Macda Gossa

Hailegirogis Kebede

David Makanka

Extract Transform Load (ETL) Project

GWU Data analytics Bootcamp

November 20, 2018

Data analysis and management requires data integration. These data may come in different formats and sources such as csv files, json, files, or html tables. Integration of various datasets is a key for data analysis. In other words, identifying the dataset, extracting and reading the data, cleaning and structuring it in the desired format, and loading the clean data into a database for storage is imperative for analysis and management. Integrating this which is an important part of working with data. For this ETL project we will use United Census Bureau website, which contains U.S. annual retail trade data in terms of total sales, sales taxes, Inventories and purchases.

**Data Format**

United Census Bureau website annual retail trade data are listed in a CSV files. The site uses North American Industry Classification System or NAICS codes. We will use these codes. There is a total of twelve CSV files, but we will be extracting four datasets with NAICS codes. These datasets are total sales, inventory, purchase, and gross margin csv files.

**Data Extraction and Cleanup/Transformation**

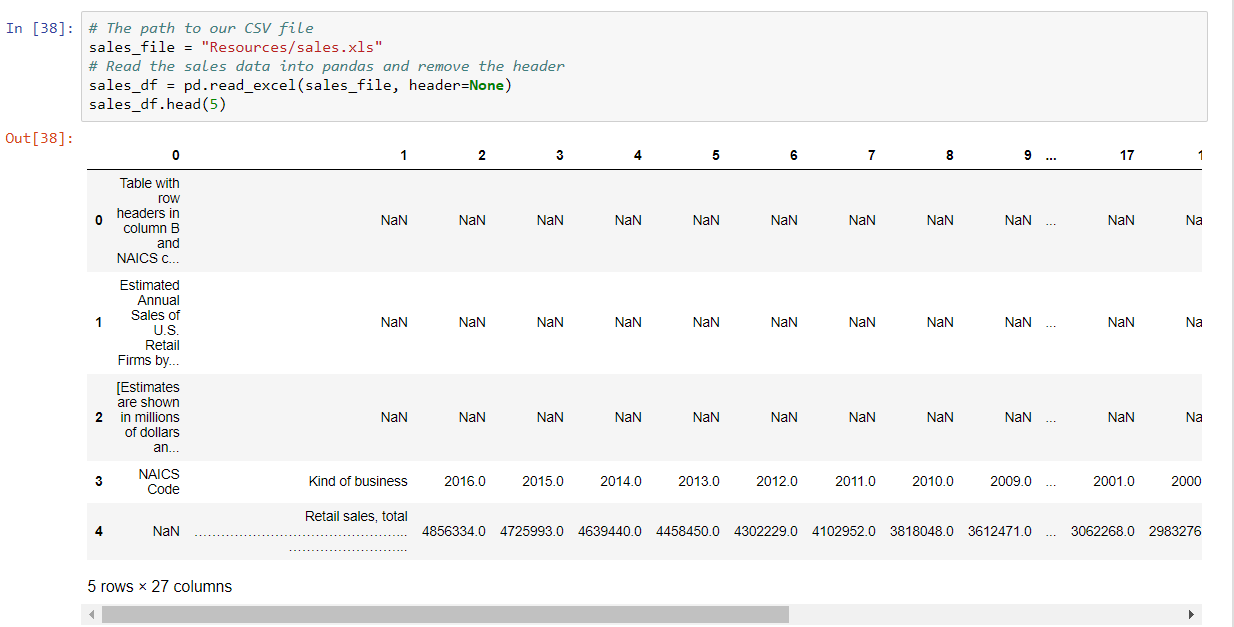
We will use pandas to extract and clean these csv files. Cleaning the data required to strip away any unnecessary variables. The data sets will require us to use the y-axis labels as our columns and we will transpose the dataset using Jupyter Notebook. Finally, we pushed the dataset into MySQL database and joined the four tables using year.

We did the data cleaning process for the sales data and then using the glob module, we wrote a function to automate the extraction and cleaning the other remaining file.

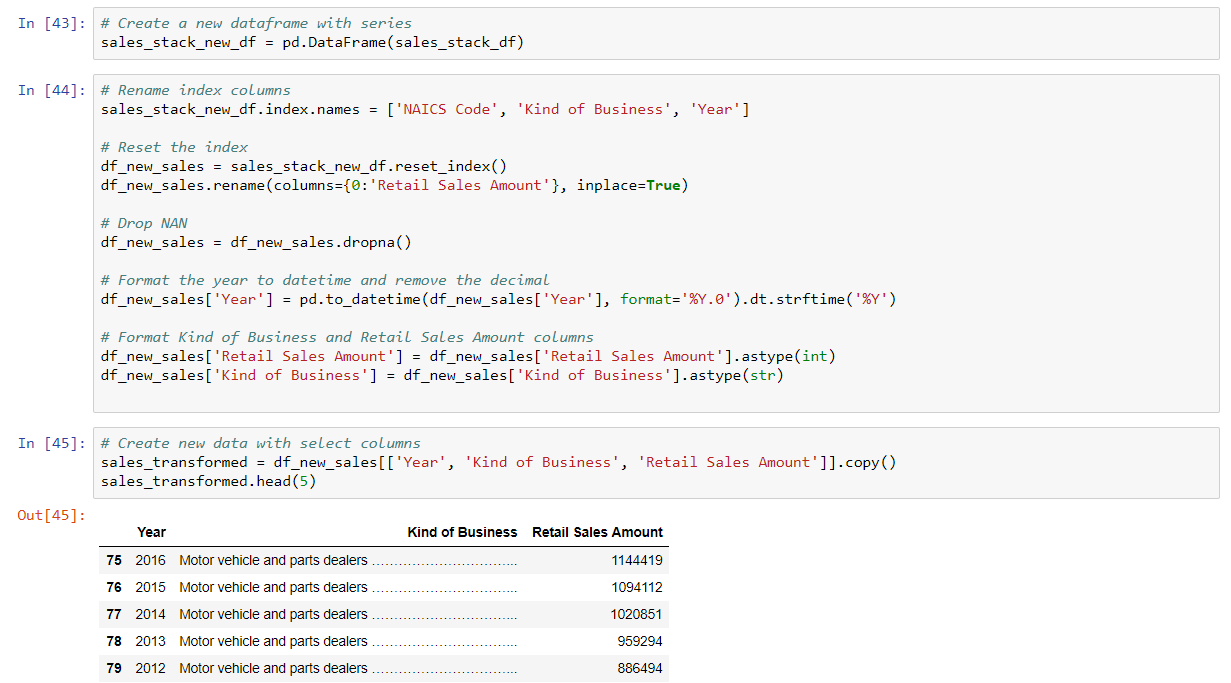
We encountered a challenge of having a single line of data during the transposition because the CSV files that we extracted were all aggregated data for the given period of years.

**Data Storage into a Database**

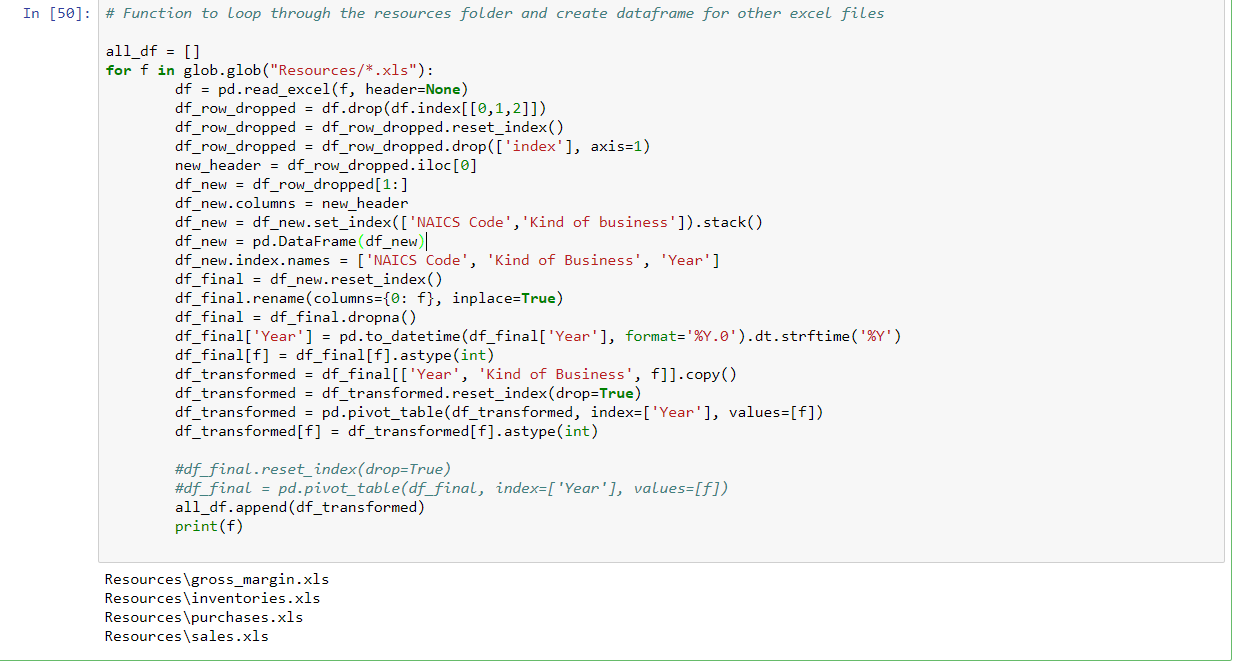
We used relational database, MySql database, for storage and created a connection using Jupyter Notebook. We then created a database, pushed and joined the tables using Year as key for joining the CSV files.

**Extraction**

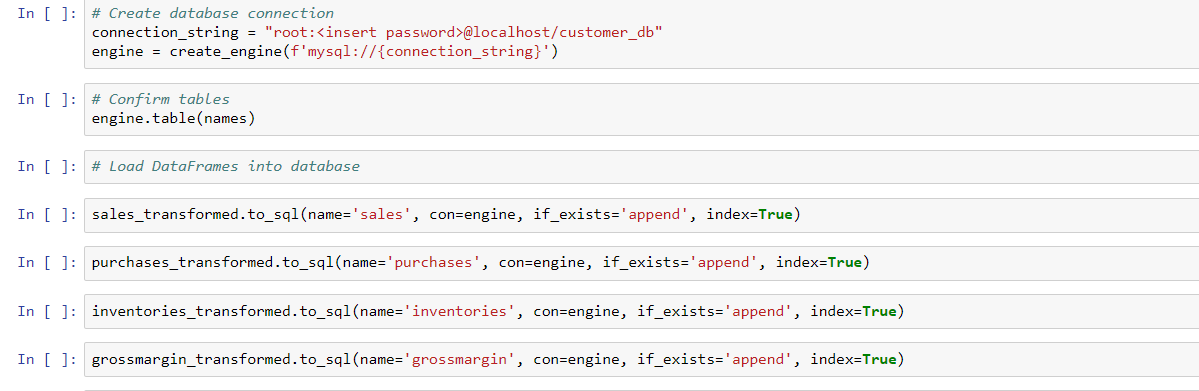
**Transformation- Screenshot 1: Data cleaning**



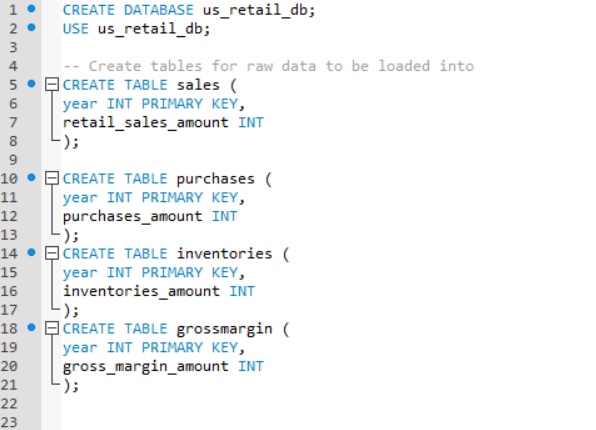
**Transformation- Screenshot 2: Cleaning all CSV files using the globe module**



**Loading to MySql - Screenshot 1: Database Connection and Data Frame Loading**



**Loading to MySql – Screenshot 2: Database and Table Creation**



**Loading to MySql- Screenshot 3: Query**

