Analyzing Retail Sales through Product and Industry Classifications

By: James Miller, Cole Ballard, Nathaniel Van Schyndel, and Gavan VanOver

We are Dev10 associates with Genesis 10 and for our capstone project we focused on retail services in the United States. Specifically, we looked at nationwide retail data, including product and industry classifications, from 1960 to 2022. Furthermore, we completed a time series analysis of retail sales with the following independent variables: consumer price index (CPI), real person income (RPI), nationwide number of retail employees (USTRADE), and nationwide number of wholesale employees (USWTRADE). The product of this time series analysis is a machine learning model that aims to predict future retail sales in the United States, based on the independent variables. This model will serve to aid our client in streamlining their supply chain logistics beyond simple tactical order fulfilment methods and practices, which will ultimately result in a lower cost of goods sold, moreover maximizing profitability. Taken all together, this project attempts to answer the following questions:

1. How have total sales in retail changed over time, as well as within in the subcategories (by NAICS codes)?
    a. Can we identify trends within the retail data?
2. What industries have the highest number of sales?
3. Is there correlation between CPI, RPI, USTRADE, or USWTRADE and retail sales?
4. Can we use historical retail sales data to predict future trends in retail sales?

We would specifically like to focus on the 2017 Census and answer the following:

5. How do product sales compare when sold in different types of establishments? (This would be looking at something like candy sales or alcohol sales and seeing how their sales compare when sold at like a gas station vs a grocery store or some other location.
    a. Which products vary the most by industry?
    b. Which industries tend to perform better?
    c. Which products sold the most in 2017

For the first four questions we will be using these datasets:

- Monthly Retail Trade Report, https://www.census.gov/retail/index.html
    - Monthly Census data from 1992 to 2020 listing the total retail sales broken down by NAICS codes
    - Accessed on: August 2, 2022
- Consumer Price Index (CPI) United States, https://www.kaggle.com/datasets/sfktrkl/consumer-price-index-cpi-united-states
    - Provides the monthly Consumer Price Index for the US since the beginning of 1960
    - Accessed on: June 29, 2022

- Macroeconomics US, https://www.kaggle.com/datasets/denychaen/us-macro?select=US_MACRO110522.csv
    - The original dataset provides over 258 different macroeconomic measures for each month since the beginning of 1960
    - Accessed on: June 29, 2022

For the product and industry-based question:

- All Sectors: Products by Industry for the U.S., https://data.census.gov/cedsci/table?q=ECNNAPCSPRD2017.EC1700NAPCSPRDIND&n=N0600.44&tid=ECNNAPCSPRD2017.EC1700NAPCSPRDIND&hidePreview=true
    - 2017 Census data broken down by both product codes and industry codes. Provides different measures such as the number of establishments and the amount of sales revenue
    - Accessed: June 29, 2022