How Inflation Has Fluctuated Over the Last Few Decades

The question I set out to answer was “Is recent inflation in the U.S rising at an alarming rate?”. After the covid-19 pandemic, we’ve seen a rise in prices across all industries. We hear this is due to inflation and most believe inflation is rising at an unreasonable rate, some say it’s the worst we’ve ever seen. I’d like to find out if this is true. For this project, I found data on the U.S annual inflation rates covering any data from 1915, with the most current year being 2022. The dataset included all countries with different categories of inflation (food, energy, etc) and the total inflation rate for each year. I will be focusing on the United States total yearly inflation rate.

To complete this data analysis, I had to obtain the dataset in a file that can be uploaded into R. I searched for a reliable database website that has datasets available for download and found data.oecd.org. I downloaded this data through a CSV file, opened R studio to begin and chose the CRAN (server) option closest to my location. When using these commands for installing packages or performing other tasks, I must take into consideration the type of device/version of R I am using as the commands may differ. I have a MacBook, so I need to use commands that are compatible with the MacOS version of R. I needed to have the ability to install a package that will let me utilize the file, so I began with installing the Tidyverse package (giving access to other common packages and data analysis tasks) and then installing a package that lets me read a CSV file. *Normally you can also load tidyverse with the library () command, but I forgot to when beginning and ended up just loading individual packages as I needed them.

Install Tidyverse: > install.packages(“tidyverse”)

Install CSV reader: > install.packages("readr")

Load CSV reader: > library(readr)

Then, I needed to locate and open this file. I had to direct R through my home page, username, location, and file name.

Upload file: > inflation_data <- "/Users/olivmck/Downloads/inflation_dataset.csv"

I received an error message:

Warning message:

One or more parsing issues, call `problems()` on your data frame for details,

e.g.:

dat <- vroom(...)

problems(dat)

I used the problems () command to investigate the issues:

> issues <- problems(inflation_dataset)

> print(issues)

This gave me results showing any rows/columns that had discrepancies. I could then see that there were more columns than expected and an empty column to the far right with no values where R expected values to be present. I needed to inspect the column names to see why there are more columns than expected and delete the column without any values. First, looking at column names:

```
> column_names <- colnames(inflation_dataset)

> print(column_names)
```

There was a single title (a file name) in the very first row and I wanted to delete that so then the correct column names could be used. When I attempted to skip the first row to use the correct column names using a skip row 1 command, it did not work and gave an error message that the file doesn't exist in the working directory. So, I made sure the working directory was set up properly so that R is accessing the correct file throughout the project (from previous command for path to open file):

```
> setwd("/Users/olivmck/Downloads")
```

Then I could attempt again for R to skip the first row to use the correct column names:

```
> inflation_dataset <- read_csv("inflation_dataset.csv", skip = 1)
```

This then showed the correct column names and did not have any issues with the amount of columns but still gave an error message to use the command "problems ()" which I still assumed was due to the empty column. So then, I could use the correct column name to delete that column containing empty values. I like to use the dplyr package command select() (a part of tidyverse) to be able to select the column I want removed (first loading dplyr):

```
> library(dplyr)

> inflation_dataset <- inflation_dataset %>%

+ select(-'Flag Codes')
```

Then I looked at a summary of the data to ensure the titles were correct:

```
> summary(inflation_dataset)
```

and checked for any missing values in all columns and rows using the is.na() function which will say TRUE if there are any missing values or FALSE if not:

```
> any_missing <- any(is.na(inflation_dataset))

> print(any_missing)

[1] FALSE
```

So then I knew there weren't any missing values and the rows/columns were set. Next, I wanted to focus on only the total yearly inflation rates within the United States, so I needed to filter out the rest of the data. I used the dplyr package to complete a filter() command and made sure both the column names and values selected were correct to be able to select only the total inflation rates for the United States.

> usa_inflation_dataset <- inflation_dataset %>%

+ filter(LOCATION == "USA" & SUBJECT == “TOT”)

And then checked that it worked by viewing the summary:

> summary(usa_inflation_dataset)

The length of each column was over 2,000 which was more than I expected for just over 100 year's worth of data. So I needed to investigate why this was happening. First, I double checked the number of rows:

> nrow(usa_inflation_dataset)

It came back with the same number. So I decided to look for duplicates:

> any_duplicates <- any(duplicated(usa_inflation_dataset))

> print(any_duplicates)

It came back “FALSE”. I decided to check on the years only for any discrepensies in case there were multiple entries per year but in this case, it will be unique values because we want only one entry per year. I used the unique() function using the correct column title:

> unique_time <- unique(usa_inflation_dataset$TIME)

> print(unique_time)

This showed me all the unique year values in that column, and it seems there were multiple entries for the same years but containing characters or additional digits for each quarter or a month of that year. I also could see that the earliest data for the USA is 1955 so I know I only need 68 values to cover 1955-2022. I wanted to only focus on the total yearly inflation, so I decided to filter out the extra entries for each year. Since the extra entries contained some type of character or symbol, I was able to filter out any years that were not strictly digits. I used the filter() function along with the grepl("^\\d+$",) function for selecting only digits.

> usa_inflation_dataset <- usa_inflation_dataset %>%

+ filter(grepl("^\\d+$", TIME))

Then check the summary to see if the length of columns is accurate:

> summary (usa_inflation_dataset)

It still showed the length to be at 135 which is too long for the 68 years I’m covering. So I checked again for duplicates but specifically for the years in case there were different types of categories I did not filter out previously.

> duplicates_time <- duplicated(usa_inflation_dataset$TIME)

> print(duplicates_time)

There were about half on the entries that came back TRUE. So I decided to just view the whole dataset now since it's a smaller number that I can glance at the find the issue. I used the print() function and n=all rows:

> print(usa_inflation_dataset, n = 135)

Here I could see the measure column had a second type of entry that duplicated all of the years. It was some type of index entry that I did not need for the actual yearly total. So I deleted all of these rows, similar to filtering out what I did want to include but instead filtering out what I did not want.

> usa_inflation_dataset <- usa_inflation_dataset %>%

+ filter(MEASURE != "IDX2015")

Then checked to ensure that the data was accurate now with the length of the columns.

> summary(usa_inflation_dataset)

It showed now that there were 67 entries, I viewed all of the rows again to see the beginning and ending year

print(usa_inflation_dataset, n = 67)

And saw it began at 1956 so there was only data from the other column for 1955. This works for my project now with the total inflation rates from 1956-2022. Now that the dataset has been cleaned up, I wanted to create a visual to begin with. A chart would let me see all of the inflation rates from beginning to end. I loaded the ggplot package to create a time series plot. Then I followed the ggplot command to create a graph with the correct values on the x and y axis.

> library(ggplot2)

> ggplot(usa_inflation_dataset, aes(x = TIME, y = Value)) + geom_line() + labs(title = "Inflation Rate in the U.S 1915-2022", x = "TIME", y = "Value")

This gave me an error message:

`geom_line()`: Each group consists of only one observation.

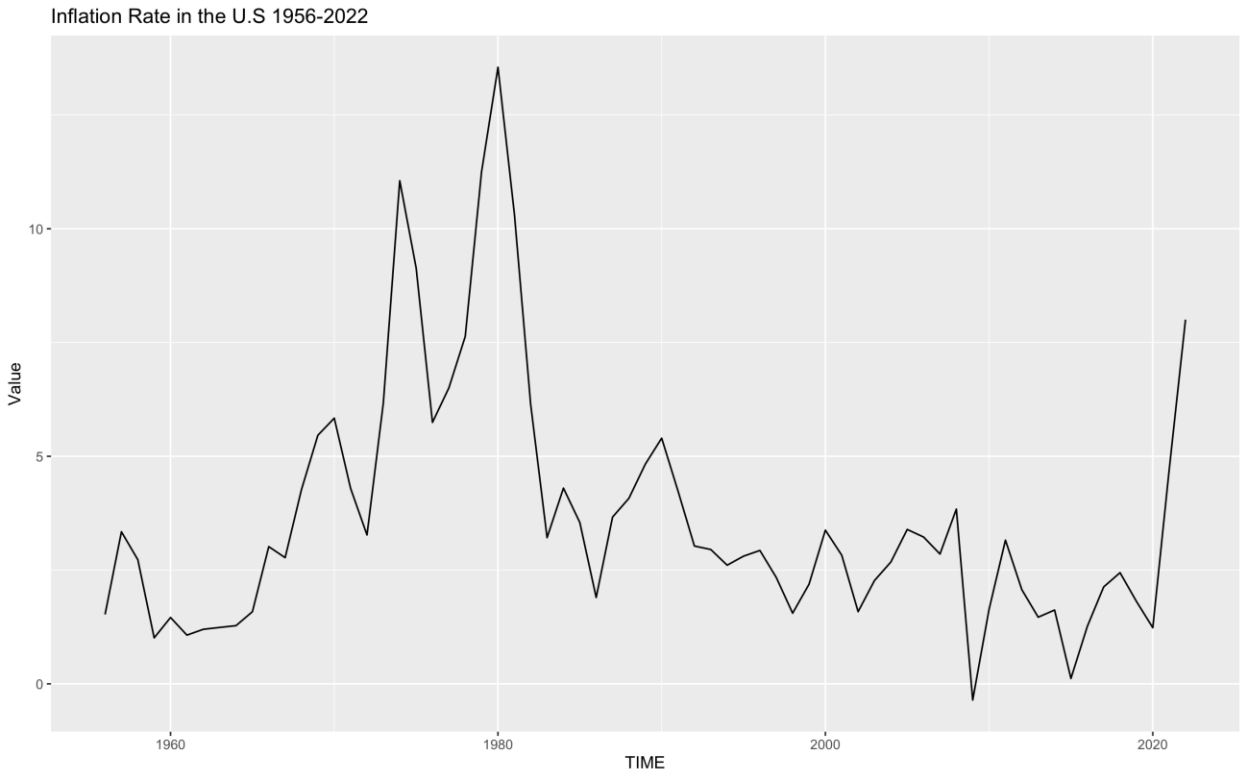
i Do you need to adjust the group aesthetic?

There also wasn't any year labels on the x axis which I realized was due to each column being recognized as characters. So I need to adjust the command to allow for the grouping of this single category but first have the year column recognized as numeric values:

> usa_inflation_dataset$TIME <- as.numeric(usa_inflation_dataset$TIME)

Then I can plot the line graph again with the single grouping:

> ggplot(usa_inflation_dataset, aes(x = TIME, y = Value, group = 1)) + geom_line() + labs(title = "Inflation Rate in the U.S 1956-2022", x = "TIME", y = "Value")



This gave me a line graph showing the changes in inflation from 1956-2022. I could see the changes over time and the very low or high points of inflation which can be attributed to recessions and financial crises. But is our recent increase in inflation consistent with the fluctuations we've seen in the past decade? First off, is our 2022 inflation rate considered an outlier? The summary() function already shows us our basic calculations of the inflation rates: minimum (-0.35), maximum (13.5492), median (2.9517), mean (3.6685), first (1.7261) and third (4.2967) quantile. The 2022 inflation rate was 8%. We know this is above average but not the highest we've seen. How unreasonable is it compared to the recent years? I calculated the z-score to determine if 2022 is an outlier using scale() and abs() functions:

```
> z_scores <- scale(usa_inflation_dataset$Value)

> outliers <- which(abs(z_scores) > 3)

> outlier_years <- usa_inflation_dataset$TIME[outliers]

> print(outlier_years)
```

This came back with only one year, 1980, which had the highest inflation rate out of all years (as seen in the line graph). I then wanted to use the IQR method as well since this type of data can change dramatically with certain events. Using the quantile() function to find both the lower and upper quartiles:

```
> Q1 <- quantile(usa_inflation_dataset$Value, 0.25)

> Q3 <- quantile(usa_inflation_dataset$Value, 0.75)

> IQR <- Q3 - Q1
```

> lower_range <- Q1 - 1.5 * IQR

> upper_range <- Q3 + 1.5 * IQR

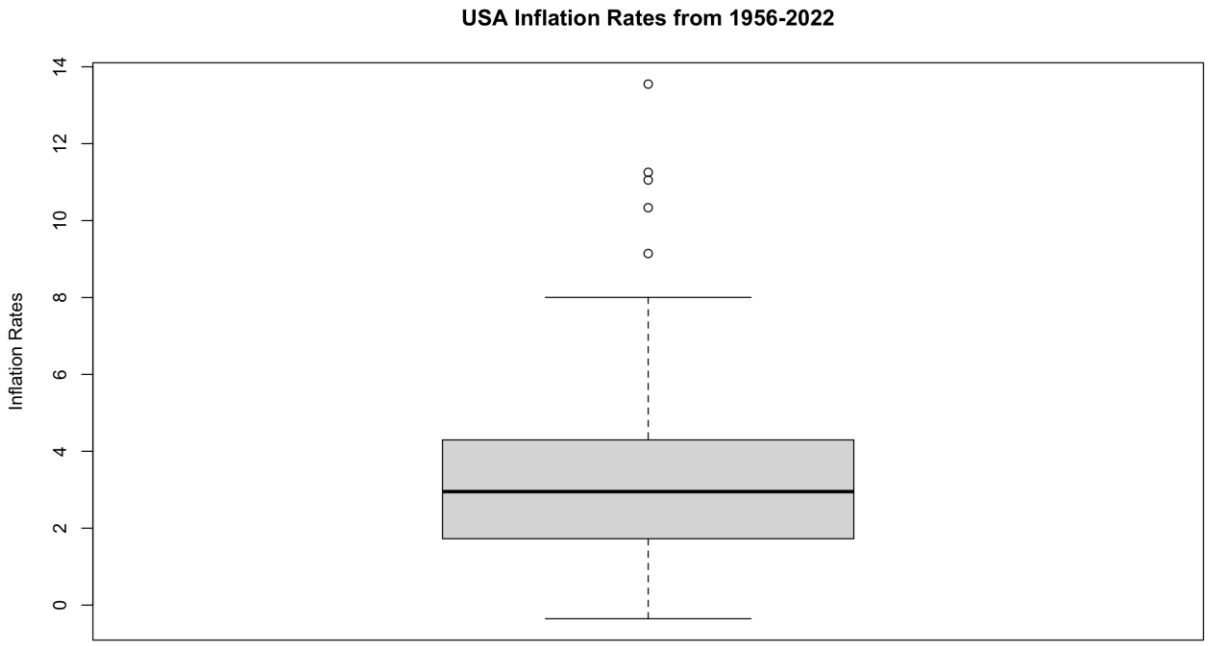
> outliers <- which(usa_inflation_dataset$Value < lower_range | usa_inflation_dataset$Value > upper_range)

> outlier_time <- usa_inflation_dataset$TIME[outliers]

> print(outlier_time)

This gave me 5 different years: 1974, 1975, 1979, 1980, 1981. The IQR results may be a better predictor of outliers for this project as it is less affected by the extreme measures of inflation that have occurred. I then created a boxplot to look further at these outliers.

> boxplot(usa_inflation_dataset$Value, main = "USA Inflation Rates from 1956-2022", ylab = "Inflation Rates")



This showed the 5 outliers, all with higher inflation rates. So, there are extreme highs but not lows. Based on this visual, it seems the majority of the inflation rates tend to sit higher on the scale. The whiskers above the average are longer than below the average.

I did not remove any outliers from this data set because it is important to the question I'm trying to answer. Based on my findings, recent years of inflation have not been more dramatic than inflation rates in the past. These visuals help to compare sudden changes and any outliers we see now and have seen in the past. We see there are similar times in history where rates increased as quickly as now, and times when inflation was higher than now.

To further my research, I would expand to global monthly inflation rates. If I considered including all countries' inflation rates over the specified time, I would better understand the movement of inflation. I would also look at monthly inflation rates to be able to identify how quickly and often rates fluctuate. I would need data including inflation rates from every country across the globe and detailed information to include each month in the year. I would need to combine many sets of data and perform data cleaning on a higher level using so much data at once. I would also need to make decisions about the analysis I perform since it would include both years and months. Adding this information would require a much more extensive cleaning and analysis of the data but would provide me with more results to compare recent inflation rates to past rates.