Alec Brooks

Eriko Sakamura

DATA101

28 November 2023

The Banana Index: A Unique Way to Look at Food Efficiency and Environmental Impact

Introduction

Food production and agriculture play a vital role in any society. The significance of the foods we consume and which foods we give economic focus to can have substantial impacts on our local and global environments. Maintaining a balanced approach to variety, coupled with an informed and nuanced understanding of the consequences of our choices, is a crucial reality in today's world. However, how we address these questions is a crucial distinction as well; often, we only consider the relative ecological impact of an entire industry on our environment. Yet, by also considering our nutritional needs, we may gain further insight into this complex and ever-evolving topic. To explore this, we will delve into "The Banana Index," examining its relevant variables and creating a unique formula to explore the relevant impact of different common foods. In doing so, we may find a new and engaging way to measure the ecological impact our choices make, and empower us to make balanced informed decisions.

Data Description

The Banana Index is a dataset created by the website *The Economist*, the creators of The Big Mac Index. A dataset showing the cost of a Mcdonald's Big Mac in different countries as a method of exploring relative purchase power. The Banana Index shows the environmental impact of producing different foods indexed to the banana. It includes 160 observations of different common foods and gives 11 unique variables, including carbon emissions per weight, calorie counts, protein, and fat contents according to how many bananas would generate the same amount of emissions for each metric. An example of this is, as mentioned in a blog post on datainnovation.org, "producing one kilogram of ground beef generates as many carbon emissions as 109 kilograms of bananas, while producing one calorie of ground beef generates as many carbon emissions as producing 54 calories of banana" (Stevens, Morgan. "Visualizing the Carbon Impact of Foods Relative to Bananas").

To answer our research question, we will be exploring the relationship between the following variables provided by the Banana Index: the total emissions per kilogram (emissions_kg), the total emissions per 1000 kilocalories (emissions_1000kcal), the total emissions per 100 grams of protein (emissions_100g_protein), and the total emissions per 100 grams of fat (emissions_100g_fat). In addition to these variables, we also have accompanying metrics looking at the total land used rather than emissions: (land_use_kg), (land_use_1000kcal), (land use per 100 grams of protein), (land use per 100 grams of fat). All the values within these variables are quantitative and indexed to the equivalent in bananas. The Banana Index also includes values for the Banana itself. But considering the nature of the dataset and our research question, these values are not necessary.

Methodology

To start the analysis, I examined the dataset's structure. This examination revealed the presence of unnecessary columns, specifically (Chart?), (type), (Banana values), and (Unnamed: 16). These variables appear to be utilized for chart generation by The Economist during export and are irrelevant to our analysis. In fact, their inclusion in the dataset could potentially lead to issues in the future. Consequently, I opted to entirely remove these columns. Additionally, while the (year) column may initially seem pertinent, all observations in this dataset share the same year value (2022). Therefore, I decided to eliminate this variable as well.

Moving forward, we conducted a check for any remaining NAs in the dataset, uncovering two observations with two NAs each, totaling four. Upon investigation, we identified these observations as "Olive oil" and "Sunflower oil," with the NA values present in the columns (emissions_100g_protein) and (land use per 100 grams of protein). The consistency of the NAs seemed relevant. A quick Google search revealed that the reason for these NAs is likely because Olive and Sunflower oil both contain 0 protein. Given the nature of our analysis, this is acceptable, and we simply assigned those NAs a value of 0. Subsequently, we concluded the cleaning and data preparation phase by standardizing all the variables.

Next, I focused on building the ranking formula. I started by reviewing a correlation table, considering (emissions_kg) as the independent variable to observe correlation values for each variable. These correlation values would serve as weights for our formula. After applying this weighted formula to create a new rankings column, I generated a list of the 25 lowest-ranking foods and the 25 highest-ranking foods. It is important to emphasize that, since our ranking is based on emissions, lower overall scores should be considered preferable, while higher scores indicate less preferable items.

Results

Table 1 is the output of the correlation function in relation to (emissions_kg). As we can see the values range a fair amount. With (land_use_kg) being the highest with 0.95.

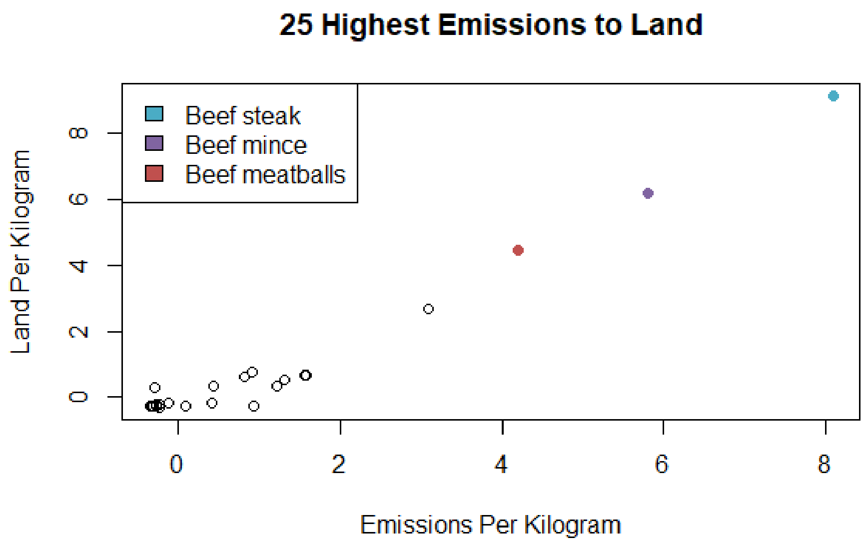
Table 1.

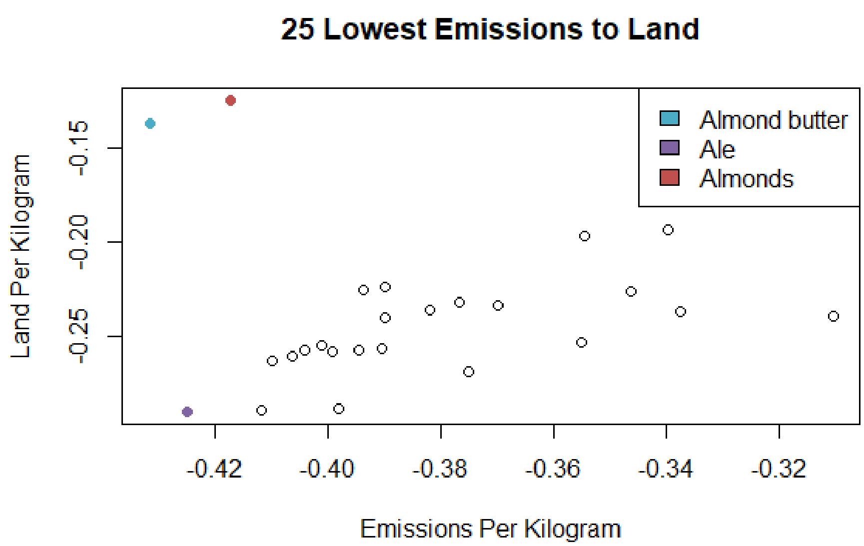
emissions_kg	1.000000000000000
emissions_1000kcal	0.882638920916087
emissions_100g_protein	0.224554871934291
emissions_100g_fat	0.29681932724674
land_use_kg	0.95374063696656
land_use_1000kcal	0.926725744832703
Land use per 100 grams of protein	0.719970991852648
Land use per 100 grams of fat	0.643778321988888

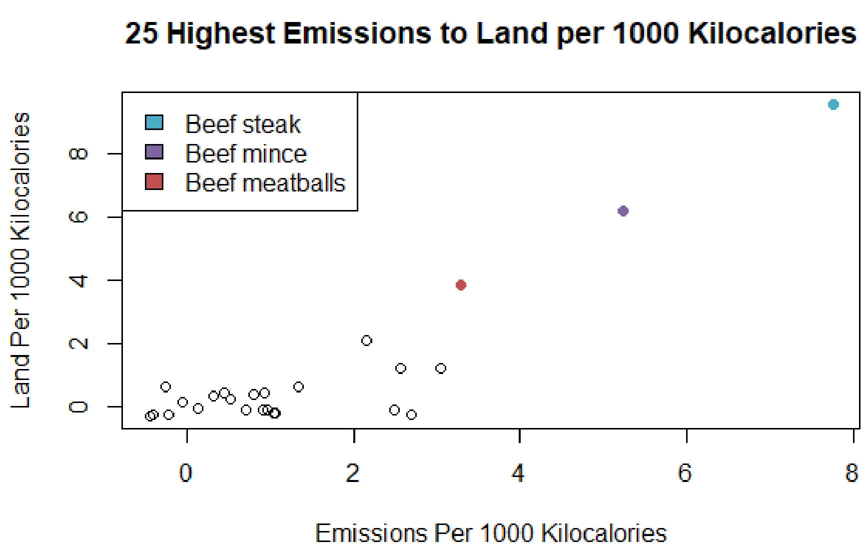
Using the values in table 1 and dividing by the tables sum resulted in the following formula:

0.1770467*emissions_kg + 0.1562683*land_use_kg + 0.03975669*land_use_1000kcal + 0.05255087*emissions_1000kcal + 0.1688566*`Land use per 100 grams of protein` + 0.1640737*`Land use per 100 grams of fat` + 0.1274685*emissions_100g_fat + 0.1139788*emissions_100g_protein.

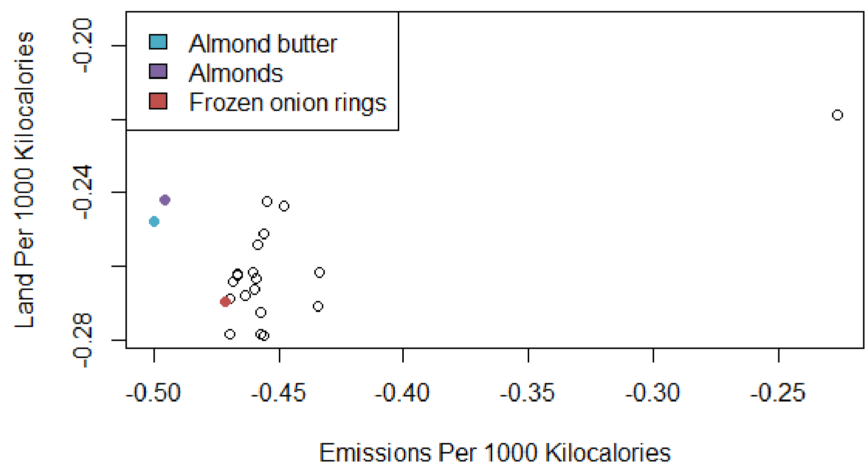
These plots illustrate the 25 highest and lowest observations for each variable, along with its corresponding land variable. The objective in examining these plots is to analyze the extremes of each metric. While the specific highest observations may vary, identifying commonalities among them can provide insights into our ranking. If consistent patterns emerge among the highest and lowest observations, it could indicate that our formula is yielding the intended results.


25 Highest Emissions to Land
Beef steak
Beef mince
Beef meatballs
Land Per Kilogram
0
2
4
6
8
0
2
4
6
8
Emissions Per Kilogram

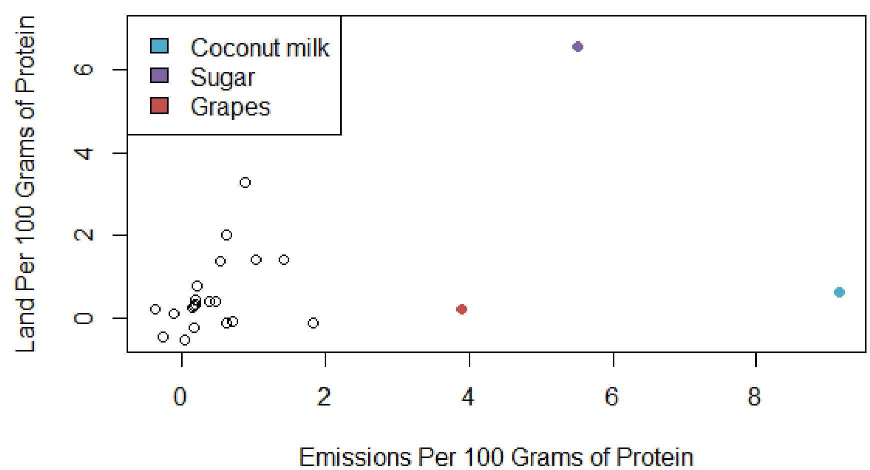

25 Lowest Emissions to Land
Almond butter
Ale
Almonds
Land Per Kilogram
-0.15
-0.20
-0.25
-0.42
-0.40
-0.38
-0.36
-0.34
-0.32
Emissions Per Kilogram


25 Highest Emissions to Land per 1000 Kilocalories
Beef steak
Beef mince
Beef meatballs
Land Per 1000 Kilocalories
0
2
4
6
8
0
2
4
6
8
Emissions Per 1000 Kilocalories

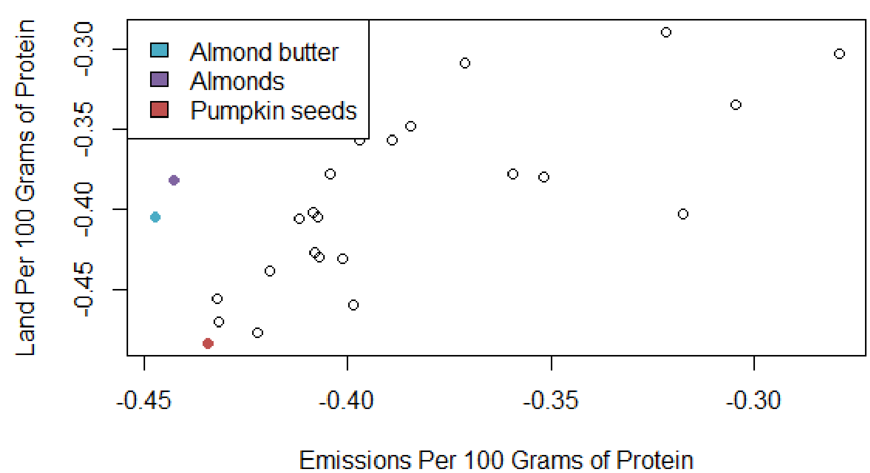
25 Lowest Emissions to Land Per 1000 Kilocalories



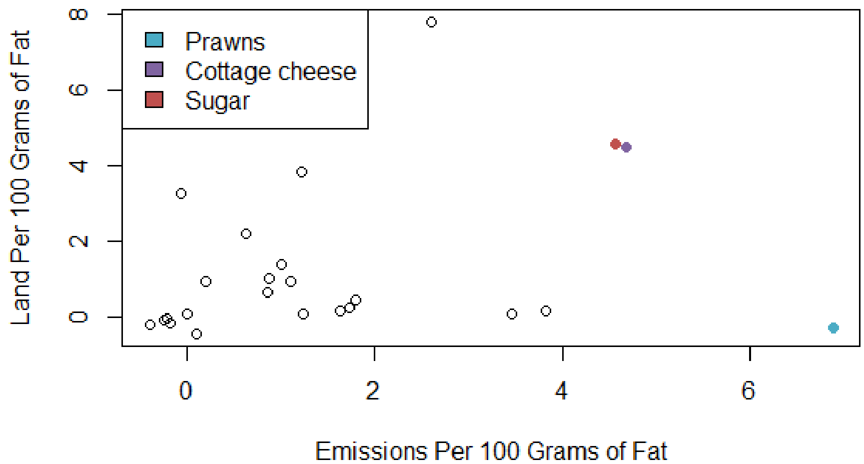
25 Highest Emissions Per 100 Grams of Protein



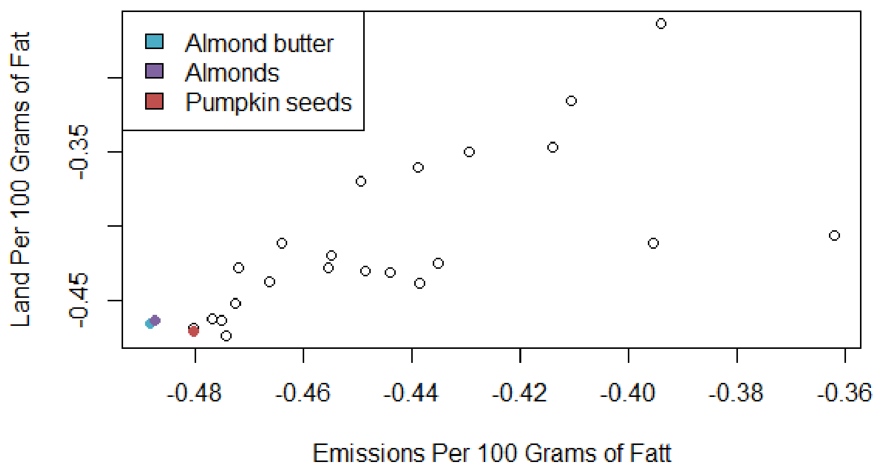
25 Lowest Emissions Per 100 Grams of Protein



25 Highest Emissions Per 100 Grams of Fat



25 Lowest Emissions Per 100 Grams of Fat



25 Most Impactful Foods

1	Beef steak	6	Beef burger	11	Raspberries	16	Lamb chops	21	Bacon
2	Beef mince	7	Coconut milk	12	Marmalade	17	Beetroot	22	Tuna
3	Sugar	8	Grapes	13	Lamb (leg)	18	Lamb burgers	23	Tomatoes
4	Beef meatballs	9	Prawns	14	Lentils	19	Chilli con carne	24	Cherry tomatoes
5	Cottage cheese	10	Lettuce	15	Dark chocolate	20	Strawberries	25	Tomato ketchup

25 Least Impactful Foods

1	Ale	6	Almond butter	11	Nut loaf	16	Pancakes	21	Sourdough bread
2	Pumpkin seeds	7	Naan	12	Bread	17	Quinoa	22	Croissants
3	Chia seeds	8	Tortilla wraps	13	Bagels	18	Popcorn	23	Soy milk
4	Tofu	9	Frozen onion rings	14	Meat-free sausages	19	Granola	24	Porridge (oatmeal)
5	Beer	10	Almonds	15	Meat-free burger	20	Shortbread biscuits	25	Apple pie

As the above tables show, when using our weighted formula, the number one most ecologically impactful food item is Beef Steaks, with other forms of beef products following. Considering the large margin shown in our plots, this should be of no surprise. It is interesting to see Sugar making such a high ranking, along with Cottage Cheese. If we draw our attention to the second chart, we see the least impactful food item is Ale. Considering its extremely low emissions to land plot, seeing it at a high ranking is not surprising either. Followed by various forms of seeds and the legendary Tofu.

Conclusion

In conclusion, our exploration of "The Banana Index" has provided valuable insights into the ecological impact of various common foods. The dataset, curated by *The Economist*, sheds light on the environmental consequences of food production. By examining crucial variables we aimed to understand the intricate relationship between food choices and their environmental footprint.

To build our ranking formula, we employed a correlation table, assigning weights to each variable based on its correlation to total emissions per kilogram. The resulting formula allowed us to generate rankings for each food item, revealing the most and least ecologically impactful choices.

It is important to note this analysis is not exhaustive, there are certainly other ways to explore the topic further. This analysis is only a step towards creating awareness and promoting informed data-driven decision-making regarding food choices. By providing a unique perspective through "The Banana Index," I hope to encourage a broader conversation about sustainable food consumption and its implications on our planet.

Appendix

Cited Works

"Visualizing the Carbon Impact of Foods Relative to Bananas", Morgan Stevens, April 20, 2023

https://datainnovation.org/2023/04/visualizing-the-carbon-impact-of-foods-relative-to-bananas/

Code

```
########Setup##################
#install.packages("readr")
#install.packages("dplyr")
library(readr)
library(dplyr)
setwd("D:/Banana_Index")
banana <- read_csv("bananaindex.csv")


########clean up phase##################
str(banana) #Review table
banana <- banana[,c(-2,-17:-11)] #remove unneeded columns
sapply(banana, function(x) sum(is.na(x))) #check for NAs
banana[is.na(banana)] <- 0 #set NAs to 0
banana_stand <- banana #scale all variables
for(i in 2:9){
  banana_stand[,i] <- scale(banana_stand[,i])
}
rm(i)
########Building Rankings##################
CorTable <- cor(banana[,-c(1)])#create a correlation table
write.csv(CorTable, "D:/Banana_Index/CorTable.csv")
banana_stand$Rank <- #building Formula and add ranking column
  (CorTable[1,1]/sum(CorTable[,1]))*banana_stand$emissions_kg +
  (CorTable[2,1]/sum(CorTable[,1]))*banana_stand$land_use_kg +
  (CorTable[3,1]/sum(CorTable[,1]))*banana_stand$land_use_1000kcal +
```

```
  (CorTable[4,1]/sum(CorTable[,1]))*banana_stand$emissions_1000kcal +
  (CorTable[5,1]/sum(CorTable[,1]))*banana_stand$`Land use per 100 grams of protein` +
  (CorTable[6,1]/sum(CorTable[,1]))*banana_stand$`Land use per 100 grams of fat` +
  (CorTable[7,1]/sum(CorTable[,1]))*banana_stand$emissions_100g_fat +
  (CorTable[8,1]/sum(CorTable[,1]))*banana_stand$emissions_100g_protein

banana_stand <- banana_stand[order(-banana_stand$Rank),]# build worst 15 list
rank_w25 <- banana_stand[1:25,]
banana_stand <- banana_stand[order(banana_stand$Rank),]# build worst 15 list
rank_b25 <- banana_stand[1:25,]

#######Charting&Tables#################

#Total emissions to Total land for 25 worst and 25 best
rank_w25 <- rank_w25[order(-rank_w25$emissions_kg),] #sort worst 25 by total emissions
plot(rank_w25$emissions_kg, rank_w25$land_use_kg, type = "p", main="25 Highest Emissions to Land",
  xlab="Emissions Per Kilogram", ylab="Land Per Kilogram")
  points(rank_w25$emissions_kg[1:3], rank_w25$land_use_kg[1:3], pch=16, col=c("#4aacc5","#8064a1",
"#c0504e"),lwd = 3)
  legend("topleft",fill = c("#4aacc5","#8064a1", "#c0504e"),
    legend = c(rank_w25$entity[1],rank_w25$entity[2],rank_w25$entity[3]))

  rank_b25 <- rank_b25[order(rank_b25$emissions_kg),] #sort Best 25 by total emissions
plot(rank_b25$emissions_kg, rank_b25$land_use_kg, type = "p", main="25 Lowest Emissions to Land",
  xlab="Emissions Per Kilogram", ylab="Land Per Kilogram")
  points(rank_b25$emissions_kg[1:3], rank_b25$land_use_kg[1:3], pch=16, col=c("#4aacc5","#8064a1",
"#c0504e"),lwd = 3)
  legend("topright",fill = c("#4aacc5","#8064a1", "#c0504e"),
    legend = c(rank_b25$entity[1],rank_b25$entity[2],rank_b25$entity[3]))

#emissions per 1000 kilocalories to land per 1000 kilocalories for 25 worst and 25 best
```

rank_w25 <- rank_w25[order(-rank_w25$emissions_1000kcal),] #sort worst 25 per 1000 kilocalorie

plot(rank_w25$emissions_1000kcal, rank_w25$land_use_1000kcal, type = "p", main="25 Highest Emissions to Land per 1000 Kilocalories",

xlab="Emissions Per 1000 Kilocalories", ylab="Land Per 1000 Kilocalories")

points(rank_w25$emissions_1000kcal[1:3], rank_w25$land_use_1000kcal[1:3], pch=16, col=c("#4aacc5","#8064a1", "#c0504e"),lwd = 3)

legend("topleft",fill = c("#4aacc5","#8064a1", "#c0504e"),

legend = c(rank_w25$entity[1],rank_w25$entity[2],rank_w25$entity[3]))

rank_b25 <- rank_b25[order(rank_b25$emissions_1000kcal),] #sort Best 25 by 1000 kilocalorie

plot(rank_b25$emissions_1000kcal, rank_b25$land_use_1000kcal, type = "p", main="25 Lowest Emissions to Land Per 1000 Kilocalories",

xlab="Emissions Per 1000 Kilocalories", ylab="Land Per 1000 Kilocalories")

points(rank_b25$emissions_1000kcal[1:3], rank_b25$land_use_1000kcal[1:3], pch=16, col=c("#4aacc5","#8064a1", "#c0504e"),lwd = 3)

legend("topleft",fill = c("#4aacc5","#8064a1", "#c0504e"),

legend = c(rank_b25$entity[1],rank_b25$entity[2],rank_b25$entity[3]))

#emissions per 100 grams of protein to land per 100 grams of protein for 25 worst and 25 best

rank_w25 <- rank_w25[order(-rank_w25$emissions_100g_protein),] #sort worst 25 per 100 grams of protein

plot(rank_w25$emissions_100g_protein, rank_w25$`Land use per 100 grams of protein`, type = "p", main="25 Highest Emissions Per 100 Grams of Protein",

xlab="Emissions Per 100 Grams of Protein", ylab="Land Per 100 Grams of Protein")

points(rank_w25$emissions_100g_protein[1:3], rank_w25$`Land use per 100 grams of protein`[1:3], pch=16, col=c("#4aacc5","#8064a1", "#c0504e"),lwd = 3)

legend("topleft",fill = c("#4aacc5","#8064a1", "#c0504e"),

legend = c(rank_w25$entity[1],rank_w25$entity[2],rank_w25$entity[3]))

rank_b25 <- rank_b25[order(rank_b25$emissions_100g_protein),] #sort Best 25 per 100 grams of protein

```
plot(rank_b25$emissions_100g_protein, rank_b25$`Land use per 100 grams of protein`, type = "p",
main="25 Lowest Emissions Per 100 Grams of Protein",
  xlab="Emissions Per 100 Grams of Protein", ylab="Land Per 100 Grams of Protein")
  points(rank_b25$emissions_100g_protein[1:3], rank_b25$`Land use per 100 grams of protein`[1:3],
pch=16, col=c("#4aacc5","#8064a1", "#c0504e"),lwd = 3)
  legend("topleft",fill = c("#4aacc5","#8064a1", "#c0504e"),
    legend = c(rank_b25$entity[1],rank_b25$entity[2],rank_b25$entity[3]))


#emissions per 100 grams of fat to land per 100 grams of fat for 25 worst and 25 best
rank_w25 <- rank_w25[order(-rank_w25$emissions_100g_fat),] #sort worst 25 per 100 grams of fat
plot(rank_w25$emissions_100g_fat, rank_w25$`Land use per 100 grams of fat`, type = "p", main="25
Highest Emissions Per 100 Grams of Fat",
  xlab="Emissions Per 100 Grams of Fat", ylab="Land Per 100 Grams of Fat")
points(rank_w25$emissions_100g_fat[1:3],rank_w25$`Land use per 100 grams of fat`[1:3], pch=16,
col=c("#4aacc5","#8064a1", "#c0504e"),lwd = 3)
legend("topleft",fill = c("#4aacc5","#8064a1", "#c0504e"),
  legend = c(rank_w25$entity[1],rank_w25$entity[2],rank_w25$entity[3]))


rank_b25 <- rank_b25[order(rank_b25$emissions_100g_fat),] #sort Best 25 per 100 grams of fat
plot(rank_b25$emissions_100g_fat, rank_b25$`Land use per 100 grams of fat`, type = "p", main="25
Lowest Emissions Per 100 Grams of Fat",
  xlab="Emissions Per 100 Grams of Fatt", ylab="Land Per 100 Grams of Fat")
  points(rank_b25$emissions_100g_fat[1:3],rank_b25$`Land use per 100 grams of fat`[1:3], pch=16,
col=c("#4aacc5","#8064a1", "#c0504e"),lwd = 3)
  legend("topleft",fill = c("#4aacc5","#8064a1", "#c0504e"),
  legend = c(rank_b25$entity[1],rank_b25$entity[2],rank_b25$entity[3]))


#Top 25 worst and best using our custome formula
rank_b25 <- rank_b25[order(rank_b25$Rank),]
rank_w25 <- rank_w25[order(-rank_w25$Rank),]
#export to csv for table
```

```
write.csv(print(rank_b25[,1],n=25), "D:/Banana_Index/25_Best.csv", row.names=FALSE)

write.csv(print(rank_w25[,1],n=25), "D:/Banana_Index/25_Worst.csv", row.names=FALSE)
```