class_04

Class Assessment (Where are you in your R journey?)

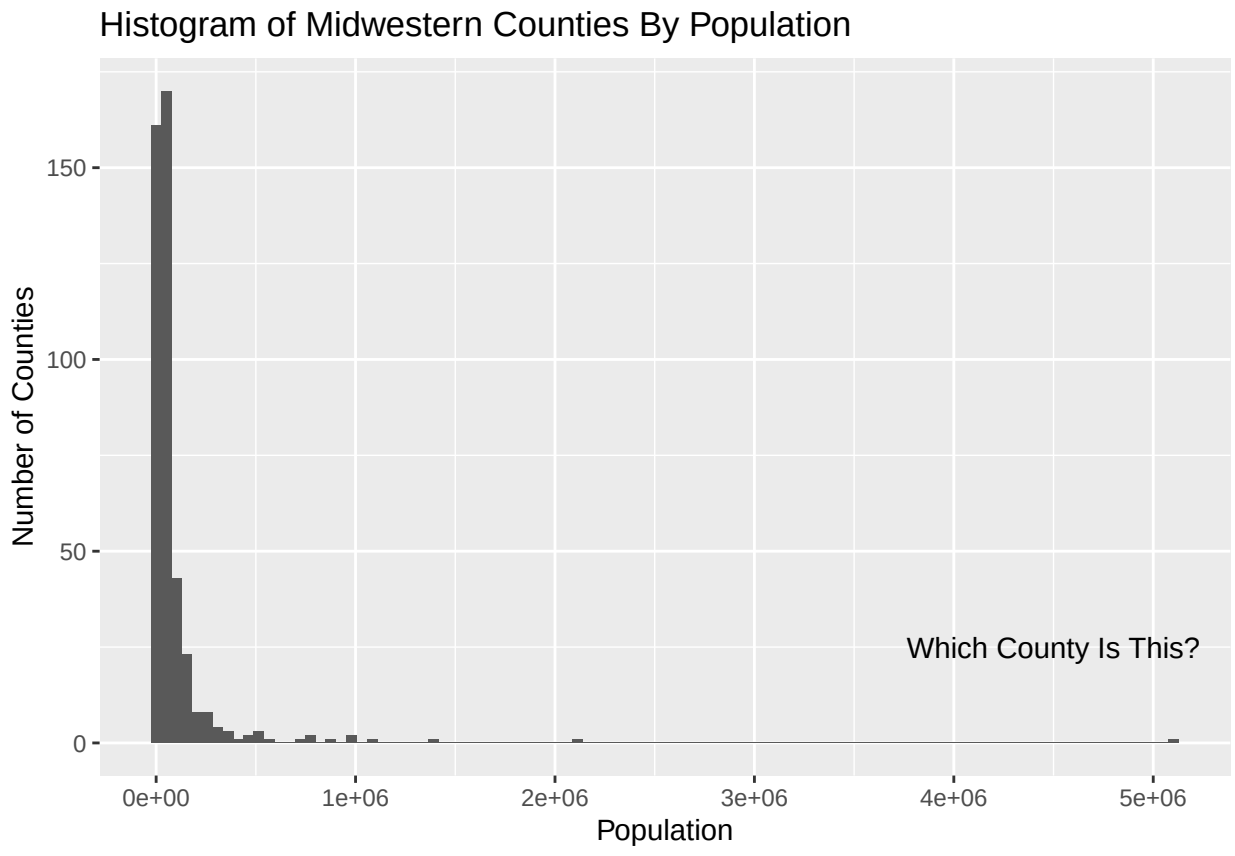
These are meant to help you assess where you are in your R journey. If you can answer both of these questions without help, then you are a in a great place. If you struggle with these, practice some more. I recommend the primers you worked on last week or the exercises at the end of the chapters in R4.DS.

```
#look at the basics of the midwest dataset
summary(midwest)
```

```
##       PID          county             state               area        
##  Min.   : 561   Length:437         Length:437         Min.   :0.00500  
##  1st Qu.: 670   Class :character   Class :character   1st Qu.:0.02400  
##  Median :1221   Mode  :character   Mode  :character   Median :0.03000  
##  Mean   :1437                                         Mean   :0.03317  
##  3rd Qu.:2059                                         3rd Qu.:0.03800  
##  Max.   :3052                                         Max.   :0.11000  
##     poptotal         popdensity          popwhite          popblack      
##  Min.   :   1701   Min.   :   85.05   Min.   :    416   Min.   :      0  
##  1st Qu.:  18840   1st Qu.:  622.41   1st Qu.:  18630   1st Qu.:     29  
##  Median :  35324   Median : 1156.21   Median :  34471   Median :    201  
##  Mean   :  96130   Mean   : 3097.74   Mean   :  81840   Mean   :  11024  
##  3rd Qu.:  75651   3rd Qu.: 2330.00   3rd Qu.:  72968   3rd Qu.:   1291  
##  Max.   :5105067   Max.   :88018.40   Max.   :3204947   Max.   :1317147  
##  popamerindian        popasian         popother        percwhite     
##  Min.   :    4.0   Min.   :     0   Min.   :     0   Min.   :10.69  
##  1st Qu.:   44.0   1st Qu.:    35   1st Qu.:    20   1st Qu.:94.89  
##  Median :   94.0   Median :   102   Median :    66   Median :98.03  
##  Mean   :  343.1   Mean   :  1310   Mean   :  1613   Mean   :95.56  
##  3rd Qu.:  288.0   3rd Qu.:   401   3rd Qu.:   345   3rd Qu.:99.07  
##  Max.   :10289.0   Max.   :188565   Max.   :384119   Max.   :99.82  
##    percblack       percamerindan        percasian        percother      
##  Min.   : 0.0000   Min.   : 0.05623   Min.   :0.0000   Min.   :0.00000  
##  1st Qu.: 0.1157   1st Qu.: 0.15793   1st Qu.:0.1737   1st Qu.:0.09102  
##  Median : 0.5390   Median : 0.21502   Median :0.2972   Median :0.17844  
##  Mean   : 2.6763   Mean   : 0.79894   Mean   :0.4872   Mean   :0.47906  
##  3rd Qu.: 2.6014   3rd Qu.: 0.38362   3rd Qu.:0.5212   3rd Qu.:0.48050  
##  Max.   :40.2100   Max.   :89.17738   Max.   :5.0705   Max.   :7.52427  
##    popadults          perchsd         percollege        percprof     
##  Min.   :   1287   Min.   :46.91   Min.   : 7.336   Min.   : 0.5203  
##  1st Qu.:  12271   1st Qu.:71.33   1st Qu.:14.114   1st Qu.: 2.9980  
##  Median :  22188   Median :74.25   Median :16.798   Median : 3.8142  
##  Mean   :  60973   Mean   :73.97   Mean   :18.273   Mean   : 4.4473  
##  3rd Qu.:  47541   3rd Qu.:77.20   3rd Qu.:20.550   3rd Qu.: 4.9493  
##  Max.   :3291995   Max.   :88.90   Max.   :48.079   Max.   :20.7913  
##  poppovertyknown   percpovertyknown percbelowpoverty percchildbelowpovert
```

```
##  Min.   :   1696   Min.   :80.90    Min.   : 2.180   Min.   : 1.919
##  1st Qu.:  18364   1st Qu.:96.89    1st Qu.: 9.199   1st Qu.:11.624
##  Median :  33788   Median :98.17    Median :11.822   Median :15.270
##  Mean   :  93642   Mean   :97.11    Mean   :12.511   Mean   :16.447
##  3rd Qu.:  72840   3rd Qu.:98.60    3rd Qu.:15.133   3rd Qu.:20.352
##  Max.   :5023523   Max.   :99.86    Max.   :48.691   Max.   :64.308
##  percadultpoverty percelderlypoverty    inmetro         category
##  Min.   : 1.938   Min.   : 3.547     Min.   :0.0000   Length:437
##  1st Qu.: 7.668   1st Qu.: 8.912     1st Qu.:0.0000   Class :character
##  Median :10.008   Median :10.869     Median :0.0000   Mode  :character
##  Mean   :10.919   Mean   :11.389     Mean   :0.3432
##  3rd Qu.:13.182   3rd Qu.:13.412     3rd Qu.:1.0000
##  Max.   :43.312   Max.   :31.162     Max.   :1.0000
```

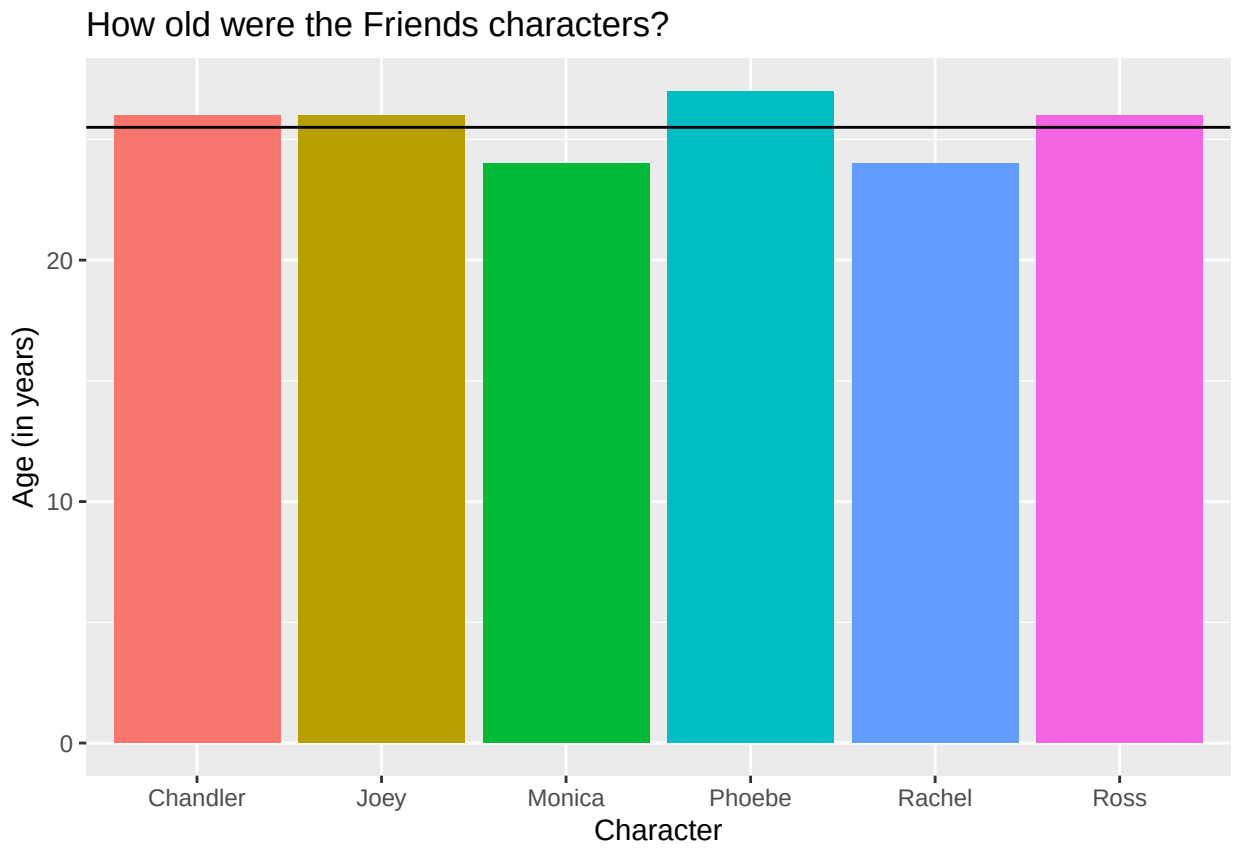
```
#building a histogram based on the total population by county
ggplot(data = midwest, mapping = aes(x = poptotal))+
  geom_histogram(bins = 100)+
  annotate("text", x = 4500000, y = 25, label = "Which County Is This?")+
  labs(title = "Histogram of Midwestern Counties By Population",
       x = "Population", y = "Number of Counties")
```



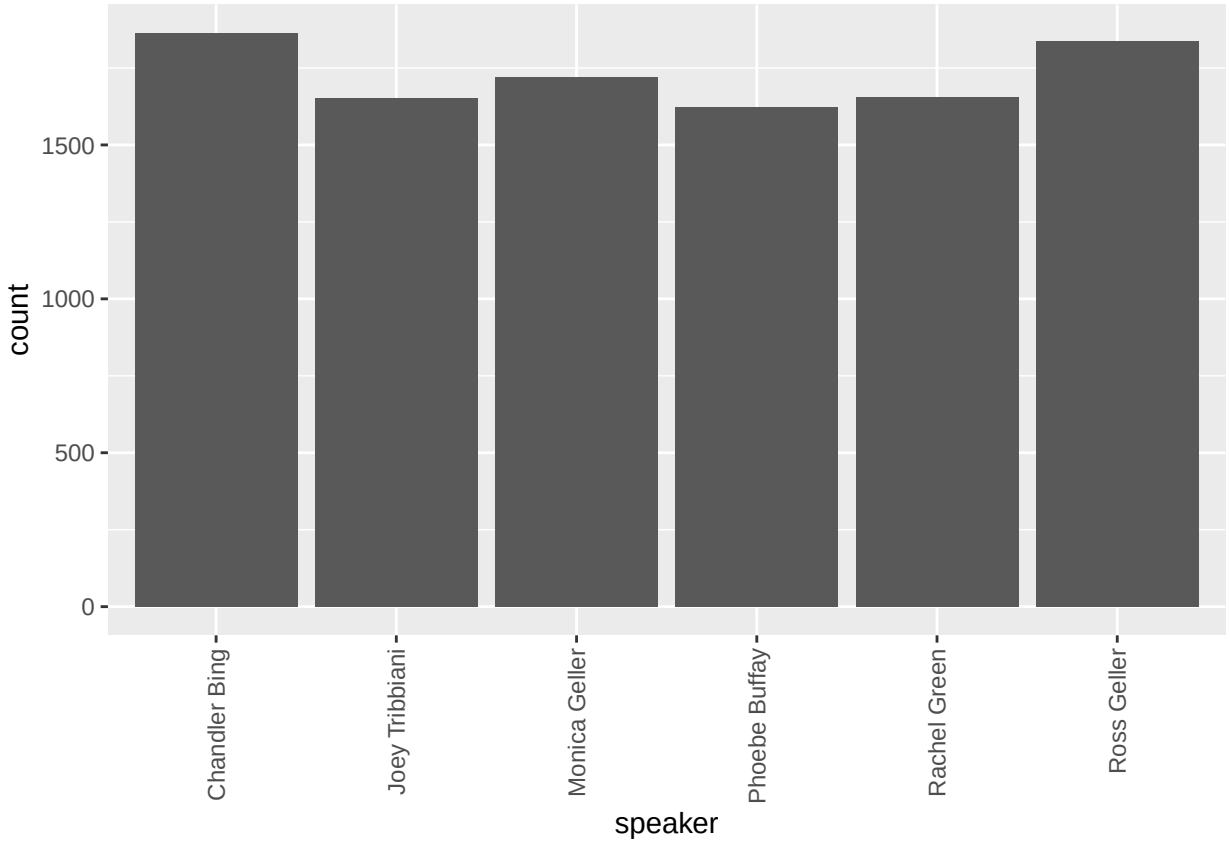

```
#build a basic dataset
ages <- tribble(
  ~speaker,      ~age,
  "Rachel",        24,
```

```
  "Phoebe",          27,
  "Ross",            26,
  "Monica",          24,
  "Chandler",        26,
  "Joey",            26
)

#create a new object that holds the mean of all the character's ages
c_avg <- mean(ages$age)

#create a bar chart that visualizes the character's age and mean
ggplot(data = ages) +
  geom_bar(mapping = aes(x = speaker,
                         y = age,
                         fill = speaker),
           stat = "identity")+
  geom_hline(aes(yintercept = c_avg))+
  labs(title = "How old were the Friends characters?",
       x = "Character", y = "Age (in years)")+
  guides(fill=FALSE)
```

```
## Warning: `guides(<scale> = FALSE)` is deprecated. Please use `guides(<scale> =
## "none")` instead.
```



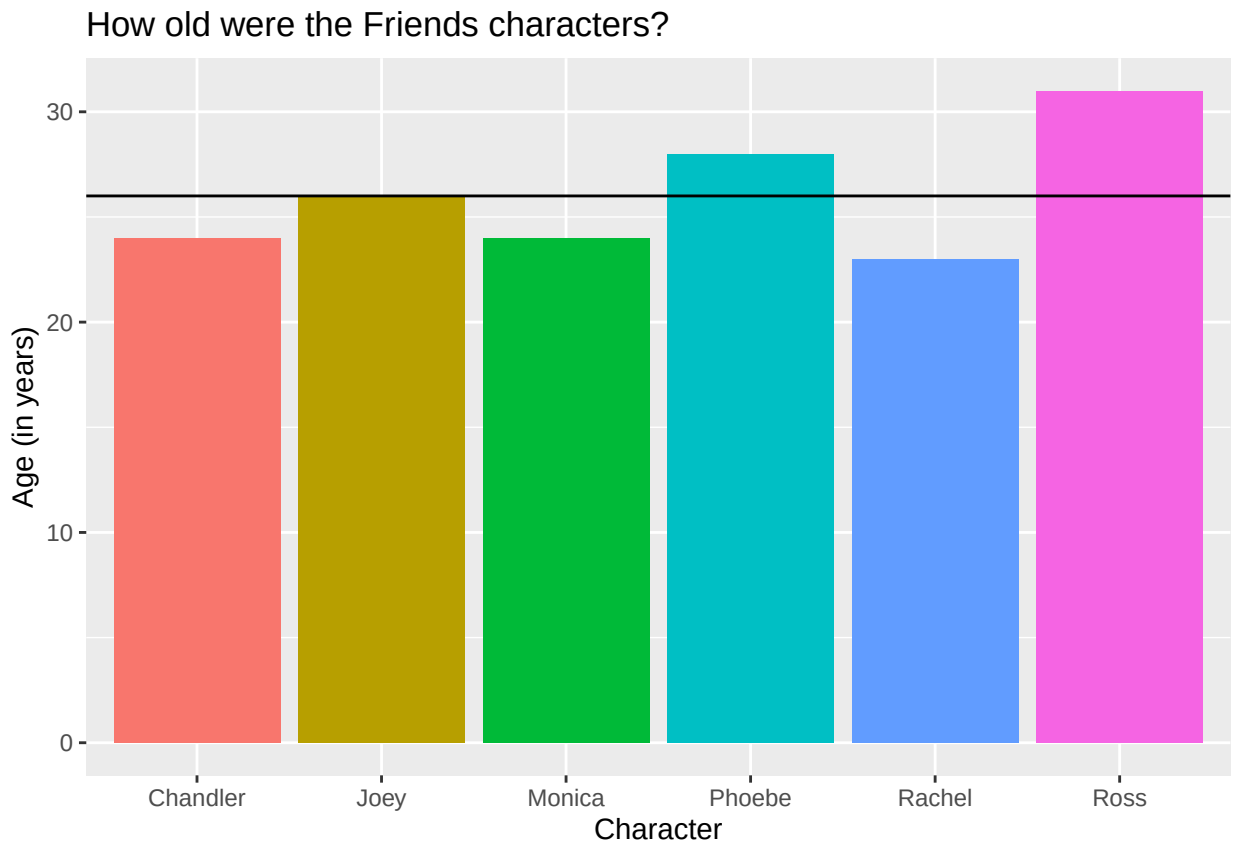
```r
#default stat transformation fro geom_bar() is stat_count()
ggplot(data = fulldata_main)+
  geom_bar(mapping = aes(x = speaker))+
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1))
```




```r
# no stat transformation
ages <- tribble(
  ~speaker,      ~age,
  "Rachel",         23,
  "Phoebe",         28,
  "Ross",           31,
  "Monica",         24,
  "Chandler",       24,
  "Joey",           26
)

c_avg <- mean(ages$age)

ggplot(data = ages) +
  geom_bar(mapping = aes(x = speaker,
                         y = age,
                         fill = speaker),
           stat = "identity")+
  geom_hline(aes(yintercept = c_avg))+
  labs(title = "How old were the Friends characters?",
```

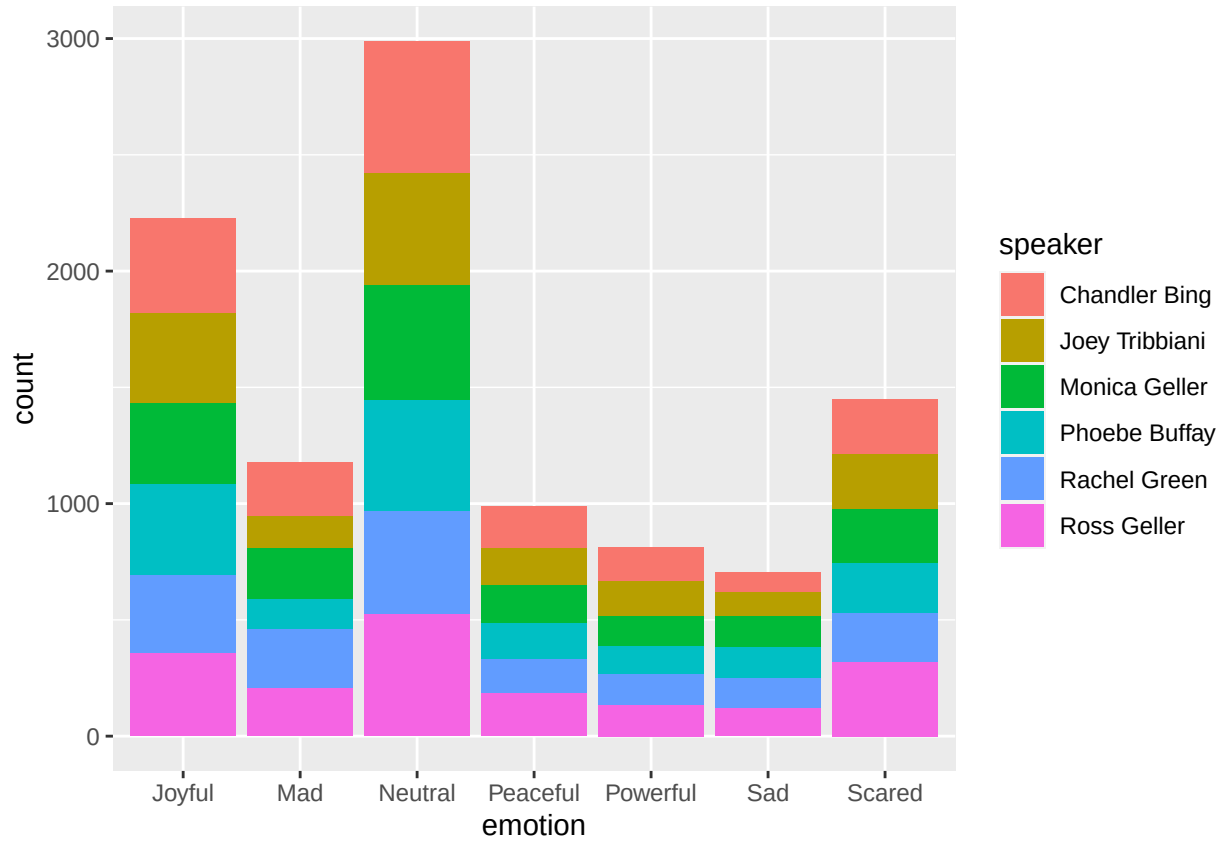
```
      x = "Character", y = "Age (in years)")+
  guides(fill=FALSE)
```

```
## Warning: `guides(<scale> = FALSE)` is deprecated. Please use `guides(<scale> =
## "none")` instead.
```

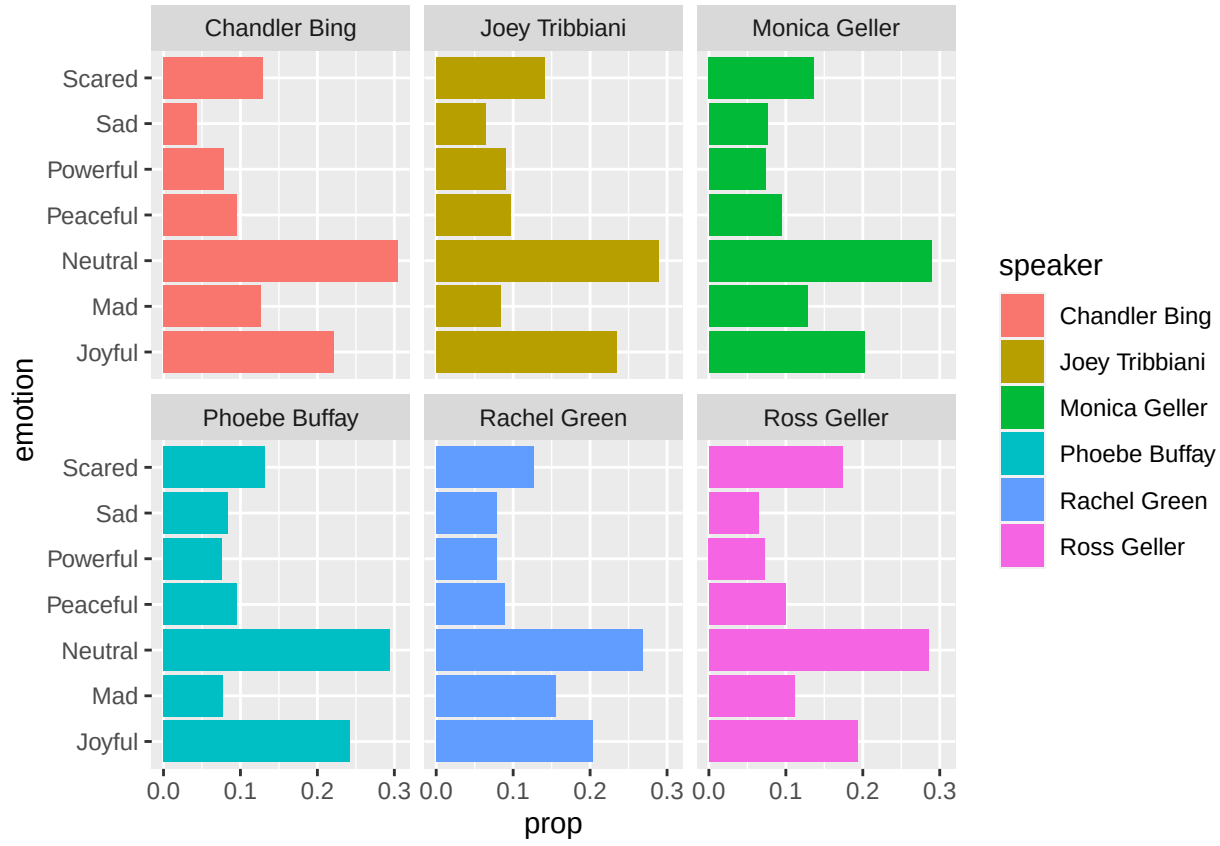


EDA

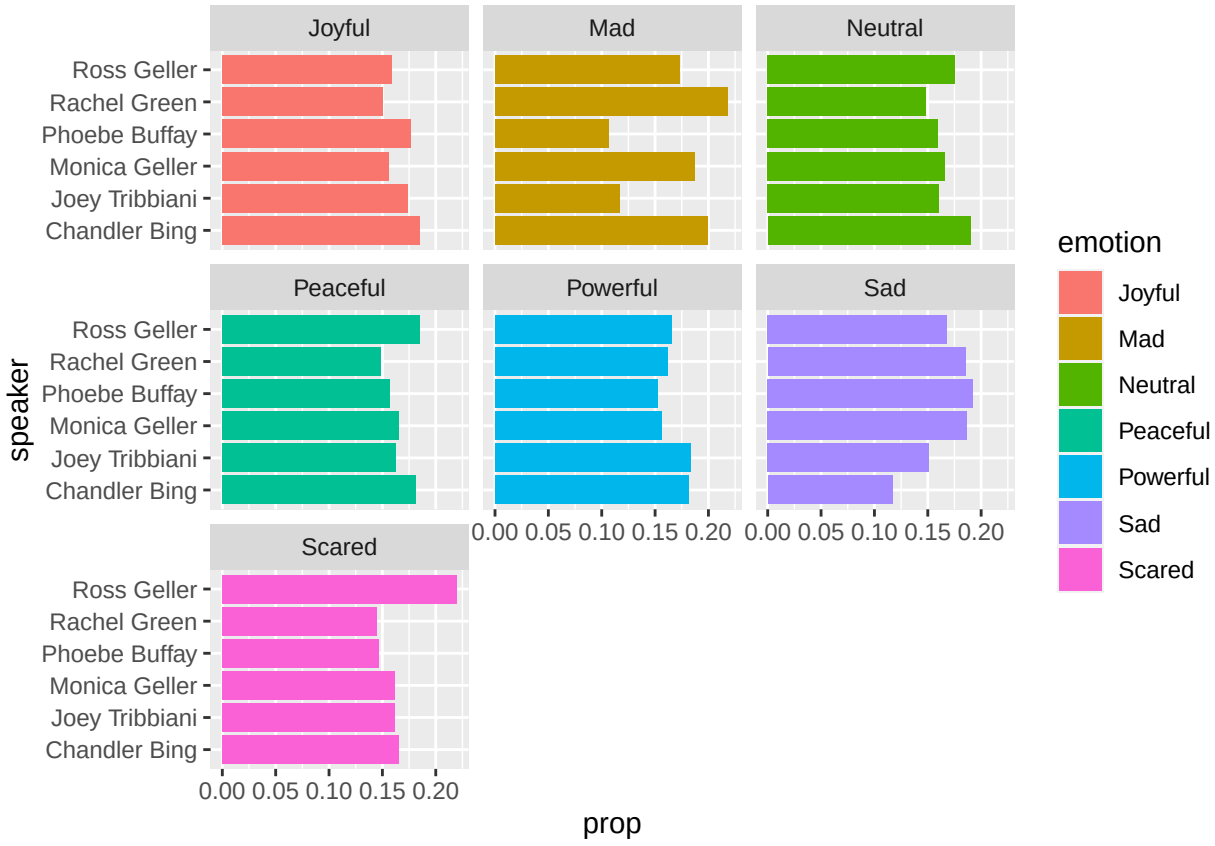
```
#stacked bar graph
ggplot(data = fulldata_main, mapping = aes(x = emotion, fill = speaker)) +
  geom_bar()
```




```
#emotion vs speaker faceted
ggplot(data = fulldata_main,
       mapping = aes(x = emotion, y = stat(prop),
                     group = 1, fill = speaker))+
  geom_bar()+
  facet_wrap(~speaker)+
  coord_flip()
```

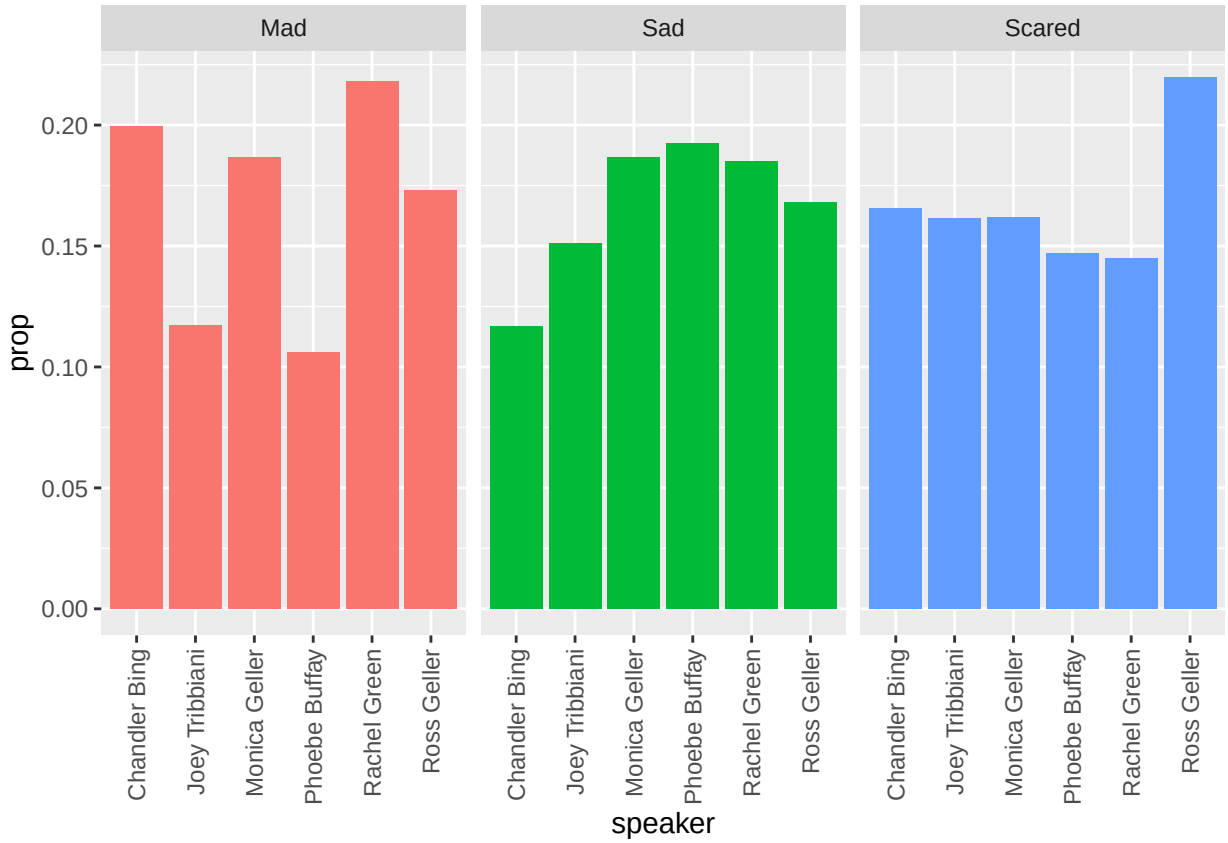



```
#speaker vs emotion faceted
ggplot(data = fulldata_main,
       mapping = aes(x = speaker, y = stat(prop),
                     group = 1, fill = emotion))+
  geom_bar()+
  facet_wrap(~emotion)+
  coord_flip()
```




```
#create list of emotions that we want to highlight
select_emotions <- c("Sad","Mad","Scared")

#visualize selected emotions and clean up graph
fulldata_main%>%
  filter(emotion%in%select_emotions)%>%
  ggplot(mapping = aes(x = speaker, y = stat(prop),
                       group = 1, fill = emotion))+
  geom_bar()+
  facet_wrap(~emotion)+
  guides(fill = FALSE)+
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1))
```

```
## Warning: `guides(<scale> = FALSE)` is deprecated. Please use `guides(<scale> =
## "none")` instead.
```



Data Transformations with dplyr

```
#simply but verbose way of filtering data and dropping NAs
full_data_new <- filter(full_data, speaker %in% target)
full_data_new_na <- drop_na(full_data_new)

#better way of filtering with pipes
full_data_new_better <- full_data%>%
  filter(speaker %in% target)%>%
  drop_na()

#creating new aggregated (n) variables
emotion_by_speaker <- full_data_new_na %>%
  group_by(speaker,emotion)%>%
  summarize(N=n())%>%
  mutate(freq = N / sum(N), pct = round((freq*100),1))
```

```
## `summarise()` has grouped output by 'speaker'. You can override using the `.groups` argument.
```

```
speaker_by_emotion <- full_data_new_na %>%
  group_by(emotion,speaker)%>%
  summarize(N=n())%>%
  mutate(freq = N / sum(N), pct = round((freq*100),1))
```

```
## `summarise()` has grouped output by 'emotion'. You can override using the `.groups` argument.
```

##Organ Donor EDA

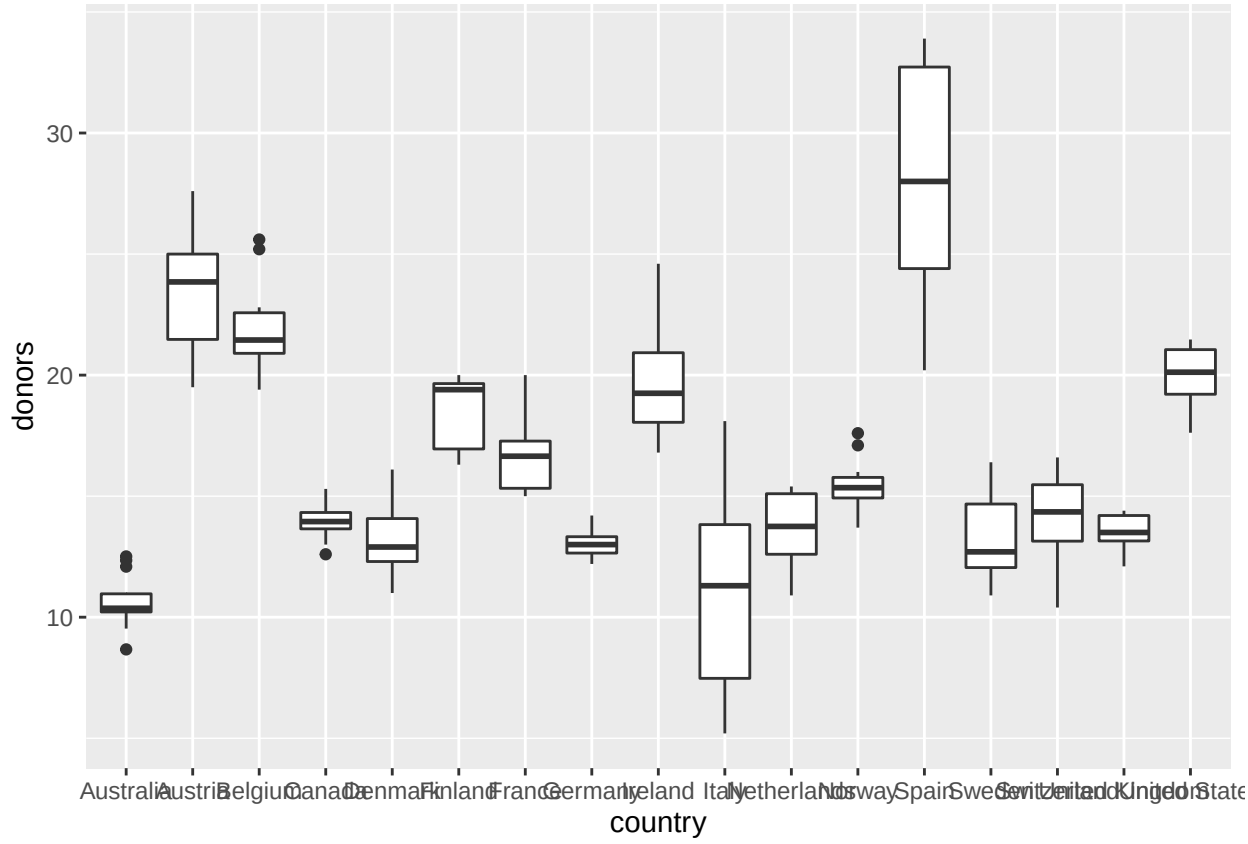
How does organ donation differ across countries? Does the type of government impact organ donation trends?

```
str(organdata)
```

```
## tibble [238 x 21] (S3: tbl_df/tbl/data.frame)
##  $ country         : chr [1:238] "Australia" "Australia" "Australia" "Australia" ...
##  $ year            : Date[1:238], format: NA "1991-01-01" ...
##  $ donors          : num [1:238] NA 12.1 12.3 12.5 10.2 ...
##  $ pop             : int [1:238] 17065 17284 17495 17667 17855 18072 18311 18518 18711 18926 ...
##  $ pop_dens        : num [1:238] 0.22 0.223 0.226 0.228 0.231 ...
##  $ gdp             : int [1:238] 16774 17171 17914 18883 19849 21079 21923 22961 24148 25445 ...
##  $ gdp_lag         : int [1:238] 16591 16774 17171 17914 18883 19849 21079 21923 22961 24148 ...
##  $ health          : num [1:238] 1300 1379 1455 1540 1626 ...
##  $ health_lag      : num [1:238] 1224 1300 1379 1455 1540 ...
##  $ pubhealth       : num [1:238] 4.8 5.4 5.4 5.4 5.4 5.5 5.6 5.7 5.9 6.1 ...
##  $ roads           : num [1:238] 137 122 113 111 108 ...
##  $ cerebvas        : int [1:238] 682 647 630 611 631 592 576 525 516 493 ...
##  $ assault         : int [1:238] 21 19 17 18 17 16 17 17 16 15 ...
##  $ external        : int [1:238] 444 425 406 376 387 371 395 385 410 409 ...
##  $ txp_pop         : num [1:238] 0.938 0.926 0.915 0.906 0.896 ...
##  $ world           : chr [1:238] "Liberal" "Liberal" "Liberal" "Liberal" ...
##  $ opt             : chr [1:238] "In" "In" "In" "In" ...
##  $ consent_law     : chr [1:238] "Informed" "Informed" "Informed" "Informed" ...
##  $ consent_practice: chr [1:238] "Informed" "Informed" "Informed" "Informed" ...
##  $ consistent      : chr [1:238] "Yes" "Yes" "Yes" "Yes" ...
##  $ ccode           : chr [1:238] "Oz" "Oz" "Oz" "Oz" ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   country = col_character(),
##   ..   year = col_integer(),
##   ..   donors = col_double(),
##   ..   pop = col_integer(),
##   ..   pop.dens = col_double(),
##   ..   gdp = col_integer(),
##   ..   gdp.lag = col_integer(),
##   ..   health = col_double(),
##   ..   health.lag = col_double(),
##   ..   pubhealth = col_double(),
##   ..   roads = col_double(),
##   ..   cerebvas = col_integer(),
##   ..   assault = col_integer(),
##   ..   external = col_integer(),
##   ..   txp.pop = col_double(),
##   ..   world = col_character(),
##   ..   opt = col_character(),
##   ..   consent.law = col_character(),
##   ..   consent.practice = col_character(),
##   ..   consistent = col_character(),
##   ..   ccode = col_character()
##   .. )
```

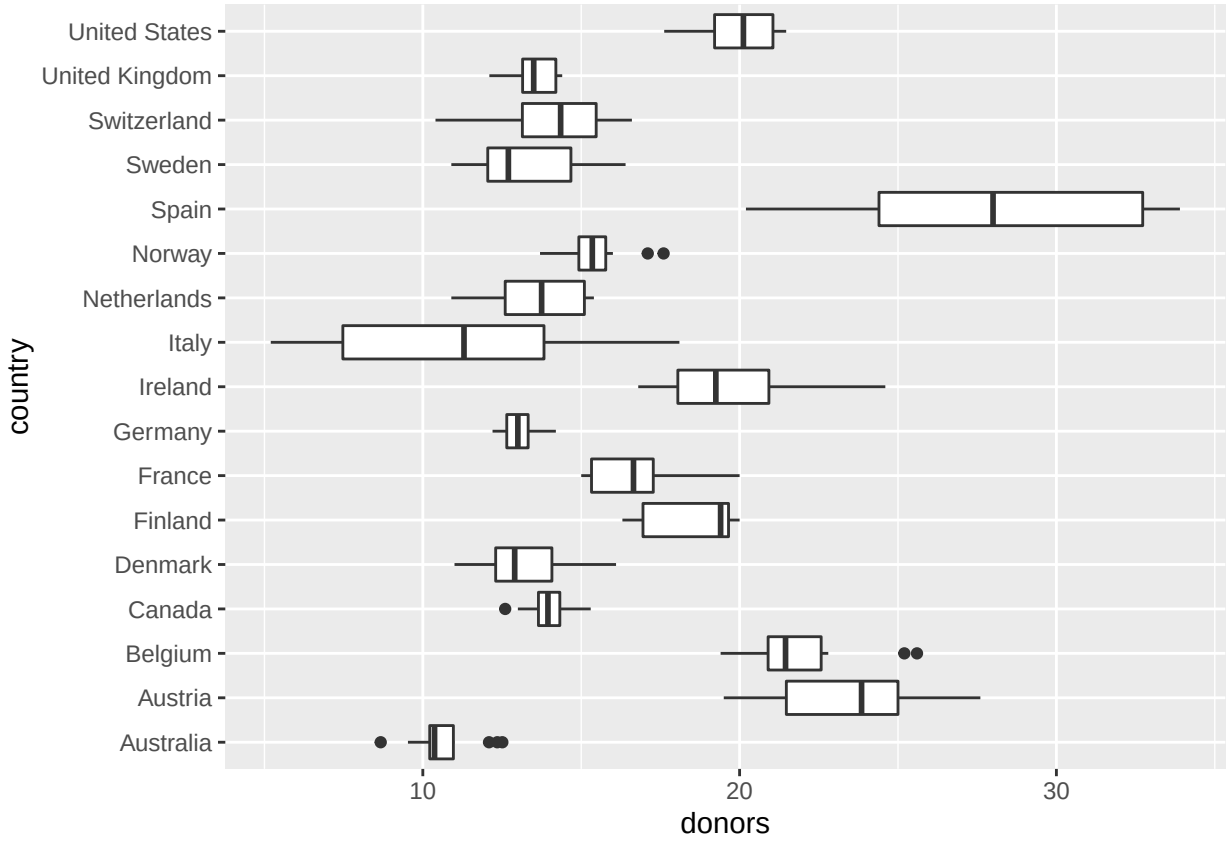
```
ggplot(data = organdata, mapping = aes(x = country, y = donors))+
  geom_boxplot()
```

```
## Warning: Removed 34 rows containing non-finite values (stat_boxplot).
```



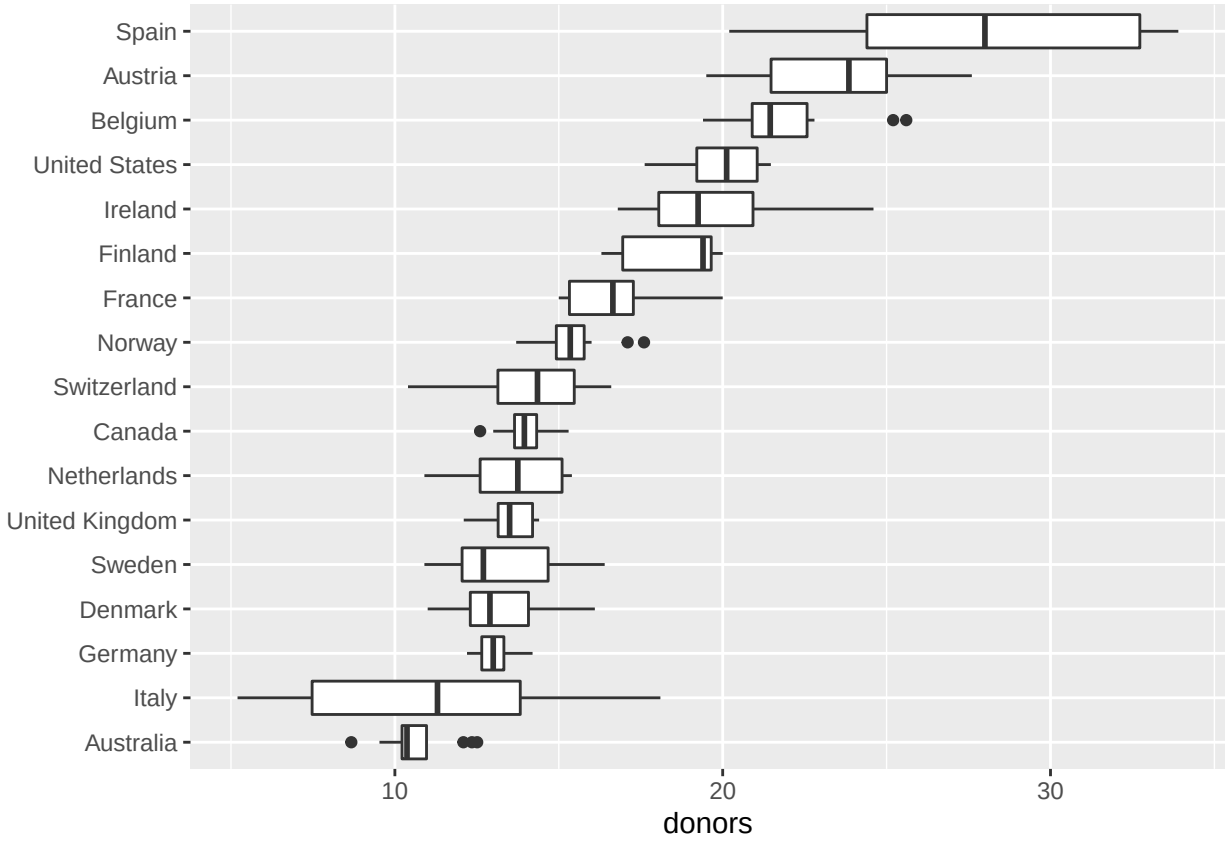

```
ggplot(data = organdata, mapping = aes(x = country, y = donors))+
  geom_boxplot() +
  coord_flip()
```

```
## Warning: Removed 34 rows containing non-finite values (stat_boxplot).
```



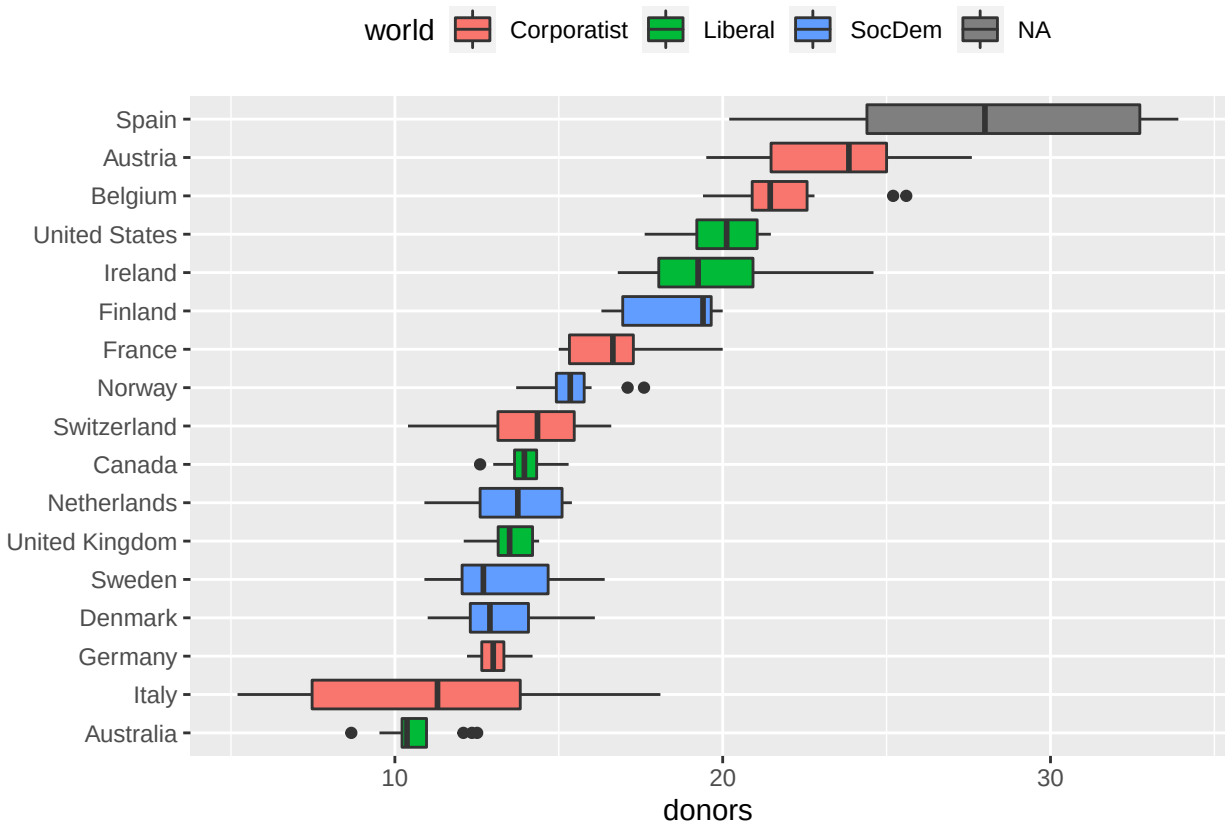

```
ggplot(data = organdata, mapping = aes(x = reorder(country, donors, na.rm=TRUE), y = donors))+
  geom_boxplot()+
  labs(x = NULL)+
  coord_flip()
```

```
## Warning: Removed 34 rows containing non-finite values (stat_boxplot).
```



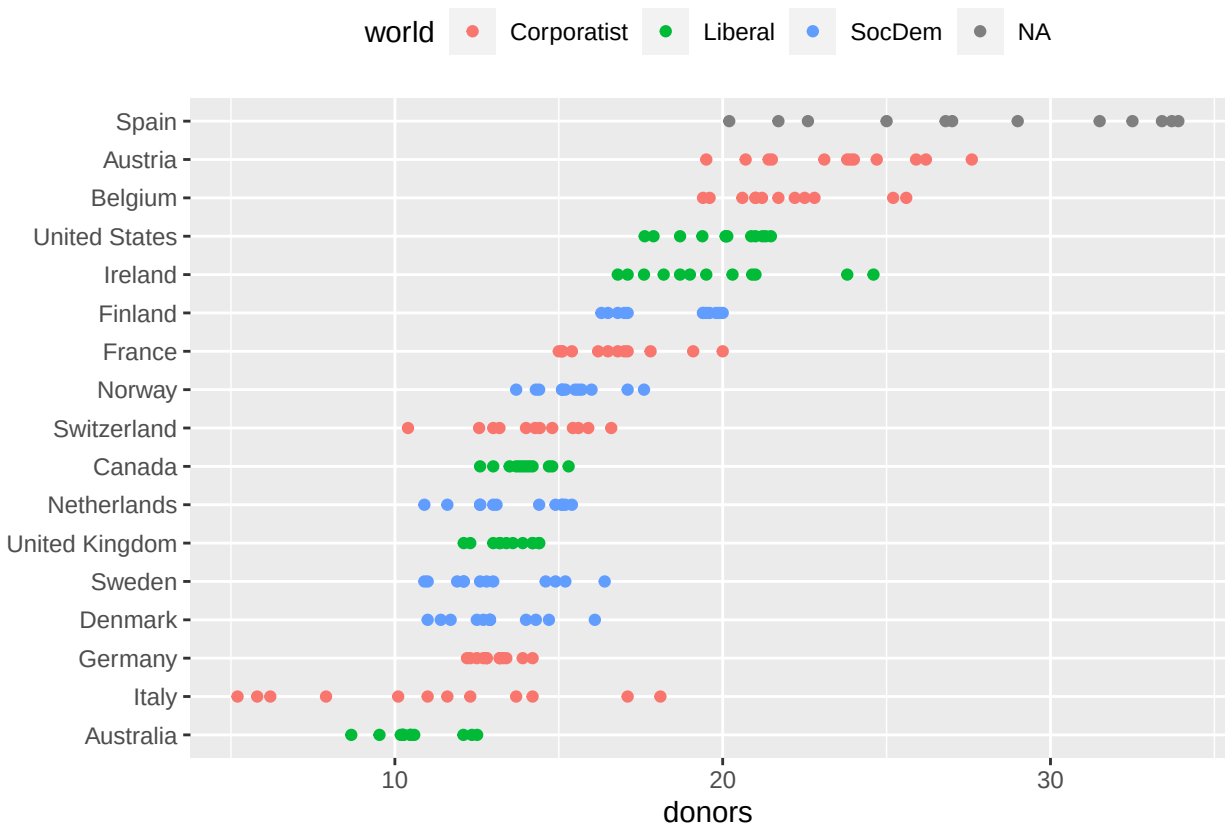

```
ggplot(data = organdata, mapping = aes(x = reorder(country, donors, na.rm=TRUE),
                                       y = donors,
                                       fill = world))+
  geom_boxplot()+
  labs(x = NULL)+
  coord_flip()+
  theme(legend.position = "top")
```

```
## Warning: Removed 34 rows containing non-finite values (stat_boxplot).
```



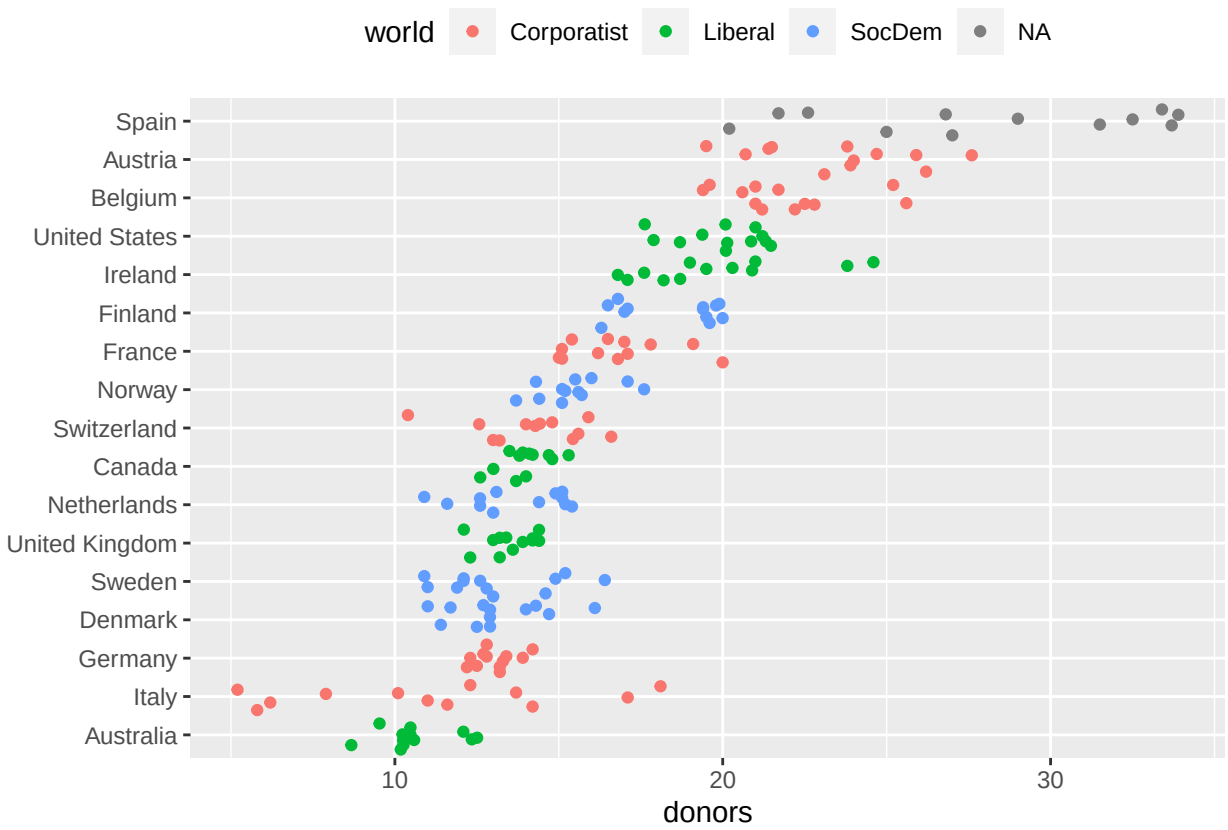

```
ggplot(data = organdata, mapping = aes(x = reorder(country, donors, na.rm=TRUE),
                                       y = donors,
                                       color = world))+
  geom_point()+
  labs(x = NULL)+
  coord_flip()+
  theme(legend.position = "top")
```

```
## Warning: Removed 34 rows containing missing values (geom_point).
```



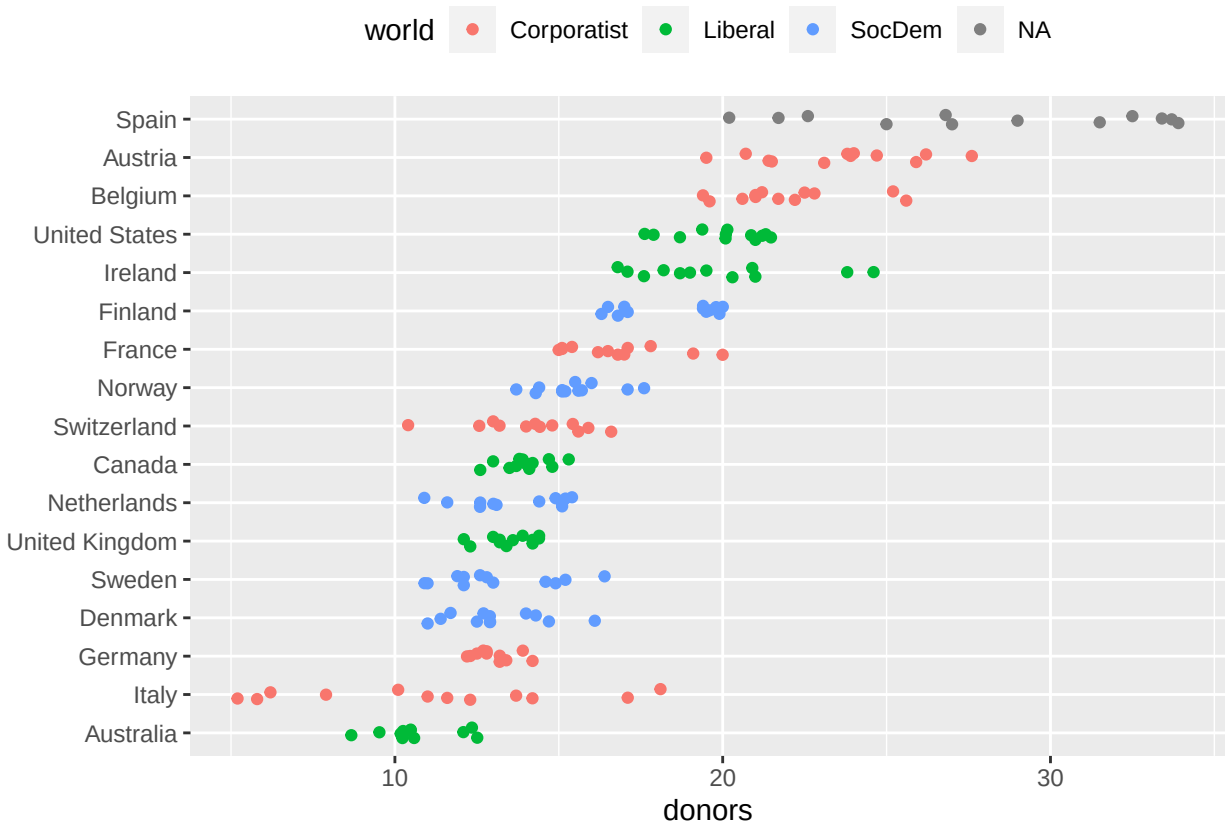

```
ggplot(data = organdata, mapping = aes(x = reorder(country, donors, na.rm=TRUE),
                                       y = donors,
                                       color = world))+
  geom_jitter()+
  labs(x = NULL)+
  coord_flip()+
  theme(legend.position = "top")
```

```
## Warning: Removed 34 rows containing missing values (geom_point).
```




```
ggplot(data = organdata, mapping = aes(x = reorder(country, donors, na.rm=TRUE),
                                       y = donors,
                                       color = world))+
  geom_jitter(position = position_jitter(width = 0.15))+
  labs(x = NULL)+
  coord_flip()+
  theme(legend.position = "top")
```

```
## Warning: Removed 34 rows containing missing values (geom_point).
```




```
#verbose way (not comprehensive - just example of first 3)
by_country_verbose <-  organdata %>% group_by(consent_law, country)%>%
  summarise(donors_mean = mean(donors, na.rm = TRUE),
            donors_sd = sd(donors, na.rm = TRUE),
            gdp_mean = mean(gdp, na.rm = TRUE))
```

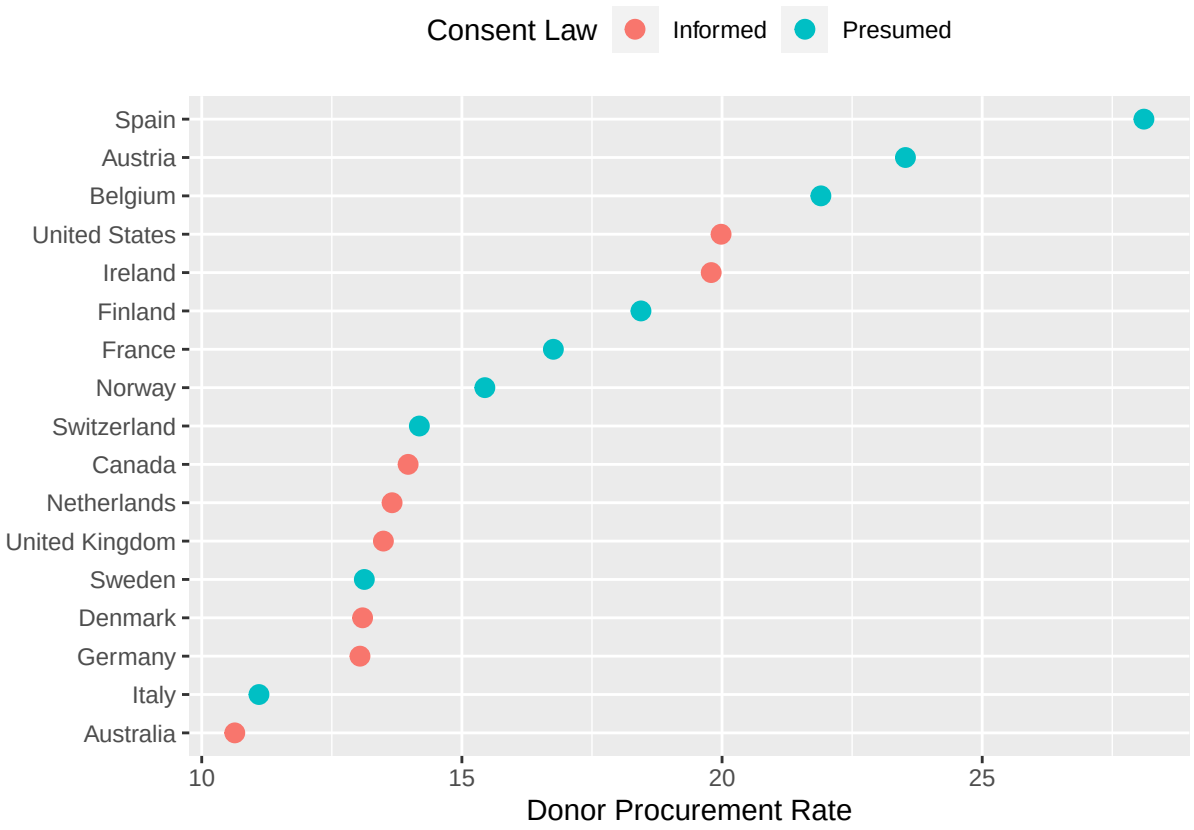
```
## `summarise()` has grouped output by 'consent_law'. You can override using the `.groups` argument.
```

```
#better way but requires summarize_if()
by_country <- organdata %>%
  group_by(consent_law, country) %>%
  summarize_if(is.numeric,
               list(~ mean(., na.rm = TRUE),
                    ~ sd(., na.rm = TRUE))) %>%
  ungroup()
by_country
```

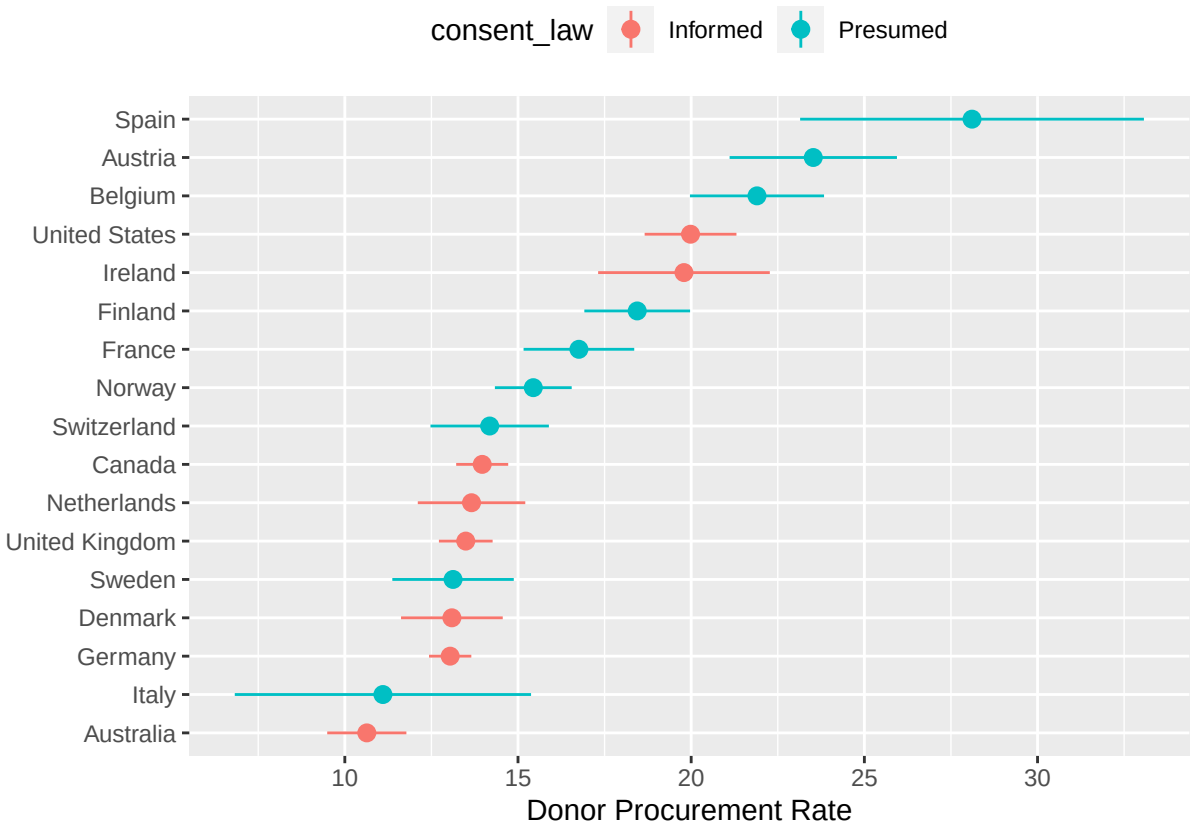
```
## # A tibble: 17 x 28
##    consent_law country  donors_mean pop_mean pop_dens_mean gdp_mean gdp_lag_mean
##    <chr>       <chr>          <dbl>    <dbl>         <dbl>    <dbl>        <dbl>
##  1 Informed    Austral~        10.6   18318.         0.237   22179.       21779.
##  2 Informed    Canada          14.0   29608.         0.297   23711.       23353.
##  3 Informed    Denmark         13.1    5257.        12.2     23722.       23275
##  4 Informed    Germany         13.0   80255.        22.5     22163.       21938.
```

```
##  5 Informed    Ireland         19.8    3674.         5.23    20824.       20154.
##  6 Informed    Netherl~        13.7   15548.        37.4     23013.       22554.
##  7 Informed    United ~        13.5   58187.        24.0     21359.       20962.
##  8 Informed    United ~        20.0  269330.         2.80    29212.       28699.
##  9 Presumed    Austria         23.5    7927.         9.45    23876.       23415.
## 10 Presumed    Belgium         21.9   10153.        30.7     22500.       22096.
## 11 Presumed    Finland         18.4    5112.         1.51    21019.       20763
## 12 Presumed    France          16.8   58056.        10.5     22603.       22211.
## 13 Presumed    Italy           11.1   57360.        19.0     21554.       21195.
## 14 Presumed    Norway          15.4    4386.         1.35    26448.       25769.
## 15 Presumed    Spain           28.1   39666.         7.84    16933        16584.
## 16 Presumed    Sweden          13.1    8789.         1.95    22415.       22094
## 17 Presumed    Switzer~        14.2    7037.        17.0     27233        26931.
## # ... with 21 more variables: health_mean <dbl>, health_lag_mean <dbl>,
## #   pubhealth_mean <dbl>, roads_mean <dbl>, cerebvas_mean <dbl>,
## #   assault_mean <dbl>, external_mean <dbl>, txp_pop_mean <dbl>,
## #   donors_sd <dbl>, pop_sd <dbl>, pop_dens_sd <dbl>, gdp_sd <dbl>,
## #   gdp_lag_sd <dbl>, health_sd <dbl>, health_lag_sd <dbl>, pubhealth_sd <dbl>,
## #   roads_sd <dbl>, cerebvas_sd <dbl>, assault_sd <dbl>, external_sd <dbl>,
## #   txp_pop_sd <dbl>
```

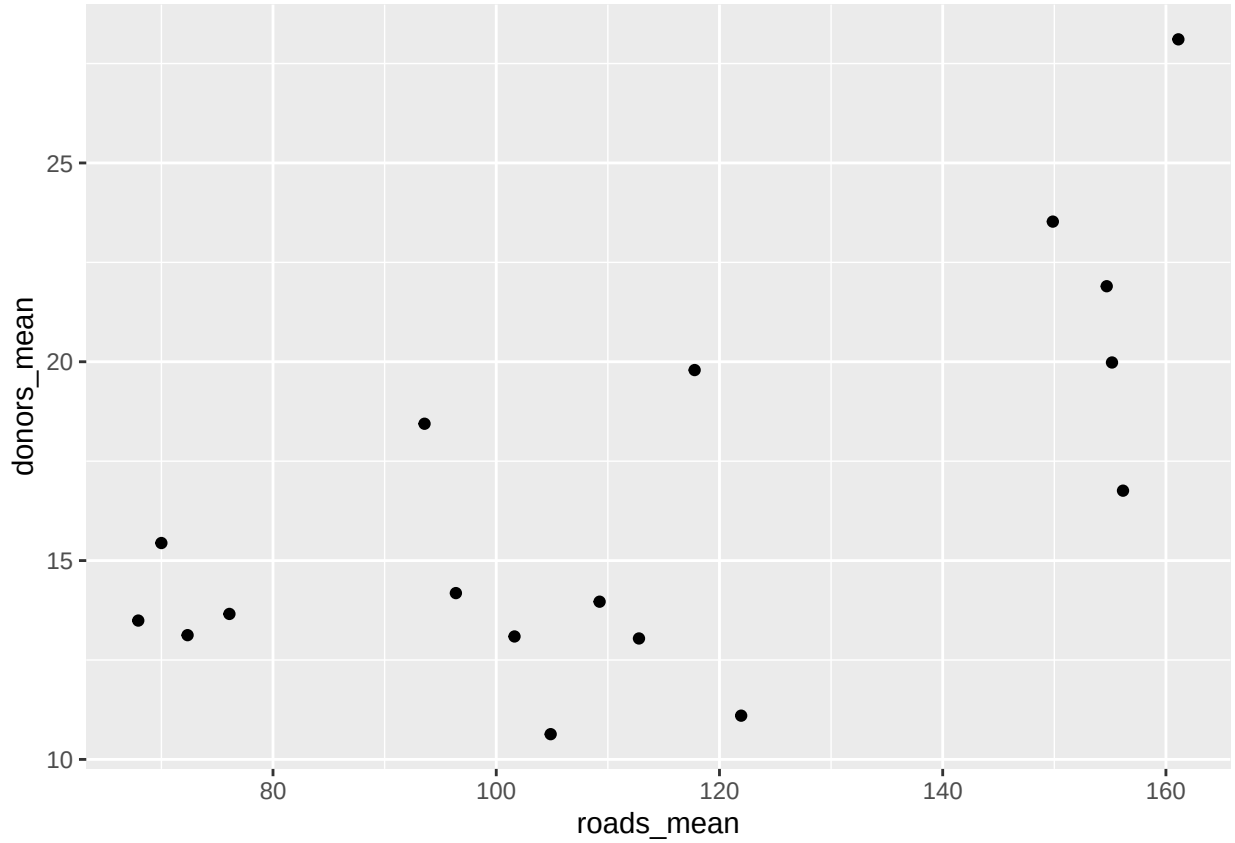
```
ggplot(data = by_country,
       mapping = aes(x = donors_mean,
                     y = reorder(country, donors_mean),
                     color = consent_law))+
  geom_point(size = 3) +
  labs(x = "Donor Procurement Rate",
       y = "",
       color = "Consent Law")+
  theme(legend.position = "top")
```



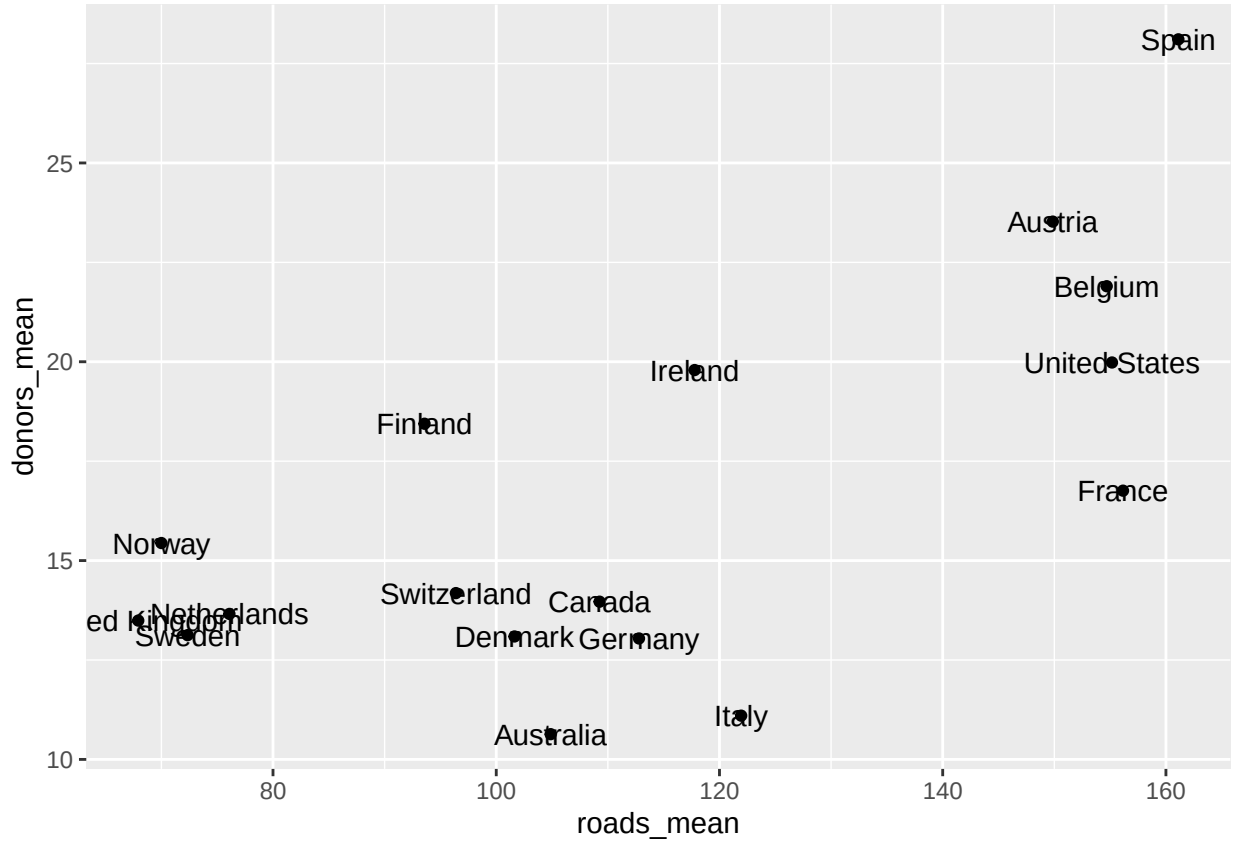

```
ggplot(data = by_country, mapping = aes(x = reorder(country, donors_mean),y = donors_mean))+
        geom_pointrange(mapping = aes(ymin = donors_mean - donors_sd,
                                      ymax = donors_mean + donors_sd,
                                      color = consent_law))+
        labs(x = "", y="Donor Procurement Rate")+
        coord_flip()+theme(legend.position = "top")
```



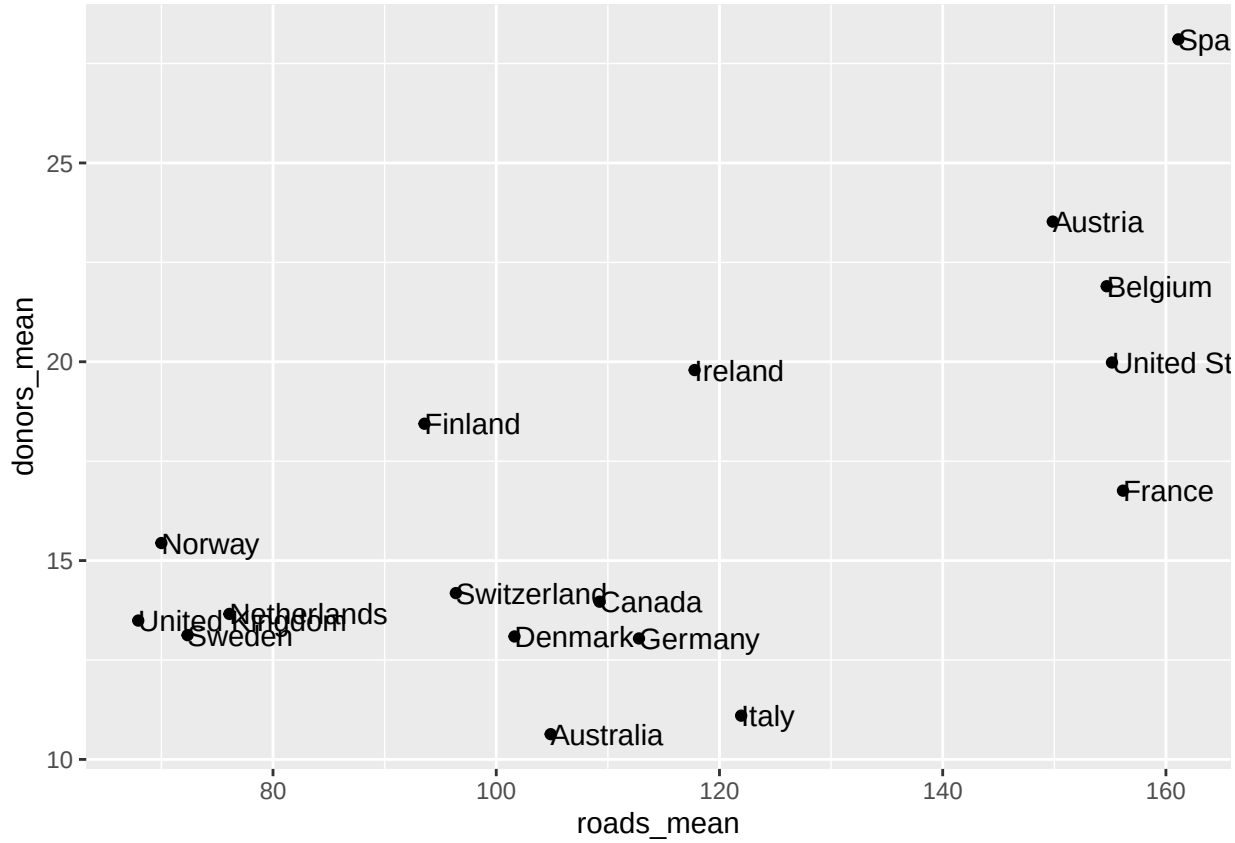

```
ggplot(data = by_country,
       mapping = aes(x = roads_mean, y = donors_mean))+
  geom_point()
```



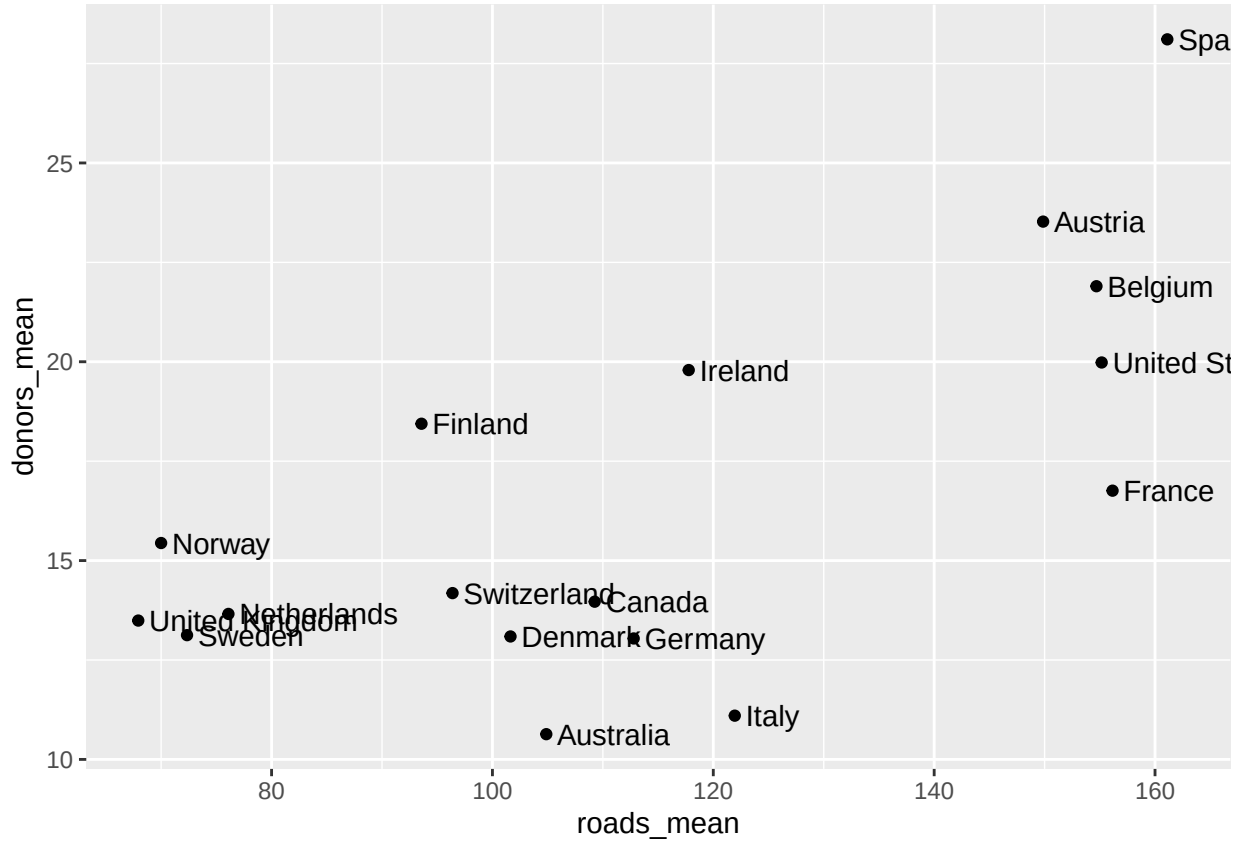

```
ggplot(data = by_country,
       mapping = aes(x = roads_mean, y = donors_mean))+
  geom_point() +
  geom_text(mapping = aes(label = country))
```



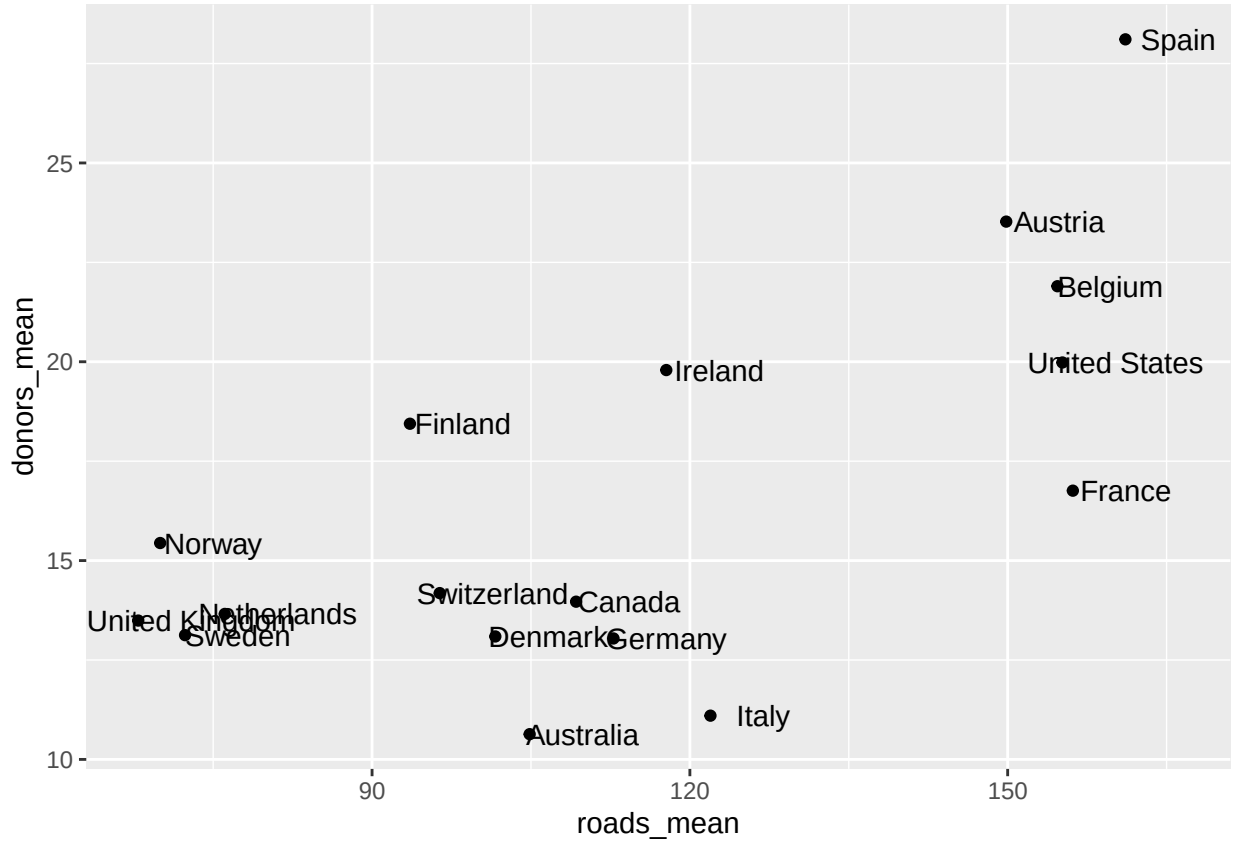

```
ggplot(data = by_country,
       mapping = aes(x = roads_mean, y = donors_mean))+
  geom_point() +
  geom_text(mapping = aes(label = country), hjust = 0)
```




```
ggplot(data = by_country,
       mapping = aes(x = roads_mean, y = donors_mean))+
  geom_point() +
  geom_text(mapping = aes(x = roads_mean + 1, label = country), hjust = 0)
```




```
ggplot(data = by_country,
       mapping = aes(x = roads_mean, y = donors_mean))+
  geom_point() +
  geom_text(mapping = aes(label = country), nudge_x = 5)
```



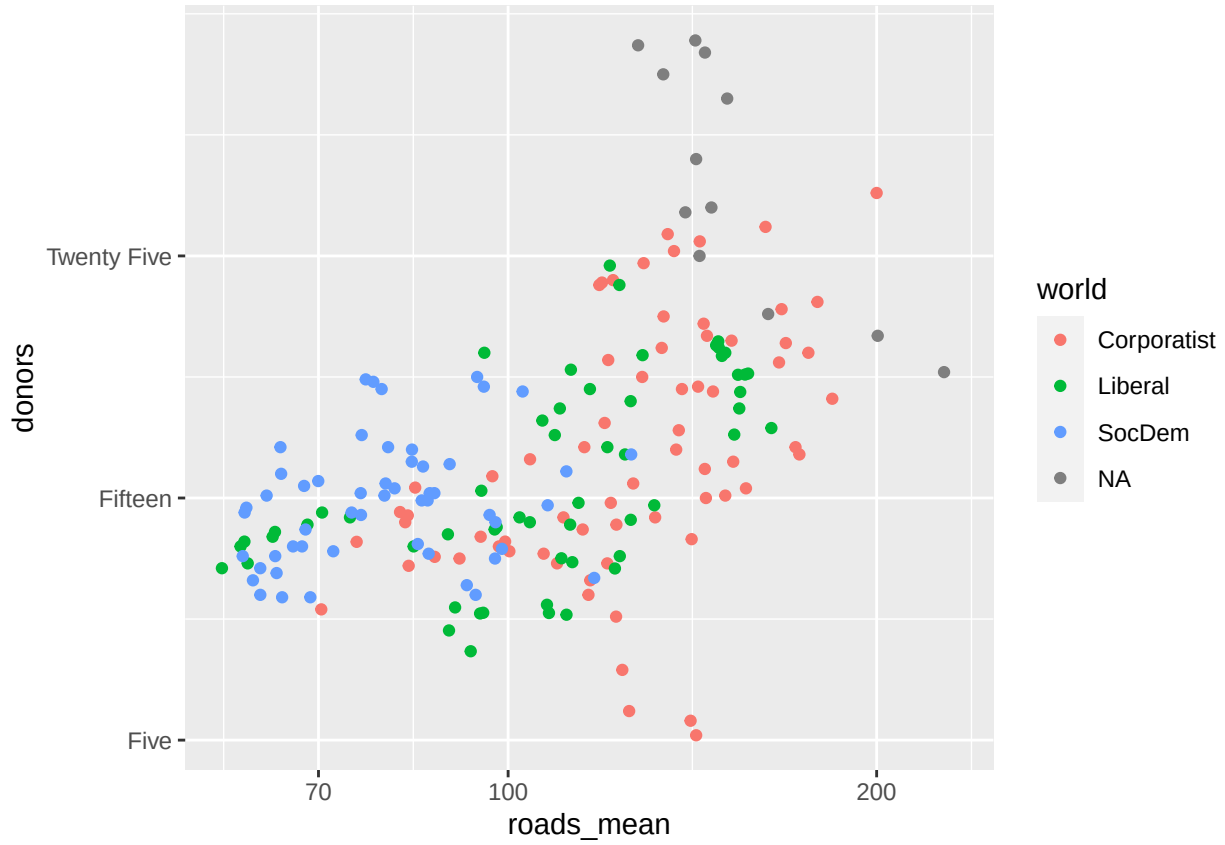

```
by_country_world <- organdata %>%
  group_by(world, country, donors) %>%
  summarize_if(is.numeric,
               list(~ mean(., na.rm = TRUE),
                    ~ sd(., na.rm = TRUE))) %>%
  ungroup()
by_country_world
```

```
## # A tibble: 206 x 27
##    world       country donors pop_mean pop_dens_mean gdp_mean gdp_lag_mean health_mean
##    <chr>       <chr>    <dbl>    <dbl>         <dbl>    <dbl>        <dbl>       <dbl>
##  1 Corporatist Austria   19.5     7968          9.50    24364        23798        1848
##  2 Corporatist Austria   20.7     7977          9.51    25423        24364        1953
##  3 Corporatist Austria   21.4     7936          9.46    21940        21119        1739
##  4 Corporatist Austria   21.5     7948          9.48    22817        21940        1865
##  5 Corporatist Austria   23.1     7841          9.35    20601        19860        1551
##  6 Corporatist Austria   23.8     8053          9.60    28842        28457        2220
##  7 Corporatist Austria   23.9     8030          9.58    28457        27738        2174
##  8 Corporatist Austria   24       8012          9.55    27738        26513        2147
##  9 Corporatist Austria   24.7     7959          9.49    23798        22817        1986
## 10 Corporatist Austria   25.9     7992          9.53    26513        25423        2069
## # ... with 196 more rows, and 19 more variables: health_lag_mean <dbl>,
## #   pubhealth_mean <dbl>, roads_mean <dbl>, cerebvas_mean <dbl>,
## #   assault_mean <dbl>, external_mean <dbl>, txp_pop_mean <dbl>, pop_sd <dbl>,
## #   pop_dens_sd <dbl>, gdp_sd <dbl>, gdp_lag_sd <dbl>, health_sd <dbl>,
```

```
## #   health_lag_sd <dbl>, pubhealth_sd <dbl>, roads_sd <dbl>, cerebvas_sd <dbl>,
## #   assault_sd <dbl>, external_sd <dbl>, txp_pop_sd <dbl>
```

```
d <- ggplot(data = by_country_world,
            mapping = aes(x = roads_mean,
                          y = donors, color = world))
d + geom_point() +
  scale_x_log10() + scale_y_continuous(breaks = c(5, 15, 25),
                                       labels = c("Five", "Fifteen", "Twenty Five"))
```

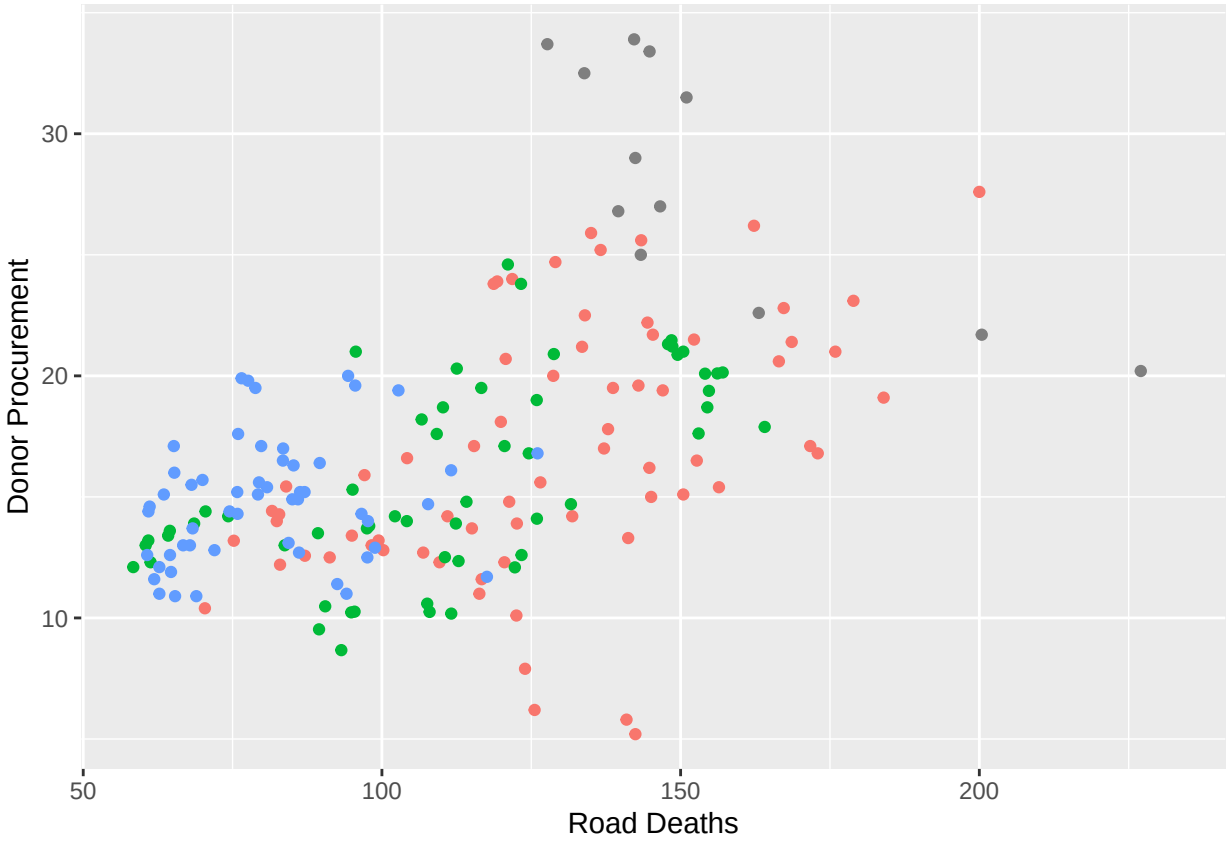
```
## Warning: Removed 17 rows containing missing values (geom_point).
```




```
e <- ggplot(data = by_country_world, mapping = aes(x = roads_mean, y = donors, color = world))
e + geom_point() + scale_color_discrete(labels = c("Corporatist", "Liberal","Social Democratic", "Uncla
  labs(x = "Road Deaths", y = "Donor Procurement",color = "Welfare State")+
  guides(color = FALSE)
```

```
## Warning: `guides(<scale> = FALSE)` is deprecated. Please use `guides(<scale> =
## "none")` instead.
```

```
## Warning: Removed 17 rows containing missing values (geom_point).
```


Donor Procurement
30
20
10
50
100
150
200
Road Deaths