World Happiness Index Week 6 Report

Evan Carr, Kairui Wang, and Lingtao Zeng,

***Abstract*—The world is a rapidly changing place. Among the fastest changing aspects are those relating to how people communicate and interact with each other. Finding that each country's life circumstances, including the social context and political institutions were such important sources to better understand people's living environment among different regions.**

I. Purpose

THE purpose of the project is to identify which sections of the world occupy the extremes of the happiness index. Once the regions are identified for the happiest and least happy parts of the world the key aspects differentiating the populations from the rest of the world will be identified. Highlighting the variables that have the highest correlation to the happiness index will allow authorities to focus on improving these areas to maximize their countries overall well-being. Finally, the project will observe which regions have changed their happiness levels the most and describe if there is a pattern for drastically improving or decreasing the happiness index. For example, most recently COVID-19 changed the world from 2019 to 2020 so searching for a worldwide impact during this time period could give an indication the pandemic has had on the happiness of the world.

II. Data Information

For analyzing the factors that influence world happiness, several data sets were found on Kaggle [1]. Each data set shows the ranking of the happiness score of each country and indicates those countries' GDP per capita, social support, freedom score, generosity score and some other features that may influence people's happiness. The happiness reports from 2015-2019 were available for analysis. The happiness score and ranking used data from Gallup World Poll. The poll asked each respondent to select a level from 0-10 as the score that they estimate their happiness to be and give a score on economic production, freedom, generosity... for how much they think this factor contributes to their happiness. All of the data in each category is numerical.

There are also scores other than the consistent yearly factors that were considered for world happiness such as family and freedom. For example in 2015, each country has an evaluation on "Dystopia Residual", which is an "is an imaginary country that has the world's least-happy people" [2]. Dystopia is the opposite word of Utopia. However, not all the reports contain the score for dystopia. There are also some factors like "Whisker.high" and "Whisker.low" in 2017 that does not have corresponding variables in 2018. Thus, in the analyzing steps the inconsistent variables will not be considered as factors.

The research subjects and contents for world happiness report are not exactly the same in each year, so there are some differences between the countries and the variables of each table. For example, there were 156 countries in the table of 2019, and 153 countries in the table of 2020. The variables for each table are also not consistent, especially the difference of names for one feature. For example, in 2015, happiness score is written as "Happiness Score", and in 2019 it is written as "Score", while in 2020, it is written as "Ladder score". Also, there are different names for health scores, freedom scores and other features. What's more, some tables contain a few more variables than others, like 2015 has the score for family, but in 2020 happiness score does not count family. Instead, it has "upperwhisker" and "lowerwhisker" as features.

The Global Database of Events, Language and Tone (GDELT) is an open source database available through Amazon Web Services that tracks events that impact the stability of every country around the world [3], [4]. Since 1979 the database has tracked more than 500 million world events [3]. The event record includes a vast array of information such as the type of event, the countries involved, the main people involved, and the Goldenstein scale. There are 267 significant events types recorded, for example protests, each of which carries a weighted Goldenstein scale value [5]. These events are placed on a scale from -10 to +10 that quantifies how intense the impact could have to help or hurt the stability of the country [5]. Some of the most extreme examples are a military attack worth -10 to the stability of the country or the acquisition of economic aid, worth +7.4 [5]. Although this number provides a baseline of the significance of the event, there is no variable for the size of the event [6]. For example, a protest of 10 or 10 thousand participants is worth -2.4 despite the large differential in participants.

To supplement the Goldstein scale, GDELT also tracks print and web media in over 100 languages from every corner of the world [3]. Starting in 2015 the database has identified approximately 900 million media items related to the events collected. This dataset collects information on the event discussed, the company publishing the article, the direct link to the article if available, the length of discussion relevant to the event, and the tone of the writing towards the event. This allows incredible insight into the perception of the event. If a protest occurs but has a neutral or insignificant tone, the event was most likely minor or had a low impact. On the other hand if the protest has a strong tone in the media, the event could be more impactful for the direction of the country or world. The tone is scored on a -100 to +100 scale from extremely negative to extremely positive [6]. Since each media mention is also recorded individually, the number of mentions could also lead to insight into how important the event was to the population. The expectation is an event that triggers increased media mentions along with a strong tone will be the events with the highest impact.

Turkmenistan	Paraguay	Paraguay	Libya	Serbia	Moldova
Mauritius	Romania	Hong Kong S.A.R., China	Philippines	Moldova	Tajikistan
Hong Kong	Estonia	Philippines	Honduras	Libya	Montenegro
Estonia	Jamaica	Serbia	Belarus	Montenegro	Russia
Indonesia	Croatia	Jordan	Turkey	Tajikistan	Kyrgyzstan
Vietnam	Hong Kong	Hungary	Pakistan	Croatia	Belarus
Turkey	Somalia	Jamaica	Hong Kong	Hong Kong	North Cyprus
Kyrgyzstan	Kosovo	Croatia	Portugal	Dominican Republic	Greece
Nigeria	turkey	Kosovo	Serbia	Bosnia and Herzegovina	Hong Kong S.A.R., of China

In this case, some data sets use "Hong Kong" while others use "Hong Kong S.A.R. of China. All will be denoted as "Hong Kong" for this analysis

III. Data Summary and Cleaning

In order to compare the happiness score and the factors for different years, we have to make the variables consistent. First, we have to make sure that all the country names are written in the same style so it is easy for selecting and filtering in further steps. For example, some tables use "Trinidad and Tobago", and the others use "Trinidad & Tobago". We choose "Trinidad and Tobago" for all the tables instead of "Trinidad & Tobago" because we do not want any symbols like "&" in the variables.

First, the country column from each table was copied and pasted into a new file for comparison. Disputable country names were checked and revised the directly in Excel. Then, python was used to contrast the set of names in each table and delete countries that are not researched for all 6 years. Finally, 146 countries remain that are useful. For the features concise format was used naming, like changing "Healthy life expectancy" to "Health", "Freedom to make life choices" to "Freedom". There are no null elements in the tables. After deleting the unnecessary columns, six factors remained that will be used for further analysis, which are "GDP", "Health", "Social support", "Freedom", "Generosity", and "Corruption". Adding the basic variables "Country", "Happiness Score", and "Happiness Rank", we have 9 columns and 146 rows for each table.

Some original tables have standard errors as the variables while the others are not. We don't know how the data was calculated, so we cannot directly use those data, but we can calculate mean and standard deviation by ourselves for visualization and further analysis.

In order to study the relationship between world happiness, the Goldstein score, and the tone of the media, the yearly average Goldstein score and tone score of each country will be used to evaluate the connection. Since the open source data is only available through the end of 2019, there are 1113 data points for both the Goldstein and media tone scores representing the average score for each country for each of the 5 years. The overall average Goldstein score is approximately 0.898, which indicates that the world as a whole is generally stable as well as the standard deviation of 1.02, showing a low variance in the stability of different countries. The 95% confidence interval for the mean of the Goldstein score is from 0.838 to 0.958 indicating that there is a strong degree of confidence the average score is greater than 0, supporting both initial assumptions. The overall distribution appears to be normally distributed albeit with a low variance.

The average tone of the media score is -2.13 indicating that despite the strong sense of stability from the Goldstein score, the media tends to portray these events negatively. The standard deviation is 1.44, indicating there is a larger discrepancy between the portrayal of world events than compared to the Goldstein scale. The 95% confidence interval for the media average ranges from -2.13 to -2.05 validating that the media tends to portray events with a higher negativity than the Goldstein scale. The tone averages also resemble a normal distribution with an increased variance, fattening the bell curve. Using the Goldstein averages to predict the tone of the media using a linear model produces a slope of 1.03 and an intercept of -3.05 each of which having a p-value of less than 2×10^{-16}. The slope shows a near direct relationship between the change in Goldstein average and change in the tone of the media. The close link compliments each other pushing the idea that both are good scales for judging the impact of an event. However, the extreme intercept demonstrates a strong digression between the Goldstein and media tone. When the Goldstein scale predicts an event to have a negligible impact on the stability of a country, the media portrays the event negatively. This offset is significant enough that despite an overwhelming 84% of the events scored positively on the Goldstein scale but only 7.3% scored positively in the media.

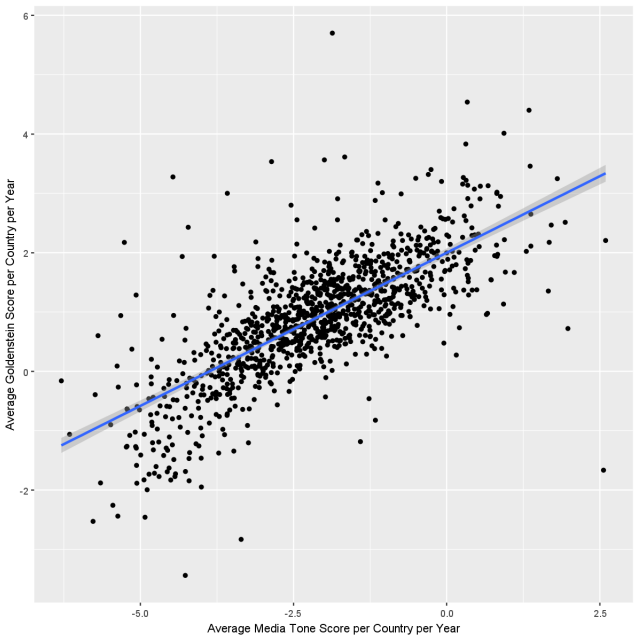

Fig. 1. Goldenstein Score vs Media Tone Score

A brief analysis comparing the Goldstein scale for world events to average yearly world happiness reveals there is a 0.31 to 1 relationship between the scale and happiness. The connection produces a statistically significant p-value of 4.37×10^{-10} strongly establishing the relationship. However, the trend also has a very low R^2 value of 0.054. So even though there is a statistically significant trend between the Goldstein scale and world happiness scores, very little of the variation in world happiness is explained by the Goldstein scale. Similarly, using the media tone scores to compare the impact of media on world happiness produces a slope of 0.198 with a p-value of 2.64×10^{-8}. Again, even though the trend is statistically significant the model accounts for very little of the variation due to a low R^2 value of 0.043.

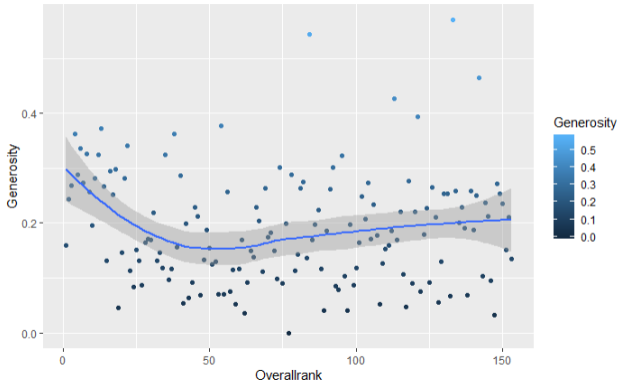

Fig. 2. Generosity vs Overall Rank

Country	R2020	Overall Rank	Compare Rank
Benin	86	136	50
Guinea	102	140	38
Kosovo	35	66	31
Niger	103	134	31
Congo (Brazzaville)	88	114	26
Liberia	124	149	25
Bosnia and Herzegovina	69	93	24
Ivory Coast	85	107	22
Philippines	52	71	19
Slovenia	33	51	18

IV. Data Analysis

The R programming language was used for initial data analysis. First, is to find the relationship between the 6 variables and their overall rank. A linear model was used to represent the relationship. The result is : there are 5 variables that have a negative linear relationship between ranking, except the "Generosity" graph shown below. Since this variable is different than the others, more analysis will be conducted in the future.

Analysis was also used to find which country improved the most on the happiness index in the past few years.

The initial hypothesis is these countries are not significantly affected COVID-19. Take Benin (Benin, officially the Republic of Benin and formerly Dahomey, is a country in West Africa) as an example. Benin has a very small number of COVID-19 patients compared to other countries around the world. There are only 5434 cases in Benin before February 28th, 2021. The rest of the countries are listed in the graph. All these countries have not faced a strong impact from COVID-19.

V. Model Proposal

Data visualization will be used on the data sets to see the trends of people's happiness in the last 5 years and their thoughts towards each social factor. Further investigation into why the variable "Generosity" is different from others since the relationship is not linear. he impact of the factors influence happiness, so the countries' happiness can be predicted through their social situation and development in further years. To achieve this goal, the data will be split into training and testing sets, and try different models like logistic regression and random forest because the data is supervised and numerical.

Another focus will be on 2019 and 2020 data to determine if COVID-19 impacted the happiness index significantly. Identifying a single factor that improves it's ranking significantly which might not be affected by COVID-19. Next step is to find some countries that have a lower ranking than before and find if there is a relationship between their rank and COVID-19 cases.

References

[1] Mar 2020. [Online]. Available: https://worldhappiness.report/
[2] "World happiness report," Nov 2019. [Online]. Available: https://www.kaggle.com/unsdsn/world-happiness
[3] "The gdelt project." [Online]. Available: https://www.gdeltproject.org/
[4] "Global database of events, language and tone (gdelt)." [Online]. Available: https://registry.opendata.aws/gdelt/
[5] "Goldstein scale for weis data." [Online]. Available: https://web.pdx.edu/~kinsella/jgscale.html
[6] "The gdelt event database data format codebook v2.0," Feb 2015. [Online]. Available: http://data.gdeltproject.org/documentation/GDELT-Event_Codebook-V2.0.pdf