Exploring Food Waste Data 🍏

Josh Houlding

2023-09-10

Introduction 📚

Food waste (https://en.wikipedia.org/wiki/Food_loss_and_waste), a pervasive global issue, refers to the disposal or wastage of edible food that is still fit for consumption. This phenomenon not only squanders precious resources but also has severe environmental, social, and economic implications. It is estimated that nearly one-third of all food produced for human consumption worldwide is lost or wasted. Such wastage contributes to greenhouse gas emissions, depletes natural resources, and exacerbates issues of food insecurity and hunger. Various factors contribute to this problem, including consumer behavior, inadequate storage, and inefficient supply chains. Understanding the complexities surrounding food waste is crucial for implementing effective strategies to mitigate its impact on both a local and global scale.

Objective 🎯

The primary objective of this project is to conduct a comprehensive exploratory data analysis on a food waste dataset sourced from Kaggle. Through this analysis, we aim to uncover valuable insights, identify significant trends, and gain a deep understanding of the factors contributing to food waste. Our goals include quantifying the extent of food waste, examining its temporal and spatial distribution, identifying key drivers of food waste, and exploring potential correlations with various socioeconomic and environmental factors. Ultimately, this EDA will enable us to make data-driven recommendations for reducing food waste and promoting sustainable practices in the food industry.

The data 📊

This dataset (https://www.kaggle.com/datasets/joebeachcapital/food-waste) contains information about food waste from every country worldwide. It has been collected from different sources, such as retailers, households, and restaurants. The data includes food waste totals, both per-capita and absolute numbers, as well as totals for household, retail and food service sources respectively. The dataset can be used to analyze the patterns and causes of food waste, as well as to propose solutions to reduce it. The dataset was created by Joe Beach Capital, based on various public datasets available on Kaggle and other platforms, and a visualization on informationisbeautiful.net (https://informationisbeautiful.net/visualizations/food-waste/) is also cited. This food waste data is from 2021.

The dataset itself is a single CSV file containing food waste data from 214 countries.

Limitations

- The data is limited to 2021, so no temporal data is available for tracking trends over time.
- The dataset does not include population data, so that will need to be acquired from another source at a later time if necessary.

Loading the data 🔄

```
# Load packages
library(tidyverse)
library(dplyr)
library(sqldf)
library(ggplot2)
library(readr)
library(knitr)

# Read the CSV file
foodWasteData <- read_csv("foodwastedata.csv")
kable(head(foodWasteData), caption="foodWasteData")
```

foodWasteData

Country	combined figures (kg/capita/year)	Household estimate (kg/capita/year)	Household estimate (tonnes/year)	Retail estimate (kg/capita/year)	Retail estimate (tonnes/year)	Food service estimate (kg/capita/year)	Food service estimate (tonnes/year)	Confidence in estimate	M49 code	Region	Source
Afghanistan	126	82	3109153	16	594982	28	1051783	Very Low Confidence	4	Southern Asia	https://www.une food-waste-inde (https://www.un food-waste-inde
Albania	127	83	238492	16	45058	28	79651	Very Low Confidence	8	Southern Europe	https://www.une food-waste-inde (https://www.un food-waste-inde
Algeria	135	91	3918529	16	673360	28	1190335	Very Low Confidence	12	Northern Africa	https://www.une food-waste-inde (https://www.un food-waste-inde
Andorra	123	84	6497	13	988	26	1971	Low Confidence	20	Southern Europe	https://www.une food-waste-inde (https://www.un food-waste-inde
Angola	144	100	3169523	16	497755	28	879908	Very Low Confidence	24	Sub-Saharan Africa	https://www.une food-waste-inde (https://www.un food-waste-inde
Antigua and Barbuda	113	74	7178	13	1244	26	2483	Low Confidence	28	Latin America and the Caribbean	https://www.une food-waste-inde (https://www.un food-waste-inde

Cleaning the data 🧹

Creating a new column

I created a new column for combined estimate in tonnes/year, since this seemed important and there was no column for it in the original dataset.

```
# Create new column for combined estimate in tonnes/year
foodWasteData <- mutate(foodWasteData, combined_estimate_tonnes_per_year = foodWasteData$`Household estimate (tonnes/year)`
+ foodWasteData$`Retail estimate (tonnes/year)` + foodWasteData$`Food service estimate (tonnes/year)`)
foodWasteData <- foodWasteData[, c(1,2,13,3,4,5,6,7,8,9,10,11,12)]
```

Renaming the columns

I changed all column names to lowercase and put underscores in them so they would be easier to feed into the sqldf package for the purposes of running SQL queries on the data.

```
# Changing column names to make them lowercase and include underscores
new_col_names <- c("country", "combined_estimate_kpcpy", "combined_estimate_tpy", "household_estimate_kpcpy", "household_est
imate_tpy", "retail_estimate_kpcpy", "retail_estimate_tpy", "food_service_estimate_kpcpy", "food_service_estimate_tpy", "est
imate_confidence", "m49_code", "region", "source")
names(foodWasteData) <- new_col_names
```

Moving the source column to reduce clutter

I didn't need the source column for my analysis, so I decided to move it instead of deleting it so it would be there if I needed it later on. The links suggested the data is originally from the UN Environment Programme, but the links in the dataset lead to pages where the data had been removed for some reason.

```
# Get source column out of the way
dataSources <- sqldf("SELECT country, m49_code, source FROM foodWasteData")
foodWasteData<- foodWasteData[, -13]
```

Analyzing the data 🔬

How much food is wasted every year in total?

```
# Find total global food waste every year
total_foodwaste <- sqldf("SELECT SUM(combined_estimate_tpy) FROM foodWasteData")[[1]]
print(total_foodwaste)
```

```
## [1] 930857271
```

In total, a staggering 930,857,271 tonnes of food is wasted every year. This is nearly a billion tonnes *every single year*.

According to Healthline (https://www.healthline.com/health/mens-health/average-weight-for-men), the average adult weighs 136.7 pounds, or 62 kg.

Thus, our total annual food waste is equivalent to:

- **15,014,000,000 (>15 billion) human adults.**
- **677,480,000 (>677 million) 2021 Toyota Corollas.**
- **209,420,000 (>209 million) African elephants.**
- **9,402.6 American Nimitz-class aircraft carriers.**

Truly staggering numbers in perspective.

Which countries have the highest and lowest combined per-capita food waste?

```
# Find the country with the highest combined per-capita food waste
highestFoodWasteKpcpy <- sqldf("SELECT country, MAX(combined_estimate_kpcpy), estimate_confidence FROM foodWasteData")
kable(head(highestFoodWasteKpcpy), align="l")
```

country	MAX(combined_estimate_kpcpy)	estimate_confidence
Malaysia	260	Medium Confidence

Malaysia has the highest combined per-capita food waste rate in the world at 260 kg/capita/year.

According to NPR (https://www.npr.org/sections/thesalt/2011/12/31/144478009/the-average-american-ate-literally-a-ton-this-year), the average American eats 1,996 lbs (905.4 kg) of food per year. Assuming that is 3 meals per day and people in every country eat about the same amount annually, we end up with an estimate of 0.827 kg/meal.

Thus, the average Malaysian wastes a massive 314 meals' worth of food every year.

```
# Find the country with the lowest combined per-capita food waste
lowestFoodWasteKpcpy <- sqldf("SELECT country, MIN(combined_estimate_kpcpy), estimate_confidence FROM foodWasteData")
kable(head(lowestFoodWasteKpcpy), align="l")
```

country	MIN(combined_estimate_kpcpy)	estimate_confidence
Slovenia	61	Medium Confidence

At the other end of the spectrum, Slovenia has the lowest combined per-capita food waste of all countries, with its citizens wasting around 74 meals a year.

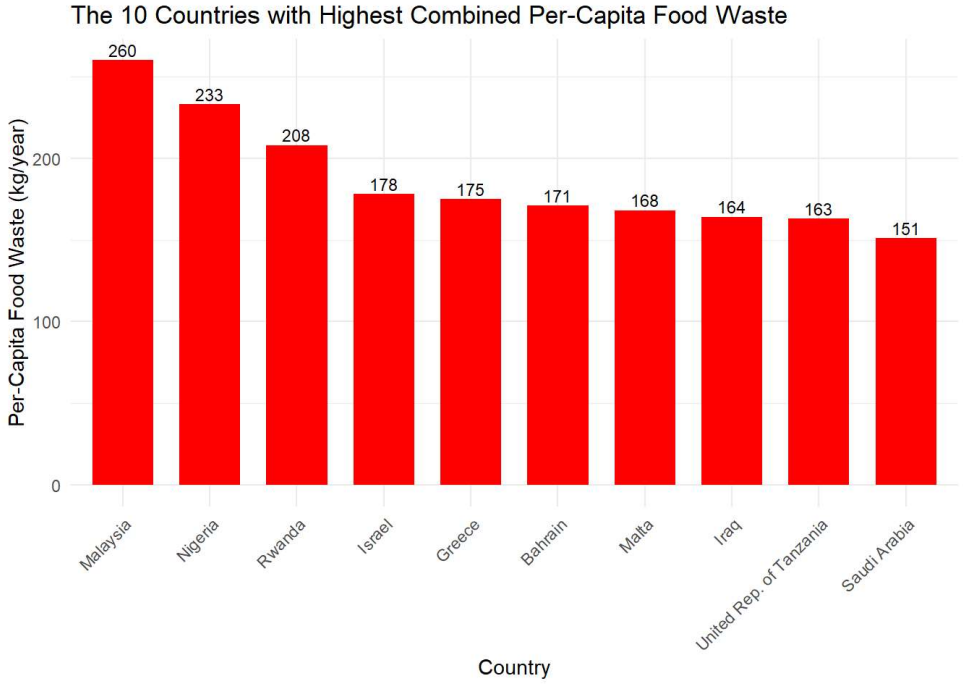
What are the top 10 and bottom 10 countries for combined per-capita food waste?

First, we aggregate the data.

```
# Subset the dataframe to select the top 10 countries with the highest and lowest combined per-capita food waste
topCountries <- tail(foodWasteData[order(foodWasteData$combined_estimate_kpcpy), ], 10)
bottomCountries <- head(foodWasteData[order(foodWasteData$combined_estimate_kpcpy), ], 10)
```

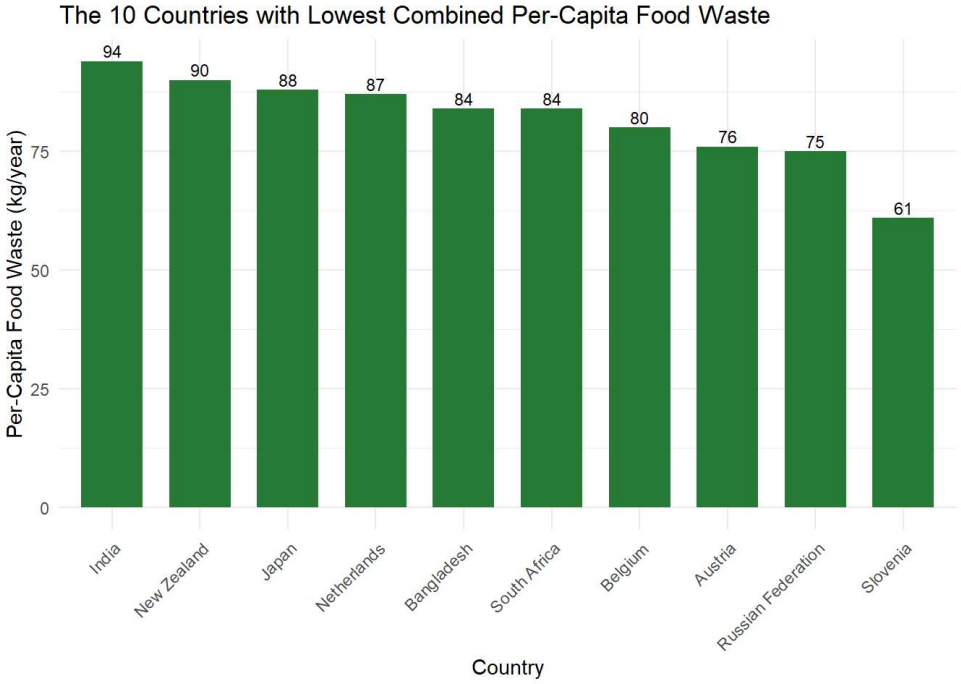
Then display it:

```
# Show the 10 countries with the highest combined per-capita food waste
ggplot(topCountries, aes(x = reorder(country, -combined_estimate_kpcpy), y = combined_estimate_kpcpy)) +
  geom_bar(stat = "identity", fill = "red", width = 0.7) +
  geom_text(aes(label = combined_estimate_kpcpy), hjust = 0.5, vjust = -0.3, size = 3) +  # Add labels
  labs(
    title = "The 10 Countries with Highest Combined Per-Capita Food Waste",
    x = "Country",
    y = "Per-Capita Food Waste (kg/year)"
  ) +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
  scale_fill_manual(
    values = c("blue" = "Combined"),
    name = "Waste Type",
    labels = c("Combined")
  ) +
  guides(fill = FALSE)  # Hide the legend
```



We can see that the top 10 countries are largely located in Africa and the Middle East, areas with lower levels of socioeconomic development.

```
# Show the 10 countries with the lowest combined per-capita food waste
ggplot(bottomCountries, aes(x = reorder(country, -combined_estimate_kpcpy), y = combined_estimate_kpcpy)) +
  geom_bar(stat = "identity", fill = "#277a36", width = 0.7) +
  geom_text(aes(label = combined_estimate_kpcpy), hjust = 0.5, vjust = -0.3, size = 3) +  # Add labels
  labs(
    title = "The 10 Countries with Lowest Combined Per-Capita Food Waste",
    x = "Country",
    y = "Per-Capita Food Waste (kg/year)"
  ) +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
  scale_fill_manual(
    values = c("blue" = "Combined"),
    name = "Waste Type",
    labels = c("Combined")
  ) +
  guides(fill = FALSE)  # Hide the legend
```



As expected, most of the countries with the lowest food waste per capita are in Europe, a highly-developed region of the world.

Which countries generate the most and least combined total food waste?

```
# Find the country with the highest combined total food waste
highestFoodWaste <- sqldf("SELECT country, MAX(combined_estimate_tpy), estimate_confidence FROM foodWasteData")
kable(head(highestFoodWaste), align="l")
```

country	MAX(combined_estimate_tpy)	estimate_confidence
China	179448659	Very Low Confidence

The country with the highest combined total food waste estimate is China, at over 179 million tonnes. However, this estimate could be inaccurate as very low confidence is given.

```
# Find the country with the lowest combined total food waste
lowestFoodWaste <- sqldf("SELECT country, MIN(combined_estimate_tpy), estimate_confidence FROM foodWasteData")
kable(head(lowestFoodWaste), align="l")
```

country	MIN(combined_estimate_tpy)	estimate_confidence
Nauru	1264	Low Confidence

Nauru is a small Micronesian island with a tiny population of 12,511 as of 2021, so it is expected that they would have a miniscule total amount of food waste.

What is the average combined per-capita food waste across all countries?

```
# Find the average combined per-capita food waste across all countries
avg_foodwaste_kpcpy <- sqldf("SELECT AVG(combined_estimate_kpcpy) FROM foodWasteData")[[1]]
print(avg_foodwaste_kpcpy)
```

```
## [1] 126.7944
```

The average person on Earth wastes about 126.8 kg of food every year, which is about 153 meals' worth given our estimate of 0.827 kg/meal earlier.

How does average per-capita food waste vary across regions?

I wanted to explore the variation in per-capita food waste across regions, and not just individual countries.

```
# Display all regions in the dataset
print(unique(foodWasteData$region))
```

```
##  [1] "Southern Asia"                  "Southern Europe"
##  [3] "Northern Africa"                "Sub-Saharan Africa"
##  [5] "Latin America and the Caribbean" "Western Asia"
##  [7] "Australia and New Zealand"      "Western Europe"
##  [9] "Eastern Europe"                 "Northern America"
## [11] "South-eastern Asia"             "Eastern Asia"
## [13] "Northern Europe"                "Melanesia"
## [15] "Polynesia"                      "Micronesia"
## [17] "Central Asia"
```

```
# Group the data by the 'region' column
foodWasteData <- foodWasteData %>%
  group_by(region)
# Calculate averages for each per-capita food waste column
averageFoodWasteByRegion <- foodWasteData %>%
  summarize(
    avg_combined_kpcpy = mean(combined_estimate_kpcpy, na.rm = TRUE),
    avg_household_kpcpy = mean(household_estimate_kpcpy, na.rm = TRUE),
    avg_retail_kpcpy = mean(retail_estimate_kpcpy, na.rm = TRUE),
    avg_food_service_kpcpy = mean(food_service_estimate_kpcpy, na.rm = TRUE)
    )
# Order the data by combined averages
averageFoodWasteByRegion <- averageFoodWasteByRegion[order(averageFoodWasteByRegion$avg_combined_kpcpy, decreasing = TRUE),
]
kable(head(averageFoodWasteByRegion, 17), align="l")
```

region	avg_combined_kpcpy	avg_household_kpcpy	avg_retail_kpcpy	avg_food_service_kpcpy
Sub-Saharan Africa	145.3958	101.64583	15.77083	27.97917
Western Asia	145.3889	101.05556	17.16667	27.16667
South-eastern Asia	137.4545	83.00000	21.18182	33.27273
Northern Africa	133.5000	89.50000	16.00000	28.00000
Central Asia	130.2000	86.20000	16.00000	28.00000
Melanesia	128.6000	85.60000	15.40000	27.60000
Micronesia	123.1429	82.00000	14.28571	26.85714
Southern Europe	123.0000	85.60000	12.60000	24.80000
Northern America	120.7500	71.50000	13.75000	35.50000
Polynesia	119.5000	76.75000	15.25000	27.50000
Latin America and the Caribbean	115.5476	73.57143	14.78571	27.19048
Southern Asia	113.1111	71.88889	16.00000	25.22222
Eastern Asia	113.0000	71.42857	13.71429	27.85714
Northern Europe	111.6667	74.66667	12.83333	24.16667
Australia and New Zealand	111.5000	81.50000	6.00000	24.00000
Eastern Europe	109.4000	68.10000	14.30000	27.00000
Western Europe	104.8889	67.11111	12.00000	25.77778

The three regions with the highest per-capita food waste are Sub-Saharan Africa, Western Asia and South-eastern Asia, at 145.4 kg/person/year for the former two and 137.5 kg/person/year for the latter. On the other hand, Western Europe is a very wealthy region and has the lowest per-capita food waste, at 104.9 kg/person/year. Northern America is near the middle of the pack, at 9th out of 17 regions. This was surprising, as I predicted that more economically-developed regions of the world would waste more. I was curious about this and wanted to learn more.

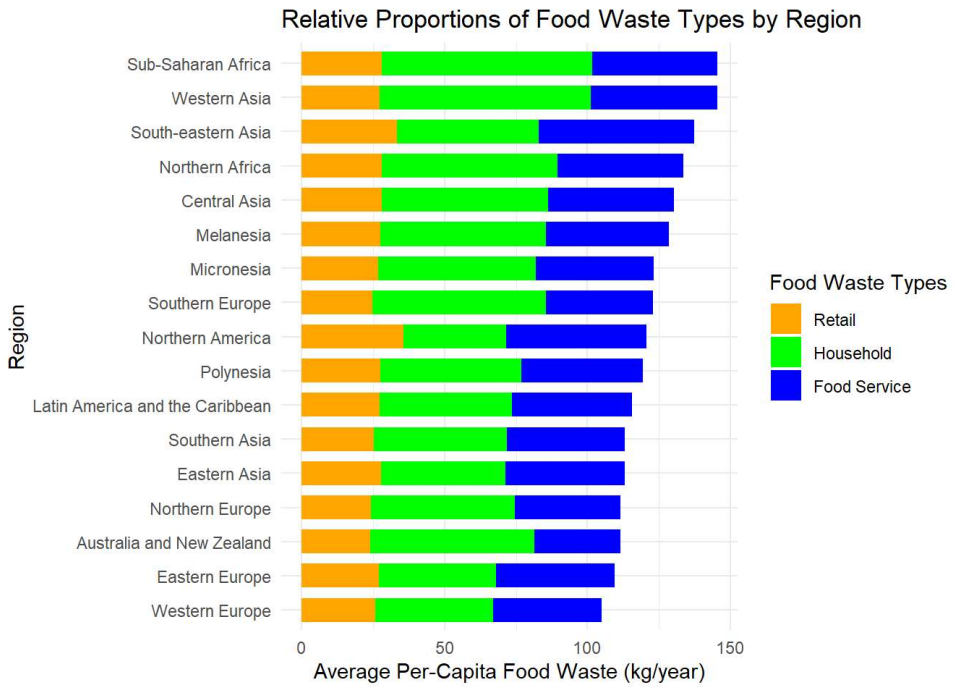
According to the World Resources Institute (https://www.wri.org/insights/reducing-food-loss-and-food-waste#:~:text=There%20used%20to%20be%20a,shown%20this%20isn%27t%20true.), poor infrastructure (unreliable roads, lack of cold storage, old equipment), suboptimal packaging, poor food management, and consumer behaviors are the most impactful factors on food waste. Less-developed regions like Sub-Saharan Africa may lack dependable roads and high-tech farming equipment that enable food to get from farm to table before going bad, as well as proper cold storage and packaging to keep food edible. Consumers in these regions may also be unaware of efficient food preparation and storage techniques, leading to tonnes of unnecessary waste.

Additional fun fact: “The environmental impact of wasted food is greater than that of packaging waste. So, while it’s important to limit this waste, it’s also important to use correct packaging to reduce food spoilage” (Goodwin, 2023). In other words, using more packaging to make sure food doesn’t go off is a worthwhile investment from an environmental perspective.

How does household per-capita food waste compare to retail and food service per-capita food waste?

Let’s take a look at the data from above in a graph.

```
# Show how different categories of per-capita food waste compare in each region
ggplot(averageFoodWasteByRegion, aes(x = reorder(region, avg_combined_kpcpy), y = avg_combined_kpcpy, fill = "Combined")) +
  geom_bar(stat = "identity", fill = "blue", width = 0.7) +
  geom_col(aes(y = avg_household_kpcpy, fill = "Household"), width = 0.7) +
  geom_col(aes(y = avg_retail_kpcpy, fill = "Retail"), width = 0.7) +
  geom_col(aes(y = avg_food_service_kpcpy, fill = "Food Service"), width = 0.7) +
  labs(
    title = "Relative Proportions of Food Waste Types by Region",
    x = "Region",
    y = "Average Per-Capita Food Waste (kg/year)"
  ) +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 0, hjust = 0.5)) +
  scale_fill_manual(
    values = c("Household" = "green", "Retail" = "blue", "Food Service" = "orange"),
    name = "Food Waste Types",
    labels = c("Retail", "Household", "Food Service")
  ) +
  guides(fill = guide_legend(title = "Food Waste Types")) +
  coord_flip()  # Flip the x and y axes for horizontal bars
```



Clearly, the biggest contributor to food waste is households for every region, and the disparity between household food waste and the other two sources is especially striking in less-developed regions like Sub-Saharan Africa and Western Asia. More developed regions like Eastern and Western Europe have a more uniform distribution of food waste contributions from all 3 sources.

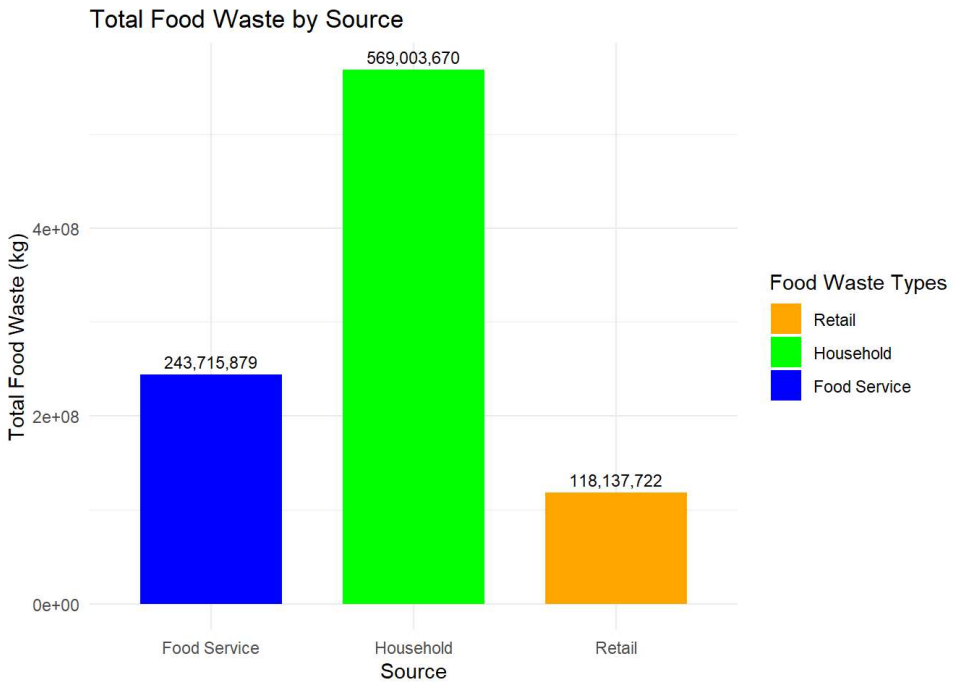
Which source contributes the most to total food waste?

Knowing the proportions of food waste from household, retail and food service in every region, I wanted to see how the totals compared.

```
# Compare food waste output of households, retail and food service
totalFoodWaste <- sqldf("SELECT SUM(household_estimate_tpy) AS total_household, SUM(retail_estimate_tpy) AS total_retail, SU
M(food_service_estimate_tpy) AS total_food_service FROM foodWasteData")
kable(head(totalFoodWaste), align="l")
```

total_household	total_retail	total_food_service
569003670	118137722	243715879

```r
# Visualize total food waste broken down by source
ggplot() +
  geom_bar(
    aes(x = c("Household", "Retail", "Food Service"),
        y = c(as.numeric(totalFoodWaste[1,1]),
              as.numeric(totalFoodWaste[1,2]),
              as.numeric(totalFoodWaste[1,3])),
        fill = c("green", "orange", "blue")),
    stat = "identity",
    position = "dodge",
    width = 0.7
  ) +
  geom_text(
    aes(x = c("Household", "Retail", "Food Service"),
        y = c(as.numeric(totalFoodWaste[1,1]),
              as.numeric(totalFoodWaste[1,2]),
              as.numeric(totalFoodWaste[1,3])),
        label = scales::comma(c(as.numeric(totalFoodWaste[1,1]),
                                as.numeric(totalFoodWaste[1,2]),
                                as.numeric(totalFoodWaste[1,3])))),
    position = position_dodge(width = 0.7),
    vjust = -0.5, size = 3
  ) +
  labs(
    title = "Total Food Waste by Source",
    x = "Source",
    y = "Total Food Waste (kg)"
  ) +
  scale_fill_manual(
    values = c("orange" = "orange", "green" = "green", "blue" = "blue"),
    name = "Food Waste Types",
    breaks = c("orange", "green", "blue"),
    labels = c("Retail", "Household", "Food Service")
  ) +
  theme_minimal()
```



Household food waste far surpasses that of food service and retail put together. This could be happening for a variety of reasons. Consumers could be accidentally purchasing too much food because of bulk buying, lack of meal planning or impulse buying, and throwing a lot of it out. They could also be unaware of the severe environmental and economic consequences of food waste, leading to wasteful behavior. As previously mentioned, many people in developing regions may not have access to refrigerators or other effective food storage units, leading to food going bad before it can be used. In some cultures, having an abundance of food is seen as a sign of hospitality or generosity, which can lead to more waste.

On the retail and food service side, many retailers around the world may reject imperfect produce or packaged food and throw it out instead of donating it. Food service establishments may have similar practices, such as pizza shops throwing out display slices at the end of the day instead of keeping or donating them.

Is there a relationship between a country's population size and its total food waste?

Answering this question required finding a dataset with population numbers for each country. However, joining it directly to the food waste dataset would be difficult because every dataset I found only had ISO-alpha3 3-letter country codes, and `foodWasteData` only had M49 country codes. I then had the idea of joining `foodWasteData` with a country code dataset containing both M49 codes *and* ISO-alpha3 codes by M49 codes, then joining the updated `foodWasteData` with a new population dataset by ISO-alpha3 codes. I found the following two datasets on Kaggle to accomplish this task:

- Standard country or area codes (M49) | Kaggle (https://www.kaggle.com/datasets/najielkotob/standard-country-or-area-codes-m49)
- 2021 World Population (updated daily) | Kaggle (https://www.kaggle.com/datasets/rsrishav/world-population)

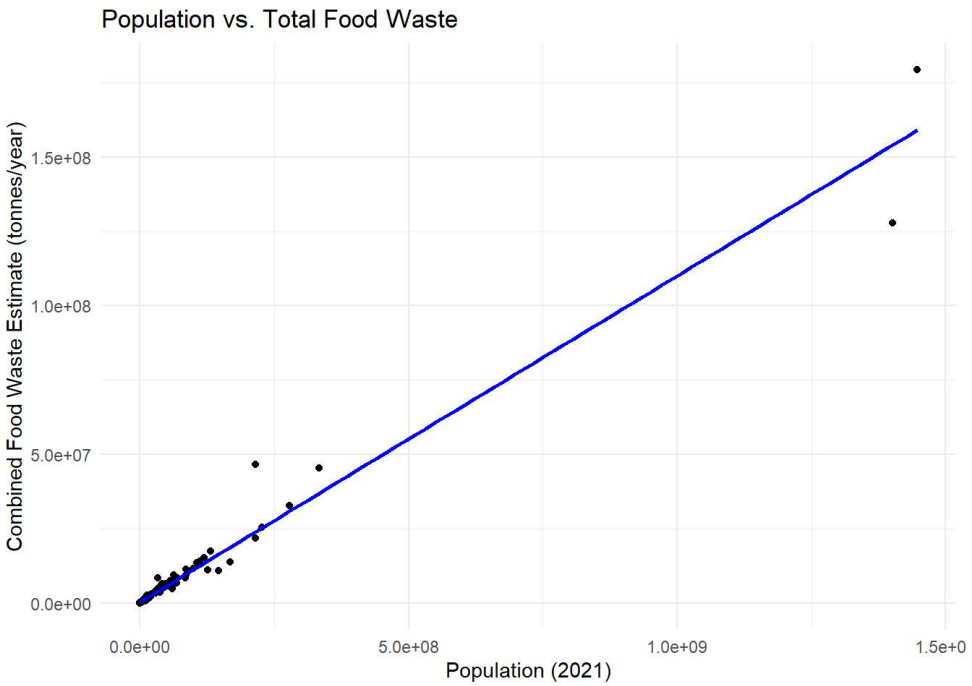
First, we join the datasets:

```
# Load dataset linking M49 codes to ISO-alpha3 codes and join with foodWasteData
codesData <- read_csv("M49 and ISO-alpha3 codes.csv")
names(codesData)[names(codesData) == 'M49 code'] <- 'm49_code'
names(codesData)[names(codesData) == 'ISO-alpha3 code'] <- 'iso_alpha3_code'
codesData <- sqldf("SELECT m49_code, iso_alpha3_code FROM codesData")
foodWasteData <- sqldf("SELECT * FROM foodWasteData INNER JOIN codesData
                ON foodWasteData.m49_code = codesData.m49_code")

# Join population dataset with foodWasteData using ISO-alpha3 codes
populationData <- read_csv("2021_population.csv")
names(populationData)[names(populationData) == 'iso_code'] <- 'iso_alpha3_code'
names(populationData)[names(populationData) == '2021_last_updated'] <- 'population_2021'
populationData <- sqldf("SELECT iso_alpha3_code, population_2021 FROM populationData")
foodWasteData <- merge(foodWasteData, populationData, by = "iso_alpha3_code", all.x = TRUE)
# Reorder foodWasteData and remove unnecessary duplicate m49 column
foodWasteData <- foodWasteData[, -14]
foodWasteData <- foodWasteData[, c(2,3,4,5,6,7,8,9,10,11,12,1,13,14)]
```

Then, we look for a correlation between population and total food waste. My prediction was that population and total food waste would be strongly correlated.

```
# Look for correlation between population size and total food waste
# Create the scatterplot with a regression line
ggplot(foodWasteData, aes(x = population_2021, y = combined_estimate_tpy)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE, color = "blue") +  # Add regression line
  labs(
    title = "Population vs. Total Food Waste",
    x = "Population (2021)",
    y = "Combined Food Waste Estimate (tonnes/year)"
  ) +
  theme_minimal()
```




```
# Calculate the regression coefficient r^2
foodwaste_regression <- lm(population_2021 ~ combined_estimate_tpy, data = foodWasteData)
r_squared <- summary(foodwaste_regression)$r.squared
print(r_squared)
```

```
## [1] 0.9660028
```

The regression coefficient r^2 is ~0.966, meaning that 96.6% of variation in a country's total combined food waste can be explained by its population. This is exactly what I expected to find.

Conclusion 👏

Recommendations

- Countries should improve education efforts around meal planning, the consequences of food waste, and proper storage of food.
- Incentives should be given to encourage people to only buy what they need or donate what they don't.
- Less-developed countries should implement policies to improve access to refrigerators and other storage units that can keep food from going bad too quickly.
- Retail and food service businesses should set up procedures to donate imperfect or extra food items to those in need instead of condemning it to the trash pile immediately. To incentivize this, governments could provide tax breaks for businesses that donate significant amounts of extra food that would have otherwise gone to waste.

To check out next

- The original dataset used in this project (https://www.kaggle.com/datasets/joebeachcapital/food-waste)

- How Wasted Food Turns into Huge Amounts of Greenhouse Gas - Scientific American (https://www.scientificamerican.com/article/how-wasted-food-turns-into-huge-amounts-of-greenhouse-gas/)
- Fight climate change by preventing food waste | Stories | WWF (https://www.worldwildlife.org/stories/fight-climate-change-by-preventing-food-waste#:~:text=But%20wasted%20food%20isn%27t,more%20potent%20than%20carbon%20dioxide.)
- My other data analysis projects:
    - Josh-Houlding-Bellabeat-Case-Study (GitHub) (https://github.com/jhould007/Josh-Houlding-Bellabeat-Case-Study)
    - Seattle-Rain-Analysis (GitHub) (https://github.com/jhould007/Seattle-Rain-Analysis)
- My entire GitHub profile (https://github.com/jhould007)
- My Kaggle profile (https://www.kaggle.com/joshhoulding)
- My LinkedIn profile (https://www.linkedin.com/in/joshuahoulding/)

Thank you for reading, and have a good one! 😀