Project Report

Vicente Santos

**Problem Statement:**

What is Happiness and how do we define it? That is a question humans as a whole have been attempting to answer for centuries. Happiness is a highly subjective experience, making being able to quantify the emotion tricky because it is something that everyone has different meanings for, However, applying a standardized definition and asking people worldwide to evaluate their life satisfaction yields fascinating patterns. This intrigue led me to the Global Happiness Index. The Global Happiness Index is based off averages obtained from survey data asking people general questions about their life and asking them to rank their life satisfaction from a scale of 1 to 10. This is then averaged by country and the data has been collected for 9 years. While this data offers extremely useful insights into global levels of happiness, it falls short of explaining the underlying reasons as to why, and I was extremely interested in learning more about the subject, but specifically what factors cause happiness.. One contentious yet popular idea I investigated is the claim: “Money doesn’t buy happiness.” This notion is frequently debated, with strong arguments on both sides. My goal was to investigate this claim and the inverse claim and identify additional variables influencing happiness, as well as their significance. This is a topic that has largely interested me because in Data Science, we think of quantifiable data as variables that exist in the world around us, not necessarily how we as humans are. I also have a passion for psychology and I have lived a large part of my life believing in the good in humanity, and trying my best to go through the world as a person who is satisfied with their life.

## Introduction to Data:

The Happiness Index Data was publicly available on their website as downloadable CSV files, and the independent variables I chose from the IMF and World Bank Data Centers . My independent variables consisted of 9 variables that I chose. These variables where chosen because I wanted to maintain a focus on variables that described a society as a whole. These variables were GDP per capita, Unemployment, Poverty, Climate, Continent (geographic), Inequality, Rule of Law, Civil Liberties, and Population. Each independent variable was publicly available in CSV files from either the World Bank or the International Monetary Fund. The happiness indexes were avaialble by year, showing a country name and their overall happiness index score from 1-10. When considering ethical implications with this topic, it is important to maintain clarity and to stay cognizant of the fact that life satisfaction has different meanings to everyone. It is vital that as observers, this topic is approached with no bias, but more fundamentally, the understanding that people around the world are different. We as humans have different values, stories and life goals, so life satisfaction naturally will be a flexible definition. On that hand, it is my personal opinion that grouping happiness by nations is an excellent way to observe this data and yields the most consistent results with reality as in nations are typically defined by shared sets of cultural values, so with that established as a variable that cannot be quantified, my focus will be shifted towards observing economic and social factors.

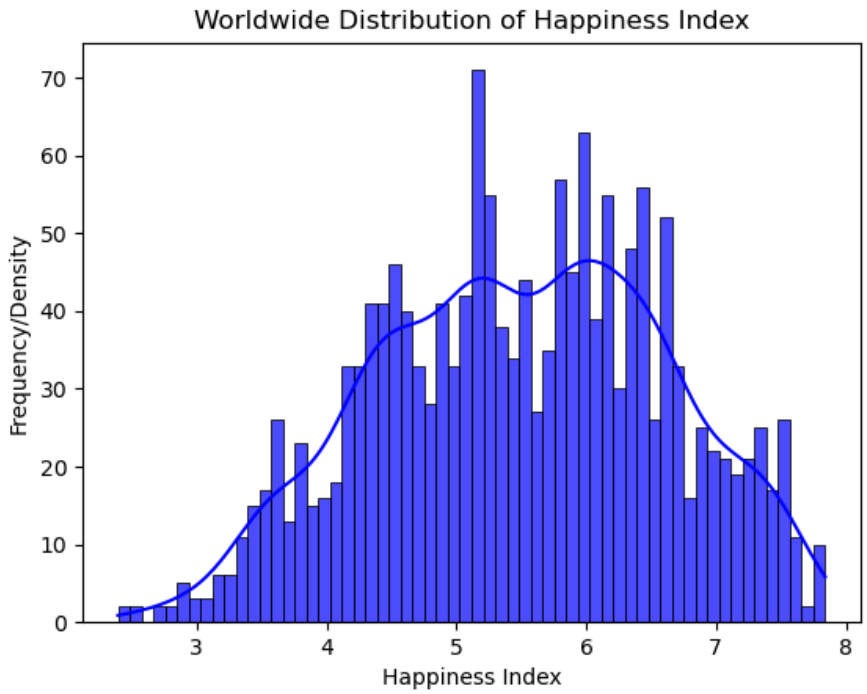
## Data Science Approaches:

The methods I had used in this project were heavily reliant on the Pandas library. I dove deep into the syntax and ended up discovering syntax that made things more efficient for me. My most “game-changing” use was the (errors=’coerce’) which was an operator that allowed any data that could not be converted to the desired type to be dropped. For my exploration I used bar

graphs and correlation matrices,  and for my analysis, I employed a mix of linear regressions, multivariable linear regressions and Kmeans clustering, I used that to determine three separate things. Which variables have the most affect on Happiness, a predictive model for the Happiness Index, and also creating clusters for the most powerful variable, GDP per capita.

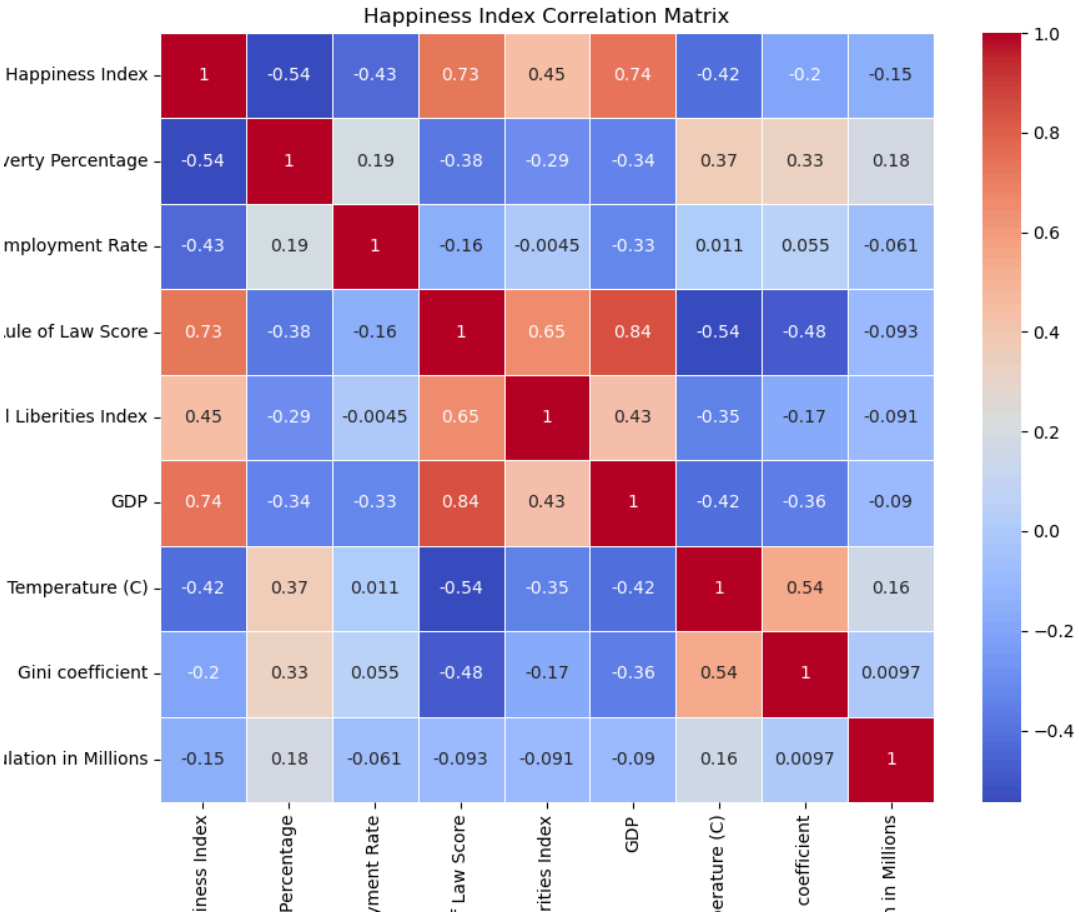
**Analysis/Results:**

My analysis started with compiling and merging all years for the Happines Index. After that, I standardized the DataFrames for my independent variables and then merged all my data into a single, uniformed DataFrame. After that, to explore the data further, I inspected it by first seeing the distribution spread for the Happiness Index. The figure I generated can be seen below.



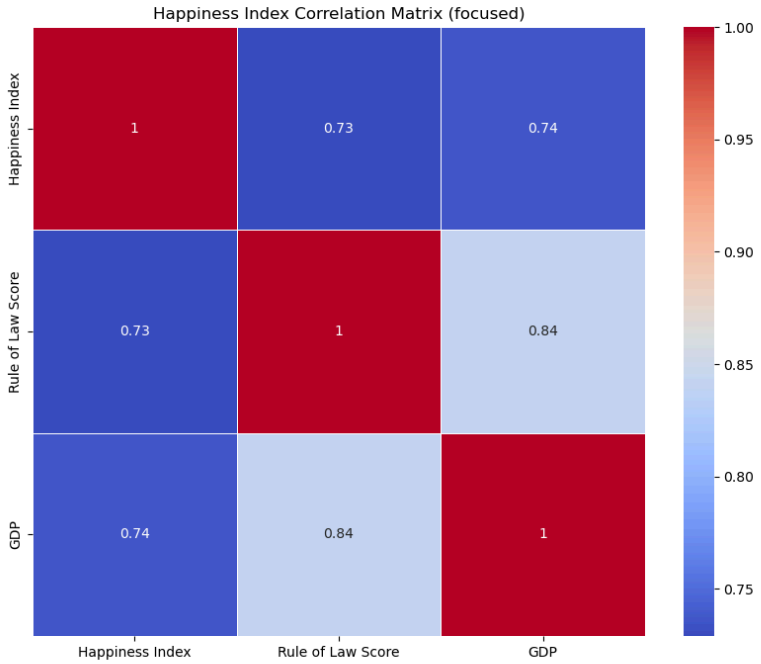

Here, there is one, big noticeable trend. No one country is above 8 on the happiness index. When finding the lowest score the one that came up was Afghanistan 2021 with a tragically low score of 1.721.  The highest score was Finland 2022 with  7.84. This shows an interesting trend of countries that do not vary that much, and with zscores under 3, they will be included in our data

and not considered outliers. An average score of 7.84 shows still signs of unhappiness, which is further exemplified in our observed average from the graph which is between 5 and 6. So the first conclusion I had arrived to was that on average, people are generally not that satisfied with their life, a sobering conclusion, but that only intrigued me more to find out the main factors as to why.  The best way to find out which factors are the most important and why. The easiest way to do this is with a correlation matrix with all 9 variables.




Upon closer examination of the correlation matrix, there are two values that stand out. GDP and Rule of Law, so to take a closer look, I focused my visualization and narrowed my  variables to

GDP and Rule of Law.



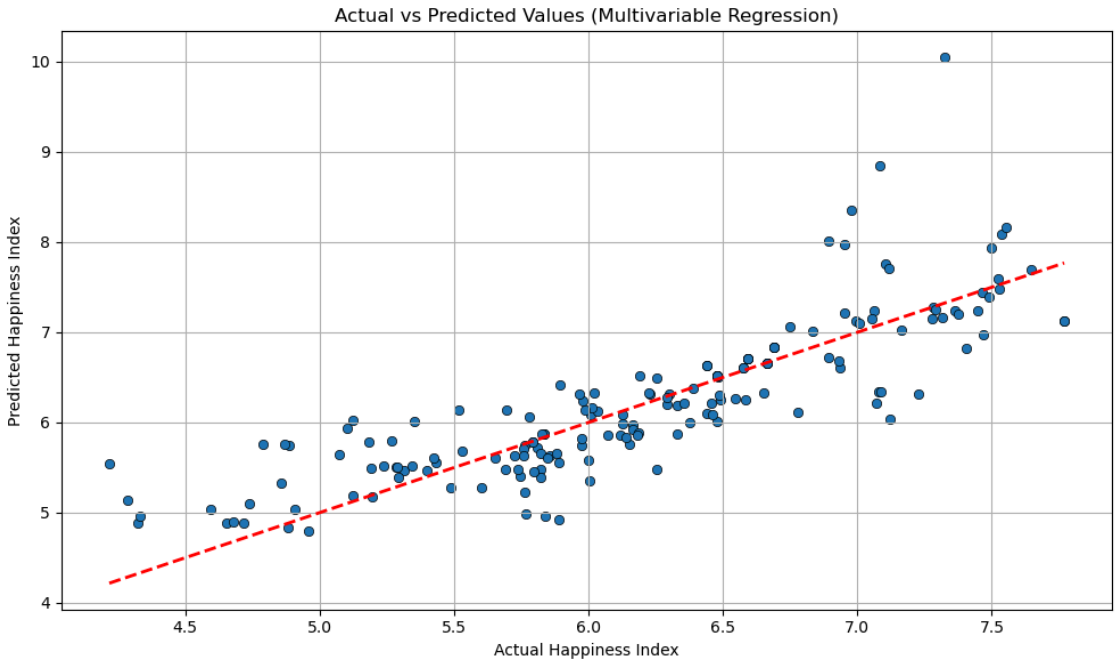

After these matrices, it is clear that by far the two most influential factors in being able to predict a happiness score is Rule of Law and GDP. With that being said, the other variables also show relative correlation as well, so it is important to not exclude them when doing a predictive analysis. To further examine these relationships I also plotted GDP related to Happiness Index, (not pictured due to space). This relationship is quite consistent showing an Interesting pattern, I later explored this pattern with a Kmeans Cluster. After that, I conducted a linear regression on every individual variable to observer their MSE and R-Squared scores to evaluate which ones to include in my Multivariable regression. This was

the results of that

| Feature | Mean Squared Error | R-Squared Score |
|---|---|---|
| Poverty Percentage | 0.708895 | 0.279156 |
| Unemployment Rate | 0.657077 | 0.127598 |
| Rule of Law Score | 0.513776 | 0.580445 |
| Civil Liberities Index | 0.963724 | 0.242382 |
| GDP | 0.546408 | 0.566906 |
| Average Temperature (C) | 0.998732 | 0.198987 |
| Gini coefficient | 0.885633 | 0.067095 |
| Population in Millions | 1.238877 | 0.018041 |

As an addition, I calculated the correlation with [‘Year’] and Happiness Index and that resulted with a correlation of 0.02, which I found extremely interesting because I had somewhat of a theory that the ‘Covid Years’ would have a large effect on global happiness, but interestingly enough, they did not. Evaluating the table above, I found that Gini Coefficient, GDP, Civil Liberties and Unemployment were my favorite metrics to use in my Multivariable regression.

After that, I did the regression with those 4 against Happiness and created a scatterplot with predictions.




Here are the stats for this model

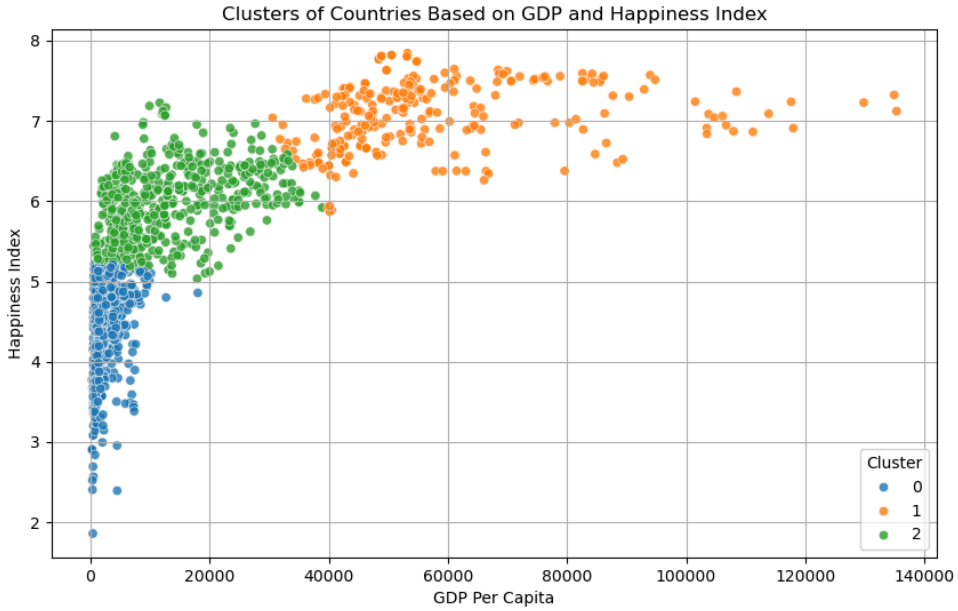
Model Evaluation:
Mean Squared Error: 0.23984281723658363 - Accurate Model
R-Squared Score: 0.6370089530539883 - Accounts for 63% of variance
Coefficients:

[-0.1328312 ] Unemployment Rate - Weak negative relationship
[ 0.35692984] Rule of Law Score - Moderate Positive relationship
[ -0.18381521] Civil liberties - Weak negative relationship
[0.64075841] GDP per capita Strong positive relationship
[ 0.31210922] Gini coefficient - Moderate positive relationship

Cross-Validated R2 Scores: [0.55322157 0.70052115 0.64727193 0.50825703 0.70784843]
Mean R2 Score: 0.6234240237476758

This model is showing good levels of accuracy, meaning that it is a good predictor of happiness levels in the low-mid range, however once we get to the upper echelon of scores, the model starts predicting with more and more variance. I found that to be very interesting, but overall the model is solid and shows good data. The conclusion i drew from this was mainly that GDP combined with the other predicted factors are the best ways of predicting happiness.

To explore this relationship with GDP further, I did a Kmeans clustering which yielded this plot




The most interesting Cluster is cluster #2, the green one. To sum it up, it can be classified as 'Low GDP, High Happiness" and this curve in general shows that at a certain happiness level between 5-6, Happiness starts to level off, but there exist anomalies. The most impressive one is the nation of Venezuela, which has a GDP< 4000 and a happiness index of above 6.7. This directly contradicts the notion that GDP is always related to Happiness, which in my opinion, is caused by unobserved factors such as cultural. 7/11 countries in cluster 2 were in Latin America, and as someone who grew up there, I can attest to the fact that general satisfaction is a widespread value there. The issue is that these factors are not really quantifiable. That exposed some other issues with this project, such that there are so many variables that can play into happiness that simply are not quantifiable. However, this is a good start, I think that models like these can be used to help guide policy, but also the Happiness Index serves as a decent metric for predicting what a country's economic conditions look like. However if a particular nation wants to raise their happiness scores, it is abundantly clear that their citizens need to be taken care of, after all it is hard to report that you are satisfied with your life whilst struggling to bring in a livable income.

CITATIONS:

Happinesss Data Index:

https://worldhappiness.report/

World Bank Data:

https://databank.worldbank.org/

IMF data:

https://www.imf.org/external/datamapper/datasets