JSC370 Final Project

2023-04-28

Contents

Introduction

Happiness is a viral topic in the modern world, as it is ultimately what everyone longs for. However, according to the World Health Organization, 1 in 8 people suffer from mental disorders, and there has been a 25% increase in depression after COVID-19. With this, I am prompted to know more about how happiness has progressed overtime and the factors that influence it. For instance, in the academic researcher, Satoshi Kanazawa's paper, "Sunshine on my shoulders makes me happy... especially if I'm less intelligent: how sunlight and intelligence affect happiness in modern society", he discovers that darkness induces fear and anxiety while exposure to sunlight increases happiness. Not only do I want to explore the effect of sunshine hours, I am also interested in other factors, such as economic status and freedom in a country.

Perhaps, with this report, I hope to provide insights to government policies or personal actions that can help improve the well-being of individuals. Therefore, combining these motivations, I hope to look at happiness from three different aspects, physiological needs (economic status), regional factors (sunshine), and emotional needs (freedom, love). My question of interest is "Do people get happier overtime? Does economic status, freedom to make life choices, and sunshine hours affect people's happiness?".

My first data set is retrieved from a renown source, World Happiness Report, which is a report written by members of the United Nations Sustainable Development Solutions Network. It makes use of the "Gallup World Poll", a global survey data that consists of over 100 questions including regional specific questions. The report provides insights in how people evaluate their own lives by country. The data set I will be using has records starting from 2005 to 2021, where the happiness score - the national average response to the question of measuring oneself on life ladders - is recorded for more than 150 countries worldwide. Other variables I will be using from the data set are log GDP per capita obtained from World Development Indicators, and the freedom to make life choices and perceptions of corruption, results obtained from Gallup World Poll.

The second data set is retrieved from Wikipedia, which contains a list of sunshine hours for 391 cities, including the yearly and monthly duration. Records of this data set are compiled from numerous sources.

The third data set I utilized is a world cities data retrieved from SimpleMaps which contains information, such as the population and geographical information (latitude and longitude) for over 40k cities worldwide. This data set is compiled from numerous reliable sources, which includes NASA, National Geospatial-Intelligence Agency, U.S. Census Bureau, and U.S. Geological Survey. This data set will specifically be used with the sunshine data set to find which cities are a better representation for the country in terms of sunshine hours.

Methods

Data Collection

The happiness data set is downloaded from the World Happiness Report website. The process is as easy as a simple download for "Data for Table 2.1", which is downloaded as an excel file. I later converted to a csv myself.

The second data set uses the "rvest" package from R to perform web scraping from Wikipedia; it uses the URL to read HTML code and CSS selectors to subsequently scrape the table.

The cities census data set is retrieved from SimpleMaps, which is available for download in csv or excel format.

Data Wrangling & Cleaning

Data 1: Sunshine Data set

This is a sample table of the data set.

Table 1: Raw Data from Wikipedia Sunshine Hours

Country	City	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec	Year	Ref.
Ivory Coast	Gagnoa	183	180.0	196.0	188.0	181.0	118.0	97.0	80.0	110.0	155.0	171	164.0	1,823.0	[2]
Ivory Coast	Bouaké	242	224.0	219.0	194.0	208.0	145.0	104.0	82.0	115.0	170.0	191	198.0	2,092.0	[2]
Ivory Coast	Abidjan	223	223.0	239.0	214.0	205.0	128.0	137.0	125.0	139.0	215.0	224	224.0	2,296.0	[2]
Ivory Coast	Odienné	242	220.2	217.3	214.7	248.8	221.8	183.5	174.5	185.4	235.8	252	242.6	2,638.6	[3]
Ivory Coast	Ferké	279	249.0	253.0	229.0	251.0	221.0	183.0	151.0	173.0	245.0	261	262.0	2,757.0	[2]

This data set contains the sunshine hours for 391 cities from 141 countries. I checked for null values in the sunshine data set and discovered that the data set is free of null values. I then converted the yearly sunshine duration from characters to numeric data. I also checked for abnormal values. For instance, I checked the locations of maximum and minimum yearly sunshine in the data set, which appear in Yuma, United States, and Tórshavn, Faroe Islands, respectively. I validated that both of these observations are accurate. To prepare for merging, since sunshine hours are reported by cities, while our happiness data set is reported in countries, I created a new data frame that records mean yearly sunshine, maximum yearly sunshine, minimum yearly sunshine for each country.

Data 2: Happiness Data

Below is a sample table of the data set.

Table 2: Raw Data from World Happiness Report

Country.name	year	Life.Ladder	Log.GDP.per.capita	Social.support	Healthy.life.expectancy.at.birth	Freedom.to.make.life.choices	Generosity	Perceptions.of.corruption	Positive.affect	Negative.affect
Afghanistan	2008	3.724	7.370	0.451	50.80	0.718	0.168	0.882	0.518	0.258
Afghanistan	2009	4.402	7.540	0.552	51.20	0.679	0.190	0.850	0.584	0.237
Afghanistan	2010	4.758	7.647	0.539	51.60	0.600	0.121	0.707	0.618	0.275
Afghanistan	2011	3.832	7.620	0.521	51.92	0.496	0.162	0.731	0.611	0.267
Afghanistan	2012	3.783	7.705	0.521	52.24	0.531	0.236	0.776	0.710	0.268

There are 36 NA values in log GDP per capita and 32 NA values in the "Freedom.to.make.life.choices" variable, which I removed. 68 observations were removed in total and 1881 observations remaining. The happiness dataset is very well compiled; the data set is free from abnormal values. As an extra validation, I checked whether freedom scores are between 0 and 1 and whether the GDP is between an acceptable range. I also checked whether there were distinct or duplicated happiness scores for the same country in the same year.

To prepare for merging the happiness data set and sunshine data set, I modified a few country names within the sunshine data set to match that of the happiness data. This is because I recognized that some countries are named differently. For instance, 'Taiwan' in the sunshine data is named "Taiwan Province of China" in the happiness data. A total of 5 names were changed.

Data 3: World Cities Data

Below is a sample table of the data set.

Table 3: Raw Data from World Cities Dataset

city	city_ascii	lat	lng	country	iso2	iso3	admin_name	capital	population	id
Tokyo	Tokyo	35.6897	139.6922	Japan	JP	JPN	Tōkyō	primary	37732000	1392685764
Jakarta	Jakarta	-6.1750	106.8275	Indonesia	ID	IDN	Jakarta	primary	33756000	1360771077
Delhi	Delhi	28.6100	77.2300	India	IN	IND	Delhi	admin	32226000	1356872604
Guangzhou	Guangzhou	23.1300	113.2600	China	CN	CHN	Guangdong	admin	26940000	1156237133
Mumbai	Mumbai	19.0761	72.8775	India	IN	IND	Mahārāshtra	admin	24973000	1356226629

There are a total of 44691 observations in this data set, and 307 of which have empty entries for the population column. However, after a thorough check, none of these empty population entries will affect our merging with the sunshine hours data, as these cities are either in countries that are not present in the happiness data set, or these cities are in countries that have other cities with non-empty population, which can be used to calculate for maximum population.

Data Merging/Filtering - Sunshine Hours Data & World Cities Data

As I have mentioned previously, the sunshine data set only contains sunshine hours for cities specifically; however, the happiness data is only recorded by country. Therefore, we need to obtain the country sunshine hours. There are many ways to do this, such as averaging the sunshine hours over all cities within the country or choose a deciding factor for which cities are a better representation for the country. I chose "population" to be the deciding factor, since in my opinion, higher population means that more people in this country are receiving this amount of sunshine hours.

Prior to merging the sunshine data set with the world cities population data set, one major modification required within the world cities data set are to match the city names. Some of them do not match that of the sunshine data set. Some changes I made include changing "Viljandi" to "Vilsandi" in the sunshine, "Luxembourg" to "Luxembourg City", and "New York" to "New York city" in the world cities data. Other name changes were also required when there were special language characters in city names, such as "Montréal" and "Kolkāta". Moreover, I changed a couple of country names. For example, I imputed the country for "Macau" which was empty previously, and changed "Saint Pierreand Miquelon" in the sunshine data set to be free of typos. One special case is "Tel Aviv" which was recorded as a city in the sunshine but a capital in the world cities data set. Therefore, to calculate the population, I summed up all population in the Tel Aviv first. A total of 30 cities were modified to prepare for merging.

After merging, I realized that there are still 28 cities that do not have population records. For cities whose country is not present in the happiness data set or cities which are in countries that already have cities with population records, it will not be a problem to obtain the max populated cities.

There were only 2 countries that require imputation for population, "Ivory Coast" and "North Macedonia", which are both not in the world cities data set. Therefore, I chose Abidjan and Skopje as the cities to represent sunshine hours after learning they are the capitals in "Ivory Coast" and "North Macedonia" respectively. Afterwards, I only kept the the highest populated cities as a representation for sunshine hours for all the countries.

To see a visualization of the selection of cities as representation for sunshine hours and the sunshine hours data set in each country geographically, please refer to Interactive Visualizations: Figure 1. /

Table 4: Country Sunshine Data by Highest Populated cities

Country.name	City	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec	Avg_sunshine
Afghanistan	Kabul	177.2	178.6	204.5	232.5	310.3	353.4	356.8	339.7	303.9	282.6	253.2	182.4	3175.1
Albania	Tirana	124.0	125.0	165.0	191.0	263.0	298.0	354.0	327.0	264.0	218.0	127.0	88.0	2544.0
Algeria	Algiers	149.0	165.0	202.0	258.0	319.0	318.0	350.0	319.0	237.0	229.0	165.0	136.0	2847.0
Angola	Luanda	219.0	208.0	213.0	199.0	233.0	223.0	175.0	150.0	145.0	164.0	199.0	212.0	2341.0
Argentina	Buenos Aires	279.0	240.8	229.0	220.0	173.6	132.0	142.6	173.6	189.0	227.0	252.0	266.6	2525.2

Data Merging - Happiness Data & Sunshine Data

After merging the data set, I discovered that there are still around 30 countries out of 161 countries which do not have sunshine hours reported. After carefully examining these countries, I recognized that they are less renown countries in the world; therefore, I decided to safely remove them from the merged data set. I also removed the columns that are not my variables of my interest, such as generosity, social support, positive effect ... etc. I renamed the columns to a more accessible format as well. My final merged data set consists of 1553 observations, and below is a sample table.

Table 5: Merged Data from World Happiness Report and Sunshine Hours Data

Country	Year	Happiness	Log_GDP	Freedom	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec	Avg_sunshine
Afghanistan	2018	2.694	7.692	0.374	177.2	178.6	204.5	232.5	310.3	353.4	356.8	339.7	303.9	282.6	253.2	182.4	3175.1
Afghanistan	2008	3.724	7.370	0.718	177.2	178.6	204.5	232.5	310.3	353.4	356.8	339.7	303.9	282.6	253.2	182.4	3175.1
Afghanistan	2012	3.783	7.705	0.531	177.2	178.6	204.5	232.5	310.3	353.4	356.8	339.7	303.9	282.6	253.2	182.4	3175.1
Afghanistan	2011	3.832	7.620	0.496	177.2	178.6	204.5	232.5	310.3	353.4	356.8	339.7	303.9	282.6	253.2	182.4	3175.1
Afghanistan	2017	2.662	7.697	0.427	177.2	178.6	204.5	232.5	310.3	353.4	356.8	339.7	303.9	282.6	253.2	182.4	3175.1

Data Exploration

I used various R functions, such as summary() and str(), to explore the variables within my data set. Below is a summary of all the variables in the data set.

Table 6: Table 4: Summary Statistics for Variables in Merged Data Set

Variable Names	Type	Q1	Median	Q3	Mean	Maximum	Minimum
Country	chr						
Year	int	2010	2013	2017	2013	2020	2005
Happiness	num	4.72	5.491	6.414	5.558	8.019	2.375
Log_GDP	num	8.594	9.6	10.473	9.479	11.648	6.635
Freedom	num	0.644	0.757	0.85	0.74	0.985	0.258
Avg_sunshine	num	1893.5	2333.9	2808.4	2361.018	3737.1	1230
Max_sunshine	num	1893.5	2333.9	2808.4	2361.018	3737.1	1230
Min_sunshine	num	1893.5	2333.9	2808.4	2361.018	3737.1	1230

I checked the distributions of each variable by plotting the respective histograms. I recognized that freedom is left-skewed, which means that the data is a better representation for countries with more freedom. Happiness score is normally distributed, which a good indication that one of the assumptions for linear regression is satisfied. As for average sunshine hours, it holds a bimodal distribution with the two peaks at around 1800 and 3000 hours. There are two explanations for this: countries with 1800 and 3000 yearly sunshine duration participate in the happiness report for a lot of the years, or there are more countries with 1800 and 3000 yearly sunshine hours.

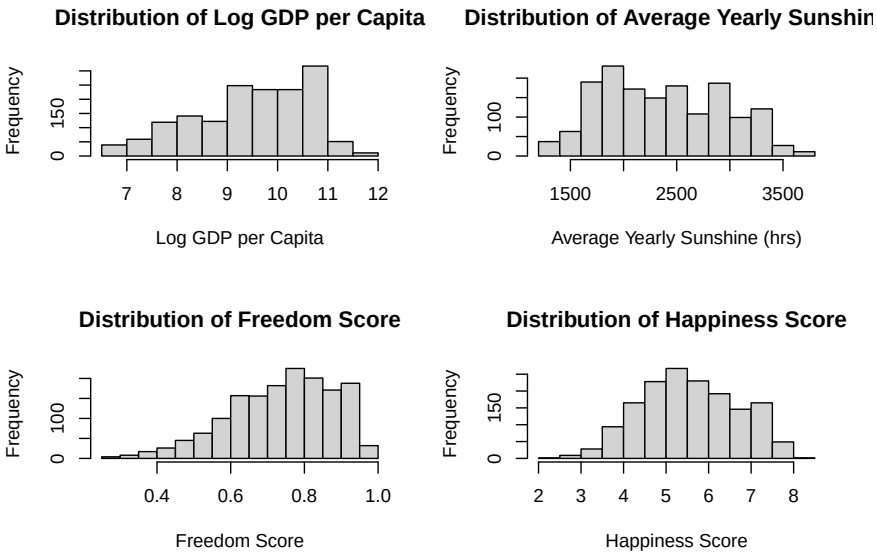

Figure 1: Distribution of interest variables, Log GDP, Sunshine Hours, Freedom, and Happiness

I also used the pairs plot function in R to explore the relationship between each of the variables and detect for possible linear relationships. As a result, I recognized that there are potential linear relationships between the variables Happiness and log GDP in particular.

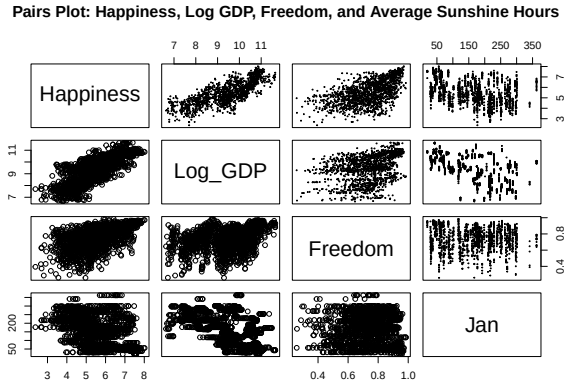

Figure 2: This is a pairs plot for variables of interest.

In addition, I realized that not every country was recorded the same number of times in the data set. This may cause a biased resultof the mean GDP or happiness score for each year, and this is something to keep in mind, which I will address later.

Table 7: Number of Countries Participated in Happiness Report Each Year

Year	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020
Countries Reported	25	69	84	90	92	101	117	116	111	113	114	116	120	114	114	79

Results

The results section is divided into 6 sections and each of them uses various techniques to answer different components of the research question. As a prelude, I will briefly introduce each of the sections. Section 1 looks at how mean happiness score changes overtime; section 2 focus on the distribution of GDP, freedom scores, and sunshine hours based on groupings determined by happiness score; section 3 shows the distribution of happiness score based on sunshine, freedom, and income levels; section 4 utilizes a linear model using happiness as a response variable; section 5 does cubic spline modeling based on a small modification on time; section 6 utilizes more complex modeling, including bagging, random forest, boosting, and XG boosting.

1. Mean Happiness Score Overtime

Please refer to interactive visualizations Figure 2 on the website for this component.

Explanation:

As mentioned previously, not every country is recorded every year between 2005 to 2020. Therefore, if we analyze the mean happiness score progression over time using all observations, there would be inaccurate results. For instance, the mean happiness score may increase because a country with high happiness score was added in the later years. Therefore, I chose to only report the mean happiness score for countries that participated in the Happiness report every year between 2010 to 2020. Furthermore, I separated the trend for mean happiness score into categories based on income levels (above or below average), freedom levels (above or below average), and sunshine hours (above or below average). An overall trend in mean happiness score for these countries without categorizing is also displayed.

As a result, we can inspect that there is no specific trend in the happiness score overtime.The overall trend is rather horizontal with small ups and downs; there is not much change over time. Also, for some categories, happiness score increases, while it decreases in other categories. For instance, countries with freedom scores above average, the mean happiness score decreases, while in the countries of sunshine hours below average, the happiness score decreases. Moreover, an interesting observation I found is that between 2011 - 2014, the mean happiness score in most categories decreases; however, for the category of freedom level above average, the mean score significantly increases.

Although these are simply trends observed in the countries that were reported every year, it is still a good indication of how happiness has progressed overtime. In fact, we can still gain brief insights regarding the effects of income, freedom, and sunshine overtime. As we can see, countries with income below the average have the lowest mean happiness score in every year, followed by countries with income below the average and sunshine duration above average. Meanwhile, countries with income above the average have the highest mean happiness score overtime, followed by freedom above averaged countries. This allows us to suggest that high income and freedom give rise to the happiness score.

2. Distribution of GDP, Freedom Score, and Sunshine Hours by Happiness Score

Table 8: Summary of GDP based on Happiness Frequency Grouping

Happiness Frequency	Average Log GDP	Max Log GDP	Min Log GDP	# of Countries
1/3 to 2/3	9.799564	11.000	8.398	12
at least 2/3	10.328373	11.648	8.193	52
less than 1/3	8.621554	10.393	6.635	63

Table 9: Summary of Sunshine Hours based on Happiness Frequency Grouping

Happiness Frequency	Average Sunshine Hours	Max Sunshine Hours	Min Sunshine Hours
1/3 to 2/3	2229.037	3187.0	1754
at least 2/3	2168.550	3508.7	1230
less than 1/3	2568.144	3737.1	1618

Table 10: Summary of Freedom Score based on Happiness Frequency Grouping

Happiness Frequency	Average Freedom Score	Max Freedom Score	Min Freedom Score
1/3 to 2/3	0.7123397	0.954	0.369
at least 2/3	0.8205863	0.985	0.458
less than 1/3	0.6698748	0.952	0.258

Explanation:

This is a summary table done by categorizing countries by their levels of happiness. I define three categories for the countries based on the frequency of their happiness scores being greater or equal to the mean happiness score of the corresponding year. There are three groups in total: less than $\frac{1}{3}$, $\frac{1}{3}$ to $\frac{2}{3}$, and at least $\frac{2}{3}$. The categories can be interpreted as follows: if a country appears in the less than $\frac{1}{3}$ category, this means that out of all the years the country participated in the World Happiness Report, less than $\frac{1}{3}$ of their records yielded a happiness scores higher or equal to the mean happiness of that same year.

As a result of the summary table, the category of "at least $\frac{2}{3}$" yields the highest average log GDP and average freedom score, whereas the highest average sunshine appears in the "less than $\frac{1}{3}$" category. Also, the maximum log GDP occurs in countries with at least $\frac{2}{3}$ being happy and the minimum occurs in the "less than $\frac{1}{3}$" category. This suggests that countries with higher GDP have higher happiness. A similar result applies on the factor freedom, which suggests that countries with higher freedom lead to more happiness. On the contrary, the category of "less than $\frac{1}{3}$" yields the maximum sunshine hours, while the minimum occurs in the "at least $\frac{2}{3}$" category. This allows us to say that sunshine hours doesn't increase happiness, but perhaps decreases.

Here, however, we can see there is an uneven distribution of categories. The $\frac{1}{3}$ to $\frac{2}{3}$ category only has 11 countries, whereas the other two groups have over 50. We may need to consider potential bias in further analysis using this table due to the under population of the $\frac{1}{3}$ to $\frac{2}{3}$ category.

As a note, I used the maximum of maximum yearly sunshine and minimum of minimum yearly sunshine to record the maximum and minimum sunshine hours for each category, rather than the maximum and minimum values of average yearly sunshine.

3. Distribution of Hppiness Score by Sunshine, Freedom, and Income Levels

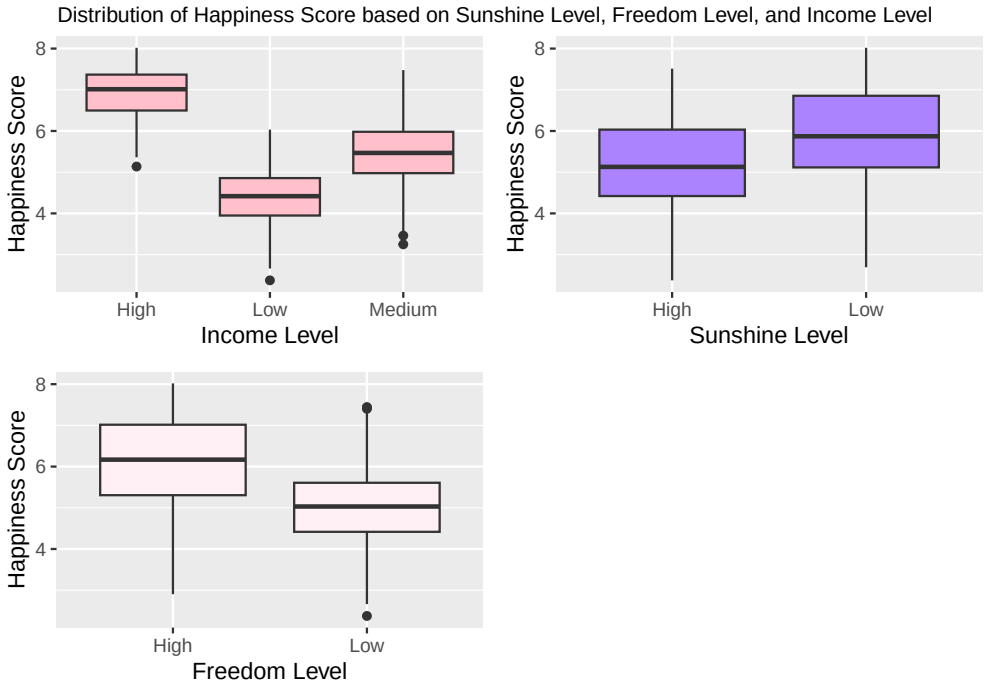

Figure 3: Distribution of happiness score based on sunshine Level, freedom level, and income Level (individually)

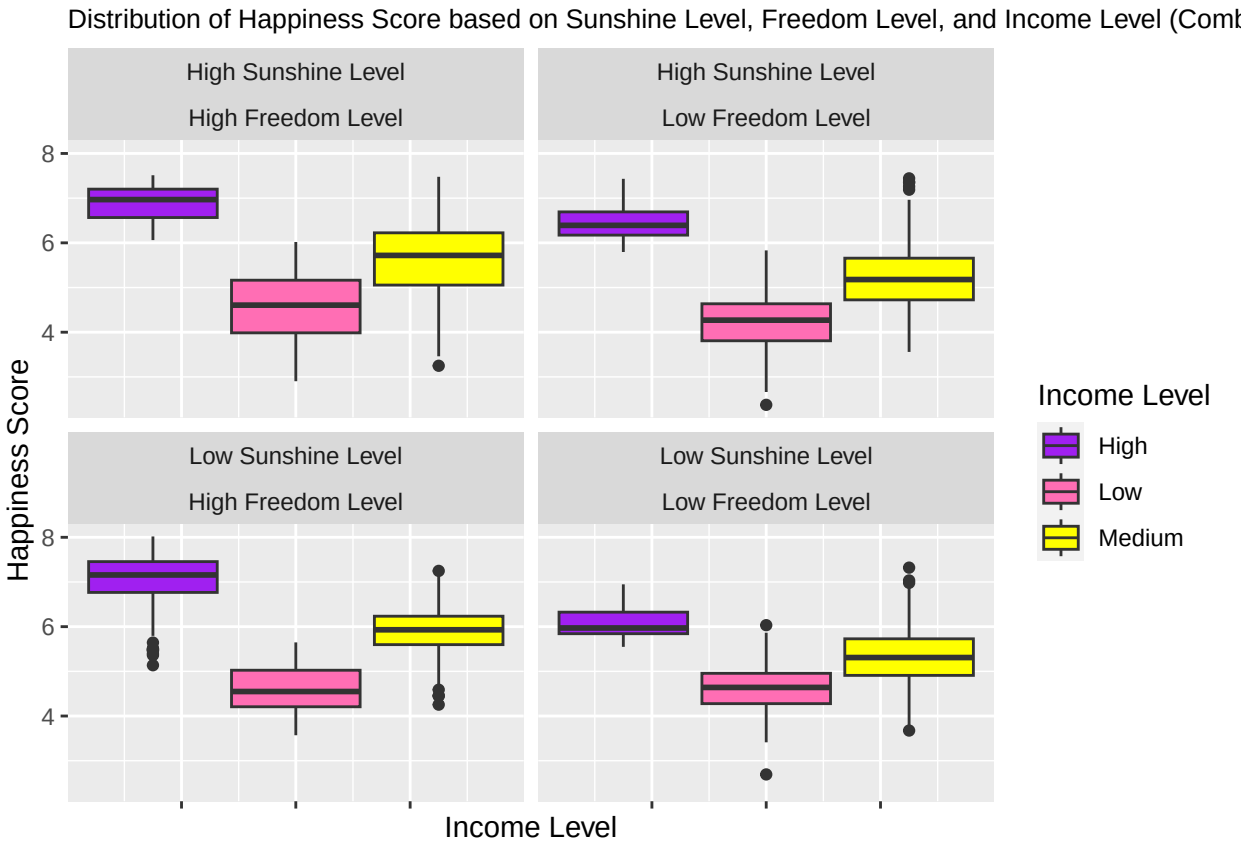

Figure 4: Distribution of happiness score with various combinations based on sunshine Level, freedom Level, and income Level

Note: For interactive visualizations of Figure 1 and 2, please refer to interactive visualizations Figure 3 on the website.

Explanation:

Income is separated into three levels: high, medium, and low. A "low" income indicates the log GDP per capita is less than or equal to the first quantile of log GDP per capita, whereas "medium" indicates that it lies between the first and third quantile, and "high" when it is above the third quantile. Freedom score and sunshine hours are classified into two groups by comparing with the mean values: high and low.

By inspecting figure 3, a high income level generally generates a higher average happiness score, and similarly for high freedom levels. As for sunshine levels, low sunshine levels typically yield a higher happiness score. The low freedom level group possesses a smaller interquartile range than that of the high freedom level group, meaning that its happiness scores has the least variations. Meanwhile, the variation between the high, medium, and low income group and the variation between the high and low sunshine level group is rather similiar. Moreover, there a few outliers with low happiness score in the high, medium, and low income, which suggests that there are still countries in each of these groups that possess a lower happiness score than the average.

Through figure 4, we can also infer about the greatness of effects between whether the decrease in happiness that sunshine causes is larger than that of the increase income and freedom causes. We can suggest that perhaps income has a larger effect than sunshine on happiness. This is because the group of low sunshine, low income yields a lower mean happiness score than the group with high sunshine and high income, holding freedom level fixed. As for comparing the effects of freedom level and sunshine level, the result is uncertain. We can see that the group of high income, high sunshine, high freedom yields a higher mean happiness score

than the group of high income, low sunshine, low freedom. However, when income is medium leveled, the results do not hold; the high sunshine, high freedom yields a lower mean happiness than that of low sunshine, low freedom group. Obviously, these results may still be biased if some subgroups are under populated.

4. Linear Model

Table 11: Linear Model Coefficients with Significance

Terms	Coefficient Estimate	Significance
Intercept	-2.1120000	Yes
Log GDP per capita	0.6432000	Yes
Freedom Score	2.4090000	Yes
Average Sunshine	-0.0000886	Yes

Please also refer to interactive visualizations Figure 4 on the website for this component.

Explanation:

In this part, I performed an analysis between happiness scores and the following predictors, freedom score, average sunshine hours, and log GDP per capita. In Figure 4, we can see that as freedom score and log GDP Per Capita increases, the happiness score increases. There is a positive association. On the other hand, there is a negative association between happiness score and average sunshine hours; as sunshine hours increase, the happiness score decreases.

Trends in the plot is also reflected in the linear model after fitting. All the predictors are deemed to be significant, which means we can interpret their impacts on happiness score using the coefficients produced. For the terms, freedom score and log GDP per capita, the coefficients are positive; as freedom score and log GDP per capita increases, happiness increases as well. As a note, a small change in freedom score causes a larger increase in happiness score than that of log GDP per capita, which I suppose is due to the small range of freedom score, 0 to 1. On the other hand, sunshine hours lower the happiness score, but the coefficient is so small that there is nearly no effect.

An adjusted R-squared of 0.7102 has also been reported. This means that 70% of the variability in happiness is being explained by our model, which means that our model is not a poor fit. After attempting to remove the predictor, Average sunshine, the adjusted R-squared only decreases by 0.0016, which is extremely low, and this suggests that sunshine hours does not directly impact happiness.

Previously, when I used the average sunshine hours over cities to represent a country's sunshine hours, sunshine hours were deemed to be insignificant and the coefficients for other terms remain extremely similar.

5. Cubic spline Model

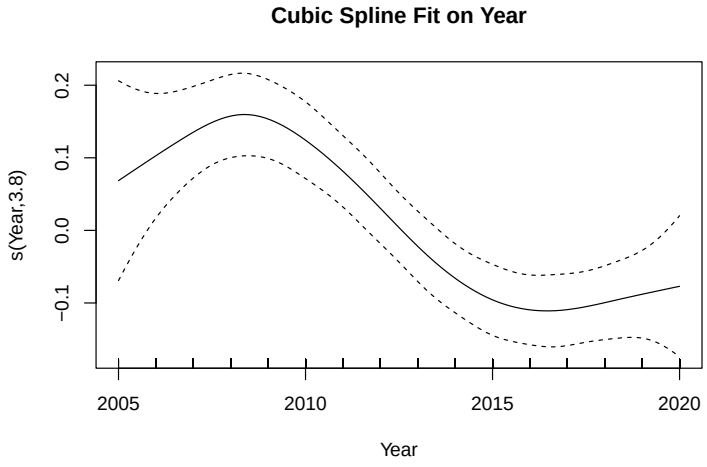

Figure 5: Cubic spline model fit on variable, year

Table 12: Cubic Spline (on Year) Model Summary

Terms	Coefficient Estimate	Significance
Intercept	-2.2469000	Yes
Log GDP per capita	0.6577000	Yes
Freedom Score	2.5760000	Yes
Average Sunshine	-0.0000468	No

Explanation:

I fit a cubic spline regression model having 10 knots on the variable, Year, along with other variables including economic status, freedom, and sunshine hours. From figure 3, we can see that there is a slight increase in happiness in between years 2005 -2008, and from 2008 onward, the happiness decreases. Until 2017, the happiness slightly rises again. Also, it is worthy to note that the trend between 2010-2020 shares a few similarities from the plot in part 1, such as the slight increase from 2017 - 2020 and the decrease from 2013 - 2015.

However, the adjusted R-squared for this model is 0.72, which isn't a very large difference from the linear model fitted in part 4. This means that happiness is not explained by time as much. In fact, the coefficients of the other variables remain extremely similar as well. Therefore, through this model, we can say that time doesn't play a big role in affecting happiness.

6. More Complex Modelling

I performed a 70/30 train test split on the data set, and built the models using the four variables of interest, `Log_GDP`, `Avg_sunshine`, `Year`, and `Freedom`, based on a selection of hyperparameters as specified below. The models include bagging, random forest, boosting, and XG boosting.

Bagging:

No specific hyperparameters were specified for Bagging, as this model seeks out different value combinations already.

Random Forest:

No specific hyperparameters were specified for Random Forest, as this model seeks out different value combinations already.

Boosting:

I set the number of trees to be 1000 and learning rate at 0.3. I tested different learning rates between 0.01 and 0.3 and recognized that 0.3 yields the smallest MSE out of these values.

XG Boosting:

To select the hyperparameters, I modeled a max depth of 3, 5, 7, 9. For each of these depths, I first ran an epoch of 1000 rounds at a learning rate of 0.2, and to select the final number of epochs, I chose the iteration where MSE stops decreasing. This is because when the MSE starts to increase slightly, this indicates over fitting of the data. After, I compared the mse and adjusted R-squared of these 4 models, and chose to use the model with a max depth of 9, nrounds of 53, and learning rate of 0.2.

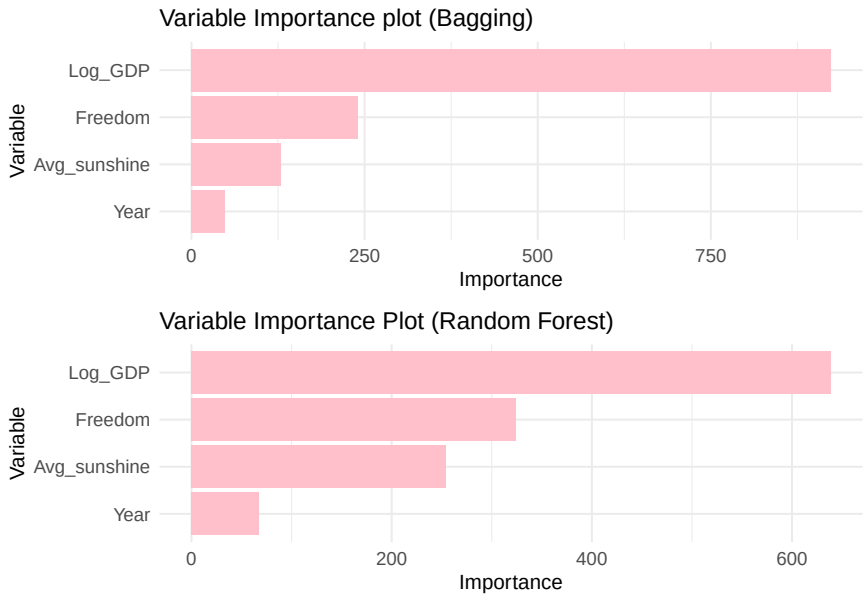

Figure 6: Variable importance plot for models, Bagging and Random Forest

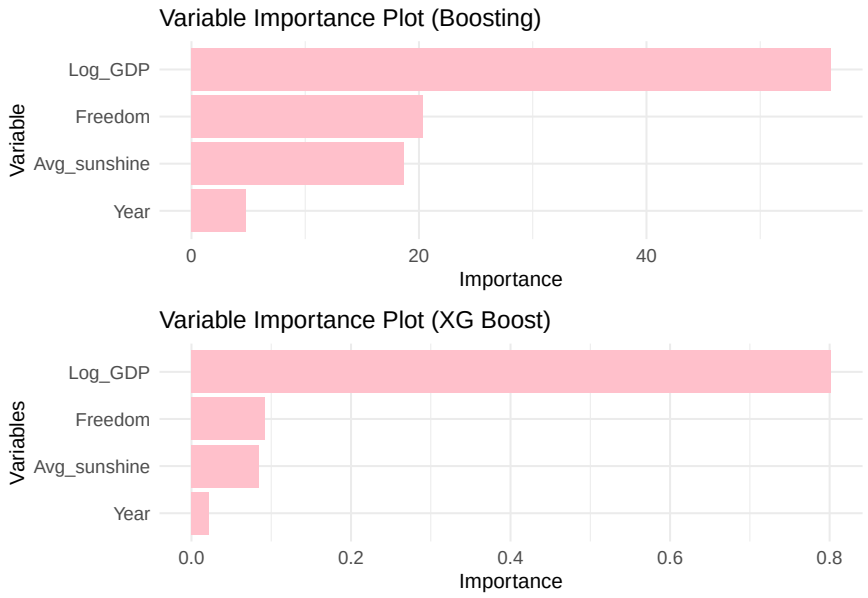

Figure 7: Variable Importance Plot for models, Xg Boosting and Boosting

Test Accuracy:

To obtain the accuracy of our models, I fitted the model on the test set and computed computed the Mean Squared Error and the R-squared produced by the four models. I did not compute the adjusted R-squared for these models because after some careful research, I recognized that adjusted R-squared is not a suitable metric for nonlinear models.

Table 13: Raw Data from Wikipedia Sunshine Hours

Method	MSE	R-squared
Bagging	0.1954331	0.853
Random Forest	0.2769128	0.792
Boosting	0.2381254	0.821
XG Boosting	0.1750634	0.869

Interpretations:

From all the models, the variable of importance is in the following order: `Log_GDP`, `Freedom`, `Avg_sunshine`, and `Year`. Out of these factors, the most important factor for a model to make accurate prediction is GDP. The models rely on it alot to make predictions. This could be an indication that it influences our response variable the most. A citizen's income can highly determine their happinesss, followed by freedom, sunshine hours received, and time period.

We can see that the XG Boosting model yields the lowest MSE, while it gives the highest R-squared. This suggests that it is the best model out of the four models. In fact, all of these models perform better than that of the linear model and spline basis fitted in part 5, as they are more complex. This is not surprising since XG Boosting is a higher level implementation of gradient boosting that uses advanced regularization (L1 & L2), and this highly improves its model generalization abilities. With an R-squared of 0.868, approximately 86% of the variability in the Happiness score is explained by the predictors within the XG boosting model. On the other hand, Random Forest yields the highest MSE and lowest R-squared, my best guess for this result is that random forest only chooses a subset of predictors on a split. Therefore, with only 4 predictors in our case, it might be lose accuracy when even less predictors are being used.

Conclusion

In this report, our primary questions are to see whether happiness increase over time and whether a citizen's happiness is impacted by their economic status, freedom, and sunshine hours received.

Based on the results of 1 and 5, we can see that there is no particular trend in the increase of happiness overtime, and it is rather similar over the years. However, there is a particular drop between 2011 - 2014 in the mean happiness score, which is worth noting. As of the factors that influence happiness, both economic status and freedom increase happiness. On the other hand, sunshine hours decreases happiness score; however, its decrease in happiness score is so minimal that it can be neglected. In terms of how much these factors impact happiness, economic status, followed by freedom, are what causes a more significance influence, compared to that of sunshine hours and time period.

Therefore, the main takeaway from this analysis is that higher economic status and more freedom given lead to happier individuals.

Limitations

One major limitation of this study is that the sunshine data set only contains the sunshine hours for particular countries. An amount of countries were removed from the study despite their happiness score was recorded. If more observations and countries are included, the effects of factors that influence happiness can be considered from more parts of the world, adding variability to our data set in terms of country backgrounds. Another limitation is that sunshine hours differs widely across countries if the country is very large. Although I used population as an indicator for the representation within the country, it may not be the best deciding factor. Therefore, it may not be the most accurate amount of sunshine a citizen receives but simply a "best guess". It is more dependent on the city of the country they are in. Also, although sunshine hours do not differ as much across different years, there are still differences where some years may have slightly higher or less sunshine duration. Therefore, using the same amount of sunshine hours for every year in the data set may still cause some biased results that we need to be careful of. To add on, as mentioned previously, freedom score is right skewed; therefore, this analysis could not be as good for applying on countries with lower freedom, since our data set contains more observations with higher freedom scores. As a last limitation, the countries participating in the world happiness report every year is not the same, and this restricts how we analyze the happiness change over time and how we reference "yearly mean happiness score".