# Technical Report: Retail Sales

Nathan Van Schyndel, Gavan VanOver,
Cole Ballard, James Miller
8/19/2022

Our project assignment was based on retail sales with an emphasis to utilize the data available through the US Census Bureau which pertain to codes 44-45 (Retail Sales). With that focus we began exploratory analysis to find the main points of prominence within the census data. What are the different classifications within the census bureau data? The census data is broken down into numerous categories that have basis in demographics: sex, ethnicity, race, veteran status etc. There are also subsections created on the basis of establishment size: number of firms, number of establishments etc. There are statistics on the job creation and loss: number of jobs created from opening establishments the last 12 months, number of jobs created from expanding and opening establishments during the last 12 months, number of jobs lost from closing establishments during the last 12 months etc. With the seemingly infinite categorizations and manipulations of the vast amount of data we chose to focus on products by industry that contain the NAPCS and NAICS codes. The NAPCS codes start with 44 and 45 in respect to the retail industry as a marker and are further broken down into different meanings with expansion on this base. The meaning of these NAPCS codes are based on the goods and production associated with them. Examples are lapidary work except for watch jewels, construction services for new swimming pools, wholesale sales of new and used automobiles, etc. This coding is further stratified into the NAICS codes which are built upon the basis of the retail outlet which delivers these goods to consumers. Some of the meanings include used car dealers, florists, shoe stores, etc. The information breakdown is the number of establishments and sales, value of shipments, or revenue of NAPCS collection code in $1000.

(Bureau, *All Sectors: Products by Industry for the U.S.: 2017* 2017)

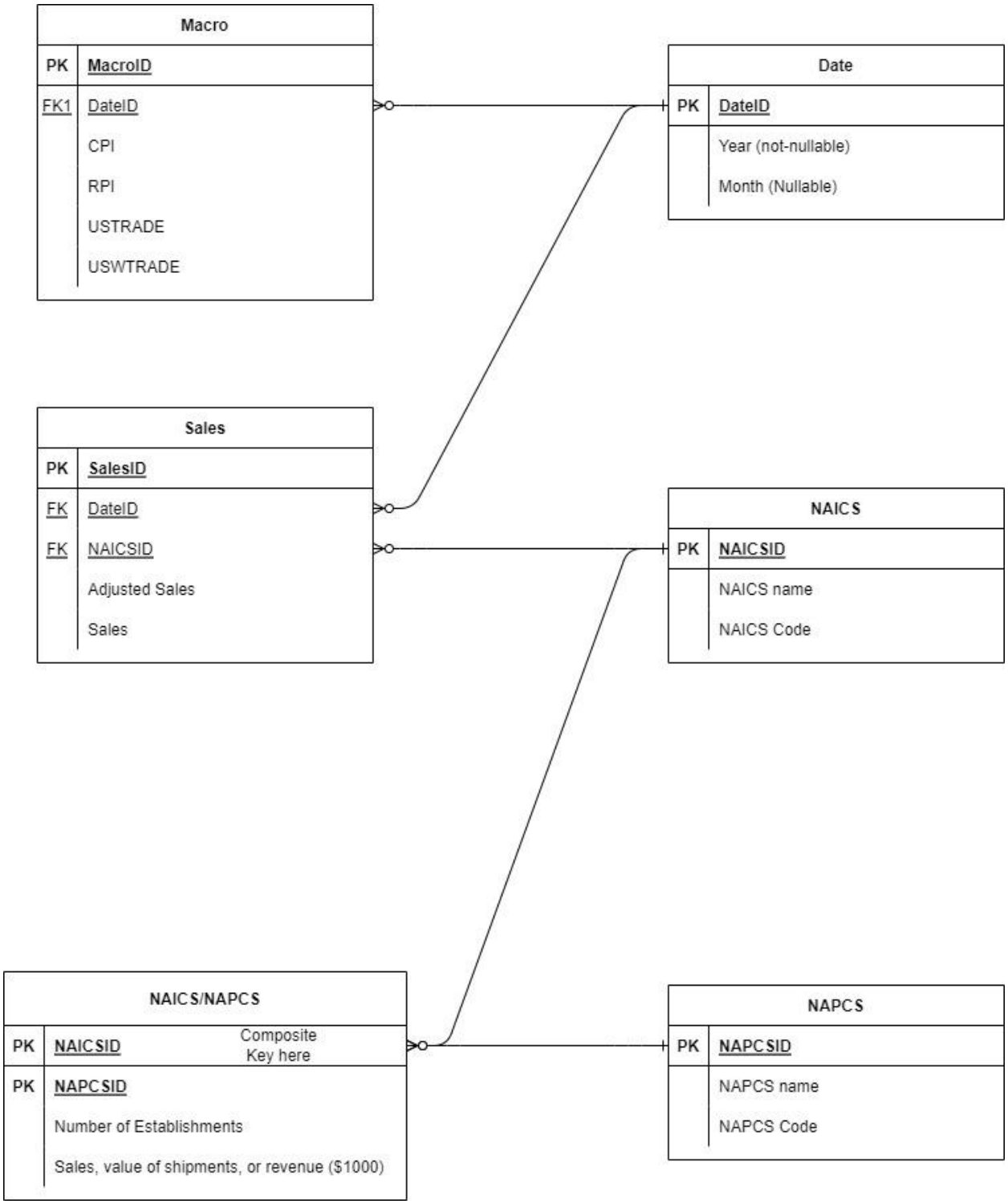
| Meaning of NAPCS collection code | 2017 NAICS code | Meaning of NAICS code |
|---|---|---|
| Wholesale sales of other goods, not elsewhere classified | 451140 | Musical instrument and sup... |
| Wholesale sales of other goods, not elsewhere classified | 451211 | Book stores |
| Wholesale sales of other goods, not elsewhere classified | 452319 | All other general merchandi... |
| Wholesale sales of other goods, not elsewhere classified | 453110 | Florists |
| Wholesale sales of other goods, not elsewhere classified | 453220 | Gift, novelty, and souvenir s... |
| Wholesale sales of other goods, not elsewhere classified | 453310 | Used merchandise stores |
| Wholesale sales of other goods, not elsewhere classified | 453910 | Pet and pet supplies stores |
| Wholesale sales of other goods, not elsewhere classified | 453920 | Art dealers |
| Wholesale sales of other goods, not elsewhere classified | 453991 | Tobacco stores |
| Wholesale sales of other goods, not elsewhere classified | 453998 | All other miscellaneous stor... |
| Wholesale sales of other goods, not elsewhere classified | 454110 | Electronic shopping and ma... |
| Wholesale sales of other goods, not elsewhere classified | 454210 | Vending machine operators |
| Wholesale sales of other goods, not elsewhere classified | 454310 | Fuel dealers |
| Wholesale sales of other goods, not elsewhere classified | 454390 | Other direct selling establis... |

With that basis established, we continued with more EDA for supplemental and relatable data. Since our initial data set was established in 2017 and covered the breadth of the US and all sales falling into the retail category we chose to continue looking for information in a very broad spectrum. With that, we were able to obtain datasets that overlapped on the timeframe, and on information obtained with the basis of NAICS code. The first was a dataset with hundreds of columns based on macroeconomic principles (Dchaen, *Macroeconomics us* 2022). The second was another dataset focused on CPI which could be joined with the macro table (Avigan, *Consumer price index (CPI)* 2020). The third data set came again from the US census bureau. This data came in the form of an excel workbook broken down into individual sheets based on year then NAICS codes under sales by month (Bureau, *Monthly Retail Trade Report* 2022). EDA was completed and numerous questions were established for exploration into the datasets and their relationships. The finalized chosen questions were:

1. How have sales changed over time?
    a. What kind of trends can be observed over time?
2. What industries are the biggest by sales?
3. Are there relationships between macroeconomic principles and retail sales?
4. Can we use previous data to predict future trends in retail sales?

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

Each set of data needed extensive ETL processing to be usable and relatable for translation into our database through the cloud services pipeline. A brief summary will be discussed here. A full summary can be read in detail in the ETL report. The excel workbook (Bureau, *Monthly Retail Trade Report* 2022) underwent extensive transformation and cleaning to reach multiple structures for different applications. All extraneous information was determined and removed. After initial processing the multiple structures were established. For use in uploading into the database the month and year columns need to be transposed and broken into separate columns for loading into the proper fields and relating to other entities in the database. For machine learning the entire dataframe needed to be transposed so dates were in the datetime data type and established as the index columnized by NAICS code or name. The macroeconomic (Dchaen, *Macroeconomics us* 2022) and CPI (Avigan, *Consumer price index (CPI)* 2020) tables were cleaned and combined to contain only four fields that most closely pertained to our retail sales data. The data in the primary census table (Bureau, *All Sectors: Products by Industry for the U.S.: 2017* 2017) was stripped of all calculable values that involved totals. Numerous numerical columns contained characters that needed to be removed for conversion. Naming conventions were changed and made more approachable. The data was then loaded into the database through cloud services in the format according to the ERD.



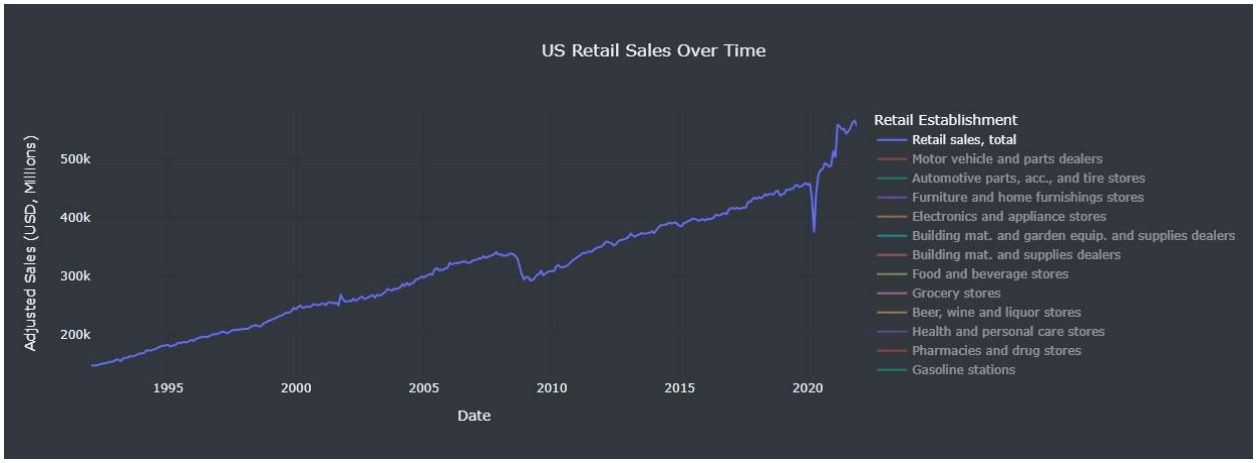

The DDL allowing the completion of the ETL was done through Azure data services and populated using Azure cloud services. A document with all code for the DDL is located in the code folder.

After producing working code for the ETL, we created a data pipeline that could automate the process. This required the creation of several objects, the first being an Azure Data Factory to house the pipeline. Our Data Factory was titled, ‘data-factory-gavanvanover’ and the pipeline created within was called “Capstone_Automation”. The first step of the pipeline was to create two data sets. The first data set was the original 2017 data for the NAICS/NAPCS sales and the second was a copy of that stored in a separate container. The pipeline would copy the original data set from our data container and create that copy in a container called “capstone-group6-producer-data”. Validation was added to make sure that the copied data set was actually where it was supposed to be. Then a topic and producer were created for reading and writing the files. The topic and producer allowed for some data cleaning and transformation to be completed prior to the data being compiled and consumed. The consumer would then gather those messages, perform some more transformations on the data, then write the data to individual .csv files in the container titled “capstone-group6-consumed_data”. After the consumer was finished, the pipeline would preemptively turn off all constraints within the SQL database so that the next databrick, “capstone-populating-database”, could run and load all of the data, not just the 2017 census data, to their respective tables. The last step of the pipeline was turning the constraints back on so that the primary and foreign keys would act as intended when querying the database. In summary, the pipeline copies the 2017 census data and moves it to a new container, ensures that the copied data is in the right place, creates a topic and producer, creates a consumer, and correctly populates the database, all while cleaning and transforming the data along the way.

With the database populated, construction of the dashboard to represent our findings with visualizations could be implemented. Original versions of the visualizations were generated in the form of napkin drawings and submitted for peer review to assure proper representation of the data and exploratory questions. After review by colleagues the visualizations were modified accordingly and put into production. While visualization production

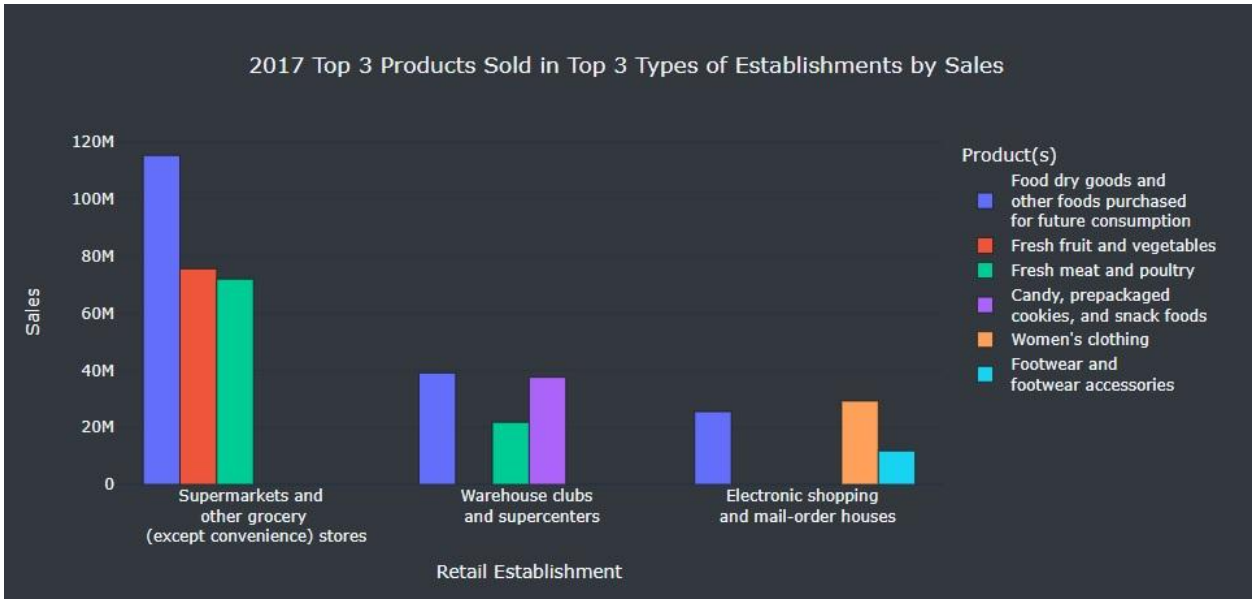
was underway we progressed in instantiating our dashboard using Dash. Dash is a python framework created by plotly for creative interactive web applications. This dashboard was translated to a live webpage using Heroku. Heroku is a PaaS that enables developers to build, run, and operate applications entirely in the cloud. The dashboard in its entirety contains all our visualizations and is hosted by the Heroku app. A napkin drawing of our dash layout was contrived and again put through multiple peer reviews for coherence while the basic structure and code was created. With review of visualizations completed, implemented critiques, full scale construction of the web hosted dashboard began. Documentation for the drawings and feedback are located in the Plans, Drawings, and Outlines folder of our github repository. With the dashboard and all other processes completed the last section of time was ensuring an intuitive schema to the entire project and tidying up any oversights that the project incurred.

While we effectuated the deliverables of the project, we focused on obtaining answers to our initial proposed questions. The most rudimentary question, our first, was how are retail sales changing over time.

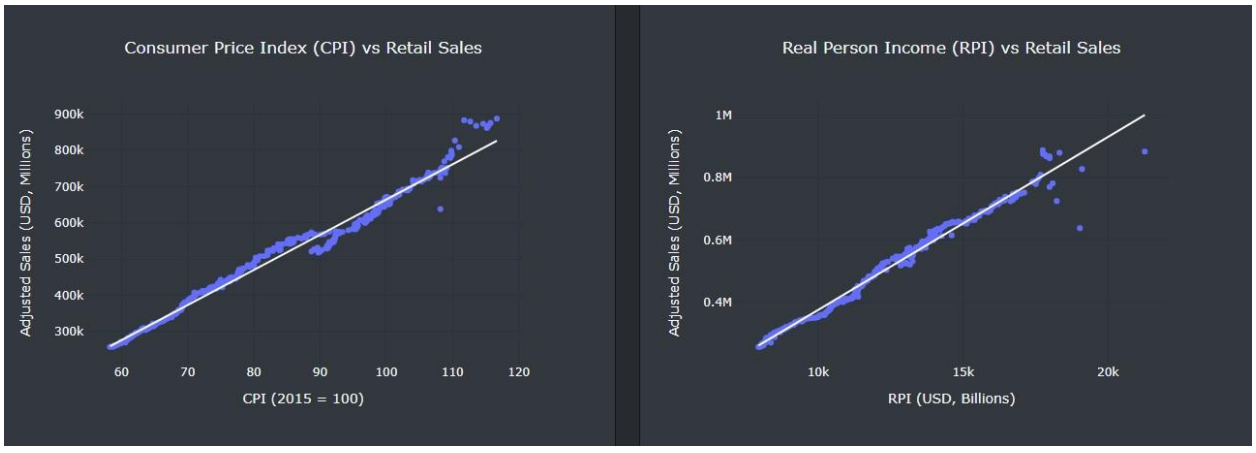



Nearly all categories that were broken town by sales saw a positive trend. In this positive sales trend a modest seasonal anomaly is observed. On a regular basis sales were seen to fluctuate with very large and notable indications happening in 2008 and 2020. These fluctuations represent the housing collapse and the COVID-19 pandemic, respectively. The second

question was which specific industries in the overarching retail industry are the largest by sales?



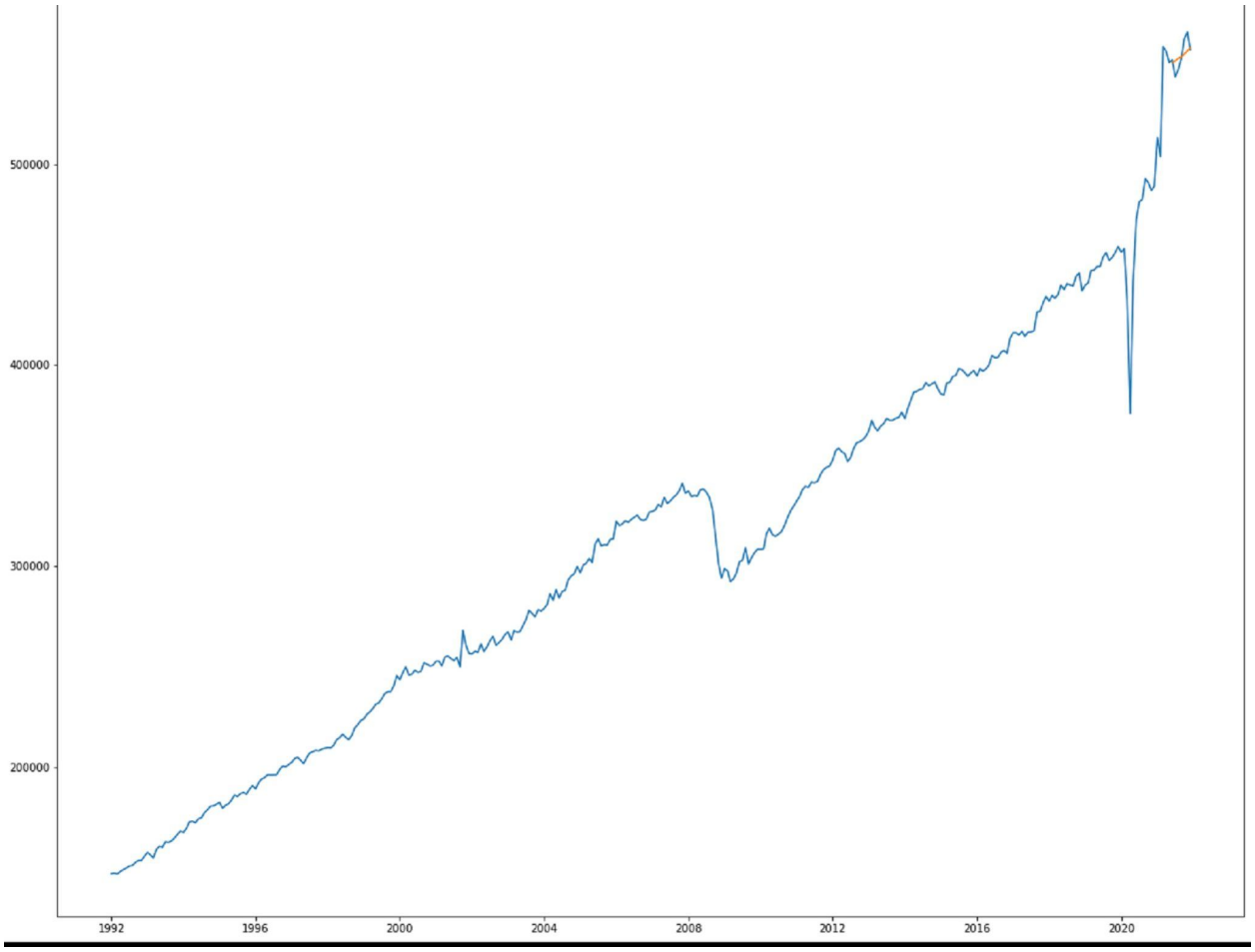

In retail, the top three largest by sales are supermarkets and other grocery (except convenience) stores, Warehouse clubs and supercenters, and lastly electronic shipping and mail order houses. All three of the categories seem like intuitive depictions for the top spots in retail. The top two being associated with food and goods that could be seen as essential. These goods could be included in the CPI bag. With the advent of COVID-19, electronic shopping has seen an aggressive increase in sales. Question three centered on the relationship between macroeconomic principles and retail sales.



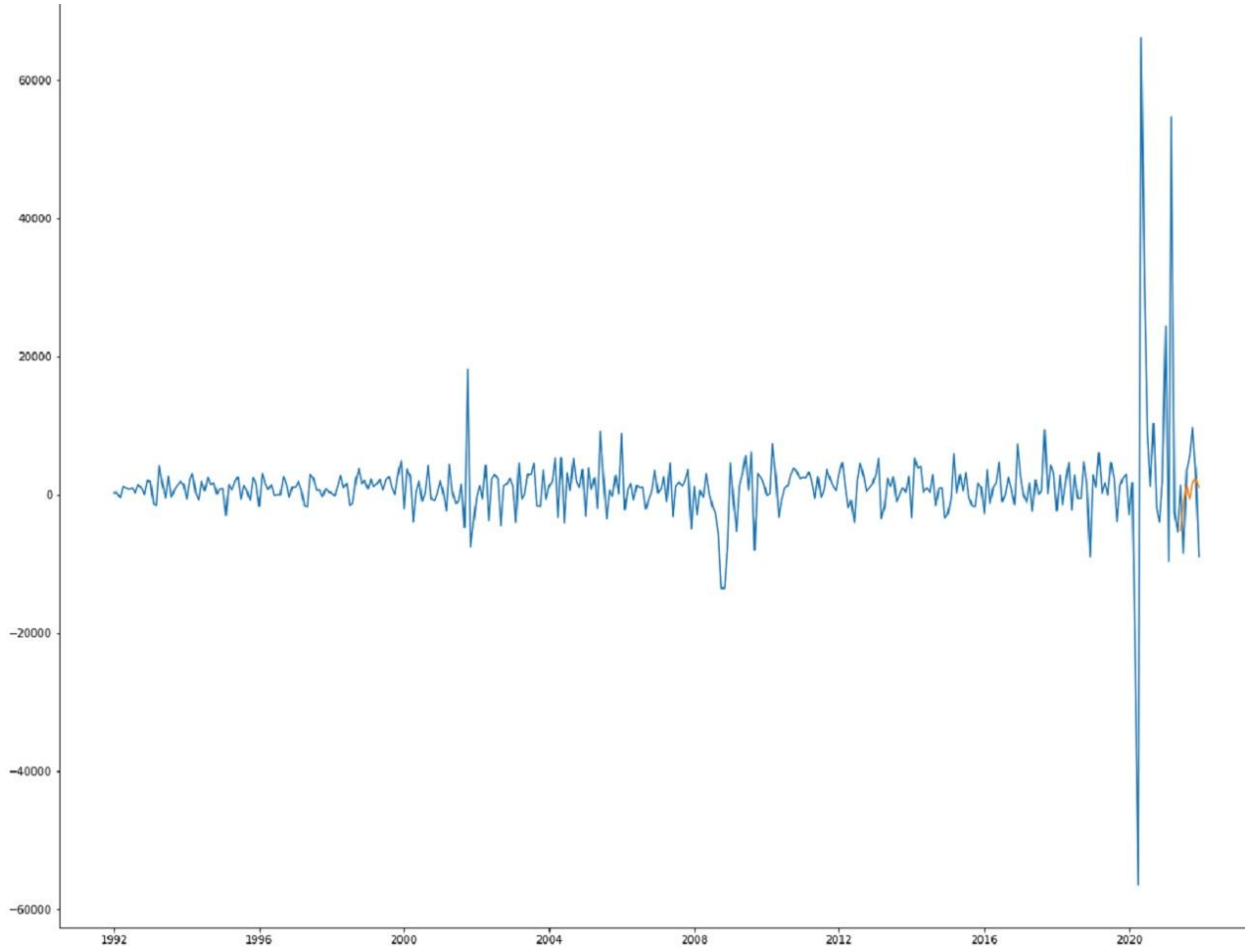
Even though the graphs show what appears to be a linear positive linear relationship, it is important to check with a correlation matrix.

| | Retail sales, total | CPI | RPI |
|---|---|---|---|
| Retail sales, total | 1 | | |
| CPI | 0.980778512 | 1 | |
| RPI | 0.987666331 | 0.982462 | 1 |

The correlation coefficient is very close to positive 1 for both the CPI to retail sales and retail sales to RPI relationships, meaning both are strong, positive linear relationships. This relationship however seems to break down as sales reach higher values. It is important to note that CPI and RPI are both related to the cost of goods. This demonstrates that the cost of goods to consumers has been relatively stable in accordance with inflation. CPI is a newer representation of RPI. CPI includes mortgage costs and RPI takes no housing costs into account. Question four is derived from our machine learning model and will be discussed in more detail later. Question five focused on how product sales compare when sold in different types of establishments. The last question that has been mentioned, but not discussed is question four: Can we predict future sales trends within the retail industry? To determine this, we were advised to look into a machine learning algorithm that related to time-series analysis. Since we had a large amount of associated data, we chose to utilize a multi-variant time series model, which is not to be confused with a multivariate machine learning model. Multi-variable linear regression models have a continuous outcome and multiple predictors. A multivariate method refers to the modeling of data that is often derived from longitudinal studies, and outcome is measured for the same individual at multiple time points. After trying multiple methodologies, python arima - pmdarima - was chosen as the model. The data frame was composed in the proper structure and cleaned, imputing any null values forward or backwards as necessary. A model was generated which demonstrated that the raw data is not stationary, as it fails the Dickey-Fuller test.

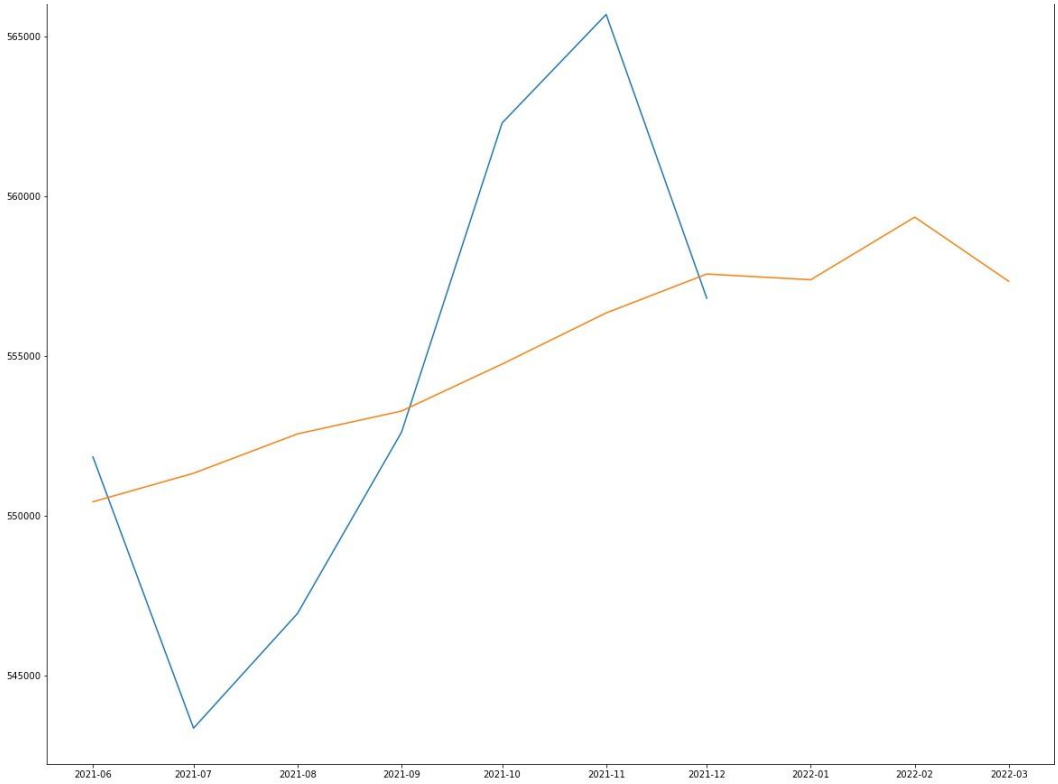



A differenced stationary model was produced alongside the non-stationary model. Both models produced modest results that included three intercepts.

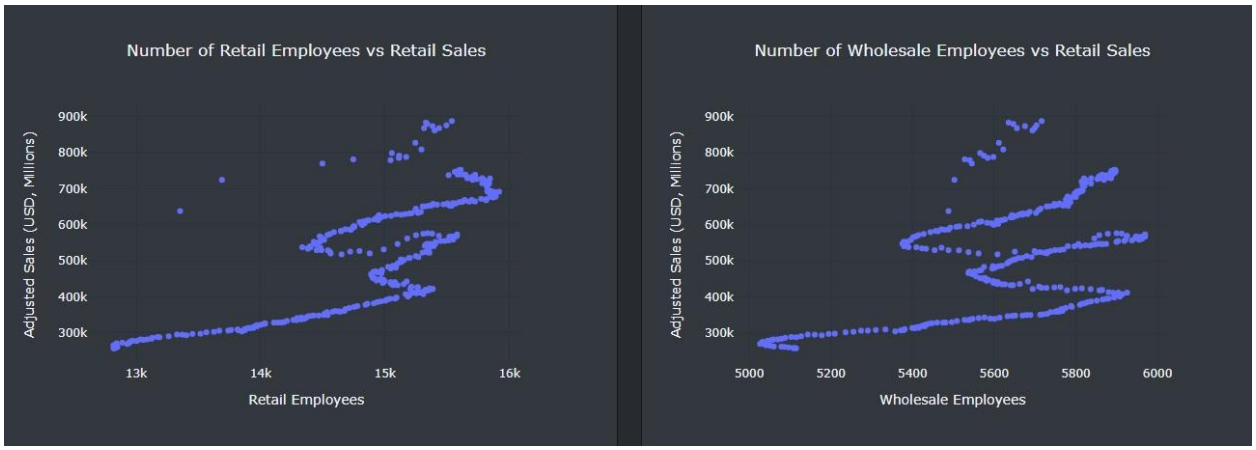



One proposed model used criterion based on AIC. While we acquired approximately 360 rows of data among 30 columns this likely does not contain enough data to be effective using a model with AIC or Akaike information criterion. Based on the summary statistics, MAE, and MAPE displayed by any of the three models generated they were within reasonable bounds to be considered acceptable. The primary factor for determining our decision was mean absolute percentage error, by which in order to considered an effective and acceptable model is generally needed to be within 5%, 6-25% to be low accuracy, above 25% unacceptable, and we found our

bounds inside the margin of low accuracy.




The images above demonstrate our model over the training set with the first quarter sales predictions into 2022. The top image is the non-stationary model and bottom stationary. Secondary information that was discovered is an association and stratification of the number of employees in wholesale and retail industries in association with sales.




As the sales increase by approximately 100-200k in sales a new trend line seems to appear. Since it was not significant in the original discussion it was not researched. However, with more time it would be interesting to find a correlation as to why these trends or pseudo trends appear.

With this research, there were different evaluations that were recognized as varying ways to assess the industry. We decided the primary focus of the project would be on sales and principles in association with sales. The specific industry of focus is retail. This project took an ample amount of time among the team over the assigned weeks. We were able to acknowledge divergent trends across the retail industry and its subcategorization according to the census data and associated data. These trends, while being recognizable, were not able to be proven substantial or corollary in the sense that they would not help one to predict very far ahead into the future. This is more or less a limitation of a time-series analysis. Take for instance, the housing market crash of circa 2008 or the economic impacts of the COVID-19 outbreak in 2020.

These would certainly throw a wrench into any attempt to forecast something such as retail sales far into the future. Our predictive analysis was in the range of low accuracy and coupled with independent factors like the ones mentioned above would not be particularly reliable for long term forecasting. That said, we found that it would be feasible to predict retail sales for a short time into the future, approximately 3 months in our case. These smaller forecasts could continue to be updated on a rolling basis as additional information becomes available. Perhaps with additional time and deeper consideration, analysis could be performed to further dissect the data into more appropriate groupings that could conceivably indicate more substantial value in forecasting, as well as market opportunity.

Sources:

(Bureau, *All Sectors: Products by Industry for the U.S.: 2017* 2017)

Bureau, U. S. C. (2017). *All Sectors: Products by Industry for the U.S.: 2017*. Explore census data. Retrieved August 3, 2022, from https://data.census.gov/cedsci/table?q=ECNNAPCSPRD2017.EC1700NAPCSPRDIND&n=N0600.44&tid=ECNNAPCSPRD2017.EC1700NAPCSPRDIND&hidePreview=true

(Avigan, *Consumer price index (CPI)* 2020)

Avigan, A. (2020, February 16). *Consumer price index (CPI)*. Kaggle. Retrieved August 4, 2022, from https://www.kaggle.com/datasets/aavigan/consumer-price-index-usa-all-items

(Dchaen, *Macroeconomics us* 2022)

Dchaen. (2022, May 11). *Macroeconomics us*. Kaggle. Retrieved August 4, 2022, from https://www.kaggle.com/datasets/denychaen/us-macro?select=US_MACRO110522.csv

(Bureau, *Monthly Retail Trade Report* 2022)

Bureau, U. S. C. (2022, July 15). *Monthly Retail Trade Report*. United States Census Bureau. Retrieved August 10, 2022, from https://www.census.gov/retail/index.html