Exploring and Analyzing Data – Lab Assignment 1

Christopher J. Llop

Submitted to:

Paul Laskowski and Guang Yang

Date: October 1, 2014

**Lab 1**

**Summary of Assignment and Approach**

Paul Laskowski and Guang Yang (hereafter “the clients”) have asked the author to assist in answering questions relating to the exploration and analysis of data. They have provided the author with a data file including non-confidential GDP data points and have asked that certain analyses be performed on this data, including:

1. Calculating the mean growth in GDP from 2011 to 2012;
2. Building a histogram to study the distribution of the mean calculated in (1); and
3. Examining how many countries have GDP above and below said mean to further explore implications on the distribution of the mean.

Furthermore, the clients have requested that the author find a new dataset to combine with the data provided to opine on an item that might be of interest. For the purposes of this study, a dataset from the World Bank investigating population by country was selected.[[1]](#footnote-1) The author solved data issues to allow this dataset to be integrated with the client’s GDP data and created several plots to show the relationship between population and GDP growth. This analysis involved splitting the countries into “positive” and “negative” GDP growth so that a logarithmic scale could be applied to both countries with GDP that grow and those with GDP that shrink.

In Section I of this report, the author provides assistance in answering the questions provided by the clients (the, “Multiple Choice” section). In Section II, the author addresses the specific questions asked relating to GDP growth and shows histograms of distributions of interest. In Section III the author discusses merging in another dataset, presents scatterplots, and opines on his findings.

# Multiple Choice

The clients have requested that their questions remain confidential and are not to be disclosed in this report. As such, the author has adopted the numbering scheme as presented by the client to answer these questions:

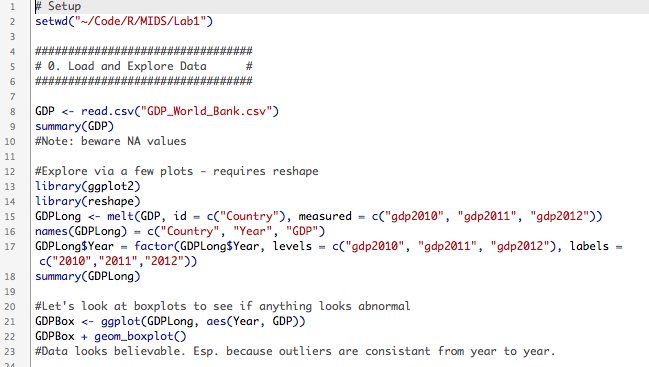
1. e) ordinal
2. a) measuring a potentially interval or ordinal variable as a binary variable
3. d) standard deviations can be directly compared to the individual deviation of one data point away from the mean
4. b) stratified random sampling
5. c) a and b are equally likely[[2]](#footnote-2)
6. b) for large samples, it suggests that the normal distribution is a good model for the distribution of the mean and other statistics
7. f) none of the above
8. d) the distribution of your age variable is platykurtic

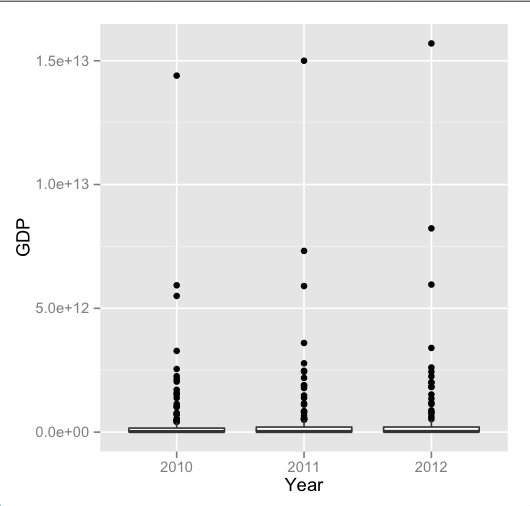
# Data Analysis and Short Answer

In this section, the author provides assessment of GDP data provided by the clients.

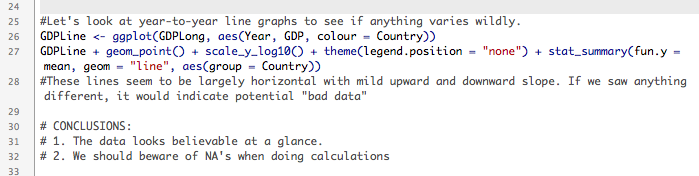
## Sanity Checks of Provided Data

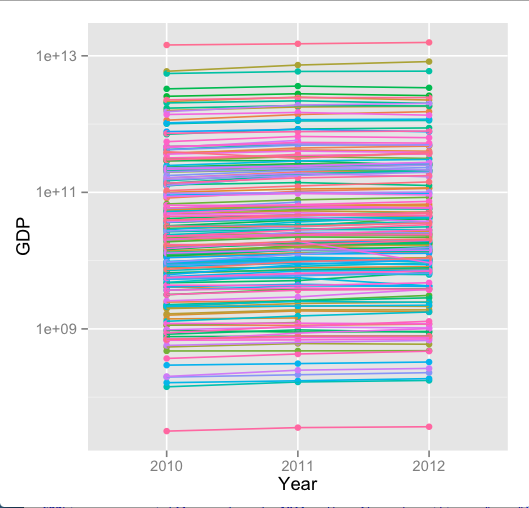
Prior to answering the questions posed by the clients, it was important to sanity check the data. The following R code was used to check for outliers and to investigate any suspicious data points:

****



These box plots show the presence of many outliers. The author later sorted the data and looked to see what the outliers were. This process showed that the outliers of the United States and China both have higher GDPs than the rest of the world, which fits in with general knowledge. As one final check, the author plotted a line graph of the GDP of each country by year. This line graph shows that no country’s GDP fluctuates wildly in a way that might warrant further review. Each countries line is relatively stable with limited growth or declines over the period studied.

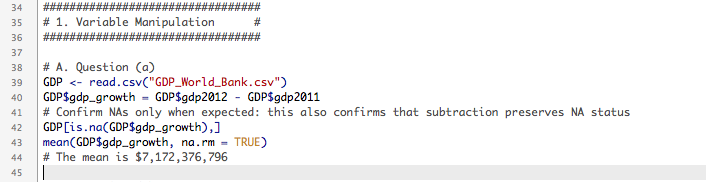
****



This investigation showed no unexpected abnormalities in the data. Furthermore, observations that might appear to be drastic outliers have been explained as representing countries that genuinely have a much larger than average GDP. The fact that these observations are consistently larger from year to year further indicates that the data seems a sound source for analysis.

## Calculating the Mean

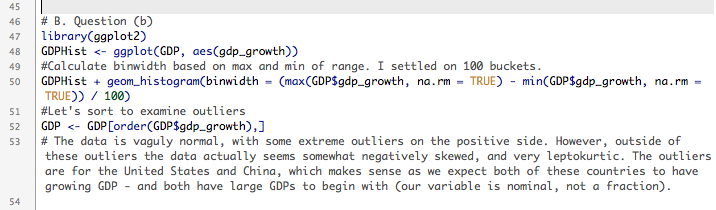
The following code was used to create a new variable, gdp\_growth, which measures the increase in GDP from 2011 to 2012, and to examine its mean:

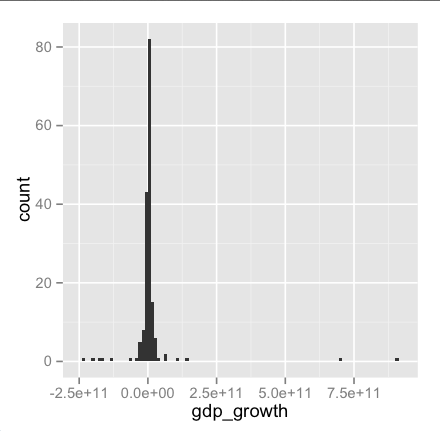
****

The mean of the new variable is reported as $7,172,376,796.

## Examination of a Histogram

The following code was used to create a histogram used to examine the distribution of the variable gdp\_growth:

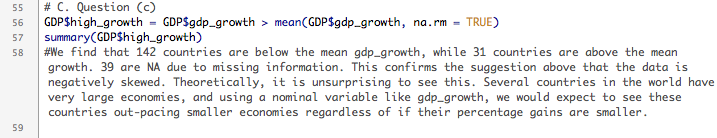
****



The distribution isvaguely normal, with some extreme outliers on the positive side. However, outside of these outliers the data actually seems somewhat negatively skewed, and very leptokurtic. The outliers are for the US and China, as discussed above, which makes sense.

## Above vs. Below Average GDP Growth

The following code was used to create a variable, high\_growth, which examined if a specific country’s GDP growth was higher than the mean:

****

This code showed that 142 (67%) countries experienced GDP growth above the mean, while 31 (17%) countries experience growth below the mean. 39 (18%) countries that were missing data in either year were excluded from the plot. This confirms our findings in Section I.C. above, pointing towards a negative skew to the data.

# Integration of Additional Data

In this section, the author examines the relationship between total population and GDP growth between 2011 and 2012.

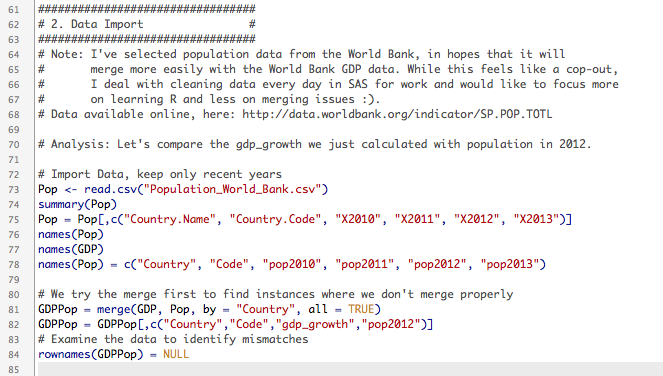
## Selection of Data

In order to expedite the process of generating this report, data sets from the World Bank were reviewed under the assumption that data from the same underlying source as the client’s data would require fewer adjustments to merge successful. Ultimately, the author decided to select a dataset showing world population (hereafter, “the population data”) from the years XXXX – YYYY for this study. After examining the data, the author decided that the year 2012 would provide a good comparison point to the variable gdp\_growth, as the gdp\_growth variable shows the change in GDP from the year 2011 to 2012.

The author undertook an inductive approach to his investigation and did not hypothesize on the results of comparing the two variables prior to generating plots.

## Issues in Merging Selected Data

The author used the following code to import the population data and perform a “first merge” to investigate how successfully the datasets could be matched:

****

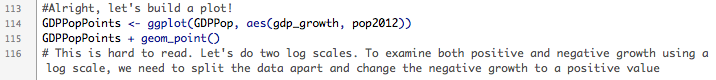
Despite the best efforts of the author to select well-merging data, several countries did not merge together easily. These countries were examined manually and renamed with the following code:

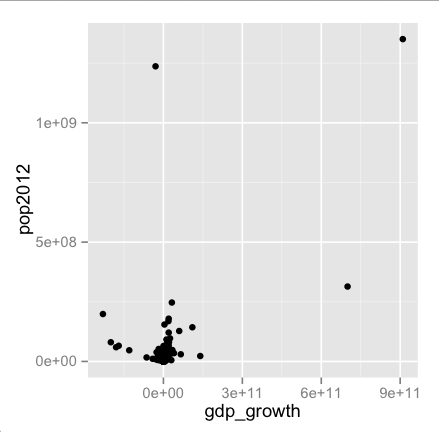
****

After performing this second merge, the author confirmed that any remaining mismatches stemmed from aggregate reporting in the population data. That is, the population data provided metrics for entire continents and regions, in addition to country-specific metrics. These additional data entries were dropped from the analysis.

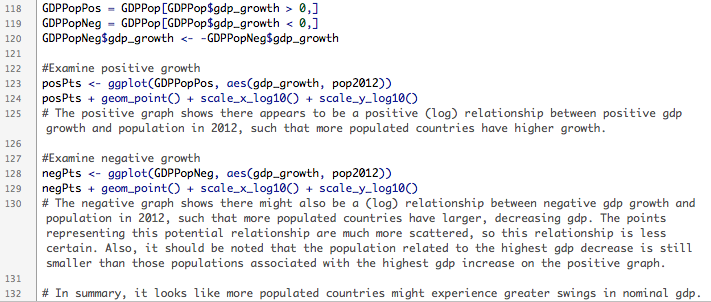
## Scatterplots of Results

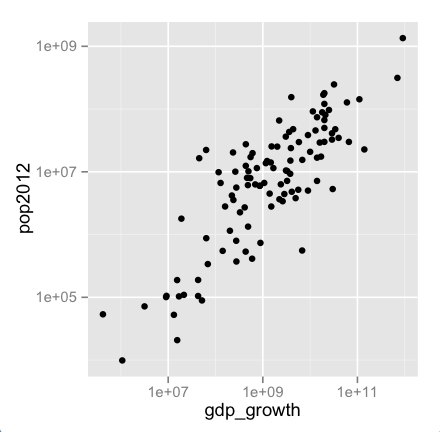
After the merge was complete, the author used the following code to create a “quick” scatterplot of the results:

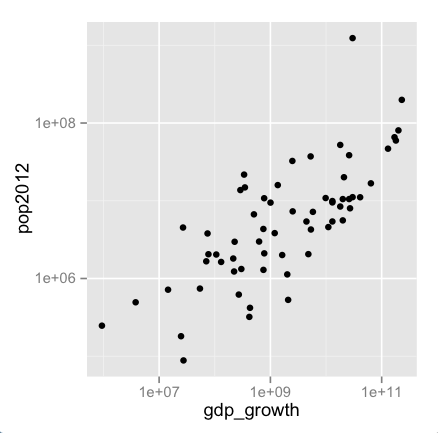
****

****

As can be seen from the above figure, the data appears tightly clustered using a linear scale. To fix this problem, the author decided to attempt a logarithmic scale. However, because a logarithmic scale ignores negative data, the author decided to split the analysis up into two parts based on the direction of the gdp\_growth variable. The following code splits the gdp\_growth variable and then displays scatterplots for each of sub-analyses:

****

****

****

## Conclusions

The positive graph (first of the two) shows that there appears to be a positive (log) relationship between positive GDP growth and population in 2012, such that more populated countries have higher growth.

However, the negative graph also shows there might be a (log) relationship between negative GDP growth and population in 2012, such that more populated countries have larger, decreasing GDP. The points representing this relationship are more scattered than the positive growth analysis, making this relationship less certain. It also should be noted that the population related to the highest GDP decrease is still much smaller than those populations associated with the highest GDP increase in the positive growth analysis.

In summary, it looks like more populated countries might be more likely to experience greater swings in nominal GDP.

1. See “sp.pop.totl\_Indicator\_en\_csv\_v2.zip”, available online at: http://data.worldbank.org/indicator/SP.POP.TOTL. [↑](#footnote-ref-1)
2. ; ; [↑](#footnote-ref-2)