For analyzing the factors that influence world happiness, we found several datasets on Kaggle. Each dataset shows the ranking of the happiness score of each country and indicates those countries’ GDP per capita, social support, freedom score, generosity score and some other features that may influence people’s happiness. We have the happiness report from 2015-2019. The happiness score and ranking used data from Gallup World Poll. The poll asked each respondent to select a level from 0-10 as the score that him/she estimate his/her happiness and give a score on economic production, freedom, generosity… for how much he/she think this factor contributes to his/her happiness.

There are also scores other than the main factors that we considered for world happiness (family, freedom…), for example in 2015’s report, each country has an evaluation on “Dystopia Residual”, which is an “is an imaginary country that has the world’s least-happy people”. Dystopia is the opposite word of Utopia. However, not all the reports contain the score for dystopia. There are also some factors like “Whisker.high” and “Whisker.low” in 2017 that does not have corresponding variables in 2018, so we are going to delete those columns in the cleaning steps.

The research subjects and contents for world happiness report are not exactly the same in each year, so there are some differences between the countries and the variables of each table. For example, there were 156 countries in the table of 2019, and 153 countries in the table of 2020. The variables for each table are also not consistent, especially the difference of names for one feature. For example, in 2015, happiness score is written as “Happiness Score”, and in 2019 it is written as “Score”, while in 2020, it is written as “Ladder score”. Also, there are different names for health scores, freedom scores and other features. What’s more, some tables contain a few more variables than others, like 2015 has the score for family, but in 2020 happiness score does not count family. Instead, it has “upperwhisker” and “lowerwhisker” as features.

In order to see the happiness trend between years, we have to make the variables consistent. First, we have to make sure that all the country names are written in the same style as it is easy to compare and contrast. For example, some tables use “Trinidad and Tobago”, and the others use “Trinidad & Tobago”. We select the country column from each table and paste them into a new file. We checked and revised the disputable country names directly in Excel. Then, we use python to contrast the set of names in each table and delete countries that are not researched for all 6 years. Finally we remain 146 countries that are useful. For the features, we choose to use the concise way for naming, like changing “Healthy life expectancy” to “health”, “Freedom to make life choices” to “freedom”.