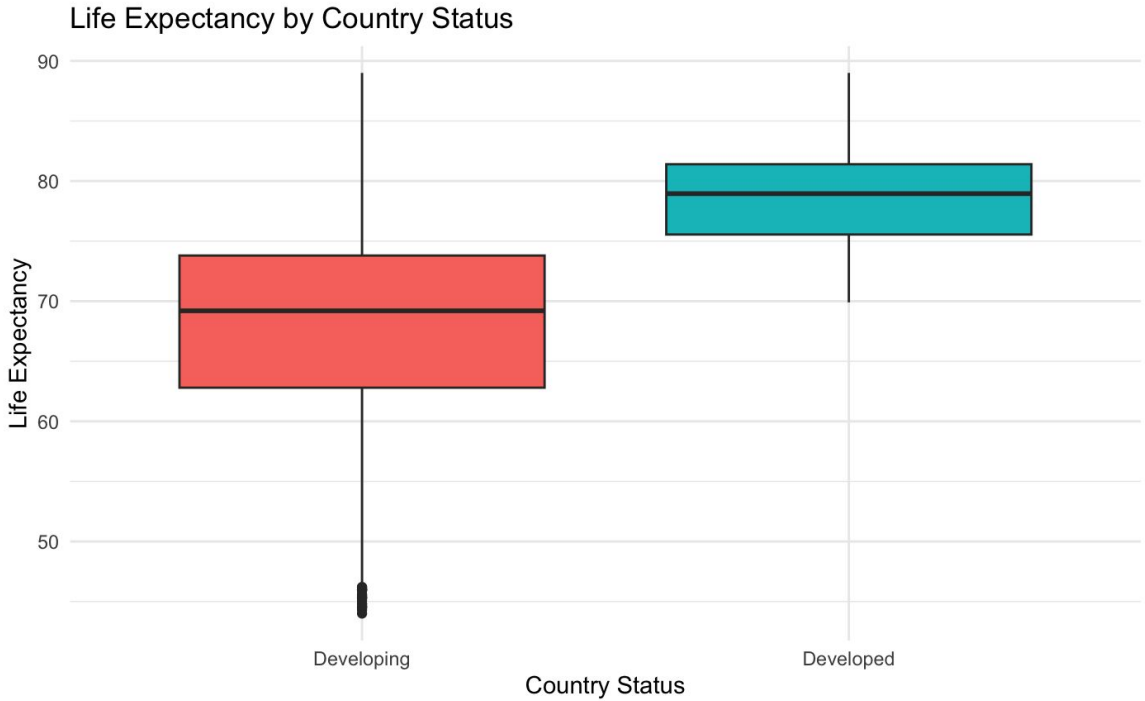
DIIG Data Challenge

F’23, Amy Liu

Question

What sector should the WHO fund to most effectively increase life expectancy?

Developing vs. Developed Disparities



T-test: Significant interactions between development status and life expectancy ($p < 2.2e-16$).

Correlation Tests

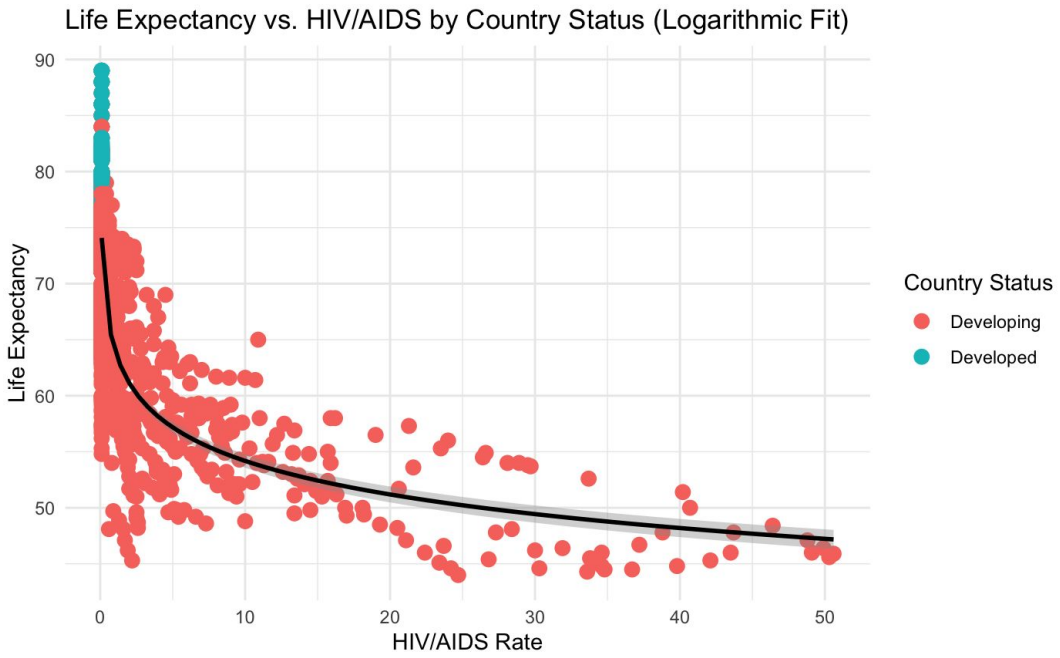
Correlation tests on variables and life expectancy that have p-value < 2.2e-16.

Variables that have the largest (in magnitude) correlation coefficients are adult mortality, HIV/AIDS, income composition of resources, schooling, and thinness (1-19 years and 5-9 years).

Variable <chr>	Correlation_Coefficient <dbl>	P_Value <dbl>
Status	0.4427976	3.922853e-80
Life expectancy	1.0000000	0.000000e+00
Adult Mortality	-0.7025231	1.381910e-245
Alcohol	0.4027183	2.517230e-65
percentage expenditure	0.4096308	9.738022e-68
BMI	0.5420416	1.400348e-126
Polio	0.3272944	1.794250e-42
Diphtheria	0.3413312	2.861774e-46
HIV/AIDS	-0.5922363	1.109672e-156
GDP	0.4413218	1.496282e-79

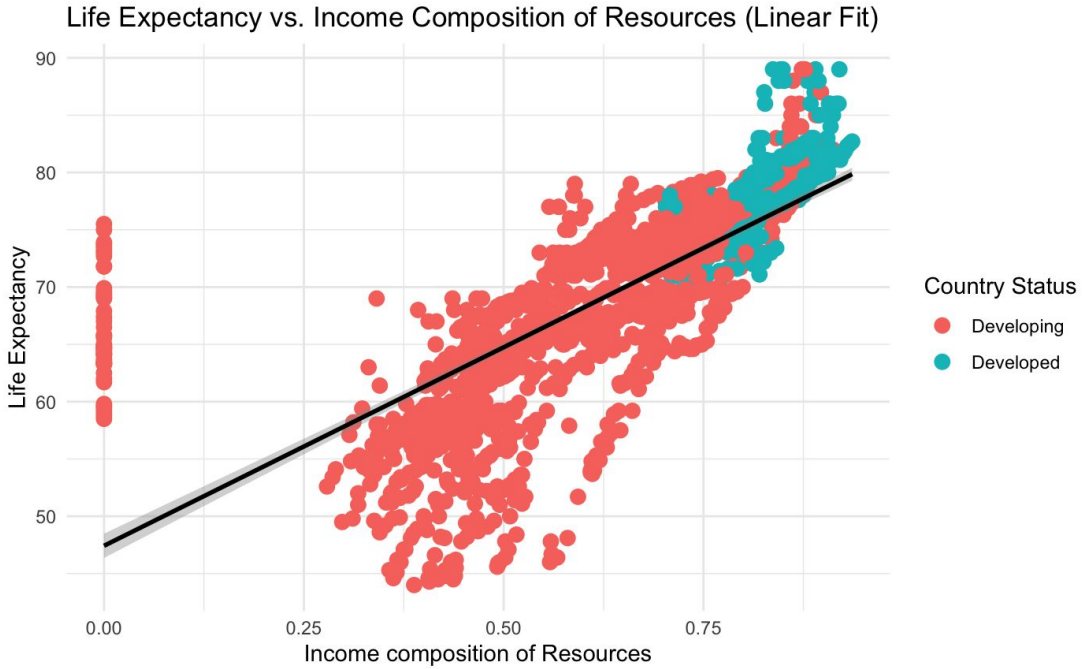
Variable <chr>	Correlation_Coefficient <dbl>	P_Value <dbl>
thinness 1-19 years	-0.4578382	3.174107e-86
thinness 5-9 years	-0.4575083	4.350603e-86
Income composition of resources	0.7210826	9.272181e-265
Schooling	0.7276300	6.694044e-272

Life Expectancy and HIV/AIDS



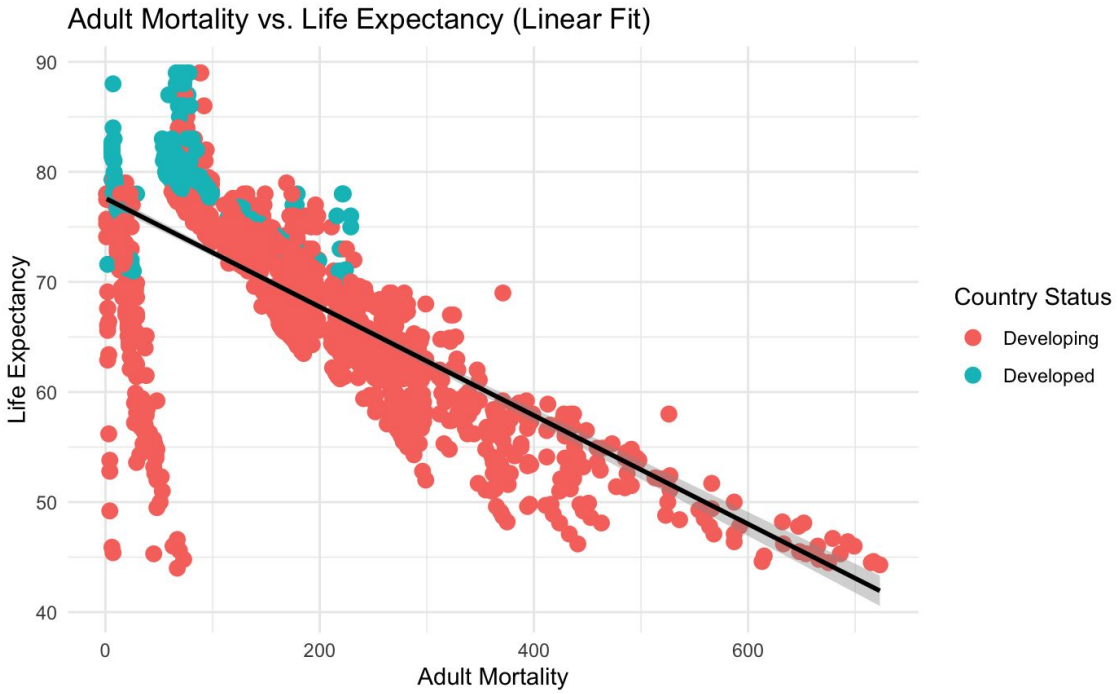
As the HIV/AIDS rate increases, life expectancy is found to decrease. We can see that developed countries are clustered towards having a low HIV/AIDS rate and higher life expectancy. There is a roughly logistic relationship between these variables.

Life Expectancy and Income Composition of Resources



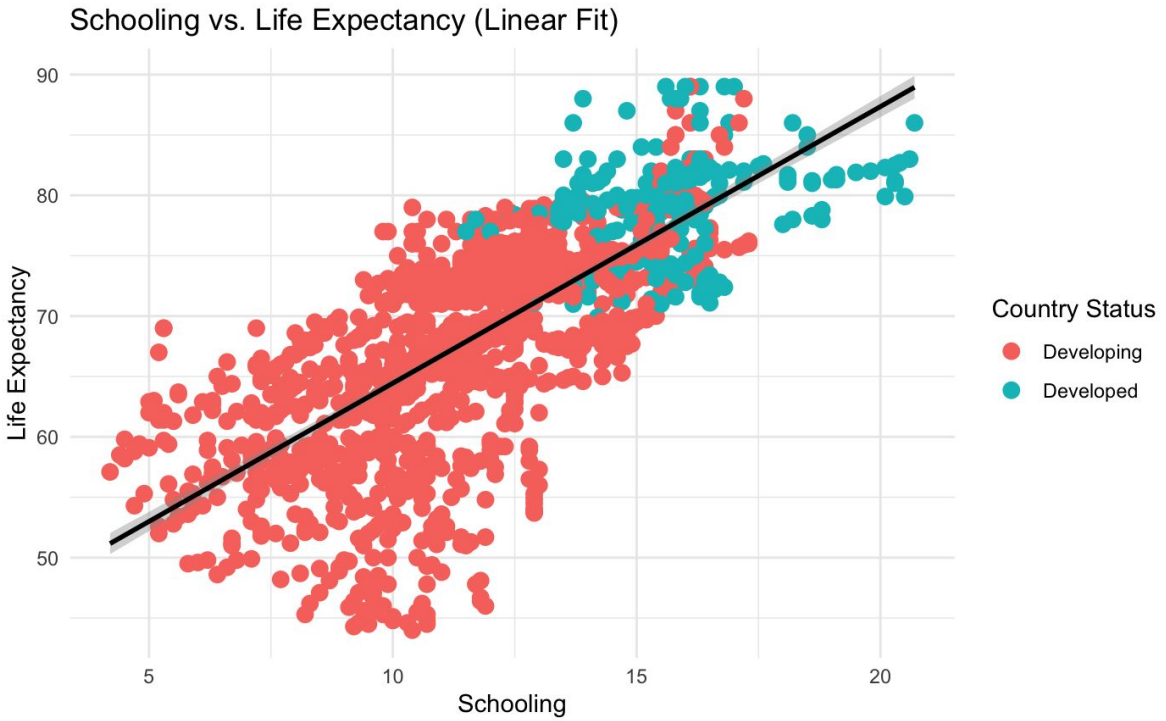
Developed countries have higher life expectancies and higher income composition of resources. There is a roughly linear and positive relationship with a few select outliers.

Life Expectancy and Adult Mortality



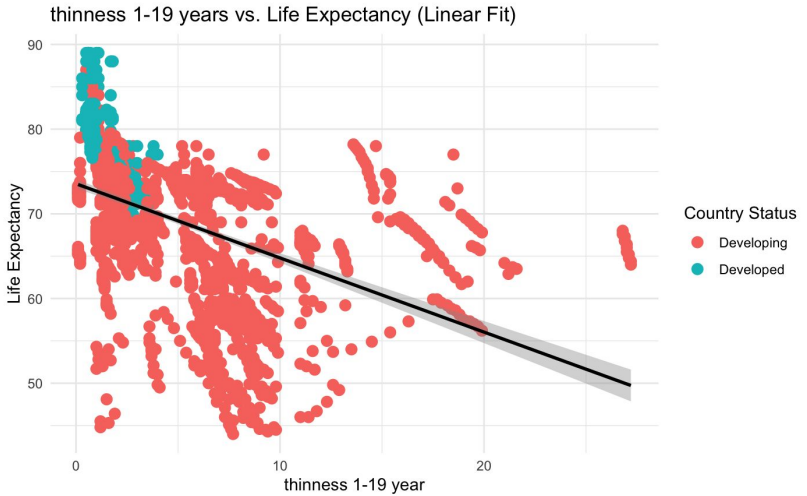
Developed countries have lower rates of adult mortality and high life expectancy (to be expected, since adult mortality is defined as “probability of dying between 15 – 60 years per 1000 population” and life expectancy is defined as “average period a person is expected to live.”

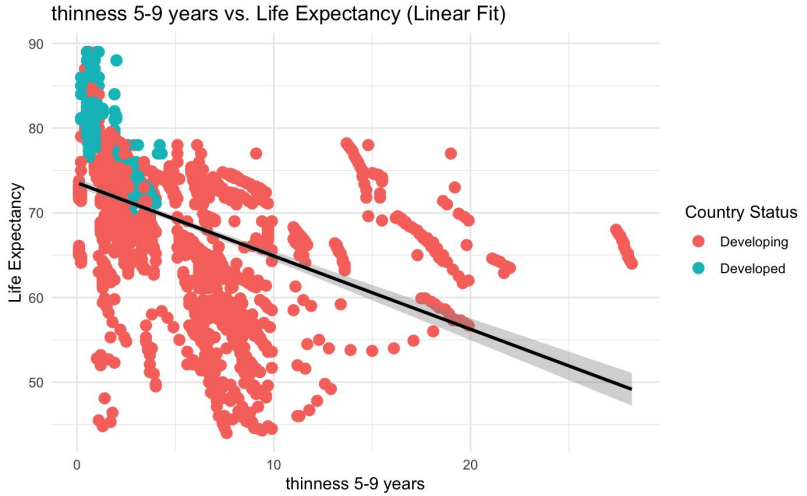
Life Expectancy and Schooling



We see a roughly linear, positive relationship between schooling and life expectancy. We also see that developed countries have higher life expectancies and tend to have more years of schooling.

Life Expectancy and Thinness





Both thinness 5–9 and 1–19 years have a negative relationship with life expectancy. Developed countries are clustered towards higher life expectancy and lower thinness.

Linear Regression Model

```
Coefficients:
                                  Estimate Std. Error t value Pr(>|t|)
(Intercept)                       48.68222    0.56712  85.840  < 2e-16 ***
poly(`HIV/AIDS`, 2)1            -157.35726    4.02814 -39.065  < 2e-16 ***
poly(`HIV/AIDS`, 2)2              72.06234    4.00779  17.981  < 2e-16 ***
`Income composition of resources`  11.94416    0.86146  13.865  < 2e-16 ***
Schooling                          1.13474    0.05696  19.923  < 2e-16 ***
thinness1.19                       0.01145    0.05650   0.203  0.83946
thinness5.9                       -0.14910    0.05523  -2.699  0.00702 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.878 on 1642 degrees of freedom
Multiple R-squared:  0.8063,    Adjusted R-squared:  0.8056
F-statistic:  1139 on 6 and 1642 DF,  p-value: < 2.2e-16
```

Decided to drop **adult mortality**, as their definitions are inverse and the relationship is highly expected.

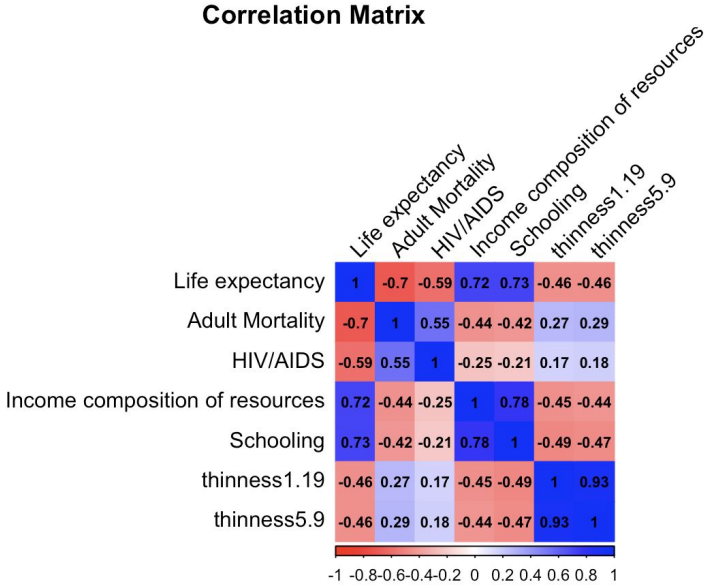
Why a linear regression model? Most of our variables seem to have a linear relationship with life expectancy (from EDA plotting).

Why the polynomial terms? Life expectancy and HIV/AIDS seem to have a logistic relationship. However, since our other variables have a linear relationship, using a polynomial term correction in our linear model allows us to capture both HIV/AIDS’ negative linear relationship with life expectancy along with the curvature in the relationship.

Checking for Collinearity

```
`HIV/AIDS` `Income composition of resources`                Schooling
  1.074197                             2.689313             2.770076
thinness1.19                         thinness5.9
  7.397473                             7.238721
```

VIF for thinness variables are above 5.



Correlation between thinness variables are very high (0.93).

Correlation between schooling and income composition of resources are also worth investigating (0.78).

Tweaking Model for Thinness

```
Coefficients:
                                  Estimate Std. Error t value Pr(>|t|)
(Intercept)                       48.70420    0.55649  87.521  < 2e-16 ***
poly(`HIV/AIDS`, 2)1            -157.37438    4.02608 -39.089  < 2e-16 ***
poly(`HIV/AIDS`, 2)2              72.06556    4.00659  17.987  < 2e-16 ***
`Income composition of resources`  11.93788    0.86065  13.871  < 2e-16 ***
Schooling                          1.13374    0.05672  19.987  < 2e-16 ***
thinness5.9                       -0.13898    0.02354  -5.903 4.34e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.877 on 1643 degrees of freedom
Multiple R-squared:  0.8063,    Adjusted R-squared:  0.8057
F-statistic:  1368 on 5 and 1643 DF,  p-value: < 2.2e-16
```

Dropping Thinness 1-19 years (less statistically significant). Now, all variables are statistically significant, and F-statistic is higher.

Comparison of Models

Model 1 (last slide)

```
Coefficients:
                                  Estimate Std. Error t value Pr(>|t|)
(Intercept)                       48.70420    0.55649  87.521  < 2e-16 ***
poly(`HIV/AIDS`, 2)1            -157.37438    4.02608 -39.089  < 2e-16 ***
poly(`HIV/AIDS`, 2)2              72.06556    4.00659  17.987  < 2e-16 ***
`Income composition of resources`  11.93788    0.86065  13.871  < 2e-16 ***
Schooling                          1.13374    0.05672  19.987  < 2e-16 ***
thinness5.9                       -0.13898    0.02354  -5.903 4.34e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.877 on 1643 degrees of freedom
Multiple R-squared:  0.8063,    Adjusted R-squared:  0.8057
F-statistic:  1368 on 5 and 1643 DF,  p-value: < 2.2e-16
```

Model 2 (no income composition)

```
Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       49.64520    0.58361   85.07  < 2e-16 ***
poly(`HIV/AIDS`, 2)1 -164.74426    4.21678  -39.07  < 2e-16 ***
poly(`HIV/AIDS`, 2)2   78.48312    4.20508   18.66  < 2e-16 ***
Schooling               1.69277    0.04218   40.14  < 2e-16 ***
thinness5.9            -0.17505    0.02473   -7.08 2.13e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.097 on 1644 degrees of freedom
Multiple R-squared:  0.7837,    Adjusted R-squared:  0.7831
F-statistic:  1489 on 4 and 1644 DF,  p-value: < 2.2e-16
```

Model 3 (no schooling)

```
Coefficients:
                                  Estimate Std. Error t value Pr(>|t|)
(Intercept)                       55.25329    0.50134 110.210   <2e-16 ***
poly(`HIV/AIDS`, 2)1            -158.33232    4.48723 -35.285   <2e-16 ***
poly(`HIV/AIDS`, 2)2              76.22026    4.45981  17.090   <2e-16 ***
`Income composition of resources`  24.15961    0.67505  35.789   <2e-16 ***
thinness5.9                       -0.24635    0.02555  -9.641   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.321 on 1644 degrees of freedom
Multiple R-squared:  0.7593,    Adjusted R-squared:  0.7587
F-statistic:  1296 on 4 and 1644 DF,  p-value: < 2.2e-16
```

RSE is lowest and **multiple r-squared** and **adjusted r-squared** are highest for our model including both income composition of resources and schooling.

Since the correlation between those two variables is below 0.8 and VIF for both are below 5, and our model's performance (RSE and multiple r-squared) when including both variables outweighs the difference in the F-statistic (1368 in Model 1 vs 1489 in Model 2), we will keep both variables in our model (Model 1).

Suggestions

HIV/AIDS

Increase sex education and awareness, access to HIV testing, condom distribution

Income composition

Reduce income inequality: social welfare programs, supporting rural development

Schooling

School infrastructure development and improvements, funding transportation to school

Thinness

Combat food insecurity: free and reduced lunch programs

Thank you!

Do you have any questions?

al577@duke.edu
https://www.linkedin.com/in/amyhliu27/