Happiness Rate Analysis. What makes you happy?

By: Solomiya Shuptar

General Idea

There are over 7 billions people on the planet now. No doubt that each one of us has their own story, each one of us lives a different life, but there is still at least one common thing between us. We all fall asleep with the same wish: to be happy.

National happiness determines a lot of things in the country and is determined by lots of factors. There exists a general perception that Ukrainians consider themselves as a constantly unhappy nation and generally are extremely unsatisfied with the things around. The idea of my project was to explore the dependencies between happiness rate of countries and different factors that actually influence them.

1. Data Description and Collection

For my research, I used mainly three sources of the data. The first one is the regular datasets available on happiness rate over countries. The datasets are in the form of cross-sectional data which contains standard factors including the happiness rate evaluations: GDP, Family, Health, Freedom, Trust, Generosity. Other data which I included as factors into my research were taken from UNESCO Institute for Statistics (UIS), the official and trusted source of internationally-comparable data on education, science, culture and communication. Specifically I used data on enrollment in education(primary, secondary, tertiary) and science development. Average income data I took from OECD data source.

2. Premiliar Data Filtering and Exploration

Firstly, I went through regular data filtering, renaming columns and making the data more comfortable to use. Fortunately, my data was quite accurate, it didn't contain any missing values or dramatic outliers, so I got down to exploration. For premiliar analysis and finding simple trends I took 2015 as a base year and started working with it.

Correlation matrices

For finding dependencies between variables, I built multiple correlation matrices and performed cor. tests, where H_0 hypothesis states that variables are independent and alternative H_1 hypothesis states that there exists a correlation and we reject H_0 hypothesis.

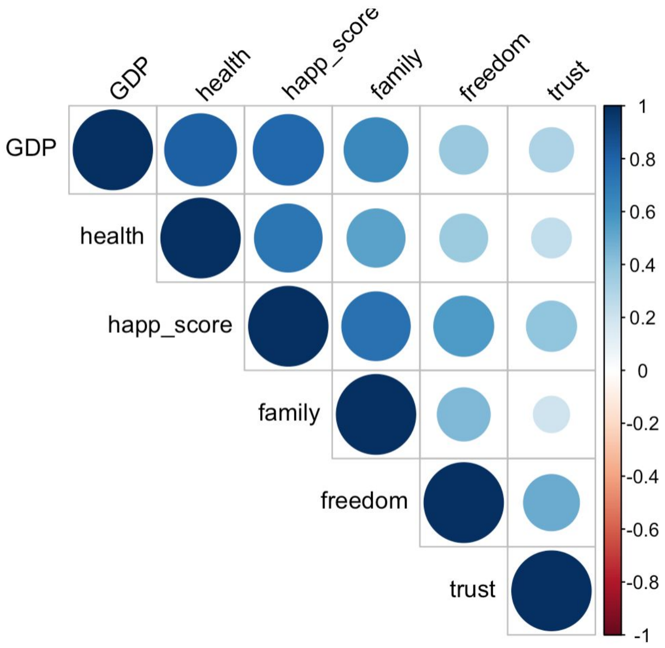
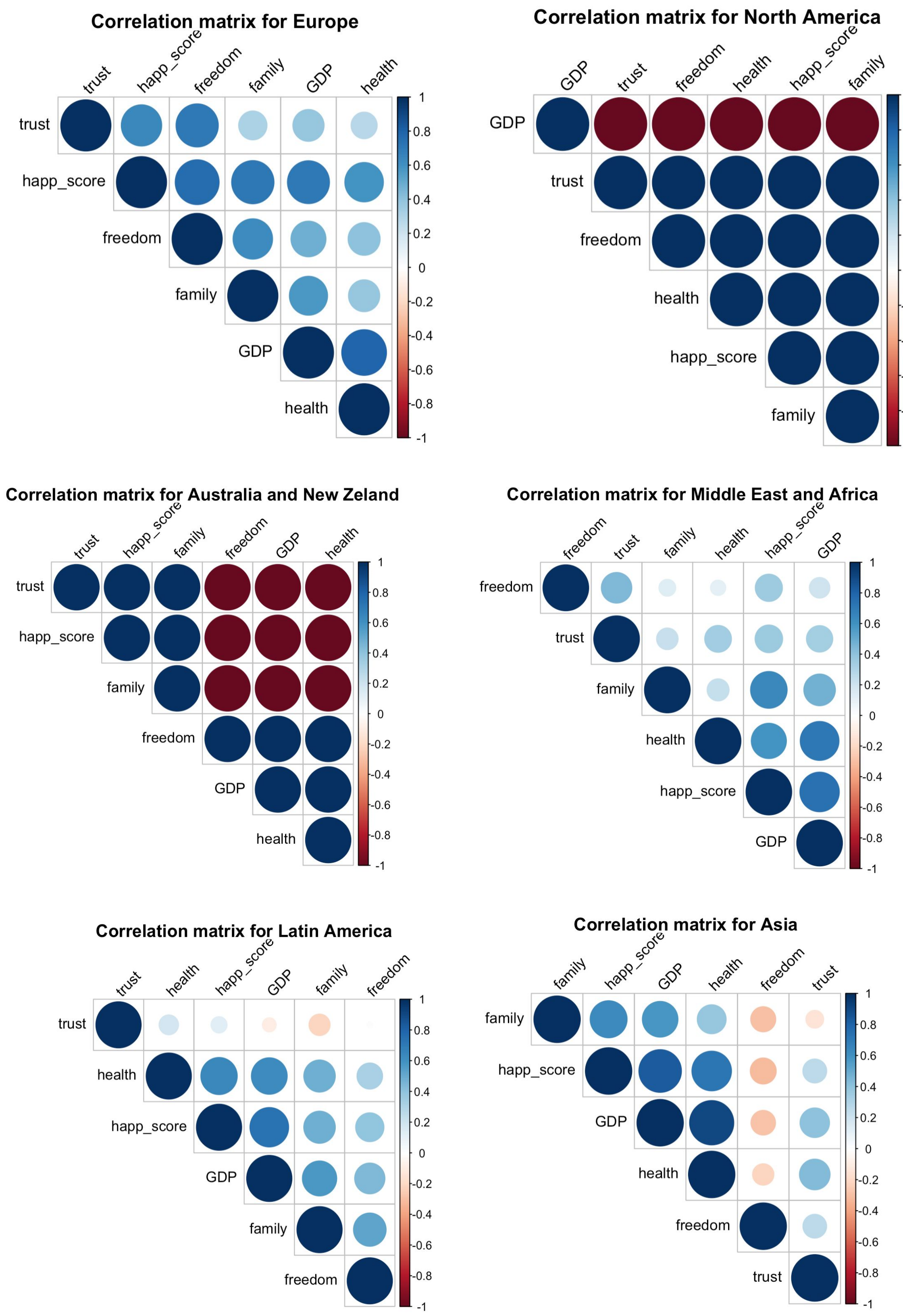


Figure 1. Correlation matrix. Full sample. 2015

I point to the strong correlation between Happiness Score and Family, Health, GDP and a little less correlation with Freedom. Let's see what happens if we cluster data by regions. Interesting thing is that in North America, Australia and New Zealand we see extreme values which indicate that the situation in those countries is somewhat different during the years considered.

Random Forest Algorithm

For further feature selection, I used the Random Forest algorithm. Accuracy of the model is 77% and it showed that all the features are statistically significant except trust, which was in the last place and appeared to be the least significant. Thus, I decided not to use it further.



Correlation matrix for Europe

Correlation matrix for North America

Correlation matrix for Australia and New Zeland

Correlation matrix for Middle East and Africa

Correlation matrix for Latin America

Correlation matrix for Asia

Figure 2. Correlation matrix. Sample clustered by Regions. 2015

Data Visualization

As expected, Figure 3 shows a positive relationship between different variables and happiness score. In the next step, I formally investigate statistical significance of these relationships by using the regression analysis.

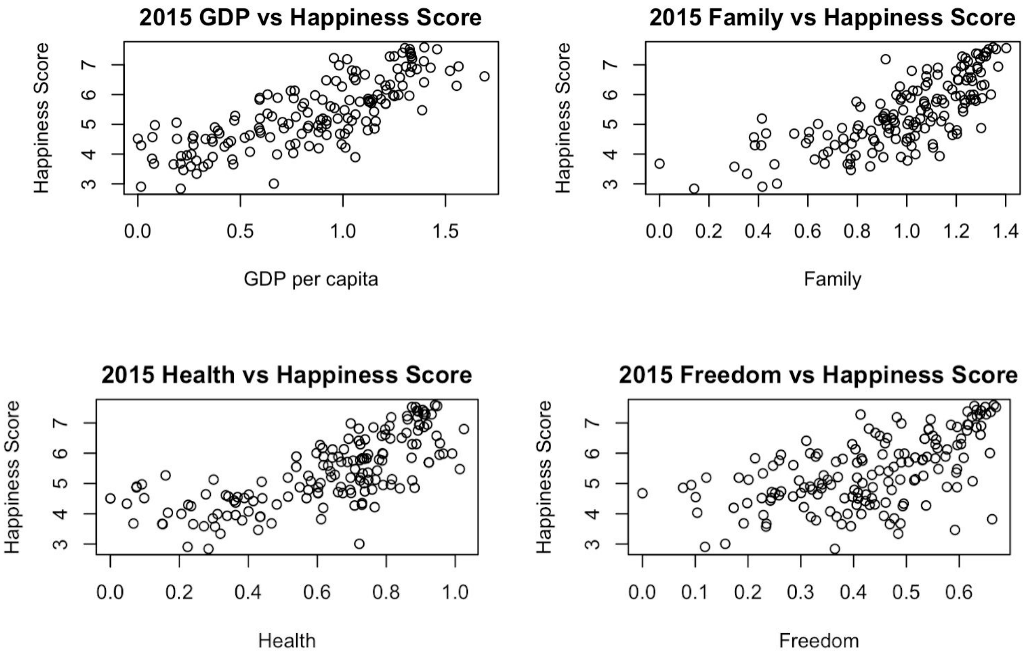


Figure 3. Plots of features. 2015

3. Time-Series Evolution and Extreme Cases Exploration

For the visualization of the time-series evolution from every set of data, I pick top countries (those having their happiness scores over 7.0 during the years considered). Figure 4 shows how happiness rate changes over the years 2015 to 2019. Here, I examine the most interesting cases separately. Figure 4 shows a big decline in happiness score in Austria in 2017 that was followed by a dramatic increase. Also, the United States has completely missing data in the time range between 2016 and 2018 which means that its value dropped under 7.0.

After plotting the curves of Happiness Score and main regressors, I observe that GDP and happiness score are not perfectly correlated. Also, GDP does not always explain the happiness rate in the country. While health almost always has the same curve as happiness rate. In the next steps, I add some factors into the model, which describe the cultural development of the country, and see if they improve it.

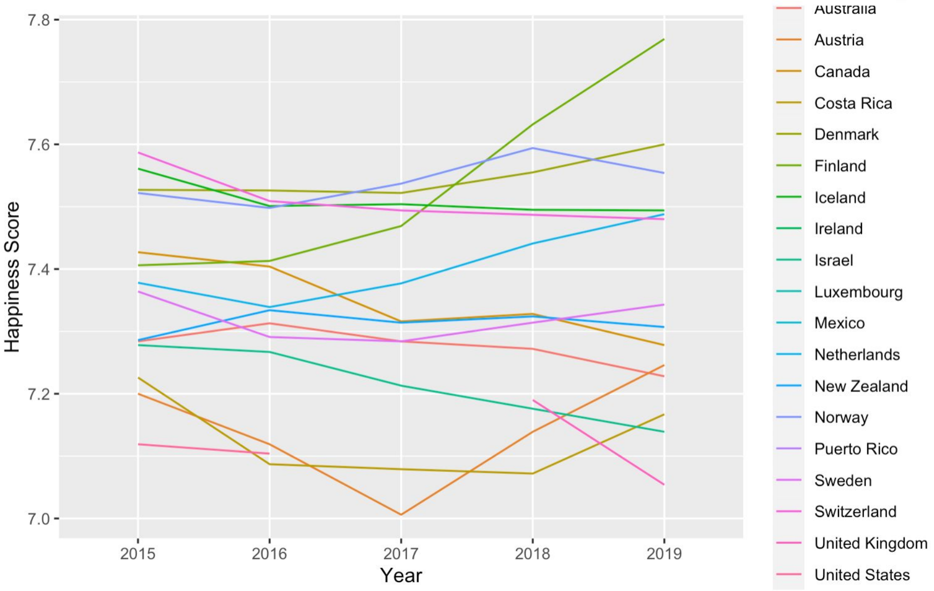
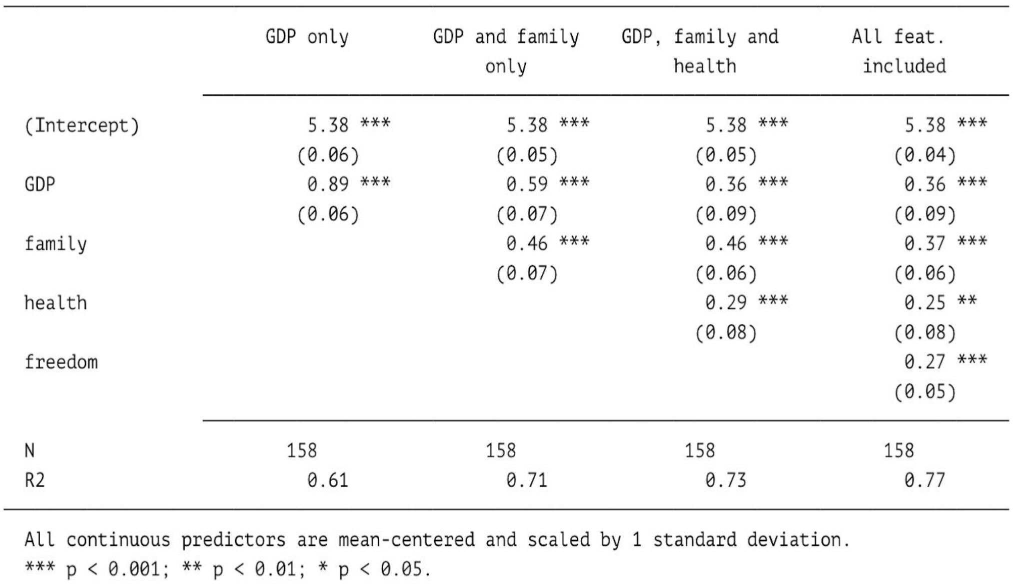


Figure 4. Time-Series Evolution of Happiness Score

4. Preliminary Results

Linear Models (Base year 2015)

Summary below describes linear models during the data for year 2015. The regressions include following features added one by one: GDP, family, health, and freedom.



	GDP only	GDP and family only	GDP, family and health	All feat. included
(Intercept)	5.38 *** (0.06)	5.38 *** (0.05)	5.38 *** (0.05)	5.38 *** (0.04)
GDP	0.89 *** (0.06)	0.59 *** (0.07)	0.36 *** (0.09)	0.36 *** (0.09)
family		0.46 *** (0.07)	0.46 *** (0.06)	0.37 *** (0.06)
health			0.29 *** (0.08)	0.25 ** (0.08)
freedom				0.27 *** (0.05)
N	158	158	158	158
R2	0.61	0.71	0.73	0.77

All continuous predictors are mean-centered and scaled by 1 standard deviation.
*** p < 0.001; ** p < 0.01; * p < 0.05.

Figure 5. Summary Statistics on simple linear models

We observe the features that we have selected are strongly significant based on the p-values. (notice that the values in brackets are standard errors). The plot with the residuals demonstrates that the residuals are distributed more or less equally around the mean and we do not observe a trend in the variance. Therefore, heteroscedasticity is not present in the regressions. Overall, summary of linear models confirm our expectations about the impact and further provide details of significance levels of the variables. In particular, adding them one by one improves the fit of the model as measured by higher R2's. Also, when freedom is added to the model, health becomes a little less significant.

However, using this simple linear model I assume that observations are not related with each other within certain groups, that is, they are independent observations across different countries. To control for the region-specific effects, I consider the regression of the happiness score on the predictive variables above allowing for clusters within the geographical regions. Indeed, previously we observed that correlation matrices of features vary from region to region. Therefore, I also report the results of linear regressions on the data clustered by region. This means that we assume independence across clusters but allow for correlation within clusters. After we clustered the data, standard errors of most estimates became lower, whereas standard errors of the coefficient in front of freedom increased.

5. Additional Features

In what follows, I add the feature which describes enrollment in education and merge it with the population data by country to obtain the ratio value. I replicate the same regressions with the extended data sample by reporting standard and clustered standard errors. Finally, I augment the data with the features contained in the Science Development Dataset, which also was transformed to the ratio value (merged with population).

6. Main Results

In this section, I describe the main results of the analysis including the robustness checks. In some specifications, I add regional dummies as factors to capture the region fixed effects. The summary below presents the description of the models and a short summary of the results.

Model with standard features given in the original dataset

It showed basic performance and accuracy (Adjusted $R^2 \sim 77\%$). Also, regressors are poorly significant.

Full model with education and science development features

The accuracy of the model increases dramatically (Adjusted $R^2 \sim 86\%$). We see better significance of the regional factors. Meanwhile, GDP, freedom and enrollment in secondary education appear to be the most signficant.

Standard coefficient significance test vs robust significance test, clustered by group

Compared to standard coefficient significance test and test, cluster by group, we see heterogeneous changes in t-statistics and p-values. For instance, some of the variables increase in significance (tertiary education, Eastern Asia Region), whereas a Western Europe factor becomes less significant. The remaining regressors including GDP, Freedom, and Secondary Education have the same results.

FE and RE model and Hausman test

The fixed effects model assumes that each group has an unobserved group-specific component. Furthermore, the model allows these unobservable effects to be correlated with the regressors. The RE model assumes this correlation is zero. Overall, the results of the FE model in our case indicate less accuracy compared to the RE specification.

Adding income as a regressor

GDP feature is volatile as a regressor in different models. Furthermore, GDP might not be an accurate measure of people's financial wellbeing. Thus, I include the average income as a proxy of the individual wealth in each country instead of GDP. I retrieved a new dataset, merged it and included it into a pooled dataset.

Full model with income instead of GDP: FE and RE specifications

Performing the panel regressions, RE and FE models give us better results: an increase in accuracy (from 86% to 92%) and better significance of the regressors. Next step I analyze the RE and FE model, and again the FE model gives relatively worse accuracy compared to the RE model.

Hausman test and LM test

I perform a Hausman test where the null hypothesis is that the preferred model is random effects vs. the alternative one is the fixed effects specification. A p-value of the test is 0.1162 and hence the RE model can be rejected in favor of FE. The final testing step is performing an LM (Breusch-Pagan Lagrange multiplier) test to decide between RE regression and simple OLS regression. P-value is < 0.05, so our RE model is preferred over a simple OLS model.