VideoGameSales

This document represents a reevaluation and refinement of a previous project I participated in earlier in 2023. While many of the ideas originate from that project, this iteration emphasizes the implementation of more suitable statistical analysis techniques, while also showcasing a substantial effort in rewriting the code to enhance readability and conciseness.

The original prompt: "Our company wants to know how video game sales react to economic conditions that may affect consumer spending habits."

```
library(tidyverse)
library(vroom)
library(DataExplorer)
library(prophet)
library(patchwork)
```

Importing Libraries

```
get_df_from_php <- function(file.php){

  # Read php to string
  php_string <- read_file(file.php)

  # Separate into relavent consoles
  DS <- c('DS', str_extract(php_string, "(?s)name:'DS',(.*)\\},\\]"))
  NS <- c('NS', str_extract(php_string, "(?s)name:'NS',(.*)\\},\\]"))
  PS3 <- c('PS3', str_extract(php_string, "(?s)name:'PS3',(.*)\\},\\]"))
  PS4 <- c('PS4', str_extract(php_string, "(?s)name:'PS4',(.*)\\},\\]"))
  PS5 <- c('PS5', str_extract(php_string, "(?s)name:'PS5',(.*)\\},\\]"))
  PSP <- c('PSP', str_extract(php_string, "(?s)name:'PSP',(.*)\\},\\]"))
  PSV <- c('PSV', str_extract(php_string, "(?s)name:'PSV',(.*)\\},\\]"))
  ThreeDS <- c('ThreeDS', str_extract(php_string, "(?s)name:'3DS',(.*)\\},\\]"))
  Wii <- c('Wii', str_extract(php_string, "(?s)name:'Wii',(.*)\\},\\]"))
  WiiU <- c('WiiU', str_extract(php_string, "(?s)name:'WiiU',(.*)\\},\\]"))
  x360 <- c('x360', str_extract(php_string, "(?s)name:'X360',(.*)\\},\\]"))
  XOne <- c('XOne', str_extract(php_string, "(?s)name:'XOne',(.*)\\},\\]"))
  Xs <- c('Xs', str_extract(php_string, "(?s)name:'XS',(.*)\\},\\]"))

  # Add to list
  console_string_list <- list(DS,
                              NS,
                              PS3,
```

```
                          PS4,
                          PS5,
                          PSP,
                          PSV,
                          ThreeDS,
                          Wii,
                          WiiU,
                          x360,
                          XOne,
                          Xs)

  # Convert to df's and join together
  first_table = TRUE
  for (console in console_string_list){
    x <- str_extract_all(console[2], '(?<=x: )\\d*(?=,)')
    y <- str_extract_all(console[2], '(?<=y: )\\d*(?=\\n)')
    df <- tibble(x = as.numeric(unlist(x)), y = as.numeric(unlist(y)))
    df <- df %>%
      group_by(x) %>%
      summarise(y = sum(y))
    colnames(df) <- c('t', console[1])
    if (first_table){
      df_all <- df
      first_table = FALSE
    }
    else{
      df_all <- df_all %>%
        full_join(df, by = 't')
    }
  }

  # Convert unix timestamp to datetime
  df_all <- df_all %>%
    mutate(t = as_datetime(t/1000))

  return(df_all)
}

# Getting hardware sales data
HardwareGlobalWeekly <- get_df_from_php('hw_date.php') %>%
  mutate_all(~replace(., is.na(.), 0)) %>%
  transmute(time = t,
            Global_h_sales = DS + NS + PS3 + PS4 + PS5 + PSP + PSV + ThreeDS + Wii + WiiU + x360 + XOne
HardwareUSAWeekly <- get_df_from_php('hw_date_USA.php') %>%
  mutate_all(~replace(., is.na(.), 0)) %>%
  transmute(time = t,
            USA_h_sales = DS + NS + PS3 + PS4 + PS5 + PSP + PSV + ThreeDS + Wii + WiiU + x360 + XOne + X
HardwareJapanWeekly <- get_df_from_php('hw_date_Japan.php') %>%
  mutate_all(~replace(., is.na(.), 0)) %>%
  transmute(time = t,
            Japan_h_sales = DS + NS + PS3 + PS4 + PS5 + PSP + PSV + ThreeDS + Wii + WiiU + x360 + XOne +

hardware_sales_yearly <- HardwareGlobalWeekly %>%
```

```
  inner_join(HardwareUSAWeekly, by = 'time') %>%
  inner_join(HardwareJapanWeekly, by = 'time') %>%
  mutate(Year = year(time)) %>%
  group_by(Year) %>%
  summarise(Global_h_sales = sum(Global_h_sales),
            USA_h_sales = sum(USA_h_sales),
            Japan_h_sales = sum(Japan_h_sales))


# Getting software sales data (from kaggle)
software_sales_yearly <- vroom('vgsales.csv') %>%
  mutate(Year = as.numeric(Year)) %>%
  group_by(Year) %>%
  summarise(Global_s_sales = sum(Global_Sales)*1000000,
            USA_s_sales = sum(NA_Sales)*1000000,
            Japan_s_sales = sum(JP_Sales)*1000000)

# Getting data for potential economic indicators of video game sales
# *These csv's were created using public data from the OECD and World Bank*
economic_indicators <- vroom('Indicators_edit.csv')

cci <- vroom('CCI.csv') %>%
  pivot_wider(names_from = LOCATION, values_from = CCI) %>%
  mutate(Year = as.numeric(substr(TIME, start = 1, stop = 4))) %>%
  group_by(Year) %>%
  summarise(Global_cci = mean(OECD),
            USA_cci = mean(USA),
            Japan_cci = mean(JPN))

# Creating one global df
omni <- hardware_sales_yearly %>%
  full_join(software_sales_yearly, by = 'Year') %>%
  left_join(economic_indicators, by = 'Year') %>%
  left_join(cci, by = 'Year') %>%
  select(-...1) %>%
  arrange(Year) %>%
  # Dropping incomplete year 2023
  filter(Year != 2023) %>%
  rename(JP_inf = `JPN Inflation (annual %)`,
         JP_gdp_change = `JPN GDP per capita growth (annual %)`,
         JP_gdp = `JPN GDP per capita`,
         JP_rate = `JPN Lending interest rate (%)`,
         JP_unemp = `JPN Unemployment`,
         JP_cpi = `JPN Consumer price index`,
         JP_cci = `Japan_cci`,
         US_inf = `US Inflation (annual %)`,
         US_gdp_change = `US GDP per capita growth (annual %)`,
         US_gdp = `US GDP per capita`,
         US_rate = `US Lending interest rate (%)`,
         US_unemp = `US Unemployment`,
         US_cpi = `US Consumer price index`,
         US_cci = `USA_cci`, )
```

```r
# Creating df's with target predictor variables

US_hardware_df <- omni %>%
  select(1, 3, 14, 15, 16, 17, 18, 19, 21)
US_software_df <- omni %>%
  select(1, 6, 14, 15, 16, 17, 18, 19, 21)
```

Data Wrangling

```r
print(omni)
```

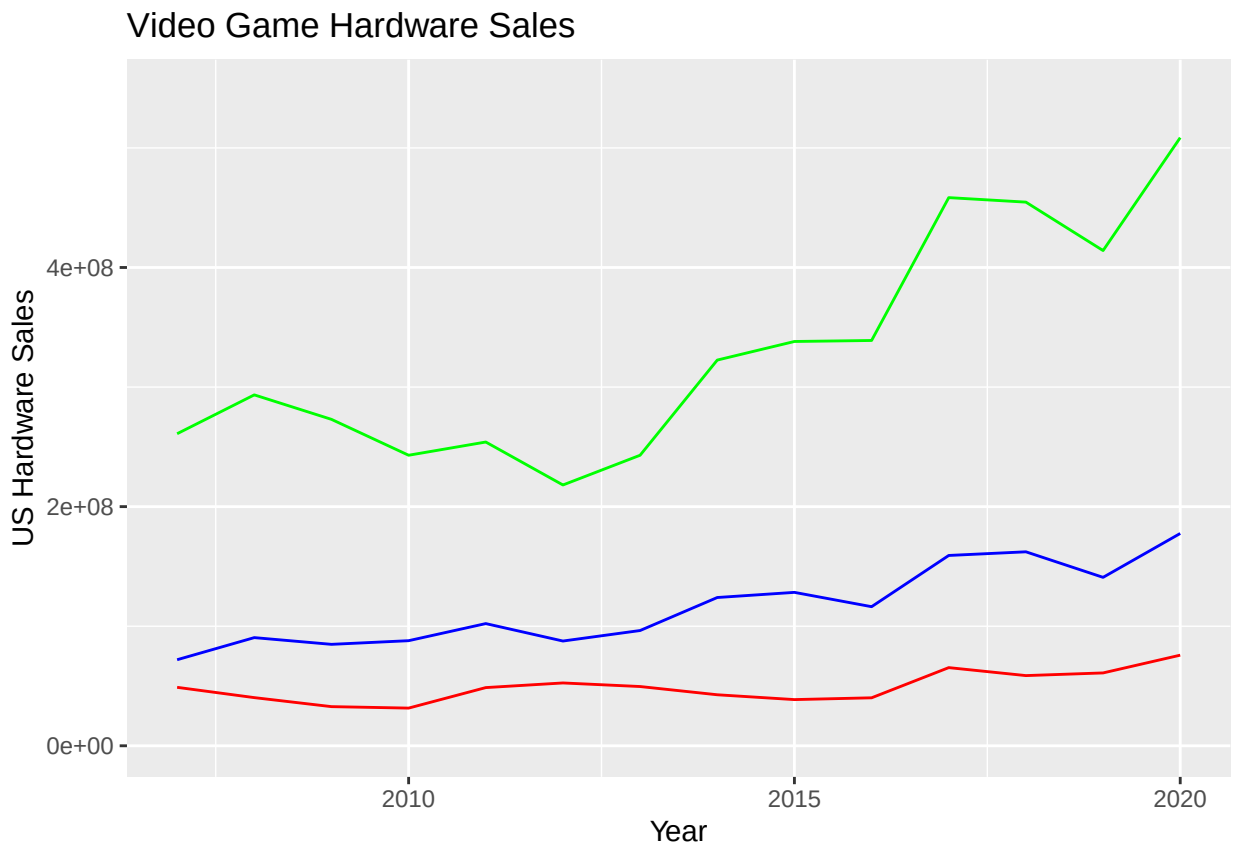
Exploratory Data Analysis

```
## # A tibble: 43 x 22
##     Year Global_h_sales USA_h_sales Japan_h_sales Global_s_sales USA_s_sales
##    <dbl>          <dbl>       <dbl>         <dbl>          <dbl>       <dbl>
##  1  1980             NA          NA            NA       11380000    10590000
##  2  1981             NA          NA            NA       35770000    33400000
##  3  1982             NA          NA            NA       28860000    26920000
##  4  1983             NA          NA            NA       16790000     7760000
##  5  1984             NA          NA            NA       50360000    33280000
##  6  1985             NA          NA            NA       53940000    33730000
##  7  1986             NA          NA            NA       37070000    12500000
##  8  1987             NA          NA            NA       21740000     8460000
##  9  1988             NA          NA            NA       47220000    23870000
## 10  1989             NA          NA            NA       73450000    45150000
## # i 33 more rows
## # i 16 more variables: Japan_s_sales <dbl>, JP_inf <dbl>, JP_gdp_change <dbl>,
## #   JP_gdp <dbl>, JP_rate <dbl>, JP_unemp <dbl>, JP_cpi <dbl>, US_inf <dbl>,
## #   US_gdp_change <dbl>, US_gdp <dbl>, US_rate <dbl>, US_unemp <dbl>,
## #   US_cpi <dbl>, Global_cci <dbl>, US_cci <dbl>, JP_cci <dbl>
```

Data for software sales appear to be reliable from 1983 - 2013.
Data for hardware sales appear to be reliable from 2007 - 2022.
Available Economic data appear cuts off at 2020.

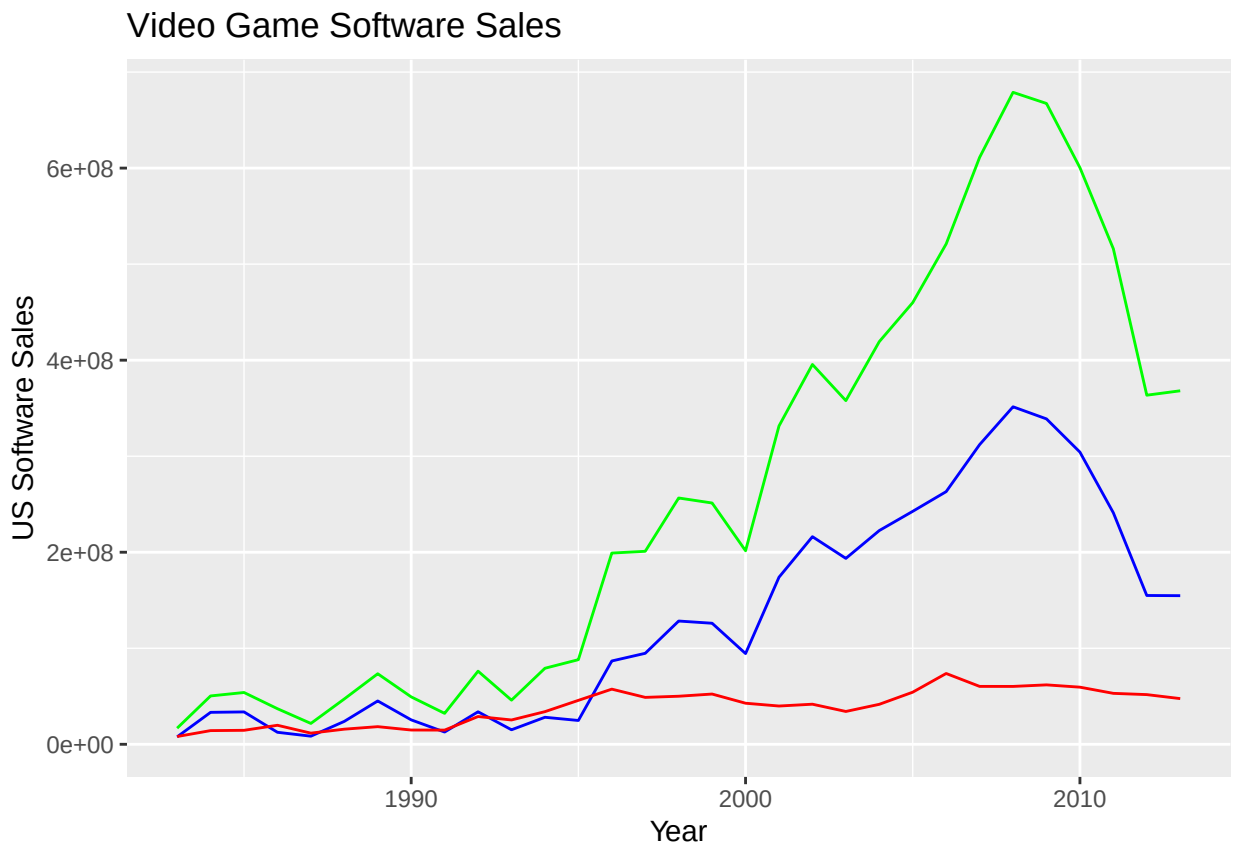
```r
US_hardware_df <- US_hardware_df %>%
  filter(Year >= 2007,
         Year <= 2020)
US_software_df <- US_software_df %>%
  filter(Year >= 1983,
         Year <= 2015)
```

```r
ggplot(data = omni) +
  geom_line(aes(x = Year, y = USA_h_sales), color = 'blue') +
  geom_line(aes(x = Year, y = Japan_h_sales), color = 'red') +
  geom_line(aes(x = Year, y = Global_h_sales), color = 'green') +
  xlim(2007, 2020) +
  ylab('US Hardware Sales') +
  labs(title = 'Video Game Hardware Sales')
```



Hardware sales data are limited, but sales appear to be somewhat consistent, increasing slightly.

```
ggplot(data = omni) +
  geom_line(aes(x = Year, y = USA_s_sales), color = 'blue') +
  geom_line(aes(x = Year, y = Japan_s_sales), color = 'red') +
  geom_line(aes(x = Year, y = Global_s_sales), color = 'green') +
  xlim(1983, 2013) +
  ylab('US Software Sales') +
  labs(title = 'Video Game Software Sales')
```

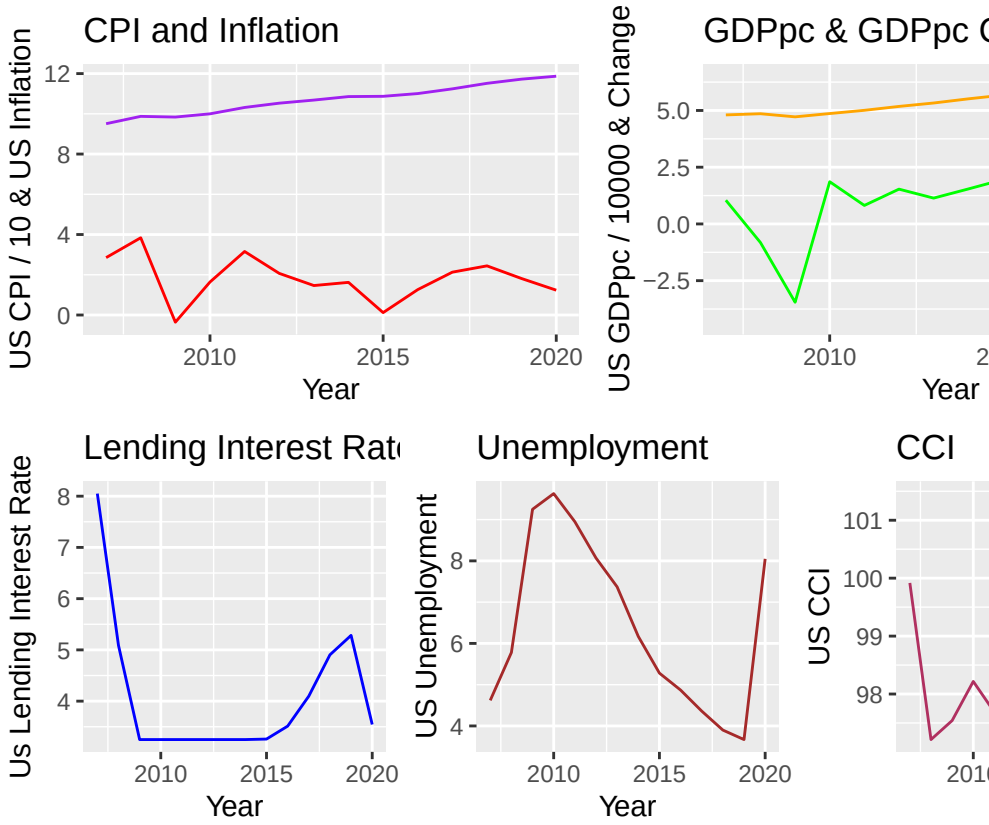


Video game software appears to be somewhat of a novelty from 1983 - 1995 and then gains popularity until about 2008. We suspect that software sales will react to economic indicators differently depending on whether people consider video games to be a novelty or an integral part of their entertainment.

We will examine three different periods of video game sales starting with the most recent: 2007 - 2020 hardware sales, 1995 - 2013 software sales, and 1983 - 1995 software sales. For Data Analysis, we will use data from the United States.

```
# Plotting US economic indicators against hardware sales
plot_1 <- ggplot(data = US_hardware_df) +
  geom_line(aes(x = Year, y = US_cpi/10), color = 'purple') +
  geom_line(aes(x = Year, y = US_inf), color = 'red') +
  ylab('US CPI / 10 & US Inflation') +
  labs(title = 'CPI and Inflation')

plot_2 <- ggplot(data = US_hardware_df) +
  geom_line(aes(x = Year, y = US_gdp/10000), color = 'orange') +
  geom_line(aes(x = Year, y = US_gdp_change), color = 'green') +
  ylab('US GDPpc / 10000 & Change') +
  labs(title = 'GDPpc & GDPpc Change')

plot_3 <- ggplot(data = US_hardware_df) +
  geom_line(aes(x = Year, y = US_rate), color = 'blue') +
  ylab('Us Lending Interest Rate') +
```

```
  labs(title = 'Lending Interest Rate')

plot_4 <- ggplot(data = US_hardware_df) +
  geom_line(aes(x = Year, y = US_unemp), color = 'brown') +
  ylab('US Unemployment') +
  labs(title = 'Unemployment')

plot_5 <- ggplot(data = US_hardware_df) +
  geom_line(aes(x = Year, y = US_cci), color = 'maroon') +
  ylab('US CCI') +
  labs(title = 'CCI')

(plot_1 + plot_2) / (plot_3 + plot_4 + plot_5)
```



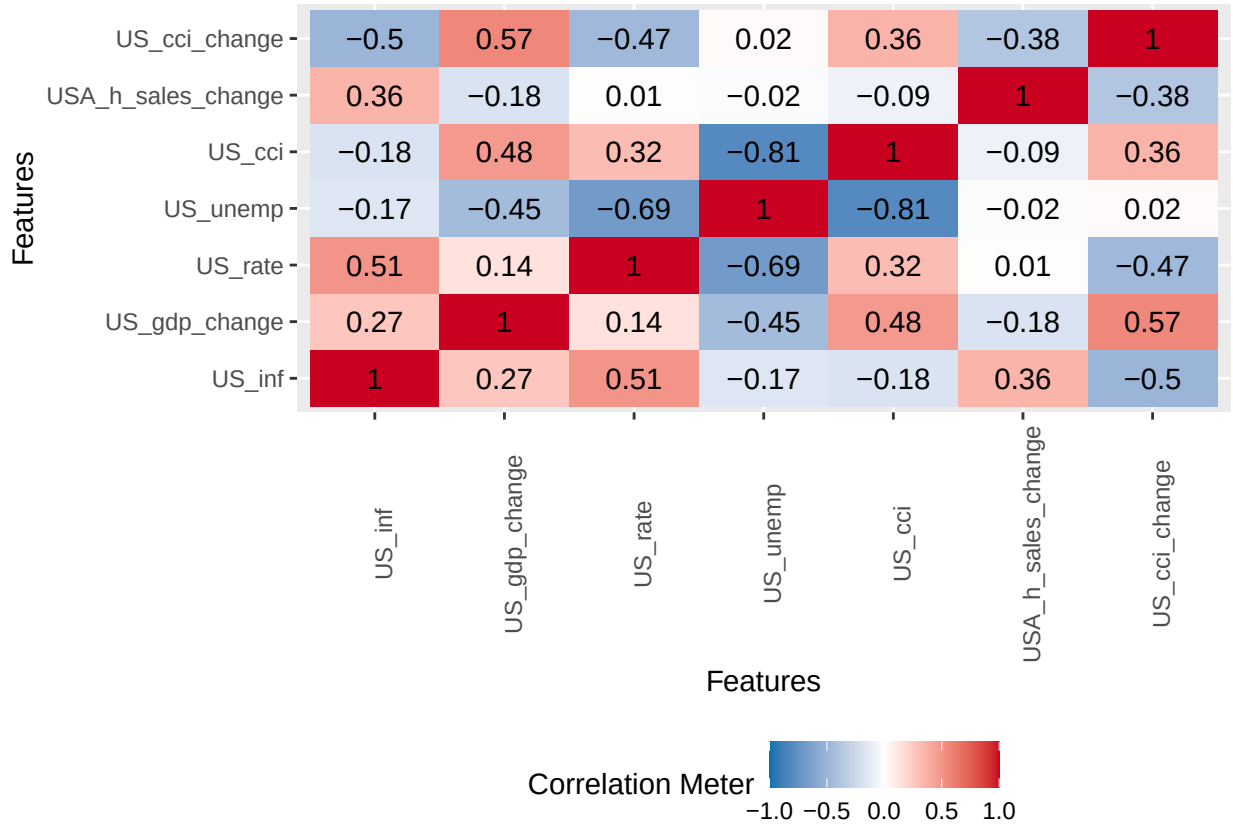
2007 - 2020 US Hardware Sales

While most variables are stationary, CPI and GDP have upward trends, which poses a problem for many statistical modeling techniques. We have Percent Change columns for both of these measurements, so we will drop CPI and GDP, and use Inflation rate and GDP per capita growth instead.

It's also evident that hardware sales has a slight upward trend. To mitigate this issue, we will try to find correlations between the percent change in sales and economic indicators, instead of the actual total sales.

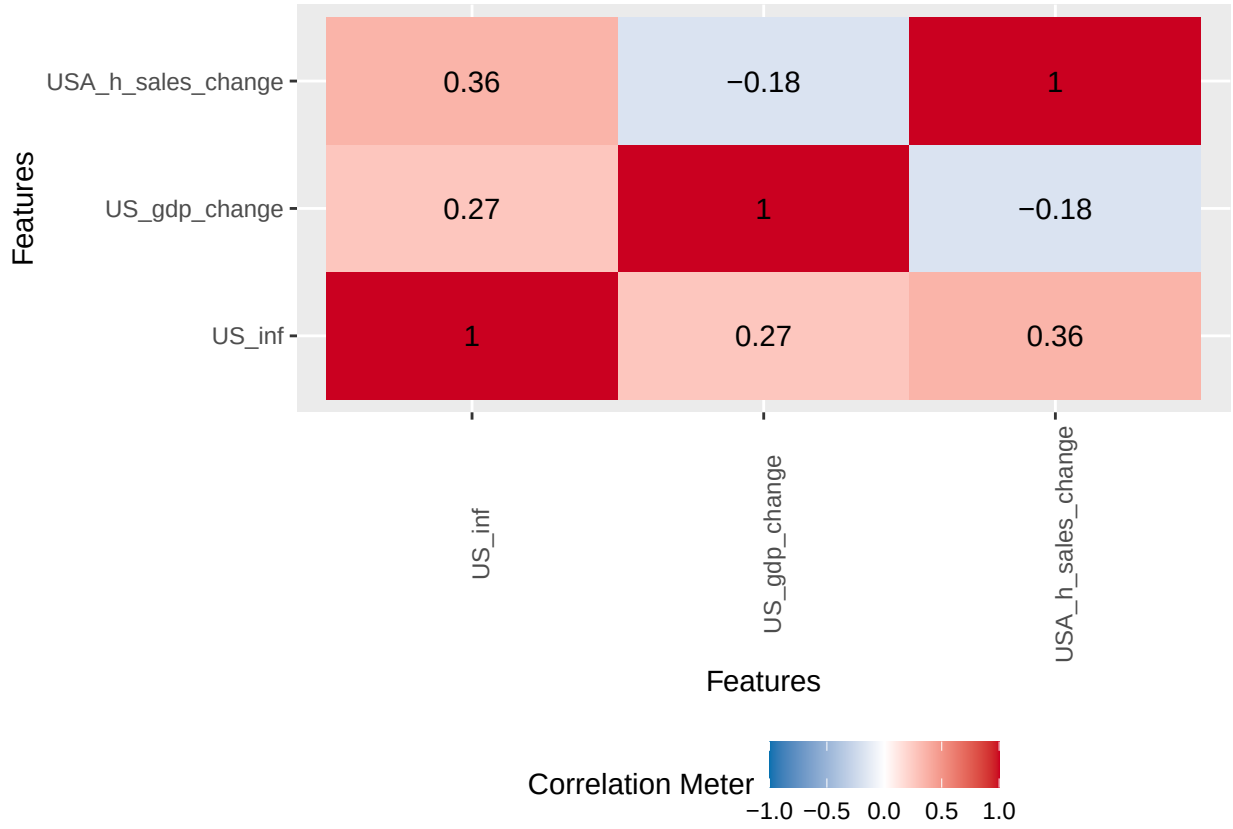
```
US_hardware_df <- US_hardware_df %>%
  mutate(USA_h_sales_change = (USA_h_sales / lag(USA_h_sales) - 1)) %>%
  mutate(US_cci_change = (US_cci / lag(US_cci) - 1)) %>%
  select(-US_gdp, -US_cpi, -USA_h_sales) %>%
  filter(Year >= 2008)
```

```
plot_correlation(US_hardware_df %>%
                   select(-Year))
```



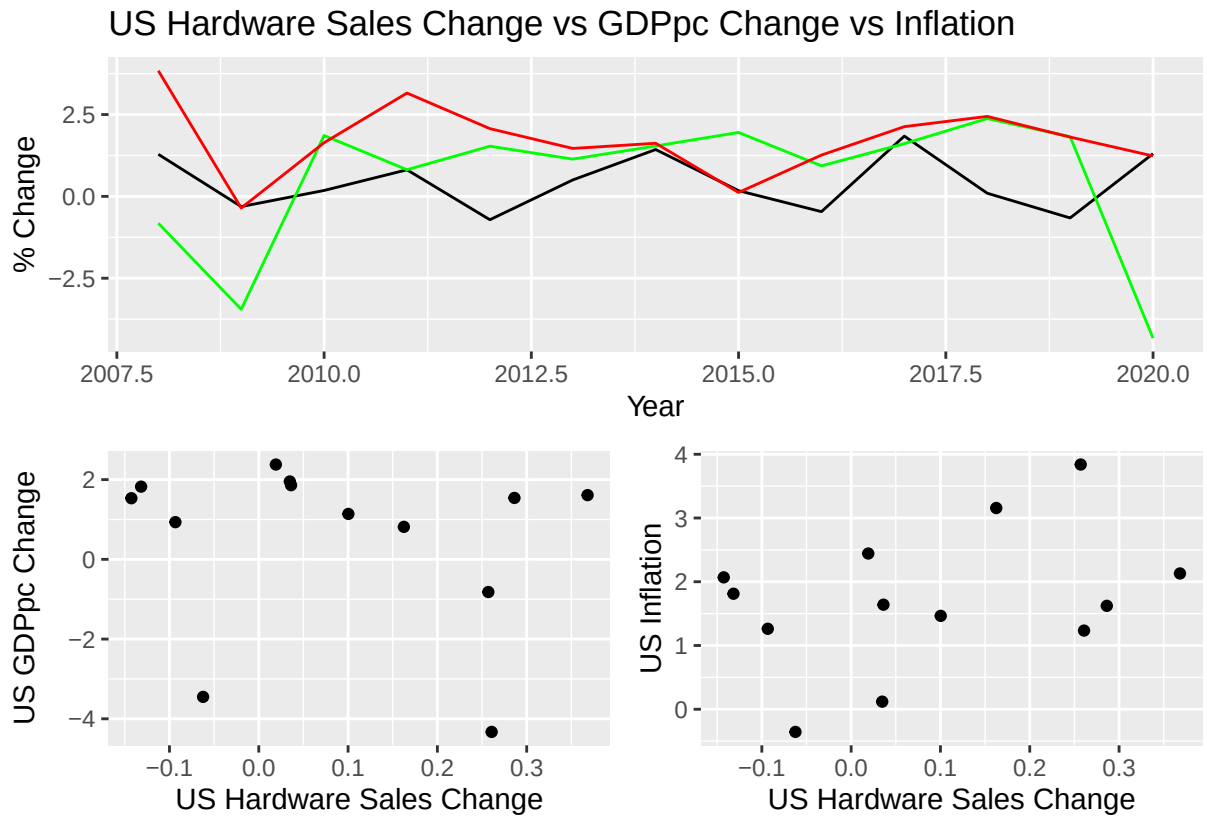
Multicollinearity exists between a few of these variables; we will discard the US lending rate, US Unemployment, US CCI, US CCI, and US CCI change.

```
US_hardware_df <- US_hardware_df %>%
  select(-US_rate, -US_unemp, -US_cci, -US_cci_change)
plot_correlation(US_hardware_df %>%
                   select(-Year))
```



We will further examine the relationship between sales and the indicators above- inflation and GDP change.

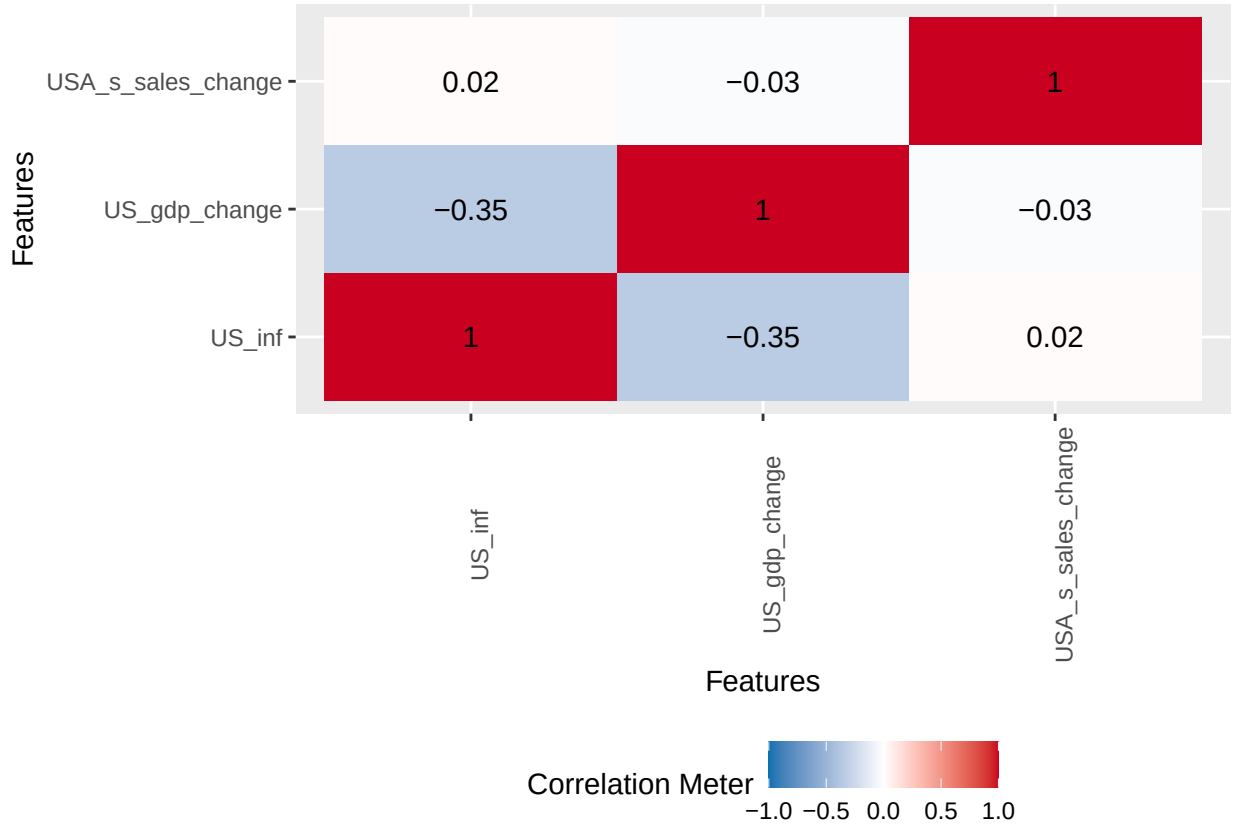
```
plot_1 <- ggplot(US_hardware_df) +
  geom_line(aes(x = Year, y = USA_h_sales_change*5), color = 'black') +
  geom_line(aes(x = Year, y = US_gdp_change), color = 'green') +
  geom_line(aes(x = Year, y = US_inf), color = 'red') +
  ylab('% Change') +
  labs(title = 'US Hardware Sales Change vs GDPpc Change vs Inflation')
plot_2 <- ggplot(US_hardware_df) +
  geom_point(aes(x = USA_h_sales_change, y = US_gdp_change), color = 'black') +
  ylab('US GDPpc Change') +
  xlab('US Hardware Sales Change')
plot_3 <- ggplot(US_hardware_df) +
  geom_point(aes(x = USA_h_sales_change, y = US_inf), color = 'black') +
  ylab('US Inflation') +
  xlab('US Hardware Sales Change')

plot_1 / (plot_2 + plot_3)
```



These relationships do not appear very correlated. We will examine other data for more insight.

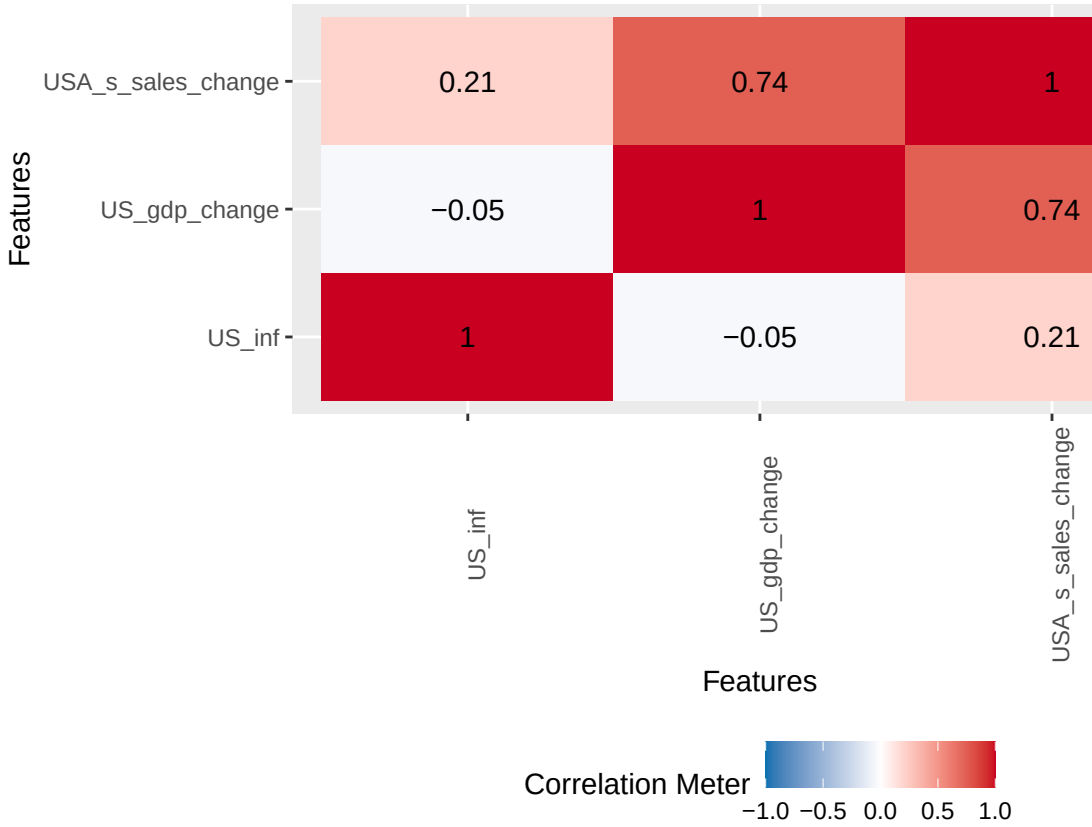
1995 - 2013 US Software We will begin this analysis by taking similar steps to mitigate dependence and multicollinearity in the data.

```
US_software_df_1 <- US_software_df %>%
  select(-US_gdp, -US_cpi) %>%
  mutate(USA_s_sales_change = (USA_s_sales / lag(USA_s_sales) - 1)) %>%
  mutate(US_cci_change = (US_cci / lag(US_cci) - 1)) %>%
  filter(Year >= 1995, Year <= 2008) %>%
  select(-USA_s_sales, -US_rate, -US_unemp, -US_cci, -US_cci_change)

plot_correlation(US_software_df_1 %>%
                   select(-Year))
```



There is virtually no correlation in this time period. This is possibly due to the rapidly increasing popularity of video games adding noise to the data. We will proceed to the next time period.

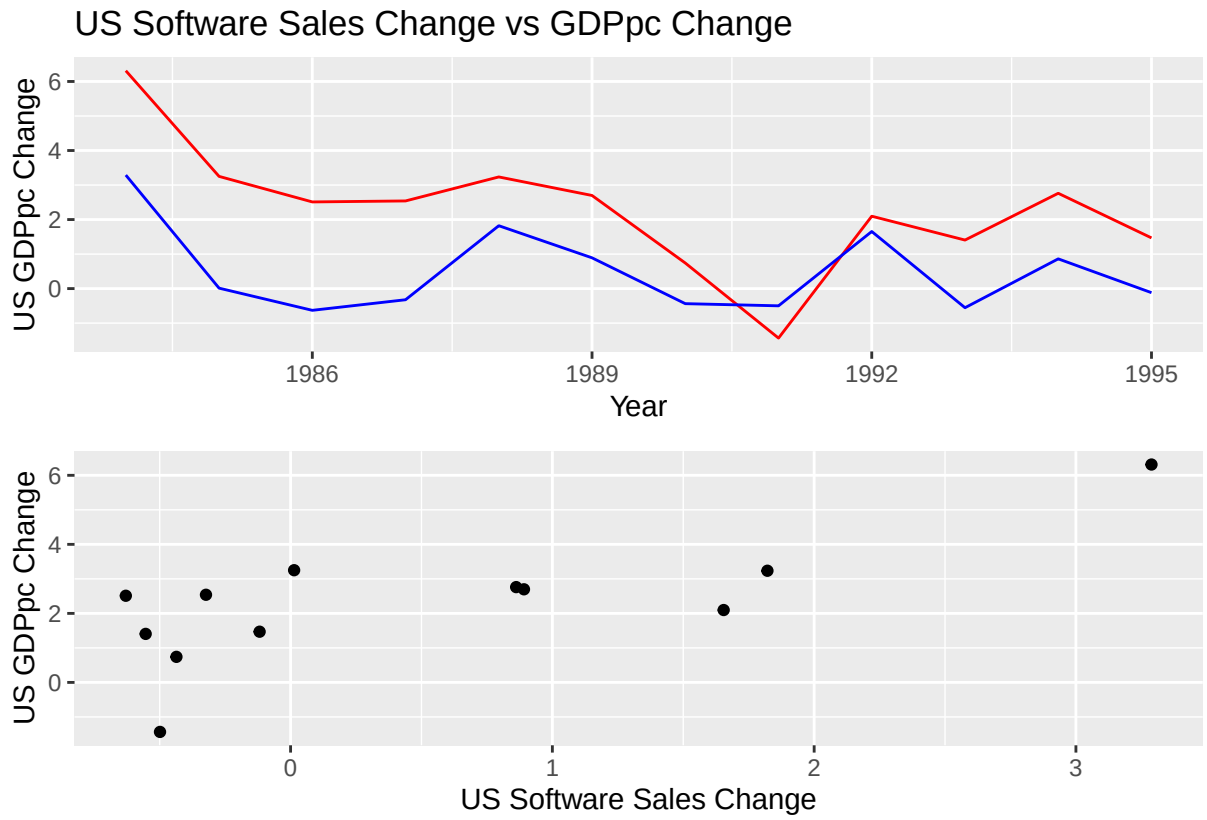
```
US_software_df_2 <- US_software_df %>%
  select(-US_gdp, -US_cpi) %>%
  mutate(USA_s_sales_change = (USA_s_sales / lag(USA_s_sales) - 1)) %>%
  mutate(US_cci_change = (US_cci / lag(US_cci) - 1)) %>%
  filter(Year >= 1984, Year <= 1995) %>%
  select(-USA_s_sales, -US_rate, -US_unemp, -US_cci, -US_cci_change)

plot_correlation(US_software_df_2 %>%
                   select(-Year))
```



1983 - 1995 US Software

There is a strong correlation between the change in US Software sales and the change in GDP per capita here. We'll examine this more closely to investigate the correlation.

```
plot_1 <- ggplot(data = US_software_df_2) +
  geom_line(aes(x = Year, y = US_gdp_change), color = 'red') +
  geom_line(aes(x = Year, y = USA_s_sales_change), color = 'blue') +
  ylab('US GDPpc Change') +
  labs(title = 'US Software Sales Change vs GDPpc Change')

plot_2 <- ggplot(data = US_software_df_2) +
  geom_point(aes(x = USA_s_sales_change, y = US_gdp_change), color = 'black') +
  ylab('US GDPpc Change') +
  xlab('US Software Sales Change')

plot_1 / plot_2
```




```
model <- lm(USA_s_sales_change ~ US_gdp_change, data = US_software_df_2)
summary(model)
```

```
##
## Call:
## lm(formula = USA_s_sales_change ~ US_gdp_change, data = US_software_df_2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.2339 -0.6866 -0.0089  0.7903  1.2584
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)    -0.6602     0.4145  -1.593  0.14231
## US_gdp_change   0.5037     0.1438   3.502  0.00571 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.866 on 10 degrees of freedom
## Multiple R-squared:  0.5508, Adjusted R-squared:  0.5059
## F-statistic: 12.26 on 1 and 10 DF,  p-value: 0.005709
```

This is the most compelling evidence for a real correlation between any economic indicator and video game sales. However, this data is 30-40 years old, so using this to generalize how video games react to economic indicators today would be irresponsible.

Key Findings After a thorough analysis of both US hardware and software sales, we do not have much evidence of economic indicators driving sales, or a change in sales, for video game hardware or software.

While it appears that video game software sales were synchronous with gdp per capita before 1995, this relationship seemingly disappears as video games surge in popularity. This could be the result of the population of video game owners changing dramatically as video games get into the hands of more people, or other variables that are more influential in determining video game sales.

With more data available, a more confident conclusion could be given about the relationship between software sales and hardware sales.