~ /> _

Missing Data

~/> previously …

MOST DATA ANALYSIS IS CLEANING & RECODING

country	year	cases	population
Afghanistan	1999	745	19987071
Afghanistan	2000	2666	20595360
Brazil	1999	37737	172006362
Brazil	2000	80488	174504898
China	1999	212258	1272915272
China	2000	213766	1280428583

variables

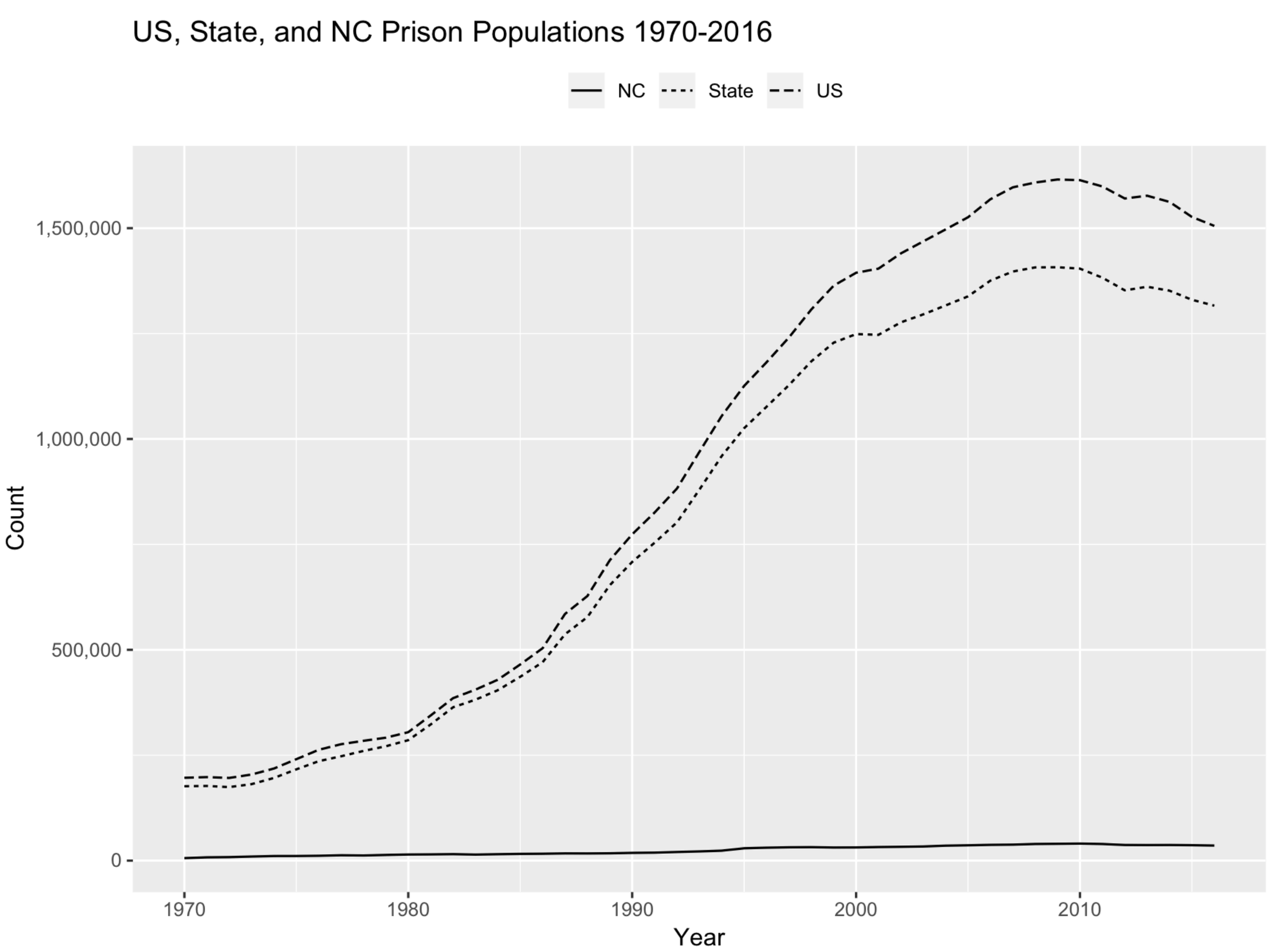
country	year	cases	population
Afghanistan	1999	745	19987071
Afghanistan	2000	2666	20595360
Brazil	1999	37737	172006362
Brazil	2000	80488	174504898
China	1999	212258	1272915272
China	2000	213766	1280428583

observations

country	year	cases	population
Afghanistan	1999	745	19987071
Afghanistan	2000	2666	20595360
Brazil	1999	37737	172006362
Brazil	2000	80488	174504898
China	1999	212258	1272915272
China	2000	213766	1280428583

values

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
Year	US_total_pris_pop	Fed_total_pris_pop	State_total_pris_pop	NC_total_pris_pop	US_total_jail_pop	US_total_pretrial	US_adult_jail_pop	US_adult_pretrial	NC_total_jail_pop	NC_total_pretrial	NC_adult_jail_pop	NC_adult_pretrial	US_pop	US_youth_pop	US_adult_pop	US_prop_adult	NC_pop
1970	196441	20038	176403	5969	160863	83079			3384				205052174	69800000	135252174	0.66	5084411
1971	198061	20948	177113	7795									207660677	69800000	137860677	0.66	5201000
1972	196092	21713	174379	8263	141600	53700							209896021	69400000	140496021	0.67	5296000
1973	204211	22815	181396	9641									211908788	68800000	143108788	0.68	5382000
1974	218466	22361	196105	10932									213853928	68000000	145853928	0.68	5461000
1975	240593	24131	216462	10993									215973199	67200000	148773199	0.69	5535000
1976	262833	26980	235853	11570									218035164	66300000	151735164	0.70	5593000
1977	276157	28650	247507	12769									220239425	65500000	154739425	0.70	5668000
1978	284149	23973	260176	12268	158783	77453	156783		2798				222584545	64800000	157784545	0.71	5739000
1979	291610	20315	271295	13461									225055487	64100000	160955487	0.72	5802000
1980	304692	19025	285667	14456	163994				3924				227224681	63700000	163524681	0.72	5898980
1981	344283	21311	322972	14754									229465714	63200000	166265714	0.72	5956661
1982	385343	21630	363713	15349	209582	119463	207853	118189					231664458	62800000	168864458	0.73	6019141
1983	405501	23836	381665	14257	223551	109567	221815	113984	3496	2515			233791994	62600000	171191994	0.73	6077066
1984	429050	24805	404245	15219	234500		234500	116331					235824902	62500000	173324902	0.73	6164014
1985	465236	29215	436021	16007	256615		250468	127059			3474		237923795	62600000	175323795	0.74	6253989
1986	503794	33135	470659	16373	274444		269179	142112					240132887	62900000	177232887	0.74	6321582
1987	585084	48300	536784	17218	295873		289495	150101					242288918	63100000	179188918	0.74	6403696
1988	627600	49928	577672	17078	343569		341893	175669	5469	4095			244498982	63200000	181298982	0.74	6480598
1989	712364	59171	653193	17454	395553	161948	393303	204291					246819230	63500000	183319230	0.74	6565467
1990	773919	65526	708393	18411	403019	207358	403019	207358					249464396	64215494	185248902	0.74	6656987
1991	825559	71608	753951	18903	424129	217671	424129	217671					252153092	65307843	186845249	0.74	6748135
1992	882500	80259	802241	20454	441780	223840	441780	223840					255029699	66501754	188527945	0.74	6831850
1993	969301	89587	879714	21892	459804	228900	455500	228900	8939				257782608	67585622	190196986	0.74	6947412
1994	1054702	95034	959668	23648	486474		479800						260327021	68631532	191695489	0.74	7060959
1995	1125874	100250	1025624	29253	507044	284100	499300						262803276	69464022	193339254	0.74	7185403
1996	1181919	105544	1076375	30647	518492	298100	510400						265228572	70225449	195003123	0.74	7307658
1997	1240659	112973	1127686	31612	567079	235200	557974	321484					267783607	70916437	196867170	0.74	7428672
1998	1307154	123041	1184113	31961	592462	331800	584372	331323					270248003	71428507	198819496	0.74	7545828
1999	1363686	135246	1228440	31086	605943	327500	596485		13279				272690813	71946360	200744453	0.74	7650789
2000	1394231	145416	1248815	31266	621100	343600	613534						282162411	72400000	209762411	0.74	8081614
2001	1404032	156993	1247039	32253	631240	364900	623628						284968955	72700000	212268955	0.74	8210122
2002	1440144	163528	1276616	32832	665475	394300	658228						287625193	72900000	214725193	0.75	8326201
2003	1468601	173059	1295542	33560	691301	414800	684431						290107933	73100000	217007933	0.75	8422501


US, State, and NC Prison Populations 1970-2016
NC
State
US
1,500,000
1,000,000
500,000
0
Count
1970
1980
1990
2000
2010
Year

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
Year	US_total_pris_pop	Fed_total_pris_pop	State_total_pris_pop	NC_total_pris_pop	US_total_jail_pop	US_total_pretrial	US_adult_jail_pop	US_adult_pretrial	NC_total_jail_pop	NC_total_pretrial	NC_adult_jail_pop	NC_adult_pretrial	US_pop	US_youth_pop	US_adult_pop	US_prop_adult	NC_pop
1970	196441	20038	176403	5969	160863	83079			3384				205052174	69800000	135252174	0.66	5084411
1971	198061	20948	177113	7795									207660677	69800000	137860677	0.66	5201000
1972	196092	21713	174379	8263	141600	53700							209896021	69400000	140496021	0.67	5296000
1973	204211	22815	181396	9641									211908788	68800000	143108788	0.68	5382000
1974	218466	22361	196105	10932									213853928	68000000	145853928	0.68	5461000
1975	240593	24131	216462	10993									215973199	67200000	148773199	0.69	5535000
1976	262833	26980	235853	11570									218035164	66300000	151735164	0.70	5593000
1977	276157	28650	247507	12769									220239425	65500000	154739425	0.70	5668000
1978	284149	23973	260176	12268	158783	77453	156783		2798				222584545	64800000	157784545	0.71	5739000
1979	291610	20315	271295	13461									225055487	64100000	160955487	0.72	5802000
1980	304692	19025	285667	14456	163994				3924				227224681	63700000	163524681	0.72	5898980
1981	344283	21311	322972	14754									229465714	63200000	166265714	0.72	5956661
1982	385343	21630	363713	15349	209582	119463	207853	118189					231664458	62800000	168864458	0.73	6019141
1983	405501	23836	381665	14257	223551	109567	221815	113984	3496	2515			233791994	62600000	171191994	0.73	6077066
1984	429050	24805	404245	15219	234500		234500	116331					235824902	62500000	173324902	0.73	6164014
1985	465236	29215	436021	16007	256615		250468	127059			3474		237923795	62600000	175323795	0.74	6253989
1986	503794	33135	470659	16373	274444		269179	142112					240132887	62900000	177232887	0.74	6321582
1987	585084	48300	536784	17218	295873		289495	150101					242288918	63100000	179188918	0.74	6403696
1988	627600	49928	577672	17078	343569		341893	175669	5469	4095			244498982	63200000	181298982	0.74	6480598
1989	712364	59171	653193	17454	395553	161948	393303	204291					246819230	63500000	183319230	0.74	6565467
1990	773919	65526	708393	18411	403019	207358	403019	207358					249464396	64215494	185248902	0.74	6656987
1991	825559	71608	753951	18903	424129	217671	424129	217671					252153092	65307843	186845249	0.74	6748135
1992	882500	80259	802241	20454	441780	223840	441780	223840					255029699	66501754	188527945	0.74	6831850
1993	969301	89587	879714	21892	459804	228900	455500	228900	8939				257782608	67585622	190196986	0.74	6947412
1994	1054702	95034	959668	23648	486474		479800						260327021	68631532	191695489	0.74	7060959
1995	1125874	100250	1025624	29253	507044	284100	499300						262803276	69464022	193339254	0.74	7185403
1996	1181919	105544	1076375	30647	518492	298100	510400						265228572	70225449	195003123	0.74	7307658
1997	1240659	112973	1127686	31612	567079	235200	557974	321484					267783607	70916437	196867170	0.74	7428672
1998	1307154	123041	1184113	31961	592462	331800	584372	331323					270248003	71428507	198819496	0.74	7545828
1999	1363686	135246	1228440	31086	605943	327500	596485		13279				272690813	71946360	200744453	0.74	7650789
2000	1394231	145416	1248815	31266	621100	343600	613534						282162411	72400000	209762411	0.74	8081614
2001	1404032	156993	1247039	32253	631240	364900	623628						284968955	72700000	212268955	0.74	8210122
2002	1440144	163528	1276616	32832	665475	394300	658228						287625193	72900000	214725193	0.75	8326201
2003	1468601	173059	1295542	33560	691301	414800	684431						290107933	73100000	217007933	0.75	8422501

~/> janitor

```r
library(tidyverse)
library(janitor)

data <- clean_names(read_csv("data/prison_jail_1970_2016.csv"))

data
```

```
# A tibble: 47 x 20
    year us_total_pris_p… fed_total_pris_… state_total_pri…
   <dbl>            <dbl>            <dbl>            <dbl>
 1  1970           196441            20038           176403
 2  1971           198061            20948           177113
 3  1972           196092            21713           174379
 4  1973           204211            22815           181396
 5  1974           218466            22361           196105
 6  1975           240593            24131           216462
 7  1976           262833            26980           235853
 8  1977           276157            28650           247507
 9  1978           284149            23973           260176
10  1979           291610            20315           271295
# … with 37 more rows, and 16 more variables:
#   nc_total_pris_pop <dbl>, us_total_jail_pop <dbl>,
#   us_total_pretrial <dbl>, us_adult_jail_pop <dbl>,
#   us_adult_pretrial <dbl>, nc_total_jail_pop <dbl>,
#   nc_total_pretrial <dbl>, nc_adult_jail_pop <dbl>,
#   nc_adult_pretrial <lgl>, us_pop <dbl>, us_youth_pop <dbl>,
#   us_adult_pop <dbl>, us_prop_adult <dbl>, nc_pop <dbl>,
#   nc_youth_pop <dbl>, nc_adult_pop <dbl>
```

```
data %>% select(year, us_total_pris_pop,
                state_total_pris_pop,
                nc_total_pris_pop)
```

```
# A tibble: 47 x 4
    year us_total_pris_pop state_total_pris_pop nc_total_pris_pop
   <dbl>             <dbl>                <dbl>             <dbl>
 1  1970            196441               176403              5969
 2  1971            198061               177113              7795
 3  1972            196092               174379              8263
 4  1973            204211               181396              9641
 5  1974            218466               196105             10932
 6  1975            240593               216462             10993
 7  1976            262833               235853             11570
 8  1977            276157               247507             12769
 9  1978            284149               260176             12268
10  1979            291610               271295             13461
# … with 37 more rows
```

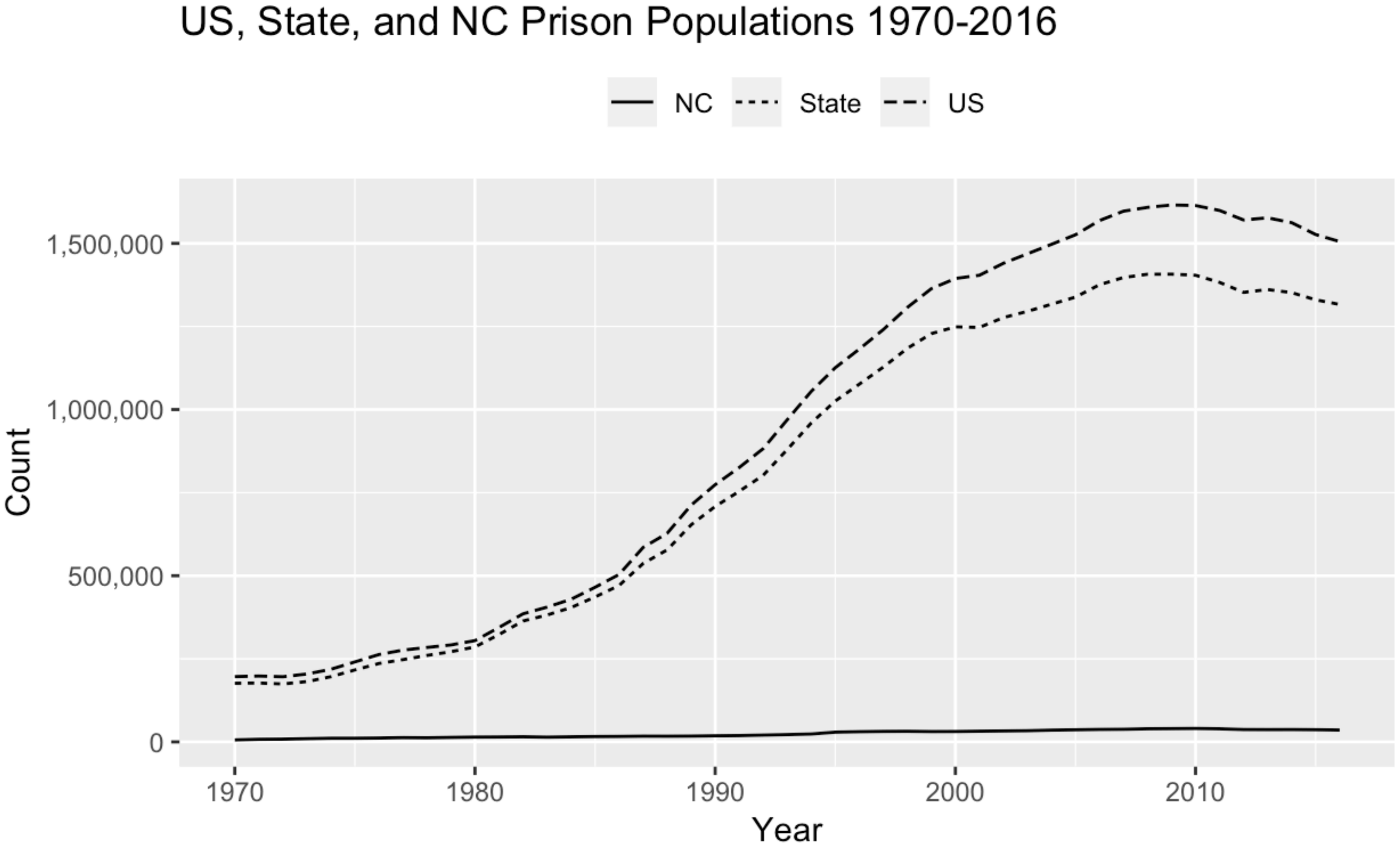
```
data %>% select(year, us_total_pris_pop,
                state_total_pris_pop,
                nc_total_pris_pop) %>%
    gather(series, count, us_total_pris_pop:nc_total_pris_pop)
```

```
# A tibble: 141 x 3
    year series              count
   <dbl> <chr>               <dbl>
 1  1970 us_total_pris_pop 196441
 2  1971 us_total_pris_pop 198061
 3  1972 us_total_pris_pop 196092
 4  1973 us_total_pris_pop 204211
 5  1974 us_total_pris_pop 218466
 6  1975 us_total_pris_pop 240593
 7  1976 us_total_pris_pop 262833
 8  1977 us_total_pris_pop 276157
 9  1978 us_total_pris_pop 284149
10  1979 us_total_pris_pop 291610
# … with 131 more rows
```

```
data %>% select(year, us_total_pris_pop,
                state_total_pris_pop,
                nc_total_pris_pop) %>%
    gather(series, count, us_total_pris_pop:nc_total_pris_pop) %>%
    mutate(series = recode(series, us_total_pris_pop = "US",
                           state_total_pris_pop = "State",
                           nc_total_pris_pop = "NC"))
```

```
# A tibble: 141 x 3
    year series  count
   <dbl> <chr>   <dbl>
 1  1970 US     196441
 2  1971 US     198061
 3  1972 US     196092
 4  1973 US     204211
 5  1974 US     218466
 6  1975 US     240593
 7  1976 US     262833
 8  1977 US     276157
 9  1978 US     284149
10  1979 US     291610
# … with 131 more rows
```

```r
data %>% select(year, us_total_pris_pop,
                state_total_pris_pop,
                nc_total_pris_pop) %>%
    gather(series, count, us_total_pris_pop:nc_total_pris_pop) %>%
    mutate(series = recode(series, us_total_pris_pop = "US",
                           state_total_pris_pop = "State",
                           nc_total_pris_pop = "NC")) %>%
    ggplot(aes(x = year, y = count, linetype = series)) +
    geom_line() +
    scale_y_continuous(labels = scales::comma) +
    labs(x = "Year", y = "Count", linetype = NULL,
         title = "US, State, and NC Prison Populations 1970-2016") +
    theme(legend.position = "top")
```


US, State, and NC Prison Populations 1970-2016
NC
State
US
1,500,000
1,000,000
500,000
0
Count
1970
1980
1990
2000
2010
Year

~ /> _

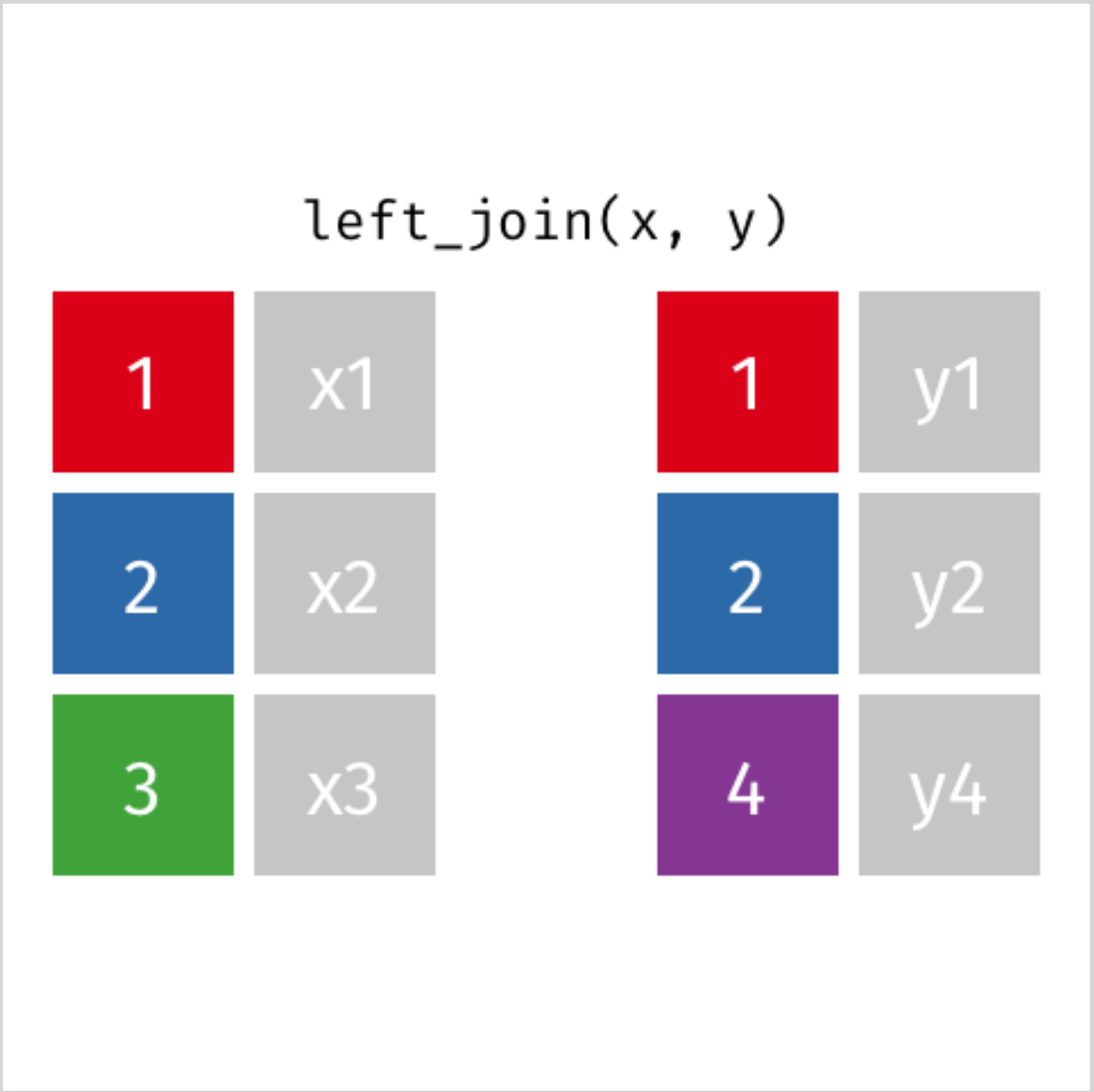
TABLE JOINS

x

1	x1
2	x2
3	x3

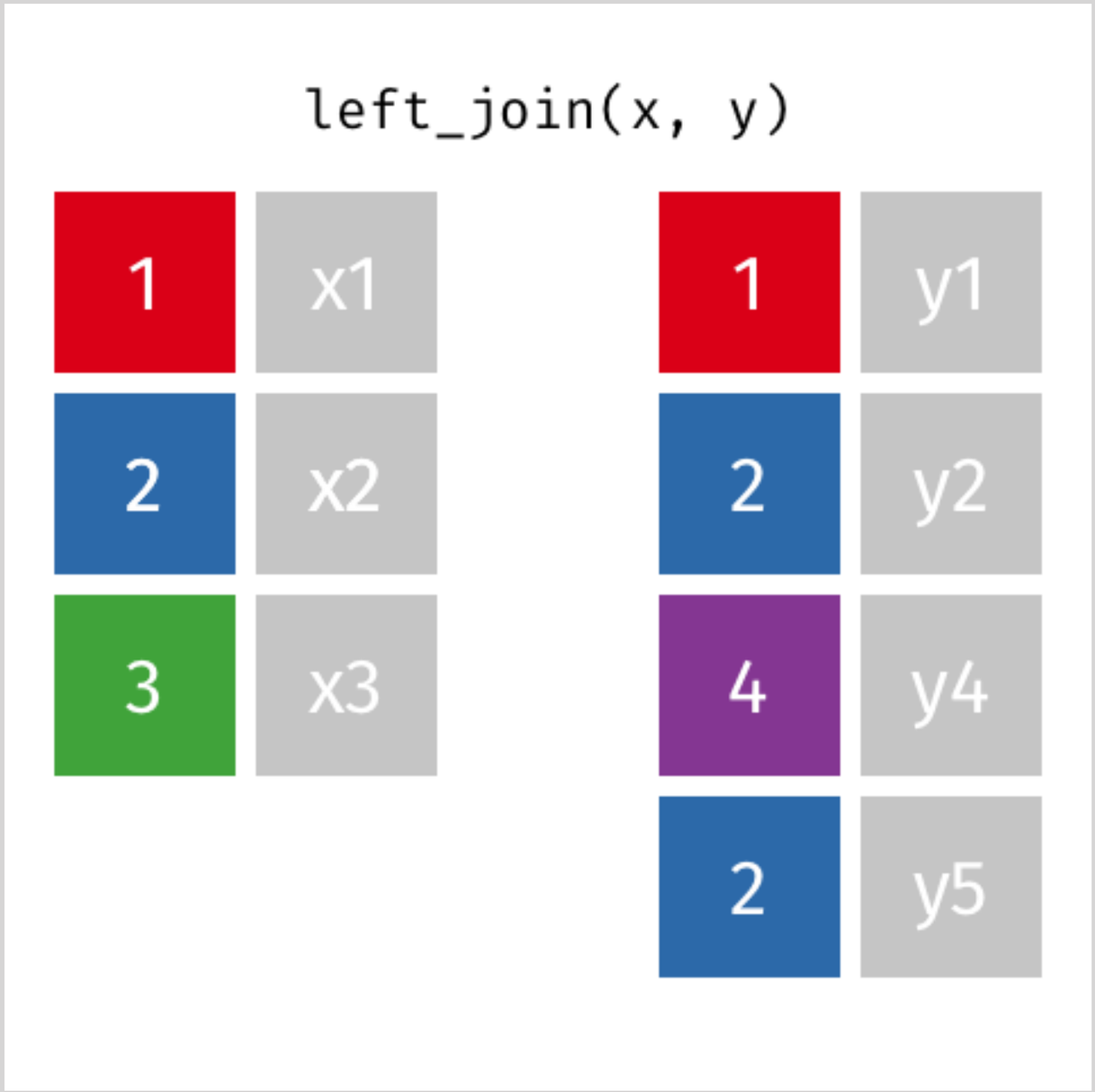
y

1	y1
2	y2
4	y4



All rows from x, and all columns from x and y. Rows in x with no match in y will have NA values in the new columns.

LEFT JOIN



If there are multiple matches between x and y, all combinations of the matches are returned.

LEFT JOIN

```r
senate <- data %>%
  filter(position == "U.S. Senator") %>%
  group_by(pid) %>%
  summarize(first = first(first),
            last = first(last),
            party = first(party),
            state = first(state),
            start = first(start),
            end = first(end))
```

```r
house <- data %>%
  filter(position == "U.S. Representative") %>%
  group_by(pid) %>%
  summarize(first = first(first),
            last = first(last),
            party = first(party),
            state = first(state),
            district = first(district),
            start = first(start),
            end = first(end))
```

```r
senate <- data %>%
  filter(position == "U.S. Senator") %>%
  group_by(pid) %>%
  summarize(first = first(first),
            last = first(last),
            party = first(party),
            state = first(state),
            start = first(start),
            end = first(end))
```

```r
house <- data %>%
  filter(position == "U.S. Representative") %>%
  group_by(pid) %>%
  summarize(state = first(state),
            district = first(district),
            start = first(start),
            end = first(end))
```

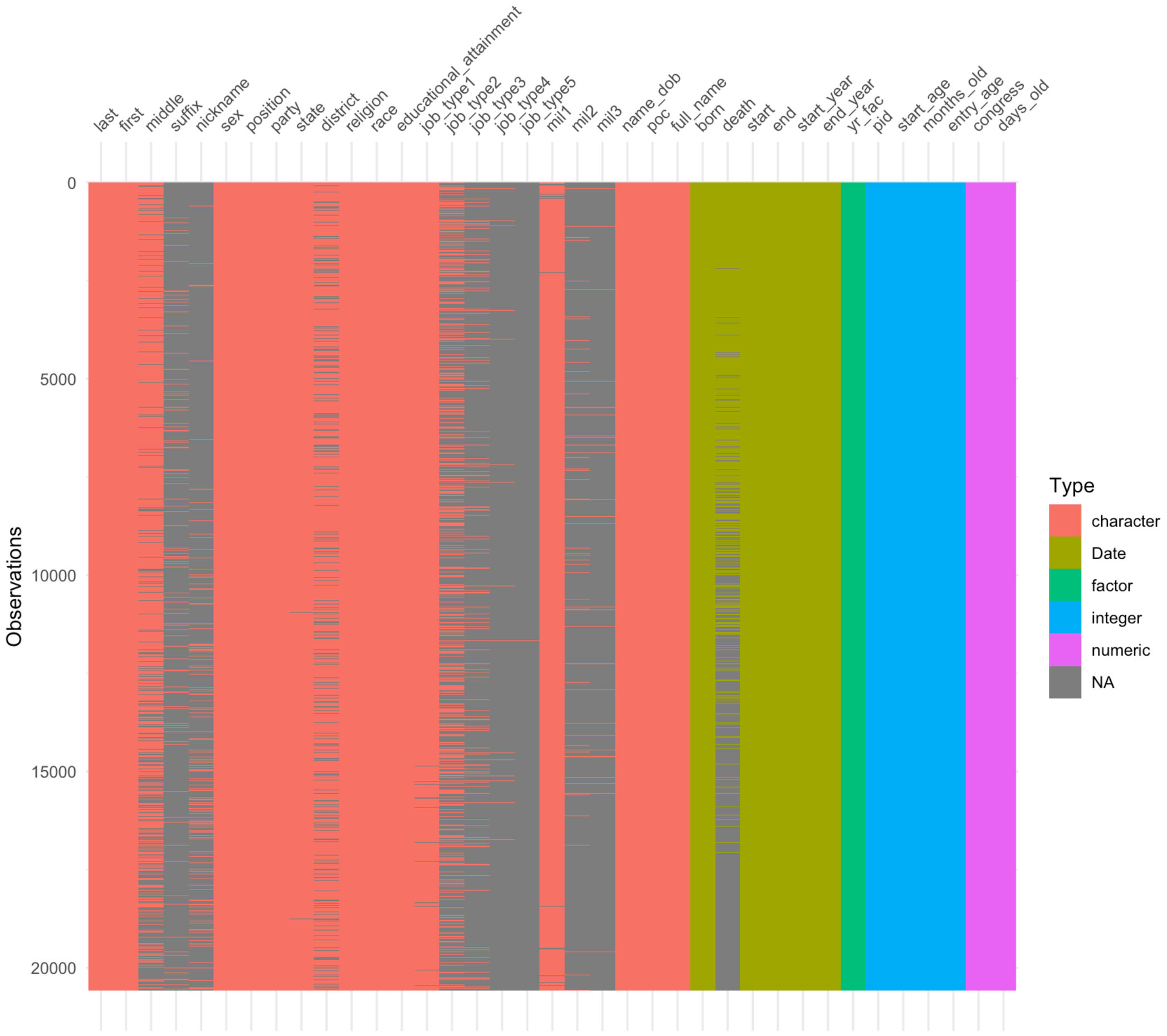
```r
sen_and_house <- inner_join(senate, house, by = "pid")

sen_and_house
```

```
# A tibble: 198 x 11
     pid first last  party state.x start.x    end.x      state.y
   <int> <chr> <chr> <chr> <chr>   <date>     <date>     <chr>
 1     8 Clin… Ande… Demo… NM      1949-01-03 1973-01-03 NM
 2    27 Frank Barr… Repu… WY      1953-01-03 1959-01-03 WY
 3    32 James Beall Repu… MD      1953-01-03 1965-01-03 MD
 4    35 Geor… Bend… Repu… OH      1954-12-16 1957-01-03 OH
 5    67 Thom… Burch Demo… VA      1946-05-31 1946-11-05 VA
 6    83 Frank Carl… Repu… KS      1950-11-29 1969-01-03 KS
 7    86 Clif… Case  Repu… NJ      1955-01-05 1979-01-03 NJ
 8    87 Fran… Case  Repu… SD      1951-01-03 1962-06-22 SD
 9    90 Virg… Chap… Demo… KY      1949-01-03 1951-03-08 KY
10    98 Earle Clem… Demo… KY      1950-11-27 1957-01-03 KY
# … with 188 more rows, and 3 more variables: district <chr>,
#   start.y <date>, end.y <date>
```

~ /> _

MISSING DATA

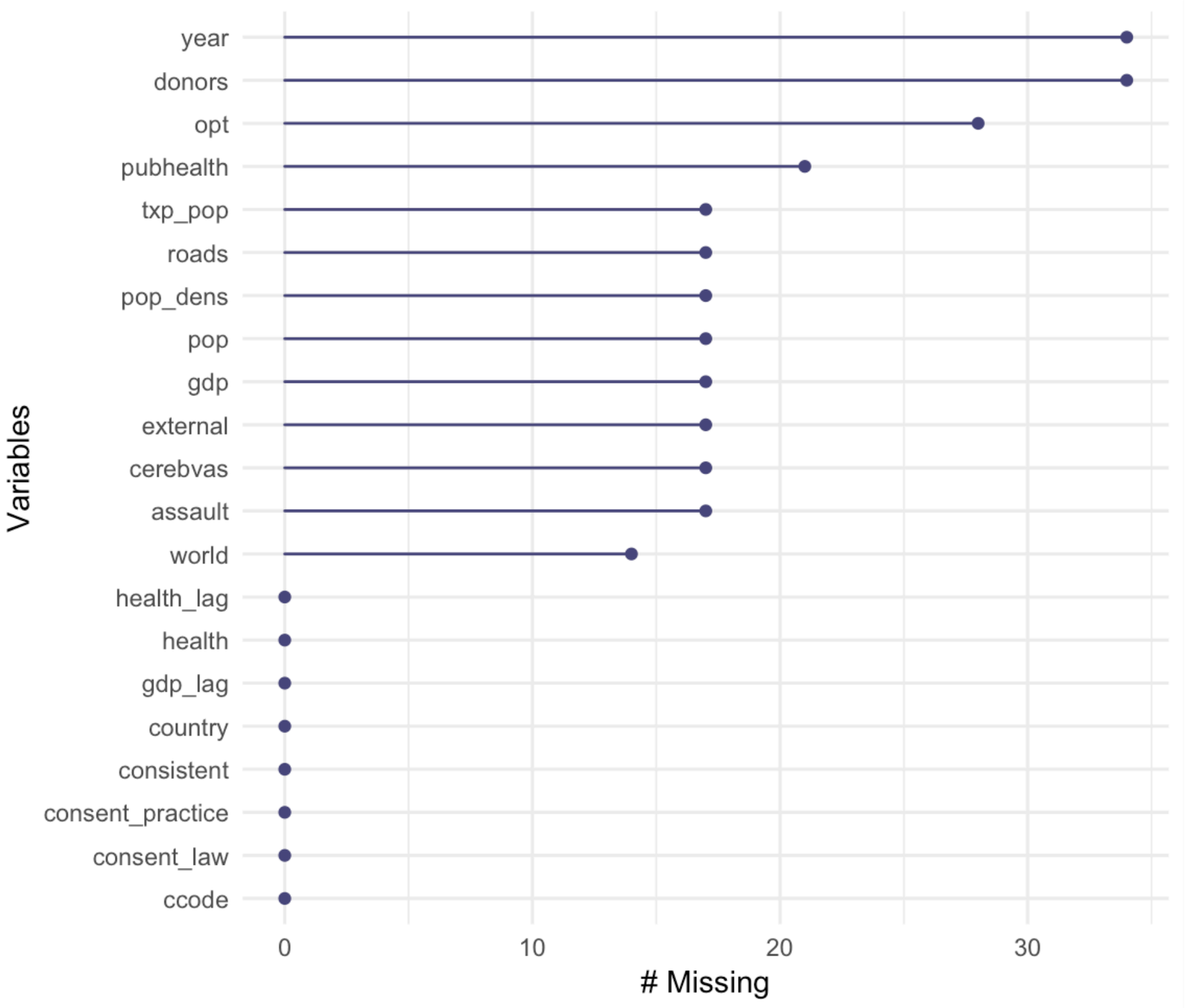
```
library(naniar)
library(visdat)

vis_dat(data)
```


last
first
middle
suffix
nickname
sex
position
party
state
district
religion
race
educational_attainment
job_type1
job_type2
job_type3
job_type4
job_type5
mil1
mil2
mil3
name_dob
poc
full_name
born
death
start
end
start_year
end_year
yr_fac
pid
start_age
months_old
entry_age
congress
days_old
0
5000
10000
15000
20000
Observations
Type
character
Date
factor
integer
numeric
NA

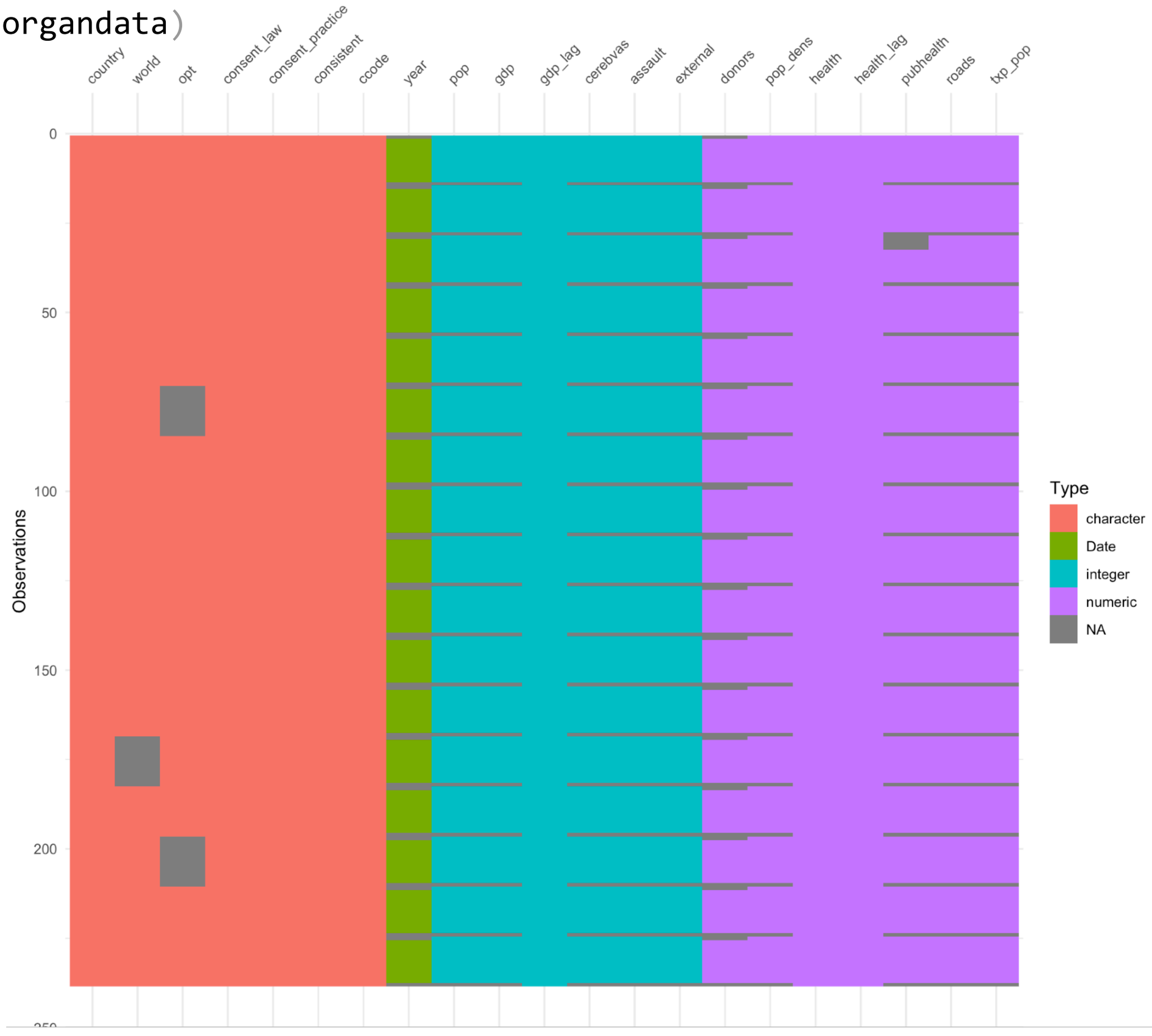
```
library(socviz)
organdata
```

```
# A tibble: 238 x 21
   country year       donors   pop pop_dens   gdp gdp_lag health
   <chr>   <date>      <dbl> <int>    <dbl> <int>   <int>  <dbl>
 1 Austra… NA           NA    17065    0.220 16774   16591   1300
 2 Austra… 1991-01-01  12.1   17284    0.223 17171   16774   1379
 3 Austra… 1992-01-01  12.4   17495    0.226 17914   17171   1455
 4 Austra… 1993-01-01  12.5   17667    0.228 18883   17914   1540
 5 Austra… 1994-01-01  10.2   17855    0.231 19849   18883   1626
 6 Austra… 1995-01-01  10.2   18072    0.233 21079   19849   1737
 7 Austra… 1996-01-01  10.6   18311    0.237 21923   21079   1846
 8 Austra… 1997-01-01  10.3   18518    0.239 22961   21923   1948
 9 Austra… 1998-01-01  10.5   18711    0.242 24148   22961   2077
10 Austra… 1999-01-01   8.67 18926    0.244 25445   24148   2231
# … with 228 more rows, and 13 more variables: health_lag <dbl>,
#   pubhealth <dbl>, roads <dbl>, cerebvas <int>, assault <int>,
#   external <int>, txp_pop <dbl>, world <chr>, opt <chr>,
#   consent_law <chr>, consent_practice <chr>, consistent <chr>,
#   ccode <chr>
```

```
gg_miss_var(organdata)
```



```
vis_dat(organdata)
```



```
miss_var_summary(organdata)
```

```
A tibble: 21 x 3
   variable  n_miss pct_miss
   <chr>      <int>    <dbl>
 1 year          34    14.3
 2 donors        34    14.3
 3 opt           28    11.8
 4 pubhealth     21     8.82
 5 pop           17     7.14
 6 pop_dens      17     7.14
 7 gdp           17     7.14
 8 roads         17     7.14
 9 cerebvas      17     7.14
10 assault       17     7.14
# … with 11 more rows
```

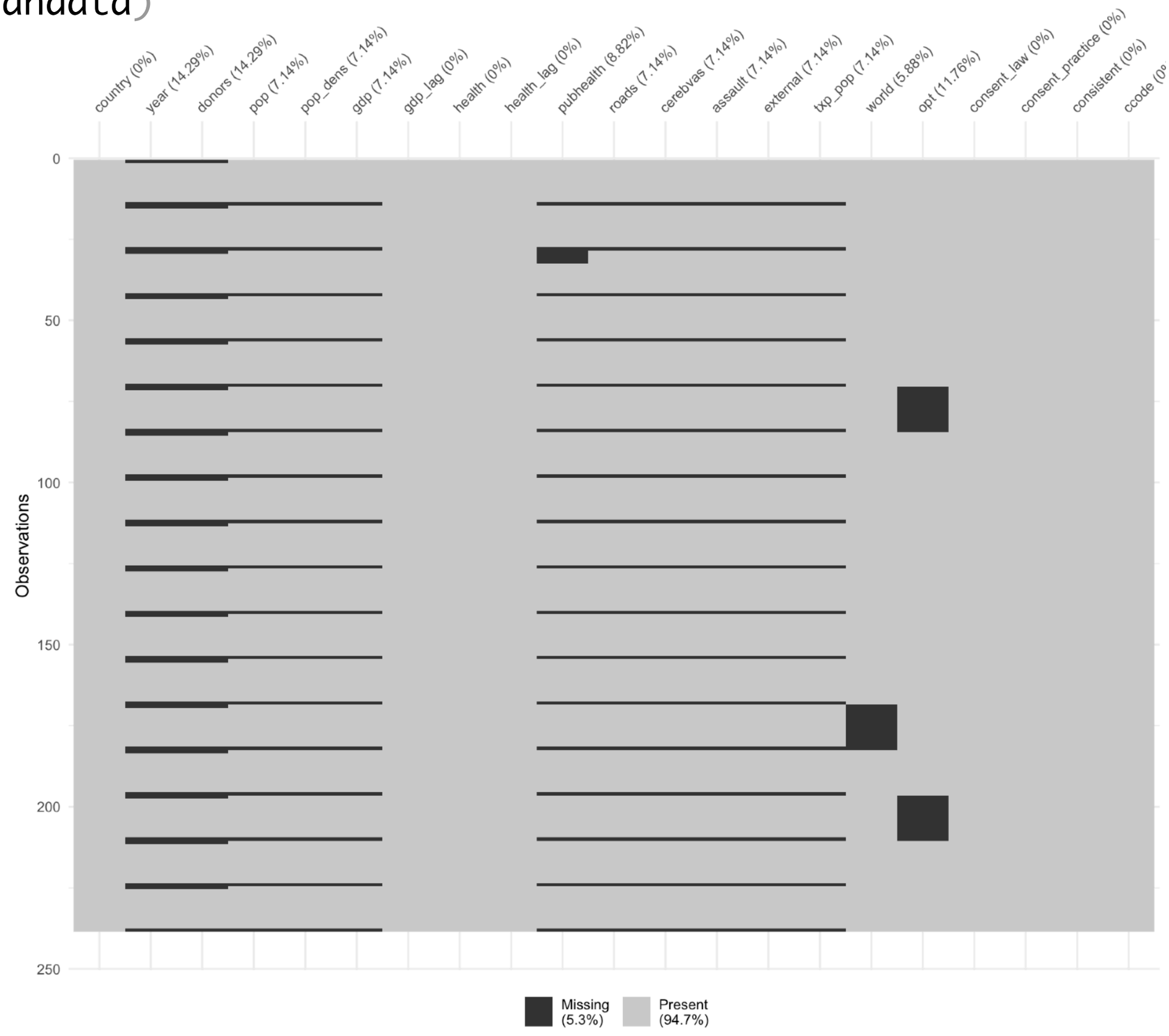
```
miss_case_summary(organdata)
```

```
A tibble: 238 x 3
    case n_miss pct_miss
   <int>  <int>    <dbl>
 1    84     12     57.1
 2   182     12     57.1
 3   210     12     57.1
 4    14     11     52.4
 5    28     11     52.4
 6    42     11     52.4
 7    56     11     52.4
 8    70     11     52.4
 9    98     11     52.4
10   112     11     52.4
# … with 228 more rows
```

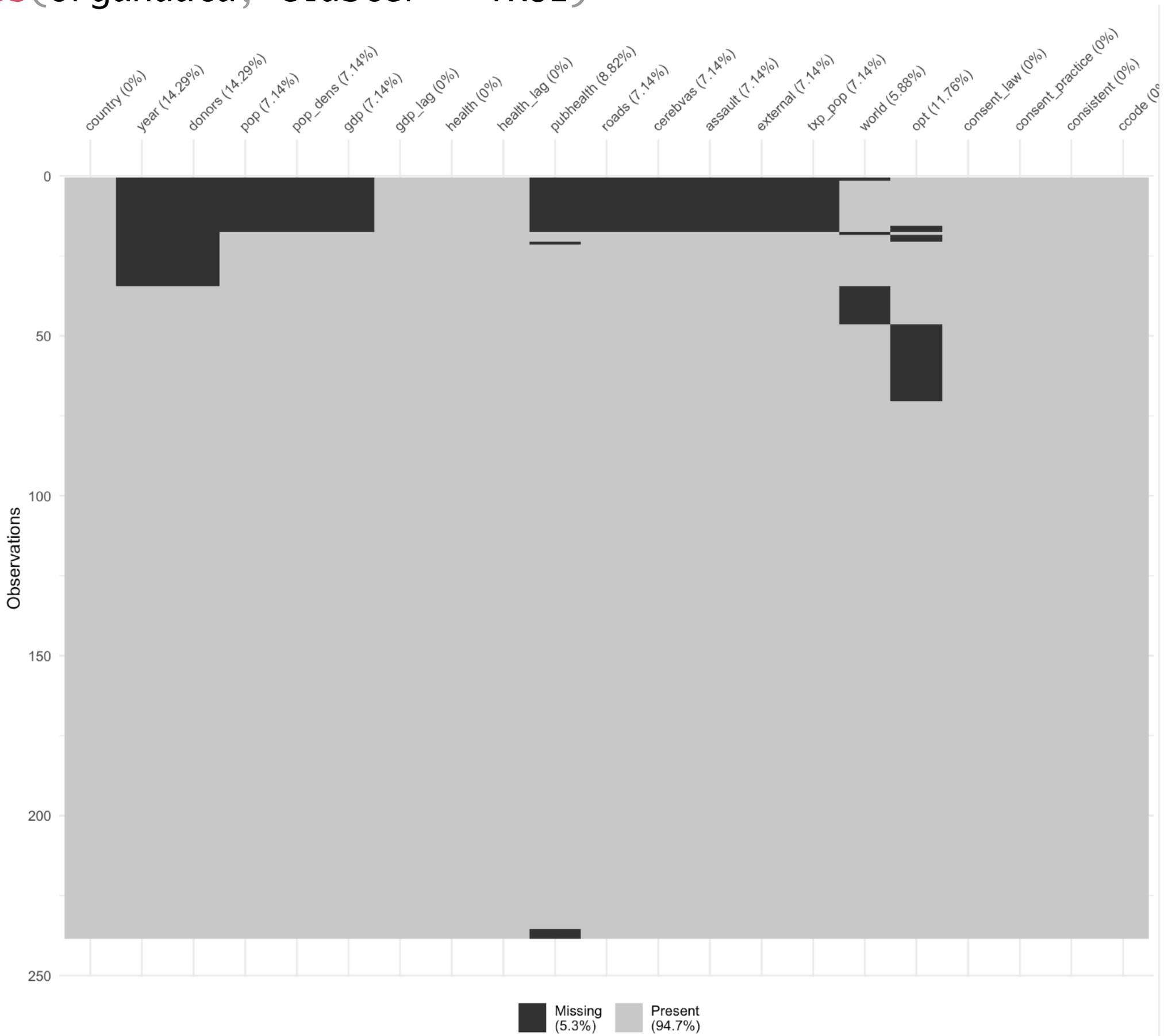
```
organdata %>%
  select(consent_law, year, pubhealth, roads) %>%
  group_by(consent_law) %>%
  miss_var_summary()
```

```
A tibble: 6 x 4
  consent_law variable  n_miss pct_miss
  <chr>       <chr>      <int>    <dbl>
1 Informed    year          16    14.3 
2 Informed    pubhealth      8     7.14
3 Informed    roads          8     7.14
4 Presumed    year          18    14.3 
5 Presumed    pubhealth     13    10.3 
6 Presumed    roads          9     7.14
```

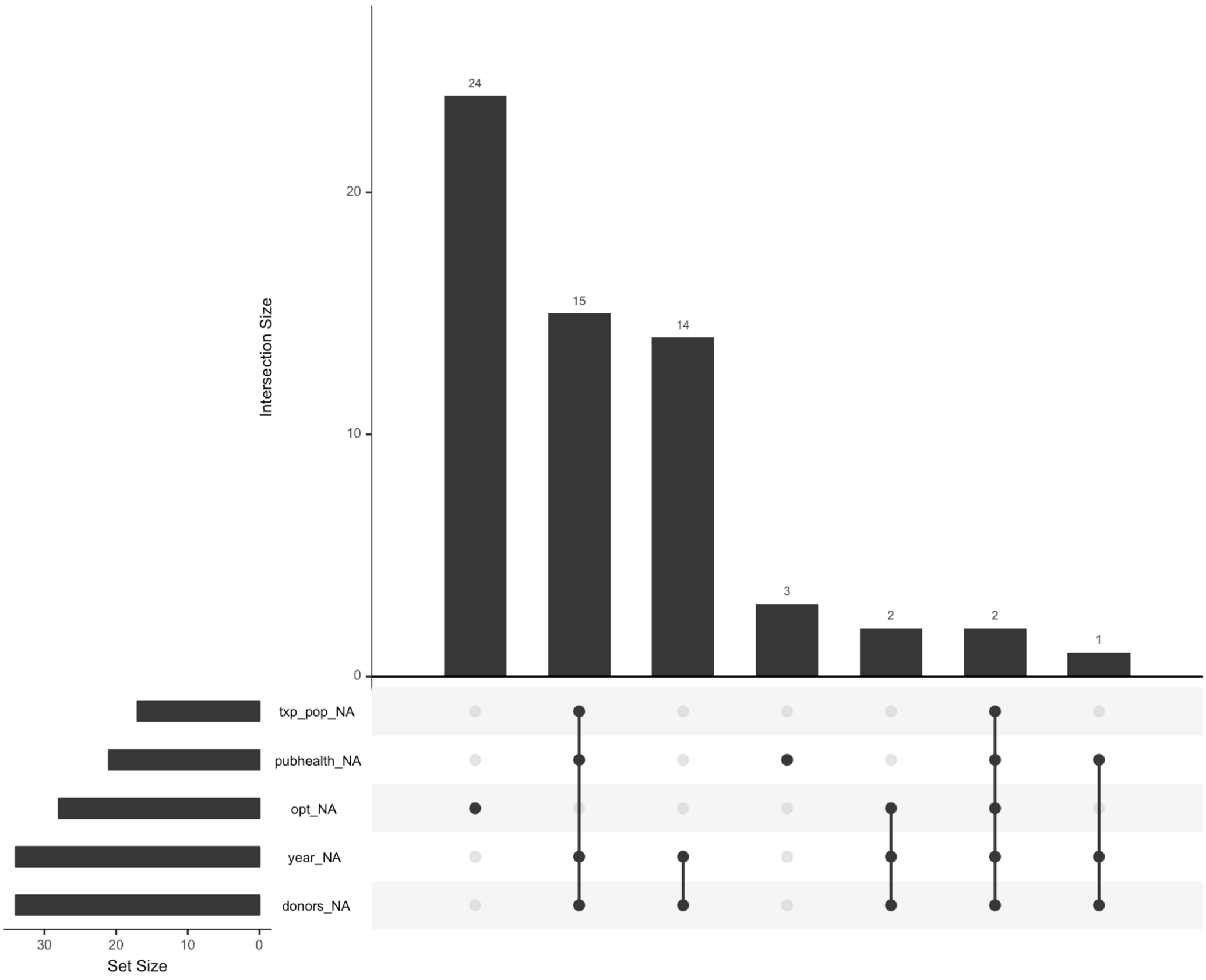
```
vis_miss(organdata)
```



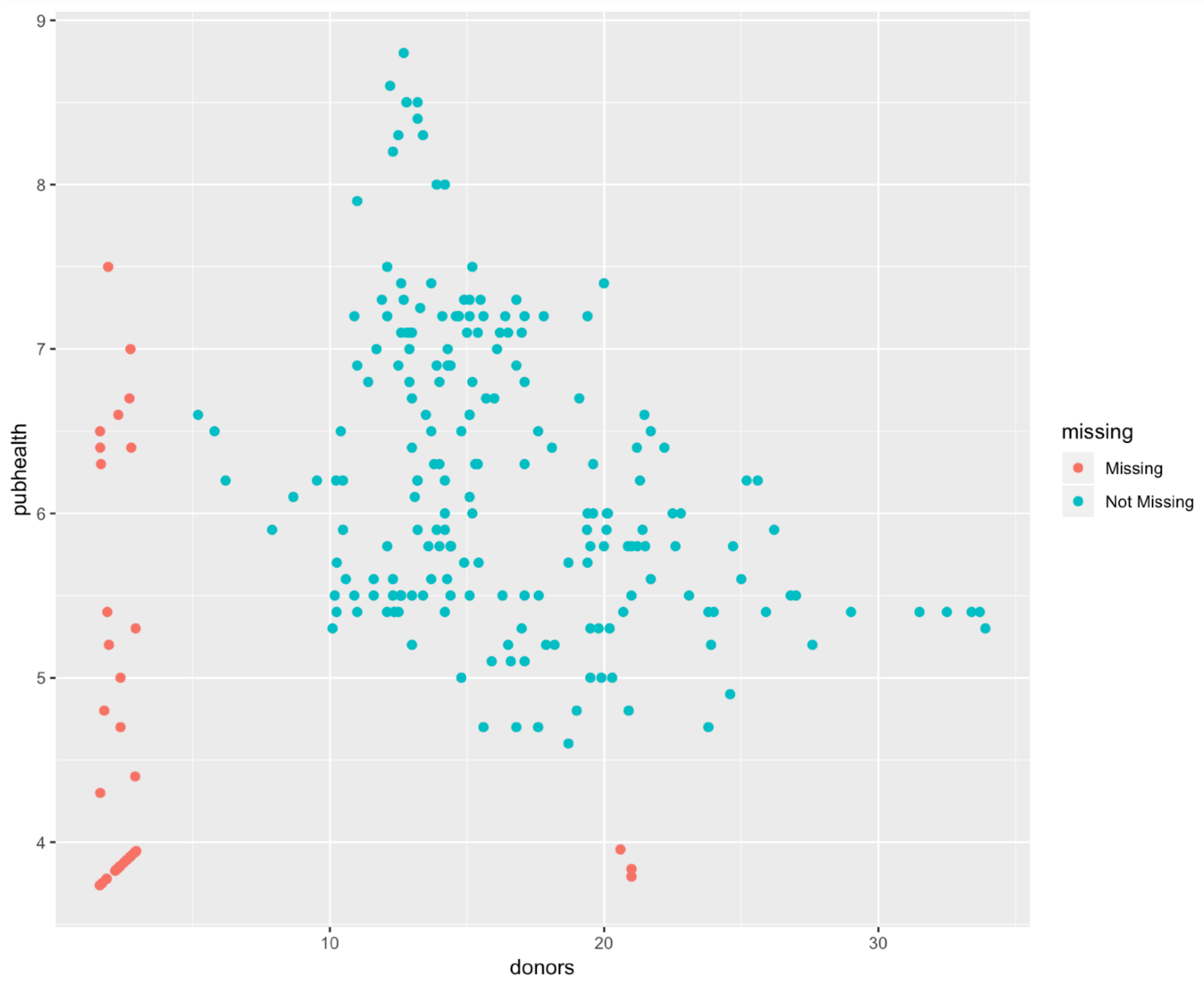
```
vis_miss(organdata, cluster = TRUE)
```



```
gg_miss_upset(organdata)
```



```
ggplot(organdata,
       aes(x = donors,
           y = pubhealth)) +
  geom_miss_point()
```



```
gg_miss_fct(x = riskfactors, fct = marital)
```