Project 2: ETL

# Group 2:

Ruby Sanchez

Vilma Diaz

Andrew Paton

Karina Horna

# Introduction:

We’ve all heard that once you drive a vehicle off the dealership lot it depreciates drastically and immediately. We thought it was interesting that in 2021 the value of used vehicles are retaining their value and selling for higher prices than what was typical in prior years. This piqued our interest and thus we worked to gather historical vehicle data and the most current pricing data to build our database for future analysis.

The purpose of this project was to extract data from multiple data sources, transform the data into a usable and meaningful format, then load the data frames into tables within our database.

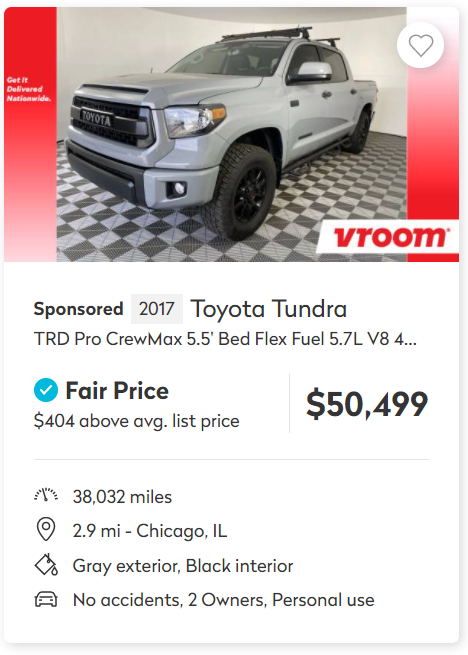
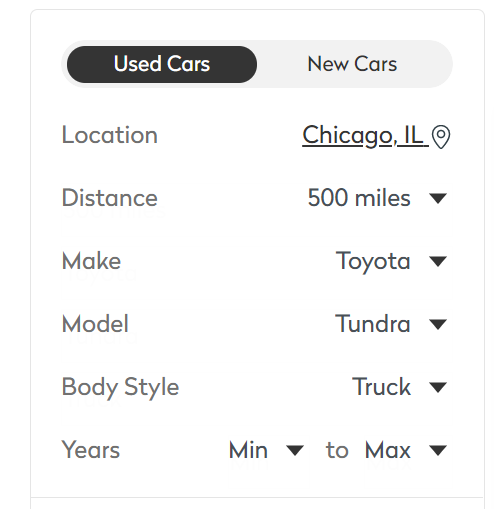
The team gathered historical data via Kaggle which was available in a csv file. Also, we built a web scrape program that gathered current data from truecar.com. With data from these different sources we determined that the transformation of the data would be most efficient using Pandas. The final step was to load the data to PostgreSQL where future analysis can be conducted.

Within Github we have stored the following:

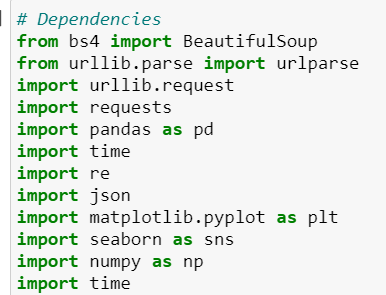
* Historical Data in a csv format (from Kaggle)
* CSV file generated from the web scrape program
* .ipynb file which includes the scrape program code and the ETL code
* .sql file which includes the queries used to establish the database tables

# Extract:

Scrape Program

We used BeautifulSoup to scrape the data for trucks currently for sale within 500 miles of Chicago, Illinois. We chose this location to match the historical data set of truck sales in Illinois from 2018-2020.

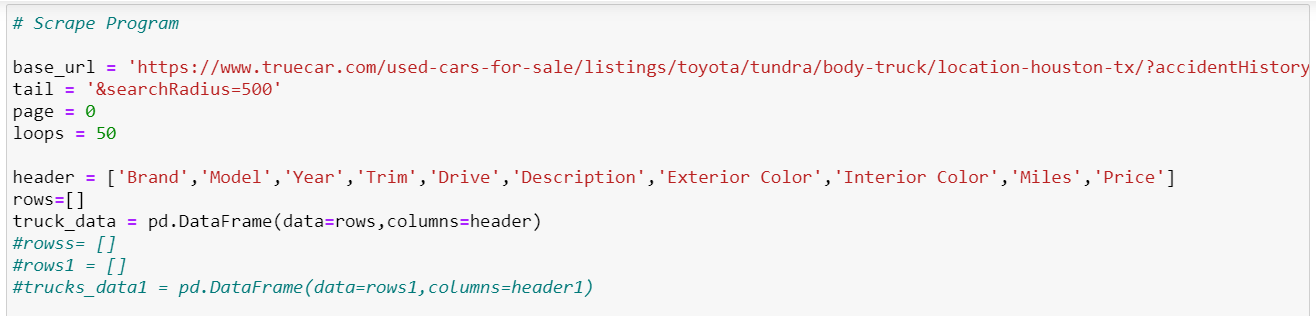
To build the scrape we built a nested “for” loop to iterate saved data for each truck then changing to the next page of listings. We created an empty list to contain the data then assigning it as rows to a Pandas data frame.

Scrape Code:

Dependencies such as BeautifulSoup

Web scrape Variables

We noticed each page URL had a page number, therefore allowing us to use a loop number to change the page and download all the trucks



Web Scrape Main For Loop

As mentioned, we changed the URL by the loop number and saved all the trucks listed on that page in “div\_list”



Web scrape Nested For Loop

The nested loop saves each truck’s information by calling the loop number record from the “div\_list”. A try and except function was used in case information wasn’t present.



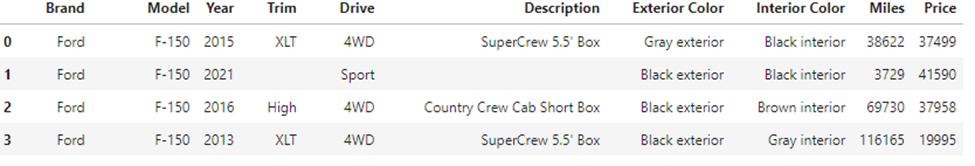
Truck Historical Sales Data from Kaggle

We found historical truck data of all sales in Illinois from 2018-2020 to compare with current listings. The data was stored as a CSV within the same repository as our web scrape program. We read the file using pd.read then proceeded to the Transform stage.

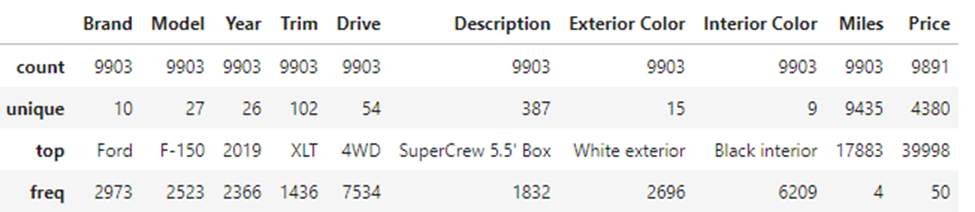
# Transform:

After acquiring the data from both our web scraper program and Kaggle, we further transformed it into similar tables layouts.

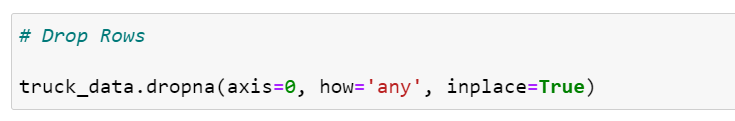
Web Scrape table example:

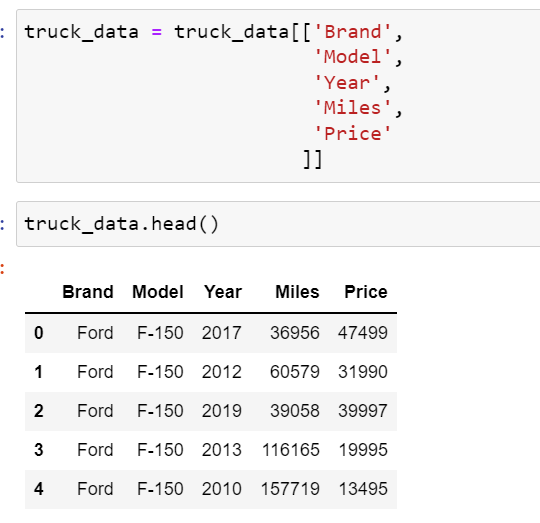


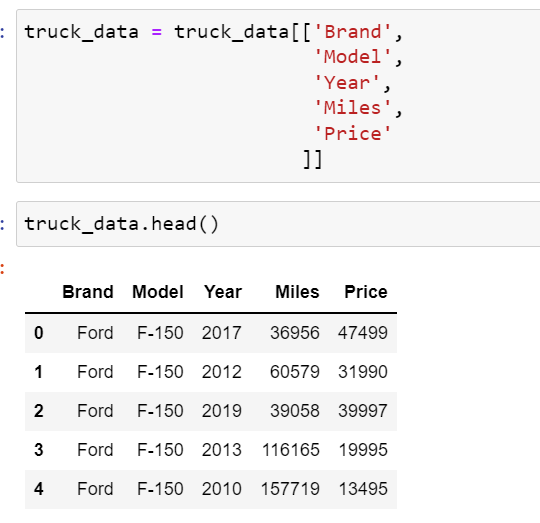
When reviewing the data, we found that some rows had missing values. This was evident by the various counts in values for the column headers.



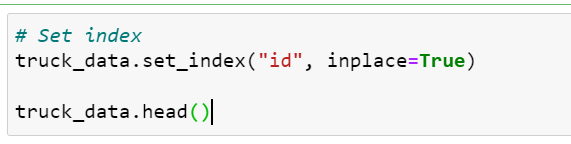
To remove incomplete data, we used “dropna”



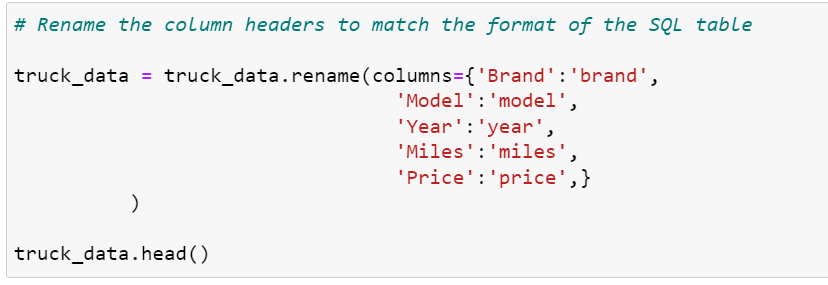
To trim down the data set to just relevant columns we created a data frame that included just the columns that we were interested in.



In order to prepare our data for the load phase, we set the index as our unique id.



Additionally, we renamed columns in order to fit the format of our PostgreSQL database tables. This allowed us to complete the loading of the data into our database.



# Load:

The final database was created in PostgreSQL 14. Before the loading of the data could happen, we had to create a database and establish the tables in which the data would be stored.

Database name: vehicle\_data

Table names: ['historical\_data', 'current\_data']

The next step was to establish the connection between our code used to extract and transform the data and the PostgreSQL database where we wanted the data stored. Once we made the connection and established that it was properly working, we were able to successfully load our data.

# Potential Analysis:

Now that the database is created, we have the ability to query both the historical data and the current data for future analysis. We could use this data to understand the price escalation seen in today's auto industry. There are many current issues that we suspect are leading to an increase in used vehicle prices, such as supply chain shortages or inflation. Once we can complete a thorough analysis of the data in our database then we could make an informed decision about future analysis and areas to research.