**Instacart Product Analysis Project**

We discovered our dataset on [Kaggle.com](http://www.Kaggle.com) through a competition Instacart was sponsoring. Instacart is a grocery app that allows consumers to remotely purchase groceries online, or through an app on their mobile device. The dataset reflects food product data sold through Instacart, including, but not limited to, quantity, type, and price. Our main goal was to identify the top 50 vegetables that were ordered through Instacart and identify any trends.

To begin, 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.

After we had our new Top 50 Veggies dataframe, 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, making it into a relational dataset. Our first trend popped out almost immediately, which was that vegetables labeled “organic” were the most preferred. As shown below, this was identified due to the fact that the top five most ordered vegetables were all organic.

|  |  |  |
| --- | --- | --- |
|  | **vegetables** | **order count** |
| **0** | Organic Baby Spinach | 241921 |
| **1** | Organic Hass Avocado | 213584 |
| **2** | Organic Avocado | 176815 |
| **3** | Organic Yellow Onion | 113426 |
| **4** | Organic Garlic | 109778 |

If any user wanted to view the order counts for every vegetable, we also included that dataframe 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.