IMAI QSS CH3

Musaab Farooqui

8/17/2021

```
data("afghan", package = "qss")
data("afghan.village", package = "qss")
```

Chapter 3.1: *"Measuring Civilian Victimization during wartime"*

```
afghan %>%
  select(age, educ.years, employed, income) %>%
  summary()
```

```
##       age          educ.years        employed         income         
##  Min.   :15.00   Min.   : 0.000   Min.   :0.0000   Length:2754       
##  1st Qu.:22.00   1st Qu.: 0.000   1st Qu.:0.0000   Class :character  
##  Median :30.00   Median : 1.000   Median :1.0000   Mode  :character  
##  Mean   :32.39   Mean   : 4.002   Mean   :0.5828                     
##  3rd Qu.:40.00   3rd Qu.: 8.000   3rd Qu.:1.0000                     
##  Max.   :80.00   Max.   :18.000   Max.   :1.0000                     
```

```
count(afghan, income)
```

```
##            income    n
## 1   10,001-20,000  616
## 2    2,001-10,000 1420
## 3   20,001-30,000   93
## 4 less than 2,000  457
## 5     over 30,000   14
## 6            <NA>  154
```

```
afghan %>%
  group_by(violent.exp.ISAF, violent.exp.taliban) %>%
  count() %>%
  ungroup() %>%
  mutate(prop = n /sum(n))
```

```
## # A tibble: 9 x 4
##   violent.exp.ISAF violent.exp.taliban     n    prop
##              <int>               <int> <int>   <dbl>
## 1                0                   0  1330 0.483
## 2                0                   1   354 0.129
## 3                0                  NA    22 0.00799
## 4                1                   0   475 0.172
## 5                1                   1   526 0.191
## 6                1                  NA    22 0.00799
## 7               NA                   0     7 0.00254
## 8               NA                   1     8 0.00290
## 9               NA                  NA    10 0.00363
```

Chapter 3.2: *“Handling Missing Data in R”*

```r
head(afghan$income, n = 10)
```

```
##  [1] "2,001-10,000"  "2,001-10,000"  "2,001-10,000"  "2,001-10,000"
##  [5] "2,001-10,000"  NA              "10,001-20,000" "2,001-10,000"
##  [9] "2,001-10,000"  NA
```

```r
head(is.na (afghan$income), n = 10)
```

```
##  [1] FALSE FALSE FALSE FALSE FALSE  TRUE FALSE FALSE FALSE  TRUE
```

```r
summarise(afghan,
          n_missing = sum(is.na(income)),
          p_missing = mean(is.na(income)))
```

```
##   n_missing  p_missing
## 1       154 0.05591866
```

```r
violent_exp_prop <-
  afghan %>%
  group_by(violent.exp.ISAF, violent.exp.taliban) %>%
  count() %>%
  ungroup() %>%
  mutate(prop = n / sum(n)) %>%
  select(-n)
violent_exp_prop
```

```
## # A tibble: 9 x 3
##   violent.exp.ISAF violent.exp.taliban    prop
##              <int>               <int>   <dbl>
## 1                0                   0 0.483
## 2                0                   1 0.129
## 3                0                  NA 0.00799
## 4                1                   0 0.172
## 5                1                   1 0.191
## 6                1                  NA 0.00799
## 7               NA                   0 0.00254
## 8               NA                   1 0.00290
## 9               NA                  NA 0.00363
```

```
violent_exp_prop %>%
  spread(violent.exp.taliban, prop)
```

```
## # A tibble: 3 x 4
##   violent.exp.ISAF     ‘0‘     ‘1‘  ‘<NA>‘
##              <int>   <dbl>   <dbl>   <dbl>
## 1                0 0.483   0.129   0.00799
## 2                1 0.172   0.191   0.00799
## 3               NA 0.00254 0.00290 0.00363
```

```
drop_na(afghan) %>% head()
```

```
##   province     district village.id age educ.years employed        income
## 1    Logar Baraki Barak         80  26         10        0  2,001-10,000
## 2    Logar Baraki Barak         80  49          3        1  2,001-10,000
## 3    Logar Baraki Barak         80  60          0        1  2,001-10,000
## 4    Logar Baraki Barak         80  34         14        1  2,001-10,000
## 5    Logar Baraki Barak         80  21         12        1  2,001-10,000
## 6    Logar Baraki Barak         80  42          6        1 10,001-20,000
##   violent.exp.ISAF violent.exp.taliban list.group list.response
## 1                0                   0    control             0
## 2                0                   0    control             1
## 3                1                   0    control             1
## 4                0                   0       ISAF             3
## 5                0                   0       ISAF             3
## 6                0                   0    taliban             1
```

```
NA
```

```
## [1] NA
```

```
NA_integer_
```

```
## [1] NA
```

```
NA_real_
```

```
## [1] NA
```

```
NA_character_
```

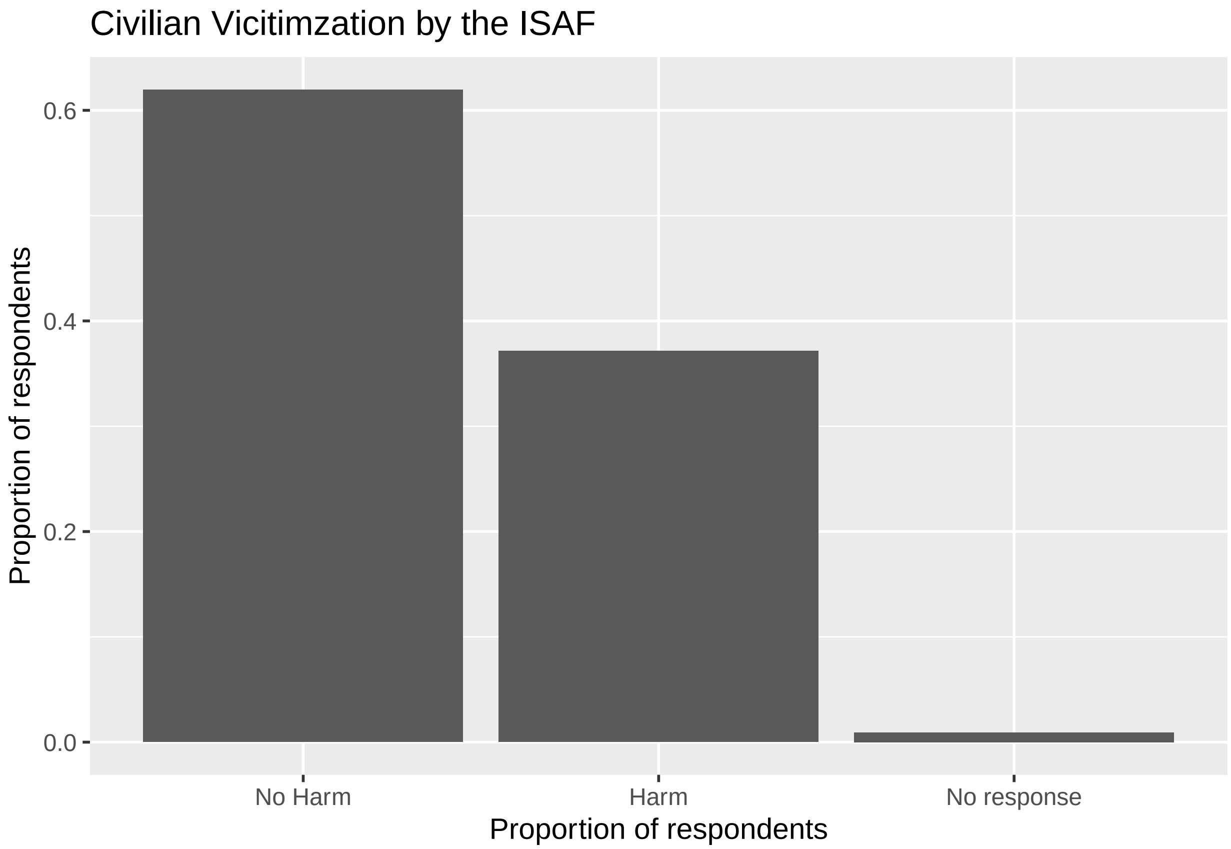
```
## [1] NA
```

```
x <- 1:5
class(x)
if_else(x<3, x,NA)
if_else(x < 3, x, NA_integer_)
```

3.3 Visualizing the Univariate Distribution

3.3.1 Barplot

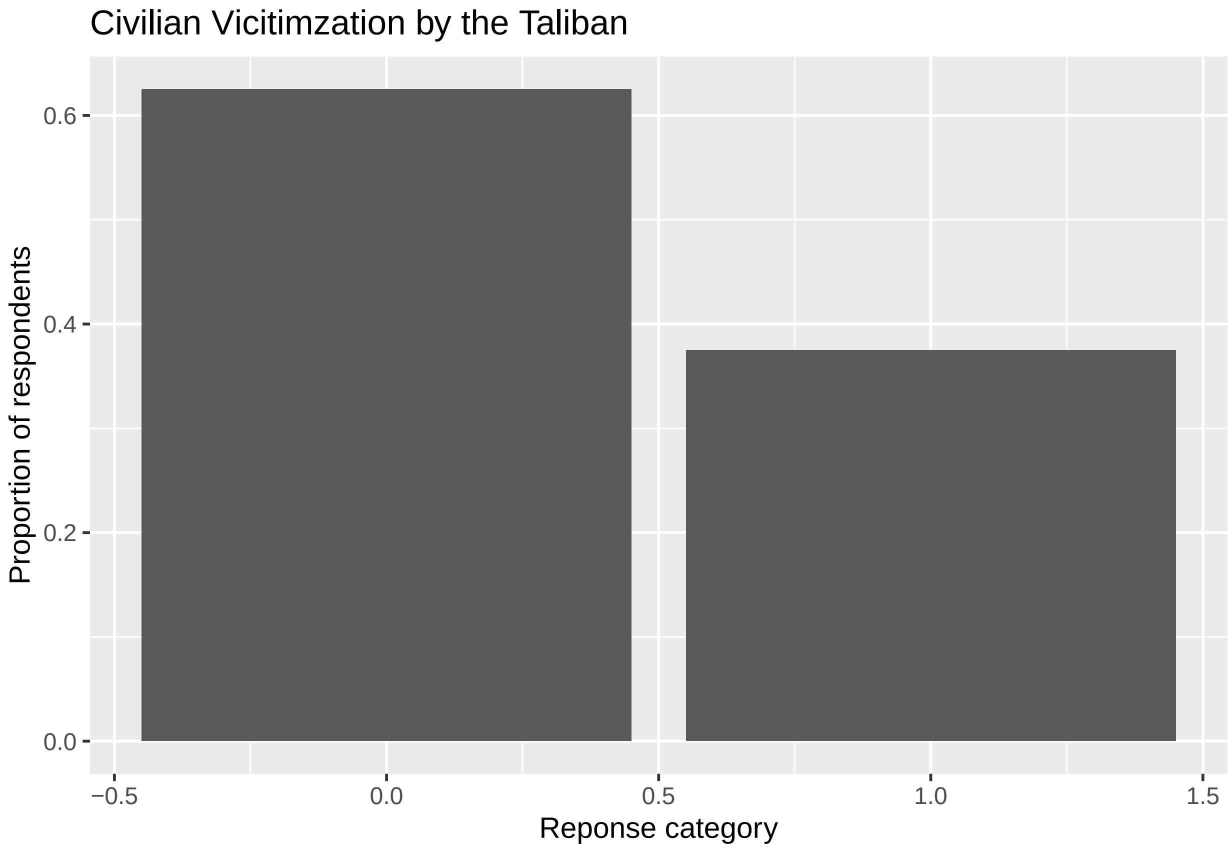
```
afghan <-
  afghan %>%
  mutate(violent.exp.ISAF.fct =
           fct_explicit_na(fct_recode(factor(violent.exp.ISAF), Harm = "1", "No Harm" = "0"),
                            "No response"))
ggplot(afghan, aes(x = violent.exp.ISAF.fct, y = ..prop.., group = 1)) +
  geom_bar() +
  xlab("Proportion of respondents") +
  ylab("Proportion of respondents") +
  ggtitle("Civilian Vicitimzation by the ISAF")
```



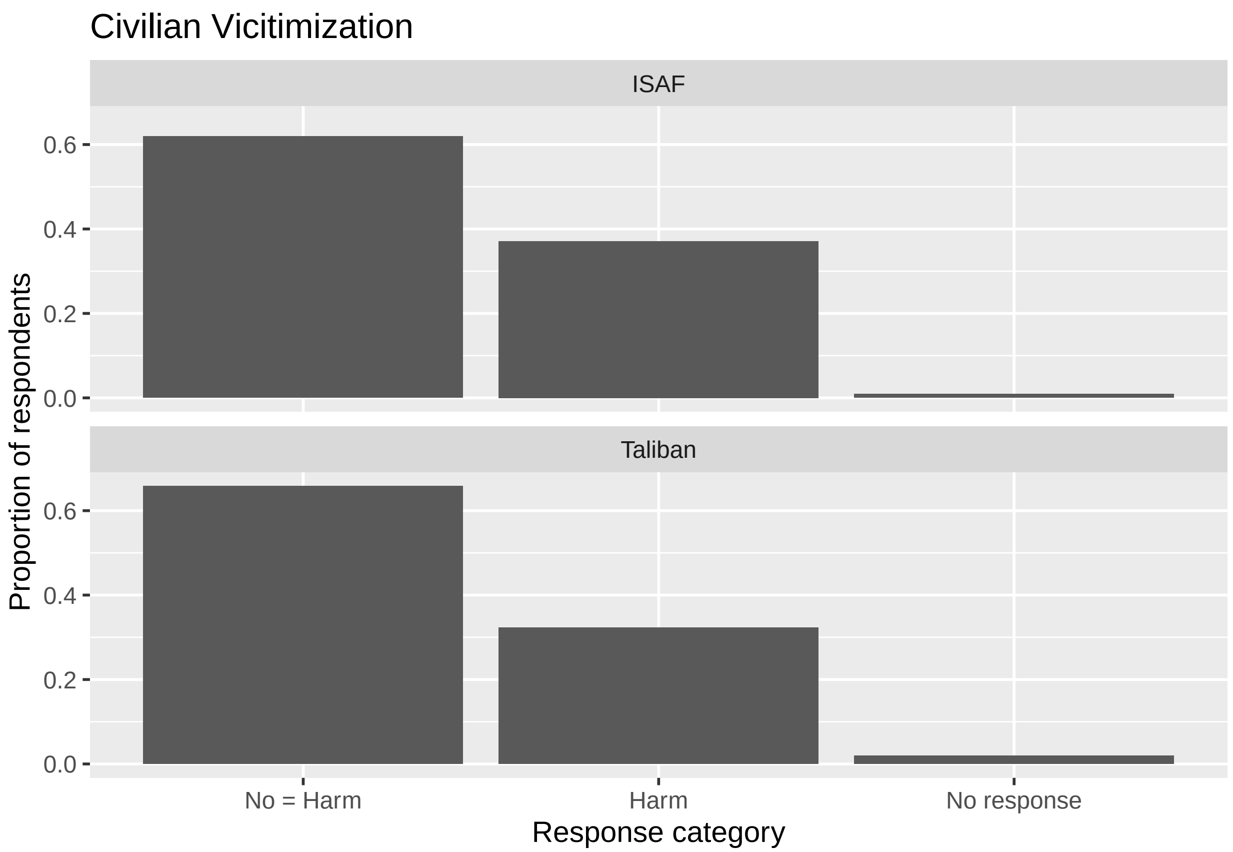
```
afghan <-
  afghan %>%
  mutate(violent.exp.taliban.fct =
           fct_explicit_na(fct_recode(factor(violent.exp.taliban), Harm = "1", "No Harm" = "0"),
                           "No response"))
```

```
ggplot(afghan, aes(x = violent.exp.ISAF, y = ..prop.., group = 1)) +
  geom_bar() +
  xlab("Reponse category")+
  ylab("Proportion of respondents") +
  ggtitle("Civilian Vicitimzation by the Taliban")
```

```
## Warning: Removed 25 rows containing non-finite values (stat_count).
```



```
select(afghan, violent.exp.ISAF, violent.exp.taliban) %>%
  gather(variable, value) %>%
  mutate(value = fct_explicit_na(fct_recode(factor(value),
              Harm = "1", "No = Harm" = "0"),
              "No response"),
  variable = recode(variable,
                    violent.exp.ISAF = "ISAF",
                    violent.exp.taliban = "Taliban")) %>%
  ggplot(aes(x = value, y = ..prop.., group = 1)) +
  geom_bar() +
  facet_wrap(~ variable, ncol = 1) +
  xlab("Response category") +
  ylab("Proportion of respondents") +
  ggtitle("Civilian Vicitimization")
```

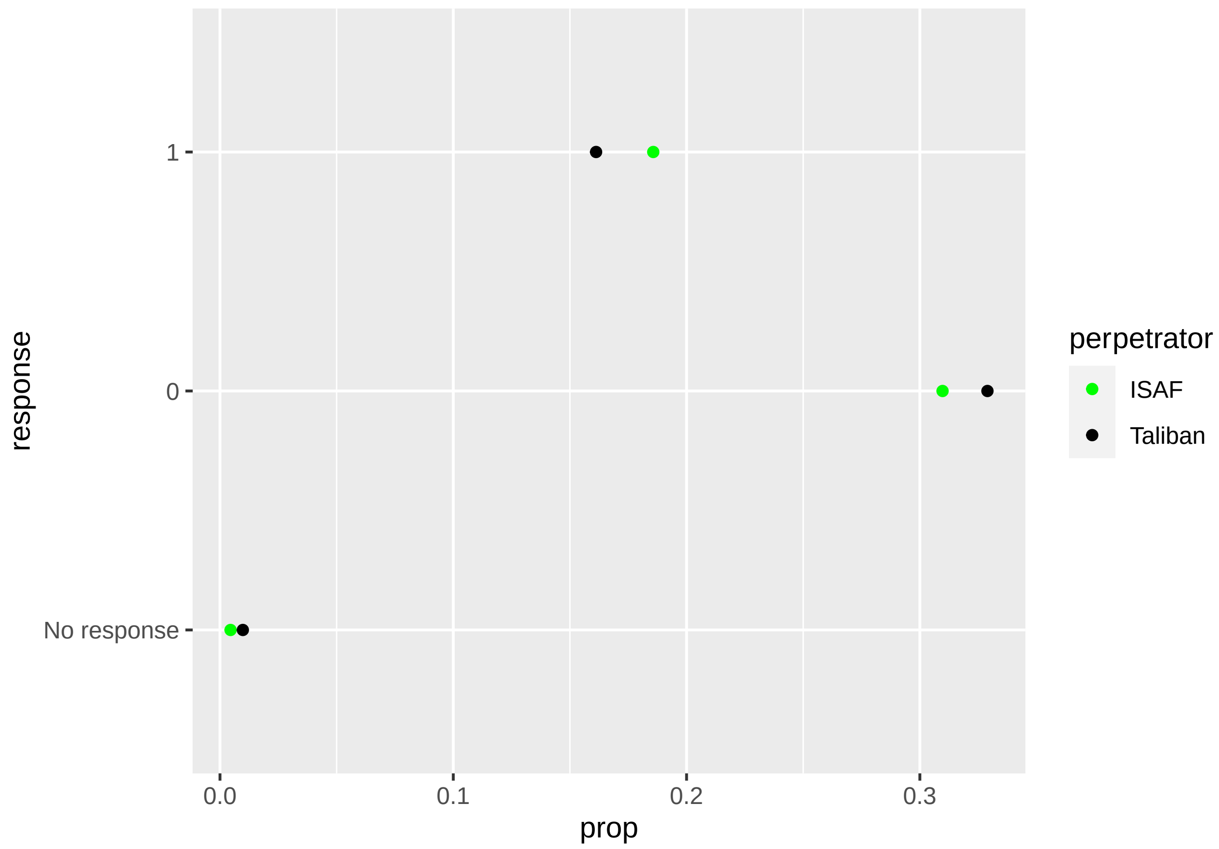


```
violent_exp <-
  afghan %>%
  select(violent.exp.ISAF, violent.exp.taliban) %>%
  gather(perpetrator, response) %>%
  mutate(perpetrator = str_replace(perpetrator, "violent\\.exp\\.",""),
         perpetrator = str_replace(perpetrator, "taliban", "Taliban"),
         response = fct_recode(factor(response), "No response"),
         response = fct_explicit_na(response, "No response"),
         response = fct_relevel(response, c("No response", "No Harm"))) %>%
  count(perpetrator, response) %>%
  mutate(prop = n / sum(n))
```

```
## Warning: Unknown levels in ‘f‘: No response
```

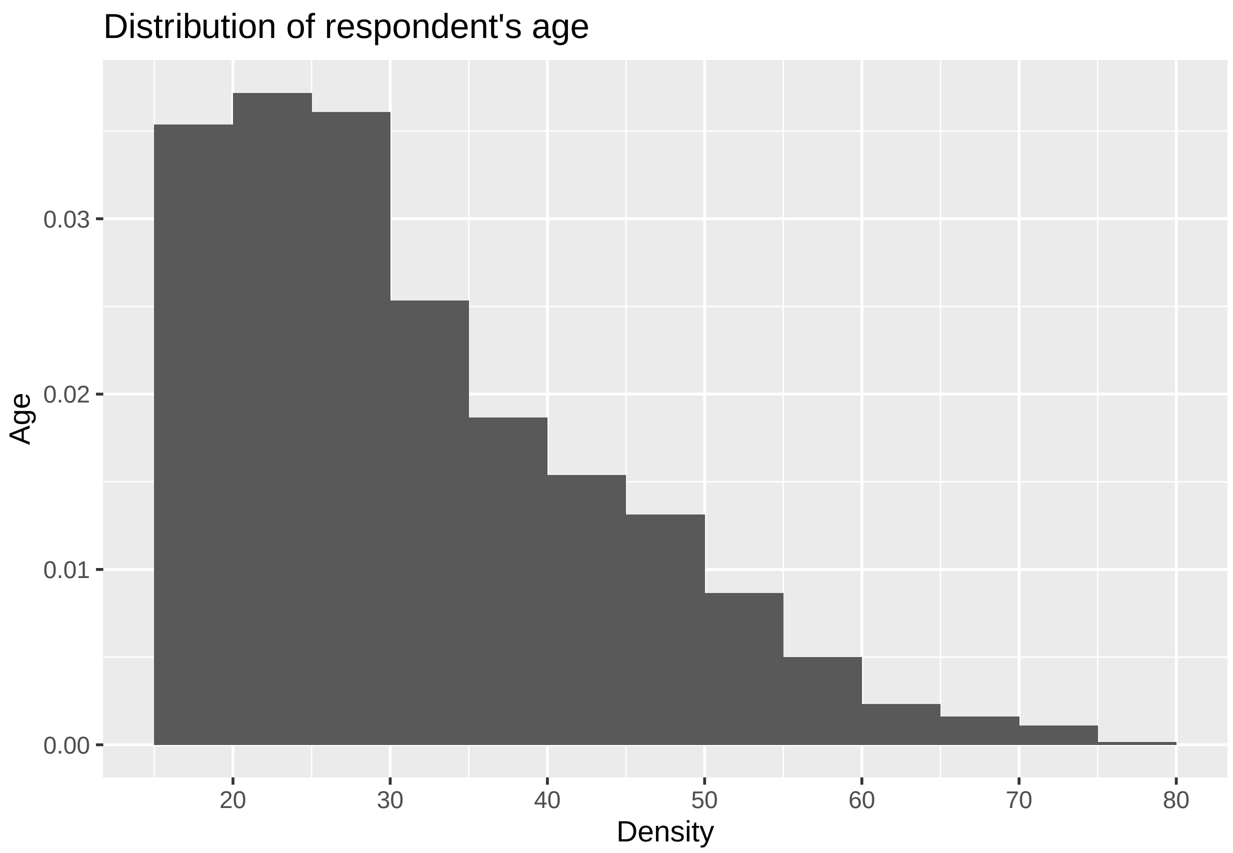
```
## Warning: Unknown levels in ‘f‘: No Harm
```

```
ggplot(violent_exp, aes(x = prop, y = response, color = perpetrator)) +
  geom_point() +
  scale_color_manual(values = c(ISAF = "green", Taliban = "black"))
```

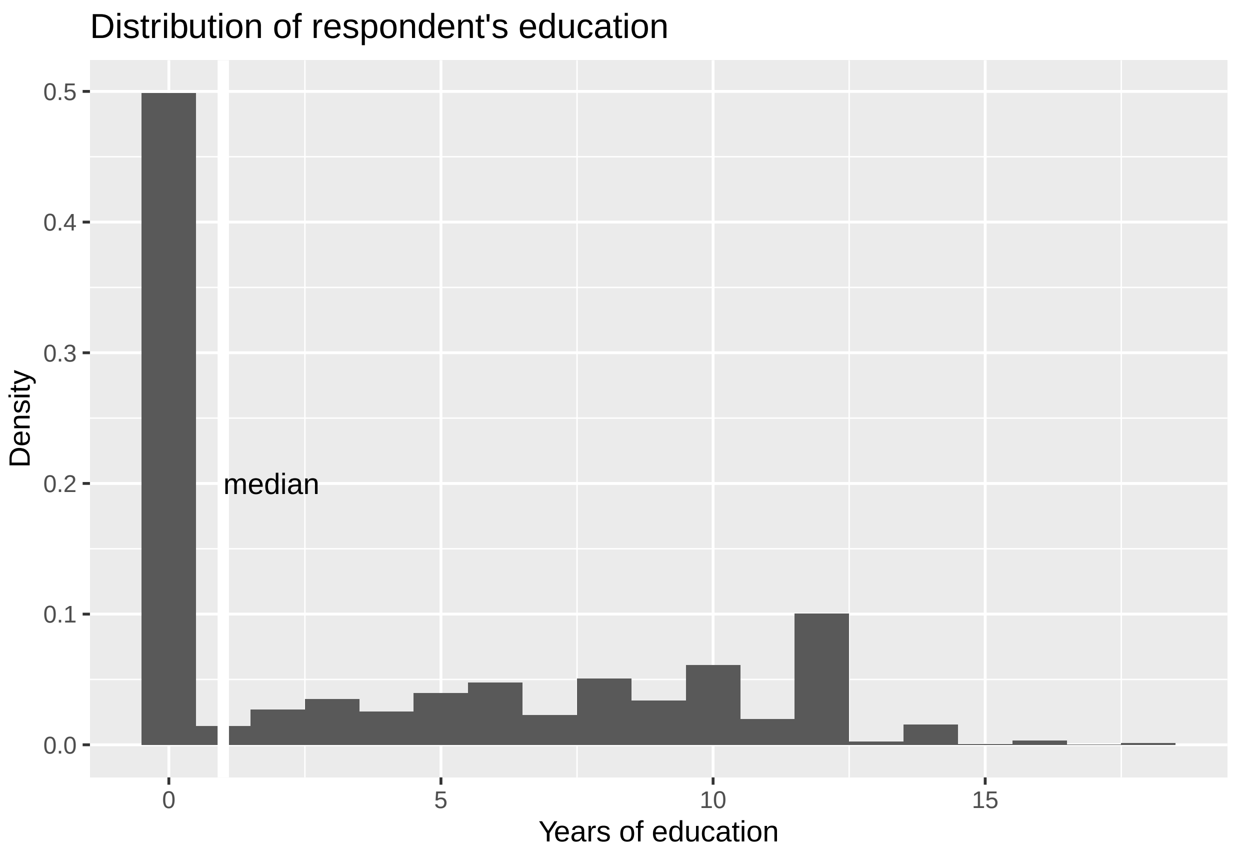


3.3.2 Histogram

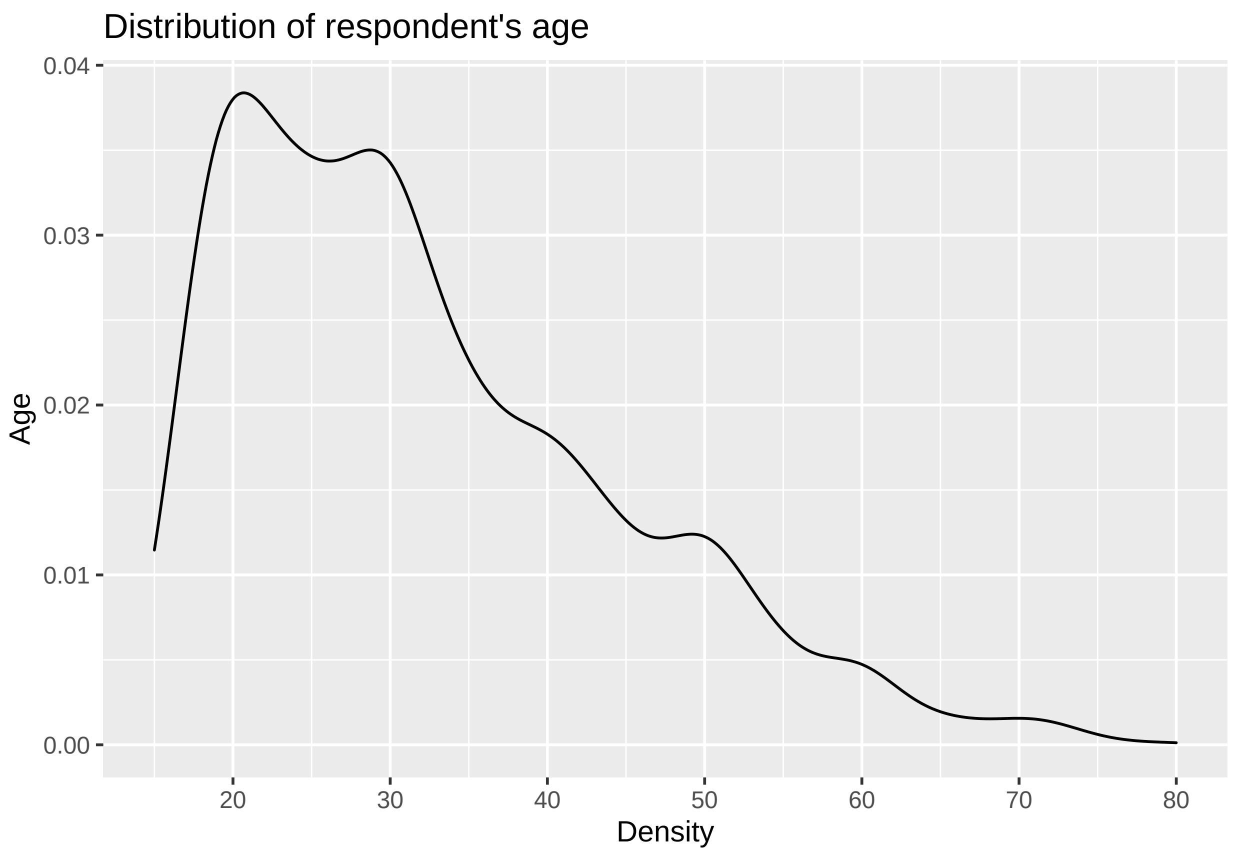
```
ggplot(afghan, aes(x = age, y = ..density..)) +
  geom_histogram(binwidth = 5, boundary = 0) +
  scale_x_continuous(breaks = seq(20, 80, by = 10)) +
  labs(title = "Distribution of respondent's age",
       y = "Age", x = "Density")
```



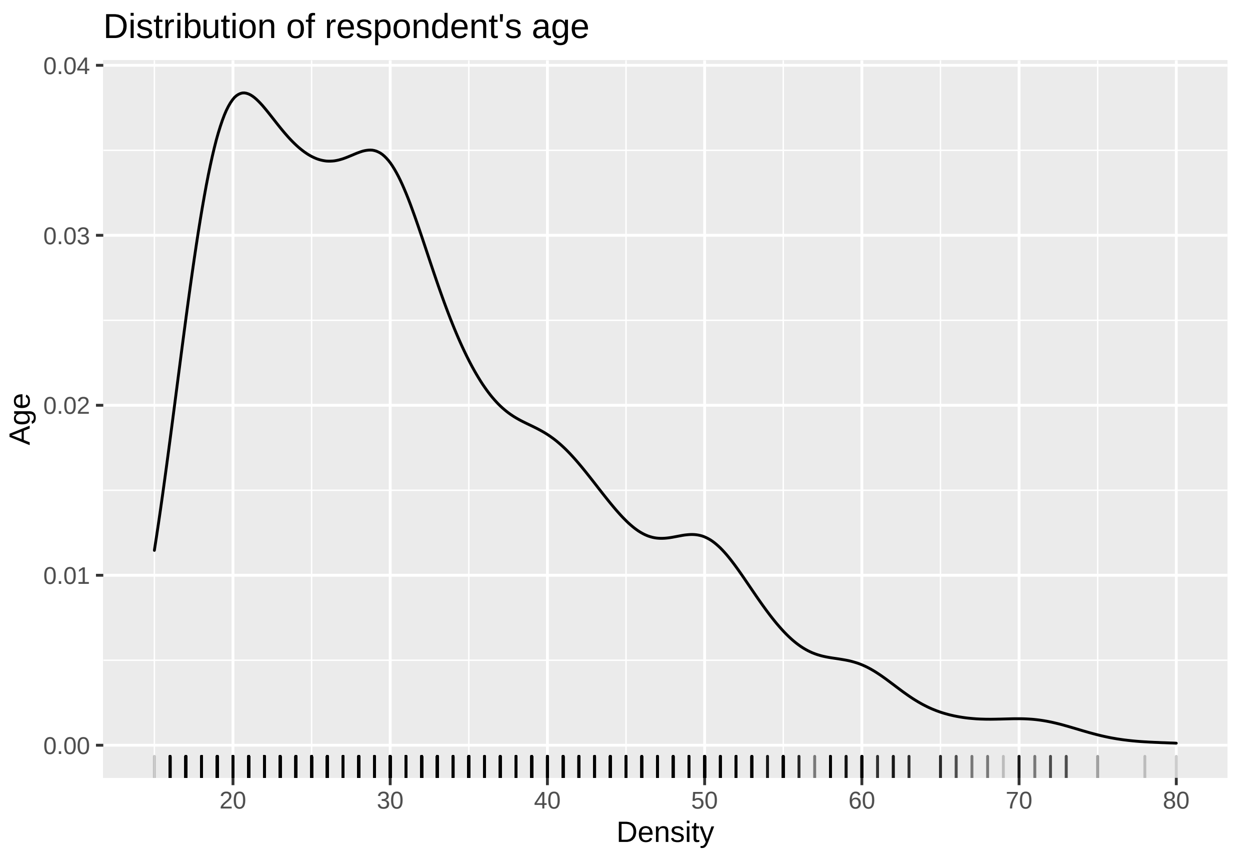
```
ggplot(afghan, aes(x = educ.years, y = ..density..)) +
  geom_histogram(binwidth = 1, center = 0) +
  geom_vline(xintercept = median(afghan$educ.years),
             color = "white", size = 2) +
  annotate("text", x = median(afghan$educ.years),
           y = 0.2, label = "median", hjust = 0) +
  labs(title = "Distribution of respondent’s education",
       x = "Years of education",
       y = "Density")
```



```
dens_plot <- ggplot(afghan, aes(x = age)) +
  geom_density() +
  scale_x_continuous(breaks = seq(20, 80, by = 10)) +
  labs(title = "Distribution of respondent’s age",
       y = "Age", x = "Density")
dens_plot
```

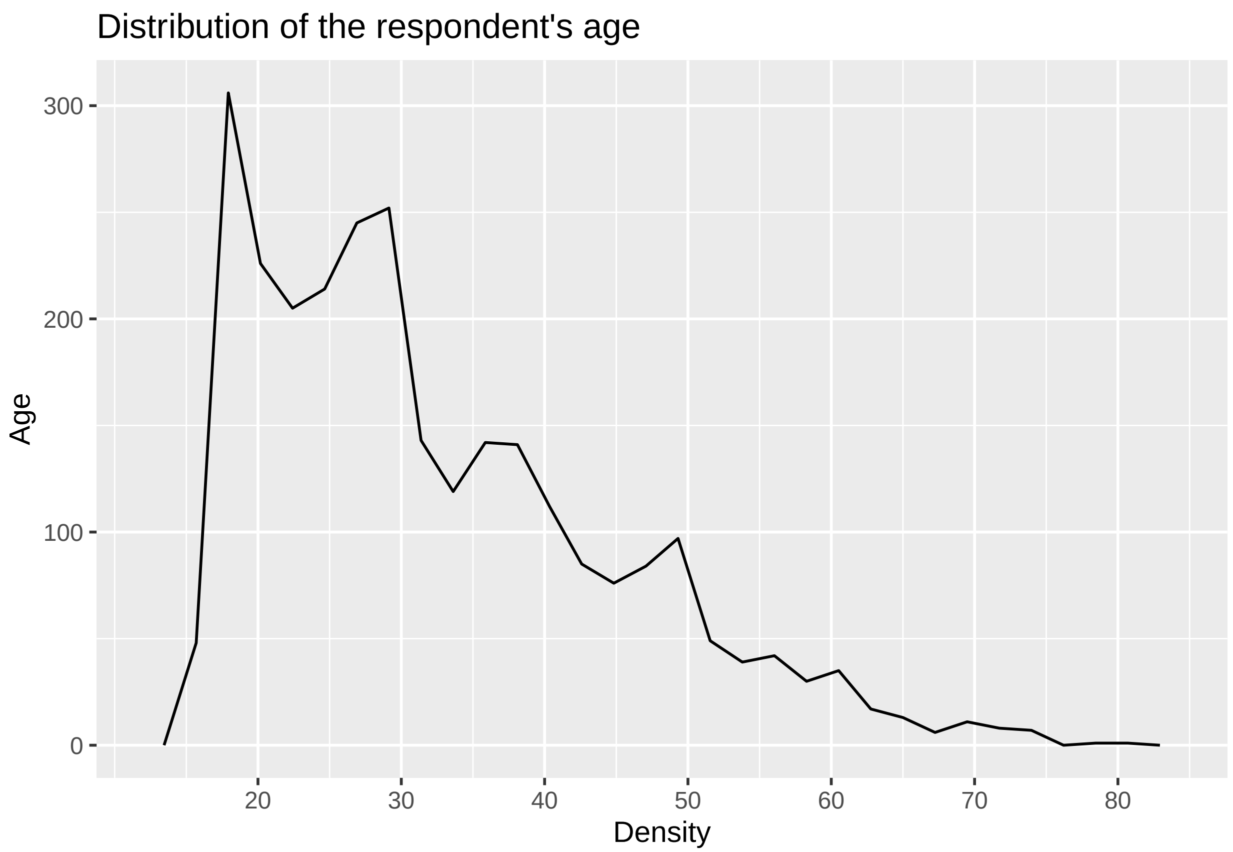


```
dens_plot + geom_rug(alpha = .2)
```



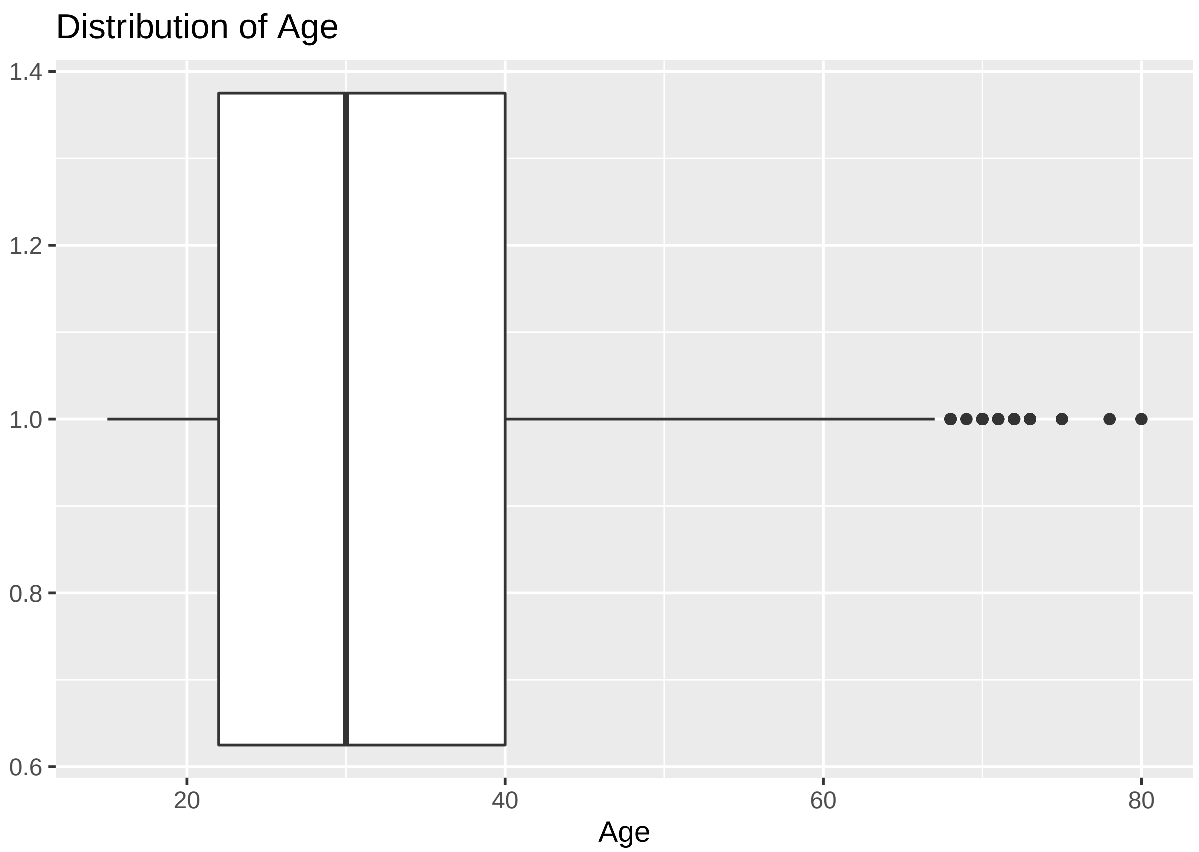
```
ggplot(afghan, aes(x = age)) +
  geom_freqpoly() +
  scale_x_continuous(breaks = seq(20, 80, by = 10)) +
  labs(title = "Distribution of the respondent’s age", y = "Age", x = "Density")
```

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

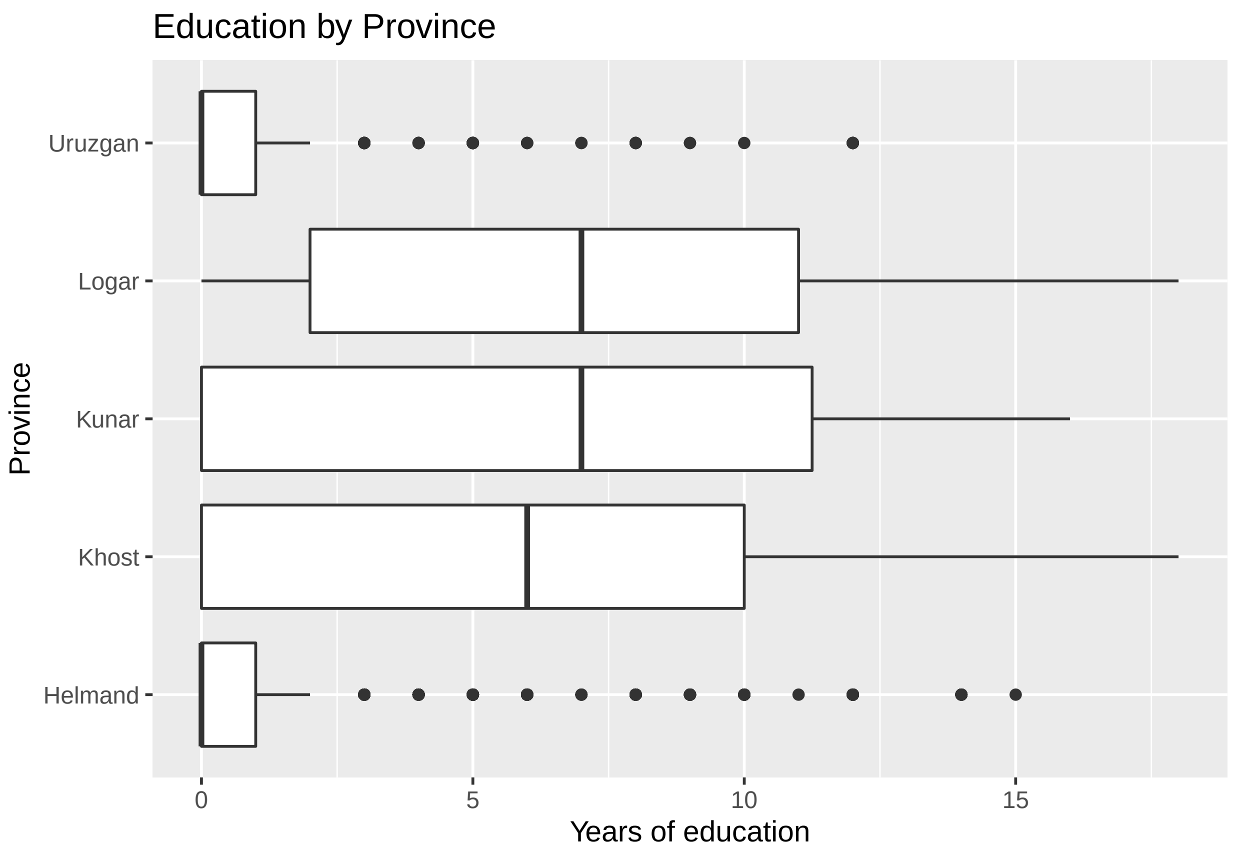


3.3.3 Boxplot

```
ggplot(afghan, aes(x =1, y = age)) +
  geom_boxplot() +
  coord_flip() +
  labs(y = "Age", x = "", title = "Distribution of Age")
```



```
ggplot(afghan, aes(y = educ.years, x = province)) +
  geom_boxplot() +
  coord_flip() +
  labs(x = "Province", y = "Years of education",
       title = "Education by Province")
```



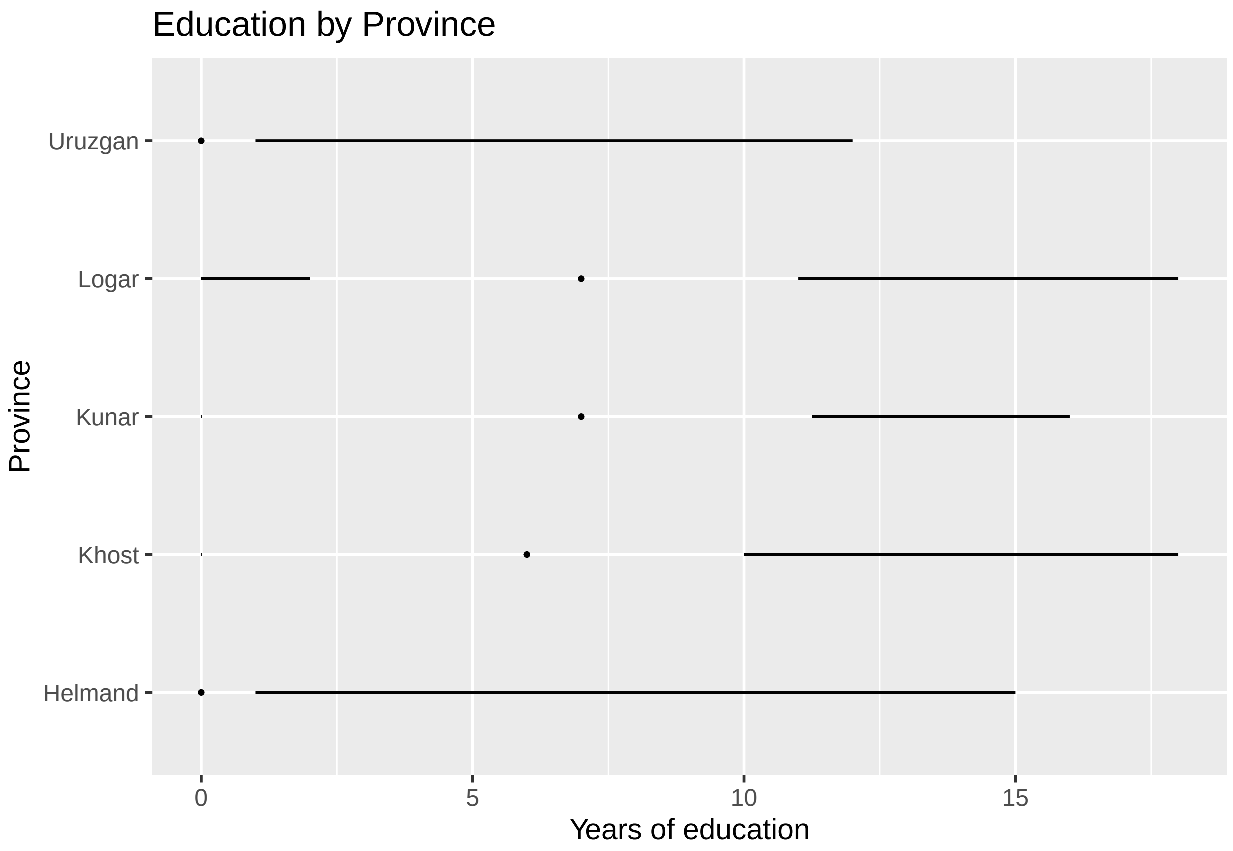
```
afghan %>%
  group_by(province) %>%
  summarise(educ.years = mean(educ.years, na.rm = TRUE),
            violent.exp.taliban =
              mean(violent.exp.taliban, na.rm = TRUE),
             violent.exp.ISAF =
               mean(violent.exp.ISAF, na.rm =TRUE)) %>%
            arrange(educ.years)
```

```
## # A tibble: 5 x 4
##   province educ.years violent.exp.taliban violent.exp.ISAF
##   <chr>         <dbl>               <dbl>            <dbl>
## 1 Uruzgan        1.04              0.455            0.496
## 2 Helmand        1.60              0.504            0.541
## 3 Khost          5.79              0.233            0.242
## 4 Kunar          5.93              0.303            0.399
## 5 Logar          6.70              0.0802           0.144
```

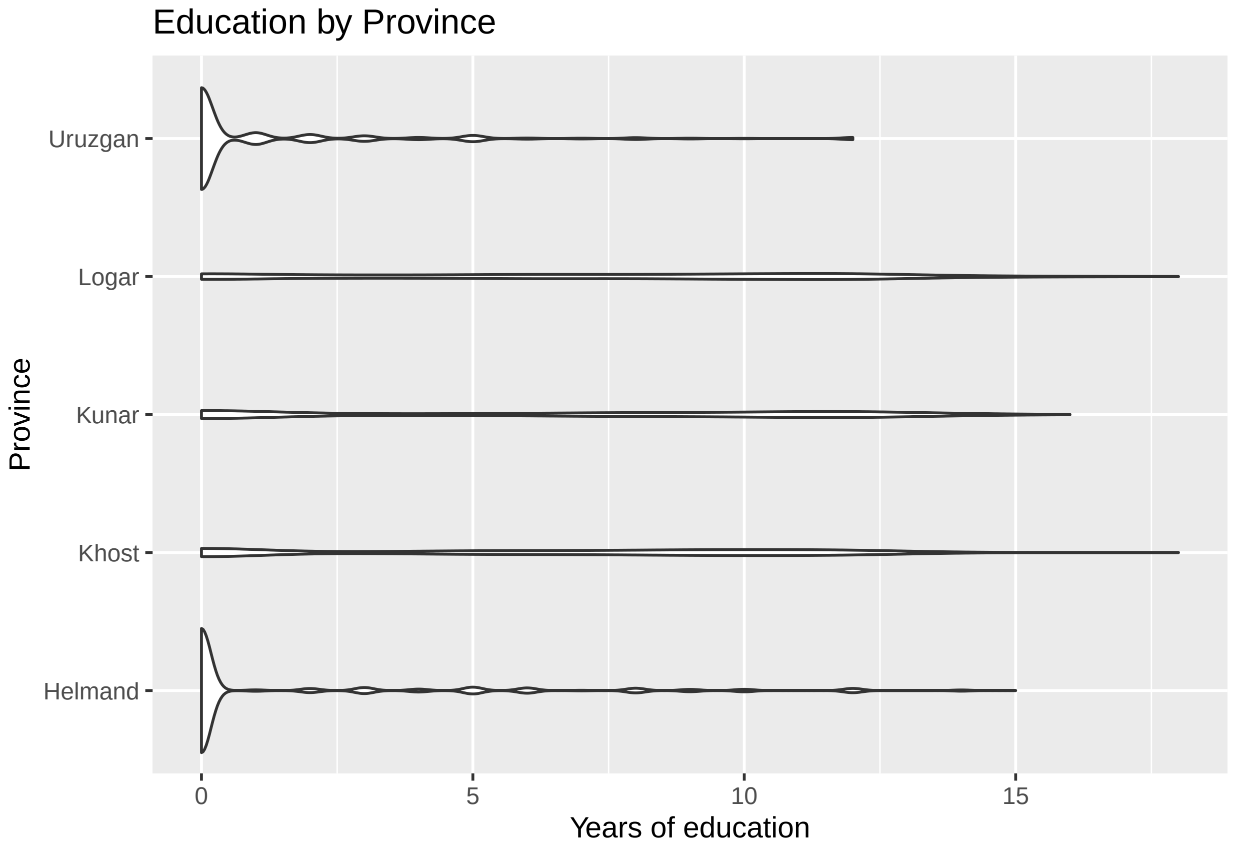
```
library(ggthemes)
```

```
## Warning: package 'ggthemes' was built under R version 4.0.2
```

```
ggplot(afghan, aes(y = educ.years, x = province)) +
  geom_tufteboxplot() +
  coord_flip() +
  labs(x = "Province", y = "Years of education",
    title = "Education by Province")
```



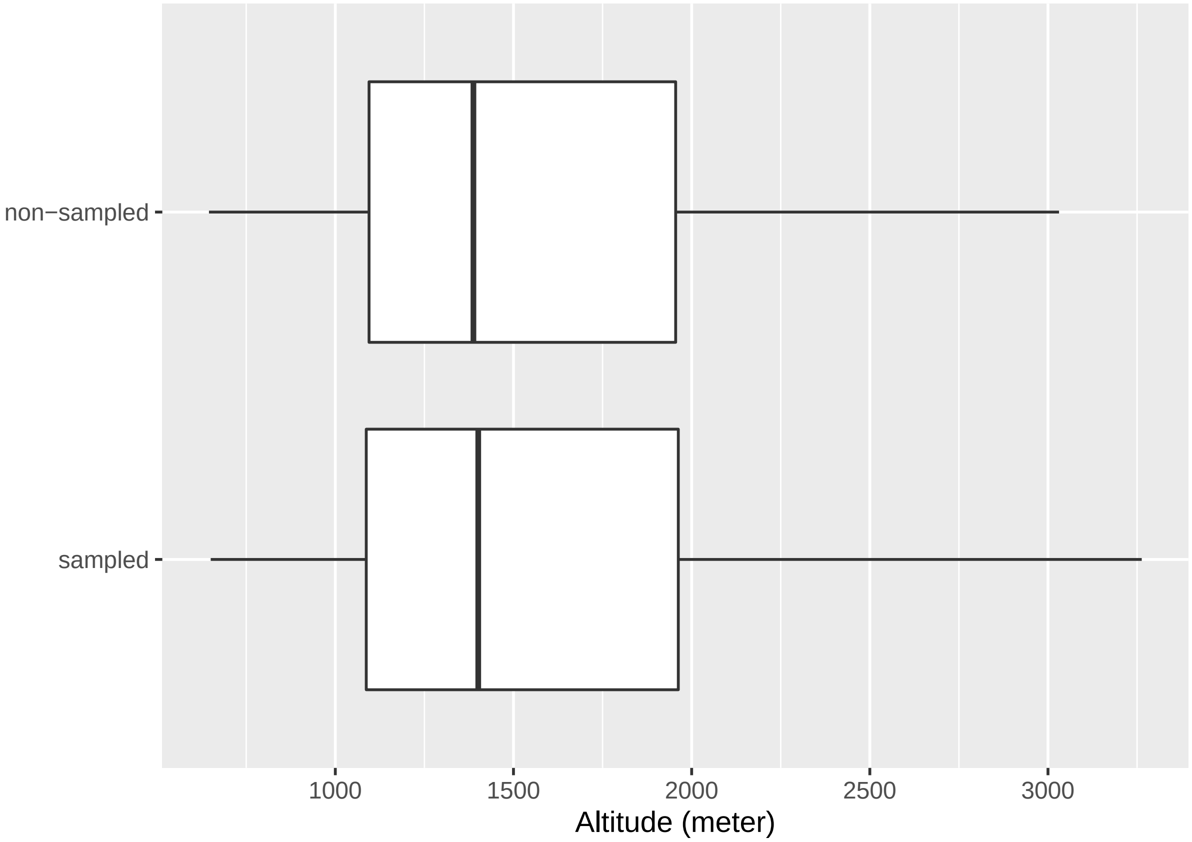
```
ggplot(afghan, aes(y = educ.years, x = province)) +
  geom_violin() +
  coord_flip() +
  labs(x= "Province", y = "Years of education", title = "Education by Province")
```



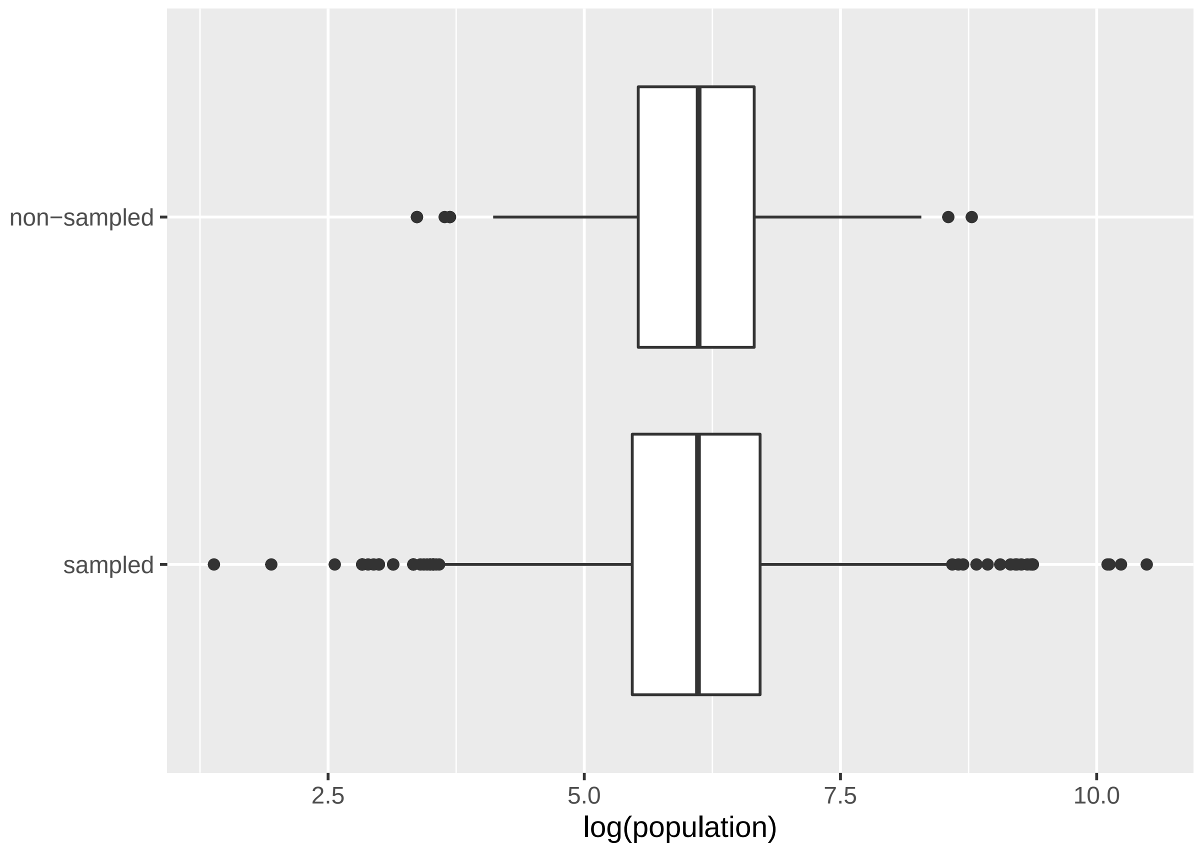
3.4 Survey Sampling

```r
data("afghan.village", package = "qss")
```

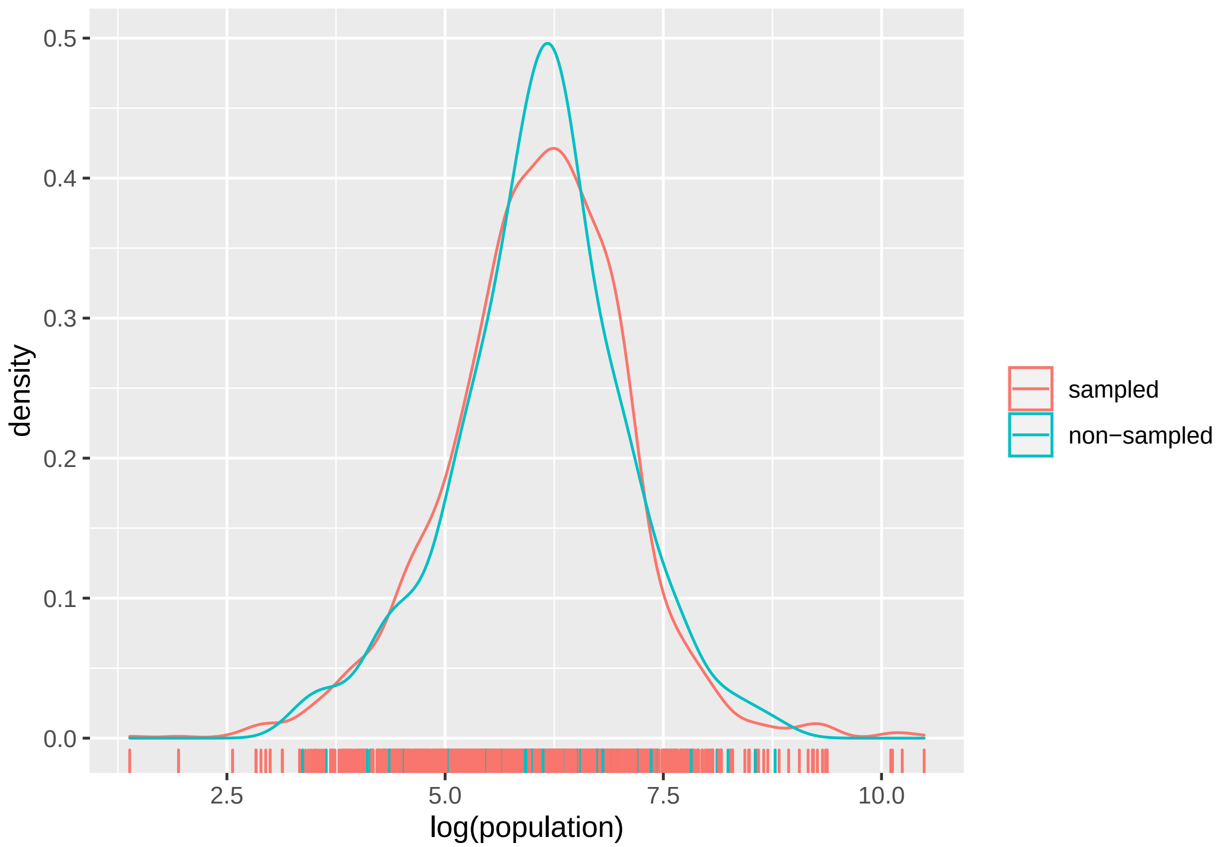
```r
ggplot(afghan.village, aes(x = factor(village.surveyed,
                                      labels = c("sampled", "non-sampled")),
                           y = altitude)) +
  geom_boxplot() +
  labs(y = "Altitude (meter)", x = "") +
  coord_flip()
```



```r
ggplot(afghan.village, aes(x = factor(village.surveyed,
                                      labels = c("sampled", "non-sampled")),
                           y = log(population))) +
  geom_boxplot() +
  labs(y = "log(population)", x = "") +
  coord_flip()
```



```r
ggplot(afghan.village, aes(colour = factor(village.surveyed,
                                          labels = c("sampled", "non-sampled")),
                            x = log(population))) +
  geom_density() +
  geom_rug() +
  labs(x = "log(population)", colour = "")
```



```r
afghan %>%
  group_by(province) %>%
  summarise(ISAF = mean(is.na(violent.exp.ISAF)),
            taliban = mean(is.na(violent.exp.taliban))) %>%
  arrange(-ISAF)
```

```
## # A tibble: 5 x 3
##   province    ISAF taliban
##   <chr>      <dbl>   <dbl>
## 1 Uruzgan  0.0207  0.0620
## 2 Helmand  0.0164  0.0304
## 3 Khost    0.00476 0.00635
## 4 Kunar    0       0
## 5 Logar    0       0
```

```r
(mean(filter(afghan, list.group == "ISAF")$list.response) -
  mean(filter(afghan, list.group == "control")$list.response))
```

```
## [1] 0.04901961
```

```r
afghan %>%
  group_by(list.response, list.group) %>%
  count() %>%
  glimpse() %>%
  spread(list.group, n, fill = 0)
```

```
## Rows: 12
## Columns: 3
## Groups: list.response, list.group [12]
## $ list.response <int> 0, 0, 1, 1, 1, 2, 2, 2, 3, 3, 3, 4
## $ list.group    <chr> "control", "ISAF", "control", "ISAF", "taliban", "contro~
## $ n             <int> 188, 174, 265, 278, 433, 265, 260, 287, 200, 182, 198, 24
```

```
## # A tibble: 5 x 4
## # Groups:   list.response [5]
##   list.response control  ISAF taliban
##           <int>   <dbl> <dbl>   <dbl>
## 1             0     188   174       0
## 2             1     265   278     433
## 3             2     265   260     287
## 4             3     200   182     198
## 5             4       0    24       0
```

3.5 Measuring Political Polarization

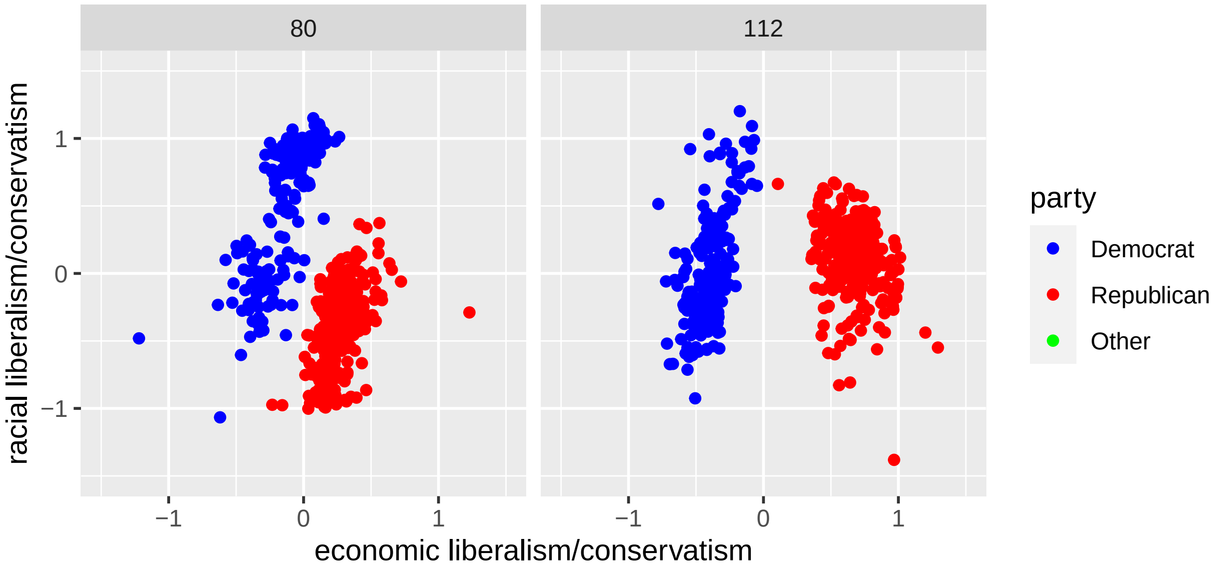
```
data("congress", package = "qss")
```

```
glimpse(congress)
```

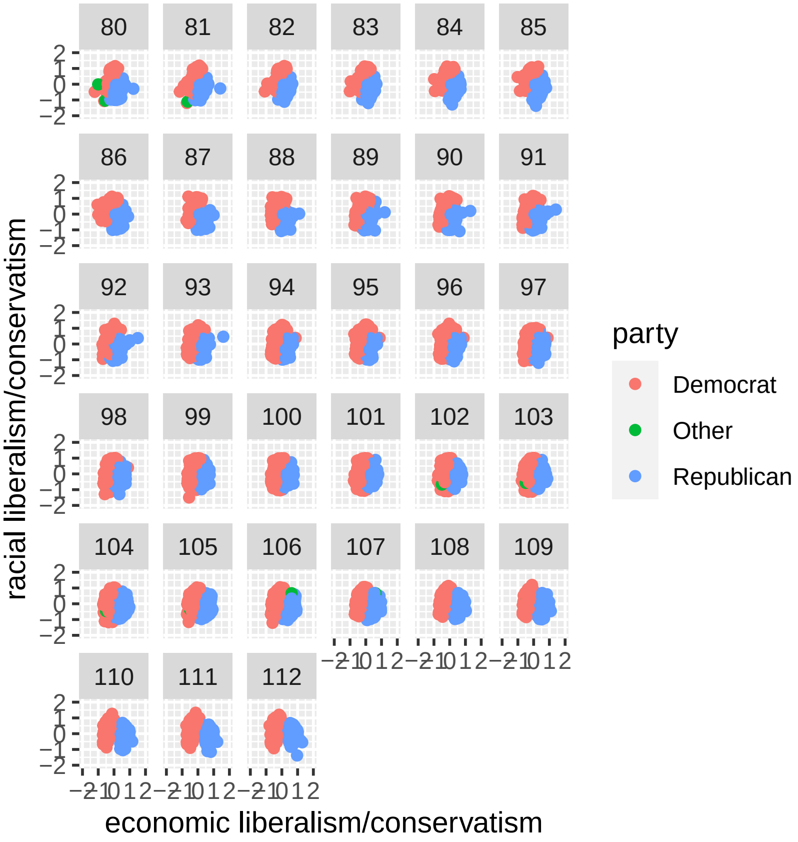
```
## Rows: 14,552
## Columns: 7
## $ congress <int> 80, 80, 80, 80, 80, 80, 80, 80, 80, 80, 80, 80, 80, 80, 80, 8~
## $ district <int> 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 98, 98, 1, 2, 3, 4, 5, 6, 7, 1,~
## $ state    <chr> "USA", "ALABAMA", "ALABAMA", "ALABAMA", "ALABAMA", "ALABAMA",~
## $ party    <chr> "Democrat", "Democrat", "Democrat", "Democrat", "Democrat", "~
## $ name     <chr> "TRUMAN", "BOYKIN  F.", "GRANT  G.", "ANDREWS  G.", "HOBBS  S~
## $ dwnom1   <dbl> -0.276, -0.026, -0.042, -0.008, -0.082, -0.170, -0.124, -0.03~
## $ dwnom2   <dbl> 0.016, 0.796, 0.999, 1.005, 1.066, 0.870, 0.990, 0.892, 0.888~
```

```
q <-
  congress %>%
  filter(congress %in% c(80, 112),
         party %in% c("Democrat", "Republican")) %>%
  ggplot(aes(x = dwnom1, y = dwnom2, colour = party)) +
  geom_point() +
  facet_wrap(~ congress) +
  coord_fixed() +
  scale_y_continuous("racial liberalism/conservatism",
                     limits = c(-1.5, 1.5)) +
  scale_x_continuous("economic liberalism/conservatism",
                     limits = c(-1.5, 1.5))
```

```
scale_colour_parties <-
  scale_colour_manual(values = c(Democrat = "blue",
                                 Republican = "red",
                                 Other = "green"))
q + scale_colour_parties
```

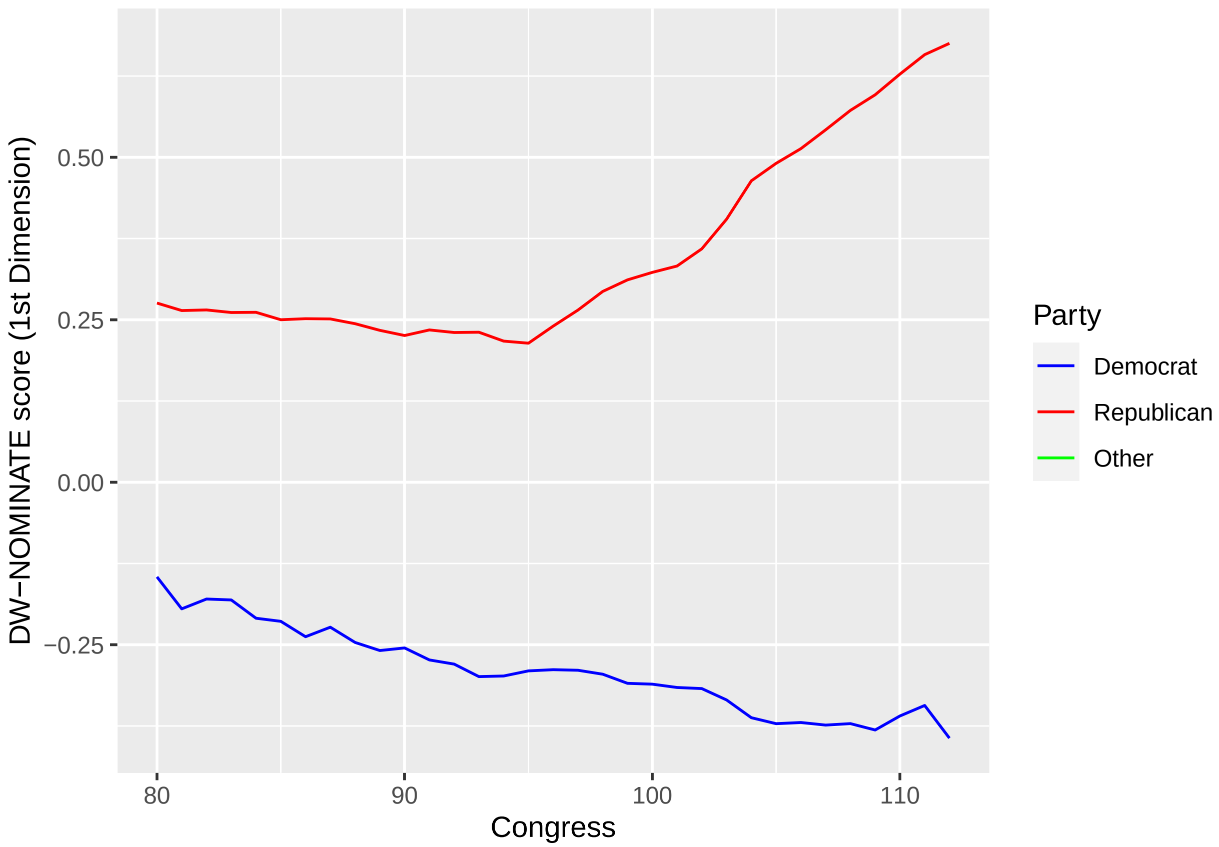


```r
congress %>%
  ggplot(aes(x = dwnom1, y = dwnom2, colour = party)) +
  geom_point() +
  facet_wrap(~ congress) +
  coord_fixed() +
  scale_y_continuous("racial liberalism/conservatism",
                     limits = c(-2, 2)) +
  scale_x_continuous("economic liberalism/conservatism",
                     limits = c(-2, 2))
```



```
congress %>%
  group_by(congress, party) %>%
  summarise(dwnom1 = mean(dwnom1)) %>%
  filter(party %in% c("Democrat", "Republican")) %>%
  ggplot(aes(x = congress, y = dwnom1,
             colour = fct_reorder2(party, congress, dwnom1))) +
  geom_line() +
  scale_colour_parties +
  labs(y = "DW-NOMINATE score (1st Dimension)", x = "Congress",
       colour = "Party")
```

```
## ‘summarise()‘ has grouped output by ’congress’. You can override using the ‘.groups‘ argument.
```

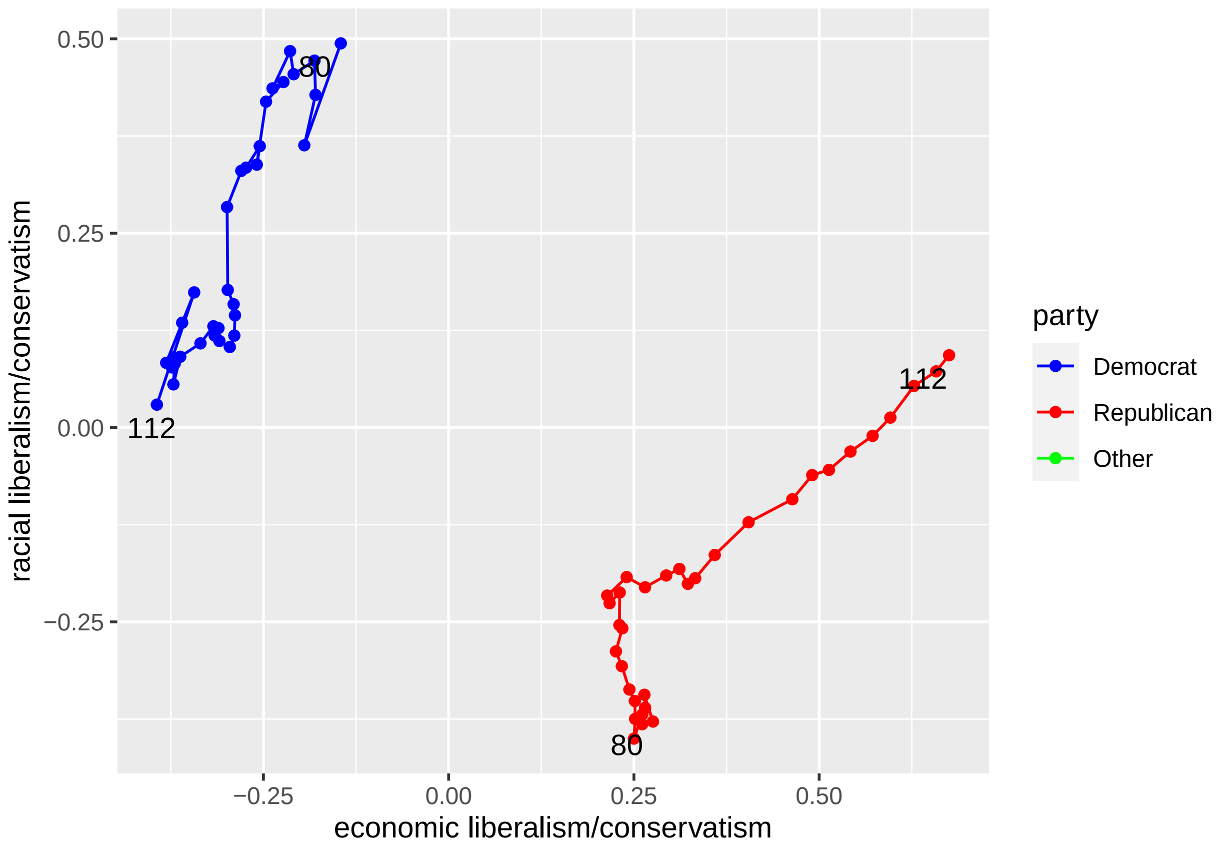


```r
party_means <-
  congress %>%
  filter(party %in% c("Democrat", "Republican")) %>%
  group_by(party, congress) %>%
  summarise(dwnom1 = mean(dwnom1),
            dwnom2 = mean(dwnom2))
```

```
## ‘summarise()‘ has grouped output by ’party’. You can override using the ‘.groups‘ argument.
```

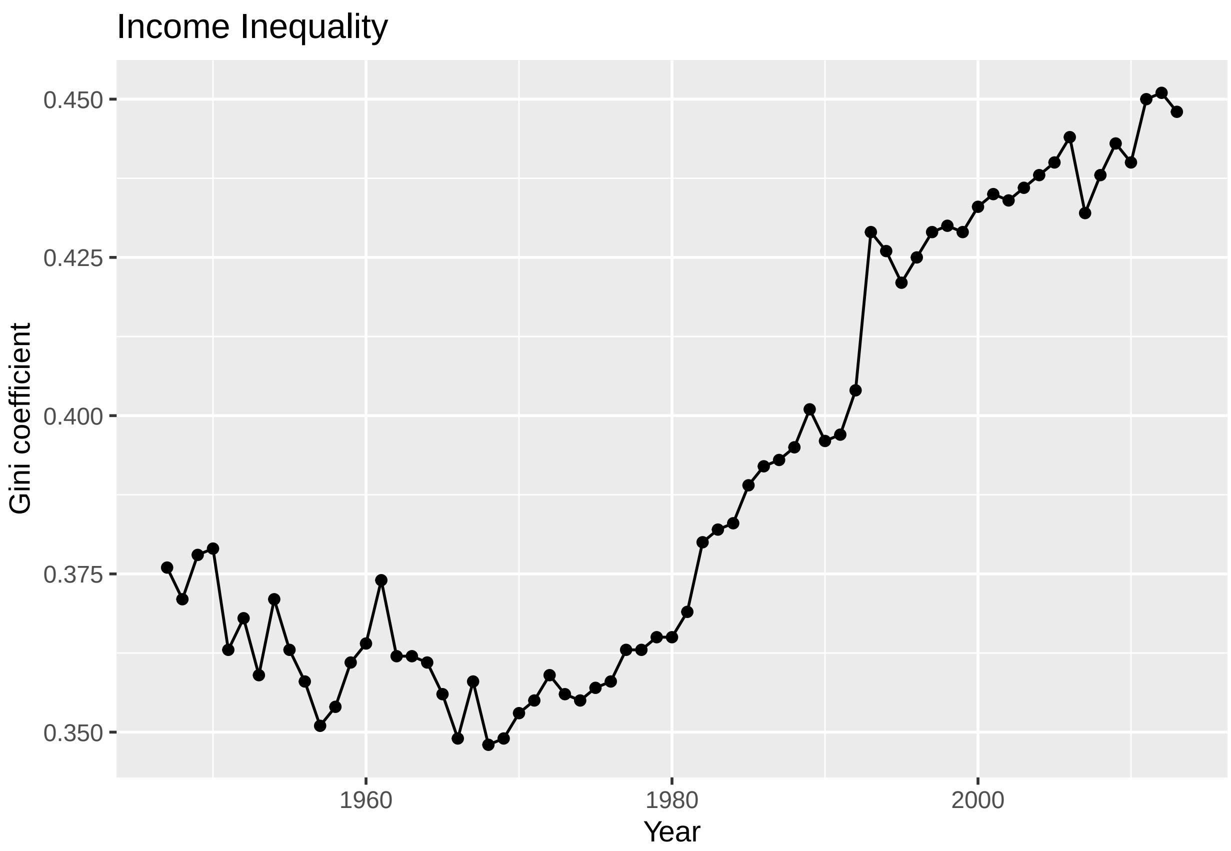
```r
party_endpoints <-
  party_means %>%
  filter(congress %in% c(min(congress), max(congress))) %>%
  mutate(label = str_c(party, congress, sep = " - "))

ggplot(party_means,
         aes(x = dwnom1, y = dwnom2, color = party,
             group = party)) +
  geom_point() +
  geom_path() +
  ggrepel::geom_text_repel(data = party_endpoints,
                           mapping = aes(label = congress),
                           color = "black") +
  scale_y_continuous("racial liberalism/conservatism") +
  scale_x_continuous("economic liberalism/conservatism") +
  scale_colour_parties
```



```
data("USGini", package = "qss")
```

```
ggplot(USGini, aes(x = year, y = gini)) +
  geom_point() +
  geom_line() +
  labs(x = "Year", y = "Gini coefficient") +
  ggtitle("Income Inequality")
```



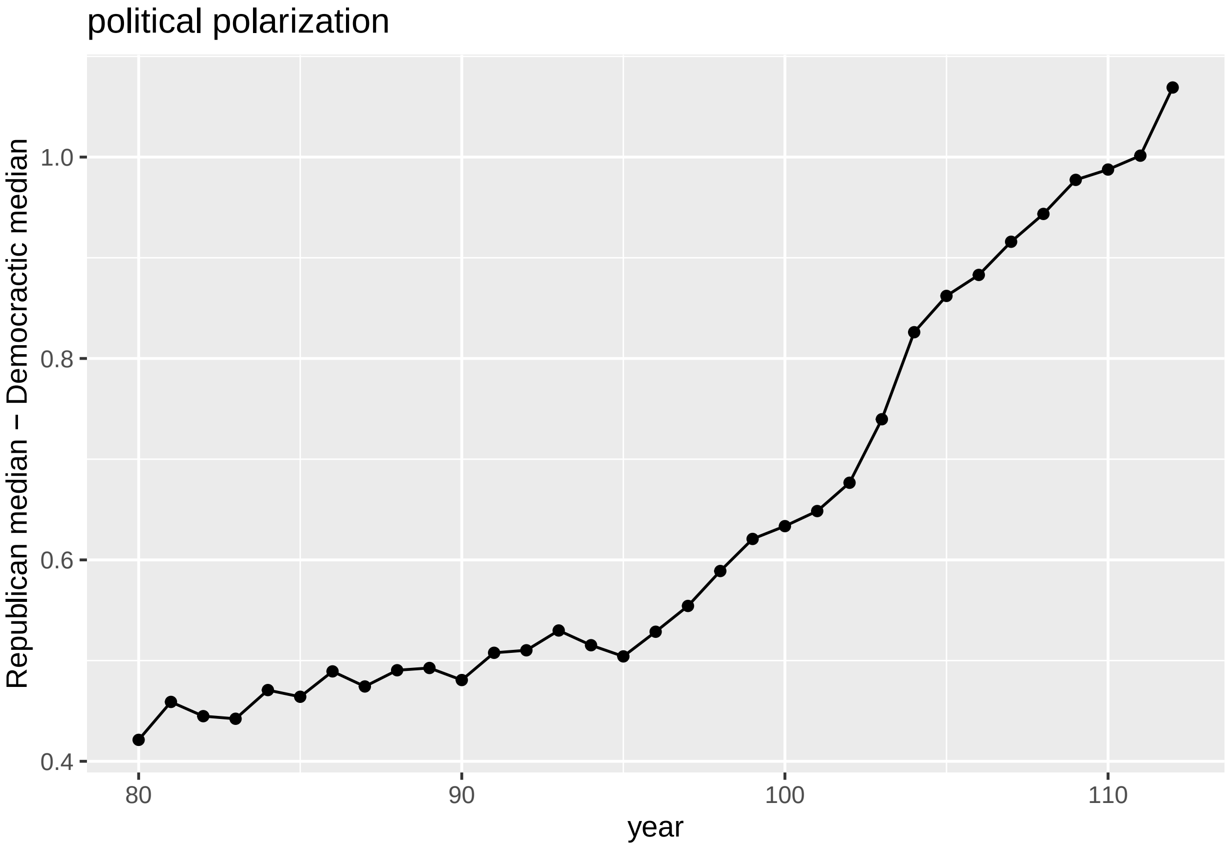
```
party_polarization <-
  congress %>%
  group_by(congress, party) %>%
  summarise(dwnom1 = mean(dwnom1)) %>%
  filter(party %in% c("Democrat", "Republican")) %>%
  spread(party, dwnom1) %>%
  mutate(polarization = Republican - Democrat)
```

```
## ‘summarise()‘ has grouped output by ’congress’. You can override using the ‘.groups‘ argument.
```

```
party_polarization
```

```
## # A tibble: 33 x 4
## # Groups:   congress [33]
##    congress Democrat Republican polarization
##       <int>    <dbl>      <dbl>        <dbl>
##  1       80   -0.146      0.276        0.421
##  2       81   -0.195      0.264        0.459
##  3       82   -0.180      0.265        0.445
##  4       83   -0.181      0.261        0.442
##  5       84   -0.209      0.261        0.471
##  6       85   -0.214      0.250        0.464
##  7       86   -0.238      0.252        0.489
##  8       87   -0.223      0.251        0.474
##  9       88   -0.246      0.244        0.490
## 10       89   -0.259      0.234        0.493
## # ... with 23 more rows
```

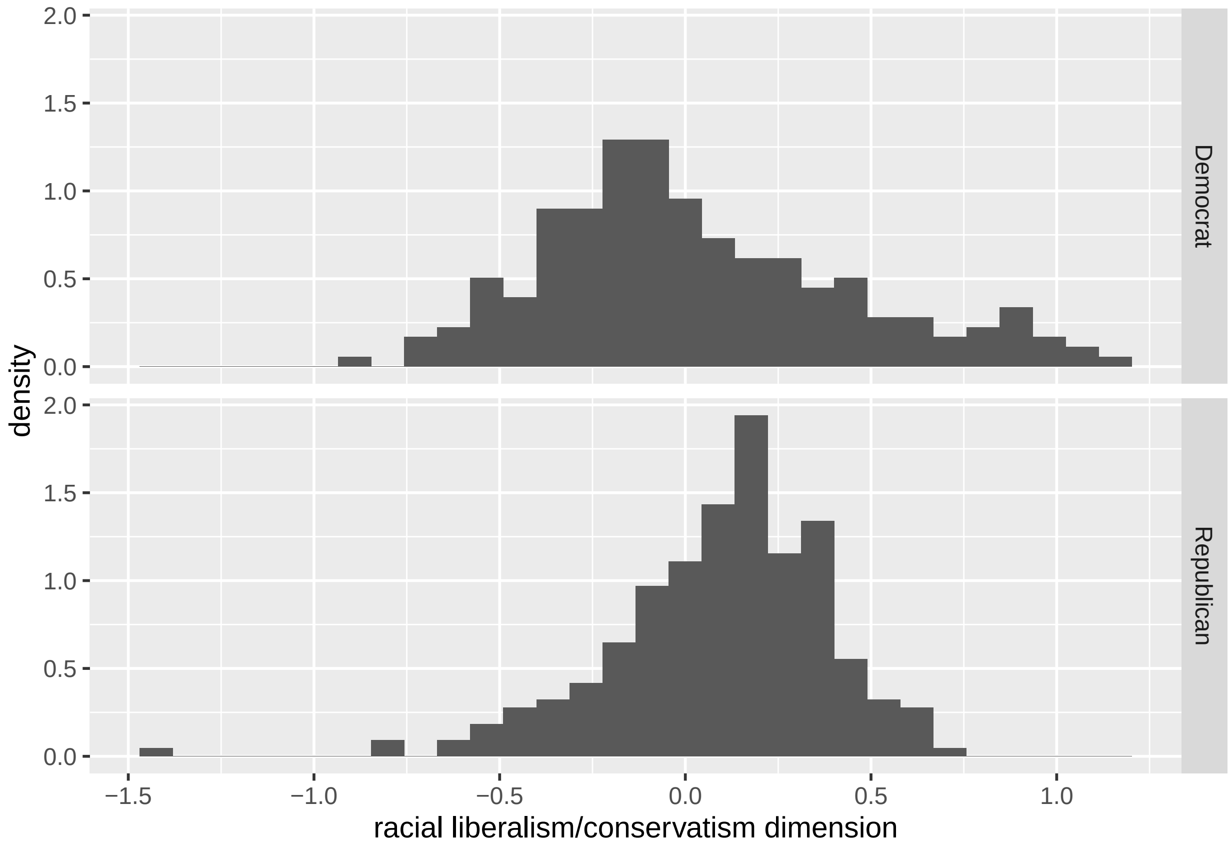
```
ggplot(party_polarization, aes(x = congress, y = polarization)) +
  geom_point() +
  geom_line() +
  ggtitle("political polarization") +
  labs(x = "year", y = "Republican median - Democractic median")
```



```
congress %>%
  filter(congress == 112, party %in% c("Republican", "Democrat")) %>%
  ggplot(aes(x = dwnom2, y = ..density..)) +
  geom_histogram(bindwith = 0.2) +
  facet_grid(party ~ .) +
  labs(x = "racial liberalism/conservatism dimension")
```

```
## Warning: Ignoring unknown parameters: bindwith
```

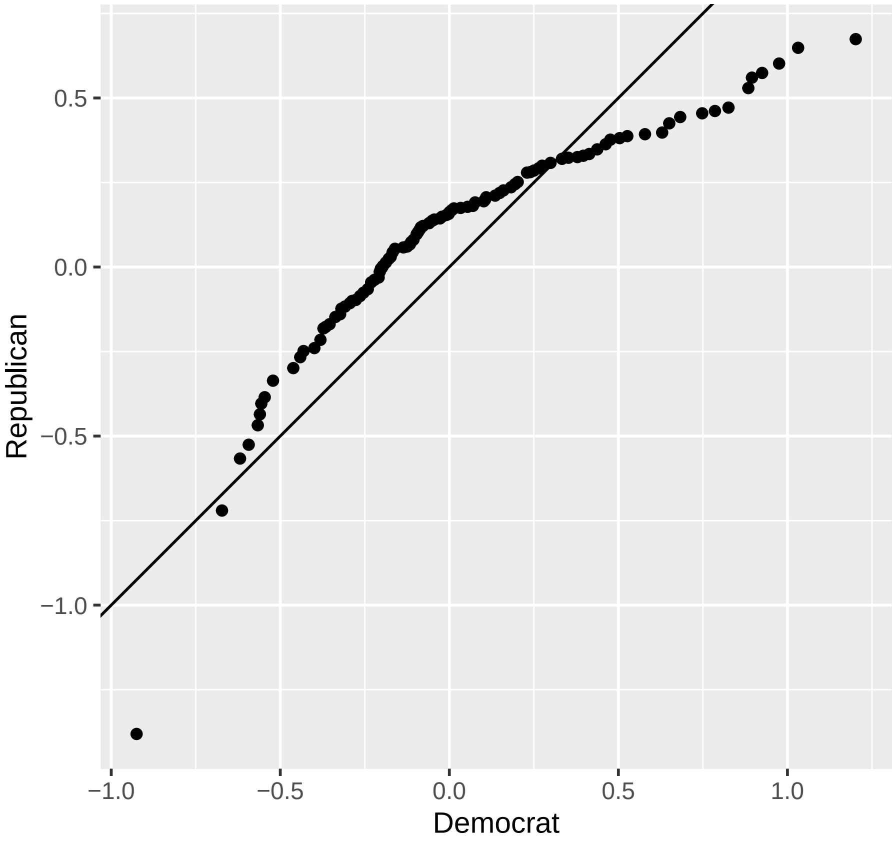
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



```
party_qtiles <- tibble(
  probs = seq(0, 1, by = 0.01),
  Democrat = quantile(filter(congress, congress == 112,
                             party == "Democrat")$dwnom2,
         probs = probs),
  Republican = quantile(filter(congress, congress == 112,
                               party == "Republican")$dwnom2,
         probs = probs)
)
party_qtiles
```

```
## # A tibble: 101 x 3
##    probs Democrat Republican
##    <dbl>    <dbl>      <dbl>
##  1  0      -0.925     -1.38
##  2  0.01   -0.672     -0.720
##  3  0.02   -0.619     -0.566
##  4  0.03   -0.593     -0.526
##  5  0.04   -0.567     -0.468
##  6  0.05   -0.560     -0.436
##  7  0.06   -0.556     -0.404
##  8  0.07   -0.546     -0.385
##  9  0.08   -0.522     -0.336
## 10  0.09   -0.462     -0.299
## # ... with 91 more rows
```

```
party_qtiles %>%
  ggplot(aes(x = Democrat, y = Republican)) +
  geom_point() +
  geom_abline() +
  coord_fixed()
```



3.6 Clustering

```
k80two.out <-
  kmeans(select(filter(congress, congress == 80),
                     dwnom1, dwnom2),
              centers = 2, nstart = 5)
```

```
congress80 <-
  congress %>%
  filter(congress == 80) %>%
  mutate(cluster2 = factor(k80two.out$cluster))
```

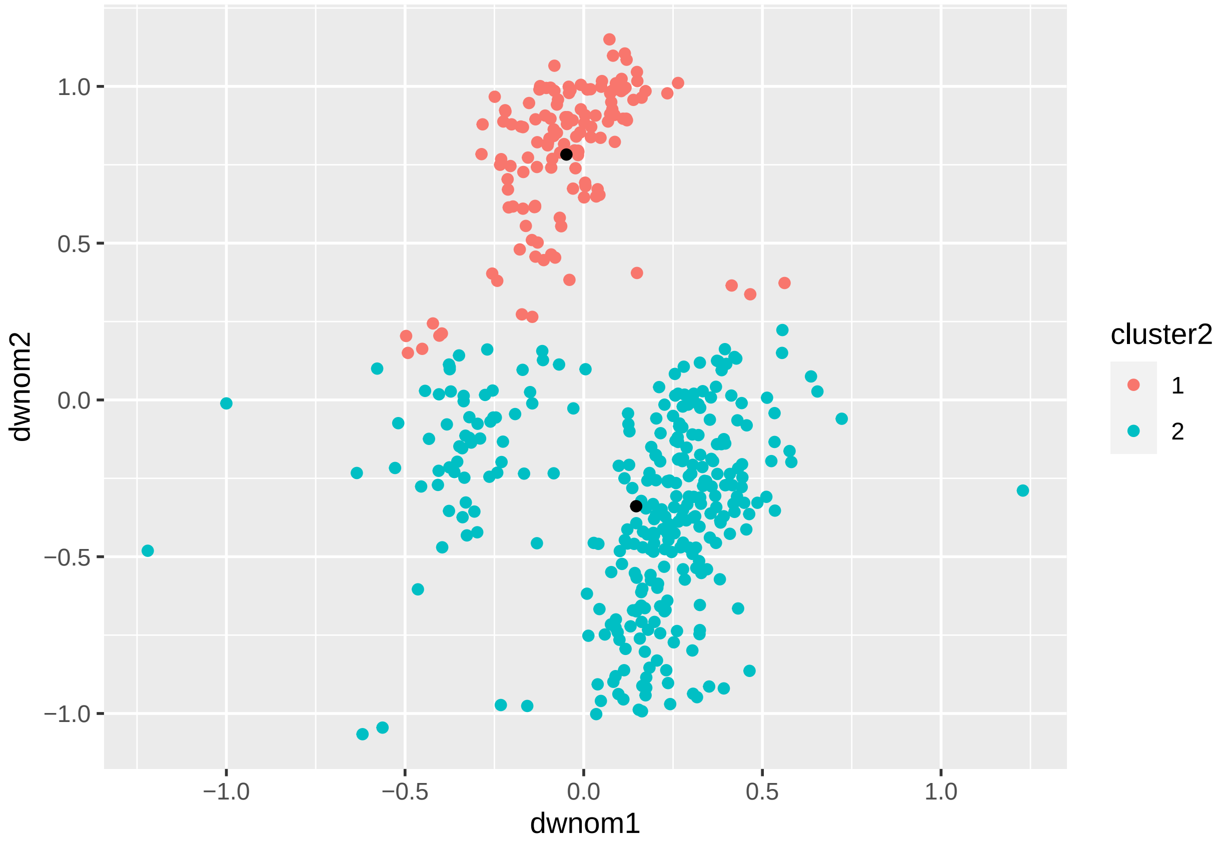
```
k80two.out$centers
```

```
##        dwnom1     dwnom2
## 1 -0.04843704  0.7827259
## 2  0.14681029 -0.3389293
```

```
k80two.clusters <- tidy(k80two.out)
k80two.clusters
```

```
## # A tibble: 2 x 5
##    dwnom1 dwnom2  size withinss cluster
##     <dbl>  <dbl> <int>    <dbl> <fct>
## 1 -0.0484  0.783   135     10.9 1
## 2  0.147  -0.339   311     54.9 2
```

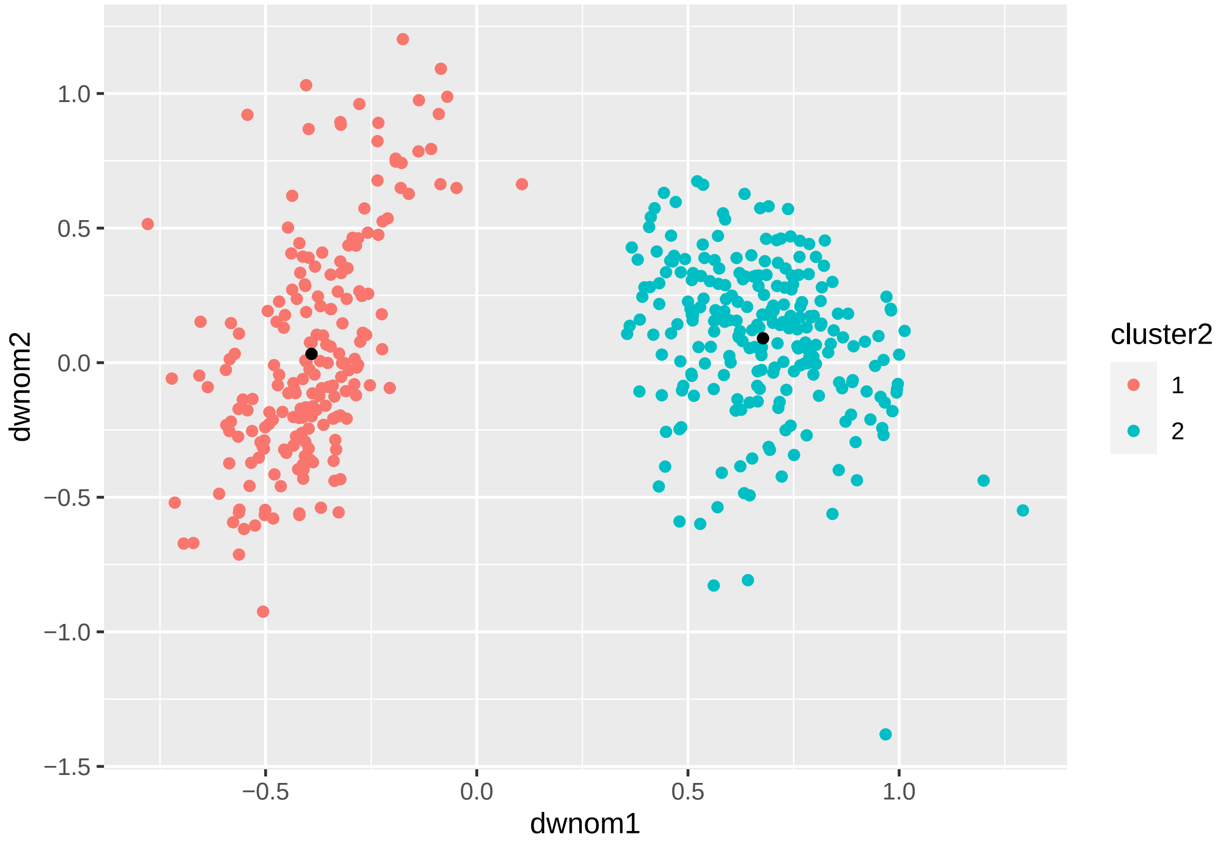
```
ggplot() +
  geom_point(data = congress80,
             aes(x = dwnom1, y = dwnom2, colour = cluster2)) +
  geom_point(data = k80two.clusters, mapping = aes(x = dwnom1, y = dwnom2))
```



```
congress80 %>%
  group_by(party, cluster2) %>%
  count()
```

```
## # A tibble: 5 x 3
## # Groups:   party, cluster2 [5]
##   party      cluster2     n
##   <chr>      <fct>    <int>
## 1 Democrat   1          132
## 2 Democrat   2           62
## 3 Other      2            2
## 4 Republican 1            3
## 5 Republican 2          247
```

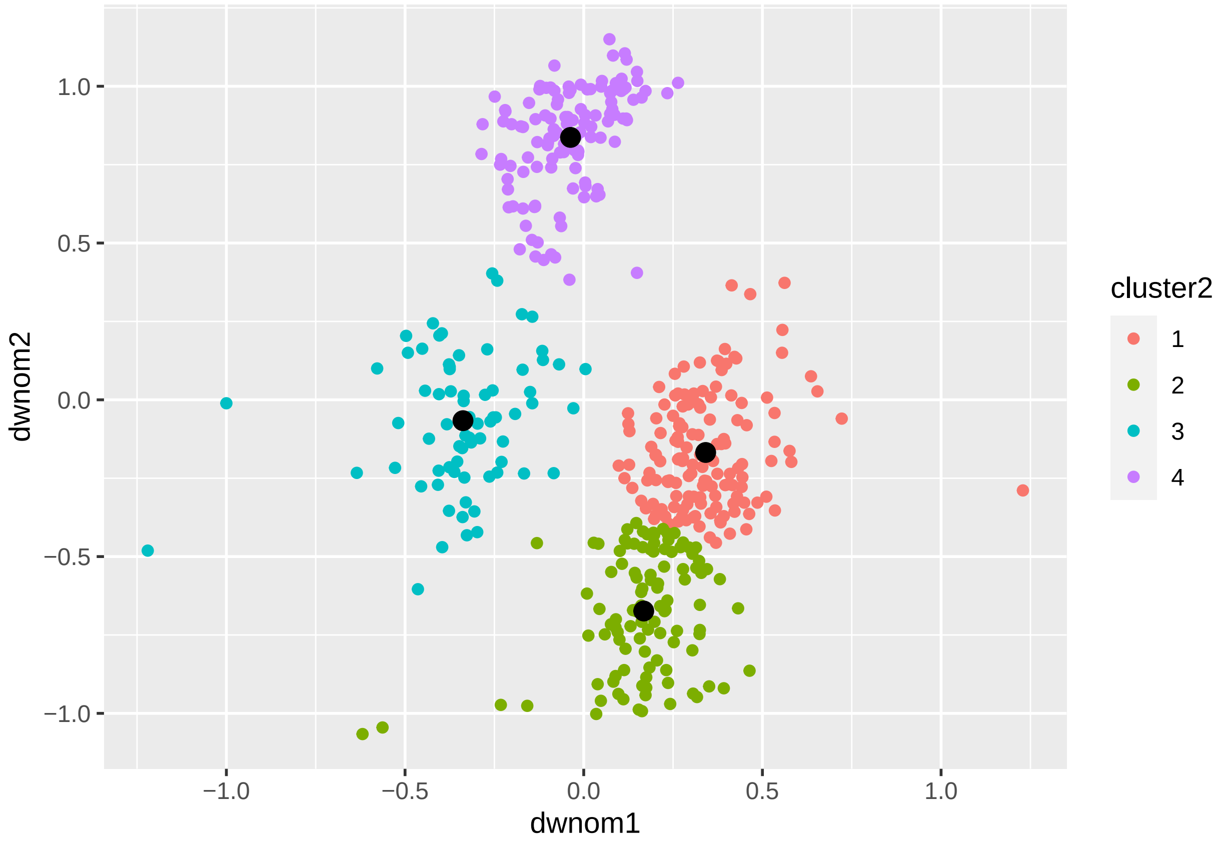
```
k112two.out <-
  kmeans(select(filter(congress, congress == 112),
                dwnom1, dwnom2),
         centers = 2, nstart = 5)
congress112 <-
  filter(congress, congress == 112) %>%
  mutate(cluster2 = factor(k112two.out$cluster))
k112two.clusters <- tidy(k112two.out)
ggplot() +
  geom_point(data = congress112,
             mapping = aes(x = dwnom1, y = dwnom2, colour = cluster2)) +
  geom_point(data = k112two.clusters,
             mapping = aes(x = dwnom1, y = dwnom2))
```



```
congress112 %>%
  group_by(party, cluster2) %>%
  count()
```

```
## # A tibble: 3 x 3
## # Groups:   party, cluster2 [3]
##   party      cluster2     n
##   <chr>      <fct>    <int>
## 1 Democrat   1          200
## 2 Republican 1            1
## 3 Republican 2          242
```

```
k80four.out <-
  kmeans(select(filter(congress, congress == 80),
                dwnom1, dwnom2),
         centers = 4, nstart = 5)
congress80 <-
  filter(congress, congress == 80) %>%
  mutate(cluster2 = factor(k80four.out$cluster))
k80four.clusters <- tidy(k80four.out)
ggplot() +
  geom_point(data = congress80,
             mapping = aes(x = dwnom1, y = dwnom2, colour = cluster2)) +
  geom_point(data = k80four.clusters,
             mapping = aes(x =dwnom1, y = dwnom2,), size = 3)
```


dwnom2
dwnom1
cluster2