Evan Analysis

Evan Meade

4/26/2021

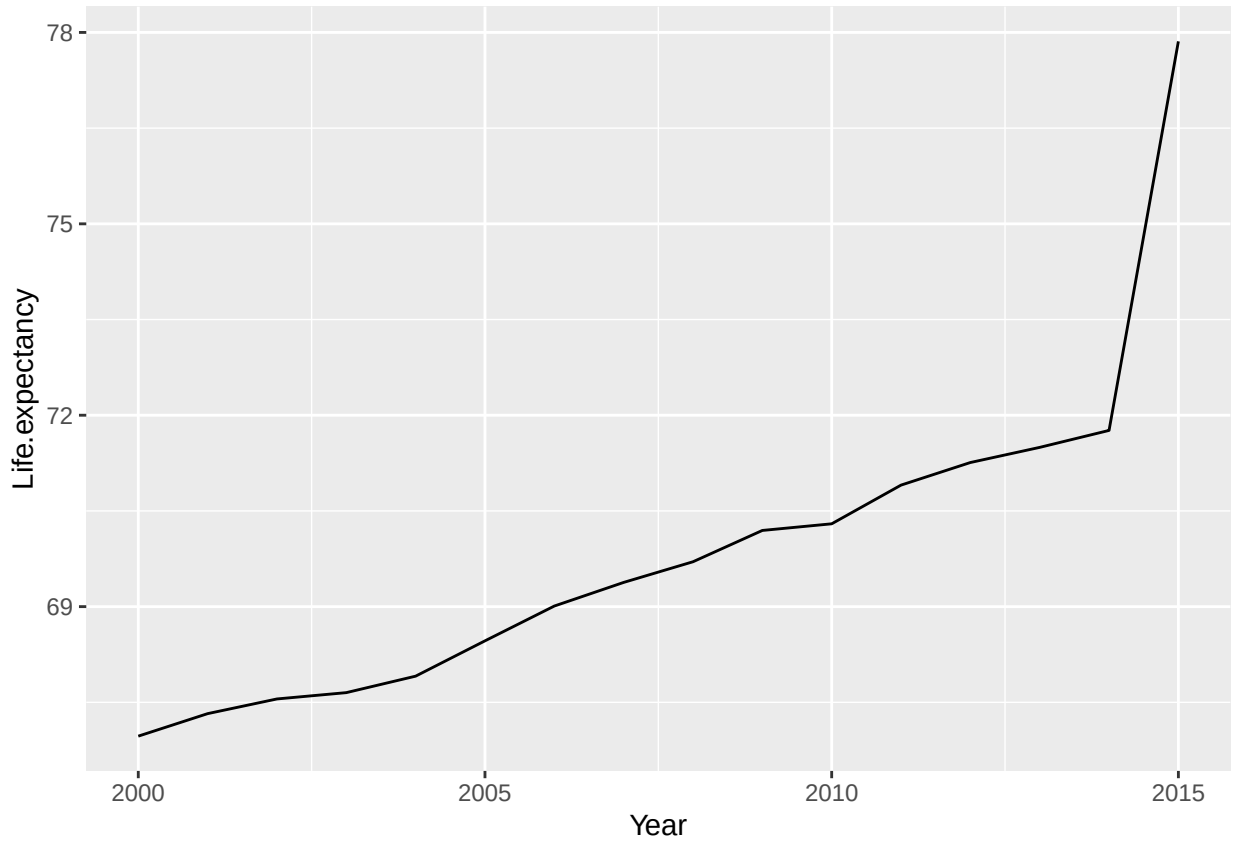
Imports

```
library(ggplot2)

data <- read.csv("life_CLEAN.csv")
str(data)
```

```
## 'data.frame':    2311 obs. of  13 variables:
##  $ Country                        : chr  "Afghanistan" "Afghanistan" "Afghanistan" "Afghanistan" ...
##  $ Year                           : int  2015 2014 2013 2012 2011 2010 2009 2008 2007 2006 ...
##  $ Status                         : chr  "Developing" "Developing" "Developing" "Developing" ...
##  $ Life.expectancy                : num  65 59.9 59.9 59.5 59.2 58.8 58.6 58.1 57.5 57.3 ...
##  $ Schooling                      : num  10.1 10 9.9 9.8 9.5 9.2 8.9 8.7 8.4 8.1 ...
##  $ GDP                            : num  584.3 612.7 631.7 670 63.5 ...
##  $ Alcohol                        : num  0.01 0.01 0.01 0.01 0.01 0.01 0.01 0.03 0.02 0.03 ...
##  $ BMI                            : num  19.1 18.6 18.1 17.6 17.2 16.7 16.2 15.7 15.2 14.7 ...
##  $ percentage.expenditure         : num  71.3 73.5 73.2 78.2 7.1 ...
##  $ Income.composition.of.resources: num  0.479 0.476 0.47 0.463 0.454 0.448 0.434 0.433 0.415 0.405
##  $ HIV.AIDS                       : num  0.1 0.1 0.1 0.1 0.1 0.1 0.1 0.1 0.1 0.1 ...
##  $ thinness.5.9.years             : num  17.3 17.5 17.7 18 18.2 18.4 18.7 18.9 19.1 19.3 ...
##  $ thinness.10.19.years           : num  17.2 17.5 17.7 17.9 18.2 18.4 18.6 18.8 19 19.2 ...
```

Temporal Analysis

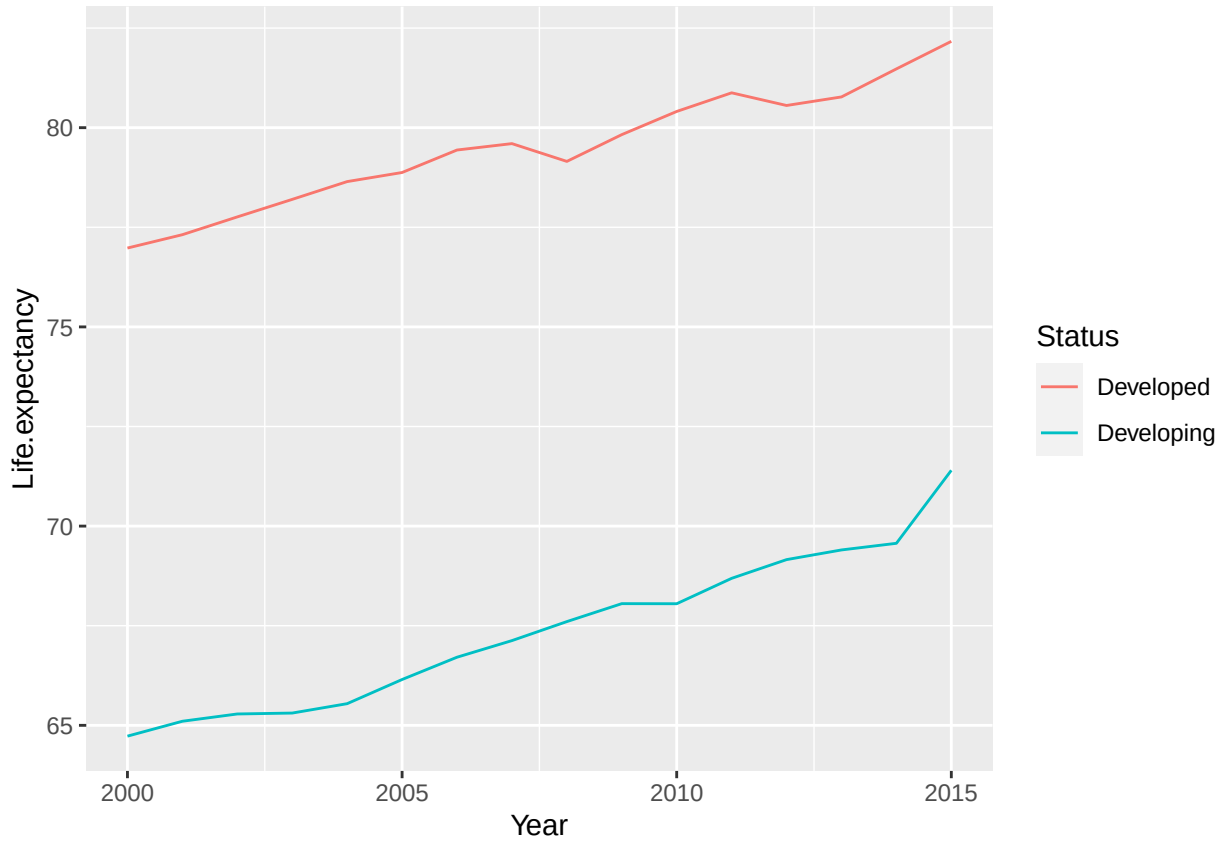
```
year_averages <- aggregate(Life.expectancy ~ Year, data = data, FUN = mean)

ggplot(data = year_averages) +
  geom_line(mapping = aes(x = Year, y = Life.expectancy))
```




```r
summary(lm(Life.expectancy ~ Year, data = year_averages))
```

```
##
## Call:
## lm(formula = Life.expectancy ~ Year, data = year_averages)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.2597 -0.6871 -0.2405  0.0802  4.3518
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -907.02839  145.05142  -6.253 2.12e-05 ***
## Year           0.48662    0.07225   6.735 9.55e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.332 on 14 degrees of freedom
## Multiple R-squared:  0.7641, Adjusted R-squared:  0.7473
## F-statistic: 45.36 on 1 and 14 DF,  p-value: 9.551e-06
```

```r
status_averages <- aggregate(Life.expectancy ~ Year + Status, data = data, FUN = mean)

ggplot(data = status_averages) +
  geom_line(mapping = aes(x = Year, y = Life.expectancy, color = Status))
```




```
data$Status.proxy <- as.integer(factor(data$Status, levels = c("Developing",
                                                               "Developed"),
                                       labels = c(0, 1))) - 1
summary(lm(Life.expectancy ~ Status.proxy + Year, data = data))
```

```
##
## Call:
## lm(formula = Life.expectancy ~ Status.proxy + Year, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -31.883  -5.658   1.239   6.384  21.901
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -657.67632   80.43631  -8.176 4.77e-16 ***
## Status.proxy   12.22552    0.44800  27.289  < 2e-16 ***
## Year            0.36112    0.04008   9.011  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.328 on 2308 degrees of freedom
## Multiple R-squared:  0.2642, Adjusted R-squared:  0.2636
## F-statistic: 414.4 on 2 and 2308 DF,  p-value: < 2.2e-16
```

So clearly country status is a large factor in life expectancy over time. On average, developed countries have

life expectancies over 12 years higher than developing countries. Additionally, every year raises the global life expectancy by about a third of a year. These coefficients are highly significant and graphically the trends look relatively constant. If we extrapolate this pattern forwards, that means each person can expect to life on average, 1.5 times as long as the life expectancy when they were born. For people born around 2000, that's about 116 years old.

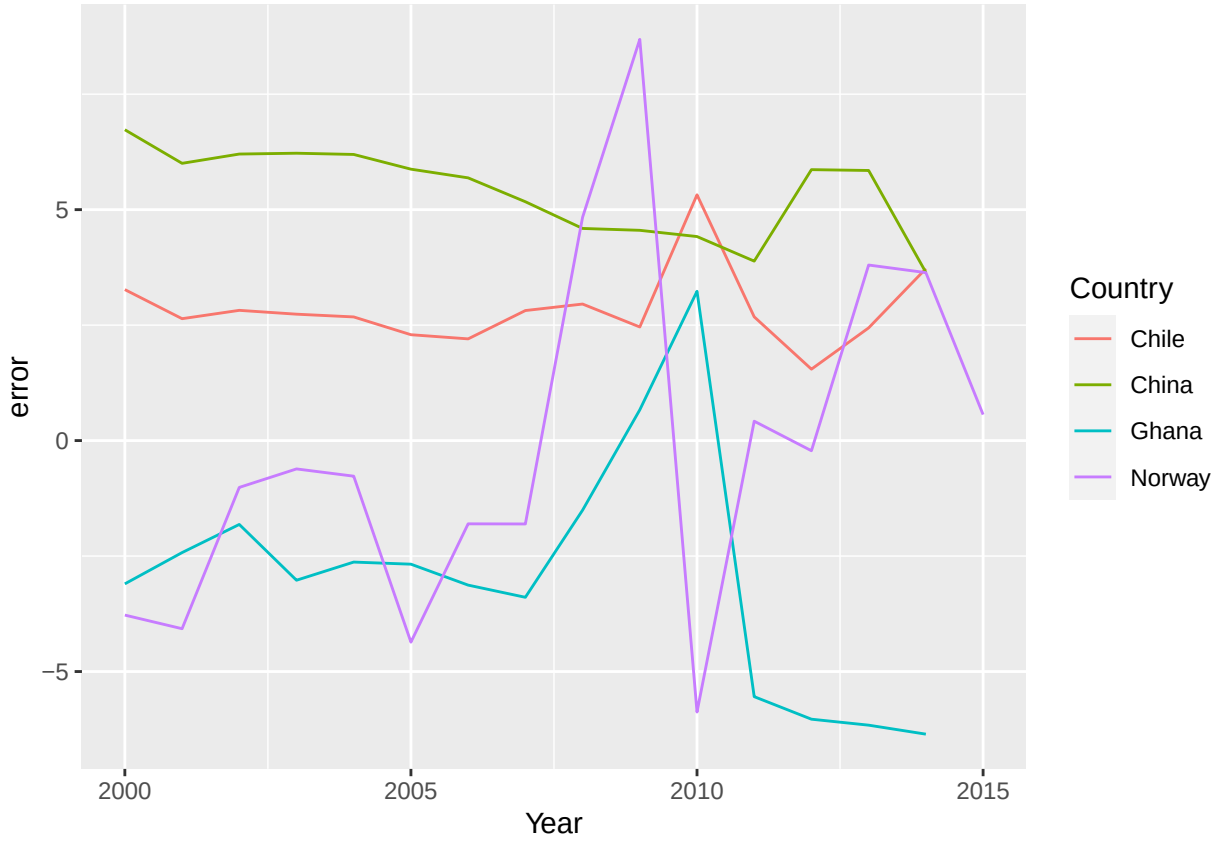
Model Fitting

```
model <- lm(Life.expectancy ~ Schooling + GDP + Alcohol + BMI +
              percentage.expenditure + Income.composition.of.resources +
              HIV.AIDS + thinness.5.9.years + thinness.10.19.years, data = data)
summary(model)
```

```
##
## Call:
## lm(formula = Life.expectancy ~ Schooling + GDP + Alcohol + BMI +
##     percentage.expenditure + Income.composition.of.resources +
##     HIV.AIDS + thinness.5.9.years + thinness.10.19.years, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -27.5813  -2.5470  -0.0207   2.6412  25.4956
##
## Coefficients:
##                                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)                      4.980e+01  4.985e-01  99.895  < 2e-16 ***
## Schooling                        1.078e+00  5.206e-02  20.702  < 2e-16 ***
## GDP                              4.434e-05  1.709e-05   2.594 0.009537 **
## Alcohol                         -1.009e-01  2.961e-02  -3.407 0.000668 ***
## BMI                              5.476e-02  6.336e-03   8.642  < 2e-16 ***
## percentage.expenditure           1.773e-04  1.111e-04   1.596 0.110670
## Income.composition.of.resources  9.957e+00  7.498e-01  13.279  < 2e-16 ***
## HIV.AIDS                        -6.609e-01  1.768e-02 -37.389  < 2e-16 ***
## thinness.5.9.years              -4.548e-02  5.638e-02  -0.807 0.419871
## thinness.10.19.years            -7.606e-02  5.757e-02  -1.321 0.186593
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.522 on 2301 degrees of freedom
## Multiple R-squared:  0.7837, Adjusted R-squared:  0.7829
## F-statistic: 926.5 on 9 and 2301 DF,  p-value: < 2.2e-16
```

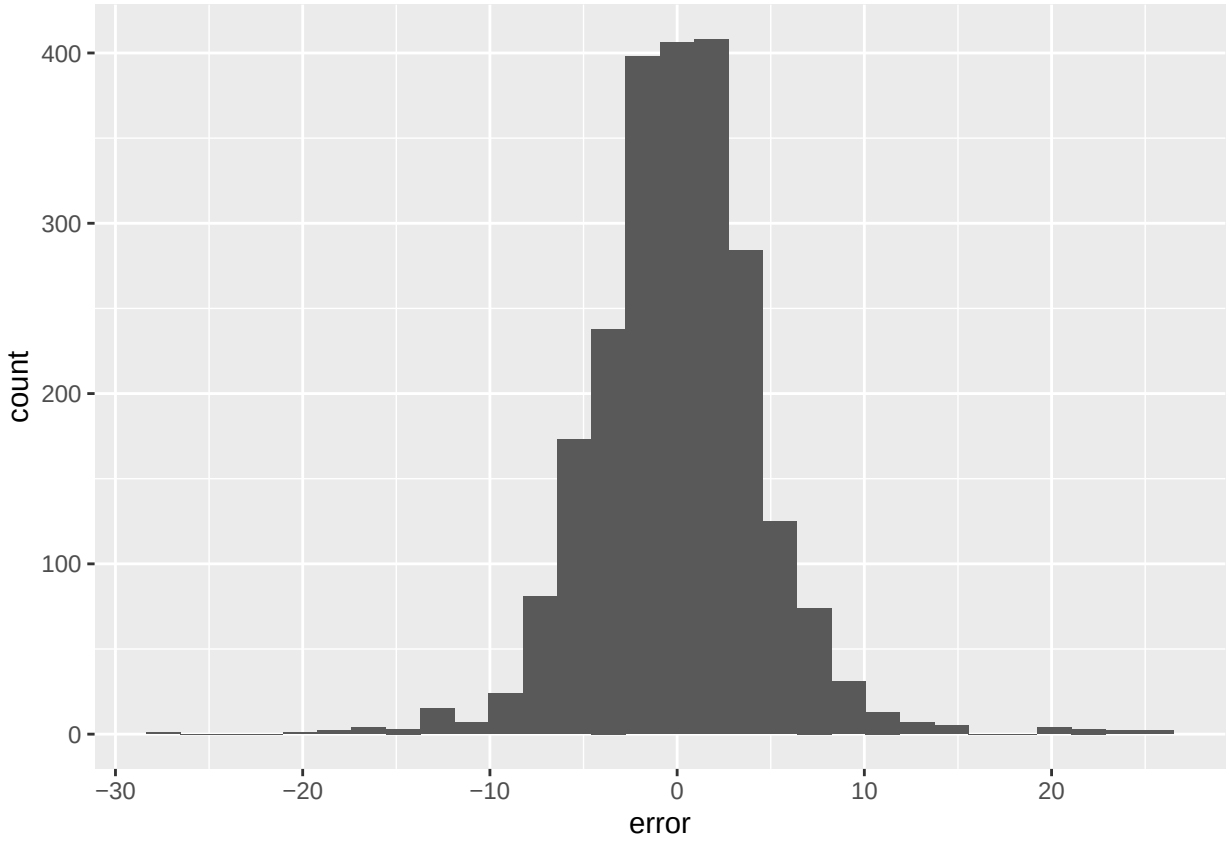
Over- and Under-Performing Countries

```
data$y.hat <- predict(model, data)
data$error <- data$Life.expectancy - data$y.hat

country_subset <- c("Norway", "China", "Ghana", "Chile", "Vietnam")
ggplot(data = data[which(data$Country %in% country_subset), ]) +
  geom_line(mapping = aes(x = Year, y = error, color = Country))
```

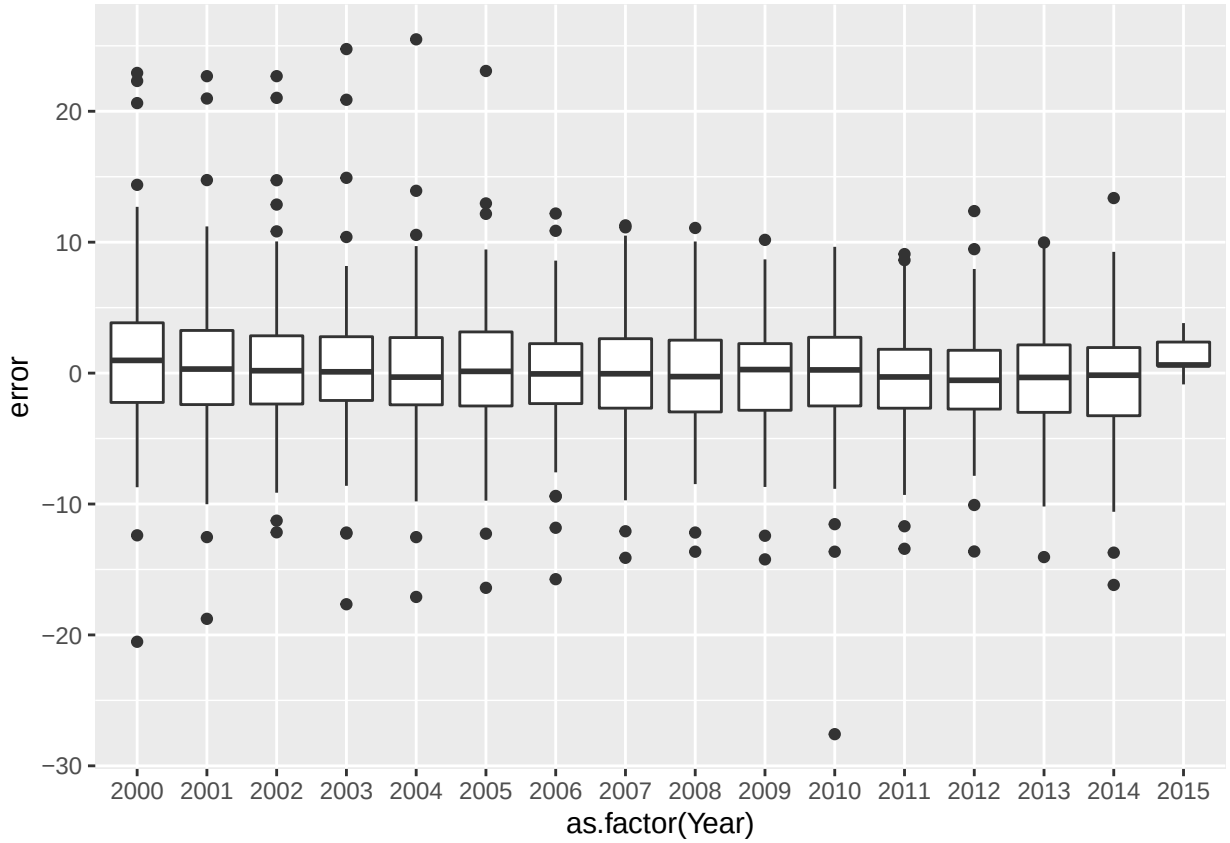



```
ggplot(data = data) +
  geom_histogram(mapping = aes(x = error))
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

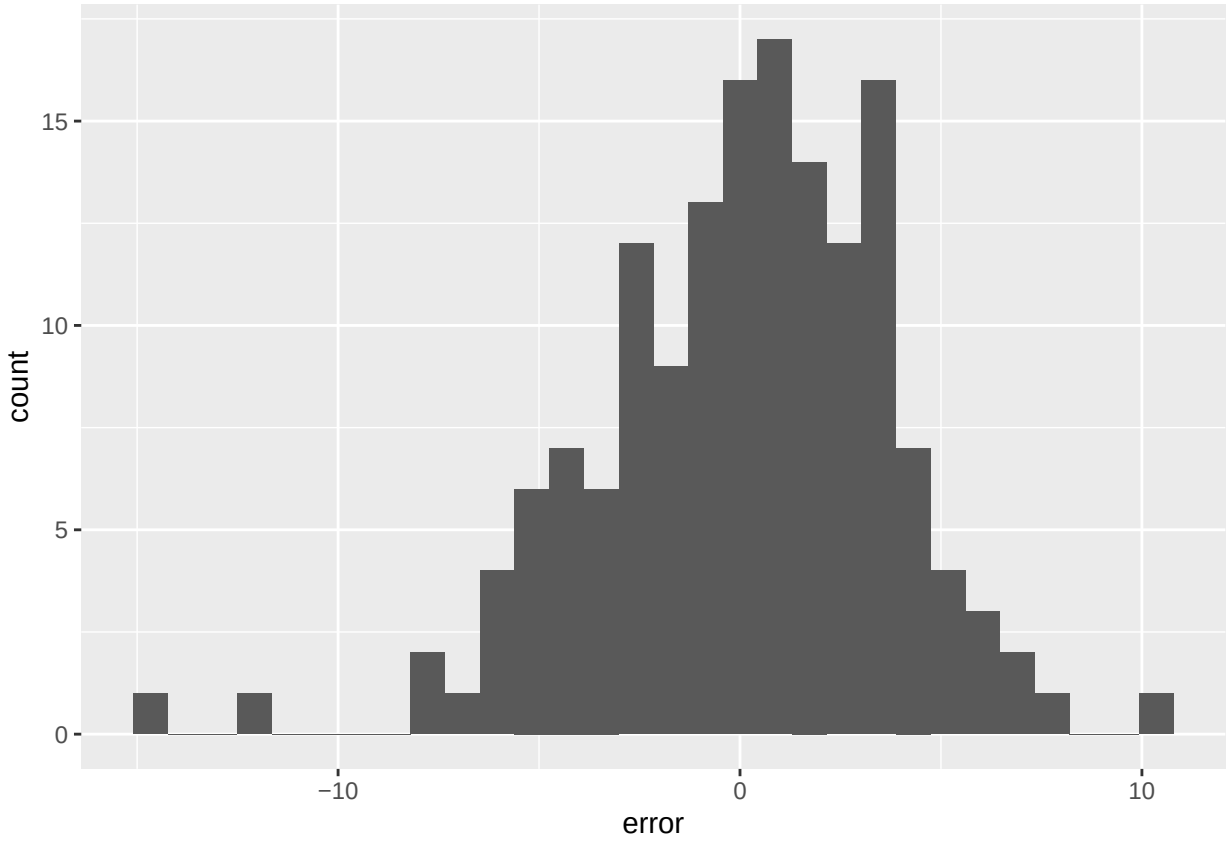



```
ggplot(data = data) +
  geom_boxplot(mapping = aes(x = as.factor(Year), y = error))
```




```
av_error <- aggregate(error ~ Country, data = data, FUN = mean)

ggplot(data = av_error) +
  geom_histogram(mapping = aes(x = error))
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



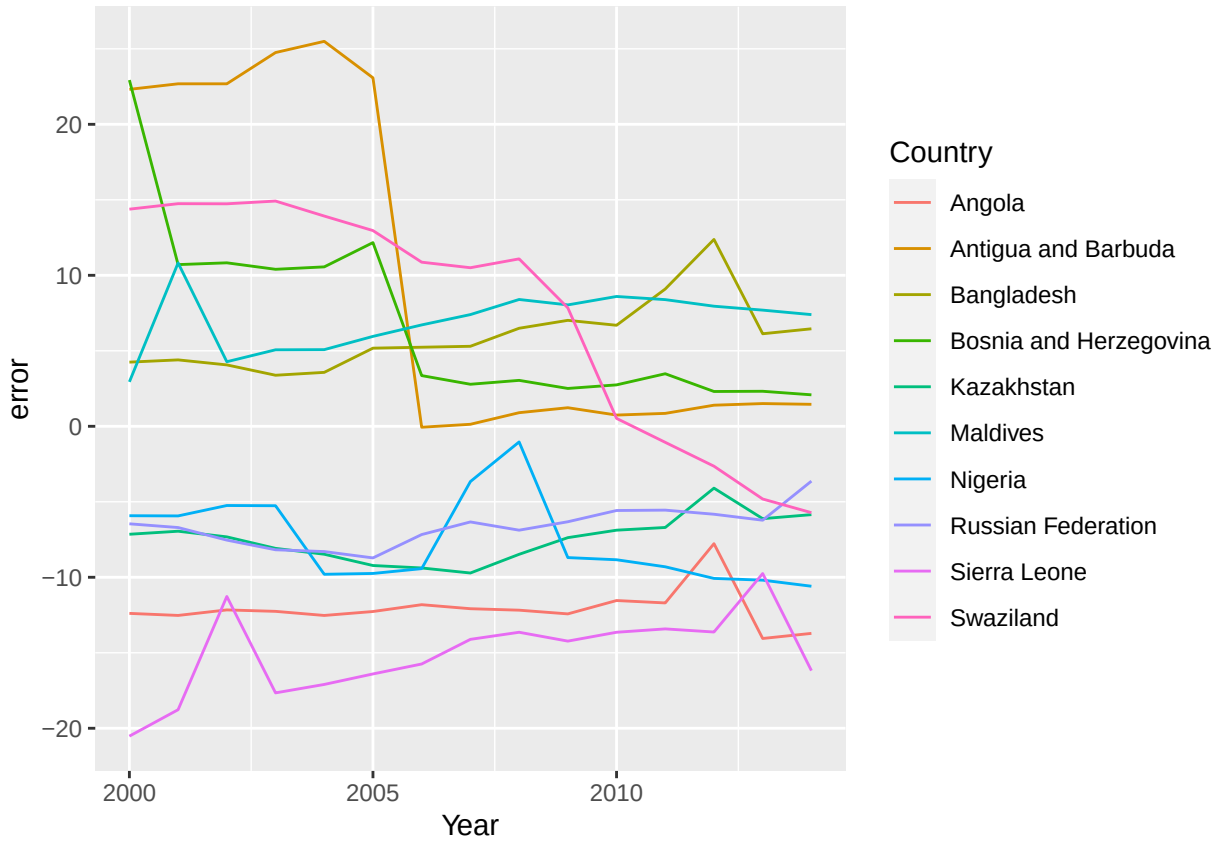

```
outlier_order <- c(head(order(av_error$error, decreasing = TRUE), 5),
                   tail(order(av_error$error, decreasing = TRUE), 5))
av_error$over.rank <- order(av_error$error, decreasing = TRUE)
outliers <- av_error$Country[outlier_order]
print(paste0(c("Top Over-perfomers: ", outliers[1:5])))
```

```
## [1] "Top Over-perfomers: "   "Antigua and Barbuda"    "Swaziland"
## [4] "Maldives"               "Bosnia and Herzegovina" "Bangladesh"
```

```
print(paste0(c("Top Over-perfomers: ", outliers[10:6])))
```

```
## [1] "Top Over-perfomers: " "Sierra Leone"         "Angola"
## [4] "Nigeria"              "Kazakhstan"           "Russian Federation"
```

```
outlier_data <- data[which(data$Country %in% outliers), ]

ggplot(data = outlier_data) +
  geom_line(mapping = aes(x = Year, y = error, color = Country))
```


20
10
0
−10
−20
error
2000
2005
2010
Year
Country
Angola
Antigua and Barbuda
Bangladesh
Bosnia and Herzegovina
Kazakhstan
Maldives
Nigeria
Russian Federation
Sierra Leone
Swaziland