W203 Week 5 Lab 1

Mohammad Jawad Habib

February 7, 2016

Part 2a. Variable Manipulations

10. Load Data and calculate gdp_growth

We'll set the working directory to where we saved the "GDP_World_Bank.csv" file and then load it as a `data.frame`.

Before running the script change the directory to where you have stored the 'GDP_World_Bank.csv'.

```
dir.string <- "C:/Users/SP4/Documents/Exploring and Analyzing Data/W203 Async/W203 Week 5/Lab 1"

# dir.string <- "C:/Users/jhabib/documents/r/W203/W203 Week 5/Lab 1"
setwd(dir.string)

gdp.worldbank <- read.csv("GDP_World_Bank.csv", header = TRUE)
```

`gdp_growth` is the difference between `gdp2012` and `gdp2011`.

```
gdp.worldbank$gdp_growth <- gdp.worldbank$gdp2012 - gdp.worldbank$gdp2011

# gdp.worldbank$gdp_growth
```

We then calculate the mean of `gdp_growth`; we will ingore `NA` values in the mean calculation.

```
gdp.growth.mean <- mean(gdp.worldbank$gdp_growth, na.rm = TRUE)
gdp.growth.mean
```

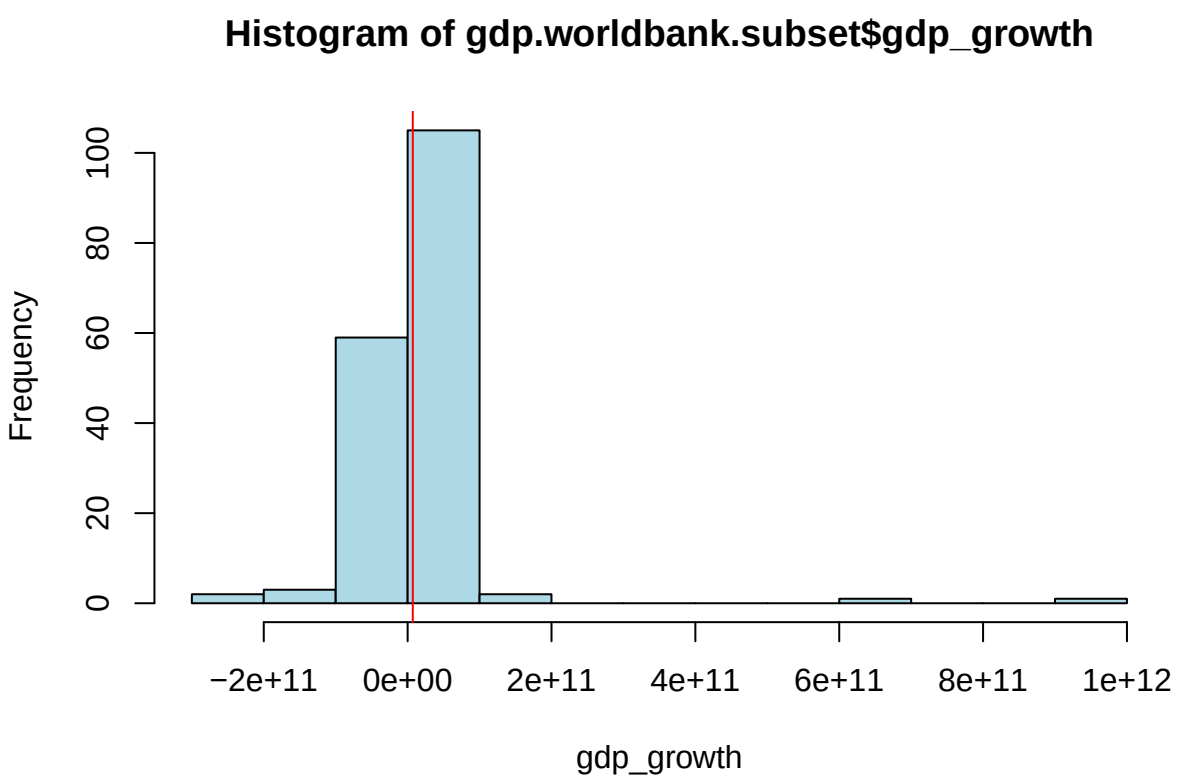
```
## [1] 7172376796
```

The calculation above reveals the `gdp.growth.mean` to be `return(gdp.growth.mean)`.

11. Create a histogram of gdp_growth

We can create a basic histogram using the base R. We will ignore the `NA` values in `gdp_growth`.

```
gdp.worldbank.subset <- gdp.worldbank[!is.na(gdp.worldbank$gdp_growth),c(1, 3, 4, 5)]

hist(gdp.worldbank.subset$gdp_growth,
     freq = TRUE,
     breaks = diff(range(gdp.worldbank.subset$gdp_growth))/1e11,
     xlab = "gdp_growth",
     ylab = "Frequency",
     col = "lightblue")
par(new = TRUE)
abline(v = mean(gdp.worldbank.subset$gdp_growth), col="red")
```

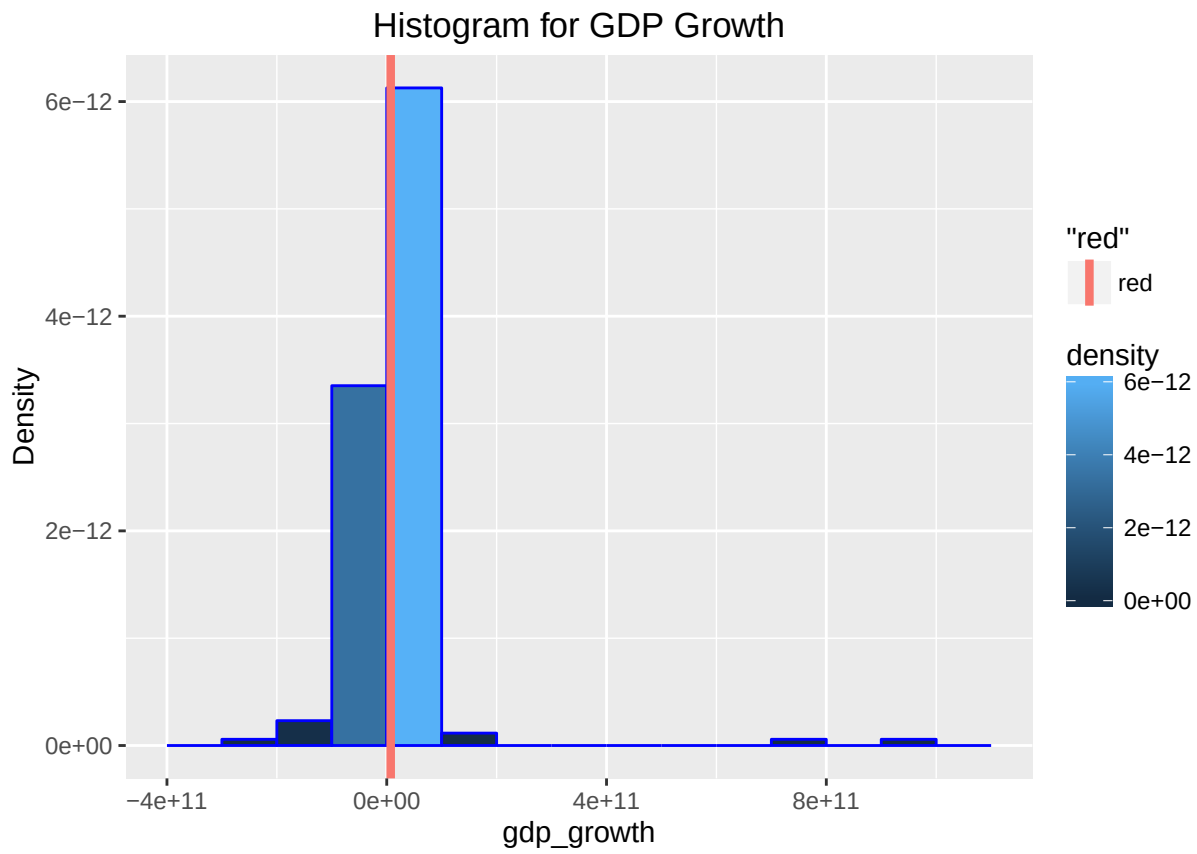


We will use the ggplot2 library to create a nicer histogram than the base R package. This time we will plot the density on y-axis.

```
require(ggplot2)
```

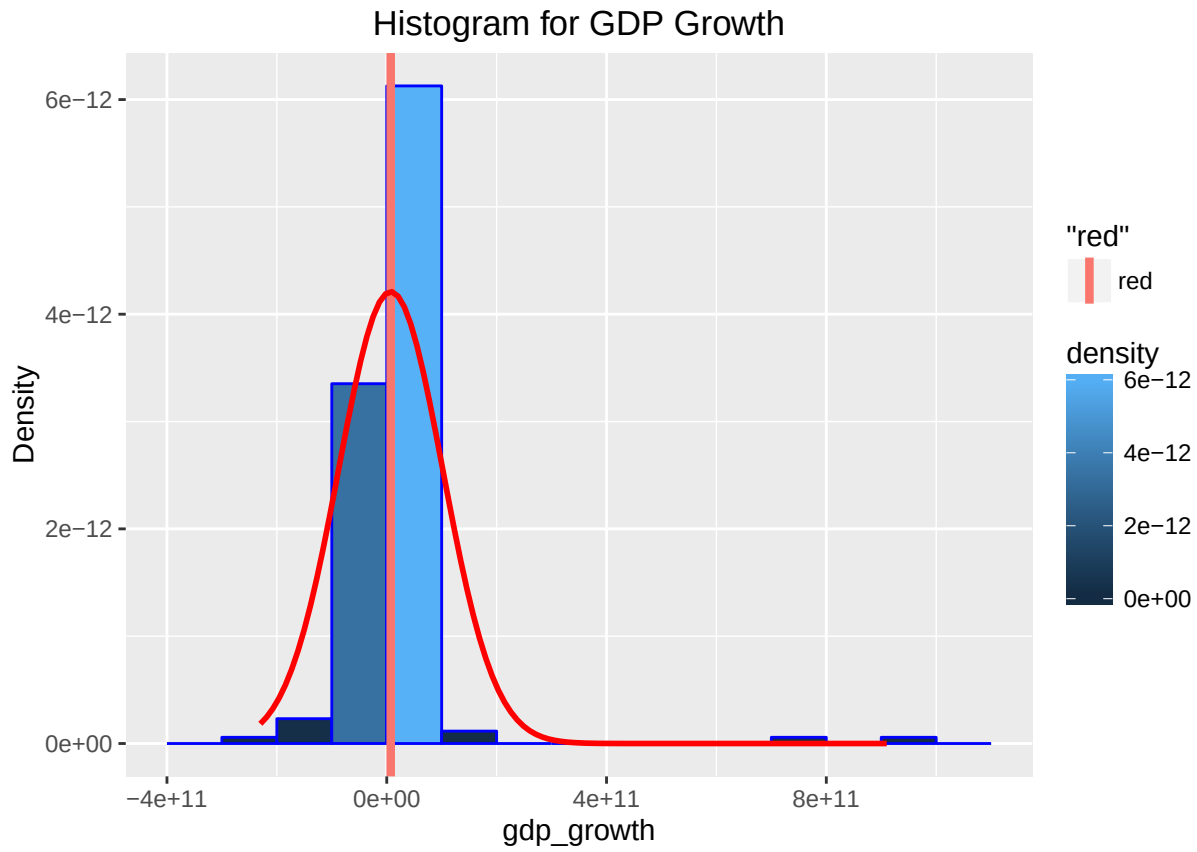
```
## Loading required package: ggplot2
```

```
gdp.growth.hist <- ggplot(gdp.worldbank.subset, aes(gdp_growth)) +
  geom_histogram(na.rm = TRUE,
                 bins = diff(range(gdp.worldbank.subset$gdp_growth))/1e11,
                 aes(fill = ..density.., y = ..density..),
                 col = "blue") +
  geom_vline(aes(xintercept=mean(gdp.worldbank.subset$gdp_growth),
                 colour = "red"),
             size = 1.5) +
  labs(title="Histogram for GDP Growth") +
  labs(x="gdp_growth", y="Density")

gdp.growth.hist
```



Let's add a normal curve to our graph to visually see how it fits.

```
gdp.growth.hist +
  stat_function(fun=dnorm,
                colour = "red",
                size = 1,
                args=list(mean=mean(gdp.worldbank.subset$gdp_growth),
                          sd=sd(gdp.worldbank.subset$gdp_growth)))
```



This looks like a positively skewed, leptokurtic distribution. Normal distribution does not seem like a good fit.

Let's create a table with some basic parameters that describe this distribution.

```
gdp.growth.summary.parameter <- c("Mean", "Median", "Minimum", "Maximum", "Standard Deviation", "Skewnes
gdp.growth.summary.values <- c(gdp.growth.mean,
                               median(gdp.worldbank.subset$gdp_growth),
                               min(gdp.worldbank.subset$gdp_growth),
                               max(gdp.worldbank.subset$gdp_growth),
                               sd(gdp.worldbank.subset$gdp_growth),
                               e1071::skewness(gdp.worldbank.subset$gdp_growth),
                               e1071::kurtosis(gdp.worldbank.subset$gdp_growth))
gdp.growth.summary <- data.frame(gdp.growth.summary.parameter, gdp.growth.summary.values)
gdp.growth.summary
```

```
##   gdp.growth.summary.parameter gdp.growth.summary.values
## 1                         Mean              7.172377e+09
## 2                       Median              2.017000e+08
## 3                      Minimum             -2.300000e+11
## 4                      Maximum              9.100000e+11
## 5           Standard Deviation              9.476377e+10
## 6                     Skewness              7.027270e+00
## 7                     Kurtosis              6.171466e+01
```

We can use the Jarque-Bera Test from `tseries` to determine goodness of fit.

```
require(tseries)
```

```
## Loading required package: tseries
```

```
jarque.bera.test(gdp.worldbank.subset$gdp_growth)
```

```
##
##  Jarque Bera Test
##
## data:  gdp.worldbank.subset$gdp_growth
## X-squared = 29579, df = 2, p-value < 2.2e-16
```

It seems that the p-value of 2.2e-16 means that our null-hypothesis of normal distribution is very unlikely.

12. Create a high_growth boolean variable

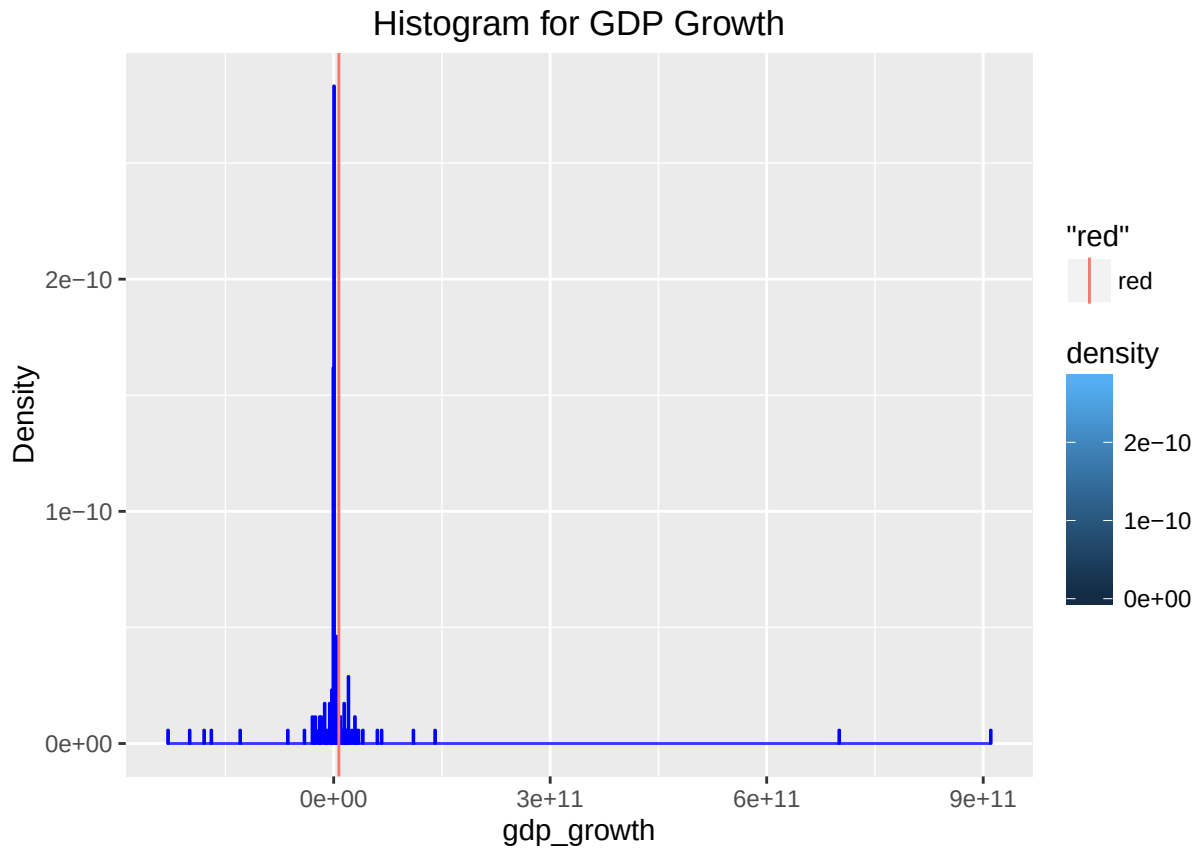
We will create a high_growth variable that returns 1 if gdp_growth is greater than mean for a country and 0 if not.

```
gdp.worldbank.subset$high_growth <- ifelse(gdp.worldbank.subset$gdp_growth > gdp.growth.mean,
                                           1,
                                           0)
high.growth.countries <- sum(gdp.worldbank.subset$high_growth)

low.growth.countries <- length(gdp.worldbank.subset$high_growth) - sum(gdp.worldbank.subset$high_growth
```

There are 31 countries with high gdp growth by our measure and 142 countries without. These results exclude NA values. This makes sense given the distribution of gdp_growth because of the positive skewness.

Consider the histogram of gdp_growth with resized bins to better show the values on either side of the mean.

```
gdp.growth.hist1 <- ggplot(gdp.worldbank.subset, aes(gdp_growth)) +
  geom_histogram(na.rm = TRUE,
                 bins = diff(range(gdp.worldbank.subset$gdp_growth))/1e9,
                 aes(fill = ..density.., y = ..density..),
                 col = "blue") +
  geom_vline(aes(xintercept=mean(gdp.worldbank.subset$gdp_growth),
                 colour = "red"),
             size = 0.5) +
  labs(title="Histogram for GDP Growth") +
  labs(x="gdp_growth", y="Density")

gdp.growth.hist1
```



Part 2b. Data Import

I would like to explore whether changes in the Rural Population of a country are correlated to the `gdp_growth` in the same time period. For this purpose I will use 'SP.RUR.TOTL.ZS' indicator from http://data.worldbank.org/indicator/SP.RUR.TOTL.ZS. This indicator shows rural population as a percent of total population by year. We will import data from years 2010, 2011 and 2012 just because our gdp data frame has the same years.

To make things simpler, we will use the `WDI` package that implements an API to obtain data from data.worldbank.org.

```
require(WDI)
```

```
## Loading required package: WDI
```

```
## Loading required package: RJSONIO
```

```
rural.pop.worldbank <- WDI(country = "all", indicator = "SP.RUR.TOTL.ZS",
                           start = 2010, end = 2012, extra = FALSE, cache = NULL)

head(rural.pop.worldbank)
```

```
##   iso2c                country SP.RUR.TOTL.ZS year
```

```
## 1    1A             Arab World       43.15100 2012
## 2    1A             Arab World       43.43492 2011
## 3    1A             Arab World       43.73413 2010
## 4    S3 Caribbean small states       58.06299 2012
## 5    S3 Caribbean small states       58.12272 2011
## 6    S3 Caribbean small states       58.17559 2010
```

We see that `rural.pop.worldbank` is in long format. We will convert it to wide format.

```
require(reshape2)
```

```
## Loading required package: reshape2
```

```
# We don't need iso2c so let's set it to NULL
rural.pop.worldbank$iso2c <- NULL

# Rename SP.RUR.TOTL.ZS to a more readable rural_pop_percent
names(rural.pop.worldbank)[names(rural.pop.worldbank) == "SP.RUR.TOTL.ZS"] <- "rural_pop_percent"

# Convert the data to wide format
rural.pop.wide <- dcast(melt(rural.pop.worldbank,
                             id.vars = c("country", "year")), country+variable~variable+year)
```

Let's merge our gdp and rural population data frames together by country. We will use our `gdp.worldbank.subset` data frame because we removed the `NA` values of gdp from it.

```
country.data.merged <- merge(gdp.worldbank.subset,
                             rural.pop.wide,
                             by.x = "Country",
                             by.y = "country",
                             incomparables = NULL)

head(country.data.merged)
```

```
##               Country      gdp2011      gdp2012  gdp_growth high_growth
## 1             Albania  12959563902  13119013351   159449449           0
## 2             Algeria 199000000000 208000000000  9000000000           1
## 3              Angola 104000000000 114000000000 10000000000           1
## 4 Antigua and Barbuda   1124586886   1176348888    51762002           0
## 5           Argentina 446000000000 475000000000 29000000000           1
## 6             Armenia  10138077996   9910387657  -227690339           0
##            variable rural_pop_percent_2010 rural_pop_percent_2011
## 1 rural_pop_percent                  47.837                  46.753
## 2 rural_pop_percent                  32.474                  31.791
## 3 rural_pop_percent                  59.903                  59.100
## 4 rural_pop_percent                  73.761                  74.333
## 5 rural_pop_percent                   9.034                   8.867
## 6 rural_pop_percent                  36.420                  36.629
##   rural_pop_percent_2012
## 1                 45.670
## 2                 31.130
## 3                 58.301
```

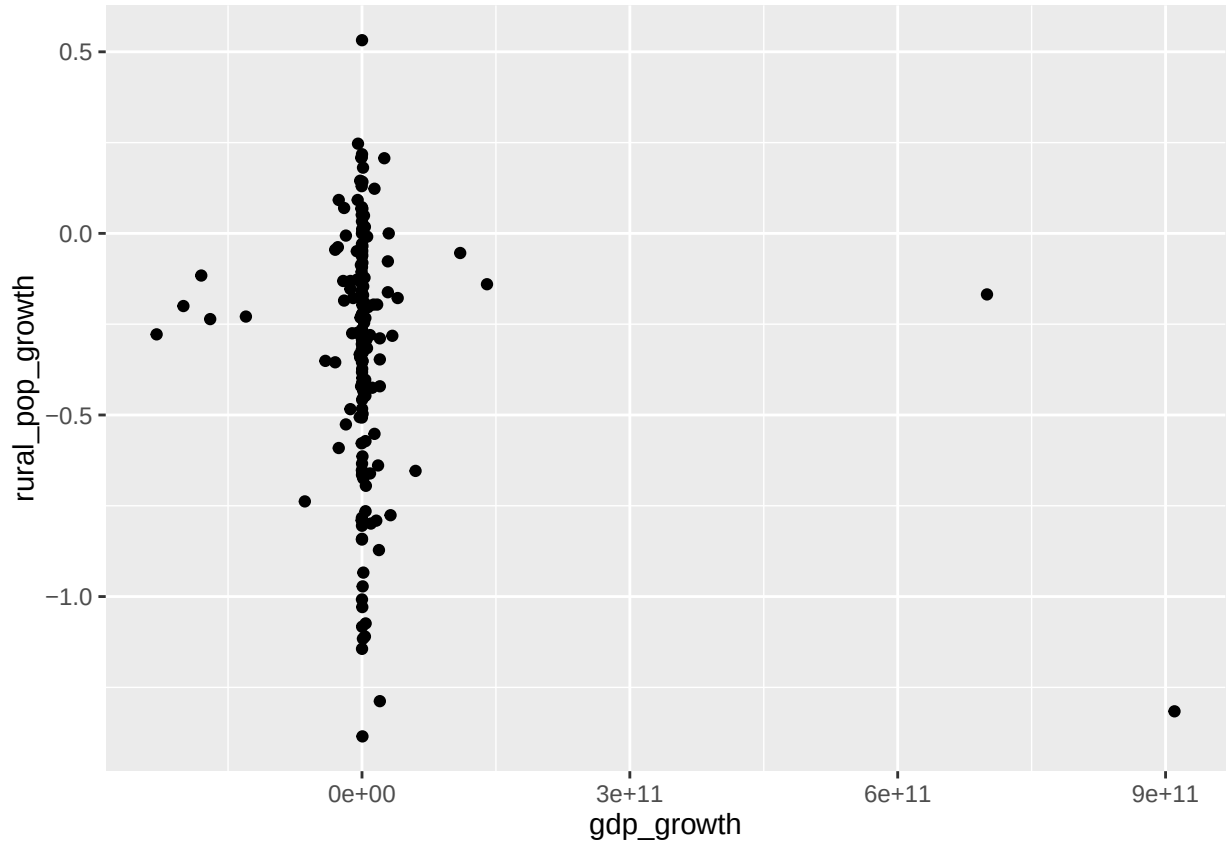
```
## 4                74.865
## 5                 8.705
## 6                36.839
```

Next we'll add a new variable to the merged data frame to get the change in percent population from 2011 to 2012.

```
country.data.merged$rural_pop_growth <- country.data.merged$rural_pop_percent_2012 -
  country.data.merged$rural_pop_percent_2011
```

Let's create a scatter plot of `rural_pop_growth` on y-axis and `gdp_growth` on x-axis to see if there are any trends.

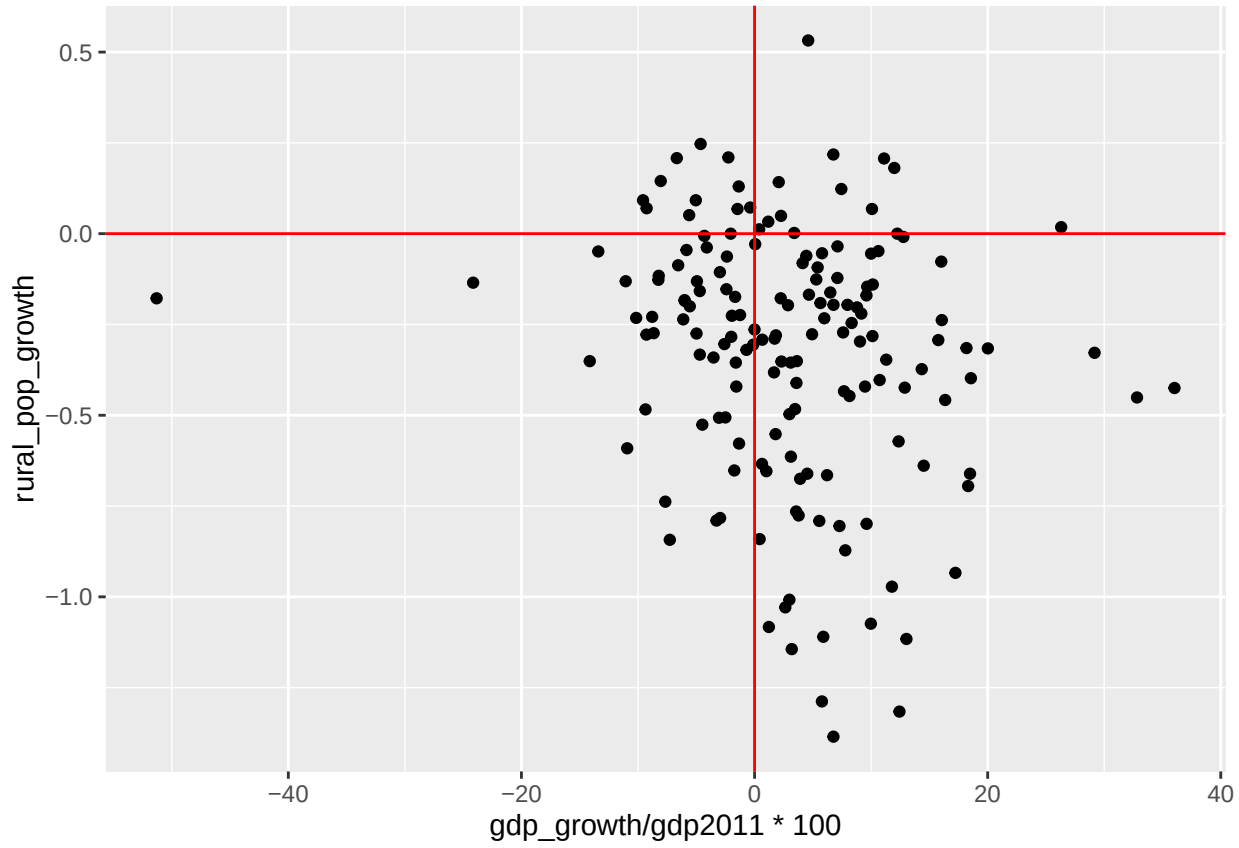
```
rur.growth.by.gdp.growth <- ggplot(country.data.merged,
                                   aes(x = gdp_growth, y = rural_pop_growth)) +
  geom_point(na.rm = TRUE)

rur.growth.by.gdp.growth
```



It's interesting to note that there does not appear to be a correlation between changes in rural populationand changes in gdp from 2011 to 2012.

May be we can read the graph better if x-axis showed `gdp_growth` as a percent of `gdp2011`.

```r
rur.growth.by.gdp.growth1 <- ggplot(country.data.merged,
                                    aes(x = gdp_growth/gdp2011*100, y = rural_pop_growth)) +
  geom_point(na.rm = TRUE) +
  geom_vline(xintercept = 0, color = "red") +
  geom_hline(yintercept = 0, color = "red")

rur.growth.by.gdp.growth1
```



This graph is readable but it still does not show any obvious correlation between changes in rural population and `gdp_growth`. It seems though that most countries saw a decrease in rural population percentage regardless of the positive or negative changes in `gdp_growth` from 2011 to 2012.