Data Analyst Internship Description

|  |
| --- |
| New York City Department of Health and Mental Hygiene (DOHMH)Office of School Health (OSH) |

|  |  |
| --- | --- |
| From: | Alexander M Vogel |
| Date: | May 28, 2021 |
| Re: | OSH Data Analyst Intern Requirements |
| In this role you will be required to aggregate and manage data from various external sources and transform the data into statistics for the Office of School Health physicians. You will need to follow the provided Readme file to update and reproduce the data analysis from Spring 2021. Building off the links and resources included, you will need to use a web browser to find the updated datasets and download the appropriate files.  Using RStudio, you should update the included R Markdown (.Rmd) file with the most current data. The code from Spring 2021 includes tasks of: data import (.xls and .csv) and data transformation (using RStudio as well as Excel or SQL to trim down larger datasets) including aggregation (left joins, mutations, summaries, pivots, counts, etc.) and recoding (mutate, factor, recode). In the case of the Youth Risk Behavior Studies (YRBS), you will need to analyze complex survey designs with strata, weights, and PSUs as well as make the necessary transformations (filters, selections, dummy columns). Finally, you will need to export the data according to the specifications of your leadership. This may include recoding column names or row values, utilizing functions and loops as necessary to avoid manual mistakes and allow for simplified future changes.  Knowledge of the *tidyverse* is required. This is a suite of instrumental packages in RStudio for tidy data, utilizing *magrittr* pipes (*%>%*) for efficient and organized code, primarily using the sub-packages *tidyr, dplyr, stringr, readr,* and *readxl*. Complex survey analysis can be completed using the *survey* package with the help of *srvyr*, which is the *dplyr* equivalent for survey data. Other optional packages used in Spring 2021 were *naniar* for dealing with missing values, *fastDummies* for easy dummy columns, and *glue* for recoding strings.  Once the data has been exported into a tidy table, you may be asked to create visualizations. You may use any software you like. For Spring 2021, Tableau was used. Each visualization included several variables, with each school district having one page to itself. Axes should be labeled, titles provided, and data should be labeled appropriately. At the completion of your analysis, you will need to ensure proper documentation, including a Readme file including data sources and inline comments on the R Markdown file to explain coding methodology. | |