Suicides in India - a project proposal

With a population of 1.3 billion, it isn’t surprising that India ranks high in suicides. As a child growing up in India, news of farmers committing suicide to avoid paying debts was common every summer. As the summers grew hotter, these suicide rates kept increasing. The government of India has made available public data caused by accidents and suicides starting from the year 1967 to 2015. A subset of this data is available on Kaggle, with a surprising amount of detail available for the cause of suicides starting from 2001.

According to this subset alone, 1.4 million suicides have occurred between 2001 and 2012. With the undeniable problem of growing numbers of suicides, we need to find a way to understand the trends and the causes behind these alarming numbers. An ideal end result of using this data set to derive insights would be identifying demographic groups in each state of India who might be at risk of committing suicide and thus in need of help. This information would be highly useful to non-profit organizations that work on preventing suicides, such as [Save Indian Farmers LLC](http://www.saveindianfarmers.org/) and [AASRA](http://www.aasra.info/).

As a start, I will attempt to identify correlations between regional suicide numbers and a few demographic indicators such as:

* Literacy rates
* Fertility rates
* Population density
* Average maximum temperatures
* Media exposure

To establish these correlations, it would be useful to also apply filters to the data before looking for correlations, because different factors will affect groups of people differently. For example, the rural male population might be at risk of higher suicide rates as compared to, say, the urban male population, due to economic or environmental factors. Once correlations have been identified, it would be easier to pinpoint regional population groups that are at risk of increased suicide rates.

Another problem worth considering is the amount of missing information under causes of suicide. Using the information available on causes of suicide, it might be possible to classify the unclassified suicides into one of the named causes. This would help reinforce the above correlations and the resulting model for identifying at-risk groups.

The deliverables for this project include the code to perform ETL on the demographic indicator data, find correlations and build a model. A final report in the form of a blog post or a PDF document will be part of the final deliverables in a GitHub repository.