Project 2: Final Report

INFO 3300: Data-Driven Web Applications

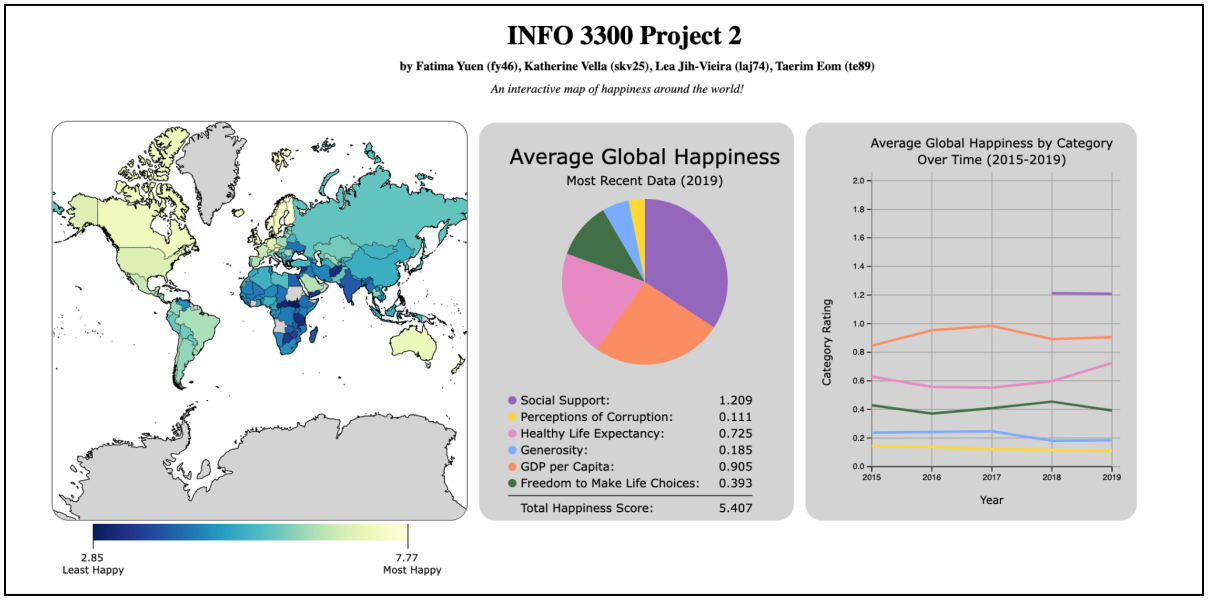
November 15, 2022

Fatima Yuen (fy46), Katherine Vella (skv25), Lea Jih-Vieira (laj74), Taerim Eom (te89)

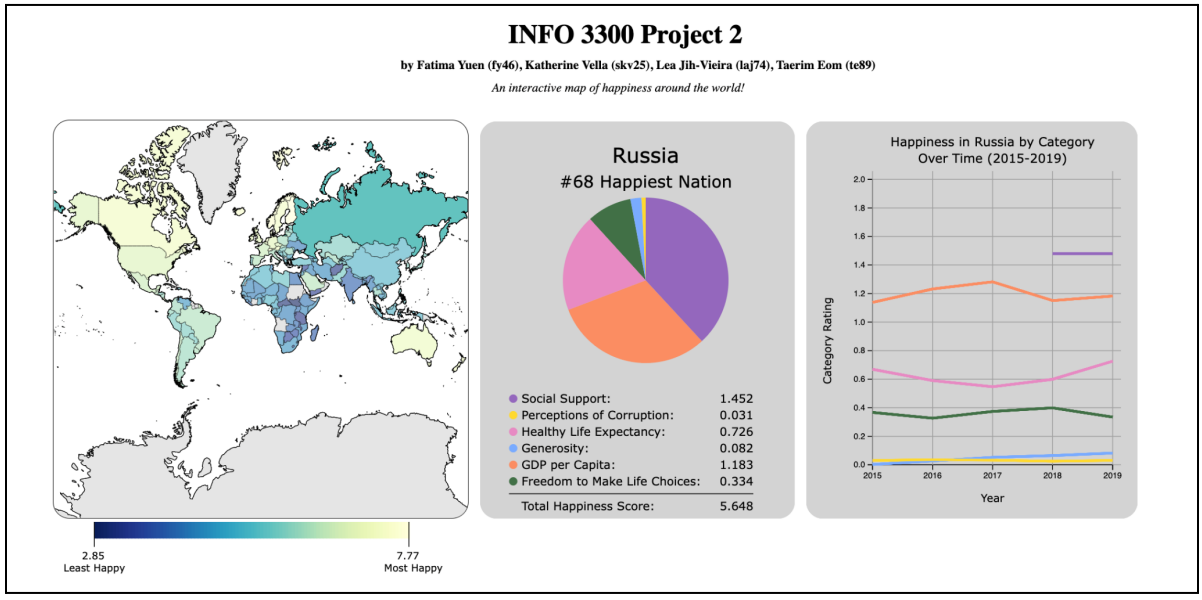
A visual analysis of happiness around the world by country.

Global Stats View (initial view and upon double clicking country):

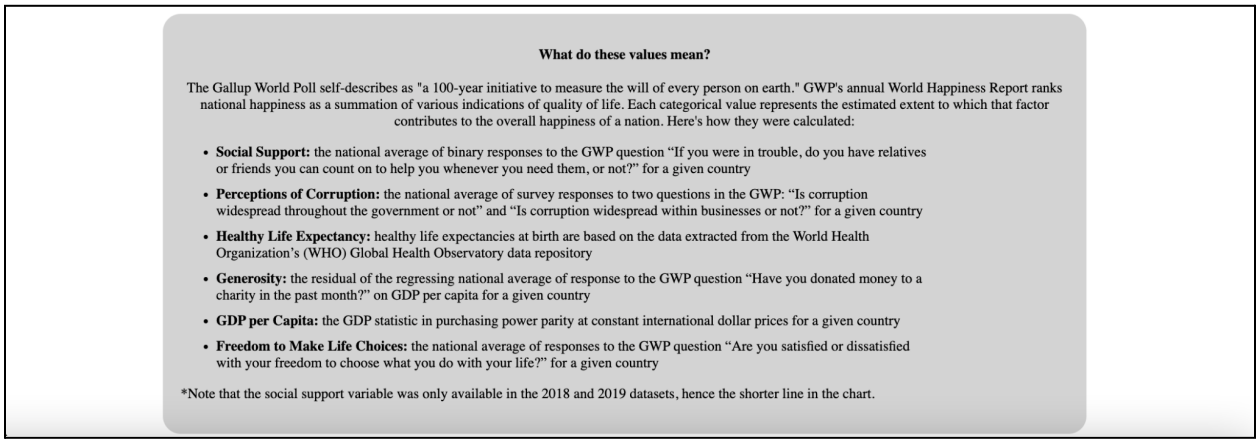
[*please view our visualizations using full-screen*]



Individual Country View (upon click of specific country):



Description below graphs:



Data Description

Objectives: The goal of our final project is to visualize happiness worldwide based on relevant factors (GDP, life expectancy, etc). We aim to demonstrate how each of the factors contributes to overall happiness and how countries compare globally in terms of their individual and continent-wide happiness rankings. This information could be useful to people looking to travel or relocate internationally because it gives insight into how citizens of a given country perceive their country.

Data Source #1: World Happiness Report

Data Description:

The world happiness dataset we selected was scraped from The Gallup World Poll and then found on Kaggle (World Happiness Report). The World Happiness Report produced by Gallup World Poll is a publication containing rankings of national happiness, aggregated from individuals' responses regarding various indications of quality of life. We used a total of 5 happiness datasets, each corresponding to a single year from 2015-2019. Each dataset consisted of 9 columns — 7 numeric and 2 categorical.

Description of the Variables:

The majority of the variable descriptions can be found directly from the World Happiness Report Statistical Appendix.

Continuous Numerical Variables:

a. Score - Happiness score based on the national average response to the question of life evaluations for a given country.
b. GDP Per Capita - Estimated extent to which GDP Per Capita contributes to Score. GDP Per Capita is the GDP statistic in purchasing power parity at constant international dollar prices for a given country.
c. Social Support - Estimated extent to which Social Support contributes to Score. Social Support represents the national average of binary responses to the GWP

question “If you were in trouble, do you have relatives or friends you can count on to help you whenever you need them, or not?” for a given country. This variable was only introduced in the 2018 dataset and was not present in any preceding dataset.

d. Healthy Life Expectancy - Estimated extent to which Healthy Life Expectancy contributes to Score. Healthy life expectancies at birth are based on the data extracted from the World Health Organization’s (WHO) Global Health Observatory data repository.
e. Freedom to Make Life Choices - Estimated extent to which Freedom to Make Life Choices contributes to Score. Freedom to Make Life Choices represents the national average of responses to the GWP question “Are you satisfied or dissatisfied with your freedom to choose what you do with your life?” for a given country.
f. Generosity - Estimated extent to which Generosity contributes to Score. Generosity represents the residual of the regressing national average of response to the GWP question “Have you donated money to a charity in the past month?” on GDP per capita for a given country.
g. Perception of Corruption - Estimated extent to which Perception of Corruption contributes to Score. Perception of Corruption is the national average of survey responses to two questions in the GWP: “Is corruption widespread throughout the government or not” and “Is corruption widespread within businesses or not?” for a given country.

Categorical Variables:

a. Overall Rank - The rank of a given country based on its Score value relative to all of the countries in the dataset.
b. Country or Region - The country/territory for which values are being measured.

Data Source #2: World Atlas

Data Description: The world atlas topojson dataset we selected was sourced from the D3 world atlas GitHub repository (World Atlas). We selected the countries-50m.json file, which contains geometry collections of countries, territories, and land. This dataset is scaled at the 1:50m medium scale.

Data Pre-Processing and Criteria for Variable Selection:

The primary cleaning that had to be done was matching country names between the two datasets. The World Atlas Dataset contained more countries/territories than the Happiness Report Dataset, leading us to have missing happiness data on many countries. This missing data, while possible to work with, required us to pay careful attention to type errors in our code to ensure that our visualization would run without issue. Additionally, the only field in common between the two datasets was the country names, a string field that contained many inconsistencies. For example, the United States may be labeled as “United States” in one dataset, while it would be labeled as

"United States of America" in the other. We identified a total of 11 countries to manually standardize names to ensure a seamless join between the datasets later on.

Overview of Interactive Elements and Design Rationale

Description of the Mapping from Data to Visual Elements:

1. **Visualization 1 – "Happiness Scores Around the World in 2019"**

For our first visualization, we created a map with countries outlined and filled them with color based on their 2019 happiness score. We were able to fill 156 countries with the dataset we had and countries without data were left a gray color. Users are able to zoom in and out of the map to get a better look at regions of the world map. Users are also able to click on specific countries and receive more information about that country's overall happiness rank or the makeup of their happiness score. To convey the clickability of countries, we created mouseover and mouse out events that widen the outline of the country the user is moused over. Widening the country's outline when it is moused/hovered over communicates to the user that the visualization affords clicking and encourages the user to interact with the visualization without needing to explicitly tell them to do so. When the user clicks on a country, the opacity of all unselected countries decreases. This helps the user identify which country they clicked on by highlighting that country with the highest opacity. If the user clicks the highlighted country again, all of the opacities on the map return to normal.

The visual channel we employed in the first visualization is the color of the countries. The outlined countries in the map are the marks. Each country's happiness score was mapped to a color based on a sequential color scale and then used to fill that country.

Many considerations were made in our decision to use the blue-green-yellow sequential color scale over other scales. First, we chose to use a sequential color scale for this visualization because the happiness score does not contain a central value of importance, making a sequential scale more intuitive to understand. Other scales such as a divergent color scale would be ill-suited to represent this data because it is continuous with values ranging from 0 to 7.769. After making the decision to use a sequential scale, we decided to use a blue-green-yellow color range. We decided to use blue as the darkest color and associated it with the lowest scores, and yellow as the lightest color associated with the highest scores. This decision to associate the darkest blue with the lowest happiness scores goes against the convention of using darker colors for high values. We went against this convention because given the context of our color scale, it makes sense to map dark blue to a lower happiness score because blue is associated with sadness/depression, whereas yellow is associated with happiness. To further clarify this to our users, we also included a color scale legend so that the user can see which colors mean a higher versus lower score.

When all of the visualizations are first loaded, all countries have the default opacity of 1 and the global summary data is displayed in the 2nd and 3rd visualizations. When a country is clicked, it maintains full opacity while the rest of the countries shift to a decreased opacity. This is

a visual signal to the user that a country has been clicked. Additionally, more information about the selected country is displayed in the 2nd and 3rd visualizations to the right of the map (more detail on the contents of these visualizations will be described later in the report). If the user wishes to view the global summary data again, they may click on the same country once more and return all visualizations to their default global values.

There were multiple tradeoffs we had to consider while developing this visualization. One potential tradeoff of our color scale is that users may first interpret the darker blue color with a higher happiness score because oftentimes darker, more intense hues indicate higher values. Despite this, we made the decision to implement the color scale the way we did because we felt that mapping blue to "sadder" countries and yellow to "happier" countries was the most intuitive use of our color scale, given our context. We embraced this trade-off so that the colors correlate with the emotion they represent rather than having them represent the opposite. Our color scale legend also helps further mitigate this tradeoff by giving the user an explicit mapping of the colors for them to reference.

Another tradeoff present is our audience cannot obtain the specific happiness scores of each country from this visualization alone. Instead, users are only able to compare countries against each other based on the colors they see. For example, we can see that Canada is a light yellow while the United States is a light green so we can conclude that Canada has a higher happiness score than the United States. The map color scale legend is also intended to help our audience compare the colors of the countries with more guidance. To manage this tradeoff of being unable to display scores, we instead displayed the country-specific happiness scores and their breakdowns in Visualization 2: "Happiness Breakdown by Country". So while our first visualization alone is unable to communicate specific scores and their breakdown, we utilized supplemental visualizations to fill this gap.

2. Visualization 2 – "Happiness Breakdown by Country"

Our second visualization, "Happiness Breakdown by Country", serves as an extension of our first visualization. This visualization is a pie chart detailing the values of the 6 different factors that contribute to the happiness score - social support, perceptions of corruption, healthy life expectancy, generosity, GDP per capita, and freedom to make life choices. A legend mapping each of these factors to its corresponding color and value is located below the chart. At the very bottom of the legend, users can see the summation of all the factors' values into the total happiness score.

When the page is first loaded, the global average happiness score and its breakdown are displayed in the visualization. When a user clicks on a country from the first visualization, our second visualization responds to the selection by displaying the happiness score and breakdown of that country. In addition to all of the information displayed as described previously, the name of the country and its overall happiness rank are also displayed at the top of the pie chart. For countries where there is no happiness data available, no pie chart is displayed. Instead, the message "Happiness data not available for this country" is displayed in place of the chart. When a

user double clicks a country to return the map visualization back to its default values, our second visualization also responds by returning to the global average happiness score and its breakdown.

The mark we implemented was the circle shape of the pie chart to convey to our audience that the happiness score is made up of many categories. In other words, the 6 factors are each part of the whole happiness score. The visual channels we employed were 6 different colors to represent the 6 features of a country that make up a country's happiness score as well as the angular size of each wedge in the pie chart. Since these fields are categorical, we chose random colors that were unrelated to each other while trying to accommodate colorblind users as well. For our final set of colors, we chose a series of random colors found on the D3 Color Schemes website. Rather than using Category10, the random color scheme we were introduced to in class, we used lighter colors that were softer on the eyes to make our visualization more appealing.

One trade-off of this visualization is that the user will not be able to see the exact percentage of how much a category makes up the happiness score. Users may be able to infer the percentages by guessing the size of the category slice in the pie chart, but may not get the exact percentage since we did not include those numbers. Instead, we chose to display the true values of each category because each country has a different total happiness score. One reason we chose to do this is if we had included percentages, the user might be misled to think that each country has the same total score and only differs in how much each factor affects the score. While users will not know the percentage contributed by each factor, they will know the exact amount each factor contributes to its country's total happiness score and they can still use the pie chart's area to infer that percentage.

3. **Visualization 3 – "Happiness Breakdown by Country Over Time"**

The third visualization displays the 4 year trend of happiness data from 2015 to 2019 and its 6 criteria that contribute to the calculation of the happiness scores. The line chart consists of the *Year* on the X-axis and *Category Rating* on the Y-axis. Similar to the second visualization on "Happiness breakdown by country," the user can interact with the line chart with a click of a country on the map. The implemented visual mark includes linear lines that depict the changing data points over the span of 5 years. Instead of point visualizations, we decided to use lines as our geometric marks for clear representation of the changing trend and its connections. We also employed vertical and horizontal gridlines as marks in this visualization. They are present to help our audience keep track of which vertical lines represent which year in time and which horizontal lines represent what rating value. A trade-off of using linear lines instead of circle points to convey the ratings throughout the year is that the user may think the ratings for each category change linearly throughout the year but that may not be the case. As we only have one data point for each category for each year and for each country, we do not know if the category tends throughout each of the years is linear. Despite this, we still chose to connect our data points linearly in order to show the average trend throughout the years. The design decision for the visual channels is made based on the second visualization as we intended to use identical category colors for both visualizations to minimize any possible confusion regarding user interaction and interpretation. For the details of our design decisions on color schemes, refer to the Visualization 2 description as we used the identical color schemes for consistency. For countries that do not

have available data for display, such as Democratic Republic of the Congo, we instead display a text message “Data (2015-2019) not available for this country.”

Some notes that are worth being discussed are regarding the validity and reliability of the category rating data. According to the appendix, the category ratings data are collected, evaluated, and calculated individually, meaning that the category ratings may not be compared across different categories. Below are a few examples of the data categories description provided in the appendix.

• *Social support (or having someone to count on in times of trouble) is the national average of the binary responses (either 0 or 1) to the GWP question “If you were in trouble, do you have relatives or friends you can count on to help you whenever you need them, or not?”*

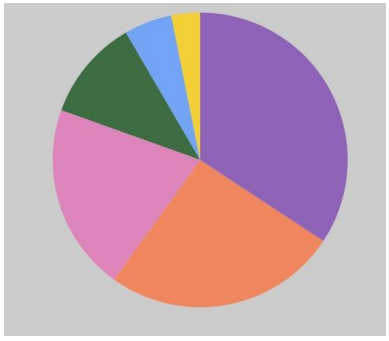
• *Freedom to make life choices is the national average of responses to the GWP question “Are you satisfied or dissatisfied with your freedom to choose what you do with your life?”*

• *Generosity is the residual of regressing national average of response to the GWP question “Have you donated money to a charity in the past month?” on GDP per capita.*

As we may notice, the category ratings are evaluated as a result of processing binary responses, responses within a given range, and etc. Hence, we may notice the users that comparing category ratings across different categories based on the value rating may not provide accurate information.

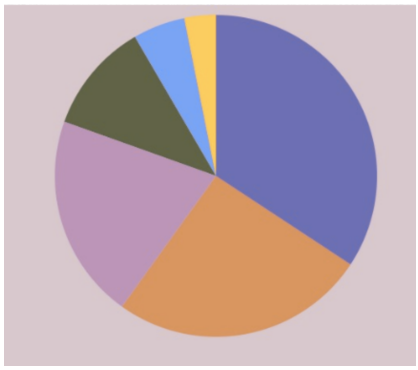
Usability of Our Color Scale Given Color Blindness:

Original:

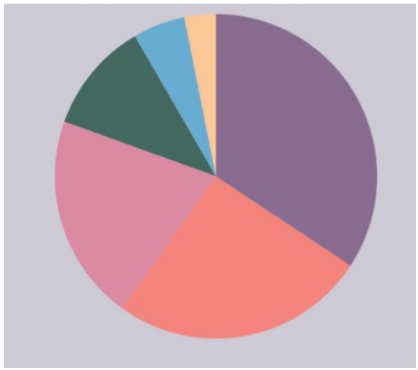


Using the Coblis Color Blindness Simulator,

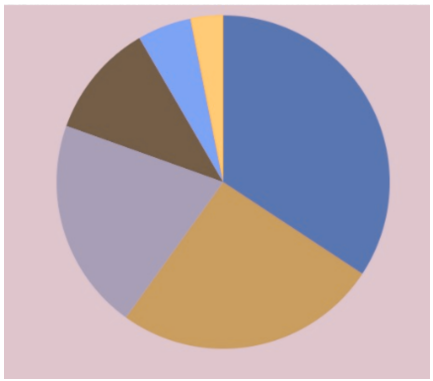
With Green-Weak/Deuteranomaly:



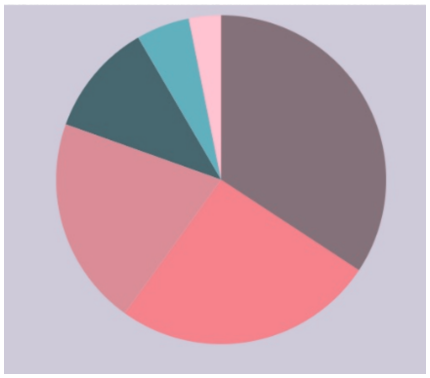
With Blue-Weak/Tritanomaly:



With Green-Blind/Deuteranopia:



With Blue-Blind/Tritanopia:



In devising our visual channels, we considered the usability of our design relative to color blindness. Using the color blindness simulator linked above, we found that our color scale would still afford perceivable differences in hue for individuals with color blindness.

Transformations:
We performed linear transformations for variable and color scales creations for each visualization.

Our Story

Our Inspiration:
Our inspiration came from thinking about potential places to travel to or even move to for a long period of time. One factor that comes to mind when thinking about what places to travel to is the overall happiness of the country. We want to know how a country's citizens feel towards their own home whether it be if they feel satisfied with the social support that is available to them or if they feel there is corruption within their government. Since we aim to help others decide what location to travel to, our first visualization helps our audience compare countries based on total happiness scores. Our second visualization aims to give more specific details on a country's happiness score such as what and how much each category makes up the happiness score. The third visualization shows how categories that make up the happiness scores have changed over the last couple of years.

What Insights Does "Happiness Scores Around the World" Convey?
The first visualization allows for our audience to make comparisons among countries all over the world. If a user is already interested in a specific region in the world, they may zoom to look more closely at a fewer set of countries. Users can only make comparisons of happiness scores by comparing colors of the country. The map legend that accompanies the map is meant to help users understand what the colors mean. The more yellow countries are most happy and closer to the max score of 7.77 while the darker blue countries are least happy and closer to the minimum score of 2.85.

What Insights Does "Happiness Breakdown by Country" Convey?
The second visualization conveys country specific information to the audience. At its default state, it displays global summary statistics or average scores of each category over all the countries we have information on. Once a country on the map is clicked on, this visualization changes to show the country's overall happiness rank in the world as well as the categories that contribute to its happiness score. The pie chart conveys the fraction of the happiness score that is made up by a certain category while the legend

beneath the pie chart conveys the exact amounts of how much a category contributes to the country's overall happiness score. The user is also able to tell the country's exact happiness score from this visualization. If a user is interested in the specific makeup of a happiness score, they may use this visualization to learn more.

What Insights Does "Happiness Breakdown by Country Over Time" Convey?
The third visualization provides information about the changing category ratings from 2015 to 2019 that contribute to the calculation of happiness score. The first two visualizations focus on providing the details of happiness data on the year 2019, whereas the last visualization provides contextual information as it shows the trend of happiness score over the past 4 years leading up to year 2019. There may be events, such as nationwide disasters, that may have had an impact on 2019 happiness data. The user will not only be able to interpret the county's happiness score data more accurately, but also will be able to tell how the happiness score has changed over time.

What Surprised Us?
While constructing our visualizations, specifically our line chart, we found that the way the surveys were conducted was not consistent. The Social Support factor was only introduced in 2018 while all other factors were there from the earliest dataset in 2015. A surprising fact we found after constructing our third visualization is that the social support of the #1 happiest nation, Finland, drastically decreased from the year 2018 to 2019. This category contributes most to the happiness score and decreases the most out of all categories for Finland. Despite this drastic decrease in social support, Finland continues to be the happiest country in the world in both these years. We also found that many country names would vary over the years due to human input in the datasets and the political nature of the nations. For example, in February 2019, Macedonia's name changed to the name North Macedonia but this was not reflected in our dataset and had to be manually matched on our end. Another example is that South Sudan seceded from Sudan in 2011 but this change in nations was not reflected in our datasets until halfway through our timeline of 2015 to 2019.

Team Contributions:

Lea: Lea created the first visualization utilizing the topojson file (which she found) of countries around the world (~6 hours). She implemented the mouseover functionality for that visualization as well (~2 hours). She also contributed to the data description and writing in general of the final report. (~3 hours).

Katherine: Katherine created the pie chart visualization and the line chart visualization, and she implemented the on-click functionality (~8 hours). She also helped proof-read the final report (20 mins). Katherine found the happiness report datasets that we used (20 mins).

Fatima: Fatima worked on finding a suitable color scale for the map visualization and reversed it to match colors with appropriate feelings. (~2 hours) She also created the map legend and significantly contributed to the final report. (~ 4 hours)

Taerim: Taerim created the GitHub repository and created a data dictionary for convenient data manipulation (30 min). She also explored potential visualizations and their implementation (6 hours) and contributed significantly to writing the final report (2 hours).