Capstone Final Project – OECD Better Life Index Study

I. Introduction

What makes life happy? This is a question people often ask. It should also be a major concern of economists of countries in the world, since economists give advices to policymakers and help them in making decisions. Traditionally, economists use some economic indicators such as GDP and CPI to measure the economic health of one country. However, criticisms rise towards such practice, main concern being if these statistics can really give a true account of people's current and future well-being. Afterall, the ultimate goal of economic and social development is to promote people's well-being. Thus, this project is motivated by a comparison of "true" life level of different countries. In this study, I would like to use some indicators more directly associated with the happiness of the people in different countries, leveraged by the Foursquare location data, to answer the two following questions: First, which aspects are more important for happiness of life ? Second, does the venue clusters share common characters with the happiness indicators cluster among different countries? The answers of these questions will be helpful for meaningful for economists to get more holistic decryption of one country and help policymaker in their policy design.

II. Data

II.1. Data description

This study will use " OECD Better Life Index" combined with Foursquare location data. Since May 2011, OECD (Organisation for Economic Cooperation and Development) begun to publish

" OECD Better Life Index". They attempt to address the criticisms mentioned above and to bring together internationally comparable measures of well-being in line with the recommendations of the Commission on the Measurement of Economic Performance and Social Progress. These indicators covers 11 topics which OECD has identified as "*essential to well-being in terms of material living conditions (housing, income, jobs) and quality of life (community, education, environment, governance, health, life satisfaction, safety and work-life balance)*."[1] Each topic is reflected by one to four specific indicators. Currently, the index cover 35 members countries of OECD, including most of the world's developed economies and a few emerging countries. The data can be downloaded from OECD's website (https://stats.oecd.org/Index.aspx?DataSetCode=BLI). OECD Better Life Index allows us to answer the first question. To see the venues in these countries and answer the second question, we will rely on Foursquare data.

II.2. Data preparation - Format reshaping

The data is downloaded from OECD's website as mentioned above in a csv format with a dimension of 2369*17. One issue should be addressed before exploring the data. The data is organized in "a list" such as countries and variables are stacked one after another in one column. While a format with country in rows and variables in columns will be more convenient for our purpose. Besides, some columns in the original data describe data attributes such as 'Measure', 'Inequality', 'PowerCode', 'Reference Period Code', 'Reference Period', 'Flag Codes' and 'Flags'. Most of them are useless therefore we delete them. We retain only 'Country', 'Indicator', 'Unit', 'Value' in our data set before data transformation. The format reshaping is done with the pivot_table function in the pandas. The data is then transformed into a 41 * 25 format, in the other words, with 41 countries and 23 indicators (the other 2 columns are country name and units to tell us if the value is a percentage or number).

II.3. Countries and variables selection

[1] Explanation from OECD, http://www.oecdbetterlifeindex.org/about/better-life-initiative/.

Before proceeding the analyze, another classical issue is dealing with the missing value. By examining the data, we notice there are 15 variables with missing values. For dealing with the missing value, we choose dropping them instead of replacing them with average value to keep the nature of the data. However, since our objective is to capture more aspects of life, we will drop some countries with too many missing values as well for keeping more variables of interest.

Concretely, the variables with the most missing values are 'Household net adjusted disposable income', 'Household net wealth', 'Labour market insecurity', 'Personal earnings' and 'Time devoted to leisure and personal care'. On the other hands, to decide which variable we can delete, we examine the correlations among the variables as well. The variables with the highest coefficient is Personal Earnings. We decide to keep it. While 'Household net adjusted disposable income' has also high correlation coefficient, we remove this variable considering it describe income aspect and it can be proxied by 'Personal earnings'. We discard all the other variables as well. An interesting remark is 'Personal earnings' is highly correlated with happiness but not 'Household net wealth'. Could we say “Job makes happy but not wealth”? Such a conclusion deserve further research and is out of the scope of this project. We then delete all the countries with too many missing values. At the end, we keep 19 variables (excluding country name and unit information) and 30 countries in our dataset.

lll.Methodology

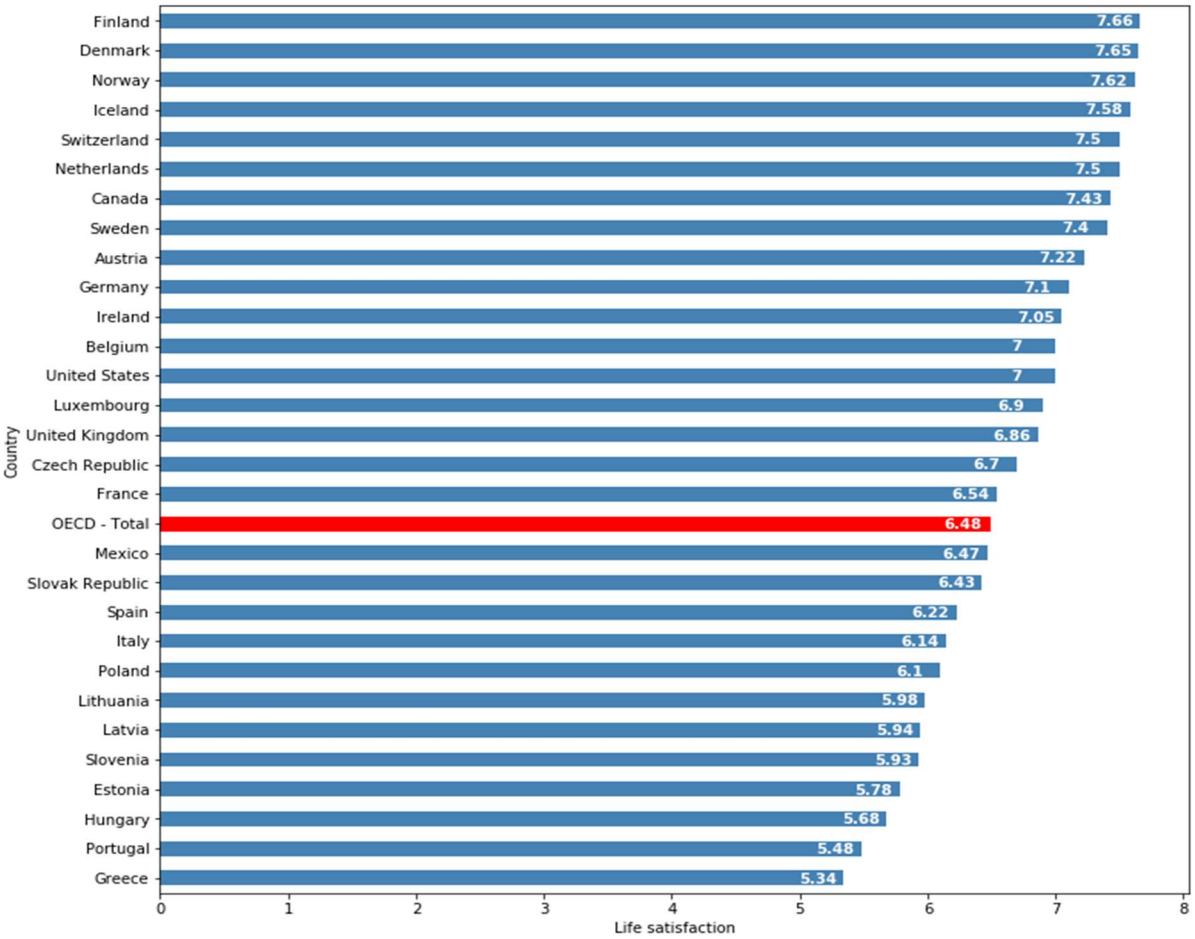
We choose “Life satisfaction” indicator as a proxy of life satisfaction level. People are asked to evaluate their life as a whole on a scale from 0 to 10 rather than their current feelings in OECD survey for this variable. Although it’s subjective, but it remains a good candidate since we are interested in measuring life satisfaction and happiness. We will use linear regression to explain the variation of this variable. The preference of linear regression over polynomial model is motivated by the limited number of observation and we want to avoid overfitting problem. We don’t chose logistic regression neither because treating a quantitative dependent variable as categorical will lose information.

Once Life satisfaction explained, we would like to find similarity among countries. Since neither hierarchical structure nor density-based clustering are required in partition, we will rely on k-means methods, which will allow us to partition these countries into groups that have similar characteristics. Two k-means clustering will be done, one relying on the OCED data and the other with Foursquare venues data. We could see if groups in the two ways share similarities.

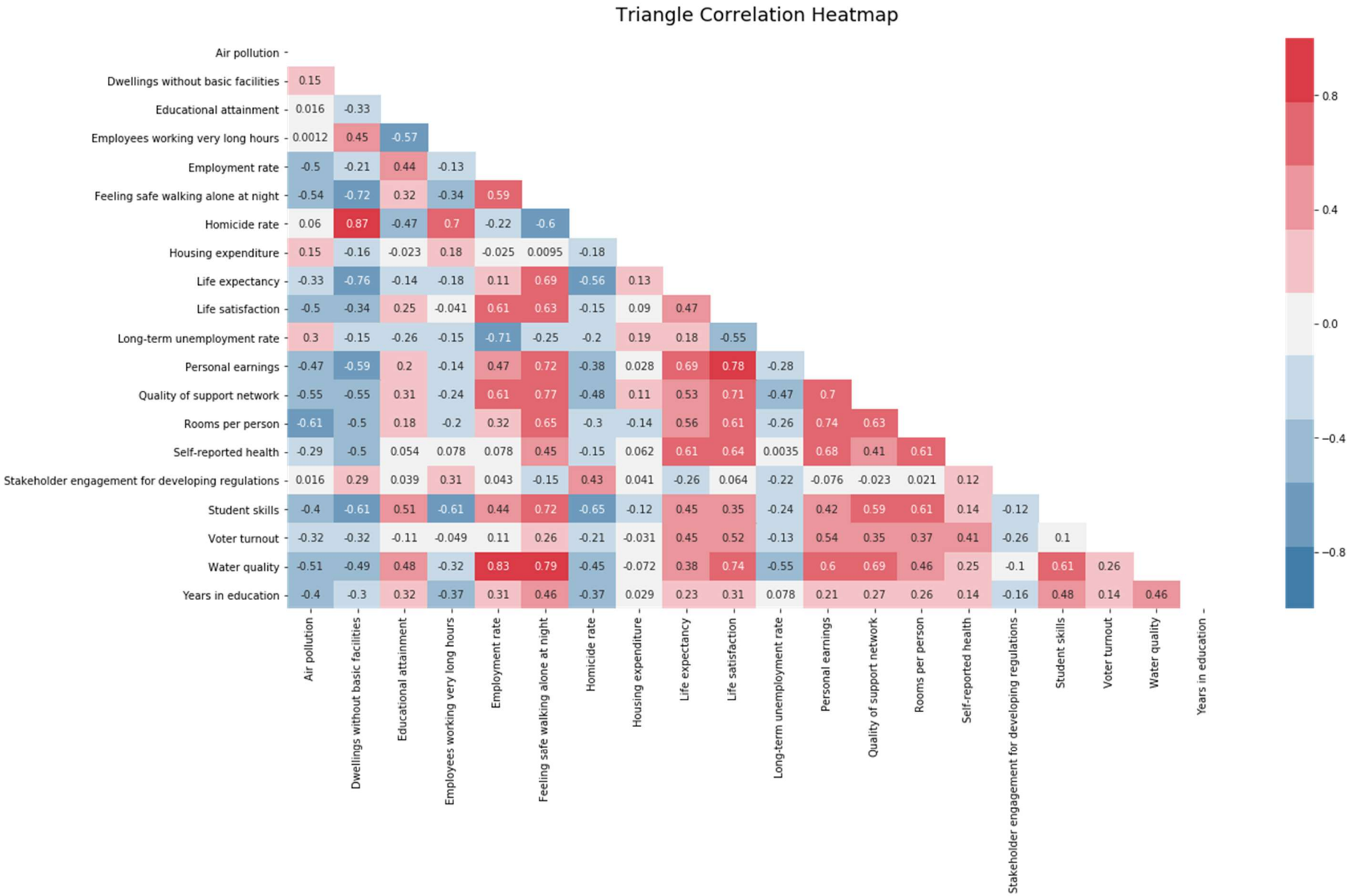
IV. Results

IV.1 Exploratory analysis

First let's look at the life satisfaction indicator across countries. As the bar plots below show, Life satisfaction is not evenly shared across the OECD countries. Scandinavians countries and Switzerland have the highest score. While Greece, Portugal, Hungary and Estonia – have a relatively low level of overall life satisfaction. On average, people across the OECD gave an evaluation of 6.49. Since we are interested in individual country, we will drop "OECD – total" row for the rest of the study.

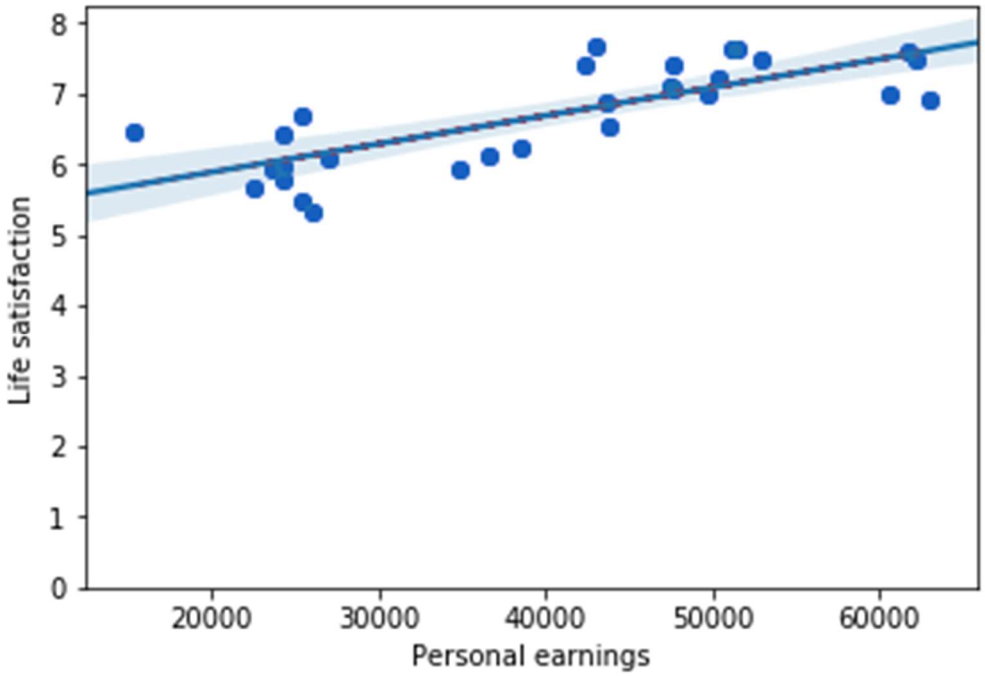


We then study the correlation relation among the variables, which are shown in the heatmap below. As we can see, the variable correlated with Life Satisfaction the most is still “Personal Earning”.


Triangle Correlation Heatmap
Air pollution
Dwellings without basic facilities
Educational attainment
Employees working very long hours
Employment rate
Feeling safe walking alone at night
Homicide rate
Housing expenditure
Life expectancy
Life satisfaction
Long-term unemployment rate
Personal earnings
Quality of support network
Rooms per person
Self-reported health
Stakeholder engagement for developing regulations
Student skills
Voter turnout
Water quality
Years in education
0.15
0.016 -0.33
0.0012 0.45 -0.57
-0.5 -0.21 0.44 -0.13
-0.54 -0.72 0.32 -0.34 0.59
0.06 0.87 -0.47 0.7 -0.22 -0.6
0.15 -0.16 -0.023 0.18 -0.025 0.0095 -0.18
-0.33 -0.76 -0.14 -0.18 0.11 0.69 -0.56 0.13
-0.5 -0.34 0.25 -0.041 0.61 0.63 -0.15 0.09 0.47
0.3 -0.15 -0.26 -0.15 -0.71 -0.25 -0.2 0.19 0.18 -0.55
-0.47 -0.59 0.2 -0.14 0.47 0.72 -0.38 0.028 0.69 0.78 -0.28
-0.55 -0.55 0.31 -0.24 0.61 0.77 -0.48 0.11 0.53 0.71 -0.47 0.7
-0.61 -0.5 0.18 -0.2 0.32 0.65 -0.3 -0.14 0.56 0.61 -0.26 0.74 0.63
-0.29 -0.5 0.054 0.078 0.078 0.45 -0.15 0.062 0.61 0.64 0.0035 0.68 0.41 0.61
0.016 0.29 0.039 0.31 0.043 -0.15 0.43 0.041 -0.26 0.064 -0.22 -0.076 -0.023 0.021 0.12
-0.4 -0.61 0.51 -0.61 0.44 0.72 -0.65 -0.12 0.45 0.35 -0.24 0.42 0.59 0.61 0.14 -0.12
-0.32 -0.32 -0.11 -0.049 0.11 0.26 -0.21 -0.031 0.45 0.52 -0.13 0.54 0.35 0.37 0.41 -0.26 0.1
-0.51 -0.49 0.48 -0.32 0.83 0.79 -0.45 -0.072 0.38 0.74 -0.55 0.6 0.69 0.46 0.25 -0.1 0.61 0.26
-0.4 -0.3 0.32 -0.37 0.31 0.46 -0.37 0.029 0.23 0.31 0.078 0.21 0.27 0.26 0.14 -0.16 0.48 0.14 0.46
0.8
0.4
0.0
−0.4
−0.8

IV.2 Linear regression results

We run the first version of regression with only “Personal Earning”. The coefficient is quite small (nearly 0) and the R-square of the regression is 0.61. he result is summarized by the figure below :



We then include other variables which have a high absolute correlation coefficients (say >0.5 in an ad hoc manner) such as “Water quality”, “Quality of support network”, “Self-reported health”, “Feeling safe walking alone at night”, “Employment rate”, “Rooms per person” and “Voter turnout” in a mutli-regression model. Then we drop the statistically insignificant variables and return We retain “Personal earnings “, “Water quality”, “Self-reported health” and “Long-term unemployment rate”. The final results are shown in the table. The most significant variables are water quality and self-reported health. While the coefficient of personal earning remains low. In the other word, environment and health aspect account most for happiness of people. “Long-term unemployment rate” are also significant and has a negative impact on life satisfaction index.

OLS Regression Results

Dep. Variable:	Life satisfaction	R-squared:	0.847
Model:	OLS	Adj. R-squared:	0.821
Method:	Least Squares	F-statistic:	33.09
Date:	Tue, 04 Aug 2020	Prob (F-statistic):	1.91e-09
Time:	09:31:34	Log-Likelihood:	-3.9320
No. Observations:	29	AIC:	17.86
Df Residuals:	24	BIC:	24.70
Df Model:	4		
Covariance Type:	nonrobust		

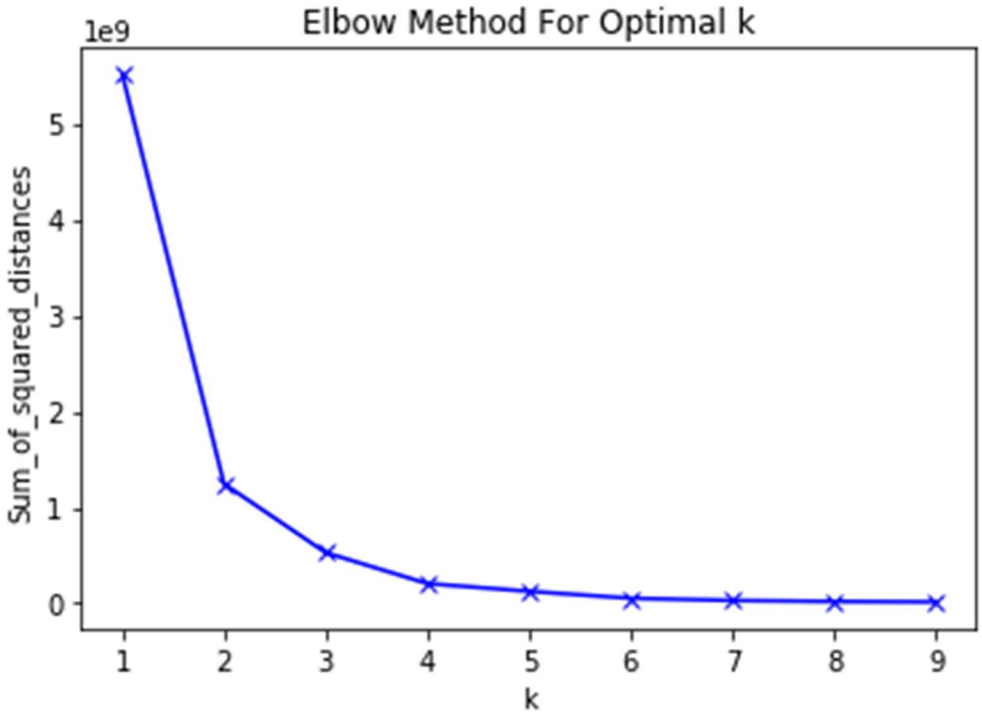
	coef	std err	t	P>|t|	[0.025	0.975]
const	2.0600	0.907	2.272	0.032	0.189	3.931
Personal earnings	1.05e-05	7.09e-06	1.480	0.152	-4.14e-06	2.51e-05
Water quality	0.0292	0.010	2.967	0.007	0.009	0.050
Self-reported health	0.0279	0.008	3.555	0.002	0.012	0.044
Long-term unemployment rate	-0.0712	0.023	-3.123	0.005	-0.118	-0.024

Omnibus:	2.130	Durbin-Watson:	2.030
Prob(Omnibus):	0.345	Jarque-Bera (JB):	1.324
Skew:	-0.521	Prob(JB):	0.516
Kurtosis:	3.106	Cond. No.	6.82e+05

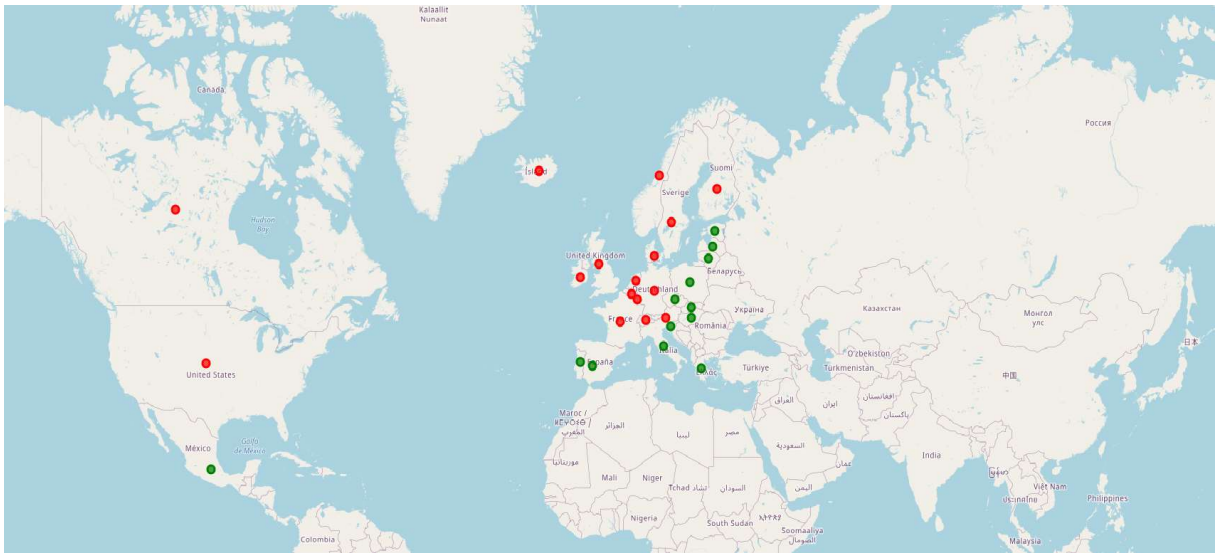
IV.2 K-means results

IV.2.1. OECD Better life results

The first step is to determine the number of clusters. By varying k from 1 to 10 clusters, we calculate the total within-cluster sum of square of distance (ssd) for each k. The curve of ssd according to the number of clusters k is shown in the figure below. We should point out that the elbow method is an heuristic method and it can be sometimes ambiguous. Here we can observe 3 bends at k=2-4 in the plot. Therefore, we run clustering with 2, 3, 4 clusters respectively and we only report the results of 2 clusters here.

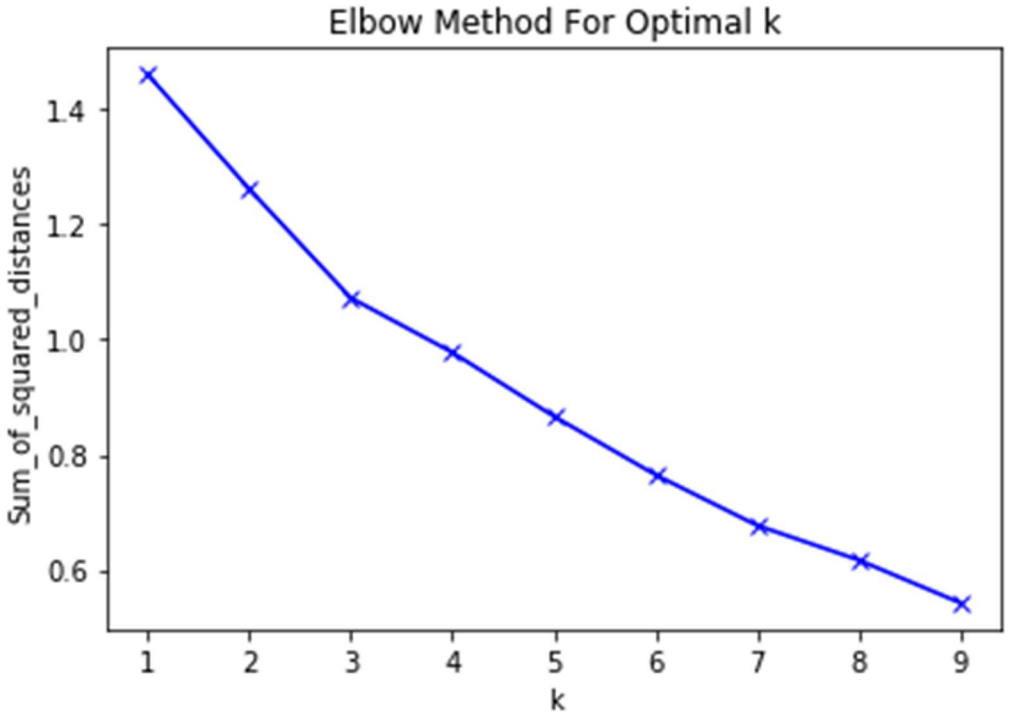


To visualizer the cluster result, we willy rely on folium.Map. The latitude and the latitude locations of each country are obtained by Nominatim in geocode. The cluster map is shown as below. As we can see, the countries are separated into 2 groups. Interestingly, the countries are separated into a above-average life satisfaction scoring countries (red circles) and below-average life satisfaction scoring countries (green circles).



IV.2.2. Foursquare results

Finally we examine if the venues characteristics of these countries share similitude in clustering. We use foursquare API to get venues information of the OECD countries. The limit is set to 500 and radius 50km (50000 m). There is no clear elbow point indicating the number of clustering and we choose 4 as above for comparison. The clustering results are shown below.



Although a (light) knee point are shown at 3, the sum of square of distance continue to decrease at a similar rate after. Therefore, we run clustering with 2, 3, 4 clusters respectively as above for comparison. The results of 2 clusters and 4 clusters are shown above. The results of 2 clusters are quite neat : one group with 3 countries (Finland, France and Norway) and the others (3 clusters grouping just classify Norway as the 2nd group and we will not report here), which is difficult to interpret. We thus report the results of 4 groups below too .

If we examine the venues characteristics of the 4 clusters, they can be (very) roughly as : Supermarket and store intensely presented cluster (Finland, France and Sweden), holiday and nature cluster (Austria, Poland, Slovak Republic and Spain), Norway and the other countries. However, we do not realize any particular links with the life satisfaction index and we cannot infer any relation at this stage.

V. Discussion

During this study, we have some interesting observations : First, Personal Earnings and Household net adjusted disposable income has high correlation coefficient with life satisfaction but not Household net wealth. Furthermore, once we include other variables in the regression, even personal earning becomes insignificant, but unemployment rate remains significant. What's the role of earning, wealth and work in the happiness? This question can be a subject of independent study.

Second, although no particular similarities are found between venues characteristic and Life Satisfaction level, a more detailed study of each capital or representative cities perhaps give more insights.

VI. Conclusion

In this project, I have used OECD Better life index and foursquare location data to study the life aspects which are important for life happiness. The results shows that environment and health aspect account most for happiness of people. Economists should pay more attention in these field. No evidence shows particular similarities between venues characteristic and Life Satisfaction level, a more detailed and targeted study can be considered.