**Final Report - What Makes People Happy: Happiness Level Prediction**

**Team members:** Yutong Li, Daniel Masters, Ramin Eghtesadi

**Background**

A study conducted on mental health before the pandemic was revisited using the same population in March to July 2020 and it was observed that there was a 90% increase in depression rates of the participants. Though the presence of the pandemic has decreased in most people’s lives, it still exists. It begs the question: are people happy now? What is the happiness level of people on a global scale? These questions were our motivation for this project and why we chose to use the datasets provided by The World Happiness Report. The World Happiness Report has been conducting surveys on global happiness in over 150 countries since 2012. They provide a general definition of happiness by using economy, health+life expectancy, family support, trust in government, freedom, generosity and dystopia residual as main indicators. In this project, we will be looking into the happiness score of each dataset from 2015 to 2022. Some of the questions that will guide our data science project are: What are the 20 happiest countries and the 20 unhappiest countries for each year? What regions do they belong to? Which countries have had the most dramatic changes in the ranking over time and what possibly caused these changes? What is the most significant contributor to happiness among these six explanatory variables? We will also further evaluate the strength of association of the six predictor variables with the region variable. From here, we will find the most significant contributor to happiness among these six explanatory variables. We will also consider many significant global events like the COVID-19 pandemic and try to ascertain how they affect the happiness score. Finally, we will predict next year’s total happiness score for some countries based on the previous data.

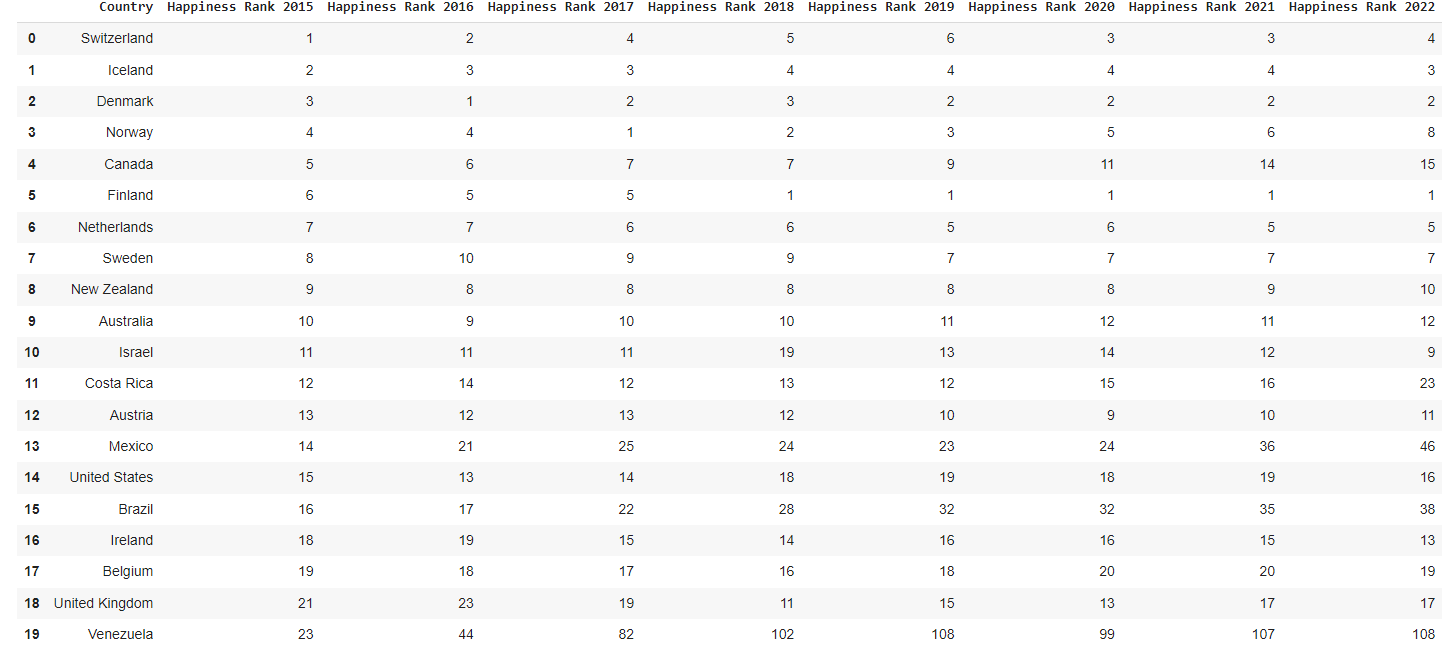
**Dataset Description**

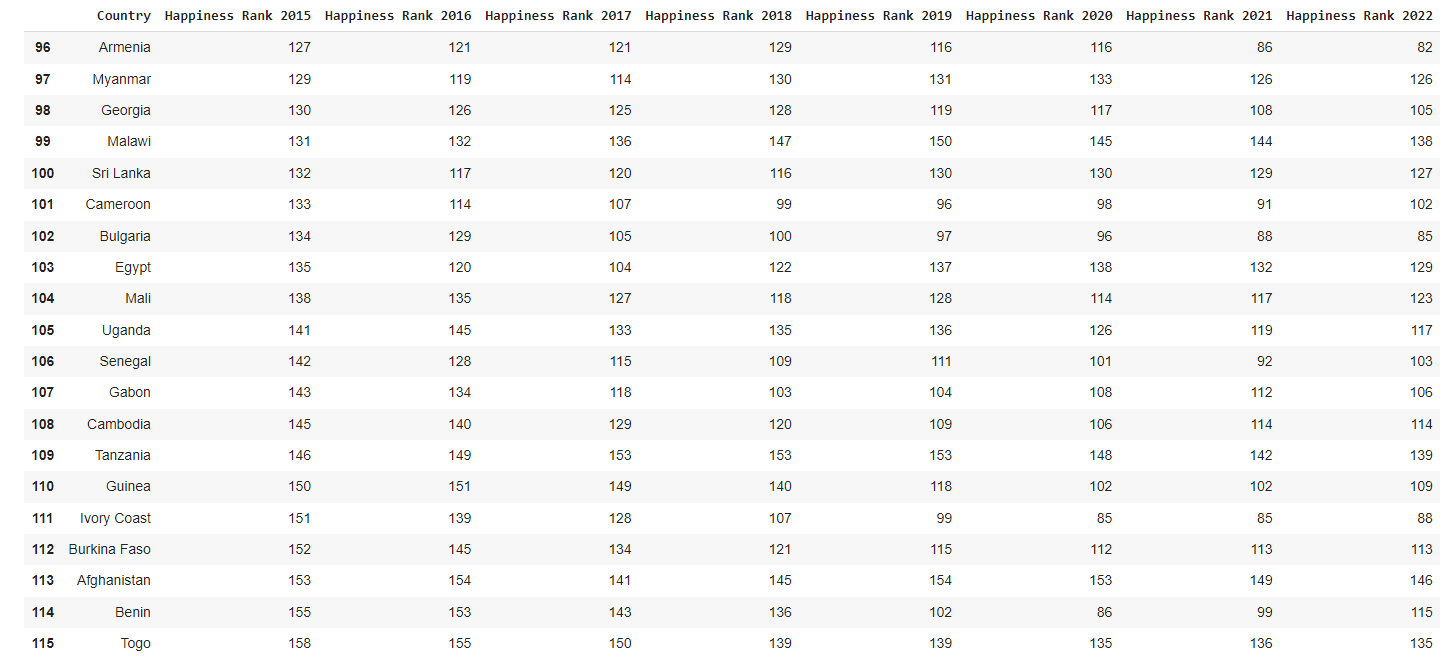
These datasets were obtained from Kaggle and all of them are open licensed. The datasets collect happiness scores from over 150 countries from 2015 to 2022 and are sourced from the Gallup World Poll. The size of each dataset is between 150 and 160 data points (countries). We will only focus on seven main happiness indicators: GDP per Capita, Health and Life Expectancy, Family Support, Trust in Government, Freedom, Generosity and Dystopia Residual. According to the world happiness report website, the participants are asked to rate their lives on a 0-10 scale and all the scores are being calculated using the Gallup weights. We will use these six main indicators to make the best predictions we can.

**Data Preprocessing and Initial Interpretation**

To preprocess the data, we read in the data and created dataframes for each of the 8 datasets using Python (pandas library). We first identified the variables that would be used throughout our project which were Country, Region, Happiness Score and the seven happiness indicator variables: GDP per Capita, Health (Life Expectancy), Family Support, Trust in Government, Freedom, Generosity and Dystopia Residual. One of the first things we noticed was that each of these variables had names that varied or were missing across all datasets. To address this issue we altered the variables by either changing its column name or appending new columns in the dataframe. These changes allowed us to have a consistent naming scheme. We also merged all datasets on the Country variable which dropped countries that were not ranked in every dataset. Lastly, we dropped all missing values from the merged list. This left us with 116 countries and we focused on the top and bottom 20. The figure below shows this.

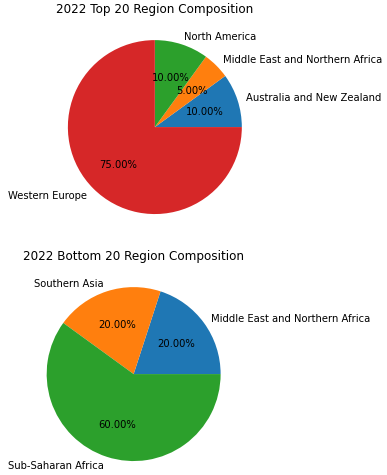
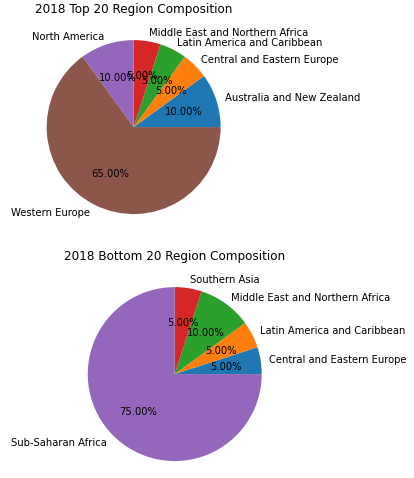
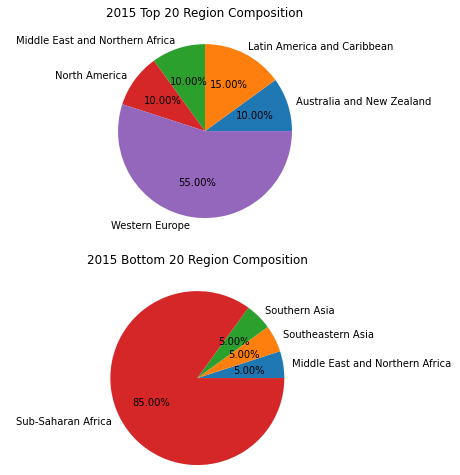
Figure 1)





We want to observe why these countries are at the top and bottom 20 of the happiness rankings. We decided to group each of these countries by region to see if we can find the connection between region and happiness score. To help us find this connection, we calculated the composition of these regions and displayed it in a pie chart. The pie charts below show the percentages that each region occupies in the top and bottom twenty countries for each year.

Figure 3) Figure 4) Figure 5)

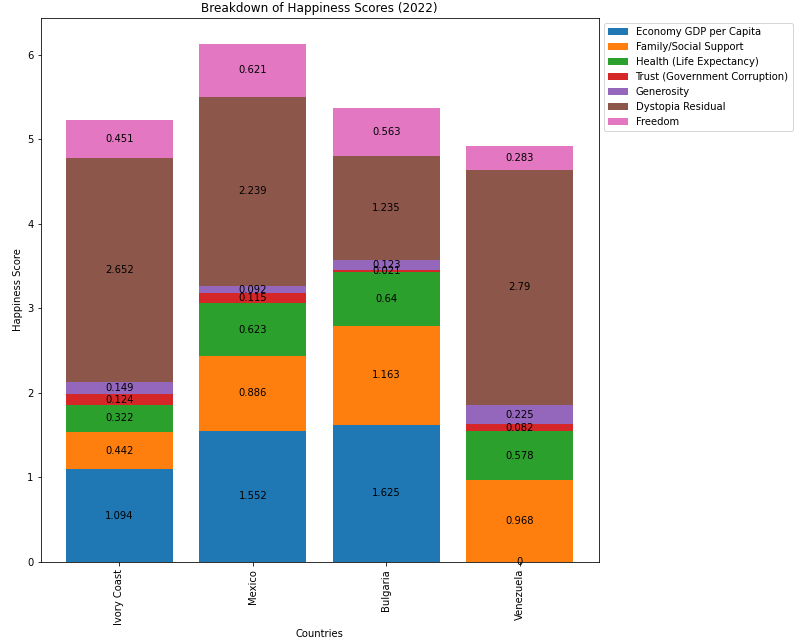
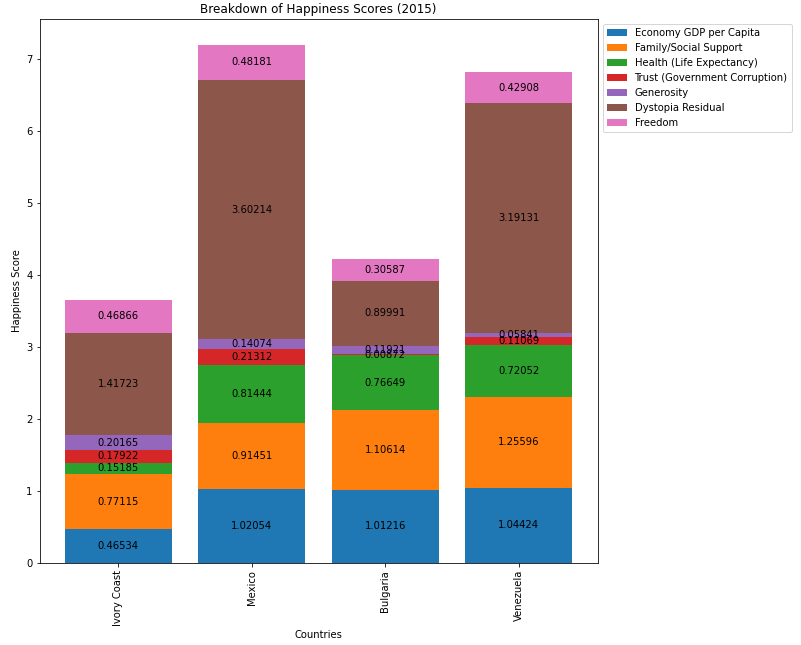


*(pie charts for 2016, 2017, 2019, 2020, 2021 can be found in project code)*

The first six pie charts show the top and bottom twenty region compositions for 2015, 2018 and 2022. The region that holds the greatest number of countries in the top twenty for 2015, 2018 and 2022 is Western Europe. In the bottom twenty for 2015, 2018 and 2022, the region that holds the greatest number of countries is Sub-Saharan Africa. The Western Europe region dominates the top 20 happiest countries by making up 55% to 75% of the composition every year from 2015-2022 (attached code shows missing pie charts to support this claim). On the opposite side, the Sub-Saharan Africa region dominates the bottom 20 happiest countries by making up 75% to 85% of the composition for every year (see attached code for missing pie charts). The trend in the top 20 happiest countries from 2015-2022 is that the country is likely to be from the Western Europe region and likewise, the trend in the bottom 20 is that the country is likely to be from the Sub-Saharan Africa region. We back these claims by providing further probabilistic calculations based on the percentages from the pie charts. On average, the probability that a top 20 country belongs to the Western Europe region is 64.375% (5.15/8 = .64375 = 64.375%). Similarly, the average probability that a bottom 20 country belongs to the Sub-Saharan Africa region is 74.375% (5.95/8 = .74375). These calculated percentages support our claims that a country in the top 20 will likely be from the Western Europe region. Such a high average probability of Western Europe being in the top 20 implies that a random country from Western Europe is more likely to have a high happiness score. Likewise, a random country from the Sub-Saharan Africa region is likely to have a low happiness score.

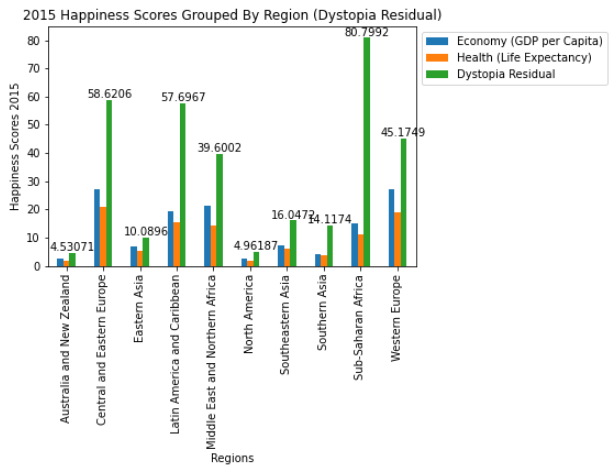
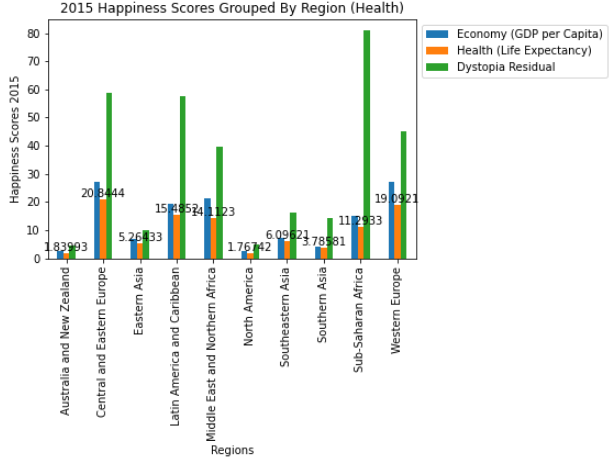
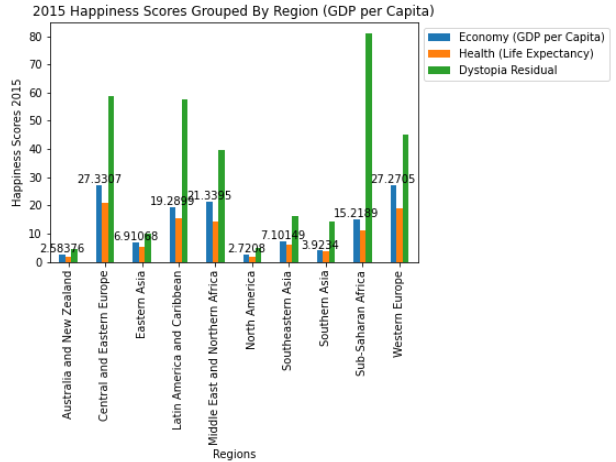
After observing the happiness ranks of each of the top twenty happiest and bottom twenty happiest countries (figure 1 and figure 2), we came across countries that we call “big risers” or “big fallers.” Big risers/fallers are calculated by the difference in initial happiness rank in 2015 and ending happiness rank in 2022. Big risers show a large increase in happiness ranking (i.e 150th in 2015 to 85th in 2022) and big fallers are those that experience a big decrease in happiness ranking (i.e 22nd in 2015 to 102nd in 2022). The countries we determine to be big risers were Ivory Coast and Bulgaria while the big fallers were Mexico and Venezuela. We were curious as to why these countries rose and fell so much in happiness ranking so we made graphs to show the composition of their happiness scores in 2015 and 2022.

Figure 6) Figure 7)



Happiness score is what determines happiness ranking and is composed of 6 focus/predictor variables. As mentioned earlier, these variables are: Economy (GDP per Capita), Family/Social Support, Health (Life Expectancy), Trust (Government Corruption), Generosity, Dystopia Residual. As one can see from the graphs, Ivory Coast and Bulgaria started with low happiness scores in 2015 and ended up with much higher ones in 2022. The opposite happened to Mexico and Venezuela. The predictor variables that contributed most to the increase in the big risers’ happiness scores were GDP Per Capita and Dystopia Residual. Dystopia Residual is the score that compares a country to a Dystopia (the least happiest country or country with 0 happiness score). The higher the Dystopia residual, the happier the country. The predictor variables that contributed most to the decrease of happiness scores for the big fallers were Health and Dystopia Residual. The rise in GDP Per Capita as well as Dystopia Residual for Ivory Coast and Bulgaria reflect their country’s growing economies. On the opposite side, the decrease in Health and Dystopia Residual for Mexico and Venezuela point to both countries still recovering from the negative effects of COVID-19. We observed that GDP per capita, Health and Dystopia Residual are the main predictor variable contributors to happiness score for Ivory Coast, Bulgaria, Mexico and Venezuela. Now, we want to measure the association of these three predictor variables across all regions through the graphs below.

Figure 8) Figure 9) Figure 10)

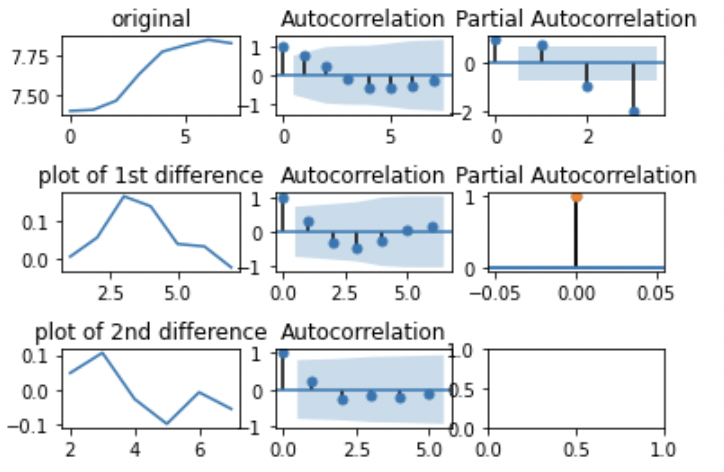
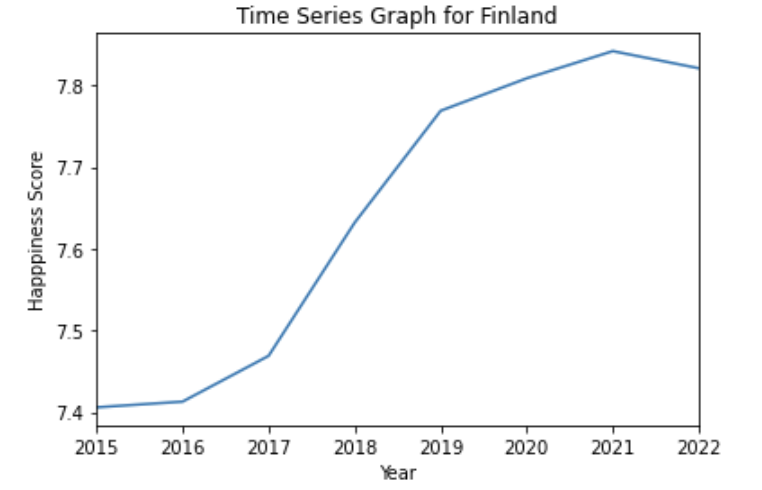
****

*(graphs for 2016, 2017, 2018, 2019, 2020, 2021, 2022 can be found in project code)*

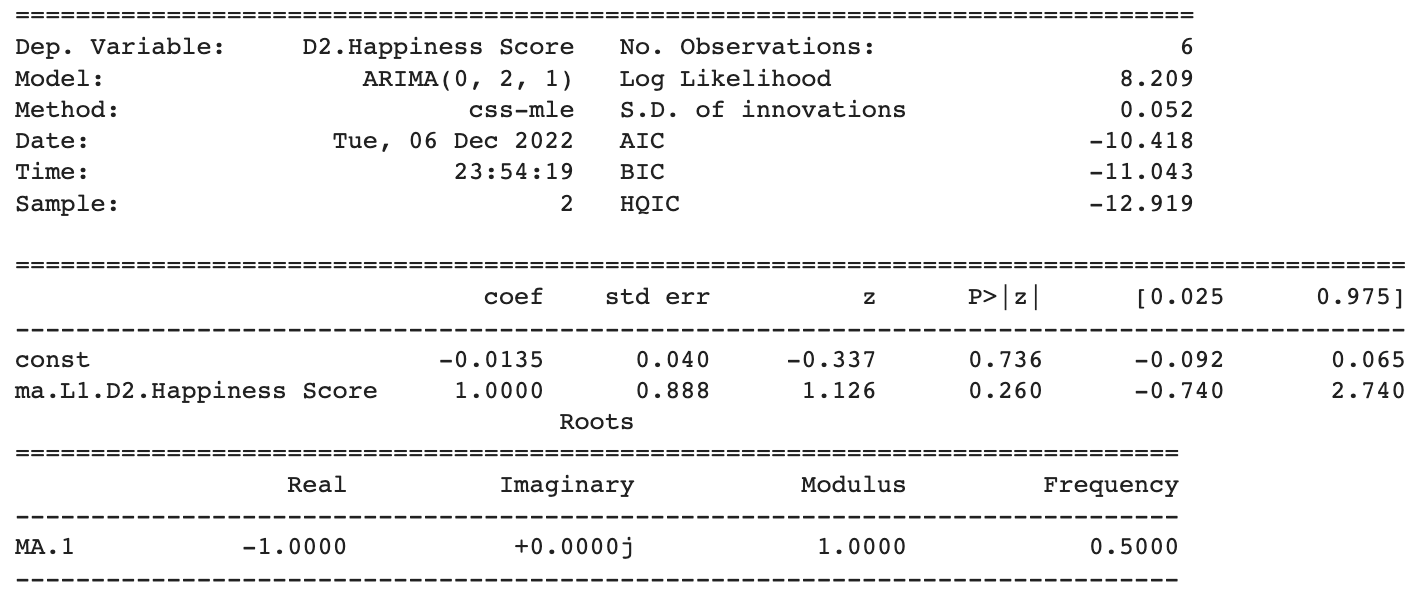
Each of these graphs show all three of our predictor variables of interest. Each graph focuses on a different predictor variable which is denoted by the numerical float labeled over each column. We see that each float value across each region is drastically different. For example, the Australia and New Zealand region’s GPA per capita is 2.58 while the Central and Eastern Europe region’s GDP per capita is 27.33. These large differences can be observed across multiple regions and points to strong association between each predictor variable and the region, which again confirms that happiness scores across regions are dissimilar.

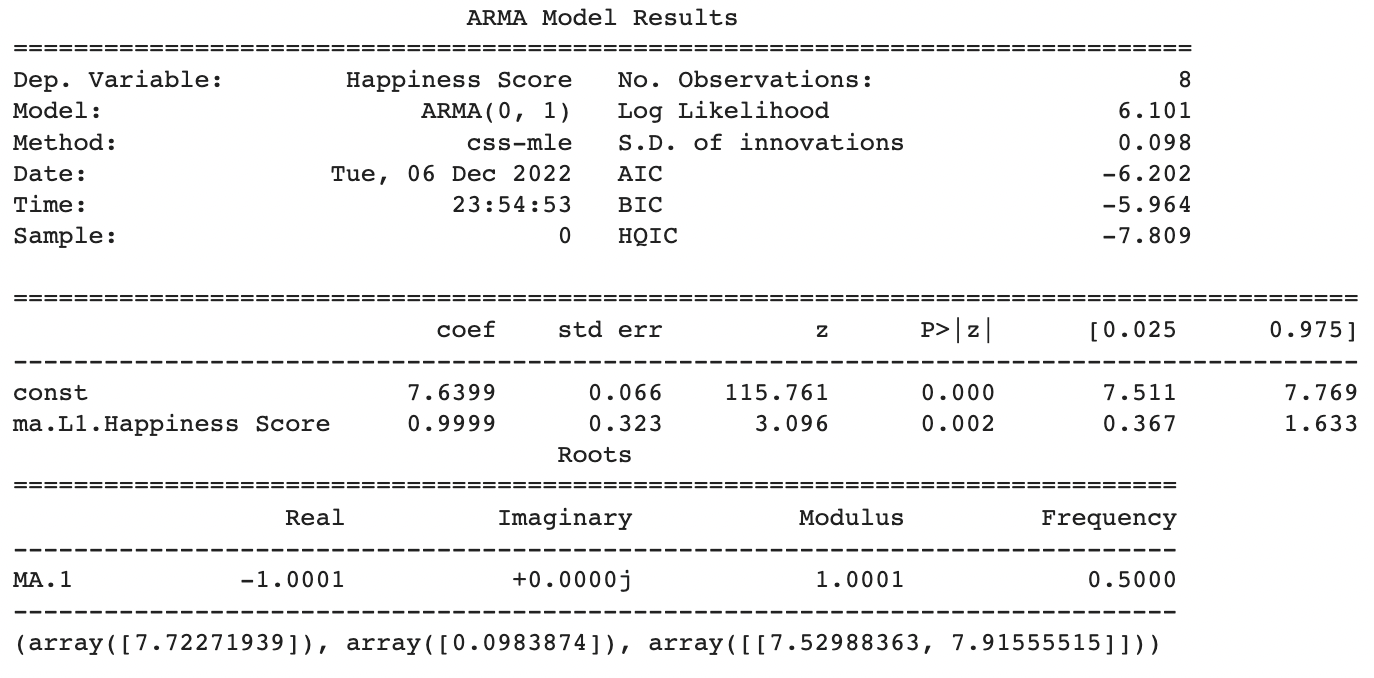
**Time Series Forecasting**

Now in order to predict the future happiness score for each country for the next year, since we have the happiness scores collected in a time ordered fashion from the past 8 years, it seems that time series forecasting is an adequate analysis to perform in this case despite the concerns of inaccuracy due to lack of data points. For demonstration purposes, we choose Finland, the top country in 2022, as an example. First, we need to determine if seasonality exists in our data. Happiness score is not seasonal data by convention, and it does not change regularly within a certain period of time, which can also be confirmed by the time series graph below (e.g., TS plot 1). Thus, we will proceed using the ARIMA model instead of the SARIMA model. To build an ARIMA model, we need to identify the parameters p, d, and q where p stands for the order of the AR () term, d represents the degree of differencing required for stationarity, and q is the order of the MA () term. Stationarity is another issue for modeling based on limited data points we have; it requires the mean and the variance to be constant over time. There appears to be an upward trend in the time series graph of Finland, so the data is clearly not stationary by simply looking at the graph. We can try to make it stationary by differencing which is the most common way to adjust non-stationarity. As we can observe from TS plot 2), time series look more stationary by differencing twice, meanwhile the sample ACF decays to zero slightly more quickly as lag increases. Hence, we can choose d=2 for our model. The ACF plot of the second differenced data cuts off at lag 1, but the PACF plot does not show due to insufficient data. By observing the pattern shown in the PACF plot of the original data and the first differenced data (TS plot 2), the PACF of the second differencing is very likely to be tailing off. Then we can say we have a MA (1) process and propose an ARIMA (0, 2, 1) model for forecasting. After fitting the model, we obtain the following result summary (TS plot 3). However, the confidence intervals for both parameters include 0, which implies they are not statistically significant and there probably exists a better model. Since the positive effect that differencing leads to is not significant for this time series, we can try to stick to the original data without differencing as the PACF is clear to read as well. Similar to the previous model, the ACF still cuts off at lag 1 and the PACF is tailing off, then we propose an ARIMA (0, 0, 1) model. Both confidence intervals do not include 0, so this model is statistically significant (TS plot 4). We can report the value that this model generates: the predicted happiness score of Finland in 2023 is 7.72 which can be found at the bottom of the result summary (TS plot 4).



Top Left : TS plot 1), Top Right: TS plot 2), Bottom left: TS plot 3)



TS plot 4)

There is a side note worth mentioning, despite the second model having statistically significant results, the AIC and BIC values of the first one are smaller than the second one, which suggests the first model is a better fit. Even though the predicted values from both models are extremely close, we still prefer the second model because of the statistically significant parameters. Considering any economic or political change has a different impact to every country, it seems to be more reasonable to propose a separate model to each of them. Due to the time limit, we repeat the above model selection process to a selected number of countries. - Denmark: it has been ranked in the top 5 countries for every past year. The proposed models are very similar to Finland which are ARIMA (0,1,1) and ARIMA (0,0,1). AIC and BIC values are very close between two models, but ARIMA (0,0,1) still generates significant results that the other one does not. ARIMA (0,0,1) gives the predicted future happiness score of Denmark in 2023: 7.62. At this point, we can also tell that the very similar predicted values of Finland and Denmark make sense because they belong to the same clustering group (details in the next section). - Ivory Coast: its happiness score has the biggest increase. The parameters of model ARIMA (0,0,2) are all significant, and we report that the predicted future happiness score of Ivory Coast in 2023 is 4.98. - Mexico: has the biggest decreased happiness score. The predicted future value of Mexico in 2023 from ARIMA (0, 0, 1) is 6.35. - The ACF, PACF and all the model outputs for the above three countries are shown in the code file.

**Clustering - TSNE Implementation**

The algorithm uses k-means clustering to pick a cluster center significantly reducing the dimensionality of the data by splitting it into clusters. was selected as 2: the list of the top 20 and bottom 20 countries. This can be seen in our general comparison between 2015 and 2022 as the cluster for the top 20 countries is relatively the same in 2022 with the exception of the loss of Mexico and the Emirates which were present in 2015. The countries consisting of the 20 least ranked countries continue to change every year, typically replaced by a different set of countries primarily in central Africa. This trend is consistent in every year of data between 2015 and 2022.

**Conclusion**

All of our collected data points to the conclusion that different predictor variables in different regions had profound effects on different countries' annual happiness scores and rankings. GDP per capita, Health, and Dystopia Residual were found to be the main cause for the drastic change in many countries’ happiness scores. As mentioned, the predictor variables that contributed most to the increase in the big risers’ happiness scores were GDP Per Capita and Dystopia Residual while those that contributed to the most significant decrease in happiness score include Health and Dystopia Residual. This variation further substantiates the general reason many western countries in Europe and the America’s continuously make the list of the top 20 countries as they continuously see the highest ranks in health and GDP per capita with the slight annual variation being the result of minor changes in each countries’ dystopia residual. On the contrary, many countries on the Sub-African continent continuously ranked the lowest of the 20 countries given very poor health and economic outcomes. As the stacked bar graphs show, The rise in GDP Per Capita as well as dystopia residual for Ivory Coast and Bulgaria reflect their country’s growing economies as the increase in GDP per capita is directly related to an increase in each respective countries residual dystopia while the decrease in health and dystopia residual for Mexico and Venezuela resulted in a significant drop of these countries’ happiness scores and rankings. While our data was able to pinpoint the main predictor variables affecting happiness, many different confounding variables shape people's perception of happiness all around the world. Different variables also come into play in different regions. This can be seen through the significant drops in the rankings of Mexico and Venezuela; Mexico’s significant drop can be attributed to COVID-19’s effect on the nation’s wealth while Venezuela’s drop can be attributed to political instability. The implementation of TSNE clustering also underscores the importance of major global events as happiness rankings significantly dropped during 2020. The clustering in 2020 was structured much differently than any other year.

**References：**

-Robert H. Shumway, David S. Stoffer. *Time Series Analysis and Its Applications With R Examples, 4th Edition*

-*World Happiness Report up to 2022.* <https://www.kaggle.com/datasets/mathurinache/world-happiness-report>

-John F.Helliwell, Haifang Huang, Shun Wang, Max Norton. *Happiness, Benevolence, and trust during Covid-19 and beyond,* World Happiness Report. [worldhappiness.report/ed/2022/happiness-benevolence-and-trust-during-covid-19-and-beyond/](https://worldhappiness.report/ed/2022/happiness-benevolence-and-trust-during-covid-19-and-beyond/)

*- How does the Gallup Poll work*?. <https://www.gallup.com/178667/gallup-world-poll-work.aspx/>