Project 5

Read in the dataset you will be working with:

```
# Food Consumption Dataset
food <- readr::read_csv('https://raw.githubusercontent.com/rfordatascience/tidytuesday/master/data/2020/2020-02-18/food_cons
umption.csv')

# Cleaning Dataset
food_grouped <- food %>%
  group_by (country) %>%
  summarize(
    total_consumption = sum(consumption), # Country Total
    total_emmission = sum(co2_emmission)) # Country Total

# Filtering Out Vegan Options
meat <- food[!(
  food$food_category=="Wheat and Wheat Products" |
    food$food_category=="Rice" |
    food$food_category=="Soybeans" |
    food$food_category=="Nuts inc. Peanut Butter"),] %>%
  group_by (country) %>%
  summarize(
    meat_consumption = sum(consumption)) # Country Total

# Combining the Datasets and Creating a % Meat Column
food_grouped_final <- left_join(food_grouped, meat) %>%
  group_by (country) %>%
  mutate(
    percentage_meat = meat_consumption/total_consumption)

# Filtering Top 15 GDP Countries
top_countries <- food_grouped_final[(
  food_grouped_final$country=="USA" |
    food_grouped_final$country=="China" |
    food_grouped_final$country=="Japan" |
    food_grouped_final$country=="Germany" |
    food_grouped_final$country=="United Kingdom" |
    food_grouped_final$country=="India" |
    food_grouped_final$country=="France" |
    food_grouped_final$country=="Italy" |
    food_grouped_final$country=="Canada" |
    food_grouped_final$country=="Sourth Korea" |
    food_grouped_final$country=="Russia" |
    food_grouped_final$country=="Brazil" |
    food_grouped_final$country=="Australia" |
    food_grouped_final$country=="Spain" |
    food_grouped_final$country=="Indonesia"),]
```

Information about the dataset

- Name: *Food Consumption and CO2 Emissions*
- Author: Kasia Kulma
- Source: https://github.com/rfordatascience/tidytuesday/blob/master/data/2020/2020-02-18/readme.md#food_consumptioncsv (https://github.com/rfordatascience/tidytuesday/blob/master/data/2020/2020-02-18/readme.md#food_consumptioncsv).

Question

Is there a correlation between the percentage of meat products a country consumes and it’s total CO2 emissions?

Introduction

The dataset used is the *Food Consumption and CO2 Emissions* Dataset, originally from *nu3* and cleaned by Kasia Kulma. It was gathered from *nu3* via webscrapping and contains 1430 observations for different countries of 4 different variables. For more information, visit the https://www.nu3.de/blogs/nutrition/food-carbon-footprint-index-2018 (https://www.nu3.de/blogs/nutrition/food-carbon-footprint-index-2018) or https://github.com/rfordatascience/tidytuesday/blob/master/data/2020/2020-02-18/readme.md#food_consumptioncsv (https://github.com/rfordatascience/tidytuesday/blob/master/data/2020/2020-02-18/readme.md#food_consumptioncsv).

In order to answer my question, I focused on the all 4 variables as follows:

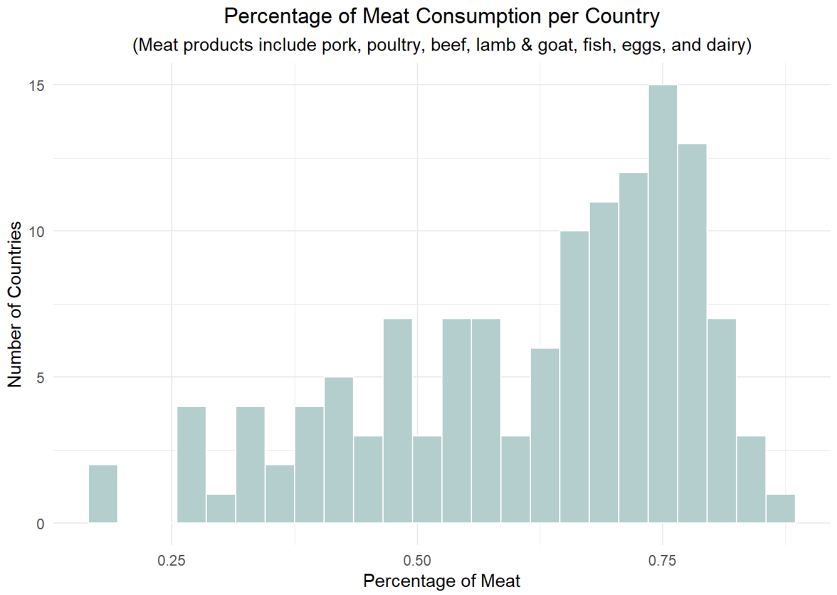
- `country` : character variable, country name
- `food_category` : character variable, filtered for only meat products (pork, poultry, beef, lamb & goat, fish, eggs, and dairy) to calculate a percentage of meat in overall diet
- `consumption` : double variable, consumption measured in kg/person/year for each type of `food_category`
- `co2_emmission` : double variable, CO2 emissions measured in kg/person/year for each type of `food_category`

Approach

My approach was broken down into the following three steps:

1. Broadly assess the variables in question: After grouping the data by country, I created two separate histograms and general summaries for both the `consumption` and `co2_emmission` variables to assess any outliers or areas of concern. Both variables appeared to be skewed but showed no areas of concern.

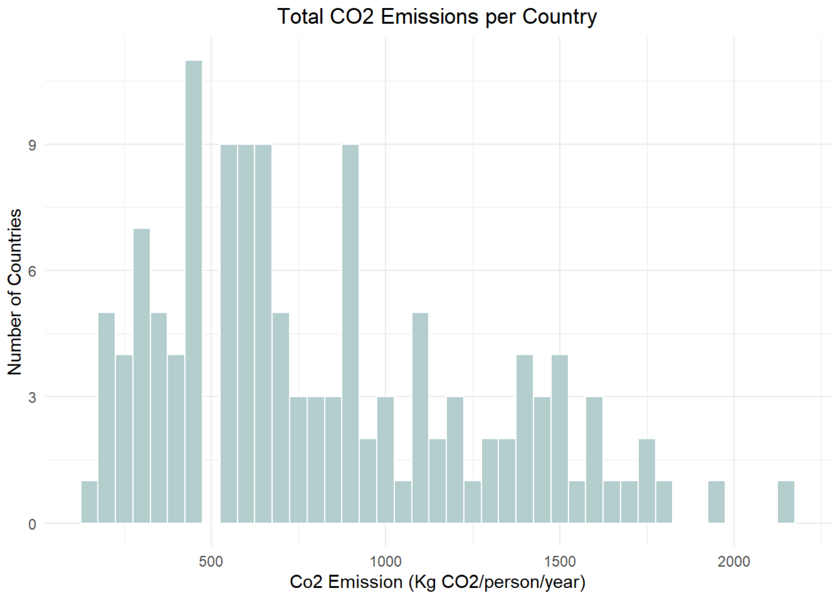
```r
# Consumption Variable
ggplot(food_grouped_final, aes(percentage_meat)) +
  geom_histogram(binwidth = .03, colour = "white", fill = "lightcyan3") +
  theme_minimal() +
  xlab("Percentage of Meat") +
  ylab("Number of Countries") +
  ggtitle("Percentage of Meat Consumption per Country",
  subtitle = "(Meat products include pork, poultry, beef, lamb & goat, fish, eggs, and dairy)") +
  theme(plot.title = element_text(hjust = 0.5), plot.subtitle = element_text(hjust = 0.5))
```



```r
summary(food_grouped_final$percentage_meat)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##  0.1786  0.5058  0.6645  0.6164  0.7413  0.8595
```

```r
# CO2 Variable
ggplot(food_grouped_final, aes(total_emmission)) +
  geom_histogram(binwidth = 50, colour = "white", fill = "lightcyan3") +
  theme_minimal() +
  xlab("Co2 Emission (Kg CO2/person/year)") +
  ylab("Number of Countries") +
  ggtitle("Total CO2 Emissions per Country") +
  theme(plot.title = element_text(hjust = 0.5))
```



```r
summary(food_grouped_final$total_emmission)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   141.4   446.7   685.9   818.2  1111.2  2172.4
```

2. Create hierarchical clusters: I ultimately settled on the maximum distance method using UPGMA, given that it produced the most reasonable and consistent groups for 3 clusters. The dendrogram is not provided given the extensive amount of observations.

3. Create a linear model: Finally, I combined the hierarchical cluster visualization with a traditional linear model. I also higlighted 15 major countries ([Top 15 Countries by GDP in 2022][https://globalpeoservices.com/top-15-countries-by-gdp-in-2022/ (https://globalpeoservices.com/top-15-countries-by-gdp-in-2022/)]) that might provide good reference points.

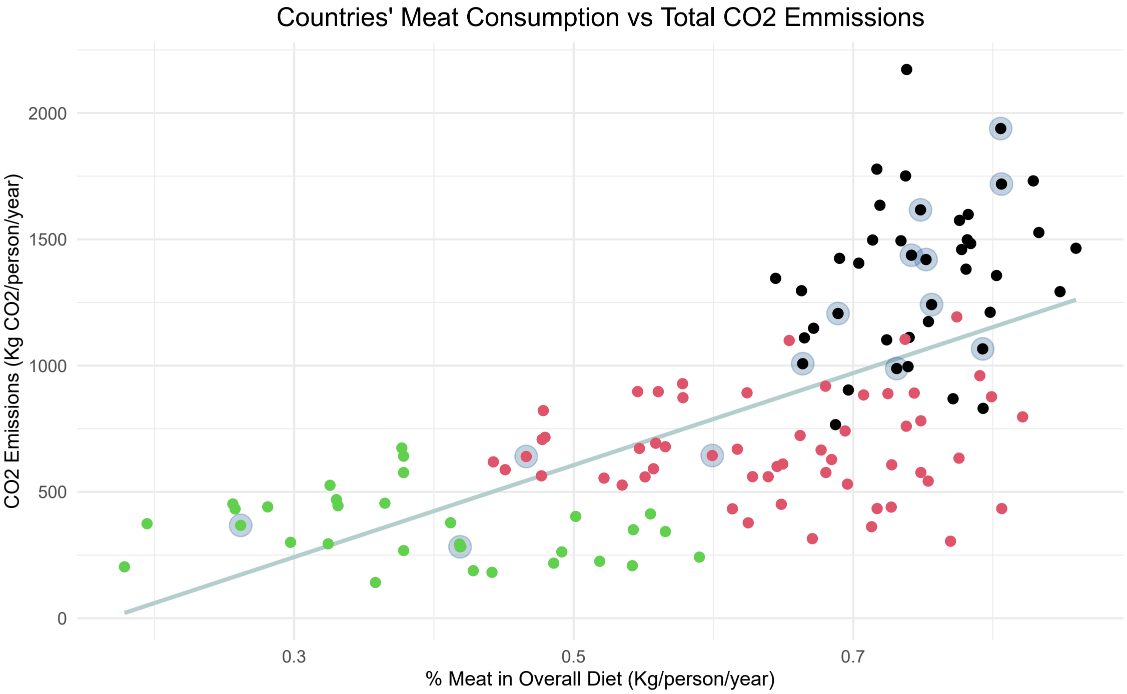
Analysis

```
# Hierarchical Clustering w/ UPGMA
dist_out <- food_grouped_final %>%
  column_to_rownames(var = "country") %>%
  scale() %>%
  dist(method = "maximum") # Only method that didn't have outliers

hc_out <- hclust(
  dist_out, method = "average") #average = UPGMA

cluster <- cutree(hc_out, k = 3) #3 clusters


# Create Linear Model
lm_out <- lm(total_emmission ~ percentage_meat, data = food_grouped_final)

summary(lm_out)
```

```
##
## Call:
## lm(formula = total_emmission ~ percentage_meat, data = food_grouped_final)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -792.6 -273.2   -8.1  224.8 1132.1
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)       -304.4      119.6  -2.545   0.0121 *
## percentage_meat   1821.3      187.6   9.706   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 347.4 on 128 degrees of freedom
## Multiple R-squared:  0.424,  Adjusted R-squared:  0.4195
## F-statistic: 94.21 on 1 and 128 DF,  p-value: < 2.2e-16
```

```
# Plot Clustering + LM + Top 15 Countries
co2_scatter <- ggplot(food_grouped_final) +
  aes(percentage_meat, total_emmission) +
  theme_minimal() +
  geom_smooth(
    method = "lm", color = "lightcyan3",
    se = FALSE) +  #suppress confidence band
  xlab("% Meat in Overall Diet (Kg/person/year)") +
  ylab("CO2 Emissions (Kg CO2/person/year)") +
  ggtitle("Countries' Meat Consumption vs Total CO2 Emmissions") +
  theme(plot.title = element_text(hjust = 0.5),
        axis.title.y = element_text(size = 10),
        axis.title.x = element_text(size = 10)) +
  geom_point( # Highlighting top 15 GDP countries for reference
    data = top_countries,
    aes(percentage_meat, total_emmission),
    color = "dodgerblue4",
    fill = "dodgerblue4",
    size = 5,
    alpha = 0.25,
    shape = 21) +
  geom_point_interactive(
    aes(data_id = country, tooltip = country),
    color=cluster,
    size = 2)

girafe(
  ggobj = co2_scatter,
  width_svg = 8,
  height_svg = 8*0.618,
  options = list(
    opts_hover(css = "fill: #000000;"),
    opts_hover_inv(css = "opacity: 0.2;")))
```

```
## `geom_smooth()` using formula 'y ~ x'
```



Discussion

The overall trend shows that with increases in meat consumption, there are correlated increased in CO2 emissions. There appears to be more variation in the data as the percentage of meat in overall diet increases However, there are also more data points as the percentage of meat in overall diet increases, inherently giving way to more variability.

The LM regression results prove to be significant with a p-value of less than 2e-16 and an R^2 value of 0.4195. This means that the percent of meat in the overall diet of a country most likely is an indicator of the country’s total CO2 emissions, but this only explains 41.95% of the variation in the data. Some limiting factors in our study that may have affected this low R^2 could be the small number of variables to look at and the unclear method of measuring the variables.

Highlighting the top 15 GDP countries provides some helpful reference points for the data, and a majority of these countries appear to fall in the high CO2 emission category (*shown in black*). It is important to note that there were no statistical tests performed for GDP and CO2 correlations. Moreover, there are likely strong compounding factors from countries’ production of goods, population, and overall development.