# **2.2 Course Project: Milestone 2--Data Selection and Project Proposal**

## **Introduction**

**Background**

**Kaggle is the subsidiary of Google LLC, is an online community of the data scientists and machine learning practitioners. For my project, I have downloaded the datasets from kaggle for further analysis, implementing machine learning techniques. I am planning to work on this data because police brutality is one of the most concern topics nowadays. So I am planning to work on it to find the actual evidence of activities (arrest, shot, killed) based on the available dataset. Basically it is providing following information on the given datasets:**

1. **Name of the person killed by police.**
2. **Date**
3. **Manner of death**
4. **Armed, Gender**
5. **Race, City, State**
6. **Sign of mental illness**
7. **Threat level**
8. **Flee**
9. **Body camera**

**More precisely, I will implement the all predictive analytical techniques to find the actual result about police involvement on arrest, shot, killed and so on.**

**Problem Statement:**

It is more or less generally accepted that “police reform” is necessary to stop to kill the people while arresting. However, there is so many factors are playing vital roles for those incidents. For example, economic status, educational status, gun control, legal boundaries, political views etc. Nowadays, lots of people are chanting about “Black lives matter” so i will also find the race wise percentage of the murder. Hence, I will go through each data points and will analyze the cause of the problems.

**Scope**

The scope of the project will focus on 2015 to 2017 of police killing datasets in USA. If I found something related data to this data, I will merge, and implement for better output and solutions.

## **Preliminary Requirements**

**Technical Approach**

This technical part consists of following steps.

1. **Data loading**: I will load the data through kaggle.com.
2. **Exploratory Data Analysis**: I will apply all EDA techniques using statistics and visualization using python and R to make it more readable and further analysis.
3. **Data Cleaning:** I will check the condition of the data and if necessary, I will clean up, transform and handle missing data too.
4. **Modeling:** I will examine the different models for best fit, applying best in class validation.
5. **Predictive Analysis:** After completing all steps, I will apply the predictive analytics for best prediction and solutions.

**Data sources or plan for data:**

Data selection is one of the most difficult parts of the project. If we select bad or insufficient data, it could lead the bad output and we might not able to find what we are looking for. Here I am planning to use kaggle data of policekilling. The source of the data is given below:

Source: <https://www.kaggle.com/kwullum/fatal-police-shootings-in-the-us?select=PoliceKillingsUS.csv>

This data has 14 columns (variables) and 2536 rows. I will implement all variables to find the solution using predictive analytic techniques.

**Analysis:**

I will analyze the main cause of the murder, rate of the murder in different cities, types of the armed, race of the murdered person, gender of the murdered person using some prediction models like regression, clustering etc.

**Requirement Development:**

Technically I need following requirements to complete this initial project plan.

1. Datasets: I need some datasets to work on it.
2. Some knowledge of EDA, Python, R, Jupiter lab and Statistics.

**Model Deployment:**

For this project model deployment is static approach because I am using this data from third party source only for the learning purpose. I guess, I cannot use this data for the deployment purpose but I will update this project to my gihub repository for future reference.

**Testing and Evaluation:**

While working on the project, testing will be performed by breaking the actual data into two sets as training and testing datasets. It will help us to check the model is well fitted or not for the future prediction. For the evaluation process I will be performing cross-validation techniques based on the training and testing datasets.

## Expected Results

I will work on the available data and will find the murdered rate based on the race to find the race based discrimination and crime rate using regression model and also find the linearity of it. Expect that, I will work on the “Armed” data to find the cause of the murdered as well as used weapons.

## Execution and Management of Project

**Project Plan**

Right now, I have no further knowledge for entire project so my plan is simply to follow the Milestone requirements and timeline.

Week 1: Milestone 1 Due (Team Information/Communication Plan)

Week 2: Milestone 2 Due (Data Selection and Project Proposal) & Peer Review

Week 3: Peer Review Week 4: Milestone 3 Due (Preliminary Analysis)

Week 5: Peer Review

Week 6: Peer Review

Week 7: Milestone 4 Due (Project Presentation & Status)

Week 8: Peer Review

Week 10: Milestone 5 Due (Final project paper and presentation) & Peer Review (Due Saturday!)

**Project Risk:**

Project risk is the uncertain event or condition that might affect the accuracy of the project output and applied models. There are no risk free projects because there are an infinite number of events that can have negative effect on the project. So we need to focus on the risk management, it is not the process of eliminating risk but process of identifying, assessing, and managing risk. While working on this project, I might encounter some sort of risk but I will identify and minimize the risk for better prediction.

**References:**

<https://pm4id.org/chapter/11-1-defining-risk/#:~:text=Project%20risk%20is%20an%20uncertain,the%20impact%20on%20the%20project.>

<https://www.kaggle.com/kwullum/fatal-police-shootings-in-the-us?select=PoliceKillingsUS.csv>