UNITED STATES DIGITAL CORPS

Thank you for taking the time to participate in the project-based assessment phase of the U.S. Digital Corps hiring process. We are excited that you are moving forward with us!

All projects are due by 11:59pm ET on Sunday, January 8, 2023.

Instructions: Please read through the below prompt thoroughly. Follow all directions and be sure to answer every question.

- You should plan to spend two hours completing this assessment.
- Please submit any code you wrote in an .ipynb file or export it as a PDF. Written responses should be submitted in a PDF of three pages or fewer. Responses received in other formats may not be reviewed and written responses longer than three pages will only have the first three pages reviewed.
- Save your project using this naming convention:
    - Code file: "LastName_FirstName_Code_Data"
    - Written responses (if applicable): "LastName_FirstName_Written_Data"
- **Email your completed assessment to usdigitalcorps+assessment@gsa.gov no later than 11:59pm ET on Sunday, January 8, 2023.**
- **Do not share any details about this assessment with anyone, both during the assessment period and after the deadline has passed.**

Evaluation: Your submission will be reviewed by subject matter experts (SMEs) in your field and evaluated against the competencies and specialized experience shared in the job posting for this track. SMEs will look to understand your responses as well as your thought process.

Support: Please reach out to usdigitalcorps@gsa.gov with any questions or **to request an accommodation.**

This assessment is used solely for U.S. Digital Corps applicant evaluation.

UNITED STATES DIGITAL CORPS

Project-Based Assessment – Data Science and Analytics

Applicant instructions

1. The data for this assessment is taken from the CDC and Behavioral Risk Factor Surveillance System.
2. A codebook describing the data variables is here: https://www.cdc.gov/brfss/annual_data/2014/pdf/CODEBOOK14_LLCP.pdf
3. We have added an identifier with the variable name 'PERSONID'.
4. You are asked to prepare the data in order to build a model to identify risk factors for diabetes. Show your data processing work in a reproducible way using a data analysis platform.
5. You may use common data science languages (R or Python) and tools (Jupyter, Google Colab, Rmd, etc.) to complete this assessment. Return a PDF report and supporting code.

Assessment prompts

1. Download, merge, and describe the dataset and its basic characteristics (e.g., shape, variable types, basic stats).
2. Choose several variables and create visualizations to show their distributions. Justify your variable selection.
3. Clean the dataset to handle any missing data and justify your decisions.
4. For building a model, would you rescale any data in this dataset? How and why or why not?
5. Build a model to identify risk factors for diabetes. Explain your choice of model and what it can predict. What metrics would you use to assess performance? For this dataset, how would you know your model is adequate?
6. Using these data, what are some identifiable risk factors for diabetes? How do you know? Explain as if you were reporting the results to a non-technical stakeholder.