The Environmental Impact of Food Production

Rylan Daniels

2021

Table of Contents

Hypothesis

As the world's population has grown tremendously, so has the global demand for food. Water and energy needs have increased tremendously to meet rising agricultural needs. All of this has had extreme consequences on the environment. Energy sources are being quickly depleted, and carbon emissions from food production are significant. What could be done to feed the world in a more sustainable way? I hypothesize that if we restrict foods that have high energy needs, high volume, and high emissions, we can minimize the environmental impact of food production. By focusing on a single production stage and other emission effect—determined during Exploratory Data Analysis—this hypothesis can be quantified more easily.

My data analysis predicts environmental emissions from the production of food using machine learning. To pursue this goal, correlations between the stages of food production (farming, transportation, etc.) and other effects (land use, greenhouse gas emissions) will be examined.

The Data

I used the "Environment Impact of Food Production" data set from Kaggle, which can be found here: https://www.kaggle.com/selfvivek/environment-impact-of-food-production

There are a total of 43 observations and 23 variables in the dataset. The observations are comprised of food products, and the variables are comprised of different types of emissions that those food products produce.

The original variables from the dataset are:

Food product – name of food category
Land use change – Greenhouse emissions in Kg CO2 equivalents per Kg of product
Animal Feed – Greenhouse emissions in Kg CO2 equivalents per Kg of product
Farm – Greenhouse emissions in Kg CO2 equivalents per Kg of product
Processing – Greenhouse emissions in Kg CO2 equivalents per Kg of product
Transport – Greenhouse emissions in Kg CO2 equivalents per Kg of product
Packaging – Greenhouse emissions in Kg CO2 equivalents per Kg of product
Retail – Greenhouse emissions in Kg CO2 equivalents per Kg of product
Total emissions – Greenhouse emissions in Kg CO2 equivalents per Kg of product
Eutrophying emissions per 1000kcal
Eutrophying emissions per kilogram
Eutrophying emissions per 100g protein
Freshwater withdrawals per 1000kcal
Freshwater withdrawals per 100g protein
Freshwater withdrawals per kilogram
Greenhouse gas emissions per 1000kcal
Greenhouse gas emissions per 100g protein
Land use per 1000kcal
Land use per kilogram
Land use per 100g protein
Scarcity-weighted water use per kilogram
Scarcity-weighted water use per 100g protein
Scarcity-weighted water use per 1000kcal

Data Transformation

While most of the data was useful for analyzing emissions of food production, I was able to transform the data set to make it more practical for addressing my hypothesis. I used the feature engineering technique of subsetting to remove variables concerning the scarcity of water usage since this was not critical to my analysis. In addition, I added a new custom variable to the data set, “Food Group”. The Food Group variable categorized the data into the five major food groups—grains, vegetables, protein, fruit, and dairy. This categorization would become useful for data analysis and data visualization.

Exploratory Data Analysis (EDA)

Through a variety of Exploratory Data Analysis techniques, including arithmetic and table functions, I was able to determine the following insights about the data set:

- There are 33 non-missing values of greenhouse gas emissions per 1000 kCals.
- There are 43 types of food in the data set.
- Statistics and percentages about the Food Group categorical variable.

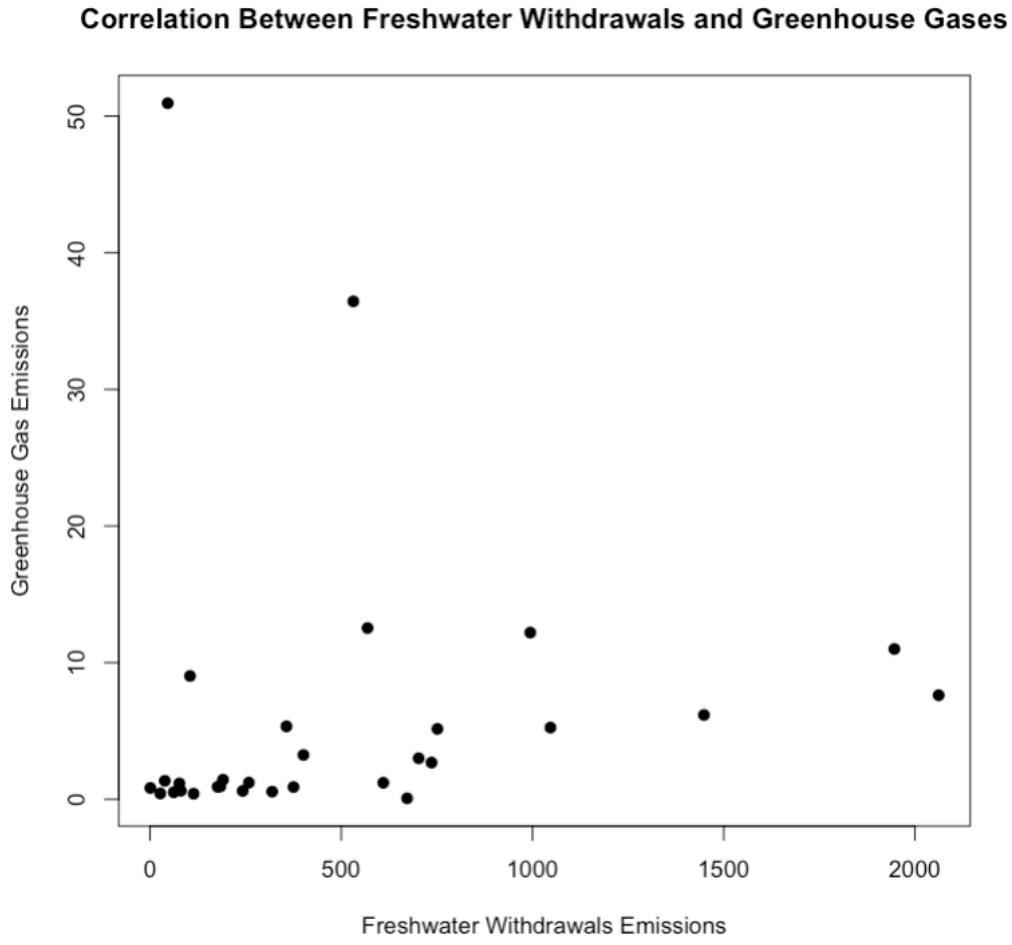
I also analyzed the data via functions providing a summary, the structure, the variable definitions, and dimensions of the data set.

I used the following functions:

```r
# exploratory data analysis (EDA)
head(foodProduction)
dim(foodProduction) # dimensions
# [1] 43 23
names(foodProduction) # variable definitions
summary(foodProduction)

str(foodProduction)
```

```r
# there are 33 non-missing values of greenhouse gas emissions per 1000 kCals
sum(!is.na(foodProductionGrouped$Greenhouse.gas.emissions.per.1000kcal..kgCO.eq.per.1000kcal.))

# there are 43 types of food in the data set
levels(as.factor(foodProductionGrouped$Food.product))

# there is just 1 of each type of food
table(foodProductionGrouped$Food.product)

# percentages of food products
prop.table(table(foodProductionGrouped$Food.product))

# amount of foods in each food group
table(foodProductionGrouped$`"Food.group"`)
# dairy     fruit     grain   protein vegetable
# 3         8         5        11        16

# each food group's percentage
prop.table(table(foodProductionGrouped$`"Food.group"`))
# dairy      fruit      grain    protein  vegetable
# 0.06976744 0.18604651 0.11627907 0.25581395 0.37209302
```

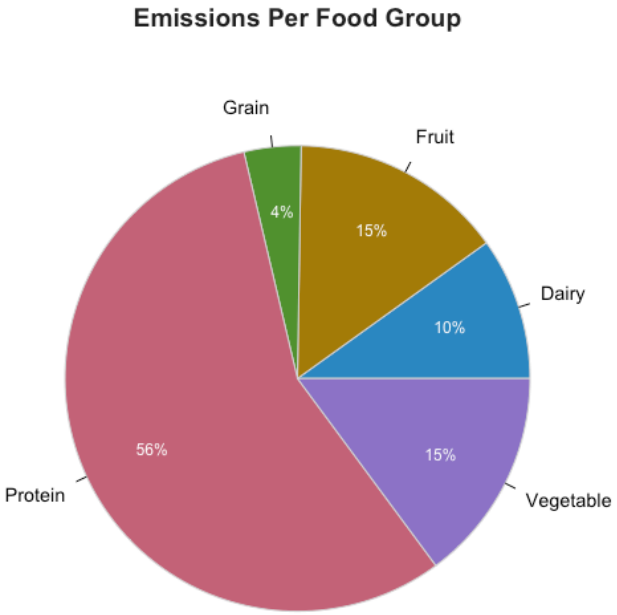
In addition, I explored the correlation between freshwater withdrawals emissions and greenhouse gas emissions in the following scatterplot:

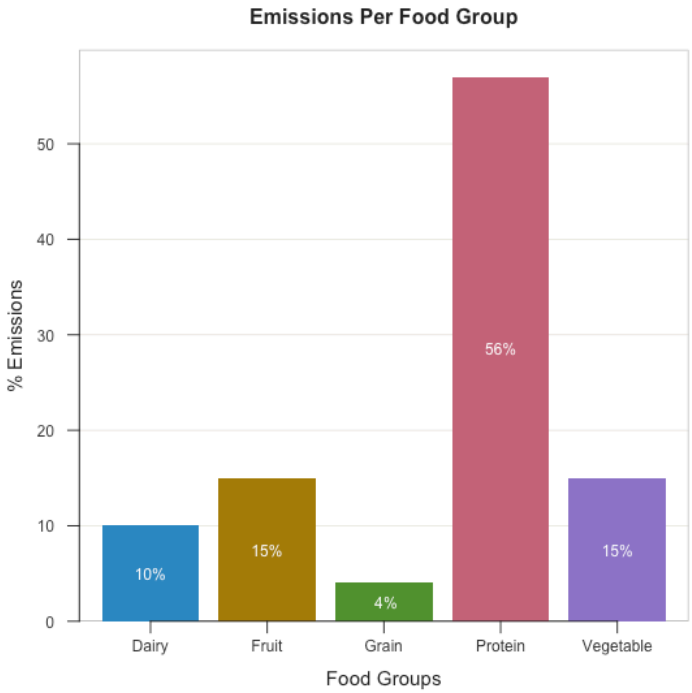


Data Visualization

By visualizing the data through several different data visualization methods, I gained a clearer understanding of the data set's most important conclusions. Both exploratory and expository data visualizations were used.

The following charts (see next page), a pie chart and a bar chart, illustrate the amount of emissions of each food group. These charts make use of the newly added Food Group variable to simplify the visualization to a few distinct sections, rather than simultaneously displaying over 40 individual segments of emissions. This makes the charts' more clear and interpretable. I used the lessR package to generate charts that enable superimposed percentages. Clearly, protein products produce the most emissions—over half! Grains are the only food group to produce less than ten percent of recorded emissions.



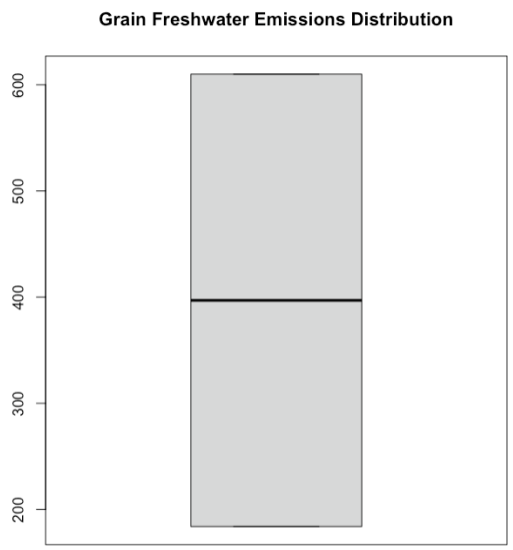


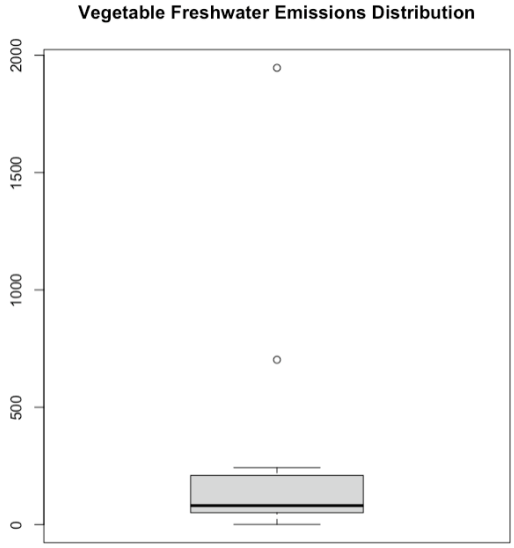
Next, I generated heat maps (see next page) to visualize levels of magnitude in the emissions data. Two separate heat maps were created to distinguish between emissions directly resulting from production processes like framing and transportation, and other effects like eutrophication and greenhouse gas emissions. Farming appeared to contribute the most to a food product's emissions across almost all the foods, while the other production processes were much closer in their emissions. By far, emissions from freshwater withdrawals were the most significant amount of emissions for most of the food products. Notably, the heat map for emission types showed blank spaces where NAs occurred in the data, indicating a lack of recorded data for certain emissions.

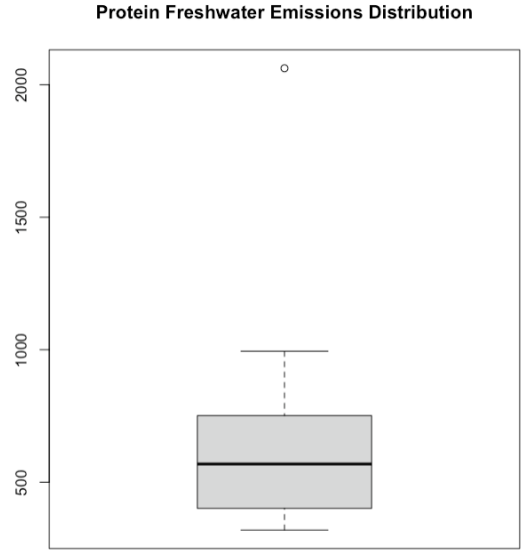
Farm
Processing
Transport
Packging
Retail
Total_emissions
Shrimps (farme
Fish (farmed)
Eggs
Cheese
Milk
Poultry Meat
Pig Meat
Lamb & Mutton
Beef (dairy her
Beef (beef herd
Dark Chocolate
Coffee
Other Fruit
Wine
Berries & Grap
Apples
Bananas
Citrus Fruit
Other Vegetabl
Brassicas
Root Vegetable
Onions & Leek
Tomatoes
Olive Oil
Rapeseed Oil
Sunflower Oil
Palm Oil
Soybean Oil
Tofu
Soymilk
Groundnuts
Nuts
Peas
Other Pulses
Beet Sugar
Cane Sugar
Cassava
Potatoes
Rice
Oatmeal
Barley (Beer)
Maize (Meal)
Wheat & Rye (

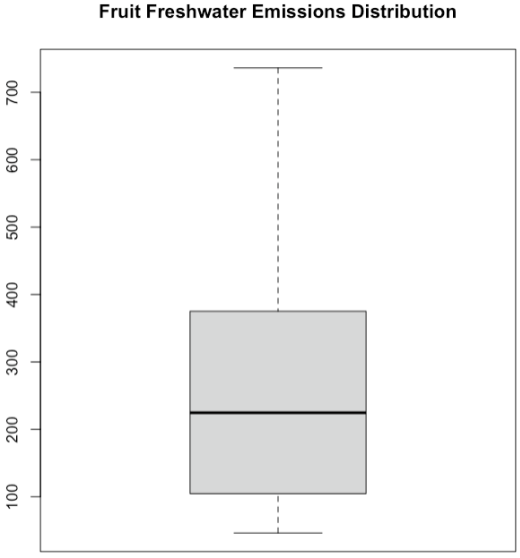
Land.use.change
Animal.Feed
Total_emissions
Eutrophying.emissions.per.1000kcal..gPO.eq.per.1000kcal.
Eutrophying.emissions.per.kilogram..gPO.eq.per.kilogram.
Eutrophying.emissions.per.100g.protein..gPO.eq.per.100.grams.protein.
Freshwater.withdrawals.per.1000kcal..liters.per.1000kcal.
Freshwater.withdrawals.per.100g.protein..liters.per.100g.protein.
Freshwater.withdrawals.per.kilogram..liters.per.kilogram.
Greenhouse.gas.emissions.per.1000kcal..kgCO.eq.per.1000kcal.
Greenhouse.gas.emissions.per.100g.protein..kgCO.eq.per.100g.protein.
Land.use.per.1000kcal..m..per.1000kcal.
Land.use.per.kilogram..m..per.kilogram.
Land.use.per.100g.protein..m..per.100g.protein.
"Food.group"
Shrimps (farme
Fish (farmed)
Eggs
Cheese
Milk
Poultry Meat
Pig Meat
Lamb & Muttor
Beef (dairy her
Beef (beef herd
Dark Chocolate
Coffee
Other Fruit
Wine
Berries & Grap
Apples
Bananas
Citrus Fruit
Other Vegetab
Brassicas
Root Vegetable
Onions & Leek
Tomatoes
Olive Oil
Rapeseed Oil
Sunflower Oil
Palm Oil
Soybean Oil
Tofu
Soymilk
Groundnuts
Nuts
Peas
Other Pulses
Beet Sugar
Cane Sugar
Cassava
Potatoes
Rice
Oatmeal
Barley (Beer)
Maize (Meal)
Wheat & Rye (

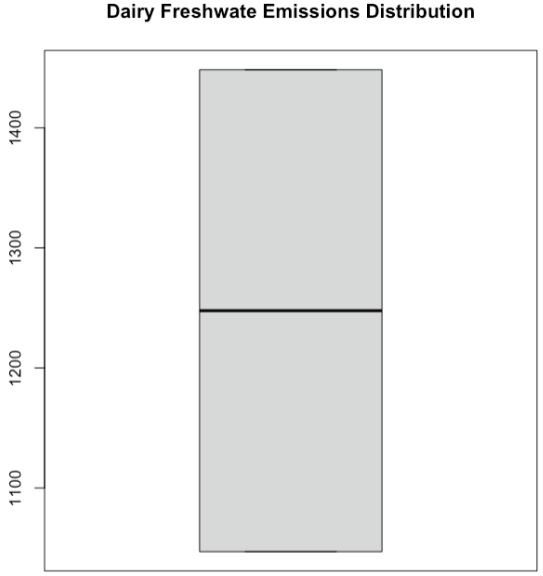
I created box plots displaying the frequency distribution of freshwater withdrawals emissions from each food group:





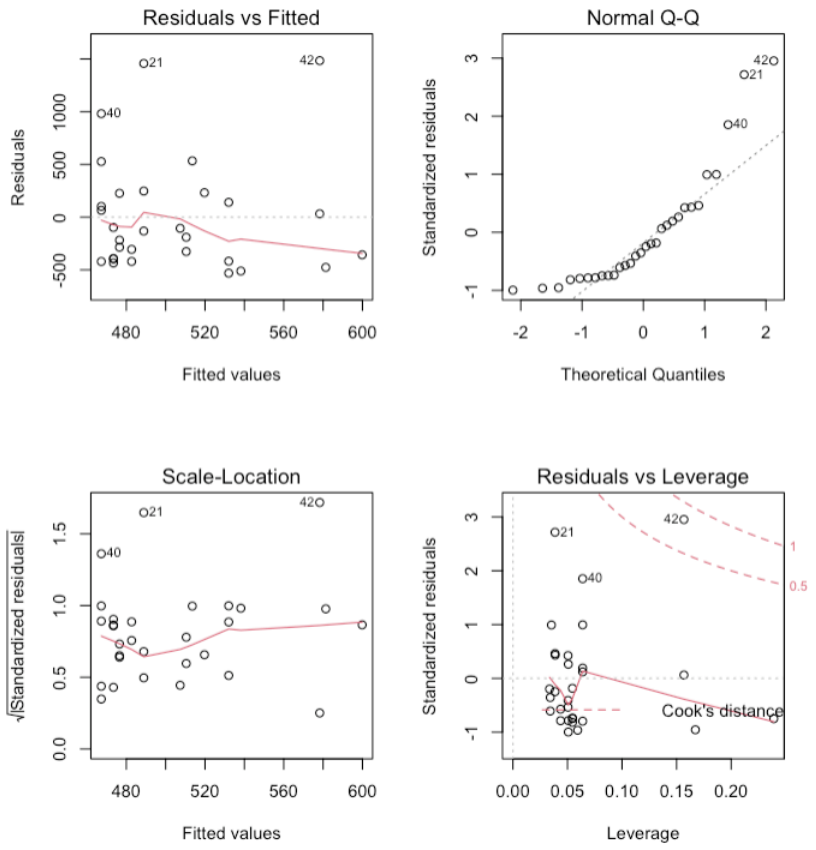




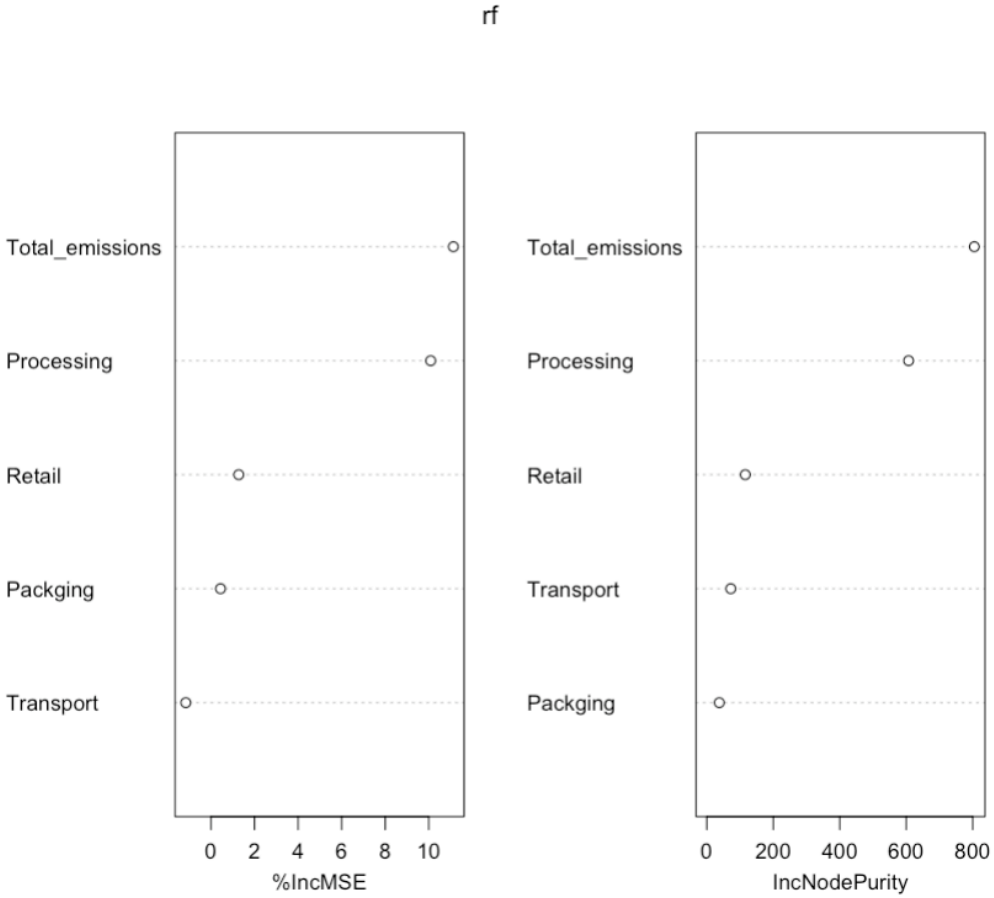


Machine Learning

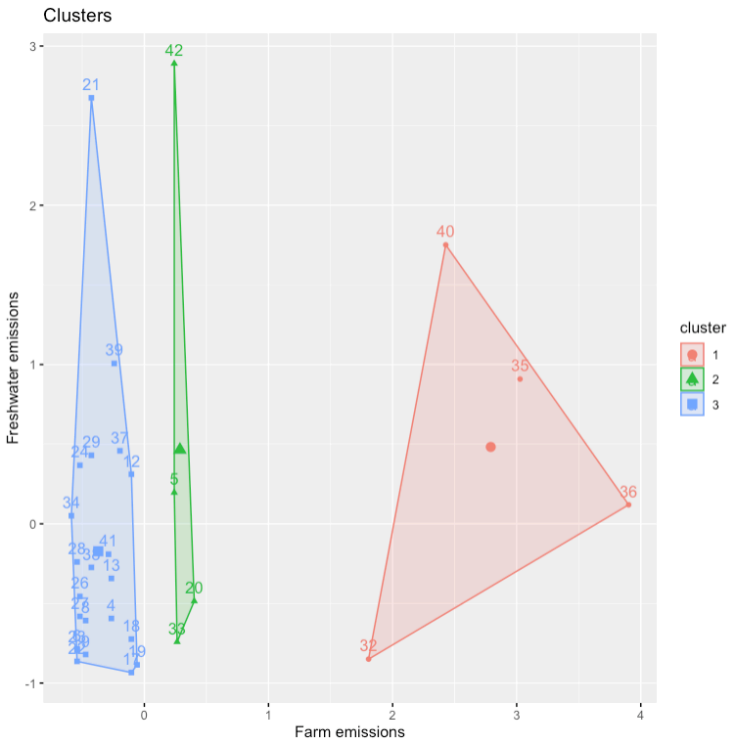
I used the simple linear regression algorithm to determine the correlation between freshwater withdrawals emissions and farm production emissions, with freshwater withdrawals emissions being the response variable and farm production emissions being the predictor. In this supervised learning, these variables were chosen based off of the EDA and data visualization work. Freshwater withdrawals emissions had the most significance impact, and farm emissions were the largest across most food products. Plotting the results of the machine learning algorithm produced the following charts:



In addition, I used another supervised learning technique in the form of random forests to continue machine learnings capabilities in predicting farm emissions. The variable importance plot produced the following chart:



Finally, I integrated the unsupervised machine learning method of the K-means algorithm. This would be used to explore the accuracy of the previously made predictions. After coordinating data clusters and centroids, the algorithm produced the following plot using the facto extra library:



Conclusion

Based on my analysis of the data, it was determined that grains have the lowest average levels of emission out of each of the five major food groups. Freshwater emissions are the most severe form of emissions caused by food production processes. Out of the entire food production process, farming is the stage with the most emissions, and this is highly consistent throughout all food groups.
In the future, my algorithm could be used when ordering at a restaurant to see which menu items are more environmentally friendly, by examining the ingredients in each dish. Chefs could use my program to design more sustainable dishes. Operations managers could use my program to implement food distribution supply chains that minimize their effects on the environment.

Algorithms may have the potential to augment the world. 🤖🌎

References

Environmental Impact of Food Production: What are the environmental impacts of food and agriculture? https://www.kaggle.com/selfvivek/environment-impact-of-food-production

Our World In Data: https://ourworldindata.org/

Our World in Data - Environmental impacts of food production: https://ourworldindata.org/environmental-impacts-of-food