CUNY SPS
DATA 607 - Fall 2023
Data Science in Context
Presentation

Marley Myrianthopoulos

Command Line Arguments

Streamlining Your Workflow for Fun and Profit

Disclaimers

- I am not an expert in using the command line.
- I am not a working data scientist.
- This presentation chronicles my personal journey of improving my workflow. Your mileage may vary.

The Situation

- My school needs a list of every student who came to school late last week.
- Operational definition of “Late”: Arrived too late to attend 2nd period.
- Data to include for each student:
  - Name
  - ID number
  - Grade level
  - How many days they came late last week
- We have a .csv file of attendance data with every student’s attendance for every period.
- Problem: There is no record for overall daily attendance, only per period.
- Solution: Students are considered present for the day if they are marked present (“P”, “T”, or “S”) in two or more classes that day.

Example of Existing Data

Date	Period	Attendance	StudentID	Name	GradeLevel	Teacher
10/23/2023	2	P	1001	Student 1	12	Teacher 1
10/23/2023	3	P	1001	Student 1	12	Teacher 2
10/23/2023	4	P	1001	Student 1	12	Teacher 3
10/23/2023	2	P	1002	Student 2	12	Teacher 1
10/23/2023	3	P	1002	Student 2	12	Teacher 2
10/23/2023	4	P	1002	Student 2	12	Teacher 3
10/23/2023	2	P	1003	Student 3	12	Teacher 2
10/23/2023	3	P	1003	Student 3	12	Teacher 4
10/23/2023	4	P	1003	Student 3	12	Teacher 5
10/23/2023	5	P	1003	Student 3	12	Teacher 1
10/23/2023	2	P	1004	Student 4	12	Teacher 1
10/23/2023	3	P	1004	Student 4	12	Teacher 2
10/23/2023	4	P	1004	Student 4	12	Teacher 3
10/23/2023	2	A	1005	Student 5	12	Teacher 1
10/23/2023	3	A	1005	Student 5	12	Teacher 1
10/23/2023	4	A	1005	Student 5	12	Teacher 2
10/23/2023	5	A	1005	Student 5	12	Teacher 4
10/23/2023	2	P	1006	Student 6	12	Teacher 2
10/23/2023	3	P	1006	Student 6	12	Teacher 4

Example of Desired Output

StudentID	Name	GradeLevel	missed2nd
1012	Student 12	12	5
1061	Student 61	12	5
1019	Student 19	12	4
1072	Student 72	12	4
1257	Student 257	10	4
1328	Student 328	9	4
1017	Student 17	12	3
1026	Student 26	12	3
1089	Student 89	11	3
1183	Student 183	12	3
1191	Student 191	11	3
1214	Student 214	10	3
1231	Student 231	10	3
1235	Student 235	10	3
1244	Student 244	10	3
1247	Student 247	12	3
1270	Student 270	10	3
1282	Student 282	11	3
1294	Student 294	9	3
1302	Student 302	9	3
1307	Student 307	9	3
1319	Student 319	9	3
1324	Student 324	9	3

Level Zero: Count Them!

Workflow:

- ➔ Make a list of each student in the school.
- ➔ Review each student's attendance for each day to determine if they arrived late.
- ➔ Tally the total number of late arrivals for each student.
- ➔ Rewrite the list in descending order by number of days missing 2nd period.

Level One: Google Sheets

Workflow:

1) Open a web browser.
2) Navigate to sheets.google.com
3) Create a new Google sheets document.
4) Import the .csv attendance file into the sheet.
5) Create a new column called “daily.” Populate the new column with the number of periods that the student in that row was present in the day for that row. Formula: =countifs(A:A,A2,D:D,D2,C:C,"P")+countifs(A:A,A2,D:D,D2,C:C,"T")+countifs(A:A,A2,D:D,D2,C:C,"S")
6) Create a new column called “StudentID”. Populate the new column with a list of the unique student ID numbers from the original StudentID column. Formula: =unique(D2:D)
7) Create a new column called “Name”. Populate the new column with the names of the students whose ID numbers appear in the column in step 5. Formula: =vlookup(I2,D:E,2,False)
8) Create a new column called “GradeLevel”. Populate the new column with the grade levels of the students whose ID numbers appear in the column in step 5. Formula: =vlookup(I2,D:F,3,False)
9) Create a new column called “missed2nd”. Populate the new column with the number of rows in the original data where the student ID number is the same as the student ID number from the column in step 5, the “Period” column is 2, the “Attendance” column is not “P”, “T”, or “S”, and the “daily” column is 2 or more. Formula: =countifs(D:D,I2,C:C,"<>P",C:C,"<>T",C:C,"<>S",B:B,2,H:H,">=2")
10) Create a new page in the Google sheets document.
11) Copy the columns from steps 4-7 and use “paste without formatting” to paste them into the new page.
12) Sort the sheet by the “missed2nd” column, descending.
13) Download the data from this new page as a .csv file.
14) Move the .csv file from your downloads folder to the desired location in your file system.

Additional Reports

- ➔ How many times each student missed school each week.
- ➔ Which students had perfect attendance last week (present in every class).

Level 2: R Program

```r
raw_data <- read.csv("/Users/marleymyrianthopoulos/Desktop/Attendance Presentation/Level 2/Week _/attendance_week__.csv")

library(dplyr)

**Enter week number before running program to complete file pathway**

modified_data <- raw_data %>%
  group_by(StudentID, Date) %>% mutate(daily = sum(case_when(Attendance %in% c("P","T","S") ~ 1, T ~ 0))) %>% ungroup()

late_students <- modified_data %>%
  filter(Period == 2 & Attendance == "A" & daily >= 2) %>% group_by(StudentID) %>% mutate(missed2nd = length(StudentID)) %>% ungroup() %>%
  select(StudentID, Name, GradeLevel, missed2nd) %>% distinct() %>% arrange(desc(missed2nd))

days_missed <- modified_data %>%
  filter(Period == 4 & daily < 2) %>% group_by(StudentID) %>% mutate(days_missed = length(StudentID)) %>% ungroup() %>%
  select(StudentID, Name, GradeLevel, days_missed) %>% distinct() %>% arrange(desc(days_missed))

perfect_attendance <- modified_data %>%
  group_by(StudentID) %>% mutate(missed_periods = sum(case_when(!Attendance %in% c("P","T","S") ~ 1, T ~ 0))) %>% ungroup() %>%
  filter(missed_periods == 0) %>% select(StudentID, Name, GradeLevel) %>% distinct() %>% arrange(GradeLevel)

write.csv(late_students, "/Users/marleymyrianthopoulos/Desktop/Attendance Presentation/Level 2/Week _/late_students.csv", row.names = FALSE)
write.csv(days_missed, "/Users/marleymyrianthopoulos/Desktop/Attendance Presentation/Level 2/Week _/days_missed.csv", row.names = FALSE)
write.csv(perfect_attendance, "/Users/marleymyrianthopoulos/Desktop/Attendance Presentation/Level 2/Week _/perfect_attendance.csv", row.names = FALSE)
```

Level 2: R Program

Workflow:

1) Open RStudio.
2) Open the .rmd file in RStudio.
3) Enter the week number in the file pathway for the attendance data .csv file.
4) Enter the week number in the file pathway for the report outputs.
5) Run the program.

Level 3: R Function

```r
level3 <- function(week_number) {

csv_pathway <- paste("/Users/marleymyrianthopoulos/Desktop/Attendance Presentation/Level 3/Week ", week_number, "/attendance_week_", week_number, ".csv", sep = "")

raw_data <- read.csv(csv_pathway)

library(dplyr)

modified_data <- raw_data %>%
  group_by(StudentID, Date) %>% mutate(daily = sum(case_when(Attendance %in% c("P","T","S") ~ 1, T ~ 0))) %>% ungroup()

late_students <- modified_data %>%
  filter(Period == 2 & Attendance == "A" & daily >= 2) %>% group_by(StudentID) %>% mutate(missed2nd = length(StudentID)) %>% ungroup() %>%
  select(StudentID, Name, GradeLevel, missed2nd) %>% distinct() %>% arrange(desc(missed2nd))

days_missed <- modified_data %>%
  filter(Period == 4 & daily < 2) %>% group_by(StudentID) %>% mutate(days_missed = length(StudentID)) %>% ungroup() %>%
  select(StudentID, Name, GradeLevel, days_missed) %>% distinct() %>% arrange(desc(days_missed))

perfect_attendance <- modified_data %>%
  group_by(StudentID) %>% mutate(missed_periods = sum(case_when(!Attendance %in% c("P","T","S") ~ 1, T ~ 0))) %>% ungroup() %>%
  filter(missed_periods == 0) %>% select(StudentID, Name, GradeLevel) %>% distinct() %>% arrange(GradeLevel)

write_pathway <- paste("/Users/marleymyrianthopoulos/Desktop/Attendance Presentation/Level 3/Week ", week_number, sep = "")
write.csv(late_students, file.path(write_pathway, "late_students.csv"), row.names = FALSE)
write.csv(days_missed, file.path(write_pathway, "days_missed.csv"), row.names = FALSE)
write.csv(perfect_attendance, file.path(write_pathway, "perfect_attendance.csv"), row.names = FALSE)

}

level3(_)
```

The input variable for the function creates the appropriate file patheways

Enter week number before running the program

Level 3: R Function

Workflow:

1) Open RStudio.
2) Open the .rmd file in RStudio.
3) Change the week number in the function input.
4) Run the program.

Level 4: Command Line Argument

Setup:

1) Include code to get the week number from the command line prompt using commandArgs.
2) Save program as a .R file.
3) Move .R file to your working directory.

Level 4: Command Line Argument

```
week_number <- commandArgs(trailingOnly = TRUE)

csv_pathway <- paste("/Users/marleymyrianthopoulos/Desktop/Attendance Presentation/Level 3/Week ", week_number, "/attendance_week_", week_number, ".csv", sep = "")

raw_data <- read.csv(csv_pathway)

library(dplyr)

modified_data <- raw_data %>%
  group_by(StudentID, Date) %>% mutate(daily = sum(case_when(Attendance %in% c("P","T","S") ~ 1, T ~ 0))) %>% ungroup()

late_students <- modified_data %>%
  filter(Period == 2 & Attendance == "A" & daily >= 2) %>% group_by(StudentID) %>% mutate(missed2nd = length(StudentID)) %>% ungroup() %>%
  select(StudentID, Name, GradeLevel, missed2nd) %>% distinct() %>% arrange(desc(missed2nd))

days_missed <- modified_data %>%
  filter(Period == 4 & daily < 2) %>% group_by(StudentID) %>% mutate(days_missed = length(StudentID)) %>% ungroup() %>%
  select(StudentID, Name, GradeLevel, days_missed) %>% distinct() %>% arrange(desc(days_missed))

perfect_attendance <- modified_data %>%
  group_by(StudentID) %>% mutate(missed_periods = sum(case_when(!Attendance %in% c("P","T","S") ~ 1, T ~ 0))) %>% ungroup() %>%
  filter(missed_periods == 0) %>% select(StudentID, Name, GradeLevel) %>% distinct() %>% arrange(GradeLevel)

write_pathway <- paste("/Users/marleymyrianthopoulos/Desktop/Attendance Presentation/Level 3/Week ", week_number, sep = "")
write.csv(late_students, file.path(write_pathway, "late_students.csv"), row.names = FALSE)
write.csv(days_missed, file.path(write_pathway, "days_missed.csv"), row.names = FALSE)
write.csv(perfect_attendance, file.path(write_pathway, "perfect_attendance.csv"), row.names = FALSE)
```

The week number comes from the argument provided in the command line prompt

Level 4: Command Line Argument

Workflow:

1) Open Terminal
2) Enter the command: Rscript [code filename].R [week number]. For example, for this program to generate the reports for week 7, I'll enter: Rscript dcislevel4.R 7
3) Celebrate!

Summary

Level 0: Count them!

Level 1: Google Sheets

Level 2: R Program

Level 3: R Function

Level 4: Command Line Argument

Thank you!

The following resources are available in a github repository here:
https://github.com/Marley-Myrianthopoulos/datascienceincontext

- ➔ These slides
- ➔ Sample attendance data
- ➔ .rmd files for the Level 2 and Level 3 code
- ➔ The .R file for the Level 4 code

DFTBA!