Predicting World Happiness

FUNMI ESURUOSO
IST 707-17403
PROFESSOR LIN

INTRODUCTION

- The World Happiness Report ranks 155 countries by their "happiness levels". Three years of data - 2015, 2016 and 2017 was acquired through the Gallup World Poll.
- The dataset variables this report focuses on include:
  - Country/Region
  - Happiness Rank
  - Happiness Score
  - Economy (GDP per Capita)
  - Family
  - Health (Life Expectancy)
  - Freedom
  - Trust (Government Corruption)
  - Generosity
  - Dystopia Residual
- Happiness score is a metric measured by asking sampled individuals: "How would you rate your happiness on a scale of 0 to 10 where 10 is the happiest."
- Happiness rank is given based on the happiness scores, and the remaining attributes.

BUSINESS PROBLEM

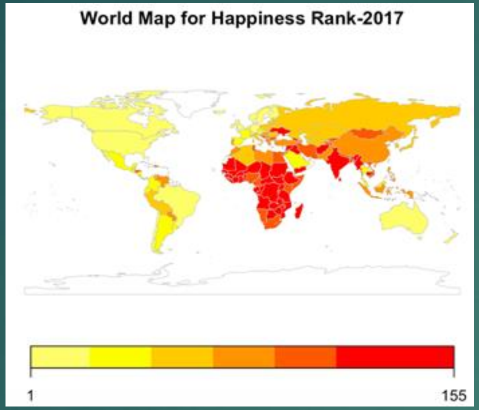
- Happiness indicators are more frequently being used by governments, organizations and society to inform policy decisions. These measures of well-being are able to assess the progress of nations and ranks them comparatively to each other.
- Our analysis seeks to explore a few questions:
  - Which attributes have the biggest impact on happiness?
  - Which countries have experienced significant increases or decreases in happiness overtime?
  - Can we predict the happiness rank/score of a country based on the provided variables?

Exploratory Analysis

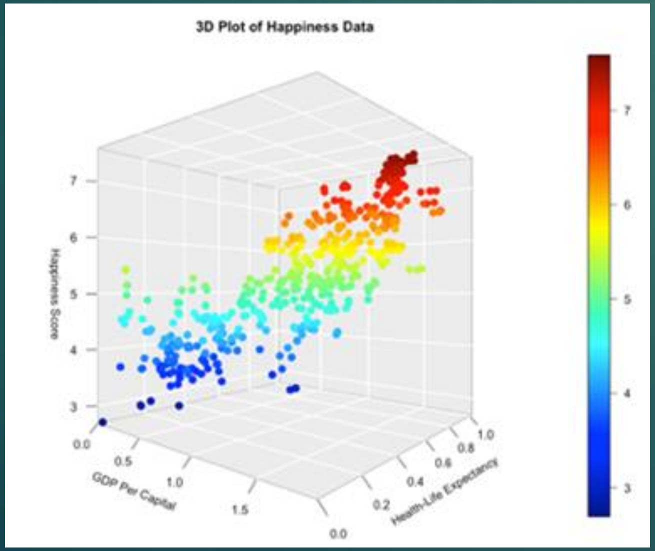
Two way Distribution & World Map

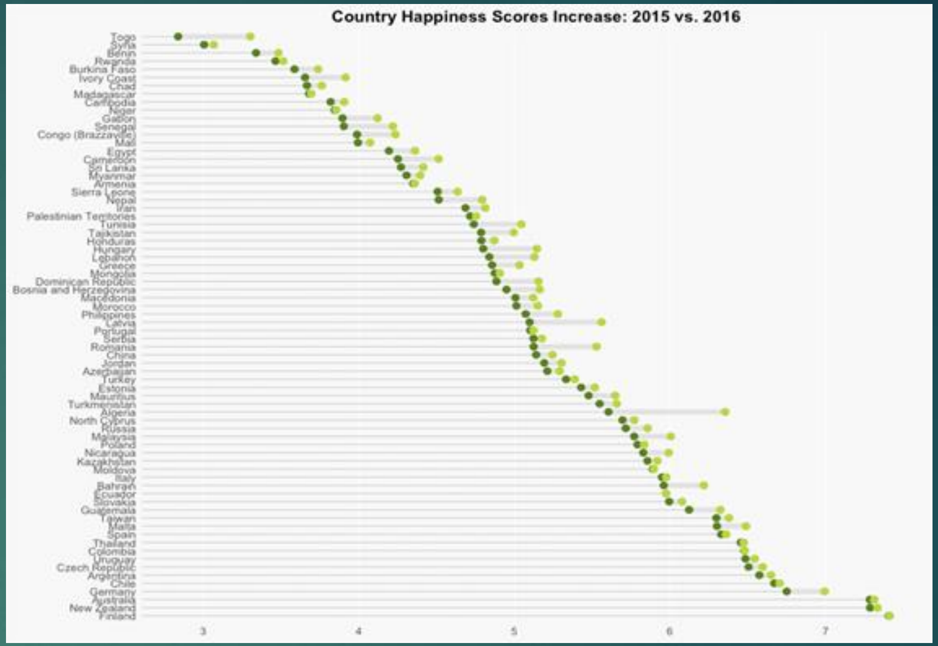
A tibble: 6 x 2

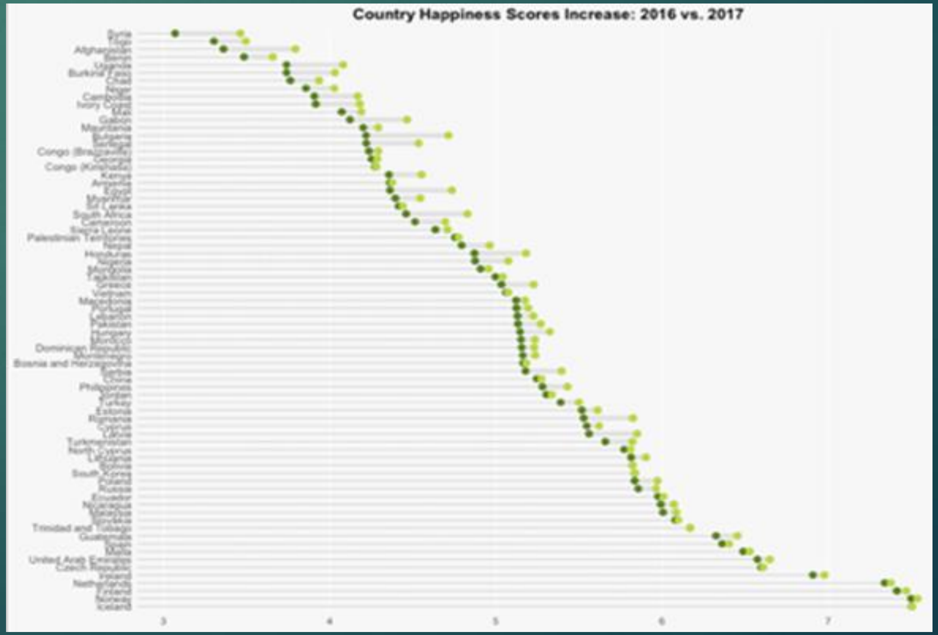
	continent	mean
	<fct>	<dbl>
1	Africa	4.24
2	Asia	5.28
3	Australia	7.30
4	Europe	6.10
5	North America	6.03
6	South America	6.10



3D Scatter Plot





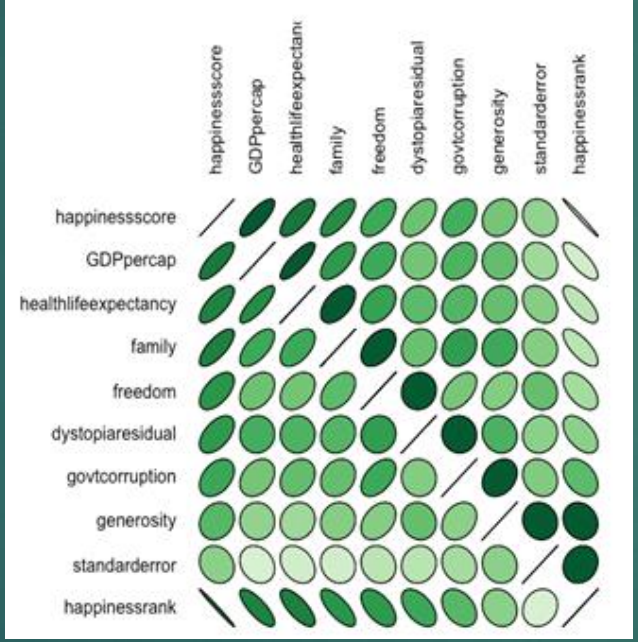


Happiness Score Change Over (2015-2017)

Exploratory Analysis cont...

Correlation Percentages & Matrix

Attribute	Correlation
Generosity	0.1635616
Trust (Government Corruption)	0.4063397
Dystopia Residual	0.4897472
Freedom	0.5603534
Family	0.636532
Health (Life Expectancy)	0.7480404
Economy (GDP per Capita)	0.7854496

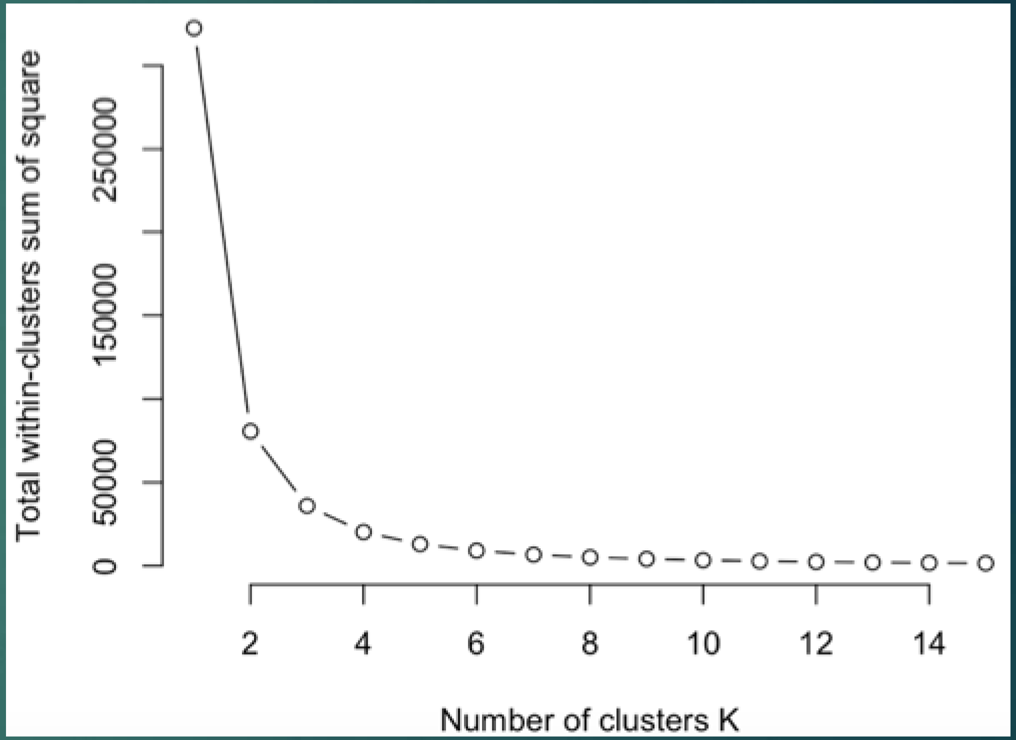


Multiple Linear Regression (MLR)

```
Call:
lm(formula = happinessscore ~ GDPpercap + family + healthlifeexpectancy +
    freedom + dystopiaresidual + govtcorruption, data = data)

Residuals:
     Min       1Q   Median       3Q      Max
-0.28853 -0.08189 -0.01691  0.06141  0.49881

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)           0.20400    0.02920   6.987  9.8e-12 ***
GDPpercap             0.90398    0.02383  37.926  < 2e-16 ***
family                0.99996    0.02288  43.707  < 2e-16 ***
healthlifeexpectancy  1.09171    0.03807  28.675  < 2e-16 ***
freedom               1.26310    0.04654  27.142  < 2e-16 ***
dystopiaresidual      0.97054    0.00986  98.434  < 2e-16 ***
govtcorruption        1.22139    0.05851  20.874  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1193 on 463 degrees of freedom
Multiple R-squared:  0.9891,    Adjusted R-squared:  0.989
F-statistic:  7028 on 6 and 463 DF,  p-value: < 2.2e-16
```

K-Means & Discretization

- Created elbow plot & HAC output to visualize how our data naturally clusters together
- Since our dataset combines information on countries from multiple years, we needed a way to categorize the variable
- Decided on an optimal number of bins (3) to discretize our class label (happiness rank)
  - "Very Happy"
  - "Happy"
  - "Less Happy"
- This discretized variable was the basis for all remaining analysis methods



Association Rules

- One of our objectives was determining which attributes contributed most to making a country "very happy"
- Through setting this specific category = RHS, we were able to generate rules
- Since arules requires only categorical attributes, we further discretized all variables to an ordered low, medium, high value
- Hyperparameters tuned: support, confidence, max length

	lhs		rhs	support	confidence	lift	count
[1]	{GDPpercap=high, freedom=high}	=>	{happinessrank=very happy}	0.1702128	0.8888889	2.678063	80
[2]	{GDPpercap=high, govtcorruption=high}	=>	{happinessrank=very happy}	0.1659574	0.8863636	2.670455	78
[3]	{healthlifeexpectancy=high, freedom=high}	=>	{happinessrank=very happy}	0.1659574	0.8478261	2.554348	78
[4]	{healthlifeexpectancy=high, govtcorruption=high}	=>	{happinessrank=very happy}	0.1574468	0.8409091	2.533508	74
[5]	{GDPpercap=high, family=high}	=>	{happinessrank=very happy}	0.1553191	0.8295455	2.499272	73

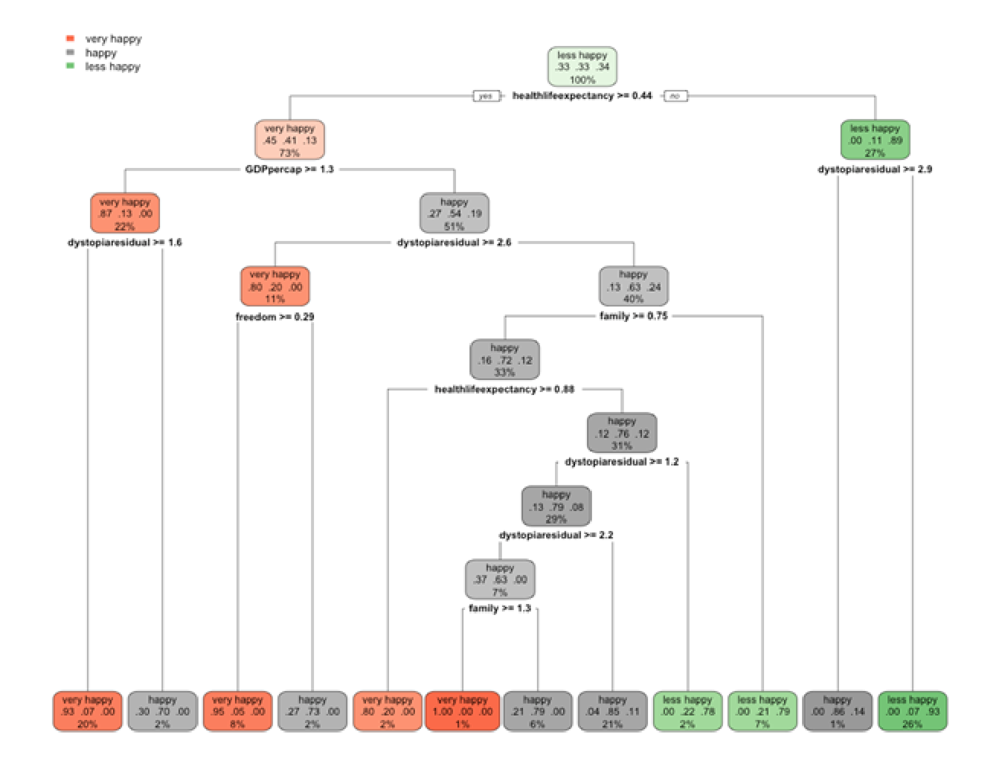
- Consistently, GDP per capita & health/life expectancy have the biggest impact
- Government corruption may be less of a factor in poorer/worse ranked countries as they place importance on things we take for granted such as freedom & family

Decision Tree

- We wanted to see if we could accurately predict a country’s happiness rank using our three class labels.
- 80% of our dataset was used for training while the remaining 20% was used to test our model’s accuracy.
- Ran several models with a two-way split and different attributes based on our descritized class label.
- Best model correctly classified 296 instances and could accurately predict a countries overall rank with 86% accuracy.

```
=== 10 Fold Cross Validation ===

=== Summary ===

Correctly Classified Instances         296               78.7234 %
Incorrectly Classified Instances        80               21.2766 %
Kappa statistic                          0.6807
Mean absolute error                      0.1704
Root mean squared error                  0.3476
Relative absolute error                 38.3543 %
Root relative squared error             73.7541 %
Total Number of Instances              376

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.840    0.104    0.802      0.840    0.820      0.728    0.901     0.804     very happy
                 0.675    0.148    0.681      0.675    0.678      0.528    0.785     0.574     happy
                 0.840    0.065    0.873      0.840    0.856      0.782    0.925     0.855     less happy
Weighted Avg.    0.787    0.105    0.788      0.787    0.787      0.683    0.872     0.748

=== Confusion Matrix ===

   a   b   c   <-- classified as
 105  19   1 |   a = very happy
  24  81  15 |   b = happy
   2  19 110 |   c = less happy
```

```
Confusion Matrix and Statistics

            Reference
Prediction   very happy happy less happy
  very happy         26     2          0
  happy               5    32          4
  less happy          0     2         23

Overall Statistics

               Accuracy : 0.8617
                 95% CI : (0.7751, 0.9243)
    No Information Rate : 0.383
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.7899
```



Naive Bayes

- Used as another supervised learning method for classification
- We sought to predict the happiness ranking of a country based on our discretized class label
- Initial model received a 91.4% accuracy
  - Less valid since one of the variables (happiness score) is highly correlated with happiness rank
- Removed happiness score & re-ran model
- New model can predict the overall general rank of a country's happiness with a 77.42% accuracy

```
             Reference
Prediction    very happy happy less happy
  very happy          23     3          0
  happy                8    23          5
  less happy           0     5         26

Overall Statistics

               Accuracy : 0.7742
                 95% CI : (0.6758, 0.8545)
    No Information Rate : 0.3333
    P-Value [Acc > NIR] : < 2.2e-16
```

Conclusion & Deployment

- A country's GDP, health/life expectancy, family, and freedom have the highest impact on a country's happiness.
- Based on our descritzed class labels or "happiness ranks"; "very happy", "happy", and "less happy", our association rules output showed that if a country's GDP, freedom, and trust in the government regarding corruption are high, the country is likely to have a very happy ranking.
- Our most efficient decision tree and naive bayes models predicted a countries overall rank with 86% and 77% accuracy respectively. Both models were great predictors of a country's happiness rank.
- Some users of our analysis and models are as follows:
  - Governments and organizations
  - Elected officials
  - Candidates seeking election

WORKS CITED

Kaggle. (2017, 02 28). *Kaggle*. Retrieved from Kaggle.com: https://www.kaggle.com/unsdsn/world-happiness/metadata