**Data Description**

|  |  |  |
| --- | --- | --- |
| **Variable** | **Description** | **Source** |
| country | Country | World Bank & Our World in Data |
| year | Year of observation | World Bank & Our World in Data |
| pct\_65 | Population ages 65 and above (% of total population) | World Bank |
| fert\_rate | Fertility rate, total (births per woman) | World Bank |
| health\_exp | Current health expenditure per capita (current US$) | World Bank |
| lit\_rate\_adult | Literacy rate, adult total (% of people ages 15 and above) | World Bank |
| pct\_rural | Rural population (% of total population) | World Bank |
| pct\_unemp | Unemployment, total (% of total labor force) (modeled ILO estimate) | World Bank |
| pct\_mdisorder | Mental heatlth Disorder prevalence (% of total population) | Our World in Data |
| pct\_depression | Depression prevalence (% of total population) | Our World in Data |
| pop | Total Population | World Bank |
| share\_suicide | Number of Suicides for 100,000 | World Health Organization |
| hdi | Human Development Index for year | United Nations Development Program |
| gdp | GDP in given year ($) | World Bank |
| gdp\_capita | GDP per Capita ($) | World Bank |
| sex | Gender | World Bank |
| age | Age | World Bank |

**Research Question**

* We plan on examining the drivers of suicides across the world, with close attention to google trends data suggesting suicide ideation
* Suicide Prediction at t+1

**Methodology**

* Interpretable Machine Learning with SHAP values: we will run Machine Learning regression models such as Decision Trees, Linear Regression, Ridge and Lasso but also ensemble methods such as Gradient Boosting, XGBoost, and Random Forest. SHAP values are model agnostic, and therefore enable us to understand and interpret the relative importance of each variables in determining the output of any model.
* Customizable Autoregressive Linear Models: we will make use of the most performant ML models we built to predict the suicide rate in a given country at a given time for t+1. We will combine the prediction of these models with autoregressive linear models such as Prophet, Bayesian Structural Time Series, or simpler models such as ARIMA, ARIMAX and SARIMAX.

**Next Steps**

* Merge with Google Trends data (and add keywords suggesting suicide ideation)
* Run Machine Learning models, compare performance, and interpret results across models.
* Data Cleaning (aggregate variables by year, country and age)

**Appendix**

**Chart, scatter chart

Description automatically generatedChart, scatter chart

Description automatically generated**

**Chart, scatter chart

Description automatically generatedChart, scatter chart

Description automatically generated**

**Chart, bar chart

Description automatically generatedChart, line chart

Description automatically generated**

**Table

Description automatically generatedTable

Description automatically generated**