What Makes People Happy?

Keir Havel

Introduction

My research questions are: what variables can be used to predict a country's subjective happiness best, and what distinguishes happy countries from unhappy countries? To answer these questions, I'll be looking at PCA and correlation matrices of the data, as well as fitting linear models, k-means clustering, kNN, and LDA+QDA analyses.

The attributes of the dataset are GDP (PPP,per capita), energy consumption(per capita), CO2 emissions (per capita), a democracy index (Economist Intelligence Unit), education expenditure (per capita), subjective happiness (UN WHR), health expenditure (per capita), homicide rates, journal articles published, population density, military expenditure (relative to GDP), index of economic freedom (WSJ), research expenditure, suicide rate, undernourishment (population percent), and woman's empowerment. These attributes were chosen to provide an estimate of country's material wealth, scientific development, social development, crime, and natural resources. There were 200 observations in the full compiled dataset, with many missing data-points due to inconsistencies in countries reported and inconsistencies in labeling (e.g. are Chinese SARs included under China or are they treated as independent entities).

Exploratory Analysis

PCA and Correlation

The data had some repeated rows and a non-numeric column, so I deleted those to prepare for the PCA.

```
data = read.table('/home/keir/Downloads/PSTAT/FINALPROJECT/data/Data.csv',quote = "", sep = ',', fill =
data1 = data[-1]
data1 = data1[-29,]
X = as.matrix(data1)
```

Next, I made measures for the center and dispersion of the data, and used those to perform a PCA.

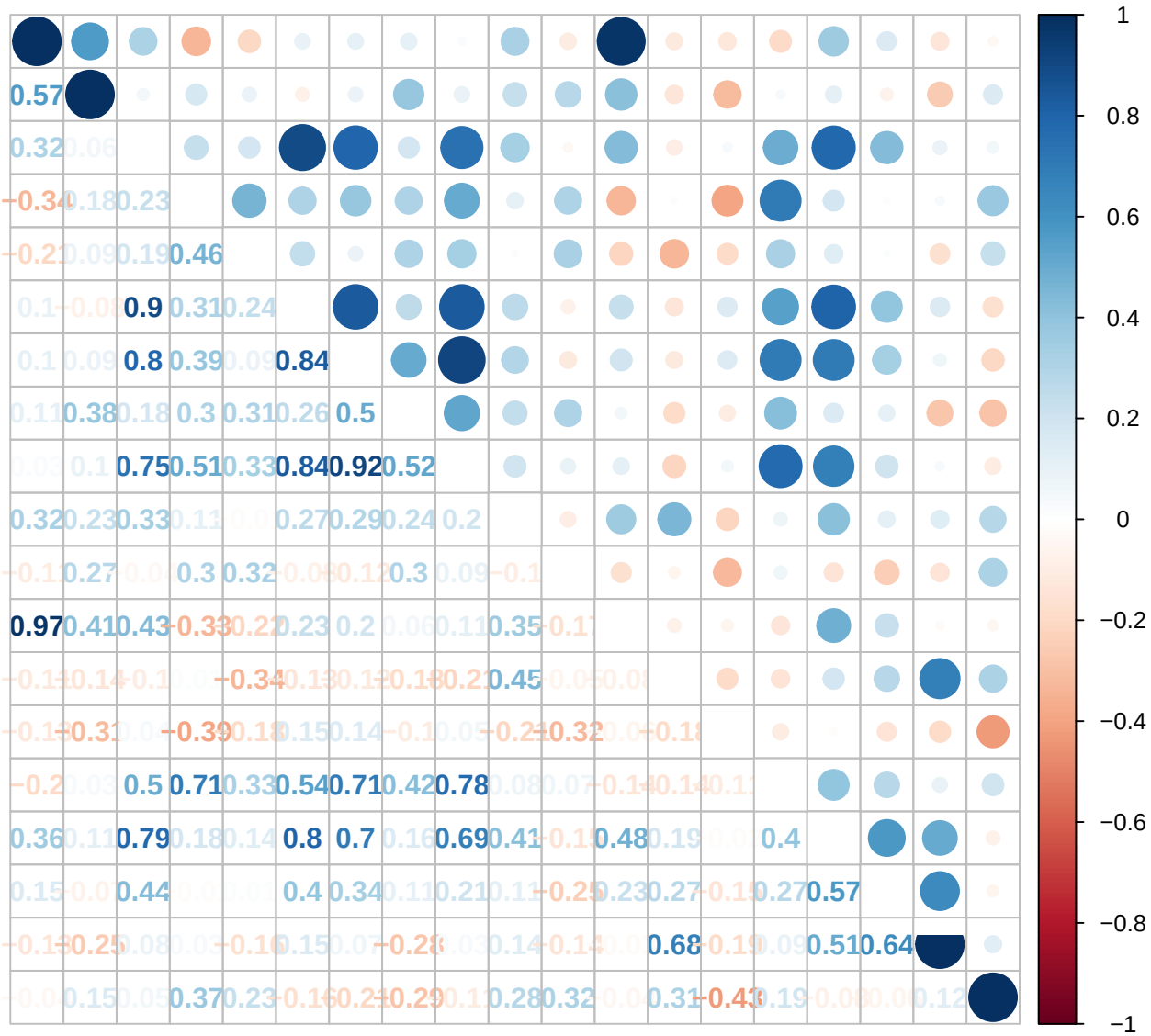
```
centr <- apply(X,2,mean, na.rm=T)
disp <- apply(X,2,sd, na.rm=T)
pca = prcomp(~.,data=data1, center = centr, scale = disp, na.action = na.omit)
summary(pca)
```

```
## Importance of components:
##                           PC1    PC2    PC3     PC4     PC5     PC6
## Standard deviation     3.6414 1.9375 1.6657 1.17502 0.98824 0.96462
## Proportion of Variance 0.5251 0.1487 0.1099 0.05468 0.03868 0.03685
## Cumulative Proportion  0.5251 0.6738 0.7837 0.83836 0.87703 0.91388
##                           PC7    PC8     PC9    PC10    PC11    PC12
## Standard deviation     0.78035 0.7212 0.62777 0.48569 0.44632 0.37159
## Proportion of Variance 0.02412 0.0206 0.01561 0.00934 0.00789 0.00547
## Cumulative Proportion  0.93800 0.9586 0.97421 0.98355 0.99144 0.99691
##                           PC13    PC14   PC15    PC16    PC17    PC18
```

```
## Standard deviation     0.20351 0.11566 0.1010 0.07604 0.06065 0.05354
## Proportion of Variance 0.00164 0.00053 0.0004 0.00023 0.00015 0.00011
## Cumulative Proportion  0.99855 0.99908 0.9995 0.99971 0.99985 0.99997
##                            PC19
## Standard deviation     0.02841
## Proportion of Variance 0.00003
## Cumulative Proportion  1.00000
```

From the PCA, 6 principle components explain >90% of the variance, and 8 explain >95%. This analysis shows that food and journal publications both have a large portion of the data's variance. To further elucidate relationships among the variables, I'll look at the correlation matrix.

```
library(corrplot)
corelate = cor(data1, use = "complete")
corelate =
corrplot.mixed(corelate, tl.pos = "n")
```



From correlation matrix, a few high (>0.75) correlations exist. Food and Journal articles, emissions and energy consumption and GDP and research, energy consumption and health expenditure and research, health

expenditure and economic freedom. Thus these 6 variables will be reduced to two in the linear models: Food and energy consumption. Furthermore, since happiness is the major variable of interest, all countries with no happiness data will be deleted; additionally countries and variables with high numbers of missing values will be deleted.

```
## Loading required package: Rcpp
```

```
## mice 2.25 2015-11-09
```

An omitted line (it was long and messy) above used the mice package, previous PCA, and correlation analysis to reduce the data.

Basic models for more exploration

To learn a little more about the (reduced) data, a simple linear model was fit.

```
model = lm(Happiness ~.,data=data3, na.action = na.omit)
summary(model)
```

```
##
## Call:
## lm(formula = Happiness ~ ., data = data3, na.action = na.omit)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.74452 -0.44137  0.06628  0.48972  1.36680
##
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)           3.150e+00  4.041e-01   7.796 1.20e-11 ***
## Food                 -9.460e-10  1.448e-09  -0.653  0.51522
## Biodiversity          6.573e-03  4.394e-03   1.496  0.13831
## Democracy             2.536e-01  5.393e-02   4.702 9.49e-06 ***
## Energy.Consum         3.439e-03  6.029e-04   5.705 1.54e-07 ***
## High.Tech.Exports     2.633e-02  9.849e-03   2.674  0.00894 **
## Homicide.Rates        7.994e-03  6.116e-03   1.307  0.19460
## Population.Density   -2.413e-04  9.847e-05  -2.450  0.01625 *
## Military.Expenditure  9.217e-02  5.185e-02   1.778  0.07891 .
## Suicide.Rate         -5.261e-03  1.186e-02  -0.443  0.65850
## Women.Empowerment     3.731e-01  6.303e-01   0.592  0.55541
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.71 on 88 degrees of freedom
##   (59 observations deleted due to missingness)
## Multiple R-squared:  0.6403, Adjusted R-squared:  0.5994
## F-statistic: 15.66 on 10 and 88 DF,  p-value: 9.953e-16
```

From this, it appears that the democracy index, energy consumption, high-tech exports, and population density are the most significant predictors of happiness.

Next, another PCA will find the variance in the reduced data, which furthermore had all its NA values removed.

```
newdata = na.omit(data2)
newdata1 = newdata[-1]
newdata1 = newdata1[-5]
X1 = as.matrix(newdata1)
centr1 <- apply(X1,2,mean, na.rm=T)
disp1 <- apply(X1,2,sd, na.rm=T)
pca1 = prcomp(~.,data=newdata1,center = centr1, scale = disp1)
summary(pca1)
```

```
## Importance of components:
##                           PC1    PC2    PC3    PC4    PC5     PC6     PC7
## Standard deviation     1.5206 1.2844 1.2661 1.0354 1.0093 0.84166 0.74919
## Proportion of Variance 0.2312 0.1650 0.1603 0.1072 0.1019 0.07084 0.05613
## Cumulative Proportion  0.2312 0.3962 0.5565 0.6637 0.7656 0.83641 0.89254
##                           PC8     PC9    PC10
## Standard deviation     0.65006 0.61532 0.52287
## Proportion of Variance 0.04226 0.03786 0.02734
## Cumulative Proportion  0.93480 0.97266 1.00000
```

Unlike the first PCA, this analysis had more diverse loadings (more equally representation across the attributes) and was more easy to interpret. The first PC seemed to be related to degree of democracy (eg high loading on democracy, women's empowerment); the second was more highly loaded on military, population density, (negatively) on food production and biodiversity, which suggests some relationship with population size.

```
DemocraticDevelopment = pca1$x[,1]
Population = -pca1$x[,2]
Happiness = newdata$Happiness
Democracy = newdata$Democracy
EnergyConsumption = newdata$Energy.Consum
Food = newdata$Food
HomicideRates = newdata$Homicide.Rates
```

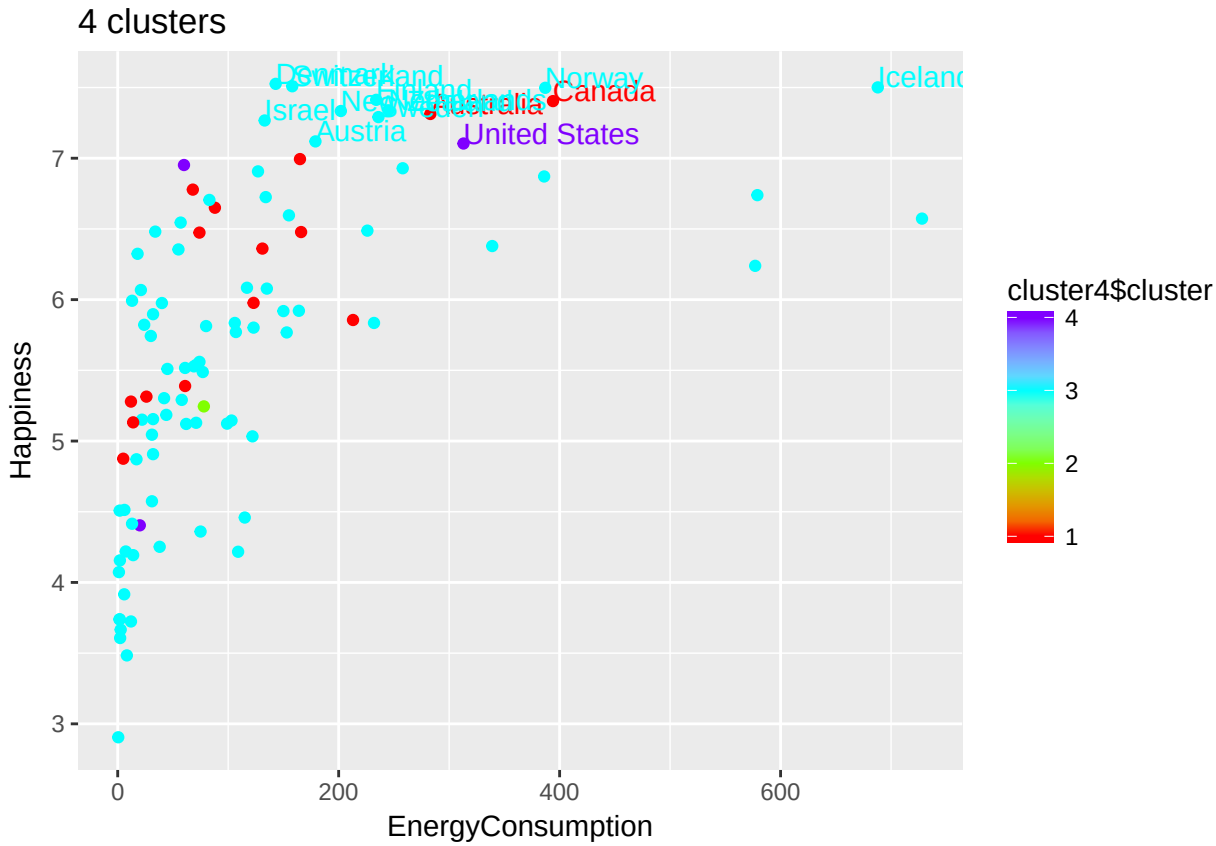
The following are graphs of the data fitted on to the first two principle components, colored by Happiness. The first graph highlights the most happy countries, the second the least.

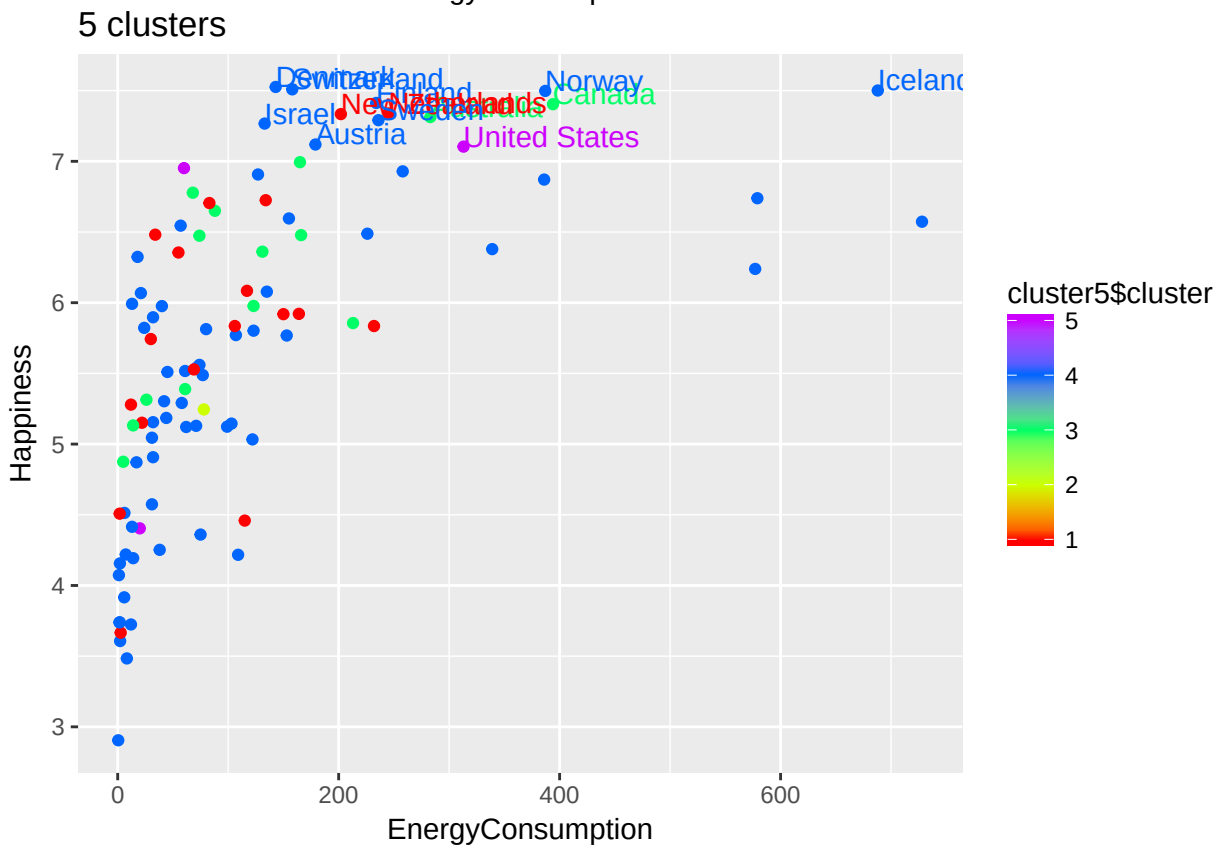
From these graphs, it appears more democratic countries are more happy.

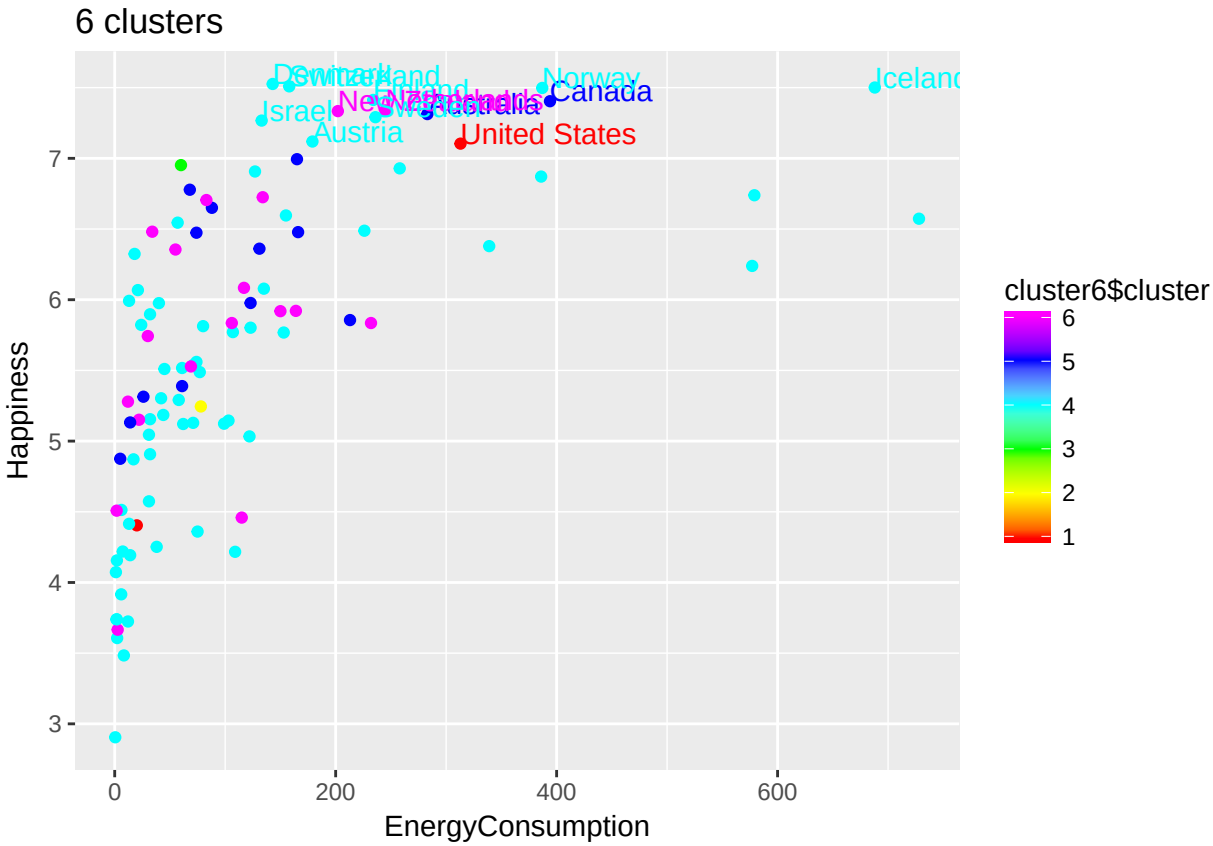
K-means clustering

The following plots show plots of energy consumption and happiness, colored by their cluster. These two attributes were chosen for their general significance to the data, as shown by the PCAs and correlation matrix. The clusters have no obvious visual relationship, but further analysis will try and elucidate them.

```
cluster5 = kmeans(newdata[,-1],5)

cluster4 = kmeans(newdata[,-1],4)

cluster6 = kmeans(newdata[,-1],6)
```


4 clusters
Norway
Canada
United States
Austria
Israel
Happiness
EnergyConsumption
0
200
400
600
3
4
5
6
7
cluster4$cluster
4
3
2
1

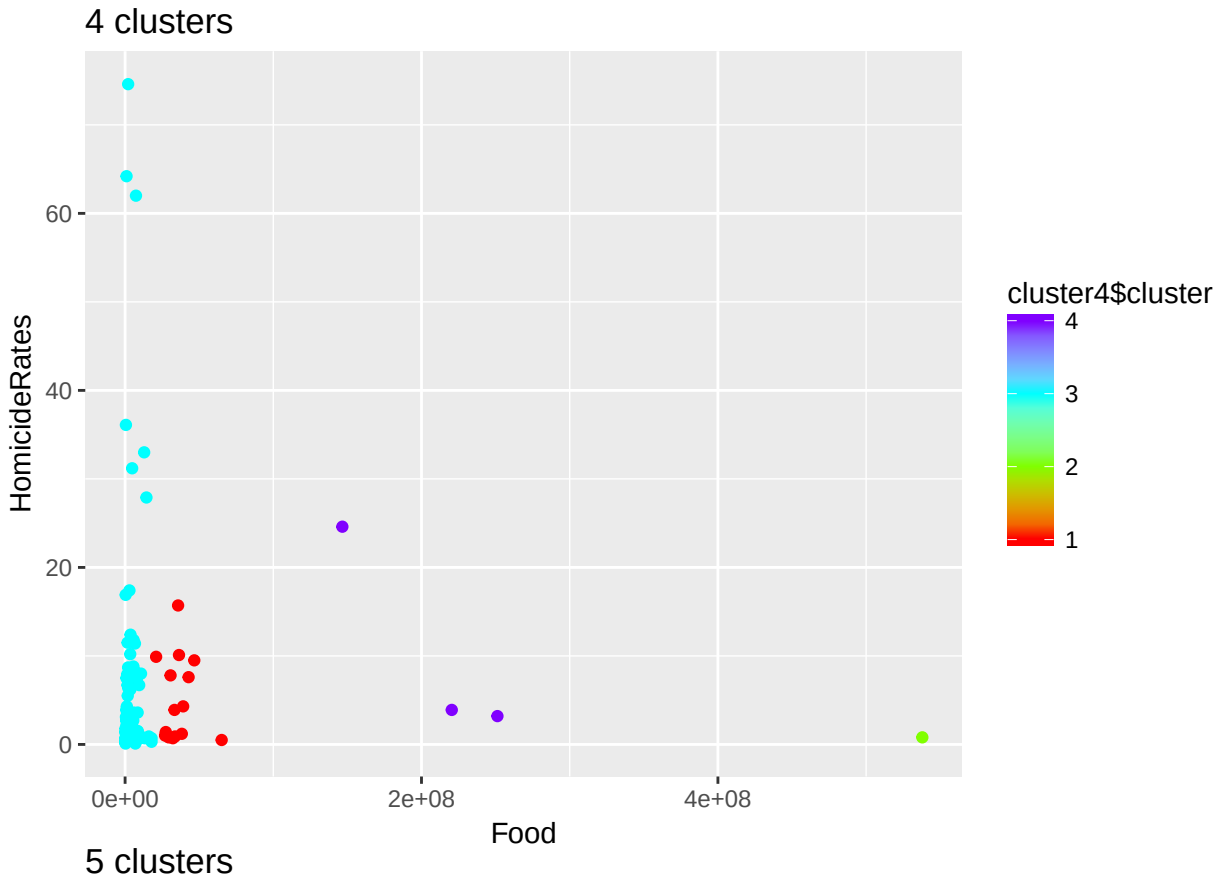

5 clusters
Norway
Canada
United States
Austria
Israel
Happiness
EnergyConsumption
0
200
400
600
3
4
5
6
7
cluster5$cluster
5
4
3
2
1

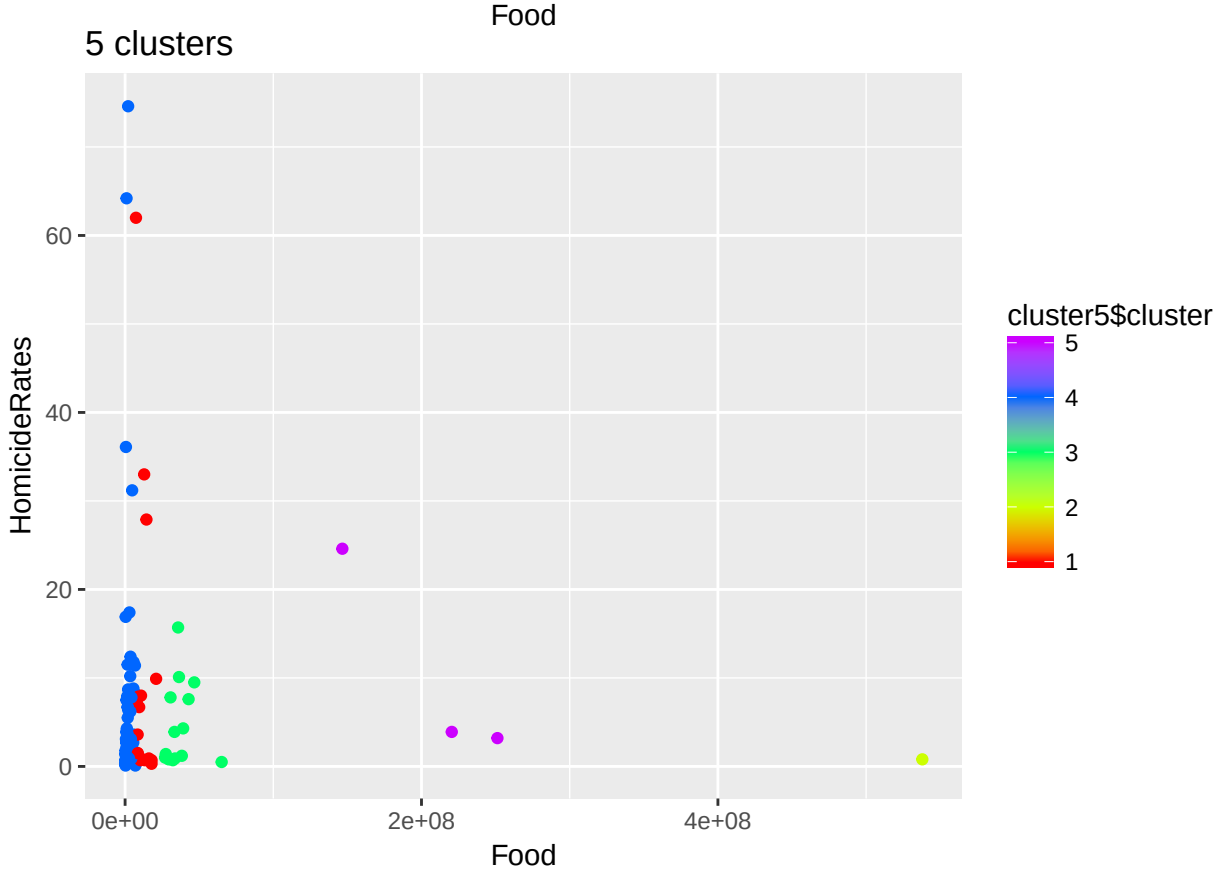


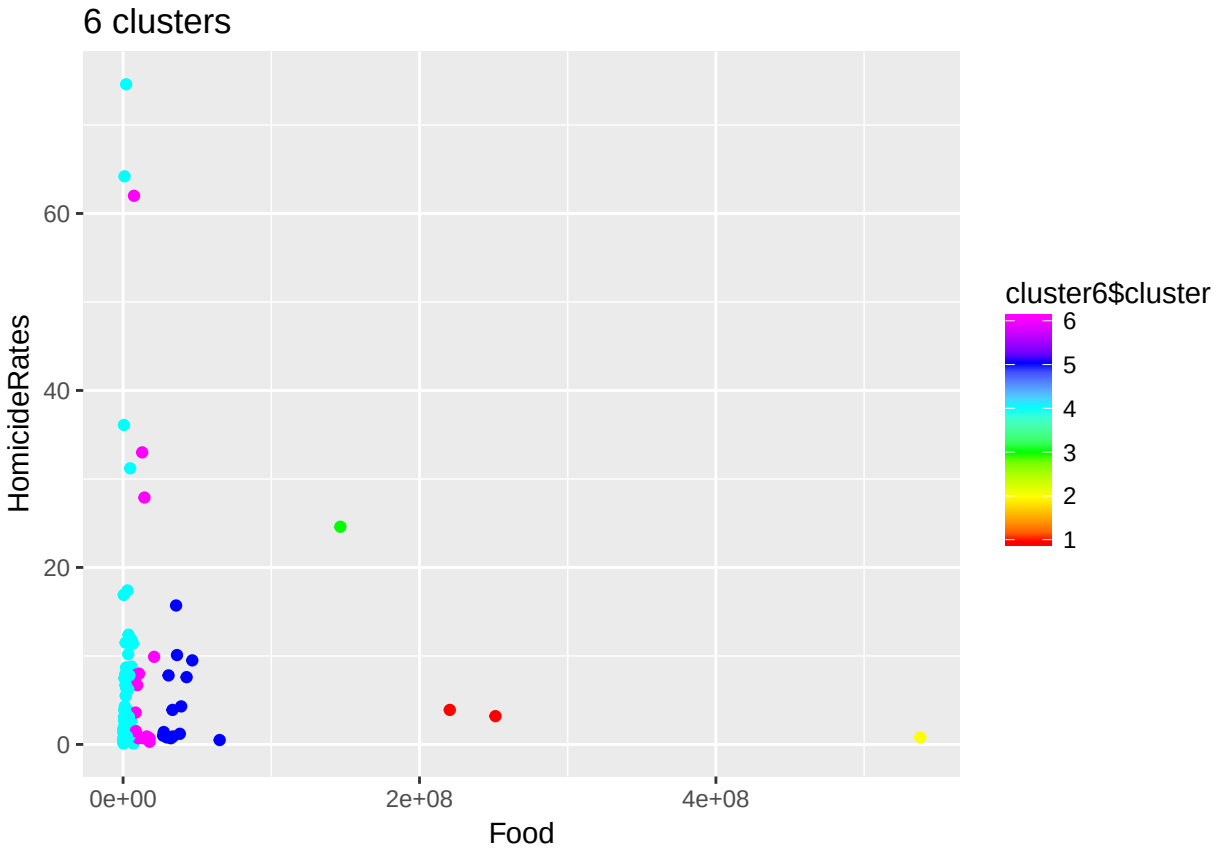
The following will calculate the location of the cluster's centers, to find what attributes they vary the most under.

```
Center4 = as.data.frame(cluster4$centers)
Center5 = as.data.frame(cluster5$centers)
Center6 = as.data.frame(cluster6$centers)
```

The previous suggests that food and homicide rates will be useful features to differentiate clusters, since the clusters differ mostly within these variables. The following charts now clearly show that homicide rates and food can be used to visually interpret the clusters. This is likely a result of the hugeness of the food production values. Next, clustering will be redone with scaled data.

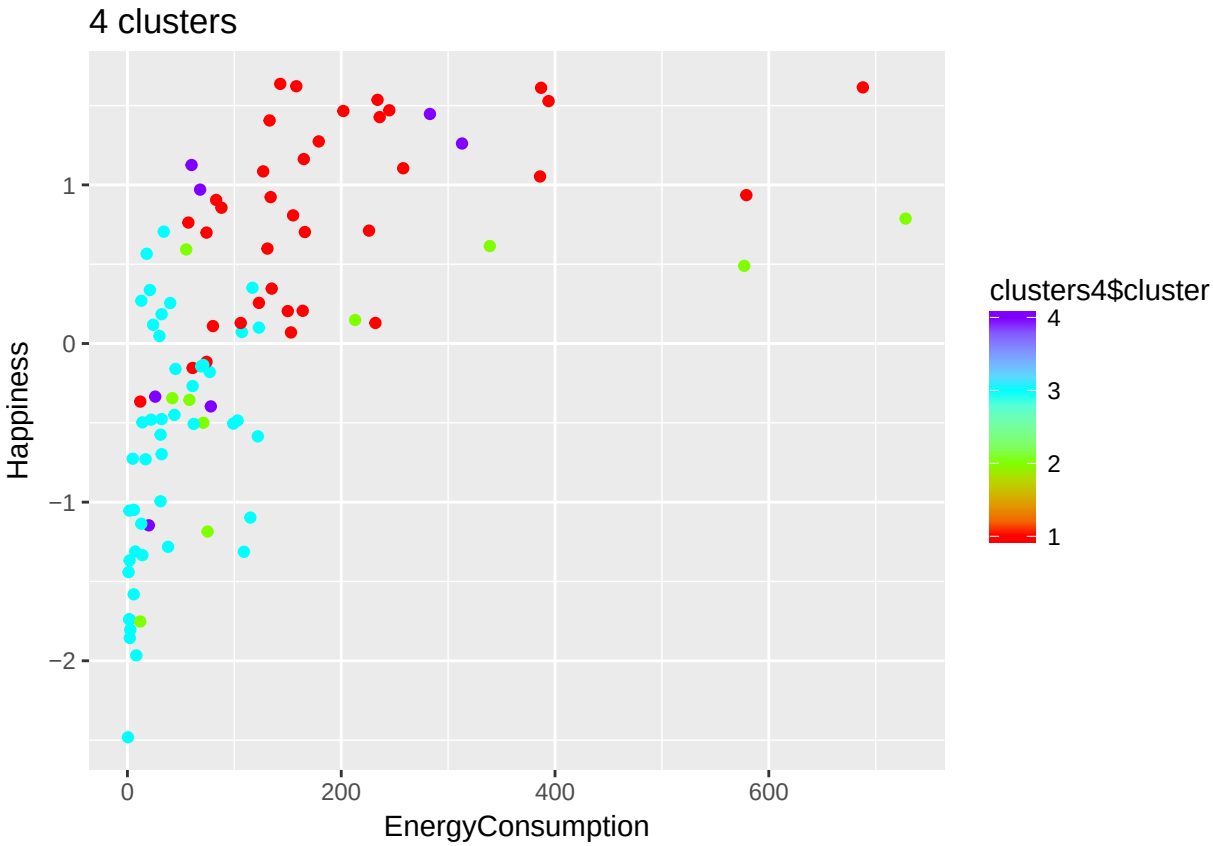

4 clusters
HomicideRates
60
40
20
0
0e+00
2e+08
4e+08
Food
cluster4$cluster
4
3
2
1

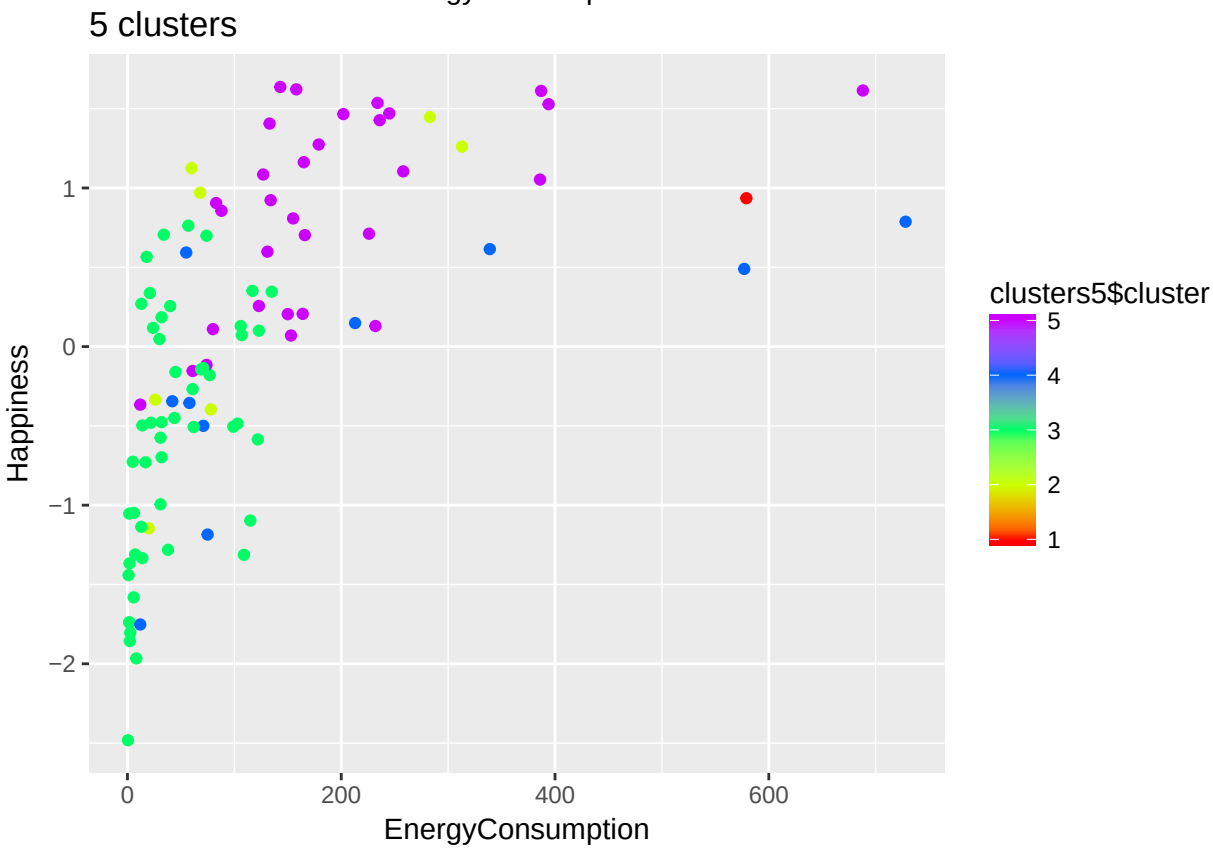

5 clusters
HomicideRates
60
40
20
0
0e+00
2e+08
4e+08
Food
cluster5$cluster
5
4
3
2
1

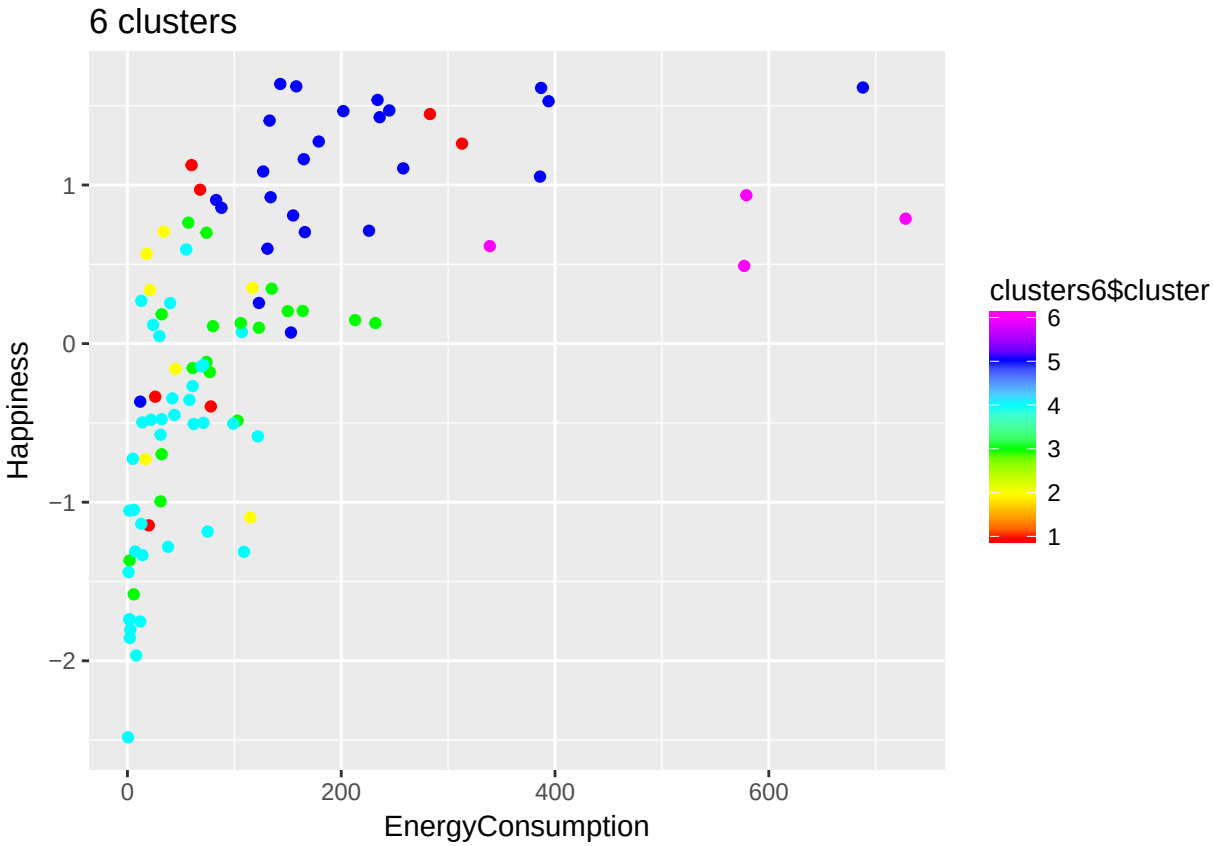


The following clusters have scaled data, which makes the clusters on the original axes more clear and interpretable. The data has a power-law distribution when plotted with energy consumption vs happiness, clustering seems to break different parts of this curve into different clusters. The "knee" of the curve, has its own cluster for all of the k-values, as does the "bottom" (left part) of the curve.

```
scaledata = newdata
scaledata[,-c(1)] = scale(scaledata[,-c(1)])

clusters5 = kmeans(scaledata[,-1],5)

clusters4 = kmeans(scaledata[,-1],4)

clusters6 = kmeans(scaledata[,-1],6)
```


4 clusters
Happiness
EnergyConsumption
clusters4$cluster
1
2
3
4
0
200
400
600
−2
−1
0
1


5 clusters
Happiness
EnergyConsumption
clusters5$cluster
1
2
3
4
5
0
200
400
600
−2
−1
0
1



k-NN

The first thing to do for kNN shall be to split the data into a training and test set.

```
library(class)
library(reshape2)
X.dat = as.matrix(scaledata[, -6])
Y.dat = as.matrix(scaledata[,6])
Y.dat = ifelse(Y.dat>0,1,0)


set.seed (1)
perc = 0.7

n <- nrow(scaledata)
train <- sample (n, size = floor (n*perc), replace = F)

X.train <- X.dat [train,]
X.test <- X.dat [-train,]

Y.train <- Y.dat[train]
Y.test <- Y.dat[-train]
```

The kNN model simply computes if a country is happier than the median or not. The error rate for k=2 is relatively high (>34%), so cross-validation will be used to find the optimal k.

```
knn_model <- knn (X.train[,-1],X.test[,-1],Y.train,k=2)
summary (knn_model)
```

```
##  0  1
## 16 14
```

```
error_train <- table(knn_model, Y.test)
train.error <- 1-sum(diag(error_train))/sum(error_train)
train.error
```

```
## [1] 0.2666667
```

Cross-validation will be done through the knn_cv function from section.

The following will graph the folds of the cross-validation with errors, and display the average error for each k.


```
##         1         2         3         4         5         6         7
## 0.3370370 0.3082716 0.3132099 0.3053086 0.2943210 0.2839506 0.2896296
##         8         9        10        11        12        13        14
## 0.2987654 0.2965432 0.3032099 0.3053086 0.3133333 0.3198765 0.3120988
##        15
## 0.3112346
```

This suggests k=6 has the lowest error for kNN model. The error rate is nonetheless high for all k, suggesting that countries happiness is difficult to predict from their neighbors (in the parameter space). This may be alleviated by using more categories, as opposed to happier than median versus less happy than the median, since more happiness categories may allow more nuance between similar countries.

```
##         1         2         3         4         5         6         7
## 0.5897531 0.5669136 0.5649383 0.5408642 0.5470370 0.5358025 0.5304938
##         8         9        10        11        12        13        14
## 0.5461728 0.5465432 0.5428395 0.5408642 0.5434568 0.5392593 0.5513580
##        15
## 0.5587654
```

Using more happiness categories did not improve accuracy, it made it far worse, as the above errors show. These error rates reflect a new response variable with 4 happiness factors.

LDA/QDA analysis

```
library(MASS)

X = as.data.frame(newdata[,-6])
Y = Y.dat

set.seed(1)
lda1 = lda(Y~., data=X[,-1], CV =TRUE)

lda_error <- table (lda1$class,Y)
lda_error
```

```
##    Y
##      0  1
##   0 35 15
##   1 11 38
```

```
error_lda <- 1-sum(diag(lda_error))/sum(lda_error)
error_lda
```

```
## [1] 0.2626263
```

CV-LDA has slightly better performance than CV-kNN for classifying whether a country's happiness is below the median or not.

```
qda1 = qda(Y~., data=X[,-1], CV =TRUE)

qda_error <- table (qda1$class,Y)
qda_error
```

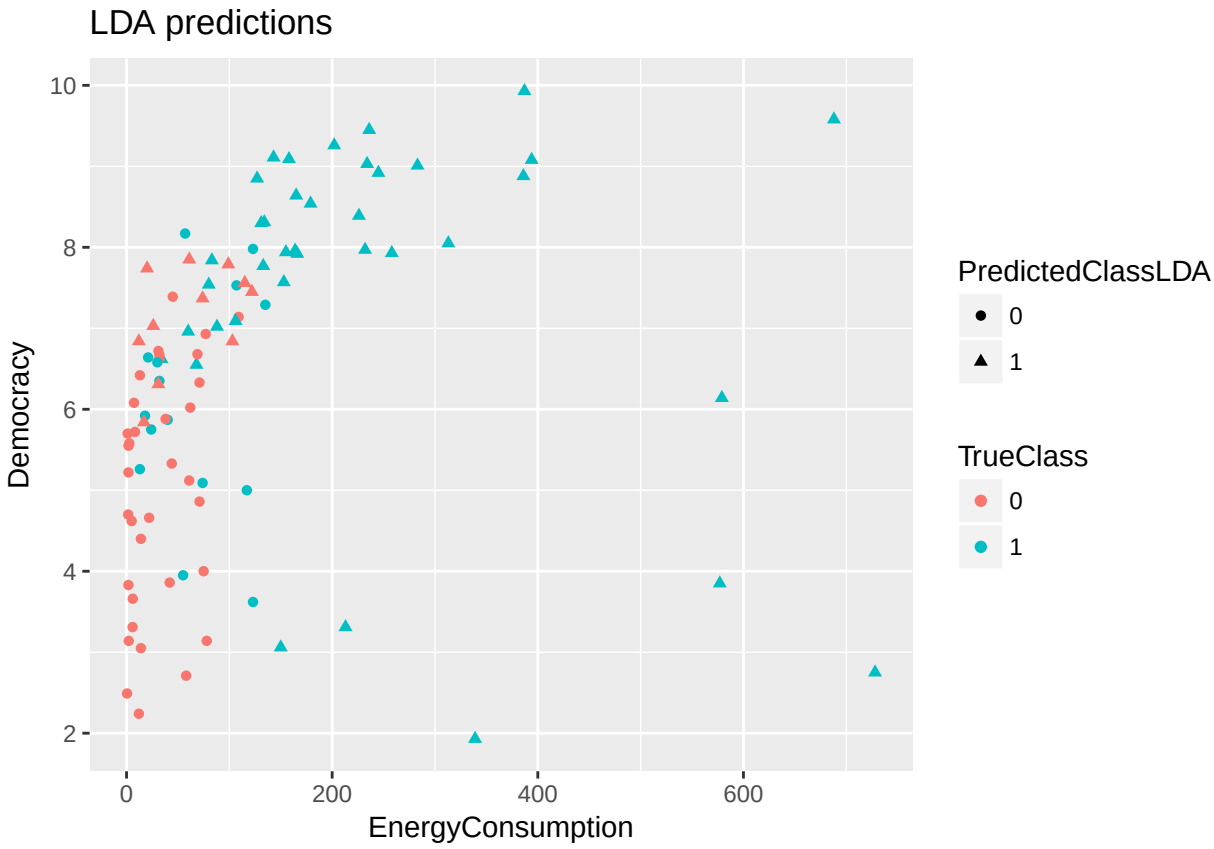
```
##    Y
##      0  1
##   0 39 15
##   1  7 38
```

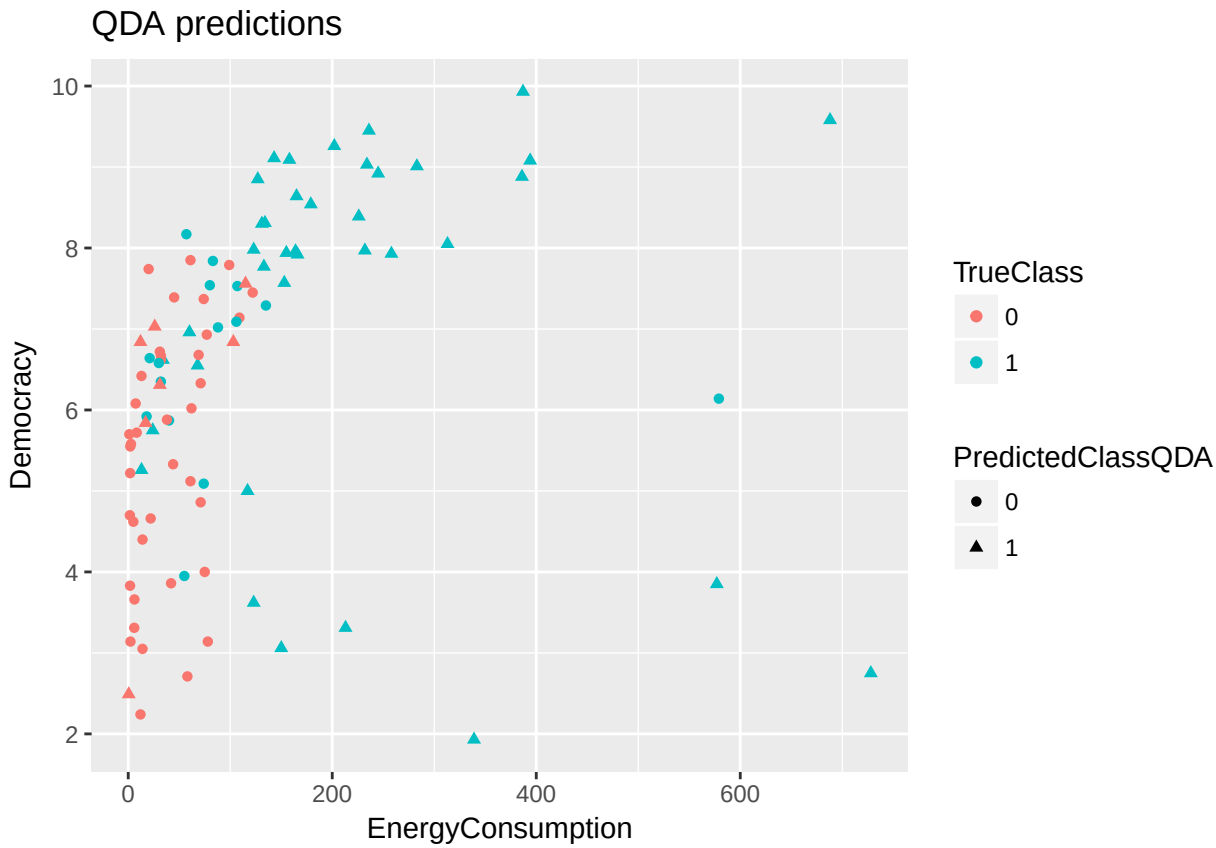
```
error_qda <- 1-sum(diag(qda_error))/sum(qda_error)
error_qda
```

```
## [1] 0.2222222
```

The above shows that QDA performs better than both LDA and kNN for classifying happy countries. Bootstrap was not used due to issues with collinearity in bootstrapping, so LOOCV was used instead.

The following plots show the LDA and QDA predictions (shape) vs the actual class values (color). A value of "1" is the class that is happier than the median, "0" less happy.





Linear Model Predictions

In this section, a quadratic model will be fit to the data (using the most significant variables from the linear model).

```
E = newdata$Energy.Consum
D = newdata$Democracy
P = newdata$Population.Density
H = newdata$High.Tech.Exports

model1 = lm(Happiness~poly(E,2)+poly(D,2)+poly(P,2)+poly(H,2), data=newdata)
summary(model1)
```

```
##
## Call:
## lm(formula = Happiness ~ poly(E, 2) + poly(D, 2) + poly(P, 2) +
##     poly(H, 2), data = newdata)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.75687 -0.39297 -0.05034  0.36844  1.44926
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.68965    0.06788  83.814  < 2e-16 ***
```

```
## poly(E, 2)1  4.98640    0.97551   5.112 1.78e-06 ***
## poly(E, 2)2 -2.50297    0.85101  -2.941 0.004156 **
## poly(D, 2)1  3.18857    0.92794   3.436 0.000895 ***
## poly(D, 2)2  0.38246    0.90555   0.422 0.673776
## poly(P, 2)1 -1.19042    0.90822  -1.311 0.193288
## poly(P, 2)2  1.56531    0.72142   2.170 0.032660 *
## poly(H, 2)1  1.40276    0.84177   1.666 0.099100 .
## poly(H, 2)2 -0.51997    0.91068  -0.571 0.569442
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6754 on 90 degrees of freedom
## Multiple R-squared:  0.6671, Adjusted R-squared:  0.6375
## F-statistic: 22.54 on 8 and 90 DF,  p-value: < 2.2e-16
```

```
summary(model)
```

```
##
## Call:
## lm(formula = Happiness ~ ., data = data3, na.action = na.omit)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.74452 -0.44137  0.06628  0.48972  1.36680
##
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)           3.150e+00  4.041e-01   7.796 1.20e-11 ***
## Food                 -9.460e-10  1.448e-09  -0.653  0.51522
## Biodiversity          6.573e-03  4.394e-03   1.496  0.13831
## Democracy             2.536e-01  5.393e-02   4.702 9.49e-06 ***
## Energy.Consum         3.439e-03  6.029e-04   5.705 1.54e-07 ***
## High.Tech.Exports     2.633e-02  9.849e-03   2.674  0.00894 **
## Homicide.Rates        7.994e-03  6.116e-03   1.307  0.19460
## Population.Density   -2.413e-04  9.847e-05  -2.450  0.01625 *
## Military.Expenditure  9.217e-02  5.185e-02   1.778  0.07891 .
## Suicide.Rate         -5.261e-03  1.186e-02  -0.443  0.65850
## Women.Empowerment     3.731e-01  6.303e-01   0.592  0.55541
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.71 on 88 degrees of freedom
##   (59 observations deleted due to missingness)
## Multiple R-squared:  0.6403, Adjusted R-squared:  0.5994
## F-statistic: 15.66 on 10 and 88 DF,  p-value: 9.953e-16
```

Comparing the adjusted-R^2 models for the data, the quadratic model explains the data better than the linear model, even when accounting for complexity of the model. Next the linear model will be used to predict happiness in all countries that didn't have happiness values in the UN data set. These will be compared to subjective happiness ratings from other surveys.

```
##        5       11       14       24       28       43       48       50
##       NA       NA       NA       NA 5.366183       NA       NA       NA
```

```
##       51       55       56       59       60       64       69       72
##       NA       NA       NA       NA 4.905071 3.933782       NA       NA
##       73       91      101      108      111      115      123      132
## 5.030045       NA       NA       NA       NA       NA       NA       NA
##      136      146      147      151      156      161      162      163
##       NA       NA       NA       NA       NA       NA       NA       NA
##      174      176      181      189      195      196      197      198
##       NA       NA       NA       NA       NA       NA       NA       NA
```

The model predicts Cabo Verde has happiness 5.4; Fiji,4.9;Gambia,3.9;Guyana,5.0. An alternative happiness report gives Guayana 51/100, or 5.1 out of 10, so this model was correct for Guyana. According to a WIN/Gallup poll, Fiji would rank 9.3/10, so its prediction is off by a large amount. The large number of missing entries resulted from the missing fields in many of the attributes in the dataset. To alleviate this problem, a reduced model will be fit using only the most complete attributes

```
df = data.frame(data3$Biodiversity,data3$Energy.Consum,data3$Population.Density,data3$Food)
names(df) = c("Biodiversity","Energy.Consum", "Population.Density","Food")
model2 = lm(data3$Happiness~., df)
summary(model2)
```

```
##
## Call:
## lm(formula = data3$Happiness ~ ., data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.19596 -0.65815  0.01158  0.59968  2.06881
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)         4.831e+00  9.705e-02  49.771  < 2e-16 ***
## Biodiversity        1.578e-02  5.187e-03   3.043  0.00277 **
## Energy.Consum       4.668e-03  5.052e-04   9.240 2.25e-16 ***
## Population.Density -2.096e-04  1.064e-04  -1.970  0.05071 .
## Food               -1.597e-09  1.741e-09  -0.917  0.36041
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9088 on 150 degrees of freedom
##   (3 observations deleted due to missingness)
## Multiple R-squared:  0.3933, Adjusted R-squared:  0.3771
## F-statistic: 24.31 on 4 and 150 DF,  p-value: 1.608e-15
```

```
new=data.frame(data4$Country,data4$Biodiversity,data4$Energy.Consum,data4$Population.Density,data4$Food
names(new) = c("Country","Biodiversity","Energy.Consum", "Population.Density","Food")
hap1=predict.lm(model2,new, na.action = na.omit)
hap1
```

```
##        1        2        3        4        5        6        7        8
## 5.329131 5.581167 5.022830 6.577579 4.884532 5.182510 4.983760 4.950215
##       10       11       13       14       15       16       17       18
## 5.308035 4.815596 5.004535 4.663516 4.972209 4.814355 5.015871 4.822194
```

```
##       20       23       25       26       27       28       29       30
## 4.714564 5.766780 5.261105 4.906799 4.806350 5.605526 4.913934 5.107600
##       31       32       33       34       36
## 4.971542 4.949680 4.805998 4.906800 4.889954
```

The reduced model predicts a high happiness score for Brunei (6.6) which has a score of 7.6 in life satisfaction in the Happy Planet Index (HPI). Gambia is predicted to have the low score of 4.7, which is one lower the HPI measurement of 5.7. Other values seem to lag behind the HPI score by about one, which could be an artifact of the survey methods or simply the inaccuracy of the model. While the reduced model doesn't accurately predict values of happiness, it does predict general features (high/low) well.

GLM

```
new1 = data.frame(data3$Biodiversity,data3$Energy.Consum,data3$Population.Density,data3$Food, data3$Happ
new1 = na.omit(new1)
names(new1) = c("Biodiversity","Energy.Consum", "Population.Density","Food", "Happiness")
medianH = median(new1$Happiness)
new1$Happiness = ifelse((new1$Happiness-medianH)>0,1,0)

n1 <- nrow(new1)
train1 <- sample (n1, size = floor (n*perc), replace = F)

newtrain = new1[train1,]
newtest = new1[-train1,]


glm1 = glm(newtrain$Happiness~.,family=binomial(link='logit'),data=newtrain)
glm.predict = predict(glm1,newtest[,-5],type='response')
glm.predict = ifelse(glm.predict > 0.5,1,0)

glm_error <- table (glm.predict,newtest$Happiness)
glm_error
```

```
##
## glm.predict  0  1
##           0 38 11
##           1  7 30
```

```
error_glm <- 1-sum(diag(glm_error))/sum(glm_error)
error_glm
```

```
## [1] 0.2093023
```

The GLM has the lowest error of all models so far, suggesting that a GLM would be the most useful model for predicting happiness from the given attributes.

Summary

Referring back to the initial research questions:

What variables can be used to predict a country's subjective happiness best?

What distinguishes happy countries from unhappy countries?

For the first, the attributes which best predicted happiness were wealth (the correlated variables energy consumption, GDP, and CO2 emission) and democratic development (score on Democracy index, Women's empowerment, economic freedom). In the linear models specifically, energy consumption and the democracy index were highly significant predictors of happiness ($P<0.001$). This shows that while "money doesn't buy happiness," wealth is highly correlated with happiness. It also suggests that democracy is much more conducive to happiness than most alternative forms of government. However, due to a number of potentially confounding variables (i.e. wealth and health are correlated, as are wealth and democracy) these results likely are biased.

To answer the second question, many models were used to distinguish happier countries from less happy countries. Classification models were used to make a simple prediction: is a country happier than the median or not? The models, from most accurate to least, were GLM, QDA, LDA, and kNN. GLM peaks around 20% accuracy, so while it is much better than guessing, it stills leaves much to be desired.The success of the GLM and QDA suggest that the association between happiness and the variables of interest are nonlinear. This suggests, as intuition would predict, that happiness is very complex and depends on a number of factors in nonlinear ways. In addition to the classification models, linear models were used to predict numerical happiness scores from the other attributes. These scores were occasionally accurate, but that is potentially due to coincidence.

Citations

CO2 Emission per capita, http://cdiac.ornl.gov/

Women's Empowerment, https://www.weforum.org/reports/global-gender-gap-report-2015/

Suicide Rate, http://www.worldlifeexpectancy.com/

Democracy Index, http://www.eiu.com/ or (paper) (http://www.yabiladi.com/img/content/EIU-Democracy-Index-2015.pdf)

Index of economic freedom, http://www.heritage.org/index/

Military expenditure per capita, http://www.sipri.org/databases/milex

Agricutural production, http://faostat.fao.org/site/613/default.aspx#ancor

Energy production,http://www.eia.gov/cfapps/ipdbproject/IEDIndex3.cfm?tid=6&pid=29&aid=12

Homicide rate, https://data.unodc.org/?lf=1&lng=en

GDP per capita, biodiversity, education expenditure per capita, health expenditure per capita, high tech exports, journal articles published, research expenditure, and population density: http://data.worldbank.org/data-catalog/world-development-indicators

World Happiness Report, http://worldhappiness.report/ed/2016/