FINAL PAPER: World Happiness

STOR 320.01 Group 25

May 05, 2021

INTRODUCTION

When looking at different countries around the world, many different and complex socioeconomic factors have contributed to varying levels of development. One would think that some of the most contributing factors to a country's well-being is GDP or its government. However, instead of measuring a country's development, can we instead measure its happiness? How can we measure something as subjective as happiness? The annual World Happiness Report seeks to answer this question, focusing on six different factors: logged GDP per capita, healthy life expectancy, perceptions of corruption, freedom to make life choices, social support, and generosity. This brings us to our first question of which of the aforementioned variables are the best predictors for a country's happiness score, or as the Gallup World Poll refers to it, the ladder score.

If we can find the answer to the above question it will enable countries that have a higher overall happiness score to know what to continue doing to maintain it, while countries with lower happiness scores can know what factors to focus on in the future to improve their overall happiness. This, in turn, can improve technology growth, educational advancements, quality of life, and most importantly happiness for their citizens. For example, governments can work towards creating laws and programs to increase those variables, or decide the order of importance for these variables so that they may efficiently delegate resources to invest in certain goals. The overall goal of a country should be to make its citizens as happy as possible. However, it's a very difficult decision to decide how to make this happen and where resources should be allocated. Finding out what are the most important predictors for happiness will aid in those decisions that affect thousands or millions of people. This is even more relevant given the COVID-19 pandemic, which has negatively affected happiness worldwide. Intuitively, many of these variables may correlate with each other, such as higher GDP associated with a higher life expectancy, but an analysis can thoroughly investigate and predict the effectiveness of these variables in regards to happiness.

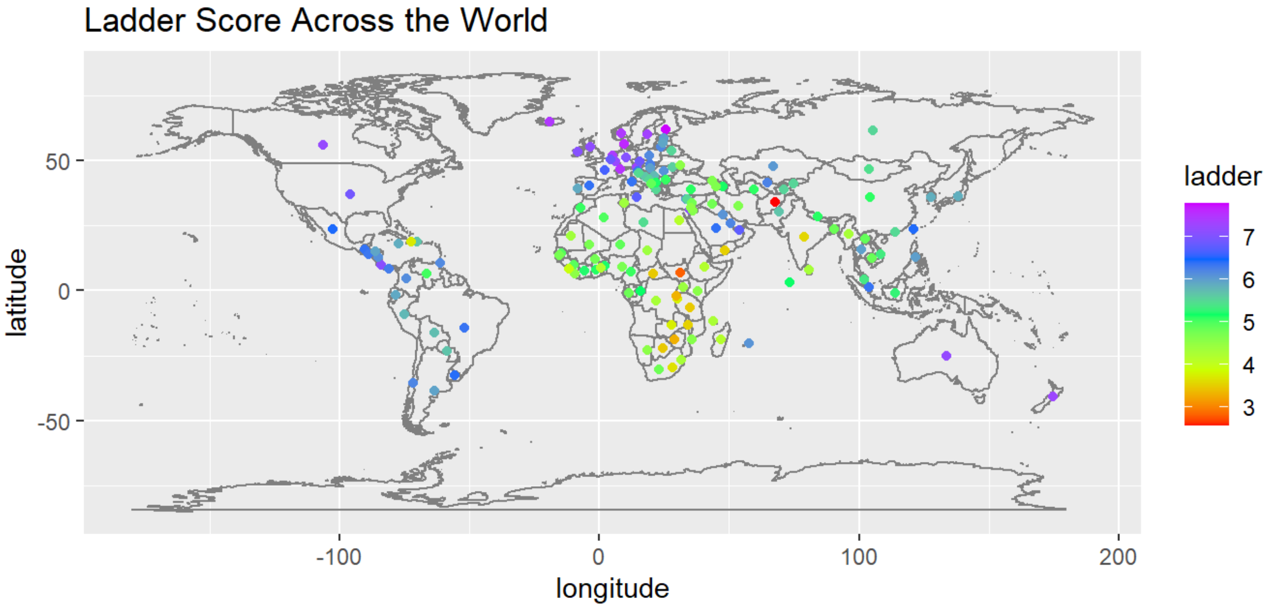
As we explored the dataset, we were interested in logged GDP per capita and its power to predict happiness. However, GDP per capita only gives an estimate about a country's overall economy and purchasing power. We wanted to scale down and look at median household income instead to answer an age-old question: can money buy happiness? More specifically, can money cause happiness? Having enough income to sustain basic survival needs is a given for improving happiness, but beyond that level we wanted to know, how does the level of household income affect the ladder score and can it predict logged GDP per capita, life expectancy, and/or social support? We also wanted to dive into how the median household income differs based on region. Investigating median household income allows for exploration between how much the wealth of its citizens has an impact on happiness and other socioeconomic factors. Knowing this, we may have another way to predict happiness, which may prove to be more accurate than the predictors from the World Happiness Report. By analyzing the median household income of a country we are aiming to explore other factors so that we can do our best to find out what makes the people of a country happier so that we can put more smiles on faces across the world.

DATA

Our data, titled, "World Happiness Report up to 2020" came from Kaggle, however, the original collectors of the data are a team of researchers, data scientists, and editors, with their base at the Sustainable Development Solutions Network (SDSN) and The Center for Sustainable Development at Columbia University, directed by Jeffrey D. Sachs. This team has also partnered up with Gallup, an American analytics and advisory company that is responsible for collecting all the worldwide survey responses for this dataset. The World Happiness Report for 2020 ranks 153 countries based on an average of three years of surveys between 2017 and 2019. The dataset contains data on 153 countries and their ladder score, logged GDP per capita, healthy life expectancy, perceptions of corruption, freedom to make life choices, social support, and generosity. The dataset also contained standard error and confidence intervals, but as they were measuring error of the ladder score, we did not use them in our analysis. For simplicity and ease of reading, ladder score was renamed to `ladder` , logged GDP per capita to `logged_GDP` , healthy life expectancy to 'life_expectancy', social support to `social_support` , freedom to make life choices to `freedom` , and perceptions of corruption to `corruption` . Below is a table of part of our data with the variables we used:

country	region	ladder	logged_GDP	social_support	life_expectancy	freedom	generosity	corruption
Finland	Western Europe	7.81	10.64	0.95	71.9	0.95	-0.06	0.20
Denmark	Western Europe	7.65	10.77	0.96	72.4	0.95	0.07	0.17
Switzerland	Western Europe	7.56	10.98	0.94	74.1	0.92	0.11	0.30
Iceland	Western Europe	7.50	10.77	0.97	73.0	0.95	0.25	0.71
Norway	Western Europe	7.49	11.09	0.95	73.2	0.96	0.13	0.26
Netherlands	Western Europe	7.45	10.81	0.94	72.3	0.91	0.21	0.36
Sweden	Western Europe	7.35	10.76	0.93	72.6	0.94	0.11	0.25
New Zealand	North America and ANZ	7.30	10.50	0.95	73.2	0.94	0.19	0.22
Austria	Western Europe	7.29	10.74	0.93	73.0	0.90	0.09	0.50
Luxembourg	Western Europe	7.24	11.45	0.91	72.6	0.91	0.00	0.37

The ladder score, or also known as the happiness score or subjective well-being, is the national average of all the Gallup survey responses to the question, "Please imagine a ladder, with steps numbered from 0 at the bottom to 10 at the top. The top of the ladder represents the best possible life for you and the bottom of the ladder represents the worst possible life for you. On which step of the ladder would you say you personally feel you stand at this time?" Logged GDP per capita is the statistics of GDP per capita in purchasing power parity (PPP) at constant 2011 international dollar prices, which is then logged. Social support is the national average of the binary responses (either 0 or 1, with 0 meaning "no" and 1 meaning "yes") to the Gallup World Poll question, "If you were in trouble, do you have relatives or friends you can count on to help you whenever you need them, or not?" Healthy life expectancy from birth is based on the data from the World Health Organization (WHO) data repository. Freedom to make life choices is defined as the national average of the binary responses to the Gallup World Poll question, "Are you satisfied or dissatisfied with your freedom to choose what you do with your life?" Generosity is the residual of regressing national average in response to the Gallup World Poll question, "Have you donated money to a charity in the past month?" on GDP per capita. Finally, perceptions of corruption is the national average of responses to the questions, "Is corruption widespread throughout the government or not" and "Is corruption widespread within businesses or not?" These survey responses were conducted through a sample of around 3,000 people for each country, and were then applied to the population. Below is a map displaying ladder score across the world.



The World Happiness dataset was also joined with a dataset that contained longitudes and latitudes for all the countries in the world in order to help visualize the pattern of happiness. The dataset, titled, "Latitude and Longitude for Every Country and State" is from Kaggle, although the original data was compiled by Google.

Additionally, to answer our second question, data about median household incomes for each country were scrapped from a website called World Population Review and joined with the World Happiness 2020 dataset. The dataset is titled "Median Income by Country 2021". The original data came from Gallup World Poll surveys. The median income is a number that comes from the middle of the nation's income distribution. In other words, half of the nation's adult residents have disposable income that is higher than the listed values, while the other half has disposable income that falls below that value. We only used two variables from this dataset, "medianHouseholdIncome" and "country". The data collected for the median household income was self-reported from a sample of about roughly 3,000 people and then applied to the population, much like the other Gallup World Poll survey question results. However, there were only 131 countries in this data, so when joined with the World Happiness 2020 dataset, some countries did not have a median household income value. To combat this we manually filled out those missing values by researching what the median household income of each missing country was.

country	medianHouseholdIncome
Luxembourg	52493
Norway	51489
Sweden	50514

country	medianHouseholdIncome
Australia	46555
Denmark	44360
United States	43585
Canada	41280
South Korea	40861
Kuwait	40854
Netherlands	38584

RESULTS

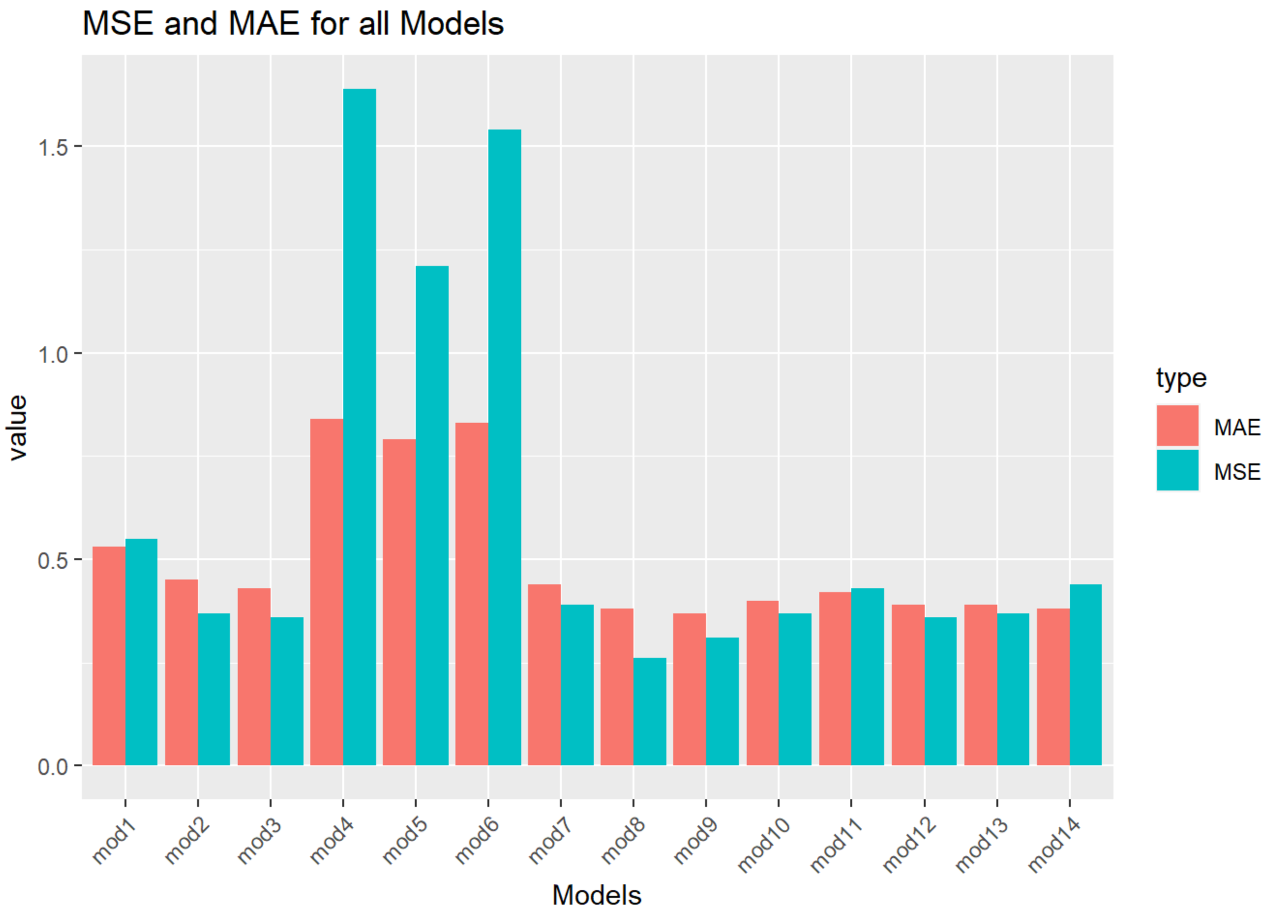
To answer our first question of “What are the most significant predictors of happiness?”, we first decided to form a total of 14 different models to help us understand which predictors are most significant in predicting happiness. The predictors we used in our models from the data were: “Healthy Life Expectancy”, “Logged GDP”, “Social Support”, “Generosity”, “Freedom” and “Perceptions of Corruption’’. Models 1 through 6 all have one predictor each. In these models there is a model for each of the 6 predictors listed above. We then made various combinations of these 6 predictors to form the rest of the 14 models.

We decided to include logged GDP in a good amount of the models because we felt logged GDP is usually a good indicator of a country’s economy and therefore thought it might be a good predictor. We also included healthy life expectancy in a lot of the models because usually healthy life expectancy is a good indication on how developed a country is. Therefore, we felt this could be a good predictor for happiness. Model 14 is interesting, however, in that we actually ran a step-wise regression with all 6 of the predictors predicting ladder score and the model that came out of this is our 14th model(mod14).

The models we chose were:

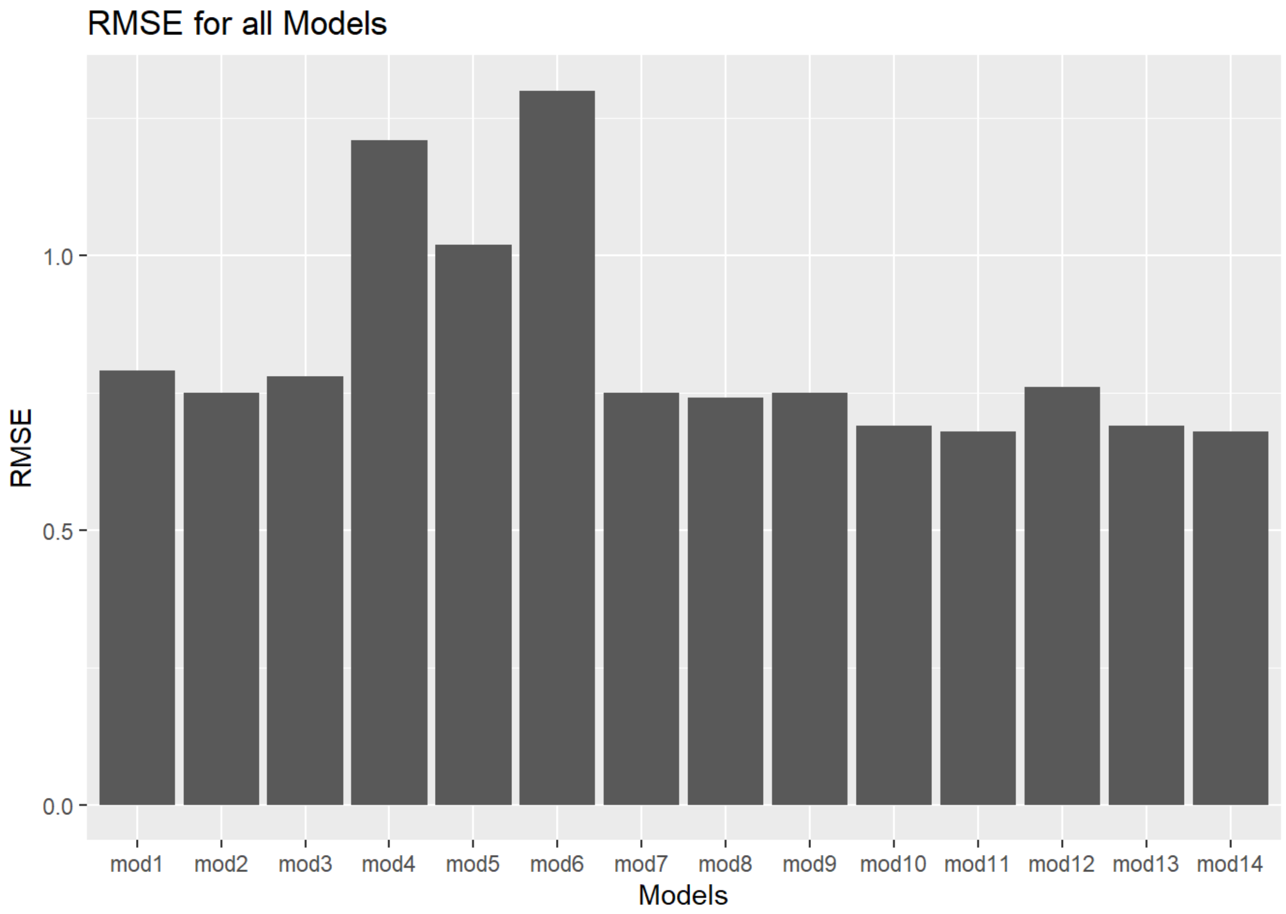
- mod1: ladder ~ Life Expectancy
- mod2: ladder ~ logged GDP
- mod3: ladder ~ Social Support
- mod4: ladder ~ Corruption
- mod5: ladder ~ Freedom
- mod6: ladder ~ Generosity
- mod7: ladder ~ logged GDP + Life Expectancy
- mod8: ladder ~ logged GDP + Social Support
- mod9: ladder ~ logged GDP + Life Expectancy, and Social Support
- mod10: ladder ~ logged GDP + Life Expectancy + Social Support + Freedom
- mod11: ladder ~ logged GDP + Life Expectancy + Freedom
- mod12: ladder ~ Life Expectancy + Social Support
- mod13: ladder ~ logged GDP + Life Expectancy + Social Support + Corruption
- mod14: ladder ~ Life Expectancy + Freedom + Social Support + Corruption

We started off by finding the mean standard error (MSE) and mean absolute error (MAE) for each model and compared the models to find which had the lowest MSE and which had the lowest MAE. In finding the MSE and MAE for each model we randomly split our data into 80% for training data and 20% for testing data. To display the results of our findings we made a clustered bar chart of the MSE and MAE values for each model.

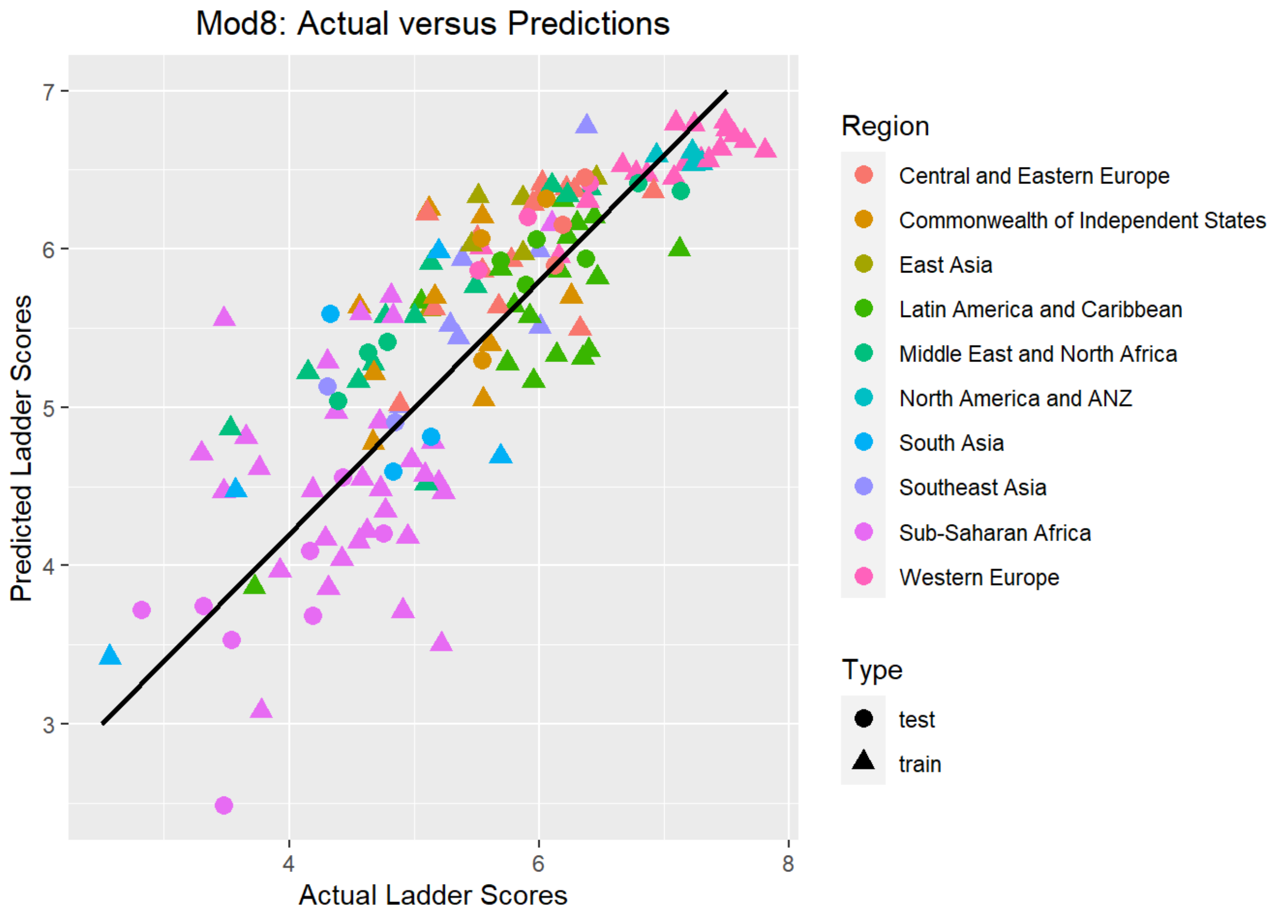


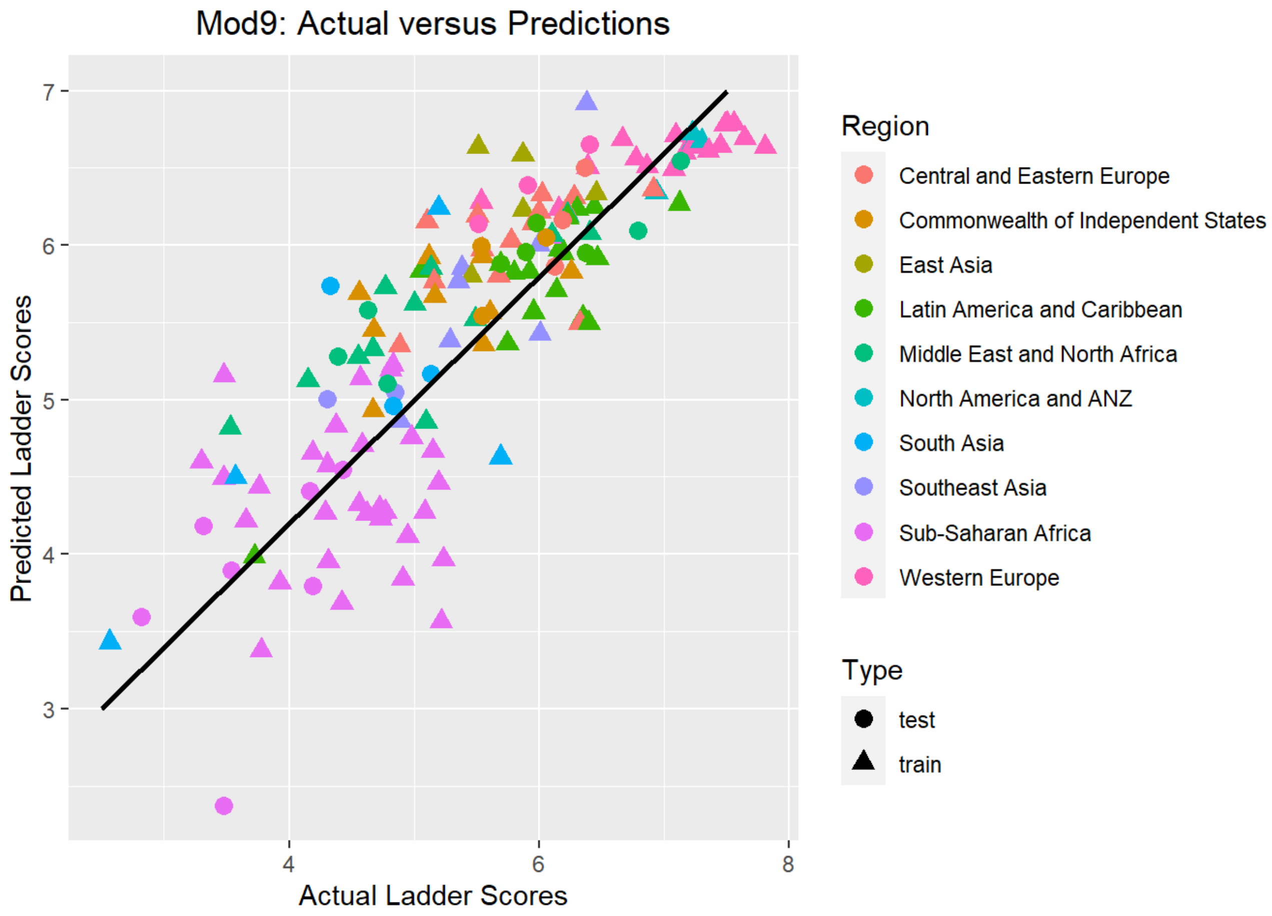
In comparing the MSE and MAE values found for the models, we discovered that mod8 and mod9 had the lowest MSE (0.26) and MAE (0.37) respectively and therefore were chosen as the two models with the highest accuracy. Mod8's variables were logged GDP and social_support, while mod9's variables were logged GDP, social support, and healthy life expectancy. Interestingly enough, mod3 which had only one predictor (social support) contained both the lowest MAE and MSE of all the single predictor models (0.43 and 0.36 respectively). These values were higher than that of mod8 and mod9 which could be due to mod8 and mod9 having more than one predictor. This could also be because mod8 and mod9 contained logged GDP which was the second most single influential variable since mod2, which had logged GDP as a single predictor, had the second smallest MSE/MAE values.

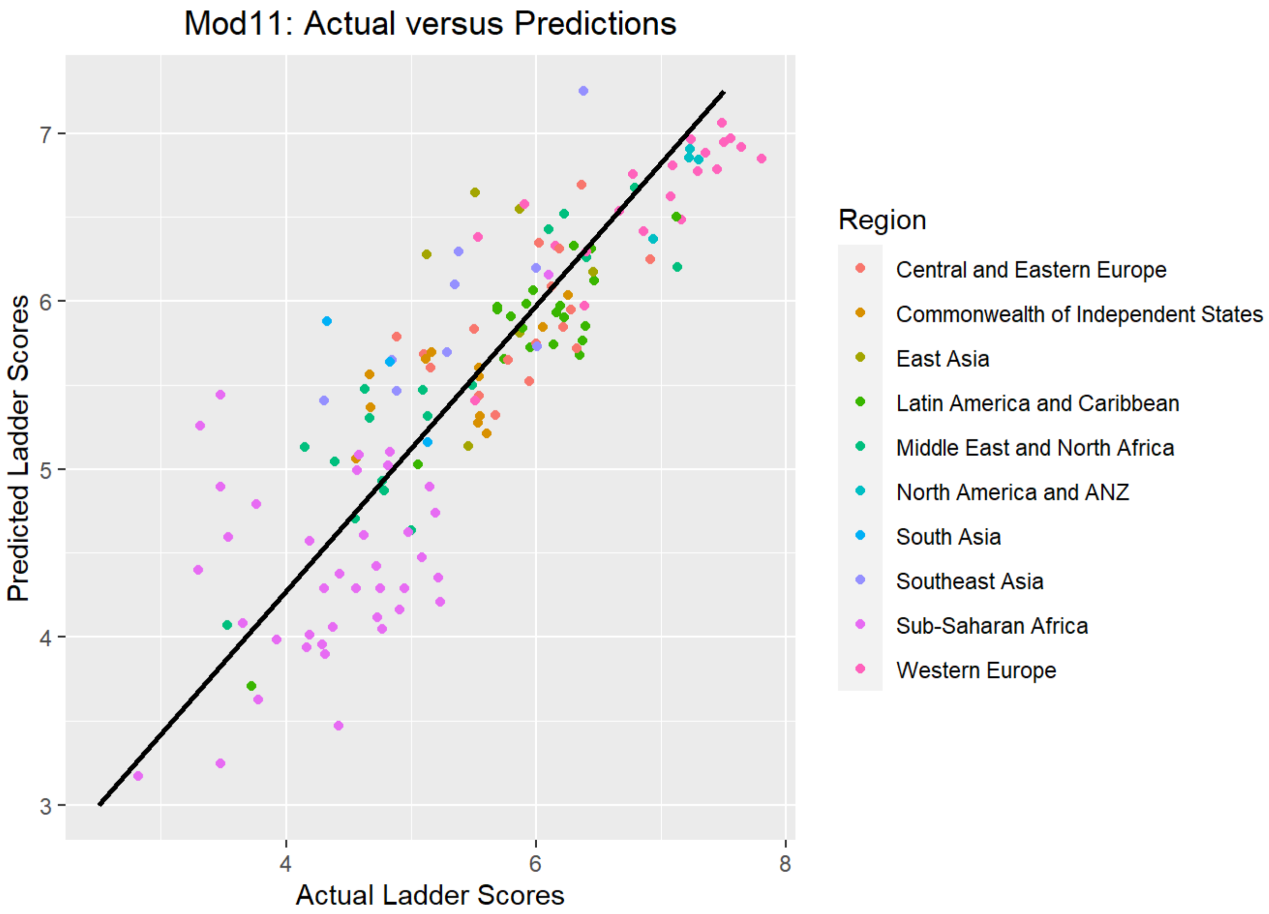
We felt one modeling technique was not enough, so another technique we utilized was cross validation, but in a different way. Instead of splitting up the data into 80% train and 20% test, we split up the data by region. So the training data would be all of the countries except the countries in one particular region and the testing data would be the one region that was left out. This was done for every region which means every region became a test set. The data was split up like this for every model. We then took the root mean squared error (RMSE) for each model to find the model that had the lowest RMSE. Below is a graphical representation of the RMSE for each model.



We found the model with the lowest RMSE value (0.68) to be mod 11. This model utilized the variables logged GDP, social support, and freedom. Using the 3 models we found we plotted the actual ladder score values against the predicted ladder score values from each model.

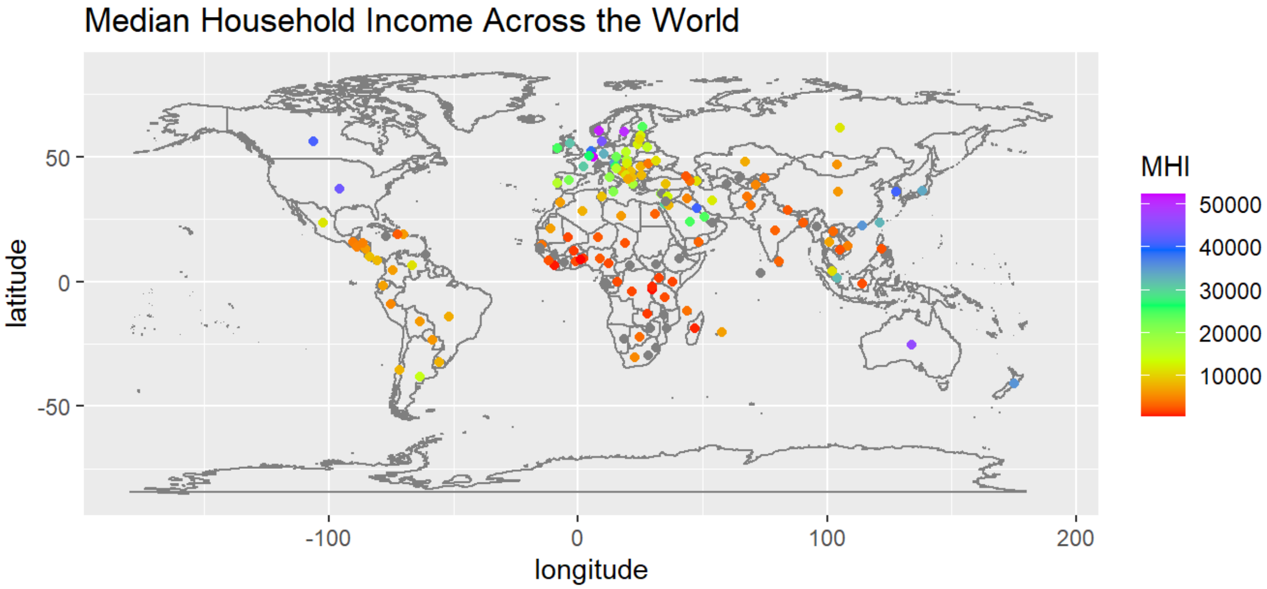

Mod8: Actual versus Predictions
Predicted Ladder Scores
7
6
5
4
3
Actual Ladder Scores
4
6
8
Region
● Central and Eastern Europe
● Commonwealth of Independent States
● East Asia
● Latin America and Caribbean
● Middle East and North Africa
● North America and ANZ
● South Asia
● Southeast Asia
● Sub-Saharan Africa
● Western Europe
Type
● test
▲ train

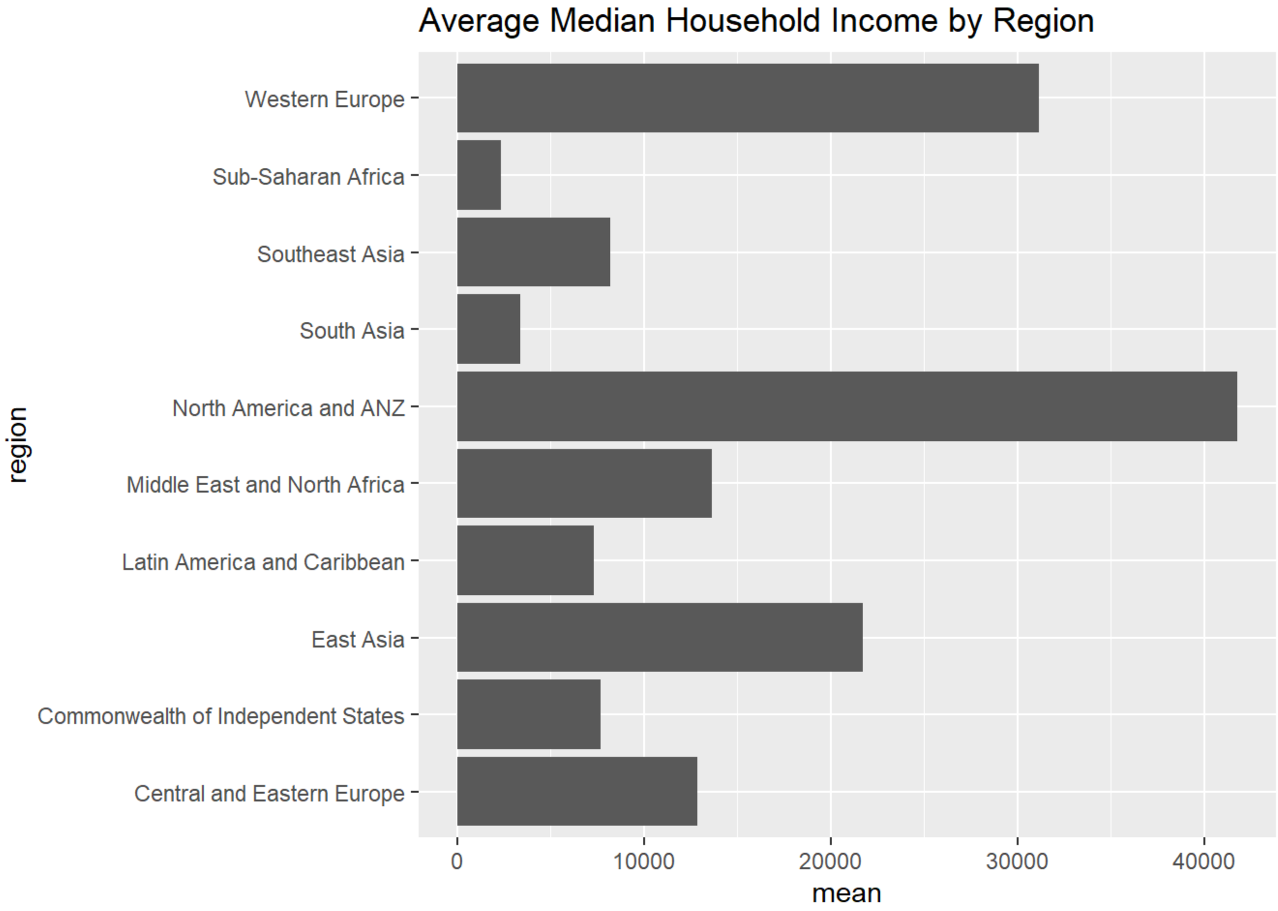

Mod9: Actual versus Predictions
Predicted Ladder Scores
7
6
5
4
3
Actual Ladder Scores
4
6
8
Region
● Central and Eastern Europe
● Commonwealth of Independent States
● East Asia
● Latin America and Caribbean
● Middle East and North Africa
● North America and ANZ
● South Asia
● Southeast Asia
● Sub-Saharan Africa
● Western Europe
Type
● test
▲ train



For all of the models we colored the points by region, but for the models with the lowest MSE and MAE we identified the test and train data by shape. You can see in all three graphs there is a very strong positive correlation between the actual and predicted ladder score values indicating that the predictions of these models are very accurate. After comparing these models, we found that all three had logged GDP, two of the three had healthy life expectancy and two of the three had social support. Therefore, we deemed logged GDP, social support, and healthy life expectancy to be the most significant factors to observe when looking to predict and improve a country's ladder or happiness score.

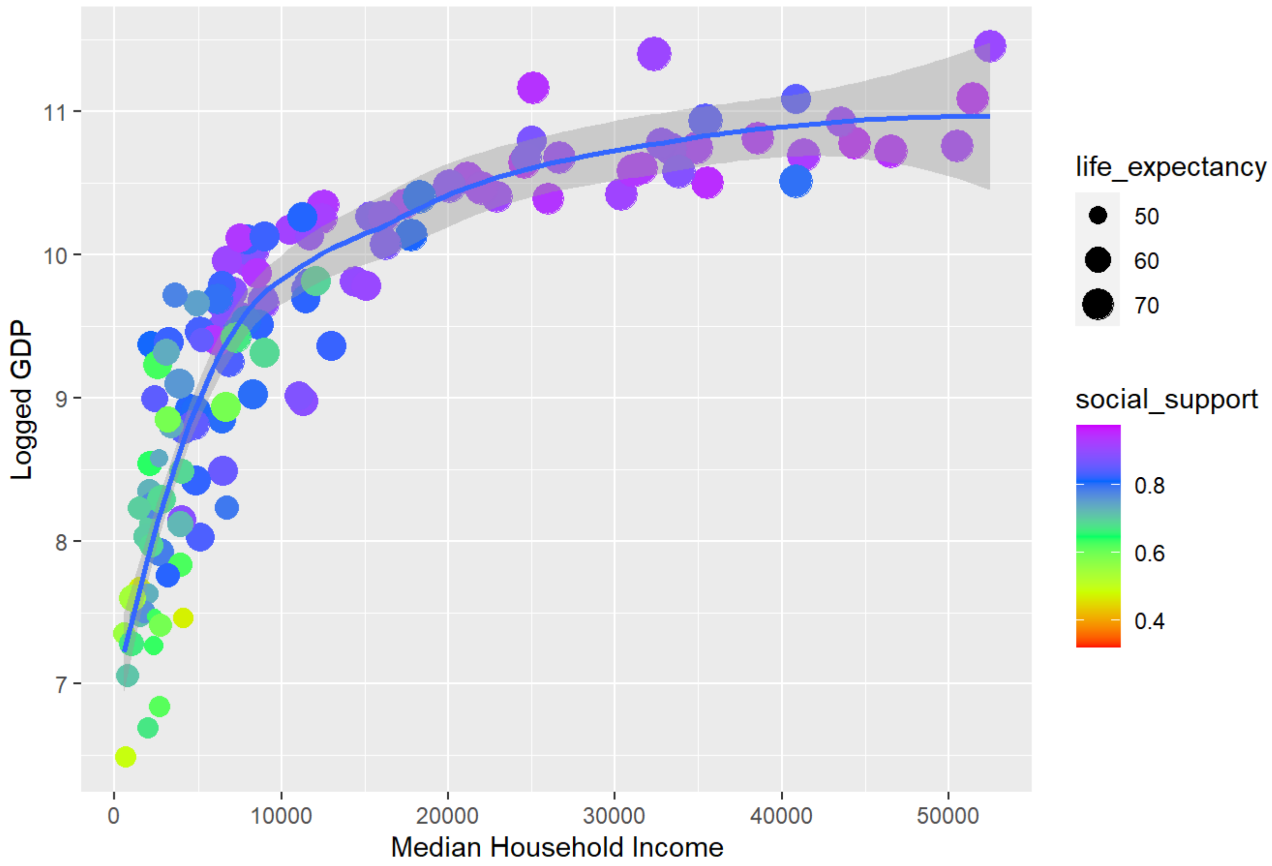
Our second question was, "Can median household income predict the ladder score, logged GDP, healthy life expectancy, and/ or social support and does it relate to each region?". When we took the top 20 happiest and top 20 unhappiest countries, based on ladder score, we found that most of the top happiest countries are in the region Western Europe and most of the unhappiest countries are in the region Sub-Saharan Africa. So, we wanted to further explore the data to see what the median household income looks like in each region.


Median Household Income Across the World
latitude
50
0
-50
longitude
-100
0
100
200
MHI
50000
40000
30000
20000
10000

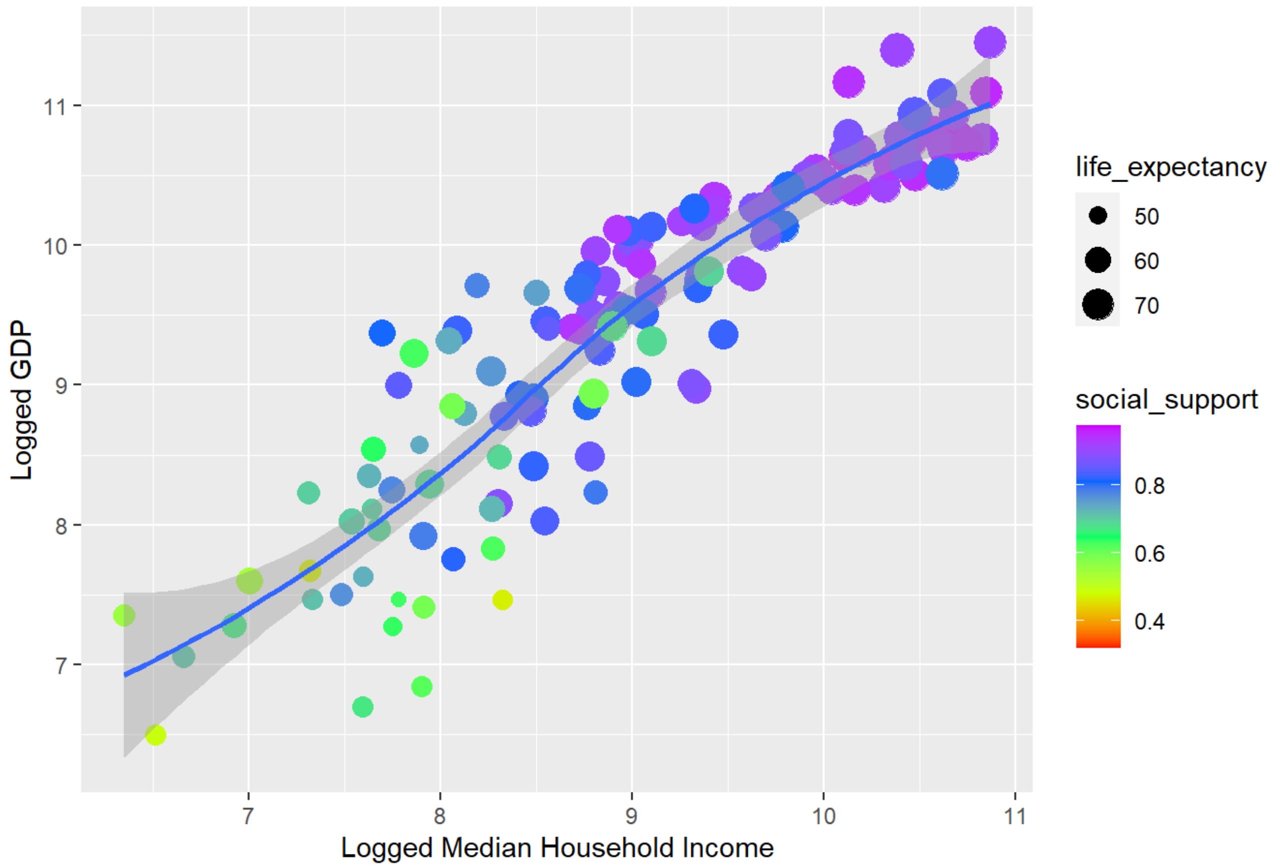

Average Median Household Income by Region
region
Western Europe
Sub-Saharan Africa
Southeast Asia
South Asia
North America and ANZ
Middle East and North Africa
Latin America and Caribbean
East Asia
Commonwealth of Independent States
Central and Eastern Europe
mean
0
10000
20000
30000
40000

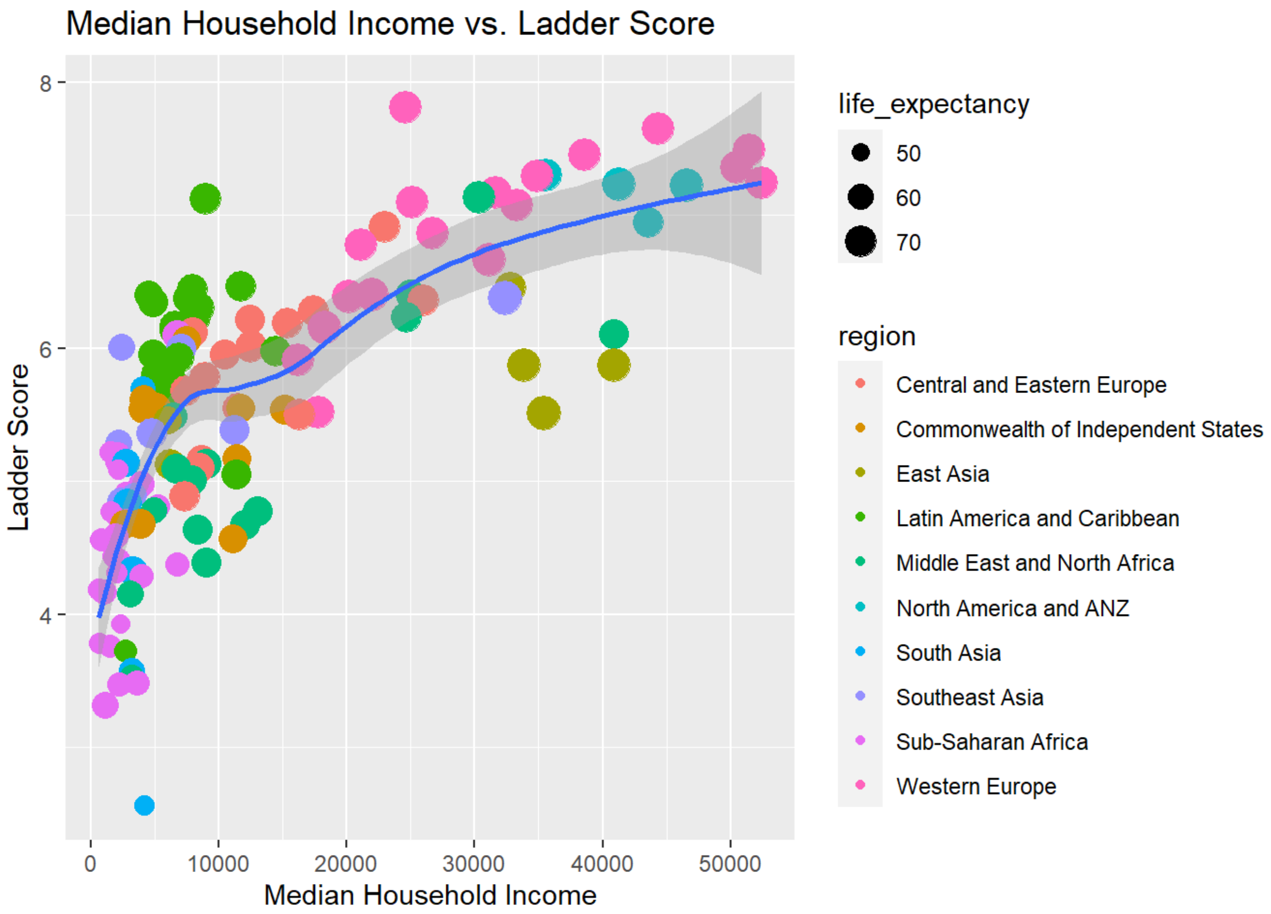
We used “MHI” to stand for median household income in the legend for our map graph. As you can see, most of the countries in Sub-Saharan Africa have a lower median household income whereas Western Europe, which had the highest amount of the top happiest countries, had an overall higher median household income when compared to the other regions. You can see from the bar graph that the regions North American and ANZ along with Western Europe had the two highest mean median household incomes, which is interesting as those countries are more developed. Now that we have a better idea of median household income across the world, to answer whether it can predict logged GDP, social support and healthy life expectancy logged GDP, we first created 4 linear models. These models had median household income predicting logged GDP, ladder score, healthy life expectancy and social support. The relationship between the variables, however, did not look linear when graphed and was even more evident in the low adjusted R-squared values from the 4 models. To combat this we logged the median household income in all 4 models and this increased the adjusted R-squared significantly. The adjusted R-squared went from 0.4905 to 0.7081 for the model predicting healthy life expectancy, 0.3295 to 0.5574 for the model predicting social support, 0.518 to 0.5913 for the model predicting ladder score and 0.5849 to 0.8128 for the model predicting logged GDP.

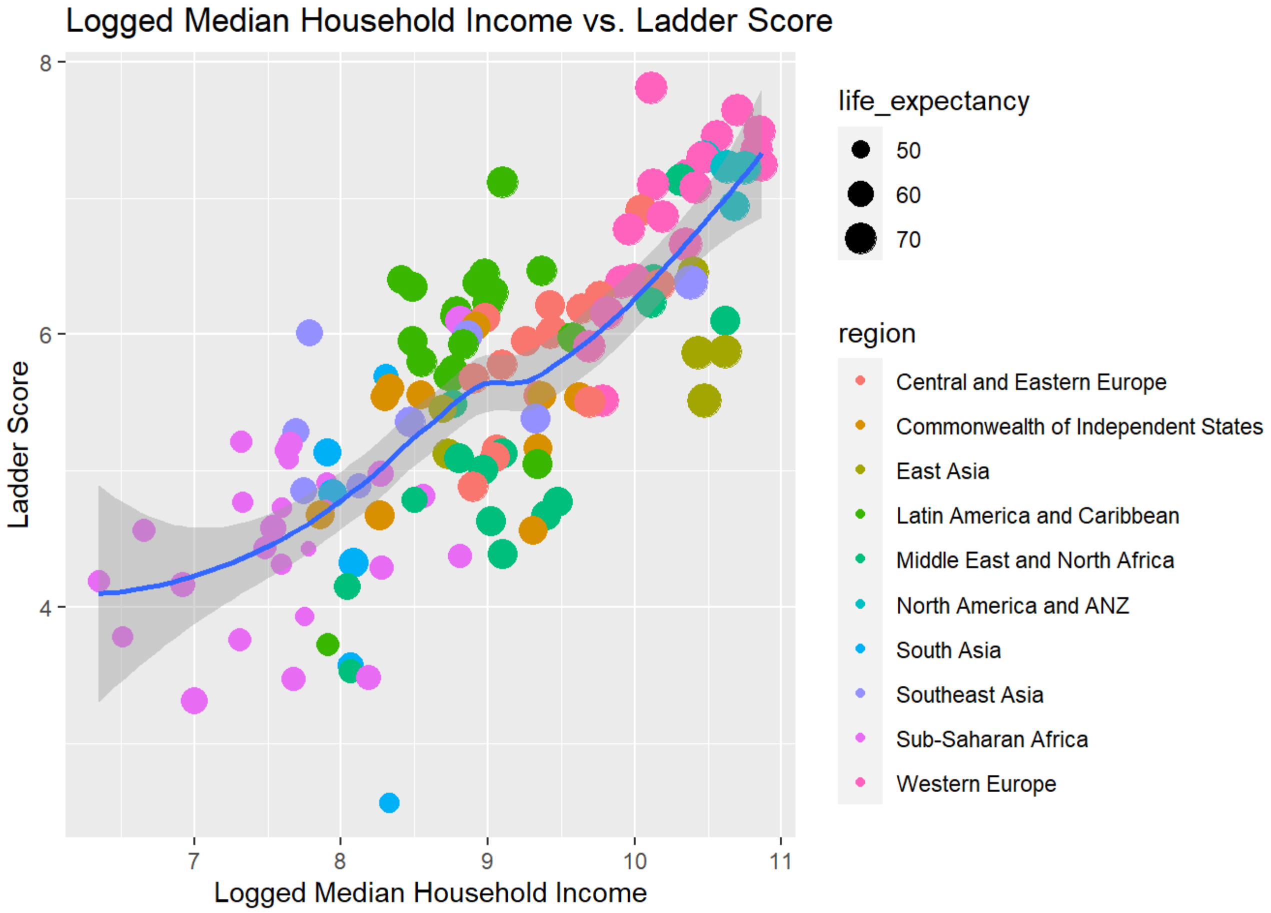
Median Household Income vs. Logged GDP



Logged Median Household Income vs. Logged GDP




Median Household Income vs. Ladder Score
Ladder Score
8
6
4
0
10000
20000
30000
40000
50000
Median Household Income
life_expectancy
50
60
70
region
Central and Eastern Europe
Commonwealth of Independent States
East Asia
Latin America and Caribbean
Middle East and North Africa
North America and ANZ
South Asia
Southeast Asia
Sub-Saharan Africa
Western Europe


Logged Median Household Income vs. Ladder Score
Ladder Score
8
6
4
7
8
9
10
11
Logged Median Household Income
life_expectancy
50
60
70
region
Central and Eastern Europe
Commonwealth of Independent States
East Asia
Latin America and Caribbean
Middle East and North Africa
North America and ANZ
South Asia
Southeast Asia
Sub-Saharan Africa
Western Europe

From the first graph above you can see as median household income increases, the logged GDP increases. It increases rapidly when median household goes from $0 to $10,000, and then you can later see as median household income gets to those higher values around $30,000 and above, logged GDP also starts hitting higher values not previously seen. The healthy life expectancy also increases as you can see the dots start to get larger as median household income increases. The social support also increases as median household income increases as you can see it go from green, which corresponds to lower social support to purple which corresponds to a higher social support value. As the median household income exceeds $10,000, the graph starts to level out. Having more money seems to become less meaningful than before. In the second graph, we logged median household income to make the relationship more linear. This makes sense to do especially since it's being plotted against a GDP that's already logged.

In the third graph above we are now plotting median household income against ladder score and made the color of the points correlate to the region. You can see that Sub-Saharan African countries have a lower median household income along with a lower ladder score, and healthy life expectancy. You can then see that the Western European countries have a higher logged GDP and healthy life expectancy. In the fourth graph we again logged median household income to show a more linear relationship.

Overall, we can see from the analysis of the plots and the adjusted R-squared values of the linear models that the median household income looks like a good predictor for ladder score, logged GDP, healthy life expectancy and social support. As median household increases, ladder score, logged GDP, healthy life expectancy and social support all increase as well. There is also a clear difference in the median household income between each region. The more developed countries seem to have a higher median household income which is interesting because these countries also have a higher ladder score. So it looks like there's strong evidence to support that money really can buy happiness.

CONCLUSION

Our first goal was to answer the question of which variables would be the best predictors of happiness. Through various modeling techniques we narrowed our variables down to logged GDP, social support, and healthy life expectancy as the most significant predictors of a country's happiness (ladder) score. It can be difficult to know how to best aid one's country as a leader, as there are so many factors and so little time and money. By identifying these variables as the primary targets for a country's happiness, a leader's worries of how to best focus their efforts is eliminated. Several avenues that these leaders could apply their efforts based on these variables are: reforming health care and access to quality living necessities, producing more goods and services, decreasing unemployment, increasing social events, and creating better platforms for communication and information. We expected logged GDP and healthy life expectancy to be quite significant predictors due to how influential they are in a country's well being and determining how developed or undeveloped a country is. The other variable that showed up as being an important predictor was social support. This was quite interesting in that there have actually been many studies done on how people who perceive their family or friends to be supportive tend to be happier. There have also been studies done on how social connections in general can boost happiness, whether in daily life or in general. Happiness is probably the most key thing people strive for in life which is what makes the investigation of, what makes the people in a country happier, all the more valuable.

There is much room for additional investigation such as observing whether access to technological developments, a country's landscape, culture, average stress levels or average work hours, would influence how happiness is determined or predicted. We didn't have access to data on average stress levels for all the countries in our data set, and the data that was available was surveyed in 2012 which was more outdated than we wanted. It could also be interesting to see how external factors such as climate or natural disasters affect the happiness score of a country. There are a variety of variables that could be influencing how happy someone is which gives a plethora of room for further investigation. There are so many factors that weren't considered that we could be missing out on that may better predict happiness. Some methods we did not use that could have been better include creating and analyzing plots of the residuals and looking at Cook’s Distance for the different models. Creating even more models with different combination of the predictors or interactions between the predictors could also help increase the accuracy of finding the best model to predict happiness.

Our second goal was to answer the question, “Can median household income predict the ladder score, logged GDP, healthy life expectancy, and/ or social support and does it relate to each region?”. We came to the conclusion that median household income can predict these variables. When median household increases, ladder score, logged GDP, healthy life expectancy, and social support all increase. We also did see a relation with median household income and region with more developed regions such as Western Europe and North American and ANZ having higher median household incomes and regions such as Sub-Saharan Africa having lower median household incomes. In the first question we narrowed down the the most important predictors for happiness to logged GDP, healthy life expectancy and social support. So, finding that median household income can predict these variables is really valuable because when world leaders are trying to figure out how to improve predictors such as logged GDP, social support and life expectancy they now know that finding ways to increase median household income could do all three thus increasing ladder score indirectly. These results are worth investigating further since the median household income shows the worth of individual wealth, rather than simply focusing on the wealth of the country alone. For example, even though a country may have a high GDP per capita, the wealth gap between the rich and poor may result in lower happiness across the country, despite the correlation GDP has with the ladder score. Countries may look into further supplementing government programs for those below the poverty line, or investing in long-term factors that generate higher average income, a prime example being education.

Using median household income to investigate happiness is merely a starting point for the wide variety of variables that could go into refining the happiness model, examples of which were stated earlier. Literacy rate, however, may go particularly well with median household income as a way to measure education, which could answer how education affects median household income, which in turn affects GDP and happiness. Other ways to improve our modeling techniques include investigating multicollinearity, since GDP per capita and median household income still deals with money, although they are on different scales. Likewise, narrowing down an accurate model with the lowest RMSE using external variables outside the dataset may serve to improve our analysis. However, the variables we have identified currently provide great guidance for how a country can seek to improve its happiness.