Tidy data

“Tidy datasets are all alike but every messy dataset is messy in its own way” — Hadley Wickham

Tidy data

Three rules:

1. Each variable forms a column
2. Each observation forms a row
3. Each type of observational unit forms a table

Example: Contingency table

	survived	**died**
drug	15	3
placebo	4	12

not tidy

Example: Contingency table

	survived	died
drug	15	3
placebo	4	12

not tidy

tidy

treatment	outcome	count
drug	survived	15
drug	died	3
placebo	survived	4
placebo	died	12

Example: Contingency table

	survived	died
drug	15	3
placebo	4	12

not tidy

tidy

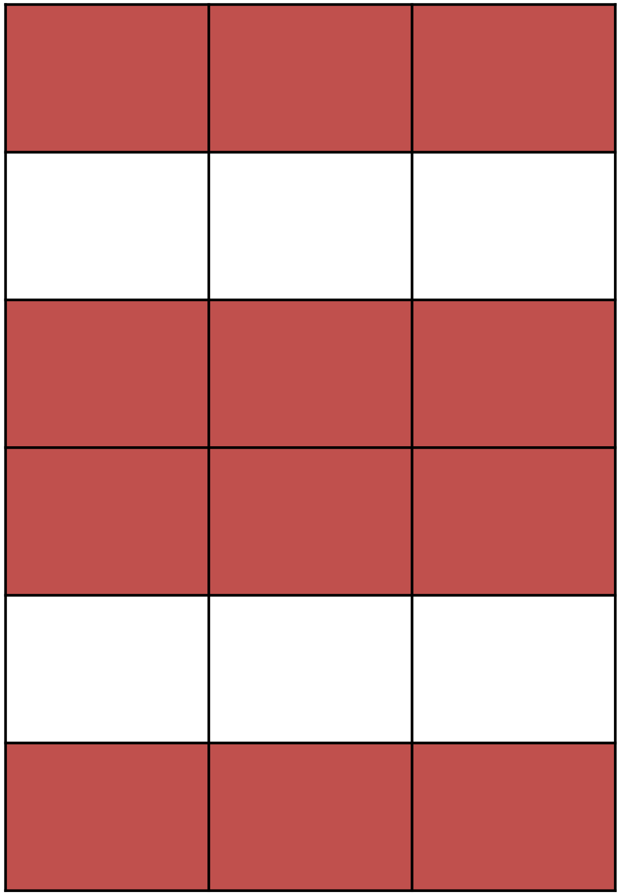
patient	treatment	outcome
1	drug	survived
2	drug	died
3	drug	survived
4	placebo	died
	⋮	

Working with tidy data in R: dplyr

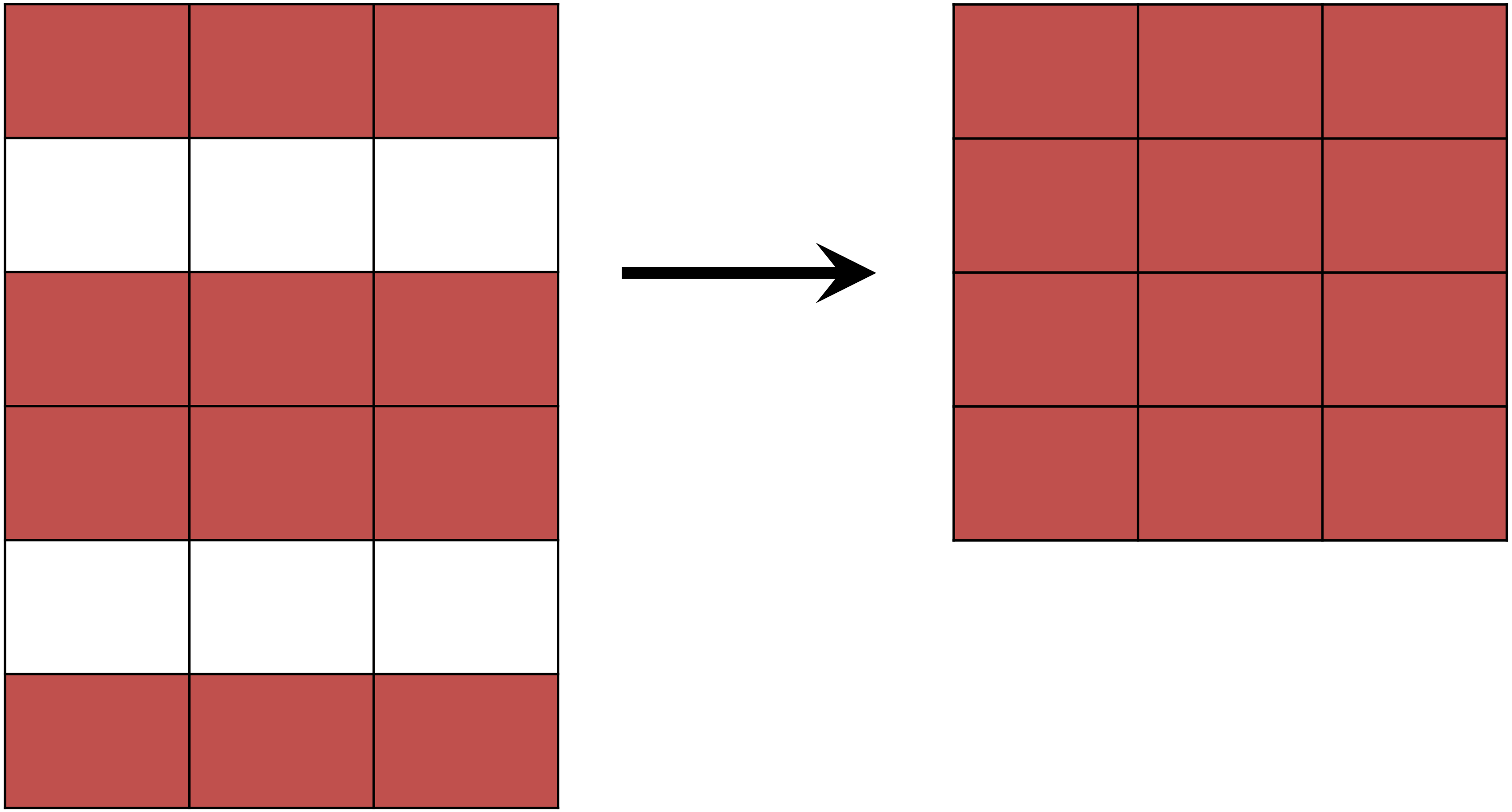
Fundamental actions on data tables:

- choose rows — `filter()`
- choose columns — `select()`
- make new columns — `mutate()`
- arrange rows — `arrange()`
- calculate summary statistics — `summarize()`
- work on groups of data — `group_by()`

`filter()`: pick rows



`filter()`: pick rows



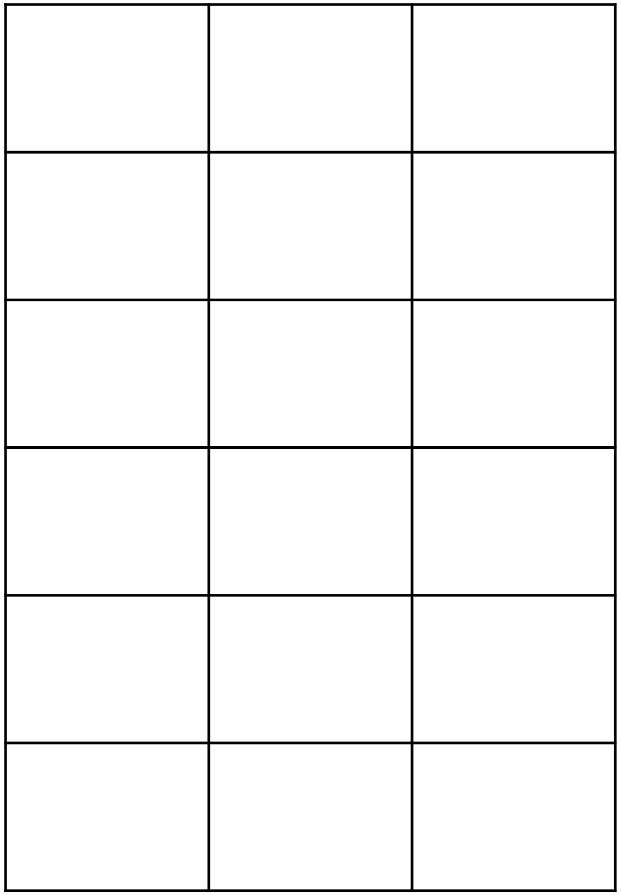
Choose rows with Sepal.Width > 4

```
> filter(iris, Sepal.Width > 4)
```

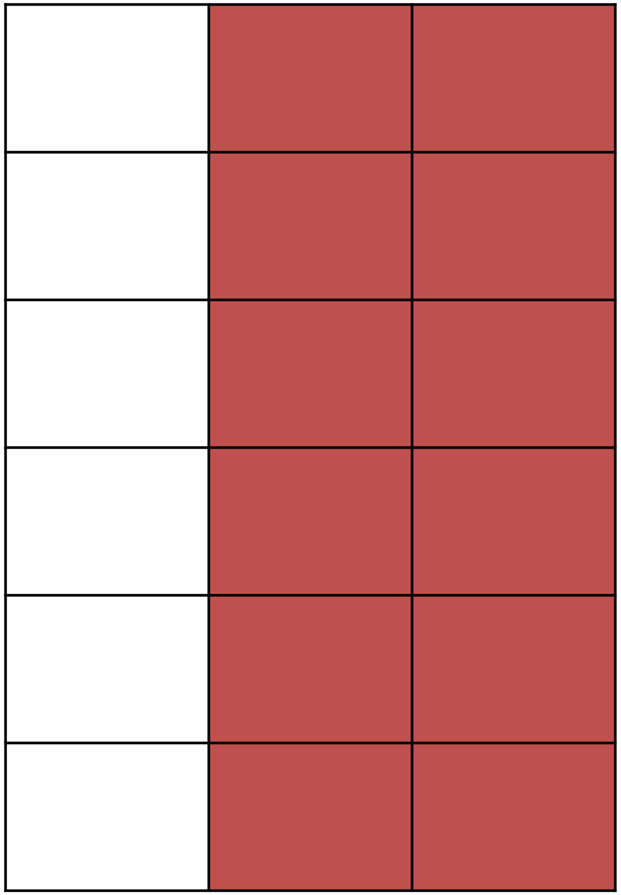
Choose rows with Sepal.Width > 4

```
> filter(iris, Sepal.Width > 4)
  Sepal.Length Sepal.Width Petal.Length Petal.Width Species
1          5.7         4.4          1.5         0.4  setosa
2          5.2         4.1          1.5         0.1  setosa
3          5.5         4.2          1.4         0.2  setosa
```

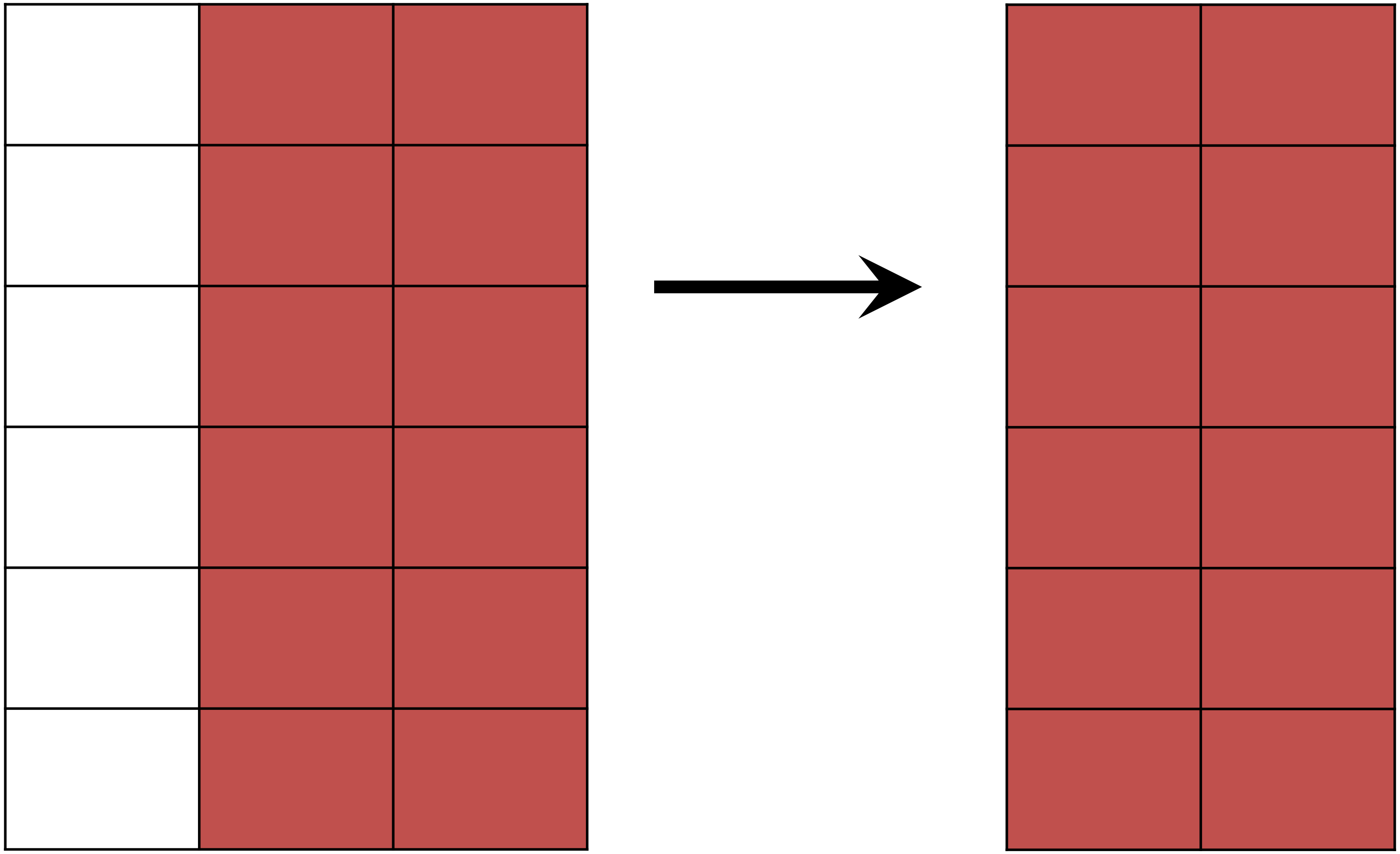
`select()`: pick columns



`select()`: pick columns



`select()`: pick columns



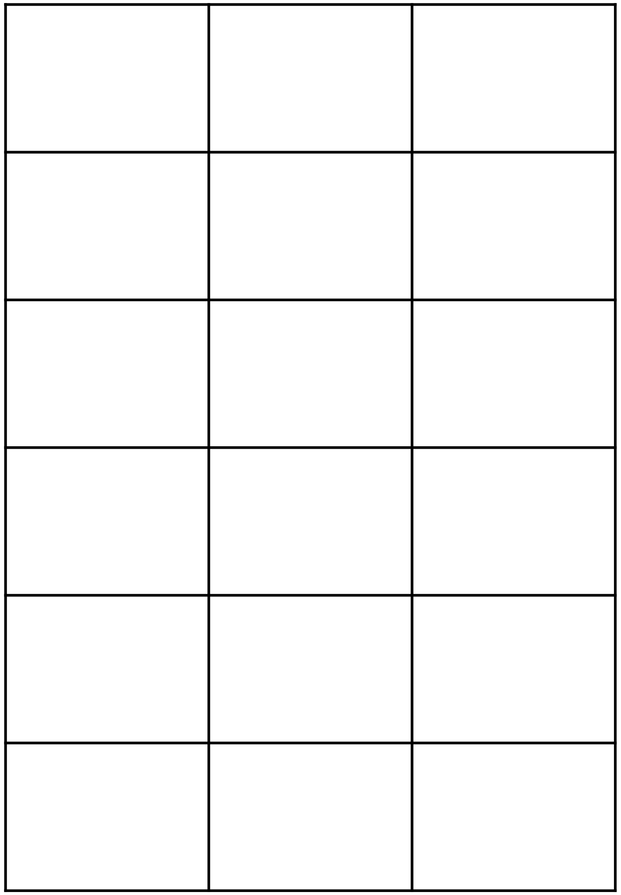
Choose the two columns Species and Sepal.Width

```
> select(iris, Species, Sepal.Width)
```

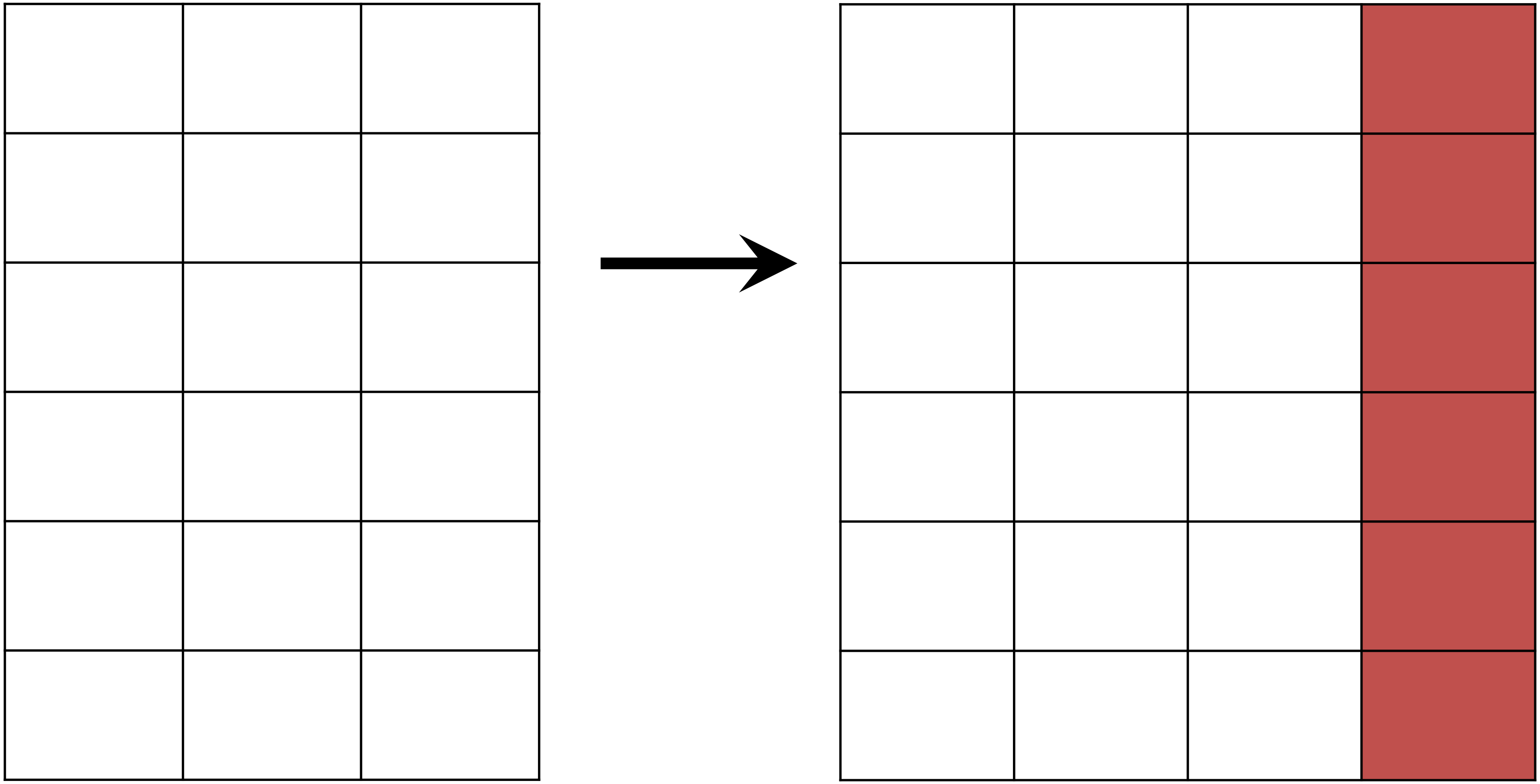
Choose the two columns Species and Sepal.Width

```
> select(iris, Species, Sepal.Width)
       Species Sepal.Width
1       setosa         3.5
2       setosa         3.0
3       setosa         3.2
4       setosa         3.1
5       setosa         3.6
6       setosa         3.9
7       setosa         3.4
8       setosa         3.4
9       setosa         2.9
10      setosa         3.1
11      setosa         3.7
12      setosa         3.4
13      setosa         3.0
14      setosa         3.0
```

`mutate()`: make new columns



`mutate()`: make new columns



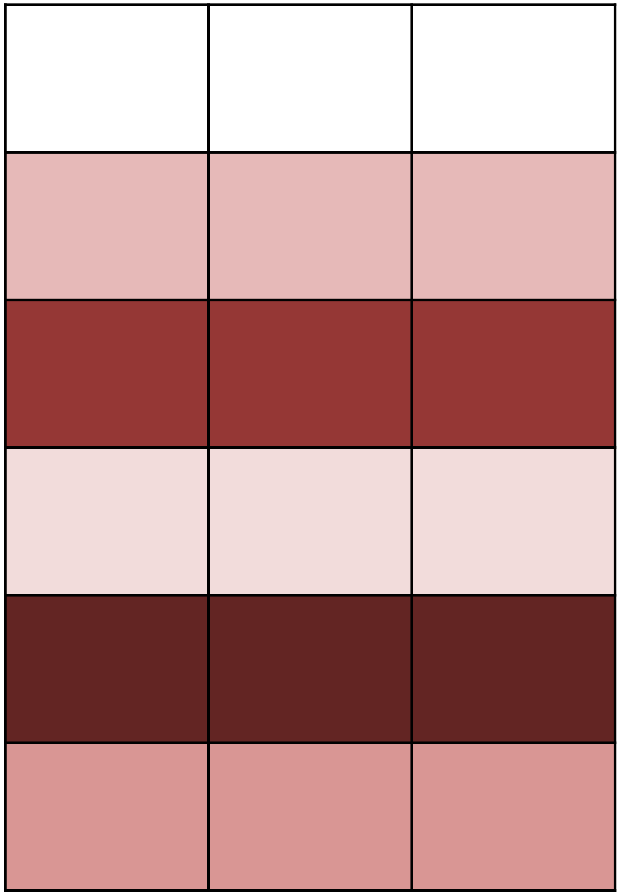
Make new column with ratio of Sepal.Length to Sepal.Width

```
> mutate(iris, Sepal.Length.to.Width = Sepal.Length/Sepal.Width)
```

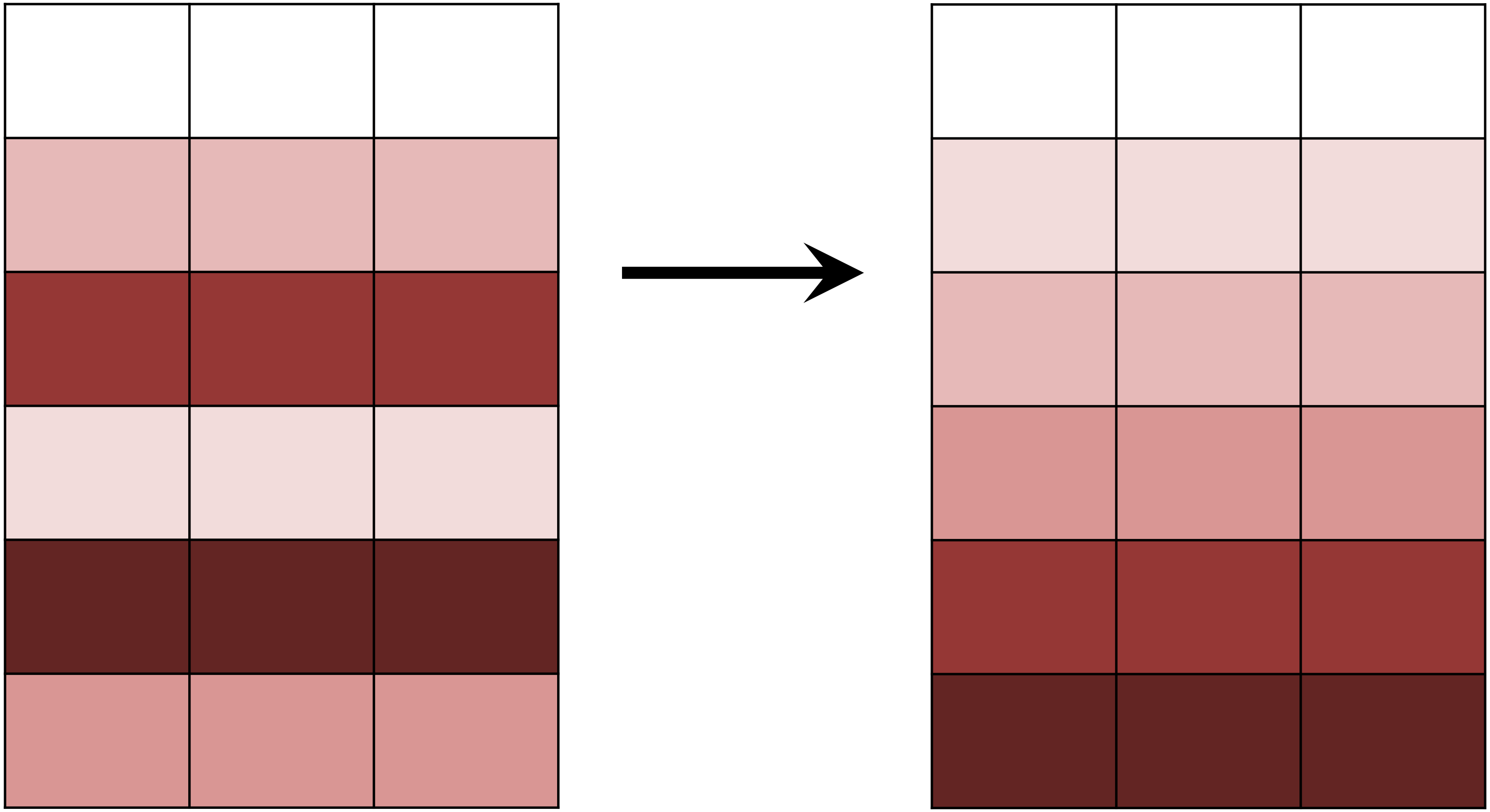
Make new column with ratio of Sepal.Length to Sepal.Width

```
> mutate(iris, Sepal.Length.to.Width = Sepal.Length/Sepal.Width)
    Sepal.Length Sepal.Width Petal.Length Petal.Width    Species Sepal.Length.to.Width
1            5.1         3.5          1.4         0.2     setosa              1.457143
2            4.9         3.0          1.4         0.2     setosa              1.633333
3            4.7         3.2          1.3         0.2     setosa              1.468750
4            4.6         3.1          1.5         0.2     setosa              1.483871
5            5.0         3.6          1.4         0.2     setosa              1.388889
6            5.4         3.9          1.7         0.4     setosa              1.384615
7            4.6         3.4          1.4         0.3     setosa              1.352941
8            5.0         3.4          1.5         0.2     setosa              1.470588
9            4.4         2.9          1.4         0.2     setosa              1.517241
10           4.9         3.1          1.5         0.1     setosa              1.580645
11           5.4         3.7          1.5         0.2     setosa              1.459459
12           4.8         3.4          1.6         0.2     setosa              1.411765
13           4.8         3.0          1.4         0.1     setosa              1.600000
14           4.3         3.0          1.1         0.1     setosa              1.433333
15           5.8         4.0          1.2         0.2     setosa              1.450000
16           5.7         4.4          1.5         0.4     setosa              1.295455
17           5.4         3.9          1.3         0.4     setosa              1.384615
18           5.1         3.5          1.4         0.3     setosa              1.457143
19           5.7         3.8          1.7         0.3     setosa              1.500000
20           5.1         3.8          1.5         0.3     setosa              1.342105
```

arrange(): change row order



`arrange()`: change row order



Sort by increasing order of Sepal.Width

```
> arrange(iris, Sepal.Width)
```

Sort by increasing order of Sepal.Width

```
> arrange(iris, Sepal.Width)
   Sepal.Length Sepal.Width Petal.Length Petal.Width    Species
1           5.0         2.0          3.5         1.0 versicolor
2           6.0         2.2          4.0         1.0 versicolor
3           6.2         2.2          4.5         1.5 versicolor
4           6.0         2.2          5.0         1.5  virginica
5           4.5         2.3          1.3         0.3     setosa
6           5.5         2.3          4.0         1.3 versicolor
7           6.3         2.3          4.4         1.3 versicolor
8           5.0         2.3          3.3         1.0 versicolor
9           4.9         2.4          3.3         1.0 versicolor
10          5.5         2.4          3.8         1.1 versicolor
11          5.5         2.4          3.7         1.0 versicolor
12          5.6         2.5          3.9         1.1 versicolor
13          6.3         2.5          4.9         1.5 versicolor
14          5.5         2.5          4.0         1.3 versicolor
```

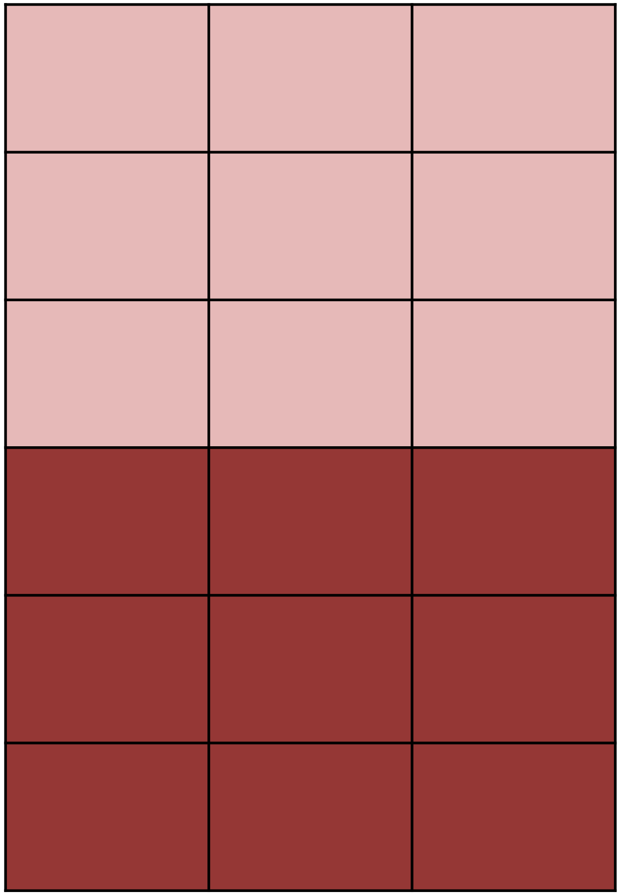
Sort by decreasing order of Sepal.Length

```
> arrange(iris, desc(Sepal.Length))
```

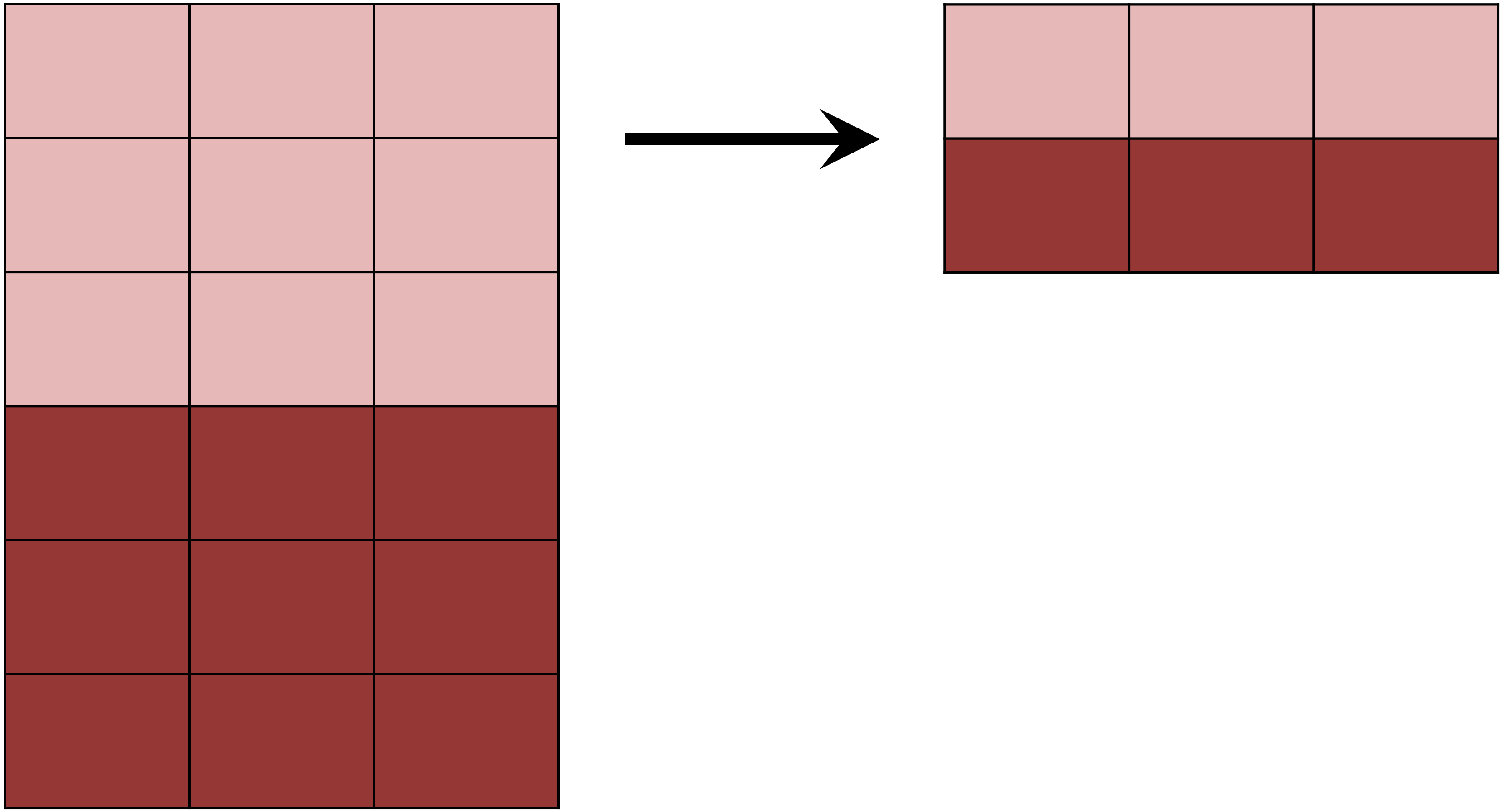
Sort by decreasing order of Sepal.Length

```
> arrange(iris, desc(Sepal.Length))
    Sepal.Length Sepal.Width Petal.Length Petal.Width    Species
1            7.9         3.8          6.4         2.0  virginica
2            7.7         3.8          6.7         2.2  virginica
3            7.7         2.6          6.9         2.3  virginica
4            7.7         2.8          6.7         2.0  virginica
5            7.7         3.0          6.1         2.3  virginica
6            7.6         3.0          6.6         2.1  virginica
7            7.4         2.8          6.1         1.9  virginica
8            7.3         2.9          6.3         1.8  virginica
9            7.2         3.6          6.1         2.5  virginica
10           7.2         3.2          6.0         1.8  virginica
11           7.2         3.0          5.8         1.6  virginica
12           7.1         3.0          5.9         2.1  virginica
13           7.0         3.2          4.7         1.4 versicolor
14           6.9         3.1          4.9         1.5 versicolor
```

`summarize()`: collapse multiple rows



`summarize()`: collapse multiple rows



Calculate mean and standard deviation of Sepal.Length

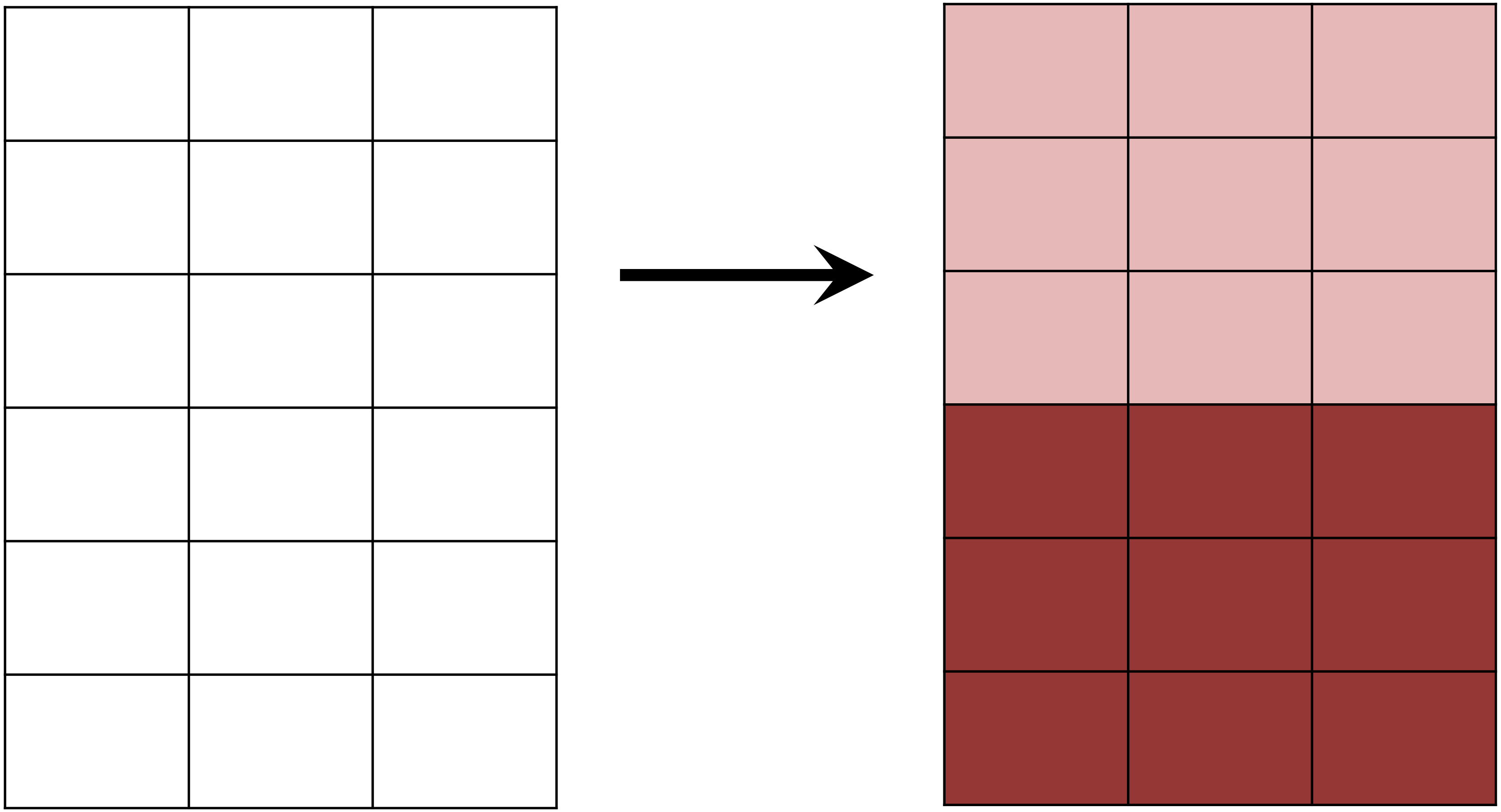
```
> summarize(iris, mean.sepal.length = mean(Sepal.Length),
                  sd.sepal.length   = sd(Sepal.Length))
```

Calculate mean and standard deviation of Sepal.Length

```
> summarize(iris, mean.sepal.length = mean(Sepal.Length),
                  sd.sepal.length   = sd(Sepal.Length))
  mean.sepal.length sd.sepal.length
1          5.843333       0.8280661
```

`group_by()`: set up groupings

`group_by()`: set up groupings



Calculate mean and standard deviation of Sepal.Length, grouped by Species

```
> summarize(group_by(iris, Species),
                  mean.sepal.length = mean(Sepal.Length),
                  sd.sepal.length   = sd(Sepal.Length))
```

Calculate mean and standard deviation of Sepal.Length, grouped by Species

```
> summarize(group_by(iris, Species),
                  mean.sepal.length = mean(Sepal.Length),
                  sd.sepal.length   = sd(Sepal.Length))
Source: local data frame [3 x 3]

     Species mean.sepal.length sd.sepal.length
1     setosa             5.006       0.3524897
2 versicolor             5.936       0.5161711
3  virginica             6.588       0.6358796
```