SI618 project1
HAPPINESS AND ECONOMIC FREEDOM
Yumin Zhang

1. Motivation

People's happiness is probably the most pursued target of all countries, societies and human beings. The factors that leads to people's happiness varies, including education, literacy, health, government stability, but probably most importantly, wealth level. Generally speaking, the freedom and ability of consumption will further impacts the ones' moods and sentiments, and determines their degree of happiness. If we express the idea of wealth in a macro angle, from society and country standard, it is the economic condition. It is reasonable to assume that the country with more economic freedom will has its citizens with higher level of happiness. This report attempts to explore more detailed the relationship between world happiness and world economic freedom. To be more specific, we will try to answer the following questions

- What is the region with highest happiness scores?
- What is the relationship between happiness and economic freedom?
- How government behaviors impact nation happiness

2. Data Sources

Economic Freedom of the World https://www.kaggle.com/gsutters/economic-freedom
This dataset measures the degree to which the policies and institutions of countries are supportive of economic freedom worldwide. It includes data of 162 countries over 1970 to 2016. The datasets are *csv* files. the economic freedom level is measured from a score on a scale of 0 to 10. it also investigates numerous factors that affects economic freedom, which can be divided into 5 major categories, size of government, legal system and property rights, sound money, freedom to trade internationally and regulation, which are all 10 scale. Size of government contains government spending, taxation, and the size of government-controlled enterprises. Legal system and property rights focus on judicial independence, legal enforcement and contracts and police system, which correlates with property protection. Sound money is about the various matrices measuring inflation. Freedom to trade internationally records situation of international buying, selling and making contracts. Regulation is on government's limitation on exchange, gain credit, hire and work.

World Happiness Report https://www.kaggle.com/unsdsn/world-happiness
The World Happiness Report is a landmark surveys the state of global happiness, ranking 155 countries by their happiness levels. The data is over 2015, 2016 and 2017. The datasets are *csv* files. It integrates across field factors - economics, psychology, national statistics, health, public policy - to evaluate the overall well-being with the progress of nations. The happiness scores and rankings use data from the Gallup World Poll, based on responses from main life evaluation from the poll. The evaluation question is Cantril ladder, which asks the respondent to imagine a best possible life for them being 10, and the worst being 0, and rate their own lives on that scale. The

columns following happiness score are respectively economy, family, health, freedom, trust, generosity and dystopia residual. These following factors do not correlate with happiness scores, but partly explain the scores difference.

3. Manipulation

Step1: Interesting Part Filtering

The datasets include data over 160 countries, 30 years and 30 different columns. We will not be able to analyze all of the data, so we will figure out which part of the data we are interested in, and filter them out with data manipulation methods. From a time perspective, we find out that World Happiness Report contains data for 2015, 2016, 2017, and Economic Freedom of the World includes data from 1970 to 2016, but with 2015 data missing, so we will filter out the part which is close, 2015 – 17 for the former and 2016 for the latter. For the World Happiness Report, we will only filter out columns country, region, happiness score. For Economic Freedom of the World, we will filter out size of government, legal system integrity, and regulation as most representative columns.

Step 2: Missing Data Handling

The Economic Freedom of the World dataset and World Happiness Report dataset is actually quite complete, thus the missing data handling is simple and straightforward. For missing data handling, only the part of data we are interested at requires processing. For example, in the Economic Freedom of the World dataset, the government size column of Angola at 1970 is missing, because Angola government was not founded at the time. The police system reliability data of Belarus at 2016 is missing, because Belarus is not open to survey for the data. However, we will not process them, because we will not research on data at 1970, and police system reliability of nations. For parts we are interested at, they are all score on a scale of 0 to 10. we will assign a 0 score to the missing part, since they are not recorded.

Step 3: Table Header Processing

In order to make *mrjob*, *Spark* and *SparkSQL* processing more convenient, we will process the table header of the datasets. The datasets are all *csv* files, which use commas to split different columns. In *SparkSQL*, due to grammar reasons, we would prefer column names to be represented in single word, so we will connect all those multiple column names with underline, like transforming “Happiness Score” into “Happiness_Score”. In *mrjob*, the algorithm will process the whole input files, so we will delete the table header line and make a new *csv* file for *mrjob* to process.

Step 4: Merging Datasets

In this project, 4 tables are included, which are all *csv* files. They are 2015.csv, 2016.csv, 2017.csv from World Happiness Report, and efw_cc.csv from Economic Freedom of the World dataset, which contains data over all time periods on one file. The merging datasets work varies for different tasks in the project, and are realize with different tools. The second task is conducted by *Spark*, the datasets included are 2016.csv and efw_cc.csv files. They have the same variables 'country', which will be used as the key to join the two datasets. The joining work are realized with *join* function in *Spark*. The third task is completed with *SparkSQL*, the datasets included are 2015.csv, 2016.csv, 2017.csv and efw_cc.csv. They are also joint with the key variables 'country', and are realized with 'on' attributes from 'select' function.

4. Large-scale Computation Tasks

Task 1: 'Happiness Points' for Every Region

In task 1, we will find out the 'happiest' region in the world, and degree of 'happiness' of respective regions. The regions are divided according to geographical locations - Western Europe, North America, South Asia and etc. The happiness score of nations are presented on a 0 to 10 scale, and we will make a simple classification - classify nations which score 6 or higher as 'Happy', and lower than 6 as 'Unhappy'. For every regions, every 'Happy' nation will add 1 to 'Happiness Points', and 'Unhappy' nation minus 1 from the points. By doing so, we will find out the degree of 'happiness' for every region.

Task 1 is realized with *mrjob*. In the mapper, for every line - each line represents a country - we judge whether the happiness score is higher than or equal to 6. If true, we will yield a pair of region name and number 1, otherwise we yield a pair of region name and minus 1. in combiner and reducer, we integrate all the pairs according to region name, add sum up their points with the same name. By doing so, we can generate a table of 'Happiness Points' for every region, as shown in Table. 1.

Table 1: Happiness Points of Every Region

Australia and New Zealand	2
Central and Eastern Europe	-25
Eastern Asia	-4
Latin America and Caribbean	4
Middle East and Northern Africa	-5
North America	2
Southeastern Asia	-3
Southern Asia	-7
Sub-Saharan Africa	-38
Western Europe	11

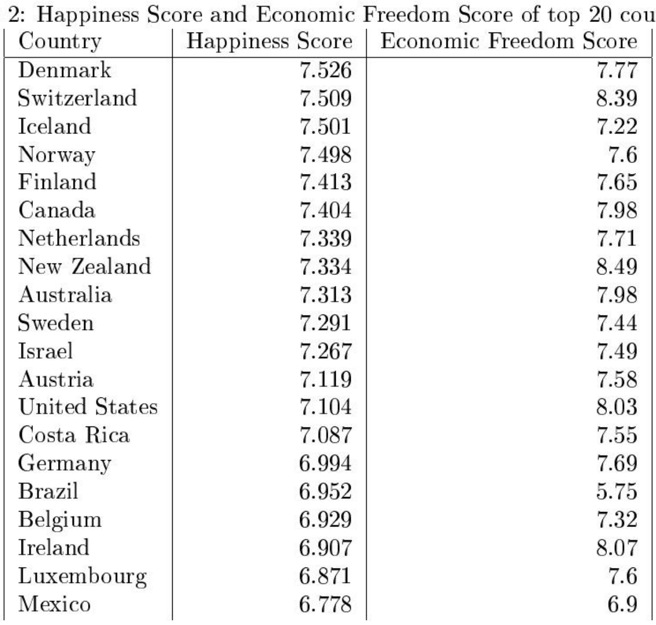
From the result table, we can clearly see that Western Europe is the happiest region throughout the world. This is consistent with common sense, since Western Europe has countries like Iceland, Denmark, Norway, which are among the most developed countries, with citizens enjoying decent living conditions, high welfare, small population and thus fewer social problems and beautiful natural scenes. Eastern and Southeastern Asia, accumulating most of population on the earth, are not very happy. They only get about -3 points, suggesting people in China, Japan and South Korea live in high pressure. The most unhappy regions in the world is Sub-Saharan Africa, with a terrible -38 points. This is reasonable, since Sub-Saharan Africa is the poorest regions suffering from wars and famine. This shows that the international community should unite to help people in Central and Eastern Europe and Sub-Saharan Africa to achieve a happy life.

Task 2: Happiness vs. Economic Freedom

It is normally assumed that nations with higher economic freedom will have their citizens enjoy a happier life. In Task 2 we will explore what the realistic situations look like. We will compare the happiness scores of each countries versus their economic freedom scores. Task 2 is conducted with *Spark*.

We will focus on data of 2016, since this is the data which World Happiness Report and Economic Freedom of the World dataset both have. We extract the columns ‘year’, ‘country’ and ‘economic freedom score’ from the *efw_cc.csv*, and then filter out the data from 2016. Then we join the data to happiness dataset according to country, and present the resulting data in descending order with respect to happiness score. The data of top 20 countries is shown in Table. 2.

Table 2: Happiness Score and Economic Freedom Score of top 20 countries



Country	Happiness Score	Economic Freedom Score
Denmark	7.526	7.77
Switzerland	7.509	8.39
Iceland	7.501	7.22
Norway	7.498	7.6
Finland	7.413	7.65
Canada	7.404	7.98
Netherlands	7.339	7.71
New Zealand	7.334	8.49
Australia	7.313	7.98
Sweden	7.291	7.44
Israel	7.267	7.49
Austria	7.119	7.58
United States	7.104	8.03
Costa Rica	7.087	7.55
Germany	6.994	7.69
Brazil	6.952	5.75
Belgium	6.929	7.32
Ireland	6.907	8.07
Luxembourg	6.871	7.6
Mexico	6.778	6.9

From Table. 2, we can clearly see that the countries with high happiness scores also have very high economic scores, so the common sense is correct. But the data presented here is relatively small size, and can not ahead to very informative conclusions, so we will conduct a visualization, showing happiness score and economic freedom score of all countries with scatter plot, as in Fig. 1.

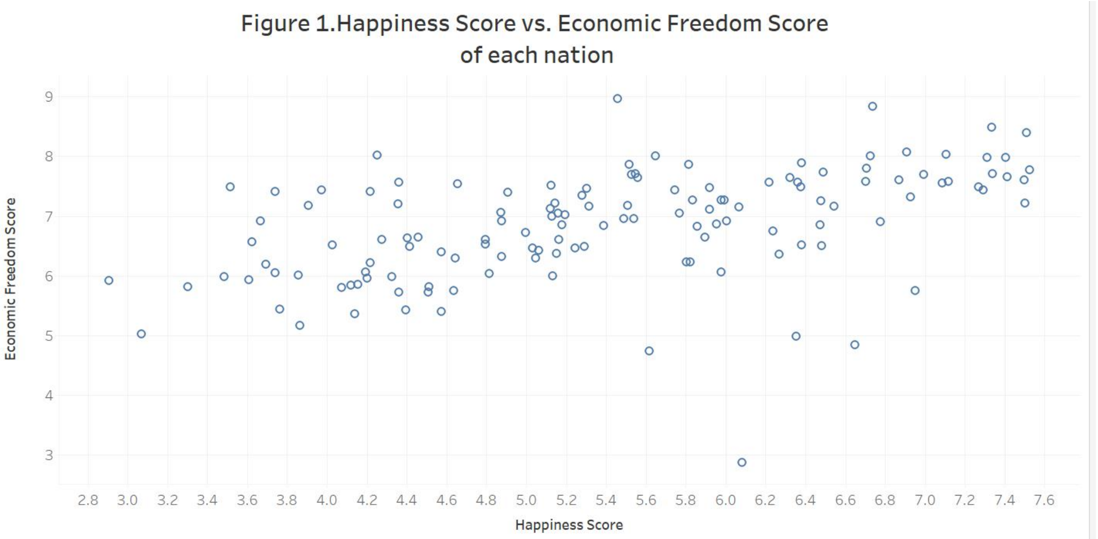


Figure 1.Happiness Score vs. Economic Freedom Score of each nation

On Fig. 1, we can draw constructive conclusions. Happiness score is indeed positively related to the economic freedom score, since the points on the graph fit to an increasing line. But we should also notice that, the increase trend is very slight. For each distinct happiness score, the economic freedom score can vary drastically. Happiness score varies on a scale significantly larger than that of economic freedom. This result implies the complexity of happiness. On one hand, improving the wealth status do impact people's level of happiness, and is probably the most important means. On the other hand, merely caring about economic issues is far from adequate. People living a rich life can also be a great pain. We should also pay attention to other factors, like environment, education, social justice, and so on, to make people live a happy life.

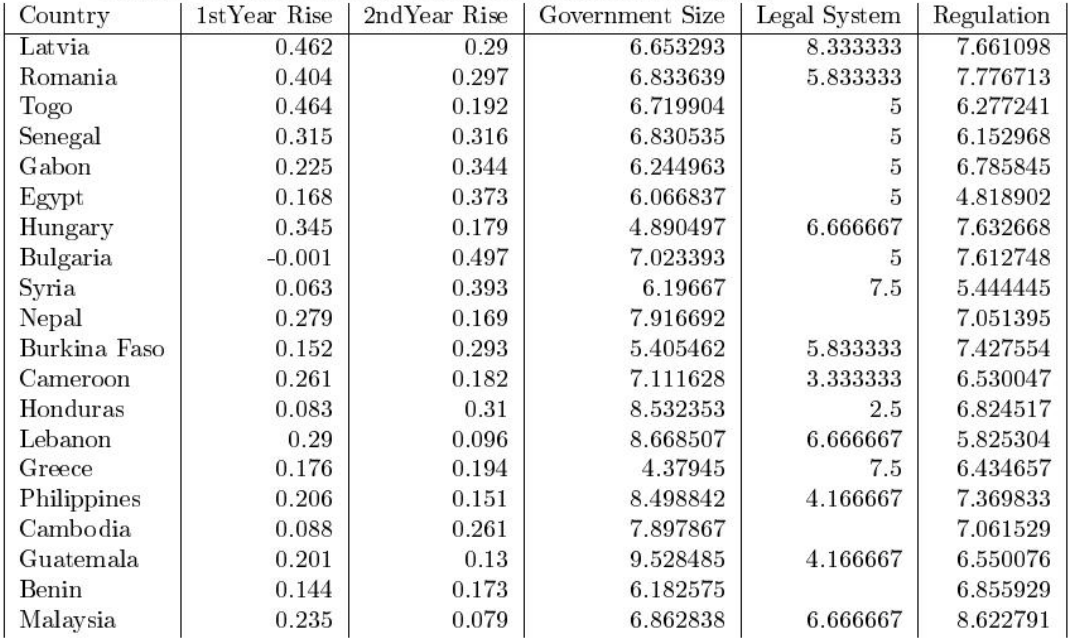
Task 3: Government's Role in Making a Happy Life

How to change the current situation and make more people's life better is an issue most people concern with. One of the most important characters in this regard is government. Government is the main controller of economic freedom, and can also release laws to impact other factors correlates with people's happiness. Governments of all the nations have the target of making people's life happy. In task 3, we will explore government's actual rule in doing so.

Task 3 will be conducted with *SparkSQL*. In Task 3, we will observe how each nation's happiness score changes from 2015 to 2016 to 2017, and we will determine the countries with highest score increasing in these three years. We will then evaluate their governments behavior which result in such increase. Government behaviors are assessed from three aspects – government size, legal system integrity, and regulation. We will observe how the nations with highest increasing perform in these three aspects.

We extract variables of country and happiness scores from 2015, 2016 and 2017, and join them to compute the scores change. We then extract country, government size, legal system integrity and regulation from economic freedom dataset, and join them to the former tables according to nations. The final results are presented with descending order of mean of two year scores increase. The data are presented in Table. 3 with the top 20 countries.

Table 3: Happiness Scores Increase & Government Behaviors



Country	1stYear Rise	2ndYear Rise	Government Size	Legal System	Regulation
Latvia	0.462	0.29	6.653293	8.333333	7.661098
Romania	0.404	0.297	6.833639	5.833333	7.776713
Togo	0.464	0.192	6.719904	5	6.277241
Senegal	0.315	0.316	6.830535	5	6.152968
Gabon	0.225	0.344	6.244963	5	6.785845
Egypt	0.168	0.373	6.066837	5	4.818902
Hungary	0.345	0.179	4.890497	6.666667	7.632668
Bulgaria	-0.001	0.497	7.023393	5	7.612748
Syria	0.063	0.393	6.19667	7.5	5.444445
Nepal	0.279	0.169	7.916692		7.051395
Burkina Faso	0.152	0.293	5.405462	5.833333	7.427554
Cameroon	0.261	0.182	7.111628	3.333333	6.530047
Honduras	0.083	0.31	8.532353	2.5	6.824517
Lebanon	0.29	0.096	8.668507	6.666667	5.825304
Greece	0.176	0.194	4.37945	7.5	6.434657
Philippines	0.206	0.151	8.498842	4.166667	7.369833
Cambodia	0.088	0.261	7.897867		7.061529
Guatemala	0.201	0.13	9.528485	4.166667	6.550076
Benin	0.144	0.173	6.182575		6.855929
Malaysia	0.235	0.079	6.862838	6.666667	8.622791

From the result, we can find some commonality between these countries with very high happiness increase. For all of the government size, legal system and regulation variables, these countries acquire a score of about 6 to 7. These scores are between medium and upper, high enough, but not too high (compared to big countries like UK and France). This result can give us a clue on how to improve citizens happiness – the government should stay strong on regulation and ruling, which can help to solve some social problems. But they should not be too aggressive, which may apply a sense of urgency to the citizens.

5. Challenge

In this project, the first problem I meet is data selection. There are various datasets in Kaggle, but trying to find out two datasets which are related and contain information of different fields so I can generate informative conclusions are very limited. I also need to consider about data processing issues, like the size, columns, format of the data, which is very challenging. The datasets I use are all csv files, which are different from the json files we use in class for *Spark*, and *SparkSQL*. This creates a lot of problems in data loading. I also find *mrjob* is very hard to use for computation and sorting. But finally, I overcome all these problems, and learn a lot from the project.