Elixir of Happiness: A Data-Driven Approach

Ilkin Bayramli, Dasha Metropolitansky, Maarten de Vries, Chang Yu

December 9, 2020

Introduction

In 1998, the prime minister of Bhutan, a country in South Asia, introduced a concept called "Gross National Happiness" to a forum at the United Nations. He proposed that instead of evaluating a country's development using economic metrics like GDP, we should measure happiness. Although the idea was initially viewed as radical, interest grew over the next ten years as people around the world felt increasingly disenfranchised from capitalism. In 2011, the UN General Assembly passed a resolution urging member nations to measure happiness and use the results in crafting public policy. In 2012, the UN issued the first World Happiness Report, which ranks countries based on surveys of their population's happiness.

We were interested in exploring how various factors, common to people in all countries, influence happiness. We had many preconceived notions about this subject. For example, our intuition was that healthy, high income, and highly educated individuals who are the primary breadwinners in their households and work a skilled job would be the happiest. We wanted to test the veracity of these assumptions. There were also many factors, such as whether a person is religious, lives in a small town or a big city, the size of their family, and whether they are immigrants, whose relationship with happiness we were unsure of and wanted to learn more about.

We decided to use data from the World Values Survey Association, which has surveyed hundreds of thousands of people around the world about their values and perspectives on social, political, economic, moral, and religious issues. Note that the UN World Happiness Report also uses the World Values Survey in developing its rankings and analyses. This paper outlines how we used the survey data to determine which factors influence happiness.

Data Collection and Wrangling

We downloaded the Wave 7 dataset from the World Values Survey website. This dataset contains data from surveys conducted in 49 countries/territories from 2017 to 2020. Each row corresponds to an individual, and the columns correspond to survey questions. The original dataset contains 69,578 rows and 536 variables. We only used 44 variables for our analysis.

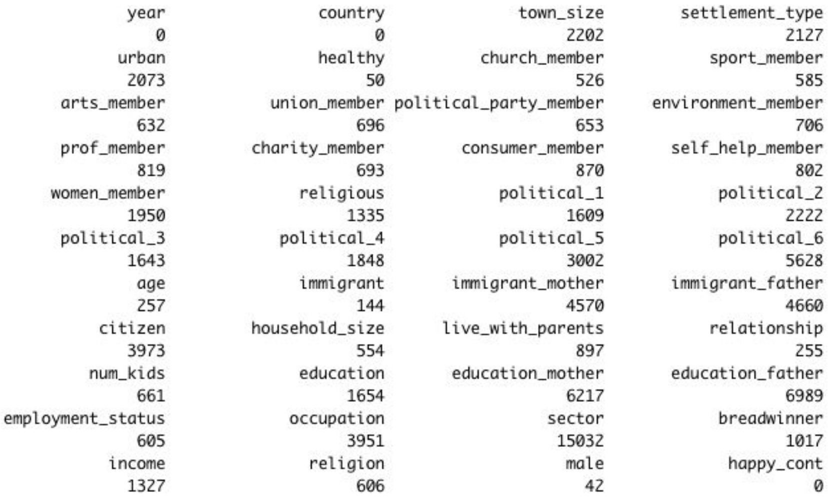
We included two main types of survey questions: those with quantifiable and objective answers, and those which would give us a well-rounded picture of people's lives (i.e. social, political, financial, and physical questions). We found that many of the questions in the full survey, while interesting, were very subjective and would be highly correlated with each other. For example, there were many questions about ethics and cultural values.

For a complete list of the variables we included in our dataset, please see the Appendix. Broadly, the topics covered include year, country, settlement size and type, health, happiness, membership of various types of organizations, religious affiliation, political engagement, gender, age, immigration and citizenship status, family's immigration and citizenship status, relationship status, number of children, family size, and occupation.

Almost all of the variables in our dataset are categorical, where every category is coded as a number. The main challenge in data wrangling was deciding which variables to convert to indicator variables. There were some clear cases. For example, our "religious" variable is based on the question "Independently of whether you go to church or not, would you say you are…" with responses of 1 = a religious person, 2 = not a religious person, and 3 = an atheist. We converted this to an indicator variable for whether the person is religious, where a response of 1 means they are, and 2 or 3 means they aren't.

Some variables were much less straightforward. For example, the "happy" variable, which is our response variable, is based on the question "Taking all things together would you say you are…" with responses of 1 = very happy, 2 = quite happy, 3 = not very happy, and 4 = not at all happy. Some members of our group thought that creating an indicator variable for happiness would be appropriate: grouping 1 and 2 into happy, and 3 and 4 into not happy. Other members thought that we'd be losing nuance in the responses by grouping them. We ultimately decided to experiment with both versions of the happiness variable.

Lastly, we had to handle missing data. For nearly all questions, we found missing observations ranging from under 100 to over 15,000. **Figure 1** shows the breakdown of missing values by variable in the dataset. We decided it would be unwise to drop observations with missing values since over half of observations had at least one missing value. Therefore, we decided to impute the column mode instead. We discuss this decision further in the Discussion section. We also removed the observations where the response variable was missing, leaving 55,341 observations.



```
             year         country              town_size     settlement_type
                0               0                   2202                2127
            urban         healthy          church_member        sport_member
             2073              50                    526                 585
      arts_member    union_member political_party_member  environment_member
              632             696                    653                 706
      prof_member  charity_member        consumer_member    self_help_member
              819             693                    870                 802
     women_member       religious            political_1         political_2
             1950            1335                   1609                2222
      political_3     political_4            political_5         political_6
             1643            1848                   3002                5628
              age       immigrant       immigrant_mother    immigrant_father
              257             144                   4570                4660
          citizen  household_size      live_with_parents        relationship
             3973             554                    897                 255
         num_kids       education       education_mother    education_father
              661            1654                   6217                6989
employment_status      occupation                 sector         breadwinner
              605            3951                  15032                1017
           income        religion                   male          happy_cont
             1327             606                     42                   0
```

Figure 1

Exploratory Data Analysis

After creating our final dataset, we conducted a preliminary analysis of the data. First, we looked at the distribution of happiness, our response variable. **Figure 2** shows that 53% of people in our dataset reported being quite happy, 32% reported being very happy, 12% reported being not very happy, and only 2% reported being not at all happy.

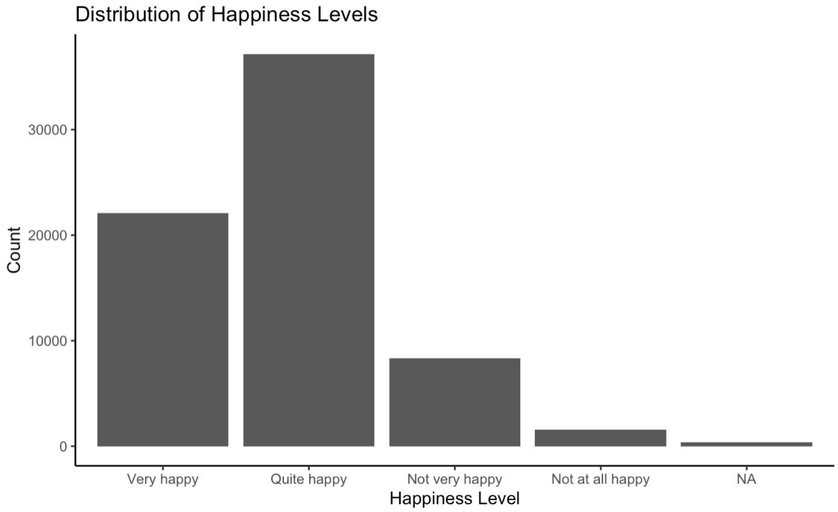
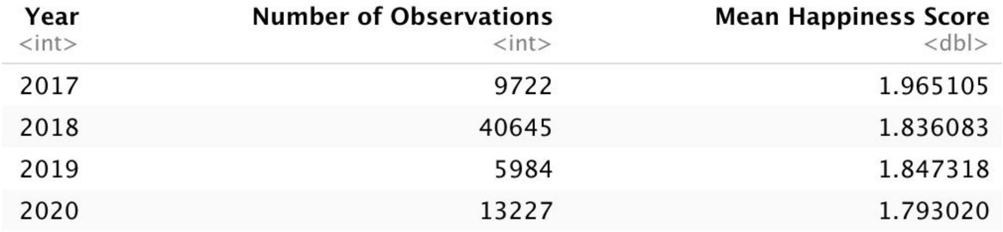


Figure 2

We also looked at happiness over time (**Figure 3**). Note that since a score of 1 corresponds to a response of "very happy" and 4 corresponds to a response of "not at all happy", a lower mean score implies higher happiness. Interestingly, we found that 2020 was the happiest year from 2017 to 2020. This result is counterintuitive and difficult to explain, especially given the large sample size. However, it is important to note that the data for 2020 isn't complete yet: it only includes surveys conducted up until September. Thus, it's possible that the mean happiness score will change with more data.



Year <int>	Number of Observations <int>	Mean Happiness Score <dbl>
2017	9722	1.965105
2018	40645	1.836083
2019	5984	1.847318
2020	13227	1.793020

Figure 3

Next, we plotted the average happiness score by country, as shown in **Figure 4**. It's clear that there are country-level differences in happiness, which indicates that a mixed effects model with country as the grouping factor would likely be appropriate.

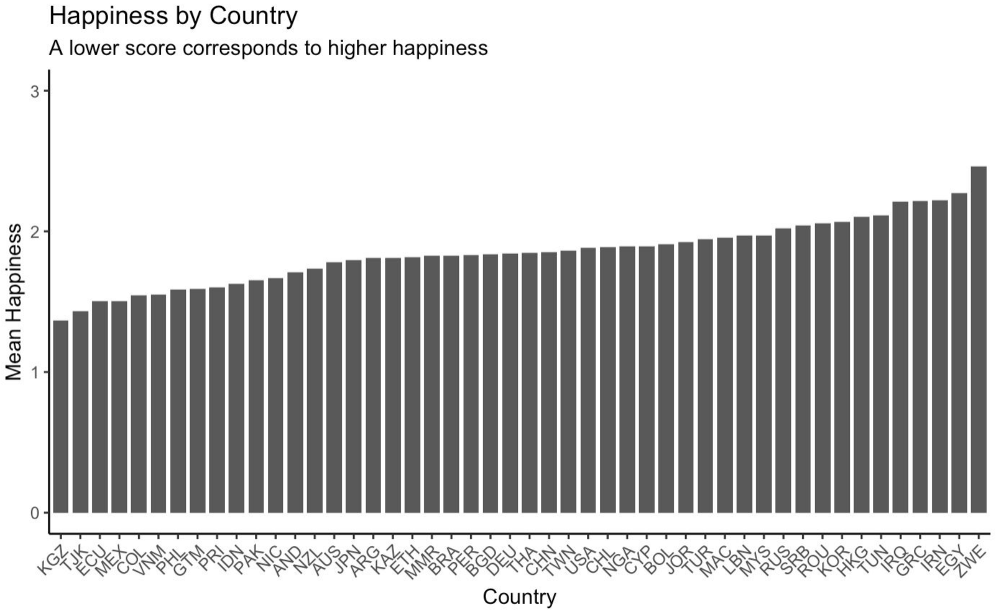


Figure 4

We also looked at the relationships between our predictor variables and happiness. There are too many variables to show here, so we've included a few that we found interesting in **Figure 5**. Firstly, health appears to have a strong relationship with happiness: the proportion of healthy people who are happy is much larger than the proportion of happy unhealthy people. The same is true for income: those in the highest income levels appear to be significantly happier than those in the lowest income levels. In contrast, and perhaps surprisingly, education level does not seem to be strongly correlated with happiness. Lastly, we see that church membership status has a fascinating relationship with happiness: although non-members and members have approximately the same proportion of happy people, the latter group has a larger proportion of very happy people.

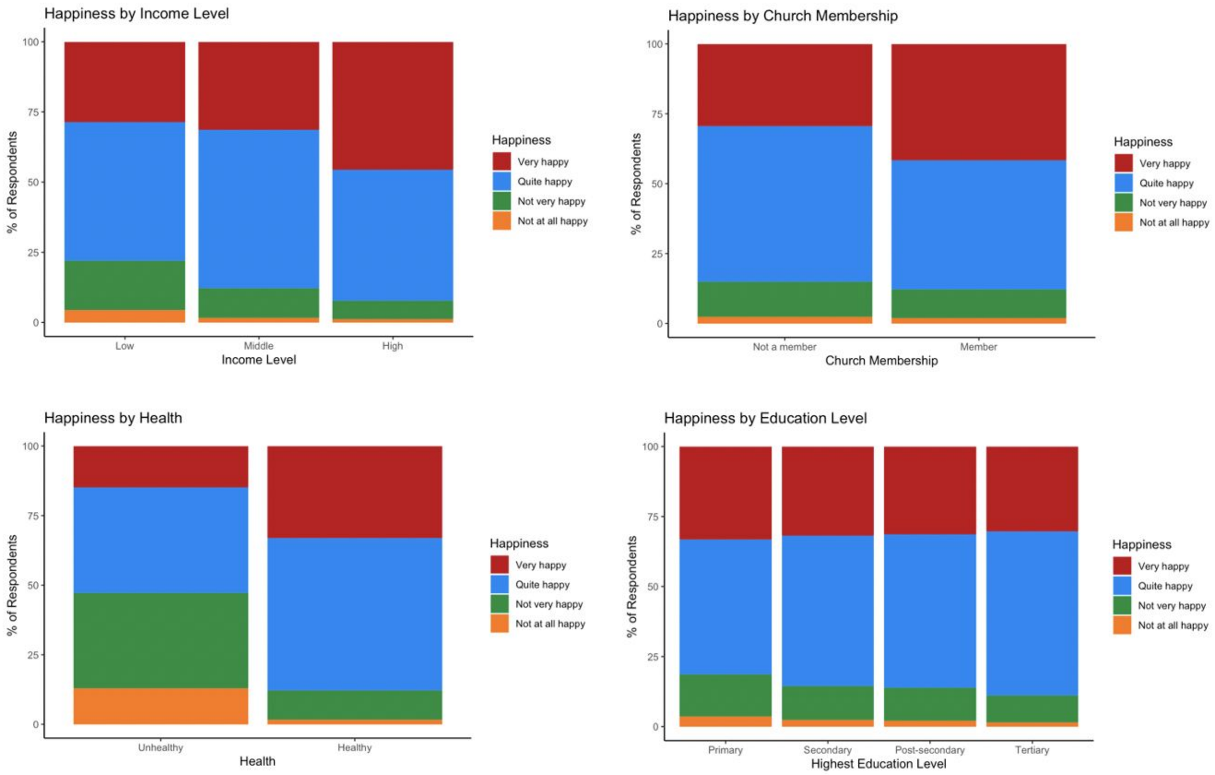


Figure 5

Finally, we plotted the distribution of the numeric variables in our dataset: age, num_kids, household_size, and ideology. **Figure 6** shows that age, household_size, and num_kids are right-skewed. As a result, we tried log-transforming these variables.

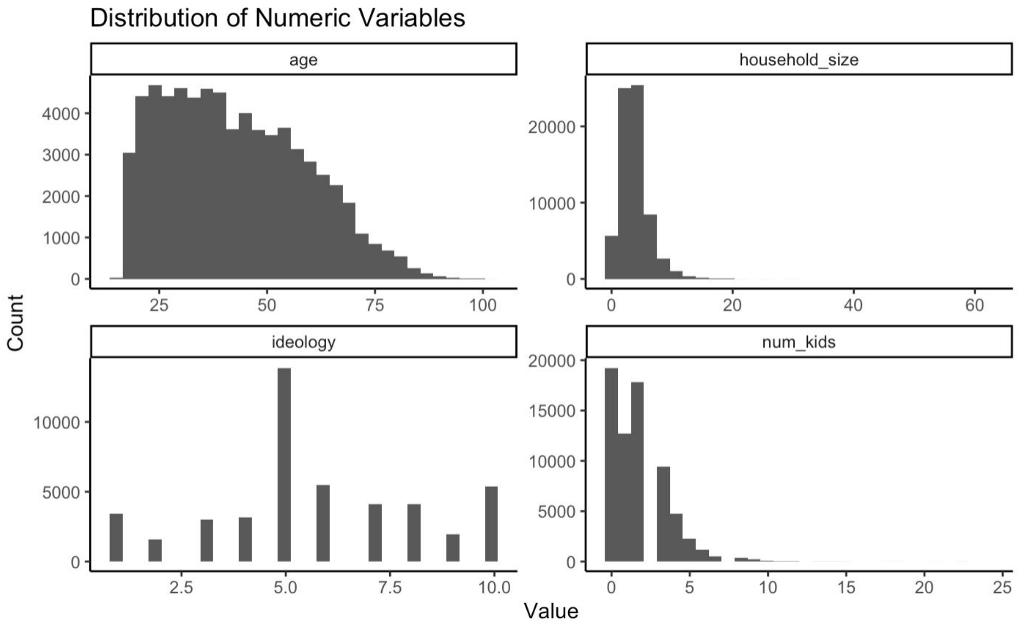


Figure 6

We see in **Figure 7** that although the log-transformed variables are now less skewed, they are still not entirely symmetric. We could try a different transformation, such as square root; however, this would sacrifice the interpretability of our model, so we decided not to do so.

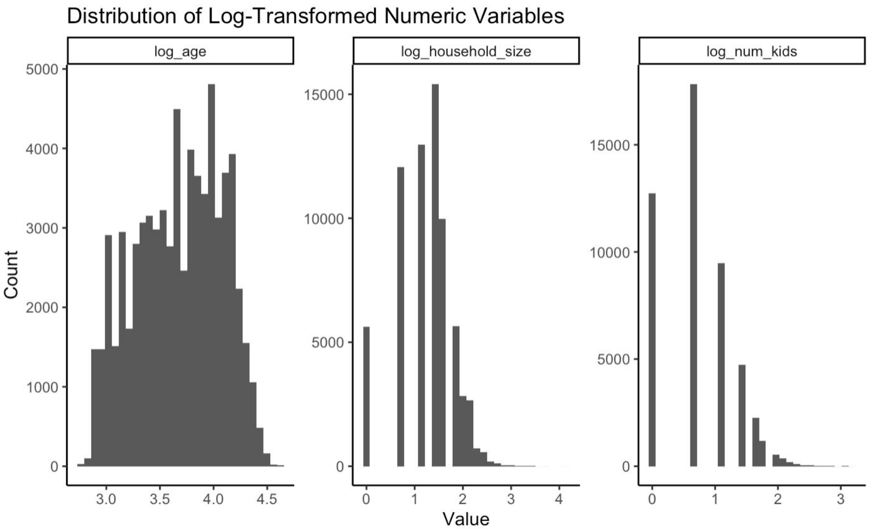


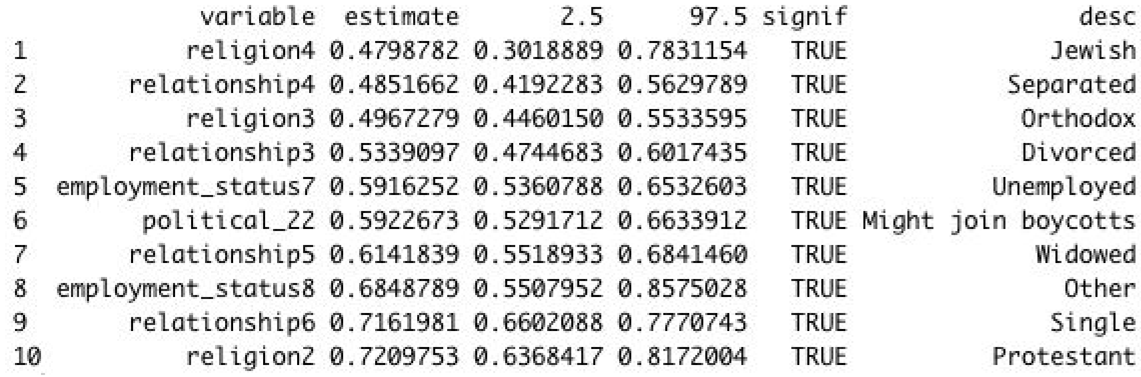
Figure 7

Models

Multiple Logistic Regression

First, we decided to perform a multiple logistic regression (model logit3) using all variables except country, year and age. We used a binary indicator for happiness as our response variable, where 1 denotes “happy” and 0 denotes “unhappy”, thus combining happiness scores 1 and 2, as well as 3 and 4.

Figure 8 shows the 10 variables with the smallest exponentiated beta coefficient estimates as well as their 95% confidence intervals (all of which were statistically significant at an alpha level of 0.05).

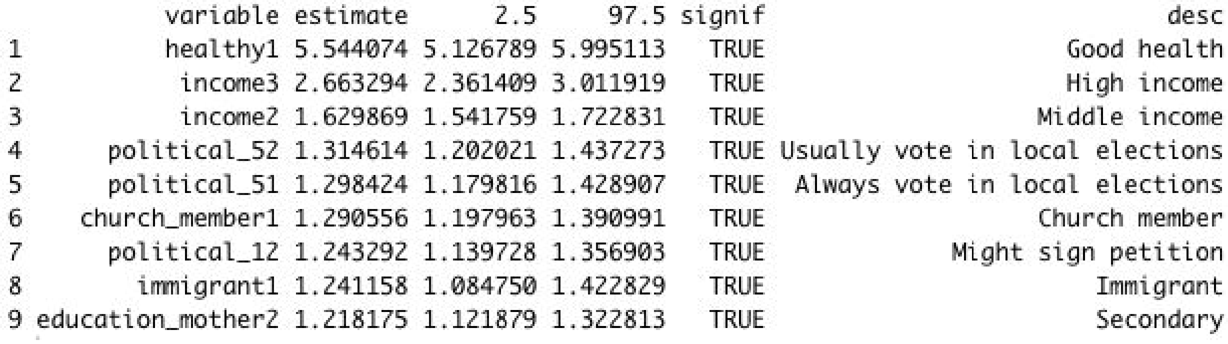


```
              variable  estimate       2.5      97.5 signif               desc
1            religion4 0.4798782 0.3018889 0.7831154   TRUE             Jewish
2        relationship4 0.4851662 0.4192283 0.5629789   TRUE          Separated
3            religion3 0.4967279 0.4460150 0.5533595   TRUE           Orthodox
4        relationship3 0.5339097 0.4744683 0.6017435   TRUE           Divorced
5  employment_status7 0.5916252 0.5360788 0.6532603   TRUE         Unemployed
6         political_22 0.5922673 0.5291712 0.6633912   TRUE Might join boycotts
7        relationship5 0.6141839 0.5518933 0.6841460   TRUE            Widowed
8  employment_status8 0.6848789 0.5507952 0.8575028   TRUE              Other
9        relationship6 0.7161981 0.6602088 0.7770743   TRUE             Single
10           religion2 0.7209753 0.6368417 0.8172004   TRUE         Protestant
```

Figure 8

The interpretation of the estimates above (e^{β_j}) is the multiplicative change in odds if the variable has indicator 1, holding other predictors fixed. We observe that according to our logistic model, the variables that correspond to the lowest odds ratios of being happy include membership of certain religious denominations, being unemployed, and not having a partner.

On the flipside, **Figure 9** shows statistically significant variables with the highest odds ratios of being happy. They include being healthy, having moderate to high income, church membership, and involvement in politics at the local level.

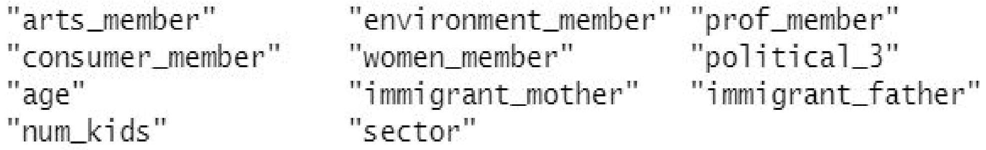


```
          variable estimate      2.5     97.5 signif                          desc
1          healthy1 5.544074 5.126789 5.995113   TRUE                   Good health
2           income3 2.663294 2.361409 3.011919   TRUE                   High income
3           income2 1.629869 1.541759 1.722831   TRUE                 Middle income
4      political_52 1.314614 1.202021 1.437273   TRUE Usually vote in local elections
5      political_51 1.298424 1.179816 1.428907   TRUE  Always vote in local elections
6    church_member1 1.290556 1.197963 1.390991   TRUE                  Church member
7      political_12 1.243292 1.139728 1.356903   TRUE             Might sign petition
8        immigrant1 1.241158 1.084750 1.422829   TRUE                      Immigrant
9 education_mother2 1.218175 1.121879 1.322813   TRUE                      Secondary
```

Figure 9

For variable selection, we preferred backwards stepwise selection over LASSO due to the inferential nature of this project. Furthermore, the glmnet implementation of LASSO turns factors into multiple dummy variables under the hood and drops the factor levels independently. As a result, some levels of a categorical variable remain in the model while others get filtered out. A more thorough discussion of drawbacks of LASSO and potential solutions to them is given in the Discussion section.

Figure 10 shows the variables that were filtered out during backwards selection starting with the multiple regression model described above.



```
"arts_member"        "environment_member" "prof_member"
"consumer_member"    "women_member"       "political_3"
"age"                "immigrant_mother"   "immigrant_father"
"num_kids"           "sector"
```

Figure 10

Mixed Effects

We hypothesize that the distribution of happiness levels are correlated for our samples at the country level. Political and economic circumstances of a country X would affect the happiness of samples from X but not those from countries Y and Z. Moreover, people living in some types of countries are potentially more likely to be happier than people living in other types of countries. Existence of scenarios like these makes “country” a natural grouping variable for a mixed effects model. The histograms in **Figure 11** support this argument.

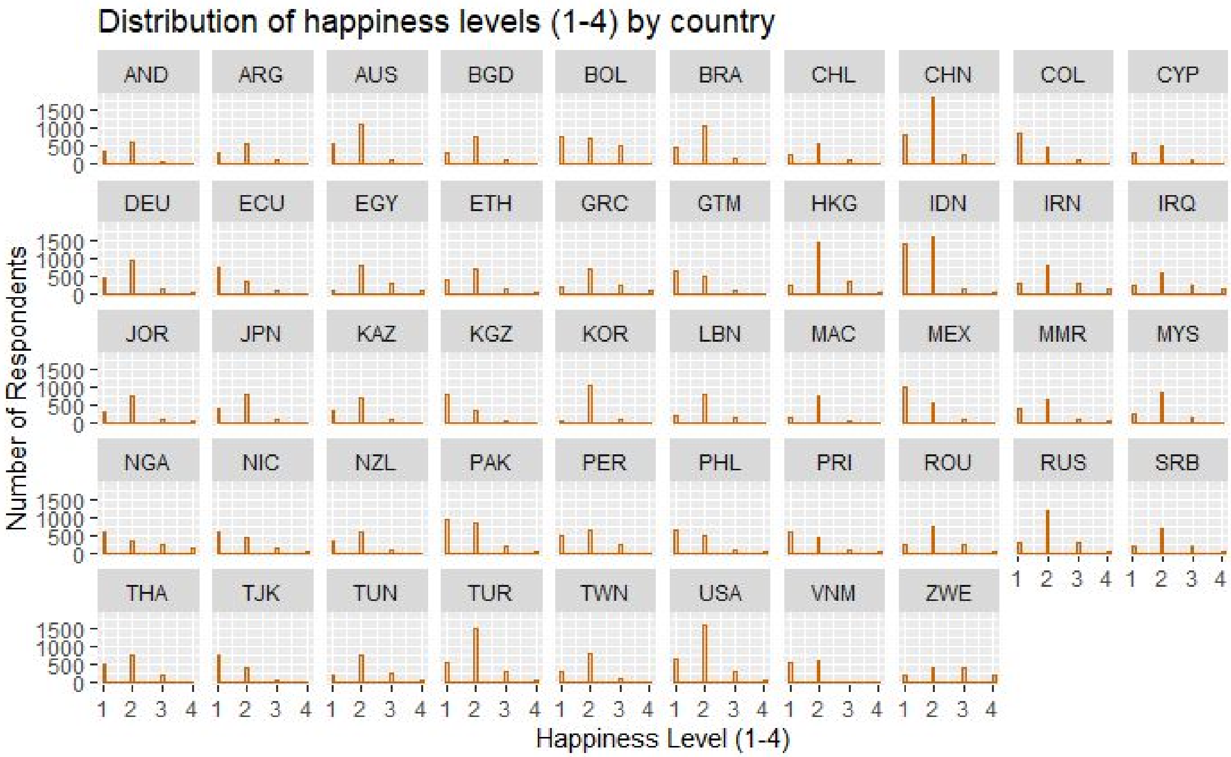


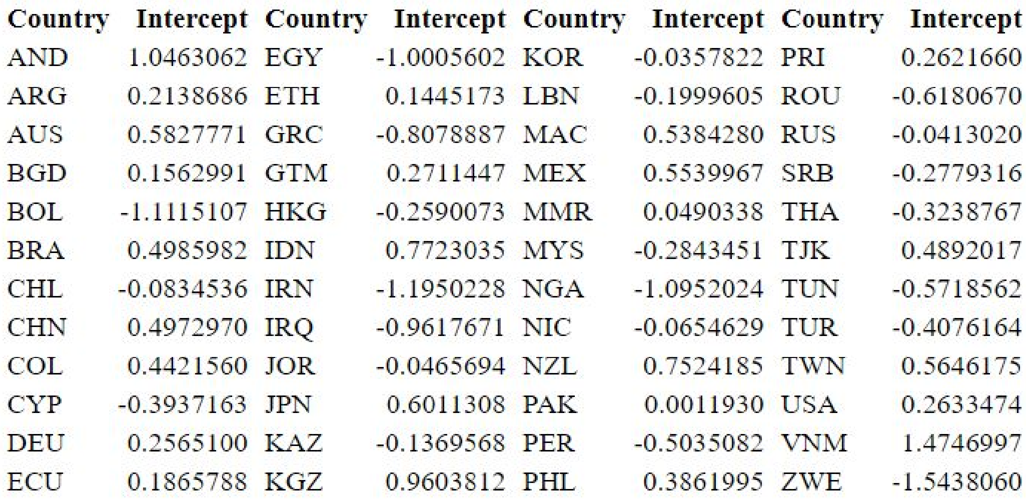
Figure 11

Here we see important distributional differences among countries. In developing countries such as BOL, COL , ECU, GTM, KGZ, MEX, NGA, NIC, PAK, PHL, PRI, TJK the mode of the happiness responses was 1. Meanwhile, most of the developed countries in the dataset have mode happiness levels of 2. This observation makes mixed effect models grouped by "country" highly relevant.

Thus, we decided to fit a logistic random intercept model predicting happiness with "country" being the source of randomness. The significance of most variables slightly decreased compared to the logit3 multiple regression model (where the "country" variable was not included at all), and some variables fell out of significance such as "political_party_member", "charity_member", "union_member", "immigrant", "citizen", "household_size", and "male". This is likely due to the collinearity of these variables with the predictor "country". For example, in developing countries membership of political parties, charitable organizations, and unions can be rare and immigration is uncommon. The signs of the variables that remained significant did not change; some variables that lost significance flipped signs. Most variables deemed significant during variable selection also remained significant in the mixed effect model[1].

Since the signs of important variables didn't change, the fixed-effect interpretations remain the same as those for the multiple regression model. **Figure 12** shows the country-level intercept estimates from our model.

[1] For multi-level factors this means that at least one level remained statistically significant..



Country	Intercept	Country	Intercept	Country	Intercept	Country	Intercept
AND	1.0463062	EGY	-1.0005602	KOR	-0.0357822	PRI	0.2621660
ARG	0.2138686	ETH	0.1445173	LBN	-0.1999605	ROU	-0.6180670
AUS	0.5827771	GRC	-0.8078887	MAC	0.5384280	RUS	-0.0413020
BGD	0.1562991	GTM	0.2711447	MEX	0.5539967	SRB	-0.2779316
BOL	-1.1115107	HKG	-0.2590073	MMR	0.0490338	THA	-0.3238767
BRA	0.4985982	IDN	0.7723035	MYS	-0.2843451	TJK	0.4892017
CHL	-0.0834536	IRN	-1.1950228	NGA	-1.0952024	TUN	-0.5718562
CHN	0.4972970	IRQ	-0.9617671	NIC	-0.0654629	TUR	-0.4076164
COL	0.4421560	JOR	-0.0465694	NZL	0.7524185	TWN	0.5646175
CYP	-0.3937163	JPN	0.6011308	PAK	0.0011930	USA	0.2633474
DEU	0.2565100	KAZ	-0.1369568	PER	-0.5035082	VNM	1.4746997
ECU	0.1865788	KGZ	0.9603812	PHL	0.3861995	ZWE	-1.5438060

Figure 12

The most interesting observation to note here is that all of the developed countries in this list with the exception of CYP and KOR have a positive intercept. The mean country effect, however, is negative at approximately -0.144, which is intuitive since most countries on this list are developing ones with negative intercepts. However, there is not much evidence of shrinking towards the overall mean in countries with smaller sample sizes (CHL, ARG, CYP). This is expected since all countries in the dataset contain at least 900 survey responses. Lastly, note that the variance of the intercept is around 0.46 which is quite high given the scale of these intercepts. This shows that indeed country of residence affects happiness levels of subjects.

We considered adding additional random intercepts to the model since the way different variables interact with happiness can vary from country to country. For example, the relationship of immigration with happiness can be different depending on a country's openness to immigrants. However, adding even one categorical variable as a random interaction effect with the variable "country" led to non-convergence during optimization. These computational issues are addressed in detail in our Discussion section.

Checking Assumptions

Given the final random intercept model, it is important to check for assumption violations before interpreting our results. Since logistic regression assumptions are difficult to check, we examine some basic ones:

1. Independence: This was accounted for by the survey design: in each country, the samples were collected using stratified random sampling[2]. We accounted for international differences in sample distributions with random intercepts.

[2] http://www.worldvaluessurvey.org/WVSContents.jsp

2. Normality: The EDA and Data Wrangling sections addressed non-normalities in data as well as transformations used for fixing them.

3. Normality of Random Intercepts: The distribution of random effects is assumed to follow a normal distribution by our mixed effects model. The histogram of computed random intercepts in **Figure 13** shows that our assumption is not violated.

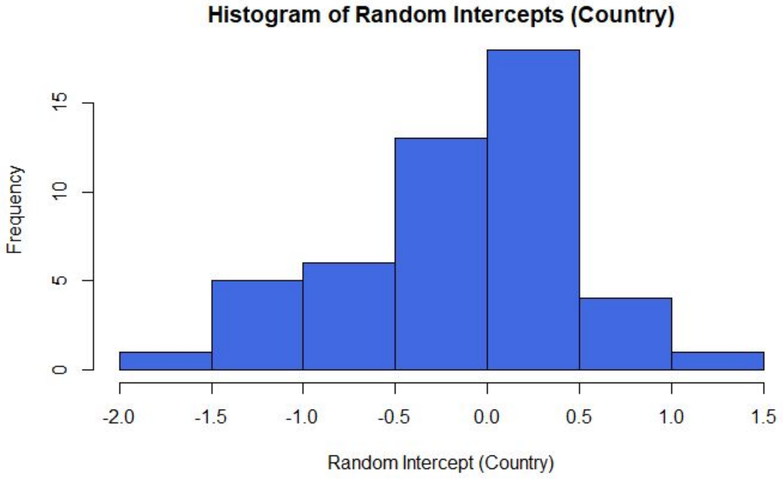


Figure 13

Results

From our models, we find that the most important predictors in determining happiness are income and health when adjusted for country effects. Other significant predictors are occupation, relationship status, and education, although their effect on happiness is smaller in magnitude. This model supports our preconceived assumptions about healthy, high income, and highly educated individuals being the happiest in a society. Other factors which we suspected, such as religion and immigrant status, were not universally significant as adjusting for country effects proved. The following paragraphs discuss this behavior more quantitatively.

Before examining the two most influential predictors, it is worth noting that the effect of countries on happiness as a grouping variable is significant due to its large standard deviation (~ 0.6389) relative to the scale of random intercepts. Therefore, we must control for this variable to see whether our predictors hold their significance intrinsically or only as a correlate of country. To this end, we first examine how the effects of health and income changed when moving from a linear model that does not account for country effects to a model that does. In the case of income, we used three brackets: income1, income2, and income3, in increasing order of value with income1 being the baseline for the other two categories to compare to. For both models, health and income were statistically very significant (p ~ 2e-16). A more interesting picture is depicted when we examine how each variable's coefficients change after adjusting for country effects:

Variable	Coefficient before controlling for country effects	Coefficient after controlling for country effects
Health	1.72	1.75
Income 2	0.51	0.54
Income 3	0.90	0.87

Table 14: Coefficients of Health and Income Variables Before and After Adjusting for Country Effects

As seen, coefficients remain nearly identical after adjusting for country effects, suggesting that the relationship among income, health, and happiness is independent of country effects. In other words, income and health are universal predictors of happiness regardless of culture. In addition, high magnitudes of these coefficients imply that health and income implies that this association is rather strong.

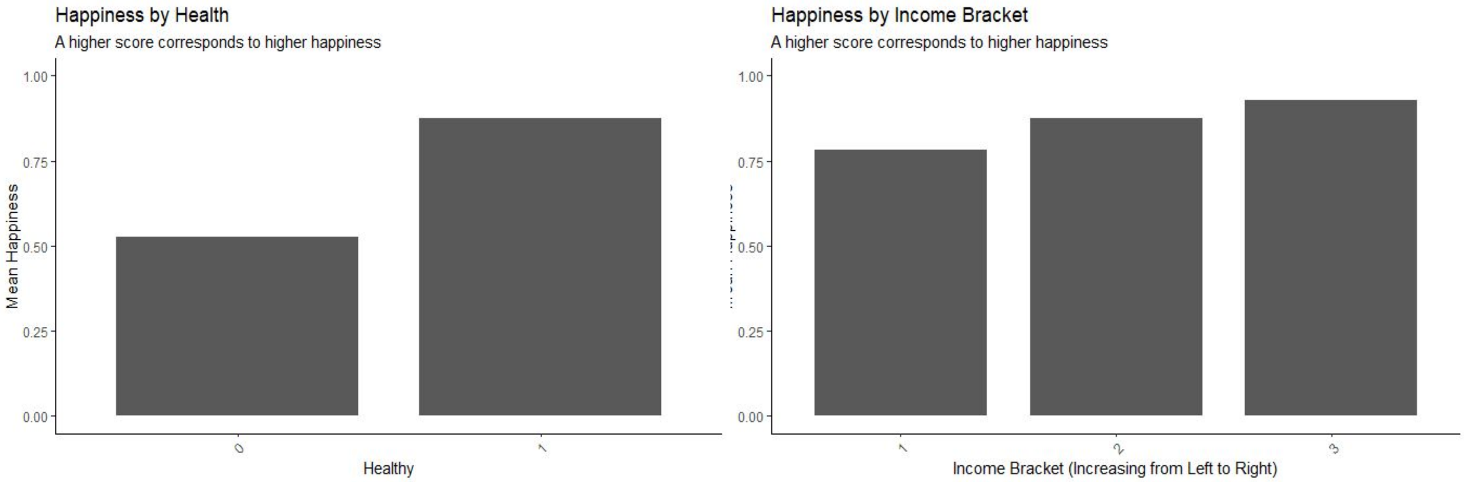


Figure 15: Mean Happiness for Different Health and Income Brackets

Several other predictors were also deemed significant both before and after controlling for country: occupation, relationship status, and education. Similar to happiness and income, these coefficients mostly did not change after controlling for country effects, implying a universal effect across different countries:

Variable (see Appendix naming keys)	Coefficient before controlling for country effects	Coefficient after controlling for country effects
occupation8	-0.21	-0.22
relationship3	-0.58	-0.65
relationship5	-0.44	-0.44
education2	0.14	0.10

Table 16: Coefficients of Several Variables Before and After Adjusting for Country Effects (see Appendix for full table).

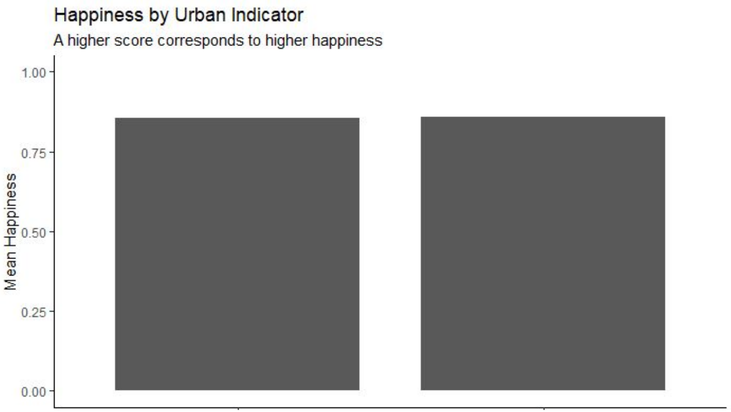
However, these coefficients were not as large as the ones from health and income, which suggests a less pronounced average effect on happiness (see Appendix for graphs). This observation, along with the understanding that skilled jobs, higher educational attainment, and relationship stability are highly correlated with higher income and better health leads us to not consider these factors as good predictors for happiness.

Finally, we consider other variables we did not priorly link to happiness: religion, urban environment, size of the family, and immigrant status. Most of these predictors were deemed significant before including country effects and insignificant afterwards. These variables are informative likely not by themselves but through the residence country of the sample that they signal. For example, immigration signals residence in a developed country (since emigration to the developing world is rare) which associates with higher happiness levels (significant positive coefficient of 0.21).

Variable **(see Appendix naming keys)**	**p-value before controlling for country effects**	**p-value after controlling for country effects**
urban1	1.42e-05	0.002
immigrant1	0.002	0.19
household_size	0.0001	0.69
religion1	0.002	0.02
religion2	5.82e-08	0.90
religion3	<<1e-15	0.12
religion4	0.001	0.78
religion5	4.66e-08	0.02

Table 17: *p*-values of Several Variables Before and After Adjusting for Country Effects

Given that immigration and household_size variables lose significance after controlling for country effects, we examine urban environment and religion as predictors of happiness. As seen in the below bar plots, the effects of these two variables on happiness are minimal with the notable exception of Jewish people (religion4):



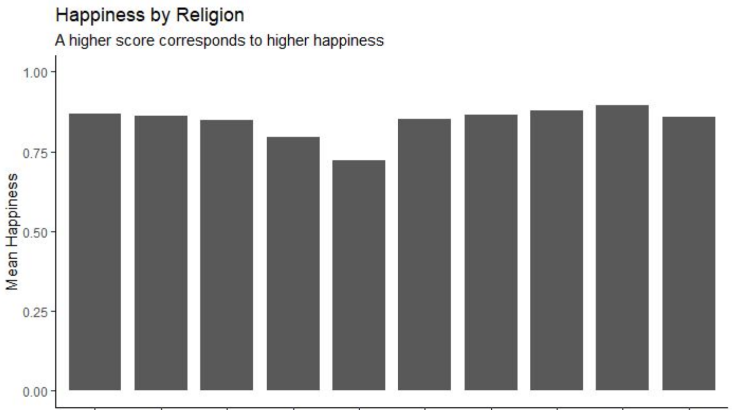


Figure 18: Mean Happiness for Urbanized Status and Religion

However, as we have noted in the table above, the effect of being Jewish by itself (religion4) is not statistically significant in the presence of country effects (p-value=0.78 >> 0.05). Thus, the lower mean happiness corresponding to religion 4 can be best explained by other factors that are correlated with being Jewish rather than being Jewish in itself. Thus, despite the significance of these factors, the small magnitude of the effect on happiness leads us to not consider them as important factors in determining whether somebody is happy.

Discussion

During our analyses, we came across a number of issues that had to be addressed.

First, during the data wrangling process, we encountered the problem of missing data. Deleting the rows with missing entries was rather infeasible since that would remove more than half of our data. Therefore, we used imputation of the mode to reconstruct the missing entries. However, this method introduces a heavy bias in our data due to the fact that the missing data is not random. Despite this bias, from an inferential standpoint, the significance of the predictors in relation to happiness should not be significantly affected.

Second challenge was getting the glmnet models to converge. When run with the default parameters, the model would take very long to run and yet, not converge. One of the main contributing factors to this was that the "glmer" function estimates the correlation matrix simultaneously, and this feature cannot be turned off by assuming a diagonal matrix. The workaround we found was to set the argument "nAGQ=0" which uses a simpler and faster approximation method for the likelihood function which cannot be exactly derived for GLMMs.[3] Although setting "nAGQ=0" results in less accurate approximations than the default Laplace approximation (nAGQ=1), in practice, this difference is rather small. A tougher challenge was adding random slopes to the mixed effects model. Even with ample time to run, the model failed to converge with a single random slope added to the model. We tried shrinking the dataset or using different variables as random slopes, but none solved the issue. Therefore, we could only account for the international differences just with a random intercept. Additionally, when trying to account for interaction terms, the models never converged under any of the aforementioned fixes, so the interaction-based model was abandoned.

Lastly, about 20% of our dataset comes from 2020 which we think does not accurately represent the true relationship between the given predictors and happiness due to the global COVID-19 pandemic. However, the mean happiness score was rather close to the previous years which we suspect is due to the fact that a large portion of the data was gathered prior to the pandemic. This helps alleviate the outlier issue with 2020 data.

Conclusion

[3] https://www.rdocumentation.org/packages/lme4/versions/1.1-25/topics/glmer

Given the above analysis, we can conclude that the country of residence highly affects a person's happiness levels. Adjusting for country effects, the most important factors in determining happiness are income and health. Not only are these factors significant in all models we considered, but their average effect on happiness were very large in comparison to other significant factors. Other factors such as education and occupation are also significant even after controlling for the country of the surveyee. However, their effect on happiness is not very pronounced compared to that of income and health. Thus, our prior assumptions about high income, high education, and good health being associated with a significant increase in happiness are correct. Other variables were either insignificant after adjusting for country or significant but with negligible association with happiness. From these findings, we see that the key to happiness doesn't come from complex interactions among society, culture, and beliefs but from being healthy, educated, and having high income. Stay happy!

Appendix

Table X: The variables in our final dataset, where asterisk (*) means we converted to indicator variable

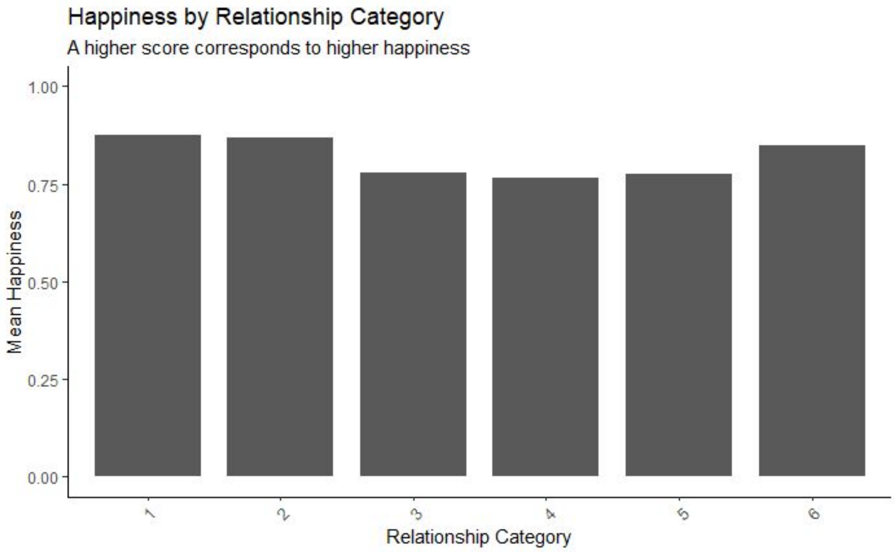
Our Variable	WVS Variable	Description
year	a_year	Year of survey (2017 - 2020)
country	b_country_alpha	ISO 3166-1 alpha-3 country code
town_size	g_townsize2	Town size (1 = under 5000, 2 = 5000 - 20000, 3 = 20000 - 100000, 4 = 100000 - 500000, 5 = 500000 and more)
settlement_type	h_settlement	Settlement type (1 = capital city, 2 = regional center, 3 = district center, 4 = another city/town that isn't a regional or district center, 5 = village)
urban*	h_urbrural	Urban (1 = yes, 0 = no)

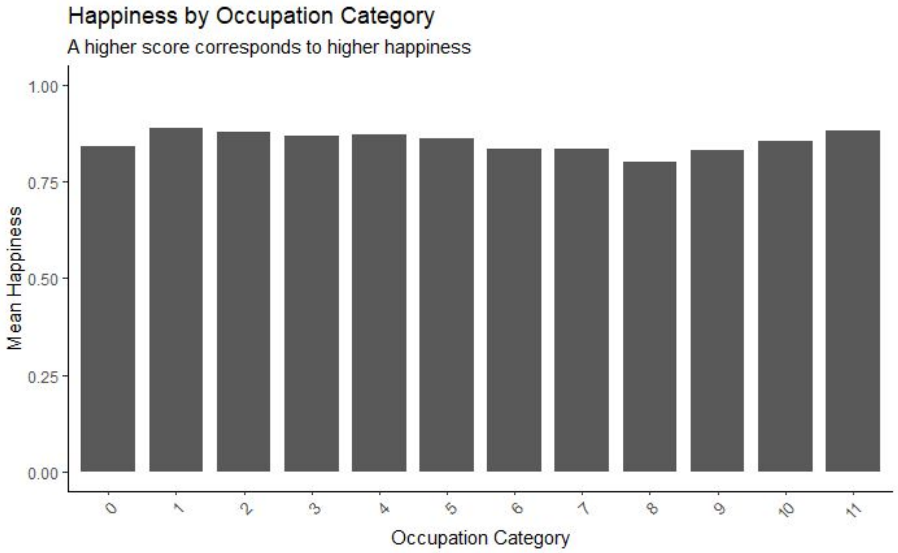
happy	q46	Happiness (1 = very happy, 2 = quite happy, 3 = not very happy, 4 = not at all happy)
healthy*	q47	Healthy (1 = yes, 0 = no)
church_member*	q94	Church member (1 = yes, 0 = no)
sport_member*	q95	Sports or recreational organisation member (1 = yes, 0 = no)
arts_member*	q96	Art, music, or educational organisation member (1 = yes, 0 = no)
union_member*	q97	Labour union member (1 = yes, 0 = no)
political_party_member*	q98	Political party member (1 = yes, 0 = no)
environment_member*	q99	Environmental organisation member (1 = yes, 0 = no)
prof_member*	q100	Professional organisation member (1 = yes, 0 = no)
charity_member*	q101	Charitable/humanitarian organisation member (1 = yes, 0 = no)
consumer_member*	q102	Consumer organisation member (1 = yes, 0 = no)
self_help_member*	q103	Self-help group member (1 = yes, 0 = no)
women_member*	q104	Women's group member (1 = yes, 0 = no)
religious*	q173	Religious (1 = yes, 0 = no)
political_1	q209	Signing a petition (2 = have done, 1 = might do, 0 = would never do)
political_2	q210	Joining in boycotts (2 = have done, 1 = might do, 0 = would never do)
political_3	q211	Attending lawful/peaceful demonstrations (2 = have done, 1 = might do, 0 = would never do)
political_4	q212	Joining unofficial strikes (2 = have done, 1 = might do, 0 = would never do)
political_5	q221	Vote in local elections (2 = always, 1 = usually, 0 = never)

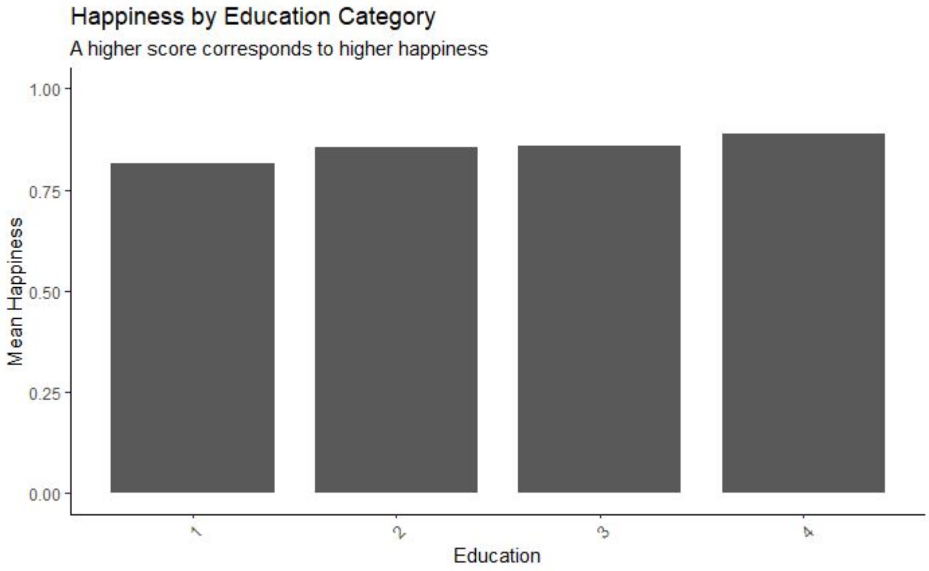
political_6	q222	Vote in national elections (2 = always, 1 = usually, 0 = never)
male*	q260	Sex (1 = male, 0 = female)
age	q262	Age
immigrant*	q263	Immigration status (1 = immigrant, 0 = not an immigrant)
immigrant_mother*	q264	Mother's immigration status (1 = immigrant, 0 = not an immigrant)
immigrant_father*	q265	Father's immigration status (1 = immigrant, 0 = not an immigrant)
citizen*	q269	Citizenship status (1 = citizen, 0 = not a citizen)
household_size	q270	Number of people in household
live_with_parents*	q271	Live with parents (1 = yes, 2 = no)
relationship	q273	Marital status (1 = married, 2 = living together as married, 3 = divorced, 4 = separated, 5 = widowed, 6 = single)
num_kids	q274	Number of children
education	q275r	Highest educational level (1 = primary, 2 = secondary, 3 = post-secondary, 4 = tertiary)
education_mother	q277r	Mother's highest educational level (1 = primary, 2 = secondary, 3 = post-secondary, 4 = tertiary)
education_father	q278r	Father's highest educational level (1 = primary, 2 = secondary, 3 = post-secondary, 4 = tertiary)
employment_status	q279	Employment status (1 = full time, 2 = part time, 3 = self-employed, 4 = retired/pensioned, 5 = housewife not otherwise employed, 6 = students, 7 = unemployed, 8 = other)
occupation	q281	Occupational group (0 = never had a job, 2 = professional and technical, 3 = higher administrative, 4 = clerical, 5 = service, 6 = skilled worker, 7 = semi-skilled worker, 8 = unskilled

		worker, 9 = farm worker, 10 = farm owner/farm manager, 11 = other)
sector	q284	Sector of employment (1 = government or public institution, 2 = private business or industry, 3 = private non-profit organization)
breadwinner	q285	Chief wage earner (1 = yes, 2 = no)
income	q288r	Income level within country (1 = low, 2 = medium, 3 = high)
religion	q289	Religious denomination (0 = do not belong to a denomination, 1 = Roman Catholic, 2 = Protestant, 3 = Orthodox, 4 = Jew, 5 = Muslim, 6 = Hindu, 7 = Buddhist, 8 = Other Christian, 9 = other)

Additional Graphs and Tables







Variable **(see Appendix naming keys)**	**Coefficient before controlling for country effects**	**Coefficient after controlling for country effects**
relationship2	-0.17	-0.18
relationship3	-0.58	-0.65
relationship4	-0.76	-0.80
relationship5	-0.44	-0.44

relationship6	-0.29	-0.25
education2	0.14	0.10
education3	0.11	0.13
education4	0.13	0.19
occupation1	0.14*	0.016
occupation2	-0.077	-0.081
occupation3	-0.10	-0.074
occupation4	-0.0089*	-0.084
occupation5	-0.015	-0.075
occupation6	-0.22	-0.18
occupation7	-0.19	-0.22
occupation8	-0.21	-0.23
occupation9	-0.12*	-0.25
occupation10	-0.011*	-0.15
occupation11	0.29*	-0.31

* Note that these coefficients were not significant in either case, so having large coefficient differences between the models does not incur additional investigation

R Code:

The R code for data wrangling, EDA, model building, and visualizations are found at this link:

https://github.com/dashamet/stat139-project