Week 3 Notes

Katz

1/31/2023

## One continuous variable (either predictor or outcome variable)

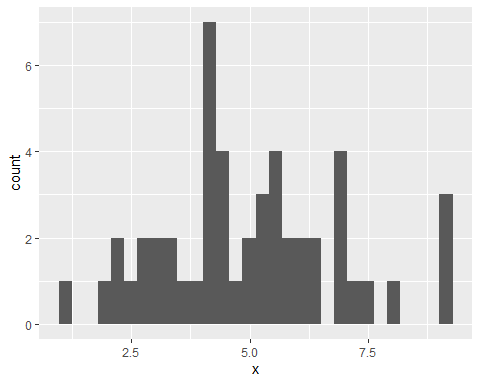
This is another comment

num <- 50  
mu <- 5  
stdev <- 2 #  
  
x <- rnorm(n = num, mean = mu, sd = stdev)  
hist(x)

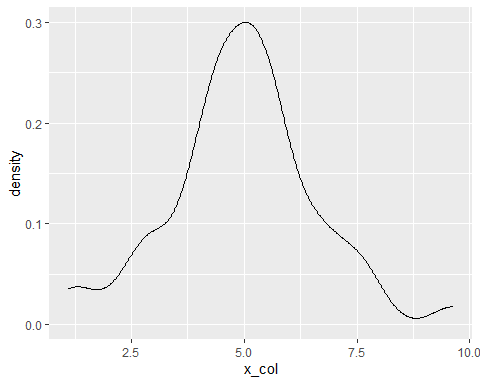
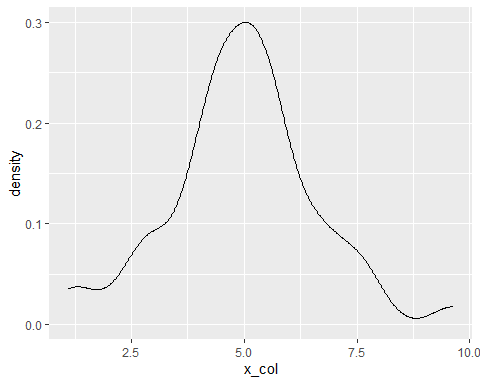
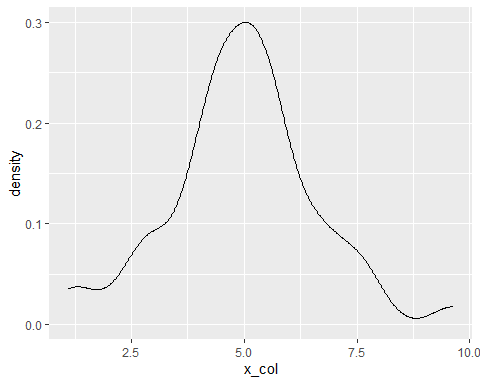
You can also do this using ggplot rather than base R graphics

num <- 50  
mu <- 5  
stdev <- 2  
  
x\_vec <- rnorm(n = num, mean = mu, sd = stdev)  
x\_df <- tibble(x\_col = x\_vec)  
  
ggplot(data = x\_df, mapping = aes(x = x)) +  
 geom\_histogram()

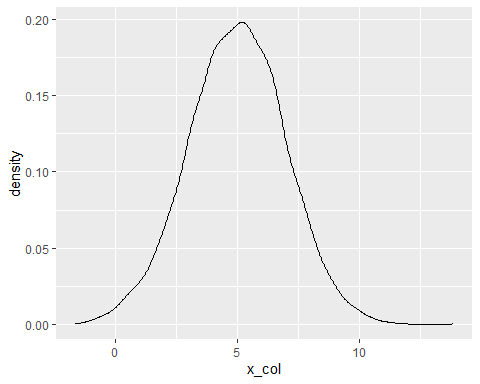
## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.



You can use geom\_density() instead of geom\_histogram() to get a smooth graph



Just for fun, look at what happens to the the plot if you increase the sample size



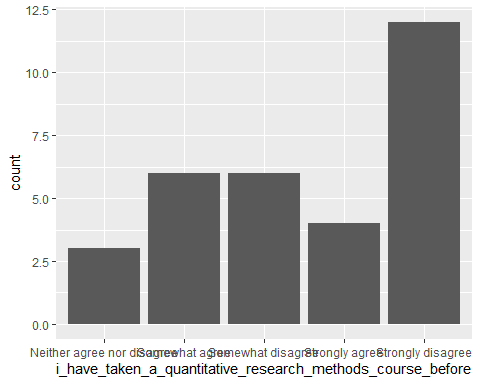
## One Discrete Variable (either predictor or outcome)

survey\_df <- read\_csv("./data/ENGE 6614 Prior Knowledge\_cleaned.csv")

## Rows: 31 Columns: 49  
## ── Column specification ────────────────────────────────────────────────────────  
## Delimiter: ","  
## chr (48): I have taken a quantitative research methods course before, I am i...  
## dbl (1): student\_id  
##   
## ℹ Use `spec()` to retrieve the full column specification for this data.  
## ℹ Specify the column types or set `show\_col\_types = FALSE` to quiet this message.

survey\_df <- survey\_df %>% clean\_names()

survey\_df %>%   
 ggplot(aes(x = i\_have\_taken\_a\_quantitative\_research\_methods\_course\_before)) +  
 geom\_bar()



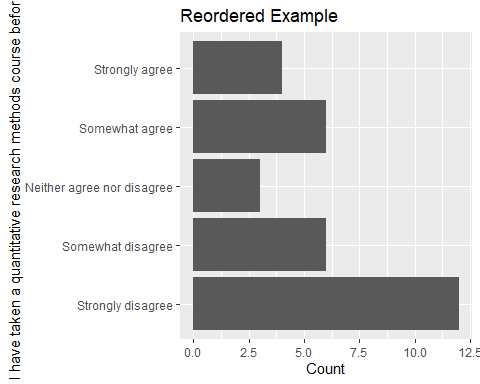
Notice that the ordering is not quite what we would want. It is alphabetical. Try to find how we can fix this.

Here is one way:

glimpse(survey\_df)

## Rows: 31  
## Columns: 49  
## $ student\_id <dbl> …  
## $ i\_have\_taken\_a\_quantitative\_research\_methods\_course\_before <chr> …  
## $ i\_am\_interested\_in\_using\_quantitative\_research\_methods\_in\_my\_own\_work <chr> …  
## $ i\_know\_what\_a\_type\_i\_error\_is <chr> …  
## $ i\_know\_what\_a\_type\_ii\_error\_is <chr> …  
## $ i\_know\_what\_a\_statistical\_confidence\_level\_is <chr> …  
## $ i\_know\_what\_a\_p\_value\_is <chr> …  
## $ i\_know\_what\_p\_hacking\_means <chr> …  
## $ i\_know\_what\_statistical\_power\_means <chr> …  
## $ i\_have\_heard\_of\_frequentist\_statistics\_before <chr> …  
## $ i\_have\_heard\_of\_bayesian\_statistics\_before <chr> …  
## $ i\_have\_heard\_the\_term\_parametric\_statistics\_before <chr> …  
## $ i\_have\_heard\_the\_term\_non\_parametric\_statistics\_before <chr> …  
## $ i\_know\_what\_a\_histogram\_is <chr> …  
## $ i\_know\_what\_a\_probability\_distribution\_is <chr> …  
## $ i\_know\_what\_a\_random\_variable\_is <chr> …  
## $ i\_know\_what\_a\_probability\_distribution\_function\_is <chr> …  
## $ i\_know\_what\_a\_cumulative\_distribution\_function\_is <chr> …  
## $ i\_know\_what\_the\_expectation\_of\_a\_random\_variable\_is <chr> …  
## $ i\_know\_how\_to\_calculate\_the\_variance\_of\_a\_random\_variable <chr> …  
## $ i\_know\_what\_a\_z\_score\_is <chr> …  
## $ i\_know\_how\_to\_calculate\_the\_correlation\_between\_two\_variables <chr> …  
## $ i\_know\_how\_to\_interpret\_the\_correlation\_coefficient\_between\_two\_variables <chr> …  
## $ i\_have\_heard\_of\_linear\_regression <chr> …  
## $ i\_know\_how\_to\_run\_a\_linear\_regression\_in\_some\_software\_or\_by\_hand\_if\_im\_feeling\_wild <chr> …  
## $ i\_know\_how\_to\_interpret\_a\_linear\_regression <chr> …  
## $ i\_have\_heard\_of\_multiple\_regression <chr> …  
## $ i\_know\_how\_to\_perform\_a\_multiple\_regression <chr> …  
## $ i\_know\_how\_to\_interpret\_a\_multiple\_regression <chr> …  
## $ i\_have\_heard\_of\_logistic\_regression <chr> …  
## $ i\_understand\_when\_to\_use\_a\_logistic\_regression <chr> …  
## $ i\_know\_how\_to\_interpret\_the\_results\_of\_a\_logistic\_regression <chr> …  
## $ i\_have\_heard\_of\_t\_tests <chr> …  
## $ i\_have\_performed\_a\_t\_test\_before <chr> …  
## $ i\_know\_how\_to\_interpret\_the\_results\_of\_a\_t\_test <chr> …  
## $ i\_have\_heard\_of\_analysis\_of\_variance <chr> …  
## $ i\_understand\_when\_to\_run\_an\_analysis\_of\_variance\_anova <chr> …  
## $ i\_know\_how\_to\_interpret\_the\_results\_from\_an\_anova <chr> …  
## $ i\_have\_heard\_of\_a\_chi\_square\_test <chr> …  
## $ i\_have\_used\_a\_chi\_square\_test\_before <chr> …  
## $ i\_know\_how\_to\_interpret\_the\_results\_of\_a\_chi\_square\_test <chr> …  
## $ i\_have\_heard\_of\_cluster\_analysis\_before <chr> …  
## $ i\_have\_used\_cluster\_analysis\_before <chr> …  
## $ i\_know\_how\_to\_interpret\_the\_results\_of\_a\_cluster\_analysis <chr> …  
## $ i\_have\_heard\_of\_factor\_analysis\_either\_exploratory\_or\_confirmatory <chr> …  
## $ i\_have\_used\_factor\_analysis\_either\_exploratory\_or\_confirmatory <chr> …  
## $ i\_know\_how\_to\_interpret\_the\_results\_of\_a\_factor\_analysis\_either\_exploratory\_or\_confirmatory <chr> …  
## $ i\_already\_have\_r\_and\_rstudio\_downloaded\_to\_my\_computer <chr> …  
## $ i\_have\_used\_r\_before <chr> …

q\_levels <- c("Strongly disagree", "Somewhat disagree", "Neither agree nor disagree",  
 "Somewhat agree", "Strongly agree")  
  
  
  
survey\_df$i\_have\_taken\_a\_quantitative\_research\_methods\_course\_before <- factor(survey\_df$i\_have\_taken\_a\_quantitative\_research\_methods\_course\_before,   
 levels = q\_levels,  
 ordered = TRUE)  
  
  
  
## Now try plotting  
  
survey\_df %>%   
 ggplot(aes(x = i\_have\_taken\_a\_quantitative\_research\_methods\_course\_before)) +  
 geom\_bar() +  
 labs(x = "I have taken a quantitative research methods course before",  
 y = "Count",  
 title = "Reordered Example") +  
 coord\_flip()



### Joining two datasets

Let’s imagine that we have a separate dataset that has information about the students who completed the pre-course prior knowledge survey.

First, we will load in that dataset

survey\_info\_df <- read\_csv("./data/survey\_student\_info.csv")

## Rows: 31 Columns: 4  
## ── Column specification ────────────────────────────────────────────────────────  
## Delimiter: ","  
## chr (3): standing, college, required  
## dbl (1): student\_id  
##   
## ℹ Use `spec()` to retrieve the full column specification for this data.  
## ℹ Specify the column types or set `show\_col\_types = FALSE` to quiet this message.

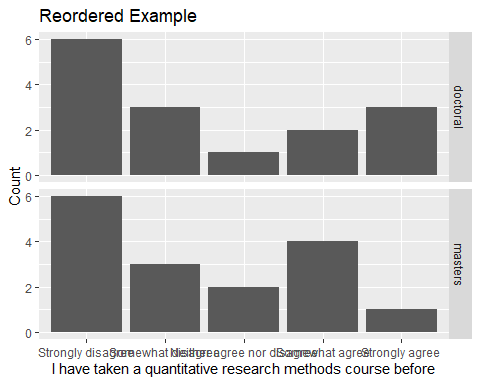
Next, let’s join the two datasets based on the student id column, which is in each of the two dataframes.

survey\_df <- survey\_df %>% inner\_join(survey\_info\_df, by = "student\_id")

Now we should have both datasets joined into one and saved as survey\_df.

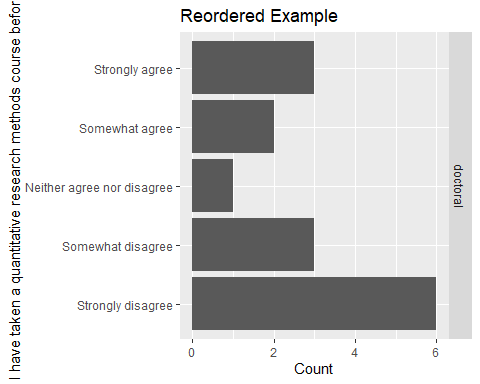
With this, we can make some nicer plots and do something like use facet\_grid() to look at students who are masters and doctoral students, for example.

survey\_df %>%   
 ggplot(aes(x = i\_have\_taken\_a\_quantitative\_research\_methods\_course\_before)) +  
 geom\_bar() +  
 facet\_grid(standing ~.) +  
 labs(x = "I have taken a quantitative research methods course before",  
 y = "Count",  
 title = "Reordered Example")



The x axis looks a little crowded. What if we try coord\_flip()

survey\_df %>%   
 filter(standing == "doctoral") %>%  
 ggplot(aes(x = i\_have\_taken\_a\_quantitative\_research\_methods\_course\_before)) +  
 geom\_bar() +  
 coord\_flip() +  
 facet\_grid(standing ~.) +  
 labs(x = "I have taken a quantitative research methods course before",  
 y = "Count",  
 title = "Reordered Example")



A quick note on filters

# these two accomplish the same thing  
filtered\_df <- filter(survey\_df, required == "yes")  
filtered\_df

## # A tibble: 16 × 52  
## student\_id i\_have\_t…¹ i\_am\_…² i\_kno…³ i\_kno…⁴ i\_kno…⁵ i\_kno…⁶ i\_kno…⁷ i\_kno…⁸  
## <dbl> <ord> <chr> <chr> <chr> <chr> <chr> <chr> <chr>   
## 1 1 Somewhat … Somewh… Strong… Strong… Somewh… Strong… Strong… Neithe…  
## 2 2 Strongly … Neithe… Somewh… Somewh… Somewh… Somewh… Strong… Strong…  
## 3 4 Somewhat … Strong… Strong… Strong… Somewh… Neithe… Somewh… Somewh…  
## 4 8 Somewhat … Somewh… Somewh… Somewh… Somewh… Somewh… Somewh… Somewh…  
## 5 9 Strongly … Strong… Somewh… Somewh… Strong… Strong… Strong… Somewh…  
## 6 11 Strongly … Strong… Strong… Strong… Strong… Strong… Strong… Somewh…  
## 7 16 Strongly … Strong… Somewh… Somewh… Somewh… Somewh… Strong… Somewh…  
## 8 17 Strongly … Strong… Strong… Strong… Somewh… Somewh… Strong… Strong…  
## 9 18 Somewhat … Somewh… Somewh… Somewh… Somewh… Somewh… Strong… Somewh…  
## 10 20 Strongly … Neithe… Neithe… Neithe… Neithe… Neithe… Strong… Strong…  
## 11 22 Strongly … Strong… Strong… Strong… Somewh… Somewh… Strong… Strong…  
## 12 23 Somewhat … Strong… Somewh… Somewh… Somewh… Strong… Somewh… Strong…  
## 13 25 Neither a… Strong… Strong… Strong… Neithe… Neithe… Strong… Strong…  
## 14 26 Somewhat … Strong… Strong… Strong… Strong… Strong… Somewh… Strong…  
## 15 28 Strongly … Neithe… Somewh… Somewh… Somewh… Somewh… Somewh… Somewh…  
## 16 30 Strongly … Strong… Somewh… Somewh… Somewh… Somewh… Somewh… Somewh…  
## # … with 43 more variables:  
## # i\_have\_heard\_of\_frequentist\_statistics\_before <chr>,  
## # i\_have\_heard\_of\_bayesian\_statistics\_before <chr>,  
## # i\_have\_heard\_the\_term\_parametric\_statistics\_before <chr>,  
## # i\_have\_heard\_the\_term\_non\_parametric\_statistics\_before <chr>,  
## # i\_know\_what\_a\_histogram\_is <chr>,  
## # i\_know\_what\_a\_probability\_distribution\_is <chr>, …

filtered\_df <- survey\_df %>% filter(required == "yes")  
filtered\_df

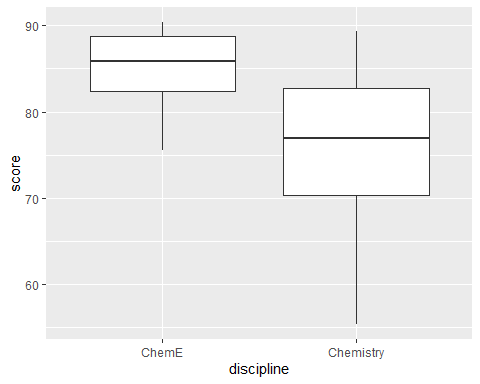
## # A tibble: 16 × 52  
## student\_id i\_have\_t…¹ i\_am\_…² i\_kno…³ i\_kno…⁴ i\_kno…⁵ i\_kno…⁶ i\_kno…⁷ i\_kno…⁸  
## <dbl> <ord> <chr> <chr> <chr> <chr> <chr> <chr> <chr>   
## 1 1 Somewhat … Somewh… Strong… Strong… Somewh… Strong… Strong… Neithe…  
## 2 2 Strongly … Neithe… Somewh… Somewh… Somewh… Somewh… Strong… Strong…  
## 3 4 Somewhat … Strong… Strong… Strong… Somewh… Neithe… Somewh… Somewh…  
## 4 8 Somewhat … Somewh… Somewh… Somewh… Somewh… Somewh… Somewh… Somewh…  
## 5 9 Strongly … Strong… Somewh… Somewh… Strong… Strong… Strong… Somewh…  
## 6 11 Strongly … Strong… Strong… Strong… Strong… Strong… Strong… Somewh…  
## 7 16 Strongly … Strong… Somewh… Somewh… Somewh… Somewh… Strong… Somewh…  
## 8 17 Strongly … Strong… Strong… Strong… Somewh… Somewh… Strong… Strong…  
## 9 18 Somewhat … Somewh… Somewh… Somewh… Somewh… Somewh… Strong… Somewh…  
## 10 20 Strongly … Neithe… Neithe… Neithe… Neithe… Neithe… Strong… Strong…  
## 11 22 Strongly … Strong… Strong… Strong… Somewh… Somewh… Strong… Strong…  
## 12 23 Somewhat … Strong… Somewh… Somewh… Somewh… Strong… Somewh… Strong…  
## 13 25 Neither a… Strong… Strong… Strong… Neithe… Neithe… Strong… Strong…  
## 14 26 Somewhat … Strong… Strong… Strong… Strong… Strong… Somewh… Strong…  
## 15 28 Strongly … Neithe… Somewh… Somewh… Somewh… Somewh… Somewh… Somewh…  
## 16 30 Strongly … Strong… Somewh… Somewh… Somewh… Somewh… Somewh… Somewh…  
## # … with 43 more variables:  
## # i\_have\_heard\_of\_frequentist\_statistics\_before <chr>,  
## # i\_have\_heard\_of\_bayesian\_statistics\_before <chr>,  
## # i\_have\_heard\_the\_term\_parametric\_statistics\_before <chr>,  
## # i\_have\_heard\_the\_term\_non\_parametric\_statistics\_before <chr>,  
## # i\_know\_what\_a\_histogram\_is <chr>,  
## # i\_know\_what\_a\_probability\_distribution\_is <chr>, …

# A little more about plotting

We are going to shift gears again and look at a few different kinds of plots. The main thing to remember here is that you want to think about whether the variables you have are nominal, ordinal, or continuous (that includes interval and ratio).

## Discrete Predictor, Continuous Outcome

group\_size <- 20  
chem\_e\_scores <- rnorm(n = group\_size, mean = 85, sd = 4)  
chem\_scores <- rnorm(n = group\_size, mean = 78, sd = 6)  
  
  
data\_df <- tibble(  
 discipline = rep(c("ChemE", "Chemistry"), each = group\_size),  
 score = c(chem\_e\_scores, chem\_scores)  
)  
  
data\_df %>%   
 ggplot(aes(x = discipline, y = score)) +  
 geom\_boxplot()

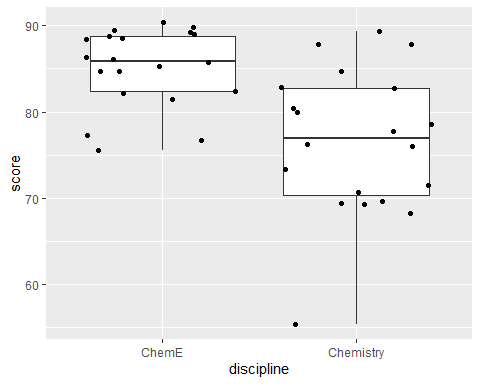


You can make a few modifications to possibly make this easier to read.

The first is to put the discrete category on the y axis instead of the x axis.

The second is to use geom\_jitter() in addition to geom\_boxplot() to show the individual points in each group.

data\_df %>%   
 ggplot(aes(y = score, x = discipline)) +  
 geom\_boxplot() +  
 geom\_jitter()



## Continuous predictor and continuous outcome

First, let’s re-do a lot of the steps in this week’s script for reading in data and transforming it a little

mydata <- read\_csv("./data/Free Reduced Lunch by Schools and Grade Structures 2008-2017\_final.csv")

## Rows: 2101 Columns: 137  
## ── Column specification ────────────────────────────────────────────────────────  
## Delimiter: ","  
## chr (136): sch\_id, div\_name, school\_num, school\_name, school\_name2, type0809...  
## dbl (1): div\_num  
##   
## ℹ Use `spec()` to retrieve the full column specification for this data.  
## ℹ Specify the column types or set `show\_col\_types = FALSE` to quiet this message.

# check the structure of the data (this output is a bit long)  
#str(mydata) # I would recommend not running this line because it takes so long to print out  
  
str(mydata$total\_2017)

## chr [1:2101] "742" "236" "60" "624" "286" "485" "583" "550" "600" "514" ...

# a faster alternative to str() is glimpse()  
glimpse(mydata)

## Rows: 2,101  
## Columns: 137  
## $ sch\_id <chr> "001-0070", "001-0080", "001-0530", "001-0540", "001-05…  
## $ div\_num <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 2, 2, 2, 2, 2, 2, 2, 2…  
## $ div\_name <chr> "Accomack County", "Accomack County", "Accomack County"…  
## $ school\_num <chr> "0070\xa0", "0080\xa0", "0530\xa0", "0540\xa0", "0580\x…  
## $ school\_name <chr> "NANDUA HIGH", "CHINCOTEAGUE ELEM", "TANGIER COMBINED",…  
## $ school\_name2 <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,…  
## $ type0809 <chr> "SCH-HIGH", "SCH-ELEM", "SCH-COMB", "SCH-HIGH", "SCH-CO…  
## $ lowgrade\_2008 <chr> "9", "PK", "KG", "9", "6", "PK", "PK", "PK", "PK", "6",…  
## $ higrade\_2008 <chr> "12", "5", "12", "12", "12", "5", "5", "5", "5", "8", "…  
## $ totalFT\_2008 <chr> "731", "263", "80", "638", "333", "536", "610", "490", …  
## $ total\_2008 <chr> "731", "263", "80", "638", "333", "536", "610", "490", …  
## $ snp\_0809 <chr> "659", "257", "80", "622", "329", "518", "569", "481", …  
## $ free\_elig\_0809 <chr> "306", "95", "38", "289", "101", "350", "394", "274", "…  
## $ free\_per\_0809 <chr> "46.43%", "36.96%", "47.50%", "46.46%", "30.70%", "67.5…  
## $ red\_elig\_0809 <chr> "64", "8", "0", "56", "9", "32", "49", "45", "61", "52"…  
## $ red\_per\_0809 <chr> "9.71%", "3.11%", "0.00%", "9.00%", "2.74%", "6.18%", "…  
## $ totalFRL\_0809 <chr> "370", "103", "38", "345", "110", "382", "443", "319", …  
## $ totalper\_0809 <chr> "56.15%", "40.08%", "47.50%", "55.47%", "33.43%", "73.7…  
## $ type0910 <chr> "SCH-HIGH", "SCH-ELEM", "SCH-COMB", "SCH-HIGH", "SCH-CO…  
## $ lowgrade\_2009 <chr> "9", "PK", "KG", "9", "6", "PK", "PK", "PK", "PK", "6",…  
## $ higrade\_2009 <chr> "12", "5", "12", "12", "12", "5", "5", "5", "5", "8", "…  
## $ totalFT\_2009 <chr> "654", "266", "78", "634", "301", "531", "625", "500", …  
## $ total\_2009 <chr> "654", "266", "78", "634", "301", "531", "625", "500", …  
## $ snp\_0910 <chr> "655", "266", "78", "635", "300", "533", "605", "494", …  
## $ free\_elig\_0910 <chr> "290", "99", "36", "286", "103", "368", "404", "296", "…  
## $ free\_per\_0910 <chr> "44.27%", "37.22%", "46.15%", "45.04%", "34.33%", "69.0…  
## $ red\_elig\_0910 <chr> "37", "14", "0", "66", "9", "28", "61", "26", "47", "32…  
## $ red\_per\_0910 <chr> "5.65%", "5.26%", "0.00%", "10.39%", "3.00%", "5.25%", …  
## $ totalFRL\_09010 <chr> "327", "113", "36", "352", "112", "396", "465", "322", …  
## $ totalper\_0910 <chr> "49.92%", "42.48%", "46.15%", "55.43%", "37.33%", "74.3…  
## $ type1011 <chr> "SCH-HIGH", "SCH-ELEM", "SCH-COMB", "SCH-HIGH", "SCH-CO…  
## $ lowgrade\_2010 <chr> "9", "PK", "KG", "9", "6", "PK", "PK", "PK", "PK", "6",…  
## $ higrade\_2010 <chr> "12", "5", "12", "12", "12", "5", "5", "5", "5", "8", "…  
## $ totalFT\_2010 <chr> "603", "268", "74", "614", "279", "536", "604", "558", …  
## $ total\_2010 <chr> "603", "268", "74", "614", "279", "536", "604", "558", …  
## $ snp\_1011 <chr> "603", "277", "74", "606", "279", "533", "597", "551", …  
## $ free\_elig\_1011 <chr> "285", "108", "32", "308", "100", "375", "415", "318", …  
## $ free\_per\_1011 <chr> "47.26%", "38.99%", "43.24%", "50.83%", "35.84%", "70.3…  
## $ red\_elig\_1011 <chr> "46", "8", "0", "50", "5", "35", "41", "30", "45", "36"…  
## $ red\_per\_1011 <chr> "7.63%", "2.89%", "0.00%", "8.25%", "1.79%", "6.57%", "…  
## $ totalFRL\_1011 <chr> "331", "116", "32", "358", "105", "410", "456", "348", …  
## $ totalper\_1011 <chr> "54.89%", "41.88%", "43.24%", "59.08%", "37.63%", "76.9…  
## $ type1112 <chr> "SCH-HIGH", "SCH-ELEM", "SCH-COMB", "SCH-HIGH", "SCH-CO…  
## $ lowgrade\_2011 <chr> "9", "PK", "KG", "9", "6", "PK", "PK", "PK", "PK", "6",…  
## $ higrade\_2011 <chr> "12", "5", "12", "12", "12", "5", "5", "5", "5", "8", "…  
## $ totalFT\_2011 <chr> "593", "276", "73", "605", "291", "533", "593", "582", …  
## $ total\_2011 <chr> "593", "276", "73", "605", "291", "533", "593", "582", …  
## $ snp\_1112 <chr> "593", "281", "73", "611", "294", "529", "578", "588", …  
## $ free\_elig\_1112 <chr> "289", "116", "31", "318", "95", "366", "426", "371", "…  
## $ free\_per\_1112 <chr> "48.74%", "41.28%", "42.47%", "52.05%", "32.31%", "69.1…  
## $ red\_elig\_1112 <chr> "50", "14", "0", "44", "9", "34", "27", "32", "33", "40…  
## $ red\_per\_1112 <chr> "8.43%", "4.98%", "0.00%", "7.20%", "3.06%", "6.43%", "…  
## $ totalFRL\_1112 <chr> "339", "130", "31", "362", "104", "400", "453", "403", …  
## $ totalper\_1112 <chr> "57.17%", "46.26%", "42.47%", "59.25%", "35.37%", "75.6…  
## $ type1213 <chr> "SCH-HIGH", "SCH-ELEM", "SCH-COMB", "SCH-HIGH", "SCH-CO…  
## $ lowgrade\_2012 <chr> "9", "PK", "KG", "9", "6", "PK", "PK", "PK", "PK", "6",…  
## $ higrade\_2012 <chr> "12", "5", "12", "12", "12", "5", "5", "5", "5", "8", "…  
## $ totalFT\_2012 <chr> "637", "258", "68", "579", "292", "505", "628", "562", …  
## $ total\_2012 <chr> "637", "258", "68", "579", "292", "505", "628", "562", …  
## $ snp\_1213 <chr> "633", "259", "68", "579", "292", "498", "625", "571", …  
## $ free\_elig\_1213 <chr> "324", "117", "21", "348", "102", "364", "459", "385", …  
## $ free\_per\_1213 <chr> "51.18%", "45.17%", "30.88%", "60.10%", "34.93%", "73.0…  
## $ red\_elig\_1213 <chr> "42", "20", "5", "33", "12", "25", "31", "25", "37", "2…  
## $ red\_per\_1213 <chr> "6.64%", "7.72%", "7.35%", "5.70%", "4.11%", "5.02%", "…  
## $ totalFRL\_1213 <chr> "366", "137", "26", "381", "114", "389", "490", "410", …  
## $ totalper\_1213 <chr> "57.82%", "52.90%", "38.24%", "65.80%", "39.04%", "78.1…  
## $ type1314 <chr> "SCH-HIGH", "SCH-ELEM", "SCH-COMB", "SCH-HIGH", "SCH-CO…  
## $ lowgrade\_2013 <chr> "9", "PK", "KG", "9", "6", "PK", "PK", "PK", "PK", "6",…  
## $ higrade\_2013 <chr> "12", "5", "12", "12", "12", "5", "5", "5", "5", "8", "…  
## $ totalFT\_2013 <chr> "670", "238", "66", "582", "301", "530", "650", "551", …  
## $ total\_2013 <chr> "670", "238", "66", "582", "301", "530", "650", "551", …  
## $ snp\_1314 <chr> "668", "239", "56", "589", "302", "524", "631", "561", …  
## $ free\_elig\_1314 <chr> "346", "102", "12", "347", "115", "414", "443", "372", …  
## $ free\_per\_1314 <chr> "51.80%", "42.68%", "21.43%", "58.91%", "38.08%", "79.0…  
## $ red\_elig\_1314 <chr> "44", "19", "4", "54", "13", "30", "42", "19", "20", "3…  
## $ red\_per\_1314 <chr> "6.59%", "7.95%", "7.14%", "9.17%", "4.30%", "5.73%", "…  
## $ totalFRL\_1314 <chr> "390", "121", "16", "401", "128", "444", "485", "391", …  
## $ totalper\_1314 <chr> "58.38%", "50.63%", "28.57%", "68.08%", "42.38%", "84.7…  
## $ type1415 <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,…  
## $ lowgrade\_2014 <chr> "9", "PK", "KG", "9", "6", "PK", "PK", "PK", "PK", "6",…  
## $ higrade\_2014 <chr> "12", "5", "12", "12", "12", "5", "5", "5", "5", "8", "…  
## $ totalFT\_2014 <chr> "685", "251", "65", "581", "290", "510", "623", "559", …  
## $ total\_2014 <chr> "685", "251", "65", "581", "290", "510", "623", "559", …  
## $ snp\_1415 <chr> "672", "239", "61", "586", "291", "506", "607", "570", …  
## $ free\_elig\_1415 <chr> "361", "93", "14", "351", "97", "391", "411", "383", "5…  
## $ free\_per\_1415 <chr> "53.72%", "38.91%", "22.95%", "59.90%", "33.33%", "77.2…  
## $ red\_elig\_1415 <chr> "40", "17", "4", "40", "15", "28", "44", "25", "26", "3…  
## $ red\_per\_1415 <chr> "5.95%", "7.11%", "6.56%", "6.83%", "5.15%", "5.53%", "…  
## $ totalFRL\_1415 <chr> "401", "110", "18", "391", "112", "419", "455", "408", …  
## $ totalper\_1415 <chr> "59.67%", "46.03%", "29.51%", "66.72%", "38.49%", "82.8…  
## $ CEP\_1516 <chr> "#NULL!", "#NULL!", "#NULL!", "#NULL!", "#NULL!", "#NUL…  
## $ type1516 <chr> "SCH-HIGH", "SCH-ELEM", "SCH-COMB", "SCH-HIGH", "SCH-CO…  
## $ lowgrade\_2015 <chr> "9", "PK", "KG", "9", "6", "PK", "PK", "PK", "PK", "6",…  
## $ higrade\_2015 <chr> "12", "5", "12", "12", "12", "5", "5", "5", "5", "8", "…  
## $ totalFT\_2015 <chr> "737", "259", "65", "621", "291", "492", "601", "582", …  
## $ total\_2015 <chr> "737", "259", "65", "621", "291", "492", "601", "582", …  
## $ snp\_1516 <chr> "728", "268", "67", "608", "293", "491", "605", "603", …  
## $ free\_elig\_1516 <chr> "362", "109", "12", "339", "92", "358", "413", "387", "…  
## $ free\_per\_1516 <chr> "49.73%", "40.67%", "17.91%", "55.76%", "31.40%", "72.9…  
## $ red\_elig\_1516 <chr> "46", "16", "4", "47", "9", "36", "35", "30", "19", "38…  
## $ red\_Per\_1516 <chr> "6.32%", "5.97%", "5.97%", "7.73%", "3.07%", "7.33%", "…  
## $ totalFRL\_1516 <chr> "408", "125", "16", "386", "101", "394", "448", "417", …  
## $ totalper\_1516 <chr> "56.04%", "46.64%", "23.88%", "63.49%", "34.47%", "80.2…  
## $ CEP\_1617 <chr> "#NULL!", "#NULL!", "#NULL!", "#NULL!", "#NULL!", "#NUL…  
## $ type1617 <chr> "SCH-HIGH", "SCH-ELEM", "SCH-COMB", "SCH-HIGH", "SCH-CO…  
## $ lowgrade\_2016 <chr> "9", "PK", "KG", "9", "6", "PK", "PK", "PK", "PK", "6",…  
## $ higrade\_2016 <chr> "12", "5", "12", "12", "12", "5", "5", "5", "5", "8", "…  
## $ totalFT\_2016 <chr> "747", "255", "66", "606", "282", "482", "596", "540", …  
## $ total\_2016 <chr> "747", "255", "66", "606", "282", "482", "596", "540", …  
## $ snp\_2016 <chr> "743", "250", "45", "605", "283", "505", "615", "531", …  
## $ free\_elig\_1617 <chr> "357", "110", "15", "367", "94", "343", "414", "356", "…  
## $ free\_per\_1617 <chr> "48.05%", "44.00%", "33.33%", "60.66%", "33.22%", "67.9…  
## $ red\_elig\_1617 <chr> "39", "9", "2", "53", "5", "29", "47", "24", "39", "34"…  
## $ red\_per\_1617 <chr> "5.25%", "3.60%", "4.44%", "8.76%", "1.77%", "5.74%", "…  
## $ totalFRL\_1617 <chr> "396", "119", "17", "420", "99", "372", "461", "380", "…  
## $ totalper\_1617 <chr> "53.30%", "47.60%", "37.78%", "69.42%", "34.98%", "73.6…  
## $ CEP\_1718 <chr> "#NULL!", "#NULL!", "#NULL!", "#NULL!", "#NULL!", "#NUL…  
## $ type1718 <chr> "High", "Elementary", "Combined", "High", "Combined", "…  
## $ lowgrade\_2017 <chr> "9", "PK", "KG", "9", "6", "PK", "PK", "PK", "PK", "6",…  
## $ higrade\_2017 <chr> "12", "5", "12", "12", "12", "5", "5", "5", "5", "8", "…  
## $ totalFT\_2017 <chr> "741", "236", "60", "624", "286", "485", "583", "550", …  
## $ total\_2017 <chr> "742", "236", "60", "624", "286", "485", "583", "550", …  
## $ snp\_1718 <chr> "783", "258", "62", "657", "292", "554", "652", "576", …  
## $ free\_elig\_1718 <chr> "364", "96", "22", "364", "96", "362", "376", "364", "4…  
## $ free\_per\_1718 <chr> "46.49%", "37.21%", "35.48%", "55.40%", "32.88%", "65.3…  
## $ red\_elig\_1718 <chr> "39", "7", "2", "56", "9", "23", "45", "34", "31", "30"…  
## $ red\_per\_1718 <chr> "4.98%", "2.71%", "3.23%", "8.52%", "3.08%", "4.15%", "…  
## $ totalFRL\_1718 <chr> "403", "103", "24", "420", "105", "385", "421", "398", …  
## $ totalper\_1718 <chr> "51.47%", "39.92%", "38.71%", "63.93%", "35.96%", "69.4…  
## $ stable <chr> "1", "1", "1", "1", "1", "1", "1", "1", "1", "1", "1", …  
## $ new <chr> "0", "0", "0", "0", "0", "0", "0", "0", "0", "0", "0", …  
## $ closed <chr> "0", "0", "0", "0", "0", "0", "0", "0", "0", "0", "0", …  
## $ close\_yr <chr> "0", "0", "0", "0", "0", "0", "0", "0", "0", "0", "0", …  
## $ reuseid <chr> "0", "0", "0", "0", "0", "0", "0", "0", "0", "0", "0", …  
## $ gradechg <chr> "0", "0", "0", "0", "0", "0", "0", "0", "0", "0", "0", …  
## $ gradechg\_yr <chr> "0", "0", "0", "0", "0", "0", "0", "0", "0", "0", "0", …  
## $ grchgyr\_2 <chr> "0", "0", "0", "0", "0", "0", "0", "0", "0", "0", "0", …

### NOTE: When you have a lot of variables, running this str() function is not a great idea - the output is a little too cumbersome

## Mutating Variables

Note that almost all of the data reads in as a “character” data type which are just strings, This can create issues.

We know that many of the columns are actually storing numbers or “numeric” values as R refers to them. We need to fix this.

Let’s tell R that these columns (at least the two we are going to use) are numeric.

We are going to see two interchangeable ways to do this.

First, we use the $ operator which lets me specify a specific column within my data frame in combination with the as.numeric() function

mydata$total\_2017<-as.numeric(mydata$total\_2017)  
mydata$totalFRL\_1718<-as.numeric(mydata$totalFRL\_1718)  
  
# some columns have a percent symbol, which you will need to remove before coercing to numeric data type  
  
mydata <- mydata %>%  
 mutate(totalper\_0809 = str\_remove(totalper\_0809, pattern = "%"))  
  
  
mydata$totalper\_0809 <- as.numeric(mydata$totalper\_0809)  
  
str(mydata$totalper\_0809)

## num [1:2101] 56.1 40.1 47.5 55.5 33.4 ...

# Second, alternatively, we can do this for a whole set of variables at once.  
# We just need to specify a matching criteria.  
  
newdf <- mydata %>%   
 mutate\_at(vars(starts\_with("total")), as.numeric)  
  
newdf <- newdf %>%   
 mutate\_at(vars(starts\_with("totalFRL")), as.numeric)  
  
# Check whether the old and new variables are stored differently   
# (old as a character, new as a numeric variable)  
str(mydata$total\_2008)

## chr [1:2101] "731" "263" "80" "638" "333" "536" "610" "490" "585" "450" ...

str(newdf$total\_2008)

## num [1:2101] 731 263 80 638 333 536 610 490 585 450 ...

## Filtering, Selecting, Grouping, and Summarizing

A basic operation we do a lot is to filter the data so that we are working with a subset of all that we have.

We can do this with the filter() function, part of the dplyr package (in the tidyverse collection of packages).

# Let's say we want to look at the schools with div\_num values less than 50  
newdf %>% filter(div\_num < 50)

## # A tibble: 800 × 137  
## sch\_id div\_num div\_name schoo…¹ schoo…² schoo…³ type0…⁴ lowgr…⁵ higra…⁶  
## <chr> <dbl> <chr> <chr> <chr> <chr> <chr> <chr> <chr>   
## 1 001-0070 1 Accomack Co… "0070\… NANDUA… <NA> SCH-HI… 9 12   
## 2 001-0080 1 Accomack Co… "0080\… CHINCO… <NA> SCH-EL… PK 5   
## 3 001-0530 1 Accomack Co… "0530\… TANGIE… <NA> SCH-CO… KG 12   
## 4 001-0540 1 Accomack Co… "0540\… ARCADI… <NA> SCH-HI… 9 12   
## 5 001-0580 1 Accomack Co… "0580\… CHINCO… <NA> SCH-CO… 6 12   
## 6 001-0590 1 Accomack Co… "0590\… PUNGOT… <NA> SCH-EL… PK 5   
## 7 001-0600 1 Accomack Co… "0600\… KEGOTA… <NA> SCH-EL… PK 5   
## 8 001-0701 1 Accomack Co… "0701\… ACCAWM… <NA> SCH-EL… PK 5   
## 9 001-0702 1 Accomack Co… "0702\… METOMP… <NA> SCH-EL… PK 5   
## 10 001-0703 1 Accomack Co… "0703\… NANDUA… <NA> SCH-MID 6 8   
## # … with 790 more rows, 128 more variables: totalFT\_2008 <dbl>,  
## # total\_2008 <dbl>, snp\_0809 <chr>, free\_elig\_0809 <chr>,  
## # free\_per\_0809 <chr>, red\_elig\_0809 <chr>, red\_per\_0809 <chr>,  
## # totalFRL\_0809 <dbl>, totalper\_0809 <dbl>, type0910 <chr>,  
## # lowgrade\_2009 <chr>, higrade\_2009 <chr>, totalFT\_2009 <dbl>,  
## # total\_2009 <dbl>, snp\_0910 <chr>, free\_elig\_0910 <chr>,  
## # free\_per\_0910 <chr>, red\_elig\_0910 <chr>, red\_per\_0910 <chr>, …

# Or, if we want to look at schools where the highest grade in 2008 was grade five, we can try:  
newdf %>% filter(higrade\_2008 == "5") # this returns a subsetted dataframe with 878 rows

## # A tibble: 878 × 137  
## sch\_id div\_num div\_name schoo…¹ schoo…² schoo…³ type0…⁴ lowgr…⁵ higra…⁶  
## <chr> <dbl> <chr> <chr> <chr> <chr> <chr> <chr> <chr>   
## 1 001-0080 1 Accomack Co… "0080\… CHINCO… <NA> SCH-EL… PK 5   
## 2 001-0590 1 Accomack Co… "0590\… PUNGOT… <NA> SCH-EL… PK 5   
## 3 001-0600 1 Accomack Co… "0600\… KEGOTA… <NA> SCH-EL… PK 5   
## 4 001-0701 1 Accomack Co… "0701\… ACCAWM… <NA> SCH-EL… PK 5   
## 5 001-0702 1 Accomack Co… "0702\… METOMP… <NA> SCH-EL… PK 5   
## 6 002-0010 2 Albemarle C… "0010\… HOLLYM… <NA> SCH-EL… PK 5   
## 7 002-0030 2 Albemarle C… "0030\… SCOTTS… <NA> SCH-EL… PK 5   
## 8 002-0040 2 Albemarle C… "0040\… MARY C… <NA> SCH-EL… PK 5   
## 9 002-0100 2 Albemarle C… "0100\… BROADU… <NA> SCH-EL… PK 5   
## 10 002-0150 2 Albemarle C… "0150\… PAUL H… <NA> SCH-EL… PK 5   
## # … with 868 more rows, 128 more variables: totalFT\_2008 <dbl>,  
## # total\_2008 <dbl>, snp\_0809 <chr>, free\_elig\_0809 <chr>,  
## # free\_per\_0809 <chr>, red\_elig\_0809 <chr>, red\_per\_0809 <chr>,  
## # totalFRL\_0809 <dbl>, totalper\_0809 <dbl>, type0910 <chr>,  
## # lowgrade\_2009 <chr>, higrade\_2009 <chr>, totalFT\_2009 <dbl>,  
## # total\_2009 <dbl>, snp\_0910 <chr>, free\_elig\_0910 <chr>,  
## # free\_per\_0910 <chr>, red\_elig\_0910 <chr>, red\_per\_0910 <chr>, …

## Note that we had to set it equal to the character value "5" rather than the numeric value 5. Why?  
  
# If we wanted to filter on numeric values instead, we would want to do something like this:  
newdf %>%   
 mutate(higrade\_2008 = as.numeric(higrade\_2008)) %>%   
 filter(higrade\_2008 == 5) # again, this returns a subsetted dataframe with 878 rows

## Warning in mask$eval\_all\_mutate(quo): NAs introduced by coercion

## # A tibble: 878 × 137  
## sch\_id div\_num div\_name schoo…¹ schoo…² schoo…³ type0…⁴ lowgr…⁵ higra…⁶  
## <chr> <dbl> <chr> <chr> <chr> <chr> <chr> <chr> <dbl>  
## 1 001-0080 1 Accomack Co… "0080\… CHINCO… <NA> SCH-EL… PK 5  
## 2 001-0590 1 Accomack Co… "0590\… PUNGOT… <NA> SCH-EL… PK 5  
## 3 001-0600 1 Accomack Co… "0600\… KEGOTA… <NA> SCH-EL… PK 5  
## 4 001-0701 1 Accomack Co… "0701\… ACCAWM… <NA> SCH-EL… PK 5  
## 5 001-0702 1 Accomack Co… "0702\… METOMP… <NA> SCH-EL… PK 5  
## 6 002-0010 2 Albemarle C… "0010\… HOLLYM… <NA> SCH-EL… PK 5  
## 7 002-0030 2 Albemarle C… "0030\… SCOTTS… <NA> SCH-EL… PK 5  
## 8 002-0040 2 Albemarle C… "0040\… MARY C… <NA> SCH-EL… PK 5  
## 9 002-0100 2 Albemarle C… "0100\… BROADU… <NA> SCH-EL… PK 5  
## 10 002-0150 2 Albemarle C… "0150\… PAUL H… <NA> SCH-EL… PK 5  
## # … with 868 more rows, 128 more variables: totalFT\_2008 <dbl>,  
## # total\_2008 <dbl>, snp\_0809 <chr>, free\_elig\_0809 <chr>,  
## # free\_per\_0809 <chr>, red\_elig\_0809 <chr>, red\_per\_0809 <chr>,  
## # totalFRL\_0809 <dbl>, totalper\_0809 <dbl>, type0910 <chr>,  
## # lowgrade\_2009 <chr>, higrade\_2009 <chr>, totalFT\_2009 <dbl>,  
## # total\_2009 <dbl>, snp\_0910 <chr>, free\_elig\_0910 <chr>,  
## # free\_per\_0910 <chr>, red\_elig\_0910 <chr>, red\_per\_0910 <chr>, …

Let’s shift gears to a different combination of operations…

Let’s go ahead and try using tidyverse to narrow to what we want. Imagine we want to see the county level aggregate numbers for FRL in the 2017-2018 school year.

We will start out with our entire data frame and then use pipes (the %>% operator) to work from there. The final result will be stored in our new data frame that we are creating, called county\_level\_aggregate.

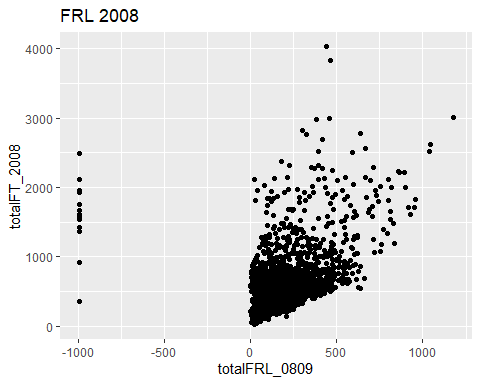
First, select will pick columns Next, group\_by and summarize work together to get us our aggregate totals.

county\_level\_aggregate <- newdf %>%   
 select(div\_name, total\_2017, totalFRL\_1718) %>%  
 group\_by(div\_name) %>%  
 summarize(totalstudents = sum(total\_2017),   
 totalFRL = sum(totalFRL\_1718))  
# now, we can compute percentages if we like and we can specify a new column by referring to  
# one that doesn't exist yet but will after we run this code.  
# We will do this two interchangeable ways.  
  
# First, the old school way:  
  
county\_level\_aggregate$percent\_FRL <- county\_level\_aggregate$totalFRL/county\_level\_aggregate$totalstudents\*100  
  
  
# Second, the tidyverse way:  
county\_level\_aggregate <- county\_level\_aggregate %>%   
 mutate(percent\_frl = totalFRL / totalstudents \* 100)  
  
  
  
# just for fun, let's see how this could have been incorporated into our summarize call  
  
county\_level\_percents <- newdf %>%   
 select(div\_name, total\_2017, totalFRL\_1718) %>%  
 group\_by(div\_name) %>%  
 summarize(percentFRL=sum(totalFRL\_1718)/sum(total\_2017) \* 100)

### Something is going to look weird with this plot

newdf %>%   
 ggplot(aes(totalFRL\_0809, totalFT\_2008)) +  
 geom\_point() +  
 labs(title = "FRL 2008", x = "totalFRL\_0809")

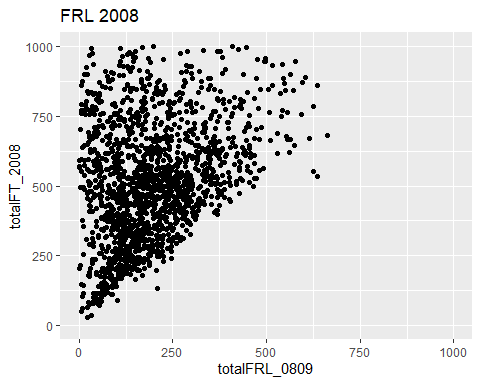
## Warning: Removed 236 rows containing missing values (`geom\_point()`).



### Let’s see if we can fix it

newdf %>%   
 filter(!is.na(totalFRL\_0809)) %>%  
 ggplot(aes(totalFRL\_0809, totalFT\_2008)) +  
 geom\_point() +  
 labs(title = "FRL 2008",   
 x = "totalFRL\_0809") +  
 xlim(0, 1000) +  
 ylim(0, 1000)

## Warning: Removed 324 rows containing missing values (`geom\_point()`).



Practice reading in .dat file from the textbook dataset files for ch. 4

narcissism\_df <- read\_delim("./data/FacebookNarcissism.dat")

## Rows: 776 Columns: 4  
## ── Column specification ────────────────────────────────────────────────────────  
## Delimiter: "\t"  
## chr (1): Rating\_Type  
## dbl (3): id, NPQC\_R\_Total, Rating  
##   
## ℹ Use `spec()` to retrieve the full column specification for this data.  
## ℹ Specify the column types or set `show\_col\_types = FALSE` to quiet this message.

Look at the dataframe

narcissism\_df

## # A tibble: 776 × 4  
## id NPQC\_R\_Total Rating\_Type Rating  
## <dbl> <dbl> <chr> <dbl>  
## 1 1 31 Attractive 2  
## 2 1 31 Fashionable 2  
## 3 1 31 Glamourous 2  
## 4 1 31 Cool 2  
## 5 2 37 Attractive 2  
## 6 2 37 Fashionable 2  
## 7 2 37 Glamourous 2  
## 8 2 37 Cool 2  
## 9 5 44.5 Attractive 3  
## 10 5 44.5 Fashionable 3  
## # … with 766 more rows

Let’s try to plot the distribution for each rating type

narcissism\_df %>%   
 ggplot(aes(x = Rating)) +  
 geom\_histogram() +  
 facet\_wrap(.~Rating\_Type) +  
 labs(y = "Count",  
 title = "Exploratory plot of Facebook Narcissism dataset")

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.

