Joe Alcini

STA 402

Dr. Wright

8 December 2021

Term Project Preliminary Review

Introduction

The consumer price index is a popular metric used by economists to judge the economy of the United States of America. The index is produced by the Bureau of Labor Statistics, and it measures the change in the price of consumer goods. The primary usage for this index is to judge if the rate of inflation is changing overtime. This metric has become more of a focal point of economists after the rise in price of certain consumer goods in recent months. Economists can use the historical trends that are present in this dataset to make estimates if the recent rise in price of consumer goods is expected or out of the ordinary. They can also determine if certain geographical regions in the United States are having different types of price changes. This is important because the economists can look at different factors than the national economy and national political policy to determine causes for these differences and better understand the economic health of the United States.

Given this information some questions economists may be interested knowing are:

Are their significant differences in the Consumer Price Index between regions?

Is the consumer price index growing at a significantly faster rate in the 21st Century compared to the 20th century?

Description of Data

The structure of the dataset provided by the Bureau of Labor Statistics has many structural challenges to overcome when trying to create a useful projection. The first is the use of multiple different files that relate to the desired data. The dataset uses codes for some types of categories of categories that are buried in other files and need to be manually parsed to understand their meaning. This requires a great deal of time to read through the data sets to find the meaningful relations and columns that need to be used to begin extracting the useful data and assigning meaning to it. A diagram has been included to help visually represent the relationships that exist in this data.

The first dataset that should be examined is the cu.area file. This dataset contains the various regions that have a corresponding Consumer Price Index. The region names appear in a column called area_name and the ones that are concern to us are: Northeast, Midwest, South, and West. The area_code column in this dataset acts as an id field that is referenced in other files.

Therefore we can now examine the cu.series data. This contains the identifiers of the types of Consumer Price Metrics that exist. This identifier falls under the series_id column and is used in other series. The area_code that was referenced in the cu.area file also appears in this file as well. The other useful coulmns are the seasonal column which tells if an adjustment is made based on the time of year, we are interested in unadjusted data, the periodicity_code which represents how often the measure is taken, we are interested in regularly measured data, and finally the series tile which describes the series measure.

Finally the cu.data.1.AllItems file contains the values for the Consumer Price Index of select series. It contains a series_id column which is identical to the cu.series series id, as well as a year, a period which represent whether it is for a specific month or year, and a value representing the value of the Consumer Price index at that time.

Due to the nature of using three different files merging of data sets on common fields must take place to ensure that the target information can be extracted. The first merge that occurs is between the area and series data because they share the same field, area_code, and both are relatively small after filtering them. The next step is to merge the result of these data sets into the allItems data set using the series_id. The data that should be kept is the area_name, value, and a combination of the period and year to create a date variable for graphing.

Strategy

The first thing to complete when filtering data is to determine what regions the user wants to view. This can be achieved by passing them as a space separated list into the macro and then storing them in a dataset called input_regions and storing the number of regions that were input. Due to an input that may not be in order the data must be sorted. After this is complete it can be merged with the area data set which is being input and cleaned to omit all entries that are not a part of the four regions specified. Then the data is merged between the area and user input_region data and only retained if they exist in the user's inputs. The fields that are retained are the area_code and area_name.

Then the series data is input from the file. The data filtered to only included series that are not adjusted by the season and are regularly measured using the seasonal and periodicity _code columns. The result is a short list of series of which some provide no use to us. The columns that are kept are the series_id and the area_code. The data can now be merged with the area data using the area_code. If the area code appears in the area data set, then the entry is kept. Once this is complete the only needed columns are the series id and the area_name.

Finally, the allItems data set is created from its input file. The first entries that are filtered are those that fall outside of the year bounds set by the user in the macro heading. Then the data is filtered by the selected period calculation by checking what period type was specified by the user and deleting entries that do not fit such as eliminating the annual average if the Monthly is selected. The quarterly data is calculated by taking the monthly value at the end of every fiscal quarter. Now that many items from this large file have been filtered out it can now be merged with the series data that has been filtered down by the series_id and only kept if it is present in the series data set. Once this is complete a new field is created called date. Date combines the period month and year and is set to the first day of the month to calculate time series. Yearly data is set to the first of the year. Once the dates are created the final columns that are kept are the area_name, date, and value.

Now the data can be graphed using sgpanel and taking in the filtered items data as its input. The graphs are paneled by region and have the x axis represented by the date variable and labeled by the interval that the user supplied, and the y axis representing the value of the consumer price index. Each panel is labeled by region and a title on top of the graph specifies the period type that was used as well as a description of the chart.

Results

The graphs output by the macro will contain the area_code to create region-based panels, the values of the cpi and dates will make up the values on the y and x axes respectively. The series lines will appear as a uniform color and the labels on the graph will be “Value of the CPI” on the y axis and “*Period* (Days from 01-01-1960)”. The title for the entire graph is "*Period* Consumer Price Index Over Time by Region". The graphs appear in one line and is set by the number of regions initially input by the user.

Discussion

When examining the graphs that are produced by the macro it does not appear that there are any significant long-term differences between the consumer price index in different regions within the United States. If these claims were to be proven a statistical test would have to be preformed but on appearance alone the visual differences in the graph would indicate any difference that does exist would be too small to serve as a significant difference. However, some trends do appear such as the Midwest and West regions and the South and Northeast regions having slightly similar shaped compared to one another. Additionally, it does not appear that there are any significant differences in the change of the consumer price index over time. All four of the regions have roughly the same rate of increase and the same value of the Consumer Price Index of every point. This results in there being no distinguishable visual difference in the Consumer Price Index changing rapidly over time and would also likely be backed up by a statistical test if it were preformed over the same data.

Overall, when examining the results some additional assumptions can be made about the United State’s economy. It appears that when something affects the American economy all regions feel similar effects to one another. This would imply that the American economy is the sum of all its parts and that if a particular region is starting to experience more inflation others will as well.

Code:

```
/*
	Name: Joe Alcini
	Course: STA 402
	Date: 12/08/2021
	File: CPI.sas

	Description: This program creates a
	macro used to display graphs of the consumer
	price index for regions specified by a user
	for a given time range and calculation interval.
	The user can input 1-4 regions that consist of:
	Midwest, Northeast, Sout and West using the regions parameter.
	The user can enter a starting and ending year to
	limit the number of data points that make up the graph by
	using the start_time and end_time parameters.
	Finally the user can choose how what calculation interval
	is used by choosing between: Month, Quarter,or Year
	using the calc_period parameter. User can specify path
	to data using the path variable before macro heading.
	Refrain from using qutation marks in the path declaration.

	Example Inputs:
	%cpi(regions="South Northeast Midwest West", start_time=1970, end_time=2021, calc_period=Year);
	%cpi(regions="Midwest South Northeast West", start_time=1975, end_time=1995, calc_period=Month);
	%cpi(regions="Midwest West", start_time=2010, end_time=2021, calc_period=Quarter);
*/


/* Set path to the folder conatining the datasets */
```

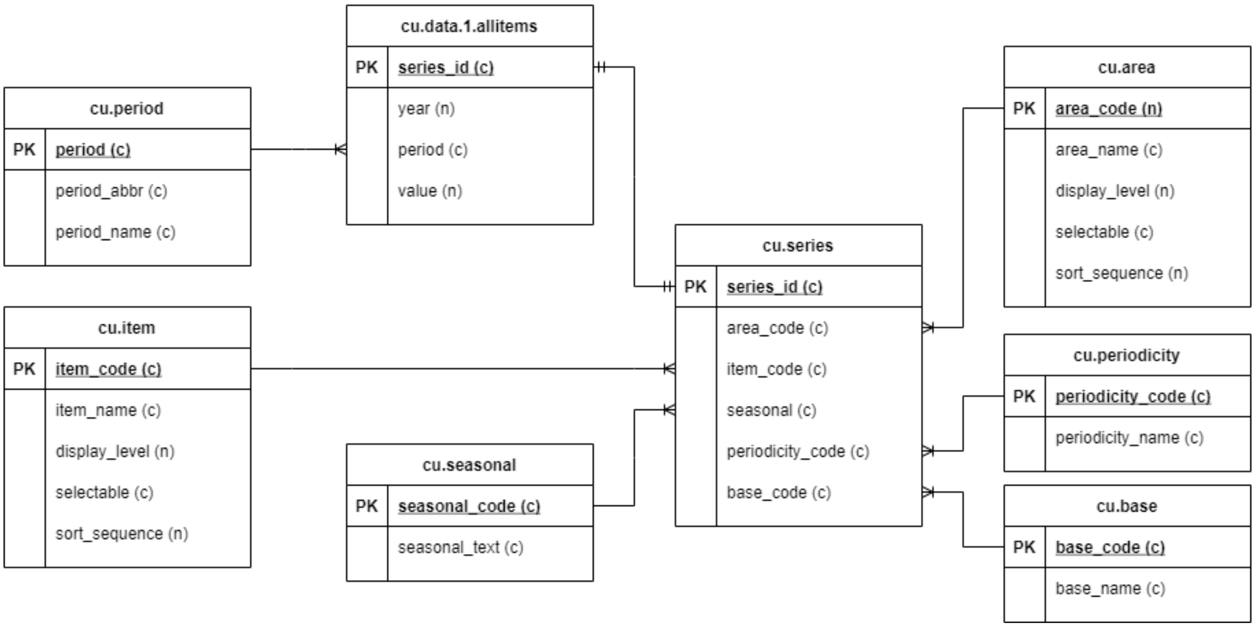
```
%let path = ;

/* Declares the macro */
%macro cpi(regions=, start_time=, end_time=, calc_period=);
      /* Take user given locations */
      data input_regions;
            /* Specify max length for a region */
            length area_name $ 9;

            /* Count number of regions */
            num = 0;

            /* Searches for names */
            do until (area_name=" ");
                  /* Searches for the region */
                  area_name = scan(&regions, num + 1);

                  /* Space deliminated finds the space*/
                  if area_name~=' ' then do;
                        num = num + 1;* adds 1 to the counter;
                        output;* outputs the region into the dataset;
                  end;
            end;

            /* Creates a macro variable storing the number of regions
*/
            call symput("num_regions", num);
      run;



      /* Sorts the user reigons for merging */
      proc sort data=input_regions;
```

```
        by area_name;
    run;

    /*  area data:
        pull out area code and name
        filter out user regions
    */
    data cu_area;
        /* Takes input from the given file */
        infile "&path\cu.area"
            firstobs=2
            obs = 14
            expandtabs;

        /* Specifies lengths of taarget variables */
        length area_code $ 4;
        length area_name $ 49;

        /* Inputs data */
        input area_code $ area_name & display_level selectable $ @;

        /* Filters out target regions */
        if area_name not = "Northeast"
            and area_name not = "Midwest 0"
            and area_name not = "South"
            and area_name not = "West" then delete;

        /* Cleans midwest data due to layout of the file */
        if area_name = "Midwest 0" then area_name =
substr(area_name, 1, 7);

        /* Merges with user specified regions*/
        merge input_regions(in=in_regions);
            by area_name;

        /* Retains entry if specified*/
        if in_regions;

        /* Keeps the following columns */
        keep area_code area_name;
    run;

    /*  series data:
        pull out series id and merge with regions area code
        filter by selected area codes in input
    */
    data cu_series;
        /* Takes input from the given file */
        infile "&path\cu.series"
            firstobs=2
            expandtabs;

        /* Specifies lengths of taarget variables */
        length series_id $ 15;
        length series_title $ 74;
        length area_code $ 4;

        /* Inputs data */
        input series_id $ area_code $ item_code $ seasonal $
periodicity_code $ base_code $ base_period $ series_title & @;

        /* Filters by target parameters */
        if seasonal = 'S' then delete;
        else if periodicity_code not = 'R' then delete;
```

```
            else if not find(series_title, 'All items') or not
find(series_title, 'not seasonally adjusted') then delete;

            /* Keeps selected columns */
            keep series_id area_code;
      run;

      /* Merges the areas and series*/
      data cu_ids;
            /* Merges with desired regions */
            merge cu_area(in=in_area) cu_series;
                  by area_code;

            /* Retains if present in user regions */
            if in_area;

            /* Keeps the following columns */
            keep series_id area_name;
      run;

      /*    allItems data:
            keep the series, period val (M--), and monatary value
            retain periods that match the period type specified
      */
      data cu_allItems;
            /* Takes input from the given file */
            infile "&path\cu.data.1.AllItems"
                  firstobs=2
                  expandtabs;

            /* Specifies lengths of taarget variables */
            length series_id $ 15;
```

```
        input series_id $ year $ period $ value @;

        /* Removes entries outside of the time bounds */
        if &start_time > year or &end_time < year then delete;

        /* Removes periods based on the period calulation */
        if "&calc_period" = "Year" and period = 'M13' then output;
       if "&calc_period" = "Month" and period not = 'M13' then
output;
        if "&calc_period" = "Quarter" then do;
            if period = 'M03' then output;
            else if period = 'M06' then output;
            else if period = 'M09' then output;
            else if period = 'M12' then output;
            else delete;
        end;
    run;

    /* final data:
        merges with series
        creates dates using the combination of year and period
    */
    data cu_allItems_final;
        /* Merges series and item data */
        merge cu_ids(in=in_series) cu_allItems;
            by series_id;

        /* Must contain only items from the series data */
        if in_series;

        /* Creates valid dates */
        if period not = 'M13' then date = input(cats(substr(period, 2, 2), '-01-', year), mmddyy10.);
        else date = input(cats('01-01-', year), mmddyy10.);

        /* Keeps the following columns */
        keep area_name date value;
    run;

    /*Titles the plot*/
    title "&calc_period.ly Consumer Price Index Over Time by Region";

    /*
        Smoothed line sgplot series x=date y=CPI_Value
        format by using axis labels
        use a different plot in the same grouping for each plot
    */
    proc sgpanel data=cu_allItems_final;
        /* Create 1 row of side by side plots */
        panelby area_name / novarname columns=&num_regions;

        /* Creates the series for each region*/
        series x=date y=value;

        /* Labels the axes */
        colaxis label = "&calc_period.s (Days from 01-01-1960)";
        rowaxis label = "Value of the CPI";
    run;
%mend cpi;
```

BLS-Consumer Price Index Data Layout

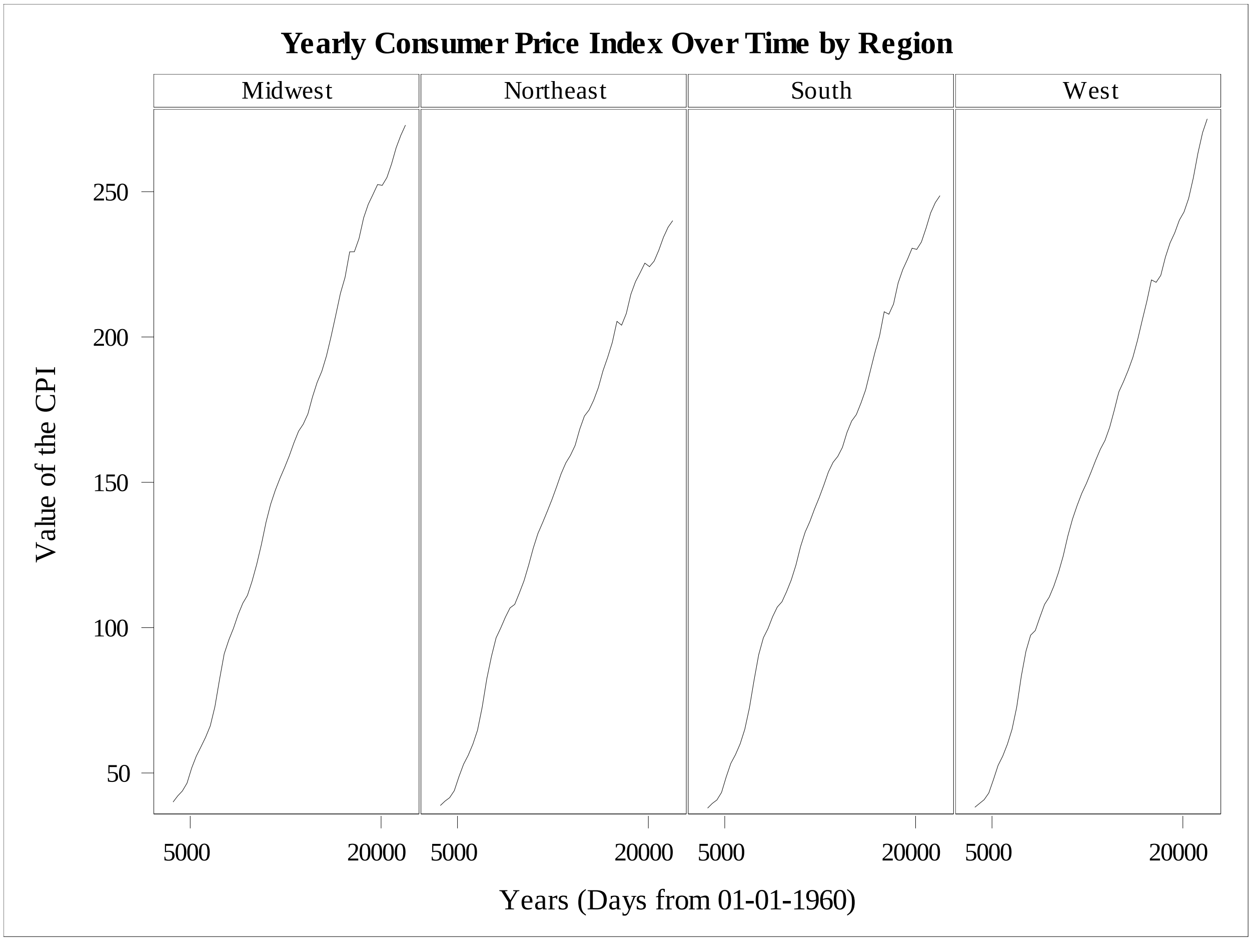


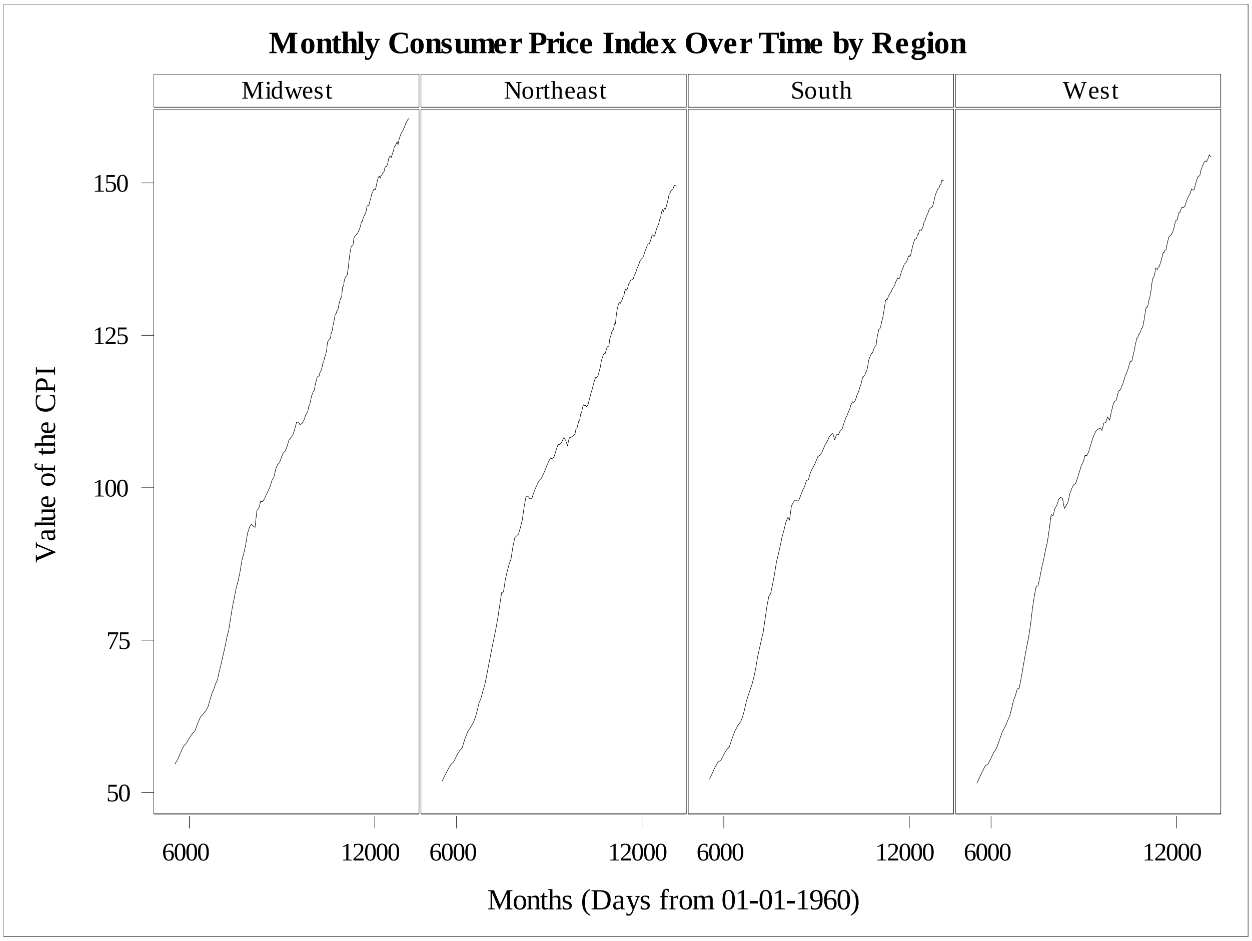
Area Data

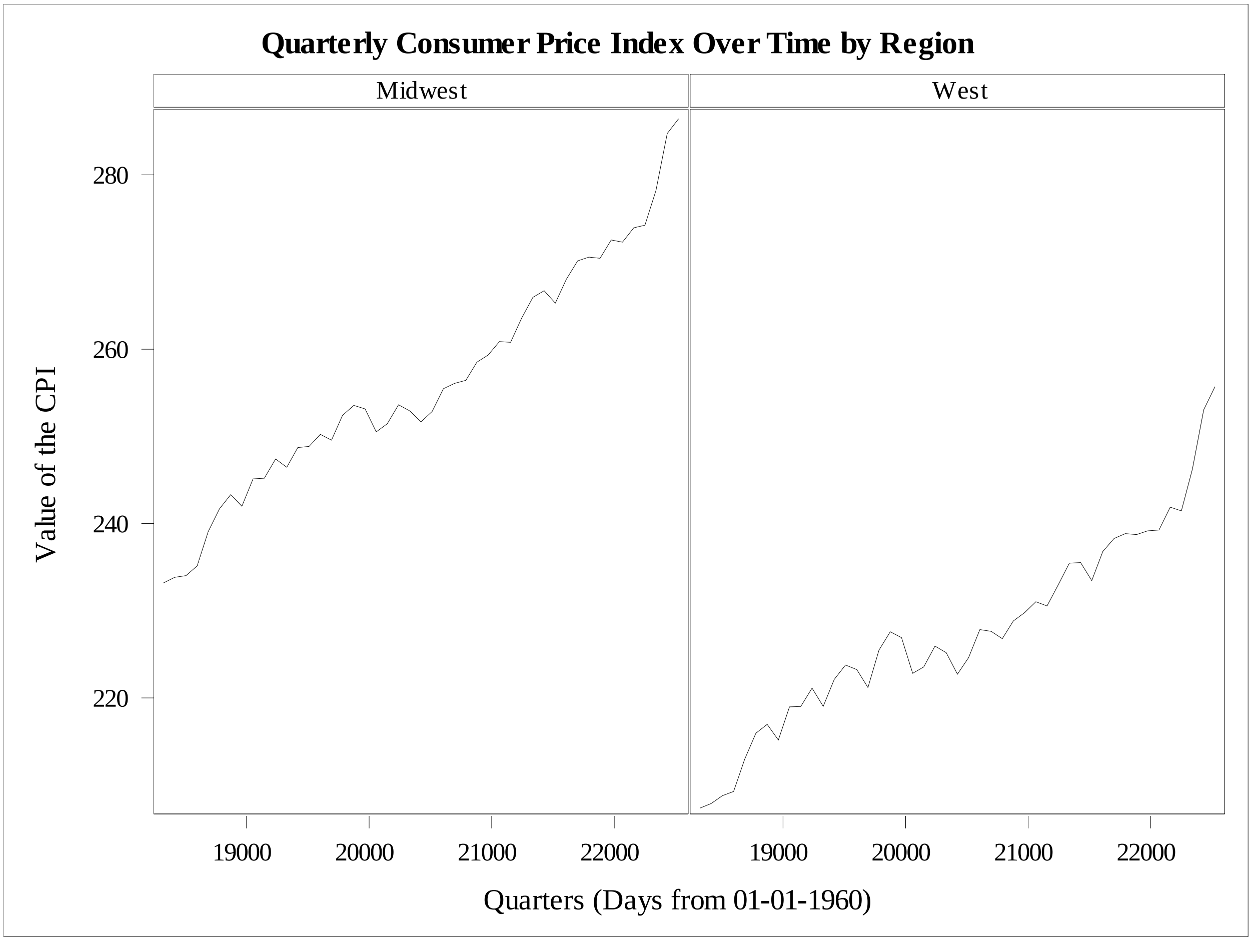
Obs	area_code	area_name
1	0100	Northeast
2	0200	Midwest
3	0300	South
4	0400	West

Series Data

Obs	series_id	area_code	item_code
1	CUUR0100SA0	0100	SA0
2	CUUR0200SA0	0200	SA0
3	CUUR0300SA0	0300	SA0
4	CUUR0400SA0	0400	SA0


Yearly Consumer Price Index Over Time by Region
Midwest
Northeast
South
West
Value of the CPI
250
200
150
100
50
5000
20000
5000
20000
5000
20000
5000
20000
Years (Days from 01-01-1960)


Monthly Consumer Price Index Over Time by Region
Midwest
Northeast
South
West
150
125
100
75
50
Value of the CPI
6000
12000
6000
12000
6000
12000
6000
12000
Months (Days from 01-01-1960)


Quarterly Consumer Price Index Over Time by Region
Midwest
West
280
260
240
220
Value of the CPI
19000
20000
21000
22000
19000
20000
21000
22000
Quarters (Days from 01-01-1960)