Poli-301:

The Political Science Discipline

TODAY'S AGENDA

1 Gathering data to make it tidy

2 Factors for categories

3 Setting up projects, getting data into R

But first...

Subset Variables (Columns)

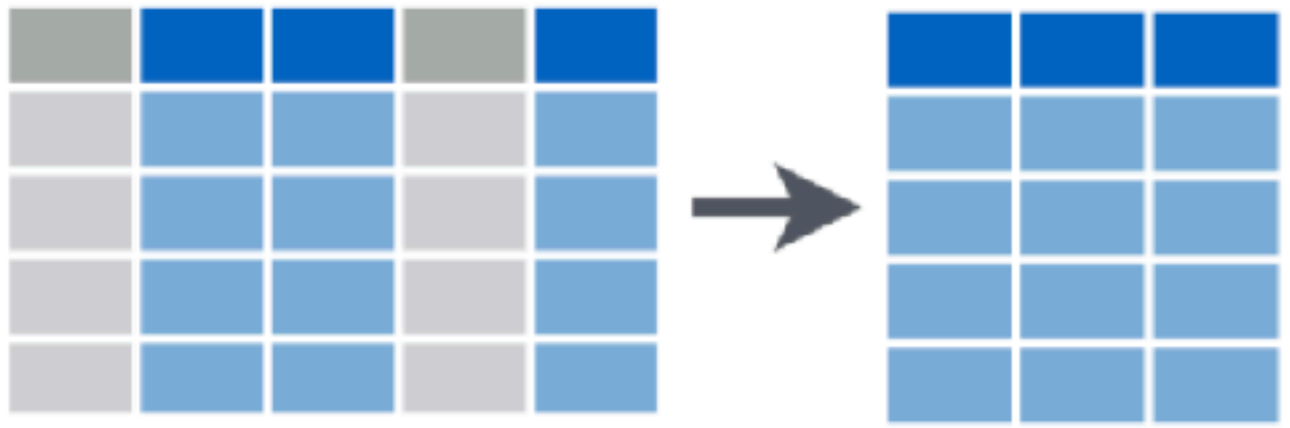
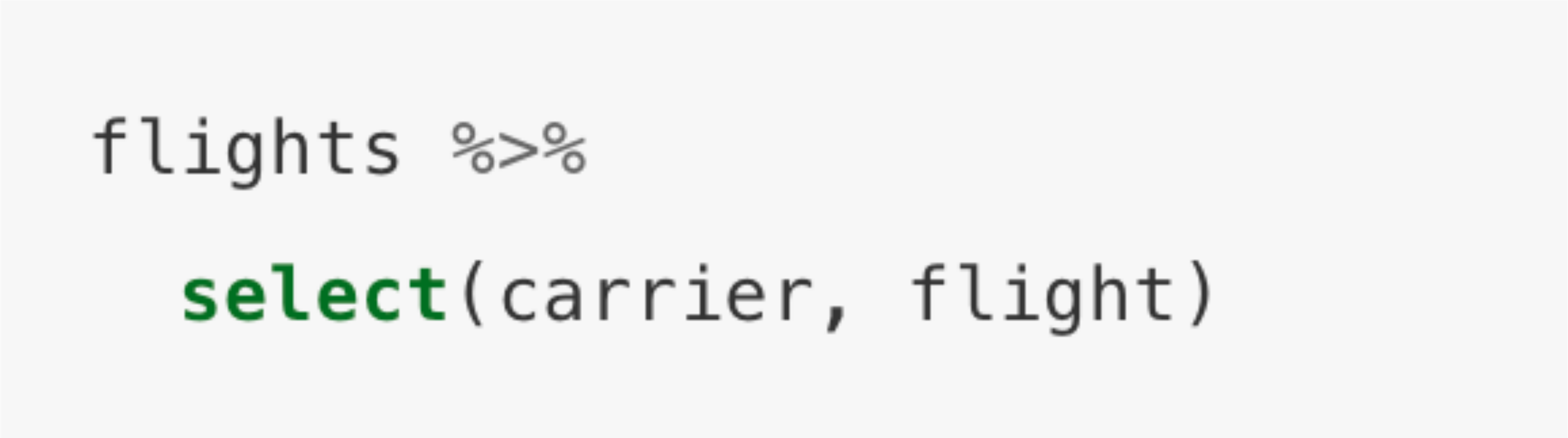
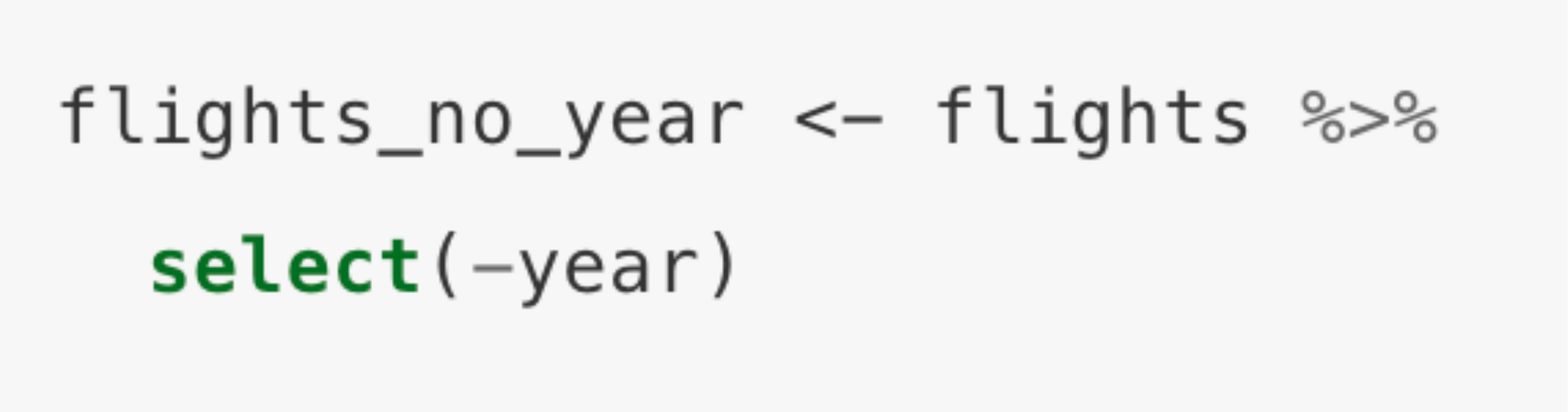


FIGURE 3.9: Diagram of select() columns.



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



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

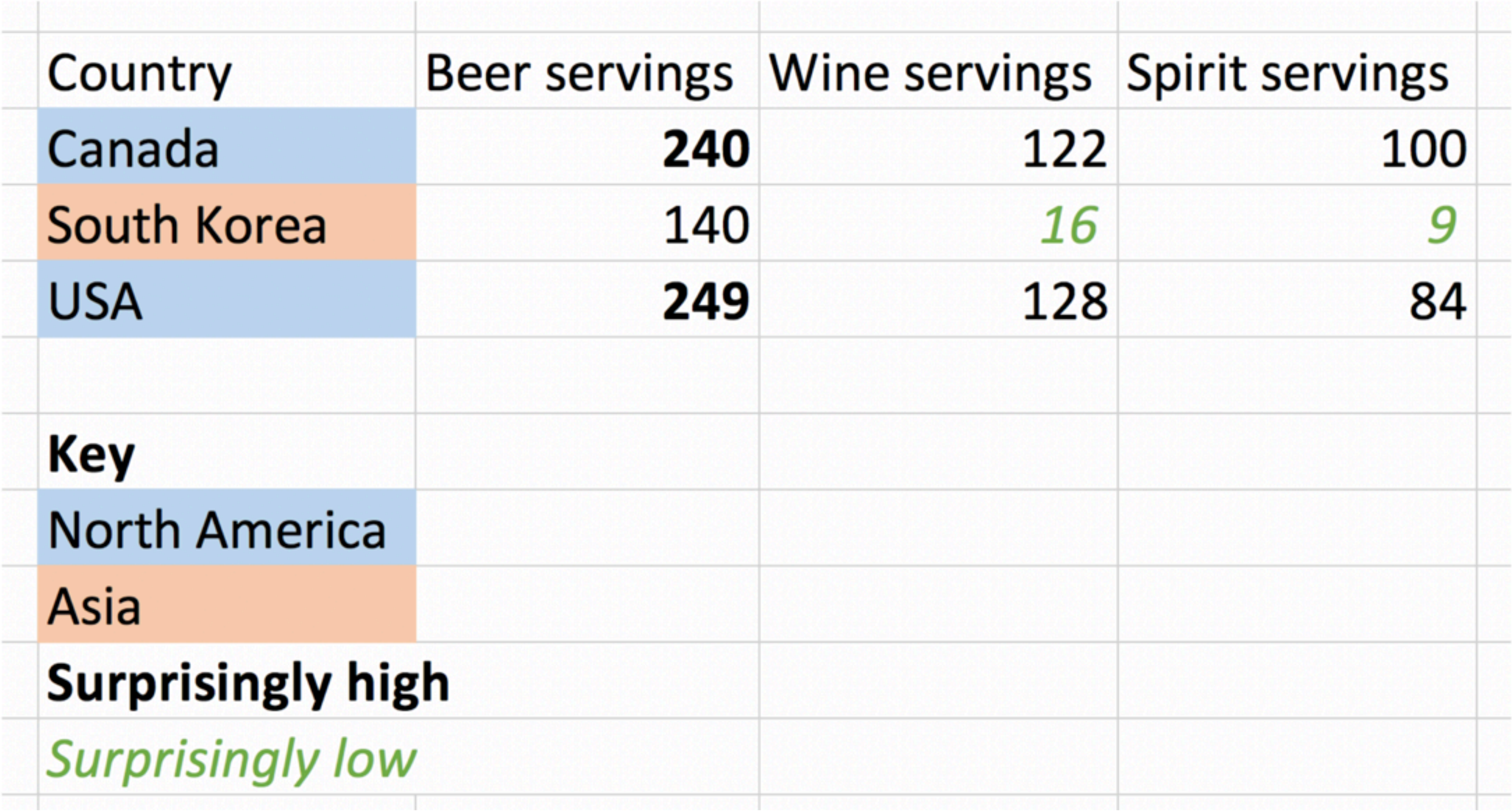
What is tidy data?

Each variable is a column

Each observation is a row

Data is “long”, not “wide”

Extremely untidy



Country	Beer servings	Wine servings	Spirit servings
Canada	**240**	122	100
South Korea	140	*16*	*9*
USA	**249**	128	84

Key

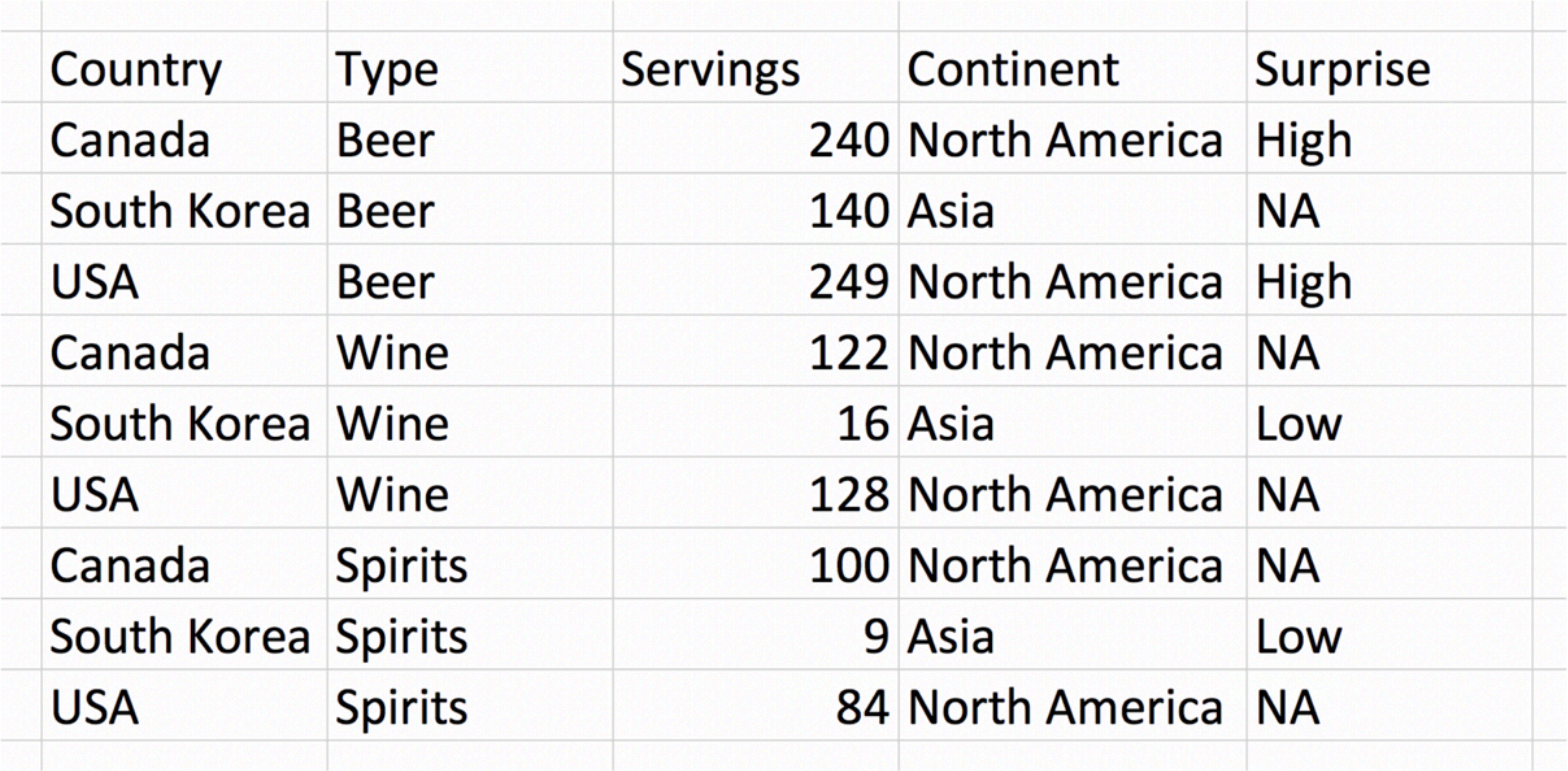
North America

Asia

Surprisingly high

Surprisingly low

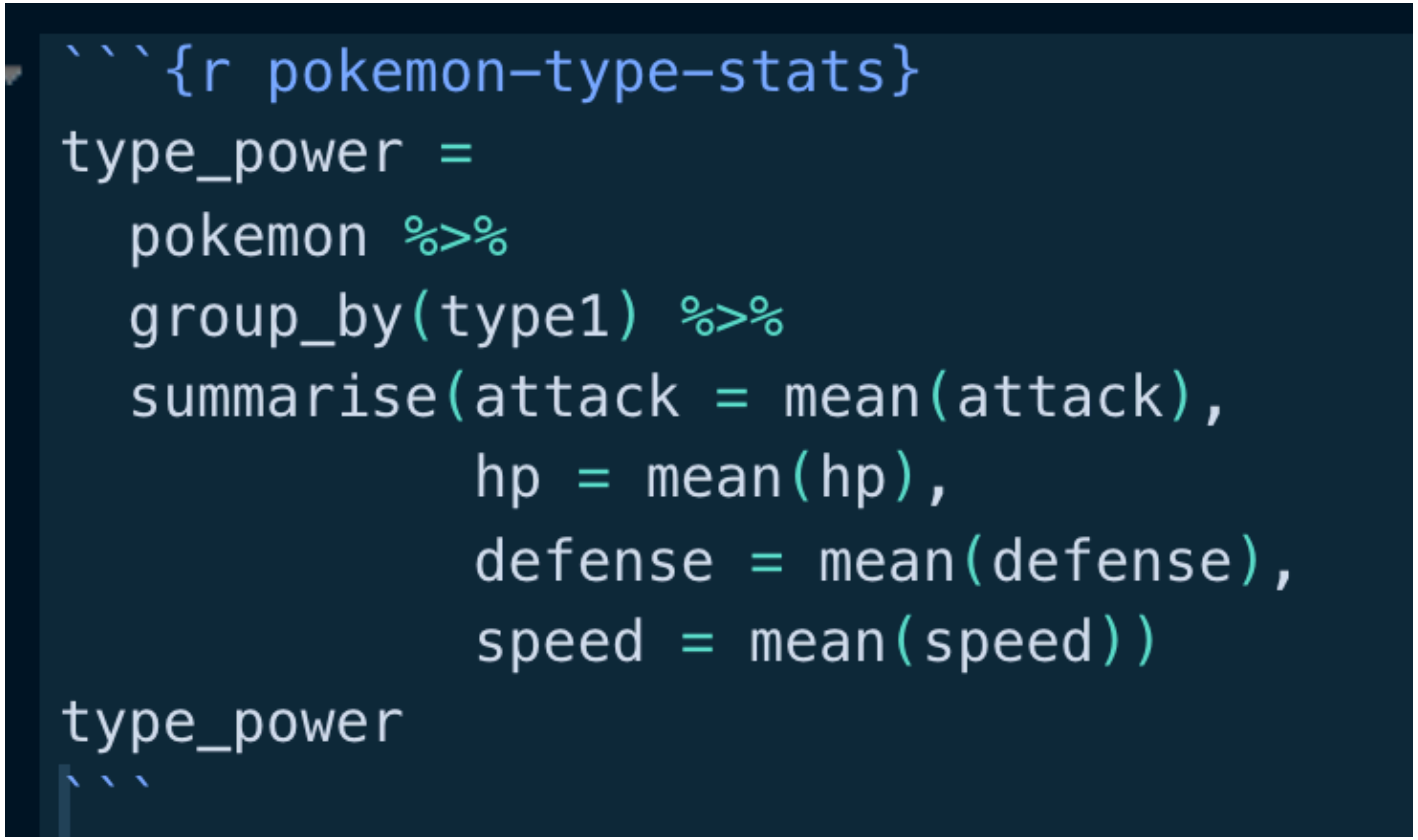
Tidy



Country	Type	Servings	Continent	Surprise
Canada	Beer	240	North America	High
South Korea	Beer	140	Asia	NA
USA	Beer	249	North America	High
Canada	Wine	122	North America	NA
South Korea	Wine	16	Asia	Low
USA	Wine	128	North America	NA
Canada	Spirits	100	North America	NA
South Korea	Spirits	9	Asia	Low
USA	Spirits	84	North America	NA

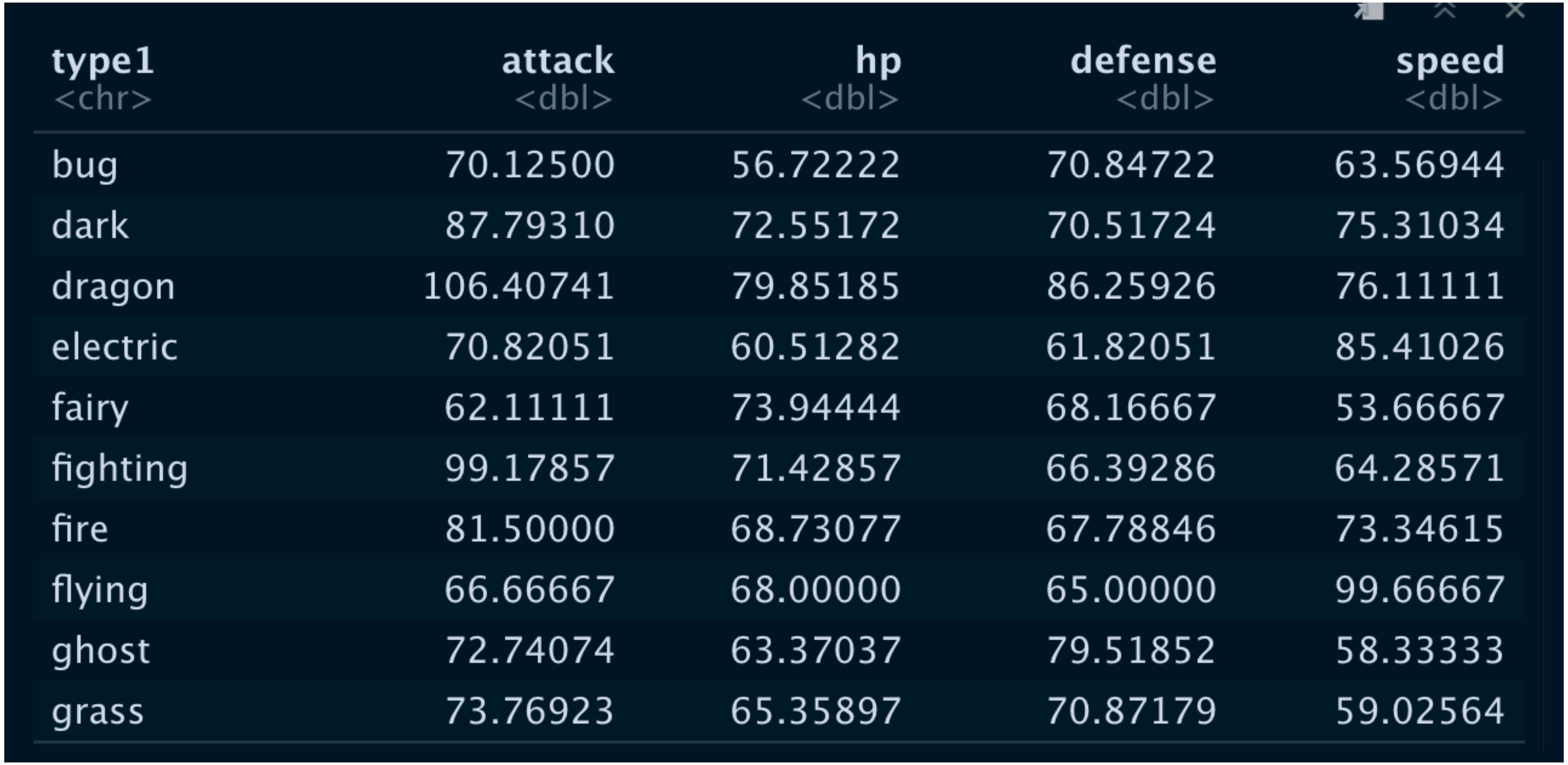
Untidy data

Average stats for each Pokemon type



~~~
```{r pokemon-type-stats}
type_power =
  pokemon %>%
  group_by(type1) %>%
  summarise(attack = mean(attack),
            hp = mean(hp),
            defense = mean(defense),
            speed = mean(speed))
type_power
```
~~~

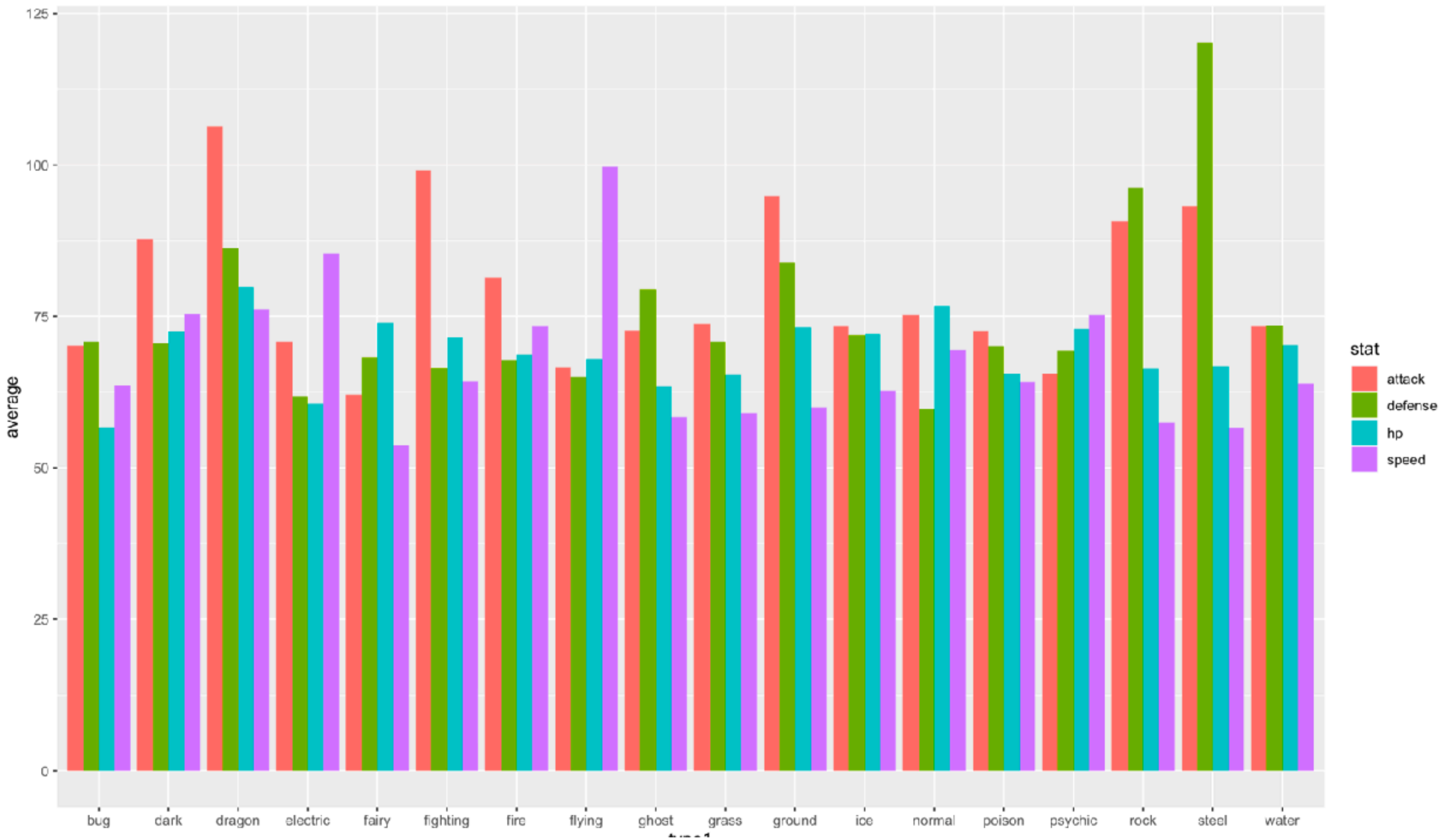
Untidy data



type1	attack	hp	defense	speed
<chr>	<dbl>	<dbl>	<dbl>	<dbl>
bug	70.12500	56.72222	70.84722	63.56944
dark	87.79310	72.55172	70.51724	75.31034
dragon	106.40741	79.85185	86.25926	76.11111
electric	70.82051	60.51282	61.82051	85.41026
fairy	62.11111	73.94444	68.16667	53.66667
fighting	99.17857	71.42857	66.39286	64.28571
fire	81.50000	68.73077	67.78846	73.34615
flying	66.66667	68.00000	65.00000	99.66667
ghost	72.74074	63.37037	79.51852	58.33333
grass	73.76923	65.35897	70.87179	59.02564

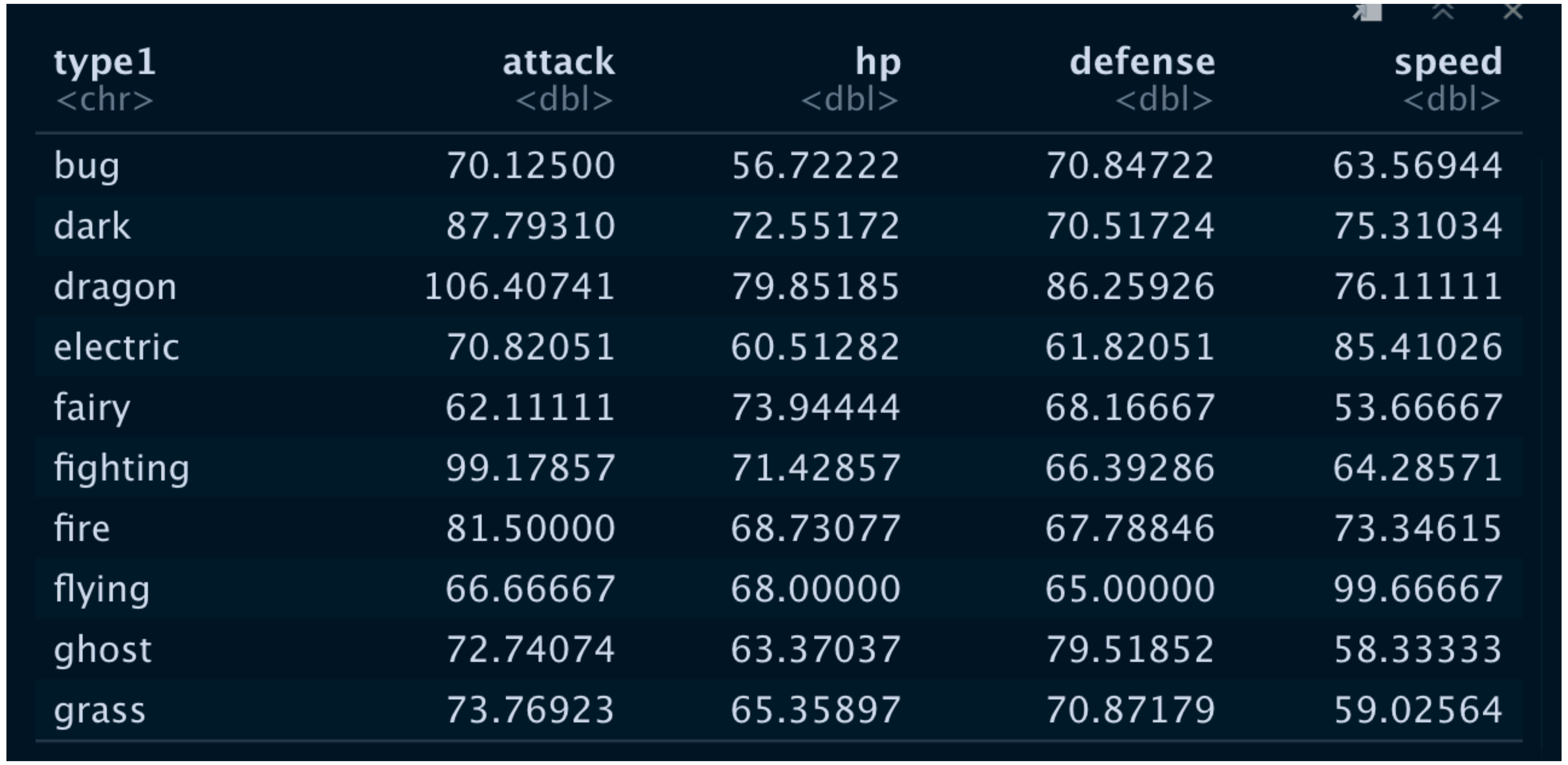
Untidy data

Say I want to make the plot below:

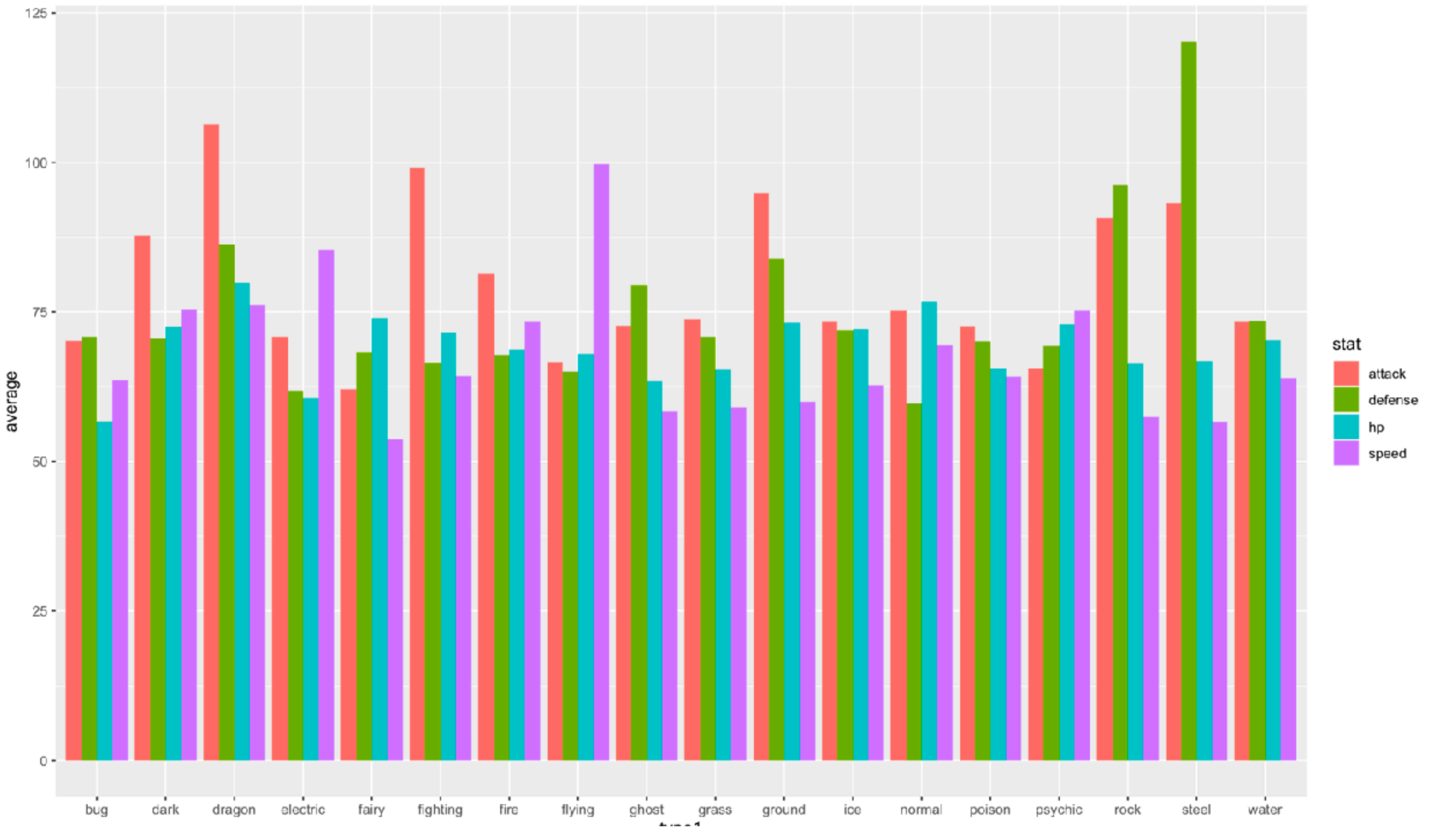


How?

```
ggplot(type_power, aes(x = ???, y = ???, fill = ???)) +
geom_col(position = “dodge”)
```



type1 <chr>	attack <dbl>	hp <dbl>	defense <dbl>	speed <dbl>
bug	70.12500	56.72222	70.84722	63.56944
dark	87.79310	72.55172	70.51724	75.31034
dragon	106.40741	79.85185	86.25926	76.11111
electric	70.82051	60.51282	61.82051	85.41026
fairy	62.11111	73.94444	68.16667	53.66667
fighting	99.17857	71.42857	66.39286	64.28571
fire	81.50000	68.73077	67.78846	73.34615
flying	66.66667	68.00000	65.00000	99.66667
ghost	72.74074	63.37037	79.51852	58.33333
grass	73.76923	65.35897	70.87179	59.02564

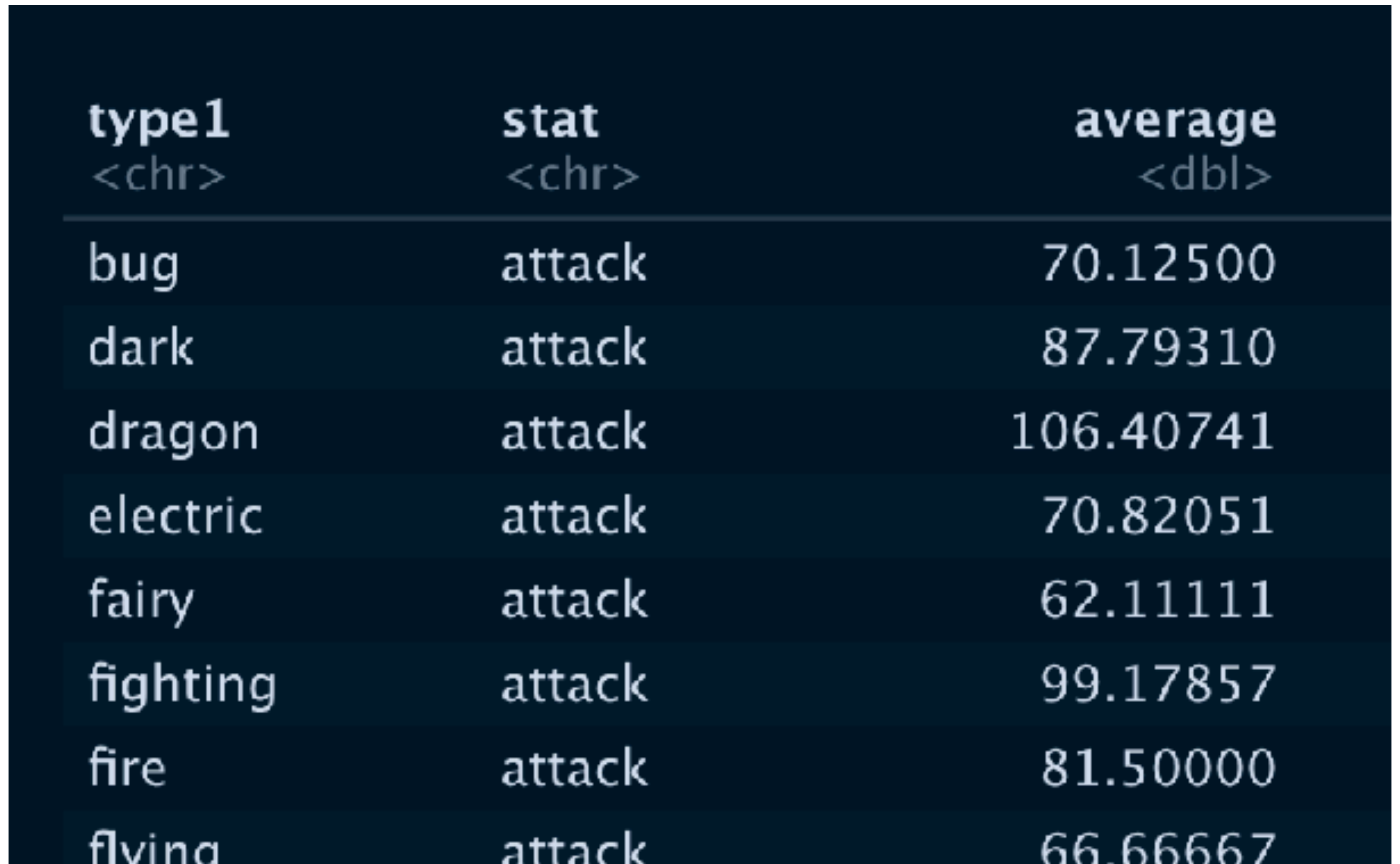


Can’t be done

Need one variable, “average” for y-axis

But I have averages for 4 Pokemon stats

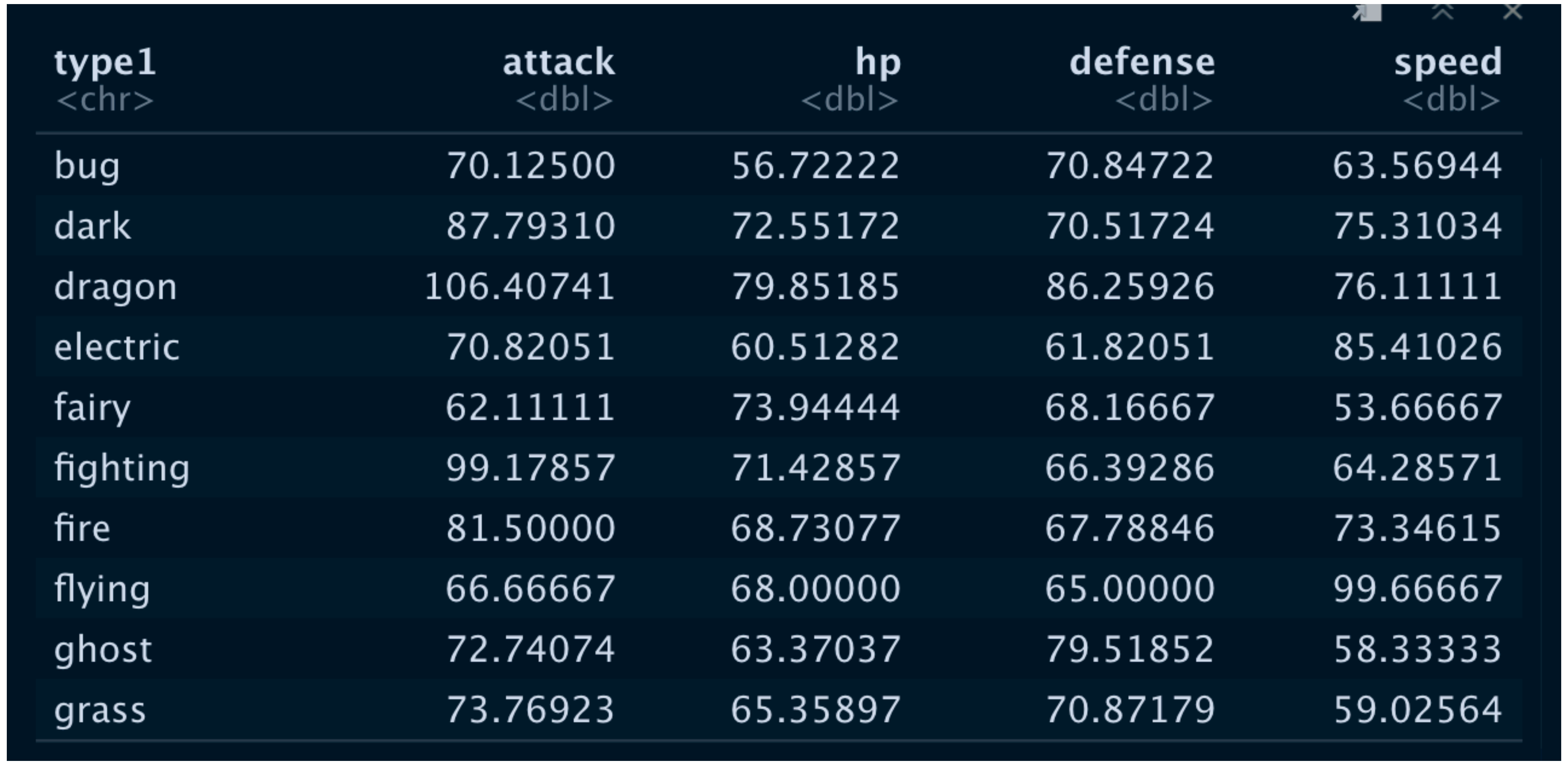
I need *one* variable, “stat”,
with values “attack, defense, speed, hp”



type1	stat	average
<chr>	<chr>	<dbl>
bug	attack	70.12500
dark	attack	87.79310
dragon	attack	106.40741
electric	attack	70.82051
fairy	attack	62.11111
fighting	attack	99.17857
fire	attack	81.50000
flying	attack	66.66667

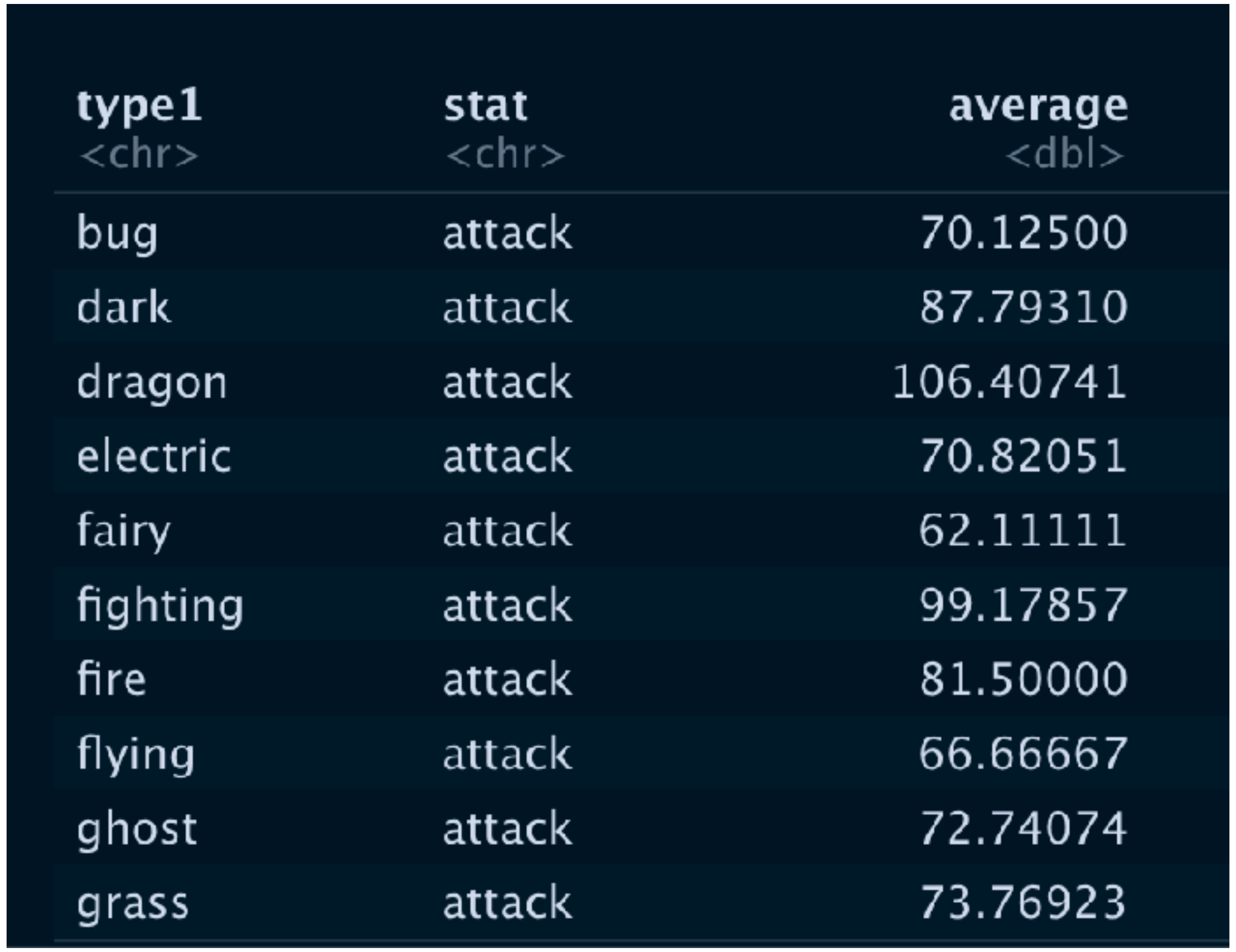
We need long data

Wide



type1	attack	hp	defense	speed
<chr>	<dbl>	<dbl>	<dbl>	<dbl>
bug	70.12500	56.72222	70.84722	63.56944
dark	87.79310	72.55172	70.51724	75.31034
dragon	106.40741	79.85185	86.25926	76.11111
electric	70.82051	60.51282	61.82051	85.41026
fairy	62.11111	73.94444	68.16667	53.66667
fighting	99.17857	71.42857	66.39286	64.28571
fire	81.50000	68.73077	67.78846	73.34615
flying	66.66667	68.00000	65.00000	99.66667
ghost	72.74074	63.37037	79.51852	58.33333
grass	73.76923	65.35897	70.87179	59.02564

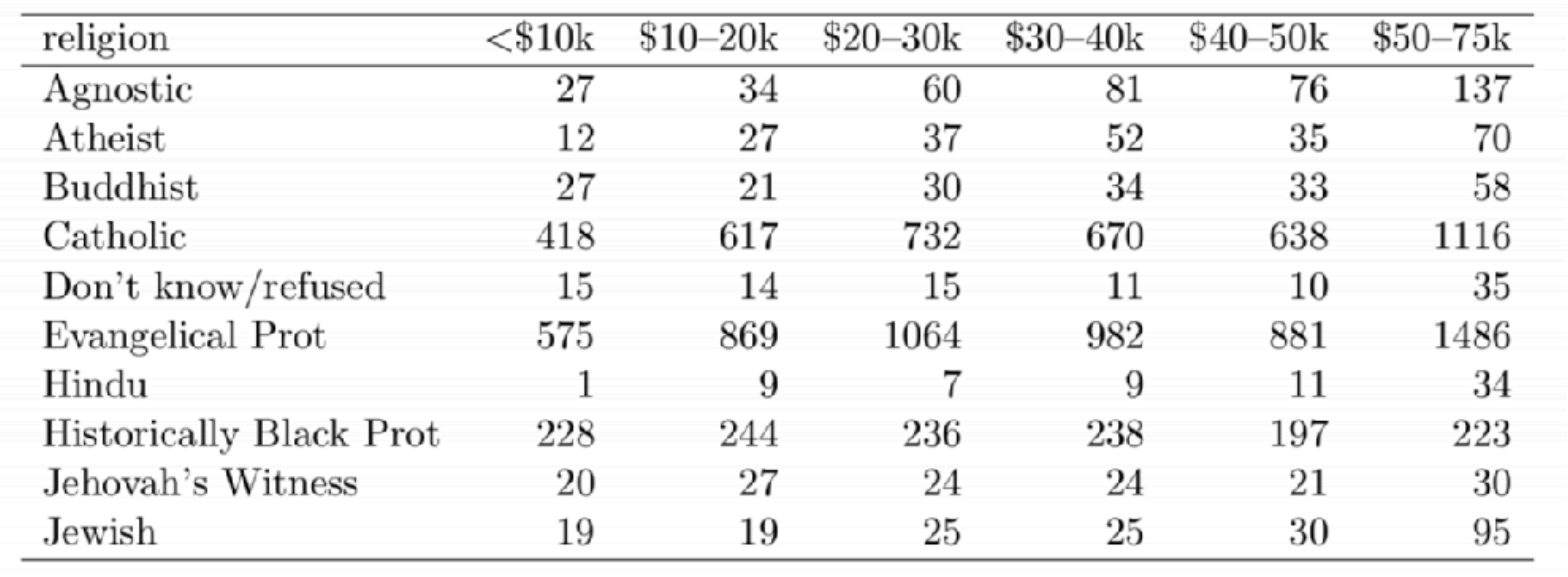
Long



type1	stat	average
<chr>	<chr>	<dbl>
bug	attack	70.12500
dark	attack	87.79310
dragon	attack	106.40741
electric	attack	70.82051
fairy	attack	62.11111
fighting	attack	99.17857
fire	attack	81.50000
flying	attack	66.66667
ghost	attack	72.74074
grass	attack	73.76923

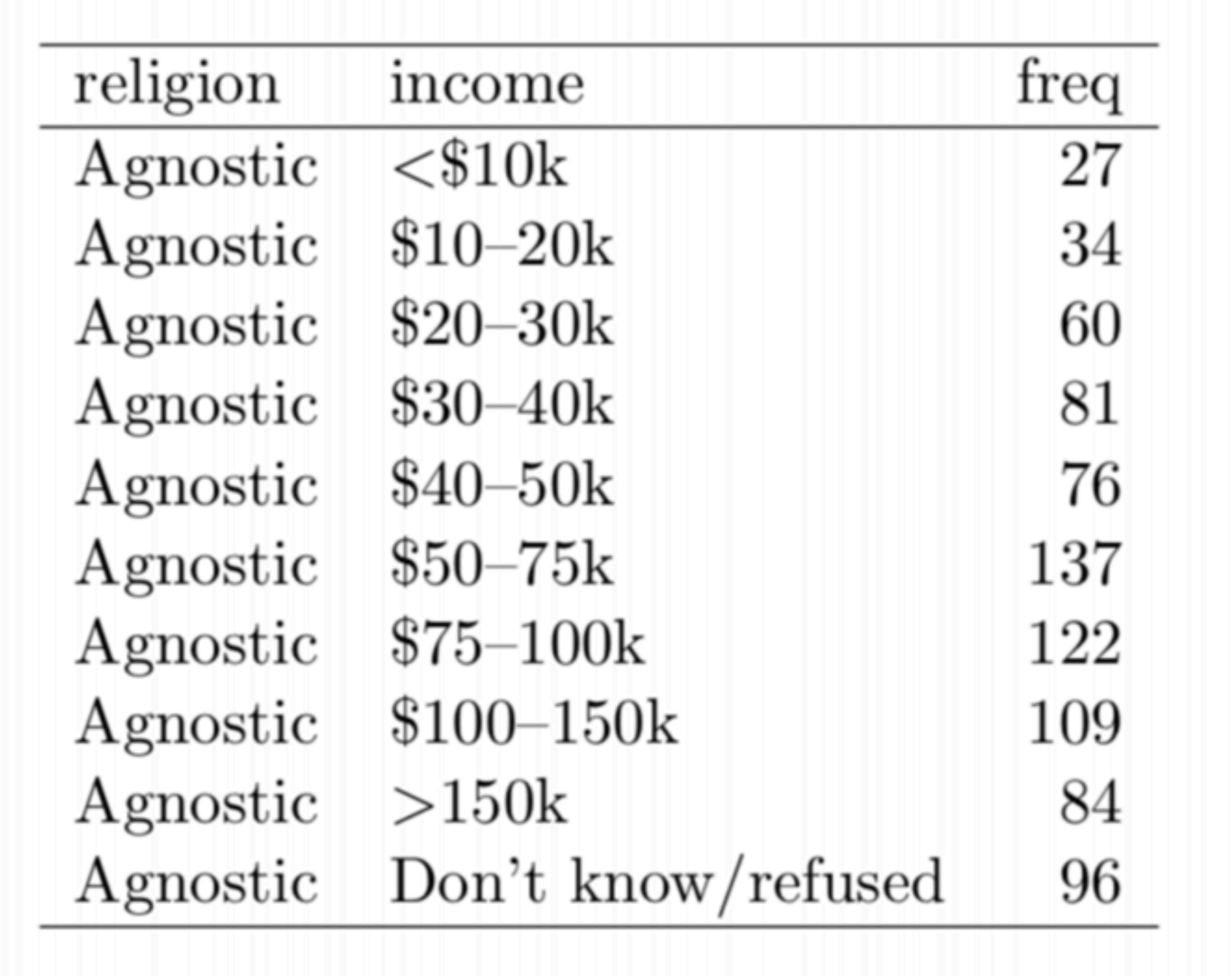
Wide vs. Long

Wide



religion	<$10k	$10–20k	$20–30k	$30–40k	$40–50k	$50–75k
Agnostic	27	34	60	81	76	137
Atheist	12	27	37	52	35	70
Buddhist	27	21	30	34	33	58
Catholic	418	617	732	670	638	1116
Don't know/refused	15	14	15	11	10	35
Evangelical Prot	575	869	1064	982	881	1486
Hindu	1	9	7	9	11	34
Historically Black Prot	228	244	236	238	197	223
Jehovah's Witness	20	27	24	24	21	30
Jewish	19	19	25	25	30	95

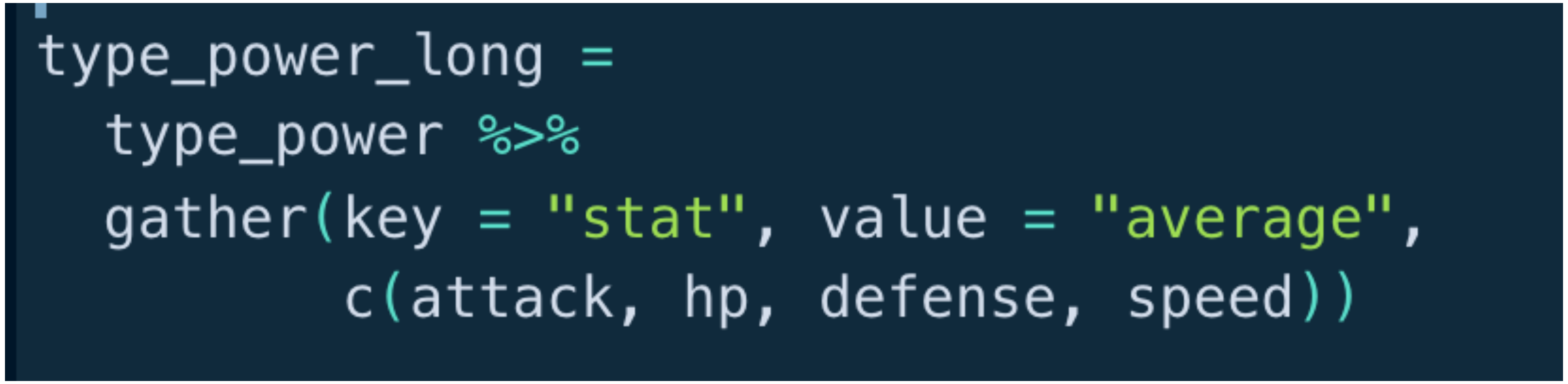
Long



religion	income	freq
Agnostic	<$10k	27
Agnostic	$10–20k	34
Agnostic	$20–30k	60
Agnostic	$30–40k	81
Agnostic	$40–50k	76
Agnostic	$50–75k	137
Agnostic	$75–100k	122
Agnostic	$100–150k	109
Agnostic	>150k	84
Agnostic	Don't know/refused	96

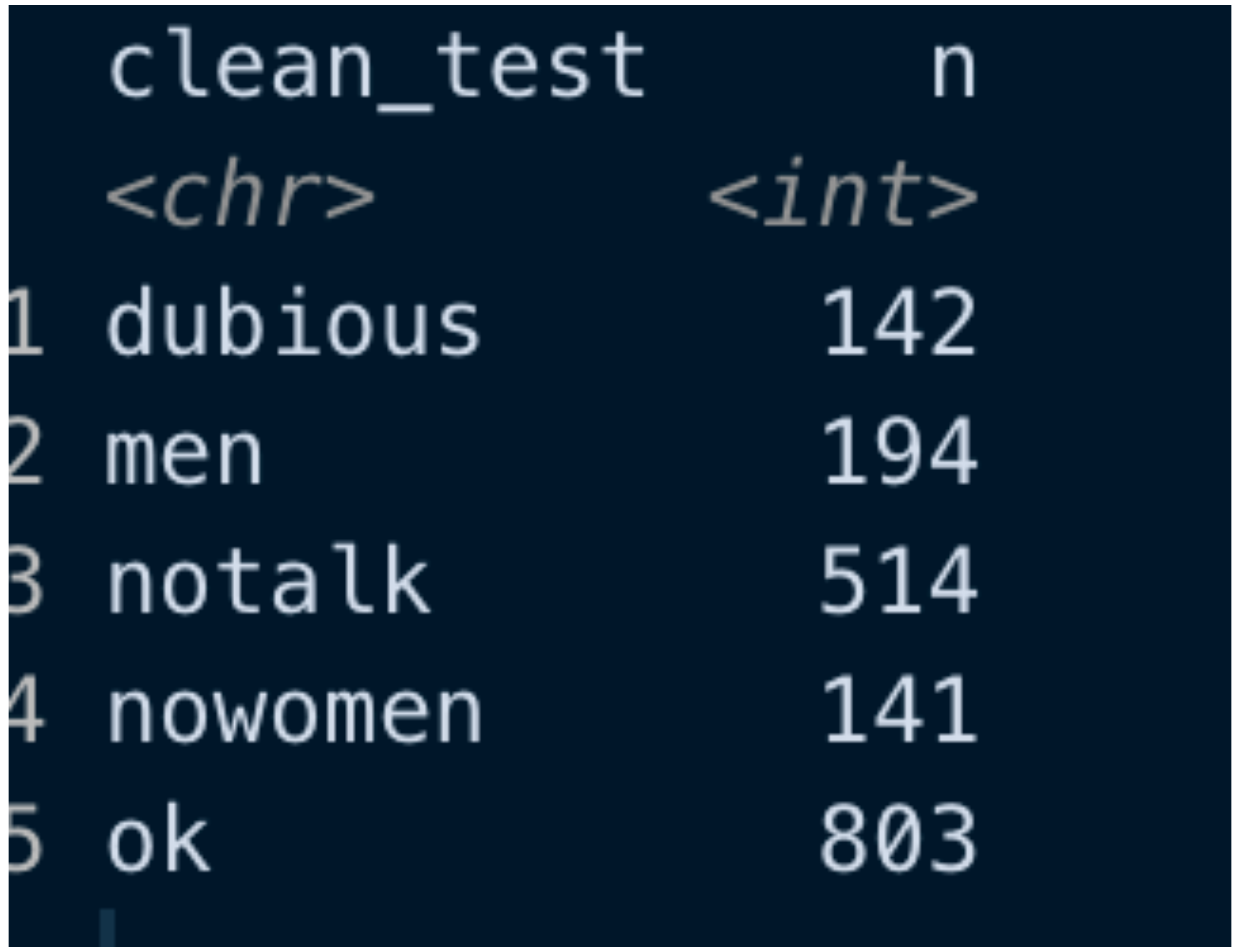
“Gathering” data from wide to long

```
Data %>%
gather(key = “new key”, value = “new value”,
c(variables, to, be, gathered))
```

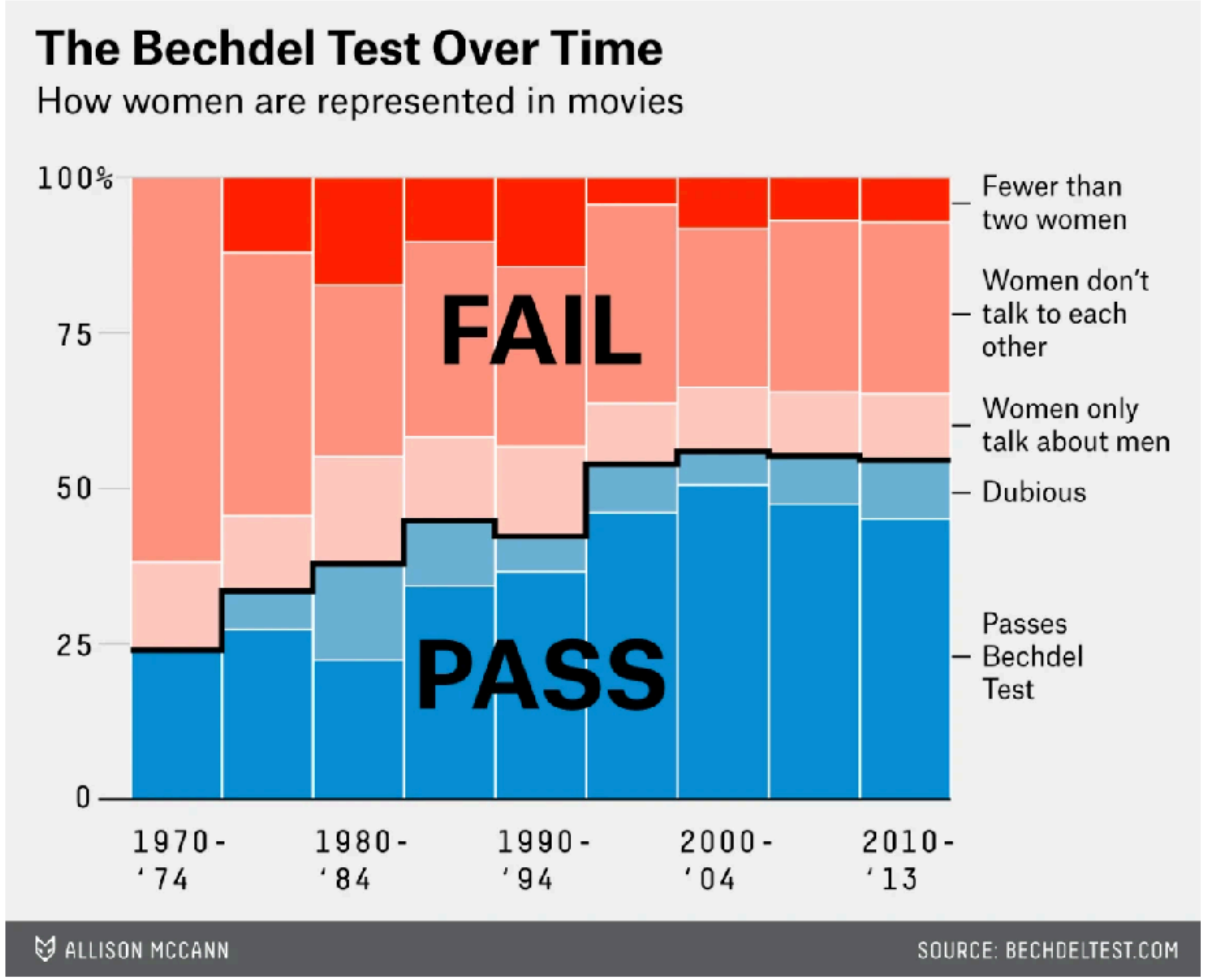


```
type_power_long =
  type_power %>%
  gather(key = "stat", value = "average",
         c(attack, hp, defense, speed))
```

Ordered categories



```
  clean_test        n
  <chr>         <int>
1 dubious         142
2 men             194
3 notalk          514
4 nowomen         141
5 ok              803
```



The Bechdel Test Over Time

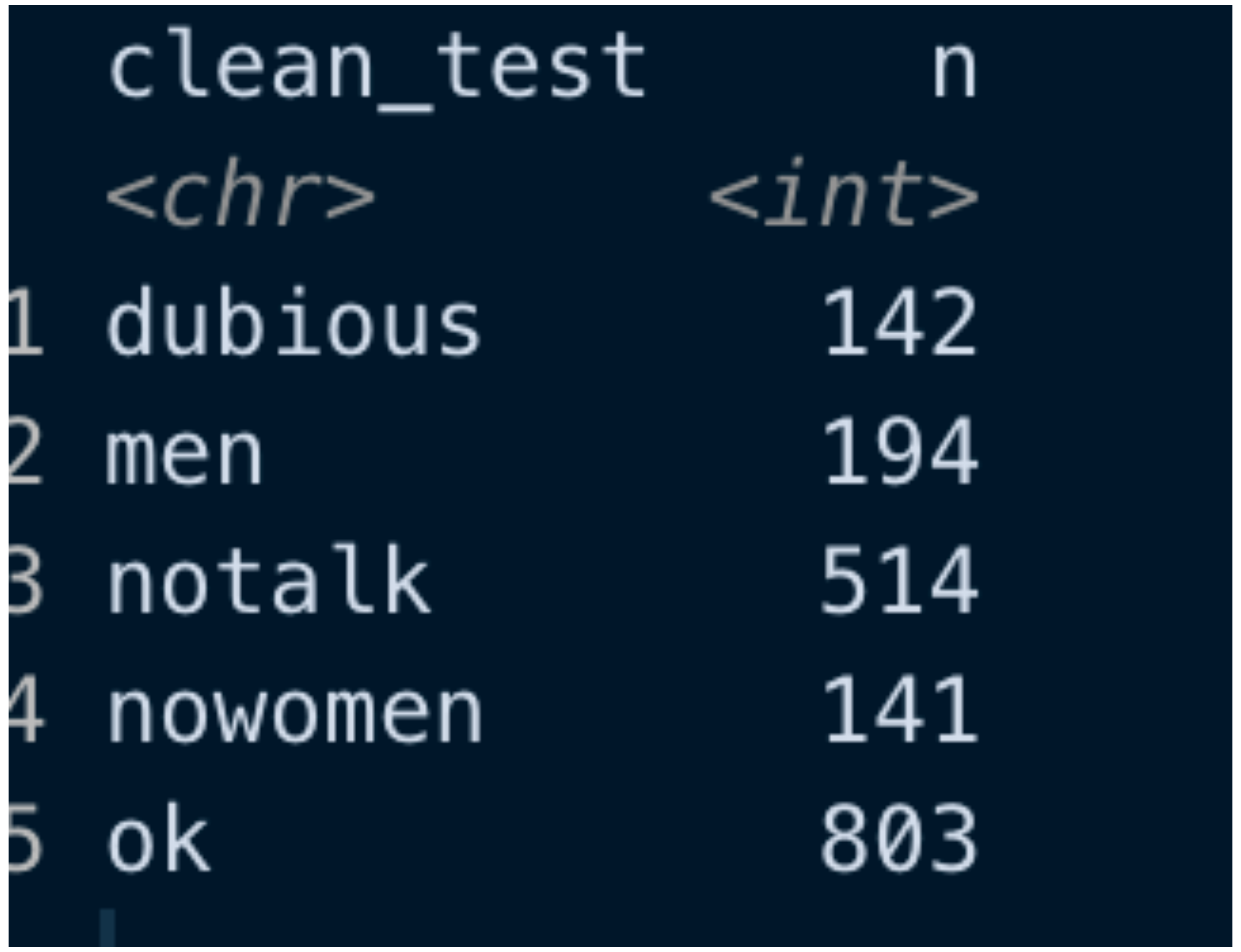
How women are represented in movies

ALLISON MCCANN

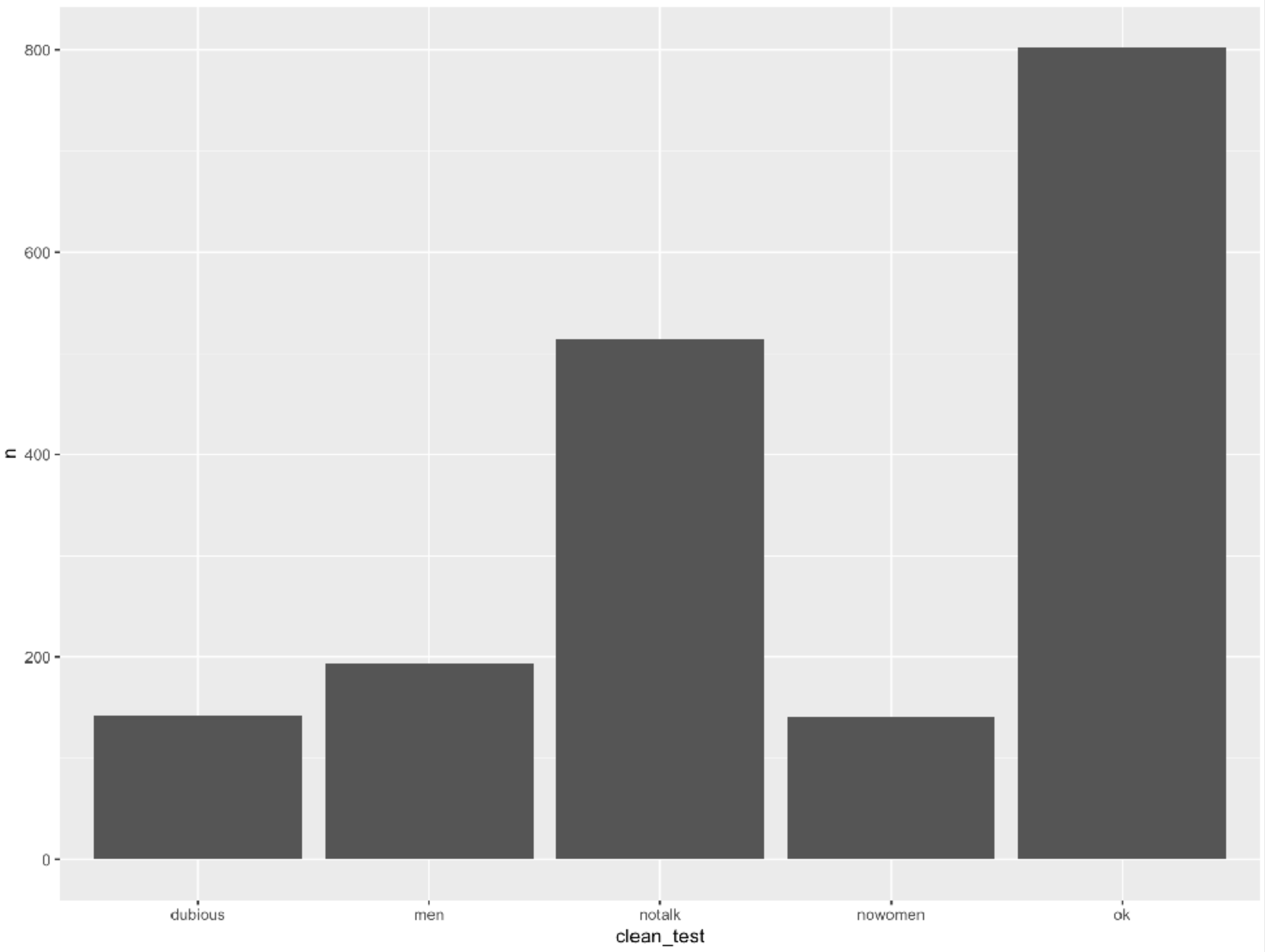
SOURCE: BECHDELTEST.COM

Ordered categories

What if we wanted the plot to reflect order of test?

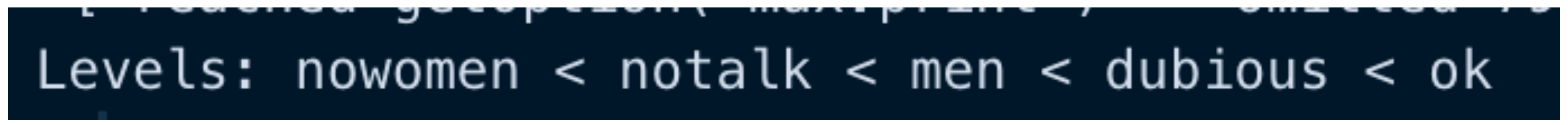


```
  clean_test      n
  <chr>       <int>
1 dubious       142
2 men           194
3 notalk        514
4 nowomen       141
5 ok            803
```

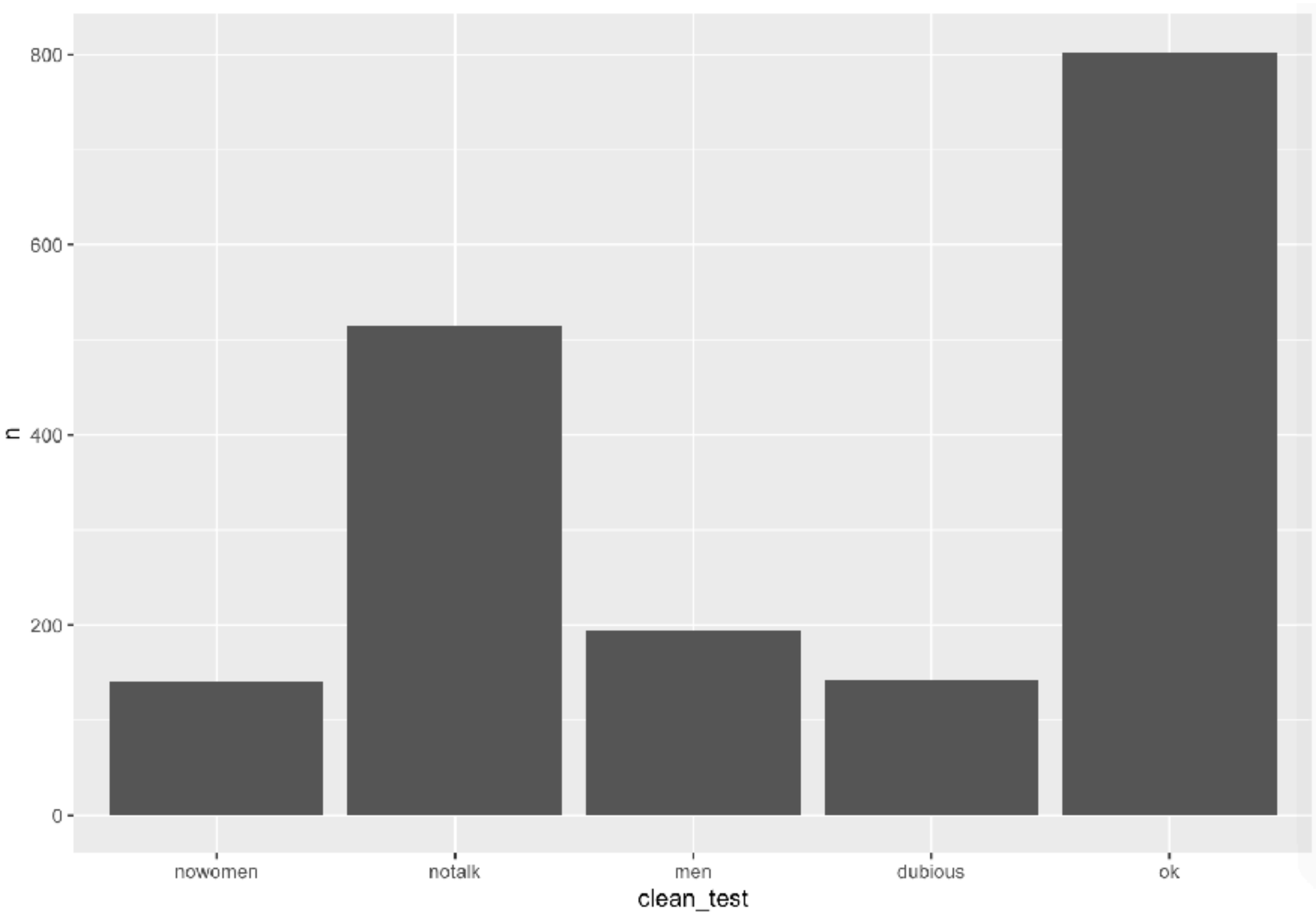


Factors

Factors let us impose order on categories



```
 [ reached getOption("max.print") -- omitted 7
Levels: nowomen < notalk < men < dubious < ok
```



Two most common ways of using factors

By hand

```
factor(variable, levels = c(“first”, “second”, …))
```

Order by frequency of value

```
fct_infreq(variable)
```

Order by another variable

```
fct_reorder(variable, 2nd variable)
```

Projects and working directories

Everything on your computer has a location

The location has an “address” called a file path

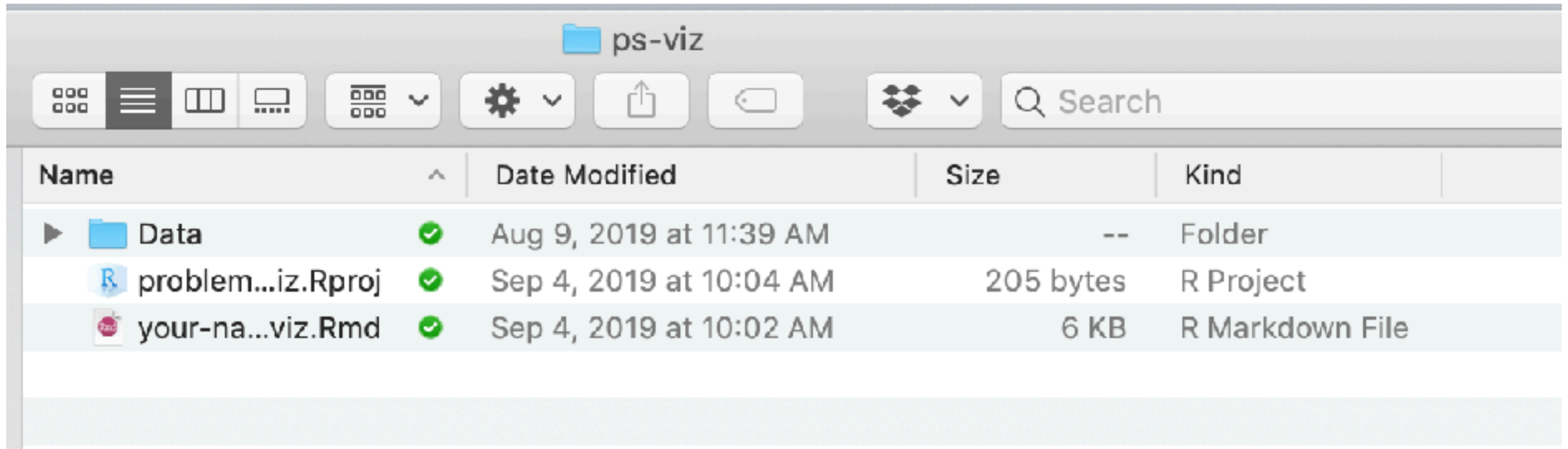
You can find the path for any file via right-click, info

```
/Users/JuamnTellez/Dropbox/poli-301/lectures/07-projects.key
```

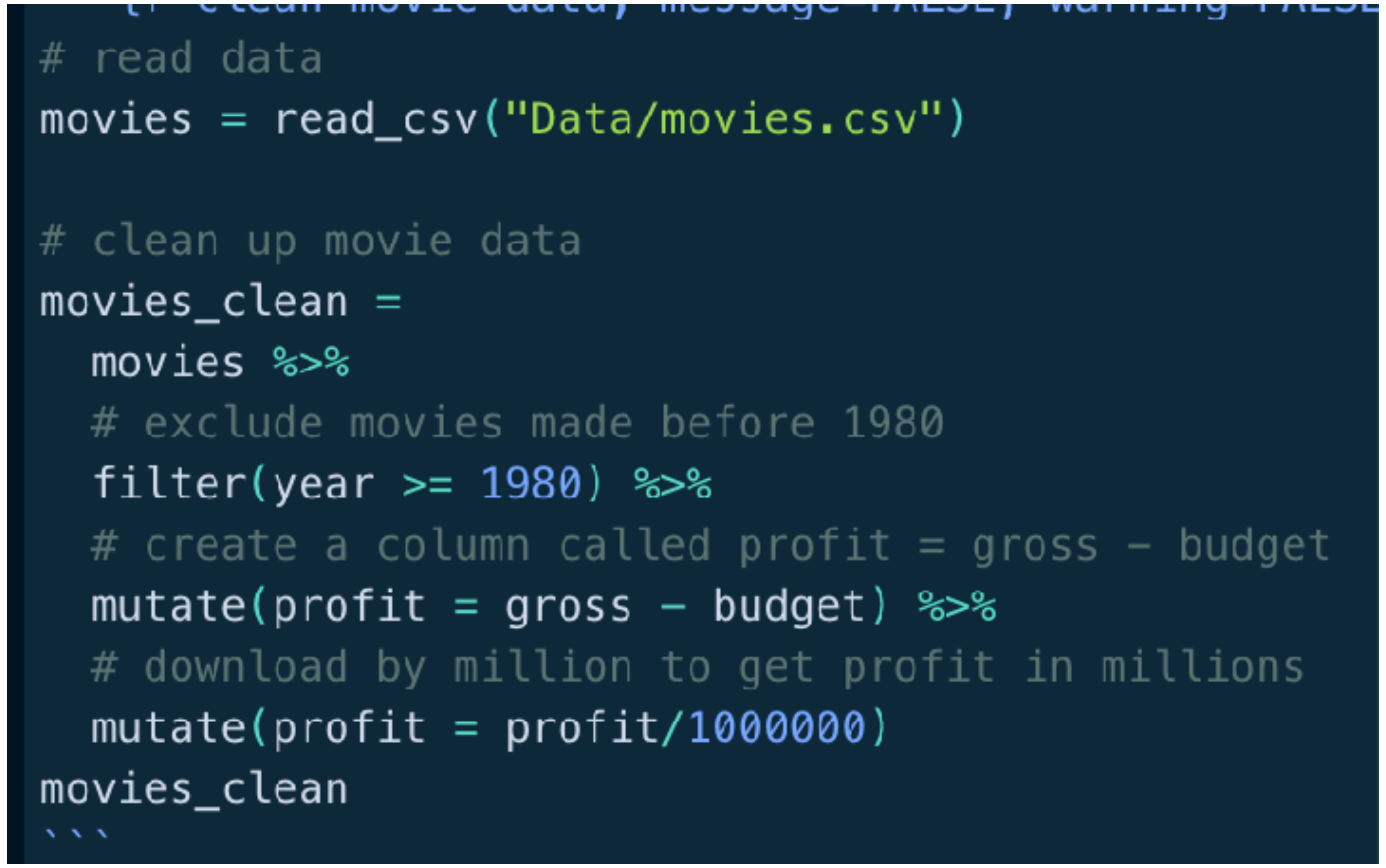
Projects and working directories

Most times we use data that lives in our computer

How does R know where it is?



Name		Date Modified	Size	Kind
▶ Data	✓	Aug 9, 2019 at 11:39 AM	--	Folder
problem...iz.Rproj	✓	Sep 4, 2019 at 10:04 AM	205 bytes	R Project
your-na...viz.Rmd	✓	Sep 4, 2019 at 10:02 AM	6 KB	R Markdown File



```
{r clean movie data, message=FALSE, warning=FALSE
# read data
movies = read_csv("Data/movies.csv")

# clean up movie data
movies_clean =
  movies %>%
  # exclude movies made before 1980
  filter(year >= 1980) %>%
  # create a column called profit = gross - budget
  mutate(profit = gross - budget) %>%
  # download by million to get profit in millions
  mutate(profit = profit/1000000)
movies_clean
```
```
```

Projects and working directories

Rstudio projects help keep you organized

Keep projects in one folder

They establish a *working directory* (a point of reference) for your file paths

```
/Users/JuamnTellez/Dropbox/poli-301/problem-sets/answers/ps1/Data/movies.csv
```

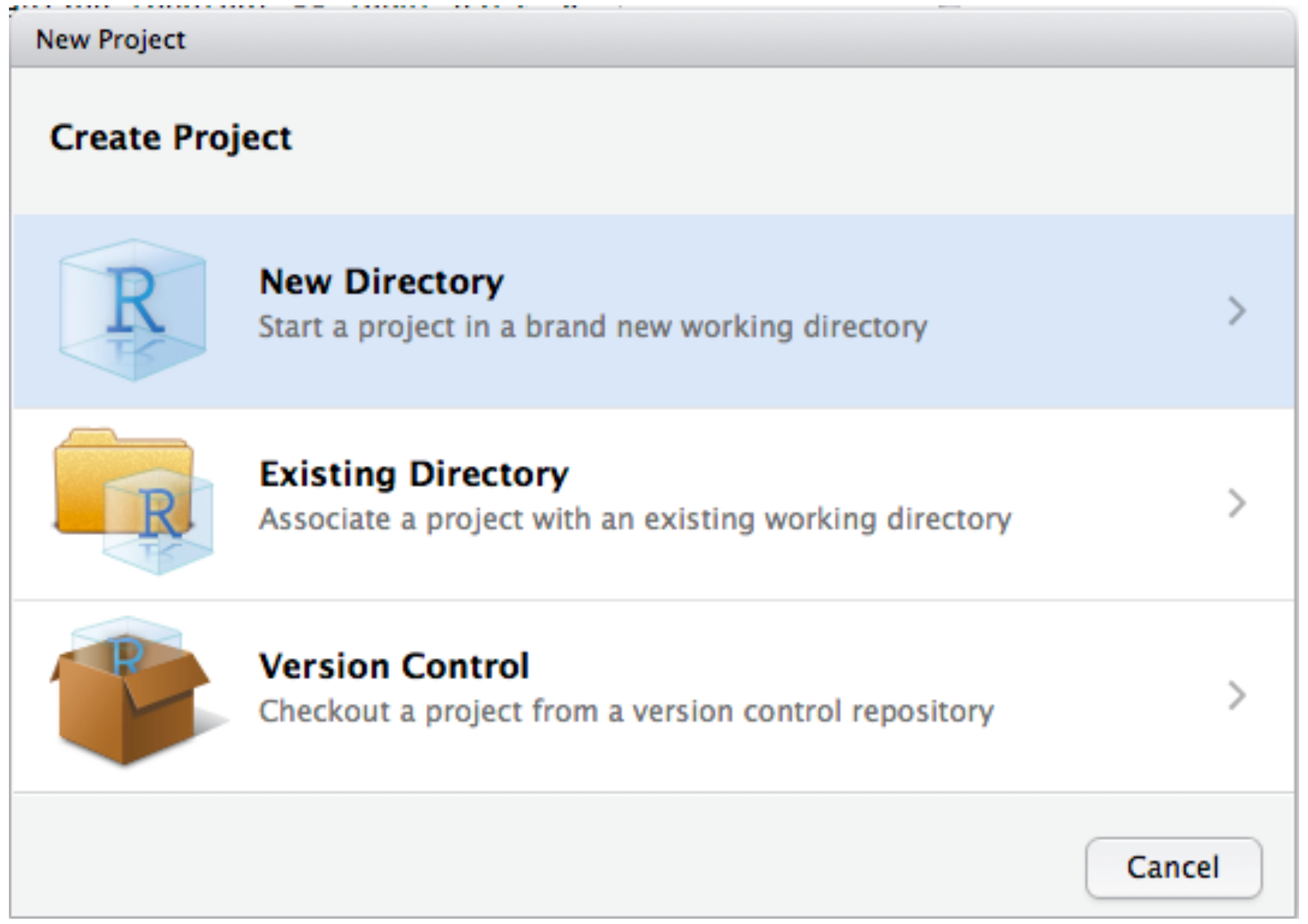
VS

```
Data/movies.csv
```

Workflow

Make new project for each homework

Put it somewhere you can find!



New Project

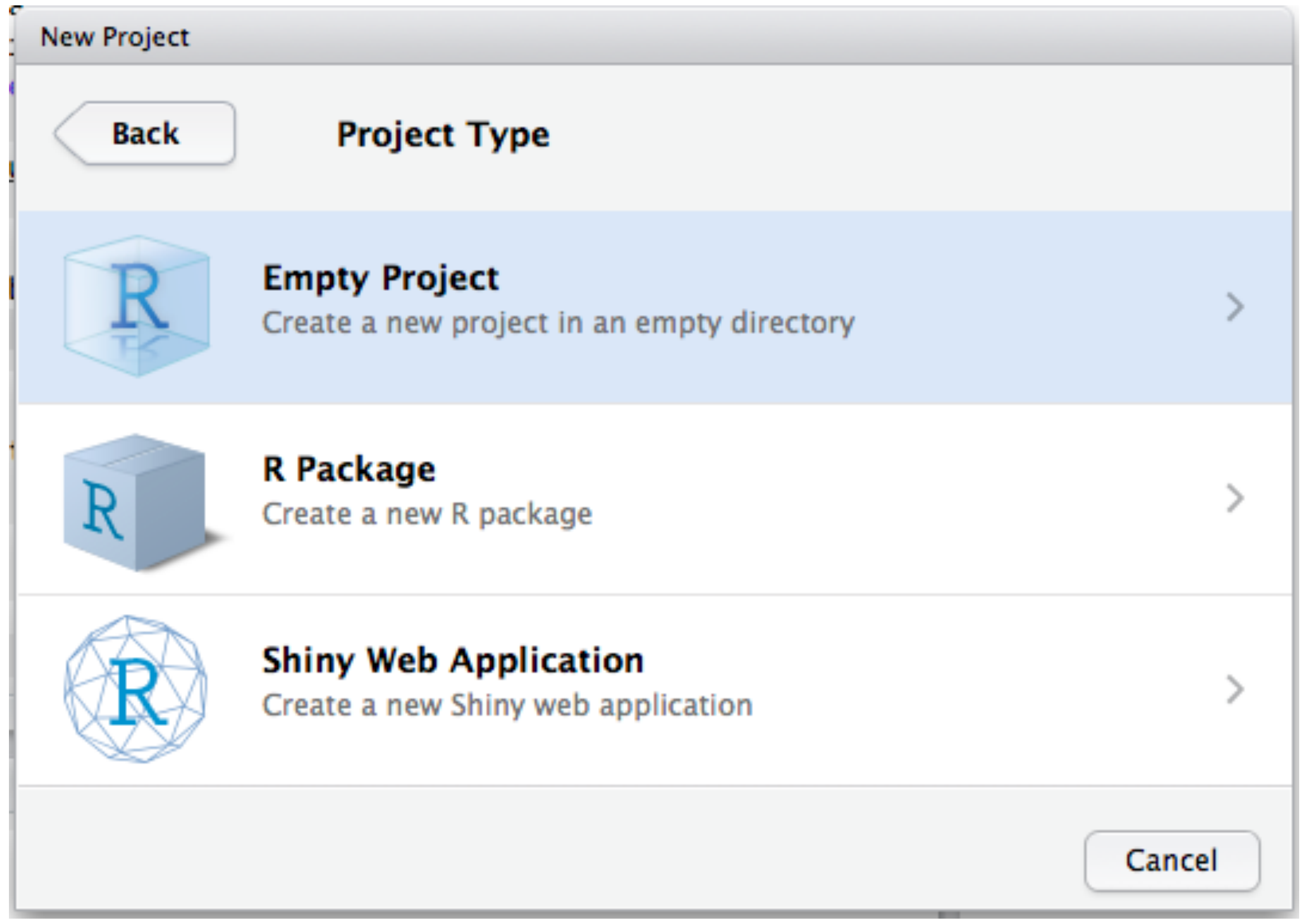
Create Project

New Directory
Start a project in a brand new working directory

Existing Directory
Associate a project with an existing working directory

Version Control
Checkout a project from a version control repository

Cancel



New Project

Back

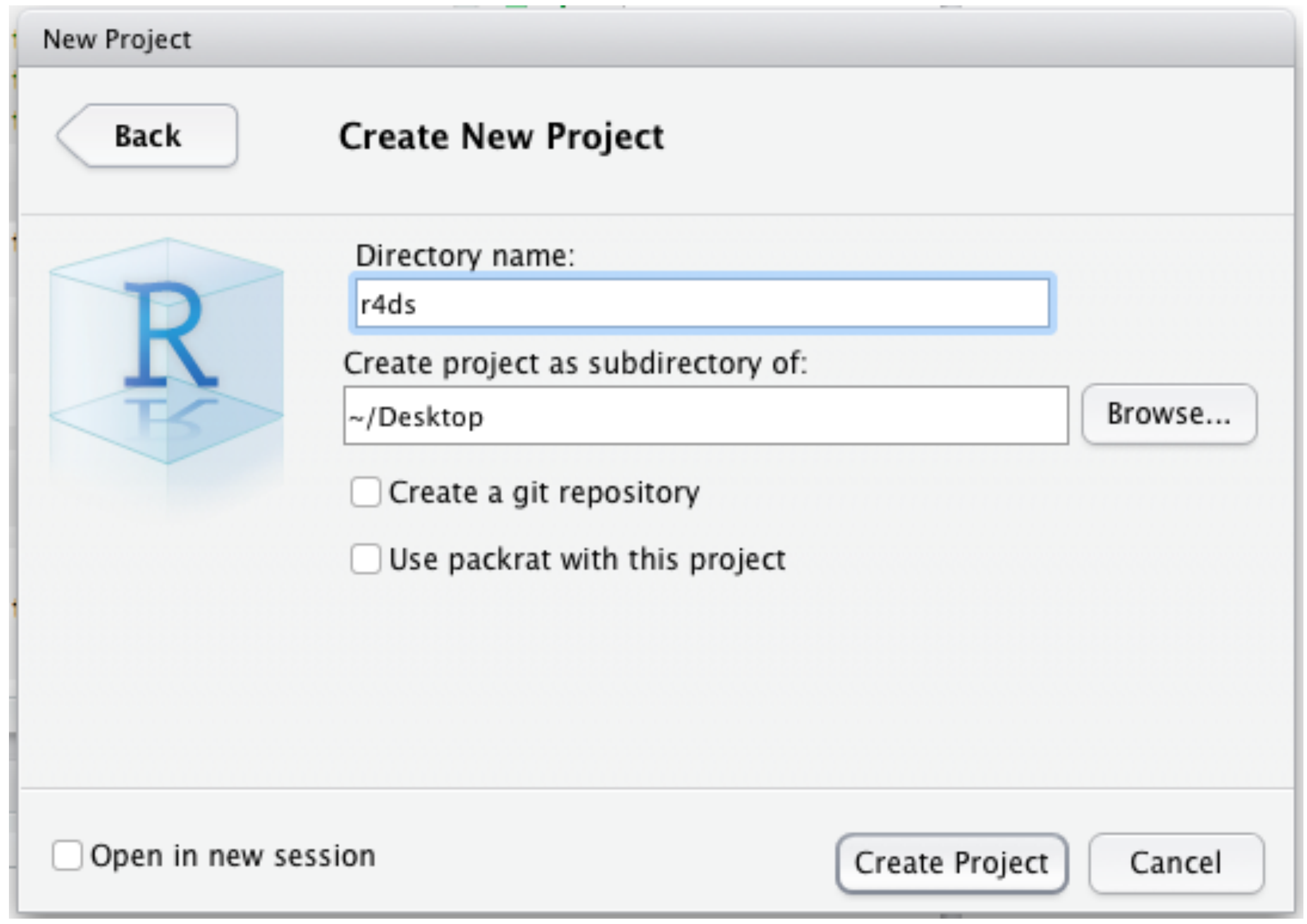
Project Type

Empty Project
Create a new project in an empty directory

R Package
Create a new R package

Shiny Web Application
Create a new Shiny web application

Cancel



New Project

Back

Create New Project

Directory name: r4ds

Create project as subdirectory of: ~/Desktop

Browse...

□ Create a git repository

□ Use packrat with this project

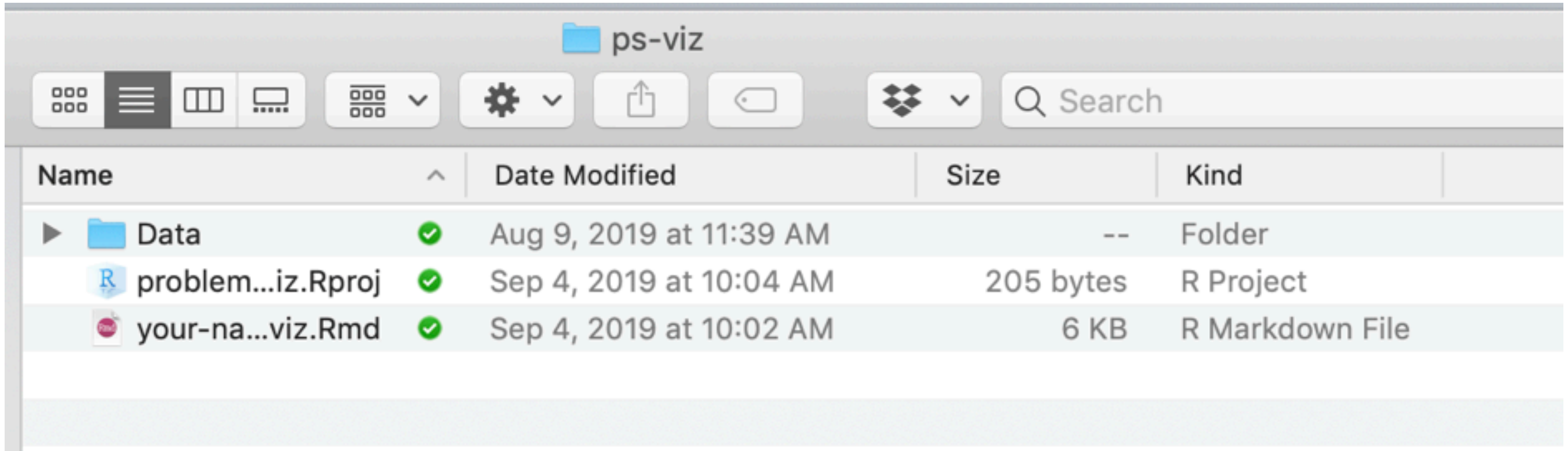
□ Open in new session

Create Project

Cancel

Workflow

Make a folder for data



Name	Date Modified	Size	Kind
Data	Aug 9, 2019 at 11:39 AM	--	Folder
problem...iz.Rproj	Sep 4, 2019 at 10:04 AM	205 bytes	R Project
your-na...viz.Rmd	Sep 4, 2019 at 10:02 AM	6 KB	R Markdown File

Double-click on .Rproj file to start coding