Data Vusualization in R with ggplot2

ryannthegeek

6/17/22

Table of contents

List of Figures

List of Tables

1 Getting started with ggplot

1.1 Data and Data Description

1.1.1 cel data

Table 1: cel data variable names and there description

Variable name	Description
thomas_name	Name of the member
congress	number of the congress (there is a new congress every two years)
year	year of the start of the congress
st_name	State abbreviation for the member's district
cd	congressional district number
dem	0/1 indicator for whether the member is a democrat
elected	year the member was elected
female	0/1 indicator for whether the member is female
votepct	the percent of the vote the MC won in the election for this congress
dwnom1	DW-Nominate score indicative member ideology. Higher is more conservative
deleg_size	How many MCs are in the member's state delegation?
speaker	Is the member the Speaker of the House? 0/1
subchr	Is the member the chair of a congressional subcommittee?
afam	Is the member African American? 0/1
latino	Is the member latino?
power	Is the member on a "powerful" committee in Congress?
chair	Is the member a chair of a full committee?
state_leg	Was the member a state legislator prior to being elected to congress?
state_leg_prof	How professionalized is the state legislature in the member's state? Higher is more professional
majority	Is the member in the majority in this congress? 0/1
maj_leader	Is the member a majority leader in this congress? 0/1
min_leader	Is the member a minority leader in this congress? 0/1
meddist	How far away is the member from the chamber median dwnom1 score?
meddist	How far away is the member from the majority median dwnom1 score?
all_bills	How many bills did the member introduce in this congress?
all_aic	How many bills did the member introduce that get action in a committee in this congress?
all_abc	How many bills did the member introduce that get action beyond the committee state in this congress?

Variable name	Description
all_pass	How many bills did the member introduce that passed out of the House in this congress?
all_law	How many bills did the member introduced that became law in this congress?
les	Volden and Wiseman's legislative effective score (LES). Higher means the member is more effective.
seniority	How many term has the member been in congress, including the current term

1.1.2 cces data

Table 2: ccess data variable names and there description

Variable Name	Description
id_number	respondent ID number
region	In which census region do you live?

```
1 Northwest

2 Midwest

3 South

4 West
```

gender	1 Male or 2 Female
educ	What is the highest level of education you have completed?

```
1 No high school

2 High school graduate

3 Some college

4 2-year

5 4-year

6 Post-grad
```

Variable Name	Description
edloan	Are you currently responisble for paying off a student loan?

```
1 Yes

2 No
```

race	What racial or ethnic group best describes you?

```
1 White

2 Black

3 Hispanic

4 Asian

5 Native American

6 Mixed

7 Other

8 Middle Eastern
```

hispanic	Are you of Spanish, Latino, or Hispanic origin or descent?

```
1 Yes

2 No
```

employ	Which of the following best describes your current employment status?

```
1 Full-time

2 Part-time

3 Temporarily laid off

4 Unemployed

5 Retired

6 Permanently disabled

7 Homemaker

8 Student

9 Other
```

Variable Name	Description
marstat	What is you marital status?

```
1 Married

2 Separated

3 Divorced

4 Widowed

5 Never married

6 Domestic/civil partnership
```

pid7	Generally speaking, do you think of yourself as a…?

```
1 Strong Democrat

2 Not very strong Democrat

3 Lean Democrat

4 Independent

5 Lean Republican

6 Not very strong Republican

7 Strong Republican
```

ideo5	In general, how would you describe your own political viewpoint?

```
1 Very liberal

2 Liberal

3 Moderate

4 Conservative

5 Very conservative
```

pew_religimp	How important is religion in yuor life?

```
1 Very important

2 Somewhat important

3 Not too important

4 Not at all important
```

Variable Name	Description
newsint	Some people seem to follow what's going on in government and public affairs most of the time, whether there's an election going on or not. Others aren't that interested. Would you say you follow what's going on in government and public affairs… 1 Most of the time 2 Some of the time 3 Only now and then 4 Hardly at all
faminc_new	Thinking back over the last year, what was your family's annual income? 1 Less than $10,000 2 $10,000 - $19,999 3 $20,000 - $29,999 4 $30,000 - $39,999 5 $40,000 - $49,999 6 $50,000 - $59,999 7 $60,000 - $69,999 8 $70,000 - $79,999 9 $80,000 - $99,999 10 $100,000 - $119,999 11 $120,000 - $149,999 12 $150,000 - $199,999 13 $200,000 - $249,999 14 $250,000 - $349,999 15 $350,000 - $499,999 16 $500,000 or more
union	Are you a member of a labor union? 1 Yes, I am currently a member of a labor union 2 I formerly was a member of a labor union 3 I am not now, nor have I been, a member of a labor union
investor	Do you personally (or jointly with a spouse), have any money invested in the stock market right now, either in an individual stock or in a mutual fund? 1 Yes 2 No

Variable Name	Description
CC18_308a	Job approval – President Trump Do you approve or disapprove of the way each is doing their job… 1 Strongly approve 2 Somewhat approve 3 Somewhat disapprove 4 Strongly disapprove
CC18_312a	Party Recall + Name Recognition - Governor Please indicate whether you've heard of this person and if so which party he or she is affiliated with… 1 Never heard of person 2 Republican 3 Democrat 4 Other Party / Independent 5 Not sure
CC18_312b	Party Recall + Name Recognition - Senator 1 Please indicate whether you've heard of this person and if so which party he or she is affiliated with… 1 Never heard of person 2 Republican 3 Democrat 4 Other Party / Independent 5 Not sure
CC18_312c	Party Recall + Name Recognition - Senator 2 Please indicate whether you've heard of this person and if so which party he or she is affiliated with… 1 Never heard of person 2 Republican 3 Democrat 4 Other Party / Independent 5 Not sure

Variable Name	Description
CC18_318d	Party Recall + Name Recognition - Representative Please indicate whether you've heard of this person and if so which party he or she is affiliated with… 1 Never heard of person 2 Republican 3 Democrat 4 Other Party / Independent 5 Not sure
CC18_325a	Taxes – Cut the Corporate Income Tax rate from 39 percent to 21 percent. Congress considered many changes in tax law over the past two years. Do you support or oppose each of the following? 1 Support 2 Oppose
CC18_325b	Taxes – Reduce the mortgage interest deduction. Allow people to deduct the interest on no more than $500,000 of mortgage debt. The previous limit was $1 million. Congress considered many changes in tax law over the past two years. Do you support or oppose each of the following? 1 Support 2 Oppose
CC18_325c	Taxes – Limit the amount of state and local taxes that can be deducted to $10,000 (previously there was no limit). Congress considered many changes in tax law over the past two years. Do you support or oppose each of the following? 1 Support 2 Oppose
CC18_325d	Taxes – Increase the standard deduction on federal income taxes from $12,000 to to $25,000 Congress considered many changes in tax law over the past two years. Do you support or oppose each of the following? 1 Support 2 Oppose

1.2 Getting started with ggplot part 1

```
require(tidyverse)
cel <- read_csv("cel_data.csv")
names(cel)
```

```
 [1] "...1"           "thomas_num"     "thomas_name"    "icpsr"
 [5] "congress"       "year"           "st_name"        "cd"
 [9] "dem"            "elected"        "female"         "votepct"
[13] "dwnom1"         "deleg_size"     "speaker"        "subchr"
[17] "afam"           "latino"         "votepct_sq"     "power"
[21] "chair"          "state_leg"      "state_leg_prof" "majority"
[25] "maj_leader"     "min_leader"     "meddist"        "majdist"
[29] "all_bills"      "all_aic"        "all_abc"        "all_pass"
[33] "all_law"        "les"            "seniority"      "benchmark"
[37] "expectation"    "TotalInParty"   "RankInParty"
```

```
dim(cel)
```

```
[1] 10262    39
```

```
table(cel$year)
```

```

1973 1975 1977 1979 1981 1983 1985 1987 1989 1991 1993 1995 1997 1999 2001 2003
 444  444  443  442  447  444  445  446  449  447  446  445  449  442  447  444
2005 2007 2009 2011 2013 2015 2017
 445  452  451  449  450  443  448
```

1.2.1 Making a scatter plot

```
#####filter the data we want
fig115 <- cel %>%
  filter(congress == 115) %>%
  select(seniority, all_pass)
head(fig115)
```

```
# A tibble: 6 x 2
  seniority all_pass
      <dbl>    <dbl>
1         2        1
2         3        2
3        11        0
4         2        3
5         2        1
6         4        1
```

```
ggplot(data = fig115, aes(x = seniority, y = all_pass)) +
  geom_point()
```

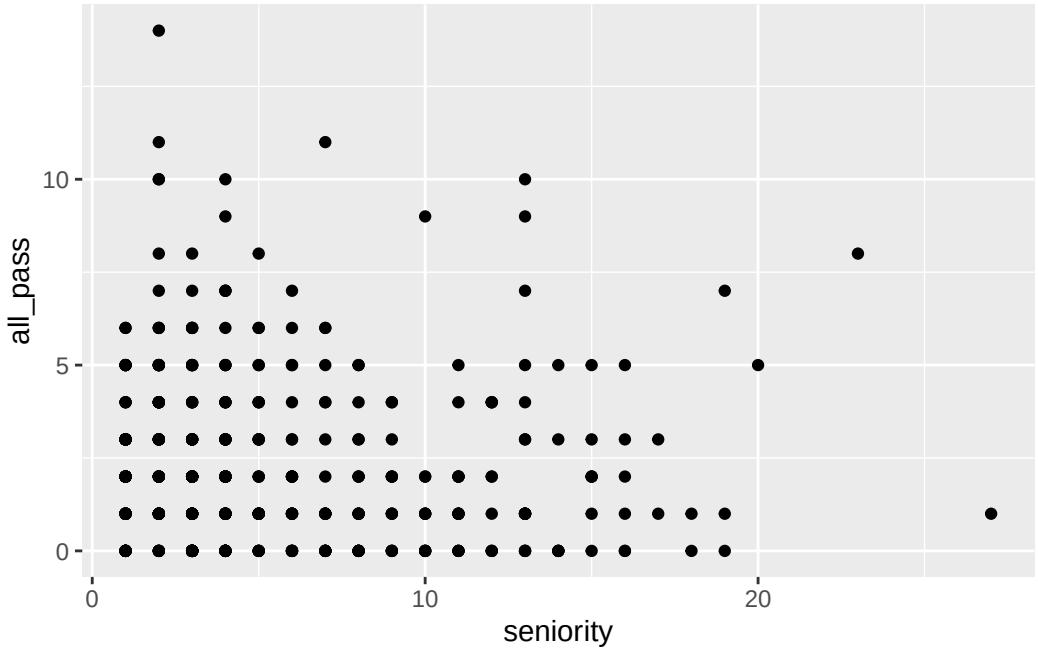

Figure 1: A plain plot of all_pass against seniority

The points here are members of Congress, so their individual rows of the data table, one row per member of the 115th Congress. And the position of the points on the plot comes from how senior member is on the x axis, that is how long they've been in Congress. And on the y axis, how many bills he or she passed out of Congress in the 115th session of Congress. Now, one problem with this chart is that there's an issue with over plotting, which means that there are cases in the data where there's more than one member that has exactly the same x and y coordinate value. So exactly the same seniority and then the same number of bills passed. For instance, if I had to members of Congress and they both had a seniority value of 6, and they passed 3 bills, then those two points would be right on top of each other in the figure. To help represent this and not have these points over plotted on top of each other, we're going to change the command slightly. So rather than use the geom point geom function, we'll use the geom jitter function instead.

1.2.2 jitter adds random noise to the data to avoid over plotting

When you use geom jitter, what you're doing is you're adding a little bit of random noise to the x and y values of the data, so the plots won't be exactly on top of each other

```
ggplot(fig115 ,aes(x = seniority, y = all_pass)) +
  geom_jitter() +
  labs(x = "Seniority", y = "Bills Passed",
       title = "Seniority and Bills Passed in the 115th Congress")
```

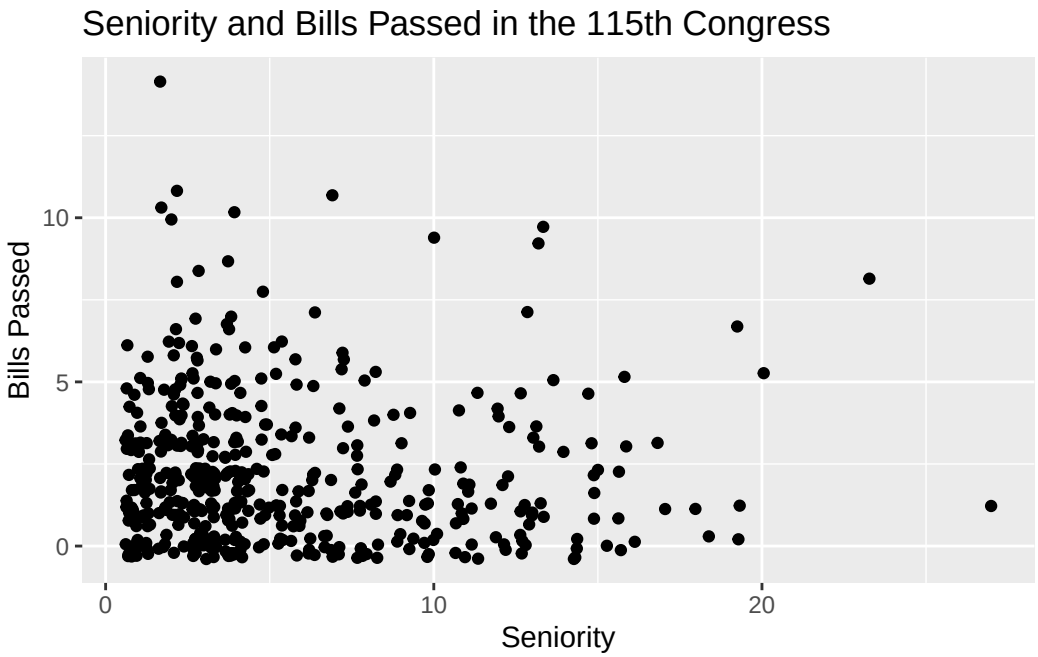

Figure 2: A jitter plot of all_pass against seniority

1.2.3 To plot with colors we modify the data first

Modify filter and select to grab "dem"
The dim variable is just a series of 1s and 0s which indicate whether or not each member is a Democrat or Republican.

```
fig115 <- cel %>%
  filter(congress == 115) %>%
  select(seniority, all_pass,  dem)

ggplot(fig115, aes(x = seniority, y = all_pass, color = dem)) +
  geom_jitter() +
  labs(x = "Seniority", y = "Bills Passed",
       title = "Seniority and Bills Passed in the 115th Congress")
```

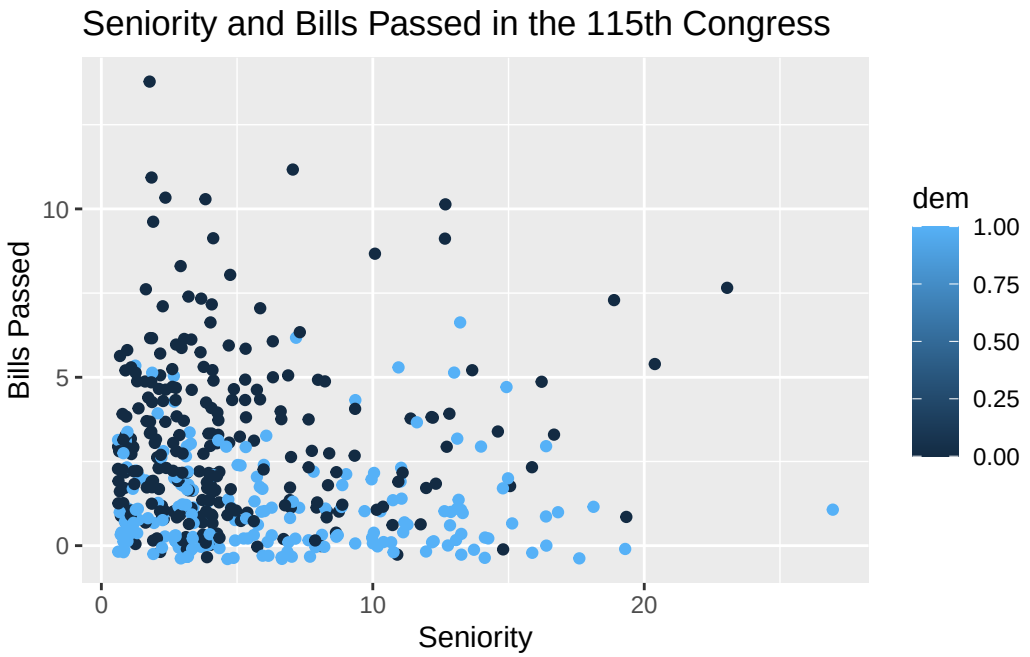

Figure 3: A color plot of all_pass against seniority

1.2.4 Colors are strange, let's fix

```
## make dem a categorical variable called "party"
party <- recode(fig115$dem, `1` = 'Democrat', `0` = 'Republican')
fig115 <- add_column(fig115, party)

ggplot(fig115, aes(x = seniority, y = all_pass, color = party)) +
  geom_jitter() +
  labs(x = "Seniority", y = "Bills Passed",
       title = "Seniority and Bills Passed in the 115th Congress")
```

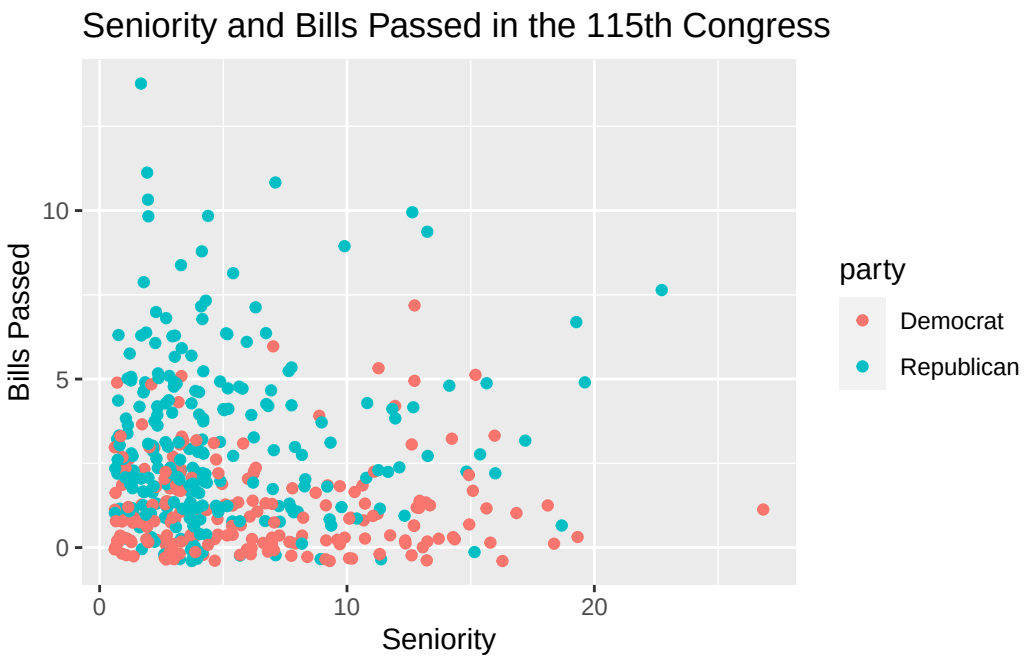

Figure 4: A modified color plot of all_pass against seniority

2 More visualizations with ggplot

2.1 Bar plots

2.1.1 Basic Bar plots

Let's say that we want a bar chart that shows counts of how many Republicans and Democrats there were in the 115th Congress. There are 435 members of Congress, and this bar plot would show you the number of Democrats and the number of Republicans as two separate bars.

```
cel %>%
  filter(congress == 115) %>%
  ggplot(aes(x = dem)) +
  geom_bar()
```

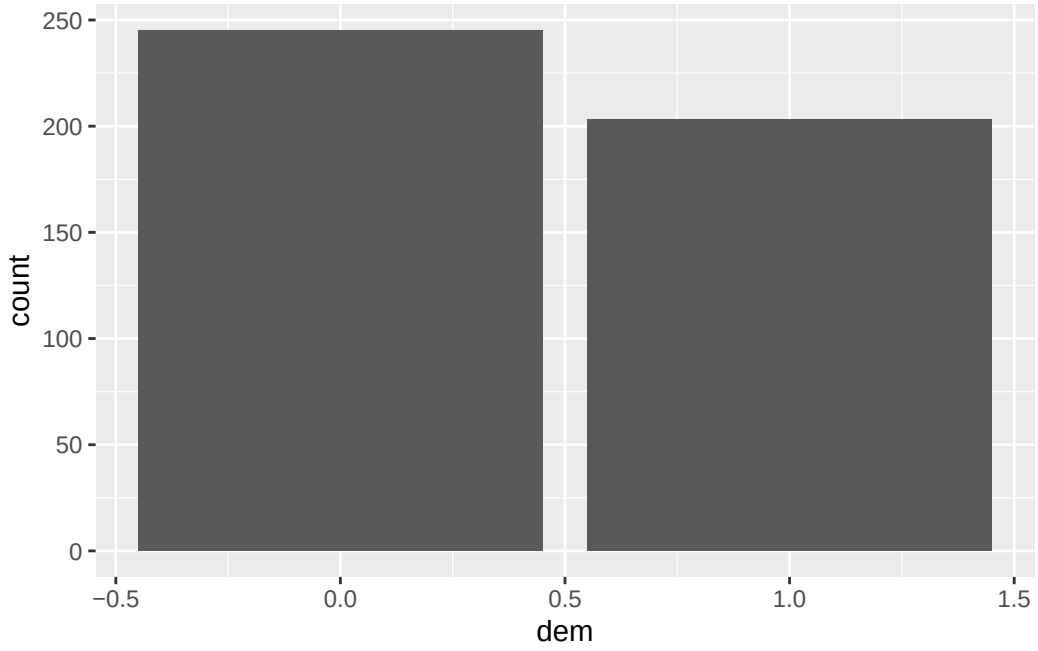

Figure 5: Basic bar plot for dem variable in the 115th Congress. 0 = Republican, 1 = Democrat

Let's prove that the bar plot is right by comparing with a frequency table:

```
table(filter(cel, congress == 115)$dem)
```

```

  0   1
245 203
```

use `st_name` instead, show how counts of how many members of Congress from each state:

```
cel %>%
  filter(congress == 115) %>%
  ggplot(aes(x = st_name)) +
  geom_bar()
```

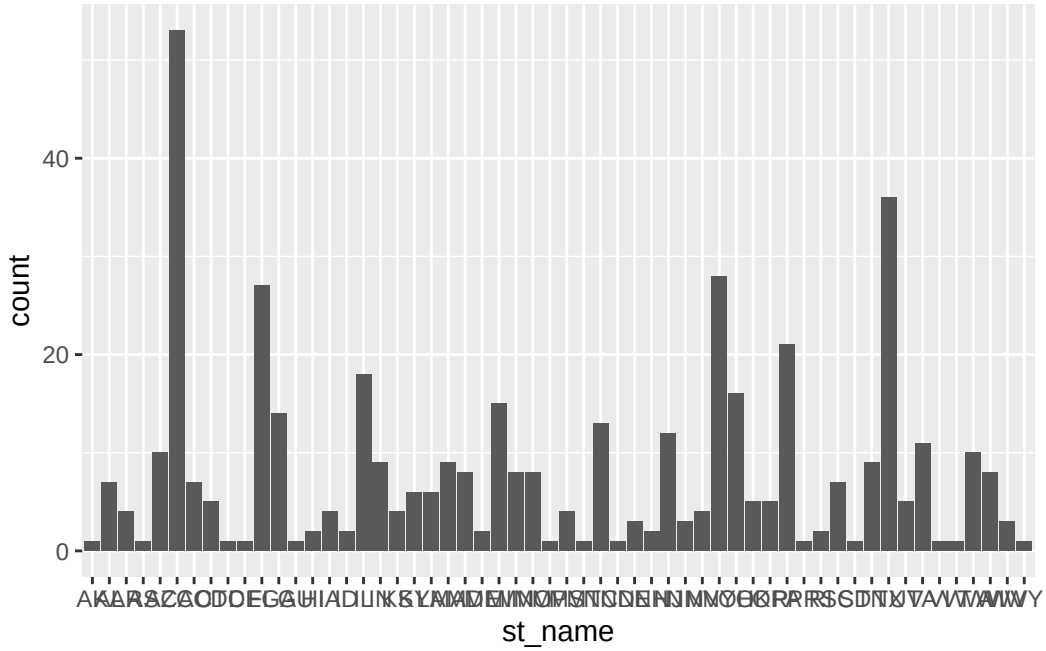

Figure 6: Counts of how many members of Congress from each state

Using party variable created previously to make some bar plots add axis labels

```
cel %>%
  filter(congress == 115)%>%
  ggplot(aes(x = party, fill = party)) +
  geom_bar() +
  labs(x = "Party",y = "Number of Members")
```

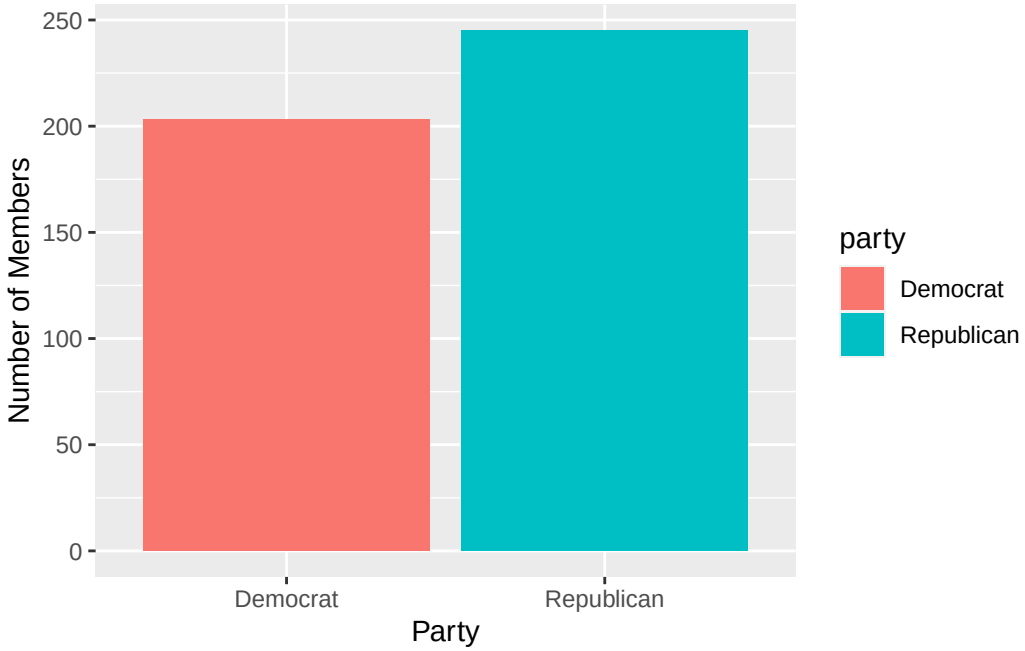

Figure 7: Bar plot using party variable

```
cel %>%
  filter(congress == 115) %>%
  ggplot(aes(x = party, fill = party)) +
  geom_bar() +
  labs(x = "Party",y = "Number of Members") +
  scale_fill_manual(values = c("blue", "red")) ## note the use of fill on the
    two instances
```

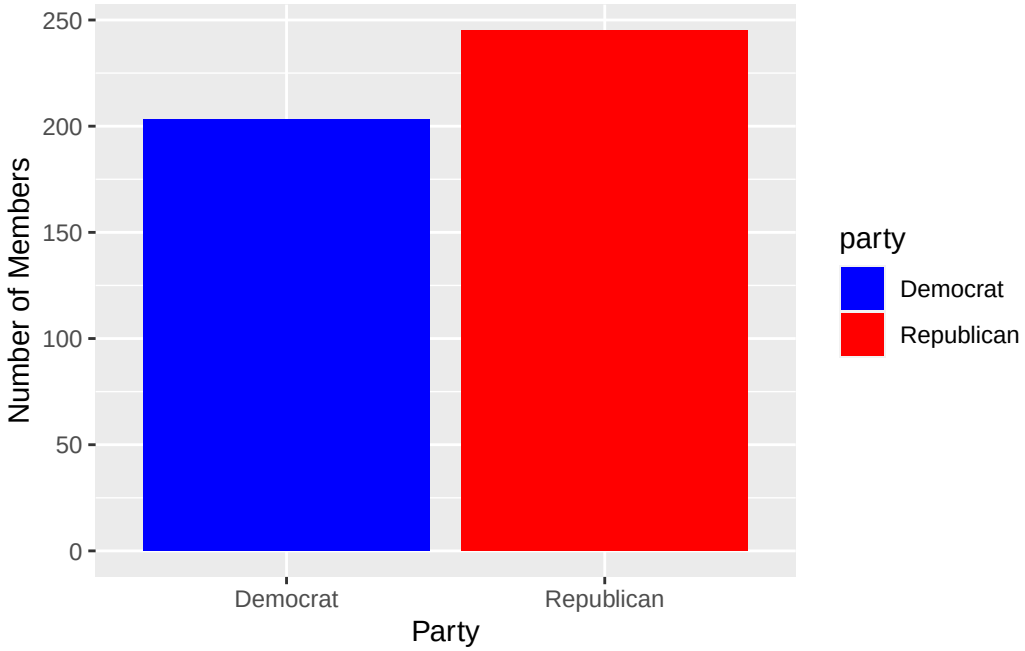

Figure 8: manually change the colors of the bars

```
cel %>%
  filter(congress == 115) %>%
  ggplot(aes(x = party, fill = party)) +
  geom_bar() +
  labs(x = "Party",y = "Number of Members") +
  scale_fill_manual(values = c("blue", "red")) +
  guides(fill = "none")
```

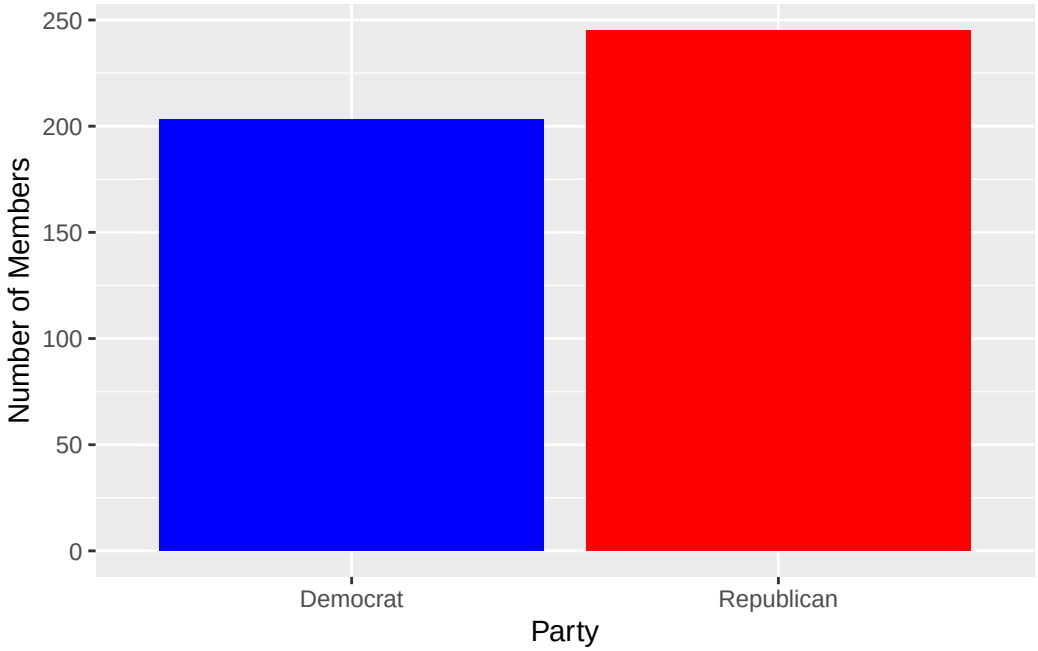

Figure 9: drop the legend with the "guides" command

2.1.2 Making more barplots and manipulating more data

Making a barplot of proportions a toy demonstration

```
#####a bowl of fruit
apple <- rep("apple",6)
orange <- rep("orange",3)
banana <- rep("banana",1)
###put together the fruits in a dataframe
###create a single columns with fruits
fruit_bowl <- tibble("fruits"=c(apple,orange,banana));fruit_bowl
```

```
# A tibble: 10 x 1
   fruits
   <chr>
 1 apple
 2 apple
 3 apple
 4 apple
 5 apple
 6 apple
 7 orange
 8 orange
 9 orange
10 banana
```

Let's calculate proportions instead

```
#####create a table that counts fruits in a second column
fruit_bowl_summary <- fruit_bowl %>%
  group_by(fruits) %>%
  summarize("count" = n())
fruit_bowl_summary
```

```
# A tibble: 3 x 2
  fruits count
  <chr>  <int>
1 apple      6
2 banana     1
3 orange     3
```

```
####calculate proportions
fruit_bowl_summary$proportion <- fruit_bowl_summary$count/sum(fruit_bowl_summary
    $count)
fruit_bowl_summary
```

```
# A tibble: 3 x 3
  fruits count proportion
  <chr>  <int>      <dbl>
1 apple      6        0.6
2 banana     1        0.1
3 orange     3        0.3
```

```
ggplot(fruit_bowl_summary, aes(x = fruits, y = proportion, fill = fruits)) +
  geom_bar(stat = "identity") +
  scale_fill_manual(values = c("red", "yellow", "orange")) +
  guides(fill = "none") +
  labs(x = "Fruits",y = "Proportion of Fruits")
```

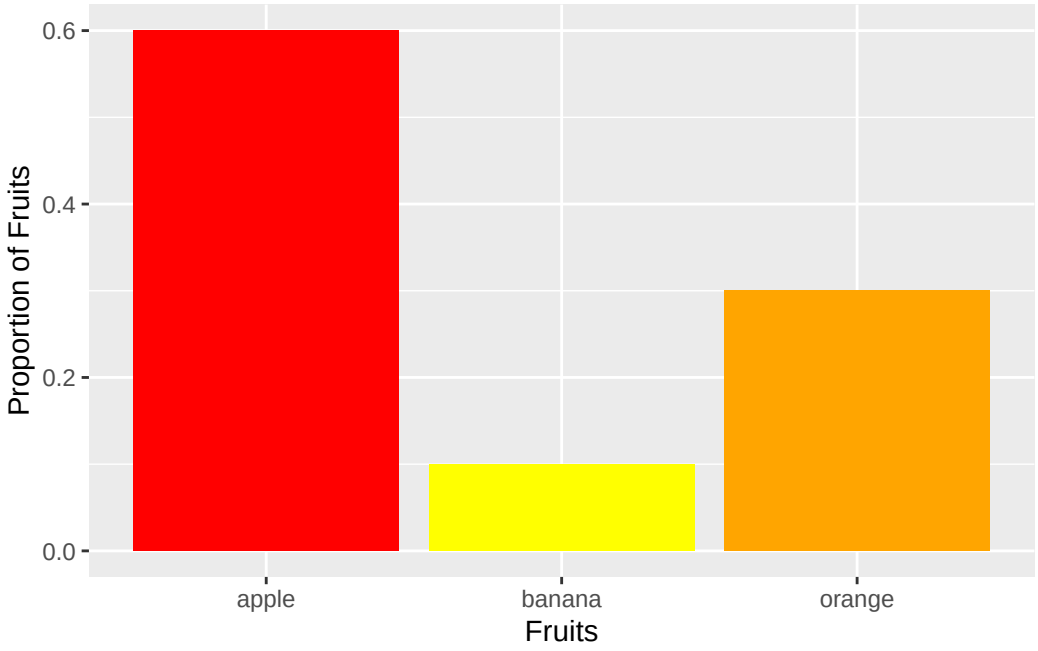

Figure 10: Bar plot using "stat" to tell command to plot the exact value for proportion

The stat function is telling the function not to do its default action, which is to use that count function, but instead plot the different values for every single row of the dataset directly based on the identity of that value in the table. That's why I mapped both the x and the y, for every row map the x-value and the y-value as the top of the bar in that bar plot.

2.1.3 More practice with bar plots

```
require(tidyverse)
cces <- read_csv("cces_data.csv")
names(cces)
```

```
 [1] "...1"         "caseid"       "region"       "gender"       "educ"
 [6] "edloan"       "race"         "hispanic"     "employ"       "marstat"
[11] "pid7"         "ideo5"        "pew_religimp" "newsint"      "faminc_new"
[16] "union"        "investor"     "CC18_308a"    "CC18_310a"    "CC18_310b"
[21] "CC18_310c"    "CC18_310d"    "CC18_325a"    "CC18_325b"    "CC18_325c"
[26] "CC18_325d"
```

```
dim(cces)
```

```
[1] 1000   26
```

This survey question is a 7-point scale, where 1-3 means that the survey respondent is a strong Democrat to lean Democrat, 4 means that the respondent is a true independent, and 5-7 means that the respondent is a lean Republican to a strong Republican. We're going to recode this Data and collapse it so it's just a three-part variable for Republican, Independent, or Democrat.

```
## create counts of Democrats, Republicans, and Independent by region
dem_rep <- recode(cces$pid7,
                  `1` = "Democrat",
                  `2` = "Democrat",
                  `3` = "Democrat",
                  `4` = "Independent",
                  `5` = "Republican",
                  `6` = "Republican",
                  `7` = "Republican")

table(dem_rep)
```

```
dem_rep
   Democrat Independent  Republican
        516         119         365
```

```
cces <- add_column(cces, dem_rep)
```

```
ggplot(cces, aes(x = region, fill = dem_rep)) +
  geom_bar(position = "dodge") +
  labs(x = "Region", y = "Count")
```

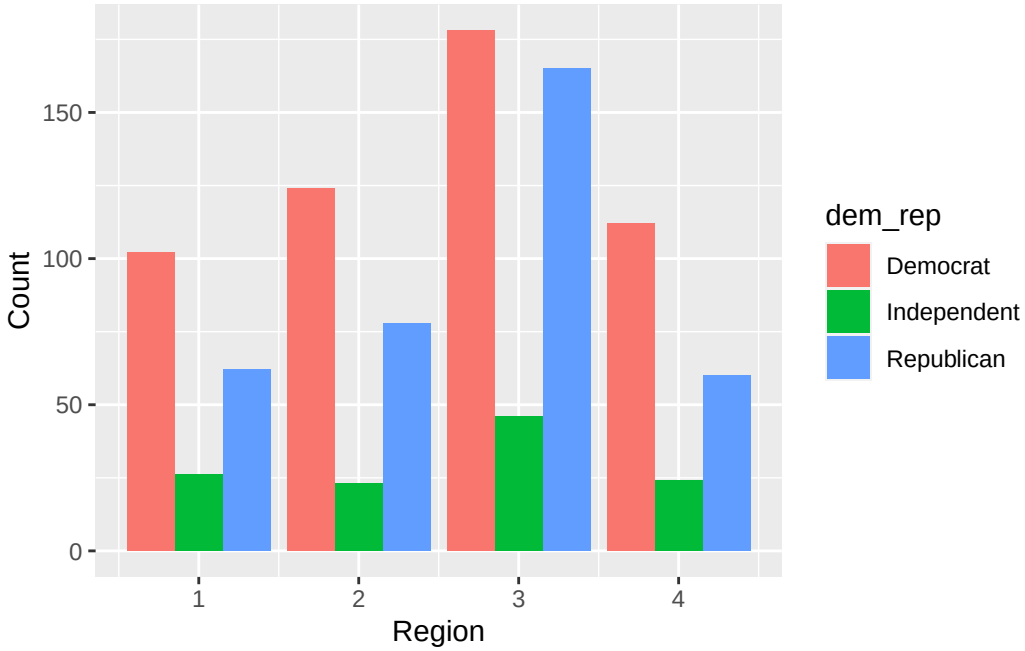

Figure 11: A plot of grouped bars

```
## The dodge function takes the stacked bar chart and breaks it apart. So you
    have 12 bars, that are divided into four groups.
```

2.2 Line plots and time series

```
require(tidyverse)
####create a sequence of years

years <- seq(from = 2001, to = 2020, by = 1)

####create "fake" data for price (note, your values will be different)
set.seed(12345)
price <- rnorm(20 , mean = 15, sd = 5)

####put years and price together

fig_data <- tibble("year" = years, "stock_price" = price)
head(fig_data)
```

```
# A tibble: 6 x 2
   year stock_price
  <dbl>       <dbl>
1  2001        17.9
2  2002        18.5
```

```
3   2003        14.5
4   2004        12.7
5   2005        18.0
6   2006         5.91
```

```
ggplot(data = fig_data, aes(x = years, y = price)) +
  geom_line()
```

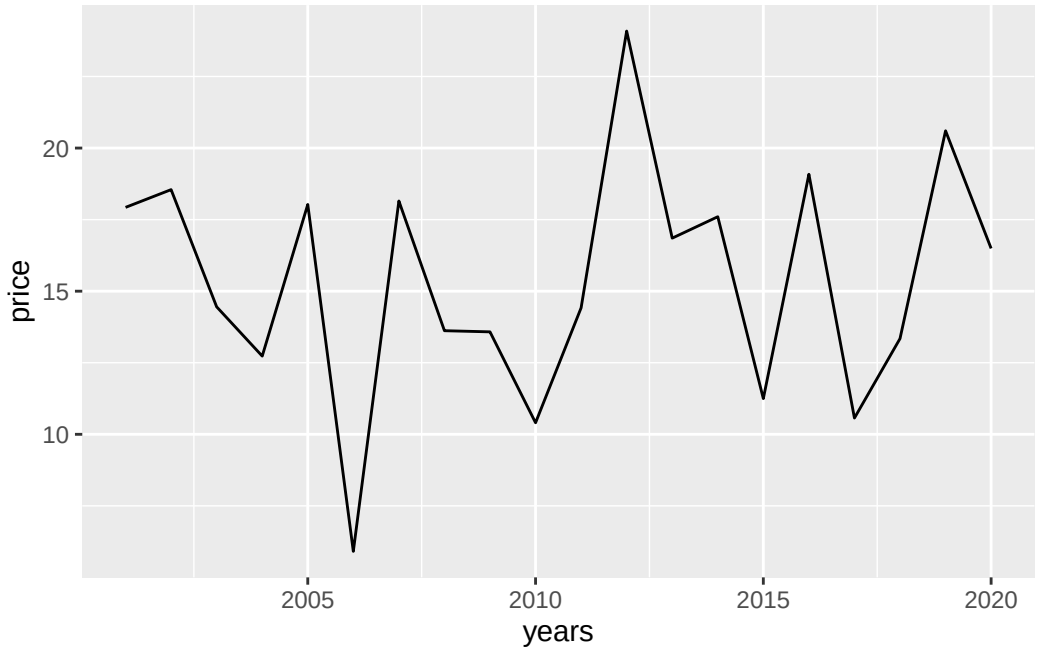

Figure 12: A time series plot of one stock

```
## make data for the first of two stocks

fig_data$stock_id <- rep("Stock_1", 20)

stock_1_time_series <- fig_data

## create data for the second company
#### same approach as with the last company
stock_id <- rep("Stock_2", 20)

years <- seq(from = 2001, to = 2020, by = 1)

price <- rnorm(20, mean = 10, sd = 3)

stock_2_time_series <- tibble("stock_id" = stock_id, "year" = years,
                              "stock_price" = price)

```

```r
## combine with bind_rows()

all_stocks_time_series <- bind_rows(stock_1_time_series,stock_2_time_series)

## make the plot, setting group to stock_id
ggplot(data = all_stocks_time_series,
       aes(x = year, y = stock_price, group = stock_id)) +
  geom_line()
```

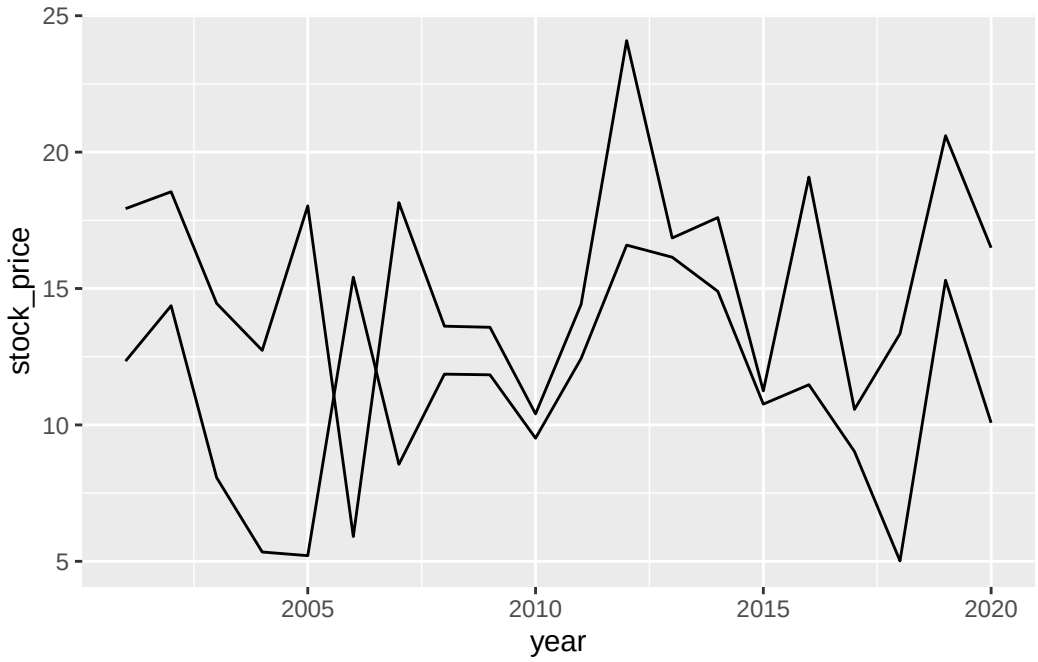

Figure 13: A time series plot of two stocks

Modify group, linetype, color, and add `facet_wrap()`

```r
ggplot(all_stocks_time_series, (aes(x = year,y = stock_price, group = stock_id,
                                    linetype = stock_id, color = stock_id))) +
  geom_line() +
  facet_wrap(~stock_id)
```

Figure 14: Using facet_wrap() to plot the two stocks

2.2.1 Practice with cel data set

```
attach(cel)
cel$Party <- recode(cel$dem,
                    `1` = "Democrat",
                    `0` = "Republican")
fig_data <- cel %>%
  group_by(Party, year) %>%
  summarize("Ideology" = mean(dwnom1, na.rm = T))
require(knitr)
kable(head(fig_data))
```

Party	year	Ideology
Democrat	1973	-0.2953566
Democrat	1975	-0.2954948
Democrat	1977	-0.2869758
Democrat	1979	-0.2855018
Democrat	1981	-0.2889877
Democrat	1983	-0.2940443

A plot of Ideology against year of republican and democrat

```r
ggplot(data = fig_data, aes(x = year, y = Ideology, group = Party, color = Party
    )) +
  geom_line() +
  scale_color_manual(values = c("blue", "red"))
```

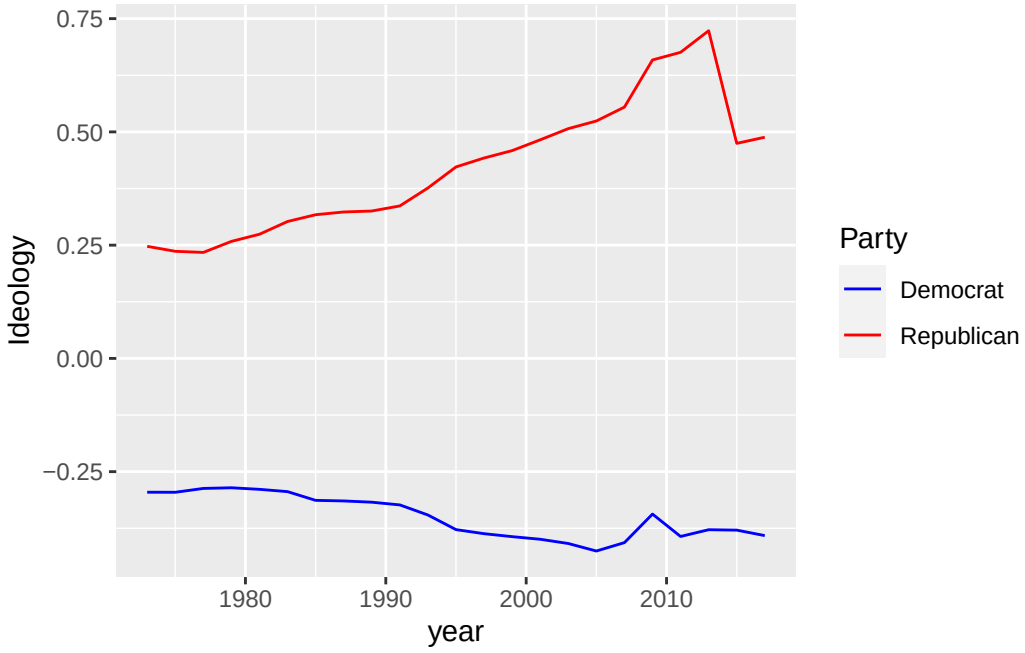

Figure 15: A plot of Ideology against year of republican and democrat

2.3 Learning new visualizations heatmaps

2.3.1 Dummy data

```r
x <- LETTERS[1:20] # assigns letters of alphabet
y <- paste0("var", seq(1, 20)) # takes the string var and pastes it to a
    sequence of 1:20

dat <- expand.grid(X = x, Y = y) # Creates a data frame from all combinations of
     the supplied vectors or factors.
require(knitr)
kable(head(dat))
```

X	Y
A	var1
B	var1

X	Y
C	var1
D	var1
E	var1
F	var1

```r
set.seed(1234)
dat$Z <- runif(400, 0, 5) # Uniform distribution pulling random numbers from the
     interval [0, 5]
```

2.3.2 Heatmap

```r
ggplot(data = dat, aes(x = X, y = Y, fill = Z)) +
  geom_tile()
```

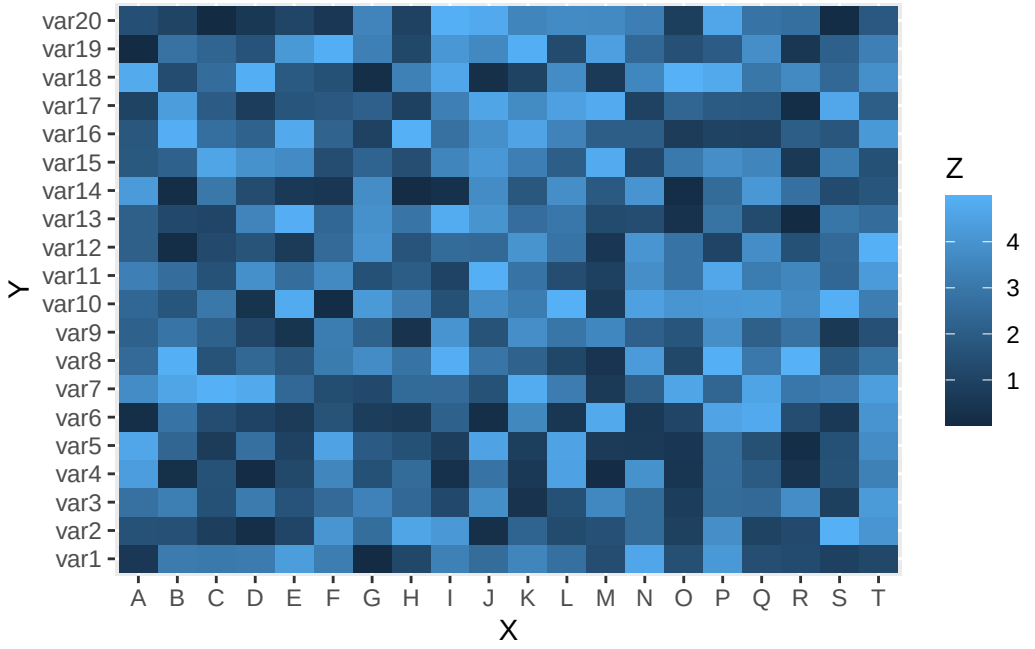

Figure 16: A plot of heatmap

2.3.3 Practice again using a more substantive example

```r
players <- c("Michael","LeBron","Kobe")
points <- c(35, 40,45)
assists <- c(10,12,5)
rebounds <- c(15,12,5)

basketball <- tibble(players, points, assists, rebounds)
kable(head(basketball))
```

players	points	assists	rebounds
Michael	35	10	15
LeBron	40	12	12
Kobe	45	5	5

standardize the values:

```r
## divide each value by the maximum in the datset so that the maximum value is
    one
basketball$standardized_points <- basketball$points/max(basketball$points)
basketball$standardized_assists <- basketball$assists/max(basketball$assists)
basketball$standardized_rebounds <- basketball$rebounds/max(basketball$rebounds)

basketball_stanardize <- select(basketball,
                                "players",
                                "standardized_points",
                                "standardized_assists",
                                "standardized_rebounds")

long_basketball_scaled <- pivot_longer(basketball_stanardize,
                                       c("standardized_points",
                                         "standardized_assists",
                                         "standardized_rebounds"),
                                       names_to = "stat", values_to = "value")
kable(head(long_basketball_scaled))
```

players	stat	value
Michael	standardized_points	0.7777778
Michael	standardized_assists	0.8333333
Michael	standardized_rebounds	1.0000000
LeBron	standardized_points	0.8888889
LeBron	standardized_assists	1.0000000
LeBron	standardized_rebounds	0.8000000

Another plot of heatmap using the basketball example

```r
ggplot(long_basketball_scaled, aes(x = players, y = stat, fill = value)) +
  geom_tile()
```

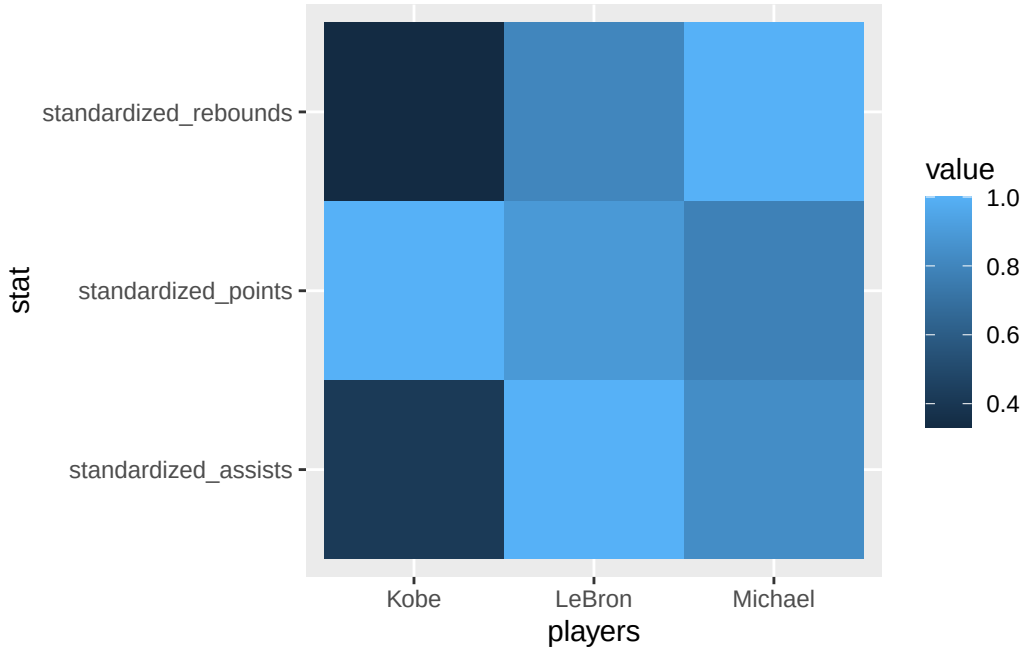

Figure 17: Another plot of heatmap using the basketball example

3 ggplot graphical elements

3.1 An example simulation

Make a scatter plot of lifetime high scores in tetris against cumulative hours playing tetris likely to see positive correlation

```
kid <- c("Nick","Jessica","Justin","Brandi","Kelly","Enrique")
time_spent <- c(40,35,25,20,10,5)
high_score <- c(100,75,85,50,25,30)

tetris <- tibble(kid,time_spent,high_score)
kable(tetris)
```

kid	time_spent	high_score
Nick	40	100
Jessica	35	75
Justin	25	85
Brandi	20	50
Kelly	10	25

kid	time_spent	high_score
Enrique	5	30

Basic point plot

```
ggplot(data = tetris, aes(x = time_spent, y = high_score)) +
  geom_point()

## adding names to the points
ggplot(data = tetris, aes(x = time_spent, y = high_score)) +
  geom_point() +
  geom_text(aes(label = kid))

####push the text away from the point
ggplot(data = tetris, aes(x = time_spent, y = high_score)) +
  geom_point() +
  geom_text(aes(label = kid), nudge_y = 5) ## move the text 5 points from the
    point
```

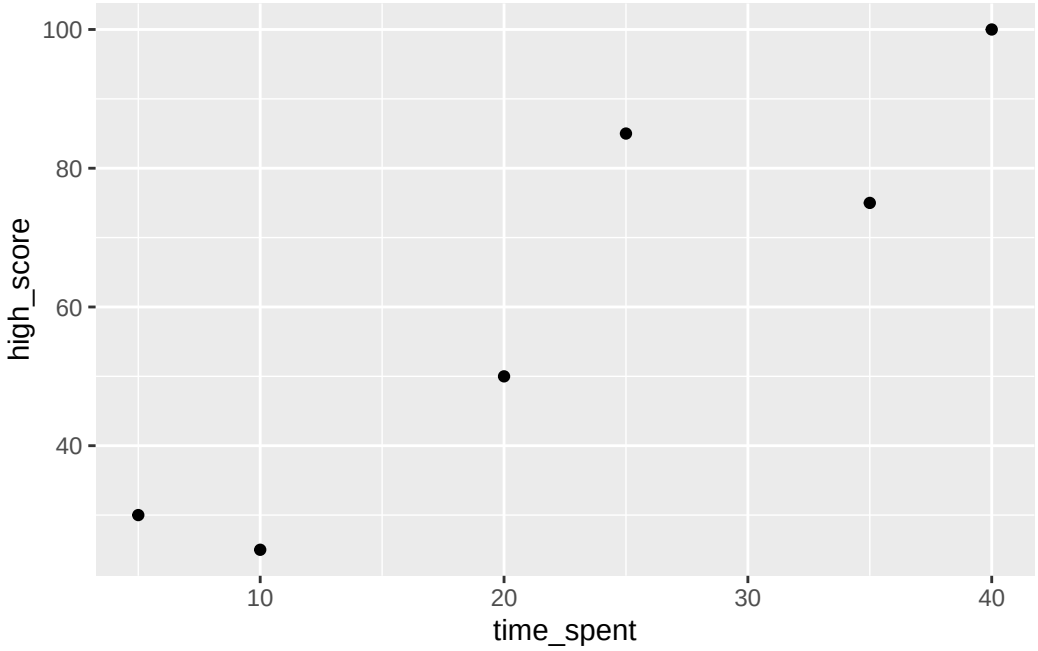

Figure 18: Basic point plot

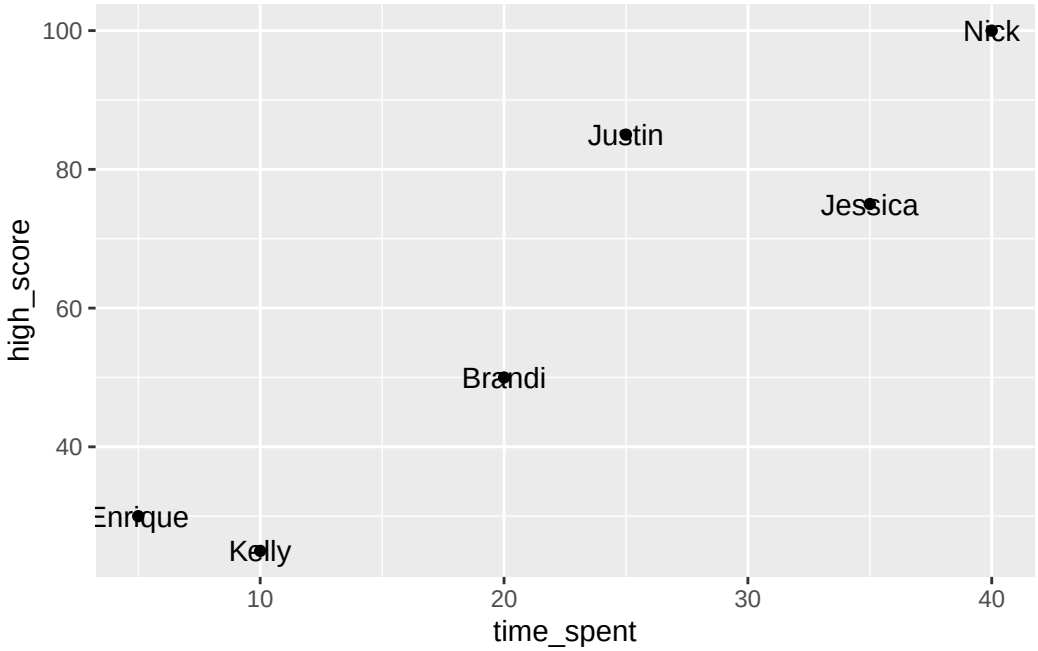

Figure 19: adding names to the points

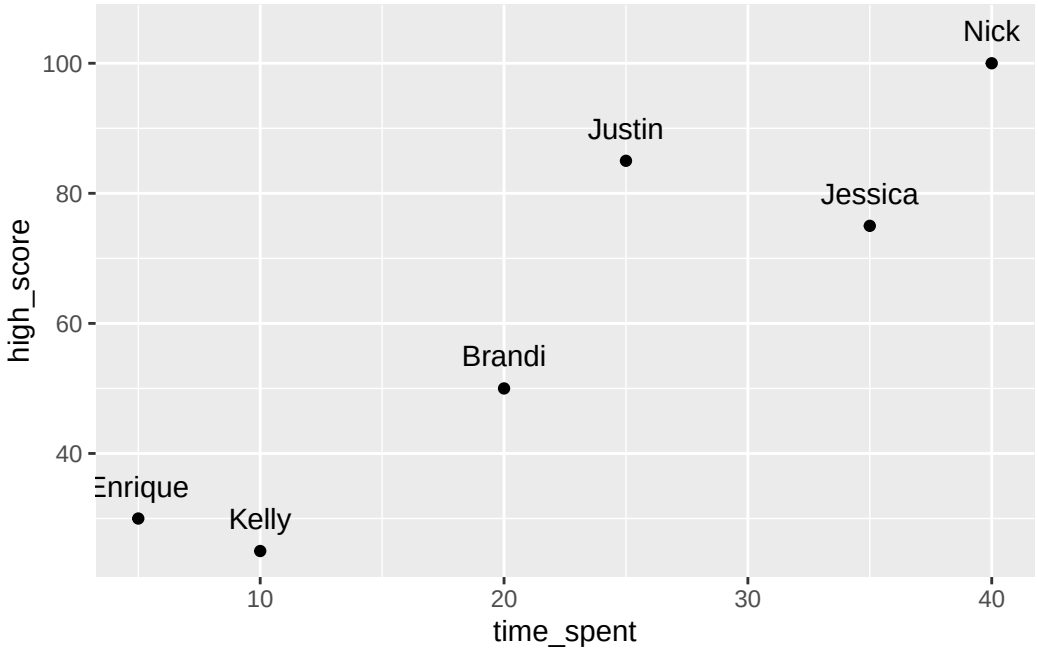

Figure 20: push the text away from the point

3.2 Make a plot using Congress (cel) data

```
attach(cel)
cel %>%
  filter(congress == 115) %>%
  ggplot(aes(x = dwnom1, y = all_pass, label = thomas_name)) +
  geom_point() +
  geom_text()

cel %>%
  filter(congress == 115) %>%
  ggplot(aes(x = dwnom1, y = all_pass, label = thomas_name)) +
  geom_point() +
  ## specify here that you only want the geom_text to apply to a subset of the
    data
  geom_text(data = filter(cel, congress == 115 & all_pass > 8))
```

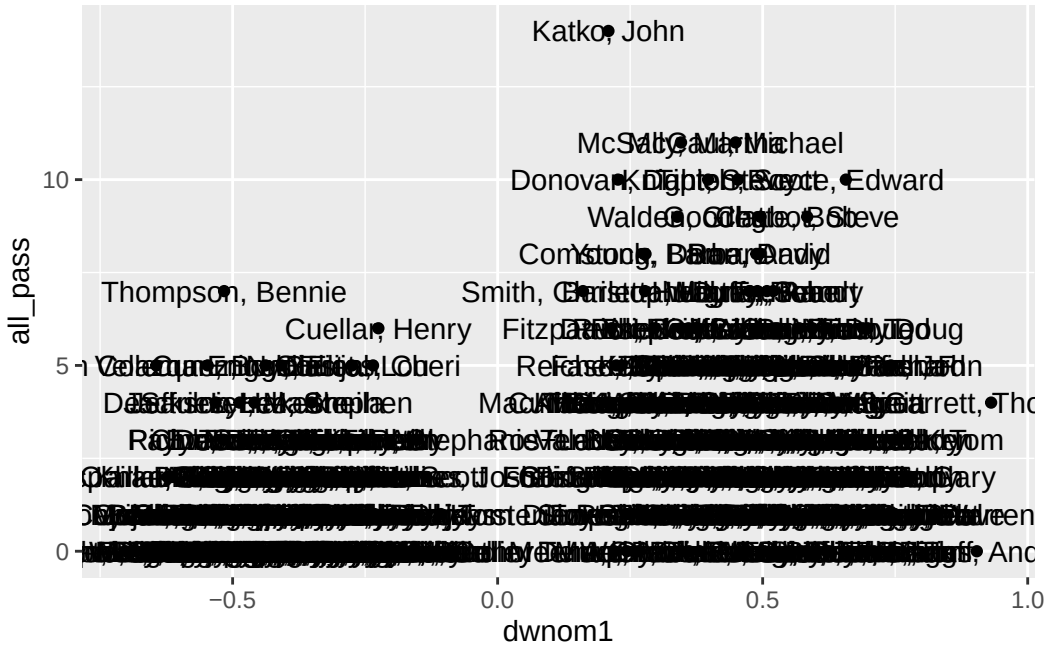

Figure 21: Basic point plot

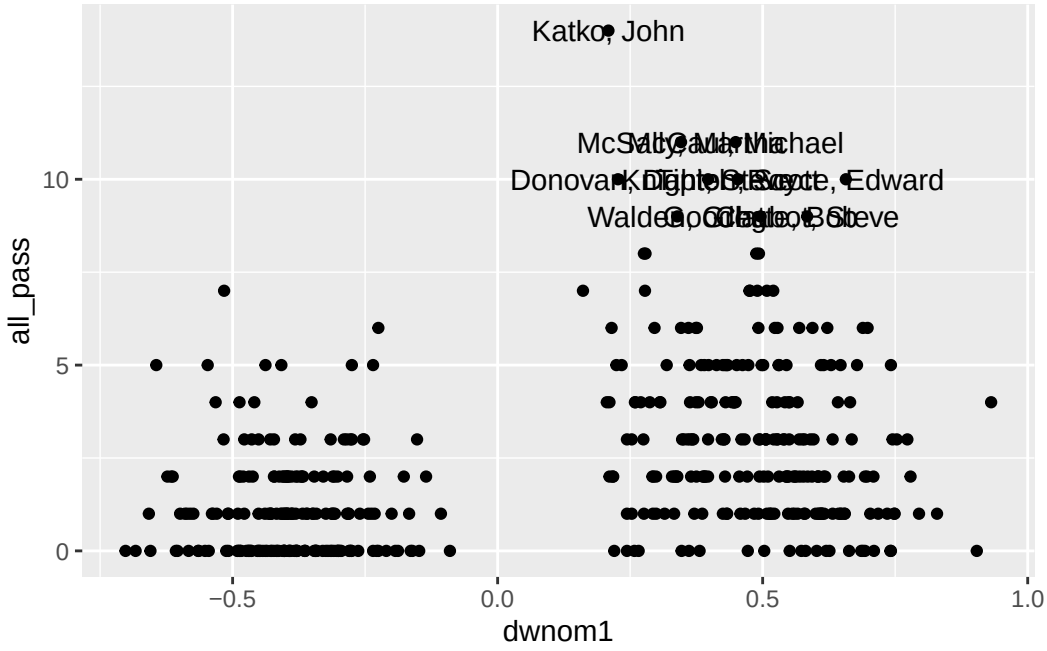

Figure 22: specification on geom_text

Point plots of all_pass against dwnom1

```
require(ggrepel)
cel %>%
  filter(congress == 115) %>%
  ggplot(aes(x = dwnom1, y = all_pass)) +
  geom_point() +
  geom_text_repel(data = filter(cel, congress == 115 & all_pass > 8),
                  aes(x = dwnom1, y = all_pass, label = thomas_name))

cel %>% filter(congress == 115) %>%
  ggplot(aes(x = dwnom1, y = all_pass)) +
  geom_point() +
  geom_text_repel(data = filter(cel, congress == 115 & all_pass > 8),
                  aes(x = dwnom1, y = all_pass, label = thomas_name)) +
  annotate("rect", xmin = .05, xmax = .4, ymin = 13, ymax = 15, alpha = .2,
           fill = "red") +
  annotate("text", x = .6, y = 14, label = "Most Passed", color = "red")
```

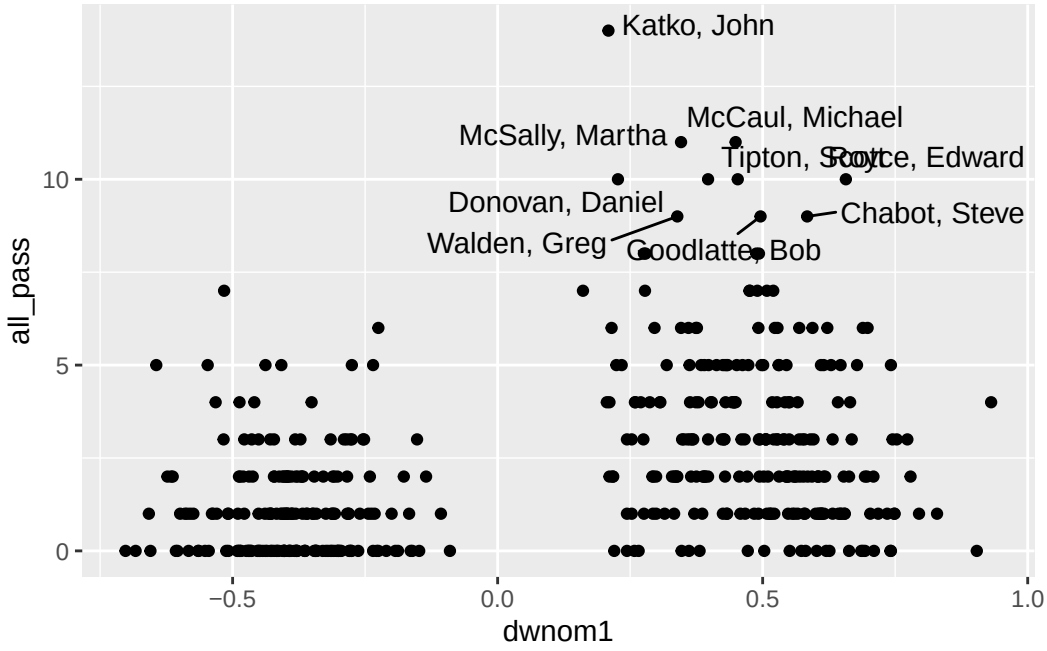

Figure 23: Basic point plot

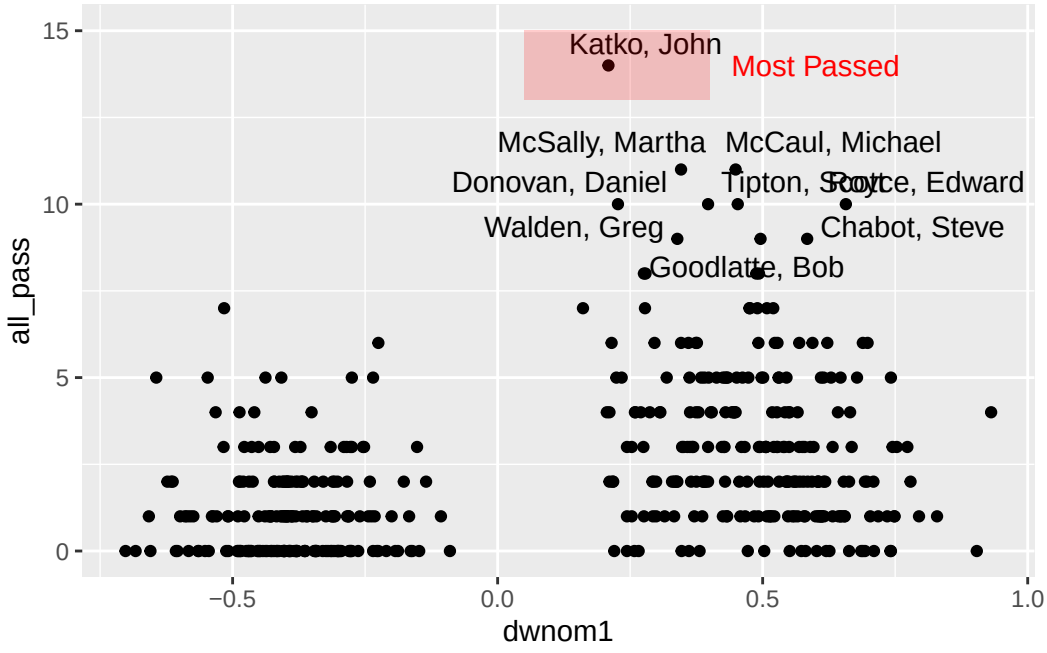

Figure 24: specification on geom_text

Using ggrepel to correct overlapping of text

3.3 Modifying visual elements and themes

3.3.1 Colors legends and themes using the cces data

```
plot_data <- select(cces, "CC18_308a", "ideo5", "educ", "faminc_new", "employ")
    %>%
  drop_na()

## color gradient is automatic here
ggplot(plot_data, aes(x = ideo5, y = CC18_308a, color = educ, size = faminc_new)
    ) +
  geom_jitter()

ggplot(plot_data, aes(x = ideo5, y = CC18_308a, color = educ, size = faminc_new)
    ) +
  geom_jitter() +
  #####use scale_color_gradient here to show continuous change in a numeric
    variable
  scale_color_gradient(low = "gray", high = "purple") ## gradients are used for
    continuous numeric variables
```

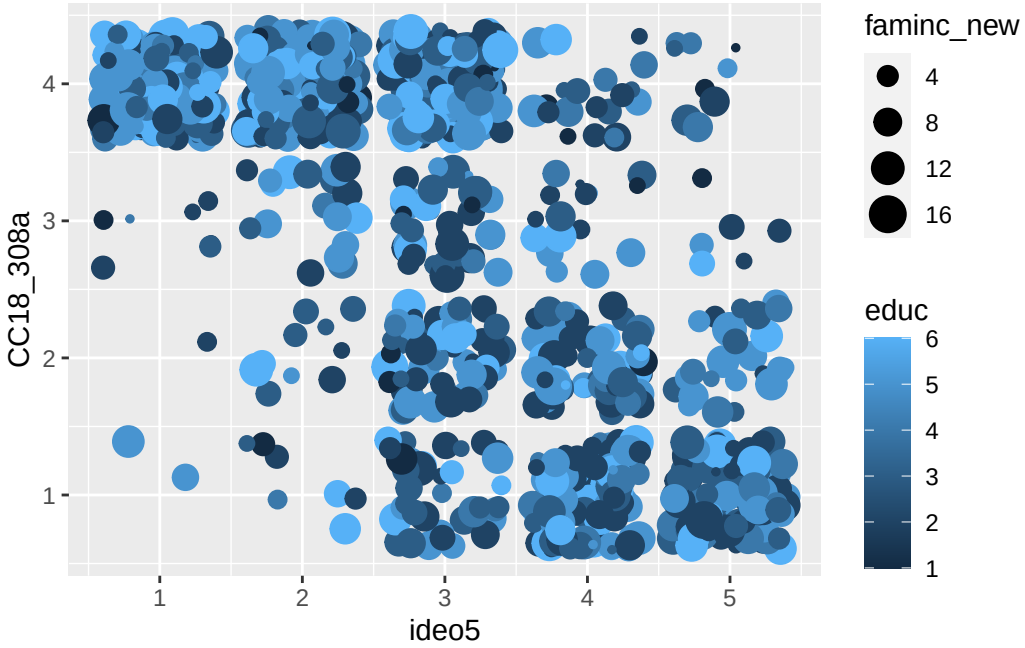

Figure 25: Automatic color gradient

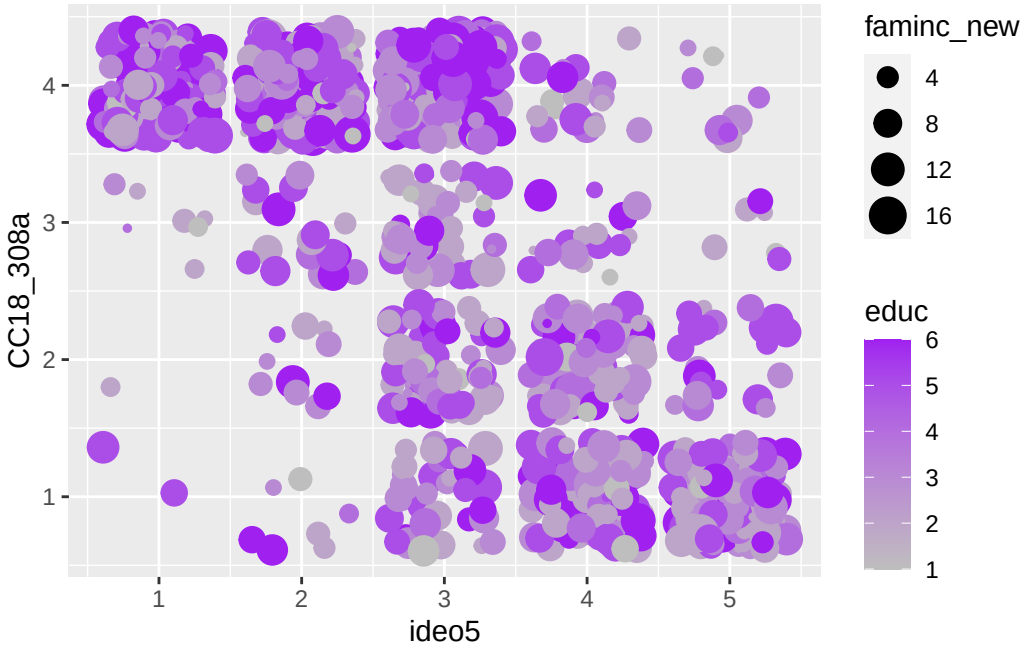

Figure 26: Manually selecting color

Using ggrepel to correct overlapping of text

use employment as categorical variable

```
plot_data$employ_cat <- recode(plot_data$employ,
                               `1` = "Full-time",
                               `2` = "Part-time",
                               `3` = "Temp. Layoff",
                               `4` = "Unemployed",
                               `5` = "Retired",
                               `6` = "Disabled",
                               `7` = "Homemaker",
                               `8` = "Student",
                               `9` = "Other")
```

instead of scale color gradient, use scale color brewer for the discrete variable created above

```
require(RColorBrewer)
ggplot(plot_data, aes(x = ideo5, y = CC18_308a, color = employ_cat)) +
  geom_jitter() +
  scale_color_brewer(palette = "RdYlGn")
```

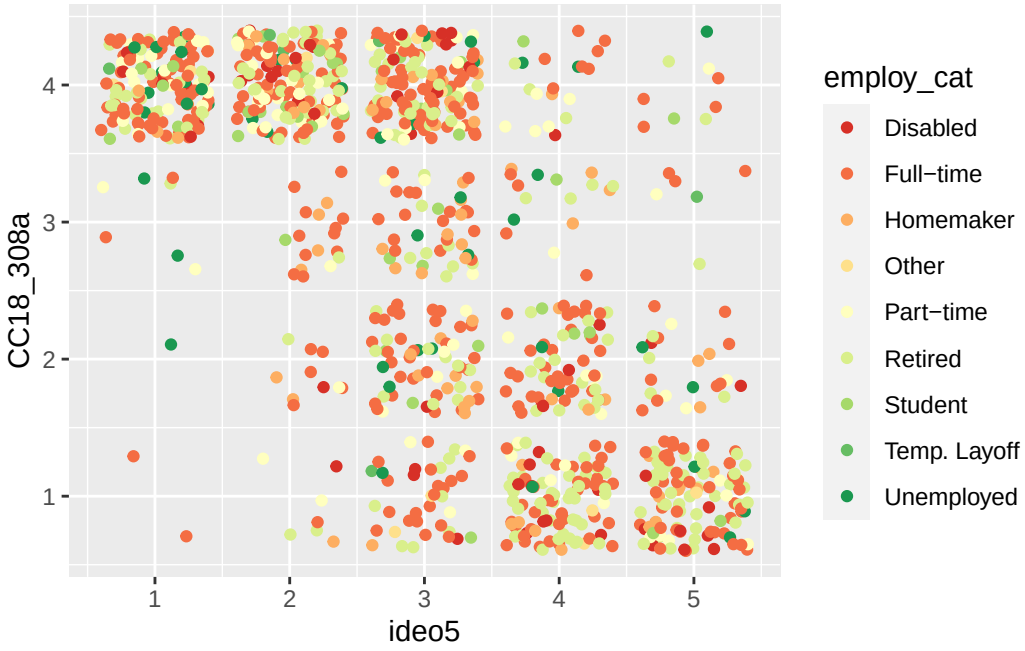

Figure 27: Using RColorBrewer package

Many other palettes can be found at palettes.

```
###renaming the employ_cat column
plot_data <- rename(plot_data, "Employment" = employ_cat) ## note the order!!
```

3.3.2 More theme modifications