PUBLIC SAFETY AT IIT

Manu Sreekumar, Mathew Athoopallil, Mikhail Rybachenko, Parker Joncus

Introduction:

Campus safety is one of the top factors that goes into consideration by students while opting for the university for their future studies. Illinois Institute of Technology, a privately funded university with an attendance of 7,266 students attracts students from across the world. The campus geographically lies on the south side of Chicago downtown, which is considered as a crime hotspot. During a time period of 2015 - 2017 , more than 7200 incidents were reported within and surrounding areas of the campus (surrounding area within two blocks of the campus). IIT Public Safety ensures the safety of students within the campus. However, more than 30% of the students are residing outside the campus in private establishments and they have to commute through the surrounding areas as well. Hence, the safety of the campus would be reflected as the safety of both the campus and the surrounding areas. Our project is aimed at collecting the records of criminal incidents that happened within and outside the campus, and attempt to predict the possible future hotspots of crime for helping the public safety officials to concentrate their efforts.

Problem Statement and Importance

The business problem that we are trying to solve using the project is: "Determine locations for deployment of public safety measures to effectively reduce incidents rates on campus and the surrounding areas". The importance of solving the business problem could be listed as follows :

- Identification of zones/sectors of serious incidents will enable public safety department to effectively deploy their forces
- The surrounding areas with higher chances of incidents should be brought under surveillance to ensure the safety of students residing in these areas.
- Effectively bring down the crime rates happening within and around the campus to improve the "safety" aspect of the university and thereby attract potential students across the globe.

Plan of Action

The following plan of action was formulated for carrying out the project execution :

- Collect records of incidents that happened within and surrounding areas of campus within two blocks for time period of 2015 - 2017
- Collect historical weather data for the incident locations for the time period

- Getting the latitudes and longitude information for the incident address
- Data munging for removing empty values, correcting errors, standardizing incident types and preparing dataset in standardized format in chronological order
- Exploratory Data analysis of the features present in the dataset
- Segregating the incident types into two classes based on Severity and build Classification model
- Train the model on 2015 - 2017 training data, and test the model accuracy and results using the recent crime data
- Create a prediction app that maps 1- Day, 3-Day and 5-Day prediction of incidents and depict the crime hotspots in IIT Campus area

Datasets Used:

The datasets that we have used for basing the results of our studies are the following :

- IIT Public Safety Crime logs for the incident records of the campus from 2015 - 2017.
- Chicago Crime data portal is used as source of incidents in the surrounding area (within two mile grid) for the same time period
- Underground Weather API for extracting the historical weather data.
- Google API for getting Latitude and Longitude information
- Google Maps for getting the location map of IIT Area

Web-Scraping:

IIT Public Safety Blog:

First, we've found out whether we need special permission to web-scrape data from IIT Public Safety Blog, and there were no limitations for that at the time of communication.
Second, we need to examine the structure we are going to web-scrape. Incidents that occur at IIT campus are published in form of blog posts, available at https://blogs.iit.edu/public_safety/.
Each post can have several incidents reported or message that there are no incidents to report.

All the archive incidents are stored by months and can be accessed from web or by adding year and month in the numerical form separated by '/', for example, to view incidents for January 2015 the link will be the following: https://blogs.iit.edu/public_safety/2015/01/
Also, if all the incidents for the selected month don't fit on one page, there will be several pages for the month which can be accessed by adding '/page/2', where '2' is page number:
https://blogs.iit.edu/public_safety/2015/01/page/2/
Another thing we need to identify is how posts are stored on a web-page. In our case each post is stored in element of class 'content', which we are going to extract from web-pages for each month we are analyzing.

And the last thing we need to identify how each post will be splitted into parts. We will extract:
- Incident type
- Location
- Date occurred
- Disposition
- Notes

Note that posts are inconsistent in formatting and information given, for example, in the earlier posts there will be no 'Disposition' information. In addition to that, location information part starts with 'ILLINOIS INSTITUTE OF TECHNOLOGY', which is not explicitly saying us that all posts location information will start from that. So, in the extraction process, we've also used an approach, when we've first extracted location+date ('ILLINOIS INSTITUTE OF TECHNOLOGY : MIES CAMPUS : CAMPUS WIDE 10/30/2017 03:30 AM') and then by subtracting date of occurence obtained the location.

Weather Data:

We found many different weather data candidates for past data, however many of them came at some cost. We determined that the best weather API to use was Underground Weather (www.wunderground.com). They did have plans that were available for purchase, but they also allowed free access as long as we stay under 500 calls per day and 10 calls per minute. One of the deciding factors for using Underground Weather's API is that they have created an R package to easily make calls to the API. The package is called 'rwunderground'. In order to get this package, you must use devtools and install from github with the command:

$$devtools :: install\_github('ALShum/rwunderground') \qquad (1)$$

You can then use commands within the package to set the key, set the location, and get the historical data given the key, location, start date, and end date. Since we had obtained a free key, we could only call the API 500 times a day. Because of this, we made sure the set the start date and end date so that there are only 500 days in between. This meant that in order to get all of the data that we wanted for the 3 years, we needed to run this code on 3 separate days and then merge the data. The history function to get the historical data had a convenient argument that limits the calls to 10 per minute so that we do not go over the free limit. We also monitored to total calls per day to ensure that we did not go over the total calls in a day. Also, since we had to do this over three days, the datasets were obtained and saved manually each day. The readme for the rwunderground package supplied easy instrctions for how to make calls to the API (https://cran.r-project.org/web/packages/rwunderground/README.html).

Crime Portal Data:

The crime data for the surrounding areas of IIT (Kent and Meis) was obtained from the Chicago Data Portal. The bounds was a 2 mile grid centered at the Mies Campus and a 1 mile grid centered at Kent Campus. The latitude bounds were 41.8603, 41.8025 longitude bounds were -87.66687, -87.58759 for Meis campus, and, latitude bounds were 41.89363, 41.86465 and longitude bounds were -87.66203, -87.62239 for Kent campus. Using these bounds all the data from 2015 onwards was pulled from the crime data portal to incorporate it into further steps including EDA and the prediction model.
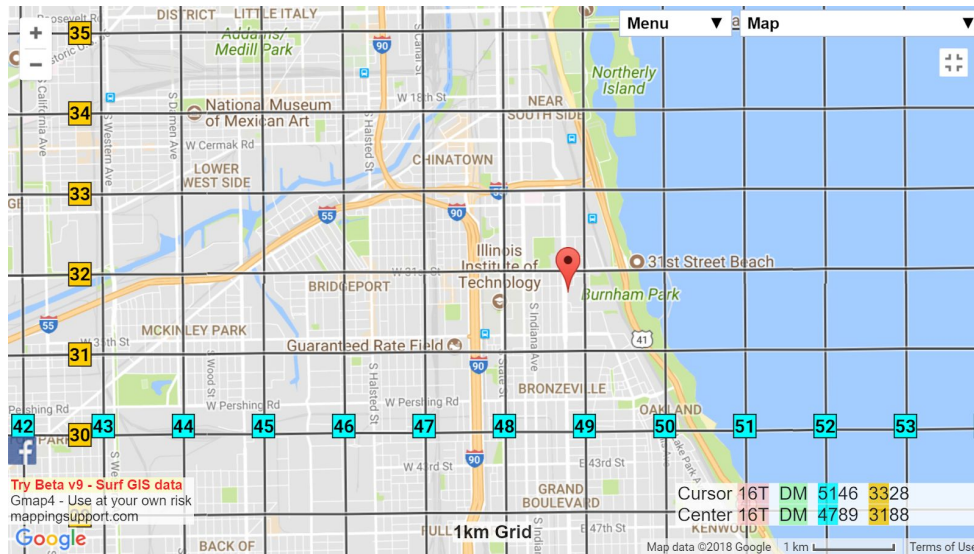
Data Preparation and Munging:

IIT Crime Logs Dataset: Creating Structured Data

Web-scraping data (*get_web_data.R*)

We will follow the steps below to extract data from IIT Public Safety blog:

1. Set the time frame for incidents we are going to web-scrape (some incidents may be posted with 2-3 days delay). We are web-scraping incidents from '2015-01-01' to '2018-01-01'.

2. Obtain links for the time frame set (without pages, yet):

#[1] "https://blogs.iit.edu/public_safety/2015/01/" "https://blogs.iit.edu/public_safety/2015/02/"
#[3] "https://blogs.iit.edu/public_safety/2015/03/" "https://blogs.iit.edu/public_safety/2015/04/"
#[5] "https://blogs.iit.edu/public_safety/2015/05/" "https://blogs.iit.edu/public_safety/2015/06/"

3. Obtain all additional pages that exist for the links for time frame:
"https://blogs.iit.edu/public_safety/2015/01/"   "https://blogs.iit.edu/public_safety/2015/01/page/2/"
"https://blogs.iit.edu/public_safety/2015/01/page/3/"
"https://blogs.iit.edu/public_safety/2015/02/"
"https://blogs.iit.edu/public_safety/2015/02/page/2/"
"https://blogs.iit.edu/public_safety/2015/03/"

4. Extract all the posts for the time frame for all links (one element (string) can have multiple posts):

#[[86]]

#[1] "\n\t\t\t\t\t"

#[2] "\r\n\t\t\tthere is no information to report for the iitpsd public crime log for 1-8-2018.\n\t\t\t\r\n\t\t"

#[3] "\r\n\t\t\tgood morning and welcome back illinois tech!\nbelow is the public crime log for friday, january 5 thought sunday january 7 2018\nincident type:harassment : phone\nillinois institute of technology : kent campus – 565 w adams 1/5/2018 10:33 am\ndisposition: information only\nnotes: iitpsd responded to a phone harassment call.\n\nincident type:alarm : fire\nillinois institute of technology : vandercook 2 – 3125 s federal 1/5/2018 05:00 pm\ndisposition: facilities notified\nnotes: iitpsd responded to a call of an actived fire alarm at vandercook #2.\nincident type:assault-aggravated\nillinois institute of technology : vandercook 2 – 3125 s federal 1/6/2018 11:40 pm\ndisposition: cleared by arrest\nnotes: iitpsd responded to 31st and federal for a wellbeing check of a member of general public.\nincident type:utility incident : water\nillinois institute of technology : \mies campus : academic/administrative buildings : siegel hall – 3301 s dearborn 1/7/2018 04:36 pm\ndisposition: housekeeping notified\nnotes: iitpsd responded to siegel hall for a utility incident involving water.\nthank you\n\n\t\t\t\r\n\t\t"

5. At this point we will obtain number of incidents for the timeframe to check if it matches with number of incidents obtained for analysis.
6. Split the extracted content to singular posts:

#[997] "incident type:assist other agency/department\nillinois institute of technology : mies campus : street locations : state-2/5/2016 08:30 pm\ndisposition: police notified\nnotes: iitpsd responded to 31st and state street for a report of shots fired in the area involving members of the general public.\n\nthank you,\n \n\t\t\t\r\n\t\t"

#[998] "\r\n\t\t\thello,\n \nthere is no information to report for the   iitpsd public crime log for 2/4/2016.\n \nthank you,\n \n\t\t\t\r\n\t\t"

7. Take the vector of posts, extract data and puts in a dataframe, removing NA's

	Incident	Location	Occured	Disposition	Notes
2449	drug/narcotic offense	illinois institute of technology : mies ca...	1/28/2018 05:02 pm	housing notified notes: iitpsd responded ...	iitpsd responded to a call of a strong odor of
2451	injured/sick person : non–emergenc...	illinois institute of technology : mies ca...	1/25/2018 10:27 am	cleared – no further action notes: iitpsd r...	iitpsd responded to metals for a non–emerge
2452	alarm : smoke	illinois institute of technology : mies ca...	1/25/2018 09:00 pm	housing notified notes: iitpsd responded ...	iitpsd responded to msv for an activated fire
2454	damage to property	illinois institute of technology : mies ca...	1/24/2018 03:40 pm	cleared – no further action notes: iitpsd r...	iitpsd responded to farr hall for a damage to
2457	injured/sick person : emergency me...	illinois institute of technology : mies ca...	1/16/2018 01:48 pm	cleared – no further action notes: cfd res...	cfd responded to 3424 s. state street for a yo
2458	administrative information	illinois institute of technology : mies ca...	1/16/2018 02:55 pm	cleared – no further action notes: iitpsd r...	iitpsd responded to center court at mtcc for a
2460	injured/sick person : emergency me...	illinois institute of technology : mies ca...	1/15/2018 12:04 pm	cleared – no further action notes: iitpsd r...	iitpsd responded to hermann hall for an emer
2461	alarm : fire	illinois institute of technology : mies ca...	1/15/2018 03:04 pm	cleared – no further action notes: iitpsd r...	iitpsd responded to the lsr for an activated fir
2463	injured/sick person : emergency me...	illinois institute of technology : mies ca...	1/13/2018 01:10 am	housing notified notes: iitpsd responded ...	iitpsd responded to 33rd and michigan for an
2464	alarm : fire	illinois institute of technology : mies ca...	1/13/2018 03:54 pm	housing notified notes: iitpsd responded ...	iitpsd responded to an active fire alarm at the

8. In the final steps of web-scraping, remove white spaces at start and end of values, remove escape characters, characters from the set [-,–,:] at the end of string and convert date of occurrence to POSIXct format.
9. Check if number of incidents obtained in point 5 above match number of incidents in the dataframe of all incidents. For the 3-years period, we've obtained 1604 incidents.
10. Save dataframe as .rda for further processing

As an output of the web-scraping step, we've obtained a data frame of incidents with the following columns:
- Incident
- Location
- Occurred
- Disposition
- Notes

Getting address for location (*iitCrime_getAddress.R*)

Once we've looked through the obtained incident's location information, we've realized, that not all incidents have the exact address, but can have an IIT Campus building name as a way to specify incident location:

Incident Type:INJURED/SICK PERSON : EMERGENCY MEDICAL TRANSPORT/AMBULANCE : STUDENT

ILLINOIS INSTITUTE OF TECHNOLOGY : MIES CAMPUS: CARMEN HALL – 5/26/2016 01:23 AM

Disposition: HOUSING NOTIFIED

Notes: IITPSD responded to Carmen Hall for an emergency medical transport of an intoxicated IIT student.

Or even abbreviated name, like 'mtcc' for 'McCormick Tribune Campus Center'.
For such cases we've built a dictionary (*iit_dict_buildings.csv*) of all IIT Buildings and abbreviated names with corresponding addresses.

Some of the incident reports have the location specified in informal way, for instance:

Incident Type: ACCIDENT : MOTOR VEHICLE

ILLINOIS INSTITUTE OF TECHNOLOGY : MIES CAMPUS : STREET LOCATIONS- 11/10/2015 02:00 PM

Notes: IITPSD responded to 34th St. and State St. for a report of an IIT student that was struck by a vehicle in a hit and run accident.

Such information can not be used straight away to determine incident's latitude and longitude, As we consider incidents on campus and surrounding area, we've built a dictionary of streets (*iit_dict_ns_streets.csv*), that go through our area of analysis from north to south (Michigan, State, Wabash etc) with 'South' add-on.

street	address
michigan	south michigan ave
wabash	south wabash ave
state	south state str
lasalle	south lasalle str
la salle	south lasalle str
federal	south federal str
dearborn	south dearborn str

Third major group of incidents without exact address are incidents occurred at parking lots, that are specified as:

Incident Type: ASSAULT-SIMPLE

ILLINOIS INSTITUTE OF TECHNOLOGY : MIES CAMPUS : GENERAL PARKING LOCATIONS : B2 LOT -11/10/2015 10:00 PM

Notes: IITPSD responded to a physical altercation involving two IIT students.

For such reports, will be using a dictionary of parking lots, but for parking lots (*iit_dict_parkings.csv*) we will return latitude and longitude straight away pointing at the center of parking lot, otherwise in case of using address, we will get an intersection of two streets near the parking lot, which is misleading for our analysis.

parking	address	Latitude	Longitude
A1 lot	2939 south state	41.83979	-87.626358
A2 lot	3040 s wabash	41.838987	-87.626262
B2 lot	W 31st street and South Federal	41.838044	-87.628779
B3 lot	3140 South Federal St	41.836824	-87.629417
B5 lot	3144 South Federal St	41.835832	-87.629165

We've also included two CTA stations in this dictionary - 'red line and 35th' and 'green line and 35th/34th'.

We will follow the steps below to obtain exact address for the location specified by Public Safety officers:
1. Load data we've web-scraped with **get_web_data.R**
2. Load dictionaries of IIT buildings, parking lots and streets ('iit_dict_buildings.csv', 'iit_dict_parkings.csv', 'iit_dict_ns_streets.csv').
3. Make sure that we don't have multiple space between words and both ends in location and notes columns we will be working with.
4. Convert dictionaries records to lower case.
5. Add new columns to the data frame of incidents: *"Address", "Latitude", "Longitude"*
6. Apply IIT buildings dictionary to location column to obtain exact address. We've managed to get an address for 1038 of 1604 observations.
7. Apply IIT parking lots dictionary to location column to obtain latitude and longitude. We've managed to get an address/or location for 1127 of 1604 observations.

8. Apply IIT buildings to notes column to obtain exact address. We've managed to get an address for 1252 of 1604 observations.
9. Apply IIT parking lots dictionary to notes column to obtain latitude and longitude. We've managed to get an address/or location for 1258 of 1604 observations.
10. Search for address in location with regex, for example, search for address of type '3424 s michigan', replace 's' with 'south' and add address value for instances with empty address. We've managed to get an address for 1322 of 1604 observations.
11. Search for intersections of streets in location and notes columns with IIT streets dictionary. We've managed to get an address for 1428 of 1604 observations.
12. Save dataframe as .rda for further processing.

Standardization of Incidents:

IIT Incident Logs Data

From the extracted data of IIT-logs the incident type in many cases was not the actual incident type. For instance, take a look at the image below, the incident type logged was Trespassing, however, going through the Notes attribute of the log it is clear that it was an Assault.

Public Crime Log for Friday, November 18th through Sunday November 20th 2016

November 21st, 2016   Logged in User   💬 Comments off

Good morning IIT,

Below is the public crime log for Friday, November 18th through Sunday November 20th 2016

Incident Type:TRESPASSING : CRIMINAL TRESPASS TO LAND
ILLINOIS INSTITUTE OF TECHNOLOGY : MIES CAMPUS : ACADEMIC/ADMINISTRATIVE BUILDINGS
: MCCORMICK TRIBUNE CAMPUS CENTER – 3201 S STATE 11/18/2016 08:22 AM
Disposition: POLICE NOTIFIED
Notes: IITPSD responded to 7/11 store for a report of a assault to a store employee.

To fix the problem text analysis was performed by identifying observations that contained key words associated to a particular incident type. For instance, cluster of words "medical", "injury", "slip", "emergency", "transport", "injur" and "fainted" were tagged as "MEDICAL INCIDENT". Below is the cluster words associated with the different incident types,

"marijuana", "narcotic", "drug" => "NARCOTICS"
"robb" => "ROBBERY"
"alarm", "smoke", "fire" => "ALARM"
"assault" => "ASSAULT"
"battery" => "BATTERY"
"theft","stolen","burglary","larceny" => "THEFT"
"water","power","utility","elevator","entrapment","surge","electricity","gas"=>"UTILITY INCIDENT"
"well", "check", "friend", "worried" => "WELL BEING CHECK"
"disturbance", "noise", "complaint","disorderly","hazing" => "DISTURBANCE"
"accident", "vehicle", "hit and run" => "ACCIDENT"

"damage", "property", "vandal","window" => "DAMAGE TO PROPERTY"
"trespass","enter" => "TRESPASS"
"alcohol","liquor" => "LIQUOR LAW VIOLATION"
"suspicious" => "SUSPICION"

Once this was completed the remaining records which had not been associated with any word clusters were assigning to new types such as "HOMICIDE", "MISSING PERSON", "OTHER", "SEXUAL CRIME", "STALKING", "WEAPON".

INCIDENT_TYPE1	Notes
ACCIDENT	iitpsd responded to 31st & wabash for a student who was s...
ALARM	iitpsd responded to msv for an active fire alarm. it was later...
ASSAULT	iitpsd responded to ycla for a simple assault against a teac...
BATTERY	iitpsd took a report of two students involved in a physical al...
DAMAGE TO PROPERTY	iitpsd responded to the mccormick tribune campus center d...
DISTURBANCE	iitpsd responded to youth connection leadership academy, f...
HOMICIDE	iitpsd responded to the a2 lot for a criminal homicide involv...
LIQUOR LAW VIOLATION	iitpsd was dispatched to mccormick student lounge east hal...
MEDICAL INCIDENT	iitpsd responded to 3424 for a emergency medical transport.
MISSING PERSON	iitpsd took a report of a missing iit student.
NARCOTICS	iitpsd responded to msv for smell of marijuana. thank you
OTHER	iitpsd responded to the 17th floor of the tower to investigat...
ROBBERY	iitpsd took a strong armed robbery report involving an iit s...
SEXUAL CRIME	iitpsd responded to unauthorized solicitation activity near 3...
STALKING	iitpsd responded to a report of stalking at crown hall.
SUSPICION	iitpsd responded to farr hall, for a report of suspicious acti...
THEFT	suspected theft from the barnes & noble bookstore kent co...
TRESPASS	iiitpsd responded to a possible criminal trespass of a mem...
UTILITY INCIDENT	iitpsd personnel responded to hermann hall for an elevator...
WEAPON	iitpsd gathered information on a shooting that occurred off...
WELL BEING CHECK	iitpsd responded at ssv middle for a well being check. tha...

Chicago Data Portal

For the final dataset to have identical incident type classification the crime data from the Chicago Data Portal was standardized by using merge operation. First a dictionary was created for the unique incident types in the crime data to be matched, merge_IIT_Crime.xlsx for Meis and merge_IITKent_Crime for Kent campuses. The merge operation was then performed resulting in a new column with Standardized incidents.

Reclassification of Incidents:
Once this standardization was performed the incidents were reclassified into two broad categories, "MILD INCIDENT" and "SERIOUS INCIDENT". Serious incident were chosen as the incidents where there is an immediate threat to life/valuables. Mild incidents on the other hand were classified on the basis that there was no immediate threat to life by another person by another person.Classification shown below.

INCIDENT_TYPE1	INCIDENT_TYPE2
ACCIDENT	SERIOUS INCIDENTS
ALARM	MILD INCIDENTS
ARSON	SERIOUS INCIDENTS
ASSAULT	SERIOUS INCIDENTS
BATTERY	SERIOUS INCIDENTS
DAMAGE TO PROPERTY	SERIOUS INCIDENTS
DISTURBANCE	SERIOUS INCIDENTS
HOMICIDE	SERIOUS INCIDENTS
KIDNAPPING	SERIOUS INCIDENTS
LIQUOR LAW VIOLATION	SERIOUS INCIDENTS
MEDICAL INCIDENT	MILD INCIDENTS
MISSING PERSON	SERIOUS INCIDENTS
NARCOTICS	SERIOUS INCIDENTS
OFFENSE INVOLVING CHILDREN	SERIOUS INCIDENTS
OTHER	MILD INCIDENTS
ROBBERY	SERIOUS INCIDENTS
SEXUAL CRIME	SERIOUS INCIDENTS
STALKING	SERIOUS INCIDENTS
SUSPICION	MILD INCIDENTS
THEFT	SERIOUS INCIDENTS
TRESPASS	SERIOUS INCIDENTS
UTILITY INCIDENT	MILD INCIDENTS
WEAPON	SERIOUS INCIDENTS
WELL BEING CHECK	MILD INCIDENTS

Standardization of Time and Date:

The date and time for both Crime Data Portal and IIT Incident log datasets were not identical. They were all converted into Posixct time with tz as America/Chicago. When data was being converted to this format, the values of the time had changed i.e. the time in the actual incident report log did not match the time in our dataset. This was because the time was being read as tz = "Europe/London". Thus 5 hours(without daylight savings) needed to be added before converting to tz = "America/Chicago".

Cleaning Weather Dataset:

Cleaning and Standardization of Weather Data:

The weather dataset that we obtained had 21 attributes. 6 of these attributes had more than half missing values, while 2 of those were missing all of the values, we discarded these attributes. There were then about 6 out of 46,000 missing values, so we discarded these records from the dataset since there were so few. The other issue we found with the weather data is that the condition attribute had 37 different factors, and many of these were redundant. As seen in the figure below, there are conditions scattered clouds, partly cloudy, and mostly cloudy. In order to reduce the factors, we created 7 buckets labeled clear, clouds, fog, ice, rain, snow, and thunderstorms. The conditions typically come with descriptive before the condition as seen below. In order to try to not lose any information, we also created another set of 3 buckets for the descriptive words labeled low, medium, and high for the severity of the current condition. The problem with these buckets is that it was based solely on the severity of the description word and not on the condition with the descriptive word. For example mostly cloudy will be bucketed as high severity because of the word 'mostly', but this does not mean that the weather itself is severe in a dangerous sense. Because of this we should not use Severity in our analysis because they cannot be compared across different conditions.

<u>Feature Extraction:</u>

Weather Data:

After cleaning the weather data, we needed to merge the weather data with the crime data based on date and time. We decided that we did not need to merge the entire weather dataset with the crime data. The attributes of weather that we chose to merge were condition, standardized condition, severity, temperature, humidity and wind. We did add both the original condition label as well as the bucketed condition to see which would give better results. The problem with merging the data sets is that the weather data is taken at discrete random time intervals (typically every hour). This meant that we could not join directly because the time in the crime data and weather data may not match up. We determined that the best estimate for the weather at the exact time of the crime is the closest weather condition in time. We used manhattan distance between the times of the crimes and the times of the weather. Our implementation of this used a nested for loop with the crime data set being the outside for loop. We take the current date/time of the crime and calculate the closest weather record. If there are two weather records of the same distance, the first weather condition was used and we merged it with the crime data set.

COND	STAND_COND	SEVERITY	TEMP	HUM	WIND	PRECIP
Clear	Clear	Low	41.0	57	5.8	0.00
Partly Cloudy	Clouds	Low	33.1	78	4.6	0.00
Scattered Clouds	Clouds	Low	48.9	48	4.6	0.00
Mostly Cloudy	Clouds	High	52.0	43	9.2	0.00
Mostly Cloudy	Clouds	High	54.0	41	6.9	0.00
Mostly Cloudy	Clouds	High	53.1	43	4.6	0.00
Mostly Cloudy	Clouds	High	46.0	56	3.5	0.00
Mostly Cloudy	Clouds	High	50.0	50	8.1	0.00
Overcast	Clouds	Medium	51.1	50	12.7	0.00
Overcast	Clouds	Medium	51.1	50	12.7	0.00
Light Rain	Rain	Low	53.1	61	13.8	0.00
Overcast	Clouds	Medium	54.0	64	23.0	0.02

Day, Month, Time-Bucket

The day, month, time-bucket were extracted from the standardized date as shown below,

OCCURED	MONTH	DAY	TIME_BUCKET
2018-01-31 11:00:00	Jan	Wed	09-12
2018-02-28 19:45:00	Feb	Wed	18-21
2018-03-09 18:30:00	Mar	Fri	18-21
2017-04-30 16:15:00	Apr	Sun	15-18
2017-05-31 20:49:00	May	Wed	18-21
2017-06-30 16:07:00	Jun	Fri	15-18
2017-07-31 17:30:00	Jul	Mon	15-18
2017-08-31 23:00:00	Aug	Thu	21-00
2017-09-30 18:22:00	Sep	Sat	18-21
2017-10-31 16:30:00	Oct	Tue	15-18
2017-11-30 21:15:00	Nov	Thu	21-00
2017-12-31 19:00:00	Dec	Sun	18-21

Latitude, Longitude, Sector and GeoHash

From the address column, the latitude and longitude were extracted using the google API. Care was taken to ensure the latitude and longitude values that were beyond the bounds of our analysis (2 mile grid and 1 mile grid for Meis and Kent respectively) were converted to NA values. All NA values were then removed. The parking lot latitude and longitude were directly used as extraction of Latitude and Longitude from google API of the address for parking lot gave the values of the intersection rather than the actual point.

Once the Latitude and Longitude were extracted, sectors were created by equally dividing the bound on latitude, i.e., 41.8276 and 41.8421, into 4 equal parts and the bound on longitude i.e., -87.6330 and -87.6170, into 2 equal parts. A total of 8 sectors were thus created.

Geohashing is used for aggregating the latitude and longitude value in to a single hash value with the required precision. In this instance, we have used precision '7' to hash all the latitude and longitude details of the incidents, and is treated as a factor variable with 133 levels.

ADDRESS	LATITUDE	LONGITUDE	SECTOR	GeoHash
001XX W 35TH ST	41.83095	-87.63070	1	dp3tvzj
035XX S DR MARTIN LUTHER KING JR DR	41.83101	-87.61732	2	dp3typp
034XX S STATE ST	41.83238	-87.62667	3	dp3typ2
034XX S MICHIGAN AVE	41.83242	-87.62329	4	dp3typ6
032XX S WELLS ST	41.83585	-87.63283	5	dp3tvzg
032XX S PRAIRIE AVE	41.83569	-87.62064	6	dp3typu
029XX S FEDERAL ST	41.84147	-87.62920	7	dp3wjcn
030XX S MICHIGAN AVE	41.83881	-87.62326	8	dp3wn0d

Type of Data:

If it was IIT-AREA/IIT-CAMPUS.

Consolidation of Datasets:

The IIT Incident log data, Crime Portal data, Weather Data were combined with the following columns,

| LOCATION	ADDRESS	OCCURED	INCIDENT_TYPE1	LATITUDE	LONGITUDE	TYPE_OF_DATA	INCIDENT_TYPE2	MONTH	DAY	TIME_BUCKET	COND	STAND_COND	SEVERITY	TEMP	HUM	WIND	PRECIP	SECTOR	GeoHash
	030XX S WENTWORTH AVE	2017-12-01 19:52:00	WEAPON	41.83897	-87.63169	IIT-AREA	SERIOUS INCIDENTS	Dec	Wed	06-09	Light Snow	Snow	Low	16.0	81	10.4	0.00	8	dp3wn0d
olumn 1: numeric	035XX S PRAIRIE AVE	2017-12-02 07:00:00	OTHER	41.83012	-87.62053	IIT-AREA	MILD INCIDENTS	Dec	Wed	18-21	Light Snow	Snow	Low	21.9	78	13.8	0.00	7	dp3wjcn
ith range 1 - 7252 OTHER	032XX S DR MARTIN LUTHER KING JR DR	2017-12-02 11:03:00	THEFT	41.83336	-87.61701	IIT-AREA	SERIOUS INCIDENTS	Dec	Thu	03-06	Light Snow	Snow	Low	17.1	81	11.5	0.01	8	dp3wn14
STREET	030XX S MICHIGAN AVE	2017-12-02 12:30:00	THEFT	41.83946	-87.62323	IIT-AREA	SERIOUS INCIDENTS	Dec	Thu	12-15	Light Snow	Snow	Low	17.1	68	15.0	0.00	2	dp3typp
STREET	033XX S MICHIGAN AVE	2017-12-02 13:30:00	THEFT	41.83388	-87.62332	IIT-AREA	SERIOUS INCIDENTS	Dec	Thu	15-18	Blowing Snow	Snow	High	16.0	65	13.8	0.00	1	dp3tvzj
STREET	032XX S PRAIRIE AVE	2017-12-02 14:30:00	DAMAGE TO PROPERTY	41.83561	-87.62064	IIT-AREA	SERIOUS INCIDENTS	Dec	Thu	15-18	Blowing Snow	Snow	High	16.0	65	13.8	0.00	1	dp3tvzj
APARTMENT	0000X W 36TH ST	2017-12-02 20:00:00	BATTERY	41.82917	-87.62710	IIT-AREA	SERIOUS INCIDENTS	Dec	Thu	15-18	Blowing Snow	Snow	High	15.1	61	17.3	0.00	8	dp3wn1p
SIDEWALK	035XX S PRAIRIE AVE	2017-12-03 17:00:00	BATTERY	41.83059	-87.62054	IIT-AREA	SERIOUS INCIDENTS	Dec	Thu	21-00	Blowing Snow	Snow	High	3.9	64	12.7	0.00	7	dp3wjc5
illinois institute of technology: mies campus: residential buil...	3333 south wabash ave	2017-12-03 22:30:00	ALARM	41.83391	-87.62464	IIT-CAMPUS	MILD INCIDENTS	Dec	Thu	21-00	Clear	Clear	Low	3.0	60	11.5	0.00	1	dp3tvye

Exploratory Data Analysis:

- There was a class imbalance between Mild Incidents and Serious Incidents.

IIT Campus and Surrounding area Incidents by Frequency
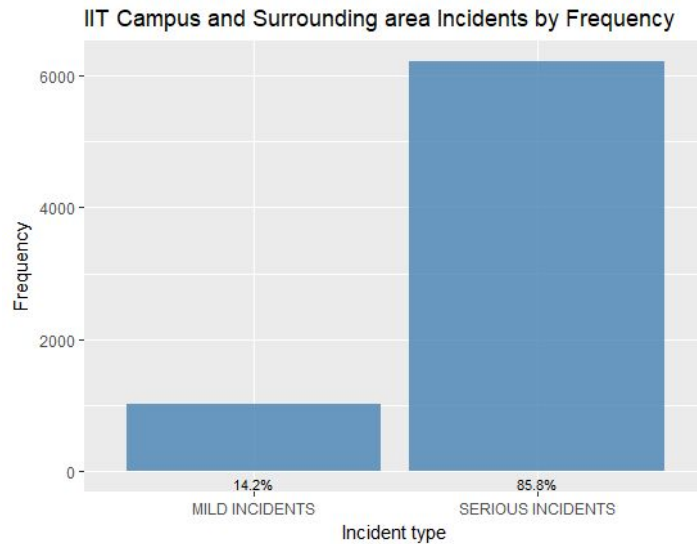
- Distribution of incidents across month of the year indicated that there were relatively low number of reported incidents in December and July, as it was Christmas Vacation and Summer Break


Frequency of Incident type by Month
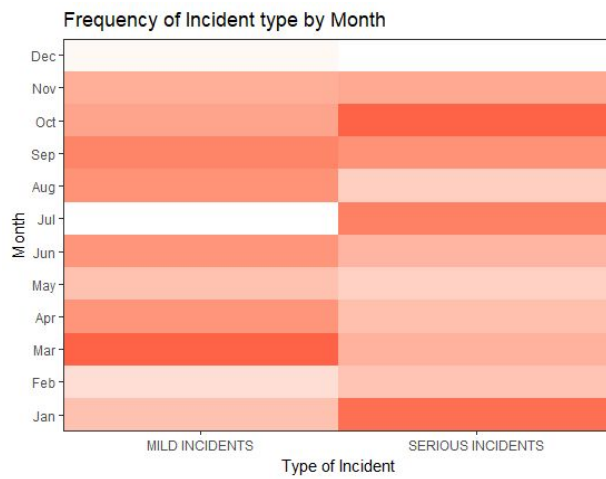
- Distribution of Incidents across the days of the week feature showed there was slightly lower incidents reported on weekends

IIT campus and Surrounding area Incidents by Days

Frequency of Incident type by Day of the Week
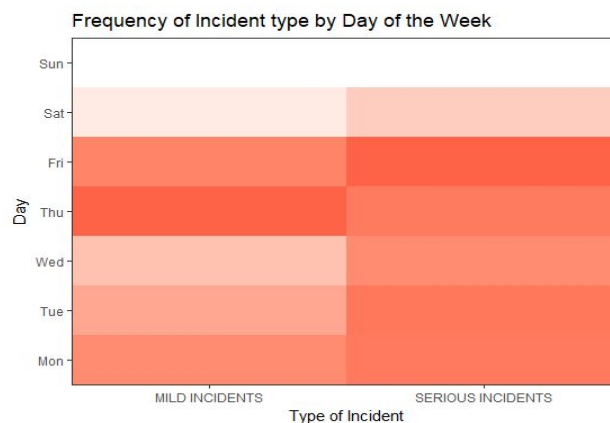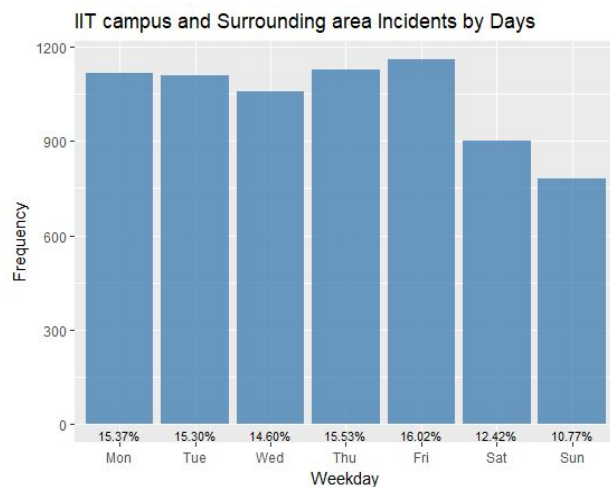
- Distribution of Incidents across the time buckets showed lower Incidents reported in 3:00 AM to 6:00 AM
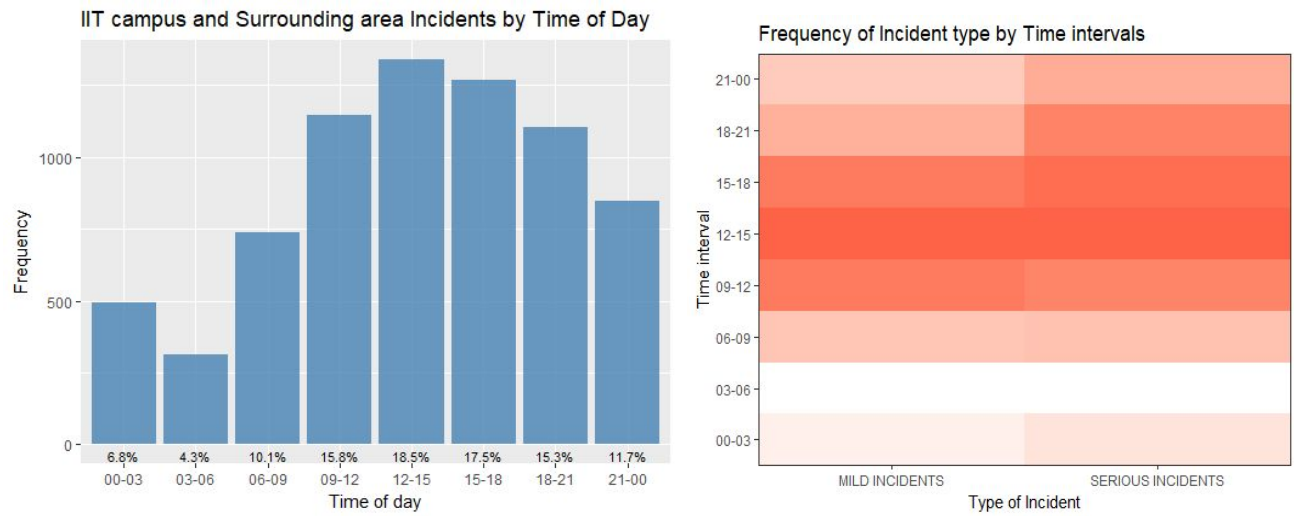
- Below are the histogram plots for weather conditions with number of incidents. The first plot from the left shows the top 5 original conditions and their frequency. The next shows the frequency after we bucketed the conditions. The last plot is a histogram for the temperature in degrees Fahrenheit.
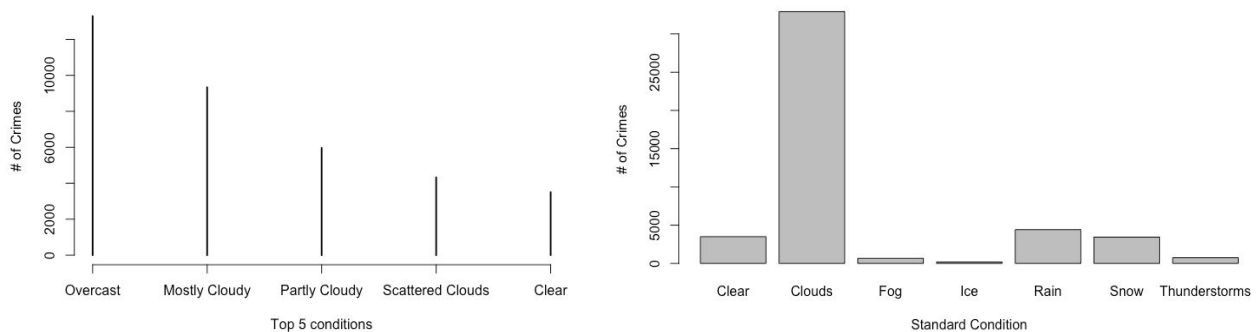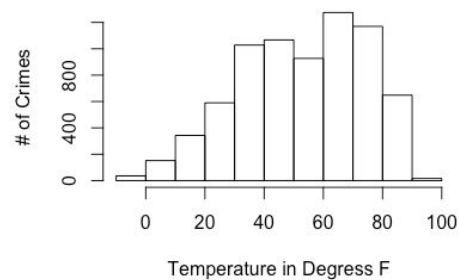
Predictive Model:

In order to answer the question of where to send a public safety officer to prevent a crime, we built many different prediction models that will predict the probability of a serious incident happening, given a date, time, and location. The prediction model will predict that a serious incident will happen if the probability is above .5 and a mild incident will happen if the probability is below .5. Mild incidents cannot typically be prevented, so we treated them as a serious incident not happening. In order to test our models, we trained the models on the first 80% of data by date and then tested on the later 20% of data. We split by date because we are trying to predict in the future, and this is a good way to simulate and test predicting in the future. We used 4 different modeling techniques to try to predict serious incidents: Naive Bayes, Decision Trees, Random Forests, and Logistic Regression. The results of the models are listed below.

Model	No Information Rate	Accuracy	Sensitivity	Specificity
Decision Trees	.85	.88	.46	.95
Random Forests	.85	.85	0	1.00
Naive Bayes	.85	.85	0	1.00
Logistic Regression	.85	.88	.46	.95

From this table, it can be easily seen that logistic regression is the only model that did not predict all serious incidents and have accuracy the same and no information rate. This is because we have so many serious incidents in comparison to mild incidents that the models want to always predict serious. This is impractical however because not all mild incidents will be reported and there are a limited number of public safety officers, so we cannot tell them to go everywhere, always. In order to fix this problem, we used the package ROSE (https://cran.r-project.org/web/pa ckages/ROSE/ROSE.pdf). This package allowed us to synthetically sample from our data in order to balance out the mild and serious incidents. We could then train our models on this data and test it on the original test data so that the models predict mild more often and not serious all the time. With the Rose package we got the following results:

Model	No Information Rate	Accuracy	Sensitivity	Specificity
Decision Trees	.85	.80	.70	.80

Random Forests	.85	.63	.42	.66
Naive Bayes	.85	.79	.71	.80
Logistic Regression	.85	.80	.69	.81

While our accuracy has fallen, we are getting models that are more realistic and will predict more specific areas that have a larger probability of a serious crime happening. We ended up deciding to choose the logistic regression model trained on the rose data as our model to predict serious crimes because it has one of the highest accuracies and it has the highest specificity. In order to deploy this model, we gave it a location, date, and time and it would give us a probability, we could then use this probability to create a heat map to show public safety officers.

Challenges Faced:

- Unstructured data from the IIT crime logs
- Ambiguous/incorrect classification of incident types
- Location Description instead of address
- Limit to calls in Weather and Google address API
- Redundant information from weather API
- Imbalance in Crime classes in IIT Area and Campus data

Deployment:

The application's goal is to provide a self-explanatory visual representation of incident's hotspots on campus based on historical data and for prediction of incidents.

Historical incidents map can be used to analyze the hotspot's locations and how they change through the time in order to make informed decisions in terms of incidents prevention and increase awareness of students which areas should be avoided.

Prediction of incidents provides potential incident's hotspots for 1, 3 and 5 days from the current date. Based on that information a recommendation to spend more time in the hotspot area during the shift could be given to a public safety officer.

**Application features (*app.r*)**

Working area splitted into side panel on the left, where user applies filters and display area to the right for visual representations with two tabs to view map with prediction and tab with historical data. In the future version, side panel will be splitted to separate filters that are applicable to prediction from filters used in historical representation.
Currently, filters in the green box used when working with incident prediction, blue box filters when plotting hotspots based on historical data.
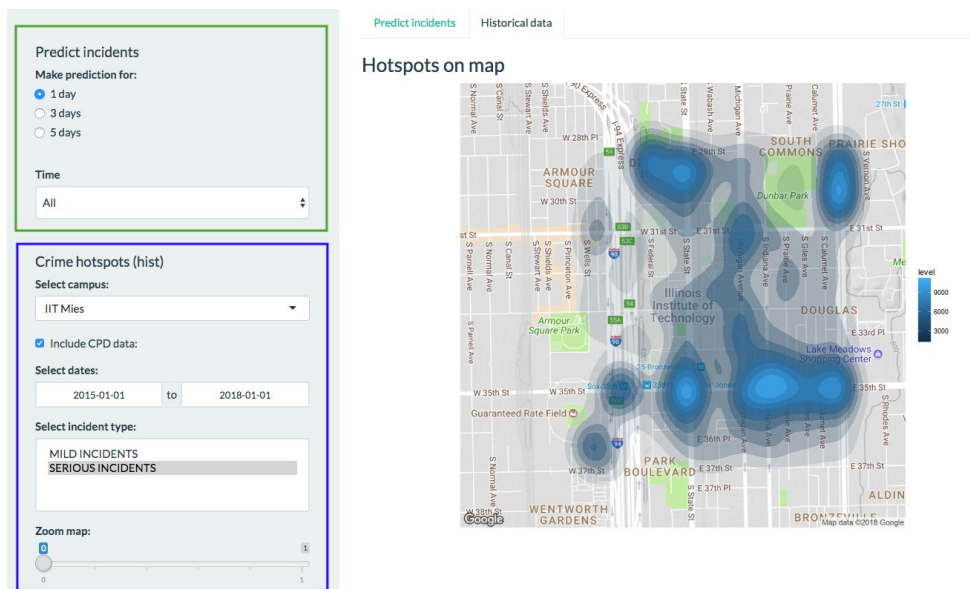
To view predictions, we choose 'Predict Incidents' tab, choose for what period of time to do a prediction (1,3 or 5 days). Also, user may apply a time filter, to view prediction for the specific time frame during the day, for instance, from 12 to 15. That can be useful to understand which areas need more public safety resources during the day.
Additionally, IIT campus area is splitted into 8 sectors, so that public safety patrols can easily split their zones of responsibility. All sectors are shown in red box below.
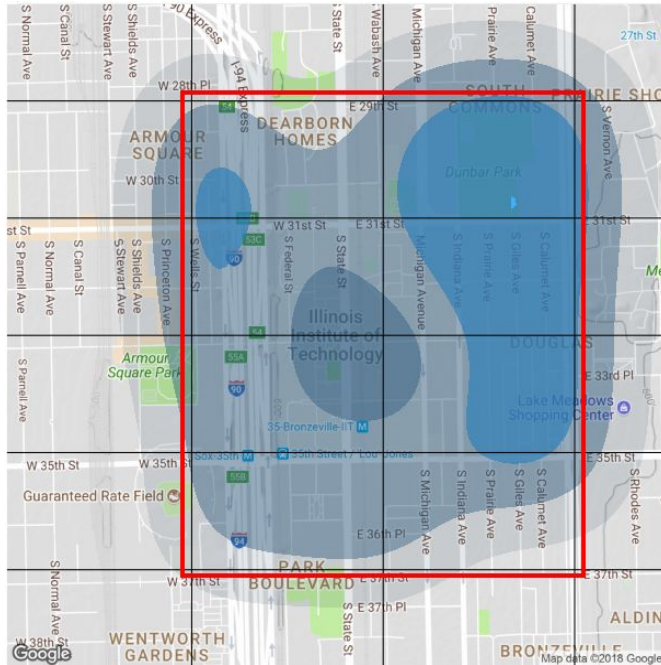
To view incident hotspots based on historical data, we choose tab 'historical data' and can apply the following filters:
- Select campus (IIT MIES or IIT KENT)
- 'Include CPD (Chicago Police Department) data' or use only IIT Public Safety incidents reports
- Set timeframe
- Select incident types to plot (multiple selection)
- Zoom in and out

Current limitations:
- Static maps. Dynamic maps planned for the next version of application
- Incident hotspots for IIT KENT can be viewed only when 'include CPD data' is checked, because IIT KENT is one building and all incidents reported to IIT Public Safety occurred in the building, not in the surrounding area, so density map can not be plotted
- Time filter can be applied only for prediction

Application available at https://mikhailry.shinyapps.io/CampusSafety/

Conclusion and Future Improvements:

The results of our studies were very useful in terms of identifying the crime hotspots within and around the campus, and identify the possible correlations of incidents with respect to time, weather

and location for predicting the locations for serious incidents in future. The location and time proved to be the most important indicators for prediction of serious crimes, and weather does not seem to play an important role considering the evidence at hand. The final standardized dataset and the analysis results should be extremely useful for IIT public safety to identify the areas of potential crime and effectively allocate their officers.

The future improvements for the study would be the following :
- A Standardardized dataset for logging, extraction and querying of incident data would be extremely helpful for future prediction modelling and studies.
- A Dashboard could be planned to be maintained for tracking the effectiveness of predictive modelling and corrective actions of public safety
- Adding new data points to the incident data like victim gender, and age would be beneficial for pursuing other crime related research questions.
- Currently, we have employed binary classification of incident types for simplicity of model building. Future studies should include Multi-class classification for incident types to add more granularity to the prediction of incidents.
- Extending the research to other university campuses in Chicago area to understand how the prediction accuracy of model varies across various campus locations.

References:

Chicago Data Portal. Crimes - 2001 to present | City of Chicago | Data Portal. (n.d.). Retrieved from https://data.cityofchicago.org/Public-Safety/Crimes-2001-to-present/ijzp-q8t2

Github. Mikhailry/campus_safety. (n.d.). Retrieved from https://github.com/Mikhailry/campus_safety.git

Rose Package. (2015). Retrieved from https://cran.r-project.org/web/pa ckages/ROSE/ROSE.pdf

Shinyapps. Mikhailry. (n.d). Retreived from https://mikhailry.shinyapps.io/CampusSafety/

Weather Underground API. (n.d.). Retrieved from https://www.wunderground.com/weather/api/

Weather Underground R API. (n.d.). Retrieved from https://cran.r-project.org/web/packages/rwund erground/README.html