Data Science for Everyone – Data Wrangling – Tidy Data

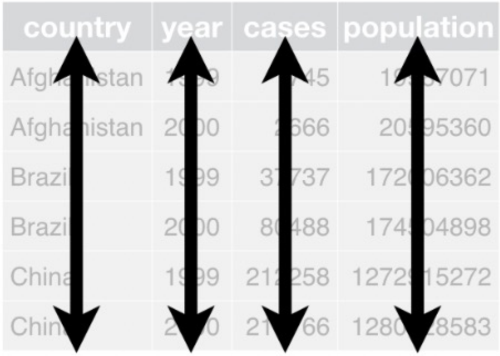
Dr. Ab Mosca (they/them)

Slides based off slides courtesy of Jordan Crouser (https://jcrouser.github.io/)

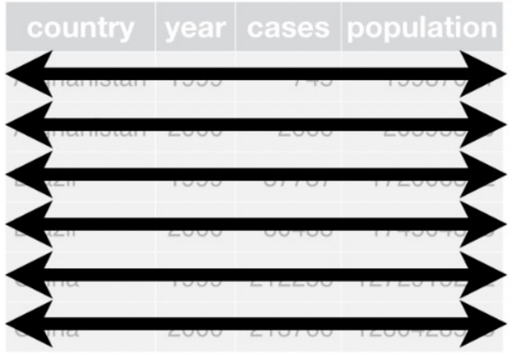
Plan for Today

- Define Tidy data
- Clean messy data

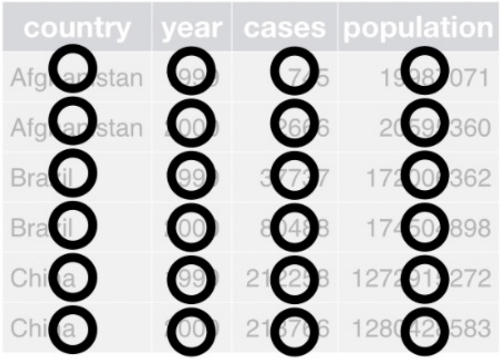
Reminder: Table Vocabulary



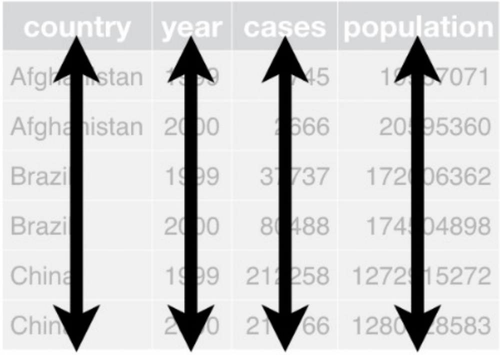
columns



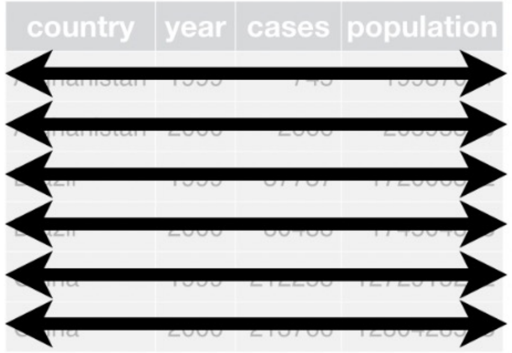
rows



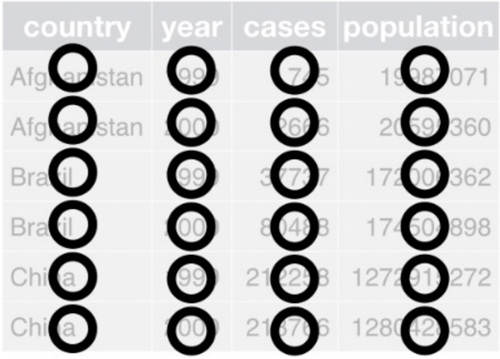
cells



variables



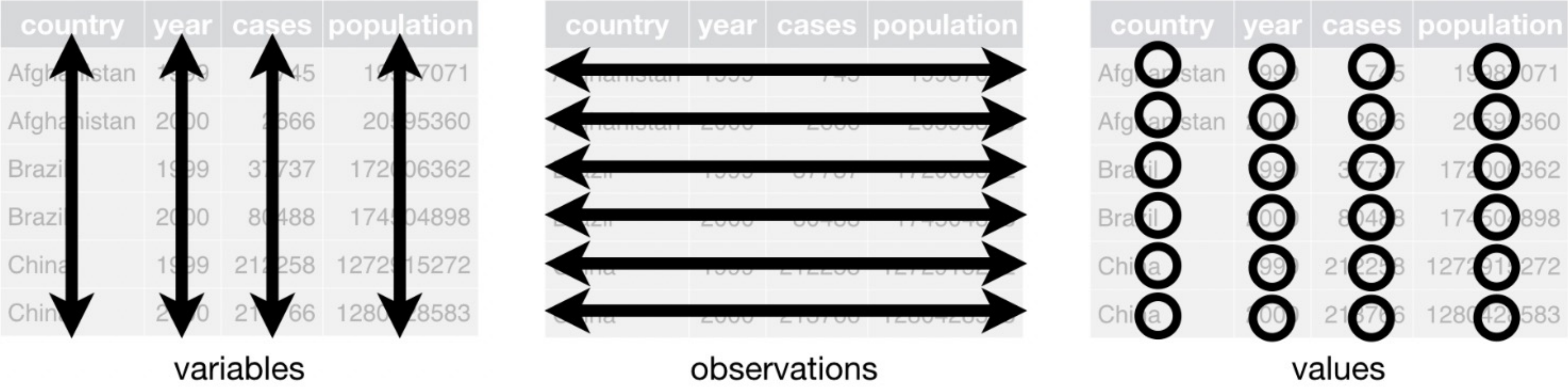
observations



values

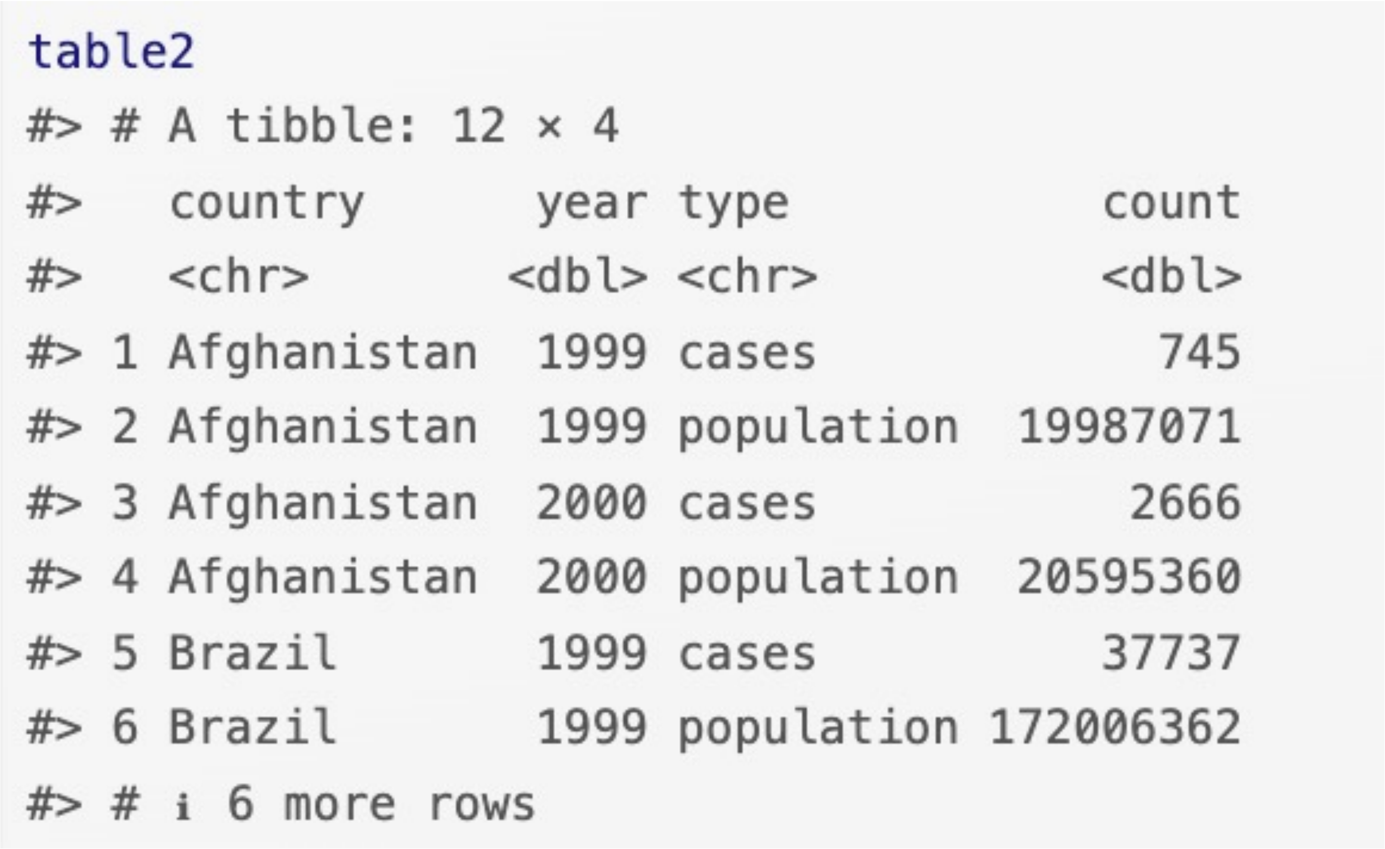
Reminder: Table Vocabulary

- When data is tidy, every column is a variable, every row is an observation, and every value has it’s own cell
- We won’t always get data in this format; sometimes data collectors record data in different ways



Messy / Tidy Data

- When data is tidy, every column is a variable, every row is an observation, and every value has it's own cell

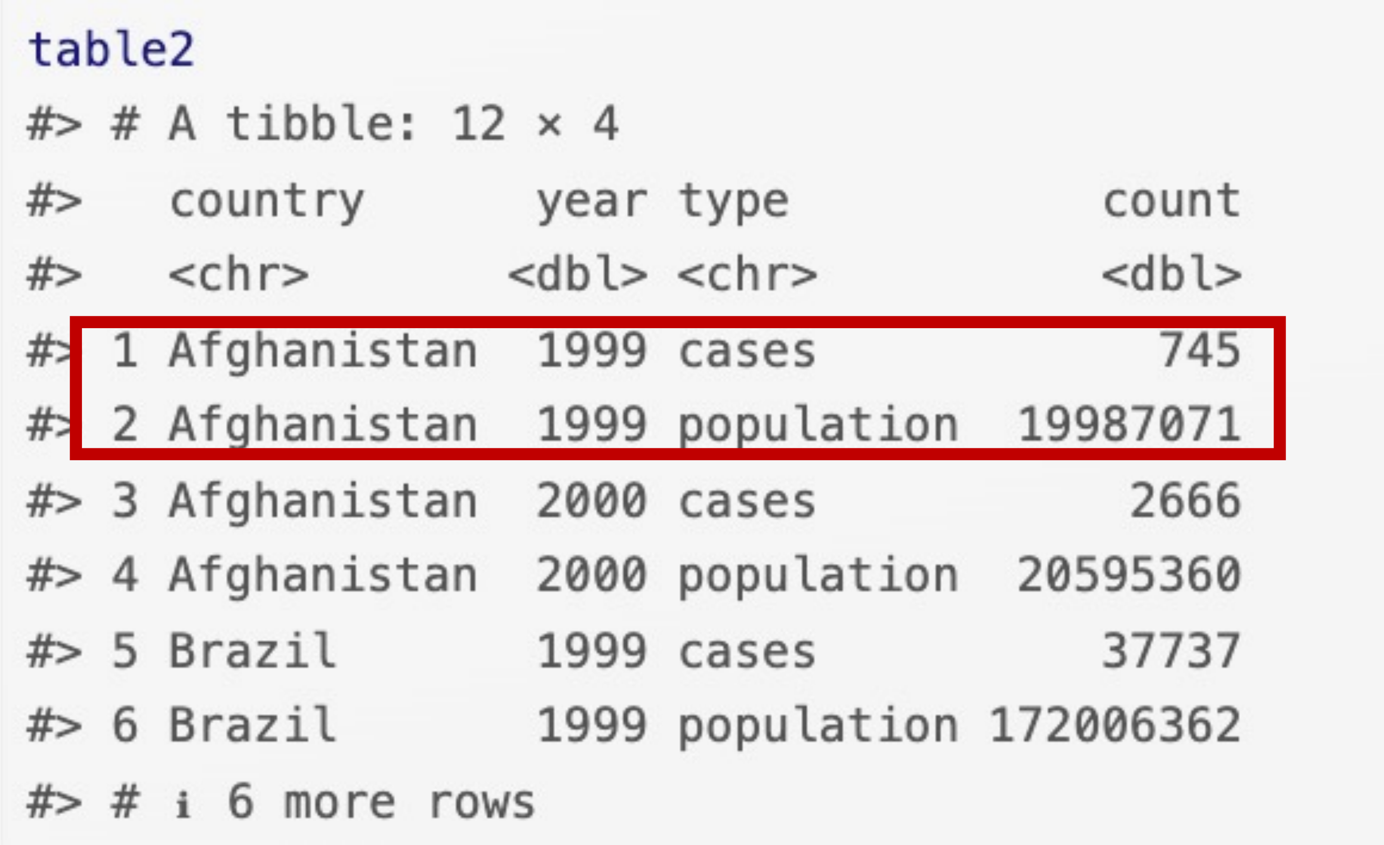


```
table2
#> # A tibble: 12 × 4
#>   country      year type            count
#>   <chr>       <dbl> <chr>           <dbl>
#> 1 Afghanistan  1999 cases             745
#> 2 Afghanistan  1999 population   19987071
#> 3 Afghanistan  2000 cases            2666
#> 4 Afghanistan  2000 population   20595360
#> 5 Brazil       1999 cases           37737
#> 6 Brazil       1999 population  172006362
#> # ℹ 6 more rows
```

Tidy or messy? Why?

Messy / Tidy Data

- When data is tidy, every column is a variable, every row is an observation, and every value has it's own cell

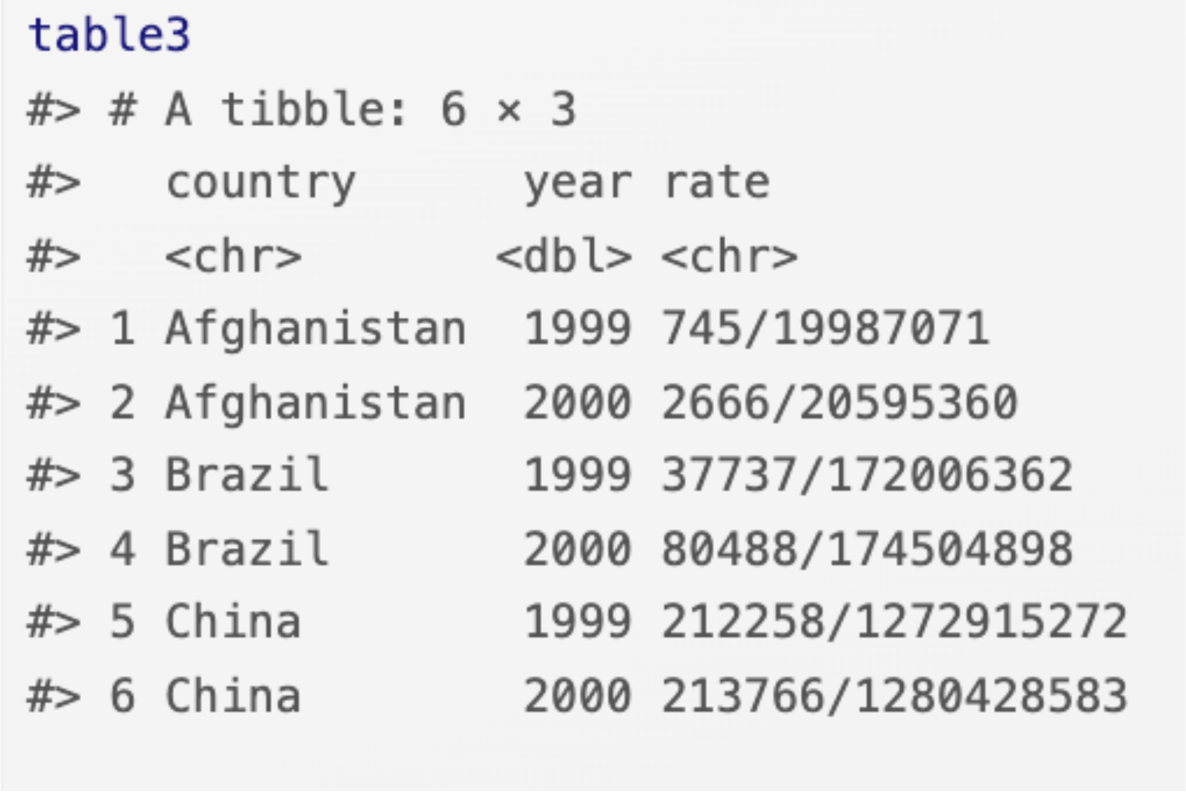


```
table2
#> # A tibble: 12 × 4
#>   country      year type            count
#>   <chr>       <dbl> <chr>           <dbl>
#> 1 Afghanistan  1999 cases             745
#> 2 Afghanistan  1999 population   19987071
#> 3 Afghanistan  2000 cases            2666
#> 4 Afghanistan  2000 population   20595360
#> 5 Brazil       1999 cases           37737
#> 6 Brazil       1999 population  172006362
#> # ℹ 6 more rows
```

Tidy o messy? Why?

Messy / Tidy Data

- When data is tidy, every column is a variable, every row is an observation, and every value has it's own cell

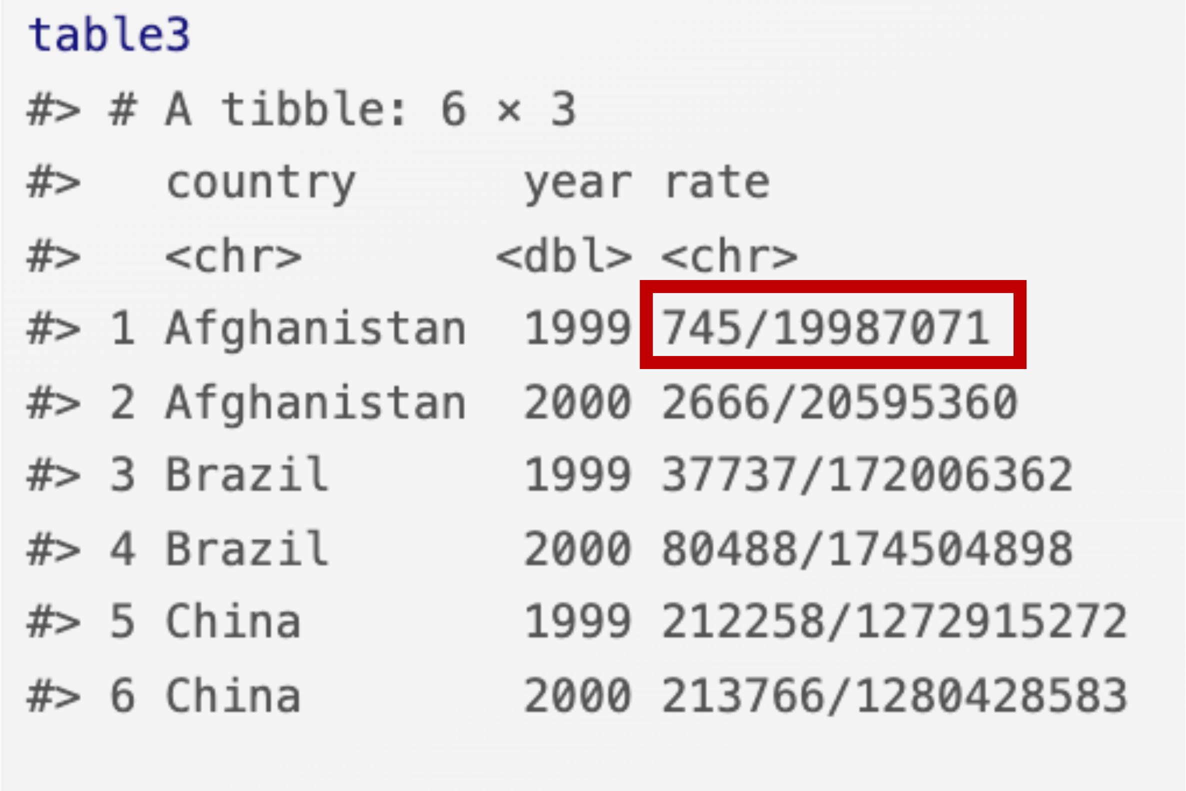


```
table3
#> # A tibble: 6 × 3
#>   country      year rate
#>   <chr>       <dbl> <chr>
#> 1 Afghanistan  1999 745/19987071
#> 2 Afghanistan  2000 2666/20595360
#> 3 Brazil       1999 37737/172006362
#> 4 Brazil       2000 80488/174504898
#> 5 China        1999 212258/1272915272
#> 6 China        2000 213766/1280428583
```

Tidy or messy? Why?

Messy / Tidy Data

- When data is tidy, every column is a variable, every row is an observation, and every value has it’s own cell

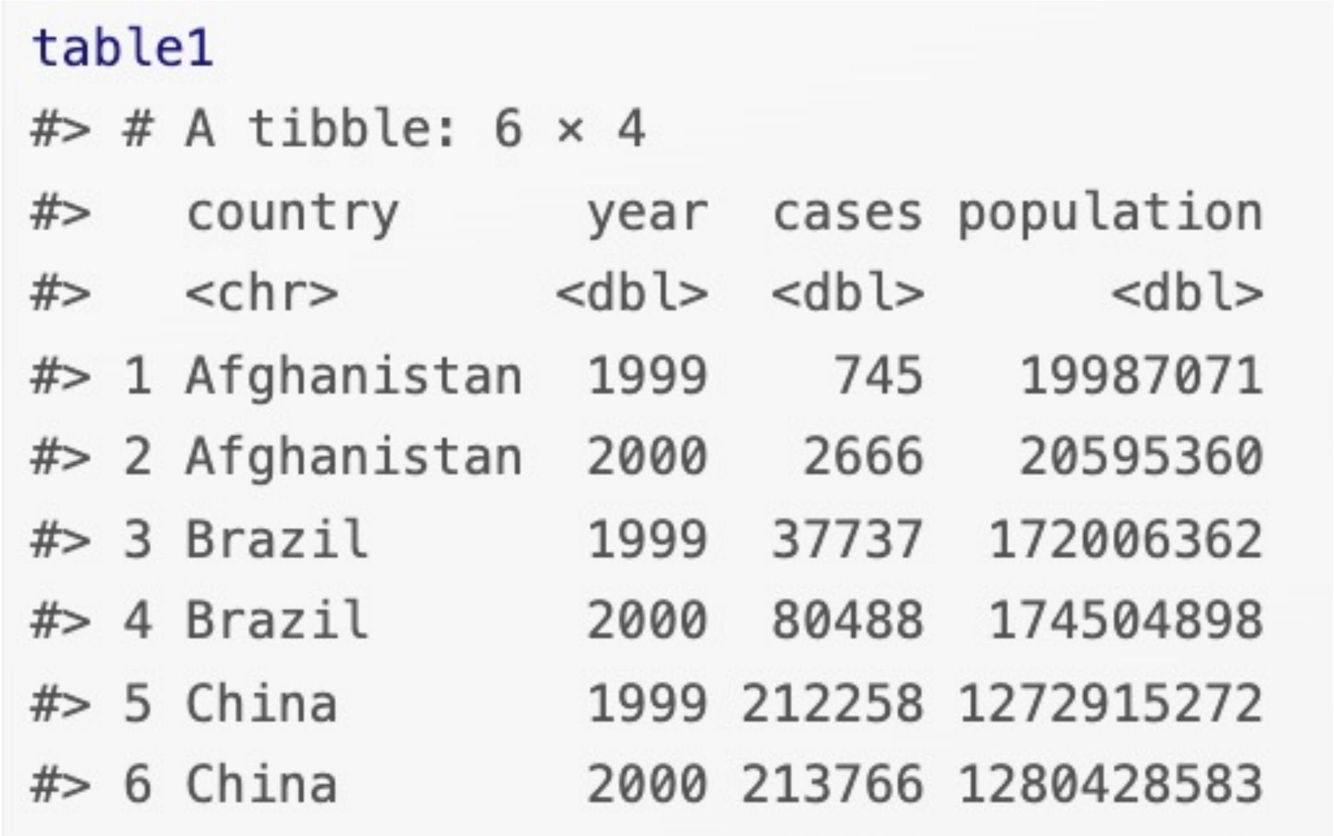


```
table3
#> # A tibble: 6 × 3
#>   country      year rate
#>   <chr>       <dbl> <chr>
#> 1 Afghanistan  1999 745/19987071
#> 2 Afghanistan  2000 2666/20595360
#> 3 Brazil       1999 37737/172006362
#> 4 Brazil       2000 80488/174504898
#> 5 China        1999 212258/1272915272
#> 6 China        2000 213766/1280428583
```

Tidy o messy? Why?

Messy / Tidy Data

- When data is tidy, every column is a variable, every row is an observation, and every value has it's own cell

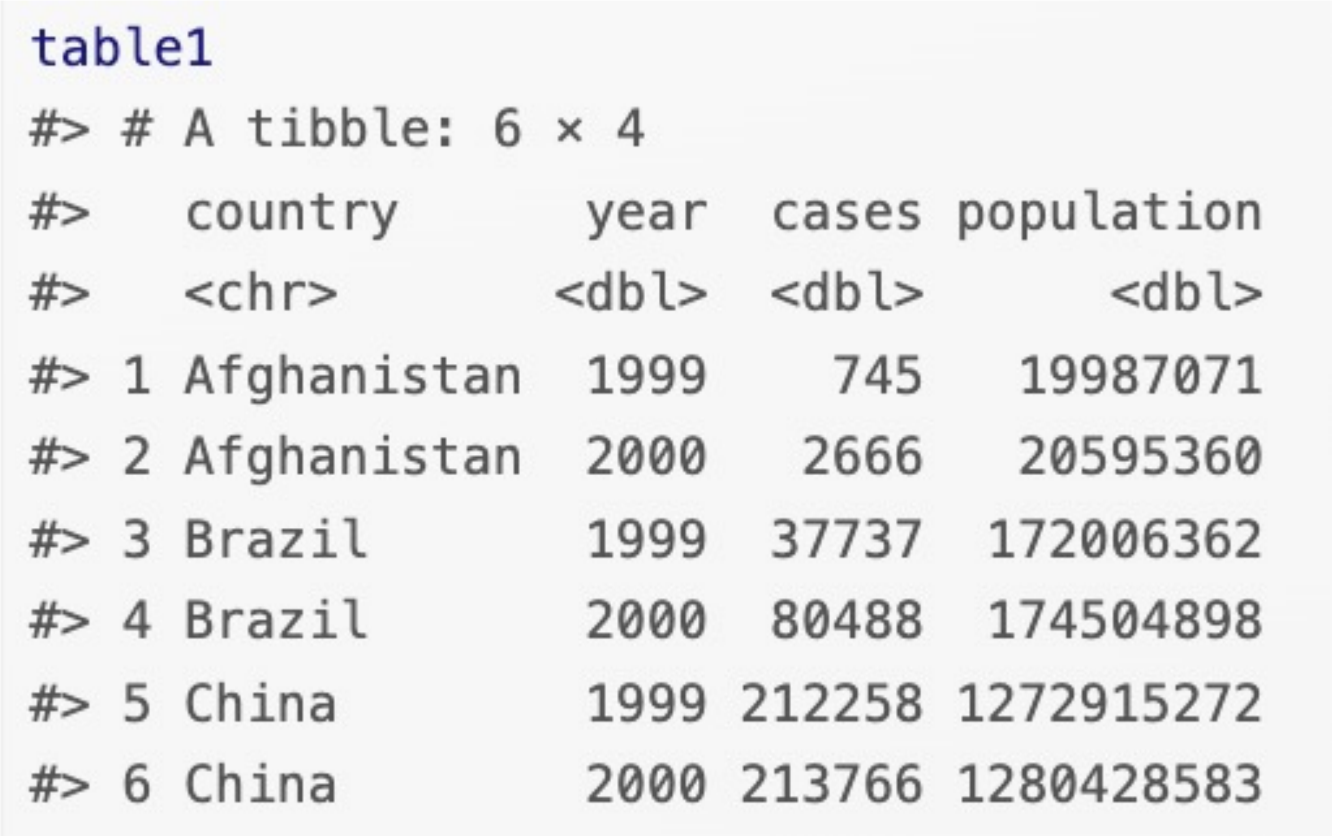


```
table1
#> # A tibble: 6 × 4
#>   country      year  cases population
#>   <chr>       <dbl>  <dbl>      <dbl>
#> 1 Afghanistan  1999    745   19987071
#> 2 Afghanistan  2000   2666   20595360
#> 3 Brazil       1999  37737  172006362
#> 4 Brazil       2000  80488  174504898
#> 5 China        1999 212258 1272915272
#> 6 China        2000 213766 1280428583
```

Tidy or messy? Why?

Messy / Tidy Data

- When data is tidy, every column is a variable, every row is an observation, and every value has it’s own cell

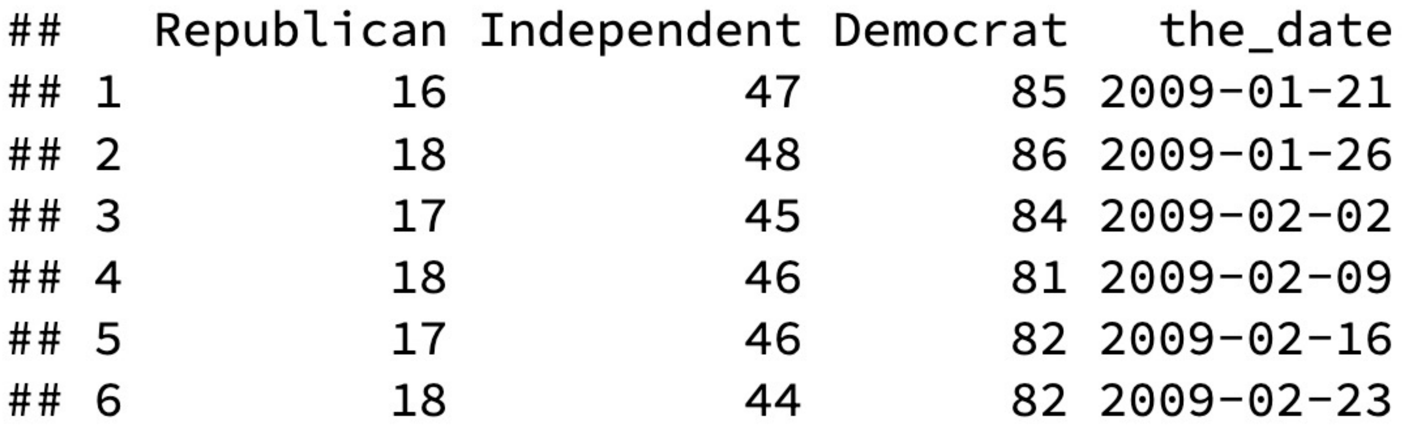


```
table1
#> # A tibble: 6 × 4
#>   country      year  cases population
#>   <chr>       <dbl>  <dbl>      <dbl>
#> 1 Afghanistan  1999    745   19987071
#> 2 Afghanistan  2000   2666   20595360
#> 3 Brazil       1999  37737  172006362
#> 4 Brazil       2000  80488  174504898
#> 5 China        1999 212258 1272915272
#> 6 China        2000 213766 1280428583
```

Tidy or messy? Why?

Messy / Tidy Data

- When data is tidy, every column is a variable, every row is an observation, and every value has it's own cell

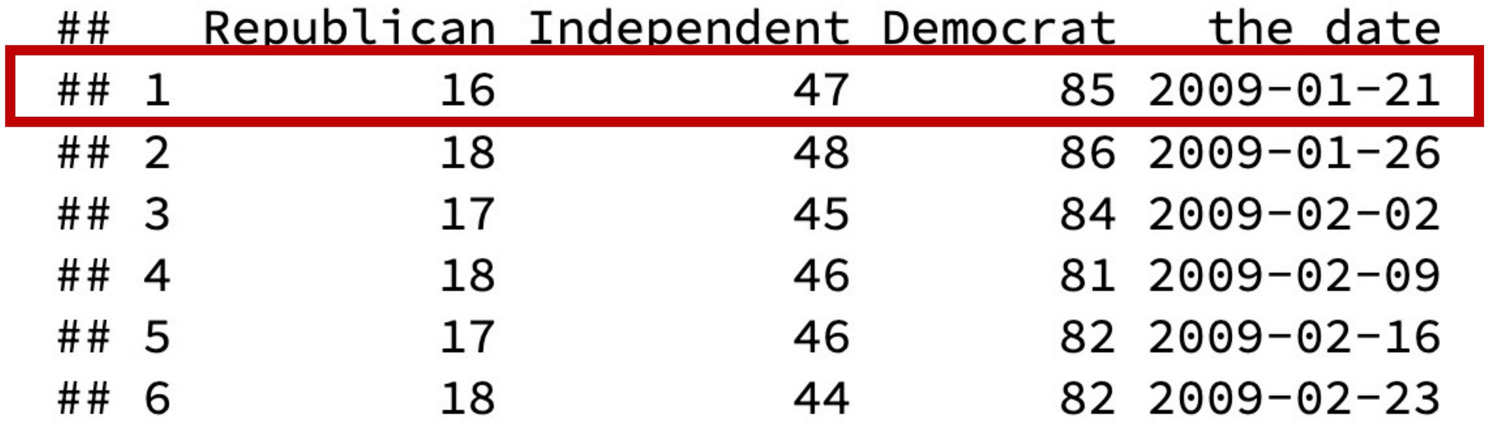


```
##    Republican Independent Democrat   the_date
## 1          16          47       85 2009-01-21
## 2          18          48       86 2009-01-26
## 3          17          45       84 2009-02-02
## 4          18          46       81 2009-02-09
## 5          17          46       82 2009-02-16
## 6          18          44       82 2009-02-23
```

Tidy or messy? Why?

Messy / Tidy Data

- When data is tidy, every column is a variable, every row is an observation, and every value has it's own cell

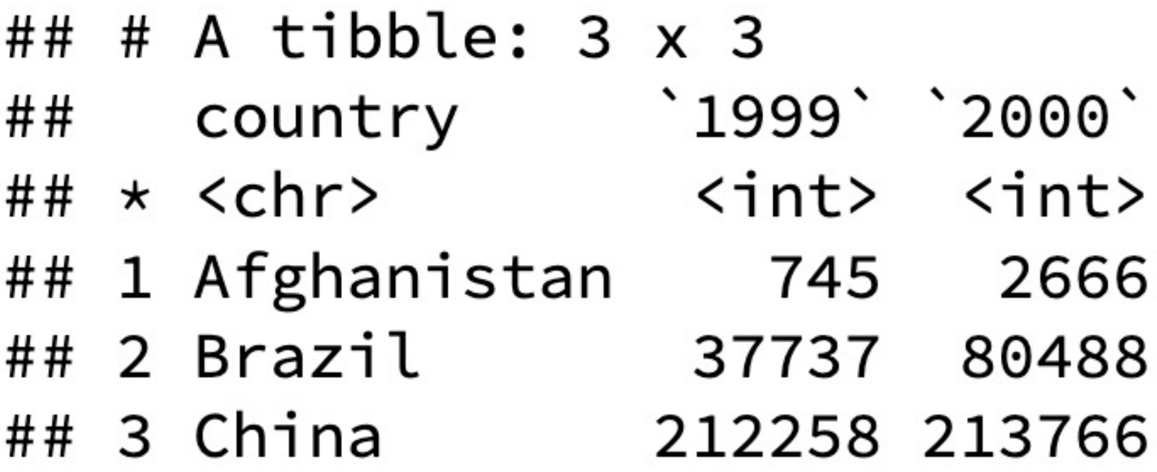


```
##    Republican Independent Democrat   the date
## 1          16          47       85 2009-01-21
## 2          18          48       86 2009-01-26
## 3          17          45       84 2009-02-02
## 4          18          46       81 2009-02-09
## 5          17          46       82 2009-02-16
## 6          18          44       82 2009-02-23
```

Tidy or messy? Why?

Messy / Tidy Data

- When data is tidy, every column is a variable, every row is an observation, and every value has it's own cell

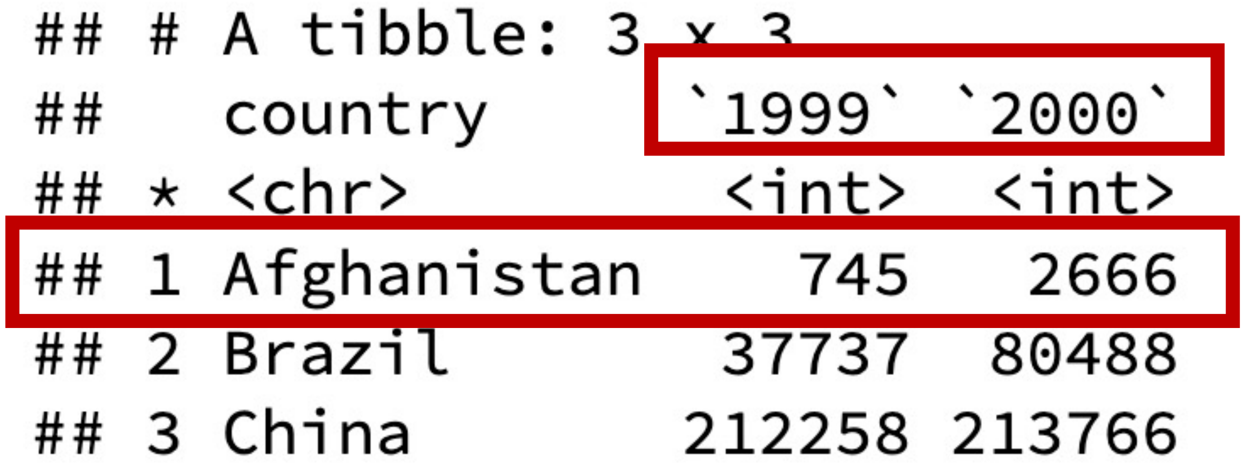


```
## # A tibble: 3 x 3
##   country     `1999` `2000`
## * <chr>        <int>  <int>
## 1 Afghanistan    745   2666
## 2 Brazil       37737  80488
## 3 China       212258 213766
```

Tidy or messy? Why?

Messy / Tidy Data

- When data is tidy, every column is a variable, every row is an observation, and every value has it’s own cell



```
## # A tibble: 3 x 3
##   country      `1999` `2000`
## * <chr>         <int>  <int>
## 1 Afghanistan     745   2666
## 2 Brazil        37737  80488
## 3 China        212258 213766
```

Tidy or messy? Why?

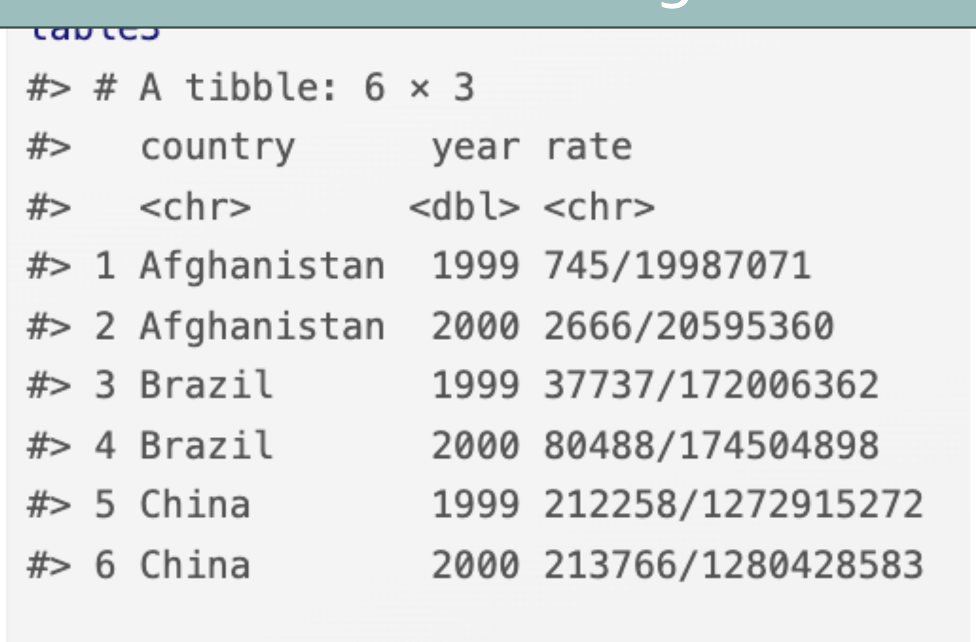
Why care?

- Uniformity makes learning tools easier
- Most functions in the tidyverse are designed to work with tidy data (**tidy**verse, get it? ☺)
    - Ex. ggplot, dplyr

Why care?

- Uniformity makes learning tools easier
- Most functions in the tidyverse are designed to work with tidy data (**tidy**verse, get it? ☺)
  - Ex. ggplot, dplyr

Can you calculate case rate with one line of code for table3?

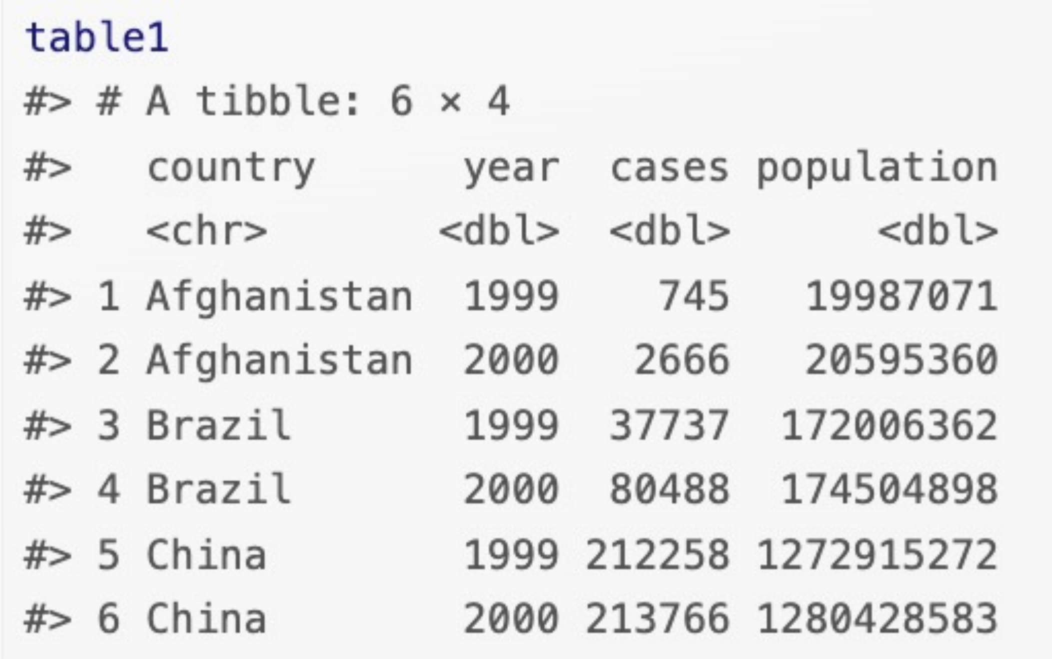


```
table3
#> # A tibble: 6 × 3
#>   country      year rate
#>   <chr>       <dbl> <chr>
#> 1 Afghanistan  1999 745/19987071
#> 2 Afghanistan  2000 2666/20595360
#> 3 Brazil       1999 37737/172006362
#> 4 Brazil       2000 80488/174504898
#> 5 China        1999 212258/1272915272
#> 6 China        2000 213766/1280428583
```

Why care?

- Uniformity makes learning tools easier
- Most functions in the tidyverse are designed to work with tidy data (**tidy**verse, get it? ☺)
  - Ex. ggplot, dplyr

What about table1?



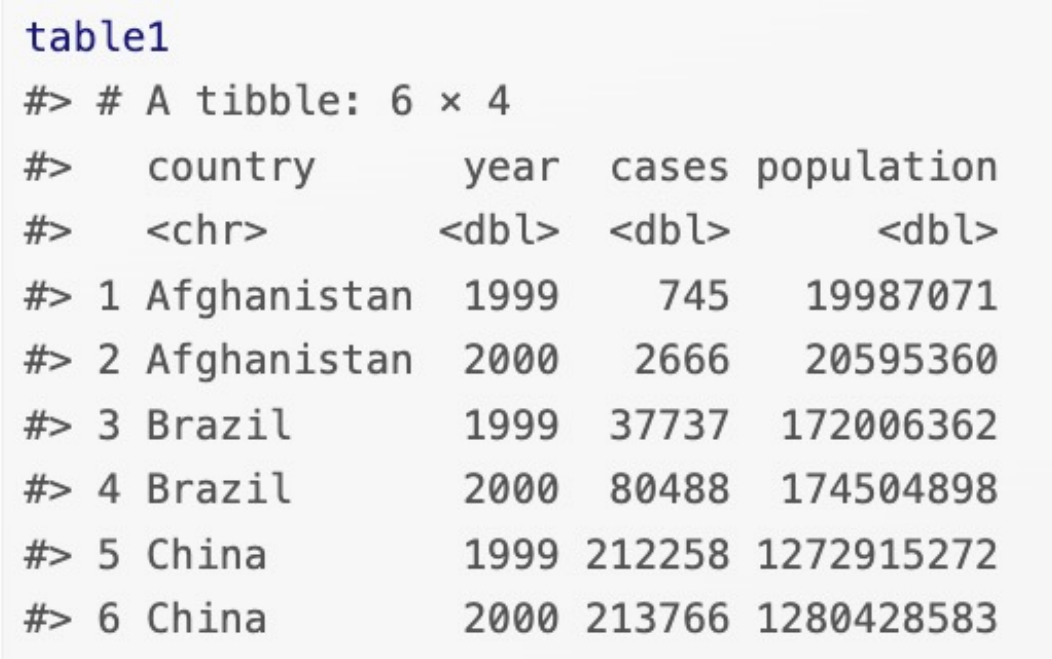
```
table1
#> # A tibble: 6 × 4
#>   country      year  cases population
#>   <chr>       <dbl>  <dbl>      <dbl>
#> 1 Afghanistan  1999    745   19987071
#> 2 Afghanistan  2000   2666   20595360
#> 3 Brazil       1999  37737  172006362
#> 4 Brazil       2000  80488  174504898
#> 5 China        1999 212258 1272915272
#> 6 China        2000 213766 1280428583
```

Why care?

```
table1 %>%
   mutate(rate = cases / population * 10000)
```

Ex: ggplot, dplyr

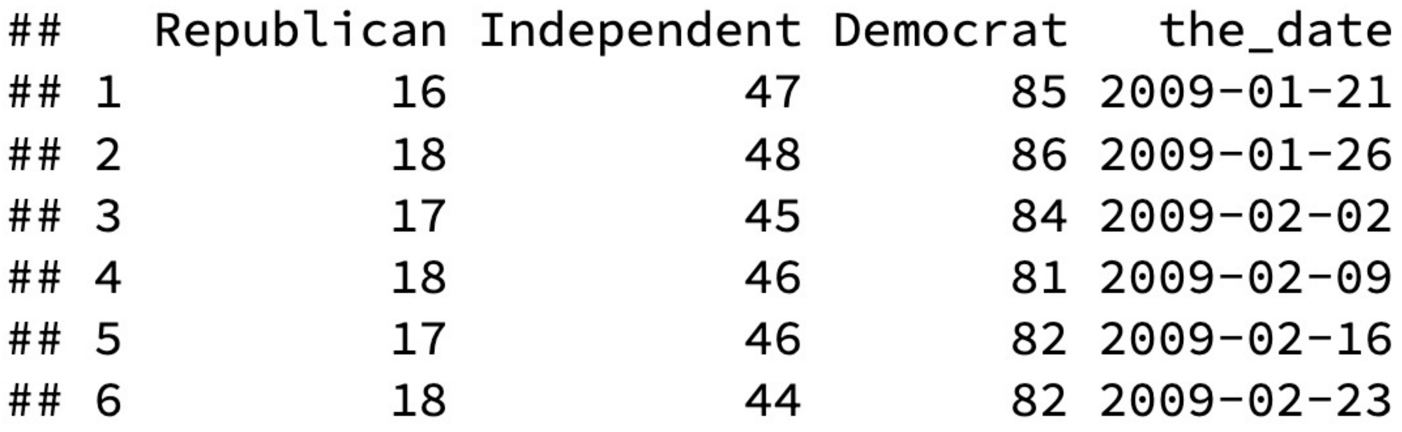
What about table1?



```
table1
#> # A tibble: 6 × 4
#>   country      year  cases population
#>   <chr>       <dbl>  <dbl>      <dbl>
#> 1 Afghanistan  1999    745   19987071
#> 2 Afghanistan  2000   2666   20595360
#> 3 Brazil       1999  37737  172006362
#> 4 Brazil       2000  80488  174504898
#> 5 China        1999 212258 1272915272
#> 6 China        2000 213766 1280428583
```

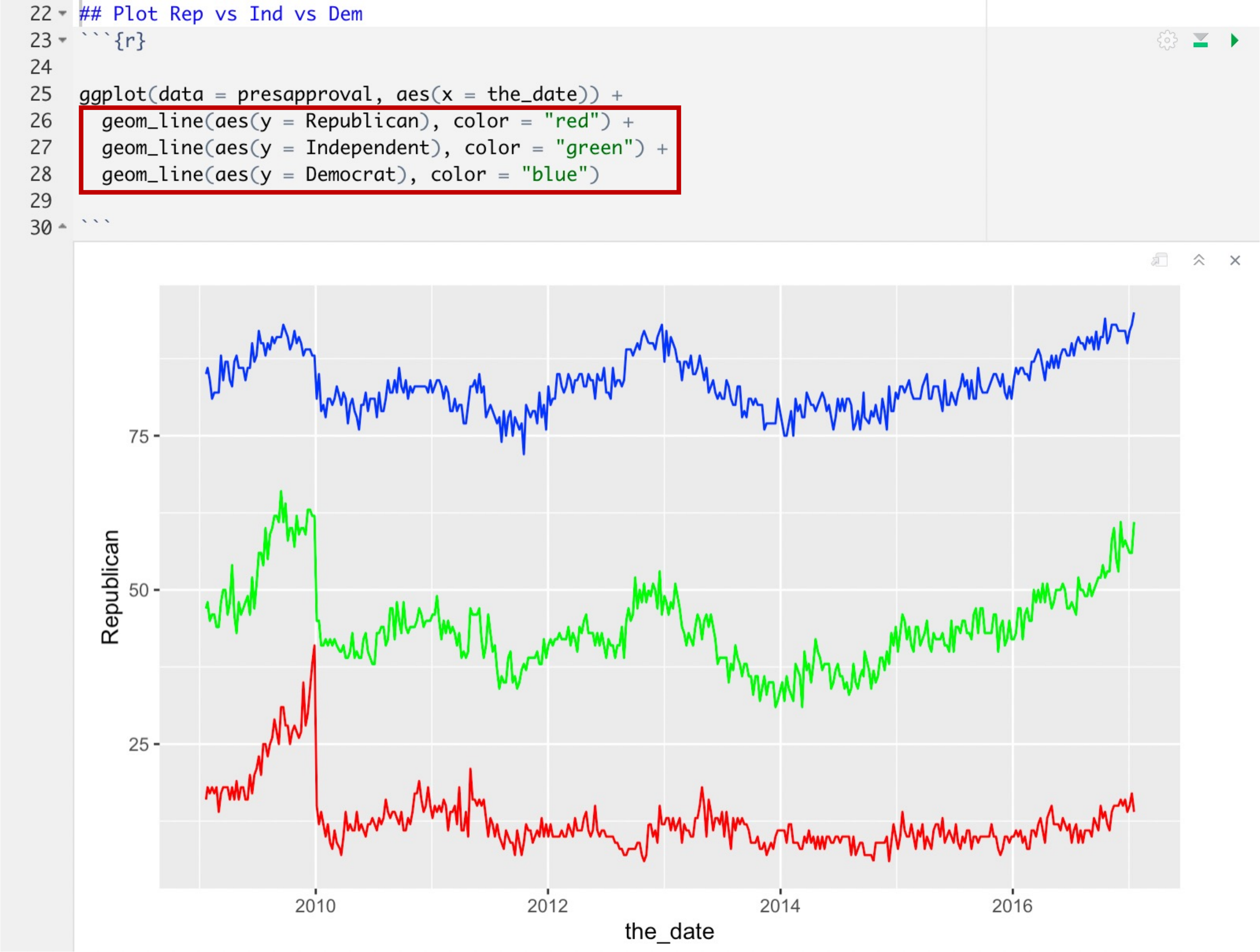
Why care?

How would you plot a line chart for this data with a line for each party?



```
##   Republican Independent Democrat   the_date
## 1         16          47       85 2009-01-21
## 2         18          48       86 2009-01-26
## 3         17          45       84 2009-02-02
## 4         18          46       81 2009-02-09
## 5         17          46       82 2009-02-16
## 6         18          44       82 2009-02-23
```

Why care?



~~~
## Plot Rep vs Ind vs Dem
```{r}

ggplot(data = presapproval, aes(x = the_date)) +
  geom_line(aes(y = Republican), color = "red") +
  geom_line(aes(y = Independent), color = "green") +
  geom_line(aes(y = Democrat), color = "blue")

```
~~~

Tidy Data

- How do we make data Tidy?

Tidy Da

- How do we make data Tidy?

Re-write this table in a tidy format

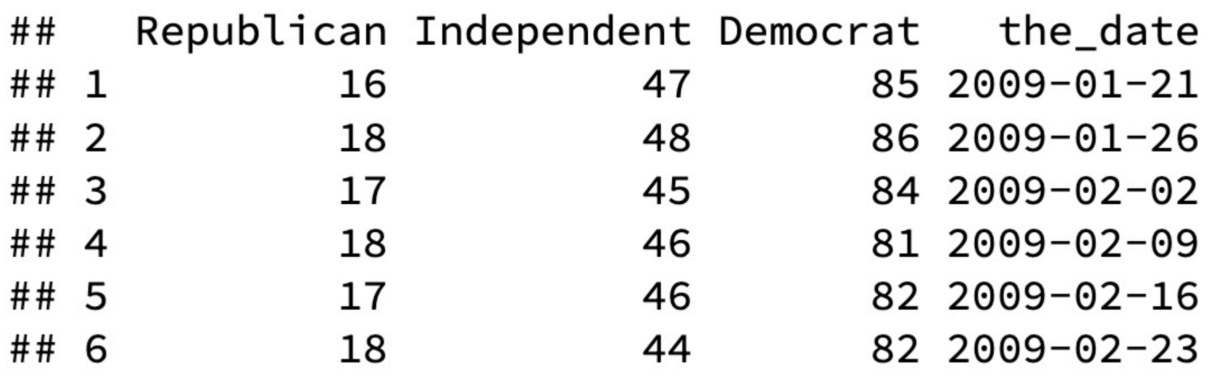
```
##    Republican Independent Democrat   the_date
## 1          16          47       85 2009-01-21
## 2          18          48       86 2009-01-26
## 3          17          45       84 2009-02-02
## 4          18          46       81 2009-02-09
## 5          17          46       82 2009-02-16
## 6          18          44       82 2009-02-23
```

These values are approval ratings

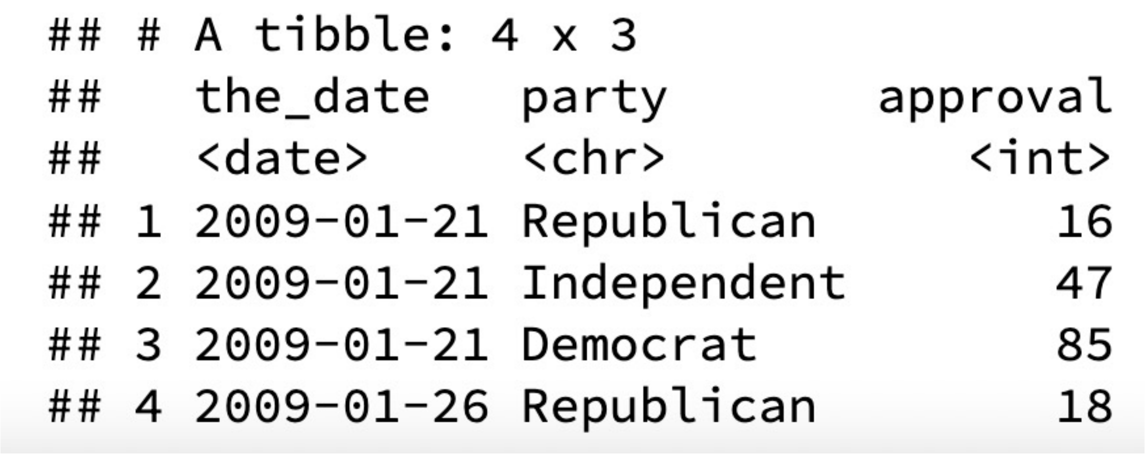
Tidy Data

- How do we make data Tidy?

Re-write this table in a tidy format



```
##   Republican Independent Democrat   the_date
## 1         16          47       85 2009-01-21
## 2         18          48       86 2009-01-26
## 3         17          45       84 2009-02-02
## 4         18          46       81 2009-02-09
## 5         17          46       82 2009-02-16
## 6         18          44       82 2009-02-23
```

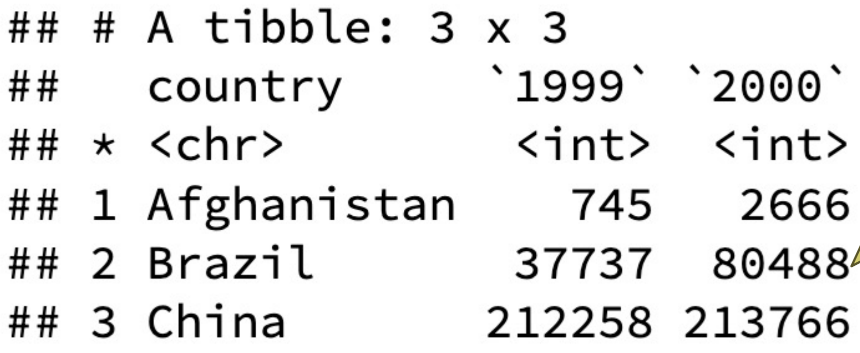


```
## # A tibble: 4 x 3
##   the_date   party       approval
##   <date>     <chr>          <int>
## 1 2009-01-21 Republican        16
## 2 2009-01-21 Independent       47
## 3 2009-01-21 Democrat          85
## 4 2009-01-26 Republican        18
```

Tidy Data

- How do we make data Tidy?

Re-write this table in a tidy format



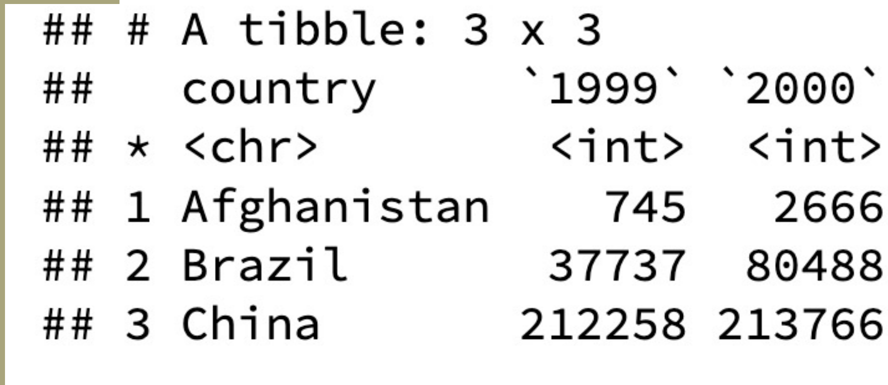
```
## # A tibble: 3 x 3
##   country     `1999` `2000`
## * <chr>        <int>  <int>
## 1 Afghanistan    745   2666
## 2 Brazil       37737  80488
## 3 China       212258 213766
```

These values are cases

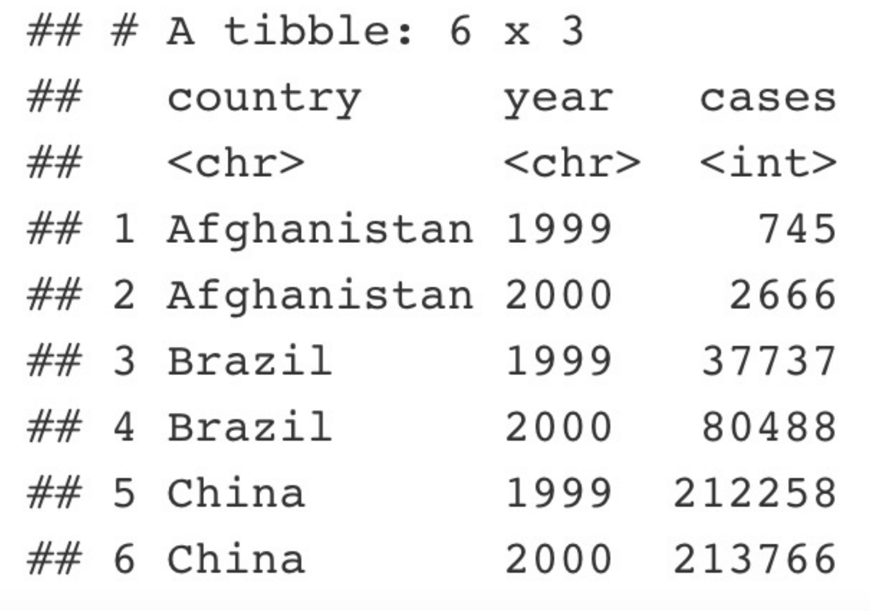
Tidy Data

- How do we make data Tidy?

Re-write this table in a tidy format



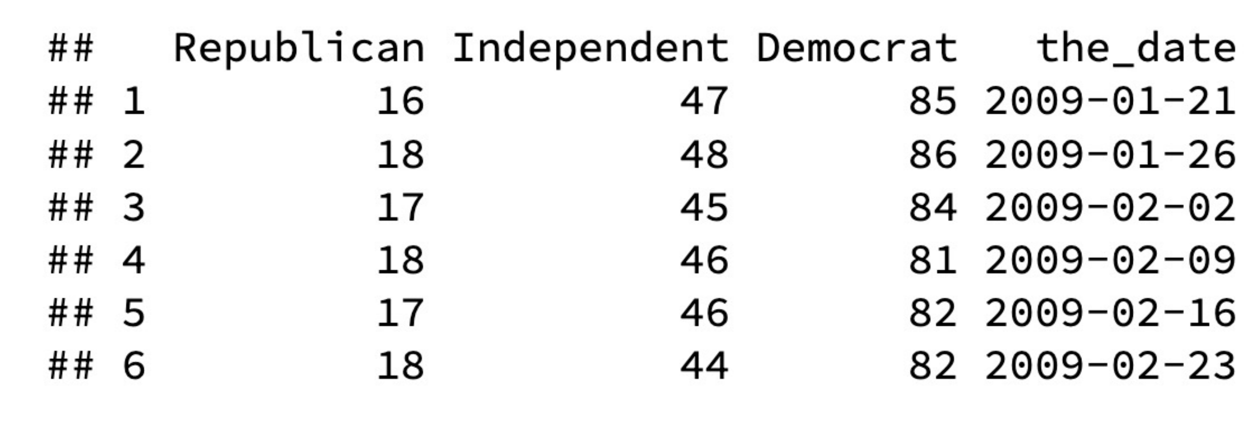
```
## # A tibble: 3 x 3
##   country     `1999` `2000`
## * <chr>        <int>  <int>
## 1 Afghanistan    745   2666
## 2 Brazil       37737  80488
## 3 China       212258 213766
```



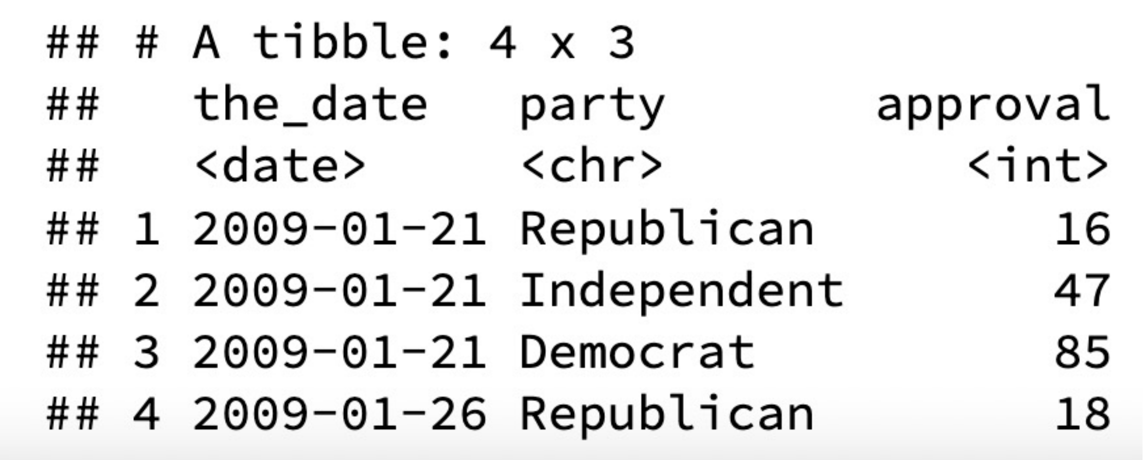
```
## # A tibble: 6 x 3
##   country     year   cases
##   <chr>       <chr>  <int>
## 1 Afghanistan 1999     745
## 2 Afghanistan 2000    2666
## 3 Brazil      1999   37737
## 4 Brazil      2000   80488
## 5 China       1999  212258
## 6 China       2000  213766
```

What is similar in these two cases?

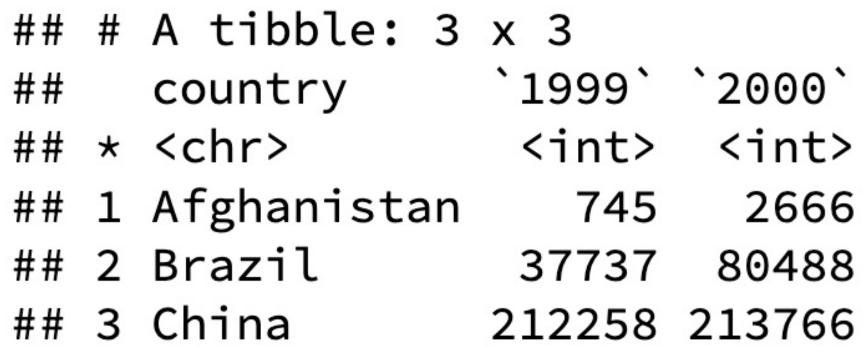
Ti



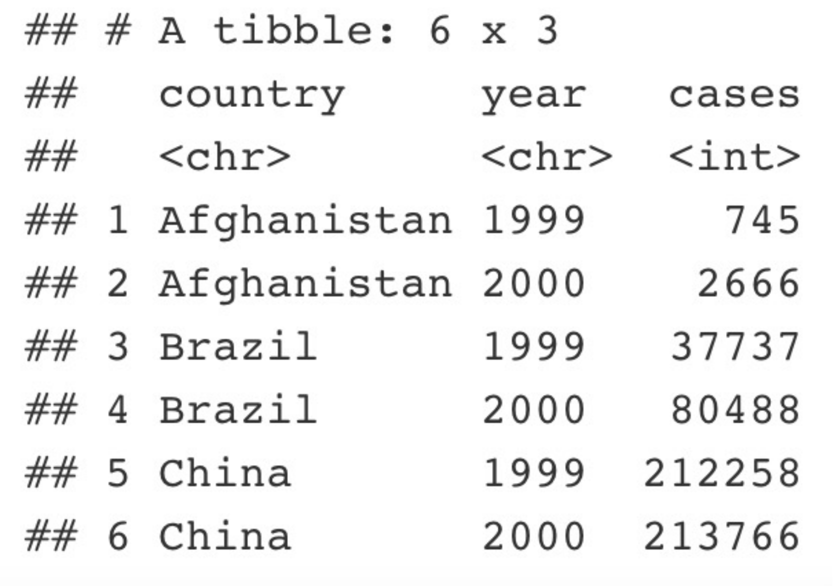
```
##   Republican Independent Democrat   the_date
## 1         16          47       85 2009-01-21
## 2         18          48       86 2009-01-26
## 3         17          45       84 2009-02-02
## 4         18          46       81 2009-02-09
## 5         17          46       82 2009-02-16
## 6         18          44       82 2009-02-23
```



```
## # A tibble: 4 x 3
##   the_date   party       approval
##   <date>     <chr>          <int>
## 1 2009-01-21 Republican        16
## 2 2009-01-21 Independent       47
## 3 2009-01-21 Democrat          85
## 4 2009-01-26 Republican        18
```



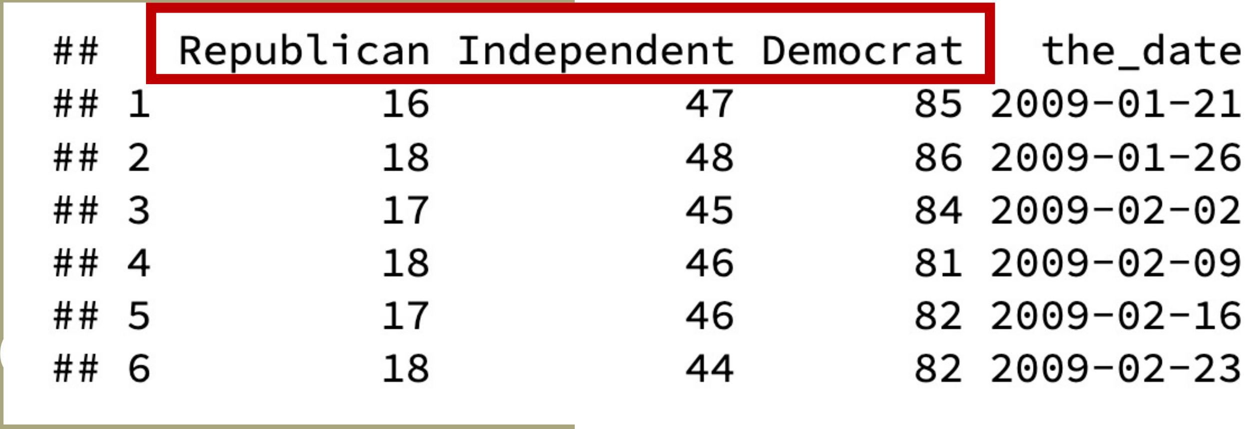
```
## # A tibble: 3 x 3
##   country     `1999` `2000`
## * <chr>        <int>  <int>
## 1 Afghanistan    745   2666
## 2 Brazil       37737  80488
## 3 China       212258 213766
```



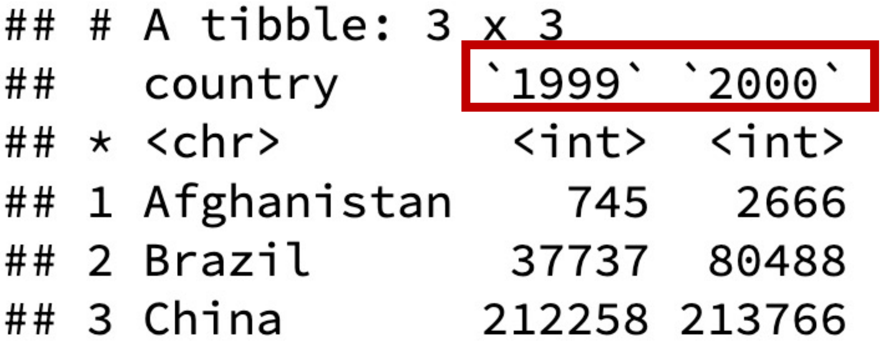
```
## # A tibble: 6 x 3
##   country     year   cases
##   <chr>       <chr>  <int>
## 1 Afghanistan 1999     745
## 2 Afghanistan 2000    2666
## 3 Brazil      1999   37737
## 4 Brazil      2000   80488
## 5 China       1999  212258
## 6 China       2000  213766
```

- Sometimes column headers contain data

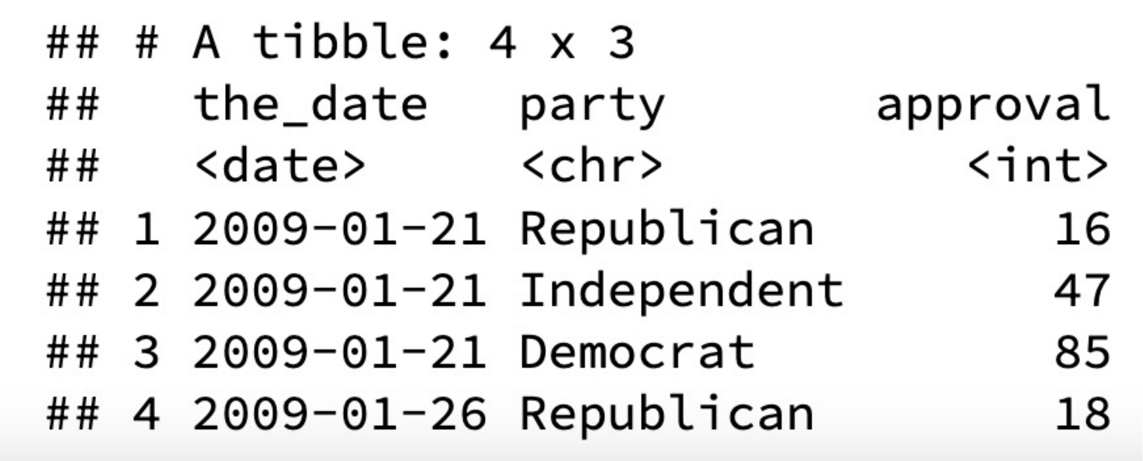
Ti



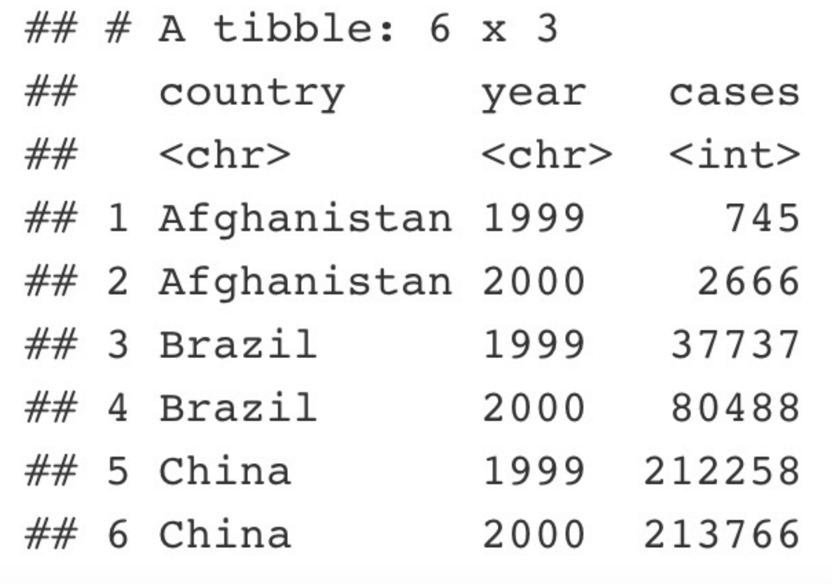
```
##   Republican Independent Democrat   the_date
## 1         16          47       85 2009-01-21
## 2         18          48       86 2009-01-26
## 3         17          45       84 2009-02-02
## 4         18          46       81 2009-02-09
## 5         17          46       82 2009-02-16
## 6         18          44       82 2009-02-23
```



```
## # A tibble: 3 x 3
##   country     `1999` `2000`
## * <chr>        <int>  <int>
## 1 Afghanistan    745   2666
## 2 Brazil       37737  80488
## 3 China       212258 213766
```



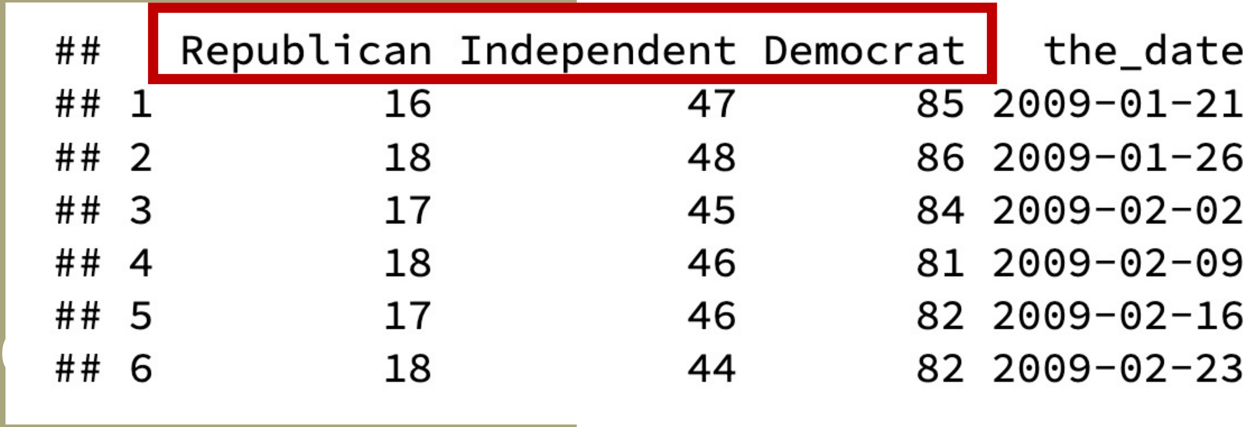
```
## # A tibble: 4 x 3
##   the_date   party       approval
##   <date>     <chr>          <int>
## 1 2009-01-21 Republican        16
## 2 2009-01-21 Independent       47
## 3 2009-01-21 Democrat          85
## 4 2009-01-26 Republican        18
```



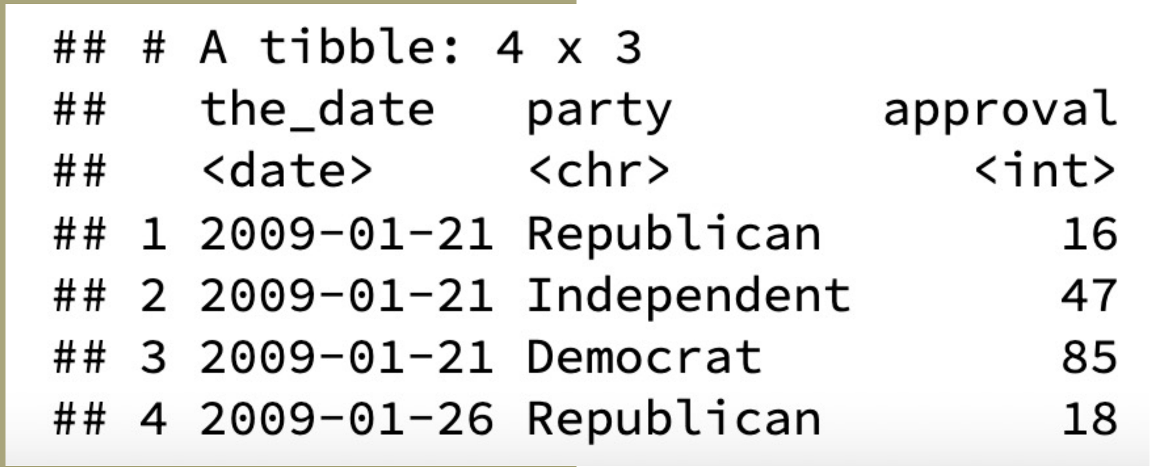
```
## # A tibble: 6 x 3
##   country     year   cases
##   <chr>       <chr>  <int>
## 1 Afghanistan 1999     745
## 2 Afghanistan 2000    2666
## 3 Brazil      1999   37737
## 4 Brazil      2000   80488
## 5 China       1999  212258
## 6 China       2000  213766
```

- Sometimes column headers contain data
- To correct this, we need to `pivot` our table

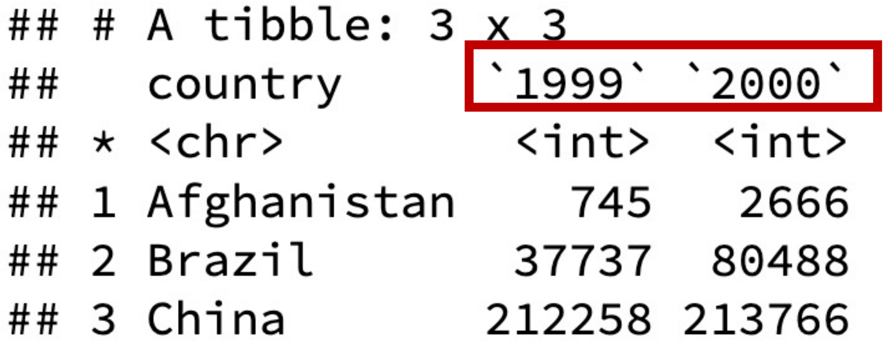
Ti



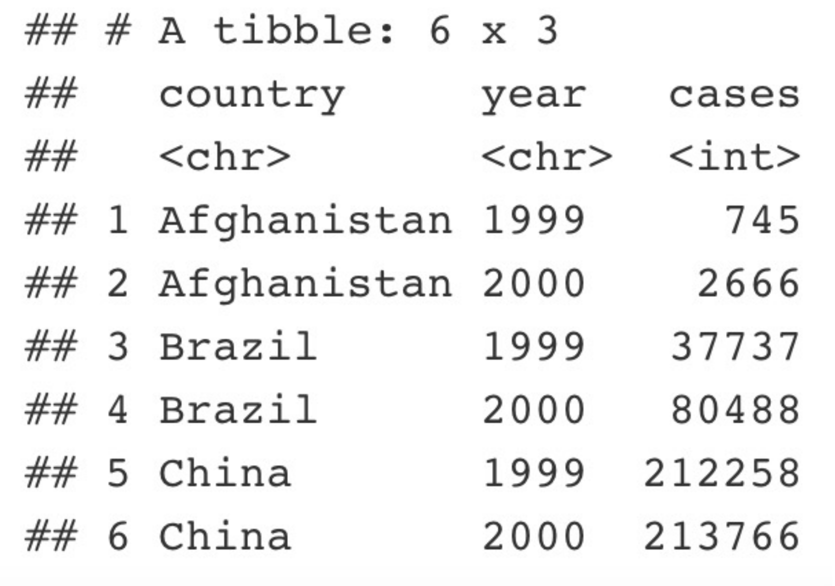
```
##   Republican Independent Democrat   the_date
## 1         16          47       85 2009-01-21
## 2         18          48       86 2009-01-26
## 3         17          45       84 2009-02-02
## 4         18          46       81 2009-02-09
## 5         17          46       82 2009-02-16
## 6         18          44       82 2009-02-23
```



```
## # A tibble: 4 x 3
##   the_date   party       approval
##   <date>     <chr>          <int>
## 1 2009-01-21 Republican        16
## 2 2009-01-21 Independent       47
## 3 2009-01-21 Democrat          85
## 4 2009-01-26 Republican        18
```



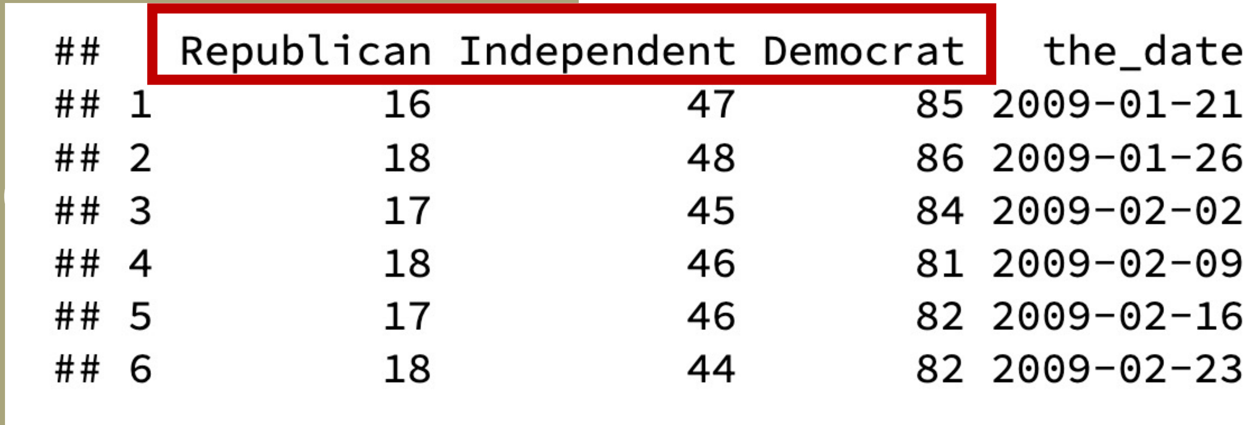
```
## # A tibble: 3 x 3
##   country     `1999` `2000`
## * <chr>        <int>  <int>
## 1 Afghanistan    745   2666
## 2 Brazil       37737  80488
## 3 China       212258 213766
```



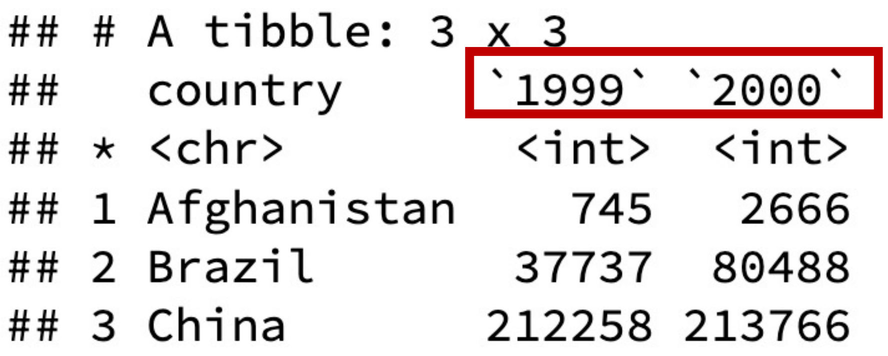
```
## # A tibble: 6 x 3
##   country     year   cases
##   <chr>       <chr>  <int>
## 1 Afghanistan 1999     745
## 2 Afghanistan 2000    2666
## 3 Brazil      1999   37737
## 4 Brazil      2000   80488
## 5 China       1999  212258
## 6 China       2000  213766
```

Ti

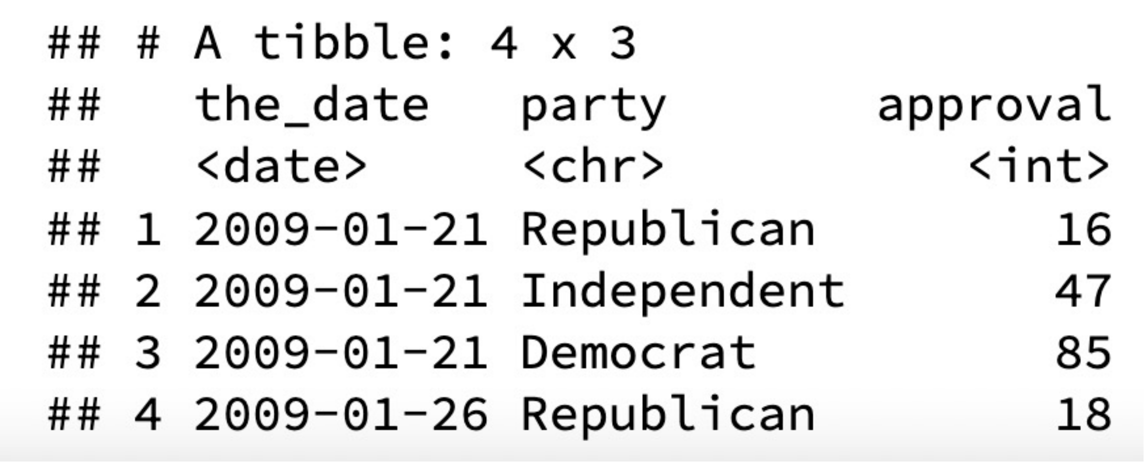
- Sometimes column headers contain data
- To correct this, we need to `pivot` our table
- When we move column headers to a variable, we `pivot_longer`



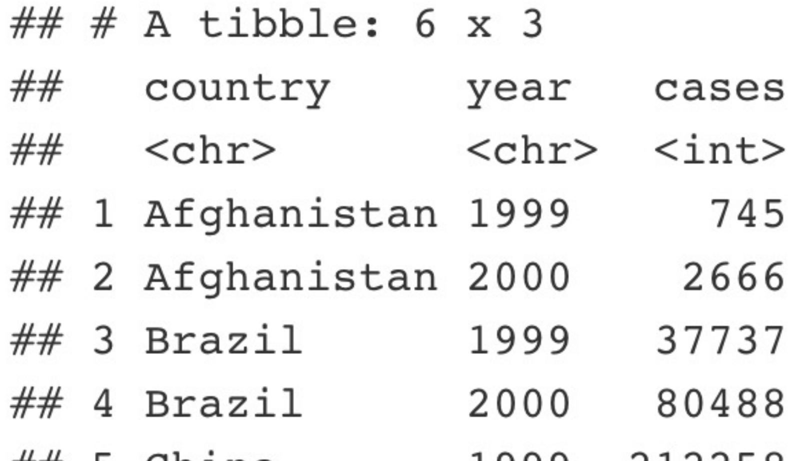
```
##   Republican Independent Democrat   the_date
## 1         16          47       85 2009-01-21
## 2         18          48       86 2009-01-26
## 3         17          45       84 2009-02-02
## 4         18          46       81 2009-02-09
## 5         17          46       82 2009-02-16
## 6         18          44       82 2009-02-23
```



```
## # A tibble: 3 x 3
##   country     `1999` `2000`
## * <chr>        <int>  <int>
## 1 Afghanistan    745   2666
## 2 Brazil       37737  80488
## 3 China       212258 213766
```



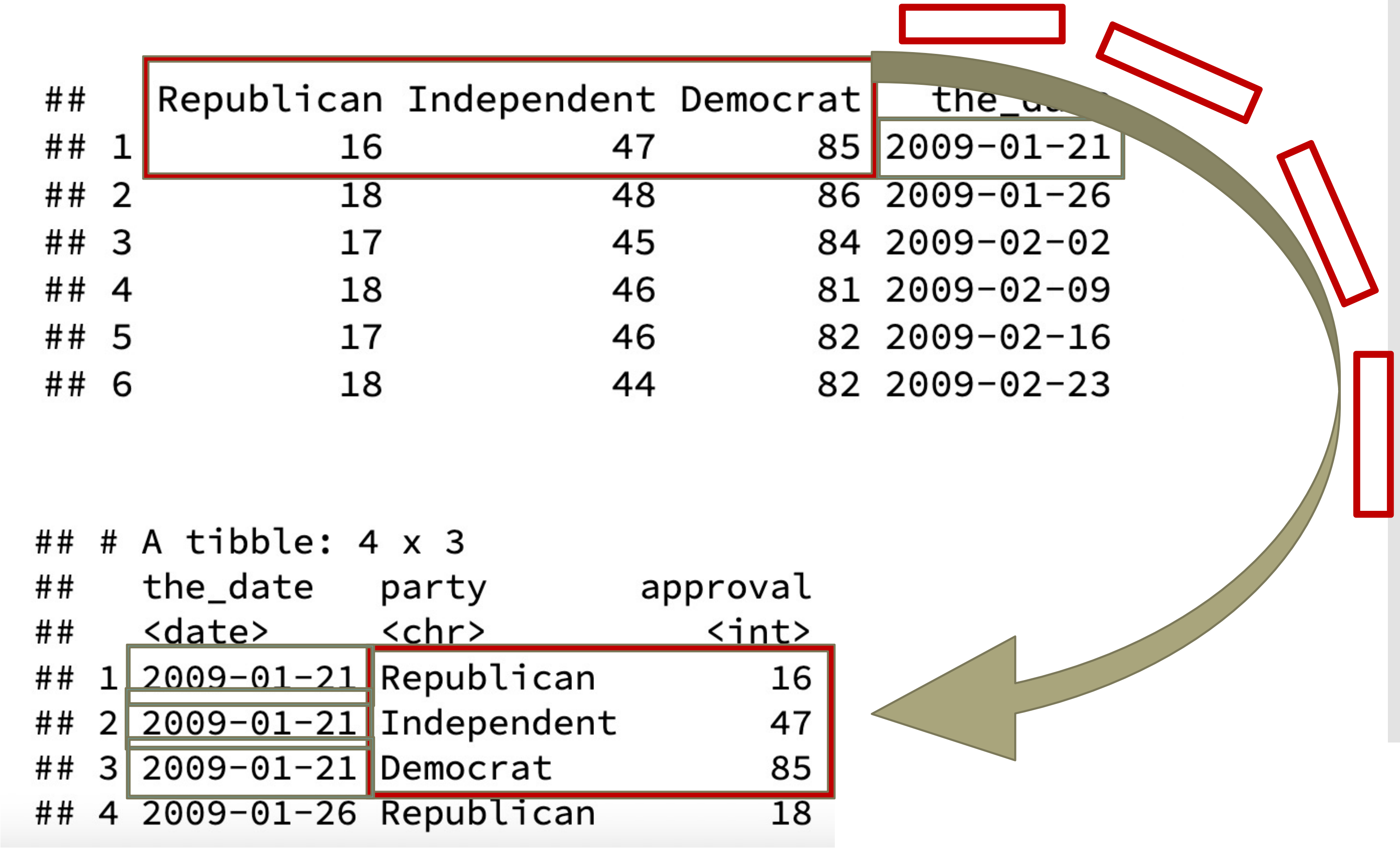
```
## # A tibble: 4 x 3
##   the_date   party       approval
##   <date>     <chr>          <int>
## 1 2009-01-21 Republican        16
## 2 2009-01-21 Independent       47
## 3 2009-01-21 Democrat          85
## 4 2009-01-26 Republican        18
```



```
## # A tibble: 6 x 3
##   country     year   cases
##   <chr>       <chr>  <int>
## 1 Afghanistan 1999     745
## 2 Afghanistan 2000    2666
## 3 Brazil      1999   37737
## 4 Brazil      2000   80488
## 5 China       1999  212258
```

Tidy Data

- `pivot_longer`
  - Each observation gets its own row
  - Number of rows increases (table gets longer)

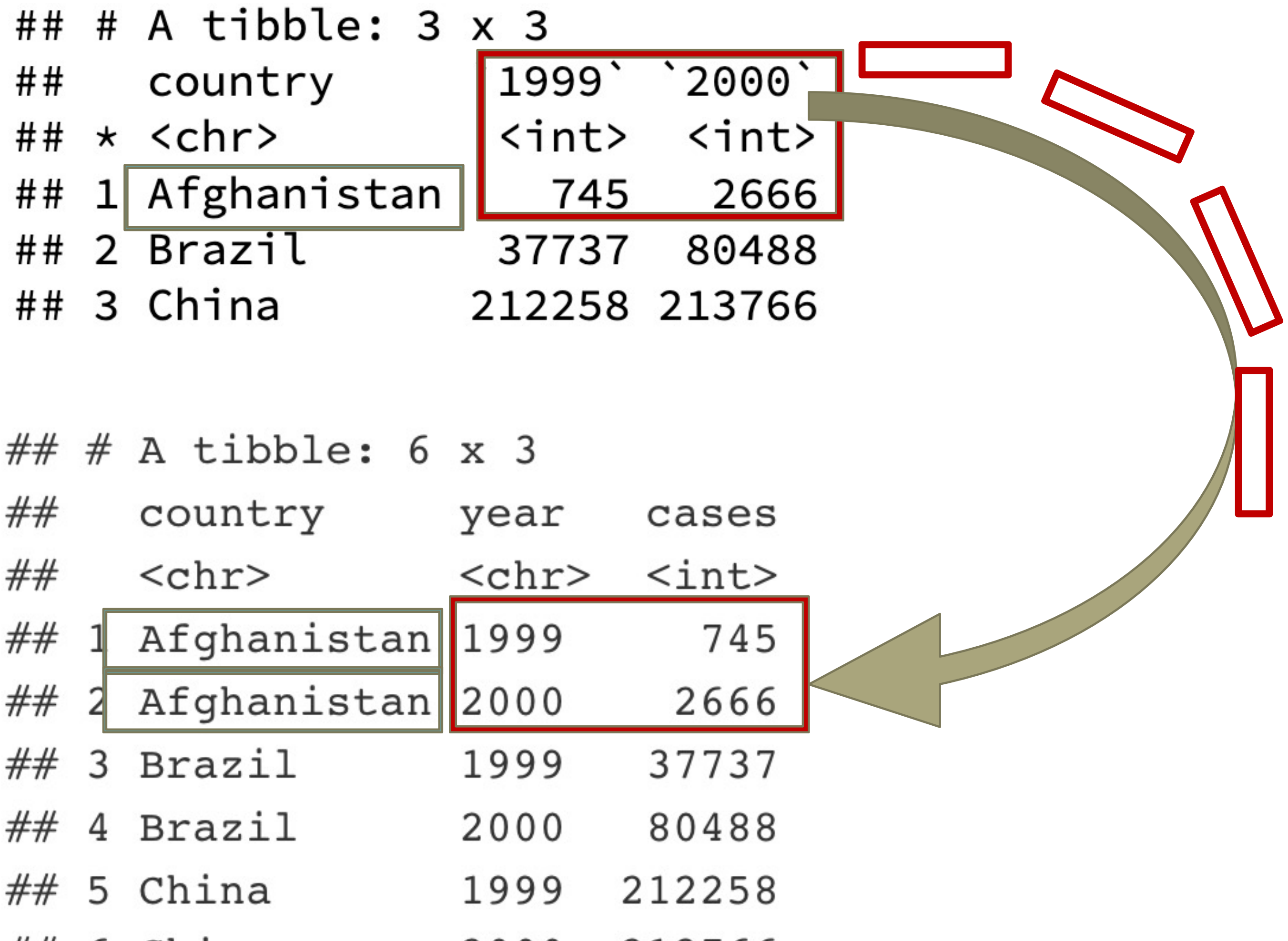


```
##    Republican Independent Democrat   the_da
## 1          16          47       85 2009-01-21
## 2          18          48       86 2009-01-26
## 3          17          45       84 2009-02-02
## 4          18          46       81 2009-02-09
## 5          17          46       82 2009-02-16
## 6          18          44       82 2009-02-23
```

```
## # A tibble: 4 x 3
##   the_date   party       approval
##   <date>     <chr>          <int>
## 1 2009-01-21 Republican        16
## 2 2009-01-21 Independent       47
## 3 2009-01-21 Democrat          85
## 4 2009-01-26 Republican        18
```

Tidy Data

- `pivot_longer`
  - Each observation gets its own row
  - Number of rows increases (table gets longer)



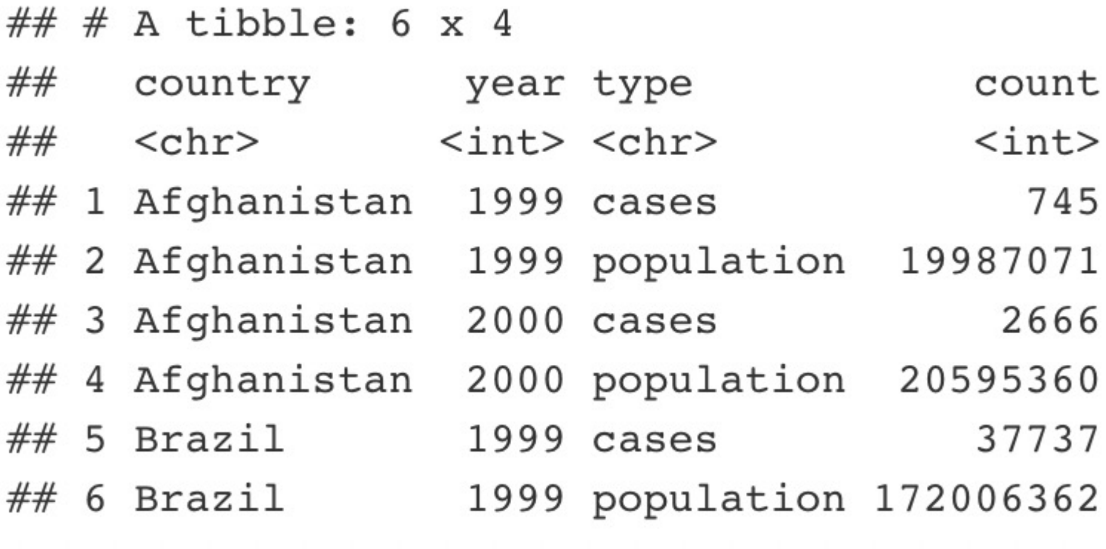
```
## # A tibble: 3 x 3
##   country      1999` `2000`
## * <chr>        <int>  <int>
## 1 Afghanistan    745   2666
## 2 Brazil       37737  80488
## 3 China       212258 213766
```

```
## # A tibble: 6 x 3
##   country     year   cases
##   <chr>       <chr>  <int>
## 1 Afghanistan 1999     745
## 2 Afghanistan 2000    2666
## 3 Brazil      1999   37737
## 4 Brazil      2000   80488
## 5 China       1999  212258
## 6 China       2000  213766
```

Tidy Data

- How do we make data Tidy?

Re-write this table in a tidy format

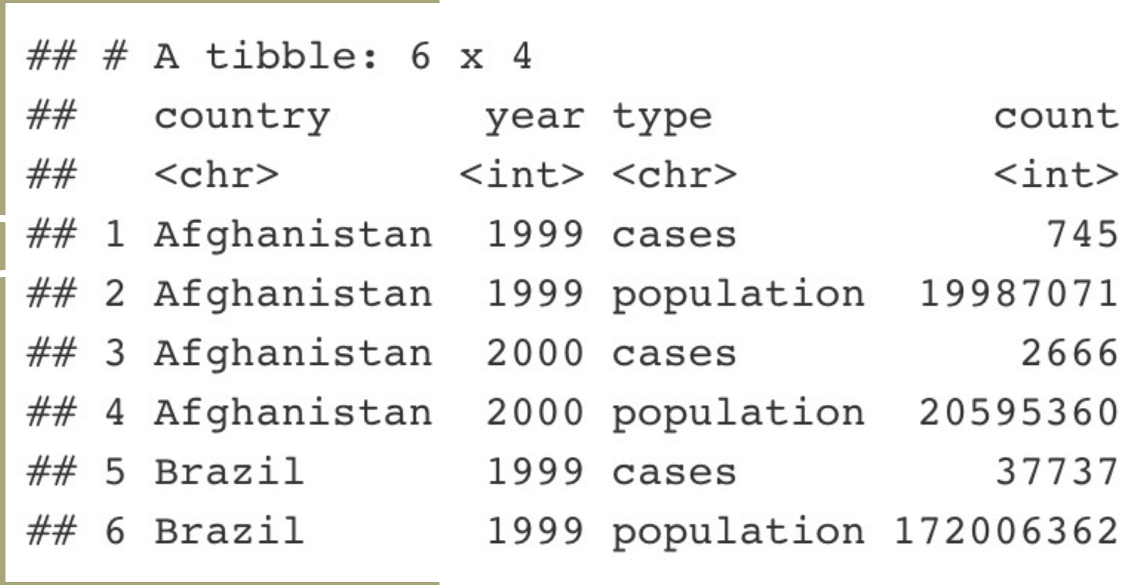


```
## # A tibble: 6 x 4
##   country      year type           count
##   <chr>       <int> <chr>          <int>
## 1 Afghanistan  1999 cases            745
## 2 Afghanistan  1999 population  19987071
## 3 Afghanistan  2000 cases           2666
## 4 Afghanistan  2000 population  20595360
## 5 Brazil       1999 cases          37737
## 6 Brazil       1999 population 172006362
```

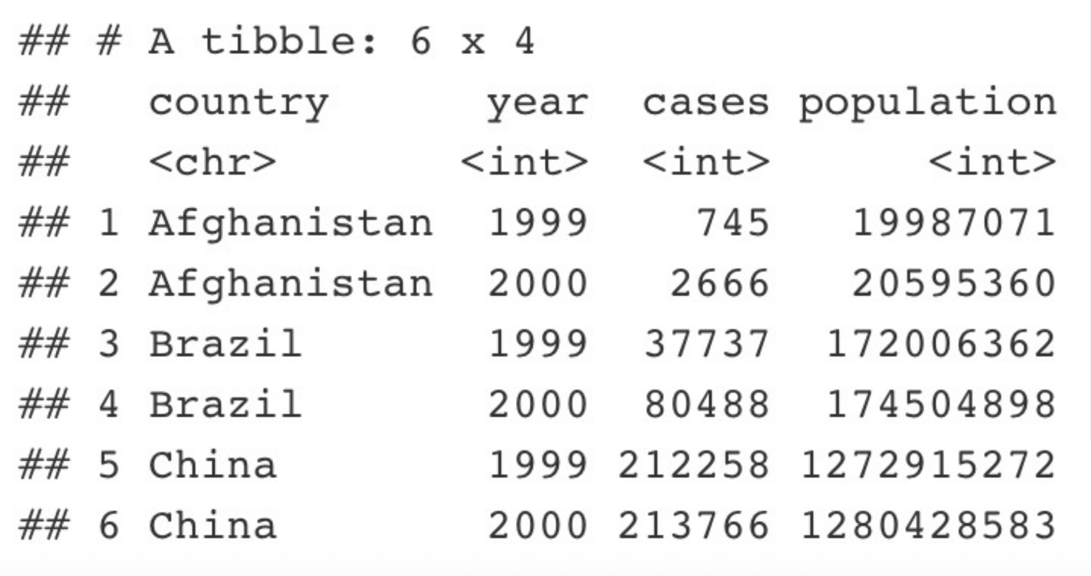
Tidy D

- How do we make data Tidy?

Re-write this table in a tidy format



```
## # A tibble: 6 x 4
##   country      year type           count
##   <chr>       <int> <chr>          <int>
## 1 Afghanistan  1999 cases            745
## 2 Afghanistan  1999 population  19987071
## 3 Afghanistan  2000 cases           2666
## 4 Afghanistan  2000 population  20595360
## 5 Brazil       1999 cases          37737
## 6 Brazil       1999 population 172006362
```

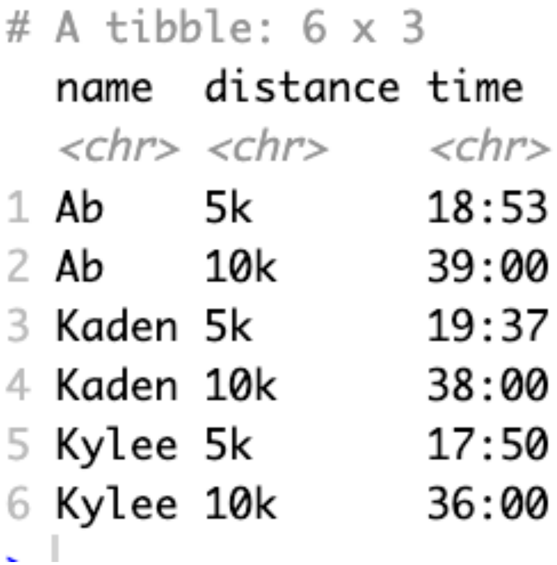


```
## # A tibble: 6 x 4
##   country      year  cases population
##   <chr>       <int>  <int>      <int>
## 1 Afghanistan  1999    745   19987071
## 2 Afghanistan  2000   2666   20595360
## 3 Brazil       1999  37737  172006362
## 4 Brazil       2000  80488  174504898
## 5 China        1999 212258 1272915272
## 6 China        2000 213766 1280428583
```

Tidy Data

- How do we make data Tidy?

Re-write this table in a tidy format

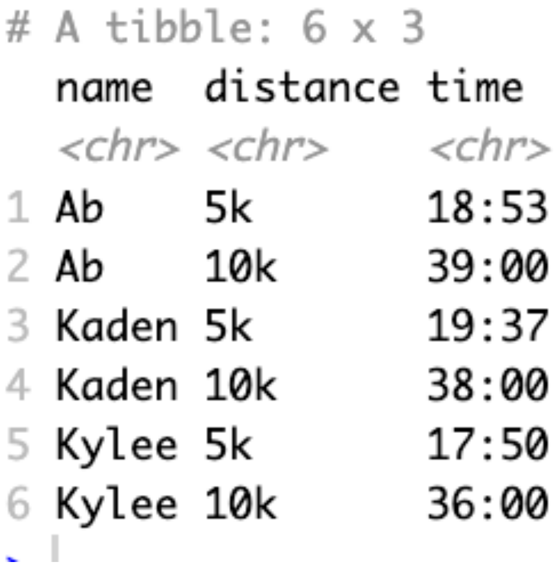


```
# A tibble: 6 x 3
  name  distance time
  <chr> <chr>    <chr>
1 Ab    5k       18:53
2 Ab    10k      39:00
3 Kaden 5k       19:37
4 Kaden 10k      38:00
5 Kylee 5k       17:50
6 Kylee 10k      36:00
```

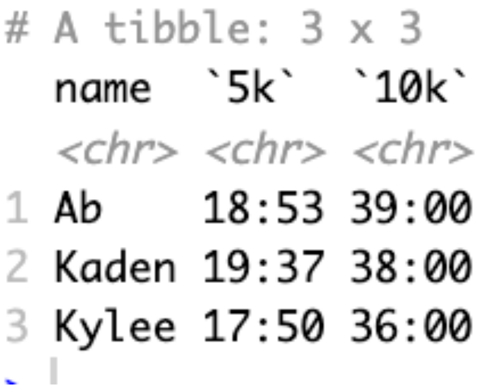
Tidy Data

- How do we make data Tidy?

Re-write this table in a tidy format



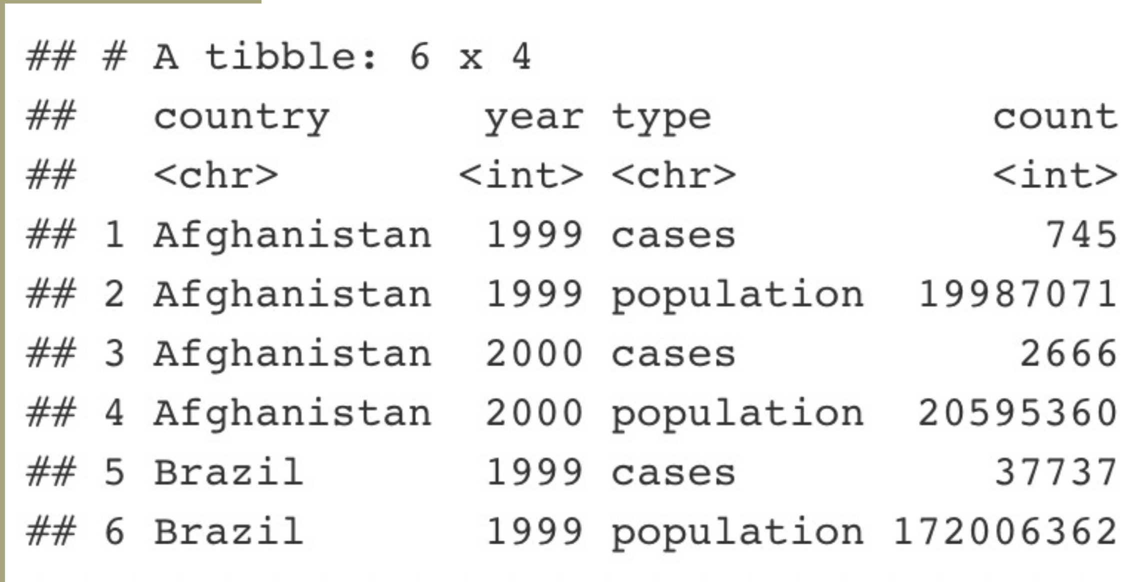
```
# A tibble: 6 x 3
  name  distance time
  <chr> <chr>    <chr>
1 Ab    5k       18:53
2 Ab    10k      39:00
3 Kaden 5k       19:37
4 Kaden 10k      38:00
5 Kylee 5k       17:50
6 Kylee 10k      36:00
```



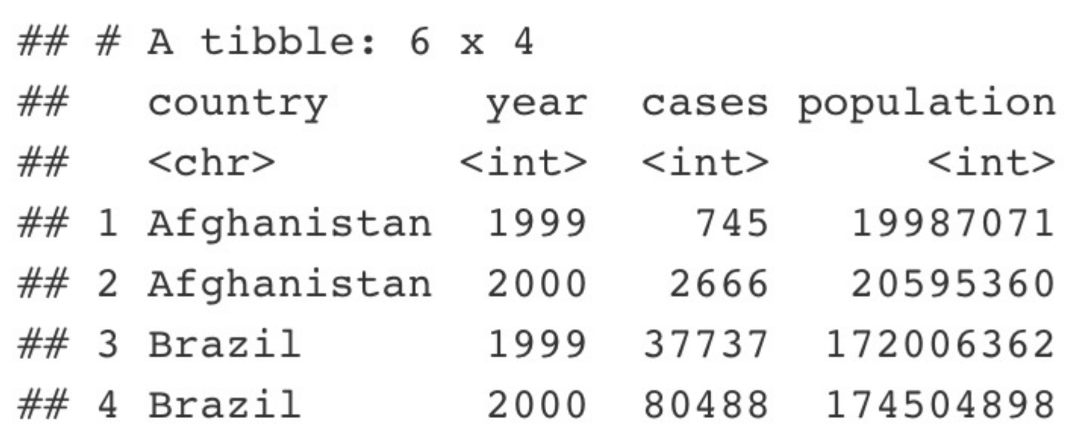
```
# A tibble: 3 x 3
  name  `5k`  `10k`
  <chr> <chr> <chr>
1 Ab    18:53 39:00
2 Kaden 19:37 38:00
3 Kylee 17:50 36:00
```

Tidy Data

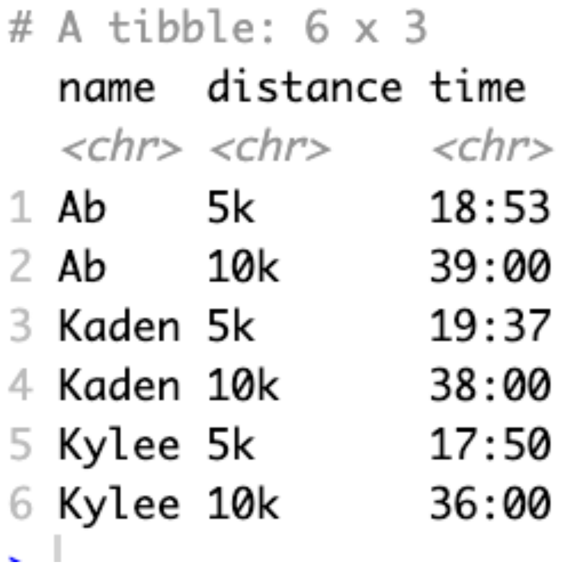
What is similar in these two cases?



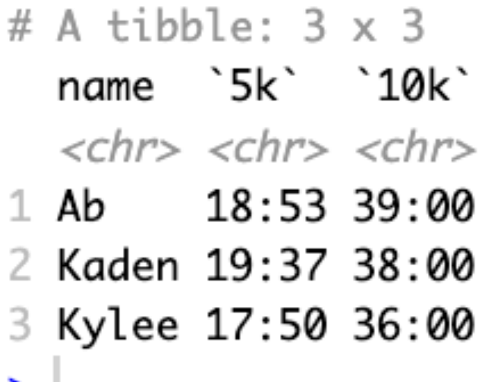
```
## # A tibble: 6 x 4
##   country      year type            count
##   <chr>       <int> <chr>           <int>
## 1 Afghanistan  1999 cases             745
## 2 Afghanistan  1999 population   19987071
## 3 Afghanistan  2000 cases            2666
## 4 Afghanistan  2000 population   20595360
## 5 Brazil       1999 cases           37737
## 6 Brazil       1999 population 172006362
```



```
## # A tibble: 6 x 4
##   country      year  cases population
##   <chr>       <int>  <int>      <int>
## 1 Afghanistan  1999    745   19987071
## 2 Afghanistan  2000   2666   20595360
## 3 Brazil       1999  37737  172006362
## 4 Brazil       2000  80488  174504898
```



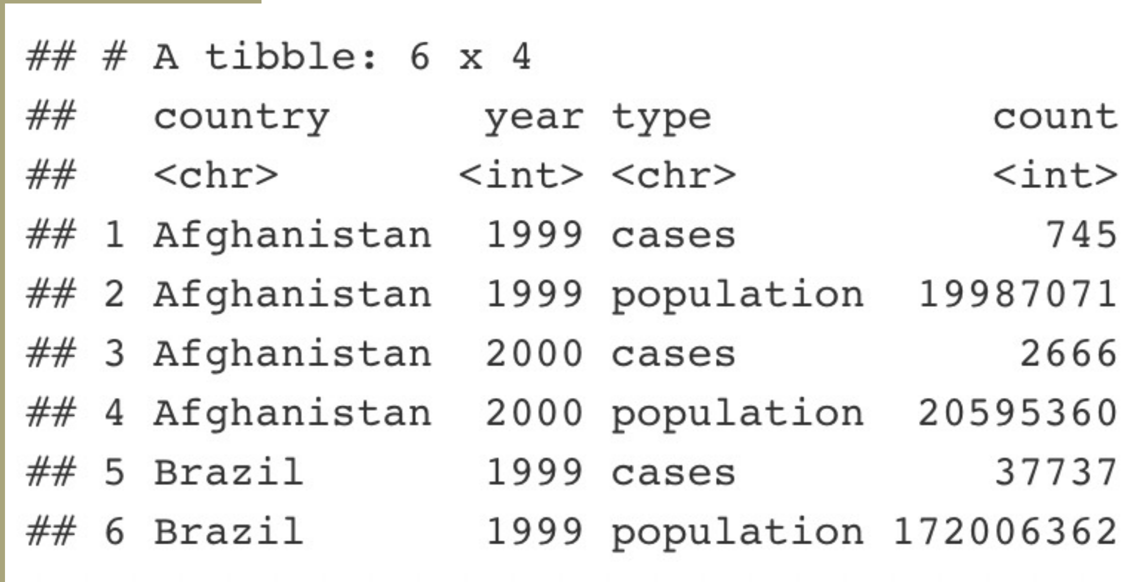
```
# A tibble: 6 x 3
  name  distance time
  <chr> <chr>    <chr>
1 Ab    5k       18:53
2 Ab    10k      39:00
3 Kaden 5k       19:37
4 Kaden 10k      38:00
5 Kylee 5k       17:50
6 Kylee 10k      36:00
```



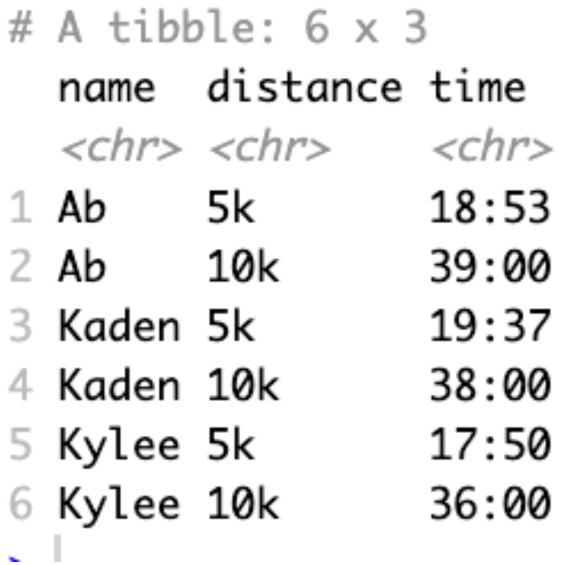
```
# A tibble: 3 x 3
  name  `5k`  `10k`
  <chr> <chr> <chr>
1 Ab    18:53 39:00
2 Kaden 19:37 38:00
3 Kylee 17:50 36:00
```

Tidy Data

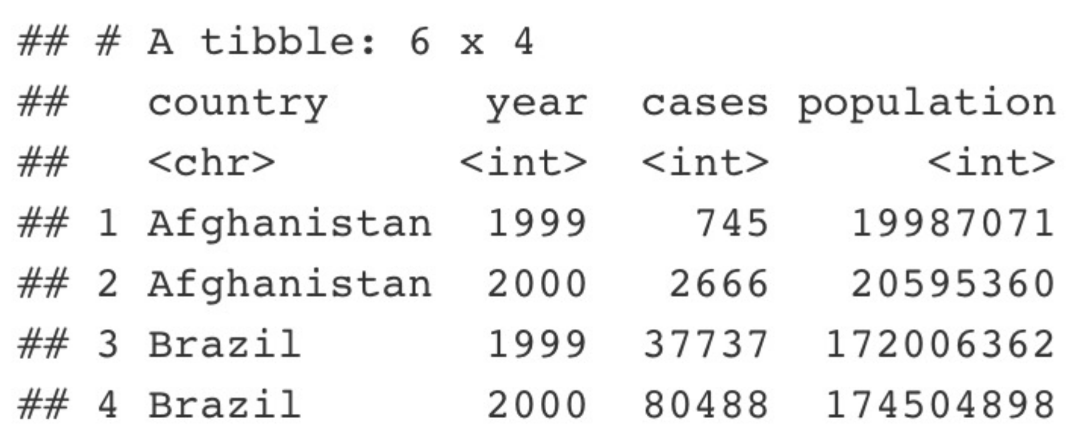
- Sometimes observations are split between rows



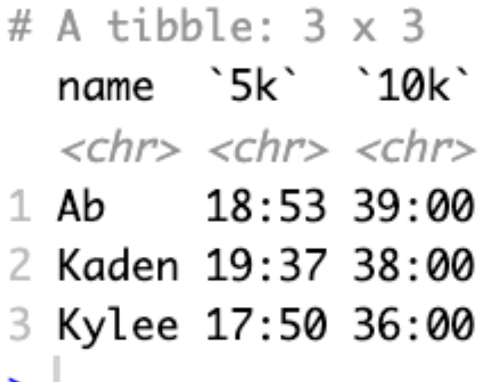
```
## # A tibble: 6 x 4
##   country      year type            count
##   <chr>       <int> <chr>           <int>
## 1 Afghanistan  1999 cases             745
## 2 Afghanistan  1999 population   19987071
## 3 Afghanistan  2000 cases            2666
## 4 Afghanistan  2000 population   20595360
## 5 Brazil       1999 cases           37737
## 6 Brazil       1999 population  172006362
```



```
# A tibble: 6 x 3
  name  distance time
  <chr> <chr>    <chr>
1 Ab    5k       18:53
2 Ab    10k      39:00
3 Kaden 5k       19:37
4 Kaden 10k      38:00
5 Kylee 5k       17:50
6 Kylee 10k      36:00
```



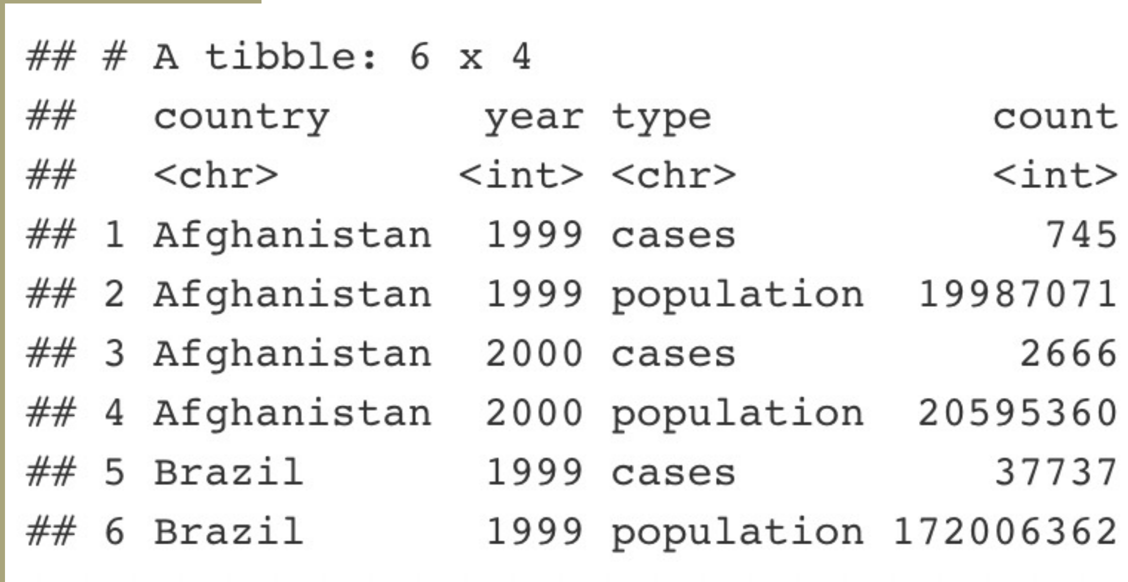
```
## # A tibble: 6 x 4
##   country      year  cases population
##   <chr>       <int>  <int>      <int>
## 1 Afghanistan  1999    745   19987071
## 2 Afghanistan  2000   2666   20595360
## 3 Brazil       1999  37737  172006362
## 4 Brazil       2000  80488  174504898
```



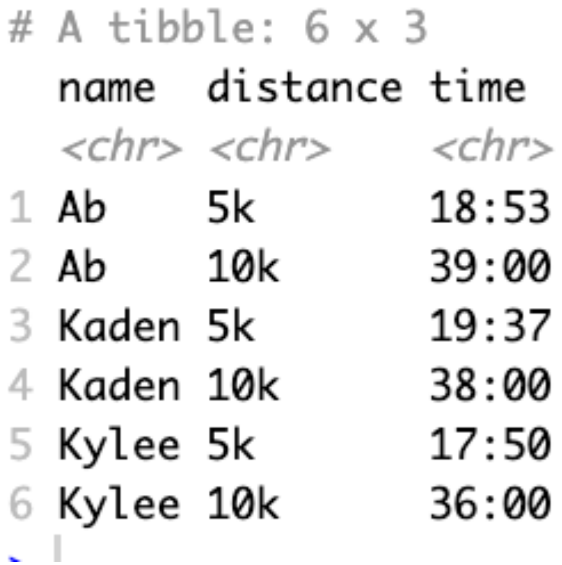
```
# A tibble: 3 x 3
  name  `5k`  `10k`
  <chr> <chr> <chr>
1 Ab    18:53 39:00
2 Kaden 19:37 38:00
3 Kylee 17:50 36:00
```

Tidy Data

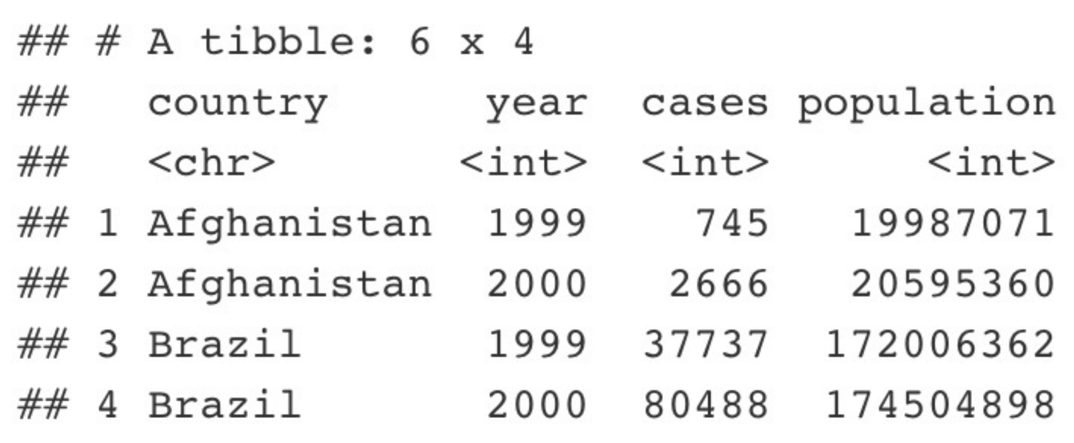
- Sometimes observations are split between rows
- To correct this we need to `pivot` the table



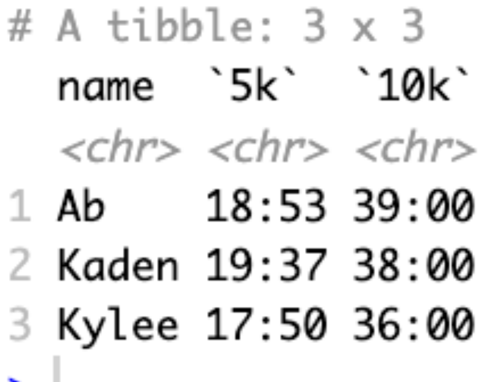
```
## # A tibble: 6 x 4
##   country      year type            count
##   <chr>       <int> <chr>           <int>
## 1 Afghanistan  1999 cases             745
## 2 Afghanistan  1999 population   19987071
## 3 Afghanistan  2000 cases            2666
## 4 Afghanistan  2000 population   20595360
## 5 Brazil       1999 cases           37737
## 6 Brazil       1999 population 172006362
```



```
# A tibble: 6 x 3
  name  distance time 
  <chr> <chr>    <chr>
1 Ab    5k       18:53
2 Ab    10k      39:00
3 Kaden 5k       19:37
4 Kaden 10k      38:00
5 Kylee 5k       17:50
6 Kylee 10k      36:00
```



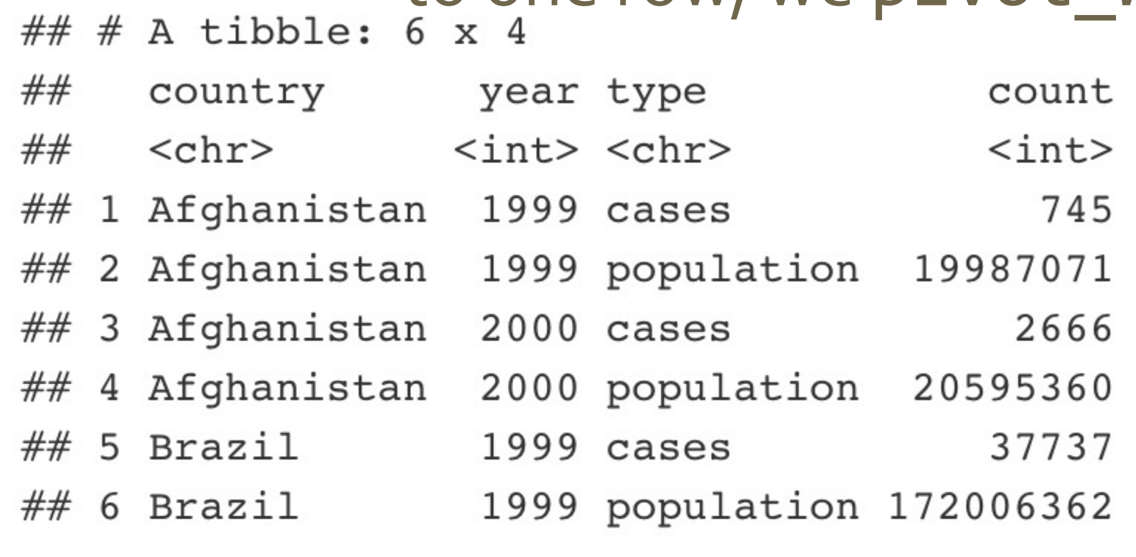
```
## # A tibble: 6 x 4
##   country      year  cases population
##   <chr>       <int>  <int>      <int>
## 1 Afghanistan  1999    745   19987071
## 2 Afghanistan  2000   2666   20595360
## 3 Brazil       1999  37737  172006362
## 4 Brazil       2000  80488  174504898
```



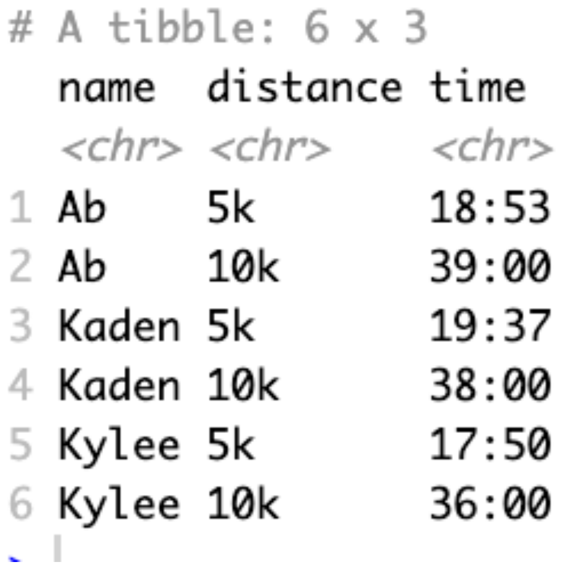
```
# A tibble: 3 x 3
  name  `5k`  `10k`
  <chr> <chr> <chr>
1 Ab    18:53 39:00
2 Kaden 19:37 38:00
3 Kylee 17:50 36:00
```

Tidy Data

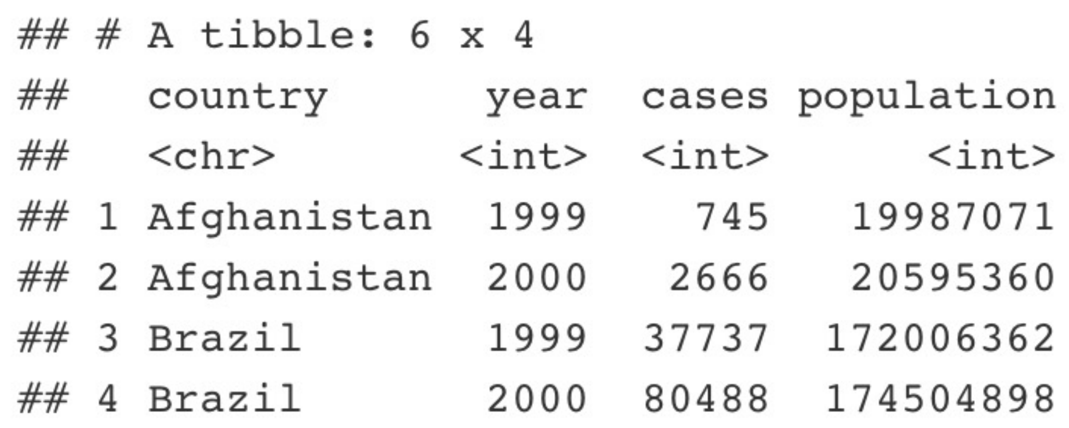
- Sometimes observations are split between rows
- To correct this we need to `pivot` the table
- When we move observations from multiple rows to one row, we `pivot_wider`



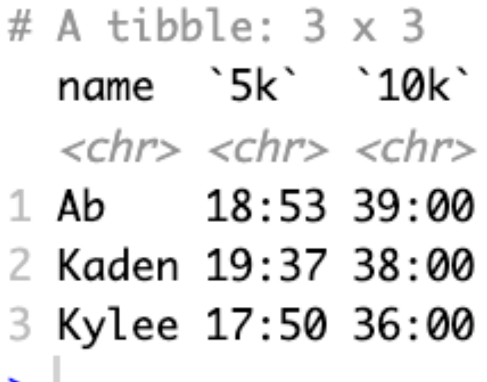
```
## # A tibble: 6 x 4
##   country      year type            count
##   <chr>       <int> <chr>           <int>
## 1 Afghanistan  1999 cases             745
## 2 Afghanistan  1999 population   19987071
## 3 Afghanistan  2000 cases            2666
## 4 Afghanistan  2000 population   20595360
## 5 Brazil       1999 cases           37737
## 6 Brazil       1999 population 172006362
```



```
# A tibble: 6 x 3
  name  distance time 
  <chr> <chr>    <chr>
1 Ab    5k       18:53
2 Ab    10k      39:00
3 Kaden 5k       19:37
4 Kaden 10k      38:00
5 Kylee 5k       17:50
6 Kylee 10k      36:00
```



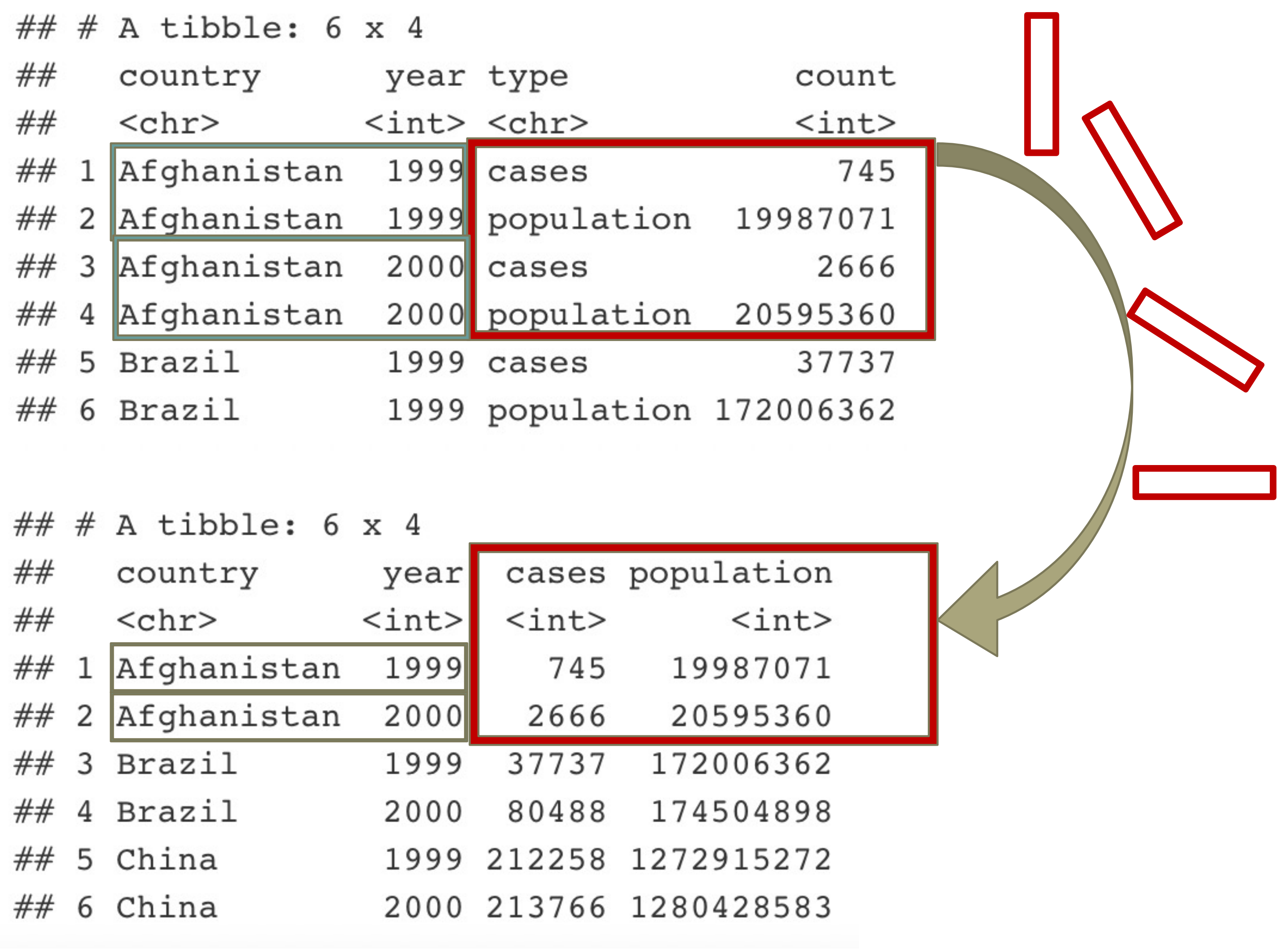
```
## # A tibble: 6 x 4
##   country      year  cases population
##   <chr>       <int>  <int>      <int>
## 1 Afghanistan  1999    745   19987071
## 2 Afghanistan  2000   2666   20595360
## 3 Brazil       1999  37737  172006362
## 4 Brazil       2000  80488  174504898
```



```
# A tibble: 3 x 3
  name  `5k`  `10k`
  <chr> <chr> <chr>
1 Ab    18:53 39:00
2 Kaden 19:37 38:00
3 Kylee 17:50 36:00
```

Tidy Data

- `pivot_wider`
  - Each observation gets its own row
  - Number of rows decreases

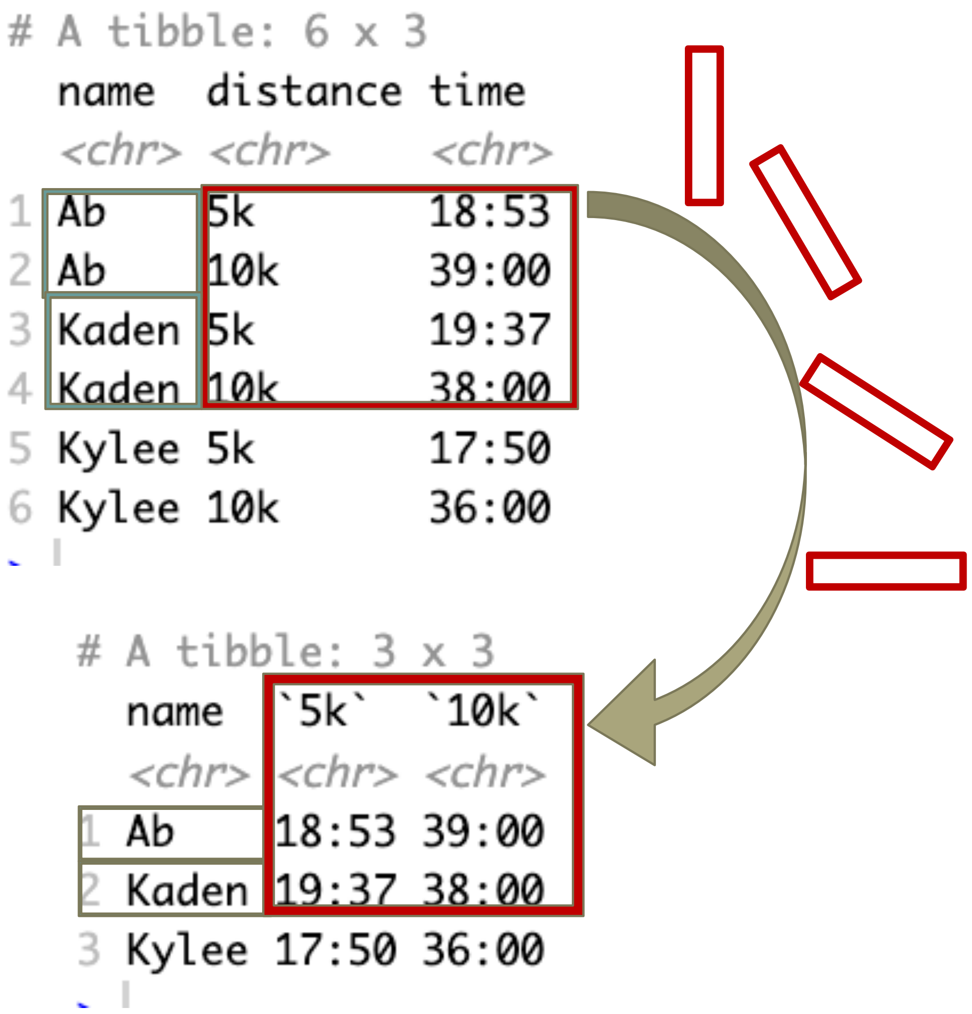


```
## # A tibble: 6 x 4
##   country      year type            count
##   <chr>       <int> <chr>           <int>
## 1 Afghanistan  1999 cases             745
## 2 Afghanistan  1999 population   19987071
## 3 Afghanistan  2000 cases            2666
## 4 Afghanistan  2000 population   20595360
## 5 Brazil       1999 cases           37737
## 6 Brazil       1999 population  172006362
```

```
## # A tibble: 6 x 4
##   country      year  cases population
##   <chr>       <int>  <int>      <int>
## 1 Afghanistan  1999    745   19987071
## 2 Afghanistan  2000   2666   20595360
## 3 Brazil       1999  37737  172006362
## 4 Brazil       2000  80488  174504898
## 5 China        1999 212258 1272915272
## 6 China        2000 213766 1280428583
```

Tidy Data

- `pivot_wider`
  - Each observation gets its own row
  - Number of rows decreases

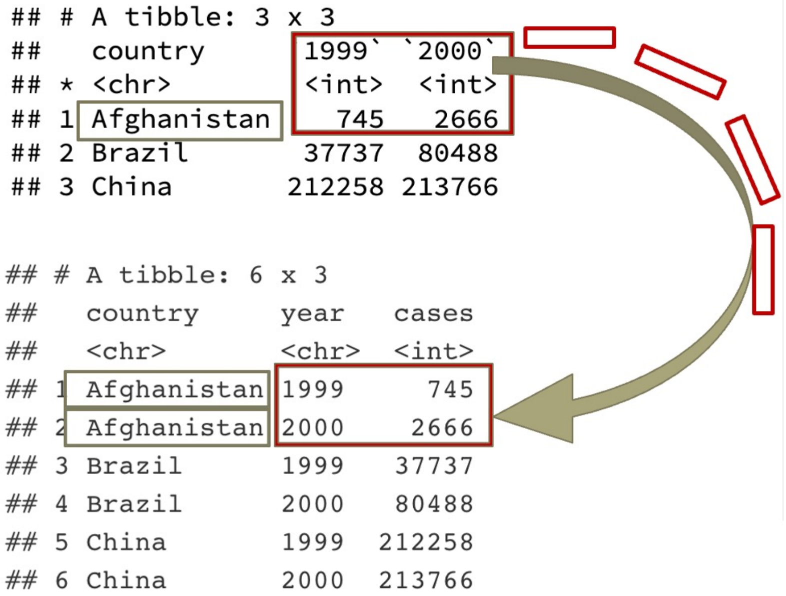


```
# A tibble: 6 x 3
  name  distance time
  <chr> <chr>    <chr>
1 Ab    5k       18:53
2 Ab    10k      39:00
3 Kaden 5k       19:37
4 Kaden 10k      38:00
5 Kylee 5k       17:50
6 Kylee 10k      36:00
```

```
# A tibble: 3 x 3
  name  `5k`  `10k`
  <chr> <chr> <chr>
1 Ab    18:53 39:00
2 Kaden 19:37 38:00
3 Kylee 17:50 36:00
```

- `tidyr`
  - R package that helps make data tidy
  - We will primarily use two functions:
    - `pivot_longer()`
    - `pivot_wider()`

- tidyr
  - pivot_longer()
    - wide → narrow

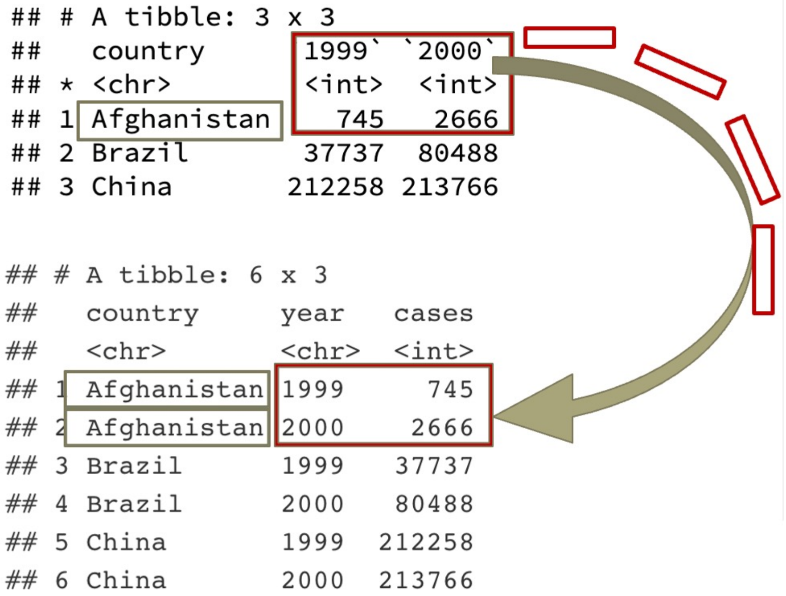


```
## # A tibble: 3 x 3
##   country     `1999` `2000`
## * <chr>        <int>  <int>
## 1 Afghanistan    745   2666
## 2 Brazil       37737  80488
## 3 China       212258 213766
```

```
## # A tibble: 6 x 3
##   country     year   cases
##   <chr>       <chr>  <int>
## 1 Afghanistan 1999     745
## 2 Afghanistan 2000    2666
## 3 Brazil      1999   37737
## 4 Brazil      2000   80488
## 5 China       1999  212258
## 6 China       2000  213766
```

```
table4a %>%
   pivot_longer(-country,
                names_to = "year",
                values_to = "cases")
```

- tidyr
  - pivot_longer()
    - wide → narrow



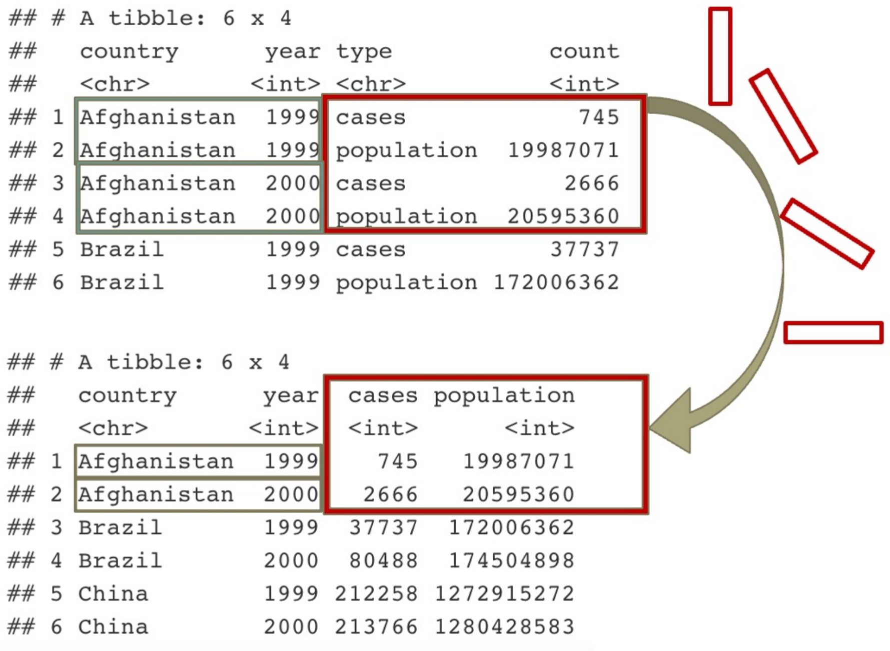
```
## # A tibble: 3 x 3
##   country     `1999` `2000`
## * <chr>        <int>  <int>
## 1 Afghanistan    745   2666
## 2 Brazil       37737  80488
## 3 China       212258 213766
```

```
## # A tibble: 6 x 3
##   country     year   cases
##   <chr>       <chr>  <int>
## 1 Afghanistan 1999     745
## 2 Afghanistan 2000    2666
## 3 Brazil      1999   37737
## 4 Brazil      2000   80488
## 5 China       1999  212258
## 6 China       2000  213766
```

```
table4a %>%
   pivot_longer(-country,
                         names_to = "year",
                         values_to = "cases")
```

- **-country**: pivot all columns except country
- **names_to = "year"**: make a new column called year (into which we'll put the pivoted column names)
- **values_to = "cases"**: make another new column called cases (into which we'll put the pivoted values)

- tidyr
  - pivot_wider()
    - narrow → wide

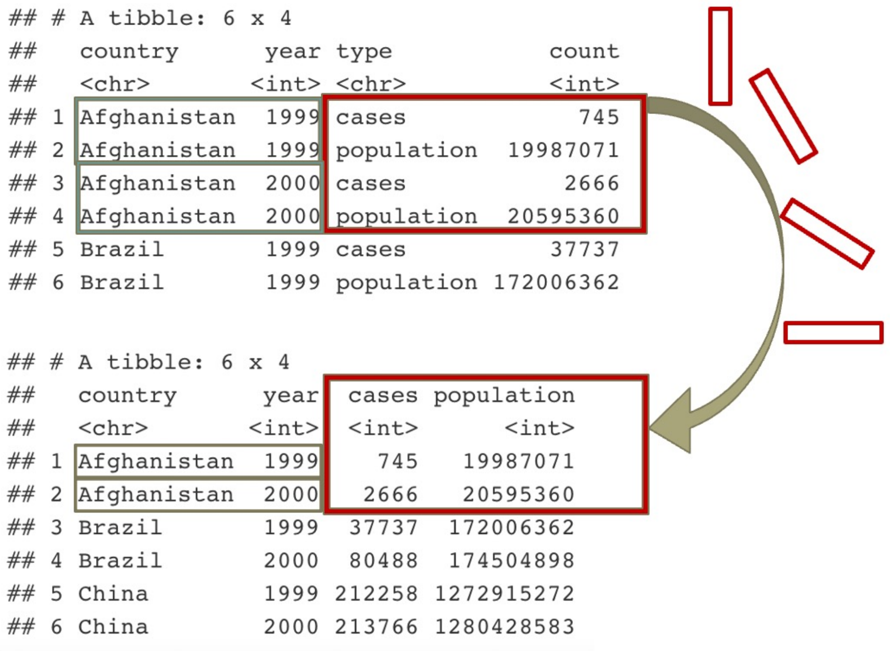


```
## # A tibble: 6 x 4
##   country      year type            count
##   <chr>       <int> <chr>           <int>
## 1 Afghanistan  1999 cases             745
## 2 Afghanistan  1999 population   19987071
## 3 Afghanistan  2000 cases            2666
## 4 Afghanistan  2000 population   20595360
## 5 Brazil       1999 cases           37737
## 6 Brazil       1999 population  172006362
```

```
## # A tibble: 6 x 4
##   country      year  cases population
##   <chr>       <int>  <int>      <int>
## 1 Afghanistan  1999    745   19987071
## 2 Afghanistan  2000   2666   20595360
## 3 Brazil       1999  37737  172006362
## 4 Brazil       2000  80488  174504898
## 5 China        1999 212258 1272915272
## 6 China        2000 213766 1280428583
```

```
table2 %>%
   pivot_wider(names_from = type,
               values_from = count)
```

- `tidyr`
  - `pivot_wider()`
    - narrow → wide



```
## # A tibble: 6 x 4
##   country      year type            count
##   <chr>       <int> <chr>           <int>
## 1 Afghanistan  1999 cases             745
## 2 Afghanistan  1999 population   19987071
## 3 Afghanistan  2000 cases            2666
## 4 Afghanistan  2000 population   20595360
## 5 Brazil       1999 cases           37737
## 6 Brazil       1999 population  172006362
```

```
## # A tibble: 6 x 4
##   country      year  cases population
##   <chr>       <int>  <int>      <int>
## 1 Afghanistan  1999    745   19987071
## 2 Afghanistan  2000   2666   20595360
## 3 Brazil       1999  37737  172006362
## 4 Brazil       2000  80488  174504898
## 5 China        1999 212258 1272915272
## 6 China        2000 213766 1280428583
```

```
table2 %>%
  pivot_wider(names_from = type,
              values_from = count)
```

- **`names_from = type`**: grab the values in the column called `type` (we'll `pivot` these values out to become the names of our new columns)
- **`values_from = count`**: grab the values in the column called count (we'll pivot these across their corresponding columns)

Tidy Data

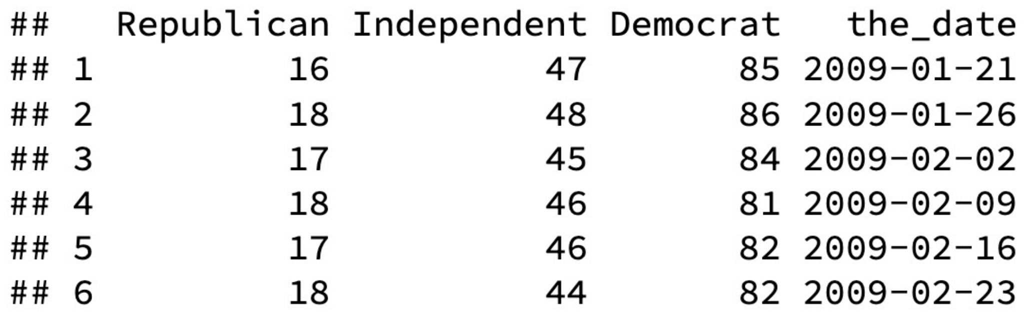
- We tend to use `pivot_longer()` most often

Fill in the missing code below to pivot presapproval from wide form to long form.

```
##    Republican Independent Democrat   the_date
## 1          16          47       85 2009-01-21
## 2          18          48       86 2009-01-26
## 3          17          45       84 2009-02-02
## 4          18          46       81 2009-02-09
## 5          17          46       82 2009-02-16
## 6          18          44       82 2009-02-23
```

```
presapproval_tidy <- presapproval %>%
   pivot_longer(-   ,
                names_to = "",
                values_to = "")
```

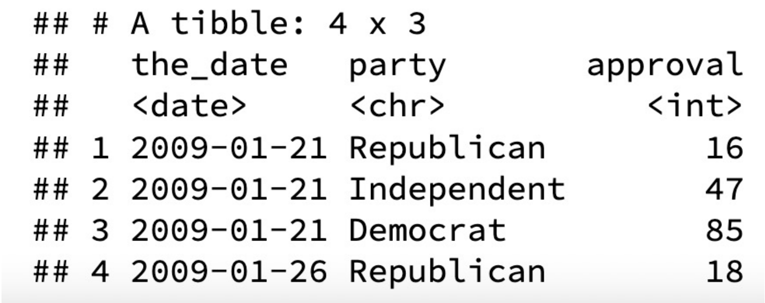
Tidy Data



```
##   Republican Independent Democrat   the_date
## 1         16          47       85 2009-01-21
## 2         18          48       86 2009-01-26
## 3         17          45       84 2009-02-02
## 4         18          46       81 2009-02-09
## 5         17          46       82 2009-02-16
## 6         18          44       82 2009-02-23
```

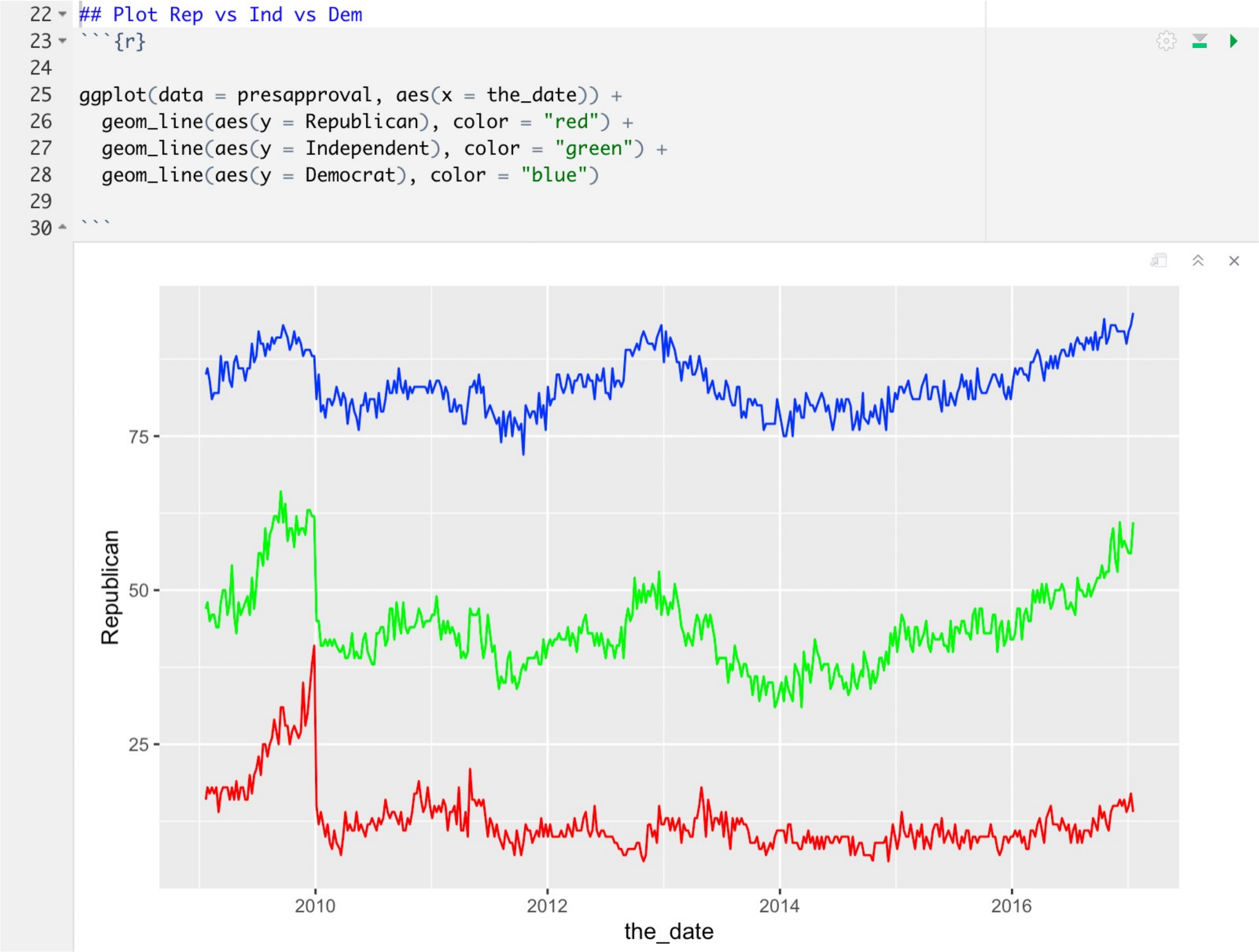
```
presapproval_tidy <- presapproval %>%
   pivot_longer(-the_date,
                names_to = "party",
                values_to = "approval")
```

- `-the_date`: pivot everything except `the_date`
- `names_to = "party"`: make a new column called party into which we'll put pivoted column names
- `values_to = "approval"`: make a new column called approval into which we'll put pivoted values



```
## # A tibble: 4 x 3
##   the_date   party       approval
##   <date>     <chr>          <int>
## 1 2009-01-21 Republican        16
## 2 2009-01-21 Independent       47
## 3 2009-01-21 Democrat          85
## 4 2009-01-26 Republican        18
```

Tidy Data

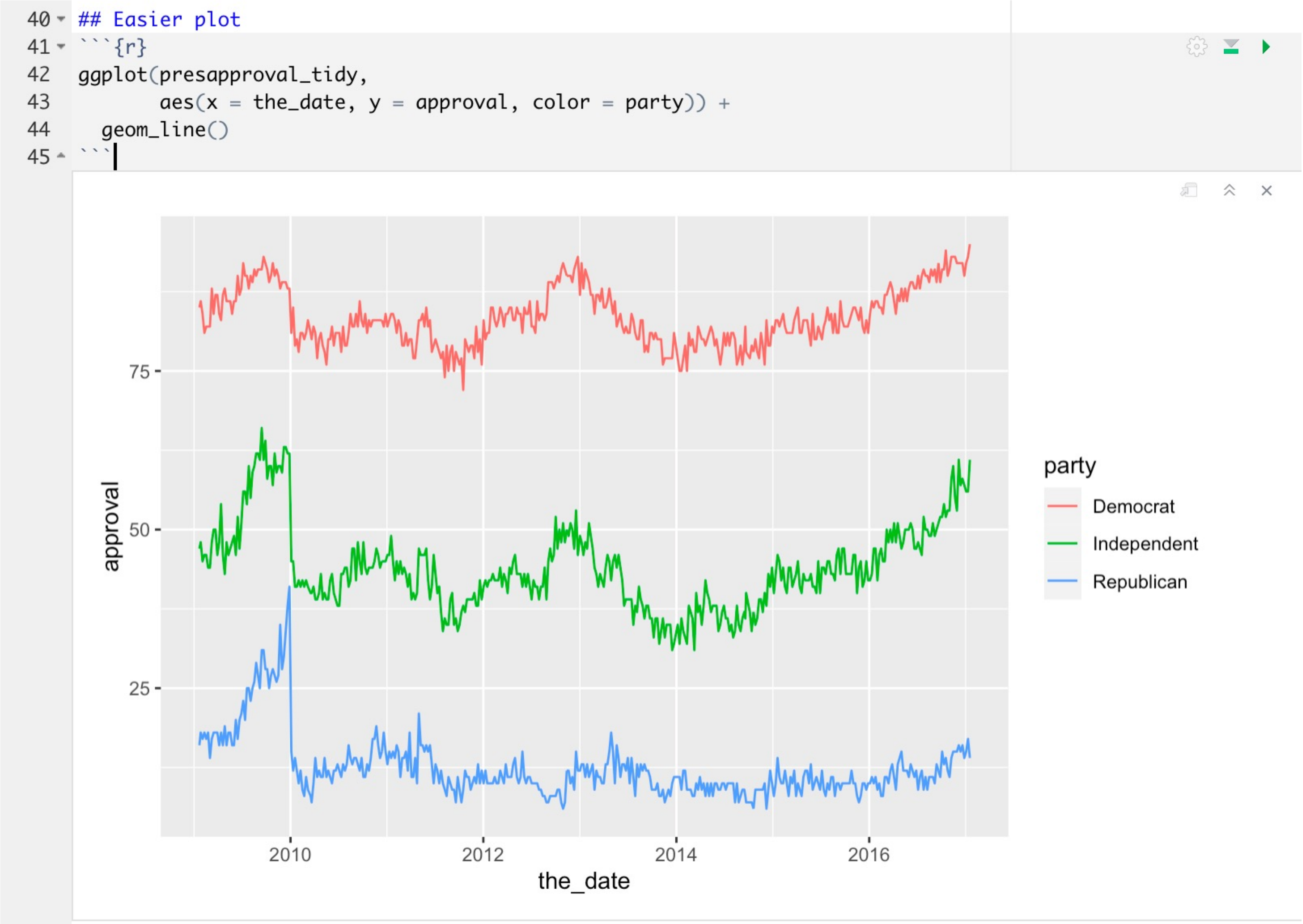


~~~
## Plot Rep vs Ind vs Dem
```{r}

ggplot(data = presapproval, aes(x = the_date)) +
  geom_line(aes(y = Republican), color = "red") +
  geom_line(aes(y = Independent), color = "green") +
  geom_line(aes(y = Democrat), color = "blue")

```
~~~

Tidy Data



~~~
## Easier plot
```{r}
ggplot(presapproval_tidy,
       aes(x = the_date, y = approval, color = party)) +
  geom_line()
```
~~~