Final Project

Guled L. Gedi

Saint Mary’s University of Minnesota

BIA640: Data Visualization & Storytelling

Dr. Timothy Kyle

March 31st, 2024

Data Visualization Proposal:

The data that I am planning on using is brain function and mental wellbeing of individuals during the COVID-19 pandemic. The dataset(s) will ideally contain some measure of brain activity (MRI readings, for example), along with age, sex, location of individuals, an evaluation of happiness, and a series of qualifying factors that aim to give a more wholistic understanding of everyone's answers (such as asking about family size, how safe they feel, etc.). It would be assumed that the questions would be asked in a survey manner, along with some recognized metric to measure the non-quantitative values of the survey (such as happiness, stress, feelings before and after COVID, life satisfaction, etc.). It would also be interesting to be able to have a column that measures if the individual notices any difference in their work quality/ethic before and after the pandemic. It seems unlikely, however, as I am unaware of any standard measure of this due to it being a recent phenomenon, and if there is, it will probably be a combined variable of different explanatory variables. If the dataset found deals with country values instead of individuals, the data collected will then be the average response of those surveyed in each country.

In most datasets of this nature, they tend to range between 100 to 200 members that are surveyed, which answers the saturation point metric. In order for a dataset to be statistically significance, it should have no less than 30 data points, which is rare to find in such a topic. Many surveys and evaluations of this nature usually range in between 100 and 300 responses at a smaller scale, and at an international scale, thousands of responses are recorded and the average score of the region/city/country is recorded as the response. The data will need to be found on a public dataset repository, such as Kaggle, and from what I've skimmed through, there is ample data available on this topic. The question remains whether I will need to use multiple datasets to give a wholistic answer to what I aim to find and answer or if I will need to adjust my sphere of focus in visualization. Since I will be primarily looking to Kaggle for data, it will be safe to say that the dataset I will select will either be a .csv or .xsl file (most likely it'll be .csv, as text files are more prevalent in data analysis than Excel files from my experience). It would be expected that of the files I peruse, all of them would at least share one KPI: happiness. However, they could also have other KPIs, such as life satisfaction, home stability, life expectancy, whether they feel safe or supported, reported feelings of stress, etc. I would have to deal with null values and errors found in the cells as well, as in a lot of larger datasets, empty calls show up more and more frequently the larger the dataset is.

Regardless of the data I use and how I adjust my focus accordingly, I hope to be able to analyze and visualize data found relating to mental wellness and the COVID-19 pandemic. With a sufficient sample size and background research on the affects of COVID during and after the pandemic, it shows promising results to be able to present and analyze data relating to mental wellness during COVID, as well as investigating the underlying influences of the additional explanatory variables and seeing how they come into play in altering responses of wellbeing.

The focus of my research will be aimed on the relationship between metal wellness and the effects of COVID-19 on the world's mental state. Much research and data was collected during 2020 and 2021 addressing aspects of this, which allows for some precedent and background research on this topic. My aim will be to focus on the neurological and psychological aspects of COVID on people, and then to discover more about how all of the data relates to each other. How each variable plays a part in affecting self-perceived happiness and/or satisfaction, what patterns can be observed in the data, and how significantly each variable impacts the response of the surveyed individuals. It is hoped that through this data visualization project, that certain patterns of behavior can be found and/or traced that can be implemented in even larger studies. This can hopefully lead to discoveries on how to help people suffering after COVID to be identified and given the necessary aid to improve their mental states to lead to more productive and fulfilling lives.

Key Performance Indicators:

The data that I am using is the World Happiness Report of 2020, with special focus on how the COVID-19 impacted the results of this year's results. The data was collected and downloaded from Kaggle as a zip file containing 3 CSV files. While I had initially planned on using the 'MortalityData' file, after closer inspection, it was not fit for use in visualization. It had too many null cells and didn't contain the metric for measuring happiness that I had planned to use as my response variable. I decided to use a different file, 'DataFor Figure2.1', which contained the happiness metric, had no missing or null values, and contained several other explanatory variables that could be used in analyzing and visualizing the data. After sifting through the data and reading the supplementary material provided in the zip file, I have now have roughly 3 key performance indicators that focus on different aspects of the data and each tell a story about how they are all intertwined. I've decided to name them Regional Happiness, Social Comfort per Country, and Life Fulfilment.

Regional Happiness is the simplest of the 3. It is a display of the average happiness score per world region (Central and Eastern Europe, Commonwealth of Independent States, East Asia, Latin America and Caribbean, Middle East and North Africa, North America and Australia/New Zealand, South Asia, Southeast Asia, Sub-Saharan Africa, and Western Europe). While initially I had planned to use all of the individual countries for this, I realized it would be too cumbersome and crowded, especially since I was planning on using a bar graph to visualize the data. This isn’t meant to be a ground-breaking use of the visualization software, but it is meant to be an introduction for the readers of the data to first understand what the data is measuring and where the data was collected from.

Social Comfort per Country is a more detailed indicator, in that it is the first visualization that attempts to actively measure/observe a relationship between variables. In this case, it is the relationship between the happiness index and Freedom to make life choices & Social support. These two explanatory variables seemed to have much overlap between them, as it can generally be assumed that if high social support is present in a society, the freedom to make life choices is greater than in a society where it is not. It is hoped that this data can be represented in a scatterplot with the individual countries as the individual data points. This indicator will shed light on how sociological and psychological pressures affect the average happiness of the countries.

The third key performance indicator is Life Fulfilment. Like the second KPI, this aims to measure the relationship between multiple explanatory variables and the happiness index, with the variables in question being healthy life expectancy and generosity. The relationship between healthy life expectancy and happiness is perhaps a given, but the relationship between generosity and happiness is something that isn’t as easily measured. Due to the simple fact that the generosity of a person greatly varies based on financial and social situations, as well as how safe the country may be and how prevalent corruption may be found in the government (which are also 2 other explanatory variables that are found in the ‘MortalityData’ file). As with the first KPI, to visualize all of the countries as individual data points would be too hard to keep track off. Using the world’s regions as a grouping mechanism, this KPI aims to see how people’s self-reported physical and social health was during the COVID-19 pandemic.

Continuing from this, I would still like to incorporate elements from the ‘MortalityData’ file, especially with regards to using more of the obscure column data such as whether the country was an island or whether the head of the country was a female. Combining the tables will prove to incorporate more of the data of the survey and give a more comprehensive image of how happiness worldwide was impacted by the COVID-19 pandemic. Most of the data taken from this table will be discrete, so not much transformation will be needed, but transformations will need to taken into consideration especially further along the process should more precise questions arise during the data analysis and visualization process. More variables could also be added to existing KPIs, or new KPIs could be added to neatly organize the many aspects of the data collected. There is also the question of formatting the data values, as the data seems to be collected in Europe where commas are used in place of decimal points, thus giving a wrong impression that the happiness index data found is in the thousands when the scale is actually between 1 and 10.

Dashboard Elements:

This dashboard aims to present the data collected for the World Happiness Report of 2020 through various key performance indicators, aimed at addressing and visualizing the relationships between the different variables present from the survey. Using a variety of graphs and figures, the data will be presented in a way to illustrate the impact of COVID-19 on happiness worldwide, along with safety, social support, freedom, finance and health. The data will be presented in a straightforward way, providing the necessary information for the readers while not being too cluttered or empty. The goal of this dashboard is to help inspire more research into the individual topics discussed and presented in the dashboard and to further analyze post-COVID physical and mental health. It is hoped that many professionals from differing fields will find the data collected here to be impactful and a starting point for new research in this topic.

The ideal audience that will have the most interest in this dashboard will be members of the psychology and sociology communities as well as members of the scientific communities, most likely those involved with social neuroscience. However, researchers in general would find this data beneficial especially since the data collected covers a vast array of topics such as financial and political elements. The goal of this dashboard is to inform relevant parties how the effects of COVID-19 rippled in all aspects of life, especially when comparing the happiness scores of years prior to 2020 to the 2020 report. This dashboard will most likely be shared in a psychological journal as the data collected and the metrics used are frequently used in similar research or perhaps be posted publicly online with the initial data source posted as well to ensure that there is data transparency.

The dashboard will start off with a basic graph displaying the average happiness per world region in 2020 to introduce the readers to the topic at hand and the extent to which the data was collected. Each column will represent each of the regions in the dataset along with their respective names. I will also consider using a geographical graph as well just to assist in the visualization of the spread of the data. It will then be followed by the Life Fulfilment graph and Social Comfort Per Country in order to begin showing the relationship between the various variables and the happiness index. There will be a scatterplot comparing the happiness index from 2017 – 2020, and finally there will be a scatterplot comparing the relationship between safety, social support, happiness, and freedom to make life decisions. Each of these graphs will focus on one aspect of the data and aim to demonstrate how each of the variables are related and how they impact happiness and each other and if there are any relationships between them.

In order to make the dashboard neat and accessible, several formatting and cleaning tactics will be used. Firstly, all of the text present in the dashboard will be of the same font, with titles being 5 font sizes larger than the standard text size or axis text. The color scheme will contain a variety of colors, but in order to accommodate for colorblind individuals, green and red will not be placed next to each other and instead be replaced with orange and blue. All of the graphs will be formatted to be the same size and lined up to make sure that they are all parallel to each other. Each graph and image will be labeled and organized according to size and relevance. The initial happiness vs world region graph will be one of the larger graphs, followed by the other 4 in descending order. If the geographical map is produceable, it will have its own area of the dashboard a bit removed from the rest as it is mainly present for conceptualizing the data presented in the happiness vs world region graph.

The goal of this dashboard is to educate and inform about the mental state of people during the pandemic, and to prompt further studies branching from all sciences. The data will be filtered and available to be adjusted to suit the reader’s needs in accessing a specific part of the data. Whether psychological surveys, economic reports, neurological studies, or social experiments, it is hoped that through observing this dashboard further focused studies can be done. Although many of the effects of the pandemic have all but disappeared, there is now a permeant shift in society from pre-covid and the world has changed to accommodate for that. Another goal of this dashboard is to encourage further research into the relationships between the variables selected and their real-life effects and forecasting capabilities.

Trend Models:

In discussing the main trend models and forecasting patterns of my dashboard, only 2 of my KPI’s deal with trends and forecasting: Life Fulfilment and Social Comfort per Country.

Both scatterplot graphs contain lines of best fit, as well as the model equations, intercepts, R-squared coefficients, and p-values to determine how statistically significant each model was.

For Life Fulfilment, the graph displays average happiness of country by 2 explanatory variables: freedom to make life choices and social support. Both of these values are taken from self-survey reports, however, the main understanding of the scale used to measure it is that the higher of a score reported, it would mean that the individual is more comfortable to make important life decisions on their own and that they have more social support in doing things that they so desire, whether they be related to the first variable or not. The equation for the freedom vs happiness graph is Happiness score = 5.75902*Freedom to make life choices + 974.012, with a P-value of less than 0.001. This indicates a strong positive relationship between the 2 variables and suggests that freedom to make life choices significantly impacts happiness scores. The equation model for the social support vs happiness is Happiness = 7.075*Social support + -231.479 and a P-value less than 0.001, which also indicates a strong positive relationship between the two variables and also suggests that social support has a significant impact on happiness.

For Social Comfort per Country region, it also visualizes the impact of 2 explanatory variables on the average happiness score: avg. healthy life expectancy & avg. generosity. Trend lines were produced, along with the model equations, intercepts, R-squared coefficients, and p-values to determine how statistically significant each model was. The first graph measures the relationship between happiness and life expectancy. The equation for the first model is Happiness = -2.74307*Avg. Healthy life expectancy + 265155 with a P-value of 0.417227. This indicates that there is a weak negative correlation between social comfort and happiness, with no significant impact on happiness. The second graph measures the relationship between happiness and generosity per country region. The second model equation is Happiness = -220.681*Avg. Generosity + 82540 with a P-Value of 0.276781, which indicates a weak negative correlation between the two variables.