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.

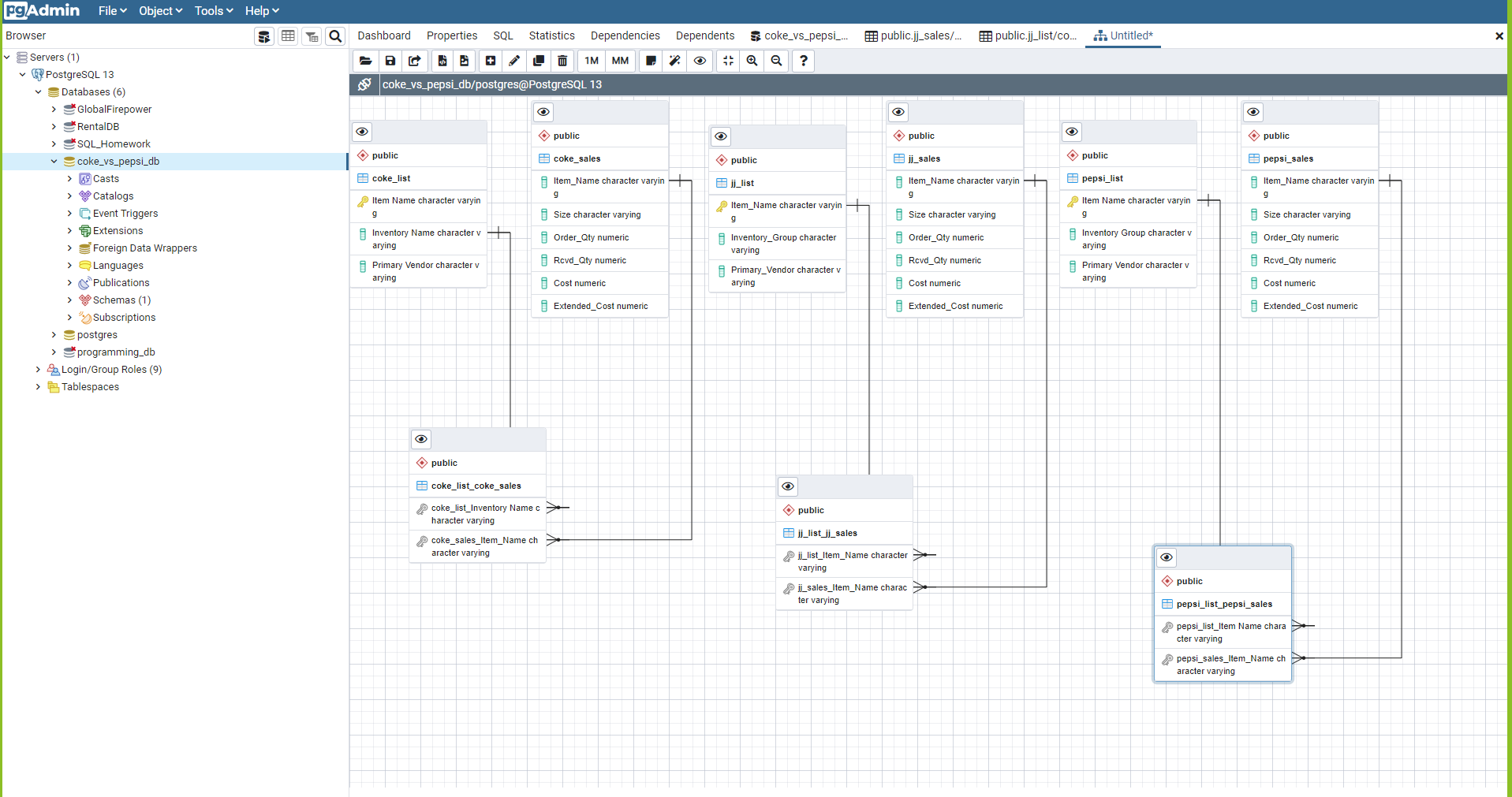


Figure 1: Entity Relationship Diagram

The data was then merged into finalized tables in SQL using queries. Using the item name as the key, the sales lists were merged with the item lists. This data can now be used for analysis looking at these two companies.

**Discussion**

The data selected was incredibly hard to work with. It had a lot of formatting within the initial download that had to be removed in order to even start to work with it. We also discovered in SQL, the data had extra spacing, leading to queries being difficult to run. We had to go back into the data, both in Excel and Python, to reformat multiple times in order to get queries to work in SQL.

**Recommendations**

From here, it would be easy to start comparing these two companies. Each data set encompasses a year. A few different things that would be interesting to look at include doing a year over year comparison of sales. You can also do a comparison of specific similar products, like Diet Coke versus Diet Pepsi