CTData Technical Exercise

Research Analyst

Overview

This technical exercise is an opportunity for you to provide an example of your work and skill set. You have a maximum of 2 hours to work on this exercise (not counting the time it takes to read these instructions). Integrity is a core value of our organization. Please use the honor system and keep to this allotted amount of time. If you can't complete all parts of the exercise in 2 hours, please make a note of this in your response and briefly describe how you would have approached the incomplete parts of the exercise if you'd had more time. When you are finished, email your response and any files used for the exercise to our Operations Manager, Wenyu Xie (wx@ctdata.org). If you need clarification for any part of this exercise, please reach out to our Senior Research Associate, Cynthia Willner (cwillner@ctdata.org).

Technical Exercise

The Research Analyst at CTData will be responsible for processing, analyzing, and reporting on data from a quarterly statewide survey of families with young children. For this technical exercise, we have created a synthetic (fake) dataset based on an actual dataset from this survey. This synthetic dataset is in the attached file named "*CTData_Research-Analyst_Technical-Exercise_DATA.CSV*." Using this synthetic dataset, we ask you to demonstrate how you would process and analyze the data **in R.**[1]

I. Review the data codebook in the attached file named "*CTData Research Analyst Technical Exercise - Codebook.xlsx*." Note that the "Display Logic" column provides the conditions under which a survey question was displayed to the respondent. If the display logic condition is false, the question was not displayed and the response is missing (coded as a blank value in the CSV file). Note that most survey questions were not forced-

[1] If you do not currently have R on your computer, you can download the software for free at https://www.r-project.org/, and you may also wish to (optionally) download the RStudio IDE here: https://posit.co/download/rstudio-desktop/. You do not need to count the time it may take you to download and install software in the 2 hours allotted for completing this exercise.

response, so missing data are also present due to non-response to individual survey questions.

II. Create a single R script, R Markdown document, or R Quarto document to conduct all of the data processing steps and analyses described below. Please document your code clearly with comments. **Share the full R script/Markdown/Quarto document (.R, .Rmd, or .qmd) *and* any outputs (e.g., rendered html or PDF documents, saved R datasets or exported CSV/Excel files, etc.) with your response**.

Data Processing Steps & Analyses:

1) Conduct basic descriptive analyses of each variable in the dataset to familiarize yourself with the data. Show and comment your work in your R script and share any output that your script generates.

2) Recode variables, create new variables, and/or restructure the dataset as necessary to support the analysis questions described below. Show and comment your work in your R script.

3) Answer the following analysis questions. You may provide your answers in any format you prefer, for example a Word document, PowerPoint slides, an R Markdown or Quarto HTML document, etc.

    a. What percentage of respondents were currently using non-parental care for *any child* at the time of the survey? Include the sample size for your analysis.

    b. What percentage of *young children* were currently receiving non-parental care at the time of the survey? Include the sample size for your analysis.

    c. What percentage of respondents within each household income category (relative to state median income) were currently using non-parental care for *any child*? Please include the sample size for your analysis and a data visualization (you may create this visualization using whatever software you like).

    d. Is the variation among household income groups in the likelihood of using non-parental care for any child *statistically significant*? Report the test(s) you used, the value(s) of the test statistic(s), the degrees of freedom, and the p-value(s). Describe this result and what it means in language that is accessible to a general audience.

Please submit the following documents in response to this technical exercise:

1) The full R script/Markdown/Quarto document (.R, .Rmd, or .qmd) used for all data processing steps and analyses

2) Any files that you may have saved, rendered or exported from R (e.g., rendered HTML or PDF documents, saved R datasets, exported CSV/Excel files, etc.). Note that you are not *required* to save or export any documents from R.

3) A document with your responses to the questions in section II(3) above. This may be an HTML or PDF document rendered using R Markdown or Quarto, a Word document, PowerPoint slides, or any other format you prefer.