Final Project Report

Title and author(s):

Title: “Happiness Factors”

Author(s): Xinyue Ma, Ziqi Chen

Summary of research questions:

1. “ What is the distribution of different Happiness Scores among different countries in the world? ”

   We are interested in this question, since people from different countries have different Happiness Scores based on culture diversities, GDP differences, etc. Thus, we are curious about the distribution of Happiness Score in the world. In other words, after analyzing data and constructing a visual map , we can identify which country/ region has a relatively highest Happiness index than others.

   Result: From the Happiness Score map, it is clear that North America and Europe have higher happiness scores than other regions. Besides, Finland is the country that has the highest happiness score.

2. “What is the relationship between Economy(GDP per capita) and Happiness Score for the top-20 Happiness level countries in 2019? ”

This question is valuable to investigate, because economic growth is a goal that most countries are pursuing at present. Thus, we are concerned about the change in Happiness Score within the change of economic development. For data analyzing, we are trying to collect data about GDP and Happiness Scores for top-20 Happiness level countries in 2019 and then use the plotly scatterplot to analyze the possible relations.

Result: From the graph, we can tell the countries with top20 happiness scores have similar GDP levels in 2019 except a few outliers. We noticed that Iceland has a super low GDP compared to other top20 happiest countries.

3. "What are the correlations between Happiness Indicators?"

We are curious about this question because correlation can help us to find the relationship not only between the score and indicator, but also one indicator with another indicator.

Result: Indicators corresponding with country capabilities such as (social support and GDP) relate closely with each other. However, subjective indicators such as generosity and freedom to make choices have small correlation with other indicators.

Motivation and Background:

As the development of society, people are increasingly caring about the level of happiness. Instead of pursuing quality of life blindly, an increasing number of citizens begin to seek ways to facilitate their happiness. With investigations on the happiness level, happiness score is increasingly popular for governments, organizations and civil society to use as indicators to inform their policy-making decisions. As college students, we are also in the stage of anxiety due to our cultural background, living standards, health, etc. Thus, we are interested in the happiness levels around the world.

For the first question, " What is the distribution of different Happiness Scores among different countries in the world? ", since we are international students, we want to view the overall happiness levels varied in different countries and different continents. We are curious about whether the happiness level matches what we originally considered in our culture acknowledgement.

The second question is the relationship between Economy and Happiness Score among the top 20 countries that have high happiness scores. Knowing this question helps us to recognize whether the home country's Economy indeed improves our happiness scores. Since we were both born in the late twentieth century, we have experienced the biggest progress of the economy in China. We indeed feel increasingly happy till now, but we are curious about whether the GDP is the key factor of facilitating happiness level. By investigating the relationship between GDP and top 20 happiness scores countries in the latest data (year 2019), we are likely to find the question for our curiosity.

The last question, "What are the correlations between Health Life Expectancy with other Happiness Indicators?"digs more in-depth investigation for

us to find the correlations between one of the happiness indicators ("life expectancy") with other indicators. As we mentioned earlier, happiness data is useful for the government to make policies. For many societies, governments are working to invest a lot into health welfare to increase citizens' happiness level. Will these policies have an impact over other aspections in other ways? The research for this question could present the correlation between healthy life expectancy and other factors so that will imply policy potential impacts.

Dataset:

URL:

1. https://www.kaggle.com/unsdsn/world-happiness?select=2019.csv(happiness level 2019)
2. https://data.worldbank.org/indicator/NY.GDP.MKTP.CD

   (2019 world GDP)
3. Ne_110m_admin_0_countries.shp

   (geo data)

The first dataset we chose is "The World Happiness Report 2019" about global happiness. File in this dataset ranks a sample of about 155 countries by their happiness levels in 2019 and lists out the extent to calculation in several dimensions. The second dataset we chose is the worldwide countries' GDP in 2019 from "The World Bank". The countries' GDP values are calculated in U.S dollars. Furthermore, the last dataset is the geospatial data for the countries in the world. The 'geometry' column in this file includes each country's geographic shapes such

as Polygon/MultiPolygon. By joining this geospatial shape file and previous happiness scores dataset together, we can plot a happiness levels map.

Methodology:

For this project, here are three parts described to analyze these problems.

Part1:

To analyze the first research question, we are going to use two datasets called “World Happiness Report 2019” and “Geo_data”. Firstly, we need to filter the relative column in both dataset to avoid complicated data. Before merging two files, we recognize the prompt of the “World Happiness Report 2019” dataset demonstrates USA as the United States, which is not the same as the name of USA in the Second dataset. Thus, we correct the name before merging . We use the merge function to left join one file with another. Finally, we will use plotly to draw a Choropleth Map that contains specific Happiness Scores in the world by matplotlib. We will use distinct colors to represent the level of Happiness Scores.(the higher the Happiness scores, the darker the color)

Part2:

To analyze the second research question, we are going to use the datasets of “World Happiness Report” and “World GDP”. Since we are only interested in countries with high levels of happiness in the latest year in the dataset, we are going to clean our dataset first. We will only tackle the file of “2019.csv” and we will only keep the columns of “Country” and “Happiness Score” in this dataset. In

order to have access to the high level happiness countries, we will sort out the Top 20 Happiness Score in this dataset. After we clean the first dataset from the “World Happiness Report”, we will need to merge this dataset with “World GDP” to get the information about these top 20 countries' economic levels. By using Plotly, we will plot a Bubble Chart where the happiness score is represented by the bubble size. The x-axis will be the country names and the y-axis will be the GDP levels.

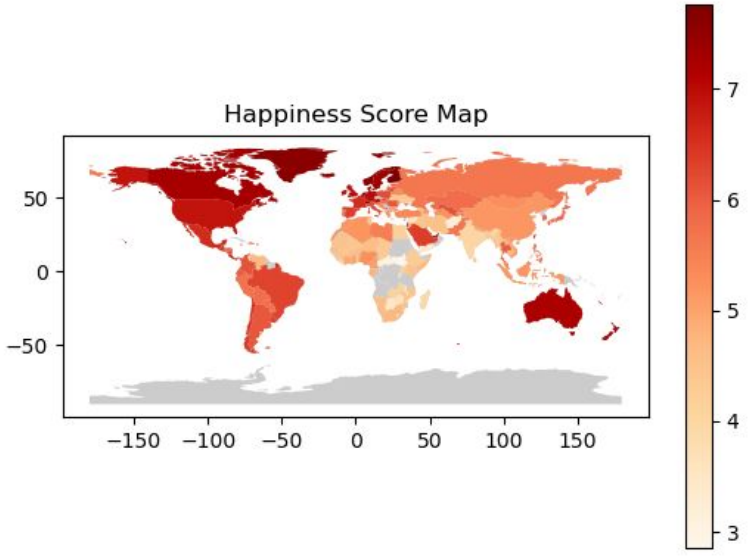
Part3:

To analyze the last research question, we will use “World Happiness Report 2019.csv” to get the correlation between Happiness indicators. To begin with, to reduce the computation of unnecessary data, we will make a copy of a dataframe that only has columns we need for later calculation. Furthermore, we need to convert each column list to a numpy array, and expand each of them together. After that, we calculate the correlation by using corr() funcion, and then store them in a new numpy array that has the shape (len(variables), len(variables)). Finally, we are going to use “Heatmap” from plotly to visualize the correlation between each factor. A darker color in Heatmap will demonstrate a high correlation between two variables. From the Heatmap, we can observe the relationship between any two Happiness variables and improve our analysis on Happiness scores.

Results:

Part1:

Visualization:

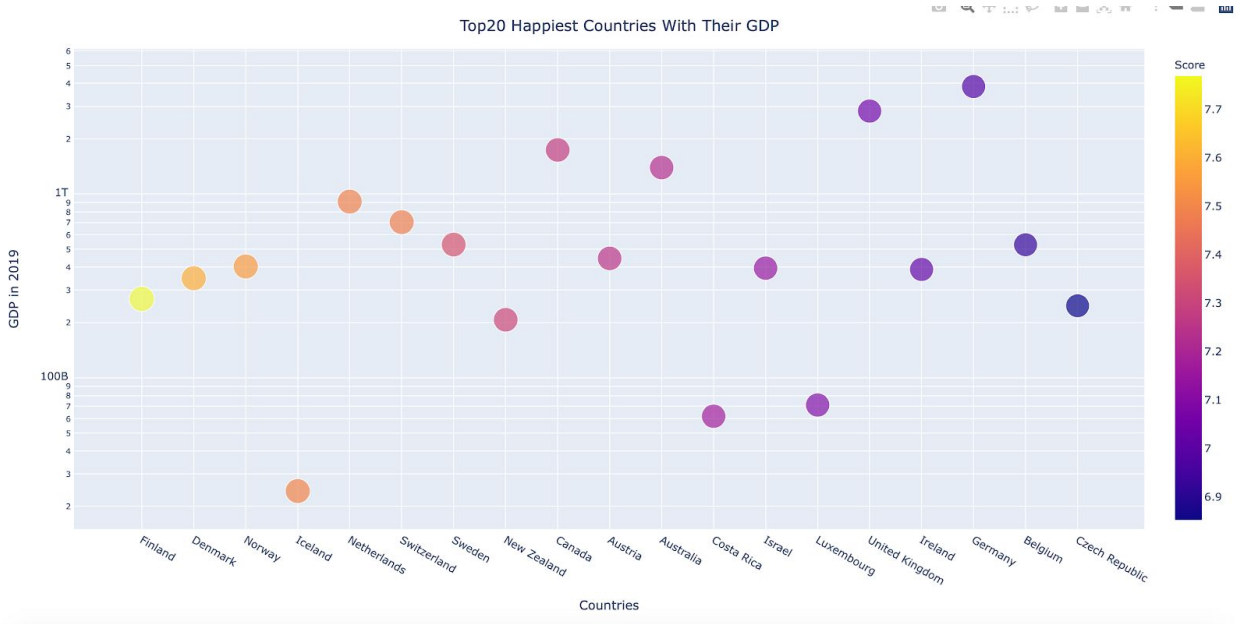


From the Choropleth Map generated by matplotlib, we find North America and Europe has average higher happiness scores than other regions. This result is reasonable, since these regions have many developed countries, which may lead to high happiness scores since people can have a higher life standard than other countries. Besides, for a particular country, Finland has the highest happiness index in the world. Thus, according to the world map, we can also recognize that the size of the region does not correlate with happiness scores. However, we Find there is

missing data for the Antarctic Pole and some countries in Africa, since the colors of these regions are grey. Thus, our analysis may not be completely thorough.

Part2:

Visualization:



Before drawing the graph, we imagined that the GDP level will present a decreasing trend as the happiness scores drop. However, in the graph, we could observe that the top 20 happiest countries shared a similar level of GDP in 2019 roughly. Canada, UK and Germany have higher GDP than other countries. Surprisingly, in the top 5 happiest countries, Iceland has a much lower GDP than the overall level among these 20 countries. By researching the analysis of Iceland 2019 economy, we noticed that the possible reasons for such low GDP may be the tourism level decreased. And also, since fishing is the only large scale of output for this country, Iceland is vulnerable to the economic shocks from all around the world. However, we noticed that there are a couple of reports discovering that local

people are optimistic about their economy and they think the economy will recover by several economic swings. Probably this leads to the high happiness level in this country in spite of lower level GDP.

Part3:

Visualization:

From the Heatmap generated by plotly, it is clear to recognize that 'Generosity' has much smaller correlation with score, GDP, social support and health life expectancy. Since correlation measures the extent of connection between variables, we can get the conclusion that Generosity has almost no connection with the indicators mentioned previously. This result may be caused by the type of indicators: Objective indicators and Subjective indicators. We define indicators with national capabilities as objective indicators, and a person's own feelling as subjective indicators. Thus, citizens might think the generosity

(subjective indicator) does not relate to the other objective indicators. One thing we need to print out is the data accuracy, since the standard for each person to measure subjective indicators might be different.

Consequences and implications: If a government wants to enhance the citizens happiness level, by investigating the "Score" column on this map, we would encourage the government to consider setting up policies in regard to the aspects from enhancing economic level, providing more social welfare like unemployment insurance, and improving the level of healthcare. These aspects show a larger correlation with happiness score. As well as these aspects have large correlation with each other, when governments enact policies, they should pay attention to the potential changes over other aspects. For example, if a government wants to improve social support by enhancing unemployment insurance, this may cause a larger unemployment rate since people find jobs with less efforts and thus a lower GDP may occur and may affect happiness level adversely.

Challenge Goals:

1. Multiple Dataset:

We are going to merge datasets in generating world happiness score 2019 distribution and finding relationship between GDP and scores in 2019. Using multiple datasets can give us richer analysis and dig deeper into the happiness factors. We think our project completes this challenge goal, since two of our tasks use a merged dataset to analyze. In the first task, we join the geospatial data and

score together in order to plot the map. And then, merging gdp.csv and scores to plot the scatter plot in the second task.

2. New Library:

Before writing the program for research question 1, we planned to use a new library called plotly to plot a Choropleth Map that contains specific Happiness Scores in the world. However, we scale back to use matplotlib to plot this figure. We searched information about plotly function and watched tutorial videos online, then found we need Geojson data to plot in plotly rather than geospatial data we had. Thus, we changed the decision, since our data did not meet the requirements of plotting a Choropleth Map by using plotly. Finally, we chose to use the matplotlib to plot the map based on geospatial data.

For research question 2 and 3, we think we fulfill the task of using "ploty". We plotted a scatter plot for GDP and happiness relations and a heatmap for correlation. The heatmap is really helpful for discovering the correlations between 7 elements.

Work Plan Evaluation:

Work Plan from Part 1:

This project is a collaborative work. And, we will use Github for each of us to access the content of our project. In order to maintain the manipulation of it

works, we construct a repository and add my group members as collaborators. We divide this project into three tasks.

For research question 1, Xinyue will work on this part and finish the visualization of the Happiness Map. To begin with, it will roughly take one hour to read international news/articles about Happiness Scores in different countries. Understanding different people's perspectives on Happiness Scores can improve our analysis on the first research question. Secondly, Xinyue will spend about another hour to load and clean the data we need for plotting the map. The clean-data process includes column extraction and merged data frame construction. Then, Xinyue will spend almost half an hour investigating a new library called library.It is significant to spend some time on reading the Document of this library and then try some examples in Jupyter before applying to our project. Finally, it probably takes 2 hours to visualize the Happiness Map through a new library called plotly.

For our research question 2, Ziqi will work on this part mainly. To clean up the data firstly, it will probably take one hour to read in, filter, sort and merge the datasets. Getting the useful dataset with the top 20 countries in 2019, Ziqi will spend about two and a half hours in calculating the relative numbers and visualizing the dataset with Plotly. After getting the visualization graph, Ziqi will spend about one hour searching online for special events that could potentially impact GDP or happiness level in these countries in 2019. And finally, it will take about half an hour for Ziqi to sum up the information from the graph and potential events.

For our research question 3, Xinyue and Ziqi will work on this problem together. Ziqi will spend about one hour and a half cleaning up the datasets by filtering out each of the useful columns and converting them into numpy arrays separately. Then, Xinyue will spend roughly one hour doing the correlation calculation and storing values into another numpy array for later plotting. Then, Xinyue and Ziqi will spend two hours working together to plot the heatmap data by Plotly and analyze the map with research on some real-world examples.

After finishing each corresponding task, Xinyue and Ziqi will push the code to github individually and they are able to check the work, discuss and test the minor problems together. In order to perfect the project, they will also need to put their own ideas for more concise code and format.

Evaluation:

For the part of estimating the time of clearing data, we are roughly correct. But for part 3, we spend more than two hours in understanding the data and figuring out what will be the correct numpy array about the correlations. We communicate by zoom chatting on the work which really enhances the efficiency. But in the part of estimating the time used for plotting, we underestimated it. Since this is our first time using plotly, we spend a great amount of time on discovering it and getting a handle on it. Finally, by the help of Github, we can modify the code easily and communicate with it in a more efficient way.

Testing:

In previous homeworks, we found the assert equals functions are pretty useful. So we import the assert equals from “cse163_uitl.py” here as well to test

our code. For each of the research questions, compared with the huge dataset, we may only care a couple of factors in the dataset. For the first research question, we need to test whether we get a cleared data frame only with the information about the geolocation of countries and the happiness scores. For the second research question, we need to test whether we have only 20 rows since we focus on top-20 happiness scores countries here and we also need to check whether the merged dataset only covers the number of factors we want. For the third research question, we need to check whether the numpy array we filtered has the correct number of shapes as we want to plot on the graph which should be 7 elements on the x-axis and 7 elements on the y-axis.

By looking at the graph we draw for research question 3, if the graph is correctly plotted, we should observe a diagonal with correlation 1 since in the diagonal we are calculating the correlation with the factor itself.

Collaboration:

No students or other people helped us besides our group mates. We obtain some information from online such as “RealPython” and “Plotly | Graphing Libraries”.