Introduction to Social Data Analytics
Class 22

Today: Plotting in R

By the end of today's lecture, you should be able to:

- Create the following plots in R: barplot, histogram, boxplot, line plots, and scatter plots
- Recall how to generate tables and which plots require tables as inputs
- Add elements to plots: titles, axis labels, ablines, text, colors, etc.
- Interpret elements of plots after creating them (e.g. quartiles in box plots)

Open class22.R if you haven't already and fill-in as we go.

On your own, load afghan.csv and explore the data

```
names(afghan)
```

```
##  [1] "province"              "district"              "village.i
##  [4] "age"                   "educ.years"            "employed"
##  [7] "income"                "violent.exp.ISAF"      "violent.e
## [10] "list.group"            "list.response"
```

```
class(afghan$violent.exp.ISAF)
```

```
## [1] "integer"
```

```
summary(afghan$violent.exp.ISAF)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##  0.0000  0.0000  0.0000  0.3749  1.0000  1.0000      25
```

What's the difference between table vs. prop.table

```
ISAF.table <- table(afghan$violent.exp.ISAF,
                    exclude = NULL)
ISAF.table
```

```
##
##    0    1 <NA>
## 1706 1023   25
```

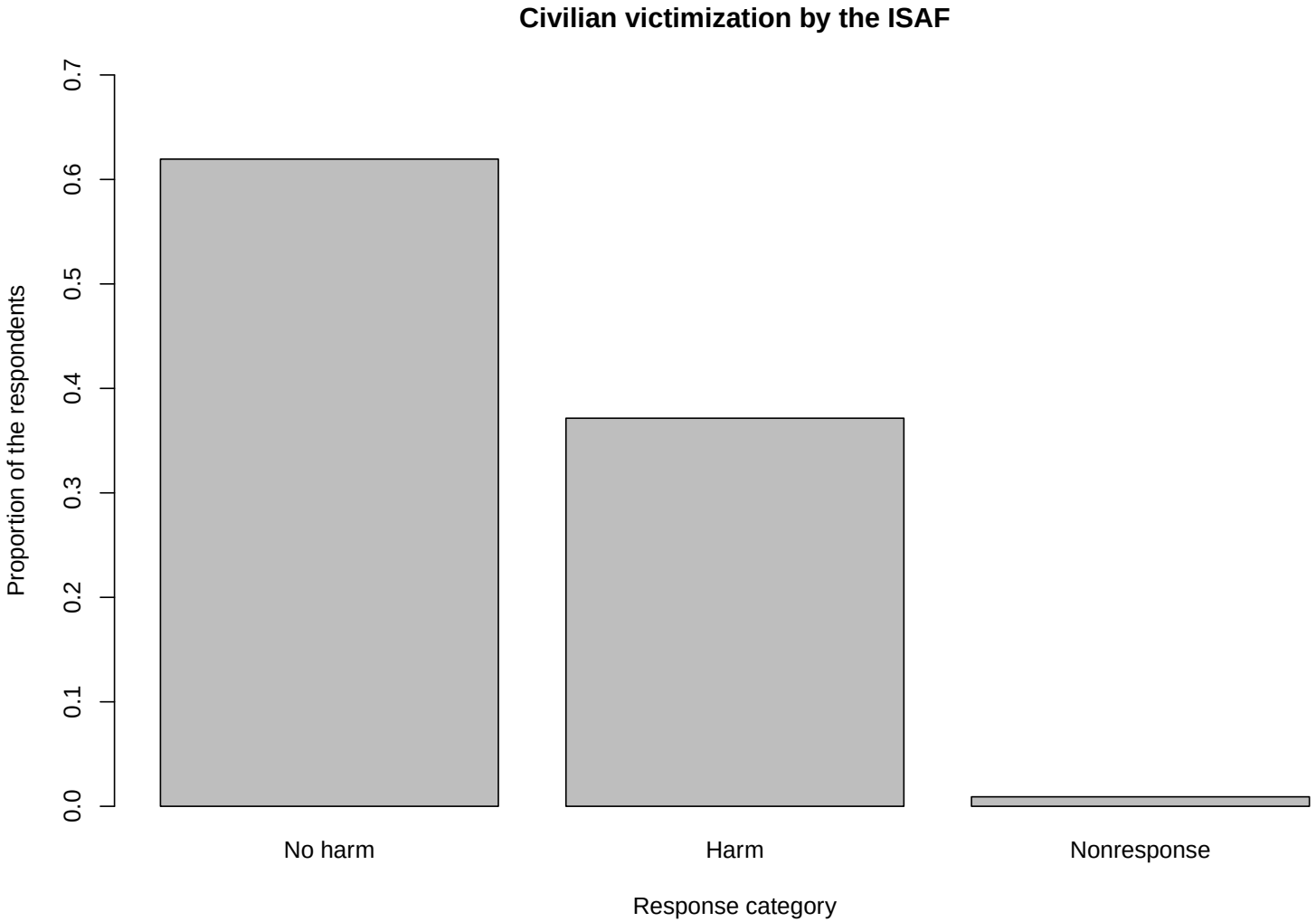
```
ISAF.ptable <- prop.table(table(afghan$violent.exp.ISAF,
                    exclude = NULL))
ISAF.ptable
```

```
##
##           0           1        <NA>
## 0.619462600 0.371459695 0.009077705
```

Create a barplot of the *percent* victimized by ISAF

```
barplot(ISAF.ptable,
        names.arg = c("No harm", "Harm", "Nonresponse"),
        main = "Civilian victimization by the ISAF",
        xlab = "Response category",
        ylab = "Proportion of the respondents",
        ylim = c(0, 0.7))
```

Your barplot should look like this.

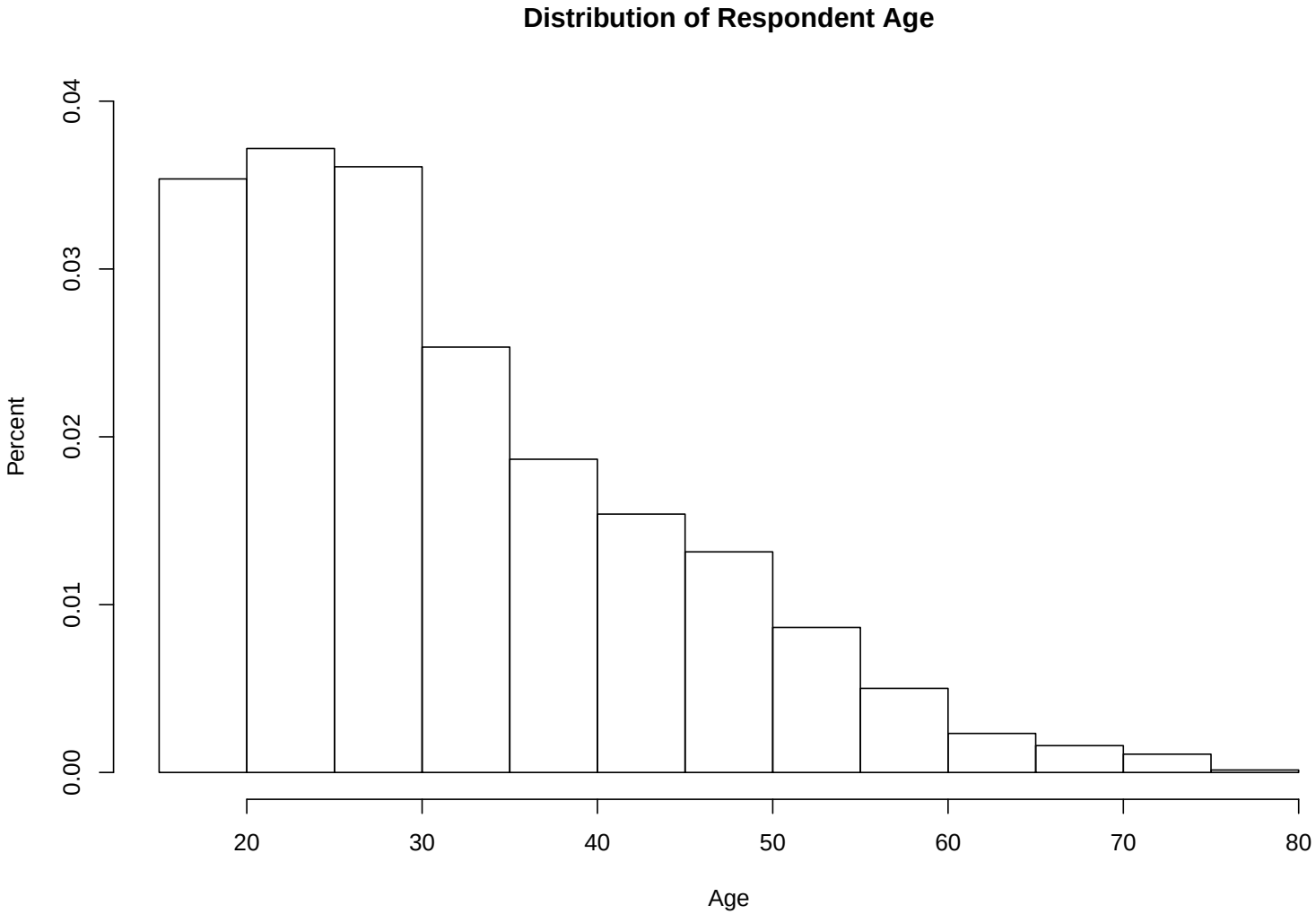


- Your turn! Create a barplot for `afghan$violent.exp.taliban`

Create a histogram of respondent ages

```
hist(afghan$age, freq = FALSE,
     ylim = c(0, 0.04),
     xlab = "Age",
     ylab = "Percent",
     main = "Distribution of Respondent Age")
```

Your histogram should look like this



- See if you can do the same for `afghan$age` (Notice `breaks()`)

Suppose you want to add a vertical line though the median...

```
# Add a vertical line at the median education level
# using abline()
abline(v = median(afghan$educ.years))

# Add a text label "median" at (x, y) = (3, 0.5)
text(x = 3, y = 0.5, "median")

# Add a vertical line at the mean using lines()
lines(x = rep(mean(afghan$educ.years), 2),
      y = c(-100, 1500))
```

- Try to add a text label "mean" in an appropriate place.

Can we create a histogram for afghan$income? Why or why not?

```
summary(afghan$income)
```

```
##   10,001-20,000    2,001-10,000   20,001-30,000 less than 2
##             616            1420              93
##     over 30,000            NA's
##              14             154
```

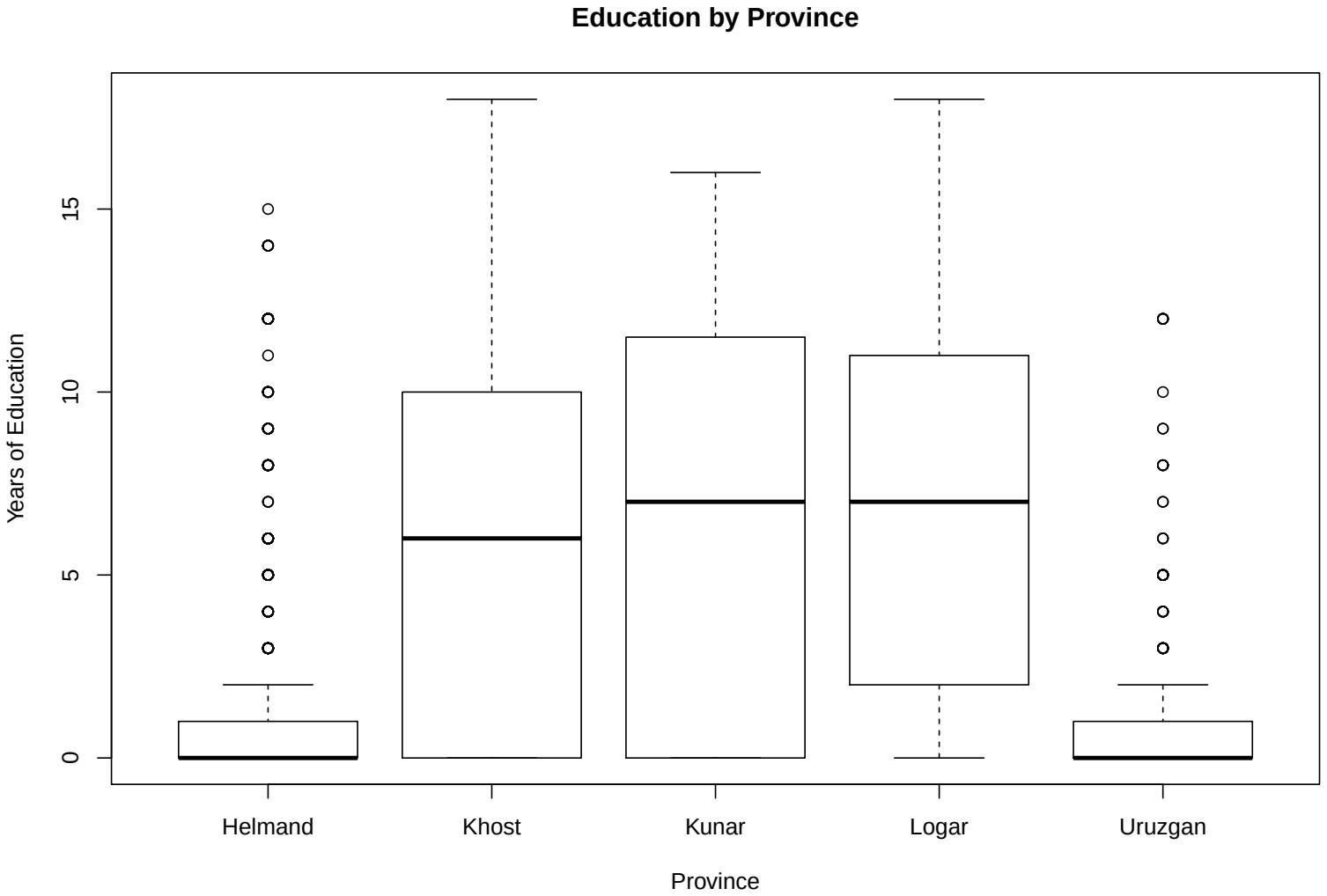
```
class(afghan$income)
```

```
## [1] "factor"
```

Make a box plot of years of education separated by province

```
boxplot(educ.years ~ province,
        data = afghan,
        main = "Education by Province",
        xlab = "Province",
        ylab = "Years of Education")
```

Which provinces are the most educated?



- Make a boxplot of age separated by each district.

Now load congress.csv and explore the data on your own

```
congress <- read.csv("congress.csv")
head(congress, 3)
```

```
##   congress district   state    party        name dwnom1 dwnc
## 1       80        0     USA Democrat     TRUMAN -0.276  0.0
## 2       80        1 ALABAMA Democrat BOYKIN  F. -0.026  0.7
## 3       80        2 ALABAMA Democrat  GRANT  G. -0.042  0.9
```

Subsetting has been done for you.

- What is `rep80`?
- What is `dem112`?

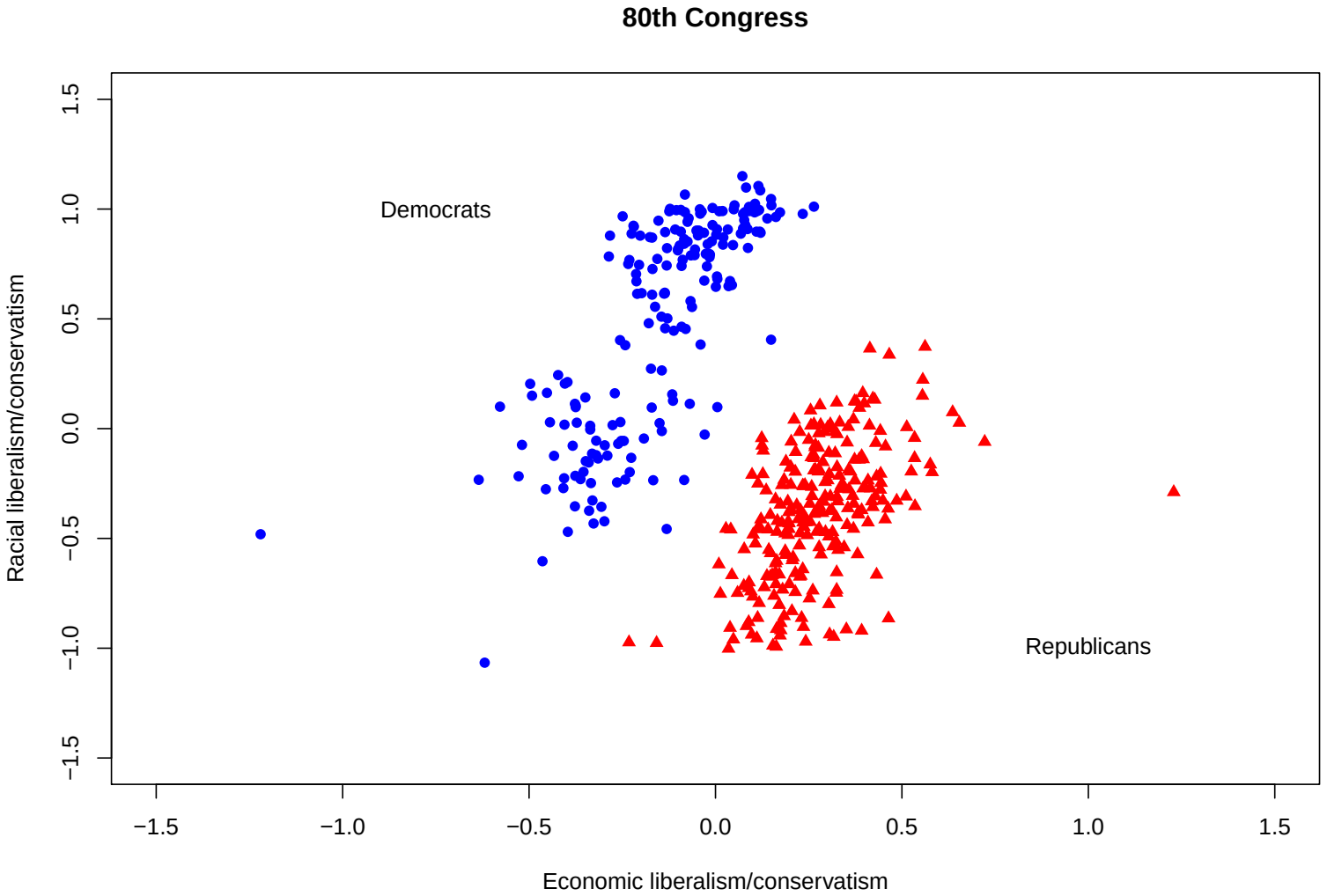
```
summary(rep80)
```

```
##     congress      district          state                party
##  Min.   :80   Min.   : 1.00   NEW YOR: 28   Democrat   :  0
##  1st Qu.:80   1st Qu.: 3.00   PENNSYL: 28   Other      :  0
##  Median :80   Median : 7.00   ILLINOI: 20   Republican:250
##  Mean   :80   Mean   :12.00   OHIO   : 20
##  3rd Qu.:80   3rd Qu.:14.75   MICHIGA: 15
##  Max.   :80   Max.   :99.00   CALIFOR: 14
##                               (Other):125
##          name          dwnom1            dwnom2
##  COLE  W. :  2   Min.   :-0.2320   Min.   :-1.0020
##  PHILLIPS :  2   1st Qu.: 0.1810   1st Qu.:-0.5833
##  ALLEN, J.:  1   Median : 0.3660   Median :-0.3505
```

Create a scatter plot demonstrating ideological division

```
plot(1, type = "n", # Type "n" specifies no plotting
     xlim = lim,
     ylim = lim,
     xlab = xlab,
     ylab = ylab,
     main = "80th Congress")
points(dem80$dwnom1, dem80$dwnom2,
       pch = 16, col = "blue") # democrats
points(rep80$dwnom1, rep80$dwnom2,
       pch = 17, col = "red") # republicans
text(-0.75, 1, "Democrats")
text(1, -1, "Republicans")
```

Your scatter plot should look like this



- Create the same plot for the 112th congress

Now we create line plots showing ideology change over time

- First, let's generate vectors of median ideology vs time for each party

```
# Calculate party median for each congress
dem.median <- tapply(dem$dwnom1, dem$congress, median)
rep.median <- tapply(rep$dwnom1, rep$congress, median)
```

Now we can plot lines

```
plot(names(dem.median), dem.median,
     col = "blue",
     type = "l",
     xlim = c(80, 115),
     ylim = c(-1, 1),
     xlab = "Congress",
     ylab = "Median ideological leaning of party")
lines(names(rep.median), rep.median, col = "red")
text(110, -0.6, "Democratic\n Party")
text(110, 0.85, "Republican\n Party")
```

Does your plot look like this?