Limitations, Problems and What to Do Next Time

1. (DATA 1) WALMART SALES 2010-2012

The limitation of the Walmart dataset is that it only includes the location details. There are other store types and store name other than Walmart and Walmart supercenter in the dataset that some of them are the subdivision of Walmart and some of them like FedEx are not related to Walmart. So, only rows with store name of Walmart and Walmart supercenter were used for analysis.

1. (DATA 2) WALMART DAILY STOCK DATA
2. (DATA 3) SALES ZIP FILE (5 YEARS)

The sales/price datasets were limited in a number of ways.

First, the data set covers 3,048 unnamed items, from 7 unnamed departments belonging to 3 unnamed categories; sold and priced at 9 unknown stores across 3 states. There is no information about how these items were selected or if the 3,048 items are all of the items that belong to those 7 departments.

Second, the year of the dates remains unknown, even though we figured out the dataset starts on January 28th. The exact year could be 2015, 2011, 2007, 2003, etc.

Third, the format of the data required extensive manipulation to create columns by day for sales data. This made extracting the daily information more difficult. After transposing this data, removing all repeated text indexes, and removing all days with no sales, the dataset was still over 500MB and over 18 million rows.

1. (DATA 4) STORE NAMES, LOCATIONS AND TIMEZONES
2. (DATA 5) 10-YR FINANCIAL DATA FROM WALMART, AMAZON, TARGET AND COSTCO

There are limitations to the data presented. A more thorough month-by-month or week-by-week breakdown of the stats would be more helpful in providing more granular tracking of each corporation's trajectory. This would also afford more opportunities to aggregate specific data for more focused ranges of time. The fact that the Target dataset has all of its values offset by a year compared to the others provides less of a complete scope of our desired comparisons. If there was a more complete set of data for all of Walmart’s competitors throughout the industry we might be able to get a clearer view of it's values as a percentage of a whole and get more specific values relating to things like market share.

1. (DATA 6) MARKET SHARE DATA BY REGION

Pre-processing was required to obtain the Walmart Market Share data. The data is originally in pdf and jpg format and needed OCR technology, to retrieve information from jpg and save it in .csv format.

1. (DATA 7) ONLINE PRODUCT DATABASE (E-COMMERCE)

The e-commerce dataset has 30,000 rows and appears to be a sample from a larger file. Without knowing how the sample was taken, it is hard to make generalizations about e-commerce products using this data. The team decided to only use the first level of the e-commerce category, since the price/sales data are categorized at a similar level.

Most limitations were specific to the datasets and data sources, but some where general issues that the team needed to address. The ETL process created of too many tables and too many large tables. Since the desired solution was to use a cloud-based PostgreSQL database, Elephant SQL, the limit of 20MB of storage resulted in samples of datasets being loaded to demonstrate the process. There were also limits to data storage in Github, so intermediary files were added to the git ignore to avoid using up space.

Additional tables with aggregate functions, calculations and merges of similar datasets were designed to be implemented easily using SQL statements. The team decided it would be simpler to separate the tables and if there was a need to have a merged dataset, it would be easy to write a merge statement on the SQL query. This helped to reduce the number of tables to only the essential dataset tables and simplified the database.