DATA606 - Data Project

Leland Randles

December 11, 2016

Introduction

The 2016 Presidential Election brought renewed attention to the marked differences between rural and urban communities throughout the US, but such disparities are not a new topic; in fact, the "mass migration" from rural communities to cities has been frequently referenced by the media, academics and politicians over the last few decades, and the migration has occurred worldwide, not just in the US.

Another topic which has received increased attention over the same time span has been income and wealth inequality; which, at least anecdotally, would appear to be related to the worldwide migration from rural communities to cities. City residents have been more likely to experience salary increases and rising property values, while rural communities have experienced stagnant wages and relatively flat property values. My research questions are an attempt to begin to expore this relationship using data sets from the World Bank web site.

Research Questions: Has there indeed been a worldwide migration from rural areas to urban communities from 1986-2015? To what degree has GNI per capita trended with urban population percentages over this same time period and does the relationship vary by World Bank country income groups?

GNI stands for 'Gross National Income', which is similar to GNP (gross national product). Despite 'national' being in the acronym, the World Bank also compiles this information at regional levels in addition to country-level numbers. When I refer to World Bank regions, I will be referring to the following categories:

Code	Region
EAS	East Asia & Pacific
ECS	Europe & Central Asia
LCN	Latin America & Caribbean
NAC	North America
SAS	South Asia
SSF	Sub-Saharan Africa

Data

Data Collection
The data source used is the World Bank's Health Nutrition and Population Statistics data set, which is part of their HealthStats database: http://data.worldbank.org/data-catalog/health-nutrition-and-population-statistics. A detailed discussion of their data collection sources and methods can be found on pages 159-172 of the pdf found here: https://openknowledge.worldbank.org/bitstream/handle/10986/23969/9781464806834.pdf. In brief, the data primarily comes from national statistical agencies, central banks, and customs services, which admittedly vary in quality by country. The World Bank also tracks the statistical capacity for each country: http://datatopics.worldbank.org/statisticalcapacity/.

Cases
The tidied data set for the region file includes 180 observations. Each observation is a region/year combination, for which there are population and GNI variables. The tidied data set for the country file includes 3,420 observations, each of which is a country/year combination.

Variables
* GNI Per Capita (in Current US Dollars) * Relative GNI Per Capita (a scaled value to eliminate overall

income differences between countries or regions) * Country/Region Code * Year * Rural Population * Rural population (% of total population) * Urban Population * Urban population (% of total) * Income Group (a country is either Low Income, Lower Middle Income, Upper Middle Income, or High Income)

Type of Study
This is an observational study. The data was collected and assembled after the fact in a way which does not interfere with how the data is collected.

Scope of Inference - Generalizability
The data set is a population data set, not a sample, though because many countries have incomplete data, I only used data for the 114 countries and 6 regions which have data for every year between 1986 and 2015.

I don't think it would be sound to generalize the conclusions drawn from my limited data set to all countries worldwide.

Scope of Inference - Causality
Because this is an observational study assmebled from data collected in the past, no causality conclusions can be drawn from the study.

Exploratory Data Analysis

```
install.packages("tidyr",repos='http://mirrors.nics.utk.edu/cran/')
install.packages("ggplot2",repos='http://mirrors.nics.utk.edu/cran/')
library(tidyr)
library(dplyr)
library(ggplot2)
```

The World Bank's Health Nutrition and Population ("HNP") Statistics data set consists of 6 files, of which two files (the country file and the data file) were joined to create two distinct raw data sets - one showing HNP stats by Region, and another showing HNP stats by Country. Below I list a key assumption and two actions taken while creating the two data sets:

1. Only countries and regions which had values for every year between 1986-2015 for GNI Per Capita, Rural Population, and Urban Population were included. Any countries or regions missing data for any of these years were discarded.
2. In addition to the measurements included on the World Bank file, another data element, called "Relative GNI Per Capita", was computed and added to the file. Relative GNI Per Capita sets the 1986 value as 1.00 (the baseline value) and then computes relative values for subsequent years. For example, if the 1986 GNI Per Capita is 1,000 and the 1987 GNI Per Capita is 1,100, then the 1986 Relative GNI Per Capita is 1.0 and the 1987 Relative GNI Per Capita is 1.1. I created this data element to eliminate differences between absolute income levels by country (for example, the US GNI Per Capita night be 5 times the Trinidad and Tobago GNI Per Capita - but we're really interested in the relative increase/decrease in GNI with respect to population changes, not the absolute GNI Per Capita amounts).

3. I also created one more data element, called "Relative Urban Popuilation Percentage" using the same kind of logic used for the "Relative GNI Per Capita" data element. The idea to for a country whose urban population percentage rose from 30% to 45% in the time span to have the same effect in a linear regression model as a country whose urban population percentage rose from 45% to 60%.

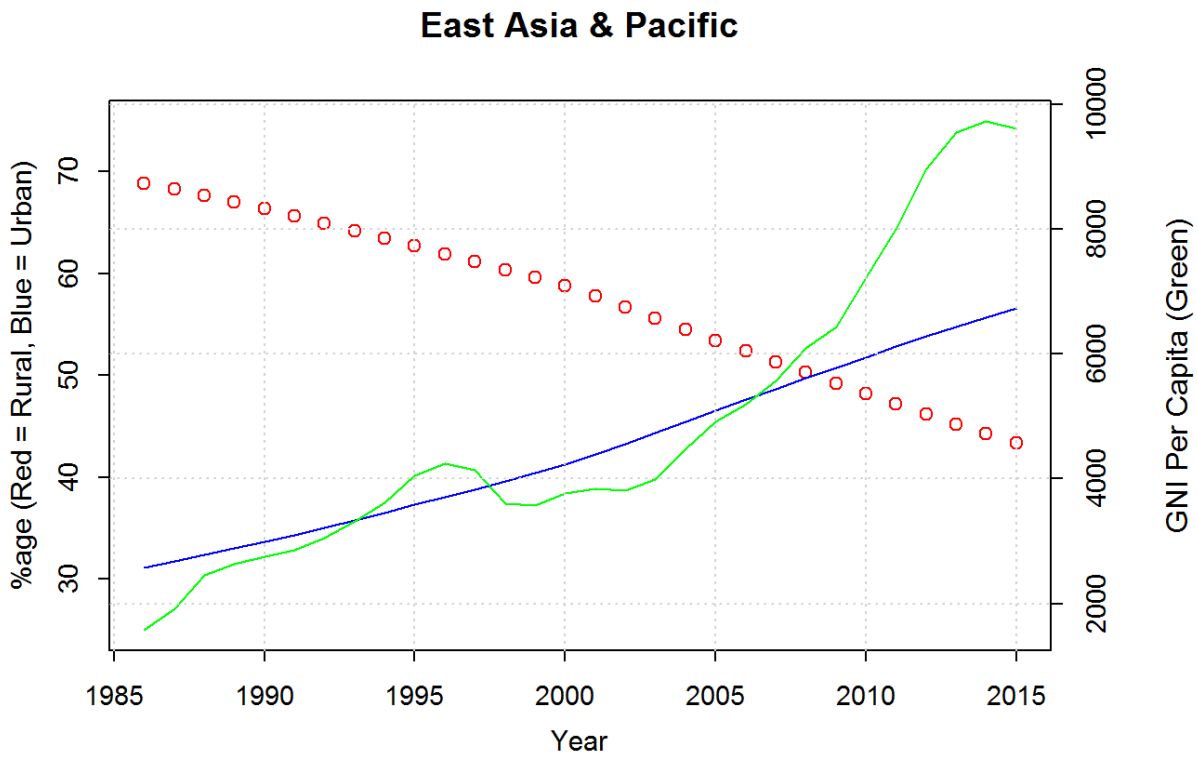
```
# load data - start by reading csv
wb_region_raw <- read.csv("C:/Users/Lelan/Documents/CUNY/DATA606/Homework/Project/Datasets/HNP_Region_Da
# clean up some column names
colnames(wb_region_raw) <- gsub("\\.", "_", colnames(wb_region_raw))
```

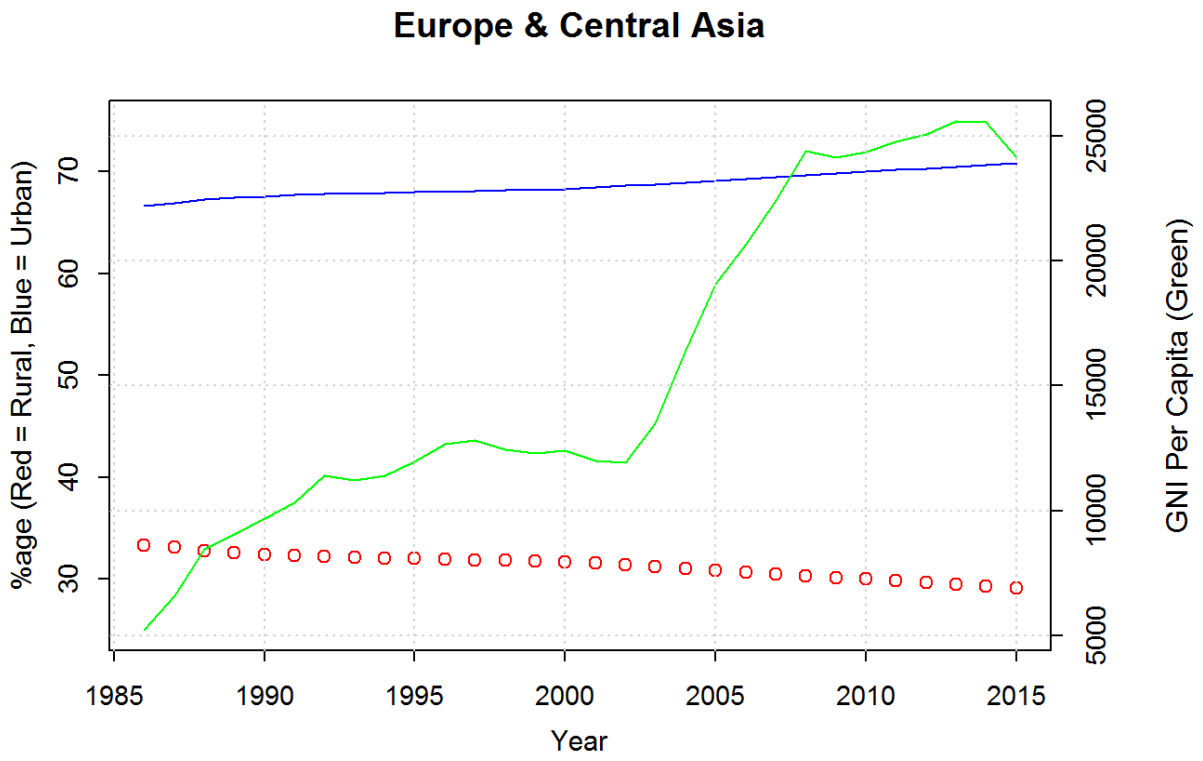
```
options(scipen=999)
# tidy and transform
wb_region <- wb_region_raw %>% gather("Year", "n", 5:34) %>% transmute(Country_Name = Country_Name, Cou
wb_region <- wb_region %>% spread(Indicator_Name, Value)
colnames(wb_region)[4:11] <- c("GNI_Per_Capita", "Relative_GNI_Per_Capita", "Rural_Population", "Rural_
```

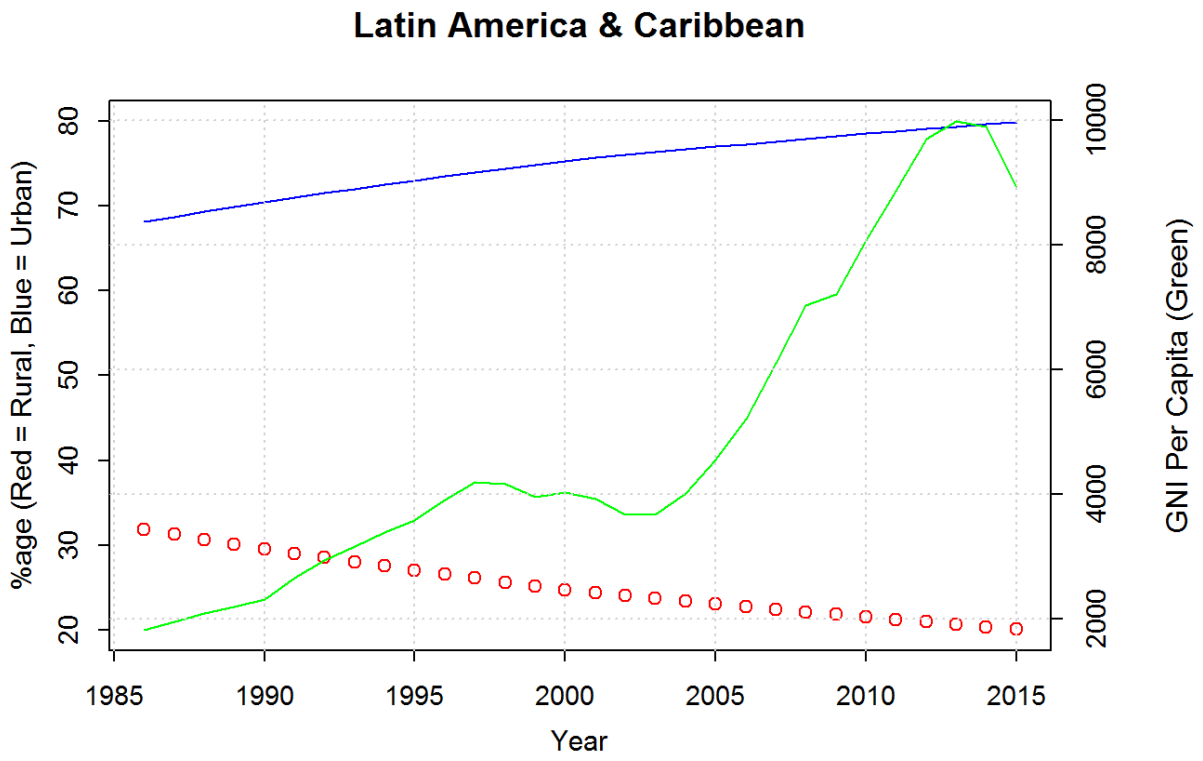
To begin exploratory data analysis, I looked at plots of each region's rural and urban population percentages and GNI Per Capita by Year to see if there has in fact been population migration from rural to city in each region, and whether GNI Per Capita has trended in concert with this migration:

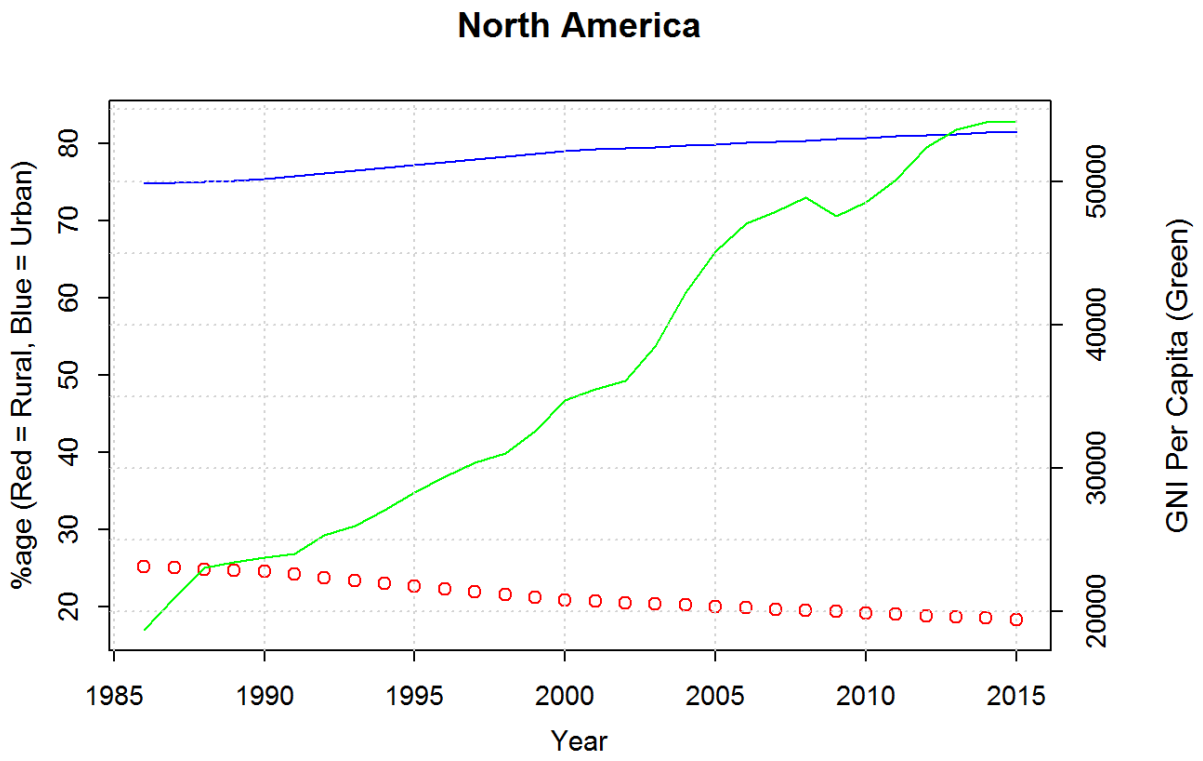
```
# Create subsets by region
wb_region_eas <- subset(wb_region, Country_Code == "EAS")
wb_region_ecs <- subset(wb_region, Country_Code == "ECS")
wb_region_lcn <- subset(wb_region, Country_Code == "LCN")
wb_region_nac <- subset(wb_region, Country_Code == "NAC")
wb_region_sas <- subset(wb_region, Country_Code == "SAS")
wb_region_ssf <- subset(wb_region, Country_Code == "SSF")

# East Asia & Pacific plot
par(mar = c(5, 4, 4, 4) + 0.3)
plot(wb_region_eas$Year, wb_region_eas$Rural_Population_Perc, xlab = "Year", ylab = "%age (Red = Rural,
lines(wb_region_eas$Year, wb_region_eas$Urban_Population_Perc, col = "Blue")
par(new = TRUE)
plot(wb_region_eas$Year, wb_region_eas$GNI_Per_Capita, type = "l", axes = FALSE, bty = "n", xlab = "", y
axis(side=4, at = pretty(range(wb_region_eas$GNI_Per_Capita)))
mtext("GNI Per Capita (Green)", side=4, line=3)
title(main = "East Asia & Pacific")
grid()

# Europe & Central Asia plot
par(mar = c(5, 4, 4, 4) + 0.3)
plot(wb_region_ecs$Year, wb_region_ecs$Rural_Population_Perc, xlab = "Year", ylab = "%age (Red = Rural,
lines(wb_region_ecs$Year, wb_region_ecs$Urban_Population_Perc, col = "Blue")
par(new = TRUE)
plot(wb_region_ecs$Year, wb_region_ecs$GNI_Per_Capita, type = "l", axes = FALSE, bty = "n", xlab = "", y
axis(side=4, at = pretty(range(wb_region_ecs$GNI_Per_Capita)))
mtext("GNI Per Capita (Green)", side=4, line=3)
title(main = "Europe & Central Asia")
grid()

# Latin America & Caribbean plot
par(mar = c(5, 4, 4, 4) + 0.3)
plot(wb_region_lcn$Year, wb_region_lcn$Rural_Population_Perc, xlab = "Year", ylab = "%age (Red = Rural,
lines(wb_region_lcn$Year, wb_region_lcn$Urban_Population_Perc, col = "Blue")
par(new = TRUE)
plot(wb_region_lcn$Year, wb_region_lcn$GNI_Per_Capita, type = "l", axes = FALSE, bty = "n", xlab = "", y
axis(side=4, at = pretty(range(wb_region_lcn$GNI_Per_Capita)))
mtext("GNI Per Capita (Green)", side=4, line=3)
title(main = "Latin America & Caribbean")
grid()

# North America plot
par(mar = c(5, 4, 4, 4) + 0.3)
```

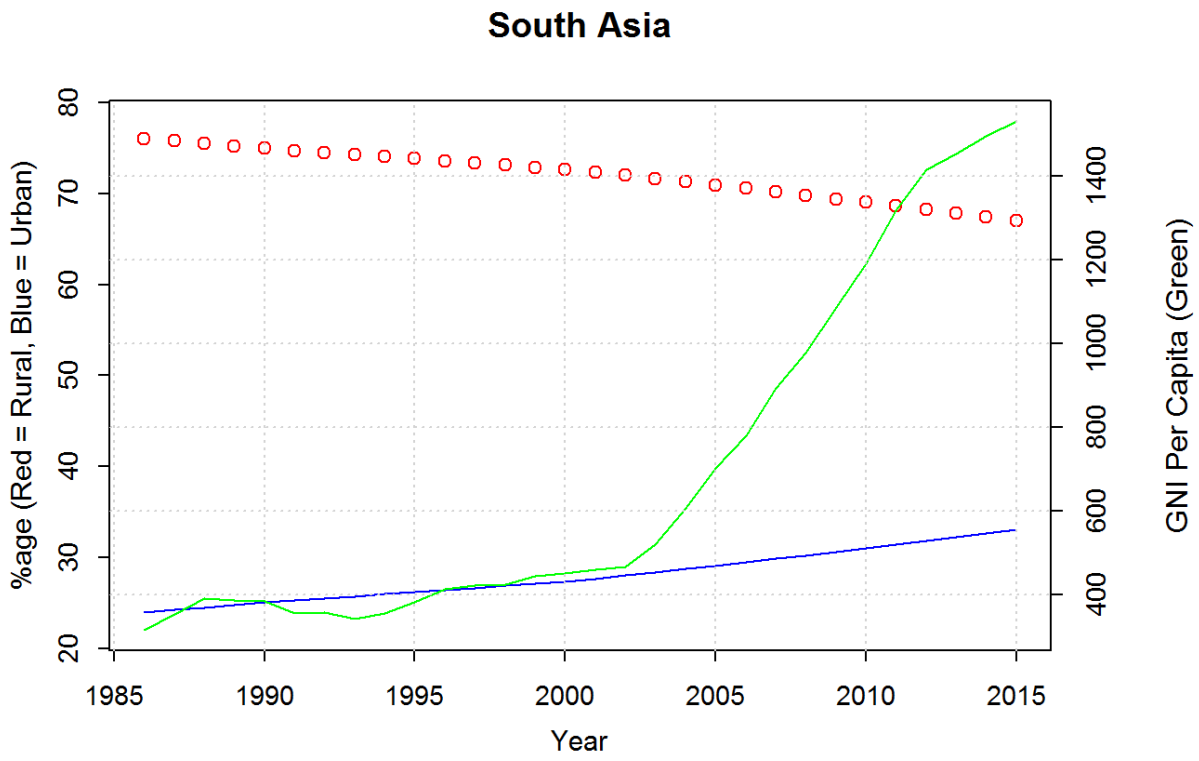
```r
plot(wb_region_nac$Year, wb_region_nac$Rural_Population_Perc, xlab = "Year", ylab = "%age (Red = Rural,
lines(wb_region_nac$Year, wb_region_nac$Urban_Population_Perc, col = "Blue")
par(new = TRUE)
plot(wb_region_nac$Year, wb_region_nac$GNI_Per_Capita, type = "l", axes = FALSE, bty = "n", xlab = "", y
axis(side=4, at = pretty(range(wb_region_nac$GNI_Per_Capita)))
mtext("GNI Per Capita (Green)", side=4, line=3)
title(main = "North America")
grid()

# South Asia plot
par(mar = c(5, 4, 4, 4) + 0.3)
plot(wb_region_sas$Year, wb_region_sas$Rural_Population_Perc, xlab = "Year", ylab = "%age (Red = Rural,
lines(wb_region_sas$Year, wb_region_sas$Urban_Population_Perc, col = "Blue")
par(new = TRUE)
plot(wb_region_sas$Year, wb_region_sas$GNI_Per_Capita, type = "l", axes = FALSE, bty = "n", xlab = "", y
axis(side=4, at = pretty(range(wb_region_sas$GNI_Per_Capita)))
mtext("GNI Per Capita (Green)", side=4, line=3)
title(main = "South Asia")
grid()

# Sub-Saharan Africa plot
par(mar = c(5, 4, 4, 4) + 0.3)
plot(wb_region_ssf$Year, wb_region_ssf$Rural_Population_Perc, xlab = "Year", ylab = "%age (Red = Rural,
lines(wb_region_ssf$Year, wb_region_ssf$Urban_Population_Perc, col = "Blue")
par(new = TRUE)
plot(wb_region_ssf$Year, wb_region_ssf$GNI_Per_Capita, type = "l", axes = FALSE, bty = "n", xlab = "", y
axis(side=4, at = pretty(range(wb_region_ssf$GNI_Per_Capita)))
mtext("GNI Per Capita (Green)", side=4, line=3)
title(main = "Sub-Saharan Africa")
grid()
```

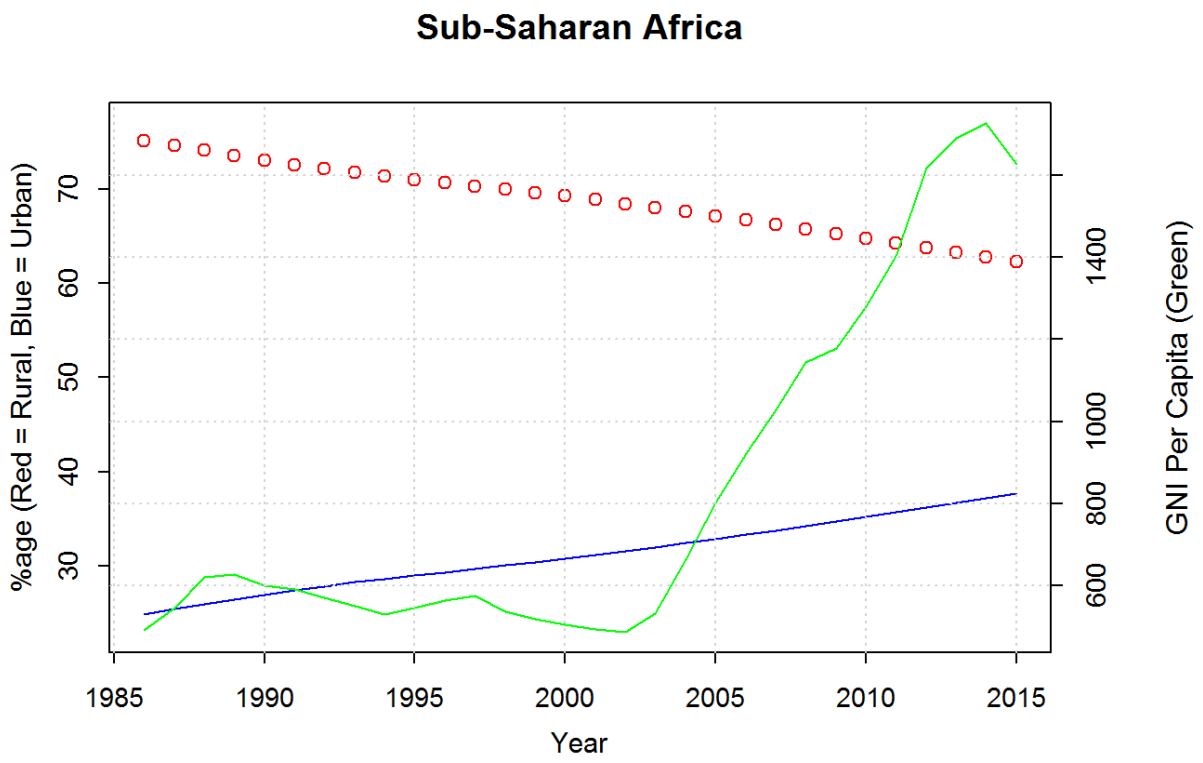

East Asia & Pacific
%age (Red = Rural, Blue = Urban)
GNI Per Capita (Green)
Year
30
40
50
60
70
2000
4000
6000
8000
10000
1985
1990
1995
2000
2005
2010
2015


Europe & Central Asia
%age (Red = Rural, Blue = Urban)
30
40
50
60
70
5000
10000
15000
20000
25000
GNI Per Capita (Green)
1985
1990
1995
2000
2005
2010
2015
Year


Latin America & Caribbean
%age (Red = Rural, Blue = Urban)
GNI Per Capita (Green)
Year
20
30
40
50
60
70
80
2000
4000
6000
8000
10000
1985
1990
1995
2000
2005
2010
2015

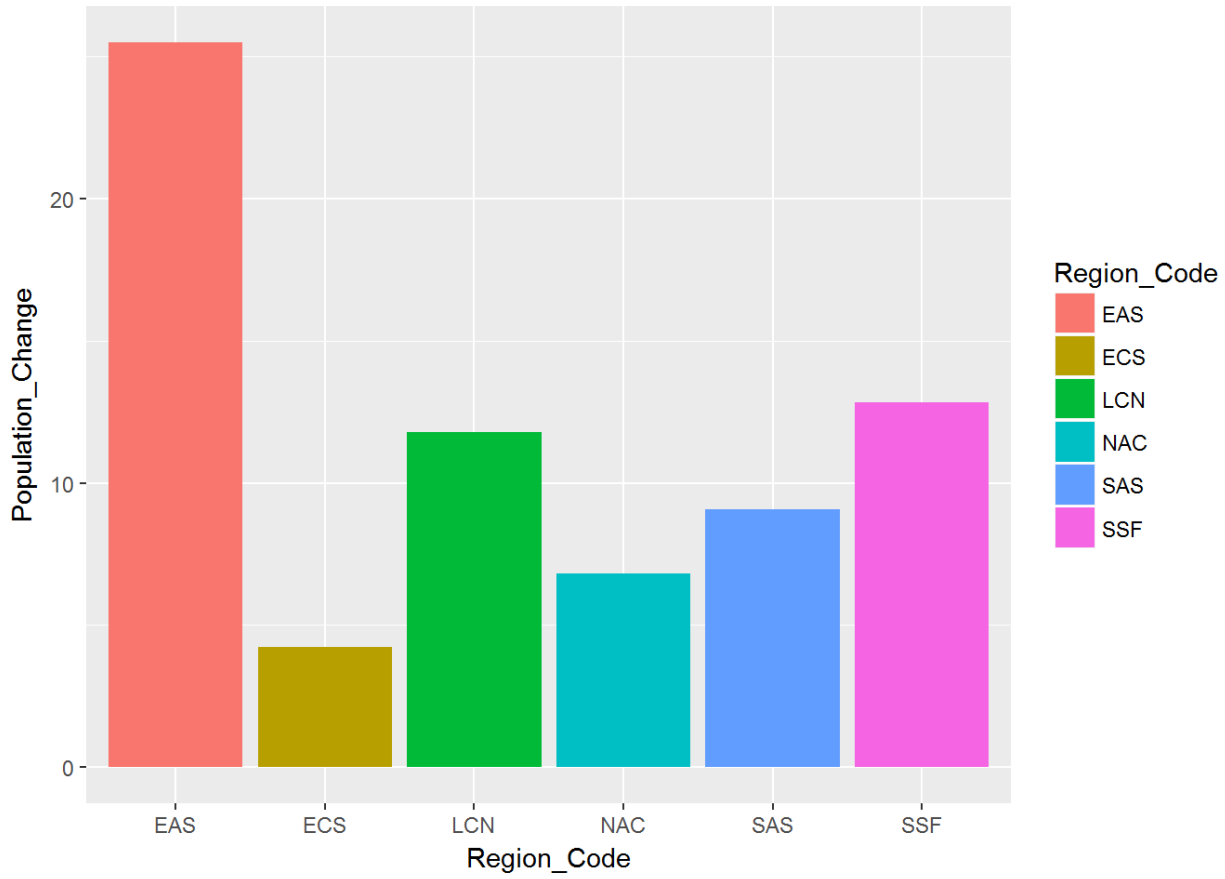

North America
%age (Red = Rural, Blue = Urban)
20
30
40
50
60
70
80
1985
1990
1995
2000
2005
2010
2015
Year
20000
30000
40000
50000
GNI Per Capita (Green)


South Asia
%age (Red = Rural, Blue = Urban)
GNI Per Capita (Green)
Year
20
30
40
50
60
70
80
400
600
800
1000
1200
1400
1985
1990
1995
2000
2005
2010
2015



Looking at the graphs, we see that there has been migration from rural to urban populations (though the degree to which varies by region), and all regions have experienced growth in their GNI Per Capita. The bar chart below shows the percentage population change for each region:

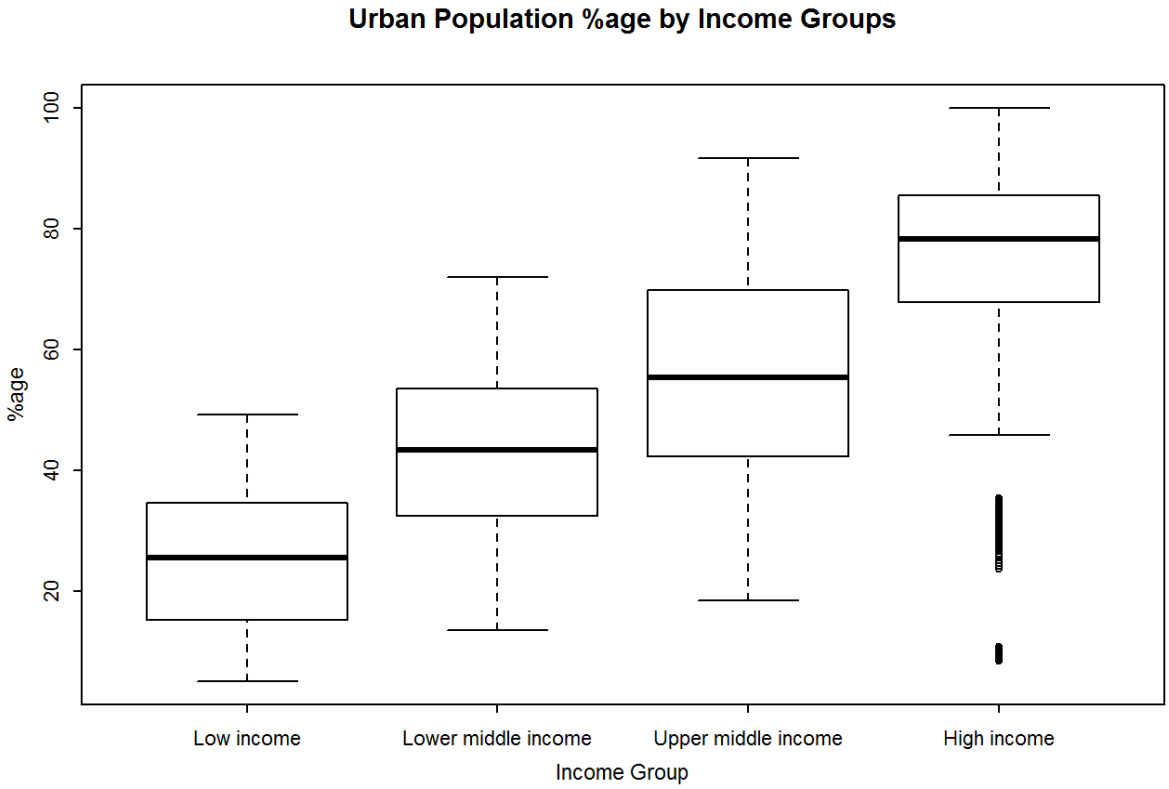
```
xa <- c("EAS","ECS","LCN","NAC","SAS","SSF")
ya <- c(max(wb_region_eas$Urban_Population_Perc) - min(wb_region_eas$Urban_Population_Perc), max(wb_reg
pop_chg <- as.data.frame(cbind(xa, as.numeric(ya)), stringsAsFactors = FALSE)
pop_chg$V2 <- as.numeric(pop_chg$V2)
colnames(pop_chg)[1:2] <- c("Region_Code","Population_Change")
ggplot(data=pop_chg, aes(x=Region_Code, y=Population_Change, fill=Region_Code)) + geom_bar(stat="identi
```

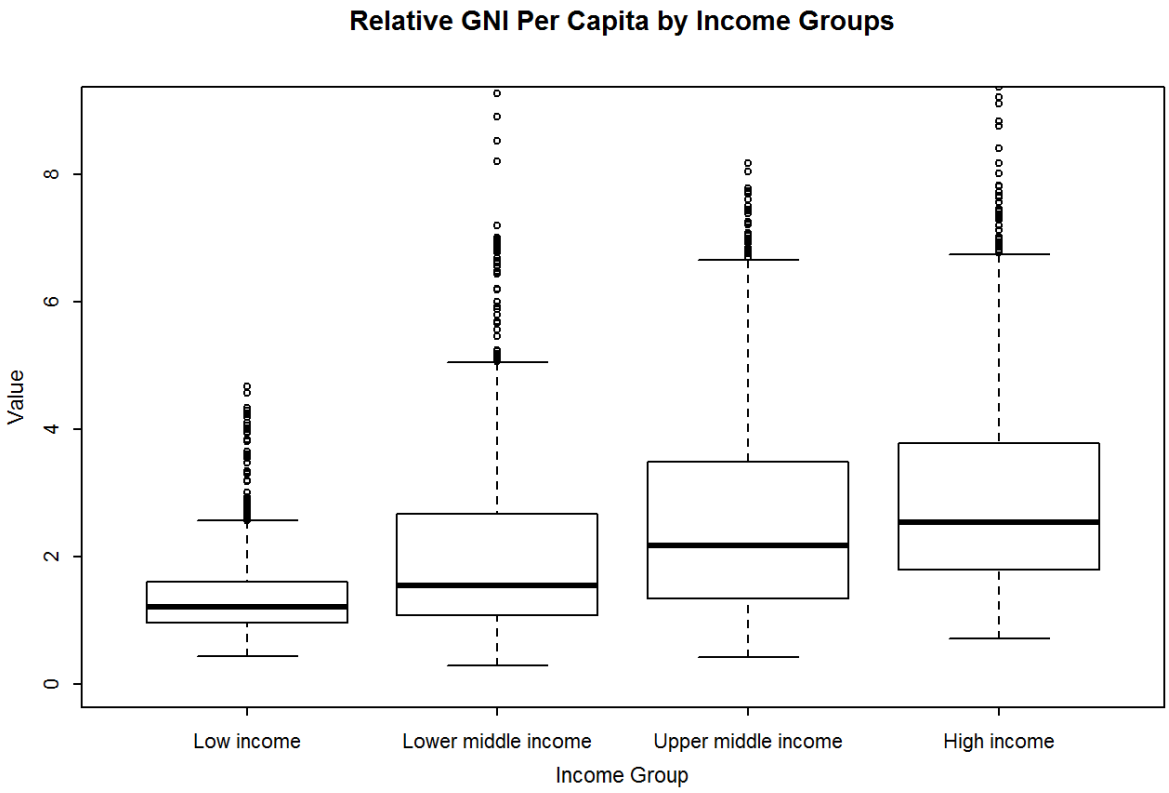


The descriptive statistics and visuals were interesting enough that I moved on to studying the Country data set. The Country data includes the "Income Group" categorical variable, which classifies a country as Low Income, Lower Middle Income, Upper Middle Income, or High Income. We can look at box plots by income group for Urban Population Percentage and Relative GNI Per Capita and see if there are differences. If the migration from rural to urban is part of what drives income inequality, we would expect to see more pronounced differences for higher income group countries than lower group income countries.

```
# load data - start by reading csv
wb_cntry_raw <- read.csv("C:/Users/Lelan/Documents/CUNY/DATA606/Homework/Project/Datasets/HNP_Country_Da
# clean up some column names
colnames(wb_cntry_raw) <- gsub("\\.", "_", colnames(wb_cntry_raw))
options(scipen=999)
# tidy and transform
wb_cntry <- wb_cntry_raw %>% gather("Year", "n", 8:37) %>% transmute(Country_Name = Country_Name, Counti
wb_cntry <- wb_cntry %>% spread(Indicator_Name, Value)
colnames(wb_cntry)[7:15] <- c("GNI_Per_Capita", "Relative_GNI_Per_Capita", "Relative_Urban_Population_Pe
```

```
inc_grp <- factor(wb_cntry$Income_Group, levels = c("Low income","Lower middle income","Upper middle inc

# Side by side boxplots by Urban Population Percentage
par(cex = 0.75, mar = c(5, 4, 5, 2) + 0.1)
boxplot(Urban_Population_Perc ~ inc_grp, data=wb_cntry, main="Urban Population %age by Income Groups", x
# Side by side boxplots by Urban Population Percentage
boxplot(Relative_GNI_Per_Capita ~ inc_grp, data=wb_cntry, main="Relative GNI Per Capita by Income Groups
```


Urban Population %age by Income Groups
%age
20
40
60
80
100
Low income
Lower middle income
Upper middle income
High income
Income Group



Inference

1st Research Question: *Has there been a worldwide migration from rural areas to urban communities from 1986-2015?*

H_0: $\mu_{diff} = 0$ There is no difference between the worldwide urban population percentage in 1986 and 2015 (null hypothesis)

H_A: $\mu_{diff} > 0$ The worldwide urban population percentage is higher in 2015 than it was in 1986 (alternative hypothesis)

I see this as a one-tailed test.

I am sampling 30 countries from the tidied data set of 114 countries in an attempt to use the techniques taught in the course. In reality the untidied data set was a population data set, not a sample.

```
# Create subsets for 1986 and 2015 values
wb_cntry.86 <- subset(wb_cntry, Year == 1986)
wb_cntry.15 <- subset(wb_cntry, Year == 2015)
# Sample of 30 from population of 114 countries
s <- sample(x = 114, 30)
# Data frame built from sample of 30
inf_df <- as.data.frame(cbind(wb_cntry.86$Country_Code[s], wb_cntry.86$Urban_Population_Perc[s], wb_cnt
# Tidying and adding difference column
inf_df$V2 <- as.numeric(inf_df$V2)
inf_df$V3 <- as.numeric(inf_df$V3)
```

```r
colnames(inf_df)[1:3] <- c("Country_Code", "1986", "2015")
inf_df$Difference <- inf_df$`2015` - inf_df$`1986`
# Calculatio of mean, standard deviation, standard error,
# z-score, p-score, and confidence interval
m <- mean(inf_df$Difference)
sd <- sd(inf_df$Difference)
se <- sd / sqrt(30)
z <- m / se
p <- 1 - pnorm(z)
p
```

```
## [1] 0.000000001293209
```

```r
ci <- c(m - qnorm(0.95) * se, m + qnorm(0.95) * se)
ci
```

```
## [1]  7.467178 13.165289
```

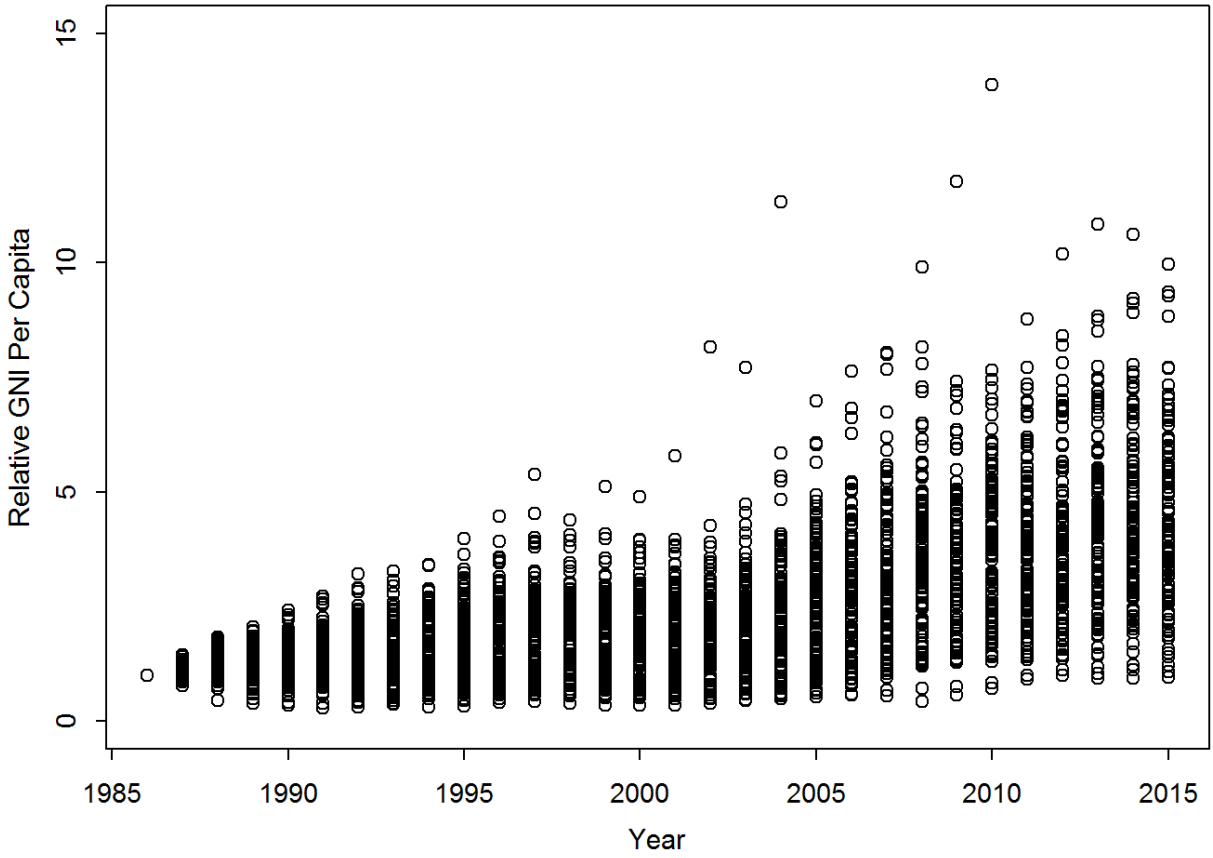
Based on the extremely low p-value and the fact that zero is way outside of the confidence interval, we reject the null hypothesis and conclude that the worldwide urban population percentage is higher in 2015 than it was in 1986.

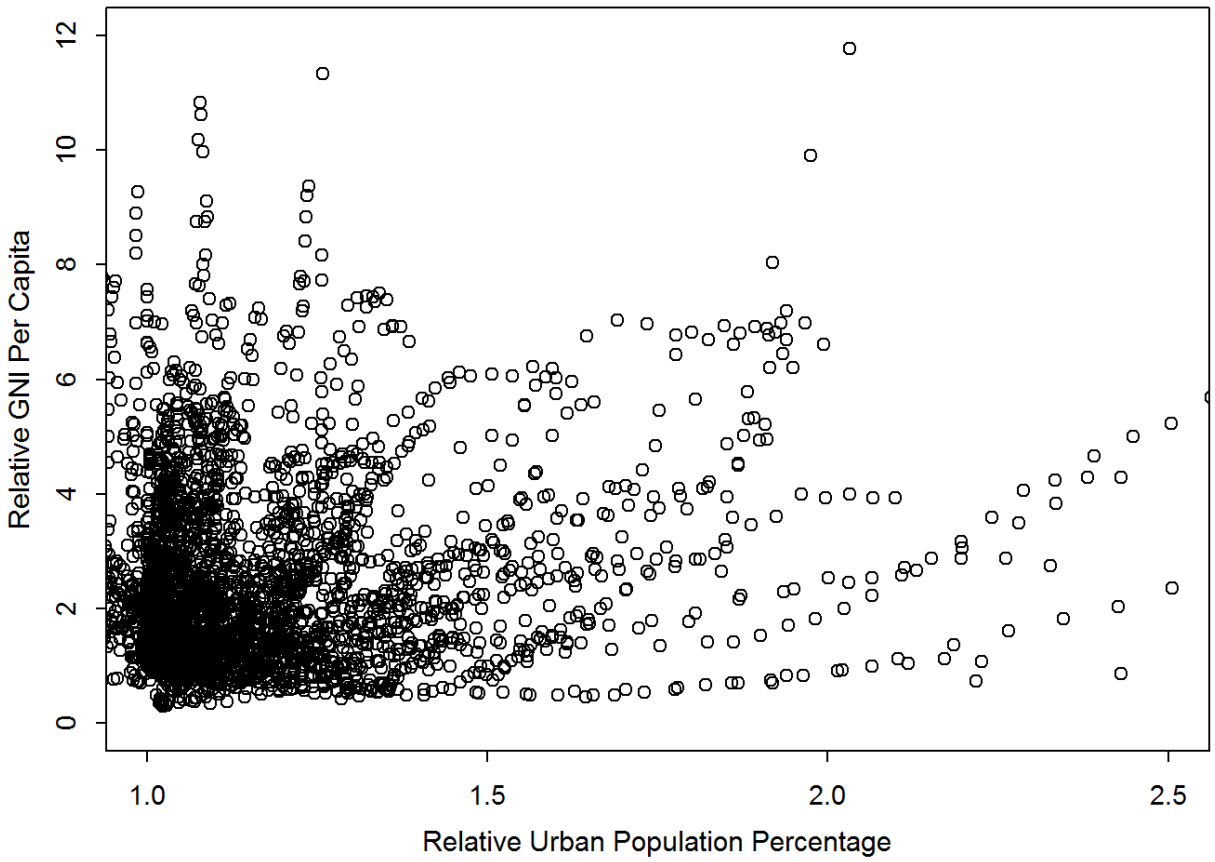
2nd Research Question: *Has GNI per capita trended with urban/rural population percentages during 1986-2015?*

To investigate this, I'll look at scatterplots, correlations, and multiple linear regression, first for Year and Relative Urban Population Percentage vs. Relative GNI Per Capita for the entire population and then by Income Group.

I am including the Year variable because it is a proxy for natural economic growth. The GNI variables are in current US dollars, so inflation is not an issue.

```r
wb_cntry$Year <- as.numeric(wb_cntry$Year)
# Scatterplot of Year vs Relative GNI Per Capita
plot(wb_cntry$Year, wb_cntry$Relative_GNI_Per_Capita, ylim = c(0,15), xlab = "Year", ylab = "Relative GN
# Scatterplot of Relative Urban Population
# Percentage vs Relative GNI Per Capita
plot(wb_cntry$Relative_Urban_Population_Perc, wb_cntry$Relative_GNI_Per_Capita, ylim = c(0,12), xlim = c
```


Relative GNI Per Capita
0
5
10
15
1985
1990
1995
2000
2005
2010
2015
Year



Looking at the scatterplots, there does appear to be a relationship between Year and Relative GNI Per Capita. The scatterplot for Relative Urban Population Percentage and Relative GNI Per Capita is less clear, however. Though there is a general relationship, the heavy cluster of plots near the x/y axis would suggest that there are quite a few countries where there has been minimal growth in either variable.

Next, I computed correlations based on the full population of 114 countries.

```
# Correlation between Year and Relative GNI Per Capita
cor(wb_cntry$Year, wb_cntry$Relative_GNI_Per_Capita)
```

```
## [1] 0.3574315
```

```
# Correlation between Relative Urban Population Percentage and
# Relative GNI Per Capita
cor(wb_cntry$Relative_Urban_Population_Perc, wb_cntry$Relative_GNI_Per_Capita)
```

```
## [1] 0.08051726
```

According to the results, the natural economic growth rate (using Year as a proxy) is more strongly correlated to Relative GNI Per Capita than the Relative Urban Population Percentage. In fact, the correlation between Relative Urban Population Percentage and Relative GNI Per Capita is weak.

Third, I ran a multiple linear regression model modeling the Relative GNI Per Capita using Year and Relative Urban Population Percentage.

```r
# Multiple linear regression
both.lm <- lm(Relative_GNI_Per_Capita ~ Year + Relative_Urban_Population_Perc, data = wb_cntry)
summary(both.lm)
```

```
##
## Call:
## lm(formula = Relative_GNI_Per_Capita ~ Year + Relative_Urban_Population_Perc,
##     data = wb_cntry)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -3.849 -1.000 -0.114  0.495 69.352
##
## Coefficients:
##                                  Estimate  Std. Error t value
## (Intercept)                   -297.800078   13.649846 -21.817
## Year                             0.150415    0.006857  21.936
## Relative_Urban_Population_Perc   -0.418248    0.180593  -2.316
##                                          Pr(>|t|)
## (Intercept)                   <0.0000000000000002 ***
## Year                          <0.0000000000000002 ***
## Relative_Urban_Population_Perc              0.0206 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.285 on 3417 degrees of freedom
## Multiple R-squared:  0.1291, Adjusted R-squared:  0.1286
## F-statistic: 253.3 on 2 and 3417 DF,  p-value: < 0.00000000000000022
```

As you can see, the slope for the Relative Urban Population Percentage in the model that was generated is **negative**, which would indicate that an increase in Relative Urban Population Percentage would result in lower Relative GNP Per Capita, which is exactly opposite of what I've theorized.

But could this be a case of the countries in the low income group weighing down the model? To examine, I created data frames for each income group and ran multiple linear regression models for each income group separately.

```r
# Create income group data frames
wb_cntry.li <- wb_cntry[wb_cntry$Income_Group == "Low income",]
wb_cntry.lmi <- wb_cntry[wb_cntry$Income_Group == "Lower middle income",]
wb_cntry.umi <- wb_cntry[wb_cntry$Income_Group == "Upper middle income",]
wb_cntry.hi <- wb_cntry[wb_cntry$Income_Group == "High income",]
# Low income group multiple regression model
li.both.lm <- lm(Relative_GNI_Per_Capita ~ Year + Relative_Urban_Population_Perc, data = wb_cntry.li)
summary(li.both.lm)
```

```
##
## Call:
## lm(formula = Relative_GNI_Per_Capita ~ Year + Relative_Urban_Population_Perc,
##     data = wb_cntry.li)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
```

```
## -1.43632 -0.51689 -0.04974  0.35018  2.52372
##
## Coefficients:
##                                  Estimate  Std. Error t value
## (Intercept)                    -106.048551    7.363340 -14.402
## Year                              0.053756    0.003697  14.542
## Relative_Urban_Population_Perc   -0.017675    0.054550  -0.324
##                                          Pr(>|t|)
## (Intercept)                    <0.0000000000000002 ***
## Year                           <0.0000000000000002 ***
## Relative_Urban_Population_Perc                0.746
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6931 on 567 degrees of freedom
## Multiple R-squared:  0.3078, Adjusted R-squared:  0.3054
## F-statistic: 126.1 on 2 and 567 DF,  p-value: < 0.00000000000000022
```

```
# Low middle income group multiple regression model
lmi.both.lm <- lm(Relative_GNI_Per_Capita ~ Year + Relative_Urban_Population_Perc, data = wb_cntry.lmi)
summary(lmi.both.lm)
```

```
##
## Call:
## lm(formula = Relative_GNI_Per_Capita ~ Year + Relative_Urban_Population_Perc,
##     data = wb_cntry.lmi)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.3302 -0.7140  0.0203  0.5544  5.6827
##
## Coefficients:
##                                  Estimate  Std. Error t value
## (Intercept)                    -220.997366   10.183227 -21.702
## Year                              0.111228    0.005133  21.668
## Relative_Urban_Population_Perc    0.464626    0.147745   3.145
##                                         Pr(>|t|)
## (Intercept)                    < 0.0000000000000002 ***
## Year                           < 0.0000000000000002 ***
## Relative_Urban_Population_Perc              0.00173 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.091 on 777 degrees of freedom
## Multiple R-squared:  0.4755, Adjusted R-squared:  0.4741
## F-statistic: 352.1 on 2 and 777 DF,  p-value: < 0.00000000000000022
```

```
# Upper middle income group multiple regression model
umi.both.lm <- lm(Relative_GNI_Per_Capita ~ Year + Relative_Urban_Population_Perc, data = wb_cntry.umi)
summary(umi.both.lm)
```

```
##
## Call:
```

```
## lm(formula = Relative_GNI_Per_Capita ~ Year + Relative_Urban_Population_Perc,
##     data = wb_cntry.umi)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -5.303 -1.667 -0.423  0.515 67.883
##
## Coefficients:
##                                 Estimate Std. Error t value
## (Intercept)                   -404.4291    48.3314  -8.368
## Year                             0.2019     0.0244   8.276
## Relative_Urban_Population_Perc   3.2677     0.9703   3.368
##                                         Pr(>|t|)
## (Intercept)                   < 0.0000000000000002 ***
## Year                           0.00000000000000044 ***
## Relative_Urban_Population_Perc             0.000789 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.817 on 927 degrees of freedom
## Multiple R-squared:  0.1206, Adjusted R-squared:  0.1187
## F-statistic: 63.59 on 2 and 927 DF,  p-value: < 0.00000000000000022
```

```r
# Hi income group multiple regression model
hi.both.lm <- lm(Relative_GNI_Per_Capita ~ Year + Relative_Urban_Population_Perc, data = wb_cntry.hi)
summary(hi.both.lm)
```

```
##
## Call:
## lm(formula = Relative_GNI_Per_Capita ~ Year + Relative_Urban_Population_Perc,
##     data = wb_cntry.hi)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.6439 -0.6080  0.0346  0.4215  6.2113
##
## Coefficients:
##                                  Estimate  Std. Error t value
## (Intercept)                   -257.658487    7.614338 -33.839
## Year                             0.129144    0.003861  33.445
## Relative_Urban_Population_Perc   2.143169    0.432345   4.957
##                                         Pr(>|t|)
## (Intercept)                   < 0.0000000000000002 ***
## Year                          < 0.0000000000000002 ***
## Relative_Urban_Population_Perc          0.000000824 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.086 on 1137 degrees of freedom
## Multiple R-squared:  0.5396, Adjusted R-squared:  0.5388
## F-statistic: 666.3 on 2 and 1137 DF,  p-value: < 0.00000000000000022
```

Looking at the 4 multiple linear regression models by Income Group, we see that the Low Income group has a negative coefficient for the Relative Urban Population Percentage, but all other income groups have positive

coefficients. The adjusted R squared values show that Year and Relative Urban Population Percentage do a solid job of explaining the variability in Relative GNP Per Capita, particularly for the Low Middle Income group and the High Income group.

Conclusion

The World Bank data set, though limited, inspires future research into questions of population migration and income/wealth disparities. We can conclude with strong confidence that there has been a migration from rural communities to urban communities, and that migration has been a worldwide phenomenon.

The linkage between this phenomenon and changes in GNI Per Capita is less clear. We see fairly strong coefficients in our linear models for the low middle income group and the high income group, but the low income group has a negative coefficient and the upper middle income group has a low (but positive) coefficient.

Additional research would require more data and better data. The World Bank data is limited in that it doesn't have a rural/urban categorical variable which would allow for the comparison of rural GNI Per Capita vs. urban GNI Per Capita. There are also inherent difficulties in examining a worldwide data set. Any single country's unsual GNI Per Capita or Urban Population Percentage pattern could skew the overall analysis. There are dozens of factors which effect GNP Per Capita beyond urban population percentage (politcal events, wars, revolutions, trade agreements, commodities/natural resources, etc) and controlling for these factors to focus on urban population percentage would be challenging. It would probably make more sense to study individual countries in depth.