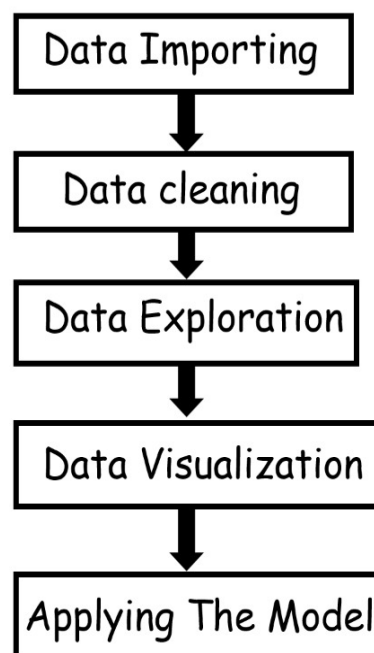# Work BreakDown

Dataset :

This dataset 'Chicago-crime-Analysis.csv' consists of information related to crimes that happened in previous years (2001 -2017). The DataSet comprises multiple columns and each column contains information like Crime Type, Crime-Location, IUCR Code(Illinois Uniform Crime Reporting) and so on. Data are extracted from the Chicago Police Department's Citizen Law Enforcement Analysis and Reporting (CLEAR) system. In order to protect the privacy of crime victims, addresses are shown at the block level only and specific locations are not identified. We typically train approximately 50000 items in each of the four Datasets and then use the Clustering model to analyze the frequency of crimes.

Attribute Information

1. ID   Unique identifier for the record   (Integer)

2. Case Number Records Division Number (Integer)

3. Date   Date when the incident occurred   (datetime)

4. Block   The partially redacted address   (Integer)

5. IUCR   The Illinois Uniform Crime Reporting code (Integer)

6. Primary Type   The primary description of the IUCR code   (Varchar)

7. Description   The Secondary description of the IUCR code   (Varchar)

8.   Location Description Description of the location (Varchar)

9.   Arrest Indicates whether an arrest was made (Varchar)

10.   Domestic Indicates whether The Incident Was Domestic-related (Varchar)

11.   Beat Indicates the beat where the incident occurred (Integer)

12.   District Indicates the police district where the incident occurred (Integer)

13.   Ward   The ward where the incident occurred   (Integer)

14.   Community Area The community area where the incident occurred (Integer)

15.   FBI Code   Special code for every crime   (Integer)

16.   X-coordinate The x coordinate of the location   (Integer)

17.   Y-Coordinate The y coordinate of the location   (Integer)

18.   Year   Year the incident occurred   (Integer)

19.   Latitude The latitude of the location where the incident occurred (Float)

20.   Longitude The longitude of the location where the incident occurred (Float)

21.   Location   The location where the incident occurred   (Datetime)

```
┌─────────────────────┐
│   Data Importing    │
└─────────────────────┘
           │
           ▼
┌─────────────────────┐
│   Data cleaning     │
└─────────────────────┘
           │
           ▼
┌─────────────────────┐
│  Data Exploration   │
└─────────────────────┘
           │
           ▼
┌─────────────────────┐
│  Data Visualization │
└─────────────────────┘
           │
           ▼
┌─────────────────────┐
│  Applying The Model │
└─────────────────────┘
```

(The Break-Down Process Diagram)

k-means clustering:

k-means is one of the simplest unsupervised learning algorithms that solve the well known clustering problem. The procedure follows a simple and easy way to classify a given data set through a certain number of clusters (assume k clusters) fixed apriori. The main idea is to define k centers, one for each cluster. These centers should be placed in a cunning way because different locations cause different results. So, the better choice is to place them as much as possible far away from each other. The next step is to take each point belonging to a given data set and associate it to the nearest center. When no point is pending, the first step is completed and an early group age is done. At this point we need to re-calculate k new centroids as bary-center of the clusters resulting from the previous step. After we have these k new centroids, a new binding has to be done between the same data set points and the nearest new center. A loop has been generated. As a result of this loop we may notice that the k centers change their location step by step until no more changes are done or in other words centers do not move any more. Finally, this algorithm aims at minimizing an objective function know as squared error function given by:

$$J(V) = \sum_{i=1}^{c} \sum_{j=1}^{c_i} \left(\left\| x_i - v_j \right\| \right)^2$$

Where:
'$\|x_i - v_j\|$' is the Euclidean distance between x_i and v_j.

'c_i' is the number of data points in i^{th} cluster.

'c' is the number of cluster centers.