**Data Science Program Final Project**

# Executive Summary

At the conclusion of the Entity Academy/ Woz U Data Science certification program, all graduates must complete a final project of their choice. This project involves six structured lessons that correlate with a given six-week timeline. A student may work alone or in a group with other students. This document is as a result of the dedicated work of Virginia Mburu Brovont and Lucy Thuku. It will explain the purpose and scope for their chosen topic: US Teen Birth Rates.

# Business Objectives

The team members will demonstrate skills they have acquired throughout the Data Science program. They will be using R, Python, Tableau, and any other relevant programs to wrangle, analyze, and visualize “Teen pregnancies” dataset that is made available in different databases from sources including but not limited to Kaggle and data.gov.

CDC research tells us that "teen childbearing can carry health, economic, and social costs for mothers and their children" [(CDC, Vital Signs, April 2015)](https://www.cdc.gov/vitalsigns/larc/index.html). This information is corroborated by the American Association of Pediatrics (AAP) which reports that "less than 40% of teenage girls who have a child before age 18 earn a high school diploma by age 22," reducing their lifetime earning potential, and that "nearly 2/3 of teenage mothers receive public assistance, and their chances of living in poverty increase as they enter adulthood." Additionally, children born to adolescent mothers are "more likely to perform poorly in school, repeat a grade, or drop out" and daughters are "more likely to repeat the cycle and become teen mothers themselves" [(AAP, "Teenage Pregnancy")](https://www.healthychildren.org/English/ages-stages/teen/dating-sex/pages/Teenage-Pregnancy.aspx).

Adolescent mothers face more risk from complications associated with pregnancy like high blood pressure, anemia, and "poor weight gain." They are also are more at risk for "premature birth, low birth weight, and infant death" when compared to older mothers [(AAP, "Helping Teen Parents and Their Children Build Healthy Futures")](https://www.healthychildren.org/English/ages-stages/teen/dating-sex/pages/Teen-Parents.aspx).

Lowering the teen birth rate is crucial for empowering young women to have better health outcomes, more opportunities for education, and, ultimately, create better financial futures.

At the end of the project, the team members should be able to explain their work in layman’s terms, and present their findings to their fellow students, faculty, staff, and potential employers, via Zoom.

# Background

The purpose of this project is to explore the trends in teenage births in the United States for a certain period.

The team members chose this topic because they were both curios to see how the number of teen births in the United States has changed over the decades and if there were any significant regional differences in the number of births. The topic relates well with them because they both gave birth to children at an early age which brought about several risks to both the mothers and the babies.

The team members hope to bring awareness of risks involved with teen pregnancies.

# Scope

The team members will be using the software taught in their school program to complete the project. They will be intentional on using tools of their interest or tools that may aid finding a job. They may choose to use additional software/tools, but that is not required.

# Functional requirements

Data Wrangling: The downloaded dataset should be successfully cleaned up for analyzing. Columns and unusable columns should be removed. As the dataset is large, the team members should consider sub-setting the dataset in a proper manner, meaning the subset should be a random selection of the data. The datatypes for each column should also be converted to a usable format for the needed analysis.

Data Analysis: The team members will familiarize themselves with the dataset. They should have a good understanding of what each column means, and how the values are measured. They will brainstorm on questions to ask, and what they might gather from the dataset. Then, they will identify the proper functions to create models, predictions, etc.

Data Visualization: Once the team members have a comprehensive understanding of and insight gathered from the dataset, they will work on visualizing the findings. They may decide to use Tableau or other graphing programs and compile the visuals and texts in a Power Point slideshow.

# Personnel requirements

The project will require a team of 2 data science student researchers and 1 product owner

(Group instructor).

They will touch base once a day via Zoom or Slack to problem-solve or to check in on work progresses. Once a week, they will review the past week workload and plan out the next week. They will take turns being the scrum master and report their progress to their instructor.

Once a week, they will meet with their instructor. They should be prepared to ask questions and seek guidance for the next steps.

They may also consult with their coding mentor.

# Delivery schedule

The project will be spread out in six weeks. Week 1 will be for project planning which will include the creation of a Kanban board on Trello, the choosing of a specific topic, a search for data on the topic, determining at least two questions using the chosen data and finally, choosing the method of analysis.

Week 2 will be for Data Wrangling. After gathering the data, there will be a bit of reformatting that will need to happen before they can perform any computation or other analyses. Much of this reformatting is attribute reclassification, which involves converting data from a string format to something more appropriate (another format such as integer, date, etc). They will also trim the dataset down to its most concise, useful form to make the rows and columns as simple as possible while still providing useful insight.

Week 3 will be for Data Exploration. Much of the goal of this project is to provide a platform for self-discovery using interactive maps and graphs. Key summary statistics and figures will be highlighted to get a basic understanding of the events and overall context of the data.

Week 4 will be for Data Analysis. The team members will perform relevant descriptive statistics including summary statistics. The comparison groups and methods will be updated continuously as the project progresses.

Week 5 will be for Data Visualization. Based on the results and findings from their analysis, the techniques that they will use for data visualizations will be updated later as they work on their project.

Week 6 will be for Data Reporting and Presentation. The findings of the project will be presented using plots, interactive visualizations and reports. The final project will then be presented to the career service team, the product owner, potential employers, and other interested parties. Each team member will get a chance to speak during the presentation and elaborate on what choices were made regarding the work, why they were made, and what the subsequent outcome was.

# Other requirements

All programs used should be free of charge. However, some group members may opt to utilize a paid version of any of the statistical or data visualization programs.

# Assumptions

The software programs and platforms that the team members use should be available, up-to-date, and not broken.

Team members will source data from reliable sources including but not limited to Google Scholar, Kaggle, Data.gov, Data.world and research gate. The team is expected to refer to their previous schoolwork and assignments for better practices during the project. The team is also expected to update their progress on the team’s Github repository for the instructor to track their work and the team to keep up with their progress.

The team is also supposed to update their Trello account for better project management practices.

# Limitations

The largest disadvantage that may be faced by the team will be time zone differences. Also, both team members have conflicting work schedules.

Constant communication will be the key to the success of the project.

# Risks

Some of the possible risks that may arise would most likely be natural occurrences/calamities, family emergencies, work scheduling conflicts and time zone differences. The team members will work together and always communicate with each other in case any of the above-mentioned risk interferes with delivering of the project on time.