Lynn Fan, Annie Lin, Leona Lu, Eric Young
Duke Computer Science 216 – Everything Data, Spring 2021
April 28, 2021

Project Final Report: Happiness Scores For Global Citizens

Part 1: Introduction and Research Questions

The World Happiness Report, first published in 2012, uses *Gross Domestic Product (GDP) per capita, family, life expectancy, freedom, generosity, and absence of government corruption* as parameters to measure happiness of citizens across the world. By reflecting the impact of big events and government policies on the happiness of populations, the report allows policymakers and researchers to evaluate past policies and to make informed future political decisions. Based on the report, we will conduct vertical comparison to observe a country's change in happiness score over time and horizontal comparison to analyze the differences in happiness scores among different groups of countries (for example, between developed countries and developing countries). Additionally, we will investigate correlation between different parameters to determine if high score in one parameter contributes to high score in another. The research questions have largely stayed unchanged throughout the project:

1. How are parameters such as GDP per capita and life expectancy related to happiness scores of citizens among all countries?
2. Do developed countries overall have a different set of statistically significant parameters than developing countries?
3. Is there any country experiencing significant increase/decrease in happiness score over the duration of the Report, and which parameters are the most significant in affecting the changes?

The research question was originally two separate questions: *how are parameters such as GDP per capita and life expectancy related to happiness scores of citizens in a particular country*, and *which parameters included in the Report are statistically significant among all countries?* After initial data exploration, however, we realized that these two questions have overlapping with the rest of the report. Therefore, we condensed it to just 1 question: *how are parameters such as GDP per capita and life expectancy related to happiness scores of citizens among all countries?*

Happiness is a critical metric that measures the relative citizen welfare of countries, and countries often attribute their happiness scores to specific regime characteristics. Therefore, we believe it is important to explore what variables can contribute to an increase or decrease in happiness score. Although causal relationships are hard to prove with regression, correlations have important implications to policy-making and the human rights agenda as well. We will do regression analysis with happiness scores as the response variable to explore what variables are statistically significant to changes in happiness scores, and use cluster analysis to understand whether being developed/developing has implications on happiness scores. We will also conduct time-series analysis on countries with multiple data entries to see if any variables are related to their changes in happiness over time. In addition, our current database has rows for each country, so it is convenient to include additional datasets that provide more demographic or political/social/economic statistics to study literature review and to see how significant our research is in the context of existing relevant study.

Part 2: Summary of Results

For question 1, we used a linear regression model to analyze what factors are statistically significant to changes in happiness scores. The final model includes five variables: social support, freedom to make life choices, generosity, log GDP per capita, and perceptions of corruption. The feature selector we employed did not choose healthy life expectancy at birth and Human Freedom score, although our EDA shows linear correlations between these variables and happiness scores. We found out that removing these two variables significantly improves multicollinearity problems among variables, which might explain why the selector did not choose these variables in the final model.

For question 2, we employed the k-means clustering algorithm with k=3 (based on the "elbow method") and a convergence criterion of 0.001 to group countries with similar characteristics into 3 distinct groups. In general, our results show that developed countries tend to have a higher happiness score, GDP, social support, life expectancy, and perception of corruption but a lower generosity level. Conversely, developing countries tend to have a lower happiness score, GDP, social support, life expectancy, and perception of corruption but a higher generosity level.

Part 3: Data Source

Dataset: World Happiness Report

The World Happiness Report dataset from Kaggle contains 5 csv files that detail happiness scores and rankings of 155 countries in 2015, 2016, 2017, 2018, and 2019 respectively. Each annual report consists of quantitative data columns representative of 6 factors - economic production, social support, life expectancy, freedom, absence of corruption, and generosity- that contribute to overall happiness. The temporal cohesiveness of the separate files within the dataset allows us to conduct time series analysis on the change of happiness over time of citizens from diverse locations. With specific changes in the aforementioned data columns noted, we can further examine the root causes of shifts in happiness ratings, perhaps related to major government policies or socio-economic phenomena.

As the project progressed, in addition to the World Happiness Report, we also found The Human Freedom Index, a dataset developed to measure economic freedom metrics of different countries, such as the freedom to trade or to use sound money. We are interested in how the human freedom index, calculated by another research group, might help explain happiness scores, and merged these two datasets for our analysis.

Part 4: Results and Methods

Research Question 1:
How are parameters such as GDP per capita and life expectancy related to happiness scores of citizens among all countries?

To answer the first question, we first conducted exploratory data analysis to see what variables could be potentially correlated to happiness scores, plotting scatterplots of these variables against the happiness scores. According to the scatterplots, among the available variables in the dataset, Log GDP per capita, social support, healthy life expectancy at birth, freedom to make life choices, perceptions of corruption, and The Human Freedom Index score all show linear or near-linear correlations. Therefore, we decided to use a linear regression model for this question.

To start off, we use the Epsilon-Support Vector Regression module sklearn.svm and feature selection module sklearn.feature_selection to select model variables. We started with Log GDP per capita, Social support, Healthy life expectancy at birth, Freedom to make life choices, Generosity, Perceptions of corruption, and Human Freedom score. We included those that did not show a strong linear trend in EDA, so that we can provide the selector with most information possible. The selector chose all of them except for Healthy life expectancy at birth and Human Freedom score. The model has a MSE of 0.36 and a R^2 score of 0.73. The coefficients of the variables selected are displayed in the table below:

Variable	Coefficient
Social support	2.018753
Freedom to make life choices	1.002960
Generosity	0.904757
Log GDP per capita	0.536794
Perceptions of corruption	-0.723674

Table 1: Coefficient Table for Question 1

For social support to decrease by 1 unit (a range of 0 to 1), with other variables remaining the same, Happiness score is predicted to decrease by 2.018753 (a range of 0 to 5).

To explore why the selector did not choose Healthy life expectancy at birth and Human Freedom score, we looked into multicollinearity as one of the potential explanations using variance inflation factor from the module statsmodels. The variance inflation factor (VIF) is a common metric for multicollinearity, as it measures the increase of the variance of the parameter estimates if an additional variable is added to the regression model. Below is the table of VIF values of those when all variables are selected, and when Healthy life expectancy at birth and Human Freedom score are removed. Multicollinearity problems improved significantly after the two variables were removed, so we believe multicollinearity issues might be why the model selector did not choose them.

All variables	VIF	Model selected variables	VIF
Social support	104.364910	Social support	103.346769
Freedom to make life choices	38.531594	Freedom to make life choices	35.637514
Generosity	1.208316	Generosity	1.206848
Log GDP per capita	376.941248	Log GDP per capita	109.251525
Perceptions of corruption	11.286572	Perceptions of corruption	10.926971
Healthy life expectancy at birth	298.922636		
HF score	146.905392		

Table 2: VIF Table for Question 1

Question 2:
Do developed countries overall have a different set of statistically significant parameters than developing countries?

I. Theory

We employed the k-means clustering algorithm to efficiently group the countries into multiple clusters with distinct characteristics. The general idea of the k-means clustering algorithm is to partition points into k-clusters c1, c2, …, ck and find the corresponding k center points z1, z2, …, zk in order to minimize the cost function.

More precisely, the k-means clustering algorithm follows the following 3 iterative steps:

1. Initialization: to start the algorithm, randomly select k data points from the dataset as initial centroids. Note that centroids are the centers of clusters. In this case, since we have no information indicating the location of cluster centers, we arbitrarily define them.
2. Cluster Assignment: calculate the Euclidean distances of each data point from each centroid, assigning all the data points that are closest to a centroid to a cluster.
3. Centroid Movement: calculate the new centroid of each cluster, which is the mean of all data points in the cluster. Repeat steps 2 and 3 until the algorithm converges to acquire the final clustering assignments.

Lastly, we used the elbow method, which runs the k-means clustering algorithm with different numbers of clusters and finds the within-cluster sum of squares (measures within-cluster variance), to determine the ideal number of clusters to choose.

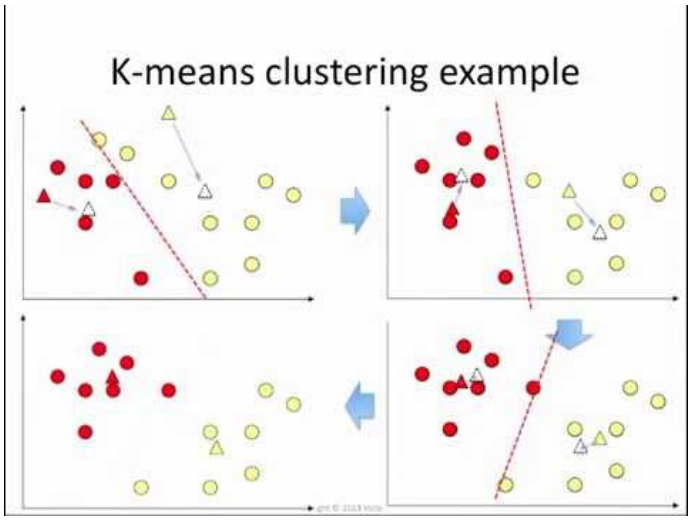

Figure 1: K-means clustering example

II. Code

1. Data Preprocessing:

Upon data mining, we normalized the data (rescaled numeric values into the range 0 and 1) to ensure that every attribute contributed equally to determining the distance between points. We chose normalization

over standardization because our input attributes rely on the magnitude of values instead of the distribution of Gaussian processes.

2. Selecting the Ideal Number of Clusters (Elbow Plot):

To select the ideal number of clusters for the k-means clustering algorithm to find, we use the elbow method and notice that “elbow of the curve” is around 3 clusters. To minimize the within-cluster sum of squares error while drawing useful conclusions, we choose 3 clusters for our algorithm.

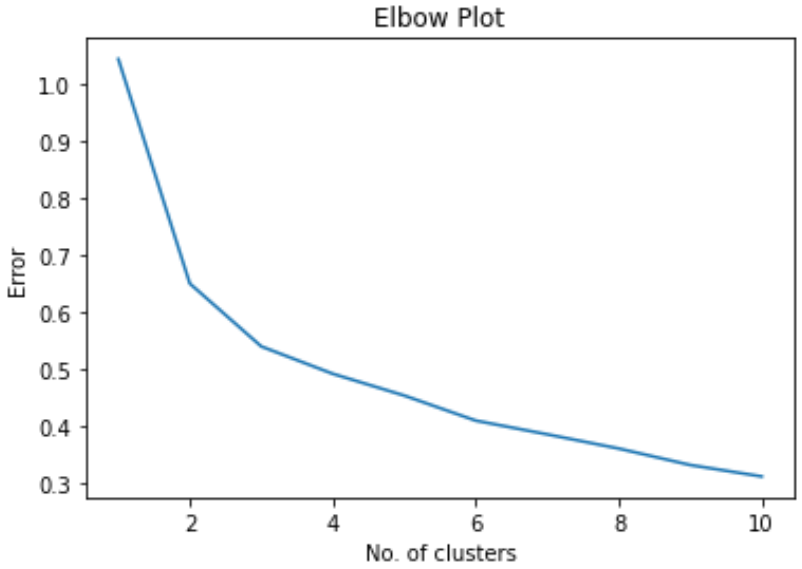

Figure 2: Elbow Plot

3. Implementing the K-Means Clustering Algorithm

We proceed to implement the k-means clustering algorithm with 3 clusters and a convergence criterion of 0.001. Note that the convergence criterion, which determines when iteration ceases, represents a proportion of the minimum distance between initial cluster centers.

III. Results

From the figure below, we notice that Group 0 contains the most developed countries and the group contains the highest average happiness score, GDP, social support, and life expectancy, which is expected. However, Group 0 also has the lowest generosity and the highest corruption perception. Conversely, Group 1 contains the lowest happiness score, GDP, social support, and life expectancy, which is expected. However Group 1 also has the highest generosity and the lowest corruption perception.

Group 0	Group 1	Group 2
'Finland', 'Denmark', 'Norway', 'Iceland', 'Netherlands', 'Switzerland', 'Sweden', 'New Zealand', 'Canada', 'Austria', 'Australia', 'Israel', 'Luxembourg', 'United Kingdom', 'Ireland', 'Germany', 'Belgium', 'United	'Pakistan', 'Tajikistan', 'Nigeria', 'Cameroon', 'Ghana', 'Ivory Coast', 'Benin', 'Congo (Brazzaville)', 'Somalia', 'Niger', 'Burkina Faso',	'Costa Rica', 'Guatemala', 'El Salvador', 'Uzbekistan', 'Nicaragua', 'Kosovo',

States', 'Czech Republic', 'United Arab Emirates', 'Malta', 'Mexico', 'France', 'Taiwan', 'Chile', 'Saudi Arabia', 'Qatar', 'Spain', 'Panama', 'Brazil', 'Uruguay', 'Singapore', 'Italy', 'Bahrain', 'Slovakia', 'Trinidad & Tobago', 'Poland', 'Lithuania', 'Colombia', 'Slovenia', 'Argentina', 'Romania', 'Cyprus', 'Kuwait', 'Thailand', 'Latvia', 'South Korea', 'Estonia', 'Mauritius', 'Japan', 'Kazakhstan', 'Hungary', 'Northern Cyprus', 'Peru', 'Portugal', 'Russia', 'Serbia', 'Libya', 'Montenegro', 'Croatia', 'Hong Kong', 'Dominican Republic', 'Bosnia and Herzegovina', 'Turkey', 'Malaysia', 'Belarus', 'Greece', 'Mongolia', 'North Macedonia', 'Turkmenistan', 'Algeria', 'Morocco', 'Azerbaijan', 'Lebanon', 'China', 'Bulgaria', 'Jordan', 'Gabon', 'South Africa', 'Albania', 'Venezuela', 'Namibia', 'Armenia', 'Iran', 'Georgia', 'Tunisia', 'Iraq', 'Sri Lanka', 'Ukraine', 'Egypt', 'India', 'Botswana'	'Guinea', 'Gambia', 'Mozambique', 'Congo (Kinshasa)', 'Sierra Leone', 'Chad', 'Togo', 'Liberia', 'Comoros', 'Burundi', 'Haiti', 'Syria', 'Malawi', 'Afghanistan', 'Central African Republic', 'South Sudan'	'Ecuador', 'Jamaica', 'Honduras', 'Bolivia', 'Paraguay', 'Philippines', 'Moldova', 'Kyrgyzstan', 'Indonesia', 'Vietnam', 'Bhutan', 'Nepal', 'Laos', 'Cambodia', 'Palestinian Territories', 'Senegal', 'Kenya', 'Mauritania', 'Bangladesh', 'Mali', 'Myanmar', 'Ethiopia', 'Swaziland', 'Uganda', 'Zambia', 'Madagascar', 'Lesotho', 'Zimbabwe', 'Yemen', 'Rwanda', 'Tanzania'
Score Mean: 5.901663043478262 Score SD: 0.9505040456558449 GDP Mean: 1.179945652173913 GDP SD: 0.2003505820854184 Social Support Mean: 1.3495760869565216 Social Support SD: 0.1969450504745606 Life Expectancy Mean: 0.8675760869565217 Life Expectancy SD: 0.14198095659933416 Generosity Mean: 0.16431521739130434 Generosity SD: 0.09228805676036805 Corruption Mean: 0.12126086956521738 Corruption SD: 0.10903488368459618	Score Mean: 4.298307692307692 Score SD: 0.7077769514598165 GDP Mean: 0.3277307692307692 GDP SD: 0.19867049598853714 Social Support Mean: 0.7398461538461538 Social Support SD: 0.2482849932432939 Life Expectancy Mean: 0.38799999999999996 Life Expectancy SD: 0.12155973523650515 Generosity Mean: 0.2185769230769231 Generosity SD: 0.05972319036568308 Corruption Mean: 0.09626923076923079 Corruption SD: 0.05333894355354522	Score Mean: 4.960351351351351 Score SD: 0.9591030217249644 GDP Mean: 0.6332702702702702 GDP SD: 0.2004677838441539 Social Support Mean: 1.1909999999999998 Social Support SD: 0.17562921630987718 Life Expectancy Mean: 0.6212972972972973 Life Expectancy SD: 0.19428842656061143 Generosity Mean: 0.21137837837837833 Generosity SD: 0.10839750454676038 Corruption Mean: 0.09605405405405405 Corruption SD: 0.07149104874830284

* Note that green = highest, yellow = middle, red = lowest

Table 3: Results Table for Question 2

Question 3:

Is there any country experiencing significant increase/decrease in happiness score over the duration of the Report, and which parameters are the most significant in affecting the changes?

To answer the third question, we first visualize the change of the happiness rating within different regions and countries between 2008 to 2018. Based on the graph, Oceania has the highest average life ladder. Asia has the fourth highest average life ladder, and Africa has the lowest.
While Oceania's happiness rating remained stable through 2008 to 2018. Europe has displayed drastic change in the year of 2008. Before 2015, Europe had the second highest average life ladder and Americas had the third highest, but Americas' average life ladder exceeds that of Europe in 2008.
It is also interesting to observe that trends of Happiness rating are similar between Asia and America, being the opposite direction of Europe.

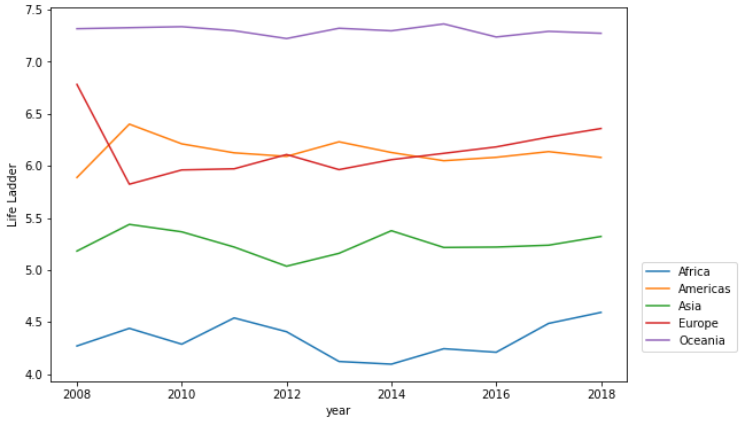

Figure 3: EDA by Region for Question 3

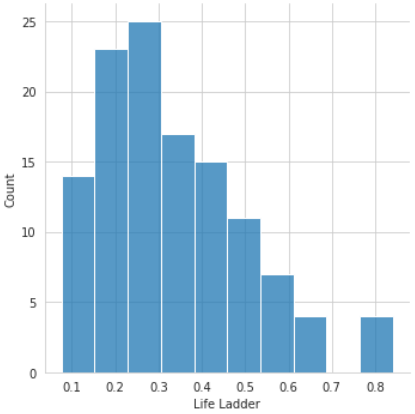

Figure 4: Life Ladder EDA for Question 3

In order to test the reason behind the drop of the happiness score in Europe, we conducted linear regression for Europe in 2008 and 2009. We use happiness score (life ladder) as response variable and explanatory variable as GDP, Social Support, Healthy Life Expectancy, Freedom to make life choices, Generosity, Perception of corruption, Positive/Negative affect, and human freedom score.

For 2008, perception of corruptions. Freedom to make life choices and Human freedom score display significance in predicting happiness score. This means for every one point increase in perception of corruptions, the overall happiness score is going to decrease by 1.77 points. For 2009, while perception of Corruption remains statistically significant, Healthy Life expectancy at Birth and Negative Effect are also important.

In order to investigate the factors contributing to the trend displayed in the previous graph, we then move to analyze the distribution of each country. We first calculated standard deviation of Life Ladder in each country from 2008 to 2018. We used standard deviation to measure the magnitude of change of Life Ladder in every country throughout the time period. Based on different standard deviation, we then grouped countries into three categories:

Group A: Country with the largest σcLL

Country name	Life Ladder	Country name	Life Ladder
Benin	0.840790	Guinea	0.779655
Angola	0.815286	Botswana	0.674009
Liberia	0.789411		

Table 4: Group A Results Table for Question 3

Group B (head): Country with the smallest σcLL

Country name	Life Ladder	Country name	Life Ladder
New Zealand	0.076092	Netherlands	0.099135
Australia	0.088326	Belgium	0.105270
Norway	0.090442		

Table 5: Group B Results Table for Question 3

Group C (head): Countries sharing the the most common range of σcLL

Country name	Life Ladder	Country name	Life Ladder
France	0.229625	Mozambique	0.247990
Paraguay	0.229815	Armenia	0.248054
Algeria	0.232318		

Table 6: Group C Results Table for Question 3

We found among all these three groups, Healthy life expectancy at birth has a significantly larger standard deviation than that of other parameters, meaning it changed the most in the given time period. This leads to the conclusion that healthy life expectancy at birth is the most important in affecting change in Life Ladder in countries with the lowest Life Ladder. In addition, for all three groups, generosity, freedom to make life choices, perceptions of corruption are the most significant parameters, which indicates their large impact on life ladder. On the other hand, social support is the least significant, which indicates it does not really affect life ladder.

Part 5: Limitations and Future Work

In the first question, multicollinearity problems still persist in the model after the model selector excludes two variables from the full set of variables. More work could be done to mitigate these issues, potentially by looking to more detailed variables and exploring why these variables are highly correlated to each other. In addition, a linear regression model might not capture higher-order variable relationships. Due to the time constraints, we did not use other models and compare their robustness with each other, and a linear regression model might not be the most suitable model for the research problem after trying more models.

For the second question, limitations for utilizing the k-means clustering algorithm include the ambiguous decision to choose 3 clusters based on a subjective interpretation of the “elbow method” and the randomness of results. First of all, after the construction of an elbow plot, we chose to interpret the “elbow of the curve” as 3 clusters; however, considering that other cluster numbers (e.g. 6) could also be interpreted as the “elbow of the curve”, our 3-cluster results might overlook some important findings present in other k-cluster assignments. Secondly, the k-means clustering algorithm is a randomized algorithm. This means that depending on the initial assignment of centroids, the results of the clustering assignments can be different. In our implementation, we specified the starting location of centroids in order to prevent different cluster assignments of countries every time we run our code, but that initialization step may potentially yield skewed results. According to the results, it is notable how developed countries tend to have a lower generosity score and a higher perception of corruption compared to developing countries.

Namely, observing that developed countries are usually high GDP capitalist countries, it makes sense that they tend to have lower generosity scores. Since developed countries have more politically educated citizens, it also makes sense that people from developed countries are able to be more aware of cases of political corruption. Further literature review on the underlying political models of developed and developing countries can be conducted in the future to serve as theoretical explanations for these results.

For the third question, in both Group A, B, and C, *healthy life expectancy at birth, generosity, freedom to make life choices, perceptions of corruption* are the most significant parameters affecting Life Ladder, and *social support* is the least significant parameter. Further literature review is needed to determine the cause. In addition, the number of countries in Group C (representing countries sharing the most common range of σcLL) is determined by the bin size of the histogram, so changing the bin size will cause slight changes in which countries are included in this group and in the relative weight of each parameter. Finally, the surveyed country has changed each year for the happiness report: while the 2008 report includes 17 European countries with a focus in Northern and Western Europe, like Austria, Norway, Netherland; 2009 data has more than 22 countries with the focus mostly in Eastern European Countries, like Poland, Romania, Serbia. Changes in surveyed countries are likely to create bias for the linear regression model. For future work, increasing the country sample and fixed surveyed country will produce results with less bias.

Code Link:
https://colab.research.google.com/drive/1cJgpIV8V-g7HWGN4xAuHvch1GAHanA9C?usp=sharing#scrollTo=eMOjZDqiDuKK
https://duke.box.com/s/ermhhotcpwv6a5rmu1s7iz3lz5c1w4gx