World Happiness Report

Happyr (Group 2): Cavan Ingram, Claire Gloss, Dominic Scerbo, Lauren O'Donnell, & Qian Shen
May 4, 2022

Abstract

Happiness scores have been determined by country based on the Gallup World Poll for the past eight years. To better understand these happiness scores, we investigated what factors most contribute to the happiness score. We analyzed the countries and regions with the best and worst happiness scores and specifically examined the United States' ratings to determine how it compared to other nations. We also investigated the relationships between the contributing factors to the happiness score: economy, health, freedom, family, trust (of the government), and generosity. We tested the data for a possible linear relationship between the variables and happiness score. Our tests and observations determined that the happiest nations resided in Northern Europe and the least happy countries lived in Sub-Saharan Africa. We also found that the United States is consistently in the top 20 happiest nations and its happiness score seemed to be increasing year after year. Overall, we found economy, health, freedom, and family to have a high correlation with happiness scores and were able to fit an accurate regression equation to predict happiness scores.

Introduction

Data Introduction

The data used for our analysis of Country Happiness was obtained from the Gallup World Poll from 2015 to 2022. The happiness scores were based on answers to the Cantril Ladder question, which asked individuals to rate their life on a scale from 0 to 10. Data was polled early in the reporting year.

The survey provided a list of variables that could factor into overall happiness. These factors were ranked by the individuals surveyed from the most influential to least influential in relation to their overall happiness. The factors included economy, trust, freedom, family, and generosity. These factors did not contribute to the overall happiness score, but were utilized by the survey as potential sources of explanation for country happiness. The data also ranked countries based on their happiness score for each year and divided each country by region of the world.

A typical survey would include around 1,000 people per country to give an accurate representation of the population aged 15 and older. Bigger countries might receive closer to 2,000 responders, whereas smaller or more dangerous countries would receive closer to 500. Gallup polled with both in-person and phone interviews where the same questions were asked every year to every participating country.

Research Question

Our initial research question was to determine what factors contributed most to overall country happiness. We further divided our question of overall happiness into four additional questions as our exploratory data analysis progressed:

- What countries or regions globally have the highest happiness scores?
- What does the US happiness score look like and how is it compared to other countries?
- What features in the data contribute to happiness the most?
- Are the determined features a good predictor of happiness?

Data Processing & Methodology

The data was originally presented in eight separate data files; one file for each observed year. The data was processed to account for missing values. Happiness rank was computed via the happiness score and missing region values were completed using a region dictionary. A year column was added to each year's data set to differentiate between years.

The column names for each data file were standardized along with the data format for each column before appending all eight files into one dataset.

Results

Analysis of World Regions

We first investigated happiness scores across the globe. We created a heat map, Figure 1, that showed the average happiness scores by country. This revealed that Canada, Australia, and Northern Europe were the happiest countries. It also showed the least happy nations were generally located in Africa and South Asia.

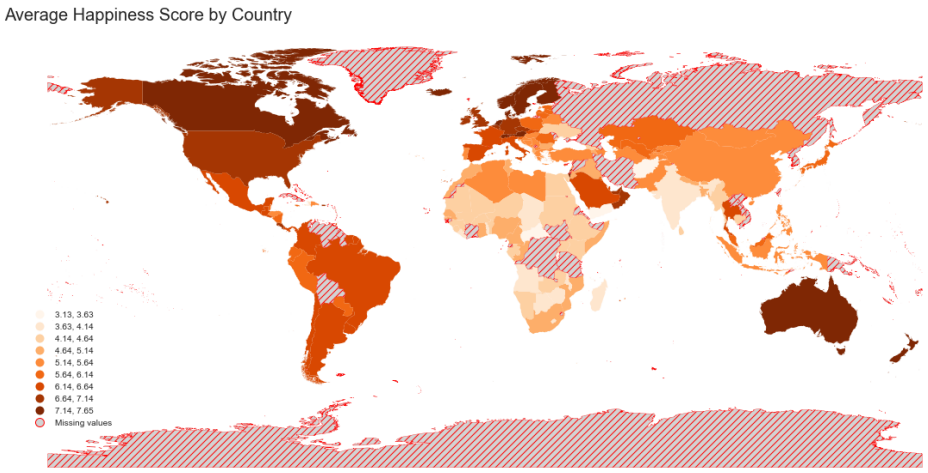

Figure 1

To identify the extremes of the happiness scores, we looked at which countries were ranked in the top and bottom five countries. Consistent with the map above, Northern European countries were consistently in the top five happiest countries. Similarly, Sub-Saharan African nations frequented the chart in the bottom five. Since the countries were relatively consistently ranked across the years, we averaged their scores and observed the overall happiest and least happy countries in the world. We compared the scores to the mean happiness score for all countries, 5.3665. In Figure 2, the happiness score for the happiest country, Finland, is 142% of the average country's happiness score. Similarly, the happiness score of the least happy country, Afghanistan, is 58% of the average score. The top five nations were averaged together, as well as the bottom five nations, and comparisons were made between them. The happiness score for the happiest nations were about 230% of the least happy countries' score, and the happiness scores for the least happy countries were about 43% of the happiest countries' score.

Rank	Countries	Happiness Scores	Delta to Mean Happiness	Delta to Average Happiest Countries	Delta to Avg Least Happy Countries
1	Finland	7.6451	142.46%	101.22%	236.69%
2	Denmark	7.5789	141.22%	100.35%	234.64%
3	Switzerland	7.5250	140.22%	99.63%	232.97%
4	Iceland	7.5213	140.15%	99.58%	232.86%
5	Norway	7.4938	139.64%	99.22%	232.00%
-5	Burundi	3.2779	61.08%	43.40%	101.48%
-4	South Sudan	3.2693	60.92%	43.29%	101.22%
-3	Rwanda	3.2680	60.90%	43.27%	101.18%
-2	Central African Republic	3.2026	59.68%	42.40%	99.15%
-1	Afghanistan	3.1322	58.37%	41.47%	96.97%

Figure 2

Mean Happiness Score
5.3665

Rank	Regions	Happiness Scores	Delta to Mean Happiness	Delta to Average Happiest Regions	Delta to Avg Least Happy Regions
1	Australia and New Zealand	7.3043	128.39%	103.81%	157.29%
2	North America	7.2635	127.68%	103.23%	156.41%
3	North America and ANZ	6.8373	120.18%	97.17%	147.23%
4	Western Europe	6.7400	118.47%	95.79%	145.14%
-4	Middle East and North Africa	5.2379	92.07%	74.44%	112.79%
-3	Southern Asia	4.5721	80.37%	64.98%	98.45%
-2	South Asia	4.4515	78.25%	63.26%	95.86%
-1	Sub-Saharan Africa	4.3140	75.83%	61.31%	92.90%

Mean Happiness Score
5.6890

Figure 3

We investigated the regions and their happiness rankings. We ranked the top and bottom four regions' average happiness scores across the eight years. Interestingly, this analysis revealed the region that is home to the top five happiest countries was not contained in the results. Figure 3 displays Australia/New Zealand as the happiest region. Based on this analysis, the Western European countries may have been better divided into smaller regions. However, the least happy regions were not surprising and aligned with the least happy nations.

Analysis of the United States

The United States was consistently in the top 20 happiest countries between 2015 and 2022. As Figure 4 shows, America's ranking started to trend down in 2018 but improved in 2022, returning to 16th place. We also observed from the table that Finland has been the world's happiest country for five consecutive years.

US Score	7.12	7.10	6.99	6.89	6.89	6.94	6.95	6.98
Rank	2015	2016	2017	2018	2019	2020	2021	2022
1	Switzerland	Denmark	Norway	Finland	Finland	Finland	Finland	Finland
2	Iceland	Switzerland	Denmark	Norway	Denmark	Denmark	Denmark	Denmark
3	Denmark	Iceland	Iceland	Denmark	Norway	Switzerland	Switzerland	Iceland
4	Norway	Norway	Switzerland	Iceland	Iceland	Iceland	Iceland	Switzerland
5	Canada	Finland	Finland	Switzerland	Netherlands	Norway	Netherlands	Netherlands
6	Finland	Canada	Netherlands	Netherlands	Switzerland	Netherlands	Norway	Luxembourg*
7	Netherlands	Netherlands	Canada	Canada	Sweden	Sweden	Sweden	Sweden
8	Sweden	New Zealand	New Zealand	New Zealand	New Zealand	New Zealand	Luxembourg	Norway
9	New Zealand	Australia	Sweden	Sweden	Canada	Austria	New Zealand	Israel
10	Australia	Sweden	Australia	Australia	Austria	Luxembourg	Austria	New Zealand
11	Israel	Israel	Israel	United Kingdom	Australia	Canada	Australia	Austria
12	Costa Rica	Austria	Costa Rica	Austria	Costa Rica	Australia	Israel	Australia
13	Austria	United States	Austria	Costa Rica	Israel	United Kingdom	Germany	Ireland
14	Mexico	Costa Rica	United States	Ireland	Luxembourg	Israel	Canada	Germany
15	United States	Puerto Rico	Ireland	Germany	United Kingdom	Costa Rica	Ireland	Canada
16	Brazil	Germany	Germany	Belgium	Ireland	Ireland	Costa Rica	United States
17	Luxembourg	Brazil	Belgium	Luxembourg	Germany	Germany	United Kingdom	United Kingdom
18	Ireland	Belgium	Luxembourg	United States	Belgium	United States	Czech Republic	Czechia
19	Belgium	Ireland	United Kingdom	Israel	United States	Czech Republic	United States	Belgium
20	United Arab Emirates	Luxembourg	Chile	ıited Arab Emirat	Czech Republic	Belgium	Belgium	France

Figure 4

To understand what contributed most to America's happiness score, we examined the data in a boxplot, separated by the different factors. It was clear that economy and family were key contributors for America's happiness score, while trust (of the government) contributed the least.

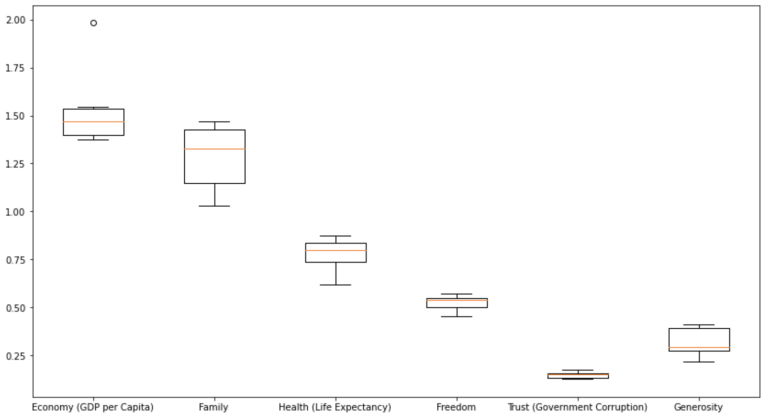

Figure 5

We then conducted trend analysis for each of the drivers and found:

- The economy rating had a significant increase in 2022, likely due to a big comeback after the pandemic. This score also drove America's overall happiness ranking back to 16th place in 2022.

- America's health scores were low due to the Covid impact during 2021 and 2022
- Interestingly, considered one of the most generous countries in the world, America's generosity score started to trend down beginning in 2018. That may be because, though the economy is strong, America's household debt had been increasing in the past few years, and that might have a negative impact on the generosity score.

Compared to the world's happiest country Finland, America lagged behind mostly in the trust score, followed by freedom and health.

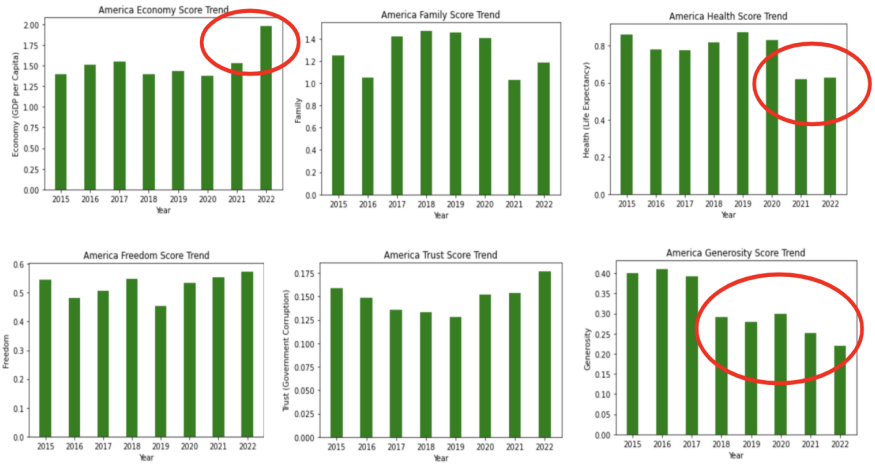

Figure 6

	Year 2022		
	Finland	United States	Difference
Rank	1	16	
Generosity	0.109	0.22	-0.111
Family	1.258	1.182	0.076
Health	0.775	0.628	0.147
Trust of Government	0.534	0.177	0.357
Freedom	0.736	0.574	0.162

Figure 7

Variable Relationships

To continue exploratory data analysis, we turned to the relationship(s) between the variables and the happiness score for each country over time.

Figure 8 displays the overall world happiness score year over year. The data revealed the overall happiness scores generally increased, except in 2017. As we continued to investigate individual variables, it became clear that world events and local events had dramatic effects on happiness scores. For example, in 2016, the world saw Great Britain vote to leave the EU, potentially resulting in a low happiness score in 2017.

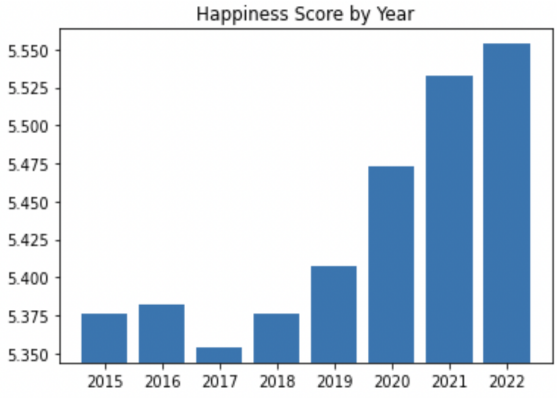

Figure 8

To determine the relationship between each variable and happiness, we plotted each variable against happiness as well as constructed a bar chart to display the change over time of the ranked importance of the happiness factor.

Economy (GDP per Capita)

The impact of economy on happiness appeared to have the strongest linear relationship, based on the scatterplot seen in Figure 9. Countries that ranked economy as most important seemed to have higher happiness scores. However, year over year, economy was not consistent in its rating, as seen in the bar chart in Figure 9. The variance could be a result of world events. For example, in 2015, the Chinese stock market crashed, but in 2017, the stock markets began to rise. 2020 saw the beginnings of COVID-19, and 2022 saw the start of the recovery from the COVID-19 lockdowns.

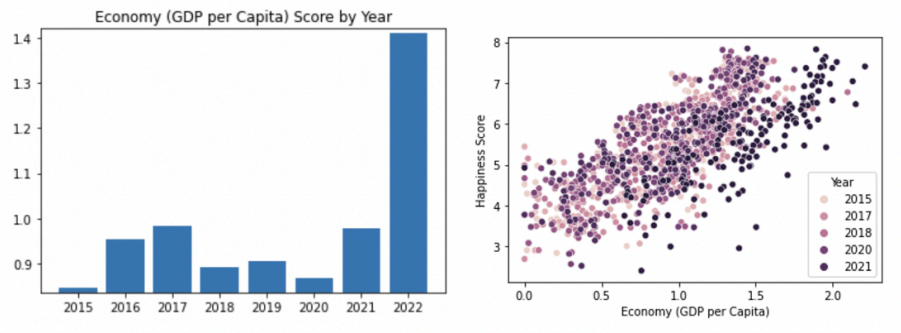

Figure 9

Health (Life Expectancy)

Health also appeared to have a linear relationship with the happiness score. Like economy, health also saw variance over the years, with 2021 seeing a steep drop likely due to the COVID-19 worldwide pandemic.

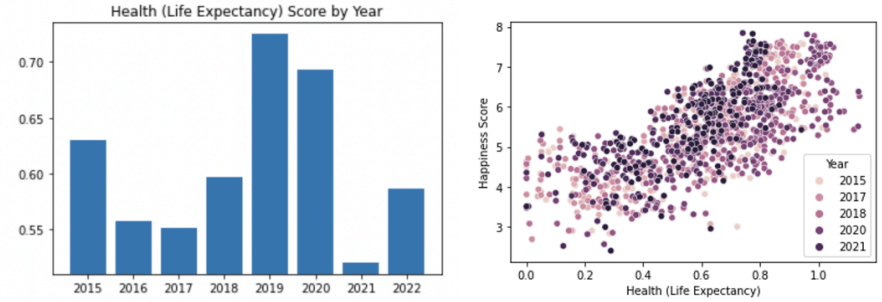

Figure 10

Freedom

Freedom also appeared to have a relatively linear relationship with happiness, though potentially weaker than that of economy and health. The data suggested variance year over year, potentially due to world events. For instance, 2016 saw the Brexit vote, Russian interference with an election, North Korean missile tests, and other impeachments/ interferences. 2019 saw many government protests worldwide.

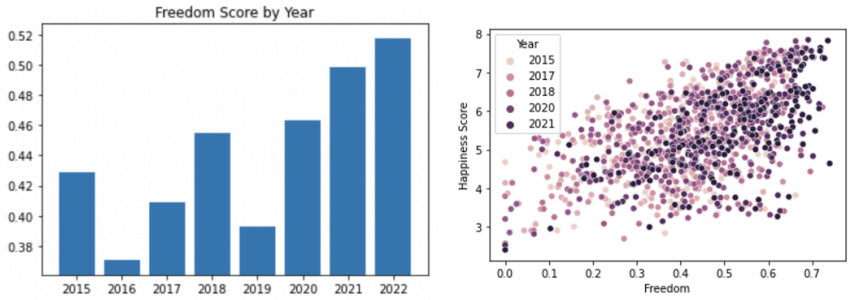

Figure 11

Family

Family's relationship with happiness score was also observed to be potentially linear. The data over the years suggested a bell-shaped curve, seeing a significant drop in 2021, which could potentially be related to families quarantining together in 2021.

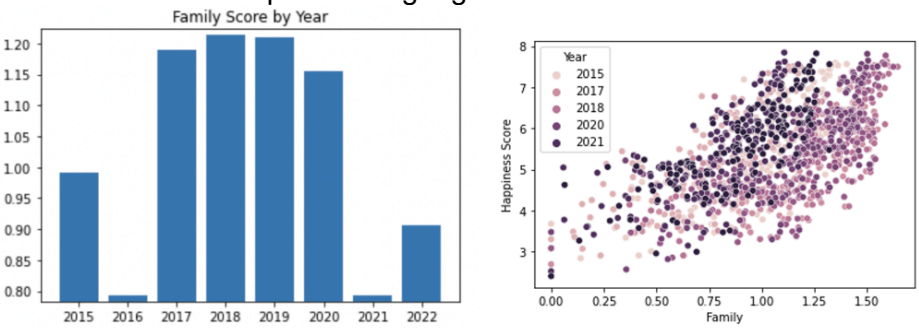

Figure 12

Trust (Government Corruption)

Trust was the first variable to have a clear non-linear relationship with the happiness score. Instead, trust appeared to have a more logarithmic shape, as seen in Figure 13. Interestingly, trust appeared to have an inverse relationship to the bell-shaped curve of family.

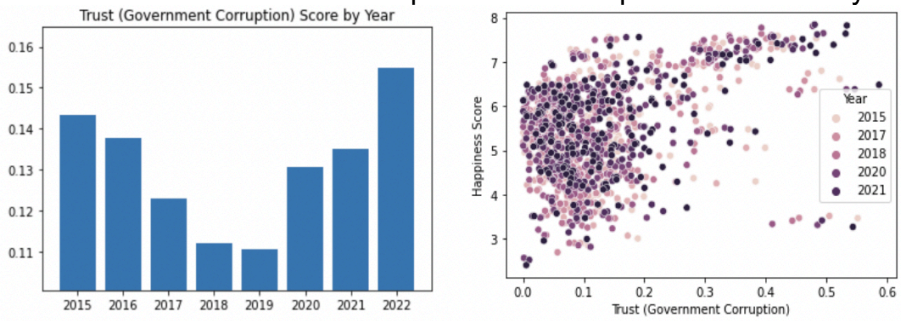

Figure 13

Generosity

Likewise to trust, generosity did not appear to have a linear relationship with happiness. Several of the years experienced major drop-offs in the importance of generosity to country happiness. Some potentially influential events include the many worldwide protests against governments in 2018 and 2019, the Me Too Movement, global warming's sharp increase, and an increase in mass shootings.

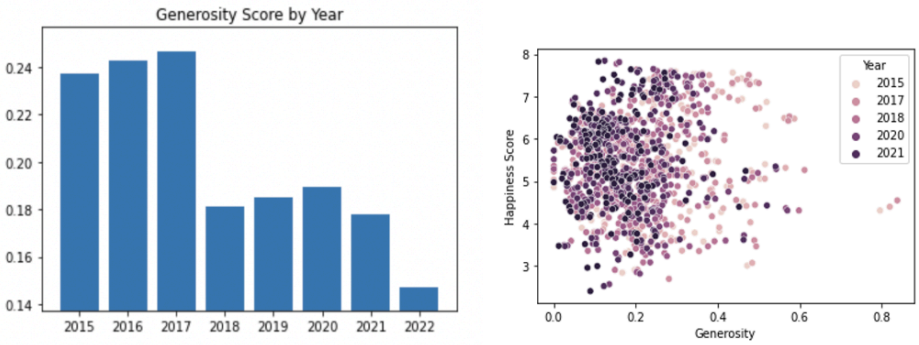

Figure 14

Multiple Linear Regression

Based on the exploratory data analysis findings of the individual variables, we explored if a multiple linear regression model could be used to predict country happiness scores. Various visualization methods were used to determine if the identified predictors were appropriate to use in a regression setting and met the basic regression assumption. We also visualized and contextually evaluated various metrics from the model in order to assess its overall performance.

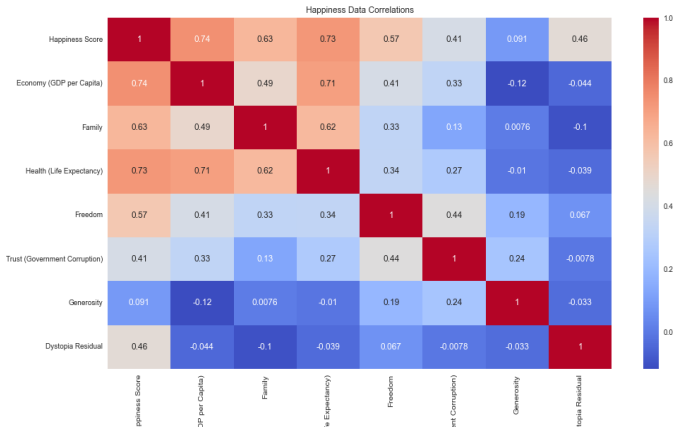

Figure 15

Using the seaborn package, we generated a heat map using the correlation coefficients between the variables, seen in Figure 15. This confirmed our initial findings that economy, health, freedom, and family have a strong positive correlation with country happiness scores and a multiple linear regression may be an appropriate model for this dataset. However, it is important to note that there were correlations between the identified predictors, a sign of multicollinearity; thus we were mindful that the model may lead to overfitting of the training data.

The correlation coefficients and the scatterplots of the potential predictors indicated that they would be appropriate to use in the model, but to further validate the utility of these predictors we individually plotted the variables against the happiness score along with a regression line and visually determined there were no clear outliers or concerning points of interest. Looking at the regression lines, we observed the variance of the data may have also had an impact on the model's performance. Therefore, prior to fitting the model, we applied the StandardScaler preprocessing function from the sklearn package to transform the data in order to reduce the variance. This method allowed easier evaluatation of the coefficients of the model contextually. Based on this assessment health, family, economy, and freedom appeared to be suitable features to evaluate for this model with the contingency that multicollinearity may have been present.

To assess potential overfitting, we split the dataset in two using the sklearn train_test_split function where 20% of the data was used as a test set. This allowed us to calculate measures of accuracy on each set to determine if the model has overfit to the training data used to fit the regression model.

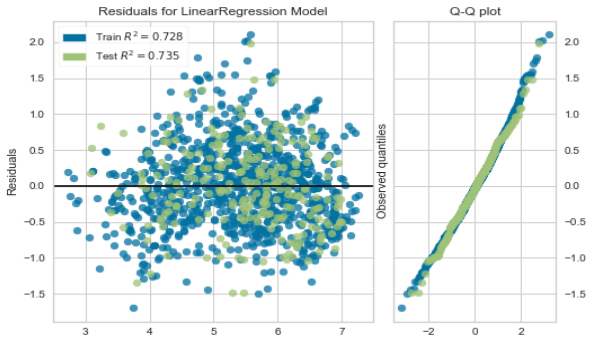

Figure 16

After fitting the model to the training data, we checked the model for the basic linear regression assumptions: 1) the mean of the error was zero for the predictor for each response 2) the variance of the errors were constant 3) the data were not collected in sequence 4) the normality assumption of the errors were met.

We quickly noted assumption 3, the data were not collected in sequence, was met. Based on the acquisition appropriate of the data collectors, we determined that the data samples are independent. After fitting the model, we were able to address assumptions 1, 2, and 4 by using the visualizations from the YellowBrick package, which generated residual and Q-Q plots for a model as seen in Figure 16. The points in each plot were colored based on their origin dataset for validation of the assumptions for each dataset simultaneously. The residual plot validated assumptions 1 and 2, indicating that the mean of the error was zero and the variances were relatively constant for a real-world setting. The points fell close to the diagonal line of the Q-Q plot meeting the normality assumption of the errors and validating assumption 4. In conclusion, for both the train and test datasets, we confirmed that we met the linear regression assumptions.

The linear regression model of this dataset was determined to be:

$$y = 5.4058499 + 0.2861x_1 + 0.2519x_2 + 0.3704x_3 + 0.2976x_4,$$

where x_1 represents the health rating, x_2 represents family, x_3 represents economy, and x_4 represents freedom.

In order to measure the overall performance of the model, we calculated the R squared value of the train and test datasets, as well as the mean absolute error, mean squared error, and the root mean squared error. R squared achieved a value of 0.73 and 0.74 for the train and test data, respectively. This indicated the model is a decent predictor of happiness score and the training data was not overfit. The Mean Absolute Error is a valuable metric for the model's predictive performance as it measures the absolute difference between the true and predicted values. In this case, the closer the metric was to zero, the better the model's performance. With a value of 0.4566, we concluded the model is a good predictor of happiness score. Similarly, the lower value of Mean Squared Error (0.3256) and Root Mean Squared Error (0.5706) indicated higher model accuracy. The Mean Squared Error measures the variance of the residuals, while the Root Mean Squared Error measures the standard deviation, applying a penalty to the larger prediction errors. In both cases, these values were below one and near-zero, indicating the model is a good predictor overall.

Unit Testing

During analysis, we conducted several unit tests. To ensure the utility and functionality of various visualization and data management processes, two unit tests were applied in Python using the `unittest` package, and two tests were conducted in R using the `testthat` library.

```
import os
import unittest
import pycountry
import geopandas as gpd

class GDFTest(unittest.TestCase):

    def test_for_geometry(self):
        fp = os.path.join(os.path.dirname(__file__),
r'data\ne_10m_admin_0_countries\ne_10m_admin_0_countries.shp')
        test_gdf = gpd.read_file(fp)[['geometry']].to_crs('EPSG:4326')
        sample_gdf = gpd.GeoDataFrame({'geometry': []}, crs="EPSG:4326")
        self.assertEqual(test_gdf['geometry'].dtype,
sample_gdf['geometry'].dtype)

    def test_country_info(self):
        country_dat = {'Germany': 'DEU', 'United States': 'USA', 'Italy': 'ITA',
'Australia': 'AUS'}
        for country_name, country_code in country_dat.items():
            result = pycountry.countries.get(alpha_3=country_code)
            result.name
            self.assertTrue(country_name==result.name, f'Expected name
{country_name} does not match {result.name}')

if __name__ == '__main__':
    unittest.main()
```

```
!python unittests.py

--------------------------------
Ran 2 tests in 0.699s

OK
```

Figure 17

The tests conducted in Python, seen in Figure 17, related to generating the Choropleth plot, Figure 1. The Choropleth plot was generated using built-in plotting functions from the `geopandas` package. To generate the geospatial plot, the function needed a reference to 1) the geospatial boundaries of the data and 2) the receptive values associated with each boundary. The geospatial boundaries were from a shapefile that contained the world's country boundaries. The `geopandas` package allows you to read the shapefile directly and structure the data similarly to a pandas data frame but with the addition of a geometry field. The data in the geometry field is automatically typed as a geometry object which allows it to be used in various geospatial analysis applications. Once the data is read into the data frame, also known as a geodataframe, you can then call the plot function to visualize the world boundaries. Before doing so, we applied a unit test to ensure that the geospatial boundaries for the countries in the geodataframe were, in fact, a geometry type and not a generic object. To do this, an arbitrary geodataframe was manually created, and a geometry field was added and manually typed as a geometry object. The shapefile was then read and directly compared the type of both geodataframe's geometry fields to ensure a match. The test passed successfully and generated geospatial visualizations using the world country boundaries shapefile.

To visualize the happiness scores associated with each country, we merged the happiness dataset with the country boundaries geodataframe. Each dataset referenced the associated country using a different key, where the happiness dataset contained the country's full name while the country shapefile contained a three-letter country code. To fuse the data, we needed to match the country name with the respective country code. To do this quickly, we leveraged the `pycountry` package. The `pycountry` package allows the user to pass in a country identifier (e.g. name, two-letter country code, three-letter country code, etc.) and returns an object with corresponding identifying information. For example, the following data represents a return object from the pycountry package with identifiers for Italy:
`Country(alpha_2='IT', alpha_3='ITA', flag='🇮🇹', name='Italy', numeric='380', official_name='Italian Republic').`
This data allowed us to coalesce the datasets with the naming conventions provided and then merge the dataset based on those associated identifying keys into a single geodataframe. In the `test_country_test unittest` function, we validated the information from the `pycountry` package by creating a sample of known country names and three-letter country codes, running the `pycountry` get function using the three-letter country code as a parameter to retrieve the country information, then compared the resulting name to the known country name. This unit test passed, which allowed us to confidently merge the dataset and ultimately enabled us to generate the Choropleth plot of the county boundaries colored by their respective happiness score.

```
##Create a function for the below code so that can meet the unit test requirement
df_yr<- function(year){
  if(year %in% seq(2015,2022,1)){
  df%>%
    filter(Year==year)%>%
    arrange(desc(Happiness.Score))
  }
  else{
    stop("year not in data set. ",.call=F)
  }

}

df2015<-df_yr(2015)
df2016<-df_yr(2016)
df2017<-df_yr(2017)
df2018<-df_yr(2018)
df2019<-df_yr(2019)
df2020<-df_yr(2020)
df2021<-df_yr(2021)
df2022<-df_yr(2022)

##Test that the function found data frame matches a manually found data frame
df2015_expected<-df%>%
  filter(Year==2015)%>%
  arrange(desc(Happiness.Score))

test_that("DF Correctness", expect_equal(df2015,df2015_expected))

##Test that the function gives an error when the year entered is not in the data frame
for(x in seq(2010,2030,1)){
  if(x %in% seq(2015,2022,1)){

  }
  else{
    test_that("DF Year",expect_error(df_yr(x)))
  }
}
```

```
Test passed
Test passed
Test passed
Test passed
Test passed
Test passed
Test passed
Test passed
Test passed
Test passed
Test passed
Test passed
Test passed
>
```

Figure 18

Additional testing was conducted to ensure that the function used to create a year-specific data frame ran correctly. The first test checked that the data frames which were outputted from the function matched those created manually. We specifically tested the data frame for the year 2015's accuracy. The test passed, indicating the function-made data frame was equivalent to the manually created data frame. The second test checked that the function identified the out-of-range years correctly. We confirmed that an error was returned when a year was entered into the function that fell outside of the years of the data set. To conserve runtime, years just outside of this range were tested instead of all years outside of the range. Each of these years passed, indicating that the function will only work for years within the range 2015 to 2022.

Conclusion

We concluded the strongest indicators of country happiness between 2015 and 2022 were economy, health, freedom, and family. These variables contributed to a strong linear regression model in predicting a happiness score for a given country. The North American, European, and Australian regions had overall higher happiness scores year over year, while African and Middle Eastern regions had the lowest scores. We observed the United States' happiness score was consistently ranked in the highest twenty happiness scores from 2015 to 2022 and increased in score each year.

Practical applications of this analysis include determining the best countries to live in based on citizen happiness and which contributing factors to happiness should be most considered to increase country-wide and worldwide happiness scores. Our findings can also be used to help humanitarian efforts by identifying the regions that may need assistance and the factors that need to be improved upon, such as healthcare and freedom.

Due to the simplicity of the world happiness data set, we were not able to control for many factors in our analysis. The Gallup World Poll was conducted via cold calling. Countries with limited telecommunication technology may have lower response totals and skewed responses, due to the increased access to such technologies in higher socioeconomic classes. Therefore, if the analysis were repeated, we recommend collecting data such as the number of surveys completed per country, number attempted, and total country population to appropriately weigh country happiness scores. Data including salaries of survey responders and overall country salaries, spending habits, and available technologies may also be considered for calculating score weights and other contributing factors.

We observed the main contributing factors sometimes differed between 2015 and 2022. We speculated that world events may have impacted the happiness score rating and ranking of these factors. Incorporation of or additional research into impacting world and local events could provide additional context to the variance of the contributing factors year to year, country to country, and region to region. This may lead to the ability to predict change in happiness over time as similar events come to fruition both within the affected countries, regions, and worldwide.