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
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
Separate tables for different observational units

Table of individual people

<table>
<thead>
<tr>
<th>Age</th>
<th>Sex</th>
<th>City</th>
</tr>
</thead>
<tbody>
<tr>
<td>37</td>
<td>male</td>
<td>Houston</td>
</tr>
<tr>
<td>19</td>
<td>male</td>
<td>Houston</td>
</tr>
<tr>
<td>8</td>
<td>female</td>
<td>Austin</td>
</tr>
<tr>
<td>78</td>
<td>female</td>
<td>Dallas</td>
</tr>
</tbody>
</table>
Separate tables for different observational units

Table of individual people

<table>
<thead>
<tr>
<th>Age</th>
<th>Sex</th>
<th>City</th>
</tr>
</thead>
<tbody>
<tr>
<td>37</td>
<td>male</td>
<td>Houston</td>
</tr>
<tr>
<td>19</td>
<td>male</td>
<td>Houston</td>
</tr>
<tr>
<td>8</td>
<td>female</td>
<td>Austin</td>
</tr>
<tr>
<td>78</td>
<td>female</td>
<td>Dallas</td>
</tr>
</tbody>
</table>

Table of cities

<table>
<thead>
<tr>
<th>City</th>
<th>Area</th>
<th>Population</th>
</tr>
</thead>
<tbody>
<tr>
<td>Houston</td>
<td>608</td>
<td>2,239,558</td>
</tr>
<tr>
<td>Austin</td>
<td>307</td>
<td>912,791</td>
</tr>
<tr>
<td>Dallas</td>
<td>386</td>
<td>NA</td>
</tr>
<tr>
<td>San Antonio</td>
<td>NA</td>
<td>1,436,697</td>
</tr>
</tbody>
</table>
Working with tidy data in R: tidyverse

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()`
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()`
• combine tables — `left_join()`, ...
left_join(): combine two tables
left_join(): combine two tables
left_join(): missing values in 2\textsuperscript{nd} table are set to NA
left_join(): missing values in 2\textsuperscript{nd} table are set to NA
left_join(): values from 2nd table are duplicated where necessary
`left_join()`: values from 2\textsuperscript{nd} table are duplicated where necessary.
Example: Joining tables

Let’s extract two tables from msleep:
Let’s extract two tables from `msleep`:

```r
> order_table <- select(msleep, name, order)
> order_table

<table>
<thead>
<tr>
<th>name</th>
<th>order</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cheetah</td>
<td>Carnivora</td>
</tr>
<tr>
<td>Owl monkey</td>
<td>Primates</td>
</tr>
<tr>
<td>Mountain beaver</td>
<td>Rodentia</td>
</tr>
<tr>
<td>Greater short-tailed shrew</td>
<td>Soricomorpha</td>
</tr>
<tr>
<td>Cow</td>
<td>Artiodactyla</td>
</tr>
<tr>
<td>Three-toed sloth</td>
<td>Pilosa</td>
</tr>
<tr>
<td>Northern fur seal</td>
<td>Carnivora</td>
</tr>
<tr>
<td>Vesper mouse</td>
<td>Rodentia</td>
</tr>
<tr>
<td>Dog</td>
<td>Carnivora</td>
</tr>
<tr>
<td>Roe deer</td>
<td>Artiodactyla</td>
</tr>
</tbody>
</table>
```
Example: Joining tables

Let's extract two tables from msleep:

```r
> awake_table <- select(msleep, name, awake)
> awake_table

<table>
<thead>
<tr>
<th>name</th>
<th>awake</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cheetah</td>
<td>11.90</td>
</tr>
<tr>
<td>Owl monkey</td>
<td>7.00</td>
</tr>
<tr>
<td>Mountain beaver</td>
<td>9.60</td>
</tr>
<tr>
<td>Greater short-tailed shrew</td>
<td>9.10</td>
</tr>
<tr>
<td>Cow</td>
<td>20.00</td>
</tr>
<tr>
<td>Three-toed sloth</td>
<td>9.60</td>
</tr>
<tr>
<td>Northern fur seal</td>
<td>15.30</td>
</tr>
<tr>
<td>Vesper mouse</td>
<td>17.00</td>
</tr>
<tr>
<td>Dog</td>
<td>13.90</td>
</tr>
<tr>
<td>Roe deer</td>
<td>21.00</td>
</tr>
</tbody>
</table>
```
Example: Joining tables

And put them back together:

> left_join(order_table, awake_table)
Example: Joining tables

And put them back together:

```r
> left_join(order_table, awake_table)
Joining by: "name"

<table>
<thead>
<tr>
<th>name</th>
<th>order</th>
<th>awake</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cheetah</td>
<td>Carnivora</td>
<td>11.90</td>
</tr>
<tr>
<td>Owl monkey</td>
<td>Primates</td>
<td>7.00</td>
</tr>
<tr>
<td>Mountain beaver</td>
<td>Rodentia</td>
<td>9.60</td>
</tr>
<tr>
<td>Greater short-tailed shrew</td>
<td>Soricomorpha</td>
<td>9.10</td>
</tr>
<tr>
<td>Cow</td>
<td>Artiodactyla</td>
<td>20.00</td>
</tr>
<tr>
<td>Three-toed sloth</td>
<td>Pilosa</td>
<td>9.60</td>
</tr>
<tr>
<td>Northern fur seal</td>
<td>Carnivora</td>
<td>15.30</td>
</tr>
<tr>
<td>Vesper mouse</td>
<td>Rodentia</td>
<td>17.00</td>
</tr>
<tr>
<td>Dog</td>
<td>Carnivora</td>
<td>13.90</td>
</tr>
<tr>
<td>Roe deer</td>
<td>Artiodactyla</td>
<td>21.00</td>
</tr>
</tbody>
</table>
Several different join functions are available

- left_join()
- right_join()
- inner_join()
- semi_join()
- full_join()
- anti_join()