DSIFinal Project

Bryce Leary

1/11/2020

Introduction

821 million people across the world suffered from hunger in 2018 according to the United Nations. Policymakers and global leaders are committed to ending hunger, and have codified this effort through the Millenium Challenge Goals, the Sustainable Development Goals, and the UN's Zero Hunger Challenge. These policymakers and global leaders face many challenges however, and this analysis provides an analysis of food availability in West Africa. We seek to understand how capital flows in the agricultural sector influence the availability of food in West African countries.

#Collection Method We are using six sets of data which originate from the Food and Agriculture Organization of the United States (FAO). All five datasets are time series and range from 2000 to 2017. The first dataset is the aggregated data of Development Aid Disbursement (DevAid), which is arranged by bilateral, multilateral and private donors across all West African countries. The DevAid data was last updated on January 26, 2019 and was obtained by FAO through the Credit Reporting System (CRS). The DevAid data is composed of data on the amount of aid disbursed for basic nutrition and food aid and food security programs, measured in 2016 USD millions.

The second dataset is the Foreign Direct Investment (FDI). which is measured in terms of 2016 US Millions of dollars. This dataset was last updated on November 11, 2019 and covers the total FDI inflow and outflows to and from developing countries in the West Africa region. This dataset was collected by the United Nations Conference on Trade and Development (UNCTAD), The International Trade Centre (INTRACEN), the Organization for Economic Co-operation and Development (OECD) and the International Monetary Fund (IMF) Balance of Payments Manual. Both the DevAid and FDI datasets are from the FAO data group of Development Flows to Agriculture.

The third dataset is the Average Dietary Energy Supply Adequacy (ADESA) from the FAO's Suite of Food Security Indicators which was last updated on October 11, 2019. It is represented in a three-year average format and is indicated as a percentage (FAO 2019). The dietary energy supply is determined by each country's average supply of calories for food consumption of the population.

The fourth and fifth datasets are presented as imports and exports of crops and livestock products. Both exports and imports include the total of aggregated agricultural products by country on an annual basis. These datasets were last updated on October 9, 2019. This agricultural trade data was collected corresponding to the Standard International Merchandise Trade Statistics Methodology; the main providers of this data are UNSD and Eurostat, but other providers are involved if needed for non-reporting countries or missing cells (FAO 2019).

The sixth and last data set is the Depth of Food Deficit (Depth) and originates from the World Bank Data Development Indicators, which were sourced from the FAO's Food Security Statistics and last updated on December 4, 2019. The key indicator in this dataset is represented in kilocalories per person per day, based on the number of calories needed per day to lift the undernourished population from this category when everything else remains the same.

#Our analysis finds that PUTFINDINGSHERE.

Prior to formally testing our hypothesis ,we executed exploratory `ggplots` to view the shape of our explanatory and response variables (figure placeholder). This initial view led us to consider a basic linear regression to fit our analysis and knowledge of statistical methods.

This regression indicated that multilateral donor flow and export flows are the two main influences on our response variables, `adesa` and `depth`.

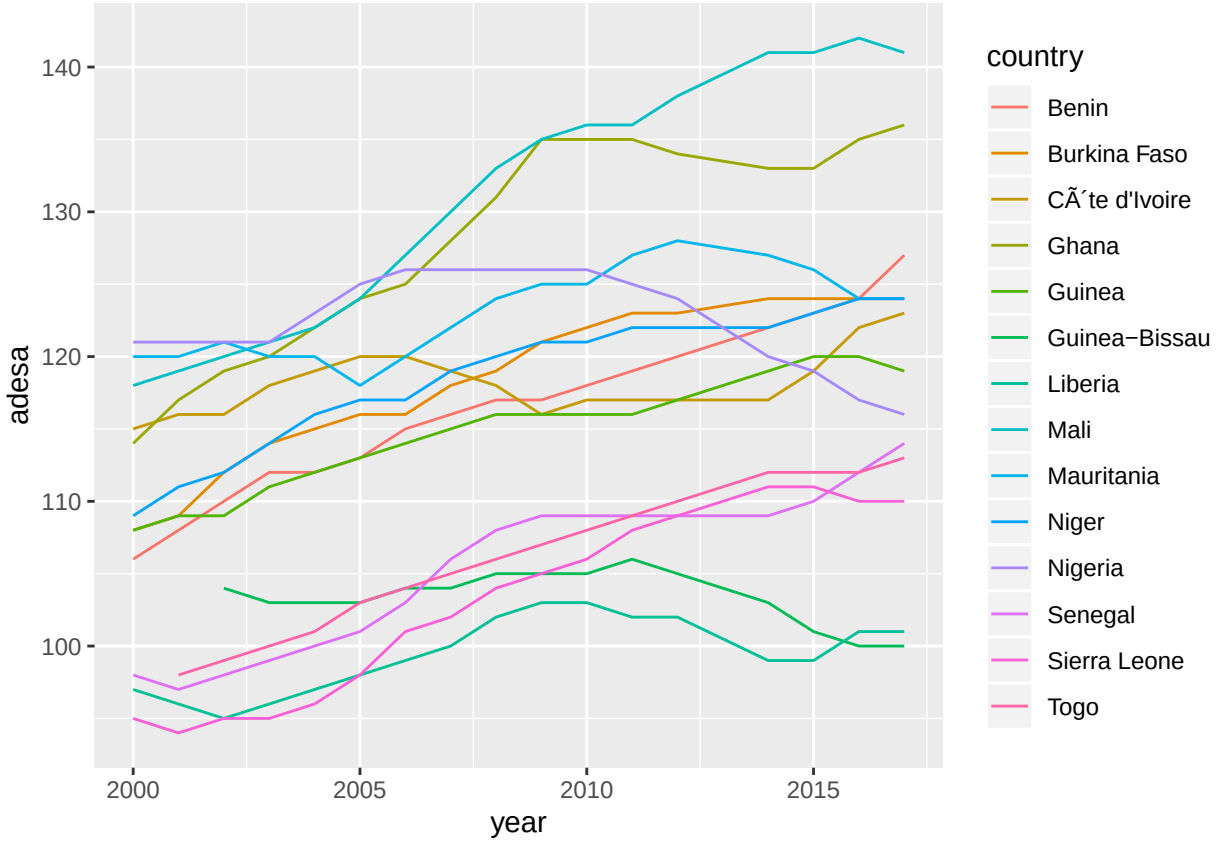
To validate this finding, we also fit the regression on a line graph with smoothing features (figure placeholder).

Together, the regression output and graph show that as the multilateral aid flow increases by approximately 79 million USD, the depth of the food deficit decreases by 1,000 calories.

The same method was executed for export values and average dietary energy supply (ADESA).

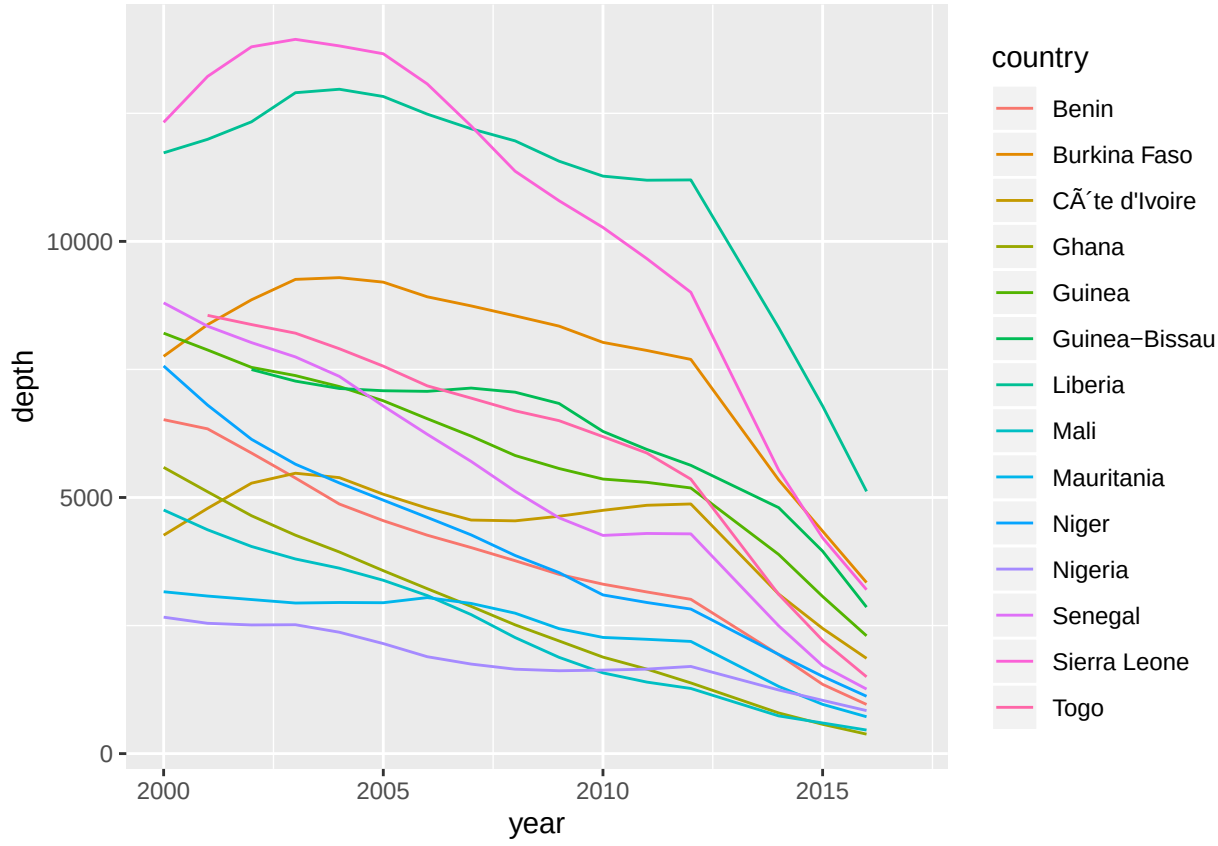
This output and graph depicts how an increase in exports reduces available food for populations experiencing hunger across the West African countries.

```
#Plot of ADESA by country over year
ggplot(agriculture, aes(x = year, y = adesa, color = country)) + geom_line()
```




```
#Plot of depth by country over year
ggplot(agriculture, aes(x = year, y = depth, color = country)) + geom_line()
```

```
## Warning: Removed 14 rows containing missing values (geom_path).
```



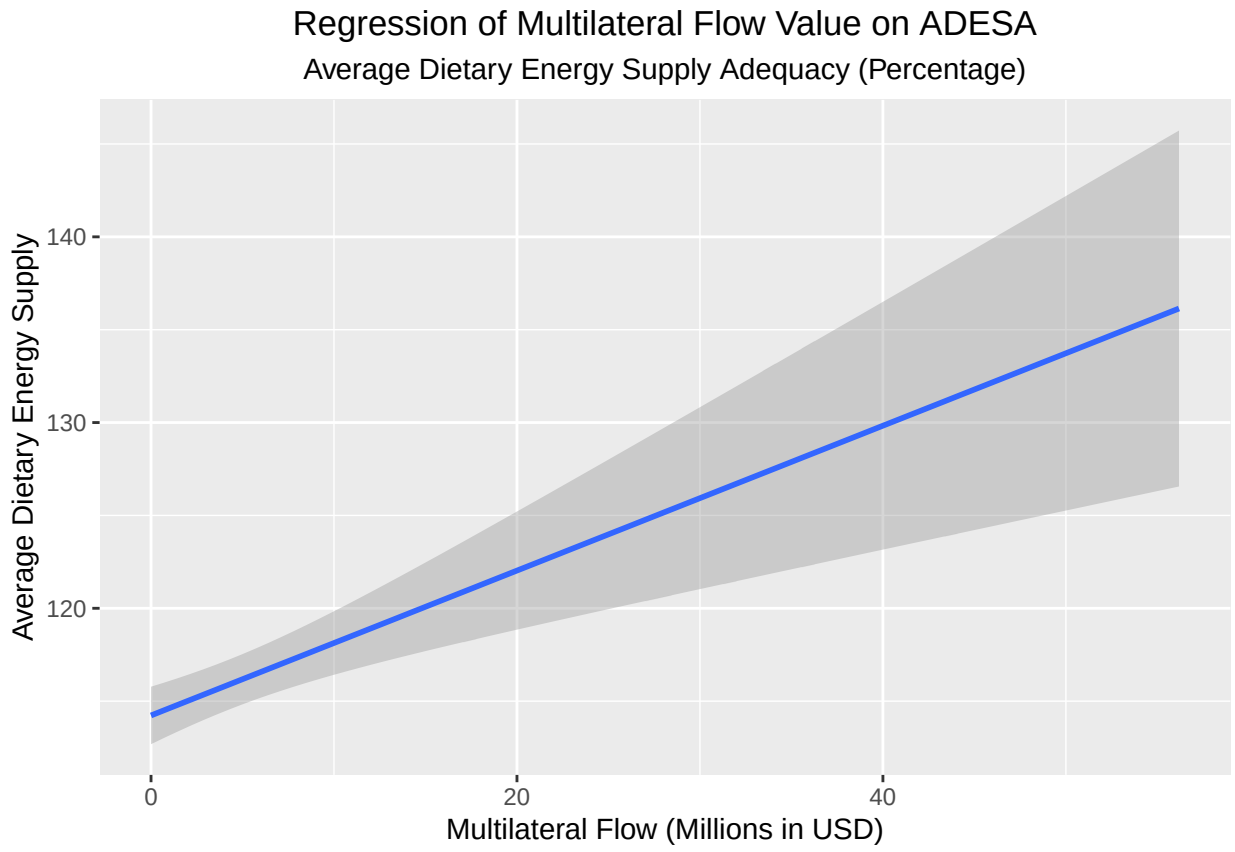

```
###Basic linear regression###
summaidfdiie <- lm(cbind(depth,adesa) ~ bilateral + multilateral + fdi_net + exp_value + imp_value, data
summary(summaidfdiie)
```

```
## Response depth :
##
## Call:
## lm(formula = depth ~ bilateral + multilateral + fdi_net + exp_value +
##     imp_value, data = agriculture)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -4949  -2053   -180   1384   7662
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  6295.3741   351.0691  17.932  < 2e-16 ***
## bilateral     -14.9188    20.1343  -0.741  0.45980
## multilateral  -71.0741    31.0665  -2.288  0.02345 *
## fdi_net        -0.5423     0.2037  -2.662  0.00855 **
## exp_value      -0.3312     0.1592  -2.081  0.03905 *
## imp_value      -0.3640     0.2386  -1.526  0.12902
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2700 on 161 degrees of freedom
```

```
##   (68 observations deleted due to missingness)
## Multiple R-squared:  0.2564, Adjusted R-squared:  0.2333
## F-statistic:  11.1 on 5 and 161 DF,  p-value: 3.333e-09
##
##
## Response adesa :
##
## Call:
## lm(formula = adesa ~ bilateral + multilateral + fdi_net + exp_value +
##     imp_value, data = agriculture)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -19.6206  -6.6908   0.1056   6.2303  24.6378
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   1.094e+02  1.143e+00  95.716  < 2e-16 ***
## bilateral     2.374e-01  6.556e-02   3.621 0.000392 ***
## multilateral  1.929e-01  1.012e-01   1.907 0.058322 .
## fdi_net       3.311e-03  6.632e-04   4.992 1.54e-06 ***
## exp_value     1.660e-03  5.183e-04   3.203 0.001638 **
## imp_value    -1.253e-03  7.767e-04  -1.613 0.108781
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.79 on 161 degrees of freedom
##   (68 observations deleted due to missingness)
## Multiple R-squared:  0.3302, Adjusted R-squared:  0.3094
## F-statistic: 15.88 on 5 and 161 DF,  p-value: 1.061e-12
```

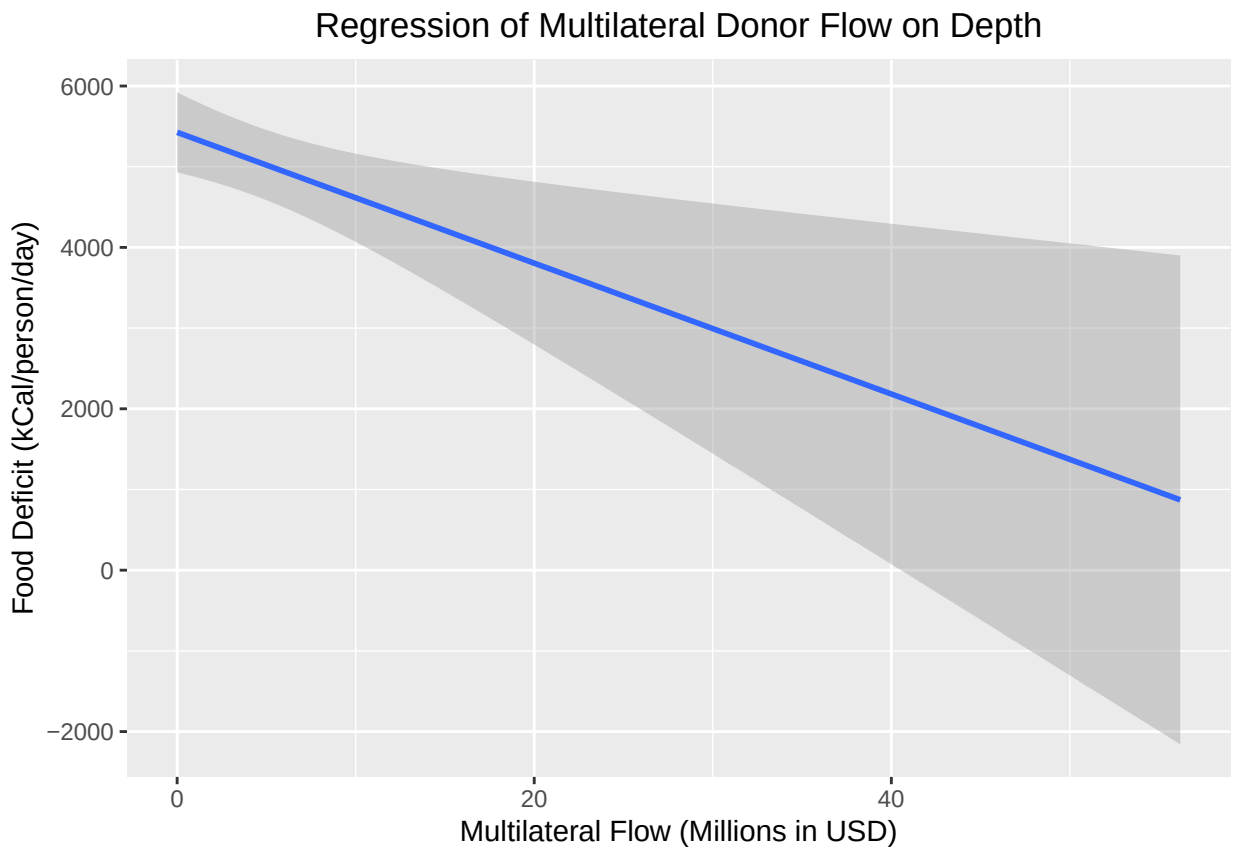
```
###Plots of Multivariate Regression###

#Multilateral on ADESA
ma <- ggplot(agriculture, aes(x = multilateral, y = adesa)) + geom_smooth(method = "lm")
ma + theme(plot.title = element_text(hjust = 0.5)) +
  theme(plot.subtitle = element_text(hjust = 0.5)) +
  labs(title = "Regression of Multilateral Flow Value on ADESA",
       subtitle = "Average Dietary Energy Supply Adequacy (Percentage)",
       x = "Multilateral Flow (Millions in USD)", y = "Average Dietary Energy Supply")
```

```
## Warning: Removed 17 rows containing non-finite values (stat_smooth).
```




```
#Multilateral on Depth
md <- ggplot(agriculture, aes(x = multilateral, y = depth)) + geom_smooth(method = "lm")
md + theme(plot.title = element_text(hjust = 0.5)) +
  labs(title = "Regression of Multilateral Donor Flow on Depth",
       x = "Multilateral Flow (Millions in USD)", y = "Food Deficit (kCal/person/day)")
```

```
## Warning: Removed 31 rows containing non-finite values (stat_smooth).
```

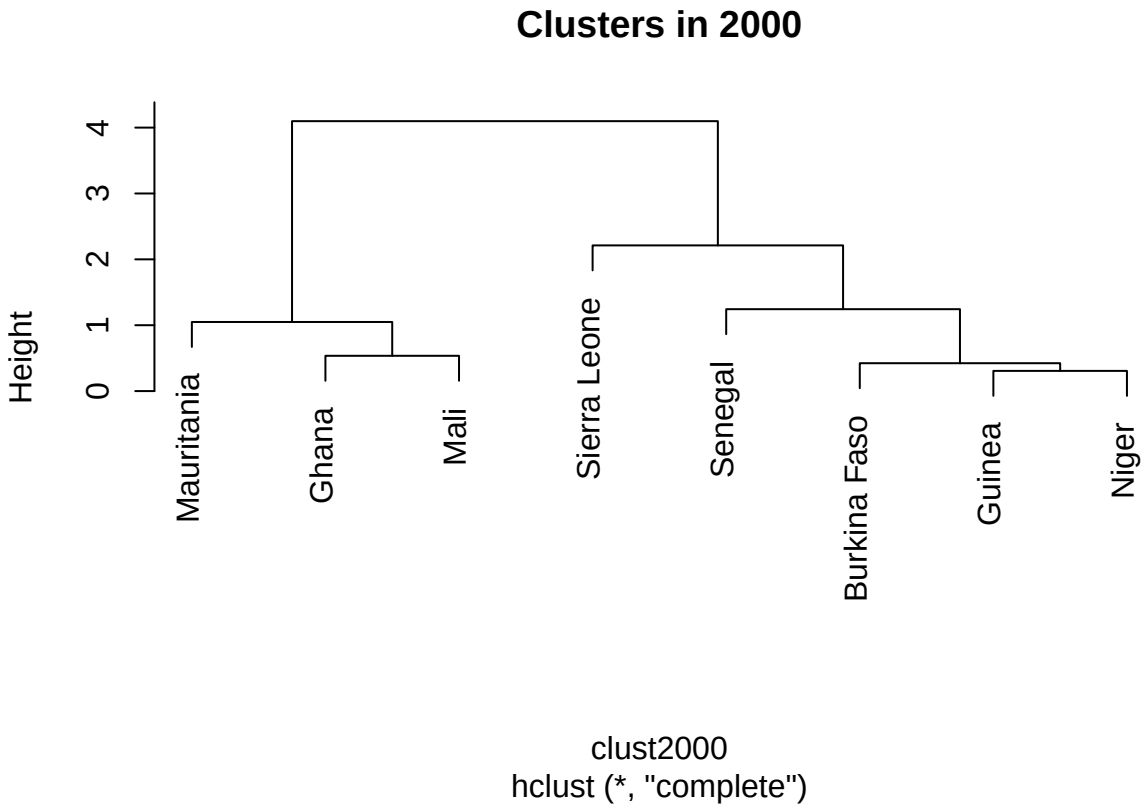


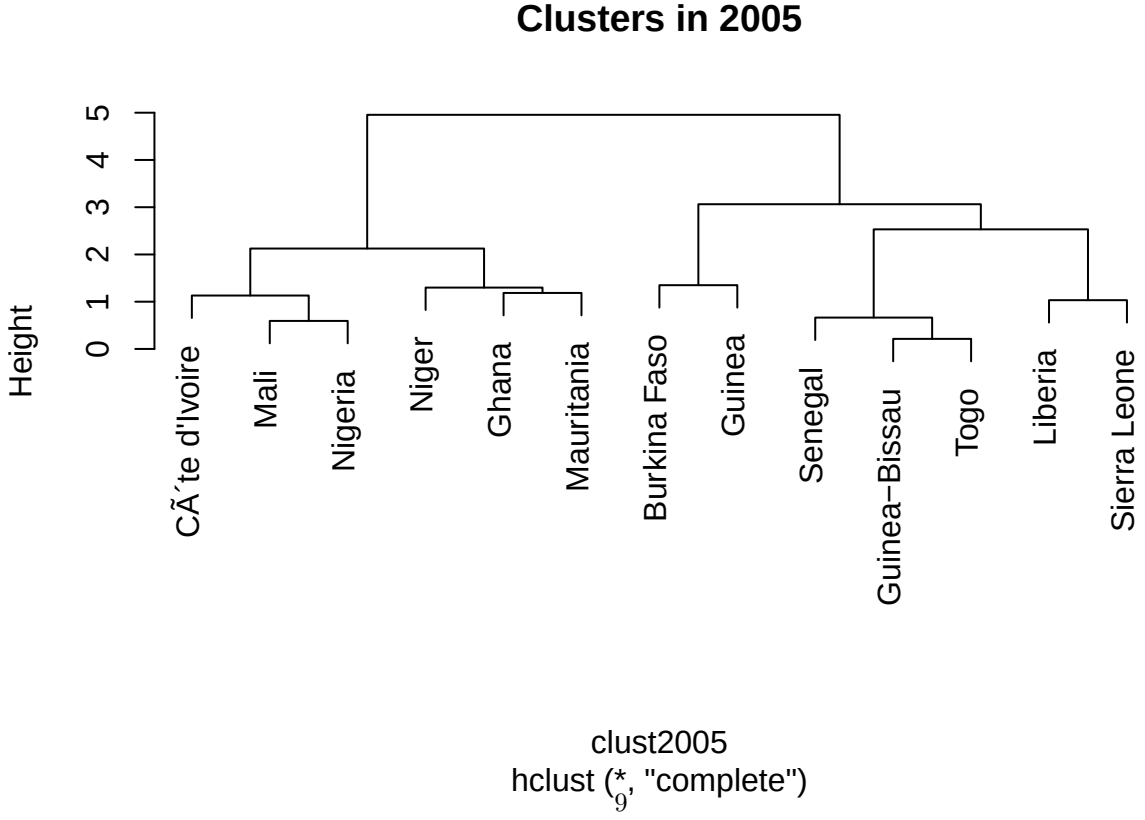

```
#Export Value on ADESA
ea <- ggplot(agriculture, aes(x = exp_value, y = adesa)) + geom_smooth(method = "lm")
ea + theme(plot.title = element_text(hjust = 0.5)) +
  theme(plot.subtitle = element_text(hjust = 0.5)) +
  labs(title = "Regression of Export Value on ADESA",
       subtitle = "Average Dietary Energy Supply Adequacy (Percentage)",
       x = "Export Value (Millions in USD)", y = "Average Dietary Energy Supply")
```

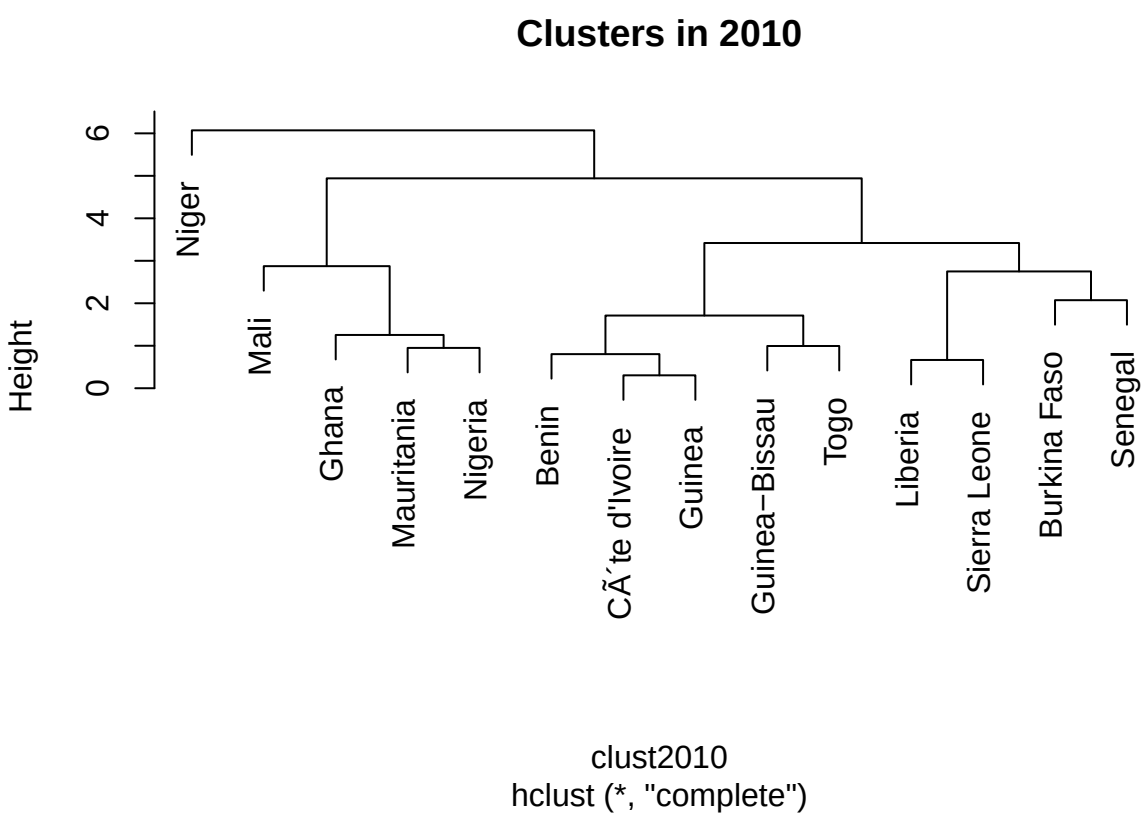

```
#Export Value on Depth
ed <- ggplot(agriculture, aes(x = exp_value, y = depth)) + geom_smooth(method = "lm")
ed + theme(plot.title = element_text(hjust = 0.5)) +
  labs(title = "Regression of Export Value on Depth",
       x = "Export Value (Millions in USD)", y = "Food Deficit (kCal/person/day)")
```

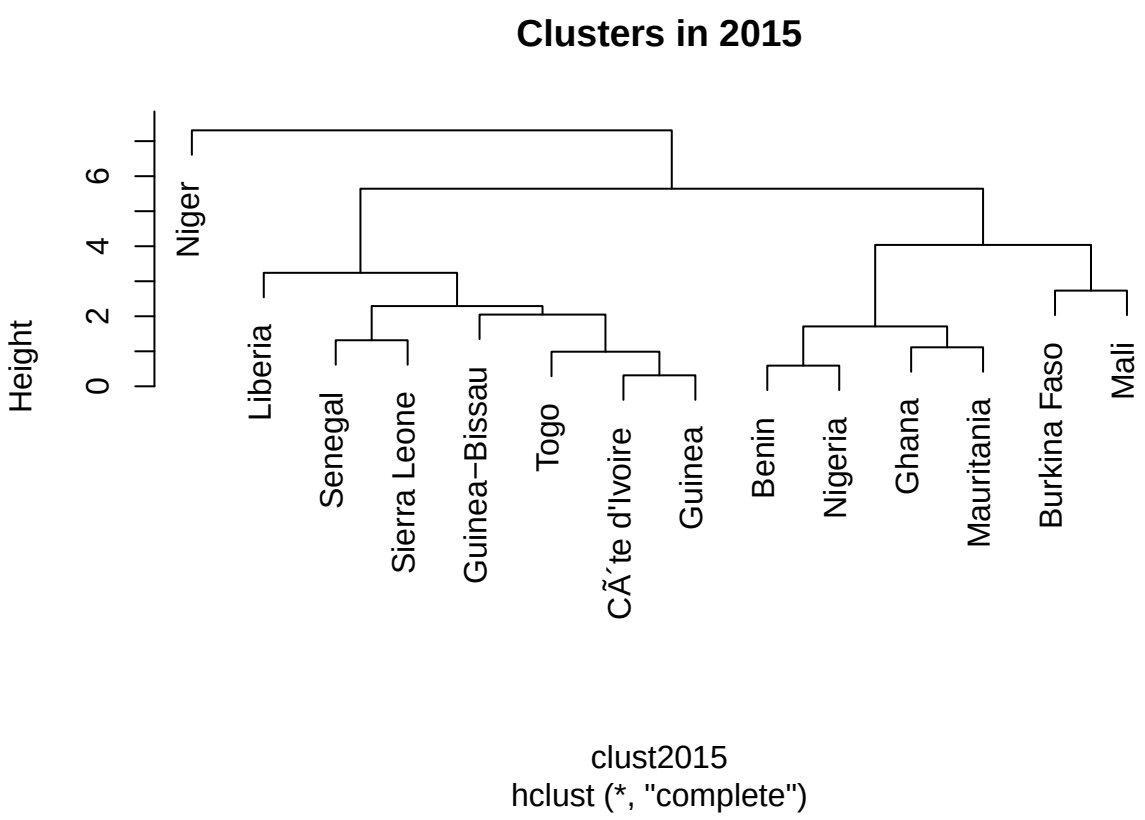
```
## Warning: Removed 14 rows containing non-finite values (stat_smooth).
```


Regression of Export Value on Depth
Food Deficit (kCal/person/day)
6000
4000
2000
0
0
2000
4000
6000
Export Value (Millions in USD)


Clusters in 2000
Height
0
1
2
3
4
Mauritania
Ghana
Mali
Sierra Leone
Senegal
Burkina Faso
Guinea
Niger
clust2000
hclust (*, "complete")


Clusters in 2005
Height
0
1
2
3
4
5
CÃ´te d'Ivoire
Mali
Nigeria
Niger
Ghana
Mauritania
Burkina Faso
Guinea
Senegal
Guinea−Bissau
Togo
Liberia
Sierra Leone
clust2005
hclust (*, "complete")


Clusters in 2010
Height
0
2
4
6
Niger
Mali
Ghana
Mauritania
Nigeria
Benin
CÃ´te d'Ivoire
Guinea
Guinea−Bissau
Togo
Liberia
Sierra Leone
Burkina Faso
Senegal
clust2010
hclust (*, "complete")



Cluster Analysis

To more closely examine the trends uncovered by the linear regression model, we decided to complete a cluster analysis to examine how the different states have shifted in relation to each other. This analysis could potentially reveal additional information on how capital flows have impacted similar states. We focused on bilateral and multilateral development flows, as they contained the most complete data and clustering is ineffective with significant amounts of missing data. Due to the relatively small sample size, we utilized a hierarchical agglomerative method clustering process as it can identify nuance in small-n datasets which are harder to trace through k-means clustering.

This process examined the shifts in clusters at four time markers: 2000, 2005, 2010, and 2015. These time periods were selected as they contain most of the data, avoid incomplete data, and prevent an oversaturation of the dendrogram. The results are summarized in dendrograms created through `hclust`, with the closeness between clusters summarized through the `height` variable on the Y-axis.

Overall, we see countries moving further apart from each other as the millennium progressed. This is represented by the maximum height value moving from 4.0979373 in 2000 to 7.3082381 in 2015, a change of -3.2103008. We can also examine Baker's Gamma correlation coefficient to determine the level of similarity between two dendrograms. Due to the varying size of the dendrograms, it is only possible to compare figure 3 and figure 4. Between these we find a Baker's Gamma correlation of 0.6811645. This correlation coefficient indicates a strong similarity between the two models, though they are not exactly the same. While this correlation coefficient doesn't account for height, we also see a change in similarity between the two years of -1.2373249 as Niger continues to pull away from additional countries. Thus, while we see most countries experiencing similar effects of multinational and bilateral development aid

#Caveats and additional notes regarding our study:

Our study includes several caveats, which we would like to address in this section. These caveats are evaluated in terms of data collection, data cleaning, and analysis.

Collection

This data was gathered by the Food and Agriculture Organization of the United Nations (FAO). While the FAO indicates that most of its data is from a third party which often has its own data integrity standards, our team does not know the exact data collection methods and decisions made in the moment. For this reason, we acknowledge that there may be unknown sources of collection bias. To help mitigate these effects, we reviewed the codebook for each data set.

Growth domestic product (GDP) is not included in this dataset. While GDP is a commonly used variable when examining country-level growth , we chose to focus on capital flows. Policymakers have several policy levers at their disposal which can impact capital flows, such as banking regulations or legislation. We included capital flows instead of GDP to focus on the relationship between policymakers and nutrition outcomes.

Cleaning

To clean this dataset, we kept `country` and `country_code` as the unique identification variables for each observation. This decision relies on the assumption that country would be the. main variable by which we would need to use numerical identification as fallback.

During the cleaning and merging process, (placeholder for Ivory coast) stood out as an observation that was inconsistently spelled across the distinct datasets. To avoid editing the excel sheet, we reconciled this inconsistent spelling by editing the `country` column form a factor to character. Then, we replaced the existing spelling with NA. After this, we replaced NA with the alternative spelling.

The USD values for the Foreign Direct Investment (FDI) dataset were adjusted from 2010 USD values to 2016 USD values. This adjustment was made in light of the Development Aid (DevAid) dataset, which included 2016 USD values. We could have adjusted the DevAid values to 2010, but decided to use the most recent value of 2016 instead. This logic also considers that our dataset includes years from 2000 to 2017.

The Crops and Livestock Import and Export (Trade) datasets included import and export values in units of thousands, while the other four datasets of this study were in units of millions. To reconcile this difference, we divided the trade values `exp_value` and `imp_export` by 1,000 so that these values would be represented in millions.

Analysis

Our analysis includes a linear regression and cluster analysis. While we did attempt to. execute a Bayesian regression at the beginning of the study, our final study does not include this analysis. We recognize that a Bayesian regression method may have been a more robust analysis method. However, due to our levels of statistical experience and background, we executed a multivariate linear regression. To continue to increase our knowledge in statistical methods, we focused on pushing our newly gained knowledge from the course via the cluster analysis.

The linear regression in our study excludes a flow type, `private`.The `private` observations were omitted from the regression due to their low n-size and minimal contribution to the exploratory Bayesian regression method mentioned above.