Final Class Project

What makes us happy

Author:
Alexander Manfred Fichtl

Final class project for the 373290: Introduction to Data Science (Spring 2020) course

14.06.2020

Table of Contents

1. Introduction

Countless variables in your life determine how happy you feel in general or on a day to day basis. While it's impossible to survey all of these variables, the "Young People Survey" from a statistics class at FSEV UK did quite an excellent job of exploring the preferences, interests, habits, opinions, and fears of young people. These variables do, in fact, contribute to a big chunk to one's happiness.

The dataset (responses.csv & columns.csv) can be found at

https://www.kaggle.com/miroslavsabo/young-people-survey

and comes with the following overview of the contents:

The data file (responses.csv) consists of 1010 rows and 150 columns (139 integers and 11 categorical). For convenience, the original variable names were shortened in the data file. See the columns.csv file if you want to match the data with the original names.

- The data contain missing values.
- The survey was presented to participants in both electronic and written form.
- The original questionnaire was in Slovak language and was later translated
- into English.
- All participants were of Slovakian nationality, aged between 15-30.

The variables can be split into the following groups:

- Music preferences (19 items)
- Movie preferences (12 items)
- Hobbies & interests (32 items)
- Phobias (10 items)
- Health habits (3 items)
- Personality traits, views on life, & opinions (57 items)
- Spending habits (7 items)
- Demographics (10 items)

This data story will discuss the results of the EDA, Data Cleansing, Feature Engineering and Machine Learning done in R Studio. The main goal of this project is to find out, whether or how strong social factors determine the "Happiness in life" of young people and to try and predict this variable of young people given the columns in the dataset. Moreover, since I am personally interested in the "Changing the past", "Religion", and "Science and technology" variables, there will be some work on those too. The variables were displayed in the survey as follows:

- "I wish, I could change the past": Strongly disagree 1-2-3-4-5 | Strongly agree (integer)

- "I am 100% happy with my life": Strongly disagree 1-2-3-4-5 | Strongly agree (integer)
- "Religion": Not interested 1-2-3-4-5 | Very interested (integer)
- "Science and technology": Not interested 1-2-3-4-5| Very interested (integer)

2. Exploratory Data Analysis (EDA)

2.1 Basic EDA

Here are some basic statistics regarding the dataset:

Name	*Value*
Rows	1,010
Columns	150
Discrete columns	11
Continuous columns	139
All missing columns	0
Missing observations	571
Complete Rows	686
Total observations	151,500
Memory allocation	629.7 Kb

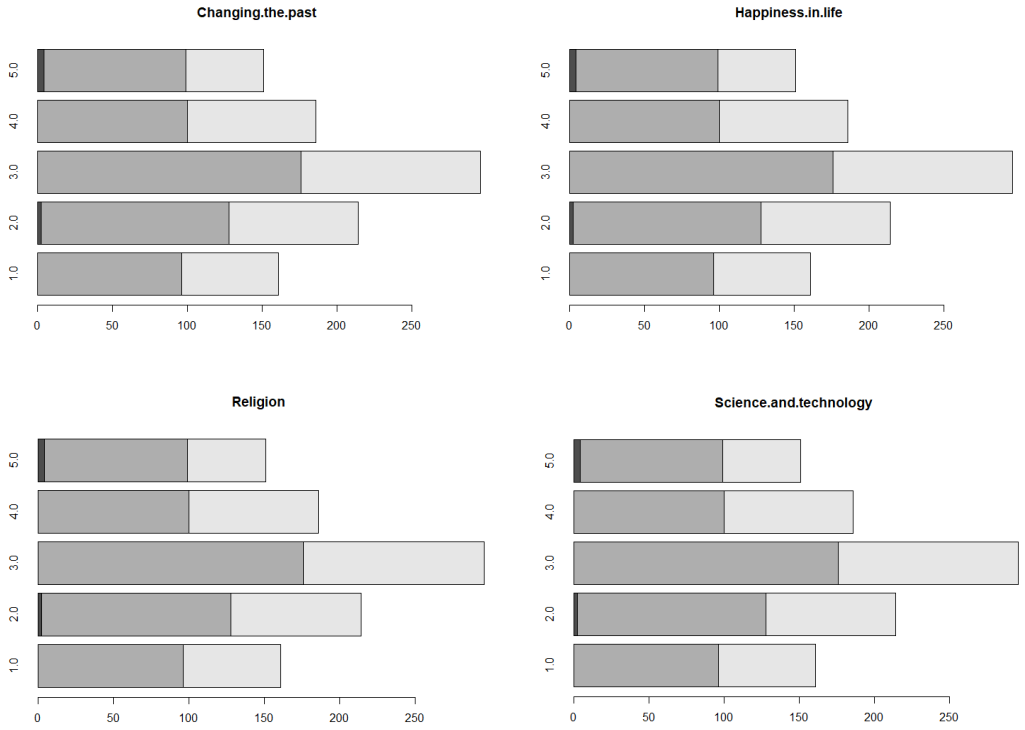
Luckily, the dataset is quite complete, with only very few missing values. At the same time, though, the dataset is rather complex, with many different types of columns.

Variable	*Numer of NA's*	*Complete Rate*	*Mean*	*Standard Deviation*
Changing the past	2	0.998	2.95	1.28
Happiness in life	4	0.996	3.71	0.824
Religion	3	0.997	2.27	1.32
Science and technology	6	0.994	3.23	1.28

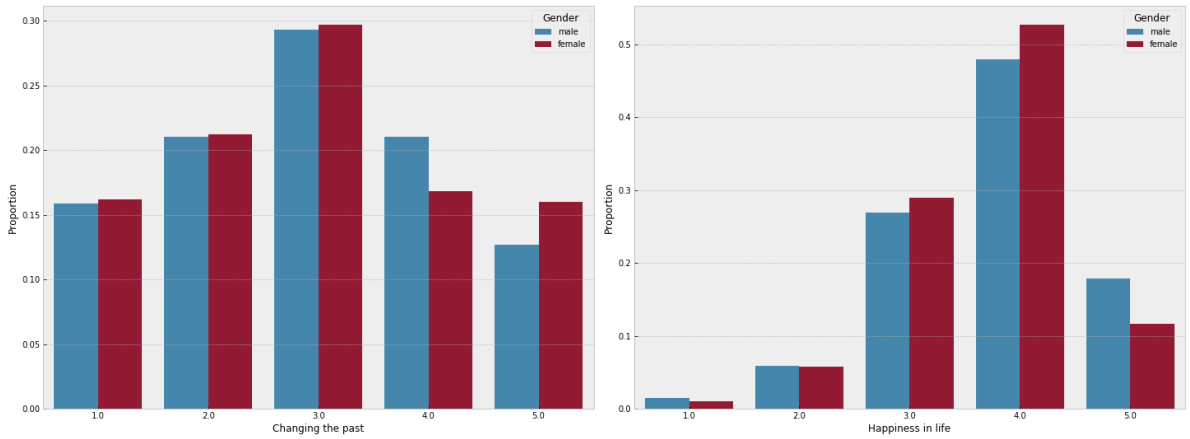
Looking at the variables of interest, the means tell us that on average, young people seem to be quite happy, while still having some regrets, be rather uninterested in religion, and somewhat interested in Science and technology. Moreover, there are missing values amongst each column, but they are very few (less than 1%).

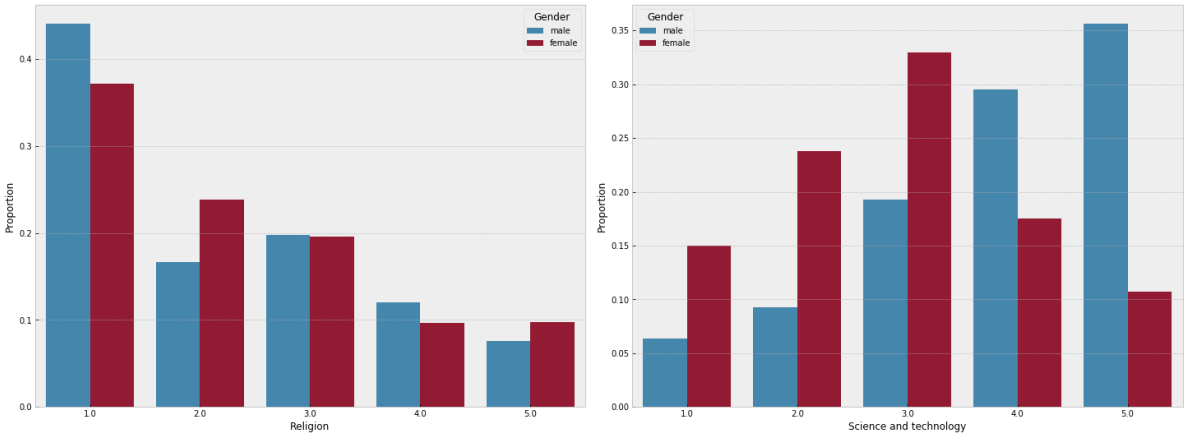
2.2 In-depth EDA

In this section, we will analyze characteristic differences and anomalies in the dataset and the correlations that are formed with our variables of interest. Let us start by looking at the distribution of the variables:



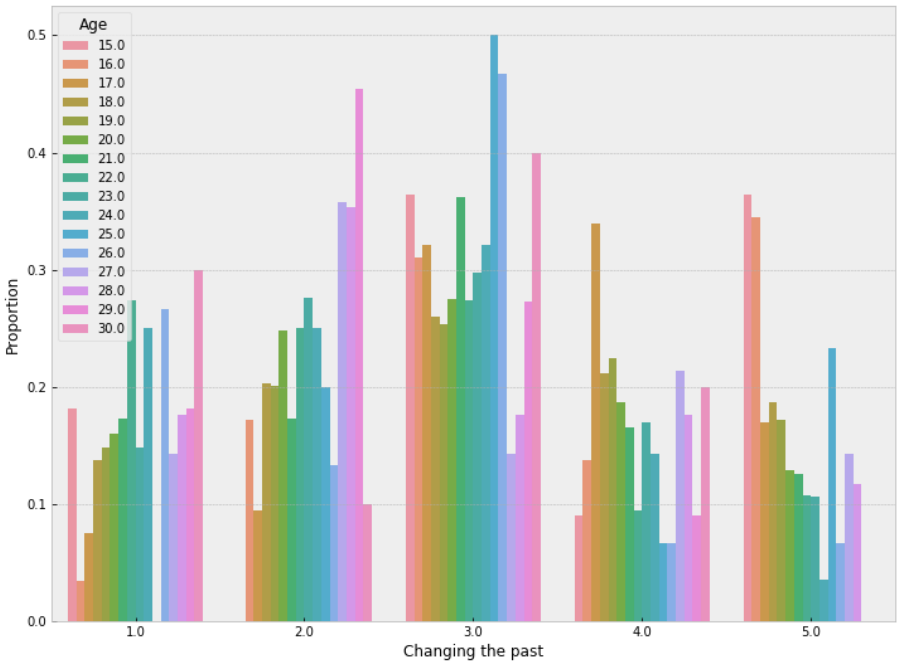
Unfortunately, the classes are highly imbalanced, especially "Happiness in life". This is a problem for a lot of machine learning models. Luckily there are some solutions that we were able to make use of, even for the multiclass classification. Let's see what the distribution of the variables between the given 2 genders looks like:





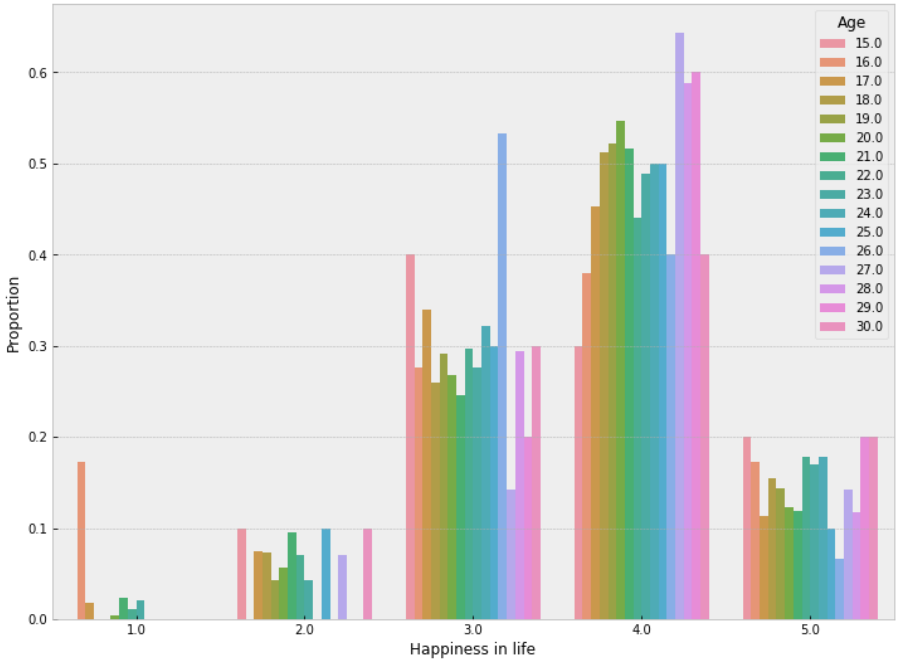
The most notable insights from the plots definitely lie within the "Science and technology" plot. Men seem to be way more interested in scientific and technological topics. This was rather expected, though. Women, for example, make up only 28% of the workforce in Science, technology, engineering, and math (STEM) as well.
(See https://www.aauw.org/resources/research/the-stem-gap/)

Next, we will check if there is anything noticeable within the age distribution amongst the variables:

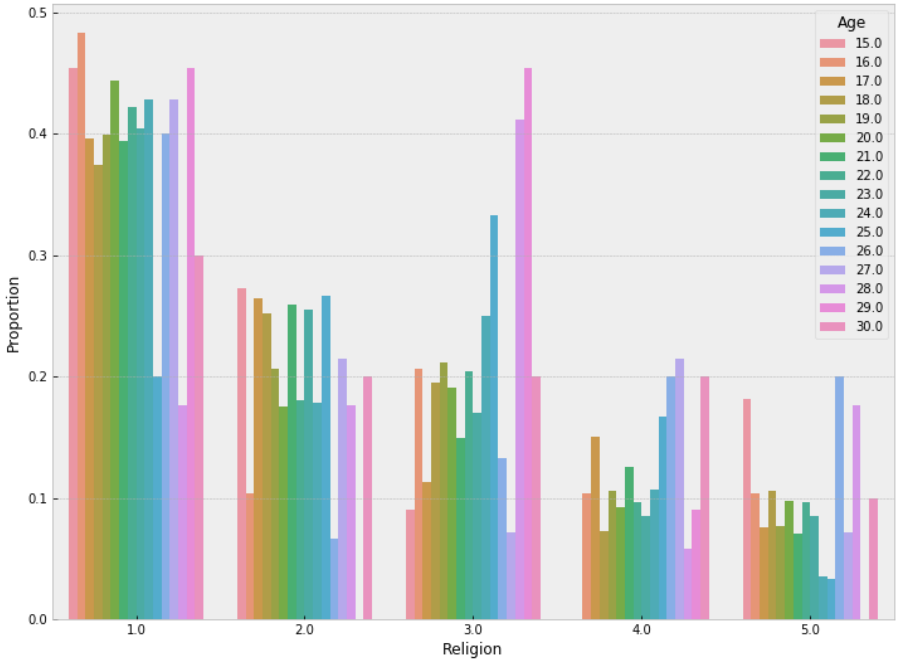


In the “Changing the past” plot, there's a peak of 15 and 16-year-olds at the 5.0 score. They seem to double as likely to strongly agree to the statement, "I wish, I could change the past" than more grown-ups are. This could be chance, though, if there are very few participants that are this young. With 40 participants being in this age range, I'd say the discovery is at least somewhat significant. An explanation might be that at the age of 15 and 16, the participants might still be in their puberty and therefore react quite emotional to what happens in their life and ”being in their "kid rebellious phase" they often overdramatize. However, more grown-ups should have an easier time accepting and living with unfortunate events in the past and more often have a grateful "this made me who I am" attitude. This fits

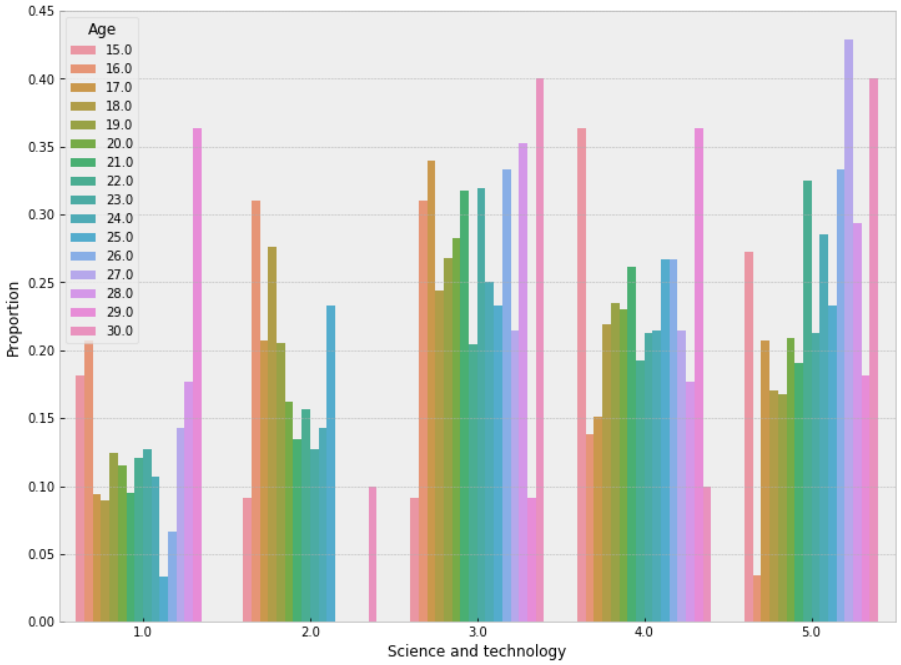
well with the 29- and 30-year-olds concentrating on the 1, 2, and 3 answers while entirely skipping the 5. But of course, chance could still play a role here.



In the “Happiness in life” plotThe peak of 16-year-olds at 1.0 might be explained by them overdramatizing a lot of things. Another interesting fact is there are no 18 or 19-year-olds that picked 1, even though they make up a considerable chunk of the participant pool. And finally, the oldest participants, the 27+-year-olds, seem to be the happiest on average.



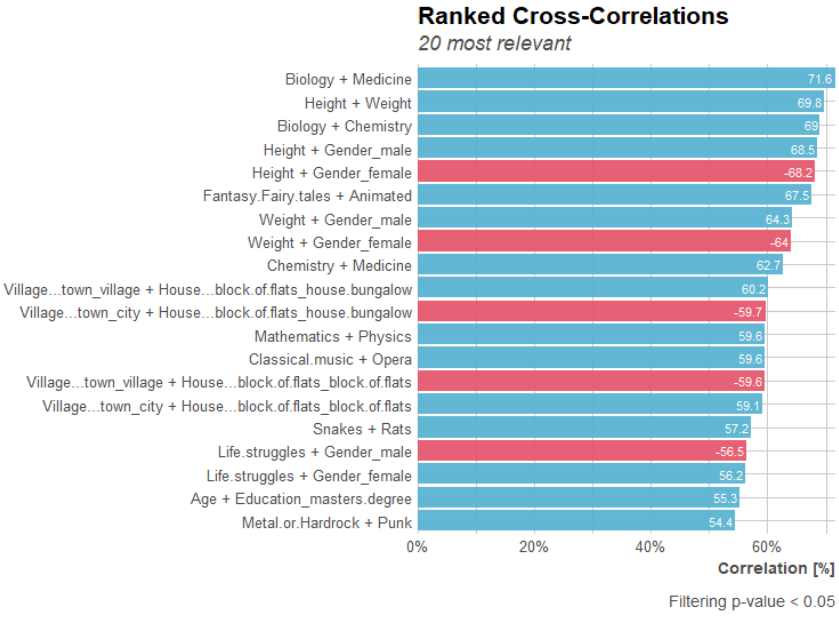
Here, only some singled out ages stand out, like 28 (low 1.0 count, high 3.0 and 5.0 count) or 29 (high 3.0 count). But since there are also only a few participants with this age, this is probably not significant. The insight we can take from this is that rather low interest in religion is common within all age groups.



Yet again, responses of the age groups with a low participant count tend to have a lot of peaks within the extremes (1.0 and 5.0). This is a common statistical observation, though. Therefore, we should not interpret too much into it. The high participant age groups are more evenly distributed with fewer peaks. Still, a slight trend towards more interest in science and technology with higher age seems to be present.

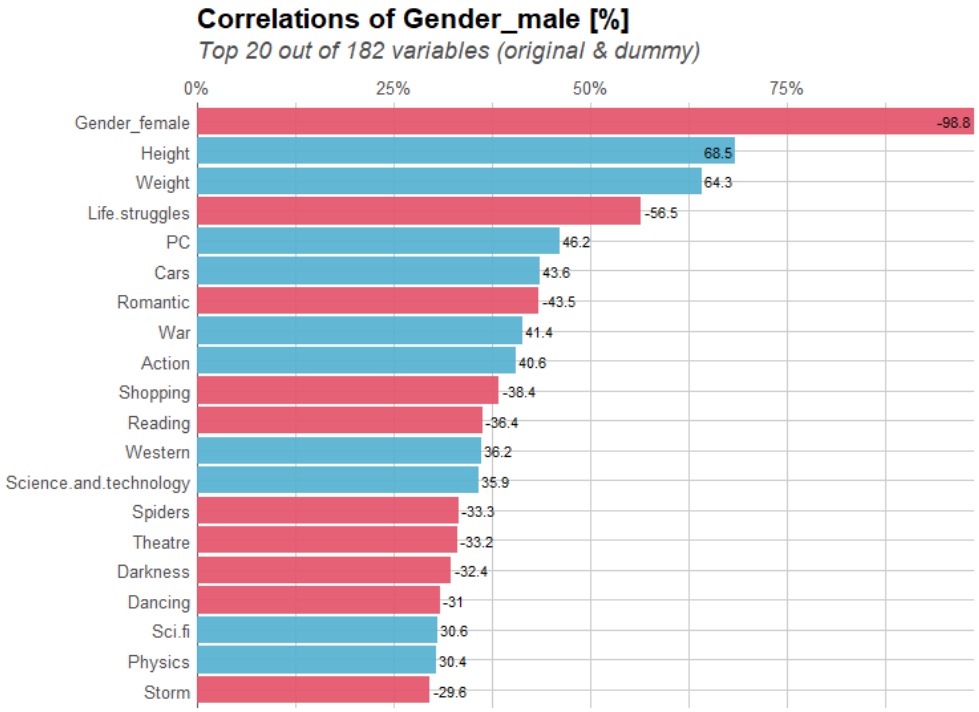
Correlations

Next, we will check for correlations. We have to be careful, though, a high correlation between two variables does not mean that the two variables are causally related. Instead, correlations only provide an initial indication that this could be the case. Therefore, we will have to take all the upcoming insights with a grain of salt. Let us start by looking at the highest general correlations in the dataset:



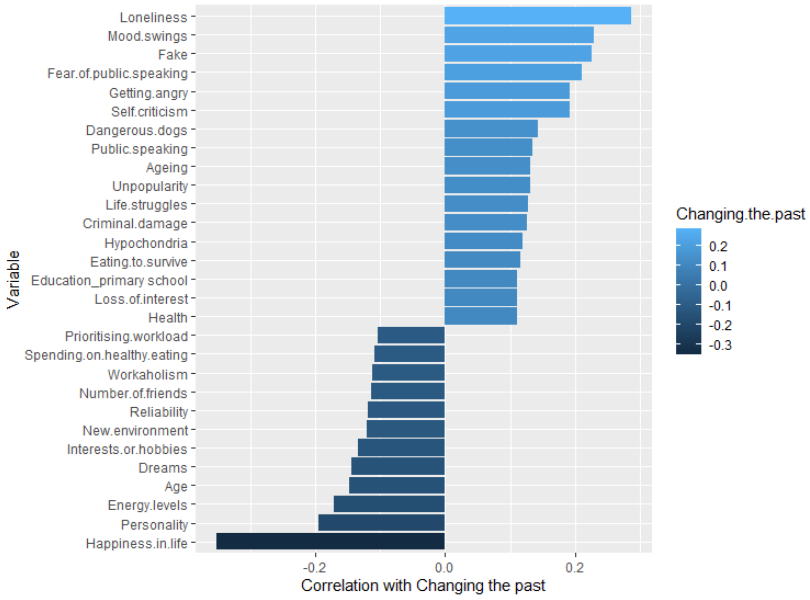
These are the most significant correlations of all variables of the dataset. Most of them are quite self-explanatory, especially the positive correlations, like Biology and Medicine or Height and Weight. The negative correlations are not that easy to understand. Amongst them, we find the Loneliness/Happiness in life and Changing the past/Happiness in life correlations (part of our variables of interest). We can ignore the negative correlations between the dummy variables of the same original categorical variable. This only indicates that, of course, you cannot pick two of them at the same time.

To give a quick input on most of the other negative correlations: They don't make immediate sense either, like "Life struggles" and "Weight" or "Reading" and "Cars". But at a closer look, one variable always seems female-typical and the other one male-typical. Let us see if there really are correlations of these variables with the gender.

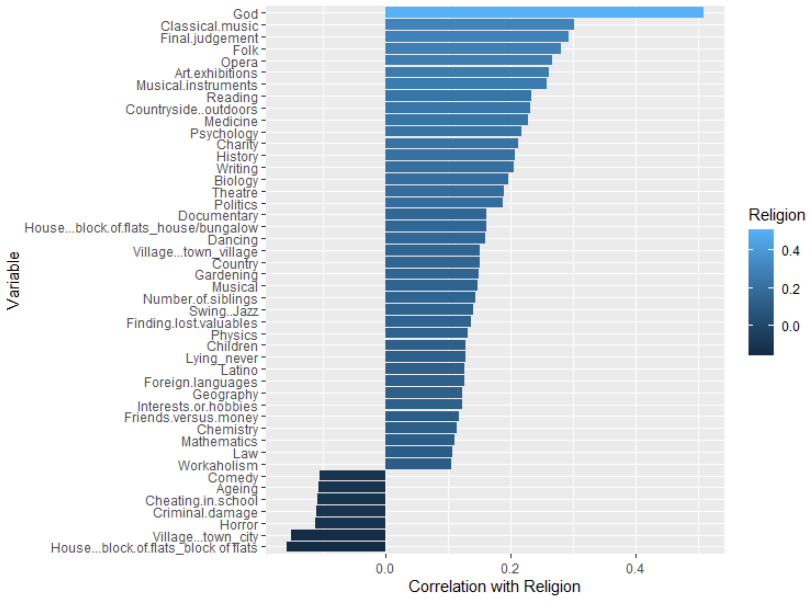


We can clearly see that all the variables have high correlations with the gender variable, indeed. This is a typical example of multicollinearity. It's not like people have fewer life struggles (life struggles being defined as "I cry when I feel down, or things don't go the right way") because they weigh more. It's just that women tend to weigh less and, at the same time, cry more often than men.

Let's move on to our variables of interest; we will use a proper plotting function for that. While we can take a brief look at the top correlations of each of the four variables, we will then focus on the results for the "Happiness in life" variable, though, to keep the scale of this project somewhat limited.

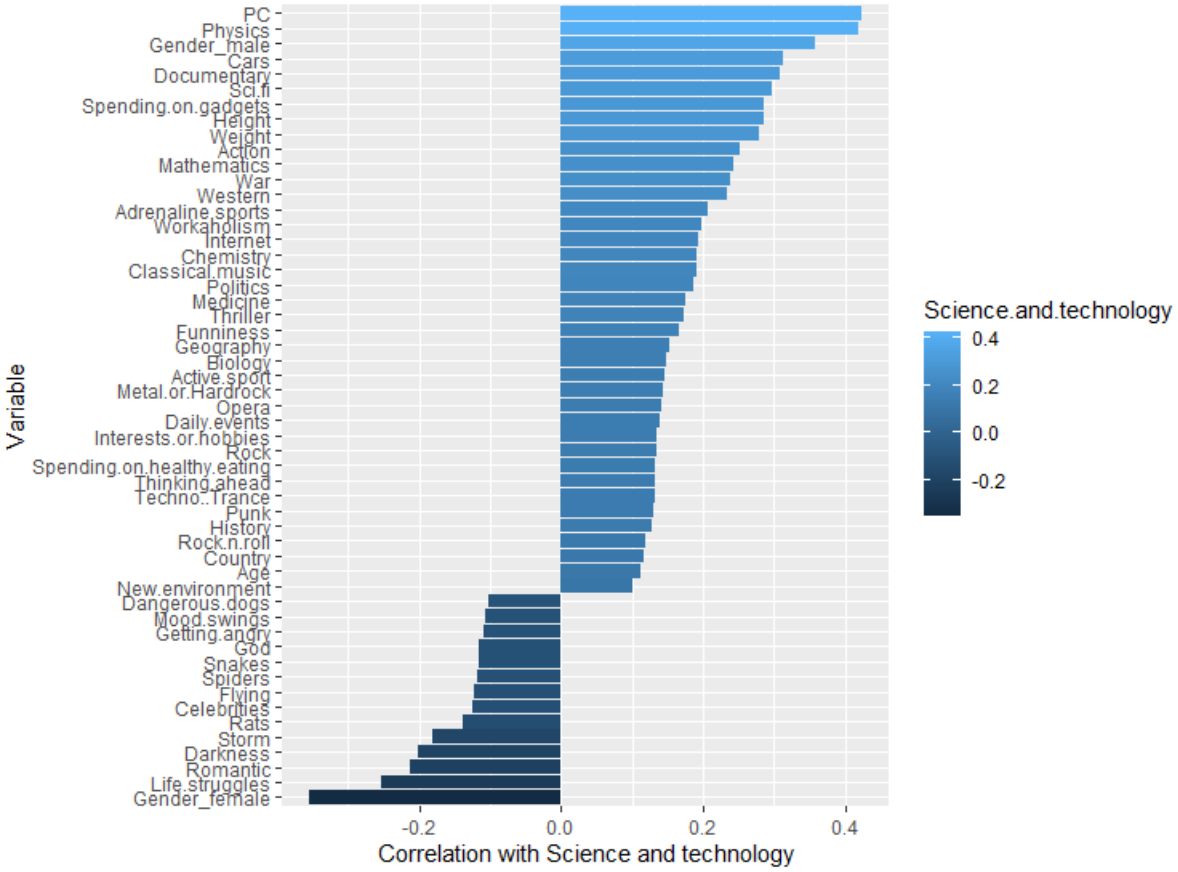


The top correlations with the “Changing the past” variable definitely make sense. If you are happy with your life and personality and are well educated, you will not regret too many of your past decisions. However, if the opposite is the case, and you feel lonely, have mood swings, and are poorly educated, it's very understandable that you might regret some of your past decisions. Maybe you regret picking up a fight with a lot of your friends and losing them, or dropping out of high school because you did not put in enough work. Overall, all the correlations are not too strong, though (around |0.3| at most).

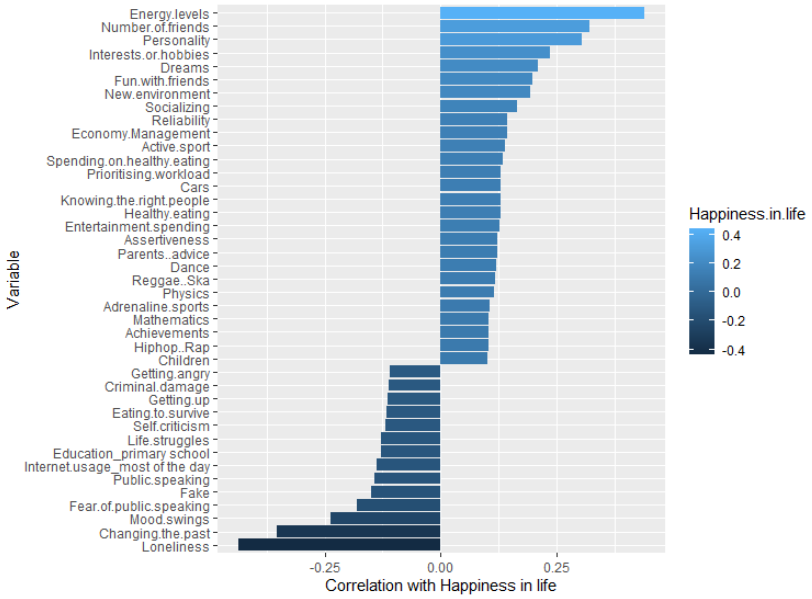


For the “Religion” variable, all the negative correlations are not very significant (around -0.1). However, the Religion variable has some strong positive correlations. First of all, the more interested one is in religion, the more he believes in god. This makes perfect sense and is not much of a surprise; the same goes for believing in final judgment. What's interesting, though, is that they will also enjoy classical music, folk, and opera more. Additionally, they will be more interested in art exhibitions, musical instruments, and even poetry. It seems like religious young people will have a bias towards "classical" culture in general. This could be

because they will get in touch with it more, especially in church. Moreover, religious people tend to live in houses out in villages. This makes sense; it is known that religion and belief in god are much more common in rural areas (where people tend to live in houses rather than in blocks) than amongst city dwellers.



Within the highest correlations of the "Science and technology" variable, we find a lot of the variables that highly correlate with the gender and, therefore, the gender variable itself as well, of course. It's not like young people are less interested in science and technology because they are romantic or cry more ("life struggle"). It's just that men are generally more interested in STEM fields and tend to be less romantic/susceptible to crying / more into cars... Real, direct (positive) correlations though are probably the one with physics and sci-fi, which obviously do make sense.



The numeric correlation values seem to be far more significant than the categorical ones. Energy levels and loneliness might play a major role in young people's happiness (corr_value > |0.4|). This makes a lot of sense: Research shows that especially chronic loneliness can have a significant impact on one's wellbeing and overall health in a very negative way. If someone is always "full of life and energy", though, this, of course, has a very positive effect on one's happiness. It's much better than always feeling down/weak and staying at home because of it. "Dreams", "Interests or hobbies", "Personality", and "Number of friends" (corr_value > 0.2) and “Changing the past” and “Mood swings” (corr_value < -0.2) are definitely a factor as well. I think the quite significant positive correlation between one's happiness in life and having good dreams is a lovely one.

The most significant corr_values for the categorical variables have to do with internet usage and education. It looks like if someone spends a lot of time online and has only been to primary school, he will be rather unhappy. This also makes sense, since well-educated people will usually have a higher standard of living. Moreover, well educated, young people typically have a bright future ahead of them. On the other hand, spending a lot of time online means less time spent in the real world where one gets vitamin D from the sun, finds friends and maybe a partner for life. This accounts for less happiness for sure.

From here on we can, without doubt, say that social factors play indeed a major role in determining young people’s happiness. But still, there are a ton of other factors that play a role as well.

3. Machine Learning

We finally made it to the machine learning section. Here we will look at the results of different machine learning models that have been trained to predict “Happiness in life” of young people.

Multinomial logistic regression

Training this model took very little time, but the results were rather mediocre: An accuracy of 50.5% on the test set only.

Neural network
Unfortunately, I do not have the experience with R yet to program a proper neural network. Something went wrong and I did not have the time to find my mistake. This is why instead, I wanted to train a Random Forest in Python, because here I do have experience.

Random Forest (Python)
The Random Forest model definitely performed way better than the other ones. While the imbalance of the "Happiness in life" variable is a big problem for all the models, we managed to somewhat fix this issue by oversampling the minority classes with SMOTE (Synthetic Minority Over-sampling Technique). With dropping some columns that might have been colinear, using proper standardization and using a random grid to search for the best hyperparameters we managed to achieve these metrics:

```
Average accuracy score on test set:  0.9238410596026491

Average f1 score on test set:  0.8040378714491199

Average precision score on test set:  0.9371340078463761

Average recall score on test set:  0.7484652202573729
```

I'd say that's quite impressive for 5 label classification problem with only 1010 rows in the dataset. And it's not just the accuracy that came out quite well (that is not an reliable metric after all), but the f1 score as well! This means, we can further confirm our thesis: Yes, it is also possible to predict the "Happiness in life" of young people, given the right data and a good ML model. Still, there is lots of room for improvement. We will come to that in the conclusion.

4. Conclusion

There are many exciting discoveries to be made within this data set, and there's a lot to learn via EDA. Unfortunately, the machine learning section in R Studio did not go too well and I did not have the time to improve the models /try different approaches. But I will invest more time into it in the upcoming weeks for sure! But within R and even with the Random Forest model there are countless things one could still try in order to improve the model / get better predictions:

- Try a different method to fix the class imbalance problem. We could, for example, try providing some bias to minority classes: We can estimate class weights in scikit_learn by using compute_class_weight and use the parameter 'class_weight' while training the model.
- Try excluding some variables (especially the dummy variables) and do more/better feature engineering in general.
- Do some PCA and see if the training provides better results with the artificial variables
- Only use the variables with significant correlations with "Happiness in life" for the training
- I figured, it could be very promising to try efficient pairwise multilabel classification. I wanted to try following this guide: https://xang1234.github.io/multi-label/. Unfortunately, I did not have any more time.

However, the most significant improvement in the overall EDA insights and the ML section could have probably been made by adding just a few more variables to the dataset. I figured, the questions that are missing most in the survey are:

- "How healthy do you consider yourself to be?"
- "What is your relationship status?"
- "What is your financial status?"
- "Do you have a lot of contact with your family?"
- "How much sport do you do?"

In my opinion, these questions could have contributed a lot of essential information to the dataset. Happiness could have probably been predicted more easily, and a lot of other insights might have come from it as well.

"Happiness is when what you think, what you say, and what you do are in harmony." — Mahatma Gandhi