### **CTL Data Analyst Interview Assignment Instructions**

### **Overview:**

*The intent of the R script I created (“CTL\_DATA\_ANALYST\_INTERVIEW\_ASSIGNMENT\_SCRIPT.R”) is to be able to create a data visualization of the number of students in attendance of CTL events per Stanford School with minimal manual excel cleaning. This does require some assumptions on how the given attendance data files are formatted, however for the purpose of this assignment I have only detailed instructions on how to reformat the provided files in order to successfully run the script.*

**Instructions:**

Reformatting Given Excel Files –

*Below I detail what modifications I made to the given excel files. A copy of an already modified folder is on my repository, with the suffix “\_MODIFIED” attached to the name of every modified excel workbook.*

1. Remove the first 3 rows of the “Zoom” sheet for the following files in order to standardize column for their “Zoom” sheet with the rest of the files
   1. Fall Quarter 2021 -> Chem Virtual Study Halls -> 2020\_10\_03\_CVSH.xlsx
   2. Fall Quarter 2021 -> Chem Virtual Study Halls -> 2020\_10\_18\_CVSH.xlsx
   3. Fall Quarter 2021 -> Chem Virtual Study Halls -> 2020\_11\_01\_CVSH.xlsx
   4. Fall Quarter 2021 -> Chem Virtual Study Halls -> 2020\_11\_15\_CVSH.xlsx
   5. Winter Quarter 2021 -> Workshops -> 2021\_01\_14\_WKS.xlsx
   6. Winter Quarter 2021 -> Workshops -> 2021\_02\_04\_WKS
   7. Winter Quarter 2021 -> Workshops -> 2021\_02\_4\_WKS2
2. Add column header “Name (Original Name)” to the “Zoom” sheet Column A of Fall Quarter 2021 -> Workshops -> 2020\_10\_07\_WKS\_UPDATED.xlsx (it is not necessary to give names the other columns as they are not used in the analysis)
3. Create a column “Stanford School” in the “Eventbrite” sheet of the listed below files. Manually categorize the Stanford School of each registered participant using their “Undergraduate Major” column (Note: I felt this was much faster than automating the process, since there weren’t many declared majors). If the student is “Undeclared” or their major is simply not available leave the corresponding “Stanford School” cell blank. Where you create the column is not important as long as you create it on the correct sheet. In addition please be sure to use the School names used by other files (ex. “Engineering”, “Law”, “Humanities and Sciences”, “Medicine”, “Business”, “Earth, Energy, and Environmental Sciences”, “Education”)
   1. Fall Quarter 2021 -> Chem Virtual Study Halls -> 2020\_10\_03\_CVSH.xlsx
   2. Fall Quarter 2021 -> Chem Virtual Study Halls -> 2020\_10\_18\_CVSH.xlsx
   3. Fall Quarter 2021 -> Chem Virtual Study Halls -> 2020\_11\_01\_CVSH.xlsx
   4. Fall Quarter 2021 -> Chem Virtual Study Halls -> 2020\_11\_15\_CVSH.xlsx
   5. Spring Quarter 2021 -> Workshops -> 2021\_05\_28\_WKS.xlsx

Running the Script –

*I decided to use R as it was the language which I’ve used for data visualization in the past the most. I used the RStudio IDE to facilitate script creation. However, obtaining RStudio is not necessary to run the script. I installed* ***RStudio*** *and* ***tidyverse*** *packages (along with tidyverse’s associated packages to run). You may run the script in RStudio or your terminal using the command:*

Rscript --vanilla Desktop/data\_analyst\_sample\_data\_MODIFIED/CTL\_DATA\_ANALYST\_INTERVIEW\_ASSIGNMENT\_SCRIPT.R

***Some thing to note:***

***Make sure the folder containing your files and script is named “data\_analyst\_sample\_data\_MODIFIED” and on your Desktop. The folder uploaded to the repository should abide by this, so long as you download it to your Desktop, the command above should work. If recreating the folder yourself using the above instructions, simply change the folder name at the end to match, and move the R script into it. Even if you change the folder name when writing in the command, the script itself still runs on the assumption that the folder you are using to store your excel files is named “data\_analyst\_sample\_data\_MODIFIED” and on your Desktop.***

*After you run the script the following graph should be saved in “data\_analyst\_sample\_data\_folder\_MODIFIED” as “Student Per School Graph.pdf” and look like the graph below:*

*Chart, bar chart

Description automatically generated*

**Notes on Process and Decision-Making:**

*This section just has general notes on my thought-process for analyzing the dataset.*

1. Determining “Attendees” –
   1. I decided to call people who attended the event those who both registered and actually showed up on the Zoom. I did so by cross-referencing the list of people who registered with those who attended, and only including them in the count if the name showed up in both lists.
      1. This may exclude people who attended an event but didn’t register. However, I assumed since all the events were on Zoom the number of people doing this would be small (since they would need to have access to the Zoom link), and in addition, the people who attended but did not register would not provide any data regarding number of attendees for a particular school, anyway.
      2. This also excludes people who registered but did not attend, however I think it makes sense to exclude people in this category.
      3. Some people may be excluded if their name on the Zoom attendance sheet and Eventbrite registration sheet is different, but I don’t think this can be avoided when automating this process.
   2. I also only counted each unique attendee once. In other words, if someone attended three events that year, they would only be counted once for the data analysis. I think this makes the most sense too, given it does not change the total number of people being attracted to CTL events per each school, and therefore will likely give us a better estimate of the information we want.
2. Determining “School” –
   1. For those who we only had majors in, I simply filtered the attendees into the school with their major. For those who are undeclared I included them in the “NA” category, the people whose schools we didn’t know.