Jeffrey_hw5_exercise1

Jack Jeffrey

HW 5

load tidyverse

```
library(tidyverse)
```

```
-- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
v dplyr     1.1.4     v readr     2.1.5
v forcats   1.0.0     v stringr   1.5.1
v ggplot2   3.5.1     v tibble    3.2.1
v lubridate 1.9.3     v tidyr     1.3.1
v purrr     1.0.2
-- Conflicts ------------------------------------------ tidyverse_conflicts() --
x dplyr::filter() masks stats::filter()
x dplyr::lag()    masks stats::lag()
i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to becor
```

load data set

```
world.data <- read.csv("/Users/jackjeffrey/Documents/Poli502_Jeffrey/Data/world.csv")
# view data
head(world.data)
```

```
            country       colony confidence decentralization dem_other
1       Afghanistan           UK         NA               NA      10.5
2           Albania Soviet Union   49.33593             0.74      63.0
3           Algeria       France   52.05573               NA      40.8
```

```
4           Andorra        Spain         NA               NA     100.0
5            Angola     Portugal         NA               NA      40.8
6 Antigua & Barbuda           UK         NA               NA      87.5
  dem_other5 democ_regime    district_size3 durable effectiveness      enpp_3
1        10%           No     single member       4      13.71158
2 Approx 60%          Yes                         3      35.46099 1-3 parties
3 Approx 40%           No 6 or more members       5      32.62411
4       100%          Yes                        NA      78.72340
5 Approx 40%           No                         3      19.14894
6 Approx 90%          Yes     single member      NA      59.81088 1-3 parties
         eu fhrate04_rev fhrate08_rev frac_eth frac_eth3 free_business
1 Not member          2.5            3   0.7693      High            NA
2 Not member            5            8   0.2204       Low          68.0
3 Not member          2.5            3   0.3394    Medium          71.2
4 Not member    Most free           12   0.7139      High            NA
5 Not member          2.5            3   0.7867      High          43.4
6 Not member            6           10   0.1643       Low            NA
  free_corrupt free_finance free_fiscal free_govspend free_invest free_labor
1           NA           NA          NA            NA          NA         NA
2           34           70        92.6          74.2          70       52.1
3           32           30        83.5          73.4          45       56.4
4           NA           NA          NA            NA          NA         NA
5           19           40        85.1          62.8          35       45.2
6           NA           NA          NA            NA          NA         NA
  free_monetary free_overall free_property free_trade gdp08 gdp_10_thou
1            NA           NA            NA         NA  30.6          NA
2          78.7         66.0            35       85.8  24.3      0.1535
3          77.2         56.9            30       70.7 276.0      0.1785
4            NA           NA            NA         NA    NA          NA
5          62.6         48.4            20       70.4 106.3      0.0857
6            NA           NA            NA         NA    NA      1.0449
  gdp_cap2 gdp_cap3 gdppcap08 gender_equal3 gini04 gini08   hi_gdp indy
1                          NA             NA     NA          1919
2      Low   Middle      7715           28.2   31.1  Low GDP 1991
3      Low   Middle      8033           35.3   35.3  Low GDP 1962
4                          NA             NA     NA          1278
5      Low   Middle      5899             NA     NA  Low GDP 1975
6     High     High        NA             NA     NA High GDP 1981
        oecd  old2006  old2003     pmat12_3    pop03 pop08    pop08_3
1 Not member       NA       NA                   NA  27.4 >=16.8 mil
2 Not member 8.479821 7.278363 Low post-mat  3169064   3.1  <=4.3 mil
3 Not member 4.578136 4.045199              31832610  34.4 >=16.8 mil
4 Not member       NA       NA                66000    NA
```

```
5 Not member 2.450295 2.930542              13522110  18.0 >=16.8 mil
6 Not member       NA 8.186610                 78580    NA
          popcat3 pr_sys protact3        regime_type3      region sources
1 Moderate (1-29m)     No                 Dictatorship Middle East      NA
2 Moderate (1-29m)     No Moderate Parliamentary democ  C&E Europe      NA
3 Moderate (1-29m)    Yes                 Dictatorship      Africa      NA
4 Small (under 1m)     No          Parliamentary democ   W. Europe      NA
5 Moderate (1-29m)    Yes                 Dictatorship      Africa      NA
6 Small (under 1m)     No          Parliamentary democ  S. America      NA
         typerel unions urban03 urban06 vi_rel3 votevap00s women05 women09
1         Muslim     NA      NA   23.28                 NA      NA    27.7
2         Muslim     NA 44.2390   46.14  20-50%      59.56     6.4    16.4
3         Muslim     NA 58.8302   63.94    >50%         NA      NA     7.7
4 Roman Catholic     NA 91.7404   90.28              20.95    14.3    35.7
5 Roman Catholic     NA 36.1806   53.96                 NA      NA    37.3
6     Protestant     NA 37.7566   39.60              76.34    10.5    10.5
  womyear       womyear2  yng2003  young06
1      NA                     NA       NA
2    1920 1944 or before 27.34834 26.35428
3    1962     After 1944 33.91887 28.94154
4    1973     After 1944       NA       NA
5    1975     After 1944 47.62524 46.32196
6    1951     After 1944 20.66509       NA
```

```
# summarize statistics
summary(world.data $ women09)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
   0.00    9.70   15.55   17.18   22.95   56.30      11
```

```
# view class
world.data $ women09 %>% class
```

```
[1] "numeric"
```

```
# check for NA's
is.na(world.data$women09)
```

```
  [1] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
 [13] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
```

```
 [25]  TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
 [37] FALSE FALSE  TRUE  TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
 [49] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE  TRUE FALSE FALSE
 [61] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE  TRUE FALSE FALSE
 [73] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
 [85] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
 [97] FALSE FALSE FALSE FALSE FALSE FALSE  TRUE FALSE FALSE FALSE FALSE FALSE
[109] FALSE FALSE FALSE FALSE FALSE  TRUE FALSE FALSE FALSE FALSE  TRUE FALSE
[121] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
[133] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
[145] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
[157] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE  TRUE FALSE FALSE
[169] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
[181] FALSE FALSE FALSE FALSE FALSE FALSE  TRUE FALSE  TRUE FALSE FALSE
```

```r
# view mean with NA's removed
world.data $ women09 %>% na.omit %>% mean
```

```
[1] 17.17722
```

```r
# view standard deviation with NA's removed
sd(world.data $ women09, na.rm = TRUE)
```

```
[1] 11.05299
```

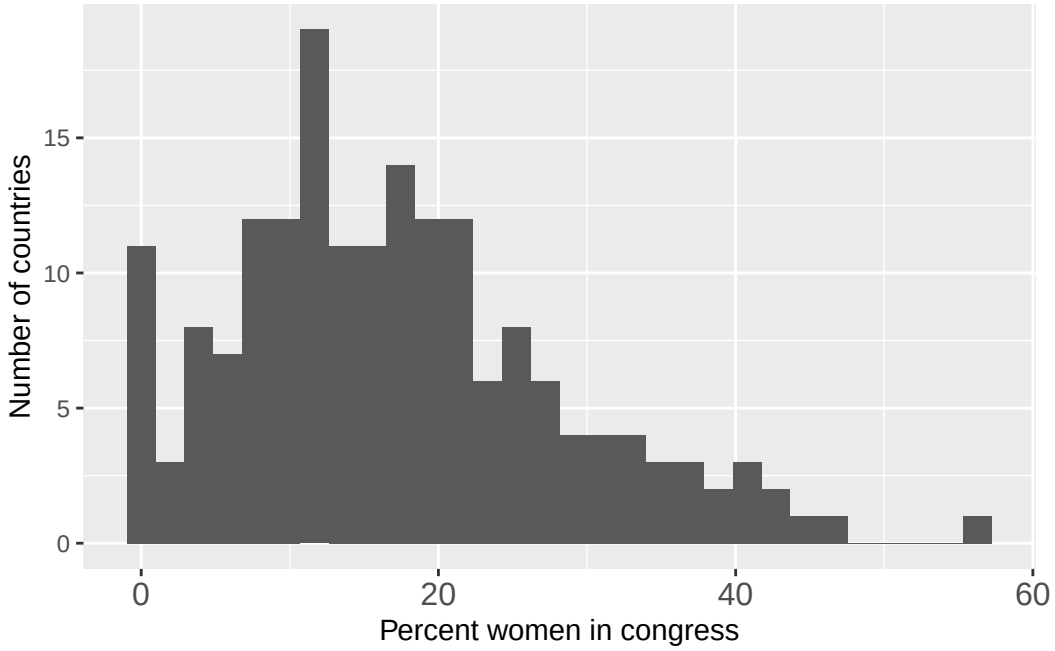
```r
# assign variable to an object with NA's removed
women09_clean <- na.omit(world.data$women09)
```

Create a histogram

```r
g <- ggplot(world.data, aes(x = women09)) +
  geom_histogram() + #geom_histogram(binwidth = 1)
  theme(axis.text.x = element_text(size = 12)) +
  xlab("Percent women in congress") + ylab("Number of countries")
g
```

```
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

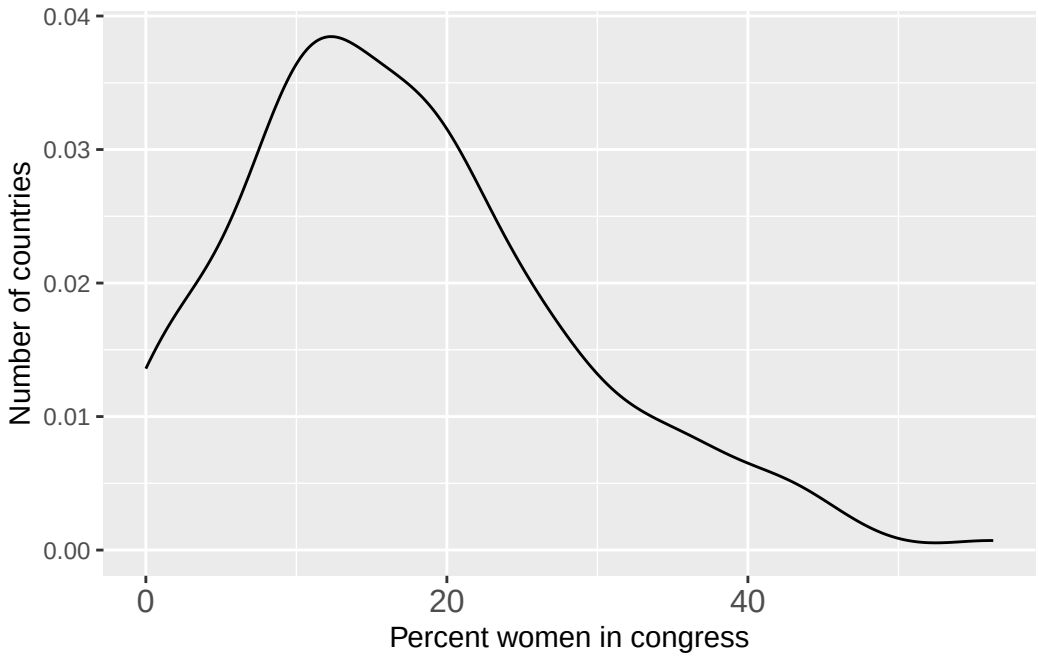
```
Warning: Removed 11 rows containing non-finite outside the scale range
(`stat_bin()`).
```



Create a density plot smoothed histogram

```
g <- ggplot(world.data, aes(x = women09))
g <- g + geom_density()
g <- g + theme(axis.text.x = element_text(size = 12))
g <- g + xlab("Percent women in congress")
g <- g + ylab("Number of countries")
g
```

```
Warning: Removed 11 rows containing non-finite outside the scale range
(`stat_density()`).
```



1. Constructing a confidence interval

```
# find point estimate
mean(world.data $ women09, na.rm = TRUE)
```

```
[1] 17.17722
```

```
# store point estimate as an object
pe <- mean(world.data $ women09, na.rm = TRUE)
# store standard deviation as an object
sd <- sd(world.data $ women09, na.rm = TRUE)
# find sample size and store it as an object
n <- length( world.data $ women09[is.na(world.data $ women09) == FALSE] )
# find standard error and store it as an object
se <- sd/sqrt(n)
# Construct a confidence interval
# construct lower bound
pe - 2 * se
```

```
[1] 15.52954
```

```
# construct upper bound
pe + 2 * se
```

```
[1] 18.82491
```

```
# 95% confidence interval equals [15.53, 18.82]
```

2. Creating a subset of dataset

```
# create a subset of data without NA's
wd.women09 <- world.data[ is.na(world.data $ women09) == FALSE , ]
head(wd.women09)
```

```
            country       colony confidence decentralization dem_other
1       Afghanistan           UK         NA               NA      10.5
2           Albania Soviet Union   49.33593             0.74      63.0
3           Algeria       France   52.05573               NA      40.8
4           Andorra        Spain         NA               NA     100.0
5            Angola     Portugal         NA               NA      40.8
6 Antigua & Barbuda           UK         NA               NA      87.5
  dem_other5 democ_regime    district_size3 durable effectiveness      enpp_3
1        10%           No     single member       4      13.71158
2 Approx 60%          Yes                           3      35.46099 1-3 parties
3 Approx 40%           No 6 or more members       5      32.62411
4       100%          Yes                         NA      78.72340
5 Approx 40%           No                           3      19.14894
6 Approx 90%          Yes     single member      NA      59.81088 1-3 parties
          eu fhrate04_rev fhrate08_rev frac_eth frac_eth3 free_business
1 Not member          2.5            3   0.7693      High            NA
2 Not member            5            8   0.2204       Low          68.0
3 Not member          2.5            3   0.3394    Medium          71.2
4 Not member    Most free           12   0.7139      High            NA
5 Not member          2.5            3   0.7867      High          43.4
6 Not member            6           10   0.1643       Low            NA
  free_corrupt free_finance free_fiscal free_govspend free_invest free_labor
1           NA           NA          NA            NA          NA         NA
2           34           70        92.6          74.2          70       52.1
3           32           30        83.5          73.4          45       56.4
```

```
4           NA           NA          NA            NA          NA         NA
5           19           40        85.1          62.8          35       45.2
6           NA           NA          NA            NA          NA         NA
  free_monetary free_overall free_property free_trade gdp08 gdp_10_thou
1            NA           NA            NA         NA  30.6          NA
2          78.7         66.0            35       85.8  24.3      0.1535
3          77.2         56.9            30       70.7 276.0      0.1785
4            NA           NA            NA         NA    NA          NA
5          62.6         48.4            20       70.4 106.3      0.0857
6            NA           NA            NA         NA    NA      1.0449
  gdp_cap2 gdp_cap3 gdppcap08 gender_equal3 gini04 gini08   hi_gdp indy
1                          NA             NA     NA          1919
2      Low   Middle      7715           28.2   31.1  Low GDP 1991
3      Low   Middle      8033           35.3   35.3  Low GDP 1962
4                          NA             NA     NA          1278
5      Low   Middle      5899             NA     NA  Low GDP 1975
6     High     High        NA             NA     NA High GDP 1981
        oecd  old2006  old2003     pmat12_3    pop03 pop08    pop08_3
1 Not member       NA       NA                   NA  27.4 >=16.8 mil
2 Not member 8.479821 7.278363 Low post-mat  3169064   3.1  <=4.3 mil
3 Not member 4.578136 4.045199              31832610  34.4 >=16.8 mil
4 Not member       NA       NA                66000    NA
5 Not member 2.450295 2.930542              13522110  18.0 >=16.8 mil
6 Not member       NA 8.186610                 78580    NA
          popcat3 pr_sys protact3        regime_type3      region sources
1 Moderate (1-29m)     No                Dictatorship Middle East      NA
2 Moderate (1-29m)     No Moderate Parliamentary democ  C&E Europe      NA
3 Moderate (1-29m)    Yes                Dictatorship      Africa      NA
4 Small (under 1m)     No          Parliamentary democ   W. Europe      NA
5 Moderate (1-29m)    Yes                Dictatorship      Africa      NA
6 Small (under 1m)     No          Parliamentary democ  S. America      NA
        typerel unions urban03 urban06 vi_rel3 votevap00s women05 women09
1         Muslim     NA      NA   23.28                 NA      NA    27.7
2         Muslim     NA 44.2390   46.14  20-50%      59.56     6.4    16.4
3         Muslim     NA 58.8302   63.94    >50%         NA      NA     7.7
4 Roman Catholic     NA 91.7404   90.28              20.95    14.3    35.7
5 Roman Catholic     NA 36.1806   53.96                 NA      NA    37.3
6     Protestant     NA 37.7566   39.60              76.34    10.5    10.5
  womyear       womyear2  yng2003  young06
1      NA                      NA       NA
2    1920 1944 or before 27.34834 26.35428
3    1962     After 1944 33.91887 28.94154
4    1973     After 1944       NA       NA
```

```
5    1975     After 1944 47.62524 46.32196
6    1951     After 1944 20.66509       NA
```

```
# find confidence interval within subset
# point estimate
pe <- mean( wd.women09 $ women09 )
# standard error
sd <- sd( wd.women09 $ women09 )
n <- length( wd.women09 $ women09 )
se <- sd/sqrt(n)
pe
```

```
[1] 17.17722
```

```
# lower bound
pe - 2 * se
```

```
[1] 15.52954
```

```
# upper bound
pe + 2 * se
```

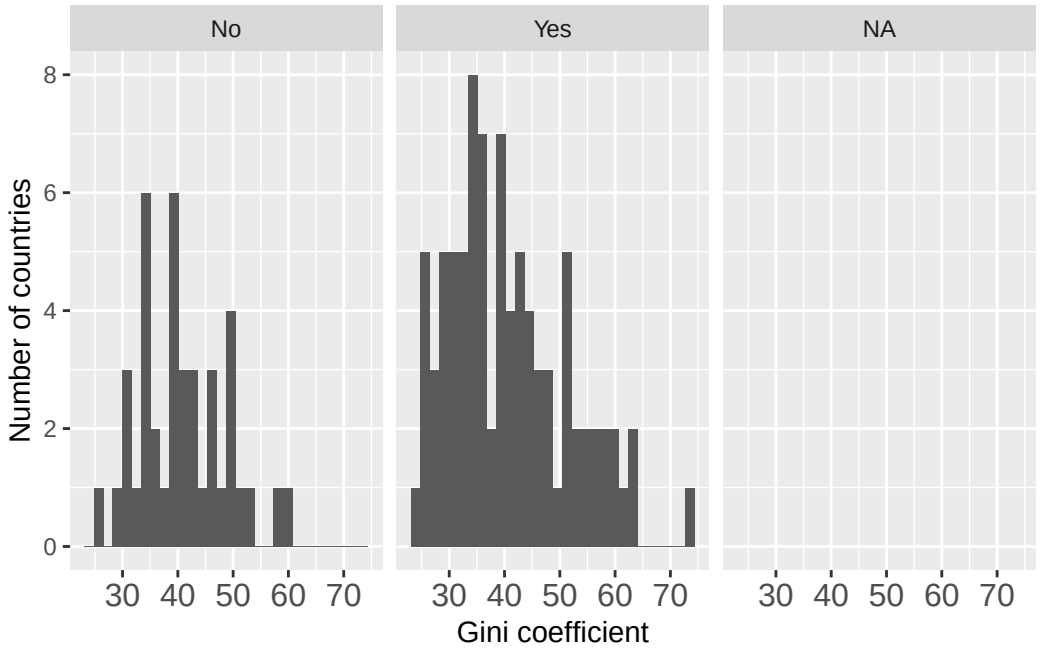
```
[1] 18.82491
```

3. Re-labeling a factor variable

```
# create a histogram using facet wrap
g <- ggplot(world.data, aes(x = gini08))
g <- g + geom_histogram()
g <- g + theme(axis.text.x = element_text(size = 12))
g <- g + xlab("Gini coefficient")
g <- g + ylab("Number of countries")
g <- g + facet_wrap( ~ democ_regime) # we can create separate histograms of gini08 for diffe
g
```

```
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

```
Warning: Removed 64 rows containing non-finite outside the scale range
(`stat_bin()`).
```

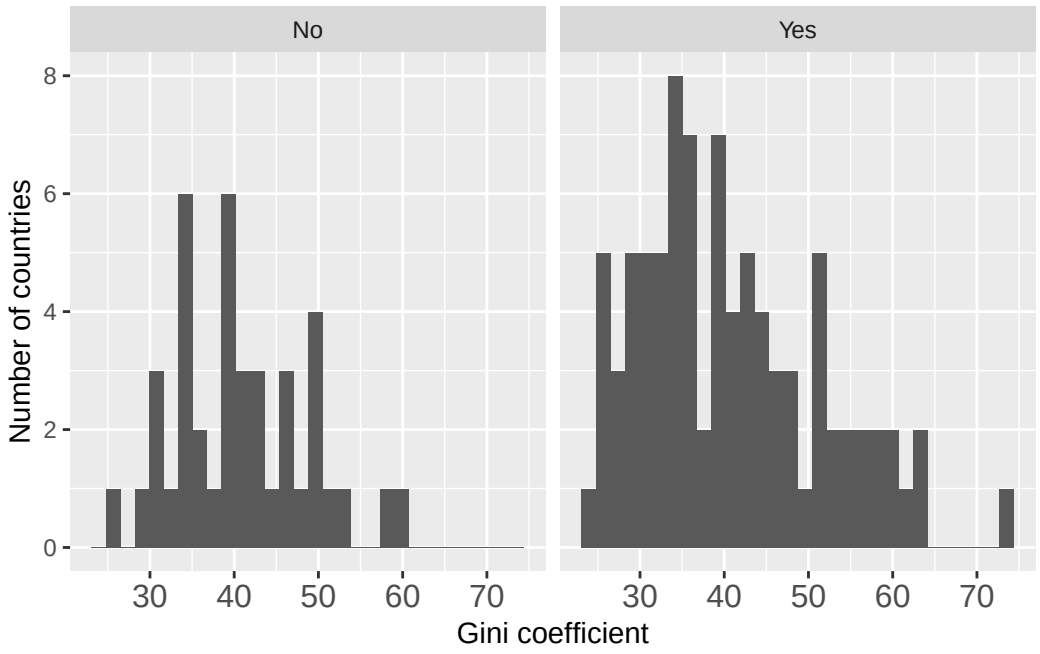


Construct a smaller dataset

```
# assign variable as an object with NA's removed
wd.dem <- world.data[ is.na(world.data $ democ_regime) == FALSE , ]
wd.dem <- world.data %>% filter(!is.na(democ_regime))
# create a histogram using the new dataset
g <- ggplot(wd.dem, aes(x = gini08))
g <- g + geom_histogram()
g <- g + theme(axis.text.x = element_text(size = 12))
g <- g + xlab("Gini coefficient")
g <- g + ylab("Number of countries")
g <- g + facet_wrap( ~ democ_regime)
g
```

```
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

```
Warning: Removed 62 rows containing non-finite outside the scale range
(`stat_bin()`).
```

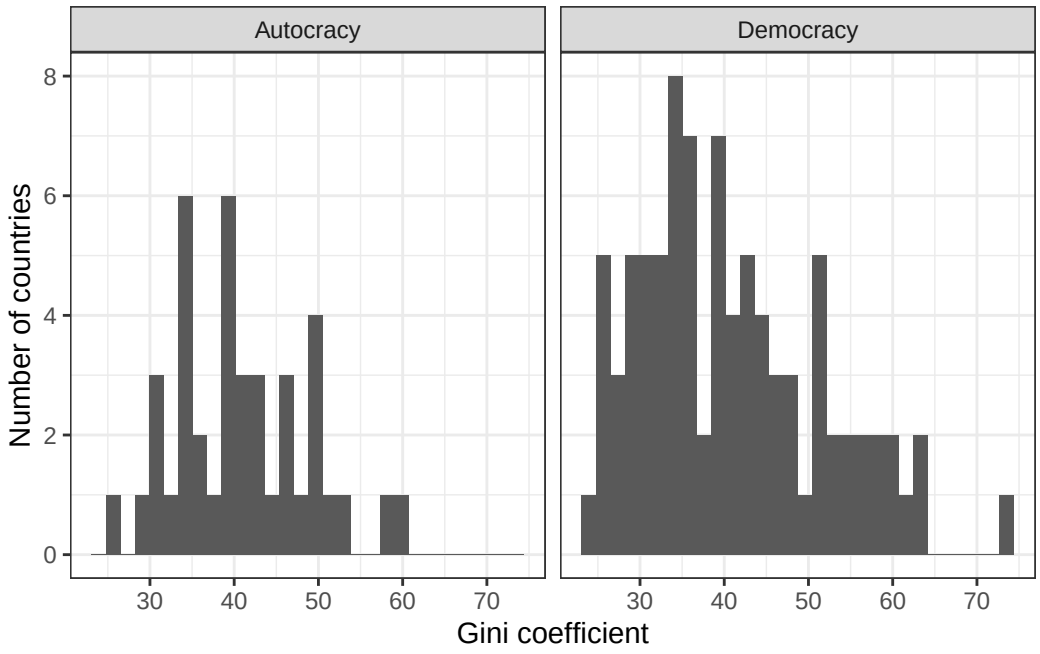


Re-create histogram with changed labels

```
wd.dem $ democ <- factor(wd.dem $ democ_regime,
                         levels = c("No", "Yes"),
                         labels = c("Autocracy", "Democracy")
                         )
# create a histogram with new factors
g <- ggplot(wd.dem, aes(x = gini08))
g <- g + geom_histogram()
g <- g + theme(axis.text.x = element_text(size = 12))
g <- g + xlab("Gini coefficient")
g <- g + ylab("Number of countries")
g <- g + facet_wrap( ~ democ)
g + theme_bw() # I like a black-white theme more
```

```
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

```
Warning: Removed 62 rows containing non-finite outside the scale range
(`stat_bin()`).
```


Autocracy
Democracy
Number of countries
0
2
4
6
8
30
40
50
60
70
Gini coefficient