

Tidying Data



Data Tidying

untidy data

Australian Bureau of Statistics										
1800.0 Australian Marriage Law Postal Survey, 2017										
Released on 15 November 2017										
Table 5 Participation by Federal Electoral Division(a), Males and Age Gender apartheid										
Year NA	18-19 years	20-24 years	25-29 years	30-34 years	35-39 years	40-44 years	45-49 years	50-54 years	55-59 years	60-64 years
Ungia(c)	Total participants	292	1,058	1,465	1,653	1,515	1,516	1,710	1,730	1,753
Merged cells	Eligible participants	572	2,910	3,789	3,996	3,607	3,506	3,645	3,331	2,960
	Participation rate (%)	51.0	36.4	38.7	41.4	42.0	43.2	46.9	51.9	59.2
Solomon	Total participants	442	1,461	2,066	2,357	2,188	2,057	2,224	2,108	2,134
	Eligible participants	750	2,991	3,994	4,155	3,634	3,398	3,427	3,066	2,931
Northern Territory (Total)	Participation rate (%)	58.9	48.8	51.7	56.7	60.2	60.5	64.9	68.8	72.8
	Total participants	734	2,519	3,531	4,010	3,703	3,573	3,834	3,838	3,887
Australian Capital Territory Divisions	Eligible participants	1,322	5,901	7,783	8,151	7,241	6,904	7,072	6,397	5,891
	Participation rate (%)	55.5	42.7	45.4	49.2	51.1	51.8	55.6	60.0	66.0
Canberra(d)	Total participants	1,764	4,789	4,817	4,973	4,626	4,453	5,074	4,826	5,169
	Eligible participants	2,260	6,471	6,448	6,509	5,983	5,805	6,302	5,902	6,044
Fenner(e)	Participation rate (%)	78.1	74.0	74.7	76.4	77.3	76.7	80.5	81.8	85.5
	Total participants	1,477	4,687	5,178	5,786	6,025	5,463	5,191	4,208	3,948
NA Year	Eligible participants	1,904	6,354	7,121	7,822	7,960	7,155	6,480	5,206	4,692
	Participation rate (%)	77.6	73.8	72.7	74.0	75.7	76.4	80.1	80.8	84.1
Australia	Total participants	5,241	9,478	9,895	10,759	10,652	9,910	10,209	9,054	9,117
	Eligible participants	4,164	12,825	13,569	14,331	13,943	12,960	12,782	11,108	10,736
Total	Participation rate (%)	77.8	73.9	73.7	75.1	76.4	76.5	80.3	81.3	84.9
	Total participants	151,297	438,166	441,658	460,548	462,206	479,360	524,620	517,693	543,449
a) The Federal Electoral Divisions are current as at 24 August 2017	Eligible participants	201,439	635,909	646,916	665,250	656,446	660,841	693,850	659,150	664,720
	Participation rate (%)	75.1	68.9	68.3	69.2	70.4	72.5	75.6	78.5	81.8
Return of the table junk										
MS Excel or Die										

tidy data

1	area	gender	age	State	Area (sq km)	Eligible participants	Participation rate (%)	Total participants	Total Participants
2	Adelaide	Female	18-19 years	SA	76	1341	83.5	1120	1120
3	Adelaide	Female	20-24 years	SA	76	4620	81.2	3750	3750
4	Adelaide	Female	25-29 years	SA	76	4897	81.8	4004	4004
5	Adelaide	Female	30-34 years	SA	76	4784	79.8	3820	3820
6	Adelaide	Female	35-39 years	SA	76	4319	79	3411	3411
7	Adelaide	Female	40-44 years	SA	76	4310	80.6	3472	3472
8	Adelaide	Female	45-49 years	SA	76	4579	81.4	3728	3728
9	Adelaide	Female	50-54 years	SA	76	4475	84.7	3791	3791
10	Adelaide	Female	55-59 years	SA	76	4622	87.3	4033	4033
11	Adelaide	Female	60-64 years	SA	76	4342	89.3	3879	3879
12	Adelaide	Female	65-69 years	SA	76	3970	90.7	3602	3602
13	Adelaide	Female	70-74 years	SA	76	3009	90.3	2716	2716
14	Adelaide	Female	75-79 years	SA	76	2156	88.5	1908	1908
15	Adelaide	Female	80-84 years	SA	76	1673	85.1	1423	1423

data
→
wrangling

dplyr functions

```
# to install and load dplyr  
install.packages("dplyr")  
library(dplyr)
```

- `%>%` - pipe operator for chaining a sequence of operations
- `glimpse()` - get an overview of what's included in dataset
- `filter()` - filter rows
- `select()` - select, rename, and re-order columns
- `rename()` - rename columns
- `arrange()` - reorder rows
- `mutate()` - create a new column
- `group_by()` - group variables
- `summarize()` - summarize information within a dataset
- `left_join()` - combining data across data frame



tidyr functions

```
# to install and load tidyr  
install.packages("tidyr")  
library(tidyr)
```

- unite() - combine contents of two or more columns into a single column
- separate() - separate contents of a column into two or more columns



janitor functions

```
# to install and load janitor  
install.packages("janitor")  
library(janitor)
```

- `clean_names()` - clean names of a data frame
- `tabyl()` - get a helpful summary of a variable



skimr functions

```
# to install and load skimr  
install.packages("skimr")  
library(skimr)
```

- `skim()` - summarize a data frame



The pipe operator: %>%

If you want to: `A --> B`

In R:

Without the pipe operator **`B (A)`**
With the pipe operator **`A %>% B`**

Data frame `A`
Function `B ()`
Function `C ()`
Function `D ()`

If you want to: `A --> B --> C --> D`

In R:

Without the pipe operator **`D (C (B (A)))`**
With the pipe operator **`A %>% B %>% C %>% D`**

Filtering Data




```
> glimpse(msleep)
```

```
Observations: 83
```

```
Variables: 11
```

```
$ name      <chr> "Cheetah", "Owl monkey", "Mountain beaver", "Greater short-tailed shrew", "Cow", "Three...
$ genus     <chr> "Acinonyx", "Aotus", "Aplodontia", "Blarina", "Bos", "Bradypus", "Callorhinus", "Calomy...
$ vore      <chr> "carni", "omni", "herbi", "omni", "herbi", "herbi", "carni", NA, "carni", "herbi", "her...
$ order     <chr> "Carnivora", "Primates", "Rodentia", "Soricomorpha", "Artiodactyla", "Pilosa", "Carnivo...
$ conservation <chr> "lc", NA, "nt", "lc", "domesticated", NA, "vu", NA, "domesticated", "lc", "lc", "domest...
$ sleep_total <dbl> 12.1, 17.0, 14.4, 14.9, 4.0, 14.4, 8.7, 7.0, 10.1, 3.0, 5.3, 9.4, 10.0, 12.5, 10.3, 8.3...
$ sleep_rem  <dbl> NA, 1.8, 2.4, 2.3, 0.7, 2.2, 1.4, NA, 2.9, NA, 0.6, 0.8, 0.7, 1.5, 2.2, 2.0, 1.4, 3.1, ...
$ sleep_cycle <dbl> NA, NA, NA, 0.133, 0.667, 0.767, 0.383, NA, 0.333, NA, NA, 0.217, NA, 0.117, NA, NA, 0....
$ awake     <dbl> 11.9, 7.0, 9.6, 9.1, 20.0, 9.6, 15.3, 17.0, 13.9, 21.0, 18.7, 14.6, 14.0, 11.5, 13.7, 1...
$ brainwt   <dbl> NA, 0.01550, NA, 0.00029, 0.42300, NA, NA, NA, 0.07000, 0.09820, 0.11500, 0.00550, NA, ...
$ bodywt    <dbl> 50.000, 0.480, 1.350, 0.019, 600.000, 3.850, 20.490, 0.045, 14.000, 14.800, 33.500, 0.7...
```



There are:

```
> glimpse(msleep)
```

```
Observations: 83
```

```
Variables: 11
```

- 83 rows
- 11 columns

```
$ name      <chr> "Cheetah", "Owl monkey", "Mountain beaver", "Greater short-tailed shrew", "Cow", "Three...
$ genus     <chr> "Acinonyx", "Aotus", "Aplodontia", "Blarina", "Bos", "Bradypus", "Callorhinus", "Calomy...
$ vore      <chr> "carni", "omni", "herbi", "omni", "herbi", "herbi", "carni", NA, "carni", "herbi", "her...
$ order     <chr> "Carnivora", "Primates", "Rodentia", "Soricomorpha", "Artiodactyla", "Pilosa", "Carnivo...
$ conservation <chr> "lc", NA, "nt", "lc", "domesticated", NA, "vu", NA, "domesticated", "lc", "lc", "domest...
$ sleep_total <dbl> 12.1, 17.0, 14.4, 14.9, 4.0, 14.4, 8.7, 7.0, 10.1, 3.0, 5.3, 9.4, 10.0, 12.5, 10.3, 8.3...
$ sleep_rem  <dbl> NA, 1.8, 2.4, 2.3, 0.7, 2.2, 1.4, NA, 2.9, NA, 0.6, 0.8, 0.7, 1.5, 2.2, 2.0, 1.4, 3.1, ...
$ sleep_cycle <dbl> NA, NA, NA, 0.133, 0.667, 0.767, 0.383, NA, 0.333, NA, NA, 0.217, NA, 0.117, NA, NA, 0....
$ awake     <dbl> 11.9, 7.0, 9.6, 9.1, 20.0, 9.6, 15.3, 17.0, 13.9, 21.0, 18.7, 14.6, 14.0, 11.5, 13.7, 1...
$ brainwt   <dbl> NA, 0.01550, NA, 0.00029, 0.42300, NA, NA, NA, 0.07000, 0.09820, 0.11500, 0.00550, NA, ...
$ bodywt    <dbl> 50.000, 0.480, 1.350, 0.019, 600.000, 3.850, 20.490, 0.045, 14.000, 14.800, 33.500, 0.7...
```



There are:

```
> glimpse(msleep)
```

```
Observations: 83
```

```
Variables: 11
```

- 83 rows
- 11 columns

```
$ name      <chr> "Cheetah", "Owl monkey", "Mountain beaver", "Greater short-tailed shrew", "Cow", "Three...
$ genus     <chr> "Acinonyx", "Aotus", "Aplodontia", "Blarina", "Bos", "Bradypus", "Callorhinus", "Calomy...
$ vore      <chr> "carni", "omni", "herbi", "omni", "herbi", "herbi", "carni", NA, "carni", "herbi", "her...
$ order     <chr> "Carnivora", "Primates", "Rodentia", "Soricomorpha", "Artiodactyla", "Pilosa", "Carnivo...
$ conservation <chr> "lc", NA, "nt", "lc", "domesticated", NA, "vu", NA, "domesticated", "lc", "lc", "domest...
$ sleep_total <dbl> 12.1, 17.0, 14.4, 14.9, 4.0, 14.4, 8.7, 7.0, 10.1, 3.0, 5.3, 9.4, 10.0, 12.5, 10.3, 8.3...
$ sleep_rem  <dbl> NA, 1.8, 2.4, 2.3, 0.7, 2.2, 1.4, NA, 2.9, NA, 0.6, 0.8, 0.7, 1.5, 2.2, 2.0, 1.4, 3.1, ...
$ sleep_cycle <dbl> NA, NA, NA, 0.133, 0.667, 0.767, 0.383, NA, 0.333, NA, NA, 0.217, NA, 0.117, NA, NA, 0....
$ awake     <dbl> 11.9, 7.0, 9.6, 9.1, 20.0, 9.6, 15.3, 17.0, 13.9, 21.0, 18.7, 14.6, 14.0, 11.5, 13.7, 1...
$ brainwt   <dbl> NA, 0.01550, NA, 0.00029, 0.42300, NA, NA, NA, 0.07000, 0.09820, 0.11500, 0.00550, NA, ...
$ bodywt    <dbl> 50.000, 0.480, 1.350, 0.019, 600.000, 3.850, 20.490, 0.045, 14.000, 14.800, 33.500, 0.7...
```

The names of
the columns



```
> glimpse(msleep)
```

```
Observations: 83  
Variables: 11
```

```
$ name      <chr> "Cheetah", "Owl monkey", "Mountain beaver", "Greater short-tailed shrew", "Cow", "Three...  
$ genus     <chr> "Acinonyx", "Aotus", "Aplodontia", "Blarina", "Bos", "Bradypus", "Callorhinus", "Calomy...  
$ vore      <chr> "carni", "omni", "herbi", "omni", "herbi", "herbi", "carni", NA, "carni", "herbi", "her...  
$ order     <chr> "Carnivora", "Primates", "Rodentia", "Soricomorpha", "Artiodactyla", "Pilosa", "Carnivo...  
$ conservation <chr> "lc", NA, "nt", "lc", "domesticated", NA, "vu", NA, "domesticated", "lc", "lc", "domest...  
$ sleep_total <dbl> 12.1, 17.0, 14.4, 14.9, 4.0, 14.4, 8.7, 7.0, 10.1, 3.0, 5.3, 9.4, 10.0, 12.5, 10.3, 8.3...  
$ sleep_rem  <dbl> NA, 1.8, 2.4, 2.3, 0.7, 2.2, 1.4, NA, 2.9, NA, 0.6, 0.8, 0.7, 1.5, 2.2, 2.0, 1.4, 3.1, ...  
$ sleep_cycle <dbl> NA, NA, NA, 0.133, 0.667, 0.767, 0.383, NA, 0.333, NA, NA, 0.217, NA, 0.117, NA, NA, 0...  
$ awake     <dbl> 11.9, 7.0, 9.6, 9.1, 20.0, 9.6, 15.3, 17.0, 13.9, 21.0, 18.7, 14.6, 14.0, 11.5, 13.7, 1...  
$ brainwt   <dbl> NA, 0.01550, NA, 0.00029, 0.42300, NA, NA, NA, 0.07000, 0.09820, 0.11500, 0.00550, NA, ...  
$ bodywt    <dbl> 50.000, 0.480, 1.350, 0.019, 600.000, 3.850, 20.490, 0.045, 14.000, 14.800, 33.500, 0.7...
```

There are:

- 83 rows
- 11 columns

The first 5 columns are
character variables

The names of
the columns




```

> glimpse(msleep)
Observations: 83
Variables: 11
 $ name      <chr> "Cheetah", "Owl monkey", "Mountain beaver", "Greater short-tailed shrew", "Cow", "Three...
 $ genus     <chr> "Acinonyx", "Aotus", "Aplodontia", "Blarina", "Bos", "Bradypus", "Callorhinus", "Calomy...
 $ vore      <chr> "carni", "omni", "herbi", "omni", "herbi", "herbi", "carni", NA, "carni", "herbi", "her...
 $ order     <chr> "Carnivora", "Primates", "Rodentia", "Soricomorpha", "Artiodactyla", "Pilosa", "Carnivo...
 $ conservation <chr> "lc", NA, "nt", "lc", "domesticated", NA, "vu", NA, "domesticated", "lc", "lc", "domest...
 $ sleep_total <dbl> 12.1, 17.0, 14.4, 14.9, 4.0, 14.4, 8.7, 7.0, 10.1, 3.0, 5.3, 9.4, 10.0, 12.5, 10.3, 8.3...
 $ sleep_rem  <dbl> NA, 1.8, 2.4, 2.3, 0.7, 2.2, 1.4, NA, 2.9, NA, 0.6, 0.8, 0.7, 1.5, 2.2, 2.0, 1.4, 3.1, ...
 $ sleep_cycle <dbl> NA, NA, NA, 0.133, 0.667, 0.767, 0.383, NA, 0.333, NA, NA, 0.217, NA, 0.117, NA, NA, 0...
 $ awake     <dbl> 11.9, 7.0, 9.6, 9.1, 20.0, 9.6, 15.3, 17.0, 13.9, 21.0, 18.7, 14.6, 14.0, 11.5, 13.7, 1...
 $ brainwt   <dbl> NA, 0.01550, NA, 0.00029, 0.42300, NA, NA, NA, 0.07000, 0.09820, 0.11500, 0.00550, NA, ...
 $ bodywt    <dbl> 50.000, 0.480, 1.350, 0.019, 600.000, 3.850, 20.490, 0.045, 14.000, 14.800, 33.500, 0.7...

```

There are:

- 83 rows
- 11 columns

The first 5 columns are
character variables

The first three names of the
animals in the dataset

The names of
the columns



Equivalent to:

```
> msleep %>%  
+   filter(order == "Primates")
```

```
filter(msleep, order == "Primates")
```

```
# A tibble: 12 x 11
```

	name	genus	vore	order	conservation	sleep_total	sleep_rem	sleep_cycle	awake	brainwt	bodywt
	<chr>	<chr>	<chr>	<chr>	<chr>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
1	Owl monkey	Aotus	omni	Prima...	<NA>	17.0	1.80	NA	7.00	0.0155	0.480
2	Grivet	Cercopithe...	omni	Prima...	lc	10.0	0.700	NA	14.0	NA	4.75
3	Patas monkey	Erythroceb...	omni	Prima...	lc	10.9	1.10	NA	13.1	0.115	10.0
4	Galago	Galago	omni	Prima...	<NA>	9.80	1.10	0.550	14.2	0.00500	0.200
5	Human	Homo	omni	Prima...	<NA>	8.00	1.90	1.50	16.0	1.32	62.0
6	Mongoose lemur	Lemur	herbi	Prima...	vu	9.50	0.900	NA	14.5	NA	1.67
7	Macaque	Macaca	omni	Prima...	<NA>	10.1	1.20	0.750	13.9	0.179	6.80
8	Slow loris	Nyctibeus	carni	Prima...	<NA>	11.0	NA	NA	13.0	0.0125	1.40
9	Chimpanzee	Pan	omni	Prima...	<NA>	9.70	1.40	1.42	14.3	0.440	52.2
10	Baboon	Papio	omni	Prima...	<NA>	9.40	1.00	0.667	14.6	0.180	25.2
11	Potto	Perodictic...	omni	Prima...	lc	11.0	NA	NA	13.0	NA	1.10
12	Squirrel monkey	Saimiri	omni	Prima...	<NA>	9.60	1.40	NA	14.4	0.0200	0.743



```
> msleep %>%  
+   filter(order == "Primates", sleep_total > 10)
```

```
# A tibble: 5 x 11
```

	name	genus	vore	order	conservation	sleep_total	sleep_rem	sleep_cycle	awake	brainwt	bodywt
	<chr>	<chr>	<chr>	<chr>	<chr>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
1	Owl monkey	Aotus	omni	Primates	<NA>	17.0	1.80	NA	7.00	0.0155	0.480
2	Patas monkey	Erythrocebus	omni	Primates	lc	10.9	1.10	NA	13.1	0.115	10.0
3	Macaque	Macaca	omni	Primates	<NA>	10.1	1.20	0.750	13.9	0.179	6.80
4	Slow loris	Nyctibeus	carni	Primates	<NA>	11.0	NA	NA	13.0	0.0125	1.40
5	Potto	Perodicticus	omni	Primates	lc	11.0	NA	NA	13.0	NA	1.10

Gives the same results: `msleep %>%
 filter(order == "Primates" & sleep_total > 10)`



```
> msleep %>%  
+   filter(order == "Primates", sleep_total > 10) %>%  
+   select(name, sleep_total, sleep_rem, sleep_cycle)  
# A tibble: 5 x 4
```

	name	sleep_total	sleep_rem	sleep_cycle
	<chr>	<dbl>	<dbl>	<dbl>
1	Owl monkey	17.0	1.80	NA
2	Patas monkey	10.9	1.10	NA
3	Macaque	10.1	1.20	0.750
4	Slow loris	11.0	NA	NA
5	Potto	11.0	NA	NA

Without the pipe operator: `select(filter(msleep, order == "Primates",
sleep_total > 10), name, sleep_total, sleep_rem,
sleep_cycle)`




```
> msleep %>%  
+   filter(order == "Primates", sleep_total > 10) %>%  
+   select(name, total=sleep_total, rem=sleep_rem, cycle=sleep_cycle)  
# A tibble: 5 x 4
```

	name	total	rem	cycle
	<chr>	<dbl>	<dbl>	<dbl>
1	Owl monkey	17.0	1.80	NA
2	Patas monkey	10.9	1.10	NA
3	Macaque	10.1	1.20	0.750
4	Slow loris	11.0	NA	NA
5	Potto	11.0	NA	NA



```

> msleep %>%
+   filter(order == "Primates", sleep_total > 10) %>%
+   rename(total=sleep_total, rem=sleep_rem, cycle=sleep_cycle)

```

```
# A tibble: 5 x 11
```

	name	genus	vore	order	conservation	total	rem	cycle	awake	brainwt	bodywt
	<chr>	<chr>	<chr>	<chr>	<chr>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
1	Owl monkey	Aotus	omni	Primates	<NA>	17	1.8	NA	7	0.0155	0.48
2	Patas monkey	Erythrocebus	omni	Primates	lc	10.9	1.1	NA	13.1	0.115	10
3	Macaque	Macaca	omni	Primates	<NA>	10.1	1.2	0.75	13.9	0.179	6.8
4	Slow loris	Nyctibeus	carni	Primates	<NA>	11	NA	NA	13	0.0125	1.4
5	Potto	Perodicticus	omni	Primates	lc	11	NA	NA	13	NA	1.1



Reordering Data



```
> msleep %>%  
+   filter(order == "Primates", sleep_total > 10) %>%  
+   select(name, sleep_rem, sleep_cycle, sleep_total)  
# A tibble: 5 x 4
```

	name	sleep_rem	sleep_cycle	sleep_total
	<chr>	<dbl>	<dbl>	<dbl>
1	Owl monkey	1.80	NA	17.0
2	Patas monkey	1.10	NA	10.9
3	Macaque	1.20	0.750	10.1
4	Slow loris	NA	NA	11.0
5	Potto	NA	NA	11.0

```
> msleep %>%  
+   filter(order == "Primates", sleep_total > 10) %>%  
+   select(name, sleep_rem, sleep_cycle, sleep_total) %>%  
+   arrange(sleep_total)
```

```
# A tibble: 5 x 4
```

	name	sleep_rem	sleep_cycle	sleep_total
	<chr>	<dbl>	<dbl>	<dbl>
1	Macaque	1.20	0.750	10.1
2	Patas monkey	1.10	NA	10.9
3	Slow loris	NA	NA	11.0
4	Potto	NA	NA	11.0
5	Owl monkey	1.80	NA	17.0

smallest
largest



```
> msleep %>%  
+   filter(order == "Primates", sleep_total > 10) %>%  
+   select(name, sleep_rem, sleep_cycle, sleep_total) %>%  
+   arrange(desc(sleep_total))
```

```
# A tibble: 5 x 4
```

	name	sleep_rem	sleep_cycle	sleep_total
	<chr>	<dbl>	<dbl>	<dbl>
1	Owl monkey	1.80	NA	17.0
2	Slow loris	NA	NA	11.0
3	Potto	NA	NA	11.0
4	Patas monkey	1.10	NA	10.9
5	Macaque	1.20	0.750	10.1

largest

smallest

```
> msleep %>%  
+   filter(order == "Primates", sleep_total > 10) %>%  
+   select(name, sleep_rem, sleep_cycle, sleep_total) %>%  
+   arrange(name)  
# A tibble: 5 x 4
```

	name <chr>	sleep_rem <dbl>	sleep_cycle <dbl>	sleep_total <dbl>
1	Macaque	1.20	0.750	10.1
2	Owl monkey	1.80	NA	17.0
3	Patas monkey	1.10	NA	10.9
4	Potto	NA	NA	11.0
5	Slow loris	NA	NA	11.0

Sorted alphabetically
by name

Sort alphabetically by name, then total sleep:
`arrange(name, sleep_total)`



Manipulating Data




```
> msleep %>%  
+   filter(order == "Primates", sleep_total > 10) %>%  
+   select(name, sleep_rem, sleep_cycle, sleep_total) %>%  
+   arrange(name) %>%  
+   mutate(sleep_total_min = sleep_total * 60)
```

```
# A tibble: 5 x 5
```

	name	sleep_rem	sleep_cycle	sleep_total	sleep_total_min
	<chr>	<dbl>	<dbl>	<dbl>	<dbl>
1	Macaque	1.20	0.750	10.1	606
2	Owl monkey	1.80	NA	17.0	1020
3	Patas monkey	1.10	NA	10.9	654
4	Potto	NA	NA	11.0	660
5	Slow loris	NA	NA	11.0	660

A whole new column!



if not already installed, you'll have to run the following line of code

```
install.packages('httr')
```

load the library

```
library("httr")
```

download file

```
GET("https://raw.githubusercontent.com/suzanbaert/RTutorials/master/Rmd_originals/conservation_explanation.csv",  
write_disk(tf <- tempfile(fileext = ".csv"))  
conservation <- read_csv(tf)
```

take a look at this file

```
head(conservation)
```



```
> conservation
```

```
# A tibble: 11 x 1
```

```
  `conservation abbreviation`
```

```
  <chr>
```

```
1 EX = Extinct
```

```
2 EW = Extinct in the wild
```

```
3 CR = Critically Endangered
```

```
4 EN = Endangered
```

```
5 VU = Vulnerable
```

```
6 NT = Near Threatened
```

```
7 LC = Least Concern
```

```
8 DD = Data deficient
```

```
9 NE = Not evaluated
```

```
10 PE = Probably extinct (informal)
```

```
11 PEW = Probably extinct in the wild (informal)
```

Tidy data violation!

Space in column name should be an underscore.

Tidy data violation!

There are two pieces of information in a column: (1) abbreviation and (2) description.




```
> conservation %>%  
+   separate(`conservation abbreviation`,  
+           into = c("abbreviation", "description"), sep = " = ")  
# A tibble: 11 x 2
```

	abbreviation <chr>	description <chr>
1	EX	Extinct
2	EW	Extinct in the wild
3	CR	Critically Endangered
4	EN	Endangered
5	VU	Vulnerable
6	NT	Near Threatened
7	LC	Least Concern
8	DD	Data deficient
9	NE	Not evaluated
10	PE	Probably extinct (informal)
11	PEW	Probably extinct in the wild (informal)

```
> conservation %>%
+   separate(`conservation abbreviation`,
+           into = c("abbreviation", "description"), sep = " = ") %>%
+   unite(united_col, abbreviation, description, sep=" = ")
# A tibble: 11 x 1
  united_col
  <chr>
1 EX = Extinct
2 EW = Extinct in the wild
3 CR = Critically Endangered
4 EN = Endangered
5 VU = Vulnerable
6 NT = Near Threatened
7 LC = Least Concern
8 DD = Data deficient
9 NE = Not evaluated
10 PE = Probably extinct (informal)
11 PEW = Probably extinct in the wild (informal)
```



```
> conservation %>%  
+   clean_names()  
# A tibble: 11 x 1  
  conservation_abbreviation  
  <chr>  
1 EX = Extinct  
2 EW = Extinct in the wild  
3 CR = Critically Endangered  
4 EN = Endangered  
5 VU = Vulnerable  
6 NT = Near Threatened  
7 LC = Least Concern  
8 DD = Data deficient  
9 NE = Not evaluated  
10 PE = Probably extinct (informal)  
11 PEW = Probably extinct in the wild (informal)
```



Adds underscore to column name


```
> msleep %>%
+   mutate(conservation = toupper(conservation)) %>%
+   left_join(conservation, by = c("conservation" = "abbreviation"))
```

```
# A tibble: 83 x 12
```

	name	genus	vore	order	conservation	sleep_total	sleep_rem	sleep_cycle	awake	brainwt	bodywt	description
	<chr>	<chr>	<chr>	<chr>	<chr>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<chr>
1	Cheetah	Acino...	carni	Carn...	LC	12.1	NA	NA	11.9	NA	5.00e ⁺¹	Least Conc...
2	Owl mon...	Aotus	omni	Prim...	<NA>	17.0	1.80	NA	7.00	1.55e ⁻²	4.80e ⁻¹	<NA>
3	Mountai...	Aplod...	herbi	Rode...	NT	14.4	2.40	NA	9.60	NA	1.35e ⁺⁰	Near Threa...
4	Greater...	Blari...	omni	Sori...	LC	14.9	2.30	0.133	9.10	2.90e ⁻⁴	1.90e ⁻²	Least Conc...
5	Cow	Bos	herbi	Arti...	DOMESTICATED	4.00	0.700	0.667	20.0	4.23e ⁻¹	6.00e ⁺²	<NA>
6	Three-t...	Brady...	herbi	Pilo...	<NA>	14.4	2.20	0.767	9.60	NA	3.85e ⁺⁰	<NA>
7	Norther...	Callo...	carni	Carn...	VU	8.70	1.40	0.383	15.3	NA	2.05e ⁺¹	Vulnerable
8	Vesper ...	Calom...	<NA>	Rode...	<NA>	7.00	NA	NA	17.0	NA	4.50e ⁻²	<NA>
9	Dog	Canis	carni	Carn...	DOMESTICATED	10.1	2.90	0.333	13.9	7.00e ⁻²	1.40e ⁺¹	<NA>
10	Roe deer	Capre...	herbi	Arti...	LC	3.00	NA	NA	21.0	9.82e ⁻²	1.48e ⁺¹	Least Conc...

```
# ... with 73 more rows
```



Summarizing Data




```
> msleep
# A tibble: 83 x 11
  name      genus  vore order  conservation sleep_total sleep_rem sleep_cycle awake brainwt bodywt
  <chr>    <chr>  <chr> <chr>    <chr>          <dbl>    <dbl>    <dbl> <dbl>  <dbl>  <dbl>
1 Cheetah  Acinonyx  carni Carnivo... lc          12.1      NA      NA    11.9  NA    5.00e+1
2 Owl monkey Aotus    omni  Primates <NA>       17.0     1.80    NA     7.00  1.55e-2 4.80e-1
3 Mountain beaver Aplodon... herbi Rodentia nt        14.4     2.40    NA     9.60  NA    1.35e+0
4 Greater short-... Blarina  omni  Soricom... lc        14.9     2.30    0.133  9.10  2.90e-4 1.90e-2
5 Cow      Bos      herbi Artioda... domesticated 4.00     0.700    0.667  20.0  4.23e-1 6.00e+2
6 Three-toed slo... Bradypus herbi Pilosa  <NA>       14.4     2.20    0.767  9.60  NA    3.85e+0
7 Northern fur s... Callorh... carni Carnivo... vu        8.70     1.40    0.383  15.3  NA    2.05e+1
8 Vesper mouse Calomys  <NA>  Rodentia <NA>       7.00     NA      NA    17.0  NA    4.50e-2
9 Dog      Canis    carni Carnivo... domesticated 10.1     2.90    0.333  13.9  7.00e-2 1.40e+1
10 Roe deer Capreol... herbi Artioda... lc         3.00     NA      NA    21.0  9.82e-2 1.48e+1
# ... with 73 more rows
```

```
> msleep %>%
+   group_by(order)
```

```
# A tibble: 83 x 11
```

```
# Groups:   order [19]
```

```
  name      genus  vore order  conservation sleep_total sleep_rem sleep_cycle awake brainwt bodywt
  <chr>    <chr>  <chr> <chr>    <chr>          <dbl>    <dbl>    <dbl> <dbl>  <dbl>  <dbl>
1 Cheetah  Acinonyx  carni Carnivo... lc          12.1      NA      NA    11.9  NA    5.00e+1
2 Owl monkey Aotus    omni  Primates <NA>       17.0     1.80    NA     7.00  1.55e-2 4.80e-1
3 Mountain beaver Aplodon... herbi Rodentia nt        14.4     2.40    NA     9.60  NA    1.35e+0
4 Greater short-... Blarina  omni  Soricom... lc        14.9     2.30    0.133  9.10  2.90e-4 1.90e-2
5 Cow      Bos      herbi Artioda... domesticated 4.00     0.700    0.667  20.0  4.23e-1 6.00e+2
6 Three-toed slo... Bradypus herbi Pilosa  <NA>       14.4     2.20    0.767  9.60  NA    3.85e+0
7 Northern fur s... Callorh... carni Carnivo... vu        8.70     1.40    0.383  15.3  NA    2.05e+1
8 Vesper mouse Calomys  <NA>  Rodentia <NA>       7.00     NA      NA    17.0  NA    4.50e-2
9 Dog      Canis    carni Carnivo... domesticated 10.1     2.90    0.333  13.9  7.00e-2 1.40e+1
10 Roe deer Capreol... herbi Artioda... lc         3.00     NA      NA    21.0  9.82e-2 1.48e+1
# ... with 73 more rows
```



```
> msleep %>%  
+   select(order) %>%  
+   summarize(N=n())  
# A tibble: 1 x 1  
      N  
  <int>  
1    83
```

Same as `nrow(msleep)`



```
> msleep %>%  
+   group_by(order) %>%  
+   select(order) %>%  
+   summarize(N=n())  
# A tibble: 19 x 2  
  order      N  
  <chr>    <int>  
1 Afrosoricida      1  
2 Artiodactyla      6  
3 Carnivora       12  
4 Cetacea           3  
5 Chiroptera        2  
6 Cingulata         2  
7 Didelphimorphia   2  
8 Diprotodontia     2  
9 Erinaceomorpha    2  
10 Hyracoidea        3  
11 Lagomorpha        1  
12 Monotremata       1  
13 Perissodactyla    3  
14 Pilosa            1  
15 Primates         12  
16 Proboscidea       2  
17 Rodentia         22  
18 Scandentia        1  
19 Soricomorpha      5
```



```
> msleep %>%  
+   group_by(order) %>%  
+   select(order, sleep_total) %>%  
+   summarize(N=n(), mean_sleep=mean(sleep_total))
```

```
# A tibble: 19 x 3
```

	order <chr>	N <int>	mean_sleep <dbl>
1	Afrosoricida	1	15.6
2	Artiodactyla	6	4.52
3	Carnivora	12	10.1
4	Cetacea	3	4.50
5	Chiroptera	2	19.8
6	Cingulata	2	17.8
7	Didelphimorphia	2	18.7
8	Diprotodontia	2	12.4
9	Erinaceomorpha	2	10.2
10	Hyracoidea	3	5.67
11	Lagomorpha	1	8.40
12	Monotremata	1	8.60
13	Perissodactyla	3	3.47
14	Pilosa	1	14.4
15	Primates	12	10.5
16	Proboscidea	2	3.60
17	Rodentia	22	12.5
18	Scandentia	1	8.90
19	Soricomorpha	5	11.1

```
> msleep %>%
```

```
+ tabyl(order)
```

	order	n	percent
1	Afrosoricida	1	0.0120
2	Artiodactyla	6	0.0723
3	Carnivora	12	0.1446
4	Cetacea	3	0.0361
5	Chiroptera	2	0.0241
6	Cingulata	2	0.0241
7	Didelphimorphia	2	0.0241
8	Diprotodontia	2	0.0241
9	Erinaceomorpha	2	0.0241
10	Hyracoidea	3	0.0361
11	Lagomorpha	1	0.0120
12	Monotremata	1	0.0120
13	Perissodactyla	3	0.0361
14	Pilosa	1	0.0120
15	Primates	12	0.1446
16	Proboscidea	2	0.0241
17	Rodentia	22	0.2651
18	Scandentia	1	0.0120
19	Soricomorpha	5	0.0602



```
> summary(msleep$awake)
```

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
4.1	10.2	13.9	13.6	16.1	22.1



```
> skim(msleep)
```

Skim summary statistics







n obs: 83

n variables: 11

Variable type: character

variable	missing	complete	n	min	max	empty	n_unique
conservation	29	54	83	2	12	0	6
genus	0	83	83	3	13	0	77
name	0	83	83	3	30	0	83
order	0	83	83	6	15	0	19
vore	7	76	83	4	7	0	4

Variable type: numeric

variable	missing	complete	n	mean	sd	p0	p25	p50	p75	p100	hist
awake	0	83	83	13.57	4.45	4.1	10.25	13.9	16.15	22.1	
bodywt	0	83	83	166.14	786.84	0.005	0.17	1.67	41.75	6654	
brainwt	27	56	83	0.28	0.98	0.00014	0.0029	0.012	0.13	5.71	
sleep_cycle	51	32	83	0.44	0.36	0.12	0.18	0.33	0.58	1.5	
sleep_rem	22	61	83	1.88	1.3	0.1	0.9	1.5	2.4	6.6	
sleep_total	0	83	83	10.43	4.45	1.9	7.85	10.1	13.75	19.9	



Filtering, Re-ordering, Manipulating, and Summarizing

