Project D-19: Kaggle World Happiness Report

Kadi Sammul, Ellen Leib, Robert Leht

Repository: https://github.com/Sabasik/IDS-2020

BUSINESS UNDERSTANDING

Our business goals

Our choice of topic is not influenced by a business standpoint, rather by an interest to learn more about the direction that our world is going in emotionally and psychologically, as we are still young people and our generation is approaching working age. We would like to learn, where in the world help is most needed and what could be done to raise the happiness score of those countries. By finding the most important contributors to the happiness score, the best solutions can be found.

The results of this project could play an important role for everyday users of the internet. With more information, people are becoming more knowledgeable about the dangers that come with the internet and the abundance of it. If this research project were to showcase a well-reasoned correlation between internet usage and lack of happiness, this could potentially make people pay more attention to their internet usage. In order for this to happen, the connection between the two features would have to be justified and visualised extremely well, otherwise it could easily be considered an accidental correlation.

We can also find correlations between happiness and wealth of a country, which is something even a government could use to assess the situation of its well-being. Our objective with this project is not to earn money or sell the results to anyone, but to learn an interesting thing or two and raise awareness in case it needs to be raised.

Assessing our situation

Our two datasets are fairly small. The internet usage dataset only holds data from one year time and that will probably make finding a proper correlation very difficult. The world happiness dataset also holds only six years worth of data about the population’s happiness and that is not much to work around with. We could assume that internet usage has grown since the time of making that database and was lower before the making of it, but that already diminishes the trustworthiness of our research. Since we’re working with all countries in the world, we will have to make judgements based off of their relative sizes and also their geographical, historical and economical statuses, which will again make the research more biased and not as objective as it ideally should be. If we had immense amounts of data from many more years, this would be a lesser problem, however the internet really has only been around for about 15 years and even less in many places in the world, so either way it would be difficult to make a fair judgement.

Working with the happiness dataset alone will likely show more promise as there is more data to assess and comparisons can be made throughout the given years. All of the data will have to be handled relative to each other and some sort of exceptions regarding some countries like city-states and developing countries will probably have to be made in order to balance the data a bit more. However, since the datasets are not big, everything will be easily calculable and should not require external resources.

We are doing all of these calculations and operations on our school laptops, using our home wireless networks as we are in quarantine, which will definitely impact the effectiveness of the project, because we can't physically work together, however we can use different platforms to communicate via voice and text messages. We all have very similar backgrounds in data science, so if one of our teammates were to become ill, the others should be able to do their part collectively. Our project fortunately does not include complicated terminology, but there will definitely be a couple of features in our datasets we will have to learn more about, like GDP for example. Since our project is done from home, the costs and benefits can't really be measured.

Our data-mining goals

We hope to find a strong correlation between features and the happiness level amongst the countries listed. We also wish to investigate the trends of the happiness score through the years to be able to assess whether the score is improving, worsening or steady. We will train a model to predict a happiness score when given other features in the dataset, such as level of social support, GDP, life expectancy etc. We would like to have an accuracy of at least 0.9.

Visualisations of the data will also be highly helpful as this will help to understand the proportions of the data better. Our goal is also to determine whether there is a correlation between internet usage and the country's happiness score and consequently, if it makes sense and isn't accidental. We will be making these assessments together.

DATA UNDERSTANDING

Gathering data

We mainly wanted data that would have the happiness score of different countries as well as some other describing factors about each country (i.e. life expectancy, GDP, social support, ...). The data had to be about several years. We also wanted some data about (mostly) the same countries about the same time period, so we could see whether there exist some interesting correlations (we chose internet usage, because it seemed the most likely thing to be tied to the happiness score) between them so that if we found any, we could try to see whether we can determine the source of the correlation. We wanted the data to be in csv format, as we have used that a lot in the previous weeks.

We found this sort of data available in Kaggle for all to use (https://www.kaggle.com/mathurinache/world-happiness-report, https://www.kaggle.com/tanuprabhu/list-of-countries-by-number-of-internet-users). The first dataset covers the years 2015-2020. The second dataset is about 2018, so we can only do that part of our project based on that one year, and therefore we must be careful to not make any rash assumptions about correlation that may seem to exist, but actually do not.

We intend to use all the files in these datasets, as there are not many of them. Since the internet usage dataset is very small we expect to use all the columns in it, although only the countries that are also in the world happiness datasets (the internet usage has ~50 more countries in it than happiness). From the happiness dataset we intend to use all the countries, but only the factors that are covered in each year, as some years have more factors brought out than others.

Describing data

The first dataset contains information from 2015-2020. It covers 153 (the actual number varies in each year) countries, containing each country's happiness score and other factors, the number of different factors varying each year from 7 - 18. In each year there should be factors: GDP per capita, life expectancy, freedom to make life choices, generosity, corruption perception. We intend to use the columns that are in all the years, so we have consistent data.

The second dataset contains 215 countries with the number of internet users, population, the population rank (how big the population is compared to the other countries in this dataset), percentage of internet users, and the users rank (how many internet users there are compared to the other countries in this dataset) from 2018. With this dataset we try to see whether there is a correlation between this and the previous dataset's year 2018, but we are aware that with this limited data we may not be able to verify the existence of correlation even if we seem to find one, as the seeming correlation may be generated by some factors that are not covered in either of these datasets, or may simply be accidental in the given year.

Exploring data

The data in internet usage dataset is mainly as strings, so to use it we will need to convert it to numerical, as well as remove random excessive data from other years (for example in the percentage column). Most of the countries have internet users up to 8000000, with less countries having even more. The population is distributed mostly in the same way, with most countries having populations up to 40 000 000 and less over that. The percentage of internet users varies more with values from 1 to 99%

The world happiness dataset is better in the sense that columns with number values are actually numeric types, not strings with numbers. The column names however are slightly different in each year file as well as the number of columns with the data given are different. There we must remove the columns that are not present in each file and rename them all so the data is presented in the same way in each case. In different years there are also different ranges on the same data in some cases. For example the life expectancy is in most cases a value around 0 and 1, but in 2020 it is up to 76.

Verifying data quality

Our data exists and we had no problems with accessing it. There are some minor issues that can be eliminated, for example from the numbers that are given as strings we can remove the commas and convert them to string as well as remove excessive columns and rename them if necessary to make sure the data is styled the same in different years. We also must make sure the data is presented with the same values/in the same ranges each year, as there are some occurrences of the same data presented differently in different years.

Overall we have some minor issues with the data that we believe we can overcome and therefore are able to use it as intended for fulfilling our goals in our project.

PLANNING

Plan

1. Clean and adjust the data so all dataset would have the same columns, all values would be numerical and in the same scale, research about the data. (10 h - Kadi)
2. Analyze and investigate the trends of the happiness score over the years. (about 30 h- Ellen)
    a. Make one big dataset from all the yearly datasets.
    b. Find all the correlations between all the features and the happiness score.
        i. Find the biggest correlation.
        ii. Analyse the reasons
    c. Illustrate the data on different plots
        i. Make maps from different years' scores.
        ii. Illustrate the overall trend over the years.
        iii. Make a map of changes to the happiness score over the years.
    d. Find the countries with the biggest change in happiness over these years
        i. Analyse the reasons
3. Train a model that predicts a country's happiness score using GDP, life expectancy, social support etc. (about 30 h- Robert)
    a. Try different machine learning algorithms and find the one with the biggest accuracy/precision/recall for the entire world.
        i. Decision trees
        ii. Random forests
        iii. Regression models (Lasso, Ridge etc)
        iv. Also try different parameters for all of these methods.
    b. Train a model based on specific countries (for example Estonia) to predict happiness in the future.

        i. Test different algorithms.

4. Investigate the correlation between internet usage and the happiness score in the world (and also maybe some other factors). (20 h- Kadi)
    a. Make a map of internet usage in the world.
    b. Bring the two datasets together so that they would be compatible.
    c. Find the correlation between internet usage and the happiness score.
        i. Analyse the result.
    d. Research some other factors that could affect the happiness score of a country.
        i. Find the correlations.
5. Summarize our work, make it presentable. (Everybody)