NewGlobe Case Study - Analyst M&E/Data Analytics Teams

Kamran Ahmed

2023-10-18

The Data

You have received four files, all in .dta and .xlsx formats, so you can use whichever format you prefer. These files are the following:

• "Lesson completion": file provided at the teacher level, meaning that there is a unique row for each teacher. The file contains the grade that each teacher teaches, and the average lesson completion rate over the term of interest.

• "Pupil attendance": file provided at the pupil level (that means that there is a unique row for each pupil). This file includes the unique school ID, unique pupil ID, the pupil's grade, the attendance records, and the present records.

○ The attendance records means the total number of times that a pupil's teacher took attendance.

○ The present records means the total number of times that a pupil was present, out of the attendance

• "Pupil scores": file provided at the pupil*subject level (that means that there are more than one row per pupil). This file includes the unique school ID, unique pupil ID, the pupil's grade, the subject for this assessment, and the score obtained in this assessment.

• "School information": file provided at the school-level. It includes the region and province where each school is located, the unique school ID, and the "treatment status" (yes/no) for a given tutoring program.

Load libraries

```
library(tidyverse)
```

```
## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──
## ✔ dplyr     1.1.2     ✔ readr     2.1.4
## ✔ forcats   1.0.0     ✔ stringr   1.5.0
## ✔ ggplot2   3.4.2     ✔ tibble    3.2.1
## ✔ lubridate 1.9.2     ✔ tidyr     1.3.0
## ✔ purrr     1.0.1
## ── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()    masks stats::lag()
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts
to become errors
```

```
library(readxl)
```

Load Data

Let's load the data and have a quick look of it to see how the data look like

```
Lesson_completion <- read_excel("Lesson completion.xlsx")
head(Lesson_completion)
```

```
## # A tibble: 6 × 4
##   school_id teacher_id grade   lesson_completion_rate
##       <dbl>      <dbl> <chr>                    <dbl>
## 1       416        505 Grade 1                  0.568
## 2       416        202 Grade 2                  0.681
## 3       416        124 Grade 3                  0.250
## 4       416        516 Grade 4                  0.359
## 5       416        145 Grade 5                  0.397
## 6       792        545 Grade 1                  0.809
```

```
Pupil_attendance <- read_excel("Pupil attendance.xlsx")
head(Pupil_attendance)
```

```
## # A tibble: 6 × 5
##   school_id pupil_id grade   attendance_records present_records
##       <dbl>    <dbl> <chr>                <dbl>           <dbl>
## 1     35175        1 Grade 1                 91              69
## 2     40580        7 Grade 2                 92              86
## 3      9342        8 Grade 5                 43              39
## 4    858450       10 Grade 5                 86              62
## 5       792       13 Grade 3                104              81
## 6    324884       14 Grade 4                 90              67
```

```
Pupil_scores <- read_excel("Pupil scores.xlsx")
head(Pupil_scores)
```

```
## # A tibble: 6 × 5
##   school_id pupil_id grade   subject     score
##       <dbl>    <dbl> <chr>   <chr>       <dbl>
## 1     35175        1 Grade 1 Fluency    65
## 2     35175        1 Grade 1 Kiswahili   0.943
## 3     35175        1 Grade 1 Math        1
## 4     40580        7 Grade 2 Math        0.933
## 5     40580        7 Grade 2 Kiswahili   0.943
## 6     40580        7 Grade 2 Fluency   117
```

```
School_information <- read_excel("School_information.xlsx")
head(School_information)
```

```
## # A tibble: 6 × 4
##   region  province school_id tutoring_program
##   <chr>   <chr>        <dbl> <chr>
## 1 Mombasa Coast       136992 No
## 2 Kilifi  Coast       687400 Yes
## 3 Mombasa Coast       609982 Yes
## 4 Eastern Eastern     223941 No
## 5 Isiolo  Eastern      34092 No
## 6 Isiolo  Eastern      46684 No
```

Step 1: Data cleaning

Please create a file at the student-level which has information about their test scores, school information, their attendance, and their teacher's lesson completion rate. Note that this is the main data set that we expect you to share with us.

Hint: note that the four data sets you will use are all presented at different "levels" of the data (e.g., "School information" is at the level of the school, but "Pupil scores" is at the level of the student). Therefore, we suggest that you start by reshaping the "Pupil scores" file so that each student only has one row in the data, with different columns for their scores in math, fluency, and Kiswahili. Use this as your "base file", and start merging all the other files to this. Be careful with how you merge things: since there are many students to a school or even a teacher, some of these merges will need to be "many-to-one" (but not all).

```
# Create a base file by reshaping the "Pupil scores" file so that each student only has
one row in the data, with different columns for their scores in math, fluency, and Kiswa
hili
Pupil_scores <- Pupil_scores %>%
  pivot_wider(names_from = subject, values_from = score)
head(Pupil_scores)
```

```
## # A tibble: 6 × 6
##   school_id pupil_id grade   Fluency Kiswahili  Math
##       <dbl>    <dbl> <chr>     <dbl>     <dbl> <dbl>
## 1     35175        1 Grade 1      65     0.943 1
## 2     40580        7 Grade 2     117     0.943 0.933
## 3      9342        8 Grade 5     144     0.850 0.700
## 4    858450       10 Grade 5     211     1     0.720
## 5       792       13 Grade 3     221     0.857 0.967
## 6    324884       14 Grade 4     267     0.921 0.900
```

```
# Merge Pupil_attendance file to the base file i.e, Pupil_scores file
pupil_df <- merge(Pupil_scores, Pupil_attendance, by = c("pupil_id", "school_id", "grad
e"))
head(pupil_df)
```

```
##   pupil_id school_id   grade Fluency Kiswahili      Math attendance_records
## 1        1     35175 Grade 1      65 0.9428571 1.0000000                 91
## 2       10    858450 Grade 5     211 1.0000000 0.7200000                 86
## 3      100     32940 Grade 2     170 0.7142857 0.7333333                 61
## 4    10000     49404 Grade 1       7 0.6285715 0.7000000                 93
## 5    10002    223941 Grade 1       0 0.6571429 0.8333333                 92
## 6    10005    822894 Grade 2     137 0.7739512 0.6362270                 92
##   present_records
## 1              69
## 2              62
## 3              49
## 4              46
## 5              44
## 6              80
```

```
# Merge teacher's data
pupil_teacher_df <- merge(pupil_df, Lesson_completion, by = c("school_id", "grade"))
head(pupil_teacher_df)
```

```
##   school_id   grade pupil_id Fluency Kiswahili Math attendance_records
## 1    108210 Grade 1     6430      41 0.4000000  0.9                 89
## 2    108210 Grade 1    10987      32 0.5428572  1.0                 83
## 3    108210 Grade 1    22350      NA 0.1428571  1.0                 85
## 4    108210 Grade 1     5572      41 0.7428572  1.0                 89
## 5    108210 Grade 1    21191      12 0.4000000  0.9                 89
## 6    108210 Grade 1    10184      33 0.8857143  1.0                 57
##   present_records teacher_id lesson_completion_rate
## 1              72        323             0.3953488
## 2              74        323             0.3953488
## 3              26        323             0.3953488
## 4              87        323             0.3953488
## 5              59        323             0.3953488
## 6              55        323             0.3953488
```

```
#merge school information
pupil_teacher_school_df <- merge(pupil_teacher_df, School_information, by = "school_id")
head(pupil_teacher_school_df)
```

```
##   school_id   grade pupil_id Fluency Kiswahili      Math attendance_records
## 1       416 Grade 1    23222      43 0.6571429 0.9666666                 85
## 2       416 Grade 1     8377      11 0.1428571 0.8666667                 85
## 3       416 Grade 1    11313      26 0.1428571 0.7666667                 85
## 4       416 Grade 1     5052      38 0.5428572 1.0000000                 85
## 5       416 Grade 1     6151      21 0.1428571 0.7666667                 85
## 6       416 Grade 1     2097      10 0.2000000 0.6666667                 85
##   present_records teacher_id lesson_completion_rate    region province
## 1              76        505             0.5684008 Kirinyaga  Central
## 2              69        505             0.5684008 Kirinyaga  Central
## 3              64        505             0.5684008 Kirinyaga  Central
## 4              77        505             0.5684008 Kirinyaga  Central
## 5              59        505             0.5684008 Kirinyaga  Central
## 6              62        505             0.5684008 Kirinyaga  Central
##   tutoring_program
## 1               No
## 2               No
## 3               No
## 4               No
## 5               No
## 6               No
```

This is the the main data set that we will work with. Let's export this as a csv file and name it "main_data"

```
write.csv(pupil_teacher_school_df, file = "main_data.csv", row.names = FALSE)
```

Step 2: Calculating KPIs

One of our main KPIs within the Schools Vertical is "Percent Pupils Present". The "layman's definition" of this KPI is "The percentage of pupils who were present, out of all pupils - across all days in the term to date". In other words, the percentage of pupils who were present (for each pupil in the "Pupil attendance" file, this is displayed in the "present_records" variable), out of pupils who had attendance records (the "attendance_records" variable in the same file).

• The first task is to translate this KPI into the data. We will calculate this KPI in two different ways. First, calculate this KPI for all pupils at once. What is the network-level average Percent Pupils Present (use two decimal points)?

```
# Network-Level Average Percent Pupils Present (All Pupils)
network_level_average_kpi <- round(sum(pupil_teacher_school_df$present_records)/sum(pupi
l_teacher_school_df$attendance_records), 2)
network_level_average_kpi
```

```
## [1] 0.76
```

• Now, please calculate this percentage for each school, and create an average at the school-level. What is the average Percent Pupils Present now (use two decimal points)?

```
# School-Level Average Percent Pupils Present
school_level_average_kpi <- pupil_teacher_school_df %>%
  select(school_id, present_records, attendance_records) %>%
  group_by(school_id) %>%
  summarise(total_present = sum(present_records), total_records = sum(attendance_record
s)) %>%
  mutate(school_kpi = round(total_present/total_records, 2)) %>%
  summarise(round(mean(school_kpi), 2))
school_level_average_kpi
```

```
## # A tibble: 1 × 1
##   `round(mean(school_kpi), 2)`
##                          <dbl>
## 1                         0.76
```

• How does the interpretation of the KPI change between the two approaches? Does it matter in this case? When would it matter, (i.e., when would one be more appropriate than the other?) 2-4 sentences max.

```
# The way we interpret the KPI shifts with these two approaches due to their scope. When
we calculate the network-level average, we're looking at the big picture, assessing how
well the entire network (Bridge Kenya programme) is performing by considering all pupils
across all schools. On the other hand, the school-level average narrows our focus to ind
ividual school performance, helping us pinpoint differences between schools. In this cas
e it doesn't matter much because we get the same values for the KPIs through both approa
ches when rounded to two decimal places. However, it would matter when there is more het
erogeneity across schools in terms of attendance rate and number of pupils. If there is
more variation in attendance rate across schools and number of pupils, merely calculatin
g percentage for each school and creating a simple average at the school-level would giv
e the same weightage to each scool regardless of the number of pupils in that school, he
nce the value will deviate from the network-level average. The choice between these appr
oaches hinges on the specific analysis or decision-making context. The network-level app
roach would one be more appropriate when we want to gauge the overall network performanc
e, whereas the school-level approach is valuable for recognizing variations and addressi
ng specific issues within each school. Ultimately, the choice depends on the specific ob
jectives of the analysis or decision-making process.
```

Step 3: Descriptives

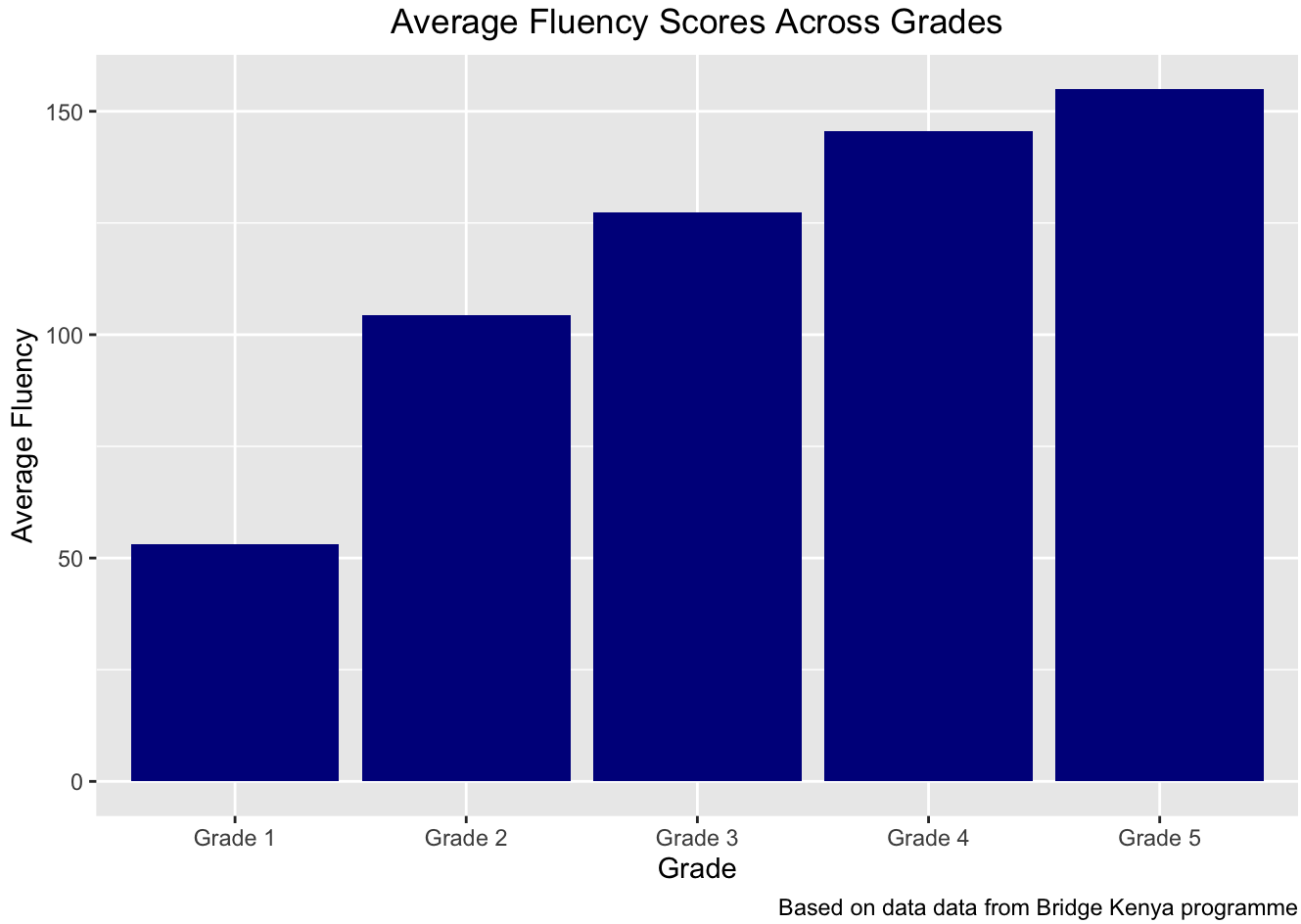
Let's dig into the reading fluency scores in your current data set. These came from the "Pupil scores" data, but you will need the data set you created in Step 1 above to answer these questions. Please answer the following questions as succinctly as possible.

Please create a figure or a table, whichever you prefer, which shows average fluency scores for each of the five grades.

```
pupil_teacher_school_df %>%
  select(grade, Fluency) %>%
  group_by(grade) %>%
  summarise(mean(Fluency, na.rm = TRUE))
```

```
## # A tibble: 5 × 2
##   grade   `mean(Fluency, na.rm = TRUE)`
##   <chr>                           <dbl>
## 1 Grade 1                          53.1
## 2 Grade 2                         104.
## 3 Grade 3                         127.
## 4 Grade 4                         145.
## 5 Grade 5                         155.
```

```
pupil_teacher_school_df %>%
  select(grade, Fluency) %>%
  group_by(grade) %>%
  summarise(avg_fluency = mean(Fluency, na.rm = TRUE)) %>%
  ggplot(aes(x = grade, y=avg_fluency))+
  geom_col(fill= "dark blue")+
  labs(title = "Average Fluency Scores Across Grades",
       x = "Grade",
       y = "Average Fluency",
       caption = "Based on data data from Bridge Kenya programme")+
  theme(plot.title = element_text(hjust = 0.5))
```



- Which regions (using the “region” variable) have the lowest and highest average fluency score across all grades?

```
pupil_teacher_school_df %>%
  select(region, Fluency) %>%
  group_by(region) %>%
  summarise(avg_fluency = mean(Fluency, na.rm = TRUE)) %>%
  filter(avg_fluency == max(avg_fluency, na.rm=TRUE) | avg_fluency == min(avg_fluency, n
a.rm=TRUE))
```

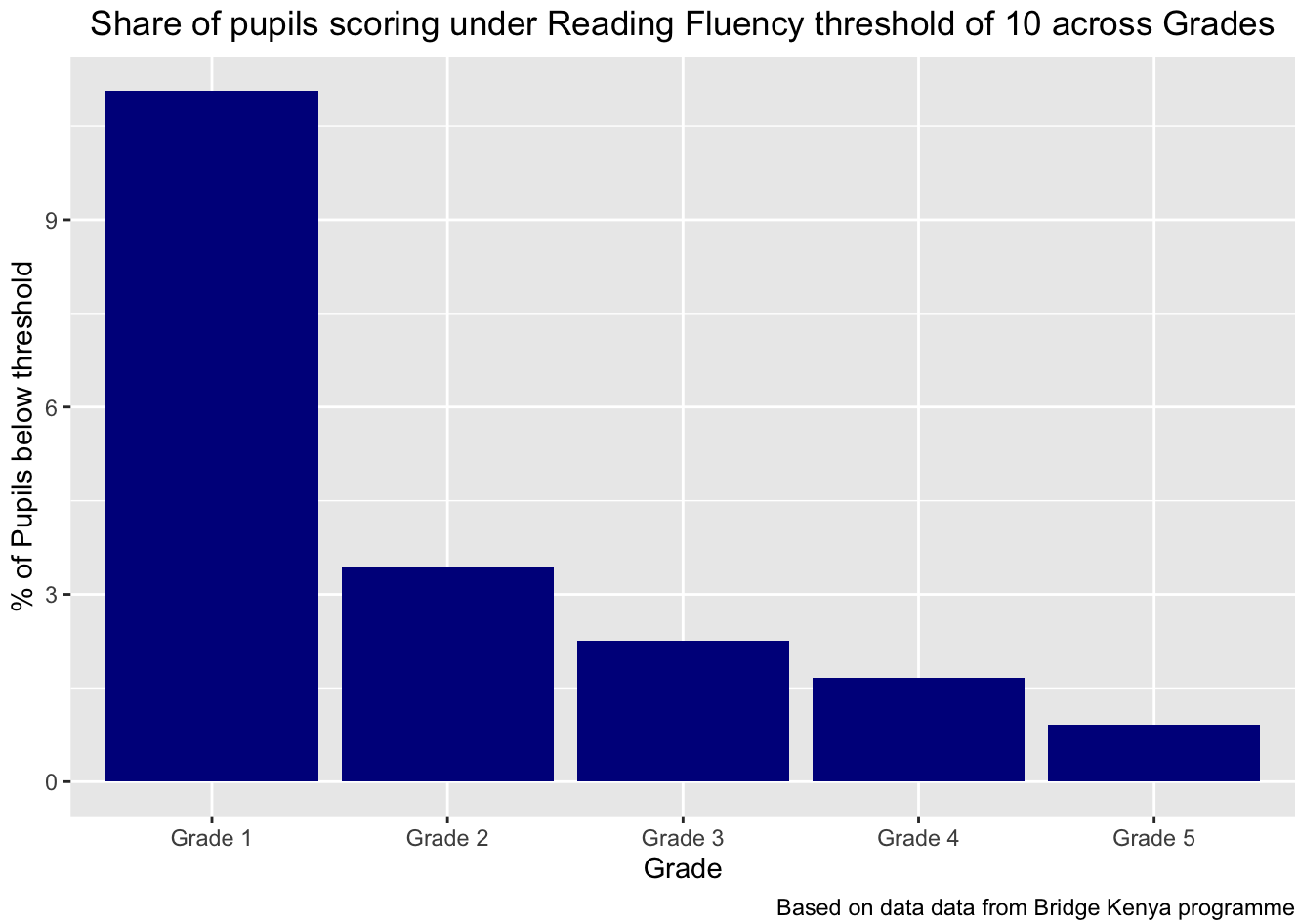
```
## # A tibble: 2 × 2
##   region    avg_fluency
##   <chr>           <dbl>
## 1 Kirinyaga        60.3
## 2 Machakos        158.
```

Kirinyaga has the lowest average fluency score across all grades. Machakos has the highest average fluency score across all grades.

• Please create a binary variable that is 1 if a given child reads at 10 or lower, and 0 otherwise. Please create a bar chart with grades on the x-axis, and the share of pupils scoring under this threshold for each grade.

```
pupil_teacher_school_df$not_fluent <- ifelse(pupil_teacher_school_df$Fluency <=10, 1, 0)

pupil_teacher_school_df %>%
  select(grade, not_fluent) %>%
  group_by(grade) %>%
  summarise(proportion_not_fluent = mean(not_fluent, na.rm = TRUE)) %>%
  ggplot(aes(x = grade, y=proportion_not_fluent*100))+
  geom_col(fill= "dark blue")+
  labs(title = "Share of pupils scoring under Reading Fluency threshold of 10 across Gra
des",
       x = "Grade",
       y = "% of Pupils below threshold",
       caption = "Based on data data from Bridge Kenya programme")+
  theme(plot.title = element_text(hjust = 0.5))
```



- What school has the highest share of pupils scoring under this threshold in grade 3?

```
pupil_teacher_school_df %>%
  filter(grade == "Grade 3") %>%
  select(school_id, not_fluent) %>%
  group_by(school_id) %>%
  summarise(proportion_not_fluent = mean(not_fluent, na.rm = TRUE)) %>%
  filter(proportion_not_fluent == max(proportion_not_fluent, na.rm=TRUE))
```

```
## # A tibble: 1 × 2
##   school_id proportion_not_fluent
##       <dbl>                 <dbl>
## 1    223941                 0.342
```

school_id 223941 is the one that has the highest share of pupils scoring under this threshold in grade 3.

Step 4: Impact evaluation

During this term, we rolled out an intensive after-school tutoring program in 55 schools. The selection to be a part of the 55 schools was randomly assigned - in other words, these schools were part of a randomized controlled trial (RCT). The “School Information” data set has a binary variable for whether each school was part of the program or not.

- Our Chief Academic Officer would like to know whether this program had any effects on test scores in math, Kiswahili, fluency, and/or student attendance. Please conduct any calculations you see fit to answer his questions.

Let's do the following calculations to transform the columns first and prepare the data for regression analysis.

```
# Create a column for attendance performance at the student level
pupil_teacher_school_df$student_attendance_score <- 100*(pupil_teacher_school_df$present
_records/pupil_teacher_school_df$attendance_records)

# The column for Math, Kiswahili and lesson completion rate currently have values betwee
n 0 and 1. Let's multiply them by 100 to show them in percentage form

pupil_teacher_school_df$Math <- pupil_teacher_school_df$Math*100
pupil_teacher_school_df$Kiswahili <- pupil_teacher_school_df$Kiswahili*100
pupil_teacher_school_df$lesson_completion_rate <- pupil_teacher_school_df$lesson_complet
ion_rate*100
```

```
summary(lm(Math ~ tutoring_program, data = pupil_teacher_school_df))
```

```
##
## Call:
## lm(formula = Math ~ tutoring_program, data = pupil_teacher_school_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -74.010 -14.375   2.735  19.402  29.402
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)          70.5982     0.2852  247.56   <2e-16 ***
## tutoring_programYes   3.7767     0.4026    9.38   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 22.13 on 12085 degrees of freedom
##   (614 observations deleted due to missingness)
## Multiple R-squared:  0.007227,   Adjusted R-squared:  0.007145
## F-statistic: 87.98 on 1 and 12085 DF,  p-value: < 2.2e-16
```

```
summary(lm(Kiswahili ~ tutoring_program, data = pupil_teacher_school_df))
```

```
##
## Call:
## lm(formula = Kiswahili ~ tutoring_program, data = pupil_teacher_school_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -80.660 -12.948   6.338  19.340  32.052
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)          67.9482     0.2923   232.4   <2e-16 ***
## tutoring_programYes  12.7121     0.4127    30.8   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 22.7 on 12099 degrees of freedom
##   (600 observations deleted due to missingness)
## Multiple R-squared:  0.07272,    Adjusted R-squared:  0.07264
## F-statistic: 948.8 on 1 and 12099 DF,  p-value: < 2.2e-16
```

```
summary(lm(Fluency ~ tutoring_program, data = pupil_teacher_school_df))
```

```
##
## Call:
## lm(formula = Fluency ~ tutoring_program, data = pupil_teacher_school_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -130.01  -56.02  -12.02   49.98  252.99
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)          97.2551     0.9118  106.66   <2e-16 ***
## tutoring_programYes  32.7602     1.2810   25.57   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 70.09 on 11974 degrees of freedom
##   (725 observations deleted due to missingness)
## Multiple R-squared:  0.05179,    Adjusted R-squared:  0.05171
## F-statistic:   654 on 1 and 11974 DF,  p-value: < 2.2e-16
```

```
summary(lm(student_attendance_score ~ tutoring_program, data = pupil_teacher_school_df))
```

```
##
## Call:
## lm(formula = student_attendance_score ~ tutoring_program, data = pupil_teacher_school
_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -77.166  -8.931   3.993  13.030  25.300
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)          74.7002     0.2232 334.613  < 2e-16 ***
## tutoring_programYes   2.4660     0.3159   7.806 6.35e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 17.8 on 12699 degrees of freedom
## Multiple R-squared:  0.004776,   Adjusted R-squared:  0.004697
## F-statistic: 60.94 on 1 and 12699 DF,  p-value: 6.348e-15
```

```
# By running simple linear regression with Math, Kiswahili, Fluency, and Attendance scor
e as the dependent variable and tutoring program as the independent variable we saw that
tutoring program has a statistically significant positive effect on students' test score
s in Math, Kiswahili, Fluency, and attendance on average. The results are highly statist
ically significant even at a significance level as low as 0.001.
```

● After conducting the impact evaluation, we have heard anecdotally that teachers in schools that received tutoring felt more motivated and were completing their lessons at a faster pace. Hence, we could worry that the effects that we see are not (solely) due to the tutoring program, but also due to the higher lesson completion rate. Does this hypothesis hold up in the data?

```
# If teachers in schools that received tutoring indeed felt more motivated and were comp
leting their lessons at a faster pace then tutoring program also has an effect on teache
rs completion rate. In that case teacher's lesson completion rate would be a confounding
variable that also effects the outcome variable and is being effected by the treatment v
ariable (tutoring program). If this is true, omitting teachers lesson completion rate fr
om the regressions would make the results biased as the effects of teachers' higher less
on completion rate would also be wrongly attributed to the tutoring program.
#Let's first run a diagnostic regression regression of teachers lesson completion rate o
n tutoring program to see if there is any effect of tutoring program on lesson completio
n rate. Then, we will modify the above regressions by controlling for the effects of les
son completion rate by including this variable in our regression analysis as a covariat
e.
```

```
summary(lm(lesson_completion_rate ~ tutoring_program, data = pupil_teacher_school_df))
```

```
##
## Call:
## lm(formula = lesson_completion_rate ~ tutoring_program, data = pupil_teacher_school_d
f)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -60.854 -16.288   4.733  17.321  39.588
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)          60.4115     0.2810 214.996   <2e-16 ***
## tutoring_programYes   0.4425     0.3976   1.113    0.266
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 22.41 on 12699 degrees of freedom
## Multiple R-squared:  9.753e-05,  Adjusted R-squared:  1.879e-05
## F-statistic: 1.239 on 1 and 12699 DF,  p-value: 0.2657
```

```
# In the above regression results, we see that tutoring program does not have a statisti
cally significant effect on lesson completion rate. This suggests that it is not a signi
ficant confounder. However, it might be confounding in combination with the tutoring pro
gram variable so we should still use it as a control variable. Below, we control for les
son_completion_rate to see if the results of regression vary from the above ones.
```

```
summary(lm(Math ~ tutoring_program+lesson_completion_rate, data = pupil_teacher_school_d
f))
```

```
##
## Call:
## lm(formula = Math ~ tutoring_program + lesson_completion_rate,
##     data = pupil_teacher_school_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -76.050 -14.325   2.933  19.017  36.814
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)            63.140229   0.609769 103.548   <2e-16 ***
## tutoring_programYes     3.733093   0.399538   9.344   <2e-16 ***
## lesson_completion_rate  0.122980   0.008907  13.807   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 21.96 on 12084 degrees of freedom
##   (614 observations deleted due to missingness)
## Multiple R-squared:  0.02265,    Adjusted R-squared:  0.02249
## F-statistic:   140 on 2 and 12084 DF,  p-value: < 2.2e-16
```

```
summary(lm(Kiswahili ~ tutoring_program+lesson_completion_rate, data = pupil_teacher_sch
ool_df))
```

```
##
## Call:
## lm(formula = Kiswahili ~ tutoring_program + lesson_completion_rate,
##     data = pupil_teacher_school_df)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -81.88 -12.76   6.12  18.44  35.05
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)            64.953350   0.629297 103.216  < 2e-16 ***
## tutoring_programYes    12.693886   0.412233  30.793  < 2e-16 ***
## lesson_completion_rate  0.049367   0.009189   5.372 7.92e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 22.67 on 12098 degrees of freedom
##   (600 observations deleted due to missingness)
## Multiple R-squared:  0.07492,    Adjusted R-squared:  0.07477
## F-statistic: 489.9 on 2 and 12098 DF,  p-value: < 2.2e-16
```

```
summary(lm(Fluency ~ tutoring_program+lesson_completion_rate, data = pupil_teacher_schoo
l_df))
```

```
##
## Call:
## lm(formula = Fluency ~ tutoring_program + lesson_completion_rate,
##     data = pupil_teacher_school_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -148.52  -55.43  -10.80   49.43  240.96
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)            115.75903    1.95379   59.25   <2e-16 ***
## tutoring_programYes     32.76266    1.27499   25.70   <2e-16 ***
## lesson_completion_rate  -0.30357    0.02839  -10.70   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 69.76 on 11973 degrees of freedom
##   (725 observations deleted due to missingness)
## Multiple R-squared:  0.06076,    Adjusted R-squared:  0.06061
## F-statistic: 387.3 on 2 and 11973 DF,  p-value: < 2.2e-16
```

```
summary(lm(student_attendance_score ~ tutoring_program+lesson_completion_rate, data = pu
pil_teacher_school_df))
```

```
##
## Call:
## lm(formula = student_attendance_score ~ tutoring_program + lesson_completion_rate,
##     data = pupil_teacher_school_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -77.481  -8.920   4.001  12.993  26.241
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)            73.759424   0.480803 153.409  < 2e-16 ***
## tutoring_programYes     2.459136   0.315868   7.785 7.49e-15 ***
## lesson_completion_rate  0.015573   0.007049   2.209   0.0272 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 17.8 on 12698 degrees of freedom
## Multiple R-squared:  0.005158,   Adjusted R-squared:  0.005001
## F-statistic: 32.92 on 2 and 12698 DF,  p-value: 5.501e-15
```

```
# By including lesson completion rate variable in our regression analysis as a covariat
e, we saw that the coefficients for tutoring program in all of the four regressions rema
ined almost the same as before and were not effected much. Although we do see a statisti
cally significant effect of lesson completion rate on students performance and attendanc
e, this variable is independently assocaited with the outcome variables and does not med
iate the effects of the tutoring program.
# The hypothesis that the effects that we saw are not (solely) due to the tutoring progr
am, but also due to the higher lesson completion rate does not hold up in the data.
```