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"*

```
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
```

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

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

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

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

```
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
```

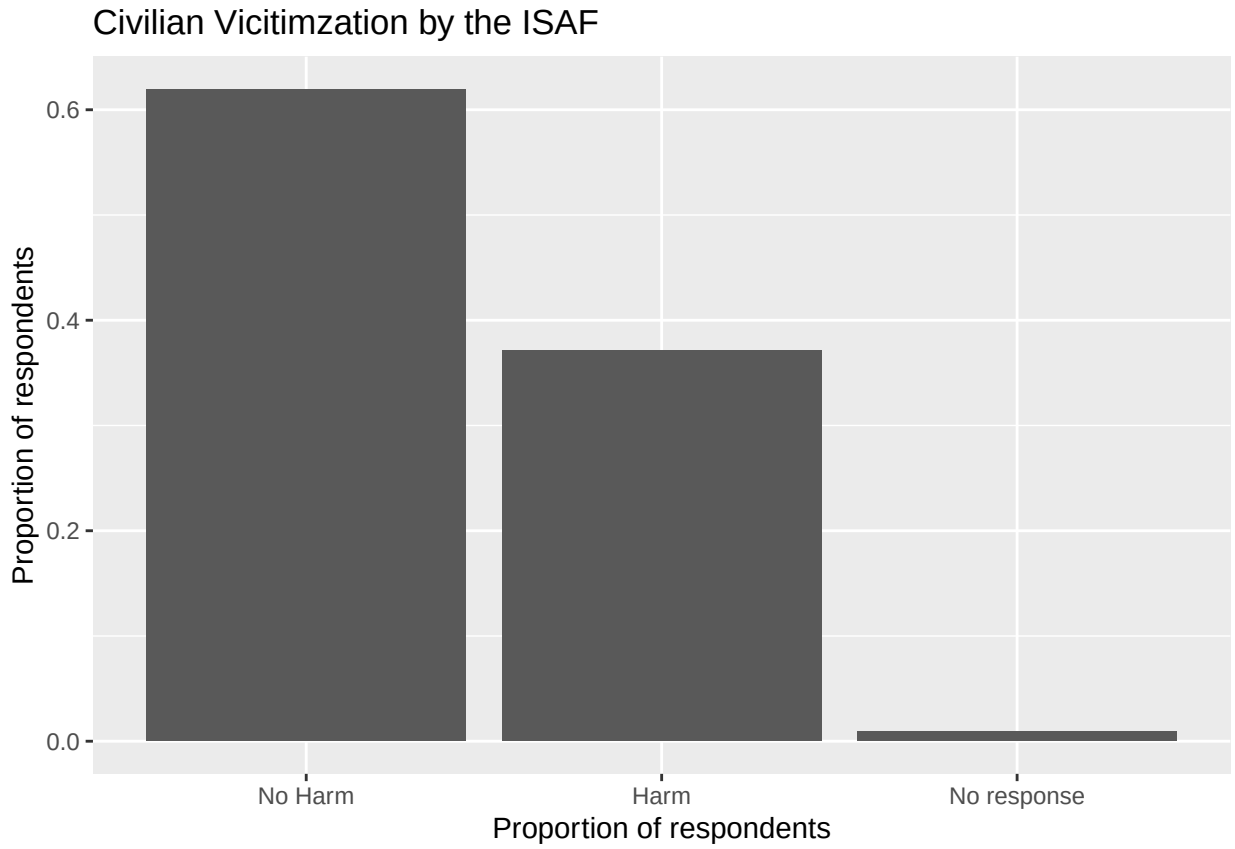
```
x <- 1:5
class(x)
if_else(x<3, x,NA)
```

```
if_else(x < 3, x, NA_integer_)
```

3.3 Visualizing the Univariate Distribution

3.3.1 Barplot

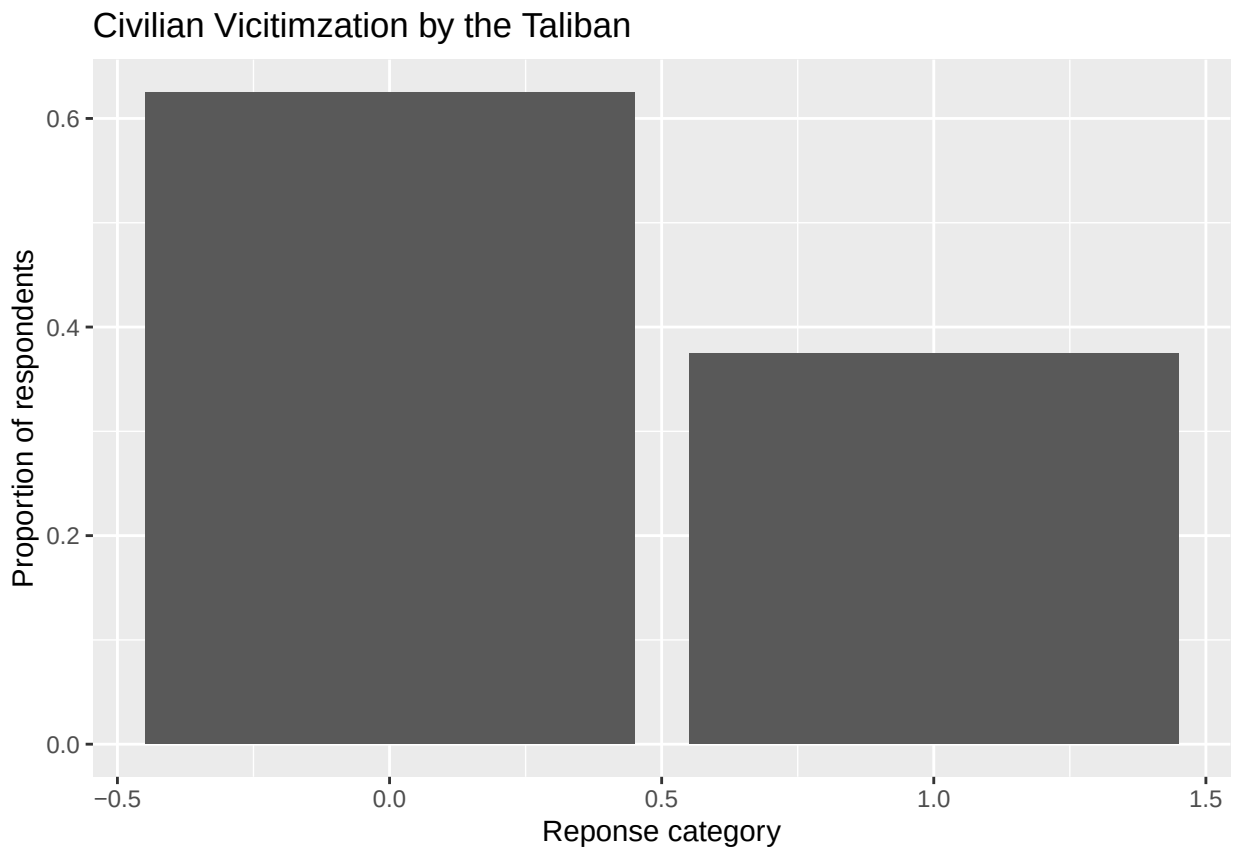
```
afghan <-
  afghan %>%
  mutate(violent.exp.ISAF.fct =
           fct_explicit_na(fct_recode(factor(violent.exp.ISAF), Harm = "1", "No Harm" = "0"),
                           "No response"))
ggplot(afghan, aes(x = violent.exp.ISAF.fct, y = ..prop.., group = 1)) +
  geom_bar() +
  xlab("Proportion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



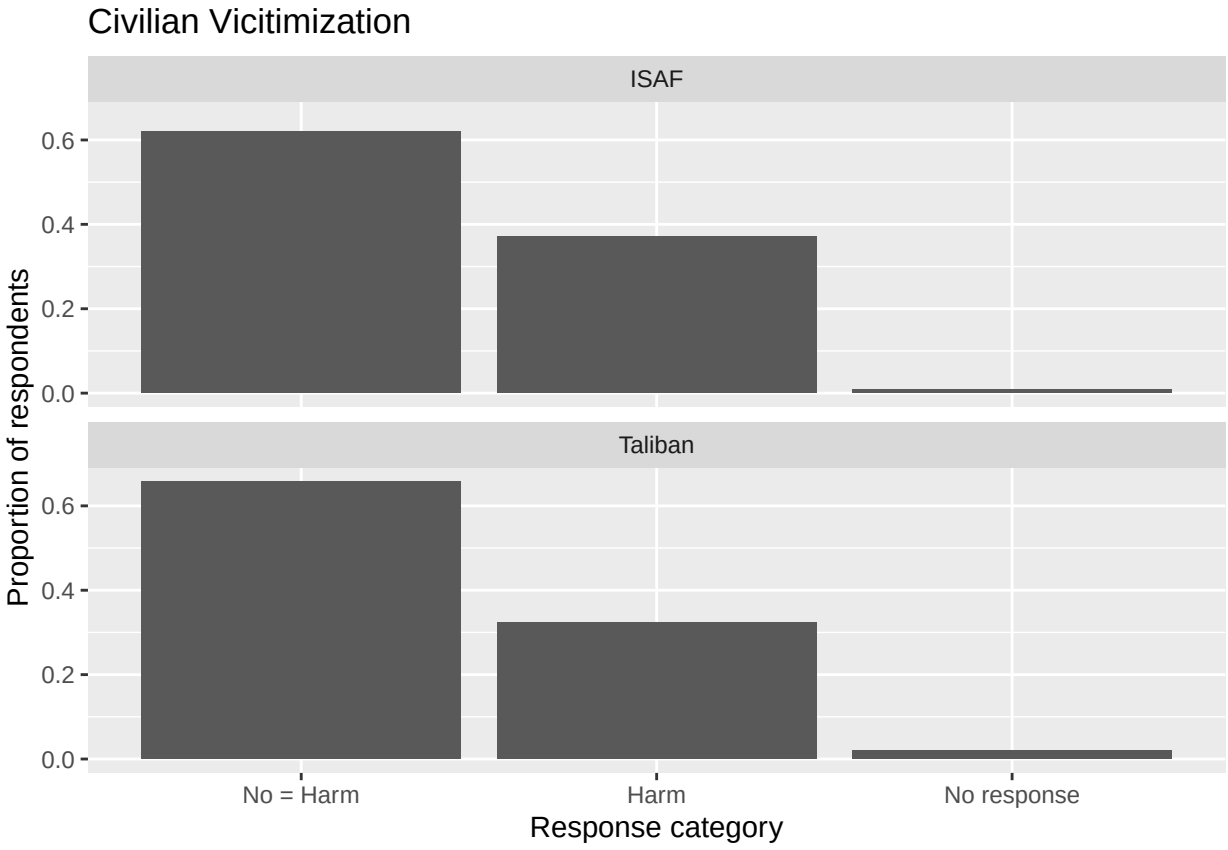

```
afghan <-
  afghan %>%
  mutate(violent.exp.taliban.fct =
           fct_explicit_na(fct_recode(factor(violent.exp.taliban), Harm = "1", "No Harm" = "0"),
                           "No response"))
```

```
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")
```

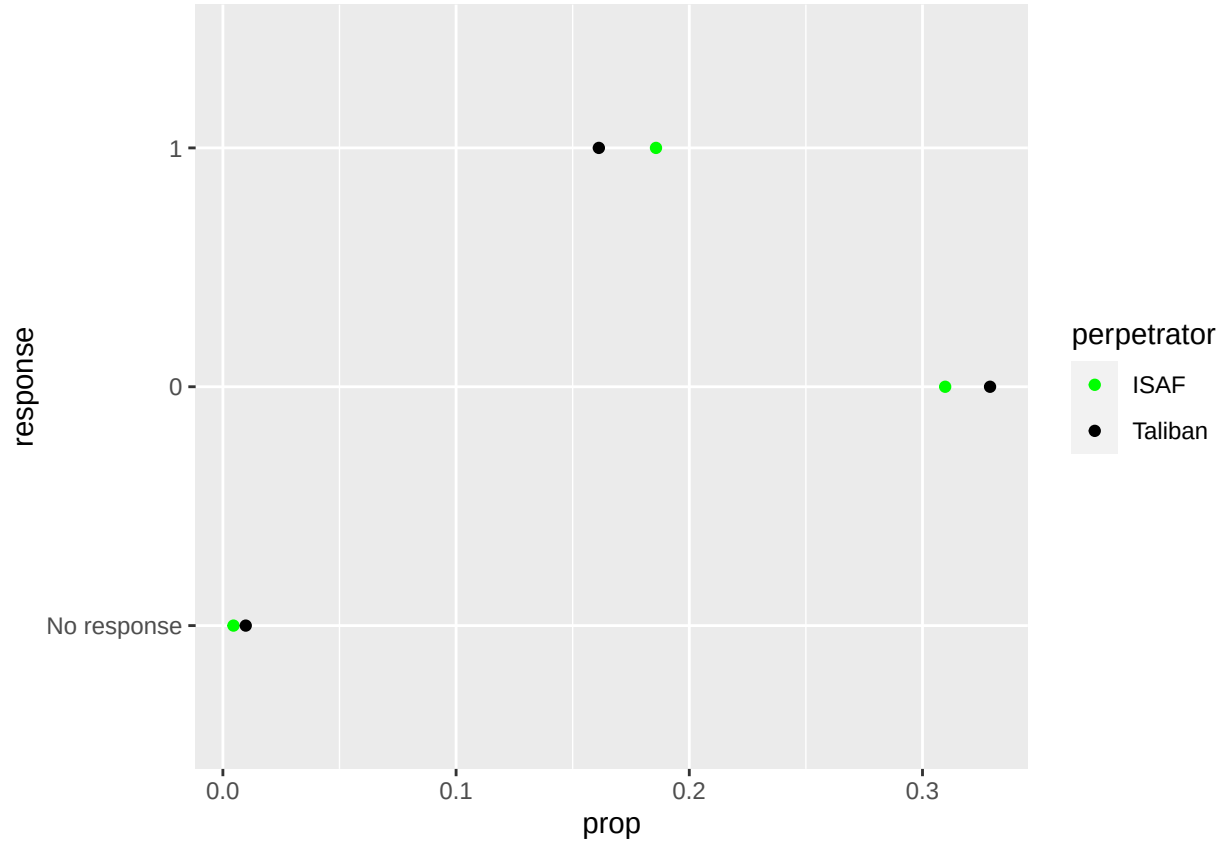



```
violent_exp <-
  afghan %>%
  select(violent.exp.ISAF, violent.exp.taliban) %>%
  gather(perpetrator, response) %>%
  mutate(perpetrator = str_replace(perpetrator, "violent\\.exp\\.",""),
         perpetrator = str_replace(perpetrator, "taliban", "Taliban"),
         response = fct_recode(factor(response), "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
```

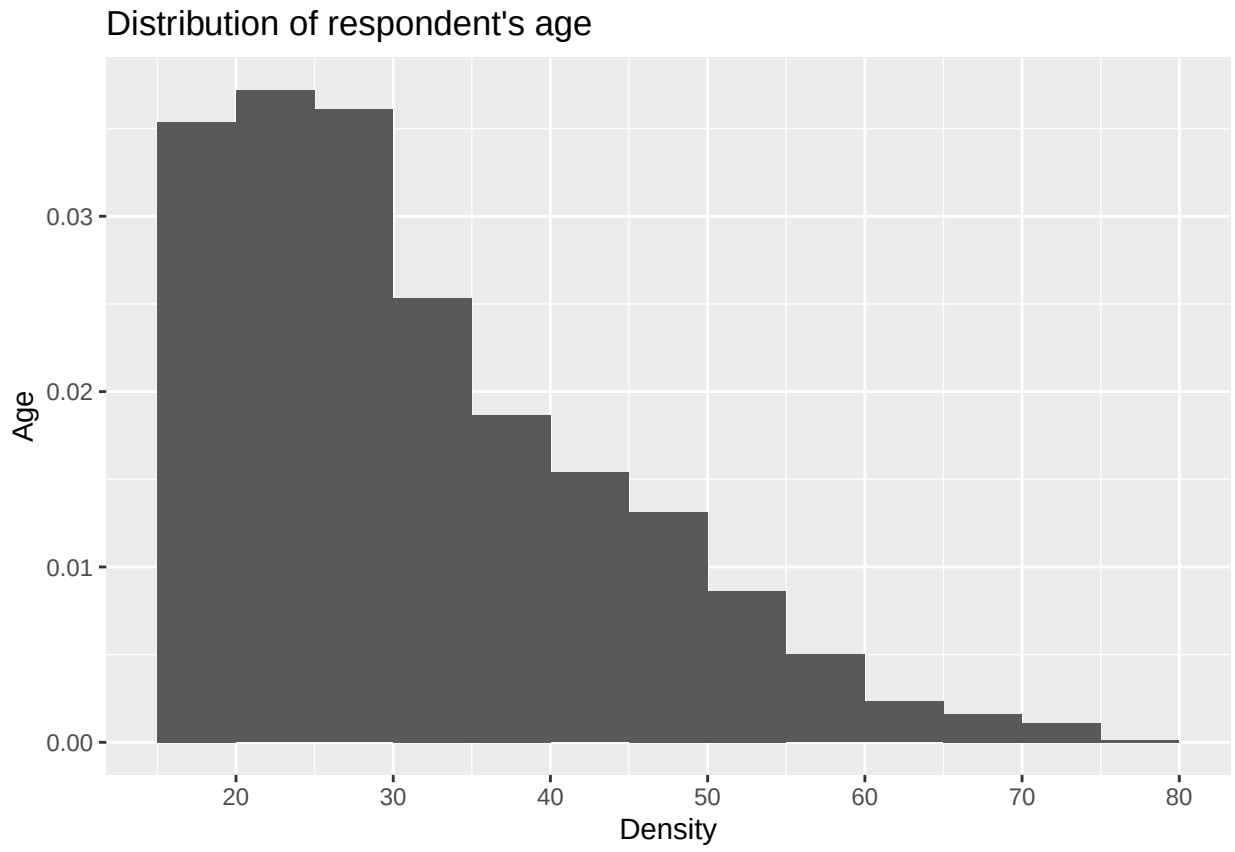
```
## Warning: Unknown levels in `f`: No Harm
```

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

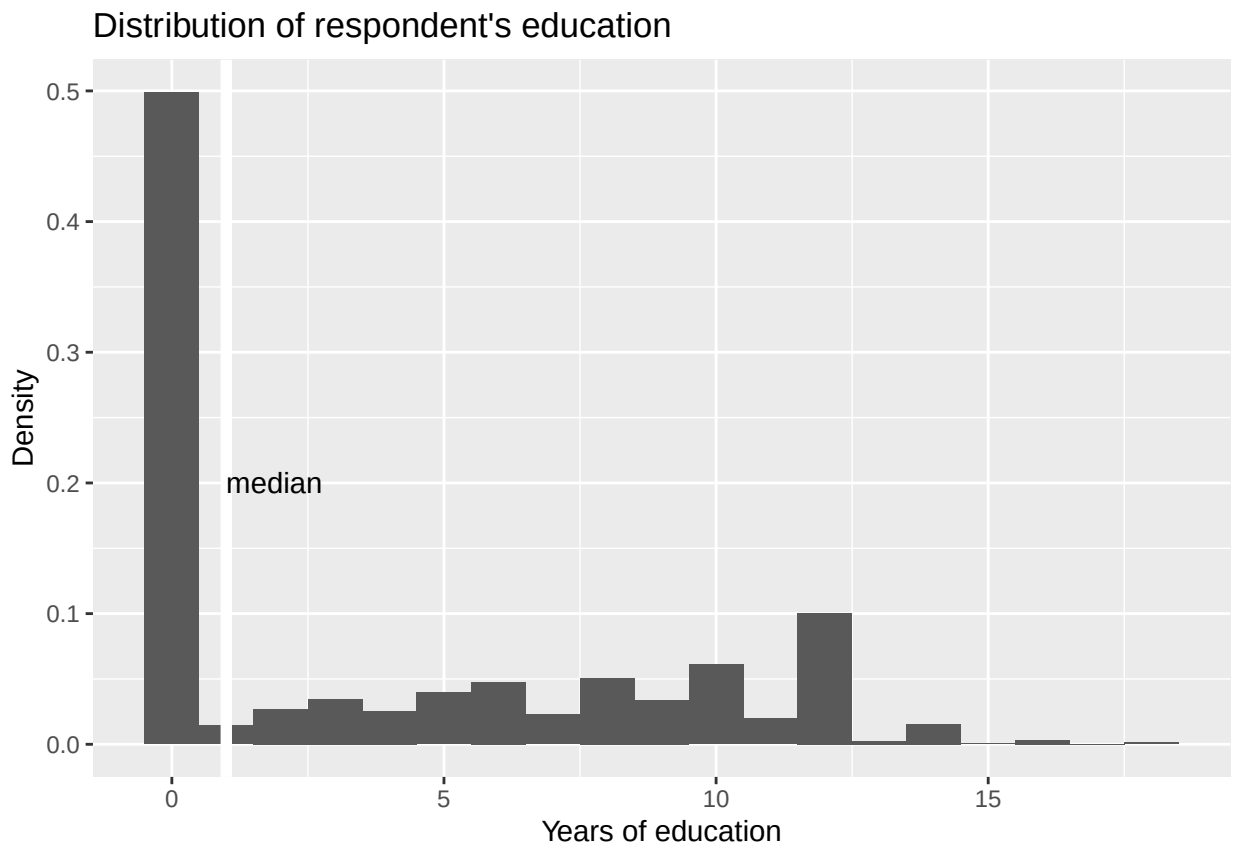


3.3.2 Histogram

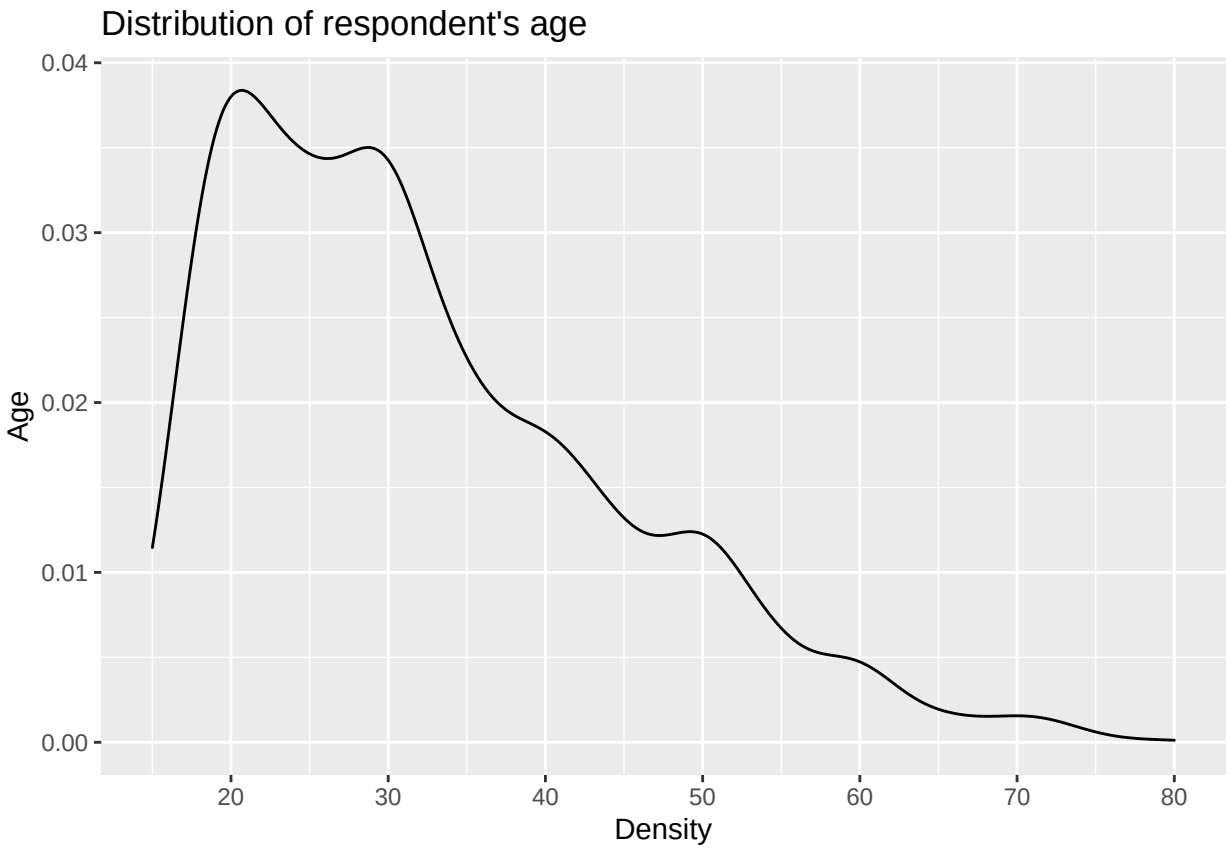
```
ggplot(afghan, aes(x = age, y = ..density..)) +
  geom_histogram(binwidth = 5, boundary = 0) +
  scale_x_continuous(breaks = seq(20, 80, by = 10)) +
  labs(title = "Distribution of respondent's age",
       y = "Age", x = "Density")
```



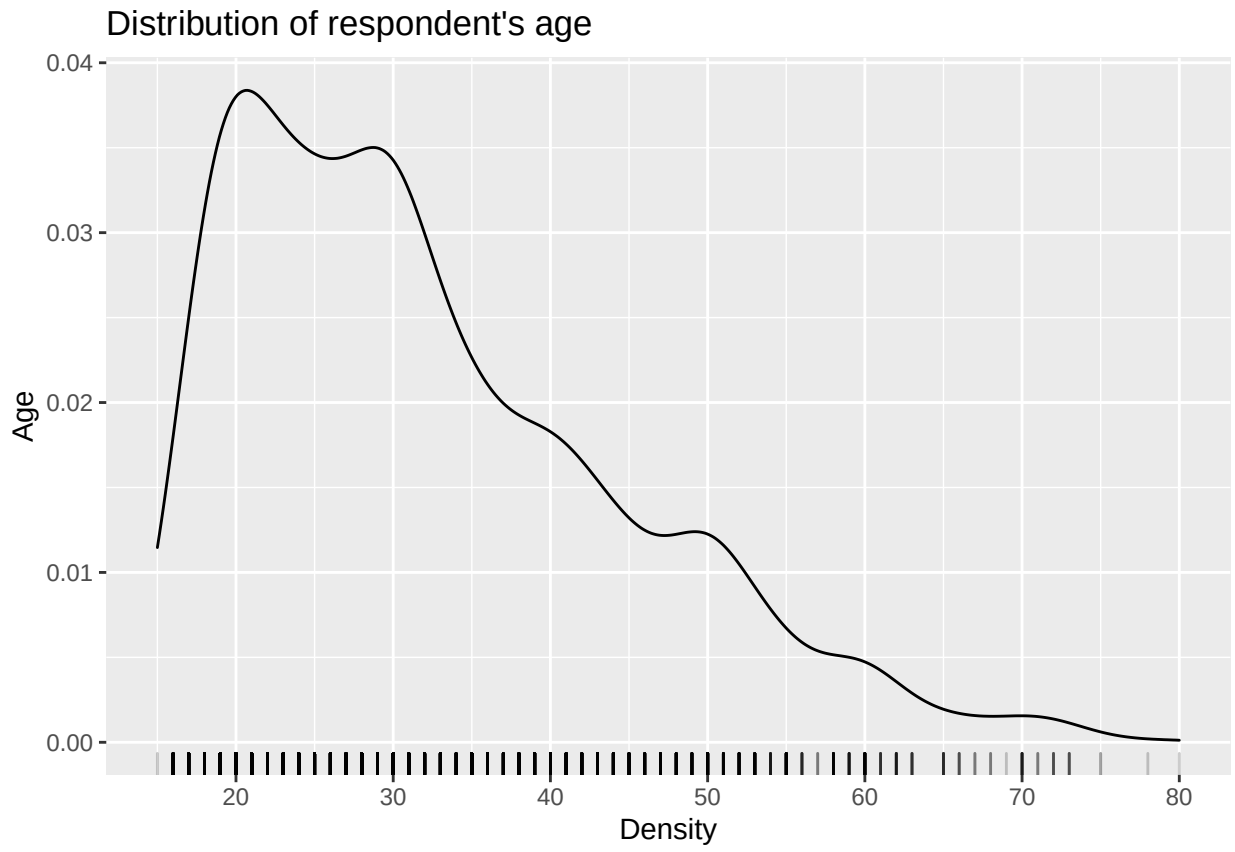

```
ggplot(afghan, aes(x = educ.years, y = ..density..)) +
  geom_histogram(binwidth = 1, center = 0) +
  geom_vline(xintercept = median(afghan$educ.years),
             color = "white", size = 2) +
  annotate("text", x = median(afghan$educ.years),
           y = 0.2, label = "median", hjust = 0) +
  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
```

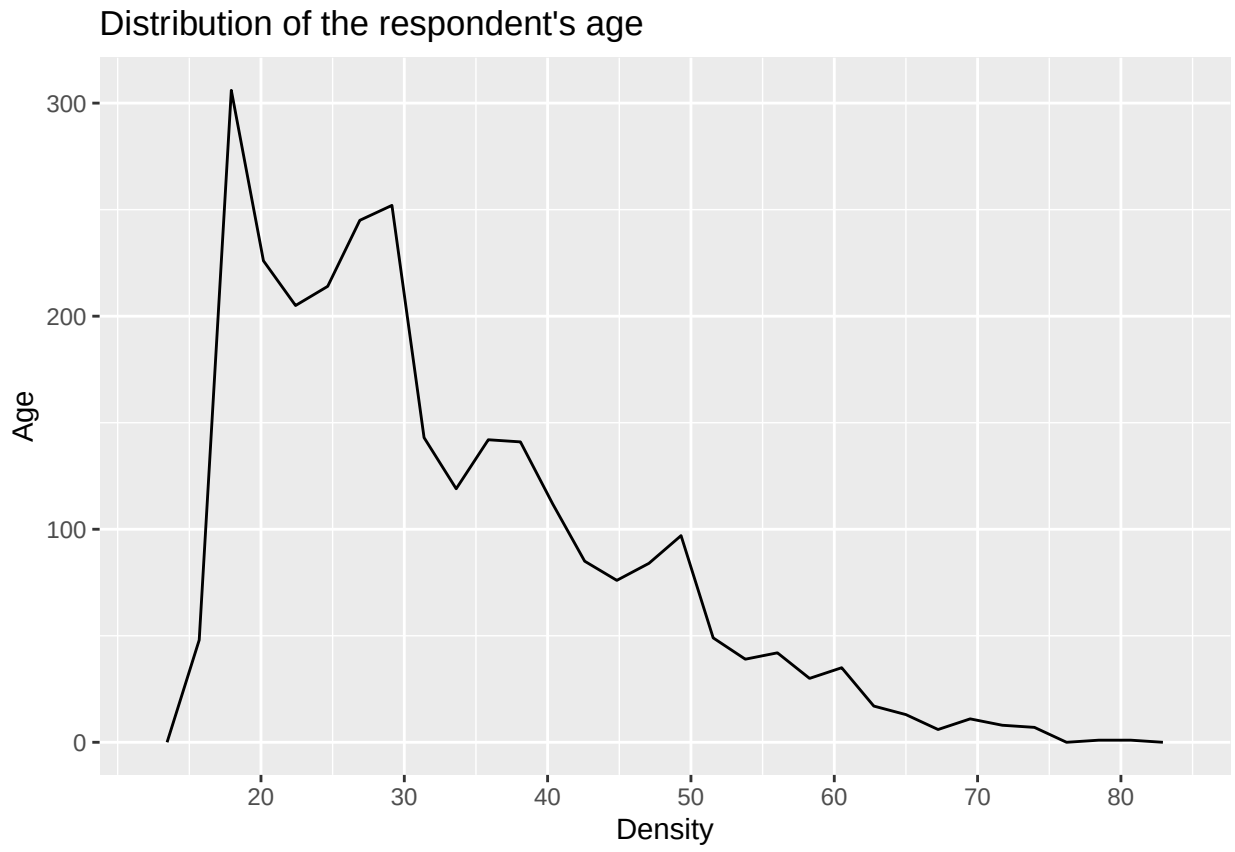



```
dens_plot + geom_rug(alpha = .2)
```



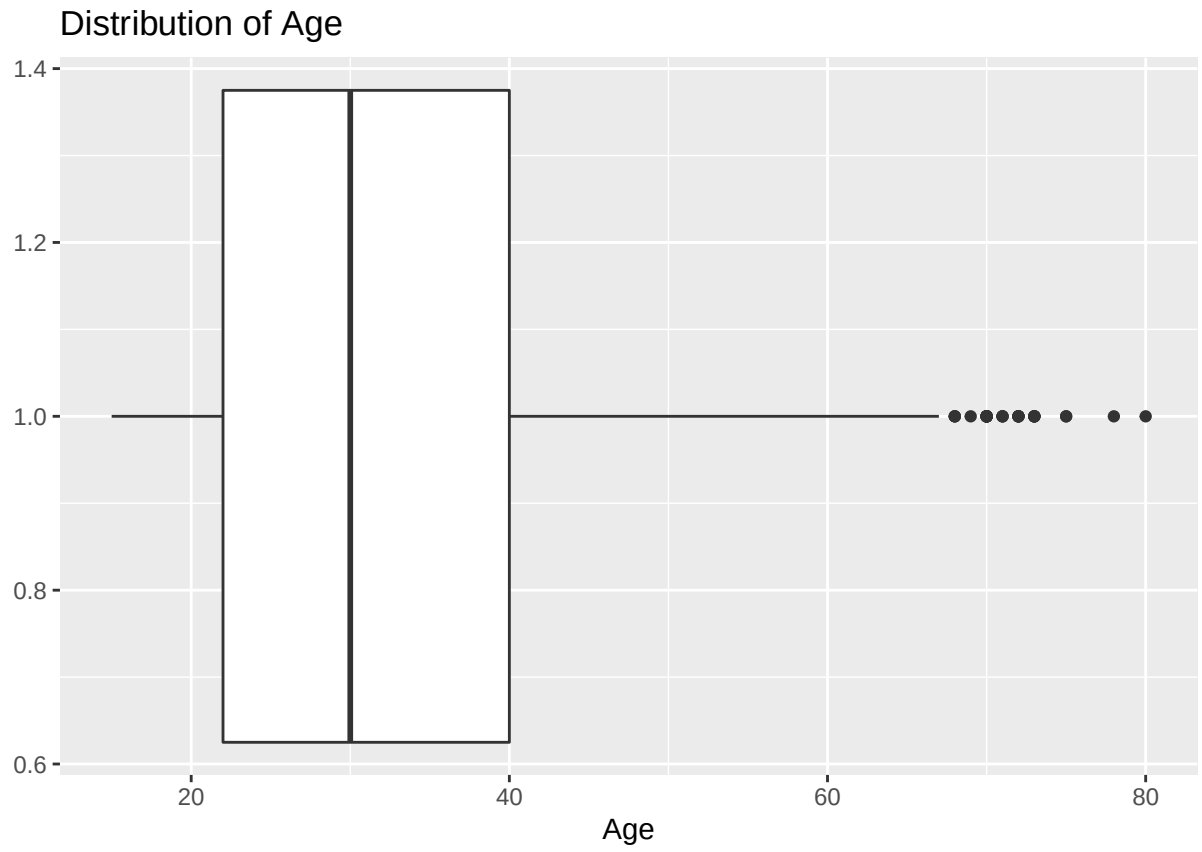

```
ggplot(afghan, aes(x = age)) +
  geom_freqpoly() +
  scale_x_continuous(breaks = seq(20, 80, by = 10)) +
  labs(title = "Distribution of the respondent's age", y = "Age", x = "Density")
```

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

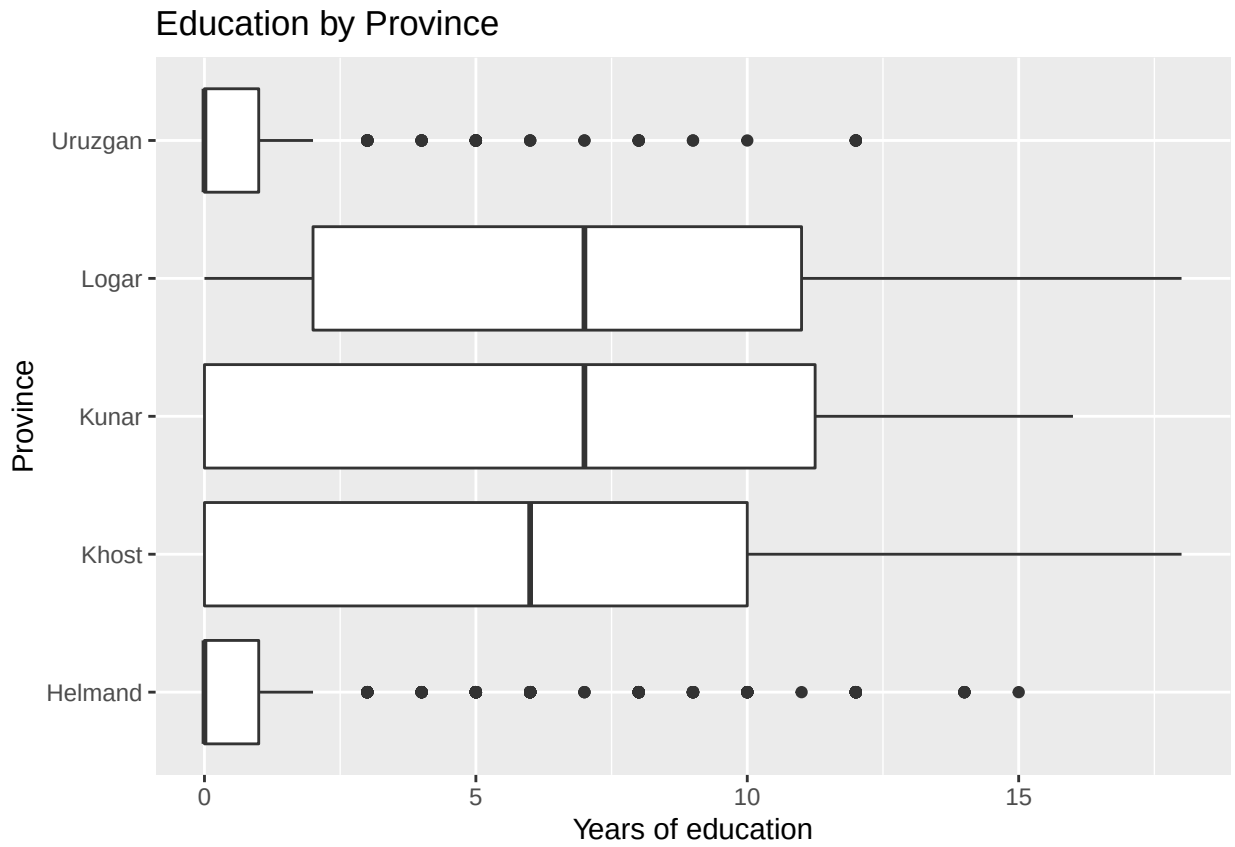


3.3.3 Boxplot

```
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)
```

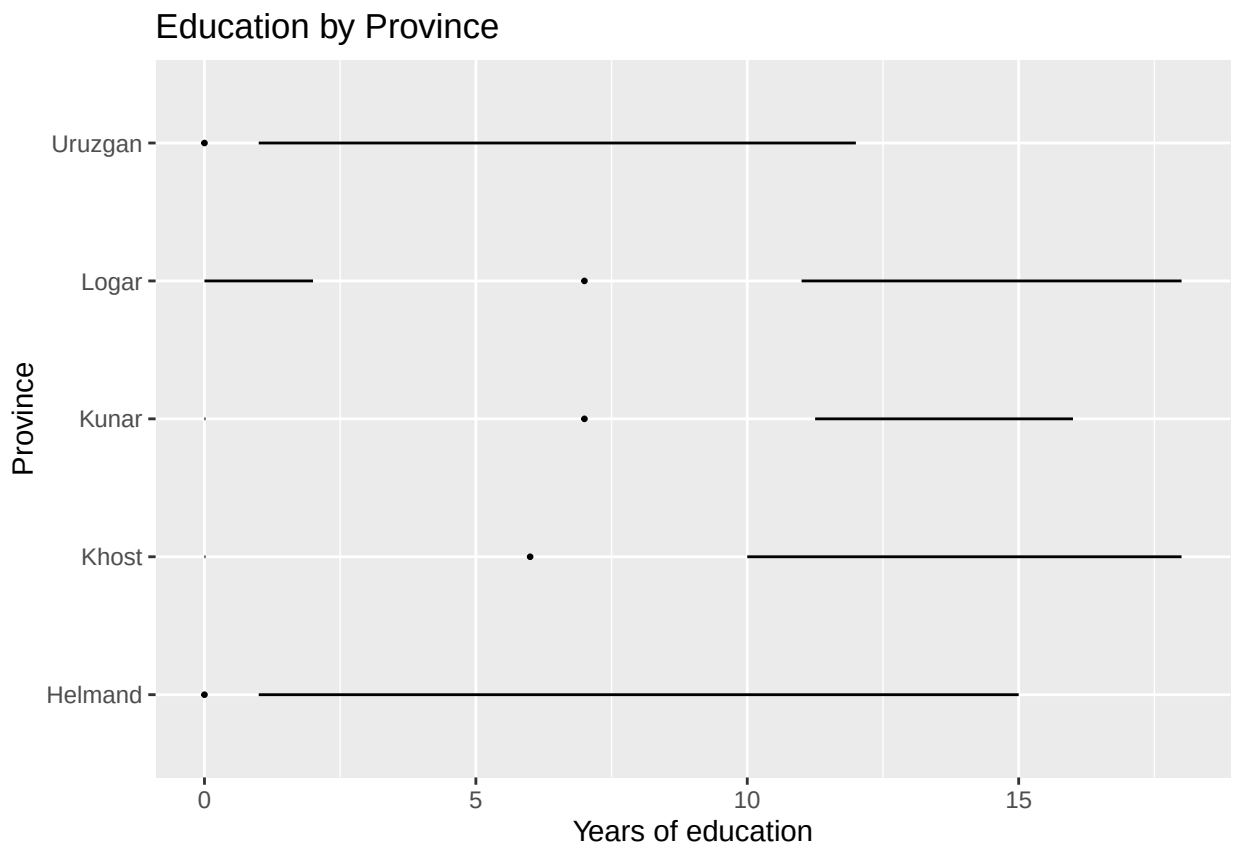
```
## # A tibble: 5 x 4
##   province educ.years violent.exp.taliban violent.exp.ISAF
##   <chr>         <dbl>               <dbl>            <dbl>
## 1 Uruzgan        1.04               0.455            0.496
## 2 Helmand        1.60               0.504            0.541
## 3 Khost          5.79   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

```
library(ggthemes)
```

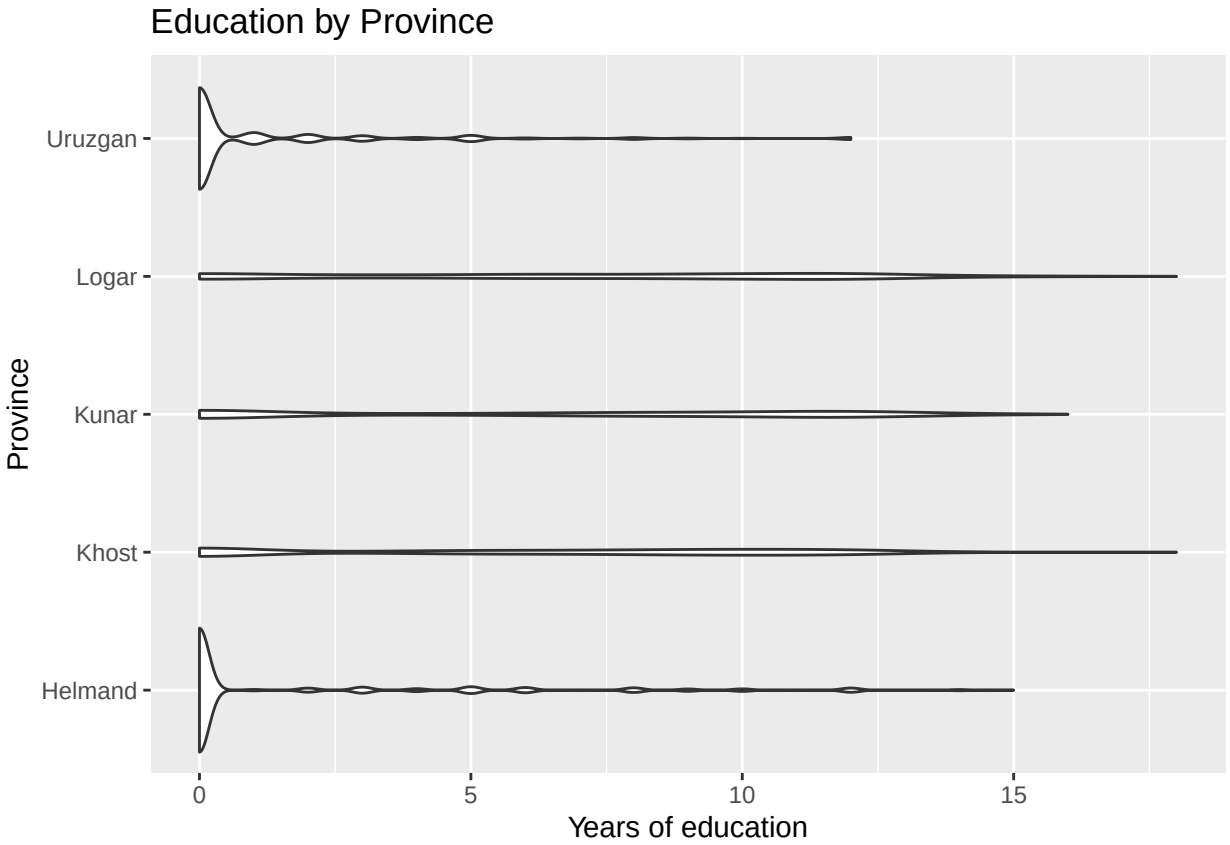
```
## Warning: package 'ggthemes' was built under R version 4.0.2
```

```
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")
```


Education by Province
Uruzgan
Logar
Kunar
Khost
Helmand
Province
0
5
10
15
Years of education