```r
1   #install.packages("pastecs")
2   library(pastecs)
3
4   library(readr) #importing csv files
5   library(dplyr) #general analysis
6   library(ggplot2) #making charts
7   library(lubridate) #date functions
8   library(reshape2) #use this for melt function to create one record for each team
9   library(tidyr)
10  library(janitor) #use this for doing crosstabs
11  library(scales) #needed for stacked bar chart axis labels
12  library(knitr) #needed for making tables in markdown page
13  library(htmltools)#this is needed for Rstudio to display kable and other html code
14  library(rmarkdown)
15  library(kableExtra)
16  library(ggthemes)
17  library(stringr)
18  library(RMySQL)
19  library(readxl) #for importing Excel files
20  library(DT) #needed for making  searchable sortable data tble
21  library(waffle)
22  library(foreign) #for importing SPSS files
23  library(jsonlite) #for exporting JSON
24  library(car)
25  library(aws.s3) #for loading to AWS server
26
27
28  options(scipen=999)
29
30  library(scales)
31
32
33  # Import Basic Skills revenue -----------------------------------------------
34
35  basicskills <-  read_csv('./data/basicskills_revenue_import.csv',
    col_types=cols(`District Number`=col_character(), `District Type`=col_character()))%>%
36    clean_names() %>% mutate(districtid=paste(district_number, district_type, '000',
    sep='-'))
37
38
39  basicskills2 <-  basicskills %>% select(-district_number, -district_type, -district)
40
41  basicskills3 <-  melt(basicskills2, id.vars='districtid')
42
43  basicskills3 <-  basicskills3 %>% mutate(datayr=substr(variable, 2, 6),
44                                        yr=
    as.numeric(paste('20',substr(variable,5,6), sep='')),
45                                        type=substr(variable, 8,100))
46
47  # import UFARS -----------------------------------------------------------
48
49
50
51  ufars06_18 <-  read_csv('./data/ufars06_18.csv',
52                      col_types=cols(.default=col_character(),
    tot_amt=col_double()))%>% rename(datayear=dat_yer,
```

```r
                     districtnum=dst_num,

                     disttype=dst_tye,

                     fund=fun_num,

                     organization=ogz_num,

                     program=prg_num,

                     finance=fna_num,

                     object=obj_num,

                     course=crs_num,

                     schoolclass=unt_cls)


codes <-   read_excel("./data/UFARS/09-ListofCodes 2019.1.xlsx", sheet="CODES",
    range="A1:D730")



# import from mysql -----------------------------------------------------



con <- dbConnect(RMySQL::MySQL(), host = Sys.getenv("host"), dbname="Schools",user=
    Sys.getenv("userid"), password=Sys.getenv("pwd"))

#list the tables in the database we've connected to
#dbListTables(con)

#list the fields in the table; change "mytablename" to the name of the table you're
    trying to connect to
#dbListFields(con,'mytablename')


#Pull DistrictList table
data1 <- dbSendQuery(con, "select * from DistrictList")

#assign it to a new data frame
district_list <- fetch(data1, n=-1)

dbClearResult(data1)


#Pull mobility data
data2 <- dbSendQuery(con, "select schoolid, fiscalyear as yr, midyr_pct as mobility,
    schoolclassification
from mobility")

#assign it to a new data frame
mobility <- fetch(data2, n=-1)
```

```
 97   dbClearResult(data2)
 98
 99
100
101   #Pull race data
102   data3 <- dbSendQuery(con, "select schoolid, schoolyear as yr,  pctminority,
      totalstudents
103   from enroll_race where schoolyear>=2007")
104
105   #assign it to a new data frame
106   race <- fetch(data3, n=-1)
107
108   dbClearResult(data3)
109
110
111   #Pull teacher data
112   data4 <- dbSendQuery(con, "select idnumber as schoolid, concat('20', right(schoolyr,2))
      as yr, totfte,
113   newteacher, avgyrsexp
114   from teacher_demographics where distnum<>'9999'")
115
116   #assign it to a new data frame
117   teachers <- fetch(data4, n=-1)
118
119   dbClearResult(data4)
120
121
122   #Pull special enrollment data (free lunch and ELL)
123   data5 <- dbSendQuery(con, "select schoolid, concat('20', right(datayear,2)) as
      yr,k12enr, freek12, redk12, lepidentifiedk12, lepservedk12
124                       from enroll_special
125                       where grade='All Grades' and datayear not like '9%'
126                       having yr>2006")
127
128   #assign it to a new data frame
129   special <- fetch(data5, n=-1)
130
131   dbClearResult(data5)
132
133
134
135   #Pull SchoolList table
136   data6 <- dbSendQuery(con, "select * from SchoolList")
137
138   #assign it to a new data frame
139   school_list <- fetch(data6, n=-1)
140
141   dbClearResult(data6)
142
143
144
145   #disconnect connection
146   dbDisconnect(con)
147
148
149
150   # clean up mysql data -------------------------------------------------
```

```r
151
152
153  special$k12enr[is.na(special$k12enr)] <- 0
154  special$freek12[is.na(special$freek12)] <- 0
155  special$redk12[is.na(special$redk12)] <- 0
156  special$lepidentifiedk12[is.na(special$lepidentifiedk12)] <- 0
157  special$lepservedk12[is.na(special$lepservedk12)] <- 0
158
159
160  #clean up district_list data frame and add a districtid number
161  district_list <- district_list %>% clean_names() %>% rename(district_name=organization)
162
163  school_list <-  school_list %>% clean_names()
164
165
166  #change the yr variable to integer in a new field and drop the yr variable
167  race <- race %>% mutate(schoolyr=as.integer(yr)) %>% select(-yr)
168
169
170  mobility <- mobility %>% mutate(schoolyr=as.integer(yr)) %>% select(-yr)
171
172  special <- special %>% mutate(schoolyr=as.integer(yr),
173                               districtid=paste(str_sub(schoolid,1,7),'000',sep="-"))
     %>% select(-yr)
174
175  teachers <- teachers %>% mutate(schoolyr=as.integer(yr)) %>% select(-yr)
176
177  # import text files -------------------------------------------------
178
179
180
181  #open enrollment
182  openenroll <-  read_csv('./data/openenroll.csv')%>% group_by(districtid, yr) %>%
183    summarise(enroll=sum(enrolled),
184             leaving=sum(LeavingToTrad)+sum(LeavingToCharter),
185             coming=sum(ComingIn),
186             resident=sum(residents)) %>%
187    mutate(pctleving=leaving/resident,
188          pctcoming=coming/enroll)
189
190
191  #attendance
192  attend <-  read_csv('./data/consistent_attendance_northstar.csv') %>%
193    mutate(schoolid=paste(districtnumber, districttype, schoolnumber, sep="-"))
194
195
196
197  #compensatory revenue / poverty concentration
198  #this one needs districtid attached to grab location information from district_list
199  revenue <-  read_csv('./data/compensatory_revenue_bysite_06_18.csv') %>%
200    clean_names() %>%
201    mutate(schoolid=paste(district_number, district_type, site_number, sep="-"),
202          yr=as.integer(str_sub(year,4,6))+2000,
203          districtid=paste(str_sub(schoolid,1,7),'000',sep="-"))
204
205
206
```

```r
207
208
209  #mca data
210  math <-  read_csv('./data/math_scores.csv', col_types=cols(.default="c",
     totaltested=col_integer(),
211                                                  level3=col_integer(),
212                                                  level4=col_integer())) %>%
213    mutate(math_totalproficient=level3+level4,
214          math_pctproficient=math_totalproficient/totaltested,
215          yr=as.integer(str_sub(datayear,4,6))+2000) %>%
216    filter(yr>=2007)
217
218
219
220  read <-  read_csv('./data/read_scores.csv', col_types=cols(.default="c",
     totaltested=col_integer(),
221                                                  level3=col_integer(),
222                                                  level4=col_integer())) %>%
223    mutate(read_totalproficient=level3+level4,
224          read_pctproficient=read_totalproficient/totaltested,
225          yr=as.integer(str_sub(datayear,4,6))+2000) %>%
226    filter(yr>=2007)
227
228
229
230  rm(data1)
231  rm(data2)
232  rm(data3)
233  rm(data4)
234  rm(data5)
235  rm(data6)
236
237
238  # IDENTIFY SCHOOLS IN NEED ------------------------------------------------
239
240  #sets a level (1 through 4) depending on math proficiency percentage
241  #for simplicity, I cut it at quartiles, with the highest number (4) being the most in
     need
242  math <-  math %>% mutate(need_level = case_when(math_pctproficient<.25~4,
243                                          math_pctproficient>=.25 &
     math_pctproficient<.5~3,
244                                          math_pctproficient>=.5 &
     math_pctproficient<.75~2,
245                                          math_pctproficient>=.75~1))
246
247
248  #create a column that counts that number of kids below proficiency
249  #(levels 1 and 2 added together)
250  math <-  math %>% mutate(number_in_need = totaltested-math_totalproficient)
251
252  #math %>% group_by(yr) %>% summarise(in_need=sum(number_in_need))
253
254
255
256
257  # MATCH MONEY -------------------------------------------------------------
258
```

```r
259  #to simplify this, going to limit data down to 2017-18 school year
260  #And only district types of 1, 3 and 7
261
262  math2018 <-   math %>% filter(yr==2018) %>%  select(schoolid, yr, need_level,
     totaltested, number_in_need, math_pctproficient) %>%
263     mutate(dist_type=str_sub(schoolid,6,7))%>% filter(dist_type=='01' | dist_type=='03' |
     dist_type=='07')
264
265  #names(revenue)
266
267  revenue <- revenue %>% filter(yr>2006) %>%  select(schoolid, districtid,
     district_number, district_type, district_name, site_number, site_name, yr,
     fall_enrollment, free_lunch_count, reduced_lunch_count,
268                                           adjusted_count, concentration,
     factor, pupil_units,
269                                           revenue_per_adjusted_count, revenue)
     %>%
270     rename(students_yr_prior=fall_enrollment)
271
272
273
274
275  df <-   left_join(revenue, school_list %>%
276                    select(school_id, metro7county, location, school_name,
     school_location_county_name, classification, grades, school_type),
     by=c("schoolid"="school_id"))
277
278
279  #names(df)
280
281  revenue2018 <- df %>% filter(yr==2018, district_type=='01' | district_type=='03' |
     district_type=='07') %>%
282     mutate(grades2= str_trim(grades))
283
284
285  #organization code 005 = districtwide spending
286
287  ufars06_18 <- ufars06_18 %>%
288     filter(finance=='317') %>%
289     mutate(schoolid=paste(districtnum, disttype, organization, sep="-"),
290                           yr=as.integer(str_sub(datayear,4,6))+2000,

     districtid=paste(str_sub(schoolid,1,7),'000',sep="-"))
291
292
293
294
295
296  ufars06_18 <-   left_join(ufars06_18, district_list %>% select(district_number,
     district_type, district_name), by=c("districtnum"="district_number",
     "disttype"="district_type"))
297
298
299
300
301  program_codes <-   codes %>% filter(top_group=='Program')
302
```

```r
303  ufars06_18 <-   left_join(ufars06_18, program_codes %>% select(code, detail, sub_group),
     by=c("program"="code"))
304
305
306
307
308  ufars2018 <-   ufars06_18 %>%
309    filter(datayear=='17-18', disttype=='01' | disttype=='03' | disttype=='07')%>%
310    group_by(yr, schoolid, disttype, districtid) %>%
311    summarise(tot_spent = sum(tot_amt))
312
313
314
315
316  match2018 <- left_join(revenue2018, ufars2018 %>% select(schoolid, tot_spent),
     by=c("schoolid"="schoolid"))
317
318  match2018 <- left_join(match2018, math2018, by=c("schoolid"="schoolid"))
319
320  match2018 %>% group_by(need_level) %>% summarise(tot_revenue =
     sum(revenue),adjusted_rev=sum(revenue_per_adjusted_count), total_spent=sum(tot_spent),
     count=n(), pupils=sum(adjusted_count))
321
322  match2018$tot_spent[is.na(match2018$tot_spent)] <- 0
323  match2018$revenue[is.na(match2018$revenue)] <- 0
324
325
326  match2018 <-   match2018 %>% mutate(poverty_level = case_when(concentration>=.8~'very
     high',
327                                          concentration>=.6 &
     concentration<.8~'high',
328                                          concentration>=.4 &
     concentration<.6~'medium',
329                                          concentration>=.2 &
     concentration<.4~'low',
330                                          concentration<.2~'very
     low'))
331
332
333  #match2018 %>% filter(tot_spent>0) %>% group_by(poverty_level) %>%
334  #  summarise(tot_revenue = sum(revenue),adjusted_rev=sum(revenue_per_adjusted_count),
     total_spent=sum(tot_spent), count=n(), pupils=sum(adjusted_count))
335
336
337  match2018 <-   match2018 %>% mutate(rev_per_need = if_else(is.na(revenue), 0,
     round_half_up(as.double(revenue/adjusted_count))),
338                               spend_per_need = if_else(is.na(tot_spent), 0,
     round_half_up(as.double(tot_spent/adjusted_count))))
339
340
341
342
343  special_district <-   special %>% group_by(districtid) %>%
     summarise(tot_enroll=sum(k12enr), totfreek12= sum(freek12), totredk12= sum(redk12)) %>%
     mutate(freelunch=totfreek12+totredk12, pctfreelunch = freelunch/tot_enroll,
     poverty_level = case_when(pctfreelunch>=.8~'very high',
```

```r
  pctfreelunch>=.6 & pctfreelunch<.8~'high',



  pctfreelunch>=.4 & pctfreelunch<.6~'medium',



  pctfreelunch>=.2 & pctfreelunch<.4~'low',



  pctfreelunch<.2~'very low'))

# DISTRICT LEVEL ANALYSIS ----------------------------------------------

dist_spent <-   ufars06_18 %>% filter(disttype=='01' | disttype=='03') %>%
  group_by(yr, districtnum, disttype) %>% summarise(spent= sum(tot_amt))

dist_rev <-   revenue %>%   filter(district_type=='01' | district_type=='03') %>%
  group_by(yr, district_number, district_type, district_name) %>%
summarise(rev=sum(revenue), poverty=sum(adjusted_count))


dist_match_allyrs <-   left_join(dist_rev, dist_spent, by=c("yr"="yr",
"district_number"="districtnum", "district_type"="disttype"))

dist_match_allyrs$rev[is.na(dist_match_allyrs$rev)] <- 0
dist_match_allyrs$spent[is.na(dist_match_allyrs$spent)] <- 0

dist_match_allyrs <- dist_match_allyrs%>% mutate(diff=round(spent-rev,2),
pct=round((diff/rev)*100,1))


dist_match_allyrs <- dist_match_allyrs %>%
  mutate(scope = case_when(pct>=14.49~'over by 15% or more',
                           pct<14.49 & pct>9.49~'over by 10%-14%',
                           pct<=9.49 & pct>0 ~'over by less than 10%',
                           pct==0 ~'even',
                           pct<0 & pct> -9.49 ~'under by less than 10%',
                           pct> -14.49 & pct< -9.49~'under by 10-14%%',
                           pct<= -14.49~'under by 15% or more',
                           TRUE~'something went wrong'),
         districtid=paste(district_number, district_type, '000', sep="-"))


dist_match_allyrs <-   left_join(dist_match_allyrs, special_district %>%
select(districtid, tot_enroll, pctfreelunch, poverty_level),
by=c("districtid"="districtid"))


#add size of district
dist_match_allyrs <-   dist_match_allyrs %>%
  mutate(size = case_when(tot_enroll>24735~4,
```

```r
                           tot_enroll<=24735 & tot_enroll>10915~3,
                           tot_enroll<=10915 & tot_enroll>5657~2,
                           tot_enroll<=5657~1))

#use this to figure out the break points in the above case_when
#quantile(dist_match_allyrs$tot_enroll)

#where located-- by location
dist_match_allyrs <-  left_join(dist_match_allyrs, district_list %>% select(id_number,
location), by=c("districtid"="id_number"))
```