- **Title:** Trends and Patterns in Instacart Orders

- **Team members**: José Gutiérrez, Keaton Hoeger, Anna Malawista, Aidan O'Connor, Michael Whitlock

- **GitHub Repository:** https://github.com/MichaelWhitlock/CSCI4502Group1.git

- **Description:** We would like to analyze patterns in the orders placed to the company Instacart. We plan to determine if there are trends in the types or rate of products ordered depending on the time of day, week and frequency of repurchases. Ideally, we would also like to implement some simple machine learning to predict future order activity. Potential questions include: "Which items are most likely to be added to an order, based on the items already in the cart?"; "Can we predict the length of time between orders for specific customers?"; "Based on purchase history, can we predict which new items a customer is most likely to purchase?"

- **Prior Work:** Instacart posted the dataset on their website and a brief description on Medium:

https://www.instacart.com/datasets/grocery-shopping-2017

https://tech.instacart.com/3-million-instacart-orders-open-sourced-d40d29ead6f2

We have found that most of the prior work on this dataset is exploratory in nature. The following two links are representative of the type of work we've come across:

https://nycdatascience.com/blog/student-works/capstone/instacart/

https://www.kaggle.com/philippsp/exploratory-analysis-instacart

Instacart's Medium post also includes two short examples of patterns derived from the dataset.

- **Datasets:** We will be using the Instacart dataset (made up of 6 tables/files) from this link: https://www.instacart.com/datasets/grocery-shopping-2017 . Each team member has downloaded the dataset.

- **Proposed work**: The dataset is very clean and simple to follow. The 6 tables available are linked by ID attributes for the orders, departments, etc. We are joining the tables by these attributes, so we can fully analyze all the order information available. We have 3.2 million data points in total for analysis. There are pre-designated subsets of data for training and testing machine learning models.

- **List of tools** We intend to primarily use python tools such as Pandas, NumPy, MySQL, and Scikit-learn.

- **Evaluation:** We want to try using as many of the following analysis methods as can be reasonably applied: data visualizations (box plots, scatter plots and trend lines, histograms, etc.), data analysis (correlation coefficient, chi-square test, and other correlation statistics, association rule), classification and prediction with machine learning. The dataset includes a test/train split, enabling us to evaluate prediction and classification models. We will continue to modify and add to our intended evaluation methods as we learn more throughout the course.