Predicting Happiness

An Attempt

Alyssa Andrichik, Eva Licht, and Joe Yalowitz

Abstract

We investigate the World Happiness Index in order to build a model that can predict a country's happiness score based on demographic and geographic factors such as literacy levels, cell-phone use, birth rate, death rate, GDP per capita, perceived corruption, etc. We build an array of linear models, simple and with interaction, and use other regression analysis tools such as ridge, lasso, and principal component analysis to understand our data. We also use regression trees, random forests, and boosted trees to develop prediction methods for our research question. We determine that a country's region is the best predictor of its happiness score.

–

Introduction

Happiness can define a country. Beyond the great importance of individual and communal well-being that accompanies happiness and positive emotions, studies show that positive emotions contribute to "broadening workers" individual mindsets, enabling them to build up their personal resources in terms of enhanced sensitivity and positive attitudes toward their workplace," and can increase productivity (De Satio 2019). Predicting a country's happiness can aid governments in supporting their citizens and ensure greater well-being.

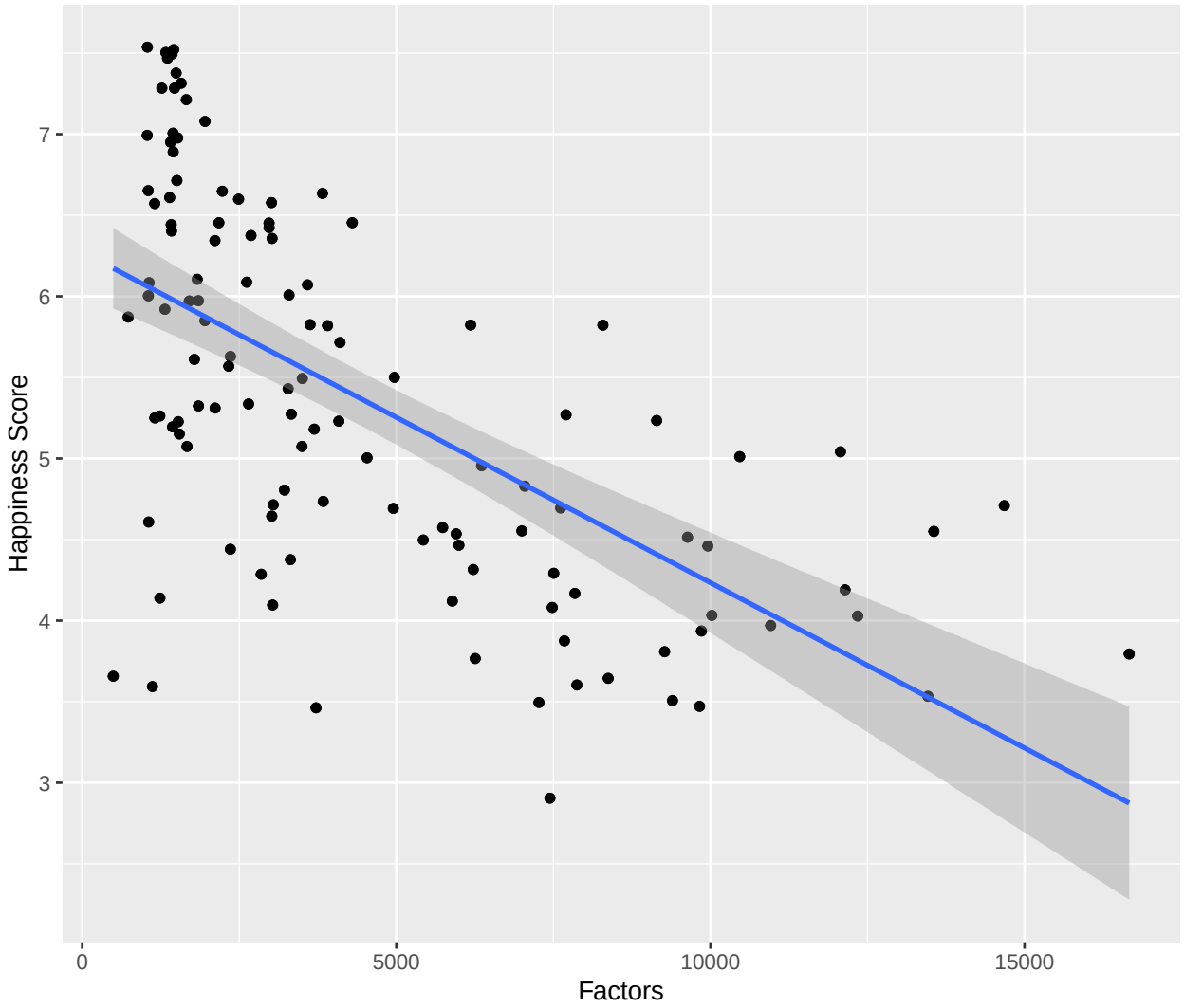
In terms of data analysis, we employed both data model analysis and algorithmic model analysis. Two primary research questions guided our project. First, can we use a mixture of demographic and geographic data to predict happiness for a country? Second, is there a significant difference in happiness score between regions of the world, and is region a significant predictor of happiness?

–

The Data

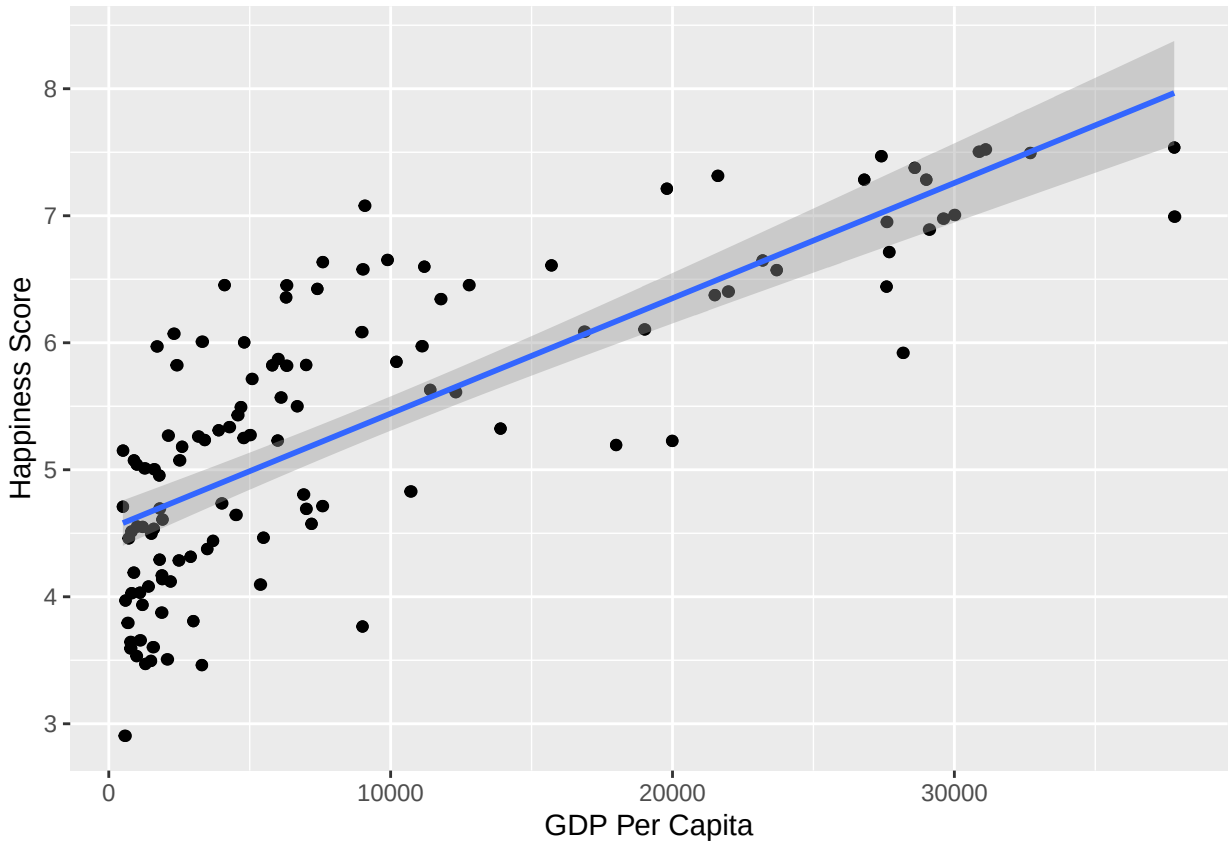
Our final dataset pulls data from a wide variety of sources. We obtained our "happiness" data from the Gallup World Poll. Our demographic data on countries came from the US Government. Data on corruption came from Transparency International. We combined these disparate sources into one data set with 120 observations and 22 variables. Each observation refers to a country of the world. These combinations created our complete dataset, but required immense data wrangling. As the sources differed, merging by country resulted in errors because each data set recorded country name differently. We had to mutate the data to eliminate these differences: changing all three data sets to reporting "United States" rather than "The United States" or "United States of America." Additionally, the data was collected by several different organizations, and some data had commas to signify decimals, rather than periods. Variable names had to be normalized and checked for any possible causes of error; for example, GDP Per Capita was originally reported as "GDP ($ per capita)" and the dollar sign prompted errors in the code. After fixing the variable names, merging, and checking the data reporting format, the data was ready for its initial analysis.

Exploratory Data Analysis



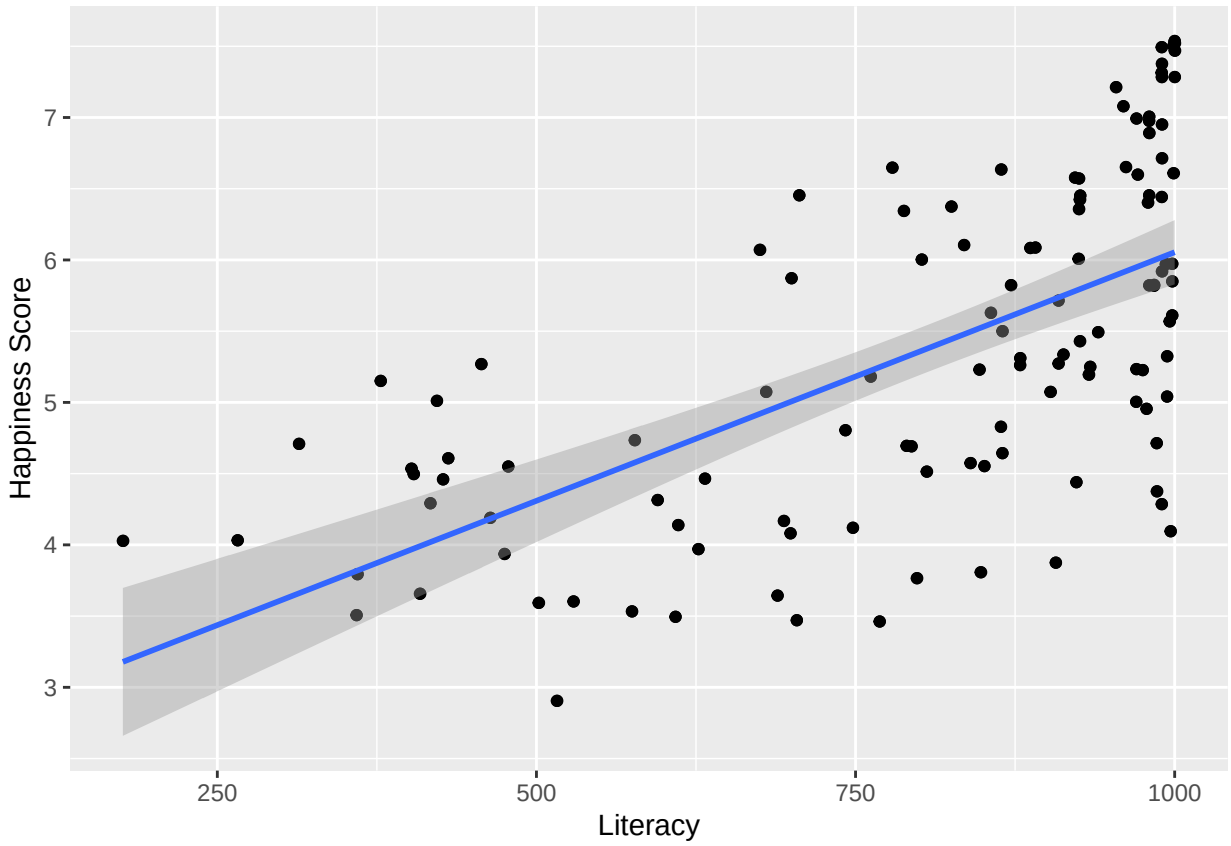
Adjusted R-squared: 0.4808

```
mod5<- lm(HappinessScore~GDP_percapita, data=happy_country)
ggplot(happy_country, aes(x=GDP_percapita, y=HappinessScore)) +
  geom_point()+
  labs(x="GDP Per Capita", y="Happiness Score") +
  geom_jitter()+
  stat_smooth(method = "lm")
```

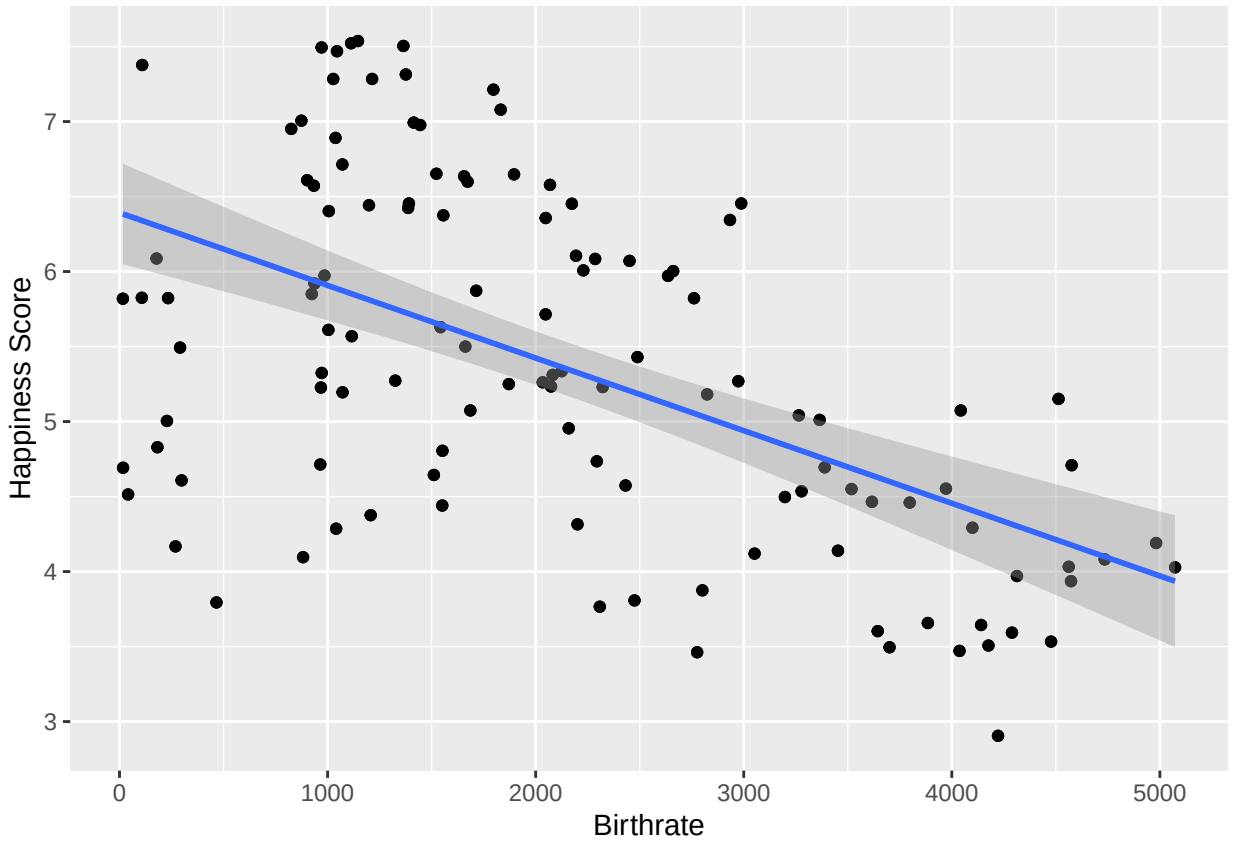


Adjusted R-squared: 0.5964

```
lm1<- lm(HappinessScore~Literacy, data= happy_country)
lmplot1<- ggplot(happy_country, aes(x = Literacy, y = HappinessScore)) +
  geom_point() +
  labs(x="Literacy", y="Happiness Score")+
  geom_jitter() +
  stat_smooth(method = "lm")
lmplot1
```

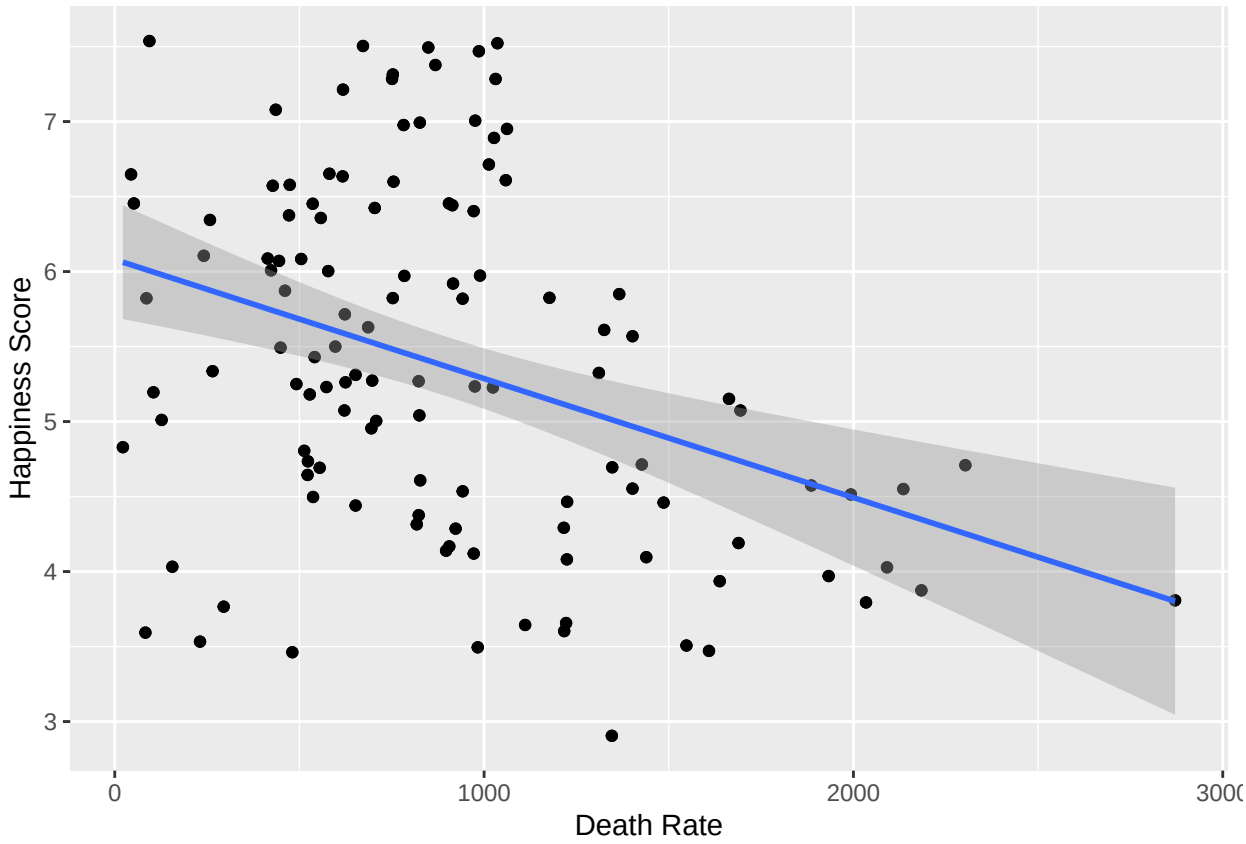



```
lm2<- lm(HappinessScore~Birthrate, data= happy_country)
lmplot2<- ggplot(happy_country, aes(x = Birthrate, y = HappinessScore)) +
  geom_point() +
  labs(x="Birthrate", y="Happiness Score")+
  geom_jitter() +
  stat_smooth(method = "lm")
lmplot2
```



Adjusted R-squared: 0.292

```
lm3<- lm(HappinessScore~Deathrate, data= happy_country)
lmplot3<- ggplot(happy_country, aes(x = Deathrate, y = HappinessScore)) +
  geom_point() +
  labs(x="Death Rate", y="Happiness Score")+
  geom_jitter() +
  stat_smooth(method = "lm")
lmplot3
```

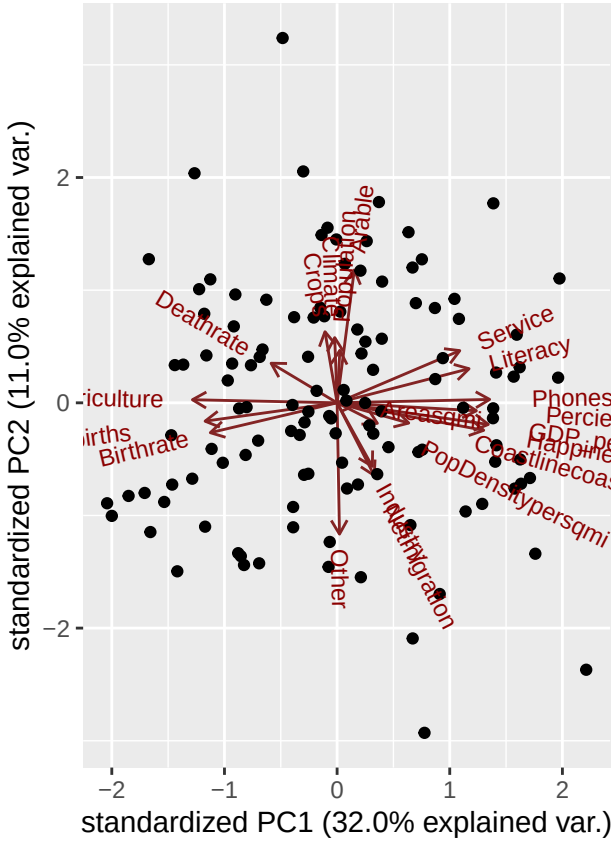


Adjusted R-squared: 0.1253

PCA

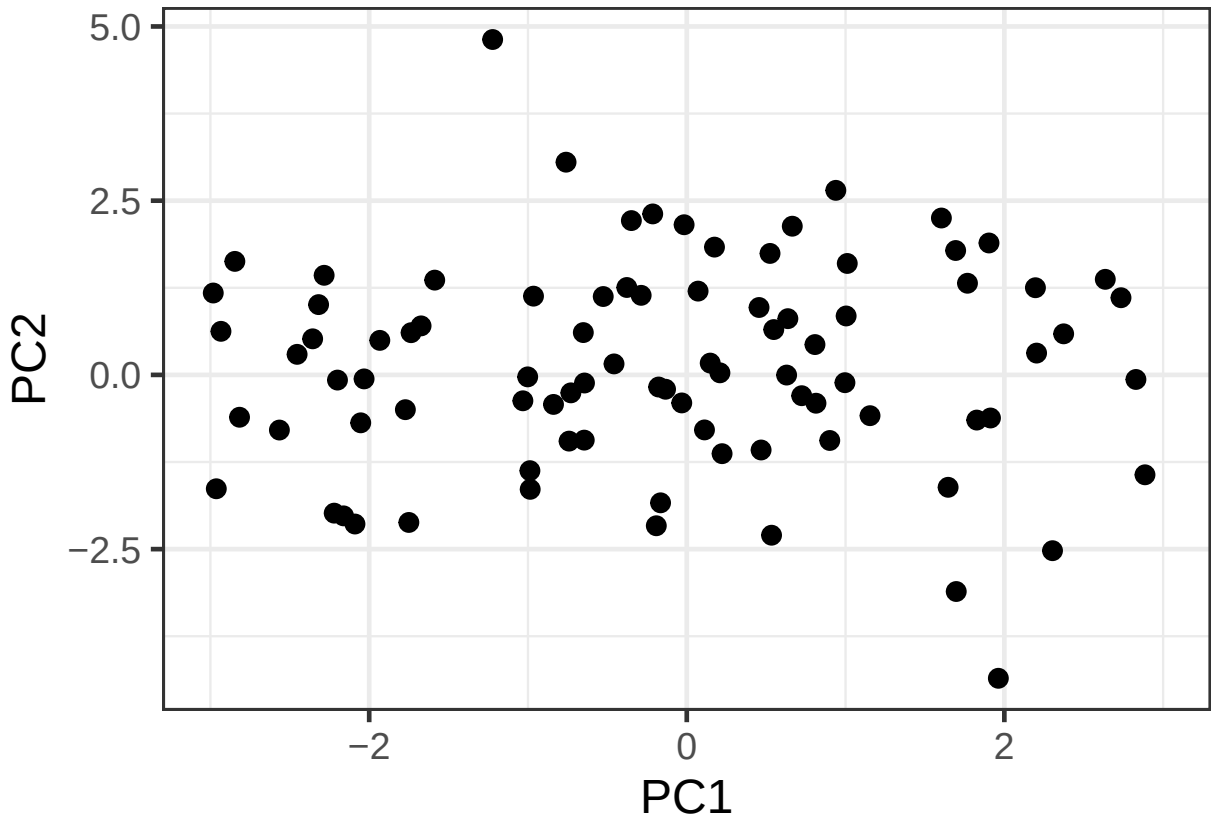
```
pca1 <- prcomp(num_happy, center=TRUE, scale. =  TRUE)
```

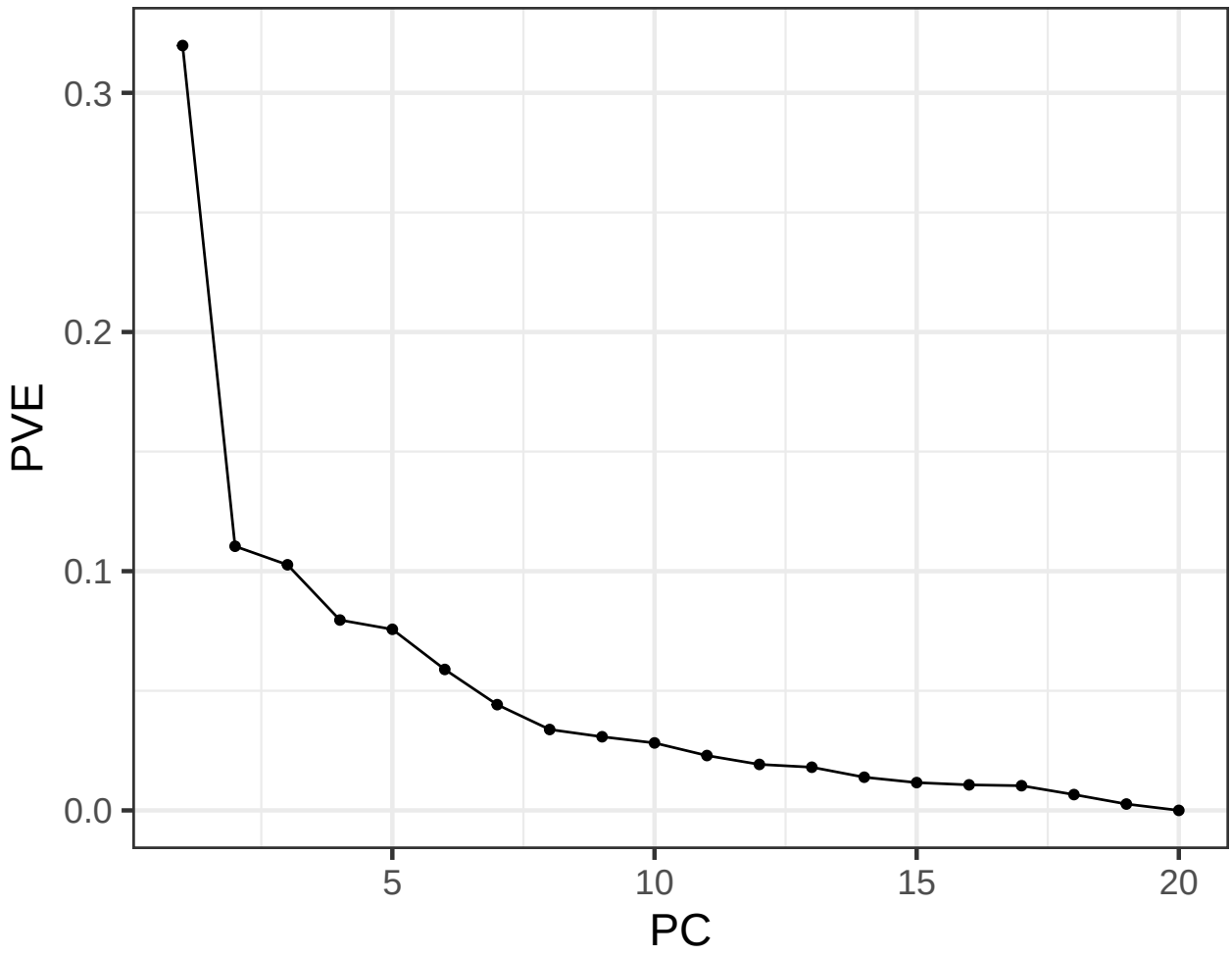
```
ggbiplot(pca1)
```




```
dat <- as.data.frame(pca1$x)
PCAPlot <- ggplot(dat, aes(x = PC1, y = PC2)) +
  geom_point(size = 3) +
  xlim(c(-3, 3)) +
  theme_bw(base_size = 18)
PCAPlot
```

```
## Warning: Removed 33 rows containing missing values (geom_point).
```


5.0
2.5
0.0
−2.5
PC2
−2
0
2
PC1



Scree Plot

Breakdown of Variables and Regions Data

Top Correlated Varaiables: Perceived Corruption (0.8832017), Net Migration (0.8384467), and Industry(0.8170564).

```
#Linear Model to Predict Eastern European Happiness
EastEurolm <- lm(HappinessScore ~ PercievedCorruptionScore + Netmigration +
                 Industry + Coastlinecoastbyarearatio, data = eastern_europe)
```

Adjusted R-squared: 0.9959 p-value: 0.0431

```
#Western Europe Region
western_europe <- filter(happy_country, Region == "WESTERN EUROPE")
western_europe_nocat <- subset(western_europe, select = -c(Region, Country, Climate))
western_europe_cor <- western_europe_nocat %>%
  cor(western_europe_nocat)
```

Top Correlated Variables: Percieved Corruption (0.870394734), GDP per capita (0.854352684), Crops (-0.829523968).

```r
#Linear Model to Predict Western European Happiness
WestEurolm <- lm(HappinessScore ~ PercievedCorruptionScore + GDP_percapita +
                 Literacy + Agriculture + Deathrate, data = western_europe)
```

Adjusted R-squared: 0.8866 p-value: 2.623e-05

```r
#Latin America and Caribbean Region
latin_america_carib <- filter(happy_country, Region == "LATIN AMER. & CARIB")
latin_america_carib_nocat <- subset(latin_america_carib, select = -c(Region, Country, Climate))
latin_america_carib_cor <- latin_america_carib_nocat %>%
  cor(latin_america_carib_nocat)
```

Top Correlated Variables: Phonesper1000 (0.7221507), GDP_percapita (0.6595595), Literacy (0.6108366)

```r
#Linear Model to Predict Latin America and Caribbean Happiness
LACablm <- lm(HappinessScore ~ Phonesper1000 + Deathrate + Crops +
                Arable, data = latin_america_carib)
```

Adjusted R-squared: 0.8094 p-value: 4.992e-06

```r
# Africa Region
happy_country2 <- happy_country
happy_country2$Region[happy_country2$Region == "NORTHERN AFRICA"] <- "AFRICA"
happy_country2$Region[happy_country2$Region == "SUB-SAHARAN AFRICA"] <- "AFRICA"
africa <- filter(happy_country2, Region == "AFRICA")
africa_nocat <- subset(africa, select = -c(Region, Country, Climate))
africa_cor <- africa_nocat %>%
  cor(africa_nocat)
```

Top Correlated Variables: Phonesper1000 (0.53851034), GDP_percapita (0.43636820), Birthrate (-0.43576128).

```r
#Linear Model to Predict Africa Happiness
Africalm <- lm(HappinessScore ~ Phonesper1000 + Birthrate + Crops +
                 PercievedCorruptionScore, data = africa)
```

Adjusted R-squared: 0.3706 p-value: 0.00113

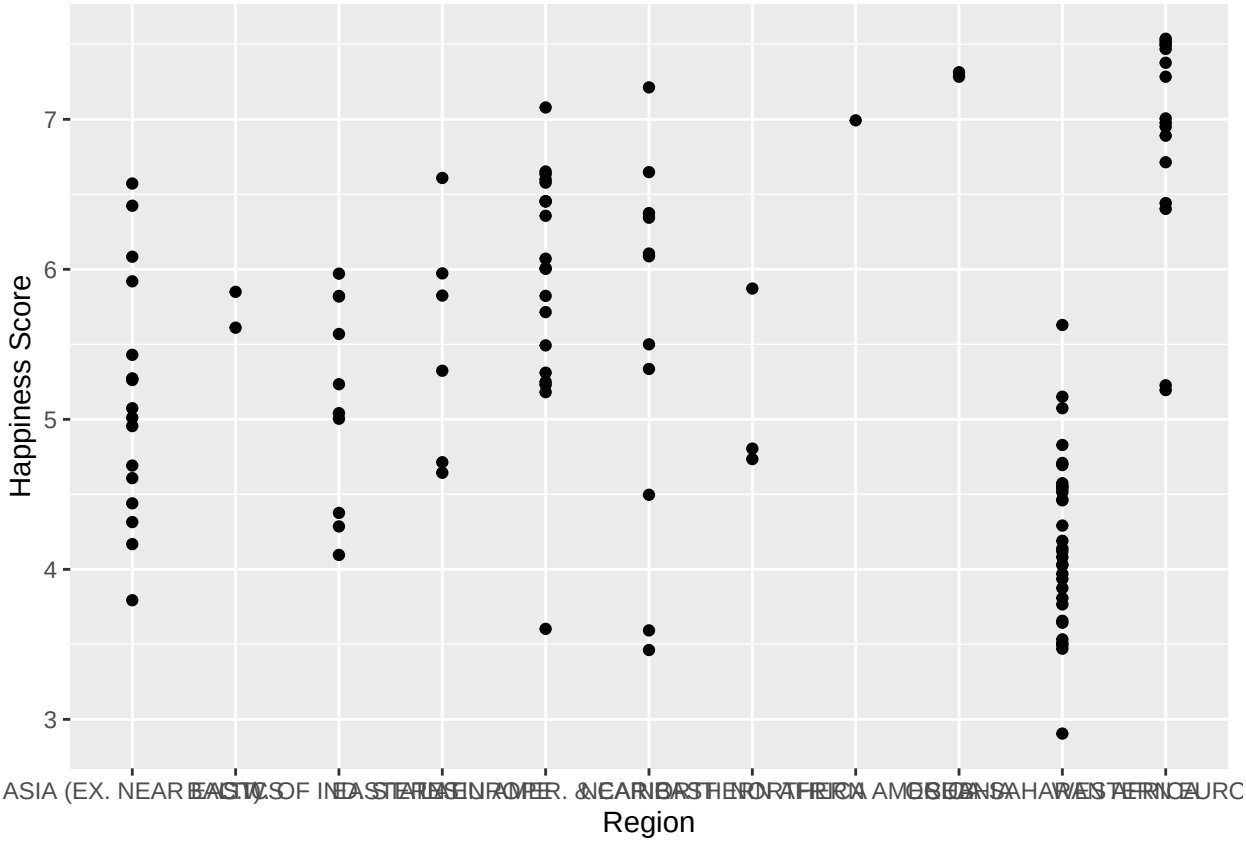
```r
# ASIA (EX. NEAR EAST) Region
asiaNE <- filter(happy_country2, Region == "ASIA (EX. NEAR EAST)")
asiaNE_nocat <- subset(asiaNE, select = -c(Region, Country, Climate))
asiaNE_cor <- asiaNE_nocat %>%
  cor(asiaNE_nocat)
```

Top Correlated Variables: Agriculture (-0.82178348), PercievedCorruptionScore (0.66801346), GDP_percapita (0.66188480).

```r
#Linear Model to Predict ASIA (EX. NEAR EAST) Happiness
asiaNElm <- lm(HappinessScore ~ Agriculture + Phonesper1000 + Birthrate +
                 GDP_percapita, data = asiaNE)
```

Adjusted R-squared: 0.6589 p-value: 0.001531

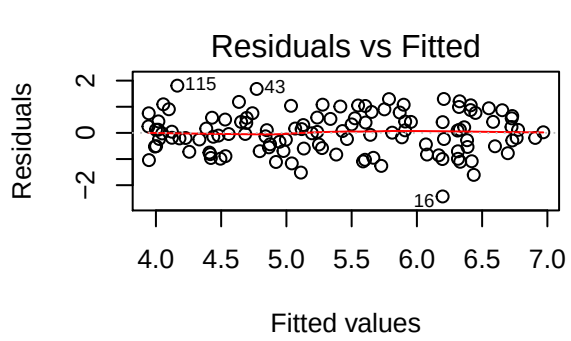
```
#Happiness Score sorted by Region
RegionPlot<- ggplot(happy_country, aes(x = Region, y = HappinessScore)) +
 geom_point() +
 labs(x="Region", y="Happiness Score")
RegionPlot
```

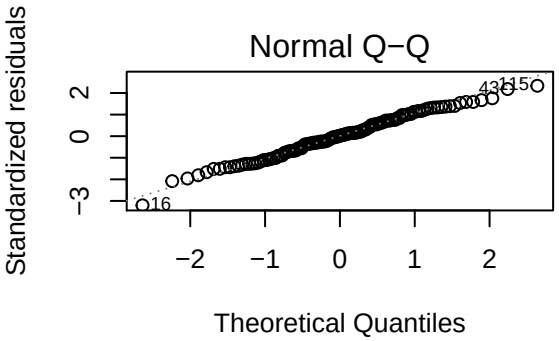


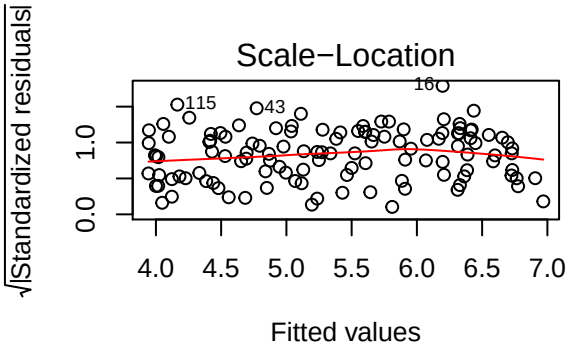
Modeling

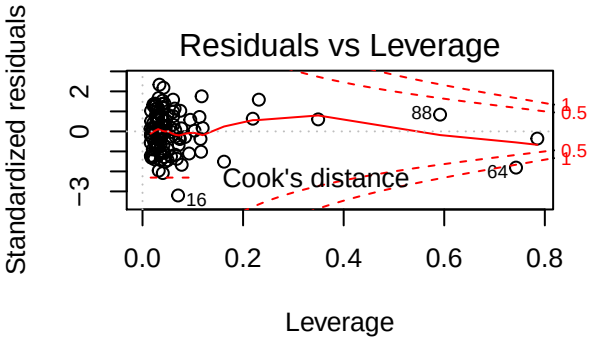
Ridge and Lasso

```
#OLS with interaction
olsmod<- lm(HappinessScore~Agriculture*Industry*Service, data=num_happy)
par(mfrow = c(2, 2))
OLS<- plot(olsmod)
```

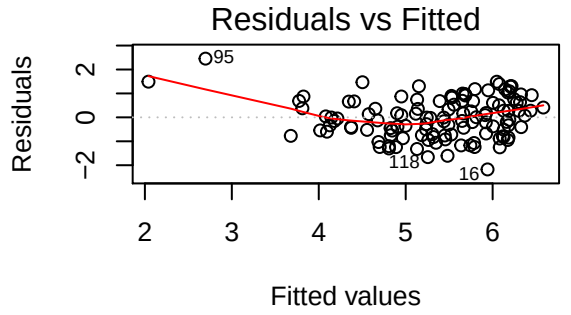


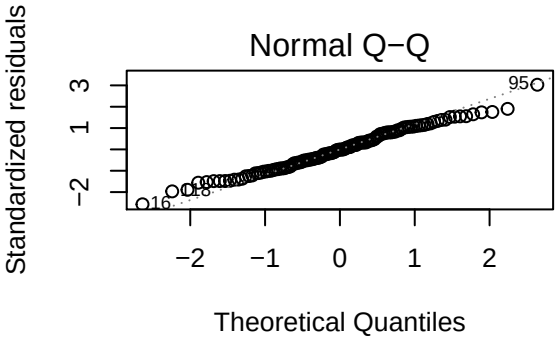


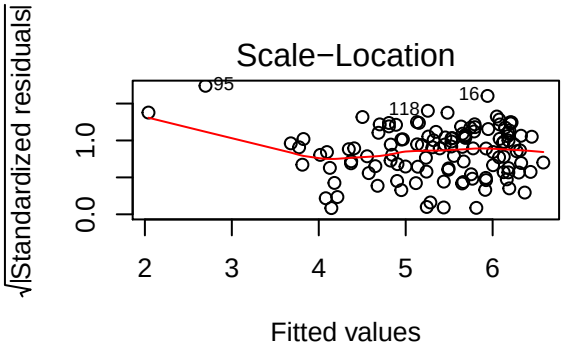


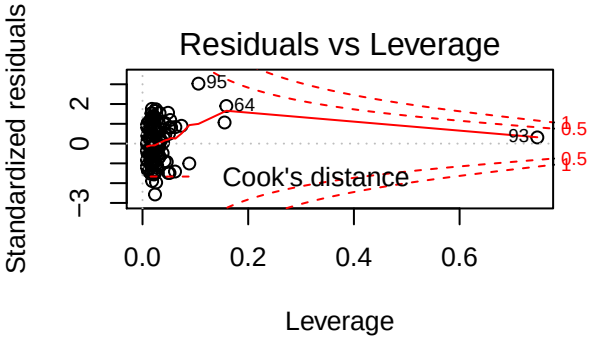



```
#OLS without interaction
olsmod2<- lm(HappinessScore~Agriculture+Industry+Service, data=num_happy)
par(mfrow = c(2, 2))
OLS2<- plot(olsmod2)
```

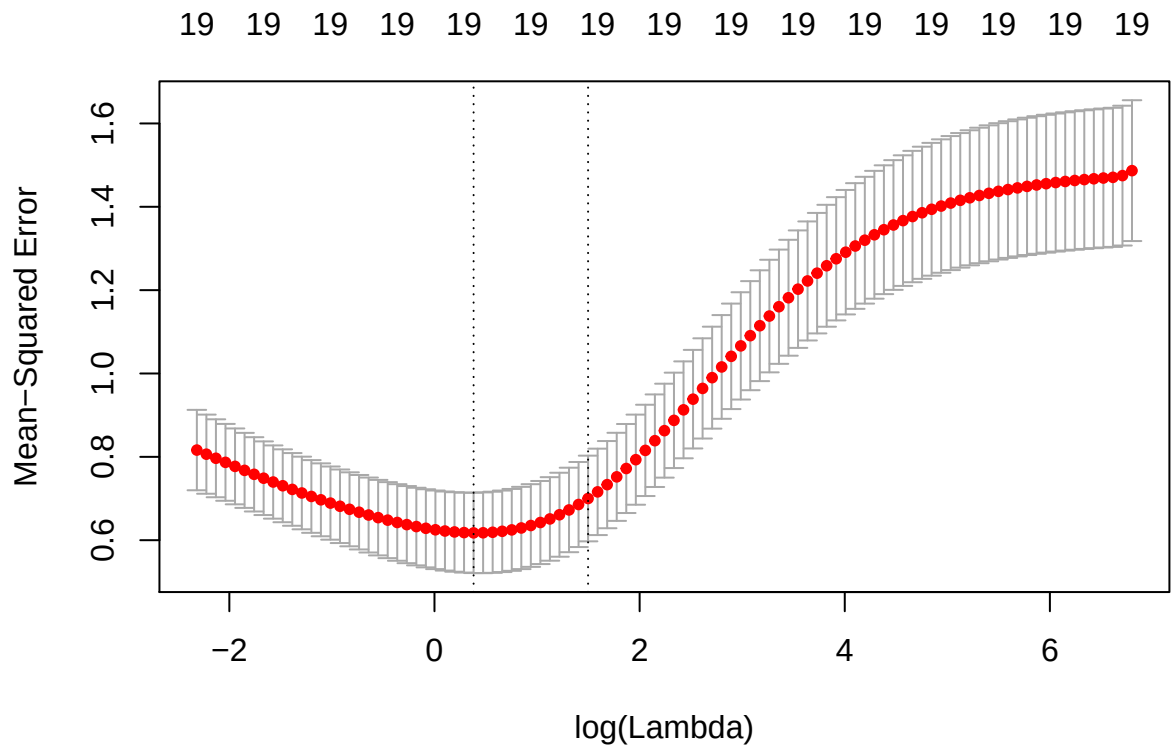




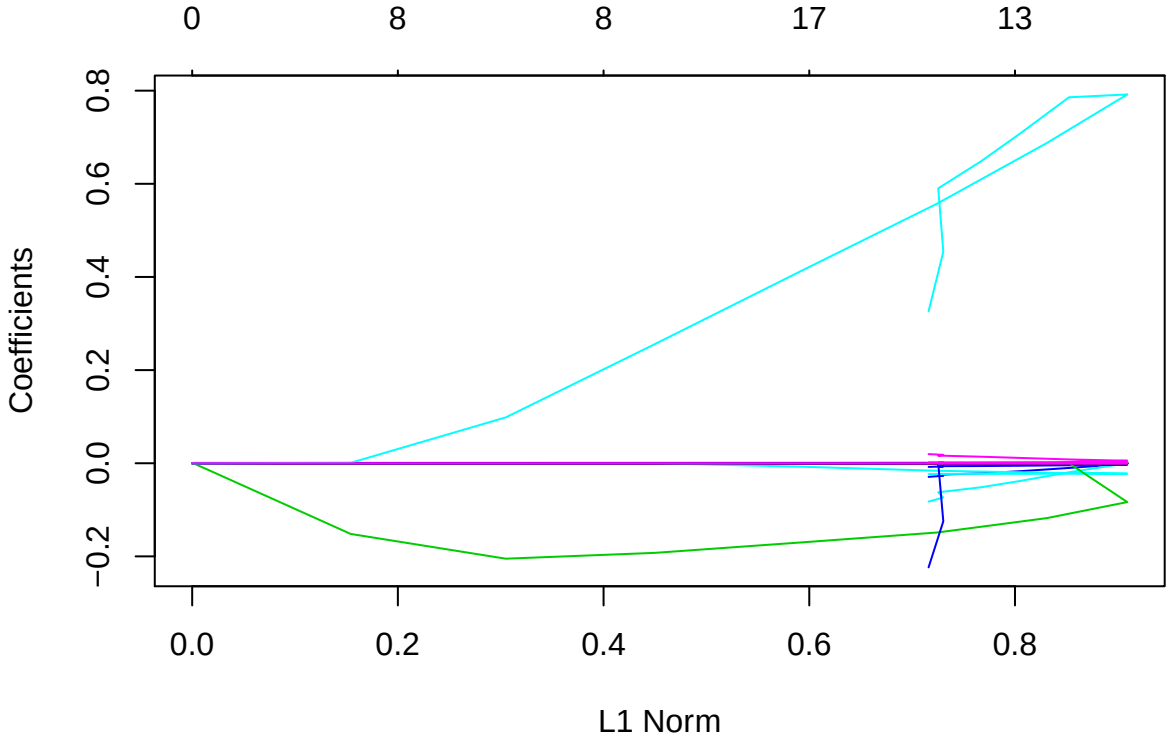




```
cv.out <- cv.glmnet(xnew[trainnew,], ynew[trainnew], alpha = 0)
plot(cv.out)
```

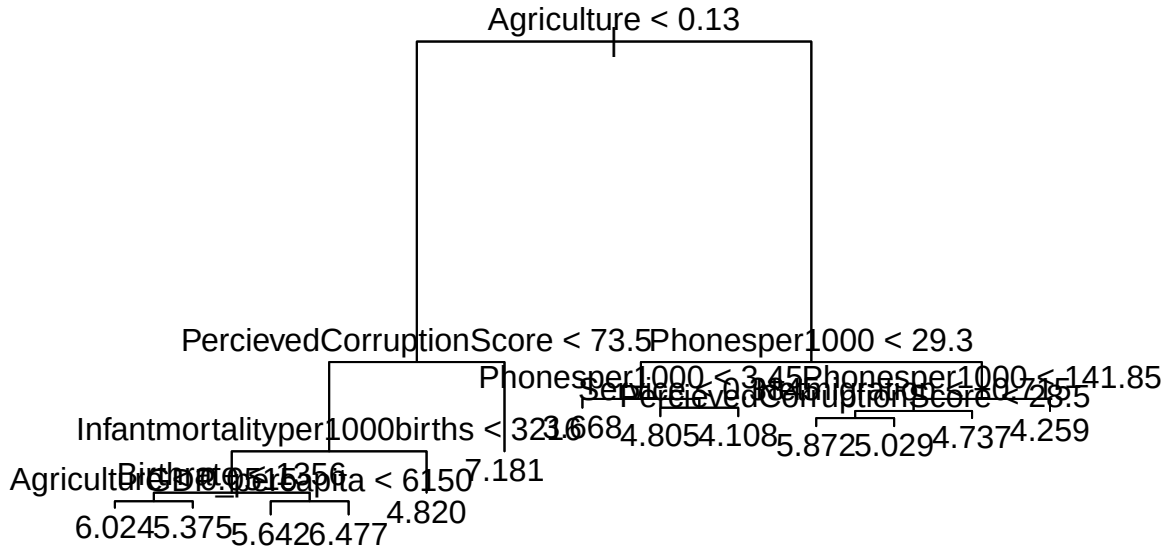



```
plot(lasso.mod)
```

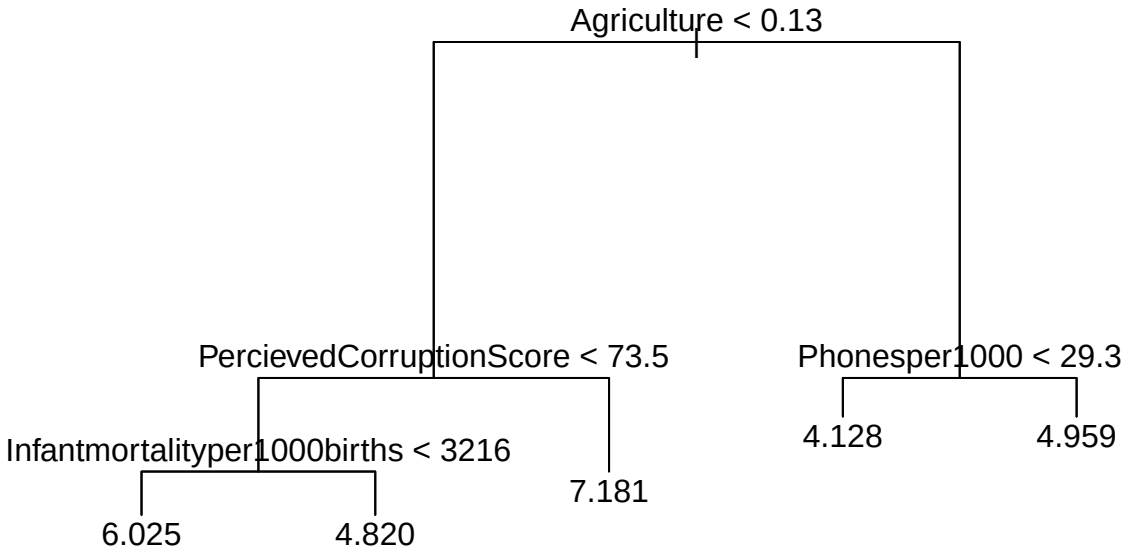


Regression Trees

```
regtree<- tree(HappinessScore ~ .-HappinessScore,data=num_happy)
plot(regtree)
text(regtree, pretty=0)
```




```
tprune <- prune.tree(regtree, best = 5)
plot(tprune)
text(tprune, pretty = 0)
```



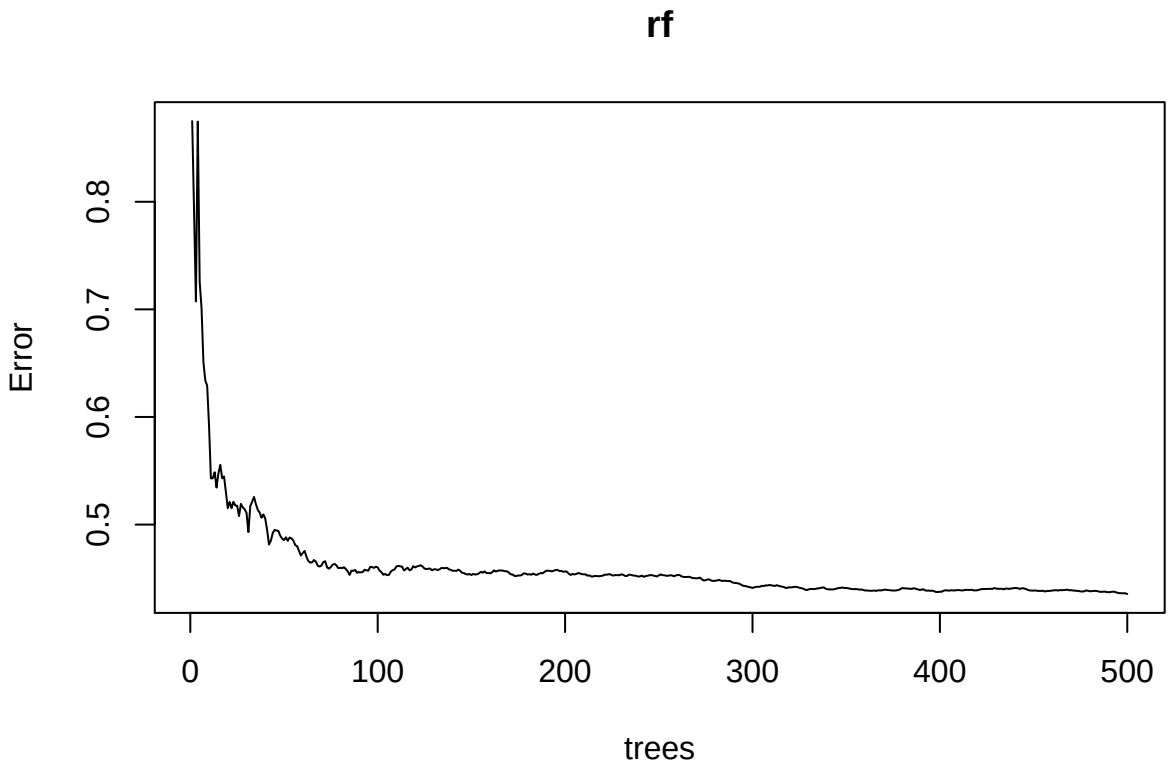

```
tree_est <- predict(tprune, newdata=num_happy)
MSE_test<- mean((tree_est - num_happy$HappinessScore)^2)
MSE_test
```

```
## [1] 0.3164792
```

Boosted Tree ## Boost

Random Forest

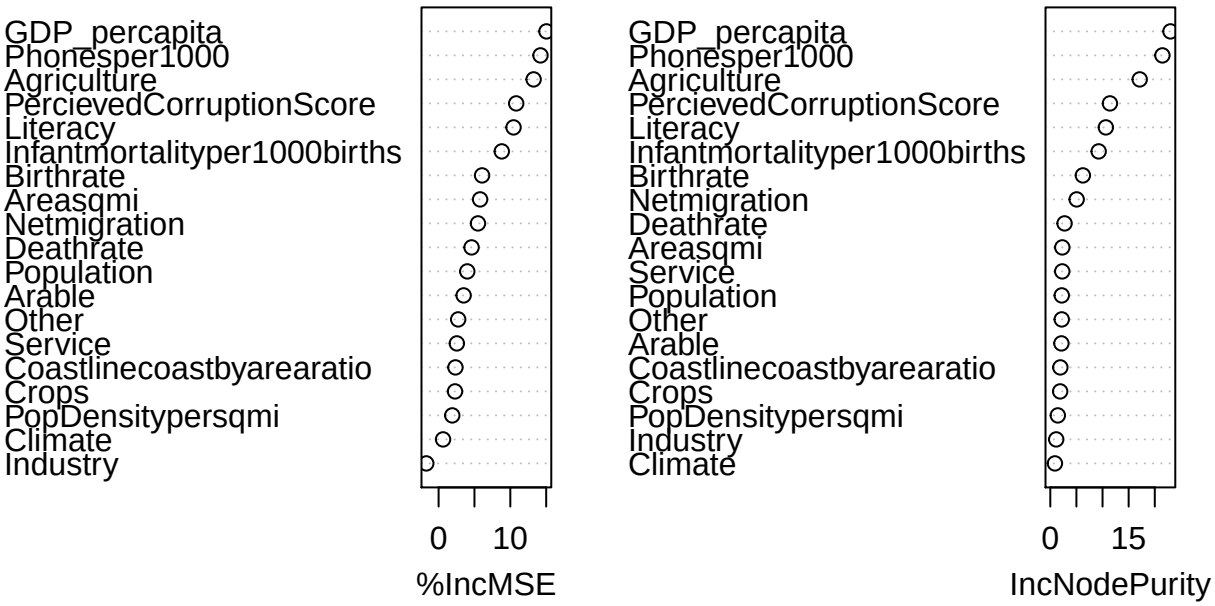
```
rf <- randomForest(HappinessScore ~ .-HappinessScore, data = traind, importance = TRUE)
plot(rf)
```




```
testrf <- predict(rf, newdata=testd)
MSE2<- mean((testrf - testd$HappinessScore)^2)
MSE2
```

```
## [1] 0.4458951
```

```
varImpPlot(rf)
```



LINEAR MODELS

```
#Region Linear Model
alymod<- lm(HappinessScore ~ Region, data = happy_country)
```

Adjusted R-squared: 0.5754

```
#GDP + Arable Land + Infant Mortality + Percieved Corruption
alymod0 <- lm(HappinessScore ~ GDP_percapita + Arable +
                Infantmortalityper1000births + PercievedCorruptionScore,
              data = happy_country)
```

Adjusted R-squared: 0.6734

```
#Region + GDP + Infant Mortality + Percieved Corruption
alymod2 <- lm(HappinessScore ~ Region + GDP_percapita +
                Infantmortalityper1000births + PercievedCorruptionScore,
              data = happy_country)
```

Adjusted R-squared: 0.7558

```
#Region + GDP + Arable Land + Percieved Corruption
alymod3 <- lm(HappinessScore ~ Region + GDP_percapita + Arable +
                PercievedCorruptionScore, data = happy_country)
```

Adjusted R-squared: 0.7590

```
#Region + GDP + Arable Land + Infant Mortality
alymod4 <- lm(HappinessScore ~ Region + GDP_percapita + Arable +
                Infantmortalityper1000births, data = happy_country)
```

Adjusted R-squared: 0.7627

```
#Region + GDP + Arable Land + Infant Mortality + Percieved Corruption
alymod5 <- lm(HappinessScore ~ Region + GDP_percapita + Arable +
              Infantmortalityper1000births + PercievedCorruptionScore,
              data = happy_country)
```

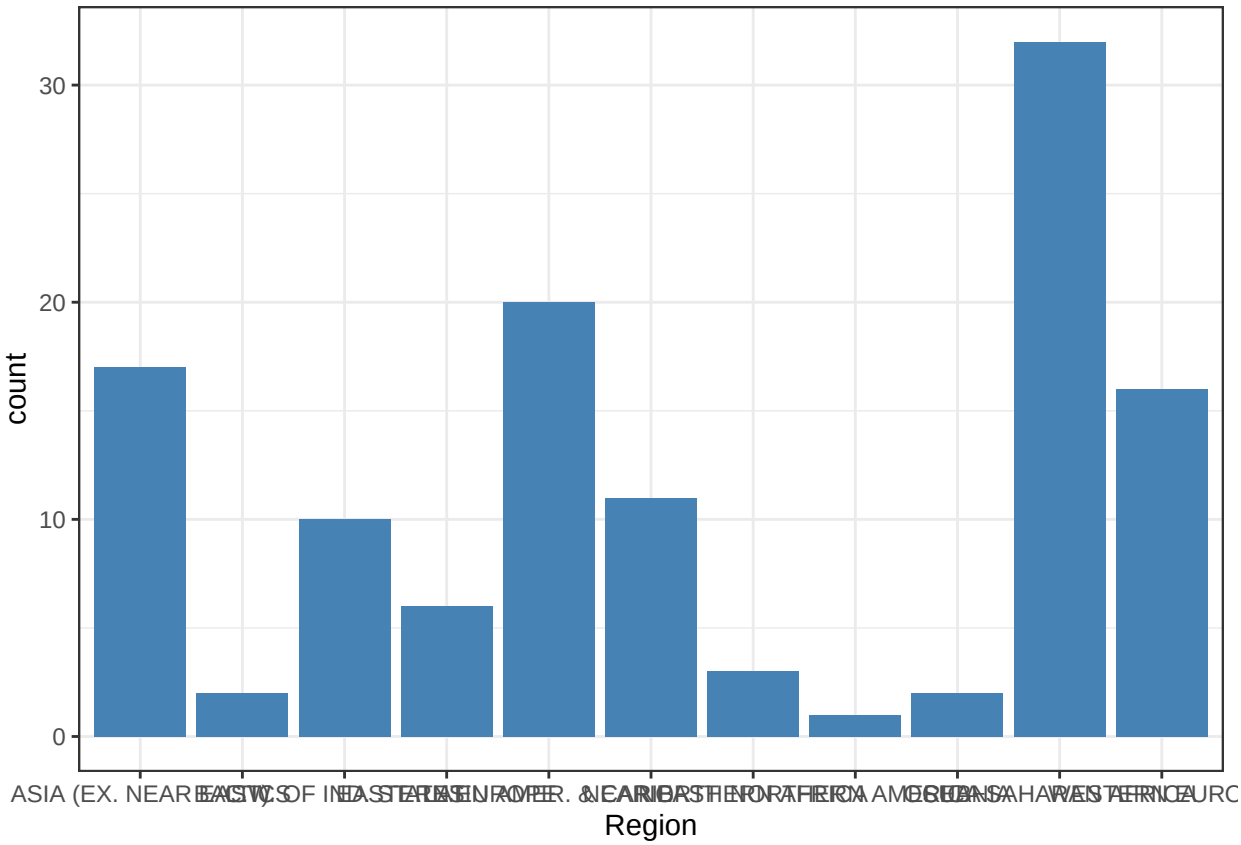
Adjusted R-squared: 0.7644

```
#Region + GDP + Arable Land + Infant Mortality + Percieved Corruption + Coastline Area
alymod6 <- lm(HappinessScore ~ Region + GDP_percapita + Arable +
              Infantmortalityper1000births + PercievedCorruptionScore +
                Coastlinecoastbyarearatio, data = happy_country)
```

Adjusted R-squared: 0.7648

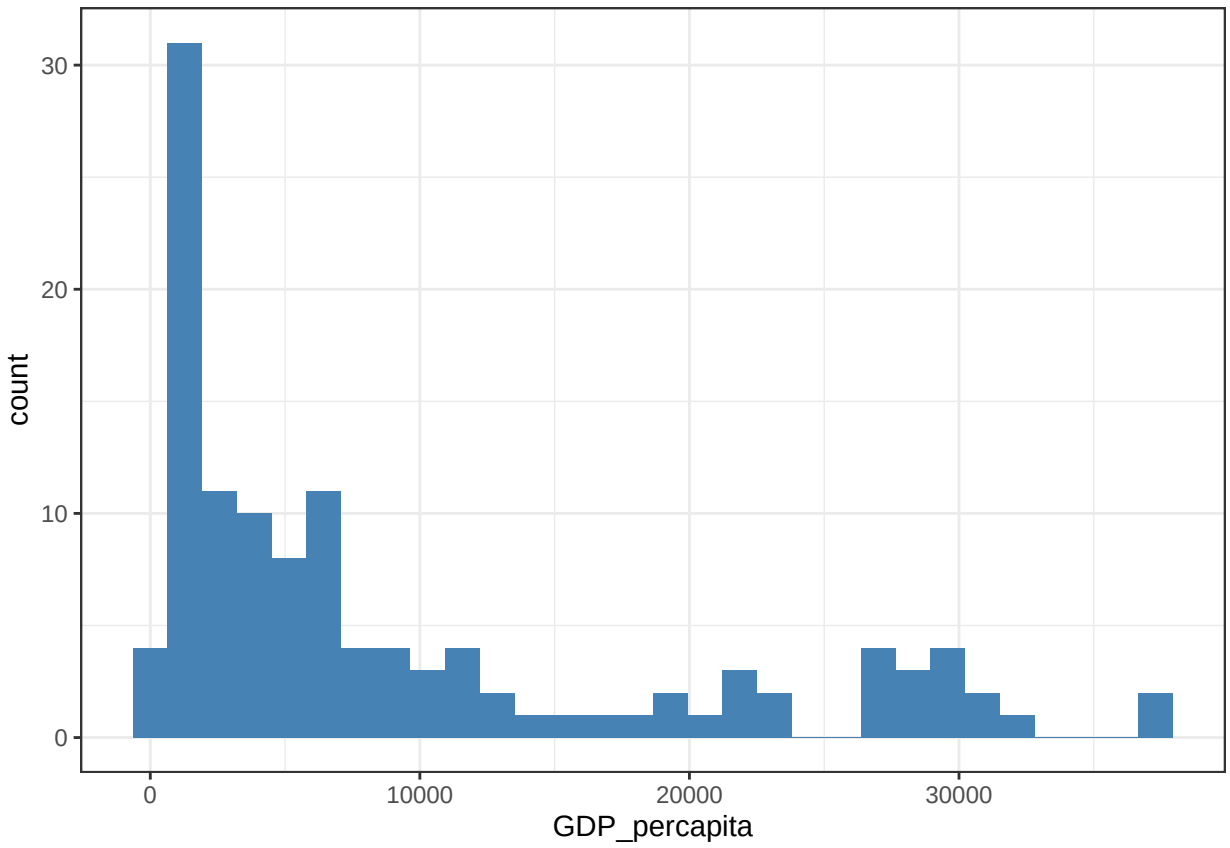
```
#Normality of Region
ggplot(happy_country, aes(x = Region)) +
geom_histogram(fill = "steelblue", stat="count") +
theme_bw()
```

```
## Warning: Ignoring unknown parameters: binwidth, bins, pad
```

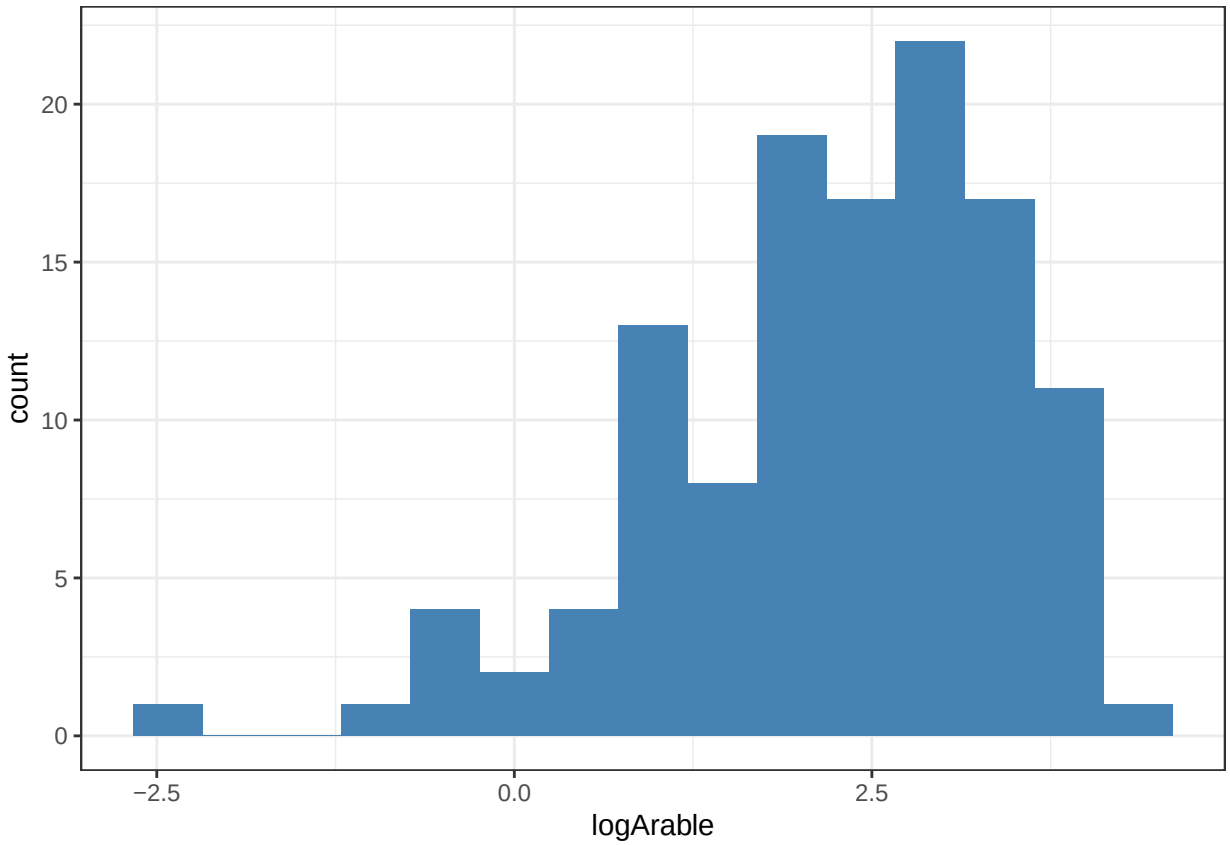


```r
#Normality of GDP per capita
logGDP <- log(happy_country$GDP_percapita)
ggplot(happy_country, aes(x = GDP_percapita)) +
geom_histogram(fill = "steelblue") +
theme_bw()
```

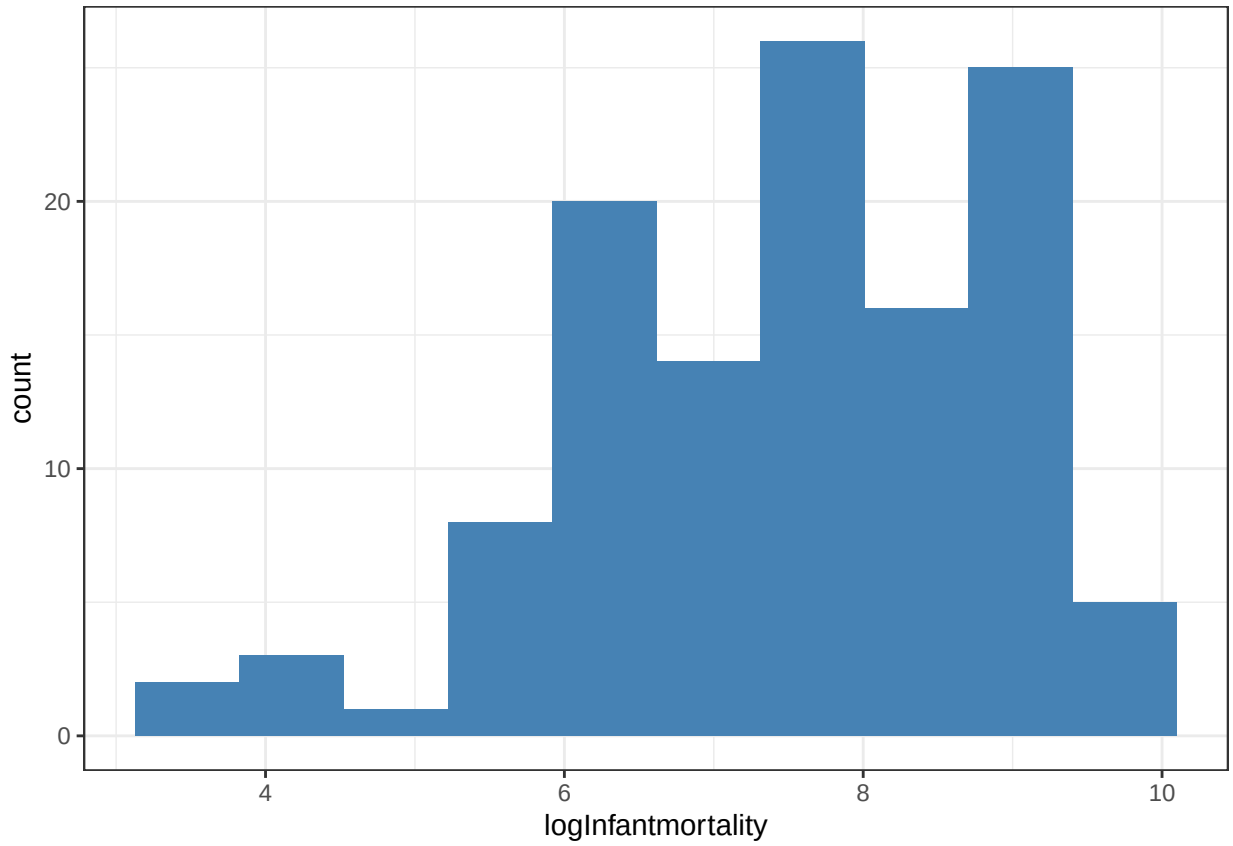
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



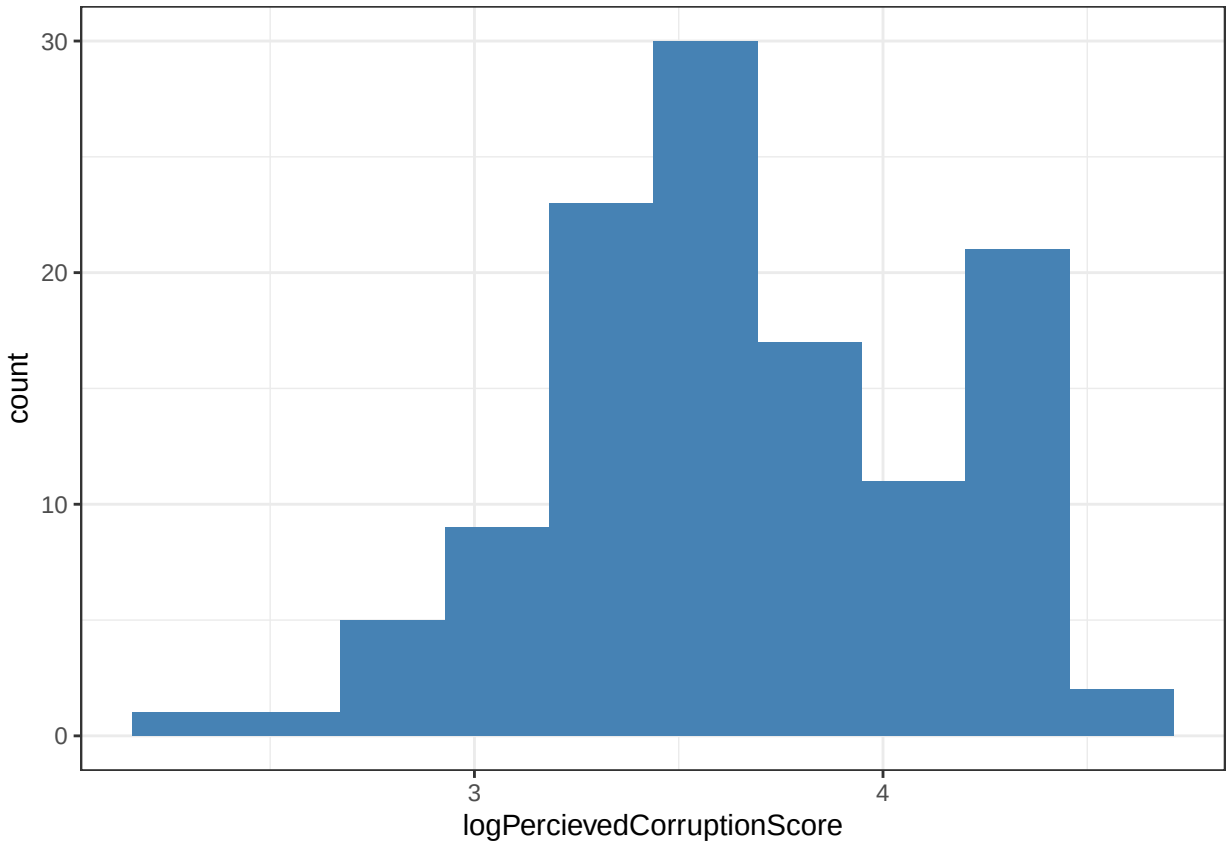

```r
#Normality of Arable
logArable <- log(happy_country$Arable)
ggplot(happy_country, aes(x = logArable)) +
geom_histogram(fill = "steelblue", bins = "15") +
theme_bw()
```




```
#Normality of Infantmortalityper1000births
logInfantmortality <- log(happy_country$Infantmortalityper1000births)
ggplot(happy_country, aes(x = logInfantmortality)) +
geom_histogram(fill = "steelblue", bins = "10") +
theme_bw()
```

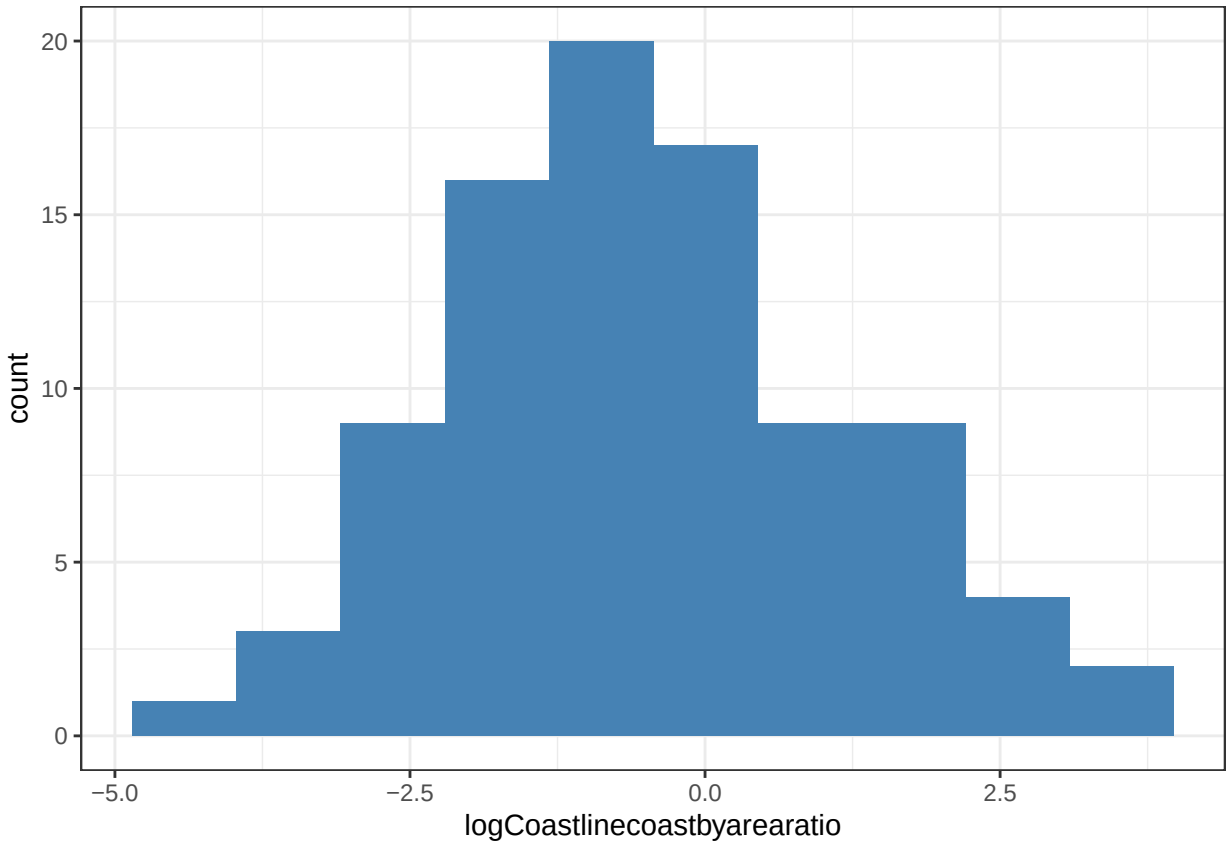



```
#Normality of PercievedCorruptionScore
logPercievedCorruptionScore <- log(happy_country$PercievedCorruptionScore)
ggplot(happy_country, aes(x = logPercievedCorruptionScore)) +
geom_histogram(fill = "steelblue", bins = "10") +
theme_bw()
```




```
#Normality of Coastlinecoastbyarearatio
logCoastlinecoastbyarearatio <- log(happy_country$Coastlinecoastbyarearatio)
lnCoastlinecoastbyarearatio <- log1p(happy_country$Coastlinecoastbyarearatio)
ggplot(happy_country, aes(x = logCoastlinecoastbyarearatio)) +
geom_histogram(fill = "steelblue", bins = "10") +
theme_bw()
```

```
## Warning: Removed 30 rows containing non-finite values (stat_bin).
```




```
#Region + GDP + log Arable Land + log Infant Mortality + log Percieved Corruption + Coastline Area
alymod7 <- lm(HappinessScore ~ Region + GDP_percapita +
                log(Arable) + log(Infantmortalityper1000births) +
                log(PercievedCorruptionScore) + Coastlinecoastbyarearatio,
              data = happy_country)
```

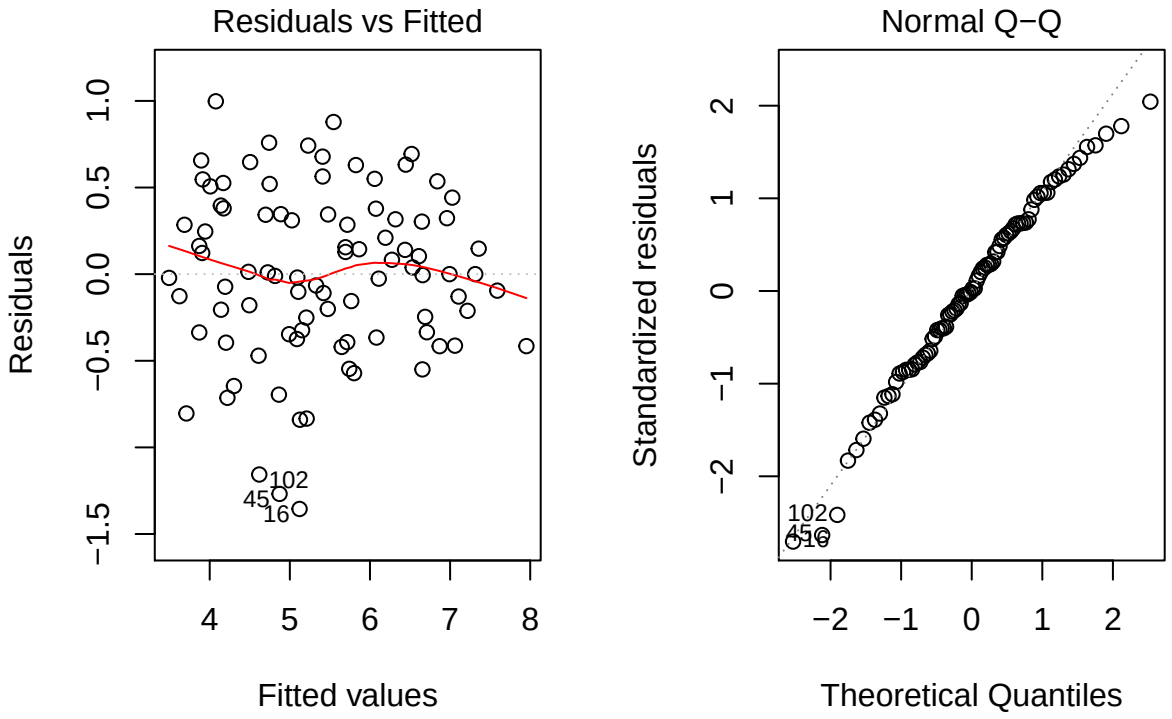
```
set.seed(1)
trainS <- sample(1:nrow(happy_country), nrow(happy_country) * .75)
trainD <- happy_country[trainS, ]
testD <- happy_country[-trainS, ]
```

```
#Train 1
linearmodel1 <- lm(HappinessScore ~ Region + GDP_percapita + Arable +
                Infantmortalityper1000births + PercievedCorruptionScore +
                  Coastlinecoastbyarearatio, data = trainD)
```

Adjusted R-squared: 0.8061

```
#Residuals 1
par(mfrow = c(1, 2))
plot(linearmodel1, 1:2)
```

```
## Warning: not plotting observations with leverage one:
##   13, 68
```

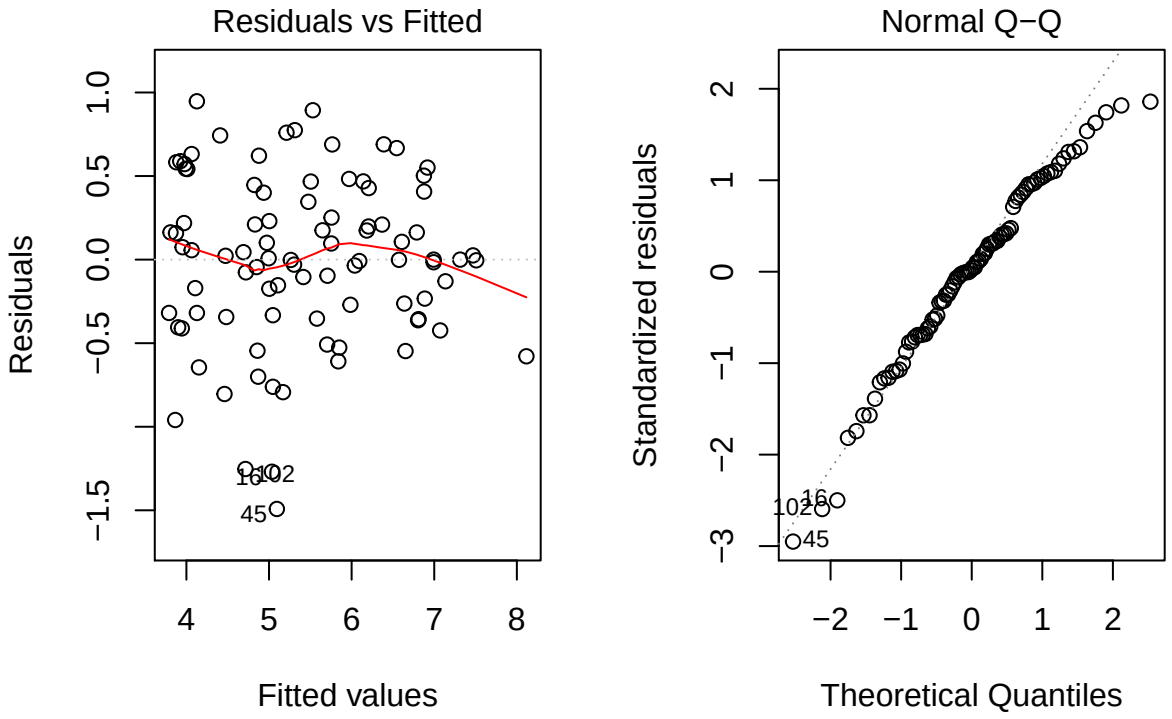



```
#Train 2 with the Log variables
linearmodel2 <- lm(HappinessScore ~ Region + GDP_percapita +
                log(Arable) + log(Infantmortalityper1000births) +
                log(PercievedCorruptionScore) + Coastlinecoastbyarearatio,
                data = trainD)
```

Adjusted R-squared: 0.7921

```
#Residuals 2
par(mfrow = c(1, 2))
plot(linearmodel2, 1:2)
```

```
## Warning: not plotting observations with leverage one:
##   13, 68
```



Discussion

References

Arafa, S. (2019, April 5). Why Governments Should Care More about Happiness. Greater Good. https://greatergood.berkeley.edu/article/item/why_governments_should_care_more_about_happiness De Stasio, S., Fiorilli, C., Benevene, P., Boldrini, F., Ragni, B., Pepe, A., & Maldonado Briegas, J. J. (2019). Subjective Happiness and Compassion Are Enough to Increase Teachers' Work Engagement? Frontiers in Psychology, 10. https://doi.org/10.3389/fpsyg.2019.02268 e.V, T. I. (n.d.). Corruption Perceptions Index 2017. Retrieved December 4, 2019, from Www.transparency.org website Lasso, Fernando. "Countries of the World Data (World Factbook US Government)." Erasmus University, 26 Apr. 2018. "World Happiness Report (Gallup World Poll)." Sustainable Development Solutions Network Updates, 28 Feb. 2017.