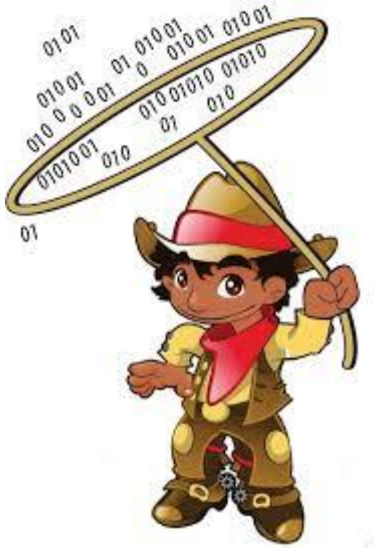
Data wrangling verbs

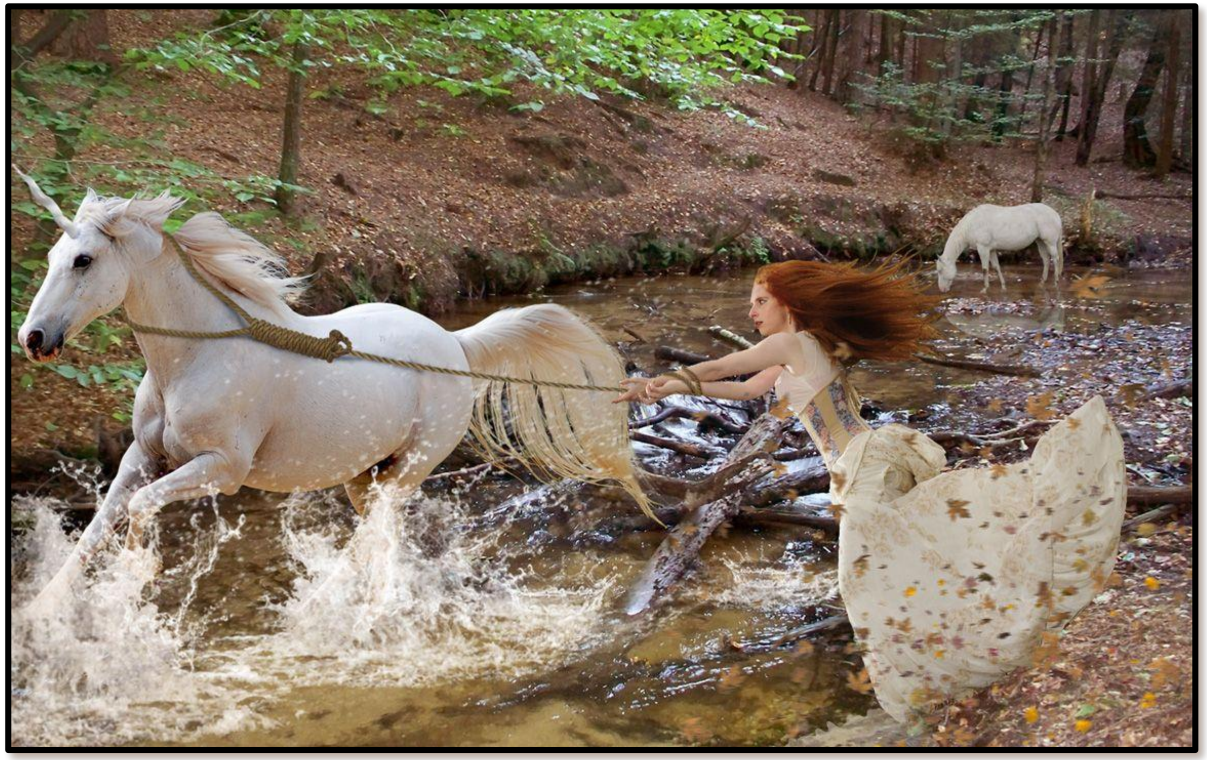
*in the **dplyr** package*

& the pipe operator

%>%

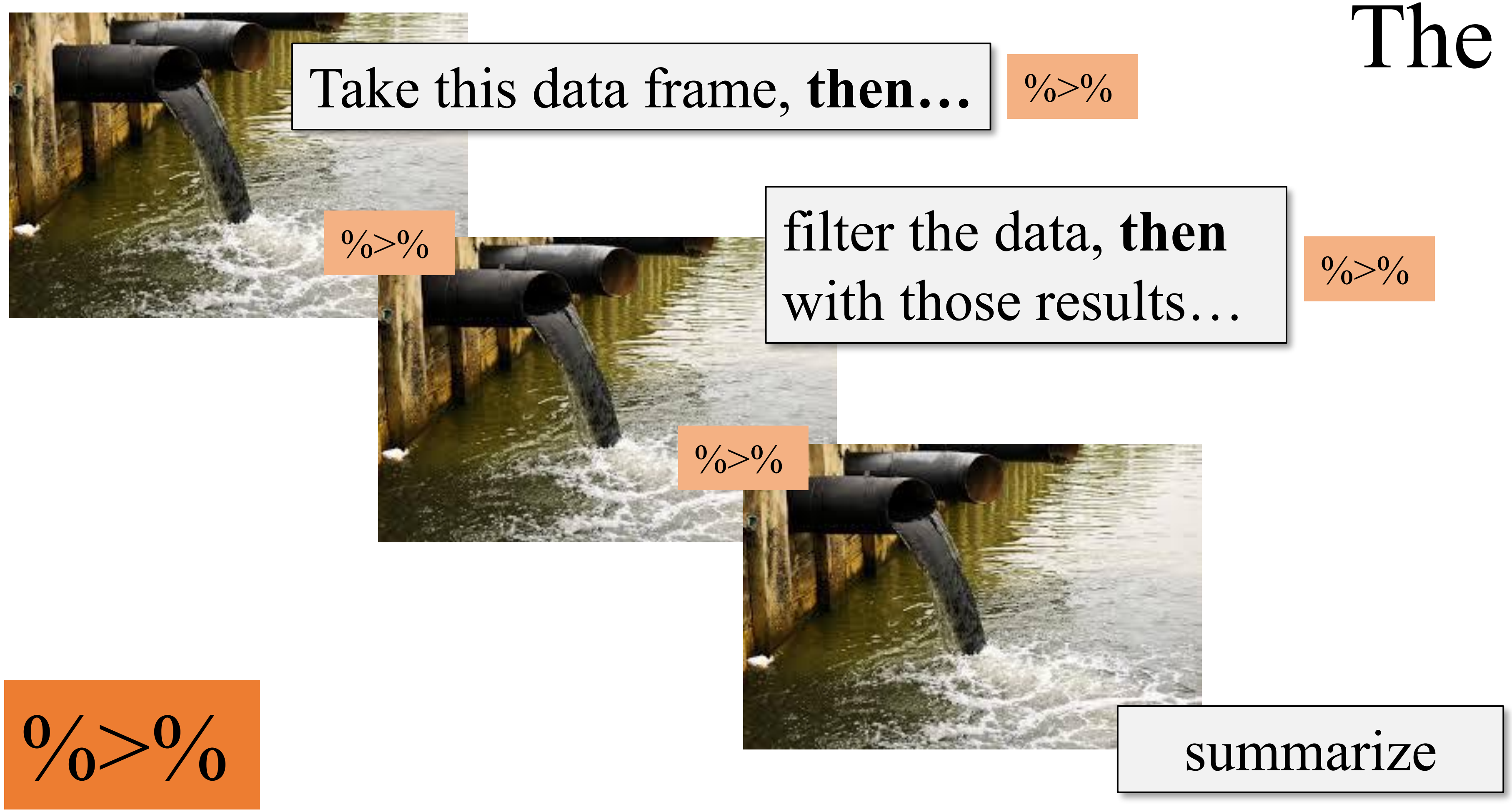
select()
filter()
summarize()
group_by()
mutate()
arrange()





The pipe operator

%>%



Take this data frame, **then…**

%>%

%>%

filter the data, **then** with those results…

%>%

%>%

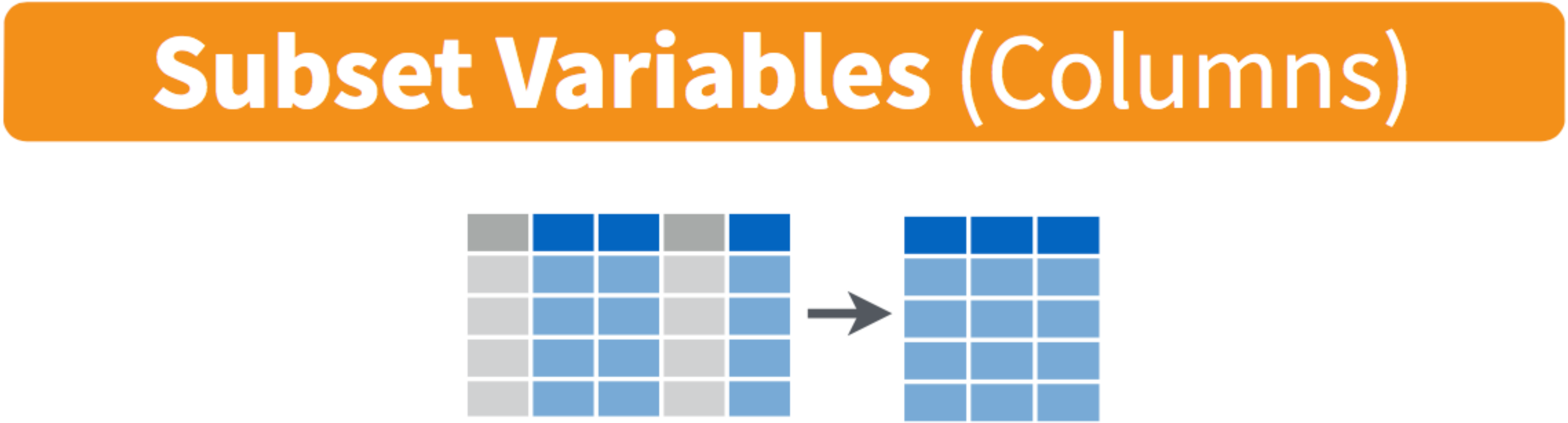
summarize

%>%

- “Pipe” a data frame into a “verb” command
- “Chain” the results from one “verb” command into another
- Think of it as the word **“then”**

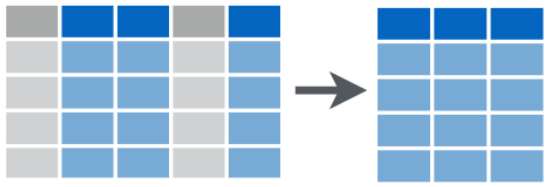
select()

Pull out just the **columns** you want in a data set, based on the column names



select()

Subset Variables (Columns)



Example

the data frame

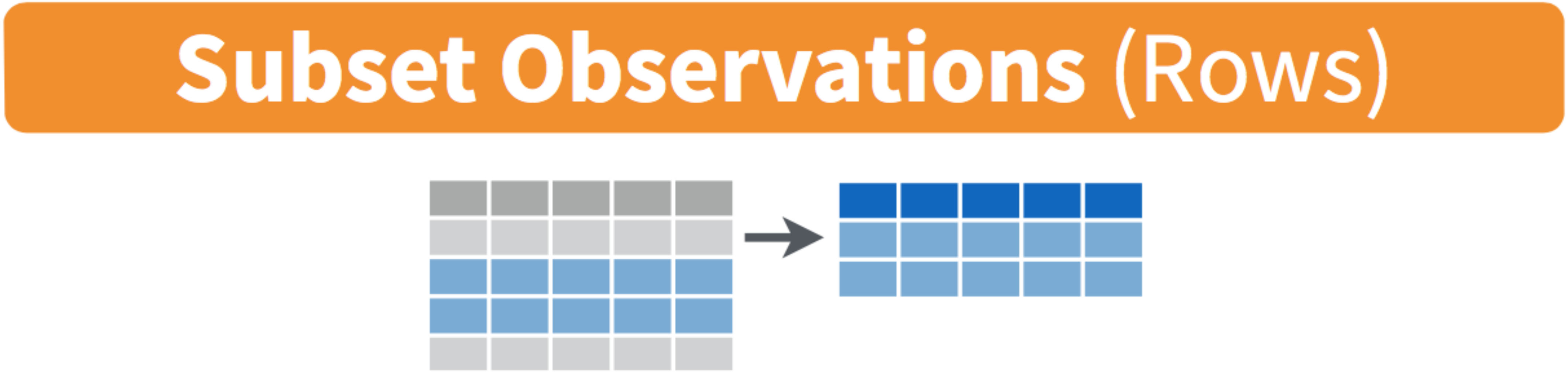
```
flights_sub<- flights %>%
     select(arr_delay, dep_delay)
```

the verb (function)

what columns (variable) to keep

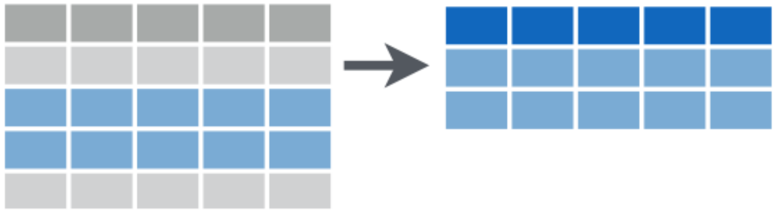
filter()

Pull out just the **rows** you want in a data set, based on some criteria



filter()

Subset Observations (Rows)



Only keep rows of the data in which the column meets this criteria

the data frame

Example

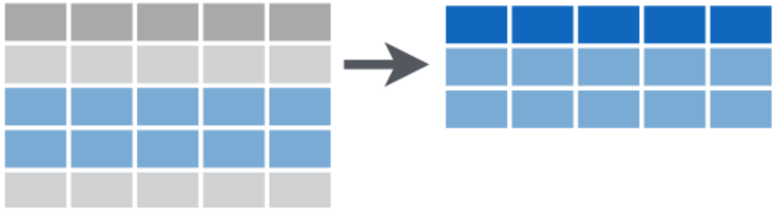
```
on_time_flights<- flights %>%
                         filter(arr_delay < 30)
```

the verb (function)

what column (variable) to use for filtering

filter()

Subset Observations (Rows)



&

|

Only keep rows of the data in which the column meets this criteria

the data frame

Another example

```
on_time_flights <- flights %>%
                   filter(origin == “AK”))
```

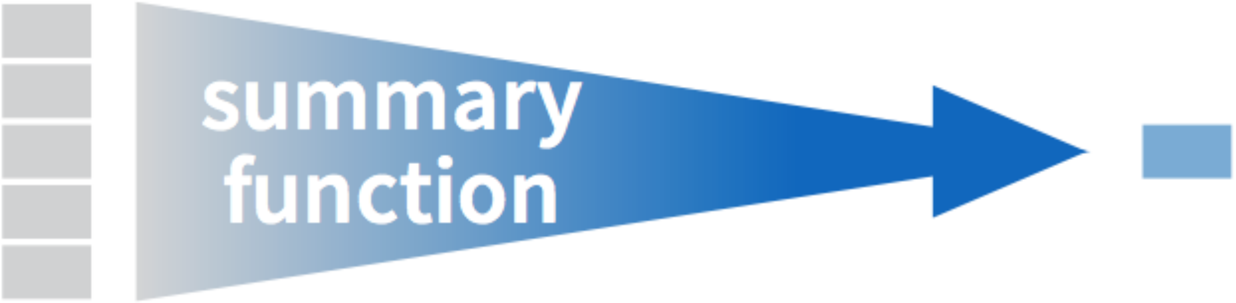
the verb (function)

what column (variable) to use for filtering

==

!=

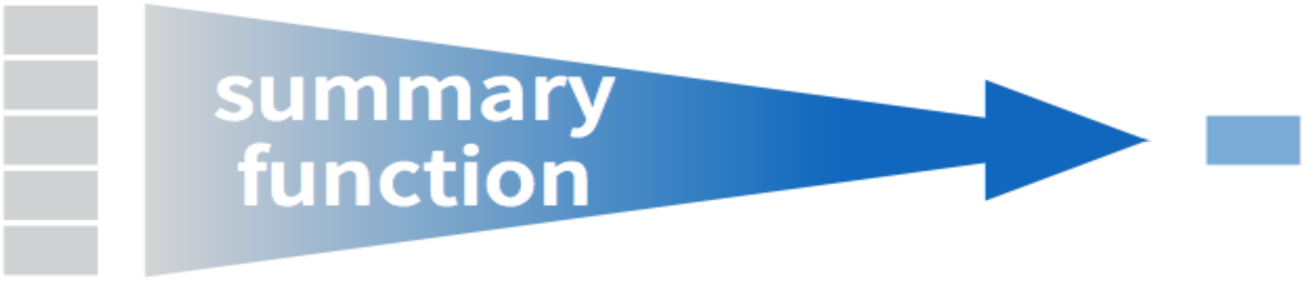
summarize()



Take a column of data from a data frame and reduce it down to a single summary statistic

Summarise Data

summarize()



Example

the data frame

what summary statistic to calculate

```
flights %>%
         summarize(av_dist = mean(distance))
```

the verb (function)

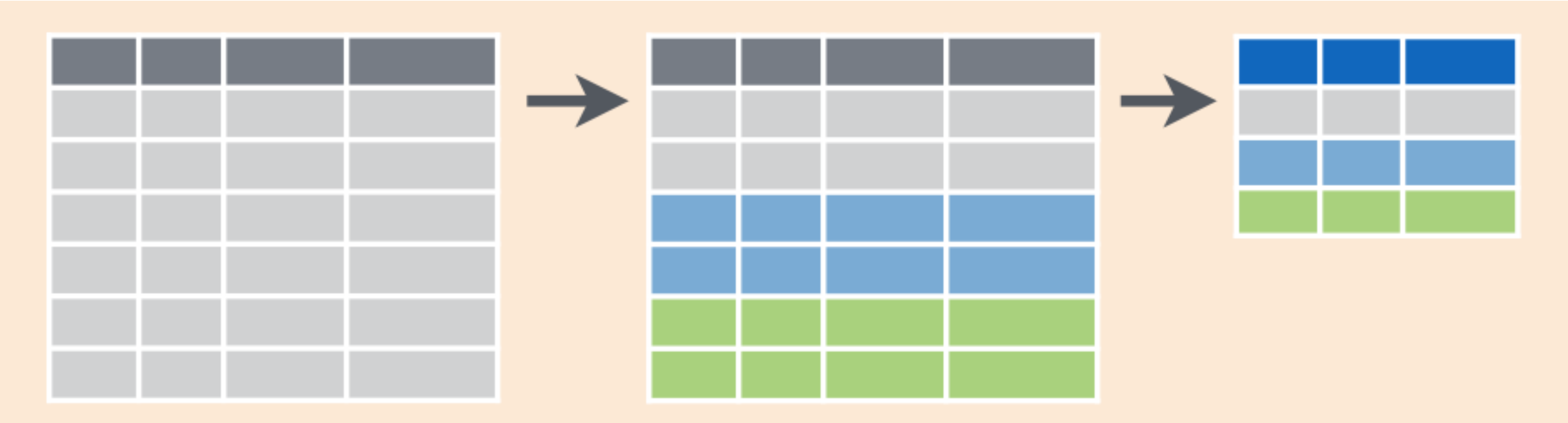
what to call the result

the column to summarize

group_by()

then

summarize()

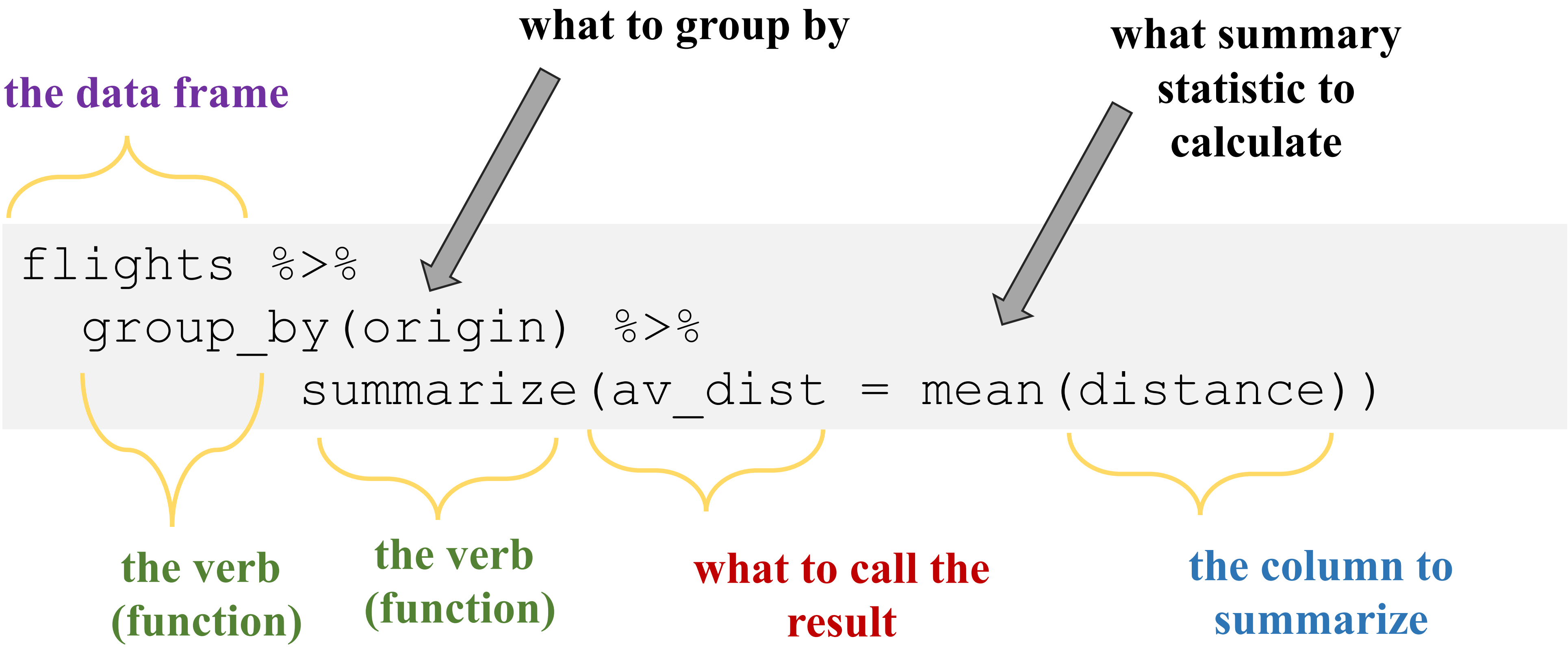


Take a column of data and reduce it down to a summary statistic, by some grouping variable

summarize()

summary
function

Example



```
flights %>%
  group_by(origin) %>%
         summarize(av_dist = mean(distance))
```

the data frame

what to group by

what summary statistic to calculate

the verb (function)

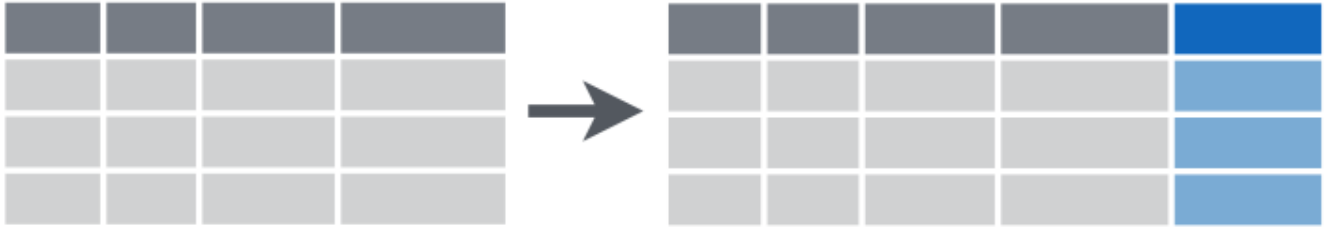
the verb (function)

what to call the result

the column to summarize

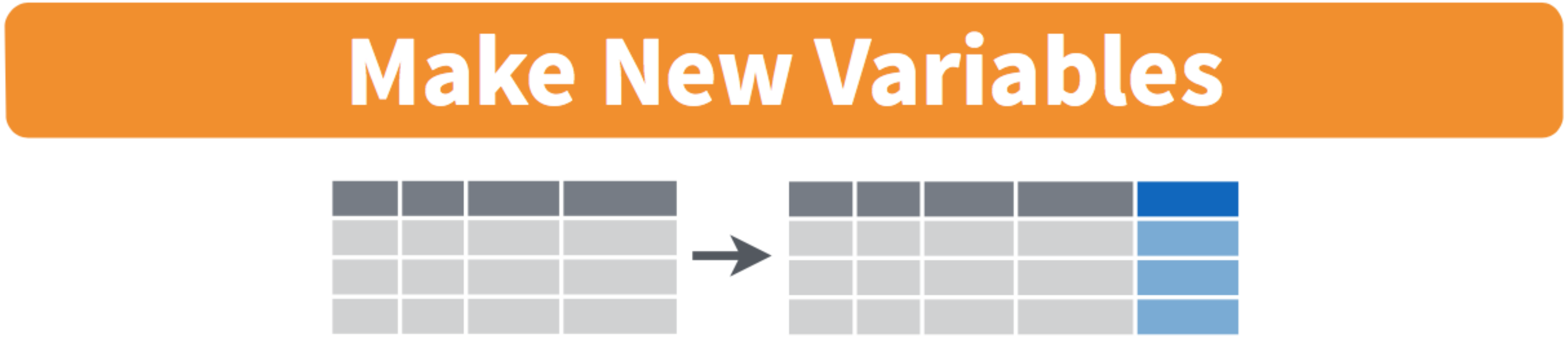
mutate()

Make New Variables



Mutant growth on a tomato

mutate()



Example

the data frame

what to call the new column (new variable)

```
flights %>%
    mutate(dist_ft = distance * 5280)
```

the verb (function)

what to put in the new column