Investigating the Prevalence of Corn Syrup and Sugar in Brands and Food Products

CS 5010 – Section 1: Professor Judy Fox

Team 6: Nitika Kataria (nk3rf), Robert Knuuti (uqq5zz), Swaroop Veerabhadrappa (sbv5dn)

May 5, 2021

Introduction & Motivation for Analysis

- Brands today have direct influences on our health with the way they have composed our food. One particular concern that's been in the front of many health conscious individuals is the prolific usage of corn syrup and sugar in many food brands and their products.
- We seek to identify which brands produce the greatest number of products that use corn syrup and sugar and perform analysis on those brands.
  - Established a ranking system based on their usage of corn syrup and sugar from their ingredients list.
  - Identified how frequently corn syrup and sugar are used in products of similar categories of food.

Data: Overview, Sources

fdc_id	ID of the food in the table
brand_owner	The brand owner of the food
gtin_upc	GTIN or UPC code identifying the food. Duplicate codes identify an update to the product, use publication_date found in the food table to distinguish when each product was updated.
ingredients	The list of ingredients (as it appears on the product label
serving_size	The amount of serving size when expressed in as grams or ml
serving_size_unit	The unit used to express the serving size (gram or ml)
household_serving_fulltext	Amount and unit of serving size when expressed in household units.
branded_food_category	The branded food category, assigned by GDSN or Label Insight
data_source	The source of the data for this food GDSN (for GS1), or LI (for Label Insight)
modified_date	The date reflects when the product was last modified by the data provider.
available_date	This is the date when the product record was available for inclusion in the database.
market_country	The primary country where the product was marketed

Data: Experimental Design

1. Pull the data from the USDA website and extract the data to a working directory (requests)
2. Import the brand data into a table for analysis (pandas)
3. Insert a new column indicating the presence of corn syrup and sugar in the ingredients list (pandas)
4. Insert a new column indicating the position of corn syrup and another column for the position of sugar in the ingredients list (pandas)
5. Group the branded items based on the position of corn syrup and sugar (pandas)
6. Produce a visualization showing how many products a company has listed which includes corn syrup and sugar, scaled by the position of corn syrup and sugar in the ingredients list. (matplotlib / seaborn)

Data: Exploratory Data Analysis

- Ingredient: Corn Syrup
  - Number of Products (rows): 96,002
  - Number of Brands: 6,126
  - Unique Ingredients: 56,928
  - Number of Food Categories: 185
  - Number of unique index position values for Corn Syrup: 105
  - Index position:
    - Mode: 0 and 2 (bimodal distribution)*
    - Mean: 9
    - Median: 5

- Ingredient: Sugar
  - Number of Products (rows): 246,773
  - Number of Brands: 17,782
  - Unique Ingredients: 156,310
  - Number of Food Categories: 218
  - Number of unique index position values for Sugar: 100
  - Index position:
    - Mode: 0 and 2 (bimodal distribution)*
    - Mean: 6
    - Median: 3

Data: Cleanup

- Raw data set contains:
  - 498,182 rows (All Food Products)
  - 13 columns (Variables/Attributes)
  - 4 columns with incorrect data type
  - 60,268 Missing Data Points
  - No Duplicate Records

- Cleansed data set contains:
  - Food products containing Corn Syrup: 96,002
  - Food products containing Sugar: 246,773
  - 12 columns (after dropping one empty column)
  - Correct data types for columns set:
    - 3 columns → datetime
    - 4 columns → categorical
  - 904 Missing Data Points

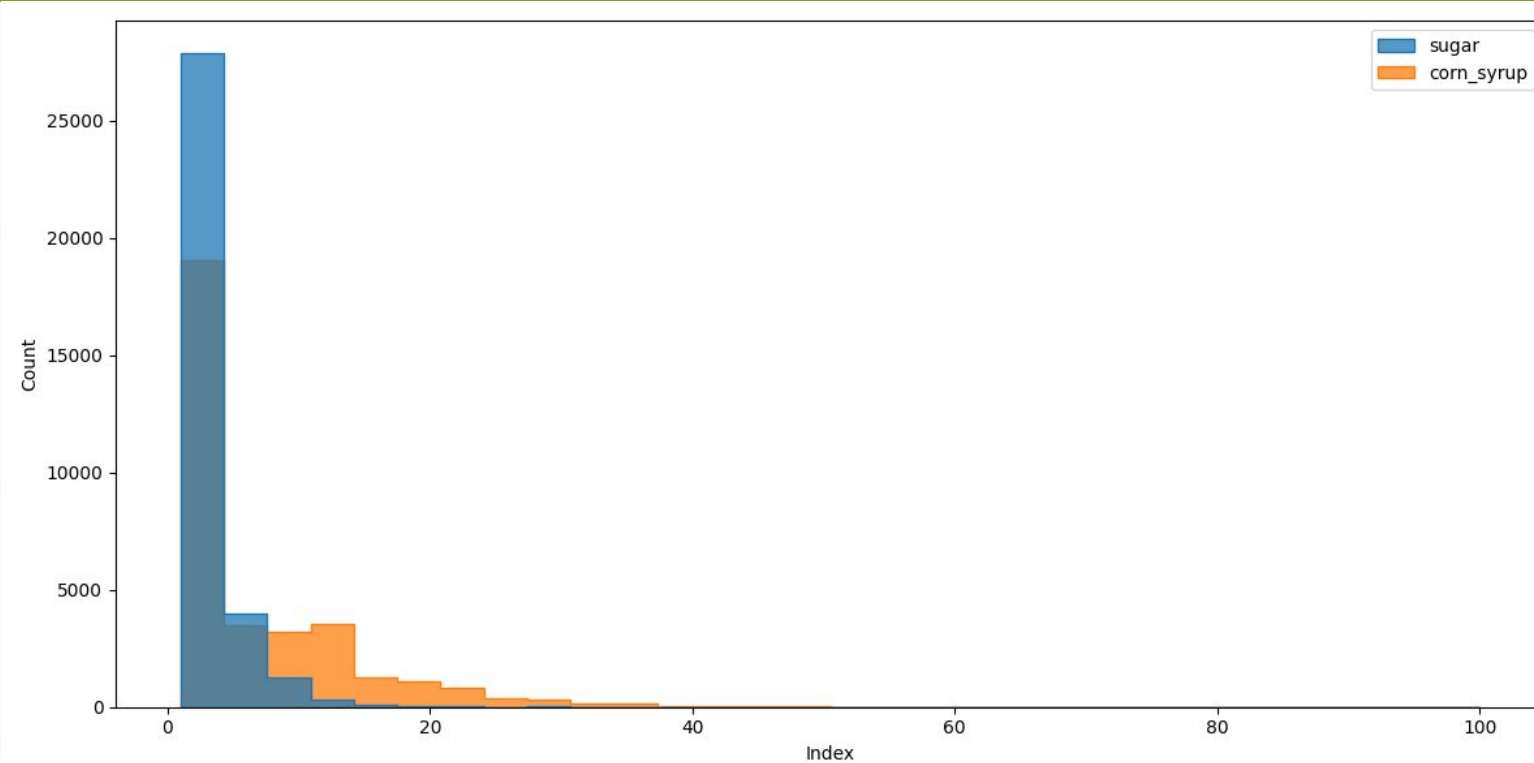# Corn Syrup & Sugar Usage in the Top 10 Brands
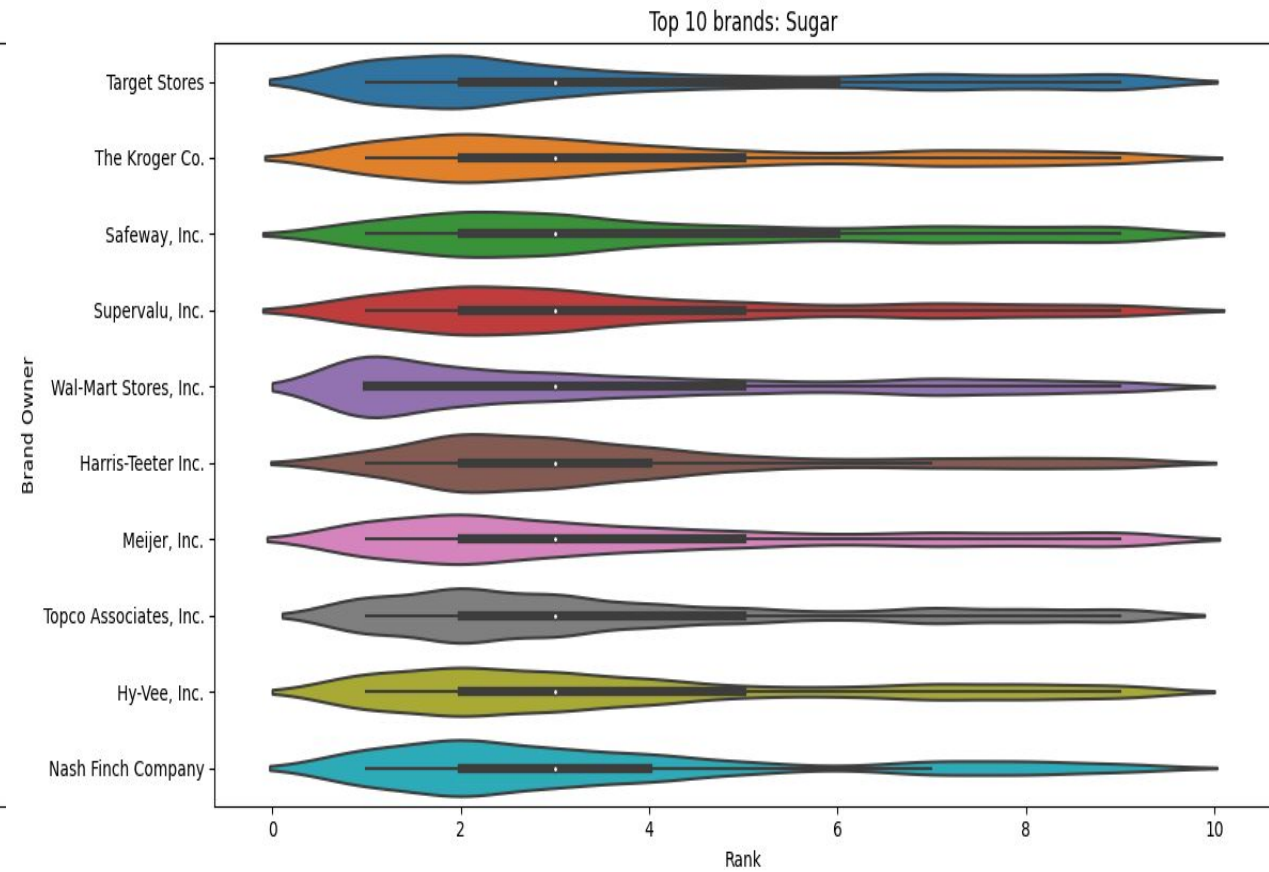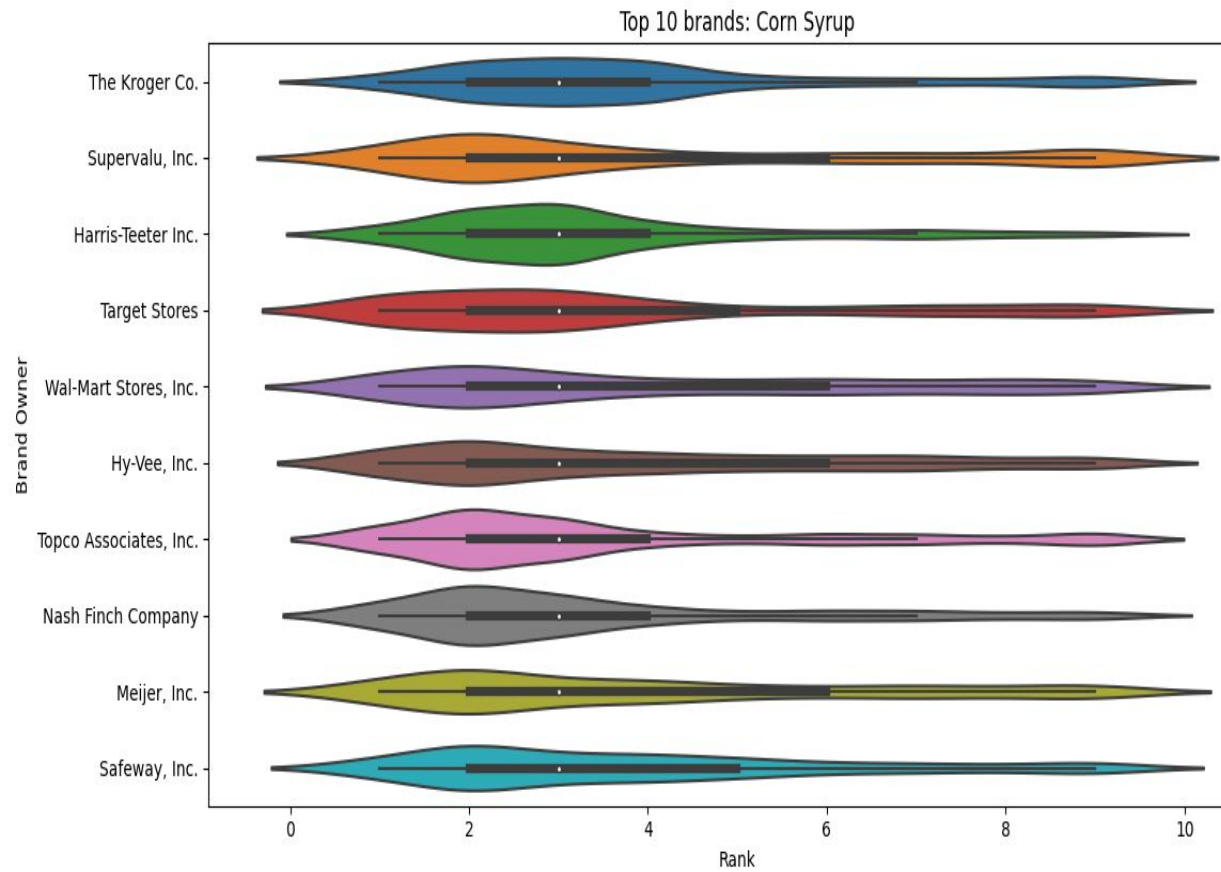
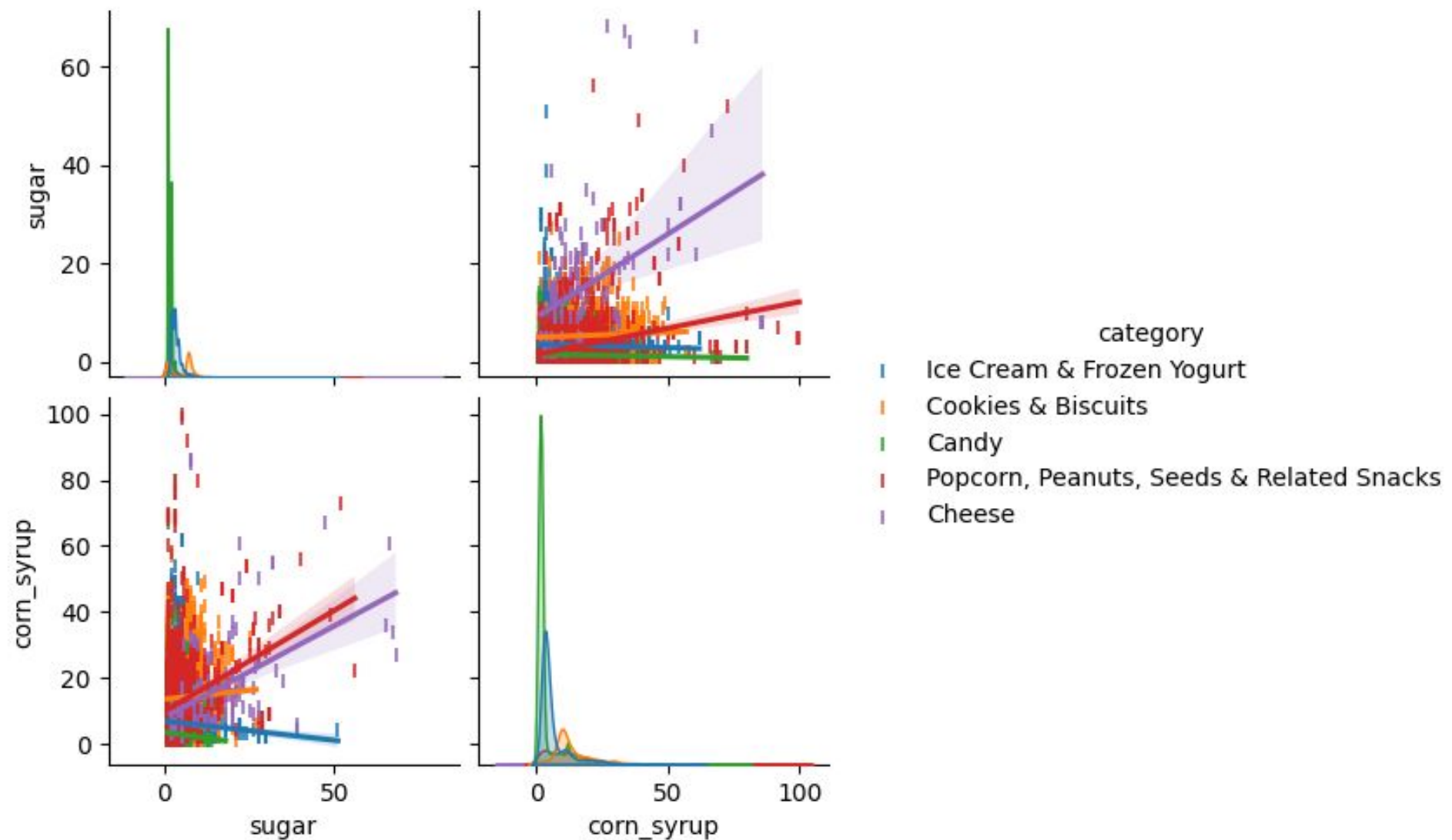# How much is corn syrup used in the top 5 food categories?

- 239
  Total Categories
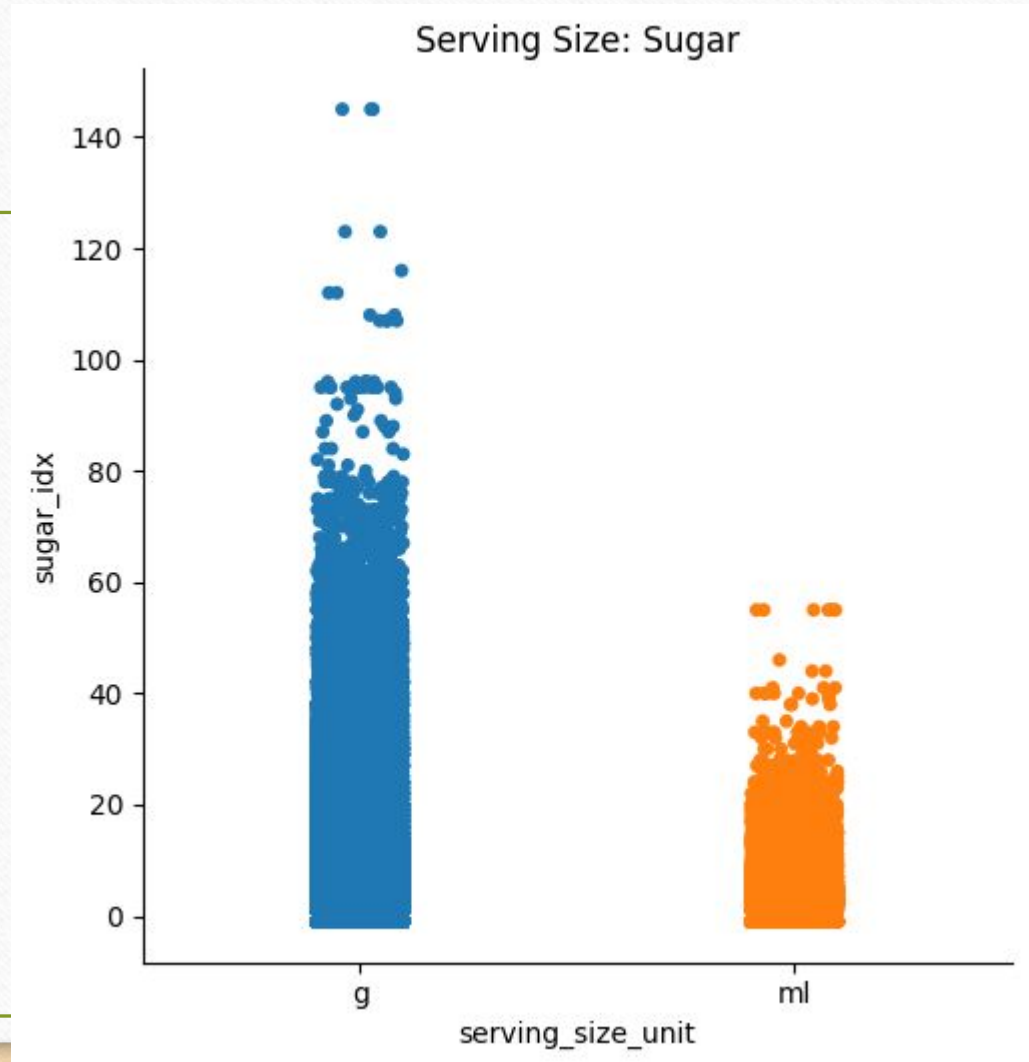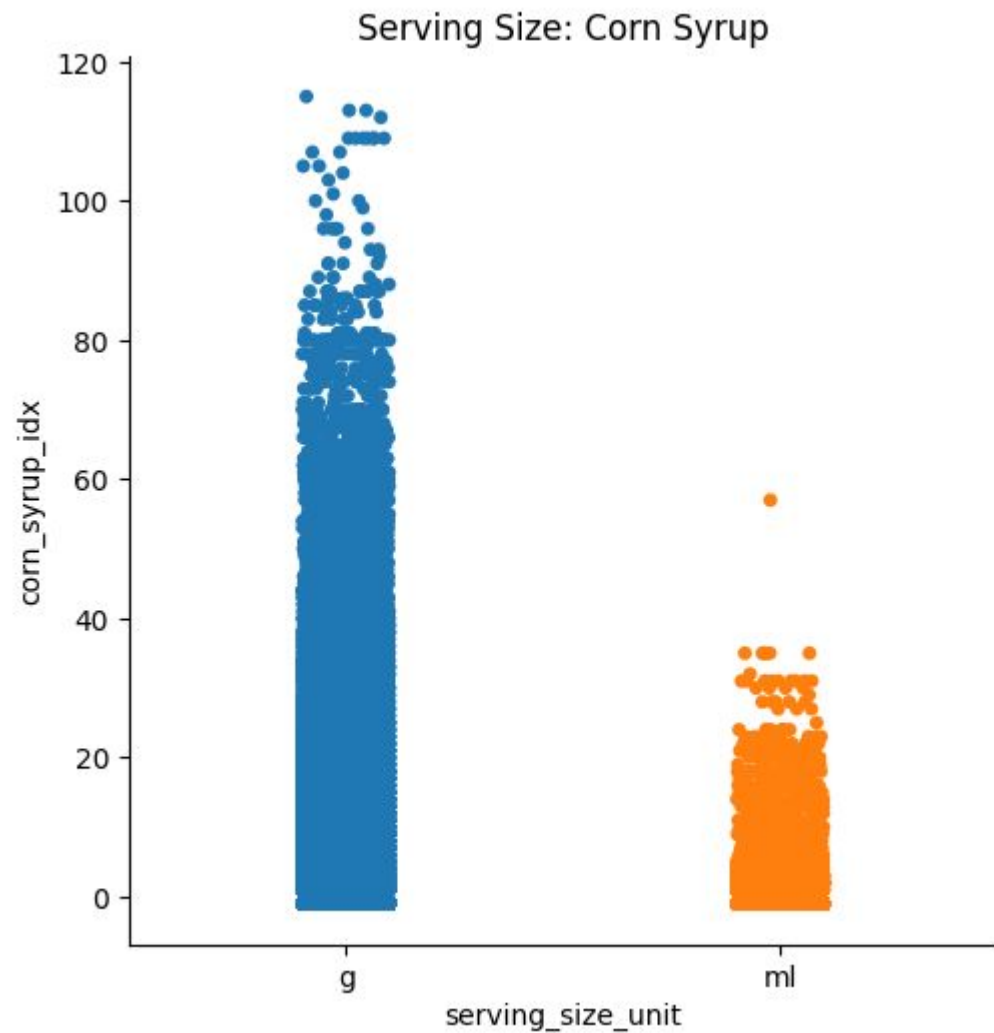
- 92,420 Food Items from top 5 subset
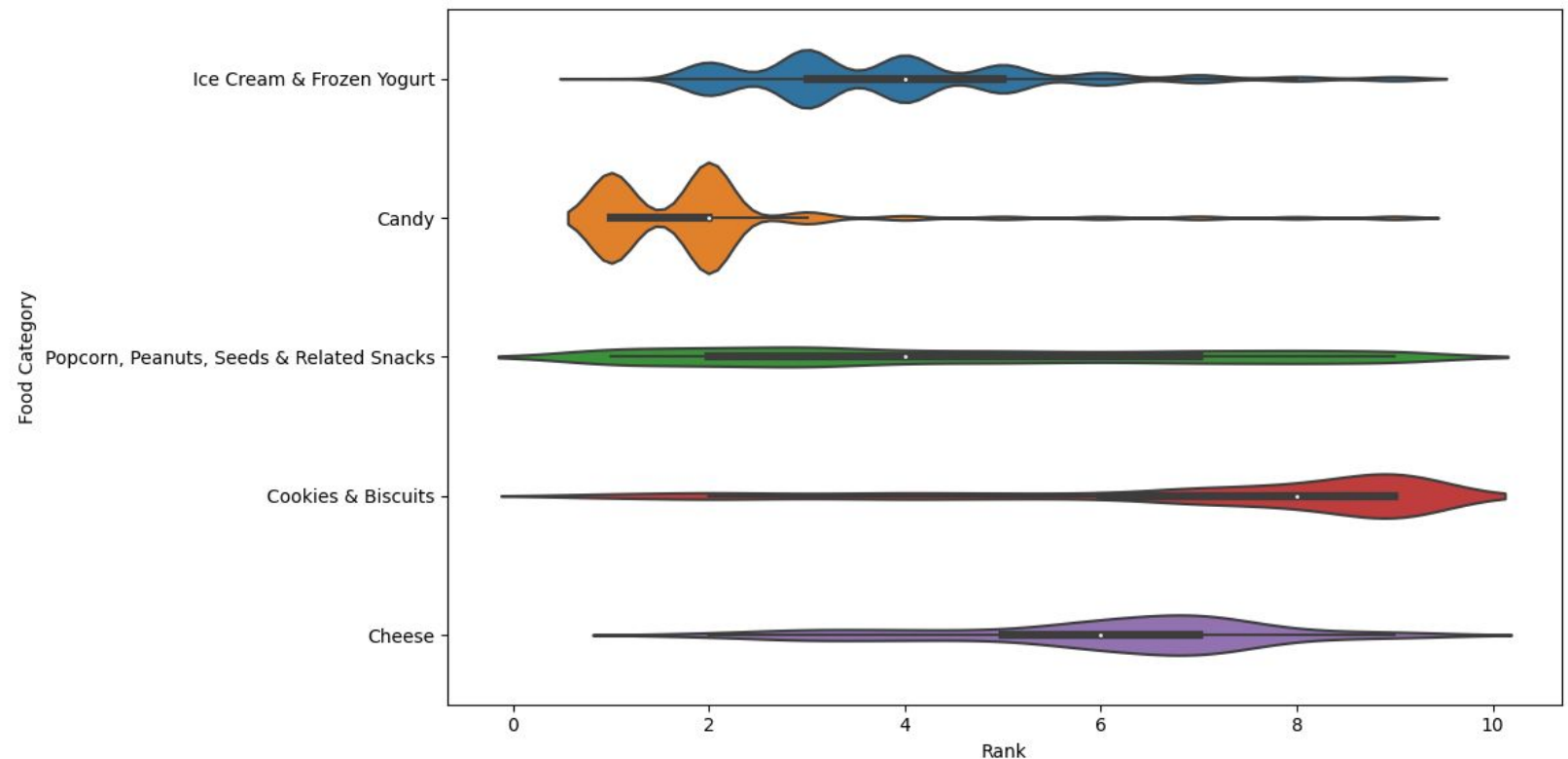
Top 5 Density of Corn Syrup vs Sugar

Correlations between Corn Syrup and Sugar

Solid Food Items vs. Liquid Food Items

Above and Beyond

- Project was built as a python module using poetry, the modern alternative to setuptools.
- Integrated Sphinx for API documentation
- Automated download and extraction of data from UDSA's website

Project Summary & Next Steps

- Nearly 25% of all food products examined contain corn syrup and 50% contain sugar
  - Amount of each ingredient is highly dependent on brand, food category, and product
  - There's no correlation between Corn Syrup and Sugar

- Ingredients can be misleading
  - Sweeteners: Corn Syrup vs. Sugar
  - Over 200 different names for sugar (maltodextrin)

- Given the importance of knowing what ingredients are in the food products we consume, we recommend additional analysis be performed, including the following:
  - Exploring the prevalence of disguised ingredients such as maltodextrin
  - Conducting time-series analysis would allow us to examine how recipes (or combinations of ingredients) have changed over time

Resources

- Third Party Python Libraries
  - Pandas: https://pandas.pydata.org/
  - Pytest: https://docs.pytest.org/en/6.2.x/
  - Requests: https://docs.python-requests.org/en/master/
  - Seaborn: https://seaborn.pydata.org/
  - Sphinx: https://www.sphinx-doc.org/en/master/