**COSC 4301/5340 Data Science and Big Data Analysis**

**Final Project**

**Topic** :

Analysis and Data Mining two large data sets using various Visualization Techniques

**Team** :

1. Rohit A.Khadse (L20407353)
2. Saurabh Mahajan (L20\_\_\_\_\_)

**Data Set to be used:**

<http://catalog.data.gov/dataset/campus-safety-and-security-survey-2013>

We are going to use the following two data sets from the above link:

1. OnCampusArrest
2. NonCampusArrest

**Description:**

1. OnCampusArrest  
   *24 attributes and 11064 tuple entries  
   The data set shows the tabular data of arrests that took place in various colleges ON CAMPUS*
2. NonCampusArrest  
   *24 attributes and 11064 tuple entries  
   The data set shows the tabular data of arrests that took place in various colleges NON CAMPUS*

**Task:**We are going to analyze the two data sets and interpret the areas where the arrest rate is high i.e. either On Campus or Non Campus. We will also use different visualization techniques to analyse datasets and variations in it.

**Techniques and Algorithms to be used:**   
We are going to perform visualization techniques like Histogram, Scatter Plot, Bar Graphs, Pie Chart and Heat Map. We will also apply k-means clustering technique.

**Implementation:**

We will be RStudio for R Programming and MySQl workbench for MySQL development.

**Output:**   
We will be submitting the datasets used, source code files, output image files and presentation.