What Influences World Happiness Scores?

An analysis of Environmental, Demographic and Fiscal Data and 2017 World Happiness Scores

William Burns

Reshma Jayaram

Michael Puerto

Jeanne Reppert

Jazarai Sturdivant

University of North Carolina at Greensboro

December 9, 2019

Introduction

Purpose and Motivation: Each year the United Nations gathers data from countries using six metrics to evaluate the “Happiness” of a selection of nations. The data from these six variables is computed using a weighted average which serves to provide a “Happiness Score” for the countries which subsequently is used to rank the countries. Although this metric is an interesting construct, one questions the validity of the choice of metrics used as there appears to be a bias towards high rankings with European and developed countries and low rankings with countries outside the western hemisphere.

Summary of Questions: Given the concern about the nature of the “Happiness Score” construct, we propose a further examination of world happiness scores considering additional environmental, demographic and fiscal factors. For example, at first glance, we noticed that countries which have significantly colder climates occupy the majority of the top twenty places in the World Happiness score rankings. This would lead us to ask if there are particular environmental factors that lead to greater World Happiness ratings. Likewise, these same countries as a whole have more government spending on items such as healthcare, education and social welfare. Given this observation we would ask if this government spending or financial stability a major factor in World Happiness scores? We wish to investigate if there is a set of additional variables that could lead to predicting similar scores, thus adding to the explanation of why countries are ranked.

Conclusions and Overall Summary: Based on our investigation, we found that there are several other factors that may indeed lead to the “Happiness Scores”. For example, in computing environmental factors we found that cloudiness (and thus degree of sunshine) is not a predictor while temperature was a strong predictor. Overall, our model built with environmental variables produced the strongest overall results. Likewise, our model built with economic measures, while not quite as strong as the environmental model produced results that indicated a reasonable amount of correlation between the variables and our target as did the model built with health related variables. Ultimately, we concluded that a variety of factors may be used to arrive at the rankings given, and that given an entirely different set of variables, it could be possible to redefine the “Happiness” construct altogether.

Data

Motivation and Context: World Happiness Reports provide insight into certain perceptions of countries throughout the world. Our primary motivation is to examine how fiscal, demographic and environmental factors correlate with these average happiness ratings and to investigate whether any additional outside factors might be predictive of 2017 World Happiness scores thus providing an alternative explanation for World Happiness ratings.

Dataset Descriptions: The World Happiness Data 2017 was our primary dataset with the "Happiness Score" variable being the question of interest. The "region" variable from the World Happiness Data (2016) was used as a reference and imported a reference file so that we could compare regions within the database. The Environmental Variables for World Countries, two files containing IMF data, and the Country Statistics - UN Data file were used to compare other country features to the World Happiness Score to ascertain potential correlation between other characteristics of countries as compared to their respective scores.

Data Sources:

World Happiness Data (2017): (https://www.kaggle.com/unsdsn/world-happiness)

The World Happiness Data, acquired from Kaggle, contain "Happiness Scores" for each year (2015-2018) that were compiled through surveys performed by a United Nations study. Participants were asked to rate their countries based on GDP, health, family, government trust, freedom and generosity. These ratings were then weighted and compiled with the countries ranked based on the resulting "Happiness Scores." The 2016 World Happiness dataset was used to get region labels.

Environmental Variables for World Countries:
https://www.kaggle.com/zanderventer/environmental-variables-for-world-countries

This kaggle dataset is a compilation of climate data for countries taken from the google earth engine. Variables include items such as average cloudy days per year, mean annual temperature, annual precipitation, elevation, percentage covered by cropland, and percentage covered by trees.

IMF Data: https://www.imf.org/external/pubs/ft/weo/2019/01/weodata/index.aspx

We imported two descriptive variables (gross national savings and government expenditures, each as a percent of GDP) from the IMF website to give additional national fiscal data for analysis.

Country Statistics - UN Data:
https://www.kaggle.com/sudalairajkumar/undata-country-profiles

This dataset from Kaggle gives a large array of data compiled from the United Nations' UNData website. There are 50 possible variable choices to be investigated for study.

Methods

Data Extracted: Data was gathered by the sources listed previously and was subsequently inspected and cleaned. Because the IMF data was able to be sorted and columns and values were selected directly from the website, there were few adjustments to be made to these two rather small datasets. Likewise, few changes were made to the Happiness Datasets as there were no missing values, inconsistencies or errors within this set. Data was then converted and stored in a shared file in order to load onto the colab site. Dataframe names are as following: worldhappy17,

WorldEnvData, CtProfile, IMF_Nat_Savings17, IMF_Gov17 and CtRg. The two IMF files were later merged into IMF_Merge.

Data Preparation: The Country Statistics dataset required a significant amount of cleaning. First, an anomaly existed in the kaggle dataset where the mobile phone statistics heading was duplicated but the actual column values were not duplicated. This resulted in the columns after the mobile phone statistics shifting and causing inaccurate values. A succinct solution to this problem was not identified in python or sql so the cell containing the additional column name was deleted from the downloaded dataset in Excel with the columns shifting to the appropriate location. This file was then saved and named “country_profile_modified.csv”. This was the only modification made to the initial csv file.

Once imported to Python, thirty-four of the fifty columns for the country statistics file that were not initially selected for use were dropped before constructing the database. Additionally, a few modifications had to be made to the remaining columns. First, the International migrant stock variable contained two values. The “as percent of total population” value was utilized for this dataset and the other number was dropped for all rows. The same problem existed for CO2 Emissions so the million tons value was dropped and the tons value was retained.

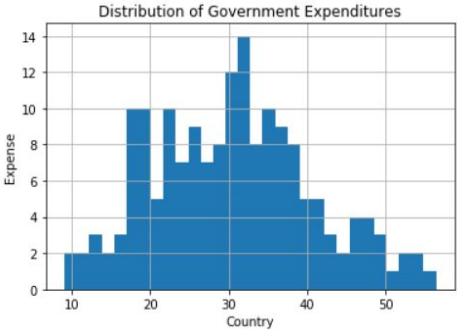
The problem of two values in a column separated by a backslash presented itself in three other variables for the Country Statistics dataset. The Life expectancy at birth variable was split into two columns representing male and females and the Population using improved sanitation facilities was split into two columns representing urban and rural populations. Likewise the Education: Tertiary gross enrollment was split into male and female variables. We were able to utilize string manipulation to convert those values to strings, then split them on the separators making sure to convert the values into their appropriate float values later.

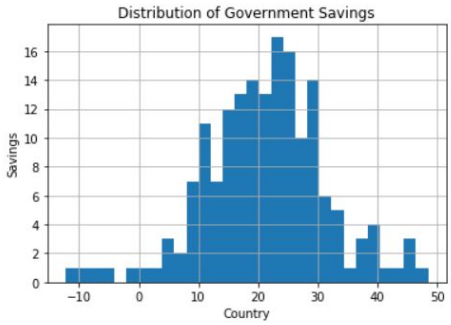
Finally, there were a few randomly scattered missing values that appeared to be unreported for some countries due to different reporting standards or because that characteristic did not apply to the specific entity. For example, countries with populations less than 550,000 people do not report very much in the way of data in the Country Statistics. Thus while it is likely that these values were not missing at random, due to their limited infrastructure to report, the values were declared as null. These missing values are also likely to be not included in the sql searches as well once data tables are joined because the World Happiness Ratings include mostly larger countries. Lastly, several data types in the Country Statistics set were changed from object types to float types. The variables were originally objects because they included characters like “~0.0”, “-99[]”.

World Happiness country names and other data sets: Another cleaning issue that presented itself was that the country names did not match between the various datasets. The World Happiness 2017 dataset was used as the master list for countries

and names were corrected in the other sets to match with the World Happiness 2017 datasets. For example, some files had United States of America whereas the World Happiness dataset had United States. We matched the country names by doing a reverse intersection on the master dataset and the dataset in question. We matched those countries whose country names referred to the same country. All other countries which were not included in the master dataset were not included in the analysis.

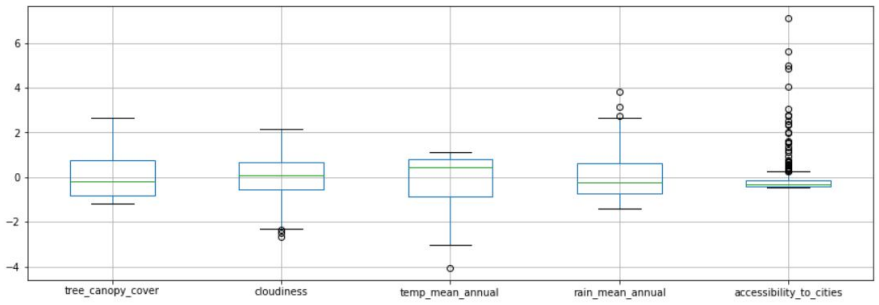
Outliers, Integrations and Transformations: Each dataset was investigated for outliers and transformations. The IMF data was integrated with the variables for government expenditures and national savings rates combined into one dataset. Both the IMF data and the World Happiness sets did not have significant outliers and their distributions were relatively normal so after investigations, no steps were taken to make any changes to these sets.



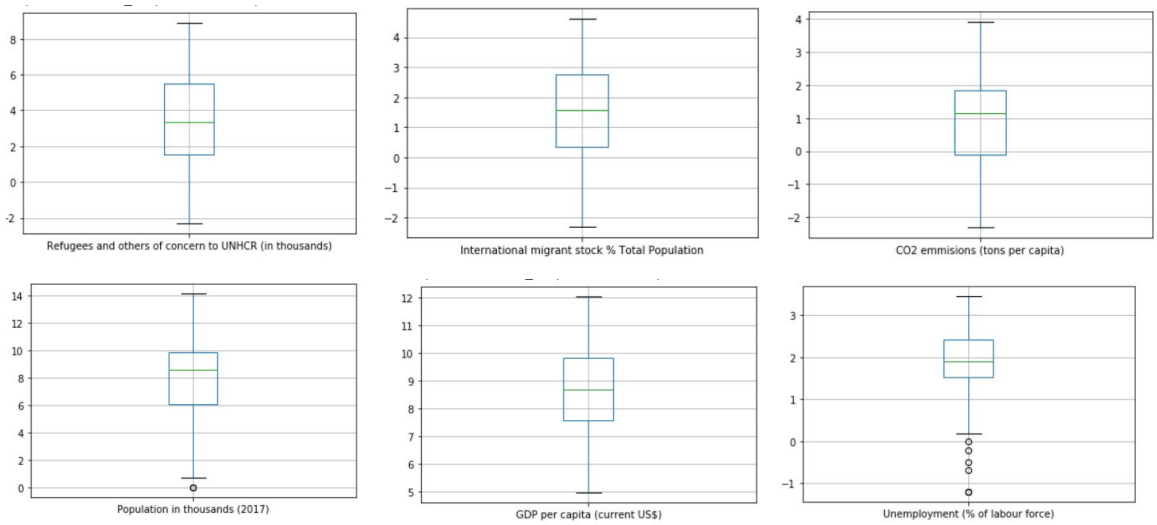


Twenty-two of the twenty-eight variables in the World Environmental dataset were dropped and the remaining ones were investigated for outliers and normality using boxplot visualizations. We observed the most significant number of outliers within the accessibility to cities variable and some outliers in the other variables as well. Understandable, the mean temperature and main rainfall variables were quite skewed.

Several transformations were attempted to correct the distributions of the variables. Ultimately the best outcome with the attempted transformations involved using a z-score normalization. This transformation improved the distribution of four of the five variables. The accessibility to cities variable still showed significant spread and variability. It was thus observed that this variable may not be suitable for possible inclusion in our model.



The Country Statistics dataset was also investigated utilizing boxplot visualizations after dropping columns and cleaning the remaining columns. Several of the variables had either significant outliers and some needed transformations. Log transformations were performed on the following variables: CO2 emissions, Population in thousands, GDP per capita, Unemployment, International migrant stock, and Refugees and others of concern to UNHCR.



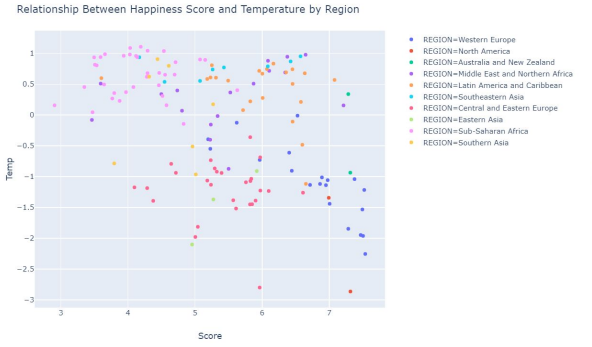
Database Description: Once all files were investigated, cleaned and transformed the appropriate SQL statements were used to assemble the files into a database named “World Happiness”. Initially a relational diagram was generated to visualize the datasets (Appendix A). The final outcome of our database is visualized in the relational diagram included in Appendix B. This database contains five entities with the primary key “Country”. The Country Region utilizes a foreign key to link regions with country names in our queries. Entity names are as follows: IMF_Combined_Data_2017, World_Happiness_2017, World_Environmental_Data, Country_Profile_Variables, and Country_Region.

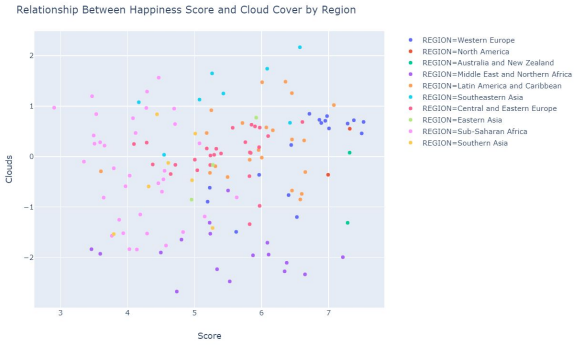
Analysis

We chose to break our analysis up by various topics. In doing so, we can model and assess which of these themes is best for predicting and estimating world happiness. We will take a split train approach, using Pyspark to generate our models using a linear regression method. For the initial queries we chose query methods that achieved a listing of attributes that could be efficiently sorted by region. This permitted visualizations that enabled us to observe trends between countries and their related areas and the studied attributes. To confirm some of our hypotheses that were made based on the first six queries and visualizations we did additional queries that showed aggregations of environmental, health and financial attributes

Environmental Analysis

In assessing the effect of Environmental factors on happiness scores, we chose to model average annual rain, average annual cloudiness, average annual temperature, CO2 emissions, tree canopy cover and accessibility to cities against world happiness scores. CO2 emissions and average annual temperature demonstrated the highest correlations between world happiness scores. Tree canopy cover and accessibility to cities showed virtually no correlation between the variables and happiness scores. We will omit these two variables from our model. Annual mean rain and cloudiness showed a weak correlation to happiness scores so for our model we will include these two variables as well.

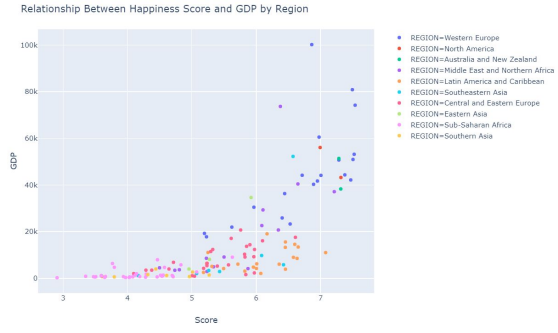




For the pyspark model for our environmental analysis we chose to use the four variables average annual rain, average annual cloudiness, average annual temperatures and CO2 emissions.

Financial

In this theme, we use government spending, government savings, GDP, labor force participation, and unemployment to measure the effect of financial factors on happiness scores. We found that in all cases there was at least some correlation between economic factors and happiness scores.

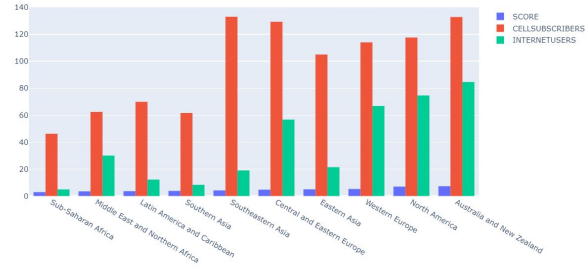


Countries appear in this visualization in order of happiness scores with the lowest score on the left and the highest scores on the right. Interestingly, the countries with the

highest government savings have the highest happiness scores. Savings rates also seem to have some correlation with happiness scores (with a couple exceptions.) Unemployment rates do not appear to have a strong correlation with happiness scores. What is most notable is the large difference between government expenditures and unemployment rates in Western Europe where many of the countries with the highest world happiness scores reside. It is likely that these two variables are useful predictors of world happiness. All of the variables studied in this section will be used in our financial model although GDP will be excluded since this was already accounted for in the UN World Happiness model and thus would be redundant for our model.

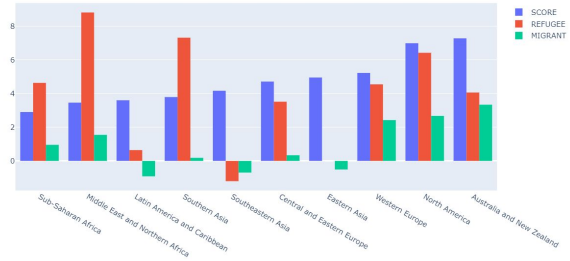
Technology

To assess the effect of technology use on happiness, we chose to model individuals using the internet and mobile cellular subscriptions against happiness.



Population, Migration and Refugees

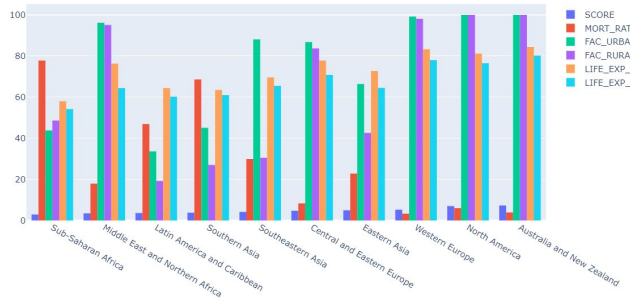
We observed that there is not a strong correlation between population and world happiness scores however, there do seem to be some relationship between migrant populations and world happiness scores (with some exceptions from one region to another.) Higher migrant populations appear to correlate with higher happiness scores. The refugee population variables shows a great deal of variation in terms of its relationship with world happiness scores. Due to these variations we will not include this variable in our model.



Health

For this category we investigated rural and urban sanitation, life expectancies for males and females, mortality rates and percent of GDP spending on healthcare as it related to World Happiness scores. We found most of the variables to be correlated. Of particular

interest was the one outlier point in the GDP spending on healthcare that relates to spending in the United States in the scatterplot. Additionally, we found that while there are some correlations with urban and rural sanitation practices, there are some confounding factors within this variable. Some regions have consistent scores with urban and rural sanitation while other regions have a greater disparity between rural and urban sanitation. For this reason, we are unsure as to whether these variables will lead to a meaningful contribution to a model without further data being gathered. We will not use these variables for our model at this time for this reason.



Results

Based on the analysis of the five categories investigated in the above discussion, three groupings of variables were constructed to utilize in linear models. The first dataset, Environmental Factors, was made up of annual mean rainfall, annual mean temperature, and CO2 emissions. The next dataset, Financial and Technology Factors, was made up of internet usage, cell phone usage, average national savings, average government expenditures, and unemployment rates. The final dataset, Health/Demographics, was made up of male and female life expectancy, migrant populations, health as total expenditure of GDP, infant mortality rates, and population in thousands.

The following results for the three models were observed:

Environmental Factors: The coefficients of the environmental model variables (mean annual rainfall, mean annual temperature, and CO2 emissions) were 0.3501, -0.1537, and 0.5259 with an intercept of 5.066494. The model had an RMSE of 0.50 and an r squared value of 0.757.

Financial and Technology Factors: The coefficients of the financial and technology factors (cell phone subscribers, internet users, government expenditures, national savings rates, and unemployment rates) were 0.0014, 0.0305, 0.0026, 0.0061, and -0.0316 with an intercept of 3.779801. The model had an RMSE of 0.686 and an r squared value of 0.552 .

Health and Demographic Factors: The coefficients of the health and demographic variables (life expectancy - female and male, health gdp, infant mortality rate, and

population) models were 0.0718, 0.0279, 0.0528, 0.0041, 0.1678 and 0.0964 with an intercept of -3.405005. The model had an RMSE of 0.643 and an r squared value of 0.576.

The model with the variables related to environmental factors had the most predictive significance with an r squared value of .757 which accounts for about 76% of the variability in the happiness score in the model. The model with the next most predictive significance was the model with variables related to health and demographic factors. The variables in this model accounted for 58% percent of the variability in happiness scores in the dataset. The last model, with variables related to financial and technology factors had the least significant r squared value with the model variables accounting for 55% of the variability in the happiness scores in the dataset. The lowest RMSE score was also found in the model built with environmental scores (0.50) versus our least effective model with an RMSE of 0.686.

Conclusions

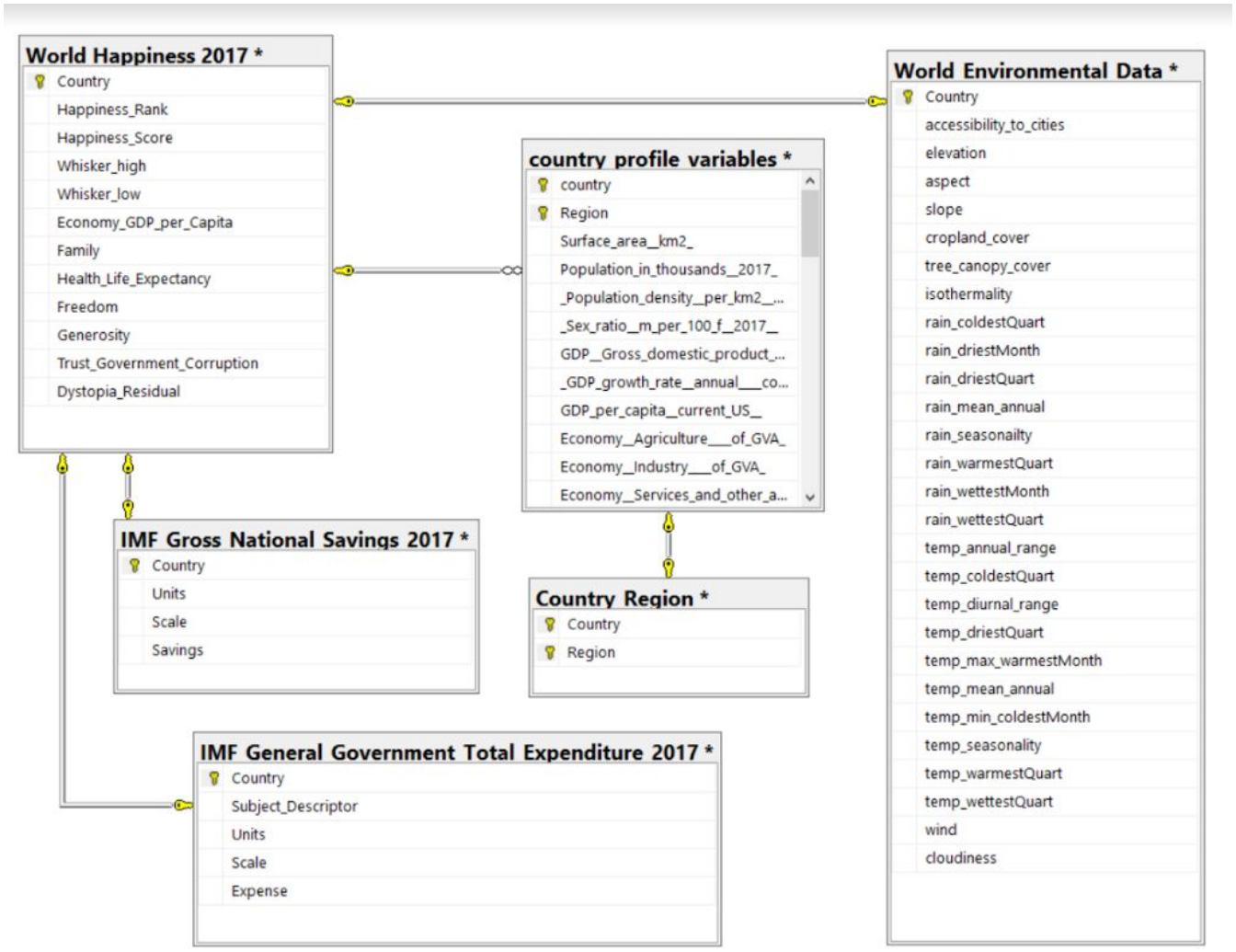
This research project aimed to unpack the “Happiness Score” construct used by the United Nations to rank countries. We wished to investigate some of the factors behind why countries received higher rankings and why other countries received lower rankings. By investigating the relationship with variables from the World Environmental sets, Country Profile set, World Happiness Set and IMF Data, we were able to pinpoint variables that could be used to predict world happiness.

The results detailed in the prior section indicate that the strongest correlation between World Happiness scores and other factors investigated in this study was between environmental factors. This confirms our initial hunch that weather appeared to be correlated with the overall listing of country rankings with the countries with the lowest overall mean temperature having some of the highest rankings. Additionally, we found that there was a strong relationship between health measures and happiness ratings with higher levels of health correlating with higher happiness scores although this was a somewhat weaker correlation with these factors accounting for 58% of the variability in Happiness Scores. Interestingly, there is still a case to be made for financial and technology measures to be correlated with Happiness Scores given an r squared value of .55.

While clearly, none of these values can be used to argue for causality, it is interesting that there are clear correlations between many other social, political, economic, and environmental factors and World Happiness Scores. Ultimately the conclusion of this study would be to recommend that the interpretation of World Happiness Scores as constructed by the United Nations be viewed with some level of caution as there may be other underlying factors that may lead to the results of this measure. It would also be recommended that there be consideration to expanding the number and types of metrics investigated in order to portray other views of the World Happiness construct that could potentially lead to significantly different results.

Recommendations for future studies: This study provided a preliminary investigation of world happiness data as it related to other demographic data in various datasets. There are a variety of future avenues of study that could be taken to further insights into this study. For example, while the educational data was considered for study, many nations do not reliably report their educational data and thus this metric would not have proved to be a valuable measure of our target despite being an important variable. Efforts could be made to acquire more data in this area in order to analyze the impact of education on World Happiness Scores. Additionally, certain environmental data may need to be investigated over a longer period of time in order to account for local variations in annual weather patterns. Finally, some demographic data would need to be reworked in order to have a meaningful contribution to our study. For example, refugee data, while interesting to study, varies so significantly (in our case we noticed an effect in the Middle East that seemed to skew the results that is likely to be attributed to the war in Syria) that this data as well might need to be studied over a longer period of time.

Appendix A - Initial Relational Database Diagram



Appendix B - Final Relational Database Diagram