Machine Learning Engineer Nanodegree

Capstone Project

Ahmed Nasser
Nov 9st, 2019

I. Definition

Project Overview

According to The World Health Organization (WHO) close to 800 000 people die by suicide every year. Furthermore, for each suicide, there are more than 20 suicide attempts. Suicides and suicide attempts have a ripple effect that impacts on families, friends, colleagues, communities and societies. Suicides are preventable. Much can be done to prevent suicide at individual, community and national levels.

So, comes my intuition to inspect the similarities between the nations in committing suicides to give some indicator on how to prevent them

Similar project on the dataset to predict suicides rates

https://github.com/olgaminguett/Suicide-Rates-Overview

Problem Statement

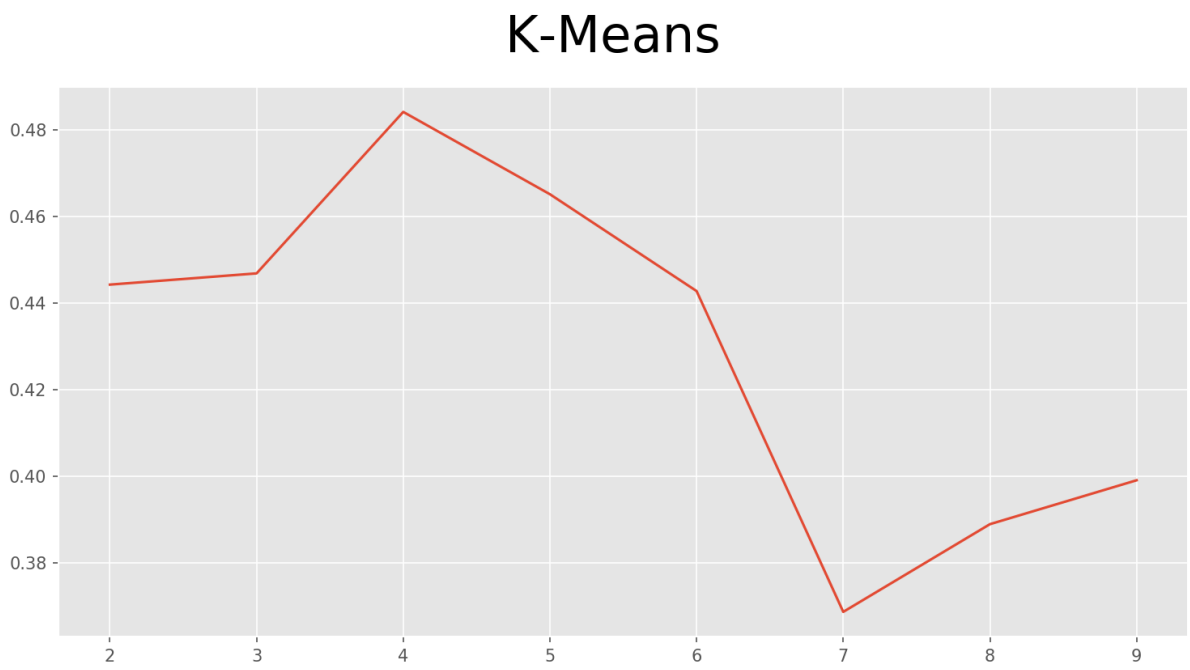
Suicide Prevention., I want to inspect the reasons behind Suicides and how it evolved through the years and similarities between countries according to those reasons to give some indicator on the right direction to prevent suicides.

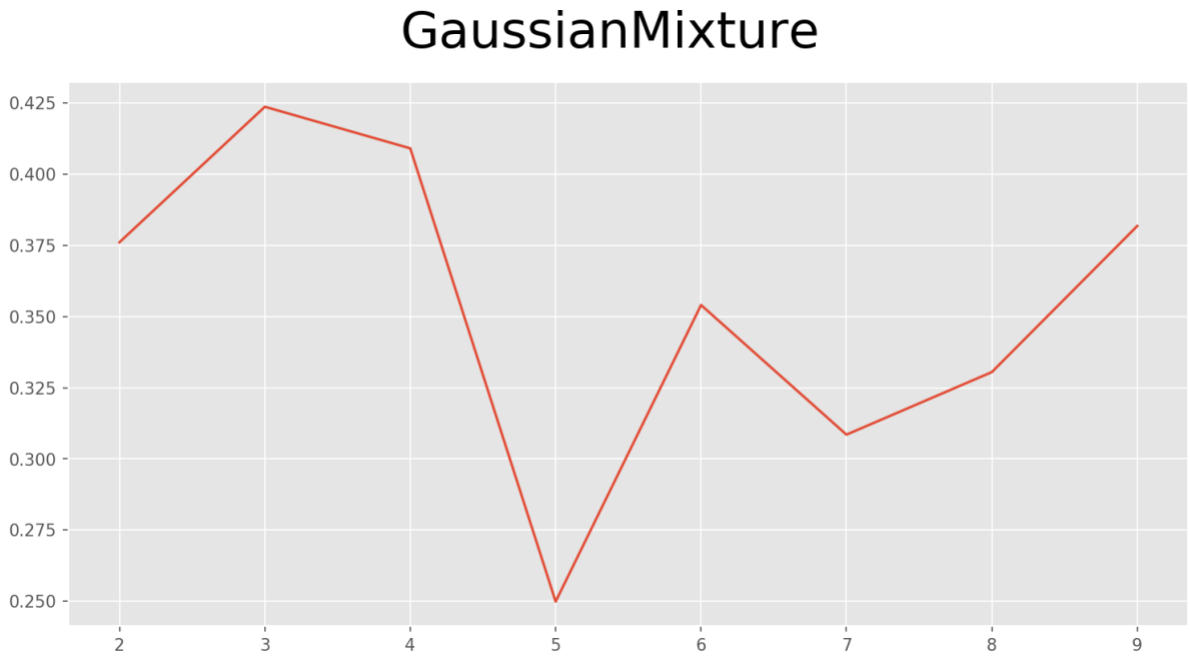
This a clustering problem where I find similarities in committing suicide between countries by clustering them according to chosen variables using k-means algorithm

Metrics

I used internal validation indices using Silhouette score as the data is not labeled to get the sense of the best k clusters numbers for k means and to compare the result of k means and GaussianMixture

As we see below k means has a better score compared to GaussianMixture





II. Analysis

Data Exploration

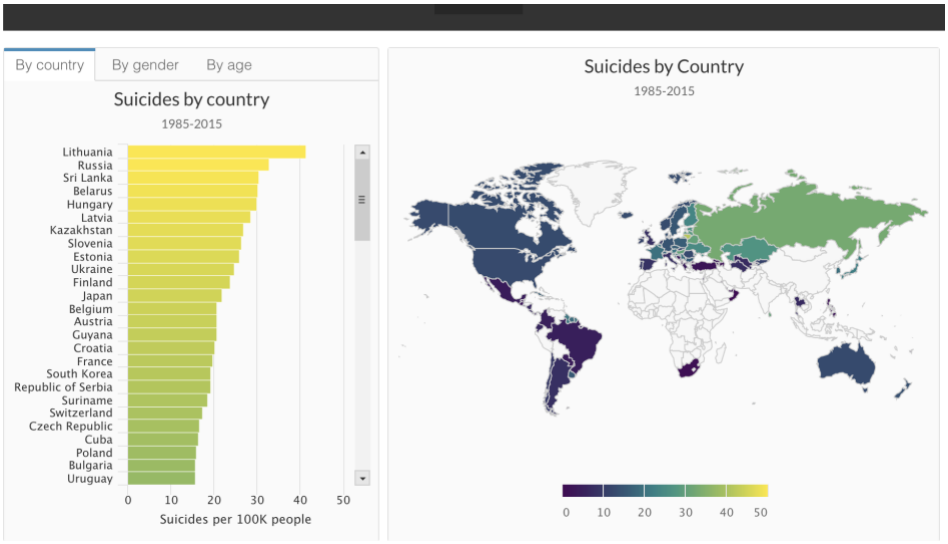
This compiled dataset pulled from four other datasets linked by time and place, and was built to find signals correlated to increased suicide rates among different cohorts globally, across the socio-economic spectrum.

Data set sample

country	year	sex	age	suicides_no	population	suicides/100k pop	country-year	HDI for year	gdp_for_year ($)	gdp_per_capita ($)	generation
Albania	1987	male	15-24 years	21	312900	6.71	Albania1987	NaN	2,156,624,900	796	Generation X
Albania	1987	male	35-54 years	16	308000	5.19	Albania1987	NaN	2,156,624,900	796	Silent
Albania	1987	female	15-24 years	14	289700	4.83	Albania1987	NaN	2,156,624,900	796	Generation X
Albania	1987	male	75+ years	1	21800	4.59	Albania1987	NaN	2,156,624,900	796	G.I. Generation
Albania	1987	male	25-34 years	9	274300	3.28	Albania1987	NaN	2,156,624,900	796	Boomers

features country , year , sex , age ,suicides_no , population , suicides/100k pop , country-year ,HDI for year ,gdp_for_year ($) ,gdp_per_capita ($) ,generation

this gives the intuition of similarities between countries

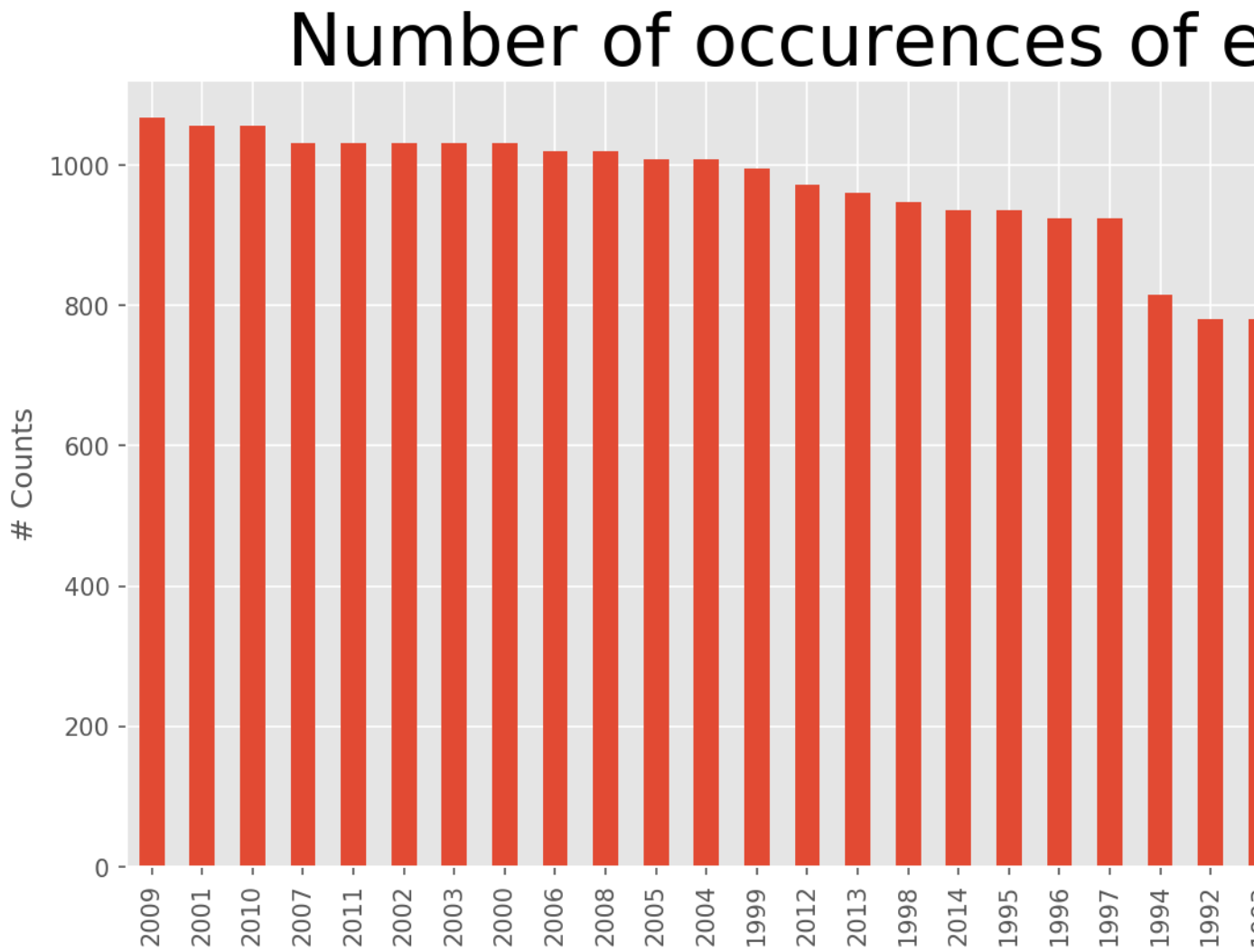


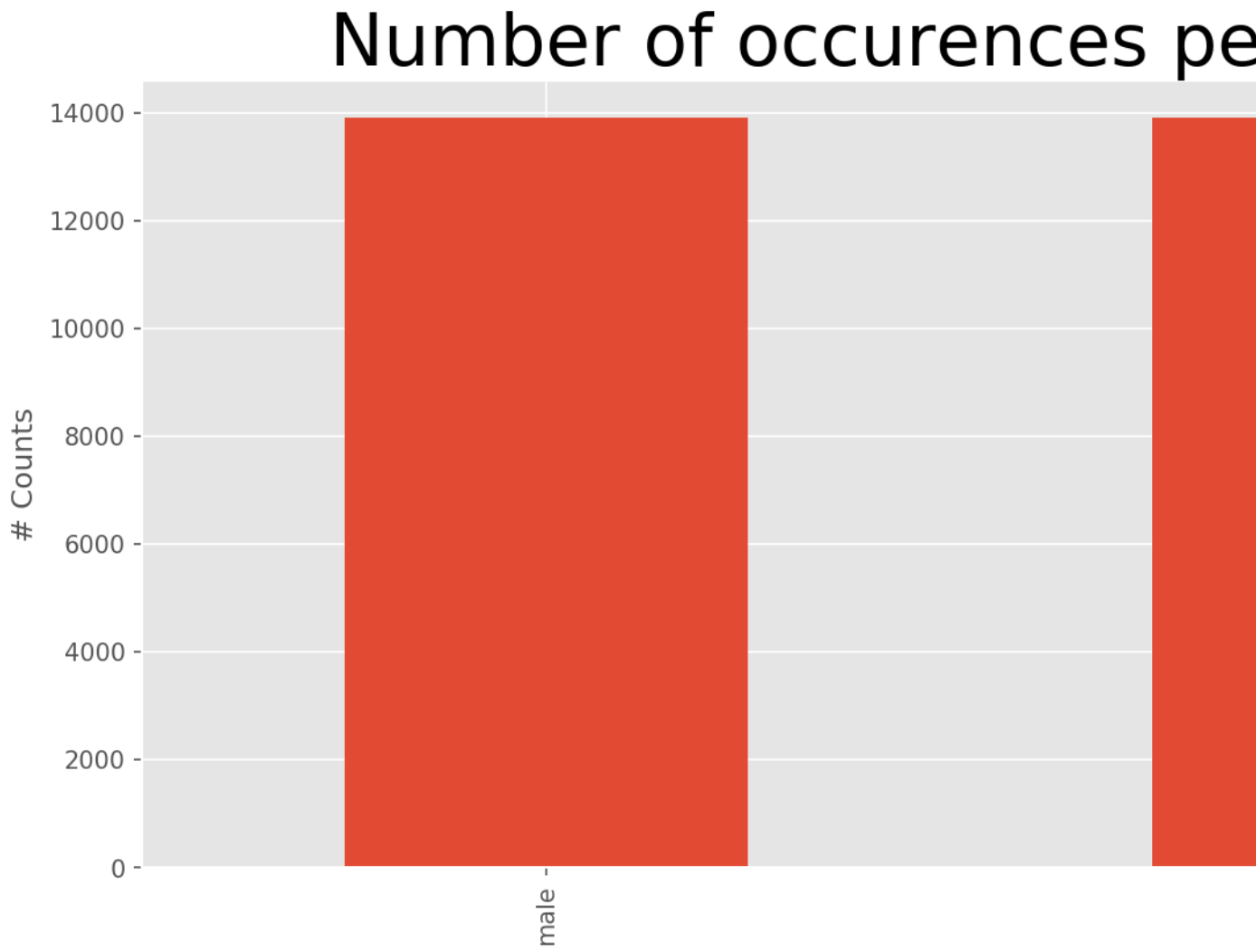
here we get some intuition about the data

	year	suicides_no	population	suicides/100k pop	HDI for y
count	27820.000000	27820.000000	2.782000e+04	27820.000000	8364.0000
mean	2001.258375	242.574407	1.844794e+06	12.816097	0.7766
std	8.469055	902.047917	3.911779e+06	18.961511	0.0933
min	1985.000000	0.000000	2.780000e+02	0.000000	0.4830
25%	1995.000000	3.000000	9.749850e+04	0.920000	0.7130
50%	2002.000000	25.000000	4.301500e+05	5.990000	0.7790
75%	2008.000000	131.000000	1.486143e+06	16.620000	0.8550
max	2016.000000	22338.000000	4.380521e+07	224.970000	0.9440

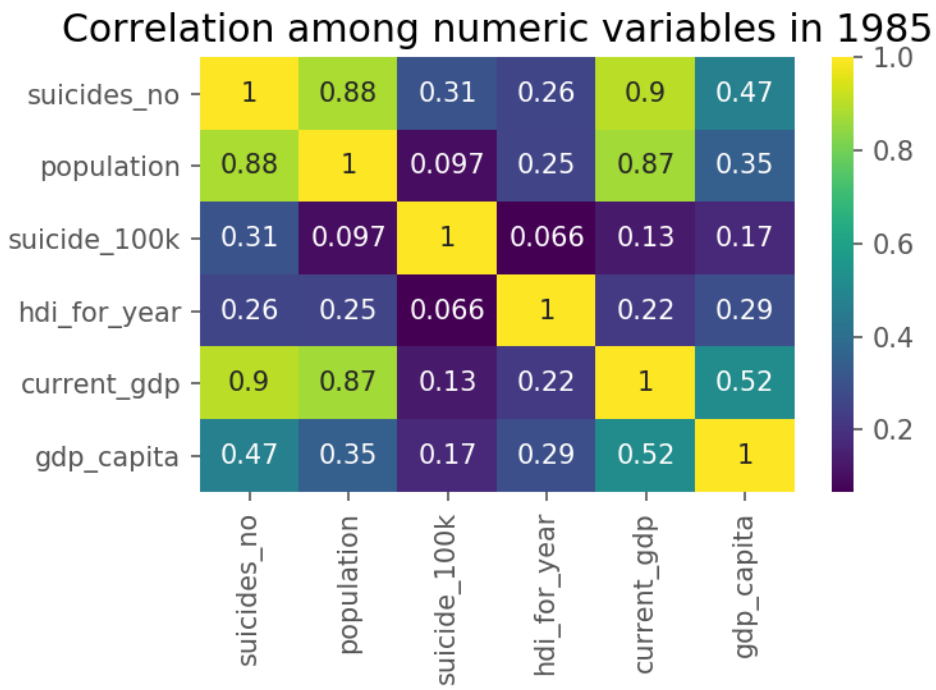
Exploratory Visualization

Exploring the suicide rates through the years and per gender I needed here to get some intuition about how the suicide rates goes when changing the variables


Number of occurences of e
Counts
0
200
400
600
800
1000
2009
2001
2010
2007
2011
2002
2003
2000
2006
2008
2005
2004
1999
2012
2013
1998
2014
1995
1996
1997
1994
1992

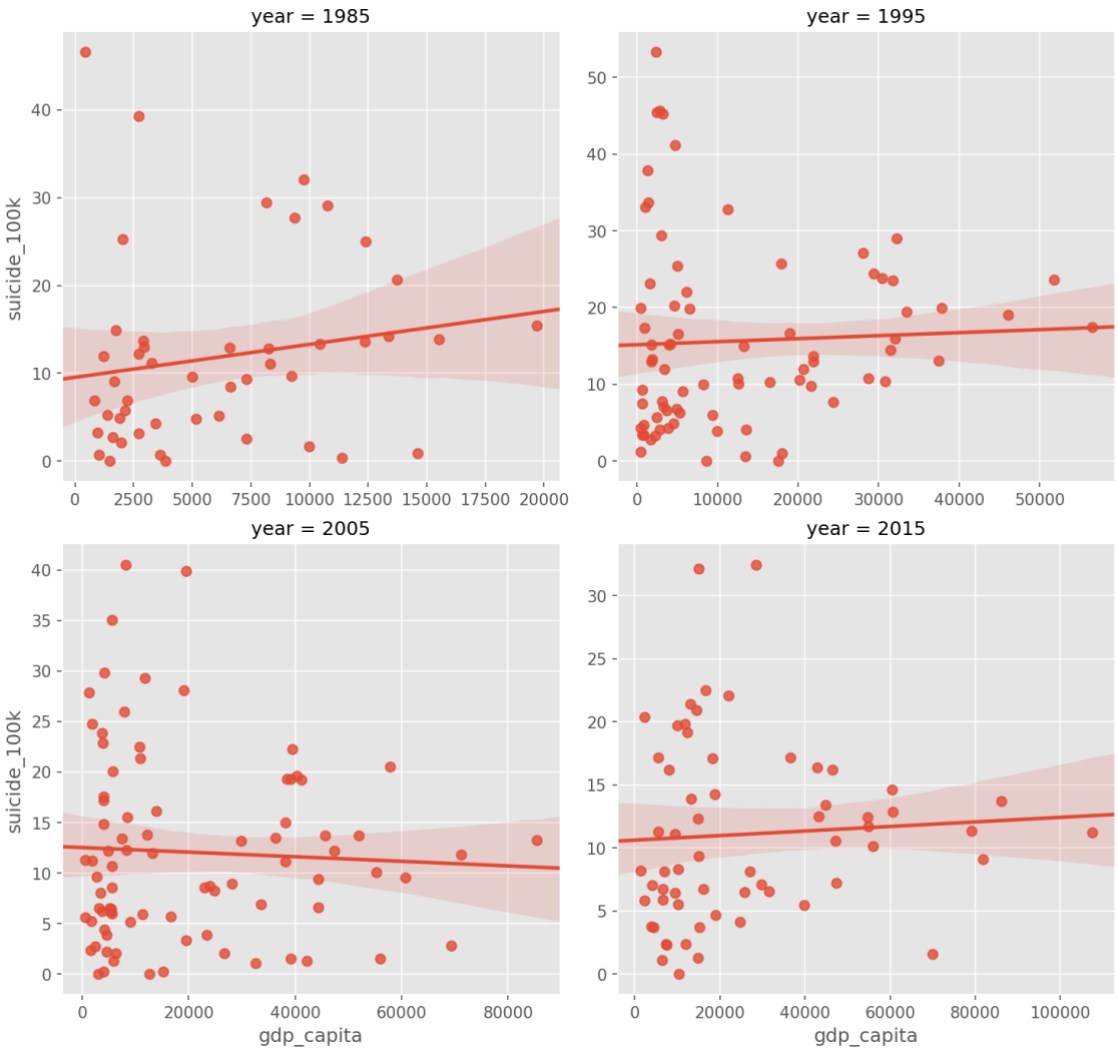


a) Correlation between data

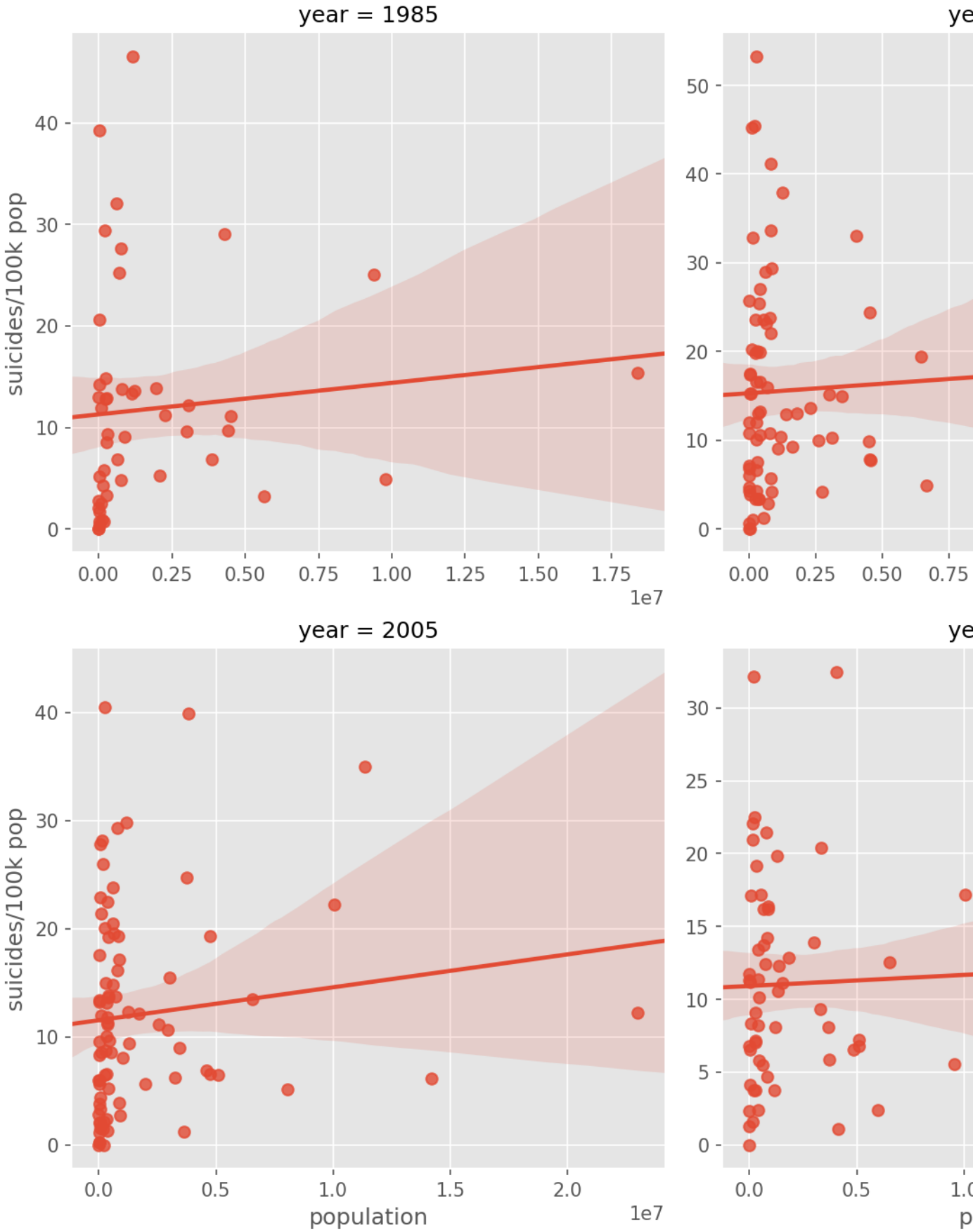


b) From the correlation mentioned earlier I would indicate that's there's a relation between GDP_capita (a measure of a country's economic output that accounts for its number of people.

) and number of suicides


year = 1985
year = 1995
year = 2005
year = 2015
suicide_100k
gdp_capita
0
2500
5000
7500
10000
12500
15000
17500
20000
10000
20000
30000
40000
50000
60000
80000
100000

link between Suicide and pop

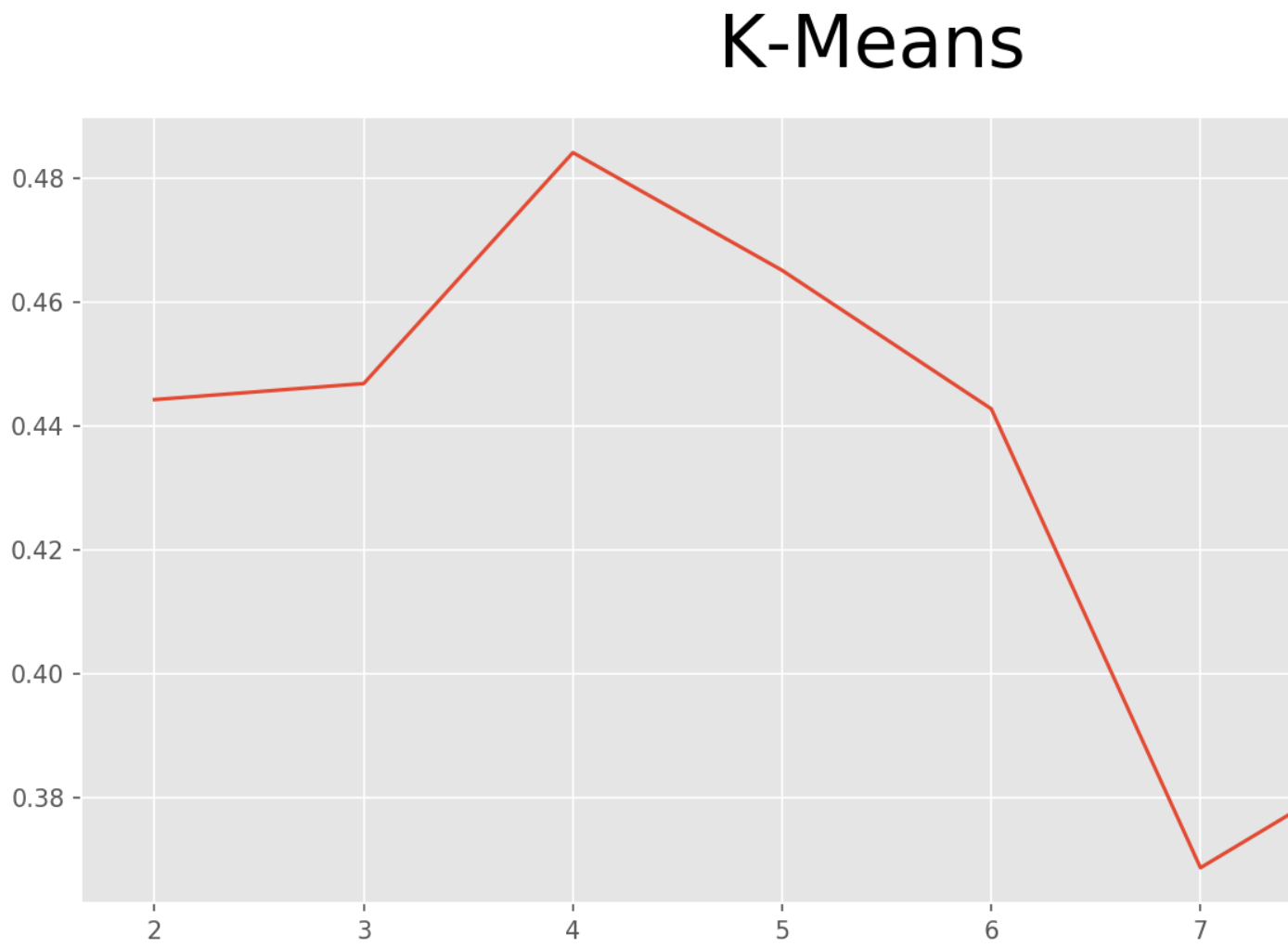


Algorithms and Techniques

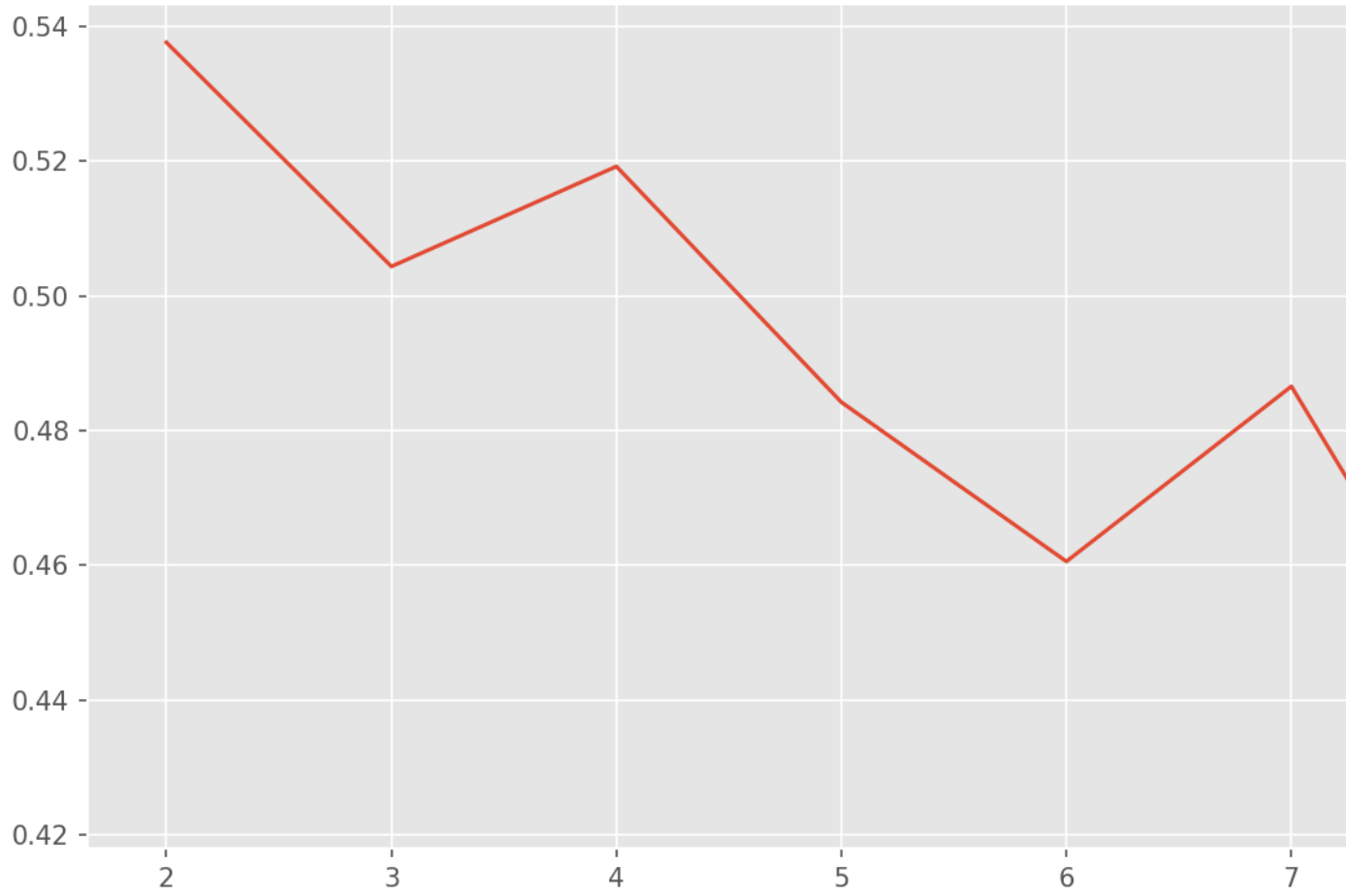
I used Cluster analysis using K-means as the data not skewed to find similarities between countries .

1. scaling the dataset so each column will have the same weight. This is important because of the distance metric the KNN algorithm uses
2. data to be cluster will depend on ('suicides_no', 'suicide_100k', 'hdi_for_year', 'current_gdp', 'gdp_capita', 'part_generation', 'Boomers','G.I. Generation', 'Generation X', 'Generation Z', 'Millenials','Silent')

the idea to see how the similarities between countries evolved through years from 1985 to 2015 I speared the data into 1985 and 2015 and as Silhouette score suggests clustering for data in 1985 fits best with 3 cluster



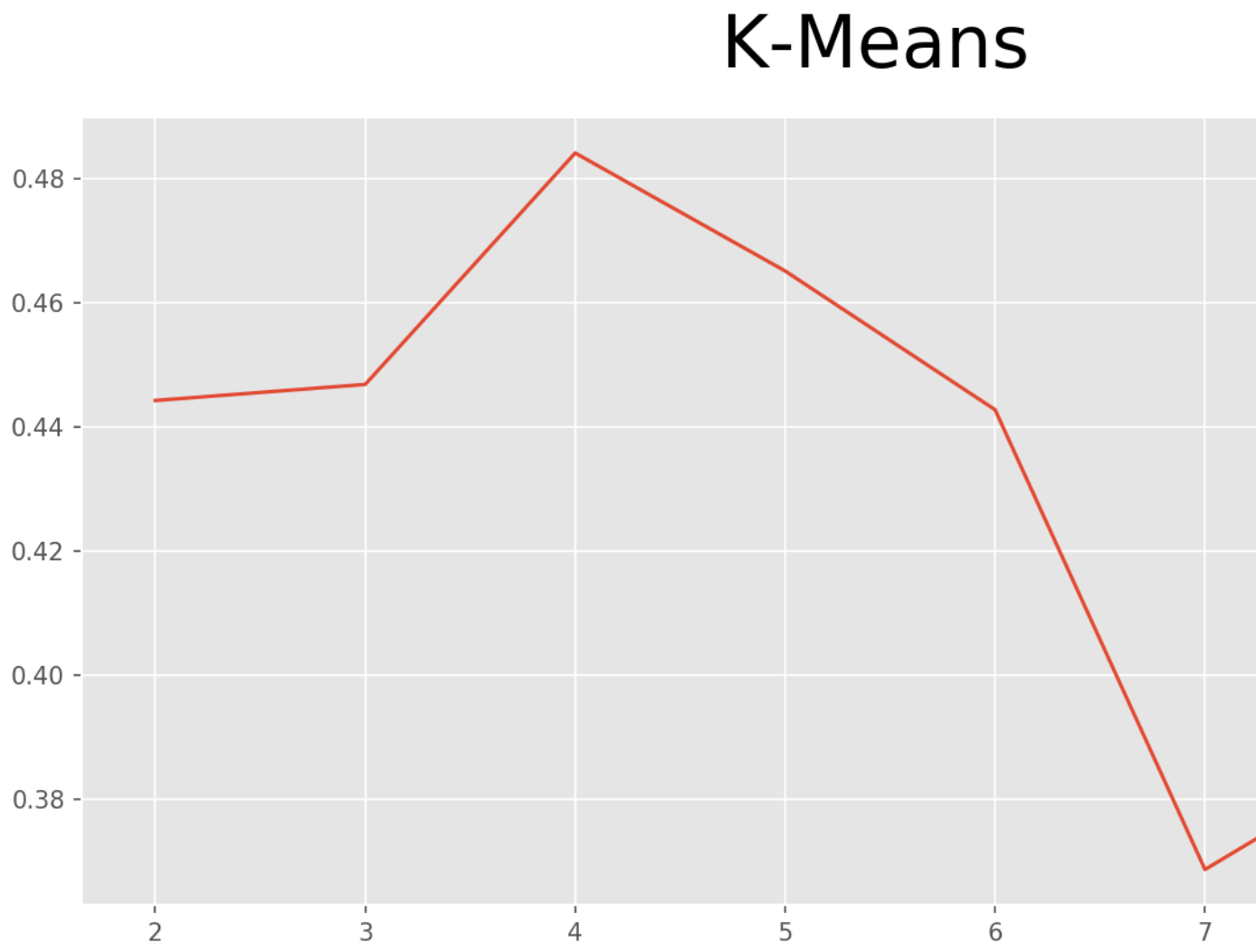
while fits 2 cluster with 2015 data



Benchmark

I dropped the population data at my solution as there's a strong correlation between population and suicide rates which gives the best clustering but I want to know if the between GDP_capita (a measure of a country's economic output that accounts for its number of people.) only gives the same intuition

Then I made another model with the population data as a Benchmark Model to my solution which eventually gave me an accuracy of 0.8125 for my model



Silhouette score for Benchmark model

III. Methodology

Data Preprocessing

I added a new column called occurrence which gives the value 1 to every row so that whenever we group and sum the numeric data we can come back to the original data by dividing it.

```
wholesale new_data dataset has 2321 samples with 8 features each.
```

Out[13]:

	year	country	suicides_no	suicides/100k pop	HDI for year	gdp_per_capita ($)	occurence	part_gene
0	1985	Antigua and Barbuda	0	0.00	0.000	46200	12	
1	1985	Argentina	1988	134.47	8.328	39168	12	
2	1985	Australia	1861	163.41	0.000	148488	12	
3	1985	Austria	2091	384.81	9.168	117108	12	
4	1985	Bahamas	1	4.76	0.000	136716	12	

Scaling the dataset

I did scaling to the dataset in order to fit the K means algorithm

Out[14]:

	year	country	suicides_no	suicides/100k pop	HDI for year	gdp_per_capita ($)	occurence	part_generation
0	1985	Antigua and Barbuda	-0.410885	-1.313957	-0.648461	-0.689247	0.0	0.083315
1	1985	Argentina	-0.129948	-0.163769	1.281230	-0.720323	0.0	0.083315
2	1985	Australia	-0.147895	0.083769	-0.648461	-0.237212	0.0	0.083315
3	1985	Austria	-0.115392	1.977512	1.475867	-0.375887	0.0	0.083315
4	1985	Bahamas	-0.410744	-1.273243	-0.648461	-0.289235	0.0	0.083315

I spirited the dataset to 1985 data and 2015 and did clustering to each see the evolution of suicide rates

```
# prepraing training data
x = new_data[new_data["year"] == 1985].drop(["year", "suicides_no"],
                                            axis = 1).set_index("country")

x.head()
```

Out[57]:

country	suicides/100k pop	HDI for year	gdp_per_capita ($)	occurence	part_generation
Antigua and Barbuda	-1.313957	-0.648461	-0.689247	0.0	0.083315
Argentina	-0.163769	1.281230	-0.720323	0.0	0.083315
Australia	0.083769	-0.648461	-0.237212	0.0	0.083315
Austria	1.977512	1.475867	-0.375887	0.0	0.083315
Bahamas	-1.273243	-0.648461	-0.289235	0.0	0.083315

Implementation

The last step in the section above is the first step of my algorithm after separating the data to 1985 data and 2015 data first I analyzed GussianMixture and K means against the data to see what fits more the data using Silhouette score which gave me k means in both cases

Then I applied k means clustering for both as follows:

1. for 1985 data I used 3 clusters which gave me those clusters

```
------------------------------

0 ['Antigua and Barbuda', 'Argentina', 'Australia', 'Bahamas', 'Bahrain', 'Barbados', 'Canada', 'Chile', 'Col
'Costa Rica', 'Dominica', 'Ecuador', 'Greece', 'Grenada', 'Iceland', 'Ireland', 'Israel', 'Italy', 'Jamaica',
t', 'Malta', 'Mauritius', 'Mexico', 'Netherlands', 'New Zealand', 'Panama', 'Paraguay', 'Portugal', 'Puerto R
epublic of Korea', 'Saint Vincent and Grenadines', 'Seychelles', 'Spain', 'Thailand', 'Trinidad and Tobago',
Kingdom', 'Uruguay']
------------------------------

1 ['Austria', 'Belgium', 'Bulgaria', 'France', 'Luxembourg', 'Singapore', 'Sri Lanka', 'Suriname']
------------------------------

2 ['Brazil', 'Japan', 'United States']
```

2. for 2015 data I used 2 clusters which gave me those clusters

```
------------------------------

0 ['Australia', 'Austria', 'Belgium', 'Denmark', 'Finland', 'Germany', 'Iceland', 'Israel', 'J
etherlands', 'Norway', 'Qatar', 'Singapore', 'Sweden', 'Switzerland', 'United Kingdom', 'Unite
------------------------------

1 ['Antigua and Barbuda', 'Argentina', 'Armenia', 'Belize', 'Brazil', 'Chile', 'Colombia', 'Cr
s', 'Czech Republic', 'Ecuador', 'Estonia', 'Georgia', 'Greece', 'Grenada', 'Guatemala', 'Hung
tan', 'Kyrgyzstan', 'Latvia', 'Lithuania', 'Malta', 'Mauritius', 'Mexico', 'Nicaragua', 'Panam
ico', 'Republic of Korea', 'Romania', 'Russian Federation', 'Saint Vincent and Grenadines', 'S
'Slovenia', 'South Africa', 'Spain', 'Thailand', 'Turkey', 'Turkmenistan', 'Ukraine', 'Uruguay
```

Refinement

I faced some troubles results first without pre processing the data and without choosing the right k for k means but after I did the Data Preprocessing and applying the Silhouette score I got a solution which I was satisfied about

IV. Results

Model Evaluation and Validation

As mentioned before I dropped the population data at my solution as there's a strong correlation between population and suicide rates which gives the best clustering but I want to know if the between GDP_capita (a measure of a country's economic output that accounts for its number of people.) only gives the same intuition

Then I made another model with the population data as a Benchmark Model to my solution which eventually gave me an accuracy of 0.8125 for my model

As my intuition is to see the evolution of similarities through years according GDP_capita

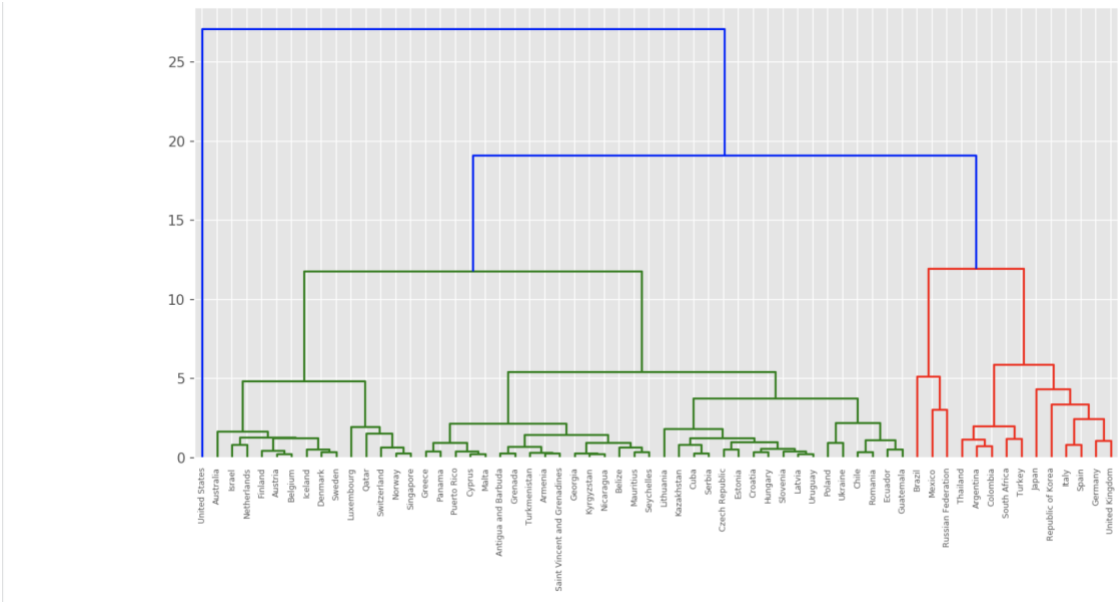
Also for validation Silhouette score got me the best clustering score for 1985 at 3 clusters and for 2015 data 2 clusters

Justification

The problem I'm seeking about is to see how the similarities evolved through the years between countries in subside rates where I depended on GDP_capita to see how can those similarities goes into some clusters

My justification is that the population doesn't affect the suicide rates and according to of 0.812 score of my solution compared to the benchmark model I think I did some progress in extracting the similarities

Here's a dendogram using Hierarchical clustering algorithm which gives a near results to my solution



V. Conclusion

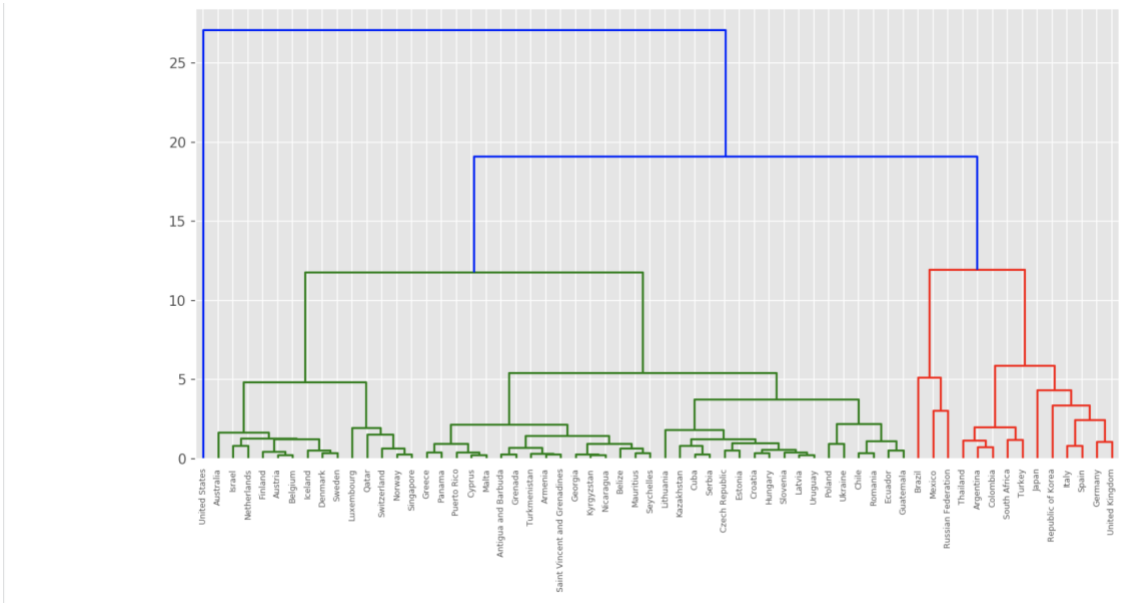
Free-Form Visualization

In the below we see how similarities in suicide rates evolved from 1985 (first image) to 2015 (second image)

```
------------------------------

0 ['Antigua and Barbuda', 'Argentina', 'Australia', 'Bahamas', 'Bahrain', 'Barbados', 'Canada', 'Chile', 'Colombia',
'Costa Rica', 'Dominica', 'Ecuador', 'Greece', 'Grenada', 'Iceland', 'Ireland', 'Israel', 'Italy', 'Jamaica', 'Kuwai
t', 'Malta', 'Mauritius', 'Mexico', 'Netherlands', 'New Zealand', 'Panama', 'Paraguay', 'Portugal', 'Puerto Rico', 'R
epublic of Korea', 'Saint Vincent and Grenadines', 'Seychelles', 'Spain', 'Thailand', 'Trinidad and Tobago', 'United
Kingdom', 'Uruguay']
------------------------------

1 ['Austria', 'Belgium', 'Bulgaria', 'France', 'Luxembourg', 'Singapore', 'Sri Lanka', 'Suriname']
------------------------------

2 ['Brazil', 'Japan', 'United States']
```

```
------------------------------

0 ['Australia', 'Austria', 'Belgium', 'Denmark', 'Finland', 'Germany', 'Iceland', 'Israel', 'Japan', '
etherlands', 'Norway', 'Qatar', 'Singapore', 'Sweden', 'Switzerland', 'United Kingdom', 'United States
------------------------------

1 ['Antigua and Barbuda', 'Argentina', 'Armenia', 'Belize', 'Brazil', 'Chile', 'Colombia', 'Croatia',
s', 'Czech Republic', 'Ecuador', 'Estonia', 'Georgia', 'Greece', 'Grenada', 'Guatemala', 'Hungary', 'I
tan', 'Kyrgyzstan', 'Latvia', 'Lithuania', 'Malta', 'Mauritius', 'Mexico', 'Nicaragua', 'Panama', 'Pol
ico', 'Republic of Korea', 'Romania', 'Russian Federation', 'Saint Vincent and Grenadines', 'Serbia',
'Slovenia', 'South Africa', 'Spain', 'Thailand', 'Turkey', 'Turkmenistan', 'Ukraine', 'Uruguay']
```

also the below dendodgram gives us some sort of this evolution



Reflection

The problem I'm seeking about is to see how the similarities evolved through the years between countries in subside rates where I depended on GDP_capita to see how can those similarities goes into some clusters

This compiled dataset pulled from four other datasets linked by time and place, and was built to find signals correlated to increased suicide rates among different cohorts globally, across the socio-economic spectrum.

References

United Nations Development Program. (2018). Human development index (HDI). Retrieved from http://hdr.undp.org/en/indicators/137506

World Bank. (2018). World development indicators: GDP (current US$) by country:1985 to 2016. Retrieved from http://databank.worldbank.org/data/source/world-development-indicators#

[Szamil]. (2017). Suicide in the Twenty-First Century [dataset]. Retrieved from https://www.kaggle.com/szamil/suicide-in-the-twenty-first-century/notebook

World Health Organization. (2018). Suicide prevention. Retrieved from http://www.who.int/mental_health/suicide-prevention/en/

I used K means clustering to find similarities in evolution of suicide rates which gave me those clusters

1985 data

```
------------------------------

0 ['Antigua and Barbuda', 'Argentina', 'Australia', 'Bahamas', 'Bahrain', 'Barbados', 'Canada', 'Chile', 'Colombia',
'Costa Rica', 'Dominica', 'Ecuador', 'Greece', 'Grenada', 'Iceland', 'Ireland', 'Israel', 'Italy', 'Jamaica', 'Kuwai
t', 'Malta', 'Mauritius', 'Mexico', 'Netherlands', 'New Zealand', 'Panama', 'Paraguay', 'Portugal', 'Puerto Rico', 'R
epublic of Korea', 'Saint Vincent and Grenadines', 'Seychelles', 'Spain', 'Thailand', 'Trinidad and Tobago', 'United
Kingdom', 'Uruguay']
------------------------------

1 ['Austria', 'Belgium', 'Bulgaria', 'France', 'Luxembourg', 'Singapore', 'Sri Lanka', 'Suriname']
------------------------------

2 ['Brazil', 'Japan', 'United States']
```

2015 data

```
------------------------------

0 ['Australia', 'Austria', 'Belgium', 'Denmark', 'Finland', 'Germany', 'Iceland', 'Israel', 'Japan', '
etherlands', 'Norway', 'Qatar', 'Singapore', 'Sweden', 'Switzerland', 'United Kingdom', 'United States
------------------------------

1 ['Antigua and Barbuda', 'Argentina', 'Armenia', 'Belize', 'Brazil', 'Chile', 'Colombia', 'Croatia',
s', 'Czech Republic', 'Ecuador', 'Estonia', 'Georgia', 'Greece', 'Grenada', 'Guatemala', 'Hungary', 'I
tan', 'Kyrgyzstan', 'Latvia', 'Lithuania', 'Malta', 'Mauritius', 'Mexico', 'Nicaragua', 'Panama', 'Pol
ico', 'Republic of Korea', 'Romania', 'Russian Federation', 'Saint Vincent and Grenadines', 'Serbia',
'Slovenia', 'South Africa', 'Spain', 'Thailand', 'Turkey', 'Turkmenistan', 'Ukraine', 'Uruguay']
```

Improvement

I need to consider more enhancements to my model and be it as a base for further studies such suicide predication and preventing using a supervised learning technique