

Calculating the Average and SD in R

`group_by()` and `summarize()`

Section 6.4 of the notes


```
# group and summarize data
grouped_df <- group_by(nominate, party, congress)
smry <- summarize(grouped_df,
                  average_ideology = mean(ideology),
                  sd_ideology = sd(ideology))
```

function that applies groups to the data frame

```
# group and summarize data
grouped_df <- group_by(nominate, party, congress)
smry <- summarize(grouped_df,
                   average_ideology = mean(ideology),
                   sd_ideology = sd(ideology))
```


1st argument: data frame to group

```
# group and summarize data  
grouped_df <- group_by(nominate, party, congress)  
smry <- summarize(grouped_df,  
                   average_ideology = mean(ideology),  
                   sd_ideology = sd(ideology))
```




2nd argument: a grouping variable

```
# group and summarize data
grouped_df <- group_by(nominate, party, congress)
smry <- summarize(grouped_df,
  average_ideology = mean(ideology),
  sd_ideology = sd(ideology))
```



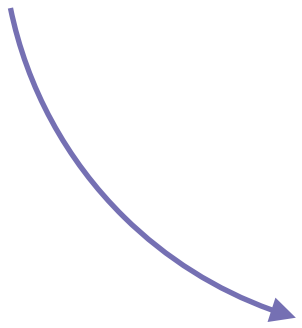
3rd argument: a(nother) grouping variable

```
# group and summarize data
grouped_df <- group_by(nominate, party, congress)
smry <- summarize(grouped_df,
  average_ideology = mean(ideology),
  sd_ideology = sd(ideology))
```



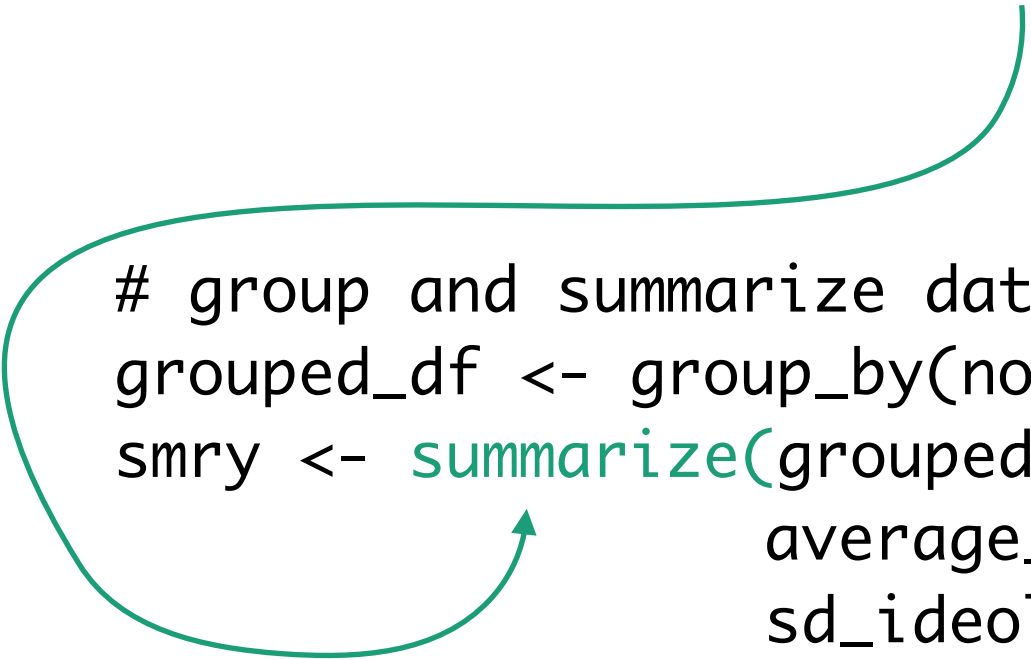
We could add a 3rd and 4th grouping variable if we wanted. Or we could have only one grouping variable.

```
# group and summarize data
grouped_df <- group_by(nominate, party, congress)
smry <- summarize(grouped_df,
                  average_ideology = mean(ideology),
                  sd_ideology = sd(ideology))
```



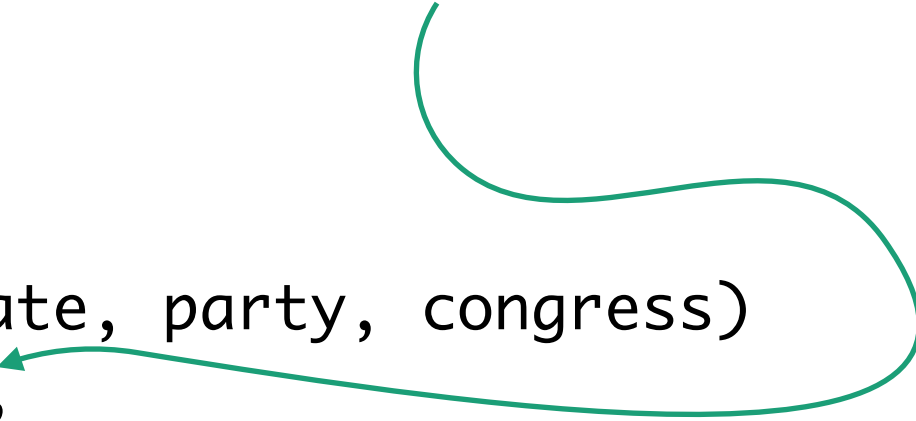
**A function that computes statistics (i.e., “summaries”)
within each group of a grouped data frame.**

```
# group and summarize data
grouped_df <- group_by(nominate, party, congress)
smry <- summarize(grouped_df,
                  average_ideology = mean(ideology),
                  sd_ideology = sd(ideology))
```



1st argument: a grouped data frame

```
# group and summarize data
grouped_df <- group_by(nominate, party, congress)
smry <- summarize(grouped_df,
                  average_ideology = mean(ideology),
                  sd_ideology = sd(ideology))
```



2nd argument: a quantity calculated using a variable in the grouped data frame. It is explicitly named, but you choose the name.

```
# group and summarize data
grouped_df <- group_by(nominate, party, congress)
smry <- summarize(grouped_df,
  average_ideology = mean(ideology),
  sd_ideology = sd(ideology))
```

3rd argument: a(nother) quantity calculated using a variable in the grouped data frame. Again, it is explicitly named, but you choose the name.

```
# group and summarize data
grouped_df <- group_by(nominate, party, congress)
smry <- summarize(grouped_df,
                   average_ideology = mean(ideology),
                   sd_ideology = sd(ideology))
```

```
# group and summarize data
grouped_df <- group_by(nominate, party, congress)
smry <- summarize(grouped_df,
                  average_ideology = mean(ideology),
                  sd_ideology = sd(ideology))
```

Question: If we run this code, what is **smry**?

```
# group and summarize data
grouped_df <- group_by(nominate, party, congress)
smry <- summarize(grouped_df,
                  average_ideology = mean(ideology),
                  sd_ideology = sd(ideology))
```

Question: If we run this code, what is **smry**?

Answer: A data frame.

```
# group and summarize data
grouped_df <- group_by(nominate, party, congress)
smry <- summarize(grouped_df,
                  average_ideology = mean(ideology),
                  sd_ideology = sd(ideology))
```

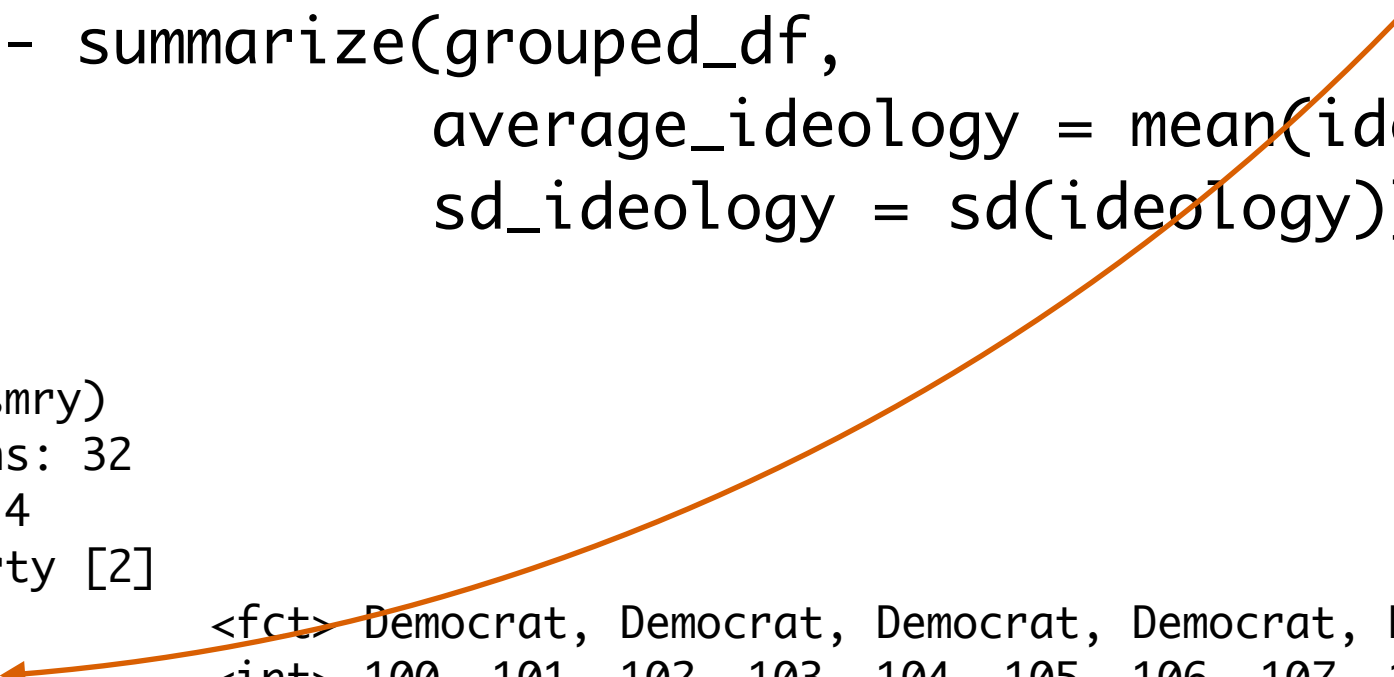
```
> glimpse(smry)
Observations: 32
Variables: 4
Groups: party [2]
$ party          <fct> Democrat, Democrat, Democrat, Democrat, Democrat, Democrat, Democrat,...
$ congress       <int> 100, 101, 102, 103, 104, 105, 106, 107, 108, 109, 110, 111, 112, 113,...
$ average_ideology <dbl> -0.3092901, -0.3130075, -0.3142407, -0.3333065, -0.3615000, -0.375533...
$ sd_ideology     <dbl> 0.1653092, 0.1664293, 0.1658089, 0.1609726, 0.1524251, 0.1377665, 0.1...
```

```
# group and summarize data
grouped_df <- group_by(nominate, party, congress)
smry <- summarize(grouped_df,
                  average_ideology = mean(ideology),
                  sd_ideology = sd(ideology))
```

```
> glimpse(smry)
Observations: 32
Variables: 4
Groups: party [2]
$ party <fct> Democrat, Democrat, Democrat, Democrat, Democrat, Democrat, Democrat,...
$ congress <int> 100, 101, 102, 103, 104, 105, 106, 107, 108, 109, 110, 111, 112, 113,...
$ average_ideology <dbl> -0.3092901, -0.3130075, -0.3142407, -0.3333065, -0.3615000, -0.375533...
$ sd_ideology <dbl> 0.1653092, 0.1664293, 0.1658089, 0.1609726, 0.1524251, 0.1377665, 0.1...
```

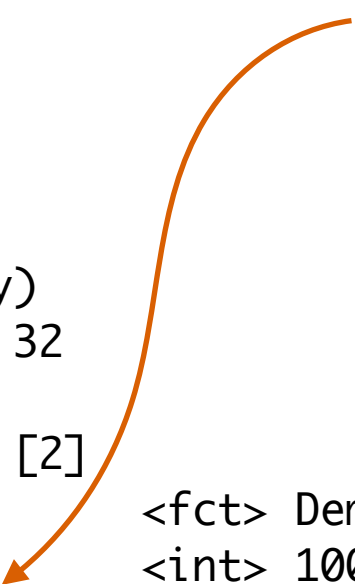
```
# group and summarize data
grouped_df <- group_by(nominate, party, congress)
smry <- summarize(grouped_df,
                  average_ideology = mean(ideology),
                  sd_ideology = sd(ideology))
```

```
> glimpse(smry)
Observations: 32
Variables: 4
Groups: party [2]
$ party      <fct> Democrat, Democrat, Democrat, Democrat, Democrat, Democrat, Democrat,...
$ congress   <int> 100, 101, 102, 103, 104, 105, 106, 107, 108, 109, 110, 111, 112, 113,...
$ average_ideology <dbl> -0.3092901, -0.3130075, -0.3142407, -0.3333065, -0.3615000, -0.375533...
$ sd_ideology <dbl> 0.1653092, 0.1664293, 0.1658089, 0.1609726, 0.1524251, 0.1377665, 0.1...
```



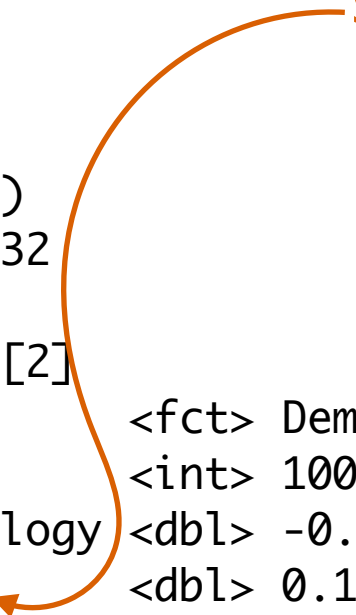

```
# group and summarize data
grouped_df <- group_by(nominate, party, congress)
smry <- summarize(grouped_df,
  average_ideology = mean(ideology),
  sd_ideology = sd(ideology))
```

```
> glimpse(smry)
Observations: 32
Variables: 4
Groups: party [2]
$ party      <fct> Democrat, Democrat, Democrat, Democrat, Democrat, Democrat, Democrat,...
$ congress   <int> 100, 101, 102, 103, 104, 105, 106, 107, 108, 109, 110, 111, 112, 113,...
$ average_ideology <dbl> -0.3092901, -0.3130075, -0.3142407, -0.3333065, -0.3615000, -0.375533...
$ sd_ideology <dbl> 0.1653092, 0.1664293, 0.1658089, 0.1609726, 0.1524251, 0.1377665, 0.1...
```



```
# group and summarize data
grouped_df <- group_by(nominate, party, congress)
smry <- summarize(grouped_df,
                  average_ideology = mean(ideology),
                  sd_ideology = sd(ideology))
```

```
> glimpse(smry)
Observations: 32
Variables: 4
Groups: party [2]
$ party      <fct> Democrat, Democrat, Democrat, Democrat, Democrat, Democrat, Democrat,...
$ congress   <int> 100, 101, 102, 103, 104, 105, 106, 107, 108, 109, 110, 111, 112, 113,...
$ average_ideology <dbl> -0.3092901, -0.3130075, -0.3142407, -0.3333065, -0.3615000, -0.375533...
$ sd_ideology <dbl> 0.1653092, 0.1664293, 0.1658089, 0.1609726, 0.1524251, 0.1377665, 0.1...
```



Key Point

Combining `group_by()` and `summarize()` creates a data frame with the following variables:

- the grouping variables
 - party
 - congress
- the summaries (argument names become variable names)
 - average_ideology
 - sd_ideology

```
# group and summarize data
grouped_df <- group_by(nominate, party, congress)
smry <- summarize(grouped_df,
                  average_ideology = mean(ideology),
                  sd_ideology = sd(ideology))
```

```
> glimpse(smry)
Observations: 32
Variables: 4
Groups: party [2]
$ party      <fct> Democrat, Democrat, Democrat, Democrat, Democrat, Democrat, Democrat,...
$ congress   <int> 100, 101, 102, 103, 104, 105, 106, 107, 108, 109, 110, 111, 112, 113,...
$ average_ideology <dbl> -0.3092901, -0.3130075, -0.3142407, -0.3333065, -0.3615000, -0.375533...
$ sd_ideology  <dbl> 0.1653092, 0.1664293, 0.1658089, 0.1609726, 0.1524251, 0.1377665, 0.1...
```

Most importantly, we can use `ggplot()` with `smry`.

```
# create line plot  
ggplot(smry, aes(x = congress, y = average_ideology, color = party)) +  
  geom_line()
```

