DSCI 510 – Fall 2025 – University of Southern California

Instructor: Dr. Alexey Tregubov

Final Project Proposal

Samira Kazimi
USC ID: 5473373539
Project Title: Correlation Between Breastfeeding Practices and Gluten Sensitivity Risk

---

Project Description:

This project investigates whether early breastfeeding duration and exclusivity are correlated with the risk of developing gluten sensitivity or celiac disease later in life. It integrates multiple datasets to examine health trends across demographic and nutritional variables.

The study will combine three data sources to ensure diversity and reliability of the analysis:

1. **HealthData.gov API:** Provides breastfeeding and maternal health data by state and demographic group (CSV via API).
2. **Kaggle – Celiac Disease (Coeliac Disease) Dataset:** Contains gluten-sensitivity data and biomarkers (CSV).
3. **Kaggle – Weight/Height and Breastfeeding Pattern of Infants Dataset:** Captures infant feeding patterns and growth metrics (CSV).

Python libraries such as **pandas**, **NumPy**, and **seaborn** will be used to clean, merge, and visualize data. Correlation coefficients and heatmaps will be generated to identify relationships between breastfeeding practices and gluten-related health outcomes.

The expected outcome is a statistically supported insight into whether early breastfeeding influences gluten-sensitivity development, contributing to evidence-based public health discussions on infant nutrition.

Data Sources Table:

Data source name	Short description	Source URL	Type	List of fields	Format	Tried collecting with Python	Est. size / points
HealthData.gov Breastfeeding Data	Breastfeeding and maternal health data by state	https://healthdata.gov/api/views/x7kq-cyv4/rows.csv	API	Various demographics, states	CSV	Yes	~1000+
Kaggle Celiac Disease Dataset	Gluten sensitivity and biomarkers	https://www.kaggle.com/datasets/jackwin07/celiac-disease-coeliac-disease	File	Gluten sensitivity markers	CSV	Yes	2206
Kaggle Infant Feeding Patterns	Infant feeding types and growth metrics	https://www.kaggle.com/datasets/chidirolex/weightheight-and-breastfeeding-pattern-of-infants	File	Feeding type, growth measurements	CSV	Yes	1000+