**ETL Project Report**

Project 2 – Dec 22, 2018

Coel Locandro and Haley Disterhoft

**Extraction:**

From Kaggle.com, we downloaded several datasets from the competition “Instacart.” Instacart is a grocery app that allows consumers to purchase food products. They have datasets that reflect what products they sell and how many times each product was purchased. Our goal was to see the top 50 vegetables that were ordered through Instacart.

**Transformation:**

First, we loaded the csv files for the product names and product categories into Pandas. We merged these files and then, in order to filter out products that weren’t vegetables, we scraped a list of vegetables from the site [www.vegetablesfruitsgrains.com](http://www.vegetablesfruitsgrains.com). We filtered our original list of products with our list of vegetables, searching for every product that contained the name of the vegetable in the name of the product. We were able to do this by running a for loop and by splitting the name of the products into separate words.

Once we had our list of Instacart vegetables, we merged this data with the Orders\_Prior dataset, to find out how many customers ordered each vegetable. We ran a values\_count on the merged data, and reduced it to the top 50 Vegetables.

At this point we had an issue with our index, because Pandas wanted our list of vegetables to be our index. We were able to reset our index, and we relabeled our column headings to accurately reflect the data.

**Loading:**

After we had our new Top 50 Veggies data frame, we sent the table to MySQL. We created a database called ETL\_Project through PyMySQL, and after troubleshooting several errors that mostly had to do with syntax, we were able to send our Top 50 Veggies table to MySQL, which makes it a relational dataset.

If any future user wanted to find out the order counts for every vegetable, we also included that data frame in our original code.

Future users could also go back and search order counts for other products, like fruits (within the produce category), or even things like snacks or frozen items. It’s an extremely large collection of data, so the opportunities are plentiful.