Hi Caity,

Nice going! You’ve found all Klint’s easter eggs and you showed some neat recoding skills with the use of across() etc. Your boxplot for Q1 was really great, the geom\_joy was a nice surprise for me, and I am impressed you thought of working with background colors to signal the difference between plots for democrat and republican voters. Here are some areas with room for improvement

* Mixing base and tidyverse. Base R slips in here and there. That’s ok, it takes time to get used to another way of coding and it is so tempting to fall back on what you already know. But… read.csv does do weird things and I think you experienced something like that with the weird entry for ppage. In tidyverse’s read\_csv the problem was much more manageable. In addition, sticking to one way of coding does improve the readability of the code. So keep trying! I’ve made suggestions troughout your code that you hopefully find helpful
* I was not a fan of the stacked bar charts for count. They were really hard to interpret. In the end, you ended up with your awesome boxplot, so I don’t want to ding you too much on it, but take I still wanted to mention this for future work. In general, avoid stacked bar charts. If you must use them (and sometimes they are helpful!) make sure to use % instead of counts. This makes comparison much easier.
* I struggled with your visualization for the forced choice Q5 item that you used for the last part of the assignment. This would be a case where a stacked bar chart with % would have worked. The boxplots were hard to decipher.

So all in all, this assignment was mostly designed to test your data wrangling skills (data viz was cherry on top) and I think you got those functions figured out. Just keep working of consistent coding!

Grade: A-

EDUC 423A/SOC 302A: Assignment 2

Caity McGinley

2/6/2021 - OAE Extension

# Honor Code Statement  
  
We strongly encourage students to form study groups and students may discuss and work on assignments in groups. We expect that each student understands their own submission. As such, students must write their submissions independently and clearly disclose the names of all other students who were part of their study group. Additionally, lifting code or solutions directly from the internet (e.g., Google, GitHub, Stack Overflow) is a violation of the [Stanford Honor Code](https://communitystandards.stanford.edu/policies-and-guidance/honor-code). We take academic honesty and Honor Code violations extremely seriously and expect the same of students. If you have questions about what may or may not constitute an Honor Code violation, please reach out the teaching team.  
  
\*\*Email:\*\*   
caitym26@stanford.edu  
  
\*\*SUID:\*\*   
06211809  
\*\*Study Group:\*\*   
  
I acknowledge and agree to abide by the Honor Code.  
  
\*\*Signed:\*\*   
caity McGinley  
  
# Setup and Data Cleaning  
  
  
```r  
# Include all code required to load packages, import, and clean data here.  
library (knitr)

## Warning: package 'knitr' was built under R version 4.0.3

library (dplyr)

## Warning: package 'dplyr' was built under R version 4.0.3

##   
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':  
##   
## filter, lag

## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

library(gridExtra)

## Warning: package 'gridExtra' was built under R version 4.0.3

##   
## Attaching package: 'gridExtra'

## The following object is masked from 'package:dplyr':  
##   
## combine

library(ggplot2)  
library(tidyverse)

library(tidytext)

## Warning: package 'tidytext' was built under R version 4.0.3

library(textdata)

## Warning: package 'textdata' was built under R version 4.0.3

library(janitor)

## Warning: package 'janitor' was built under R version 4.0.3

##   
## Attaching package: 'janitor'

## The following objects are masked from 'package:stats':  
##   
## chisq.test, fisher.test

library(stringr)  
library(stringi)  
library(cowplot)

## Warning: package 'cowplot' was built under R version 4.0.3

library(ggjoy)

## Warning: package 'ggjoy' was built under R version 4.0.3

## Loading required package: ggridges

## Warning: package 'ggridges' was built under R version 4.0.3

## The ggjoy package has been deprecated. Please switch over to the  
## ggridges package, which provides the same functionality. Porting  
## guidelines can be found here:  
## <https://github.com/clauswilke/ggjoy/blob/master/README.md>

I see you are loading a lot of additional packages. Some of them are on their way out (‘deprecated’). Not a biggie for this assignment, but try to stay up to date with package developments.

setwd("C:/Users/cmcgi/Downloads/Soc 302A\_Lab1")  
  
#loading data  
votedata <- read.csv("nonvoter\_data\_v2.csv", header=T, sep=",")

I notice you use base R for reading your data. That is fine, but be aware that it sometimes behaves differently than read\_csv from tidyverse.   
#I need age, voting status, experience, engagement. What variables deal with that?   
  
#Need to pick variables I want to keep and explore for answering the question. These are the ones I have selected from the codebook.  
selectdata <- votedata %>%   
select (ppage,gender, race, educ, voter\_category, Q5,Q14, Q15, Q16, Q17\_1, Q17\_2,Q17\_3,Q17\_4, Q20, Q21, Q22, Q23, Q24, Q26, Q27\_1,Q27\_2, Q27\_3, Q27\_4, Q27\_5, Q27\_6, Q28\_1,Q28\_2,Q28\_3, Q28\_4, Q28\_5, Q28\_6, Q28\_7,Q28\_8,Q29\_1, Q29\_2, Q29\_3, Q29\_4, Q29\_5, Q29\_6, Q29\_7, Q29\_8, Q29\_9, Q29\_10, Q30)  
  
descriptions <- selectdata %>%   
select (ppage, gender, race, educ, voter\_category)  
  
  
#Confidence variables need to be numerically recoded.   
  
table(selectdata$Q17\_1) #weird space in somewhat confident Nice catch!

##   
## Not at all confident Not very confident Somewhat confident   
## 216 473 2225   
## Very confident   
## 2868

#Recode numerically.

selectdata <- selectdata %>%   
 mutate(across(starts\_with("Q17"),   
 ~ (case\_when(. == "Not at all confident" ~ 1,   
 . == "Not very confident" ~ 2,   
 . == "Somewhat confident" ~ 3,   
 . == "Very confident" ~ 4)))) %>%   
 mutate(across(starts\_with("Q28"),   
 ~ (case\_when(. == "Not Selected" ~ 0,  
 . == "Selected" ~ 1)))) %>%   
 mutate(across(starts\_with("Q29"),   
 ~ (case\_when(. == "Not Selected" ~ 0,  
 . == "Selected" ~ 1))))   
 Nice usage of across.

I would use if\_else() for conditional statements with only 2 options. It is shorter! I use case\_when() mostly when there are more than 2 categories I’m coding into.   
#Don't necessarily need to recode others since I am really only interested in these. Ok! Nice to have this explicitly stated  
  
#Renaming selected variables so that they are easily interpretable  
nameddata <- dplyr::rename (selectdata, attitude\_toward\_voting = Q5, Republican\_party\_wants = Q14, Democrat\_party\_wants = Q15, how\_hard\_to\_vote = Q16, confidence\_method\_person\_machines = Q17\_1, confidence\_method\_person\_ballots = Q17\_2, confidence\_method\_mail\_ballots = Q17\_3, confidence\_method\_electronic\_online = Q17\_4, voter\_registration\_status = Q20, voting\_plans\_2020= Q21, why\_not\_registered = Q22, voting\_for = Q23, how\_voting= Q24, general\_voting\_behavior = Q26, voting\_behavior\_multi\_yr\_2018 = Q27\_1, voting\_behavior\_multi\_yr\_2016 = Q27\_2, voting\_behavior\_multi\_yr\_2014 = Q27\_3, voting\_behavior\_multi\_yr\_2012 = Q27\_4, voting\_behavior\_multi\_yr\_2010 = Q27\_5, voting\_behavior\_multi\_yr\_2008 = Q27\_6, why\_do\_you\_vote\_1 = Q28\_1, why\_do\_you\_vote\_2 = Q28\_2, why\_do\_you\_vote\_3 = Q28\_3, why\_do\_you\_vote\_4 = Q28\_4, why\_do\_you\_vote\_5 = Q28\_5, why\_do\_you\_vote\_6 = Q28\_6, why\_do\_you\_vote\_7 = Q28\_7, why\_do\_you\_vote\_8 = Q28\_8, why\_do\_you\_not\_vote\_1 = Q29\_1, why\_do\_you\_not\_vote\_2 = Q29\_2, why\_do\_you\_not\_vote\_3 = Q29\_3, why\_do\_you\_not\_vote\_4 = Q29\_4, why\_do\_you\_not\_vote\_5 = Q29\_5, why\_do\_you\_not\_vote\_6 = Q29\_6, why\_do\_you\_not\_vote\_7 = Q29\_7, why\_do\_you\_not\_vote\_8 = Q29\_8, why\_do\_you\_not\_vote = Q29\_9, why\_do\_you\_not\_vote\_10 = Q29\_10, political\_affiliation = Q30)

It might be the word knitting, but this block of code is really hard to read. You may want to consider indenting by tabs. R studio does this automatically.   
  
#didn't name descriptive stuff so I need to merge.  
data <- merge(descriptions, nameddata)   
   
  
#Exploring data. I need age and likelihood to vote.)  
  
table(data$ppage, useNA = "always") #weird â€œ40â€\u009d symbol, need to remove.

##   
## 22 23 24 25 26 27 28 29   
## 579 262 349 547 728 675 967 1678   
## 30 31 32 33 34 35 36 37   
## 310 226 472 197 302 403 358 346   
## 38 39 40 41 42 43 44 45   
## 419 300 240 598 400 277 300 223   
## 46 47 48 49 50 51 52 53   
## 134 213 275 248 291 336 352 509   
## 54 55 56 57 58 59 60 61   
## 412 624 542 541 602 754 449 579   
## 62 63 64 65 66 67 68 69   
## 673 515 720 774 732 530 610 501   
## 70 71 72 73 74 75 76 77   
## 449 540 505 853 220 247 173 288   
## 78 79 80 81 82 83 84 85   
## 186 101 115 94 88 58 22 35   
## 86 87 88 89 90 91 92 94   
## 25 18 8 8 7 1 3 1   
## â\200œ40â\200\235 <NA>   
## 1 0

data <- data %>%  
 mutate(ppage = str\_replace\_all(ppage, "â€œ40â€\u009d", ""))  
 table(data$ppage, useNA = "always")

##   
## 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36   
## 1 579 262 349 547 728 675 967 1678 310 226 472 197 302 403 358   
## 37 38 39 40 41 42 43 44 45 46 47 48 49 50 51 52   
## 346 419 300 240 598 400 277 300 223 134 213 275 248 291 336 352   
## 53 54 55 56 57 58 59 60 61 62 63 64 65 66 67 68   
## 509 412 624 542 541 602 754 449 579 673 515 720 774 732 530 610   
## 69 70 71 72 73 74 75 76 77 78 79 80 81 82 83 84   
## 501 449 540 505 853 220 247 173 288 186 101 115 94 88 58 22   
## 85 86 87 88 89 90 91 92 94 <NA>   
## 35 25 18 8 8 7 1 3 1 0

class(data$ppage)

## [1] "character"

data$ppage <- as.numeric(data$ppage)  
  
  
#checking  
table(data$political\_affiliation, useNA = "always") #returns REPUBILCAN!!! Need to fix You’ve found another easter egg ☺

##   
## Another party, please specify Democrat   
## 542 8525   
## Independent No preference   
## 7056 2713   
## Repubilcan Republican   
## 44 8057   
## <NA>   
## 181

#fixing   
data$political\_affiliation <-   
 stri\_replace\_all\_regex(data$political\_affiliation, "Repubilcan", "Republican", vectorize\_all = FALSE)

This works but you can also use the more general if\_else

I also notice that you are mixing base R coding with tidyverse. It works, so that is fine, but a consistent coding style helps with readability for yourself and collaborators

Consider this for consistent coding



table(data$political\_affiliation, useNA = "always") # fixed

##   
## Another party, please specify Democrat   
## 542 8525   
## Independent No preference   
## 7056 2713   
## Republican <NA>   
## 8101 181

#last thing  
data$voter\_category <- factor(data$voter\_category, levels = c("rarely/never",   
 "sporadic", "always"))

How does the likelihood of voting vary by age? Does this vary across the two main political parties? For people of different ages, what seem to be the reasons they do (or do not) vote? How does confidence in the electoral process vary by age?

After you’ve cleaned the data, you will use your variable creation and visualization skills to address the question of how voters of different ages experience and engage in voting.

# How do voters of different ages experience and engage in voting?

**How does the likelihood of voting vary by age?**

# Include all code required to generate your visualization here.  
  
# Exploring and checking

min(data$ppage, na.rm = TRUE) #22

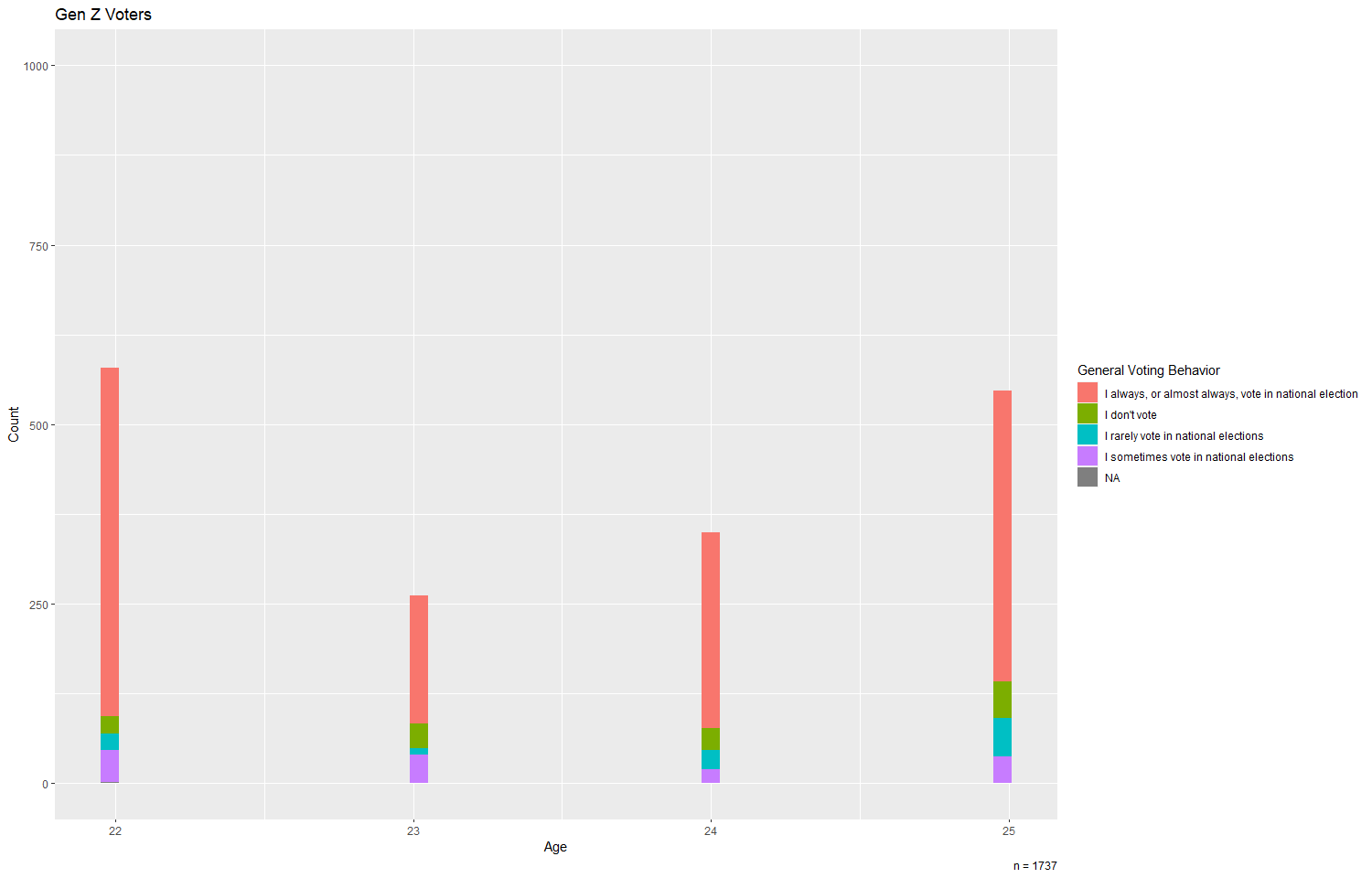
## [1] 22

max(data$ppage, na.rm = TRUE) #94

## [1] 94

#Filtering and making data subsets based off of age generation according to: <https://www.beresfordresearch.com/age-range-by-generation/> Nice!  
  
genzvoters <- filter (data, ppage %in% 22:25)  
  
millenialvoters <- filter (data, ppage %in% 26:40)  
  
genxvoters <- filter (data, ppage %in% 41:56)  
  
boomervoters <- filter (data, ppage %in% 57:75)  
  
postwarvoters <- filter(data, ppage %in% 76:94)  
  
  
#Labels for graphs   
comp.labels <- labs(x = "Age", y = "Count")  
  
   
  
genz <- ggplot(genzvoters, aes(x=ppage, fill=general\_voting\_behavior)) +  
 geom\_histogram(bins = 50, binwdith = 5) +  
 ylim(0,1000) +  
 comp.labels +  
 labs(fill = "General Voting Behavior") +  
 labs(title ="Gen Z Voters", caption = "n = 1737")

## Warning: Ignoring unknown parameters: binwdith

genz

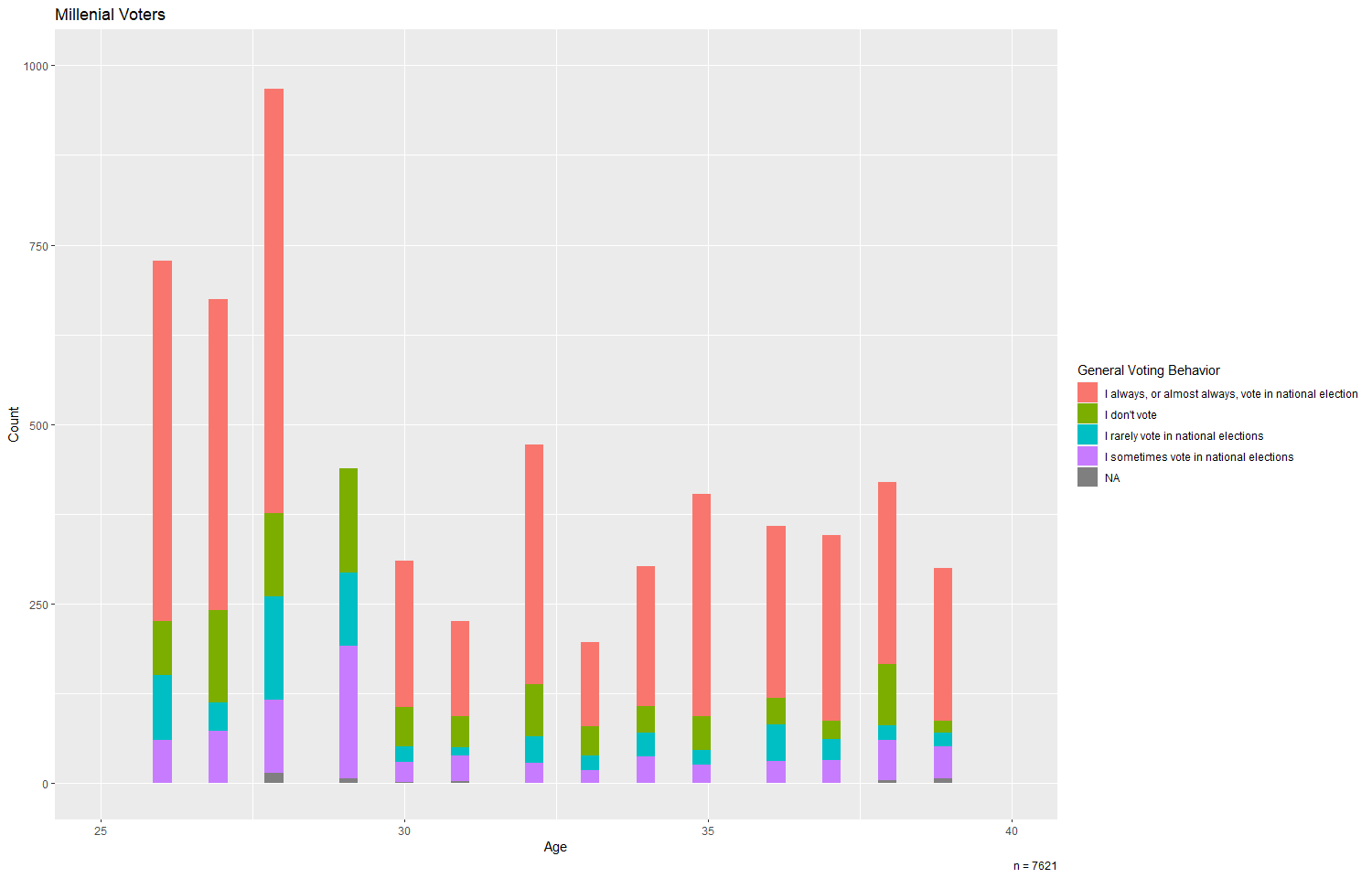
This is a good start, but I am having a hard time to compare between the ages. I have to assess both group size and the length of the bars, which is difficult because they are not plotted side by side. And, do we want to compare within generations or rather between generations. An alternative would be a bar chartwith position = dodge of percentages?

millenials <- ggplot(millenialvoters, aes(x=ppage, fill=general\_voting\_behavior)) +  
 geom\_histogram(bins = 50, binwdith = 5) +   
 xlim (25, 40) +  
 ylim(0,1000) +  
 comp.labels +   
 labs(fill = "General Voting Behavior") +  
 labs(title ="Millenial Voters", caption = "n = 7621")

## Warning: Ignoring unknown parameters: binwdith

millenials

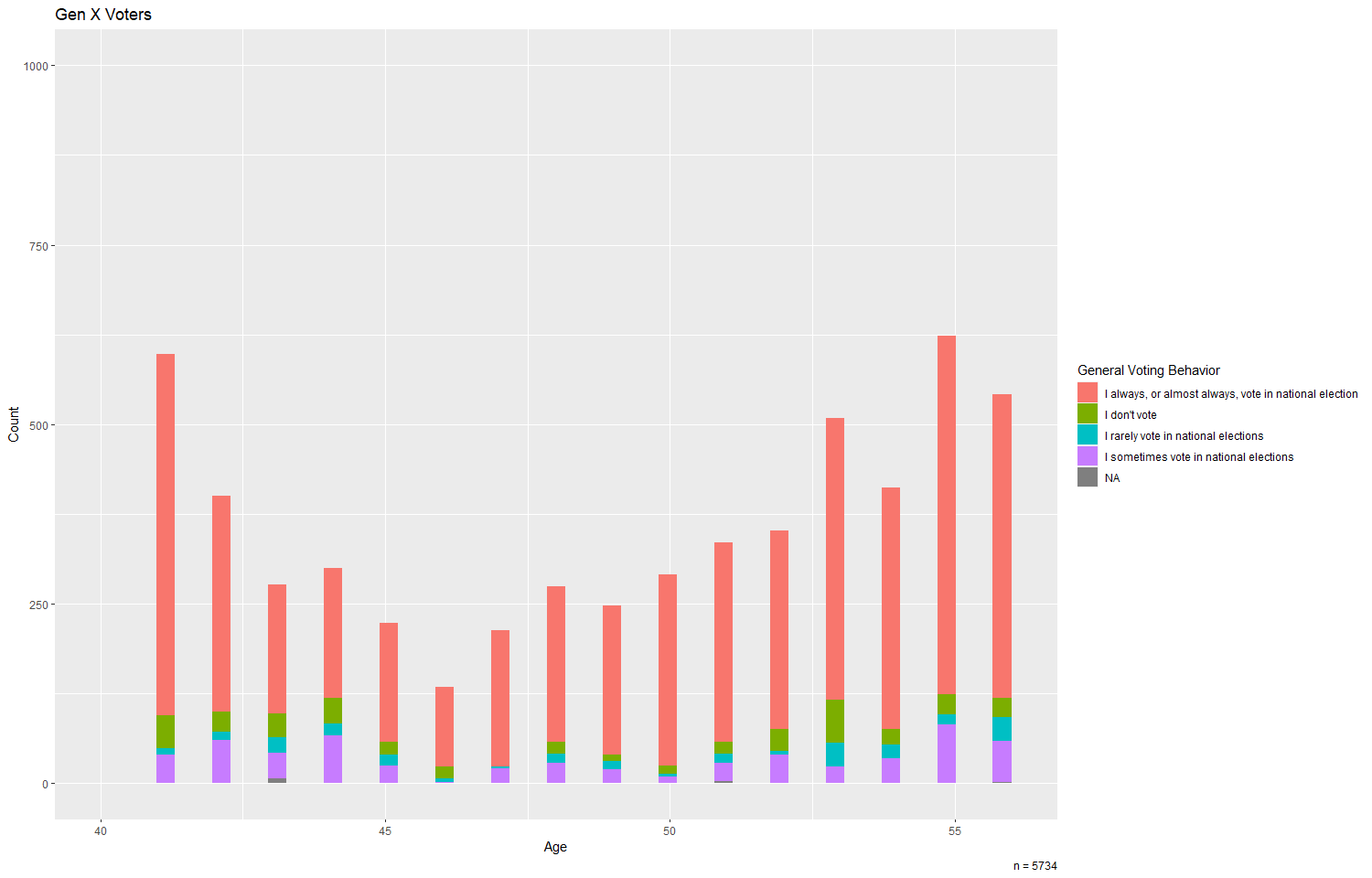
## Warning: Removed 11 rows containing missing values (geom\_bar).



genx <- ggplot(genxvoters, aes(x=ppage, fill=general\_voting\_behavior)) +  
 geom\_histogram(bins = 50, binwdith = 5) +   
 xlim(40,56) +  
 ylim(0,1000) +  
 comp.labels +   
 labs(fill = "General Voting Behavior") +  
 labs(title ="Gen X Voters", caption = "n = 5734")

## Warning: Ignoring unknown parameters: binwdith

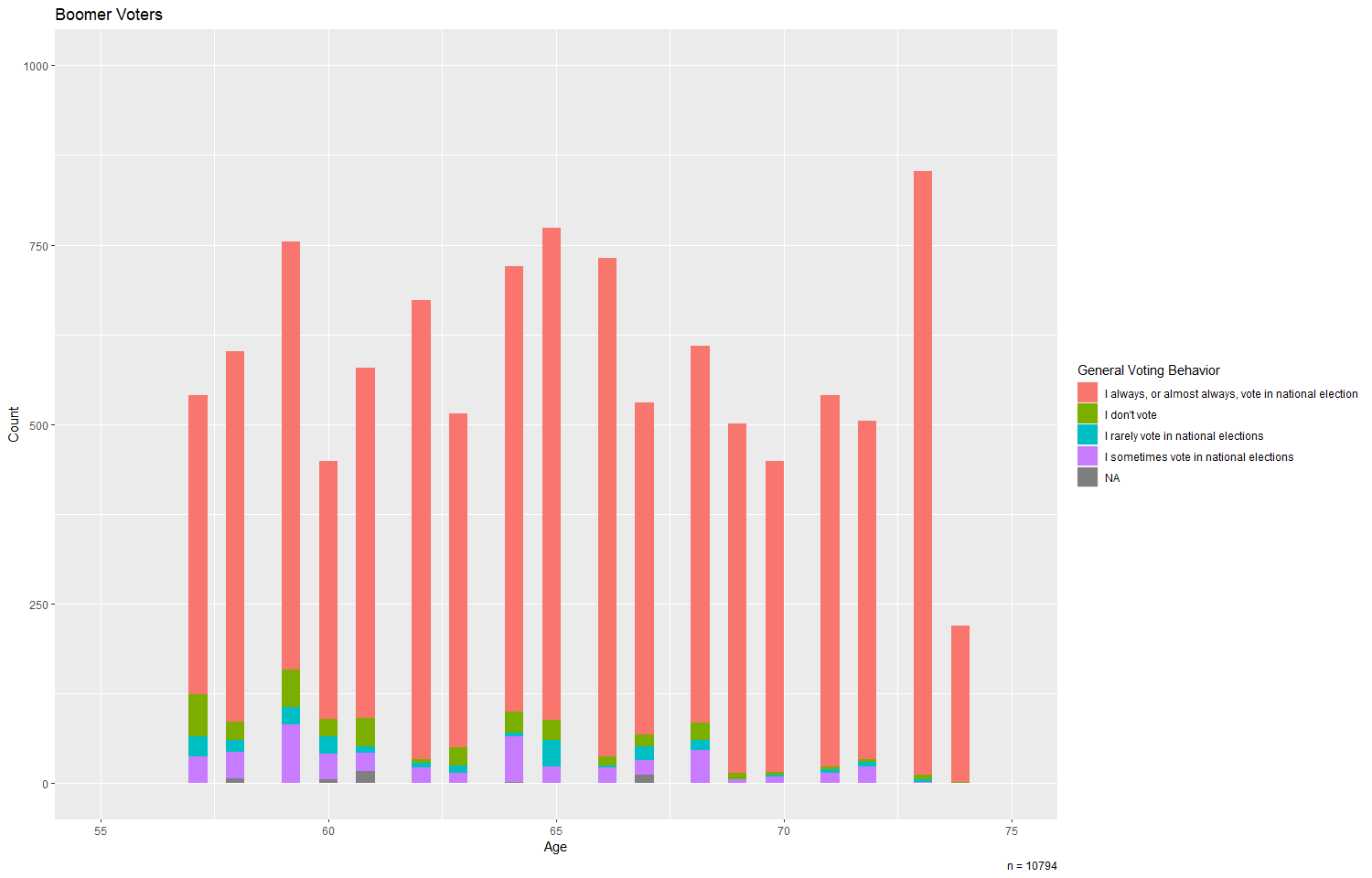
genx

## Warning: Removed 5 rows containing missing values (geom\_bar).

boomers <- ggplot(boomervoters, aes(x=ppage, fill=general\_voting\_behavior)) +  
 geom\_histogram(bins = 50, binwdith = 5) +   
 xlim(55,75) +  
 ylim(0,1000) +  
 comp.labels +  
 labs(fill = "General Voting Behavior") +  
 labs(title ="Boomer Voters", caption = "n = 10794")

## Warning: Ignoring unknown parameters: binwdith

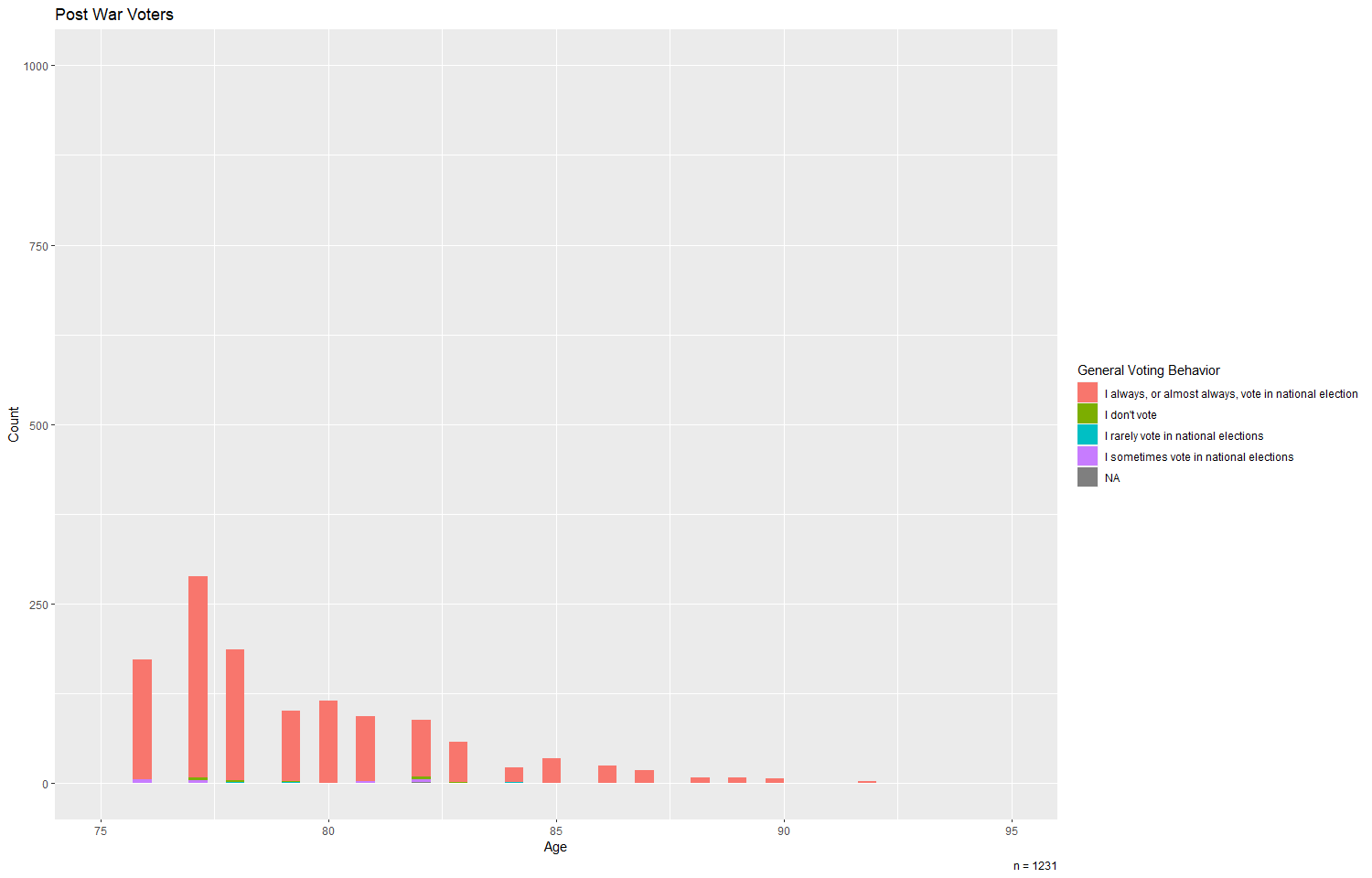
boomers

## Warning: Removed 10 rows containing missing values (geom\_bar).

postwar <- ggplot(postwarvoters, aes(x=ppage, fill=general\_voting\_behavior)) +  
 geom\_histogram(bins = 50, binwdith = 5) +  
 xlim(75,95) +  
 ylim(0,1000) +  
 comp.labels +   
 labs(fill = "General Voting Behavior") +  
 labs(title ="Post War Voters", caption = "n = 1231")

## Warning: Ignoring unknown parameters: binwdith

postwar

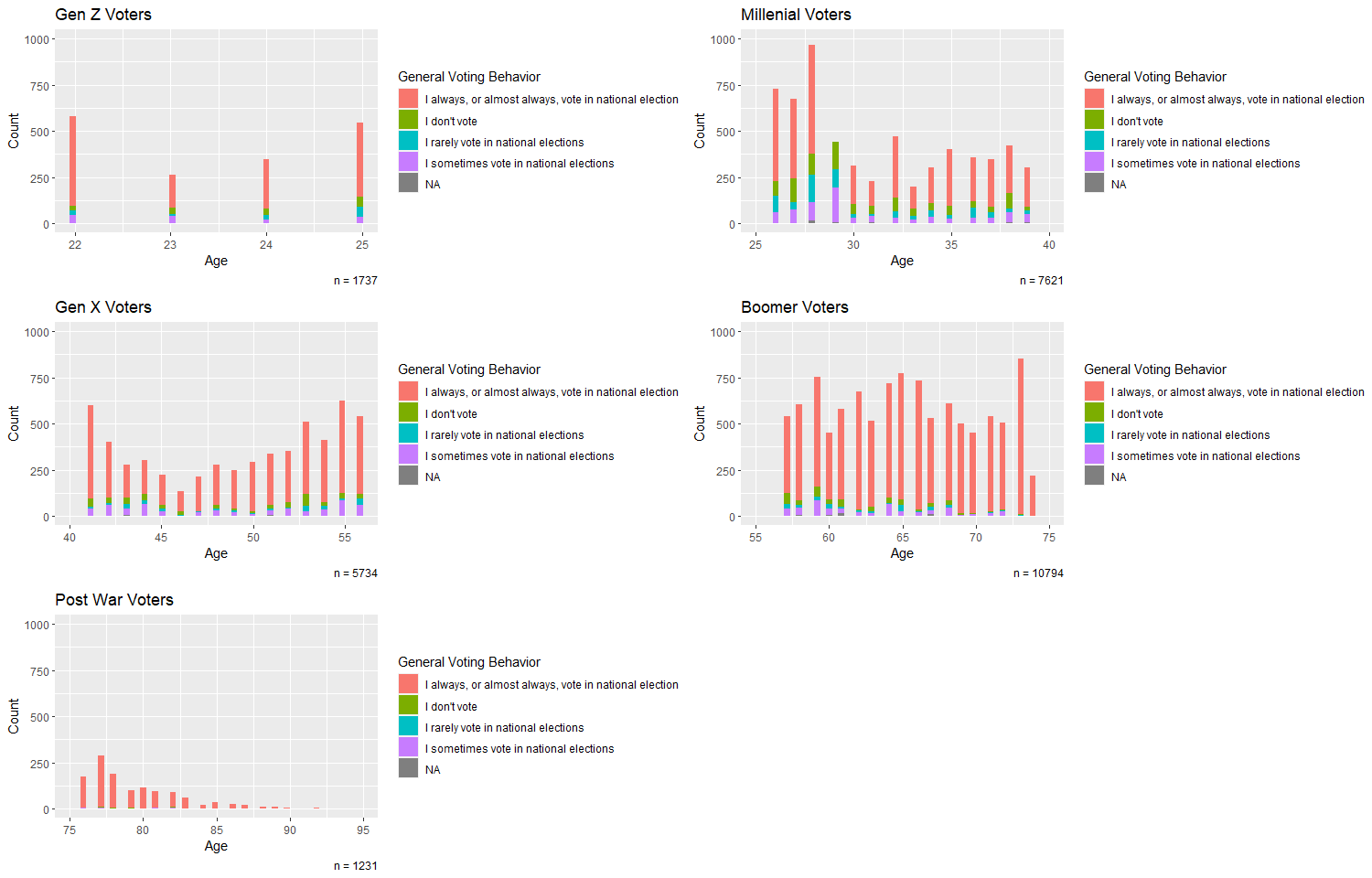
## Warning: Removed 10 rows containing missing values (geom\_bar).

plot\_grid(genz, millenials, genx, boomers, postwar, ncol = 2, nrow = 3)

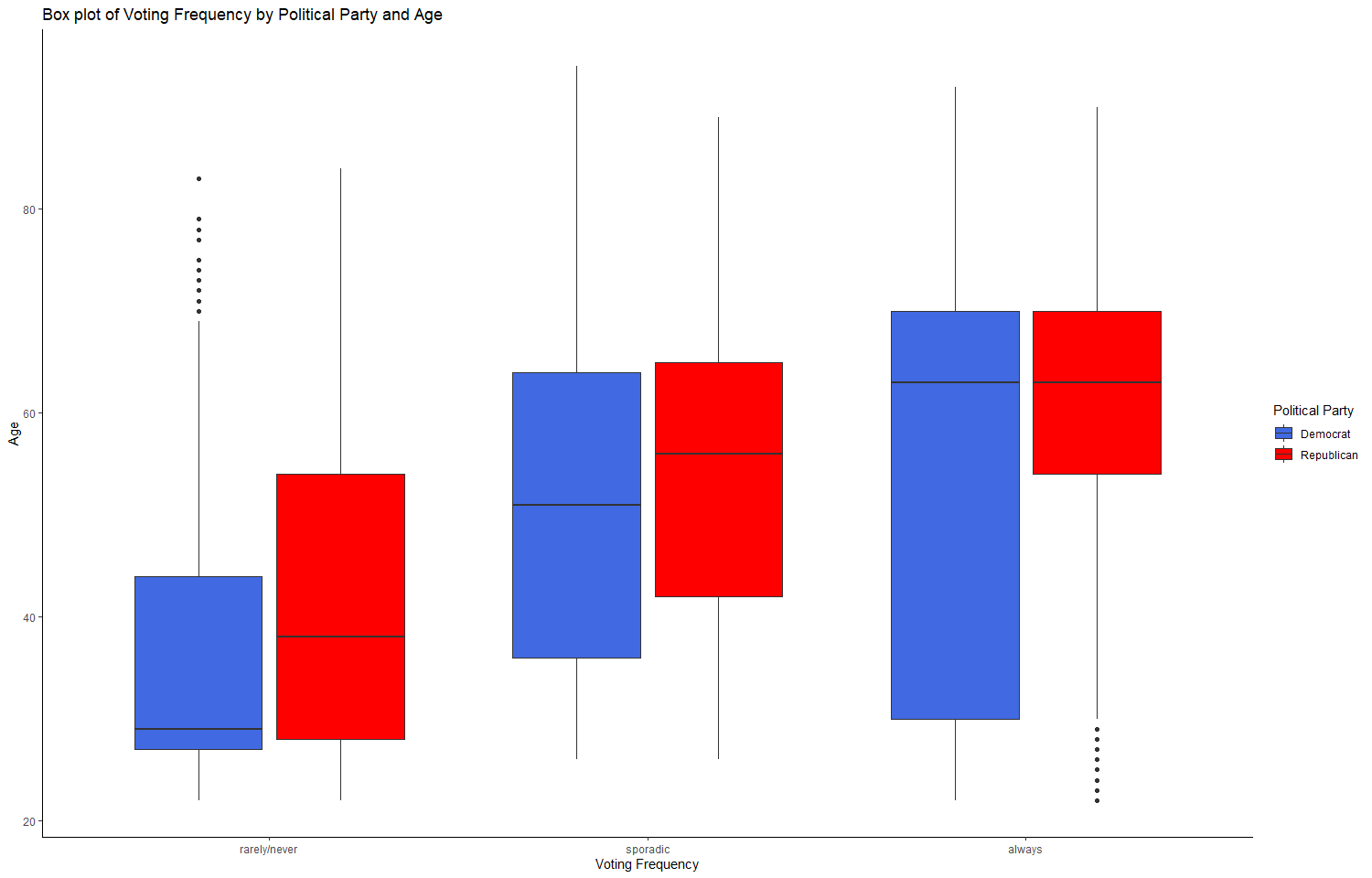
## Warning: Removed 11 rows containing missing values (geom\_bar).

## Warning: Removed 5 rows containing missing values (geom\_bar).

## Warning: Removed 10 rows containing missing values (geom\_bar).  
  
## Warning: Removed 10 rows containing missing values (geom\_bar).

Ah ok, here comes the generation comparison. It is hard to compare all these figures. Would there have been a way where you could have made a variable of generation rather than separate data objects. That way you could have made a dodged bar chart by age group. Another option would have been to create a boxplot or violin plot with age as Y and voting group as X?

bipartisan <- data %>%   
 filter(political\_affiliation %in% c("Democrat", "Republican"))  
   
 bipartisan <- bipartisan %>%  
 mutate(generation = case\_when(ppage %in% 22:25 ~ "Gen Z",  
 ppage %in% 26:40 ~ "Millenials",  
 ppage %in% 41:56 ~ "Gen X",  
 ppage %in% 57:75 ~ "Boomers",  
 ppage %in% 76:96 ~ "Post War"))   
   
  
# set factor levels in logical order  
bipartisan$generation <- factor(bipartisan$generation, levels = c("Gen Z", "Millenials", "Gen X", "Boomers", "Post War")) Yeah for ordered factor! Nice!

Oh this plot is much better than the histograms you showed before. This is an EXCELLENT visualization for the research question.   
bipartisan %>%  
 ggplot() +  
 geom\_boxplot(aes(x = voter\_category, y = ppage, fill = political\_affiliation)) +  
 labs(x = "Voting Frequency",  
 y = "Age",  
 fill = "Political Party",   
 title = "Box plot of Voting Frequency by Political Party and Age") +  
 scale\_fill\_manual(values = c("royalblue", "red")) +  
 theme\_classic()

**For people of different ages, what seem to be the reasons they do (or do not) vote?**

# Include all code required to generate your visualization here.  
data <- data %>%  
 mutate(generation = case\_when(ppage %in% 22:25 ~ "Gen Z",  
 ppage %in% 26:40 ~ "Millenials",  
 ppage %in% 41:56 ~ "Gen X",  
 ppage %in% 57:75 ~ "Boomers",  
 ppage %in% 76:96 ~ "Post War"))

Ooooh I like the use of %in% here. That makes the code really concise and clean looking.

table(data$generation)

##   
## Boomers Gen X Gen Z Millenials Post War   
## 10794 5734 1737 7621 1231

# set factor levels in logical order  
data$generation <- factor(data$generation, levels = c("Gen Z", "Millenials", "Gen X", "Boomers", "Post War"))  
  
  
  
  
# I'm going to need to pivot here!   
data\_why\_vote <- data %>%   
 pivot\_longer(names\_to = "qs",  
 values\_to = "answers",   
 cols = starts\_with("why\_do\_you\_vote"))

Ah you are ahead of the course here. We didn’t expect you to know about this and would have been fine with you focusing on a particular reason. Kudos for figuring this out.   
#checking  
head(data\_why\_vote$qs)

## [1] "why\_do\_you\_vote\_1" "why\_do\_you\_vote\_2" "why\_do\_you\_vote\_3"  
## [4] "why\_do\_you\_vote\_4" "why\_do\_you\_vote\_5" "why\_do\_you\_vote\_6"

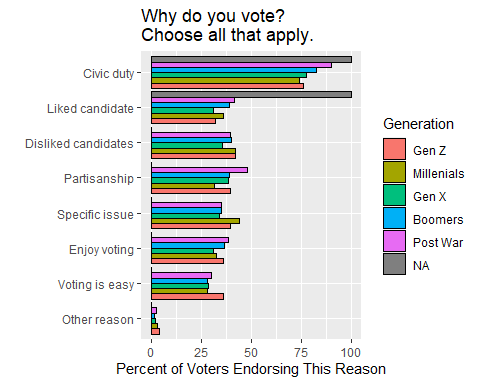
head(data\_why\_vote$answers)

## [1] 0 0 0 0 0 0

#Need to create means  
voters <- data\_why\_vote %>%   
 group\_by(generation, qs) %>%   
 summarise(mean = mean(answers, na.rm = TRUE) \* 100) %>%   
 ggplot(aes(x = reorder(qs, mean), y = mean, fill = generation)) +  
 geom\_col(position = "dodge", color = "black") +   
 scale\_x\_discrete(labels = c("Other reason", "Voting is easy", "Enjoy voting",   
 "Specific issue", "Partisanship", "Disliked candidates",   
 "Liked candidate", "Civic duty")) +  
 coord\_flip() +  
 labs(title = "Why do you vote? \nChoose all that apply.",  
 y = "Percent of Voters Endorsing This Reason",  
 x = "",  
 fill = "Generation")

## `summarise()` regrouping output by 'generation' (override with `.groups` argument)

voters



# I can't figure out how to remove NA!!!!!!!!! Why!!!! I don’t get why when I get rid of the weird symbols attached to forty, I suddenly have an NA when I didn’t before? It must be introduced via coercion, but even when I try to na.omit, it doesn’t work. It has to be something with the strings. There’s like maybe 4 NAs.

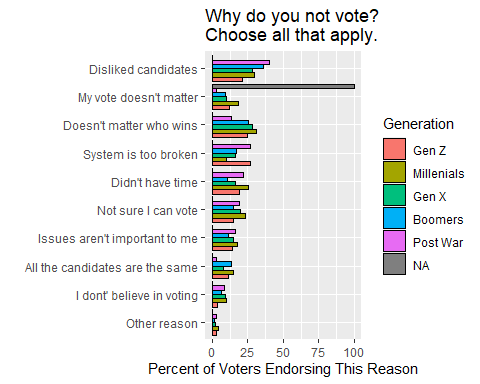
I suspect reading the data with read,csv instead of read\_csv messed things up,. With read\_csbv the forty values will be double quoted “40” which you then have to mutate to 40 (as numeric, not character. For you it showed up as "â€œ40â€\u009d" which seems not to be the literal character value, but to be an encoding issue.

Avoid read.csv. It just does weird things!

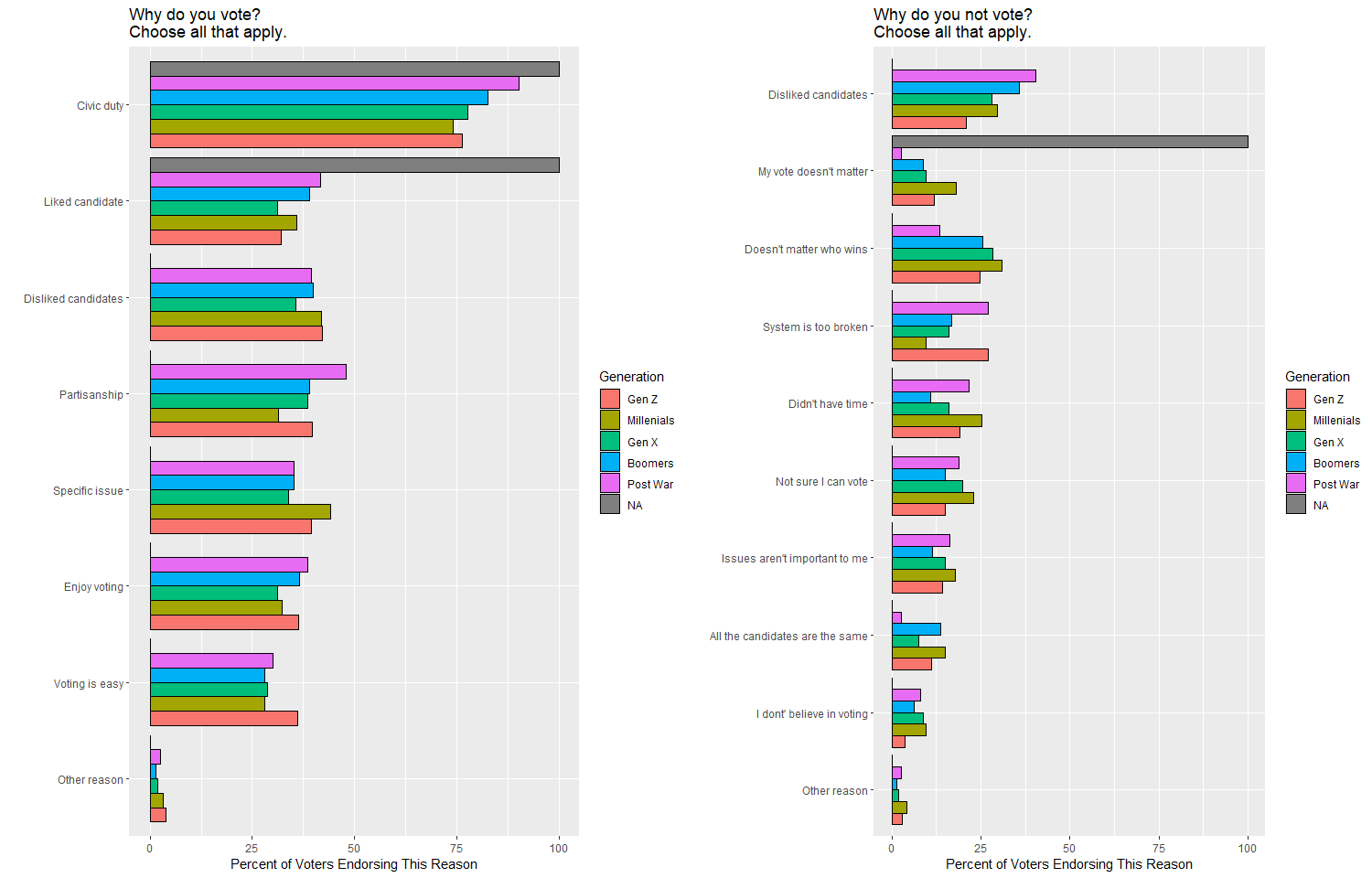
# I'm going to need to pivot here again!   
data\_why\_not\_vote <- data %>%   
 pivot\_longer(names\_to = "qss",  
 values\_to = "answerss",   
 cols = starts\_with("why\_do\_you\_not\_vote"))  
   
 nonvoters <- data\_why\_not\_vote %>%   
 group\_by(generation, qss) %>%   
 summarise(mean = mean(answerss, na.rm = TRUE) \* 100) %>%   
 ggplot(aes(x = reorder(qss, mean), y = mean, fill = generation)) +  
 geom\_col(position = "dodge", color = "black") +   
 scale\_x\_discrete(labels = c("Other reason", "I dont' believe in voting", "All the candidates are the same", "Issues aren't important to me", "Not sure I can vote", "Didn't have time",   
 "System is too broken", "Doesn't matter who wins", "My vote doesn't matter", "Disliked candidates")) +  
 coord\_flip() +  
 labs(title = "Why do you not vote? \nChoose all that apply.",  
 y = "Percent of Voters Endorsing This Reason",  
 x = "",  
 fill = "Generation")

## `summarise()` regrouping output by 'generation' (override with `.groups` argument)

nonvoters



plot\_grid(voters, nonvoters, ncol = 2)

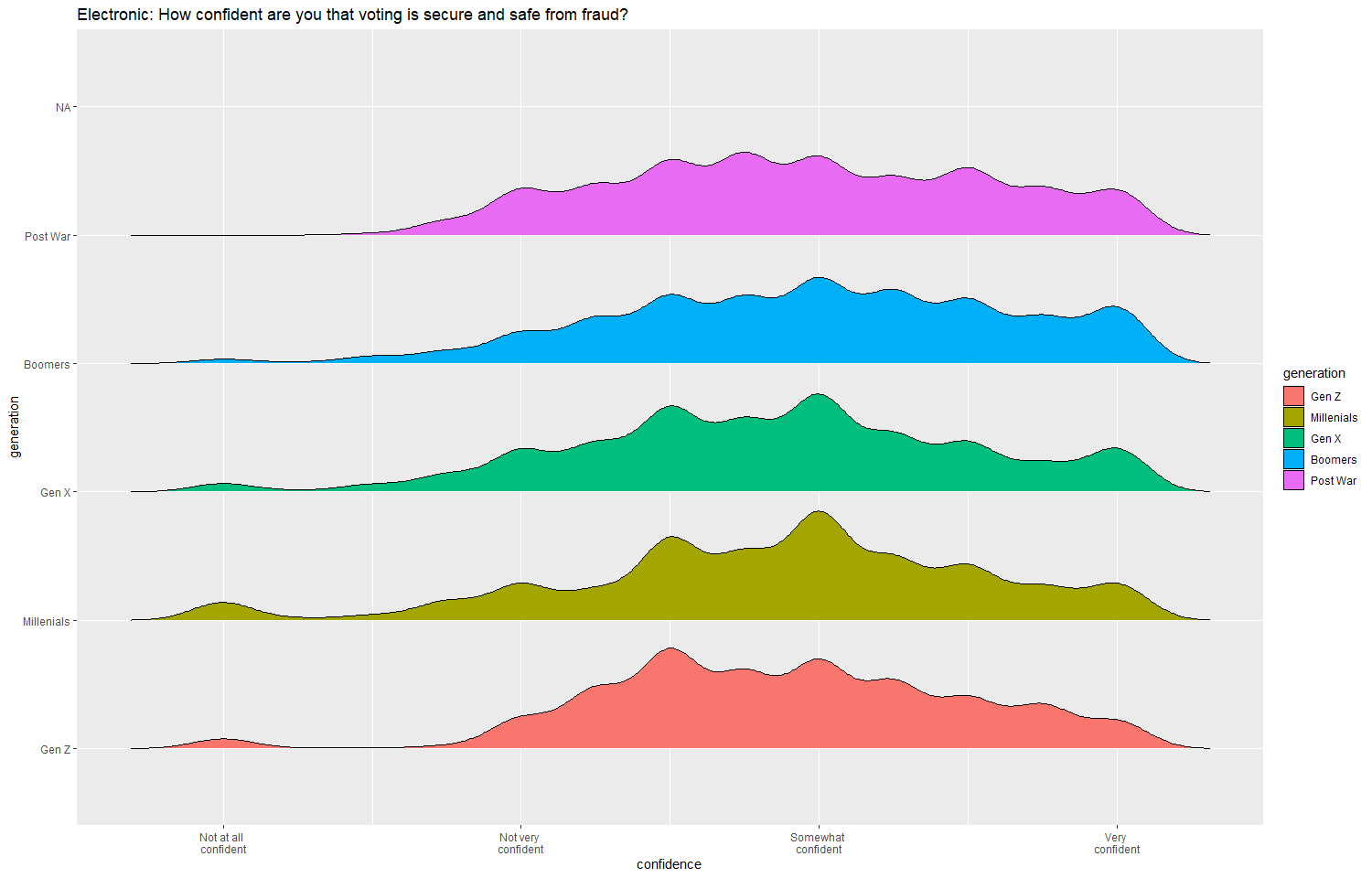
Nice work on this one!

**How does confidence in the electoral process vary by age?**

# Include all code required to generate your visualization here.  
  
  
  
data <- data %>%   
 mutate(confidence = rowMeans(select(data, starts\_with("confidence"))))  
  
data %>%   
 group\_by(generation) %>%   
 summarise(confidence = mean(confidence, na.rm = T))

## `summarise()` ungrouping output (override with `.groups` argument)

## # A tibble: 6 x 2  
## generation confidence  
## <fct> <dbl>  
## 1 Gen Z 2.88  
## 2 Millenials 2.85  
## 3 Gen X 2.84  
## 4 Boomers 2.98  
## 5 Post War 2.94  
## 6 <NA> 3.75

electronic <- data %>%   
 ggplot(aes(x=confidence, y=generation, fill = generation, height=..density..)) +  
 geom\_joy(scale=0.85) +  
labs(title = "Electronic: How confident are you that voting is secure and safe from fraud?") +   
scale\_x\_continuous(breaks = c(1, 2, 3, 4), labels = c("Not at all \nconfident",  
 "Not very \nconfident",  
 "Somewhat \nconfident",  
 "Very \nconfident"))  
electronic

oooh I like this geom\_joy().

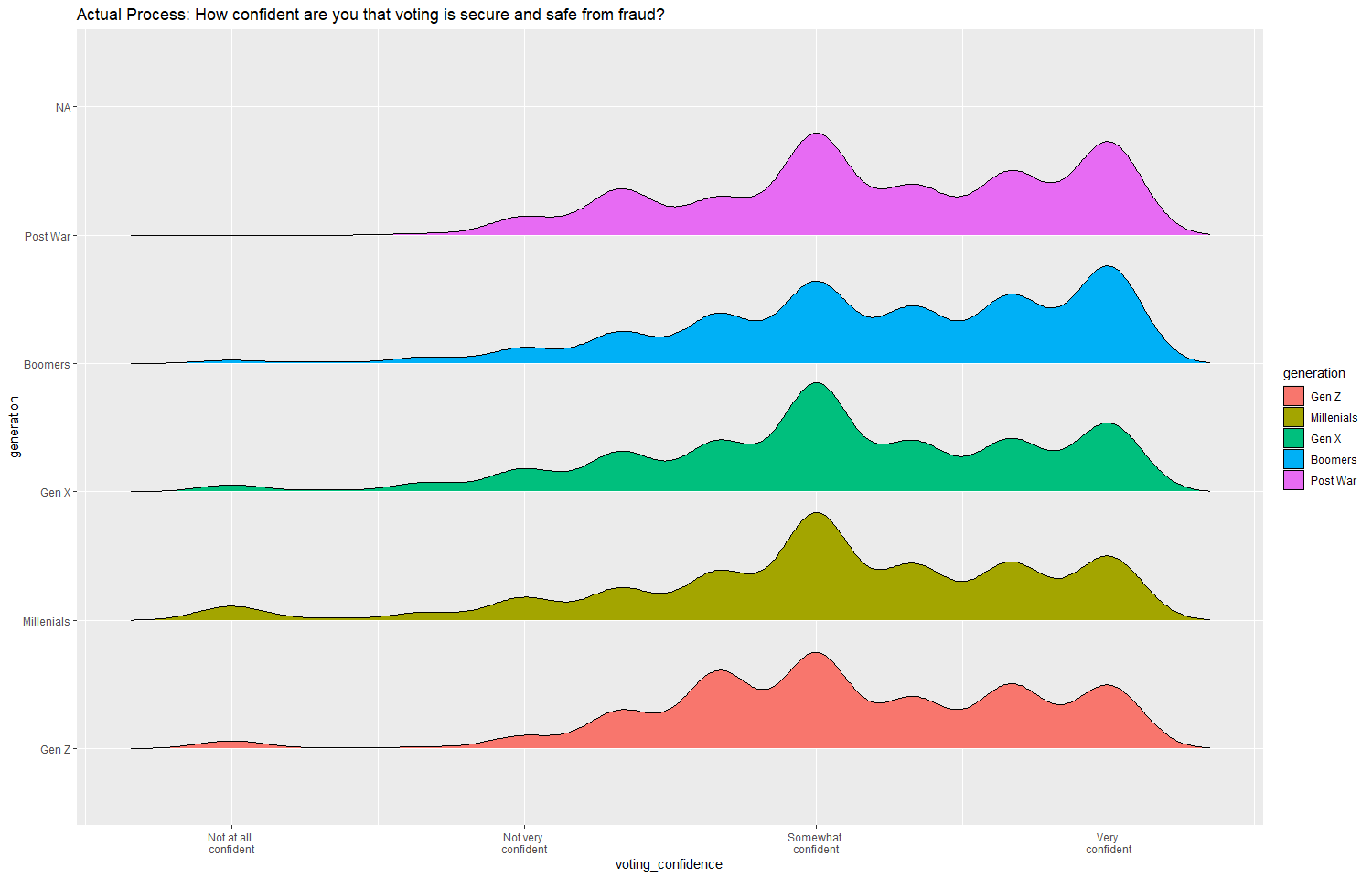
data <- data %>%   
 mutate(voting\_confidence = rowMeans(select(data,"confidence\_method\_person\_machines", "confidence\_method\_person\_ballots", "confidence\_method\_mail\_ballots")))  
   
 data %>%   
 group\_by(generation) %>%   
 summarise(voting\_confidence = mean(voting\_confidence, na.rm = T))

## `summarise()` ungrouping output (override with `.groups` argument)

## # A tibble: 6 x 2  
## generation voting\_confidence  
## <fct> <dbl>  
## 1 Gen Z 3.10  
## 2 Millenials 3.04  
## 3 Gen X 3.06  
## 4 Boomers 3.22  
## 5 Post War 3.21  
## 6 <NA> 4

actualprocess <- data %>%   
 ggplot(aes(x=voting\_confidence, y=generation, fill = generation, height=..density..)) +  
 geom\_joy(scale=0.85) +  
labs(title = "Actual Process: How confident are you that voting is secure and safe from fraud?") +   
scale\_x\_continuous(breaks = c(1, 2, 3, 4), labels = c("Not at all \nconfident",  
 "Not very \nconfident",  
 "Somewhat \nconfident",  
 "Very \nconfident"))

actualprocess



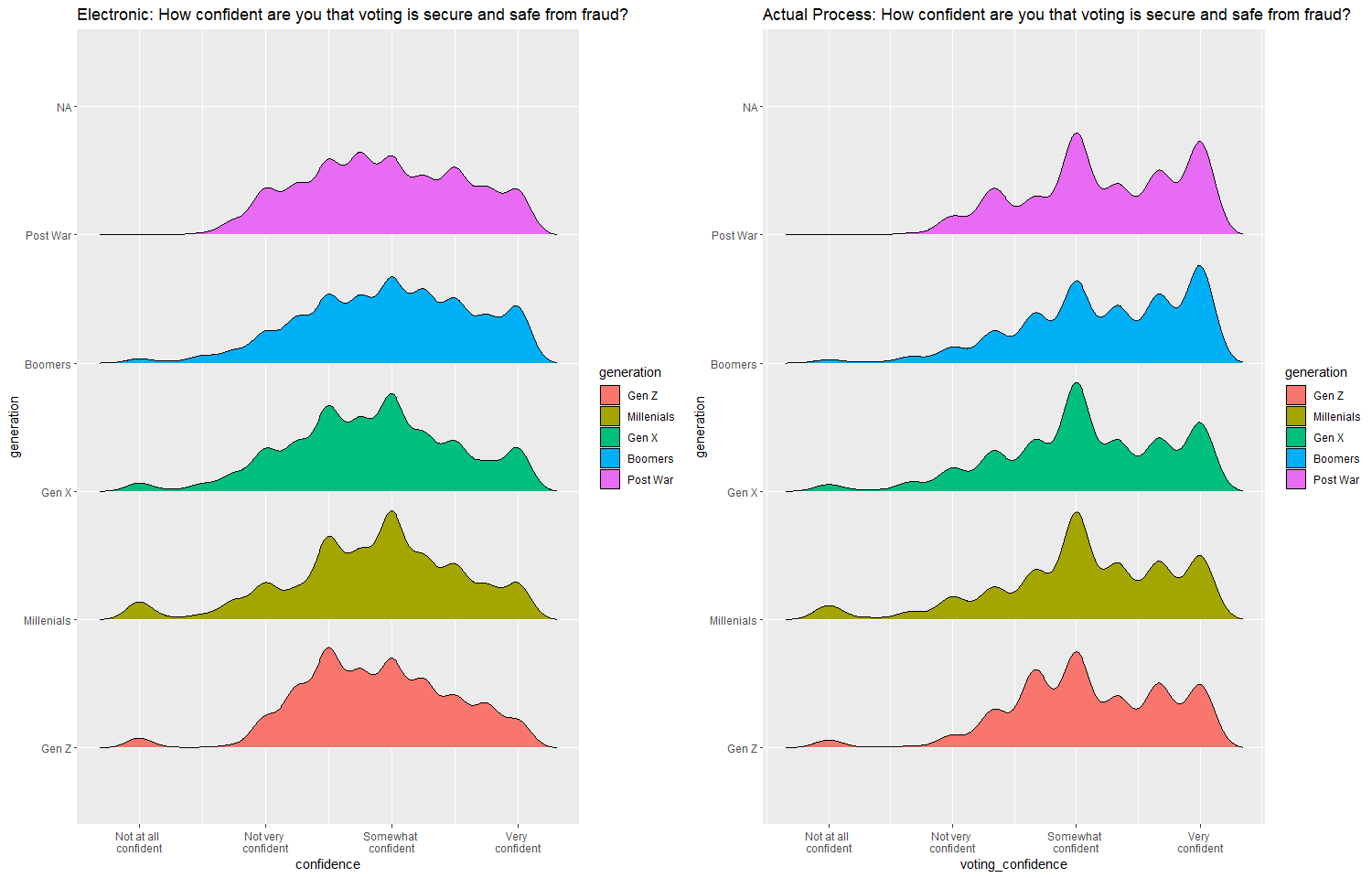
plot\_grid(electronic, actualprocess, ncol = 2)

## Picking joint bandwidth of 0.103

## Warning: Removed 310 rows containing non-finite values (stat\_density\_ridges).

## Picking joint bandwidth of 0.115

## Warning: Removed 272 rows containing non-finite values (stat\_density\_ridges).



**Discussion**

*Referencing your visualizations from above, answer the three research questions. For question (3), be sure to describe and justify how you came up with your measure “confidence in the electoral process.”*

I wanted to create a graphic that cleanly showed the age breakdown of the data by generation, in adddition to accurately showcasing the variability inside the generation as well. That is a good goal! Boomers are the largest group in the sample, followed by Millenials, Gen X, Gen Z, and Post War. Boomers also seem to have the most consistent engagement in voting compared to the other generations, while Millenials show the least amount of voter turnout considering their size. For those of Post War generation (a considerably smaller group), most of the members that are still alive consistently vote.

My boxplot comparing voting frequency by age and political party is interesting. This was an excellent graph It looks like in the “always voting” category, there is a much larger variation in age for Republicans than Democrats, possibly showing that young (more likely middle age) a violin plot would have showed that better as it gives you the actual distribution of age Republicans are more likely to vote than middle age Democrats. There means are the same however. In the “sporadic” category, the mean age of Democrats is higher than that of Republicans. It also the same for the “rarely, never” category. In this latter actegory, there looks to be a positive skew, meaning the tail on the right side of the distribution is longer or fatter. The mean and median will be greater than the mode. The mean will be greater than the median.

This contrasts to the “always” category where it is negatively skewed.

I decided to keep looking at the difference between generations for this question because I think it gives me interesting insight to what matters to people in each age bracket. Here is what I found for each specific group. It would be helpful to reference which figure this refers to

**Gen Z:**

* Do Vote: It seems that most people from Gen Z vote because they view voting as their civic duty. Compared to all other generations, they found voting to be the most easy.
* Don’t Vote: It seems that those who don’t from Gen Z either feel like the system is too broken or it doesn’t matter who wins.

**Millenials:**

* Do Vote: It seems that most Millenials vote because they view voting as their civic duty. Compared to all the other generations, Millenials are the most motivated to vote because they care about a specific issue.
* Don’t Vote: Millenials that don’t vote most often disliked candidates or felt that it didn’t matter who won.

**Gen X:**

* Do Vote: It seems that most people from Gen X vote because they view voting as their civic duty.
* Don’t Vote: Gen X people that don’t vote most often disliked candidates or felt that it didn’t matter who won.

**Boomers:**

* Do Vote: It seems that most Boomers vote because they view voting as their civic duty.
* Don’t Vote: Boomers that don’t vote most often disliked candidates.

**Post War:**

* Do Vote: It seems that most Post War people vote because they view voting as their civic duty. Post War people also displayed the most partisanship as a reason to vote.
* Don’t Vote: The Post War members that don’t vote most often disliked candidates.

Very thorough!

3.

For this question, I calculated people’s confidence in the electoral process by averaging their answers from Question 17, which consisted of a list of the following:

1. **In-person voting machines**
2. **Paper ballots cast in person**
3. **Paper ballots submitted by mail**
4. **Electronic votes submitted online or by email**

Even though we aren’t able to vote by electronic ballot, I think it is important to take this measure into consideration. For many members of Gen Z, electronic voting is a definite reality of the future. Failing to consider how this measure plays into the overall electoral process would be a disservice–I am also highly curious to see generational attittudes could be constructed aroudn this issue. Just because America does not yet use electronic voting does not mean that people can not have an opinion on it that is thoughtfully constructed. I agree with that Averaging the respondent’s answers gives me an easy way to interpret the differences between generation.Therefore, I created two plots, one that considered electronic voting and one that did not. Good!

Almost all generations have some outliers that have no confidence in the voting system whatsoever, creating a tale on the density plot. In this electronic voting graph, all the plots appear platykurtic, albeit, it is shaped more like a normal distribution than the graph considering only person-machine, person-ballot, and mail-ballot voting. However, in the electronic graph, most respondents across every generation seem to be somewhat confident in the electoral process. The electronic graph has fewer “extreme” events than the actual process graph. Eliminating electronic voting from the graph brought to light new insights:

* **all generations became skewed to become more confident in the electoral process. More people fell into the “very confident” bucket compared to the electronic graph.**
* **Boomers and Post War generations seem the most confident in the electoral process, while it appears that Gen X, Millenials, and Gen Z have the most outliers that have no confidence in the system. Well that is a bit harsh. They have less confidence, but many do have confidence!**
* **it seems that including electronic voting makes people less confident in the electoral system (rightfully so, after so much hacking on election day!)**

Overall takeaway, I believe that Boomers are the most actively engaged in the political system while also believing in it. Millenials, on the other hand, are the most disillusioned and the least participatory in terms of their ratio in the population. good!

# How do people experience voting differently?

# Include all code required to generate your visualizations here.  
  
#Creation of new variable  
data <- data %>%   
 mutate(agerange = case\_when(ppage < 30 ~ "under 30",   
 ppage >= 30 & ppage <= 65 ~ "30-50",  
 ppage > 50 ~ "over 50"),

Agerange = factor(data$agerange, levels = c("under 30", "30-50", "over 50")))  
data$agerange <- factor(data$agerange, levels = c("under 30", "30-50", "over 50"))

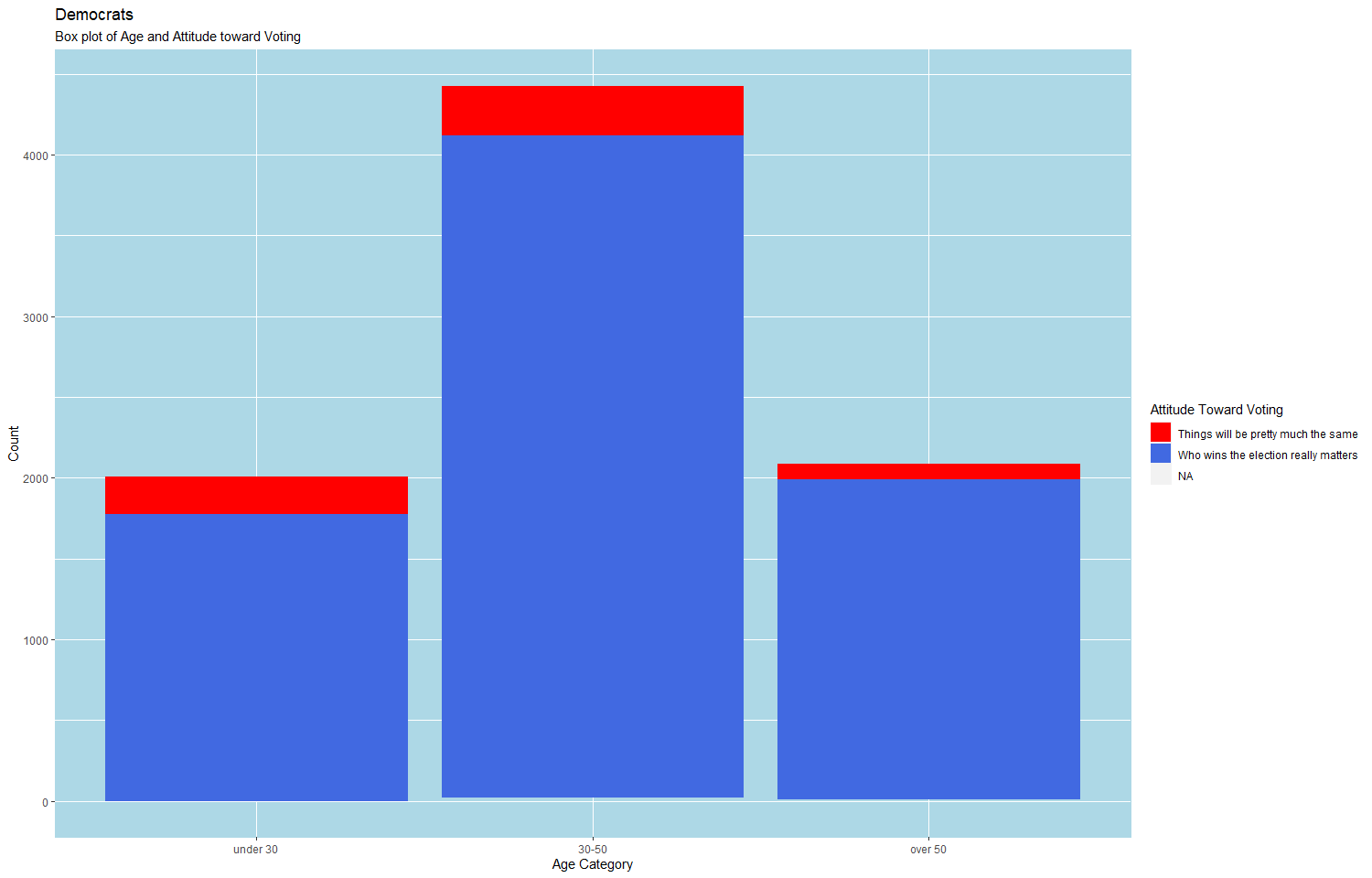
You can easily put the factor conversion in the block of mutations  
  
Democrats <- data %>%   
 filter(political\_affiliation == "Democrat")  
   
Republicans <- data %>%   
 filter(political\_affiliation == "Republican")  
  
  
   
 ggplot(Democrats, aes(x=agerange, fill=attitude\_toward\_voting)) +  
 geom\_histogram(stat = "count") +  
 labs(x = "Age Category",  
 y = "Count",  
 fill = "Attitude Toward Voting",   
 title = "Democrats",  
 subtitle = "Box plot of Age and Attitude toward Voting") +  
 scale\_fill\_manual(values = c("red", "royalblue")) +   
 theme(  
 panel.background = element\_rect(fill = "lightblue", color = "lightblue"))

## Warning.background = element\_rect(fill = "lightblue", color = "lightblue"))

ggplot(Republicans, aes(x=agerange, fill=attitude\_toward\_voting)) +  
 geom\_histogram(stat = "count") +  
 labs(x = "Age Category",  
 y = "Count",  
 fill = "Attitude Toward Voting",   
 title = "Republicans",  
 subtitle = "Box plot of Age and Attitude toward Voting") +  
 scale\_fill\_manual(values = c("red", "royalblue")) +  
 theme(  
 panel.background = element\_rect(fill = "#FF6347AA", color = "#FF6347AA"))

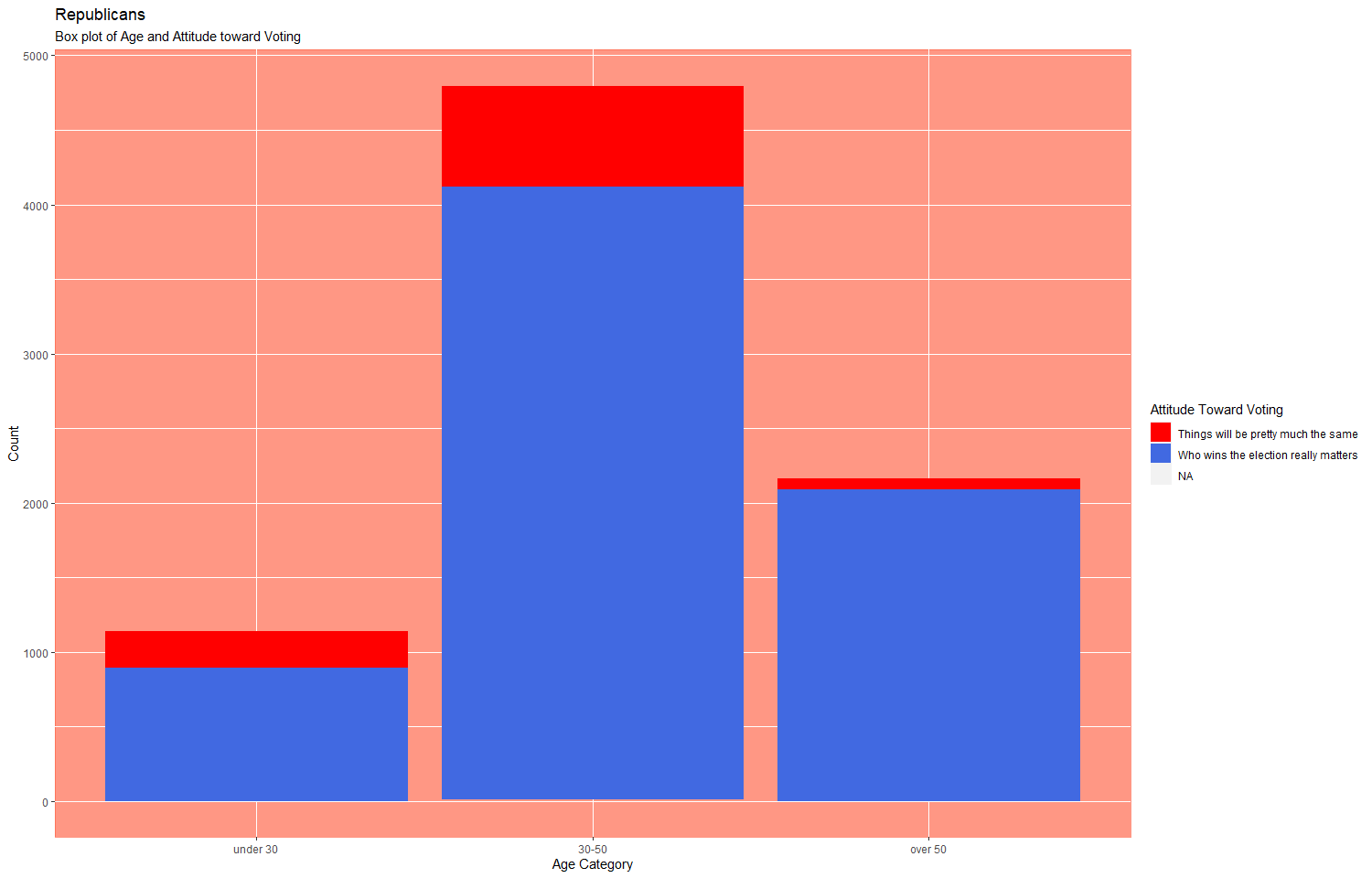
## Warning: Ignoring unknown parameters: binwidth, bins, pad

**ggplot**(Republicans, **aes**(x=agerange, fill=attitude\_toward\_voting)) **+**  
 **geom\_histogram**(stat = "count") **+**  
 **labs**(x = "Age Category",  
 y = "Count",  
 fill = "Attitude Toward Voting",   
 title = "Republicans",  
 subtitle = "Box plot of Age and Attitude toward Voting") **+**  
 **scale\_fill\_manual**(values = **c**("red", "royalblue")) **+**  
 **theme**(  
 panel.background = **element\_rect**(fill = "#FF6347AA", color = "#FF6347AA"))

## Warning: Ignoring unknown parameters: binwidth, bins, pad

I see what you are doing here with the background. I applaud it for being innovative, but it is not the most pretty graph, right?

Because Q5 is a forced choice (pick things will be same, Who wins the election, and didn’t cchoose(NA)) I wonder if the results would have been more clear if you had done percentages instead of counts. So each bar stretches to a 100% and you see % that picks what. This would make this forced choice variable character more intuitive if you do not know what kind of item Q5 is.

ggplot(Republicans, aes(x=agerange, fill=attitude\_toward\_voting)) +  
 geom\_histogram(stat = "count") +  
 labs(x = "Age Category",  
 y = "Count",  
 fill = "Attitude Toward Voting",   
 title = "Republicans",  
 subtitle = "Box plot of Age and Attitude toward Voting") +  
 scale\_fill\_manual(values = c("red", "royalblue")) +  
 theme(  
 panel.background = element\_rect(fill = "#FF6347AA", color = "#FF6347AA"))

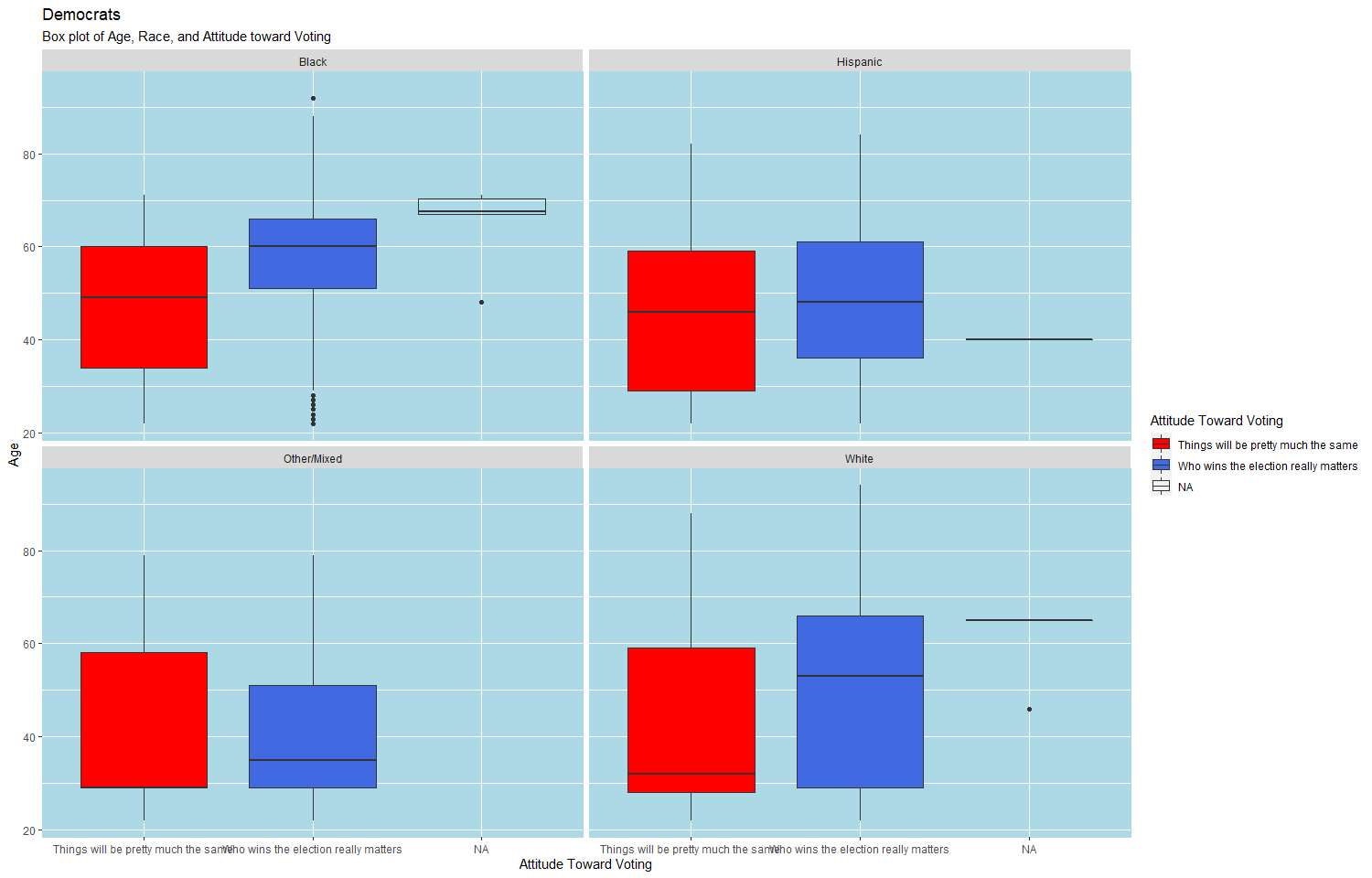
## Warning: Ignoring unknown parameters: binwidth, bins, pad

dems\_race <- Democrats %>%  
 ggplot() +  
 geom\_boxplot(aes(x = attitude\_toward\_voting, y = ppage, fill = attitude\_toward\_voting, na.rm = TRUE)) +  
 labs(x = "Attitude Toward Voting",  
 y = "Age",  
 fill = "Attitude Toward Voting",   
 title = "Democrats",   
 subtitle = "Box plot of Age, Race, and Attitude toward Voting") +  
 scale\_fill\_manual(values = c("red", "royalblue")) +  
 facet\_wrap (~ race)

## Warning: Ignoring unknown aesthetics: na.rm

Notice that na.rm is not something that aes() works with. But that’s ok, because ggplot will kick out NAs by default

dems\_race + theme(  
 panel.background = element\_rect(fill = "lightblue", color = "lightblue"))

I’m struggling with this graph. I would have done a coord\_flip so you can read the labels of attitude better. And you can hide the legend, because the axes already show what’s what.

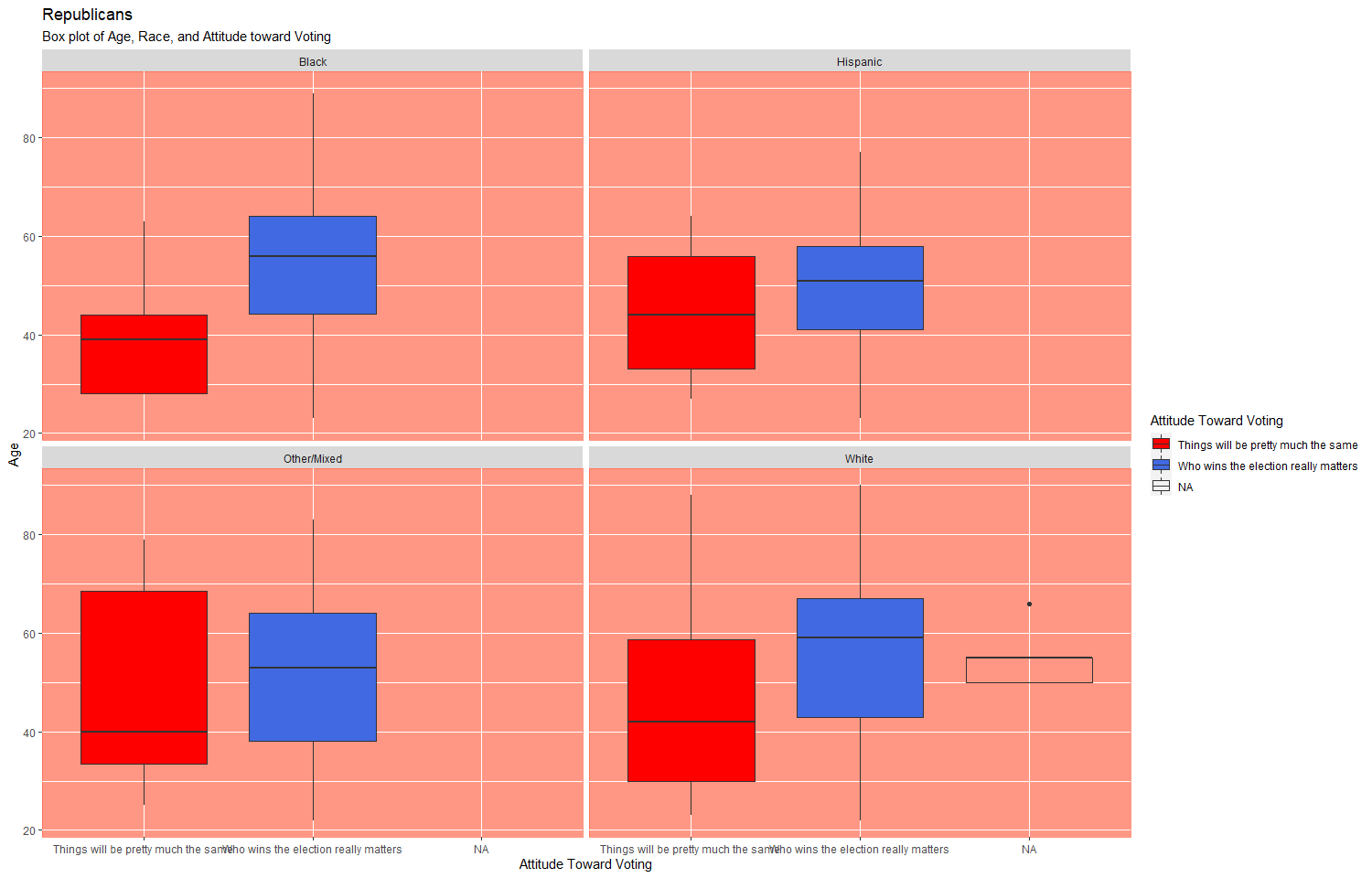
How would you interpret the graph? The median age of people who choose “things will be the same” over “who wins the election really matters” is close to 50. The median age of people who choose “who wins the election really matters” over “things will stay the same” is 60. This implies that older respondents had a more positive outlook on the impact of the elections than younger respondents.

^ This graph is ok, but it took me a while to interpret it. I wonder if a stacked bar chart with % would be more intuitive, where the bars are grouped by age group and racial group. Or just age group and facet\_wrap by race.

repubs\_race <- Republicans %>%  
 ggplot() +  
 geom\_boxplot(aes(x = attitude\_toward\_voting, y = ppage, fill = attitude\_toward\_voting, na.rm = TRUE)) +  
 labs(x = "Attitude Toward Voting",  
 y = "Age",  
 fill = "Attitude Toward Voting",   
 title = "Republicans",   
 subtitle = "Box plot of Age, Race, and Attitude toward Voting") +  
 scale\_fill\_manual(values = c("red", "royalblue")) +  
 facet\_wrap ( ~ race)

## Warning: Ignoring unknown aesthetics: na.rm

repubs\_race + theme(  
 panel.background = element\_rect(fill = "#FF6347AA", color = "#FF6347AA"))

**Demographic Group:**

Instead of exploring generations, I decided to look at a range of ages. Why? I created a new variable called “agerange” that is composed of: under 30 (because I feel like this age range still constitutes as “young”, 30 to 50 (middle-age), and 50 and over (older people).

I decided to look at how age and political party influence attitude toward voting. I’ve generally interested in how bipartisanship influences people’s perspective on the outcome. Firstly, I created two new data frames solely looking at Democrats and Republicans using filter. I then plotted age on the x-axis as a continuous variable with y as a count. The histograms Where are the histograms? show that a vast majority of voters believe that who wins the election really matters. However, a larger amount of Republicans, specifically in the 30-50 mid-life range believe it doesn’t matter because things will stay pretty much the same. From these histograms, we can also see that a lot of more young people under 30 vote Democrat than Republican. From this, we can see that 30-50 Republicans tend to be more disillusioned by the electoral system than any other group.

I additionally decided to create a second plot comparing age, race and attitude toward voting by party.The race categories consisted of: Black, Hispanic, White, and Other/Mixed. This is totally ignoring mainly Asian Americans and other individuals of color, but I unfortunately don’t have access to more information on the participants’ identities. That is a good observation. These people may have been undersampled for some reason? Definitely a problem!

**Discussion** Interestingly, older people in both parties tend to view that who wins the election really matters. There is slightly more variability in the age of Democrats who think who wins the election really matters than Republicans (perhaps more younger Democrats are motivated to vote because of this? Further analysis would be needed.)

**Republicans:**

Black:

* things will pretty much be the same: median age around 40.
* who wins really matters: mean age around 55.

Hispanic:

* things will pretty much be the same: mean age around 45.
* who wins really matters: mean age around 50.

Other:

* things will pretty much be the same: mean age around 40. Larger IQR, more variability. Perhaps because “other” encompasses many distinctly different groups.
* who wins really matters: mean age around 55.

White:

* things will pretty much be the same: mean age around 40.
* who wins really matters: mean age around 60.

Verdict— Older Republicans of all race categories tend to feel that who wins the election really matters while young Republicans of all race categories believe that things pretty much stay the same.

Perhaps older people, since they have lived longer, are able to see the impact each presidency has on their life?

The “other” category shows the most positive skew, meaning the tail on the right side of the distribution is longer or fatter. The mean and median will be greater than the mode.

**Democrats:**

Black:

* things will pretty much be the same: mean age around 50.
* who wins really matters: mean age around 60.

Hispanic:

* things will pretty much be the same: mean age around 45.
* who wins really matters: mean age around 50.

Other:

* things will pretty much be the same: mean age around 30.
* who wins really matters: mean age around 35.

White:

* things will pretty much be the same: mean age around 35.
* who wins really matters: mean age around 55.

Verdict — It seems that younger Democrats tend to feel that things pretty much stay the same as compared to their Republican counterparts. There seems to be the most variability between the two groups for white individuals.

The other category shows the most positive skew, meaning the tail on the right side of the distribution is longer or fatter. The mean and median will be greater than the mode. This is probably because so many different ethnic groups were lumped into one.

Overall, for both parties, there was the least amount of variability in age between the two stances among Hispanics.