

ETL Project Proposal:

Coke vs Pepsi “The Soda Wars”

**Introduction and Scope:**

Our company recently switch from having Coke as its prime beverage vendor to Pepsi. This created a perfect opportunity to compare sales between both companies using real world data. We will compare data from FY ‘19-’20 with FY ’20-’21. In consideration of business effects due to COVID, data will be normalized using YoY sales. We will assume normal proportion of beverage sales vs total sales.

**Data sources:**

-Pull purchasing data for Pepsi and Coke over the two-year span. We will also need to pull in purchasing data from one other vendor (Jack & Jill) which carried the fresh line of coke beverages.

- Pull general sales data to create a metric for normalizing year over year data.

**Data cleaning:**

-Remove headers and other unnecessary clutter from data

-Drop unnecessary columns

-trim unit information into simple numeric value, fix any incorrect units

**Data Joining, Filtering, Aggregating:**

-merge Coke and Jack & Jill data to create full Coke dataset. Only Odwalla and Simply Juice purchasing history is needed from Jack & Jill to complete Coke dataset.

-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