Extension of Alex's Vignette of ggplot2

Jian Quan Chen

2023-04-24

INTRODUCTION

ggplot2 was the first of Hadley Wickham's tidy packages and was intended to simplify and streamline the appearance of R graphics. In this vignette, we will walk through key plots in ggplot2 using the 'congress_age' dataset from fivethirtyeight and best tidy practices.

First, load the fivethirtyeight package and the congress_age dataset:

```
# install.packages("fivethirtyeight")
library(fivethirtyeight)
```

```
## Some larger datasets need to be installed separately, like senators and
## house_district_forecast. To install these, we recommend you install the
## fivethirtyeightdata package by running:
## install.packages('fivethirtyeightdata', repos =
## 'https://fivethirtyeightdata.github.io/drat/', type = 'source')
```

```
data("congress_age")
str(congress_age)
```

```
## Classes 'tbl_df', 'tbl' and 'data.frame':    18635 obs. of  13 variables:
##  $ congress  : int  80 80 80 80 80 80 80 80 80 80 ...
##  $ chamber   : chr  "house" "house" "house" "house" ...
##  $ bioguide  : chr  "M000112" "D000448" "S000001" "E000023" ...
##  $ firstname : chr  "Joseph" "Robert" "Adolph" "Charles" ...
##  $ middlename: chr  "Jefferson" "Lee" "Joachim" "Aubrey" ...
##  $ lastname  : chr  "Mansfield" "Doughton" "Sabath" "Eaton" ...
##  $ suffix    : chr  NA NA NA NA ...
##  $ birthday  : Date, format: "1861-02-09" "1863-11-07" ...
##  $ state     : chr  "TX" "NC" "IL" "NJ" ...
##  $ party     : chr  "D" "D" "D" "R" ...
##  $ incumbent : logi  TRUE TRUE TRUE TRUE FALSE FALSE ...
##  $ termstart : Date, format: "1947-01-03" "1947-01-03" ...
##  $ age       : num  85.9 83.2 80.7 78.8 78.3 78 77.9 76.8 76 75.8 ...
```

Load the tidyverse and ggplot2 packages:

```
library(ggplot2)
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v tibble  3.1.8      v dplyr   1.0.10
## v tidyr   1.2.1      v stringr 1.5.0
## v readr   2.1.3      v forcats 0.5.2
## v purrr   1.0.0
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(dplyr)
```

What are the top 10 most common first names of congresspeople?

First, we need to use the dplyr package to count and then sort the number of first names in the dataset.

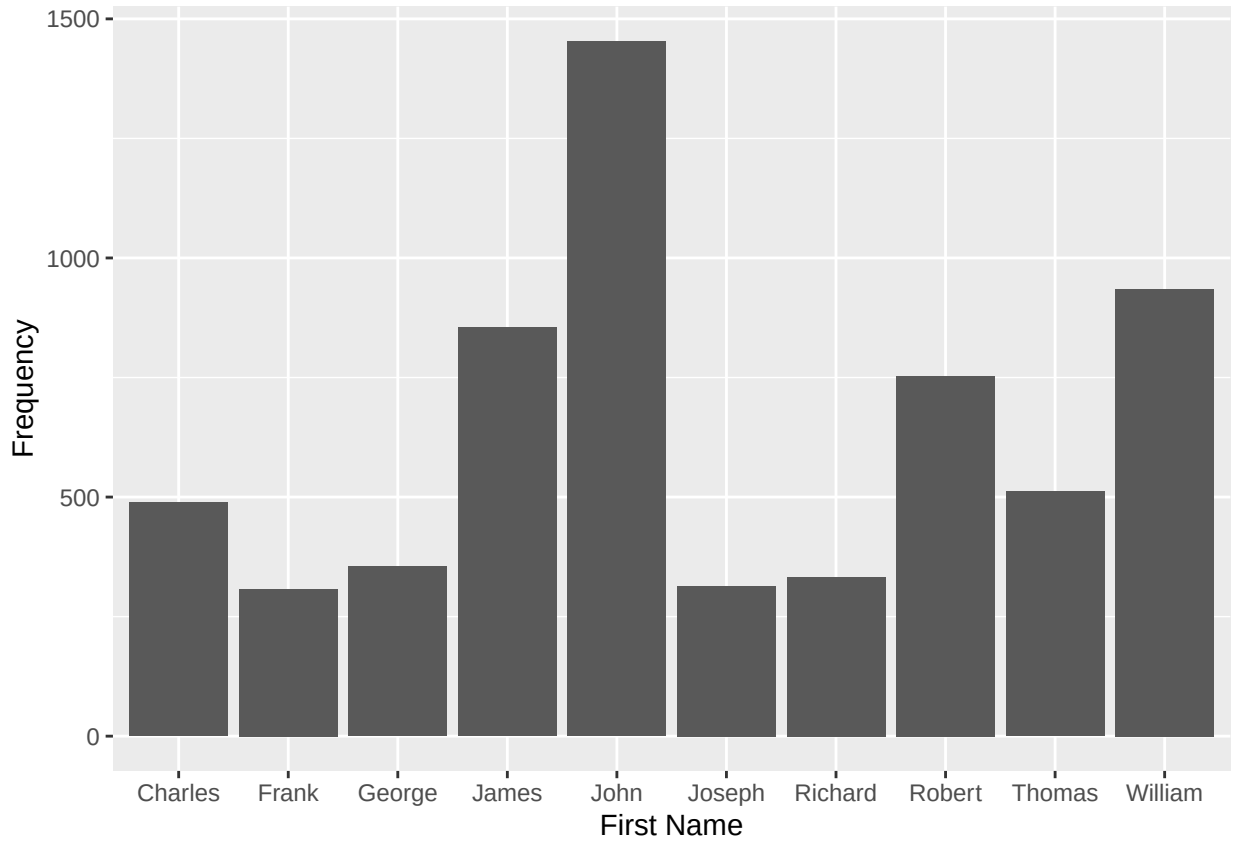
```
first_names <- congress_age %>%
  group_by(firstname) %>%
  count(firstname) %>%
  arrange(desc(n))
head(first_names)
```

```
## # A tibble: 6 x 2
## # Groups:   firstname [6]
##   firstname     n
##   <chr>     <int>
## 1 John       1453
## 2 William     935
## 3 James       855
## 4 Robert      753
## 5 Thomas      512
## 6 Charles     488
```

This uses the **group_by** function to group the congresspeople by their first names so they we can **count** them, and then we **arrange** them in descending (**desc**) order by the count (**n**) we generated.

Barplot with geom_bar Barplots with geom_bar are a very quick way to look at summary data like counts. Although geom_bar will do the counting for you, here I are passing a dataframe that has already been summarized in counts so I will use the **stat="identity"** parameter inside of geom_bar.

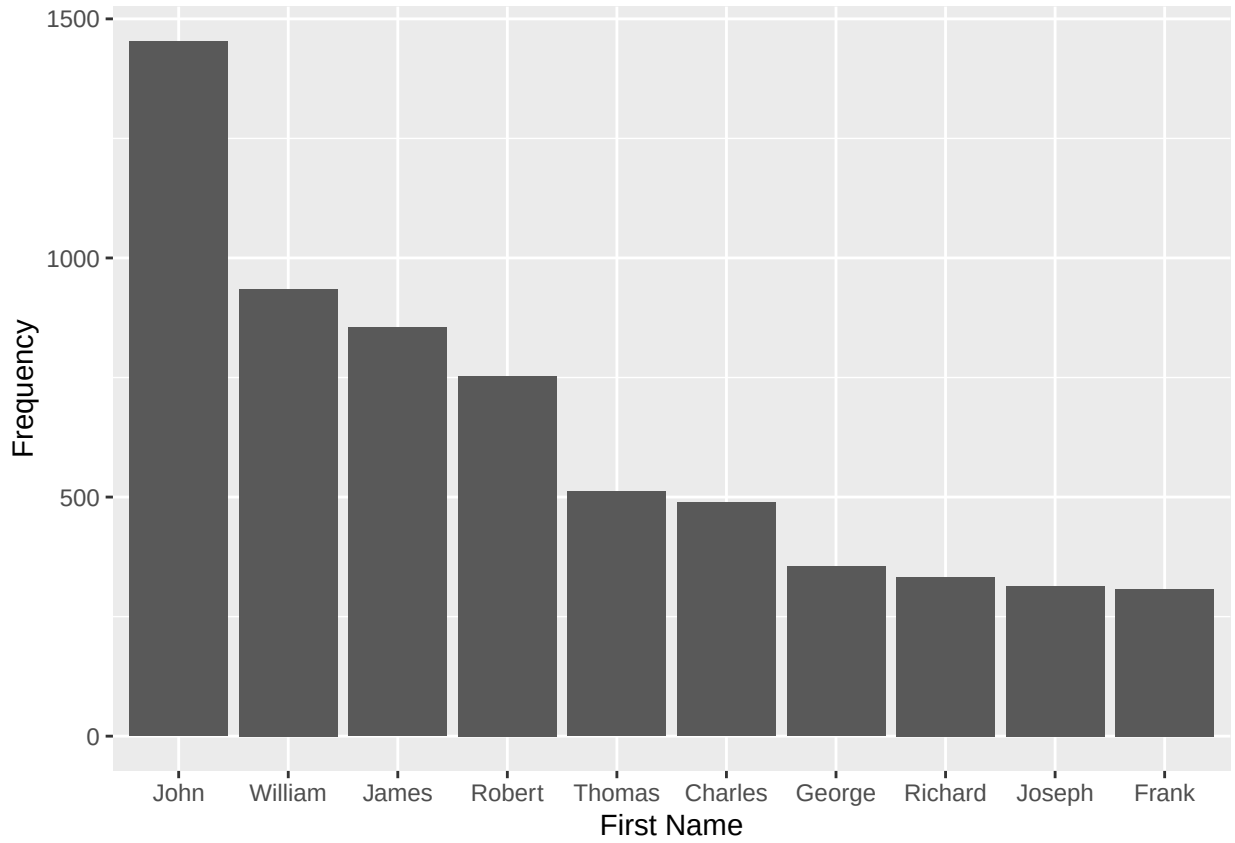
```
first_names[1:10,] %>%
  ggplot(aes(y = n, x = firstname)) +
  geom_bar(stat = "identity") +
  labs(x = "First Name", y = "Frequency")
```



Please note that x and y labels are added by using the **labs()** function. Unlike with the dplyr or tidyverse, ggplot requires + signs rather than a %>% to separate the statements. For all ggplots the aesthetic mapping **aes()** is vital as well as some form of geom statement. What is passed through the aesthetic determines what is on the x and y axis.

For example, by default ggplot will place the x axis into alphabetical order rather than take the order provided by the table. To fix this I can pass an additional parameter **scale_x_discrete()**:

```
level_order <- first_names[1:10, "firstname"]
first_names[1:10,] %>%
  ggplot(aes(y = n, x = firstname)) +
  geom_bar(stat = "identity") +
  labs(x = "First Name", y = "Frequency") +
  scale_x_discrete(limits = level_order$firstname)
```



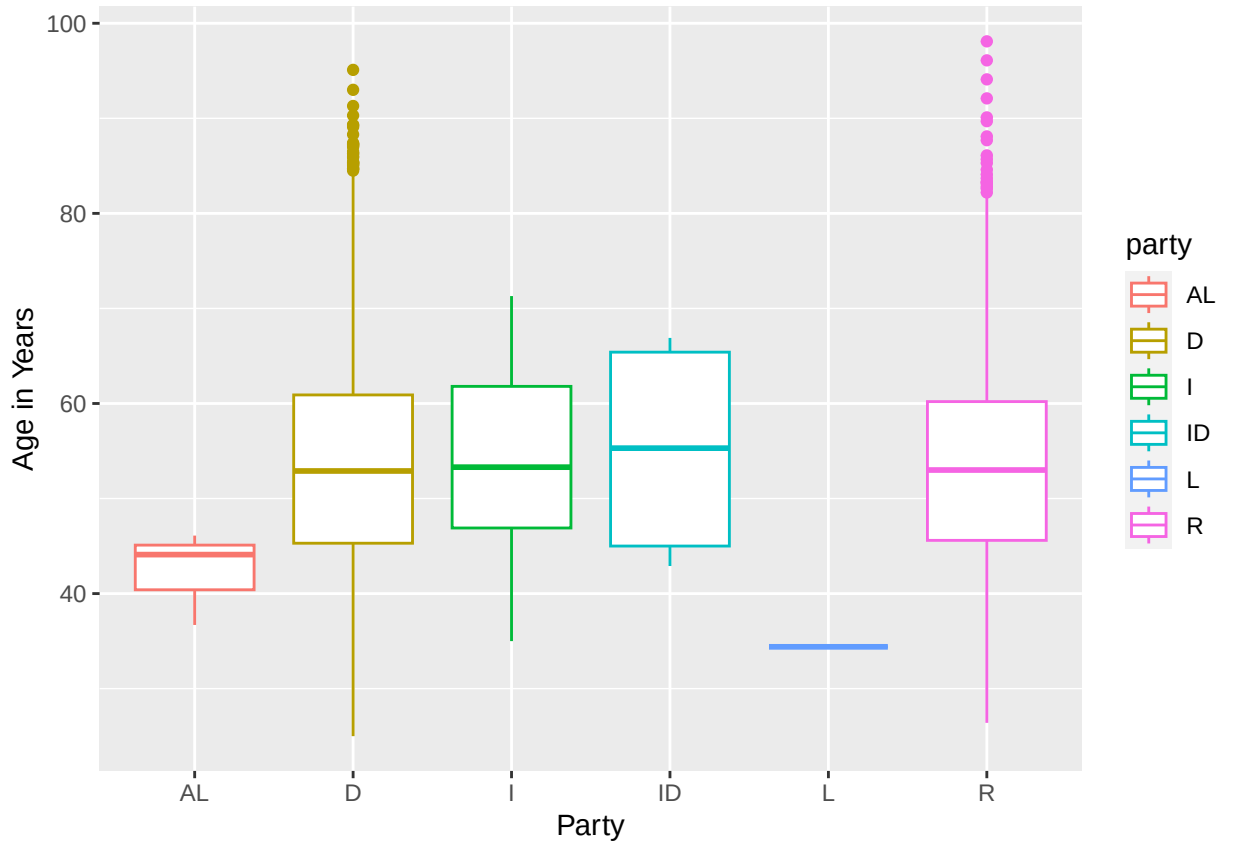
In the history of the US congress the frequency of the name John has outstripped other first name with William, James, and Robert not far behind.

Does the median age of a congress persons differ by political party?

In order to answer this question I will use a box plot on the raw dataset without any tidy manipulation.

Box plot with geom_boxplot. This will create a conventional box and whisker plot.

```
congress_age %>%
  ggplot(aes(x = party, y = age, colour = party)) +
  geom_boxplot() +
  labs(x = "Party", y = "Age in Years")
```

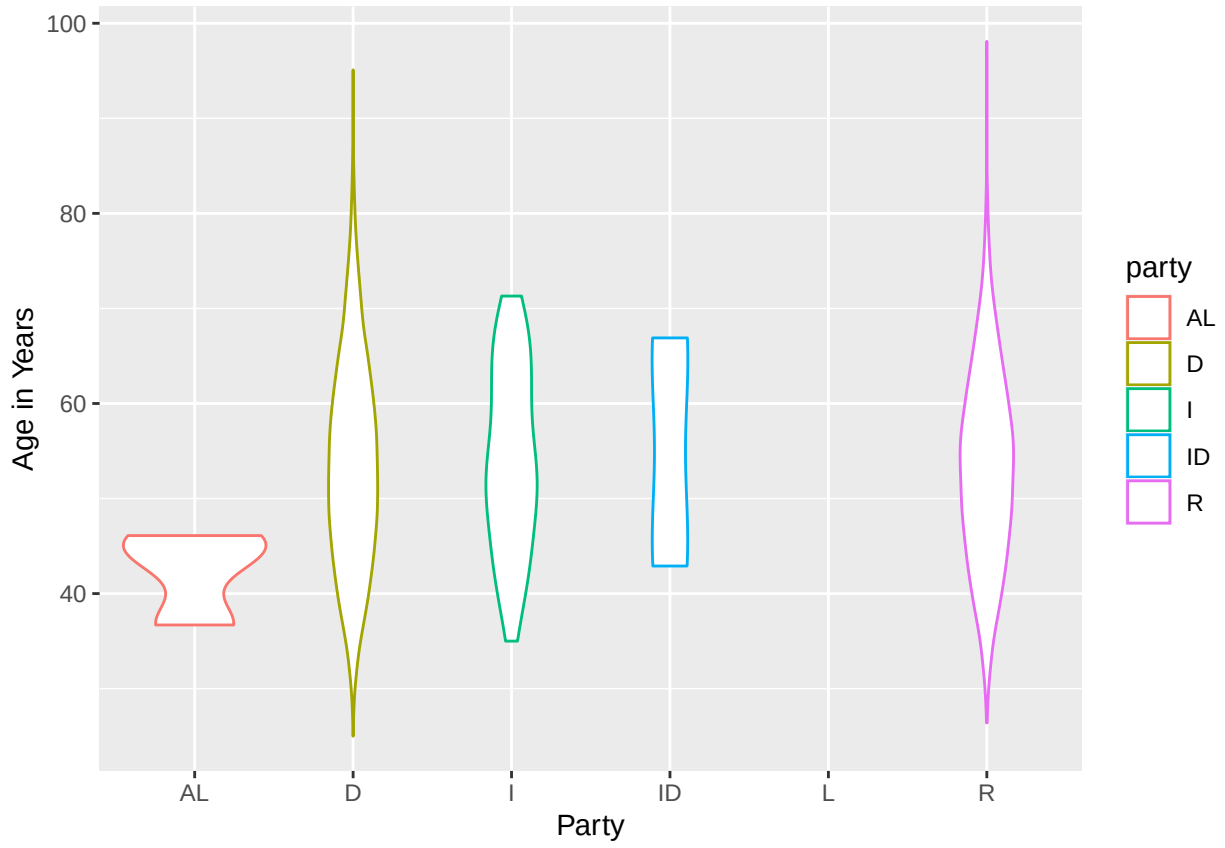


The median age is similar across all political parties, except the Libertarian(L) and the American Independent party(AL).

Violin plot with geom_violin. Violin plots are an alternative to box plots that add more information than a box plot in terms of the underlying distribution of the data. I will create a violin plot with the same data as above to demonstrate the additional information that can be obtained.

```
congress_age %>%
  ggplot(aes(x = party, y = age, colour = party)) +
  geom_violin() +
  labs(x = "Party", y = "Age in Years")
```

```
## Warning: Groups with fewer than two data points have been dropped.
```

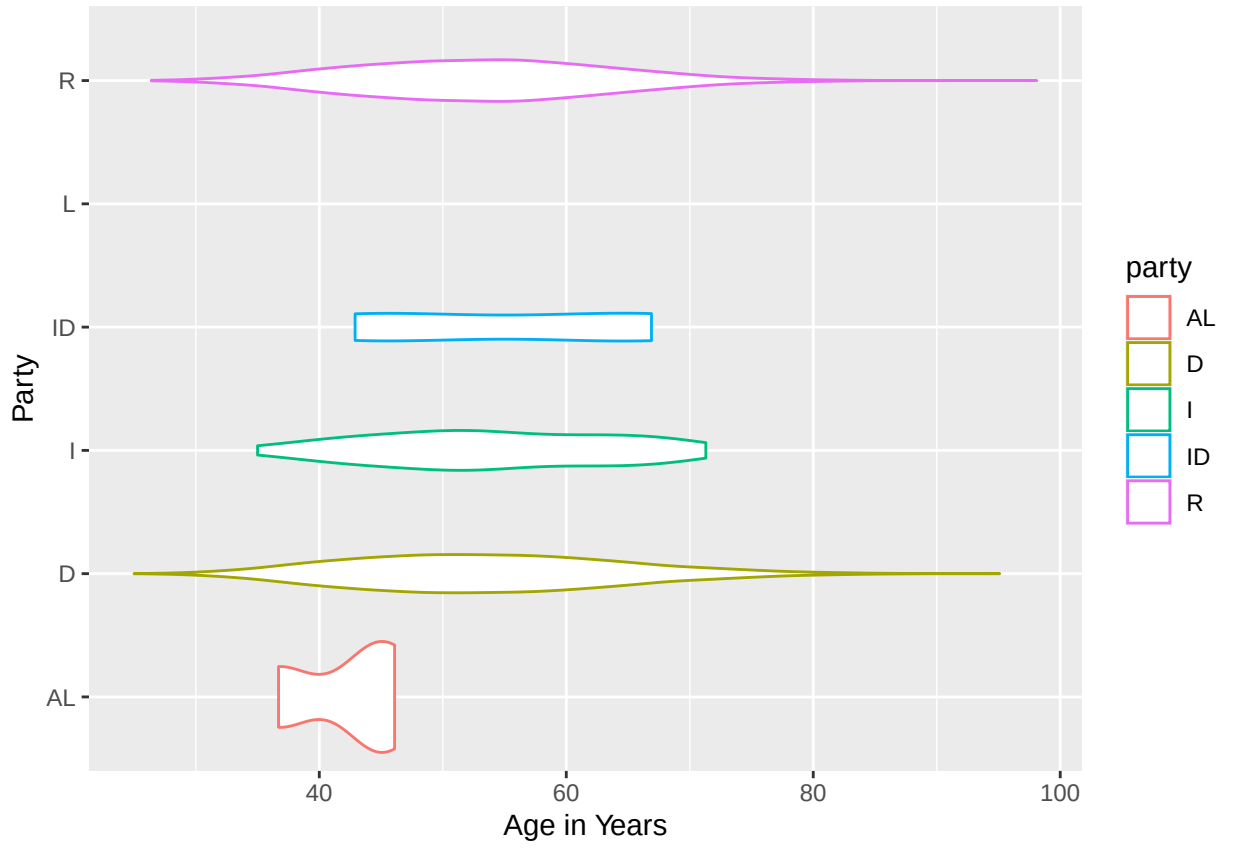


From this plot I can that the distribution of ages is most similar between Democrats(D) and Republicans(R). The Libertarian group is not shown because there was only one in the dataset.

What if we would like to visualize this plot horizontally instead? I can employ **coord_flip()** to flip the coordinates of the plot:

```
congress_age %>%
  ggplot(aes(x = party, y = age, colour = party)) +
  geom_violin() +
  labs(x = "Party", y = "Age in Years") +
  coord_flip()
```

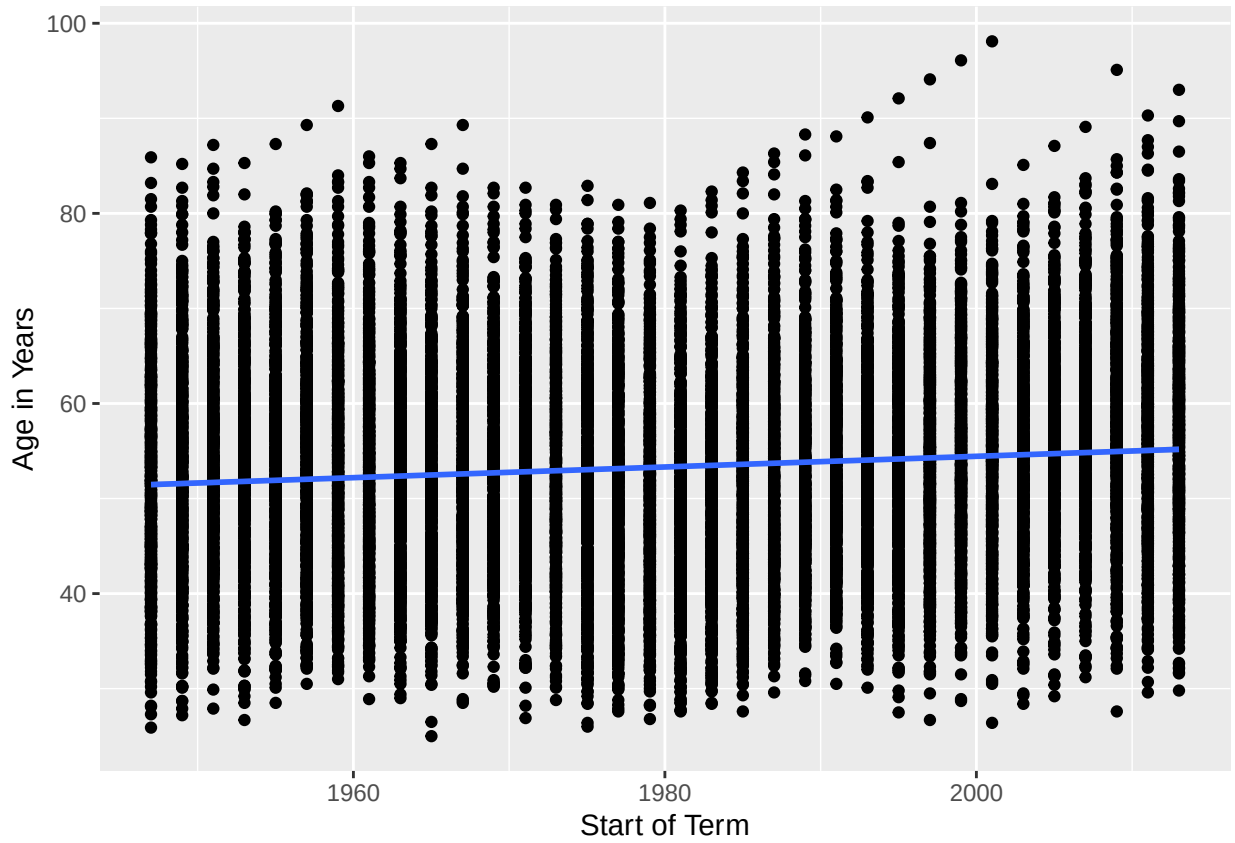
```
## Warning: Groups with fewer than two data points have been dropped.
```



Scatterplot with geom_point and geom_smooth. Scatterplots in ggplot are accomplished with **geom_point()** function and one can choose to add an an optional regression line to the data using either **geom_smooth()** or **geom_abline()**. However geom_abline requires that you have already calculated the line of best fit or another line before ploting. Geom_smooth is the ggplot replacement for baseR abline().

```
congress_age %>%
  ggplot(aes(x = termstart, y = age)) +
  geom_point() +
  geom_smooth(method = "lm") +
  labs(x = "Start of Term", y = "Age in Years")
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



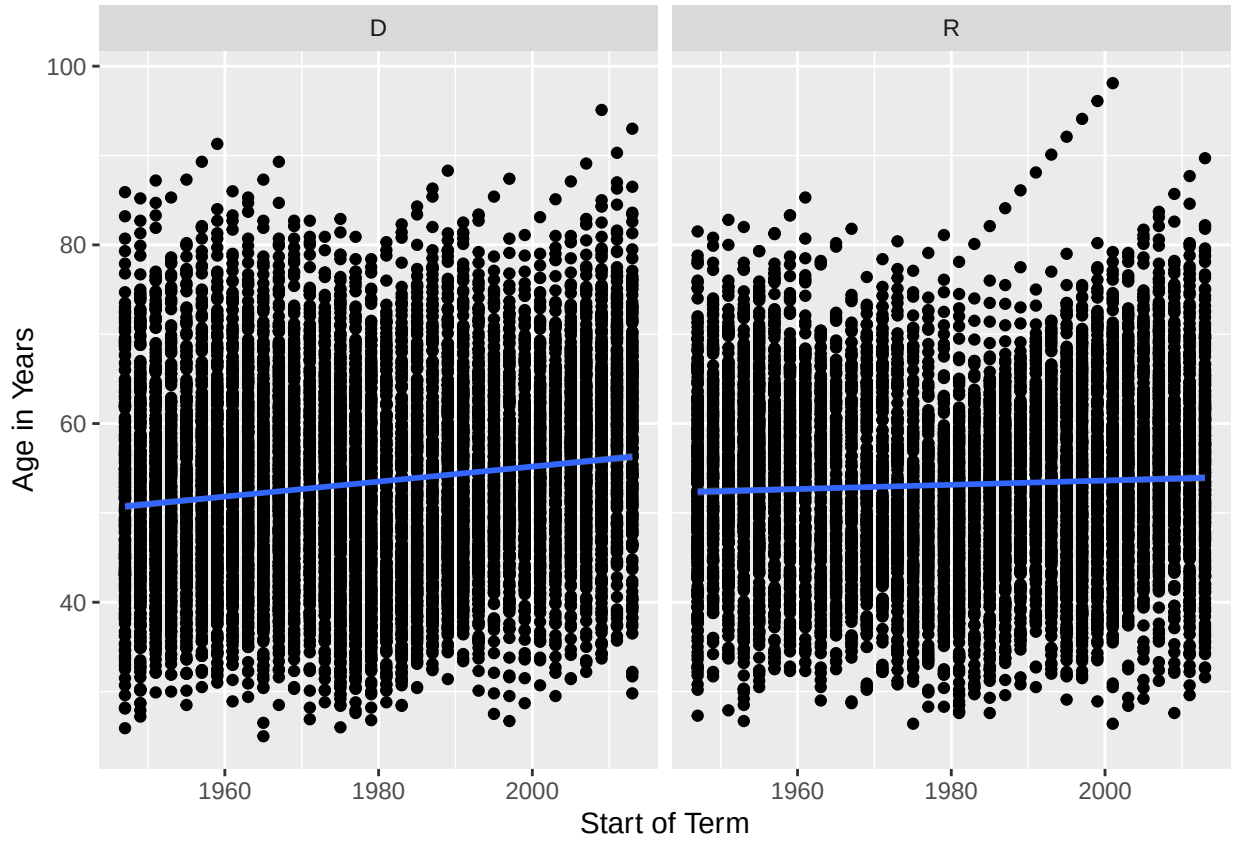
The scatterplot and regression line demonstrate that over time we are electing older people to congress.

Facet_wrap is one way to create multiple plot pannel within the same plot.

Lets use the above regression plot to test whether there is a difference between democrats and republicans at age at start of term. To create panels in a ggplot one can use either **facet_wrap()** or **facet_grid()**. Both functions perform similarly although facet_grid will create plots even for missing data where as facet_wrap will not. Here I used facet_wrap to demonstrate how the wraping works by adding "~z" where z is grouping variable.

```
congress_age %>%
  filter(party == "D" | party == "R") %>%
  ggplot(aes(x = termstart, y = age)) +
  geom_point() +
  geom_smooth(method = "lm") +
  labs(x = "Start of Term", y = "Age in Years") +
  facet_wrap(~party)
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



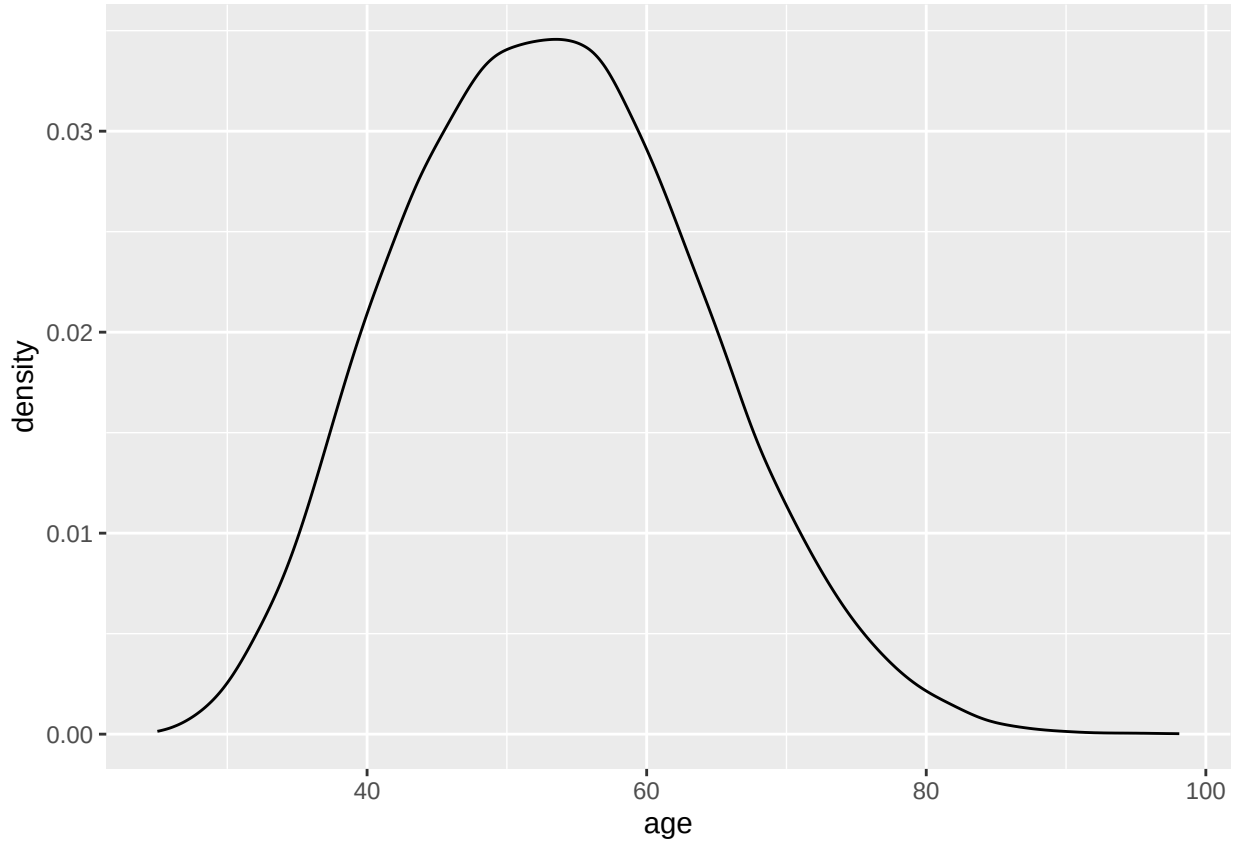
Extend

This is my extension to Alex Khaykin's vignette on the `ggplot2` package in the tidyverse. His "Create" assignment looked at key plots in `ggplot2` using the 'congress_age' dataset from fivethirtyeight. So far, he has demonstrated how to create a bar plot, boxplot, violin plot, and a scatterplot. I will expand on this by creating a density plot and histogram as well as some components in the ggplot2 package to improve data visualization.

Density Plot of Congress Age

The function `geom_density` plots a curve that shows the distribution of a continuous variable. The data is more densely distributed where the curve is highest. From the density plot, it looks like most of congress is around 50 years old.

```
congress_age %>%
  ggplot(aes(x = age)) +
  geom_density()
```

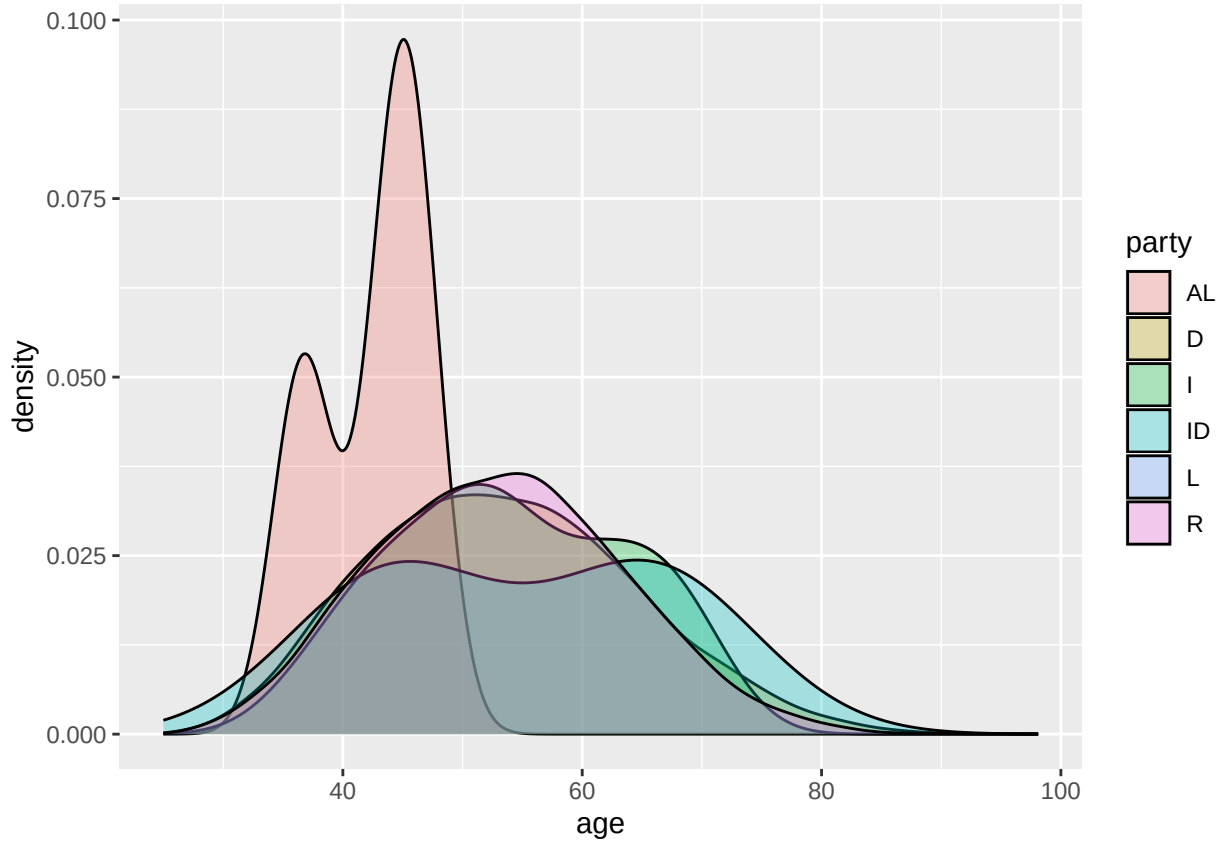


This density plot can be augmented to show the distribution of age in each party by adding the `fill` argument in `aes()`. I also changed the transparency of the curves by setting alpha to 0.3 in `geom_density()`.

```
congress_age %>%
  ggplot(aes(x = age, fill = party)) +
  geom_density(alpha = 0.3)
```

```
## Warning: Groups with fewer than two data points have been dropped.
```

```
## Warning in max(ids, na.rm = TRUE): no non-missing arguments to max; returning
## -Inf
```

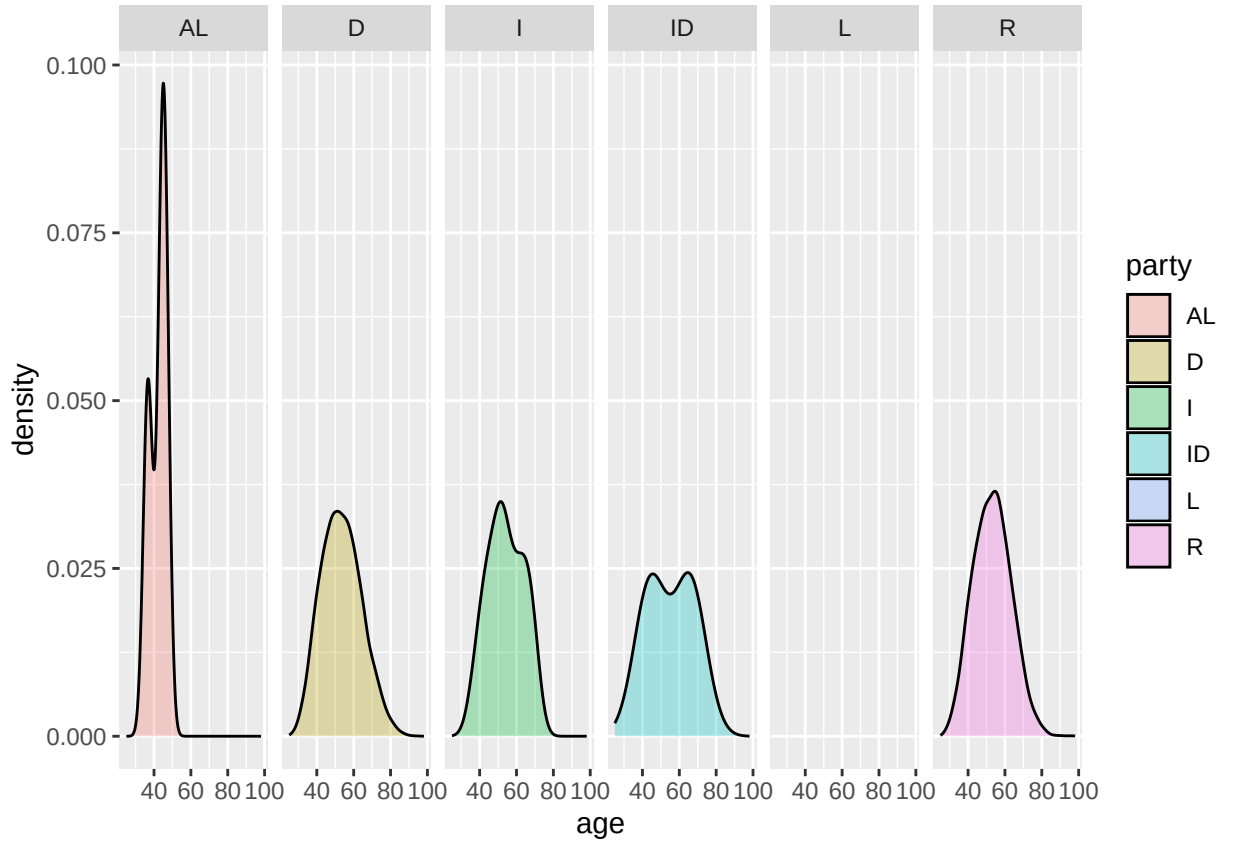


This density plot is a bit hard to read as all the curves are on top of each other. We can further clean it up by using the `facet_grid` function to create a grid of plots based on a categorical variable. Compared to the other parties, AL seems to be have the highest percentage of young congress members.

```
congress_age %>%
  ggplot(aes(x = age, fill = party)) +
  geom_density(alpha = 0.3)+
  facet_grid(~party,)
```

```
## Warning: Groups with fewer than two data points have been dropped.
```

```
## Warning in max(ids, na.rm = TRUE): no non-missing arguments to max; returning
## -Inf
```

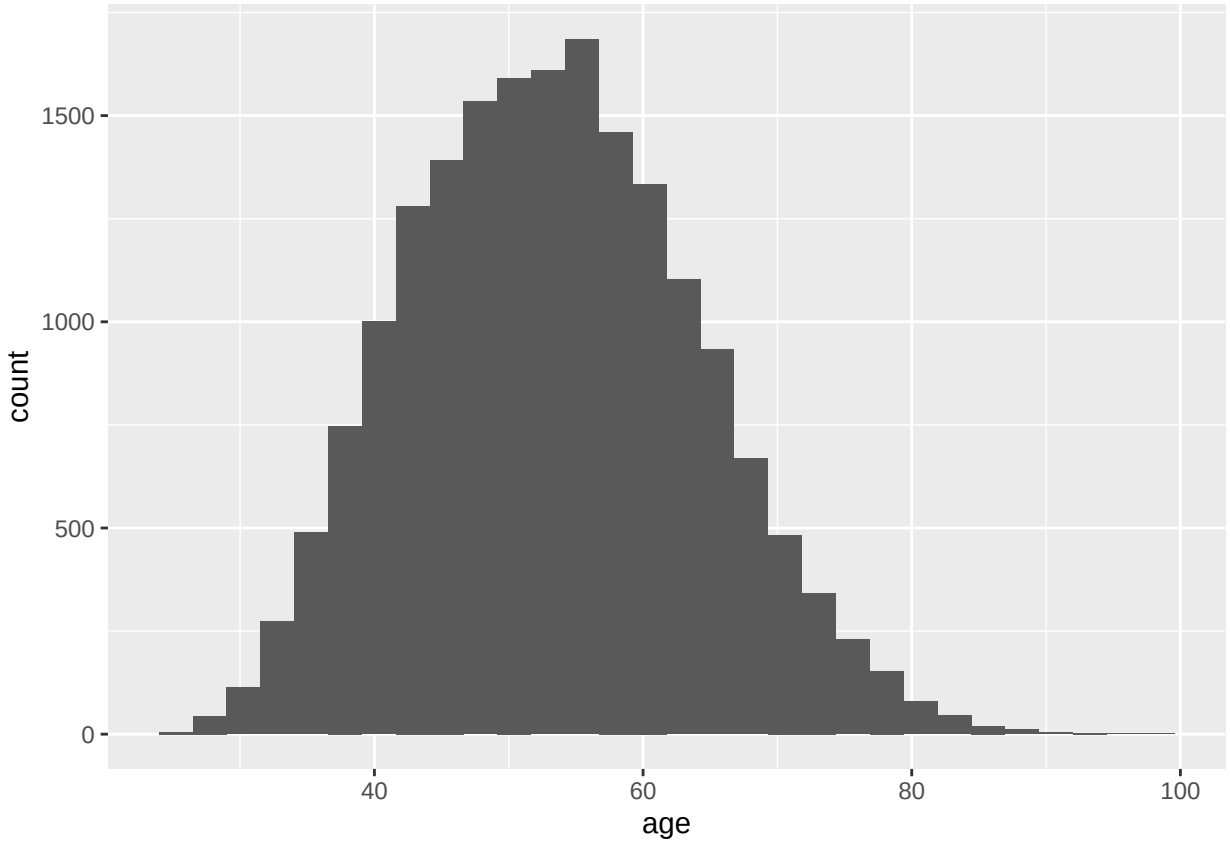


Histogram of Ages

Using `geom_histogram()`, histograms are another way to visualize the distribution of a continuous variable by grouping it into bins and counting the number of observations that fall into each bin. From the histogram, it looks like the most common age of congress members is around 50 years old.

```
congress_age %>%
  ggplot(aes(x= age))+
  geom_histogram()
```

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

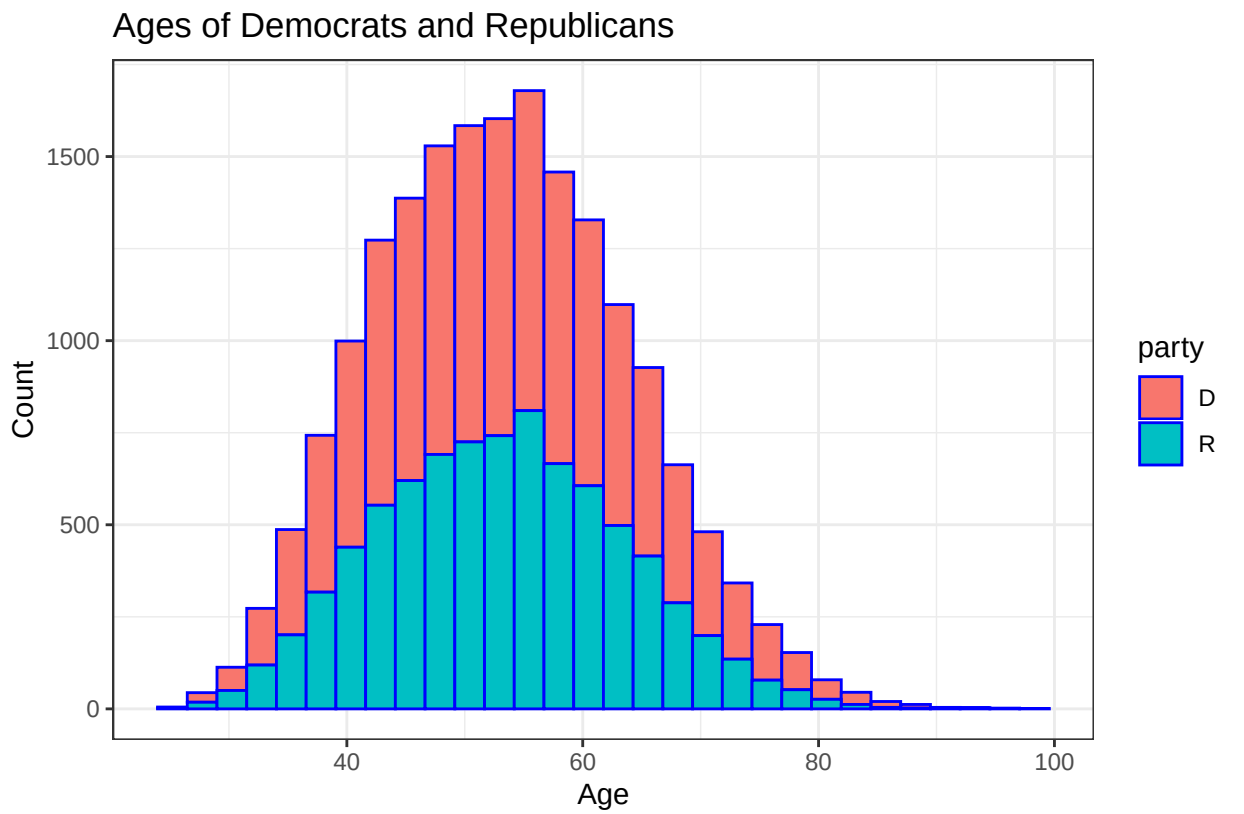


Adding Themes, Labels, and Color to Histogram

The `ggplot2` package has several built-in themes to modify the appearance of the plot. Some of my favorites are `theme_minimal()`, `theme_bw()`, and `theme_light()`. Color can also be added to the plot to enhance its visual appeal and may aid in conveying more information. One way to do this is by adding the `color` argument in aesthetic mapping or in `geom_histogram`. This can be used to help visualize categorical data. In addition, labels allow the reader to easily interpret the data and plot. This is primarily done using the `labs()` function.

```
congress_age %>%
  filter(party == "R" | party == "D") %>%
  ggplot(aes(x = age, fill = party), na.rm = TRUE) +
  geom_histogram(color = "blue") +
  labs(title = "Ages of Democrats and Republicans",x = "Age", y = "Count", caption = "A histogram of th
  theme_bw()
```

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



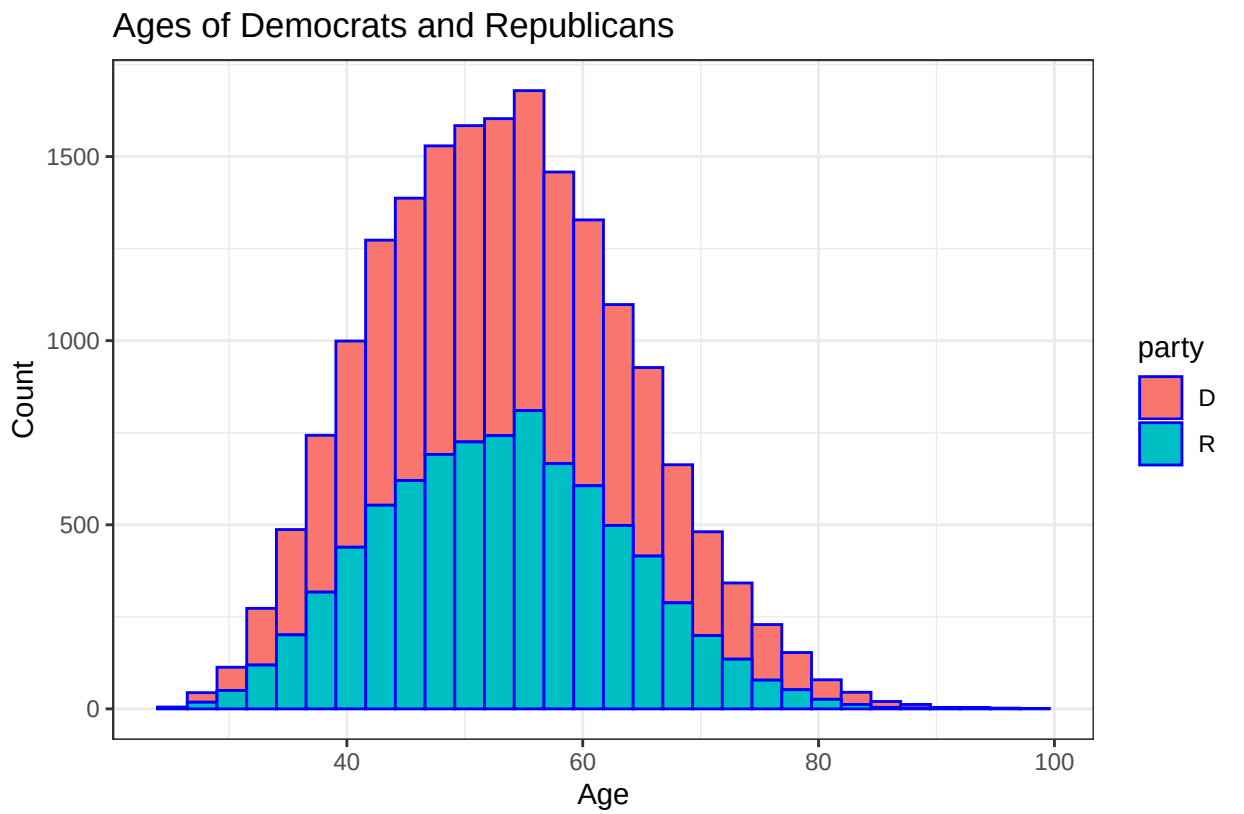
A histogram of the ages of Democrats and Republicans in Congress

Embedding and Saving the Plot The histogram can be saved to an object, which can be embedded into a document. Finally, we can save this plot to our working directory using the `ggsave()` function. We can also adjust the width and height of the plot.

```
age_of_congress <-
  congress_age %>%
    filter(party == "R" | party == "D") %>%
    ggplot(aes(x = age, fill = party), na.rm = TRUE) +
    geom_histogram(color = "blue") +
    labs(title = "Ages of Democrats and Republicans",x = "Age", y = "Count", caption = "A histogram of
    theme_bw()

age_of_congress
```

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



A histogram of the ages of Democrats and Republicans in Congress

```
ggsave("Age of Congress.pdf", age_of_congress, width = 10, height = 5)
```

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

Conclusion

In conclusion, the `ggplot2` package in the tidyverse is an important tool in visualizing data analysis. It has a wide range of functions that allow the user to create and customize plots. In this example, we were able demonstrate many uses of `ggplot2` in a large dataset of congress ages. From the plots, it is clear that the average age of congress is around 50 years old. Interestly enough, the curve of congress ages seems to follow a normal distribution.