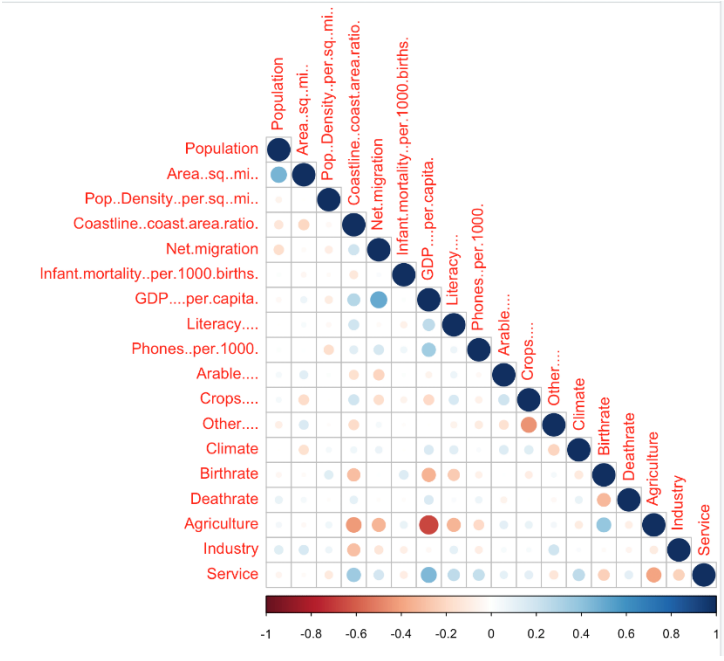
From 19th Century, because of the development of transportation and the establishment of modern government, an increasing number of people have been to another country for studying and working. Although migration can also be defined as internal migration, global migration, which means people from their mother countries to another country with long distance permanently or temporarily, is the major research field in this report. Without specific and strong reasons, people won't leave their familiar place and go to others strange cultural environment. The reason can be various, like natural disaster and war in original countries and better living environment and more well-paid job opportunities in the immigrant country, our research focused on why people want to leave their original countries.

From our experiences, the high economical performance will be a key reason why people want to go a country instead of another one. Because it means higher earnings, lower unemployment rate and bigger market. A common and excellent standard is GDP per capita. Thus, we want to develop a model to test their statistical relationship between Net migration and GDP per capita. Also, there will be a good improvement by testing more highly related variables and building a multiple linear model.

The raw data is from Kaggle and US government. 227 countries' population, area, migration, GDP per capita, etc. are showed. There is only one missing value with west Sahara, which is a country in dispute.

We continue to develop our multiple linear regression model. At first, correlation plot is a basic reference for choosing additional variables. Because it will help us to have a basic understanding for the correlation of numeric variables.



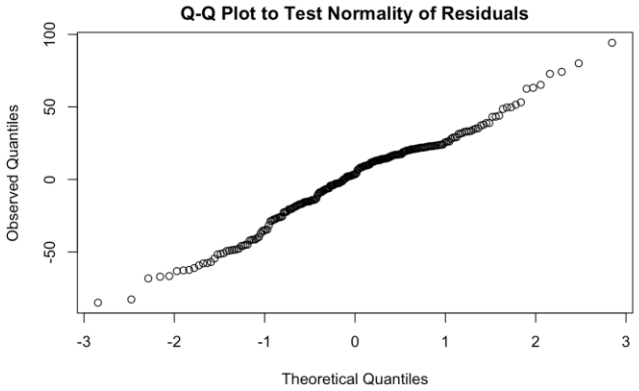
From the net migration row of correlation plot, there are two medium-related variables: GDP per capita and agriculture. The agriculture means the ratio that agriculture account for the three industries

including agriculture, industry and services. But agriculture is also high-related to GDP per capita, which violates the assumption of multicollinearity. Thus, agriculture is not acceptable. Finally, after tens of attempts, we pick the best composition of independent variables including GDP per capita, coastline and region.

```
> coun_model <- lm(net_mig~region+gdp+coastline,data=coun_data)
> summary(coun_model)

Call:
lm(formula = net_mig ~ region + gdp + coastline, data = coun_data)

Residuals:
    Min      1Q  Median      3Q     Max
-84.972 -19.108   3.542  21.093  94.224

Coefficients:
                                      Estimate Std. Error t value Pr(>|t|)
(Intercept)                          6.417e+01  7.148e+00   8.977  < 2e-16 ***
regionBALTICS                       -3.486e+01  1.979e+01  -1.762  0.07958 .
regionC.W. OF IND. STATES           -5.878e+00  1.154e+01  -0.509  0.61108
regionEASTERN EUROPE                -5.867e+00  1.147e+01  -0.511  0.60961
regionLATIN AMER. & CARIB           -2.066e+01  7.918e+00  -2.609  0.00971 **
regionNEAR EAST                      2.017e+01  1.023e+01   1.972  0.04990 *
regionNORTHERN AFRICA               -3.695e+01  1.591e+01  -2.323  0.02113 *
regionNORTHERN AMERICA               9.349e+00  1.690e+01   0.553  0.58071
regionOCEANIA                        3.471e+00  9.725e+00   0.357  0.72152
regionSUB-SAHARAN AFRICA             1.986e+00  7.957e+00   0.250  0.80314
regionWESTERN EUROPE                 2.033e+01  1.079e+01   1.885  0.06077 .
gdp                                  1.377e-03  3.413e-04   4.035  7.6e-05 ***
coastline                            9.159e-02  5.526e-02   1.657  0.09891 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 32.52 on 213 degrees of freedom
  (1 observation deleted due to missingness)
Multiple R-squared:  0.3782,    Adjusted R-squared:  0.3432
F-statistic:  10.8 on 12 and 213 DF,  p-value: < 2.2e-16
```

The residuals look normally shaped. The residual standard error is 32.52 and range of dependent vaiable is 157, so the percentage error is (any prediction would still be off by) 32.52/157*100% = 20.71%.The adjusted R square is 0.3432, which means that 34.32% of the net migration can be explained by the independent variables. It is higher than that of the simple linear regression model and the multiple linear regression is better to explain why people want to move to a country. The p-value is tiny and absolutely smaller than 0.05. So, there is a significantly statistical relationship between net migration and the composition of GDP per capita, region and coastline.

Finally, for the assumption of residuals and multicollinearity, we use qqplot, VIF and gvlma test. The points in qqplot is distributed around 45 degree angle. Thus, the residuals are normally shaped. The VIF is smaller than 2 and some of vif test is larger than 2, but it doesn’t matter because they are factor variables. All of assumptions of gvlma is acceptable. Therefore, we can say this is a qualified model.




```
> vif(coun_model)
regionBALTICS                                 regionC.W. OF IND. STATES
                    1.096200                                  1.431027
regionEASTERN EUROPE                                regionLATIN AMER. & CARIB
                    1.413816                                  2.135788
regionNEAR EAST                         regionNORTHERN AFRICA
                    1.470300                                  1.169942
regionNORTHERN AMERICA                  regionOCEANIA
                    1.320023                                  1.703254
regionSUB-SAHARAN AFRICA                regionWESTERN EUROPE
                    2.363462                                  2.697600
                         gdp                                 coastline
                    2.501513                                  1.510178
> VIF(coun_model)
[1] 1.608209
```

	Value <dbl>	p-value <dbl>	Decision <chr>
Global Stat	2.75756275	0.5991801	Assumptions acceptable.
Skewness	1.68187009	0.1946761	Assumptions acceptable.
Kurtosis	0.02256526	0.8805931	Assumptions acceptable.
Link Function	0.94805155	0.3302158	Assumptions acceptable.
Heteroscedasticity	0.10507584	0.7458214	Assumptions acceptable.

Conclusion:

1, Most of countries' GDP are in the low level and a small number of rich developed countries raise the mean of GDP. But net migration is pretty normally shaped.

2, From the result of simple linear regression model, we can say that there is a significant linear relationship between net migration and GDP per capita. In other words, people are likely to move to the country with higher economic performance, which is more obvious when GDP per capita is larger than 6000 dollars.

3, From the result of multiple linear regression model, we can say that there is a significant linear relationship between net migration and composition of region, GDP per capita, coastline. In other words, people are likely to move to the country with higher economic performance, higher percent coastline area and specific regions. For regions, people prefer to move to West Europe and North America, and not Latin America and Caribbean, BAlTICS and Northern Africa.