Group 2:

Team Members: Omar Castillo, Velmarini Vasquez, Oleg Mironov

Proposal:

Sources:

1. API- FDA Food
2. TSV File- Open Food Facts (Kaggle.com)
3. Adverse Food Events(Kaggle.com)

ETL Overview:

1. Extraction of FDA Data:
   1. We began by connecting to the **FDA Food API Endpoints**, where we recalled food enforcements reports from the time period of 2010 thru 2019.
   2. From the initial pull, we focused on extracting FDA column data **‘recalling\_firm’, ‘product\_description’, and ‘reason\_for\_recall’.**
2. Extraction of Open Food Facts CSV:
   1. Open Food Facts gathers information and data on food products from around the world.
   2. Our initial TSV file from Kaggle.com had (163) columns worth of data. Our final data frame was composed of (4) column categories **(‘product\_name’, ‘brands’, ‘ingredients\_text’, and ‘country’**). Lastly, we filtered by country, focused only on US products.
3. Extraction of Adverse Food Events CSV:
   1. The CFSAN Adverse Event Reporting System(CAERS) database was used to pull a CSV file that contained 90K reactions that were recorded from 2004-mid 2017, with 12 columns of information regarding type of reaction and related event details
   2. Our final data frame extracted and renamed (4) column categories called **‘brand’, Industry, ‘Adverse Reaction Outcome’, ‘report\_date’.**

Why This Data?

For this project, our team is using the sources listed above to store food recall information in a SQL Database to be used for the following analysis:

1. Are there ingredients in certain products that are popular amount recalled food?
2. For potential investors, how safe would it be to invest in a potential product with specific ingredients?
3. For doctors, what type of recalled food has been shown to cause issues in the past?
4. What areas are most effected by a particular outbreak or product recall?
5. Which companies or firms that have reported to the FDA can be found in the New York Times Article?