Will Davis

Tricia Dudley

Olivia Sylvester

Time Series Analysis of US Industries

A time series can be defined as “a sequence of data points that occur in successive order over some period. In particular, a time series allows one to see what factors influence certain variables from period to period.” [1] Visualizing data in this way allows for comparison of multiple data sets, allowing for the combination of the time series resulting from this data in many powerful ways. The goal is to work with a smaller number of data sets, and the methods we have chosen to select these small “clusters” of data will be further discussed. Time series are often utilized in financial environments, such as stock markets, businesses, and even used to measure the business activity of entire industries, which is the subject of analysis for this project.

The data sets being used in this project come from the United States Census Bureau, specifically historical data on US industries. There are 6 different sets of data, each with a different way of measuring the over 80 industries to which information is available. These measurement categories are as follows: Shipments, New Orders, Unfilled Orders, Total Inventories, Inventories to Shipments, and Unfilled Orders to Shipments [2]. The task at hand is to compare these industries based on their varying measures from 1992 to 2021 to observe any previously unknown similarities.

In order to analyze and work with the data, it is necessary to wrangle it. “Data wrangling refers to a variety of processes designed to transform raw data into more readily used formats.” [3] In other words, we need to format the data to be compatible in the software program we are using for this project, R-Studio.

First, we downloaded all six excel sheets. Second, since the industry names and methods of measurement are classified based on a unique ID code, we created an array that stores each ID code as an element and initialized a data frame to store this wrangled data. Then, we used the industry ID array to filter data specific to each industry. Next, we filled in the initialized table with each industry as a column, and data listed underneath. After that, we converted the type of data to numeric for each column. Last, a final column was added to include dates for each data entry, concluding the data wrangling process.

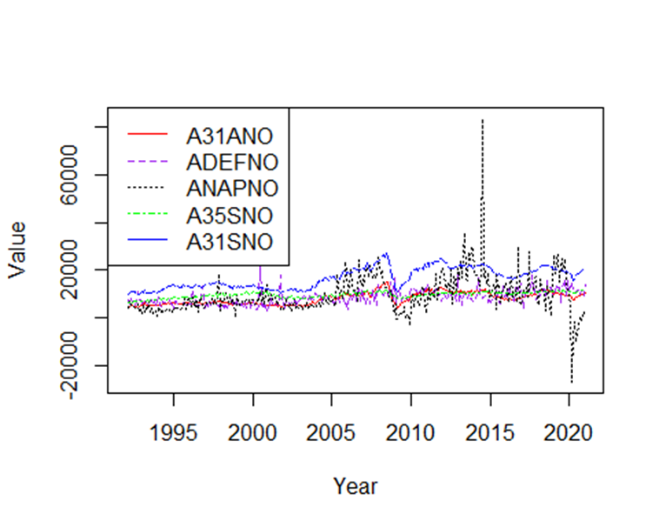
The second part of time series analysis is clustering. Clustering is the process of grouping similar sets of data based on certain similarities and criteria determined by the person analyzing the data. It is used to both cut down on the amount of data being worked with, while also serving as a means of analyzing said data [4]. Clustering is almost always done using an algorithm due to the data sets being too vast and numerous for it to be done by hand. While there are several different ways to cluster data, we have chosen three methods: mean, median, and standard deviation.

In order to cluster each industry, R packages were heavily utilized, specifically one called “matrixStats.” From this package, the function “colMeans” was used to find the mean value for each industry column of the data frames. The industry means were then converted into a list and added as a new row to the bottom of the data frames. The columns of each data frame were then rearranged according to their mean value in ascending order, along with separating the data frames into “adjusted” and “unadjusted” industries. Furthermore, a function was written that plots each industry column as a time series from the sorted data frames, by clustering the industries into groups of five. Axis titles, a legend, different line types, and color were added to each individual time series to make the plots both easier to understand and nicer to look at, concluding the process of clustering by mean.

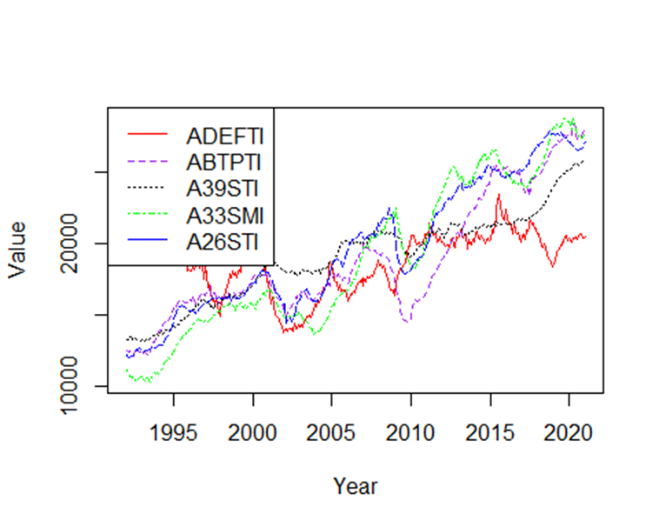
The method for clustering the median and standard deviation both closely follow the process for clustering the mean. A new row was added to each data frame to include the median for each column, which was found using the “matrixStats” function “colMedians.” The same was done for standard deviation, utilizing the function “colSds” instead. The columns of each data frame were then put into ascending order based on their respective clustering types. Once again, a function was written to plot time series in clusters of 5, with the same previous design aspects added for elucidation.

It is important to note that there are instances where values are missing in places in many industries across the excel sheets that contain the data. The code was written to include only the values that were ascertained. As such, the actual number of values used to calculate each statistical measure will vary. Furthermore, the number of industries in each table was not always divisible by five. Thus, some of the plots of the clusters may have slightly more or less than five industries included, however, the orders determined by mean, median, and standard deviation values have been preserved.

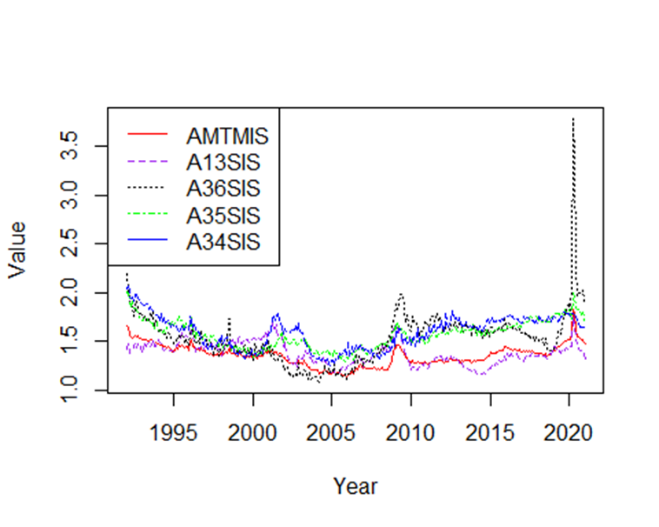
Included below are selected clusters that were particularly accurate in grouping industry time series by their respective measures:



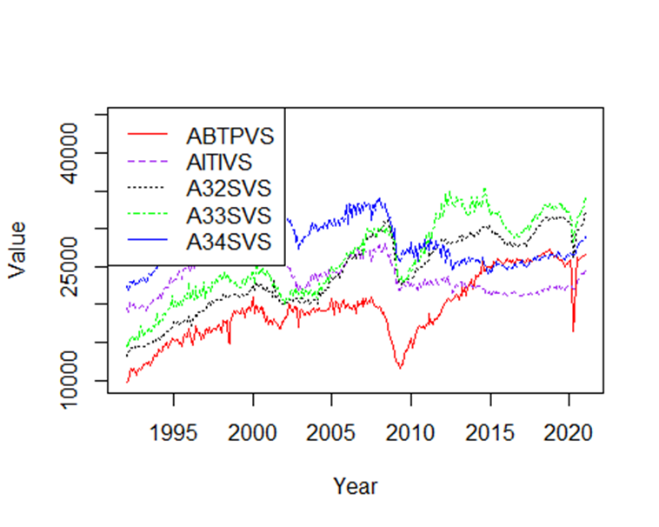
This cluster is from the New Orders data set and includes the following industries in the order of the list in the legend, clustered by their mean value: Iron and Steel Mills and Ferroalloy and Steel Product Manufacturing from Purchased Steel, Defense Capital Goods, Nondefense Aircraft and Parts, Electrical Equipment, Appliances, and Components, Primary Metals. These industries all seem to involve steel products, which is likely the cause of their like characteristics. Notice that excluding a few spikes and drops in the Nondefense Aircraft and Parts industry, the similar shape and proximity of the time series in the cluster is what we are looking to find.



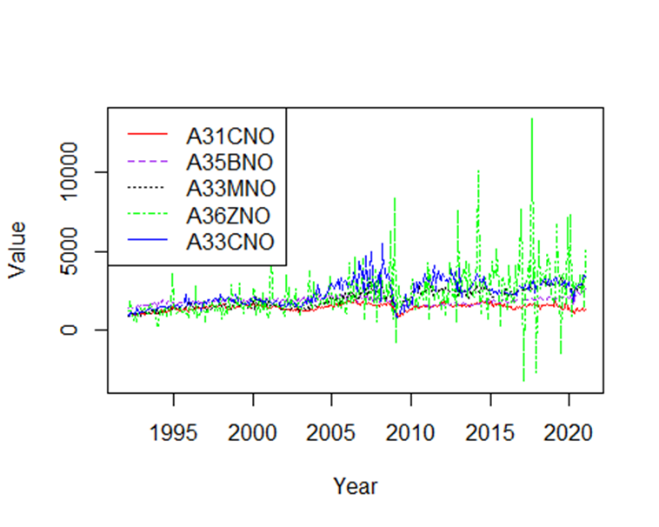
This cluster is from the Total Inventories data set and includes the following industries in the order of the list in the legend, clustered by their mean value: Defense Capital Goods, Motor Vehicle Bodies, Trailers, and Parts, Miscellaneous Products, Machinery (Materials and Supplies Inventories), Plastic and Rubber Products. This cluster is a little more difficult to interpret, but it may be that all these industries supply parts and materials to build larger machines and goods. The time series are again in close proximity to each other and similar in shape.



This cluster is from the Inventories to Shipments data set and includes the following industries in the order of the list in the legend, clustered by their mean value: Total Manufacturing, Textile Mills, Transportation Equipment, Electrical Equipment, Appliances, and Components, Fabricated Metal Products. This is another cluster of industries that are likely providing materials and goods for further manufacturing. What is interesting about this cluster is the spike that occurred in 2020 for all the industries, none more pronounced than Transportation Equipment. We suspect that this massive boom in the industry came from the sudden need to ship medical supplies all over the world to fight the COVID-19 pandemic.



This cluster is from the Shipments data set and includes the following industries in the order of the list in the legend, clustered by their median value: Motor Vehicle Bodies, Trailers, and Parts, Information Technology Industries, Fabricated Metal Products, Machinery, Computer and Electronics Products. These industries seem to be centered around electronic goods and the means of transporting them. While the time series in this cluster may not be very near each other, the display a very similar shape.



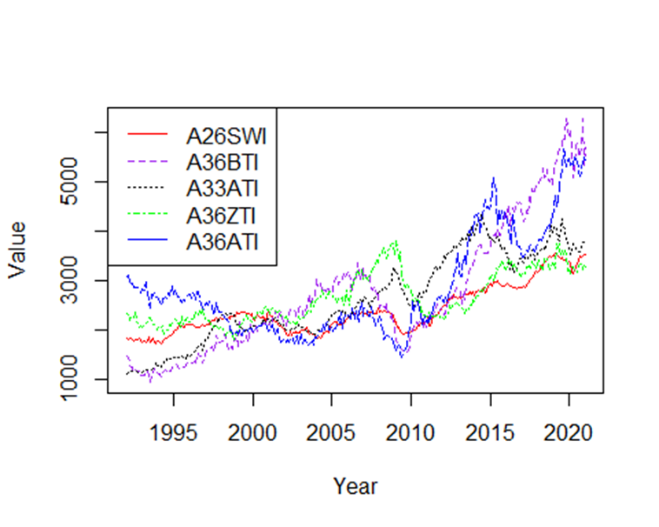
This cluster is from the New Orders data set and includes the following industries in the order of the list in the legend, clustered by their median value: Ferrous Metal Foundries, Household Appliance Manufacturing, Material Handling Equipment Manufacturing, Ships and Boats, Construction Machinery Manufacturing. These industries all seem to be centered around construction of new homes, machines, and modes of transportation. This is our first example of how our clustering methodologies may be flawed. Notice that almost all the industries in the cluster have a similar shape and proximity, except for the Ships and Boats industry in green. This industry is a lot more volatile, yet its mean was similar enough to the other industries that it was included in this cluster. When observing the industries in the cluster, it even seems that Ships and Boats is apart from the others in nature. This is a prime example of why clustering can be so difficult to do efficiently.

Now that we have imparted the characteristics we desire in each cluster, the following examples will include only names of the industries that have been included in the cluster. See if you can determine what makes the cluster “good” and what might make the cluster “bad”.

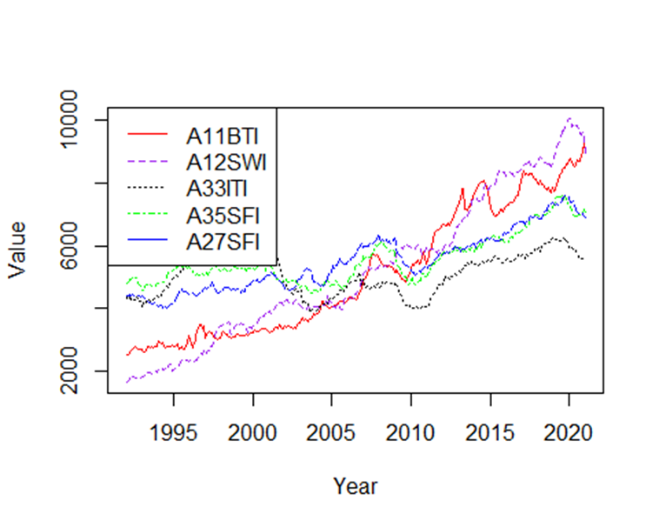
Chart, histogram

Description automatically generated

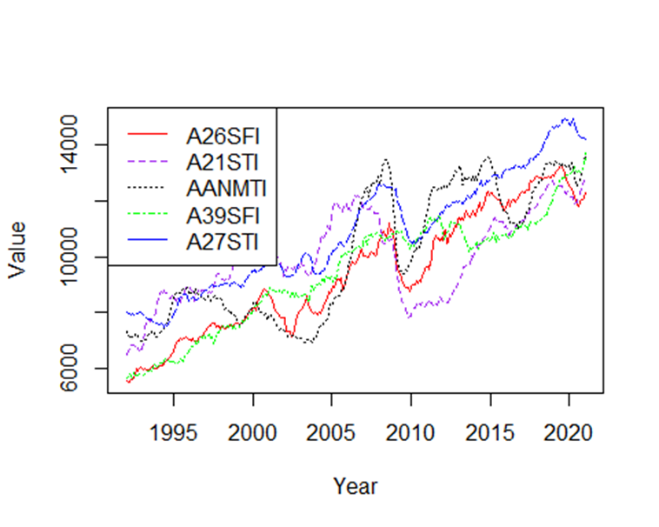
This cluster is from the New Orders data set and includes the following industries in the order of the list in the legend, clustered by their median value: Iron and Steel Mills and Ferroalloy and Steel Product Manufacturing from Purchased Steel, Defense Capital Goods, Nondefense Aircraft and Parts, Electrical Equipment, Appliances, and Components, Primary Metals.



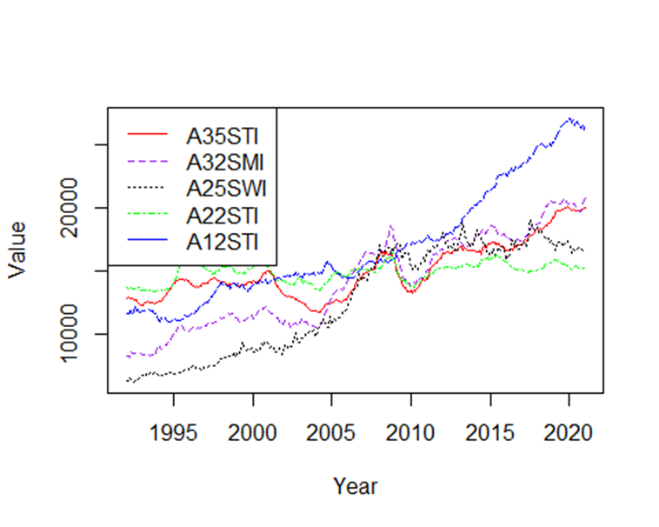
This cluster is from the Total Inventories data set and includes the following industries in the order of the list in the legend, clustered by their median value: Plastic and Rubber Products (Work in Process Inventories), Light Truck and Utility Vehicle Manufacturing, Farm Machinery and Equipment Manufacturing, Ships and Boats, Automobile Manufacturing.



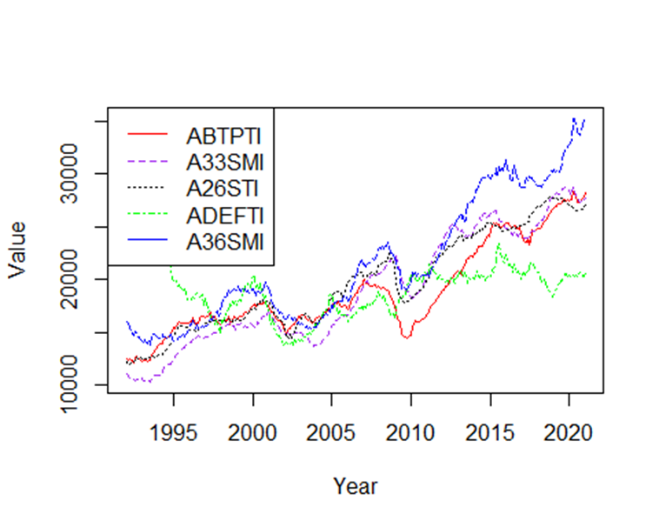
This cluster is from the Total Inventories data set and includes the following industries in the order of the list in the legend, clustered by their median value: Dairy Product Manufacturing, Beverage and Tobacco Products (Work in Process Inventories), Metalworking Machinery Manufacturing, Electrical Equipment, Appliances, and Components (Finished Goods Inventories), Nonmetallic Mineral Products (Finished Goods Inventories).



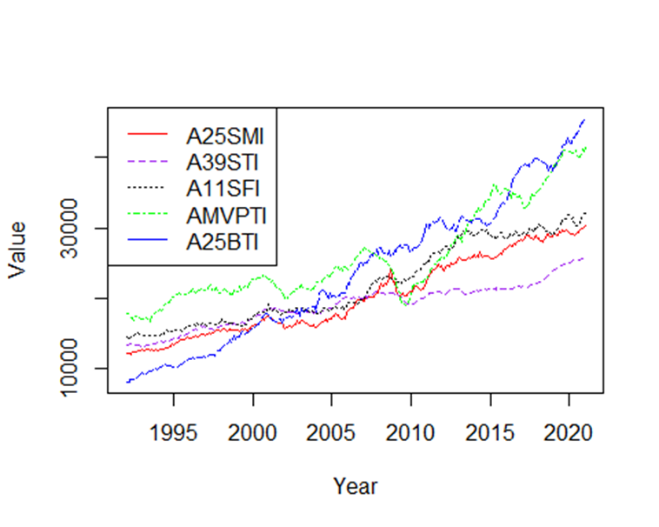
This cluster is from the Total Inventories data set and includes the following industries in the order of the list in the legend, clustered by their median value: Plastic and Rubber Products (Finished Goods Inventories), Wood Products, Aluminum and Nonferrous Metal Products, Miscellaneous Products (Finished Goods Inventories), Nonmetallic Mineral Products.



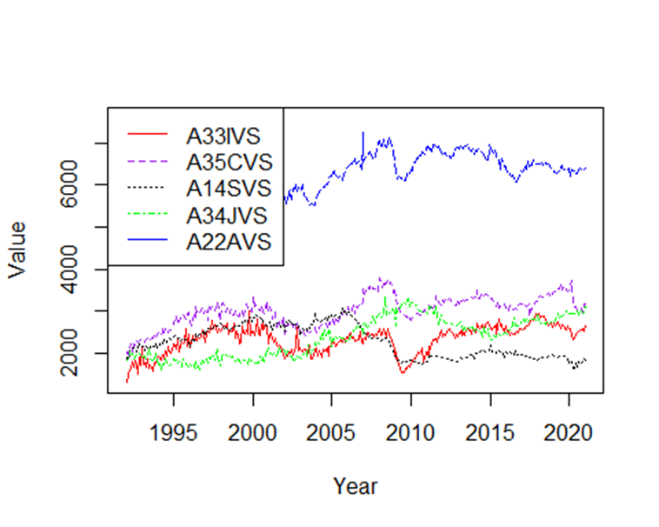
This cluster is from the Total Inventories data set and includes the following industries in the order of the list in the legend, clustered by their median value: Electrical Equipment, Appliances, and Components, Fabricated Metal Products (Materials and Supplies Inventories), Chemical Products (Work in Process Inventories), Paper Products, Beverage and Tobacco Products.



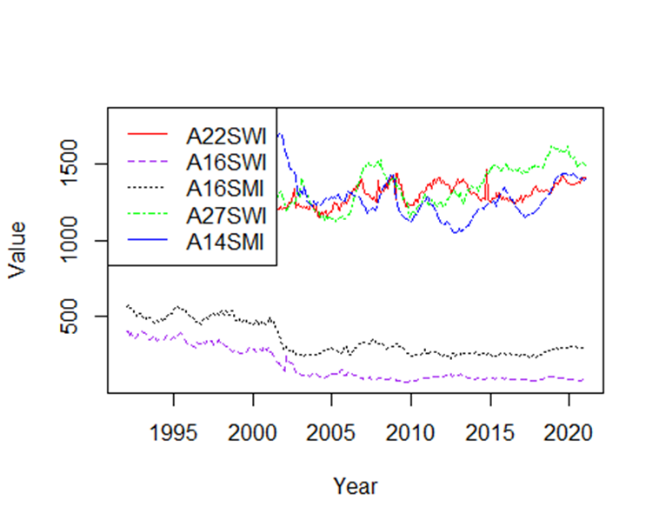
This cluster is from the Total Inventories data set and includes the following industries in the order of the list in the legend, clustered by their median value: Motor Vehicle Bodies, Trailers, and Parts, Machinery (Materials and Supplies Inventories), Plastic and Rubber Products, Defense Capital Goods, Transportation Equipment (Materials and Supplies Inventories).



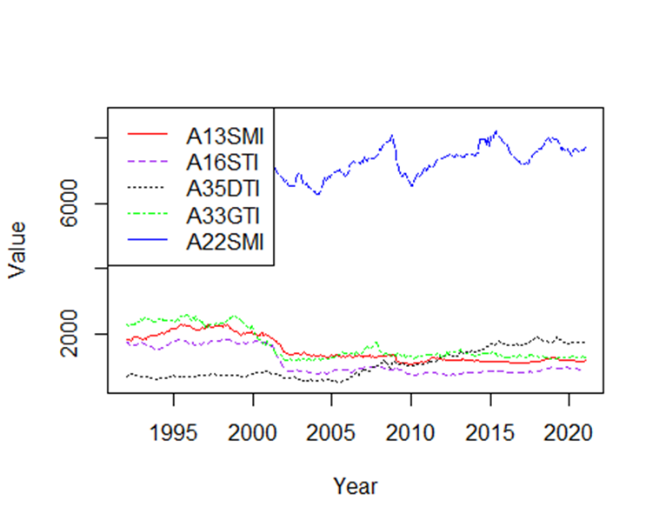
This cluster is from the Total Inventories data set and includes the following industries in the order of the list in the legend, clustered by their median value: Chemical Products (Materials and Supplies Inventories), Miscellaneous Products (Finished Goods Inventories), Food Products, Motor Vehicle and Parts, Pharmaceutical and Medicine Manufacturing.



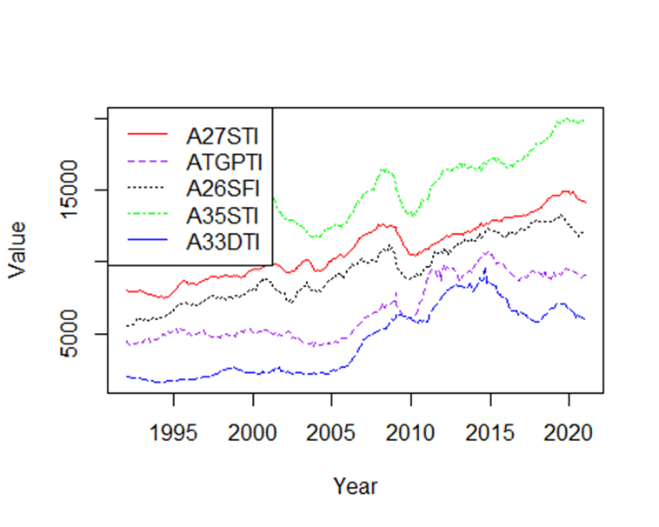
This cluster is from the Shipments data set and includes the following industries in the order of the list in the legend, clustered by their median value: Metalworking Machinery Manufacturing, Electrical Equipment Manufacturing, Textile Product Mills, Search, Detection, Navigation, Guidance, Aeronautical, and Nautical System and Instrument Manufacturing, Defense, Pulp, Paper, and Paperboard Mills.



This cluster is from the Total Inventories data set and includes the following industries in the order of the list in the legend, clustered by their standard deviation value: Paper Products (Work in Process Inventories), Leather and Allied Products (Work in Process Inventories), Leather and Allied Products (Materials and Supplies Inventories), Nonmetallic Mineral Products (Work in Process Inventories), Textile Product Mills (Materials and Supplies Inventories).



This cluster is from the Total Inventories data set and includes the following industries in the order of the list in the legend, clustered by their standard deviation value: Textile Mills (Materials and Supplies Inventories), Leather and Allied Products, Battery Manufacturing, Photographic Equipment Manufacturing, Paper Products (Materials and Supplies Inventories).



This cluster is from the Total Inventories data set and includes the following industries in the order of the list in the legend, clustered by their standard deviation value: Nonmetallic Mineral Products, Turbines, Generators, and Other Power Transmission Equipment, Plastic and Rubber Products, Electrical, Equipment (Finished Goods Inventories), Appliances, and Components, Mining and Oil and Gas Field Machinery Manufacturing.

While the clusters above were selected because of their interesting properties, there was no discernable connection across measurements or method of clustering for any of these industry clusters. However, there were certain clusters that exhibited a strong connection across the Shipments, New Orders, Unfilled Orders, and Total Inventories tables no matter the clustering method. These clusters included

Cluster #1: Nondefense Capital Goods Excluding Aircraft, Durable Goods (Materials and Supplies Inventories), Durable Goods (Work in Process Inventories), Consumer Nondurable Goods, Nondefense Capital Goods.

Cluster #2: Total Manufacturing (Work in Process Inventories), Capital Goods, Consumer Goods, Total Manufacturing (Materials and Supplies Inventories), Total Manufacturing (Finished Goods Inventories).

Cluster #3: Nondurable Goods, Durable Goods Excluding Transportation, Manufacturing with Unfulfilled Orders, Durable Goods Excluding Defense, Durable Goods.

Note that the various time series graphs associated with these clusters have been included in the Appendix, and that the ordering of these clusters is not the same for each measurement. It may vary slightly.

Not only were these clusters consistently the similar, both in shape and proximity, but they were always the last three clusters generated by the code, meaning that these industries had some of the largest mean, median, and standard deviation values. Why? The key lies in the fact that these industries are not really industries at all. They are measures of all the industries included as a whole. Many of these “industries” are even the same measurement minus a key element. For example, in Cluster #2 there are three different versions of “Total Manufacturing” dependent upon different aspects of the measure. You may then ask why we are still interested in these clusters if they do not give insight into the connections of standalone industries. While this is true, the main reason is that these clusters provide proof that our algorithms for creating clusters work well on a broad scale. When we began this research, we had no idea that there were data points that were intended to track economic activity across more than one industry, let alone many of them. It was only after we implemented our methodologies that we found this connection, which was exactly the point of clustering this time series data in the first place. This is the true power of the cluster, the ability to take large data sets that cannot be combed through by hand and find these connections that may not be obvious from the start.

To conclude, we consider the methods of clustering we employed to be successful. There were many clusters that we did not include in this report, as they did not present any useful information about the industries they contained. This is to be expected, as not all industries will have hidden connections waiting to be found. Still, we did find some clusters that were able to show us the characteristics we were looking for and allowed us to make inferences about those connections. While the methods we employed in this study were rudimentary, there are more and more advanced methods of clustering being developed every day to find even more deeply hidden connections across the vastness of our national economy. As the efficiency of these algorithms increases, the ability to make effective business and economic policy decisions increases as well. These decisions and their consequences will allow us to strengthen our economy across all industries and build toward a better tomorrow.

References

[3] *Data wrangling: What it is & why it's important*. Business Insights Blog. (2021, January 19). Retrieved March 29, 2022, from https://online.hbs.edu/blog/post/data-wrangling

[1] Hayes, A. (2022, February 8). *Understanding time series*. Investopedia. Retrieved March 24, 2022, from https://www.investopedia.com/terms/t/timeseries.asp

[2] US Census Bureau: Adriana Stoica (M3 Section Chief). (2009, September 30). *US Census Bureau Manufacturer' shipments, inventories, and orders main page*. United States Census Bureau. Retrieved March 29, 2022, from https://www.census.gov/manufacturing/m3/historical\_data/index.html

[4] *What is clustering?* NVIDIA Data Science Glossary. (n.d.). Retrieved March 31, 2022, from https://www.nvidia.com/en-us/glossary/data-science/clustering/

Appendix

Included below are the graphs associated with Cluster #1, Cluster #2, and Cluster #3:

Shipments by Mean (Adjusted)

Chart, histogram

Description automatically generated Chart, histogram

Description automatically generated Chart, histogram

Description automatically generated

New Orders by Mean (Adjusted)

Chart, histogram

Description automatically generated Chart, histogram

Description automatically generated Chart, histogram

Description automatically generated

Unfilled Orders by Mean (Adjusted)

Chart

Description automatically generated Chart, histogram

Description automatically generated Chart, histogram

Description automatically generated

Total Inventories by Mean (Adjusted)

Chart, histogram

Description automatically generated Chart, histogram

Description automatically generated Chart, histogram

Description automatically generated

Shipments by Median (Adjusted)

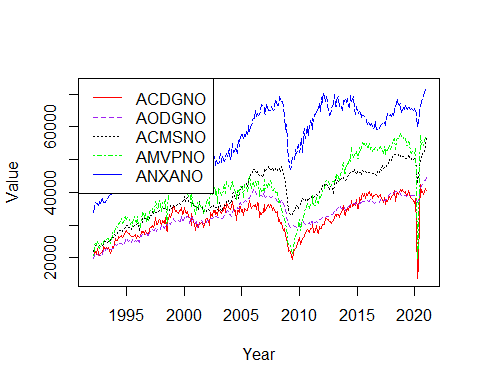
Chart, histogram

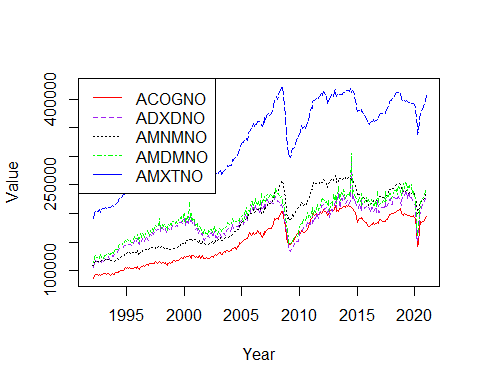
Description automatically generated Chart, histogram

Description automatically generated Chart, histogram

Description automatically generated

New Orders by Median (Adjusted)

 Chart, histogram

Description automatically generated 

Unfilled Orders by Median (Adjusted)

Chart

Description automatically generated Chart, histogram

Description automatically generated Chart, histogram

Description automatically generated

Total Inventories by Median (Adjusted)

Chart, histogram

Description automatically generated Chart, histogram

Description automatically generated Chart, histogram

Description automatically generated

Shipments by Standard Deviation (Adjusted)

Chart, histogram

Description automatically generated Chart, histogram

Description automatically generated Chart, histogram

Description automatically generated

New Orders by Standard Deviation (Adjusted)

Chart, histogram

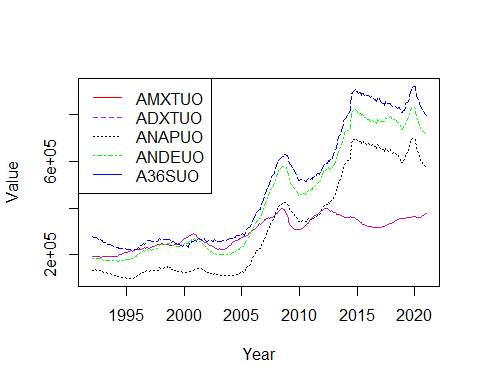
Description automatically generated Chart, histogram

Description automatically generated Chart, histogram

Description automatically generated

Unfilled Orders by Standard Deviation (Adjusted)

Chart, line chart

Description automatically generated  Chart, histogram

Description automatically generated

Total Inventories by Standard Deviation (Adjusted)

Chart, histogram

Description automatically generated Chart, histogram

Description automatically generated Chart, histogram

Description automatically generated