

Progress report

Project Scope Update

This project studies if there is a link between 'how long babies are breastfed' and 'getting gluten problems like celiac disease'. The goal is to see if breastfeeding can help protect children from gluten intolerance later in their life. The main idea and plan for the project have stayed the same since the beginning.

Data Sources

Three groups of data are used in this project:

- **HealthData.gov Breastfeeding Data:** This online data shows how many mothers breastfeed, for how long, and information about mother's age, race, and where they live. The data is downloaded using website tools (API) and is saved as a table (CSV) for studying.
- **Kaggle Celiac Disease Data:** This data shows which people have gluten problems and have medical details about them. It helps measure how common gluten illness is. The data is downloaded as a table (CSV) from Kaggle website.
- **Kaggle Infant Feeding Data:** This table shows what kinds of foods are given for babies (only breast milk, mixed, or only formula) and how fast they grow. It helps to compare feeding style and growth. This data comes as a table (CSV) from Kaggle.

Problems and Difficulties

- **Cleaning Data:** Some tables do not have all the same columns. So matching data about people or groups of people is making it hard. There are also missing values, and some data uses words instead of numbers, so they need to be changed.
- **Combining Data:** Connecting three different tables from websites and files is hard work and requires very careful work.
- **Potential issues:** If there are too few records for some groups, it will be hard to make good charts or conclusions.