Prince Amponsah, Eliot Cleveland,

Nick Hancock, Katie Blood

Abstract

Is there a preference of Coke vs Pepsi to a consumer? This ETL project looks at sales reports for both companies, each for one year. These were extracted, transformed, and loaded (ETL) in order to perform different inquiries.

pepsi vs. coke:

soda wars

[Beverage Sales Technical Report]

**Introduction**

The company we are contracted for has recently switched its primary beverage vendor from Coco-Cola Company as its primary beverage vendor to Pepsi. This provided an opportunity to compare sales between these rival companies using real world data. The data that will be compared is from Fiscal Year 2019-2020 and Fiscal Year 2020-2021. In consideration of the effect COVID had on businesses during this time period, data will be normalized using year over year sales. It is assumed that there is a normal proportion of beverage sales vs total sales.

**Methodology**

Our datasets came from purchasing data for Pepsi and Coke over the two-year span selected for comparison. Purchasing data from the vendor Jack and Jill were also brought in. This is due to the face this vendor carried the fresh line of Coke beverages. These datasets were downloaded as Excel spreadsheets.

To begin the data cleaning, we removed the headers and some formatting using Excel. Then, they were saved as CSV files. Next, the datasets were loaded into a Jupyter Notebook using read\_csv. The columns were renamed so the related datasets were all the same. The relevant columns were kept and the rest were dropped. Using the DROPNA method, any columns missing data were removed. Duplicates were also removed using the drop\_duplicates method. These new data frames were then saved as new CSV files using to\_csv.

The data was then loaded into Postgres SQL and the relationships between all tables was diagrammed using an entity relationship diagram (Figure 1). Using a query, Odwalla and Simply Juice purchasing history was extracted from the Jack & Jill data. This was then merged with the Coke dataset as they are also Coke products.

-EDA filtering to determine appropriate comparisons

-aggregate items by category, etc. - for side-by-side comparisons between vendors

**Loading into Relational Database with at least these three tables:**

-Cleaned base data sets by vendor

-Year over year sales data for normalization

-comparison table with key metrics explained

**3: Results**

A description of the results obtained without any explanation or interpretation of them. Visual aids such as graphs, tables and maps may be used to summarise the results. If raw data is to be included in the report, it should be placed in an appendix rather than in the results section.

**4: Discussion**

The results of the study are interpreted and implications of these results in terms of the project objectives are discussed. A discussion of how the results relate to other studies is also common. An evaluation of the methodology used may also be undertaken.

**5: Conclusions**

Summarises specific conclusions drawn from the results in terms of the project objectives and working hypotheses.

**6: Recommendations**

Recommendations as to future management actions and/or studies are made based on the findings of the study.

# Appendices

May contain detailed information such as data tables, laboratory reports or photographs.