EXPLORATORY DATA ANALYSIS: CASE STUDY

Joining datasets

Processed votes

Each row is one roll call/country pair

```
> votes_processed
# A tibble: 353,547 × 6
    rcid session  vote ccode  year             country
   <dbl>   <dbl> <dbl> <int> <dbl>               <chr>
1     46       2     1     2  1947       United States
2     46       2     1    20  1947              Canada
3     46       2     1    40  1947                Cuba
4     46       2     1    41  1947               Haiti
5     46       2     1    42  1947 Dominican Republic
6     46       2     1    70  1947              Mexico
7     46       2     1    90  1947           Guatemala
8     46       2     1    91  1947            Honduras
9     46       2     1    92  1947         El Salvador
10    46       2     1    93  1947           Nicaragua
# ... with 353,537 more rows
```

Descriptions dataset

```
> descriptions
# A tibble: 2,589 × 10
```

hr = Human rights

	rcid	session	date	unres	me	nu	di	hr	co	ec
	<dbl>	<dbl>	<dttm>	<chr>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
1	46	2	1947-09-04	R/2/299	0	0	0	0	0	0
2	47	2	1947-10-05	R/2/355	0	0	0	1	0	0
3	48	2	1947-10-06	R/2/461	0	0	0	0	0	0
4	49	2	1947-10-06	R/2/463	0	0	0	0	0	0
5	50	2	1947-10-06	R/2/465	0	0	0	0	0	0
6	51	2	1947-10-02	R/2/561	0	0	0	0	1	0
7	52	2	1947-11-06	R/2/650	0	0	0	0	1	0
8	53	2	1947-11-06	R/2/651	0	0	0	0	1	0
9	54	2	1947-11-06	R/2/651	0	0	0	0	1	0
10	55	2	1947-11-06	R/2/667	0	0	0	0	1	0

```
# ... with 2,579 more rows
```

Columns in common with votes_processed

Information on topics

inner_join()

```
> votes_processed %>%
    inner_join(descriptions, by = c("rcid", "session"))
# A tibble: 353,547 × 14
    rcid session  vote ccode  year            country       date   unres    me
   <dbl>   <dbl> <dbl> <int> <dbl>              <chr>     <dttm>   <chr> <dbl>
1     46       2     1     2  1947      United States 1947-09-04 R/2/299     0
2     46       2     1    20  1947             Canada 1947-09-04 R/2/299     0
3     46       2     1    40  1947               Cuba 1947-09-04 R/2/299     0
4     46       2     1    41  1947              Haiti 1947-09-04 R/2/299     0
5     46       2     1    42  1947 Dominican Republic 1947-09-04 R/2/299     0
6     46       2     1    70  1947             Mexico 1947-09-04 R/2/299     0
7     46       2     1    90  1947          Guatemala 1947-09-04 R/2/299     0
8     46       2     1    91  1947           Honduras 1947-09-04 R/2/299     0
9     46       2     1    92  1947        El Salvador 1947-09-04 R/2/299     0
10    46       2     1    93  1947          Nicaragua 1947-09-04 R/2/299     0
# ... with 353,537 more rows, and 5 more variables: nu <dbl>, di <dbl>,
#   hr <dbl>, co <dbl>, ec <dbl>
```

EXPLORATORY DATA ANALYSIS: CASE STUDY

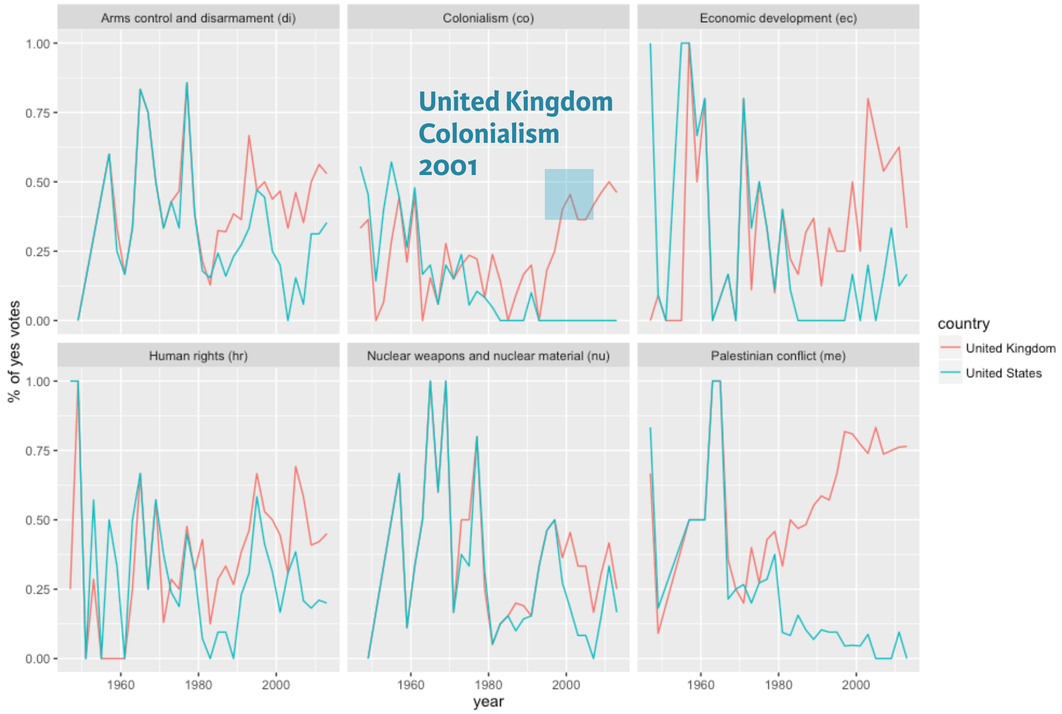
Let’s practice!

EXPLORATORY DATA ANALYSIS: CASE STUDY

Tidy data


Arms control and disarmament (di)
Colonialism (co)
Economic development (ec)
Human rights (hr)
Nuclear weapons and nuclear material (nu)
Palestinian conflict (me)
United Kingdom
Colonialism
2001
% of yes votes
1.00
0.75
0.50
0.25
0.00
1960
1980
2000
year
country
United Kingdom
United States

Tidy data: topic is a variable



Country	Year	Topic
United States	1999	co
United States	2001	co
United States	1999	nu
United States	2001	nu
United Kingdom	1999	co
United Kingdom	2001	co
United Kingdom	1999	nu
United Kingdom	2001	nu

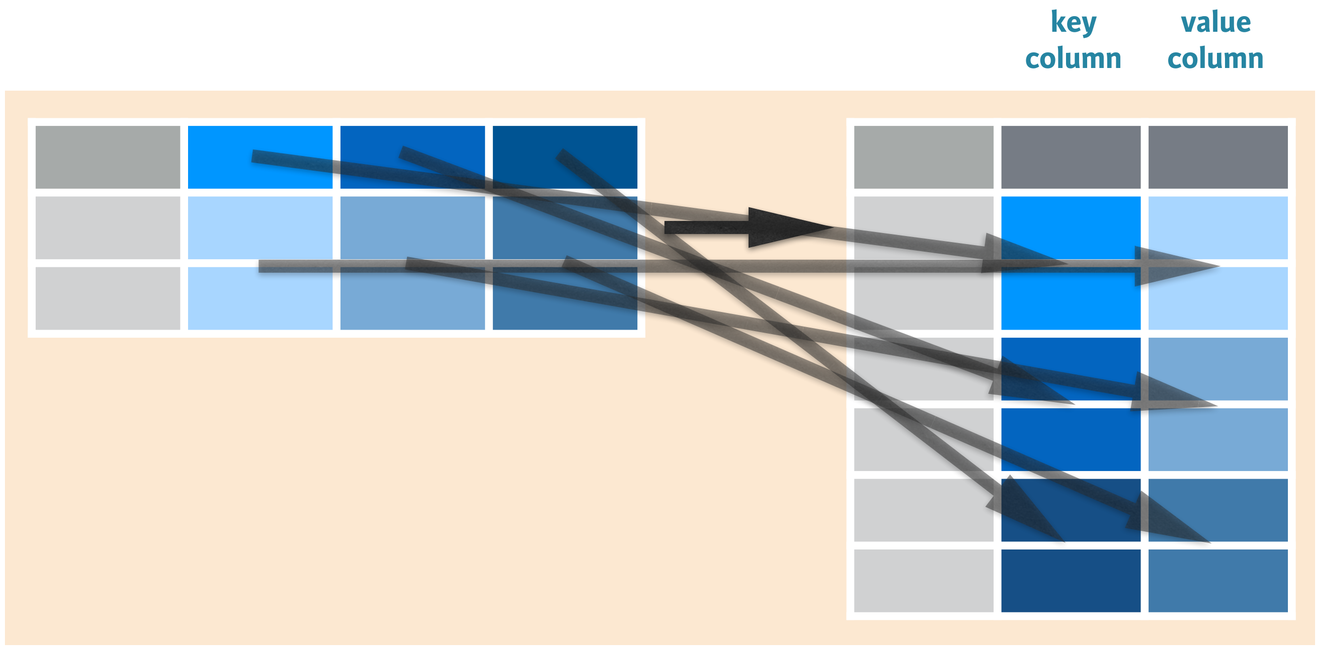
Topic is spread across six columns

```
> votes_joined %>%
    select(rcid, session, vote, country, me:ec)
# A tibble: 353,547 × 10
    rcid session  vote             country    me    nu    di    hr    co    ec
   <dbl>   <dbl> <dbl>               <chr> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
1     46       2     1       United States     0     0     0     0     0     0
2     46       2     1              Canada     0     0     0     0     0     0
3     46       2     1                Cuba     0     0     0     0     0     0
4     46       2     1               Haiti     0     0     0     0     0     0
5     46       2     1  Dominican Republic     0     0     0     0     0     0
6     46       2     1              Mexico     0     0     0     0     0     0
7     46       2     1           Guatemala     0     0     0     0     0     0
8     46       2     1            Honduras     0     0     0     0     0     0
9     46       2     1         El Salvador     0     0     0     0     0     0
10    46       2     1           Nicaragua     0     0     0     0     0     0
# ... with 353,537 more rows
```

Each topic has one column, so combine into a single variable: topic

Use `gather()` to bring columns into two

`gather()` brings multiple columns into just key and value



Use `gather()` to bring columns into two variables

```
> library(tidyr)
> votes_joined %>%
    gather(topic, has_topic, me:ec)
# A tibble: 2,121,282 × 10
    rcid session  vote ccode  year            country       date   unres topic has_topic
   <dbl>   <dbl> <dbl> <int> <dbl>              <chr>     <dttm>   <chr> <chr>     <dbl>
1     46       2     1     2  1947      United States 1947-09-04 R/2/299    me         0
2     46       2     1    20  1947             Canada 1947-09-04 R/2/299    me         0
3     46       2     1    40  1947               Cuba 1947-09-04 R/2/299    me         0
4     46       2     1    41  1947              Haiti 1947-09-04 R/2/299    me         0
5     46       2     1    42  1947 Dominican Republic 1947-09-04 R/2/299    me         0
6     46       2     1    70  1947             Mexico 1947-09-04 R/2/299    me         0
7     46       2     1    90  1947          Guatemala 1947-09-04 R/2/299    me         0
8     46       2     1    91  1947           Honduras 1947-09-04 R/2/299    me         0
9     46       2     1    92  1947        El Salvador 1947-09-04 R/2/299    me         0
10    46       2     1    93  1947          Nicaragua 1947-09-04 R/2/299    me         0
# ... with 2,121,272 more rows
```

Gathered “me” through “ec”

Use `gather()` to bring columns into one variable

```
> library(tidyr)
> votes_joined %>%
    gather(topic, is_topic, me:ec) %>%
    filter(has_topic == 1)
# A tibble: 350,032 × 10
    rcid session  vote ccode  year              country       date    unres topic has_topic
   <dbl>   <dbl> <dbl> <int> <dbl>                <chr>     <dttm>    <chr> <chr>     <dbl>
1     77       2     1     2  1947        United States 1947-11-06 R/2/1424    me         1
2     77       2     1    20  1947               Canada 1947-11-06 R/2/1424    me         1
3     77       2     3    40  1947                 Cuba 1947-11-06 R/2/1424    me         1
4     77       2     1    41  1947                Haiti 1947-11-06 R/2/1424    me         1
5     77       2     1    42  1947 Dominican Republic 1947-11-06 R/2/1424    me         1
6     77       2     2    70  1947               Mexico 1947-11-06 R/2/1424    me         1
7     77       2     1    90  1947            Guatemala 1947-11-06 R/2/1424    me         1
8     77       2     2    91  1947             Honduras 1947-11-06 R/2/1424    me         1
9     77       2     2    92  1947          El Salvador 1947-11-06 R/2/1424    me         1
10    77       2     1    93  1947            Nicaragua 1947-11-06 R/2/1424    me         1
# ... with 350,022 more rows
```

EXPLORATORY DATA ANALYSIS: CASE STUDY

Let’s practice!

EXPLORATORY DATA ANALYSIS: CASE STUDY

Tidy modeling by topic and country

Detecting a trend by topic

Tidy modeling by country

```
> library(tidyr)
> library(purrr)
> library(broom)
> country_coefficients <- by_year_country %>%
    nest(-country) %>%
    mutate(model = map(data, ~ lm(percent_yes ~ year, data = .)),
           tidied = map(model, tidy)) %>%
    unnest(tidied)
> country_coefficients
# A tibble: 399 × 6
       country        term      estimate    std.error statistic      p.value
         <chr>       <chr>         <dbl>        <dbl>     <dbl>        <dbl>
1  Afghanistan (Intercept) -11.063084650 1.4705189228 -7.523252 1.444892e-08
2  Afghanistan        year   0.006009299 0.0007426499  8.091698 3.064797e-09
3    Argentina (Intercept)  -9.464512565 2.1008982371 -4.504984 8.322481e-05
4    Argentina        year   0.005148829 0.0010610076  4.852773 3.047078e-05
5    Australia (Intercept)  -4.545492536 2.1479916283 -2.116159 4.220387e-02
6    Australia        year   0.002567161 0.0010847910  2.366503 2.417617e-02
7      Belarus (Intercept)  -7.000692717 1.5024232546 -4.659601 5.329950e-05
8      Belarus        year   0.003907557 0.0007587624  5.149908 1.284924e-05
9      Belgium (Intercept)  -5.845534016 1.5153390521 -3.857575 5.216573e-04
10     Belgium        year   0.003203234 0.0007652852  4.185673 2.072981e-04
# ... with 389 more rows
```

Tidy modeling by country and topic

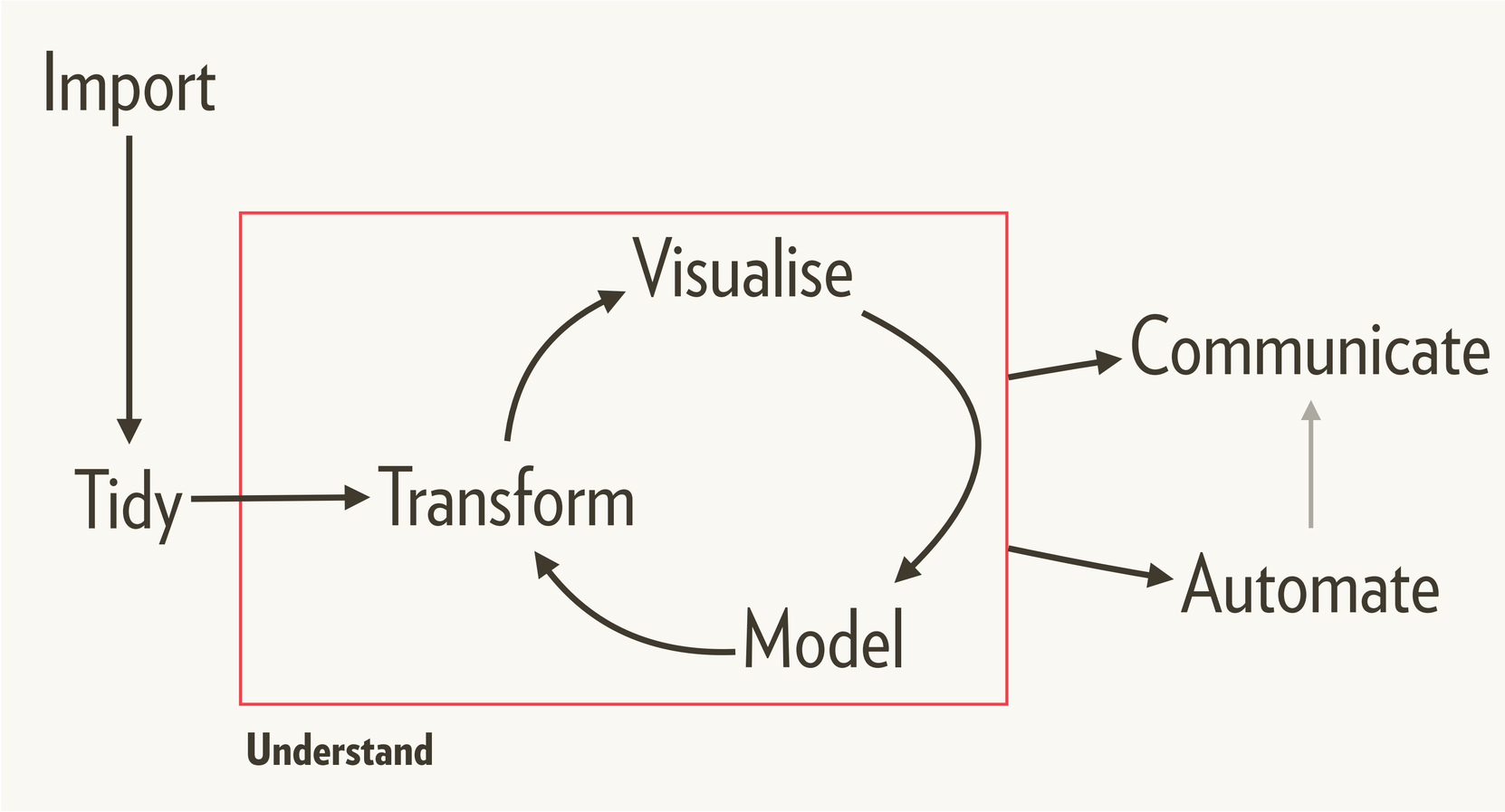
```
> library(purrr)
> library(broom)
> country_topic_coefficients <- by_year_country_topic %>%
    nest(-country, -topic) %>%
    mutate(model = map(data, ~ lm(percent_yes ~ year, data = .)),
           tidied = map(model, tidy)) %>%
    unnest(tidied)
# A tibble: 2,383 × 7
       country                          topic        term      estimate    std.error
         <chr>                          <chr>       <chr>         <dbl>        <dbl>
1  Afghanistan                     Colonialism (Intercept)  -9.196506325 1.9573746777
2  Afghanistan                     Colonialism        year   0.005106200 0.0009885245
3  Afghanistan            Economic development (Intercept) -11.476390441 3.6191205187
4  Afghanistan            Economic development        year   0.006239157 0.0018265400
5  Afghanistan                   Human rights (Intercept)  -7.265379964 4.3740212201
6  Afghanistan                   Human rights        year   0.004075877 0.0022089932
7  Afghanistan           Palestinian conflict (Intercept) -13.313363338 3.5707983095
8  Afghanistan           Palestinian conflict        year   0.007167675 0.0018002649
9  Afghanistan Arms control and disarmament (Intercept) -13.759624843 4.1328667932
10 Afghanistan Arms control and disarmament        year   0.007369733 0.0020837753
# ... with 2,373 more rows, and 2 more variables: statistic <dbl>, p.value <dbl>
```

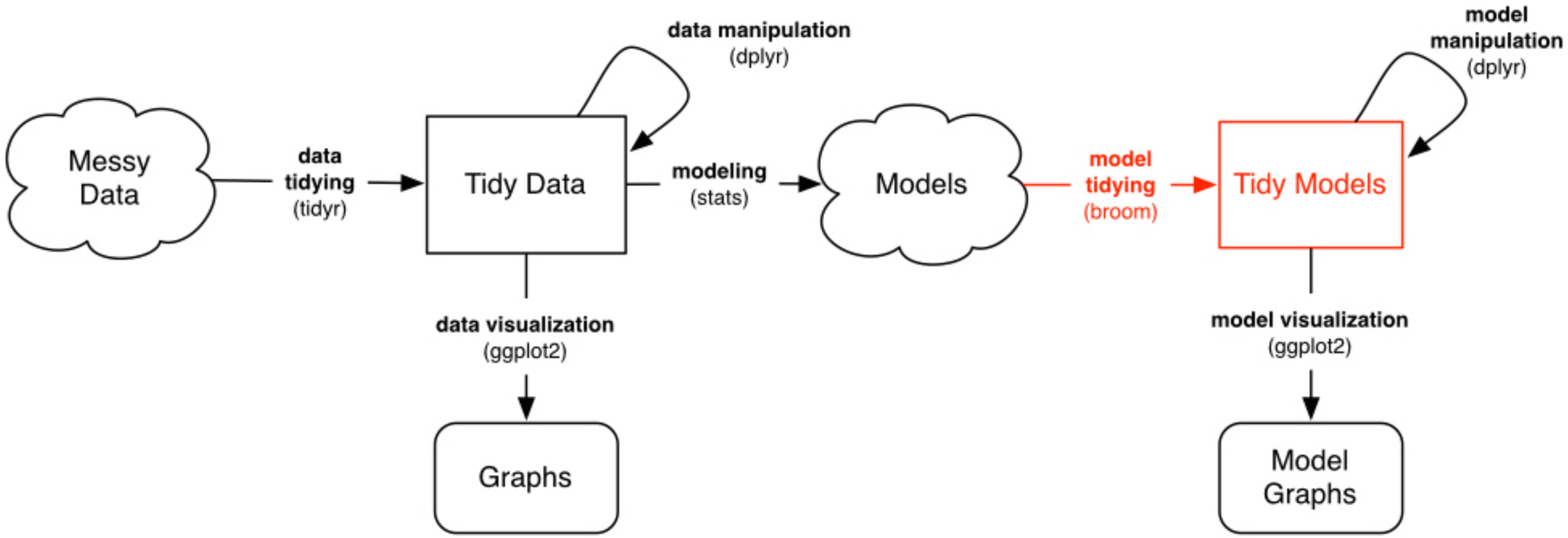
EXPLORATORY DATA ANALYSIS: CASE STUDY

Let’s practice!

EXPLORATORY DATA ANALYSIS: CASE STUDY

Conclusion


Import
Tidy
Transform
Visualise
Model
Understand
Communicate
Automate


Messy Data
data tidying (tidyr)
Tidy Data
data manipulation (dplyr)
modeling (stats)
Models
model tidying (broom)
Tidy Models
model manipulation (dplyr)
data visualization (ggplot2)
Graphs
model visualization (ggplot2)
Model Graphs

Thanks!