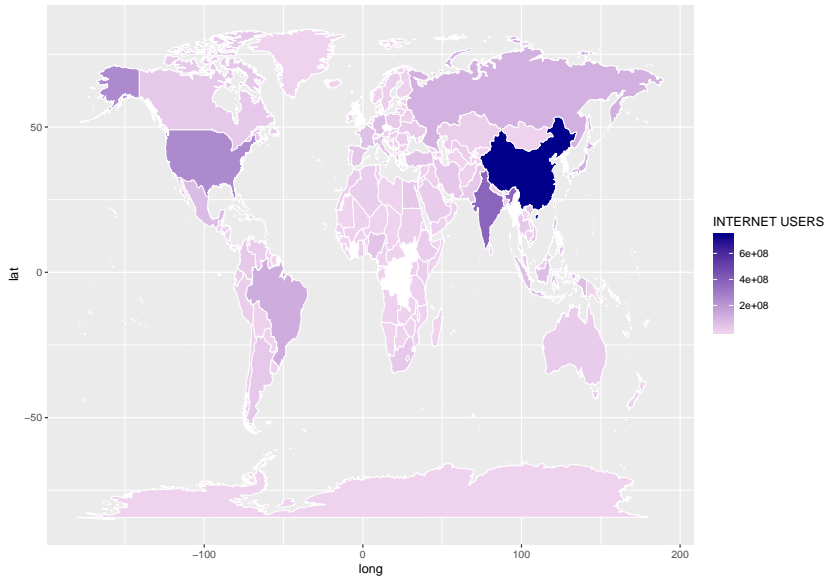# STA130 Final Project
Group 0206R2

Larry Shi, Anthony Wong, Yan Zhang, Mike Wang

Introduction

- ▶ According to the map above, there is a huge difference in the distribution of internet users around the world.
- ▶ What kind of factors are in play that create such a phenomenon?
- ▶ Analyze the correlation between Internet users and

1. Population
2. World Region
3. Life Expectancy (Health)
4. Economy & GDP
5. Goverment's Expenditure on Education
6. Political System
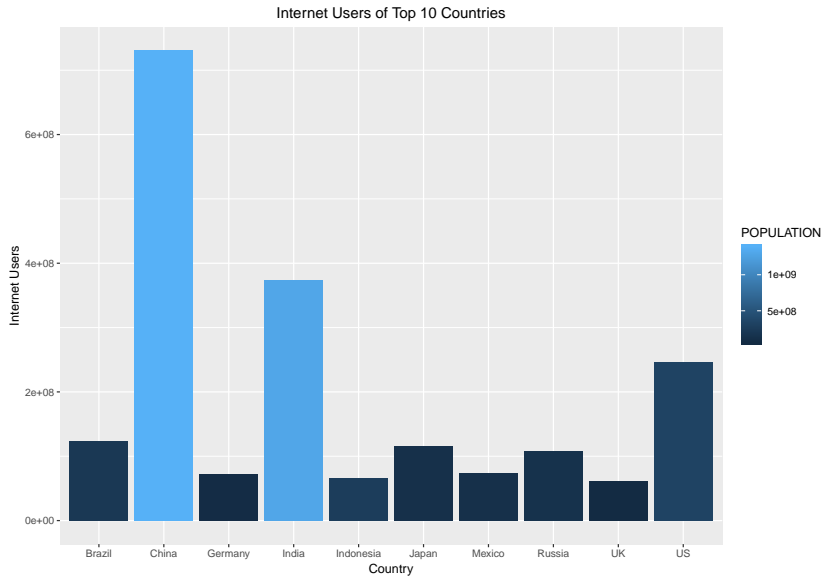7. Family Income Distribution

Statistical Methods

From the initial analysis of the population data against internet user data, we found that:
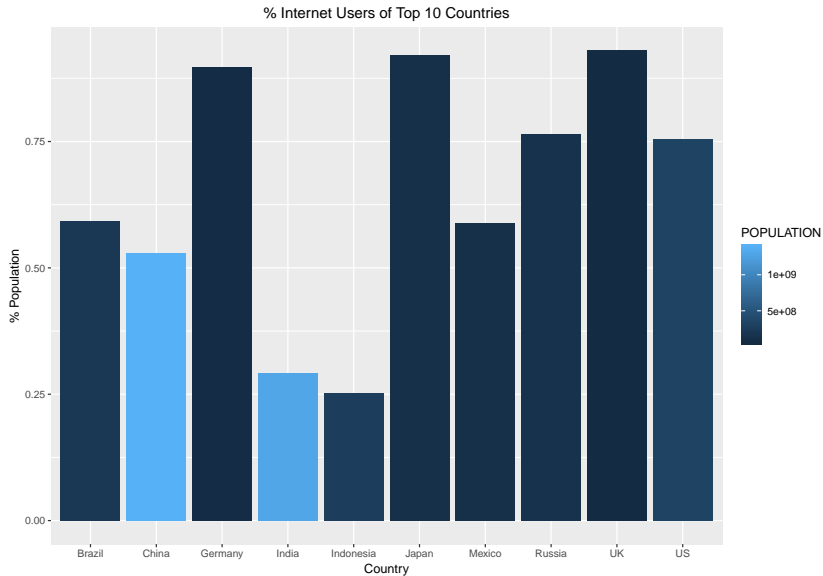
- ▶ Total population of a country doesn't really reflect the correlation between the two variables.
- ▶ Created a new variable, `Percentage = INTERNET USERS / POPULATION`, reflects the percentage of the population that uses internet
- ▶ Split `Percentage` into category of low, medium and high to allow us to classify internet users in terms of density.
- ▶ Split the GDP into 5 categories: Above $40000, between $40000 and $21000, between $21000 and $11500, between $11500 and $3700, and below $3700.
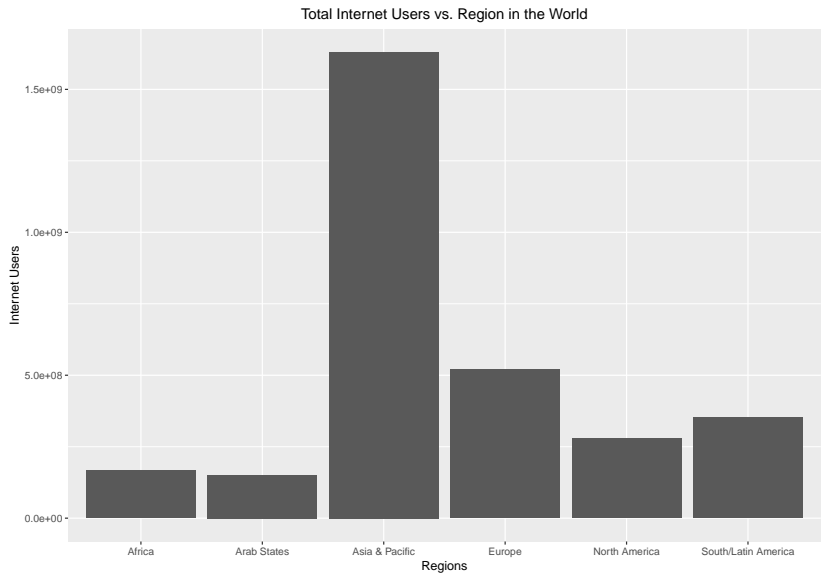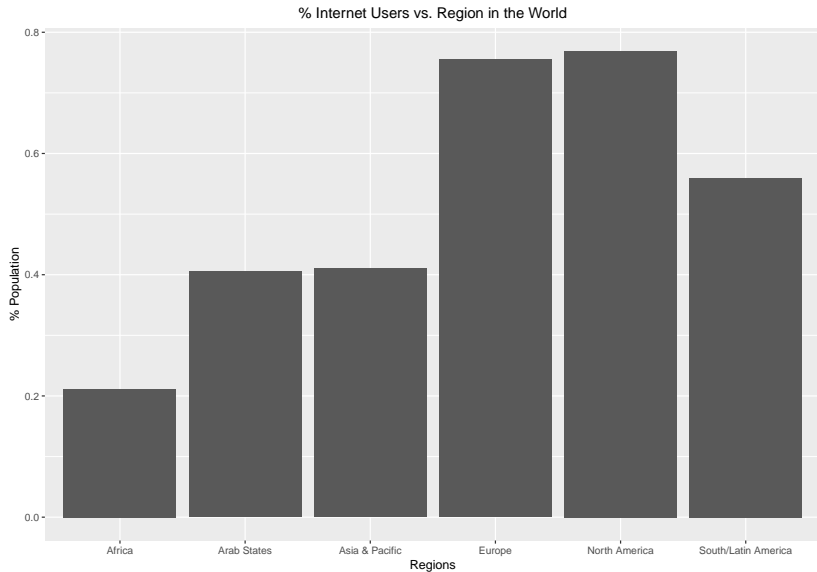
Results

Internet Users and Population

Internet Users of Top 10 Countries

Internet Users and Population

% Internet Users of Top 10 Countries

Internet Users and World Region

Total Internet Users vs. Region in the World

Internet Users and World Region

% Internet Users vs. Region in the World

Internet Users and Health

Life Expectancy vs. Total Number of Internet Users

Internet Users and Health

Life Expectancy vs. % Internet Users
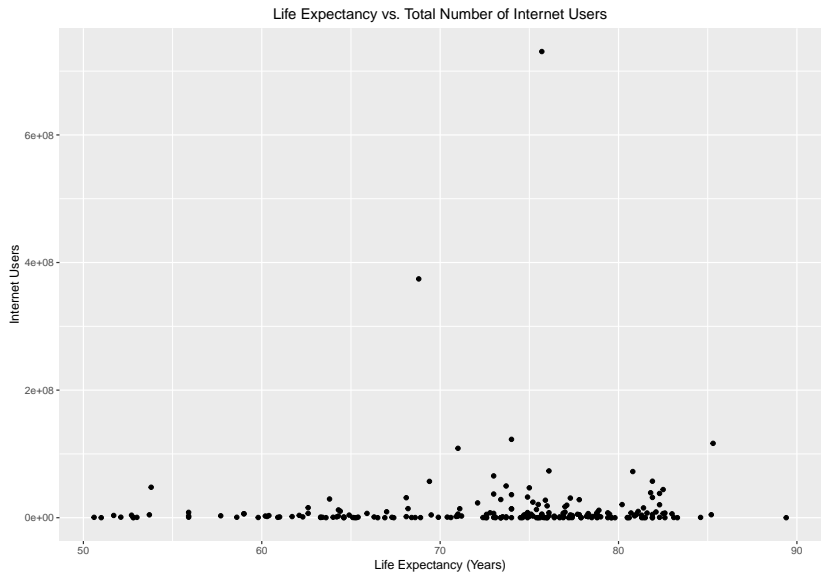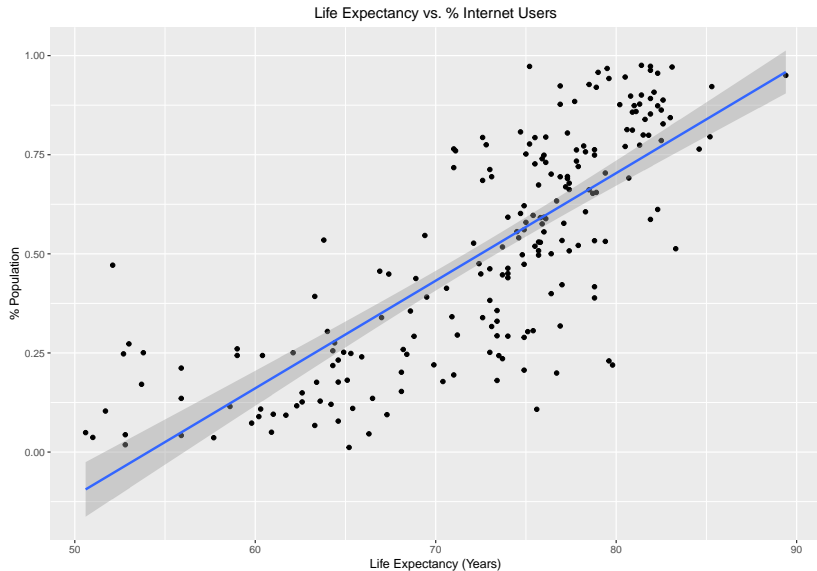
Internet Users and Health

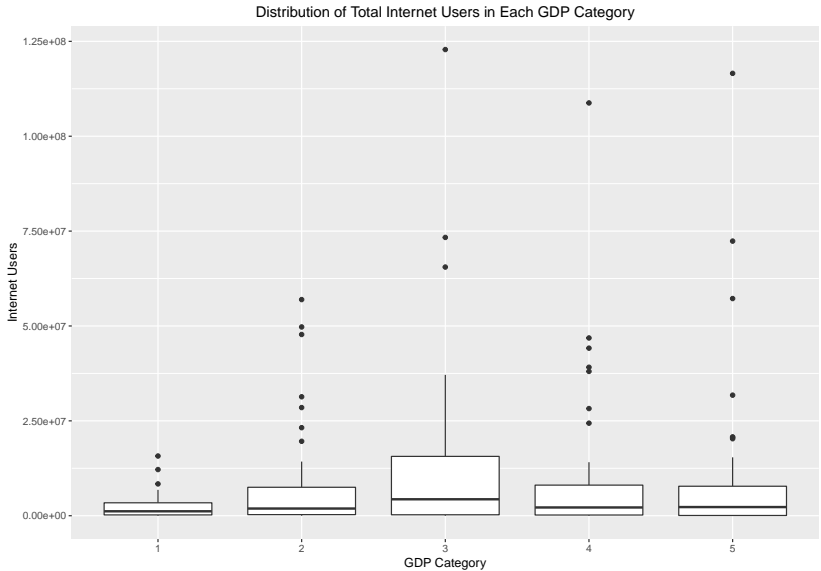Let H_0 be $\beta_1 = 0$ and H_a be $\beta_1 \neq 0$

```
##                 Estimate  Std. Error   t value      Pr(>|t
## (Intercept)  -1.46723360 0.108483328 -13.52497 1.991727e-
## `(YEARS)`     0.02713535 0.001482263  18.30670 1.864972e-
```

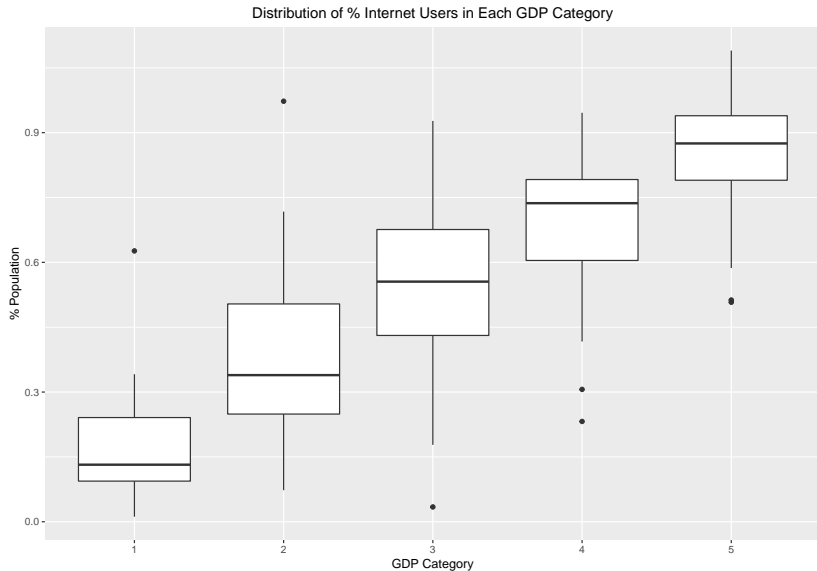Small p-value indicates strong evidence against the null hypothesis.

R squared value is rather higher indicating that the model is a good fit.

```
## [1] 0.6136489
```

Internet Users and GDP

Distribution of Total Internet Users in Each GDP Category

Internet Users and GDP

Distribution of % Internet Users in Each GDP Category

Internet Users and Education

% of GDP on Education vs. Total Number of Internet Users

Internet Users and Education
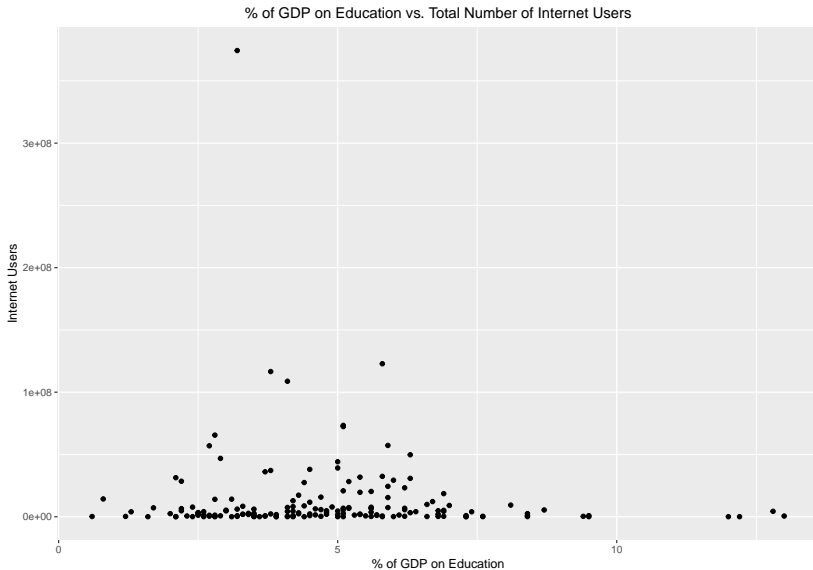
% of GDP on Education vs. % Internet Users
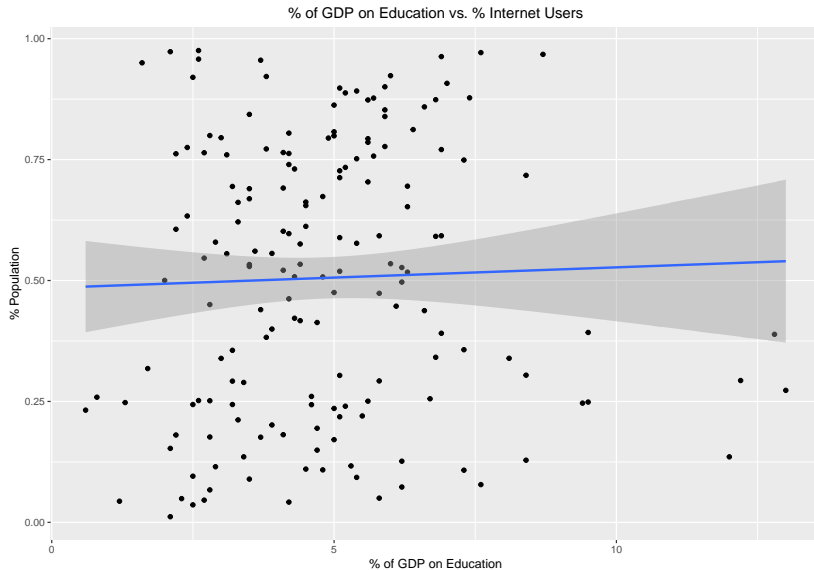
Internet Users and Education

Let H_0 be $\beta_1 = 0$ and H_a be $\beta_1 \neq 0$

```
##                    Estimate Std. Error    t value      Pr(>|t
## (Intercept)     0.48487301 0.05326949  9.1022642 2.600223e-1
## `(% OF GDP)`    0.00423642 0.01008778  0.4199555 6.750578e-0
```

Large p-value indicates no evidence against the null hypothesis.

Really smaller R squared value indicating that the model is a bad fit.
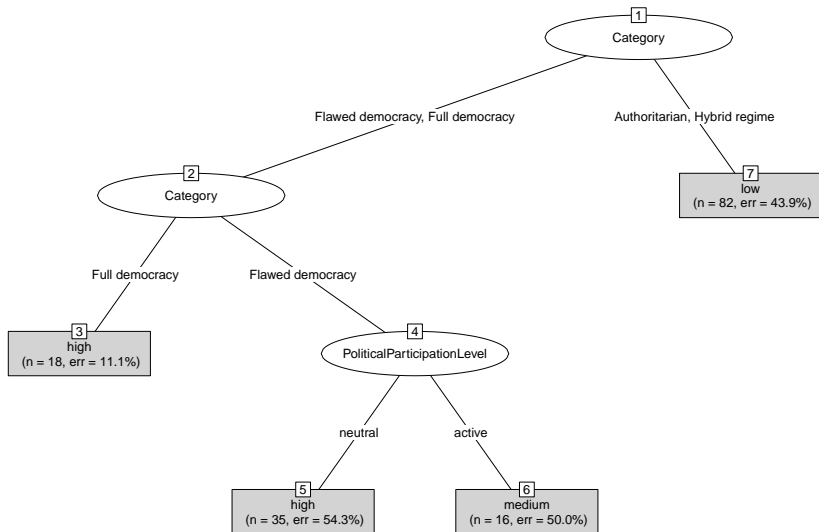
```
## [1] 0.001054949
```

Internet Users and Democracy

Internet Users and Democracy
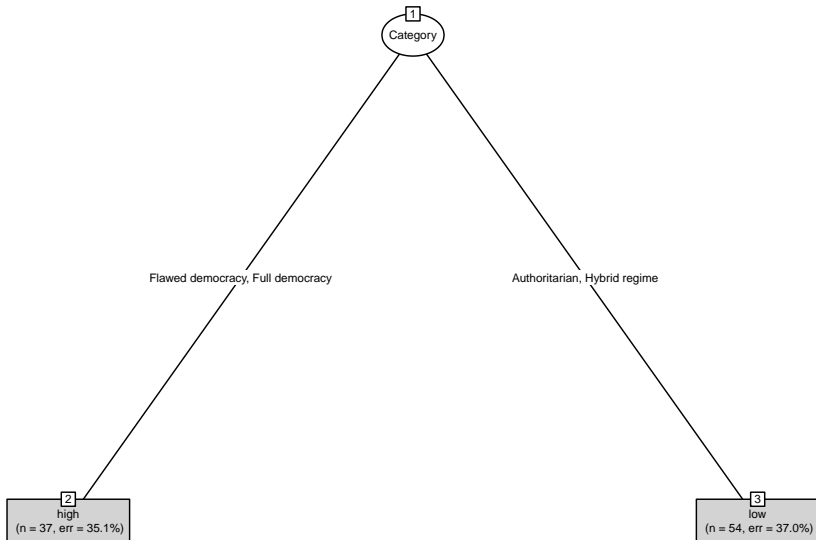
Internet Users and Democracy

```
## 
## predicted_tree high low medium
##        high     13    5     14
##        low       7   12      9
##        medium    0    0      0

## [1] 0.5081967
```

Internet Users and Family Income Distributions

Family Income (Gini Index) vs. % Internet Users

Internet Users and Family Income Distributions

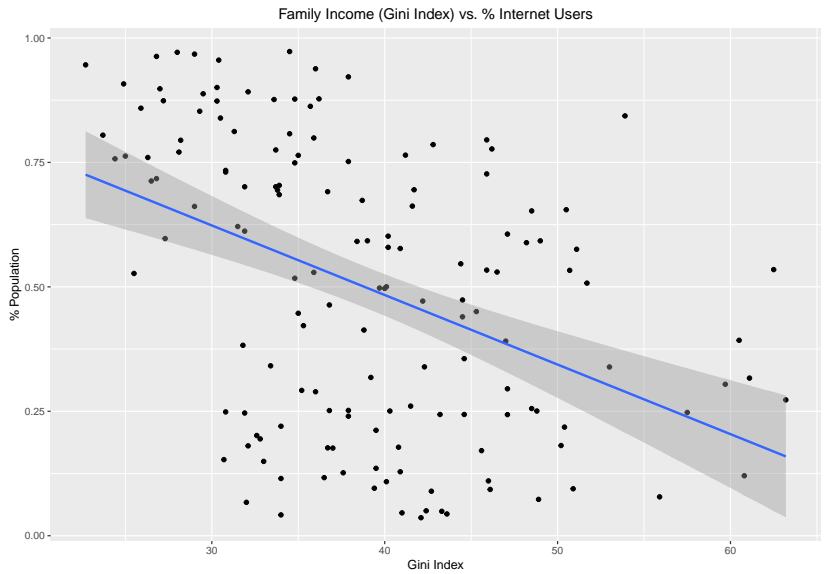Let H_0 be $\beta_1 = 0$ and H_a be $\beta_1 \neq 0$

```
##                    Estimate  Std. Error   t value      Pr(>
## (Intercept)     1.04265442 0.095616171 10.904582 1.087663e
## `Gini Index`   -0.01397581 0.002398027 -5.828046 3.379064e
```

Very small p-value indicates that there is strong evidence against the null hypothesis.

Small R squared value indicates that the model is a bad fit.

```
## [1] 0.1866618
```

Conclusion (Population and Regions)

- The number of internet users seems to increase as the population increase.
- However, some countries have significantly higher population than others causing inaccurate result.
- % internet users is used instead showing there are not "real" correlation between population and internet users.

- Asia & Pacific region appeared to have most internet users due to its high total population.
- A different graph plotted using % internet users shows that North America and Europe have much internet usage.
- Both North America and Europe are mostly consistent with developed countries, while Asia, Africa and South America have many developing countries.

Conclusion (Life Expectancy and GDP)

- ▶ Initialial analysis showed that there is no correlation between life expectancy/GDP and the number of internet users.
- ▶ Using % internet users, the plot shows a strong positive correlation between life expectancy/GDP and % internet users.

- ▶ Developed countries usually have better health care system resulting in higher life expectancy/GDP, and according to the world region analysis, this further shows that developed countries tends to have higher internet usage.

- ▶ Developed countries tends to have higher GDP which also result in higher % internet users, whereas developing countries with lower GDP have smaller % internet users.
- ▶ Some developing countries that are making an effort to be a developed country (Countries have medium GDP, e.g. Brazil, Mexico) have the large variation of internet users domestically.

Conclusion (Education, Political System and Family Income)

▶ Both total number of internet users and % internet users shows no correlation between % GDP spent on education and internet users.

▶ Countries that are open to democracy (i.e. full/flawed democracy) tend to have greater % of internet users, while countries with more restricted government (i.e. authoritarian & Hybrid) tend to have lower % of internet users.

▶ However, the accuracy of the classification tree explained by the confusion matrix shows that the training set doesn't fit well on the testing set.

▶ There exists a negative correlation between percentage of internet users and family income. Since the coefficient of determination is low, the linear model doesn't fit well and thus the data is non-linear.