Capstone Project
“Climate and Happiness”

Author: Delf Lachmund, 07/2019

I. Introduction

Every year a *world happiness report* is published. It gathers several statistics on wellbeing and happiness with the help of respondents of over 100 countries. In this project we will collect geojson and venue data of the wordwide happiness and climate data, and discuss how strongly the climate influences the happiness of a country. It is to be expected that there are stronger factors than climate when it comes to overall happiness of the people. However, we will specifically investigate if there is any influence at all.

To answer these questions we plan to do the following:

1) Gather the happiness table
2) Retrieve geojson information of the countries
3) Gather climate data, we will focus on avg. annual temperature and precipitation data
4) Merge the tables
5) Visualize the world maps using folium
6) Investigate the correlations between happiness, temperature, and precipitation
7) Build a multilinear regression model
8) Analyze the model with scatter plots, histograms, and expected variance

We will use unsupervised machine learning techniques and formulate different assumptions leading us through the analysis. Obviously, more specific knowledge could provide a better model - certain factors will always be out of scope. That being said, any considerations are made as realistic as possible.

The *target audience* is everyone interested in climate, as well as people who want to know more about which countries are the happiest.

II. Data

1) World Happiness Report Data

The data we use comes from the official world happiness report, compare https://en.wikipedia.org/wiki/World_Happiness_Report .

It is based on annual samples for the world as a whole, we will use the average data for the years 2016-2018. Three main happiness measures are reported, here we will only consider the first one of them: individual life evaluations. The life evaluation used is the Cantril Ladder, which asks survey respondents to place the status of their lives on a “ladder” scale ranging from 0 to 10, where 0 means the worst possible life and 10 the best possible life.

In the report the effects of country wide factors as health, life expectancy, gdp, and individual factors as income, social support, generosity, and corruption perception, have been considered in detail. Here we will discard those factors and merge the happiness scores with climate scores for the respective countries. More informations about the data can be found on https://s3.amazonaws.com/happiness-report/2019/WHR19_Ch2A_Appendix1.pdf .

	Country	Happiness score
0	Finland	7.7689
1	Denmark	7.6001
2	Norway	7.5539
3	Iceland	7.4936
4	Netherlands	7.4876
5	Switzerland	7.4802
6	Sweden	7.3433
7	New Zealand	7.3075
8	Canada	7.2781
9	Austria	7.2460
10	Australia	7.2280
11	Costa Rica	7.1674
12	Israel	7.1387
13	Luxembourg	7.0903
14	United Kingdom	7.0537
15	Ireland	7.0211
16	Germany	6.9850
17	Belgium	6.9230
18	United States of America	6.8923
19	Czech Republic	6.8521

2) Temperature and precipitation data

We will import tables with typical temperature and precipitation dates. The world bank provides historic data for the years 1961-1999. Here we will use the respective data and claim that the general climate won't change significantly within 20 years. The data can be found on 'http://databank.worldbank.org/data/download/catalog/cckp_historical_data_0.xls' .

	ISO_3DIGIT	Annual_temp
0	AFG	12.921455
1	AGO	21.510933
2	ALB	11.269800
3	ARE	26.825609
4	ARG	14.215225

	ISO_3DIGIT	Annual_precip
0	AFG	311.321856
1	AGO	991.305683
2	ALB	1053.235184
3	ARE	67.812025
4	ARG	559.045871

There is the problem that the country column is represented by a 3 digit country code. So for merging (“inner join”) the happiness, temperature, and precipitation tables, we need to convert the codes to full country names. For this we will use the additional table 3*.

3*) Regional Codes Data

We found this additional table on https://raw.githubusercontent.com/lukes/ISO-3166-Countries-with-Regional-Codes/master/all/all.csv .

	name	alpha-2	alpha-3	country-code	iso_3166-2	region	sub-region	intermediate-region	region-code	sub-region-code	intermediate-region-code
0	Afghanistan	AF	AFG	4	ISO 3166-2:AF	Asia	Southern Asia	NaN	142.0	34.0	NaN
1	Åland Islands	AX	ALA	248	ISO 3166-2:AX	Europe	Northern Europe	NaN	150.0	154.0	NaN
2	Albania	AL	ALB	8	ISO 3166-2:AL	Europe	Southern Europe	NaN	150.0	39.0	NaN
3	Algeria	DZ	DZA	12	ISO 3166-2:DZ	Africa	Northern Africa	NaN	2.0	15.0	NaN
4	American Samoa	AS	ASM	16	ISO 3166-2:AS	Oceania	Polynesia	NaN	9.0	61.0	NaN

4) Country-wide geojson information will be collected.

III. Methodology

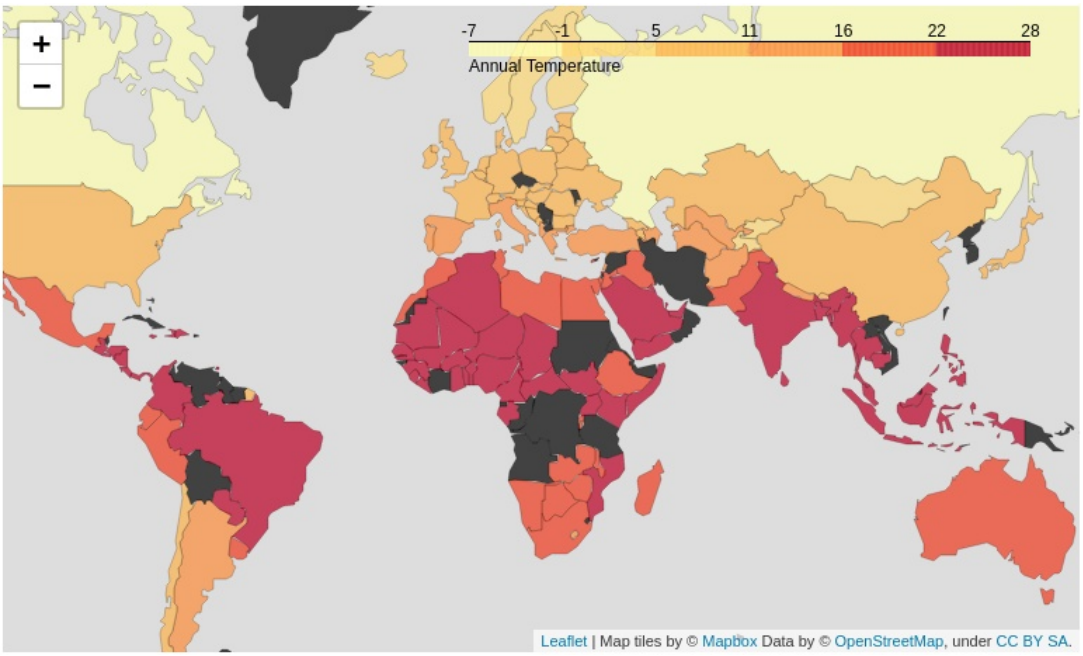
First we merge the tables into the following data frame with an “inner join”:

	Country	region	sub-region	Happiness score	Annual_temp	Annual_precip	Annual_precip_sqrt
0	Finland	Europe	Northern Europe	7.7689	1.347861	524.701104	22.906355
1	Denmark	Europe	Northern Europe	7.6001	7.818978	698.649857	26.431985
2	Norway	Europe	Northern Europe	7.5539	0.753892	1051.840741	32.432094
3	Iceland	Europe	Northern Europe	7.4936	1.351409	1039.821209	32.246259
4	Netherlands	Europe	Western Europe	7.4876	9.201738	766.068878	27.677949
5	Switzerland	Europe	Western Europe	7.4802	4.749057	1646.412455	40.576008
6	Sweden	Europe	Northern Europe	7.3433	1.546351	621.150333	24.922888
7	New Zealand	Oceania	Australia and New Zealand	7.3075	9.975832	1790.945674	42.319566
8	Canada	Americas	Northern America	7.2781	-7.144580	457.469053	21.388526
9	Austria	Europe	Western Europe	7.2460	6.186013	1161.183869	34.076148
10	Australia	Oceania	Australia and New Zealand	7.2280	21.506676	473.091500	21.750667
11	Costa Rica	Americas	Latin America and the Caribbean	7.1674	23.850929	3268.266166	57.168752
12	Israel	Asia	Western Asia	7.1387	19.698000	248.689445	15.769890
13	Luxembourg	Europe	Western Europe	7.0903	8.881900	848.805908	29.134274
14	United Kingdom	Europe	Northern Europe	7.0537	8.310239	1154.299742	33.974987
15	Ireland	Europe	Northern Europe	7.0211	9.107924	1110.628609	33.326095
16	Germany	Europe	Western Europe	6.9850	8.504452	713.499730	26.711416
17	Belgium	Europe	Western Europe	6.9230	9.514241	874.297532	29.568523
18	United States of America	Americas	Northern America	6.8923	6.754627	654.536673	25.583914
19	United Arab Emirates	Asia	Western Asia	6.8245	26.825609	67.812025	8.234806

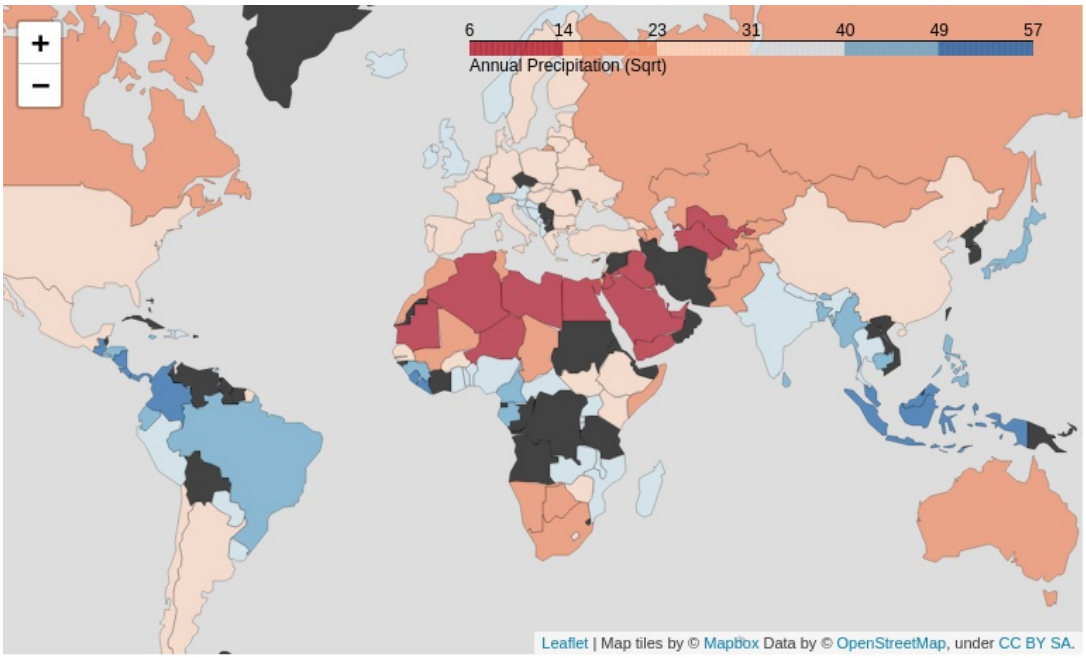
This table comprises happiness scores, as well as annual temperature and precipitation values for 133 countries.

In combination with the geojson data and folium we visualize all three variables as choropleth maps:

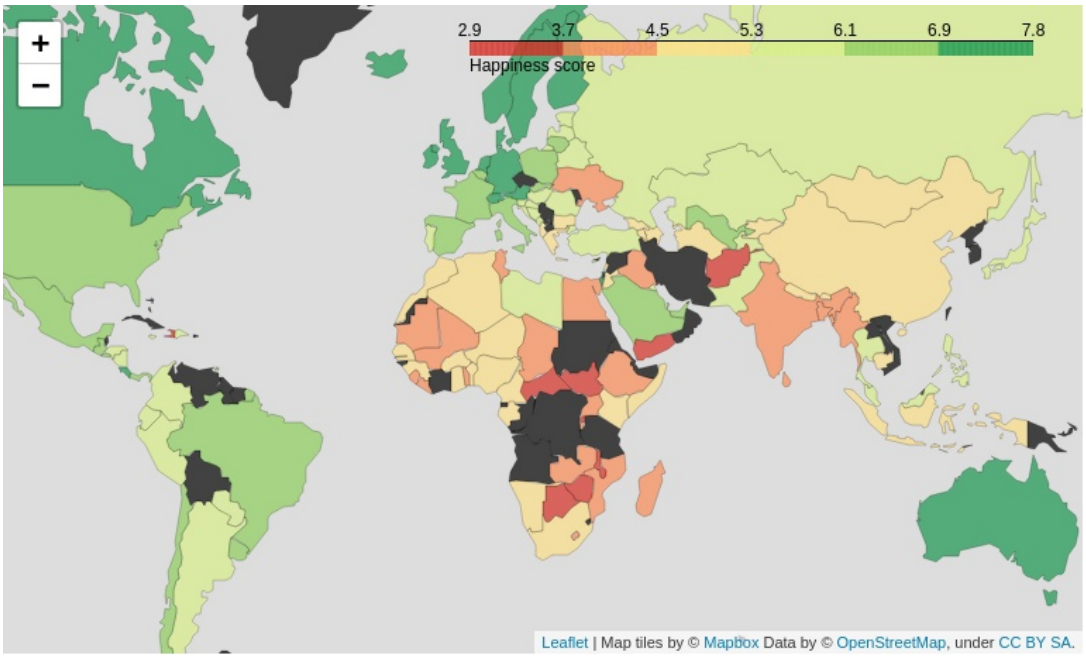
1) Temperature



2) Precipitation



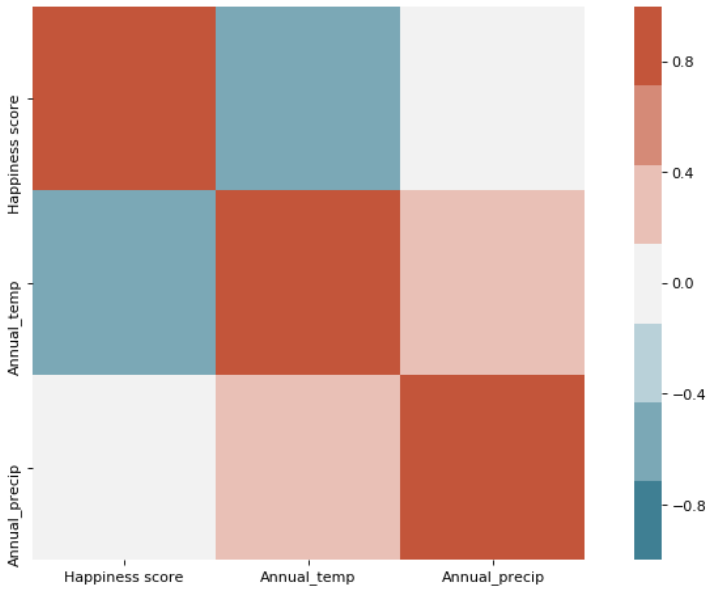
3) Happiness



We can see that some countries are not covered by the table (shown as black). For all three variables certain countries in Africa show the most extreme values, more specifically high temperatures, low precipitation, and low happiness.

We now correlate the three variables happiness, temperature, and precipitation:

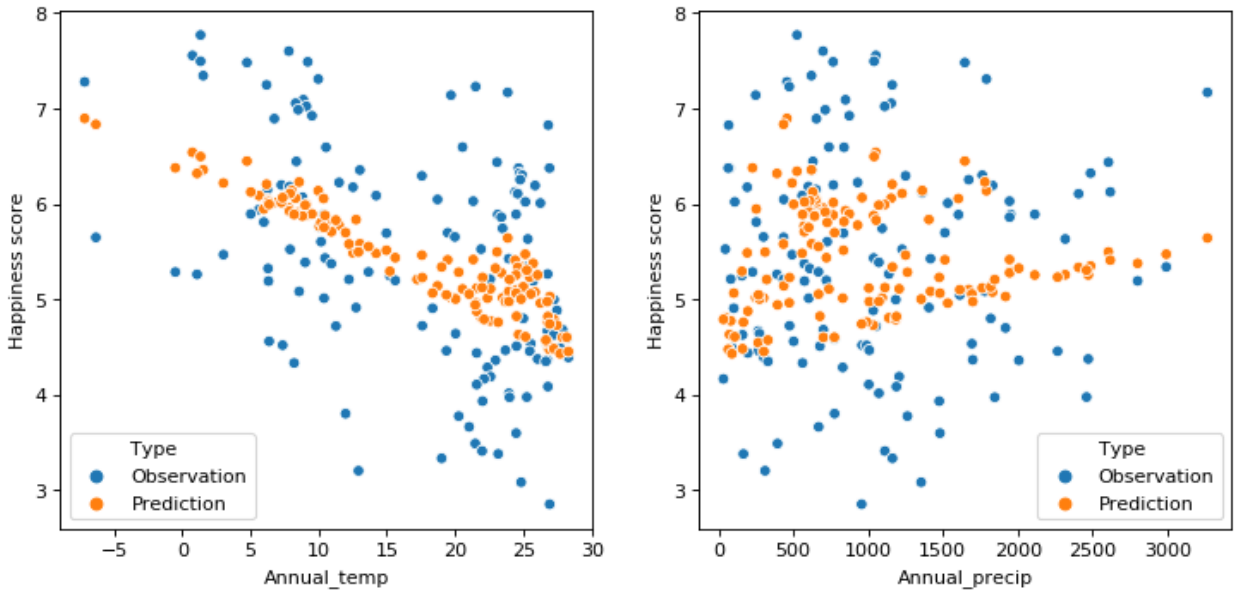
	Happiness score	Annual_temp	Annual_precip
Happiness score	1.000000	-0.458210	0.028344
Annual_temp	-0.458210	1.000000	0.309285
Annual_precip	0.028344	0.309285	1.000000



This is interesting! While happiness is negatively correlated with the temperature, with a decent correlation coefficient c of c=-0.46, precipitation and happiness are not correlated (c=0.03), even though there is a slight positive correlation between temperature and precipitation (c=0.31).

Let us now build a simple linear regression model to see how well we can predict the happiness in a country, only by knowing its annual temperature and precipitation.

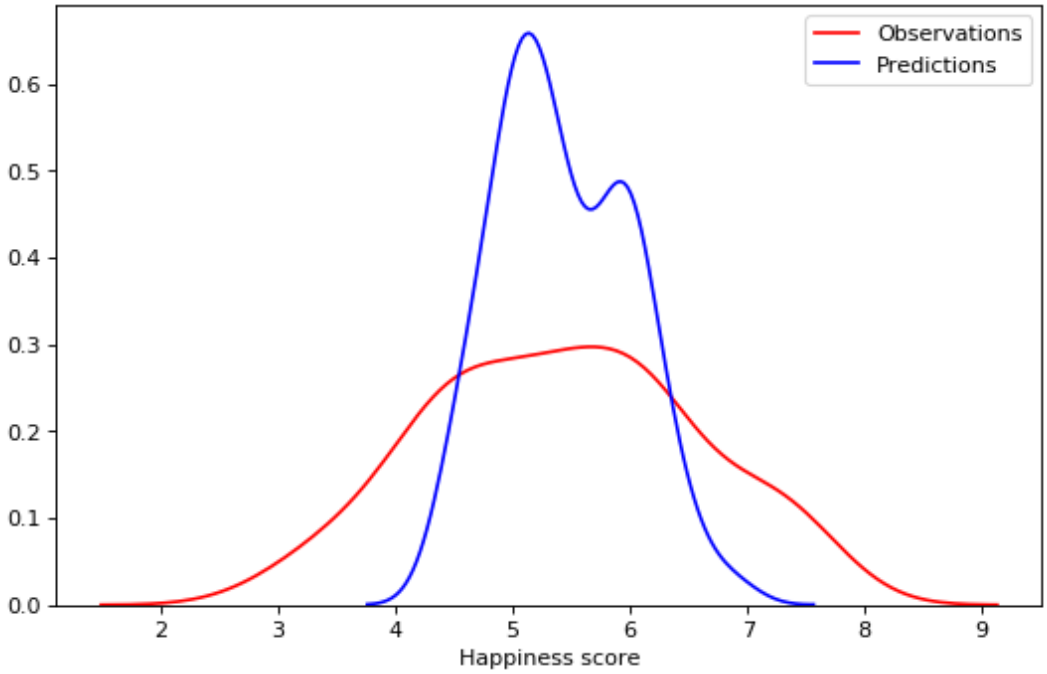
This is expected to work fairly poorly. Here, it is not the intention to build a model that predicts happiness of new countries, and splitting into train / test sets, cross-validation for parameter tuning do not make much sense. The simple 2d multilinear regression model only serves as a simple approach to judge if additional dependent variables are needed to explain the observed response variable happiness.



Again, these two plots in a way reproduce the decent negative correlation of temperature and happiness, and also show that the model assumes a small positive impact of the precipitation. It is apparent that very few countries have a very high precipitation or a very low temperature. Either way, the

temperature clearly is the major indicator of happiness among the two.

How strongly predictive are the two climate factors temperature and precipitation? Previously we noticed that the model produces a relatively tight band of predictions, i.e. extreme values of happiness are not predicted or are predicted poorly. Let us take a better look at this fact by plotting the individual histograms of the observations, and the predictions.



At last, we determine the explained variance of our model.

Explained variance of model: 0.24.

This is clearly not a high value. The considered values in the report as health, life expectancy, gdp, income, etc. allow for much better models.

IV. Results

1a) Exemplary happy countries
Finland, Denmark, Norway, Iceland, and the Netherlands are the world's happiest countries. The United Kingdom follows on 15th , Germany on the 17th place, and The United States of America follow on the 19th place.

1b) Exemplary unhappy countries
Yemen, Rwanda, Aghanistan, the Central African Republic, and South Sudan are the world's unhappiest countries.

2) Countries with high temperature
In general, it can be observed that equatorial regions are warmer than polar regions.

3) Countries with high precipitation
In general, the northern part of South America, certain west coast countries in Africa, and South East Asia, particularly the Phillipines, are the regions with the highest precipitation.

4) Correlations
There is a decent negative correlation between temperature and happiness (c=-0.46). Precipitation and happiness are uncorrelated (c=0.03).

5) Multilinear Regression
The scatter plots confirmed the negative trend of happiness over temperature, while the slightly positive trend of happiness over precipitation is negligible. The model appears to predict within a tight band, i.e. extreme happiness values cannot be predicted.

6) Histograms and explained variance
The histograms explicitly show the mentioned tightness of the predictions by the regression model. The value of explained variance is very low, exp_var = 0.24.

V. Discussion

1a) It is not surprising that many European and North American countries belong to the happiest countries. Advances in technology and long periods of peace have allowed the population to live happy lives overall. It is interesting to note that the absolute happiest countries belong to the Nordic countries in Europe. This might indicate that GDP of a country constitutes less to overall happiness than e.g. social programs of the state and an environemnt rich of seas and trees.

1b) The mentioned five countries in Africa and Southern / Western Asia are regions of hunger, heat, and war. Therefore the bad results on the happiness scale are not surprising.

2) This is a well known effect explained by higher radiation due to different sunlight angles.

3) These observations are much harder to explain. In general coastal and island regions seem to benefit due to surrounding water. Furthermore, ocean currents like the Gulf Stream and the Noth Atlantic Drift, as well as monsoons, and typhoons will have significant impact on annual precipitation in affected areas.

4) A possible explanation is the following: in the rich countries water supply is not a problem due to advanced pipe systems while in the poor countries humans settle at places where water supply is a solvable problem most of the time, i.e. close to coast lines, rivers, or seas. The issue of extreme heat on the other hand leads to many heat deaths every year, even in the rich countries, and certainly impacts overall comfort and happiness.

5) The observations regarding impact of temperature and precipitation resonate with the findings of the correlation analysis. The tightness of the prediction band can be explained with the low dimensionality of the model. It is only a rough explanation of what is observed, and without considering factors like GDP, health, or income, the general happiness cannot be explained.

6) The notion of tightness is substantiated by the histogram plot. The low value of explained variance supports the need for further variables and the existence of a much better non-linear regression seems very unlikely.

VI. Conclusion

In this project we have collected different data sets about happiness, temperature, and precipitation of in total 133 countries. We merged the data and visualized the three variables on a world choropleth map using folium.

We then investigated the correlations. We found that temperature and happiness show a negative correlation of c=-0.46 while precipitation and happiness show no explicit correlation.

We built a linear regression model of the two independent variables temperature and precipitation, and happiness as the dependent variable. We used scatterplots to compare the predictions with the observations. It was apparent that the model was performing poor in general and only reproduced rough trends - extreme values of happiness could not be predicted and many outliers were visible.

We confirmed this assessment by comparing histograms of observations and predictions, and by calculating the explained variance, which resulted in a poor value of exp_var = 0.24.

It is not to be expected that there is a hidden non-linear connection between temperature, precipitation, and happiness. Instead different factors like health, life expectancy, gdp, income, etc. are needed to produce better regression models.

The goal of the project was not to find the best regression model but to understand if temperature and precipitation have an impact at all if it comes to happiness of a country.
The data analysis suggests that there is indeed a negative correlation between happiness and temperature, while the effect of precipitation is negligible.

A possible explanation is the following: in the rich countries water supply is not a problem due to advanced pipe systems while in the poor countries humans settle at places where water supply is a solvable problem most of the time, i.e. close to coast lines, rivers, or seas. The issue of extreme heat on the other hand leads to many heat deaths every year, even in the rich countries. Furthermore, heat strongly impacts people who need to travel to work and when surpassing a certain threshold is anything but comfortable. In Spain and other countries, during summer time it is normal for people to stop working in the early afternoon for 2-3 hours and to take a long nap instead, since other actions are not possible in the heat.

Considering the global warming, humanity might need to come up with new ideas to make living in heat a bit more comfortable and less threatening.