Zeyan Du
DATA 1030
10/19/2019

Project Proposal

Introduction

Nowadays, in the eyes of the government all over the world, GDP is a strong indicator of the development of a country and the overall life quality of the citizens. In general, one would expect the citizens living in countries with higher GDP per capita live a happier life. However, is GDP per capita really a solid indicator of the mental well-being of the people? Does higher values of GDP per capita lead one to value their lives more? In order explore the answers, I plan to look deeper into how suicide rate is related to GDP per capita.

The target variable is "suicide rate" which is a normalized data that describes the rate of suicides in each country and within each gender and age group. The problem is regression-based and it is important in the way that it tells us whether GDP per capita should be a standard for measuring overall life quality and happiness of the people.

In order to explore whether GDP should be a standard for happiness, I'll evaluate the relationship between suicide rates and GDP for each country. By looking into how GDP changes with year within each country, I can evaluate how changes in GDP affects the suicide rates. In order to control for other features, I'll examine whether the other features such as age, gender, generation and sex are correlated with GDP. If these features are not correlated with GDP, I will not take them into consideration because they will not affect how GDP impacts suicide rate. If they are correlated with GDP, I will control for these features in order to examine how GDP affects suicide rates.

Original Dataset

The dataset is obtained from Kaggle. (https://www.kaggle.com/russellyates88/suicide-rates-overview-1985-to-2016). The dataset was compiled from four other datasets and contains 27820 entries and 12 features where both numerical and categorical data are documented. It was aimed at exploring signs that indicates increase in suicide rates in the global spectrum.

Numerical Data:

- HDI for year: the human development index which is a statistic composite index of life expectancy, education, and per capita income indicators and used to rank countries into four tiers of human development.
- year: year ranging from 1985 to 2016
- suicide_no: number of people that suicided
- population: population
- suicides/100k population: suicide rate for every 100k population
- gdp_for_year ($): GDP in a specific year
- gdp_per_capita ($): GDP per capita

Zeyan Du
DATA 1030
10/19/2019

Categorical Data:
- country: countries
- sex: sex of people
- age: different age range of people, categorized into six groups: 5-14 years, 15-24 years, 25-34 years, 35-54 years, 55-74 years, and 75+ years.
- country-year: documents the country and the year of suicides
- generation: generation

Preprocessed Dataset

After preprocessing the dataset, there are 27820 entries and 115 features documented in the data frame with GDP per capita being the target variable.

Numerical Data:

The variables population, suicide number, suicide rate and GDP per capita are preprocessed using Standard Scaler because these variables are continuous and follow a tail distribution.
- population: representing population of within each age group, gender, year and country.
- suicide number: number of suicides within each age group, gender, year and country.
- suicide rate: number of suicides per 100k population within each age group, gender, year and country.
- GDP per capita: GDP per capita of each country

The variable year is preprocessed using MinMax Encoder, because the variable is numerical and bounded.
- year: this variable is rescaled using the MinMaxEncoder. The values get larger as the years get more recent.

Categorical Data:

The variables country, sex and generation are preprocessed using OneHotEncoder because these variables are categorical and cannot be ordered.
- country: after preprocessing the variable, it now has 101 columns with each column representing a country.
- sex: after preprocessing the variable it now has two columns, with one column representing male and another representing female.
- generation: the variable is preprocessed into six columns, with each column representing the generations: Boomers, G.I. Generation, Generation X, Generation Z, Millennials, and Silent.

The variable age is preprocessed using Ordinal Encoder because this variable is categorical and can be ordered:
- age: this variable was categorized into six age groups in the original dataset. After preprocessing the variable, the age groups are now ranked from 0 to 5, with 0 representing the youngest age group and 5 representing the eldest age group.

(Link to Github: https://github.com/daisydu97/DATA1030-Project.git)