Machine Learning Engineer Nanodegree

Capstone Proposal

Ahmed Nasser
Nov 9st, 2019

Proposal

Domain Background

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

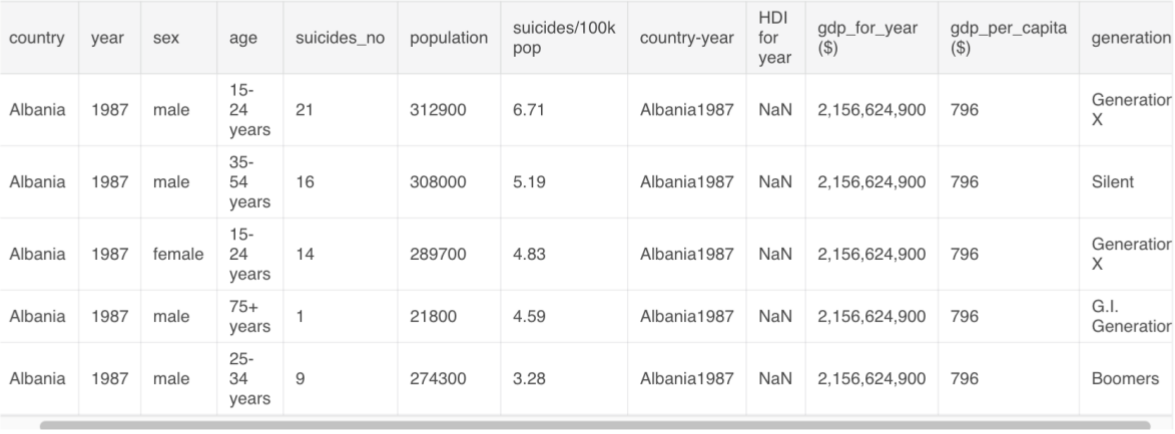
Suicide Prevention., I want to inspect the reasons behind Suicides and how it evolved through the years and similarities between countries according to those reasons to give some indicator on the right direction to prevent suicides.

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

Datasets and Inputs

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.

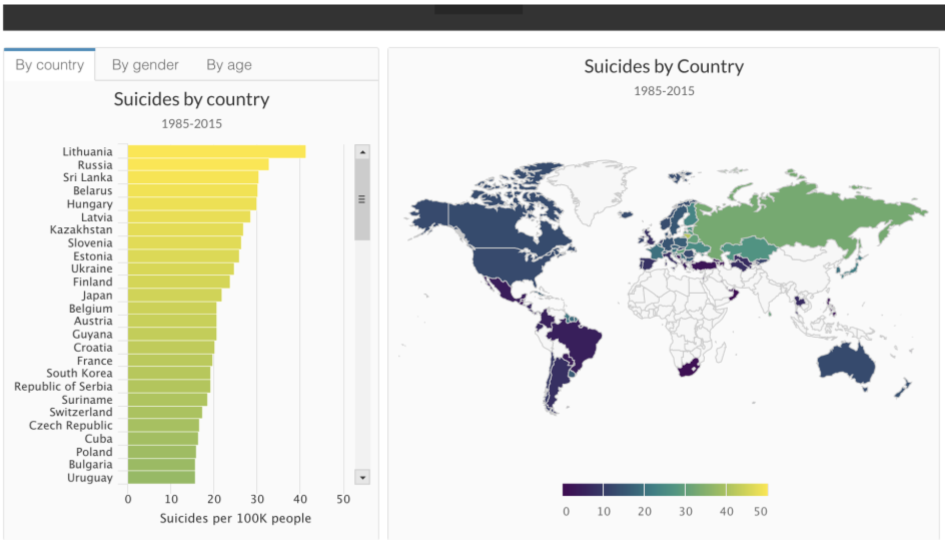
Data set sample



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	Generatior 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	Generatior X
Albania	1987	male	75+ years	1	21800	4.59	Albania1987	NaN	2,156,624,900	796	G.I. Generatior
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



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/

Solution Statement

The main objective of the project is to build a model to Cluster the countries to show how the similarities evolved since 1985. Given data about the countries, the model can cluster the countries according to their similarities using K means

Benchmark Model

I will drop 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'll make another model with the population data as a Benchmark Model to my solution

Evaluation Metrics

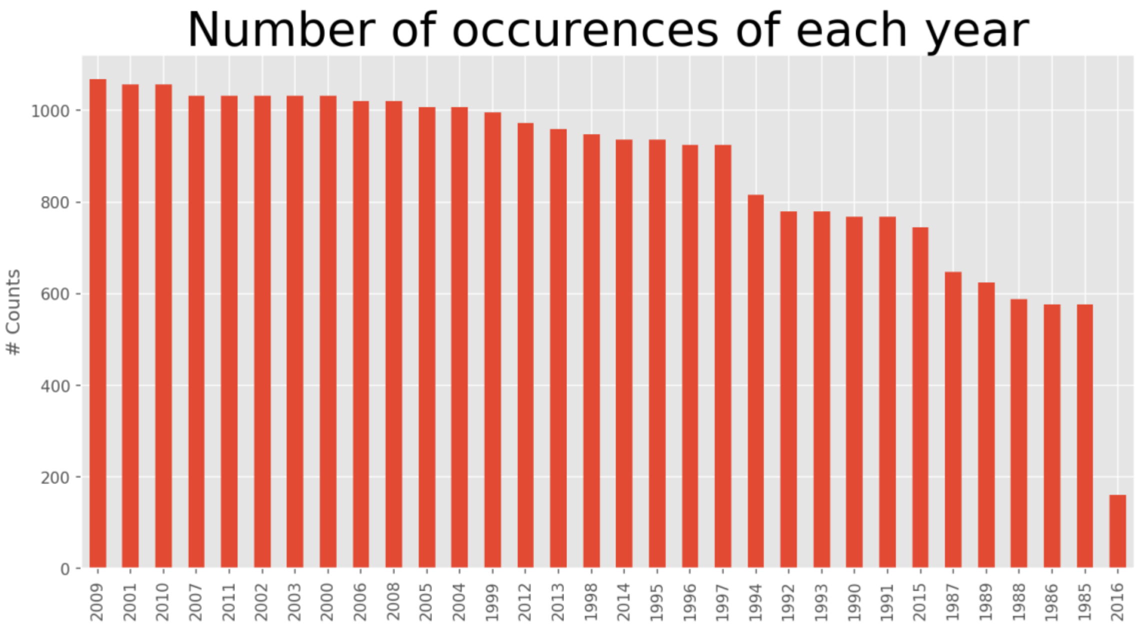
I'll use internal validation indices using Silhouette score as the data is not labeled to get the sense of the best k clusters numbers for k means

Project Design

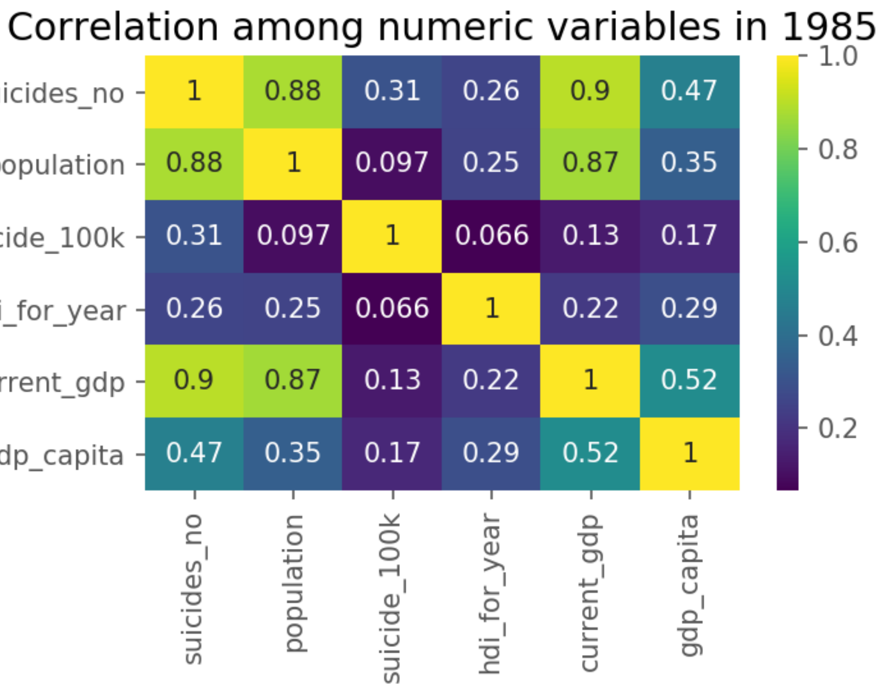
The project will consist of the following steps:

1) Data processing , Exploration and correlation

   a) Explore the data to find some statistics like

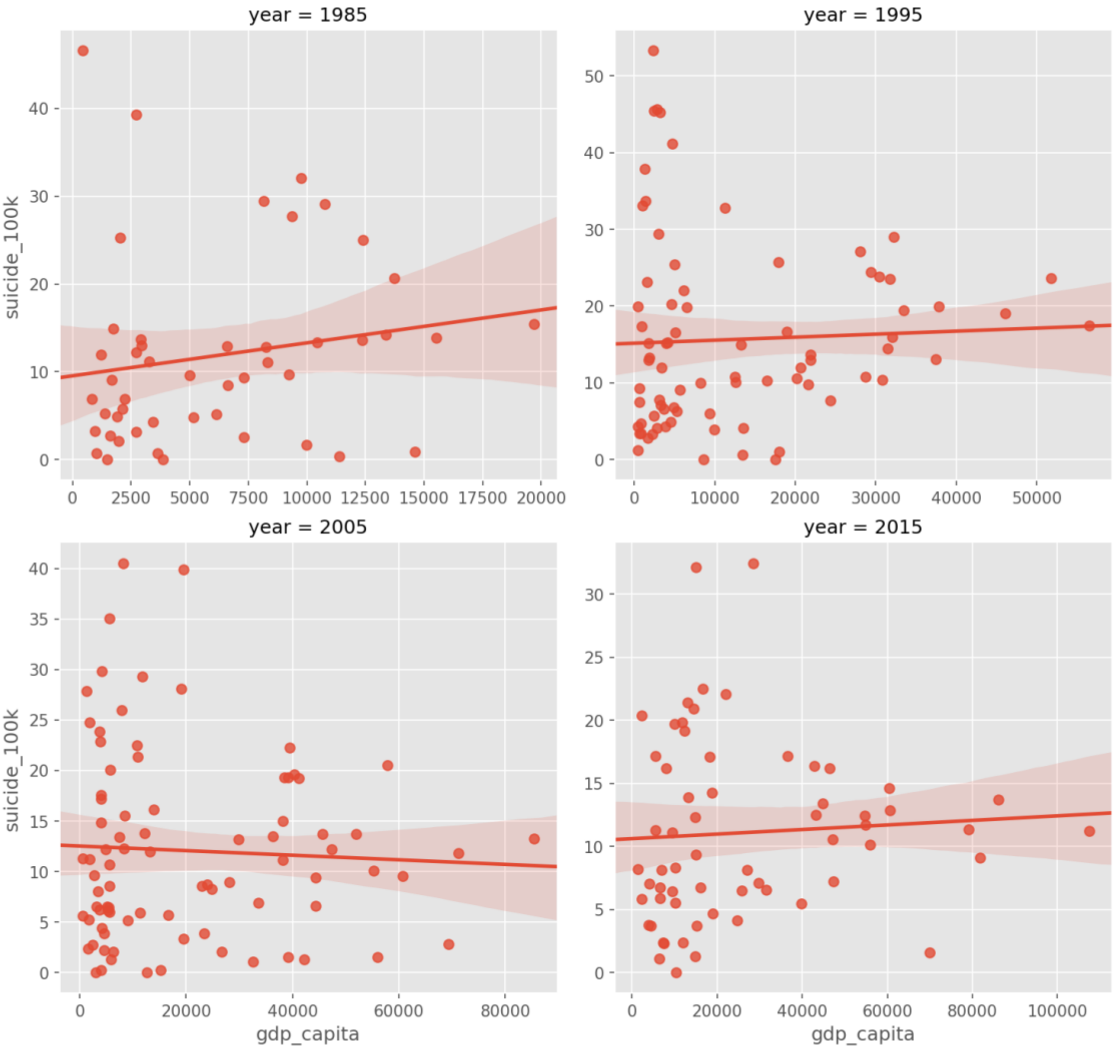


   b) Correlation between data



c) 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



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

a) scaling the dataset so each column will have the same weight. This is important because of the distance metric the KNN algorithm uses

b) 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')

3) I ‘ll use python as a languages

a) Libraries

i) pandas

ii) matplotlib.pyplot

iii) seaborn

iv) collections

v) StandardScaler

vi) KMeans