Mapping the United States through TDA

Instructor - Rubana Syed, ECON 111-800

Ryan Johnson

2024-12-11

Table of Contents

1 Introduction[1]

Since the late 19th Century, statistical techniques have been applied to the field of Economics. Common methodologies include generalized linear models or time series analysis. Although technology has advanced we still use these traditional methods. However, increasingly, new methods are being tested and developed.

The goal of this paper is examine the macroeconomic landscape of the United States of America (US) through a lens of Topological Data Analysis (TDA). Specifically, we are going to apply a TDA algorithm called BallMapper (TDABM) developed by Pawel Dlotko. This new algorithm simplifies traditional TDA methods by reducing the need for the user to pass data through multiple functions. Dlotoko demonstrates that, mathematically, TDABM produces a near similar output to that of traditional TDA methods via a fuctor map (Paweł Dłotko 2019). Thus, one benefit of TDABM compared to the traditional methods is, for those who have the tools and introductory understanding in Topology or TDA, TDABM could be a good entry point to the field.

What follows below is primarily a proof-of-concept for TDA BallMapper and closely follows other papers that have applied TDABM. We will begin with a general overview of TDABM: current published literature and useful concepts to know. We then will describe our data and reasoning behind their selection. Our Methodology and Results will touch on more theoretical concepts behind TDABM and then adress a couple of findings, respectively.

Our hope here is for the reader to walk away understanding the general motivation behind TDA BallMapper, how it differs from other common statistical methods, and how methods such as TDA could be additive to our current body of mathematical tools.

[1] Supplementary materials referenced are available upon request via email: johnson.ryan1019@gmail.com.

2 Literature

TDA is a relatively new field and is increasingly being applied to various areas of research such as genetics, financial fraud, biology, and image detection. From the small body of literature we have reviewed, TDABM has removed many of the parameters that are required for more traditional methods of TDA. Generally speaking, traditional TDA contains four broad steps, each with multiple user-defined parameters. Conversely, TDABM has reduced this process to an the user selecting their data, a coloring variable and an epsilon value – details on this follow below. What should be noted, although TDABM reduces the number of steps needed to produce similar outcomes of traditional TDA methods, we lose control of being able fine tune out outputs. We also find that interpreting results becomes more difficult. However, TDABM is still in its infancy, so there is not a large body of research on interpretation. (Paweł Dłotko 2019) One major motivation for this paper is a paper written by Paweł Dłotko, Rudkin, and Qiu (2019) that applied TDABM to a global macroeconomic dataset to compare how countries have evolved over time, their transformation from the Great Depression Era, and various views on wealth and inequality. We could not find other TDA literature found specifically focused on the macroeconomic economy of singular countries. Hence, our topic of choice.

3 Data

3.1 Sources

Three major sources of data were used for analysis: The Bureau of Economic Analysis (BEA), The Federal Reserve Board (FRB), and The Bureau of Labor Statistics (BLS).[2] We initially approached these sources[3] by selecting commonly reported data about the economy, both on an annual and quarterly basis. Since this paper is introductory and limited in nature, we chose to focus on annual data that represented overall macroeconomic measures, covered a sufficient time period, and included a range of historical economic events.

[2] The Federal Reserve Economic Data (FRED) website was used to download BLS data.

[3] See *Data Sources*

3.2 Construction

A primary challenge we encountered when selecting variables for analysis was the variability of time period coverage. Some data spanned from Depression Era 1929 all the way up to 2023. Looking at our final nine variables (see Figure 2]) and the range of years available by by source (Figure 1), we limited our years to 1958 - 2023 due to *CPIA* data having the smallest range of years.

Some of the data was only available by month. In these cases we calculated annual changes by finding the difference between January of the current year and January of the previous year. Note, this reduced the number of years in our data set by one. We also standardized the data by converting non-rate based data to percent changes from the previous year to ensure consistency across all variables (Equation 1). One feature of TDABM is that it can be sensitive to data with large scale differences. This not only reduces how precise you can be when selecting a parameter, but can also significantly increase computation time.

Looking at our correlation table (Figure 2), we observe moderate to strong correlations between the pairwise combinations of GDP, PCE, Ig, and Imports. We also see a strong correlation between CPI and Interest Rates. Since these correlations are moderate to high, it's worth noting that the shape of our output may be reflected by these correlations. They also could express themselves with similar coloring of TDABM graphs (Paweł Dłotko, Rudkin, and Qiu 2019).

```
-- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
v dplyr     1.1.4     v readr     2.1.5
v forcats   1.0.0     v stringr   1.5.1
v ggplot2   3.5.2     v tibble    3.3.0
v lubridate 1.9.4     v tidyr     1.3.1
v purrr     1.1.0
-- Conflicts ------------------------------------------ tidyverse_conflicts() --
x dplyr::filter() masks stats::filter()
```

```
x dplyr::lag()    masks stats::lag()
i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become
Loading required package: viridisLite


Attaching package: 'kableExtra'


The following object is masked from 'package:dplyr':

    group_rows


Loading required package: data.table


Attaching package: 'data.table'


The following objects are masked from 'package:lubridate':

    hour, isoweek, mday, minute, month, quarter, second, wday, week,
    yday, year
```

```
The following objects are masked from 'package:dplyr':

    between, first, last
```

```
The following object is masked from 'package:purrr':

    transpose
```

```
Note: As of February 2018, beaGet() requires 'TableName' for NIPA and NIUnderlyingDetail data
```

```
Packages loaded. BEA API loaded

[1] "FFDF158B-A64F-401C-B19D-A48365418843"
```

4 Methodology

4.1 TDA BallMapper

TDABM, and in general TDA, believes that data contains a “shape” to it. It pulls ideas from the branch of Mathematics, Topology. At a high-level, Topology focuses on the properties of geometric shapes when you deform them without breaking them, e.g. bend, twist, scrunch. The following is a basic overview of TDABM [4].

[4] Some very good examples and explanations can be found in (Dlotko, Qiu, and Rudkin 2022) and (Paweł Dłotko, Qiu, and Rudkin 2019) which touch on the finer details of TDABM algorithm. A very good illustrative example can be seen in the pre-print (P. Dłotko, Qiu, and Rudkin 2021).

4.2 BallMapper Algorithm

Before going over the algorithm there are a few For the following algorithm some useful definitions are as follows: - We will use the word pointcloud instead of dataset. - Conventionally the Euclidean distance metric is used; let $d = dist() is used as our distance metric, denote $ The core steps of the TDABM algorithm are as follows:

1. From your dataset (pointcloud) select a random point, α_n, and draw circle with radius epsilon, ϵ.
2. For some distance metric where d is the distance from, any points inside the circle will be associated with point α_n.
3. Repeat steps 1 - 3 until there are no more points to select and you have a set α_1,...,α_n.
4. Draw an edge between α_i and α_j if they share a point(s) for our set α_1,...,α_n.

Although the above list five steps, it should be noted that this is just the algorithm and is all done with one function call.[5] From the steps above, the only thing the user of TDABM is responsible for is creation of a pointcloud and choosing an appropriate ϵ.

4.3 Cloud Building

In our Data section, we noted that some calculations were needed before our analysis. Additionally, for the Federal Funds Rate, the Unemployment Rate, and CPI we had to take extra steps and considerations while constructing out pointcloud. Since the Federal Funds Rate and the Unemployment Rate are already a rate-based measurements, instead of taking a percent change from the previous year, we instead took the standard non-weight average rate by each year. For CPI, this data was given on a monthly-basis, we calculated it using our General Percent Change Formula with $k = 12$.

[5] There are packages available both in R and Python under the names BallMapper and pyBallMapper, respectively.

5 Results

The analysis reportes here are guided by Paweł Dłotko, Rudkin, and Qiu (2019)'s approach. We will be referencing our TDABM output in Figure 3 for the remainder of this section.

5.1 Interpretation of TDA BallMapper Graphs, An Abridged Version

In Figure 3 we are presented with what looks like a graph, in the mathematical sense; a display of points (nodes) and edges –some connected while others are not connected at all (satellites). If we do not change the poincloud or ϵ of our TDABM graph, we will get the same "shape" of the graph regardless of the coloration we choose.

Nodes are colored depending on the users interest. In (**?**), we are focusing color by year. Coloration calculated by taking the average of the data that is "inside" each node, in our case the average value of the years. Further, if there is only one data point in a node (i.e. a satellite points), the node will appear small and the coloration will reflect as seen on the legend. Conversely, if the node is large, this indicates more data is "inside" the node, and the average value could be reflecting the mean values of a large or small variance.

5.2 General Figure 3 Remarks

We initially notice a large connected component[6] (component A) on the right side of our graph, as well as a smaller connected component (component B) on the lower left side. Noting the coloration, there are three main sections: the right, bottom, and upper-left. On the bottom and sweeping upward to the left we see there are multiple satellite points. These can be of interest because they may indicate outliers in our data.[7] Focusing back towards component A, we notice that the two largest nodes, 28 and 22, which seem to represent the late 2010s and the dot-Com boom, respectively. We also note that there is a arm coming off the upper-right portion of component A, as well as smaller, lollipop, features emerging at the bottom and top of component A.[8]

[6] In graph theory, a connected component is one in which there exists a path from a node to every other node for a set of nodes and edges.

[7] We have not found any literature yet on whether this observation is empirically true.

[8] The observations mentioned above good starting places for interpreting BallMapper graphs.

5.3 Time Travel

Looking at Figure 3, we see that our graph is colored by year.[9] In this section we will only highlight a few observations in detail due to the nature of this paper. Additional areas for investigation can be found in the Conclusion.

Mentioned above, we observed the satellite points in the lower half of the graph. If we look at the upper-right quadrant, we see that it consists of nodes $\{1, 16, 25, 27\}$. Looking at Table 2 we see that it covers the 2020 Pandemic and the Financial Crisis and its aftermath. Seeing a general coloration from the recent 10 - 15 years we would expect to see some of the recent outlier economic events here. Indeed, node 25 consists of two of the most recent economic downturns, 2009 and 2020. However, what is interesting is that node 27 consists of three years closely following the 2008-2009 Financial crisis. However, the years following the 2020 COVID Pandemic can be found scattered across nodes $\{6, 8, 24, 28, 29\}$. Looking at node 27 we find that the common thread is high unemployment. In contrast, the years following the Pandemic are not the same year to year.[10]

Shfiting our focus towards component A, it consists of data most similar to the economic years between the 1990s - 2010s. Meanwhile, looking at component B and the general lower area of our graph, we see that it represents the Great Inflation period of the 1970s - early 1980s [ivestopedia article]. Looking at nodes $12-20$, we get exactly ten years of data[11] where there was know high inflation, high unemployment, and in general know to be bad time economically for the US (Table 3). Something of interest to note is that nodes 14 and 15 consist of 1976 - 1978. This three year period based on data seems to show more normal economic conditions. On the other hand, we see that all the other years in this grouping (nodes $12-20$) show signs of a struggling economy in some way.

[9] Supplementary graphs displaying coloration based on different variables can be found in Supplement_1.pdf.

[10] *2021:* An outlier, we see higher levels of unemployment and inflation. We see high PCE, Ig, and imports; *2022:* Nodes $\{8, 28\}$ we see that there is higher than normal inflation, export, and imports; *2023:* Nodes $\{6, 24\}$, we see low inflation and low unemployment.

[11] 1974 - 1984

6 Conclusion

This paper serves as a proof-of-concept of the usefulness of TDABM as a methodology for exploratory data analysis. By no means is this a totally comprehensive method to gain insight into data, but instead TDA BallMapper proves itself to be useful tool wehen employed in addition to other traditional analyses.

In our Results section we briefly went over two ways to interpret TDABM graphs. As seen in many other, longer papers written by Dlotko and colleagues, there is a lot of room to expound our TDABM graph here,though that is beyond the scope of this paper. In general, though, we can see that TDBM can give insight into our data that we might not see otherwise. Seeing that there is data that is not like the others gives an indication that there is something to investigate. Consider our comparison of the 08' - 09' financial crisis and the 2020 COVID Pandemic: We see learned that the years following the financial crisis were similar economically. Another way to view this is, clustering revels that the financial crisis was a singular problem that spurred the collapse of the financial system. Conversely, the COVID pandemic was a singular issue but exposed many different weaknesses in our current world.

For the future development of TDABM, and TDA in general, there are many possibilities of future research and one glaring downside. The downside to TDA is that it is a high barrier to entry to understand the methodologies inner workings. Those with a technical background will have an easier time, but nonetheless a higher barrier than methods such as linear regressions. Subject wise, TDA has an advantage in some of the social sciences because it can be somewhat of a bridge between qualitative and quantitative analyses. TDABM has been used to analyze the Brexit Vote data, seeking to understand quantitative and qualitative motivations behind its outcome Rudkin et al. (2023). TDA also has the advantages of dimension reduction which could be useful to the life-sciences. Fields such as genetics, chemistry, and biology have many fields of data, and being able to consolidate it into an understandable form could benefit the body of knowledge greatly.

Endnotes

General Percent Change Formula

$$\Delta p_n = \frac{p_n - p_{n-k}}{p_{n-k}} \text{ for some } p \in P \text{ and } n > 0 \tag{1}$$

where P is a column of our data with $x_1, ..., x_n$ observations and k is a lag integer such that $k > 0$.

Appendix

Table 1: Summary Statistics (%)

measure	Minimum	Q25	Median	Q75	Maximum	Mean	SD	VAR
GDP	-2.60	2.15	2.95	4.40	7.20	3.03	2.16	4.68
PCE	-2.50	2.32	3.10	4.35	8.80	3.23	1.97	3.89
Ig	-21.00	-0.05	6.00	9.15	27.30	4.25	8.38	70.30
Exports	-3.20	0.52	1.90	3.27	8.70	1.94	2.32	5.39
Imports	-13.50	2.32	6.55	8.80	18.80	5.25	6.17	38.10
Gov	-12.60	2.45	5.25	10.60	24.30	5.67	6.61	43.69
InterestRateA	0.08	1.85	4.56	6.56	16.39	4.72	3.57	12.76
CPIA	0.66	1.97	2.71	4.69	12.15	3.68	2.53	6.41
AvgUnrateA	3.49	4.86	5.61	6.98	9.71	5.92	1.60	2.55

Table 2: COVID-19+ (% change)

	GDP	PCE	Ig	Gov	Exports	Imports	FFR	CPI	UnRate
1958	-0.7	0.9	-7.2	3.3	-13.5	4.7	1.57	2.047782	6.841667
1979	3.2	2.4	3.5	1.8	9.9	1.7	11.20	11.323529	5.850000
2009	-2.6	-1.3	-21.0	3.6	-8.3	-12.6	0.16	1.823672	9.283333
2011	1.6	1.7	6.7	-3.2	7.2	4.8	0.10	2.276663	8.933333
2012	2.3	1.4	11.0	-2.1	4.0	2.4	0.14	1.899694	8.075000
2013	2.1	1.7	7.4	-2.4	3.0	1.2	0.11	1.740857	7.358333
2020	-2.2	-2.5	-4.5	3.4	-13.1	-9.0	0.37	1.623938	8.091667

Table 3: The Great Inflation (% change)

	GDP	PCE	Ig	Gov	Exports	Imports	FFR	CPI	UnRate
1974	-0.5	-0.8	-6.6	2.2	7.9	-2.3	10.51	11.349036	5.641667
1975	-0.2	2.3	-16.2	2.2	-0.6	-11.1	5.82	6.730769	8.475000
1976	5.4	5.6	19.1	0.5	4.4	19.6	5.05	6.126126	7.700000
1977	4.6	4.2	14.3	1.2	2.4	10.9	5.54	6.451613	7.050000
1978	5.5	4.4	11.6	2.9	10.6	8.7	7.94	8.452951	6.066667
1979	3.2	2.4	3.5	1.8	9.9	1.7	11.20	11.323529	5.850000
1980	-0.3	-0.3	-10.1	1.8	10.8	-6.7	13.35	12.153237	7.175000
1981	2.5	1.4	8.8	1.0	1.2	2.6	16.39	9.540636	7.616667
1982	-1.8	1.5	-12.6	1.8	-7.7	-1.3	12.24	4.516129	9.708333
1983	4.6	5.7	9.3	3.7	-2.6	12.6	9.09	4.732510	9.600000
1984	7.2	5.3	27.3	3.5	8.2	24.3	10.23	4.911591	7.508333

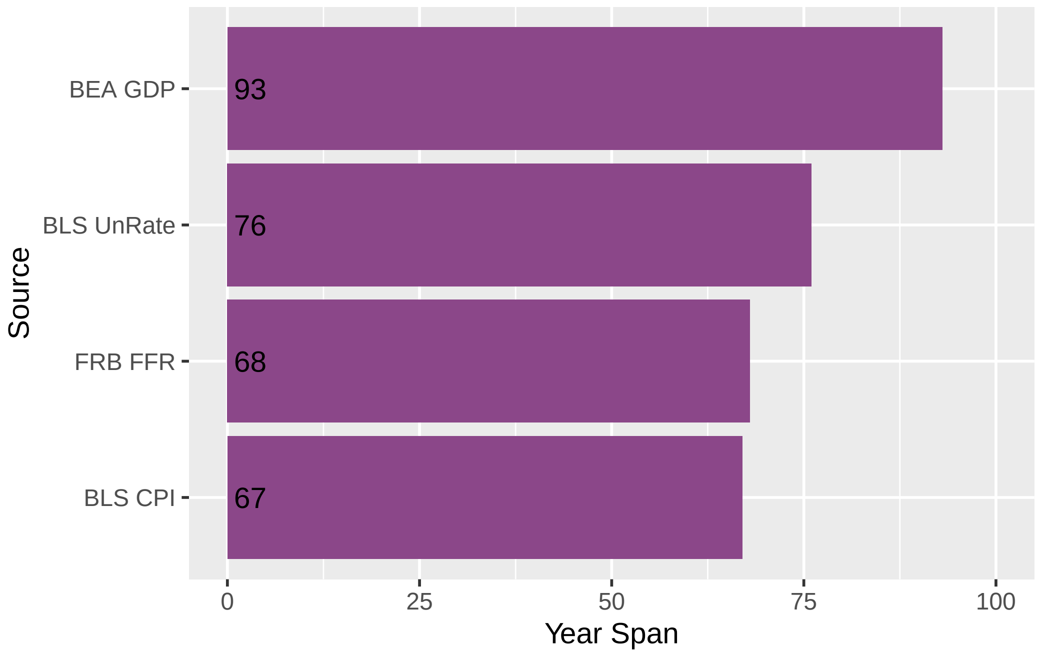

Figure 1: Year Coverage by Data Source

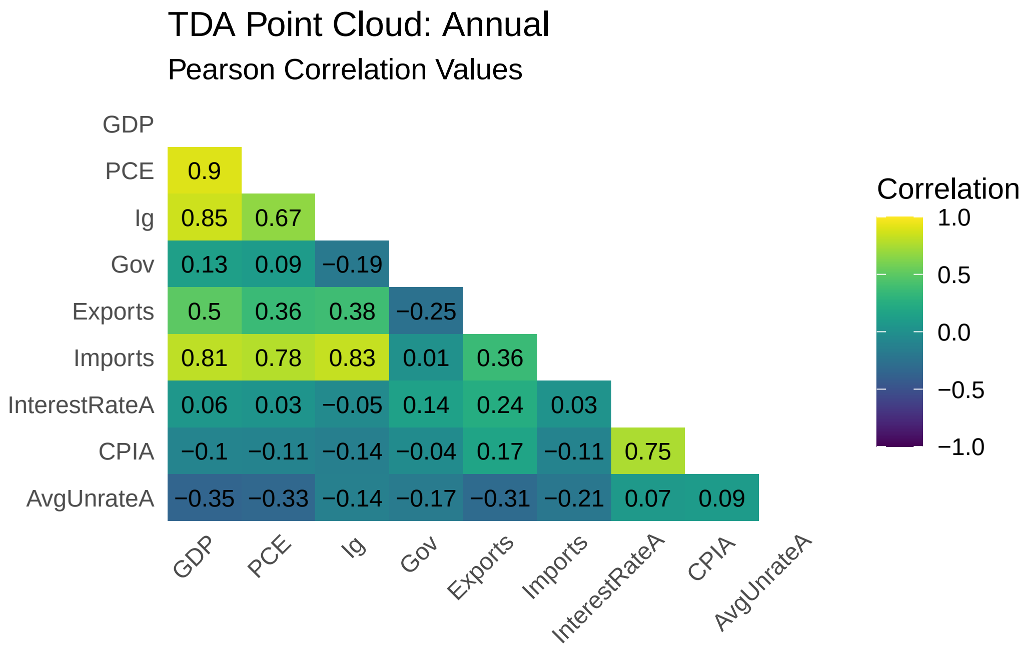

Figure 2

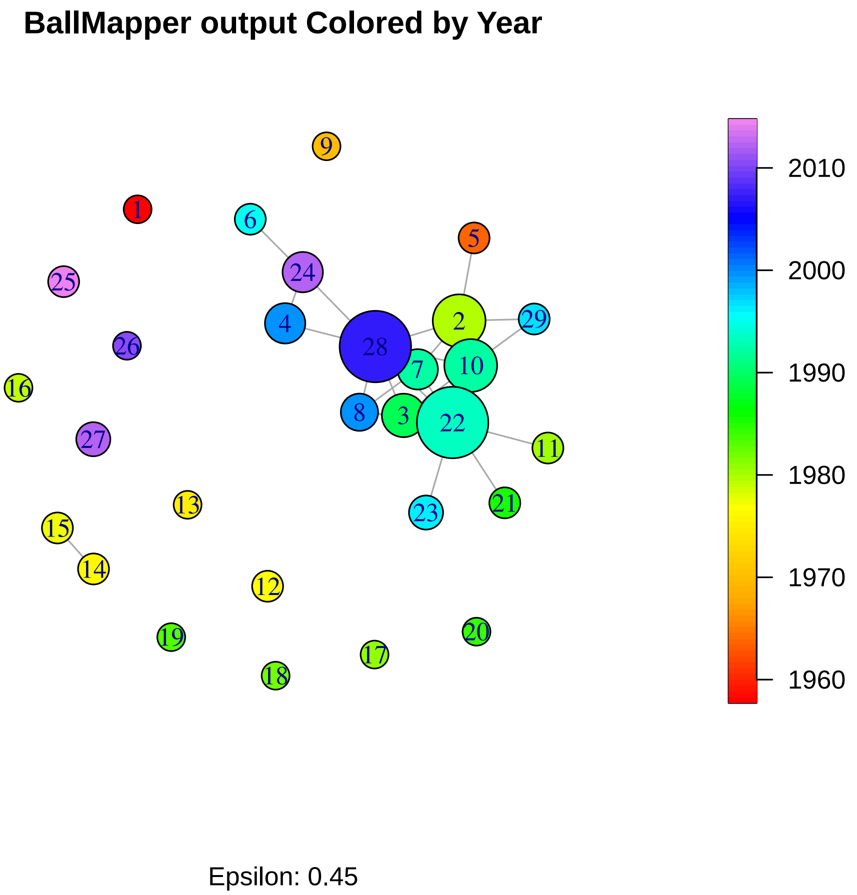

Figure 3: Data is normalized on a [0,1] scale due to all varialbe not being normally distributed See Supplement_2.pdf.

References

Dlotko, Pawel, Wanling Qiu, and Simon Rudkin. 2022. "Topological Data Analysis Ball Mapper for Finance." https://arxiv.org/abs/2206.03622.

Dłotko, Paweł. 2019. "Ball Mapper: A Shape Summary for Topological Data Analysis." https://arxiv.org/abs/1901.07410.

Dłotko, Paweł, Wanling Qiu, and Simon Rudkin. 2019. "Financial Ratios and Stock Returns Reappraised Through a Topological Data Analysis Lens." https://arxiv.org/abs/1911.10297.

Dłotko, Paweł, Simon Rudkin, and Wanling Qiu. 2019. "Topologically Mapping the Macroeconomy." https://arxiv.org/abs/1911.10476.

Dłotko, P., W. Qiu, and S. T. Rudkin. 2021. "Financial Ratios and Stock Returns Reappraised Through a Topological Data Analysis Lens." *The European Journal of Finance* 30 (1): 53–77. https://doi.org/10.1080/1351847x.2021.2009892.

Rudkin, Simon, Lucy Barros, Paweł Dłotko, and Wanling Qiu. 2023. "An Economic Topology of the Brexit Vote." *Regional Studies* 58 (3): 601–18. https://doi.org/10.1080/00343404.2023.2204123.

Data Sources

- Bureau of Economic Analysis
  - Real Exports and Imports
  - New Foreign Direct Investment in the United States
  - Real GDP
  - Personal Income and Its Disposition
- Federal Reserve Bank
  - Selected Interests Rates

- Bureau of Labor Statistics
  - CPI
  - Employment Cost Index
  - PPI
  - Housing Starts
  - Unemployment Rate