Link to GitHub Repository: https://github.com/CamYench/DATASCI200Project2.git

Happiness and the S&P 500

DATASCI 200 Project 2 Report

April 19, 2023

Hannah Grace Smith, Cameron Yenche, Matthew Zhang

Overview

The purpose of this project is to compile happiness data, examine major trends, and determine the relation–if any–to the trends of the S&P 500 stock market index. From this report, we intend to explore and find data-driven backing for lifestyle choices and external factors that are generally associated with greater levels of happiness. By providing an overview of data pertaining to socioeconomic factors, we hope to distill trends and insight from the data into practical recommendations that can be incorporated into life to improve one's subjective well-being.

Data

The report will employee data analysis techniques to explore two datasets:

2005–Present World Happiness Report
The world happiness report dataset contains several variables, including Log GDP Per Capita, Health Life Expectancy at Birth, Social Support, Freedom to Make Life Choices, Generosity, Perceptions of Corruption, Positive Affect, Negative Affect, Date, Country Name, and Regional Indicator

The variables Log GDP Per Capita, Health Life Expectancy at Birth, Social Support, Freedom to Make Life Choices, Generosity, are all independent variables that we hypothesize to be positively correlated with happiness. Perceptions of Corruption and Negative Affect are also an independent variable, but we hypothesize it to be negatively correlated with happiness.

S&P 500 Index Market Trend Data
The second dataset that we plan to use is the historical market trend data from S&P 500 over the 2005-present time frame, which contains Close/Last, Market Close Value of S&P 500, Open/High/Low Index value, and Date.

For the purpose of our report, we will focus primarily on the Market Close Value of the S&P 500 and the corresponding date values, but we note insight can also be gleaned from Open/High/Low index values when taken as a metric for daily market volatility.

Key Questions

What variable(s) is/are correlated most strongly and consistently with increased subjective well-being worldwide?

Are there regional differences in how the variables in the World Happiness report relate to subjective well-being?

What is the relationship with the state of the economic market, as measured by the S&P 500, on worldwide and regional measures of subjective well-being?

Are there differences in how the economic market, as measured by the S&P 500, relates to various regions' subjective well-being ratings?

Methodology

Initial Exploration and Cleaning

Before merging the 2005 - Present World Happiness Report data with the S&P 500 Index Market Trend data, our data set has 2199 rows and 19 columns. With exceptions for Country Name, Regional Indicator, and Year, most columns contain floats. Most data manipulation will address missing values and grouping to explore the data set by region/country.

Sanity Checking the Data

We began our investigation by first ensuring that it is valid and establishing that the following conditions were met:

- Log GDP Per Capita, Health Life Expectancy at Birth
    - Float type, no values are overly high or overly low, all values are positive
- Social Support, Freedom to Make Life Choices, Perceptions of Corruption, Positive Affect, Negative Affect
    - Float type, values between 0 and 1, all values are positive
- Generosity
    - Float type
- Country Name, Regional Indicator
    - Object/string type, countries and regions exist, countries and regions correspond in a consistent and accurate manner
- Year
    - Datetime or integer type, year ranges align with 2005 through 2023

All values from each of the columns were within the realm of our expectations. After we confirmed that the data set did not contain any duplicate values, we were ready to dive into cleaning the data.

Reconciling Dates

As these are two unrelated datasets, we cannot immediately assume that the same date range can be used for both; after conducting a brief examination of the data, we found that the S&P 500 index data and World Happiness dataset start dates needed to be reconciled. Therefore, we removed all data from both datasets before 2014 and all data from the S&P 500 dataset that was after the end of 2022 (i.e., December 30, 2023).

Missing Values

In our exploration, we discovered that there were some missing values that needed to be dealt with. Interestingly, the column with the most missing data is Confidence in National Government, which might be explained by challenges with data collection.

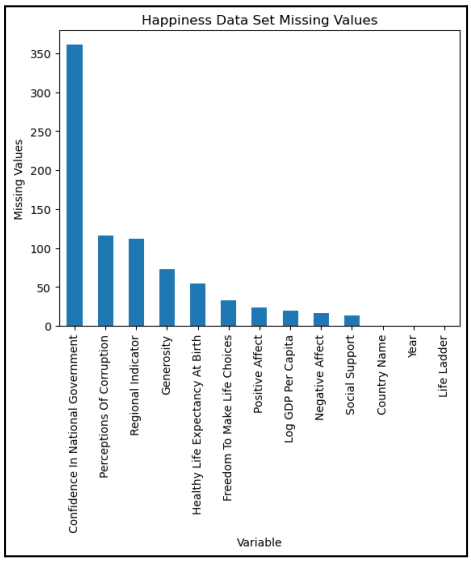


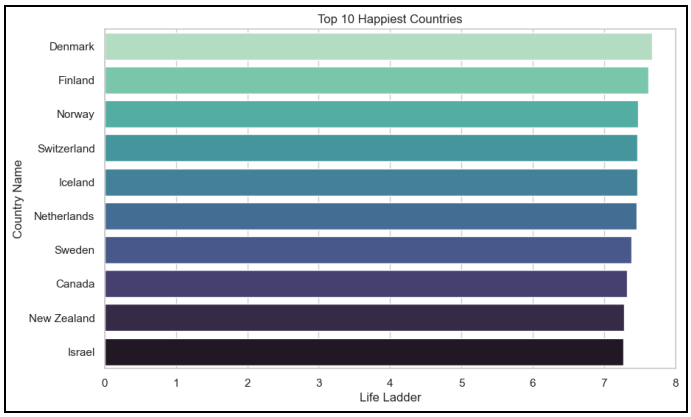
Figure 1: This figure depicts the top ten variables with the highest number of missing values

Regarding missing values for Regional Indicator, we addressed this issue by using a dictionary taken from Kaggle to assign the correct Regional Indicator for missing null values in that column by Country Name. For remaining numerical columns, we used imputation to address missing values, filling the nulls for each column based on the mean of existing values for that respective Country Name through “groupby” clauses. All other rows with missing values were dropped.

Exploratory Data Analysis

What countries appear to be the happiest? The least happy?

According to the data, the top happiest countries are Denmark, Finland, Norway, Switzerland, Iceland, Netherlands, Sweden, Canada, New Zealand, and Israel. Inversely, the countries that ranked lowest in terms of happiness are Afghanistan, Central African Republic, Burundi, Rwanda, Togo, Tanzania. Zimbabwe, Comoros, Yemen, and Botswana.



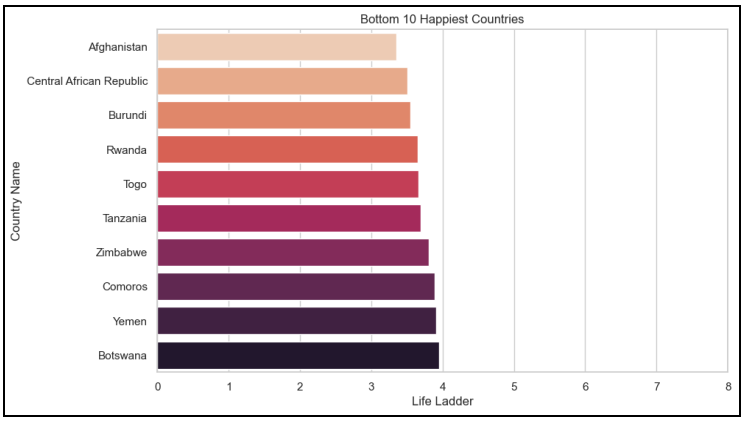


Figure 2 (left): This figure depicts the top ten countries with the highest life ladder ratings

Figure 3 (right): This figure depicts the top ten countries with the lowest life ladder ratings

After looking at these country groupings, we realized that there seems to be a geographic pattern with the data. Through a visual representation of the regions, we can see that the happiest regions appear to be around North America and Western Europe, while the least happy regions appear to be South Africa and Southeast Asia.

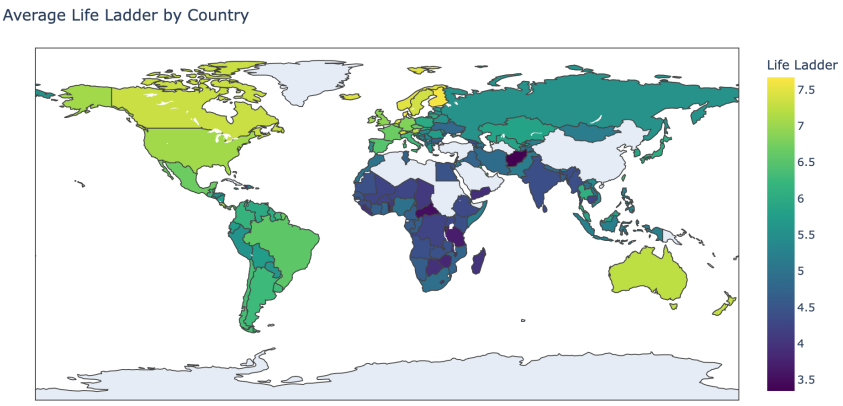
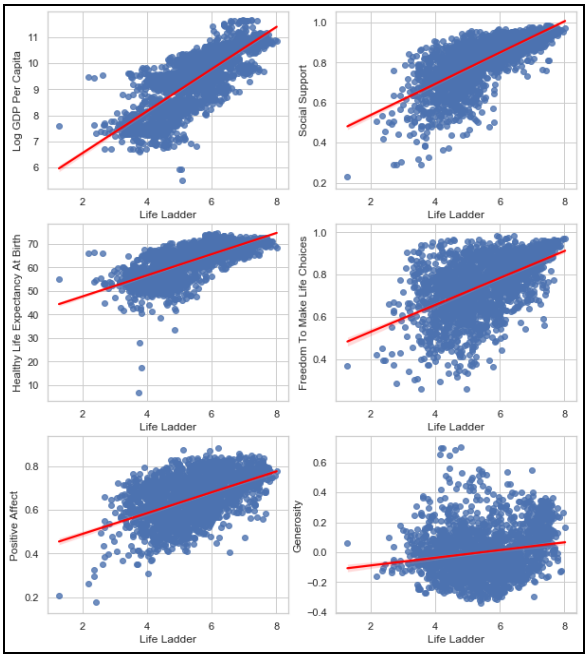


Figure 4: This figure depicts the each country's average life ladder rating

What variables show the strongest correlations with self-identified happiness level?

In our research, we determined that the variables with the strongest positive correlations with perceived happiness are Log GDP Per Capita, Social Support, Healthy Life Expectancy at Birth, Freedom to Make Life Choices, Positive Affect, and Generosity.



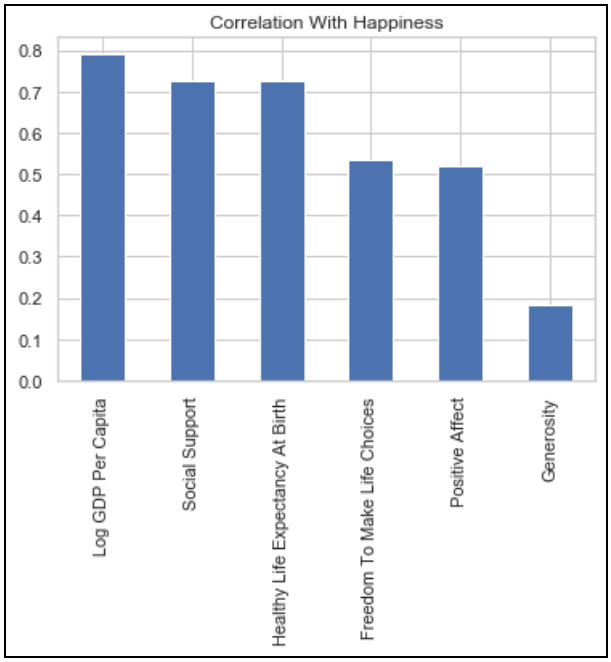


Figure 5 (left): This figure depicts scatter plots with lines of best fit to indicate correlation between each variable and life ladder

Figure 6 (right): This figure depicts the top six variables with the highest correlation with life ladder

Which of the happiest countries are consistent in the Top 10 ranks for other variables?

We can see that many of the top 10 happiest countries rank in the top 10 of the variables that are highly correlated with subjective well-being. These countries tend to remain consistent with other variables such as Log GDP Per Capita and Social Support. For example, we can see that Ireland ranks in the top 10 happiest countries as well as the top 10 for both Log GDP Per Capita and Social Support. Similarly, for Finland, Denmark, Norway and Switzerland.

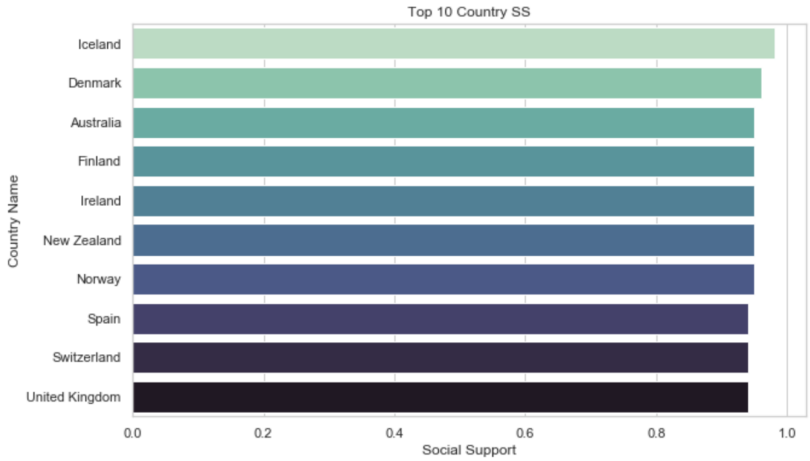


Figure 7: This figure depicts the top ten countries with the highest social support

When looking at these columns in a correlation matrix, we also notice that each of these variables appear to be more highly correlated with each other, not just with our variable of interest, happiness. This demonstrates that, although our correlations might draw our attention when we look at them independently of each other, there may be some underlying relationship between many of the variables. The cluster in the right left corner below further illustrates this.

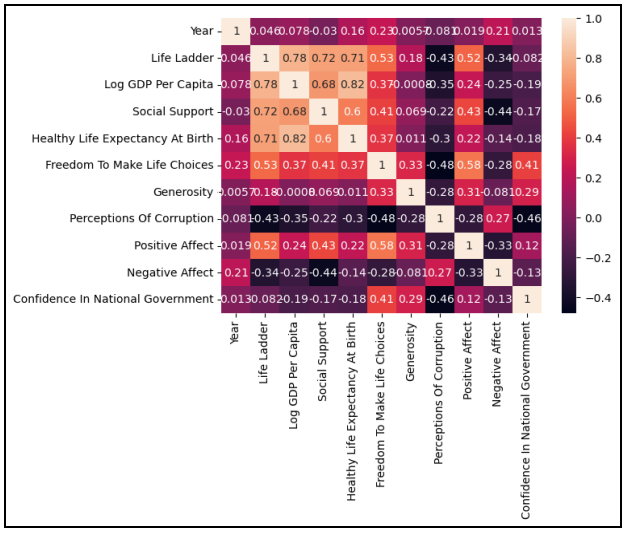
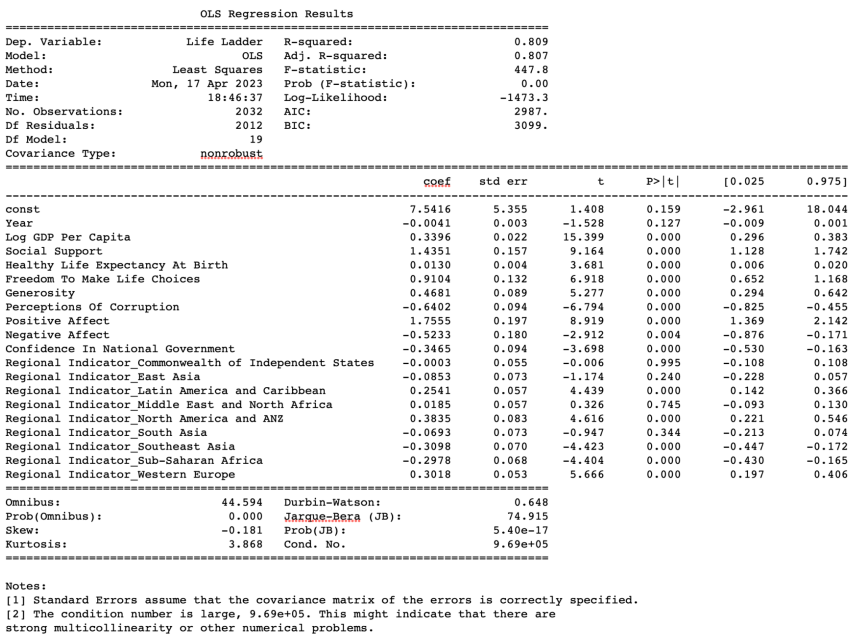


Figure 8: This figure depicts the top ten countries for the least perceived levels of corruption

When accounting for other variables, what variables are most correlated with happiness?

After spotting the relationship between the variables that appear most highly correlated with happiness, we wanted to further explore the interrelationship between each of these by using a multivariate regression. For this, we decided to drop Country and created dummy variables for Regional Indicator, as we believed that this column would yield more valuable information due to its broadness. We do not intend to use this as a predictive tool but rather as a way to further explore and corroborate our findings from before. Ultimately, this suggested the same conclusions as in our prior EDA, even when we control for all the variables.



OLS Regression Results

Dep. Variable:	Life Ladder	R-squared:	0.809
Model:	OLS	Adj. R-squared:	0.807
Method:	Least Squares	F-statistic:	447.8
Date:	Mon, 17 Apr 2023	Prob (F-statistic):	0.00
Time:	18:46:37	Log-Likelihood:	-1473.3
No. Observations:	2032	AIC:	2987.
Df Residuals:	2012	BIC:	3099.
Df Model:	19		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	7.5416	5.355	1.408	0.159	-2.961	18.044
Year	-0.0041	0.003	-1.528	0.127	-0.009	0.001
Log GDP Per Capita	0.3396	0.022	15.399	0.000	0.296	0.383
Social Support	1.4351	0.157	9.164	0.000	1.128	1.742
Healthy Life Expectancy At Birth	0.0130	0.004	3.681	0.000	0.006	0.020
Freedom To Make Life Choices	0.9104	0.132	6.918	0.000	0.652	1.168
Generosity	0.4681	0.089	5.277	0.000	0.294	0.642
Perceptions Of Corruption	-0.6402	0.094	-6.794	0.000	-0.825	-0.455
Positive Affect	1.7555	0.197	8.919	0.000	1.369	2.142
Negative Affect	-0.5233	0.180	-2.912	0.004	-0.876	-0.171
Confidence In National Government	-0.3465	0.094	-3.698	0.000	-0.530	-0.163
Regional Indicator_Commonwealth of Independent States	-0.0003	0.055	-0.006	0.995	-0.108	0.108
Regional Indicator_East Asia	-0.0853	0.073	-1.174	0.240	-0.228	0.057
Regional Indicator_Latin America and Caribbean	0.2541	0.057	4.439	0.000	0.142	0.366
Regional Indicator_Middle East and North Africa	0.0185	0.057	0.326	0.745	-0.093	0.130
Regional Indicator_North America and ANZ	0.3835	0.083	4.616	0.000	0.221	0.546
Regional Indicator_South Asia	-0.0693	0.073	-0.947	0.344	-0.213	0.074
Regional Indicator_Southeast Asia	-0.3098	0.070	-4.423	0.000	-0.447	-0.172
Regional Indicator_Sub-Saharan Africa	-0.2978	0.068	-4.404	0.000	-0.430	-0.165
Regional Indicator_Western Europe	0.3018	0.053	5.666	0.000	0.197	0.406

Omnibus:	44.594	Durbin-Watson:	0.648
Prob(Omnibus):	0.000	Jarque-Bera (JB):	74.915
Skew:	-0.181	Prob(JB):	5.40e-17
Kurtosis:	3.868	Cond. No.	9.69e+05

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 9.69e+05. This might indicate that there are
strong multicollinearity or other numerical problems.

Figure 9: This figure depicts the output of a multivariate regression for our dependent variable, life ladder

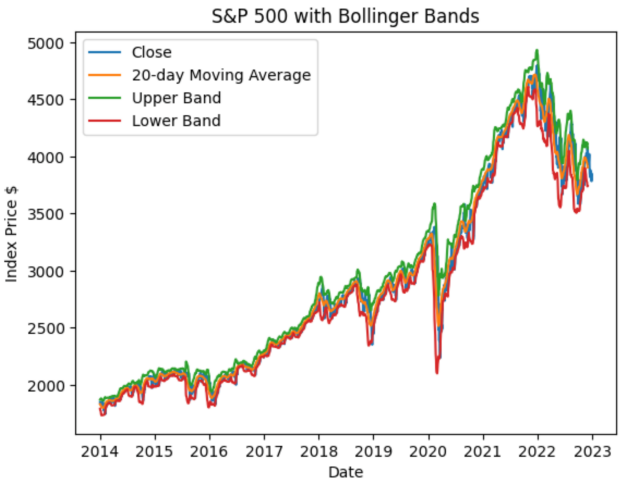
What happens when we bring in the S&P 500 data?

To integrate the data from the S&P 500, we must create yearly metrics that are compatible with the yearly data found in the World Happiness Report Data. For this report, metrics of interest are how many the S&P 500 experienced a large price decline and a large price increase (behavior that is akin to a small recessionary pull-back or a relative boom in index performance) and the volatility of the index over each year.

We are taking the S&P 500 index to be an approximation of the health of the American stock market (as is common) and proxy for the state of the American economy, which has a heavy influence on the world economy. To distill the daily market close price of the index into yearly metrics, we will apply the concept of Bollinger Bands to track the relative trends in the index's pricing over the course of a year. Bollinger Bands are formed by taking the moving average (we chose to use the standard value of 20-days as a moving average window) and plotting two standard deviations for that same moving average window above and below that moving average trend line.

Using these bands, we can develop concrete metrics for quantifying yearly cyclical trends and volatility. For example, when the close price breaks above or below the Bollinger bands, this can be taken as an indicator that the index is being overbought or oversold, which is indicative of general market booms and busts. These band breaks can then be compiled by year and used as a metric for market performance for each year. Similarly, the width between the bands can be taken as a measure of volatility, with a large bandwidth corresponding with a period of high volatility and a small bandwidth indicating a period of low volatility. The Bollinger bandwidth can then be averaged for each 20-day window in the year to generate a yearly metric for volatility. A

plot of the S&P 500's index trendline with Bollinger bands as well as a plot of yearly volatility are each depicted in the plots below. Note that the relative tightness of the Bollinger bands makes it difficult to visually interpret, so we have included a more detailed plot for each year within the appendix.



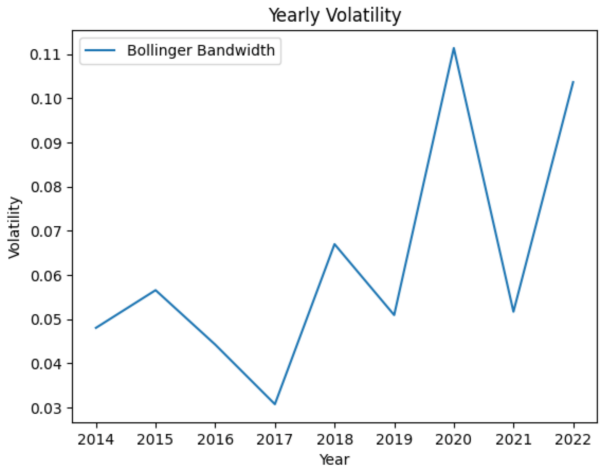
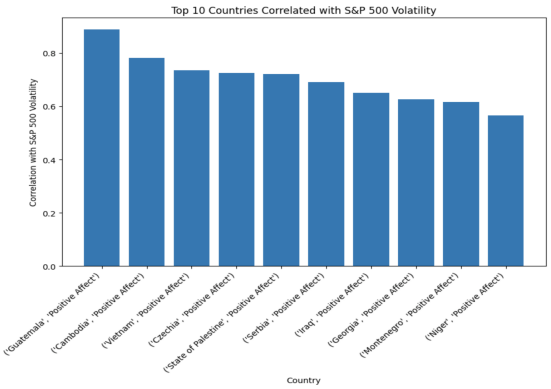


Figure 10 (left): This figure depicts the S&P 500 Market Index closing price with moving average and Bollinger Bands for the years 2014-2022

Figure 11 (right): This figure depicts the S&P 500 volatility as measured by Bollinger Bandwidth averaged for each year from 2014-2022

From these plots, we can easily recognize general trends in volatility, with most notable volatility occurring in the period associated with the COVID-19 pandemic (i.e., 2020-2022). Now we will examine the countries, whose positive affect is most strongly correlated positively and negatively with volatility:



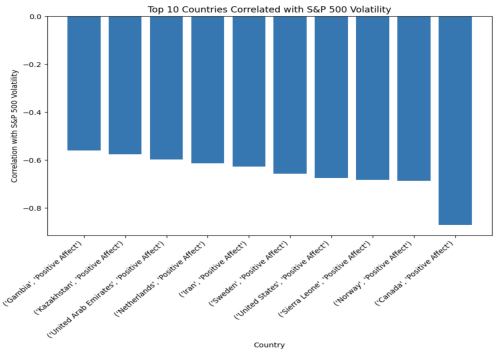


Figure 12 (left): This figure displays the top 10 countries whose positive affect is most strongly positively correlated with volatility in the S&P 500

Figure 13 (right): This figure displays the top 10 countries whose positive affect is most strongly negatively correlated with volatility in the S&P 500

From these two plots, we can readily identify that both strong positive correlations and strong negative correlations exist. Let's first overview the top 10 countries whose positive affect is correlated positively. Since the S&P 500 is an index used to represent the American economy, we can prognosticate that many of these countries may experience an increase in positive affect when there is increased volatility in the American Stock market either due to their historical amiability to the United States or their relative economic isolation to the effects of this increased isolation. As we take a closer look at the countries whose positive affect is negatively correlated with volatility in the S&P 500, we countries that align more closely with predictions, including the United States itself along with known allies and trading partners, such as Canada and the UAE. Interestingly, Canada has the most extreme negative correlation coefficient rather than the United States itself.

As we start to uncover patterns with countries' positive affect and S&P 500 that relate to the economic relationships with the United States, we will also want to explore whether regional relationships exist. To examine this, we compiled countries' correlation coefficient between positive affect and the volatility of the S&P 500, shown in the figure below:

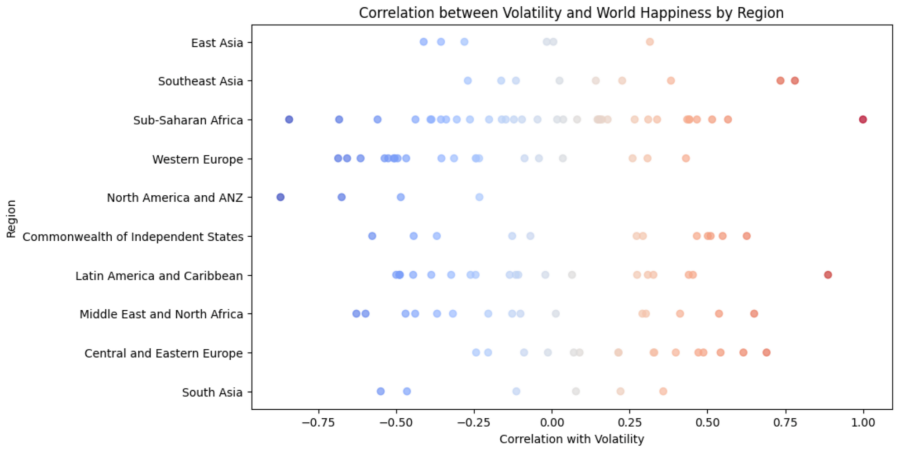


Figure 14: This figure displays a mapping of correlation coefficients for the relationship between positive affect and volatility of the S&P 500 index organized by countries' regions.

From this visualization, we can gather that there is no generalizable relationship between countries' positive affect and the volatility of the S&P 500 by region. However, this plot does highlight outliers with either a strong negative or positive correlation. This leads us to the conclusion that it is most effective to analyze this relationship on a country-by-country basis, rather than on a regional level.

Conclusion

Our original goal was to compile World Happiness Report (WHR) data to observe any underlying trends and determine if there existed a relation to the trends of the S&P 500 stock market index. To ensure that we conducted a complete report of the two datasets, we devised a structure for our analysis. In data cleaning/preprocessing and exploratory data analysis (EDA) we were able to draw conclusions to answer our key questions.

We began by cleaning the WHR dataset. While a majority of the data was present, we found that the existing null values had the potential to skew our results. The 3 variables with the most amount of null values in Confidence In National Government, Perceptions of Corruption, and Regional Indicator from most to least null values, respectively. We decided that the best way to combat these null values was to impute by country mean in order to maintain the variance in our data; then to drop any remaining rows. As for the S&P 500 stock market data, there were no null values. However, we had to conduct preprocessing as there was only data down to the year 2014. Thus, we had to subset our data to match the two datasets together and meet our desired timeframe from 2014-2022.

Moving on to our EDA, we wanted to answer our questions both visually and quantitatively. We found that the factors that have the biggest impact on overall happiness (measured by Life Ladder) were Log GDP Per Capita, Social Support, Life Expectancy At Birth, and Freedom To Make Life Choices. We found that these factors are generally indicative of a country's economic performance and accompanied by overall wealth, infrastructure and financial stability. Furthermore, we were able to confirm our findings by implementing a correlation heatmap and multiple regression. We found that the "happiest" region was North America and ANZ, while the "least happy" region was Sub-Saharan Africa. However, at a country level, the results differ. It was interesting to find that the "happiest" country was Denark, while the "least happy" region was Botswana.

There is a notable decrease in S&P 500 Stock Market prices in the years 2015 and 2022 from the previous year. Similarly, comparing the means of our variables for each of these years, it is evident that there is a decrease in Generosity and Social Support. We aimed to represent recessions through evidence of large price declines and found that the biggest trends in volatility were noticeable during the period of the COVID-19 pandemic (2020-2022). In our further analysis, we found that there is no generalizable relationship between countries' positive affect and the volatility of the S&P 500 by region.

From this report, we intend to explore and find statistical backing for lifestyle choices and external factors that are generally associated with greater levels of happiness. We can conclude that we have been able to answer these questions through our analysis. For future analysis, we hope to implement an optimized multiple regression to predict potential happiness for specified countries. We hope that this data can direct attention to a worldwide demand for more happiness and reflects well-being as a criteria for government policy and economic stability. We found that the variables most highly associated with happiness are indicators of the variation between different countries as demonstrated by individual reports. Ultimately, we have identified the happiest countries and regions in the world as well as the key variables that determine that happiest; we can recommend the positive aspects in governmental policy and the economic modalities of these countries to improve the overall happiness of our own.

Appendix

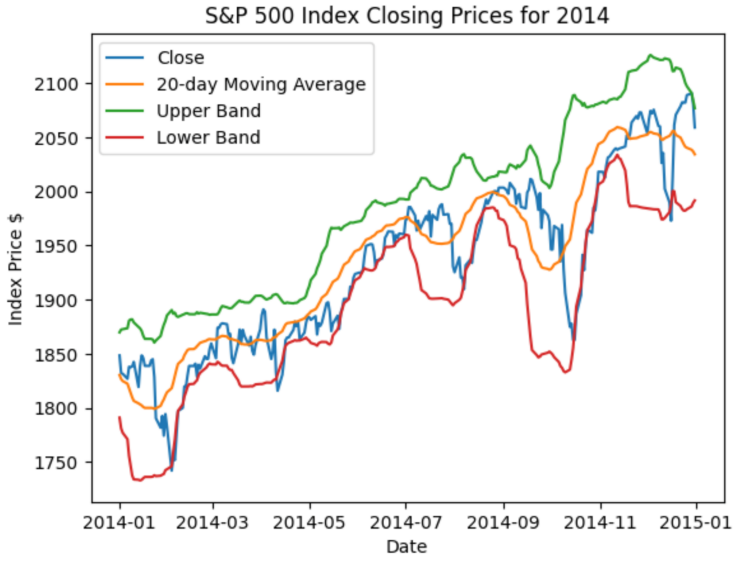


Figure 16: This figure depicts the S&P 500 Market Index closing price with moving average and Bollinger Bands for the year 2014

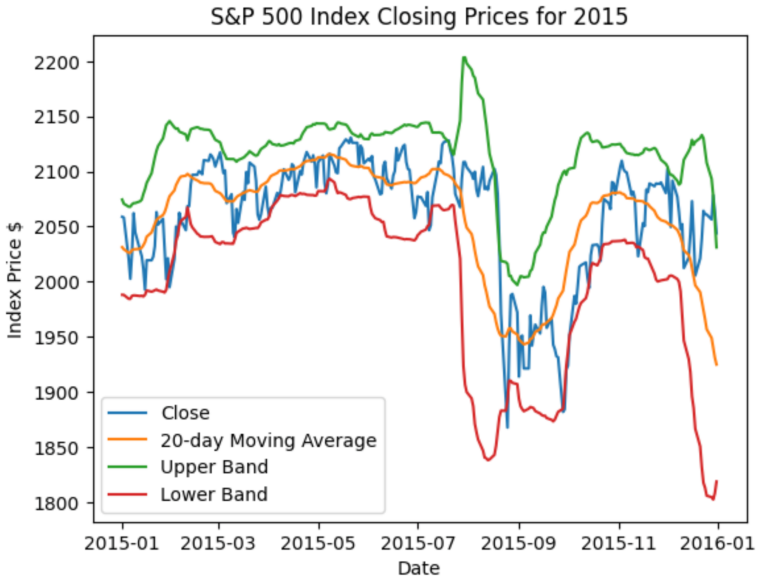


Figure 17: This figure depicts the S&P 500 Market Index closing price with moving average and Bollinger Bands for the year 2015

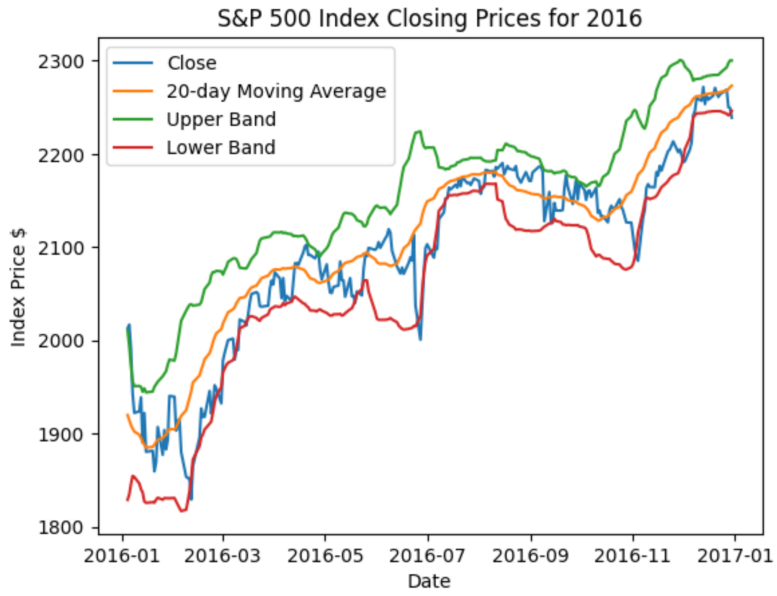


Figure 18: This figure depicts the S&P 500 Market Index closing price with moving average and Bollinger Bands for the year 2016

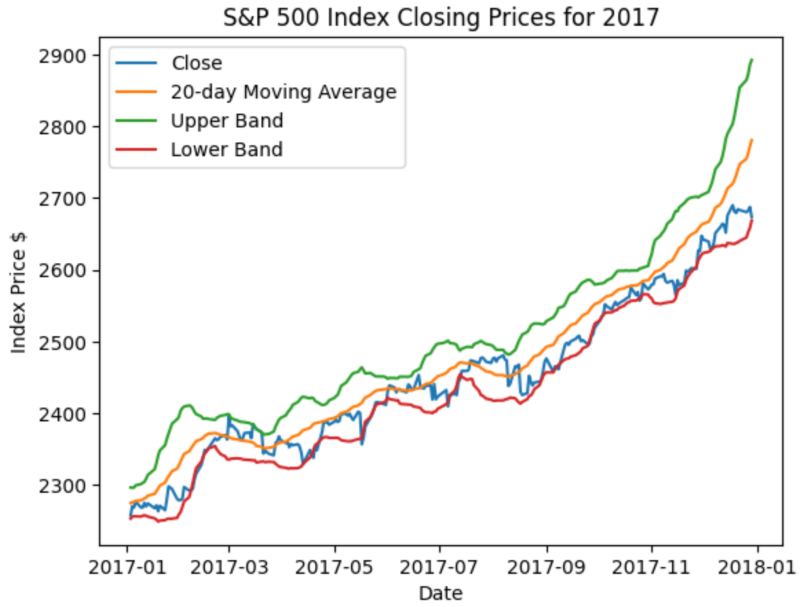


Figure 19: This figure depicts the S&P 500 Market Index closing price with moving average and Bollinger Bands for the year 2017

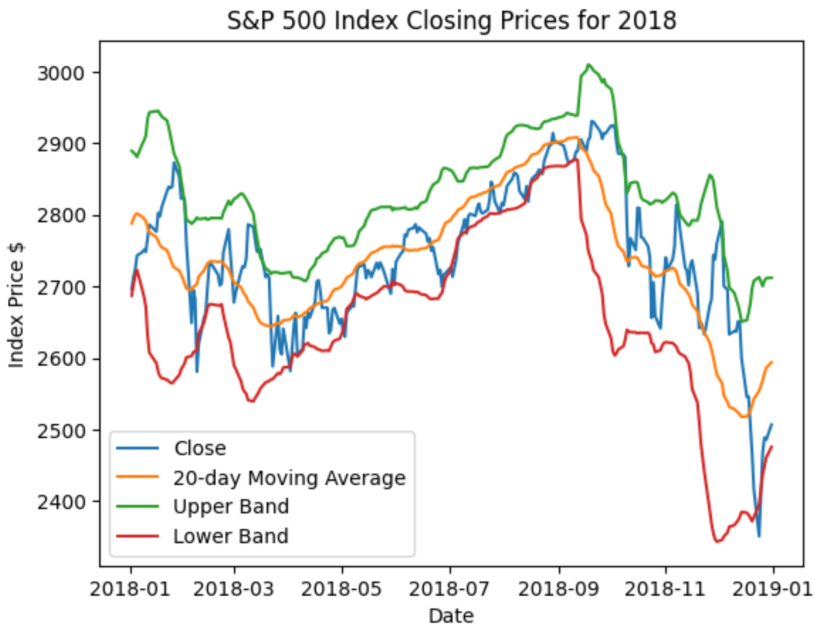


Figure 20: This figure depicts the S&P 500 Market Index closing price with moving average and Bollinger Bands for the year 2018

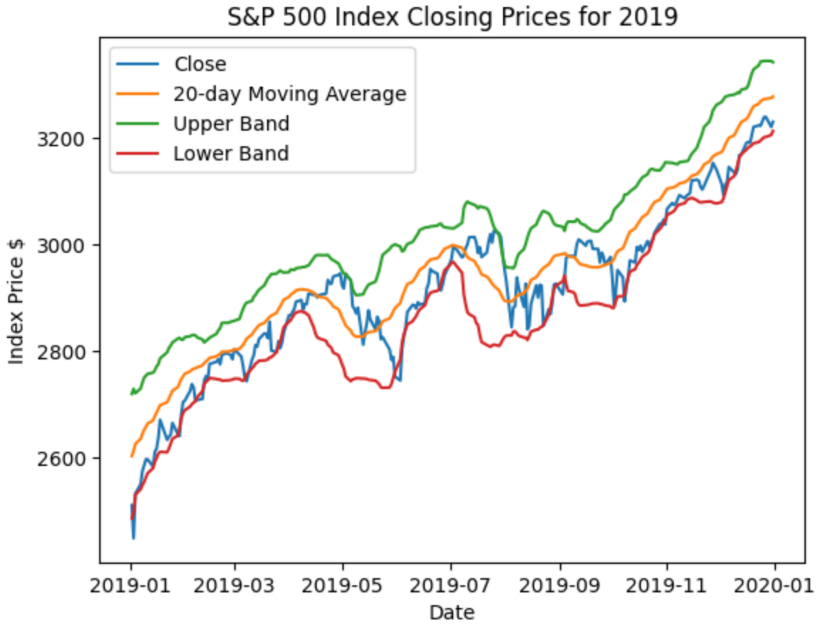


Figure 21: This figure depicts the S&P 500 Market Index closing price with moving average and Bollinger Bands for the year 2019

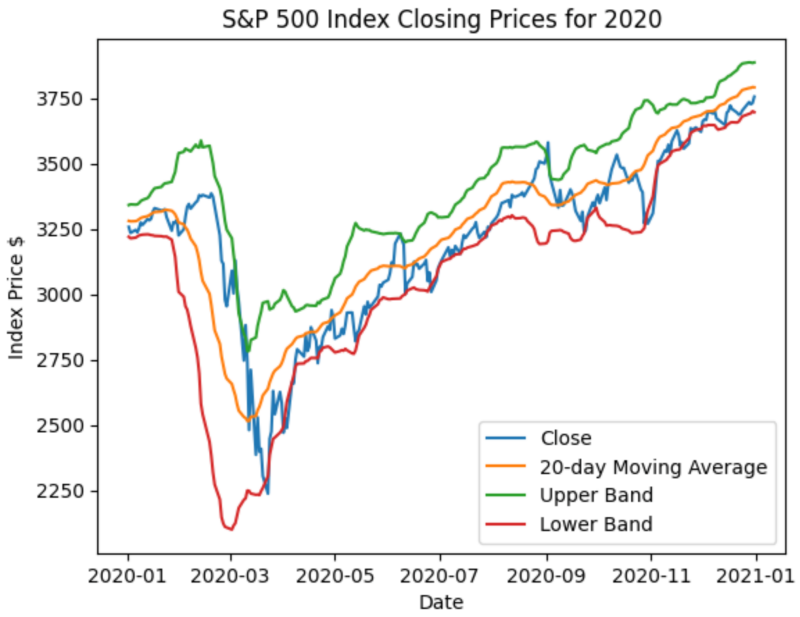


Figure 22: This figure depicts the S&P 500 Market Index closing price with moving average and Bollinger Bands for the year 2020

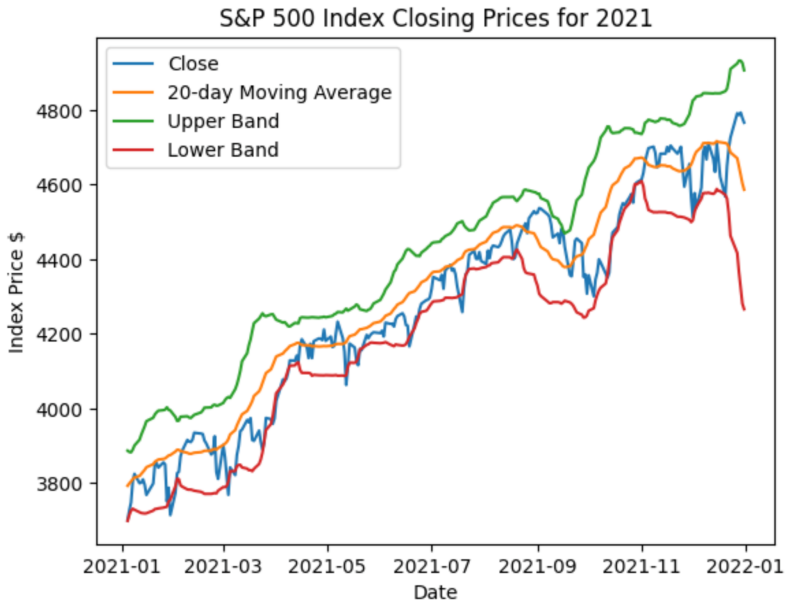


Figure 23: This figure depicts the S&P 500 Market Index closing price with moving average and Bollinger Bands for the year 2021

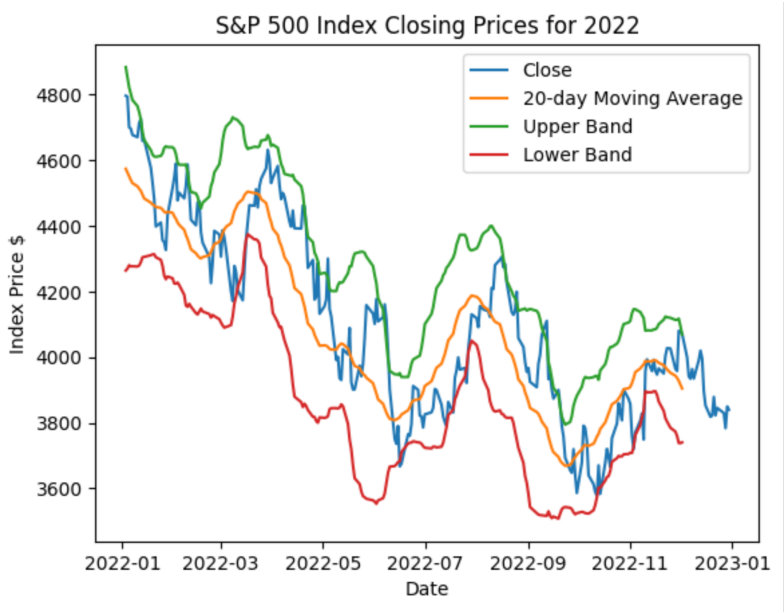


Figure 24: This figure depicts the S&P 500 Market Index closing price with moving average and Bollinger Bands for the year 2022

References

Data Sources

https://www.kaggle.com/datasets/usamabuttar/world-happiness-report-2005-present/code

https://www.nasdaq.com/market-activity/index/spx/historical