* **TITLE**

**Data Cleaning Activities**

***Analyzing New York City Crime Data and Predicting Future Crime Location***

* **INTROUDUCTION**

This document explains the Data cleaning activities performed on the Dataset mentioned below in order to come up with a predictive model that could predict the possible crime location in New York City

* **DATASET**

Data available in NYC Public Data

* After school activities NYC :

[https://data.cityofnewyork.us/Education/DYCD-after-school- programs/mbd7-jfnc](https://data.cityofnewyork.us/Education/DYCD-after-school-%20%20%20programs/mbd7-jfnc)

* NYPD 2019 Arrest Dataset NYC :

<https://data.cityofnewyork.us/Public-Safety/NYPD-Arrest-Data-Year-to-Date-/uip8-fykc>

* **MODELING APPROACH**

* In The Arrest Dataset – calculated the Percentage of missing Values for each of the key columns :

PD\_CD ---> 0.02 %

PD\_DESC ---> 0.03 %

KY\_CD ---> 0.03 %

OFNS\_DESC ---> 0.03 %

LAW\_CAT\_CD ---> 0.99 %

* Since the percentage was less than 1%, the rows which had at least one NaN value, were dropped to create a Clean dataset:

Count BEFORE cleaning - 140413

Count AFTER cleaning – 138987

* For After School Dataset similar logic has been applied and rows which have at least 1 missing value were dropped from the Dataset.
* Cleaned up the BOROUGH\_COMMUNITY to get the exact BOROUGH Name from the Zip Code