ETL Project 2 Final Report

Kai M. Hall

16 Oct 2020

Description:

This project is an example the beginning steps for a data analyst. ETL or extraction, transform, and load requires the analyst to extract data from the internet with web scraping or download csv files. The transform process included uploading the csv into jupyter notebook and clean the data. The cleaning process involves changing or choosing specific columns, setting an index, and converting and merging the data frames into one table.

At this point, the analyst searches through the data frame by checking for null values, data types that inconsistent, and converting data types. Upon cleaning completed, a database is created to hold the new transformed data frame. PgAdmin is used to create the database, the table, and the specific columns that match the columns in the data frame. Lastly, the data frame is uploaded the database ready for analyst from other team members.

Data Cleanup & Analysis

No one likes to clean up messy data. However, this process has a flow or process. My process began searching for quality data. Most website would charge for their data or the data was already cleaned to perfection. I used Kaggle and Data.world for my data sources. I choose data on alcohol consumption by country and fat intake by country. Fortunately, the fat intake data have extra information about Covid-19.

Once the data sets were collected, I began by cleaning each table’s column name and selected needed columns from the entire dataset. Next, was merging the data frames together and setting the index as id for the primary key in the database. Onward, checking the data values for nulls and data types. I did have an object in my data with a comparison character; this would prevent me from making any mathematical calculations. Therefore, the “to\_numeric” method was used to convert the column into a float type.

Lastly, I loaded the data frame to the local database. First, creating the database in pgAdmin to house the data and creating the table and columns that would match my data frame. The final step was double checking the successful transfer with panda read sql query method.

In conclusion, the personal perspective of the ETL process is alike any other checklist for maintenance. With a common focus on specific parts of data set to ensure the data is ready or analyses. Though difficult, rewarding in the end.