## Part 1 : Variable Manipulation

a. We created a variable internet growth which would capture an increase in percentage of users and calculated its value for all countries. We captured the value in the Countries dataframe by adding a column “percincr” to store the percentage increase for each country.

We also calculated basic statistics for the percincr variable namely mean , median and standard deviation.

**mean(percincr) = 21.6828**
**median(perc_incr) = 11.32075**
**stddev(perc_incr) = 33.23521**

b. We are supposed to find the country with the highest percentage growth. We used aggregate function max() to calculate the max value of the percentage. Through finding the index of the row which contained this value , we were able to identify **“Myanmar”** as the country with the highest percentage growth , a result that quite frankly surprised us. Upon further investigation, we found some interesting observations.

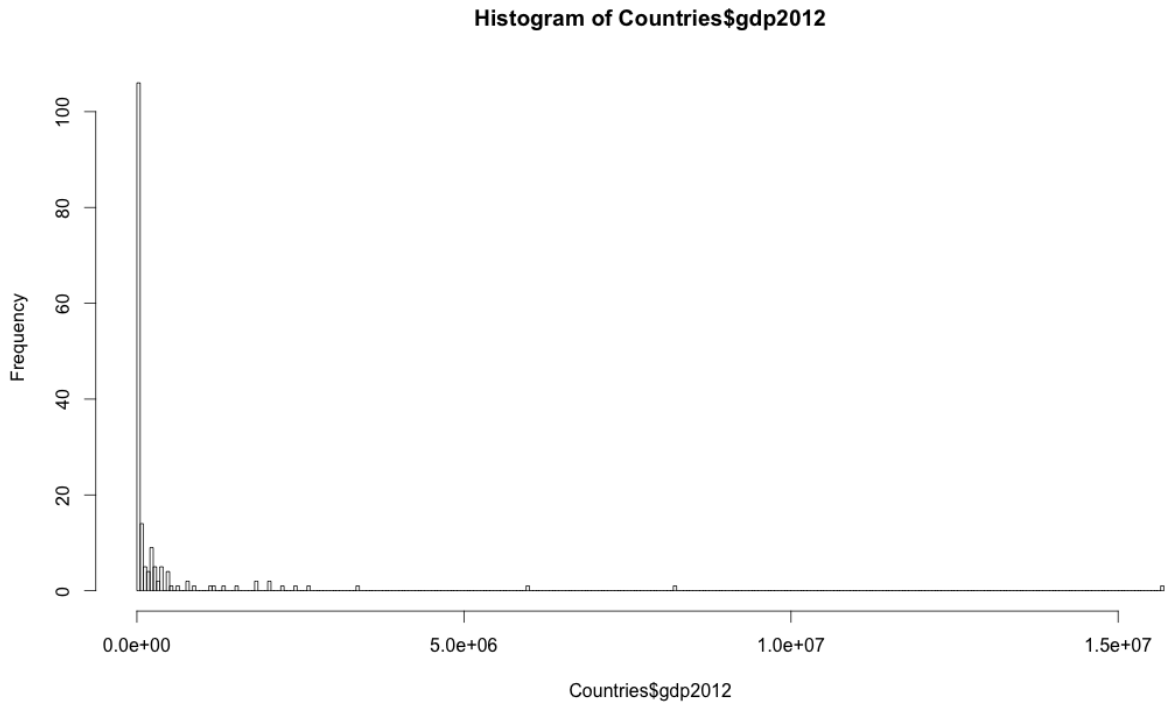
The % users was very small to begin with, the base value against which the percentage increase is captured . A small bump in number of users made myanmar the highest value in terms of percentage increase.

| Country | % in 2011 | % in 2012 | perc increase |
|---|---|---|---|
| Myanmar | 0.25 | .98 | 292.00 |

c. We made a boolean index to capture whether each of the perc increase value is above or below the mean.

We calculated the mean as **394106.8**
We found **149** countries to be below the mean value
we found **25** countries to be above the mean value.

We also calculated the median as **25243.02**. Median << Mean. The distribution is a right skewed distribution. We also used histogram to plot the gdp’s , we forced nclass parameter = 500 to see more number of bins to be able to observe the data. The histogram confirmed the nature of the distribution.




d. We calculated the mean percentage increase of internet usage in countries with higher than average gdp and countries with lower than average gdp.

Mean ( percentage increase in internet usage) = **9.537224** for countries with higher than average gdp.

Mean ( percentage increase in internet usage) = **22.14695** for countries with lower than average gdp.

This helps us understand an important phenomenon. The growth of percentage of internet users in countries with higher gdp is lower because of saturation, there are a lot of users already using the internet, hence the growth which may be significant does not show when calculated as a percentage increase. The lower gdp nations , even though have a lesser percentage of population using the internet are showing promising growth . One such example we did see in Myanmar - 0.28 % to 0.95 % computes to a percentage growth of 292 % because of the base percentage value being very low.

## Part 2 : Data Import

We decided to choose the world bank data Source -

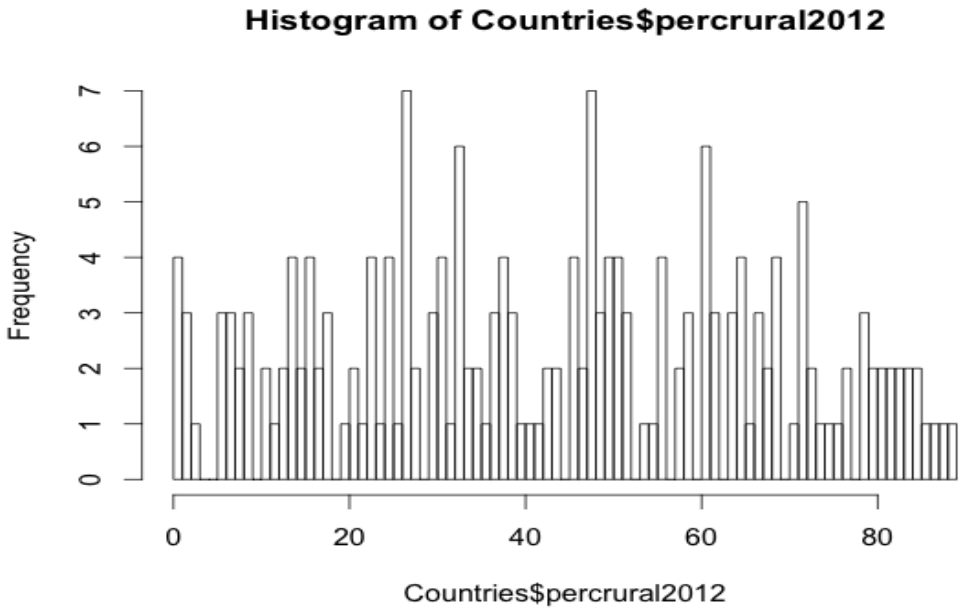
http://data.worldbank.org/indicator/SP.RUR.TOTL.ZS

The dataset contains country wise percentage rural population from years 1953 to 2011. We chose to take 2011 data and merge it with our existing data set. We did the merge comparing country names. We chose this dataset as it would be interesting to see the relationship of percentage of rural population for the country to gdp and internet usage and other factors in our existing dataset.

We had to modify the datafile after inspecting it in R. It was a csv file with the first two lines indicating the data source. The csv format to be read correctly in R required the column names to be on the header.

## Part 3 : Graphs

a) We plotted the histogram for percentage rural population data.

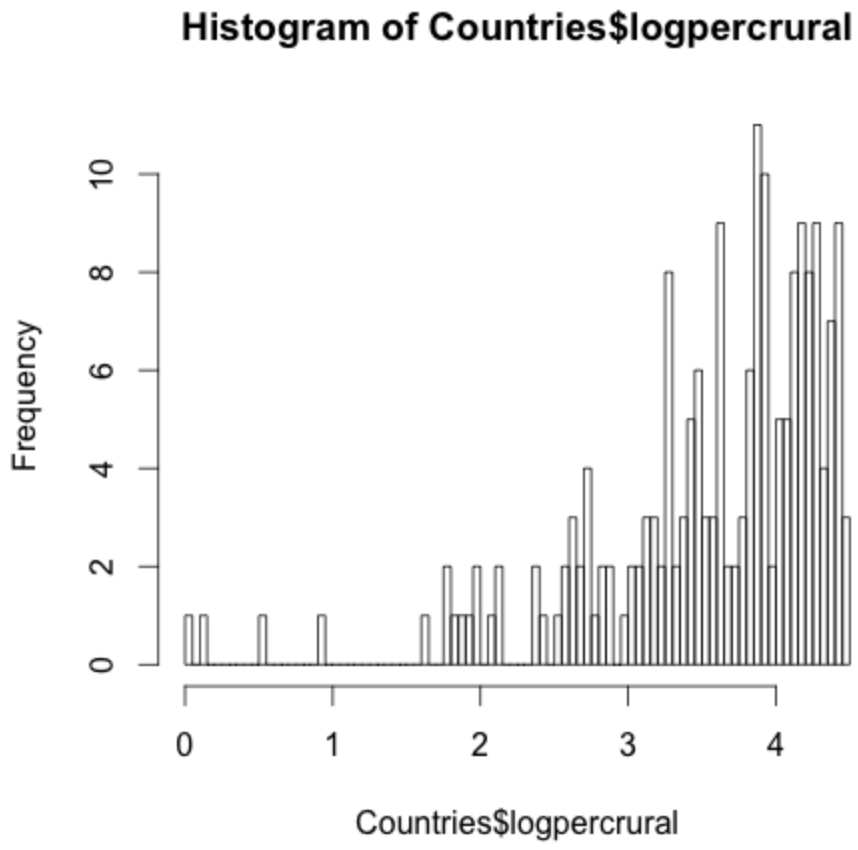


The distribution is not normal , its multimodal and not distributed uniformly around the mean. We calculated summary statistics for the data. # the summary shows mean, median, first and third quartile.

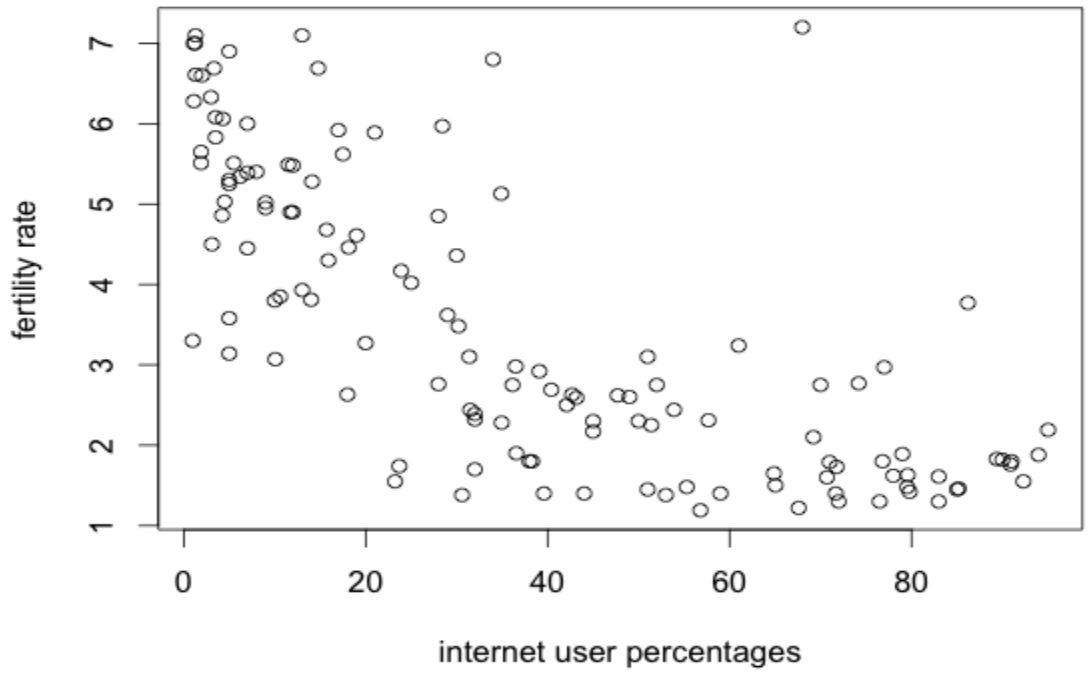
```
# Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
# 0.00   24.27   43.41   42.88   63.03   88.79       4
```

We calculated the SD = **24.13106.**
Mean to Mean + sd = 34 % of the data. The mean + df value is close to the third quartile which is also a good indicator of the data distribution being not normal.

Taking logarithm values makes the graph left skewed but not normal.

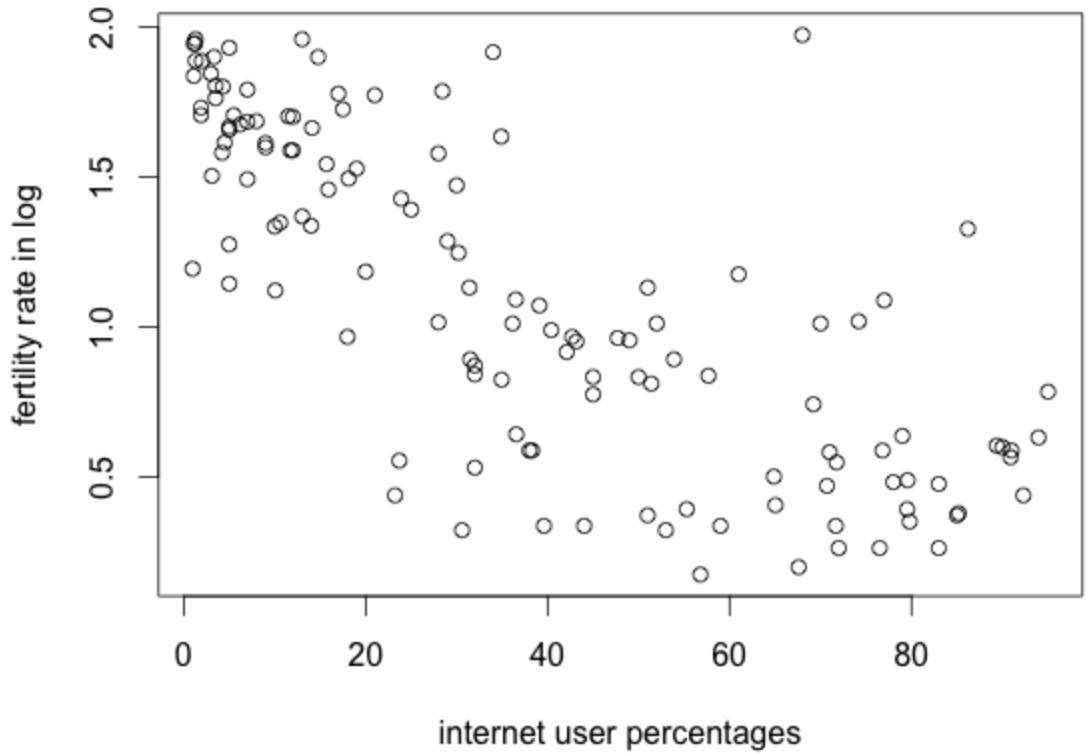


b) We plotted the scatter plot between fertility rate and percentage of internet users.



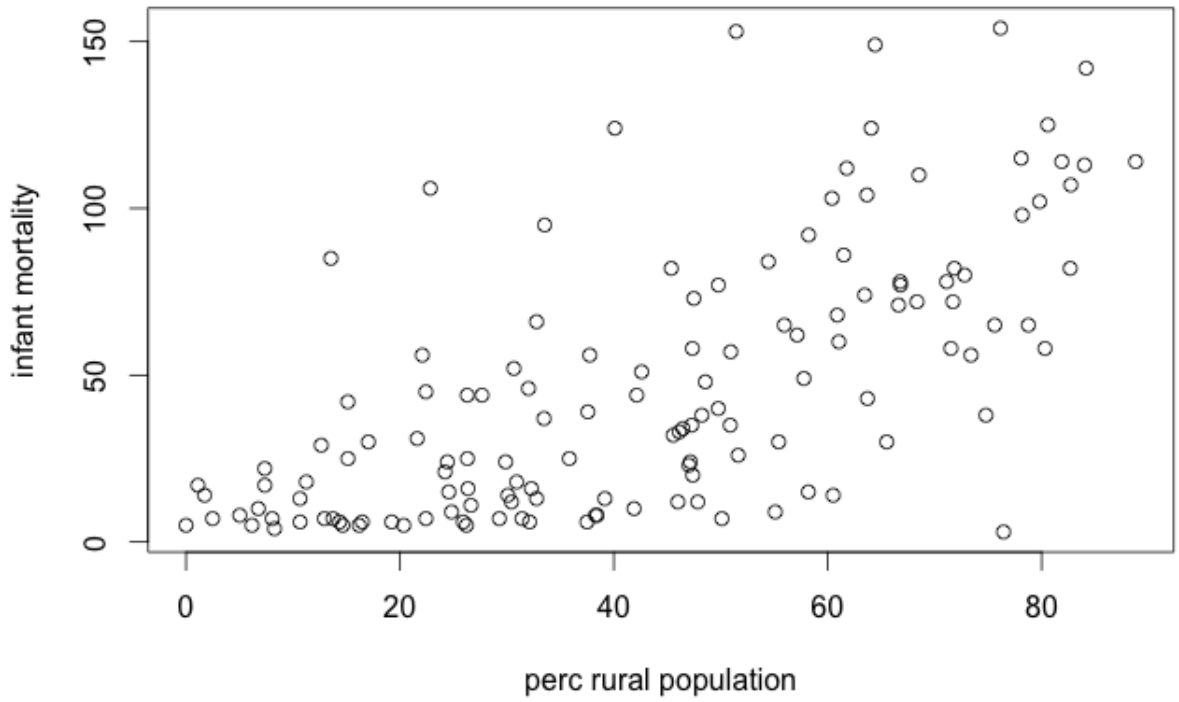

The relation although not strictly linear does indicate **negative correlation** between the two variables. It can be observed that nations with lower internet usage have higher fertility rate.

We calculated the correlation to be **- 0.69.**




the logarithm does help improve correlation to be **- 0.77.** The transformation function is making the relationship more linear as indicated by the correlation coefficient.

c) We created a scatter plot for percentage rural population and infant mortality. Please look at the graph below. The graph shows positive correlation between the two factors. We calculated the correlation coefficient as **+0.67.** We also ran summary statistics on the two datasets.




The results of the summary were -

| percrural | mortality |
|---|---|
| Min. : 0.00 | Min. : 3.00 |
| 1st Qu.:24.27 | 1st Qu.: 12.75 |
| Median :43.41 | Median : 35.00 |
| Mean :42.88 | Mean : 46.28 |
| 3rd Qu.:63.03 | 3rd Qu.: 72.25 |
| Max. :88.79 | Max. :154.00 |
| NA's :4 | NA's :62 |