Bayesian Machine Learning: Assessed Coursework – Project

WHICH FACTORS INCREASE LIFE-SATISFACTION?

Pegah Maham

May 9th, 2019

Table of Contents

Introduction

For policy makers, psychologists or social scientists, the precondition of being able to increase life satisfaction amongst a population or individuals, is the knowledge on what influences it. Since decades, data on subjective life-satisfaction and related well-being measures have been collected through polls and analysed (Diener & Suh 1997, Oishi et al., 2009). Measures of subjective well-being can be obtained by a self-report on a numerical scale (Diener et al., 2002).

Besides the empirical analyses, theoretical frameworks about life-satisfaction have been developed. Maslow (1943) proposes a classification hierarchy of human needs which is known as the *Maslow's hierarchy of needs*. Based on this classification hierarchy, Maslow et al. (1970) introduces a need-gratification theory of life satisfaction. He suggests that someone's life satisfaction correlates with the degree of the fulfilment of the personalized needs. The satisfaction of these needs differs across nations, as for example people in less wealthy nations lack financial security more often than people in wealthier societies. Subsequently, life satisfaction has been found to be varying across nations along with the fulfilment of these needs (Veenhoven, 1991; Diener & Diener, 2009; Oishi et al., 2009).

These global comparisons are important for international development policies and inter-cultural research. However, for national policies and domestic insights, life-satisfaction and correlating factors within a country are of interest. In this study, predictive factors on life-satisfaction for the contemporary German population are analysed. Moreover, the sample is clustered into two subgroups which are described and on which a more customized life-satisfaction analysis is conducted.

Data Description

To find relevant factors to life-satisfaction, survey data from Germany from 2013 is analysed. The survey was conducted as a part of the the World Value Survey which is a longitudinal multi-national survey research project, started in 1981 (Inglehart, 2014). In the analysed survey, around 2000 randomly selected participants were interviewed on questions about politics, moral values, traditions, religion and their state of life. Overall, around two hundred questions are asked. One of the questions is on personal life-satisfaction, where the answer could be given on a scale from 1 to 10, where 10 is the highest self-reported life-satisfaction. The distribution of this variable is shown below (Figure 1). As one can see, most people are self-reporting numbers from 6 to 9.

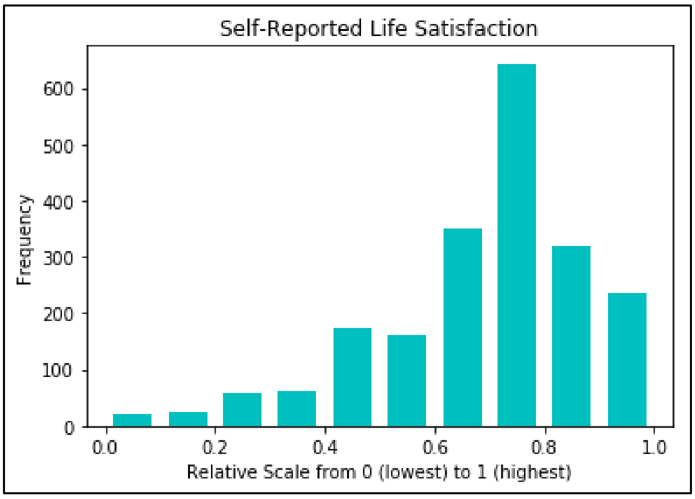

Figure 1 Distribution of the variable of interest. The x-axis is relative.

Life-Satisfaction Analysis

The goal of this analysis is to find and select the most important factors that are correlated to self-reported life-satisfaction. Potential factors are all matters that are captured in the survey questions.

Methods

The first approach to answer the question is linear regression. Linear regression is a method to analyse the relation between two variables while controlling for the influence of other variables simultaneously. Its goal is to find variables than increase or decrease, and thus correlate, with a target variable due to underlying mechanisms.

The control for other variables is necessary because variables can be correlated through other variables. For example, when people eat more ice cream, more shark attacks occur. When this relation is controlled for the season, the effect vanishes, since the correlation between the first two variables can be fully explained by the season of the year: During summer, people eat more ice and swim in the ocean. The ice cream has no effect on sharks' attacks on its own. To decrease the likelihood of falsely associating two variables, it is beneficial to include as many related variables as possible.

Unfortunately, too many variables, given a limited amount of observations, introduce some challenges. These challenges are related to the ratio of the number of observations and the number of variables (also named features). Data can be classified into *low-dimensional*, when the number of observations is much greater than the number of features and *high-dimensional* when the number of features exceeds the number of observations. In cases where the data is not low dimensional, the problem of "overfitting" can occur. Overfitting describes the case when a model that is supposed to be able to make general statements about the population, is only making statements for the available sample data. The sample data is used as input to detect patterns. Some patterns that do not occur in the population can occur in the sample data by random chance. This can happen more likely when the number of observations is low, because random outliers cancel out with a higher probability if the number of observations is high enough. Rolling a dice three times, it is imaginable to see three “sixes”. This is a rate of 100%. When rolling a dice 1000 times, it extremely unlikely that 100% of the outcomes are a “six”.

When generalizing random fluctuations in the sample data as general patterns, the data gets misinterpreted. To cope with this problem, it has been shown that *shrinkage* and *variable selection* methods are of advantage to prevent overfitting (James et al., 2013). Shrinkage methods, as the name indicates, shrink the parameters of the model. Here, the so-called “ridge” regression method is used. Variable selection methods only use a subset of the available variables and thus decrease the dimensionality of the data. Here, the “lasso” regression method is used. How many variables are selected in the lasso regression depends on a parameter named *alpha*. The higher this parameter, the stronger the selection.

To evaluate how well a model is able to generalize patterns and use insights from the input data to predict attributes of new data, an artificial new “test” dataset is created. Test data sets are not used while creating a prediction model. This leads to a trade-off. When too many data points are used as a test set, the number of observations for the model building decreases and makes it harder to identify patterns. On the other hand the test set should be representative enough to be able to evaluate the model. Here, 20% of the data is used as a test set. The remaining 80% are the so-called “training set”.

One common way to evaluate the model with the help of the test set, is to measure the mean squared error (MSE) that the model produces when it tries to predict the target variable on the test set. The MSE is the average of the squared error between prediction and true value. The error is being squared to prevent errors with opposing signs to cancel out each other. The higher the absolute errors, the higher the MSE.

Depending on the context, other measurements are of importance too. If only the accuracy of the prediction is relevant, the MSE on multiple rotating test and train sets can be sufficient. In contexts, where the interpretability of the model is relevant, the number of variables that are selected in the model becomes crucial: A model with too many variables is cumbersome to interpret.

Results

The linear regression yields to a model with an MSE of around 0.16. This method uses all 265 variables to create a model for predicting and explaining the target variable.

To make this model more comprehensible, a subset of variables that are highly significant is filtered out and presented. When interpreting any statistical model with coefficients, two observations are important for each coefficient estimate. The *direction* and *size* describe the type (negative or positive) and the strength of the correlation between the corresponding variable and the target variable (column *"coef"*). Secondly, the *significance* of the coefficient is important (column *"P>|t|"*). It captures how likely it is, that the effect is not due to random noise but rather due to an underlying correlation. The lower the significance level, the less likely it is that in reality the effect of the variable on the target is equal to zero. Below, a list of all variables with p-values under 0.1% and their coefficients is presented (Table 1).

Table 1 Results of the linear regression on the full data set.

	coef	std err	t	P>|t|	[0.025	0.975]
V10: Feeling of happiness	-0.3646	0.022	-16.553	0.000	-0.408	-0.321
V11: State of health (subjective)	-0.0677	0.017	-4.037	0.000	-0.101	-0.035
V26: Active/Inactive membership: Sport or recreational organization	0.0252	0.010	2.601	0.009	0.006	0.044
V36: Would not like to have as neighbors: Drug addicts	0.0272	0.010	2.810	0.005	0.008	0.046
V55: How much freedom of choice and control over own life	0.0936	0.019	5.012	0.000	0.057	0.130
V56: Do you think most people would try to take advantage of you if they got a chance, or would they try to be fair?	0.0547	0.020	2.732	0.006	0.015	0.094
V59: Satisfaction with financial situation of household	0.1501	0.019	7.943	0.000	0.113	0.187
V160D: I see myself as someone who: is relaxed, handles stress well	0.0745	0.018	4.218	0.000	0.040	0.109
V190: In the last 12 month, how often have you or your family: Gone without needed medicine or treatment that you needed	0.0920	0.033	2.757	0.006	0.027	0.157
V203: Justifiable: Homosexuality	0.0522	0.016	3.226	0.001	0.020	0.084
V213: I see myself as part of my local community	-0.0474	0.017	-2.720	0.007	-0.082	-0.013
V216: I see myself as an autonomous individual	0.0429	0.015	2.832	0.005	0.013	0.073
V217: Information source: Daily newspaper	0.0365	0.014	2.657	0.008	0.010	0.063

The largest coefficient is corresponding to the self-reported *happiness* variable with -0.36. This is not surprising, since these two questions ask for very similar things. The happier people describe themselves, the lower the number on this variable. The life-satisfaction variable is coded vice versa. This explains the negative coefficient for this question. The happier people are, the higher the life-satisfaction. The same applies for the negative coefficient and interpretation on the *state of health* and the identification with a *local community*.

This variable is followed by the satisfaction of the *financial situation* with a coefficient of 0.15. With a value of 0.09, the third most explaining predictor is the amount of *freedom of choice and life control*. Close to that are answers indicating necessary *medical treatment* can be afforded which is similar to financial safety. With a value of 0.07 the answer to the question on how well *stress* is handled, seems to play a role in life-satisfaction as well. Further six variables have rather small coefficients, while being significant and can be obtained from the list.

Using ridge regression, the mean squared error remains approximately the same (0.16). Since neither the MSE is decreasing, nor are variables selected, the results are not displayed here, as they are not superior to the linear regression.

The lasso regression decreases the MSE slightly to 0.15% and helps selecting variables. The latter is especially valuable because it improves the interpretability of the analysis. How many variables are selected, depends on the *alpha* parameter. Table 2 shows how the MSE changes with an increasing alpha, and the respective number of selected parameters. As one can see, the MSE does not change strongly and remains in the same order while having the lowest MSE values at alpha levels of around 0.002 with around 50 variables. Again, the focus lies in interpretability, the model selection is strongly determined by the reduction of variables.

Table 2 MSE values and number of coefficients for different alpha values of the lasso regression.

MSE	Number of Coefficients	Alpha
0.165434	237	0.0001
0.156983	68	0.0012
0.157440	34	0.0023
0.159182	21	0.0034
0.161324	17	0.0045
0.163684	11	0.0056
0.166225	10	0.0067
0.168827	6	0.0078
0.171406	5	0.0089
0.173928	5	0.0100

As a trade-off between low MSE and a reduced number of variables an alpha value of 0.067 is chosen which leads to a model with ten features. Below the output of this lasso regression is shown (Table 3). Running a linear regression on the lasso-selected variables only, some turned out to be not or less significant. This includes the *Marital Status* and the *State of Health*.

The four variables of *Happiness, Freedom of Choice, Financial Satisfaction* and *Stress Resilience* show a strong relation on life-satisfaction and are highly significant, as seen in the linear regression. The effect of the next strongest variable is only around 25% as large, compared to the former four. This makes these four variables stick out in the analysis.

Table 3 Results of the lasso regression.

	coef	std err	t	P>|t|	[0.025	0.975]
V10: Feeling of happiness	-0.1848	0.023	-8.059	0.000	-0.230	-0.140
V11: State of health (subjective)	0.0296	0.016	1.793	0.073	-0.003	0.062
V24: Most people can be trusted	0.0449	0.009	4.878	0.000	0.027	0.063
V26: Active/Inactive membership: Sport or recreational organization	0.0561	0.010	5.362	0.000	0.036	0.077
V55: How much freedom of choice and control over own life	0.3552	0.018	19.741	0.000	0.320	0.390
V57: Marital status	-0.0015	0.011	-0.131	0.896	-0.024	0.021
V59: Satisfaction with financial situation of household	0.4258	0.017	25.144	0.000	0.393	0.459
V160D: I see myself as someone who: is relaxed, handles stress well	0.2859	0.016	17.448	0.000	0.254	0.318
V213: I see myself as part of my local community	0.0446	0.017	2.670	0.008	0.012	0.077
V227: Vote in elections: National level	0.0272	0.013	2.040	0.042	0.001	0.053

Cluster Analysis

Having over two hundred variables, it is difficult to describe the data set in full. Theoretically, one could describe every single of the 2046 surveyed people in each of the questions. This would be both cumbersome and the resulting information hard to interpret. It is helpful to assign the 2046 observations into subgroups that show similarities and can be more easily described. This process of finding similar subgroups is known as *cluster analysis*.

Unfortunately, there is no label assigned to each person regarding their “group”. Moreover, it is not clear what exactly these groups look like. Over 200 questions are asked in the survey. It is not obvious which questions determine group memberships. It is helpful to have groups within which people are similar and between which people differ. The higher the within-similarity and the between-difference the clearer the groups are distinguishable and the more explanatory power is obtained by the clustering.

Methods and Results

There are different methods of finding optimal clusters. Here, two methods are implemented where the number of clusters is defined and one where the method can produce different number of clusters.

The first method to create cluster models is the *Gaussian Mixture Model* (GMM). To evaluate these models under different parameter settings, the *Bayesian Information Criterion* (BIC) is used. Besides the explanatory power of the model, this criterion penalizes additional clusters. This way, it helps to find the optimal trade-off between minimizing the model complexity (in this case the number of clusters) and its explanatory power. The lowest BIC value to cluster the World Value Survey data is found for 19 clusters. Since the clusters are supposed to be used for a more detailed regression analysis, 19 clusters are not practical. The GMM is built with two clusters instead, recognising it results in suboptimal BIC values.

The second clustering approach is the *Bayesian Gaussian Mixture Model* (BGMM). The advantage of this method is that the number of clusters does not have to be determined beforehand. With a tuning parameter, the tendency to reduce the number of clusters can be controlled. Using this method, two clusters are constructed.

While the first two methods assign probabilities for each observation to belong to any cluster, the third method, *K-Means Clustering*, assigns them deterministically to a cluster. The former can be useful to identify observations that are not strongly associated with any cluster. Here, for the purpose of creating clusters, each observation is assigned to the cluster with the highest probability.

Evaluation

For all classifications, the larger cluster is significantly, at least 50%, larger than the smaller cluster. The k-means observations are divided 1200:846, the BGMM divides into 1227:819. The GMM produces the most unbalanced division with 1407:639. As one can see in Table 4, the k-means cluster assignment correlates moderately with both the GMM assignments and the BGMM assignments. The latter two are not correlated with each other.

Table 4 Correlation of the three cluster assignments. The letters at the end stand for the corresponding initials of the method.

	cluster_ID_G	cluster_ID_K	cluster_ID_B
cluster_ID_G	1.000000	0.281017	-0.015525
cluster_ID_K	0.281017	1.000000	-0.228186
cluster_ID_B	-0.015525	-0.228186	1.000000

Ideally, the clusters have high similarity within, and large differences between the clusters. With over two hundred variables, it is a challenge to describe the clusters with regard to every variable. Rather, to obtain a description of the cluster differences, previous findings and theoretical frameworks are used.

Regarding the World Value Survey data, two attributes (called dimensions) have been found that can explain cross-cultural differences to a large amount (Inglehart & Baker, 2000). The first dimension is along "*survival*" and "*self-expression*" values. The second dimension captures "*traditional*" versus "*secular-rational values"*. Each society can be located on a global map (Figure 2) of cross-cultural variation based on these two dimensions (Inglehart 1997, Inglehart & Baker, 2000).

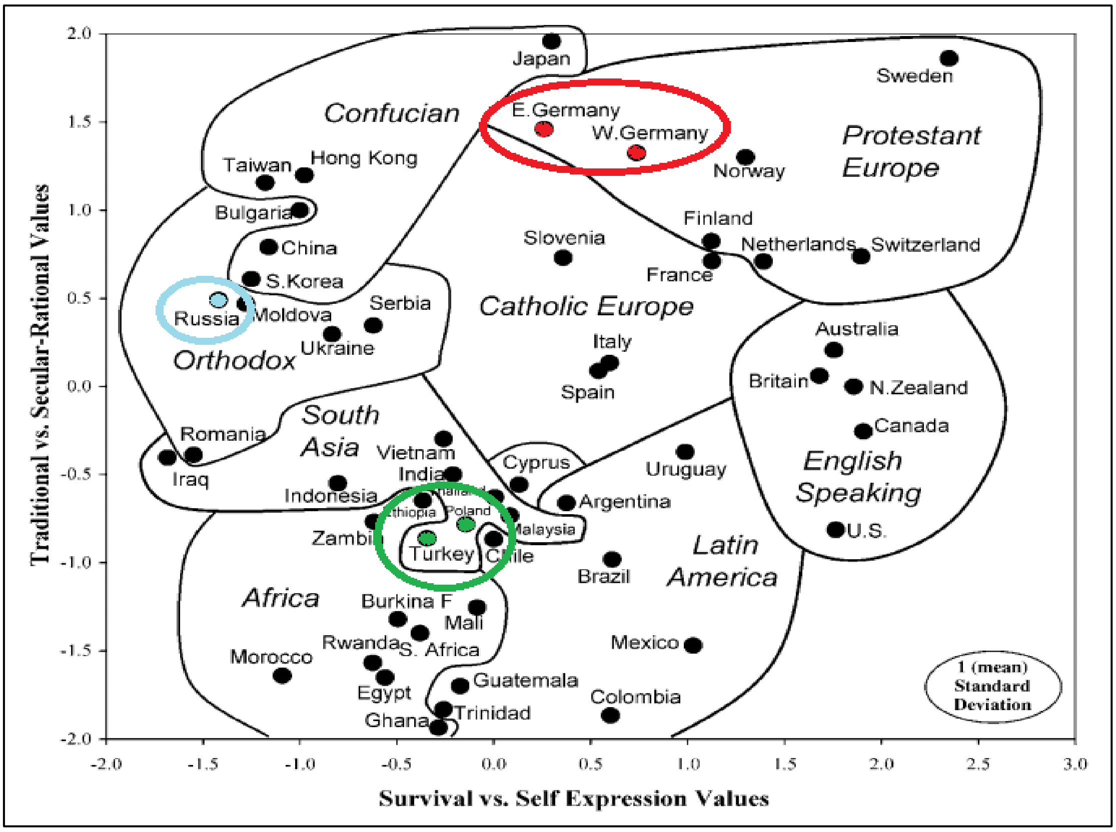

Figure 2 World map (modified, based on Inglehart & Welzel, 2010) on the two axes "survival" vs. "self-expression" and "traditional" vs. "rational-secular". Germany is highlighted in red, Russia in blue und Turkey and Poland in green. Germany's position is high in "secular-rational" values and above average in "self-expression" values.

The survival vs. self-expression dimension describes the polarization between people who are on the one end not financially secured and are less open to new experience and forms of living (immigration, homosexuality), and on the other end people who are financially secured and seek new experience. While the former group is mainly concerned with economic safety, the latter values tolerance and liberty. The secular-rational vs. traditional dimension captures the difference with regard to religiousness and tradition (Inglehart & Baker, 2000).

Overall Germany is located on the more secular and self-expression side, compared to other nations. The two dimensions can be found within societies as well, although cross-national differences are usually larger than within-national differences (Inglehart & Baker, 2000).

In the following, the resulting clusters of each clustering method are compared with regard of their ability to distinguish the observations on the two dimensions. For that reason, nine items of the survey are selected. These questions are suited for the German context and known (question 1-8) to be correlated with the two dimensions (Inglehart & Baker, 2000). Three of them correlate on the secular-rational vs. traditional scale and six correlate on the survival vs. self-expression scale.

Regarding the secular-rational (SR) vs. traditional (T) dimension, the following three variables are selected as indicators. The group which correlates to positive answers is indicated in parentheses.

1. abortion can be justified (SR)
2. politics plays an important role (SR)
3. religion is important (T)

Regarding the survival (S) vs. self-expression (SE), the following six items are examined:

4. rejection of homosexuals as neighbors (S)
5. homosexuality can be justified (SE)
6. job priority for men over women (S)
7. happiness (SE)
8. rejection of neighbors of different religion (S)
9. father immigrant (S)

The immigration background (item 9) is additionally used to the known associated variables. In the German context, most people with a migration background are of Turkish, Polish or Russian descent (German Federal Agency for Civic Education). These are nations with high survival and low self-expression value (see Figure 2). Thus, this variable can be used as a proxy, since it is shown that children are influenced by their parents in their values (Hoge, 1982). Moreover, it is known that both first and second-generation immigrants in Germany are less wealthy than non-immigrants (Algan et al, 2010), which is a further indicator on the survival scale.

For all nine variables, the mean within each cluster is calculated and plotted (Figure 3-5). For the GMM clusters, the direction of all nine differences indicates a correlation of cluster one with higher survival and traditional values (Figure 3). The differences are rather small, especially when compared with the groups created with the two following clustering methods.

Cluster 2, formed by the Bayesian Gaussian Mixture model, shows higher secular-rational values in the first row, with the exception of religiousness (Figure 4). Regarding the self-expression vs. survival scale, the second cluster visibly shows higher values on questions targeting self-expression. Especially striking is the seventh subplot, where one can see that no one in cluster 2 opposes neighbors of a different religion. For homosexual neighbors, cluster 1 opposes multiple times more often.

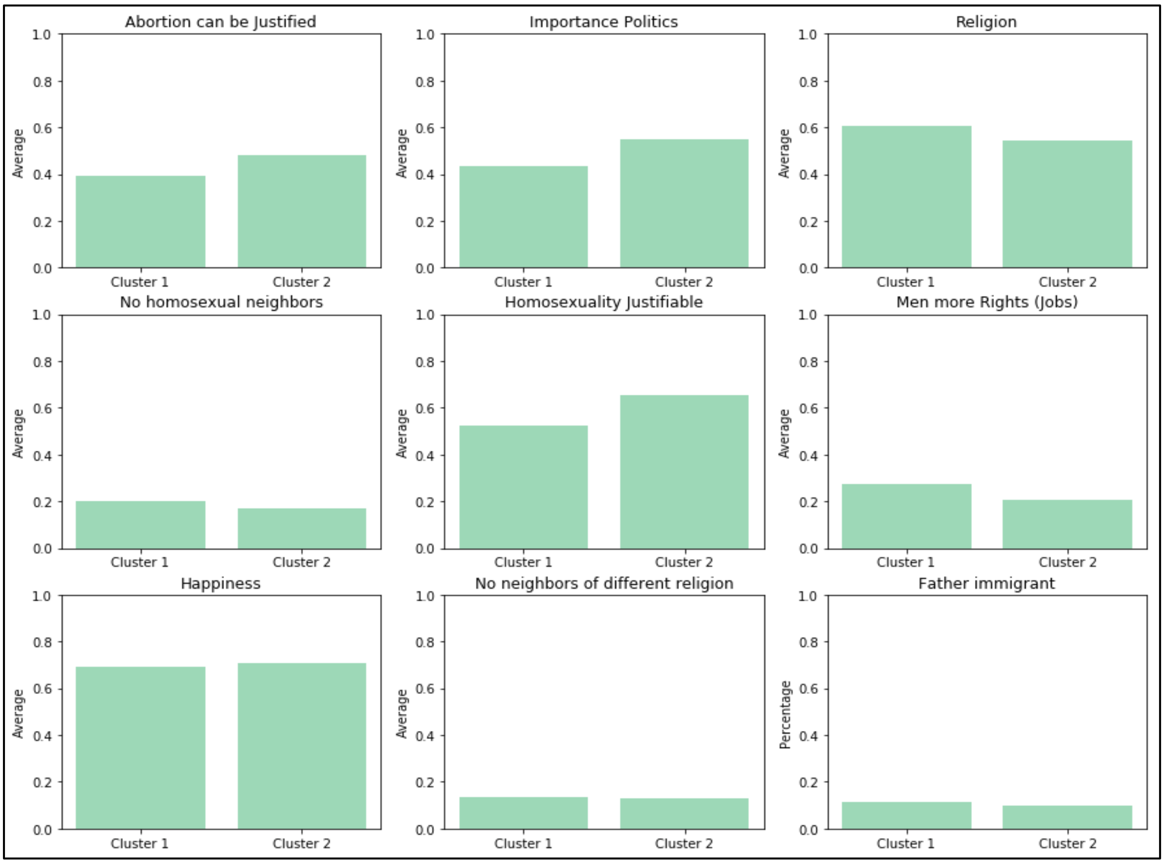

Figure 3 GMM Clusters along nine tested items.

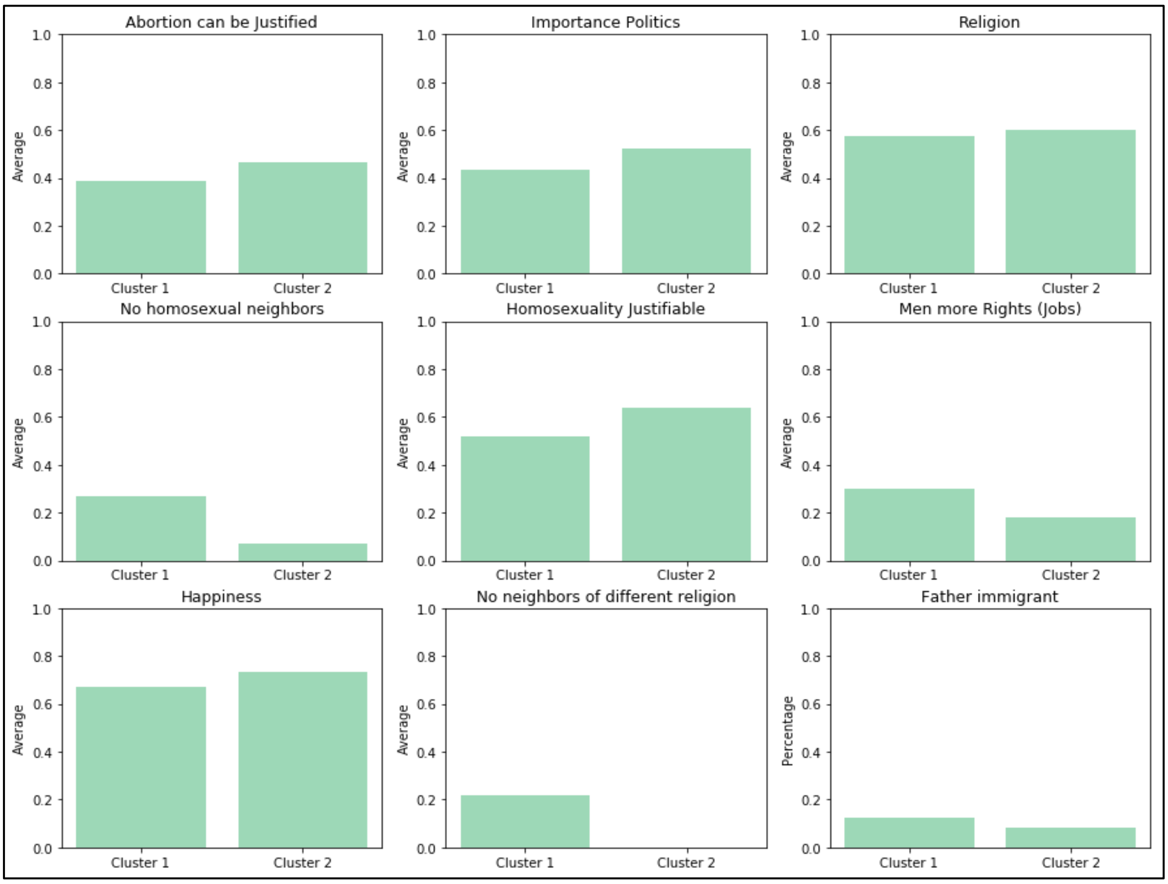

Figure 4 BGMM Clusters along nine tested items.

The two subgroups formed by the k-means clustering algorithm show a throughout distinction (Figure 5). Cluster 1 correlates on all nine variables with secular-rational and self-expression values and cluster 2 with the opposing.

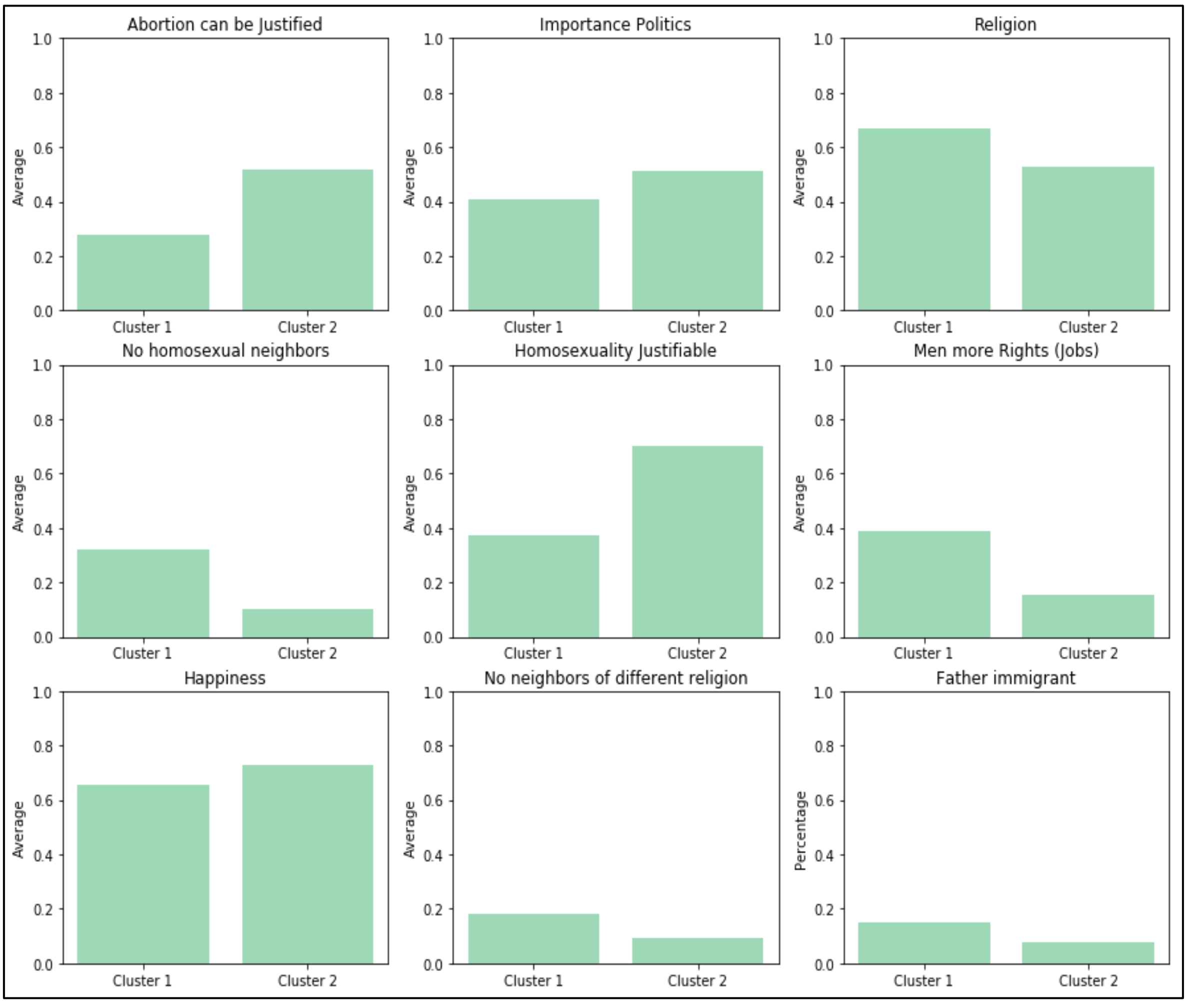

Figure 5 K-means Clusters along nine tested items.

Given these results, a hypothesis can be formulated that the k-means and BGMM clusters separate along the survival vs. self-expression dimension.

Life-Satisfaction Analysis on Clusters

The above regression results on life-satisfaction are average effects and do not apply to all German citizens. How strongly which variables influence people's life-satisfaction varies. Apart from individual differences, subgroups of the population can have different conditions. It is likely that life-satisfaction depends on different aspects of life in these subgroups.

The theory of Maslow states that life-satisfaction depends on the satisfaction of hierarchical needs. Whenever basic needs are satisfied, finer needs and their satisfaction become relevant for the subjective well-being. Multiple empirical studies

(Deaton 2008; Howell & Howell, 2008) have confirmed that life-satisfaction hinges on different variables in wealthy and less wealthy countries. It is possible that these effects occur as well in the within-nation clusters.

It is likely that people in a society located at different poles of the two dimensions have diverging needs and face different levels of satisfaction of those and thus differ in their factors that most influence their life-satisfaction. As shown, the clusters found here separate the observations with regard of the two dimensions to some extent. Thus, the two clustering methods that produce clusters best align with the two dimensions are chosen for a more detailed analysis. A lasso regression based on each of these subgroups is conducted.

The results of the lasso regression on both subgroups created by the BGMM and the k-means clustering are shown in Tables 5-8. As one can see, some factors remain the same across all four analyses and some factors change.

In all four cases, the two variables with the highest coefficient effects are the *financial satisfaction* and the *freedom of choice and control over own life.* More positive answers to both of these questions are strongest correlated amongst all variables and are selected in all four lasso regressions. These factors have also been found in the first part on the combined data set.

Another variable that is selected via lasso from over 200 possible variables in all four cases is the question on whether *people can be trusted* or not. This variable is coded such that higher numbers indicate lower trust to people. Since all four coefficients are positive, although rather small, this means that, the more trust people have, the lower their life-satisfaction. As seen in Part I, the *happiness* variable correlates with the target as well.

Table 5 Results of the lasso regression on the BGMM "self-expression" cluster.

	coef	std err	t	P>|t|	[0.025	0.975]
V10: Feeling of happiness	-0.1717	0.041	-4.222	0.000	-0.252	-0.092
V24: Most people can be trusted	0.0400	0.014	2.814	0.005	0.012	0.068
V26: Active/Inactive membership: Sport or recreational organization	0.0502	0.016	3.161	0.002	0.019	0.081
V55: How much freedom of choice and control over own life	0.3799	0.030	12.611	0.000	0.321	0.439
V59: Satisfaction with financial situation of household	0.4617	0.029	15.827	0.000	0.404	0.519
V231: Nature of tasks: manual vs. intellectual	0.1379	0.021	6.524	0.000	0.096	0.179
V234: Are you supervising someone	0.0520	0.015	3.442	0.001	0.022	0.082
V235: Are you the chief wage earner in your house	0.0426	0.015	2.809	0.005	0.013	0.072

Table 6 Results of the lasso regression on the BGMM "survival" cluster.

	coef	std err	t	P>|t|	[0.025	0.975]
V10: Feeling of happiness	-0.2111	0.030	-7.026	0.0	-0.270	-0.152
V24: Most people can be trusted	0.0596	0.013	4.460	0.0	0.033	0.086
V26: Active/Inactive membership: Sport or recreational organization	0.0815	0.015	5.356	0.0	0.052	0.111
V55: How much freedom of choice and control over own life	0.4051	0.023	17.622	0.0	0.360	0.450
V59: Satisfaction with financial situation of household	0.4617	0.024	19.492	0.0	0.415	0.508
V181: Worries: Losing my job or not finding a job	0.1846	0.019	9.724	0.0	0.147	0.222
V237: Family savings during past year	0.0905	0.021	4.269	0.0	0.049	0.132

Table 7 Results of the lasso regression on the k-means "self-expression" cluster.

	coef	std err	t	P>|t|	[0.025	0.975]
V10: Feeling of happiness	-0.2661	0.036	-7.327	0.000	-0.337	-0.195
V24: Most people can be trusted	0.0723	0.017	4.373	0.000	0.040	0.105
V55: How much freedom of choice and control over own life	0.2901	0.028	10.519	0.000	0.236	0.344
V59: Satisfaction with financial situation of household	0.4432	0.030	15.022	0.000	0.385	0.501
V160D: I see myself as someone who: is relaxed, handles stress well	0.3227	0.028	11.352	0.000	0.267	0.378
V160I: I see myself as someone who: gets nervous easily	0.0824	0.026	3.171	0.002	0.031	0.133
V237: Family savings during past year	0.0884	0.026	3.397	0.001	0.037	0.140

Table 8 Results of the lasso regression on the k-means "survival" cluster.

	coef	std err	t	P>|t|	[0.025	0.975]
V10: Feeling of happiness	-0.1283	0.031	-4.162	0.000	-0.189	-0.068
V11: State of health (subjective)	0.0603	0.023	2.662	0.008	0.016	0.105
V24: Most people can be trusted	0.0333	0.012	2.818	0.005	0.010	0.057
V55: How much freedom of choice and control over own life	0.4890	0.024	20.505	0.000	0.442	0.536
V59: Satisfaction with financial situation of household	0.4230	0.022	19.201	0.000	0.380	0.466
V181: Worries: Losing my job or not finding a job	0.1508	0.017	8.876	0.000	0.117	0.184

Cluster 2 was shown to be relatively more “self-expressing” and “rational-secular” than cluster 1. And indeed, the relevant variables suggest that this description is accurate. Looking at the two “survival” groups, a variable that additionally correlates with life-satisfaction is the “*worry of losing or not finding a job*” with rather large coefficients. The coefficient size of around 0.15/ 0.18 is the fourth/third largest in these regressions, respectively. The more worry about jobs, the less life-satisfaction is reported for these clusters. Note, that this variable is not selected for the “self-expression” groups.

The “self-expression” groups show different additional variables depending on the clustering methods. In the k-means “self-expression” group, two variables on mental well-being which capture *stress resilience* and *nervousness* are correlated with life-satisfaction. The higher the stress resilience and the ability to cope with stress, the higher the life-satisfaction.

In the BGM “self-expression” group, three work and self-actualization related variables are selected. *Intellectual labour* compared to manual labour correlates positively with life-satisfaction. *Supervising* someone or being the *chief wage earner* in the household are negatively correlated with life-satisfaction.

These finding support the hypothesis that the clusters are formed along the survival / self-expression dimension and furthermore confirm Maslow’s theory on personal hierarchical needs. People in the “self-expression” clusters are financially secured and their life-satisfaction depends on subjective well-being (stress resilience) or work-life-balance (not supervising, not chief earner). For people who are not financially secured yet, this aspect (no fear of job loss) is the main indicator for a satisfied life.

Conclusion and Outlook

Overall, some variables are consistent across all regressions and clusters and seem to robustly be correlated with life satisfaction in contemporary Germany: *Happiness, Freedom of Choice* and *Financial Satisfaction* stick out in this analysis. When analysing life-satisfaction on clusters, the findings are different. While the above-mentioned factors persist, different new items are found in the clusters. These factors confirm the hypothesis that the clusters correlate with the “survival” vs. “self-expression” dimension.

Given current political discussions on measurements against climate change in Germany, the findings of this analysis can help understand the opposing sides when it comes to economic risk. The importance of not fearing to lose one’s job is only relevant for a subgroup of the population: people on the “survival” side.

Moreover, this analysis shows how life-satisfaction factors not only differ across nations, but also within nations, such as Germany. The similarity of the patterns indicate universal underlying mechanisms and confirm Maslow’s theory of hierarchical need.

Further research could shed light on the composition of the clusters and repeat the analysis on more clusters to get a more detailed insight. Currently, the clusters mainly distinguish along the “survival” vs. “self-expression” scale, a second cut can incorporate the second dimension as well.

References

Algan Y., Dustmann C., Glitz A., Manning A. (2010). The Economic Situation of First and Second- Generation Immigrants in France, Germany and the United Kingdom, *The Economic Journal*, Volume 120, Issue 542, pages F4–F30

Deaton, A. (2008). Income, health, and well-being around the world: Evidence from the Gallup World Poll. *Journal of Economic perspectives*, *22*(2), 53-72.

Diener, E., & Suh, E. (1997). Measuring quality of life: Economic, social, and subjective indicators. *Social indicators research*, *40*(1-2), 189-216.

Diener, E., Lucas, R. E., & Oishi, S. (2002). Subjective well-being: The science of happiness and life satisfaction. *Handbook of positive psychology*, *2*, 63-73.

Diener, E., & Diener, M. (2009). Cross-cultural correlates of life satisfaction and self-esteem. In *Culture and well-being* (pp. 71-91). Springer, Dordrecht.

German Federal Agency for Civic Education https://www.bpb.de/gesellschaft/migration/laenderprofile/262758/historical-and-current-development-of-migration-to-and-from-germany (last accessed May, 6th 2019)

Hoge, D. R., Petrillo, G. H., & Smith, E. I. (1982). Transmission of religious and social values from parents to teenage children. *Journal of Marriage and the Family*.

Howell, R. T., & Howell, C. J. (2008). The relation of economic status to subjective well-being in developing countries: A meta-analysis. *Psychological bulletin*, *134*(4), 536.

Inglehart, R. (1997). *Modernization and postmodernization: Cultural, economic, and political change in 43 societies*. Princeton university press, page 81-98. ALSO WEBSITE

Inglehart, R., & Baker, W. E. (2000). *Modernization, Cultural Change, and the Persistence of Traditional Values. American Sociological Review, 65(1), 19.*

Inglehart, R., & Welzel, C. (2010). Changing Mass Priorities: The Link between Modernization and Democracy. *Perspectives on Politics*, 8(02), 551–567.

Inglehart, R., C. Haerpfer, A. Moreno, C. Welzel, K. Kizilova, J. Diez-Medrano, M. Lagos, P. Norris, E. Ponarin & B. Puranen et al. (eds.). 2014. World Values Survey: Round Six - Country-Pooled Datafile Version: www.worldvaluessurvey.org/WVSDocumentationWV6.jsp. Madrid: JD Systems Institute.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning* (Vol. 112, p. 18). New York: springer.

Maslow, A. H. (1943). A theory of human motivation. *Psychological review*, *50*(4), 370.

Maslow, A. H., Frager, R., Fadiman, J., McReynolds, C., & Cox, R. (1970). Motivation and personality. Harper & Row New York. *McClelland, DC, & Burnham, DH (1976). Power is the great motivator. Harvard Business Review*, *25*, 159-166.

Oishi, S., Diener, E., Lucas, R. E., & Suh, E. M. (2009). Cross-cultural variations in predictors of life satisfaction: Perspectives from needs and values. In *Culture and well-being* (pp. 109-127). Springer, Dordrecht.

Veenhoven, R. (1991). Is happiness relative? *Social Indicators Research*, 24, 1–34.