Elyssa Irizarry, Kwadwo Asante, Marco Lopez, Brandilyn Hall
Data Science Bootcamp
Group 3 Project 2

# Baby Diaper ETL Writeup

Diapers are a necessity for any parent with babies due to the number of diapers one baby can go through a day. Knowing the best prices and brands is top-of-mind for parents or future parents. With two mothers of babies in the group, we were inspired to collect data on and analyze diapers and their prices across different shopping websites. The goal is to find the best prices by brand and website. We will perform this analysis by first extracting the data from our chosen websites, transforming the extracted data, and then loading the transformed data into a database. In the following report, we will walk through our ETL process and the conclusions of our analysis.

The first step we took to determine the best value diapers was to scrape the sites of three major online stores; eBay, Amazon, and Walmart, and extract the information we need to perform our analysis. To scrape the sites, we used Splinter and Chrome Driver. This allowed us to access each of the sites and successfully extract the information that we needed. We then used ". find_all", ".find", ".text", and ".strip()", to locate the information we wanted from the HTML code, turn it into readable text and strip characters from the data. This allows us to manipulate and analyze the data easier. The information we extracted was the title, price, and reviews. Some of the data was very nested within the code and took a bit of "drilling" down to get the desired information. We used dot notation to access the appropriate information in our nested data. Once we found the information for a single item, we then looped through the rest of the diapers on the page to get all the data we needed. Once we retrieved the needed information, we created a data frame for each website.

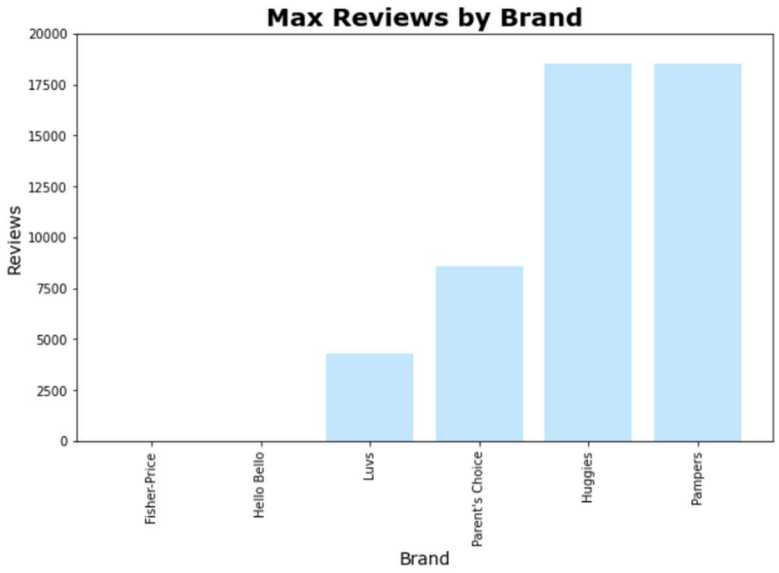
Once the data was extracted and put into data frames, we then needed to transform and clean the data. The price and review amounts came over as strings for all 3 websites, so we cast them to floats to better work with the data in the future (e.g., aggregations and statistical analysis). In our eBay data frame, our prices were displaying as a range of prices. To make it easier to manipulate and analyze, we used "split" to remove the word "to" and the dollar amounts after it, which left us one amount to use for our analysis. In our Walmart data frame, some diapers didn't have reviews. When this happened, the word "Sponsored" or number "18843" would display in the Reviews column. We decided to perform a ".replace" for both and displayed a "0" instead. Additionally, we stripped the dollar sign from in from to f the dollar amounts to also aid in ease of analysis later. Now that we cleaned and normalized our data, we were ready to begin the load process.

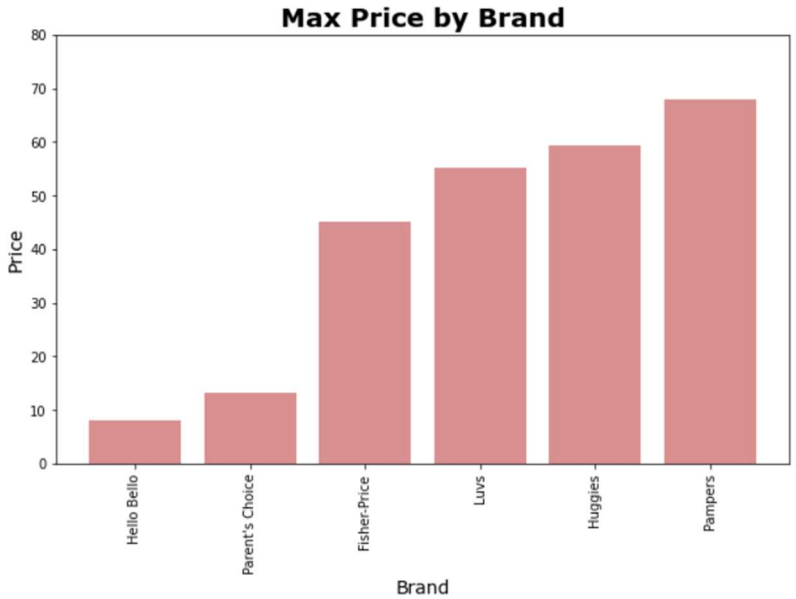
We began to think about our database design and what tables we needed to have. We discovered early on that we needed to make a website table and create a website column as well. We first created the table, then read it into Pandas and looped through our data frame to make the new column. We did the same thing for manufacturer. We also created an ERD diagram to help us easily create our database and tables. We identified and created three tables: Website, Diapers, and Manufacturer. We used quickdatabasediagrams.com to create our ERD, identify our foreign

and primary keys in each table, and add “last updated” for each table. From there, we obtained and saved the PNG, SQL file to create the tables, and the documentation. Once we were complete, we modified our create tables SQL file and pasted it into Postgres.

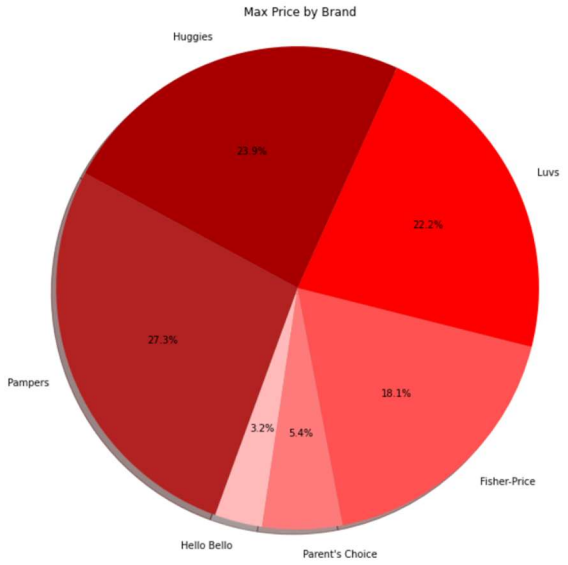
We then wrote our data frames to SQL so we can populate our tables. Once we created an SQL file for our extracted and transformed data, we pasted into Postgres and were able to view the information. We also were able to conduct some queries on the data including, max and min price, average price, group by brand, counts of brand, etc.

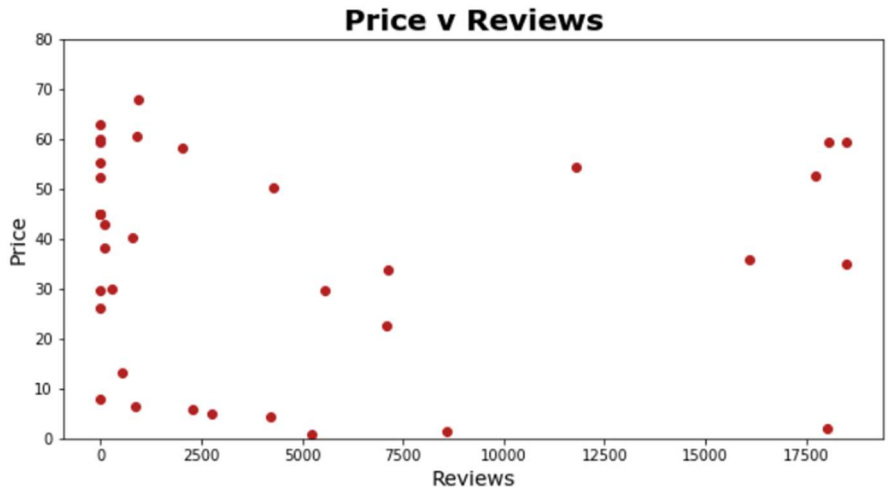
We were able to determine with a bar graph that Pampers brand was the most expensive; however, we determined LUVS had the highest average price. With another bar graph, Huggies and Pampers stood out as the two brands with the most reviews given. Notwithstanding, we also found no substantial correlation between prices and reviews to give any type of trend.



Elyssa Irizarry, Kwadwo Asante, Marco Lopez, Brandilyn Hall
Data Science Bootcamp
Group 3 Project 2




In the future, we would like to perform more analysis into the prices, by size, type, quantity, etc. We would also like to review the ratings and include that in our analysis. Eventually, we would like to do a statistical regression analysis to determine if there is a correlation between price and satisfaction and quality. The limitations we had, beyond time, was not including other merchandisers in our analysis. We also did not obtain data from other sources like, API's or datasets. It was also a challenge with the prices, as many of the websites we looked at had ranges displayed, not singular prices.

Elyssa Irizarry, Kwadwo Asante, Marco Lopez, Brandilyn Hall
Data Science Bootcamp
Group 3 Project 2

***Bonus Question 2***

To run a nightly job, one will need to set up task scheduler, create executable files, scheduler should run the website scraping python scripts, web scraped data will then be allocated to then the relational database management software.