```r
#This script is for working with the full Basic Skills revenue and expenditure data
#this encompasses both compensatory revenue and English Learner revenue
#We received data on the revenue side (the "basicskills" file) on April 18, 2019 from MDE
#The UFARS file we had received earlier and that shows all the spending (program code 219 is spending for English Learners)




#install.packages("pastecs")
library(pastecs)

library(readr) #importing csv files
library(dplyr) #general analysis
library(ggplot2) #making charts
library(lubridate) #date functions
library(reshape2) #use this for melt function to create one record for each team
library(tidyr)
library(janitor) #use this for doing crosstabs
library(scales) #needed for stacked bar chart axis labels
library(knitr) #needed for making tables in markdown page
library(htmltools)#this is needed for Rstudio to display kable and other html code
library(rmarkdown)
library(kableExtra)
library(ggthemes)
library(stringr)
library(RMySQL)
library(readxl) #for importing Excel files
library(DT) #needed for making  searchable sortable data tble
library(waffle)
library(foreign) #for importing SPSS files
library(jsonlite) #for exporting JSON
library(car)
library(aws.s3) #for loading to AWS server
options(scipen=999)
library(scales)


# traditional districts ---------------------------------------------
#this is the primary file MDE gave us showing total Basic Skills amounts for each district
#But it only includes the traditional districts (types 1 and 3)

#it's set up really wide with all the years going across, with multiple values for each year
#so this next series of code pulls it in and then rearranges it
basicskills <-  read_csv('./data/basicskills_revenue_import.csv', col_types=cols(`District Number`=col_character(), `District Type`=col_character()))%>%
  clean_names() %>% mutate(districtid=paste(district_number, district_type, '000', sep='-'))

#leave out a couple columns we don't need
basicskills2 <-  basicskills %>% select(-district_number, -district_type)

#normalize the data using melt() function
```

```r
basicskills3 <-  melt(basicskills2, id=c("districtid", "district"))

#add some new columns
basicskills3 <-  basicskills3 %>% mutate(datayr=substr(variable, 2, 6),
                                         yr=
as.numeric(paste('20',substr(variable,5,6), sep='')),
                                         type=substr(variable, 8,100))



#this turns it back into wide table
#also eliminates districts that had $0 basic skills revenue in a given year
#these appear to all be districts that consolidated or were somehow closed
#but had money in earlier years
basicskills_final <- dcast(basicskills3 %>%
                             select(districtid, district, yr, type, value), districtid
+ yr + district ~ type) %>%
  filter(total_basic_skills_revenue>0)


#MDE failed to include in that file some of the "extra" compensatory
#money that goes to some districts
#pilot money is for a handful of suburban districts (plus Rochester)
#The one-time is for only one year in the past
#early learning is kind of sporadic

#additional compensatory -- pilot, one-time and early learning
additional_comp <-  read_xlsx('./data/Additional district level compensatory
data.xlsx', sheet='Compensatory Revenue Values', range='B4:K6611') %>%
  clean_names() %>% mutate(yr=as.integer(str_sub(year,4,6))+2000,
districtid=paste(district_number_type, '000', sep='-'))


#merge the additional comp fields into basicskills_final
basicskills_final <-  left_join(basicskills_final, additional_comp %>%
                                  select(districtid, yr,  compensatory_one_time,
x1st_year_vpk_srp_compensatory, compensatory_pilot, total_compensatory),
                                by=c("districtid"="districtid", "yr"="yr"))


#get rid of null values

basicskills_final$el_revenue[is.na(basicskills_final$el_revenue)] <-  0
basicskills_final$el_concentration_revenue[is.na(basicskills_final$el_concentration_rev
enue)] <-  0
basicskills_final$total_basic_skills_revenue[is.na(basicskills_final$total_basic_skills
_revenue)] <-  0
basicskills_final$total_compensatory_revenue[is.na(basicskills_final$total_compensatory
_revenue)] <-  0
basicskills_final$compensatory_one_time[is.na(basicskills_final$compensatory_one_time)]
<-  0
basicskills_final$x1st_year_vpk_srp_compensatory[is.na(basicskills_final$x1st_year_vpk_
srp_compensatory)] <-  0
basicskills_final$compensatory_pilot[is.na(basicskills_final$compensatory_pilot)] <-  0
basicskills_final$total_compensatory[is.na(basicskills_final$total_compensatory)] <-  0

#the column called total_compensatory_revenue is missing the pilot money
```

```r
#and some others
#so use the one called total_compensatory


# charter schools ------------------------------------------------------



#Now we need to pull compensatory revenue for charter schools

#this is revenue by building for all district types (the first file that MDE sent us)
#this may not include the pilot money (not sure)
#need to use this to pull charter school compensatory revenue for the online chart

revenue <-  read_csv('./data/compensatory_revenue_bysite_06_18.csv') %>%
  clean_names() %>%
  mutate(schoolid=paste(district_number, district_type, site_number, sep="-"),
         yr=as.integer(str_sub(year,4,6))+2000,
         districtid=paste(str_sub(schoolid,1,7),'000',sep="-"))



#grab only the district type 2 and 7 compensatory revenue - summarized to district level
comp_rev_charters <-  revenue %>%
  filter(district_type=='02' | district_type=='07') %>%
  group_by(districtid, district_name, yr) %>%
  summarize(comp_rev_total = sum(revenue))


#this is EL revenue for charter schools - district types 2 and 7
#sent by MDE as a separate file
el_rev_charters <-  read_csv('./data/LEPTypes2and7.csv',
col_types=cols(dst_num=col_character(), dst_tye=col_character(),

 lep_rev=col_double(), lep_cnc_rev=col_double())) %>%
  clean_names() %>%
  mutate(districtid = paste(dst_num, dst_tye, '000', sep="-"),
         total_el = lep_rev+lep_cnc_rev,
         yr=as.integer(str_sub(dat_yer,4,6))+2000)


#create file that has basic skills revenue totals for charter schools
charters_rev <-  left_join(comp_rev_charters, el_rev_charters, by=c("yr"="yr",
"districtid"="districtid") ) %>%
  mutate(basicskills_total = total_el + comp_rev_total)



# for online ------------------------------------------------------------



#This next set of code creates a file for a table to go with the story online
#it has all districts for the 17-18 school year
#showing compensatory revenue, EL revenue and a total basic skills amount
```

```r

#first need to pull the right fields for the traditional districts from basicskills_final

foronline <-  basicskills_final %>% filter(yr==2018) %>%
  mutate(district=toupper(district), el_total = el_revenue+el_concentration_revenue,
         pilot = case_when(compensatory_pilot>0~'y', TRUE~'n'),
         basicskills_total = total_compensatory+el_total) %>%
  select(districtid, district, pilot, total_compensatory, el_total, basicskills_total)


#next pull out the fields we need for charter schools
charters_foronline <-  charters_rev %>%
  filter(yr==2018) %>%
  mutate(district=toupper(district_name), pilot='n') %>%
  rename(total_compensatory=comp_rev_total,
         el_total=total_el)
select(districtid, district,comp_rev_total, total_el, basicskills_total)



#append the traditional schools file and charter schools file together using bind_rows()
foronline <-  bind_rows(foronline, charters_foronline)

#spit out a csv file to use in DataWrapper
write.csv(foronline, './output/district_totals_2018_foronline.csv', row.names=FALSE)




# ANALYSIS ----------------------------------------------------------------


#this shows the total basic skills money going to charters each year
charters_rev %>%
  filter(dst_tye=='07') %>%
  group_by(yr) %>% summarise(count=n(), tot= sum(basicskills_total), el = sum(total_el), comp=sum(comp_rev_total))







# import UFARS spending data ----------------------------------------------------------------
#this shows how Basic Skills money was spent for 2005-06 through 2017-18

ufars06_18 <-  read_csv('./data/ufars06_18.csv',
                        col_types=cols(.default=col_character(), tot_amt=col_double()))%>%
  rename(datayear=dat_yer,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)

```

```r
#this is the list of codes (this gets used in the richfield.rmd)
codes <-  read_excel("./data/UFARS/09-ListofCodes 2019.1.xlsx", sheet="CODES", range="A1:D730")


# import from mysql ----------------------------------------------------
#this imports names and other info on the districts


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)


#disconnect connection
dbDisconnect(con)
rm(data1)


#clean up district_list data frame
district_list <- district_list %>% clean_names() %>% rename(district_name=organization)


#add some fields to the ufars data
#need to limit to finance code 317 (basic skills) because they accidentally gave us some bad records
ufars06_18 <- ufars06_18 %>%
  filter(finance=='317') %>%
  mutate(schoolid=paste(districtnum, disttype, organization, sep="-"),
         yr=as.integer(str_sub(datayear,4,6))+2000,
         districtid=paste(str_sub(schoolid,1,7),'000',sep="-"))


#split UFARS into two files
#note we're excluding the 2005-06 data from UFARS cause we don't have matching revenue data
#first one is for English Learner spending under Basic Skills
el_spent <-  ufars06_18 %>% filter(disttype=='01' | disttype=='03', program=='219', yr>2006) %>%
  group_by(yr, districtid) %>% summarise(el_spent= sum(tot_amt))

#second one is for all non-EL spending under Basic Skills
```

```r
comp_spent <-  ufars06_18 %>% filter(disttype=='01' | disttype=='03', program!='219', yr>2006) %>%
  group_by(yr, districtid) %>% summarise(comp_spent= sum(tot_amt))
```

```r
#Start putting everything together
```

```r
#Join what we have so far with compensatory spending
match <-  full_join(basicskills_final, comp_spent,
                    by=c("districtid"="districtid", "yr"="yr"))
#join with English learner spending
match <- full_join(match, el_spent,
                   by=c("districtid"="districtid", "yr"="yr"))
```

```r
#add districtname and other info about the district
#filter out district without a name (MENTOR - no money either)
compare_districts <-  left_join(match, district_list %>% select(id_number, district_name, county, metro7county, location), by=c("districtid"="id_number")) %>%
  filter(district_name!='NA')
```

```r
#fill in null values
compare_districts$total_compensatory_revenue[is.na(compare_districts$total_compensatory_revenue)] <- 0
compare_districts$el_revenue[is.na(compare_districts$el_revenue)] <- 0
compare_districts$el_concentration_revenue[is.na(compare_districts$el_concentration_revenue)] <- 0
compare_districts$comp_spent[is.na(compare_districts$comp_spent)] <- 0
compare_districts$el_spent[is.na(compare_districts$el_spent)] <- 0
```

```r
#add fields

#calculate total spent
#different between spending and revenue
#pct spent is the percentage of revenue that was spent
compare_districts <-  compare_districts %>%
  mutate(tot_spent = comp_spent+el_spent,
          diff=round(total_basic_skills_revenue-tot_spent,2),
         pctspent = (tot_spent/total_basic_skills_revenue)*100)
```

```r
#add a column that puts that pct spending into buckets
compare_districts <-  compare_districts %>%
  mutate(scope = case_when(pctspent==0 ~ 'none',
                           pctspent==100~'100%',
                           pctspent>100 ~ 'over spent',
```

```
                          pctspent<100  ~'under spent',
                          TRUE ~'check'))

#calculate the difference in spending as a percentage of revenue
#then put them into buckets
#anything that is 15% or more over or under are potentially problematic (according to MDE)
compare_districts <- compare_districts %>%  mutate(diffpct=round((diff/total_basic_skills_revenue)*100,1),
                                                   diffscope = case_when(diffpct>=14.49~'over by 15% or more',
                                                                         diffpct<14.49 & diffpct>9.49~'over by 10%-14%',
                                                                         diffpct<=9.49 & diffpct>0 ~'over by less than 10%',
                                                                         diffpct==0 ~'even',
                                                                         diffpct<0 & diffpct> -9.49 ~'under by less than 10%',
                                                                         diffpct> -14.49 & diffpct< -9.49~'under by 10-14%%',
                                                                         diffpct<= -14.49~'under by 15% or more',
                                                                         TRUE~'something went wrong'))


#Count up how many fell in each bucket in 2018
compare_districts %>% filter(yr==2018) %>% group_by(diffscope) %>% summarise(count=n())


write.csv(compare_districts, './output/districts_basicskills_totals.csv', row.names=FALSE)
```