Shawn McClain

Analyzing and Predicting Suicide Rates in Europe

Problem Statement:

Suicide remains a growing concern worldwide and a silent killer. In the U.S., suicide rates are at their highest since World War 2. But is this also the case in Europe? The goal of this project is to explore multiple X inputs as potential predictors and influencers of suicide rates in Europe.

Data

The data for this project comes from the World Health Organization, Eurostat, and the World Bank. Part of the data is from a Kaggle dataset called 'Suicide Rates Overview 1985 to 2016'. This data is grouped by country, gender, population and age group along with the corresponding rates. In my project I will expand the scope of potential factors with the help of data from the World Bank and Eurostat. This data includes potential factors such as average country latitude, mean divorce rates per 1000 people, population density, mobile subscriptions, electric power consumption, and healthcare expenditure.

I will start by cleaning the data to make it easy to search and analyse. Then, I will pursue exploratory data analysis with multiple factors as they related to suicide rates and look for patterns in the data. Finally, I will narrow my focus to factors with noticeable patterns and try different modeling techniques to help make a predictive model.

Clients

Suicide Prevention concerns society as a whole, but specifically governments and the healthcare industry. It's widely suggested that mental problems are affecting more people, but the answers to why this is happening are not so clear and seem to suggest external influence is a factor by the fluctuations over the decades. Answers to help mitigate suicide from external factors can efficiently direct government resources and give sound background information to medical professionals who are advising mental health patients. For example, if it is found that population

density is correlated with suicide rates, governments and medical professionals can advise at-risk individuals to find accomodation in areas that are less dense.

Deliverables

For the casual reader or recruiter who does not want to sift through data, I will provide a 'lite' version that has many visualizations and descriptions for my process on each section of the project. I'll also provide a full notebook that consists of all of my code along with annotated notes for my thought process at each part for those interested.

Data Wrangling

For my Capstone project regarding factors that are related to suicide rates, I first did some research on possible factors from pre-exisiting theories. Some pre-existing research suggests that divorce rates are correlated with an increase in suicide rates (Kposowa). This makes sense because divorce can be very traumatic and debilitating. But, will country divorce data predict high suicide rates?

Other research suggests a hypothesis that seems counter-intuitive. Recent research found that suicide rates peak in the spring and summer where countries in the Northern hemisphere experience the most sunlight (White, Azrael). This is surprising because sunlight has been associated with uplifting mood and increasing metabolism. I added latitude data to explore this further on a country scale. Does proximity to the North influence suicide prevalence?

Along with these pre-existing theories, I gathered data from the World Bank on a variety of factors that could contribute to suicide rates. Namely, electric consumption, population density, mobile subscriptions and broadband subscriptions. These are relatively new and modern factors that raise a general question of whether or not modernization is associated with increased suicide rates.

For my main dataset, I first read in the CSV file and dropped columns that were of no use to this project. This dataset included all countries for many factors, and a lot of years were missing numbers. I first used the replace method to replace missing values with a NaN value to ease the cleaning process. My dataset had years in the columns and series factors in a single

column, stacked vertically. I wanted to swap the location of these two, so I first named the columns ‘Year’. Then I stacked ‘year’ and unstacked ‘series name’ to the columns. With my new NA values. I wanted to drop rows of country years that were missing a lot of data, so I used dropna with a certain threshold. The remaining NA values I filled to 0 for the time-being, although I may fill it differently as the project progresses.

Because my data is numerical, I converted the values from strings to floats. I then reset, and sorted the index to give each column a unique key, and move the year and country name to columns. I decided that many smaller and obscure countries might make my data too broad and compromise the quality of the dataset, so I narrowed the countries to countries in Europe. Because these countries share more cultural similarities and are well-studied, I figured it may help hold other factors as constants and improve the quality of the data.

I also read in a dataset I found from Kaggle on suicide data and an excel speadsheet I found on Eurostat for divorce rates. Because the Kaggle dataset has age and gender filters, I decided to keep this dataset separate for the time-being. For the divorce dataset, I dropped unuseful rows and columns. As with my main dataset, I replaced filled NA values and dropped columns that were exceedingly deplete. I then filled rows using a forward fill on countries with less than 4 missing years. To finalize, I cleaned up the country names and created a new column that averaged all the years for each country. I merged a simple latitude dataset onto the divorce rate data.

I may need to take more data cleaning steps in the future, but my work thus far should provide a good foundation for my project and allow for some early exploratory analysis.

Pre-existing Research Sources:

Kposowa AJ, Divorce and suicide risk, *Journal of Epidemiology & Community Health* 2003;**57:**993.

White RA, Azrael D, Papadopoulos FC*, et al*Does suicide have a stronger association with seasonality than sunlight?*BMJ Open* 2015;**5:**e007403. doi: 10.1136/bmjopen-2014-007403

Data Story

With my code and datasets clean, my first step for analyzing my project was getting a general overview for the suicide rates and general trends between different variables. I grouped by age, country, and year to find the top 10 largest suicide per 100k results. Seven of the top 10 results were in Hungary. I then grouped the dataframe by age to count the sum total of suicides in each age bracket. The age bracket of 35-54 years old had the highest total number at 815,428 people, followed by 55-74 year olds at 624,545 people. Although 35-54 year olds had the highest total numbers, when I looked at the mean suicides per 100k population by age, the 75+ age group had the highest average at 32.238 people per 100k population. These results suggest that there are less people over 75 years old in the population when compared to other groups, but their rate of suicide is higher than others. These results were plotted in bar plots.

Moving on, I aggregated the data frame by year and the mean total numbers of suicides throughout all of the countries. Suicide rates jumped in 1995, but gradually dropped thereafter until 2016 when they spiked again. I determined that 2016 was an outlier with not enough data points as it only included 10 European countries, so it's likely it was skewing results. I used a line plot to show the regression between years.

Next I aggregated the dataset by year and gender to see if gender made a difference in suicide rates. I used the seaborn lmplot and used the hue argument to separate the graph by gender. What I found was that in both genders, the average seems to be on a steady decline, but men were approximately three times as likely to commit suicide when compared to women.

For this part of my analysis, I wanted to add more variables into the equation to see if there were any correlations between suicide rates and those variables. To do this, I merged my divorce, and country latitude dataframe onto my suicide dataframe. I compared each country by their latitude, mean divorce rate and mean suicide rate using a pairplot. I also used the .corr method on the different pairwise relationships to see if any had a significant correlation to each other. The strongest correlation was the mean divorce rate with suicide rates at 0.575.

In my Europe dataframe, I had some other variables that I also spliced into my analysis. I merged data regarding electric consumption, mobile cellular subscriptions, and population density. Grouping by year, I found that the suicide rate was strongly negatively correlated with the year and the mobile cellular subcriptions at almost -1 for each variable. I used the subplots framework to compare mobile subscriptions with suicide rates by year.

Moving forward, I think I will look to do more multivariate analysis on the project and try to hold some variables as constant to get a better picture as to what variables are better predictors, and then possibly predicting suicide rates in the future based on those factors. I hypothesise that mobile cellular subscription may help suicide rates to an extent becuase they help connect us to people and that may help prevent some suicides from happening.

Inferential Statistics

Thus far in my capstone project, I have already done some statistical analysis with my exploratory analysis. In my exploratory analysis I compared suicide rates in Europe over time with a variety of other measurements in Europe over the time period. Using the .corr() method, I found that some variables were correlated with suicide rates during the time period, and some not so much. When comparing suicide rates with electric power consumption for each country, my corr score was 0.097, which is almost 0, signifying a very weak correlation. Divorce rates and country latitude scored at 0.575 and 0.545 respectively, signifying a decent correlation. The year and mobile subscriptions had the largest correlation in either direction at -0.952 and -0.967 respectively. These two numbers are very close to -1, which signifies a strong negative correlation.

To expand upon this statistical insight, I wanted to first focus on a variable pair with the highest correlation. I chose to plot the year on the x-axis and avg. suicide rates per 100k on the y-axis. Because of the strong correlation, I felt comfortable making a linear regression to give an estimate for suicide rates in Europe in the future. I used np.polyplot which measures the mean squared errors between points and gives a linear slope along with an intercept. I used np.linspace to plot theoretical x values (years) for the y value (suicide rates). From my model, I plotted the regression along with the scatter plot. From plugging 2025 in my regression, I predict suicide rates to continue to decrease and land at about 9.59 suicides per 100k people in Europe by the year 2025.

Because divorce rates and country latitudes had only a moderate correlation with suicide rates, I thought it would be a good opportunity to use bootstrap simulations for the linear regression model for both variables vs. suicide rates. My null hypothesis in both cases was that latitude and divorce rates had 0 or a negative relationship with suicide rates. To test this hypothesis, I imported a function from the most recent Data Camp course called draw_bs_pairs_linreg. The

function passes in two arrays for the two variables we are comparing. It also passes in the number of replicates we want to create for the regression. The function iterates through the x values to compile the number of indices to use. It then uses np.random.choice to choose with replacement random indices. These indices are then used to randomly select x and y values corresponding to each index and the results are stored in arrays. It then uses np.polyfit to fit a slope and intercept for the randomized data, and this process is repeated for the amount of random samples we want. I chose 1000 for both divorce rates and country latitudes.

For divorce rates, I found a slope of 0 or less in 0% of my results, which made my p-value 0. Because I didn't see a 0 or less slope in any of my iterations, I decided to reject the null hypothesis that the slope between divorce rates and suicide rates was 0 or less. For country latitude, I found a slope of 0 or less in 0.9% of my results, which made my p-value 0.009. This means it was extremely rare to find a negative or 0 slope in my iterations. I decided to also reject the null hypothesis that the slope between divorce rates and suicide rates was 0 or less.

In-Depth Analysis

Building from my statistical analysis, I had a pretty good idea as to which variables were most predictive for suicide rates thus far. Latitude and divorce rates seemed to be strong predictors from my bootstrapped regression models as well as year and mobile subscriptions with a high r coefficient for year. But, for my prior analysis, my data had been aggregated by year and country separately. In order to use my data to perform machine learning, I needed to disaggregate my data to provide more data points for train-test-splits and predictive modeling.

My first step was manipulating my dataframes by taking a previous dataframe with all of the columns to use more variables for prediction, and melting all of my dataframes into one. Once my machine learning dataframe was cleaned, melted and merged properly, I wanted to get an overview of good predictors with StatsModels Ordinary Least Squares package. Using just divorce rates, I fit a linear regression to my target (suicides per 100k). With only divorce rates, my model accounted for an R2 value of 0.185 and an F-Statistic of 71.92. This model confirmed to me

that divorce rates were a significant predictor by alone being able to account for 18.5% of the variance.

Building off of my first OLS model, I tried adding a number of other variables and getting the summary of my OLS model. Some predictors had p-values that were greater than 0.05, so I filtered those features out of my model because they diminished the model's F-Statistic and weighed down the model's significance. Using mobile subscriptions per 100k, year, divorce rate, latitude, gdp per capita, and population, I constructed a model with an F-Stat of 43.38 and an R2 score of 0.455. Using these 6 predictors increased my R2 score and also decreased the Akaike Information Criterion from 2104 to 1985. All of my predictors had a p-value less than or equal to 0.001 except for year, which had a p-value of 0.04.

To check if my OLS model had any bias towards a non-linear regression, I plotted the residuals vs. the fitted values. For this graph, success is finding no patterns or relationships in the model. This is also useful to check and see if there are any outliers that are skewing the data. In my residual plot, the graph shows a slight propensity to overestimate vs. underestimate, but the data looks to be scattered well around the graph, indicating a linear model.

Another tool for testing linearity is with a Quantile Quantile plot. If our model is linear, we'd expect the residuals to be normally distributed, and follow the quantiles of a normal distribution corresponding to their value. A successful result is when our residuals are hugging the QQ line. In my graph, the residuals do hug the QQ line for the most part with only one or two outliers.

To mitigate against high leverage outliers that can skew our regression, I used an Influence plot. This plot also uses the residuals and plots the studentized residual value vs. the leverage of the point. High leverage points are points that are unusual to the bulk of the distribution, but not necessarily highly influential. High influence points have a relatively high leverage and a relatively large residual value. In the plot, these high influence points are indicated based on their relative bubble size. Looking at my Influence plot, there were some points with high influence, but looking deeper into the points did not indicate that these points were invalid or errors in the data.

After using the Stats Models library, I wanted to try some different methods with the Sci-Kit Learn library. For this model, I included the variables for my model because I planned on using regularization models to tune my model. I divided my predictors into the X data and my target variable (suicides per 100k) into Y. I used a train-test-split on X and Y with a test size of 0.3. This allowed me to test unknown variables on my own training data, from the same sample. My

initial R2 score when tested off of the training model was 0.5542, and improvement from the Stats Models OLS model from before. My mean squared error for this model was 22.2319.

Using seaborn's lmplot, I plotted my predicted Y values vs. my actual Y values. In a perfect model, we would expect a 1:1 ratio between predicted values and my actual values. However, in my model, below 18 suicides per 100k, my model tended to underestimate suicides and over 18 suicides per 100k, model tended to overestimate suicides.

To improve my model, I used some regularization algorithms. I fitted my training data with a Lasso Cross-Validation model. This model penalizes predictors that don't help the fit of the model with small coefficients and minimizes them towards 0 in the OLS regression. In turn, this shrinks my model and leaves the best predictors. Using the lasso regression and tuning the alpha parameter for regularization, I increased the R2 to 0.61417 and decreased the MSE to 19.245.

I also tried a Ridge Regression, which is a similar relative to the Lasso Regression. In Ridge Regression, the coefficients are squared and multiplied by penalty, then added to the OLS. This regularization technique rewards models with a lot of good predictors, because generally, coefficients will not be set to 0. Using the ridge regression and tuning the alpha parameter for regularization, my R2 improved from the bare bones model, but did not improve upon the Lasso model with an R2 of 0.6032 and a MSE of 19.791.

Finally I tried an Elastic Net model, which is a combination of the Lasso and Ridge regression. I used a grid search to sort through different combinations of tuning parameters, and came up with my best parameters for the model. My best result for the Elastic Net model was an R2 score of 0.6136 and a MSE of 19.274. Elastic Net outperformed the Ridge model, but the Lasso model still had the highest R2 and the lowest MSE.

Findings

From our initial data analysis, it was discovered that men are over 3 times as likely to commit suicide in Europe. While 35-54 year olds have the highest crude suicide totals, per capita, seniors 75+ have the highest rates. Through the lens of countries, former Soviet and Eastern European countries have the highest region rate. Latitude is correlated with suicides in Europe, with Mediterranean countries having some of the lowest rates. Divorce rates have an even stronger

link to suicides with r equal to .575. But mobile phones had the strongest correlation to suicides with r equal to -0.93.

From the more in-depth analysis, I was able to model more variables and deaggregate my data for detail and to control against year and country. Divorce rates alone were able to account for 0.185 of the variance. Adding per capita mobile subscriptions, year, latitude, gdp per capita, and population further increased the explained variance score to 0.455. Using all of my variables and the Lasso regression, which sets non-predictive variables to 0, the model achieved an explained variance score of 0.6142. Using partial dependence, which judges how predictive variables are against each other, I was able to come upon the most influential variables. Year, gdp per capita, the crude death rate, urban population percentage, divorce rate and latitude were the most predictive variables.

One interesting takeaway of this project is that going into the project, I assumed because rates in the U.S. were at their highest since World War 2 that this would also be the case in Europe. With this assumption, I expected variables of modernization like GDP per capita, mobile phone subscriptions, and urban population percentages to negatively affect suicides. My results suggest that suicides in Europe have been steadily declining over the past 20 years and with that, variables of modernization tend to be associated with a decrease in suicide rates.

This may not prove that modernization decreases suicide rates, but rather that increasing rates in the U.S are not because of modenization. There must be other variables at play in the U.S that are causing rates to increase while rates in other parts of the world steadily decline.