Institute of Data

An analysis on food production and the environmental impact

Vagner Bauer

Capstone Project

IOD Cohort AU-NZ 25-OCT-22
Data Science & AI
Lead Trainer: Amin Khatami
Assistant Trainer: Ricky Nguyen and Sebastian Giunta

Auckland, New Zealand, 15th April 2023

Table of Contents

1. Project Context

Food production is a crucial component of human survival, providing sustenance for the ever-growing global population. In November of 2022, the world's population has reached 8 billion, and it is projected to reach 9.7 billion by 2050, according to the United Nations[1]. To meet the demands of this rapidly expanding population, food production has intensified, leading to significant environmental consequences.

Agriculture and livestock production are major contributors to deforestation, water pollution, greenhouse gas emissions, and soil degradation. The environmental impact of food production has become a major concern globally, as the need to produce more food for the growing population clashes with the need to protect the environment for future generations.[1]

As a result, there is a growing need for sustainable agriculture practices that can reduce the impact of food production on the environment while still meeting the world's food demands.

The first goal of this project is to identify the major producers of food worldwide and explore the major foods produced in Australia and New Zealand. Secondly, it aims to identify the group of foods that have bigger impacts on the environment.

2. Business Question

1. What countries are the largest food producers?
2. What foods are the most produced in Australia and New Zealand?
3. What foods have the most negative impact on the environment when they are produced?
4. What is the environmental impact of shifting the protein source in a diet from beef to poultry?

5. Data Question

Can we group in clusters different types of food based on their environmental impact?

6. Data Summary

Two different datasets were used in this project. The first was extracted from (Food and Agriculture Organization of the United Nations (FAO) 2023)[2] and contains a comprehensive picture of the pattern of a country's food production during a specified 2010 and 2020. There are 169234 observations and 14 features.

The second dataset was retrieved from Our World in Data (Ritchie, Rosado and Roser 2022, Ritchie, Rosado and Roser 2022)[3] and has information on the environmental impact of 38 most common foods grown across the globe. It consists of 38 observations and 7 features.

7. Data Preprocessing

Both datasets were downloaded as csv files and loaded into Jupyter Lab. All the necessary libraries were loaded and the first step after this was to create dataframes to observe the data and make transformations as needed. None of the datasets had missing values.

7.1 Cleaning and transforming the data – food production dataset

After the first dataset was loaded, it was possible to observe that some of the features were only codes with object type (figure 1). These were removed from the dataframe as they were not relevant for this analysis. After removing those, the dataframe was not reduced to 4 columns only: country, item, year and production (figure 2). It was also necessary to rename, remove the spaces and lower case the columns labels.

Another important step in this process was to create a new feature for country code. This step was done in order to be able to perform geo plots using the plyplot library.

	Domain Code	Domain	Area Code (M49)	Area	Element Code	Element	Item Code (CPC)	Item	Year Code	Year	Unit	Value	Flag	Flag Description
169229	FBS	Food Balances (2010-)	716	Zimbabwe	5142	Food	S2899	Miscellaneous	2016	2016	1000 tonnes	33.0	I	Imputed value
169230	FBS	Food Balances (2010-)	716	Zimbabwe	5142	Food	S2899	Miscellaneous	2017	2017	1000 tonnes	15.0	I	Imputed value
169231	FBS	Food Balances (2010-)	716	Zimbabwe	5142	Food	S2899	Miscellaneous	2018	2018	1000 tonnes	16.0	I	Imputed value
169232	FBS	Food Balances (2010-)	716	Zimbabwe	5142	Food	S2899	Miscellaneous	2019	2019	1000 tonnes	38.0	I	Imputed value
169233	FBS	Food Balances (2010-)	716	Zimbabwe	5142	Food	S2899	Miscellaneous	2020	2020	1000 tonnes	14.0	I	Imputed value

Figure 1: dataframe with 14 features

	country	item	year	production	alpha_3
0	Afghanistan	Wheat and products	2010	4924.0	AFG
1	Afghanistan	Wheat and products	2011	4894.0	AFG
2	Afghanistan	Wheat and products	2012	4924.0	AFG
3	Afghanistan	Wheat and products	2013	5215.0	AFG
4	Afghanistan	Wheat and products	2014	5293.0	AFG

7.2 Cleaning and transforming the data – emissions dataset

For the second dataset, a similar approach was performed. The columns had their names changed to facilitate the manipulation. This dataframe consisted of 2 categorical and 5 continuous variables (figure 3).

	item	origin	ghg_emissions_per_kg	land_use_per_kg	freshwater_withdrawals_per_kg	scarcity-weighted_water_use_per_kg	eutrophying_emissions_per_kg
0	Apples	Plant	0.43	0.63	180.1	12948.6	1.45
1	Bananas	Plant	0.86	1.93	114.5	661.9	3.29
2	Barley	Plant	1.18	1.11	17.1	696.4	2.33
3	Beef (beef herd)	Animal	99.48	326.21	1451.2	34732.5	301.41
4	Beef (dairy herd)	Animal	33.30	43.24	2714.3	119805.2	365.29

Figure 3: second dataframe after transformation

Since there was a good number of continuous variables, the describe method was used to return the description of the dataframe (figure 4). It's important to note that the scale of values of the 5 features were very different, so one step that was taken later in the project was to standardize the features by removing the mean and scaling to unit variance.

	count	mean	std	min	25%	50%	75%	max
ghg_emissions_per_kg	38.0	10.023947	19.172390	0.39	0.9800	1.800	8.5700	99.48
land_use_per_kg	38.0	27.865789	78.807730	0.33	0.9450	2.955	11.4425	369.81
freshwater_withdrawals_per_kg	38.0	951.971053	1354.366157	0.00	105.5000	408.100	1253.4000	5605.20
scarcity-weighted_water_use_per_kg	38.0	34974.528947	54150.639614	0.00	3325.0750	13563.250	34395.7750	229889.80
eutrophying_emissions_per_kg	38.0	49.231579	88.483731	0.69	3.3375	7.515	45.2925	365.29

Figure 4: dataframe description

8. Exploratory Data Analysis (EDA)

Data visualization is a crucial aspect of data science as it enables us to gain insights and make sense of complex data. Visualizing data can make it easier to identify patterns and trends that may be difficult to see in tabular form.

8.1 EDA on food production dataset

The first graph plotted was to compare total food production over the years for different countries. Plotly library has a function to plot geographic scatter plot (figure 5), which was perfect for this scenario.

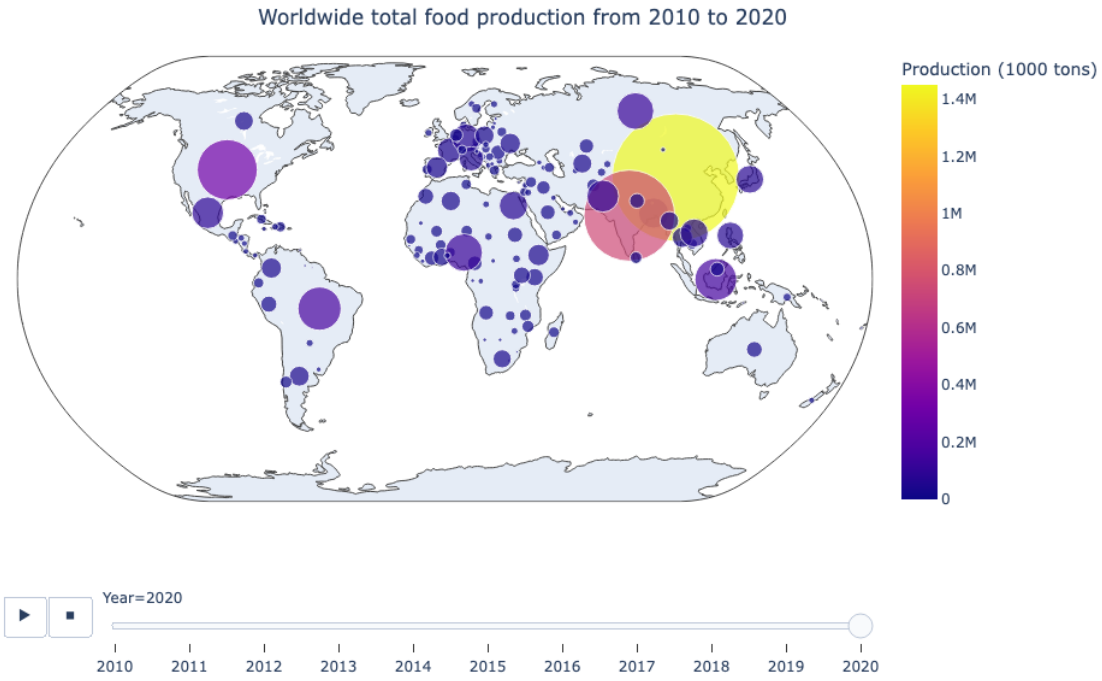

Figure 5: geographic scatter plot (world)

From the plot above, we can see that China, India and the USA were the major food producers in 2020, as defined by the bubble size and color. Australia and New Zealand had a very small contribution to the world total food production.

Figure 6 shows a comparison only in between Australia and New Zealand. In this plot, we see that there is a big difference in food production between the two countries.

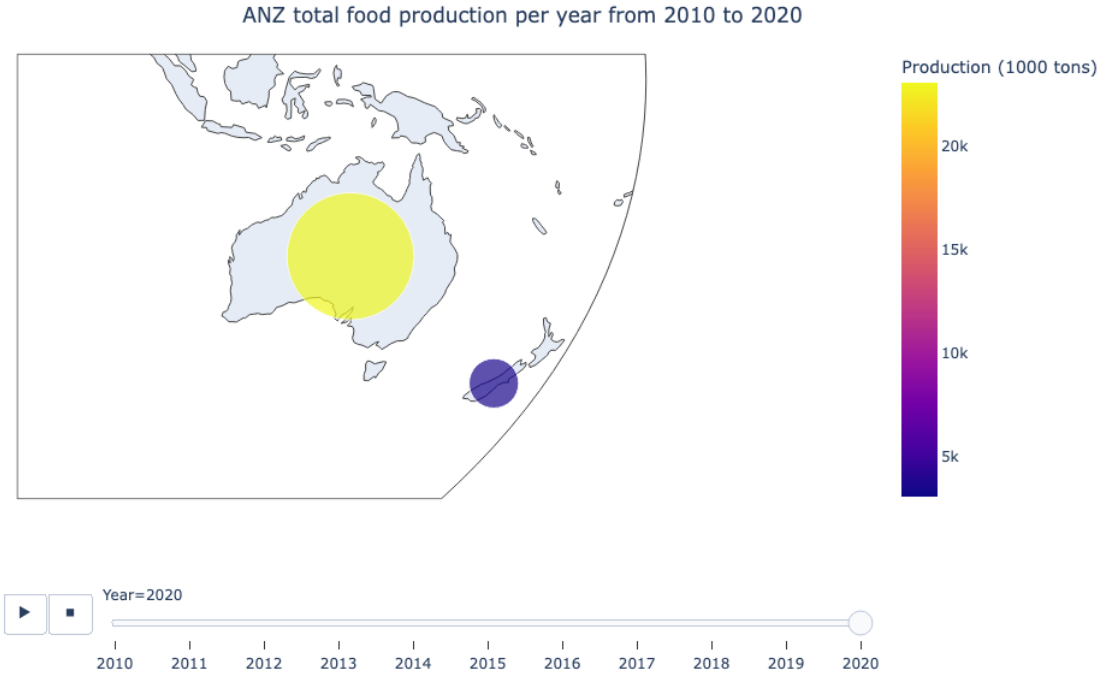

Figure 6: geographic scatter plot (Australia and New Zealand)

In 2020, Australia produced more than 20 million tons of food, whereas New Zealand produced around 4 millions tons.

Figure 7 presents a comparison between top 10 food products manufactured in Australia and New Zealand. Both countries had milk as top one over the period of 2010 to 2020.

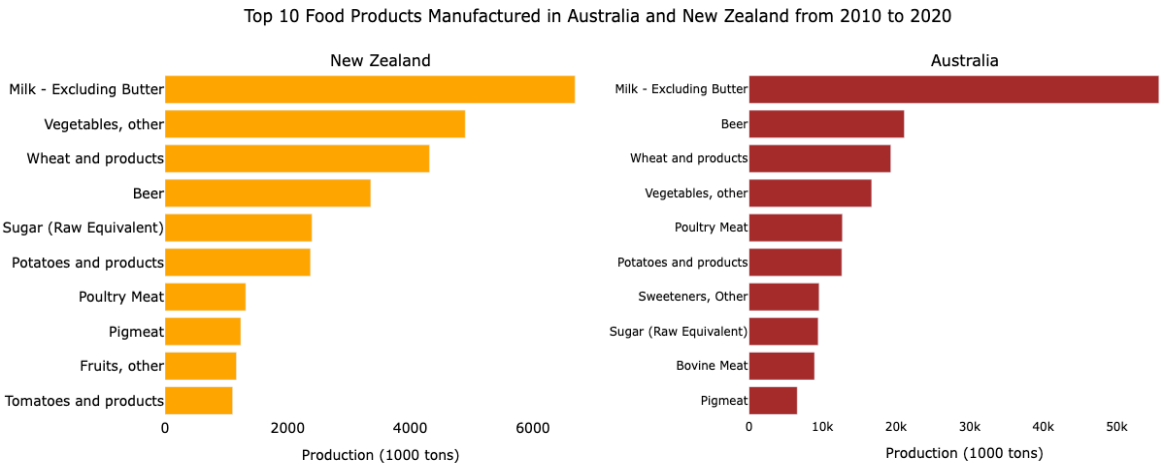

Figure 7: Top 10 food products manufactured in Australia and New Zealand

Regular expression was used in this project in order to divide the data into food from animal and plant origin. Among foods from animal origin category, the top 3 products manufactured in New Zealand are milk, poultry meat and pig meat. Among the plant origin, the top 3 products are vegetables, wheat and products and beer (figure 8).

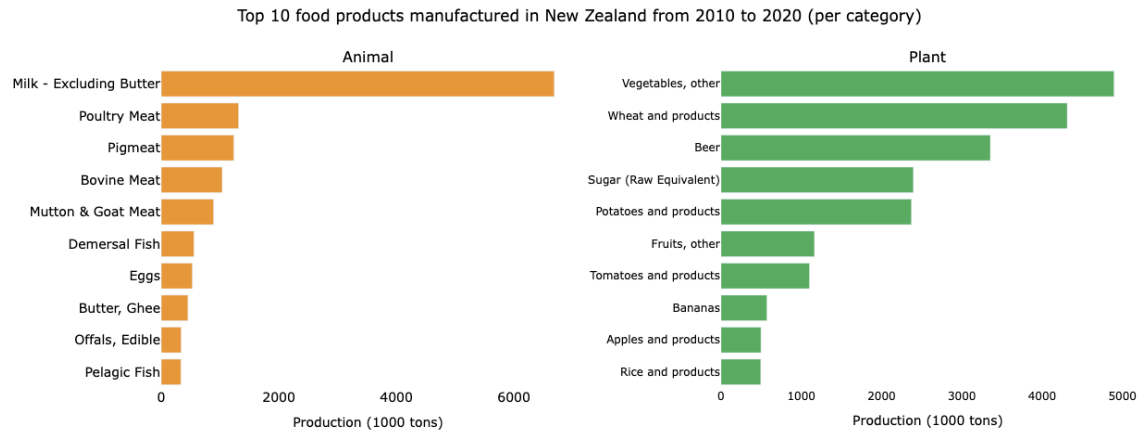

Figure 8: Top 10 food products manufactured in New Zealand divided into animal and plant origin

The same plot was generated for Australia. For animal origin category, the top 2 products are the same as in New Zealand, however the third one is bovine meat. Looking at the plant origin category, the top 1 product manufactured in Australia is beer, followed by wheat and products and vegetables (figure 9).

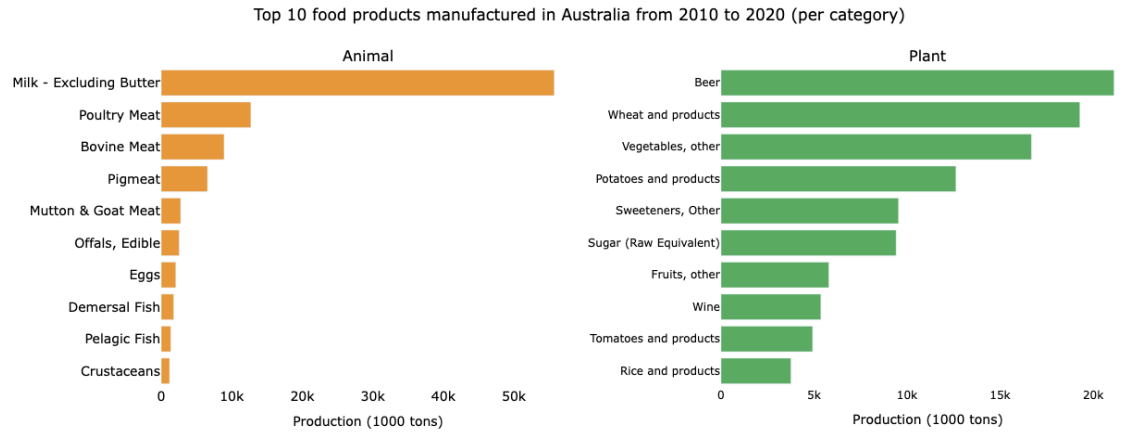

Figure 9: Top 10 food products manufactured in Australia divided into animal and plant origin

8.2 EDA on emission dataset

For the second dataset, horizontal bar plots were used to visualize the foods and their environmental impact based on the 5 features of the dataframe: greenhouse gas emissions, land use, fresh water withdrawal, scarcity-weighted water use and eutrophying emissions. (Ritchie, Rosado and Roser 2022)[3] define these features as:

- Greenhouse gas emission: measured in carbon dioxide-equivalents. This means non-CO2 gases are weighted by the amount of warming they cause over a 100-year timescale.
- Land use: measured in meters squared (m^2) per kilogram of a given food product.
- Fresh water withdrawal: measured in liters per kilogram of food product.
- Scarcity-weighted water: represents freshwater use weighted by local water scarcity. This is measured in liters per kilogram of food product.
- Eutrophying emissions: represent runoff of excess nutrients into the surrounding environment and waterways, which affect and pollute ecosystems. They are measured in grams of phosphate equivalents (PO_4eq).

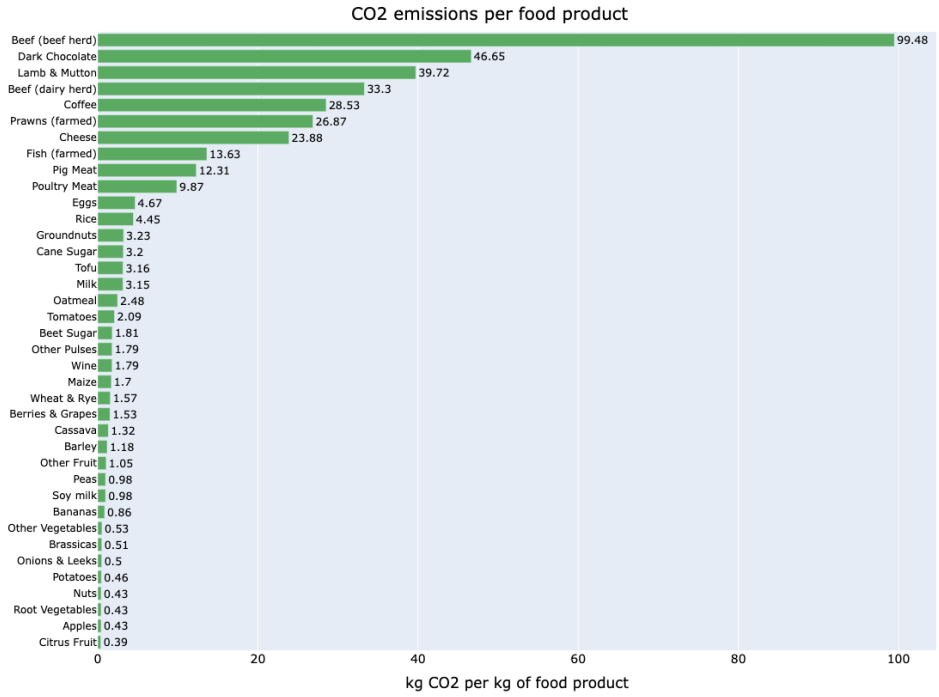

Figure 10: Greenhouse gas emissions per kg of food product

Figure 10 shows that beef, dark chocolate and lamb and mutton are the top 3 products in terms of greenhouse gas emissions (measured as kg of CO2 eq) per kg of product. Figure 11 represents the amount of area used (measured in square meters) to product one kilogram of food product. Lamb and mutton is the top one with 360.81 sqm necessary to produce 1 kg of meat. This is followed by beef and cheese, with 326.21 sqm and 87.79 sqm respectively.

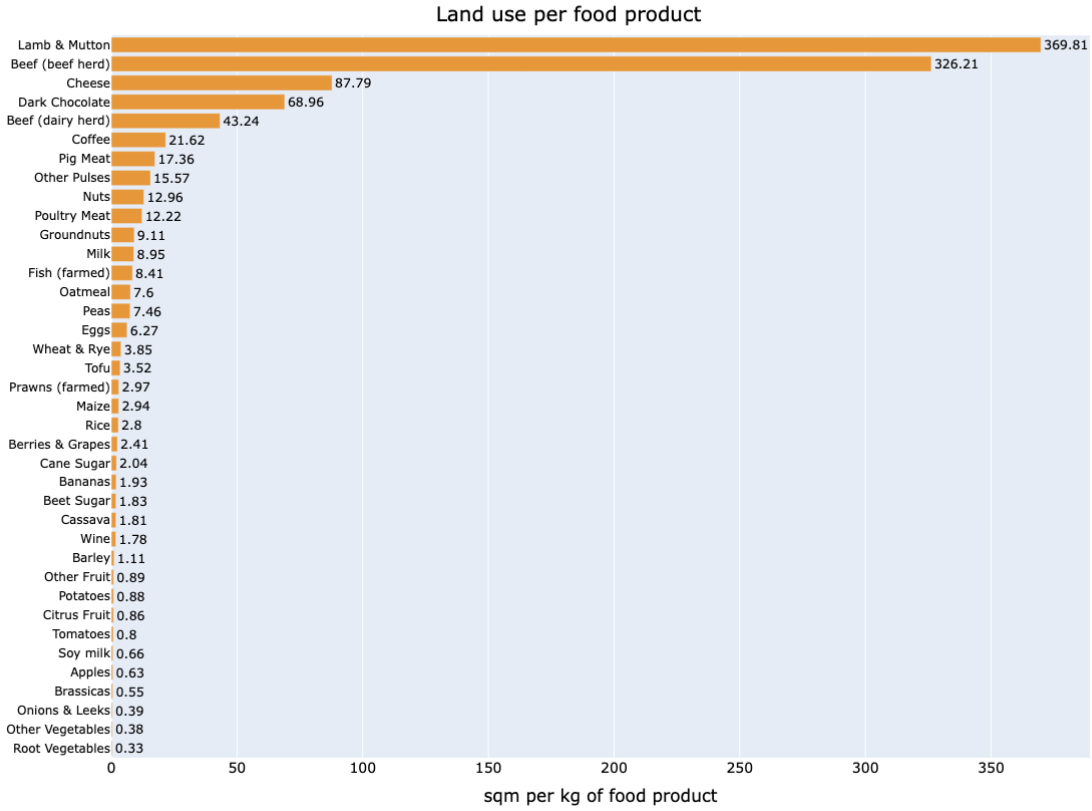

Figure 11: land use (sqm) per kg of food product

In terms of freshwater withdrawal, the production of cheese, nuts and fish are the ones with higher water usage (figure 12). The results are similar when checking scarcity-weight water use (freshwater use weighted by local water scarcity), with nuts, cheese and lamb and mutton in the top 3 (figure 13). Lastly, figure 14 shows the numbers for eutrophying emissions, which represent runoff of excess nutrients into the surrounding environment and waterways. In number one is beef (dairy herd) followed by beef (beef herd) and fish (farmed).

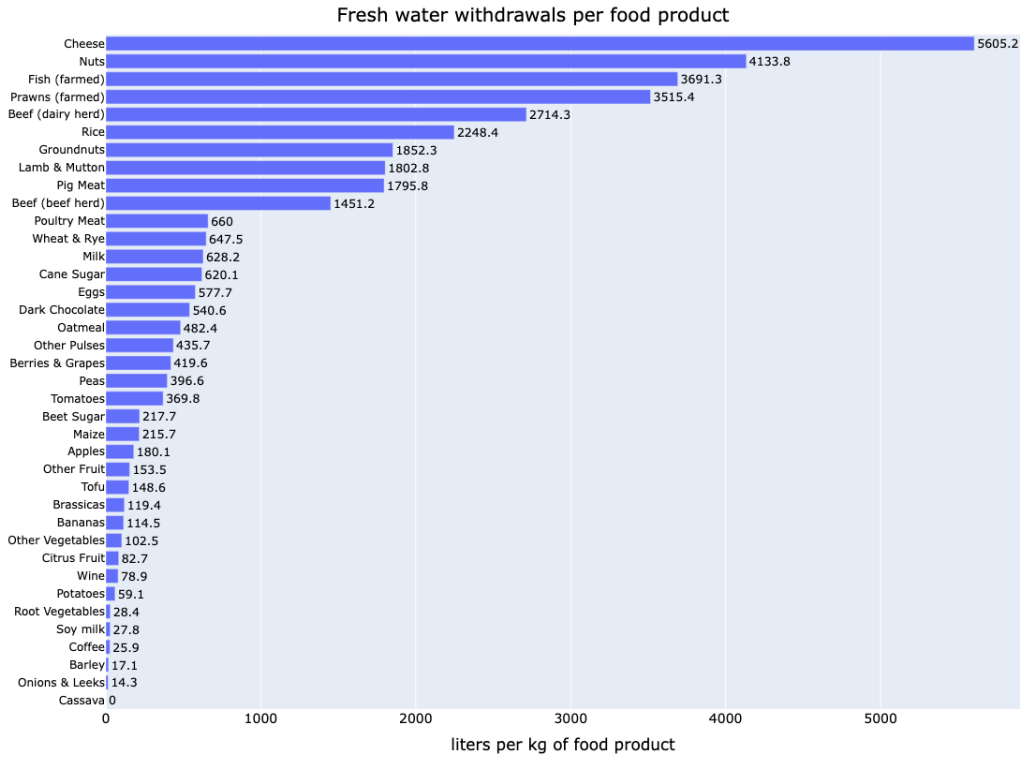

Figure 12: freshwater withdrawal (liters) per kg of food product

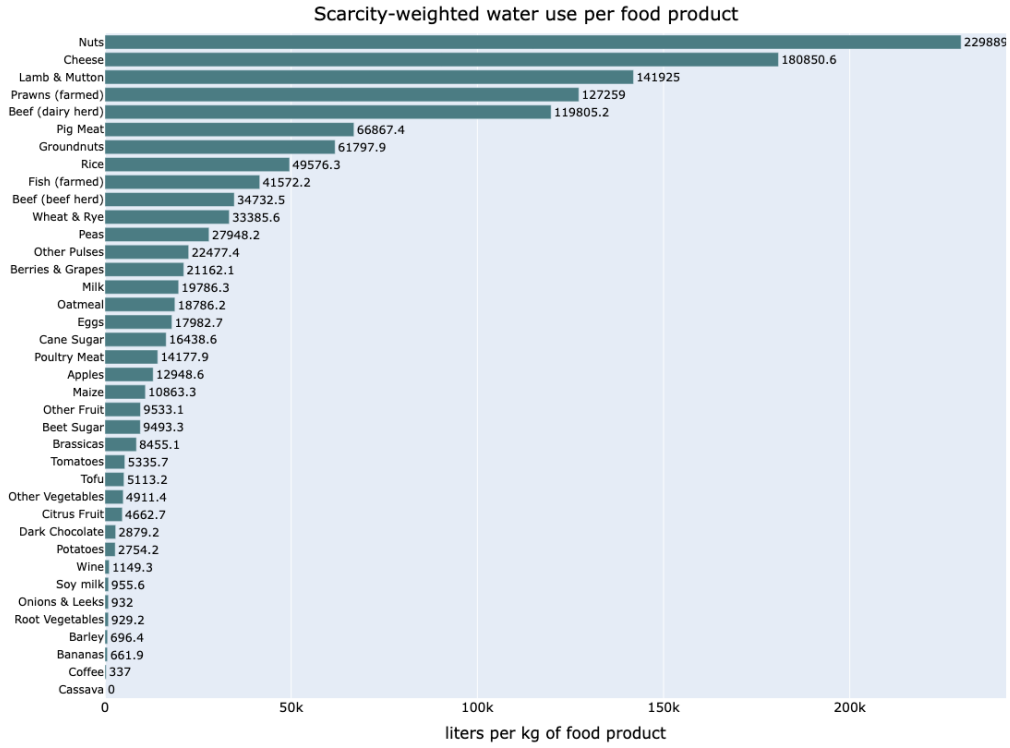

Figure 13: scarcity-weighted withdrawal (liters) per kg of food product

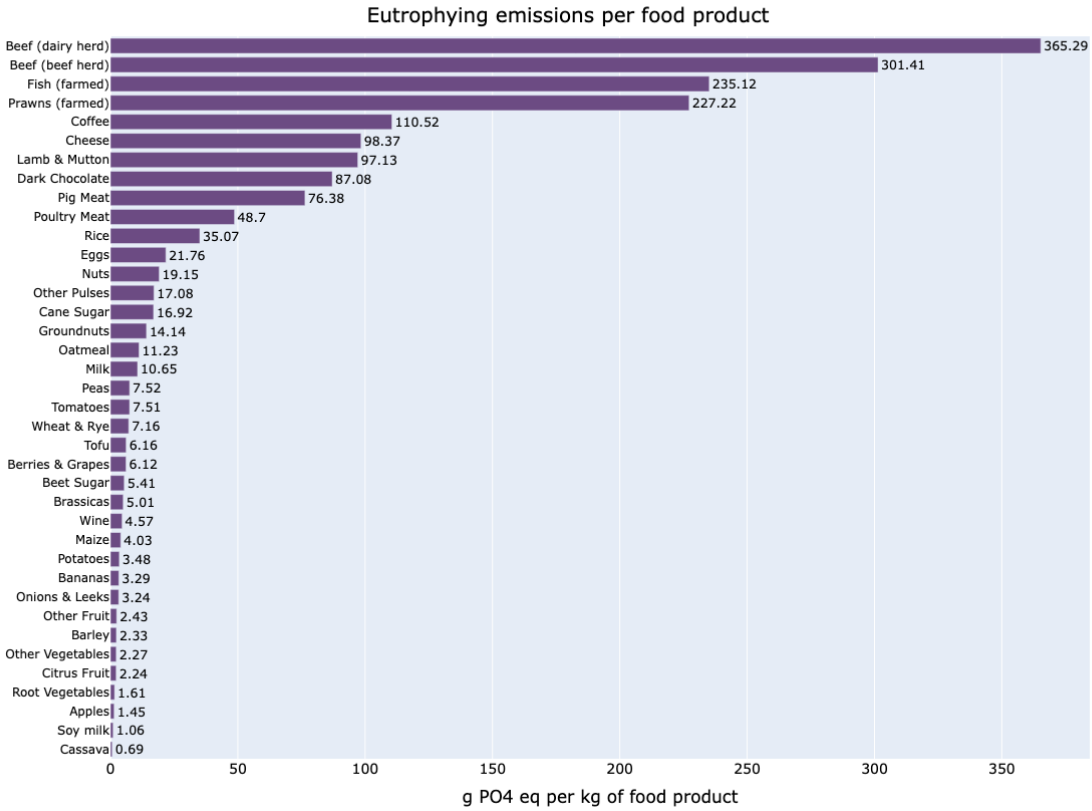

Figure 14: eutrophying emissions (g PO4 eq)) per kg of food product

9. Unsupervised learning

In order to perform clustering analysis, the first step was to remove the categorical variable origin, as it is not useful in this case. Then, the dataframe was reindex to have the food product (item) as the index, leaving only the continuous variable available for clustering.

The data was scaled using StandartScaler() from Sklearn library as mentioned previously and Principal Component Analysis (PCA) was used to reduce the dimensionality of the data while retaining as much of the original variation as possible.

Next, to perform k-means clustering, the Elbow method was used to determine the optimal number of clusters (figure 15). The plot of the sum of the square distance between points in a cluster and the cluster centroid (WCSS) against the number of clusters shows that 3 is the optimal number of clusters.

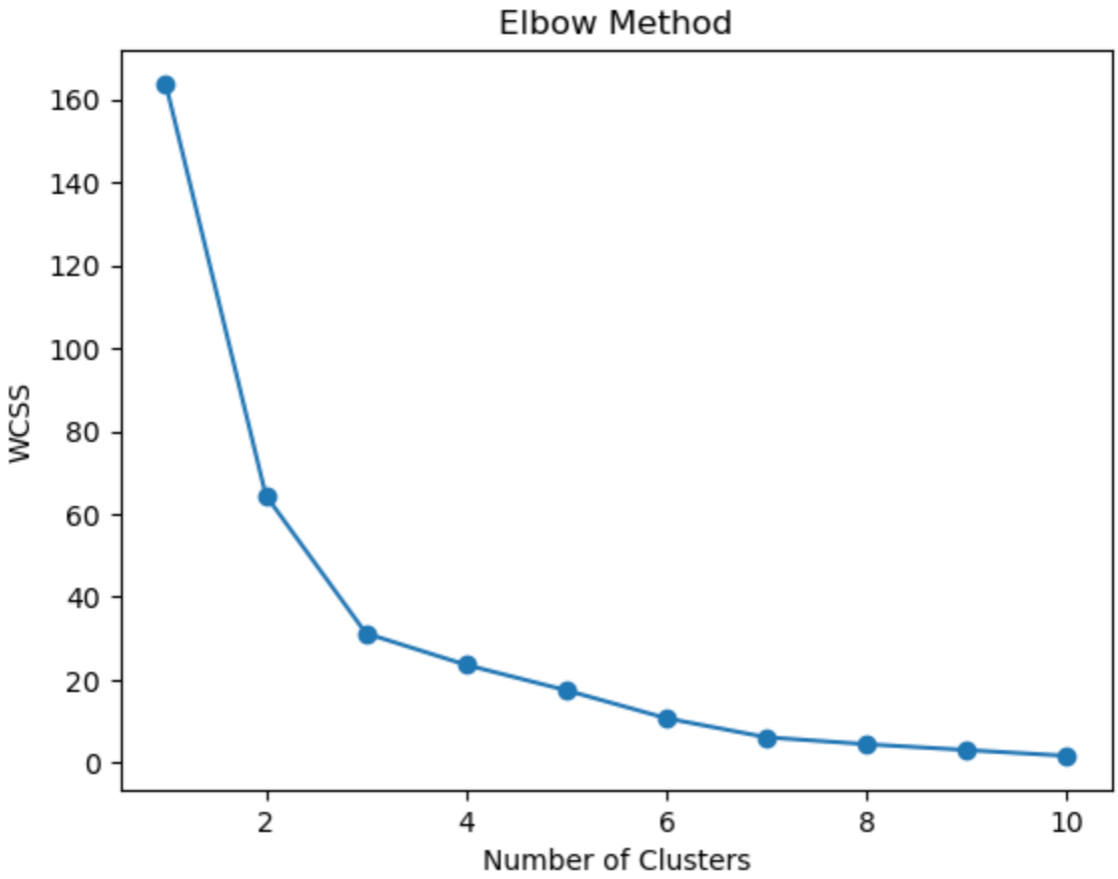

Figure 15: Elbow method

This was confirmed by performing the silhouette method, which measures how similar a data point is within-cluster (cohesion) compared to other clusters (separation). The silhouette coefficient for each data point ranges from -1 to 1, where a value of 1 indicates that the data point is very well matched to its own cluster and poorly matched to other clusters. To apply the silhouette method, the data is clustered into different numbers of clusters (minimum 2) and then the average silhouette coefficient is calculated for each cluster solution. The solution with the highest average silhouette coefficient is considered to be the best choice for the number of clusters in the dataset[4]. Figure 16 shows that the highest average silhouette coefficient 0.7354 when using 3 clusters.

```
For n_clusters = 2, the average silhouette_score is 0.7275
For n_clusters = 3, the average silhouette_score is 0.7354
For n_clusters = 4, the average silhouette_score is 0.6070
For n_clusters = 5, the average silhouette_score is 0.6346
For n_clusters = 6, the average silhouette_score is 0.6804
For n_clusters = 7, the average silhouette_score is 0.6656
For n_clusters = 8, the average silhouette_score is 0.6414
For n_clusters = 9, the average silhouette_score is 0.5569
For n_clusters = 10, the average silhouette_score is 0.5597
```

Figure 16: silhouette scores for different clusters

The k-means algorithm was then fitted to the data and the clusters were plotted using a scatter plot (figure 17).

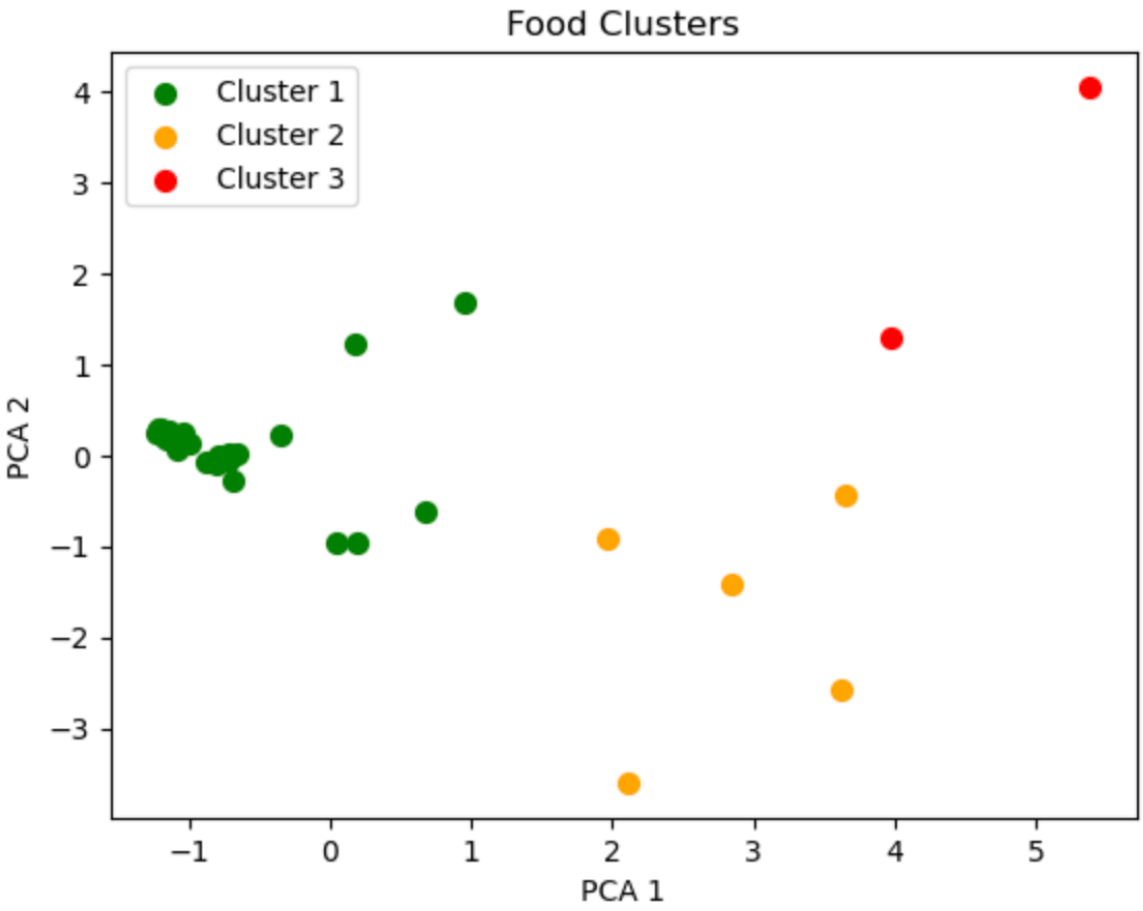

Figure 17: food clusters

The list of food items in each cluster is plotted below (figure 18).

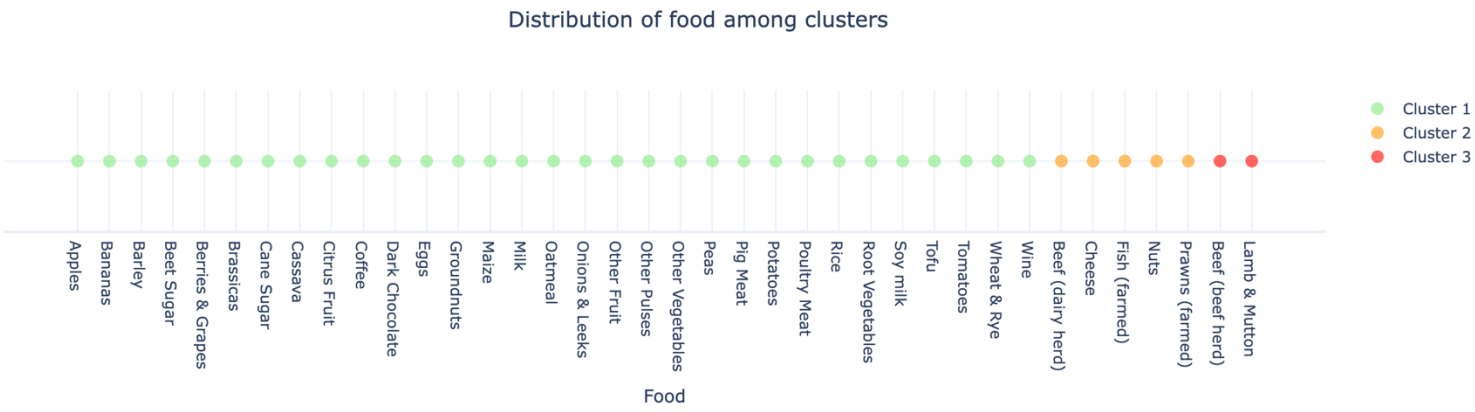

Figure 18: distribution of food among clusters

10. Summary and conclusions

The EDA for the food production dataset has shown that China, India and USA are the biggest food producers worldwide. When checking countries in Oceania, the food production in Australia was approximately 5 times higher than in New Zealand in 2020. The top 3 food products manufactured in Australia from 2010 to 2020 was milk, beer and wheat. For the same period in New Zealand, these were milk, vegetables and wheat.

On the other hand, the EDA on the emissions dataset has shown that beef, dark chocolate and lamb and mutton are the products that lead to higher CO2 emissions when produced. Lamb and mutton, beef (beef herd) and cheese are the ones that utilize more land to produce 1kg of product. When looking at water use and scarcity-weighted water use, products like cheese, nuts, lamb and mutton and fish are the ones that require more water. Finally, beef (both beef and dairy herd) and fish (farmed) are the product that have generates higher runoff of excess nutrients into the surrounding environment and waterways.

The k-means clustering algorithm was successfully implemented to create 3 different groups of foods according to their environmental impact. Beef (beef herd) and lamb and mutton are the food products with highest environmental impact, while poultry meat, pig meat, fruits and vegetables tend to have a much lower impact on environment.

11. Business answer

1. China, India and USA
2. Australia: milk, beer and wheat (2020 | New Zealand: milk, vegetables and wheat (2020)
3. Lamb and mutton and beef (beef herd)
4. The average meat consumption per capita in Australia in 2020 was 37kg of beef and 48kg of poultry. Supposing that everyone in Australia decides to swap 10kg of beef for 10kg of poultry in their diet throughout the year, each individual would be saving approximately 900kg of greenhouse gas emissions and approximately 14,000 liters of water.

12. Data answer

Yes, 3 clusters were successfully created using k-means algorithm.

13. Next steps

This project has a very detailed EDA that could be used in different sectors of the food industry to persuade consumers with environmental consciousness to change their eating habits.

As of the next steps, it would be good to have a more detailed analysis on the environmental impact of foods from animal vs plant origin. Another interesting approach would be to create clusters using foods from the same origin. For example, would bananas, apples, wheat, and beer be in the same cluster or different ones?

14. References

1. United Nations. (2021). World Population Prospects 2019: Highlights (ST/ESA/SER.A/423). https://population.un.org/wpp/Publications/Files/WPP2019_10KeyFindings.pdf

2. FAO.Food Balance Sheet. License: CC BY-NC-SA 3.0 IGO. Extracted from: https://www.fao.org/faostat/en/#data/FBS. Date of Access: 01-04-223

3. Hannah Ritchie, Pablo Rosado and Max Roser (2022) - "Environmental Impacts of Food Production". Published online at OurWorldInData.org. Retrieved from: 'https://ourworldindata.org/environmental-impacts-of-food' [Online Resource]

4. https://www.analyticsvidhya.com/blog/2021/05/k-mean-getting-the-optimal-number-of-clusters/#:~:text=The%20silhouette%20coefficient%20or%20silhouette,scikit%2Dlearn%2Fsklearn%20library.