**Project #2 ETL: Cereal Final Analysis**

Contributors: Grace Cote, Natalie Lassen, and Jessica Johnson

***Presentation:***

<https://docs.google.com/presentation/d/1pPoYYNaMPpD-_SVLiaH54SJVF9U9-3hz4itVURIAGEQ/edit?usp=sharing>

***Github*** ***Repo:***

[GitHub - natlass/ETL-Project-Cereal: A project exploring extracting, transforming, and loading data using various platforms, focusing on cereal brands.](https://github.com/natlass/ETL-Project-Cereal)

***Data*** ***Sources:***

StatInvestor.com (https://statinvestor.com/data/5322/ready-to-eat-cereal-brands-market-share-united-states/)

Data.World

(https://data.world/makeovermonday/2020w36)

***Extract****:*

For this project, we drew our data from two sources online. The nutritional data was extracted directly from “Data.world” as a CSV file. The market share table was scraped using an Excel tool, from StatInvestor.

***Transform:***

Our group started forming our project by first renaming all of our columns when creating the tables in Pandas. Since the nutritional data was formatted with brand name and cereal name in separate columns but the market share had both as one attribute, a new attribute needed to be created in order to join the two datasets properly. Before this was done, we changed the brand “Kelloggs” in the nutrition table to “Kellogg’s”, adding an apostrophe, so it would match the name in the market share table. In addition, we changed the cereal name “Frosted Mini-Wheats” in the nutrition table to “Frosted Mini Wheats”, removing the dash, so it would match the name in the market share table. Once all brands and cereal names matched in formatting, we could create a new field in the nutrition table, which consisted of the brand and cereal name. This allowed us to join the two dataframes.

***Load:***

After formatting and joining the dataframes, we created a database and uploaded it to Postgres. We did this through a simple line of code in our Jupyter Notebook, creating a database connection and engine connection. After establishing our database, we created the tables that the CSV files would export to, and connected these tables to our pandas dataframes. Then working through Postgres, we did an inner join on the two tables grouped by the cereal and manufacturers name, and called this new joined data table “full.” From here, we could query on the combined data, and view correlations between things such as rating and market shares. We chose a SQL database with Postgres since it would store all of our data easily, and would work best for the queries that we believed would be necessary with the data after loading.