Is American Surplus Milk Produced Reaching Record High? And Why? Let the Data Speak.

William Banda

Github Link: GitHub Repository

```
## 00 Setup
# Get working directory
getwd() # Prints working directory in Console
```

```
[1] "C:/Users/WilliamBanda1/Documents/C7083"
```

```
## Set working directory
setwd("C:/Users/WilliamBanda1/Documents/C7083")

# 01 Import data
# Import the Milk Cow Facts Data
Data <- readr::read_csv("C:/Users/WilliamBanda1/Documents/C7083/milkcow_facts.csv")
```

```
Rows: 35 Columns: 11
-- Column specification --------------------------------------------------------
Delimiter: ","
dbl (11): year, avg_milk_cow_number, milk_per_cow, milk_production_lbs, avg_...

i Use `spec()` to retrieve the full column specification for this data.
i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
# Import the State_Milk_Production Data
Data2 <- readr::read_csv("C:/Users/WilliamBanda1/Documents/C7083/state_milk_production.csv")
```

```
Rows: 2400 Columns: 4
-- Column specification --------------------------------------------------------
```

```
Delimiter: ","
chr (2): region, state
dbl (2): year, milk_produced

i Use `spec()` to retrieve the full column specification for this data.
i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
# Import the Fluid_Milk_sales_Data
Data3 <- readr::read_csv("C:/Users/WilliamBanda1/Documents/C7083/fluid_milk_sales.csv")
```

```
Rows: 387 Columns: 3
-- Column specification --------------------------------------------------------
Delimiter: ","
chr (1): milk_type
dbl (2): year, pounds

i Use `spec()` to retrieve the full column specification for this data.
i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
# Import the Clean Cheese Data
Data4 <- readr::read_csv("C:/Users/WilliamBanda1/Documents/C7083/clean_cheese.csv")
```

```
Rows: 48 Columns: 17
-- Column specification --------------------------------------------------------
Delimiter: ","
dbl (17): Year, Cheddar, American Other, Mozzarella, Italian other, Swiss, B...

i Use `spec()` to retrieve the full column specification for this data.
i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
# Import the Milk_Products_Facts Data
Data5 <- readr::read_csv("C:/Users/WilliamBanda1/Documents/C7083/milk_products_facts.csv")
```

```
Rows: 43 Columns: 18
-- Column specification --------------------------------------------------------
Delimiter: ","
dbl (18): year, fluid_milk, fluid_yogurt, butter, cheese_american, cheese_ot...

i Use `spec()` to retrieve the full column specification for this data.
i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
# 02 Data Tidying
# Rename the variables in Clean Cheese Data
Data4 <- Data4 %>%
  rename(
    total_american_cheese = `Total American Chese`,
    total_italian_cheese = `Total Italian Cheese`,
    total_natural_cheese = `Total Natural Cheese`,
    total_processed_cheese_products = `Total Processed Cheese Products`
  )
```

According to Northeastern University political review, American farmers dump millions of gallons of excess milk a year. In 2016, the American dairy industry discarded 43 million gallons of excess milk into fields, animal feed, or anaerobic lagoons. In the short-sighted interest of keeping farms in business, the federal government purchases billions of dollars' worth of excess milk, which is stored as cheese. As of 2019, the USDA had 1.4 billion pounds of surplus cheese. The pace of milk production began to exceed the rates of consumption, says Andrew Novakovic, professor of agricultural economics at Cornell University.

Milk cow facts.

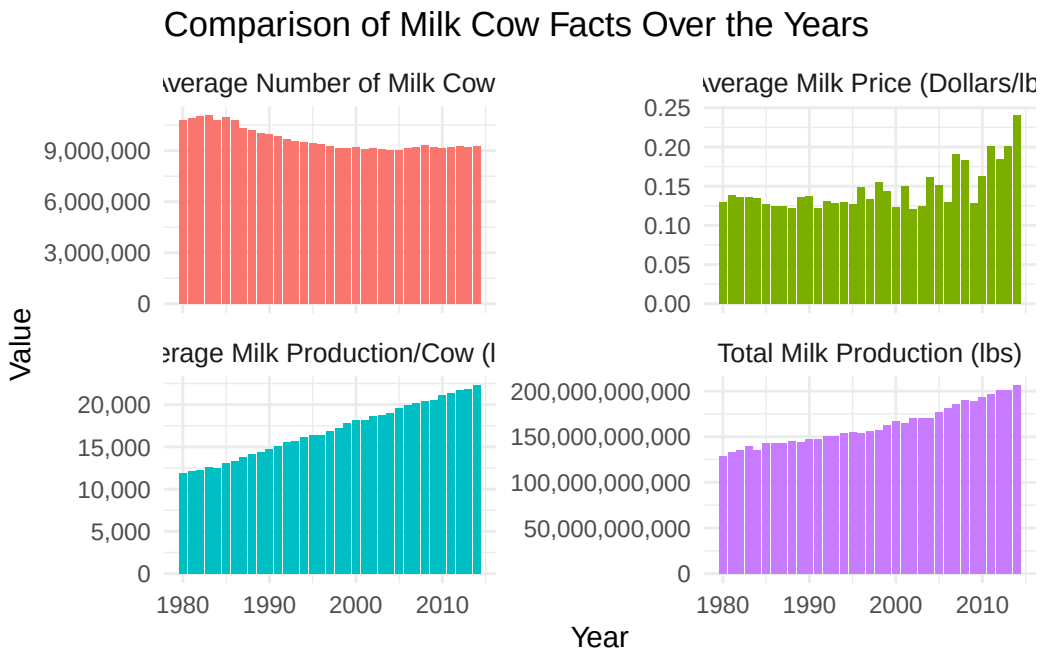
Is it a dodgy time for the American dairy industry? Well data speaks, let's take a deep dive to understand the trends in milk and milk related products and how they compare by looking at the figures from the USDA dairy data. There has been a notable decrease in the average number of cows in the US over the years, according to Farm Aid blog, a combination of poor industry regulation, bad policy, price fluctuations in the market, and a lower demand for dairy have all contributed to dairy's decline. The multi-panel bar plot below shows the trends of average number of milk cows, average milk price, average milk production and total milk production from 1980-2010.

```
# Visualise the milk cow facts data using a multi-panel bar plot
# firt select only the specified variables in the data set
Data_subset <- Data[, c("year", "avg_milk_cow_number", "milk_per_cow", "milk_production_lbs"

# Reshape the data from wide to long format
Data_long <- pivot_longer(Data_subset, cols = -year, names_to = "Variable", values_to = "Valu

# Create a multi panel bar plot using ggplot comparing each variable to year
ggplot(Data_long, aes(x = year, y = Value, fill = Variable)) +
  geom_bar(stat = "identity", position = "dodge") +
  facet_wrap(~Variable, scales = "free_y", labeller = labeller(Variable = c(
    year = "Year",
    avg_milk_cow_number = "Average Number of Milk Cows",
```

```
  milk_per_cow = "Average Milk Production/Cow (lbs)",
  milk_production_lbs = "Total Milk Production (lbs)",
  avg_price_milk = "Average Milk Price (Dollars/lb)"
)), ncol = 2) +
labs(title = "Comparison of Milk Cow Facts Over the Years",
     x = "Year",
     y = "Value") +
scale_y_continuous(labels = scales::comma) +  # Format y-axis labels as comma-separated
theme_minimal() +
theme(legend.position = "none",  # Remove legend
      strip.text = element_text(size = 10),  # Customize facet titles
      strip.background = element_blank())  # Remove facet title background
```



Source: Author's own

Despite a decline in the average number of milk cows, there has been an increase in average milk production per cow over the years. This can be attributed to improvements in animal welfare, which is linked to improvements in technology, among other factors. The increase in milk production per cow is one factor that has led to an increase in total milk production in the US. Despite the increase in production there is a notable increase in milk prices in the US. According to National Public Radio (NPR), over the past 10 years, milk production has increased by 13 percent because of high prices.

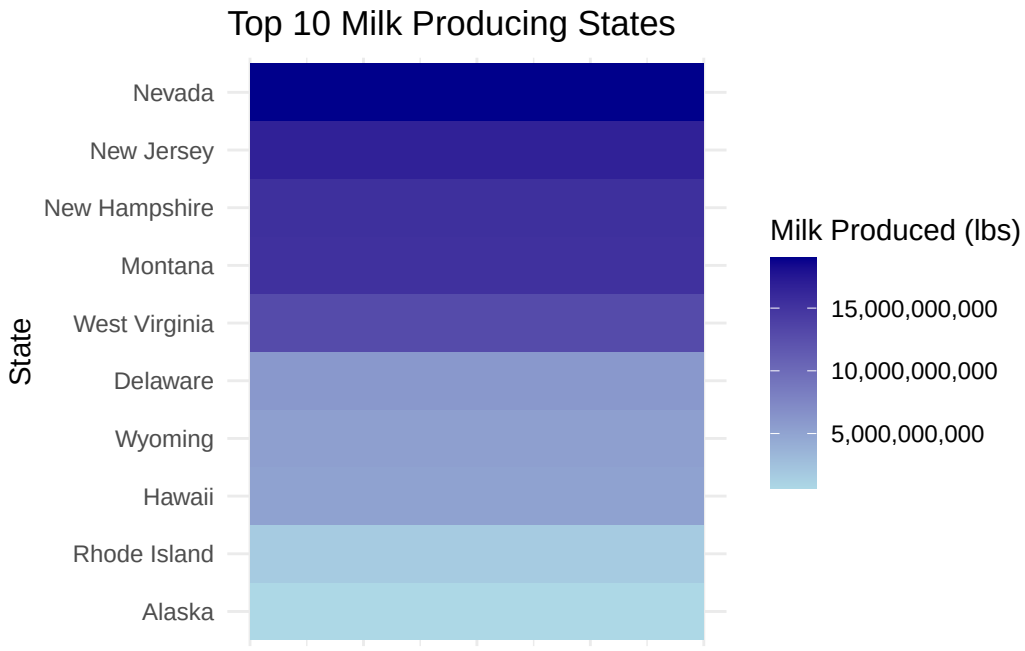
Fluid milk production

As we look at the variability of factors influencing milk production in the United States, we may also note that milk production levels are not the same across different states, this is attributed to many factors, for instance some states are more agricultural oriented than others when it comes to economic activity. Looking at the heatmap below we can see the top ten milk producing states over the years 1975-2017. Nevada tops the list while Alaska records the lowest production in comparison to the other states being in the 5 billion pounds category. The highest milk producing states fall in the 15 billion pounds category which is a huge amount in comparison. But are dairy farmers failing to realize that Americans are drinking less milk?

```
# Now lets plot a heat map of the top ten producing milk states in the US
# Calculate the total milk production by state
milk_by_state <- aggregate(milk_produced ~ state, Data2, sum)

# Sort the states by milk production in ascending order and select the top 10
top_states <- head(arrange(milk_by_state, milk_produced), 10)


# Create the heatmap
ggplot(top_states, aes(x = 1, y = reorder(state, milk_produced), fill = milk_produced)) +
  geom_tile() +
  scale_fill_gradient(low = "lightblue", high = "darkblue", labels = scales::comma) +
  labs(title = "Top 10 Milk Producing States",
       x = NULL,
       y = "State",
       fill = "Milk Produced (lbs)") +
  theme_minimal() +
  theme(axis.text.x = element_blank(),
        axis.ticks.x = element_blank())
```



Source: Author's own

Fluid milk sales

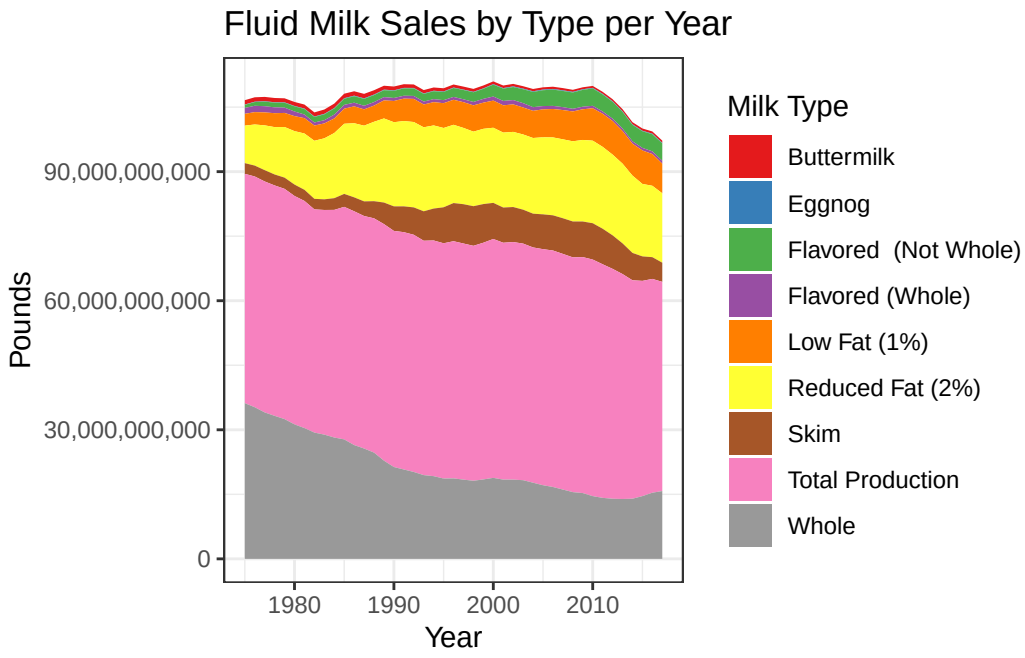
Another important aspect linked to milk production is amount of sales for a product, this determines whether there will be surplus or not in terms of stocks. Sales are influenced by demand. Just like in all other parts of the world Milk is sold in various forms in the US. Among the reasons why different forms of milk sell less than other forms includes industrial uses and dietary restrictions. The area chart below shows fluid milk sales by type per year compared to the total milk production. The stacked area chart clearly shows a decline in sales for all milk types over the years.

```
# Create an area chart of the Fluid Milk Sales Data

# Assigning the dataframe 'Data3' to a variable named 'df'
df <- Data3

# Create the area chart
ggplot(df, aes(x = year, y = pounds, fill = milk_type)) +
  geom_area(stat = "identity") +  # Use "identity" for stacked area
  labs(title = "Fluid Milk Sales by Type per Year",
       x = "Year", y = "Pounds", fill = "Milk Type") +  # Restore y-axis label and rename leg
  scale_fill_brewer(palette = "Set1") + # Use the same palette for fill and color
```

```
  theme_bw() +
  scale_y_continuous(labels = scales::comma_format(), name = "Pounds")  # Format y-axis with
```



Source: Author's own

Notably over the years whole milk sold more compared to other types of milk and Reduced fat milk was second in sales while butter milk recorded the lowest of sales. If dairy farmers are realizing that milk sales are going down in the US, could it be that they are resolving to look at other ways of utilizing excess milk stocks?

Cheese Production

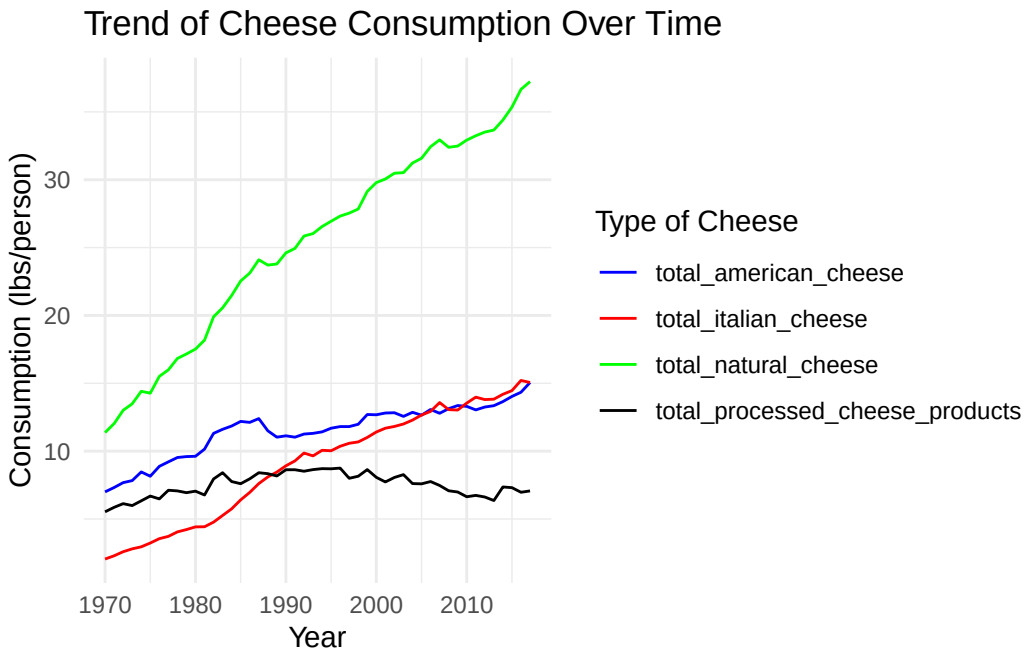
National Public Radio (NPR) has reported that suppliers turn that extra milk into cheese because it is less perishable and stays fresh for longer periods. However, Americans are increasingly shunning processed cheese slices, NPR recorded. The consumption of Natural cheese has proved to be the largest of consumed cheese in the U.S. Looking at the line plot below, American cheese remains strong in the market, but far less compared to natural cheese, while Italian cheese consumption seemed to have increased as well from being the lowest to almost equalling American cheese. Total processed cheese product consumption is seemingly declining.

```r
# Lets Visualise the clean cheese data using a line plot to observe the trends
# Filter the dataset to include only the specified variables
Data4_filtered <- Data4 %>%
  select(Year, total_american_cheese, total_italian_cheese,
         total_natural_cheese, total_processed_cheese_products)

# Melt the filtered data into long format
Data4_long <- pivot_longer(Data4_filtered, cols = -Year, names_to = "Variable", values_to = "


# Define custom colors for each line
custom_colors <- c("total_american_cheese" = "blue",
                   "total_italian_cheese" = "red",
                   "total_natural_cheese" = "green",
                   "total_processed_cheese_products" = "black")

# Create a line plot with custom colors
ggplot(Data4_long, aes(x = Year, y = Value, color = Variable)) +
  geom_line() +
  scale_color_manual(values = custom_colors) +  # Apply custom colors
  labs(title = "Trend of Cheese Consumption Over Time",
       x = "Year",
       y = "Consumption (lbs/person)",
       color = "Type of Cheese") +
  theme_minimal()
```



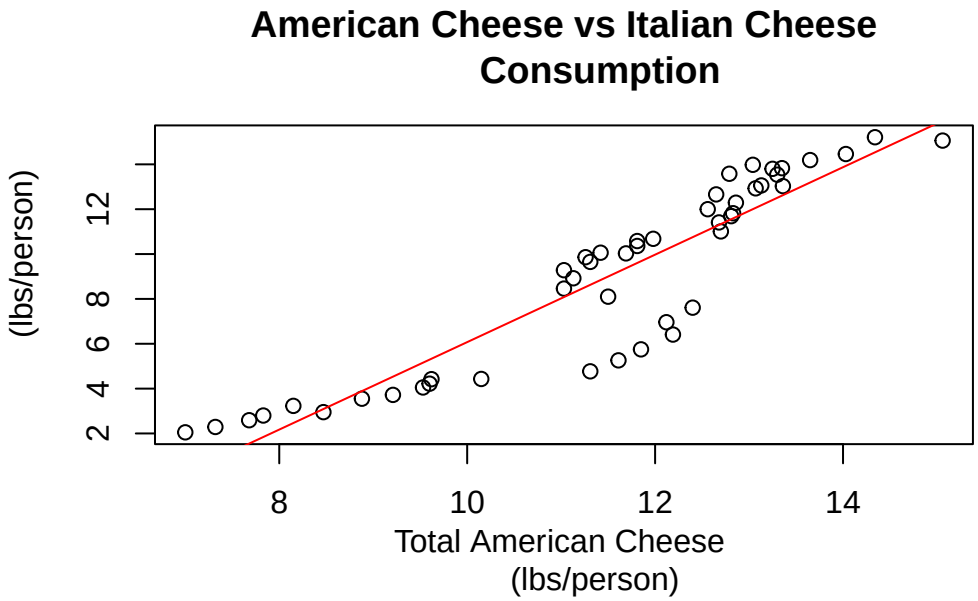
Source: Author's own

Observing the trends further by comparing total American cheese consumption and total Italian Cheese consumption using the scatter plot below, one would be interested to understand, whether there is a relationship between how these two products are being consumed. Looking at the scatter plot below one can conclude that there is no negative relationship between consumption of the two products since as the consumption of one product increases the consumption of the other seems to increase as well. However, one cannot conclude the causation of increase in consumption of one product by the other unless further analysis is conducted. But we must ask, is cheese the only product of interest in regard to milk production in the US?

```
# Lets create a scatter plot using base R to compare consumption of 2 type of Cheese
 plot(Data4$total_american_cheese,                          # Specify data
      Data4$total_italian_cheese,                               # Specify data
      main = "American Cheese vs Italian Cheese
      Consumption", # Add title
      xlab = "Total American Cheese
      (lbs/person)",                        # Add x-axis label
      ylab = "Total Italian Cheese
      (lbs/person)")                              # Add y-axis label

 # Fit linear regression model
 lm_model <- lm(total_italian_cheese ~ total_american_cheese, data = Data4)
```

```r
# Add regression line
abline(lm_model, col = "red")
```



Source: Author's own

Milk products facts.

Other exciting and very important dairy products to look at are Butter, Fluid Yoghurt, Dry Whey and Frozen Ice cream. Let's dive deep to see how these products are being used in the US in relation to milk production using the interactive plot below. Fluid Yoghurt has proved to be the most consumed product compared to Butter, Dry Whey and Frozen Ice Cream (Regular). The second most consumed product among the four products is ice cream. Butter is third on the list but much lower than Fluid yoghurt and Ice cream, this is apparently because butter is consumed with other foods and mostly used as an ingredient not consumed on its own. The lowest consumed product is dry whey. But the most important to note on these trends is the recent decline in the consumption of Dry Whey, Fluid Yogurt, Frozen Ice Cream (Regular) while Butter seems to remain constant in its consumption.

```r
# Create an interactive plot of Dairy Products Consumption Over Time
# Filter by specific years (optional)
Data5_filtered <- Data5 %>%
  filter(year >= 2010 & year <= 2020)  # Example filter
```

```
    # Create the line chart
plot_ly(Data5_filtered, x = ~year) %>%
  # Add product lines
  add_lines(x = ~year, y = ~butter, color = "Butter", name = "Butter") %>%
  add_lines(x = ~year, y = ~dry_whey, color = "Dry Whey", name = "Dry Whey") %>%
  add_lines(x = ~year, y = ~fluid_yogurt, color = "Fluid Yogurt", name = "Fluid Yogurt") %>%
  add_lines(x = ~year, y = ~frozen_ice_cream_regular, color = "Frozen Ice Cream (Regular)", n
  # Add labels on lines
  add_text(x = ~year[10], y = ~butter[10], text = "Butter", showlegend = FALSE, textposition
  add_text(x = ~year[20], y = ~dry_whey[20], text = "Dry Whey", showlegend = FALSE, textposit
  add_text(x = ~year[30], y = ~fluid_yogurt[30], text = "Fluid Yogurt", showlegend = FALSE, t
  add_text(x = ~year[40], y = ~frozen_ice_cream_regular[40], text = "Frozen Ice Cream (Regula
  # Customize plot labels and legend
  layout(title = "Dairy Products Consumption Over Time",
         yaxis = list(title = "Consumption (lbs/person)"),
         legend = list(title = list(text = "Product")))
```

Analysts have also expressed concerns regarding the potential adverse effects of U.S. trade tensions with both China and Mexico on the dairy sector. However, the U.S. Dairy Export Council suggests that the repercussions of retaliatory tariffs on American dairy goods have been relatively minor. Moreover, it's worth noting that the U.S. exports only around 6 percent of its cheese. The US government and related stakeholders might have to look for permanent solutions in regard to this excess milk that's being wasted since milk production contributes a significant amount to the country's economy and can be used better!

Good Graph and Bad Graph Examples

1. Good Graph

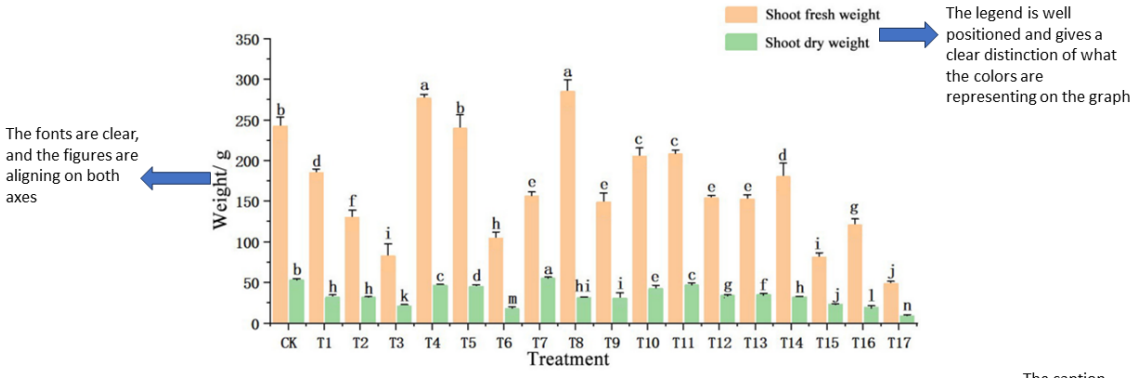
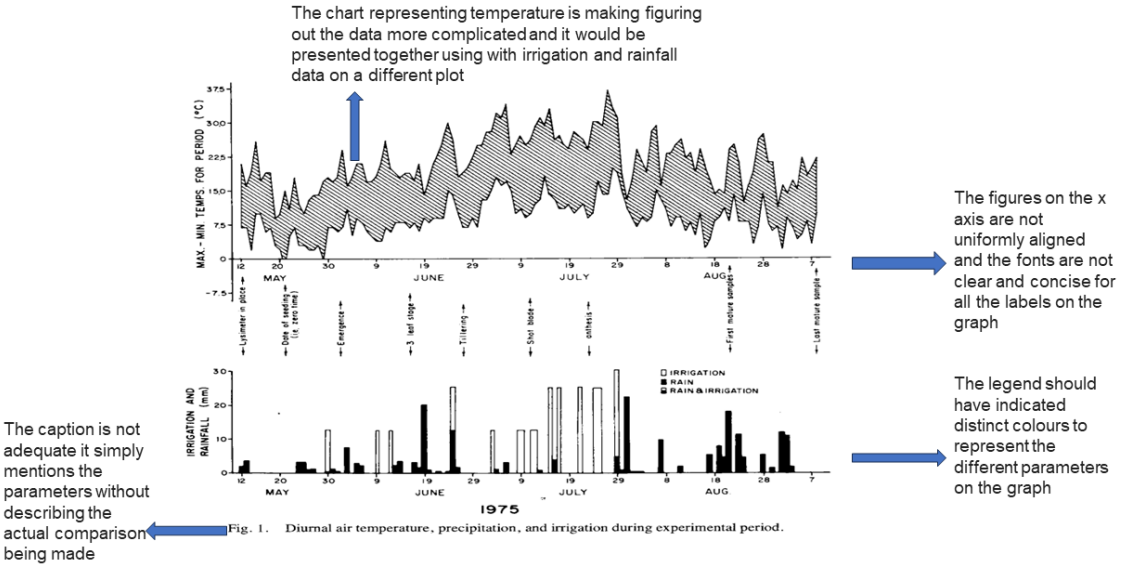


Figure 2. Effect of substitute substrates on the biomass of potted chrysanthemums. The vertical coordinate represents the weight (g). The horizontal coordinates represent the 18 treatment groups CK, T1–T17. Data are mean ± standard error ($n = 3$). Different lowercase letters indicate significant differences ($p < 0.05$), as determined by Duncan's multiple range test.

The caption gives a clear picture of the comparison being made

The above bar plot that compares the effect of substitute substrates on the biomass of potted chrysanthemums is a good example of visualising data in that it is letting the data to speak, this bar plot makes it easy to decode numerical information like weight/g. The plot avoids unnecessary thrills and it's accurate. In the plot the data stands out while using visually distinct symbols i.e., colour for the two different groups which are brown and green for shoot fresh weight and shoot dry weight respectively. The bar plot avoids unnecessary chart junk like heavy grid lines that might distract the audience from the data. The bar plot facilitates comparison by emphasising on the important comparison visually which is shoot weights either dry or fresh per treatment. The graph has well aligned figures, has a clear and concise font for all the labels. The graph has a well-positioned legend that gives a key to what each colour on the bars represents and it doesn't need a second thought to obtain meaning. The x and y axes are well labelled and that gives a clear and straight forward picture of what each of the parameters mean i.e. easy to tell that the vertical coordinates represent the weight in grams and the horizontal coordinates represent the 18 treatments groups. The point is conveyed into context by using an adequate caption and appropriate annotations.

2. Bad Graph



The data visualisation above that shows diurnal air temperature, precipitation, and irrigation during experimental period is a bad example of visualising data in that it is not letting the data to speak loudly, the figure makes it hard to decode numerical information like maximum to minimum temperatures for the period. The plot has a lot of unnecessary thrills and it's not accurate. Data is not standing out and visually the symbols like colour are not distinct. The chart representing temperature is making figuring out the data more complicated and it would be presented together using with irrigation and rainfall data on a different plot. It's hard to tell why the variables are being plotted by just looking at the graph. The figures on the x axis are not uniformly aligned and the fonts are not clear and concise for all the labels. Distinct colours should have used to represent the rainfall, irrigation and rainfall and irrigation parameters i.e. the legend should have indicated those colours. The point of the whole plot is not conveyed into context since the caption is not adequate it simply mentions the parameters without describing the actual comparison being made.

References

Link 1: US Dairy Exporter

Link 2: Farm Aid

Link 3: NPR

Link 4: NU Political Review

Link 5: Article on MDPI

Link 6: Another Article on CDN Science Publishing