C7083 Assignment: Data Visualisation and Analytics

A T Chamberlain Student ID 22394500

Data Visualisation Portfolio and Critique

Background

The data sets supplied refer to milk production and consumption in the USA over the past 20 - 30 years. Background press articles have been supplied at:

https://www.npr.org/2019/01/09/683339929/nobody-is-moving-our-cheese-american-surplus-reaches-record-high

and

https://www.washingtonpost.com/news/wonk/wp/2018/06/28/americas-cheese-stockpile-just-hit-an-all-time-high/?noredirect=on

Observations

- The theme of both articles was that there is a growing milk/cheese surplus with the Washington Post reported the stockpile had grown from 481 million in 1993 to 1390 billion pounds in 2018. They considered this was due to increasing production and a falling consumer demand for milk and cheese. Milk does not store well so any surplus is made into cheese for storage and export. About 14% of national milk production is exported but trade issues with Mexico, China and elsewhere may limit future export potential. As stockpiles grow and stores fill there is pressure not to make cheese and so cheese and thus milk prices fall.

Objectives

This report will use a range of graphical techniques from base-R and ggplot2 to create static and dynamic graphs and charts to investigate these data sources. Some of the dynamic charts will be rendered in GIF files that can be found on the GitHub repo and a Flexdashboard will be presented in a seperate R file for running.

Data

The databases on the TidyTuesday source are historic and only run up to 2017. The GitHub readme document gives the source of the data (USDA databases) and, where possible, the original data has been updated to include more recent years. The USDA datasets were available as Excel files but these were heavily formatted with column merges, repeated rows etc. Excel was used to convert the data into Tidy format with a Dictionary. The Excel and associated csv files are stored on the GitHub repo.

Data wrangling

Address the three databases in turn.

- Milkcow_facts : Express milk production in millions of pounds, cow numbers in thousands
- Milk_prod : No changes needed
- State_milk_production :Expressed milk production per state in millions of pounds

Where data expressed in very big numbers these are adjusted. Several variable names contained the '$' symbol and these were changed to 'D'

Conversion of units of measurement.

Americans measure milk production in pounds (lbs) - all other countries use kg or litres (which are very similar for milk). This is a hurdle for non-Americans and results in much mental arithmetic and introduction of error. As this report is for a non-USA audience all pounds will be converted to kilograms and the variables re-named. Prices are all referenced in dollars but as the dollar/sterling pound conversion is not constant this will not be converted.

Data dictionaries

Final dataset structure after wrangling to change units, remove dollar signs, adjust scale of units and add mutated columns.

Milk cow facts dictionary

Table 1. Milking cow numbers, yields an costs of inputs and outputs by year (35 rows).

Column number	Name	Type
1	year	num
2	avg_milk_cow_number_thousands	num
3	milk_per_cow_kg	num
4	milk_production_billion_kg	num
5	avg_price_milk_D_per_kg	num
6	dairy_ration_price_D_per_kg	num
7	milk_feed_price_ratio	num
8	milk_cow_cost_per_animal	num
9	milk_volume_to_buy_cow_in_kgs	num
10	alfalfa_hay_price_D/kg	num
11	slaughter_cow_price_D_per_kg_liveweight	num
12	milk_billions_kg_delta	num
13	lagged_milk_price	num

Milk production by state dictionary

Table 2. Amounts of milk produced by year and by state (2600 rows).

Column number	Name	Type
1	region	chr
2	state	chr
3	year	num
4	milk_produced_millions_kg	num
5	Million Tonnes	num

Milk production dictionary

Table 3. Milk production details (cow numbers and yields) by year and state (52 observations)

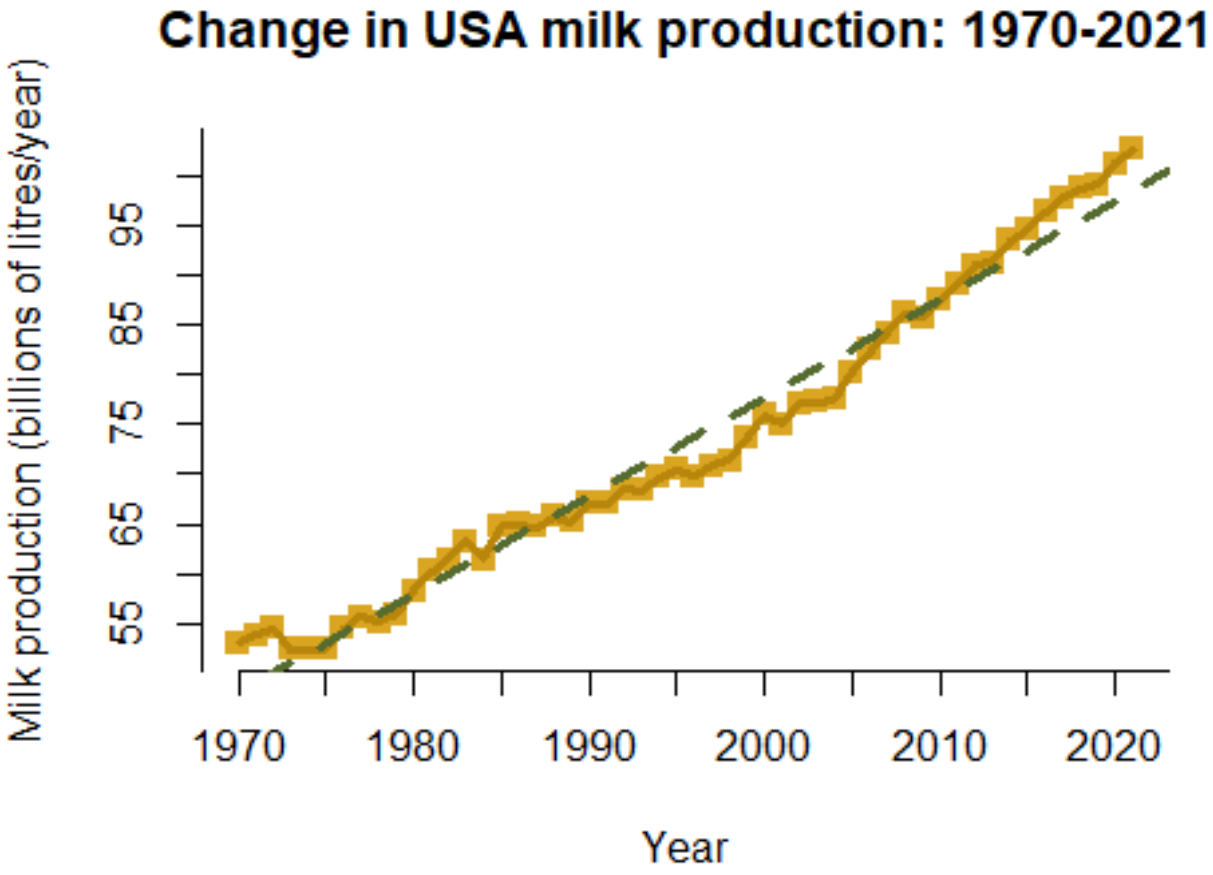
Column number	Name	Type
1	Year	num
2	Northeast_cows	num
3	Lake_States_cows	num
4	Corn_Belt_cows	num
5	Northern_Plains_cows	num
6	Appalachia_cows	num
7	Southeast_cows	num
8	Delta_States_cows	num
9	Southern_Plains_cows	num
10	Mountain_cows	num
11	West_Coast_cows	num
12	Other_States_cows	num
13	United_States_cows	num
14	Northeast_yield_kg_per_cow	num
15	Lake_States_yield_kg_per_cow	num
16	Corn_Belt_yield_kg_per_cow	num
17	Northern_Plains_yield_kg_per_cow	num
18	Appalachia_yield_kg_per_cow	num
19	Southeast_yield_kg_per_cow	num
20	Delta_States_yield_kg_per_cow	num
21	Southern_Plains_yield_kg_per_cow	num
22	Mountain_yield_kg_per_cow	num
23	West_Coast_yield_kg_per_cow	num
24	Other_States_yield_kg_per_cow	num
25	United_States_yield_kg_per_cow	num
26	Northeast_million_kg_produced	num
27	Lake_States_million_kg_produced	num
28	Corn_Belt_million_kg_produced	num
29	Northern_Plains_million_kg_produced	num
30	Appalachia_million_kg_produced	num
31	Southeast_million_kg_produced	num
32	Delta_States_million_kg_produced	num
33	Southern_Plains_million_kg_produced	num
34	Mountain_million_kg_produced	num
35	West_Coast_million_kg_produced	num
36	Other_States_million_kg_produced	num
37	United_States_million_kg_produced	num

Data processed to take into account the above changes and now ready and can produce graphics as required.

Change in national milk production volumes over time.

How has country-wide milk production varied over the years? will use the data in the milkcow_prod data-frame to plot total milk production by year. THis will be acheived using base R with a scatter and line plot

Figure 1. National milk production (billions litres per year) by calendar year over the past 50 years.



The simple linear plot with fitted line shows that milk production generally increased throughout the period from 1970 to 2020 and does not seem to dip very much at any time though milk prices will vary. There are two possible factors involved here:

- The costs of milk production can be considered as split into fixed and variable costs. Fixed costs (eg labour) do not vary with small changes in output, variable costs (eg feed) do vary. In dairy enterprises the split between fixed and variable costs is about 50:50. If milk prices fall slightly then producers will increase output as long as it does not affect fixed costs so that they dilute the fixed costs per liter. If milk prices rise many producers will increase herd fixed costs (labour, buildings) to increase income.

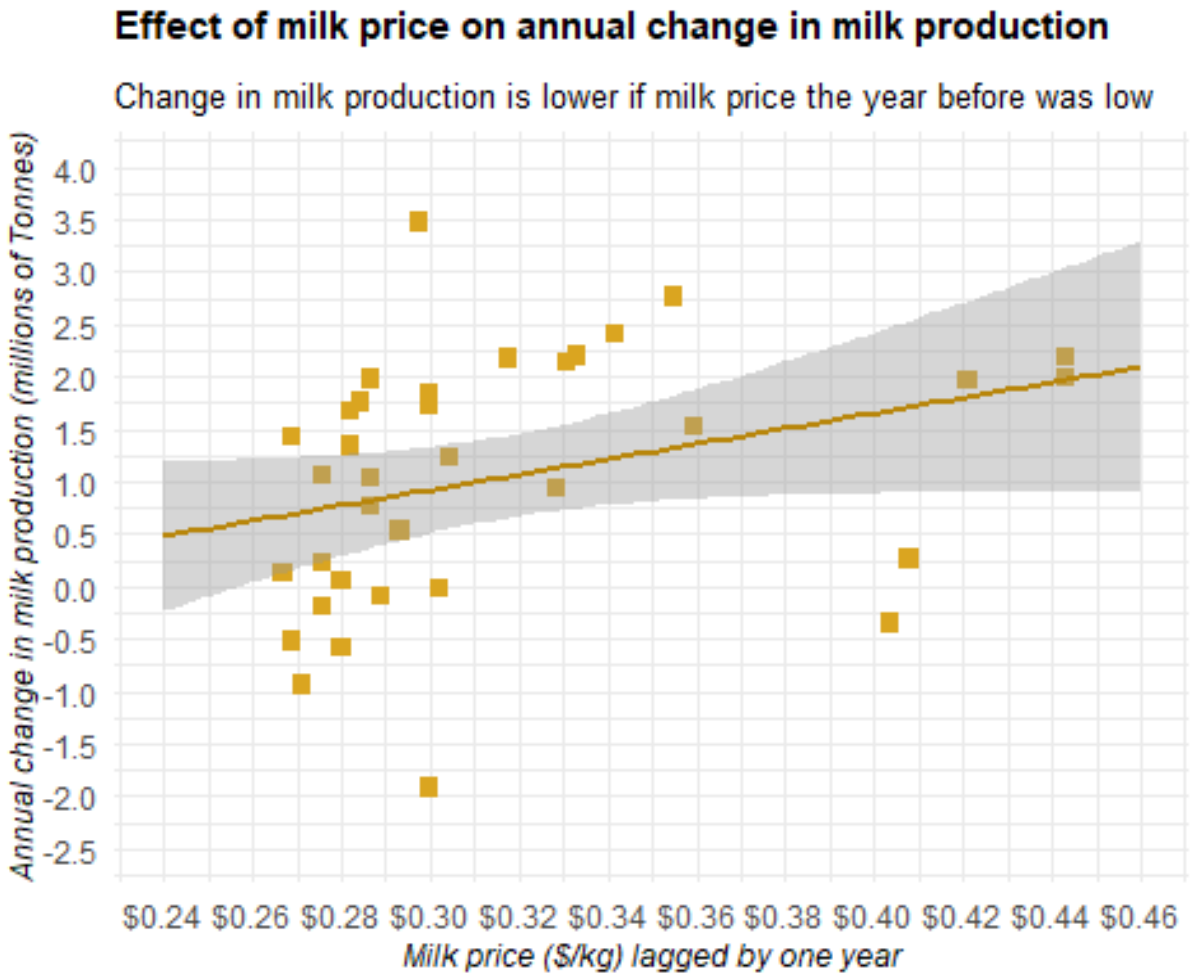
- The USA dairy sector benefits from a raft of subsidies and grants. If milk price falls these tend to pay out more money to stabilise the market and limit down falls.

It is interesting to note that milk production in the USA is a free market with no controls - producers just keep on increasing output and rely on the market to take everything produced. The EU and (more importantly Canada) have operated a controlled market with centrally enforced limits on production with fines, etc. for over production.

Is milk production affected by milk price - possibly with a lag between change in price and change in output?

It may be that the rate of increase in production is affected by milk price - ie when prices are low then the rate of increase is lower. This will be investigated by comparing the annual change in overall milk production with the milk price. The hypothesis being assessed is that milk price in the previous year affects the rate of change in annual milk production. So if milk price is low in year n-1 then the milk production increase in year n will be low.

Figure 2. Relationship between lagged milk price ($/kg) and the annual change in milk production (millions Tonnes/year).

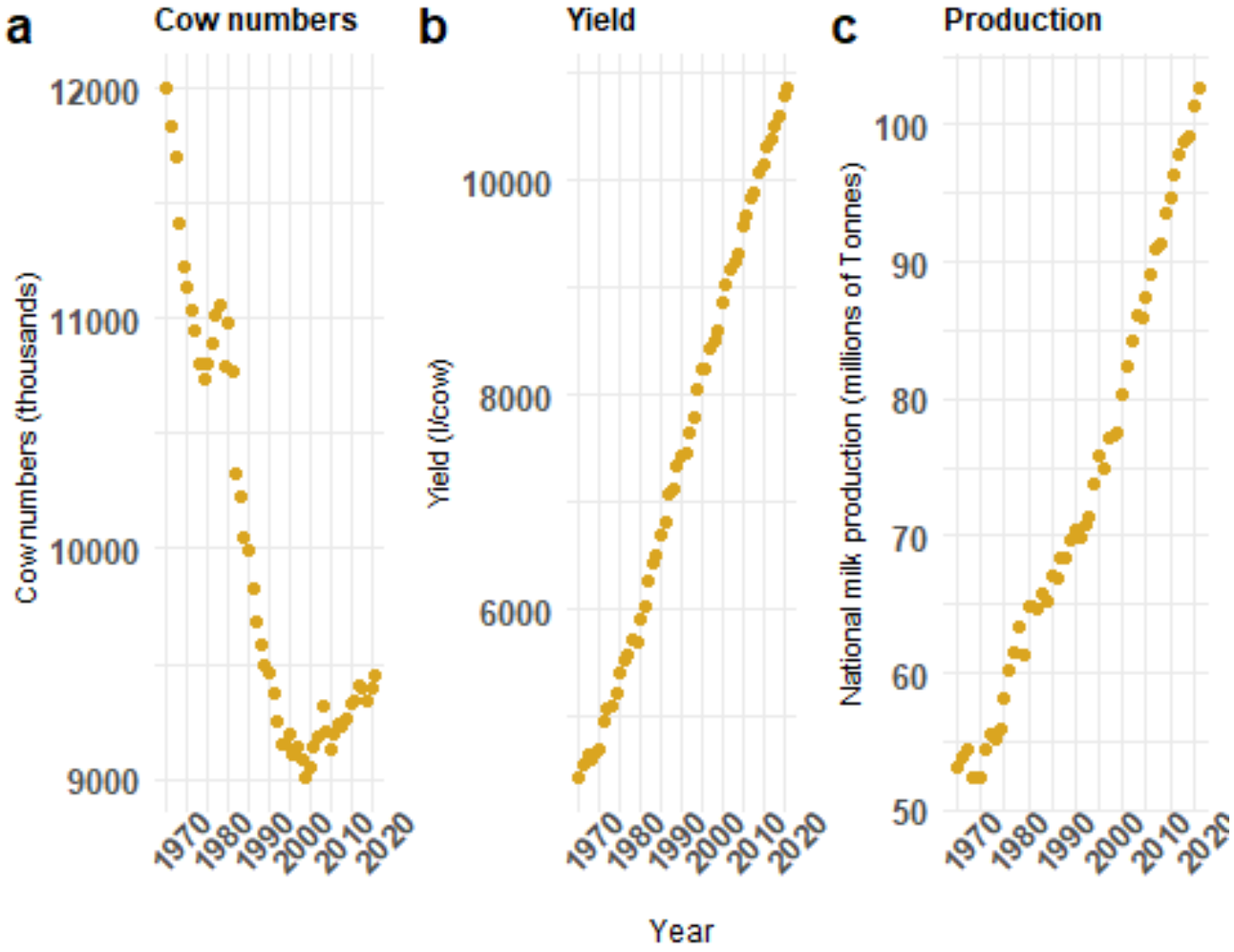


The change in milk output increases as the lagged (1 year) milk price increases. Whilst the regression line has a positive gradient the slope is not significantly different to zero. There are several notable outliers which may be linked to global events such as the 2008 recession, etc.

What factors are affecting the change in milk price?

Why did milk output increase - was it because cow numbers increased or because yield per cow increased. Will look at the country wide data and plot changes over time.

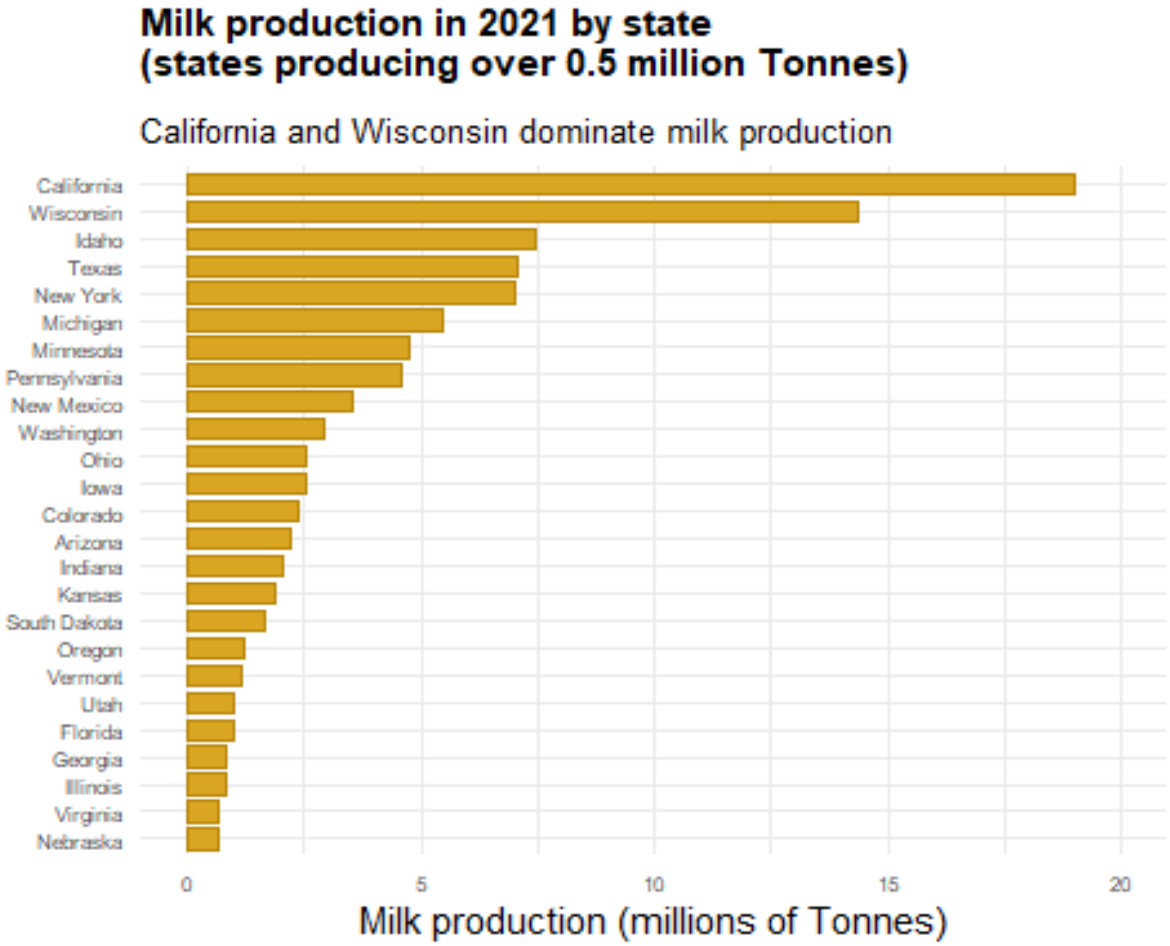
Figure 3. Changes in national cow numbers, milk yield (l/cow/year) and national milk production (millions Tonnes) over time for period 1970 to 2021.



Cow numbers dropped from 1970 to 2000 but are now increasing again. These changes have not affected milk yield per cow which has increased steadily from just over 4000 l/year to almost 11000; in contrast the current average in UK is around 9000 l/cow. The rise in yield per cow has over-ridden the change in cow numbers and milk production has just about doubled in the last 50 years.

Which states produce the most milk in 2021?

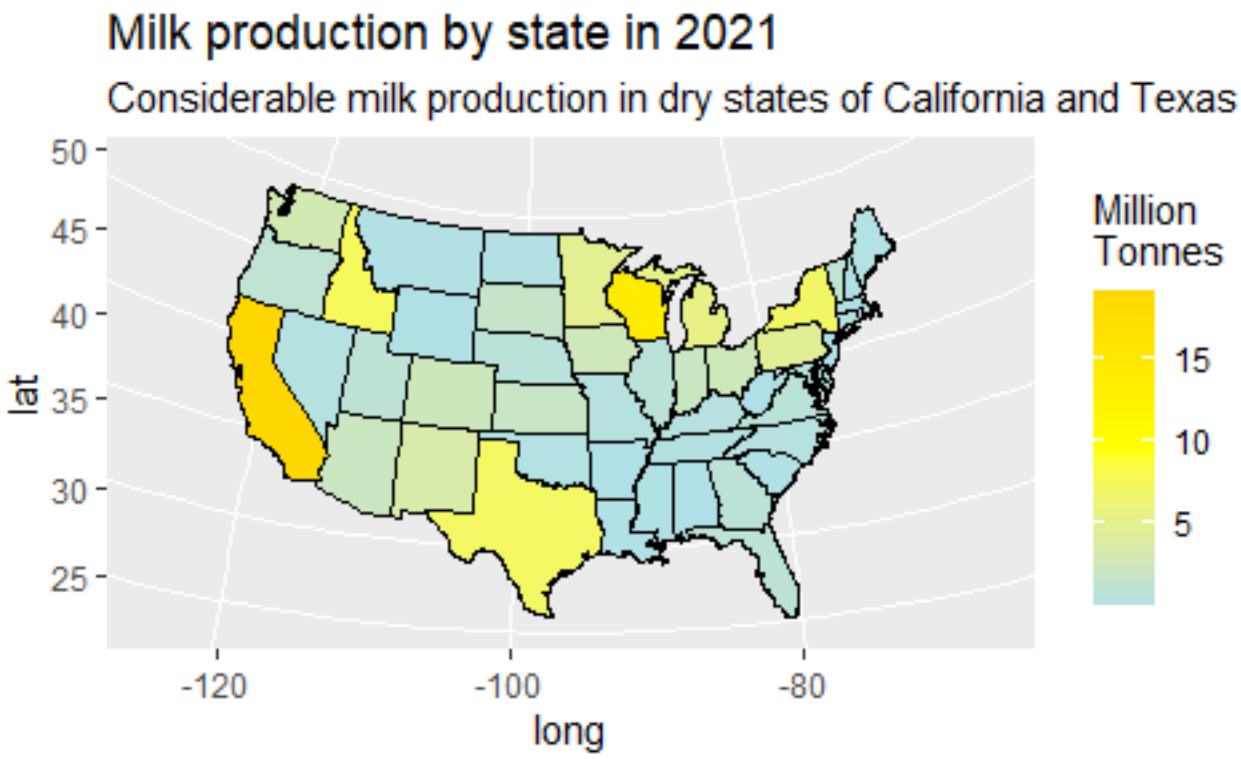
Figure 4. Milk production in major producing states (over 0.5 million Tonnes/year) in 2021



California dominated milk production in 2021 - despite their long term drought and dependence on irrigation. Wisconsin is second and has a long history of milk production. In 2021 Idaho was third with Texas just behind - social media reports that Texas is now in third place in 2023 despite its dry climate. In 2022 Texas had 8.2 inches (205 mm) of rain (the lowest ever recorded). By contrast the Midlands had 810 mm. Milk production uses a lot of water to support the cows and to maintain hygenic production. One has to question if dairy farming in Texas is sustainable, if ground water irrigation is being used how long will this last?

Overview of how milk industry statistics have changed over time.

Figure 5. Milk production in USA continental states in 2021



Changes in production and the costs and income associated with milk production.

A Flexdashboard was created to show how various parameters governing the dairy industry have changed from 1980 to 2014. This allows changes to be plotted over time, and for inter-relationships to be inspected. The R markdown file called flexdashboard_v01_20Feb23_1200Rmd can be found on the GitHub repo.

Figure 5. Dashboard showing how key parameters have changed over time (1980 to 2014) and how they inter-relate)

Run program flexdashboard_v01_20Feb23_1200Rmd from repo in R

Changes in patterns of milk production over time (1970 - 2021)

The final graphic explores how milk production has moved between different states in continental USA in the past five decades. The first version of the chart colour codes the states according to amount of milk produced.The GIF file called milkByState06.gif can be found on the repo.

Figure 6. Dynamic map showing how milk production has varied by state from 2070 to 2021

See file called milkByState06.gif on repo

In this plot how overall production is increasing is not shown well - and it doubles in this time period (see above). A further plot adds bubbles to the plot so that can see overall size of production increases and is stored on the Repo as milkByState_dots_02.gif .

Figure 6a. Dynamic map showing how milk production has varied by state from 2070 to 2021 with bubbles to represent tonnes produced by each state.

See file called milkByState_dots_02.gif on repo

At the start of the time series (1970) the main area of milk production was Wisconsin, with lesser centers in California and New York State with some production in neighbouring states. Through the 1970's and 1980's production in California increased and was static in other states. This was the situation when I was a vet student - all the research was coming from Wisconsin (Madison), New York (Cornell) and California (Davis). In the 1990's production rose in Texas - often thought of as a dry country. By the late 1990's California was dominating and the NW Pacific states were producing more. In the 2010's California and Wisconsin were the main states. This continued in the 2020's with California dominating followed by Wisconsin but also a notable amount produced in the NW Pacific states and in Texas and New Mexico.

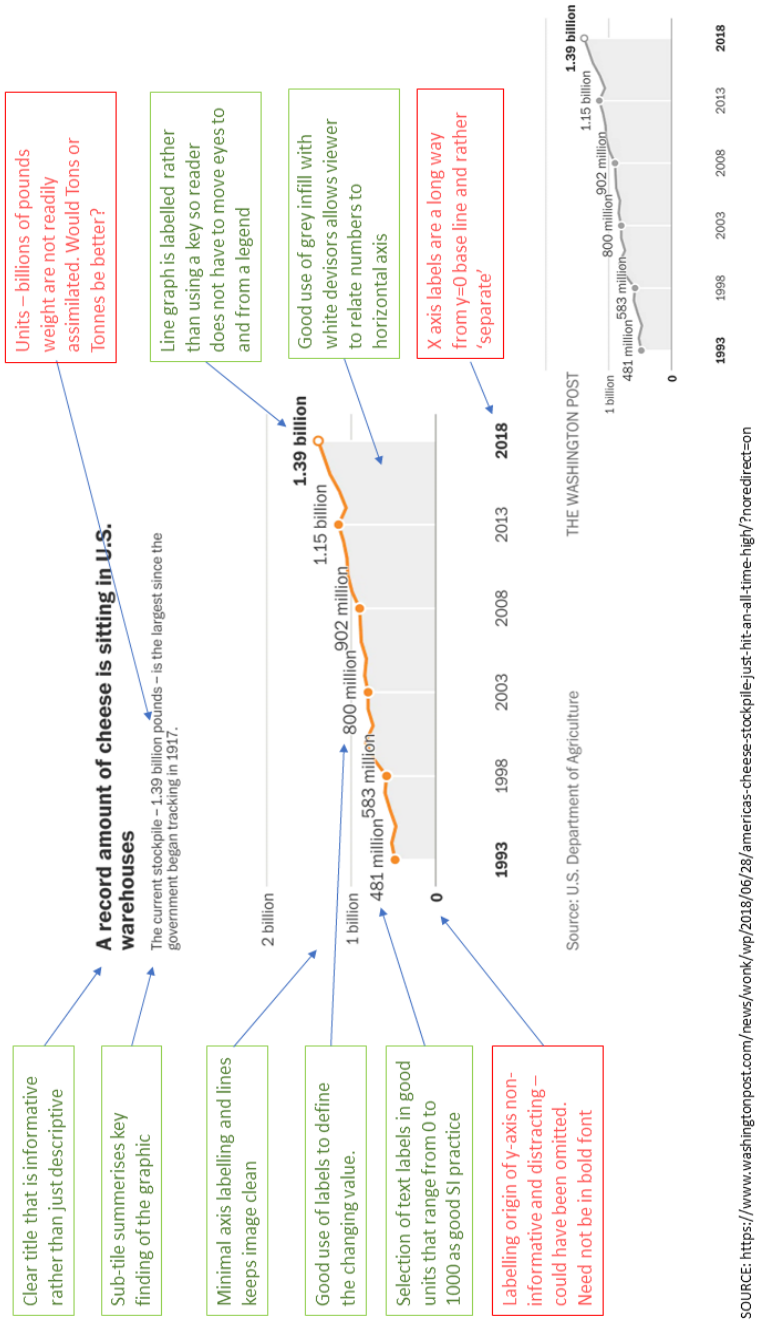
Graph Critique

Good Graphic

Washington Post has a 'good' image of cheese stockpiles over time that can be assessed.

Source: https://www.washingtonpost.com/news/wonk/wp/2018/06/28/americas-cheese-stockpile-just-hit-an-all-time-high/?noredirect=on

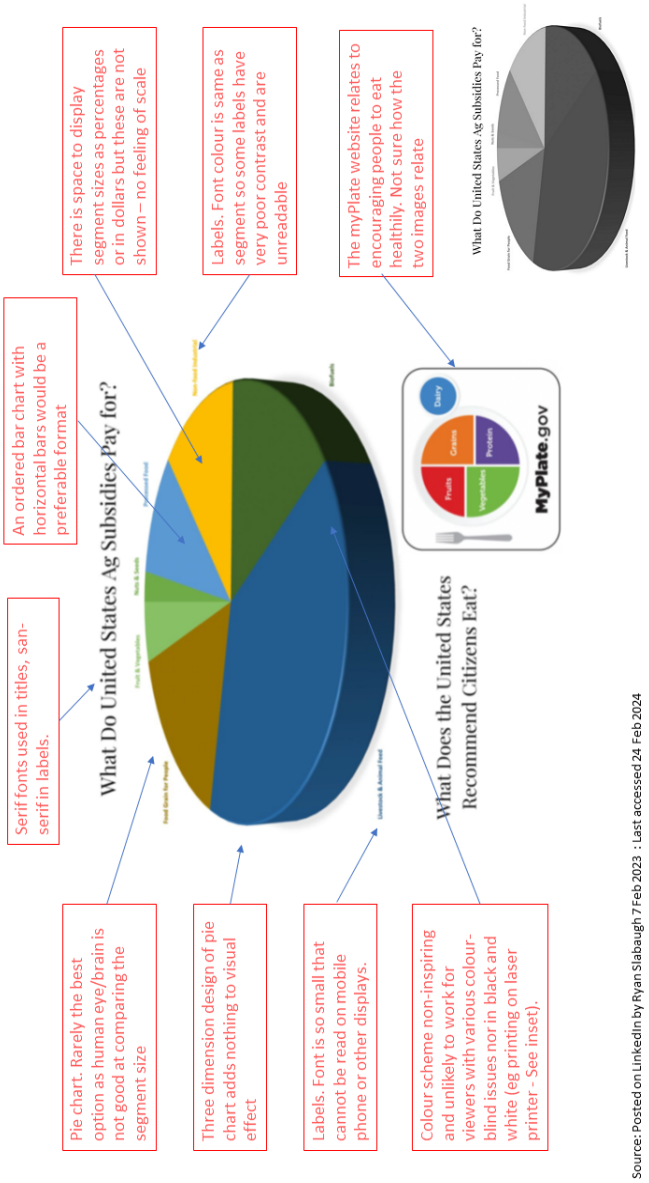
Figure 7. Graphic from Washington Post showing changes in cheese stocks annotated to highlight good and bad features of the image. (*Graphic_good_24Feb23.png on repo)*



Poor Graphic

On 7 February Ryan Slabaugh (a "Mission-driven regenerative leader" from the Quad Cities in central USA, with 3000 followers on Linked In) posted the following graphic (Figure 8) on LinkedIn which will be used for an appraisal of a 'bad' graphical image.

Figure 8. Graphic from LinkedIn feed how USA agricultural subsidies are allocated. Annotated to highlight good and bad features of the image. (*Graphic_poor_24Feb23.png on repo)*



Critique (500 words)

Figure 7 is taken from the Washington Post website and refers to an article about the rising cheese stocks in the United States. The graphic is of a minimalist style with a high ratio of white paper to 'ink'. The graph is essentially a line-graph showing how the amount of stored cheese has increased since 1993. The title and sub-tile are clear with a descending font size. The text is concise and is more informative than merely descriptive so as to act a short summary of the information in the graph. There is possibly too much white space between the sub-title and the graph and then the graph and the x-axis labels but we are not looking at the graphic in the context of the article layout. The use of colours is minimal with just an orange line to show the trend. The y-axis and data are clearly labelled so that no legend is required and sensible SI ranges (0-1000) are used for the units. However expressing the stockpiles in billions of pounds (lb) of cheese is a little unusual. Maybe Tons (this is for an American audience) would have been better – is referring to billions falsely exploiting the impact of 'very big numbers'.

The graphic has a clear 'infill' below the line with white segmentation allowing easy comparison with the x axis. Similarly thin, light grey lines allow the viewer to relate to the y-axis without being distracting. Lastly the selected colour palette should not cause confusions for people with colour vision issues and it also works well on a grey scale (see inset) which help maintain impact when printed on a B+W laser printer.

Figure 8 is from a LinkedIn post of 7 February 2023 by an American called Ryan Slabaugh. The Linkedin article is untitled but relates to the imbalance between USA food and agricultural policies. There are two titles and, I think, the aim is to show that there is an imbalance between the allocations of Ag subsidies to agricultural sub-sectors and the sub-classification of what citizens are recommended to eat. However, the sub-sector categories are not comparable. Dairy will be a part of the Livestock segment, protein will be from many segments. The graphic from MyPlate.gov is actually their logo rather than a numerical recommendation and the main pie chart is also not very numerical.

Considering the main graphical image, this is a pie-chart which is a very poor way of representing sub-categories of a whole; the human eye-brain axis is not good at comparing segments. Furthermore the pie-chart is shown as three dimensional which adds nothing but extra ink and 'dross' to the image. An ordered bar-chart with labelled, horizontal bars would have been preferable. The segments are labelled in the colour of the segment in a very small font such that they are very difficult to read. No numerical units (US dollars or percentages) are assigned to each segment although there is space. Colour palette selection is poor. In the original image the segment colours start to merge (eg the two greens) and will probably cause issues to people with colour blindness issues. When rendered in black and white (see inset) several segments (including the biggest) start to merge. Font selection is poor with serif fonts used for the titles and san-serif for the segment labels. One suspects the main aim of this graphic is NOT to inform!