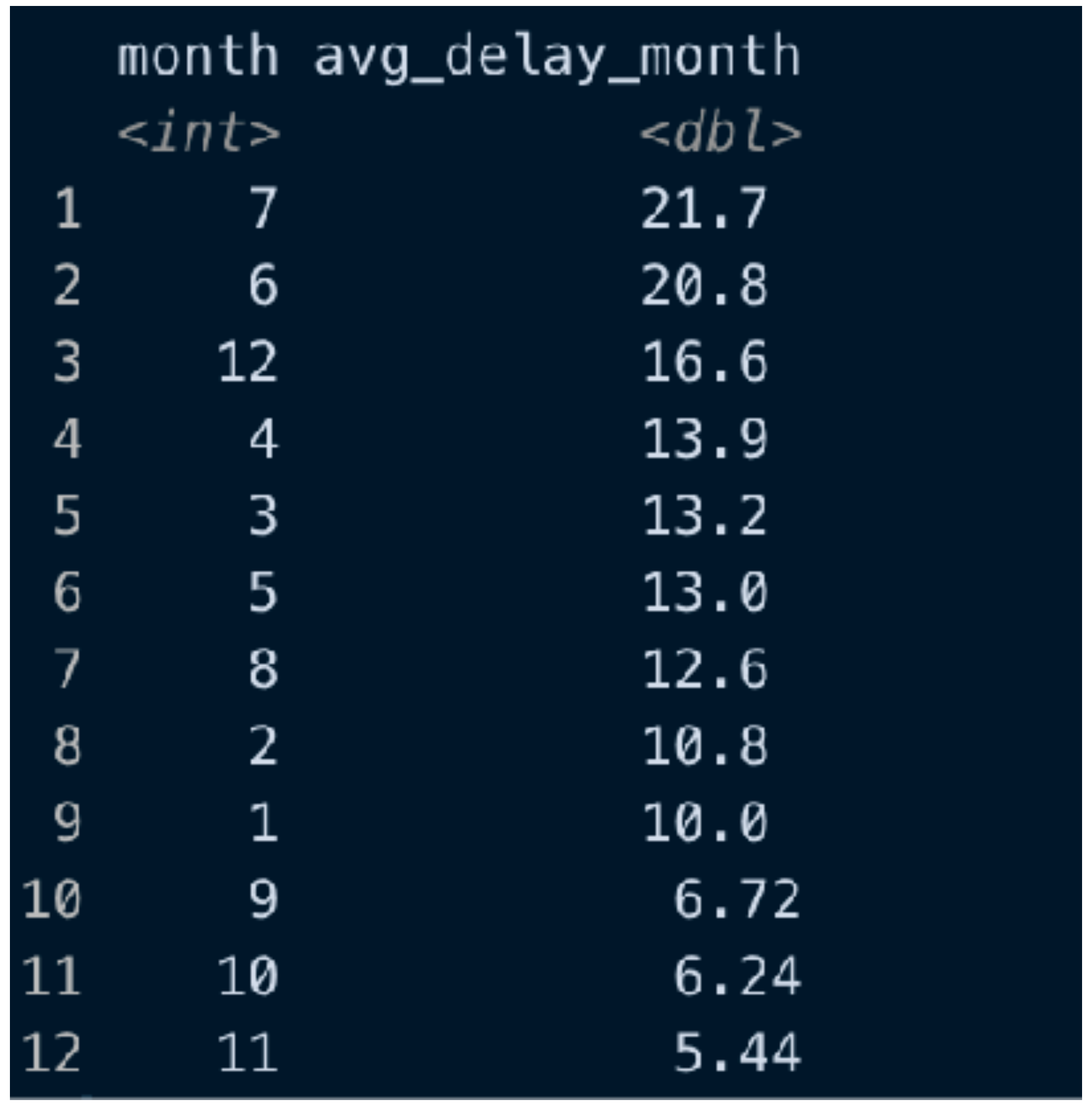


```
# A tibble: 12 x 2
   month avg_delay_month
   <int>           <dbl>
 1     1           10.0
 2     2           10.8
 3     3           13.2
 4     4           13.9
 5     5           13.0
 6     6           20.8
 7     7           21.7
 8     8           12.6
 9     9            6.72
10    10            6.24
11    11            5.44
12    12           16.6
```

Decreasing [desc()]



```
flights %>%
  group_by(month) %>%
  summarise(avg_delay_month = mean(dep_delay, na.rm = TRUE)) %>%
  arrange(desc(avg_delay_month))
```



```
   month avg_delay_month
   <int>           <dbl>
 1     7           21.7
 2     6           20.8
 3    12           16.6
 4     4           13.9
 5     3           13.2
 6     5           13.0
 7     8           12.6
 8     2           10.8
 9     1           10.0
10     9            6.72
11    10            6.24
12    11            5.44
```

Housekeeping

New Homework!
It's more difficult

Come to OH