Food Consumption and CO2 Emissions

Brendan Octaviano

23/01/2022

Introduction

Carbon Dioxide, or more commonly referred to as CO2, is a greenhouse gas resulting from a number of human activities. When overproduced, it contributes to what is known as the greenhouse gas effect - a natural process that warms the Earth's surface. What generally happens is when the sun radiates heat and energy, it is either reflected back into space, or it penetrates the earth's atmosphere, trapped by greenhouse gas to help keep Earth warm enough to sustain life. However, due to the overproduction of greenhouse gases like CO2, this contributes to the levels of greenhouse gas, trapping unnecessary heat within our atmosphere.

This is the underlying cause of what is known as climate change, and is resulting in detrimental effects to the Earth's climate. One contributor to climate change is the effect of food production. According to a report from Ritchie and Roser (2020), one-quarter of the world's greenhouse gas emissions come from the production of food. This suggests that turning our attention on both food and food consumption is integral to targeting the underlying issue of climate change.

Our Data

nu3, a German company focusing on supplying supplements and nutritional food products to European countries, have released a dataset exploring food consumption for different food categories, along with their corresponding CO2 emissions, for 130 different countries in 2018. As a result, by utilising the data available, we would like to explore the following questions:

1. Which countries should we focus on because of their CO2 emissions, both as a whole, as well as for different food categories?
2. Does non-animal product consumption minimise the effects of CO2 emissions from food consumption?

Fortunately, this data was scraped and cleaned by Kasia Kulma, so no further cleaning was required.

```
#Set seed for reproducability
set.seed(1)

#Loading our packages
library(tidyverse)
library(repr)
library(ggrepel)
library(knitr)

#Loading the data
food.consumption <- read_csv(
  'https://raw.githubusercontent.com/rfordatascience/tidytuesday/master/data/2020/2020-02-18/food_consu

head(food.consumption) %>%
  kable()
```

country	food_category	consumption	co2_emmission
Argentina	Pork	10.51	37.20
Argentina	Poultry	38.66	41.53
Argentina	Beef	55.48	1712.00
Argentina	Lamb & Goat	1.56	54.63
Argentina	Fish	4.36	6.96
Argentina	Eggs	11.39	10.46

Methodology: K-Means Clustering

Through the use of K-Means clustering, we are aiming to group our countries and food categories together so that we can effectively answer the questions we have posed.

K-Means clustering is an unsupervised machine learning technique that aims to partition a datasets observations into K different clusters thorough multiple iterations, in such a way that each observation belongs to only a single group that has similar qualities. Clustering in general takes on an important concept of distance. Distance, in a literal sense, measures how far one thing is from another, however in the world of machine learning, it refers to how similar (or dissimilar) two observations are. The larger the distance between two observations, the less similar they are, and vice versa.

By calculating the Euclidian distance between objects based on defined variables, K-means looks to minimise the "within sum of squares", while also maximising the "between sum of squares". In other words, we are looking to cluster our observations in such a way where are minimising the within-cluster dispersion, and maximise the between-cluster dispersion, ensuring that we have clear segments.

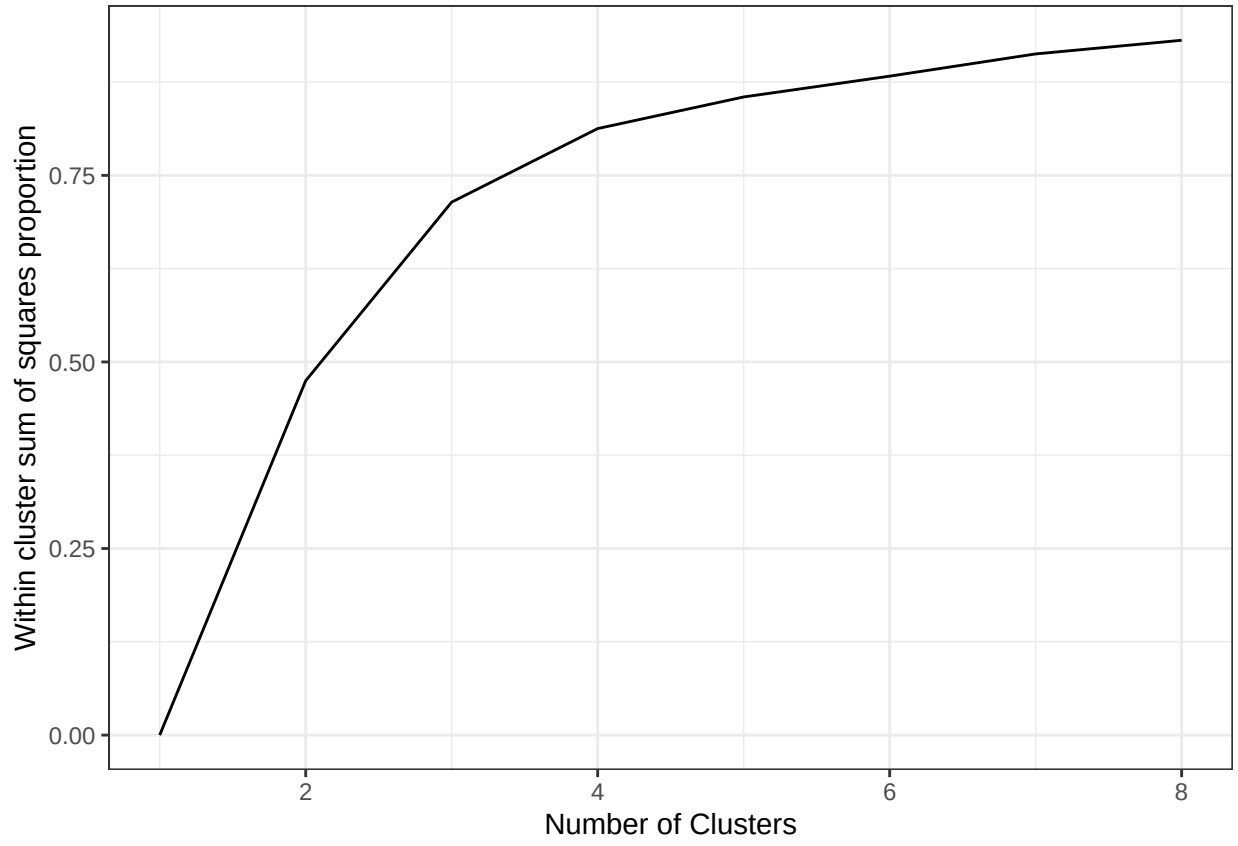
Applying K-Means clustering to our data will give us the ability to group our countries that have similar levels of consumption per food group, as well as CO2 emissions, so we can see firstly which countries we should turn our attention to, as well as what food categories in countries we should focus on.

Evaluating Model

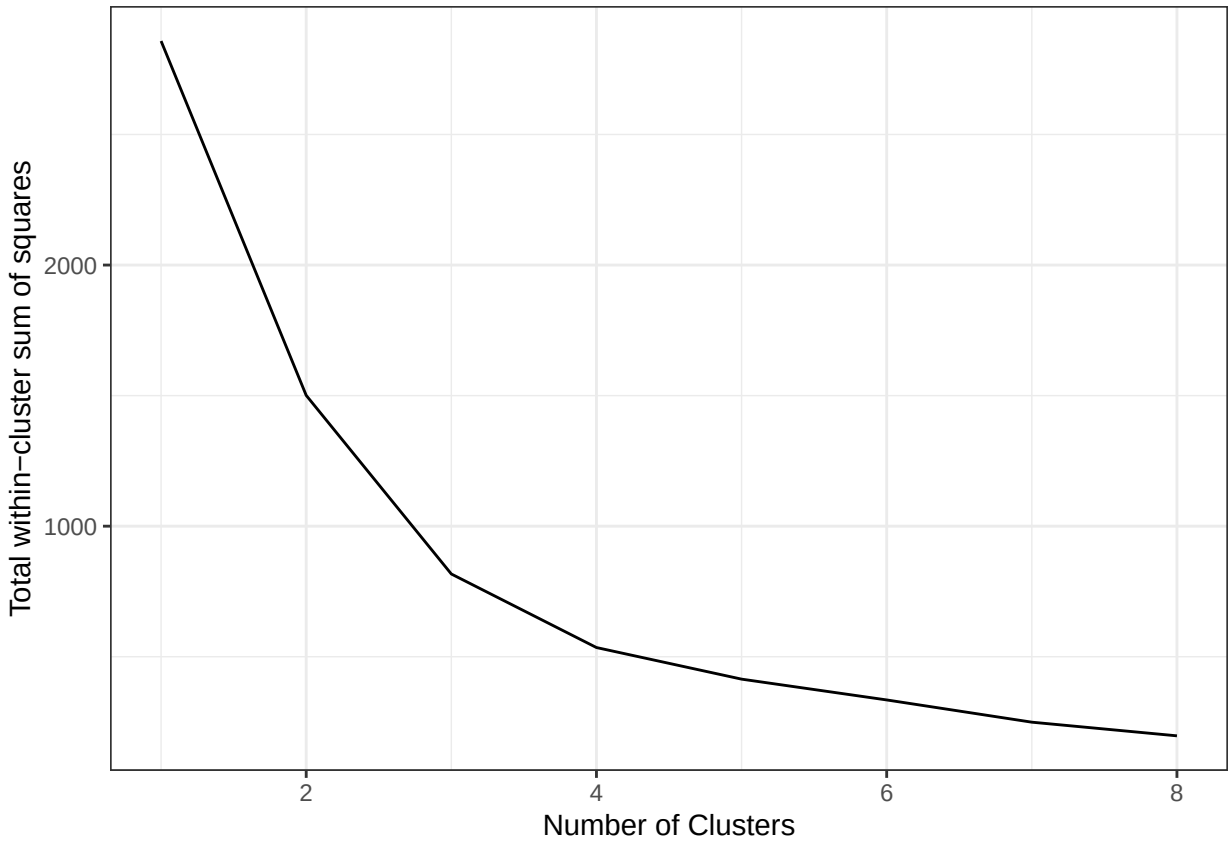
After running our K-Means clustering with a number of different clusters using a for loop, we are able to visibly see which number of clusters is optimal, where optimal is a balance between minimising the "within sum of squares", while also maximising the "between sum of squares".

```
food.consumption.scaled <- food.consumption %>%
  #Selecting only our numerical columns in the data so that we can cluster based
  #on those values
  select(consumption, co2_emmission) %>%
  #Important to scale numerical values. While looking at the metadata, the values
  #are already standardised, however consumption tends to be on a much
  #lower scale compared to Co2 emissions
  scale

#In order to choose an optimal number of clusters, we can use a for loop and calculate
#the total within sum of squares, as well as the between
#sum of squares / total sum of squares

n <- 8
ss.df <- data.frame(k = numeric(n),
                    ss = numeric(n),
                    tot = numeric(n))
```

```r
for(i in 1:n){
  km <- kmeans(food.consumption.scaled, centers = i, nstart = 25)
  ss.df$k[i] <- i
  ss.df$ss[i] <- km$betweenss / km$totss
  ss.df$tot[i] <- km$tot.withinss
}

#Plotting these against the corresponding number of clusters wil allow us to decide
#the optimal number of clusters for our analysis
ss.df %>%
  ggplot(aes(x = k,
             y = ss)) +
  geom_line() +
  labs(x = "Number of Clusters",
       y = "Within cluster sum of squares proportion") +
  theme_bw()
```




```r
ss.df %>%
  ggplot(aes(x = k,
             y = tot)) +
  geom_line() +
  labs(x = "Number of Clusters",
       y = "Total within-cluster sum of squares") +
  theme_bw()
```



As a general rule of thumb, we can choose the number of clusters where the gradient begins to flatten looking at both plots, which is when we choose 4 clusters. This was the case for both of our K-Means clustering, which will be detailed below.

Furthermore, after assessing both of our K-Means clustering, our *between sum of squares to total sum of squares ratio was both above 0.8, or 80%*. This value tells us how good the clustering that our K-Means has found, giving us a measure of the total variance in our data set that is explained by our clustering, indicating that we have split our data fairly accurately.

The two K-Means clustering we performed was for the purpose of answering our first question:

Which countries should we focus on because of their CO2 emissions, both as a whole, as well as for different food categories?

Analysis for countries as a whole

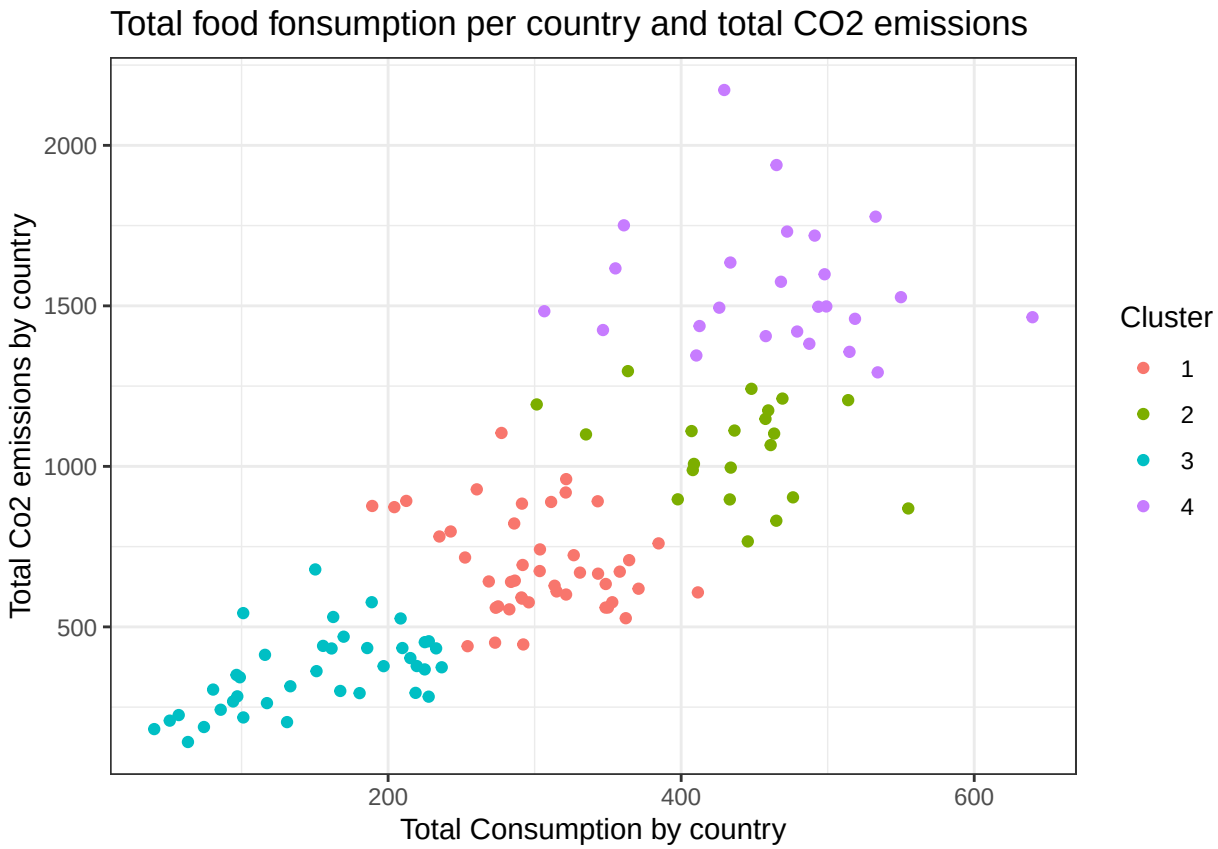
Creating our first K-Means clusters allows us to see very clearly the countries that emit the highest amount of CO2. This was achieved by adding up each of the listed CO2 emissions for each country, and creating our clusters based on consumption and emission. It is clear that those countries in cluster 3 are ones that we should focus on, since they are producing a large amount of CO2 emissions – with their average being 1560.16kg/person/year.

```
#Grouping our countries and calculating totals for consumption and CO2 Emissions so
#we can run our K-Means clustering based on these variables
food.consumption.by.country <- food.consumption %>%
  group_by(country) %>%
  summarise(total_consumption = sum(consumption),
            total_emission = sum(co2_emmission))
```

```r
#Again, important to scale our data
food.consumption.country.scaled <- food.consumption.by.country %>%
  select(total_consumption, total_emission) %>%
  scale

#Again, 4 clusters are the most optimal
km.country.final <- kmeans(food.consumption.country.scaled, centers = 4, nstart = 25)

tibble(
  "Between sum of squares to total sum of squares ratio" =
    km.country.final$betweenss / km.country.final$totss) %>%
  kable()
```

Between sum of squares to total sum of squares ratio
0.8528108

```r
food.consumption.country.clusters <- food.consumption.by.country %>%
  mutate(Cluster = as.factor(km.country.final$cluster))

food.consumption.country.clusters %>%
  ggplot(aes(x = total_consumption,
             y = total_emission,
             color = Cluster)) +
  geom_point() +
  labs(x = "Total Consumption by country",
       y = "Total Co2 emissions by country",
       title = "Total food fonsumption per country and total CO2 emissions") +
  theme_bw()
```




```
#Calculating the mean consumption for each cluster, along with the mean
#Co2 emissions for each cluster
food.consumption.country.clusters %>%
  select(-country) %>%
  group_by(Cluster) %>%
  summarise_all(mean) %>%
  arrange(desc(total_emission)) %>%
  kable()
```

Cluster	total_consumption	total_emission
4	463.3288	1560.1632
2	435.2190	1053.1590
1	302.6400	694.6707
3	150.1605	358.6797

This is gives us a large group of countries:

```
#Listing the countries in cluster who emit the highest amount of CO2 as a whole
countries.to.target <- food.consumption.country.clusters %>%
  filter(Cluster == 2) %>%
  select(country) %>%
  rename("Country" = country)

countries.to.target %>%
  kable()
```

Country
Armenia
Austria
Bahamas
Belgium
Chile
Croatia
Estonia
Germany
Italy
Lithuania
Maldives
Malta
Oman
Portugal
Romania
Russia
Slovenia
Spain
Turkey
United Arab Emirates
United Kingdom

Analysis for countries and specific food categories

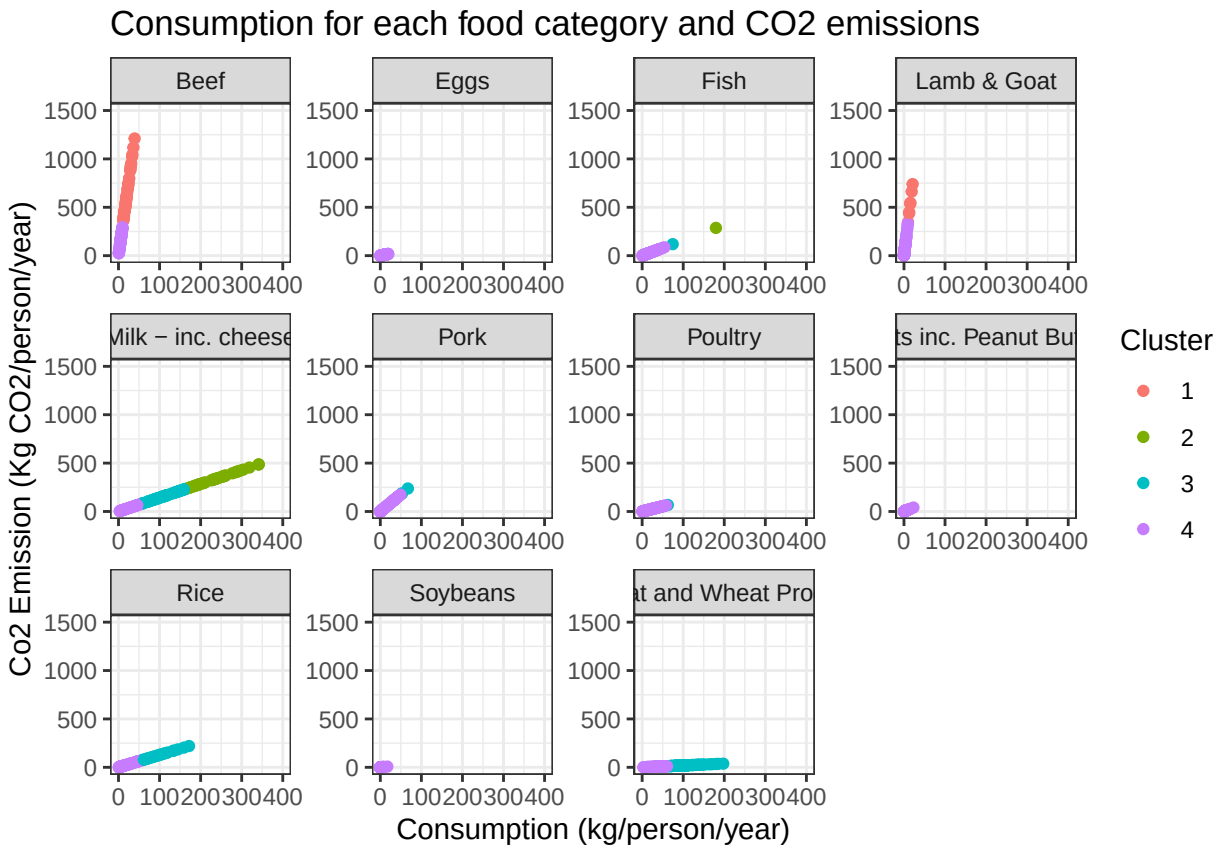
However, we want to make use of the data we have and look at countries and their individual food categories. Again, running our K-Means clustering based on both total consumption and CO2 emissions per country and food categories, we are able to see what food categories in each country are responsible for high levels of CO2 emissions.

```
#Assessing the plots above, we can see that 4 clusters is the optimal number
km.final <- kmeans(food.consumption.scaled, centers = 4, nstart = 25)

tibble(
  "Between sum of squares to total sum of squares ratio" =
    km.final$betweenss / km.final$totss) %>%
  kable()
```

Between sum of squares to total sum of squares ratio
0.8126998

```
#After getting our clusters, we are now appending it to our original dataframe
food.consumption.clusters <- food.consumption %>%
  mutate(Cluster = as.factor(km.final$cluster))

#Plotting out data
food.country.plot <- food.consumption.clusters %>%
  mutate(food_category = factor(
    food_category,
    levels = c(
      "Beef", "Eggs", "Fish", "Lamb & Goat", "Milk - inc. cheese", "Pork", "Poultry", "Nuts inc. Peanut
```

```
  ggplot(aes(x = consumption,
             y = co2_emmission,
             colour = Cluster)) +
  geom_point() +
  facet_wrap(~food_category,
             scales = "free") +
  scale_y_continuous(limits = c(0, 1500)) +
  scale_x_continuous(limits = c(0, 400)) +
  labs(x = "Consumption (kg/person/year)",
       y = "Co2 Emission (Kg CO2/person/year)",
       title = "Consumption for each food category and CO2 emissions") +
  theme_bw()

food.country.plot
```




```
#Calculating the mean consumption for each cluster, along with the mean
#Co2 emissions for each cluster
food.country.cluster.analysis <- food.consumption.clusters %>%
  select(-country, -food_category) %>%
  group_by(Cluster) %>%
  summarise_all(mean) %>%
  arrange(desc(co2_emmission))

food.country.cluster.analysis %>%
  kable()
```

Cluster	consumption	co2_emmission
1	20.50238	639.67460
2	242.95000	346.83175
3	105.99878	89.88864
4	11.53086	33.03619

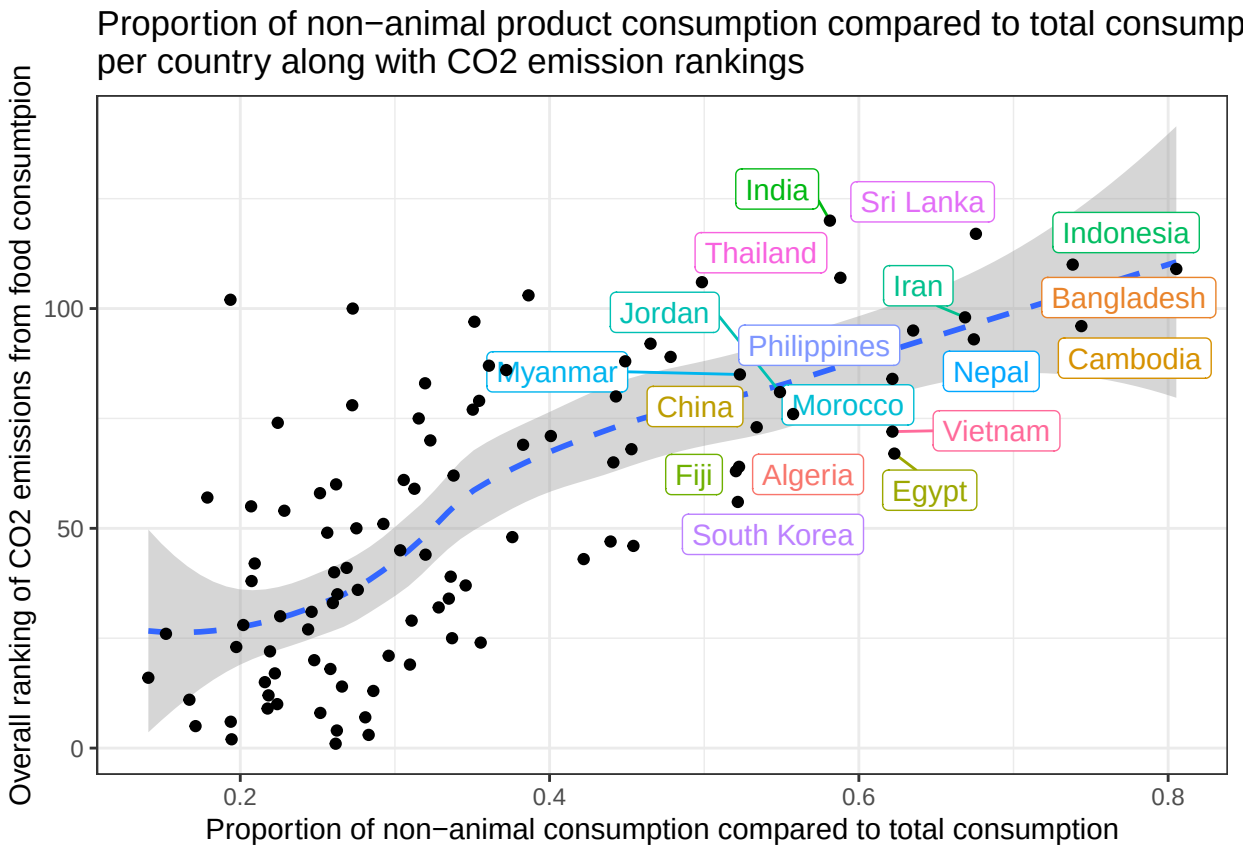
It is evident from our plots for each food category that countries and food categories, cluster 4 are responsible for a high number of CO2 emissions. The food category that is primarily included in this cluster is Beef, but we can also see some Lamb & Goat, from very large number of countries, similar to above. It was found that in this cluster, for each country and food category combination, there was an average of approximately *20kg/person/year consumed, resulting in almost 640kg/person/year of CO2 emissions*, while all other clusters exhibited a considerably smaller proportion (cluster 1 was the next largest, showing only a 3 times difference between the particular food category consumption and CO2 emissions).

Analysis of non-animal product consumption

To answer our second research question, it was important to firstly look at the countries that consumed the highest number of non-animal products. In order to account for different levels of consumption between countries, looking at the *proportion of non-animal consumption compared to total consumption* was determined to be a more meaningful measure. By plotting 100 of our 130 countries below, we are able to see the relationship between the proportion of their non-animal consumption compared to total consumption, and their overall ranking in terms of total CO2 emissions from food consumption shows a positive relationship between the two. Labels were also added to countries who had a proportion of non-animal consumption compared to total consumption greater than 0.5. To strengthen this, we have also calculated Pearson's correlation coefficient between the two variables, attaining a value of 0.71, indicating a moderately strong positive relationship.

```
#Ranking each country based on total CO2 emissions
country.rankings <- food.consumption.by.country %>%
  arrange(desc(total_emission)) %>%
  select(country) %>%
  mutate(ranking = seq(from = 1, to = 130))

#By removing the bottom 30 based on consumption, we are accounting only for countries
#who consume a reasonable amount for the purpose of our analysis
top.countries.total.emission <- arrange(
  food.consumption.by.country, desc(total_consumption))[1:100, ]

#We will now calculate for all countries which countries consume the most
#amount of non-animal products
top.nonanimal.consumption <- food.consumption %>%
  #Only looking at countries who consume a reasonable amount for the
  #purpose of our analysis
  filter(country %in% top.countries.total.emission$country) %>%
  filter(
    food_category == "Wheat and Wheat Products" | food_category == "Rice" | food_category == "Soybeans"
  group_by(country) %>%
  #Getting our total non-animal consumption for each country
  summarise(total_nonanimal_consumption = sum(consumption))%>%
  #Adding our total consumption calculated eariler into this table
  left_join(food.consumption.by.country) %>%
  #Calculating the proportion of non-animal consumption compared to total consumption
  mutate(non_animal_proportion = total_nonanimal_consumption / total_consumption) %>%
```

```
  select(country, non_animal_proportion) %>%
  #Sort them based on total non-animal product consumption
  arrange(desc(non_animal_proportion))

#By completing a left join, we are able to compare non-animal products
#consumption and their rankings compared to all countries in terms of CO2 emissions
non.animal.proportion.rankings <- left_join(
  top.nonanimal.consumption, country.rankings) %>%
  arrange(desc(non_animal_proportion))

#Plotting the data
non.animal.proportion.rankings.plot <- non.animal.proportion.rankings %>%
  ggplot(aes(x = non_animal_proportion,
             y = ranking)) +
  geom_smooth(stat = "smooth",
              linetype = "dashed") +
  geom_label_repel(data = subset(
    non.animal.proportion.rankings, non_animal_proportion > 0.5),
                   aes(label = country,
                       color = country)) +
  geom_point() +
  labs(x = "Proportion of non-animal consumption compared to total consumption",
       y = "Overall ranking of CO2 emissions from food consumtpion",
       title = "Proportion of non-animal product consumption compared to total consumption \nper country
  theme_bw()

non.animal.proportion.rankings.plot + theme(legend.position = "none")
```




```
#Calculating Pearson's correlation coefficient
tibble(
  "Pearson's Correlation Coefficient" = cor(
    non.animal.proportion.rankings$non_animal_proportion,
    non.animal.proportion.rankings$ranking)) %>%
  kable()
```

Pearson's Correlation Coefficient
0.7088274

Conclusion

After performing our analysis, we were able to gain a clear idea of what countries and food categories were responsible for a considerable amount of CO2 emissions. While we were able to gain a list of countries that exhibited high levels of CO2 emissions, what was more insightful to focus on was the different food categories. Beef, lamb and goat were the food categories that were responsible for high levels of CO2 emissions in a number of countries.

Moreover, there is a clear relationship between the proportion of non-animal consumption compared to total food consumption, along with CO2 emissions. Our analysis strongly suggests that countries who consume more non-animal products tend to emit lower amounts of CO2 relative to other countries.

With this analysis we hope that countries can understand the importance of their food choices, and ideally encourage people to consider what they regularly to consume in hopes of slowing down climate change. This could be achieved through a number of different ways, including *decreasing the price of non-animal products, raising awareness on the benefits of consuming non-animal products for both the individual as well as the*

climate, and also improving accessibility of these products.