CPI Data Analytics Using PySpark, Pandas, and Matplotlib

John Martin
Georgia State University
Atlanta, Georgia 30303
jmartin103@student.gsu.edu

***Abstract*—The economy plays an essential role in our everyday lives. It has experienced many ups and downs throughout the history of time in its journey to get to where it is today. It is responsible for the salaries that we make at our jobs. When the economy is in a contraction, then there is less money in the economy to pay employees for their work. On the other hand, when it is in an expansionary phase, there is substantial money in the economy. One essential characteristic in economics is the consumer price index (CPI), which measures the increase and decrease in the price measure of numerous expenditures, such as food, transportation, and housing. The purpose of this paper is to develop a plot to analyze average, minimum, and maximum consumer price index (CPI) data in the United States throughout the past five years.**

Keywords—United States economy, Spark, big data

1 Introduction

The United States economy is the core principle of trade, marketing, and employment. There are many different types of economies in our nation. For example, the market economy determines the prices of goods and services being traded, and it is where decisions involving investment, production, and distribution are made. These factors are dependent on the concept of supply and demand. One example of these factors is the consumer price index (CPI). The CPI examines the weighted average of the prices of various goods and services, such as transportation, food, and gasoline. It is determined by averaging the price changes for each item. The CPI is among the most frequently used data to analyze periods of inflation or deflation in the economy.

In this paper, a new method for analyzing the change in the average, minimum, and maximum consumer price index of the United States throughout the past few years, using big data, will be introduced. This method will use PySpark, Pandas, NumPy, and Matplotlib to analyze the change in the consumer price index (CPI) over time. This data will be used for the CPI analysis, and this will be plotted using Matplotlib and Pandas. The data will use 2013-2017 as the base year for determining the average CPI, and it will implement the use of the Bureau of Labor Statistics to determine the change in the CPI. The BLS has plenty of tables that contain the CPI of many different expenditures, and these tables will be read and analyzed.

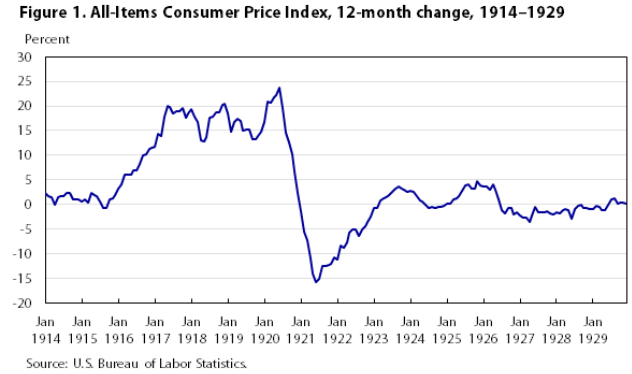

Figure 1: Example of CPI Plot [1]

2 Motivation and Related Work

There are numerous websites that give CPI statistics for the United States over the years. Such statistics include the percent change in CPI over a certain period of time, the CPI for many different types of expenditures, such as food, energy, transportation, and medical care, and the CPI can be used to predict inflation and deflation in the economy. The annual percent change and the average yearly CPI are the tools most critically used to predict and determine inflation and deflation.

The United States Bureau of Labor Statistics [1] gives a broad overview of the CPI percent change between 1914-1929. This table implements the use of the CPI of all expenditures. Among the expenditures analyzed in this graph were food, rent, apparel, fuel and electricity, and house furnishings. From the graph, we can see that there was a sharp acceleration in inflation from 1916 to 1918, and then again from 1919 to 1920. However, in the early 1920s, the economy suffered a significant deflation, as shown by the steep drop in the CPI between 1920 and 1921. The data suggests that during the World War I era, there was a steady amount of inflation, and then in the war's aftermath, the economy suffered a recession. By the time the recession ended in 1922, prices started to gradually increase, and remain more stable throughout the remainder of the 1920s [1].

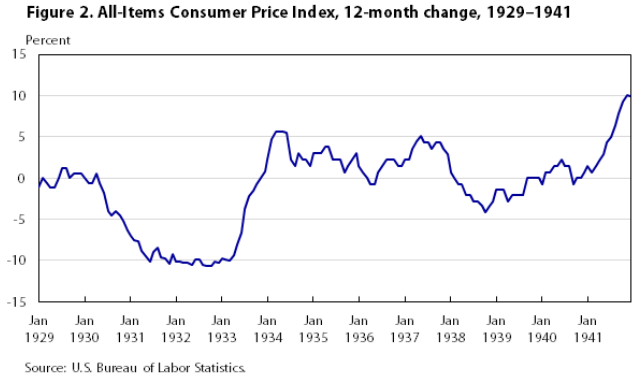

Figure 2: CPI Change During the Great Depression (1929-1941) [1]

After the price stability during the mid-to-late-1920s, there was once again growing concern with price decreases with the onset of what would become the Great Depression. During this time period, prices were consistently falling, and the unemployment rate was at an all-time high. Princes began decreasing during the early 1930s, and resulted in a very deep economic deflation, reaching an all-time low by 1933. Afterwards, prices began to dramatically rise again, due to the National Industrial Recovery Act, as well as the inventions of newer technology. However, despite the significant inflation during this time, unemployment was still very high. By the late 1930s, prices began to drop again, though this drop was not as significant as the beginning of the depression. Prices would still remain generally stable through 1940 and the start of World War II in 1941, when the Depression ended, and prices began to significantly rise again [1].

3 Method/Description of Frameworks/Libraries

In this project, I used the PySpark framework, along with a few of its libraries to graph the CPI data. Following are a few of the frameworks/libraries that I used for this project.

- PySpark is an Apache Spark Application Programming Interface (API) written in the Python programming language. This framework has been shown to process big data at significantly faster speeds than other frameworks, such as Hadoop and MapReduce. PySpark relies on the use of Resilient Distributed Datasets (RDDs) to operate data in parallel. These RDDs can reference any type of dataset in an exterior storage system, such as Hadoop and MapReduce. Furthermore, RDDs are considered fault-tolerant, and can be persistent in caching data in memory across nodes, which allows for more efficient data processing. PySpark also allows for the use of DataFrames to store and process data. DataFrames are a distributed collection of rows under columns. Essentially, it is the equivalent of a table within a relational database, such as Structured Query Language (SQL). Like RDDs, DataFrames are distributed across different clusters, and are designed for processing large data at a very efficient rate. In DataFrames and RDDs, transformation of data is accomplished by using lambda functions.
- Pandas is a Python software library used for data manipulation and analysis. Specifically, it offers the use of data structures and tables to manipulate data and time series. Like PySpark, Pandas offers the use of a DataFrame to process and manipulate data. Furthermore, it allows users to insert and delete columns from a DataFrame, and to read and write data between data structures and various file formats. It is highly optimized for performance [2].
- Matplotlib is a Python software library that allows developers to plot data from a Pandas DataFrame. It implements Python's numerical mathematics extension NumPy for plotting data. Matplotlib provides an interface similar to MATLAB, with the ability to plot data in a bar graph, line graph, histogram, scatter plot, or 3-D plot. It comes with the ability to name an x-axis and a y-axis, and is a very efficient tool when it comes to graphing and visualizing data [3].

4 Problem Statement

4.1 Problem Description

This project classifies the change in the consumer price index (CPI) over the course of 2013-2017. The data will be extracted from the Bureau of Labor Statistics [4], which provides data in the form of tables showing the CPI for each month and year. The purpose of this project is to use the CPI to determine whether the economy is in a state of inflation, or a state of deflation. If the CPI is increasing, then the economy is showing signs of inflation; otherwise, if the CPI decreases over time, then the economy is in deflation. The CPI can also remain steady over a certain period, showing a stable economy, and no signs of either inflation or deflation.

4.2 Data Requirement

The data must be from a reliable source, such as the Bureau of Labor Statistics. In this case, some data was extracted using the BLS data tables [4] provided on their website. These tables can be used to extract and query the CPI, and this can be used to identify whether the economy is in good shape or bad shape. Furthermore, transformation of data must be completed, if necessary, before being able to be queried. This transformation can be done using PySpark.

4.3 Technical Requirements

The tools must be easy to use and maintain, and should be able to use with any hardware and any operating system, and must work with both Windows and Mac. It must have excellent scalability, which is to say that it must be able to handle a growing amount of data, and must be able to process significant amounts of big data. PySpark can accomplish these tasks.

5 Data

5.1 Data Definition

The project aims to calculate the average consumer price index (CPI) over the past five years. Variables that are used for this are the CPI of all items, such as food, transportation, energy, and medical care. The BLS [4] has Excel tables that store all of this data for each month. The CPIs from each month of the year are then used to calculate the average CPI for that year. Furthermore, the minimum and maximum CPIs of each year are determined. Then, the averages of each year are combined into one DataFrame to visualize the increase or decrease of the CPI over time.

5.2 Data Visualization

A plotted line graph is used to visualize the change in CPI over time.

6 Extraction/Procedure

The process of data extraction involves analyzing and breaking down the data into smaller pieces of data to retrieve certain pieces of information. The BLS' data tables from each month of each year are combined into one Comma Separated Values (CSV) file, and the CPIs of each month are extracted; these are circled in red (Figure 3). These CPIs are stored into a Resilient Distributed Dataset (RDD), and then the RDDs are converted into DataFrames for the calculation of the average, minimum, and maximum CPI of each year from 2013-2017. Then, all of the Spark DataFrames are converted to Pandas DataFrames for plotting the data into a line graph, using Matplotlib.

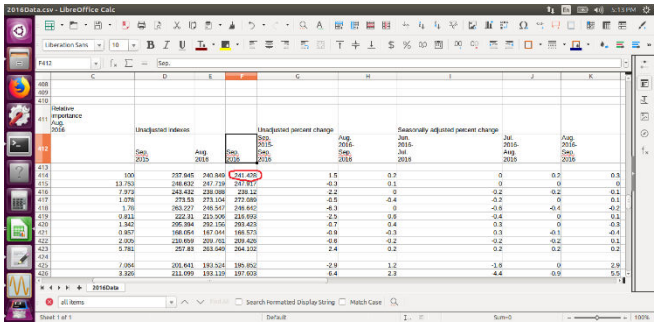

Figure 3: The CPI for all items (circled in red) in each month is extracted and stored into an RDD, which is then converted to a DataFrame.

7 Results

7.1 CPI Data

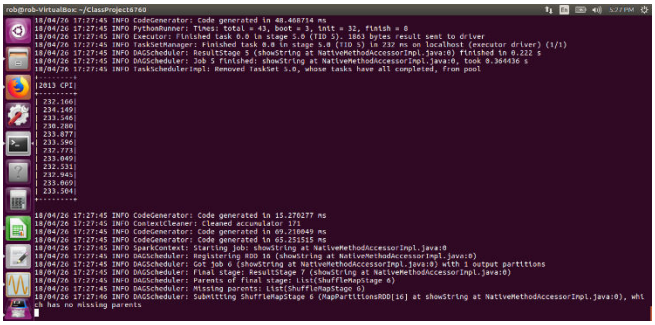

Figure 4a: 2013 CPI data from Excel spreadsheet

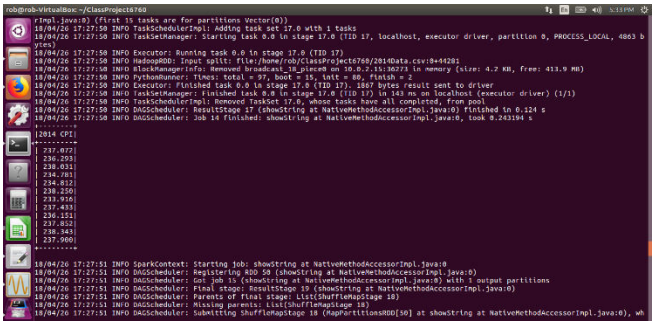

Figure 4b: 2014 CPI data from Excel spreadsheet

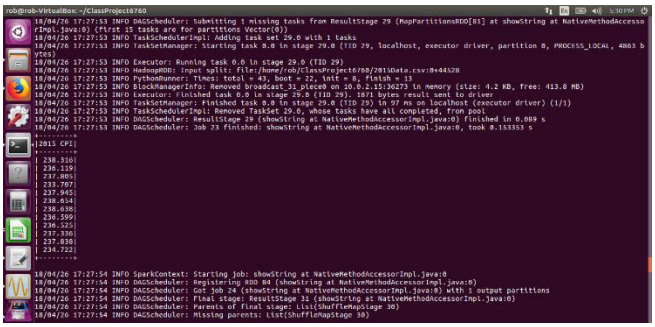

Figure 4c: 2015 CPI data from Excel spreadsheet

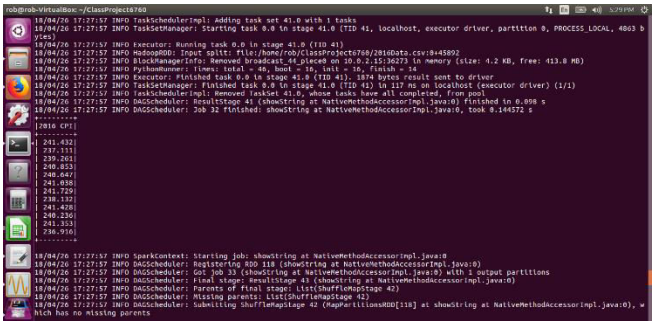

Figure 4d: 2016 CPI data from Excel spreadsheet

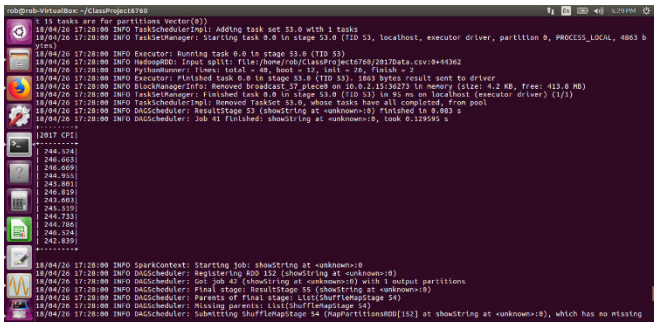

Figure 4e: 2017 CPI data from Excel spreadsheet

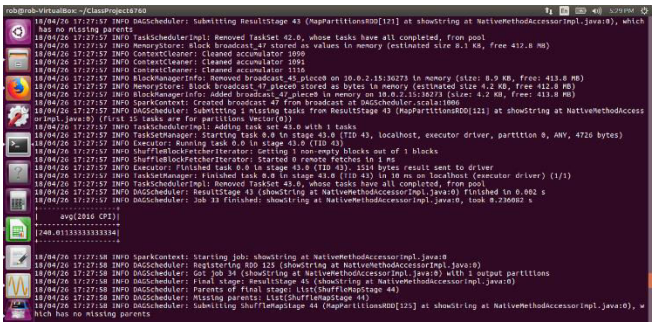

Figure 5d: 2016 average CPI calculated from data

7.2 Average CPIs from Each Year

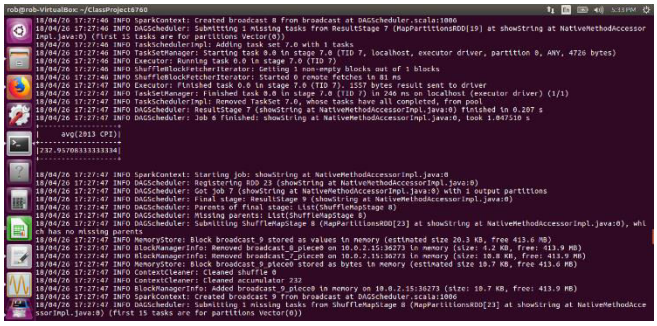

Figure 5a: 2013 average CPI calculated from data

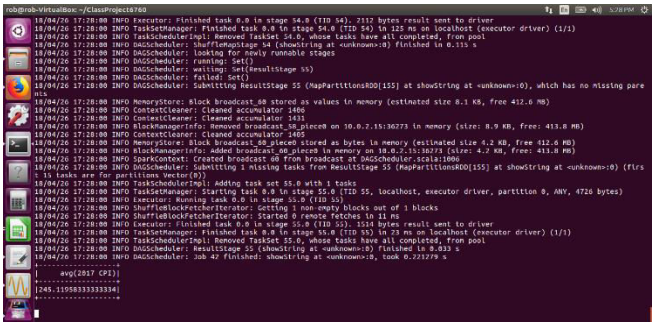

Figure 5e: 2017 average CPI calculated from data

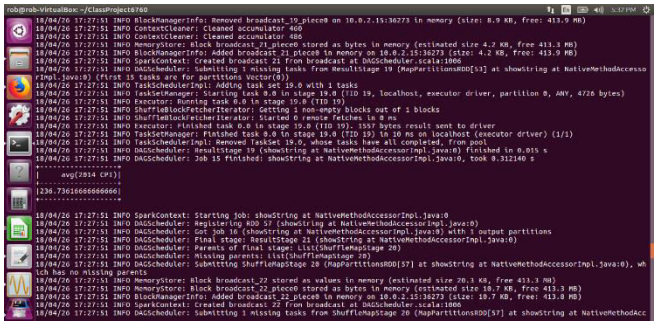

Figure 5b: 2014 average CPI calculated from data

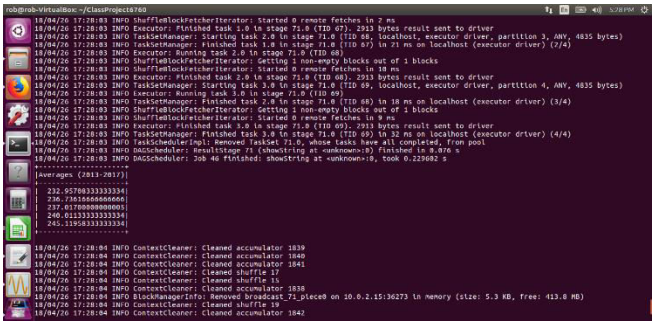

Figure 5f: All averages combined into one DataFrame

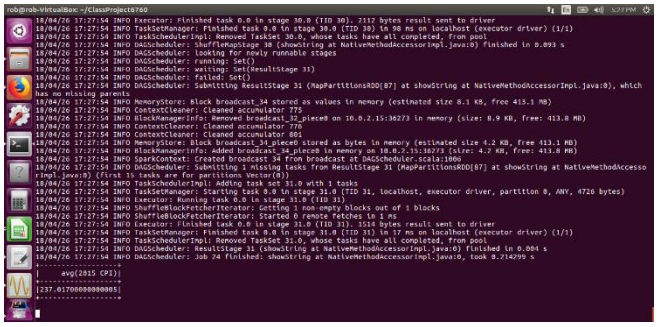

Figure 5c: 2015 average CPI calculated from data

7.3 Graph Result from Matplotlib

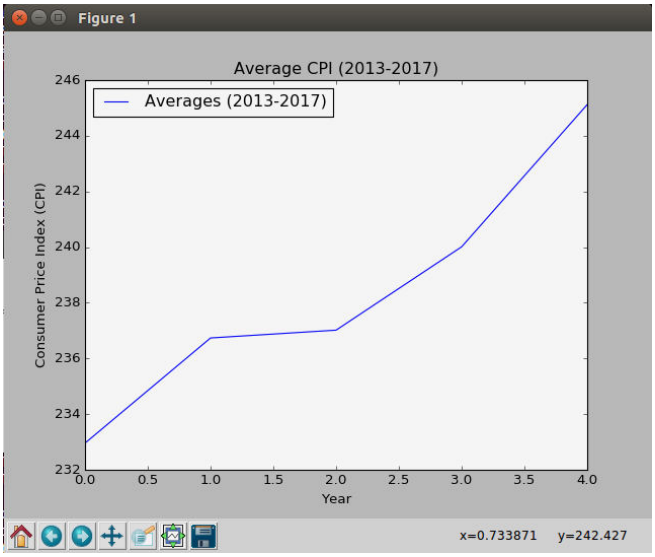

Figure 6: Line graph obtained from average CPI data

7.4 Explanation of Results

From the results gathered in Figures 4a-e, these are the CPIs of each month (albeit not in complete order by month). In PySpark, the aggregate average function is applied to each dataset to obtain the average CPI of each year (Figures 5a-f). Figure 6 displays the plotted line graph of each average CPI, accomplished by Matplotlib. From this graph, we can infer that the economy has been in inflation over the past five years because the CPI has increased over this time period, with the most significant increase being between 2016 and 2017.

8 Conclusion/Future Work

From this project, we can conclude that the economy is in inflation and in great shape because the CPI has steadily increased over the past five years, and is continuing to increase with time. This project also proves that PySpark is very efficient and fast with breaking down and processing big data, even more so than other modern big data frameworks. Its use of an RDD and a DataFrame help to visualize the data into a table. Moreover, the Pandas and Matplotlib libraries are excellent choices in helping to visualize data in a graph. They are extremely effective in both time and space when it comes to manipulating and processing data.

In the future, I plan to find a way to plot years on the x-axis of the resulting bar graph. Furthermore, I plan to discover a way to expand on this project in order to implement a dashboard to stream real-live CPI data using the BLS API. I believe this will be very beneficial for economists to analyze whether or not the economy is in good shape.

9 Acknowledgements

I would like to thank Dr. Yubao Wu for his help and support throughout this project. It was truly a good opportunity to learn with hands-on experience in handling big data, which was highly beneficial.

10 References

[1] Bureau of Labor Statistics. 2014. One Hundred Years of Price Change: The Consumer Price Index and the American Inflation Experience. (April 2014). Retrieved April 25, 2018 from https://www.bls.gov/opub/mlr/2014/article/one-hundred-years-of-price-change-the-consumer-price-index-and-the-american-inflation-experience.htm

[2] Pandas. Python Data Analysis Library: Pandas. Retrieved April 25, 2018 from https://pandas.pydata.org/

[3] Matplotlib. Matplotlib Documentation and Installation. Retrieved April 26, 2018 from https://matplotlib.org/index.html

[4] Bureau of Labor Statistics. Archived Consumer Price Index Supplemental Files. Retrieved April 25, 2018 from https://www.bls.gov/cpi/tables/supplemental-files/home.htm