# CMPE-255 Term Project

(Submission by- Anay Naik, Jui Thombre and Kajal Dhanotia)

# Gun Violence Data in the USA

## Introduction

- **Motivation:**

  Gun violence in the US results in tens of thousands of deaths and injuries annually. In 2013, there were 73,505 nonfatal firearm injuries which included 11,208 homicides, 21,175 suicides, 505 deaths due to accidental or negligent discharge of a firearm, and 281 deaths due to firearms use with "undetermined intent". Gun violence is prevalent in America and has even been called a public health crisis. It is a problem unique to America where gun ownership is much more common than in other countries. This is something which is a serious issue and hence we decided to pursue this topic and predict if certain health factors have an effect on the incident rates.

- **Objective:**

  We are interested in the factors that may lead to higher incidents of gun violence. We have chosen to investigate this from the angle of health factors. There is a perception among the media and the public that mass shootings are related to mental illness, although this has largely been shown not to be the case. In addition, suicide is the most common cause of gun deaths. We want to further investigate not only mental health but all health factors to see if there is a relationship with gun violence.

- **Approach:**

1. First, we joined data from two sources to analyse this issue. These are 1) Gun Violence Data produced from data originally compiled by the non-profit organisation Gun Violence Archive and the 2018 County Health Rankings compiled by The County Health Rankings & Roadmaps program.
2. In order to join these datasets together in a reliable way, we used the county FIPS code. This is not available in the gun violence data, but the latitude and longitude values are available, and these can be used with a web API to obtain the respective FIPS codes.
3. After adding the FIPS code, we aggregated the gun violence data by county.
4. Next, we merged the data sets together, joining on county FIPS, to do further analysis.

5. We checked for outliers and imputed missing values. We also plotted feature distributions and transformed some of the data.
6. We then applied feature engineering on our dataset. We used PCA and ranked correlations for feature selection.
7. Finally, we built and tested two multiple regression models with the most important features to predict gun incidents by county health.

- **Literature/Market review:**

  All existing studies/prediction models on the gun violence data in the USA majorly focus on predicting the future of crimes or the predict the crime trend based on gun laws in a particular state but we specifically wanted to analyse if certain health factors are responsible for increased/decreased gun violence in a particular county. Few similar researches are listed below:

  https://www.kaggle.com/duttadebadri/gun-violence-in-usa-insights-forecast

  https://www.kaggle.com/shivamb/deep-exploration-of-gun-violence-in-us

  https://greenet09.github.io/datasophy/2019/08/06/forecasting-gun-violence-in-America-using-R.html

---

## System Design & Implementation

- **Algorithm(s) considered/selected**

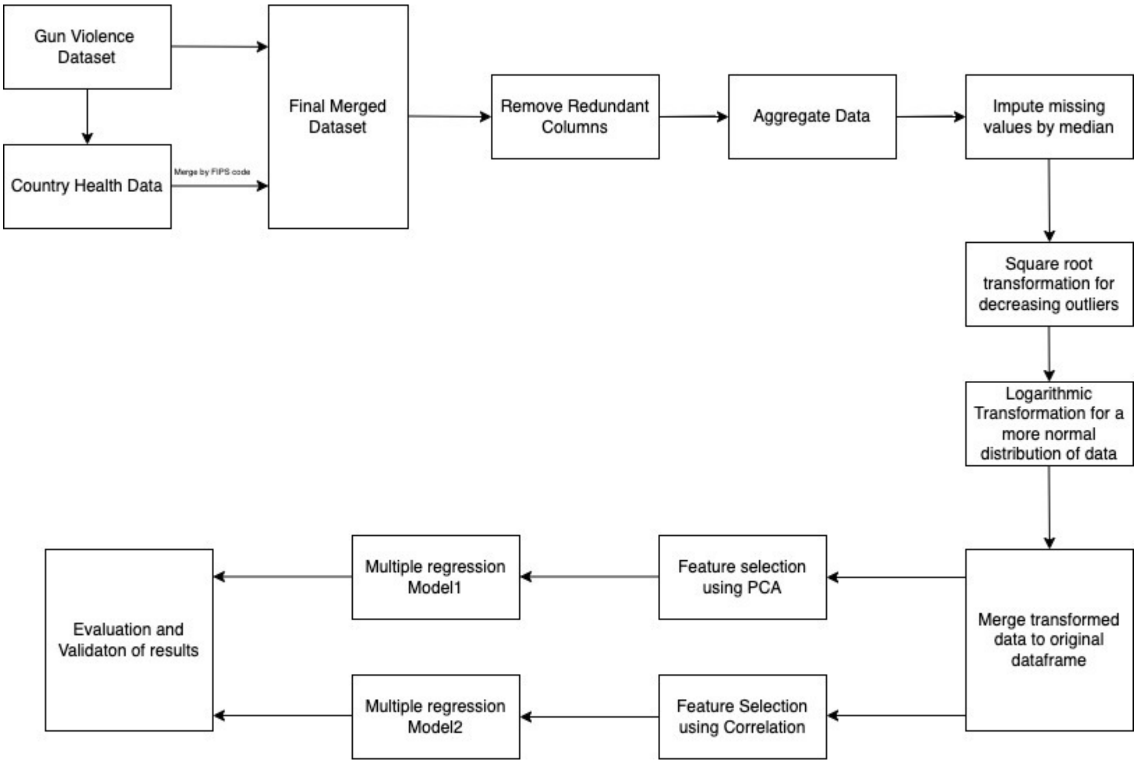
1. Aggregation
2. Standardization
3. Normalization
4. PCA
5. Correlation

- **Technologies & tools used (and why)**

1. Jupyter
2. Python libraries
3. Pandas
4. Numpy
5. Matplotlib
6. SciPy Statistics
7. Plotly

8. Calendar
9. Folium

- **System (and subsystems if needed) design/architecture/data flow:**


Gun Violence Dataset
Country Health Data
Merge by FIPS code
Final Merged Dataset
Remove Redundant Columns
Aggregate Data
Impute missing values by median
Square root transformation for decreasing outliers
Logarithmic Transformation for a more normal distribution of data
Merge transformed data to original dataframe
Feature selection using PCA
Feature Selection using Correlation
Multiple regression Model1
Multiple regression Model2
Evaluation and Validaton of results

---

## Experiments / Proof of Concept Evaluation

- **Dataset(s) used (name, source, type of data, size of data, # of instances/statistics, any preprocessing performed, etc.)**

  We are combining the gun violence dataset with the county health data.. The gun violence data was downloaded from gunviolencearchive.org. It has 239,677 observations and 29 columns for the years 2013–2018 with a row for each gun incident. To limit the scope, we filtered the file to use only the 2017 gun violence data, which contains 59,881 observations. The county health file has data for a similar timeframe (2010- 2017) but is aggregated by state and county. This file has 3,142 observations with 164 columns and 99 columns over two tabs.

- **Methodology followed (e.g., n-fold-cross validation, number of folds, size of training/test set etc.):**

We built two multiple regression models to predict the percent of gun violence per county(per population) and compared both to find the best model.

For the first model, we built a multiple regression model using the most important features of PCA. We started training the model by using the first most important feature from PCA and then eventually added more features one by one, until the value of R-squared no longer increased. Thus, we finally selected the top 8 features:

- Primary Care Provider Rate: 0.62
- Dentist Rate: 0.48
- % With Access: 0.35
- Preventable Hosp. Rate: -0.31
- Teen Birth Rate: -0.21
- Child Mortality Rate: -0.19
- Age-Adjusted Mortality 10s: -0.15
- MV Mortality Rate: -0.12

We split the data into 10 folds and used cross-validation to check the mean squared error and R-squared value. The results were consistent over the 10 folds with error of around 0.017 and R-squared at 0.86. This shows the model explains 86% of the variance in the target. Next, we used the model to predict test values, giving us an average error of 2.46%

For the second model, we built another multiple regression model with the most important features from correlation ranking. Based on the correlation matrix, the highest ranked features correlated to % gun incidents are (in descending order):

- Strong positive correlation: Chlamydia (r=0.45)
- Moderate positive correlation: % Food Insecure (r=0.37), HIV Prevalence Rate (r=0.36), % Low Birth Weight (r=0.35);
- Moderate negative correlation: Food Environment Index (r = -0.38)
- Weak Positive correlation: Other PCP Rate (r=0.285), Infant Mortality Rate (r=0.281), Child Mortality Rate (r=0.28)

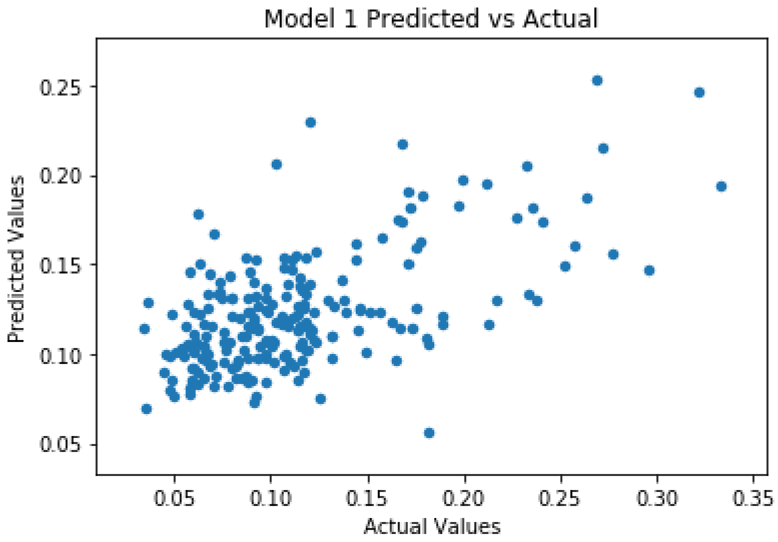
We observed that none of the features that were selected by PCA overlapped with the features of correlation ranking.

While fitting this multiple regression model we started off with the top 1 feature from the correlation ranking and kept adding more features one at a time, until we got the best values for the mean square error and R-squared value, after which it plateaued.

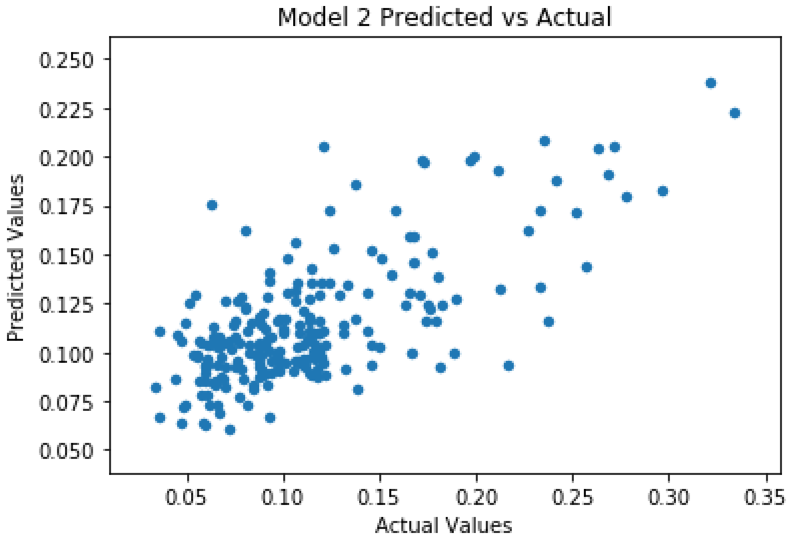
Similar to the previous model, we split the data into 10 folds and used cross-validation to check the mean squared error and R-squared value. The results were consistent over the 10 folds with error of around 0.017 and R-squared around 0.88. This shows the model explains 88% of the variance in the target. Finally, we used the model to predict test values, giving us an average error of 1.50%, which is approximately 1% lower than the previous model.

- **Graphs showing different parameters/algorithms evaluated in a comparative manner, along with some supportive text:**

  Scatter plot of the actual vs the predicted values for the multiple regression model using the most important features of PCA.



  Scatter plot of the actual vs the predicted values for the multiple regression model using the most important features of correlation ranking.



- **Analysis of results**

  Both models performed well, but Model 2 performs slightly better with a higher R-squared and lower error value.

  In Model 1, the most important features are: Primary Care Physician Rate, Dentist Rate, % With Access to Exercise Opportunities, Preventable Hospital Stays Rate, Teen Birth Rate, Child Mortality Rate, Age-Adjusted Mortality and Motor Vehicle Mortality Rate.

These eight features represent 86% of the variance in the gun violence data with an average test error of 2.5%.

In Model 2, the most important features are: Chlamydia Rate, % Food Insecure, HIV Prevalence Rate, % Low Birth Weight, Food Environment Index, Other Primary Care Providers Rate and Infant Mortality Rate. These six features represent 88% of the variance in the gun violence data with an average test error of 1.5%.

---

## Discussion & Conclusions

- **Decisions made:**

1. The first and most important decision we had to make was what to do with such a huge dataset that we had and how to make maximum sense out of it so that we can do some prediction and draw conclusions. Analysing the gun violence dataset, we realised that this dataset alone wasn't sufficient to make a prediction, so we decided to merge it with the county health data so that we are in a position to comment about the relation of gun violence incidents with the health factors affecting it.
2. We were able to find an appropriate dataset within the same time range as of gun violence data but then we had to decide how to merge those two in the most reliable way possible.
3. And the next major decision we had to make was to decide what features to remove completely. We dropped some columns such as environmental factors etc which had nothing to do with our prediction model.
4. Then we made a decision regarding which methods to use for feature selection so that we are able to build a model that gives us maximum accuracy.
5. The last decision was to build two models using the multiple regression method using features selected by PCA and correlation respectively.

- **Difficulties faced:**

1. Figuring out a way to find FIPS code from latitude and longitude data given in the gun violence dataset to merge the two datasets.
2. Aggregation of columns based on redundancy.
3. Some outliers and missing values were present and a few columns needed to be transformed to have a more normal distribution.
4. Rescale four columns to bring them in line with the rest of the data so PCA would be more accurate.

- **Things that worked well:**

1. Getting another dataset having the same time range as the gun violence dataset.
2. Model accuracy >90% for both the models built.

- **Things that didn't work well:**

1. Time and efforts taken for data cleansing.
2. Time taken to figure out transformation methods for a more normal distribution of data.

- **Future work identified based on your experience with this project**

  Identifying the other parameters on which the gun violence data may highly vary for instance food insecurity and poverty.

- **Conclusion**

  The following county health factors, grouped by categories from the original data, are related to gun violence in the US:

  1. Health Outcomes: Age-Adjusted Mortality, % Low Birth Weight, Child Mortality Rate, HIV Prevalence Rate and Infant Mortality Rate.
  2. Health Behaviours: Chlamydia Rate, Food Environment Index, % Food Insecure, Teen Birth Rate, % With Access to Exercise Opportunities and Motor Vehicle Mortality Rate.
  3. Clinical Care: Primary Care Physician Rate, Dentist Rate, Preventable Hospital Stays Rate and Other Primary Care Providers Rate.

---

## Project Plan / Task Distribution

- **Who was assigned to what task?**

  All three of us decided to divide the entire project in three equal parts, i.e. Data Preprocessing, Visualization and Model Building. We picked one section each and were responsible for end to end completion of those sections for each document (ppt, jupyter notebook and documentation). We decided on this method to distribute work to ensure we make use of our personal and academic strengths to complete and deliver the project on time.

- **Who ended up doing what task (justify as applicable):**

  The work was ultimately done as below:

- Anay Naik: Data Preprocessing and Dimensionality Reduction
- Kajal Dhanotia: Data Visualisation and Feature Engineering
- Jui Thombre: Model Building and Result Evaluation.

All three people did the presentation/jupyter and report according to the sections as mentioned above.