Project1

10-21-2022

Project 1: Inflation Over the Years of Basic Necessities

By: Ji Hwan Park and Mohammed Furqan Mehboob

Introduction

Eggs, milk, gasoline, and electricity have become basic necessities in order for people to live in the United States of America. Eggs and milk are consumed in huge amounts daily to make the food that people eat every single day. Gasoline is used to power non-electric cars that enable people to go from one place to another. Electricity is vital as it used to watch television, power our phones, and to see the interiors of our rooms with a power of a switch. We have become unaware of how different our lives would be without these necessities.

However, how have the prices of these necessities changed over the course of nearly a decade? Since America is always going into or coming out a recession, we became curious and wanted to look into how prices of these basic necessities changed over the years. We were also curious as to why there might be huge dips or surges in those prices.

The data for this project is from the U.S. Bureau of Labor and Statistics (BLS) and the data sources we used can be seen with this link: https://data.bls.gov/cgi-bin/surveymost?ap. The data sources we selected from the link provided are as follows: Eggs, grade A, large, per doz. (APU0000708111), Milk, fresh, whole, fortified, per gal. (APU0000709112), Electricity per KWH (APU000072610), and Gasoline, unleaded regular, per gallon (APU000074714). All this data is averaged for all US cities. Each row in these data sets represents the year for which each category was recorded. For example, the Electricity per KWH for 2012 or the average price of gasoline for 2016 are some of the observations. The variables within these data sets are all numerical, or consist of numbers. Before tidying, all these data sets had 13 variables and 10 observations.

The data sources found will be joined by year and the month ID variables. These variables were chosen because all the four data sets have these ID variables and as a result, they can be merged neatly together. Some potential relationships we can hypothesize are that prices for eggs, milk, and electricity will increase while gasoline prices will drop dramatically during the years of 2020 and 2021. This would be due to the corona virus pandemic that hit the economy heavily during 2020, increasing prices of groceries by a significant margin. Since people were locked inside, gas prices probably decreased as a result of a decrease in demand.

Tidying the Data

The first task is to dive into the data sources themselves to see whether if they are tidy. Tidy data is better for data analysis as each variable has their own column and each observation has their own row. This makes the data a lot easier to work with.

```
# The following code is a basic setup of options for your document
knitr::opts_chunk$set(echo = TRUE, eval = TRUE,
                      warning = FALSE, message = FALSE,
```

```
                      fig.align = "center",
                      R.options = list(max.print=100))
# loading tidyverse
library(tidyverse)
```

```
## Warning: package 'tidyverse' was built under R version 4.1.3
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v ggplot2 3.3.6      v purrr   0.3.5
## v tibble  3.1.8      v dplyr   1.0.10
## v tidyr   1.2.1      v stringr 1.4.1
## v readr   2.1.3      v forcats 0.5.2
```

```
## Warning: package 'ggplot2' was built under R version 4.1.3
```

```
## Warning: package 'tibble' was built under R version 4.1.3
```

```
## Warning: package 'tidyr' was built under R version 4.1.3
```

```
## Warning: package 'readr' was built under R version 4.1.3
```

```
## Warning: package 'purrr' was built under R version 4.1.3
```

```
## Warning: package 'dplyr' was built under R version 4.1.3
```

```
## Warning: package 'stringr' was built under R version 4.1.3
```

```
## Warning: package 'forcats' was built under R version 4.1.3
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
# looking at the data
head(eggs, 10)
```

```
## # A tibble: 10 x 13
##     Year   Jan   Feb   Mar   Apr   May   Jun   Jul   Aug   Sep   Oct   Nov   Dec
##    <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
##  1  2012  1.94  1.80  1.77  1.83  1.69  1.67  1.65  1.88  1.89  1.96  1.96  2.01
##  2  2013  1.93  1.96  1.92  1.92  1.87  1.86  1.83  1.84  1.90  1.92  1.92  2.03
##  3  2014  2.01  2.00  2.06  2.12  2.00  1.95  1.95  1.98  1.97  1.95  2.03  2.21
##  4  2015  2.11  2.09  2.13  2.06  1.96  2.57  2.57  2.94  2.97  2.81  2.66  2.75
##  5  2016  2.33  2.27  2.08  1.79  1.68  1.49  1.55  1.46  1.47  1.39  1.32  1.38
##  6  2017  1.60  1.46  1.40  1.41  1.41  1.33  1.33  1.37  1.42  1.54  1.51  1.82
##  7  2018  1.77  1.76  1.83  2.08  1.99  1.63  1.72  1.62  1.65  1.66  1.60  1.60
##  8  2019  1.55  1.56  1.54  1.46  1.36  1.20  1.24  1.22  1.38  1.28  1.40  1.54
##  9  2020  1.46  1.45  1.52  2.02  1.64  1.55  1.40  1.33  1.35  1.41  1.45  1.48
## 10  2021  1.47  1.60  1.62  1.62  1.62  1.64  1.64  1.71  1.84  1.82  1.72  1.79
```

```
head(electricity, 10)
```

```
## # A tibble: 10 x 13
##     Year   Jan   Feb   Mar   Apr   May   Jun   Jul   Aug   Sep   Oct   Nov   Dec
##    <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
##  1  2012 0.128 0.128 0.127 0.127 0.129 0.135 0.133 0.133 0.133 0.128 0.127 0.127
##  2  2013 0.129 0.129 0.128 0.128 0.131 0.137 0.137 0.137 0.137 0.132 0.13  0.131
##  3  2014 0.134 0.134 0.135 0.131 0.136 0.143 0.143 0.143 0.141 0.136 0.134 0.135
##  4  2015 0.138 0.138 0.136 0.137 0.137 0.143 0.142 0.142 0.141 0.136 0.134 0.133
##  5  2016 0.134 0.134 0.134 0.134 0.133 0.138 0.139 0.139 0.139 0.134 0.131 0.133
##  6  2017 0.134 0.135 0.134 0.135 0.137 0.142 0.143 0.142 0.142 0.137 0.136 0.136
##  7  2018 0.135 0.135 0.135 0.134 0.136 0.139 0.139 0.139 0.138 0.136 0.134 0.135
##  8  2019 0.135 0.136 0.135 0.135 0.136 0.139 0.14  0.139 0.139 0.136 0.133 0.133
##  9  2020 0.134 0.134 0.134 0.133 0.134 0.137 0.137 0.137 0.137 0.135 0.136 0.136
## 10  2021 0.136 0.137 0.138 0.139 0.14  0.142 0.143 0.144 0.144 0.142 0.142 0.142
```

```
head(gasoline, 10)
```

```
## # A tibble: 10 x 13
##     Year   Jan   Feb   Mar   Apr   May   Jun   Jul   Aug   Sep   Oct   Nov   Dec
##    <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
##  1  2012  3.40  3.57  3.87  3.93  3.79  3.55  3.45  3.71  3.86  3.79  3.49  3.33
##  2  2013  3.35  3.69  3.74  3.59  3.62  3.63  3.63  3.6   3.56  3.38  3.25  3.28
##  3  2014  3.32  3.36  3.53  3.66  3.69  3.70  3.63  3.48  3.40  3.18  2.89  2.56
##  4  2015  2.11  2.25  2.48  2.48  2.78  2.83  2.83  2.68  2.39  2.29  2.18  2.06
##  5  2016  1.97  1.77  1.96  2.13  2.26  2.36  2.22  2.15  2.21  2.24  2.19  2.23
##  6  2017  2.35  2.30  2.32  2.42  2.39  2.34  2.28  2.37  2.63  2.48  2.55  2.46
##  7  2018  2.54  2.58  2.57  2.74  2.91  2.91  2.87  2.86  2.87  2.89  2.67  2.41
##  8  2019  2.29  2.35  2.56  2.84  2.90  2.75  2.78  2.66  2.63  2.67  2.62  2.59
##  9  2020  2.57  2.46  2.27  1.88  1.88  2.08  2.18  2.18  2.19  2.16  2.09  2.17
## 10  2021  2.33  2.50  2.79  2.84  2.97  3.15  3.23  3.26  3.26  3.38  3.48  3.41
```

```
head(milk, 10)
```

```
## # A tibble: 10 x 13
##     Year   Jan   Feb   Mar   Apr   May   Jun   Jul   Aug   Sep   Oct   Nov   Dec
##    <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
##  1  2012  3.58  3.52  3.50  3.47  3.43  3.40  3.43  3.47  3.47  3.52  3.54  3.58
##  2  2013  3.53  3.48  3.43  3.43  3.44  3.46  3.45  3.45  3.43  3.46  3.49  3.50
##  3  2014  3.55  3.56  3.67  3.69  3.74  3.63  3.64  3.67  3.73  3.77  3.86  3.82
##  4  2015  3.76  3.50  3.46  3.40  3.39  3.37  3.43  3.39  3.39  3.34  3.30  3.31
##  5  2016  3.31  3.23  3.19  3.16  3.16  3.12  3.06  3.14  3.23  3.29  3.28  3.29
##  6  2017  3.32  3.3   3.32  3.26  3.24  3.21  3.22  3.17  3.21  3.16  3.15  3.16
##  7  2018  2.96  2.92  2.90  2.92  2.92  2.88  2.84  2.87  2.89  2.91  2.88  2.85
##  8  2019  2.91  2.9   2.94  2.98  2.96  3.05  3.03  3.04  3.10  3.12  3.19  3.19
##  9  2020  3.25  3.20  3.25  3.27  3.21  3.20  3.26  3.41  3.45  3.38  3.42  3.54
## 10  2021  3.47  3.37  3.35  3.45  3.50  3.56  3.63  3.56  3.58  3.66  3.67  3.74
```

From a look at the data sets, it can be confirmed that the data is not tidy because there is no months variable. Instead, all of the months are separated into their own columns (variables), making the data harder to work with. As a result, all the data sets need to be tidied up!

```r
# tidy up all the data using pivot function
eggs <- eggs %>%
  pivot_longer(cols = c('Jan':'Dec'),  # columns to combine
               names_to = 'Month',     # name of new column
               values_to = 'Egg_Price')  # move values to a new column

electricity <- electricity %>%
  pivot_longer(cols = c('Jan': 'Dec'),
               names_to = 'Month',
               values_to = 'Electricity_Price')
gasoline <- gasoline %>%
  pivot_longer(cols = c('Jan': 'Dec'),
               names_to = 'Month',
               values_to = 'Gas_Price')
milk <- milk %>%
  pivot_longer(cols = c('Jan': 'Dec'),
               names_to = 'Month',
               values_to = 'Milk_Price')
#looking at the data
head(eggs, 10)
```

```
## # A tibble: 10 x 3
##     Year Month Egg_Price
##    <dbl> <chr>     <dbl>
##  1  2012 Jan        1.94
##  2  2012 Feb        1.80
##  3  2012 Mar        1.77
##  4  2012 Apr        1.83
##  5  2012 May        1.69
##  6  2012 Jun        1.67
##  7  2012 Jul        1.65
##  8  2012 Aug        1.88
##  9  2012 Sep        1.89
## 10  2012 Oct        1.96
```

```r
head(electricity, 10)
```

```
## # A tibble: 10 x 3
##     Year Month Electricity_Price
##    <dbl> <chr>             <dbl>
##  1  2012 Jan               0.128
##  2  2012 Feb               0.128
##  3  2012 Mar               0.127
##  4  2012 Apr               0.127
##  5  2012 May               0.129
##  6  2012 Jun               0.135
##  7  2012 Jul               0.133
##  8  2012 Aug               0.133
##  9  2012 Sep               0.133
## 10  2012 Oct               0.128
```

```
head(gasoline, 10)
```

```
## # A tibble: 10 x 3
##     Year Month Gas_Price
##    <dbl> <chr>     <dbl>
##  1  2012 Jan        3.40
##  2  2012 Feb        3.57
##  3  2012 Mar        3.87
##  4  2012 Apr        3.93
##  5  2012 May        3.79
##  6  2012 Jun        3.55
##  7  2012 Jul        3.45
##  8  2012 Aug        3.71
##  9  2012 Sep        3.86
## 10  2012 Oct        3.79
```

```
head(milk, 10)
```

```
## # A tibble: 10 x 3
##     Year Month Milk_Price
##    <dbl> <chr>      <dbl>
##  1  2012 Jan         3.58
##  2  2012 Feb         3.52
##  3  2012 Mar         3.50
##  4  2012 Apr         3.47
##  5  2012 May         3.43
##  6  2012 Jun         3.40
##  7  2012 Jul         3.43
##  8  2012 Aug         3.47
##  9  2012 Sep         3.47
## 10  2012 Oct         3.52
```

Using the `pivot_longer` function, we were able to grab the columns from January to December and combine them into a new column, `Month`. This created a new categorical variable that we did not originally have. The values for those months went to a new column called `Milk Price` for the milk data or `Gas Price` for the gasoline data. Due to these functions, each of the data sets went from originally having ten observations to having over 120. Furthermore, the variables reduced from 13 to 3. After implementation, every observation now has its own row and every variable has its own column. Now that the data is all tidy, we can now start joining the data sets together!

Joining/Merging

As mentioned in the Introduction section, the data sources will be joined by the `Month` and `Year` ID variables. The function that will be used to do this is the `full_join` function that merges all the data that is present within the data sources. Since all the data sources have the same ID variables (there are no unique ID variables), the function will work well.

```
# merging eggs with electricity
eggs_elec <- eggs %>%
  full_join(electricity, by = c('Year', 'Month'))
# merging eggs and electricity with gasoline
```

```
eggs_elec_gas <- eggs_elec %>%
  full_join(gasoline, by = c('Year', 'Month'))
# merging milk with previous data set to create final one
cpi <- eggs_elec_gas %>%
  full_join(milk, by = c('Year', 'Month'))
# Looking at the data
cpi
```

```
## # A tibble: 120 x 6
##     Year Month Egg_Price Electricity_Price Gas_Price Milk_Price
##    <dbl> <chr>     <dbl>             <dbl>     <dbl>      <dbl>
##  1  2012 Jan        1.94             0.128      3.40       3.58
##  2  2012 Feb        1.80             0.128      3.57       3.52
##  3  2012 Mar        1.77             0.127      3.87       3.50
##  4  2012 Apr        1.83             0.127      3.93       3.47
##  5  2012 May        1.69             0.129      3.79       3.43
##  6  2012 Jun        1.67             0.135      3.55       3.40
##  7  2012 Jul        1.65             0.133      3.45       3.43
##  8  2012 Aug        1.88             0.133      3.71       3.47
##  9  2012 Sep        1.89             0.133      3.86       3.47
## 10  2012 Oct        1.96             0.128      3.79       3.52
## # ... with 110 more rows
```

From the code above, it can be seen that first, the eggs and electricity data sets were merged together to create the `eggs_elec` data set. The `egg_elec` data was then used to merge with the gasoline data set to create the `eggs_elec_gas` data source. The final merge was between the `eggs_elec_gas` and `milk` data to create the `cpi` data set.

After merging all the data sources together, the `cpi` (or consumer price index) data set now contains 120 observations along with 6 variables. Originally, all the tidy data sources only had 3 variables, `Year`, `Month`, and `Price` along with 120 observations. No observations were added or dropped when joining the data sets together. Furthermore, there were no ID variables left out because the two used (`Year` and `Month`) were common variables present in all the data sources and there were no unique ID variables.

Wrangling

```
# Numerical
cpi %>%
  group_by(Year) %>% # group by the year
  select(Egg_Price) %>%
  summarize(egg_mean_price = mean(Egg_Price, na.rm = T),
            egg_sd_price = sd(Egg_Price, na.rm = T)) %>% # mean and std
  arrange(desc(egg_mean_price))
```

```
## # A tibble: 10 x 3
##     Year egg_mean_price egg_sd_price
##    <dbl>          <dbl>        <dbl>
##  1  2015           2.47       0.373
##  2  2014           2.02       0.0786
##  3  2013           1.91       0.0550
##  4  2012           1.84       0.123
```

```
##  5  2018           1.74       0.156
##  6  2016           1.68       0.355
##  7  2021           1.67       0.106
##  8  2020           1.51       0.183
##  9  2017           1.47       0.136
## 10  2019           1.40       0.136
```

From the table above, it can be seen that egg prices in American cities in 2015 were the highest during the period between 2012-2021 at 2.46 US dollars per dozen. This makes sense because in 2015, America was in a recession due to multiple factors (stock market crash, slow investments, etc.). On the opposite side of the spectrum, the year 2019 was the cheapest year for eggs with only 1.39 US dollars per dozen because of a slight increase in production in 2018, which in turn caused an increase in supply of eggs in the market.

```
# Numerical
cpi %>%
  filter(Year == c(2016, 2017, 2018, 2019, 2020)) %>% # filter years
  group_by(Year) %>%
  summarise(mean_gas = mean(Gas_Price, na.rm = T)) %>%
  arrange(desc(mean_gas), 12) # descending order
```

```
## # A tibble: 5 x 2
##    Year mean_gas
##   <dbl>    <dbl>
## 1  2019     2.79
## 2  2018     2.71
## 3  2017     2.35
## 4  2016     2.06
## 5  2020     2.03
```

From this table, it can be seen that gas prices were slowly increasing between the years of 2016-2019 because the economy was slowly expanding during those years. Due to the Federal Reserve's policy to increase the money supply in the market, prices slowly started to increase. However, due to the corona virus in 2020, the gas prices went down as demand for gas decreased dramatically due to lock down. The prices above clearly demonstrate the trends described.

```
# Numerical
cpi %>%
  group_by(Month) %>% # group by Month
  summarise(egg_mean_price = mean(Egg_Price), # egg mean price
            milk_mean_price = mean(Milk_Price)) %>% # milk mean price
  arrange(desc(egg_mean_price), 12)  # descending order
```

```
## # A tibble: 12 x 3
##    Month egg_mean_price milk_mean_price
##    <chr>          <dbl>           <dbl>
##  1 Dec             1.86            3.40
##  2 Apr             1.83            3.30
##  3 Jan             1.82            3.36
##  4 Feb             1.79            3.30
##  5 Mar             1.79            3.30
##  6 Sep             1.78            3.35
##  7 Oct             1.77            3.36
```

```
##  8 Nov             1.76           3.38
##  9 Aug             1.73           3.32
## 10 May             1.72           3.30
## 11 Jun             1.69           3.29
## 12 Jul             1.69           3.30
```

The table above demonstrates that between 2012-2021, the average egg prices per dozen ($1.85) and milk prices per gallon ($3.3975) were the highest for American cities during the month of December. This makes sense because during December, people are getting ready for Christmas, where a lot of food has to be cooked. As a result of more food, demand for ingredients such as milk and eggs increases which in turn increases prices. However, I was surprised by the fact that the month of November does not have higher mean prices for eggs ($1.750) because of Thanksgiving. This could potentially be due to the fact that Thanksgiving is typically only one day while Christmas is a celebration that takes place over the course of December. However, milk prices for the month of November ($3.37) is right under the month of December. Another interesting thing is that average egg price for the month of April ($1.83) is second to December. This result could once again be due to the holiday of Easter, where eggs play a huge part in the celebration. Overall, this table was able to demonstrate how food ingredients per month over the course of nearly a decade are impacted.

```
# Categorical
mean_milk_price = mean(cpi$Milk_Price, na.rm = T) # total mean price over all the years

cpi %>%  # categorize prices by low, high, and average using average price.
  mutate(price_category = case_when(Milk_Price <= mean_milk_price ~ "Low",
                                    Milk_Price > mean_milk_price ~ "High")) %>%
  group_by(Year, price_category) %>%
  summarise(number_of_prices = n())  %>% # count total # of high and low prices per Year
  arrange(desc(number_of_prices))
```

```
## # A tibble: 12 x 3
## # Groups:   Year [10]
##     Year price_category number_of_prices
##    <dbl> <chr>                     <int>
##  1  2012 High                         12
##  2  2013 High                         12
##  3  2014 High                         12
##  4  2016 Low                          12
##  5  2017 Low                          12
##  6  2018 Low                          12
##  7  2019 Low                          12
##  8  2021 High                         12
##  9  2015 High                         10
## 10  2020 Low                           7
## 11  2020 High                          5
## 12  2015 Low                           2
```

```
cpi %>%  # categorize prices by low, high, and average using average price.
  mutate(price_category = case_when(Milk_Price < mean_milk_price ~ "Low",
                                    Milk_Price > mean_milk_price ~ "High",
                                    Milk_Price == mean_milk_price ~ "Average")) %>%
  group_by(price_category) %>%
  summarise(number_of_prices = n())  %>% # count total # of high and low prices
  arrange(desc(number_of_prices))
```

```
## # A tibble: 2 x 2
##   price_category number_of_prices
##   <chr>                     <int>
## 1 High                         63
## 2 Low                          57
```

We created a price category where High refers to the months with a price that is greater than the average, and vice versa. The code above shows that there are more months that have higher monthly mean prices of milk. The years of 2012, 2013, 2014, 2021, a few months of 2015, and a few months of 2020 are considered to be higher than the average. For instance, in the year 2012, all 12 months have prices that are higher than the average. These years make sense because America was experiencing a recession during those years. There were a total of 63 months that were considered Higher than the average price of milk per gallon. The years of 2016, 2017, 2018, 2019, a few months of 2020, and a few months of 2015 were considered to be lower. Such patterns in these years can be explained by the fact that America recovered from the recession and started to increase the production of milk. There are total of 57 months that are considered to be lower than the average.

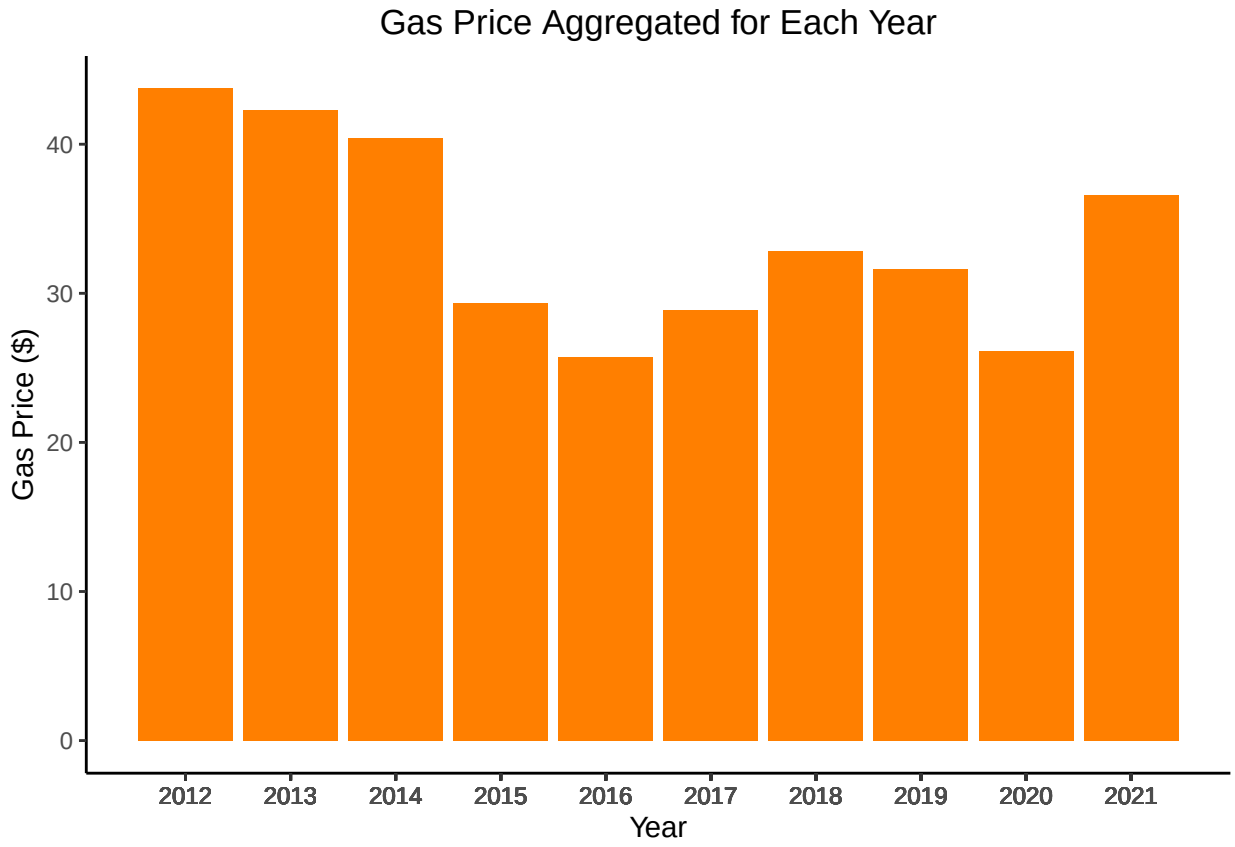
```
# Categorical
cpi %>%
  # Determining the season by the month
  mutate(seasons = case_when(Month == 'Jun' | Month == 'Jul' |
                               Month == 'Aug' ~ "Summer",
                             Month == 'Sep' | Month ==  'Oct' |
                               Month == 'Nov' ~ "Fall",
                             Month == 'Dec'| Month == 'Jan' |
                               Month =='Feb' ~ "Winter",
                             Month == 'Mar'| Month == 'Apr'|
                               Month =='May'  ~ "Spring")) %>%
  group_by(seasons) %>%
  # find mean elec. per season
  summarise(mean_elec = mean(Electricity_Price, na.rm = T),
            sd_elec = sd(Electricity_Price, na.rm = T)) %>%
  arrange(desc(mean_elec))
```

```
## # A tibble: 4 x 3
##   seasons mean_elec sd_elec
##   <chr>       <dbl>   <dbl>
## 1 Summer      0.140 0.00306
## 2 Fall        0.136 0.00417
## 3 Spring      0.134 0.00342
## 4 Winter      0.134 0.00327
```

The table above shows that the mean price of electricity per KWH (kilowatts per hour) is the highest for American cities during the summer season at 0.139 US dollars. This result is logical because during the summer, kids tend to stay at home to either watch television, watch movies, or play video games. This increase in the use of electricity would increase demand, and as a result, the price of electricity would increase. It was quite surprising that the Fall season has a larger mean electricity price ($0.136) than the Winter season ($0.1339). It would be assumed that during the Winter, people would require more electricity as a result of using heaters. However, since the price difference is so minimal, it can not be confirmed there is a significant difference among the prices.

Visualization

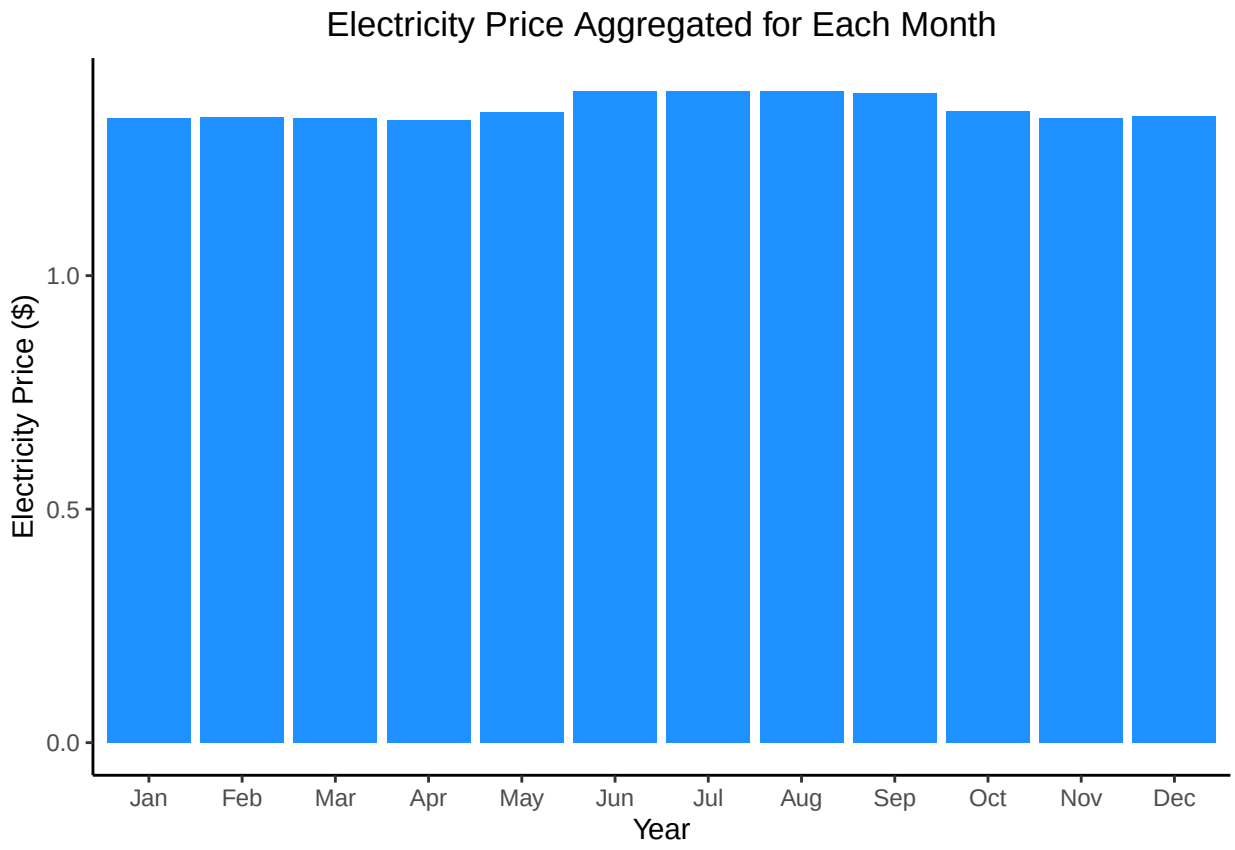
```
# One Variable
cpi %>%
  ggplot(aes(x = Year, y = Gas_Price)) +
  geom_bar(stat = 'identity', fill = 'darkorange1') + # create a bar plot
  scale_x_continuous(breaks = cpi$Year) + # map x-axis
  labs(x= "Year", y = 'Gas Price ($)',   # add axis titles
       title = 'Gas Price Aggregated for Each Year') +
  theme_classic() + # add theme
  theme(plot.title = element_text(hjust = 0.5))
```



The plot above shows how the aggregated gas price in the US changed across the years. The gas price is an unique index that is greatly affected by the global demand and supply chain in the oil market. The gas price hit the record high in the year 2012 since the recession due to the financial crisis and stayed over $40 until 2014. According to American Automobile Association (AAA), the factors that contributed to such record high gas price are hurricanes, refinery outages, and tensions in the Middle East. From the year 2015, the oil supply went back to settle down to a normal range and hovered around $30.

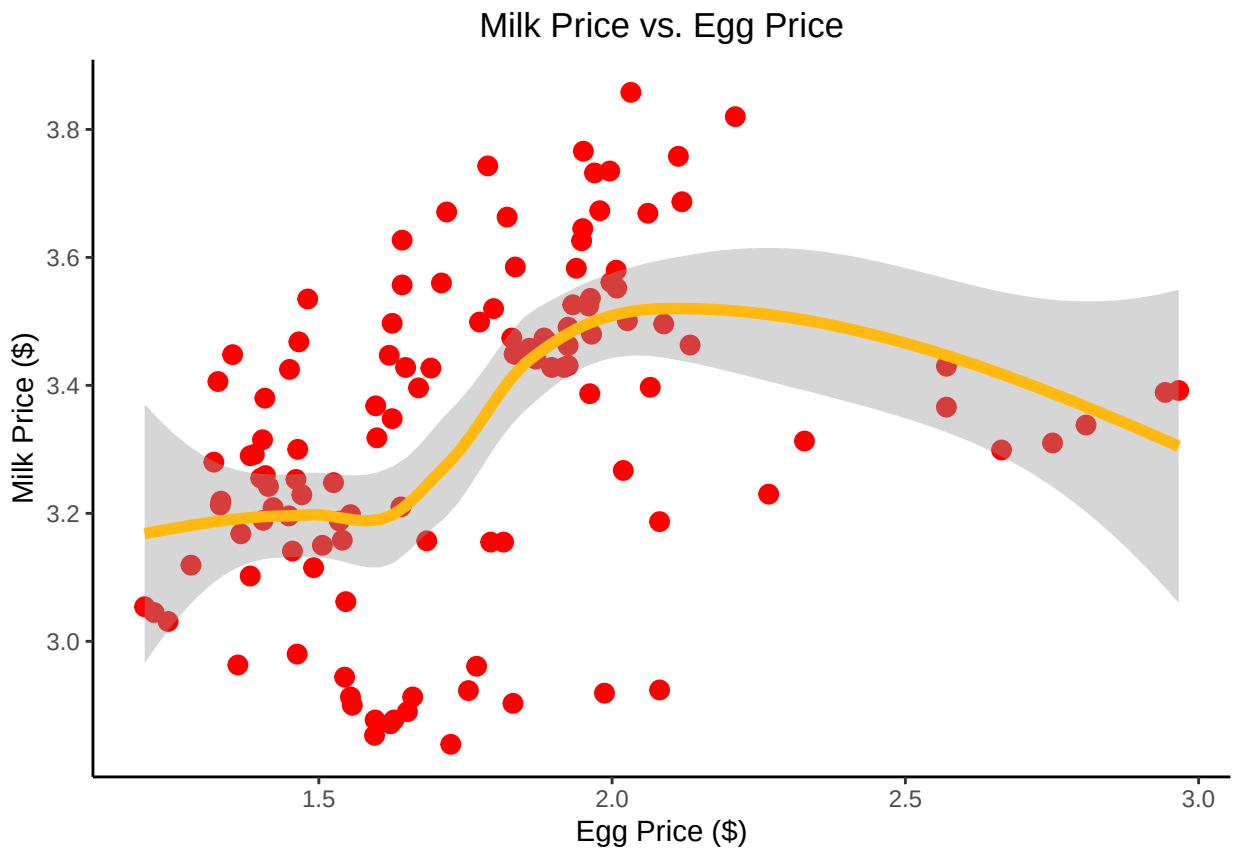
```
# One Variable
cpi %>%
  group_by(Month) %>%  # group by the month
  ggplot(aes(x = Month, y = Electricity_Price)) +
  geom_bar(stat = 'identity', fill = 'dodgerblue') + # create a bar plot
```

```
  labs(x= "Year", y = 'Electricity Price ($)',   # add axis titles
       title = 'Electricity Price Aggregated for Each Month')  +
  scale_x_discrete(limits = c("Jan","Feb","Mar", "Apr", "May",
                              "Jun", "Jul", "Aug", "Sep", "Oct",
                              "Nov", "Dec")) +
  theme_classic() + # add theme
  theme(plot.title = element_text(hjust = 0.5))
```



The plot above shows the unique nature of the aggregated electricity price in the US. It is clear that it does not change significantly across the months for each year. One major factor that prevents a huge fluctuation of electricity price is the fact that the Federal Energy Regulatory Commission (FERC) regulates the electricity price. There is only a small increase of electricity price during the summer season (from June to September), yet no drastic change in specific season or months due to the price regulation policy of the federal government.

```
# Two variable
cpi %>%
  ggplot(aes(x = Egg_Price, y= Milk_Price)) +
  geom_point(color = 'red', size = 3) +  # add scatterplot
  geom_smooth(color = 'darkgoldenrod1', size = 2) + # add a trendline
  scale_y_continuous(breaks = seq(3.0, 3.9, 0.2)) + # map y - axis
  labs(x= "Egg Price ($)", y = 'Milk Price ($)', # add axis labels
       title = 'Milk Price vs. Egg Price') +
  theme_classic() +  # add theme
  theme(plot.title = element_text(hjust = 0.5))
```

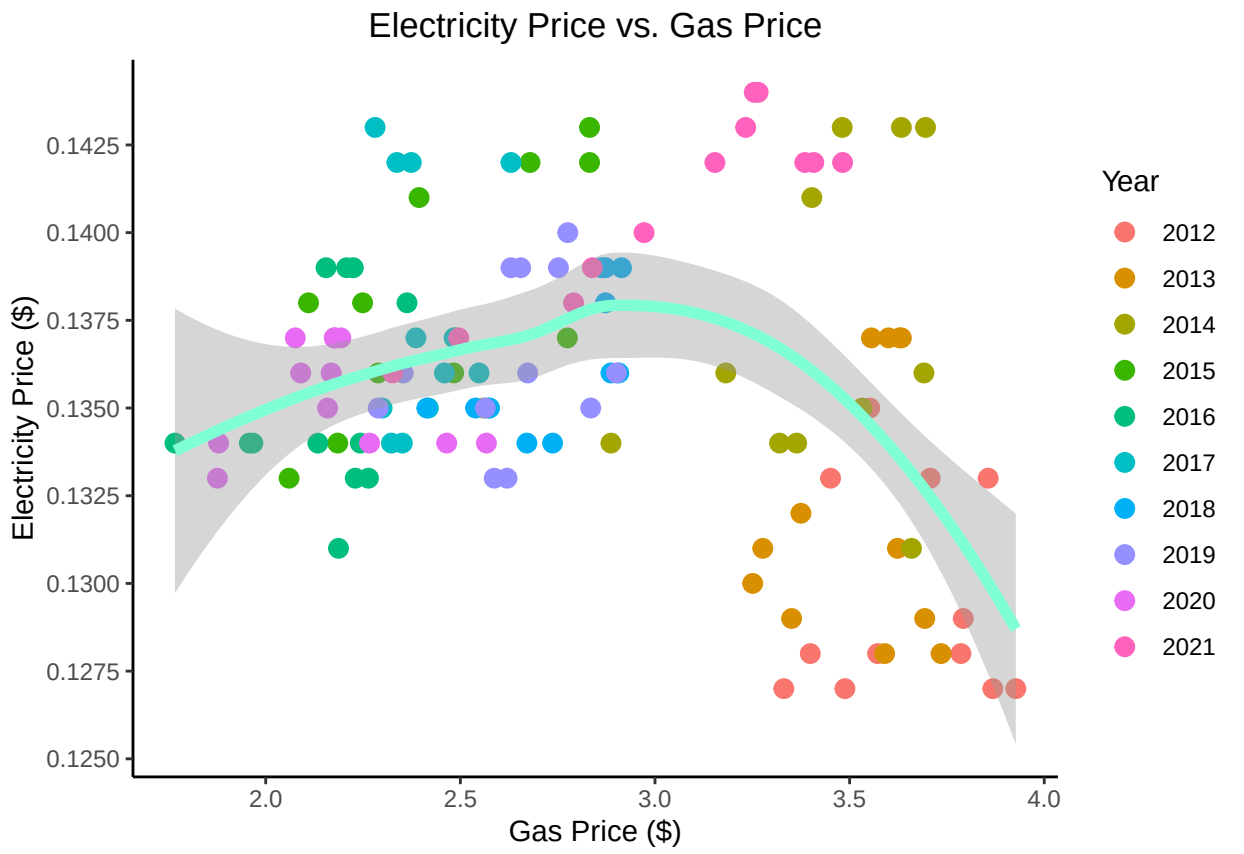



```
cor(cpi$Egg_Price, cpi$Milk_Price) # find correlation
```

```
## [1] 0.3658371
```

There is no obvious pattern between the milk and egg price from the graph above. Furthermore, the correlation coefficient for egg and milk price is only 0.3658, which tell us that there is only a small correlation between the two variables. However, it can bee seen that the egg price has wider range of price variation from ($1.0 to $3.0) compared to that of milk ($2.9 to $3.8). This shows that the egg prices are more of a sensitive index to economic situations such as inflation. Milk price normally does not vary much compared to other products due to its overproduction by the US farms.

```
# Two variables
cpi %>%
  ggplot(aes(x = Gas_Price, y= Electricity_Price)) +
  geom_point(aes(color = as.factor(Year)), size = 3) + # add scatterplot
  geom_smooth(color = 'aquamarine', size = 2) + # add trendline
  scale_y_continuous(breaks = seq(0.125, 0.150, 0.0025)) + # map y - axis
  scale_x_continuous(breaks = seq(1.5, 4.2, 0.5)) +  # map x - axis
   labs(x= "Gas Price ($)", y = 'Electricity Price ($)', # add axis labels
       title = 'Electricity Price vs. Gas Price',
       colour = "Year") +
  theme_classic() + # add theme
  theme(plot.title = element_text(hjust = 0.5))
```



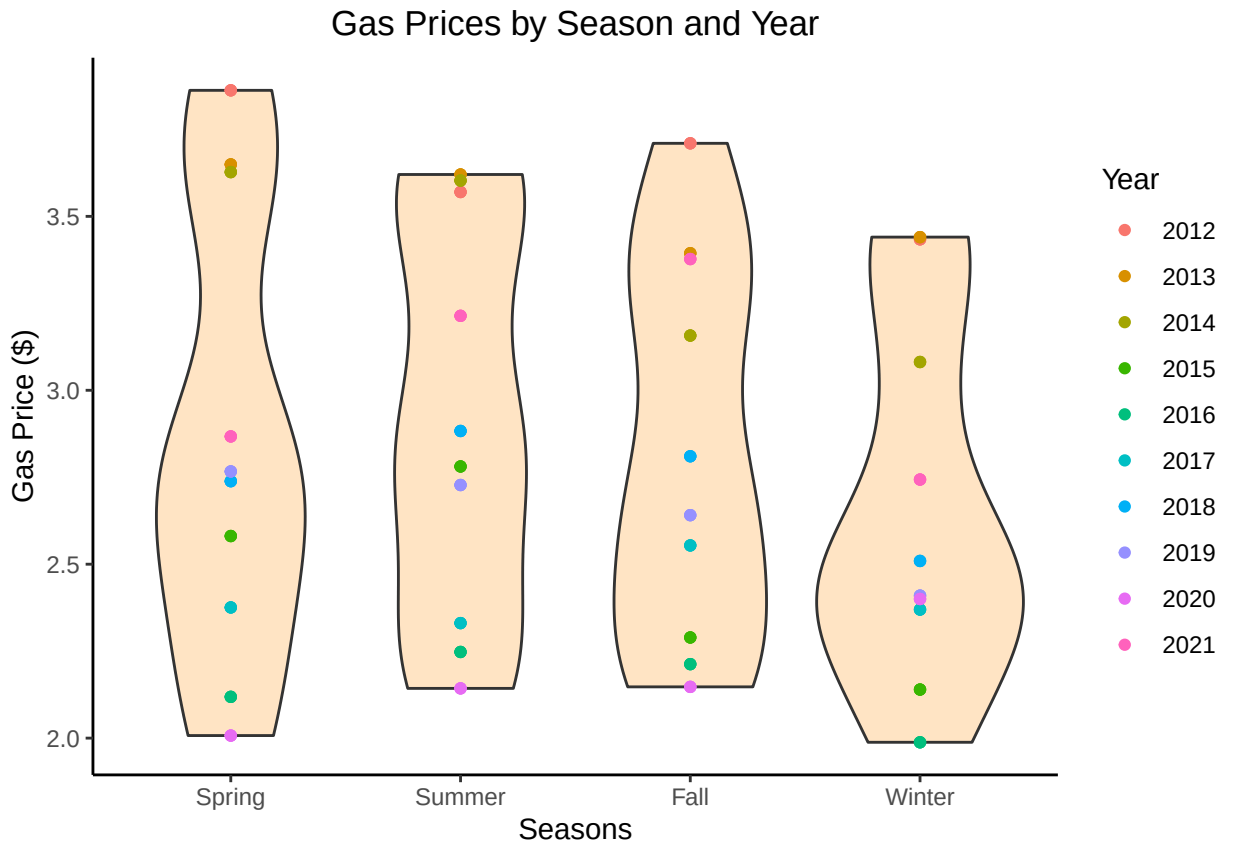

```
cor(cpi$Gas_Price, cpi$Electricity_Price) # find correlation
```

```
## [1] -0.1902064
```

The plot above shows that there is no obvious pattern between electricity price and gas price. This is further supported by the correlation analysis, which is only -0.1902. This is due to the fact that the gas price is greatly affected by the external factors such as global oil demand and supply and the electricity is regulated by the federal government. Due to the different factors that influence both of these variables, a common trend or relationship can not be found.

```
# 3 variables
cpi_season <- cpi %>%
  # create a seasons variable
  mutate(seasons = case_when(Month == 'Jun' | Month == 'Jul' |
                               Month == 'Aug' ~ "Summer",
                             Month == 'Sep' | Month ==  'Oct' |
                               Month == 'Nov' ~ "Fall",
                             Month == 'Dec'| Month == 'Jan' |
                               Month =='Feb' ~ "Winter",
                             Month == 'Mar'| Month == 'Apr'|
                               Month =='May'  ~ "Spring"))
cpi_season %>%
  group_by(Year, seasons) %>%
  mutate(gas_mean = mean(Gas_Price, na.rm = T))  %>% # find the mean
  ggplot(aes(x = seasons, y = gas_mean)) +
```

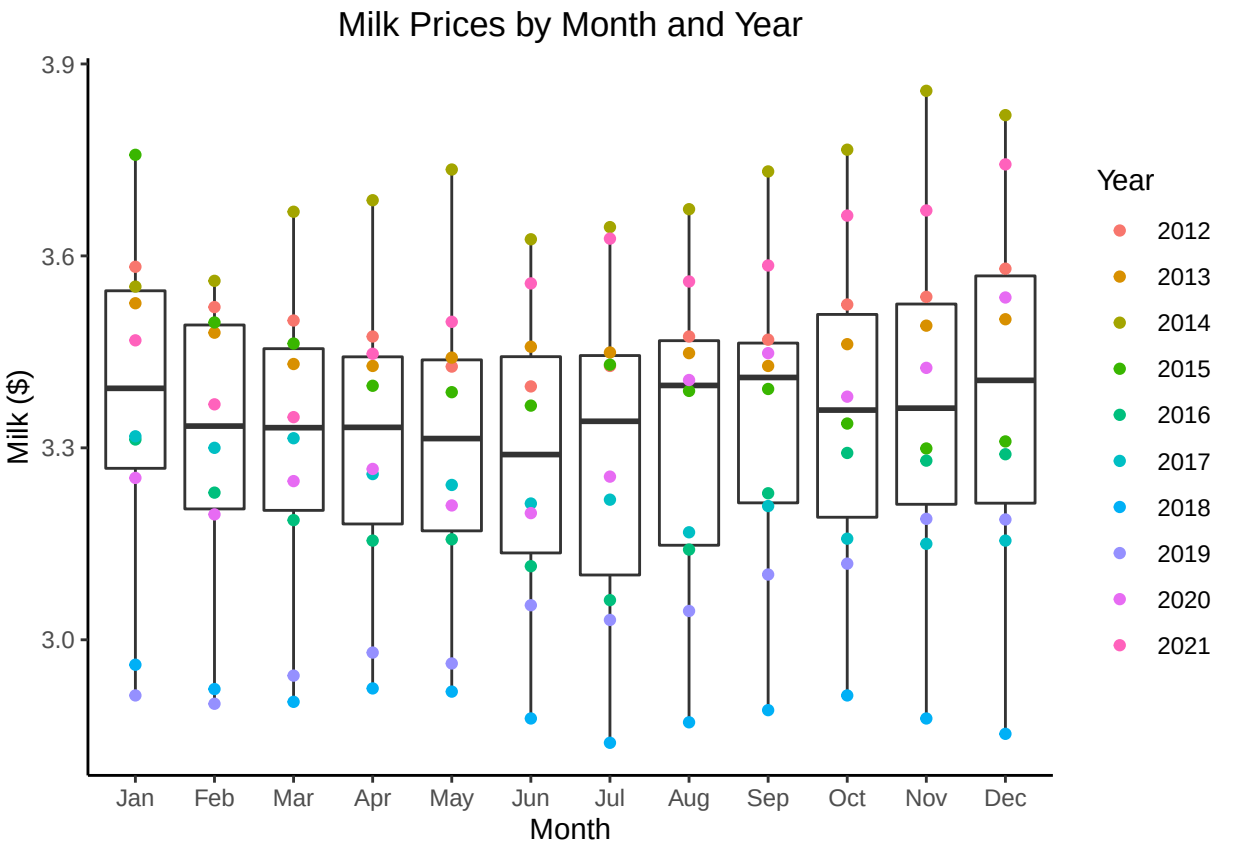
```
  geom_violin(fill = 'bisque') +  # create violin plot
  geom_point(aes(color = as.factor(Year))) + # create scatterplot
  scale_x_discrete(limits = c("Spring", "Summer", # map x - axis
                              "Fall", "Winter"))  +
  labs(x= "Seasons", y = "Gas Price ($)", # add axis titles
       title = 'Gas Prices by Season and Year',
       colour = 'Year') +
  theme_classic() + # add theme
  theme(plot.title = element_text(hjust = 0.5))
```



As it can be noticed from the plot above, the year 2012 was the year at which gas prices were typically the highest across all the seasons. As explained above, the year 2012 contained problems such as hurricanes, refinery outages, or just increased tensions with the Middle East that influenced gas prices. Another trend that can be seen from the graph is that Spring and Winter Gas Prices tended to be lower during the post-2016 era. This could potentially be due to better relations with the Middle East and oil prices returning back to their previous prices. From the violin plots, it can be seen that most of gas prices were between the 2-3 US dollar range. Furthermore, it seems like the Spring has more variation in gas prices across the years.

```
# 3 variables
cpi %>%
  ggplot(aes(x = Month, y = Milk_Price)) +
  geom_boxplot() + # add boxplot
  geom_point(aes(color = as.factor(Year))) + # add scatterplot
  scale_x_discrete(limits = c("Jan","Feb","Mar", "Apr", "May", # map x-axis
                              "Jun", "Jul", "Aug", "Sep", "Oct",
```

```
                                "Nov", "Dec")) +
labs(x = "Month", y = 'Milk ($)',
     title = 'Milk Prices by Month and Year',
     colour = 'Year') + # add axis labels
theme_classic()  + # add theme
theme(plot.title = element_text(hjust = 0.5))
```



The plot above shows that there is no drastic change of milk price across the months. The milk price surprisingly hit the record high in the year of 2014 due to its external high exports demand in that year. The milk derivative product such as cheese was under high export demand. The milk price is a good consumer product index that gives a hint how well the economy is doing. From 2015 to 2019, when the economy is strong the milk price generally stayed under the median, whereas when the economy is uncertain, the price of milk starts to have a greater fluctuation.

Conclusion

After tidying the data, merging the data, and doing comprehensive analysis on the merged data set, a lot of trends were found. The general idea derived from these trends are that prices for any of these categories are hard to predict. They are in a constant state of change and are regulated by multiple factors that are not entirely dependent on one another. Overall, this was a great look into how prices are in a constant state of flux for basic necessities in the United States.

Formatting

Contributions:

Mohammed Furqan Mehboob and Ji Hwan Park worked together to wrangle the data and create the visualizations. Mohammed wrote the introduction and tidied the data. Ji Hwan Park merged the data together and wrote the conclusion.

Comment your code, write full sentences, and knit your file!

```
Sys.info()
```

```
##           sysname           release           version          nodename
##         "Windows"          "10 x64"     "build 22621" "DESKTOP-NUBTKGG"
##           machine             login              user    effective_user
##          "x86-64"          "furqa"           "furqa"           "furqa"
```