Group 2:

Team Members: Omar Castillo, Velmarini Vasquez, Oleg Mironov

Proposal:

**Sources:**

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

**ETL Overview:**

1. Extraction of Original Data:
   1. We began by connecting to the **FDA Food API Endpoints**, where we recalled food enforcements reports from the time period of 2013 thru 2019.
   2. Open Food Facts gathers information and data on food products from around the world. Our initial TSV file from Kaggle.com had (163) columns and over (300k) data points.
2. Transformation of Data:
   1. From the initial pull, we focused on extracting FDA column data **(state, city, report\_date, product\_description, product\_quantity, reason\_for\_recall, brand, recall\_number)**.
   2. We combined our API responses from the FDA Food API Endpoints, and we removed redundancies by creating new relational tables (‘state and state id’, ‘city’, ‘reason for recall’, and ‘brands).
   3. Within our TSV file from Open Food Facts we filtered by country (US), and dropped all null records. We then established a relationship with **brand\_id** table by replacing a brand name with corresponding ID index. Our final table columns included (**brand\_id, product\_name, ingredients\_text).**
3. Loading of Final Data:
   1. For our ETL project, we decide to use MySQL database because we can remove redundancies and establish relationships between our final (6) tables.
   2. We also decided to use MySQL in order to directly correlate recalled products with popular products to other dataset.
   3. We connected to local instance of MySQL in order to create a database called **fda\_db.**
   4. We loaded all of the tables inside **fda\_db.**
   5. Our loaded dataset doesn’t have direct correlation between tables. All relationships have been established within MySQL manually.

**Why our data?**

1. FDA Food API: The FDA provides our database a reliable source of live data that is updated on a weekly basis. Through the API connection, we can maintain more accurate and trusted government information over food recalls, the responsible firms, as well as which of their products are being recalled.
2. Open Food Facts: While our FDA Food API provides us a live source of data with a high-level description of the products being recalled, the Open Food Facts csv from Kaggle provides us a large source of static information regarding ingredients within those products that may be recalled. This is an added level of detail from our initial FDA Food API, as we can further analyze the specific ingredients within the products that are being recalled.
3. Our two dataset focuses on brand and their products that have a history of being recalled. We can compare recalled products with successful product and their risk in terms of financial investment and payoff.

**What’s our use case?**

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? If investors only see recalls from certain products, they might re-consider investing in those products.
3. An analysis that can be born from the database we have created, is further analysis of the stock market prices for companies based on the amount of recalls they have.
4. For doctors, what type of recalled food has been shown to cause issues in the past?
5. What areas are most effected by a particular outbreak or product recall?
6. Which companies or firms that have reported to the FDA can be found in the New York Times Article?