**Data Science Program Final Project**

**Executive Summary**

At the conclusion of the ENTITY Academy/ Woz U Data Science bootcamp program, all prospective graduates must complete a final project on the topic of their choice. This project involves six structured lessons that correlate with a given six-week timeline, at the end of which the students must give their final presentation. A student may work alone or choose to work in a group with other students. This document is dedicated to the conclusive work of Natalin Williams Brady, Virginia Mburu Brovont, A. Nicole Mosley, Ivette Sierra, and Lucy Thuku (“The Fab 5 Data Scientists”). It will explain the purpose and scope for their chosen work which asks the question: “*What factors influence food insecurity in the United States?*”

**Business Objectives**

The team members will demonstrate skills they have acquired throughout the Data Science program. We will be using Python, Tableau, Machine Learning and any other relevant programs to wrangle, analyze, and visualize our “Food Insecurities” dataset made available by various databases from sources including but not limited to the World Bank, the U.S. Department of Agriculture (USDA), and Feeding America.

Food insecurity is defined as a lack of consistent access to provide enough food for every person in a household to live an active, healthy life; this can characterize a temporary situation for a specific household or one which lasts over a longer period of time.[[1]](#footnote-0) Food insecurity is one way we can measure how many people cannot afford to purchase food. According to the USDA, in 2020, 10.5% of households experienced food insecurity.[[2]](#footnote-1) Some of the contributing factors to food insecurity include poverty, unemployment, the pricing of food and other commodities over time, proximity to urban centers, ethnicity or race, and the number of people in a household. Although we will not be examining the direct data behind other factors such as chronic health conditions, a lack of access to healthcare, and difficult financial decisions for the average American consumer (such as choosing between paying for one’s food, heat, electricity, rent, or transportation), these facets also shape the dynamics of food insecurity at all levels of society.Food insecurity is a condition that leads to hunger. Working families across America face countless situations that can result in food inse

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**

In the process of selecting a subject for our respective presentation, we found it imperative that all of the individuals in our group demonstrated a personal interest in the topic of discussion. Initially, during the brainstorming process of our project, one of our team members, Natalin Brady, came up with the idea of researching the concept of food insecurity in the United States. Natalin took an interest in this area due to her experience in the pharmaceutical industry. Due to the recent covid-19 pandemic, her industry saw a rapid decrease in the number of patients who were able to purchase much-needed medications from their providers. As a result of high unemployment and increasing inflation, many Americans have seen a drastic change in their quality of life especially with regard to their ability to provide basic necessities – such as **food** – for their families. Natalin’s background as a single mother of a teenage son and the daughter of an aging parent has particularly made these national crises relevant for her. The same can be said for the other individuals in the group: Lucy, Virginia, Ivette, and Nicole, all of whom have seen their livelihoods affected in varying degrees by the increasing financial difficulty caused by the pandemic. Each member of the group has personally known or heard of individuals and families who are being adversely impacted by this nationwide crisis. Given that hourly wage growth in the United States has remained stagnant and uneven over the last four decades, inflation has put a dent in the wallet of many lower to middle-income families as they seek to consistently secure and maintain basic necessities such as food, clothing, and housing for their households.[[3]](#footnote-2) All of us hope to gain greater insight from our subsequent research to bring greater awareness to the prevalence of food insecurity throughout the United States and understand the various dynamics that contribute to this phenomenon at the state level.

**Scope**

Team members will be utilizing the data science and statistics software taught in the program to complete the project. They will be intentional about using any tools of interest or resources (i.e. databases, infographics, official government websites) that will aid in this endeavor and which may aid in locating a job. Additional software, tools, and outside resources may be incorporated as desired, even if not required.

Team members will be using the software introduced in the program to complete the project. They will be intentional and purposeful about using tools of interest and any additional resources that may aid them in securing a job. Incorporating additional software or tools is optional, but that is not required.

**Functional requirements**

Data Collection: In order to ensure that project has a proper scope and thorough analysis, an adequate range, variety, and number of credible datasets must be acquired so that team members can fully and accurately examine the question: “Which factors influence food insecurity across the United States?”

Data Wrangling: Data will be manually converted and mapped from a raw form into a more convenient, organized, and consumable form. The downloaded datasets should be successfully cleaned and fitted for analyzing. Columns and unusable columns should be removed. As the datasets may be fairly large, the team’s data wranglers 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 wrangling tools should include R and Python.

Data Analysis: Statistical and logical techniques will be applied to describe, condense, recap, and evaluate data, thus providing a way for team members to draw inductive references and distinguish the signal (phenomenon of interest) from the noise (statistical fluctuations) that exist in the data.[[4]](#footnote-3) All team members will familiarize themselves with the datasets. They should have a good understanding of what each column means, how the values are measured, what each indicator is representative of, the source of each dataset, and so on. They will brainstorm on questions to ask, and infer what overall insights they might gain from the dataset. Then, they will identify the proper functions to create models, predictions, etc. Data analysis tools should include R and Python.

Data Visualization: Once the team members have a comprehensive understanding of the information gleaned from the dataset, they will work on visualizing the findings that most stand out in their research and which best demonstrate how the data answers their questions. They may decide to use Tableau or other graphing programs, and compile the visuals and texts in a PowerPoint slideshow.

Presentation: Working with the WozU instructors and mentors, all team members will schedule a time to present their findings via Zoom. They should be able to communicate in a clear, concise, and easy-to-understand manner; basically, in layman's terms. The presentation should have a duration of approximately 20 minutes. All group members should be dressed professionally for this occasion.

**Personnel requirements**

The project will require a team of 5 data science student researchers and 1 product owner (group instructor).

The estimated time for the project to be completed will be 6 weeks with an average of approximately 20 hours per week per person. Each week, a new sprint begins on Monday and ends on Sunday, when the Retrospective Meeting is held.

There will be regular weekly meetings that will be led by a Scrum Master. The weekly group meetings will be held every Sunday at 5 P.M. CST. There will also be weekly 1-on-1 sessions with the product owner which will be held every Wednesday at 7.30 PM CST.

There will be a daily (or weekly) stand up hosted by the Scrum Master for each respective week. The Scrum Master should be elected at the end of a sprint on a rotation basis until each member of the team has held the position.

**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, and choosing the method of analysis.

Week 2 will be for Data Wrangling. After gathering our data, there will be a bit of reformatting that will need to happen before we 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). We 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. We 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 our analysis, the techniques that we will use for data visualizations will also be updated later as we work on our project but it is most likely that we will utilize Tableau as our primary tool for data visualization.

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**

Team members will source for data from credible sources including but not limited to Google, Google Scholar, Kaggle.com, Data.world, and official government websites. It is expected that major and vital documents on the team’s progress will be regularly added to the team’s Github repository in order to accurately reflect our current and up-to-date work. Any specific software programs and platforms utilized should be widely available to the public, up-to-date, and not broken. The Google Docs website will also be used as a solid backup platform in the event of a loss of crucial documents and work from other websites (such as Github).

**Limitations**

The primary limitation that we may face as a group may be overcoming scheduling conflicts due to the fact that we reside in different geographic areas and thus live in different time zones. Consequently, the work schedules of various group members may conflict often, and some of us may face occasional and unforeseen changes in our respective work schedules. The discrepancies in our respective time zones may affect our ability as individuals to complete and deliver assignments, if any obstacles arise that require immediate intervention. Adaptability, flexibility, and openness will be needed. Lastly, we may require help from our program instructor when he is unavailable, a complication which could considerably slow down our work.

**Risks**

Possible risks and complications that may arise would most likely include family emergencies, illness experiences by a group member, natural events/calamities, power outages, work scheduling conflicts, and/or slow, poor internet connection. Should any risk arise in the foreseeable future, any team member affected should promptly and urgently communicate the details of their situation to their Scrum Master and teammates so that they can best accommodate the needs of the project. Any tasks that cannot be completed in a timely manner should be brought to the Scrum Master’s attention as soon as possible to facilitate the efficiency of the project timeline.

1. *What is Food Insecurity?* (n.d.). Feeding America. Retrieved February 20, 2022, from https://www.feedingamerica.org/hunger-in-america/food-insecurity [↑](#footnote-ref-0)
2. USAFacts. (2021, November 23). What is food insecurity and how does the government combat it? *USAFacts*. https://usafacts.org/articles/what-is-food-insecurity-and-how-does-the-government-combat-it/?utm\_source=google&utm\_medium=cpc&utm\_campaign=ND-Education-Childcare&gclid=CjwKCAiA6seQBhAfEiwAvPqu188nJASL\_XV-CuaRtrT1XOfk-8CiK7bsS0kMvGCmAGHiGhufWougjxoCNEAQAvD\_BwE [↑](#footnote-ref-1)
3. Gould, E. (2020, February 20). *State of Working America Wages 2019: A story of slow, uneven, and unequal wage growth over the last 40 years*. Economic Policy Institute. https://www.epi.org/publication/swa-wages-2019/ [↑](#footnote-ref-2)
4. ORI - The Office of Research Integrity. (n.d.). *Responsible Conduct in Website Management: Data Analysis*. Retrieved February 20, 2022, from https://ori.hhs.gov/education/products/n\_illinois\_u/datamanagement/datopic.html [↑](#footnote-ref-3)