Illinois Institute of Technology

Homicide Rate Predictor Application

Team Data Squad

Patrick Sood (A20460126)
GVS Goutham(A20450688)
Kevin(A20454613)
Rishikesh Jangam (A20448930)
Jason Yeoh(A20457826)

CSP 571 - Data Preparation and Analysis
Prof. Adam McElhinney
May 5th, 2020

Introduction:

Crime is a bane of human society but it is ubiquitous in all cultures and countries. Having a low crime rate in society is extremely important for the well-being of its people and the progress of the country. Therefore, lowering crime rate is a highly desirable objective. In order to achieve the objective of lowering crime, we need to be able to predict crime. Collection of data to create a model to predict crime is an expensive and time intensive process. However, we observe that socio-economic indicators are routinely measured for many other purposes such as part of economic planning, for quantifying targets to be achieved and as measures of effectiveness of existing policies. So, this data already exists. Because machine learning algorithms have the capability to reliably capture associations, we can squeeze even more utility from this data (than its original intended purpose) by using it as input into machine learning algorithms and discovering associations between hitherto unrelated variables. **The objective of our project is to utilize indicators of socio-economic progress and machine learning to predict crime. This predictive capability then provides the means to lower crime by controlling and directing these socio- economic factors in desirable directions. The predictive capability also provides informed advice on the most appropriate distribution of economic resources amongst the various sectors that would have the maximum impact on lowering crime**

Problem Statement:

Crime is a complex and multifaceted phenomenon. So, for the purpose of this class project, we will like to first focus our query on the following:

We will like to predict **homicide rates** at **country level** based on the **economic indices of the country's prosperity**.

This question is significant because even in our everyday experience, in our wonderful city of Chicago, we have some inkling that certain neighborhoods of Chicago are more prone to crime than others. A formal, data-based study to predict homicide rates at country level, based on the economic indices of the country is, therefore, important. It is quite evident that crime is a multidimensional problem and is an enormous area of research by itself. Therefore, for our goal of a class project-which is to be finished within a limited time-we would explore only certain aspects that can help us predict crime. However, the methods, models and techniques that we would introduce in this analysis are quite general in nature and are applicable for more complex analysis of crime prediction as well as other predictive analytics problems.

The research that we carry out in this project will be very useful to the economic planners, governments and law enforcement agencies of the country who would like to know the answers to the following questions:

Which factors or combinations of factors are better predictors of homicide rate?

The answer to the above question will help the various stakeholders to answer the following questions:

Of the three sectors that we can invest in, which investments or a combination of investments have the biggest impact on lowering homicide rates?

In the context of lowering homicide rates, should one focus more on improving economic prosperity, on woman empowerment or on the educational development in the country or combine some of these factors to get the maximum impact for lowering homicide rate?

Because there are large number of predictors of economic prosperity (but the number of countries and the countries for which data is available are limited-and we have to do a training and test split of the dataset as well), we will have to limit the feature space to 5-7 predictors for our **primary data set** so as to not run into the problem of "curse of dimensionality". Then we will be adding predictors from the first secondary dataset on woman empowerment in a country and second secondary dataset on educational development in a country to understand whether this additional information from woman empowerment sector and educational sector enhances our model.

Data Acquisition:

We will be using the data from the following source: https://ourworldindata.org/countries
This data source contains data on large number of socio-economic indicators for almost all the countries in the world under various categories including economic, gender and education. For example, for countries such as US and China more than 700 indicators have been provided. As noted above, from the viewpoint of data, the bottleneck in our analysis is not the limited number of socio-economic indicators. It is the number of countries compared to the number of possible socio-economic indicators so that we do not run into the "curse of dimensionality" problem. This data source was chosen because the data is reliable, open source and the sources of data have been cited on the website. Another reason for selecting this data source is the quantitative nature of the predictors and the target variable-the economic and social predictors that we will be using in our model have an **interpretable, physical basis** and make business sense.

This data source contains indicators on economic development, educational development, health and women's empowerment for almost all the countries in the world from 1950's to 2017. However, we note that the data on the socio-economic indicators of the countries is by no means complete. This is understandable because collection of data is a costly exercise and requires many other factors to be present concurrently such as a stable government, priorities of the government, the kind of data that sovereign - governments are interested in collecting in their respective countries. For our primary dataset, we have chosen the economic indicators for a country. In this data source, data for a specific indicator is available in a csv file. However, the csv file needs to be suitable preprocessed to extract relevant information to construct a tidy dataset that can used for analysis. For this project, we chose 2012 as the year for analysis because based on our research on the dataset, we found that the dataset for the years 2010-2014 was relatively more complete than the other years. We have gathered the following raw data files from the above-mentioned data source and from these files, data on indicators of economic progress, as described below, was extracted:

Target Variable:

homicides-per-100000-people-per-year.csv (Homicide_Rate): Homicides per 100,000 people per year.

Primary Dataset (Economic Indicators):

1. **annual-healthcare-expenditure-per-capita.csv**: Total health expenditure is the sum of public and private health expenditures as a ratio of total population. Data are in international dollars converted using 2011 purchasing power parity (PPP) rates.
2. **gdp-per-capita-worldbank.csv:** GDP per capita adjusted for price changes over time (inflation) and price differences between countries – it is measured in international-$ in 2011 prices
3. **infant-mortality.csv:** The share of newborns who die before reaching one year of age.
4. **life-expectancy.csv:** Life expectancy at birth is defined as the average number of years that a newborn could expect to live if he or she were to pass through life subject to the age-specific mortality rates of a given period.
5. **malnutrition-death-rates.csv:** Deaths from protein-energy malnutrition per 100,000 people.
6. **median-age:** The median age divides the population in two parts of equal size: that is, there are as many persons with ages above the median age as there are with ages below the median ages

7. **size-poverty-gap-countries.csv:** The poverty gap is the amount of money that would be theoretically needed to lift the incomes of all people in extreme poverty up to the international poverty line of $1.90 a day. These estimates are expressed in international dollars using 2011 PPP conversion rates. This means that figures account for differences in prices levels, as well as for inflation.
8. **public-health-expenditure-share-GDP-OWID.csv:** Public health expenditure includes: recurrent and capital spending (central and local levels), external borrowing and grants (including donations from international agencies and NGOs), and social or compulsory insurance funds.

It is important to note that this problem is significantly constrained in terms of available data. This is because there are only about 200 countries for which data has been reported. However, the reporting of data from these countries has been inconsistent over time and we encountered significant difficulties in terms of missing data. Besides, we were constrained by requirements that the available data should be available in the time span close to the year of analysis which as mentioned earlier was chosen to be 2012. In addition, there are differences in the interplay of socio-economic factors in different countries. So, one variable which would be an important predictor for the target variable in one group of countries, would not correlate well with the target variable in another group of countries. So, as we selected groups of countries that had relatively the same features that correlated to the target, we anticipated running into the “curse of dimensionality” problem where the ratio of number of records to the number of predictors is low and which affects statistical inference.

As secondary datasets, we were able to get relevant data for one indicator on women empowerment and one indicator on educational development, which are described below:

1. **fertility-rate-complete-gapminder.csv:** Total fertility rate represents the number of children that would be born to a woman if she were to live to the end of her childbearing years and bear children in accordance with age-specific fertility rates of the specified year.
2. **Government Expenditure on education per-capita:** It is the product of GDP per capita and government expenditure on education as percentage of GDP per capita variables.

These were the economic, women’s empowerment and educational indicators used in our analysis for modeling the relationship between socio-economic factors and homicide rate.

Data Cleaning and Exploratory Data Analysis:

As mentioned earlier, we are not limited in this problem by the number of indicators but by the number of countries so as to not run into the problem of “curse of

dimensionality". While there are many predictors in the data source, for the sake of modeling the relationship between the predictor and homicide rate, an extensive exercise in choosing predictors that have high importance to the homicide rate (target variable) was carried out. To obtain predictors with high importance to the target variable, homicide rate, both Pearson and Spearman correlation coefficients between the predictors and target variable were calculated. Before doing this, we needed to clean the data. To get high quality, tidy data, it makes sense to choose only those indicators for which the number of missing values were lower. Based on our analysis of the available data in the dataset, we fixed this threshold at 15 percent. As mentioned earlier, when we explored the dataset, we started looking for the years for which there were less number of missing values (and the data set is relatively complete) and found that for year 2012 the datasets were relatively complete compared to other years. This may be because with the coming of internet in the past two decades, communication has become less costly and more efficient. Because, more information is now exchanged electronically, it is easier to mine data. So, we started looking for data for the countries in the year 2012, and even for 2012, we found that the information for all countries was not available. During data analysis, we found that the correlation between the predictors and target variable was low indicating low predictor importance to the target. An in-depth data analysis of this phenomenon showed that because of different interplay between the various socio-economic factors, there was no uniform set of predictors that correlated with the target variable for all countries across all continents. This suggested that there was a natural clustering in the data which is reasonable because of social and cultural differences between the countries on different continents. **This insight from exploratory data analysis allowed us to select countries for which predictors had a strong correlation to the target variable.** Our exploratory data analysis showed that the same predictors correlated quite well with the target variable, homicide rate, for countries in Asia and Europe. It was also found that the data for Asian and European countries has been more consistently and completely reported. AS a result of this extensive EDA, our final dataset consists of 87 countries from Asia and Europe for the year 2012.

After finalizing the countries based on this extensive EDA exercise, we were to work with only those indicators for which the number of missing values was less than the threshold (15%). It was found that data for the predictor, poverty gap, had higher number of missing values. So, we have excluded this indicator from the model. It is worthwhile to mention here that the number of countries are also reduced now, being limited to Asia and Europe, and so we need to work with less number of predictors now.

Our finalized dataset consists of 9 indicators- 7 from primary dataset and I each from the two secondary datasets.

Filling Missing Values:

- **Recover missing values:** In order to get missing values, we used additional data sources. The additional data source that we have used is: https://knoema.com/.

  Using this data source**,** we recovered the missing values for the countries of year 2012 and manually entered them. Given below is an example of recovering

```
s2_filtered_copy$gov_expen[s2_filtered_copy$Country == 'Albania'] <- 3.54
s2_filtered_copy$gov_expen[s2_filtered_copy$Country == 'Andorra'] <- 0.71
s2_filtered_copy$gov_expen[s2_filtered_copy$Country == 'Belgium'] <- 11.71
s2_filtered_copy$gov_expen[s2_filtered_copy$Country == 'Bhutan'] <- 14
s2_filtered_copy$gov_expen[s2_filtered_copy$Country == 'Cyprus'] <- 15.5
s2_filtered_copy$gov_expen[s2_filtered_copy$Country == 'Croatia'] <- 9
s2_filtered_copy$gov_expen[s2_filtered_copy$Country == 'Hungary'] <- 9
s2_filtered_copy$gov_expen[s2_filtered_copy$Country == 'Italy'] <- 8.3
s2_filtered_copy$gov_expen[s2_filtered_copy$Country == 'Kazakhstan'] <- 12
s2_filtered_copy$gov_expen[s2_filtered_copy$Country == 'Myanmar'] <- 0.42
s2_filtered_copy$gov_expen[s2_filtered_copy$Country == 'North Korea'] <- 10.7
s2_filtered_copy$gov_expen[s2_filtered_copy$Country == 'Oman'] <- 11.1
s2_filtered_copy$gov_expen[s2_filtered_copy$Country == 'Timor'] <- 9.6
s2_filtered_copy$gov_expen[s2_filtered_copy$Country == 'Uzbekistan'] <- 7.28
```

  missing values for the predictor in the secondary dataset: Government Expenditure on education per-capita**.** Similarly, procedure was followed for other predictors and our effort was to maintain the originality in data without any mean/median imputations.

- **Mean/Median Imputation:** In order to understand the distribution of values for a variable, box plots were plotted for the variable. After recovering missing values from additional data sources for a predictor, if the boxplot showed outliers, we used median to impute the missing values. If the boxplot did not show any outlier, mean was used to impute the missing values. This is consistent with the insensitive nature of the median to the outlier whereas the mean is sensitive to the presence of the outlier. This step is followed for fertility rate, annual health care per capita, GDP per capita and some other indicators which have missing values.

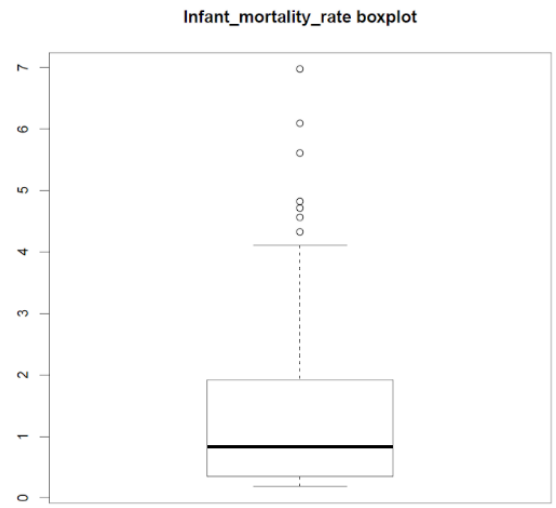
After filling in the missing values using additional data sources and imputing missing values using median/mean imputation, we observe that there are outliers for the predictors. In the figure below, one can see that there are outliers for infant mortality rate, annual health care per capita indicators. Similar situation holds for some other indicators as well. Our analysis showed that this is happening due to a greater range of values that are present for the specified indicator. For example, annual health per capita values are range from 0 to 6000 and due to this, we are getting outliers. So, we have applied transformations on the data to overcome this problem.

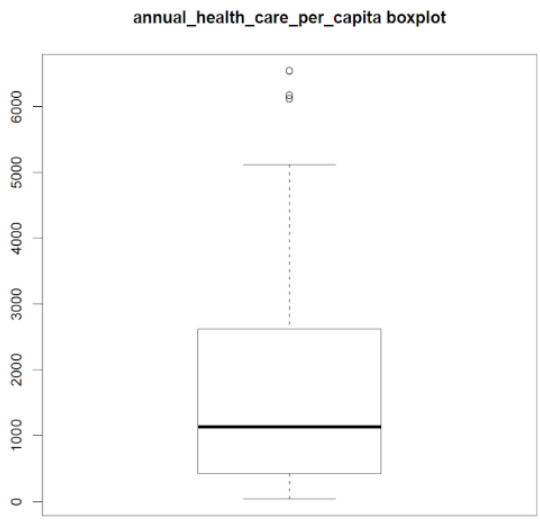
Transformations:

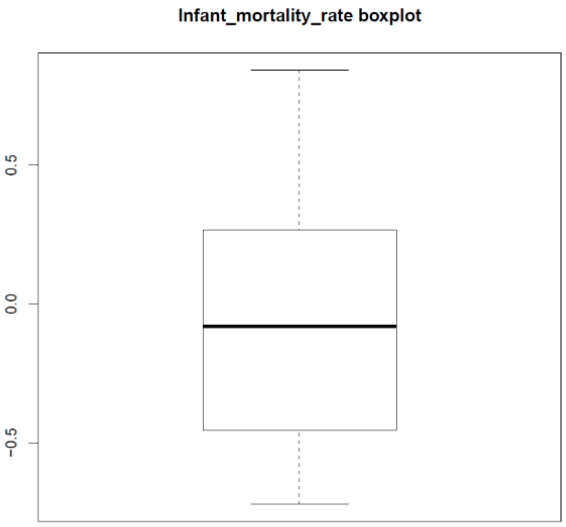
- **Log or root transformations:** By plotting the histogram for the predictor, we can check for skewness in the predictor. If the data is right skewed, we apply log

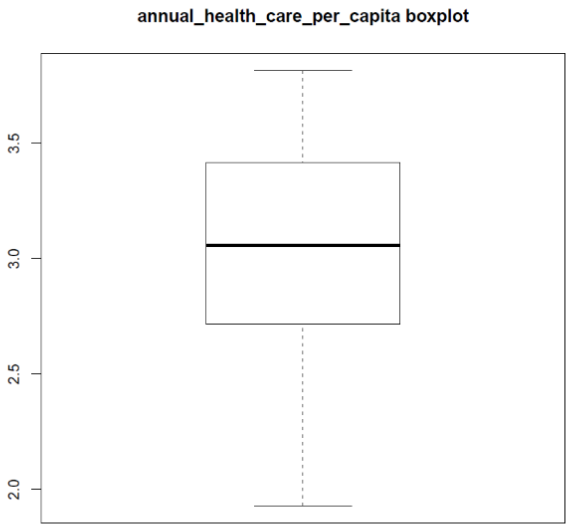
transformation check see whether the skewness is reduced and the data is normally distributed.

- **Square transformations:** If the distribution of the predictor is left skewed we apply square transformation and check whether the skewness is reduced and the distribution is nearly symmetric. Below are the examples before and after the transformations.





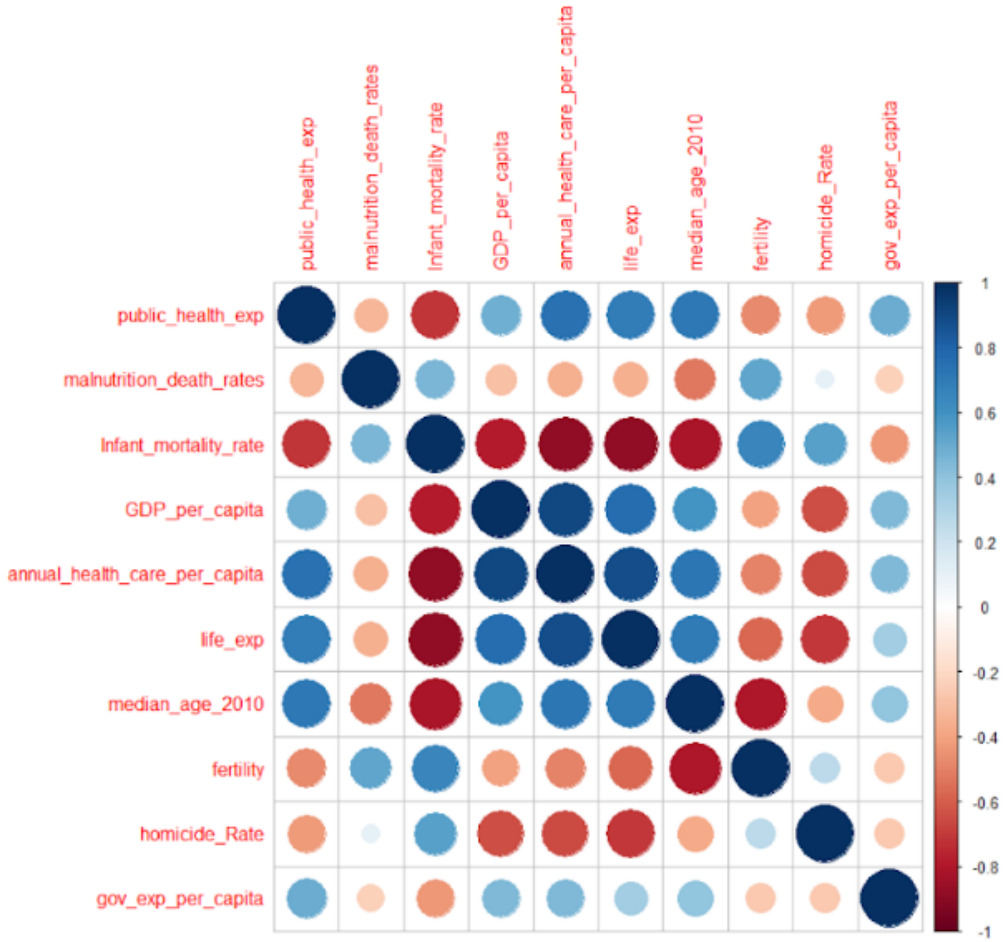




Correlations:

After cleaning the data, filling the missing values using additional data sources and median/mean imputation and suitably transforming the predictors to reduce skewness, we obtained the correlation coefficient between the predictors (in the primary and secondary datasets) with the target variable homicide rate, to ascertain their importance

to the target variable. Below are the correlation values with the target variable and correlation plot for all the indicators we are gonna use in the model building and evalutation From the correlation plot, we observe strong correlation between target and predictor values.



.

	fertility	homicide_Rate
public_health_exp	-0.4701092	-0.4259150
malnutrition_death_rates	0.5228293	0.1074761
Infant_mortality_rate	0.6598650	0.5443617
GDP_per_capita	-0.4006581	-0.6414420
annual_health_care_per_capita	-0.4963519	-0.6540179
life_exp	-0.5745626	-0.7061858
median_age_2010	-0.8044126	-0.3755759
fertility	1.0000000	0.2648918
homicide_Rate	0.2648918	1.0000000
gov_exp_per_capita	-0.2655480	-0.2641613

Data Modelling:

In choosing a predictive modeling technique for analysis, there is a tradeoff between predictive power and interpretability of the model. A complex modeling technique can capture more intricate associations between the variables but is in general also less interpretable. There is also an inherent bias-variance tradeoff and one is forced to accept a certain optimum combination of bias and variance. Keeping these considerations in mind, and because our target variable is numeric in nature, we will be using the following statistical modeling techniques in this project:

1. **Linear Regression**: Linear regression is more interpretable than ANN but depending upon the problem, may not be able to fully capture a highly complex relationship.
2. **K-Nearest Neighbor (KNN):** KNN has a simple interpretation that the target value for the input observation is the mean of the target values of its neighbors.
3. **Lasso**: Lasso is a regularization technique that is used to reduce overfitting. Lasso can also act as a variable selection technique because it can result in the coefficients of certain variables to be set to zero.
4. **Artificial Neural Networks (ANNs)**: ANN are capable of capturing extremely nonlinear relationships between the variables but they are in general less interpretable compared to methods that relate variables in an equation.

Linear Regression:

Linear regression modeling gives insight into the linear relationship between the target variable and predictors. Linear regression models are highly interpretable. In addition, because our target variable is continuous, this makes linear regression modeling a good first modeling technique to our project. Multiple linear regression was performed to obtain a relationship between the predictor variables and the target variable, homicide rate. The results of our basic linear model are presented below and it is observed that our first linear regression model had a low RMSE value.

```
Residuals:
    Min      1Q  Median      3Q     Max
-3.8499 -1.0863 -0.2425  0.7393  9.5929

Coefficients:
                              Estimate Std. Error t value Pr(>|t|)
(Intercept)                    15.3914     3.5585   4.335 5.35e-05 ***
public_health_exp              -1.9771     1.8610  -1.062   0.2921
malnutrition_death_rates        0.5153     1.2885   0.400   0.6905
Infant_mortality_rate          -7.8629     3.1568  -2.491   0.0154 *
GDP_per_capita                 -7.4359     2.8529  -2.606   0.0114 *
annual_health_care_per_capita   6.2478     3.9132   1.597   0.1154
life_exp                      -11.9214     2.8363  -4.203 8.46e-05 ***
median_age_2018                -1.4705     1.9393  -0.758   0.4511
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.142 on 63 degrees of freedom
Multiple R-squared:  0.4486,    Adjusted R-squared:  0.3873
F-statistic: 7.322 on 7 and 63 DF,  p-value: 2.04e-06
```

To increase the performance of linear regression model, several techniques were employed. First, we used the stepwise variables selection to see what variables were useful for predicting the target variable. But this was not helpful because we had only a few independent variables. We then looked back at the EDA to check for correlation between the predictor and target variables in our dataset. This showed that the independent variables were significantly correlated to our target variable and there were no outliers. After that we applied some transformations and tested some variations of our first basic regression model to check how these various variations of the regression model performed on our dataset. The results of testing various variations of the basic regression model are summarized in the table below:

Transformation	R**2 Train set	R**2 Test set	RMSE Test set	MAE Test set
Basic LM model with Intercept	0.448	0.158	3.606	2.184
Basic LM model without intercept	0.620	0.142	4.520	2.718
Log + Intercept	0.445	0.130	3.607	2.188
Log without Intercept	0.647	0.164	4.836	2.864
Range normalization + Intercept	**0.448**	**0.443**	**3.700**	**2.679**
Range normalization without Intercept	0.669	0.368	3.990	2.681

This table shows the relevant transformations we have made such as log transformation, range normalization, with and without intercept. We have used 10-fold cross validation to pick the best model out of these variants. The calculated R^2 value shows that all the transformations, except range normalization, result in overfitting. Hence, we have selected the model with range normalization as our final regression model. The 10-fold CV results of using this model for the primary, primary +first secondary dataset and primary +first secondary dataset+ second secondary dataset are presented below:

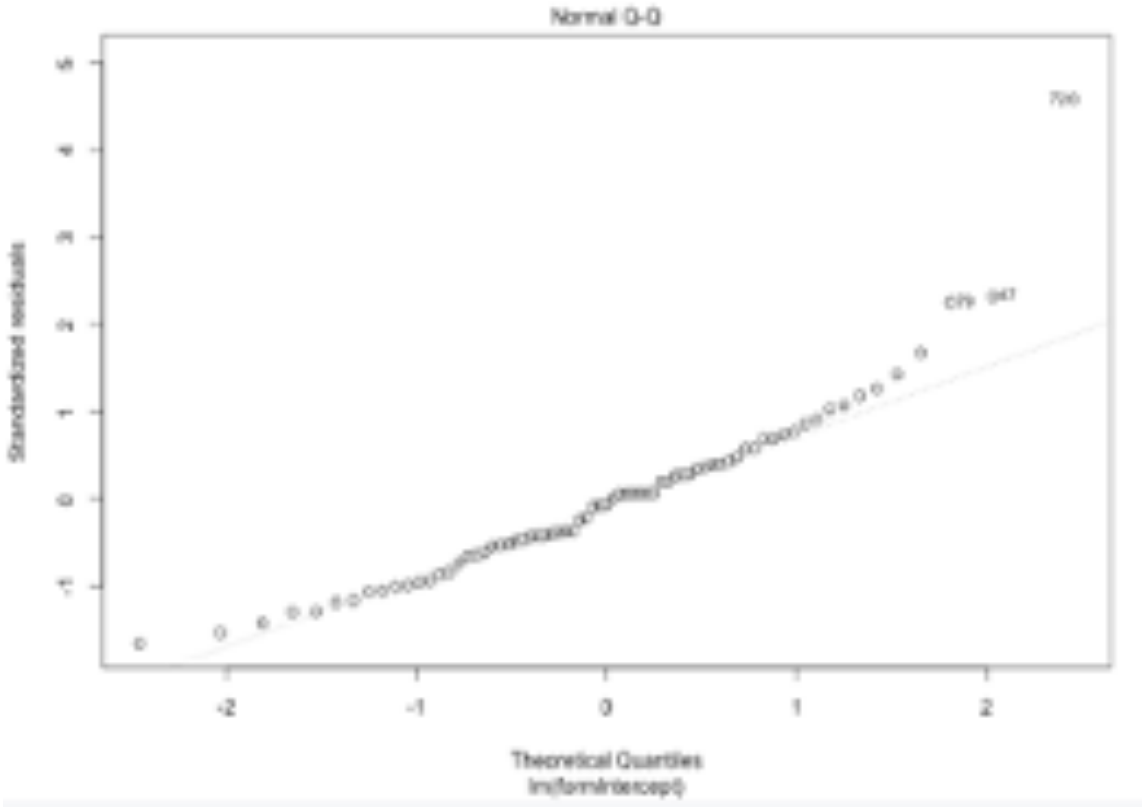
	Model 1 Primary variables	**Model 2**	**Model 3** Primary & all secondary variables

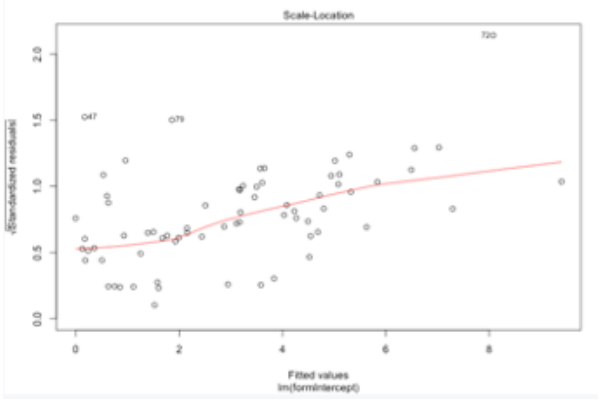
		Primary & first secondary variables	
R2**	0.82	0.92	0.94
RMSE	3.27	3.05	3.09
MAE	2.28	2.32	2.34

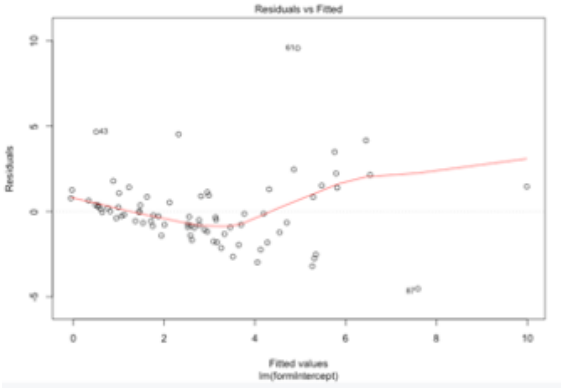
It is observed that the inclusion of information from the secondary datasets effectively boosts the performance of our model. The results from the final model are presented below:

```
Residuals:
    Min      1Q  Median      3Q     Max
-4.5312 -0.9493 -0.2415  0.8535  9.5688

Coefficients:
                              Estimate Std. Error t value Pr(>|t|)
(Intercept)                   10.31329    4.01433   2.569 0.012394 *
public_health_exp             -3.96442    1.95949  -2.023 0.046987 *
malnutrition_death_rates      -0.03266    1.25446  -0.026 0.979303
Infant_mortality_rate         -6.90449    2.90899  -2.373 0.020452 *
GDP_per_capita                -4.46978    3.14157  -1.423 0.159370
annual_health_care_per_capita  3.27543    4.30014   0.762 0.448871
life_exp                     -10.80785    3.08754  -3.500 0.000825 ***
median_age_2010                4.28099    2.70567   1.582 0.118239
fertility                      6.96647    2.74873   2.534 0.013570 *
gov_exp_per_capita             0.85409    1.56667   0.545 0.587423
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.125 on 68 degrees of freedom
Multiple R-squared:  0.4848,    Adjusted R-squared:  0.4166
F-statistic: 7.109 on 9 and 68 DF,  p-value: 3.511e-07
```

After building the model by applying range normalization to relevant variables from all of our datasets, analysis was performed on the resulting model to ensure there was no violation of the normality and the linearity requirements. Regarding the multivariate normal relationship, based on the QQplot, we can reasonably conclude that our residuals are normally distributed.







Considering autocorrelation and homoscedasticity, we observe that our residuals and our predicted values almost have a linear pattern and the variance of our residuals is almost well distributed. However, due to the fact that the performance of our model on our train set is still very low, less than 0.5, we concluded that Linear regression was not the best model for our problem.

K-Nearest Neighbors (kNN):

KNN is a technique in which predicted value for a given input is the mean of the nearest neighbors for given input. Nearest neighbors are obtained by calculating the distance between the observations. For kNN, an appropriate metric for the distance needs to be chosen. Of all distance metrics, we found that euclidean distance is performing better. Because our target variable is continuous, we use the version of kNN that is suitable for regression problems. Because distance metric is sensitive to the scale of the variable, the variables in the dataset are normalized before training the model. To normalize the data, we have applied range normalization so that all values for a variable are between 0 and 1.

An appropriate value of k needs to be chosen. k-value is determined by repeating kNN for different values of k say 1 to 30 and choosing the k value which has minimum error rate on the independent test dataset.

Next, we will split the dataset into a training and test partition using stratified random sampling. Training dataset will contain 80% of the observations and the test set will have the remaining 20%. On the test dataset we measure the model performance using Root Mean Square Error(RMSE) and Mean Absolute Error(MAE) which are defined below:

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{N}(predicted_i - Actual_i)^2}{N}}$$

$$\text{MAE} = \frac{\sum_{i=1}^{N}|predicted_i - Actual_i|}{N}$$

where N = number of obs. for the predictor.

Predicted = Predicted target Value.

Actual = Actual target Value.

We construct 3 models: model 1 trains only on primary dataset, model 2 primary + first secondary dataset, model 3 on the primary and 2 secondary datasets. The RMSE and MAE values for the model trained on the 3 datasets and tested on an independent test set are presented in the table below:

	Model 1 Primary Dataset	**Model 2 Primary&First Secondary dataset**	**Model 3 Primary and both secondarydatasets**
RMSE	4.290423	4.220351	4.191167
MAE	2.51835	2.38127	2.423485

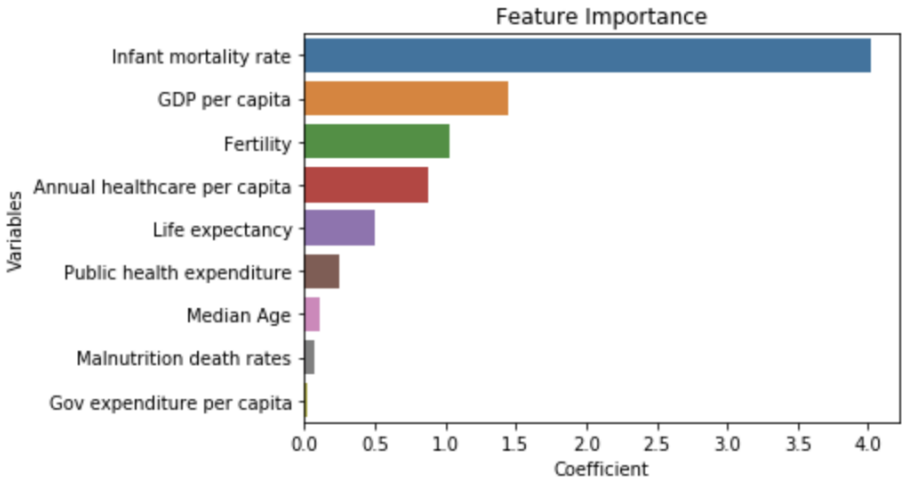
Best model is considered as the one which has lowest RMSE value. After looking at the above results, we can clearly see that model 3 has best perfomance with lowest RMSE value of 4.19 of all the models. We have used 10-fold cross validation to train the model. The results of the training model using 10-fold CV and best k value are provided below . The K value of 21 has lowest RMSE value while training so k = 21 is the final k considered.

```
k-Nearest Neighbors

71 samples
 9 predictor

No pre-processing
Resampling: Bootstrapped (25 reps)
Summary of sample sizes: 71, 71, 71, 71, 71, 71, ..
Resampling results across tuning parameters:

  k   RMSE      Rsquared   MAE
   1  3.176735  0.1117978  1.889438
   2  3.050320  0.1108825  1.870458
   3  2.919651  0.1384668  1.866654
   4  2.804012  0.1490512  1.848042
   5  2.710133  0.1789365  1.798221
   6  2.650429  0.1859210  1.771302
   7  2.568311  0.2074030  1.746238
   8  2.527453  0.2085964  1.728224
   9  2.490342  0.2182612  1.708938
  10  2.466893  0.2225903  1.691932
  11  2.473613  0.2167998  1.699658
  12  2.438888  0.2299816  1.686097
  13  2.443351  0.2263988  1.695125
  14  2.452143  0.2166183  1.710488
  15  2.438398  0.2181031  1.711089
  16  2.437664  0.2146897  1.706291
  17  2.419982  0.2218424  1.692146
  18  2.402536  0.2268575  1.682895
  19  2.402366  0.2275262  1.684685
  20  2.397100  0.2292883  1.680391
  21  2.384754  0.2401820  1.669002
  22  2.390638  0.2360477  1.667459
  23  2.391806  0.2361655  1.673871
  24  2.405795  0.2268856  1.680243
  25  2.391168  0.2368083  1.666642
  26  2.398462  0.2336247  1.674520
  27  2.403579  0.2301148  1.679687
  28  2.404915  0.2302883  1.679190
  29  2.408377  0.2271279  1.686948
  30  2.417252  0.2194820  1.691865

RMSE was used to select the optimal model using th
 smallest value.
The final value used for the model was k = 21.
```

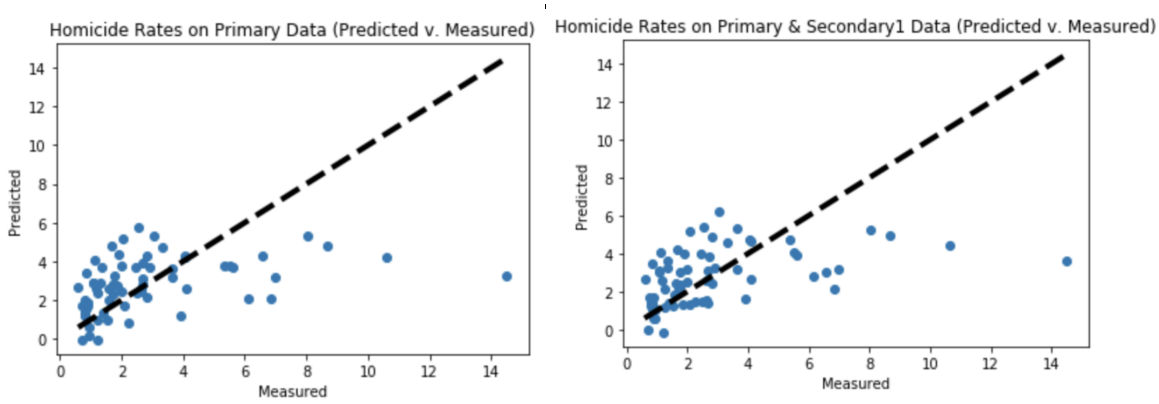
LASSO Regression:

Lasso is a regression method that is also capable of variable selection by setting coefficients of some of the variables in model to zero. Lasso imposes a constraint on the model parameters that causes regression coefficients to approach toward zero. Lasso (or L1 regularization) adds a penalty alpha to the loss function known as L1 norm. Because Lasso results in an equation, lasso results are highly interpretable. Variables with high regression coefficients represent high association with the target variable. This ensures that our model is not overly complex, is capable of doing variable selection and at the same time prevents the model from over-fitting which can result in a biased and inefficient model.

For Lasso, the **alpha** hyperparameter needs to be optimized. The parameters on our Lasso model are optimized by a 10-fold cross validated grid-search over a parameter grid. In terms of performance gains, there is approximately 3% increase in R^2 and 4% decrease in RMSE and MAE.



The chart above shows a list of features ascendingly sorted by absolute value of the regression coefficient. The correlation between the predicted and observed values for homicide rate for the Lasso model is presented below:

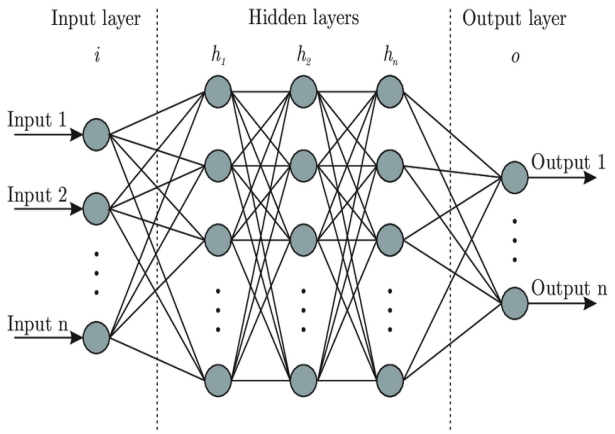


We observe that there is a good correlation between the observed and predicted values at lower values of homicide rate but large deviations from the correlation are observed at higher homicide rate.

Artificial Neural Networks:

As the name suggests, ANN algorithm is inspired by the working of brain in living organisms. ANN neural networks are capable of capturing highly non-linear relationships between the predictor and response variables and are therefore very useful for modeling highly complex relationships between inputs and outputs of the problem domain. The flip side of using ANNs is the loss in the interpretability of the model results. However, in many situations, predictive capability can be more important than interpretability. For such situations, ANN offer a highly potent solution for modeling the relationship between input and output variables and making reliable predictions even when the relationship between predictor and target variables is highly non-linear.

The structure of an ANN is presented in the schematic below:



Artificial neural network consists of neurons and interconnections between neurons. Neurons are arranged in layers: input layer, output layer and hidden layers. The input values enter into the ANN at the input nodes, these inputs are processed by the hidden layers and the output is obtained at the output nodes. ANNs are supervised learning algorithms- they need to be trained which means that the values of the parameters of the algorithm need to be optimized. These parameters are called weights in the parlance of ANNs. The optimized values of weights of an ANN are obtained by minimizing a cost function which quantifies the difference between the actual value of the output and the value of the output predicted by ANN. Multiple algorithms exist in literature to train ANN networks but the algorithm that is most commonly used is the back-propagation algorithm or its variants. In this work, we have used a variant of back propagation algorithm to train the ANN.

For a given problem, multiple ANNs can be used with different number of layers and with different number of neurons per layer. The number of hidden layers and the number of neurons per hidden layer are the hyper parameters of the ANN algorithm and need to be suitably chosen so as to avoid both overfitting and underfitting the data. The

most optimum choice of number of hidden layers and number of neurons per hidden layer for an ANN, for a given dataset, is determined by tuning these hyperparameters using 10-fold cross validation (10-fold CV). In 10-fold cross validation, the dataset is subdivided into an independent test set and a training set using stratified random sampling to maintain proper class balance between the raining and test set. The training set is further subdivided into 10 parts or folds. ANNs with different number of hidden layers and with different number of hidden neurons per hidden layer are trained on 9 of the 10 folds and then RMSE or another performance metric is calculated on the remaining fold. The process is repeated 10 times, each time taking one of the 10 folds as the validation set and the remaining 9 folds as the training set. The 10 RMSE values so obtained are averaged to get a representative RMSE value for a given anatomy of an ANN. The purpose of 10-fold CV is to control the bias variance tradeoff via averaging the results. The results are finally validated by computing the RMSE value on an independent test set to ensure that the selected ANN performs as expected from the 10- fold CV results. We have used this methodology in this project to tune the hyper parameters of the ANN and select the most optimum combination of number of hidden layers and number of neurons per hidden layer for our dataset.

The hyper parameters that were tuned for this project were the number of hidden layers and the number of nodes per hidden layers. The values of hyperparameters for ANNs that were explored using 10-fold CV in this work are presented in the table below:

Number of Hidden Layers	**Number of nodes per hidden layer**
1	1
1	2
1	3
1	4
1	5
1	6
1	7
2	(2,2)
2	(2,3)
2	(3,3)

For this work, more than 2 hidden layers were not considered because the number of predictors is quite small compared to the number of adjustable weights that are available in an ANN with large number of hidden nodes. Values of other relevant parameters of the ANNs that were used in this work are presented in the table below:

Parameter	Value
Intercept	Included
Activation function	“logistic”
Cost function	Sum of square errors
Optimization Algorithm	Resilient backpropagation with weight backtracking

In order to validate the approach as well as to test the cross-validation subroutine developed specifically for this project, the mean rmse obtained in the 10-fold CV for each of the 10 hyper-parameter values was compared with the rmse obtained on an independent test set. The results of this comparison are shown in the table below:

Number of Hidden Layers	Number of nodes per hidden layer	Mean RMSE CV-10	RMSE on test set
1	1	0.12067	0.20702
1	2	0.14408	0.11256
1	3	0.13062	0.14591
1	4	0.19020	0.14471
1	5	0.29302	0.25282
1	6	0.17163	0.21207
1	7	0.23124	0.28196
2	(2,2)	0.13929	0.23365
2	(2,3)	0.12166	0.18302
2	(3,3)	0.14277	0.21374

It is evident from the above table that 10-fold CV reliably predicts performance on unseen and independent dataset. Also as expected, the 10-fold CV rmse is lower than the rmse on the test set.

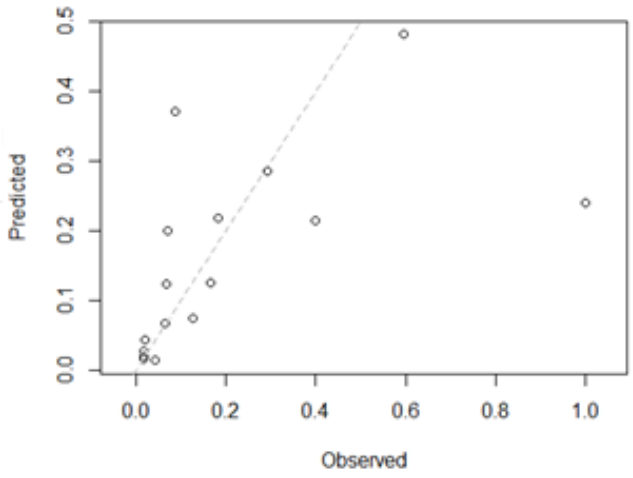
The mean RMSE values obtained from 10-fold CV as well as the RMSE values obtained on the independent test dataset are presented in table below for three different datasets:

It is observed that the RMSE values obtained for the ANN are an order of magnitude smaller than those obtained from other regression algorithms used in this work. The magnitudes of the RMSE values are consistent with one another which demonstrates the correctness of underlying computation.

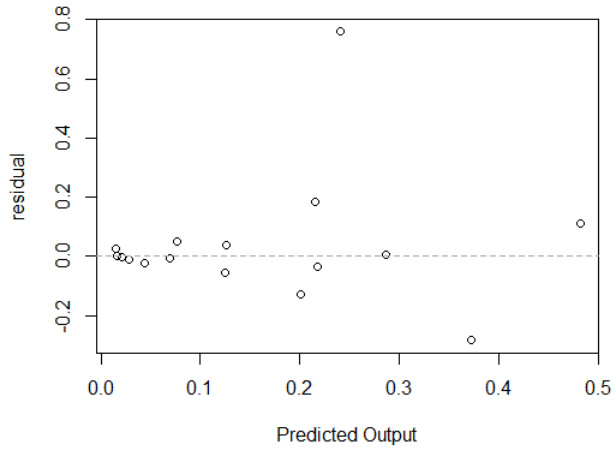
We also observe that the addition of new predictor does not have much effect on RMSE. There are several explanations for this. Firstly, ANN by their very nature have many more degrees of freedom and the dataset in this project is small compared to the number of free parameters available even with a single hidden later. Secondly, this agrees with the results we have obtained from other models used in this work which also do not show much improvement with the addition of secondary dataset. The correlation value between the predicted and actual values for an independent and unseen dataset was found to be 0.5902732 which indicates a high degree of correlation, considering the size of the dataset relative to the cultural and social diversity across different countries.

	Model 1 primary dataset	**Model 2** Primary and first secondary	**Model 3** All datasets
Mean RMSE (CV-10)	0.12067	0.13112	0.14092
RMSE (Test)	0.20702	0.19766	0.20712

The plot of observed vs the predicted values for an independent dataset is presented below:



From the above plots, it is evident that the predictions differ more from the observed values at higher observed values. The plot of residuals versus the predicted values for the ANN model is presented below:



It is observed that the residuals deviate from 0 at higher predicted values. This indicates that the model tend to be less accurate at higher predicted values which may be due to differences in the underlying socio-economic dynamics. The model predictions at low predicted values are, however, expected to be quite accurate.

Results and Discussion:

The RMSE values for the four models used in this work are presented below for comparison. We observe that ANN has the lowest RMSE amongst the four models which is one order of magnitude lower than the other modeling methods used in this work.

Model	RMSE on Test Set
Linear Regression	3.09
K-Nearest Neighbor	4.19117
Lasso Regression	3.3910
Artificial Neural Network	0.20712

This demonstrates the complexity of the socio-economic dynamics and its relationship with the predictor variable. This also justifies our choice of the four modeling techniques from the most interpretable (linear regression) to the least interpretable (ANNS) but capable of modeling highly complex relationships. Apriori, we did not had information about the extent of complexity of relationship between the socio-economic factors and homicide rate. Hence, this entire spectrum of models with tradeoff between interpretability and predictive capability was necessary. While the ANN model is not directly as interpretable as the linear regression or lasso model is, its high predictive power definitely provides useful capabilities to the various stakeholders. They can use this model to predict changes in homicide rate as the various input variables are modified. This in turn gives their decision making an informed advice regarding the investment of scarce economic resources in various socio-economic sectors. Based on its high predictive capability for this dataset, ANN model is the final chosen model for deployment

Model Deployment:

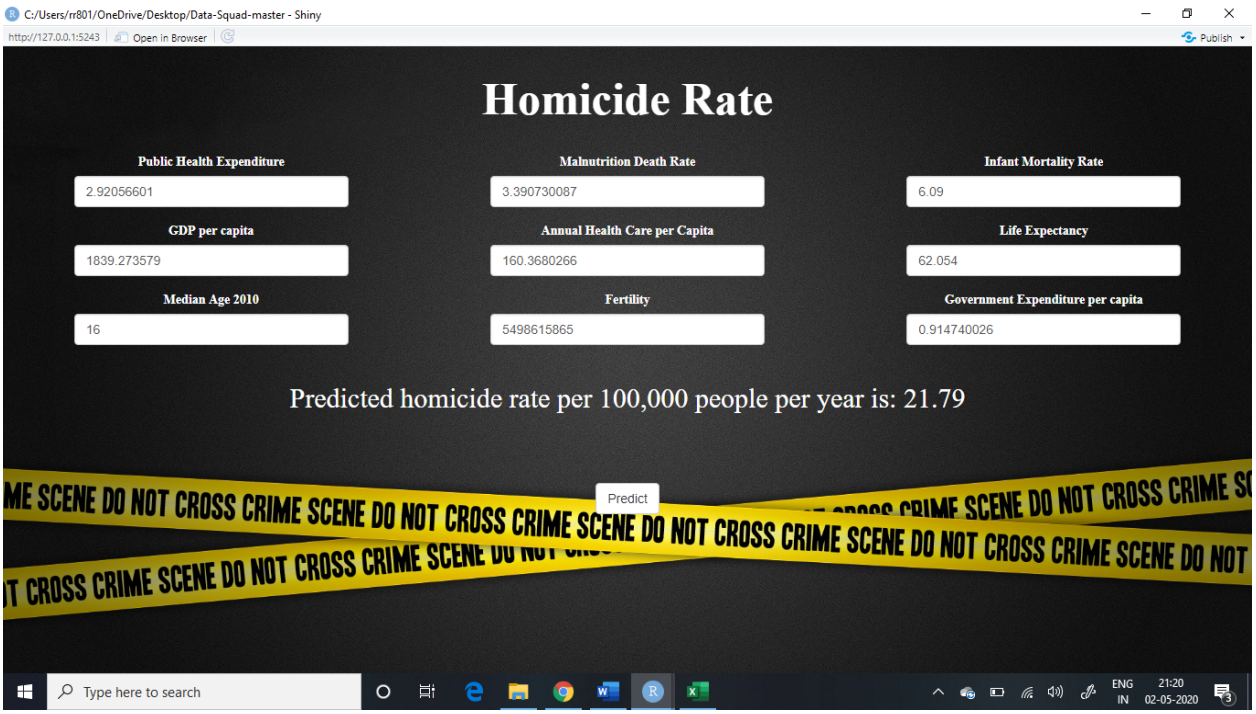
We have used Shiny for our model deployment. Shiny is an open package from RStudio, used to build interactive web pages with R. We will walk through all the steps involved in creating a shiny app. Any Shiny app is built using two components:

- **UI**: UI creates the user interface in a shiny application. It provides interactivity to the shiny app by taking the input from the user and dynamically displaying the generated output on the screen.
- **Server**: Server contains the series of steps to convert the input given by the user into the desired output to be displayed.

There are several advantages of using Shiny**:**

- **Complete Automation of the app**: A shiny app can be automated to perform a set of operations to produce the desired output based on input.
- **Knowledge of HTML, CSS or JavaScript not required**: It requires no prior knowledge of HTML, CSS or JavaScript to create a fully functional shiny app.
- **Advanced Analytics**: Shiny app is very powerful and can be used to visualize even the most complex of data like 3D plots, maps, etc.
- **Open Source**: Building and getting a shiny app online is free of cost.

When we run our code, the following web page is displayed:



We enter inputs for all nine predictors and on clicking the Predict button our model predicts the homicide rate per 100,000 people per year.

Conclusions:

In this project, we have developed a highly accurate, predictive model to predict homicide rates based on socio-economic factors. We were able to utilize already existing data, which was collected for other purposes, and leverage machine learning algorithms to squeeze higher level of utility from existing data. Although the predictive model was developed at the country level and for homicide rates, it demonstrates a general methodology that can be adapted for other units of analysis such as at state level and also for different kinds of crimes. Although we would have loved to do more, but this was a class project and we were limited by the time and resources available in a short spring semester. But in this short time, we were able to experience first hand the entire workflow and life cycle of a data science project all the way from problem formulation to deployment of a finished and working model. So, from this standpoint, this was a huge learning opportunity for us.