·      **Extraction:**

1.    We decided to add more data that includes information on Walmart stores around US cities with specific information on location, latitude and longitude. So, we get a json file from <https://www.reddit.com> and the file include 5821 rows of data for various store type including Walmart Supercenter, Walmart, Sam’s Club, Neighborhood Market, FedEx Office and so on. The data also included city, state, address, postal code, longitude, latitude.

2.    The us\_state.csv is originally a Python package for working easily with US and state metadata (https://pypi.org/project/us) and it contains US states and their territories with the FIPS codes which is uniquely identifies the geographic areas. So, another csv file was extracted from that for the 50 states of the US and their FIPS codes to use it as an index for US states.

·      **Transformation**

1.     A function was to see the data types and total NAN values and their percentage for each column. The function showed that the Walmart.json has around 15% NAN values for openDate and less than a percent null values for timezone and since these two columns will not be included in our analysis, they left intact to use other information of those rows.

2.    Then the Walmart data has been filtered by storeType to include just the Walmart Supercenter and Walmart.

3.    The column state was renamed to STATE and then it was merged with us\_state on STATE.

4.    Relevant columns had selected from the merged walmart\_state\_df and they renamed for further analysis.

·      **Load**

After creating the engine, the walmart\_state\_df was loaded to ElephantSQL using a connection string which is in a config.py file. A gitignore file was added to the root folder to ignore the config file in checkpoints. Then the Walmart\_state\_df was altered and a primary key added for the “store id” column which was unique.

·      **Limitations/Problems**

The limitation of the Walmart dataset is that it only includes the location details

There are other storeType 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.