DC Data Cleaning Record

Max Anthenelli and Jiseung Yoo

2024-04-22

Main Data Source comes from Excel sheet

```
excel_file <- file.path( box_directory, "Data", "Raw",
                         "Student Data through SY22-23 Updated.xlsx")
demog_student.year <-
  read_excel(excel_file, sheet = "Demographic")

enrollment_student.school.year <-
  read_excel(excel_file, sheet = "Enrollment and Attendance")

courses_student.school.year <-
  map(map_vec(c(19:22), ~glue("Courses SY{.x}-{.x+1}")),
      ~excel_file %>% read_excel(sheet = .x)) %>%
  list_rbind()

tests_student.year <-
  read_excel(excel_file, sheet = "PARCC")

assessments_student.year <-
  read_excel(excel_file, sheet = "Formative Assessments") %>%
  filter(Assessment_Type %in% c("DIBELS", "RI", "TRC"))
```

```
english_course_titles <- c(glue("English {6:8}"), glue("Advanced English {6:8}"),
  glue("Pre-AP English {6:8}"), glue("English FT {6:8}"),
  glue("English & Humanities {6:8}"),
  "English I", "Language, Culture and Literacy", glue("Language Arts {6:8}"),
  "English as a Second Language I", "English as a Second Language II",
  "IB MYP English II") # does Journalism count??

supplemental_english_titles <- c(
  glue("Reading Resource MS{6:8}"),
  glue("Reading Workshop {c(6:8, 'MS')}"),
  "Reading Support MS",
  "Newc Engl Lit Devt MS", "Newc Oral LangDevt MS",
  "Beginning ESL MS", "Intermed ESL MS", "Advanced ESL MS",
  "Extended Literacy MS",
  "Reading Lab", "LL: Miixed-Model Reading MS7",
  glue("LL: Mixed-Model Reading MS{6:8}"),
  glue("AVID Grade {6:8}") # this isn't specifically literacy but I found their website
)
```

```r
unrelated_subject_codes <-
  c("ADM", "ART", "CARR", "CTE", "EE", "FL", "MAT", "MU", "NULL", NULL, "PE",
    "SCI", "SS", "WL")

stari_schools <-
  c(
    'Brookland MS' = 347,
    'Browne EC' = 404,
    'Cardozo EC' = 454,
    'Deal MS' = 405,
    'Eliot-Hine MS' = 407,
    'Hart MS' = 413,
    'Johnson' = 416,
    'Mckinley' = 435,
    'Stuart-Hobson MS' = 428,
    'Wells MS' = 1071
  )
stari_schools_tibble <- tibble(
  School = names(stari_schools),
  schoolid1 = stari_schools
)
```

```r
STARI <-
  file.path(box_directory, "Data", "Raw", "STARI") %>%
  list.files(pattern = "\\.xlsx$", full.names = TRUE) %>%
  map(~{read_excel(.x) %>% as_tibble()})
```

```
## New names:
## * `` -> `...4`
## * `` -> `...5`
```

```r
binder <-
  file.path(box_directory, "Data", "Raw", "STARI", "binder_info.xlsx") %>%
  read_excel()

given_stari_teachers <-
  file.path(box_directory, "Data", "Raw", "STARI", "STARI Teachers.xlsx") %>%
  read_excel() %>%
  filter(Grade != "9th-12th") %>%
  left_join(stari_schools_tibble, join_by(School))

binder_and_given_teachers <-
  split_names(binder, "Teacher", order = "FML") %>%
  full_join(split_names(given_stari_teachers, "Teacher", order = "FML"),
            join_by(first_name, last_name, other_names, schoolid1)) %>%
  select(first_name, last_name, other_names, schoolid1) %>%
  mutate(
      last_name = ifelse(last_name == 'holloway', 'holloway-mcclendon', last_name),
      first_name = ifelse(first_name == 'ayeisha', 'ayeesha', first_name))

## Celestine Holloway-Mcclendon vs Holloway
## Ayeesha vs Ayeisha Louis
# binder_and_given_teachers
```

```
stari_teachers <-
  split_names(courses_student.school.year %>%
              filter(courses_sy_start == 2022,
                     Grade %in% c('6', '7', '8'),
                     School %in% stari_schools) %>%
              select(Teacher, Employee_Number, School),
            "Teacher", order = "LFM") %>%
  inner_join(binder_and_given_teachers,
    join_by(first_name, last_name, other_names, School == schoolid1)) %>%
  distinct() %>%
  pull(Employee_Number)
```

```
## Warning: There were 327 warnings in ‘mutate()‘.
## The first warning was:
## i In argument: ‘name_split = ‘%>%‘(...)‘.
## Caused by warning in ‘.f()‘:
## ! no spaces in value(s) of name_var
## i Run ‘dplyr::last_dplyr_warnings()‘ to see the 326 remaining warnings.
```

```
## Warning in inner_join(., binder_and_given_teachers, join_by(first_name, : Detected an unexpected man
## i Row 193 of ‘x‘ matches multiple rows in ‘y‘.
## i Row 19 of ‘y‘ matches multiple rows in ‘x‘.
## i If a many-to-many relationship is expected, set ‘relationship =
##   "many-to-many"‘ to silence this warning.
```

```
# stari_teachers
```

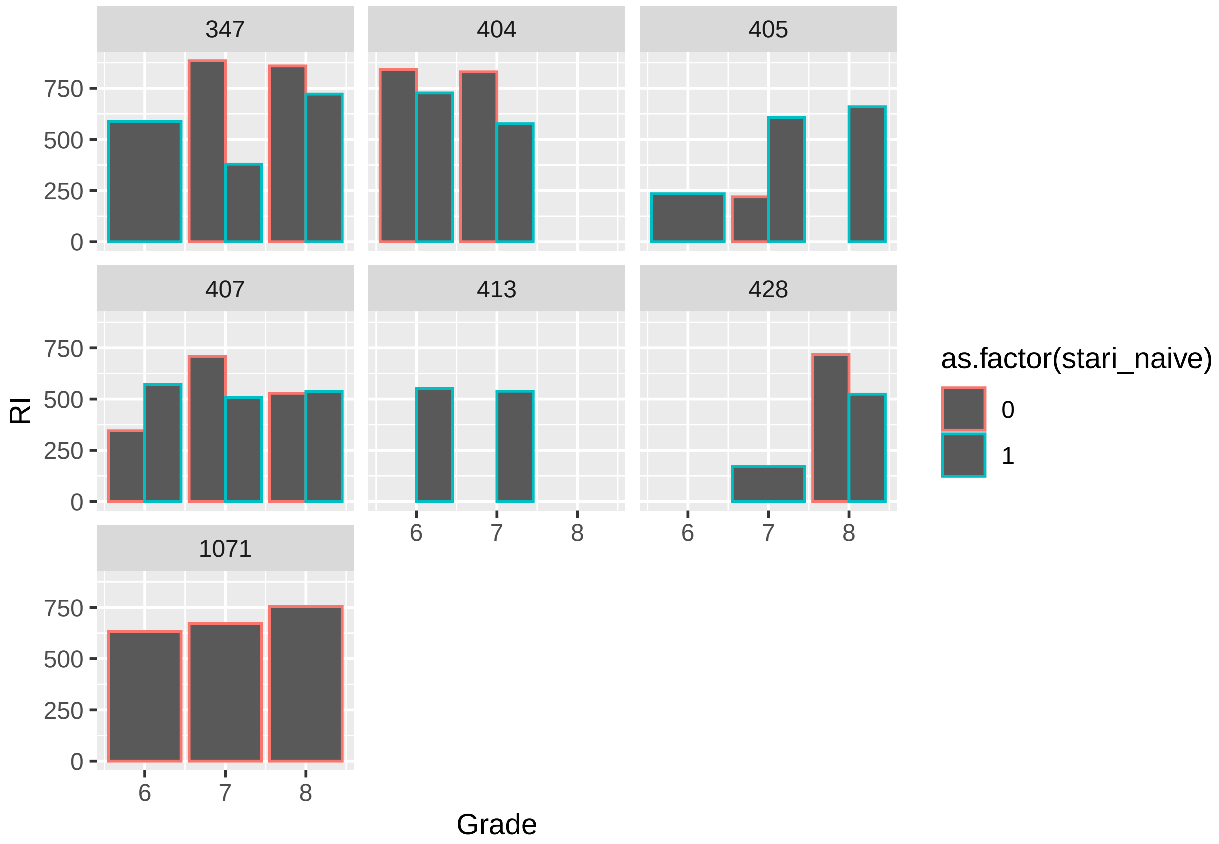
```
stari_students_naive <-
  map(4:9,~STARI[[.x]]) %>%
  list_rbind() %>%
  pull(USER_NAME)

stari_stdents_school_match <-
  map(4:9,~STARI[[.x]]) %>%
  list_rbind() %>%
  mutate(SCHOOL_NAME = case_match(SCHOOL_NAME,
    "Elliot-Hine MSS" ~ "Eliot-Hine MS",
    "Ida B Wells MSS" ~ "Wells MS",
    "Hart MSS_1" ~ "Hart MS",
    .default = SCHOOL_NAME
       )) %>%
  inner_join(stari_schools_tibble, join_by(SCHOOL_NAME == School)) %>%
  select(USER_NAME, schoolid1, GRADE) %>%
  distinct()
```

```
raw_test_score <-
  assessments_student.year %>%
  filter(School_Year_Start == 2022, Grade %in% 6:8) %>%
  select(StudentID, Grade, s_sri_ss_f, s_sri_ss_m, s_sri_ss_s)
```

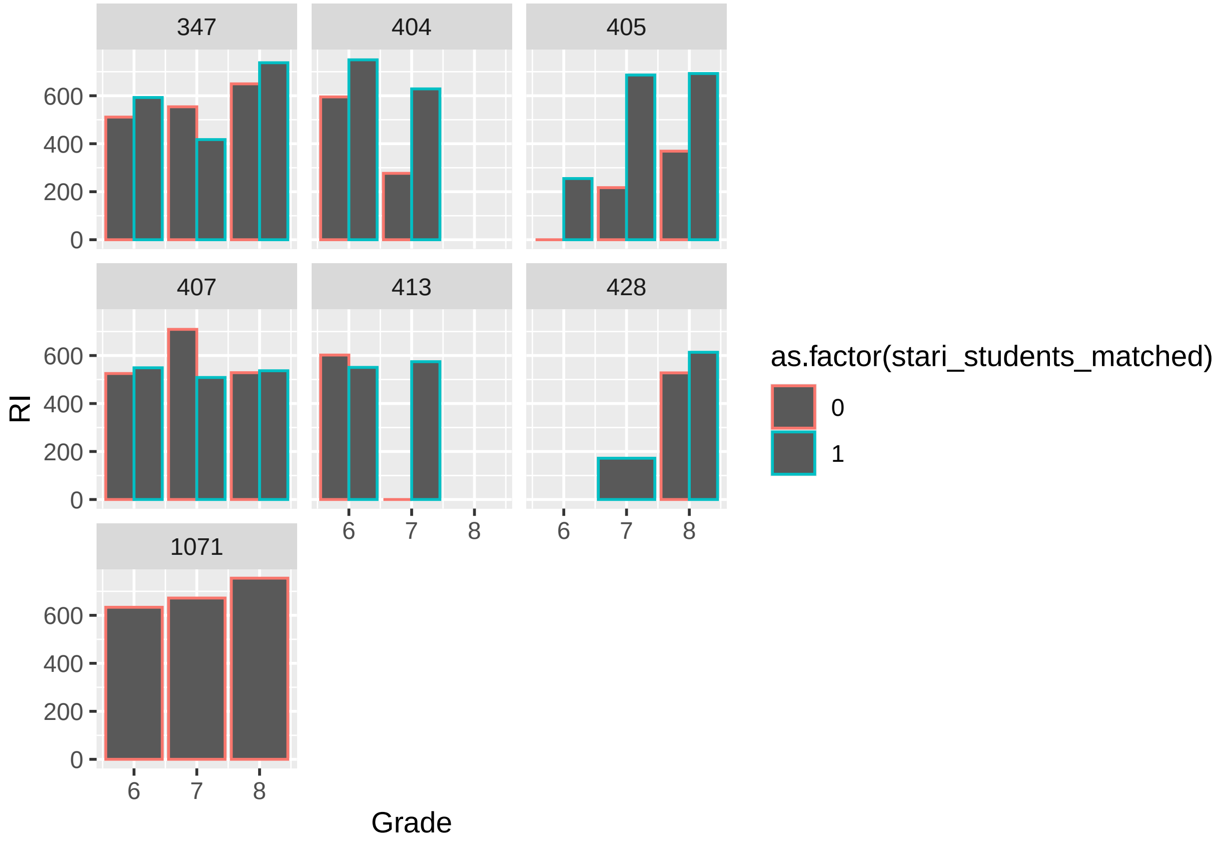
```r
course_df <-
  courses_student.school.year %>%
  filter(courses_sy_start == 2022, Grade %in% c('6', '7', '8'),
         School %in% stari_schools,
         Employee_Number %in% stari_teachers) %>%
  mutate(Grade = parse_number(Grade)) %>%
  group_by(StudentID, Title, Section, Grade, Term_Code) %>%
  distinct() %>%
  ungroup()

stari_df <-
  course_df %>%
  left_join(stari_stdents_school_match %>%  mutate(stari_students_matched = 1),
            join_by(School == schoolid1, StudentID == USER_NAME, Grade == GRADE)) %>%
  full_join(course_df %>%
  mutate(stari_naive = ifelse(StudentID %in% stari_students_naive, 1, 0)),
  by = names(course_df) , relationship = "one-to-one") %>%
  mutate(stari_students_matched = ifelse(is.na(stari_students_matched), 0, stari_students_matched)) %>%
  left_join(raw_test_score, join_by(StudentID, Grade))
```

```r
stari_df %>%
  filter(Subject_Code == "ERL", !Title %in% english_course_titles) %>%
  group_by(Title, Section, Term_Code, Grade, School, Employee_Number, stari_naive) %>%
  reframe(RI = mean(s_sri_ss_f, na.rm = TRUE),
          n= n()) %>%
  group_by(Grade, stari_naive, School) %>%
  reframe(RI = mean(RI, na.rm = TRUE),
          n = mean(n)) %>%
  ggplot(aes(x = Grade, y = RI, color = as.factor(stari_naive))) +
  geom_col(position = "dodge") +
  facet_wrap(~School)
```

```
## Warning: Removed 3 rows containing missing values or values outside the scale range
## (‘geom_col()‘).
```



```r
stari_df %>%
  filter(Subject_Code == "ERL", !Title %in% english_course_titles) %>%
  group_by(Title, Section, Term_Code, Grade, School, Employee_Number, stari_students_matched) %>%
  reframe(RI = mean(s_sri_ss_f, na.rm = TRUE),
          n= n()) %>%
  group_by(Grade, stari_students_matched, School) %>%
  reframe(RI = mean(RI, na.rm = TRUE),
          n = mean(n)) %>%
  ggplot(aes(x = Grade, y = RI, color = as.factor(stari_students_matched))) +
  geom_col(position = "dodge") +
  facet_wrap(~School)
```



```
stari_df %>%
  filter(Subject_Code == "ERL", !Title %in% english_course_titles) %>%
  group_by(Title, Section, Term_Code, Grade, School, Employee_Number) %>%
  count(stari_naive) %>%
  mutate(naive_percent = n/sum(n)) %>%
  mutate(naive_percent = ifelse(stari_naive == 0, 0, naive_percent)) %>%
  reframe( n = sum(n), naive_percent = max(naive_percent)) %>%
  ungroup() %>%
full_join(
stari_df %>%
  filter(Subject_Code == "ERL", !Title %in% english_course_titles) %>%
  group_by(Title, Section, Term_Code, Grade, School, Employee_Number) %>%
  count(stari_students_matched) %>%
  mutate(matched_percent = n/sum(n)) %>%
  mutate(matched_percent = ifelse(stari_students_matched == 0, 0, matched_percent)) %>%
  reframe(n = sum(n), matched_percent = max(matched_percent)) %>%
  ungroup(),
join_by(Title, Section, Term_Code, Grade, School, Employee_Number, n)) %>%
  gt()
```

Title	Section	Term_Code	Grade	School	Employee_Number	n	naive_percent	mat
Extended Literacy MS	6A	S1	6	347	46306	15	1.0000000	
Extended Literacy MS	7E RI	S2	7	347	118770	21	0.9523810	
Extended Literacy MS	8A	S2	8	347	114757	19	0.0000000	

Extended Literacy MS	8B	S1	7	347	114757	1	0.0000000
Extended Literacy MS	8B	S1	8	347	114757	21	0.0000000
Extended Literacy MS	Ext Lit 8	S1	8	347	46306	10	1.0000000
Extended Literacy MS	Ext.Lit8S2	S2	8	347	46306	10	1.0000000
Reading Resource MS7	7 - SLS	FY	7	405	124448	1	0.0000000
Reading Resource MS7	7 - SLS	FY	7	405	87259	7	0.0000000
Reading Support MS	45505	FY	8	405	50591	8	0.8750000
Reading Support MS	45506	FY	8	405	50591	7	1.0000000
Reading Support MS	6	FY	6	405	50591	6	1.0000000
Reading Support MS	6A	S2	6	404	82271	1	0.0000000
Reading Support MS	6B	S2	6	404	82271	24	0.9166667
Reading Support MS	7	FY	7	405	50591	12	1.0000000
Reading Support MS	7A	S2	7	347	46306	1	0.0000000
Reading Support MS	DW.SP7	FY	7	347	46306	6	1.0000000
Reading Support MS	Pineda	FY	6	1071	114578	8	0.0000000
Reading Support MS	Pineda	FY	7	1071	114578	2	0.0000000
Reading Support MS	Pineda	FY	8	1071	114578	1	0.0000000
Reading Support MS	RDG Suprt8	S1	8	347	46306	3	1.0000000
Reading Support MS	RS-8-DW	S1	8	347	46306	5	1.0000000
Reading Support MS	RS-8-DW2	S2	8	347	46306	5	1.0000000
Reading Support MS	RS-8B-DW	S2	8	347	46306	3	1.0000000
Reading Workshop 6	2-TThF2.TA	FY	6	1071	69905	28	0.0000000
Reading Workshop 6	2-TThF2.TC	FY	6	1071	102000	21	0.0000000
Reading Workshop 6	3-TThF2.TA	FY	6	1071	69905	26	0.0000000
Reading Workshop 6	3-TThF2.TC	FY	6	1071	102000	29	0.0000000
Reading Workshop 6	4-TThF2.TA	FY	6	1071	69905	27	0.0000000
Reading Workshop 6	4-TThF2.TC	FY	6	1071	102000	27	0.0000000
Reading Workshop 6	6	FY	6	405	50591	6	1.0000000
Reading Workshop 6	61	FY	6	407	109407	9	0.8888889
Reading Workshop 6	6A	S1	6	404	82271	14	1.0000000
Reading Workshop 6	6B	S1	6	404	82271	23	0.9130435
Reading Workshop 6	BOEING	FY	6	413	118489	18	1.0000000
Reading Workshop 6	BOMBERS	FY	6	413	118489	22	0.9090909
Reading Workshop 6	FIGHTERS	FY	6	413	118489	20	1.0000000
Reading Workshop 6	REDTAILS	FY	6	413	118489	20	0.8500000
Reading Workshop 6	STEALTH	FY	6	413	118489	22	1.0000000
Reading Workshop 7	2-TThF2.TA	FY	7	1071	95557	29	0.0000000
Reading Workshop 7	2-TThF2.TE	FY	7	1071	118306	26	0.0000000
Reading Workshop 7	3-TThF2.TA	FY	7	1071	95557	27	0.0000000
Reading Workshop 7	3-TThF2.TE	FY	7	1071	118306	28	0.0000000
Reading Workshop 7	4-TThF2.TA	FY	7	1071	95557	24	0.0000000
Reading Workshop 7	4-TThF2.TE	FY	7	1071	118306	22	0.0000000
Reading Workshop 7	71	FY	7	407	62597	7	1.0000000
Reading Workshop 7	72	FY	7	407	109407	12	0.9166667
Reading Workshop 7	7B	S1	7	404	82271	18	0.8888889
Reading Workshop 7	7C	S1	7	404	82271	7	1.0000000
Reading Workshop 7	REDTAILS	FY	7	413	78137	22	0.9545455
Reading Workshop 8	2-MWF1.TV	FY	8	1071	124366	31	0.0000000
Reading Workshop 8	2-TThF2.TO	FY	8	1071	123909	20	0.0000000
Reading Workshop 8	3-MWF1.TV	FY	8	1071	124366	22	0.0000000
Reading Workshop 8	3-TThF2.TO	FY	8	1071	123909	24	0.0000000
Reading Workshop 8	4-MWF1.TV	FY	8	1071	124366	27	0.0000000
Reading Workshop 8	4-TThF2.TO	FY	8	1071	123909	20	0.0000000
Reading Workshop 8	8001	T1	8	428	92796	6	0.8333333

Reading Workshop 8	8002	T2	7	428	92796	1	1.0000000
Reading Workshop 8	8002	T2	8	428	92796	6	0.8333333
Reading Workshop 8	8003	T3	7	428	92796	1	1.0000000
Reading Workshop 8	8003	T3	8	428	92796	6	0.8333333
Reading Workshop 8	8004	T4	7	428	92796	1	1.0000000
Reading Workshop 8	8004	T4	8	428	92796	6	0.8333333
Reading Workshop 8	81	FY	8	407	62597	9	0.6666667