Check point

Background information:

i. NYPD MVC done this before.
ii. It is important because we can learn from data that when and where is the most accidents happened, and we can send more police officers to the place at that time to control the traffic which can do not waste the police officers time and reduce accidents happening.
iii. Challenges: 1. how to reduce the attributes and dimensional of data.
    2. how to deal with miss data.
    3.
iv. No ethical considerations.
v. we are helping NYPD to predict the accidents happen and make them using less police officers to reduce more number of accidents.
vi. Too much attributes, and some data are missing, what attributes did not show on websites but we need to consider like weather, how to deal with special case(outlier).
vii. From website data.cityofnewyork.us which the calculation by government.
viii. data is mess and need to be cleaned so that can be used to predict.
ix. I plan to use cross-correlation, PDM and agglomerative clustering to reduce attributes. Then using K-NN deal with missing data or just delete it.
x. When data were cleaned, using K-mean to find the number of clustering by time and place. Therefore, I will use the algorithms of K-means.
xi. How many data or which year of data we using for analysis.
xii. For location, we use L1 norm distance metric to compare the latitude and longitude. For time, we use hour in a day, day in a week, and month in a year to see when is the most number of accidents.

Plot out graphs (at least this graphs should be plot out):

1. Number of accidents vs time. (daily, monthly, yearly)
2. Number of accidents vs place.
3. Number of accidents vs types of accidents