Methodology

1. Data Sets

The datasets used in this study came from two separate research corpuses. The first was the United Nations (UN) yearly World Happiness Report (WHR). The dataset consists of metrics based on survey questions distributed by the United Nations to respondents from each of the countries listed. The overarching metric used to define happiness, and the one which became the dependent variable and focal point of the present study, is the "Life Ladder". The metric was a question which asked respondents to rank their happiness as if it were a ladder of 10 rungs where rung 10 was the highest (happiest) the person felt about their given situation. Additional metrics were maintained to strengthen the eventual analysis which developed the results of this study. Some of these included, GDP, rate of corruption, and social support metrics as measured also by the UN.

The other four datasets researched for this study came from an institution called Our World in Data. The datasets consisted of an index on stringency of policy response to the COVID-19 pandemic, an index of facial covering regulations from countries across the globe, an index of public event cancellation policies, and an index of public gathering restrictions, the latter two being part of the same evaluations category listed on the Our World in Data site. As the stringency index was a metric that combined many others, the dataset turned out to be quite large and contain many columns with data pertaining to hospital bed vacancy numbers, population, and human development index, to name a few. These would have to be cleaned in our data transformation phase. The other three datasets from Our World in Data consisted just of the dates (each day of the year), the name of the country, and the metric in question.

2. Cleaning and Joining Data

To clean the datasets, it was determined it would first be necessary to find out which columns were vital among the numerous for the datasets. Specifically, the dataset for stringency index contained many of the metrics that went into creating the index number causing the file to be quite large. The csv file downloads from each dataset were imported to google colab.

To read the files and turn them into iterable datasets, pandas library was imported to the colab file. Using class lecture files and powerpoints from Foundations of Cultural and Social Data Analysis proved valuable as resources for determining which codes would help better understand the datasets.

In the cleaning stage, rows (countries) containing incomplete data for the necessary column being measured were removed. The datasets for stringency index, facial covering requirements, public event cancellations, and public gathering restrictions were collapsed down to their most necessary components. Additionally, the date formats needed to be changed in the datasets to datetime so that the month and later, year, could be extracted. For each dataset from Our World in Data, the files contained day-by-day analysis. This was collapsed into monthly averages for each of the datasets. Then, for the stringency index, the average per year was extracted from the data. Meanwhile, in the face covering, public event cancellations, and public gathering datasets, the mode of the data per year was determined to be the best measure in understanding the data. The UN data was broken down by year to include the Life Ladder metric as the main variable for study. The data from the UN's WHR was being cleaned and the determinations of which columns to keep were in flux. Eventually, the determination was made that several additional columns would help moderate the regression analysis to come.

Once the datasets were cleaned uniformly research was conducted, once again using class lecture resources from the previous course Foundations of Cultural and Social Data Analysis, and applied to joining the datasets. The inclusion of both country and year across all datasets was crucial in allowing the joining of the datasets on those two columns, further strengthening the final combined dataset resulting from the coding. Using an inner merge, the default merge method in the pandas library, helped clean the dataset further by leaving us with just the countries and years that included all data necessary for the final analysis and excluding those that would have left us with incomplete data. This was the case because we had two different sources of data and although the data from Our World in Data on COVID-19 restrictions contained the same countries and years across all four datasets, the UN's report contained different countries.

3. Hypothesis Testing

In our hypothesis testing, we conducted both correlation and regression analysis to examine the relationship between policy measures and national happiness.

Correlation analysis is a statistical technique used to evaluate the strength and direction of the relationship between two variables. The strength of the correlation is typically measured on a scale from -1 to 1, where a negative value indicates a negative relationship, and vice versa. The purpose of conducting analysis in this study is to provide an initial overview of the relationship between the two variables without

In contrast, regression analysis enables us to directly examine how changes in the independent variables affect the dependent variable, which is essential for our hypothesis testing. Additionally, regression analysis allows for the inclusion of control variables to isolate the unique impact of the independent variable (i.e. policy measure) on dependent

variables (i.e. national happiness). In our study, we extracted four variables, namely GDP, Social Support, freedom to make life choices, and perceptions of corruption, from the dataset of happiness. These variables were treated as control variables in our analysis based on their statistically significant effects on national happiness, with p-values below 0.01.